Devetintrideseta
delavnica o telekomunikacijah

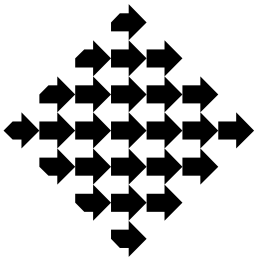
UPRAVLJANJE S PODATKI V DOBI UMETNE INTELIGENCE

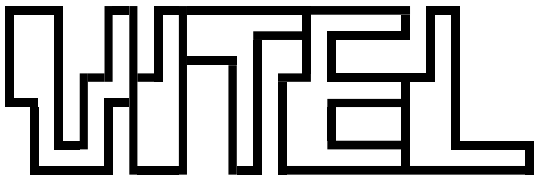
DATA MANAGEMENT IN THE AGE OF ARTIFICIAL INTELLIGENCE

16. in 17. maja 2024 Bled

Slovensko društvo za elektronske komunikacije
Elektrotehniška zveza Slovenije

SLOVENSKO DRUŠTVO ZA ELEKTRONSKE KOMUNIKACIJE
ELEKTROTEHNIŠKA ZVEZA SLOVENIJE





Devetintrideseta delavnica o telekomunikacijah

39th Workshop on Telecommunications

UPRAVLJANJE S PODATKI V DOBI UMETNE INTELIGENCE

DATA MANAGEMENT IN THE AGE OF ARTIFICIAL INTELLIGENCE

ZBORNIK REFERATOV

PROCEEDINGS

16. in 17. maja 2024 Bled, Slovenija

© 2024
Slovensko društvo za elektronske komunikacije
Elektrotehniška zveza Slovenije
Stegne 7
1521 Ljubljana, Slovenija
www.drustvo-sikom.si

39. delavnica o telekomunikacijah VITEL

ZBORNIK REFERATOV

39th Workshop on Telecommunications VITEL

PROCEEDINGS

Organizirata / *Organised by*: Slovensko društvo za elektronske komunikacije

Elektrotehniška zveza Slovenije

Pokrovitelj / *Sponsored by:* IEEE Communications Society

Uredil / *Editor*: Tomi Mlinar

Naslovnica / *Cover design:* Nikolaj Simič, Filip Samo Balan, Aleksander Vreže

Izdajatelj / *Publisher:* Slovensko društvo za elektronske komunikacije

ISSN 1581–6737

Programski in organizacijski odbor delavnice

Programme and organizing committee

Andrej Košir

Boštjan Batagelj

Ivica Kranjčević

Nikolaj Simič

Tomi Mlinar

Vesna Prodnik

Zgodovina delavnic o telekomunikacijah VITEL

History of Workshops on Telecommunications VITEL

1993: 1. *ISDN omrežja in storitve v Sloveniji*, Brdo pri Kranju
1994: 2. *Mobilne in brezvrvične telekomunikacije*, Brdo pri Kranju
1995: 3. *Podatkovna omrežja in storitve v Sloveniji*, Brdo pri Kranju
1995: 4. *Načrtovanje, upravljanje in vzdrževanje komunikacijskih omrežij*, Brdo pri Kranju
1997: 5. *Varnost in zaščita v telekomunikacijskih omrežjih*, Brdo pri Kranju
1997: 6. *Zbliževanje fiksnih in mobilnih omrežij ter storitev*, Brdo pri Kranju
1998: 7. *Telekomunikacije in sprejetje Slovenije v Evropsko unijo*, Brdo pri Kranju
1999: 8. *Omrežja IP, internet, intranet, ekstranet*, Brdo pri Kranju
1999: 9. *Upravljanje omrežij in storitev*, Brdo pri Kranju
2000: 10. *Mobilnost v telekomunikacijah*, Brdo pri Kranju
2001: 11. *Dostop do telekomunikacijskih storitev*, Brdo pri Kranju
2002: 12. *Poslovne telekomunikacije*, Ljubljana
2002: 13. *Kakovost storitev*, Brdo pri Kranju
2003: 14. *Varnost v telekomunikacijskih sistemih*, Brdo pri Kranju
2003: 15. *Mobilni internet*, Brdo pri Kranju
2004: 16. *Pametne stavbe*, Brdo pri Kranju
2005: 17. *Telefonija IP (VoIP)*, Brdo pri Kranju
2005: 18. *Storitev trojček = Triple play*, Ljubljana
2007: 19. *Brezžični širokopasovni dostop*, Brdo pri Kranju
2007: 20. *Optična dostopovna omrežja*, Brdo pri Kranju
2008: 21. *Povsem IP–omrežja*, Brdo pri Kranju
2009: 22. *Širokopasovna mobilna omrežja*, Brdo pri Kranju
2009: 23. *Konvergenčne storitve v mobilnih in fiksnih omrežjih*, Brdo pri Kranju
2010: 24. *Prehod na IPv6*, Brdo pri Kranju
2011: 25. *Internet stvari*, Brdo pri Kranju
2011: 26. *Komunikacije in računalništvo v oblaku*, Brdo pri Kranju
2012: 27. *Telekomunikacije in zasebnost,* Brdo pri Kranju
2012: 28. *Pametna mesta,* Brdo pri Kranju
2013: 29. *Infrastruktura za izpolnitev digitalne agende in kaj po tem – primer Slovenije;* Brdo pri Kranju
2014: 30. *Omrežja prihodnosti,* Brdo pri Kranju
2015: 31. *Kritična infrastruktura in IKT,* Brdo pri Kranju
2016: 32. *Pametna omrežja informacijske družbe,* Brdo pri Kranju
2017: 33. *Omrežja 5G za digitalno preobrazbo,* Brdo pri Kranju
2018: 34. *Zaupanja vreden internet,* Brdo pri Kranju
2019: 35. *Uporabna vrednost interneta vsega,* Brdo pri Kranju
2021: 36. *Vloga tehnologije 5G v vertikalah in vloga vertikal v omrežju 5G, Zoom*
2022: 37. *Povečanje odpornosti kritične infrastrukture z uporabo naprednih rešitev IKT, Zoom*
2023: 38. *Oblačno zasnovana omrežja, Bled*

Zgodovina mednarodnih simpozijev VITEL

History of International Telecommunication Symposium VITEL

1992: *VITEL*, Ljubljana
1994: *Subscriber Access*, Ljubljana
1996: *Broadband Communications Prospects and Applications*, Ljubljana
1998: *Mobility and Convergence Communication Technologies*, Ljubljana
2000: *Technologies and Communication Services for the Online Society*, Ljubljana
2002: *NGN and Beyond*, Portorož
2004: *Next Generation User*, Maribor
2006: *Content and Networking*, Ljubljana
2008: *DVB–T and MPEG4*, Bled
2010: *Digital Television Switchover Process*, Brdo pri Kranju

Uvodnik

Tema 39. delavnice o telekomunikacijah z naslovom Upravljanje s podatki v dobi umetne inteligence izhaja iz vloge podatkov v razvoju sodobnih tehnologij na različnih področjih, ne le umetne inteligence. Umetna inteligenca, kot radi imenujemo sodobne rešitve strojnega učenja in statističnega modeliranja, je vstopila iz tehnoloških in raziskovalnih krogov v vsakdanje življenje in tudi vsakdanje novice. Današnja odmevnost tem s tega področja na družbenih omrežjih je bila pred nekaj leti povsem nepredstavljiva. Globoko učenje, generativni modeli in veliki jezikovni modeli so desetletja naporov na področju računalniških in avtomatiziranih rešitev sestavili v uporabne rešitve. Slabi dve desetletji dolgo obdobje spletnih iskalnikov in Googlovih tehnologij sedaj zamenjava ChatGPT in sorodne rešitve. Rezultati so hkrati fascinantni in vsaj delno napačni. Na eni strani so rešitve, ki jih daje npr. ChatGPT, polne napak in tudi nesmislov, a realen prihranek časa do uporabnega rezultata je vendarle očiten. Boljše rešitve in rešitve, ki so boljše od človeških, so le vprašanje kratkega časa.

Rešitve umetne inteligence na posameznih področjih, od prava in ekonomije do umetnosti že prinašajo velike spremembe v delovnih procesih in posledično v družbi. Uporabnost rešitev po področjih je pravi korak naprej.

Glavni izziv na področju umetne inteligence niso več le izboljšani algoritmi, ampak legalno dostopni, dovolj zanesljivi podatki za učenje modelov strojnega učenja. Razvoj tako torej ni več le tehnološki, ampak tudi pravni in etični izziv. Gledano širše, pametne tehnologije danes neobhodno vključujejo občutljive podatke. Ker pravniki in etiki v preteklosti niso mogli reševati problemov, ki jih še ni bilo, je potreben vzporeden razvoj tehnoloških in pravno etičnih vidikov. Razvoj avtonomno vozečih vozil je tehnološko blizu rešitve, pravno pa še zelo daleč. Nekatera etična načela, ki se dotikajo oskrbe starejših, so očitno neetična in potrebujemo nove rešitve na področju etike.

Sestava programa naše delavnice sledi zgornjim izhodiščem. V prvi sekciji prvega dne delavnice osvetlimo tehnološke trende z Markom Štefančičem (Gartner), zaupanje v podatke s Samom Zorcem (MDP), varstvo zasebnosti z Evo Kalan Logar (Informacijska pooblaščenka) in odprto znanost z Uršo Opara Kraševec (Fakulteta za elektrotehniko UL). V drugi sekciji nadaljujemo s podatkovnimi skladišči. Svoje poglede bodo predstavili Robert Korošec (ORCLE), Rok Rogelj (Microsoft) in Katarina Gašperlin Stepančič (IBM). Pogled dobaviteljev opreme daje tretja sekcija s tremi predavanji, Ivanom Lesićem (Nokia), Dejanom Rutolo (Huawei) in Srđanom Kneževićem (Ericsson). Teme prvega dneva bomo soočili v okrogli mizi, ki jo bo vodila Vesna Prodnik (SIKOM) z udeleženci Matjažem Beričičem, Zoranom Vehovarjem, Samom Zorcem in Mitjo Trampužem.

Drugi dan delavnice bo prva sekcija posvečena virom podatkov in zaščiti, osvetlili jo bodo Ajda Brlec (Telekom), Mitja Luštrek (IJS) ter Matej Rabzelj (Fakulteta za elektrotehniko UL). Druga sekcija sledi uporabni vrednosti podatkov po področjih. Izbrali smo pametni dom (Mark Umberger, ENTIA), poenotenje podatkov (Mateja Sajovic) ter podatke v visokošolskem izobraževanju (Jože Rugelj, Pedagoška fakulteta UL). V tretji sekciji se bomo posvetili še podatkom in sistemom, kjer bodo uporabo umetne inteligence v tovarnah prihodnosti predstavil Urban Zaletel (Kontron), trajnostni mobilni sistem Janez Sterle (Internet Institut) ter Digitalni dvojček Blaž Peternel (Smartis).

Menimo, da preplet tehnoloških vidikov, pravnih vidikov in uporaba posameznih področij osvetli kar se da kompletno sliko vloge podatkov in umetne inteligence v smislu trendov in uporabnosti.

dr. Andrej Košir, Bled, 16. maja 2024
vodja programskega odbora VITEL 2024

Foreword

The topic of the 39th Workshop on Telecommunications VITEL 2024 entitled Data Management in the Age of Artificial Intelligence stems from the role of data in the development of modern technologies in various fields, not only artificial intelligence (AI). Artificial intelligence, as we like to call modern solutions of machine learning and statistical modeling, has entered from technological and research fields into everyday life and even everyday news. Today's resonance of topics in this area on social networks was completely unimaginable a few years ago. Deep learning, generative models, and large language models have combined decades of computing and automation efforts into useful solutions. The nearly two-decade-long era of web search engines and Google technologies is now being gradually replaced by ChatGPT and related solutions. The results are both fascinating and at least partially wrong. On the one hand, there are solutions given by e.g. ChatGPT, full of errors and nonsense, but the real saving of time to a useful result is nevertheless obvious. Better and better-than-human solutions are only a matter of time.

Artificial intelligence solutions in individual fields, from law and economics to art, are already bringing about major changes in work processes and, consequently, in society. The applicability of solutions field by field is a step forward in the development.

The main challenge in the field of artificial intelligence is no longer just improved algorithms, but legally available reliable enough data to train machine learning models. Development is thus no longer only a technological, but also a legal and ethical challenge. Broadly speaking, smart technologies today inevitably include sensitive data. Since lawyers and ethicists in the past could not solve problems that did not exist yet, the parallel development of technological and legal-ethical aspects is necessary. The development of autonomous vehicles is technologically close to a solution, but legally it is still a long way off. Some ethical principles affecting the care of the elderly are clearly unethical and we need new solutions in the field of ethics.

The structure of VITEL workshop program follows the above starting points. In the first section of the first day of workshop we shed light on technological trends with Marko Štefančič (Gartner), trust in data with Sam Zorc (MDP), privacy protection with Eva Kalan Logar (Information Commissioner) and open science with Urša Opara Kraševec (Faculty of Electrical Engineering UL). In the second section, we continue with data warehouses. Robert Korošec (ORCLE), Rok Rogelj (Microsoft) and Katarina Gašperlin Stepančič (IBM) will present their views. The view of vendors is covered in the third section with Ivan Lesić (Nokia), Dejan Rutola (Huawei) and Srđan Knežević (Ericosson). The topics of the first day will be discussed at a round table chaired by Vesna Prodnik (SIKOM) with participants Matjaž Beričič, Zoran Vehovar, Sam Zorc and Mitja Trampuž.
On the second day of the workshop, the first section will be devoted to data sources and protection, and will be highlighted by Ajda Brlec (Telekom), Mitja Luštrek (IJS) and Matej Rabzelj (Faculty of Electrical Engineering UL). The second section tracks the utility value of the data by selected fields. We chose the smart home (Mark Umberger, ENTIA), unification of data (Mateja Sajovic) and data in higher education (Jože Rugelj, Faculty of Education UL). In the third section, we will also focus on data and systems, where AI inn the factories of the future will be presented by Urban Zaletel (Kontron), the sustainable mobile system by Janez Sterle (Internet Institute) and the Digital Twin by Blaž Peternel (Smartis).

We believe that the interweaving of technological aspects, legal aspects and the use of data in individual fields illuminates as complete a picture as possible of the role of data and artificial intelligence in terms of trends and usability.

dr. Andrej Košir, *Bled, May 16th, 2024*
Head of the VITEL 2024 Programme Committee

Kazalo prispevkov

Table of Contents

16. 5. 2024

17. 5. 2024

PRISPEVKI

ARTICLES

16. 5. 2024

Gartnerjevi ključni tehnološki trendi za leto 2024

Gartner Top Strategic Technology Trends for 2024s

Marko Štefančič

Gartner Slovenija & Central Asia

POVZETEK

V obdobju gospodarske in družbene nestabilnosti so organizacije lahko uspešne z naravnanostjo k trajnostni učinkovitosti. Podjetja tako morajo prepoznati, kako najbolje izkoristiti nove tehnologije, ki bi lahko v naslednjih letih povzročile znatne pretrese, ter hkrati ostati odporna na zunanje nestabilnosti in prilagodljiva za nove modele dela.

To predavanje bo ponudilo vpogled v najpomembnejše strateških tehnoloških trende za leto 2024 in predstavilo ključna vodila CIO-tom in IT strokovnjakom:

- prednost dajte trajnostnim tehnološkim trendom, uskladite praktično kibernetsko varnost z vrednostjo poslovanja,
- investirajte v tehnologijo, ki ustreza specialističnim delavcem in z umetno inteligenco podprtemu razvoju programske opreme,,
- vključite stroje kot stranke, razširitev in umetno inteligenco v izobraževalne in delovne procese.

SUMMARY

During economic and social volatility, organizations achieve success through sustainable efficiency. Organizations must determine how to make the best possible use of new technology developments that could cause significant disruption over the next three years, such as generative AI, and embrace the connected workforce nature while remaining resilient. This presentation will provide overview of Gartner Top Strategic Technology Trends for 2024 and how CIOs and IT leaders can use them to ensure future success:

- Prioritize sustainable technology trends, aligning practical cybersecurity with business value
- Invest in technology that fits specialist workers and AI-powered software development.
- Incorporate machine customers, augmentation, and AI-powered intelligence in training and work processes

---

O AVTORJU

Marko Štefančič je regijski direktor in principal Gartnerja za Slovenijo in Centralno Azijo, kjer je zadolžen za upravljanje poslovanja, strateško načrtovanje ter sodelovanje s ključnimi strankami in uporabniki Gartnerjevih storitev v regiji. Na trgu informacijskih tehnologij in rešitev ima Marko Štefančič več kot 25 let bogatih izkušenj, tako z vodenjem in upravljanjem, marketingom, svetovanjem in prodajo. Je tudi avtor večjega števila člankov in prispevkov o IKT ter soavtor knjige. Za uspešno delo v Gartnerju je prejel več nagrad in priznanj.

ABOUT THE AUTHOR

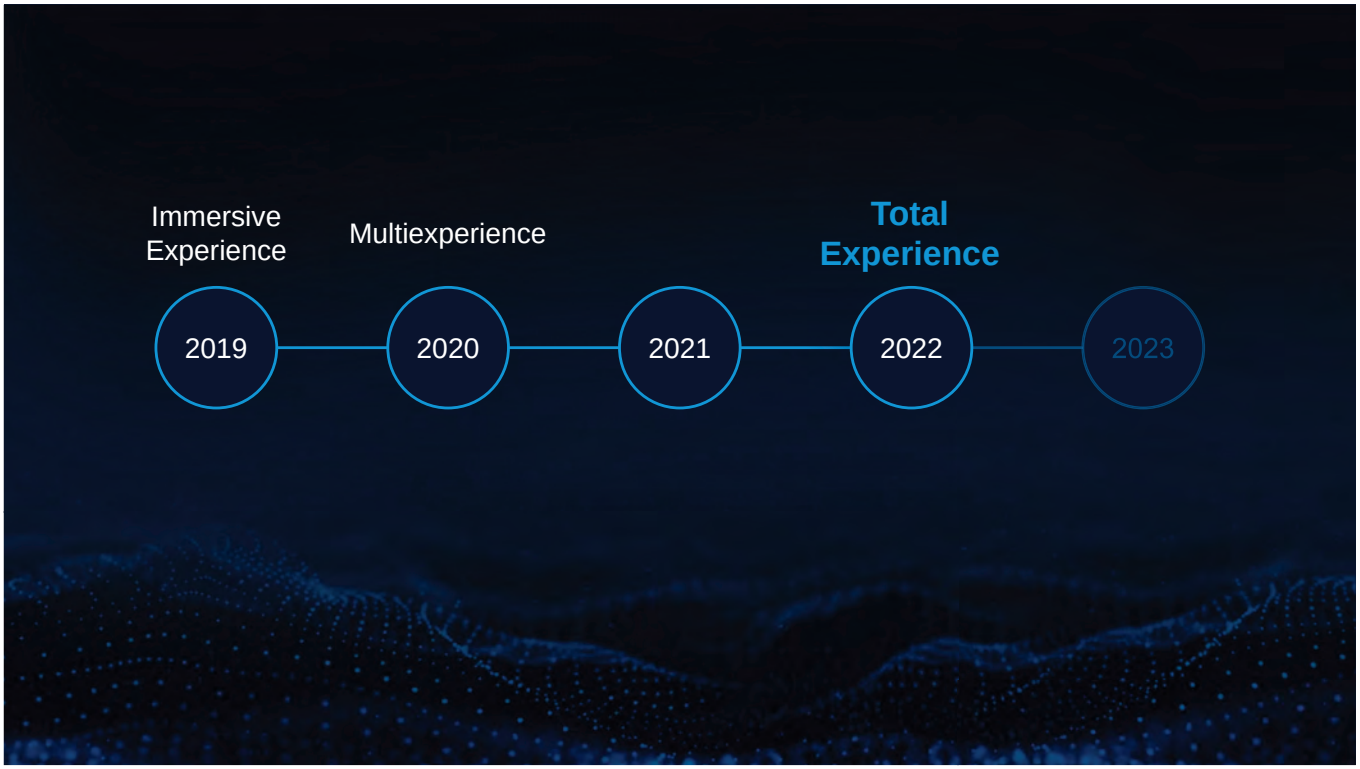
Marko Štefančič is the regional director and principal of Gartner for Slovenia and Central Asia, where he is responsible for business management, strategic planning and cooperation with key customers and users of Gartner's services in the region. Marko Štefančič has more than 25 years of rich experience in the market of information technologies and solutions, both in leadership and management, marketing, consulting and sales. He is also the author of a large number of articles and papers on ICT and co-author of a book. He received several awards and recognitions for his successful work at Gartner.

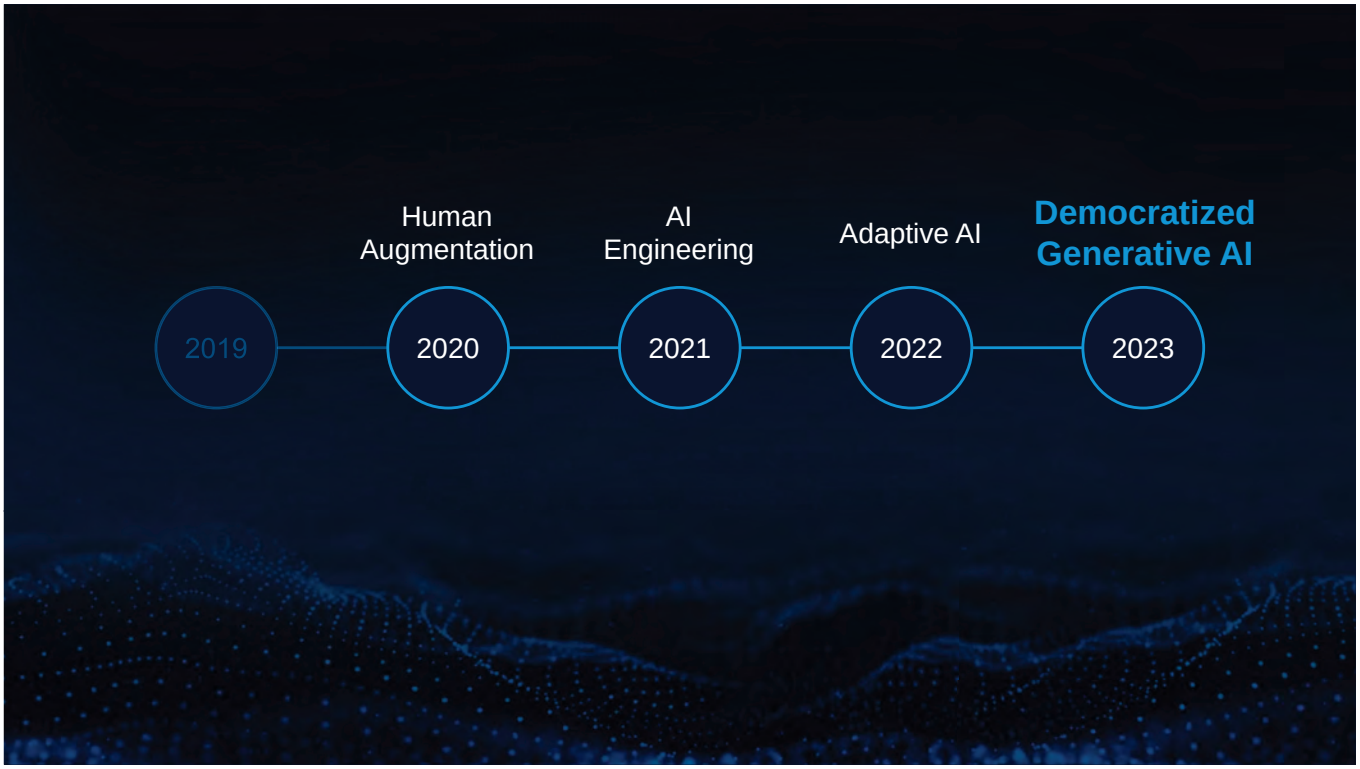

Gartner
VITEL
Top Strategic Technology Trends for 2024
Marko Štefančič
2024

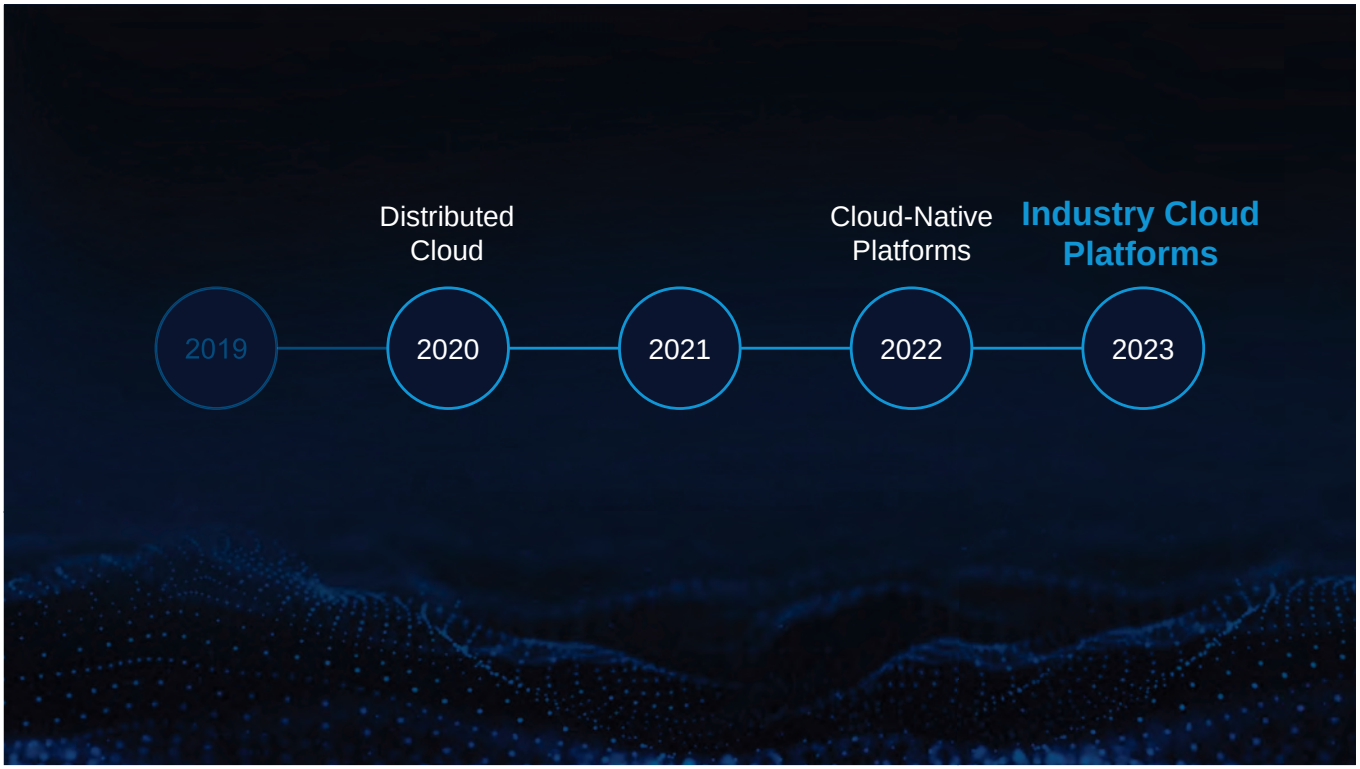

0 7 3 7
1 8 4 8
2 9 5 9
Two Centuries Ago


Vision


No One Is Ready for the “Future ... ”


Immersive Experience
Multiexperience
Total Experience
2019
2020
2021
2022
2023


Human Augmentation
AI Engineering
Adaptive AI
Democratized Generative AI
2019
2020
2021
2022
2023


Distributed Cloud
Cloud-Native Platforms
Industry Cloud Platforms
2019
2020
2021
2022
2023


2019
2020
2021
2022
2023
The road forward should fit you.

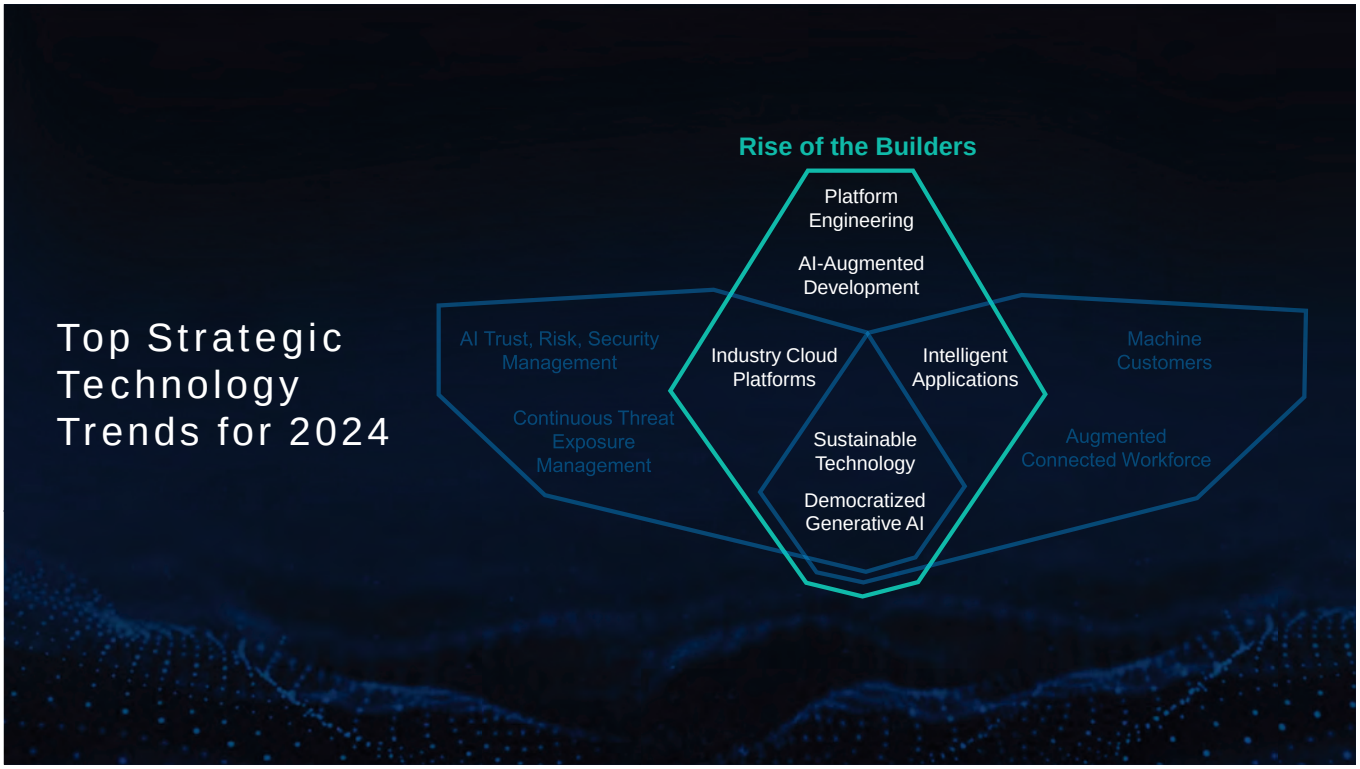

Everything Is Trending,
Everywhere, All at Once


Autocode
No-Code
Digital Arbitrage
AI Risk
AI-Augmented Development
Superapps
Digital Immune System
Platform Engineering
Tiny IT
Digital Nationalism
Generative AI
Digital National Infrastructure
Digital Spending Governor
Connected Employee
Industry Cloud Platforms
Democratized Generative AI
Connected Applications
Core Business Metrics
Business Operations Intelligence
Embedded AI
Metaverse
Low-Code
Omnipresent Cloud
AI Trust, Risk, Security Management
Intuitive Contextualization
Process Mining
Software Supply Chain Security
Connected Manager
Machine Customers
Augmented Connected Workforce
Sustainable Technology
Operational Experience
Process Alchemy
Wireless Platforms
Converged Composable Ecosystems
Adaptive AI
Rise of Builders
Customer Technology Platforms
Crowdsourcing
Digital Employee Experience
Intelligent Applications
Quantum
Applied Observability
Digital Foundations
Continuous Threat Exposure Management


We Help You
Cut Through Noise
Platform Engineering
AI-Augmented Development
AI Trust, Risk, Security Management
Industry Cloud Platforms
Intelligent Applications
Machine Customers
Continuous Threat Exposure Management
Sustainable Technology
Augmented Connected Workforce
Democratized Generative AI


Top Strategic Technology Trends for 2024
Rise of the Builders
Protect Your Investment
Deliver the Value
Platform Engineering
AI-Augmented Development
AI Trust, Risk, Security Management
Industry Cloud Platforms
Intelligent Applications
Machine Customers
Continuous Threat Exposure Management
Sustainable Technology
Augmented Connected Workforce
Democratized Generative AI


Top Strategic Technology Trends for 2024
Protect Your Investment
AI Trust, Risk, Security Management
Continuous Threat Exposure Management
Industry Cloud Platforms
Platform Engineering
AI-Augmented Development
Intelligent Applications
Machine Customers
Sustainable Technology
Democratized Generative AI
Augmented Connected Workforce


Top Strategic Technology Trends for 2024
Rise of the Builders
AI Trust, Risk, Security Management
Continuous Threat Exposure Management
Industry Cloud Platforms
Platform Engineering
AI-Augmented Development
Intelligent Applications
Machine Customers
Sustainable Technology
Democratized Generative AI
Augmented Connected Workforce

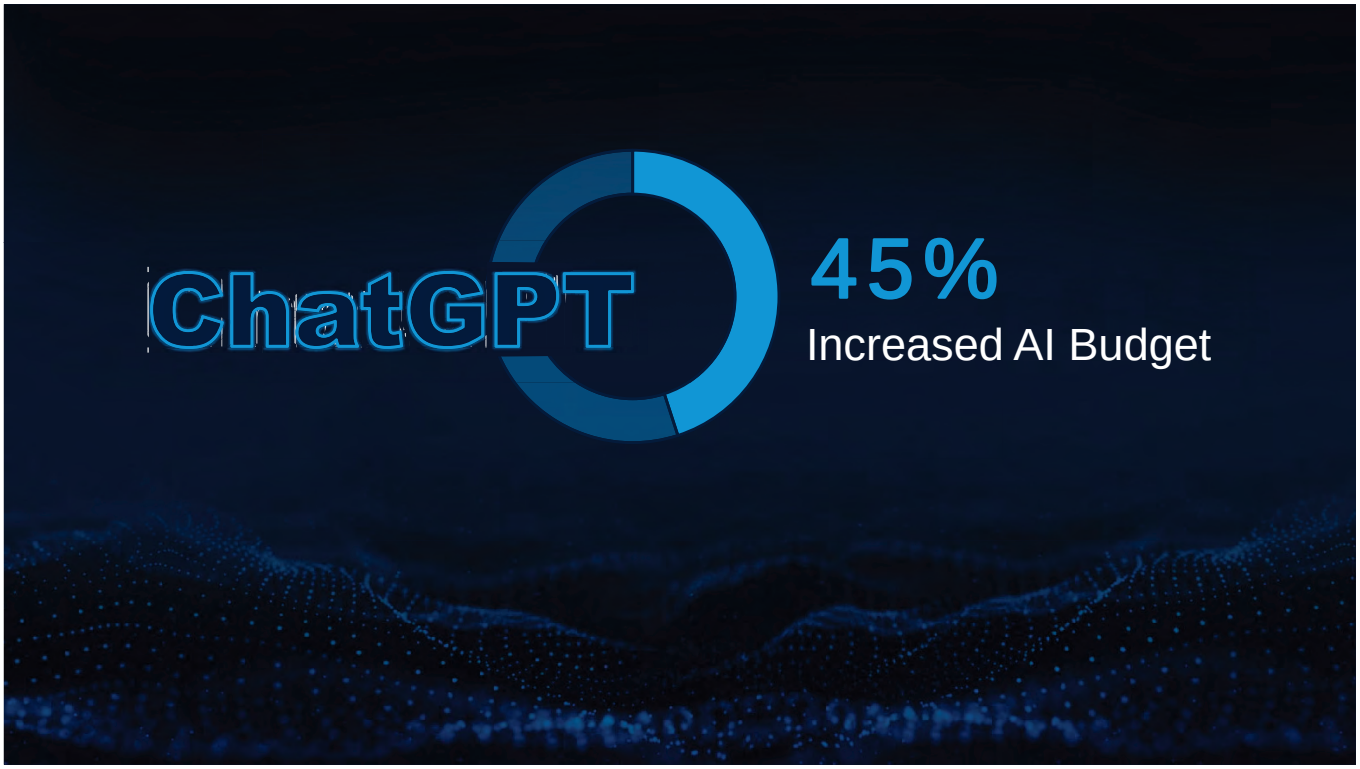

Top Strategic Technology Trends for 2024
Platform Engineering
AI-Augmented Development
AI Trust, Risk, Security Management
Industry Cloud Platforms
Intelligent Applications
Deliver the Value
Machine Customers
Continuous Threat Exposure Management
Sustainable Technology
Augmented Connected Workforce
Democratized Generative AI

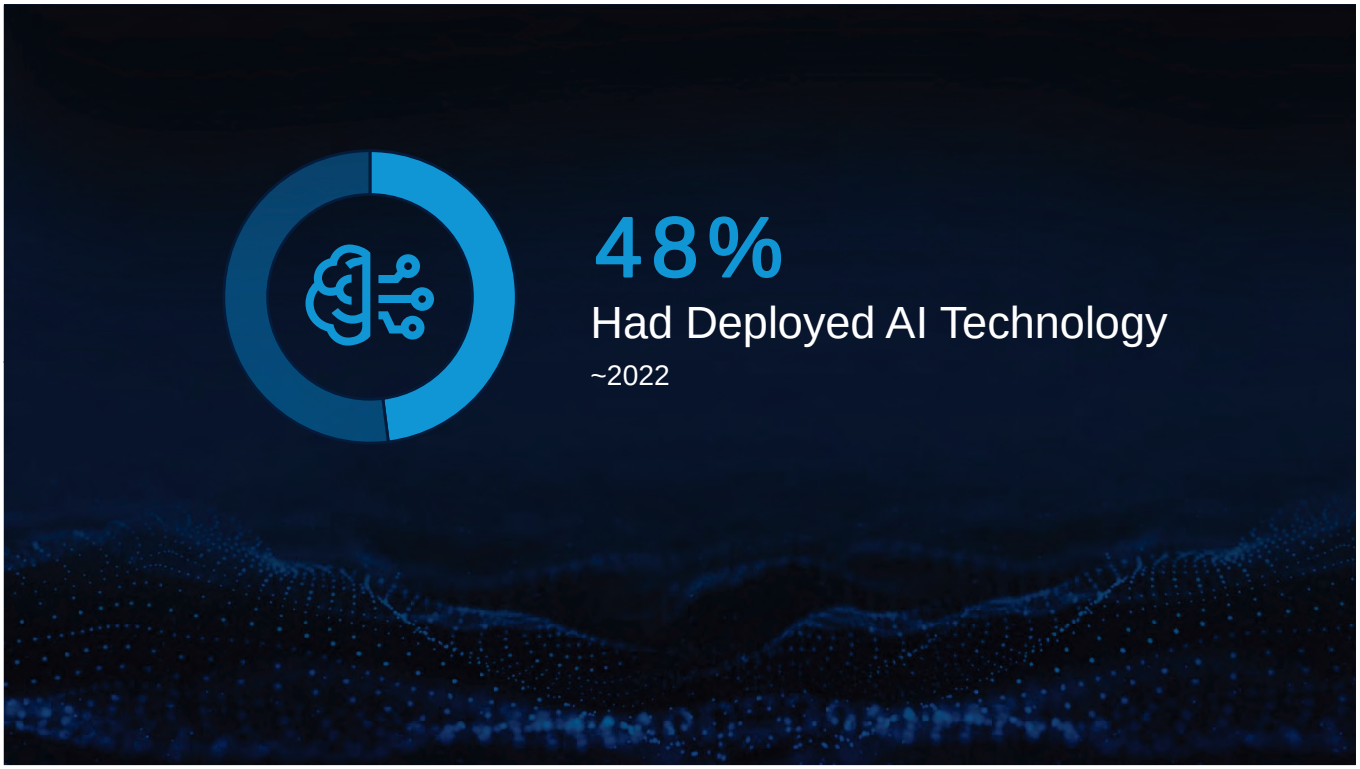

Protect Your Investment
Platform Engineering
AI-Augmented Development
AI Trust, Risk, Security Management
Industry Cloud Platforms
Intelligent Applications
Machine Customers
Continuous Threat Exposure Management
Sustainable Technology
Augmented Connected Workforce
Democratized Generative AI

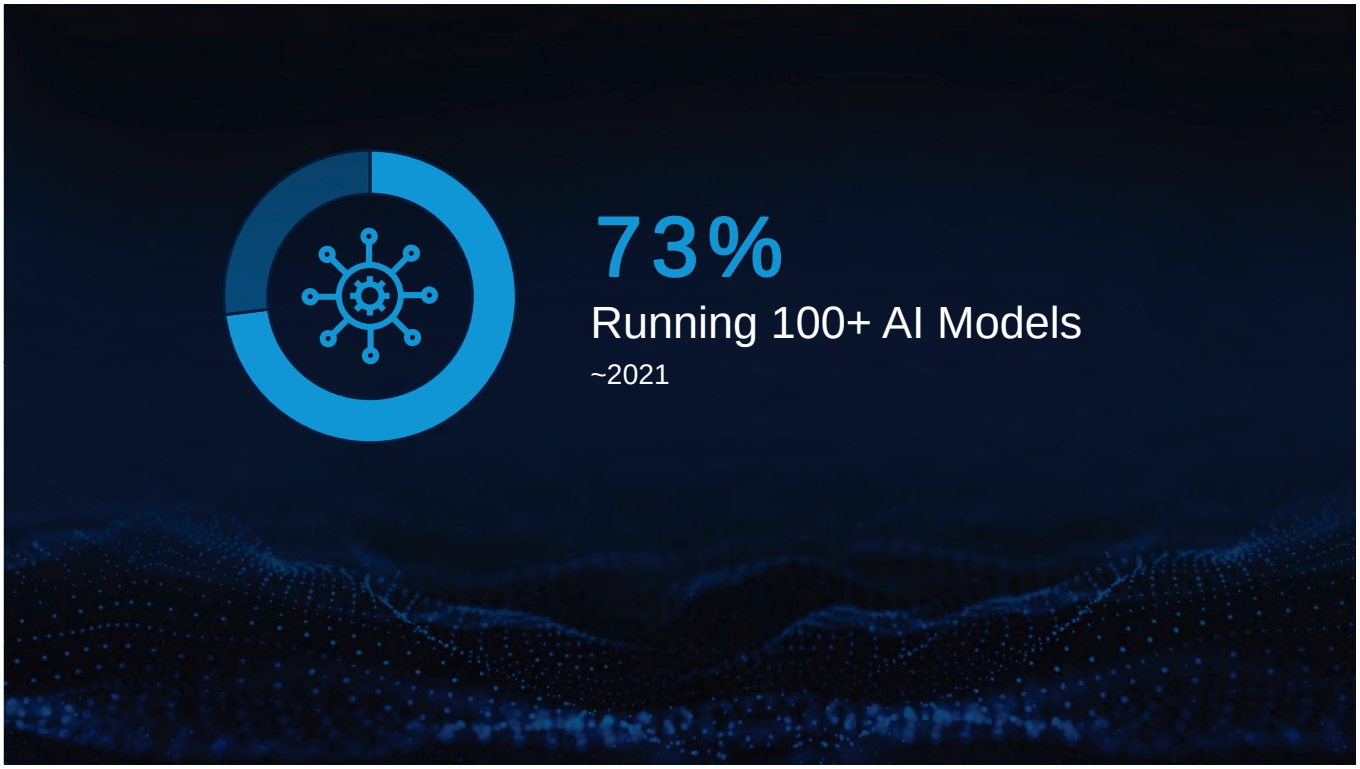

1
AI as Partner


1
AI as Partner
AI Trust, Risk,
Security Management

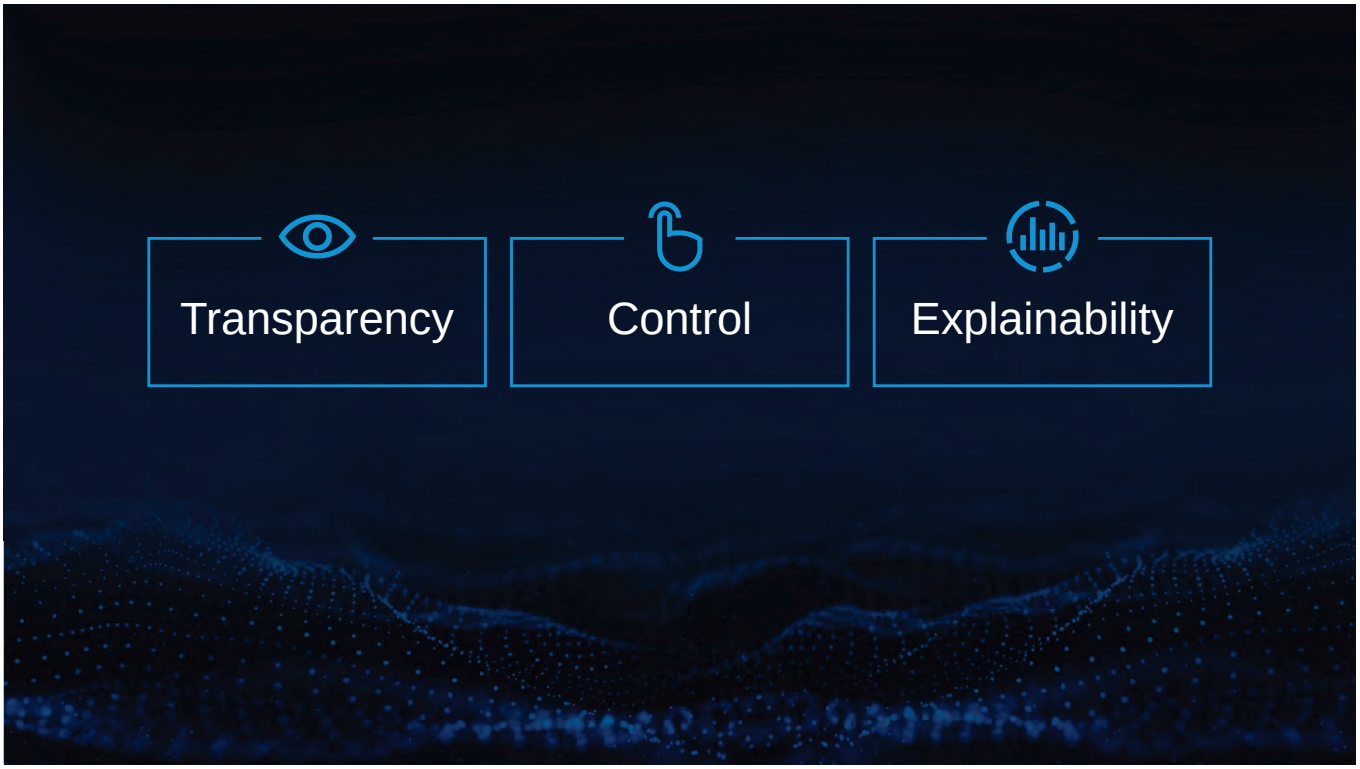

ChatGPT


ChatGPT
45%
Increased AI Budget


48%
Had Deployed AI Technology
~2022


73%
Running 100+ AI Models
~2021


40%
Suffered From Privacy
or Security Issues

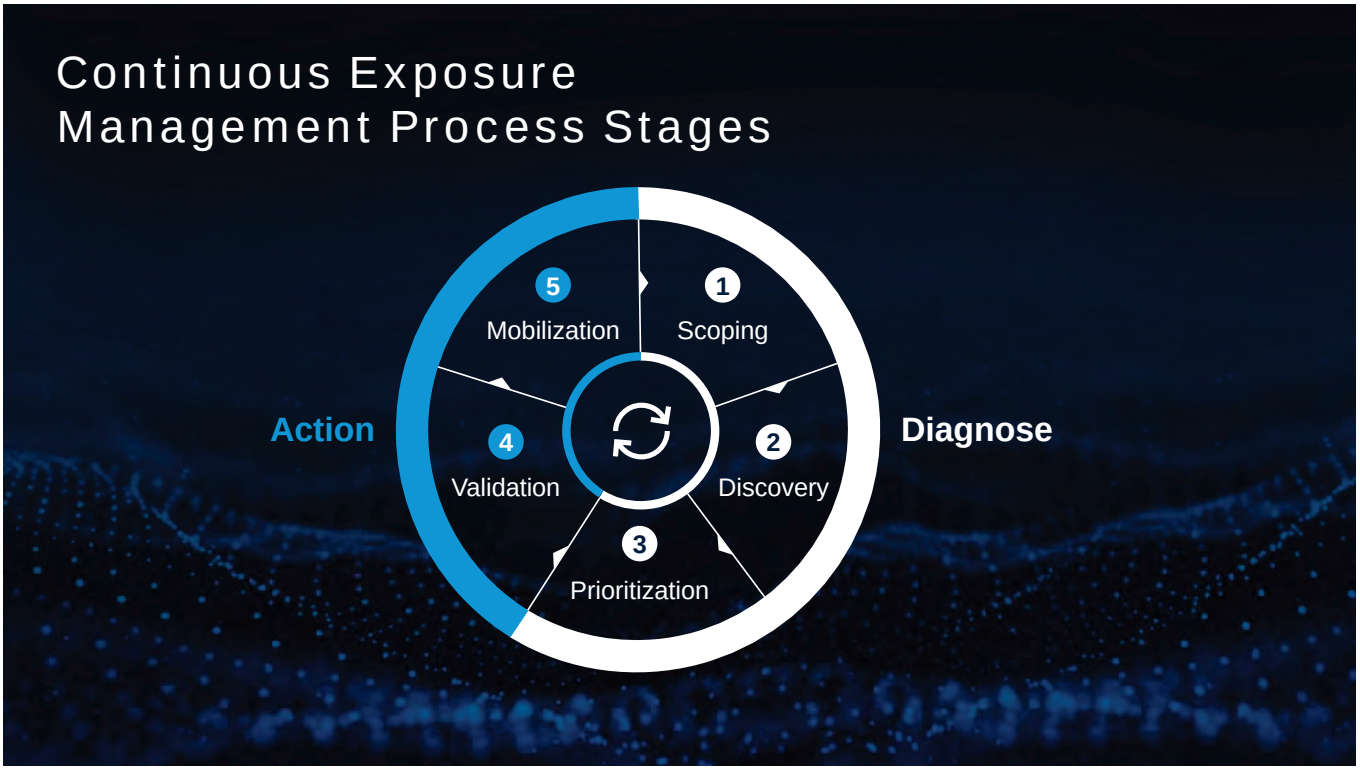

Transparency
Control
Explainability


AI TRiSM
Eases the Road to Compliance


2
Be Safe


2
Be Safe
Continuous Threat
Exposure Management


From Vulnerability Remediation
to Exposures Prioritization

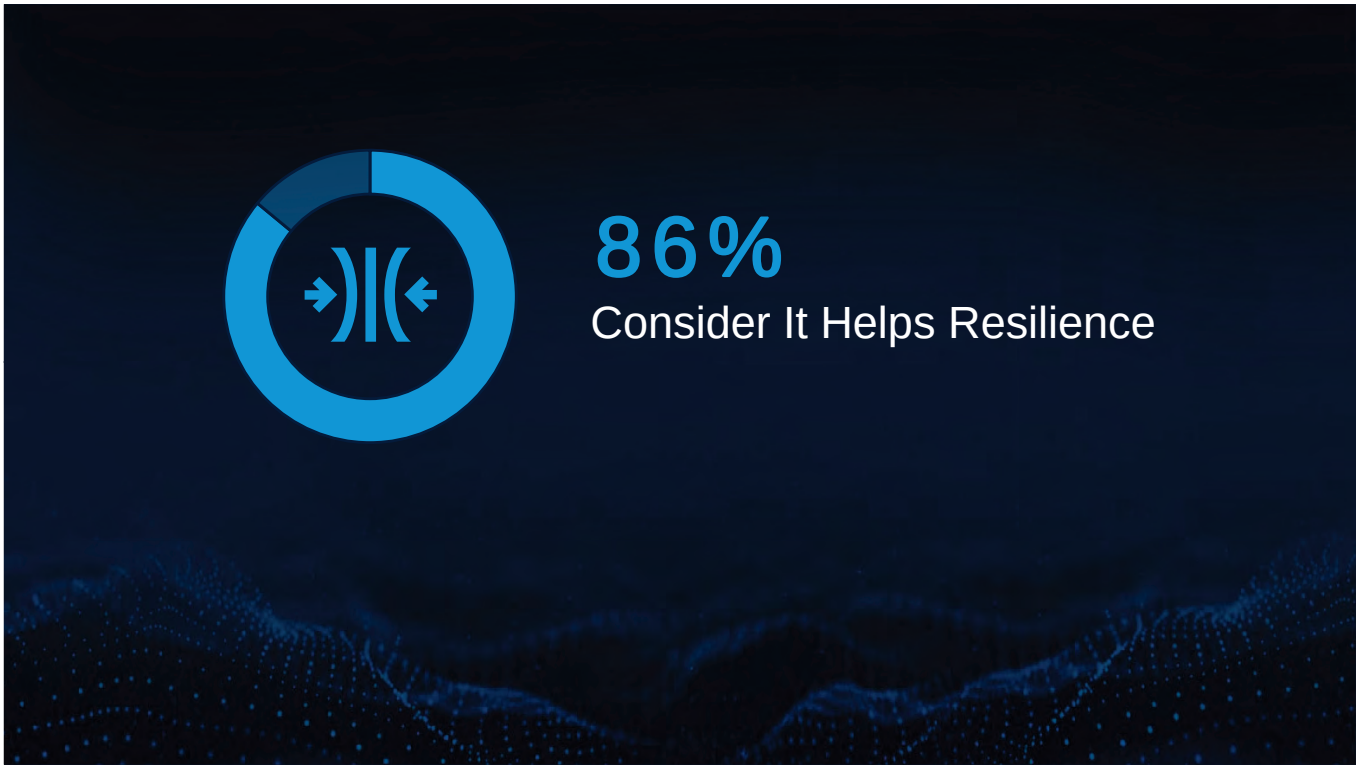

Continuous Exposure
Management Process Stages
1
Scoping
2
Discovery
3
Prioritization
4
Validation
5
Mobilization
Diagnose
Action


Three Times
Less Likely to Suffer From a Breach

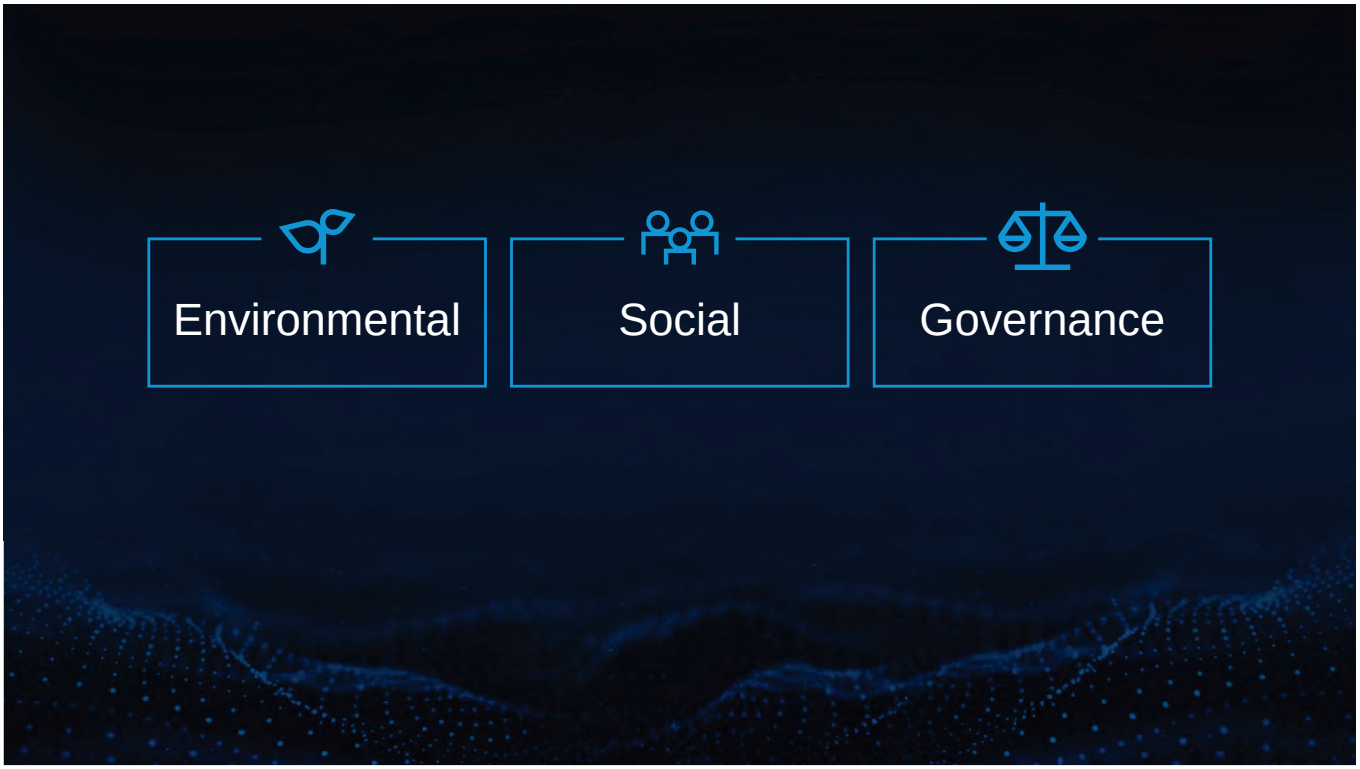

3
Protect the Future


3
Protect the Future
Use Sustainable Technology


74%
Sustainability Increases
Digital Maturity


86%
Consider It Helps Resilience


Within 2 Years,
75%
CIOs Will Be Responsible for
Sustainable Technology Outcomes

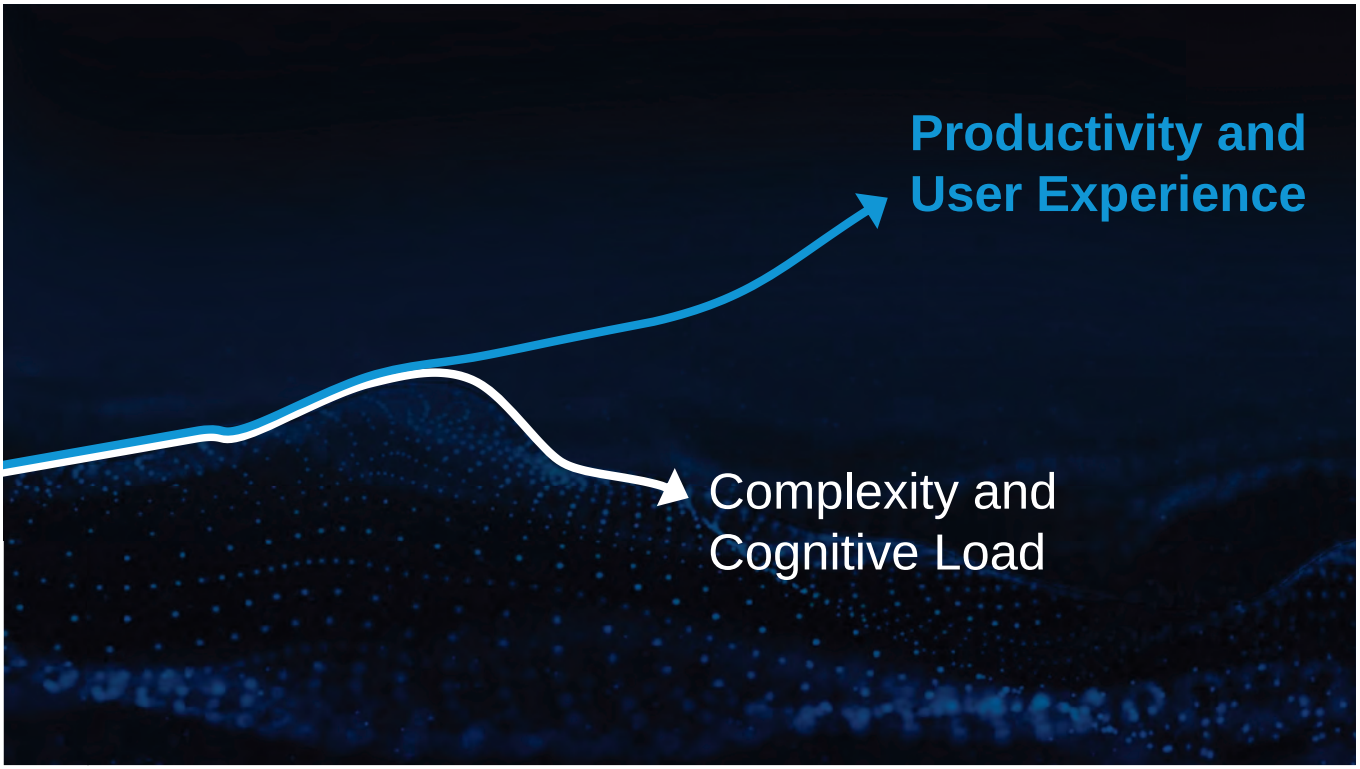

Environmental
Social
Governance


Equilibrium
of Result and Footprint

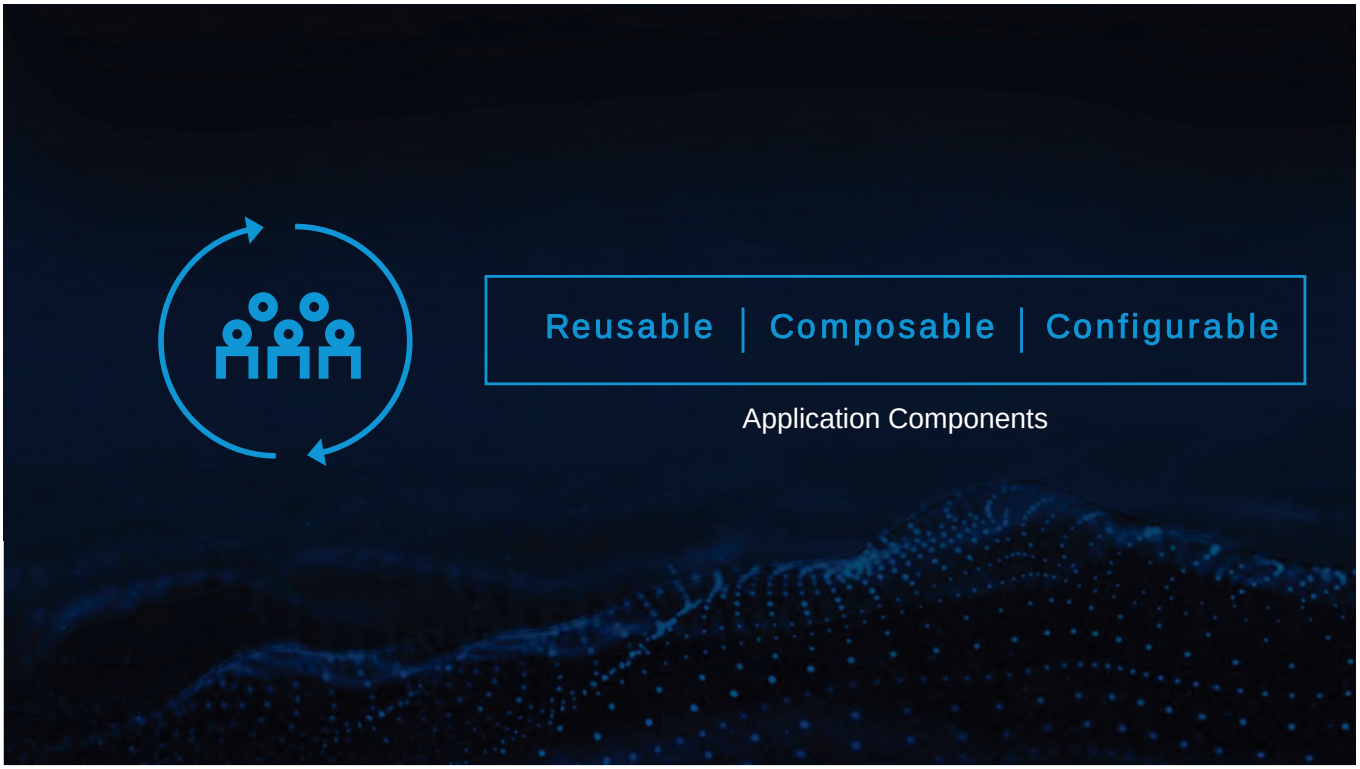

Rise of the Builders
Platform Engineering
AI-Augmented Development
AI Trust, Risk, Security Management
Industry Cloud Platforms
Intelligent Applications
Machine Customers
Continuous Threat Exposure Management
Sustainable Technology
Augmented Connected Workforce
Democratized Generative AI


4
Developer-Driven Self-Service


4
Developer-Driven Self-Service
Platform Engineering


Productivity and
User Experience
Complexity and
Cognitive Load


Establish Dedicated
Platform Teams


Reusable | Composable | Configurable
Application Components

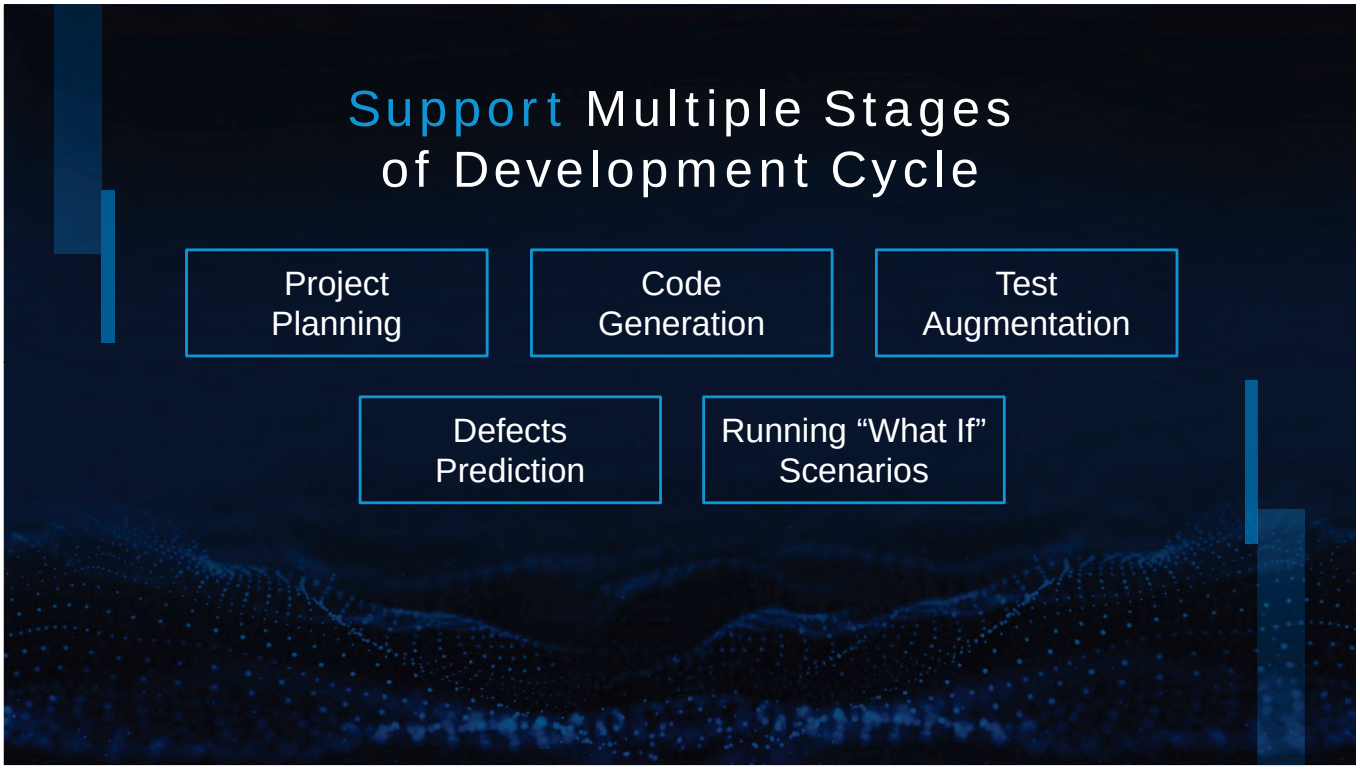

True
Product Management Culture


5
Accelerate Creation


5
Accelerate Creation
AI-Augmented Development


Design
Develop
Test

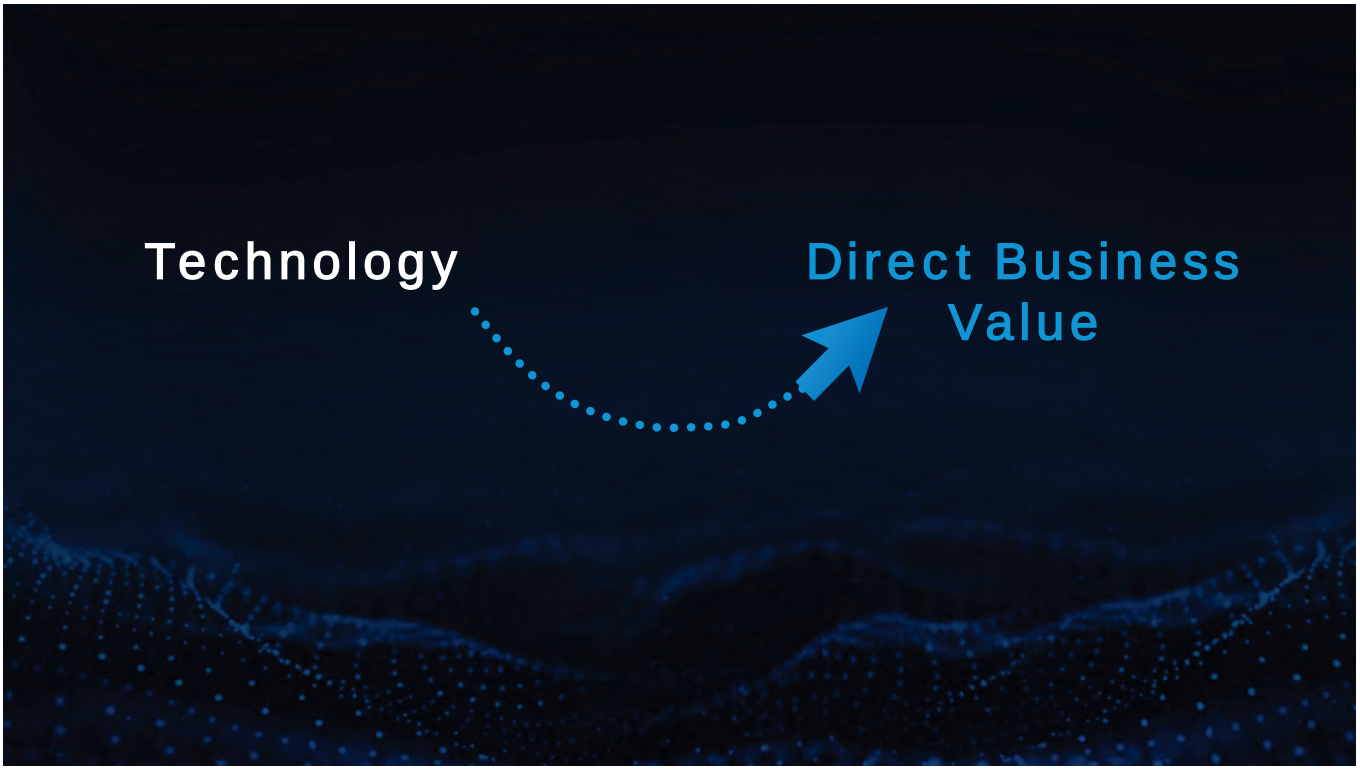

Retain Talent


Support Multiple Stages of Development Cycle
Project Planning
Code Generation
Test Augmentation
Defects Prediction
Running “What If” Scenarios


6
Tailor Your Tailor’s Work


6
Tailor Your Tailor's Work
Industry Cloud Platforms


Software
Platform
Infrastructure
Composable Product


Technology
Direct Business Value


Adaptable
Support Faster Innovation
Reduce Redundancy
Capture Value


270 Industry Cloud Platforms
Serving Two Dozen Industry Groups


Complement Your
Existing Infrastructure


Deliver the Value
Platform Engineering
AI-Augmented Development
AI Trust, Risk, Security Management
Industry Cloud Platforms
Intelligent Applications
Machine Customers
Continuous Threat Exposure Management
Sustainable Technology
Augmented Connected Workforce
Democratized Generative AI


7
Optimize Decision Making


7
Optimize Decision Making
Intelligent Applications

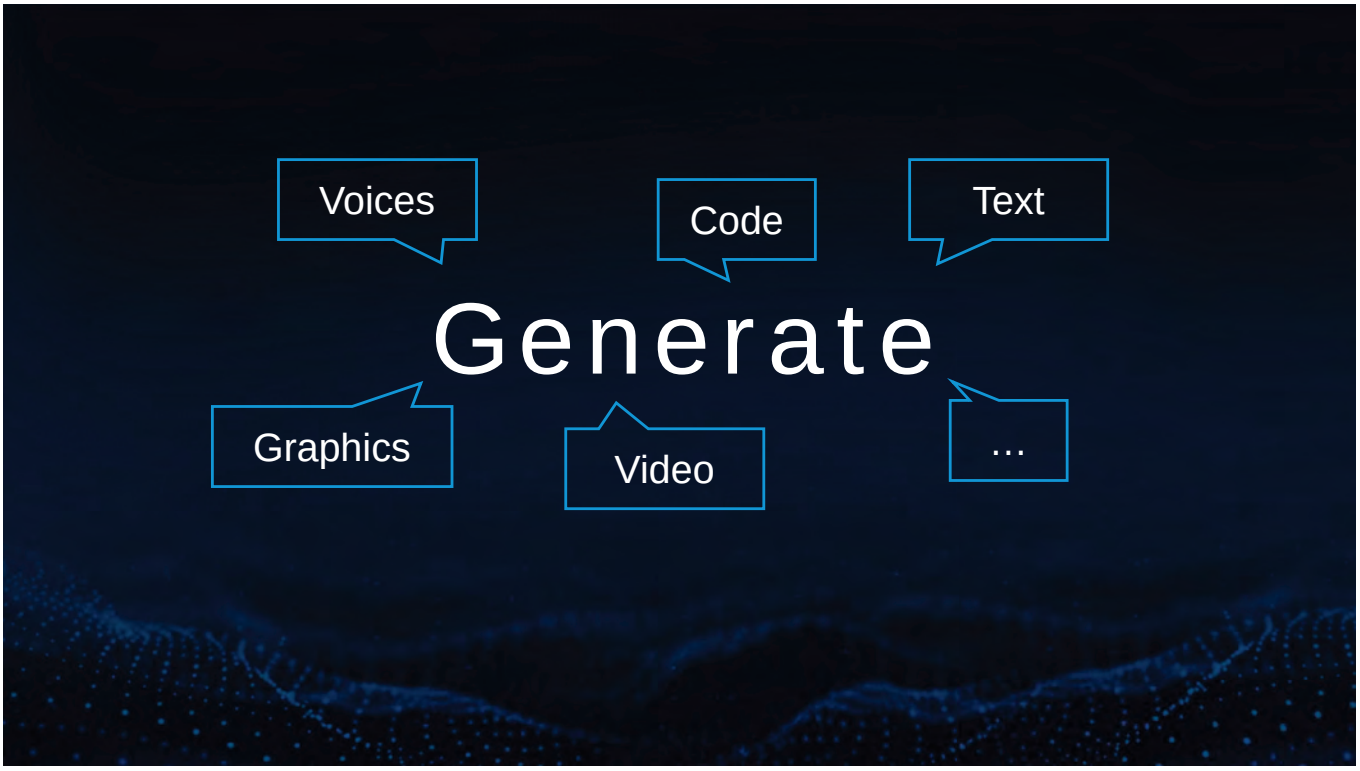

Embedded Architecture

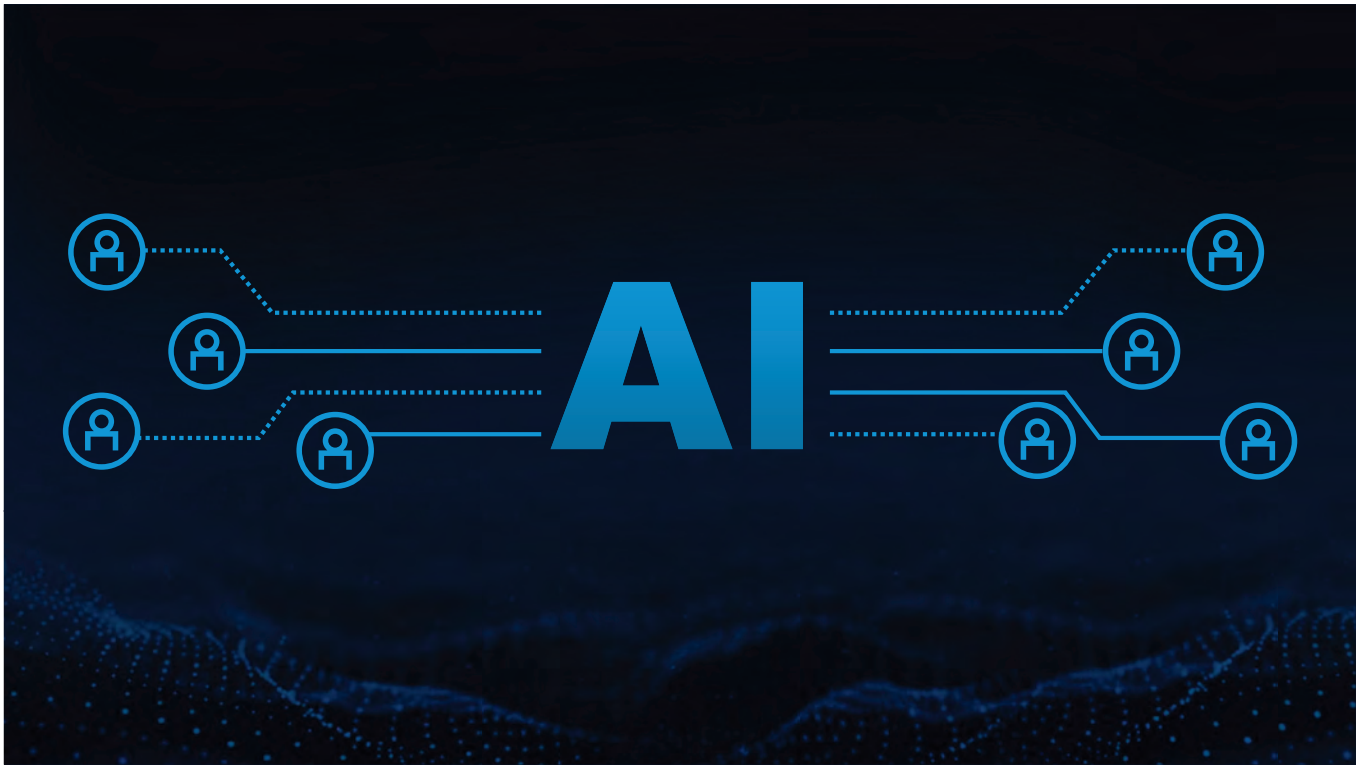

Embedded Architecture
Provides Adaptive and Contextualized Experiences


Future
Autonomous Self-Prompting Agents


8
Power and Responsibility

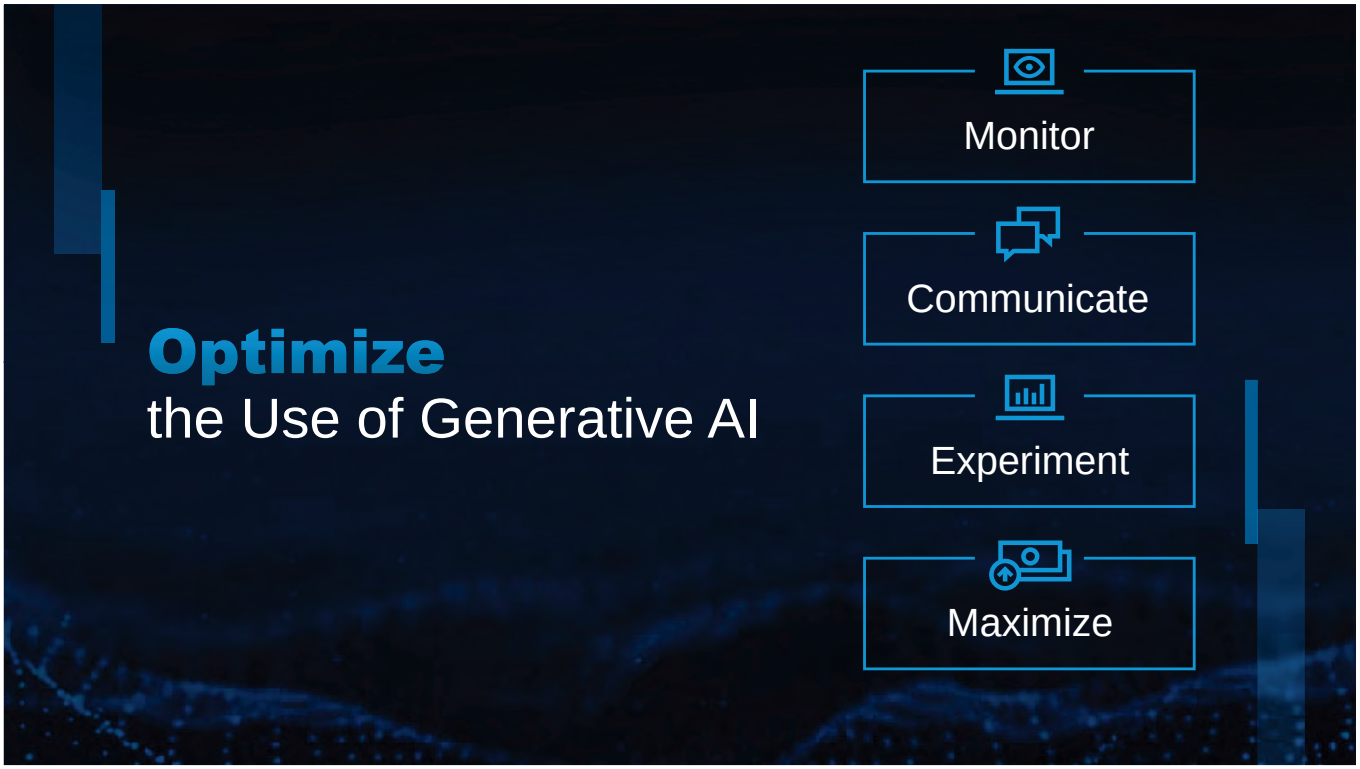

8
Power and Responsibility
Democratized
Generative AI


Voices
Code
Text
Generate
Graphics
Video
...


AI


Cost Savings
New Product Development
Improved Models
Increased Productivity
Enhanced Customer Experiences
Increased Accuracy
...

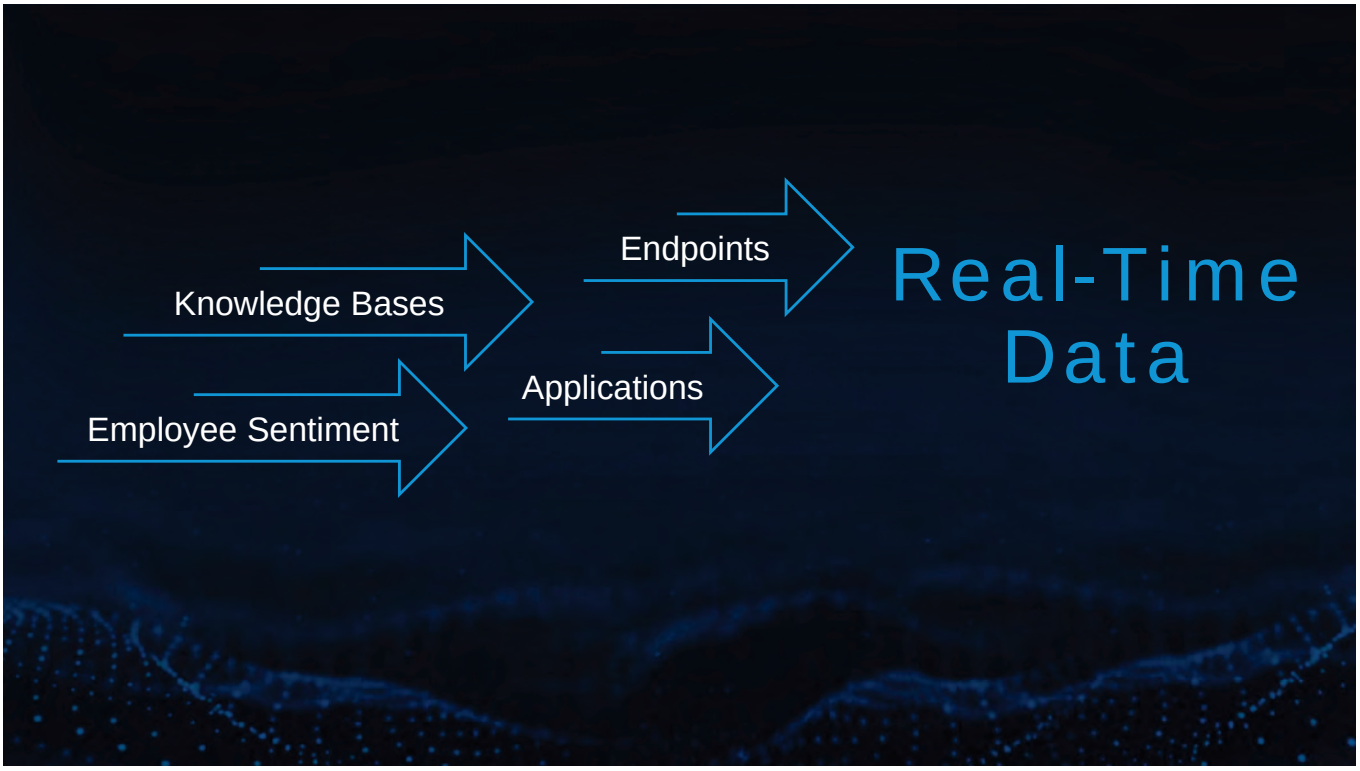

Risk
Reward

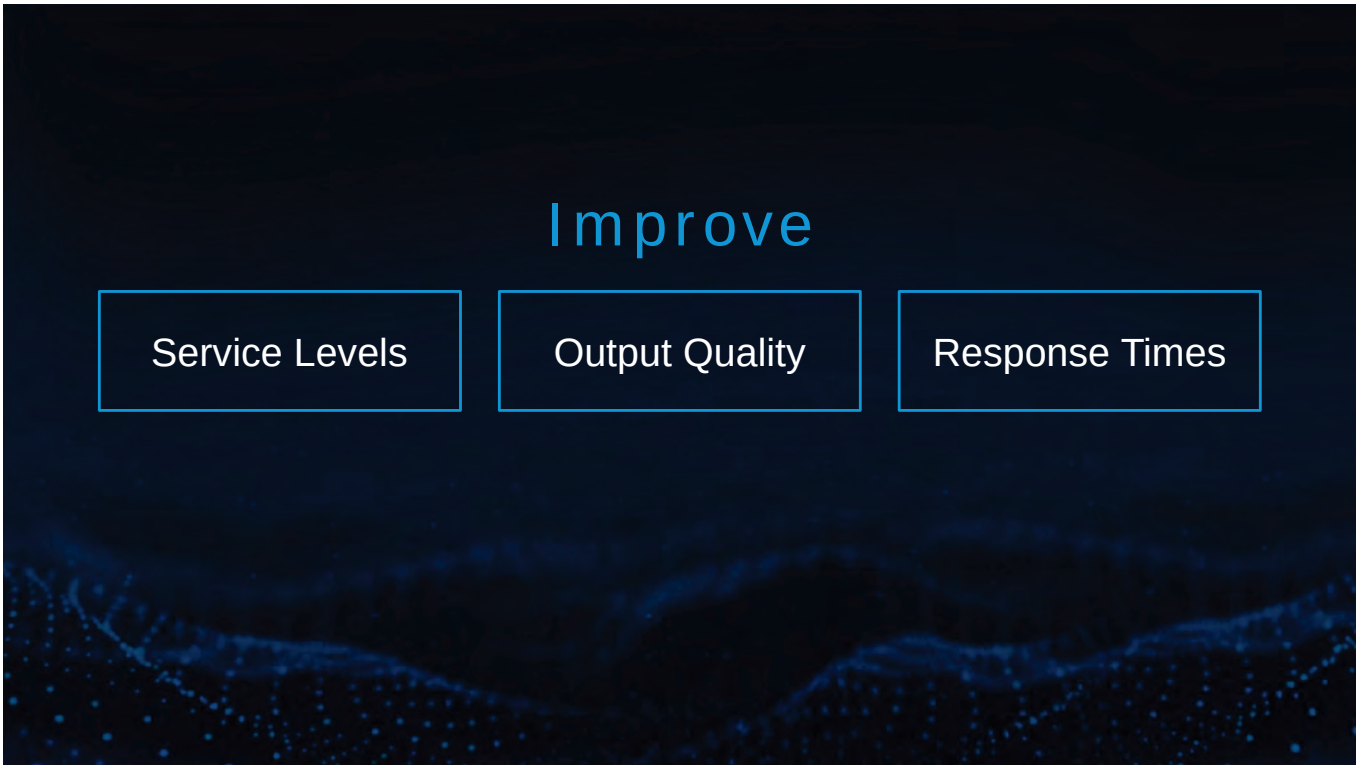

Optimize
the Use of Generative AI
Monitor
Communicate
Experiment
Maximize

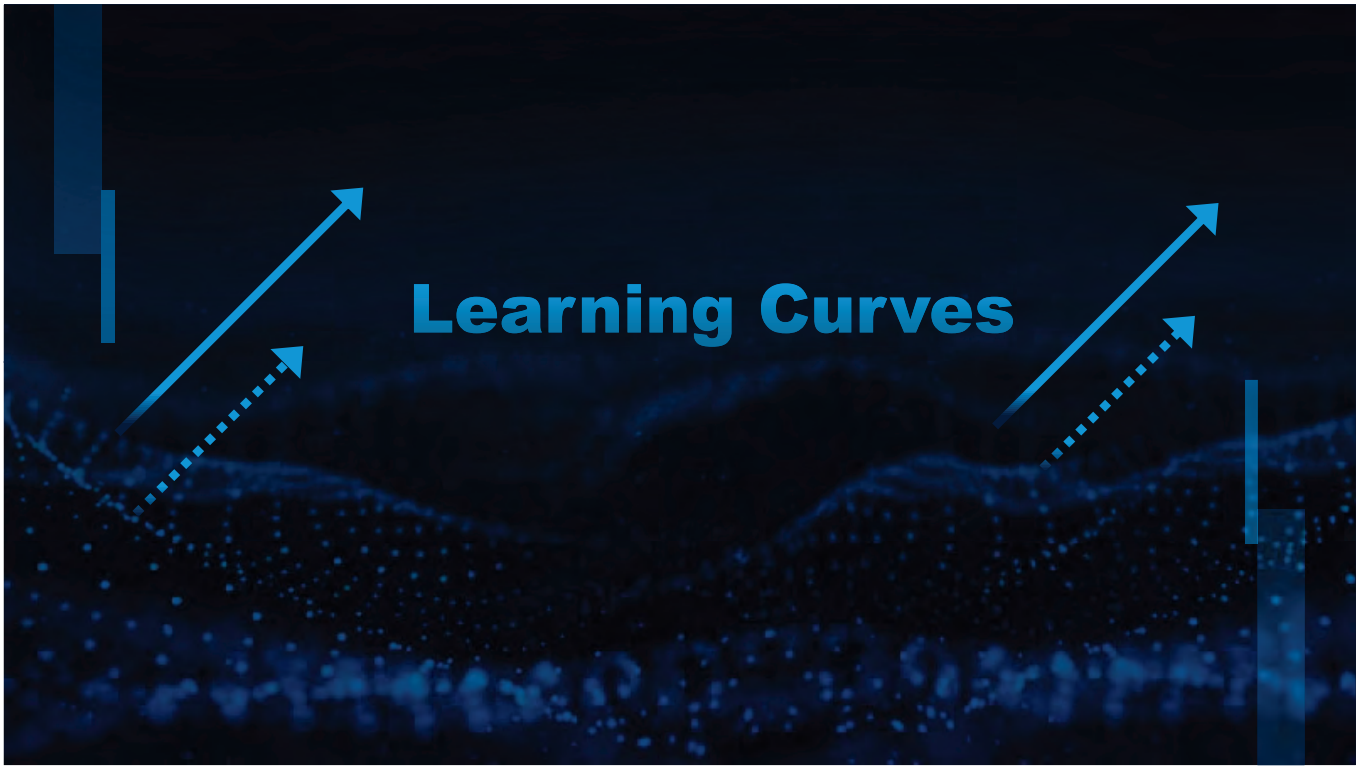

9
Push the Pioneers


9
Push the Pioneers
Augmented
Connected Workforce


DEX
Digital Employee Experience

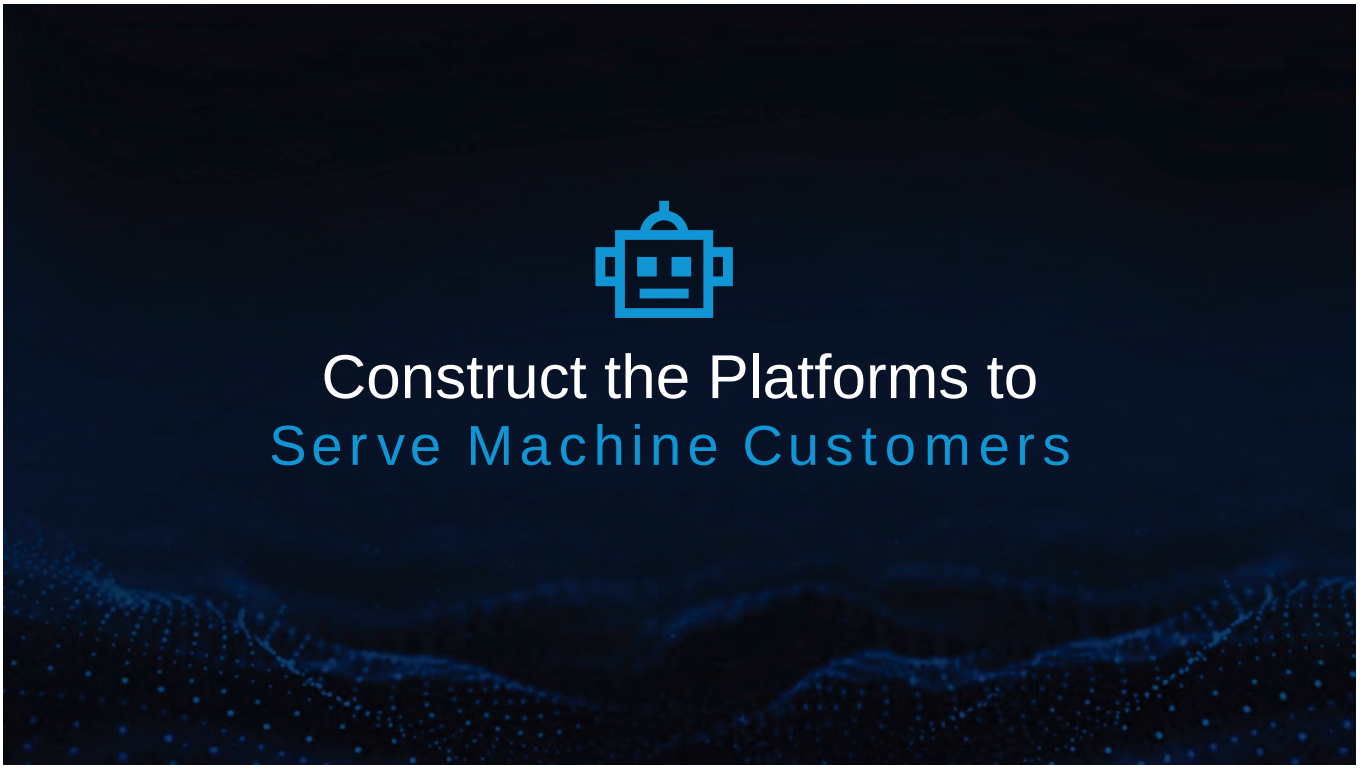

Knowledge Bases
Endpoints
Employee Sentiment
Applications
Real-Time Data


Improve
Service Levels
Output Quality
Response Times


Learning Curves


10
Buyers With Byte(s)


10
Buyers With Byte(s)
Machine Customers


Construct the Platforms to
Serve Machine Customers

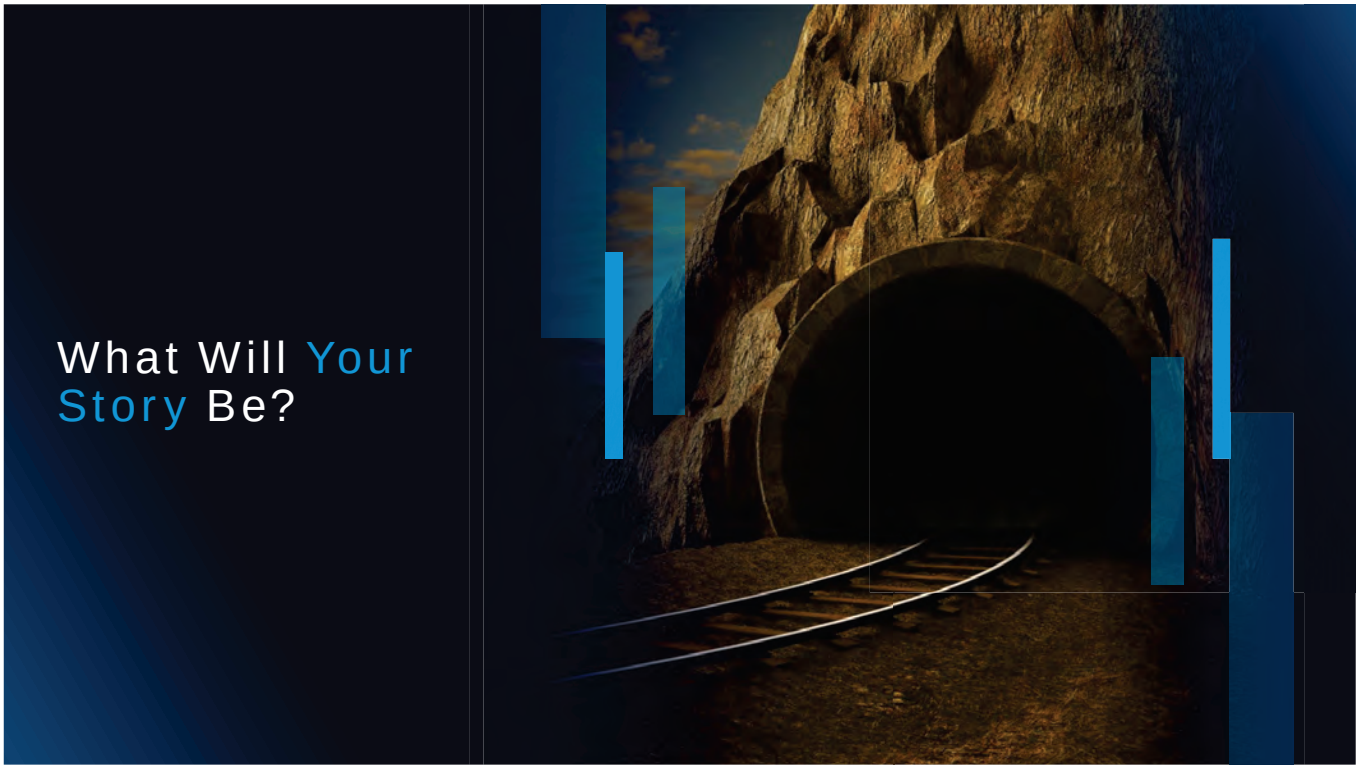

Emotional Customer Care
Price
Supply Chain Reliability


Opportunity
to Create Your Own Machine Clients


Top Strategic Technology Trends for 2024
Rise of the Builders
Platform Engineering
AI-Augmented Development
Protect Your Investment
Deliver the Value
AI Trust, Risk, Security Management
Industry Cloud Platforms
Intelligent Applications
Machine Customers
Continuous Threat Exposure Management
Sustainable Technology
Augmented Connected Workforce
Democratized Generative AI


Technology
= Means
≠ Purpose


What Are You Addressing?
Rise of the Builders
Platform Engineering
AI-Augmented Development
Protect Your Investment
Deliver the Value
AI Trust, Risk, Security Management
Industry Cloud Platforms
Intelligent Applications
Machine Customers
Continuous Threat Exposure Management
Sustainable Technology
Augmented Connected Workforce
Democratized Generative AI


What Will Your Story Be?


On-Demand Webinar
Gartner Top Strategic Technology Trends for 2024
https://www.gartner.com/en/webinar/548963
marko.stefancic@gartner.com
https://www.linkedin.com/in/mstefancic/

Zaupanje v storitve v svetu umetne inteligence

Trust in services in an AI world

Samo Zorc

Ministrstvo za digitalno preobrazbo Republike Slovenije

POVZETEK

Z vse večjo sposobnostjo in uporabnostjo različnih sistemov umetne intelligence (UI), posebej s sedanjim hitrim razvojem generativne umetne inteligence, so se v zadnjih letih poleg nekaterih revolucionarnih dosežkov, kot je razvoj novih antibiotikov, novih diagnostičnih postopkov ter reševanje nekaterih kompleksnih problemov pojavili tudi negativni učinki, ki so pomembno vplivali na življenja ljudi. Dejstvo je, da bo umetna inteligenca, kot splošno uporabna tehnologija, uporabna na vseh področjih in bo del življenja posameznika, podjetij in družbe. Generativna umetna inteligenca bo oblikovala digitalni svet na način, da ne bo enostavno mogoče razumeti kaj je resnično in kaj ne.

Ti sistemi se razvijajo na podlagi obstoječih digitalnih podatkov, predvsem tudi dostopnih na internetu, ki odslikavajo naravo stanja v družba, zato je jasno, da bo njihovo delovanje vključevalo tudi vse probleme, ki jih tovrstni podatki implicitno predstavljajo – od problemov s pristranskostjo, neresničnostjo, nepopolnostjo in nepravilnostjo delovanja sistemov. Hkrati lahko pričakujemo, da bodo ti sistemi namensko in nenamensko uporabljeni tako v dobro človeku kot tudi za nezakonite namene.

Da bi sistemi delovali na zaupanja vreden način je potrebno zagotoviti njihovo uporabo z osredotočanjem na človeka, da se zagotovi njegovo varnost, zdravje in človekove pravice, s čimer bomo zagotovili tudi prepotrebno zaupanje ljudi, potrebno za njihovo dejansko uporabo. Da bi to zagotovili, je potrebno umetno inteligenco ustrezno regulirati in spodbujati njeno družbeno sprejemljivo uporabo skladno z etičnimi načeli, hkrati pa sankcionirati nepravilno in nezakonito uporabo z negativnim učinkom na posameznika in družbo.

Akt o umetni inteligenci EU naslavlja to kompleksno problematiko sistemsko in horizontalno v okviru regulative za zagotavljanje varnosti proizvodov, ki se dajejo na trg. Cilj akta je zagotoviti tako varnost proizvodov in storitev samih kot tudi njihovo pravilno in ustrezno uporabo, predvsem pa enotno preglednost in predvidljivost pravil za vse različne sisteme UI in vsa različna področja njihove uporabe.

Predavanje pregledno predstavi osnovne koncepte regulacije UI in ključne zahteve za sisteme UI same ter zahteve za deležnike, ki sisteme razvijajo in uporabljajo.

SUMMARY

With the increasing power and usability of various AI systems, especially with the current rapid development of generative AI, in addition to some revolutionary achievements such as the development of new antibiotics, new diagnostic procedures and the solution of some complex problems, negative effects have also occurred in recent years that have significantly affected people's lives. The fact is that AI, as a universally applicable technology, will be useful in all areas and will be part of the lives of individuals, the economy and society. Generative AI will shape the digital world in such a way that it will not be easy to understand what is real and what is not.

These systems will be developed based on existing available digital data, which is in large extent available on the internet and reflects the state of society. It is therefore clear that their implementation and functioning will also include all the problems that this data implicitly entails – from the problems of bias, falsity, incompleteness and faulty execution. At the same time, it is to be expected that these systems will be used intentionally or unintentionally both for the benefit of people and for illegal purposes.

For the systems to function in a trustworthy manner, their use must be ensured by putting people at the centre to ensure their safety, health and human rights, thus ensuring the trust of people necessary for actual use of AI. To ensure this, it is necessary to adequately regulate AI and promote its socially acceptable use in line with ethical principles, while sanctioning misuse and unlawful use with negative impacts on individuals and society.

The EU AI Act addresses this complex issue in a systematic and horizontal way within the framework of assuring product safety on the internal market. The aim of the act is to ensure both that the products and services are safe themselves and at the same time also that users use them in the correct and proper way, but above all the uniform transparency and predictability of the rules for

all the different AI systems and their various fields of application.

In this presentation, the basic concepts of AI regulation, including requirements for AI systems themselves as well as the requirements for the actors who develop and use the systems, are presented.

O AVTORJU

Mag. Samo Zorc je sekretar odgovoren za politiko na področju umetne inteligence pri Ministrstvu za digitalno preobrazbo Republike Slovenije, kjer je imel ključno vlogo pri oblikovanju slovenskega pristopa k umetni inteligenci na nacionalni in mednarodni ravni. Vodil je pripravo Nacionalnega programa za umetno inteligenco – NpUI (2021) ter medresorsko delovno skupino, ki je odgovorna za pripravo in zastopanje stališč Slovenije v pogajanjih o Aktu o umetni inteligenci EU. Prav tako zastopa Slovenijo v mednarodnih organizacijah, kot so Delovna skupina za upravljanje umetne inteligence - AIGO pri OECD, Odbor za umetno inteligenco - CAI pri Svetu Evrope ter Globalno partnerstvo za umetno inteligenco – GPAI.

ABOUT THE AUTHOR

Samo Zorc, M.Sc. is the secretary responsible for policy in the field of artificial intelligence at the Ministry of Digital Transformation of the Republic of Slovenia, where he played a key role in shaping the Slovenian approach to artificial intelligence at the national and international level. He led the preparation of the National Program for Artificial Intelligence - NpUI (2021) and the interdepartmental working group responsible for the preparation and representation of Slovenia's positions in the negotiations on the EU Artificial Intelligence Act. It also represents Slovenia in international organizations, such as the Working Group for the Management of Artificial Intelligence - AIGO at the OECD, the Committee for Artificial Intelligence - CAI at the Council of Europe and the Global Partnership for Artificial Intelligence - GPAI.

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Zaupanje v storitve v svetu umetne inteligence

mag. Samo Zorc
koordinator priprave Nacionalnega programa za umetno inteligenco (NpUI)
vodja MDS za pripravo stališč Slovenije do Akta o umetni inteligenci EU (Akt o UI)
član odbora CAI za umetno inteligenco (Svet Evrope)
podpredsednik DS AIGO (OECD)
predstavnik Slovenije v Globalnem partnerstvu za umetno inteligenco (GPAI)

Ministrstvo za digitalno preobrazbo

16-17. 5.2024, Delavnica VITEL - Upravljanje s podatki v dobi umetne inteligence

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Agenda

1. **Umetna inteligenca (UI) spreminja svet**
2. **Kaj je UI?**
3. **Pristop EU k zagotavljanju zaupanja vredne UI**
4. **Akt o umetni inteligenci (Akt o UI)**

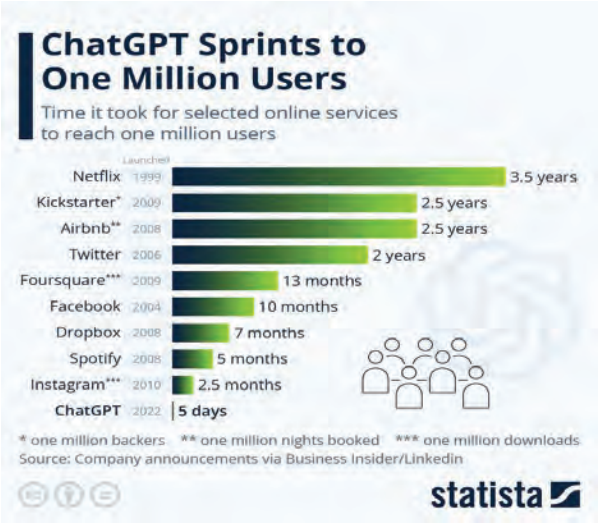
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

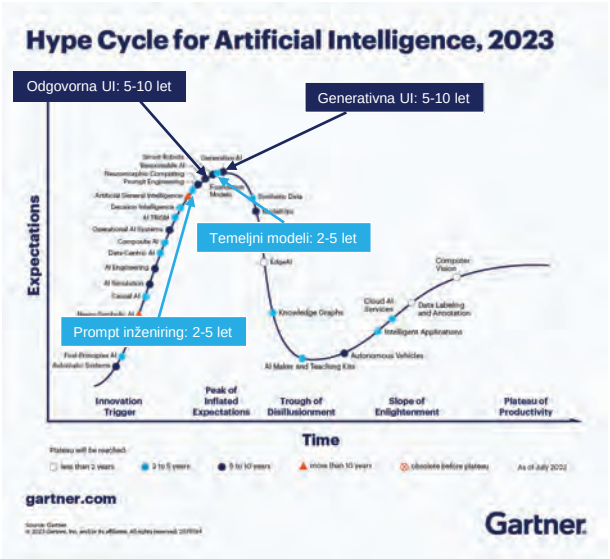
Je UI kot globalno segrevanje?

Pazite se!

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Generativna UI spreminja digitalni svet





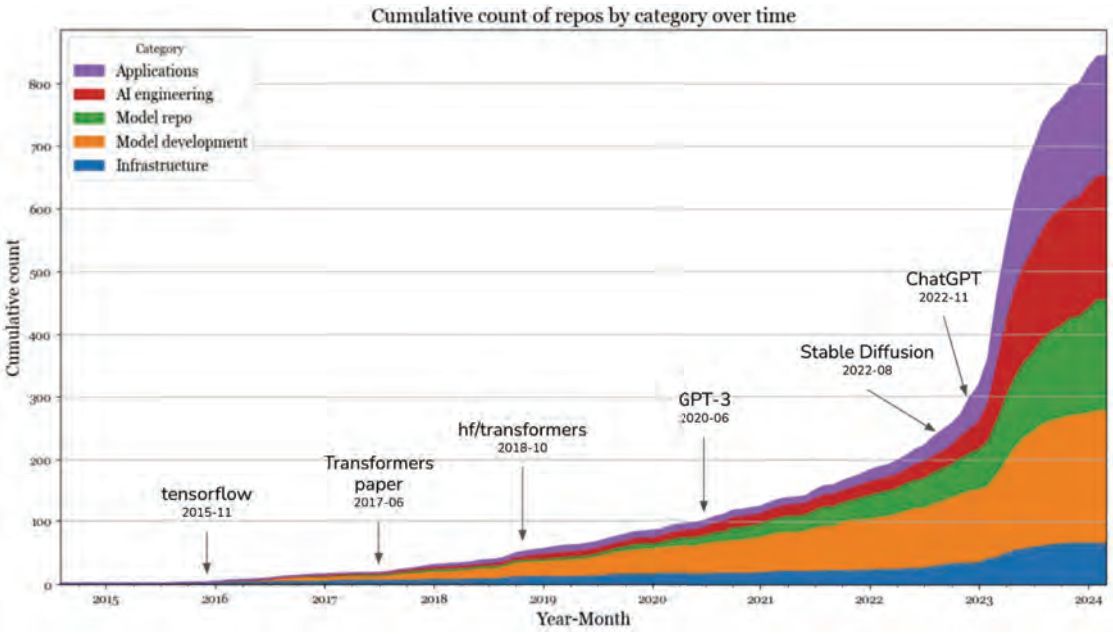
https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle

https://www.visualistan.com/2023/02/new-artificial-intelligence-bot-chatgpt-reaches-to-one-million-users-infographic.html

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Generativna UI – vročični val ali tsunami?

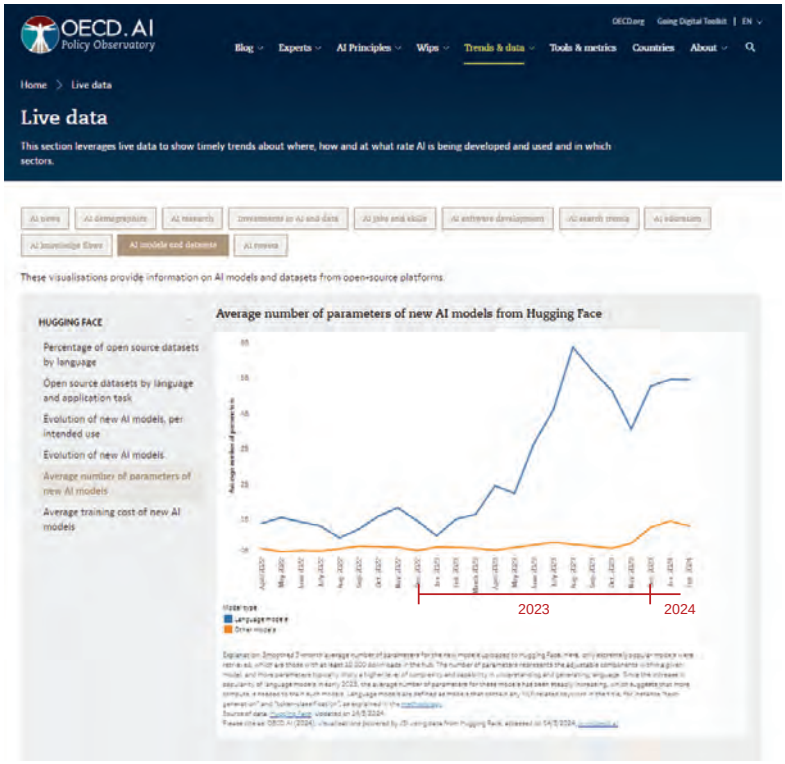
GitHub



© Chip Huyen, **What I learned from looking at 900 most popular open source AI tools,** Mar 14, 2024
https://huyenchip.com/2024/03/14/ai-oss.html

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Kje bodo pristali veliki modeli?



© OECD Observatorij za UI
https://oecd.ai/en/data?selectedArea=ai-models-and-datasets&selectedVisualization=evolution-of-number-of-parameters-from-hugging-face-models

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Kje bodo pristali mali modeli?

More Winning from Small models: Recent addition of Phi-3 from Microsoft

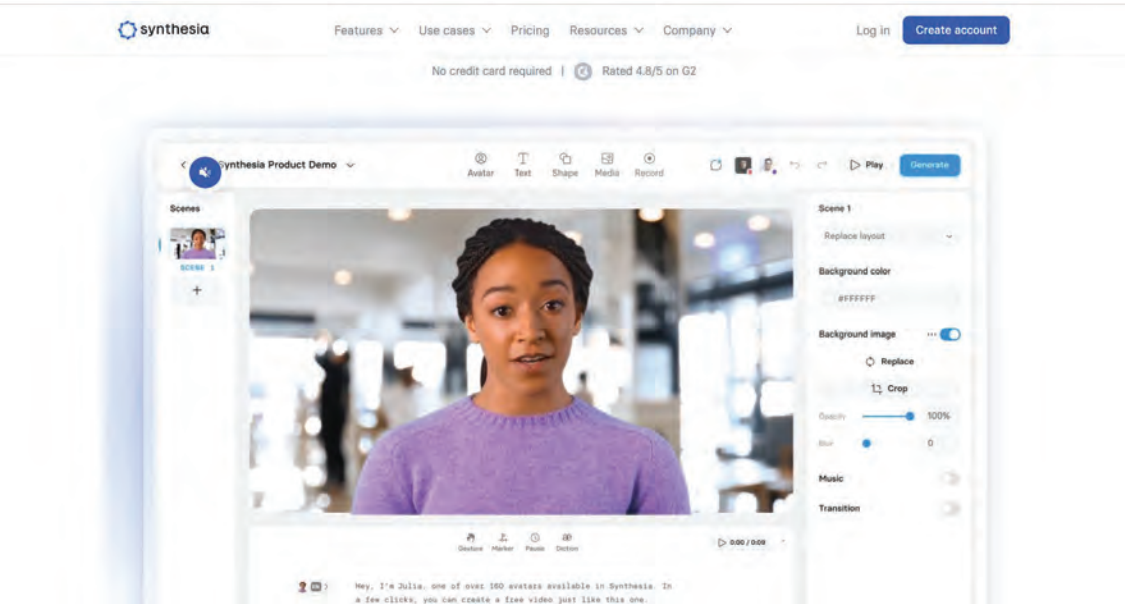
Phi-3-mini compact language model with 3.8 billion parameters trained on an extensive dataset of 3.3 trillion tokens, demonstrating performance comparable to larger models like Mixtral 8x7B and GPT-3.5. Notably, phi-3-mini achieves impressive scores of 69% on MMLU and 8.38 on MT-bench, all while being small enough for deployment on mobile devices.

	Phi-3-mini 3.8b	Phi-3-small 7b (preview)	Phi-3-medium 14b (preview)	Phi-2 2.7b	Mistral 7b	Gemma 7b	Llama-3-In 8b	Mixtral 8x7b	GPT-3.5 version 1106
MMLU (5-Shot) [HBK+21]	68.8	75.3	78.2	56.3	61.7	63.6	66.0	68.4	71.4
HellaSwag (5-Shot) [ZHB+19]	76.7	78.7	83.0	53.6	58.5	49.8	69.5	70.4	78.8
ANLI (7-Shot) [NWD+20]	52.8	55.0	58.7	42.5	47.1	48.7	54.8	55.2	58.1
GSM-8K (0-Shot; CoT) [CKB+21]	82.5	88.9	90.3	61.1	46.4	59.8	77.4	64.7	78.1
MedQA (2-Shot) [JPO+20]	53.8	58.2	69.4	40.9	49.6	50.0	58.9	62.2	63.4
AGIEval (0-Shot) [ZCG+23]	37.5	45.0	48.4	29.8	35.1	42.1	42.0	45.2	48.4
TriviaQA (5-Shot) [JCWZ17]	64.0	59.1	75.6	45.2	72.3	75.2	73.6	82.2	85.8
Arc-C (10-Shot) [CCE+18]	84.9	90.7	91.0	75.9	78.6	78.3	80.5	87.3	87.4
Arc-E (10-Shot) [CCE+18]	94.6	97.1	97.8	88.5	90.6	91.4	92.3	95.6	96.3
PIQA (5-Shot) [BZGC19]	84.2	87.8	87.7	60.2	77.7	78.1	77.1	86.0	86.6
SociQA (5-Shot) [BZGC19]	76.6	79.0	80.2	68.3	74.6	65.5	73.2	75.9	68.3

© Linkedin
https://www.linkedin.com/feed/update/urn:li:activity:7188756865147666434?utm_source=share&utm_medium=member_desktop

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

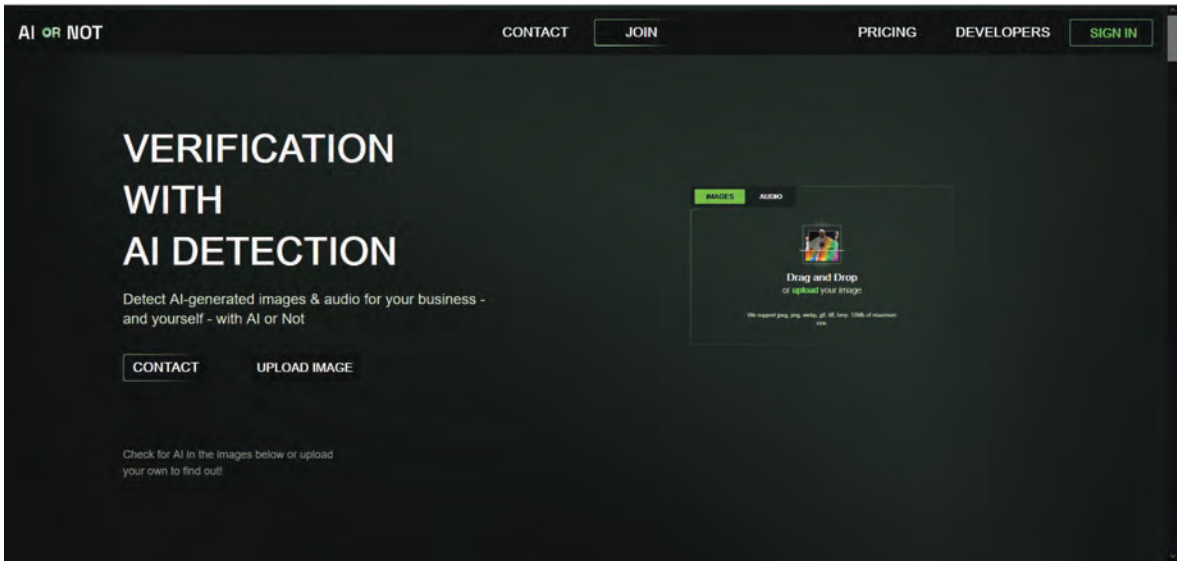
Bodo digitalni avatarji prevzeli naš posel?



https://www.synthesia.io/

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

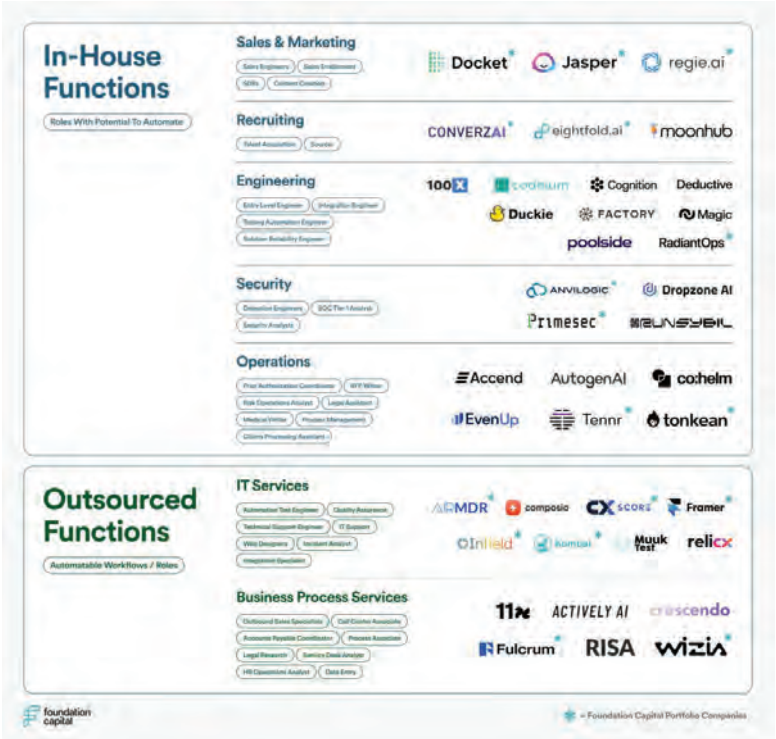
Bomo še vedeli kaj je res in kaj ne?



https://www.aiornot.com/

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

UI vodi v novo paradigmo „service-as-software“



https://foundationcapital.com/ai-service-as-software/

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Kaj pa pravzaprav je UI?

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Kaj je UI – Taksonomija EU JRC

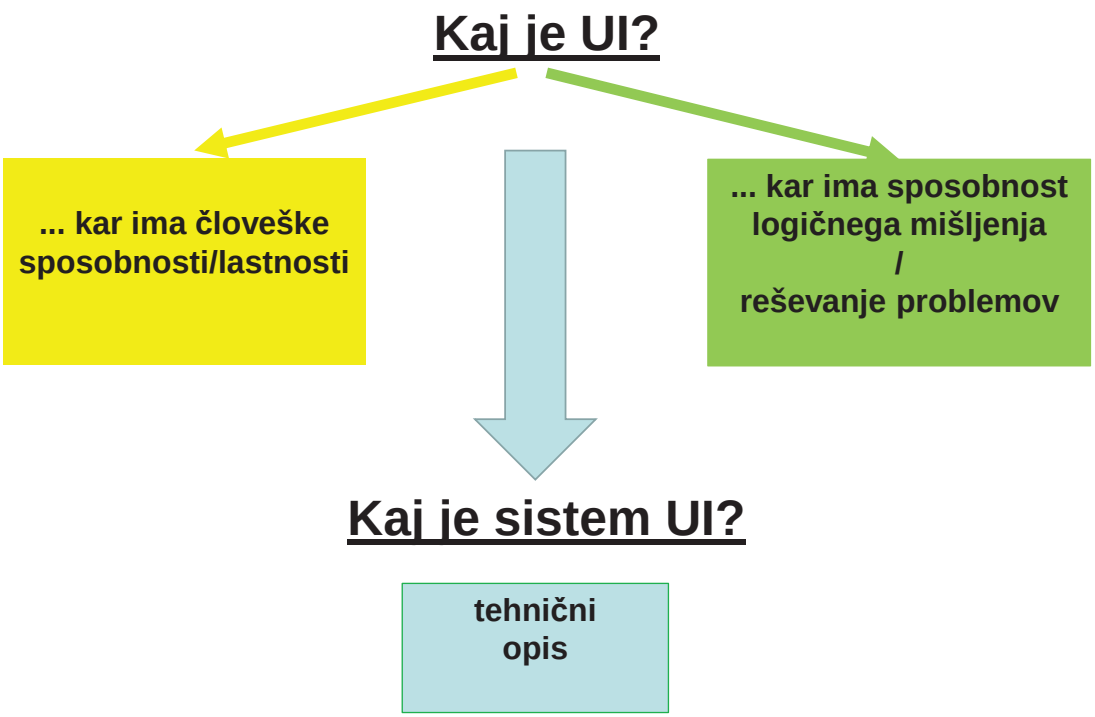
	AI taxonomy	
	AI domain	**AI subdomain**
Core	**Reasoning**	Knowledge representation
		Automated reasoning
		Common sense reasoning
	Planning	Planning and Scheduling
		Searching
		Optimisation
	Learning	Machine learning
	Communication	Natural language processing
	Perception	Computer vision
		Audio processing
Transversal	**Integration and Interaction**	Multi-agent systems
		Robotics and Automation
		Connected and Automated vehicles
	Services	AI Services
	Ethics and Philosophy	AI Ethics
		Philosophy of AI

Source: Authors' elaboration

© Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., and Delipetrev, B., AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence, EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, doi:10.2760/382730, JRC118163.

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Od antropomorfne k operativni definiciji



REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Kaj je UI – Taksonomija OECD

An AI system is a machine-based system that, for **explicit or implicit** objectives, **infers**, from the input it receives, **how to generate** outputs such as predictions, content, recommendations, or decisions that [can] influence physical or virtual environments. Different AI systems vary in their levels of **autonomy** and **adaptiveness** after deployment.

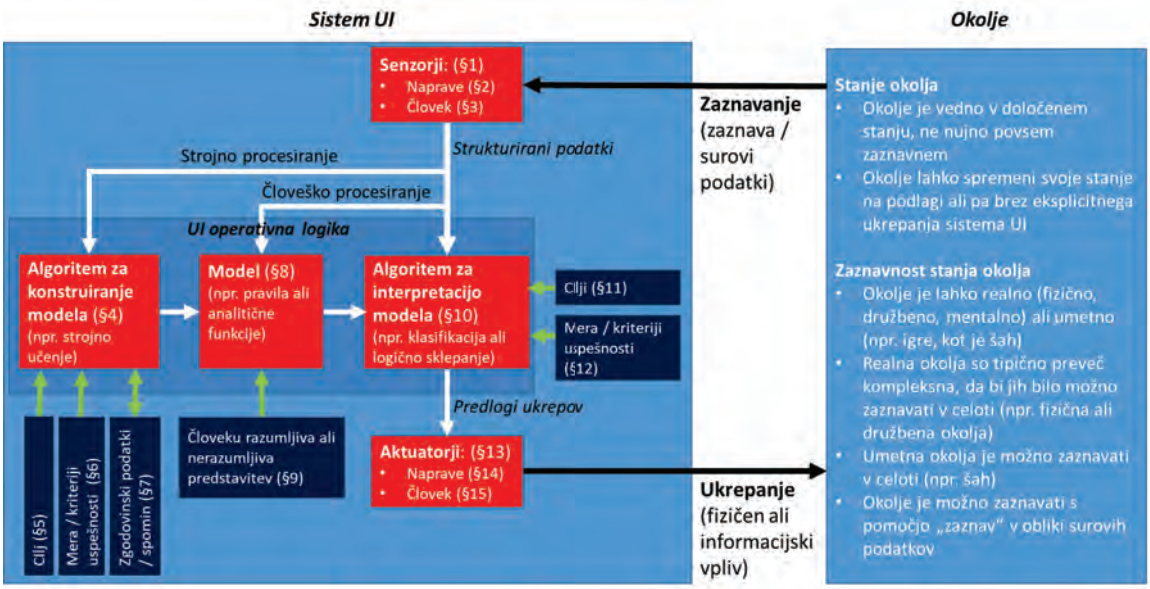
Explanatory memorandum:

- ROLE OF HUMANS, AUTONOMY AND ADAPTIVENESS
- ENVIRONMENT (OR CONTEXT)
- AI SYSTEM OBJECTIVES
- INPUT, INCLUDING DATA
- BUILDING AI SYSTEMS AND MODELS
- **INFERENCING IN AI SYSTEMS**
- OUTPUT(S)

https://oecd.ai/en/wonk/ai-system-definition-update

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Kaj je UI – Taksonomija NpUI (OECD)



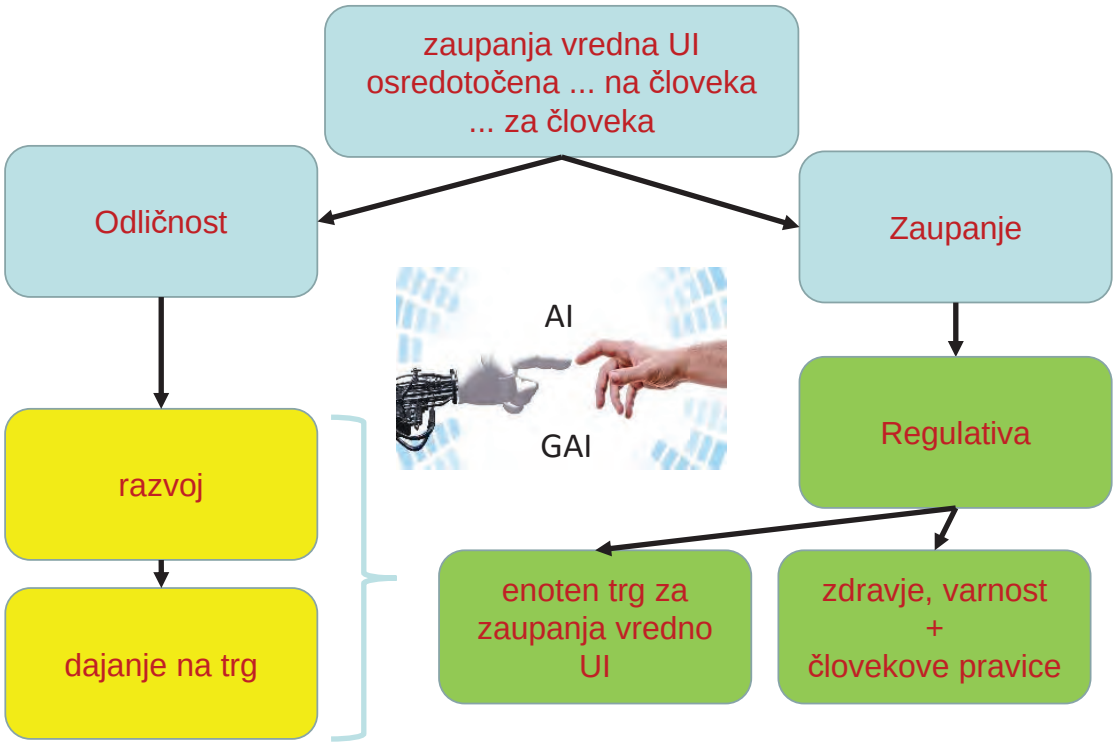
© Marko Grobelnik, IJS, vodja delovne skupine OECD, 2018

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Pristop EU k zagotavljanju zaupanja vredne UI

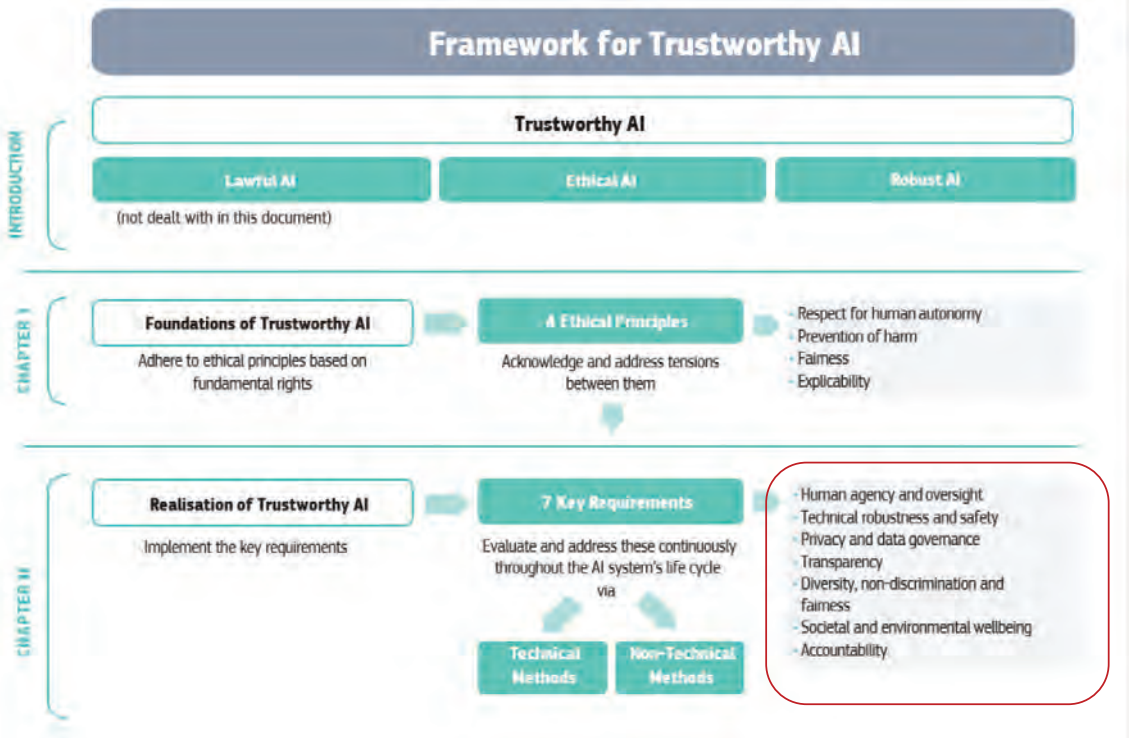
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Pristop EU – odličnost in zaupanje



REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Kaj je zaupanja vredna UI?



Etične smernice EU AI HLEG

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Pristop EU – Akt o UI

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Namen, cilji in izzivi uredbe

Zagotoviti pravilno delovanje notranjega trga z določitvijo **enotnih harmoniziranih pravil**, zlasti o razvoju, dajanju na trg EU ter uporabi **proizvodov in storitev**, ki uporabljajo tehnologije UI ali se zagotavljajo kot samostojni sistemi UI.

Specifični izzivi in cilji:

- ***Določiti zahteve za sisteme UI*** *ter s tem v zvezi* ***obveznosti vseh deležnikov*** *v celotni verigi vrednosti,* ***da se zagotovi****, da so sistemi UI, ki so dani na trg in se uporabljajo,* ***varni*** *ter* ***spoštujejo obstoječo zakonodajo o človekovih pravicah in vrednotah*** *EU.*
- ***Zagotoviti pravno varnost za olajšanje naložb in inovacij*** *na področju UI, tako da se jasno določi, katere zahteve, obveznosti ter postopke za skladnost je treba upoštevati za* ***uvedbo ali uporabo*** *sistema UI na trgu EU.*
- ***Izboljšati upravljanje in učinkovito izvrševanje obstoječe zakonodaje o človekovih pravicah in varnostnih zahtevah****, ki se uporabljajo za sisteme UI, z zagotavljanjem novih pooblastil, virov in jasnih pravil za ustrezne organe glede postopkov ugotavljanja skladnosti in naknadnega spremljanja ter* ***delitve nalog upravljanja in nadzora med nacionalno ravnjo in ravnjo EU****.*
- ***Olajšati razvoj enotnega trga*** *za* ***uporabo zakonite, varne in zaupanja vredne sisteme UI*** *ter* ***preprečiti razdrobljenost trga*** *s sprejetjem ukrepov EU za določitev minimalnih zahtev za sisteme UI, ki se* ***dajejo in uporabljajo na trgu EU*** *v skladu z* ***veljavno zakonodajo o človekovih pravicah in varnosti****.*

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Kaj je UI – Taksonomija Akt o UI

OECD:

An AI system is a machine-based system that, for **explicit or implicit** objectives, **infers**, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that [can] influence physical or virtual environments. Different AI systems vary in their levels of **autonomy** and **adaptiveness** after deployment.

Akt o UI (predlog):

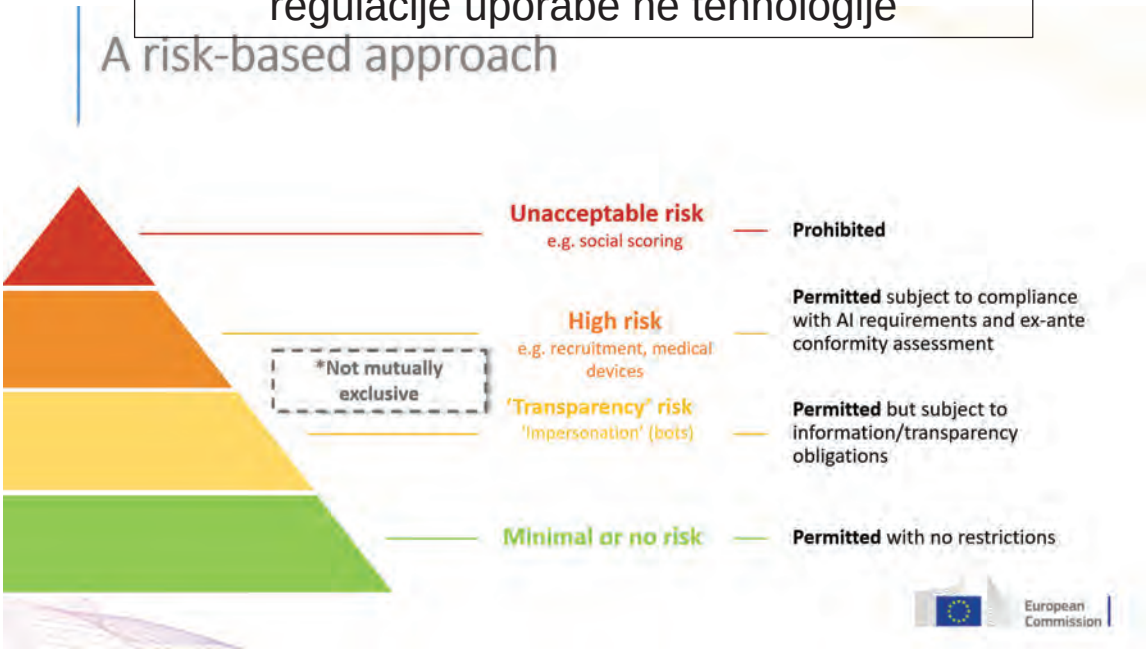
'AI system' means a **machine-based system** that is designed to operate with varying levels of **autonomy** and that may exhibit **adaptiveness** after deployment, and that, for **explicit or implicit** objectives, **infers**, **from the input it receives**, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;

'general-purpose AI system' means an AI system which is **based on a general-purpose AI model** and which has the capability to serve a **variety of purposes**, both for direct use as well as for integration in other AI systems;

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Pristop EU – zaupanje

Akt o UI –> notranji trg za zaupanja vredno UI
regulacije uporabe ne tehnologije

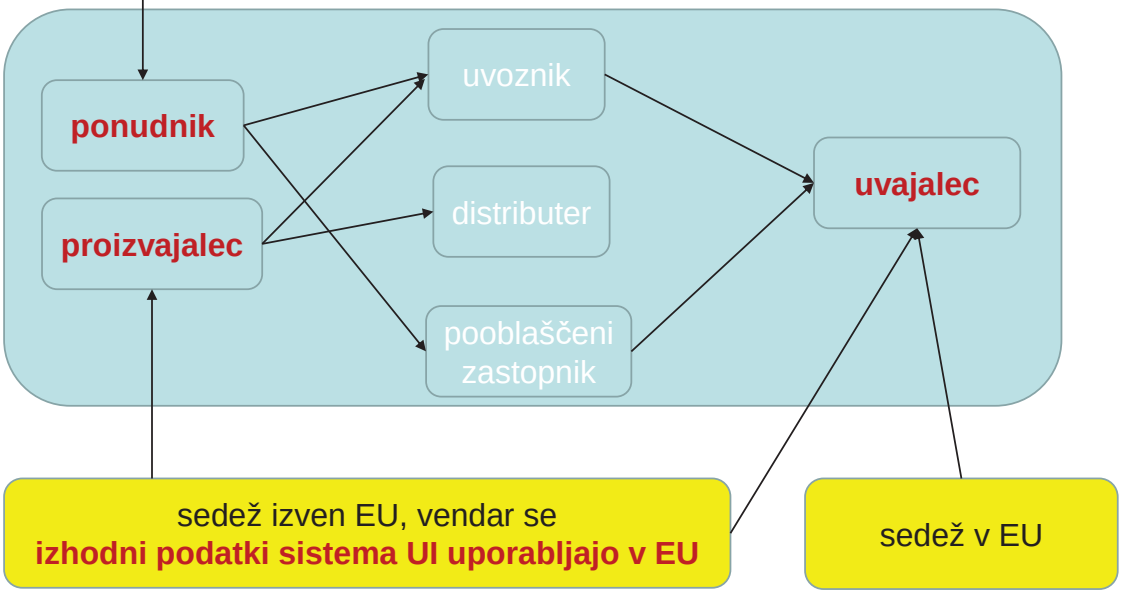


Vir: Evropska komisija, 2021

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Akt o UI – kaj ureja?

dajanje na trg ALI **dajanje v uporabo sistem UI -> za predviden namen**
placing on the market OR putting into service AI system-> for intended purpose
Ne glede na sedež!!



REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Akt o UI – česa ne ureja?

- vojska, obramba, nacionalna varnost
- javne organe drugih držav ali mednarodne organizacije če sodelujejo z EU na podlagi mednarodnih sporazumov za sodelovanje na področju preprečevanja, odkrivanja, preiskovanja in pregona kaznivih dejanj ter za sodelovanje na področju pravosodja
- sistemi UI in modeli UI za namen znanstvenih raziskav in razvoja
- raziskovalne, preizkusne ali razvojne dejavnosti v zvezi z sistemi UI ali modeli, preden se dajo na trg ali v uporabo
- zasebna neprofesionalna uporaba fizične osebe
- sistemi UI objavljeni pod FOSS licenco, razen če so dani na trg ali v uporabo kot prepovedani, visoko tvegani ali transparentni sistemi

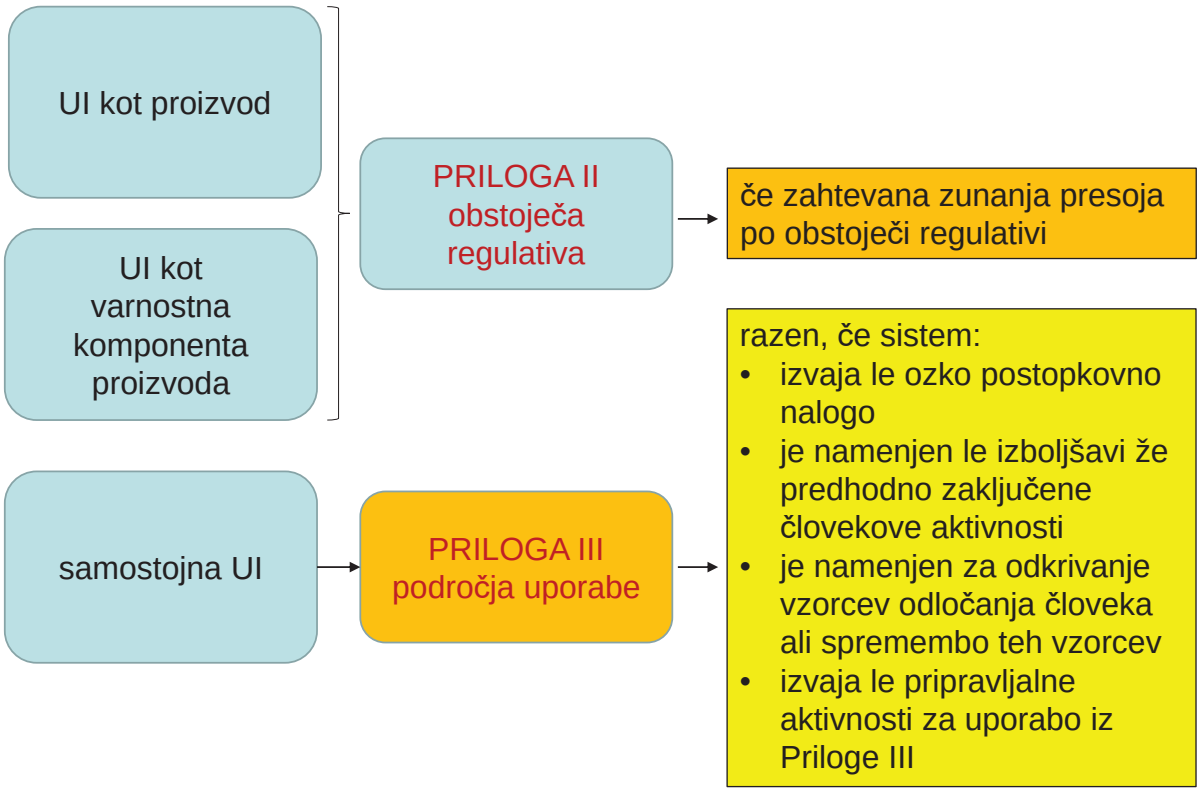
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Akt o UI - prepovedane prakse

- škodljiva uporaba podzavestne tehnike ali namerno manipulativne ali zavajajoče tehnike s ciljem bistvenega izkrivljanja vedenja
- izkoriščanje ranljivosti določene osebe ali skupine s ciljem bistvenega izkrivljanja vedenja
- socialno rangiranje
- oddaljena biometrična identifikacija "v realnem času" na javno dostopnih mestih za preprečevanje, odkrivanje, preiskovanje in pregon kaznivih dejanj- razen za opredeljene aktivnosti pod striktnimi pogoji
- sistemi biometrične kategorizacije posameznih fizičnih oseb glede rase, političnih prepričanj, članstva v sindikatih, verskega ali filozofskega prepričanja, spolnega življenja ali spolne usmerjenosti
- napovedovanje kaznivih dejanj oseb na podlagi osebnih značilnosti
- sistemi UI, ki kreirajo ali razširjajo baze za razpoznavanje obrazov iz neciljanih slik z interneta ali CCTV kamer
- razpoznavanje čustev na delovnem mestu ali izobraževalnih ustanovah, razen za medicinske namene

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Akt o UI – visoko tvegana UI



REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Visoko tvegana UI – Priloga II (NLF)

- Direktiva 2006/42/ES - o **strojih**
- Direktiva 2009/48/ES o **varnosti igrač**
- Direktiva 2013/53/EU o **plovilih za rekreacijo in osebnih plovilih**
- Direktiva 2014/33/EU o harmonizaciji zakonodaje držav članic v zvezi z **dvigali in varnostnimi komponentami za dvigala**
- Direktiva 2014/34/EU o harmonizaciji zakonodaj držav članic v zvezi z **opremo in zaščitnimi sistemi, namenjenimi za uporabo v potencialno eksplozivnih atmosferah**
- Direktiva 2014/53/EU o harmonizaciji zakonodaj držav članic v zvezi z **dostopnostjo radijske opreme na trgu**
- Direktiva 2014/68/EU o harmonizaciji zakonodaje držav članic v zvezi z omogočanjem dostopnosti **tlačne opreme na trgu**
- Uredba (EU) 2016/424 o **žičniških napravah**
- Uredba (EU) 2016/425 o **osebni varovalni opremi**
- Uredba (EU) 2016/426 o **napravah, v katerih zgoreva plinasto gorivo**
- Uredba (EU) 2017/745 o **medicinskih pripomočkih**
- Uredba (EU) 2017/746 o **in vitro diagnostičnih medicinskih pripomočkih**

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Visoko tvegana UI – Priloga II (pred NLF)

- Uredba (ES) št. 300/2008 o skupnih pravilih na področju **varovanja civilnega letalstva**
- Uredba (EU) št. 168/2013 o odobritvi in tržnem nadzoru **dvo- ali trikolesnih vozil in štirikolesnikov**
- Uredba (EU) št. 167/2013 o odobritvi in tržnem nadzoru **kmetijskih in gozdarskih vozil**;
- Direktiva 2014/90/EU o **pomorski opremi**
- Direktiva (EU) 2016/797 o **interoperabilnosti železniškega sistema** v EU
- Uredba (EU) 2018/858 o odobritvi in tržnem nadzoru **motornih vozil in njihovih priklopnikov ter sistemov, sestavnih delov in samostojnih tehničnih enot, namenjenih za taka vozila**
- Uredba (EU) 2019/2144 o zahtevah za **homologacijo motornih vozil in njihovih priklopnikov ter sistemov, sestavnih delov in samostojnih tehničnih enot, namenjenih za taka vozila, v zvezi z njihovo splošno varnostjo in zaščito potnikov v vozilu ter izpostavljenih udeležencev v cestnem prometu**
- 8. Uredba (EU) 2018/1139 o skupnih pravilih na področju **civilnega letalstva in ustanovitvi Agencije Evropske unije za varnost v letalstvu**

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Visoko tvegana UI – Priloga III

1. biometrija
2. kritična infrastruktura (digitalna, voda, elektrika, plin, ogrevanje, promet)
3. izobraževanje in poklicno usposabljanje
4. zaposlovanje, upravljanje delavcev in dostop do samozaposlitve
5. uživanje bistvenih zasebnih in javnih storitev in ugodnosti ter dostop do njih
6. preprečevanje, odkrivanje in preiskovanje kaznivih dejanj
7. upravljanje migracij, azila in nadzora meje
8. pravosodje in demokratični procesi

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Visoko tvegana UI – Priloga III

1. biometrija:
   a) sistemi za oddaljeno **biometrično identifikacijo**;
   b) biometrična **kategorizacija** na osnovi senzitivnih podatkov;
   c) **razpoznava** čustev;

2. kritična infrastruktura:
   a) sistem UI za uporabo kot varnostne komponente pri upravljanju in delovanju **kritične digitalne infrastrukture**, **cestnega prometa** ter pri **oskrbi z vodo**, **plinom**, **ogrevanjem** in **električno energijo**;

4. zaposlovanje, upravljanje delavcev in dostop do samozaposlitve:
   a) sistem UI za **zaposlovanje ali izbor** fizičnih oseb, zlasti za ciljno oglaševanje delovnih mest, **analizo in filtriranje prijav** za zaposlitev ter **ocenjevanje kandidatov**;
   b) sistem UI za **odločanje o pogojih delovnih razmerij, napredovanju ali prenehanju delovnih pogodbenih razmerij, dodeljevanju nalog** na podlagi **vedenja posameznika oziroma osebnostnih lastnosti ali značilnosti** ter spremljanju in ocenjevanju uspešnosti in vedenja oseb v takih razmerjih;

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Visoko tvegana UI – Priloga III

5. uživanje bistvenih zasebnih in javnih storitev in ugodnosti ter dostop do njih:
   a) sistem UI, ki naj bi jih uporabljali javni organi ali naj bi se uporabljali v njihovem imenu za **ocenjevanje upravičenosti fizičnih oseb do bistvenih ugodnosti in storitev javne pomoči**, vključno z **zdravstvenimi storitvami**, ter za **dodelitev, zmanjšanje, preklic ali povračilo** takih ugodnosti in storitev;
   b) sistem UI za **ocenjevanje kreditne sposobnosti fizičnih oseb** ali določanje njihove **kreditne ocene**, razen sistemov UI, ki se uporabljajo za odkrivanje finančnih goljufij;
   c) sistem UI za **oceno tveganja in oblikovanje cen** v zvezi s fizičnimi osebami v primeru **življenjskega in zdravstvenega zavarovanja**;
   d) sistem UI za **ocenjevanje in razvrščanje nujnih klicev** fizičnih oseb ali za napotitev služb za ukrepanje ob nesrečah, vključno s **policijo, gasilci in medicinsko pomočjo**, ali določanje prednosti pri njihovi napotitvi, ter sistemi za nujno zdravstveno triažo pacientov;

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

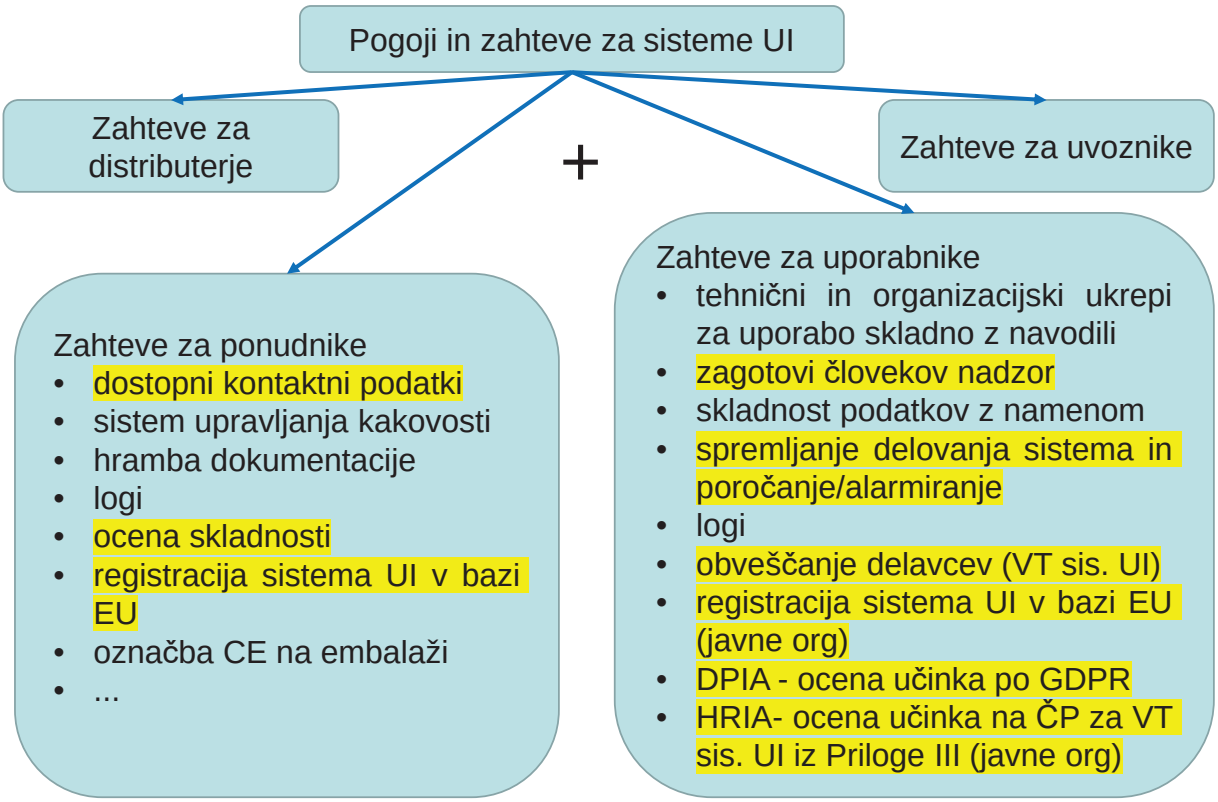
Visoko tvegana UI – zahteve za sistem UI

Pogoji in zahteve:

- Sistem za obvladovanje tveganj
- Podatki in podatkovno upravljanje
- Tehnična dokumentacija
- Vodenje evidenc
- Preglednost in zagotavljanje informacij uvajalcem
- Človeški nadzor
- Točnost, robustnost in informacijska varnost

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Visoko tvegana UI – zahteve za deležnike



REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Srednje tvegana UI - transparentnost

- **Biometrična kategorizacija**
- **Razpoznava čustev**
- **Globoki ponaredki**
- **Generirane vsebine** s strani sistemov UI ali UI splošnega namena (označevanje vsebine)
- Generiranje ali spreminjanje objavljenega **teksta** namenjenega **informiranju javnosti o stvareh javnega interesa**, razen za zagotavljanje javne varnosti ali v primeru objavljenih vsebin, ki morajo biti uredniško pregledane

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Modeli za splošne namene

1. Model UI za splošne namene

- „ model UI za splošne namene“ pomeni **model UI**, tudi kadar se takšen model UI uči z veliko količino podatkov z uporabo samonadzora v velikem obsegu, ki **kaže znatno splošno uporabnost** in je ne glede na to, kako je dan na trg, sposoben kompetentno **opravljati širok nabor različnih nalog** in ki ga je mogoče vključiti v različne sisteme ali aplikacije nižje v verigi, razen modelov UI, ki se uporabljajo za dejavnosti raziskav, razvoja ali izdelave prototipov, preden se dajo na trg;

2. Model UI za splošne namene s sistemskim tveganjem

- ima zmogljivosti **z visoko učinkovitostjo**, ocenjene na podlagi ustreznih tehničnih orodij in metodologij, vključno s kazalniki in merili uspešnosti;
  -> učenje zahteva > 10^25 FLOPS
- ima zmogljivosti ali **učinek, ki so enakovredni tistim iz točke (a)**, ob upoštevanju meril iz Priloge XIII, na podlagi odločitve Komisije, in sicer po uradni dolžnosti ali na podlagi kvalificiranega opozorila znanstvenega odbora.

Priloga XIII

a) **število parametrov modela**;
b) **kakovost ali velikost nabora podatkov**, merjeno na primer z žetoni;
c) **količino izračuna**, uporabljeno za učenje modela, merjeno s FLOP ali izraženo s kombinacijo drugih spremenljivk, kot so ocenjeni stroški učenja, predvideni čas, potreben za učenje, ali ocenjena poraba energije za učenje;
d) vhodne in izhodne načine modela, kot so besedilo v besedilo (veliki jezikovni modeli), besedilo v podobo, **večmodalnost** in najsodobnejši pragovi za določanje zmogljivosti z visoko učinkovitostjo za vsak način, ter posebna vrsta vhodnih in izhodnih podatkov (npr. biološka zaporedja);
e) **merila uspešnosti in ocenjevanja zmogljivosti modela**, vključno z upoštevanjem števila nalog brez dodatnega učenja, prilagodljivostjo za učenje novih, različnih nalog, njegovo stopnjo avtonomije in nadgradljivosti ter orodji, do katerih ima dostop;
f) ali ima velik vpliv na notranji trg zaradi svojega dosega, na katerega se domneva, ko je bil dan na voljo **vsaj 10 000 registriranim poslovnim uporabnikom s sedežem v Uniji**;
g) število registriranih končnih uporabnikov.

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Modeli za splošne namene – zahteve za ponudnike

1. Model UI za splošne namene

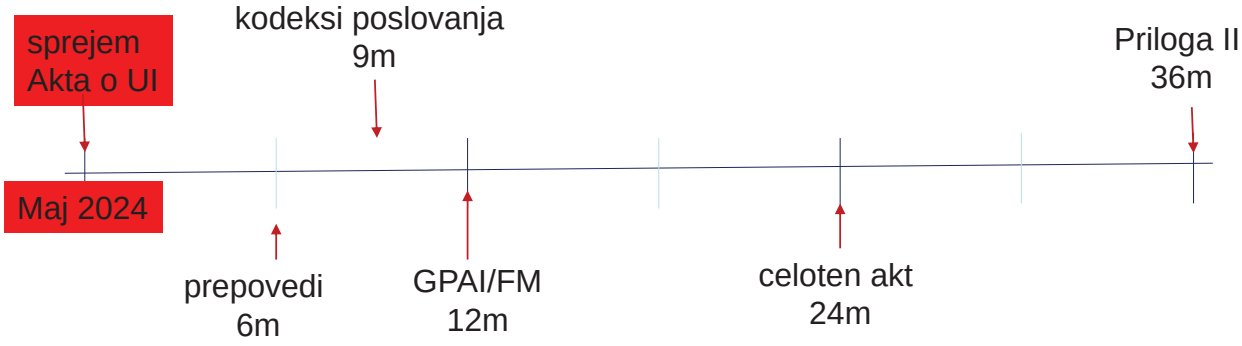
- pripravijo in posodabljajo **tehnično dokumentacijo** modela
- pripravijo, posodabljajo in **dajejo na voljo informacije in dokumentacijo za ponudnike sistemov UI**, ki nameravajo model UI za splošne namene integrirati v svoje sisteme UI
- vzpostavijo politiko za doseganje **skladnosti z avtorskim pravom** Unije
- pripravijo in objavijo dovolj **podroben povzetek vsebine, ki se uporablja za učenje modela UI** za splošne namene
- po potrebi **sodelujejo s Komisijo** in pristojnimi nacionalnimi organi

2. Model UI za splošne namene s sistemskim tveganjem:

- s **sedežem v tretji državi** pred dajanjem modela UI za splošne namene na trg Unije **s pisnim pooblastilom imenujejo pooblaščenega zastopnika s sedežem v Uniji**
- vse zahteve za Modele UI za splošne namene
- **izvedejo ocene modelov v skladu s standardiziranimi protokoli in orodji**, ki upoštevajo najsodobnejše tehnološke dosežke, vključno z izvajanjem in dokumentiranjem preizkušanja modelov po načelu kontradiktornosti z namenom prepoznavanja in zmanjšanja sistemskega tveganja;
- ocenijo in zmanjšajo morebitna **sistemska tveganja na ravni Unije**, vključno z njihovimi viri, ki lahko izhajajo iz razvoja, dajanja na trg ali uporabe modelov UI za splošne namene s sistemskim tveganjem;
- spremljajo in dokumentirajo **hude incidente** ter informacije o njih in morebitne popravne ukrepe za njihovo odpravo brez nepotrebnega odlašanja sporočajo Uradu za umetno inteligenco in po potrebi pristojnim nacionalnim organom;
- **zagotavljajo ustrezno raven kibernetske zaščite** za model UI za splošne namene s sistemskim tveganjem in fizično infrastrukturo modela

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Časovnica uveljavitve



REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Standardizacija – SIST TO za UI

SIST - TO za UI
- vodja Samo Zorc
- https://www.gov.si/novice/2023-03-24-poziv-k-sodelovanju-pri-razvoju-standardov-za-umetno-inteligenco/

Podlaga:
- Uredba 1025/2021 o standardizaciji
- Zakon o standardizaciji

Mandat/naloge:
- sodelovati pri sprejemanju standardov na EU in ISO ravni
- identificirati in sprejeti nacionalne standarde
- sekretarka TO ga. Petra Berčič (e-pošta: petra.bercic@sist.si)

Vabljeni k sodelovanju!

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA DIGITALNO PROBRAZBO

Hvala za pozornost!

mag. Samo Zorc

samo.zorc@gov.si

Kako v dobi UI zagotoviti varstvo zasebnosti? Odgovor ponuja GDPR.

How to ensure privacy protection in the age of AI? GDPR provides the answer.

Eva Kalan Logar

Republika Slovenija, Urad informacijskega pooblaščenca

POVZETEK

Sodobnega življenja, poslovanja, šolanja, pa tudi druženja si skoraj ni več mogoče predstavljati brez uporabe storitev vsaj kakega od znanih komercialnih oblačnih ponudnikov. Marsikaj nam oblak olajša, s sabo pa prinaša tudi tveganja za varstvo osebnih podatkov, ki so povezana z razpršenostjo podatkov in različnimi lokacijami njihove obdelave. V prispevku bomo poskusili odgovoriti na naslednja vprašanja: Kako dobro nam gre z varovanjem zasebnosti in osebnih podatkov v oblaku? Katera so tveganja? Na katere določbe zakonodaje o varstvu osebnih podatkov moramo biti še posebej pozorni? Kaj kaže praksa Informacijskega pooblaščenca?

SUMMARY

It is almost impossible to imagine modern life, business, schooling, and socializing without using the services of at least one of the well-known commercial cloud providers. The cloud makes many things easier for us, but it also brings with it risks for the protection of personal data, which are related to the dispersion of data and the various locations of their processing. In this article, we will try to answer the following questions: How well are we doing with protecting privacy and personal data in the cloud? What are the risks? Which provisions of the legislation on the protection of personal data should we pay particular attention to? What does the practice of the Information Commissioner show?

O AVTORJU

Mag. Eva Kalan Logar je poklicno pot začela na Ministrstvu za zunanje zadeve, Informacijskemu pooblaščencu pa se je pridružila leta 2008. Kot državna nadzornica za varstvo osebnih podatkov se je več let ukvarjala z inšpekcijskimi in prekrškovnimi postopki, prenosi podatkov v tretje države in različnimi mednarodnimi aktivnostmi pooblaščenca. Trenutno kot vodja državnih nadzornikov koordinira delo državnih nadzornikov ter sodeluje pri oblikovanju usmeritev postopkov inšpekcijskega nadzora in pri usklajevanju inšpekcijske prakse. Sodelovala je že pri različnih EU projektih in redno predava na različnih dogodkih.

ABOUT THE AUTHOR

M.Sc. Eva Kalan Logar started her professional career at the Ministry of Foreign Affairs and joined the Information Commissioner in 2008. As a national supervisor for the protection of personal data, she dealt with inspection and misdemeanour procedures, data transfers to third countries and various international activities of the Commissioner for many years. Currently, as the head of state inspectors, she coordinates the work of state inspectors and participates in the formulation of guidelines for inspection control procedures and in the coordination of inspection practice. She has already participated in various EU projects and regularly lectures at various events.

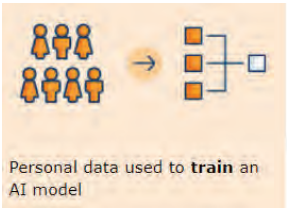
UI in GDPR

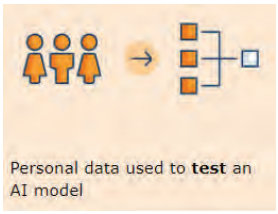
- GDPR ne vsebuje posebnih določb o umetni inteligenci, zato opredelitev definicije „osebni podatek“ temelji na splošni opredelitvi iz 4(1) člena GDPR:

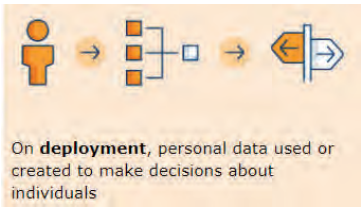
  OP = katerakoli informacija v zvezi z določenim ali določljivim posameznikom; določljiv posameznik je tisti, ki ga je mogoče neposredno ali posredno določiti, zlasti z navedbo identifikatorja, kot je ime, identifikacijska številka, podatki o lokaciji, spletni identifikator, ali z navedbo enega ali več dejavnikov, ki so značilni za fizično, fiziološko, genetsko, duševno, gospodarsko, kulturno ali družbeno identiteto tega posameznika.

UI in GDPR

- GDPR ureja zbiranje in uporabo OP. Kadar umetna inteligenca uporablja (za usposabljanje, testiranje ali uvajanje sistemov umetne inteligence) in ustvarja OP (sprejema odločitev o posameznikih, tudi če gre le za napovedi ali sklepe*), za tako obdelavo veljajo določbe GDPR!



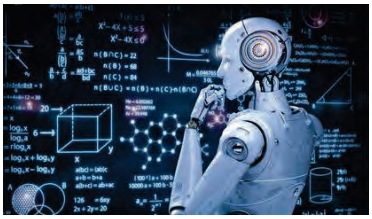




- V vseh teh primerih UI spada na področje uporabe zakonodaje o varstvu podatkov.
- Kadar UI ne vključuje uporabe OP, ne spada v področje uporabe zakonodaje o varstvu podatkov (npr. uporaba umetne inteligence za napovedovanje vremena ali astronomijo).

Kako zakonodaja o varstvu OP omenja UI?

- Pravo varstva OP je tehnološko nevtralno - ne sklicuje se neposredno na umetno inteligenco ali katere koli povezane tehnologije, kot je strojno učenje.
- Vendar pa se GDPR precej osredotoča na
  - obsežno avtomatizirano obdelavo OP,
  - profiliranje in
  - avtomatizirano sprejemanje odločitev.



Temeljna načela varstva OP

Tudi kot govorimo o obdelavi OP v svetu UI se vse začne pri osnovah varstva OP: temeljna načela varstva osebnih podatkov iz GDPR (5. člen)

Zakonitost

1. **zakonitost**
2. poštenost
3. preglednost
4. omejitev namena
5. minimizacija
6. točnost
7. omejitev hrambe
8. celovitost in zaupnost
9. odgovornost upravljavca

(a) PRIVOLITEV

- nedvoumna, informirana, specifična, svobodna
- ne more biti podlaga če obstaja pomembno nesorazmerje moči (npr. JS, delovno razmerje)

(b) POGODBA

- obdelava je potrebna za izvajanje pogodbe, katere pogodbena stranka je posameznik ali za izvajanje ukrepov na zahtevo takega posameznika pred sklenitvijo pogodbe
- le OP, ki so *potrebno* za izvajanje pogodbe/njeno sklenitev

(c) ZAKON

- izpolnitev zakonske *obveznosti*, ki velja za upravljavca
- ZVOP-2!

(d) ZAŠČITA ŽIVLJENJSKIH INTERESOV

(e) OPRAVLJANJE NALOGE V JAVNEM INTERESU ALI PRI IZVAJANJU JAVNE OBLASTI

- ZVOP-2!

(f) ZAKONITI INTERESI

- v poštev pride le za ZS; v JS le, če ne gre za izvajanje javnih (oblastnih) nalog

Poštenost in preglednost obdelave

1. zakonitost
2. **poštenost**
3. **preglednost**
4. omejitev namena
5. minimizacija
6. točnost
7. omejitev hrambe
8. celovitost in zaupnost
9. odgovornost upravljavca

- OP morajo biti obdelani pošteno in na pregleden način.
- Informiranje posameznikov ob vsakem pridobivanju njihovih OP: kdo ste, kdo je vaš DPO, katere OP obdelujete in zakaj, na kateri pravni podlagi, komu OP posredujete, rok hrambe, …
- V primeru obdelave OP, ki temelji izključno na avtomatizirani obdelavi in ima pravne ali podobno pomembne učinke za posameznika, na katerega se OP nanašajo, mora upravljavec posameznike obvestiti dodatno še o obstoju takega sprejemanja odločitev (tudi o logiki odločanja), o razlogih za to ter o predvidenih posledicah takega odločanja.
- Pri UI so glede transparentnosti pomembni še drugi vidiki, npr. da upravljavec pove, kdaj posameznik komunicira z UI in ne s človekom (če se uporabljajo virtualni asistenti) in v primeru ustvarjanja vsebin s pomočjo generativnega sistema UI da so te označene kot take.
- Smernice EDPB:
  - o transparentnosti
  - o avtomatiziranem odločanju in oblikovanju profilov

Pravice posameznikov po GDPR

- Pravica do informiranosti (12. in 13. člen) in dostopa do lastnih OP (15. člen)
- Pravica do popravka (16. člen) in izbrisa (17. člen)
- Pravica od omejitve obdelave (18. člen)
- Pravica do prenosljivosti (20. člen)
- Pravica do ugovora (21. člen)
- Pravice v zvezi z avtomatizirano obdelavo in profiliranjem (22. člen)

https://tiodlocas.si/

V primeru kršitve ali neodziva – pritožba pri IP.

Omejitev namena, minimizacija in točnost podatkov

1. zakonitost
2. poštenost
3. preglednost
4. **omejitev namena**
5. **minimizacija**
6. **točnost**
7. omejitev hrambe
8. celovitost in zaupnost
9. odgovornost upravljavca

- OP morajo biti zbrani za določene, izrecne in zakonite namene ter se ne smejo nadalje obdelovati na način, ki ni združljiv s temi nameni (nadaljnja obdelava v namene arhiviranja v javnem interesu, v znanstveno- ali zgodovinskoraziskovalne namene ali statistične namene v skladu s členom 89(1) ne velja za nezdružljivo s prvotnimi nameni).
- Študija EP o vplivu GDPR na UI (junij 2020) namiguje, da je zahteva po omejitvi namena obdelave OP združljiva z UI in velikimi podatki pod pogojem združljivosti namenov. Ponovna uporaba OP je tako dopustna, če to ni nezdružljivo z nameni, za katere so bili podatki prvotno zbrani oziroma uporabljeni. Domneva se, da je ponovna uporaba OP za statistične namene vselej združljiva (razen če vključuje nesprejemljiva tveganja za posameznika).
- OP morajo biti ustrezni, relevantni in omejeni na to, kar je potrebno za namene, za katere se obdelujejo.
- OP morajo biti točni in, kadar je to potrebno, posodobljeni; sprejeti je treba vse razumne ukrepe za zagotovitev, da se netočni osebni podatki brez odlašanja izbrišejo ali popravijo ob upoštevanju namenov, za katere se obdelujejo.

Omejitev hrambe podatkov

1. zakonitost
2. poštenost
3. preglednost
4. omejitev namena
5. minimizacija
6. točnost
7. **omejitev hrambe**
8. celovitost in zaupnost
9. odgovornost upravljavca

- OP morajo biti hranjeni v obliki, ki dopušča identifikacijo posameznikov, na katere se nanašajo OP, le toliko časa, kolikor je potrebno za namene, za katere se OP obdelujejo; (OP se lahko shranjujejo za daljše obdobje, če bodo obdelani zgolj za namene arhiviranja v javnem interesu, za znanstveno- ali zgodovinskoraziskovalne namene ali statistične namene v skladu s členom 89(1), pri čemer je treba izvajati ustrezne tehnične in organizacijske ukrepe iz te uredbe, da se zaščitijo pravice in svoboščine posameznika, na katerega se nanašajo OP)

Celovitost in zaupnost podatkov

1. zakonitost
2. poštenost
3. preglednost
4. omejitev namena
5. minimizacija
6. točnost
7. omejitev hrambe
8. **celovitost in zaupnost**
9. odgovornost upravljavca

- OP se obdelujejo se na način, ki zagotavlja ustrezno varnost OP, vključno z zaščito pred nedovoljeno ali nezakonito obdelavo ter pred nenamerno izgubo, uničenjem ali poškodbo z ustreznimi tehničnimi ali organizacijskimi ukrepi.
- Uporaba robustnih varnostnih mehanizmov, kot so šifriranje podatkov, avtentikacija uporabnikov, nadzor dostopa…

Načelo odgovornosti

1. zakonitost
2. poštenost
3. preglednost
4. omejitev namena
5. minimizacija
6. točnost
7. omejitev hrambe
8. celovitost in zaupnost
9. **odgovornost upravljavca**

- Za skladnost z načeli varstva OP je odgovoren upravljavec, ki mora biti to skladnost tudi zmožen dokazati.

HVALA ZA POZORNOST

gp.ip@ip-rs.si
www.ip-rs.si

Pot do odgovorne odprte znanosti

The path to responsible open science

Urša Opara Krašovec

Univerza v Ljubljani, Fakulteta za elektrotehniko

POVZETEK

Odprta znanost je globalna pobuda, ki se je razvila kot odgovor na številne zapletene izzive sodobne znanosti in družbe, da bi spodbudila odprto izmenjavo znanja in sodelovanje ter izboljšala kakovost, preglednost in vpliv znanstvenega dela. Z odprto znanostjo lahko znanstveni proces postane bolj pregleden, vključujoč in demokratičen. Za izvajanje odgovorne odprte znanosti je potreben celosten pristop, ki upošteva etične, družbene in tehnične vidike. Obravnavane bodo prednosti, ključni izzivi in ključni akterji za implementacijo ekosistema odprte znanosti ter kako k implementaciji načel odprte znanosti pristopa Slovenija. Izpostavljene bodo pogodbene obveznosti med financerji raziskav in raziskovalno organizacijo, povezane z odprto znanostjo, ter nujnost reforme sistema vrednotenja znanosti. Na koncu se bomo dotaknili še vprašanja prednosti in problemov, ko se umetna inteligenca sreča z znanostjo.

SUMMARY

Open Science is a global initiative that has developed in response to many of the complex challenges of modern science and society, with the aim of promoting open knowledge sharing, collaboration and improving the quality, transparency and impact of scientific work. Open science can make the scientific process more transparent, inclusive and democratic. The implementation of responsible open science requires a holistic approach that takes into account ethical, social and technical aspects. The benefits, the key challenges and key actors for the implementation of an open science ecosystem will be discussed, as well as how Slovenia is approaching the implementation of the principles of open science. Contractual obligations between research founding and research performing organisation associated with open science and the need for reform of the science evaluation system will be highlighted. In conclusion we will touch upon the question of advantages and issues when the artificial intelligence meets science.

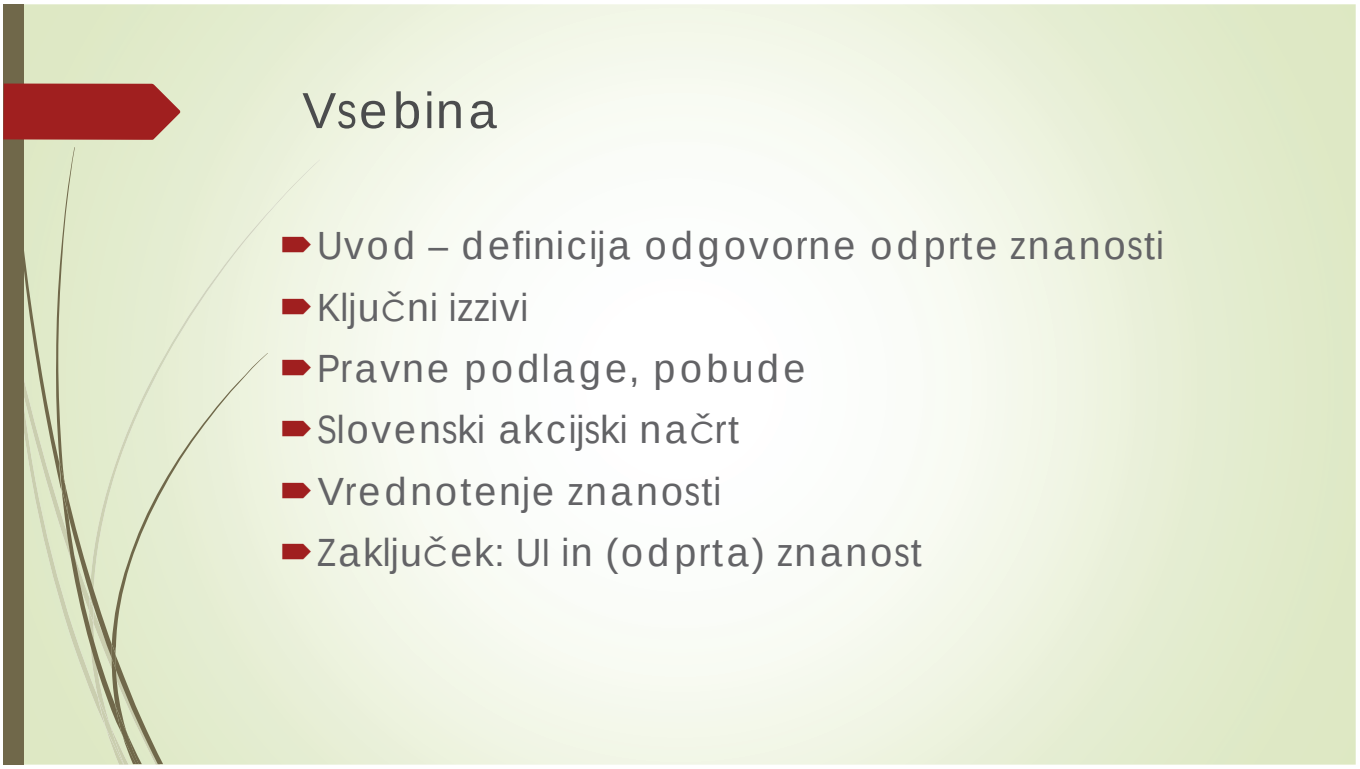
O AVTORJU

Dr. Urša Opara Krašovec je leta 1999 doktorirala iz kemije in je znanstvena svetnica na Fakulteti za elektrotehniko Univerze v Ljubljani. Njeno glavno raziskovalno področje je razvoj materialov za izkoriščanje, pretvorbo in shranjevanje sončne energije. Od leta 2003 dela kot ocenjevalka za Evropsko komisijo. Je aktivna zagovornica enakih možnosti v znanosti in raziskovalne integritete. V okviru projekta Spoznaj sodeluje pri oblikovanju ekosistema odprte znanosti in deluje kot skrbnica podatkov.

ABOUT THE AUTHOR

Dr. Urša Opara Krašovec received her PhD in Chemistry in 1999 and is a Scientific Advisor at the Faculty of Electrical Engineering, University of Ljubljana. Her main research area is the development of materials for solar energy harvesting, conversion and storage. She has been working as an evaluator for the European Commission since 2003. She is an active advocate for equal opportunities in science and research integrity. Through the project Spoznaj she is engaged in the creation of the Open Science Ecosystem and acts as a data steward.


SPOZNAJ
PODPORA PRI UVAJANJU NAČEL ODPRTE ZNANOSTI V SLOVENIJI
UNIVERZA V LJUBLJANI
Pot do odgovorne odprte znanosti
Urša Opara Krašovec
Vitel 2024, 16. in 17. maj 2024, Bled, Slovenija

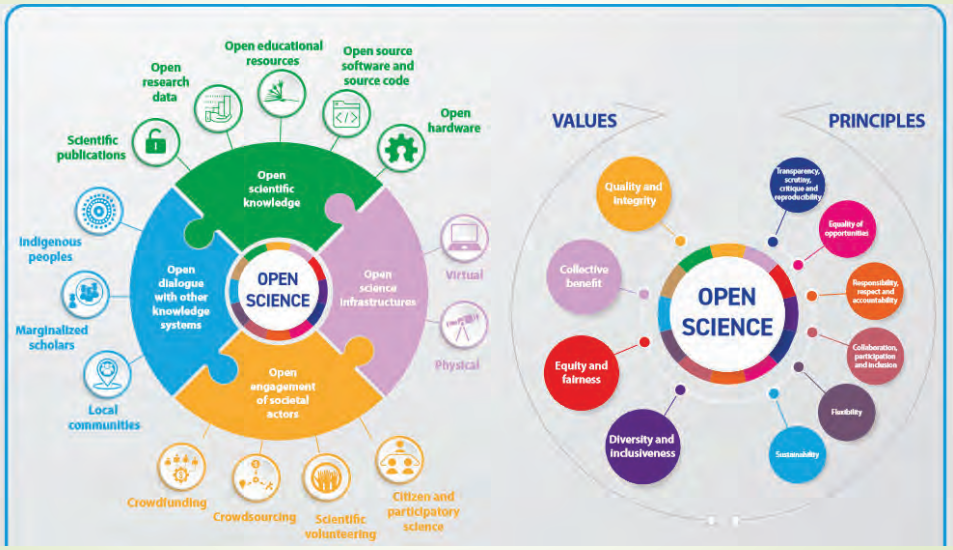

Vsebina
Uvod – definicija odgovorne odprte znanosti
Ključni izzivi
Pravne podlage, pobude
Slovenski akcijski načrt
Vrednotenje znanosti
Zaključek: UI in (odprta) znanost

Zakaj pobuda o odgovorni odprti znanosti?

- Splošno načelo v znanosti: vse teorije je mogoče postaviti pod vprašaj in ugotovljene rezultate je možno ponoviti.
- Prvi znanstveni reviji (1665): *Journal des sçavans* in *Philosophical Transactions of the Royal Society,*

- Večja preglednost,
- večja vključenost in
- demokratičnost znanstvenega procesa.

UNESCO Recommendation on Open Science, the first international standard setting instrument on open science, was adopted by 193 countries in 2021.



https://doi.org/10.54677/UTCD9302

Ključni akterji - prednosti

- Povečana vidnost
- Celovit profil

- Ekonomski prihranek
- Večja ponovljivost raziskav
- Spoštovanje določil financerjev

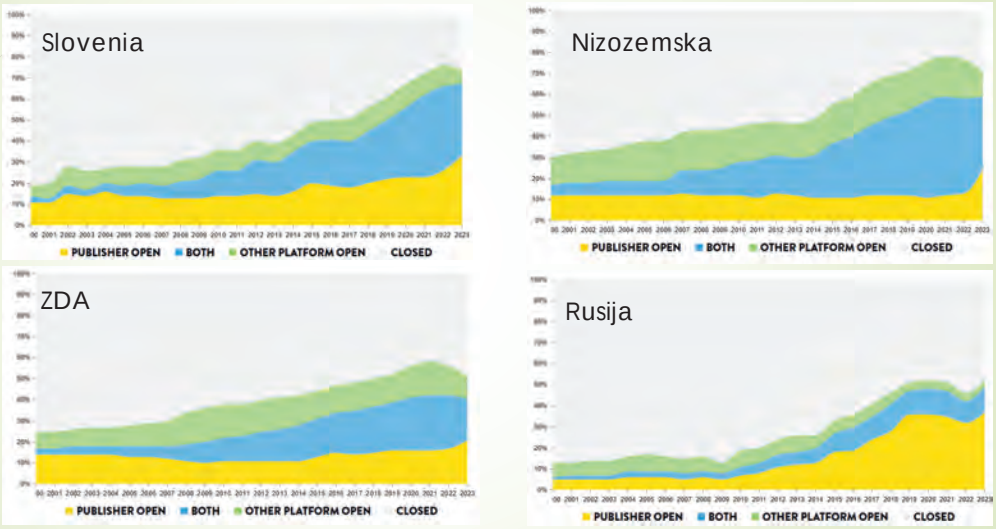
- Pregled nad vloženimi sredstvi
- Potrditev družbenega vpliva raziskav
- Hitrejši prenos idej in pospešitev inovacij

- Za kaj so porabljena proračunska sredstva
- Dostop do rezultatov raziskav, vključenost
- Večje zavedanje o družbenih izzivih

Glavni izzivi odprte znanosti

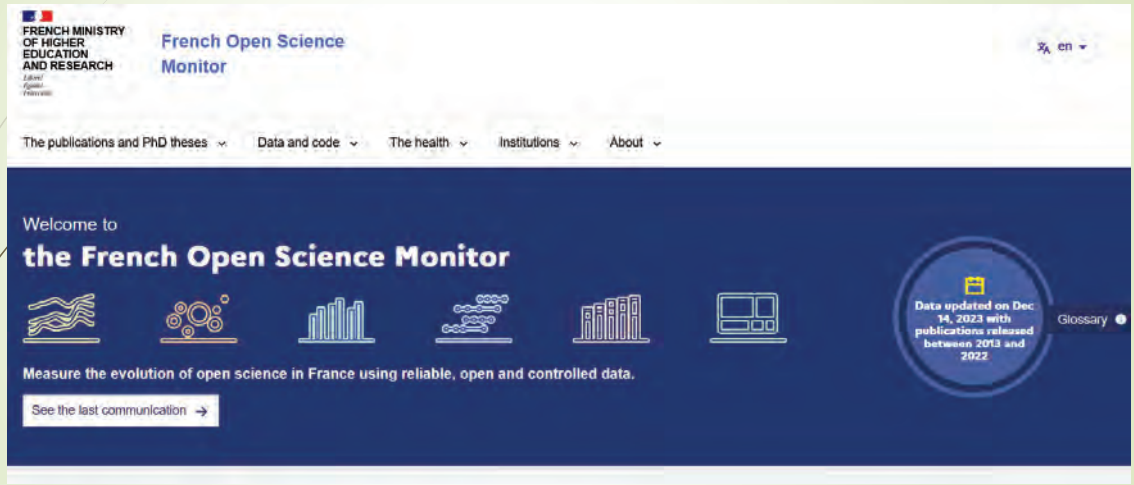
- **Kulturno - sociološki**: (ne)poznavanje prednosti OS; spreminjanje trenutnih praks in delovnih tokov, sistem nagrajevanja.
- **Tehnični:** infrastruktura za varno in dolgoročno hranjenje podatkov, IKT tehnologije za podporo procesov, spremljanje; življenjski cikel podatkov (NRRP, hranjenje, deljenje, metapodatki)
- **Politični:** politična zaveza, razvoj politik in strateško načrtovanje.
- **Organizacijski:** delovno okolje - kako odprto je raziskovanje.
- **Ekonomski:** vložek v razvoj tehničnega, političnega in organizacijskega ekosistema za odprto znanost.
- **Pravni:** GDPR, zasebnost, lastništvo podatkov, avtorske pravice, odgovornost v raziskavah. Zakonodaja, raziskovalne smernice.

COKI Open Access Dashboard - % odprto dostopnih objav skozi čas https://open.coki.ac/;



Diprose, J., Hosking, R., Rigoni, R., Roelofs, A., Chien, T., Napier, K., Wilson, K., Huang, C., Handcock, R., Montgomery, L., & Neylon, C. (2023). A User-Friendly Dashboard for Tracking Global Open Access Performance. The Journal of Electronic Publishing 26(1). doi: https://doi.org/10.3998/jep.3398

https://frenchopensciencemonitor.esr.gouv.fr/



Zakaj bi se ukvarjali z odprto znanostjo, če se nagrajujejo „tradicionalni“ način raziskovanja (zbiranje SICRIS točk)?

- Sprememba vrednotenja
- Pogodbena obveza

Pravne podlage za odprto znanost (geografska razvrstitev)

1) Priporočilo UNESCO o odprti znanosti
2) Zaključki Sveta EU o vrednotenju raziskav in odprti znanosti
3) Obzorje Evropa (Evropska komisija, 2021–2027), 17. čl. pogodbe o sofinanciranju projekta
4) Resolucija o znanstvenoraziskovalni in inovacijski strategiji Slovenije 2030, predvsem poglavje 6.2.
5) Akcijski načrt za odprto znanost
6) Zakon o znanstvenoraziskovalni in inovacijski dejavnosti, čl. 40.-42.
7) Zakon o dostopu do informacij javnega značaja, 6.č čl.
8) Uredba o izvajanju znanstvenoraziskovalnega dela v skladu z načeli odprte znanosti
9) Pravilnik o pogojih za izvajanje knjižnične javne službe, 45(3) čl.
10) Statut UL, 6. in 133. čl.
11) Strategija UL 2022-2027, točka 7.2, kazalec uspešnosti RP2-K6
12) Pravilnik o stabilnem financiranju znanstvenoraziskovalne dejavnosti UL, 25. in 26. čl.
13) Pravilnik o doktorskem študiju UL, predvsem 50. čl.

Dr. Mojca Kotar, UL

ERA, Obzorje Evropa, vrednotenje projektnih vlog, pogodbena obveznost

What ?	How ?	Mandatory in all calls/recommended
Early and open sharing of research	Preregistration, registered reports, preprints, etc.	Recommended
Research output management	Data Management Plan (DMP)	**Mandatory**
Measures to ensure reproducibility of research outputs	Information on outputs/tools/instruments and access to data/results for validation of publications	**Mandatory**
Open access to research outputs through deposition in trusted repositories	• Open access to publications • Open access to data • Open access to software, models, algorithms, workflows, etc.	**Mandatory** for peer-reviewed publications **Mandatory** for research data but with exceptions ('as open as possible, ...') Recommended for other research outputs
Participation in open peer-review	Publishing in open peer-reviewed journals or platforms	Recommended
Involving all relevant knowledge actors	Involvement of citizens, civil society and end-users in co-creation of content (e.g., crowd-sourcing, etc.)	Recommended

Pogodbene obveznosti

- Takojšnji odprti dostop do recenziranih publikacij z licenco CC-BY preko repozitorijev,
- Načrt ravnanja z raziskovalnimi podatki upoštevajoč načelo **FAIR** (F – findable, A – accessible, I – interoperable, R – reusable),
- Shranitev raziskovalnih podatkov v repozitorij in odprti dostop z licenco CC-BY (ali CC-0); „odprto kolikor je mogoče zaprto kolikor je nujno“
- Odprta znanost je del raziskovalne metodologije,

Življenski krog raziskovalnih podatkov
Načrt ravnanja z raziskovalnimi podatki

FAIR
Findable,
Accesible,
Interoperable,
Reusable

OPEN
„as open as possible – as closed as needed“

Vede o življenju

- Elixir, https://rdmkit.elixir-europe.org/data_life_cycle

Družboslovje

- Arhiv družboslovnih podatkov, 2023 https://www.adp.fdv.uni-lj.si/usposobi/ZKG/

RISS 2030 – Akcijski načrt za odprto znanost, Ukrep 6.2 (I)

- 6.2.1 Mednarodno skladen nacionalni ekosistem odprte znanosti
- 6.2.2 Vrednotenje znanstvenoraziskovalne dejavnosti v skladu z načeli odprte znanosti
- 6.2.3 Skladnost rezultatov znanstvenih raziskav z načeli FAIR in odprtost
- 6.2.4 Nacionalna skupnost za odprto znanost
- 6.2.5 Občanska znanost
- 6.2.6 Odprto dostopno nacionalno znanstveno založništvo

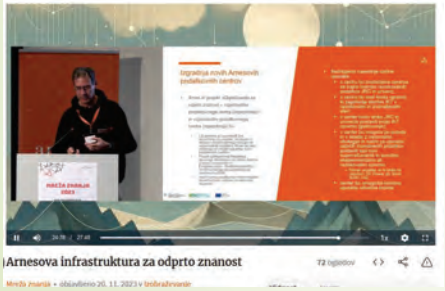
RISS 2030 – Akcijski načrt (II)

Preglednica 2: Akcijski načrt za odprto znanost za izvedbo Resolucije o znanstvenoraziskovalni in inovacijski strategiji Slovenije 2030 — Ukrep 6.2 Odprta znanost za izboljšanje kakovosti, učinkovitosti in odzivnosti raziskav

Zap. št.	Aktivnost	Kazalniki	Nosilec in sodelujoči	Predviden začetek izvajanja	Predviden zaključek izvajanja	Potrebna finančna sredstva v EUR za celotno trajanje aktivnosti	Vir financiranja
U6.2.1	Mednarodno skladen nacionalni ekosistem odprte znanosti						
U6.2.1/1	Prilagoditev nacionalnega ekosistema odprte znanosti za delo po načelih odprte znanosti						
	A6.2.1/1.1: Priprava in sprejem Uredbe o izvajanju znanstvenoraziskovalnega dela v skladu z načeli odpre znanosti kot podzakonski akt Zakona o znanstvenoraziskovalni in inovacijski dejavnosti.	K6.2.1./1.1.1: Uredba o izvajanju znanstvenoraziskovalnega dela v skladu z načeli odprte znanosti je sprejeta na Vladi Republike Slovenije.	Vlada Republike Slovenije, MVZI	2023	2023		
U6.2.1/2	Prilagoditev JRO za delovanje po načelih odprte znanosti						
	A6.2.1/2.1: Izvedba projekta za prilagoditev JRO in CTK za delovanje po načelih odprte znanosti.	K6.2.1/2.1.1: Ustrezno načrtovan, objavljen in izveden projekt z vključitvijo čim večjega števila JRO.	MVZI, JRO, CTK	2023	2026	(4.071.720)	NOO (MVZI)
	A6.2.1/2.2: Podpora kadrovskemu razvoju JRO na področju podpornih struktur, neobhodnih za izvajanje aktivnosti odprte znanosti (podatkovni svetovalci, podatkovni knjižničarji, arhivisti, vzdrževalci infrastruktur za odprto znanost).	K6.2.1/2.2.1: Stabilno financiranje JRO preko institucionalnega stebra je načrtovano tudi z vidika izvedbe tega ukrepa. Sklenjene pogodbe z JRO ji vključujejo tudi neobhodne kadrovsko	MVZI, ARIS, JRO	2026	2030	4.000.000	ARIS

RISS 2030 – Akcijski načrt (III)

U6.2.1/3	Infrastruktura za odprto znanost						
	A6.2.1/3.1: Vzpostavitev sveta nacionalne infrastrukture odprte znanosti za podporo razvoju nacionalnega ekosistema infrastrukture odprte znanosti (glej: U6.2.4/2).	K6.2.1/3.1.1: Vzpostavljen svet nacionalne infrastrukture odprte znanosti.	MVZI, SSOZ (UKM UM, Arnes, IZUM, JRO, SLING)	2023	2023		
	A6.2.1/3.2: Razvoj, vzdrževanje in delovanje nacionalne infrastrukture odprte znanosti.	A6.2.1/3.2.1: Delovanje, vzdrževanje in nadaljnji razvoj nacionalne infrastrukture odprte	UM FERI, (ARNES, IZUM)	2023	2030	1.800.000	ARIS, MVZI
	A6.2.1/3.9: Vzpostavitev dveh ločenih podatkovnih centrov za dolgotrajno hrambo raziskovalnih podatkov.	K6.2.1/3.9.1: Vzpostavljena bosta dva ločena podatkovna centra za dolgotrajno hrambo raziskovalnih podatkov v okviru javnega infrastrukturnega zavoda Arnes. Oba vzpostavljena repozitorija za raziskovalne podatke bosta skladna z določili EOSC.	ARNES (UM FERI, IZUM)	2023	2026	(11.500.000)	NOO (MVZI)



20. november 2023, **Mreža znanja**, Arnesova infrastruktura za odprto znanost
https://video.arnes.si/watch/84p2jmt88d02

https://projekt-spoznaj.si/

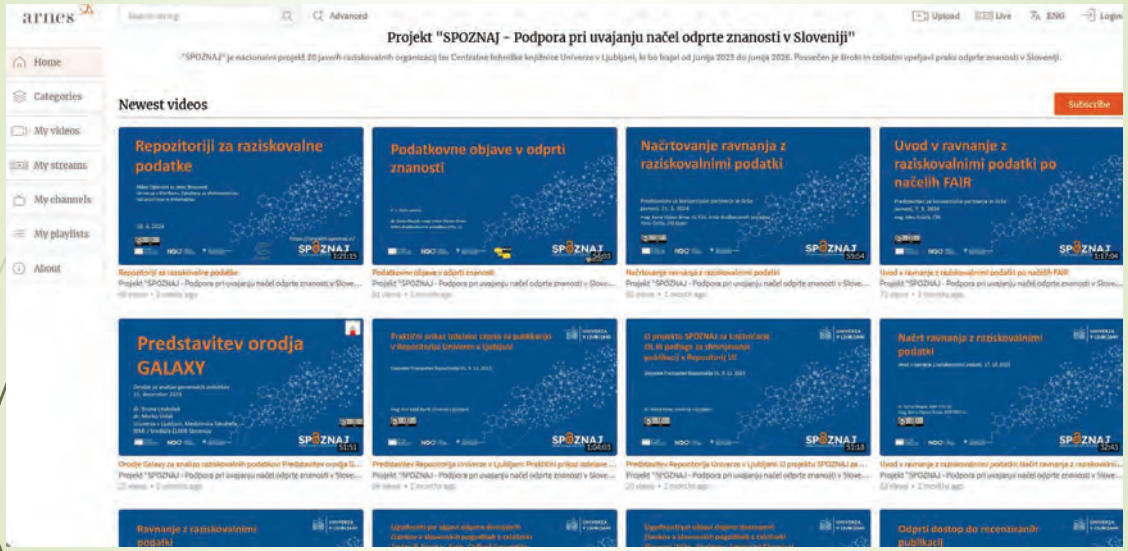
Delovni sklopi projekta SPOZNAJ

- DS 1: Upravljanje in koordinacija konzorcija
- DS 2: Ozaveščanje, **promocija** in komuniciranje dejavnosti konzorcija
- DS 3: Izvedba **usposabljanj** o praksah odprte znanosti in deljenju raziskav po načelih FAIR
- DS 4: Priprava in izvedba **specialističnih izobraževanj** za delovanje po načelih odprte znanosti
- DS 5: Priprava in distribucija **priročnika** za delovanje po načelih odprte znanosti s poudarkom na ravnanju z raziskovalnimi podatki
- DS 6: **Celostna prilagoditev delovanja po načelih odprte znanosti** in evalvacija rezultatov konzorcijskih partnerjev

UL: Podatkovni skrbniki – Data-steward

https://projekt-spoznaj.si/

https://video.arnes.si/en/?channel=9lcqz6j71lg7



Vrednotenje znanosti

- Kvantiteta pred kvaliteto objav, „*publish or perish*“
- IF revije kot merilo kvalitete objave
- SICRIS točke – slovenska posebnost

POSLEDICE:

- Neponovljivost raziskav (prirejanje, ponarejanje, plagiatorstvo, druge slabe prakse), umik objav, škandali – nezaupanje v znanost
- Objavljanje le pozitivnih rezultatov raziskav
- „Znanstvena bolha“
- …

https://coara.eu/agreement/the-commitments/

Znanstvena Bolha: kupim - prodam avtorstvo na znanstveni objavi (lažni ali resnični)

Cena avtorstva:
od nekaj sto do nekaj tisoč evrov, odvisno od mesta avtorja na članku in revije v kateri bo članek objavljen.

Razvoj trga:

- način vrednotenja: število in ne na kakovost objav (zbirateljstvo točk), „publish or perish“
- koruptivne prakse znanstvenikov v vlogi urednikov
- zloraba umetne inteligence za generiranje lažnivih znanstvenih objav
- omenja se tudi prodaja avtorstva v zameno za kritje stroška objave založniku

„In May 2022, Springer Nature retracted a paper for the first time over suspicions that some of the authorships were paid-for, …, since then (januar 2023) retracted 11 papers in a further 5 journals over similar concerns. More investigations are under way.“


NEWS | 16 January 2023

Multimillion-dollar trade in paper authorships alarms publishers

Journals have begun retracting publications with suspicious links to sites trading in author positions.

Holly Else

Buying a fraudulent authorship can cost hundreds or thousands of dollars. Credit: Getty

https://www.nature.com/articles/d41586-023-00062-9

Reforma vrednotenja znanosti tudi v podporo odprti znanosti

- San Francisco deklaracija, DORA 2013, Leiden Manifesto, HK principles,
- Plan S – financerji, 2018
- COARA – Coalition for Advancing Research Assessment, 2022 - 724 organizacij, vsi kontinenti

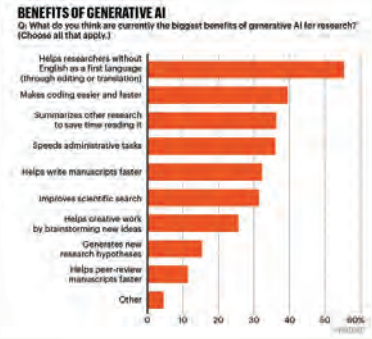
- Obzorje Evropa (ERC razpisi) – ocenjevanje
- Uredba o izvajanju znanstvenoraziskovalnega dela v skladu z načeli odprte znanosti (2023), ARIS / NAKVIS – 2024, Raziskovalne organizacije - 2027

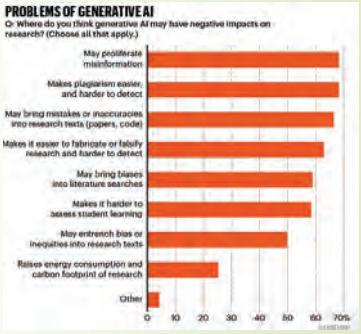
https://coara.eu/agreement/the-commitments/

UI in (odprta) znanost

?

https://www.nature.com/articles/d41586-023-02980-0





COARA WG „Ethics and Research Integrity Policy in Responsible Research Assessment for Data and Artificial Intelligence (ERIP)“

Hvala za vašo pozornost!

Baze podatkov v lokalnem računalniškem oblaku – od transakcijskih in analitičnih podatkov do vektorske baze za velike jezikovne modele

Database in the local cloud – from transactional and analytical data to vector database for LLM applications

Robert Korošec

Oracle Slovenija

POVZETEK

Oblak in oblačne storitve prinašajo veliko koristi, kot so elastičnost, plačilo po uporabi, administrirano okolje, veliko število že narejenih storitev in maksimalna avtomatizacija celotnega IT sklada. Po drugi strani se morajo nekatere pomembne storitve izvajati na lokaciji uporabnika ali čim bližje uporabniku. Razlogi so lahko tehnične narave (latenca, kompleksnost sistema), zakonske zahteve in regulative ali pomisleki glede skrbništva in upravljanja s podatki.

Oracle Cloud@Customer rešitev je privatni oblak pri uporabniku – po eni strani je na delu avtomatizacija celotnega IT sklada, kot je to v oblaku, po drugi strani je postavitev kar na lokaciji uporabnika. Rešitev je sestavljana iz najema strojne in programske opreme ter Oracle baze podatkov, ki omogoča izvajanje transakcijskih, analitičnih aplikacij ter uporabo ostalih sodobnih struktur podatkov (JSON, graf, prostorski in vektorski podatki). Prednost »vse v enem« je v tem, da se podatki v različnih strukturah nahajajo v eni sami bazi, kar poenostavi visoko razpoložljivost, okrevanje po katastrofi, varnost podatkov in administracijo celotnega sistema.

SUMMARY

The cloud and cloud services bring many benefits, such as elasticity, pay-as-you-go, managed environment, a large number of already-made services and maximum automation of the entire IT stack. On the other hand, some important services must be performed at the user's location or as close as possible to the user. The reasons may be of a technical nature (latency, system complexity), legal and regulatory requirements, or concerns regarding data custody and management.

The Oracle Cloud@Customer solution is a private cloud at the user's location - on the one hand, the automation of the entire IT stack is at work, as it is in the cloud, on the other hand, it is set up right at the user's location. The solution consists of renting hardware and software and an Oracle database, which enables the implementation of transactional and analytical applications and the use of other modern data structures (JSON, graph, spatial and vector data). The advantage of "all-in-one" is that data in different structures resides in a single database, which simplifies high availability, disaster recovery, data security and administration of the entire system.

O AVTORJU

Robert Korošec je od leta 1999 zaposlen v podjetju Oracle Software, kjer je delal kot svetovalec za Oracle baze podatkov in Oracle Exadata sisteme. V zadnjem letih deluje na področju Oracle Cloud platforme, primarno na uporabi oblačnih baz podatkov, prediktivne analitike in analize velepodatkov (BigData).

ABOUT THE AUTHOR

Robert Korošec has been employed at Oracle Software since 1999, where he worked as a consultant for Oracle databases and Oracle Exadata systems. In recent years, he has been working in the field of the Oracle Cloud platform, primarily on the use of cloud databases, predictive analytics and BigData data analysis.

ORACLE

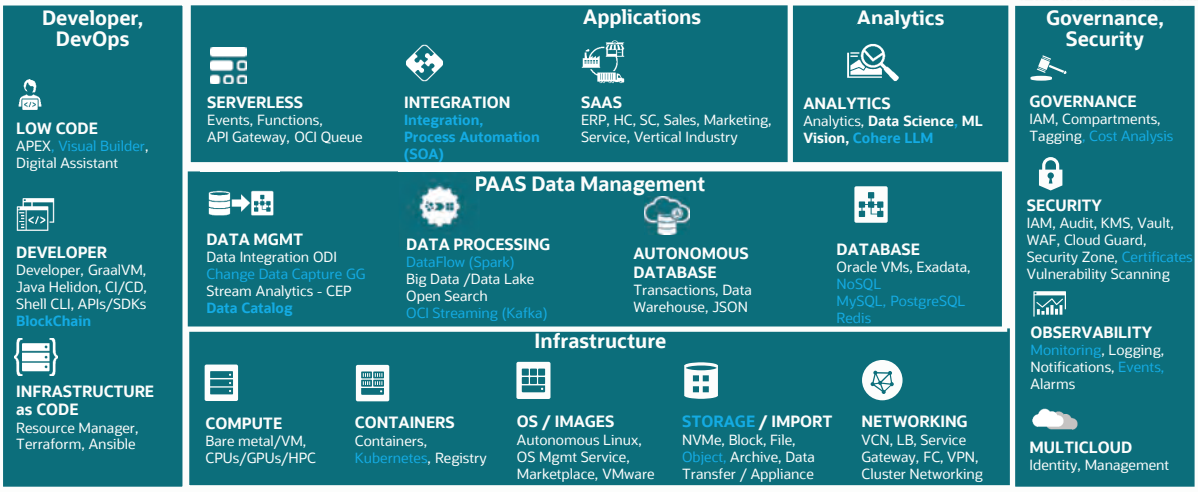
Baze podatkov v lokalnem računalniškem oblaku

Od transakcijskih in analitičnih podatkov do vektorske baze za velike jezikovne modele

Robert Korošec
Technical Consultant
Oracle
Maj 2024

Oracle Cloud Infrastructure (OCI)

Complete cloud services



2

Deploy in our public, government, or sovereign regions

$16B+
Cloud Applications
Cloud Infrastructure
run rate

57%
cloud infrastructure
consumption growth

42
cloud regions

$8.2B
CapEx in the
last 12 months

100% renewable energy by 2025

3 Copyright © 2023, Oracle and/or its affiliates

INTRODUCING

Oracle Sovereign Cloud for European Union

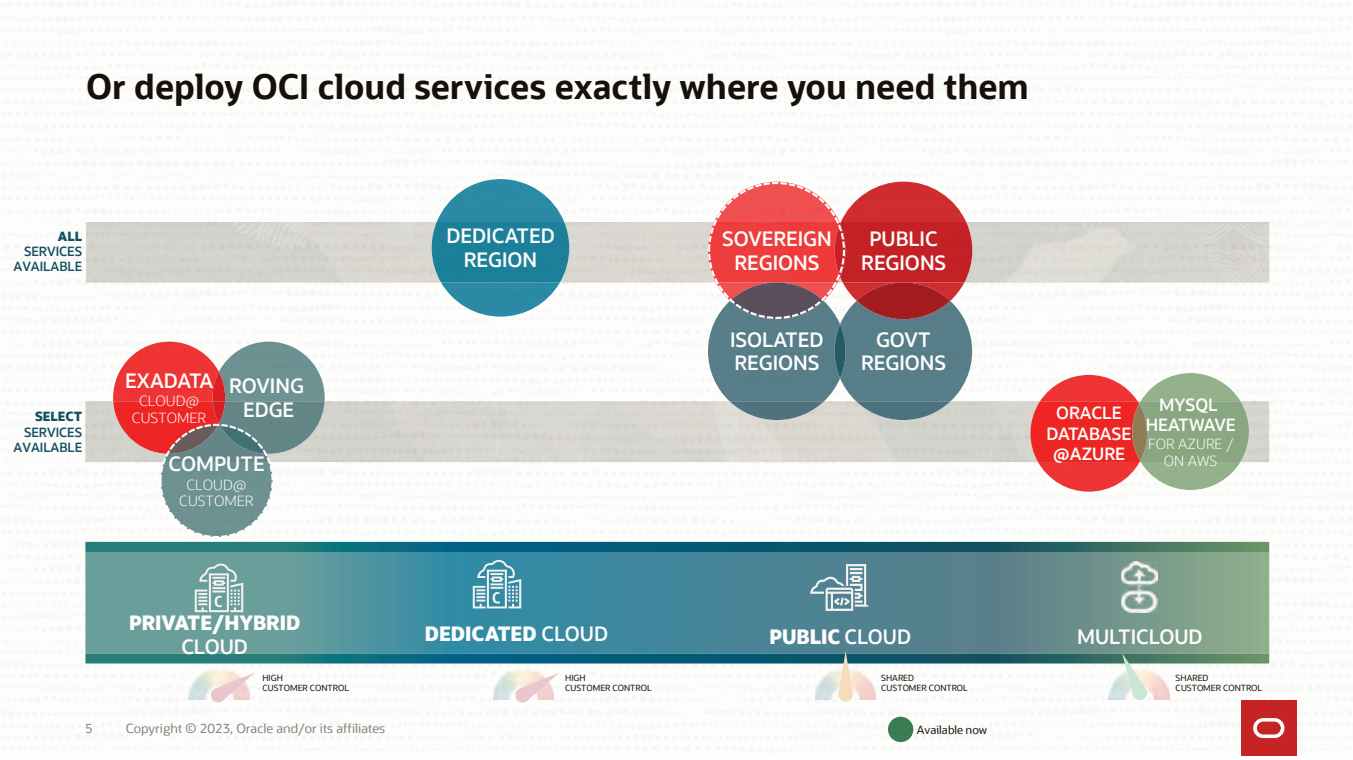
A new, EU-specific cloud offering that is **isolated** from Oracle's global public cloud regions

Located in the EU, operated by EU residents

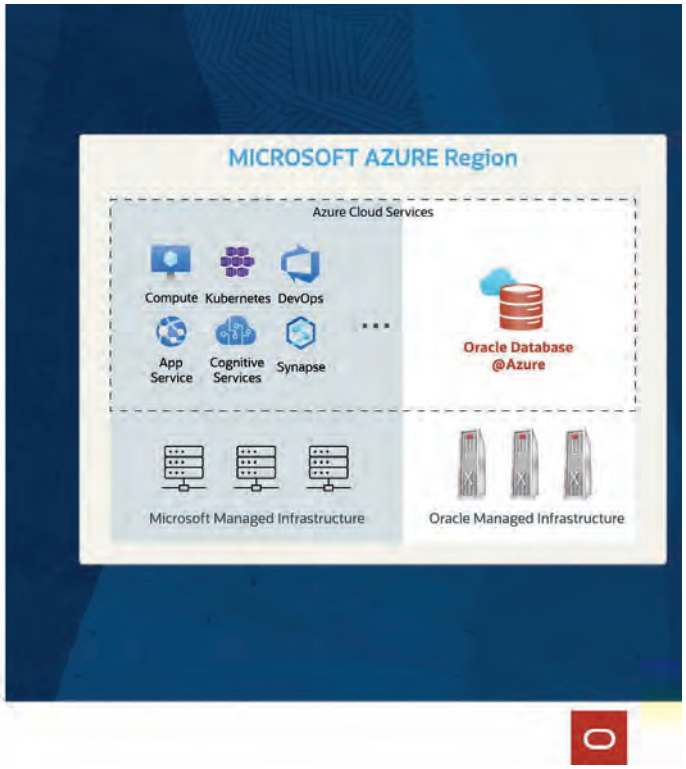
Offers data security and data sovereignty

Access to the same services, value, and innovation as Public Cloud

4


Or deploy OCI cloud services exactly where you need them
ALL SERVICES AVAILABLE
SELECT SERVICES AVAILABLE
DEDICATED REGION
SOVEREIGN REGIONS
PUBLIC REGIONS
ISOLATED REGIONS
GOVT REGIONS
EXADATA CLOUD@ CUSTOMER
ROVING EDGE
COMPUTE CLOUD@ CUSTOMER
ORACLE DATABASE @AZURE
MYSQL HEATWAVE FOR AZURE / ON AWS
PRIVATE/HYBRID CLOUD
DEDICATED CLOUD
PUBLIC CLOUD
MULTICLOUD
HIGH CUSTOMER CONTROL
HIGH CUSTOMER CONTROL
SHARED CUSTOMER CONTROL
SHARED CUSTOMER CONTROL
5
Copyright © 2023, Oracle and/or its affiliates
Available now


Announcing
Oracle Database@Azure
Oracle and Microsoft will deliver Oracle database services on OCI in Microsoft Azure datacenters
ORACLE
Microsoft
Copyright © 2023, Oracle and/or its affiliates

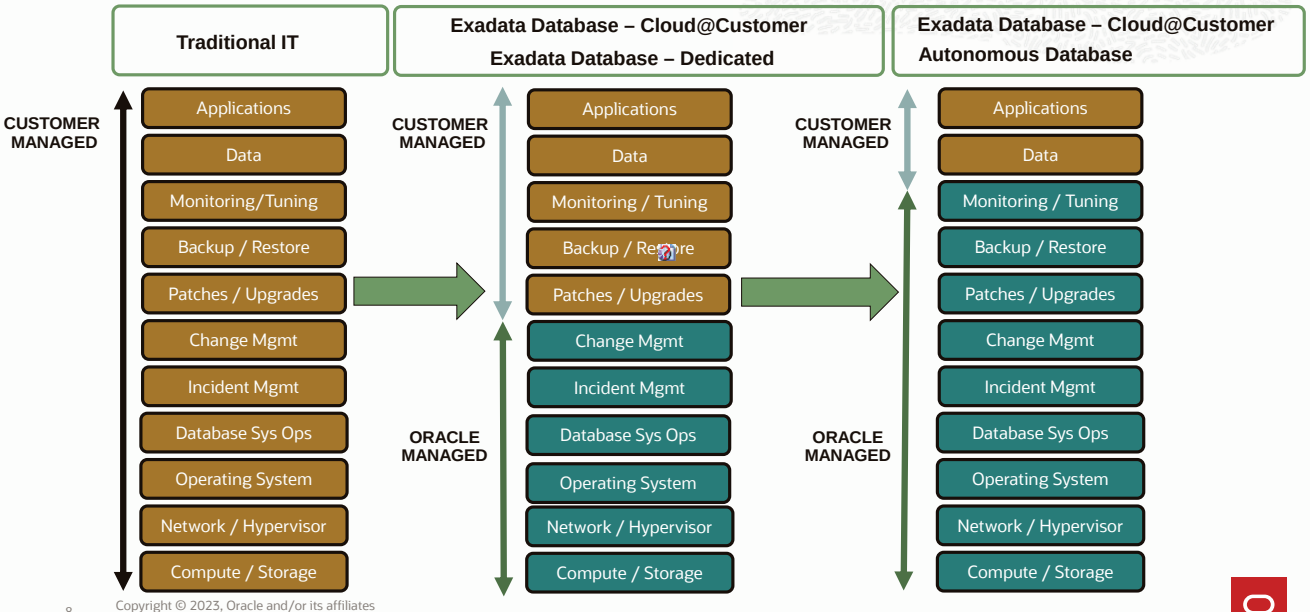

MICROSOFT AZURE Region
Azure Cloud Services
Compute
Kubernetes
DevOps
App Service
Cognitive Services
Synapse
Oracle Database @Azure
Microsoft Managed Infrastructure
Oracle Managed Infrastructure

Customers want to move database workloads to the cloud

- Simplified, automated management
- Cloud-native application development
- Pay-as-you-go economics
- With high performance, availability, scalability and security

Copyright © 2023, Oracle and/or its affiliates

Cloud Automatization



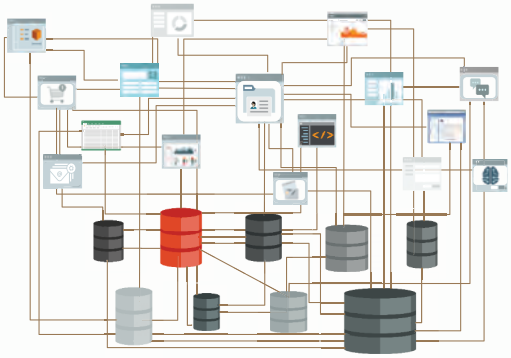
Copyright © 2023, Oracle and/or its affiliates

But not every organization or workload can easily use the public cloud

Data Residency and Security
- Regulations or policies require data to be local
- Requirements to protect data in specific ways

Response Time
- Real-world systems require low latency
- Hard to disentangle one system from others

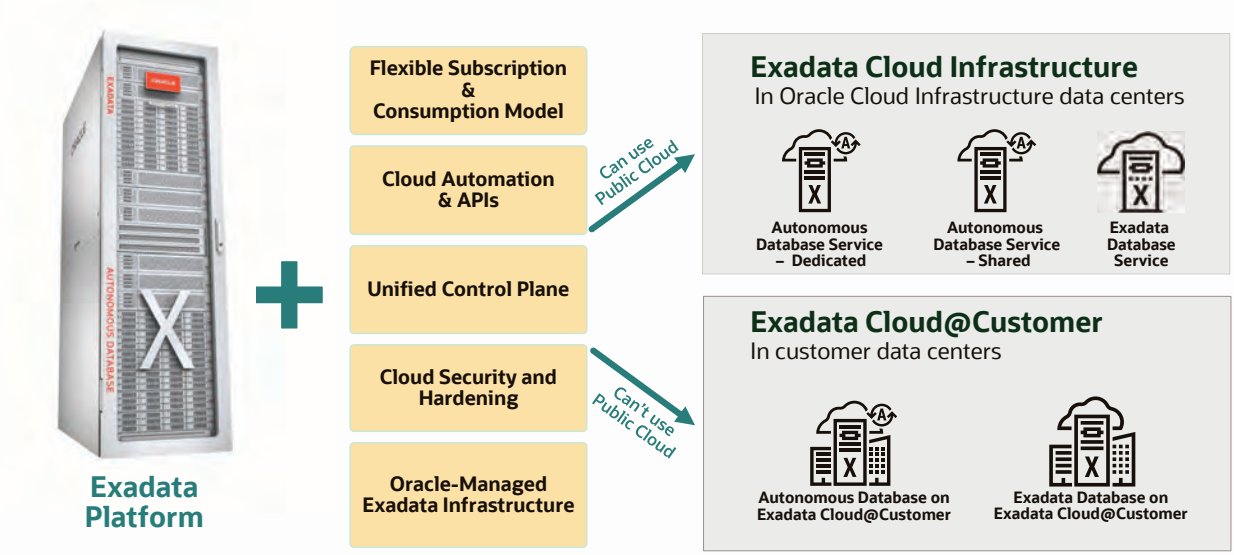
Perceived Risk
- Concerns about multi-tenant cloud
- Concerns about cloud provider access to data



9 Copyright © 2023, Oracle and/or its affiliates

The best database platform for any deployment

The Same Exadata in OCI Public Cloud and Customer Data Centers

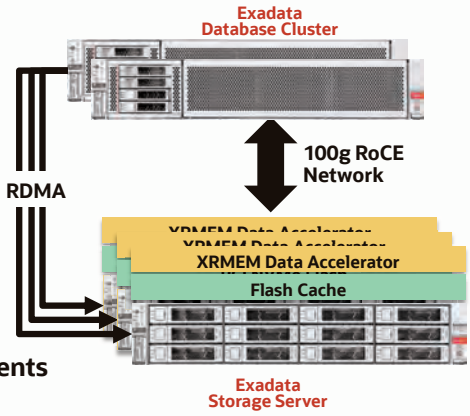


10 Copyright © 2023, Oracle and/or its affiliates

Fastest OLTP

Unique: Innovative, modern technologies tuned for high-volume, low-latency random I/O

- 100 Gb/s RDMA over Converged Ethernet 100 (RoCE) network fabric
- Exadata RDMA Memory (XRMEM) Data Accelerator
- Scale-out Database & Storage servers
- Automatic data tiering between XRMEM, Flash Cache and Disk



Unique: Elimination of DB cluster coordination bottlenecks

- Direct-to-Wire Protocol = 3x faster inter-node OLTP messaging
- Smart Fusion Block Transfer eliminates inter-node log write
- RDMA protocol coordinates transactions between nodes

Unique: Instant detection, handling of failed/failing components

- Automatic discovery of server failures without timeout
- Sub-second redirection of I/Os around sick devices

11 Copyright © 2023, Oracle and/or its affiliates

Fastest Cloud Analytics

Smart Scan (SQL Offload)

- Data-intensive processing* runs in massively parallel Exadata Storage, bypassing network bottlenecks and freeing up DB CPUs

Tiered Flash Cache

- Active data is automatically cached on PCI NVMe Flash, inactive data on low cost, high-capacity disks

Storage Indexes

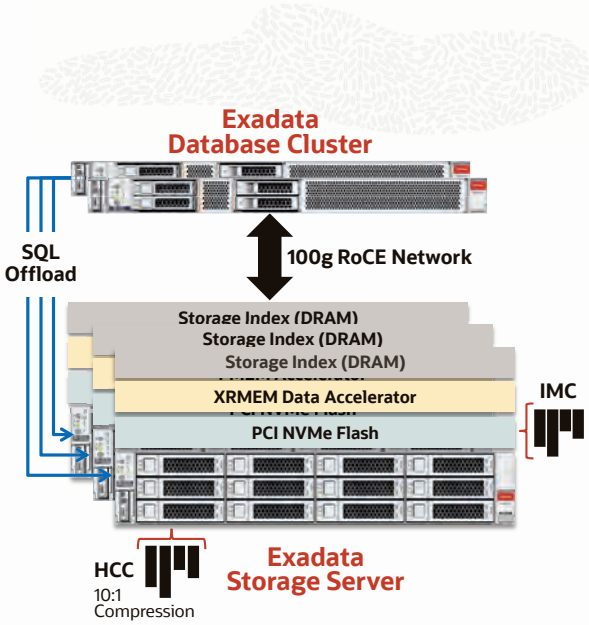
- Eliminates I/O not relevant to a particular query

Hybrid Columnar Compression (HCC)

- Compressed, columnar format in storage, saving space, reducing I/O, speeding analytic queries

In-Memory Columnar (IMC)

- Extends In-Memory database performance to higher capacity Flash memory in storage



*Includes long-running SQL queries, backups, decryption, aggregation, data mining

12 Copyright © 2022, Oracle and/or its affiliates

Exadata Cloud@Customer is Easy to Adopt

The simplest way for mission-critical apps to use cloud databases

Maintain governance and control at all times

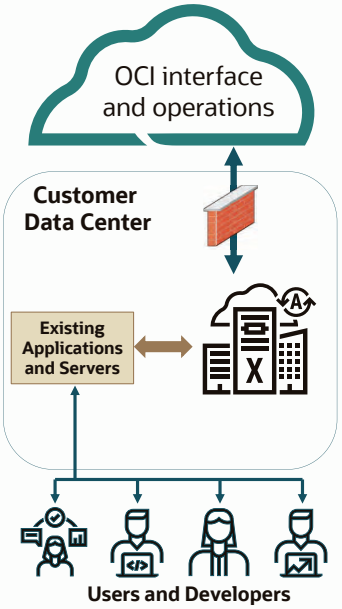
- Data in customer data centers, behind customer firewalls
- **Centralized governance and security applied to all databases**
- **Oversight over patch timing and remote Oracle cloud operators**
- Oracle MAA and Recovery Appliance integration for business continuity and data protection

Lift-and-shift existing on-premises databases

- Existing applications simply connect and run
- No application or database changes needed
- Runs Autonomous Database and Exadata Database concurrently

Accelerate time to value, convert CapEx to OpEx, and reduce TCO

- Online consumption scaling for pay-per-use vCPU economics (automated with Autonomous Database)
- No infrastructure administration
- No database administration when running Autonomous Database
- Self-service database cloud for developers



13 Copyright © 2023, Oracle and/or its affiliates

Operator Access Control (OpCtl)

Prevent, Detect, and Respond

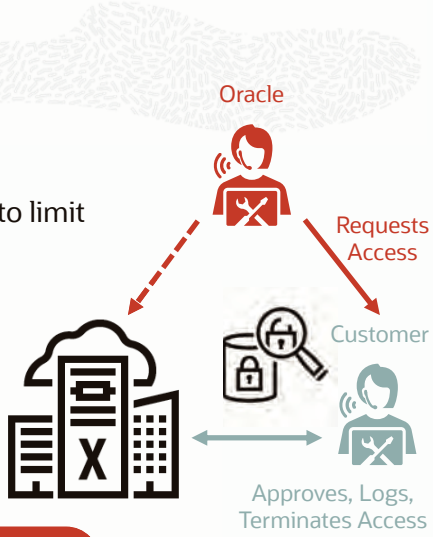
OpCtl enables customers to grant, audit, and revoke access to Exadata Cloud@Customer infrastructure managed by Oracle

Customers control access to infrastructure by Oracle operator to limit

- when they have access
- components they can access
- commands they can execute

Observe and record Oracle operator commands and keystrokes that Oracle staff execute

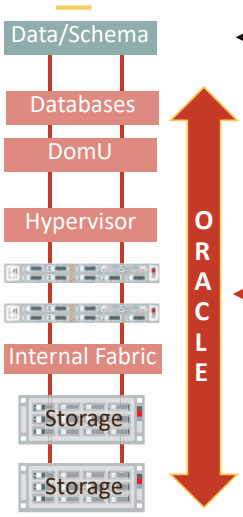
Terminate Oracle operator connections at discretion



Enhanced Security for Regulated Industries

14 Copyright © 2024, Oracle and/or its affiliates, All Rights Reserved.

Autonomous Management Model in Your Data Center

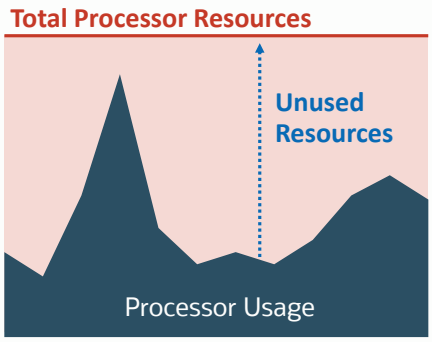


- Customer Controls DB Users, Schemas, Encryption Keys
  - Oracle Database 19c or later only
- Oracle manages, and controls
  - Hypervisor, DB servers, storage servers, InfiniBand network, etc.
  - DomU, Container Databases
  - No customer access
  - Oracle Owns all issues

 Copyright © 2020 Oracle and/or its affiliates.

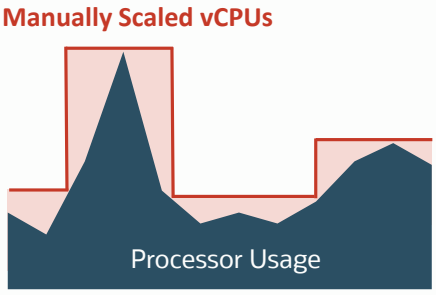
Online, Elastic Scaling with Exadata Cloud@Customer

Pay Only for What You Use



On-Premises – Static

Purchase server processors and software licenses for highest projected peak load



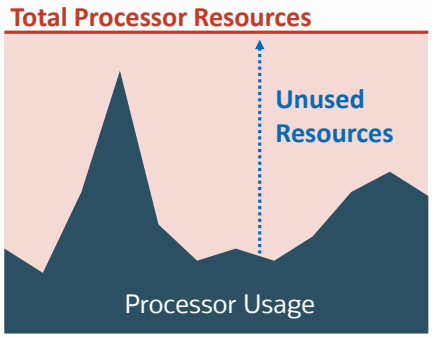
Exadata Database Service – Elastic

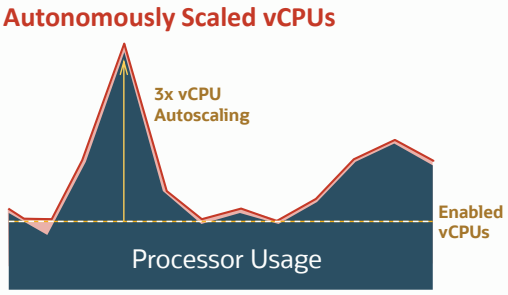
Adjust enabled vCPUs to match actual workload via APIs and web UI - vCPUs are charged per second

 Copyright © 2023, Oracle and/or its affiliates

Online, Elastic Scaling with Exadata Cloud@Customer

Even better with Autonomous Database





On-Premises – Static

Purchase server processors and software licenses for highest projected peak load

Autonomous Database – Self-scaling

Automatically scales vCPU consumption based on dynamic workload demands, in real-time

17 Copyright © 2023, Oracle and/or its affiliates

Cost-Effective Software Licensing Models

Subscribe to infrastructure and choose License Included or Bring Your Own License (BYOL)

License Included Pricing

Ideal for organizations with new workloads and dynamic utilization

- Includes Oracle Database Enterprise Edition with all options and management packs at one low price
- Consumption-based pricing for software and vCPUs, includes software support and paid for with Universal Credits

Bring Your Own License Pricing

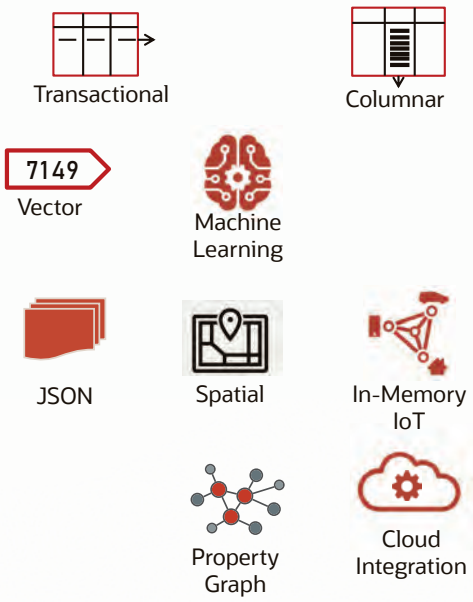
Ideal for organizations moving existing workloads with consistent usage to the cloud

- Utilize existing on-premises licenses and pay software support for them
- Very-low, compute-only consumption pricing, paid for with Universal Credits
- Includes Transparent Data Encryption, Data Safe, Oracle Machine Learning, and select management packs at no additional cost

18 Copyright © 2023, Oracle and/or its affiliates

Oracle Autonomous Database
Under the hood – A converged database

- Transactional for OLTP Applications
- Columnar for DWH
- Vector Store for Vector Embeddings
- AutoML for simple integrated Machine Learning
- Native JSON for Document Data
- Spatial for Geo-coded and Spatial data
- Persistent Memory Store for Lowest Latency IOT
- Property Graph for connected data
- Cloud SQL for integrating Object Store Data Lake



19 Copyright © 2021, Oracle and/or its affiliates

Vectors in AI represent semantics of unstructured data such as images, documents, videos, etc.

A vector is a sequence of numbers, called dimensions, used to capture the important "features" of the data

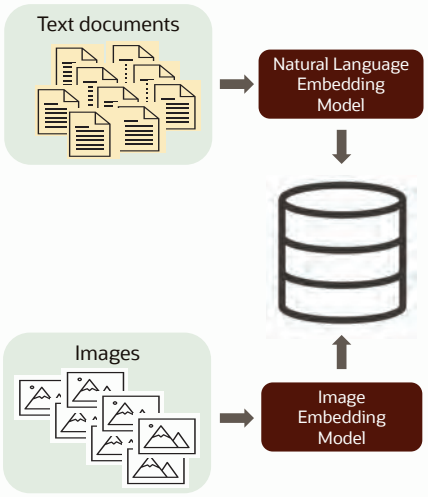
Vectors represent the semantic content of data, not the underlying words or pixels

Vectors generated using deep learning embedding models

20 Oracle DatabaseWorld @ CloudWorld Copyright © 2023, Oracle and/or its affiliates

Generating embeddings

Use an embedding model



Text Vector Table

id	vector	text
1	[0.8, 0.5, 1.6, -2.5, ...]	"It was the best of times, it was the worst of times, it was.."
2	[1.1, 0.3, 0.6, -1.3, ...]	"It is a truth universally acknowledged, that a single man.."
3	[1.3, 0.1, 0.2, -1.1, ...]	"It was a bright cold day in April, and the clocks were striking.."
...	...	...

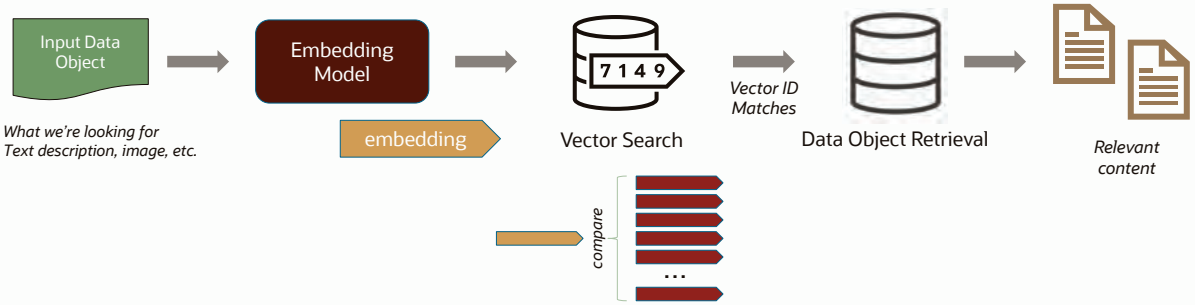
Image Vector Table

id	vector	Image
1	[0.5, 1.5, 2.6, -1.1, ...]	
2	[1.0, 0.9, 1.6, -1.3, ...]	
3	[0.6, 1.1, 1.3, -0.9, ...]	
...	...	...

 Copyright © 2023, Oracle and/or its affiliates

Simple vector search

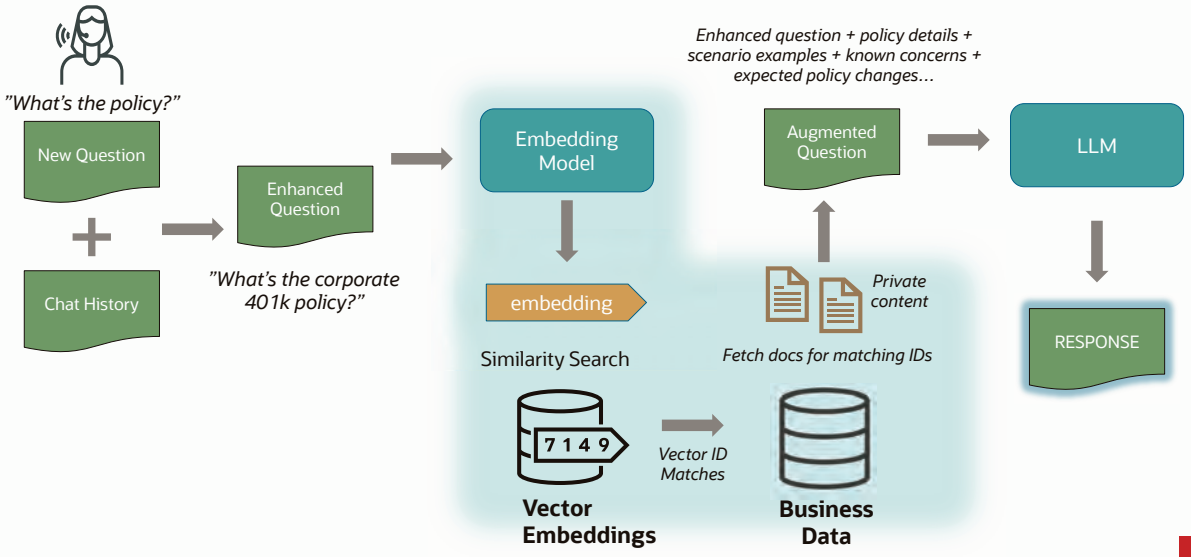
Pipeline



 Copyright © 2023, Oracle and/or its affiliates

LLM-based Chatbot with “Enterprise Knowledge”

RAG – Retrieval Augmented Generation



 Copyright © 2023, Oracle and/or its affiliates

Database cloud services in customer data centers address critical needs

- High security and full data residency behind customer-controlled firewalls
- Low latency connectivity with existing applications and data center resources
- Reduced management and self-service for database developers via cloud automation
- The same consumption model and economics as the public cloud

 Copyright © 2023, Oracle and/or its affiliates

Our mission is to help people see
data in new ways, discover insights,
unlock endless possibilities.

Primeri uvedbe GenAI projektov nad poslovnimi podatki in procesi v praksi

Examples of implementation of GenAI projects on business data and processes in practice

Rok Rogelj

Microsoft

POVZETEK

Prispevek obravnava ključne strategije in investicije Microsofta v GenAI in jezikovne modele ter načrte za razvoj končnih produktov v sodelovanju s partnerji. Poudarek je na varovanju zaupnosti in dostopnosti podatkov, kar predstavlja ključno prednost Microsoftovih rešitev Azure OpenAI v primerjavi z drugimi ponudniki velikih jezikovnih modelov, ki so sicer odlični za zasebno uporabo.

Predstavljen je primer globalne transformacije delovnih procesov z uporabo tehnologije GenAI, s posebnim poudarkom na dveh večjih konkretnih letošnjih projektih v regiji - v zavarovalništvu in javnem sektorju. Razprava zajema različne pristope k nadzoru podatkov in njihovi izpostavljenosti.

Prispevek osvetljuje trenutne ponudbe različnih jezikovnih modelov, vključno z razlikami med velikimi (LLM) in manjšimi jezikovnimi modeli (SLM), ter pogostimi primeri uporabe AI tehnologij v praksi, ki se delijo na “izboljšave interne produktivnosti zaposlenih” in “izboljšanje produktov za končne stranke”.

Zaključek je namenjen najnovejšim razvojnim načrtom na podlagi OpenAI za Enterprise okolja, vključno s generiranjem videa, slik in zvočnih obdelav, tudi v slovenskem jeziku, ki obljubljajo nadaljnje izboljšave v interakciji med človekom in strojem. Predstavitev ponuja tudi vpogled v praktične aplikacije AI tehnologij in njihov vpliv na poslovne procese in podatkovne operacije.

SUMMARY

This article discusses key strategies and Microsoft's investments in GenAI and language models, as well as plans for developing final products in collaboration with partners. The focus is on safeguarding confidentiality and data accessibility, which represents a key advantage of Microsoft's Azure OpenAI solutions compared to other providers of large language models, which are otherwise excellent for private use.

Subsequently, the article presents examples of global transformation of work processes using GenAI technology, with special emphasis on two major concrete projects in the region this year - in the insurance and public sectors. The discussion coveres various approaches to data control and their exposure.

The lecture also illuminates current offerings of different language models, including differences between large (LLM) and smaller language models (SLM), and common use cases of AI technologies in practice, divided into "improvements in internal employee productivity" and "product enhancements for end customers."

The concluding part is dedicated to the latest development plans based on OpenAI for enterprise environments, including the generation of video, images, and audio processing, even in the Slovenian language, promising further improvements in human-machine interaction. The presentation offers insights into the practical applications of AI technologies and their impact on business processes and data operations.

O AVTORJU

Rok Rogelj ima več kot 15 let izkušenj na področju informacijskih tehnologij. Jasno razume poslovno vrednost sodobnih digitalnih tehnologij v več panogah, zlasti v maloprodaji, proizvodnji in logistiki. Strankam pomaga pri optimizaciji, rasti, širitvi in reševanju izzivov digitalne transformacije. G. Rogelj je navdušen nad uvajanjem tehnologij, ki opolnomočijo ljudi, organizacije in njihove poslovne potrebe. Ima močno ozadje pri korporativnih in podjetniških strankah. Njegova trenutna strast so tehnologije v oblaku, zlasti različne baze podatkov, napredna analitika in uporaba umetne inteligence v teh zbirkah podatkov. G. Rogelj ima dokazano razvite veščine in sposobnost usmerjanja ekip pri sodelovanju s strankami.

ABOUT THE AUTHOR

Rok Rogelj has over 15 years of experience in IT. He has a clear understanding of the business value of modern digital

technologies in several industries, especially in retail, manufacturing and logistics. He helps customers to optimize, grow, expand and solve their digital transformation challenges. Mr. Rogelj is passionate about deploying technologies which empower people, organization and their business needs. He has strong background in Corporate and Enterprise customers. His current passions are cloud technologies, especially different Databases, Advanced Analytics and AI on top of Data in these Databases. Mr. Rogelj has a proven track record, highly developed skills and the ability to drive teams in customer engagements continuously.

Microsoft Azure

Primeri uvedbe GenAI projektov nad poslovnimi podatki in procesi v praksi

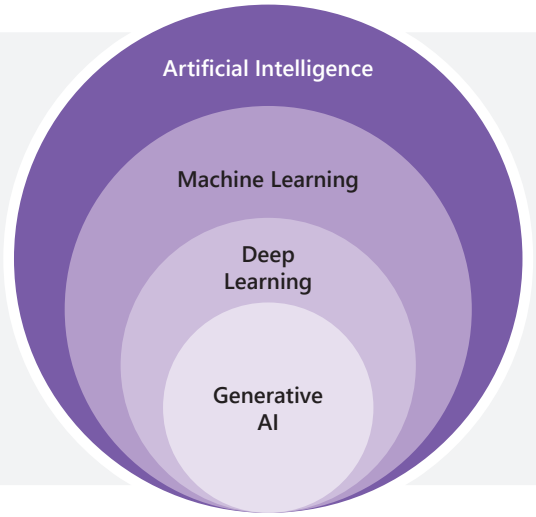
Maj, 2024

Rok Rogelj, Data&AI Specialist

Agenda

Where we are Today	Introduction to generative AI, the Microsoft Copilot ecosystem, Azure AI portfolio, our commitment to Responsible AI and Protecting your Data
Models & Capabilities	What is behind the scenes? Overview of Azure OpenAI Service and its cutting-edge models, features and solutions with other Azure AI Services
Use Cases	Top use cases for generative AI including industry highlights and highly requested customer scenarios
Sample Architecture	Example diagrams showcasing select use cases and scenarios featuring Azure OpenAI Service integrated with other Azure AI Services
Customer Inspiration	Publicly available Azure OpenAI Service customer stories featuring a wide range of use cases and company types

The journey continues with generative AI



1956 — Artificial Intelligence
The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence.

1997 — Machine Learning
Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions.

2012 — Deep Learning
A machine learning technique in which layers of neural networks are used to process data and make **decisions.**

2021 — Generative AI
Create new written, visual, and auditory content given prompts or existing data.

Microsoft and OpenAI partnership

OpenAI

Ensure that artificial general intelligence (AGI) benefits humanity

Microsoft

Empower every person and organization on the planet to achieve more

Azure OpenAI Service

GPT-4, 4-Turbo and 3.5-Turbo	GPT-4 Vision	Babbage, Davinci, GPT 3.5 Turbo	DALL·E 3	Whisper
Language	Multi-Modal	Fine Tuning	Images	Transcription & Translation

On Your Data | **Azure AI Studio** | **Assistants**

Microsoft Q4 2023 Earnings Call – Customer Focus

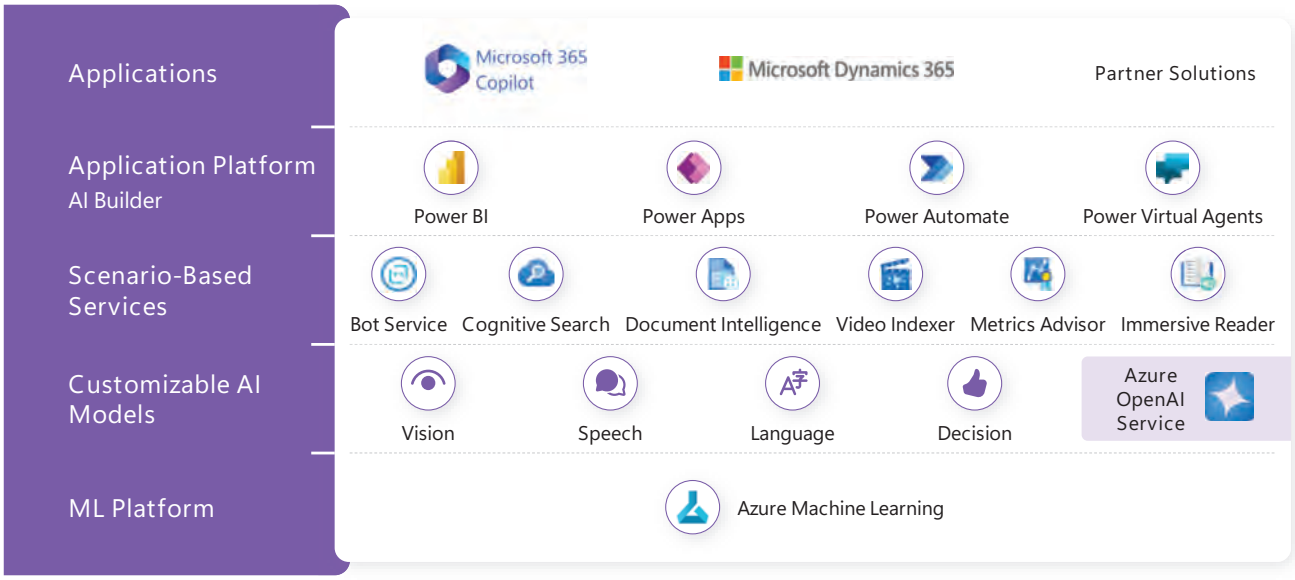
- Azure again took share this quarter, with our AI advantage.
- Azure offers the top performance for AI training and inference and the most diverse selection of AI accelerators, including the latest from AMD and NVIDIA, as well as our own first party silicon, Azure Maia.
- And, with Azure AI, we provide access to the best selection of foundation and open-source models, including both LLMs and SLMs, all integrated deeply with infrastructure, data, and tools on Azure.
- We now have 53,000 Azure AI customers. Over one-third are new to Azure over the past 12 months.
- And we have great momentum with our Azure OpenAI Service. This quarter, we added support for OpenAI's latest models, including GPT-4 Turbo, GPT-4 with Vision, Dall-E 3, as well as fine-tuning.
- We are seeing increased usage from AI-first startups like Moveworks, Perplexity, SymphonyAI, as well as some of the world's largest companies. Over half of the Fortune 500 use Azure OpenAI today, including Ally Financial, Coca-Cola, and Rockwell Automation.
- For example, at CES this month Walmart shared how it's using Azure OpenAI Service, along with its own proprietary data and models, to streamline how more than 50,000 associates work, and transform how its millions of customers shop.

-Satya Nadella, Chairman & CEO

Earnings Report

We are inspired by our customers

Microsoft is powered by Azure AI



Research Breakthroughs

2016	Object recognition *Human parity*
2017	Speech recognition *Human parity*
2018	Machine reading comprehension *Human parity*
2018	Machine translation *Human parity*
2019	Conversational QnA *Human parity*
2020	Image captioning *Human parity*
2021	Natural Language Understanding *Human parity*
2021	Commonsense Question Answering *Human parity*
2022	ChatGPT
2023	GPT-4

High Level Performance

GPT-4 test scores	Result Estimated percentile
AP biology	5 85th—100th
Uniform bar exam	298/400 ~90th
LSAT	163 ~88th
SAT reading & writing	710/800 ~93rd
SAT math	700/800 ~89th

Source: OpenAI. (2023). GPT-4: Scaling up deep learning. Retrieved from https://openai.com/research/gpt-4.

On Your Data – Use Cases

Use Cases	Details
Automated Customer Assistance	Provide quick responses to frequently asked questions and guide users through common troubleshooting steps based on the customers' data.
Real-time Document Search	Provide real-time support to product specifications and guide users through common troubleshooting steps based on product manuals
Citizen Service	Provide quick responses to frequently asked public service questions and guide users through specific steps based on citizen service support
Learning Assistant	Offer explanations and examples to help users understand academic concepts or learn new skills based on specific curricula
Legal Review	Quick access to legal insights from existing and upcoming legislation to properly advise clients
Marketing Insights	Tap into internal and external resources to respond to internal and external marketing inquiries
Software Development	Generate sample code based on the customer's needs
HR Support	Provide quick responses to frequently asked HR questions based on the customers' HR policy
Industry/Competitive Insights	Tap into publicly available resources to gain insights on the industry and competitors
Health Advice	Provide general information on symptoms, first aid, or healthy living.
Predictive Maintenance	Provide predictive maintenance and customer support based on customer's historic data

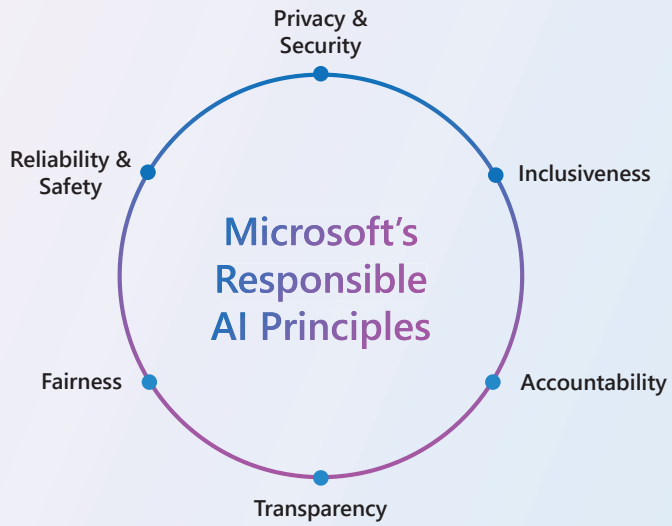
You can have confidence when using Azure OpenAI Service

When you use Azure OpenAI Service, your prompts (inputs) and completions (outputs), your embeddings, and your training data

- Are NOT available to other customers.
- ARE NOT available to OpenAI.
- Are NOT used to improve OpenAI models.
- Are NOT used to improve any Microsoft or 3rd party products or services.
- Are NOT used for automatically improving Azure OpenAI models for your use in your resource (The models are stateless, unless you explicitly fine-tune models with your training data).
- Your fine-tuned Azure OpenAI models are available exclusively for your use.

The Azure OpenAI Service is fully controlled by Microsoft; Microsoft hosts the OpenAI models in Microsoft's Azure environment and the Service does NOT interact with any services operated by OpenAI (e.g., ChatGPT, or the OpenAI API).

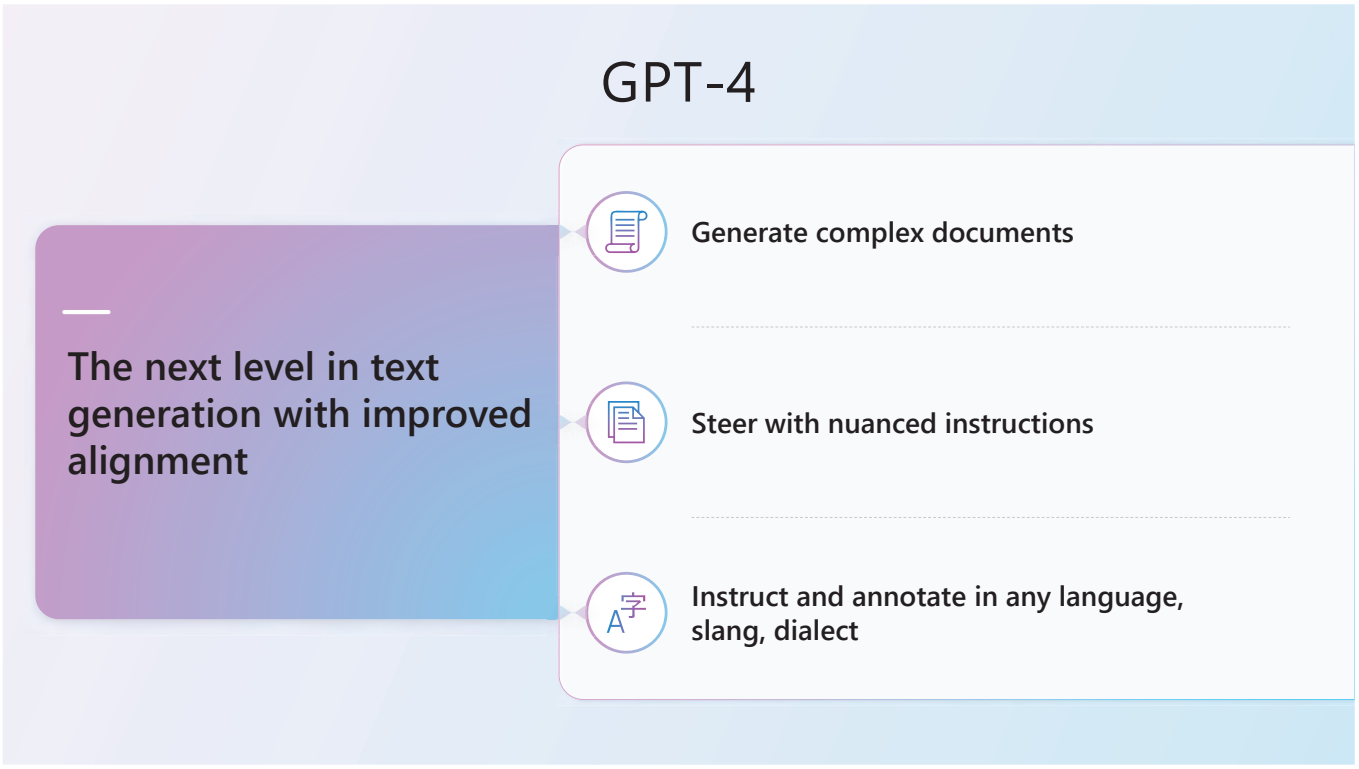
Microsoft's Responsible AI Principles



Building blocks to enact principles

- Tools and processes
- Training and practices
- Rules
- Governance

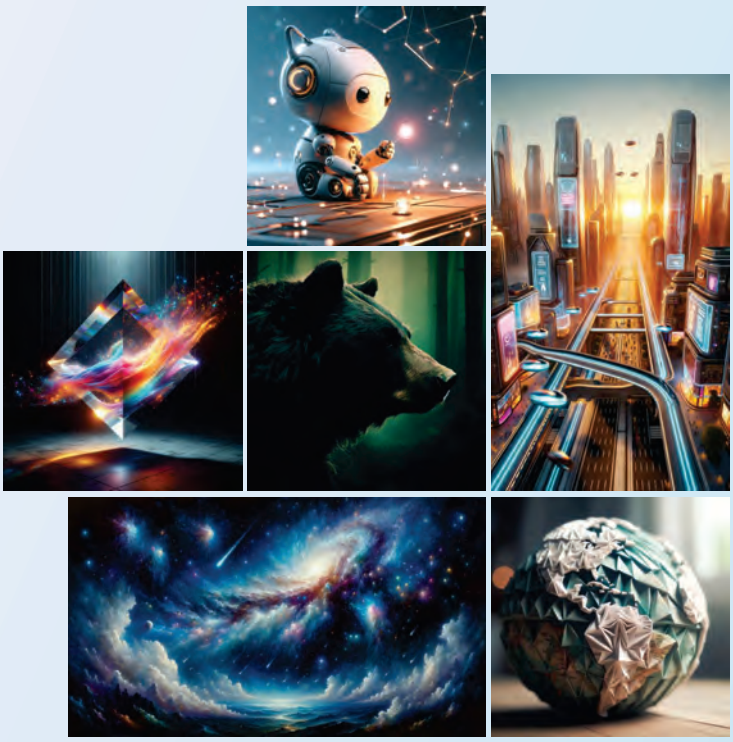

Models and Capabilities


GPT-4
The next level in text generation with improved alignment
Generate complex documents
Steer with nuanced instructions
Instruct and annotate in any language, slang, dialect

DALL·E 3

In Preview: Azure OpenAI Service

DALL-E 3 is an image generation model that allows you to generate images from text prompts



Use Cases for DALL·E 3

LOGO & BRANDING: QUICK CONCEPT GENERATION.

CREATIVE INSPIRATION: OVERCOME DESIGN BLOCKS.

CONTENT ILLUSTRATIONS: UNIQUE IMAGES FOR BLOGS/ARTICLES.

AD CAMPAIGNS: VISUALIZE MARKETING CONCEPTS.

PRODUCT VISUALIZATION: GAUGE INTEREST & FEEDBACK.

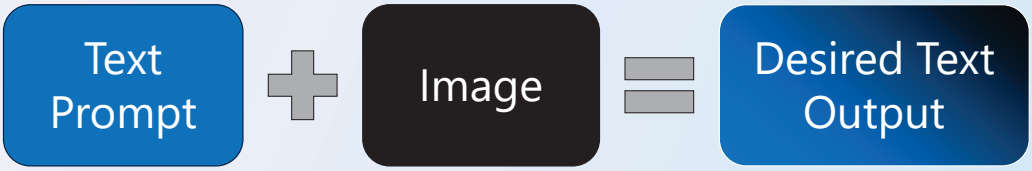
EDUCATION: CUSTOM IMAGERY FOR COURSES.

FASHION DESIGN: VISUALIZE CLOTHING PATTERNS.

GAMING: CHARACTER & ENVIRONMENT CONCEPTS.

What GPT-4V Offers:

GPT-4 with Vision (GPT-4V) is a multimodal model developed by OpenAI that accepts both image and text inputs and generates text outputs.



Note: GPT-4V doesn't generate image outputs

Whisper

The next level in transcription and translation

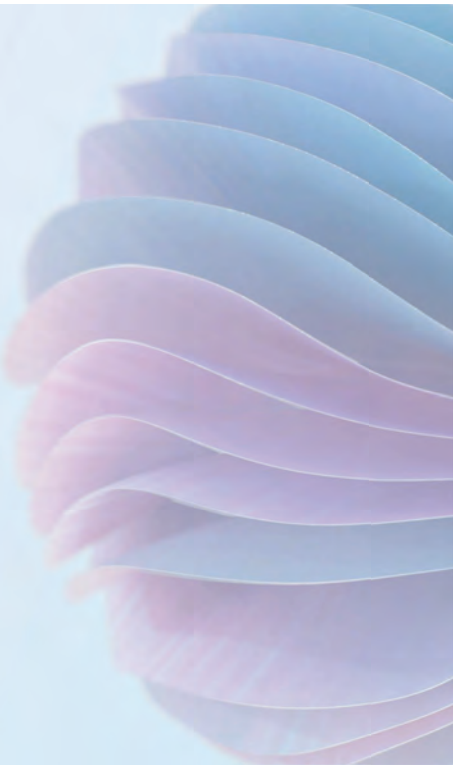
- Transcribe
- Translate

Generative AI + Document Intelligence

The next level of document intelligence integrated with Azure OpenAI Service

- Extract Structured and Unstructured Data
- Summarize Content
- Generate Summaries

Use Cases & Customer Requests



Introduction to top use cases

Business Problem	Productivity is lagging	Need for process Automation	Degraded Customer Experience	Creating Content is Time Consuming
Business Needs	Increase Productivity	Automate Processes	Improve Customer Experience	Build Creative Content
Solutions	• Conversational Search/Knowledge Insights • Code Generation and Documentation • Trend Forecasting • Report Summarization & Generation	• Document Processing • Workflow Management • Fraud Detection • Supply Chain Optimization	• Intelligent Contact Center • Agent/Employee Assistance • Virtual Assistance • Call Analytics • Call Summarization	• Marketing/Sales Content Generation • Personalized Content Generation • Product Design & Development • Digital Art
What can Generative AI Do?	Generate New Revenue Streams Deliver Differentiated Customer Experiences Modernize Internal Processes			

Top capabilities and scenarios

1.0

Content generation | Summarization | Code generation | Semantic search

2.0

Examples of advanced use cases

Telecommunications

Media Workflows, Cross Content Linking, Content Creation for Media, Speech Analytics, Analytics for B2C Contact Center, Cognitive Contact Center, Skilling Automation, Realtime Speech Transcriptions & Summarization

Manufacturing & Industrials

ChatGPT Enabled Technical Support, Customer Sentiment Analysis, Customer Service Knowledge Mining, Digital Proposal Assistant, Customer Journey Analytics, Consumer Insights Advanced Analytics, Records Summarization, Anomaly Detection, Virtual Agents with Copilot

Automotive, Mobility & Transportation

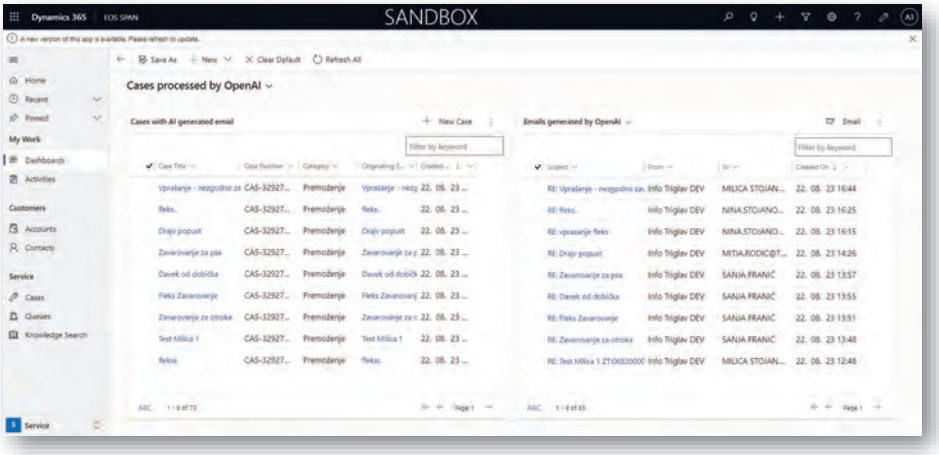
Marketing Content Generation, Contextual Contact Center, Customer Feedback Loop, Smart Incident Manager, Customer Comms, Text Summarization & Analytics

Customers Bringing it to Life

STRABAG Coca-Cola KPMG Icertis ODP CARMAX zammo.ai

Dashboard in D365 application

Starting point for Triglav Customer Service agents



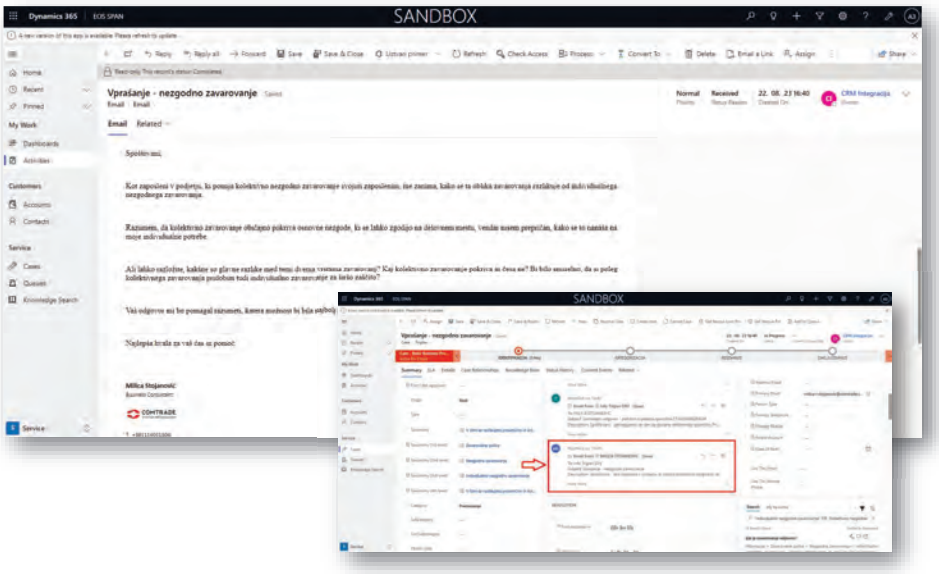
The dashboard consists of the following two sections:

- Cases with AI-generated emails: Cases processed through the OpenAI engine with generated email response
- Emails generated by OpenAI – automatically created email replies with OpenAI suggested content ready for agents to review and send

25

Cases processed by OpenAI

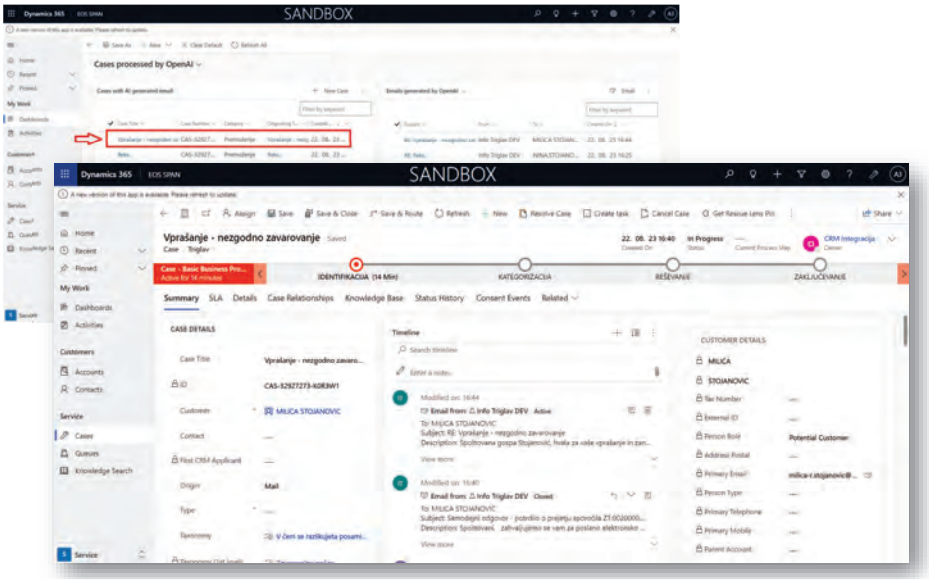
First email from customer – trigger for initiating the process



- The process is triggered when a new case is created as a result of an incoming email from a customer.
- The case is created based on the existing email-to-case process in the D365 application.

26

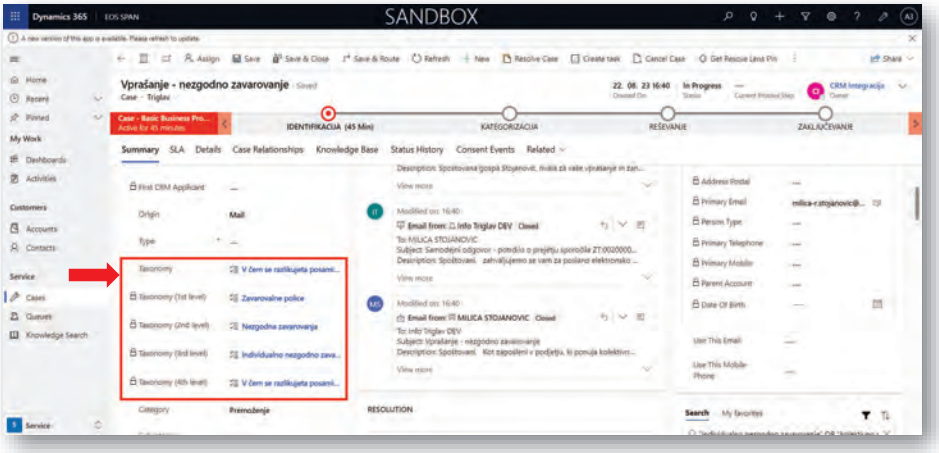
Cases processed by OpenAI
Review of the Taxonomy and Email response



- Agent can pick the case from a list in the dashboard and review both the content of the generated email responses and the Taxonomy populated based on OpenAI suggestions.

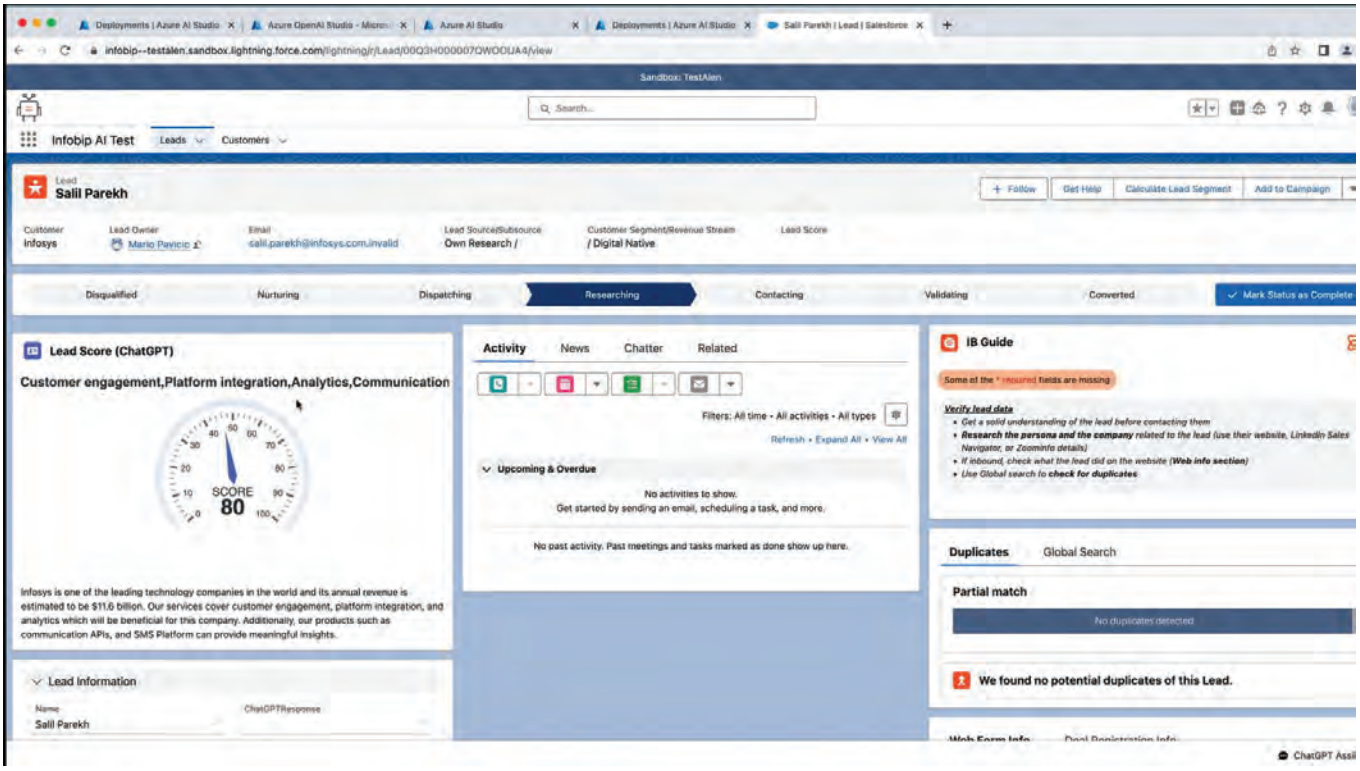
27

Cases processed by OpenAI
Case categorization – Review of the Taxonomy population



- After the population of suggested Taxonomy value, other Taxonomy fields from level 1 to level 4 are automatically populated (due to an existing process) based on the Taxonomy entered in the first field.

28



Industry Trends 1/3

Industry	Azure OpenAI Top Patterns	Public References
Automotive, Mobility & Transportation	Contact Center – Extract rich insights from call transcripts. E2E Call Center analysis: Classification, sentiment, summarization and email generation. Automatically generate responses to customer inquiries. Summary of customer support conversation logs. Improved customer satisfaction. Improve time to resolution. Smart Virtual Agents/Customer & Employee Communication – Intelligent automated responses to customers and employees through the value chain. Generate human-like responses for live chat/voice applications. Automated Documents Processing – Automate Invoice processing for large volumes using Microsoft Cognitive Services, and Azure OpenAI. Incident Reporting & Prediction – Text summarization and pattern recognition to automatically generate reports on equipment and process failures, etc. Generate automated responses and notifications. Automated Content Generation – Generative AI can be used to automatically generate content on a variety of topics, saving time and money for content creators. Generate more relevant content with summarization. Improve SEO (search engine optimization).	CARMAX BMW Group
Financial Services	Contact Center – Extract rich insights from call transcripts. E2E Call Center analysis: Classification, sentiment, summarization and email generation. Automatically generate responses to customer inquiries. Summary of customer support conversation logs. Improved customer satisfaction. Improve time to resolution. Smart Virtual Agents/Customer & Employee Communication – Intelligent automated responses to customers and employees through the value chain. Generate human-like responses for live chat/voice applications. Automated Documents Processing – Ability to ingest forms and other unstructured data as part of an automated workflow, supporting Claims and other areas of the business using Microsoft Cognitive Services, and Azure OpenAI. Digitize documents, extract key information through Semantic or Cognitive Search and summarize (financial document summarization, RFI/RFPs processing, etc.).	AMERICAN FIDELITY nationale nederlanden

Industry Trends 2/3

Industry	Azure OpenAI Top Patterns	Public References
Manufacturing & Industrials	Contact Center – Extract rich insights from call transcripts. E2E Call Center analysis: Classification, sentiment, summarization and email generation. Automatically generate responses to customer inquiries. Summary of customer support conversation logs. Improved customer satisfaction. Improve time to resolution. ChatGPT Enabled Technical Support – Contextual human like responses to customers' technical questions and inquires with better recommendations based on internal and external content databases. Customer Service Knowledge Mining – Ingest unstructured and structured data and draw better insights around key issues and patterns. Anomaly Detection – Provide NLP queries on the vision anomalies and causes. Use Synthetic Imaging to augment baseline ML training image volumes to improve ML accuracy. Improve quality insurance pass rates and reduce operational costs by catching anomalies faster, reducing amount of inventory that needs to be scrapped. Virtual Agents with Copilot – Intelligent automated responses to customers and employees through the value chain. Generate human-like responses for live chat/voice applications.	STRABAG
Telecommunications	Contact Center – Extract rich insights from call transcripts. E2E Call Center analysis: Classification, sentiment, summarization and email generation. Automatically generate responses to customer inquiries. Summary of customer support conversation logs. Improved customer satisfaction. Improve time to resolution. Cross Content Linking – Ingest content for different sources in various formats to improve knowledge mining and draw better insights. Content Creation for Marketing/Media – Generate marketing emails automatically based on personalized customer 360 information. Realtime Speech Transcriptions & Summarization – Perform entity extraction and generate summaries of audio transcripts.	AT&T

Industry Trends 3/3

Healthcare & Life Sciences

Streamline administrative tasks
Azure OpenAI Service can be used to automate administrative tasks such as scheduling, patient registration, and insurance verification, freeing up healthcare staff to focus on patient care.

Improve supply chain management
Healthcare providers can use Azure OpenAI Service to help analyze data and help optimize inventory levels of medical supplies, reducing waste and ensuring that essential supplies are always available when needed.

Predictive maintenance of medical equipment
Azure OpenAI Service can help monitor the performance of medical equipment enabling healthcare providers to predict when maintenance is needed and prevent unexpected downtime.

Automated medical coding
Azure OpenAI Service can be used to automatically analyze medical records and recommend medical codes for billing and reimbursement purposes, reducing errors and improving efficiency.

Enable data insights
By analyzing large amounts of data, Azure OpenAI Service can help healthcare providers identify patterns that enables more informed decisions about their practice.

Fraud detection
Healthcare providers can use Azure OpenAI Service to detect fraudulent insurance claims. Providers can use machine learning algorithms to analyze large volumes of data and identify patterns that indicate potential fraud. By detecting fraudulent claims early, healthcare providers can prevent losses and protect patients from unnecessary treatments.

Epic brings GPT-4 to Electronic Health Records

- *We are working closely with Epic to integrate & modernize Electronic health records with GPT-4.*
- *Two-way impact: 1) less time at keyboard entering information and 2) help investigate data in a more conversational way.*

Odprite “črno skrinjico” z upravljanjem umetne inteligence

Open the “black box” with AI governance

Katarina Gašperlin Stepančič

IBM

POVZETEK

Zaradi vedno večje uporabe umetne inteligence in prihajajočih predpisov o umetni inteligenci se je pojavila potreba po usmerjanju, upravljanju in spremljanju uporabe umetne inteligence v organizacijah. Pomembno je omeniti, da tako generativna umetna inteligenca kot pa tudi običajna (konvencionalna) umetna inteligenca zahtevata upravljanje. In navkljub temu, da generativna umetna inteligenca v zadnjem času dobiva več pozornosti, se obe vrsti umetne inteligence nenehno uporabljata v organizacijah za podporo poslovnim procesom. Upravljanje umetne inteligence je sistem pravil, praks, procesov in orodij, ki organizaciji pomagajo pri uporabi umetne inteligence v skladu z njenimi vrednotami in strategijami, obravnavajo zahteve glede skladnosti in spodbujajo zanesljivo delovanje. Učinkovit process upravljanja umetne inteligence mora vključevati tako netehnične kot tudi tehnične deležnike v organizaciji. Uspešen pristop k upravljanju AI lahko razdelimo na tri stebre. Prvi steber je upravljanje življenjskega cikla, prek katerega dosledno upravljamo, katalogiziramo in spremljamo vse modele umetne inteligence. Drugi steber je upravljanje s tveganji, s katerim upravljamo tveganja, težave in ukrepe za ublažitev v skladu s poslovnimi standardi. Organizacije morajo obravnavati več elementov tveganj umetne inteligence, ki jih je mogoče povzeti kot regulativno tveganje, tveganje izgube ugleda in organizacijsko tveganje. Tretji steber je skladnost s predpisi, ki zagotavlja, da organizacije upoštevajo zunanje predpise o umetni inteligenci za revizijo in skladnost.

SUMMARY

Due to more and more use of artificial intelligence as well as the upcoming AI regulations the need to direct, manage and monitor the artificial intelligence activities in the organizations has emerged. It is important to mention that both, generative AI as well as conventional AI require governance as even though generative AI is lately getting more attention, both types of AI are continuously used in the organizations to support business processes. AI Governance is a system of rules, practices, processes and tools that help an organization use AI in alignment with its values and strategies, address compliance requirements and drive trustworthy performance. It brings together both, non-technical as well as technical stakeholders. Successful AI governance approach can be divided into three pillars. The first pillar is lifecycle governance through which we consistently govern, catalog, and monitor all AI models. The second pillar is risk management through which we manage model risks, issues, and mitigation actions to business standards. There are multiple elements of AI risks that organizations need to address, and these can be summarized as regulatory risk, reputational risk, and organizational risk. The third pillar is regulatory compliance that ensures organizations adhere to external AI regulations for audit & compliance.

O AVTORJU

Katarina Gašperlin Stepančič je specialistka za umetno inteligenco in analitično arhitekturo v podjetju IBM. V regiji EMEA s strankami sodeluje pri operacionalizaciji tako generativne kot tudi konvencionalne umetne inteligence v organizacije ter zasnovi analitične arhitekture, ki je zato potrebna. Katarina je univerzitetno diplomo pridobila na Fakulteti za elektrotehniko Univerze v Ljubljani. Leta 2019 ji je IBM podelil nagrado IBM Cloud Europe Young Talent, prav tako pa je prejemnica več nagrad za izjemne tehnične dosežke s končno vrednostjo za stranke (IBM Client Value Outstanding Technical Achievement Awards).

ABOUT THE AUTHOR

Katarina Gašperlin Stepančič is a specialist in artificial intelligence and analytical architecture at IBM. In the EMEA region, she cooperates with clients to operationalize both, generative and conventional artificial intelligence into the organization processes and design the underlying architecture. Katarina obtained a university education at the Faculty of Electrical Engineering, University of Ljubljana. In 2019 IBM awarded her with the IBM Cloud Europe Young Talent award. She also received numerous awards for outstanding technical achievements (IBM Client Value Outstanding Technical Achievement Awards).

‘Break open the black box’ with AI governance

Balancing the rewards and risks of AI in a world of ethics and regulation – and doing that at scale

Katarina Gašperlin Stepančič
AI Engineer, EMEA Client Engineering watsonx core team
katarina.gasperlin@si.ibm.com

The impact of generative AI
The opportunity

The speed, scope, and scale of generative AI impact is unprecedented

Sources: Statista; Reuters; Goldman Sachs; IBM Institute for Business Value; Gartner. Scale Zeitgeist: AI Readiness Report, a survey of more than 1,600 executives and ML practitioners

watsonx.governance

Massive early adoption

80%

of enterprises are working with or planning to leverage foundation models and adopt generative AI

Broad-reaching and deep impact

+7%

Generative AI could raise global GDP by 7% within 10 years

Critical focus of AI activity and investment

30%

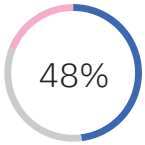
Generative AI expected to represent 30% of overall market by 2025

Enterprise considerations

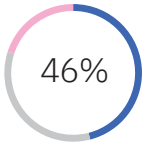
Business leaders face challenges in scaling AI across the enterprise with trust

80% of business leaders see at least one of these ethical issues as a major concern

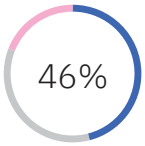
watsonx.governance



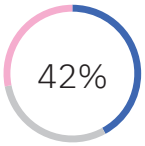
Explainability
Believe decisions made by Generative AI are not sufficiently explainable.



Ethics
Concerned about the safety and ethical aspects of Generative AI.



Bias
Believe that Generative AI will propagate established biases.



Trust
Believe Generative AI cannot be trusted.

Agree Neutral Disagree

Businesses everywhere are navigating an increasingly complex global regulatory landscape with respect to AI

> Less than 60% of executives believe their organization is prepared for AI regulation[1]

> 72% of executives are choosing to forgo generative AI over concerns about AI ethics and safety[1]

[1]IBM Institute for Business Value, The CEO's Guide to Generative AI: Responsible AI & ethics, 2023

Office of Privacy and Responsible Technology
© 2024 IBM Corporation

Privacy, data, and AI regulations and enforcement activities are increasing

Privacy laws and regulations

15 US states have passed comprehensive privacy bills, three of them will become effective in 2024.

India's Digital Personal Data Protection Act will become effective in 2024, including fines for non-compliance up to $30M.

US Executive Order limits bulk transfer of specified types of sensitive data to certain countries.

AI laws and regulations

The EU AI Act was passed in March 2024 and includes fines of up to **7% of a company's annual revenues** for noncompliance.

Canada, Brazil and Korea progress toward adoption of AI laws.

A number of US states have already adopted AI laws, and AI laws are progressing in several others.

US Executive Order directs new standards for AI safety and security.

Office of Privacy and Responsible Technology
© 2024 IBM Corporation

5

Rapid evolution in AI drives the need for *comprehensive* and *open* governance of that AI

Generative AI

65/35
Split in software spend on non-generative AI (65%) and generative AI (35%). (1)

50
Amplified and *new* risk with generative AI. (2)

Govern both forms of AI in a consistent manner.

Extend your governance to account for the new aspects of generative AI.

Innovation in models

~5000
New foundation models posted on Hugging Face every week.

Continuous innovation in open source and commercial offerings gives you an increasing range of options and trade-offs.

Govern the onboarding of new models.

Govern the trade-offs in use cases.

Consumption models

60/40
Split in software spend on AI platforms (60%) and AI embedded in enterprise applications (40%). (1)

70%
Of independent software vendors will have embedded generative AI capabilities. (1)

Govern all AI, regardless of how and by whom it's created or consumed.

Legislation

200+
Pages in the EU AI Act.

10
Standardization requests in the EU AI Act.

Translate your responsibilities into controls and workflows.

Risk-assess your use cases.

Adopt technical standards.

© IBM 2024

(1) Gartner – Forecast Analysis: AI Software, 2023-2027, Worldwide
(2) IBM AI Risk Atlas

6

AI needs governance

↓

The process of directing, monitoring and managing the AI activities of an organization

watsonx.governance

AI governance

A system of rules, practices, processes and tools that help an organization

→ use AI in alignment with its values and strategies

→ address compliance requirements

→ drive trustworthy performance

Generative AI and applications make AI governance a necessity, as using pretrained AI models billions of times sharpens risk concerns.

Gartner – Hype Cycle for Data and Analytics Governance, 2023

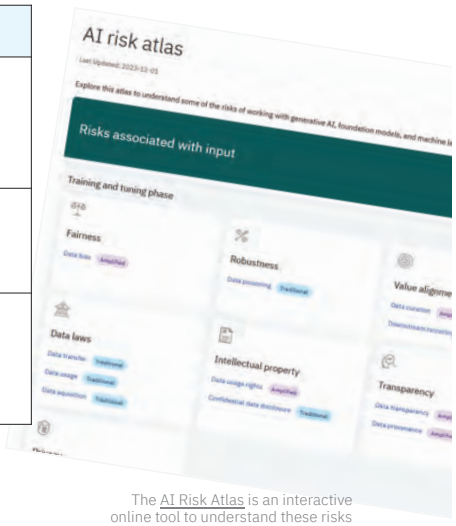
(c) 2022-2023, IBM Corporation

Data & Trust Alliance, IBM – The urgency of AI governance, 2023
https://dataandtrustalliance.org/news/perspectives-from-the-front

8

Traditional, amplified and new risks

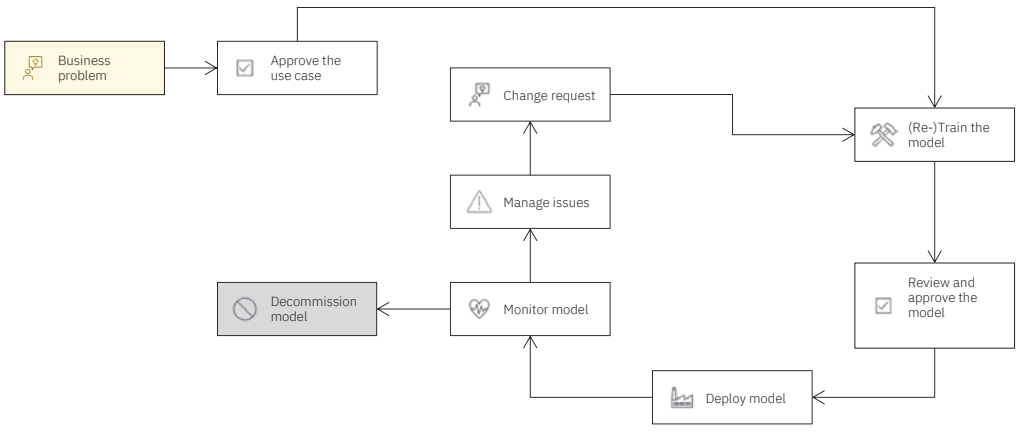
	Traditional risks	Amplified risks	New risks
Risks associated with input	Legal restrictions on moving or using (training) data.	Copyright and other IP issues with the content.	Vulnerabilities to new types of adversarial attacks such as prompt injecting.
Risks associated with output	Performance disparity across individuals or groups.	Challenges in explaining why output was generated.	Hallucination—false content generation.
Challenges	Documenting data and model details, purpose, potential uses and harms.	Increased carbon emission due to high energy requirement to train and operate.	Homogenizing of culture and thoughts.



The AI Risk Atlas is an interactive online tool to understand these risks

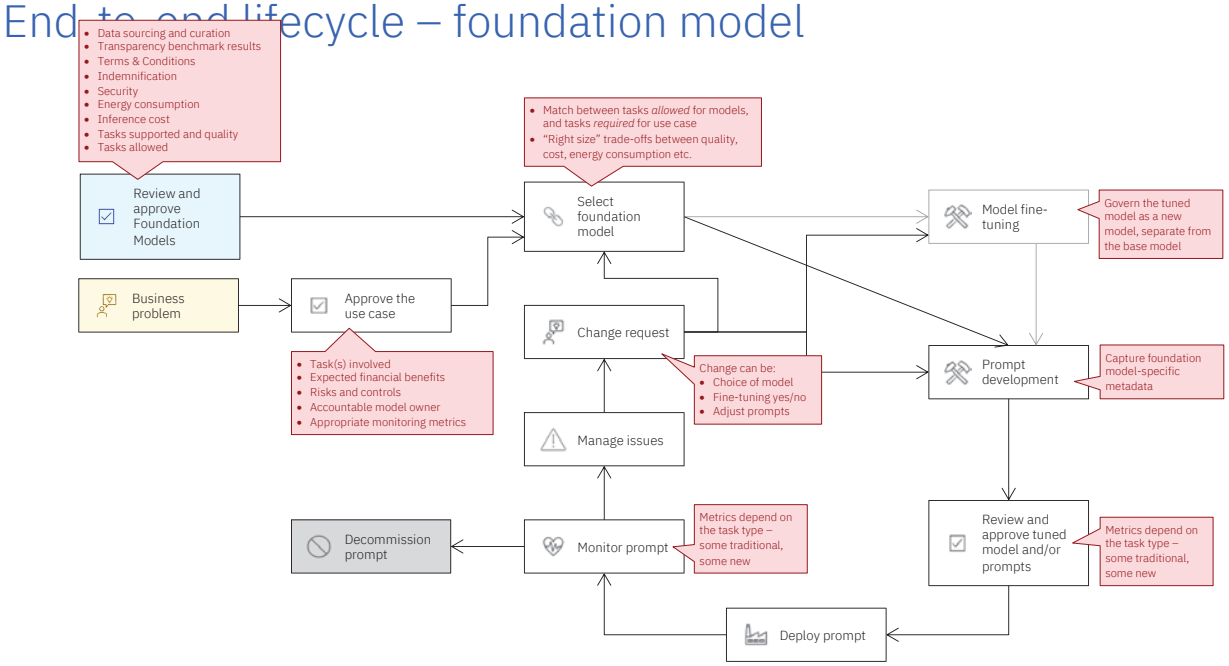
(c) 2022-2023, IBM Corporation

End-to-end lifecycle – conventional ML model



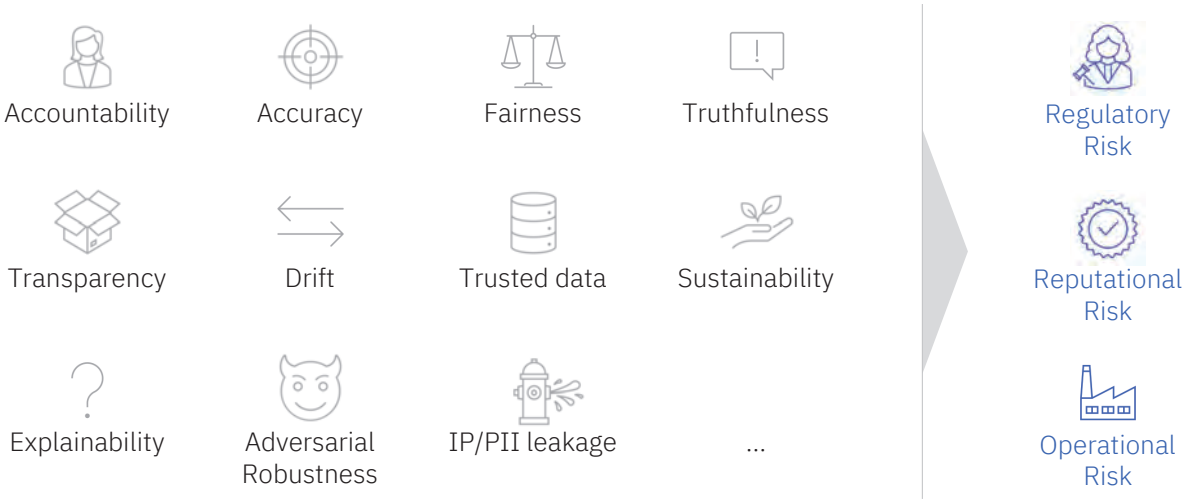
(c) 2022-2023, IBM Corporation

End-to-end lifecycle – foundation model



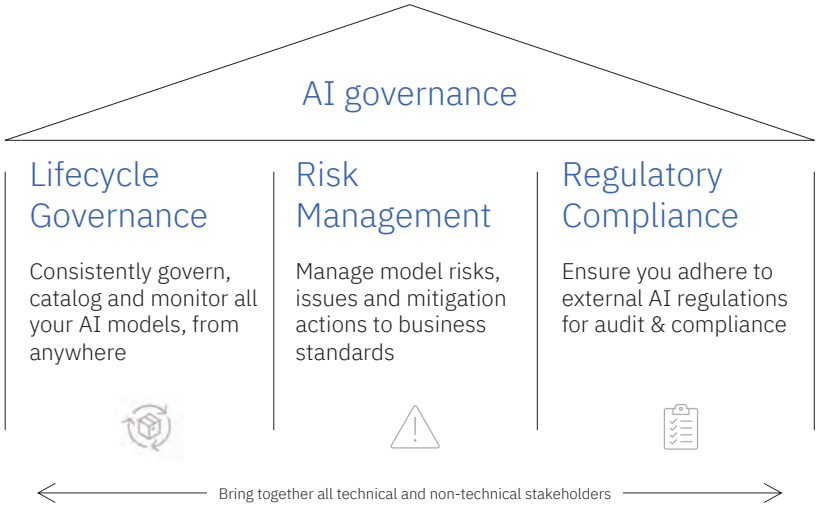
(c) 2022-2023, IBM Corporation

Elements of AI risk



(c) 2022-2024, IBM Corporation

Three pillars of AI governance



(c) 2022-2023, IBM Corporation

Governing foundation models *across* use cases

- Model approval (for pre-trained, 3rd party models) moves from “micro” to “macro” level
- Update enterprise risk framework for new and amplified risks and new metrics
- Update internal education for new and amplified risks
- Integrate model governance into vendor management

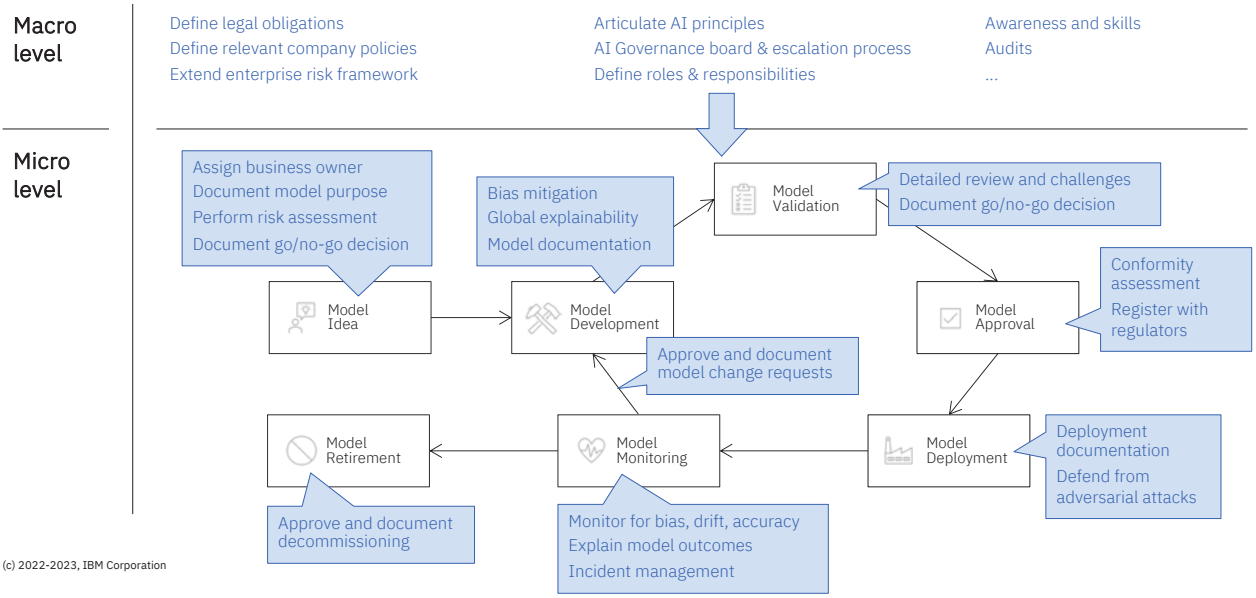
(c) 2022-2023, IBM Corporation

Which is easier said than done

Common challenges		Optimization approach
AI governance collaboration requires lots of **manual work**; amplified by changes in data and model versions.	→	**Automate** the governance activities as much as possible.
Companies have AI in **multiple tools, applications and platform**, developed inside and outside the organization	→	**Consolidate** as much as possible in one governance platform.
Governance is **not a one-size-fits-all** approach.	→	**Configure** to your specific situation.
Constrain technical teams in their choice of technology.	→	**Open** architecture to wrap around tooling of choice.

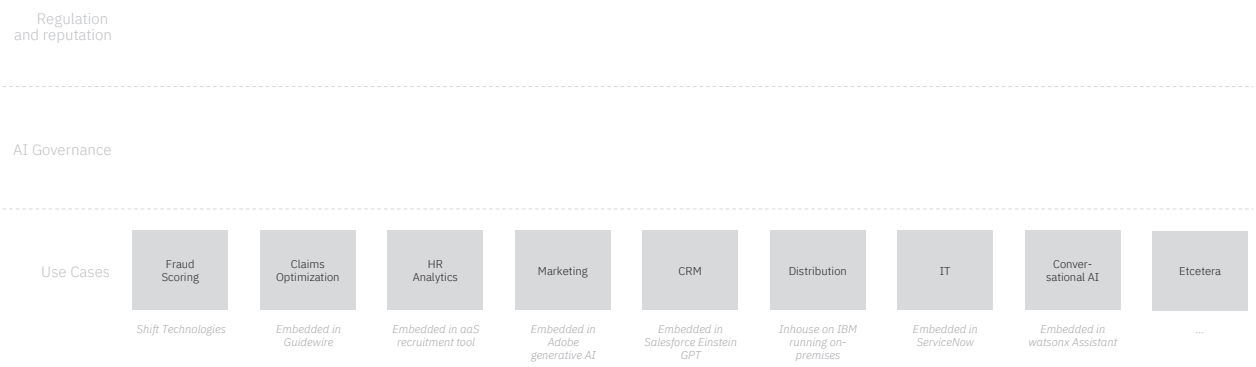
(c) 2022-2023, IBM Corporation

In practice, AI governance means...



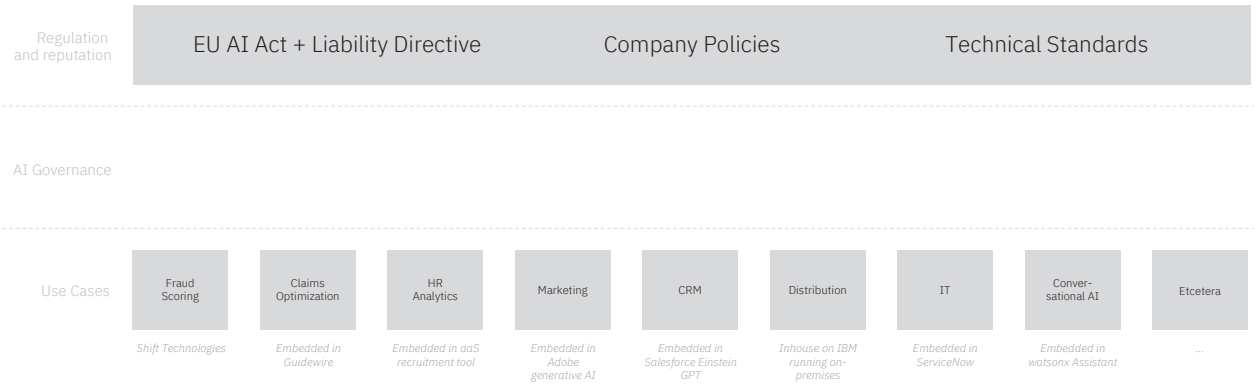
(c) 2022-2023, IBM Corporation

Consistent and efficient governance across all use cases and technologies – Insurance example



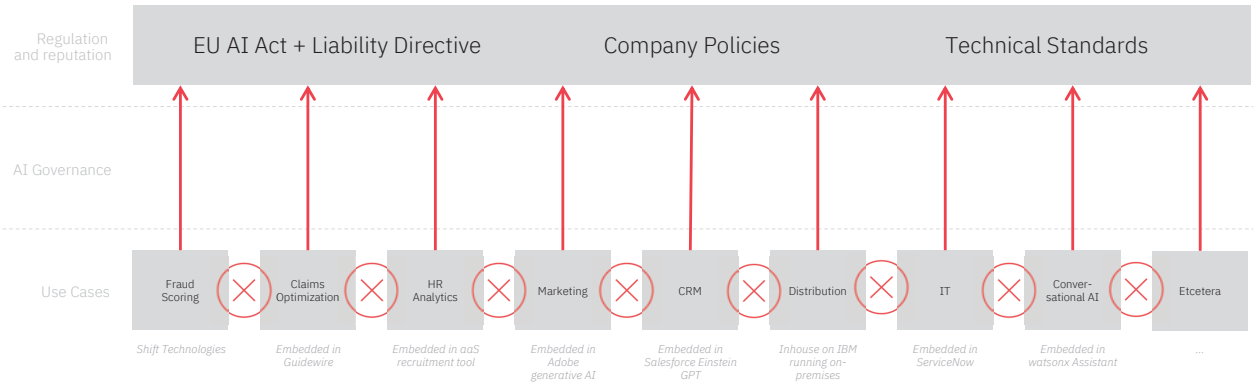
(c) 2022-2023, IBM Corporation

Consistent and efficient governance across all use cases and technologies – Insurance example



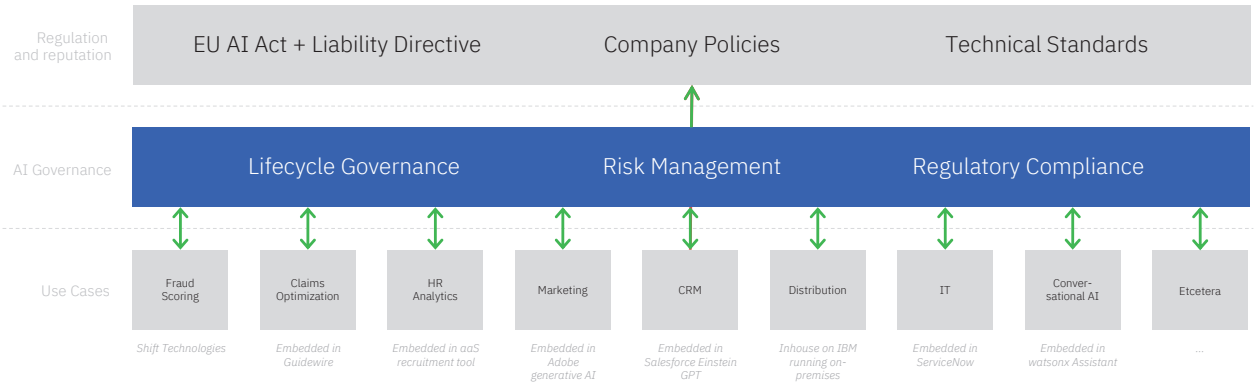
(c) 2022-2023, IBM Corporation

Consistent and efficient governance across all use cases and technologies – Insurance example



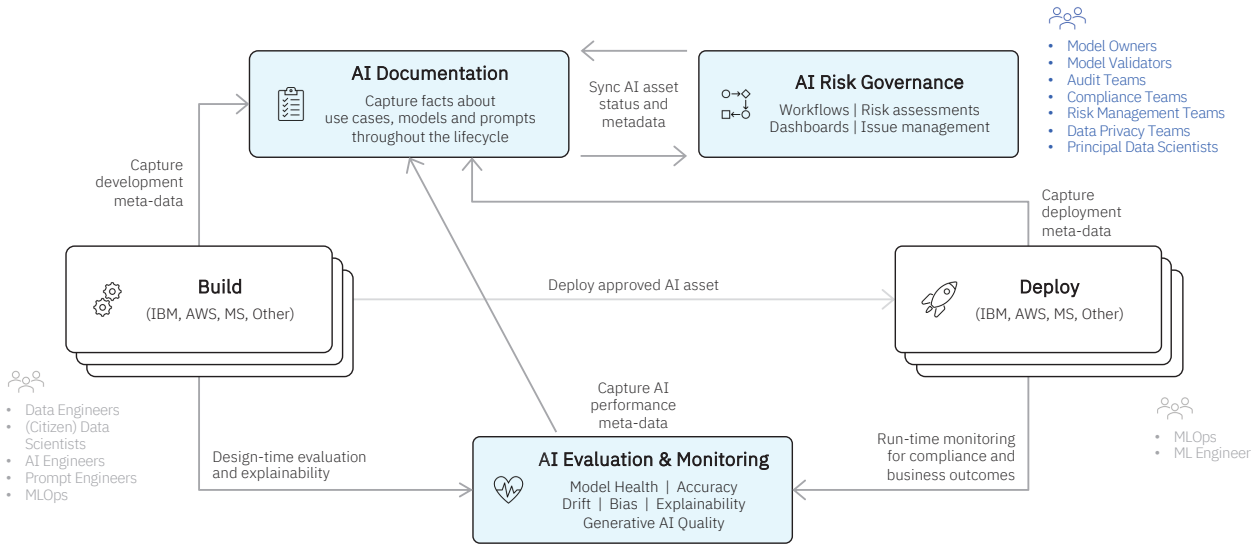
(c) 2022-2023, IBM Corporation

Consistent and efficient governance across all use cases and technologies – Insurance example



(c) 2022-2023, IBM Corporation

IBM watsonx.governance



(c) 2022-2023, IBM Corporation

IBM customers are creating more mature AI governance

Prepare for audit and regulatory compliance

- North American Bank, multiple data science stacks, 1000s of models.
- Manual audit process took months of work.
- Invested in IBM software for its completeness and its ability to work with existing technology.

Proactively mitigate bias in the hiring process

- Leaders at a North American retailer wanted their company to meet commitments as a fair employer.
- Invested in IBM software to monitor and actively seek out potential bias in their hiring systems.

US Open: Look at AI bias in a novel way

- AI-assisted curation of match highlights, available 2 minutes after the match ends.
- Excitement score is biased by player rank and the court where the match is played.
- Post-processing de-biasing applied to increase court fairness from 71% to 82% without impacting overall accuracy.

"More than 80% say that they'll commit 10% or more of their total AI budget to meeting regulatory requirements by 2024 and 45% are planning to spend at least 20%."

Accenture
From AI compliance to competitive advantage, 2022

watsonx.governance

Umetna inteligenca v radijskih dostopovnih omrežjih

Artificial Intelligence in Radio Access Networks

Ivan Lesić

Nokia

POVZETEK

V prispevku so predstavljeni glavni evolucijski koraki razvoja AI in njegov prispevek k razvoju omrežja RAN.

Ker povpraševanje po podatkih nenehno narašča, mora radijsko dostopovno omrežje raziskati in sprejeti nove storitve in funkcionalnosti, ki temeljijo na AI, da bi zagotovili optimalno izkušnjo za končnega uporabnika. Ogledali si bomo trenutne standardizacijske načrte, mejnike implementacije in kako bo RAN, podprt z umetno inteligenco, pomagal zagotoviti boljše storitve telekomunikacijski industriji in končnim uporabnikom.

SUMMARY

During the presentation, major evolutionary steps of AI development and contribution to RAN network evolution will be outlined.

As demand for data is constantly growing, radio access network has to explore and embrace new AI based services and functionalities in order to ensure optimal end user experience. We will take a look at current standardization plans, implementation milestones and how AI supported RAN will help deliver better service to telecom industry and end consumers.

O AVTORJU

Ivan Lesić je tehnični direktor v podjetju Nokia za srednjo in južno Evropo. Je višji telekomunikacijski strokovnjak, osredotočen na najnovejše tehnologije, razvoj omrežij in strateško načrtovanje ter implementacijo naprednih rešitev in storitev. Kot izvršni svetovalec je vsakodnevno vključen v strateške razprave, pri čemer mobilnim operaterjem in ponudnikom storitev pomaga definirati optimalne strategije razvoja omrežij ter pripraviti njihove izdelke in storitve na prihodnost.

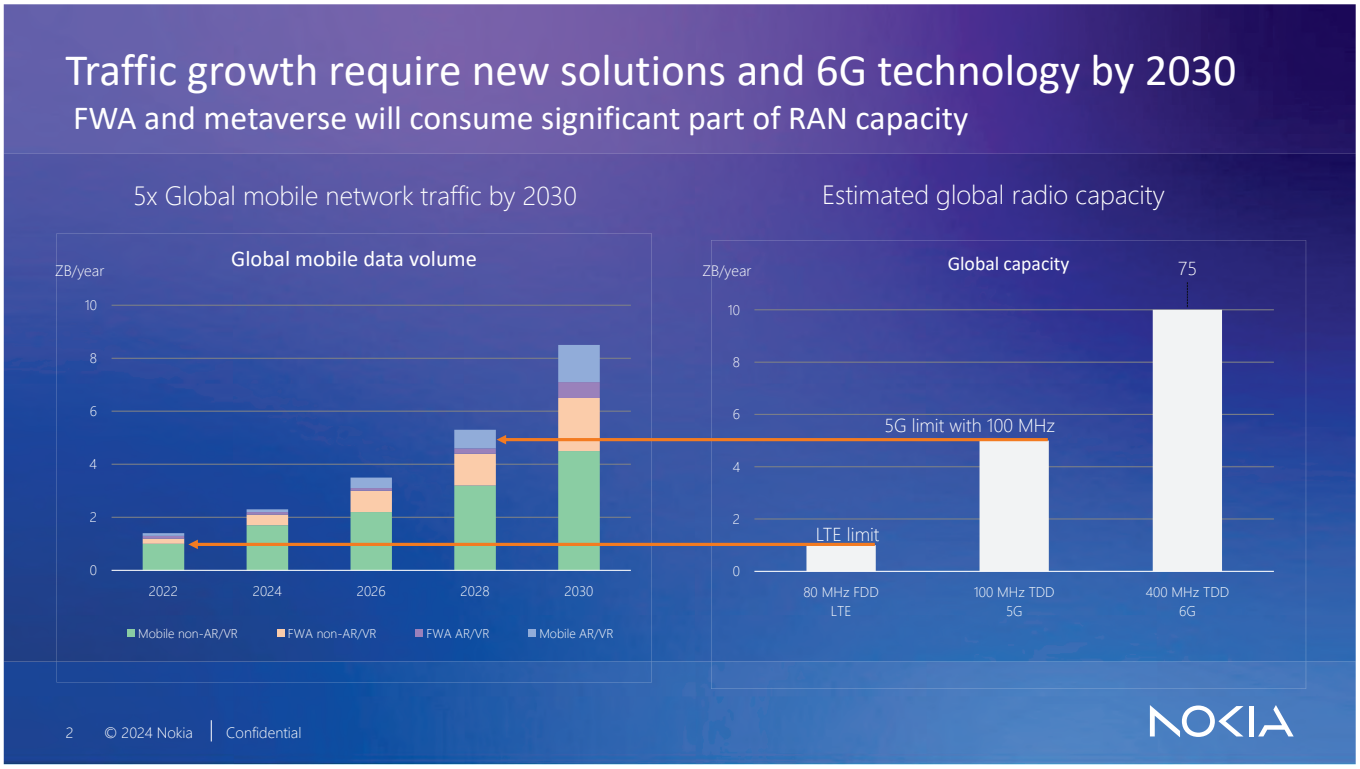
Močno ga zanimajo oblak, virtualizacija in napredna avtomatizacija omrežnih funkcionalnosti in storitev.

ABOUT THE AUTHOR

Ivan Lesić is a CTO at Nokia for Central & South Europe. He is senior telecom professional focused on latest technologies, network evolution and strategic planning and implementation of advanced solutions and services. Acting as executive consultant, he is on a daily basis engaged in strategic discussions, helping mobile operators and service providers to define optimal network development strategies, making their products and services future ready.

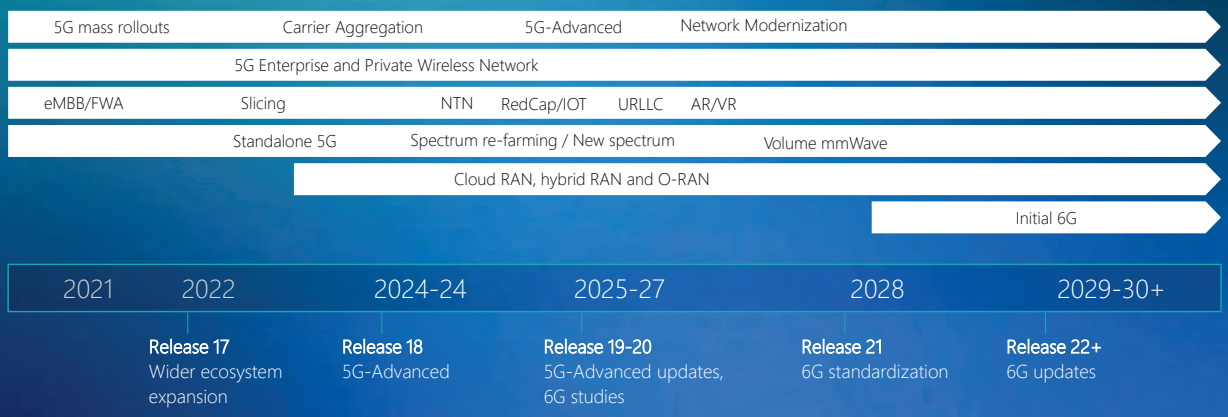
He is deeply interested in cloud, virtualization and advanced automation of network functionalities and services.


NOKIA
Vitel 2024
39’th telecommunication workshop
Artificial intelligence in radio access network
Ivan Lesić
CTO, Nokia Mobile Networks
May 2024
1 © 2024 Nokia


Traffic growth require new solutions and 6G technology by 2030
FWA and metaverse will consume significant part of RAN capacity
5x Global mobile network traffic by 2030
Global mobile data volume
ZB/year
10
8
6
4
2
0
2022
2024
2026
2028
2030
Mobile non-AR/VR
FWA non-AR/VR
FWA AR/VR
Mobile AR/VR
Estimated global radio capacity
Global capacity
ZB/year
75
5G limit with 100 MHz
LTE limit
80 MHz FDD LTE
100 MHz TDD 5G
400 MHz TDD 6G
2 © 2024 Nokia | Confidential
NOKIA

Evolving technology landscape

Private Wireless, 5G-Advanced and Cloud RAN will create new opportunities



Release 17 Wider ecosystem expansion

Release 18 5G-Advanced

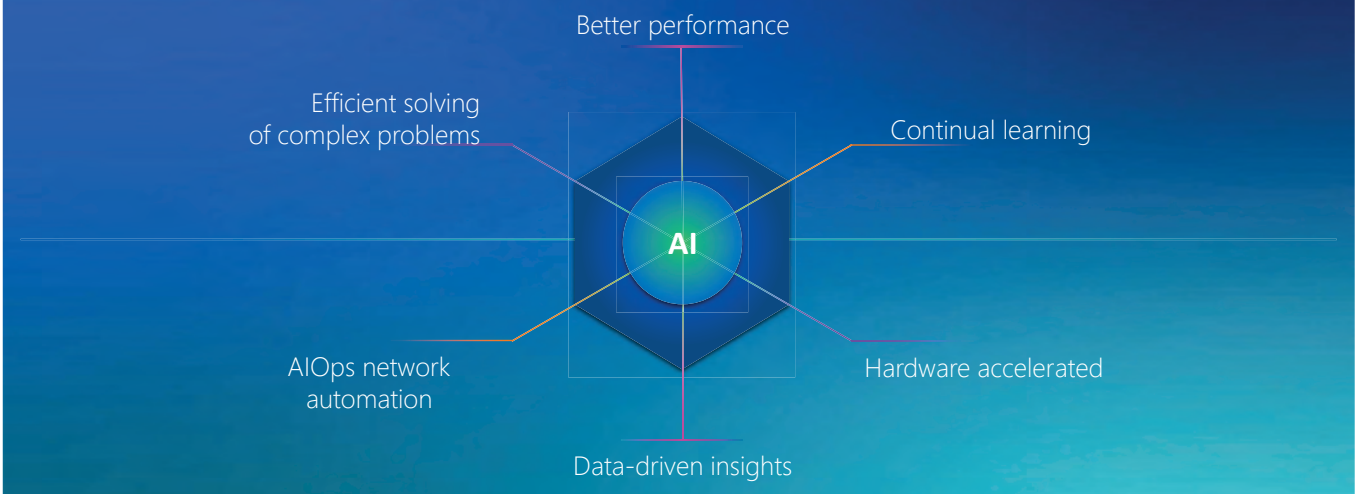
Release 19-20 5G-Advanced updates, 6G studies

Release 21 6G standardization

Release 22+ 6G updates

NOKIA

The promise of AI-Native Networks



NOKIA

Towards the AI-native system

5G: First steps

- Proprietary AI/ML solutions
- Standardized features for Network Automation

5G-Advanced: Standardized enhancements for AI/ML

- AI/ML in air interface: use cases and general framework
- Enhancements for data collection and automation

6G: AI-native system

- AI/ML integrated in all network domains and layers
- Trusted AI/ML
- Distributed and federated AI/ML

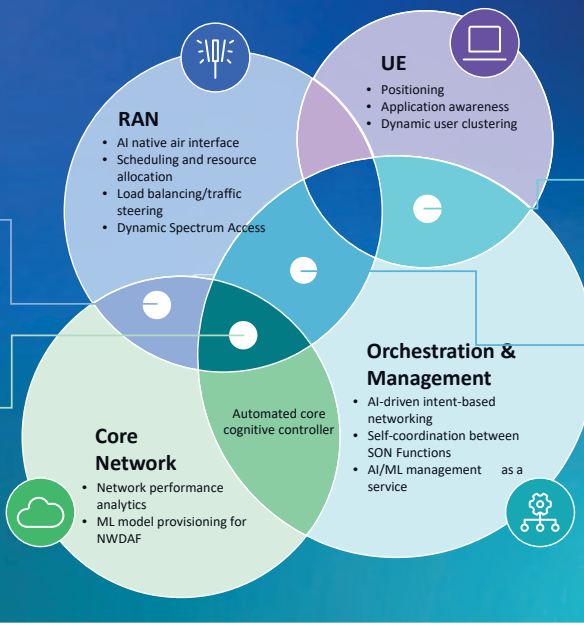
5G 5G 6G

5 © 2024 Nokia | Confidential Orange Only

NOKIA

AI nativeness, at all device and network layers

Illustrative examples



- Energy saving/efficiency
- Mobility prediction
- Slicing
- Flexible function enablement and placement

- Network optimization & maintenance
- Network robustness and resilience

- Device Management
- Robotics
- Predictive maintenance in manufacturing
- Smart homes

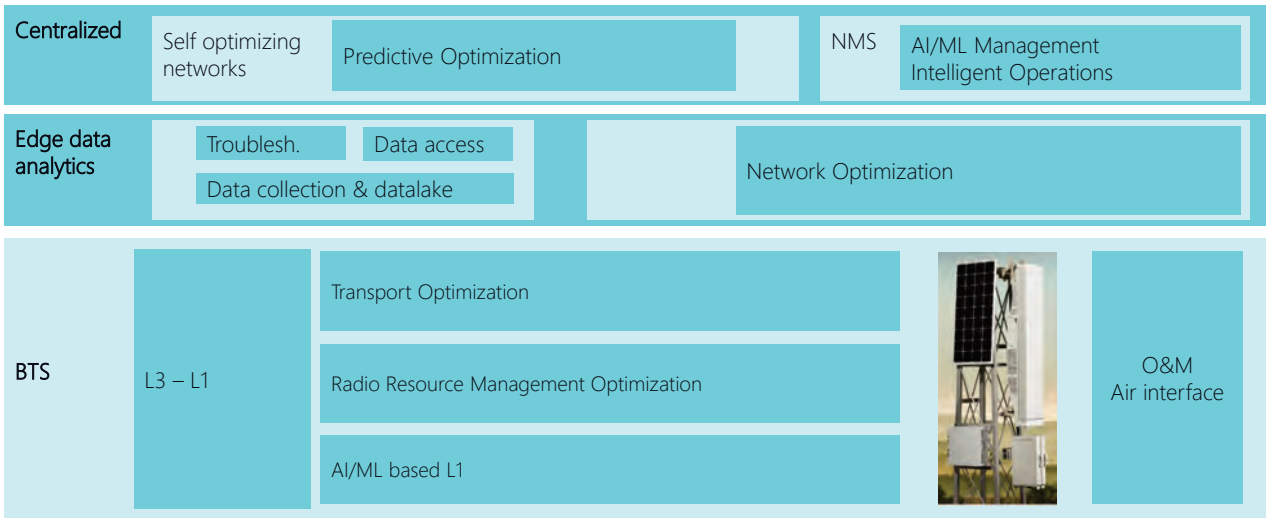
- Beam based coverage optimization
- Mobility optimization
- Multi-agent team learning

6 © 2024 Nokia | Confidential

NOKIA

Radio access networks

Where AI can help



7 © 2024 Nokia | Confidential

NOKIA

AI/ML framework will enable Native AI/ML capability in RAN

Based on 3GPP study items

Creating a future proof ML framework for

- AI for air interface
- AI-enabled energy saving
- AI-enabled load balancing
- AI-enabled mobility optimization
- AI-enabled beam pattern optimization

Key AI/ML use cases in RAN

Native AI/ML (Air interface)

- AI native air interface
- UE Positioning accuracy
- Application awareness
- Beam pattern optimization

RAN Performance and efficiency (BTS embedded L3/L2)

- Load balancing/traffic steering
- Mobility prediction and optimization
- Dynamic Spectrum Access
- Scheduling and resource allocation
- Transport optimization
- RAN Energy efficiency

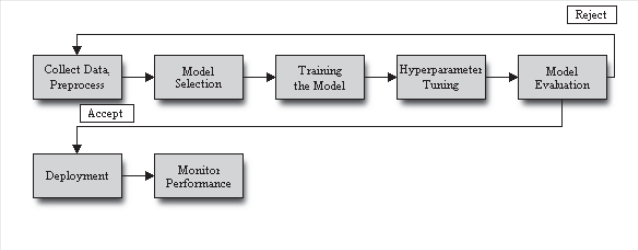
Automation (Network Management and orchestration)

- AI-driven intent-based networking
- AI/ML in SON Functions
- Non-real / Near-real time operation
- Network Energy efficiency

8 © 2024 Nokia | Confidential

NOKIA

How networks are learning?



- Machine learning can be divided into two phases: training and inference.
- The training phase involves feeding a large amount of data to a model and adjusting its parameters so that it can accurately make predictions on new, unseen data.
- The training phase can be computationally costly. This phase is also referred to as model fitting
- The inference phase involves using the trained model to make predictions on new data. The model uses the parameters learned during the training phase and generally is computationally less expensive than the model training.

Machine Learning (ML) model training pipeline typically involves

1. data collection and preprocessing in a suitable format for ML models
2. selecting the ML model architecture and algorithm suitable for solving the problem
3. training the model by using the training data iteratively until it fits the training data well
4. model evaluation using testing data to assess performance
5. model hyperparameter tuning and retraining the model in case the performance is not satisfactory
6. deployment of the trained model in the real-world setting.

5G Advanced I AI/ML in Air Interface (AI-AI)

Rel-18: Three Use Cases for AI/ML Study

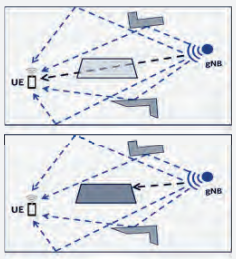
AI/ML-assisted Beam Management

- Predicting the best beam(s) in spatial or time domain
- Purpose: save overhead and reduce latency of beam measurements and improve throughput

- UE sided model: offline training at the UE
- NW-side model: offline training at the NW-side

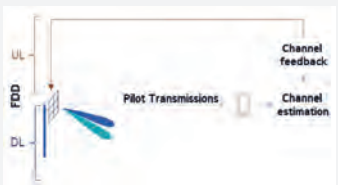
AI/ML positioning

- Positioning enhancements in heavy NLOS (non-Light-of-Sight), capability to process more data
- Purpose: Improve accuracy, save overhead



AI/ML-based CSI feedback

- Leveraging AI/ML models to compress and/or predict the CSI (Channel State Information)
- Purpose: Improve performance, reduce overhead (no channel aging)



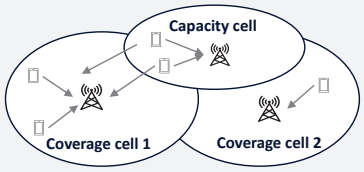
- Two-sided approach, joint and separate training

5G Advanced I NG-RAN

3GPP Study Items

AI/ML Energy Saving

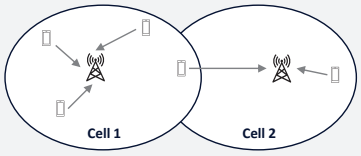
- Energy saving decisions to switch off/on cells
- Based on UE measurements, neighbor node, predict UE trajectory and handover cells



- Additional Capacity cell in high load
- Offload traffic to neighbor cells in low load and switch-off

AI/ML Load Balancing

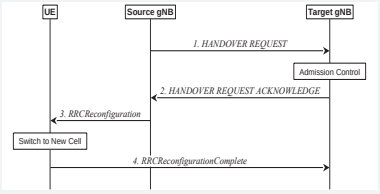
- Distribute load among cells evenly and avoid congestion
- Based on UE measurements, neighbor node, predict UE trajectory, own load and neighbor's load



- Predict best target for AI/ML load balancing and UEs to participate in handover

AI/ML Mobility Enhancements

- Optimize number of unintended events, i.e. handovers, in mobility scenario
- Based on UE measurements, neighbor node, predict UE trajectory and handover cells



3GPP TS 38.300

NOKIA

The challenges of AI/ML applied to mobile networks

Data

- Access
- Complexity
- Governance
- Privacy

Scale

- Model training
- Infrastructure
- Cost
- Sustainability

Ecosystem & Standards

- Interoperability & common framework
- Requirements and capabilities
- Testing

Trust & governance

- AI trustworthiness
- AI security
- AI ethics
- AI regulations

Understanding and mastering AI challenges is fundamental for success
A joint effort of industry, academia, and regulators

NOKIA

AI-RAN alliance intent

Industry alliance to drive AI research and innovation

Our mission

- Unlocking new growth opportunities with AI, enabled by 5G and 6G RAN
- Strengthen network operations by enhancing network efficiency and reducing power consumption

AI-RAN ALLIANCE

AI on RAN
deploying AI services at the network edge to increase operational efficiency and new mobile services

AI for RAN
advancing RAN capabilities to improve spectral efficiency

Alliance members to focus on areas of research and innovation

AI and RAN
integrating processes to utilize infrastructure to generate new AI-driven revenue opportunities

NOKIA Microsoft ERICSSON SAMSUNG T-Mobile SoftBank
DEEPSIG arm aws NVIDIA Northeastern University 東京大学

13 © 2024 Nokia | Confidential

NOKIA

Concluding remarks

- AI/ML in 5G-Advanced is the foundation for AI-native 6G
- A journey of the whole Industry towards AI-native mobile networks
- Standardization is a key enabler to unlock the full potential of AI in telecoms

14 © 2024 Nokia | Confidential Orange Only

NOKIA

Omrežna infrastruktura za pospešitev industrijske inteligence

Network infrastructure to accelerate industrial intelligence

Dejan Rotula

Huawei

POVZETEK

Lahko rečemo, da bo umetna inteligenca preoblikovala družbo v celoti. S samo 7 % leta 2021 naj bi stopnja prodora umetne inteligence v industrijo dosegla 30 % leta 2026. In ker osnovni modeli pospešujejo inteligentno preoblikovanje industrije, bo stopnja prodora do leta 2030 presegla 50 %, kar je 10-kratno povečanje. Hiter razvoj umetne inteligence bo še naprej spodbujal digitalno preobrazbo industrije in zagotovil nove priložnosti za omrežne inovacije.

Do danes se je približno 70 % podjetij preselilo v oblak, pri čemer je porazdeljeni hibridni multi-oblak pretežno priljubljena izbira za takšno migracijo. Ko tradicionalne industrije, kot so energetika, transport in finance, vstopajo v fazo hitre preobrazbe, ki temelji na oblaku, postavljajo različne transportne zahteve za omrežja, zaradi česar morajo omrežja zagotavljati prilagojene zmogljivosti zagotavljanja kakovosti. Za učinkovito podporo oblakov v podjetju se morajo omrežja razvijati v smeri boljše elastičnosti, agilnosti, varnosti in zanesljivosti.

Povečanje računalniške moči umetne inteligence bo spodbudilo vse večje povpraševanje po izgradnji globalnih omrežij podatkovnih centrov (DCN) in preoblikovanju omrežnih tehnologij. Glede na to, da lahko samo 0,1-odstotna stopnja izgube paketov poslabša računalniško zmogljivost za 50 %, je treba za popolno sprostitev računalniške moči zgraditi DCN z visoko prepustnostjo in ničelnimi zastoji.

Digitalna preobrazba industrij zahteva visokohitrostno in stabilno omrežno okolje kampusa. Kampusne storitve se hitro spreminjajo, pri čemer sta mobilna pisarna in videokonference dva glavna razvojna trenda. Približno 80 % prometa v kampusnih omrežjih bo prišlo iz avdio in video aplikacij, kar pomeni, da bodo kampusna omrežja vstopila v dobo, osredotočeno na izkušnje. Glede na te dejavnike je treba obstoječa omrežja nadgraditi, na primer z Wi-Fi 4/5 na Wi-Fi 6/7 in z GE na 10GE dostop.

Z nenehnim razvojem in uporabo tehnologij, kot sta računalništvo v oblaku in IoT, se postopoma materializira inteligentna družba, ki vključuje povezljivost vsega, vsega zaznavanja in vsega inteligentnega.

Kot taka se bodo poslovna omrežja razširila iz pisarne v proizvodnjo, prešla iz statične konfiguracije na prilagajanje na zahtevo in se preoblikovala iz upravljanja ene domene v sodelovanje v celotnem omrežju. Poleg tega se bodo meje omrežja razširile, atributi kakovosti omrežja se bodo povečali, O&M omrežja pa se bodo kvalitativno spremenila. Glede na Huaweijevo vizijo globalne industrije (GIV), bo 97 % velikih podjetij do leta 2025 uporabljalo umetno inteligenco. Omrežja, ki vključujejo zmogljivosti umetne inteligence, lahko presežejo omejitve učinkovitosti ročnega O&M in dosežejo avtonomno vožnjo z visoko stopnjo avtomatizacije in inteligence. Ko so zgrajena, lahko omrežja za avtonomno vožnjo (ADN) utrejo pot inovacijam digitalnih storitev podjetij in agilnemu delovanju.

SUMMARY

We can say that AI will reshape society in its entirety. From a mere 7% in 2021, the industry penetration rate of AI is expected to hit 30% in 2026. And as foundation models accelerate the intelligent transformation of industries, the penetration rate will exceed 50% by 2030, an increase of 10 times. The rapid development of AI will further promote the digital transformation of industries and provide new opportunities for network innovations.

To date, about 70% of enterprises have migrated to the cloud, with distributed hybrid multi-cloud predominantly being the favored choice for such migration. As traditional industries, such as energy, transportation, and finance, enter a phase of rapid cloud-based transformation, they pose differentiated transport requirements on networks, bringing about the need for networks to provide customized quality assurance capabilities. To effectively support enterprise cloudification, networks must evolve toward better elasticity, agility, security, and reliability.

The surge in AI computing power will drive increasing demand for constructing global data center networks (DCNs) and transforming network technologies. Given that a packet loss rate of only 0.1% can deteriorate computing performance by 50%, a DCN with high throughput and zero congestion must be constructed to fully unleash computing power.

The digital transformation of industries requires a high-speed and stable campus network environment.

Campus services are undergoing rapid transformation, with mobile office and video conferencing being the two major development trends. What's more, about 80% of traffic on campus networks will come from audio and video applications, meaning that campus networks will enter the experience-centric era. Given these factors, existing networks need to be upgraded, for example, from Wi-Fi 4/5 to Wi-Fi 6/7 and from GE to 10GE access.

With the continuous development and application of technologies such as cloud computing and IoT, an intelligent society featuring connectivity of everything, all things sensing, and all things intelligent is gradually materializing. As such, enterprise networks will extend from office to production, shift from static configuration to on-demand adjustment, and transform from single-domain management to network-wide collaboration. Furthermore, network boundaries will expand, network quality attributes will increase, and network O&M will undergo a qualitative change. According to Huawei's Global Industry Vision (GIV) 2025, 97% of large enterprises will be using AI by 2025. Networks that integrate AI capabilities can overcome the efficiency limitations of manual O&M, achieving autonomous driving with high levels of automation and intelligence. Once built, autonomous driving networks (ADNs) can pave the way for enterprises' digital service innovation and agile operations.

O AVTORJU

Dejan Rotula je produktni vodja v podjetju Huawei Enterprise IP Network Solutions, odgovoren za regionalno tržno širitev izdelkov, tehnično zasnovo in svetovanje strankam pri rešitvah. Končal je študij elektrotehnike Univerze v Beogradu, smer telekomunikacije. Na trgu informacijskih tehnologij in rešitev ima več kot 20 let izkušenj tako na področju inženiringa kot vodenja, marketinga in prodaje. Moč digitalne preobrazbe in inovacij ga nenehno zanima, kako lahko tehnologija spodbudi spremembe in podjetjem omogoči povečanje rasti.

ABOUT THE AUTHOR

Dejan Rotula is the product manager for Huawei Enterprise IP Network Solutions, responsible for regional market expansion of products, technical design and solution consultancy with customers. He completed his studies in electrical engineering University of Belgrade, majoring in department of telecommunications. In the market of information technologies and solutions, he has more than 20 years of experience both in the field of engineering and management, marketing and sales. The power of digital transformation and innovation makes him constantly curious about how technology can drive change and enable businesses to enhance growth.

Data Network Infrastructure to Accelerate Industrial Intelligence

Networks Accelerate AI and AI Redefines Networks

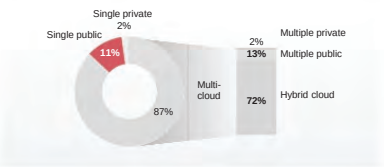
Building a Fully Connected, Intelligent World

Contents

Multi-Cloud Has Become the New Normal for Enterprise Digitalization

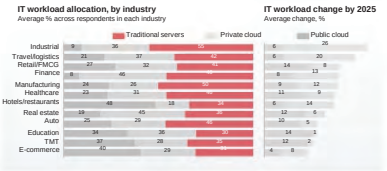
- **Multi-cloud has become the new normal for enterprises.** Enterprises are gradually transitioning from public clouds to hybrid clouds, industry clouds, edge clouds, and distributed clouds in order to achieve cost savings, data security, and cloud technology integration. Multi-cloud allows enterprises to flexibly use different cloud services based on their business needs and distribute multiple workloads across different cloud platforms. According to the Flexera 2023 State of the Cloud Report, 87% of surveyed enterprises have already been using multi-cloud services, and this shows that multi-cloud has become the new normal for enterprise digital transformation.
- **Cloud migration in traditional industries is accelerating.** With advances in cloud computing, big data, and artificial intelligence (AI), cloud services have become a driving force for service innovation and enterprise upgrade in a wider range of fields. More enterprises are embracing the cloud in order to keep pace with technological transformation and seek new development opportunities. In addition to the Internet industry, cloud service practitioners are now also seen in traditional, non-digital-native industries, such as industrial, education, healthcare, government, energy, and finance. Digital transformation is driving cloud migration across industries. Take China as an example. According to the McKinsey 2021 China Cloud Computing Survey, the share of traditional industries' IT workloads running in the cloud will increase significantly by 2025.
- **Cloud migration across industries is dominated by distributed, hybrid multi-cloud.** Cloud migration across industries emerges as enterprises continue to pursue higher efficiency, cost-effectiveness, and business growth. Hybrid multi-cloud combines the advantages of public and private clouds, and not only ensures enterprise data security but also provides a flexible cloud architecture. This makes it a popular choice for enterprises. In addition, a large number of emerging business applications require massive data analysis and computing capabilities. A multi-layer and distributed cloud computing model is going mainstream across industries. Examples include the deployment of "three DCs in two cities" in the financial industry and "distributed DCs + public clouds" in the energy industry.

87% of surveyed enterprises are already using multi-cloud in 2023



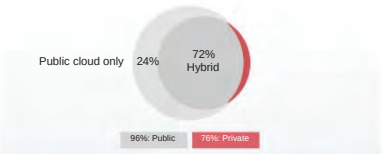
Source: Flexera 2023 State of the Cloud Report

The share of traditional industries' IT workloads running in the cloud will increase significantly by 2025



Source: McKinsey 2021 China Cloud Computing Survey

72% of surveyed customers use hybrid cloud



Source: Flexera 2023 State of the Cloud Report

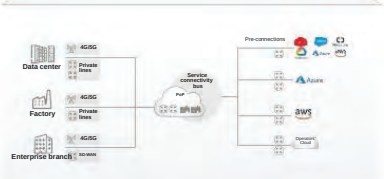
Carriers Offering Multiple Innovative Multi-Cloud Network Models

Traditional carrier networks cannot meet industries' multi-cloud service requirements. Cloud computing—built on networks—and its applications are growing rapidly. Its network requirements are shifting from simple private line access to multi-cloud networks with elasticity, agility, real-time visualization, and reliable experiences. However, carriers traditionally focus on the deployment and O&M aspects of networks, and cannot meet enterprise requirements for business network provisioning speed, adjustment flexibility, and intelligence.

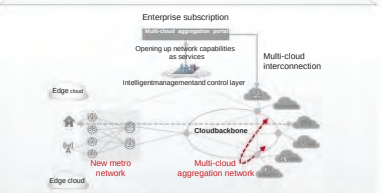
Carriers have started to innovate in multi-cloud networks. Carriers have huge network infrastructure, which is advantageous as it allows them to provide various services for enterprises. However, it may also be a burden. When there is a new business need, the sheer volume of network assets means that carriers can only implement network transformation in stages.

- **Cloud Private Line Mode, adding Packages to Increase Profits.** Traditional networking or Internet private lines are replaced by site-to-cloud private lines and multi-cloud connectivity private lines, and SD-WAN is deployed to achieve the agile provisioning of site-to-cloud private lines. Gateway capabilities that support any access methods and pre-connections to multi-cloud resource pools are enabled based on point of presence (POP) resource pools. Site-to-cloud private lines connect enterprises to multi-cloud on demand through POP.
- **Multi-cloud aggregation mode, preempting a unified procurement entry.** An SRv6 cloud backbone is built based on site-to-cloud private lines to achieve pre-connections to local different cloud resources. Carriers offer E2E SRv6 capabilities from enterprises to backbone networks to enable the automated orchestration of enterprise networks in the cloud based on business needs, providing enterprises with access to flexible and high-quality multi-cloud connectivity. Carriers build a multi-cloud aggregation platform with APIs that can be connected to third-party clouds so that they can resell third-party cloud services. This innovative business model fully captures the benefits of accessing multiple clouds with one single line.
- **Industry private network mode, ensuring experience for valued customers.** As office work and production continue their migration to the cloud, high-value sectors like finance, government, and education require isolation from public services for security reasons, and require high-quality networks to deliver a premium experience. To meet these requirements, carriers have started to deploy network slices or industry-specific physical networks based on the SRv6 cloud backbone. In addition, the SLA performance of services running on the cloud is analyzed and displayed in real time through tenant network traffic and performance indicators. This allows tenants to understand the service quality of their private lines in real time and this information can inform SLA monetization efforts. The centralized monitoring of each tenant's SLA performance helps promptly identify incidents like cloud-based network traffic and performance indicator deterioration, so that optimization and targeted maintenance can be performed to enhance cloud service experiences.

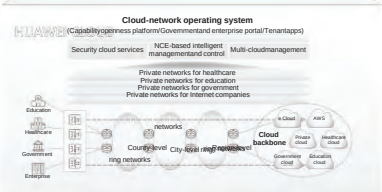
Model 1: Cloud Private Line Mode, Adding Packages to Increase Profits



Model 2: Multi-cloud aggregation mode, preempting a unified procurement entry



Model 3: Industry private network mode, ensuring experience for valued customers

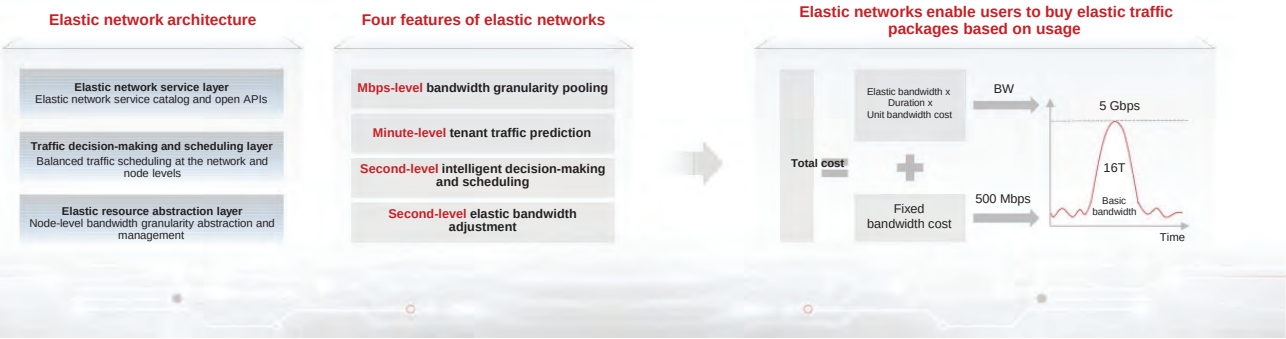


Key Feature 1: Elasticity and Agility to Support the Flexible Scheduling of Computing and Storage Resources

Purchasing private lines with insufficient bandwidth can negatively affect service experiences. Maintaining high-bandwidth private lines long term can be costly. Temporary high-bandwidth services include high-bandwidth real-time communications and periodic data migration. High-bandwidth real-time communications are instantaneous. They are triggered primarily by specific events, and last between several hours and several days. The bandwidth shortage problem cannot be solved by arbitrarily stretching communication time as this compromises the instant communication experience. Periodic data migration does not have high real-time requirements, but does have requirements for total migration time.

Elastic site-to-cloud private lines meet the bandwidth requirements of enterprise services with significant traffic peaks. Elastic computing, as a part of cloud computing, has matured. From the perspective of consumers, elastic services fulfill their needs at an optimal cost. These services are designed to meet the requirements across different business scenarios and are offered under the pay-as-you-use (PAYU) model, which makes them highly cost-effective. From the perspective of suppliers, elastic services are essentially efficient management of resources to maximize their effectiveness. Elastic site-to-cloud private lines learn from the concept of elasticity in cloud computing, where network bandwidth resources are pooled and tenant service traffic is detected and predicted in real time to facilitate the flexible scheduling of bandwidth resources across the entire network, ensuring the service experiences of elastic site-to-cloud private lines.

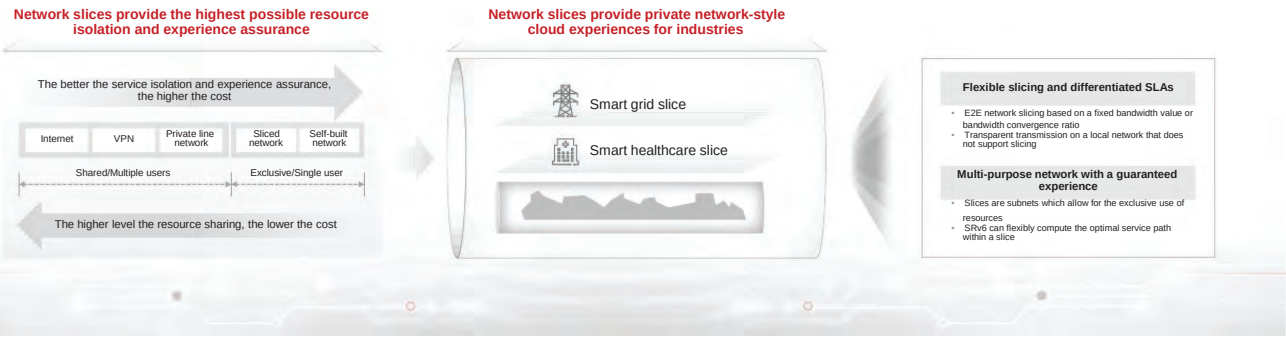
Elastic private lines guarantee user experiences and help monetize network resources. Elastic private lines mean enterprises can temporarily increase bandwidth or purchase a traffic package based on their business needs while retaining a fixed-bandwidth private line, with the temporarily increased bandwidth or purchased traffic package offering the same quality as the fixed-bandwidth private line does. This truly allows for on-demand purchase and pay-per-use. Carriers can make full use of idle bandwidth resources to maximize network value.



Key Feature 2: Service Isolation to Guarantee the Quality of Critical Cloud-Based Services

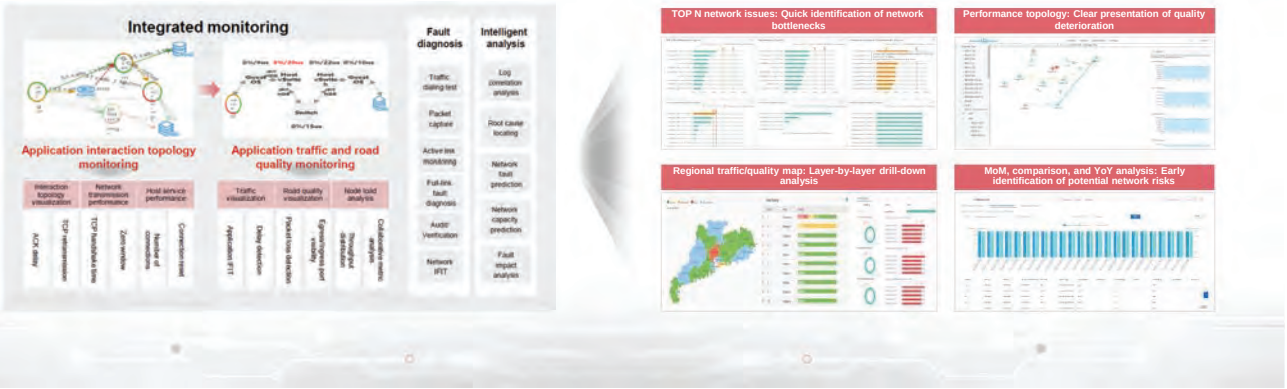
Cloud migration across industries raises higher requirements for security isolation and service experiences. Network slicing can be used to meet the security isolation and differentiated assurance requirements of different services on the same network.

- **Resource and security isolation:** From the perspective of service quality, the purpose of IP network slice isolation is to prevent a service burst or abnormal traffic in a slice from affecting other slices in the same network, which ensures that services in different network slices do not affect each other. This is especially important for vertical industries, such as smart grids, which have strict requirements on latency and jitter and whose performance is highly sensitive to the impact of other services. From the perspective of security, if information about services (private line services, such as finance and government services) in an IP network slice is not expected to be accessed or obtained by users in other network slices, effective security isolation measures need to be taken between different slices.
- **Differentiated SLA assurance:** Network slicing enables carriers to expand beyond selling traffic and provide differentiated services in the form of slices for tenants from different industries. On-demand, customized, and differentiated services will be the main business model for carriers in the future, and will also be a new value-generating area of growth for them.
- **High reliability:** High-value services and ultra-reliable low latency communications (URLLC) require IP networks to provide high availability. Millisecond-level fault recovery has become a basic requirement for IP networks. SRv6-based network slicing provides local protection against any faults on the IP network, such as Topology-Independent Loop-Free Alternate (TI-LFA) and midpoint protection. These protection technologies can help significantly improve the effectiveness of protection and enhance the reliability of IP network slices. In addition, link failover in a network slice can be performed within the slice without affecting other slices.



Key Feature 3: Real-Time Visualization to Monitor Service Quality from End to End

- **A lack of network visibility leads to inefficient O&M.** The complexity of enterprise networks will grow exponentially. This is reflected in the following aspects: (1) As hybrid office becomes popular, there will be more interconnected branches and access locations; (2) The convergence of office networks and the Internet of Things (IoT) will lead to a surge in connections; (3) Cloudification and new applications will raise higher and more volatile requirements for network performance; (4) A growing number of network equipment varieties and manufacturers can significantly scale up equipment management workloads; and (5) Requirements for network assurance will be higher, and will move from connectivity-oriented to experience-oriented. Meanwhile, the number of O&M assurance engineers will not increase proportionally, if at all, which means that more work will need to be done by fewer people. As a result, the pain points of network O&M will become more evident. There is no unified view to assess the health of enterprise networks. Users may have poor network experiences, resulting in an increased number of fault complaints. Fault recovery can be slow. To sum up, network O&M is not even close to keeping pace with enterprise digital transformation.
- **Network visualization enables the real-time detection of network changes.** Network visualization comprises real-time, dynamic, and HD network-wide resource visualization capabilities. Key technologies like big data computing engines, AI, search algorithms, route simulation, and verification algorithms are used to achieve multi-dimensional visibility, path navigation, searching and locating, and deterministic application experience assurance. Real-time network quality visibility, demarcation and locating, and self-healing capabilities are also provided to help customers switch from the traditional O&M model using static topology to one using dynamic HD electronic maps. In other words, digital maps are used to give an intuitive view of networks and significantly boost network O&M efficiency.

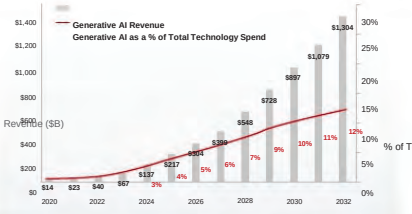


Contents

AIGC Creates Trillion-Dollar Industry Market Spaces and Accelerates Global Computing Infrastructure Construction

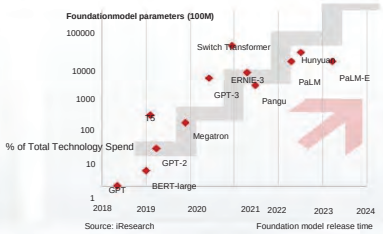
- **Foundation models proliferate and the AIGC era arrives.** Sparked by the release of ChatGPT by OpenAI in November 2022, the AI industry has quickly entered a new era of Artificial Intelligence Generated Content (AIGC), creating a new paradigm for human and production interaction. This has also ignited a content productivity revolution in the AI era. According to the latest report from Bloomberg Intelligence, the market space of generative AI will reach US$1.3 trillion by 2032, a significant jump from US$40 billion in 2022, hitting a compound annual growth rate (CAGR) of 42%.
- **AIGC will trigger in-depth transformation across the entire industry.** AIGC is penetrating a wide range of industries at an accelerating rate. Generally, AIGC mainly affects content creation and human-machine interaction. As industries become more online-oriented and the proportion of content in the value chain grows, the more obvious the disruptive effect of AIGC will be. For example, the e-commerce, gaming, and advertising industries are particularly online-oriented, and the content quality directly determines the value created. Therefore, AIGC can maximize its benefits only when it is widely applied across industries.
- **Global computing infrastructure construction is speeding up.** The parameter quantity of OpenAI's GPT series models already exceeded 117 million in June 2018. That quantity has since mushroomed, reaching billions or even trillions. On average, the number of model parameters doubles every three to four months, bringing an increasing demand for training computing power. For each 1% increase in the computing power index, the digital economy and Gross Domestic Product (GDP) increases on average by 3.5‰ and 1.8‰, respectively. Computing power is becoming a key factor that affects a nation's comprehensive strength. The construction of computing infrastructure has become a strategic initiative for the high-quality development of a nation's digital economy. According to IDC, global enterprises' investment in AI infrastructure and services is expected to exceed US$200 billion by 2025, far surpassing enterprises' digital transformation (DX) and GDP.

Generative AI market revenue forecast (2020–2032)



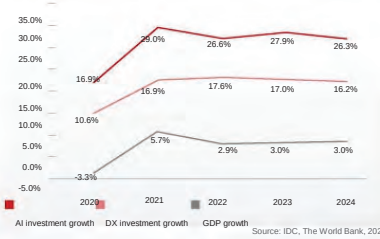
Source: Bloomberg Intelligence

Parameter quantity change trend of global foundation models



Source: iResearch

Global AI investment growth far exceeding DX and GDP

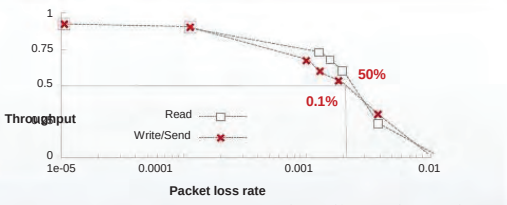


Source: IDC, The World Bank, 2022

Network Determines Training Efficiency, and Conventional Networks Cannot Meet AI Requirements

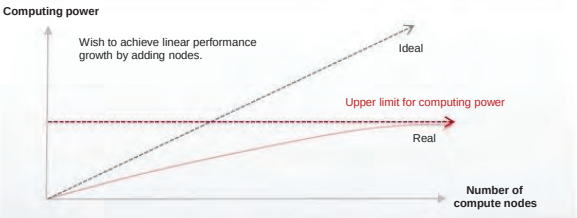
- **Conventional Ethernet cannot meet the requirements of AI DCs**. According to IDC, Ethernet accounts for more than 95% in mainstream DCs. However, in AI training scenarios, conventional Ethernet doesn't perform well in terms of network throughput, latency, and packet loss prevention. The communication mode of AI applications brings a new challenge to CPU and GPU servers as well as the existing underlying network infrastructure. Conventional Ethernet is natively prone to packet loss and cannot meet the requirements of DCs in today's AI era, where data loss cannot be tolerated during AI training.
- **An ultra-large network is required for a compute cluster with 10k GPUs**. To launch foundation models more quickly and meet the growth requirements of model parameter and token quantities, the cluster scale has increased from 1k GPUs to 10k GPUs. Take OpenAI's GPT-4 as an example. It uses thousands of GPUs to train 1.8 trillion parameters. A large-scale training network is required.
- **An ultra-high-throughput network is required for models with trillions of parameters.** Foundation models are trained in distributed mode to improve training quality and speed. In this mode, vast quantities of parameters are distributed to multiple GPUs housed in multiple servers. As such, thousands or even tens of thousands of GPUs are required to train dozens of TBs or more of data. Heavy communication traffic between many GPUs is prone to network load imbalance, which then decreases the network throughput. As a result, the overall AI training performance deteriorates.
- **A highly reliable network is required for long-term stable training.** Foundation model training is a complex project. Ensuring that systems run stably is of vital importance to the entire training process, spanning data preparation, model pre-training, and model training. Network infrastructure is the key to long-term stable training. Take a model with hundreds of billions of parameters as an example. Its total training duration is 65 days, but due to system faults causing the model to restart more than 50 times, the effective training duration is only 33 days. Typically, the training duration of foundation models is long with many interruptions.

Just 0.1% packet loss rate can reduce network throughput by 50%



Source: Congestion Control for Large-Scale RDMA Deployments

Network performance causes an upper limit for computing power, leading to serious ROI imbalance



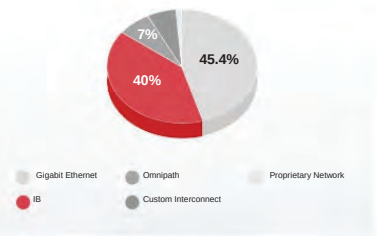
Continuous Innovation of Ethernet, Promoting AI DCs from Being Closed to Being Open

- **Active industry layout:** In July 2023, the Linux Foundation joined hands with multiple vendors to establish the Ultra Ethernet Consortium (UEC), with the aim of improving the data transmission speed and network performance so as to better adapt to ever-growing AI and HPC workloads. According to the UEC's chairperson, the project is built on Ethernet technology because it is the best representative of durable, flexible, and adaptable basic network technologies in the industry.
- **Wide application**: InfiniBand (IB) features high bandwidth and low latency, and is typically the solution of choice for building conventional HPC networks. However, the IB architecture is closed and has poor scalability, leading to high network deployment and maintenance costs. As Ethernet advances, it is finding wider applications in HPC and AI fields. According to the latest statistics, 45.5% of the world's top 500 HPC systems use Ethernet interconnection, surpassing IB. Remote Direct Memory Access over Converged Ethernet (RoCE) networks have been widely used in the compute clusters for foundation models, such as PCL-G, Huawei PanGu-Σ, and Baidu ERNIE Bot.
- **Continuous implementation of innovative solutions:** Industry players continue to innovate based on Ethernet technology. Since Huawei released the AI Fabric ultra-fast Ethernet solution in October 2018, multiple mainstream vendors have been proactive in promoting technical breakthroughs, and have launched high-speed interconnection products and solutions tailored for the HPC and AI fields one after another.

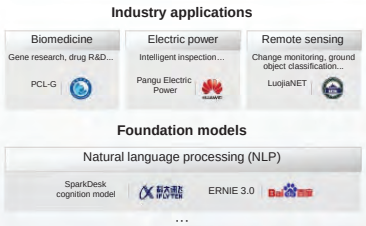
Mainstream players release new products based on Ethernet technology innovations one after another

Time	Vendor	Event
October 2018	Huawei	Released the AI Fabric ultra-fast Ethernet solution.
August 2020	HPE	Released Slingshot — Ethernet interconnection technology for HPC.
April 2022	Inspur	Released the RoCE-based lossless Ethernet solution.
May 2023	NVIDIA	Released the high-performance Ethernet architecture — Spectrum-X.
July 2023	Microsoft, Broadcom, AMD, Intel, etc.	Jointly established the UEC.

Among global top 500 HPC systems, Ethernet deployment exceeds IB deployment



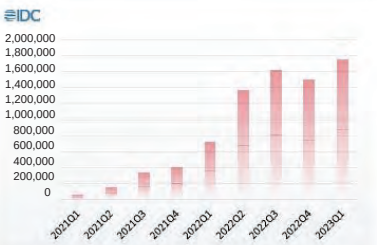
Ethernet-based RoCE networks are widely used in foundation models



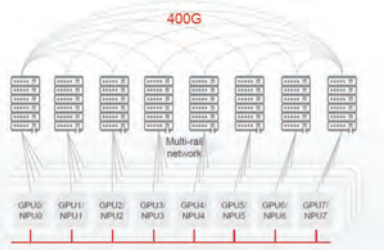
400GE Switches Enter the Batch Deployment Phase, Supporting Ultra-Broadband Simplified Architectures

- **Open ecosystem and rapid technology development**: Ethernet has always been an open ecosystem, laying a solid environmental foundation for network technology iteration. After more than 40 years of development, Ethernet has rapidly evolved from offering 10 Mbit/s to now offering 400 Gbit/s, providing ultra-broadband channels for high-speed transmission of massive amounts of data in AI scenarios. According to IDC, from 2021 to 2023, the CAGR of 400GE port shipments was 46%. In the first quarter of 2023, the 400GE port shipments reached 1.73 million.
- **Mature industry spanning standards and products:** In 2013, the 400G Ethernet standards project was officially initiated. In 2017, the Ethernet standards defined by IEEE 802.3bs were approved, indicating the maturity of the 400GE standards. As of today, all mainstream vendors are capable of offering 400GE switches. In 2019, Huawei released the industry's first highest-density 400GE DC switch built for the AI era — CloudEngine 16800.
- **400GE switches are ideal for building an ultra-broadband simplified architecture**: So far, AI clusters generally use high-performance network interface cards (NICs), each of which can provide 200G or even 400G bandwidth. This brings an increasingly urgent need for 400G access and interconnection. By adopting high-bandwidth Ethernet switches, it is possible to build a flexible network architecture to meet networking requirements in different service scenarios. Two important architectures that have attracted wide attention in the industry are the multi-rail network architecture and the Clos network architecture. In the distributed AI training scenario, the multi-rail network architecture only needs to build multiple independent network planes to connect GPUs/NPUs with the same number. This architecture, compared with traditional ones, can effectively reduce the number of network layers, number of data forwarding hops, and network construction costs. The Clos architecture uses a network-wide oversubscription-free design, helping to build a non-blocking large-capacity network featuring on-demand horizontal expansion. This architecture has higher universality and scalability and supports larger-scale networking requirements.

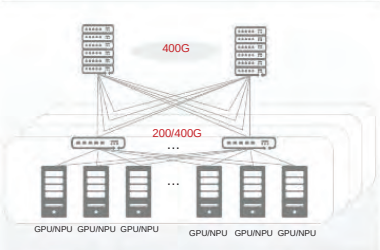
IDC: 400GE port shipment statistics



Multi-rail network architecture, reducing network layers and network construction costs



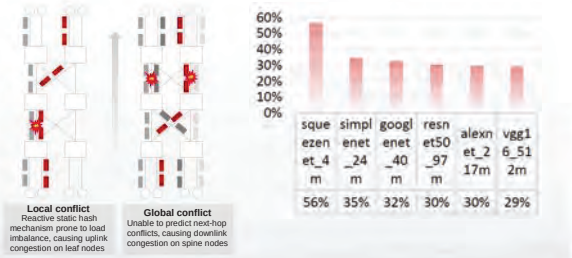
Clos network architecture, featuring high universality and scalability



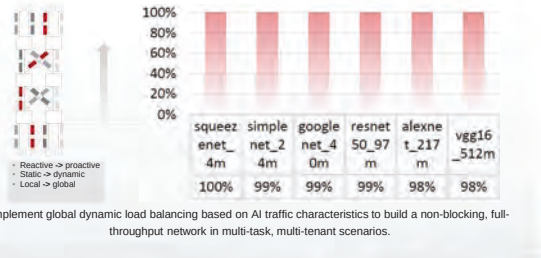
Zero Packet Loss -> Zero Blocking, Improving AI Training Efficiency

- **Network throughput is the key to AI training efficiency.** Although mainstream vendors already have ways to address Ethernet packet loss, zero packet loss is the minimum network requirement of AI computing. Furthermore, we need to find ways to improve network throughput. There is an inherent problem in network transmission: The receiver will trigger traffic retransmission upon receiving out-of-order traffic, leading to a reduction in speed. AI training mainly produces elephant flows (100 MB to several GB) — while the number of these flows is small, each one involves a large amount of data. If such flows are transmitted in the conventional load balancing mode, network nodes steer traffic only from their own perspectives. As a result, load imbalance can easily occur (the annual network throughput is just 50%). The next round of communication can start only after the slowest flow in this round reaches its destination, meaning that the slowest flow determines the overall network performance. On a network with no global load balancing technique, the overall communication efficiency is between 30% and 50%. This means that half of the network performance is wasted, and the computing power utilization of the entire cluster is only 30% to 50%.
- **Network scale load balancing (NSLB) improves network throughput.** To improve network throughput, mainstream industry players prefer to perform in-depth collaboration and adaptation of devices, networks, and protocols. This allows them to implement network-wide load balancing and achieve over 90% network throughput in addition to facilitating the adaptation of RoCE networks to foundation model training requirements. Huawei's NSLB technology is an ideal choice here. It enables collaboration between the network controller and AI scheduler to perform global path calculation based on the traffic congestion status of network-wide switches and the network topology, and obtains the communication matrix based on the training tasks assigned by the AI scheduler. In addition, NSLB can identify the optimal path based on the communication library, network topology, bandwidth, and congestion status, and automatically deliver the optimal path to network switches. Service flows are then transmitted along the path. All of this helps to improve the network throughput to over 90%.

Conventional solutions are prone to load imbalance, causing network congestion and affecting the training speed



Huawei's NSLB technology improves network throughput to over 90%



Implement global dynamic load balancing based on AI traffic characteristics to build a non-blocking, full-throughput network in multi-task, multi-tenant scenarios.

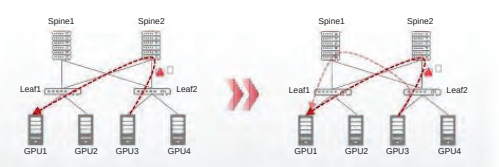
Fault Rectification Within Milliseconds, Ensuring High Cluster Stability and Reliability

- **High network reliability is fundamental to cluster system reliability.** The AI computing center network on which the foundation model relies serves as the core hub of service traffic. The stability of this network directly affects the stability of the entire cluster system. Because the network fault domain is large, failure of a single network node will affect the connectivity of dozens (or more) of compute nodes. In addition, unlike a single GPU or server that is easy to be isolated, the network shares resources as a cluster. Any performance fluctuation can affect the utilization of all computing resources. As such, continuous network stability is vital to foundation model training. Improving the fault rectification capability and O&M efficiency of the network is therefore an urgent problem that needs to be resolved.
- **Technological innovation directions for high network reliability:**

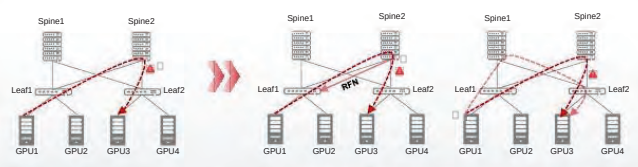
(1) Fast hardware awareness, fault rectification in sub-milliseconds: In the AI training scenario, each host communication task takes only a matter of milliseconds to complete. However, the conventional route convergence mechanism takes seconds, potentially interrupting multiple rounds of AI host communications and significantly affecting AI efficiency. One method to solve this problem is to leverage the Data Plane Fast Recovery (DPFR) technology. This technology provides capabilities such as rapid fault detection and rapid fault rectification on the local side and remote side, thereby implementing rapid link failover in sub-milliseconds with no impact on the training task.

(2) Pre-training intelligent self-check and in-training intelligent operations and maintenance (O&M): Generally, 90% of faults on high-performance networks are caused by incorrect configurations. And, as the AI training cluster scale grows, the configuration complexity keeps increasing. The computing-network collaboration mechanism — widely considered as an important technology for stable delivery of AI clusters — provides an AI scenario-oriented network model to achieve automatic generation, delivery, and detection of network configurations. Furthermore, foundation models are characterized by heavy traffic and short periods. The conventional polling and packet sampling mechanism cannot support visualization of AI network traffic indicators, and the entire network is regarded as a black box. Through network performance measurement in milliseconds as well as collective communication performance measurement in the computing-network collaboration solution, high service visibility, poor-QoE issue analysis, and rapid fault demarcation are implemented. In addition, the cluster computing O&M platform is used for unified resource scheduling in order to rapidly rectify network faults. This is another exploration direction of industry players.

Fast fault rectification on the local side



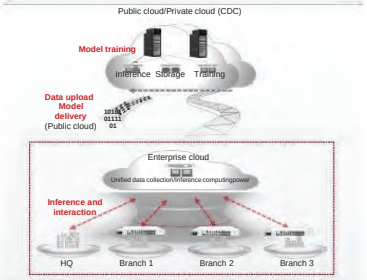
Fast fault rectification on the remote side



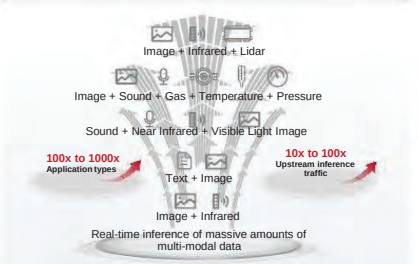
WANs Are Evolving Towards Elastic and Intelligent, Accelerating AI Inference

- **AI inference brings new network requirements.** As AI technologies and industrial intelligentization gain momentum, the network plays an increasingly important role in connecting human beings, things, and conventional applications, and supporting intelligent applications throughout their lifecycles spanning foundation model training, distribution, inference, and iteration. On the one hand, model training on the cloud and inference off the cloud bring massive amounts of data transfer, requiring high network bandwidth and throughput. On the other hand, as vast numbers of AI inference terminals and applications enter the core production system of enterprises, the number of applications grows 100-fold. Different AI applications have diverse network requirements. Take a typical industrial campus network as an example. AOI machine vision requires real-time inference and interaction, software package download requires high bandwidth, and video conferencing requires stable bandwidth. All of this brings a new challenge: how to enable fine-tuned and differentiated experience assurance on the network.
- **Building 400GE/800GE elastic and intelligent WANs.** The industry is exploring how to use 400GE/800GE devices to build an ultra-broadband network and leverage the network-terminal-computing collaboration technology and intelligent scheduling algorithm to perform intelligence awareness and analysis of applications, accurately predict network traffic change trends, and intelligently optimize network resource allocation by application type so as to remove network congestion in advance, ensure efficient transmission of massive amounts of training data, and meet differentiated service assurance requirements of applications.

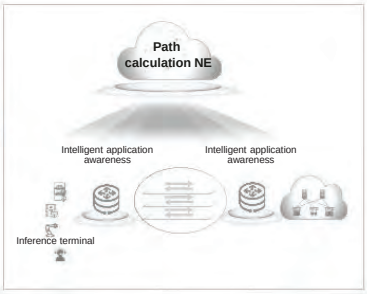
PB-level training data and TB-level model files delivery, bringing massive data transmission requirements



AI enters the production system and applications increase 100-fold, networks face challenges



Elastic and intelligent WAN, accelerating inference and interaction



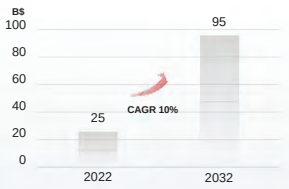
Contents

Emerging Services Drive Campus Network Upgrade to Better Support Enterprise Digital Transformation

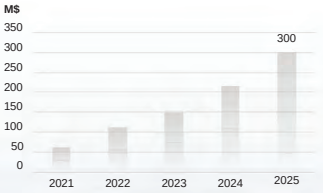
With the acceleration of enterprise digital transformation, new services and applications emerge, which not only improves enterprise office and production efficiency, but also poses new demands on campus networks. Currently, the main driving forces are as follows:

- **Widely-used video conferencing:** Video conference becomes an important tool for remote communication and hybrid office. The video conference market is expected to grow by 10% every year globally and reach US$95 billion by 2032. Take Huawei as an example. Video conferences connect nearly 400,000 users from employees to partners, covering more than 1000 office sites in 170 countries. The number of online users reaches 60,000 during peak hours, and more than 600,000 conferences are held every month. The quality of video transmission directly determines the communication efficiency.
- **Massive deployment of IoT applications:** In addition to traditional office terminals, enterprise digitalization will lead to massive IoT terminals, including devices for asset management, electronic shelf labels, and environment sensors. Take the retail industry as an example. A large number of supermarkets begin to replace traditional paper shelf labels (PSLs) with electronic shelf labels (ESLs) that support remote and real-time price update. The global ESL market will exceed US$3 billion by 2025. The rapid popularization of IoT applications and devices will further complicate enterprise IT systems.
- **Upcoming terminal upgrading:** Countries successively granted the 6 GHz frequency band, which will drive the upgrading of Wi-Fi terminals. By the first half of 2023, the number of terminals supporting 6 GHz has reached 2064, an increase of 260% compared with the first half of 2022. 67 devices support Wi-Fi 7, including 22 phones, 30 routers and gateways, 11 access points, and 4 laptops. In the process of upgrading terminals, enterprises will also expedite to upgrade WLAN for better accessing new terminals.
- **Emerging services (such as immersive experience):** As computing power and video display technologies mature, immersive applications are emerging, including holographic projection, naked-eye 3D, and metaverse office. If enterprises recently consider to innovate such technologies, campus networks will also be affected.

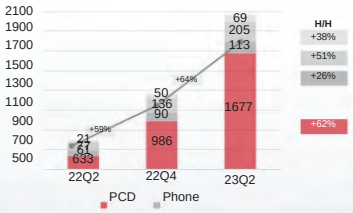
Enterprise video conferencing market forecast



Global ESL market forecast (2021 to 2025)



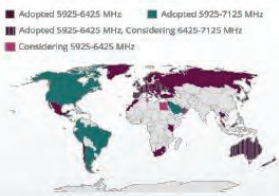
Rapid popularization of 6 GHz terminals



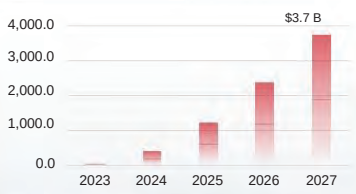
WLAN Enters the Wi-Fi 7 Era, Accelerating Wireless Network Upgrade for Campuses

- **Wi-Fi 7 has entered the phase of commercial use.** Draft 4.0 of the Wi-Fi 7 standard (802.11be) was released in July 2023, and is expected to be finalized and officially released in Q1 of 2024. In terms of chips, vendors such as Qualcomm, Broadcom, MTK, and Intel have released Wi-Fi 7 chips since 2022. In terms of spectrum, 54 countries around the world have granted 6 GHz frequency band for Wi-Fi, covering Europe, Asia Pacific, Middle East, and Latin America. Huawei will officially launch the industry's first enterprise-level Wi-Fi 7 in September 2023. Most other vendors will launch Wi-Fi 7 products to the market in 2024.
- **The enterprise-level Wi-Fi 7 market is about to enter a period of rapid development.** According to Gartner's prediction, the shipment of enterprise-level Wi-Fi 7 APs will reach 12.4 million by 2027, equivalent to 27% of the total number of APs. In particular, more than 30% of organizations in the manufacturing market will upgrade to Wi-Fi 7, which will introduce more use cases to their business processes. At the same time, the combination of Wi-Fi 7 and time-sensitive networking (TSN) greatly improves network bandwidth and reliability, which can support key business processes in manufacturing and warehousing, and accelerates the adoption of next-generation Wi-Fi technologies.
- **Wi-Fi upgrade will also drive wired network upgrade.** The peak rate of Wi-Fi 7 exceeds 10 Gbps, and 2.5GE will become the minimum requirement for Wi-Fi 7 APs, which means that traditional GE switches cannot match the bandwidth demands of next-generation enterprise wireless networks. That is why access switches are being upgraded from GE to 2.5/5GE, which will drive the delivery of 25GE aggregation switches and 100GE core switches. According to the market share data released by IDC, the 2.5/5GE, 25GE, and 100GE port shipment grew 108%, 78%, and 62% in 2022, respectively. It is estimated that the shipment will continue to grow rapidly over the next few years, driving enterprise campus networks into the 10GE era.

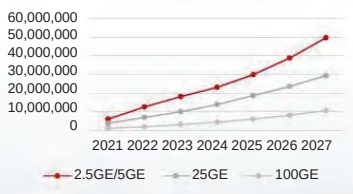
6 GHz frequency band granted in 50+ countries



Global enterprise Wi-Fi 7 market revenue forecast



Multi-GE/25GE/100GE port shipment forecast

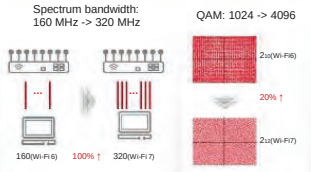


Bandwidth & Reliability Improvement, Accelerating Scenario-Specific Application of Wi-Fi 7 Across Industries

- **Higher bandwidth:** Compared with Wi-Fi 6, Wi-Fi 7 supports the 6 GHz frequency band in addition to the 2.4 GHz and 5 GHz frequency bands. This reduces signal interference while providing wider spectrum resources, enabling 160 MHz continuous networking. Together with 4096-QAM, the bandwidth is increased by 2.4 times, meeting high-bandwidth needs of applications such as 4K video, AOI-based HD quality inspection, in-vehicle software installation, and AR/VR.
- **Lower latency:** Compared with Wi-Fi 6, Wi-Fi 7 uses multi-resource unit (RU) to flexibly combine RUs over the air interface and allocates multiple RUs to a single user, improving air interface resource utilization. In this way, the average latency is reduced by over 25%. Given this, Wi-Fi 7 is especially suitable for high-quality office scenarios, which can provide better assurance for delay-sensitive services, such as HD video conferencing, interactive office, and cloud multimedia rendering.
- **Higher reliability:** Wi-Fi 7 has been greatly improved in terms of link reliability and user experience assurance. With the multi-link operation feature, multiple data connections (2.4 GHz, 5 GHz, and 6 GHz) can be set up between terminals and APs. The three links can transmit and receive data at the same time to increase the link bandwidth. They can also transmit and receive the same data (multi-fed and selective receiving) to improve the link reliability. In addition, they support data link matching based on application identification to ensure differentiated experience. These features will provide better choices for AGV-based smart warehousing and flexible manufacturing.
- **Application scenarios:** Wi-Fi 7 can be widely used in wireless terminal reconstruction scenarios, such as smart production lines, smart warehousing, industrial terminal control, vehicle drive tests, and future metaverse.

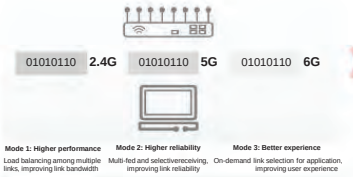
Higher wireless bandwidth

Peak rate of a single terminal: 5 Gbps



Higher link reliability

Single Link -> Multiple links

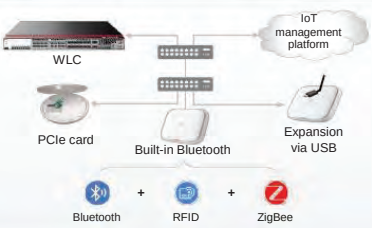


More application scenarios

Multiple Networks -> One Converged Network, Optimizing Enterprise Network Investment

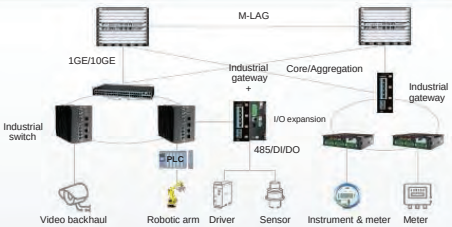
- **Wi-Fi & IoT convergence:** The emergence of IoT applications in campuses increases the procurement, deployment, and O&M demands of enterprises for IoT base stations. Given IoT base stations and APs work in the similar way, plus, WLAN is the preferred access mode in the fully-wireless era, APs are widely used in enterprise office and production areas. Therefore, IoT application expansion based on WLAN APs becomes a better choice. Currently, IoT applications can be implemented through methods such as PCIe cards, USB dongle, and built-in Bluetooth. This solution greatly reduces the network construction and O&M costs.
- **Convergence of the production and office networks:** The campus network usually provides multiple services, such as office, video conferencing, security, production, and IoT. The dedicated network construction mode requires high costs, large equipment room space, and massive integrated cabling. However, the network utilization rate is generally lower than 5%. With the introduction of VXLAN into campus networks, converged bearing of multiple services becomes possible. In this way, one physical network can carry multiple services at the same time. In addition, technologies such as application identification and network slicing are used to provide differentiated priority scheduling for different services, ensuring service quality and user experience.
- **Application scenarios:** IoT application and multi-service converged bearing can be widely deployed and applied in industry scenarios such as education, healthcare, retail, and large enterprises.

Converged IoT and Wi-Fi networking



WLAN APs as access points to share wired backhaul resources, reducing network construction costs by over 30%

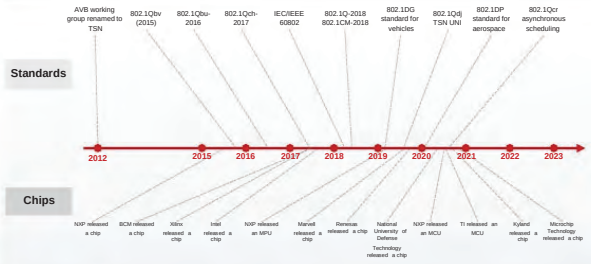
Converged production and office network

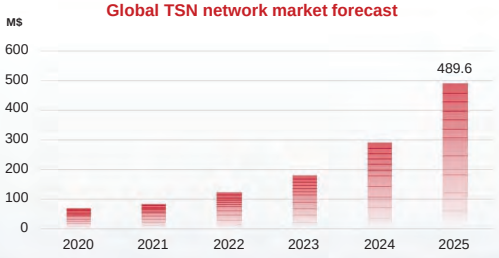


One physical network for multiple services, securely isolating services and reducing network construction costs by over 50%

Non-Deterministic Latency -> Deterministic Latency, Achieving IT/OT Convergence via TSN

- **Rapid development of TSN:** Currently, Industrial bus and industrial Ethernet are the mainstream connection modes in the manufacturing industry. Although traditional IP/Ethernet networks have advantages such as good openness, good interoperability, mature industry, high bandwidth, and low costs, they can only provide best-effort services based on statistical multiplexing, which fail to ensure deterministic low latency required by the industry. The scenario-specific industrial Ethernet network uses specific methods to achieve bounded latency, but has poor interoperability and scalability. Worse yet, the usage of dedicated software and hardware leads to higher costs. TSN combines the advantages of both traditional and industrial Ethernets. It provides users with a network infrastructure featuring low costs, high bandwidth, and statistical multiplexing, solving the problem of difficult interworking between various buses and industrial Ethernet protocols. In addition, it boasts performance highlights such as bounded latency, ultra-low latency, automated network configuration, and high reliability.
- **Accelerated commercial use of TSN:** With the help of time synchronization and precise scheduling, TSN uses the periodic network transmission mechanism to ensure microsecond-level deterministic latency for services. This meets the transmission demands of time-sensitive services in production, manufacturing, and transportation scenarios. Currently, the technical standards (bounded latency, resource management, time synchronization, and high reliability) of TSN have been released. Alongside many chip vendors have launched chips that meet these standards, more than 10 device vendors, including Huawei, have released products and solutions, through which the interoperability between different vendors has been fully verified. In this context, TSN has been put into commercial use in North America and China. In addition, with obvious advantages in bandwidth, latency, and reliability, Wi-Fi 7 can provide flexibility and scalability for TSN. The combination of Wi-Fi 7 (wireless) and TSN (wired) provides more possibilities for applications such as industrial automation and robots, as well as accelerating their commercial use.

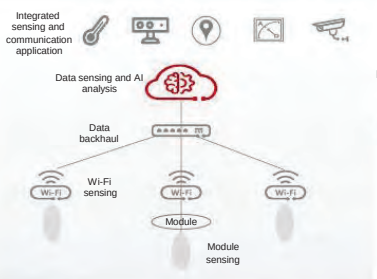




Network Communication -> Integrated Sensing and Communication, Building Networks Featuring All-Round Sensing

- **Wi-Fi devices develop towards intelligent sensing:** Wi-Fi can be used not only for communication, but also for sensing. Wi-Fi sensing uses Wi-Fi waves for motion and presence detection, and then applies machine learning algorithms to promote advanced applications. Wireless devices can be converted into sensors to perform high-precision body positioning and action recognition in a wireless manner, convert the recognition result into instructions, and backhaul them to the control system in real time. The control system then calculates the interference and reflection of signals in the physical space of people and objects, and collects data about them. Wi-Fi devices will participate in the network interaction for determining the locations of people and objects in a specific area.
- **Progress in Wi-Fi sensing standards:** In September 2020, Task Group IEEE 802.11bf was formed for standard research of integrated sensing and communication, which was used for sensing instead of data communication. Standards draft 1.0 and 2.0 were released in January and July of 2023, respectively. In addition, draft 3.0 is planned to be released in November 2023 and draft 4.0 as well as initial products and solutions will be released in January 2024. Besides, formal standards will be released in 2025.
- **Expanded application scenarios:** Integrated sensing and communication improves the performance and efficiency of wireless systems. This feature has the potential to be applied to more scenarios, such as high-precision physiological fall detection for health monitoring, and presence detection for energy saving and carbon reduction.

Sensing-oriented Wi-Fi networks



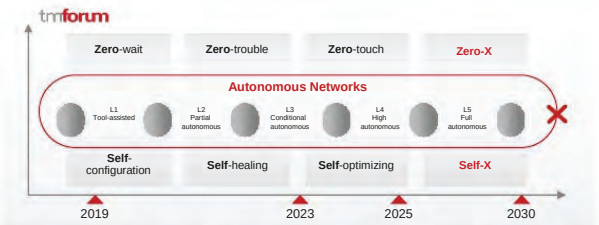
Innovation in extensive application scenarios

Contents

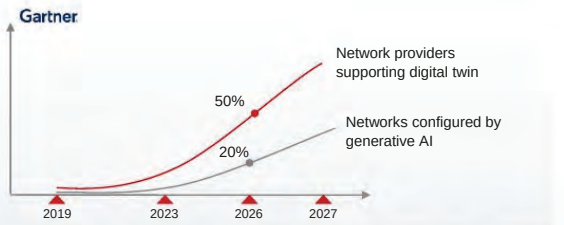
The Rapid Development of AI Marks a Watershed Moment for the Large-Scale Deployment of Network Intelligence

- **Definition and development of network intelligence:** Network intelligence refers to quickly detecting and isolating problems through real-time data collection, prediction, and association, eliminating the need for network personnel to learn advanced network configuration and troubleshooting skills. AI plays an increasingly important role in coping with the increasing network complexity and demonstrates tremendous potential in transforming traditional network O&M, significantly improving productivity. In 2019, TM Forum proposed the concept of Autonomous Networks (AN), defined five AN levels (L1 to L5), and set the goal of achieving "full Autonomous Networks." Most communications networks are currently somewhere between L2 and L3.
- **Challenges to network intelligence:** Valid data is a crucial ingredient to network intelligence. Traditional NEs do not have advanced analysis capabilities, making it difficult for O&M engineers to identify valuable suggestions from a large number of logs and alarms. And even experienced network engineers could not provide accurate intents as the input. Even if reliable network suggestions are offered, O&M engineers have doubts about implementing them due to a lack of methods for visualizing the network in a comprehensive manner. Simply put, O&M engineers do not fully trust AI before they see the actual data and results. So far, most AI network applications are developed to solve specific problems and are difficult to deploy at the system level. The application rate of AI on the entire network is less than 10%.
- **Rapid development of network intelligence**: In recent years, a rapid increase in computing power has led to the birth of the digital twin and the large-scale application of various foundation models, significantly improving data validity and the effect of visualizing the impact on services. This will facilitate the rapid deployment of network intelligence. According to Gartner, by 2026, 50% of network providers will incorporate the digital twin into their solutions, and 20% of initial network configurations will be completed by generative AI; by 2027, the proportion of enterprises that implement AI automated network O&M will increase from 10% to 90%. The rapid development of AI will promote the large-scale deployment of network intelligence.

Definition and timeline of L5 AN

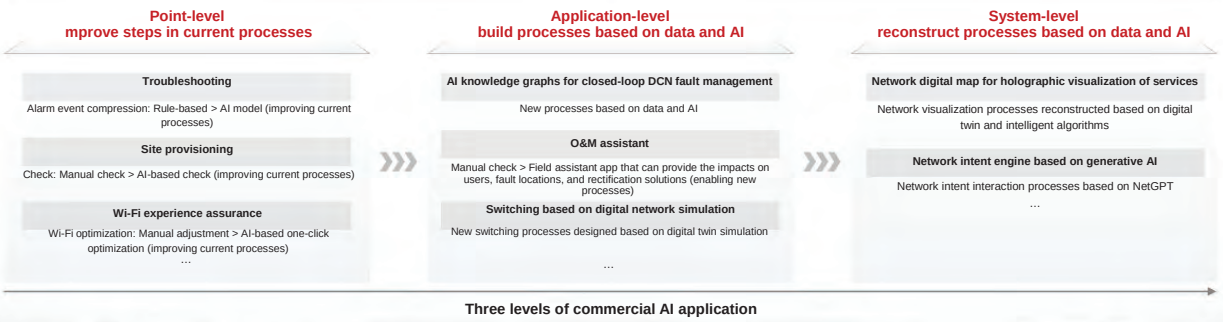


AI network innovation trends



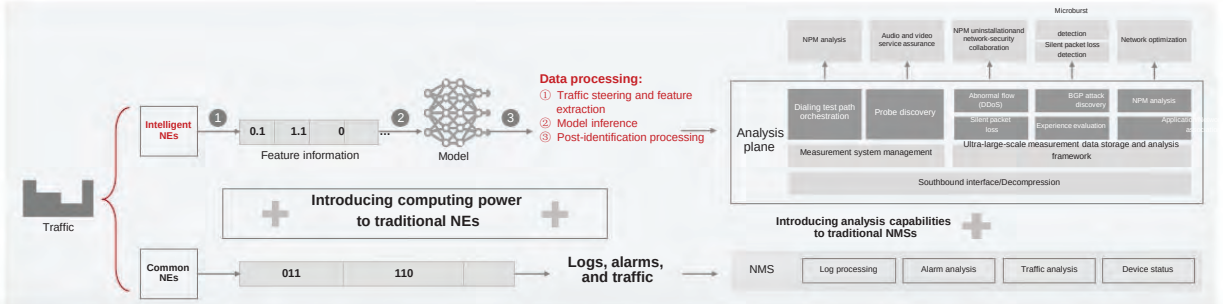
System-Level Solutions for Datacom Network Intelligence Relies on Digital Twin and Generative AI

- **System-level network intelligence is key to large-scale deployment:** AI is unable to be deployed overnight, no matter which industry. Instead, it is usually deployed at three levels. The first level is point level, where AI is used to solve specific problems and improve current processes. This type of AI can be deployed independently without changing the system. Examples include alarm compression, Wi-Fi experience assurance, and automatic site provisioning. The second level is application level, where is used to solve many different problems and enable new processes that can be deployed independently without changing the system. Examples include automatic problem identification, locating, and solving through knowledge graphs on DCNs. This type of AI can adapt to some scenarios but cannot reconstruct the entire system. The third level is system level, where AI can improve multiple processes at the same time or enable new processes by transforming interdependent processes. In the datacom network domain, digital twin and generative AI are cutting-edge technologies that can reconstruct various processes. These technologies have been deployed in system-level solutions to support the large-scale deployment of network intelligence.
- **System-level solution 1:** digital twin. The network digital map based on digital twin technology was first deployed on campus DCNs in 2018 to visualize the network and rectify some of the network faults. Now it has been deployed on a large scale to leverage the network digital twin as a basic O&M platform, lowering the trial-and-error cost, accelerating innovation and iteration, and making network O&M more intelligent.
- **System-level solution 2:** generative AI. As ChatGPT made generative AI more popular at the end of 2022, network models started to develop rapidly. Generative AI will be further applied on networks to create detailed configurations and troubleshooting procedures based on manual inputs without any templates. The objective is to provide the key capability of converting service intents into network requirements and develop a service intent engine.



From Receiving to Generating Intents, Intelligent NEs Lay a Solid Foundation for Network Intelligence

- **Intelligent NEs lays a solid foundation for network intelligence.** Network data, which is mainly generated by NEs, is the key to network intelligence. If network data contains only logs and alarms, it is difficult for O&M personnel to generate accurate network intents in most cases. The lack of accurate network intents results in passive network O&M. An intelligent network, which has become mainstream, is impossible without intelligent NEs. On an intelligent network, NEs will transform from receiving intents to generating intents, perform self-analysis, and collect, preprocess, and report multidimensional data. Based on traffic changes, the network generates intents that humans alone cannot provide, offering a solid data foundation for AI.
- **Crucial benefits of NE intelligence:** Intelligent NEs can use deep learning models to classify data flows based on packet traffic behavior features and take actions on the flows based on the inference result, improving key device capabilities, including service assurance and security detection. Take a financial backbone network as an example. Abnormal application traffic, which may affect high-value services and cause service faults, can be detected and reported by NEs alone. As such, network devices are upgraded to intelligent ones to identify abnormal traffic within seconds and perform closed-loop traffic limiting and steering, ensuring the SLA performance of high-value services. In terms of DDoS attack defense, the second-level DDoS attack detection mechanism of routers depends only on the change of the flow rate, and is unable to identify the types of attacks or accurately report them, including multi-to-one switching caused by link faults, whereas the device protocols are vulnerable to attacks. To address these issues, intelligent devices are used to identify the types of attacks from pass-by traffic, report attacks within seconds, and clean device traffic quickly for closed-loop management.



Large-Scale Using of Network Digital Map, Extending from Multidimensional Visualization to Optimization Simulation

- **Full-scale application of the network digital map:** As the digital mirrors of physical network infrastructures, digital twins maintain almost the same topologies, services, and traffic data models as the physical network does. As refined full-lifecycle, multidimensional duplicates of the physical network, the digital twins provide a digital verification environment for network O&M. Unlike traditional simulation technologies, the network digital map is not a static snapshot of the physical network; rather, it is updated in real time based on the physical network status. With the help of AI technologies and self-learning, the network digital map will evolve by itself based on online pre-verification feedback and provide higher authenticity and reliability.
- **Technical trend 1: From one-dimensional to multidimensional visualization with effective service-network association.** Traditional network management systems (NMSs) visualize network quality from a single dimension only. However, in network O&M scenarios, it's quite often the case that network faults cannot be located although service faults did occur, or the network cannot quickly prove its own innocence. For example, a bank was struggling with low O&M efficiency because there was no visualized association between its hundreds of applications and tens of thousands of NEs. As such, achieving visualization at the network layer alone is far from enough — the network digital twin engine also needs to build a digital duplicate of the unified physical network model. In terms of O&M, multi-domain and multidimensional simulation and verification must be implemented to support network planning, construction, maintenance, and optimization activities, including displaying intra-/inter-application access relationships and detecting service exceptions in seconds.
- **Technical trend 2: From offline simulation to real-time simulation, reducing risks associated with network changes.** By scenario, network simulation technologies fall into two categories. The first category is offline simulation, which was prevalent in network planning in the past. However, as the network automation level continues to increase, the closure periods of network faults, self-service application, and other services are shortened (to minutes or even seconds as expected by customers). This drives the emergence and prevalence of the second category — real-time simulation. Based on NE configuration data, real-time simulation technology simulates the control plane and forwarding plane behaviors of devices' routing protocols and accurately generates NE protocol-specific routing tables and the global routing table. By analyzing entries in these routing tables, real-time simulation technology is able to learn and verify impact on the network.

Digital Transit Map and Navigation

Network Digital Map

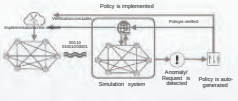
- Holographic network visualization:Network-wide SLAs are visualized.
- Optimal network performance:Network latency is reduced by 30% (tested by P3).
- Autonomous network optimization:The path optimizationtime is shortened from 3 months to 3 minutes.

Multidimensional Visualization

- One map visualizing the entire network
- Real-time topology restoration
- Mutual visibility between applications and networks

Real-Time Network Simulation

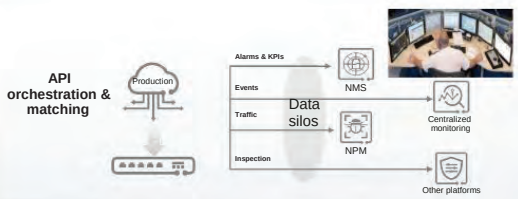
- One-click path navigation
- Simulation of multi-cloud changes
- Holographic proactive awareness



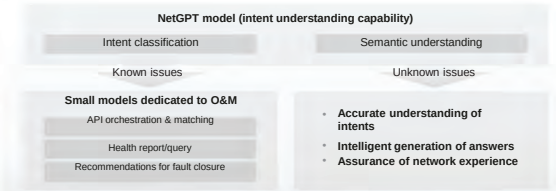
Generative AI Creates Opportunities for NetGPT

- **Generative AI drives the development of NetGPT.** The personalized development of hundreds of models and thousands of modalities creates unparalleled opportunities for the emergence of large communications network models. Large models at the application layer must be connected to networks so that they can reach end users, but communications networks feature a "thin waist of the hourglass" and therefore expect one unified large communications network model (NetGPT) to process different network services at lower service complexity. This is the trend. Currently, NetGPT is still in its infancy, and its application scenarios are still under preliminary exploration — several known scenarios are knowledge management platform enablement and interactive network intent engine processing. Looking ahead, NetGPT may be deployed in a cloud-edge collaboration manner with the aim to effectively orchestrate heterogeneous distributed communications and computing resources, and this is a key step for NetGPT to leverage its full potential.
- **Scenario 1: Knowledge management platforms significantly increase the accuracy of interactive Q&A.** Traditionally, AI Q&A assistants are built on knowledge graph technology. If the keywords of a user's question are inaccurate, assistants tend to give inaccurate answers. To address this issue, the NetGPT model has been constructed using abundant datacom network corpus, and is retrained for knowledge expansion and adjusted based on tasks. This enables the model to accurately understand user intents, improve interaction efficiency, and effectively collaborate with dedicated small models in intelligent Q&A and network experience assurance scenarios. The accuracy of feature-related answers increases from 20% to 80%.
- **Scenario 2: The manual intent engine fully understands network intents.** Although we had NMSs in the past, there were too many network O&M tools and GUIs. When handling issues, we did not know where to start. Moreover, application/service flow data was not associated with network data, and only experienced experts could identify the association between them. Due to the ability to accurately identify, distinguish, and understand user intents, the NetGPT model can convert user intents into network requirements. Based on the input of maintenance personnel, it creates detailed configurations and troubleshooting procedures without any fixed templates, improving O&M efficiency and reducing professional dependency for O&M teams.

Before: Complicated troubleshooting, difficult to get started



After: Intelligently identifying intents and recommending closed-loop measures

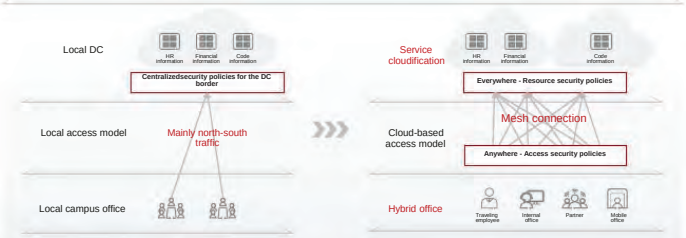


Contents

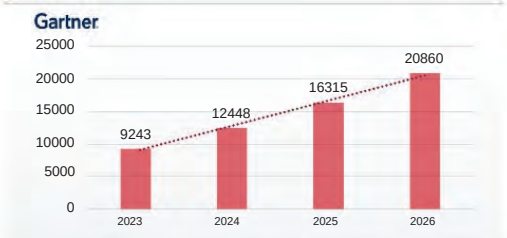
Cloudification and Hybrid Office Break Security Boundaries, and Network-Security Collaborative Defense Becomes the Mainstream Choice

- **Enterprise cloudification breaks the defense boundary:** The traditional network security architecture focuses on the enterprise intranet and establishes a layered defense system at the enterprise boundary to safeguard data security. With more and more enterprises embracing cloud migration, the traditional closed architecture has evolved into a multi-cloud and multi-branch interconnected system. This has led to the breakdown of enterprise boundaries and increased exposure to network security risks. Moreover, there is a growing need for high reliability and security in production. The original centralized service access and security system have become ineffective and cumbersome. Traditional network security technologies are unable to address the increasingly sophisticated threats and vulnerabilities faced by the network periphery. As external access to the cloud accelerates, enterprises must implement advanced access control measures to ensure they have the capability to handle related network security needs and risks.
- **Hybrid office increases security risks:** Enterprise employees are not confined to fixed workplaces, and hybrid office has become a common practice. This means that employees may access the enterprise network through the insecure Internet anytime and anywhere. The original security architecture based on the enterprise LAN border is no longer effective, which brings new challenges to enterprise data security defense. Enterprises need to consider how to ensure secure access to the enterprise headquarters and multi-cloud platforms anytime and anywhere.
- **Network-security convergence represented by SASE has become a trend:** To address these changes, the adoption of zero-trust-based secure access service edge (SASE) has emerged as a popular trend, ushering in a new era of service evolution for branch networks and security convergence. SASE offers a range of converged network and security-as-a-service functions from a single cloud delivery platform, including zero-trust network access, cloud access security proxy, secure web gateway, firewall, and SD-WAN. This enables secure and seamless connections for any application across any network, location, or device. Gartner predicts that by 2026, 80% of enterprises will implement SASE solutions for architecture and networking reconstruction, with a market size of US$21 billion.

Impact of enterprise cloudification and hybrid office on the network security architecture



SASE end user expenditure forecast (US$)



Ransomware Attacks Become the Norm, Making It Crucial to Establish a Comprehensive Defense System to Create a Robust Line of Defense

- **Ransomware attacks become the norm:** In recent years, ransomware attacks have emerged one after another and have severely affected sectors such as public service, finance, education, healthcare, manufacturing, and energy around the world. In some incidents, attackers hijack critical infrastructure to claim ransom, which may even affect the normal operation of a country. So far, the average service interruption caused by ransomware has reached 16 days. An organization is attacked by ransomware every 11 seconds, and the largest ransom is up to US$70 million. Large companies with large and complex digital infrastructure have become one of the main targets of ransomware cyber criminals. According to an IDC report, 35% of global organizations have experienced 3 to 4 ransomware incidents. A successful ransomware attack requires an average ransom of about US$150,000, causing service interruption for five days on average.
- **The protection system moves towards defense-in-depth.** Ransomware attack tactics and variants are evolving. Traditional data backup, network border protection devices, and traditional antivirus software that relies on signature detection have become invalid. The number of ransomware variants increased exponentially, from 5400 in 2021 H2 to 10666 in 2022 H1, an increase of 98%. The encryption speed and permission theft speed of ransomware are very fast. The time window for administrators to handle ransomware is very short. The fastest time for ransomware to infiltrate the system to obtain permissions is 45 minutes, while the average encryption speed of 100,000 files is only 43 minutes. In addition, latest-generation ransomware attacks target backup systems, devices, and VMs. As a result, more than 46% of organizations that pay ransom after being attacked cannot fully restore data. New service changes and frequent new threats make security protection more professional and complex. Also, more intelligent security protection methods are required. Deploying multiple types of security products has been evolved to trusted infrastructure networks. The construction of an in-depth defense system has become a hot topic for enterprise investment.

Continuously evolving of ransomware threats: rapid rise of variants, long service interruption, and frequent attacks

57x

In 2021, ransomware causes a loss of US$20 billion, which is 57 times the loss in 2015. The loss is predicted to be US$265 billion in 2031.

Source from: Cybersecurity Ventures

16 days

Ransomware attacks result in an average of 16 business days of system shutdowns.

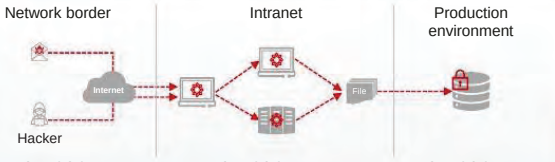
Source from: ZDNet

11s

In 2021, a ransomware attack took place every 11 seconds. By 2031, the interval is predicted to be just 2 seconds.

Source from: Cybersecurity Ventures

Building an in-depth defense system for enterprises from the perspective of ransomware attacks



Line of defense 1
Network border intrusion prevention

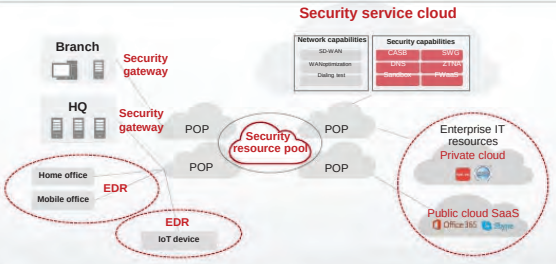
Line of defense 2
Non-proliferation within the network

Line of defense 3
Anti-encryption in the production environment

From Network-Security Separation to Network-Security Integration, the Converged Architecture Comprehensively Improves the Overall Security Posture of Enterprises

- **Converged network-security architecture:** For enterprises, the primary network requirements include branch Internet access, interconnectivity between branches and the headquarters, and SaaS service access from branches. With the increasing cloud migration of enterprises and the growing use of mobile office by employees, a significant number of users, devices, applications, and data are now located outside enterprise DCs and networks. To address this, a converged network-security architecture deploys network and security protection capabilities on corresponding network nodes, implements flexible and distributed overlay networks through software definition, applies security protection capabilities to nearby entities, and provides unified policies and security posture awareness through collaboration with the operations brain. This architecture meets the network security interconnection requirements in various enterprise scenarios.
- **Advantages of network-security convergence technologies:** Compared to traditional network security architecture, network-security convergence represented by SASE has four key features: zero trust access, cloud-native architecture, support for all edges, and global distribution. These features enable SASE to better meet the increasing demands of enterprises for cloud application services and cloud network security products. Furthermore, these features reflect the integrated deployment of network and security. SASE offers comprehensive network and security services, establishing flexible overlay networking capabilities for branches to access the Internet, headquarters, and the cloud. It also establishes an end-to-end management mechanism and security protection measures based on overlay connections, eliminating physical network restrictions and simplifying complex networks. With its identity-centric approach, SASE provides ubiquitous defense capabilities. The centralized operation service simplifies policy management and security incident handling, providing customers with a simple, efficient, secure, and stable network access and service deployment experience.

SASE-based converged network-security architecture and user benefits



- **Reduced complexity and costs:** A single service provider reduces the number of physical or virtual devices at the branch border and the number of agents.
- **Enhanced performance/latency:** The SASE supplier provides POPs around the world to optimize the access latency and route selection.
- **Low OPEX:** Enterprises are no longer affected by hardware capacity expansion and EOL device updates. In addition, enterprises can quickly defend against new threats without paying attention to signature database updates.
- **Zero trust:** Multiple threat signals and context signals are used to ensure secure access to internal resources and the Internet.
- **Improved efficiency of network and network security personnel:** Build enterprise security strategies on a single platform.

Umetna inteligenca v mobilnih omrežjih

Artificial Intelligence in Mobile Networks

Srđan Knežević

Ericsson

POVZETEK

Osrednje področje prispevka je umetna inteligenca (UI) v mobilnih telekomunikacijah. Prispevek poskuša dati pregled nad UI v telekomunikacijah z odgovori na naslednja vprašanja:

- Kaj je UI v telekomunikacijah?
- Zakaj je UI uporabna v telekomunikacijah?
- Kateri so ključni elementi UI v telekomunikacijah?

Poleg tega prispevek poskušala podrobneje opisati nekatere od ključnih aplikacijskih področij v telekomunikacijah, kjer se lahko UI izkoristi, s poudarkom na naslednjih področjih:

- Radijsko dostopovno omrežje (RAN)
- Upravljanje omrežja
- Onkraj omrežja

SUMMARY

The main scope of the presentation is on mobile telecom artificial intelligence (AI). The presentation gives an overview of telecom AI by answering the following questions:

- *What is telecom AI?*
- *Why is AI useful in telecommunications?*
- *What are the critical elements of telecom AI?*

Further, the presentation gives a close look at some key application areas in telecom where AI can be leveraged by focusing on the following:

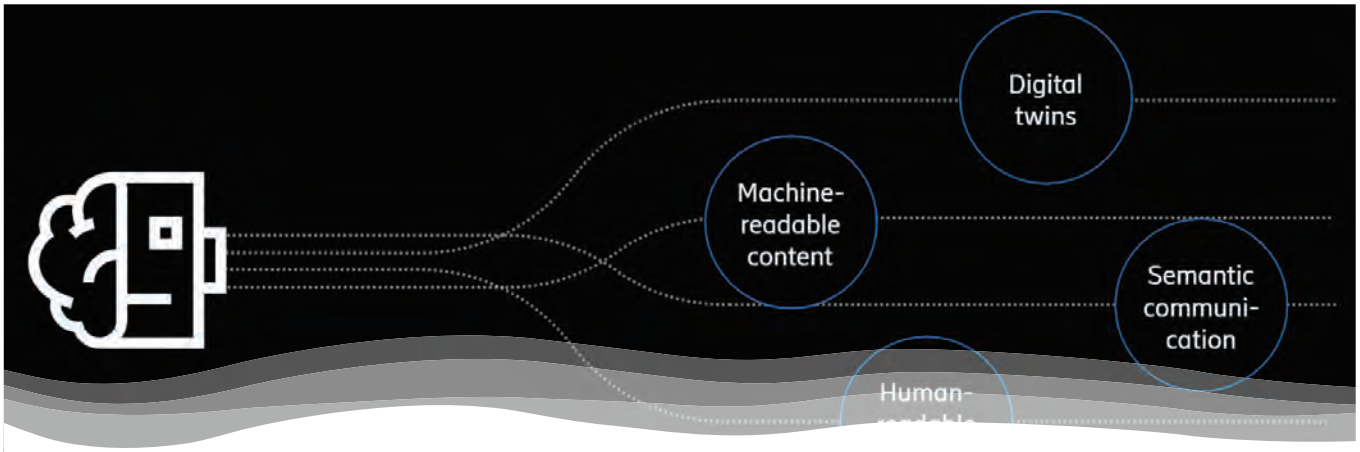
- *Radio Access Network (RAN),*
- *Network Management,*
- *Beyond network.*

O AVTORJU

Srđan Knežević je v podjetju Ericsson srkrbnik ključnih strank, s strokovnim znanjem na področju radijskih dostopovnih omrežij. S skoraj 20-letnimi izkušnjami na področju mobilnih telekomunikacij je sodeloval pri številnih 3G, 4G in 5G projektih in predstavitvah, ki pokrivajo slovenski trg.

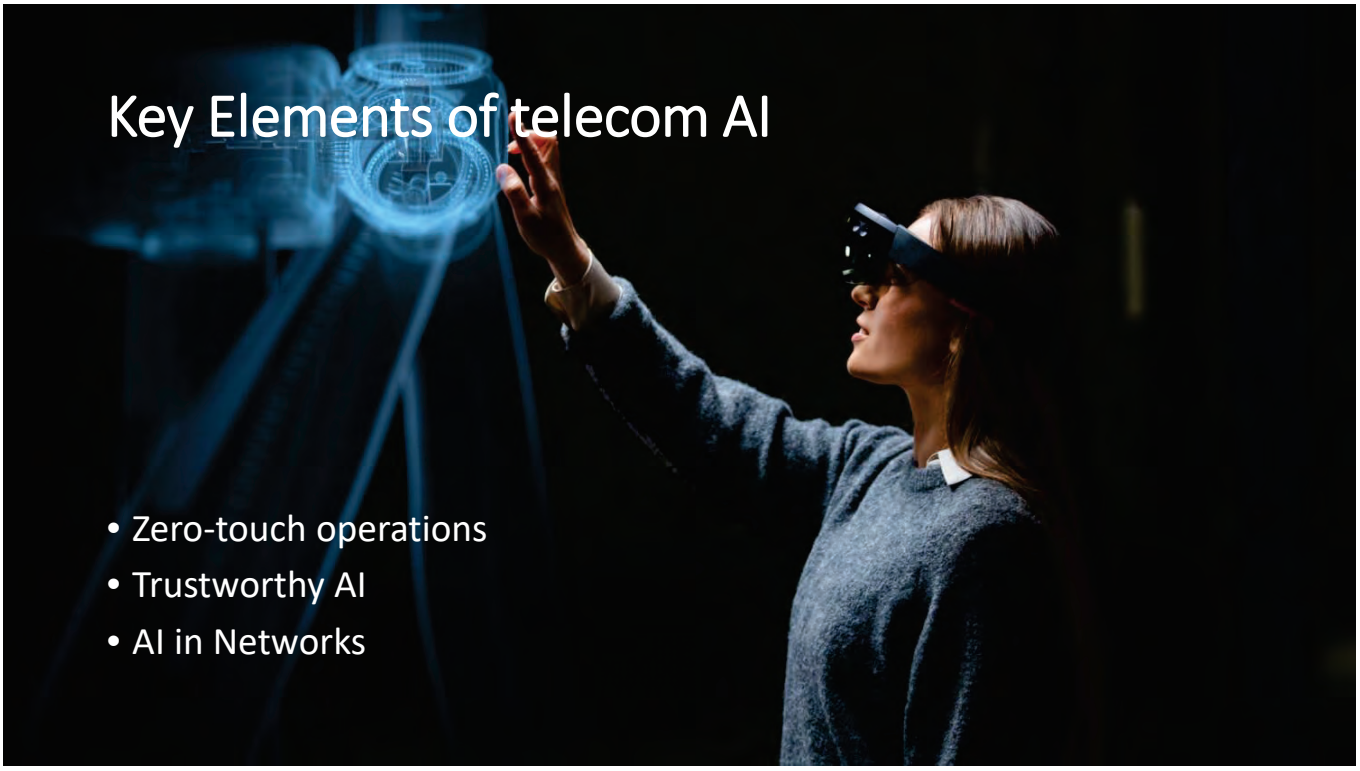
ABOUT THE AUTHOR

Srđan Knežević is an Ericsson Account Manager with expertise in Radio Access Networks. With almost 20 years of experience in mobile telecommunications, he has participated in numerous 3G, 4G, and 5G projects and presentations covering the Slovenian market.



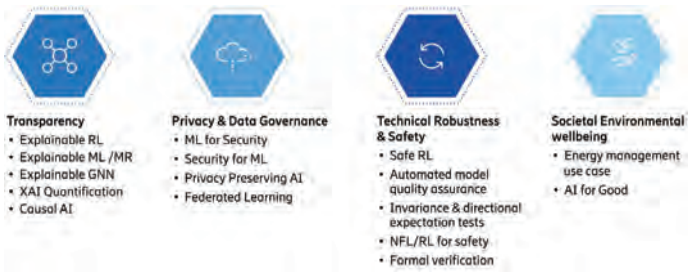
How will AI transform telecom

- Human-readable content
- Machine-readable content
- Semantic communication
- Digital twins made simple



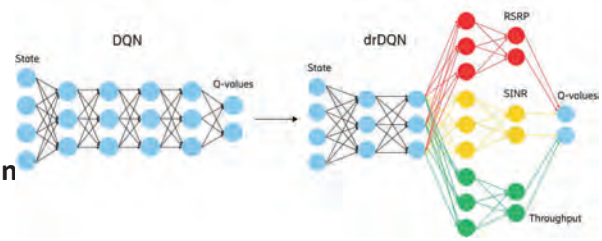
Trustworthy AI - What it means for telecom

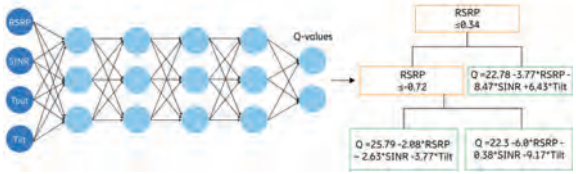
- **Human agency and oversight**
- **Transparency**
- **Privacy and data governance**
- **Diversity, non-discrimination and fairness**
- **Technical robustness and safety**
- **Social and environmental well being**



Transparency - What it means for telecom

- **Different Explainable AI (XAI) methods and techn**
- **Explainability of Machine Reasoning (MR)**
- **Explainability of Reinforcement Learning (RL)**





Applying generative AI in mobile networks

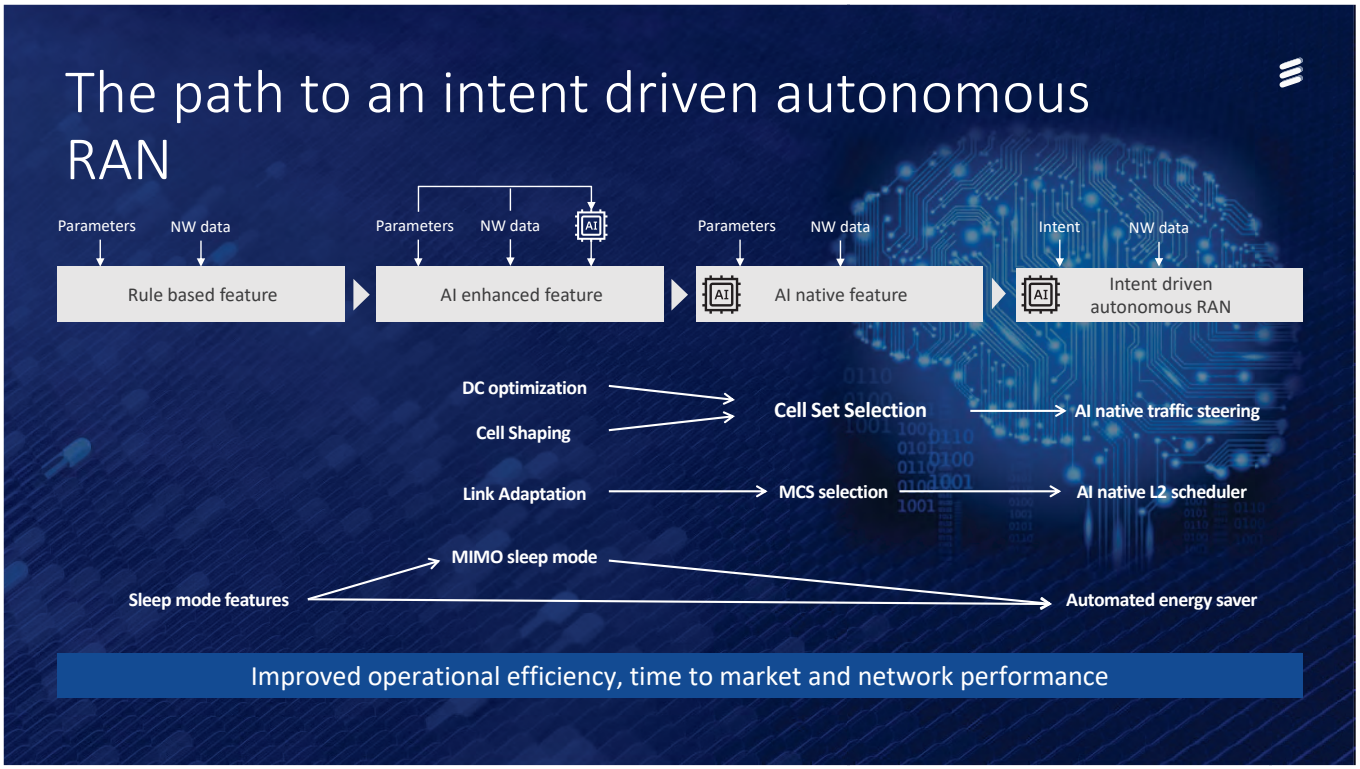
- **Radio Access Network (RAN)**
  - Wireless channel modeling
  - Spectrum sensing
  - Channel quality estimation
  - Hybrid beamforming
  - Network traffic generation
  - Network traffic analysis
  - Anomaly detection
  - Network selection
- Network Management
- Beyond Networks

Applying generative AI in mobile networks

- Radio Access Network (RAN)
- **Network Management**
  - Customer incident management
  - Network planning
  - Deployment and configuration
  - Network operations support
  - Fault diagnosis
  - Resource utilization and allocation
  - Network security
  - Threat detection
- Beyond Networks

Applying generative AI in mobile networks

- Radio Access Network (RAN)
- Network Management
- **Beyond Networks**
  - Digital twins
  - Business documentation creation
  - Generation from text-based documentation
  - Intuitive dialogue-based interfaces
  - Data generation for application support
  - Software implementation support
  - Learning services and reference chatbots
  - XR, UAVs, autonomous vehicles, remote surgery…

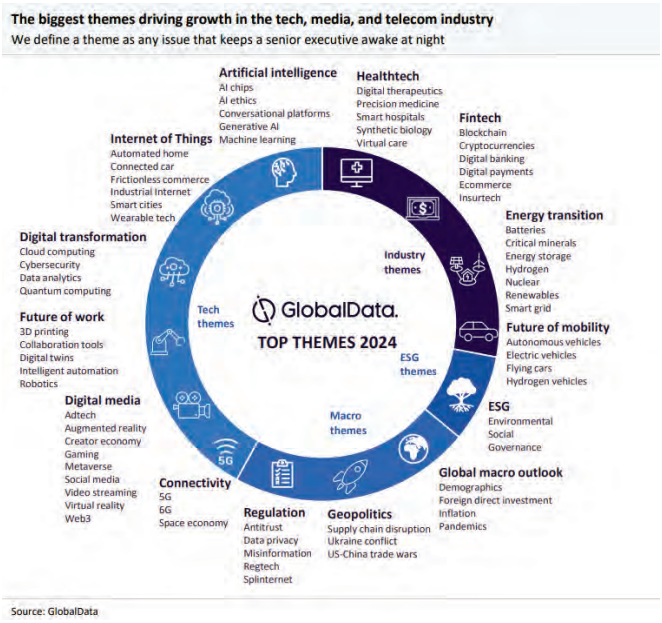


Biggest themes in the tech, media and telecom industry

Old tech and telecom topics running in a circle:

- Digital transformation
- Connectivity
- Internet of Things

… together with Regulations, Geopolitics and Global macro outlook



PRISPEVKI

ARTICLES

17. 5. 2024

Kako izkoristiti moč podatkov in umetne inteligence za ustvarjanje dodane vrednosti

How to harness the power of data and artificial intelligence to create added value

Ajda Brlec

Telekom Slovenije

POVZETEK

V Človeštvo je bilo v zadnjih stoletjih priča številnim tehnološkim revolucijam, od katerih je vsaka oblikovala tok zgodovine. V zadnjih nekaj letih je prišlo do izjemne preobrazbe digitalne pokrajine. Vstopamo v dobo, v kateri naj bi umetna inteligenca spremenila način našega dela, uporabe podatkov in povečala produktivnost. Pogosto zastavljeno vprašanje je: "Kaj lahko storimo z AI?" Vse se začne z jasno vizijo in pravimi podatki za podporo poslovnih poizvedb. Če želimo ustvariti vrednost za svoje podjetje, moramo zagotoviti, da naše delo z umetno inteligenco (AI) postane strateški proces – in ne le kup naključnih pobud. Razmišljajmo veliko, začnimo z majhnim in se hitro razširimo na preostala področja.

Štirje mejniki za ustvarjanje dodane vrednosti z umetno inteligenco:

1. Razumevanje tehnologij - Da bi izkoristili moč umetne inteligence za poslovni uspeh, je ključno najprej razumeti različne tehnologije, ki spadajo pod njeno okrilje. Umetna inteligenca ni univerzalni koncept, ki bi ustrezal vsem primerom enako. V prezentaciji bomo spoznali kateri tipi AI obstajajo in čemu so namenjeni.
2. Investicije v kakovost podatkov - Preden se potopimo v svet AI, mora podjetje imeti urejene osnove. Dandanes so le nekatera podjetja v celoti zaključila svojo digitalno preobrazbo in imajo vzpostavljeno zanesljivo podatkovno strategijo. Kakovost podatkov določa natančnost rezultatov, ki jih ustvarijo modeli AI. Vlaganje v pravo kakovost podatkov je zelo pomembno za zagotovitev razpoložljivosti, zanesljivosti in varnosti podatkov za AI.
3. Prepoznavanje priložnosti - Ob uvajanju AI, je potrebno prepoznati priložnosti v podjetju, na katere lahko umetna inteligenca pomembno vpliva. Ali obstaja postopek, ki bi ga lahko avtomatizirali ali optimizirali? Ali obstajajo vpogledi v stranke, ki bi jih lahko pridobili iz analize podatkov? Ko prepoznate te priložnosti, je čas, da določite posebne primere uporabe za AI. Kako lahko reši problem ali izboljša učinkovitost v vsakem scenariju? V prezentaciji bomo spoznali konkretne izzive, ki jih lahko rešuje AI: prodaja, poprodaja, finance, trženje, nabava itd.
4. Razširitev uporabe AI - Ko uspešno zaženemo pilotne projekte AI in vidimo pozitivne rezultate, je čas, da popeljemo svoje pobude AI na višjo raven – razširitev. Tukaj lahko resnično izkoristimo moč umetne inteligence za poslovni uspeh. Povečanje vključuje razširitev implementacije tehnologij umetne inteligence v različne oddelke ali funkcije v organizaciji. Zahteva pa skrbno načrtovanje, koordinacijo in sodelovanje med različnimi ekipami.

SUMMARY

In Humanity has witnessed multiple technological revolutions throughout the past centuries, each shaping the course of history. There has been a remarkable transformation of the digital landscape in the last few years. We are now entering an era where Artificial Intelligence is set to revolutionize the way we work, utilize data, and boost productivity. The question often asked is, "What can we do with AI?" It all starts with having a clear vision and the right data to support business inquiries. By aligning AI investments with business strategy, it has the power to drive your company forward. But to create value for the company, you have to take control and ensure that your work with AI becomes a strategic process - and not just a stack of random initiatives. Think big, start small, and scale fast.

Four milestones to create added value with artificial infteliigence:

1. *Understanding the Technologies - To harness the power of AI for business success, it is crucial first to understand the different technologies that fall under its umbrella. Artificial Intelligence is not a one-size-fits-all concept. In the presentation, we will learn what types of AI exist and what they are intended for.*

2. *Invest in data quality - Before diving into AI, the fundamentals must be there. Nowadays, only some businesses have fully completed their digital transformation exercise and have a robust data strategy in place. The quality of data determines the accuracy of results generated by AI models. Investing in the right data quality is super important to ensure the availability, reliability, and security of data for AI.*
3. *Identifying the Opportunities - You must identify the opportunities within your business that AI can significantly impact. Is there a process that could be automated or optimized? Are there customer insights that could be gleaned from data analysis? Once you've identified these opportunities, it's time to determine the specific use cases for AI. How can it solve the problem or improve efficiency in each scenario? In the presentation, we will learn about concrete challenges that AI can solve: sales, after sales, finance, marketing, procurement, etc.*
4. *Scaling Up - Once you have successfully launched your AI projects and seen positive results, it's time to take your AI initiatives to the next level — scaling up. This is where you can truly harness the power of artificial intelligence for business success. Scaling up involves expanding the implementation of AI technologies across different departments or functions within your organization. It requires careful planning, coordination, and collaboration between various teams.*

O AVTORJU

Ajda Brlec, diplomirana matematičarka, je direktorica Digitalizacije v Telekomu Slovenije. Odgovorna je za razvoj naprednih rešitev, ki podpirajo digitalizacijo in poenostavljajo življenja, kot so rešitve e-zdravja, e-commerce in naprednih plačilnih storitev. Storitve, ki jih prinaša digitalizacija, izboljšujejo kakovost javnih storitev ter imajo pomembno vlogo pri ustvarjanju trajnostne družbe in razvoju tehnoloških inovacij. Svojo kariero je kot analitik pričela v družbi TSmedia, nato je bila vodja zastopniške mreže in teleprodaje ter direktorica marketinga in prodaje. Telekomu Slovenije se je pridružila v letu 2020 kot vodja programskih pravic.

ABOUT THE AUTHOR

Ajda Brlec, a graduate in mathematics, is the Director of Digitization at Telekom Slovenije. She is responsible for the development of advanced solutions that support digitization and simplify lives, such as e-health solutions, e-commerce and advanced payment services. The services brought about by digitization improve the quality of public services and play an important role in creating a sustainable society and developing technological innovations. She started her career as an analyst at TSmedia, then she was the manager of the representative network and telesales and director of marketing and sales. She joined Telekom Slovenije in 2020 as head of program rights.

Kako izkoristiti moč podatkov in umetne inteligence za ustvarjanje dodane vrednosti?

Ajda Brlec, direktorica Digitalizacije, Telekom Slovenije

Vstopamo v dobo, v kateri naj bi umetna inteligenca spremenila način našega dela, uporabe podatkov in povečala produktivnost.

Pogosto zastavljeno vprašanje je: **"Kaj lahko storimo z UI?"**

Vse se začne z jasno vizijo in pravimi podatki za podporo poslovnih poizvedb.

Če želimo ustvariti vrednost za svoje podjetje, moramo zagotoviti, da naše delo z umetno inteligenco (UI) postane strateški proces – in ne le kup naključnih pobud.

Štirje mejniki ustvarjanja dodane vrednosti z umetno inteligenco

3

TelekomSlovenije

Štirje mejniki ustvarjanja dodane vrednosti z umetno inteligenco

4

TelekomSlovenije

Razumevanje tehnologij

- Gre za način, da se računalniki učijo iz podatkov in izkušenj, ne da bi jim bilo treba natančno povedati, kaj naj naredijo.
- Temelji na statističnih tehnikah.

- Globoko učenje je podmnožica strojnega učenja.
- Uporablja umetne nevronske mreže, ki so sestavljene iz več plasti.
- Globoko učenje je sposobno rešiti zelo zapletene probleme

- Generativna umetna inteligenca se uči iz obstoječih podatkov in s tem ustvarjanja novo in izvirno vsebino.
- Lahko ustvari nekaj povsem novega

- Zmožnost računalnika razumeti in poustvariti človeški jezik.
- Usposablja se na ogromnih količinah besedilnih podatkov
- LLM lahko deli ustrezne zgodbe ali ustvarja nove zgodbe na podlagi vseh razpoložljivih informacij.

5

TelekomSlovenije

Štirje mejniki ustvarjanja dodane vrednosti z umetno inteligenco

6

TelekomSlovenije

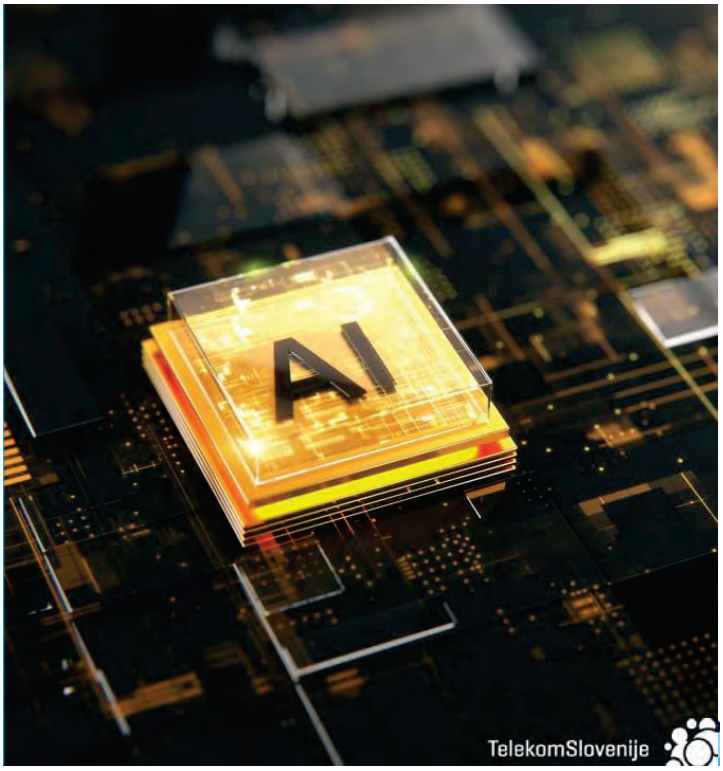
Izkoriščanje podatkov v podjetju: ključ do uspeha v digitalni dobi

Zakaj so podatki tako pomembni?

- Podatki so kot zlato v digitalni dobi.
- Ogromen potencial za razkritje dragocenih uvidov o strankah, trgih, produktih in procesih.
- izboljšava poslovanja, prilagoditve strategije, povečanje konkurenčnosti na trgu.

Kako izkoristiti podatke v podjetju?

- Podjetje želi imeti urejene osnove - vzpostavljena zanesljiva podatkovna strategija.
- Kakovost podatkov določa natančnost rezultatov, ki jih ustvarijo modeli UI.
- Prvi korak je zbiranje in shranjevanje relevantnih informacij ter ustrezna infrastruktura za zbiranje podatkov iz različnih virov.
- Naslednji korak je analiza zbranih podatkov z uporabo naprednih analitičnih metod



7

Štirje mejniki ustvarjanja dodane vrednosti z umetno inteligenco

1. Razumevanje tehnologij

2. Investicije v kakovost podatkov

3. Prepoznavanje priložnosti za UI

4. Podatkovne ekipe

8

TelekomSlovenije

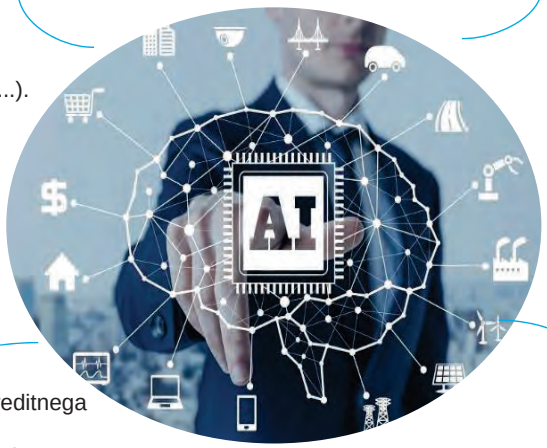
Podatkovni izzivi, ki jih lahko rešujemo z UI

Prodaja

- Optimizacija prodaje z napovednimi modeli.
- Pomoč prodajalcem z avtomatiziranimi odzivi (priporočila paketov, storitev...).
- Avtomatizirani odgovori za stranke.
- Razvrščanje in klasificiranje dohodne pošte in klicev.
- Predvidevanje zamenjave potrošnega materiala pri strankah.

Marketing

- Prilagoditve marketinških strategij: z analizo podatkov o vedenju strank lahko podjetje bolje razume njihove preference, navade in potrebe ter smiselno oblikuje kampanje.
- Zmanjševanje stroškov za ustvarjanje marketinških vsebin.
- Avtomatizacija segmentacije strank in personalizacija trženja.
- Diagnostika izboljšave izdelkov in storitev ter zapolnitev tržnih vrzeli.



Finance

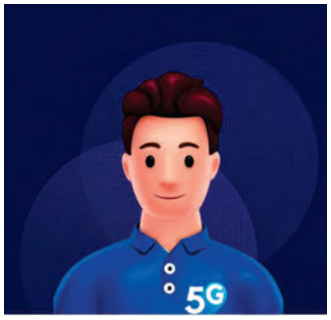
- Avtomatizirano odkrivanje kreditnega tveganja.
- Napovedni model denarnih tokov za optimizirano upravljanje likvidnosti.
- Izboljšanje obravnavanje plačnikov in neplačnikov.

Nabava

- Samodejno spremljanje tveganja dobavitelja na podlagi javnih podatkov.
- Optimizacija obravnave življenjskega cikla upravljanja pogodb.

TelekomSlovenije

Prepoznavanje priložnosti za UI v Telekomu Slovenije



Maks, digitalni svetovalec

Idejo o klepetalnem robotu smo zasnovali in realizirali sodelavci Telekoma Slovenije. Vse skupaj se je začelo že leta 2017, od takrat dalje pa Maks rešuje enostavne in ponavljajoče tehnične zahteve naših uporabnikov. Digitalni svetovalec Maks nadomešča dva svetovalca. Ob večjem izpadu storitev je Maks opravil tudi do 950 stikov oziroma nadomestil 17 svetovalcev.

UI pomočnik za boljša darila

7.000 edinstvenih sporočil in številni predlogi izdelkov iz E-trgovine, ki jih je v prazničnem času ustvaril naš UI-pomočnik. Kar 50.000 uporabnikov je s pomočjo najsodobnejših orodij in umetne inteligence uspeli obogatiti izkušnjo obdarovanja.

TelekomSlovenije

Prepoznavanje priložnosti za UI v Telekomu Slovenije

NEO – UI pri analizi podatkov

NEO uporabniki izgovorijo že več kot 15 milijonov glasovnih ukazov mesečno. UI se uporablja pri analizi podatkov, in sicer za postavitev grupiranja glasovnih ukazov, s čimer se nato išče relevantne vsebinske sklope, ki tem glasovnim ukazom odgovarjajo za boljše vsebinsko iskanje / priporočilni sistem.

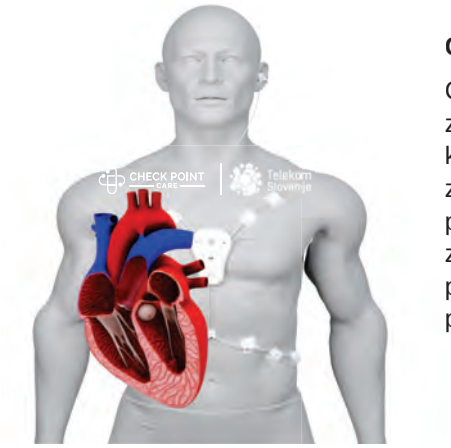
Napovedni model za prodajo

Cilj napovednih modelov je predvsem dvig učinkovitosti in konverzije.

- Identifikacija strank z večjo verjetnostjo za xsell, upsell, churn.
- Kakovostne prodajne adreme.

TelekomSlovenije

Prepoznavanje priložnosti za UI v Telekomu Slovenije



Centralni monitoring

Centralni monitoring pacientov se uporablja za kontinuirano merjenje zdravstvenih parametrov (EKG, krvni tlak, saturacija, v prihodnje tudi krvni sladkor) v bolnišnici ali na pacientovem domu ter intervencije zdravstvenega osebja v primeru odstopanj. V sodelovanju s partnerjem smo razvili storitev, ki omogoča bistveno natančnejše zaznavanje fiziološkega stanja pacienta in hitrejše napovedovanje poslabšanja v primerjavi s klasičnimi metodami (meri 19 vitalnih parametrov, zgodnje alarmiranje je podprto z umetno inteligenco).

TelekomSlovenije

Štirje mejniki ustvarjanja dodane vrednosti z umetno inteligenco

13

TelekomSlovenije

Senzorji in umetna inteligenca za razumevanje uporabnikov in prilagajanje storitev v zdravstvu in socialni oskrbi

Sensors and artificial intelligence to understand users and adapt services in health and social care

Mitja Luštrek

Institut Jožef Stefan

POVZETEK

Zaradi staranja prebivalstva in nezdravega načina življenja se soočamo z vedno večjimi potrebami po zdravstveni in socialni oskrbi. Da bi jo nudili vsem, ki jo potrebujejo, nam manjka tako denarja kot ljudi, zato se za pomoč obračamo k tehnologiji. Ključni pri tem so senzorji in umetna inteligenca: senzorji lahko spremljajo fiziološke signale in vedenje uporabnikov, umetna inteligenca pa senzorske odčitke interpretira in uporabniku ponudi nasvete, prilagojene njegovim trenutnim potrebam.

Ogledali si bomo dva primera. Prvi je aplikacija za obvladovanje srčnega popuščanja, povezana s senzorsko zapestnico. Senzorji v zapestnici spremljajo delovanje srca, prevodnost kože in gibanje. Iz teh podatkov je s pomočjo strojnega učenja mogoče izluščiti uporabnikov krvni tlak in psihološko stanje. Aplikacija potem ponudi uporabniku prilagojen program telovadbe in ga opozori, kdaj je zaradi zdravstvenega stanja bolje ne telovaditi, ter mu svetuje glede prehrane, opozarja na zdravila in še več. Poleg tega uporabnika skuša prepričati, da upošteva nasvete aplikacije, s pomočjo kognitivnega vedenjskega pristopa, ki je prilagojen njegovemu psihološkemu stanju.

Drugi primer je sistem za pomoč starostnikom, ki jim peša spomin, pri hišnih opravilih – osredotočili smo se na kuhinjske. V nasprotju s prvim primerom ne uporablja senzorjev na telesu temveč v prostoru. S pomočjo računalniškega vida prepozna interakcije med uporabnikom in gospodinjskimi pripomočki ter nekatere druge aktivnosti. Nato skuša določiti, katerega opravila se uporabnik loteva in – če zapazi napako – nanjo opozori in predlaga pravilno ravnanje.

Tovrstne rešitve lahko pripomorejo k boljšemu zdravju in samostojnemu življenju, čeprav seveda ne morejo in ne smejo povsem nadomestiti človeške pomoči.

SUMMARY

Due to the aging of the population and unhealthy lifestyle, we are facing increasing needs for health and social care. To provide it to everyone in need, we lack both financial and human resources, thus we turn to technology for help. Key technologies for this are sensing and artificial intelligence: sensors can monitor physiological signals and user behaviour, while artificial intelligence interprets sensor readings and offers advice tailored to the user's current needs. We look at two examples.

The first is an application for self-management of heart failure, connected to a sensing wristband. The sensors in the wristband monitor heart activity, skin conductivity and movement. From this data, machine learning can extract the user's blood pressure and psychological state. The application then offers a personalised exercise program and alerts the user when it is better not to exercise due to their health state, provides advice on diet, reminds about medications, and more. Additionally, it tries to persuade the user to follow its advice using a cognitive-behavioural approach tailored to their psychological state.

The second example is a system to assist elderly people with memory impairment at household tasks – with a focus on the kitchen. Unlike the first example, this system does not use wearable but ambient sensors. Using computer vision, it recognizes interactions between the user and household appliances, and other activities. It then attempts to determine which task the user is undertaking and – if it notices an error – provides an alert and suggests the correct action.

Such solutions can contribute to better health and independent living, although they cannot and should not completely replace human assistance.

O AVTORJU

Mitja Luštrek je vodja Skupine za ambientalno inteligenco na Institutu "Jožef Stefan". Njegovi glavni raziskovalni področji sta uporaba umetne inteligence v zdravstvu in interpretacija senzorskih podatkov s strojnim učenjem. Bil je nosilec več mednarodnih projektov s tega področja in je dosegel uspehe na tekmovanjih, kakršno je XPrize Pandemic Response Challenge.

ABOUT THE AUTHOR

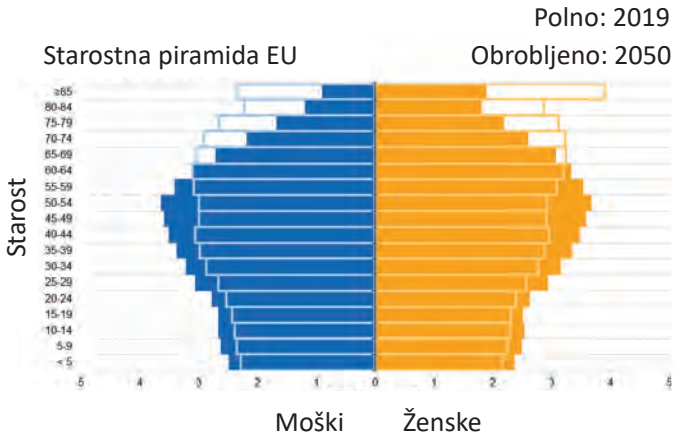
Mitja Luštrek is the head of the Ambient Intelligence Group at Jožef Stefan Institute. His main research interests are application of artificial intelligence in healthcare and interpretation of sensor data with machine learning. He has been the principal investigator in many international projects from this area and was successful in competitions such as XPrize Pandemic Response Challenge.

Senzorji in umetna inteligenca za razumevanje uporabnikov in prilagajanje storitev v zdravstvu in socialni oskrbi

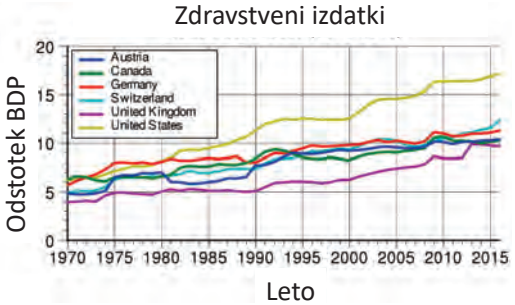
Mitja Luštrek
Odsek za inteligentne sisteme
Institut „Jožef Stefan“

Problem

Starostna piramida EU



Zdravstveni izdatki



Rešitev

- Tehnologija opolnomoči bolnike/starejše, da si pomagajo sami
  - Prilagojena posamezniku
  - Prilagojena trenutnim okoliščinam
  - Prepričevalna, kadar je potrebno
- Senzorji in umetna inteligenca zaznajo in razumejo posameznikove potrebe in kontekst
- Umetna inteligenca ponudi pravi nasvet/storitev
- To stori na (psihološko) ustrezen način
- Ambientalna inteligenca

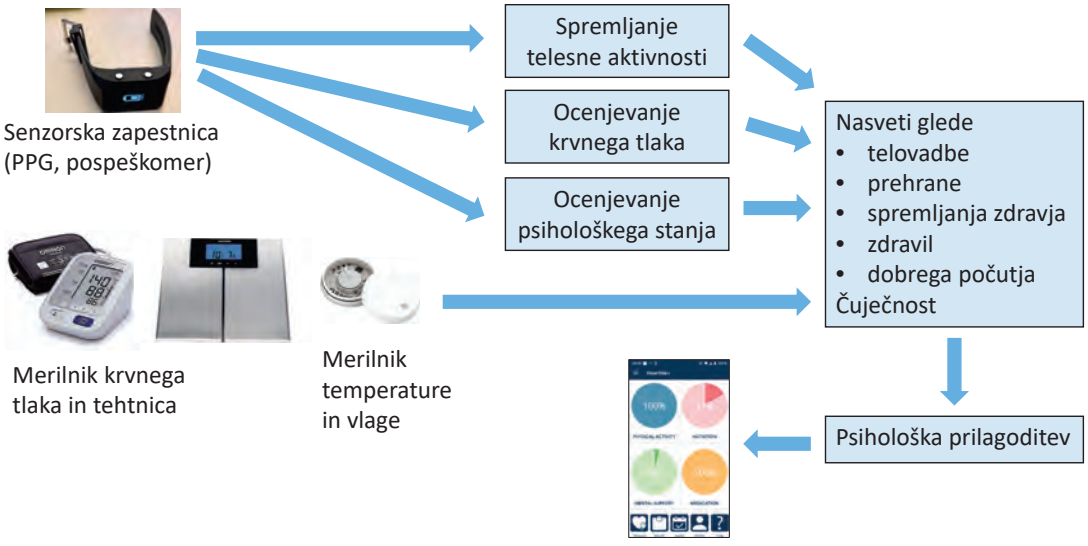
Aplikacija s senzorsko zapestnico za obvladovanje srčnega popuščanja

Obvladovanje srčnega popuščanja

- Bolezen
  - Srce po krvi po telesu ne pošilja preskrbelo s kisikom in hranili
  - Simptomi: otežena telesna aktiv zastajanje tekočine...
  - Doleti 20 % ljudi, pričakovana ži pomemben razlog za sprejem v
- Obvladovanje
  - Več različnih zdravil
  - Primerna telovedba in prehrana
  - Redno spremljanje telesne mase, krvnega tlaka, srčnega utripa ...

- Kognitivno zahtevno
  - Starost pri diagnozi 74 let
  - 25-80 % bolnikov ima kognitivne težave
- Psihološko zahtevno
  - Zdrav življanjski slog
  - 30 % bolnikov ima depresijo ali anksioznost

Shema aplikacije HeartMan



Spremljanje telesne aktivnosti

Senzorska zapestnica
(PPG, pospeškomer)

Spremljanje telesne aktivnosti

Ocenjevanje krvnega tlaka

Ocenjevanje psihološkega stanja

Nasveti glede
- telovadbe
- prehrane
- spremljanja zdravja
- zdravil
- dobrega počutja

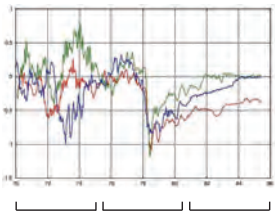
Čuječnost

Merilnik krvnega tlaka in tehtnica

Merilnik temperature in vlage

Psihološka prilagoditev

Spremljanje telesne aktivnosti



Pospeški
Srčni utrip

Okna

a_1	a_2	a_3	a_4	...

Značilke iz pospeškov

Model za prepoznavanje aktivnosti

Aktivnost

a_1	a_2	a_3	...	HR_1	HR_2	...	akt

Značilke iz pospeškov, srčnega utripa

Model za ocenjevanje porabe energije

Poraba energije

Spremljanje telesne aktivnosti

Prepoznana aktivnost

Prava aktivnost	Tek	Stanje	Počitek	Hoja	Jedenje	Premikanje rok
Tek	**367**	0	0	54	0	47
Stanje	1	**222**	96	22	0	32
Počitek	0	24	**1786**	8	60	121
Hoja	15	16	46	**2501**	80	377
Jedenje	0	0	46	16	**461**	95
Premikanje rok	6	4	32	269	129	**1557**

Spremljanje telesne aktivnosti

Prepoznana aktivnost

Prava aktivnost	Tek	Stanje	Počitek	Hoja	Jedenje	Premikanje rok
Tek	**367**	0	0	54	0	47
Stanje	1	**222**	96	22	0	32
Počitek	0	24	**1786**	8	60	121
Hoja	15	16	46	**2501**	80	377
Jedenje	0	0	46	16	**461**	95
Premikanje rok	6	4	32	269	129	**1557**

Kontekst = počitek Kontekst = hoja Kontekst = jedenje

Ocenjevanje krvnega tlaka

Senzorska zapestnica
(PPG, pospeškomer)

Spremljanje telesne aktivnosti

Ocenjevanje krvnega tlaka

Ocenjevanje psihološkega stanja

Nasveti glede
- telovadbe
- prehrane
- spremljanja zdravja
- zdravil
- dobrega počutja

Čuječnost

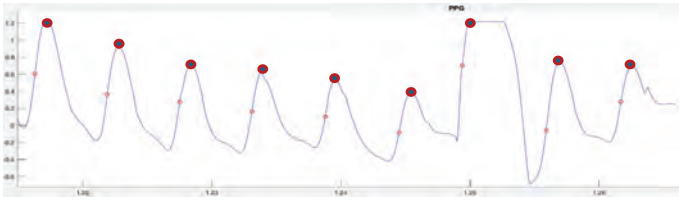
Merilnik krvnega tlaka in tehtnica

Merilnik temperature in vlage

Psihološka prilagoditev

Ocenjevanje krvnega tlaka

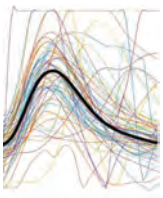
PPG



LED

Foto-detektor

Žile

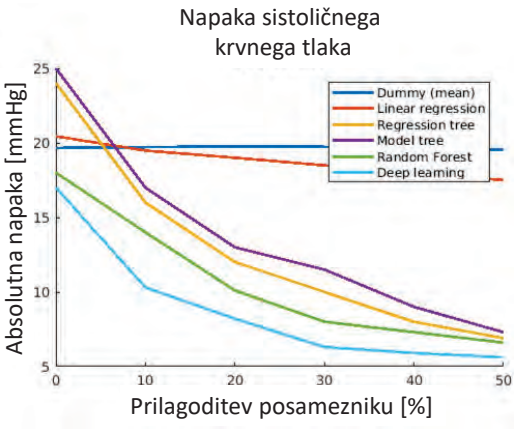


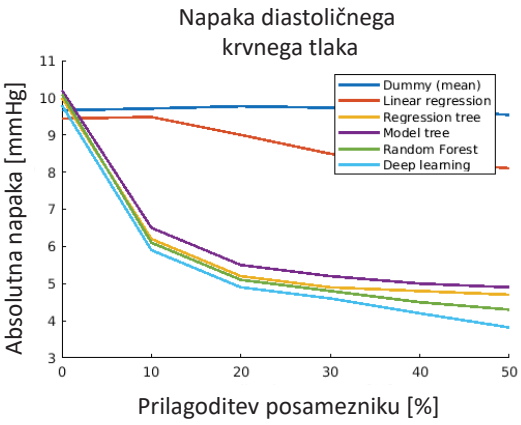
Krvni tlak

Model za ocenjevanje tlaka

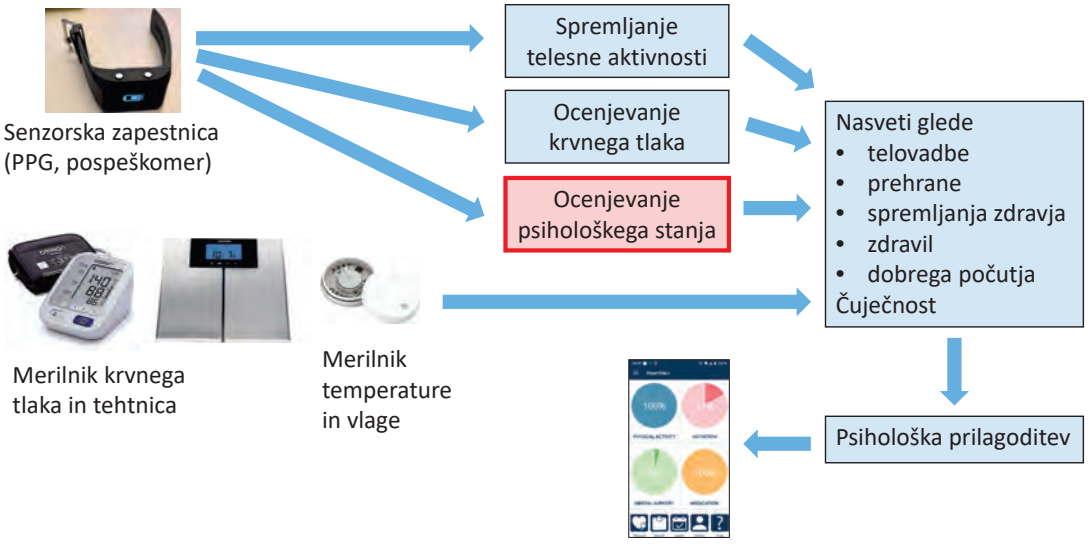
PPG_1 | PPG_2 | ...

Ocenjevanje krvnega tlaka

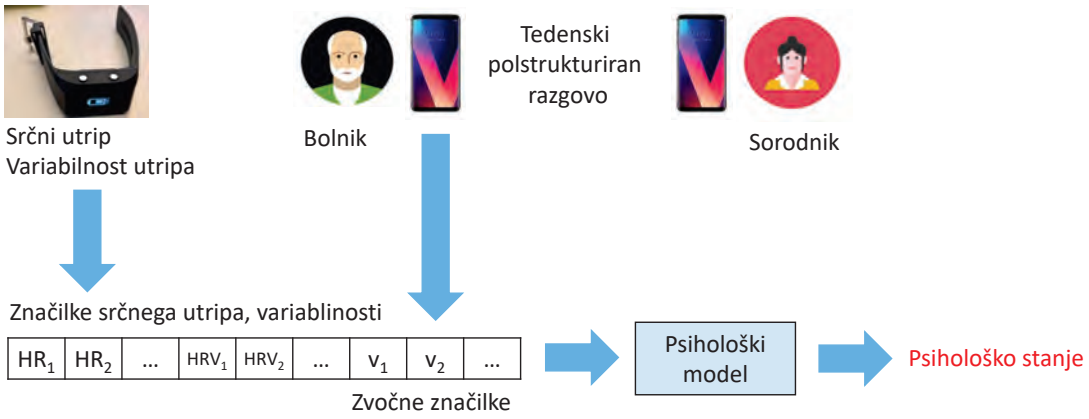




Ocenjevanje psihološkega stanja



Ocenjevanje psihološkega stanja



Ocenjevanje psihološkega stanja

		Ocenjeno stanje			
		Motivirano	Anksiozno	Depresivno	Točnost
Pravo stanje	Motivirano	**1193**	41	34	94 %
	Anksiozno	48	**882**	130	83 %
	Depresivno	31	95	**888**	88 %

Nasveti in storitve

Senzorska zapestnica
(PPG, pospeškomer)

Spremljanje telesne aktivnosti

Ocenjevanje krvnega tlaka

Ocenjevanje psihološkega stanja

Nasveti glede
- telovadbe
- prehrane
- spremljanja zdravja
- zdravil
- dobrega počutja

Čuječnost

Merilnik krvnega tlaka in tehtnica

Merilnik temperature in vlage

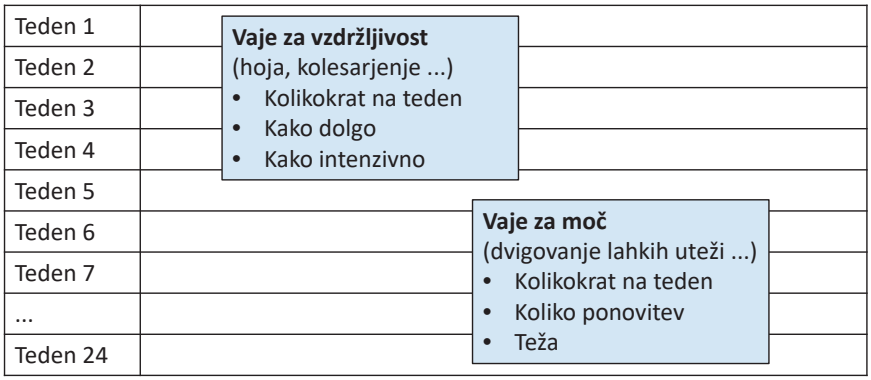
Psihološka prilagoditev

Program telovadbe

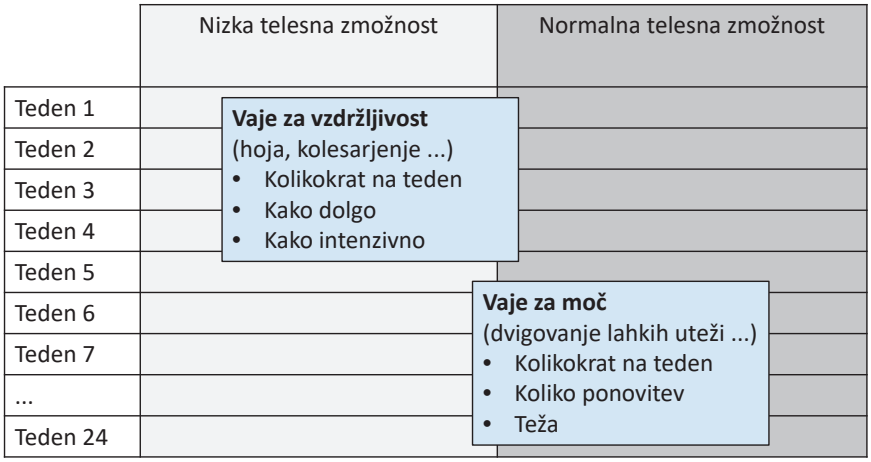
Vaje za vzdržljivost
(hoja, kolesarjenje ...)
- Kolikokrat na teden
- Kako dolgo
- Kako intenzivno

Vaje za moč
(dvigovanje lahkih uteži ...)
- Kolikokrat na teden
- Koliko ponovitev
- Teža

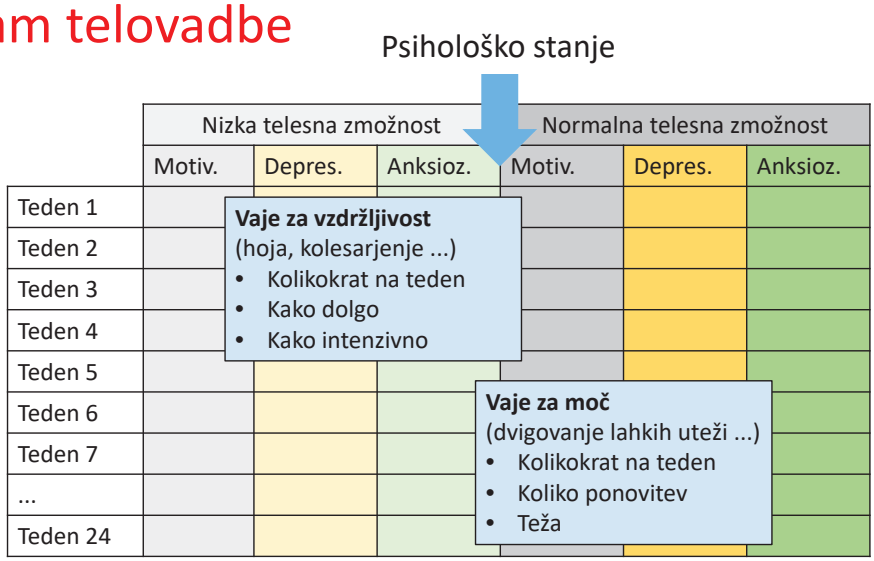
Program telovadbe



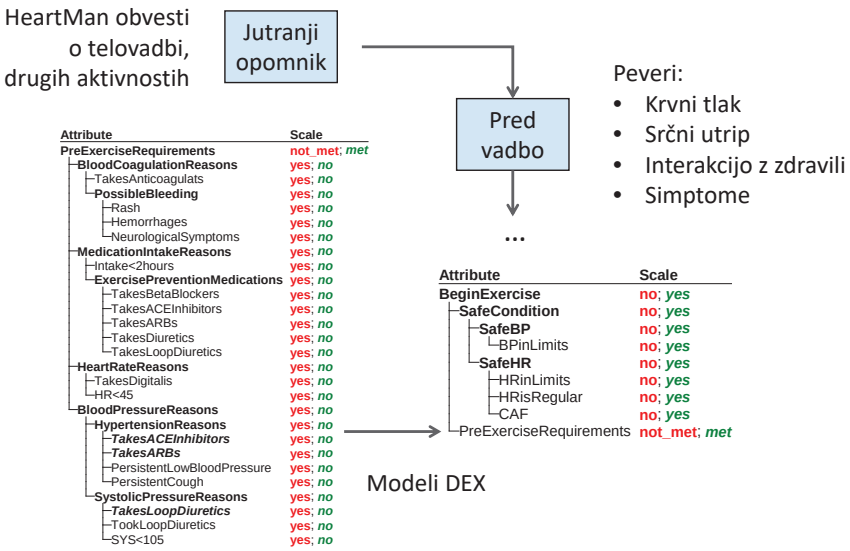
Program telovadbe



Program telovadbe



Dnevna vadba



Ocenjevanje psihološkega stanja

Senzorska zapestnica
(PPG, pospeškomer)

Spremljanje telesne aktivnosti

Ocenjevanje krvnega tlaka

Ocenjevanje psihološkega stanja

Nasveti glede
- telovadbe
- prehrane
- spremljanja zdravja
- zdravil
- dobrega počutja

Čuječnost

Merilnik krvnega tlaka in tehtnica

Merilnik temperature in vlage

Psihološka prilagoditev

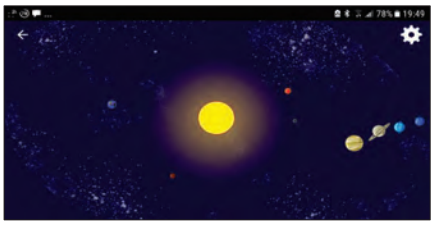
(Kontekstu prilagojena) čuječnost

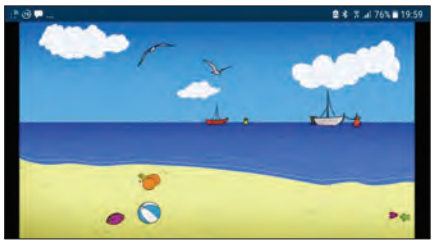
Srčni utrip
(biološka povratna zanka, biofeedback)

TAKE E MINDFUL MOMENT now

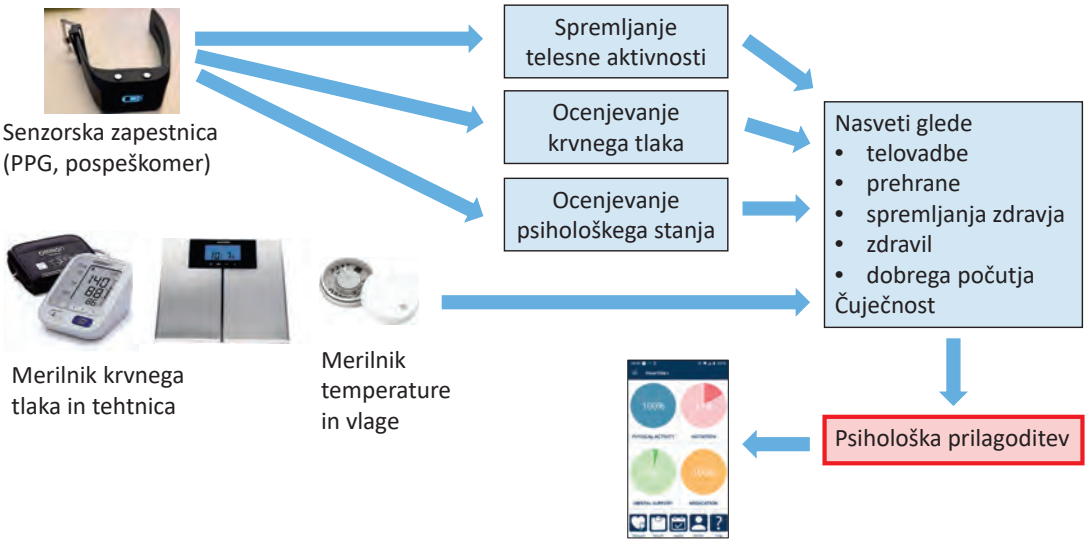
MINDFUL WALKING: focus on each step, from your heel hitting the ground to the ball of your foot to your toes. Dont't worry about where you have to get to, just focus on the act of walking. It's almost like being a toddler again.

Zaznan kontekst = hoja

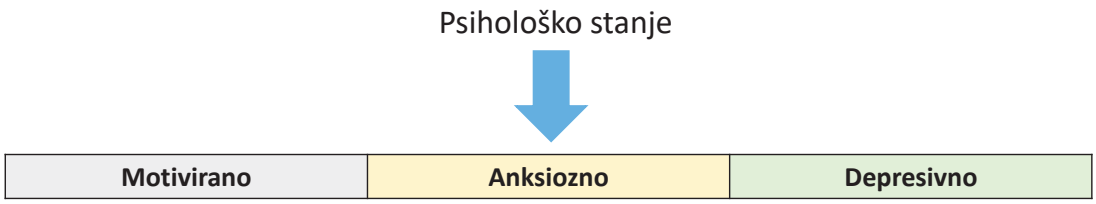




Aplikacija HeartMan



Psihološko prilagojena sporočila



Psihološko prilagojena sporočila

Psihološko stanje

	Motivirano	Anksiozno	Depresivno
	Prisiljena skladnost		
Telovedba	Gibanje je dobro za moje zdravje, zato se bom več gibal.		
Prehrana	...		

Psihološko prilagojena sporočila

Psihološko stanje

	Motivirano	Anksiozno	Depresivno
	Prisiljena skladnost	*Svobodna izbira*	
Telovedba	Gibanje je dobro za moje zdravje, zato se bom več gibal.	10-minuten sprehod in gledanje televizije sta dva načina sproščanja. Sprehod izboljša zdravje, televizija pa ne.	
Prehrana	...	...	

Psihološko prilagojena sporočila

Psihološko stanje

	Motivirano	Anksiozno	Depresivno
	Prisiljena skladnost	*Svobodna izbira*	*Utemeljitev truda*
Telovedba	Gibanje je dobro za moje zdravje, zato se bom več gibal.	10-minuten sprehod in gledanje televizije sta dva načina sproščanja. Sprehod izboljša zdravje, televizija pa ne.	Če bom grem na 10-minutni sprehod, bom deležen podobne koristi kot od zdravil.
Prehrana	...	...	...

Ambientalni sistem za spremljanje in pomoč pri vsakdanjih opravilih

Pomoč za ljudi s kognitivno motnjo

- 16 % ljudi nad starostjo 50 ima blago kognitivno motnjo, 5 % pa demenco
- Nezmožnost opravljanja vsakdanjih opravil krni njihovo kakovost življenja, obremenjuje skrbnike in jih na koncu prisli v institucionalizacijo
- Z računalniškim vidom lahko razumemo njihove aktivnosti in pomagamo, če želijo

Shema sistema CoachMyLife

Kamera

Prepoznavanje uporabnika

Prepoznavanje predmetov

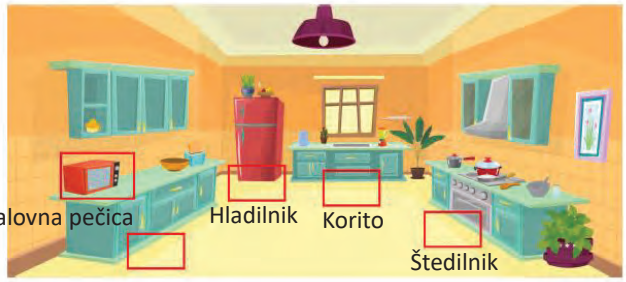
Shema sistema CoachMyLife

Kamera

Prepoznavanje uporabnika

Prepoznavanje predmetov

Označevanje kuhinjskih lokacij



Shema sistema CoachMyLife

Kamera

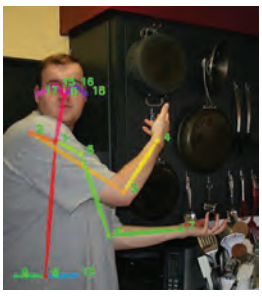
Prepoznavanje uporabnika

Prepoznavanje predmetov

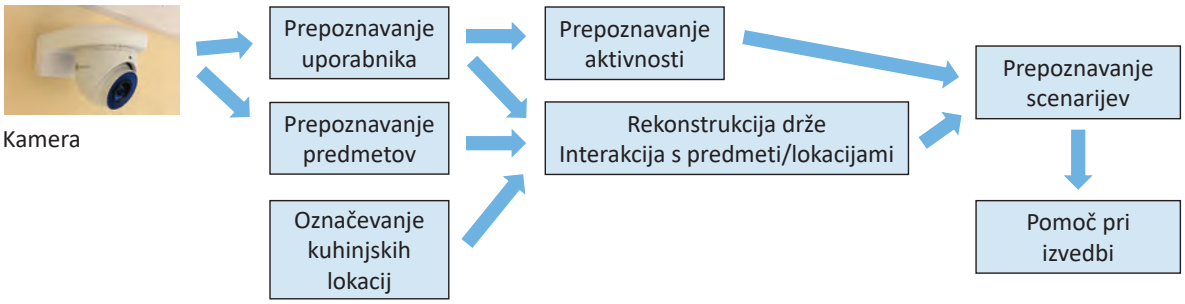
Označevanje kuhinjskih lokacij

Prepoznavanje aktivnosti

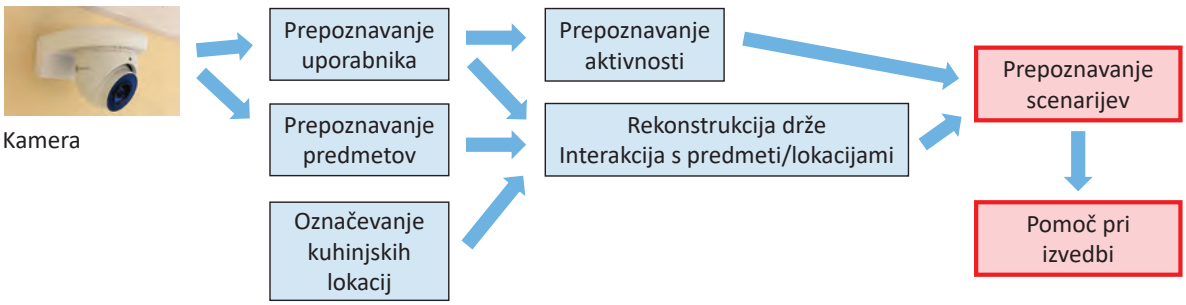
Rekonstrukcija drže
Interakcija s predmeti/lokacijami



Shema sistema CoachMyLife



Prepoznavanje scenarijev in pomoč



Prepoznavanje scenarijev in pomoč

Prepoznavanje scenarijev

Scenarij 1
Korak 1: vzemi mleko iz hladilnika
Korak 2: vzemi skodelico iz omare (lahko zamenjano)
Korak 3: stopi do kavomata in pripravi kavo
Z napako: najprej stopi do kavomata

Scenarij 2
Priprava čaja

Scenarij 3
Gretje hrane v mikrovalovni pečici

	Pravilno			**Z napako**		
	Scenarij 1	Scenarij 2	Scenarij 3	Scenarij 1	Scenarij 2	Scenarij 3
Korak 1	60 %	100 %	100 %	100 %		100 %
Korak 2	80 %	100 %	100 %	80 %	100 %	100 %
Korak 3	100 %	100 %	100 %	70 %	100 %	100 %

Uporaba sistemov kibernetskih vab za varovanje podatkov in oceno tveganja

Using deception technology to protect data and perform risk assessment

Matej Rabzelj

Univerza v Ljubljani, Fakulteta za elektrotehniko

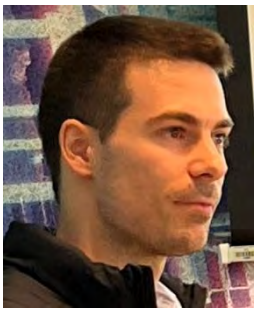
POVZETEK

Digitalizacija in razmah informacijsko-komunikacijskih tehnologij in razvoj umetne inteligence poleg novih oblik poslovanja prinašata tudi skokovit razvoj kibernetskih groženj. S pospešenim sklapljanjem naprav v internet ter s shranjevanjem in obdelavo občutljivh podatkov v oblaku namreč vsakodnevno ustvarjamo idealen poligon za kibernetske napadalce. Orodja, tehnike in znanja, uporabna za izvedbo napadov sedaj dopolnjujejo še rešitve umetne intiligence, ki lahko koristijo tako napadalcem kot kibernetski obrambi. Zaščita IKT sistemov v prvi vrsti terja dobro razumevanje kibernetskih napadov in tehnik napadalcev. Za namen identifikacije in analize tehnik, taktik in procedur napadalcev smo razvili visoko-interaktiven sistem porazdeljenih kibernetskih vab (angl. honeynet), ki s posnemanjem različnih naprav in storitev na internetu prestrezajo realne kibernetske napade. Razvite rešitve za zajem, obdelavo in vizualizacijo podatkov omogočajo identifikacijo ključnih podatkov o ciljih, orodjih in metodah zlonamernih akterjev ter ponujajo vpogled v potek kibernetskih groženj v realnem času in analizo vedenja napadalcev na osnovi agregacije podatkov.

SUMMARY

The proliferation of digital technologies, coupled with advances in artificial intelligence, has transformed business operations. Unfortunately, this transformation has also precipitated a dramatic escalation in cyber threats. As devices are increasingly connected to the internet and sensitive data is processed and stored in the cloud, we are unwittingly forging an ideal playground for cyber adversaries. The tools, techniques, and knowledge necessary for conducting attacks are now further augmented by artificial intelligence solutions, benefiting both the attackers, as well as cyber defence. Protecting ICT systems demands an understanding of cyberattacks and techniques employed by adversaries. To aid understanding of cyberattacks, we have developed a highly interactive, distributed system of network honeypots that capture real cyberattacks by mimicking various internet-connected devices and services. Our solutions for data capture, processing, and visualisation enable the identification of crucial information regarding the techniques, tactics, and procedures of malicious actors. They provide real-time insights into the dynamics of cyberthreats and enable a nuanced analysis of attacker behaviour using data aggregation.

O AVTORJU



Matej Rabzelj je doktorski študent v Laboratoriju za telekomunikacije na Fakulteti za elektrotehniko Univerze v Ljubljani, kjer se ukvarja s področjem kibernetske varnosti. Njegovo raziskovalno delo zajema obravnavo kibernetske varnosti računalništva v oblaku in sistemov kritičnih infrastruktur ter razvoj rešitev za kibernetsko obrambo na osnovi omrežnih senzorjev.

ABOUT THE AUTHOR

Matej Rabzelj is a doctoral student in the Laboratory for Telecommunications at the Faculty of Electrical Engineering in Ljubljana, where he works in the field of cyber security. His research work covers the cyber security of cloud computing and critical infrastructure systems and the development of cyber defense solutions based on network sensors.

Uporaba sistemov kibernetskih vab za varovanje podatkov in oceno tveganja

Matej Rabzelj

Laboratorij za telekomunikacije,
Fakulteta za elektrotehniko

17. maj 2024

FE UNIVERZA V LJUBLJANI Fakulteta za elektrotehniko | LTFE Laboratorij za telekomunikacije

O nas

UNIVERZA V LJUBLJANI
FE Fakulteta za elektrotehniko
Katedra za informacijsko komunikacijske tehnologije
Laboratorij za telekomunikacije

www.ltfe.org @ltfe @ltfe.org @ltfe.ul

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Nekaj referenc s področja kibernetske varnosti

- **R&D projekti**
  - AdaHon (FED4FIRE+, development and deployment of adaptive honeynets)
  - CRP V2-2125 (ARIS & URSIV; development of specialized honeypots)
  - CRP V2-2378 (ARIS & MORS & URSIV; KV obrambnih sistemov in kritičnih infrastruktur)
  - 5G Safety (MIZŠ & ERDF; cybersecurity aspects in 5G environments)
  - PASEI 2 (EDA; Protection of Autonomous Systems Against Enemy Interference)
  - INDY (EDF; Energy Independent and Efficient Deployable Military Camps)
  - HIBROM (Varnost pametnih agregatov v vojaški domeni)

- **Industrijski projekti**
  - Izobraževanje (ICT Academy, razvoj tečajev in strokovnih vsebin)
  - Svetovanje (cybersecurity, S-SDLC, vulnerability management)
  - Varnostni pregledi produktov in risk assessment

- **Pedagoške aktivnosti**
  - Razvoj novega študijskega programa na temo kiberntske varnosti
  - Prvi predmeti že potekajo (v sklopu projekta NOO ULTRA)

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Naše raziskovalno delo

- Ukvarjamo se z zaščito sistemov
  - Za zaščito je nujna ocena tveganja

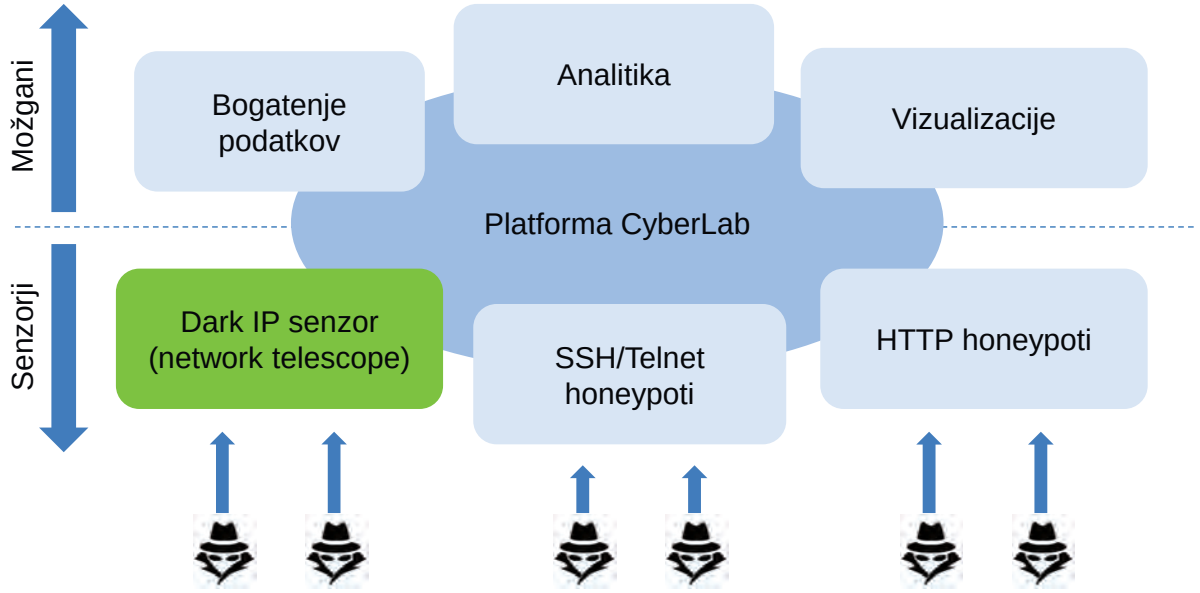
- Zato moramo vedeti kaj napadalci počnejo
  - Koliko prometa lahko pričakujemo in kakšen je?
  - Kako lahko klasificiramo napadalce glede na tehničen skillset?
  - Kaj je baseline (in kako se bo povečal zaradi LLM in drugih orodij)?

- Vse to je vedno bolj pomembno zaradi zaskrbljujočih trendov napadov
  - Nova zakonodaja nalaga ponudnikom bistvenih storitev še večjo odgovornost

- Informacije zbiramo neposredno od napadalcev
  - To nam omogoča neposreden vpogled v TTP napadalcev

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

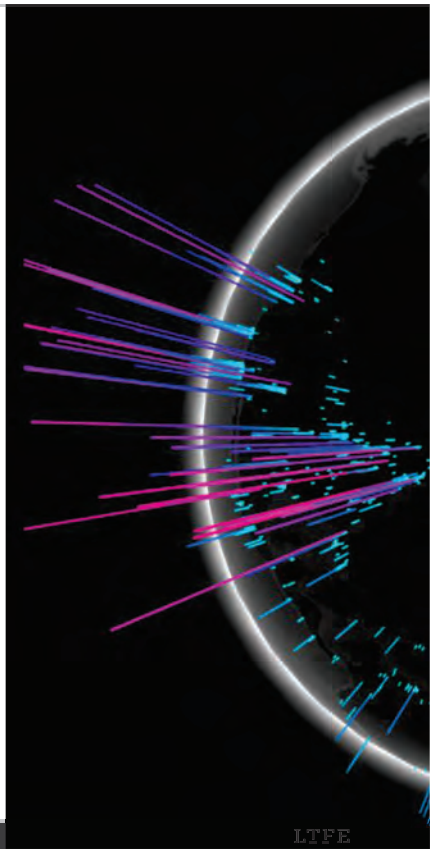
Širša slika



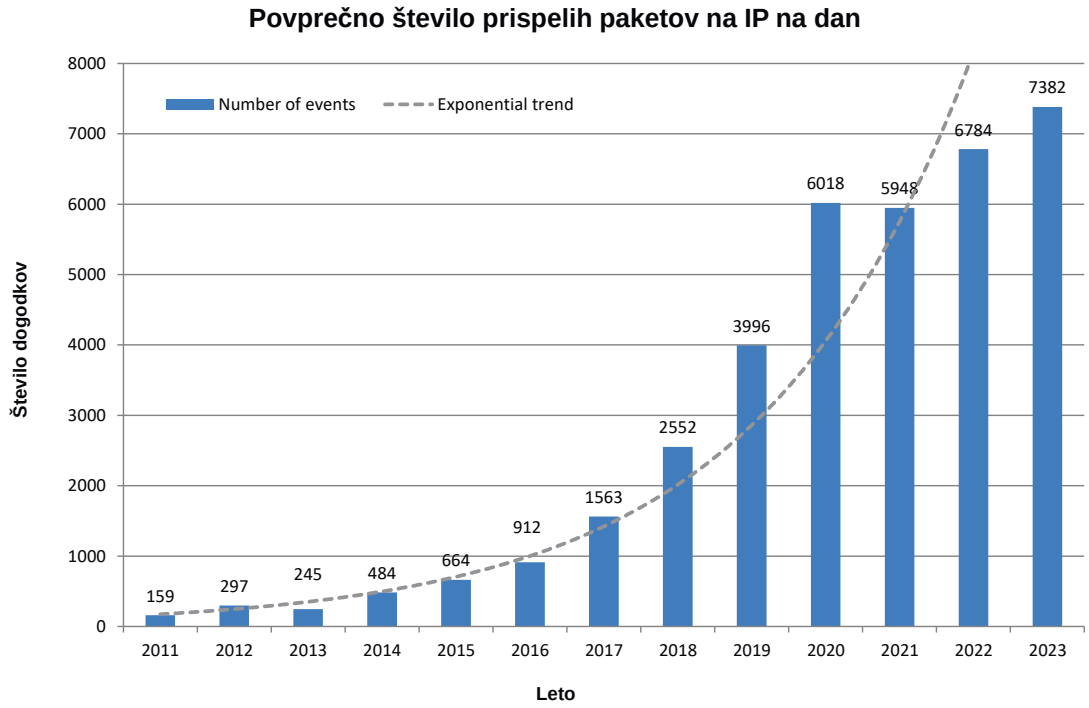
© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Omrežni teleskop

- Subnet /24
  - Nikoli uporabljen (dark IP range)
  - Listen-only
- Analogija: naselje praznih hiš
  - Samo zato, da vidimo, če kdo pride trkat
  - In da vidimo, katera vrata so bolj popularna
- Promet beležimo že 12 let



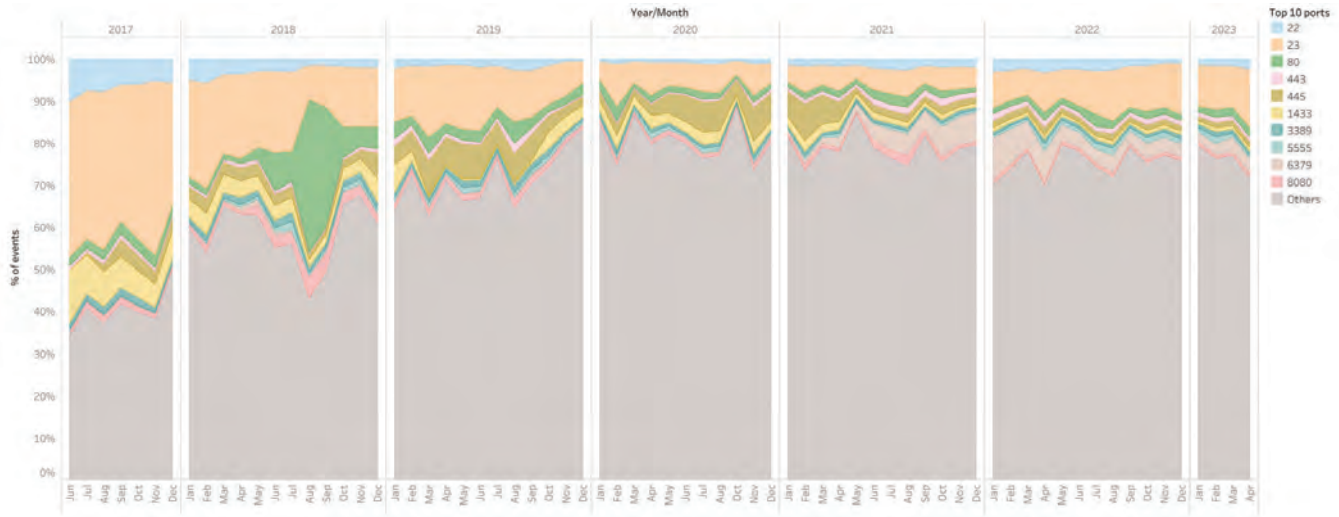
© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

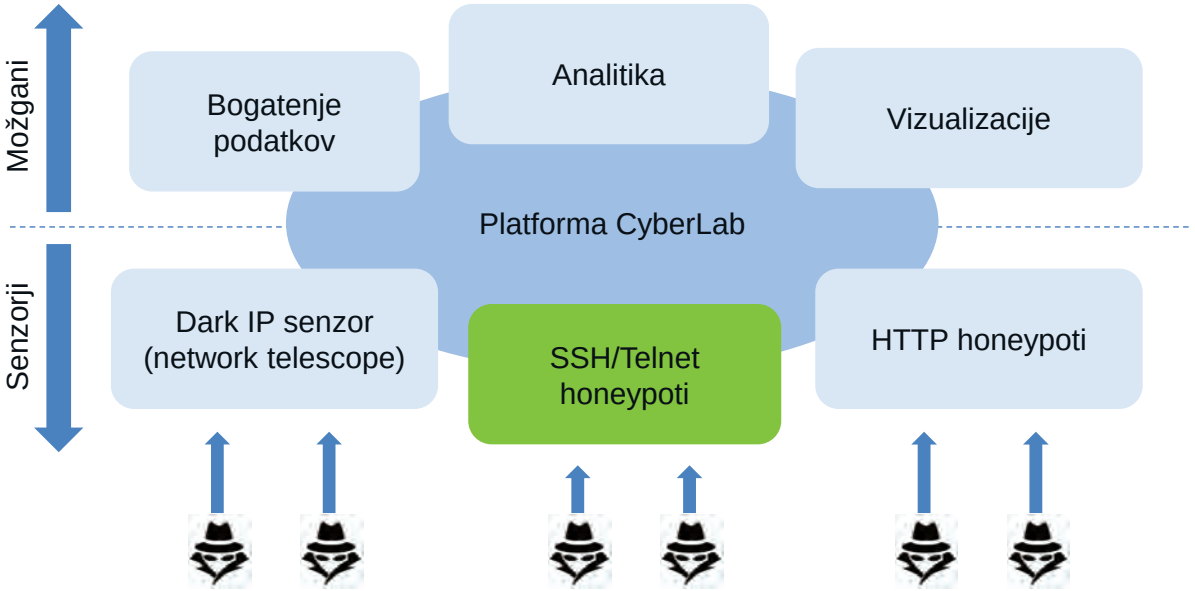
Analiza trendov in backscatter prometa

- Relativni deleži ciljnih vrat
  - 2017: 50% SSH in Telnet
  - danes: večinoma "other"



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

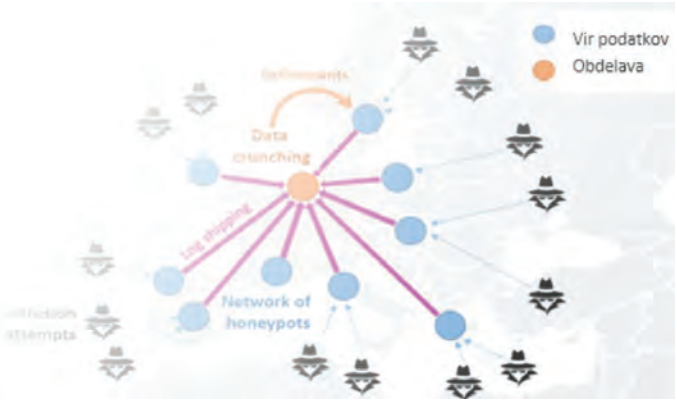
Širša slika



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Kaj je honeypot?

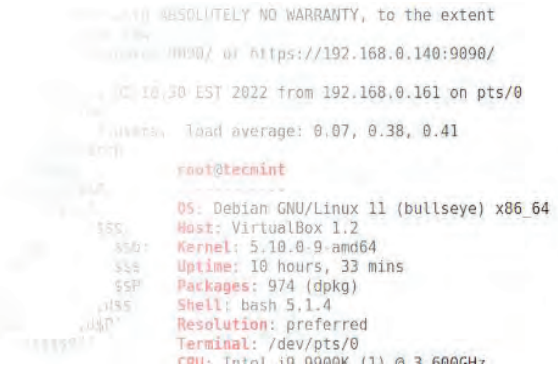
- Storitev, ki se uporablja samo za opazovanje ali zavajanje napadalcev
- Upravljamo ducate lastnih honeypotov po celem svetu
- Vse podatke centralizirano beležimo in obdelujemo
- Posebno pozornost namenjamo prepričljivosti storitev, a hkrati omejujemo možnost odskočnih napadov



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Telnet in Secure Shell (SSH)

- Protokola za oddaljen konzolni dostop
  - Privzeti način upravljanja z *nix strežniki
  - Možna enostavna avtomatizacija
- Kako lahko vidimo kaj napadalec počne?
  - Metapodatki o napadu
    (IP naslovi, verzije odjemalcev, IP geolokacija)
  - Podrobna vsebina seje
    (seznam ukazov, časovno sosledje)
  - Analiza škodljivih tovorov
    (skripte, binary datoteke)



Zbrani podatki

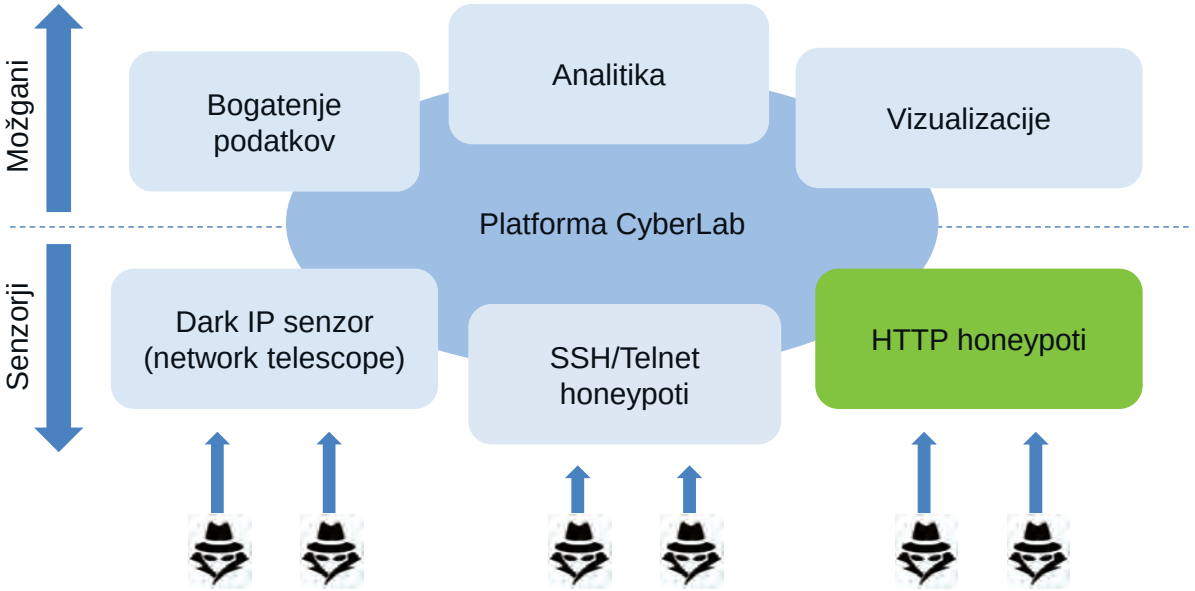
Vzorec dela zbranih podatkov:

Field	Description	Example
session_id	Unique ID of the session	7efa843ba2d1
dst_ip_identifier	Obfuscated (pseudonymized) destination public IPv4 address of the honeypot node	b1e84f073e89a5a22482d2d112acfacb774d478f49a53951369ce029f6956339
dst_host_identifier	Obfuscated (pseudonymized) name of the honeypot node	0b6413ddbb3aebd1e5b4f226a60016af8c6138386a8195d453c7dbda2cbbef38
src_ip_identifier	Obfuscated (pseudonymized) IP address of the attacker	b1e84f073e89a5a22482d2d112acfacb774d478f49a53951369ce029f6956339
eventid	Event id of the session in the cowrie honeypot	cowrie.session.connect
timestamp	UTC time of the event	2019-11-01T00:00:24.107618Z
message	Message of the cowrie honeypot; any public source and IP address mentions are replaced with the pseudonym. Other IPs, including private infrastrcture addresses are left as is.	New connection: b1e84f073e89a5a22482d2d112acfacb774d478f49a53951369ce029f6956339:44502 (192.168.144.2:2222) [session: 7efa843ba2d1]
protocol	Protocol used in the cowrie honeypot; either ssh or telnet	ssh
geolocation_data/postal_code	Source IP postal code as (determined by the logstash node)	52100
geolocation_data/continent_code	Source IP continent code (as determined by the logstash node)	EU
geolocation_data/country_code3	Source IP country code3 (as determined by the logstash node)	IT

src_port	Source TCP port	44502
sensor	Sensor name, which serves to identify our experiment configuration	ubuntu-ssh
arch	Represents the CPU/OS architecture emulated by cowrie	null
duration	Session duration in seconds	null
ssh_client_version	Attacker's SSH client version	null
username	Username used when attempting to log in; only set at the login attempt at the beginning of session, otherwise null	admin
password	Password used when attempting to log in; only set at the login attempt at the beginning of session, otherwise null	12345
macCS	HMAC algorithms supported by the client (SSH MAC supported in the session)	["hmac-sha1", "hmac-md5", "hmac-sha2-256"]
encCS	Encryption algorithms supported by the client	["3des-cbc", "aes256-ctr", "aes256-cbc", "aes192-ctr", "aes192-cbc", "aes128-cbc", "blowfish-cbc", "aes128-ctr", "cast128-cbc"]
kexAlgs	Key exchange algorithms supported by the client	["diffie-hellman-group14-sha1", "diffie-hellman-group1-sha1", "diffie-hellman-group-exchange-sha1", "diffie-hellman-group-exchange-sha256"]
keyAlgs	Public key algorithms supported by the client	["ssh-rsa", "ssh-dss"]

Podatke nato obogatimo v nadaljnjem podatkovnem cevovodu.

Širša slika



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

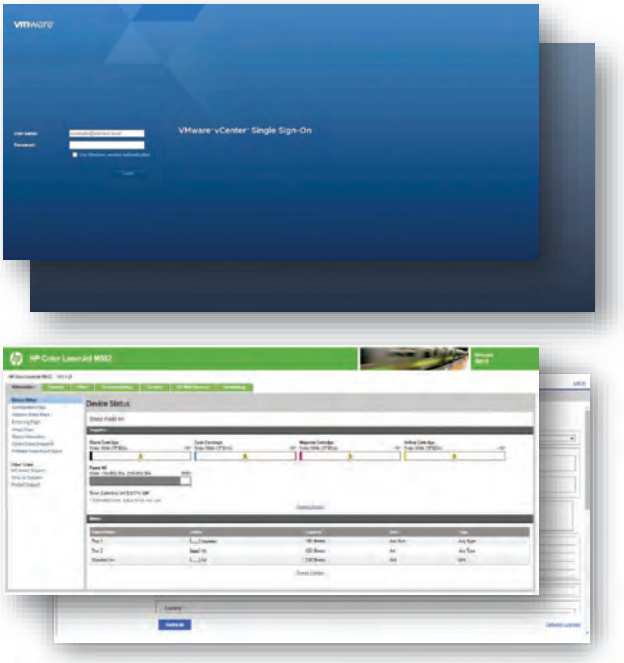
Univerzalnost protokola HTTP

- Spletne strani in interaktivne aplikacije
- IoT naprave
- Oblačne storitve in API vmesniki
- Infrastruktura kriptovalut

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Razviti modeli HTTP storitev

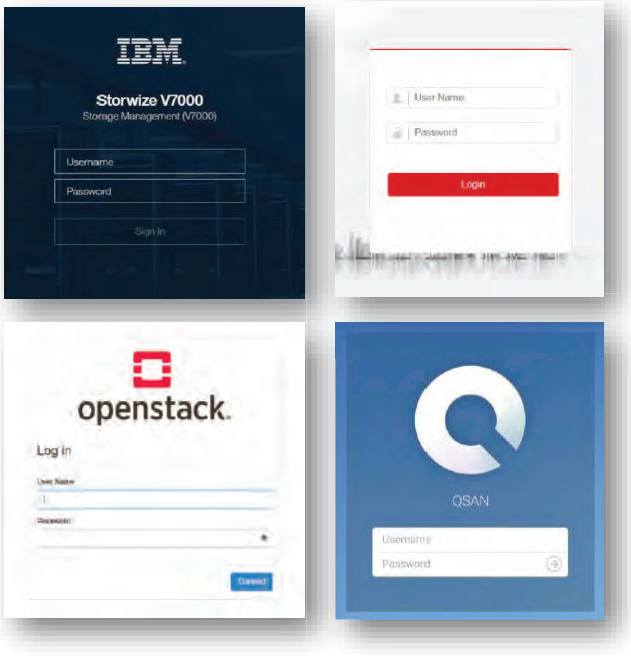
- Različni nivoji interaktivnosti
  - High-interaction honeypots
  - Low-interaction honeypots
- Primeri podprtih storitev
  - Kamere
  - Tiskalniki
  - Storage sistemi
  - Cloud in virtualizacija
  - Podatkovne baze
  - CMS systemi
- Več kot raznolikih 15 storitev



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Razviti modeli HTTP storitev

- Različni nivoji interaktivnosti
  - High-interaction honeypots
  - Low-interaction honeypots
- Primeri podprtih storitev
  - Kamere
  - Tiskalniki
  - Storage sistemi
  - Cloud in virtualizacija
  - Podatkovne baze
  - CMS systemi
- Več kot raznolikih 15 storitev



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Zbrani podatki

General request metadata

index	80434
honeypot_host_ip	[censored]
honeypot_id	[censored]
session_id	b4d833296d2bebe6
timestamp	2023-06-06 17:08:48.885000
src_ip	[censored]
src_port	40922
host_id	[censored]
honeypot_type	http
dst_ip	[censored]
dst_port	80
honeypot_host_id	cyberlab-honeypot-01
method	GET
request_url	http://[censored]/
request_body	
honeypot_technology	storwize_v7000
response_served	true
response_status_code	200

HTTP header data and classification

headers	{"accept-encoding": "gzip, deflate", "accept-language": "en-US", "upgrade-insecure-requests": "1", "connection": "keep-alive", "user-agent": "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/86.0.4 240.111 Safari/537.36", "host": "[censored]", "accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8,application/signed-exchange;v=b3;q=0.9"}
user_agent	Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/86.0.4 240.111 Safari/537.36
headers_hash	f3f38cb3ed49b10015e451ec05af0c2b
headers_signature	bc651dc0130f7a4f
headers_software	Chrome 86
headers_operating_system	Linux
headers_device	unknown
headers_device_type	pc
headers_is_weird	false

Browser fingerprinting data

- {}
  - document_title:V7000 - Log in - IBM Storwize V7000
  - local_storage_uuid:9qwq08t2nhjhvw6b2bg4yj3tlr9yjdd
  - {} navigator_data
    - hardware_concurrency:56
    - cookie_enabled:true
    - language:en-US
    - [] languages
      - en-US
    - max_touch_points:0
    - [] mimetypes
    - webdriver:true
    - user_agent:Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/86.0.4 240.111 Safari/537.36
    - vendor:Google Inc.
    - product:Gecko
    - [] plugins
    - platform:Linux x86_64
    - online:true
  - {} screen_data
  - {} selenium_data
    - detected_via_keys:false
    - detected_via_webdriver:true
  - time_zone:Asia/Shanghai
  - time:1686071335258
  - src_ip:[censored]
  - host_id:[censored]
  - timestamp:2023-06-06T17:08:55.782

IP reputation and scanner info

rdns	unknown
noise	true
whitelisted	true
classification	malicious
name	unknown
uniqueness_alert	HIGH-MEDIUM
known_scanner	true
scanner_id	zoomeye-scan
scanner_name	ZoomEye
scanner_category	scanner
scanner_trust_level	3
ip_trust_level	5
is_datacenter	false
is_tor	false
is_proxy	false
is_vpn	false
is_abuser	true

Autonomous System information

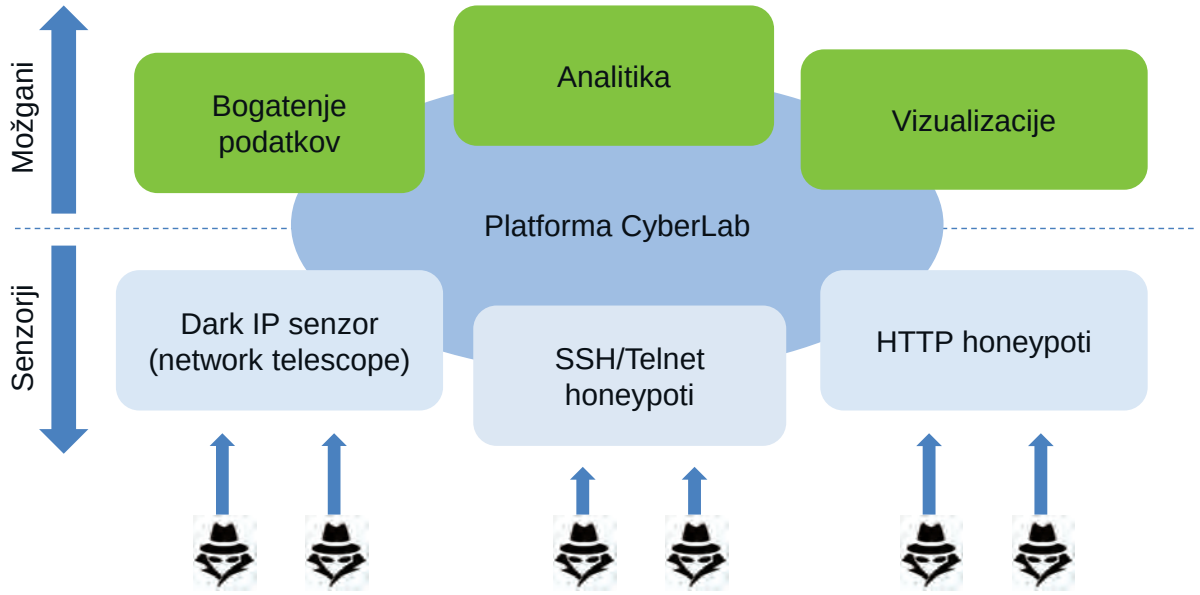
asn	38283
company_name	CHINANET Sichuan province network
datacenter_name	
country	China
city	Deyang
asn_type	isp
asn_organization	CHINANET SiChuan Telecom Internet Data Center
org_description	CHINANET-SCIDC-AS-AP CHINANET SiChuan Telecom Internet Data

IP geolocation

country_name	China
latitude	30.6498
longitude	104.0555
continent_code	AS
region_name	Sichuan
city_name	Chengdu

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Širša slika



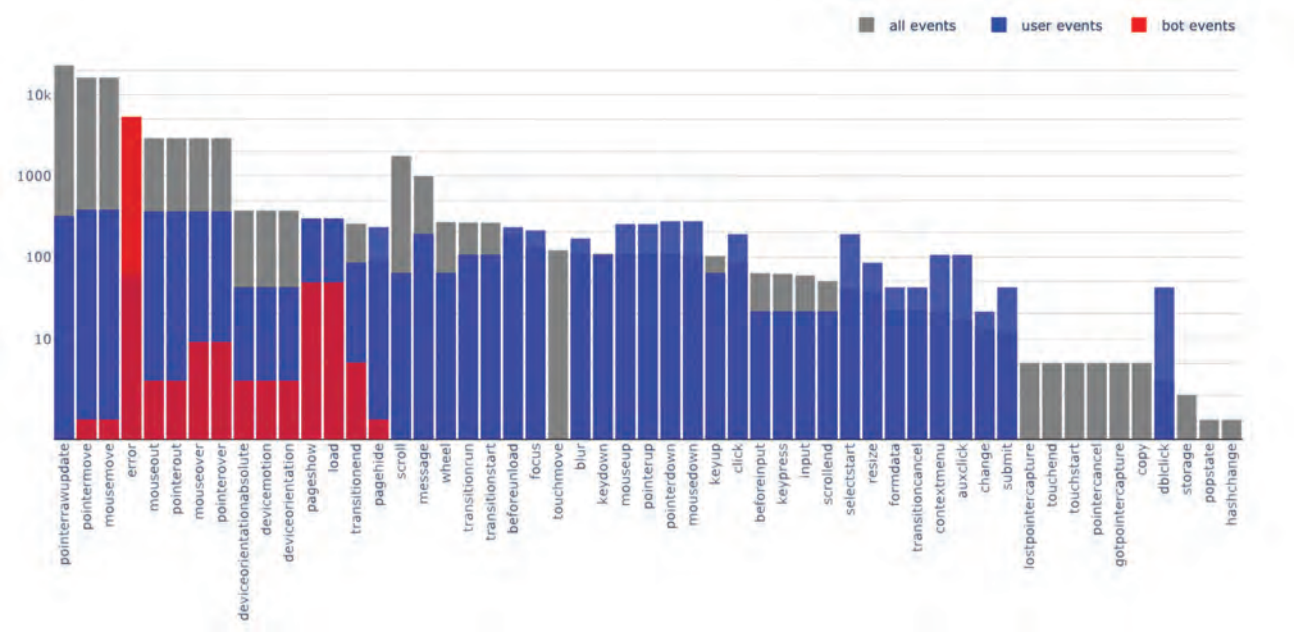
© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Analitika

- Glavna dodana vrednost je v analitiki
  - Klasifikacija in odstranitev “šuma ozadja” (scannerji, crawlerji)
  - Orodja za interaktivno eksploracijo in vizualizacijo podatkov
  - Detekcija botov in avtomatiziranih skript
  - Klasifikacija in modeliranje napadalcev
  - Primerjava med sektorji in ponudniki

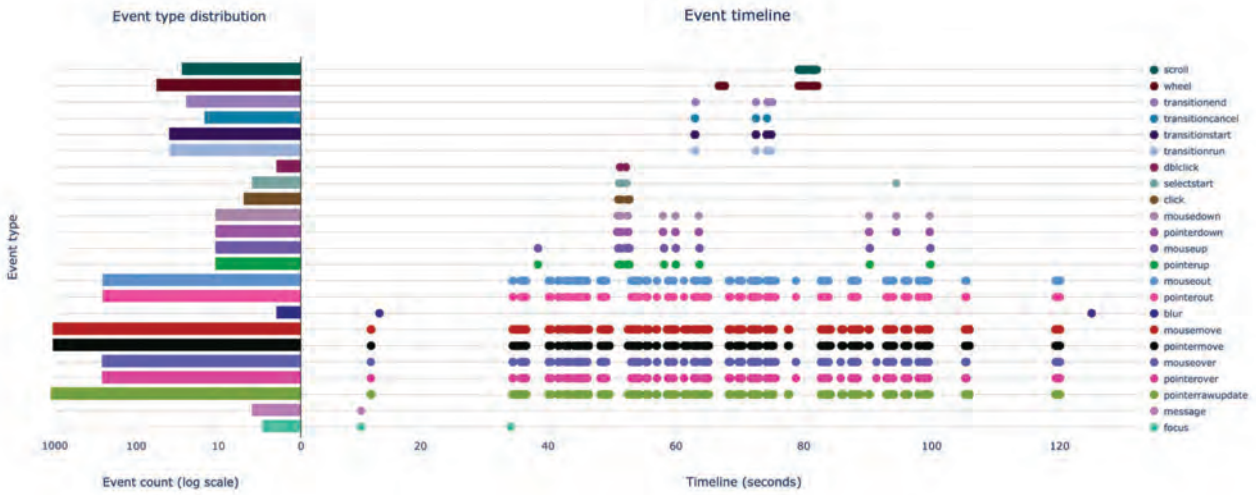
© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Primer: detekcija botov (HTTP)



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Primer: detekcija botov (HTTP)



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Primer: klasifikacija skenerjev (HTTP)

- Velik del prometa prihaja iz "spletnih iskalnikov" (ranljivosti)
  - Uporabljamo različne napredne tehnike za zaznavo in klasifikacijo
  - Ocenjujemo škodljivost njihovih zahtev
  - Identifikacija lažnih skenerjev

- Trenutno beležimo več kot 8000 IP naslovov, ki pripadajo 45 različnim skenerjem (> 15% prometa).

SHODAN ZoomEye

censys LeakIX

	classification		
actor_name	benign	malicious	unknown
Academy for Internet Research	3,880		
Alpha Strike Labs	122		
BinaryEdge	481	18	9
BitSight Technologies, Inc.	1,499		
Censys	4,328		
Criminal IP	429		
CyberResilience	74		
Driftnet	575	25	
Fofa Huashun Xin'an Technology		225	151

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Primer: specializacija napadalcev

Honeypot Technology

% of Total Count: 0.000% – 10.000%

SNORT classtype	APC SmartUPS 20kVA	IBM Storwize v7000	Joomla 3	Joomla 4	MongoDB v2.4	MongoDB v3.2	NAS QNAP v3.3.6	NAS QSAN 3.3.1	NAS Thecus 2.03.06	Openstack 17.0.5	PhpMy-Admin v5. 1.1	Printer EPSON C20600	Printer HP Color LaserJet ..	VMware vCenter Server 6.5
None	85.71%	82.43%	86.56%	81.59%	90.13%	89.68%	82.24%	82.54%	79.04%	85.01%	85.65%	68.51%	83.90%	72.75%
attempted-admin	1.90%	1.48%	1.04%	1.56%			1.32%	1.48%	1.38%	1.42%	0.99%	0.98%	0.99%	5.73%
attempted-dos	0.15%	0.38%	0.25%	0.39%			0.54%	0.36%	0.38%	0.35%	0.29%	0.29%	0.33%	0.59%
attempted-recon	7.15%	8.86%	6.36%	8.62%			8.69%	7.72%	8.15%	7.48%	6.92%	7.09%	8.61%	2.08%
attempted-user	0.80%	1.98%	1.13%	3.07%	4.88%	5.60%	2.43%	1.46%	6.51%	1.31%	0.96%	19.89%	1.00%	7.75%
default-login-attempt								0.01%						
denial-of-service			0.01%	0.01%			0.02%	0.02%	0.01%		0.01%		0.02%	1.11%
misc-activity		0.03%												
misc-attack	0.02%	0.03%	0.04%	0.08%			0.10%	0.03%	0.01%	0.07%		0.06%	0.02%	0.06%
network-scan	0.05%	0.24%	0.04%	0.06%			0.05%	0.05%	0.05%	0.22%	0.18%	0.04%	0.05%	
non-standard-protocol	0.89%	0.97%	0.33%	0.37%	2.79%	2.63%	0.38%	1.31%	0.40%	0.62%	0.27%	0.24%	1.07%	3.38%
policy-violation	1.43%	0.95%	1.80%	1.81%	1.16%	1.10%	1.15%	1.76%	1.31%	1.20%	2.43%	0.87%	1.41%	1.49%
trojan-activity		0.01%												
web-application-activity	1.01%	1.09%	1.39%	1.03%	1.05%	0.99%	1.22%	1.28%	1.00%	0.95%	0.96%	0.87%	1.12%	1.62%
web-application-attack	0.89%	1.55%	1.06%	1.40%			1.86%	1.97%	1.77%	1.37%	1.34%	1.17%	1.49%	3.43%
Grand Total	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

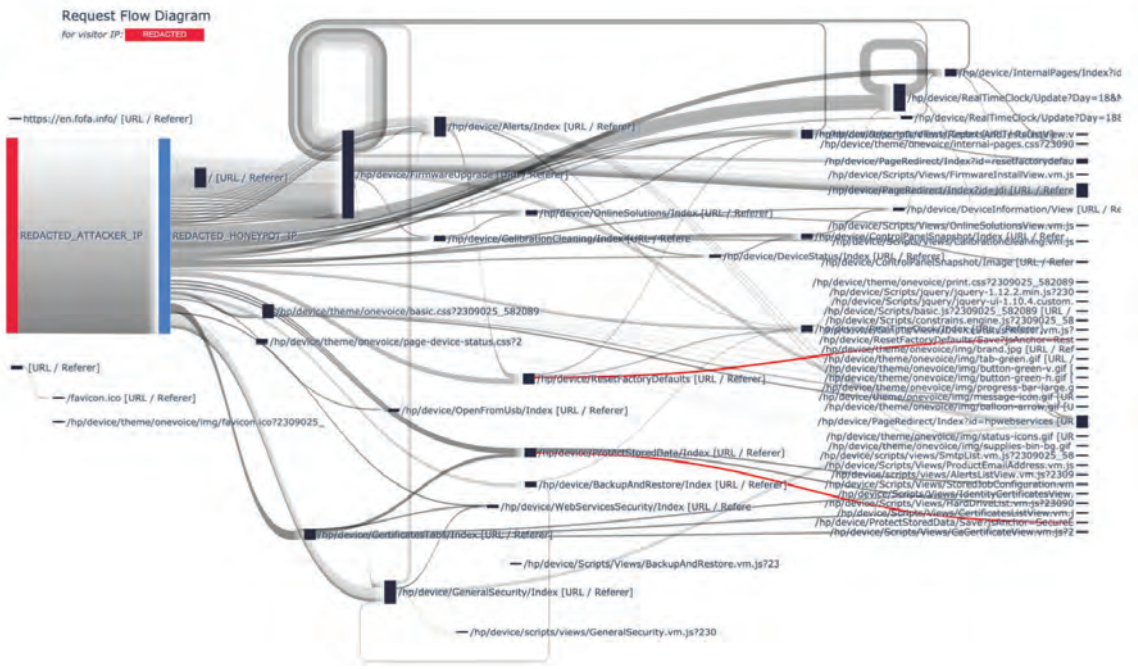
SNORT message	Honeypot technology: VMware vCenter Server 6.5
SERVER-APACHE Oracle WebLogic Apache Connector buffer overflow attempt	611
SERVER-WEBAPP VMWare vSphere Client remote code execution attempt	610
SERVER-IIS Microsoft ASP.NET bad request denial of service attempt	122
SERVER-WEBAPP Microsoft Exchange Server remote code execution attempt	69
SERVER-WEBAPP apache directory disclosure attempt	26
SERVER-WEBAPP Microsoft Exchange autodiscover server side request forgery attempt	23
SERVER-WEBAPP VMware vCenter Server file upload attempt	23
SERVER-WEBAPP Pulse Secure SSL VPN arbitrary file read attempt	5
SERVER-WEBAPP VMware vRealize Operations Manager SSRF attempt	4
SERVER-APACHE Apache Struts wildcard matching OGNL remote code execution attempt	3

Count of Source IP: 1 – 611

Top 10 vulnerabilities, unique to the VMware vCenter honeypot

LTFE

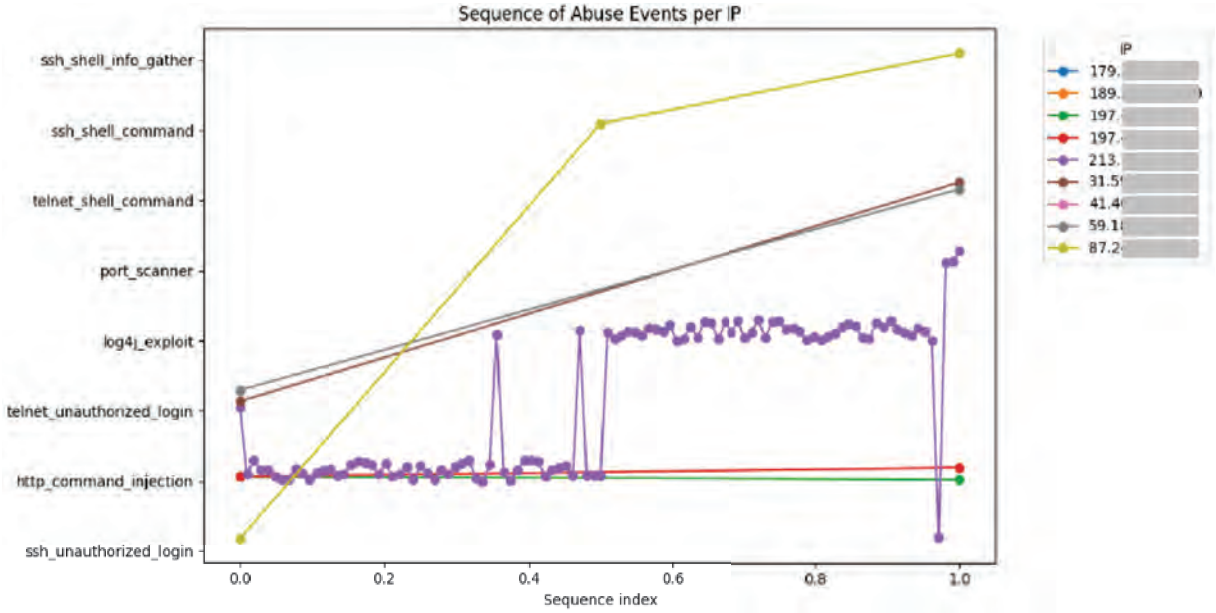
Primer: zaznava škodljivih zahtev



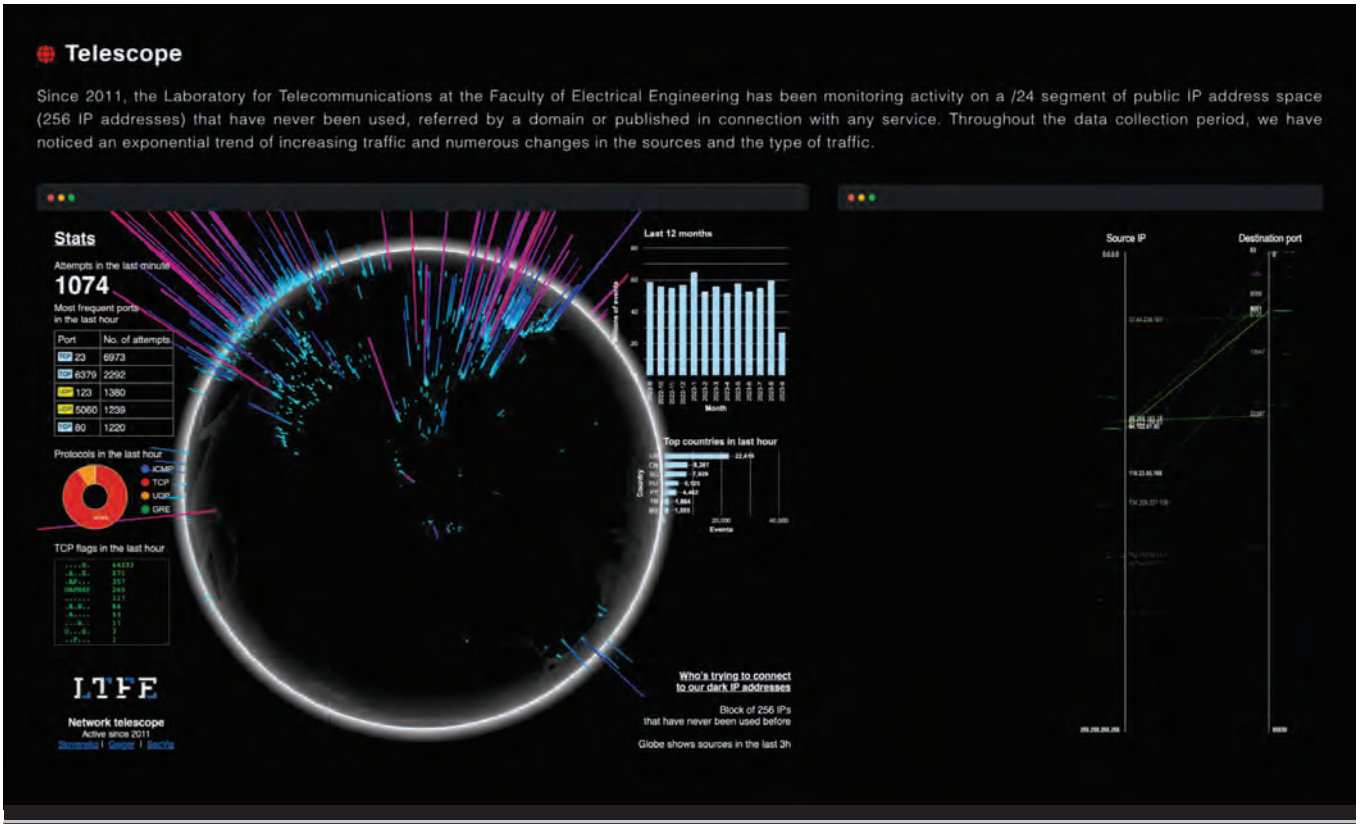
© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko

LTFE

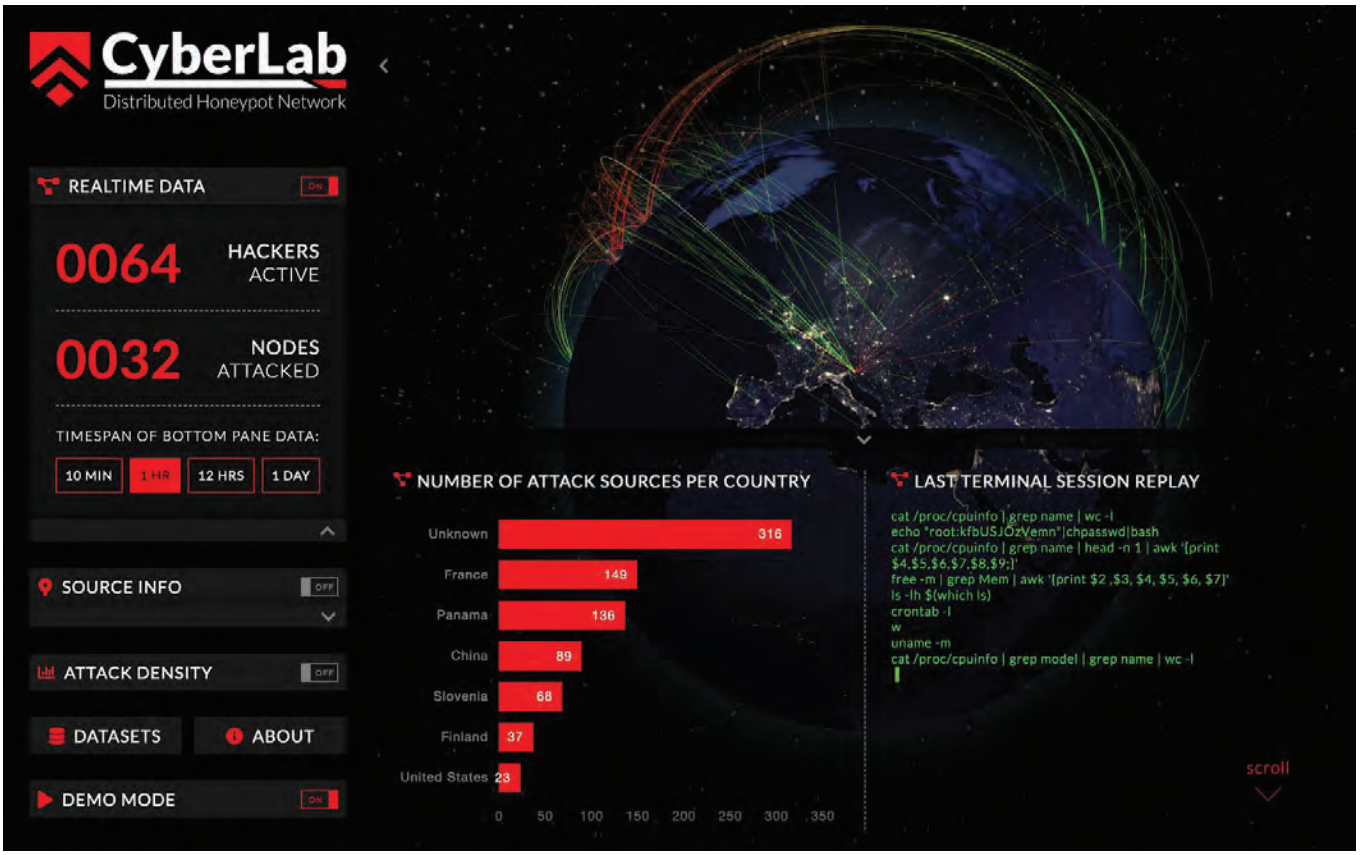
Primer: sekvenca dejanj akterjev

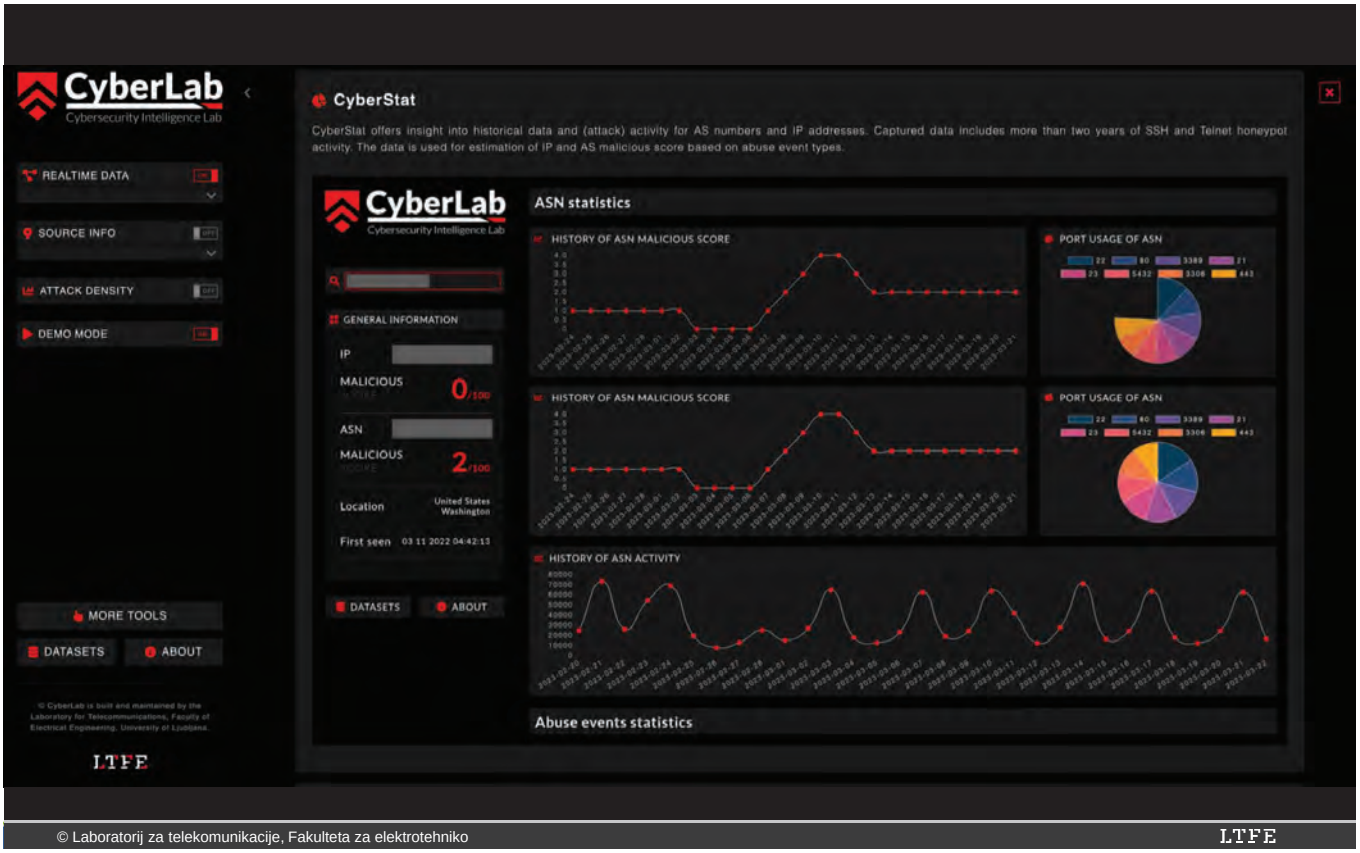


© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

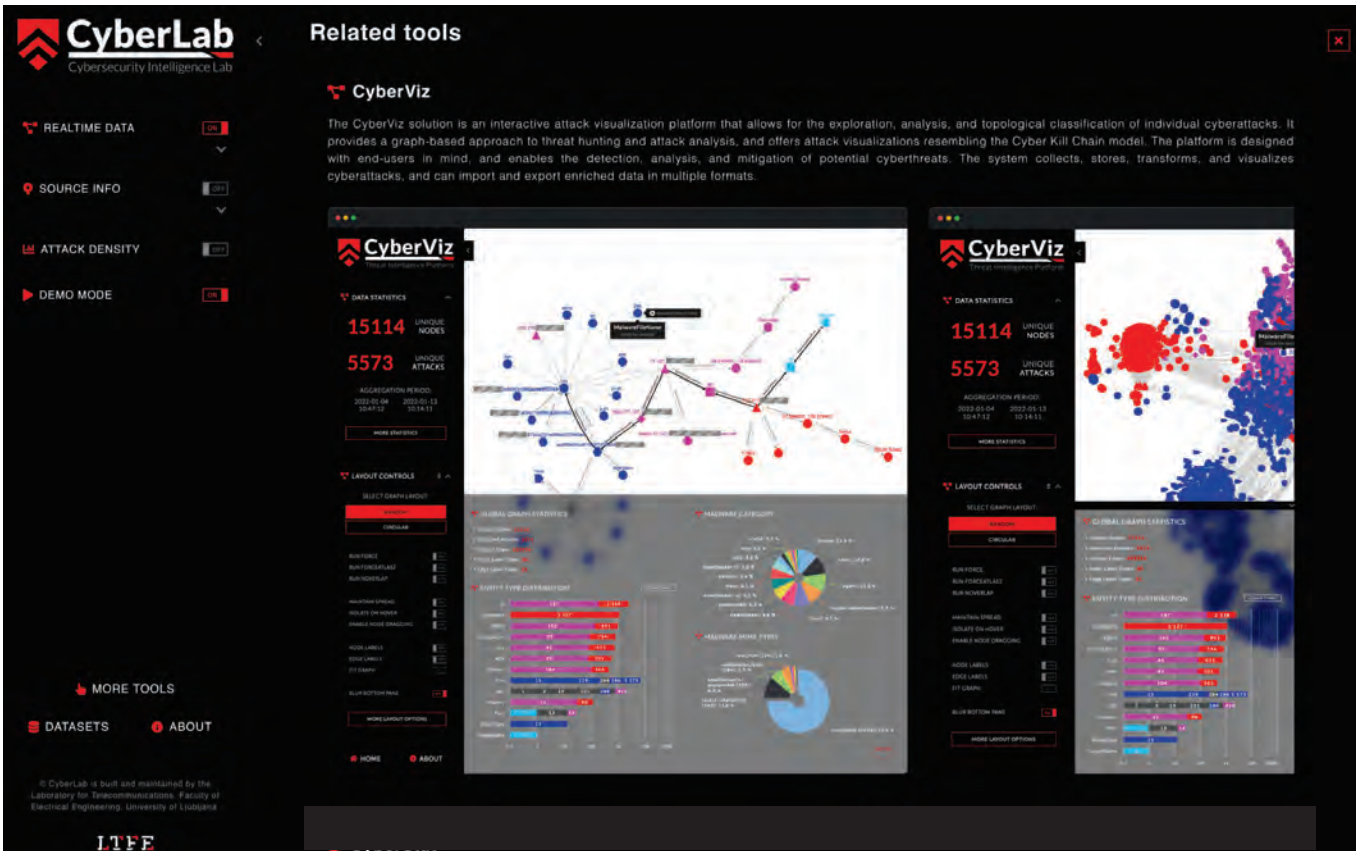


© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE



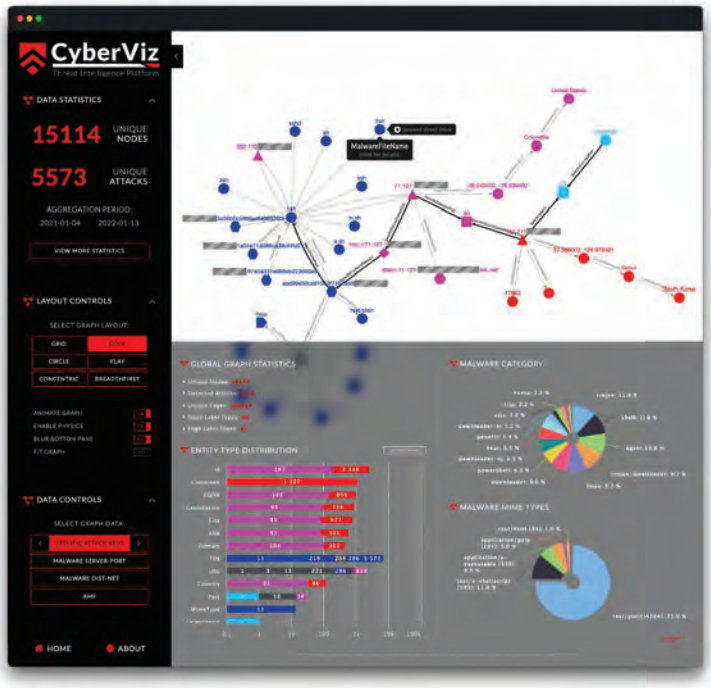


© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE



Orodje CyberViz

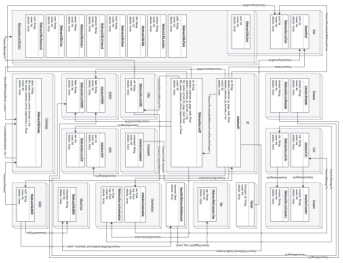
- Preslikava entitet napadov verižnemu modelu dejanj Cyber Kill Chain
  - entitete – vozlišča
  - dejanja – robovi
  - metapodatki – lastnosti
- Dekompozicija postopkov in dejanj napadalcev
- Obširna statistika škodljivih datotek

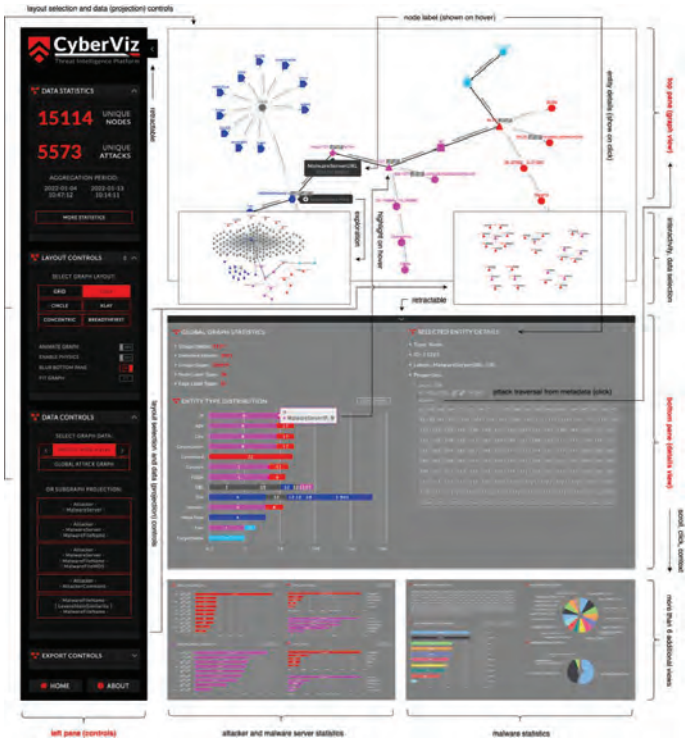


© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Orodje CyberViz

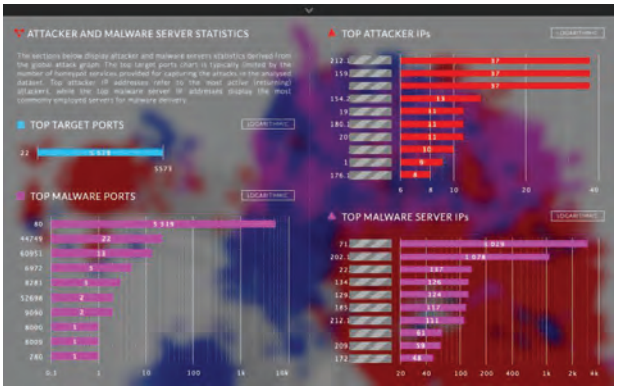
- Interaktivno orodje za eksploracijo podatkov in analizo napadov
- Omogoča razkrivanje prikritih povezav med entitetami (GP, GDS)

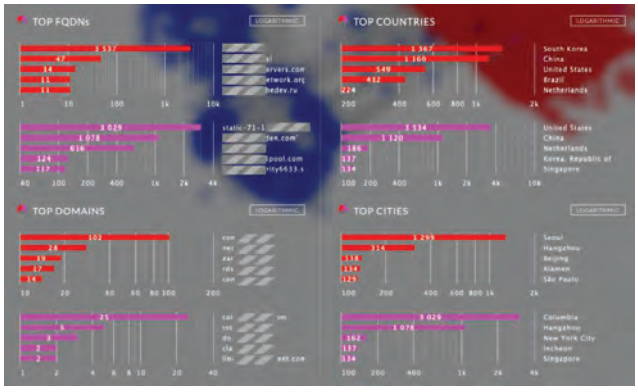


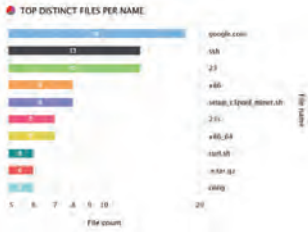


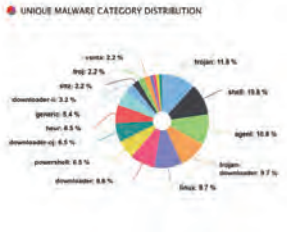
© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

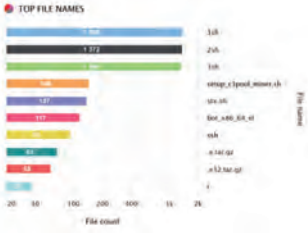
Vizualizacije in analitika







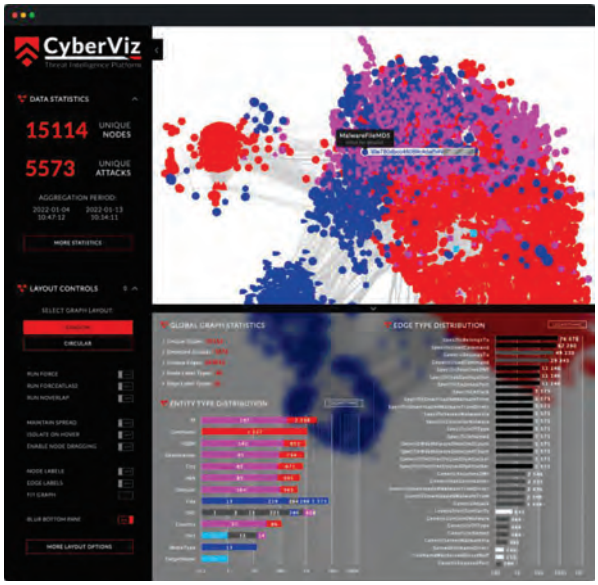


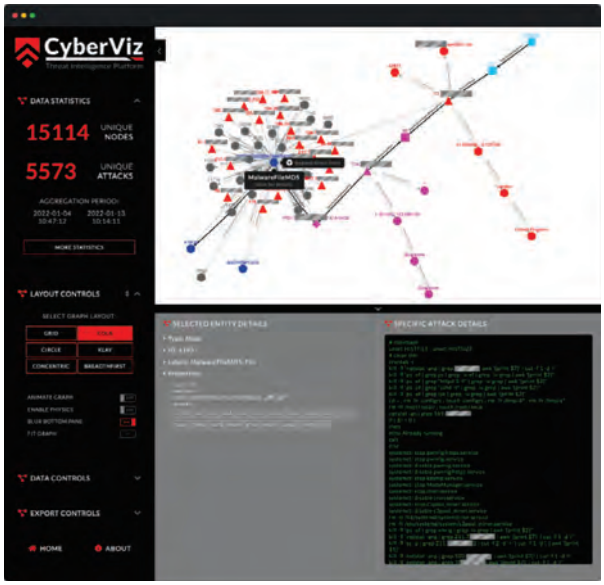


© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Vizualizacije in analitika

- Globalni graf napafov
- Razkrivanje relacij, iskanje, filtriranje

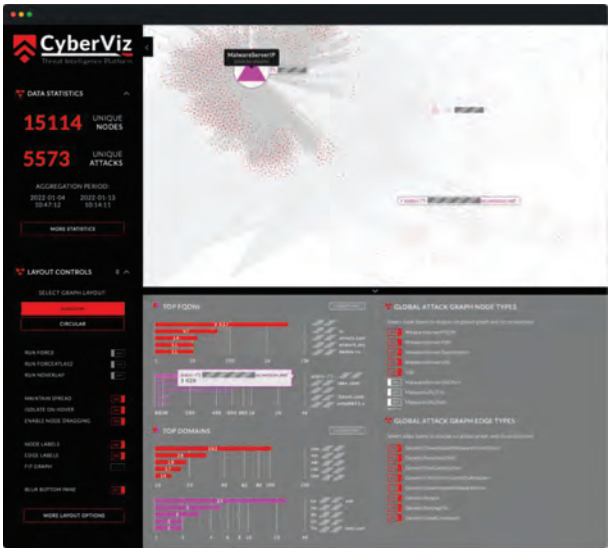


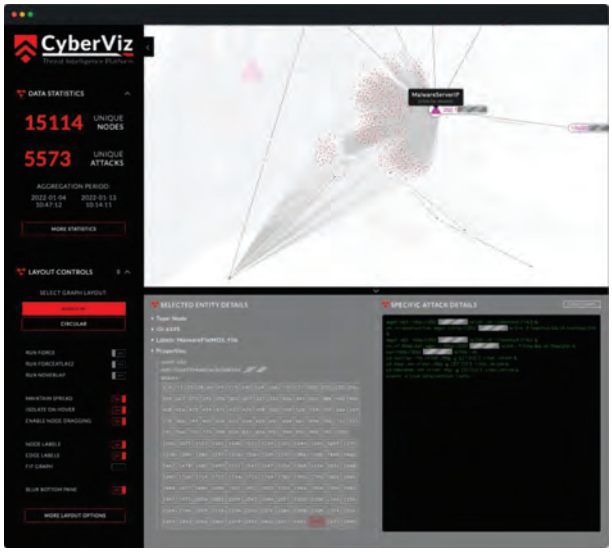


© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Vizualizacije in analitika

- Drill-down analiza
- Ogled celotne seje napada

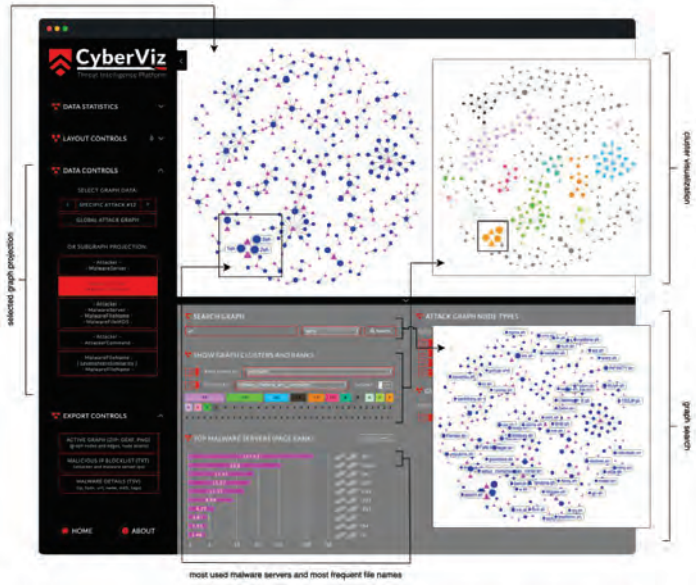




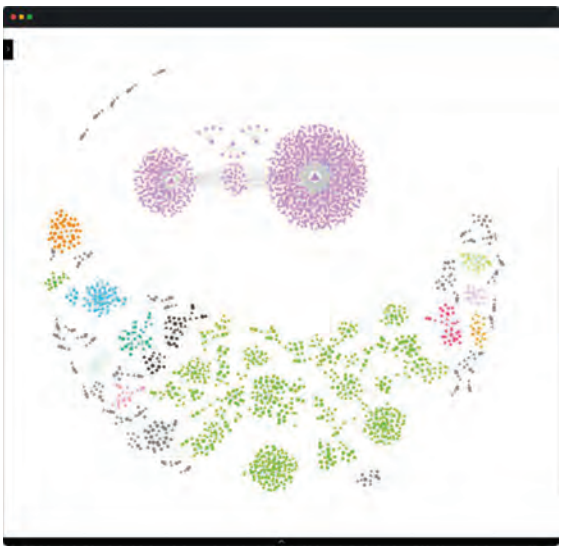
© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Vizualizacije in analitika

- Podprojekcije izbranih razmerij



- Botnet detection



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

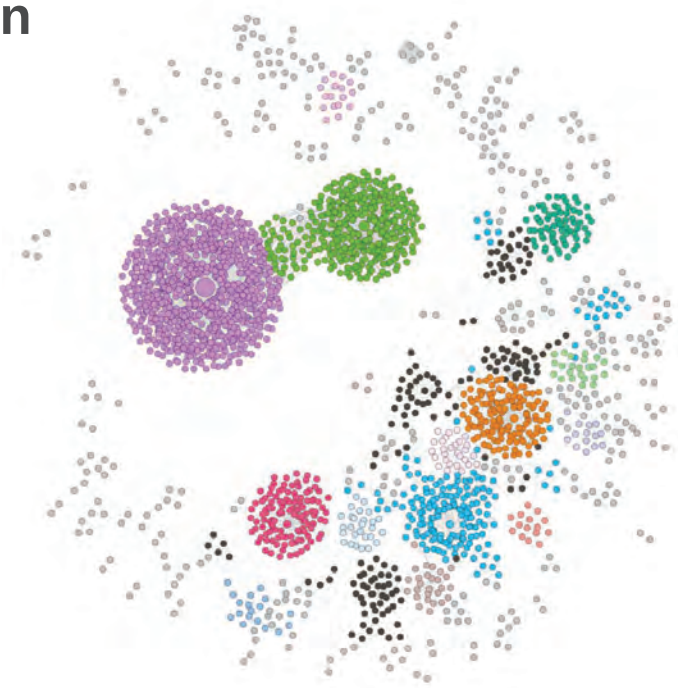
Primer: botnet detection

- Iskanje povezanih skupnosti s podprojekcijo globalnega grafa napadov s porazdeljenih senzorjev
  - Clustering na podlagi skupnih lastnosti
  - Najpreprostejši primer z uporabo virtulanih relacij:
    - AttackerServerIP
    - MalwareServerIP

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Primer: botnet detection

- Avtomatizacija procesa z uporabo GDS algoritmov
  - na podlagi parametrov (IP reputation, ipd.)
  - WCC/Louvain, PageRank



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Vabilo k sodelovanju

- Kibernetska varnost obrambnih sistemov in kritičnih infrastruktur
  - Financerji: MORS, URSIV in ARIS (ARRS)
  - Trajanje 3 leta
- Namen: poglobitev in razširitev naše platforme v smeri kritičnih infrastruktur in ponudnikov bistvenih storitev
  - Novi tipi senzorjev, glede na vaše potrebe
  - Novi postopki analitike, glede na vaše potrebe

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

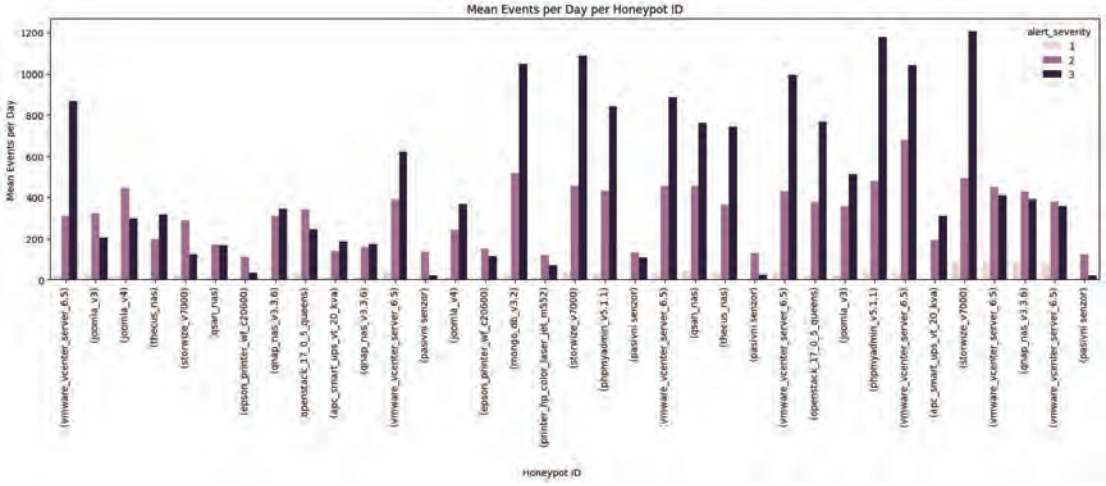
Vabilo k sodelovanju

- Lahko postanete del projekta
  - Brez obveznosti
- Možnost postavitve honeypota “pred vašimi vrati”
  - zbiranje podatkov o napadalih, ki jih vidi vsak vaš javen IP naslov*
    - * izven vaše mreže s kritičnimi storitvami, brez nevarnih premostitev

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Kaj ponujamo v projektu

- Vašim analitikom posredujemo analize
  - Primerjava z drugimi sektorji in relativna izpostavljenost
  - Poročilo o aktivnosti napadalcev, najpogostejših napadih, izvorih, ipd.



© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Kaj ponujamo v projektu

- Vašim analitikom posredujemo analize
  - Primerjava z drugimi sektorji in relativna izpostavljenost
  - Poročilo o aktivnosti napadalcev, najpogostejših napadih, izvorih, ipd.

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Kaj ponujamo v projektu

- API dostop do obogatenih podatkov
  - Iskanje agregiranih podatkov po IP naslovih
  - Ogled zgodovine aktivnosti napadalcev
  - Ogled zaznanih kategorizacij skenerjev
  - Whitelisted akterji, abuse history, TOR/VPN/proxy detection
  - Zaznana zlonamerna programska koda
  - Zgodovina HTTP/SSH/Telnet sej
  - Poskusi prijav z razlitimi podatki

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Trenutne namestitve

- Trenutno razpolagamo z
  - 256 temnimi IP naslovi (namensko raziskovalno omrežje)
  - 12 različnih tipov honeypotov (lastno omrežje)
  - 9 honeypotov (univerzitetno omrežje)
  - 20 honeypotov v javnem oblaku (4x Nemčija, 4x VB, 4x Japonska, 8x ZDA)

- V okviru projekta smo sodelovanje že razširili s
  - 3 honeypoti na IP naslovih ponudnika bistvenih storitev
  - 3 honeypoti na večjem omrežnem operaterju

- V številkah
  - več TB podatkov, več milijard zajetih dogodkov
  - preko 20 podsistemov za analizo

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Povežimo se

- Kaj pridobite?
  - Dostop do najnovejših raziskav in inovacij
  - Dostop do podatkov oz. poročil za vaš segment

- Kaj iščemo?
  - Ponudnike storitev, ki so pripravljeni deliti svoje izkušnje in izzive
  - Kandidate za postavitev senzorjev

- Kontaktirajte nas na:
  - matej.rabzelj@fe.uni-lj.si

© Laboratorij za telekomunikacije, Fakulteta za elektrotehniko LTFE

Upravljanje energije primarnih in sekundarnih porabnikov v hladilnih sistemih s pomočjo edge IoT sistema ReCool

Energy management of primary and secondary loads in refrigeration systems using edge IoT ReCool system

Mark Umberger

ENTIA

POVZETEK

Sistem ReCool rešuje problem visoke porabe električne energije v trgovskih centrih s pomočjo upravljanja energije primarnih in sekundarnih porabnikov v hladilnih sistemih. Trgovski centri so v osnovi specifični objekti z zelo veliko porabo električne energije in sicer v povprečju med 400 in 600 kWh/m2 na leto. Glavni porabnik je hladilna tehnika oz. hladilni sistem za prehrambne izdelke, ki v povprečju predstavljajo 35%-55% porabe celotne električne energije. Rezultati raziskav so pokazali, da poraba energije v trgovskih centrih predstavlja 2%-4% celotne porabe električne energije v posamezni državi. Sistem ReCool je sestavljen iz večih tehnoloških sklopov in sicer strojne opreme za upravljanje hladilne tehnike, programske opreme za podporo komunikaciji med hladilnimi sistemi na eni strani ter aplikacijami in algoritmi za upravljanje hladilne tehnike na drugi strani ter uporabniške aplikacije za izvajanje nadzorovanja in upravljanja sistema ReCool. V osnovi s sistemom ReCool upravljamo primarne in sekundarne porabnike elektrike v trgovskih centrih. Primarni se nanašajao na hladilni sistem, sekundarni pa na ostale porabnike v hladilnikih (razsvetljava, grelniki ASH, ventilatorji, itd.). Meritve so pokazale, da sekundarni porabniki porabijo celo do 30% porabe električne energije za hladilnike. Rezultati pilotnega preizkusa delovanja sistema ReCool v pilotnem supermarketu so pokazali, da je bil celoten prihranek električne energije z uporabo sistema ReCool 7,42%, - od tega je prihranek pri upravljanju sekundarnih porabnikov znašal 33,7%, medtem ko je največji prihranek električne energije na posameznem porabniku dosežen pri upravljanju grelnikov ASH (84,6 %). Na letni ravni bi tako lahko, ob predpostavki doseganja teoretično dokazanih prihrankov tudi iz naslova upravljanja primarnih porabnikov (v povprečju približno 7 %), v supermarketu z uporabo sistema ReCool prihranili več kot 14% električne energije. To bi v pilotnem supermarketu pomenilo nižjo porabo oziroma nižje stroške električne energije v obsegu 1.726 kWh tedensko ali 89.748 kWh letno, kar po takratnih cenah električne energije za komercialne porabnike (0,107 po kWh) pomeni prihranek v višini 9.603 EUR/letno.

SUMMARY

The ReCool system addresses the issue of high electricity consumption in shopping centers by managing the energy usage of primary and secondary loads in refrigeration systems. Shopping centers are fundamentally specific buildings with very high electricity consumption, averaging between 400 and 600 kWh/m2 per year. The main consumer is refrigeration system for food products, which typically accounts for 35% to 55% of the total electricity consumption. Research results have shown that the energy consumption in shopping centers represents 2% to 4% of the total electricity consumption in each country. The ReCool system consists of several technological components, namely hardware for managing refrigeration system, software for supporting communication between refrigeration systems on one side and applications and algorithms for managing refrigeration system on the other side, and user applications for implementing control and management of the ReCool system. In essence, the ReCool system manages primary and secondary electricity consumers in shopping centers. Primary consumers refer to the refrigeration system, while secondary consumers refer to other consumers in refrigerators (lighting, ASH heaters, fans, etc.). Measurements have shown that secondary consumers can account for up to 30% of electricity consumption for refrigerators. Results from the pilot test of the ReCool system in a pilot supermarket showed that the total electricity savings using the ReCool system were 7.42%. Of this, savings in managing secondary consumers amounted to 33.7%, while the largest electricity savings on individual consumers were achieved in managing ASH heaters (84.6%). On an annual basis, assuming theoretically proven savings are achieved also from managing primary consumers (on average approximately 7%), the use of the ReCool system in a supermarket could save more than 14% of

electricity. In the pilot supermarket, this would mean lower electricity consumption or lower electricity costs totaling 1,726 kWh weekly or 89,748 kWh annually, which at the then-current electricity prices for commercial consumers (€0.107 per kWh) translates to savings of €9,603 annually.

O AVTORJU

dr. Mark Umberger je ustanovitelj in direktor podjetja Entia. Svet avtomatizacije in digitalizacije zgradb ga je pritegnil že med študijem na Fakulteti za elektrotehniko v Ljubljani, kjer je iz tega področja tudi diplomiral in doktoriral. Danes vodi visokotehnološko podjetje Entia, ki pod svojo blagovno znamko ENTIAliving razvija in proizvaja strojno in programsko opremo za avtomatizacijo in digitalizacijo procesov v zgradbah. Z lastnimi rešitvami so uspešno opremili že več kot 2000 objektov v 17 različnih državah.

ABOUT THE AUTHOR

dr. Mark Umberger is the founder and CEO of Entia. The world of automation and digitization of buildings attracted him already during his studies at the Faculty of Electrical Engineering in Ljubljana, where he also received a bachelor's degree and a PhD in this field. Today, he heads the high-tech company Entia, which under its ENTIAliving brand develops and produces hardware and software for the automation and digitization of processes in buildings. Th company has already successfully equipped more than 2,000 buildings in 17 different countries with its own solutions.

Upravljanje energije primarnih in sekundarnih porabnikov v hladilni tehniki s pomočjo Edge IoT sistema ReCool

Mark Umberger
CEO

entia Living

Povzetek

- ***O podjetju***
  - Entia d.o.o., Dunajska 167 1000 Ljubljana, Slovenia, www.entia.si, prodaja@entia.si
- ***Produkt***
  - ENTIAliving – sistem za pametni dom
  - Kompatibilen z multimedijo Sonos, Apple HomeKIT, Google Home in Amazon Echo
  - ReCool – energetsko učinkovito upravljanje hladilnikov v trgovskih centrih

works with Google Home
Works with Apple HomeKit
Works with amazon alexa
Works with SONOS

- ***Zakonodaja za pametni dom PURES***
  - Pravilnik o učinkoviti rabi energije v stavbah
  - Evropski standard za pametni dom *SRI – Smart Readiness Indicator, kjer Entia dosega vse točke*
  - Vse točke po standardu SRI omogočajo in povečajo vrednost nepremičnin na dolgi rok

- ***Največji tržni delež v Sloveniji***
  - Na trgu že od leta 2009 v Sloveniji
  - Hrvaška: na trgu od leta 2021, uspešno zaključili večji projekt za VMD Group
  - Reference v tujini: Hrvaška, Avstrija, Latvija, Romunija, Anglija, Kurdistan, Bangladeš, Irak, Gruzija
  - Več kot 2.000 zaključenih pametnih domov

entia Living

Poraba električne energije po segmentih

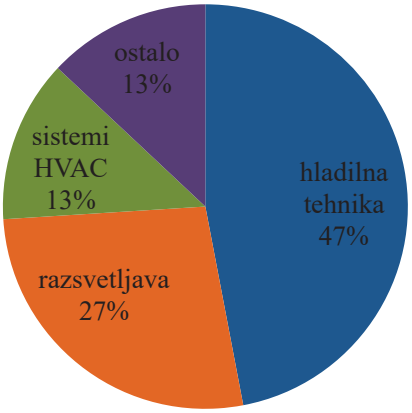
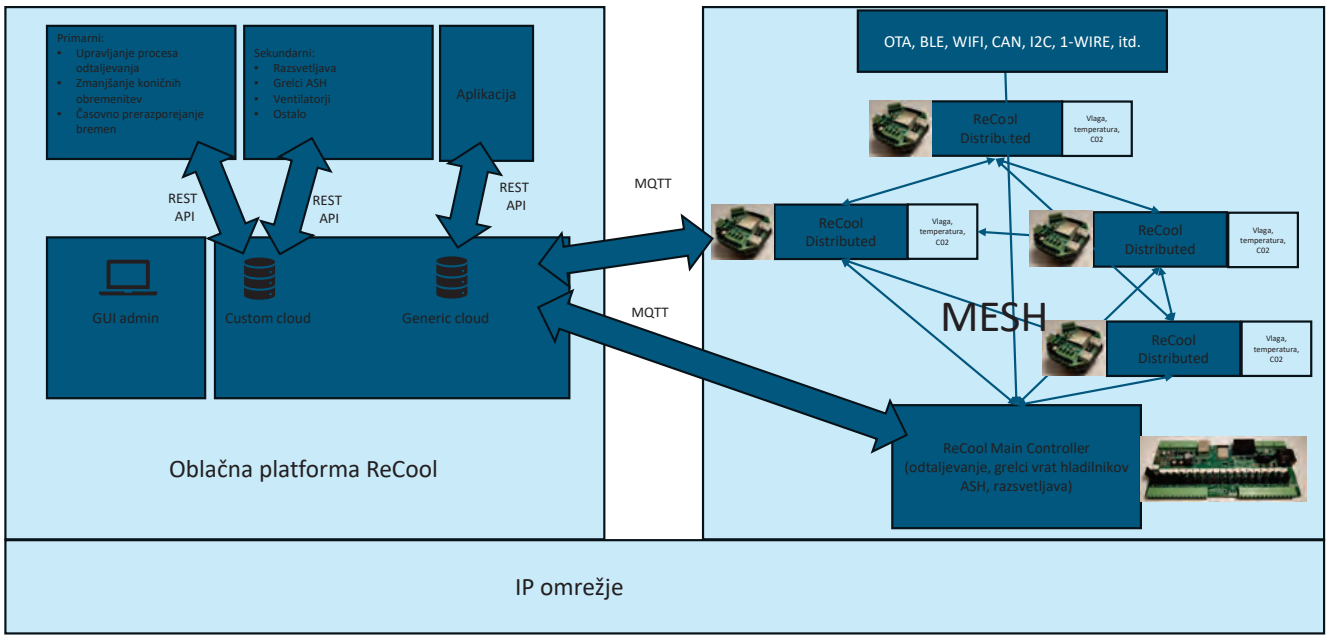


Diagram: Sestava porabe EE v enem od nakupovalnih centrov na Švedskem

entia Living

ARHITEKTURA SISTEMA RECOOL



entia Living

RECOOL NAPRAVE

ReCool Main Controller

- 20 x triac aktuatorjev za neposredno krmiljenje manjših (primarnih in sekundarnih) porabnikov hladilne tehnike do 230 V/1A;
- 2 x relejska aktuatorja za krmiljenje večjih (primarnih in sekundarnih) porabnikov hladilne tehnike do 230 V/10 A ali kontaktorjev za porabnikov brez omejitev;
- 2 x možnost priklopa zunanjih razširitvenih IO modulov (npr. 2 x 16 triacov, 2 x 8 relejev) z enakovredno funkcionalnost, kot je predstavljena zgoraj;
- 1 x DALI krmilnik za krmiljenje razsvetljave hladilne tehnike (prižiganje, ugašanje, prilagajanje jakosti = dimanje svetlobnih virov)
- 1 x RS485 vmesnik s podporo protokolu MODBUS za upravljanje kompresorjev hladilne tehnike in zajem parametrov iz njih
- 2 x DAC vmesnik za pretvorbo digitalnih signalov v analogne (0 .. 12 V)
- 3 x ADC vmesnik za pretvorbo analognih signalov v digitalne
- 1 x RTC z rezervnim napajanjem (kondenzator)
- 1 x osem bitna nastavitev naslova modula preko dveh rotary switch-ev
- komunikacija do oblaka preko Ethernet LAN povezave (RJ45 vmesnik)
- komunikacija do oblaka preko WiFi povezave z dostopno točko
- komunikacija z drugimi moduli preko MESH WiFi omrežja
- komunikacija z drugimi moduli preko vodila CAN
- opcijsko: komunikacija preko BLE in LORA
- opcijsko: komunikacija GPIO in I2C za potencialne dodatne razširitve

ReCool Distributed

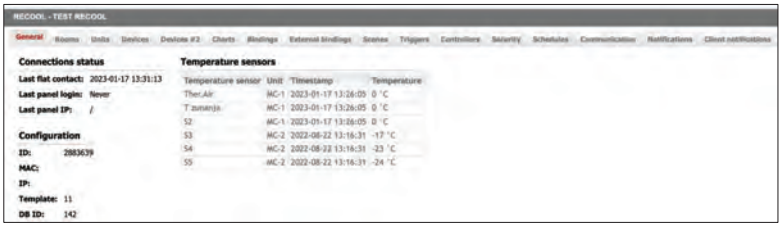
- 4 x triac aktuatorji za neposredno krmiljenje manjših (primarnih in sekundarnih) porabnikov hladilne tehnike do 230 V/1A;
- 4 x digitalni vhodi (npr. tipkala za upravljanje s hladilno tehniko)
- 1 x digitalni vhod za senzor gibanja PIR
- IR izhod za opcijsko upravljanje s klima napravo
- OneWire / DHT22 vhod za senzor temperature, temperature + vlage
- 6-bitna nastavitev naslova modula preko enega rotary switch-a in dveh DIP switch-ev
- komunikacija z drugimi moduli preko MESH WiFi omrežja

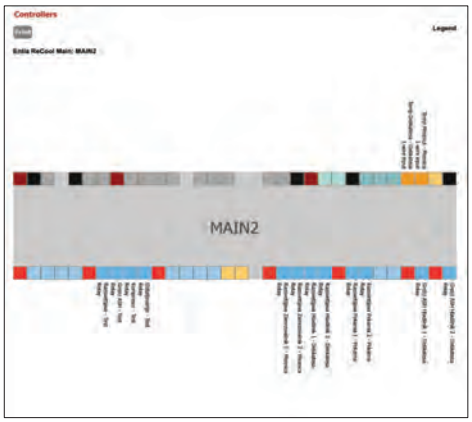
- opcijsko: komunikacija z drugimi moduli preko vodila CAN
- opcijsko: komunikacija preko BLE in LORA
- opcijsko: komunikacija do oblaka preko WiFi povezave z dostopno točko
- opcijsko: komunikacija I2C za potencialne dodatne razširitve

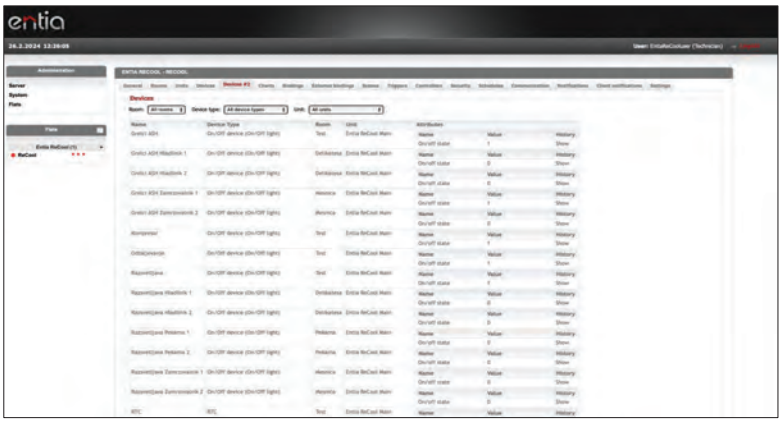
entia Living

RECOOL CUSTOM CLOUDE

- Agregacija podatkov
- Oddaljena konfiguracija sistema
- Povezljivost s fieldbus-om
- Monitoring oz. nadzor nad sistemom



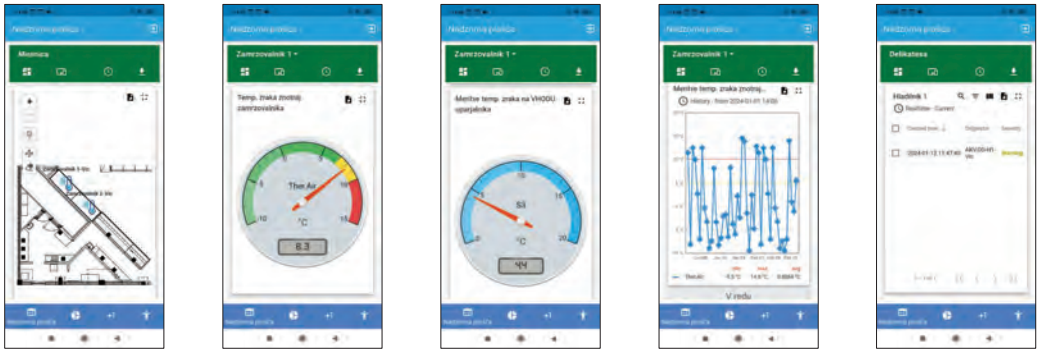




entia Living

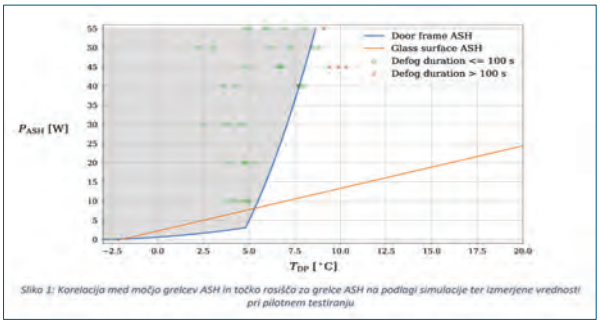
RECOOL GENERIC CLOUDE / MOBILE APP

- Analitika in hramba podatkov (BI/DWH)
- Izvajanje algoritmov
- Operativna logika
- Poslovna logika (BPM)
- Kibernetska varnost

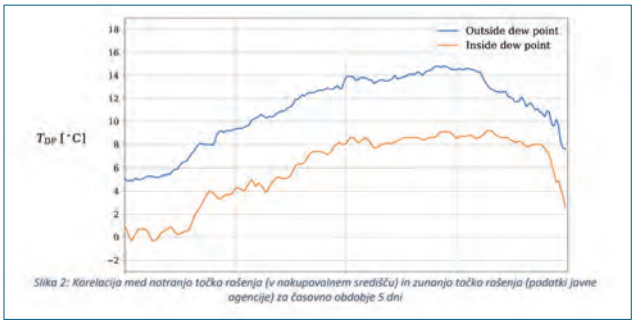


entia Living

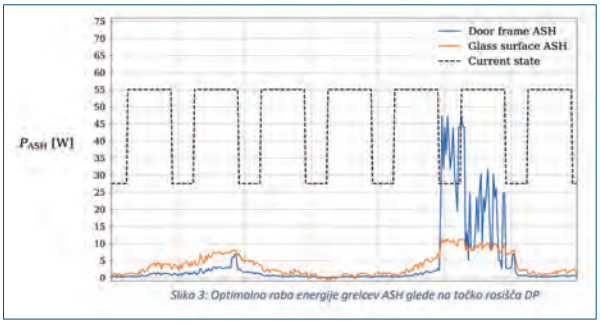
REZULTATI



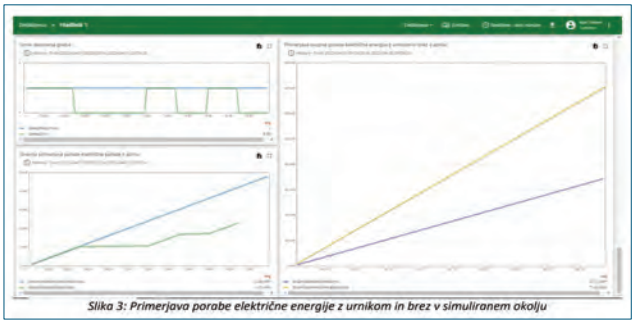
Slika 1: Korelacija med močjo grelcev ASH in točko rosišča za grelce ASH na podlagi simulacije ter izmerjene vrednosti pri pilotnem testiranju



Slika 2: Korelacija med notranjo točko rosenja (v nakupovalnem središču) in zunanjo točko rosenja (podatki javne agencije) za časovno obdobje 5 dni



Slika 3: Optimalna raba energije grelcev ASH glede na točko rosišča DP



Slika 3: Primerjava porabe električne energije z urnikom in brez v simuliranem okolju

Pomen poenotenja in enakega razumevanja podatkovnih zapisov ter njihova zaščita

Data Harmony: Achieving Consistency and Security Across Records (Data Standardization as a Foundation)

Mateja Sajovic

Agencija Republike Slovenije za javnopravne evidence in storitve

POVZETEK

Prezentacija izpostavlja ključno vlogo standardizacije, zaščite in upravljanja podatkov v zdravstvu ter širše. Poudarek je na umestitvi baze podatkov v ločen nivo arhitekture informacijskega sistema, ločen od aplikacij, kar omogoča bolj nadzorovano upravljanje podatkov.

V uvodu se citira Gartner Group, ki tudi izpostavlja nezadovoljstvo z monolitnimi sistemi v zdravstvu ter poudarja potrebo po hitrem odzivu na nove klinične in regulatorne zahteve. Trenutno stanje opisuje razdrobljenost monolitnih aplikacij, izzive ogromnega in težko obvladljivega števila integracij, pomanjkanje standardizacije podatkov ter težave s posodabljanjem obstoječih informacijskih sistemov.

Predstavljeni cilji vključujejo potrebo po zmanjšanju odvisnosti od obstoječih izvajalcev, vzpostavitev standardiziranih zapisov zdravstvenih podatkov, vključno z zdravstvenim kartonom, pomen spodbujanja inovacij ter omogočanja enostavnejšega sodelovanja med timi.

Arhitekturna umestitev podatkov v posebnem nivoju, ločenem od aplikacij, omogoča med drugim bolj pregledno upravljanje podatkov s strani internih IT ekip, večji nadzor nad dostopi in varovanjem podatkov ter nenazadnje enostavnejše dodajanje aplikacij in tudi zamenjave obstoječih aplikacij z bolj sodobnimi. Modularna arhitektura v predstavljenem bolnišničnem ekosistemu bi omogočala neodvisno upravljanje s podatki, uporabo API vrat za množico mikroservisov ter integracijski bus za izmenjavo podatkov med sistemi. Poudarek je na prehodu iz modela razvoja po naročilu v nakup namenskih procedur ter optimizaciji platform za bolj nevtralno vključevanje novih in obstoječih (starih) sistemov.

Razdelek o zaščiti podatkov poudarja pomen stroge politike dostopa, rednega usposabljanja osebja, nadzora in revizije ter sodelovanja z izkušenimi eksperti za informacijsko varnost. Opisani so tudi primeri napadov z izsiljevalsko programsko opremo, ki spet pokažejo nujnost zaščite podatkov v informacijskih sistemih in tudi nujnost pripravljenosti na ukrepanje v primerih, ko pride do napadov.

Pomen osredotočanja na podatke je opisan za zdravstvene sisteme, opisane arhitekturne paradigme pa so pomembne na vseh kritičnih področjih, od nacionalnih baz podatkov poslovnih subjektov ali državljanov do vseh ostalih registrov nacionalnega pomena.

SUMMARY

The presentation underscores the pivotal role of standardization, protection, and data management within healthcare and beyond. It emphasizes the separation of the database into a distinct layer within the information system architecture, independent of applications, facilitating more streamlined data management.

The opening citation from the Gartner Group echoes concerns over monolithic healthcare systems and underscores the imperative of promptly adapting to new clinical and regulatory demands. The current landscape delineates challenges stemming from fragmented monolithic applications, the complexities of numerous and unwieldy integrations, the absence of standardized data, and hurdles in updating existing information systems.

Outlined objectives include mitigating vendor-lock scenarios, establishing standardized health data records, fostering innovation, and facilitating seamless collaboration. By situating data within a dedicated layer separate from applications, the architecture enables enhanced data oversight by internal IT teams, greater access control, robust data protection, and facilitates the seamless integration of new applications and the gradual replacement of legacy ones.

The modular architecture within the hospital ecosystem enables autonomous data management, leveraging API ports for myriad microservices, and an integration bus for seamless data exchange between systems. The focus is on transitioning from bespoke development models to procuring specialized procedures

and optimizing platforms for seamless integration of new and legacy systems.

The data protection segment underscores the necessity of stringent access policies, ongoing staff training, vigilant monitoring, auditing, and collaboration with seasoned information security experts. Instances of ransomware attacks underscore the imperative of robust data protection measures and preparedness to respond to cyber threats.

The narrative underscores the criticality of prioritizing data within healthcare systems and underscores the relevance of architectural paradigms across all critical domains, including national databases of businesses, citizens, and other nationally significant registries.

O AVTORJU

Mag. Mateja Sajovic je diplomirala na Fakulteti za računalništvo in informatiko Univerze v Ljubljani ter magistrirala na Fakulteti za organizacijske vede Univerze v Mariboru. Trenutno deluje kot CIO v Agenciji Republike Slovenije za javnopravne evidence in storitve (AJPES) ter je od leta 2014 naprej ocenjevalka projektov v Raziskovalni agenciji Evropske unije (REA) v Bruslju. S širokimi izkušnjami v vodenju informacijskih sistemov tako v gospodarstvu kot javni upravi, vključno z delom v Univerzitetnem kliničnem centru (UKC) Ljubljana, Zavodu za zdravstveno zavarovanje Slovenije (ZZZS) ter Javni agenciji za zdravila in medicinske pripomočke (JAZMP), je sposobna hitro prebrati situacijo, jo analizirati ter ustvariti celovito sliko. Njena motivacija izhaja iz priložnosti za razvoj in implementacijo sodobnih ter kompleksnih informacijskih sistemov.

ABOUT THE AUTHOR

Mateja Sajovic graduated from the Faculty of Computer and Information Science (University of Ljubljana) and obtained her master's degree from the Faculty of Organizational Sciences (University of Maribor). She currently serves as the Chief Information Officer (CIO) at the Agency of the Republic of Slovenia for Public Legal Records and Related Services (AJPES), and has been evaluating projects at the Research Executive Agency of the European Union (REA) in Brussels since 2014. With extensive experience in leading information systems in both the private and public sectors, including roles at the University Medical Centre Ljubljana (UKC), the Health Insurance Institute of Slovenia (ZZZS), and the Agency for Medicinal Products and Medical Devices of Slovenia (JAZMP), she excels in quickly grasping situations, analysing them, and forming a big picture. Her motivation stems from the opportunity to develop and implement modern and complex information systems.

Pomen poenotenja in enakega razumevanja podatkovnih zapisov in njihova zaščita

Data Harmony: Acieving Consistency and Security Across Records
(Data Standardisation as a Foundation)

Bled, 17.5.2024

Mateja Sajovic

mag. Mateja Sajovic

- **Trenutna vloga:**
  - CIO v Agenciji Republike Slovenije za javnopravne evidence in storitve (AJPES)
  - Ocenjevalka projektov SME v Raziskovalni agenciji Evropske unije (REA) v Bruslju
- **Izkušnje v področju zdravstva:**
  - V času povabila na to konferenco opravljanje dolžnosti CIO v Univerzitetnem kliničnem centru (UKC) Ljubljana
  - Izkušnje kot CIO v širšem področju zdravstva, vključno z Zavodom za zdravstveno zavarovanje Slovenije (ZZZS) in Javno agencijo za zdravila in medicinske pripomočke (JAZMP)
- **Strokovna področja:**
  - Izkušnje vodenja v različnih poslovnih okoljih javne uprave in gospodarstva, z osredotočenostjo na razvoj informacijskih sistemov in njihovo povezovanje
- **Motivacija:**
  - Navdihujejo me okolja, ki omogočajo razvoj, implementacijo in produkcijsko delovanje sodobnih in kompleksnih informacijskih sistemov.
- **Izobraževanje:**
  - Diplomirala na Fakulteti za računalništvo in informatiko, Univerza v Ljubljani
  - Magistrirala na Fakulteti za organizacijske vede, Univerza v Mariboru
  - PRINCE 2 Cert.

Pregled predstavitve

- Ključni del digitalizacije zdravstva so podatki
- Pomen standardov pri zapisu podatkov
- Modularnost aplikacij in večja možnost izbire izvajalcev informacijskih sistemov
- Primeri posodobitev informatizacije zdravstva v Evropi
- Zagotovitev zaščite in razpoložljivosti občutljivih osebnih podatkov

Podatki v zdravstvu

- Zdravstvo potrebuje podatke
- Zagotavljanje izmenjave podatkov ni dovolj
- Potrebno je zagotoviti **razpoložljivost podatkov**

„The **dissatisfaction** with **monolithic EHR** systems and their **inability to quickly respond** and support new clinical and regulatory **requirements** has never been more **obvious**.” (Gartner Group)

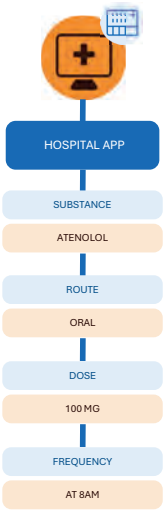
Različni zapisi istih podatkov

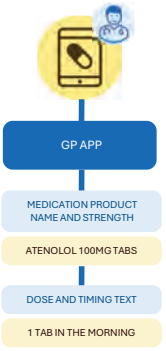
Cilji:

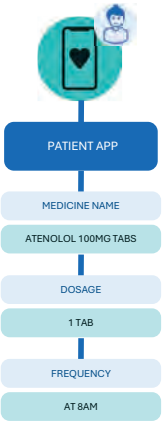
Zmanjšati odvisnost od obstoječih izvajalcev

Zmanjšati ovire za vstop novih IT izvajalcev

Vzpostavitev standardiziranega zdravstvenega kartona za celoten življenjski cikel pacienta







Cilji:

Spodbujanje inovacij

Omogočanje enostavnega sodelovanja med timi

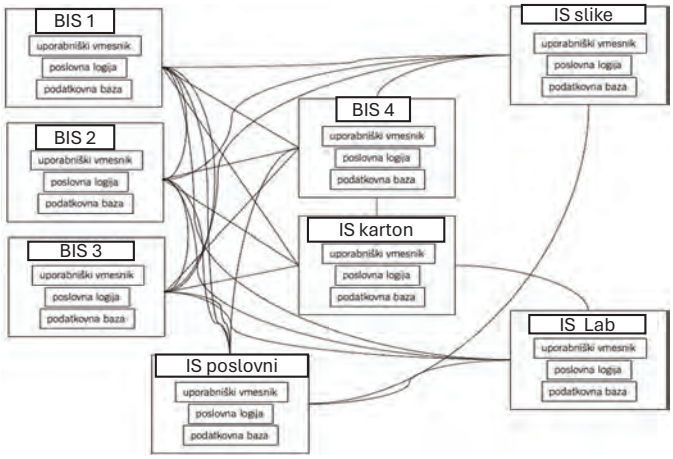
Poenostavitev sekundarne uporabe podatkov

Omogočiti pacientom kontrolo nad svojimi podatki

Današnje stanje

- Več različnih bolnišničnih/zdravstvenih IS
- Integracija aplikacij vsaka z vsako
- Večina IS ne uporablja (mednarodne) standardizacije podatkov
- Razvoj aplikacij po naročilu
- „Vendor lock-in“

Potrebna je tesnejša povezava med IS in IS na nacionalni ravni



Današnje stanje z drugega zornega kota

Ključni standardi pri zapisu podatkov

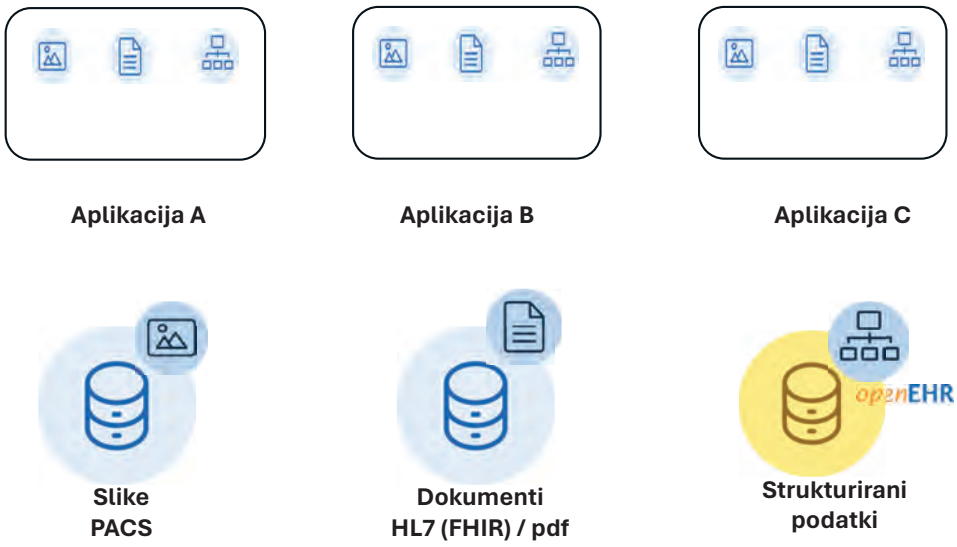
Cilj standardov:
Manjša odvisnost od ponudnika SW rešitve

Aplikacija A **Aplikacija B** **Aplikacija C**

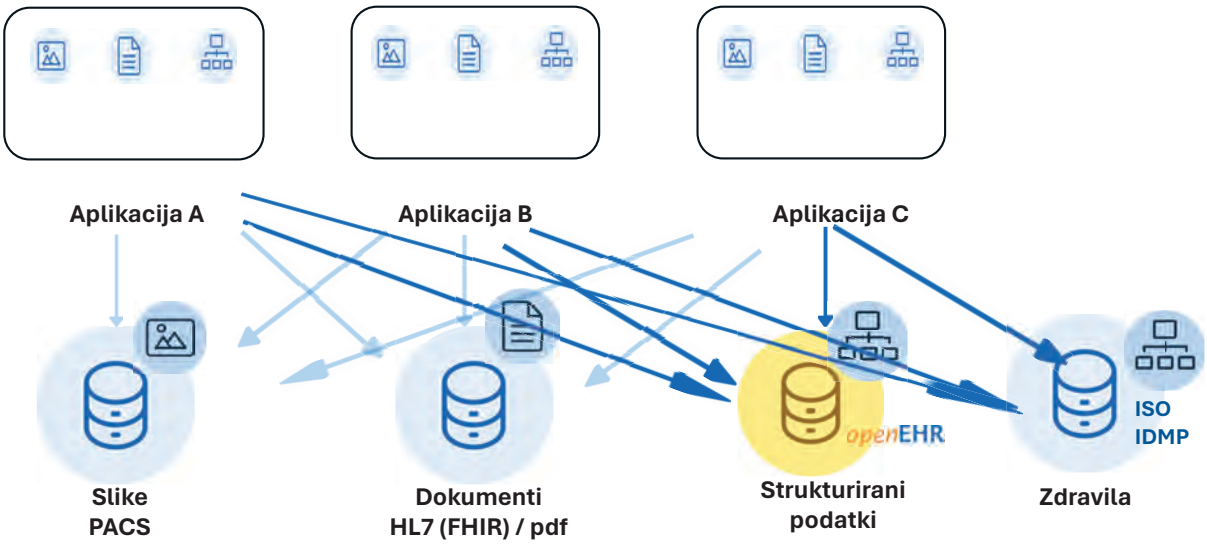
Cilj standardov:
Manjša odvisnost od ponudnika SW rešitve

Aplikacija A **Aplikacija B** **Aplikacija C**

Cilj standardov:
Manjša odvisnost od ponudnika SW rešitve



Cilj standardov:
Manjša odvisnost od ponudnika SW rešitve



Modularnost IS in večja izbira ponudnikov

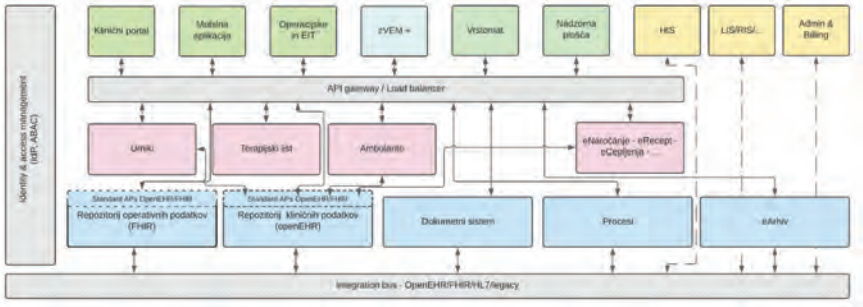
Ko govorimo o izjemnem pomenu podatkov, njihovi varnosti, upravljanju in uporabi, moramo razmišljati širše in iz vidika ciljem prilagojene arhitekture aplikacij

Cilji digitalizacije v zdravstvu

- Zagotavljati upravljanje s *podatki*, ki so enotno in standardizirano zapisani, v izključnem upravljanju izvajalcev zdravstvenih storitev in v bazi podatkov, ki je neodvisna od ponudnikov aplikacij.
- Zagotoviti sledljivost zdravljenja
  - Zagotoviti visoko kakovost in ažurnost podatkov
  - Omogočiti vpogled v potek zdravljenja (pacientom in svojcem)

  =>Digitalizirati obstoječe ključne in podporne procese (zdravljenje, oskrba kot hrana, zdravila, oblačila,…)
- Zagotoviti varno poslovanje
- Zagotoviti razpoložljivost podatkov in njihovo zaščito pred nepooblaščenimi dostopi
- Uvesti zdravstvene storitve s področja tele-medicine in digitalnega zdravja
- Zagotoviti varnost in kakovost obravnav
- Zagotoviti sporočanje opozorilnih nevarnih dogodkov (OND)
- Optimizirati stroške poslovanja

Primer ekosistema v slovenski bolnišnici



Modularna in razširljiva arhitektura:

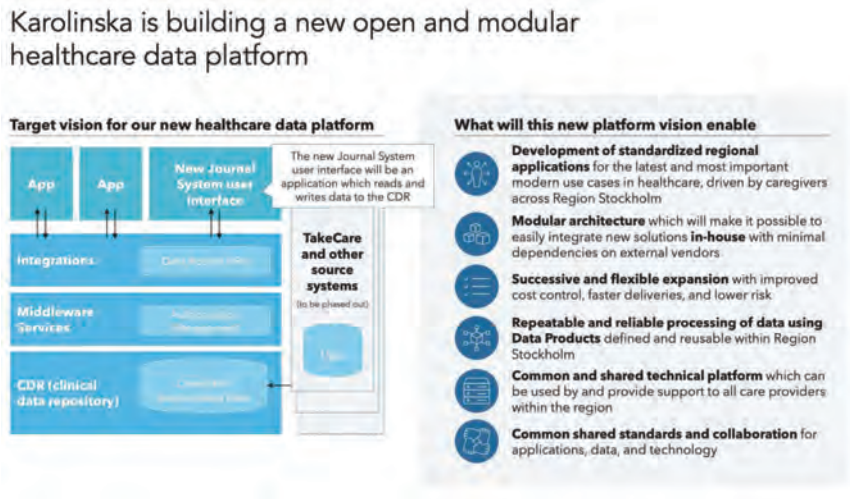
- Nivo podatkov, ki je ločen od aplikacij -> razširljivost, varnost podatkov, posodobitve brez vpliva na logiko aplikacij.
- API gateway -> vstopna točka za množico mikroservisov, upravljanje zahtev, pretok prometa, avtentikacija.
- Integration bus -> komunikacijska hrbtenica za izmenjavo podatkov med različnimi sistemi.

Prednosti modularne arhitekture - ekosistema

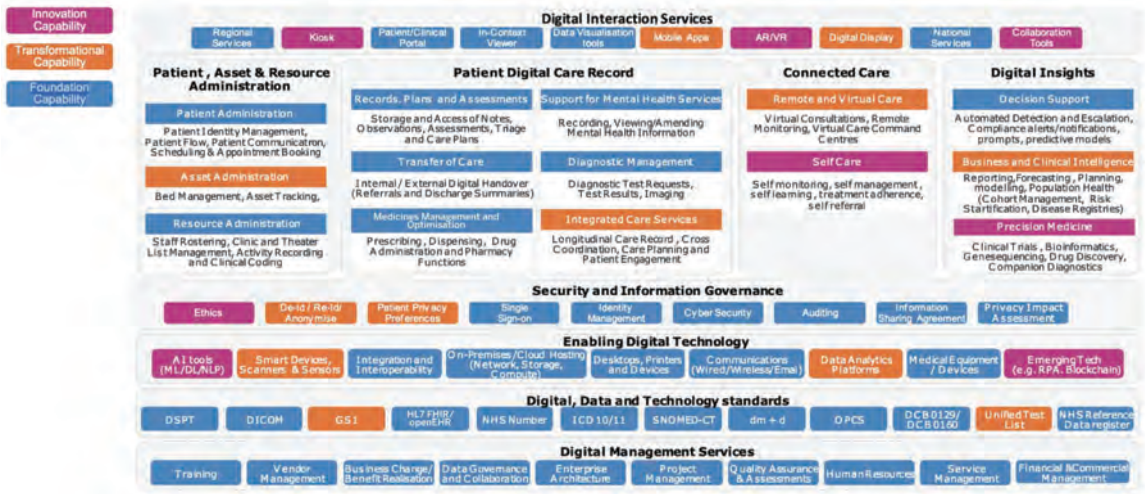
- Interna IT ekipa se fokusira na vsebine in upravljanje podatkov
- Prehod iz modela „razvoj po naročilu“ v model nakupa (ali razvoja) namenskih procedur
- Nov način razmišljanja o podatkih:
  - Podatki niso več “zaklenjeni” v izoliranih sistemih.
  - Poenotenje različnih informacij iz več virov osmisli njihov pomen.
  - Zajem in povezovanje vseh relevantnih podatkov.
- Optimalna platforma bo ločevala vsebino in tehnologijo ter bo postala bolj nevtralna glede prodajalca – vključevala bo lahko nove sisteme različnih izvajalcev in stare (legacy) sisteme.

Nekaj primerov v Evropi

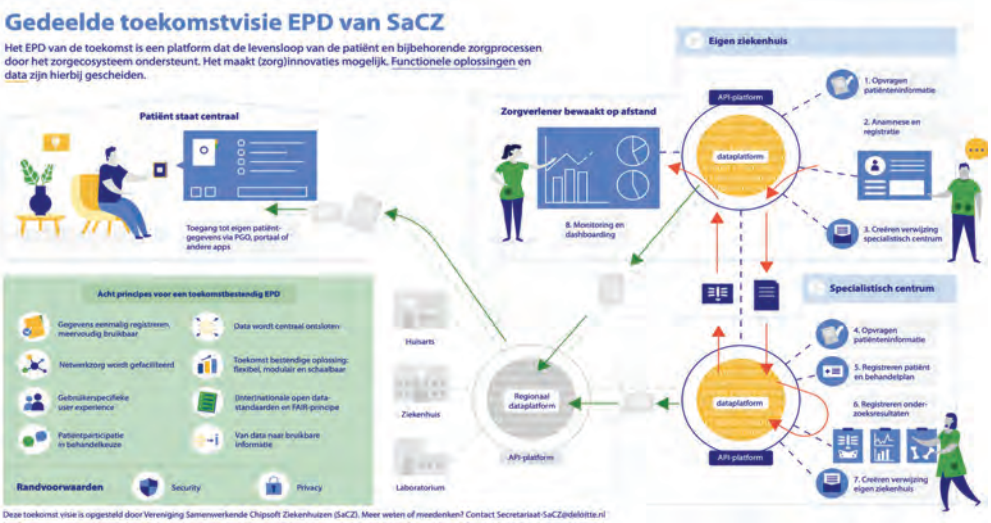
Primer: Karolinska hospital



Primer: UK NHS – arhitektura v uvajanju



Primer: Nizozemska – vizija



Zaščita podatkov

Zaščita podatkov

Poleg ustrezne varnostne infrastrukture, ki ni omejena samo na požarne zidove, protivirusne programe, enkripcije podatkov, nadzor dostopa in rešitve za preprečevanje vdorov ter rednega posodabljanja sistemov, je ključnega pomena za zaščito podatkov:

- **Stroga politika dostopa do podatkov**: dostop do zdravstvenih podatkov morajo imeti **samo** pooblaščene osebe. Sistem dodeljevanja pravic temelji na načelu najmanjših privilegijev (principle of least privilege).
- **Usposabljanje osebja**: Redno usposabljanje zdravstvenega osebja o varnosti podatkov in varnostnih postopkih je ključno, da se preprečijo napake in zlorabe.
- **Nadzor in revizija**: Redni nadzor, pregledi varnostnih postopkov in revizije sistema so potrebni za odkrivanje morebitnih varnostnih incidentov ali kršitev varnosti podatkov.
- **Zavarovanje podatkov v mobilnih napravah**: ustrezno zavarovanje podatkov, na primer z uporabo možnosti za oddaljeni izbris podatkov v primeru kraje ali izgube.
- **Sodelovanje z izkušenimi eksperti za informacijsko varnost**

Zaščita podatkov v zdravstvu je nenehen proces, ki zahteva stalno pozornost in nenehno izboljševanje, da se zagotovi varnost in zaupnost podatkov pacientov.

Napadi z izsiljevalsko programsko opremo

- 11. septembra 2020, Univerzitetna bolnišnica Düsseldorf v Nemčiji: prvi dokumentiran primer smrti zaradi ransomware napada. Bolnišnični IS onemogočeni do te mere, da ni več mogla sprejemati novih sprejemov na svoj oddelek za nesreče in urgentne službe.
- Oktober 2020, Vastaamo (finski razvijalec IS za področje psihoterapije): napadalec zahteval od izvajalca (tj.: Vastaamo) plačilo 450.000EUR v zameno, da ne objavi podatkov: kopija vseh podatkov o sistemu, vključno z imeni, naslovi, številkami socialnega zavarovanja, e-poštnimi naslovi, zapiski terapevta o vsaki zasebni seji. Vastaamo ni plačal odkupnine, posledično so plačali nekateri pacienti za svoje podatke (200-500EUR). Napadalec je na temnem spletu objavil 10 gigabajtov velik arhiv, ki je vseboval klinične zapiske približno 2000 bolnikov.
- 4.avgust 2022, Advanced (angleški razvijalec IS za NHS), Velika Britanija: onemogočen dostop do zdravstvenih nasvetov po telefonu, ki jo je vsak dan uporabljalo na tisoče ljudi in je bila ključna storitev v času COVID-19. Onemogočeno je bilo delovanje ordinacij splošnih zdravnikov, domov za ostarele in storitve duševnega zdravje, za katere je bil izvajalec Advanced. Tudi tu gre, kot v Düsseldorfu, za napad na razpoložljivost.

Vir: https://www.bcs.org/articles-opinion-and-research/biggest-healthcare-cyber-attacks-this-decade/

Zavarovanje

Skupno vsem trem napadom:

- Vse so povezane z zdravstveno oskrbo.
- Vse imajo jasen vpliv na zagotavljanje oskrbe.
- Vse spadajo v kategorijo “ransomware” (izsiljevalska programska oprema)

Večino napadov z izsiljevalsko programsko opremo izvajajo organizirane kriminalne združbe. Zakaj torej ciljati na zdravstveno varstvo:

- Zdravstvena oskrba je sestavni del naše blaginje in preživetja.
- Naši zdravstveni zapisi so zaupni.
- Naša življenja so odvisna od zdravstvenih delavcev, ki lahko zanesljivo dostopajo do sistemov za zagotavljanje te oskrbe.

Zelo malo je gospodarskih področij, ki tako neposredno vplivajo na našo blaginjo. Ko lahko napadalci ogrožajo tisto, kar nas ohranja žive, varne in zdrave – **zakaj ne bi plačali odkupnine**?

Postaja običajno, da se od izvajalcev informacijskih sistemov zahtevajo sklenjena visoka zavarovanja za primere napadov in za primere kakršnekoli škode v zvezi s podatki.

Podatki v ostalih institucijah

V predstavitvi je bilo izpostavljeno oblikovanje modularne arhitekture aplikacij, osredotočene na celovitost in varnost podatkov zlasti v kontekstu zdravstvenih sistemov.

Uporaba teh arhitekturnih paradigem z osredotočenostjo na podatke, njihovo zanesljivost, razširljivost in varnost je enako pomembna na ostalih kritičnih področjih, kot so nacionalne baze podatkov poslovnih subjektov, državljanov ali kateri koli drugi obsežni registri.

Hvala lepa.

http://linkedin.com/in/matejasajovic
Matea.Sajovic@gmail.com

Podatki in študenti pedagoških študijskih programov

Data and students in teacher training programmes

Jože Rugelj

Univerza v Ljubljani, Pedagoška fakulteta

POVZETEK

Podatki so za izvedbo pedagoškega procesa zelo pomembni, saj z njihovo uporabo učitelji načrtujemo, izvajamo in analiziramo poučevanje ter usmerjamo in podpiramo učenje. Kako učinkovito poučujemo, lahko ocenjujemo predvsem tako, da ugotavljamo, kako uspešno so se učenci učili in gradili novo znanje. V tem procesu učenci podatke v različnih pojavnih oblikah, kot so besedilo, slike, animacije, video in zvočni posnetki, in ki jim rečemo učna gradiva, zaznavajo s svojimi čutili in jih v miselnih procesih z uporabo predhodnega znanja oblikujejo v informacije in v novo znanje.

Učitelji pa zbiramo povratne informacije o učinkovitosti poučevanja oziroma učenja učencev z njihovim opazovanjem, s spremljanjem njihovih odzivov v pogovorih in drugih aktivnostih v učnem procesu ter z različnimi oblikami preverjanja znanja. Z uvajanjem digitalnih tehnologij v učni proces se te možnosti še povečajo, saj digitalna učna okolja in različne oblike interaktivnih učnih gradiv poleg vseh tradicionalnih oblik omogočajo še zbiranje velikih količnin drugih podatkov, ki so tudi povezani z učnim procesom. Pri tem je mogoče precej teh podatkov zbrati na načine, ki jih učenci sploh ne zaznajo in so zato manj moteči v procesih učenja.

Če je v tradicionalnih oblikah počevanja učitelj vse podatke, ki smo jim takrat rekli opažanja, zbral in obdelal v svoji glavi, iz njih pridobil pomembne informacije ter se na njihovi osnovi običajno tudi sproti odzival, pa je velike količine digitalnih podatkov potrebno ustrezno obdelati s pomočjo ustrezne analitike in potem iz dobljenih rezultatov dobiti potrebne informacije.

Zelo pomembno v obeh primerih je dejstvo, da je količina opažanj in podatkov, ki omogočajo koristne povratne informacije o procesu učenja, odvisna od aktivnih oblik učenja učencev. Take oblike pa so po ugotovitvah stroke tudi sicer ključne za učinkovito učenje. Tudi digitalne tehnologije v vlogi učne tehnologije pridobijo pravo vlogo šele takrat, ko so učenci pri učenju aktivni, saj šele takrat pridejo do izraza vse možnosti, ki jih ponujajo: podpora kognitivnim procesom, podpora komunikaciji in sodelovanju, učinkoviti kreativni procesi in organizacijo aktivnosti. V nasprotnem primeru služijo pretežno za distribucijo učnih gradiv in morda še za zbiranje nalog učencev.

Iz vsega povedanega lahko ugotovimo, da je za učinkovito uvajanje digitalnih tehnologij v učni proces predvsem pomembno, da učitelj uporabi ustrezne sodobne učne pristope, ki temeljijo na aktivnih oblikah učenja, pri katerih učenci rešujejo različne probleme, ter na sodelovanju med učenci in učiteljem in med učenci samimi.

Seveda pa mora učitelj poleg znanja s predmetnega področja in didaktičnega znanja imeti tudi znanja s področja uporabe digitalne tehnologije v izobraževanju, kar so strokovnjaki predstavili s pomočjo modela TPACK. Digitalne kompetence, ki jih mora imeti učitelj, so zbrane v Evropskem okviru digitalnih kompetenc izobraževalcev DIGCOMPEDU, ki so ga na osnovi ugotovitev širokega kroga strokvnjakov z različnih področij oblikovali pri Evropski komisiji. Definirali so 22 ključnih digitalnih kompetenc za učitelje in jih razvrstili v 6 skupin. Ta referenčni model je danes osnova za pripravo izobraževanja učiteljev, iz njega pa lahko razberemo tudi vsa področja v izobraževanju, kjer se učitelji srečujejo z zbiranjem, shranjevanjem in obdelavo podatkov ter z uporabo informacij, ki jih lahko iz teh podatkov izluščijo. Tudi mi bomo pogledali, kako so podatki povezani s temi kompetencami in kakšna znanja in veščine mora imeti učitelj, da jih lahko koristno in učinkovito uporabi v učnem procesu.

Pri tem vedno večji pomen pridobivajo tudi nove tehnologije, ki omogočajo obdelavo velikih količin podatkov in ki s pomočjo strojnega učenja in z uporabo kompleksnih algoritmov opravljajo naloge, ki običajno zahtevajo človeško inteligenco. V našem primeru gre lahko za reševanje problemov, učenje, ali sprejemanje odločitev. Ob tem pa morajo učitelji poznati omejitve in potencialne izzive pri razvoju in uporabi umetne inteligence.

SUMMARY

Data is essential to the delivery of the teaching process as teachers use it to plan, deliver and analyse teaching, and to guide and support learning. We can assess how effectively we are teaching primarily by how well students are learning and building new knowledge.

Learners use their senses to perceive information in various forms, such as text, images, animations, video and audio clips, which we refer to as learning material, and use their thought processes to process this information into information and new knowledge.

Teachers obtain feedback on the effectiveness of teaching or learning by observing students, monitoring their responses in conversations and other learning activities, and undertaking various forms of assessment. The introduction of digital technologies into the learning process expands these opportunities, as digital learning environments and various forms of interactive learning materials enable the collection of large amounts of other data related to the learning process in addition to all traditional forms. Much of this data can be collected in a way that is imperceptible to the learner and therefore less disruptive to the learning process.

Whereas in traditional forms of inquiry the teacher collected and processed all the data - what we used to call observations - in their head, extracted meaningful information from it, and usually acted on it, large amounts of digital data need to be properly processed using appropriate analytics and the results then extracted to provide the necessary information.

Very important in both cases is the fact that the number of observations and data that provide useful feedback on the learning process depends on active forms of learning by learners. Such forms are also seen by the profession as crucial for effective learning. Digital technologies also play their role as learning technologies only when learners are actively involved in their learning, because only then do they unfold their full potential: to support cognitive processes, to support communication and collaboration, to support effective creative processes and to organise activities. Otherwise, they are mainly used for distributing learning materials and perhaps also for the collection of learners' assignments.

From all this, we can conclude that for the effective introduction of digital technologies into the learning process, it is essential that the teacher uses appropriate modern teaching approaches based on active forms of learning in which students solve different problems and on collaboration between students and the teacher as well as between the students themselves.

In addition to subject-specific and didactic knowledge, a teacher must of course also have knowledge of the use of digital technologies in the classroom, which the experts presented using the TPACK model. The digital competences that teachers must have are summarised in the DIGCOMPEDU European Framework of Digital Competences for Educators, which was developed by the European Commission based on the findings of a large number of experts from various fields. They have defined 22 key digital competences for teachers and divided them into 6 clusters. This reference model is now the basis for teacher training and feeds into all areas of education where teachers are confronted with the collection, storage and processing of data and the use of the information gained from it. We will also look at how data relates to these competencies and what knowledge and skills a teacher needs to have to use it meaningfully and effectively in the learning process.

New technologies that enable the processing of large amounts of data and use machine learning and complex algorithms to perform tasks that normally require human intelligence are also becoming increasingly important. In our case, this could be problem solving, learning or decision making. At the same time, teachers need to be aware of the limitations and potential challenges of developing and using AI.

O AVTORJU

Dr. Jože Rugelj je redni profesor računalništva v izobraževanju na Univerzi v Ljubljani in predstojnik Katedre za didaktiko računalništva na Pedagoški fakulteti. Raziskovalno se ukvarja z uporabo IKT v izobraževanju, kognitivnimi vidiki multimedijske podpore učenju, uporabo didaktičnih iger v izobraževanju in inovativnimi pristopi k poučevanju računalništva. Rezultate raziskovalnega dela je objavil v 29 člankih v mednarodnih revijah, 68 prispevkih v zbornikih mednarodnih konferenc in 17 poglavjih v znanstvenih monografijah.

ABOUT THE AUTHOR

dr. Jože Rugelj is a full professor of computer science in education at the University of Ljubljana and head of the Department of Computer Science Didactics at the Faculty of Education. His research deals with the use of ICT in education, cognitive aspects of multimedia learning support, the use of didactic games in education and innovative approaches to teaching computer science. He published the results of his research work in 29 articles in international journals, 68 contributions in the proceedings of international conferences and 17 chapters in scientific monographs.

Podatki in študenti pedagoških študijskih programov

Jože Rugelj
Pedagoška fakulteta
Univerza v Ljubljani

Pomen podatkov za izvedbo učnega procesa

- Z uporabo podatkov učitelj
  - načrtuje,
  - izvaja in
  - analizira **poučevanje**;
  - usmerja in
  - podpira **učenje**.
- **Kakovost poučevanja** merimo z doseganjem **učnih ciljev**.
- **Učna gradiva** so **podatki** v različnih pojavnih oblikah:
  - besedilo,
  - slike,
  - animacije,
  - video posnetki,
  - zvočni posnetki.

- Učitelji zbirajo **podatke** o **učinkovitosti poučevanja** oziroma učenja učencev
  - z njihovim **opazovanjem**,
  - s **spremljanjem** njihovih **odzivov** v učnem procesu in
  - z različnimi **oblikami preverjanja znanja**.
- Z uvajanjem digitalnih tehnologij se **možnosti povečajo**, saj digitalna učna okolja in različne oblike **interaktivnih učnih gradiv** omogočajo zbiranje velikih količnin podatkov.
- Precej podatkov je lahko zbranih tako, da **učenci** tega sploh **ne zaznajo**, zato je zbiranje **manj moteče**.

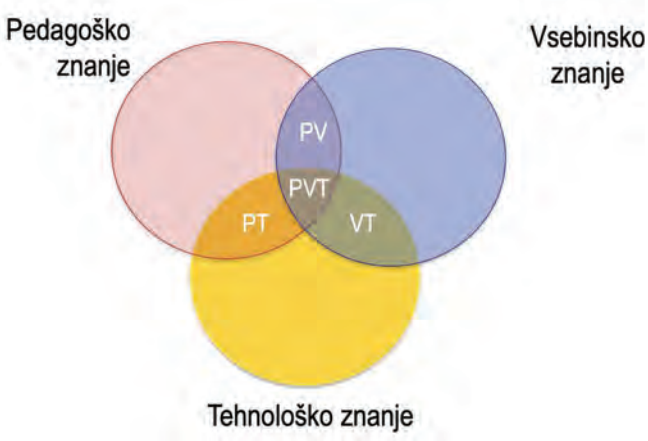
Tradicionalne in nove oblike zbiranja podatkov

- Pri tradicionalnih oblikah je učitelj vse podatke (opažanja) **zbral** in „**obdelal**“ v **svoji glavi**.
- Iz njih je **sproti** pridobil pomembne **informacije** ter se na njihovi osnovi običajno tudi **odzival**.
- Velike količine **digitalnih podatkov** je potrebno ustrezno **obdelati** s pomočjo ustrezne **analitike** in iz dobljenih rezultatov dobiti **iskane informacije**.
- **Količina** opažanj in podatkov, ki omogočajo pridobivanje koristnih informacij o učenju, je **odvisna** od **stopnje aktivnosti** oblik **učenja**.
- **Aktivne oblike** pa so po ugotovitvah stroke **tudi sicer ključne** za učinkovito učenje.

- **Digitalne tehnologije** v vlogi učne tehnologije pridobijo **pravo vlogo** šele takrat, ko so učenci pri učenju **aktivni**.
- To dosežemo z novimi **konstruktivističnimi didaktičnimi pristopi**, kjer učenci **rešujejo probleme** ter pri tem **sodelujejo**.
- Takrat pridejo **do izraza** vse **možnosti**, ki jih ponujajo:
  - podporo **kognitivnim procesom**,
  - podporo **komunikaciji** in **sodelovanju**,
  - učinkovite **kreativne procese** ter
  - **organizacijo** učnih **aktivnosti**.
- Sicer pa služijo pretežno samo za
  - **distribucijo** učnih **gradiv**,
  - **zbiranje** oddanih **nalog** učencev.

TPCK model *(Mishra & Koehler, 2006)*

- Učitelj mora poleg znanja s **predmetnega področja** in **didaktičnega** znanja imeti tudi znanja s področja **uporabe digitalne tehnologije v izobraževanju**, kar so strokovnjaki predstavili s pomočjo **modela** TPCK.



Digitalne kompetence učiteljev

- Digitalne kompetence **učiteljev** so zbrane v Evropskem okviru digitalnih kompetenc izobraževalcev DIGCOMPEDU.
- Na osnovi ugotovitev **širokega kroga strokovnjakov** z različnih področij so ga **oblikovali** pri Evropski komisiji.
- Ta **referenčni model** je danes **osnova** za **pripravo izobraževanja** učiteljev ne tem področju.
- Iz njega razberemo **vsa področja** v izobraževanju, kjer se učitelji srečujejo z **zbiranjem**, **shranjevanjem** in **obdelavo podatkov** ter z uporabo **informacij**, ki jih lahko **iz teh podatkov** izluščijo.

Digitalne kompetence učiteljev

Nekaj manj poznanih načinov uporabe podatkov

1. **Vrednotenje in spremljanje napredka**
   Podatki iz formativnega in sumativnega ocenjevanja za ocenjevanje učenja in napredka učencev, za ugotavljanje močnih in šibkih področij in prilagajanje pouka potrebam učencev.
2. **Personalizirano učenje**
   Podatki omogočajo ustvarjanje prilagojenih učnih izkušenj za učence z zbiranjem podatkov o njihovih interesih, učnih stilih in uspešnosti z zanimivim in učinkovitim učnim okoljem.
3. **Diferencirano poučevanje**
   Podatki podlaga za diferencirano poučevanje, pri katerem z analizo podatkov o uspešnosti učencev prilagodijo svoje učne strategije in zagotovijo ciljno usmerjeno podporo in obogatitev za čim boljše učne rezultate učencev.

4. Intervencija in podpora

Podatki omogočajo prepoznavanje učencev z učnimi težavami ali njihovo potrebo za dodatno podporo. S spremljanjem podatkov lahko učitelji izvajajo ciljno usmerjene intervencije, zagotavljajo korekcijo in ponujajo dodatne učne vire.

5. Sprejemanje odločitev na podlagi ‘dokazov‘

Na podatkih temelječe odločanje omogoča učiteljem, da sprejemajo utemeljene odločitve, ki temeljijo na dokazih in ne na intuiciji.

Priložnosti in nevarnosti uporabe UI

- Vedno večji pomen novih tehnologij, ki omogočajo obdelavo velikih količin podatkov.
- Uporaba teh tehnologij, ki s pomočjo strojnega učenja in s kompleksnimi algoritmi opravljajo naloge, ki običajno zahtevajo človeško inteligenco.
- V našem primeru gre lahko za reševanje problemov, učenje, ali sprejemanje odločitev.
- Ob tem pa morajo učitelji spoznati omejitve in potencialne izzive pri razvoju in uporabi umetne inteligence.

Zaključki

- Podatki so za učitelja pomembni pri pripravi, izvedbi in analizi poučevanja ter pri spremljanju učenja učencev.
- Digitalne tehnologije ponujajo mnoge nove možnosti za bolj kakovostno izvedbo izobraževanja.
- Če se pri usmerjanju učnih strategij, oblikovanju učnih načrtov in dodeljevanju sredstev opirajo na podatke, lahko učitelji in odločevalci zagotovijo, da so njihove odločitve utemeljene na objektivnih informacijah in bodo privedle do pozitivnih rezultatov.

Pomen podatkov in umetne inteligence/strojnega učenja za zasebna mobilna omrežja v tovarnah prihodnosti

Harnessing data and ai/ml for private mobile networks in factories of the future

Urban Zaletel

Kontron

POVZETEK

Zadnji razmah umetne intelligence je odprl vrata inovacijam na mnogih področjih. V industriji vlada konsenz, da je to verjetno ena pomembnejših tehnologij do sedaj. Umetna inteligenca pomaga pri izboljšanju različnih uporabniških izkušenj, kot so izkušnje na športnih tekmah, koncertih, programiranju, službi in učenju. Na primer, z uporabo naprednih algoritmov za obdelavo naravnega jezika (NLP) lahko računalnik razume kontekst in namensko prilagodi izkušnjo gledalca ali udeleženca dogodka oz. pomaga uporabniku pri opravilih. To bi lahko vključevalo personalizirane analize tekme za športne navdušence, interaktivne učne vsebine za učence ali prilagojeno avdiovizualno doživetje za obiskovalce koncertov, ter še mnogo več.

Z razvojem determinističnih omrežij (DetNet) na ravni sloja 3 se odpirajo nove možnosti za deterministično prenašanje podatkov, kar je ključnega pomena za aplikacije, ki zahtevajo nizko latenco in zelo majhno izgubo paketkov. Umetna inteligenca, še posebej v obliki naprednih algoritmov za optimizacijo omrežij, lahko igra ključno vlogo pri optimizaciji in upravljanju takih omrežij, kar omogoča boljše izkoriščanje njihove zmogljivosti in zagotavljanje kakovostne uporabniške izkušnje.

V prispevku je predstavljeno omrežje DetNet – deterministično omrežje na nivoju 3 OSI refenčnega modela in prikazano kdaj potrebujemo sploh tako omrežje in v katerih primerih je tako omrežje nesmiselno. Predstavljena je arhitektura omrežja DetNet. V zaključku je povzeto, kako se v trenutnih produktih v Kontronu že uporablja UI in kakšne so praktične koristi in potencial umetne intelligence pri izboljšanju različnih vidikov sodobnih tehnoloških rešitev.

SUMMARY

The recent surge in artificial intelligence has opened doors to innovation across various fields. There's industry consensus that this is arguably one of the most significant technologies to date. Artificial intelligence aids in enhancing various user experiences, such as those in sports events, concerts, and learning. For instance, by utilizing advanced natural language processing (NLP) algorithms, computers can grasp context and tailor the viewer's or participant's experience accordingly. This could involve personalized match analyses for sports enthusiasts, interactive learning content for students, or customized audiovisual experiences for concert attendees.

With the development of deterministic networks (DetNet) at the Layer 3 level, new possibilities for deterministic data transmission emerge, crucial for applications requiring low latency and minimal packet loss. Artificial intelligence, particularly in the form of advanced network optimization algorithms, can play a vital role in optimizing and managing such networks, enabling better utilization of their capabilities and ensuring a high-quality user experience.

The presentation introduces DetNet – a deterministic network at the Layer 3 OSI reference model – and demonstrates when such a network is needed and in which cases it might be impractical. It is presented only the architecture of the DetNet network and which 3GPP release is the be part of. Finally, reader learns how we already utilize AI in current products at Kontron, showcasing the practical benefits and potential of artificial intelligence in improving various aspects of modern technological solutions.

O AVTORJU

Urban Zaletel je izkušen vodja razvoja in raziskav v Kontron-u za mobilna zasebna omrežja. Povezuje tehnologijo s poslovnimi odločitvami in pridobljeno znanje preliva v končne izdelke za stranke. Rad ustvarja produkte od konceptne faze do končne produktivizacije. Njegovo delo vključuje raziskave, načrtovanje arhitekture programske opreme, izbiro platforme in drugo. Urban je organizator dveh največjih konferenc posvečenih razvijalcem v Sloveniji, ki potekata vsako leto v Portorožu, MakeIT in JCON.

ABOUT THE AUTHOR

Urban Zaletel is an experienced development and research manager at Kontron for mobile private networks. It connects technology with business decisions and transforms the acquired knowledge into final products for customers. He likes to create products from the concept phase to the final production. His work includes research, software architecture planning, platform selection, ... Urban is also the organizer of the two largest conferences dedicated to developers in Slovenia, which take place every year in Portorož, MakeIT and JCON.

kontron

Harnessing Data and AI/ML for Private Mobile Networks in Factories of the Future

Urban Zaletel

kontron

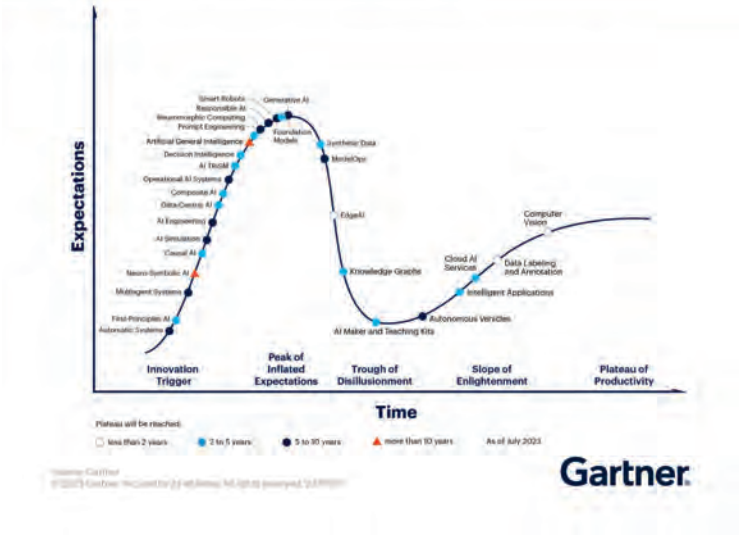
01 AI (GenAI)

02 Network and Use cases

03 DetNet

04 K8sGPT

Hype Cycle for Artificial Intelligence, 2023



kontron

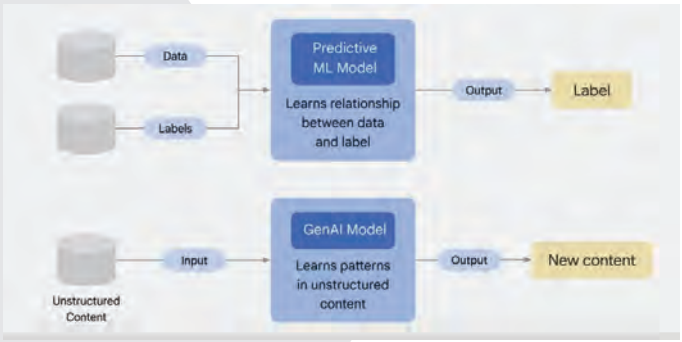
› "Is this the most important new computer technology ever? Probably."

Larry Ellison Chairman and CTO, Oracle (Cloudworld 2023)

How is Generative AI different than other AI approaches?

kontron

- Seminal 2017 paper on transformer model
  - › Blog: Transformer: A Novel Neural Network Architecture for Language Understanding
  - › Paper: Attention is all you need
- Processes whole sequences of tokens (words) not just individual words
- Faster to understand context of language
  - Improves efficiency, scale to LLM
- Started era of Generative AI



One and a half year ago: ChatGPT 3.5 was born

kontron

The baby talks!

› Multi-billion parameter large language model (LLM)
› Trained on trillions of data elements… (Wikipedia + public internet)
ChatGPT's Scale Enabled it to Understand and Generate Language!

› To train the model you need huge superclusters.
› RDMA network: Connect NVDIA GPUs into Huge Superclusters
  › Time is money: Speed and less half the cost

Kontron SI Group Presentation

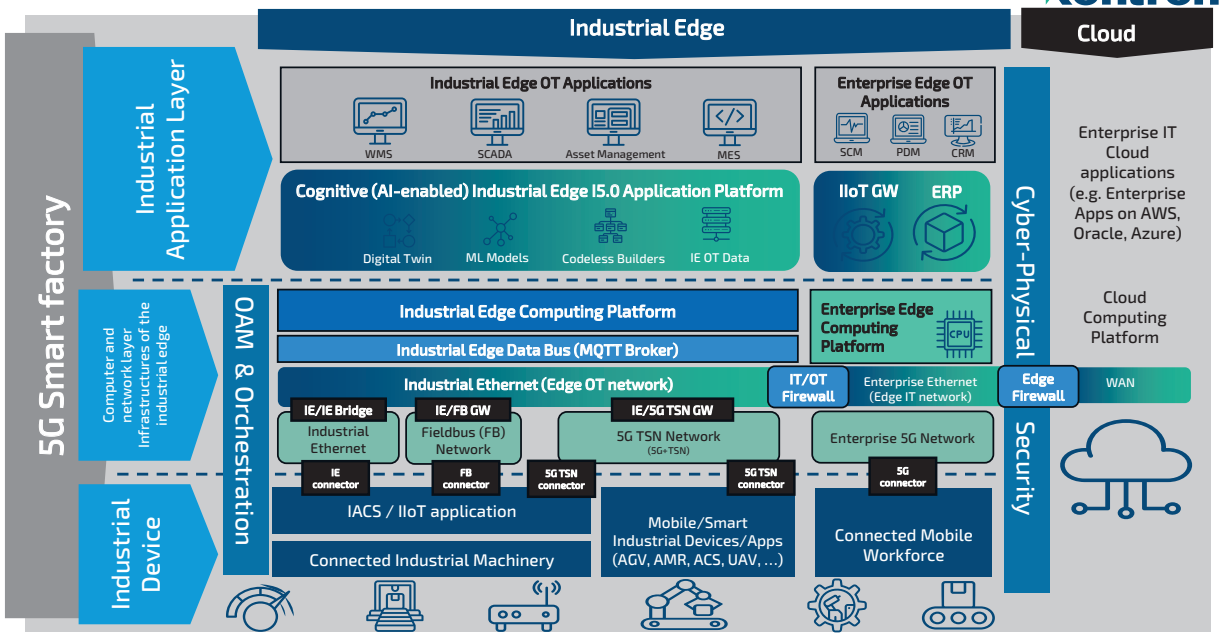
Immersive Media Sector by Enabling New Narrative Visions

kontron

› Cases like:
  › Enhance gaming experience (sports, gaming, …)
  › Enhance music concert experience
  › Cultural heritage
  › Education
  › Retail
› combined with AI will require new faster, extremely low packet loss and low end-to-end latency network.

Kontron SI Group Presentation

Overview 5G Smart Factory and AI interaction



DetNet

Deterministic IP communication over 5G network

- Defined by IETF, Deterministic Networking (DetNet) “provides a capability for the delivery of data flows with extremely low packet loss and bounded end-to-end delivery latency
- Operates at:
  - IP
  - Multiprotocol label switching (MPLS)
- 3GPP Rel. 18. (Study on Extensions to TSC framework to support DetNet)
  - 5GS DetNet node will be acting as DetNet transit node

DetNet Use Cases

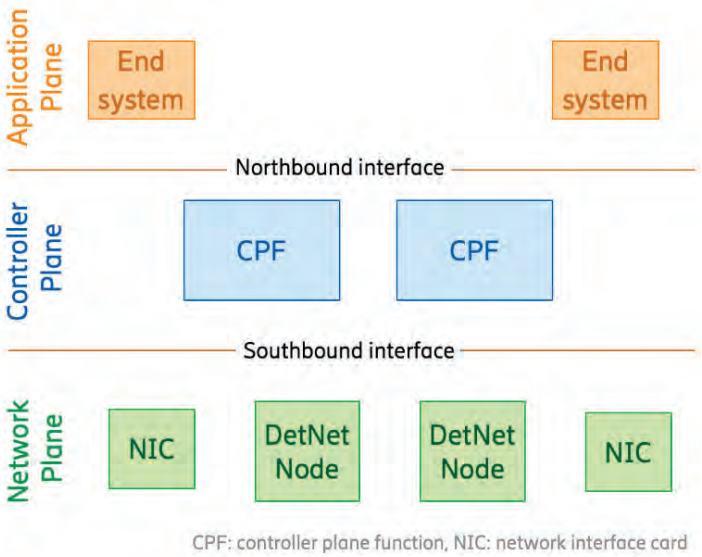
kontron

Use case	Span area	IP-capable devices	Interconnection of subnetworks	Edge cloud	DetNet applicability
Motion control (cycle 500 us)	**50m x 10m**	**No (often industrial Ethernet)**	**No**	**No (close proximity of controller and sensor/actuators)**	**No (strict short cycles may require TSN Qbv)**
Slower Motion control (cycle 0.5 – 2ms)	50m x 10m	No (often industrial Ethernet)	No	No (close proximity of controller and sensor/actuators)	Possible (unless TSN Qbv is required)
Control-to-control communication	100m x 30m	Possible	IP required to interconnect L2-based subnets	Possible	Yes
Mobile robots	1km x 1km	Possible	Unspecified	Yes (edge cloud processing)	Yes
Mobile robots: Cooperative carrying	10m x 10m or 50m x 5m	Possible	Unspecified	Possible	Possible (need to determine if sidelink can be supported with 5G DetNet)
Example: Flexible modular production	1km x 1km	Yes	Yes, wireless 5G DetNet interconnects wired subnets	Possible	Yes
Augmented reality and AI	50m x 10m	Yes	Unspecified	Yes (compute offload)	Yes
Closed-loop control	100m x 100m	Possible	Unspecified	Unspecified	Yes, especially for large-scale closed-control loop
Process and asset monitoring	10km x 10km	Yes	Unspecified	Yes	Yes
Plant asset management	10km x 10km	Yes	Unspecified	Yes	No need, requirements are relaxed. 5G alone is enough.

Kontron SI Group Presentation

DetNet Architecture

kontron

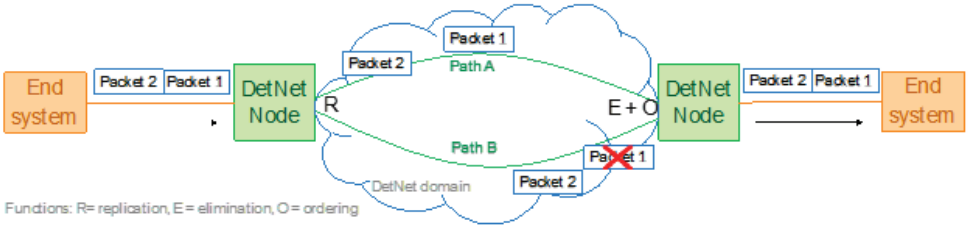


Kontron SI Group Presentation

DetNet Network Plane

kontron

› Functionalities are decomposed into two sub-layers:
  › DetNet Service sub-layer
    › PREOF (packet replication, elimination, and ordering functions)

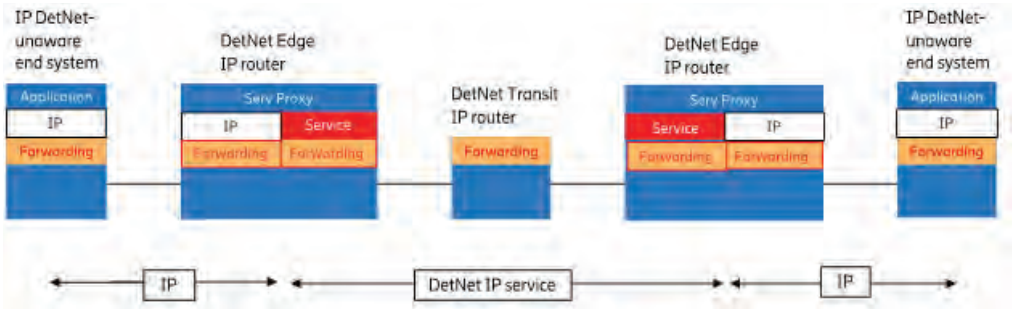


  › DetNet Forwarding sub-layer
    › employs explicit routes and resource allocation.

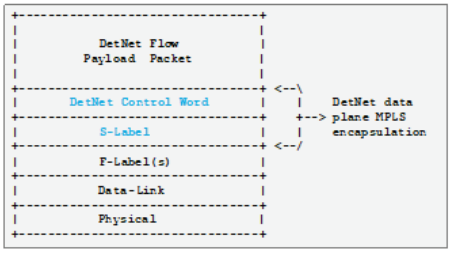
Encapsulation

kontron

› IP



› MPLS



TSN vs. DetNet features

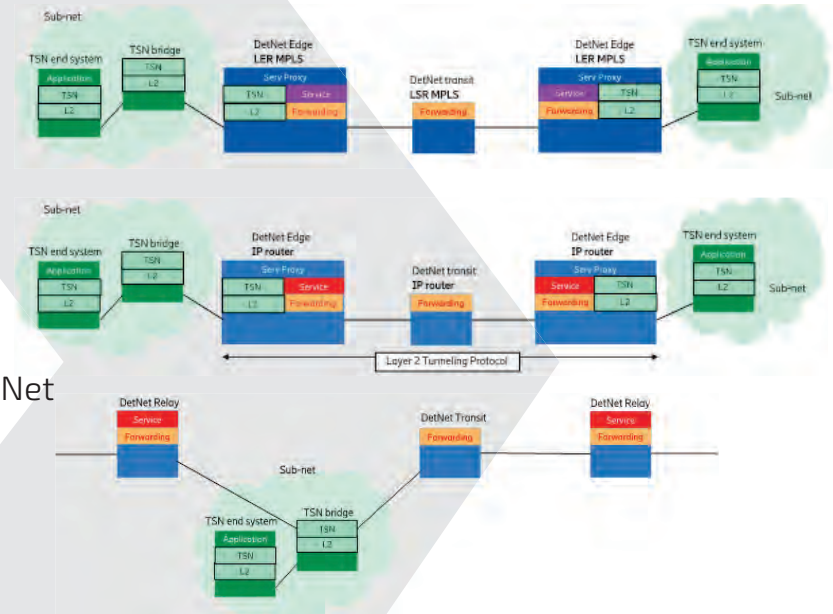
kontron

IEEE 802.1 TSN	IETF DetNet
Bounded latency	Bounded latency
Extremely low frame loss	Extremely low packet loss
Frame Replication and Elimination for Reliability (FRER)	Packet Replication, Elimination, and Ordering Functions (PREOF)
Ingress policy/Time-based (PSFP/IEEE 802.1Qci)	Ingress policing provided by IETF tools. PSFP needs to be provided by subnetwork (see [11], Note 1 and 2)
Ingress policy/Rate-based (PSFP/IEEE 802.1Qci)	Rate-limiting/policing functions supported (e.g., various formats of access control list (ACL) implemented by all major router vendors)
Scheduled traffic (IEEE 802.1Qbv)	Needs to be provided by subnetwork (see [11] and Note 1 and 2)
Time Synchronization (IEEE 802.1AS)	Data plane specific time synchronization applies.
Strict Priority	Multiple strategies supported including strict priority (Note 3)
Configuration (IEEE 802.1Qcc)	Network configuration with existing protocols and using YANG models
Frame preemption (IEEE 802.1Qbu)	Not defined for Layer 3. If needed, the Layer 2 capabilities can be used.

TSN & DetNet

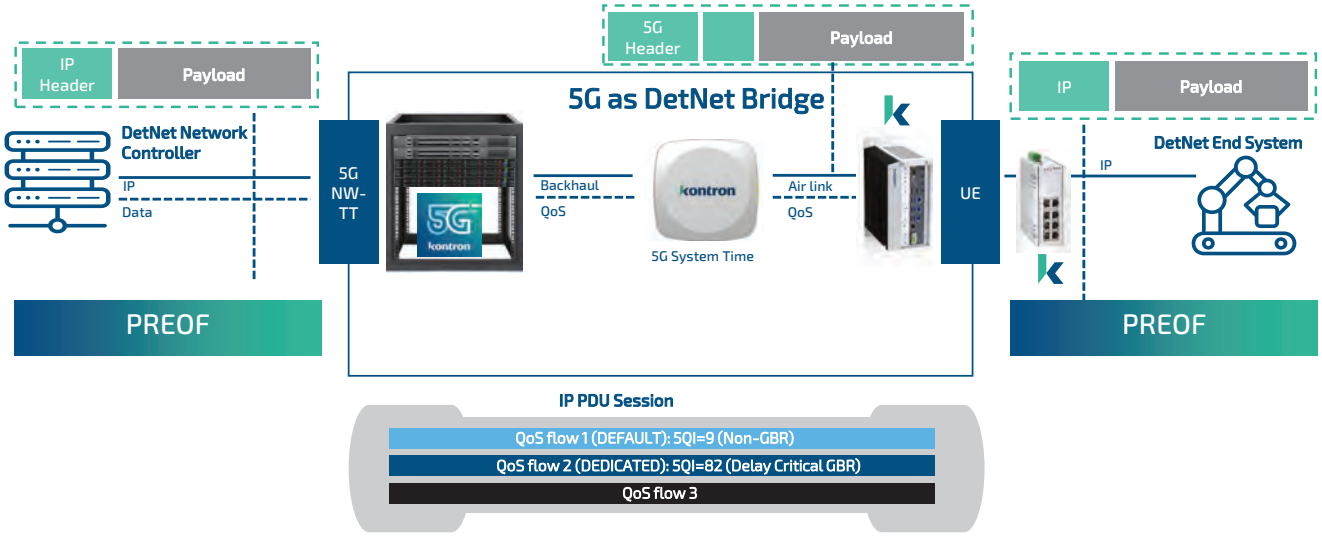
kontron

- Possible scenarios:
  - TSN over DetNet (MPLS)
  - TSN over DetNet (IP)
  - TSN is subnetwork of DetNet



DetNet over 5G

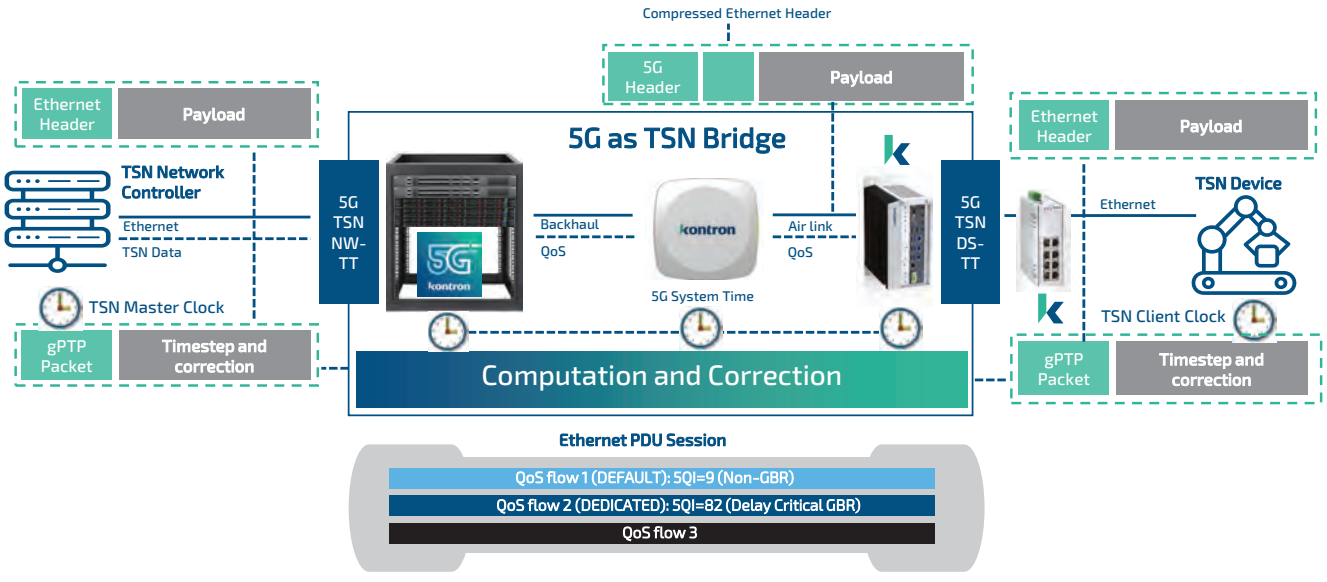
kontron



Kontron SI Group Presentation

TSN over 5G

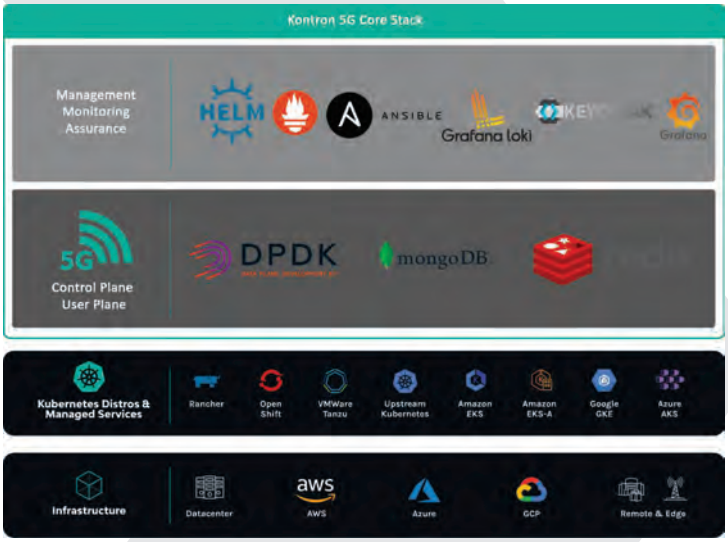
kontron



Kontron SI Group Presentation

Leveraging GenAI in 5G stack

kontron



What is k8sGPT?

Work Smart, Not Hard

kontron

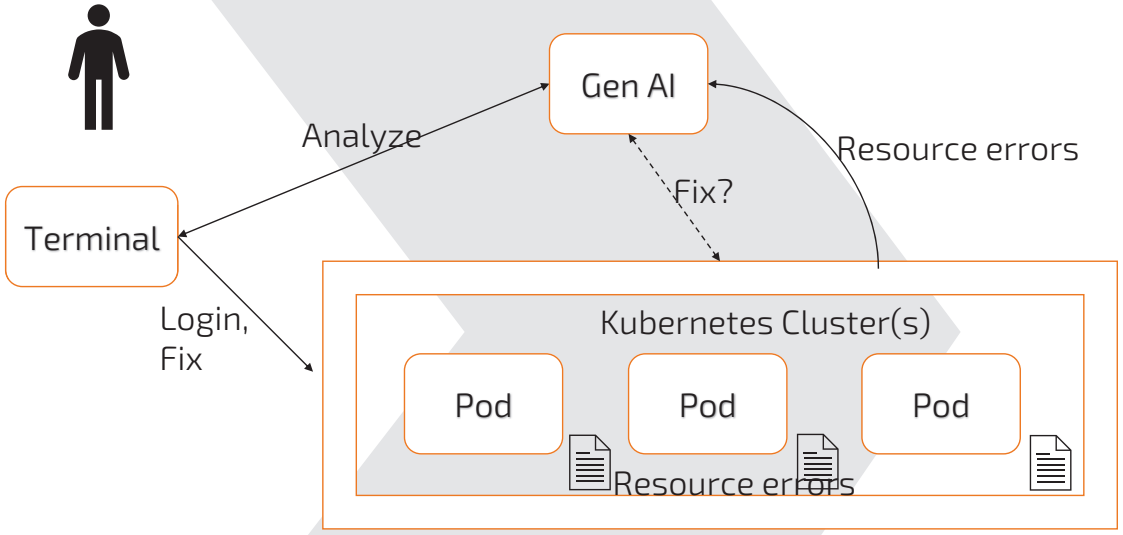
K8sGPT uses analyzers to triage and diagnose issues in your cluster. It has a set of analyzers that are built in, but you will be able to write your own analyzers.

Who is it for?

kontron

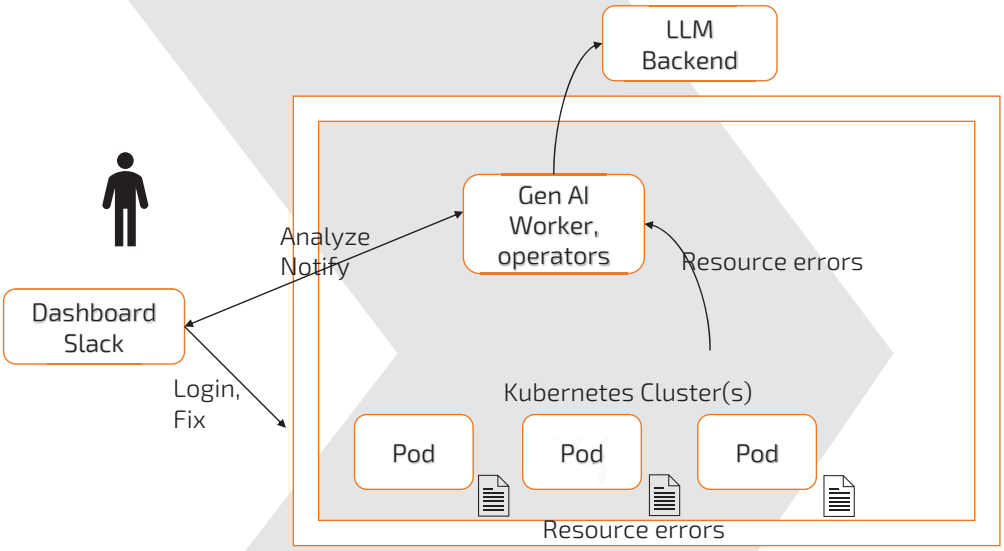
- Use Gen AI inside k8s to detect, diagnose errors, vulnerabilities
  - Codified SRE Knowledge Knows What To Search For
  - Translating k8s log errors to plain English* and generating possible solutions
- Real time monitoring of k8s clusters using the k8sgpt-operator
- Security CVE review using Trivy integration
- Caching results to S3 object storage or compatible API in Object Storage

kontron



Kontron SI Group Presentation

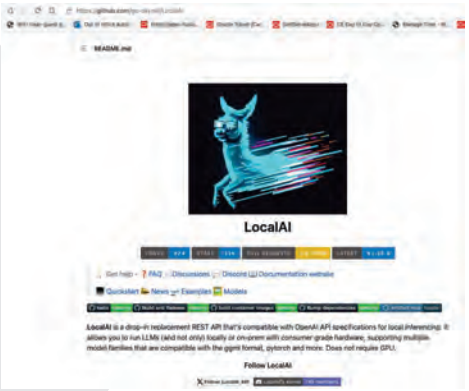
kontron



What is LocalAI?

kontron

- Local, OpenAI drop-in alternative REST API
- Run LLM within your k8s cluster
  - › llama.cpp and ggml to run inference
  - › E.g. Vicuna, Alpaca, LLaMA, Cerebras, GPT4ALL,
  - › GPT4ALL-J, koala, etc. from huggingface.com
- Perfect for edge computing that needs local AI
- Can use compute/GPU shapes for LLM
  - › Training needs compute, bandwidth, data
- Train using all the data on the internet



```yaml
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx-deployment
spec:
  selector:
    matchLabels:
      app: nginx
  replicas: 2 # tells deployment to run 2 pods matching the template
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx:1.14.2
        ports:
        - containerPort: 80
```

```
ubuntu@zaletel-workstation:~/k8s$ kubectl apply -f application.yaml
deployment.apps/nginx-deployment created
ubuntu@zaletel-workstation:~/k8s$ k8sgpt analyze --explain --filter=Pod --namespace=default
AI Provider: openai

No problems detected

ubuntu@zaletel-workstation:~/k8s$
ubuntu@zaletel-workstation:~/k8s$
ubuntu@zaletel-workstation:~/k8s$
ubuntu@zaletel-workstation:~/k8s$ kubectl get pods
NAME                                READY   STATUS    RESTARTS   AGE
nginx-deployment-86dcfdf4c6-2lf4v   1/1     Running   0          49s
nginx-deployment-86dcfdf4c6-rqw5k   1/1     Running   0          49s
```

```yaml
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx-deployment
spec:
  selector:
    matchLabels:
      app: nginx
  replicas: 2 # tells deployment to run 2 pods matching the template
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx:5.14.2
        ports:
        - containerPort: 80
```

```
ubuntu@zaletel-workstation:~/k8s$ kubectl get pods
NAME                                READY   STATUS             RESTARTS   AGE
nginx-deployment-695b6d9bbd-kf2sq   0/1     ImagePullBackOff   0          3m54s
nginx-deployment-86dcfdf4c6-2lf4v   1/1     Running            0          9m41s
nginx-deployment-86dcfdf4c6-rqw5k   1/1     Running            0          9m41s
```

GenAI explain

Instructions what to do and check

kontron

```
ubuntu@zaletel-workstation:~/k8s$ k8sgpt analyze --explain
 100% |
AI Provider: openai

0: Deployment default/nginx-deployment()
- Error: Deployment default/nginx-deployment has 2 replicas but 3 are available
Error: Deployment default/nginx-deployment has 2 replicas but 3 are available.
Solution:
1. Check the current status of the deployment using the command: kubectl get deployments
2. Scale down the deployment to match the desired number of replicas: kubectl scale --replicas=2 deployment/nginx-deployment
1: Pod default/nginx-deployment-695b6d9bbd-kf2sq(Deployment/nginx-deployment)
- Error: Back-off pulling image "nginx:5.14.2"
Error: Back-off pulling image "nginx:5.14.2"
Solution:
1. Check the image name and version for typos.
2. Ensure the image exists in the specified repository.
3. Verify network connectivity to the repository.
4. Restart the Kubernetes pod to retry pulling the image.
```

Kontron SI Group Presentation

Change language

kontron

```
ubuntu@zaletel-workstation:~/k8s$ k8sgpt analyze --language "slovenian" --explain
 100% |
AI Provider: openai

0: Deployment default/nginx-deployment()
- Error: Deployment default/nginx-deployment has 2 replicas but 3 are available
Error: Deployment ima 2 replike, vendar so na voljo 3.
Solution:
1. Preverite trenutno stanje replik s kubectl get pods.
2. Odstranite dodatno repliko s kubectl delete pod <ime_replike>.
3. Preverite, ali se je število replik uskladilo s kubectl get pods.
1: Pod default/nginx-deployment-695b6d9bbd-kf2sq(Deployment/nginx-deployment)
- Error: Back-off pulling image "nginx:5.14.2"
Error: Napaka pri prenašanju slike "nginx:5.14.2".
Rešitev:
1. Preverite povezavo do registra slike.
2. Preverite pravilnost imena in različice slike.
3. Ponovno zaženite postopek prenašanja slike.
```

Kontron SI Group Presentation

Trajnosten mobilen sistem: podatki za iskanje ravnotežja med zmogljivostjo omrežja in njegovo porabo

Sustainable Mobile System: Data for Balancing Network Performance and Energy Consumption

Janez Sterle, Luka Koršič, Rudolf Sušnik

Internet Institute

POVZETEK

V prispevku bomo podali, kako so metrike zmogljivosti omrežja in porabe energije, skupaj s ključnimi kazalniki uspešnosti, ključnega pomena pri oblikovanju trajnostnih mobilnih sistemov 5G in 6G. Poglobili se bomo v strategije za optimizacijo prenosa podatkov in zanesljivosti ob zmanjšanju porabe energije ter zagotavljanju trajnostnega pristopa k načrtovanju in delovanju omrežja. Razprava vključuje inovativne tehnike za merjenje in izboljšanje učinkovitosti omrežja, s poudarkom na ohranjanju zagotavljanja visoke kakovosti storitev.

SUMMARY

The paper presents how network performance and energy consumption metrics, alongside key performance indicators (KPIs), are critical in designing sustainable 5G and 6G mobile systems. Article delves into strategies for optimizing data throughput and reliability while minimizing energy usage, ensuring a sustainable approach to network design and operation. The discussion includes innovative techniques for measuring and improving network efficiency, with a focus on maintaining high-quality service delivery.

O AVTORJU

Janez Sterle je soustanovitelj in direktor podjetja INTERNET INSTITUT d.o.o. Magistriral in doktoriral je s področja telekomunikacij na Fakulteti za elektrotehniko, Univerze v Ljubljani. Njegovo glavno področje dela je načrtovanje, razvoj in upravljanje omrežij ter storitev, testiranje in in verifikacija tehnologij 5G/6G, NFV, IPv6, QoS in QoE; PPDR in NATO podprtih taktičnih komunikacijskih sistemov; preskušanje, merjenje in preverjanje najsodobnejših protokolov in tehnologij. Ima uveljavljene mednarodne izkušnje na področju raziskav in razvoja ter industrijskih projektov v različnih sektorjih (telekomunikacije, logistika, varnost in zaščita) vključno s projekti H2020 Evropske komisije na področju 5G in 6G tehnologij 6Green, Exigence, 5G-LOGINNOV, Int5Gent, EVOLVED-5G, 5GASP, 5G-INDUCE, 5G-IANA, MATILDA-5G in 5GINFIRE. Tesno sodeluje z industrijskimi akterji, regulatornimi in zakonodajnimi organi tako na strateški kot tehnični ravni. Ima industrijske certifikate in različne patente na področju mobilnih sistemov.

ABOUT THE AUTHOR

Janez Sterle is a co-founder and CEO of INTERNET INSTITUTE Ltd. He received his M.Sc. and Ph.D. degrees in telecommunications from the University of Ljubljana, Slovenia. His main area of work concerns network design, planning, service management, testing and implementation in production networks for 5G/6G, NFV, IPv6, QoS and QoE, PPDR and NATO enabled tactical communication system, testing, measurement and verification of state-of-the-art protocols and technologies. He has an established track record of R&D and production-grade projects in communications, safety, and security sectors, including EC's H2020 projects 6Green, Exigence, 5G-LOGINNOV, Int5Gent, EVOLVED-5G, 5GASP, 5G-INDUCE, 5G-IANA, MATILDA-5G and 5GINFIRE on the topic of 5G and 6G, and cooperates closely with the respective industries, practitioners, regulatory and legislative bodies on strategic and technical levels. He holds industrial certification and various US patents in the field of mobile systems.

Sustainable Mobile System: Data for Balancing Network Performance and Energy Consumption

dr. Janez Sterle, mag. Luka Koršič in dr. Rudolf Sušnik | INTERNET INSTITUT d.o.o.
janez.sterle@iinstitute.eu

VITEL 2024

Bled, Slovenija, 2024

Company Profile

- Company facts
  - Startup established in 2014
  - Located in Ljubljana, Slovenia
  - 100% IPR ownership
  - First employees Q4 2017 (7, +10 associates)
  - Trusted partner in EU R&I
- Core Expertise: development, engineering and operation of telco grade Quality Assurance (QA) and Critical Communications Systems (CCS)
- Main technologies verticals
  - QA | Quality assurance of mobile, fixed and cloud systems | *www.qmon.eu*
  - CCS | Solutions for 5G/IoT-based critical communications | *5gsafety.net*

© 2024 INTERNET INSTITUTE. All Rights Reserved.

We Live 5G/6G

This projects received funding from the European Union's Horizon 2020 research and innovation programme grant agreements No. 761898, 732497, 957400, 957403, 101016448, 101016608, 101016941 and 101016427.

- Operational 5G Network | SA
- 5G qMON | 5G Test Automation
- 5G IoT System | NSA/SA

- Orchestrating 5G Network
- Orchestrating 5G Test Automation
- Orchestrating 5G IoT Backend

Page 3 | © 2024 INTERNET INSTITUTE. All Rights Reserved.

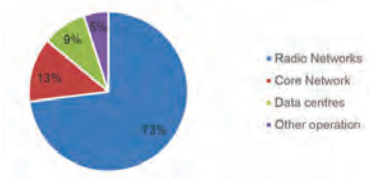
Sustainability & Mobile Systems

- Sustainability Challenge
  - Sustainability is crucial for all sectors, including ICT
  - ICT sector's commitment to aggressive sustainability targets (e.g. 6G KPI/KVI)
- Mobile Network Operators Goals
  - Carbon Footprint Reduction
  - Net Zero Emissions
- Determinants of Carbon Footprint
  - Energy Consumption: Primary factor from network operation
  - Emission Intensity: Impact of energy sources on carbon footprint

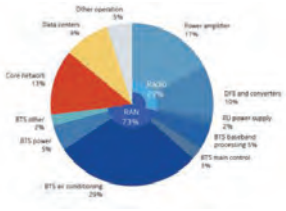
Page 4 | © 2024 INTERNET INSTITUTE. All Rights Reserved.

Mobile System Energy Consumption Breakdown

- Energy Consumption
  - Directly linked to mobile system load
  - Load fluctuates significantly over a 24-hour & weekly cycle
- Peak Load*
  - Occurs only for a few hours daily
- Mobile System Dimensioning*
  - Designed for busy hour capacity
- Resource Utilization at MNO*
  - An estimated 70% of radio resources idle on average
  - 80% of sites handle only 20% of traffic
- Significant Energy Efficiency Opportunity due to idle resources



GSMA 2021



*Nokia Bell Labs based on NGMN data

Page 5 | © 2024 INTERNET INSTITUTE. All Rights Reserved.

Cloud Native 5G/6G | Data Warehouse

Connecting energy consumption, system operational metrics and quality of experience data

© 2024 INTERNET INSTITUTE. All Rights Reserved.

5G/6G Pilot Environment | Ljubljana | n77 | 3.8 GHz

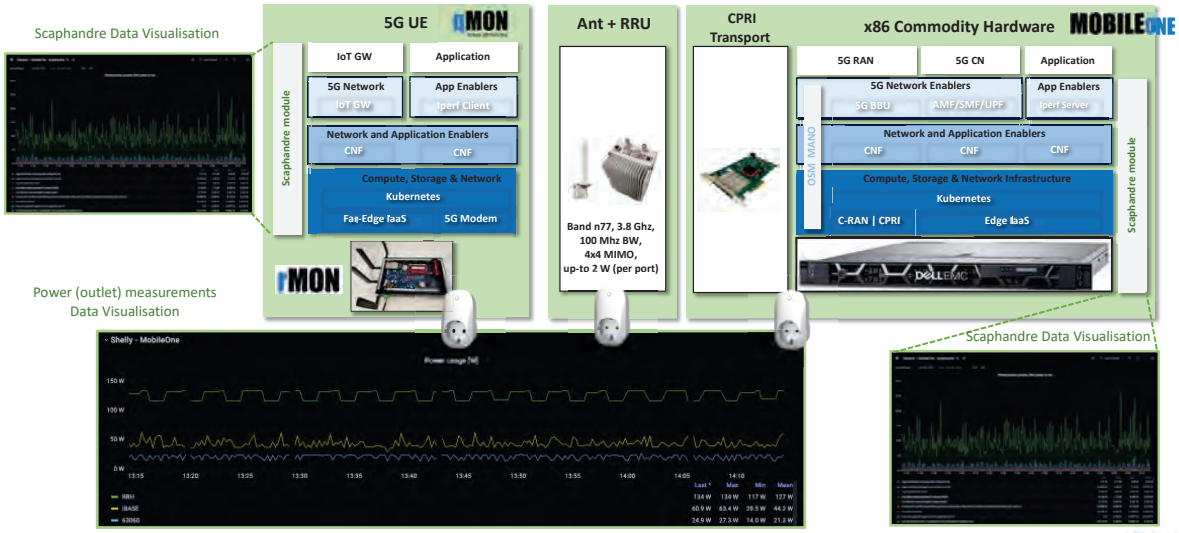
- Automated 5G System (SA) deployment MOBILEONE
  - 5G CN (10s)
  - 5G BBU/RRU (30s)
  - Edge / Kubernetes | MANO/OSM orchestrated
- 5G UE (SA) rMON
  - Industrial 5G GW (SA)
  - Far-edge (K3s)
- 5G Test Automation system qMON
  - Generating real application traffic in controlled fashion E2E
  - Collecting radio, network and application performance KPIs
    - radio sampling interval 1,5 s (RSRP, RSRQ, SINR, Tx Power…)
- Power consumption measurements
  - Power outlet level
    - RRU HW, Edge Server, 5G UE HW
  - HW component level
    - Server (input & output , RRU input, RAN output)
  - Socket level
    - CPU
  - Process level
    - Network e.g., BBU, 5G CN
    - Application

Process (level) power consumption measurements

Outlet (level) power consumption measurements

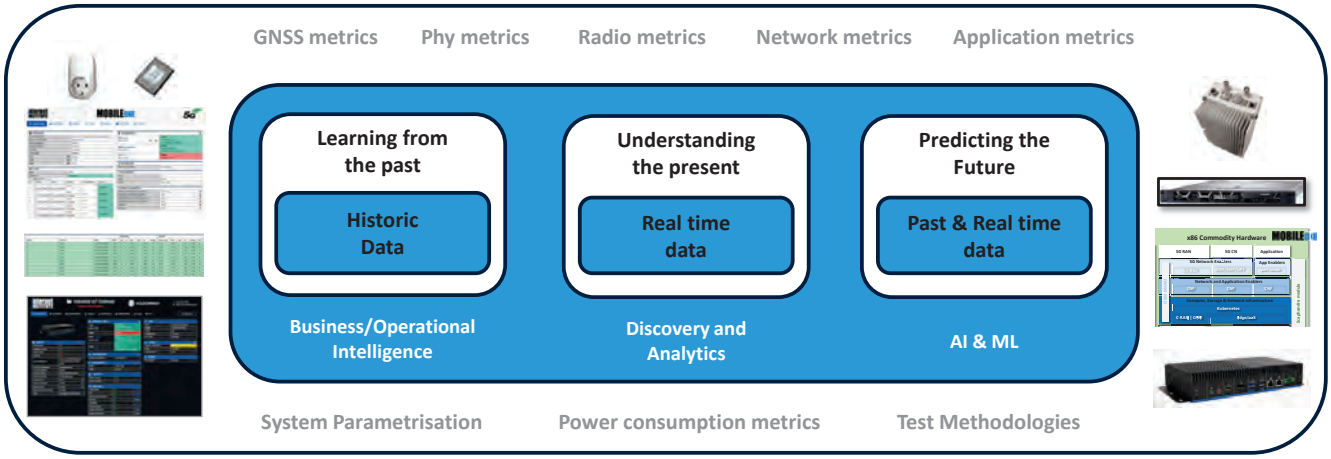
Page 7 | © 2024 INTERNET INSTITUTE. All Rights Reserved.

5G/6G Pilot Environment | Ljubljana | n77 | 3.8 GHz



Page 8 | © 2024 INTERNET INSTITUTE. All Rights Reserved.

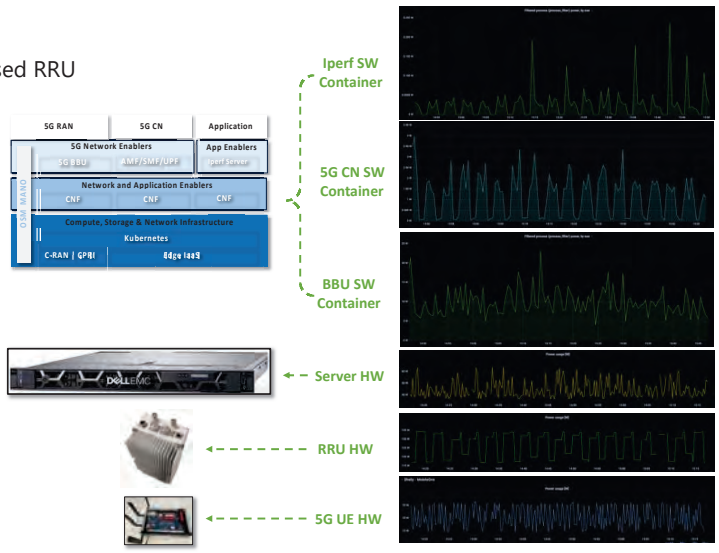
5G/6G Pilot Environment | Data Warehouse



Source: ININ (based on Qubole 2024)

Linking UE Throughput & Energy consumption 1/2

- Private 5GS | Edge IaaS extended with CPRI based RRU
  - Virtual 5G BBU
  - Virtual 5G CN
  - Edge cloud
- 5G UE (1) connected to Private 5GS
  - Industrial GW with Far-edge Cloud
- Application | Iperf cyclic test
  - Iperf Server deployed on Edge cloud
  - Iperf Client deployed on 5G UE
- Methodology
  - TCP based DL throughput test (14 sessions)
  - Duration 2 min (active load time)
  - Idle time 1 min

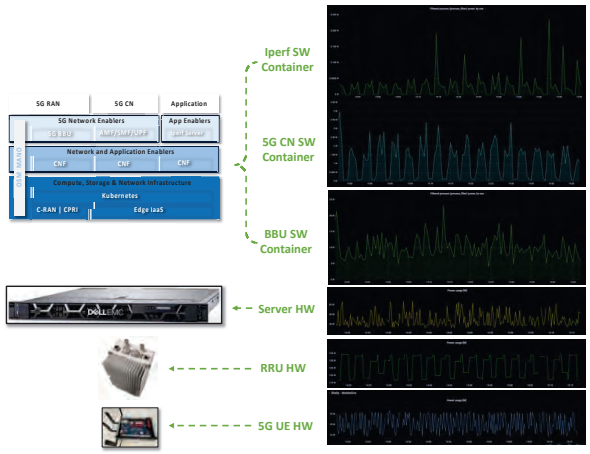


Linking UE Throughput & Energy consumption 2/2

- Test Results | Power Consumption (Mean)
  - Edge Server HW: 44,6 W
  - RRU HW: 127 W
  - BBU SW: 9,86 W
  - 5G CN: 1,1 W
  - Iperf SW: 0,02 W

  } HW level = 171,6 W

  } SW level = 10,98 W
- Power Consumption (Mean) | Active test only (DL)
  - 178 W (mean) – RRU HW + Server HW (includes all SW!)
  - 11,35 W (mean) – BBU SW + 5G CN SW + Iperf SW
- Even if the user is Idle, 5G system and 5G UE consumes power
  - HW standby mode (RRU, Server)
  - 5G NR and 5GCN standby mode (BBU, AMF, SMF,...)
  - Application standby mode
- Power consumption (RRU, 5G CN, Iperf Server) can have a distinctive patterns

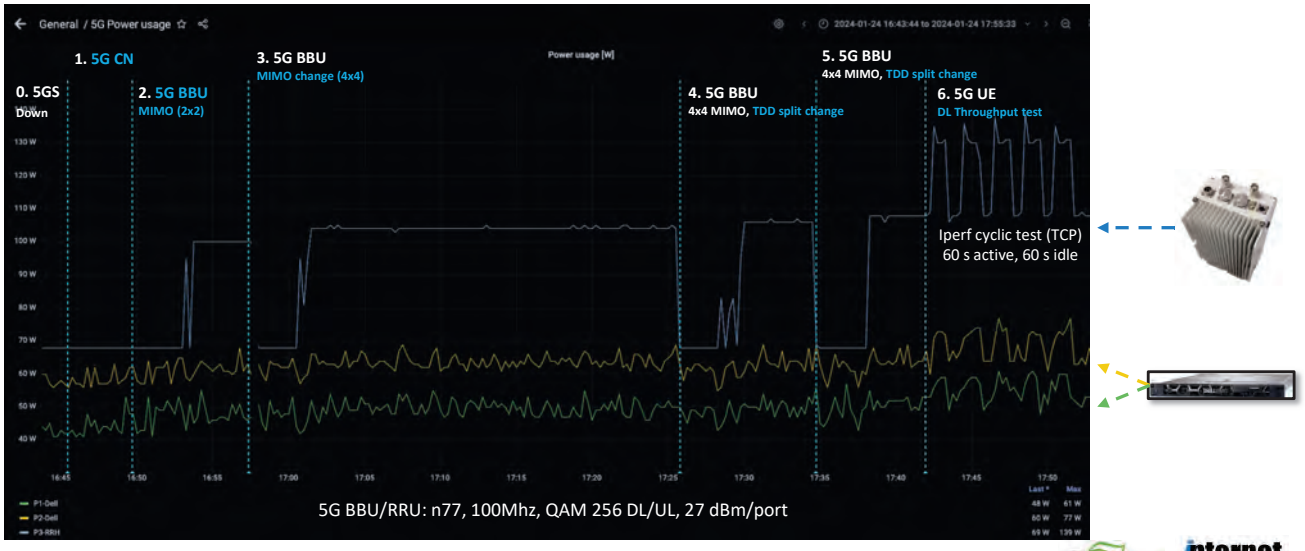


RAN: 80 % power consumption
RRU HW: 127 W
BBU SW: 9,8 W

Page 11 | © 2024 INTERNET INSTITUTE. All Rights Reserved.

Energy Consumption Patterns | 5G HW

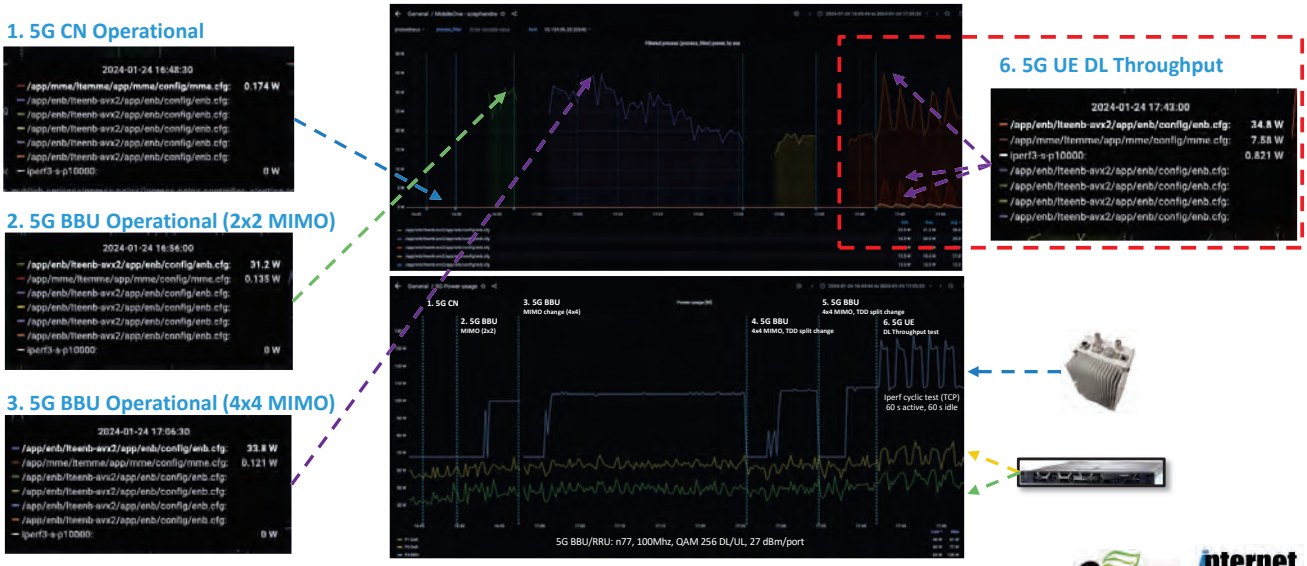
NR configuration (MIMO Level, TDD Split Ratio, etc.) and user traffic patterns (slicing, OSS user profile, etc.) have a direct impact on 5G RRU and IaaS power usage!



Page 12 | © 2024 INTERNET INSTITUTE. All Rights Reserved. **Source ININ: Power Usage - HW Level**

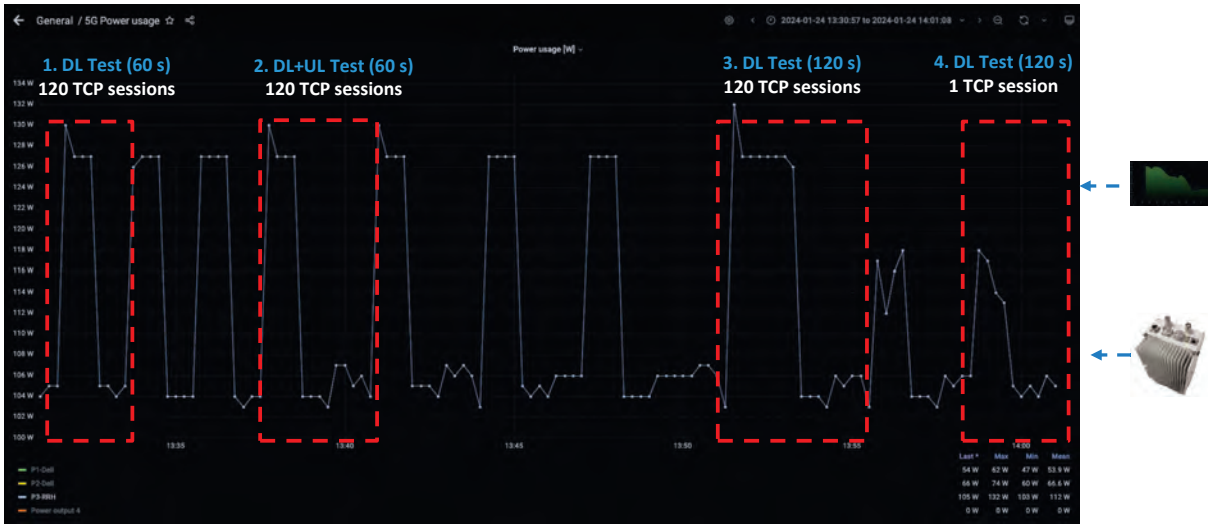
Energy Consumption Patterns | 5G SW

User traffic patterns (slicing, OSS BW user profile, etc.) have a direct impact on 5G BBU, 5G CN and App power usage!



 © 2024 INTERNET INSTITUTE. All Rights Reserved. Source ININ: Power Usage - Process Level

Energy Use Patterns | User Behaviour & Application design



 © 2024 INTERNET INSTITUTE. All Rights Reserved.

Changes in user behaviour and application design directly impact HW power usage of the deployed 5GS

Reducing mobile system power usage

A. Shut down cell?

B. Decrease power per radio port?

C. Optimise radio BW?

D. Optimise MIMO level?

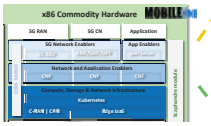
E. UE traffic shaping?

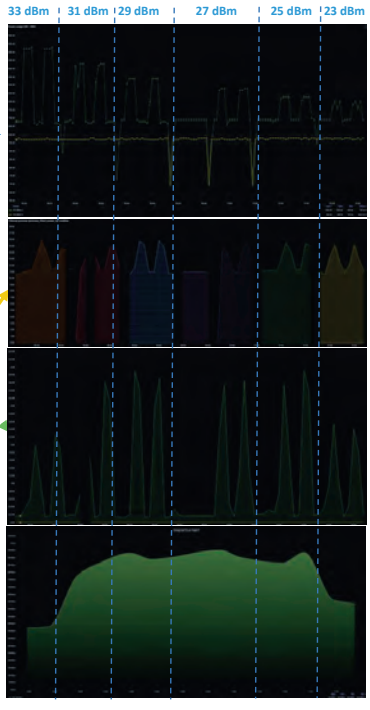
F. …

Decrease power per radio port

- RRU & BBU baseline
  - N77, 3800 MHz
  - qam256 DL/UL
  - 4x4 MIMO DL
  - BW 100 Mhz

- Results
  - RRU (HW) power optimisation
    - up to 21 % (peak)
  - BBU (SW) power optimisation
    - No optimisation
  - 5G CN (SW) power optimisation
    - up to 48 % (peak) – throughput dependent
  - User Experience
    - 430 Mbps (Max power, 33 dBm)
    - 580 Mbps (Min Power, 23 dBm)
    - 950 Mbps (27 dBm) – optimal?

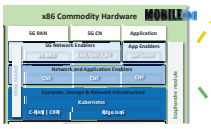


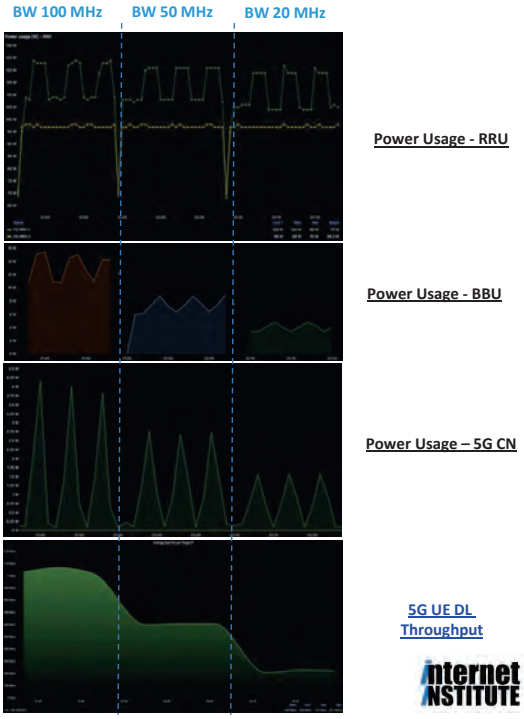


Radio BW Optimisation

- RRU & BBU baseline
  - N77, 3800 MHz
  - qam256 DL/UL
  - 4x4 MIMO DL
  - 25 dBm / port

- Results
  - RRU (HW) power optimisation
    - up to 4 % (peak)
  - BBU (SW) power optimisation
    - up to 69 % (peak)
  - 5G CN (SW) power optimisation
    - up to 62 % (peak) – throughput dependent
  - User Experience
    - 1 Gbps Throughput decreased to 230 Mbps



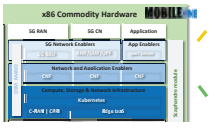


 © 2024 INTERNET INSTITUTE. All Rights Reserved.

MIMO Level Optimisation

- RRU & BBU baseline
  - N77, 3800 MHz
  - BW 100 MHz
  - qam256 DL/UL
  - 25 dBm / port

- Results
  - RRU (HW) power optimisation
    - up to 16 % (peak)
  - BBU (SW) power optimisation
    - up to 43 % (peak)
  - 5G CN (SW) power optimisation
    - up to 55 % (peak) – throughput dependent
  - User Experience:
    - 1 Gbps Throughput decreased to 390 Mbps
    - Lower MIMO level assures more stable radio link



 © 2024 INTERNET INSTITUTE. All Rights Reserved.

Conclusion 1/2

- **End-to-End Data Quality:** Collecting high-quality data across all performance and operational metrics is essential for historical analysis, current network status comprehension, and future system status forecasting
- **Energy and Load Correlation:** There's a direct correlation between energy consumption and mobile system load, emphasizing the need for efficient resource management
- **RAN's Role:** The Radio Access Network is the primary power consumer within mobile systems, highlighting the importance of its optimization
- **Efficiency Paradox:** Using “less is more” strategies (reducing radio power, MIMO level and MCS) can improve Quality of Experience (QoE) and overall system efficiency

Page 19 | © 2024 INTERNET INSTITUTE. All Rights Reserved.

Conclusion 2/2

- **Multidimensional Power Dependencies:** Power consumption is influenced by various factors including user behavior, application design, 5G OSS user profiles (considering Downlink/ Uplink Bandwidth and network slicing), and traffic patterns (balancing Uplink vs Downlink)
- **RRU Power Dynamics:** Remote Radio Unit (RRU) power per port is determined by the TDD profile, MIMO levels, and bandwidth utilized, necessitating careful configuration
- **Operational Diversity:** Mobile systems encounter a range of operational scenarios, each with unique requirements and objectives, requiring adaptable strategies
- **AI/ML Optimization:** AI and ML are pivotal in deciphering the complex, multidimensional puzzle of network optimization, offering solutions that balance multiple variables for optimal performance

Page 20 | © 2024 INTERNET INSTITUTE. All Rights Reserved.

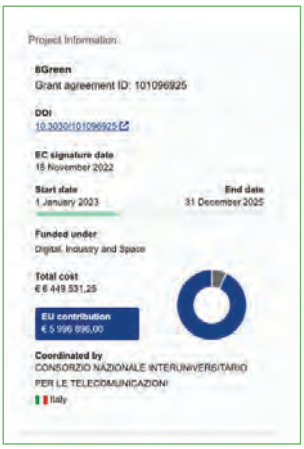
More information

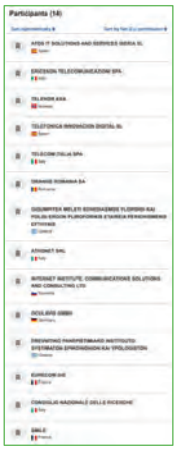
- Building sustainable mobile networks
  - 6Green SNS JU Project | www.6green.eu
- Measuring mobile system CO2 impact
  - Exigence SNS JU Project | projectexigence.eu

Page 21 | © 2024 INTERNET INSTITUTE. All Rights Reserved. Co-funded by the European Union

Co-funded by the European Union

Green Technologies for 5/6G Service-Based Architectures

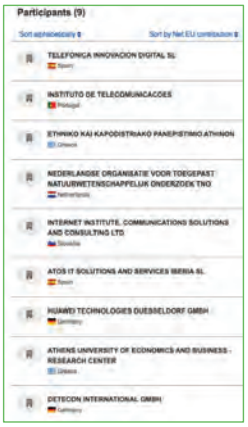


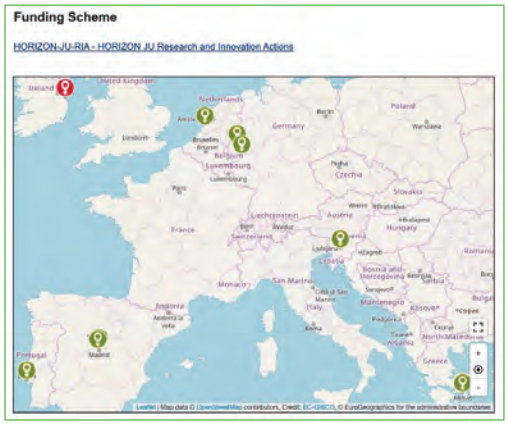


Page 22 | © 2024 INTERNET INSTITUTE. All Rights Reserved. https://cordis.europa.eu/project/id/101096925

Co-funded by the European Union

Devise & explore a novel approach for energy consumption and carbon footprint reduction of ICT services in the era of next-generation mobile telecommunications (6G)





 © 2024 INTERNET INSTITUTE. All Rights Reserved. **https://cordis.europa.eu/project/id/101139120**

Digitalni dvojček, senca ali model?

Digital twin, shadow or model?

Blaž Peternel

SMARTIS

POVZETEK

Digitalni dvojčki so nujni, ko želimo širokemu krogu uporabnikov in različnih programskih rešitev omogočiti varen in obvladljiv dostop do realnega sveta, kot so naprave, krmilni sistemi, merilna oprema in drugo. Pomembno je, da je digitalni dvojček uparjen s svojim realnim dvojčkom. Razvoj rešitev digitalnih dvojčkov pripelje v začetnih fazah predvsem do izziva usklajevanja potreb naročnikov z možnimi smernicami razvoja glede na razpoložljive tehnologije, razpoložljive podatkovne vire, integracijske standarde in potencialne omejitve obstoječega sistema. Digitalni dvojčki lahko predstavljajo podsisteme ali gradnike, praksa pa kaže, da so končne rešitve največkrat sestavljene iz več dvojčkov, ki smiselno tvorijo celoto, oz. dvojčka sistema sistemov. V razvojni fazi dvojčkov moramo tako upoštevati procese, lokacije in nabor karakterističnih spremenljivk in opremo, kjer se aktivno izvajajo funkcionalnosti dvojčkov ter smeri podatkovnih tokov. Dvojčki od roba do oblaka, ki večinoma združujejo IoT rob na eni strani (kot del OT omrežij – t.i. Operational Technology) in obstoječe IT sisteme na drugi, odpirajo tudi mnogo novih vektorjev kibernetskih groženj. Na koncu pa se je vseeno treba vprašati ali potrebujemo modele, digitalne sence ali digitalne dvojčke.

SUMMARY

We need digital twins when we want to provide a wide range of users and different software solutions with safe and manageable access to the real world, such as devices, control systems, measuring equipment and more. It is important that the digital twin is paired with its real twin. The development of digital twin solutions leads, in the initial stages, to the challenge of harmonizing the needs of clients with possible development guidelines based on available technologies, available data sources, integration standards and potential limitations of the existing system. Digital twins can represent subsystems or building blocks, but practice shows that the final solutions most often consist of several twins that meaningfully form a whole, or twin system of systems. In the development phase of twins, we must therefore take into account the processes, locations and set of characteristic variables and equipment where the functionality of twins and the directions of data flows are actively implemented. Twins from the edge to the cloud, which mostly combine the IoT edge on the one hand (as part of OT networks - the so-called Operational Technology) and existing IT systems on the other, also open many new vectors of cyber threats. In the end, however, one must ask whether we need models, digital shadows or digital twins.

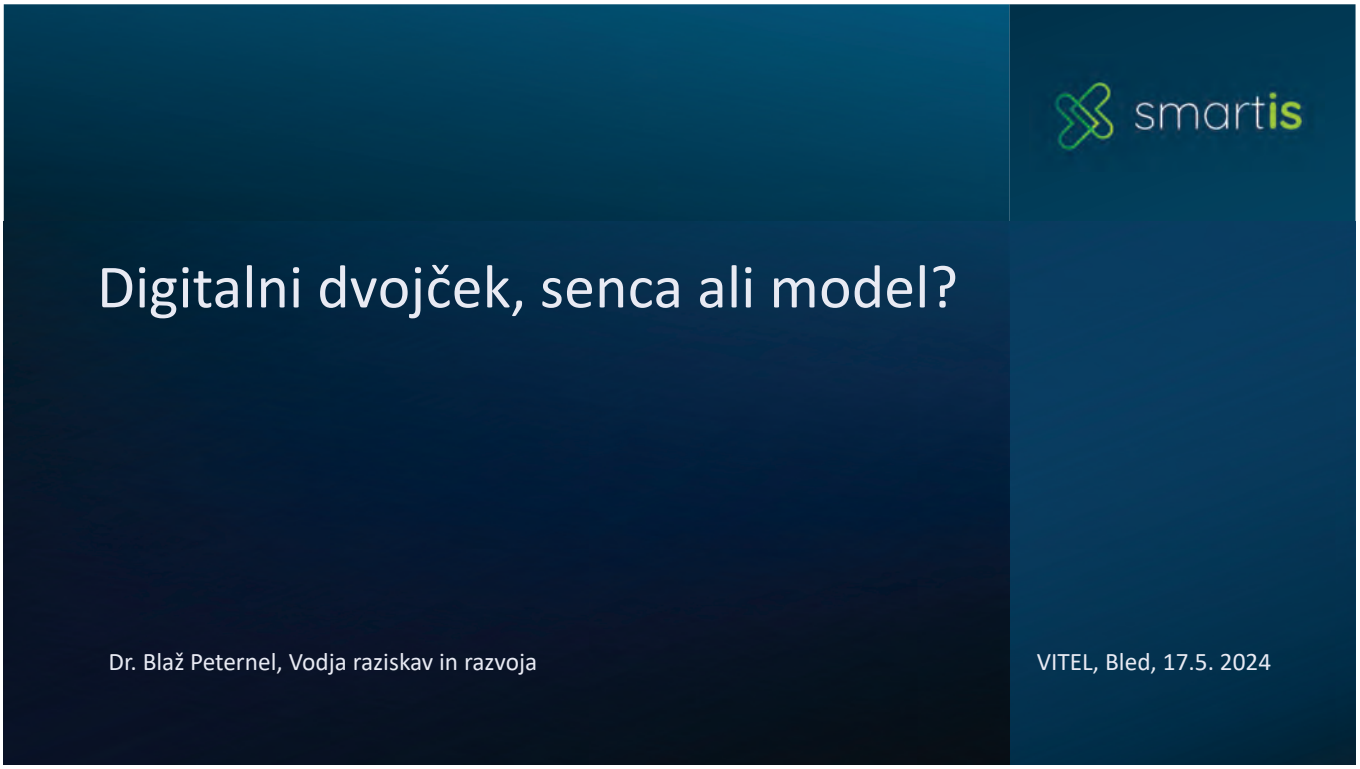
O AVTORJU

Blaž Peternel je doktoriral je iz področja elektrotehnike na Univerzi v Ljubljani. Deset let je vodil razvoj različnih tehnoloških projektov pri operaterju telekomunikacij, šest let pa vodi področje raziskav in razvoja v podjetju SmartIS. Njegove raziskovalno-razvojne aktivnosti so usmerjene v širok spekter interdisciplinarnih področij rešitev digitalnih platform (digitalni dvojčki, energetsko upravljanje, prediktivno upravljanje, pametna mesta, skupnosti in domovi).

ABOUT THE AUTHOR

Blaž Peternel received his doctorate in the field of electrical engineering from the University of Ljubljana. For ten years, he managed the development of various technological projects at a telecommunications operator, and for six years he headed the area of research and development at the company SmartIS. His research and development activities are aimed at a wide range of interdisciplinary areas of digital platform solutions (digital twins, energy management, predictive management, smart cities, communities and homes).


smartis


smartis
Digitalni dvojček, senca ali model?
Dr. Blaž Peternel, Vodja raziskav in razvoja
VITEL, Bled, 17.5. 2024

Potek predavanja

- SmartIS
- Potrebe po rešitvah digitalnih dvojčkov
- Od roba do oblaka
- IT pogled in sistem sistemov
- Praksa

smartis
PAMETNI INFORMACIJSKI SISTEMI

IT business services
- Enterprise Asset Management
- Software Asset Management
- Services and Asset Management
- Development and Project Management

IT technology services
- Hyperconverged Infrastructure
- High Performance Computing
- AIOps & APM
- Backup & Archiving

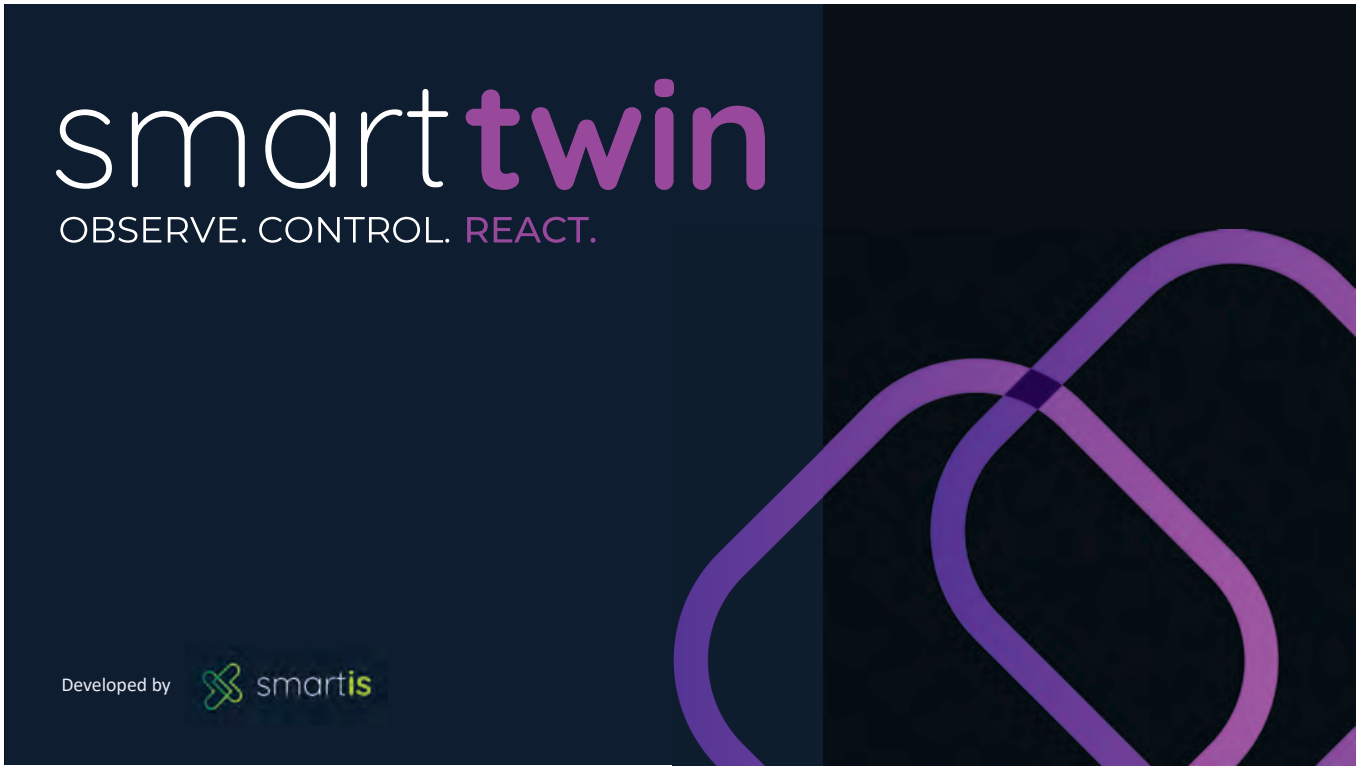
Cyber security
- Secure Gateway
- Identity & Access
- Information Protection
- Detection and Response
- Security Operations
- Critical Infrastructure Protection

Custom development
- Custom Software Development Services
- Smart DPO
- IoT/ IIoT Platform
- Digital Twin Platform
- Research and development

Employees
70+

Experience
14+ years

Locations
2



OUR BACKGROUND

- 10 years of development
- Platform's third iteration
- Beginning in enterprise energy management

OUR VISION

- Ensure reliable and easy to use IoT platform
- Expansion to industry sector
- Continuous development based on user feedback

smarttwin

OBSERVE. CONTROL. REACT.

Next-gen IoT platform for digital twinning intended to establish central monitoring system of operational and maintenance processes.

Developed by smartis

What is digital twin

Digital twin bridges physical assets and the digital world.

Input data
(SCADAS, REST APIs, DBs, ..)

Visulization
(3D models, schemas, ...)

Files
(docs, videos, ...)

Data processing engine
Near-real-time data ingestion, validation, interpretation, quality, science, ...
Dedicated high efficiency data processing service.

Data storage
Hot storage for long term data retention. Conceptualized on openness.
Ready for integration.

Data visualization
Personalized custom dashboard long-term changing data (KPIs).
Personalized monitor dashboard for near-real-time monitoring.

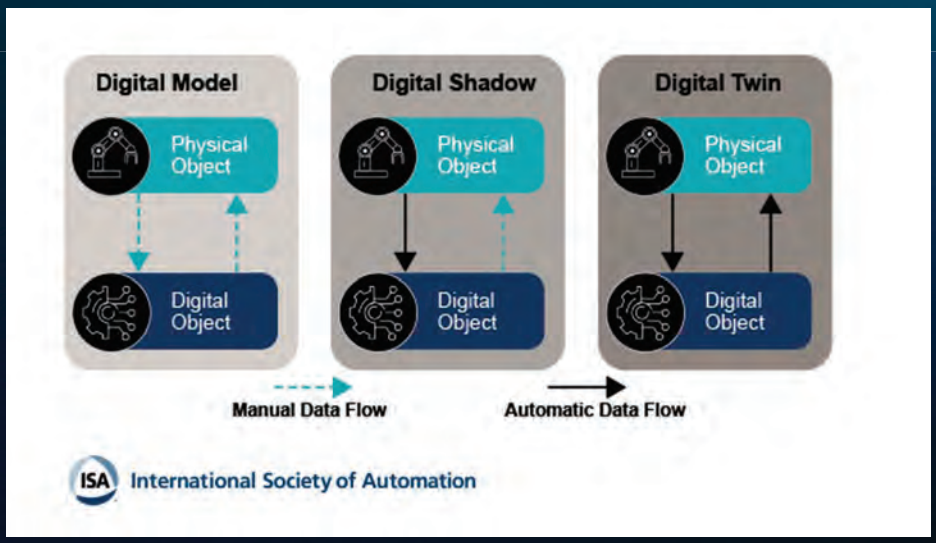
Reporting
Personalized custom reports (manual, scheduled, alarm triggered).

Tasks and notifications
Simple task management and notifications via SMS, APP, e-mail.

Maintenance (WIP)
Standard operational procedure workflow.
Standard maintenance procedures for preventive maintenance.

Visualizations
Micro or macro 3D models.
2D schemas.

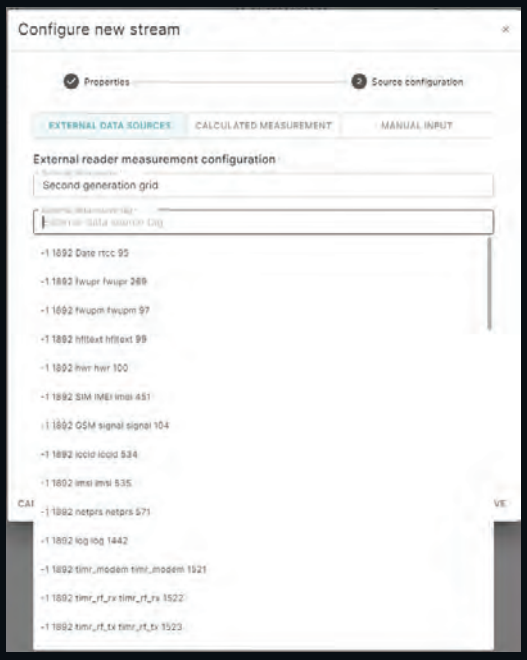
Model, senca, dvojček?



Integrations

Ingest from any input or push to any output.

- Agnostic and no vendor lock-in
- MQTT broker
- Kafka



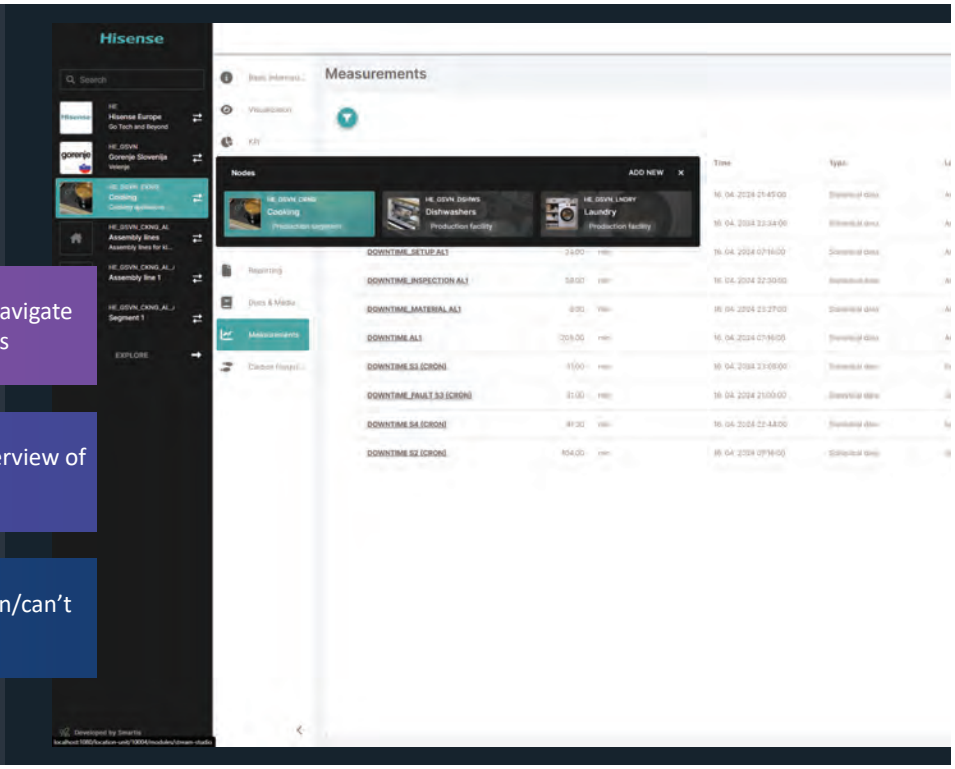
Hierarchy

Build a hierarchy that resembles real world operations.

Gives structure and ability to navigate through complex systems

Gives user ability to have an overview of sub systems

Authorization of what users can/can't see



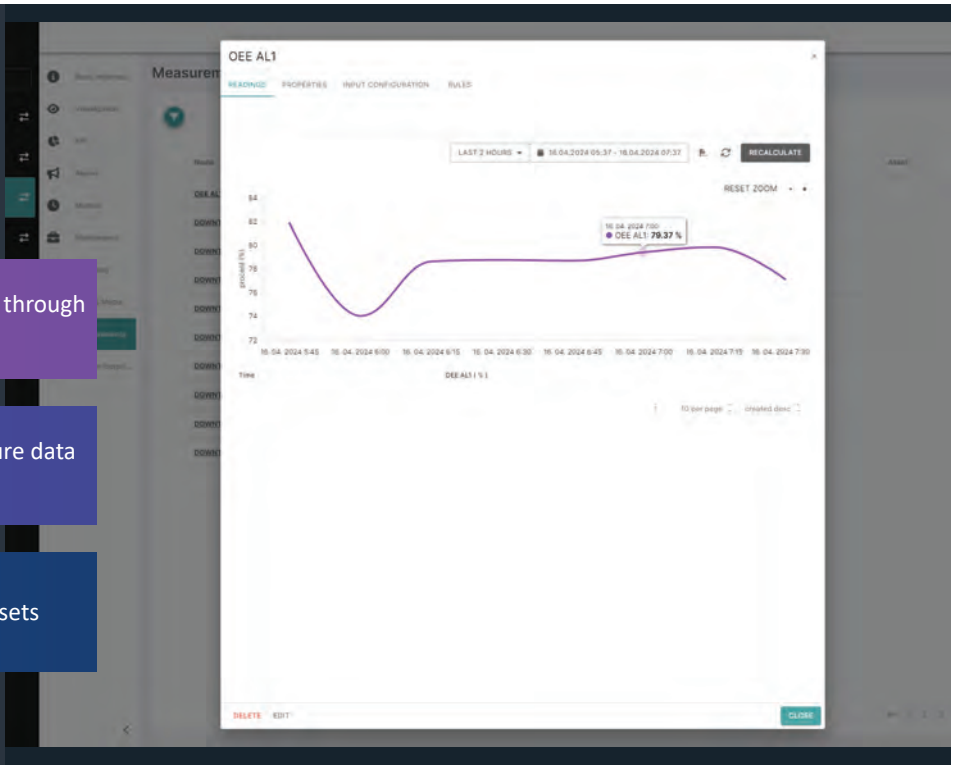
Data Streams

Connect to data sources and organize them into data streams.

Organize data into data streams through integrations

Set stream properties and ensure data quality

Create data flows and data sets



Processing

Extract value from the data though data transformation.

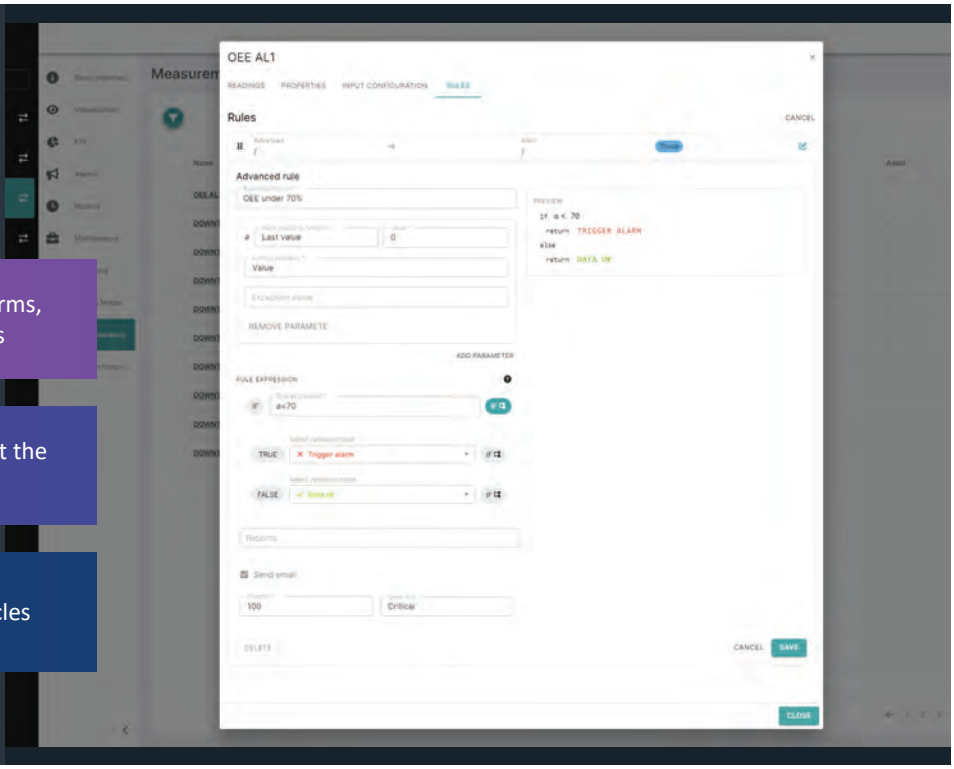
- Near real-time data processing
- Big data
- Open to receive and send data

- Data normalization
- Data interpretation
- Data validation
- Meta data
- Interpolation
- Extrapolation
- Mapping

Rules

Validate da through simple or complex rules.

- Rules create triggers for alarms, notifications and reports
- All responsible personnel get the notification
- Severity levels and rule cycles



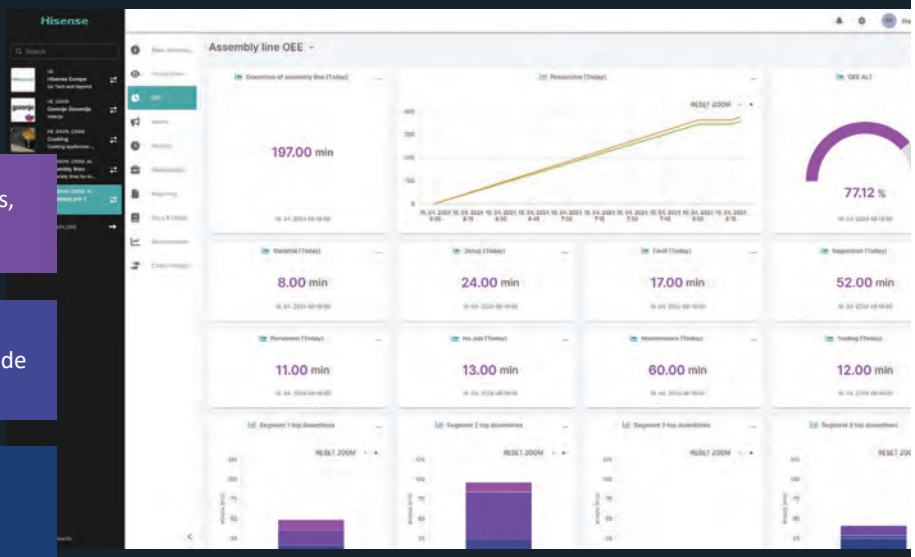
Dashboards

Create multiple personalized dashboards.

Create custom cards (graphs, pies, gauges, charts, …)

Color pallets for light and dark mode

Monitoring dashboards and KPI dashboards



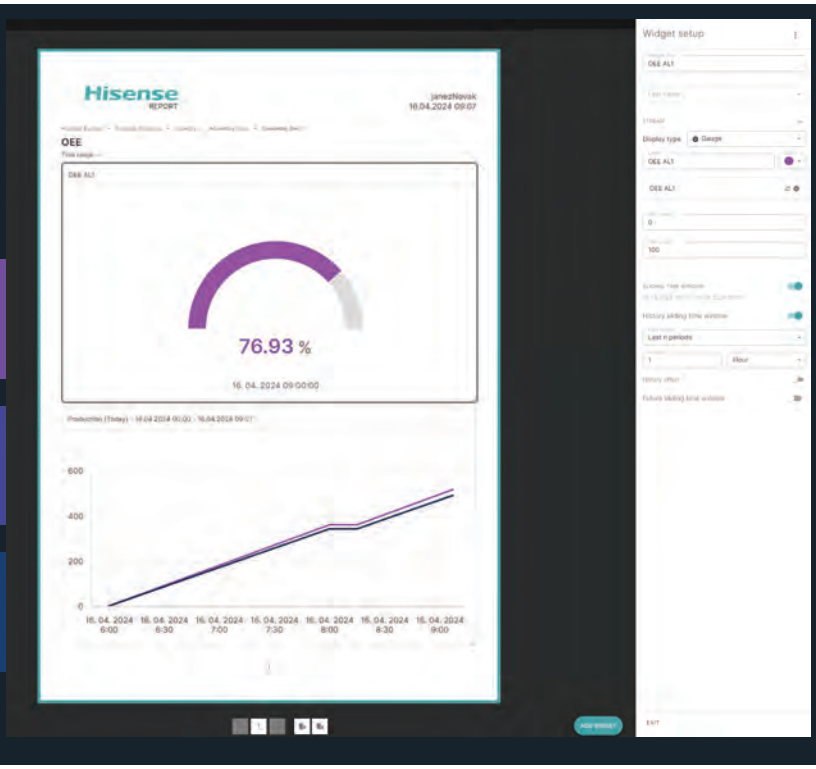
Reports

Personalized reports for reporting statuses.

Manual, schedule and rule triggered reports.

Same concept as dashboards.

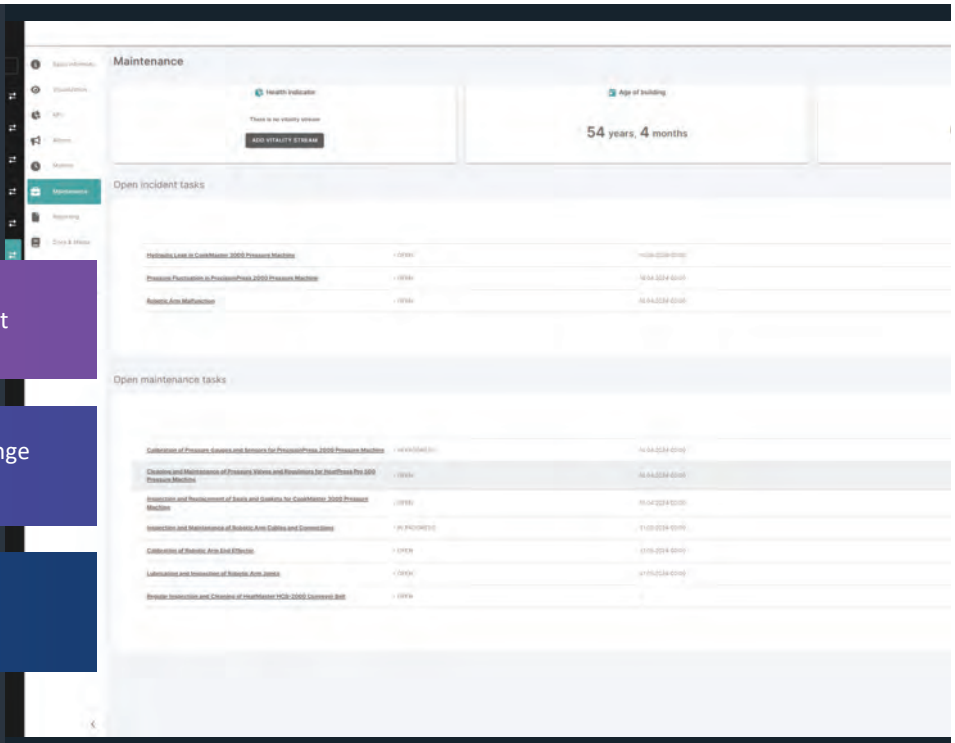
Report presets and templates.



Maintenance

Complete workflow.

- Simple task management
- Task status and status change notifications
- SMP and SOP (WIP)



Visualization

Visualize you operations.

- Support 3D models (BIM concept) and 2D schemas
- Add labels of alarms, assets/units, measurements, status.
- BIM concept

Podporniki

Sponzors

TelekomSlovenije

ERICSSON

HUAWEI

Univerza v Mariboru

Fakulteta za elektrotehniko, računalništvo in informatiko

Univerza *v Ljubljani*
Fakulteta *za elektrotehniko*

ZANESLJIV PARTNER ZA VSE IKT-IZZIVE

CELOVITE 360° IKT-REŠITVE ZA DIGITALNO PRIHODNOST

V Telekomu Slovenije ne pristajamo na kompromise in smo podjetjem zanesljiv partner pri digitalizaciji poslovanja.

Naše napredne **IKT-rešitve** omogočajo, da procesi in komunikacija v vašem podjetju deluje hitro, varno in stabilno. Za zanesljivost in **kibernetsko varnost** vašega poslovanja skrbijo naši visoko kvalificirani strokovnjaki 24 ur na dan in vse dni v letu. Z rešitvami **IoT** že danes prispevamo k digitalni preobrazbi celotne družbe. Na krilih tehnologije **5G** razvijamo nove specializirane rešitve za digitalno prihodnost, za še boljšo povezljivost in višjo učinkovitost poslovanja.

Vse 360° poslovne storitve in znanja prihodnosti na enem mestu, za še boljši, varnejši in bolj povezan jutri. Vedno na boljše.

POSLUJTE VARNO. POSLUJTE DIGITALNO.

Telekom Slovenije, d.d., Ljubljana.

TelekomSlovenije
Vedno na boljše.


ERICSSON
AI business potential
understanding the value of AI for telecom operations

Slovensko društvo za elektronske komunikacije

Elektrotehniška zveza Slovenije