EUROPEAN
LANGUAGE
DATA SPACE

The Role of Data in the Conversational AI

Jan Cuřín (The MAMA AI, SE, Czechia)
jan.curin@themama.ai

02-12-2024 LDS Country Workshop, ÚFAL MFF UK, Charles University, Prague, Czechia
https://language-data-space.ec.europa.eu

The MAMA AI Introduction and Team History

The MAMA AI (2021-present)

- Company that build secure, practical and sustainable AI
- AI Products and Services for on-prem and SaaS
- Trusted and top-skilled international partner

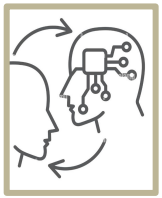
IBM Watson (2014-2021)

- Conversational, Speech & NLP on IBM Cloud
  - Watson Assistant
  - Watson Speech To Text
  - Watson Text To Speech
  - Watson Language Translator

IBM Research (1992-2021)

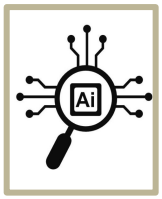
- Basic research in the fields of Speech recognition, Speech synthesis, Machine Translation, Conversational technologies

MAMA AI Fields of Interest



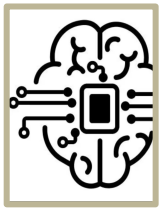
Communication AI

virtual assistants
artificial voices



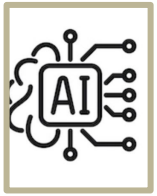
Search AI

semantic search, RAG, compliance,
molecule search (chemistry)



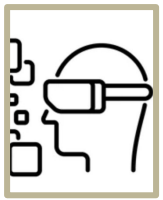
Reasoning & Generative AI

local LLMs, digital twins,
predictive maintenance



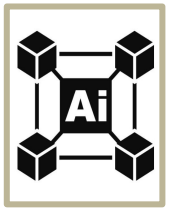
Personalized AI

smart reporting &
recommendations



Metaverse AI

virtual humans
complex simulations



Blockchain AI

digital data
tokenomics

Language and Speech Technologies

- Language Technologies are directly impacted by recent advances in AI
- Deep learning applies to many language technologies:
  - Automatic Speech Recognition (ASR)
  - Text to Speech Systems (TTS) - Deep voices
  - Dialog Systems / Chatbots / Voice Bots
  - Question-Answering Systems (QA)
  - Named Entity Recognition (NER)
  - Text Summarisation
  - Sentiment Analysis
  - Machine Translation (MT)

Data for Conversational AI

- **Dialog Systems / Chatbots / Voice Bots**
  - Bots need good, clean data for the best performance
  - Safeguards to avoid hallucination (collecting data)
  - Fine tuning of on-prem LLMs for particular domain or language (Czech …)
  - Good metadata associated with data sources
  - Data augmentation using LLMs
- **Text to Speech Systems (TTS) - Deep voices**
  - High quality voice clones (digital twins) needs high quality recordings
  - TTS Normalization (numbers, dates, names, abbreviations,)

AI needs IA

There's No AI (Artificial Intelligence) without IA (Information Architecture)

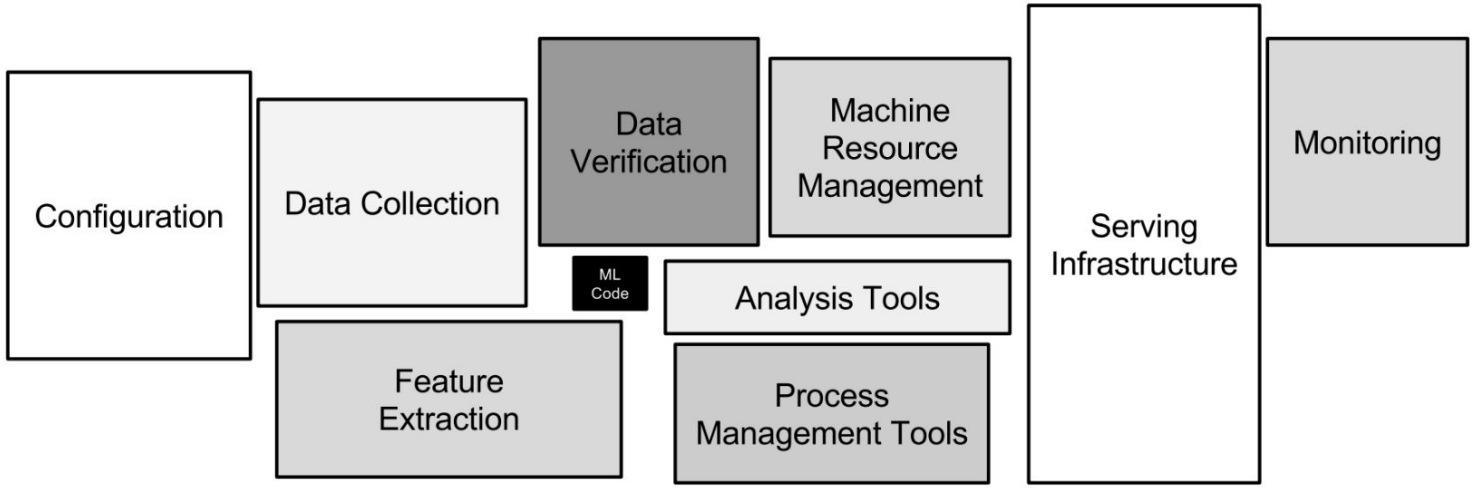
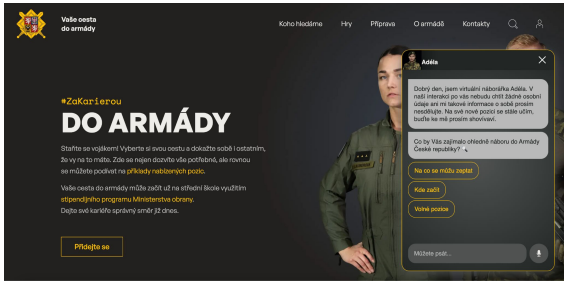


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown

Source: Scully & team: Hidden Technical Debt in Machine Learning Systems - pdf

Examples of cooperation with research projects and government

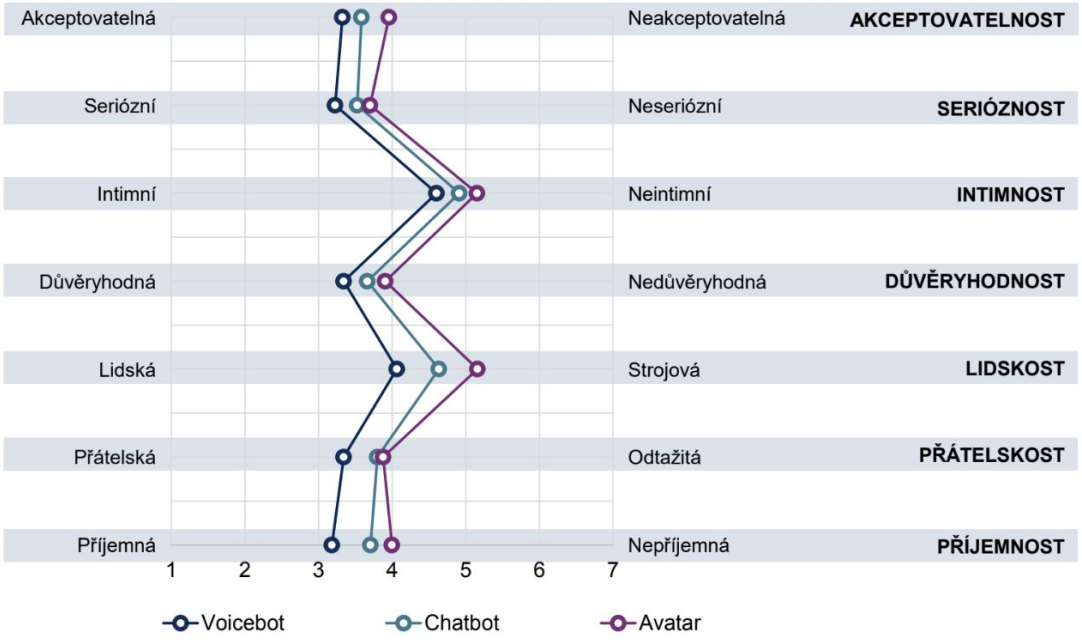
- Public sector
  - **Virtual Recruiter Adéla** for Ministry of Defence & Armed Forces
    - 1st place in AI Awards 2024 in the category AI for Government
  - **Covid Reports** with Institute of Health Information and Statistics (ÚZIS)
    - 2020-2021 - more than 2M reports generated (6250 municipalities)
    - Everyday reports on Mobilní Rozhlas platform (now Munipolis)
    - Open Data access: Open Data initiative - important for SMEs and startups



- Cooperation with Universities and Public institutions
  - **Newsroom AI**: public service in the era of automated journalism
    - Technology Agency of the Czech Republic (ID TQ01000100)
    - Charles University (FSV), CVUT, Czech Radio
    - Virtual assistant: chat vs. audio vs. avatar research with 1200 respondents in April 2024
    - iRozhlas - feedback on article recommendation in November 2024

HODNOCENÍ VIRTUÁLNÍ ASISTENTKY: VOICEBOT, CHATBOT A AVATAR

Ze tří forem virtuální asistentky je nejlépe vnímán voicebot, naopak avatar se kloní spíše k druhému, negativnímu spektru. V rámci přátelskosti však dohání chatbot.



- Za intimního považují voicebota častěji lidé se základním vzděláním (37 %), zatímco virtuální asistentku pak lidé s vyučením (28 %)
- Příjemný přijde chatbot častěji lidem ve věku 25-34 let (57 %). Seriózně chatbot působí častěji na lidi do 34 let.
- Obecně lidé starší 65 let zaujímají negativní postoje. Nedůvěryhodný přijde chatbot častěji právě jim. Častěji pak na ně odtažitě působí avatar (41 %). A neakceptovatelný jim přijde jak chatbot (38 %), tak avatar (43 %).

__Pozn.:__ Zobrazeno pomocí sémantického diferenciálu, který zachycuje vnímání respondenta a jeho hodnotovou orientaci.
__Báze:__ 04/2024: n=1232
__Otázka:__ HMC2. Nyní, prosím, do jaké míry si myslíte, že je virtuální asistentka v podobě voicebota/chatbota/avatara…

Doporučujeme

‚Šestihlavicová saň.' Kreml hrozí jadernou válkou, kvůli nové raketě aktivoval horkou linku s USA

28. 11. 2024

Kauza Stoka se vrací na začátek. Vrchní soud v Olomouci v neveřejném jednání zrušil rozsudek

28. 11. 2024

Mladí muži a ženy se vzdalují, trend přichází i do Česka. ‚Řeší se identity, ale nejde o válku pohlaví'

28. 11. 2024

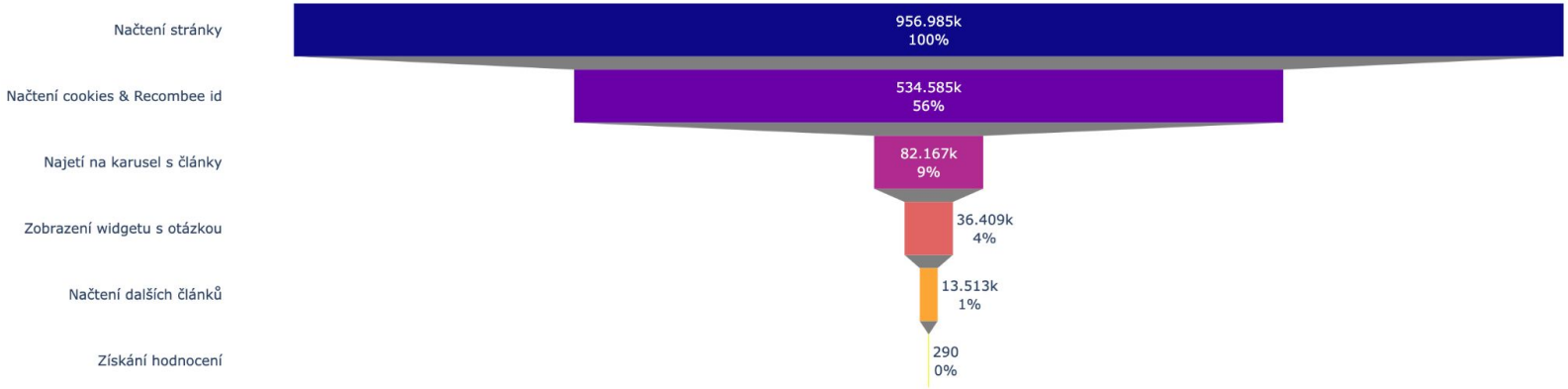
Rubl se od počátku propadá. Dodavatel Ruska kvůli tomu r

28. 11. 2024

Jste spokojeni s doporučenými články?

ANO NE

Funnel získávání zpětné vazby pro rekomendace (poslednich 10 dnu)



Data Sharing from Private Companies

Data sharing works well for the (at least partially) publicly funded projects when the data are one of the expected outcome or when the results are presented in the aggregate/statistical manner (such as for research agencies)

But in general, it is quite difficult for private companies (those in the EU) to share data publicly:

- **Data Protection Laws**: The EU has strict data protection laws, particularly under the General Data Protection Regulation (GDPR)
- **Intellectual Property**: Data may also be considered as a form of intellectual property, particularly if it has been derived from the company's own research and development activities
- **Data Quality and Management**: Once published, ensuring that data is accurate, up-to-date, and properly formatted for public consumption can be a time-consuming and costly process
- **Third-Party Agreements**: Companies usually have agreements with third parties (like suppliers, partners, or customers) that prohibit them from disclosing their data

In summary, while data sharing can have numerous benefits, it must be balanced with these considerations to ensure legal compliance and protect the company's interests.

Common European Language Data Space

Thank you!

A Common European Language Data Space – funded under contract LC-01936389 with the European Union.

Jan Cuřín (The MAMA AI, SE, Czechia)
jan.curin@themama.ai

02-12-2024 Language Data Space Workshop, ÚFAL MFF UK, Charles University, Prague, Czechia
https://language-data-space.ec.europa.eu