Bundesministerium
Klimaschutz, Umwelt,
Energie, Mobilität,
Innovation und Technologie

Bundesministerium
Finanzen

1. Vernetzungstreffen im Rahmen des AI Policy Forums

KI in der öffentlichen Verwaltung

Mittwoch, 19. Oktober 2022
Festsaal Technisches Museum Wien

AIM AT 2030
Artificial Intelligence Mission Austria

Themensession 1: **Daten und KI in der Verwaltung**

- **Alexander Schindler**, AIT - *Daten und KI - eine Einführung*
- **Alexandra Ebert**, mostly.ai - *Synthetische Daten*
- **Mihai Paunescu**, Predictive Analytics Competence Center, BMF - *Predictive Analytics in der Finanzverwaltung*
- **Manfred Gruber**, Bundeskanzleramt - *Digitales Gedächtnis Österreich*

Daten und Künstliche Intelligenz

Eine Einführung

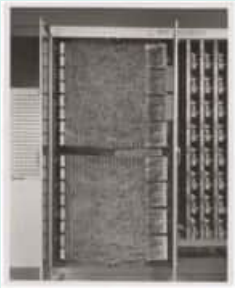
ALEXANDER SCHINDLER
Thematic Coordinator Researchfield Datascience
Data Science & Artificial Intelligence
Center for Digital Safety & Security

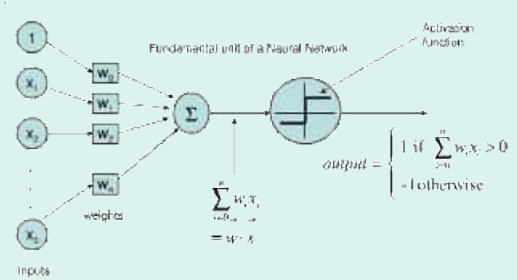
AIT Austrian Institute of Technology GmbH
Giefinggasse 4 | 1210 Vienna | Austria
T +43 50550-2902 | M +43 664 8251454
alexander.schindler@ait.ac.at | www.ait.ac.at

Was ist Künstliche Intelligenz?

Dartmouth Conference (1956)

- Namensgebung: „Artificial Intelligence“
- Multi-disziplinär
  - Philosophy (z.B. Descartes, Leibnitz)
  - Logik / Mathematik (z.B. Gödel)
  - Informatik (z.B. Turing, von Neumann)
  - Psychologie / Kognitionswissenschaften (Wissensrepresentationen)
  - Biologie / Neuro-Wissenschaften (Konnektivismus, Neural Networks)
  - Evolution (Genetic Programming)





Perceptron (Frank Rosenblatt, 1958)

- Grundbestandteil von Neuronalen Netzwerken

HISTORISCHER ÜBERBLICK

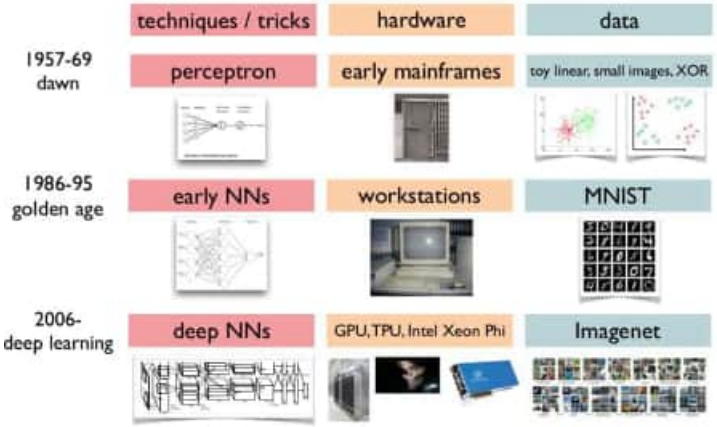
AUSTRIAN INSTITUTE OF TECHNOLOGY

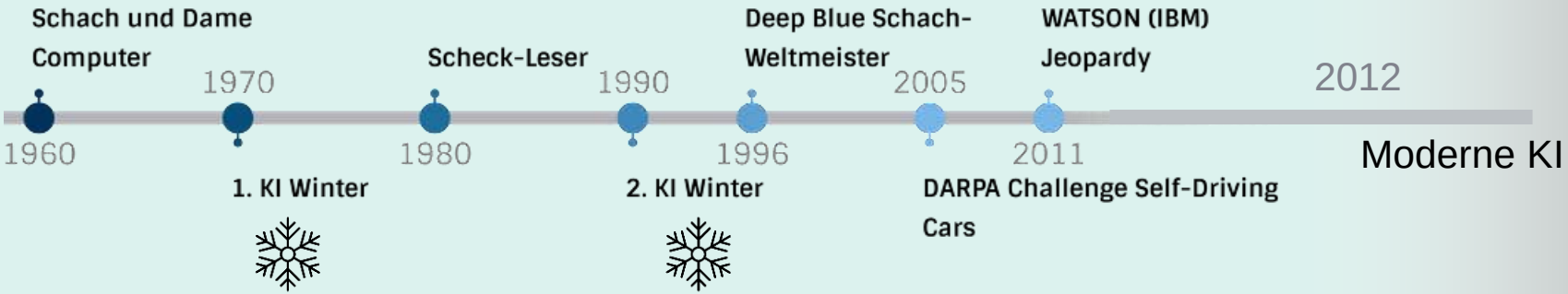
Hohe Erwartungshaltung

- Großes Investment vom Militär
- Utopische Vorstellungen

Mangelhafte Leistung

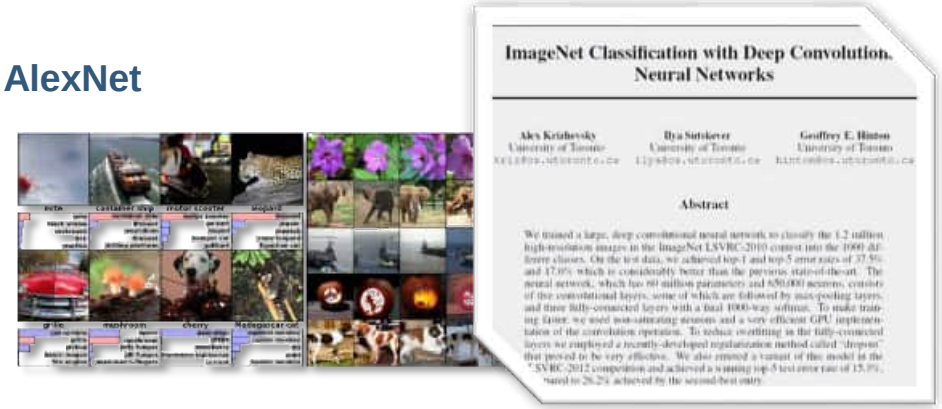
- Langsame Computer
- Kleine Datensets / Teure Datenspeicher
- Viele Probleme noch nicht gelöst
- Zu wenige „Experten“

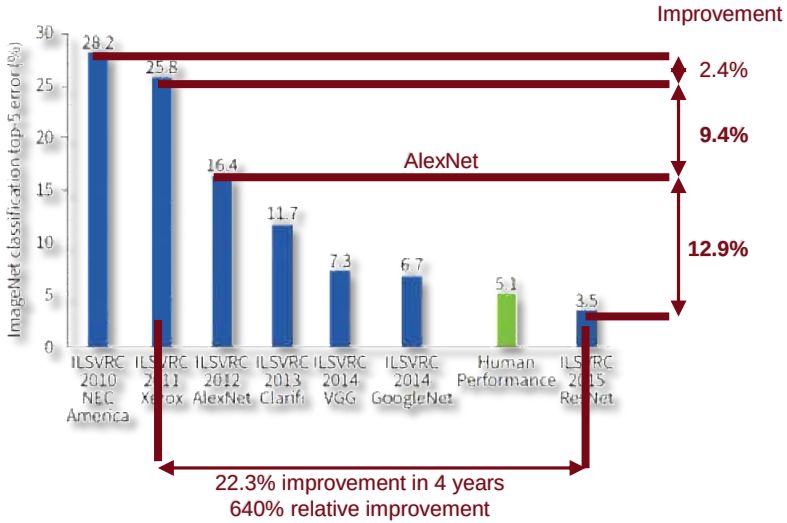




MODERNE KI

AlexNet





GPUs

Verbraucher Hardware

- Leistbare Hardware statt Supercomputer
- ➔ große Rechenleistung für PhD Studenten

KI – WISSENSCHAFT VS. HYPE

Künstliche Intelligenz …

… in der Industrie

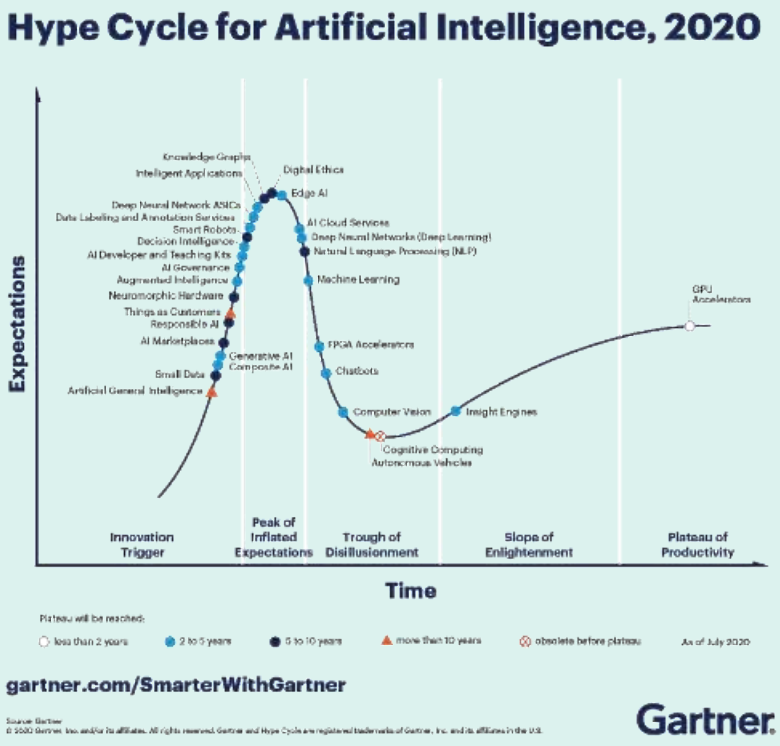
- Künstliche Intelligenz ⇔ Machine Learning

… in der öffentlichen Verwaltung

- Künstliche Intelligenz ⇔ Digitalisierung 2.0

… in der Werbung

- Die nächste Superlative
- Smart, Intelligent, …



DEFINITION
KÜNSTLICHE INTELLIGENZ

Definition von High-Level Expert Group on Artificial Intelligence (EU)

Artificial intelligence (AI) systems are **software** (and possibly also hardware) **systems** designed by humans(2) that, given a **complex goal**, act in the physical or digital dimension by **perceiving their environment through data acquisition**, **interpreting the collected structured or unstructured data**, **reasoning on the knowledge**, or processing the information, derived from this data and **deciding the best action**(s) to take to **achieve the given goal**. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions

Definition von Österreichischer Rat für Robotik und Künstliche Intelligenz (AT)

"… artificial intelligence (AI) refers to **computer systems** that **exhibit intelligent behavior**, i.e., that are capable of performing tasks that **in the past required human cognition and decision-making skills**. Artificial intelligence-based systems **analyze their environment** and **act autonomously** to **achieve specific goals**. … They operate through rule knowledge created by experts or based on statistical models derived from data (machine learning, e.g., deep learning). The term AI includes both pure software, but can also include hardware, such as in the case of autonomous robots. …"

Table 1. AI domains and subdomains constituting one part of the operational definition of AI

	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

Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2020). AI Watch Defining Artificial Intelligence. Publications Office of the European Union.

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

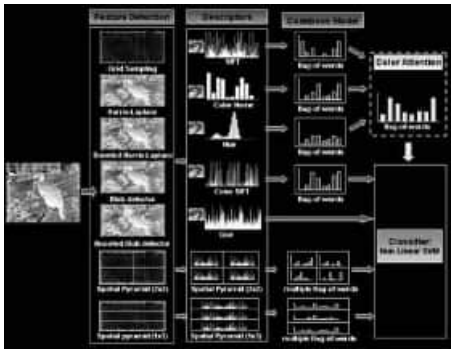
WIESO BAUCHT EINE KI SO VIELE DATEN?

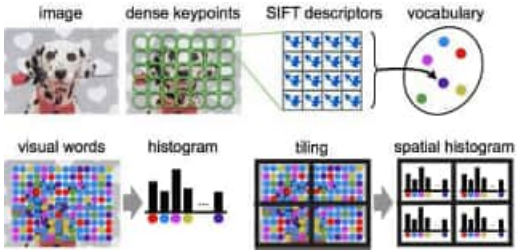
Oder: Was sind Selbstlernende Systeme?

BEISPIEL:
KLASSISCHE BILDVERARBEITUNG

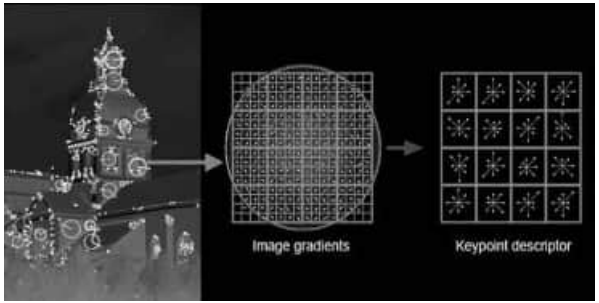
Objekterkennung

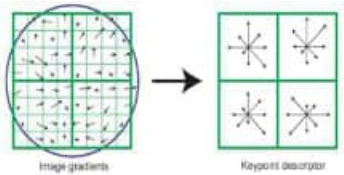
- Bildmerkmale werden von Experten definiert und automatisiert extrahiert
- Machine Learning Modell auf diesen Merkmalen trainiert um Objekte zu erkennen





Ahuja, Sarthak, and Anchita Goel. "Scene Recognition using Bag-of-Words." Forest 100: 228.



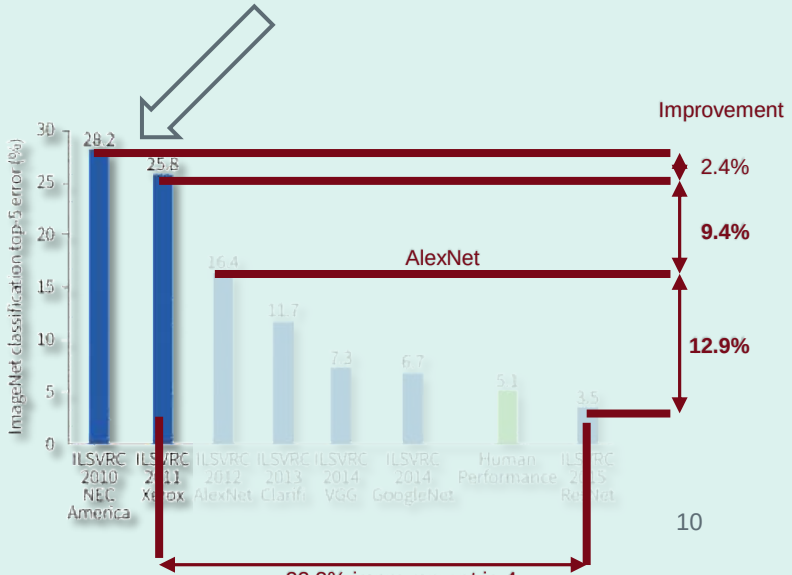


Vorteil

- Funktioniert mit wenigen Daten

Nachteil

- Benötigt hoch-spezifisches Expertenwissen
- Erreicht nicht die Leistung von State-of-the-art Deep Learning Methoden



DEEP LEARNING

Selbstlernende Systeme

- Modell „lernt“ welche Merkmale zum Lösen der Aufgabe relevant sind
- Bessere Anpassung an die Daten/Aufgabe

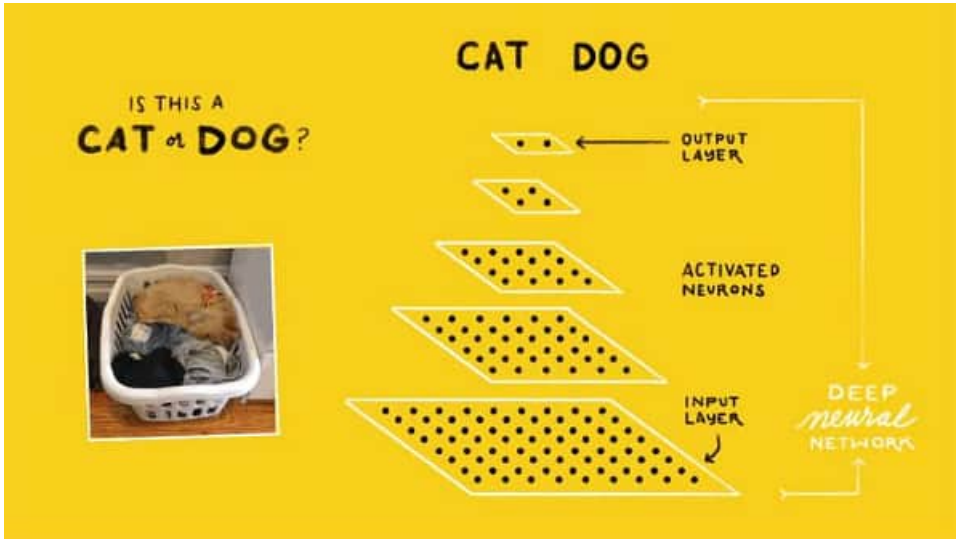
Vorteil

- Benötigt kein domänenspezifisches Expertenwissen (z.B. Bildverarbeitung)
- Sehr hohe Genauigkeit

Nachteil

- Benötigt sehr viele Daten zum Trainieren

Image



predict

Cat / Dog?

DEEP LEARNING BRAUCHT DATEN

AUSTRIAN INSTITUTE OF TECHNOLOGY

Lernen anhand Beispiele

- Deep Learning benötigt viele diverse Daten um die relevanten Merkmale eines Objekts zu erkennen.
  - Unterschiedliche Hunderassen, Automarken, Vogelarten, Kleidung, etc.

Overfitting

- „Sieht“ ein Modell nur rote Autos, geht es davon aus, dass Autos rot sind.
- Sind die Trainingsdaten zu wenig divers, kann es nicht auf „ungesehene“ Daten generalisieren („es erkennt keine blauen Autos“)

Probleme

- Bias: Annotator‘s Bias, Vorurteile, Stereotype, etc. bilden sich im Annotierungs- / Datensampling-prozess ab
- Persönliche Daten: Identifizierende Informationen in Text, Bild, Ton
  - ➔ Daten Anonymisieren, Synthetische Daten
- Ethische Implikationen: Welche Daten dürfen verwendet werden (z.B. Hautfarbe, Religion, sexuelle Orientierung, Konsumverhalten, etc.)
  - ➔ Ethik Richtlinien, gesetzliche Richtlinien

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

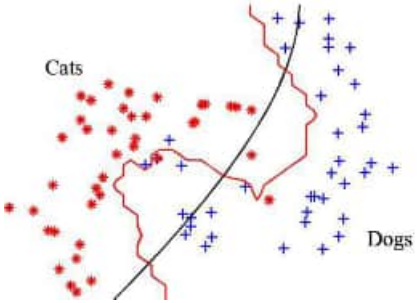
IN DER VERWALTUNG GIBT ES SO VIELE DATEN

Kann man die nicht einfach nutzen?

WIE SYSTEME LERNEN

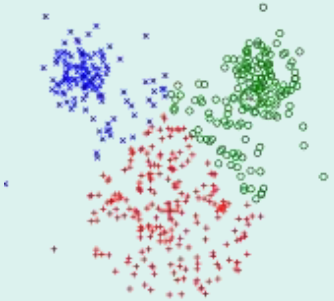
Supervised

- Annotierte Daten (Labelled Data)
  - Z.B. Kategorien (Rechnung, Vertrag, Korrespondenz)
- KI lernt von Assoziationen
  - Input ⇔ Output
- **Nachteil**
  - Daten Annotieren ist sehr aufwändig / teuer
- Beispiele
  - Klassifikation (z.B. Dokumentenkategorien)
  - Regression: Wert vorhersagen / Trendanalyse



Unsupervised

- Unlabelled Data
- KI lernt durrch inherente Strukuren
- **Nachteil**
  - Es können keine direkten Vorhersagen getroffen werden
- Beispiele
  - Clustering: Gruppen ähnlicher Datenpunkte / Dokumente
  - Data Mining: Mustererkennung



ARTEN VON DATEN

Strukturierte Daten

- Z.B. Tabellarische Daten
- Semistrukturierte Daten: XML, JSON

	A	B	C	D
1				
2	**Datum**	**Von**	**Bis**	**Dauer**
3	23.12.2020	23.12.2020 06:44	23.12.2020 13:48	7:04:00
4	13.11.2021	13.11.2021 09:23	13.11.2021 12:03	2:40:00
5	13.11.2021	13.11.2021 14:54	13.11.2021 20:58	6:04:00
6	15.11.2021	15.11.2021 08:12	15.11.2021 18:10	9:58:00
7	16.11.2021	16.11.2021 08:55	16.11.2021 19:24	10:29:00
8	17.11.2021	17.11.2021 07:05	17.11.2021 16:59	9:54:00
9	18.11.2021	18.11.2021 07:55	18.11.2021 17:16	9:21:00

Unstrukturierte Daten

- Z.B. Emails, Textdatein, Social Media, etc.

https://workspacetips.io/tips/gmail/send-an-email-as-an-attachment-in-gmail/

Heterogene Daten

- Daten aus verschiedenen Quellen und Formaten
- Z.B. Word Dokumente, Excel Tabellen, Web Seiten

Multi-Modale Daten

- Daten aus verschiedenen Medien-Formaten
  - Text
  - Bild / Video
  - Ton
  - Sensor

Rangliste nach Schwierigkeitsgrad (einfach → schwer)

1. Strukturierte Daten
2. Unstrukturierte Daten
3. Heterogene Daten
4. Multi-Modale Daten

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

WAS GEHT GUT UND WAS GEHT NICHT?

Fortschritte durch KI

WAS IST KI NICHT

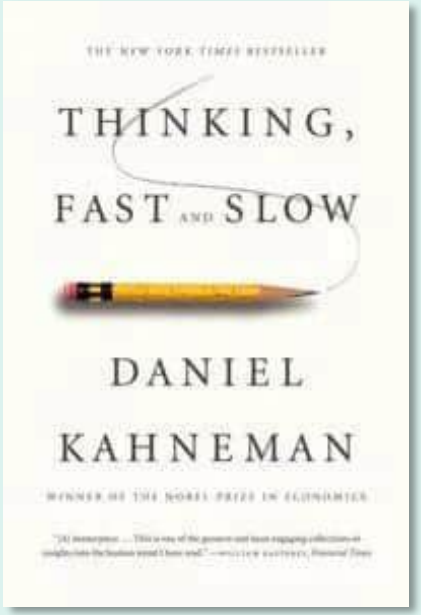
Ersatz für Menschliche Intelligenz

- Keine echte Intelligenz
- KI sind (größtenteils) statistische Lernverfahren (Maschinelles Lernen)
  - ➔ Vorhersagen / Entscheidungen sind Wahrscheinlichkeiten
  - Wie verlässlich sind diese

- Kahneman:
  - „System 1“ – Bauchentscheidung, schnell, automatisch
  - „System 2“ – Ab/herleiten, Schließen, langsam

- Lösungs-/Aufgaben-orientiert
  - Kann eine bestimmte Aufgabe sehr gut lösen
  - Aber, kein Allgemeinwissen

    ➔ *General Artificial Intelligence* gibt es nicht



GROßE FORTSCHRITTE IN DER KI

Textverarbeitung

- Textverständnis
- Textinterpretation
  - Sentiment, Hate Speech
- Automatische Übersetzung
- Suche, Sortieren
- Text Generieren
  - Zusammenfassung
  - Antworten
- Query Answering
  - Fragen im Klartext stellen

Bildverarbeitung

- Komplexe Objekterkennung / Tracking
- Segmentierung
  - Z.B. Fahrbahn, Gehweg, Autos
- Bild Generierung
  - Deep Fakes, Dall-e

Audio-Analyse

- Spracherkennung
  - Multi-Lingual, Dialekte
- Automatische Transkription
- Audio Events erkennen
- Audio Szenen erkennen
- Musik Interpretieren

Vielen Dank

für Ihre Aufmerksamkeit.

ALEXANDER SCHINDLER
Thematic Coordinator Researchfield Datascience
Data Science & Artificial Intelligence
Center for Digital Safety & Security

AIT Austrian Institute of Technology GmbH
Giefinggasse 4 | 1210 Vienna | Austria
T +43 50550-2902 | M +43 664 8251454
alexander.schindler@ait.ac.at | www.ait.ac.at

MOSTLY·AI

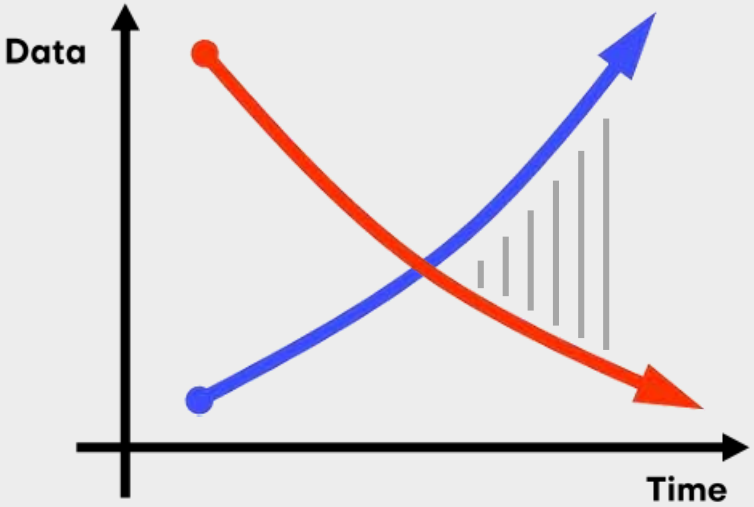
Synthetic Data for Privacy, AI Fairness and Democratizing Data Access

Alexandra Ebert
Chief Trust Officer | MOSTLY AI
Host of the Data Democratization Podcast
Chair of the IEEE Synthetic Data IC Expert Group

> *[synthetic data] can become the unifying bridge between policy support and computational models, by unlocking the potential of data hidden in silos; thus becoming the key enabler of artificial intelligence in business and policy applications in Europe.*
>
> **-European Commission's JRC, 2022**

AI needs synthetic data! But why?

Reason #1: Today's data fuel crisis

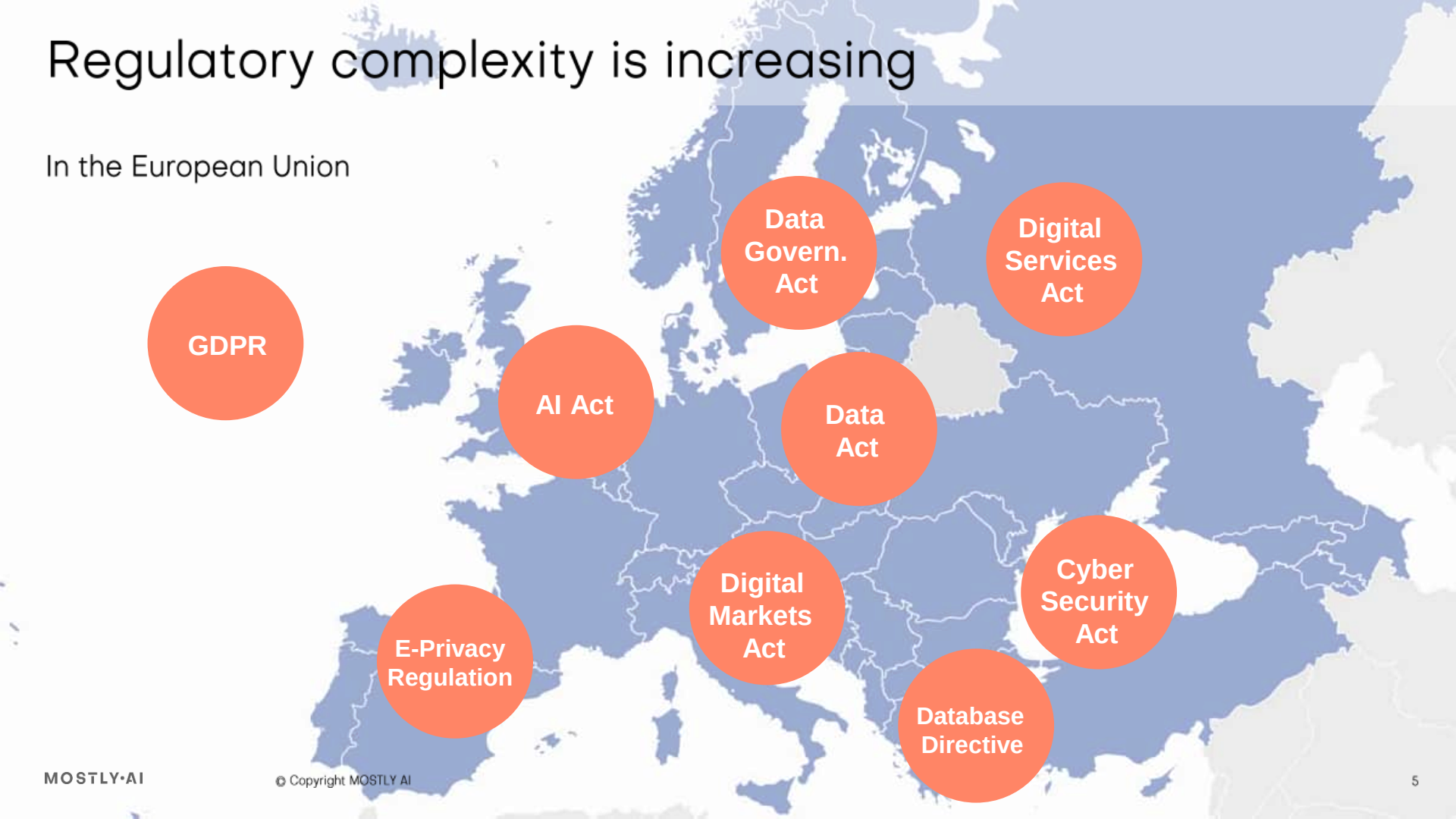


AI, analytics & automation drive demand for data

Regulation, risk & ethics restrict availability of data

Regulatory complexity is increasing

In the European Union



Regulatory complexity is increasing

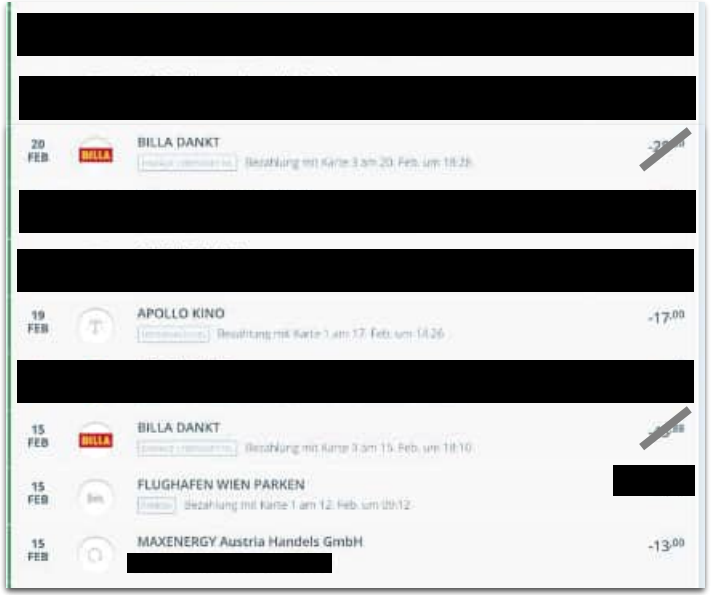
In the European Union – and globally

"By 2023, 65% of the world's population will have its personal information covered under modern privacy regulations" (Gartner)

Reason #2: Legacy anonymization fails for big data

You can either have useful or private data. But not both!

User #3dcf29717a9f9b39



20-30€

40-50€

User #71f7c3014d2ced27

Legacy anonymization destroys data utility

You can either have useful or private data. But not both!

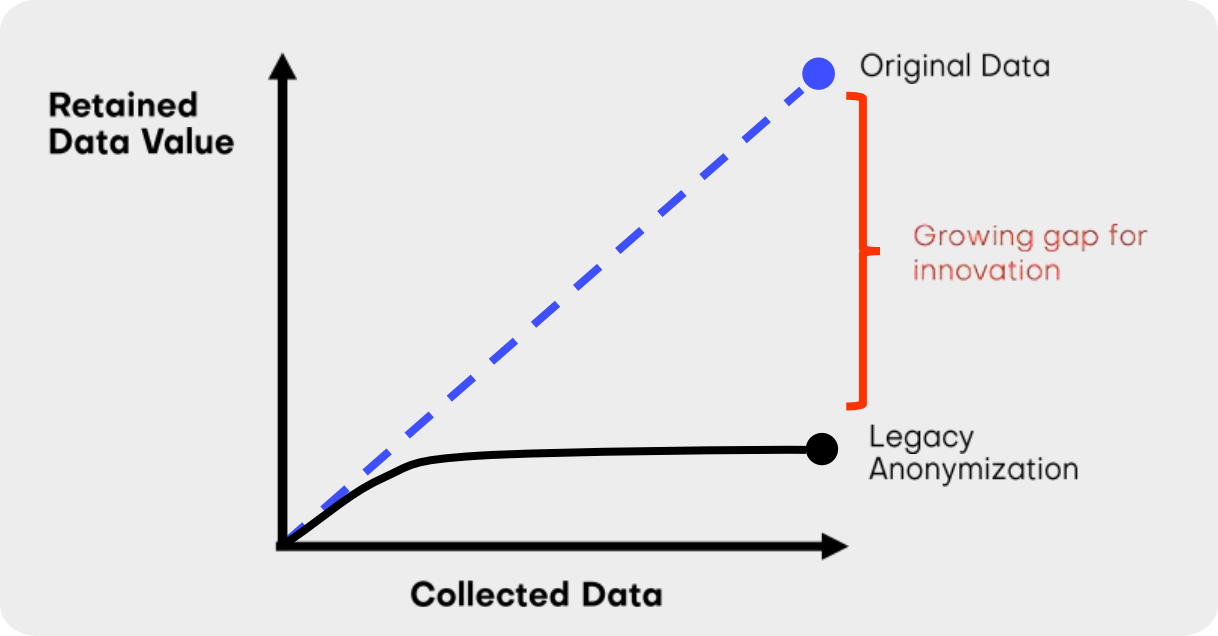
User #3dcf29717a9f9b39

User #71f7c3014d2ced27

→ **The 1:1 relationship between supposedly "anonymous" & real data allows re-identification**

Legacy anonymization is easy to re-identify

Legacy anonymization fails for big data



What is AI-generated synthetic data, actually?

What is AI-generated synthetic data?

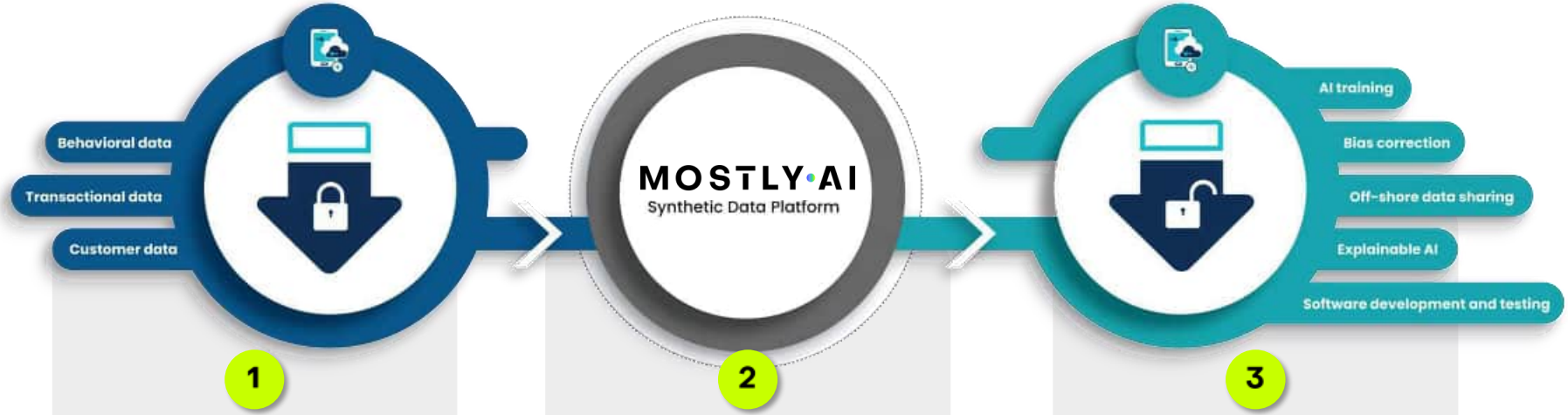
AI-generated synthetic data is an **anonymization technology.**

It's **highly realistic & statistically representative**, yet **fully anonymous** artificial data.

Synthetic data helps you to **unlock data** while **protecting privacy**.

How is AI-generated synthetic data created?

The process for structured synthetic data generation is easy, fully automated and industry-agnostic.



1

Original transaction data

- Highly valuable, but
- Privacy-sensitive
- Locked away (GDPR/CCPA)

2

Synthetization

- Fully automated with privacy mechanisms
- Unsupervised ML process – AI learns patterns of input data
- Results in model containing all statistical info of input data

3

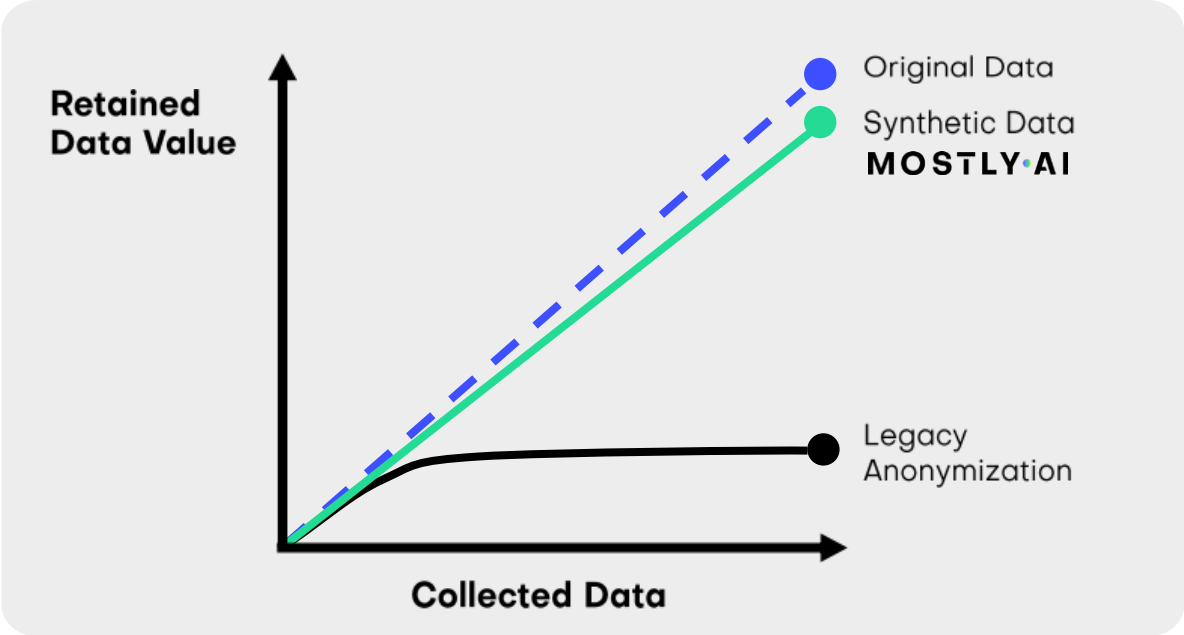
Synthetic transaction data

- As-good-as-real
- Fully anonymous
- Free to use (GDPR/CCPA compliant)

Synthetic data – the solution for privacy & data utility

- Synthetic data is highly realistic & statistically representative
- Yet, there are no 1:1 relations and
- No re-identification risk

→ It protects privacy while preserving granular statistics



What is the business value of synthetic data?

1

Eliminate Privacy Risk
by minimizing the need to touch actual customer data because of irreversible AI anonymization

2

Be Faster
by reducing time-to-data and time-to-market of your data projects

3

Be More Accurate
by working with granular synthetic data that retains structure, correlations and time-dependencies perfectly

4

Collaborate more
by sharing your data freely and safely within and across organizations

5

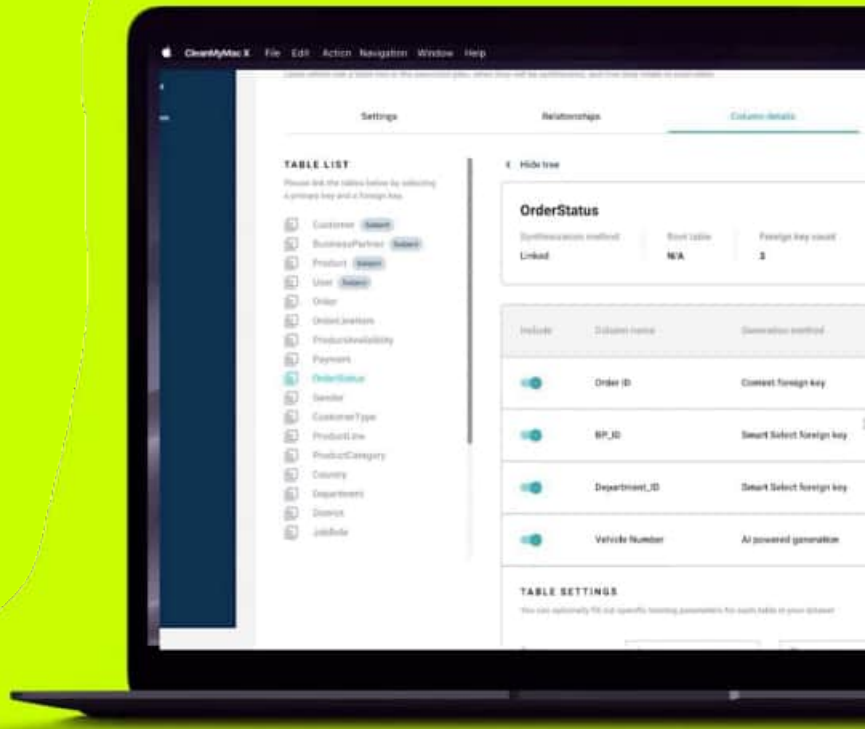
Automate & Save Costs
by streamlining legal, compliance and risk management procedures

6

Improve AI Performance by Upsampling Rare Events
by generating more balanced synthetic data (e.g., with more fraud cases) to boost accuracy of downstream AI/ML models

The top synthetic data use cases

- **AI training** & advanced analytics
- **3rd party data-sharing** & collaboration
- Digital **product development**
- Software **testing**
- **Cloud migration**
- **Open data** sharing & democratizing data access
- **AI bias** detection & mitigation
- **Responsible AI governance** & algorithmic auditing

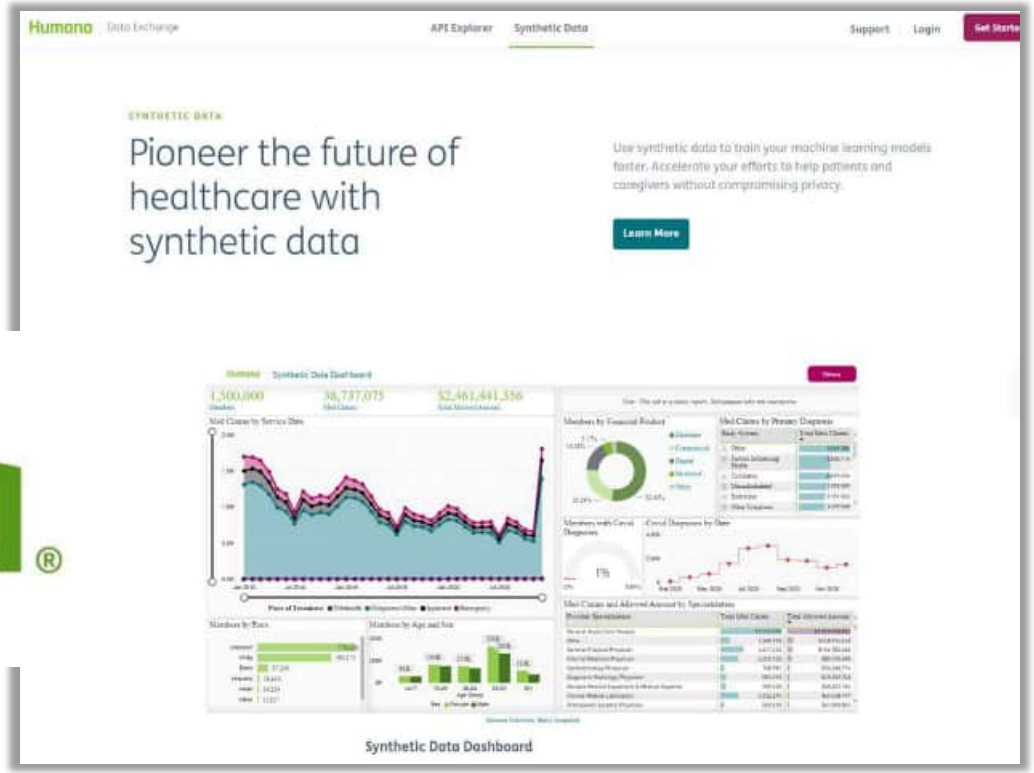


citi SWIFT Telefónica ERSTE Group Humana City of Vienna

Democratizing synthetic data = democratizing innovation

Humana's open synthetic healthcare data

Synthetic healthcare data containing 1.5Mio customer records – openly available for vendors, researchers, and partners



https://developers.humana.com/syntheticdata

“

On our way to be the digitalization capital, we actively shape the digital transformation. Through cooperation with companies such as MOSTLY AI, we take an important step to enable data-driven innovation by providing even more valuable Open Data while ensuring full anonymization of personal information through data synthetization.

Brigitte Lutz

Data Governance Coordinator

City of Vienna

Synthetic healthcare data in Germany

07.06.2022

InGef vergibt Unterauftrag im Projekt KI-FDZ zur Synthetisierung von Gesundheitsdaten an MOSTLY AI

Hier geht es zur Pressemitteilung.

„KI-FDZ soll die sichere Nutzung von Gesundheitsdaten zu Forschungszwecken in Deutschland nachhaltig verbessern […] und dabei gleichzeitig die Nutzbarkeit der Daten wahren."

KI-FDZ

Anonymisierte und synthetisierte Daten zur Verbesserung der Gesundheitsversorgung in Deutschland

Berlin, 07. Juni 2022 – Das InGef - Institut für angewandte Gesundheitsforschung Berlin (InGef) plant, im Rahmen des Forschungsprojektes „Künstliche Intelligenz am Forschungsdatenzentrum – Erforschung von Anonymisierungsmöglichkeiten und AI-readiness (KI-FDZ)", KI-basierte synthetische, identifikationsgeschützte Gesundheitsdaten zu generieren. Die Förderung des Projektes hat das Bundesministerium für Gesundheit (BMG) übernommen, die datentechnische Umsetzung erfolgt gemeinsam mit MOSTLY AI, einem Unternehmen spezialisiert auf die Generierung synthetischer Daten. Das Projekt KI-FDZ hat unter anderem das Ziel, Routinedaten der gesetzlichen Krankenversicherung in Deutschland durch geeignete Verfahren zu synthetisieren.

Gefördert durch:
Bundesministerium für Gesundheit
aufgrund eines Beschlusses des Deutschen Bundestages

ingef FDZ Gesundheit BIH Berlin Institute of Health @Charité Fraunhofer MEVIS MOSTLY·AI

What if...all researchers had open access to a synthetic version of the EU cancer dataset?

- JRC trained the MOSTLY AI synthetic data generator on 500k data subjects out of **50 million EU cancer patients**. " **The results,** available in the accompanying archive both as a PDF report and CSV files with the data correlations, **are impressive**."
- "Synthetic data have **proven great potential** and are the go-to methods **ready to be deployed in real-life scenarios**. Policy applications can now be researched and developed with little risk involved."
- "More important than focusing on how to synthesize data is **what can we achieve with the new data available at scale**, how to **convince data owners to unleash their coveted data** to the broadest audience, and how to accommodate this **massive new ability into the policy formulation and assessment**."

Why is synthetic data essential for Responsible AI?

How does it help with AI Fairness?

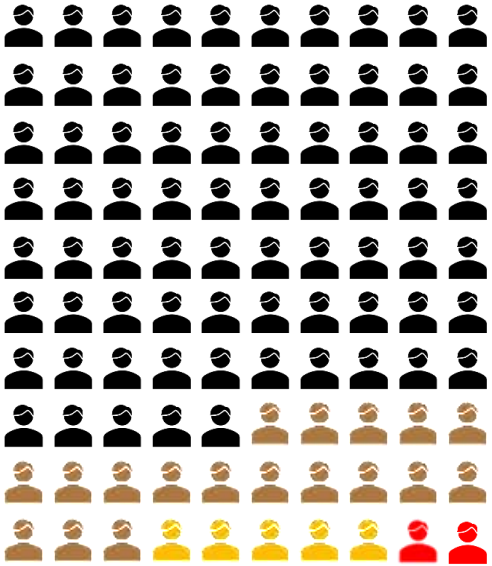
AI fairness and why bias in AI is a problem

10 (out of many) reasons for bias in AI

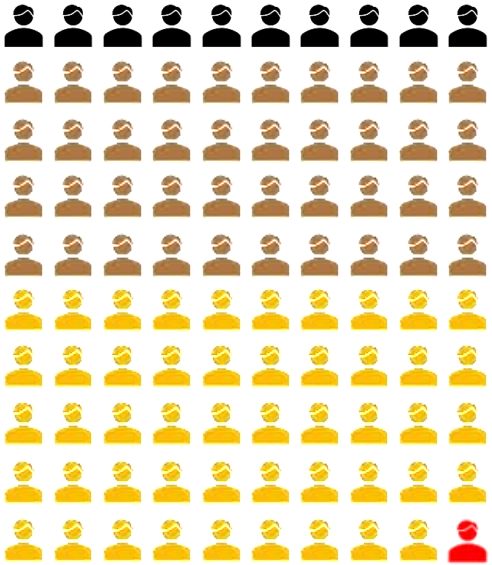
1. Insufficient training data (e.g. facial recognition systems, UK health app,…)
2. Humans are biased – and so is the data AI is trained on (e.g. Amazon's HR algorithm)
3. De-biasing data is exceptionally hard to do (removing sensitive attributes is not a good option!)
4. De-biasing AI models is very difficult too
5. Diversity amongst AI professionals is not as high as it should be
6. Fairness comes at a cost (that companies may not be willing to pay)
7. External AI audits could help – if privacy would not be an issue
8. Fairness is hard to define
9. What was fair yesterday could be biased tomorrow (e.g. Microsoft's Tay)
10. The vicious bias cycle: biased AI will lead to more bias in data

Case study: hidden biases in facial detection

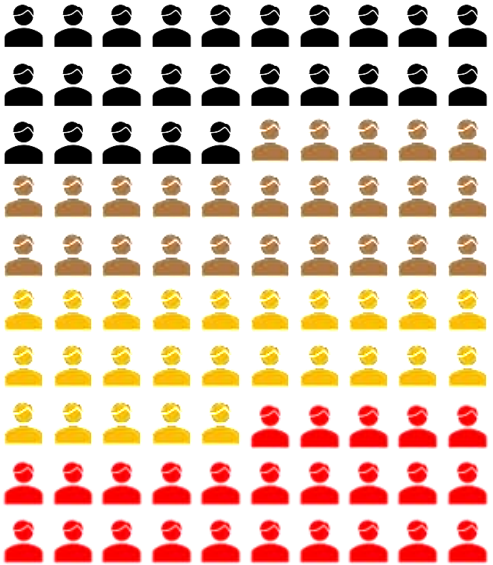
Real world



“Gold-Standard” Dataset



Balanced Dataset



Black hair Brown hair Blonde hair Red hair

How can Fair Synthetic Data help?

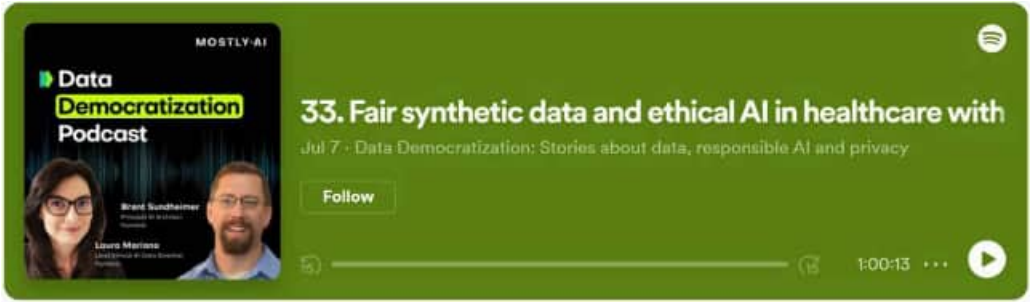
Illustrating the power of fair synthetic data with NVIDIA's StyleGAN

Synthetic data for AI fairness & bias mitigation

Create fair synthetic data that reflects the world not as is but as society would like it to be

- → Detect biases
- → Mitigate biases
- → Access to sensitive attributes
- → Access to data from different geographic regions
- → Overrepresent minorities
- → Share it with AI bias auditors
- → ...

Recommendation: Check out our Fairness in AI blog series



Recommendation: Fair synthetic data podcast episode with Humana

What could you achieve with making **synthetic data** available **at scale**?

Alexandra Ebert
Chief Trust Officer | MOSTLY AI
alexandra.ebert@mostly.ai

Predictive Analytics in der Finanzverwaltung

Wien, September 2022

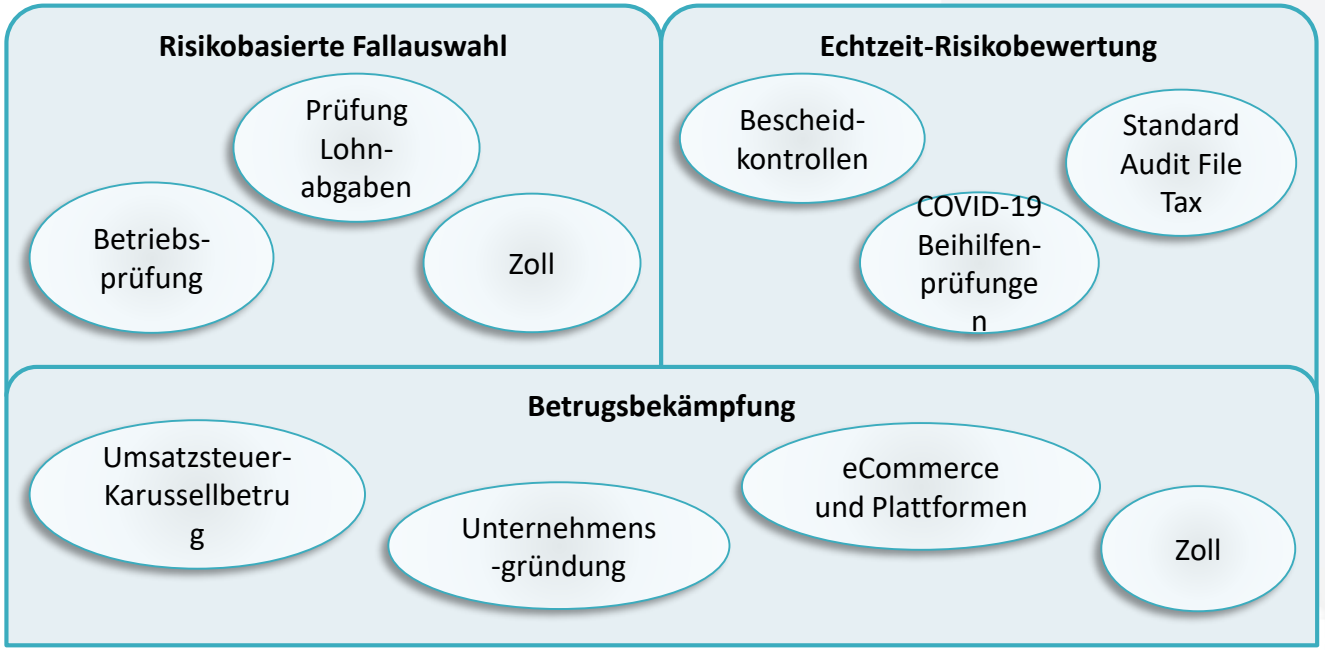
Risikobewertung und Optimierung mit Machine Learning

Geschäftsfälle identifizieren, die von einem Mitarbeiter überprüft werden sollten, weil ähnliche Fälle in der Vergangenheit mit einem hohen Risiko verbunden waren.

Systematische Auswahl von Geschäftsfällen mit hohen Risiko führt zur optimalen Allokation von Prüfungskapazitäten.

Voraussetzung: Historische Daten zu Geschäftsfällen und Risiken

Projekte des Predictive Analytics Competence Center



Bescheidkontrolle in der Arbeitnehmerveranlagung

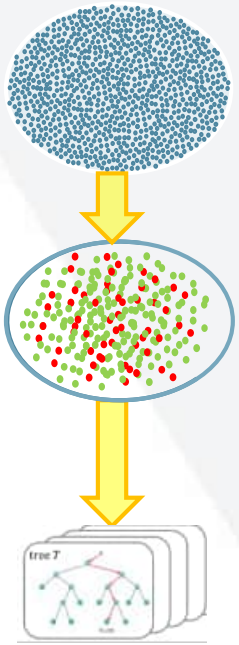
Geschäftsfall: Einkommenssteuererklärung

Non-Compliance Risiko: Bereiche der Erklärung nicht korrekt erklärt (z.B. Werbungskosten, Pendler, Kinder)

Historische Daten: Erklärung vs. Korrektur durch Mitarbeiter

Werbungskosten

		Erklärt	Korrektur
Berufsbezeichnung		Linienpilot	
Pendlerpauschale	718	3.672,00	1.224,00
Pendlereuro	916	869,00	54,67
Gewerkschaftsbeiträge	717	396,00	
Arbeitsmittel	719	433,15	411,54
Reisekosten	721	1.587,34	L
Doppelte Haushaltsführung	723	269,00	
Sonstige Werbungskosten	724	607,92	250,92



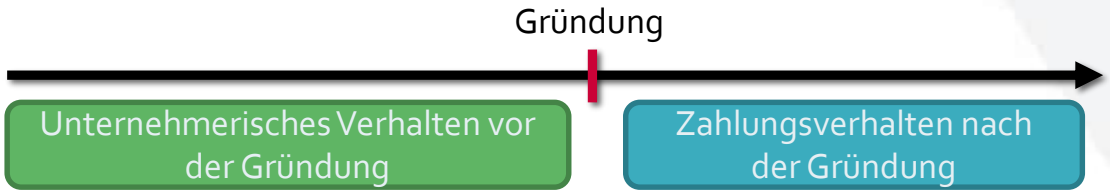
Machine Learning: Zusammenhang zwischen Risikoindikatoren (z.B. Erklärungskennzahlen, Vorjahreswerte, Differenzen, Verhältnisse) und Korrekturen

Unternehmensgründung

Geschäftsfall: Neugründung

Non-Compliance Risiko: Abgaben werden nicht bezahlt werden

Historische Daten:



Machine Learning: Zusammenhang zwischen Verhalten vor der Gründung (Schätzungen, Rückstände, Insolvenzen) und Risiko für Non-Compliance

Umsatzsteuerkarussellbetrug

Geschäftsfall: Unternehmen

Risiko: Missing Trader (Umsatzsteuer nicht bezahlt)

Historische Daten:

- keine ausreichende Zahl von Missing Trader
- Innergemeinschaftliche Erwerbe und Umsatzsteuervoranmeldung

Machine Learning: Identifikation von Ausreißer / außergewöhnlichen Verhalten indem aus dem historischen Daten die normalen Zusammenhänge definiert

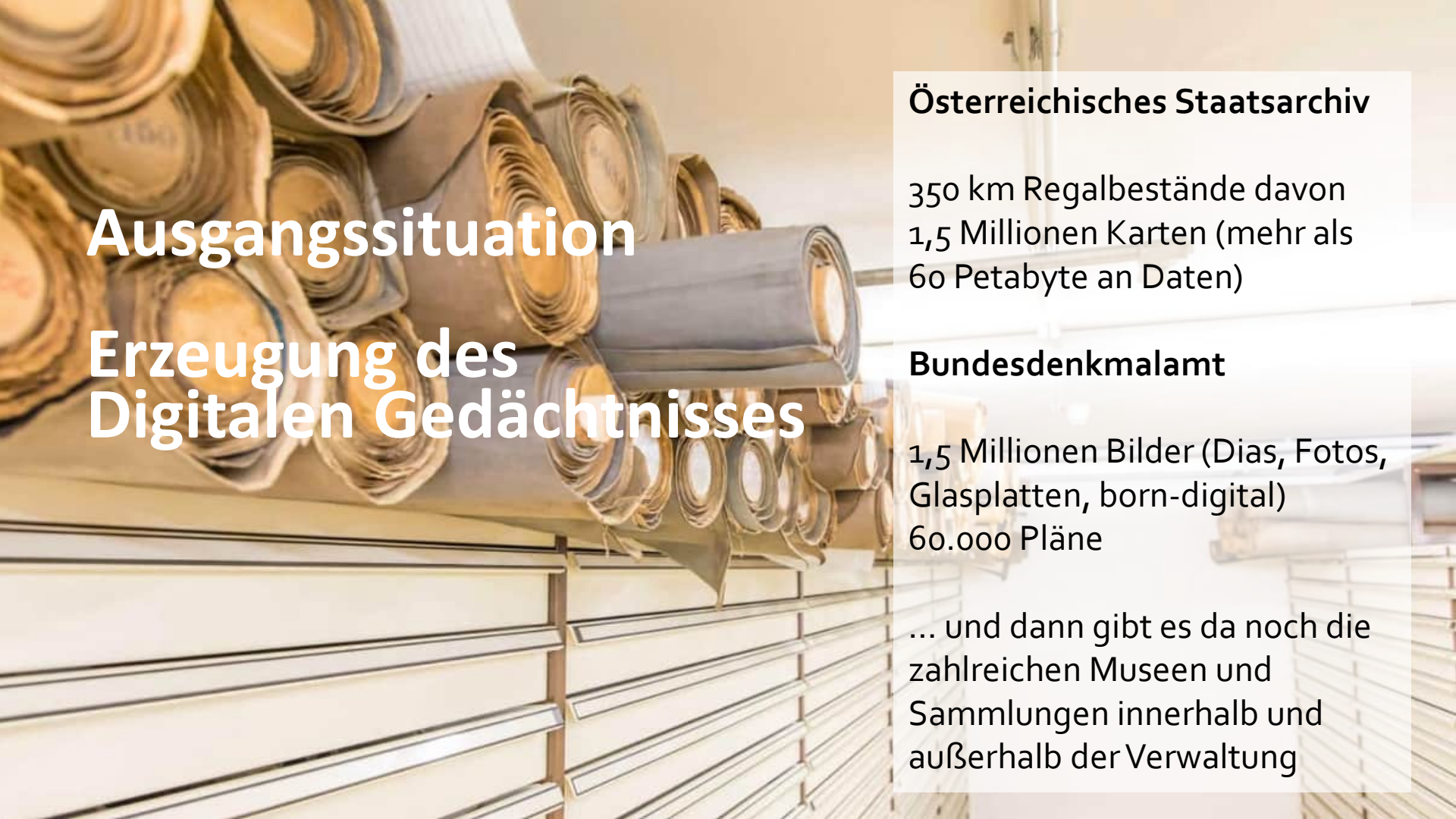
Erfolgsfaktoren für Predictive Analytics Projekte

- Historische Datenverfügbarkeit und Datenqualität
- Fachliche Initiative und Projektleitung
- Starke Unterstützung auf Management Ebene
- Predictive Analytics ist effektiver bei der Optimierung von Entscheidungen je
  - eindeutiger die Auswahlmöglichkeiten der Entscheidung
  - häufiger die Entscheidung getroffen wird
  - mehr Personen die Entscheidung treffen
  - komplexer die Daten für die Entscheidung sind
  - mehr Daten für die Entscheidung zur Verfügung stehen
  - zeitaufwändiger die Entscheidung
  - wichtiger Erfahrung für die Entscheidungsfindung
  - höher der Wert / Konsequenz der Entscheidung

Daten-Plattformen am Beispiel des Digitalen Gedächtnisses Österreichs

Manfred Gruber
Bundeskanzleramt

Wien, 19. Oktober 2022



Ausgangssituation

Erzeugung des Digitalen Gedächtnisses

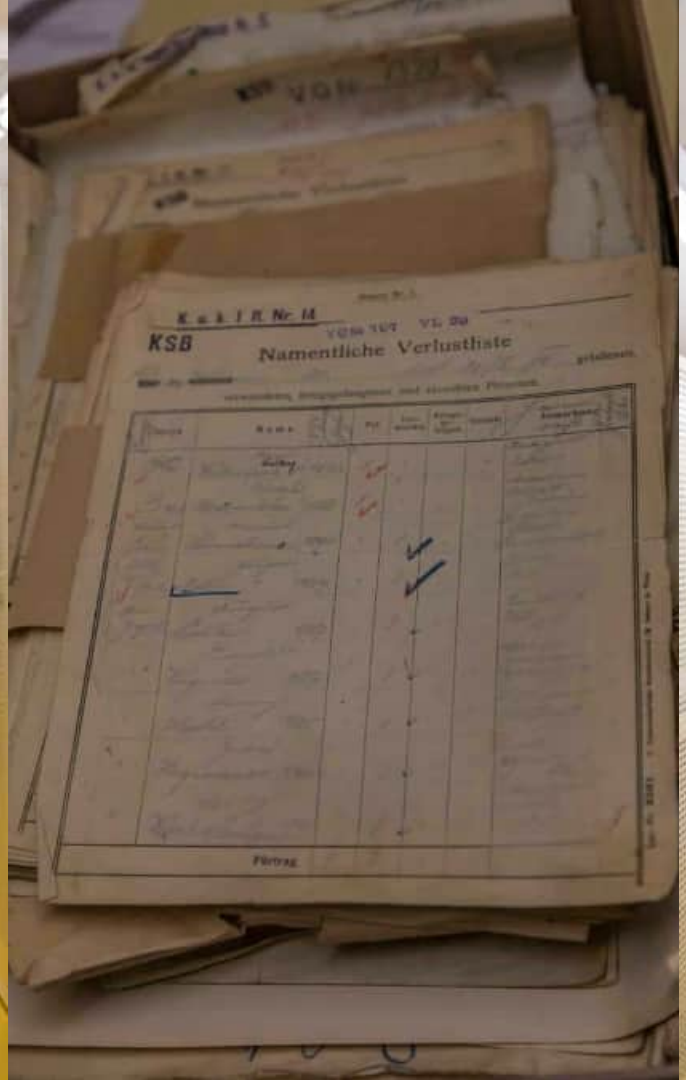
Österreichisches Staatsarchiv

350 km Regalbestände davon 1,5 Millionen Karten (mehr als 60 Petabyte an Daten)

Bundesdenkmalamt

1,5 Millionen Bilder (Dias, Fotos, Glasplatten, born-digital)
60.000 Pläne

… und dann gibt es da noch die zahlreichen Museen und Sammlungen innerhalb und außerhalb der Verwaltung


KSB
Namentliche Verlustliste


England.
EB
Liste Nr. 144
Exh. Nr.
176409 R.
Name: Krizsan László
Alter oder Geburtsjahr 1899
Heimatsgem.
Heimatsbez.: Nagybánya, Szatmármegye
Angehörige: Krizsan Péter Nelson u. 19
Charge: Inf.
Grundb.-Nr.:
A.-J.:
Truppenkörper: Hir. 72
Unterabt.: 2. Bun. 5. K
am:
Gefangen in:
Gesundheitszustand: Durch eine Bombe verschüttet
Interniert in: Casualty Clearing Station France
Russland


161
162
164


STADTSCHLAITHING


446

Bis 2025 entstehen alleine durch die momentan laufenden Projekte rund 2,1 Petabyte an Daten!

… und sind somit eine Grundlage für neue Projekte.

Derzeit laufende Projekte

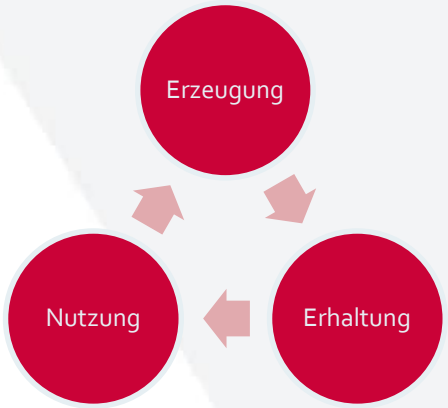
- BDA: 150.000 Dias gescannt und aufbereitet
- BDA: 60.000 Pläne scannen und aufbereiten
- BDA: 1,5 Mio. Bilder – Born digital
- BDA: Kleiner Bestand Glasplatten
- ÖStA: 20.000 Kriegskarten
- ÖStA: 300.000.000 Laufmeter Mikrofilme
- ÖStA: 70.000 Karteikarten (Ansiedlerkarteien)
- ÖStA: Findmittel, Indexbücher
- ÖStA: Sammlung Beer (1880-1914) - Glasplatten
- BHÖ: 2000 Gipsskulpturen abgeschlossen
- … viele Projekte in der Queue bzw. in Planung

Herausforderungen

- Fehlendes technisches Know-how bzw. fehlende Ressourcen in den Institutionen
- Nicht vorhandene Service-Ebenen (jeder muss sich die Werkzeuge selbst bauen)
- Wert der Daten wird nicht erkannt (Grundlage AI, Vernetzung der Daten, neue Methoden, Verbindung mit anderen Domänen) und sie werden nicht geteilt
- Teilweise umfangreiche und komplexe Prozesse um brauchbare Digitalisate zu erzeugen (Metadatenanreicherung, 3D-Scans, …)
- Immer größere Datenbestände (oft im hohen Terabyte-Bereich) -
- Vermischung von Archivdaten mit operativen Daten
- Immer mehr born-digital Bestände (→ Schwarzes Loch der Fachanwendungen)
- Rahmenbedingungen (Schutz, Rechtliche Grundlagen, …)

Datenmanagementportal BKA

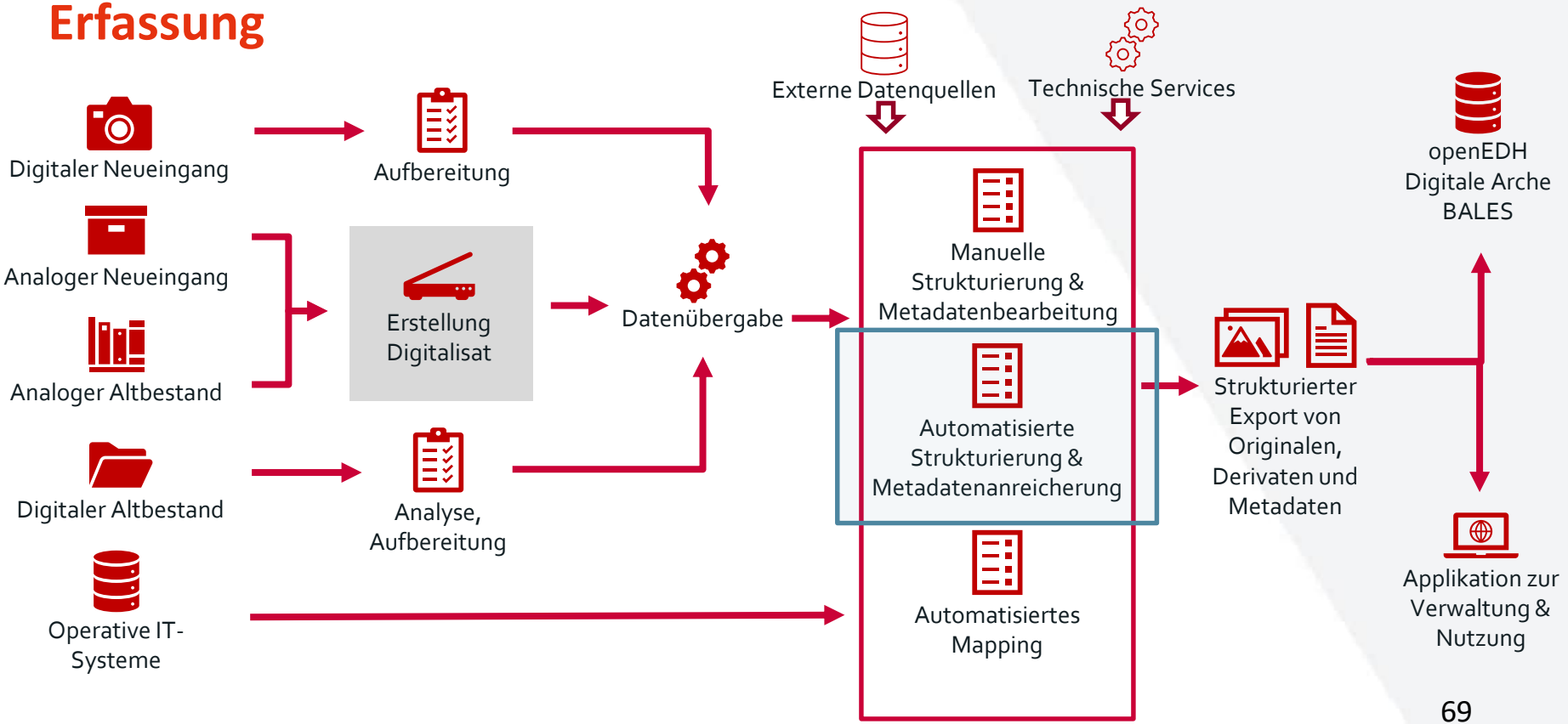
Erzeugung	(Er-)Haltung	Nutzung
• Analog zu Digital • Digital zu Digital • Automatische Prozesse (Silent Archiving) • AI Unterstützung • Qualitätsstandards • Projektbegleitung	• Datenkatalog • Daten-Management • Enorme Datenmengen • Sichere Verwahrung • Langzeiterhaltung	• Vermittlung • Forschung • Grundlage für Verfahren • Grundlage für AI • Prozessverbesserung
• Goobi Workflow • KI Services	• open Enterprise Data Hub (openEDH) • Bundesarchiv- und Langzeiterhaltungs-system (BALES) • Digitale Arche	• Goobi Viewer • Forschungsplattform • BALES



Grundprinzipien

- Daten-Souveränität als oberstes Prinzip
- Einsatz und Erzeugung von Open Source zur Steigerung der Nachhaltigkeit
- Verbinden bestehender Lösungen (jede Software macht das, was sie besonders gut kann)
- Einfache und selbsterklärende Datenstrukturen
- Langzeiterhaltung und digitale Archivierung von Anfang an mitgedacht
- Nachvollziehbarkeit
- Motivation zu (Metadaten)-Standards, FAIR und Open (Linked) Data

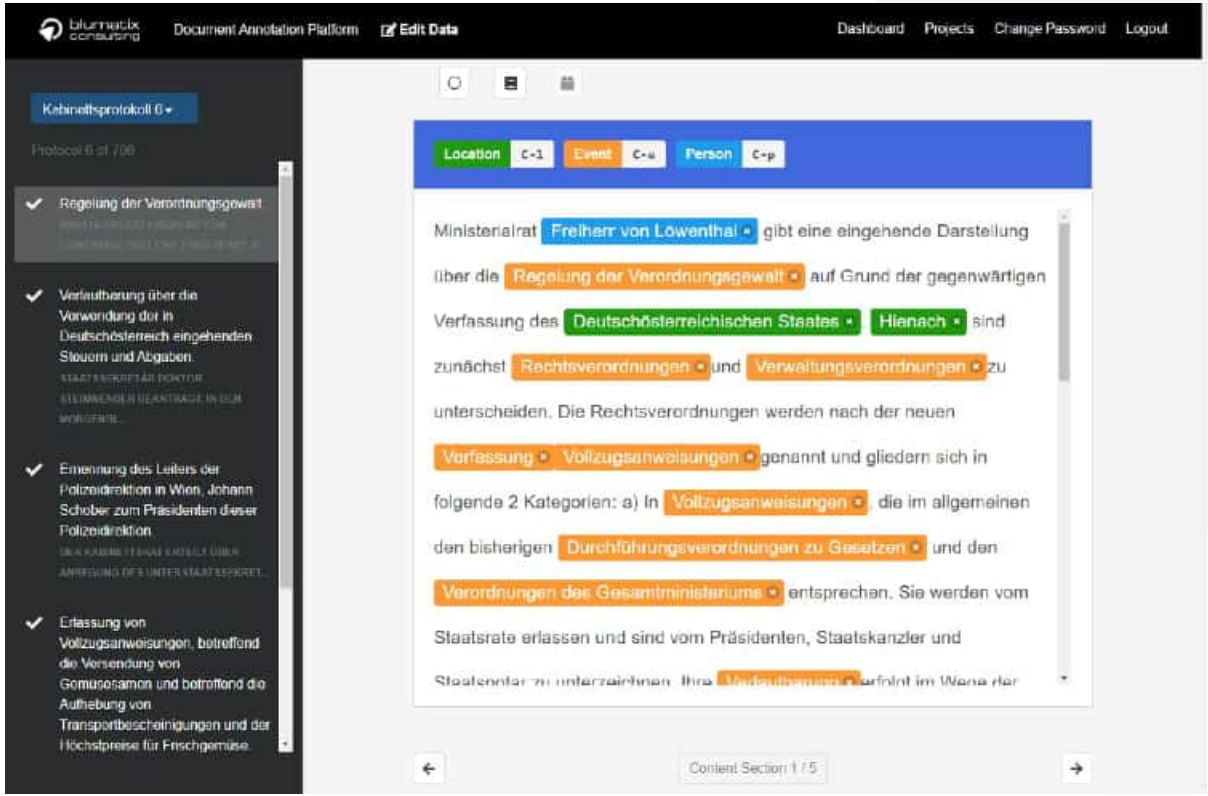
Erfassung


Digitaler Neueingang
Analoger Neueingang
Analoger Altbestand
Digitaler Altbestand
Operative IT-Systeme
Aufbereitung
Erstellung Digitalisat
Analyse, Aufbereitung
Datenübergabe
Externe Datenquellen
Technische Services
Manuelle Strukturierung & Metadatenbearbeitung
Automatisierte Strukturierung & Metadatenanreicherung
Automatisiertes Mapping
Strukturierter Export von Originalen, Derivaten und Metadaten
openEDH
Digitale Arche
BALES
Applikation zur Verwaltung & Nutzung

Beispiel Bildklassifikation MAK (POC)

Kategorie	Wahrscheinlichkeit
Ornamentik	71.6%
Natur	67.5%
Blume	95.1%
Pflanze	89.2%
Druckgrafik	100%

Beispiel Textannotation (POC)

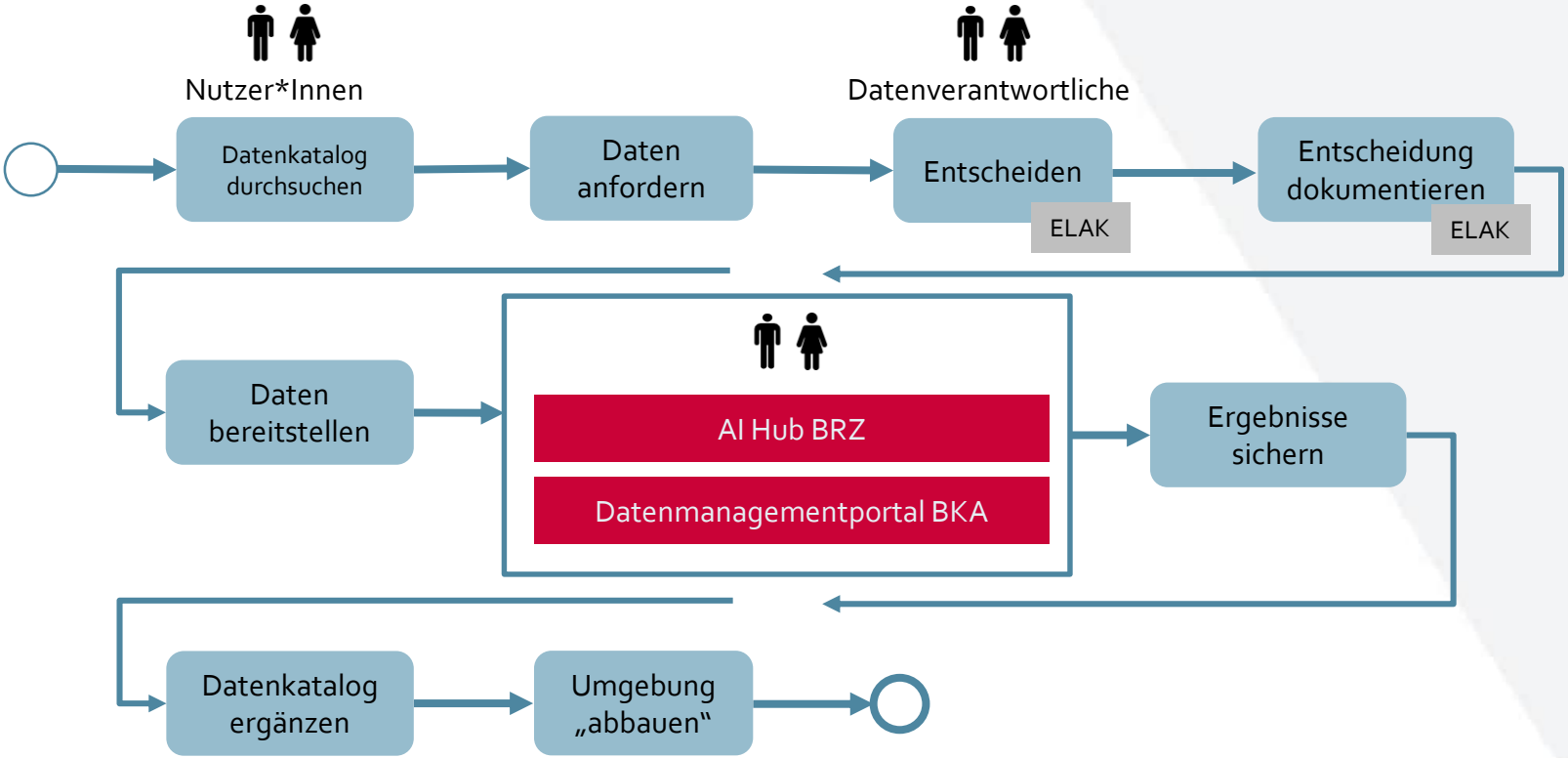


Beispiel Gesichtsdetektion

KI Service zur Gesichtsdetektion

Im Rahmen eines Projektes im Bundesdenkmalamt mussten 150.000 Bilder auf DSGVO-relevante Inhalte durchsucht und anschließend beurteilt werden.
Das in den Prozess eingebettete Service markiert automatisch gefundene Gesichter und schreibt einen Vermerk in die Metadaten.

Nutzen der Daten



Danke für Ihre Aufmerksamkeit!

Manfred Gruber
Bundeskanzleramt
+43 664 811 70 52
manfred.gruber@bka.gv.at