



KI-Systeme schützen durch Privacy-Preserving Machine Learning

Dr. Matthias Dorner

DATEV eG | AI Office

22.10.2024

it-sa Expo & Congress 2024



- **DATEV eG als Genossenschaft** mit Hauptsitz in Nürnberg
- **Führendes Softwarehaus für Steuerberater, Wirtschaftsprüfer, Rechtsanwälte und deren Mandanten**
 - Jahresumsatz 2023: 1,44 Mrd. EUR
 - Genossenschaftsmitglieder: >40,000
 - Kunden: >700,000
 - Beschäftigte: 9,000, davon >2,000 in FuE
 - Produkte: insb. Steuern, Finanz-/Lohnbuchhaltung und IT-Services
- **Höchste Anforderungen an Unternehmens-/Informationssicherheit und Datenschutz**
- Entwicklung und Einsatz von **Softwarelösungen mit künstlicher Intelligenz (KI)** und **Big Data**

Herausforderung sichere KI-Systeme

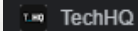


Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

17.03.2018




Business illustrations by Storyset



China data breach hosted on Alibaba with 1 billion records+

The exposure of a billion of China's citizens' data is the biggest breach to date. The information has been unprotected for 14 months.

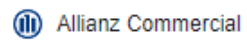
07.07.2022



The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.


27.12.2023



Neue Datenschutzrends treiben Frequenz und Kosten von Cyber-Schäden

Cyber-Schadensfälle haben im vergangenen Jahr weiter zugenommen, was zum großen Teil auf eine Zunahme von Datenschutzverletzungen...

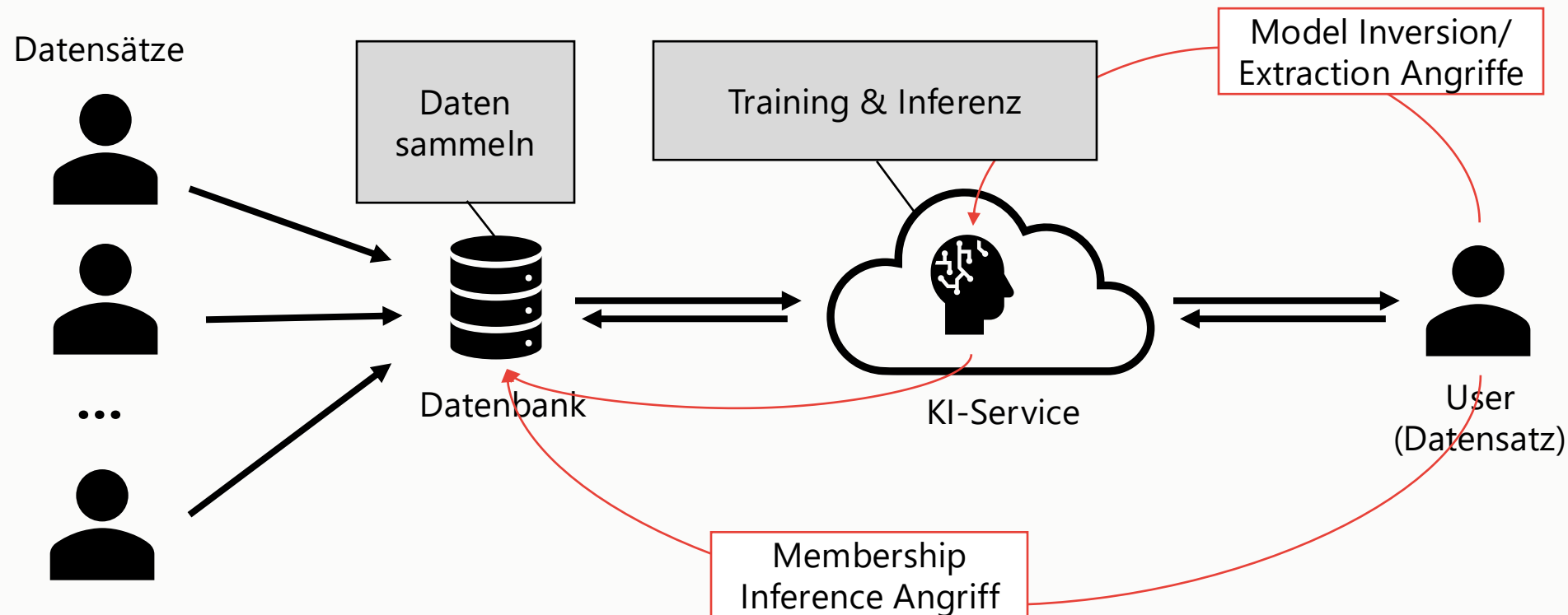
09.10.2024



- (1) Welche Risiken gibt es in KI-Systemen?
- (2) So schützt Privacy Preserving Machine Learning
- (3) DATEV Personal-Benchmark online
- (4) Zusammenfassung

Welche Risiken gibt es in KI-Systemen?

- „Klassische“ IT-Sicherheitsrisiken sowie zusätzlich
 - Leistungsfähige KI-Systeme benötigen große Mengen an oftmals sensiblen (Trainings-)Daten
 - Regulatorische Anforderungen für sichere und vertrauenswürdige KI-Systeme
 - Architektur von KI-Systemen verursachen Anfälligkeit für spezifische Angriffsszenarien



Privacy Preserving Machine Learning (PPML)

Überbegriff zu Technologien und Praktiken, die (sensible) Daten während Training und Inferenz in KI-Systemen schützen

- Erhaltung der Leistungsfähigkeit des gesamten KI-Systems bei Wahrung des Datenschutzes
- Integration und Skalierbarkeit von IT-Infrastrukturen
- Systematisierung für ganzheitliches Privacy Accounting & Management
- Bsp. Homomorphic Encryption, SPMC, Federated Learning, **Differential Privacy**, Data Anonymization

Differential Privacy (DP)

Garantiert, dass eine Datenabfrage keine sensiblen Informationen über einzelne Datensätze einer Datenbank preisgibt

- **Statistisches Rauschen** verschleiert Einfluss einzelner Datenpunkte auf Ergebnis („plausible deniability“)
- Steuerung über **Privatsphäre Budget** (ϵ bzw. ϵ und δ)
 - No free lunch! Privatsphäre kostet Modellgüte
- Keine spezifischen Angreifermodelle nötig. Mathematische Garantie umfasst starke Angreifer, die alle Daten kennen
- Einsatz bei i) Preprocessing (Daten verrauschen), ii) **Training** (PATE, **DP-SDG**), iii) Inferenz (Ergebnisse verrauschen, Modelkompression)

DATEV Personal-Benchmark Online

Beschreibung Use Case



Ein faires Gehalt einfach vermitteln und anbieten

Mit DATEV Personal-Benchmark online ermitteln Steuerberater und Mandant das optimale Gehalt für neue oder bestehende Mitarbeiter im Handumdrehen.

Gehaltsanalyse

Auf Basis von 5 Millionen echten Lohnabrechnungen aus den DATEV-Programmen



Marktwertprognose berechnen ⁱ

Was wäre ein üblicher Marktwert für den (potenziellen) Mitarbeitenden?

Mandant/in (optional)
Firmenname, Beraternummer (BNR), Mandanten-Nr. (Mdt)

Mitarbeiter/in
Nachname, Vorname, Personal-Nr. (PNR)

[?] Weitere Infos zu [Mitarbeitenden](#) im DATEV Hilfe-Center.

Beruf *
Anlagenbuchhalter/in

Ort des Unternehmens *
Nürnberg, Mittelfr

Arbeitsstunden (Woche) *
40

Branche *
C: Verarbeitendes Gewerbe

Anzahl der Mitarbeitenden *
11-25

Berufserfahrung *
15

Ausbildungsabschluss *
Bachelor

Betriebszugehörigkeit *
1-3 Jahre

[?] Weitere Infos zu [Beruf](#), [Stunden](#), [Berufserfahrung](#) und [Branche](#) im DATEV Hilfe-Center.

**1. Mitarbeiterauswahl
oder Benutzereingabe
Stellenbeschreibung**

**2. KI-basierte Marktwert-
prognose für Jahresbrutto
mit Gehaltsbestandteilen
(nicht im Bild)**



Angriff 1 Gehalt eines Nachbarn (Person)

- Membership Inference: Nachbar in Trainingsdaten enthalten
- Attribute Inference: Extraktion des tatsächlichen Gehalts

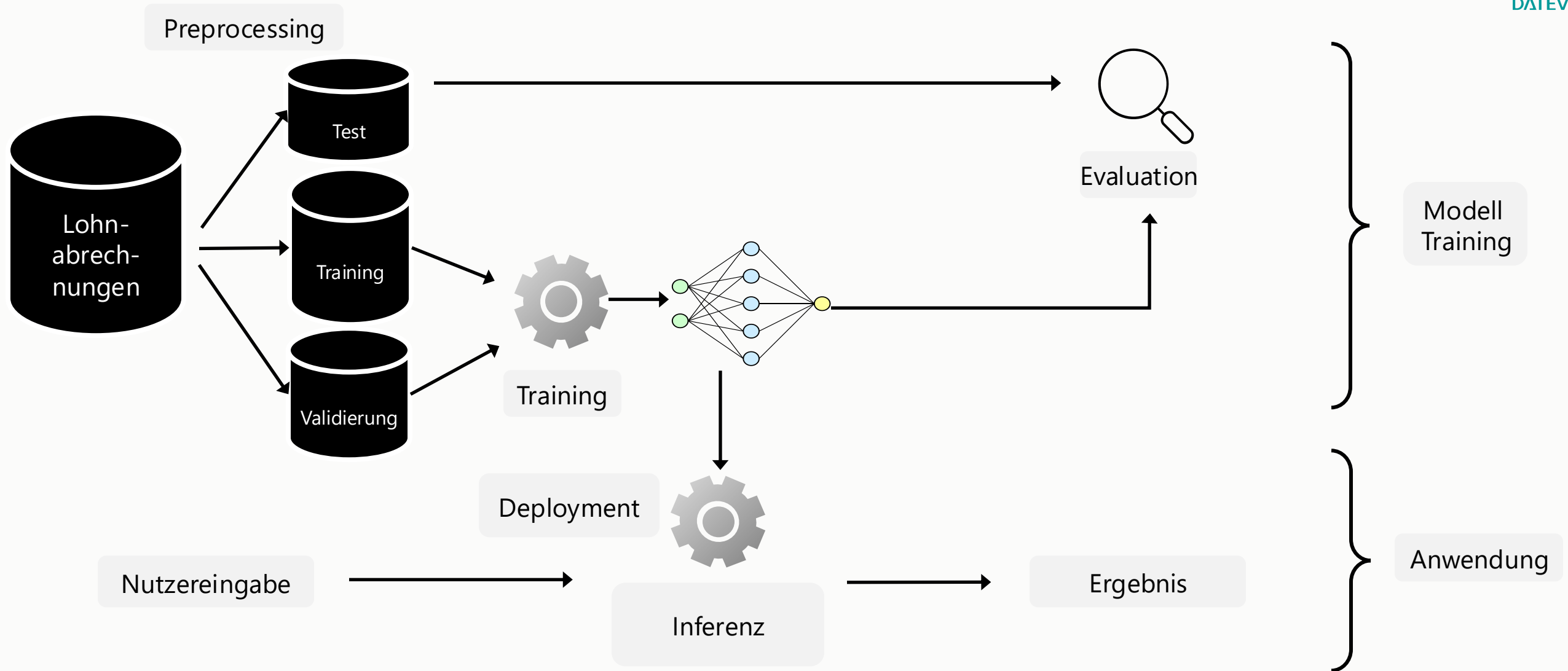


Angriff 2 Gehaltsstruktur eines Konkurrenten (Unternehmen)

- Membership Inference: Lohnabrechnungen des Konkurrenten in Trainingsdaten
- Multi-Attribute Inference: Extraktion Gehaltsstruktur eines Mandanten

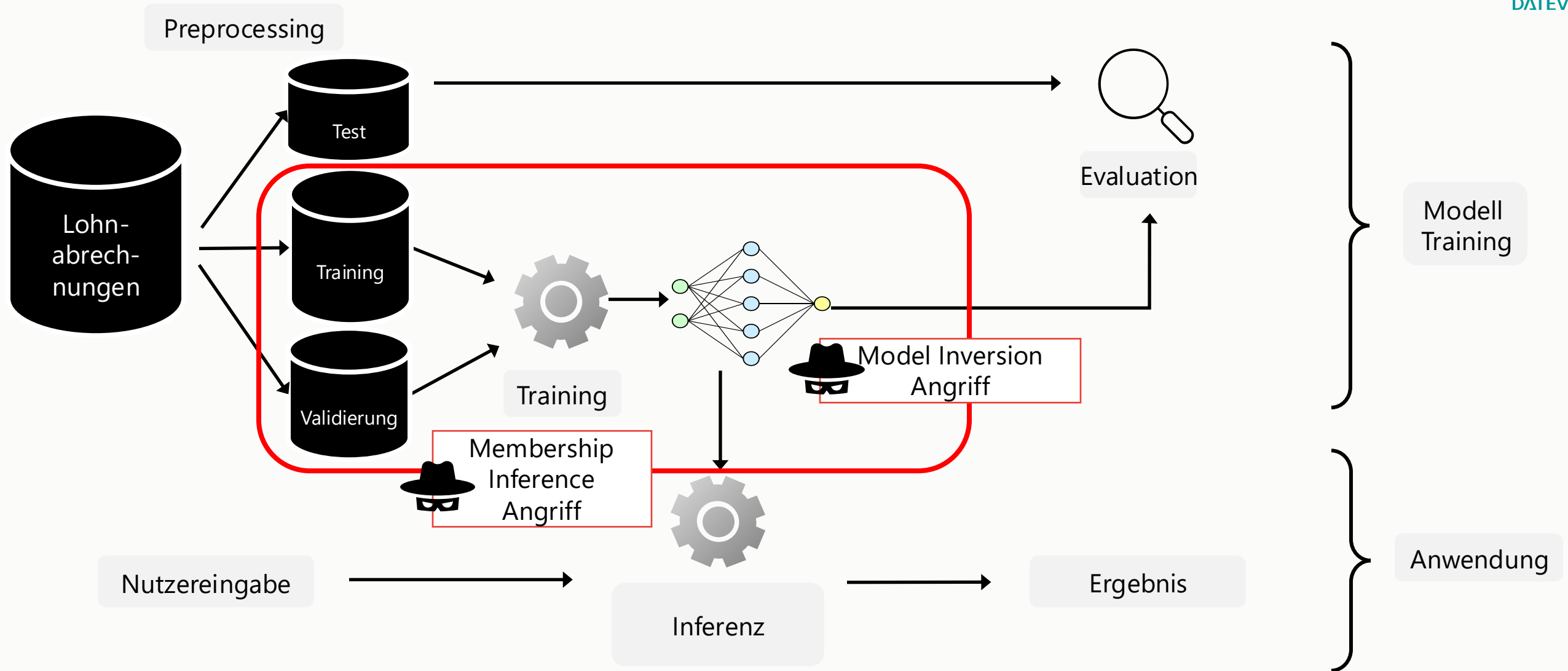
DATEV Personal-Benchmark Online

Architektur KI-System – Modell ohne Privacy



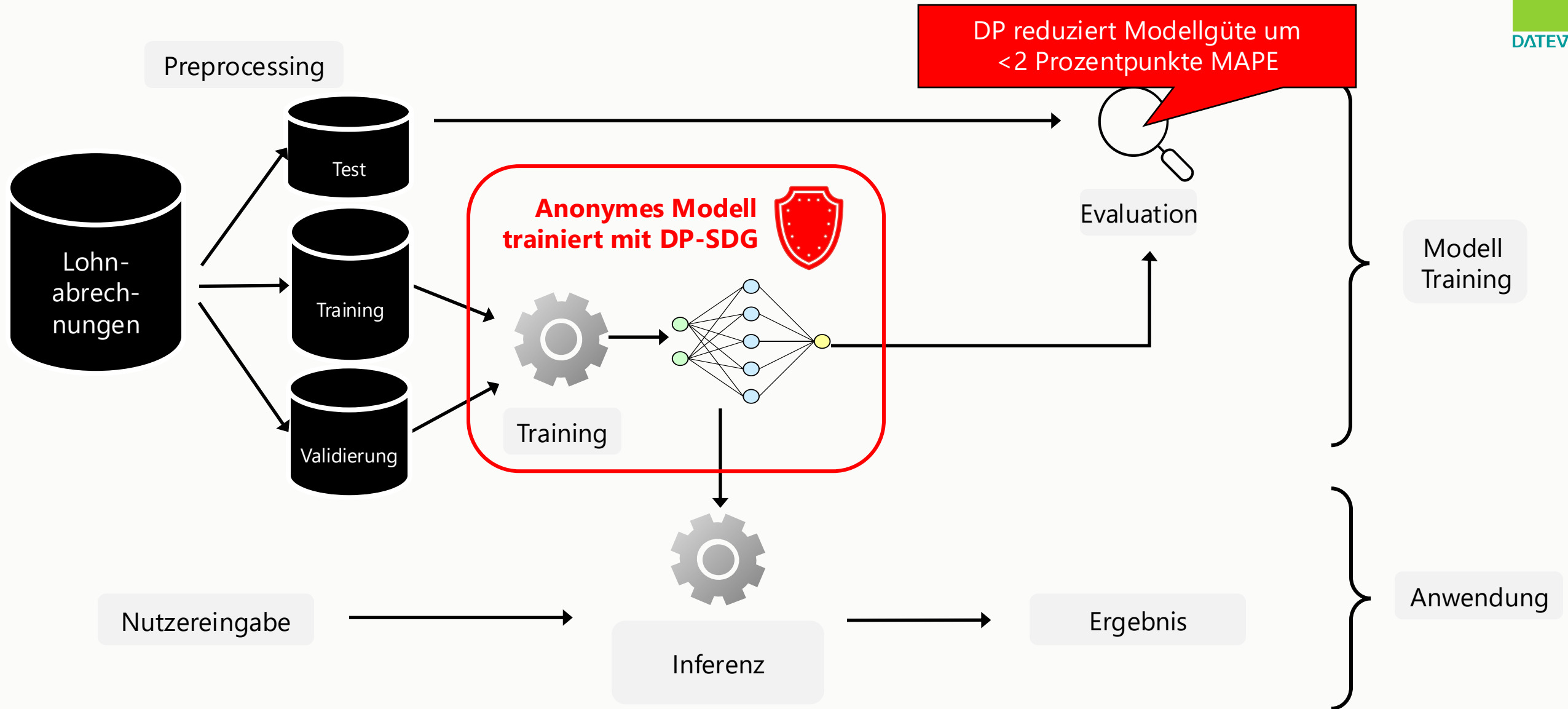
DATEV Personal-Benchmark Online

Architektur KI-System - Angriffe



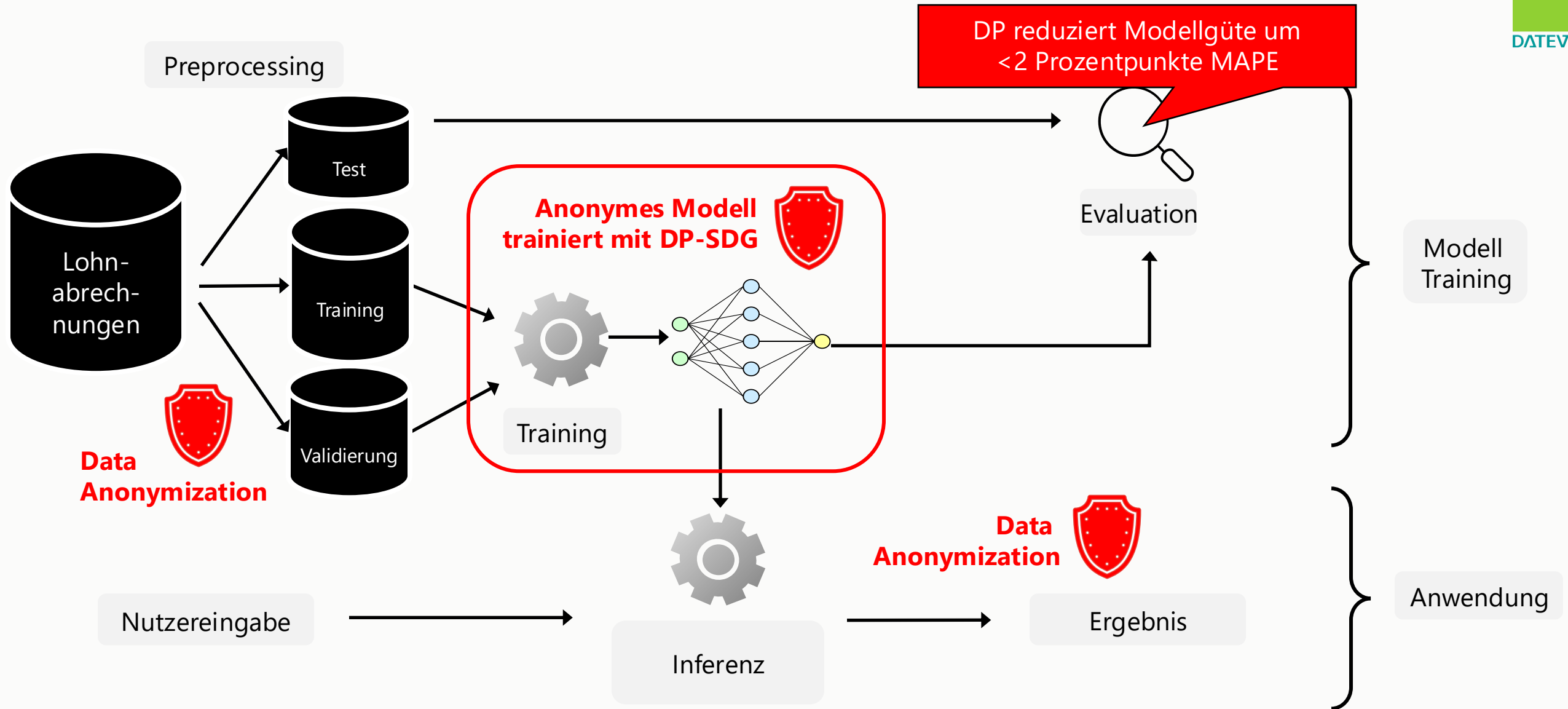
DATEV Personal-Benchmark Online

Architektur KI-System – Modell mit DP



DATEV Personal-Benchmark Online

Architektur KI-System – Modell mit DP und Privacy Amplification



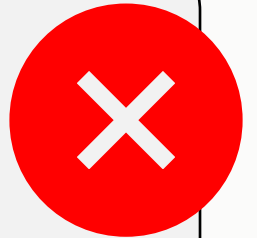
DATEV Personal-Benchmark Online

Angriffsszenarien auf die Privatsphäre



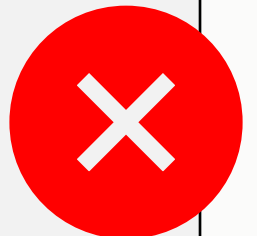
Angriff 1 Gehalt eines Nachbarn (Person)

- Membership Inference: Nachbar in Trainingsdaten enthalten
- Attribute Inference: Extraktion des tatsächlichen Gehalts



Angriff 2 Gehaltsstruktur eines Konkurrenten (Unternehmen)

- Membership Inference: Lohnabrechnungen des Konkurrenten in Trainingsdaten
- Multi-Attribute Inference: Extraktion aussagekräftige Gehaltsstruktur



DATEV Personal-Benchmark Online

Bewertung

PPML-Technologien für Verwendung von Lohnabrechnungen in KI-System

- Differential Privacy im Modelltraining (DP-SDG)
- Wirksamkeit in Penetrationstest mit externem Auditor belegt

Zusätzliche Absicherung des KI-Systems

- Privacy Amplification durch Anwendung Data Anonymization (u.a. k-Anonymität, Random Sampling)
- Geheimhaltung des (anonymen) Modells
- Technische und organisatorische Maßnahmen

BayLDA: Schutz des KI-Systems wirksam und entspricht dem Stand der Technik



Data illustrations by Storyset

Modernste Standards für sichere und vertrauenswürdige KI-Systeme durch

- Enge Zusammenarbeit von interner Produktentwicklung, IT-Sicherheit und Datenschutz
- Forschungsk Kooperationen mit Partnern in Industrie und Wissenschaft

Ausblick: Daran arbeiten wir

- Optimierte Informationssicherheit bei generativer KI
- Explainable AI für transparente und erklärbare KI in DATEV Software
- Management von KI-Chancen und Risiken durch Privacy Accounting



Worker illustrations by Storyset

Haben Sie Fragen? Interesse an einem Einstieg?

Besuchen Sie uns am **DATEV eG Stand in Halle 9 | 9-228**
oder **online**





Zukunft gestalten. Gemeinsam.

Allgemein

- (1) Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (pp. 308-318).
- (2) Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.
- (3) Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1322-1333).
- (4) Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.
- (5) Xu, R., Baracaldo, N., & Joshi, J. (2021). Privacy-preserving machine learning: Methods, challenges and directions. arXiv preprint arXiv:2108.04417.

Veröffentlichungen zu DATEV Personal-Benchmark online

- (1) Eichinger, F., Mayer, M. (2022). Predicting Salaries with Random-Forest Regression. In: Alyoubi, B., Ben Ncir, CE., Alharbi, I., Jarboui, A. (eds) Machine Learning and Data Analytics for Solving Business Problems. Unsupervised and Semi-Supervised Learning. Springer, Cham. https://doi.org/10.1007/978-3-031-18483-3_1
- (2) Eichinger, F., Kiesel, J., Dorner, M., Arnold, S. (2023). Estimations of Professional Experience with Panel Data to Improve Salary Predictions. In: Bramer, M., Stahl, F. (eds) Artificial Intelligence XL. SGA1 2023. Lecture Notes in Computer Science (vol 14381). Springer, Cham. https://doi.org/10.1007/978-3-031-47994-6_46