Datenschutz für KI aus technischer Perspektive

Joshua Stock
Sicherheit in verteilten Systemen (SVS)
https://svs.informatik.uni-hamburg.de

Arbeitsbereich Sicherheit in Verteilten Systemen (SVS)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Prof. Dr. Hannes Federrath
hannes.federrath@uni-hamburg.de
https://svs.informatik.uni-hamburg.de

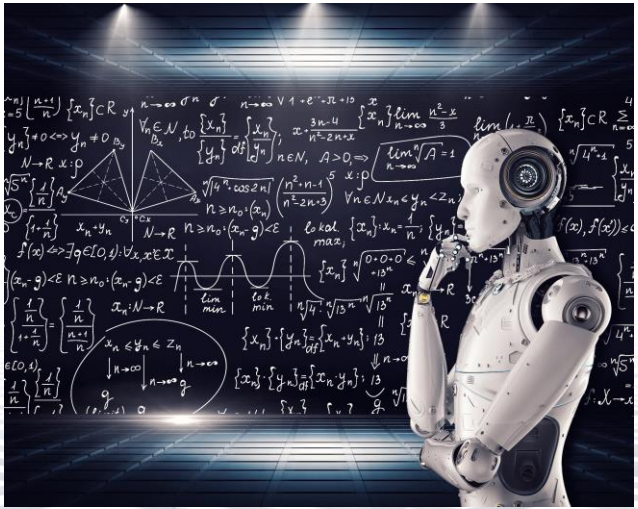
Joshua Stock
Wissenschaftlicher Mitarbeiter
joshua.stock@uni-hamburg.de

Themen:

- **Sicherheit und Privatsphäre** in Anwendungen
- **KI**: Privatsphäreangriffe und juristische Fragestellungen
- Datenschutzfreundliche **Angriffserkennung**
- **Anonymisierung** und Pseudonymisierung
- **IoT** Sicherheit und Forensik
- Sicherheit und Privatsphäre in der **Cloud**

Agenda

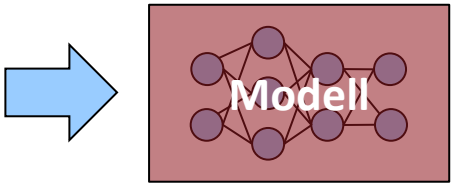
- KI entzaubern: Was bedeutet Künstliche Intelligenz?
- Datenschutzrisiko Privatsphäreangriffe
- Privatsphäreschützende Maßnahmen



Was ist Künstliche Intelligenz (KI)?

- Dieser Vortrag: KI synonym verwendet mit **Maschinellem Lernen**
  - „schwache KI“: Jeweils **eine** Aufgabe, kein allumfassendes System

- Beispiel:
  - Erkennung von Hirntumoren auf CT-Scans



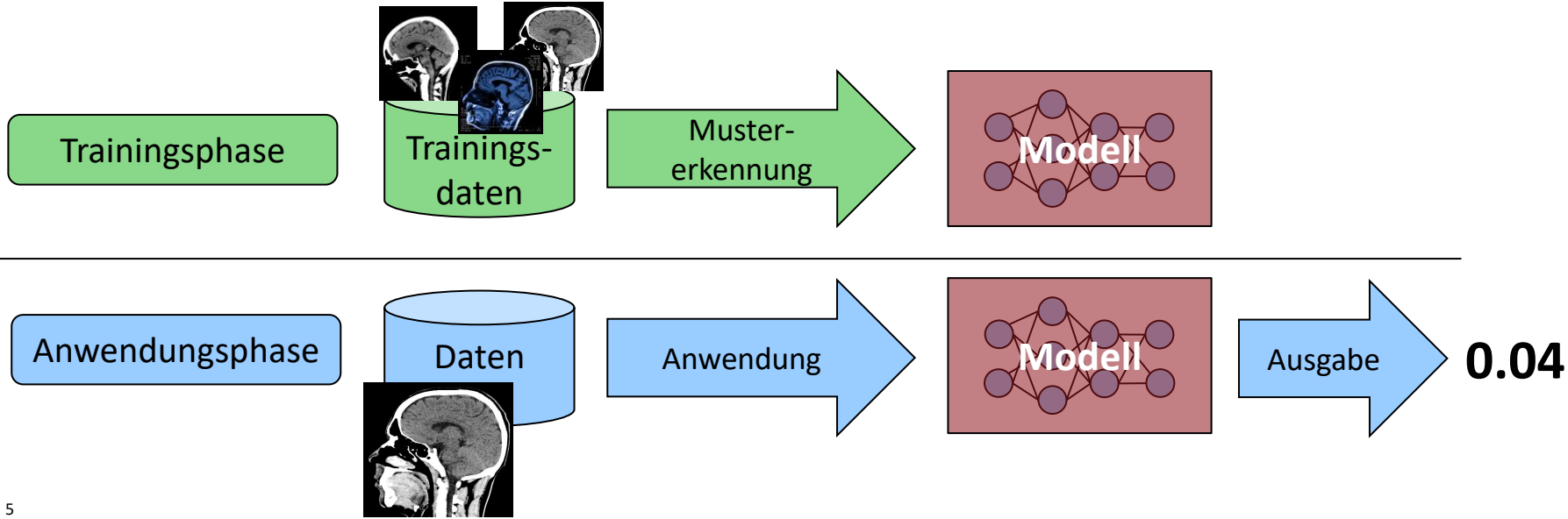
0.04

Kein Tumor erkannt,
Restwahrscheinlichkeit 4%

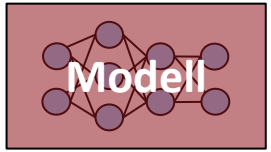
Was ist Künstliche Intelligenz (KI)?

- Dieser Vortrag: KI synonym verwendet mit **Maschinellem Lernen**
  - „schwache KI“: Jeweils **eine** Aufgabe, kein allumfassendes System

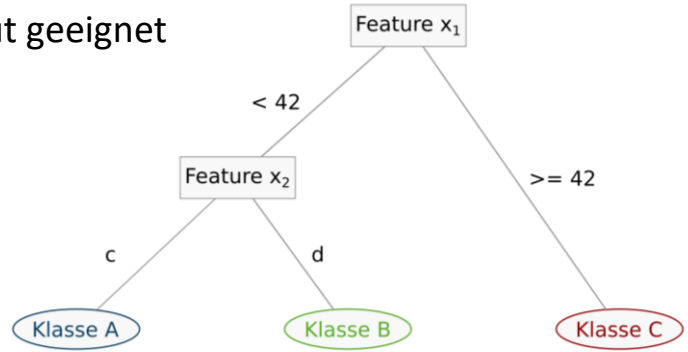
- Klasse von Algorithmen, die aus großen Datenmengen **Muster** extrahieren
- Im Anschluss: Anwendung auf **neue Daten**



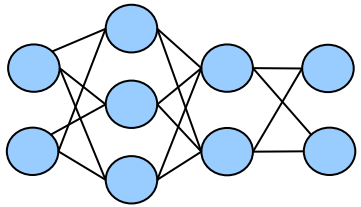
Was ist ein KI-Modell?



- Viele verschiedene Modelltypen
  - Je nach Aufgabengebiet und Aufgabe unterschiedlich gut geeignet



- Diverse Clustering-Algorithmen
- Entscheidungsbäume
- Ensemblemethoden wie Random Forests
- Lineare/Logistische Regression
- Support Vector Machines
- Künstliche Neuronale Netze
  - u.a. für Bildverarbeitung



- Gemeinsamkeit: Erst Training (Mustererkennung), dann Anwendung

Technische Datenschutzrisiken im KI-Kontext

- Naive Annahme: „Trainiertes Modell enthält nur abstrakte Muster, keine konkreten Daten“
- Modelle lernen oft mehr Informationen als für Aufgabe notwendig
- Extraktion von (sensiblen) Daten aus Modellparametern: **Privatsphäreangriffe**
  - Property Inference
  - Model Extraction
  - Membership Inference
  - Model Inversion

Privatsphäreangriffe 1/2

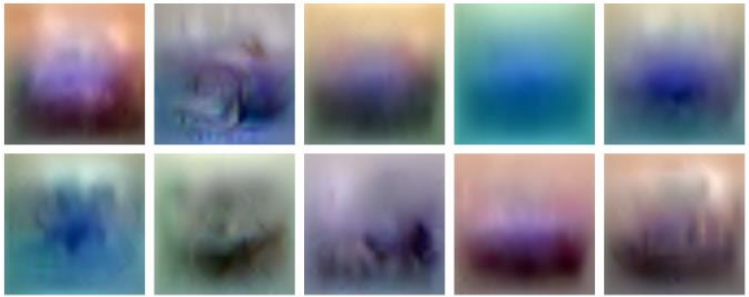
- **Property Inference** (auch: Distribution Inference) [Ga+18]
  - Zielt nicht auf Einzeldaten ab, sondern statistische Verteilungen
  - Beispiel: Wie hoch ist der Männeranteil in den Trainingsdaten?
  - Relevant, wenn Kund*innen-/Patient*innendemographie geheim bleiben soll
    - Weniger relevant, wenn Verteilungen transparent sind (klinische Studien)

- **Model Extraction** (auch: Model Stealing) [OSF19]
  - Betrifft eher Geschäftsgeheimnisse als Datenschutz
  - Wenn nur **Zugang zum Modell**, nicht Modell(parameter) selbst veröffentlicht werden
    - „Black box“: Anfragen an das Modell, ohne Interna zu kennen
  - Ermöglicht Rekonstruktion der Modellparameter durch wiederholte Anfragen

Privatsphäreangriffe 2/2

- **Membership Inference** [SS+17]
  - Zeigt, ob einzelne Datenpunkte Teil der Trainingsdaten waren
  - Relevant, wenn Zugehörigkeit zu Trainingsdaten bereits sensible Information darstellt

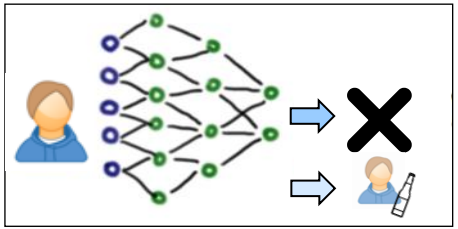
- **Model Inversion** [ZH20]
  - Rekonstruktion von Trainingsdaten
  - Vor allem für Klassifizierungsmodelle anwendbar mit wenig Diversität in Klassen
  - Auch: Kreditkartennummern aus Sprachmodell



Beispiel 1: *Membership Inference* Angriff

- Patient Herr Müller: alkoholkrank
- Einwilligung in Suchtklinik zu Verarbeitung seiner anonymisierten Patientendaten
  - Training eines KI-Modells der Firma *AlcScore*

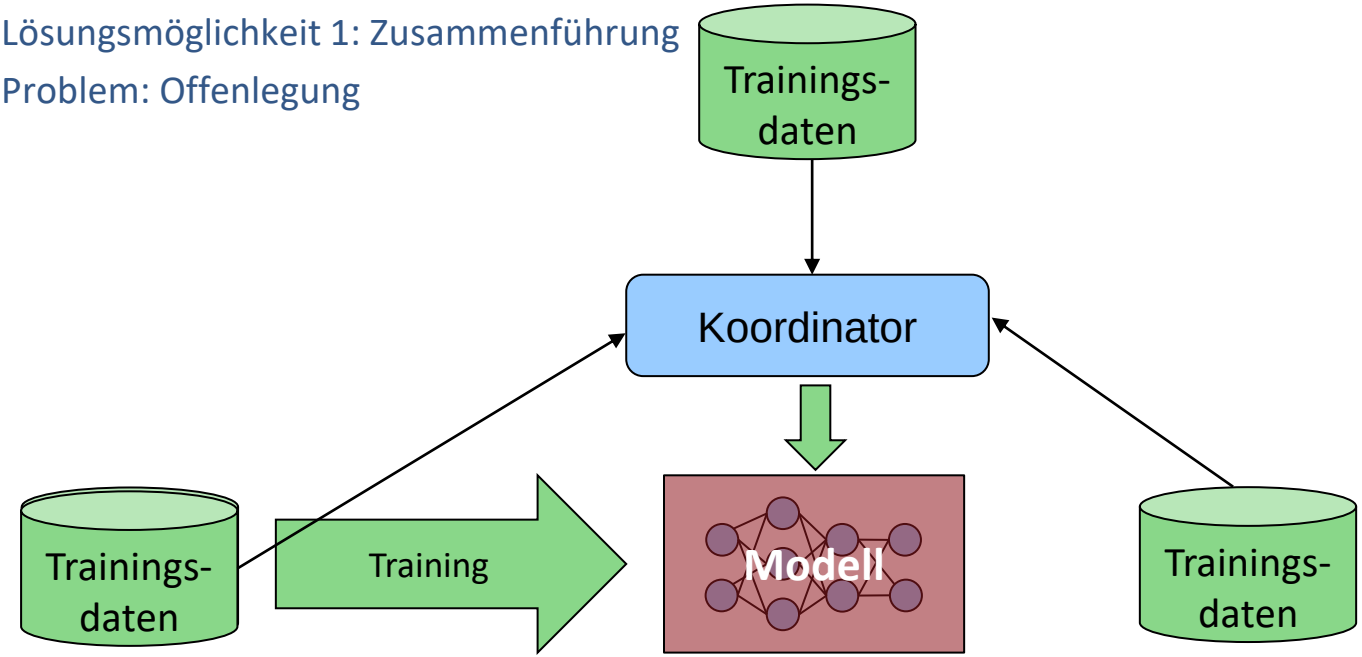
- Idee von *AlcScore*:
  - Anhand von Patientendaten Risikofaktoren extrahieren
  - Eingabe in KI-Modell: Anonymisierte Patientendaten
  - Ausgabe: Risikoscore Neigung zu Alkoholismus



- Jahre später: Herr Mülller möchte Lebensversicherung abschließen und muss Gesundheitsdaten angeben
  - Versicherung hat Modell von *AlcScore* akquiriert
  - Kann anhand Herrn Müllers Daten nicht nur Score berechnen
  - Durch *Membership Inference:* Herr Müller war Teil der *AlcScore*-Trainingsdaten
  - Diagnostizierte Sucht → Versicherung lehnt Vertrag mit Herrn Müller ab

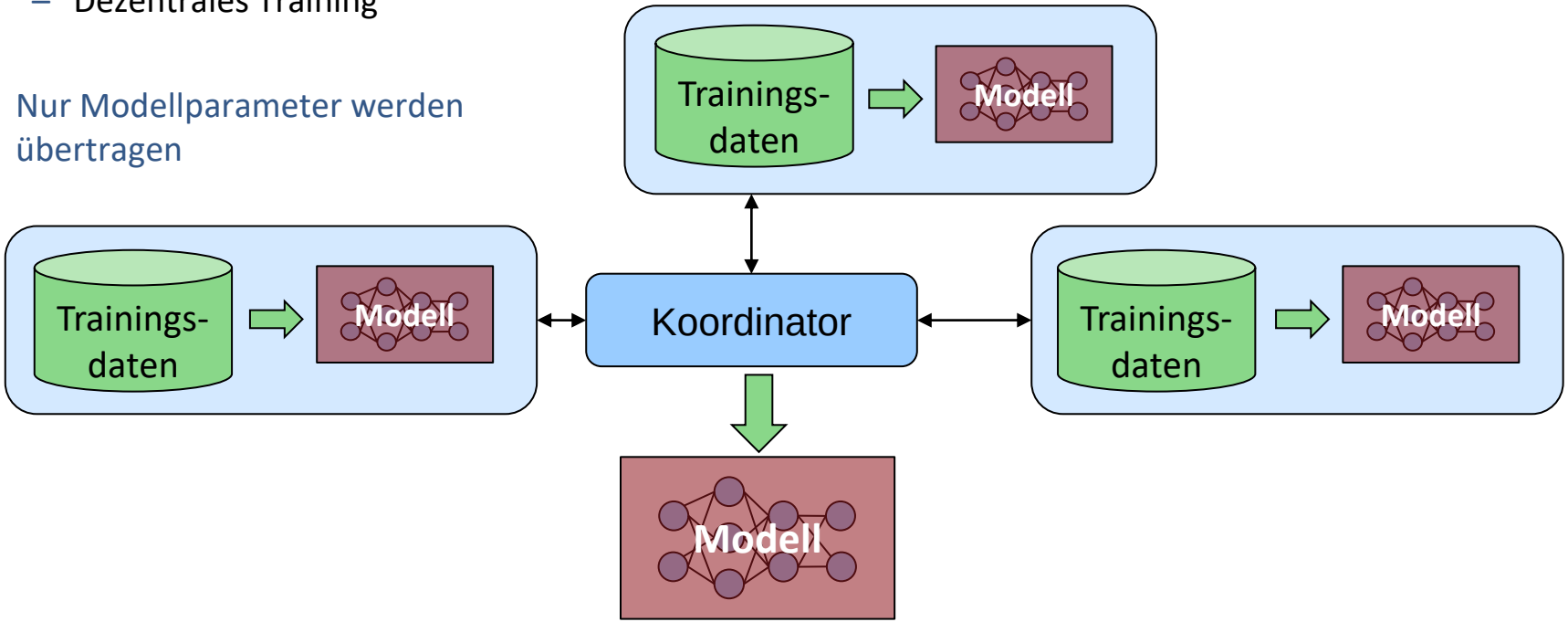
Verteilte Trainingsdaten

- Problem: Trainingsdaten liegen verteilt vor
  - Z.B. MRT-Scans in mehreren Krankenhäusern
- Lösungsmöglichkeit 1: Zusammenführung
- Problem: Offenlegung



Verteilte Trainingsdaten: Federated Learning

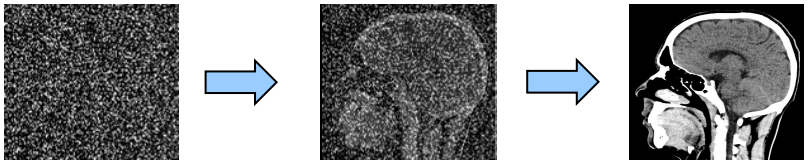
- Lösung 2: Federated Learning
  - Dezentrales Training
- Nur Modellparameter werden übertragen



Beispiel 2: Federated Learning Rekonstruktionsangriff

- **Federated Learning** (FL): Trainingsdaten für Modell sind auf mehrere Standorte verteilt
  - Vorteil: Daten müssen nicht ausgetauscht werden
  - Zentraler Server koordiniert
    - Jeder Standort trainiert lokales Modell
    - Schickt aktualisierte Modellparameter an Server
    - Server aggregiert Parameter, aktualisiert und verteilt globales Modell

- Beispiel: Gesundheitsdienstleister bietet FL-Dienst für gemeinsame Modellentwicklung an
  - Hirntumorerkennung auf CT-Bildern
  - Kliniken nehmen teil: lediglich Modellupdates und keine CT-Bilder werden übertragen
  - Dienstleister kann aus Modellupdates einzelner Klinik Bilder rekonstruieren [ZH20]
    - Start bei zufälligem Rauschen; iterativer Prozess, bis Modellparameter 1:1 übereinstimmen



Mögliche Schutzmaßnahmen (*Privacy Preserving Machine Learning*)

- Verteilter Datensatz: **Federated Learning**

- **Vorverarbeitung** von Trainingsdaten: oft nicht ausreichend
  - Zum Beispiel Zensur von sensiblen Merkmalen
    - KI-Modelle können u.U. Merkmale rekonstruieren

- **Differential Privacy**: Unschärfe in Modellen
  - Schutz von Individuen mit mathematischen Garantien
  - Gut erforscht und in der Praxis angekommen [VeBe22]
  - Problem: Privatsphäre oft schwer quantifizierbar
    - Wie viel Unschärfe ist nötig?

Mögliche Schutzmaßnahmen (*Privacy Preserving Machine Learning*)

- **Homomorphe Verschlüsselung**
  - Erlaubt Berechnungen auf verschlüsselten Daten
  - Trainingsphase: Modellerzeugung auf Fremdserver ohne Datenzugriff
  - Anwendungsphase: Datenauswertung auf Fremdserver ohne Dateneinsicht

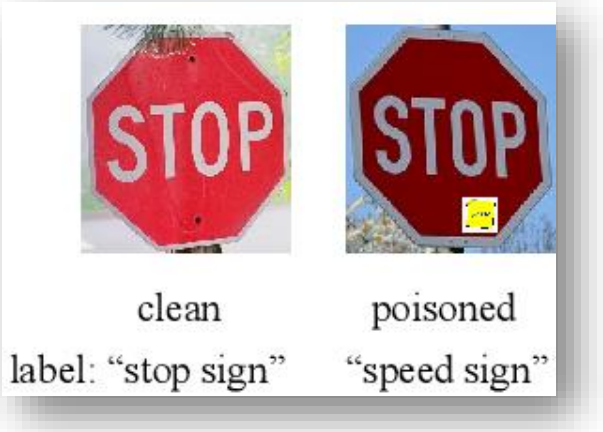
- **Secure Multi-Party Computation**
  - Gemeinsame Modellerzeugung im Federated Learning
  - Ohne Offenlegung individueller Beiträge zum Gesamtmodell

- Beide obenstehenden Ansätze: Sehr rechenintensiv [CP21]
  - Noch nicht praxistauglich → Zukunftsthema

Exkurs: Poisoning

- **Poisoning-Angriffe**: „Vergiften“ von Modellen führt zu Fehlern im Modell [DS19]
  - Z.B. Sticker auf Stoppschild führt zu Erkennung als Geschwindigkeitsbegrenzung

- Setzen **während** des Trainings an
- Privatsphäreangriffe: Nach dem Training



Zusammenfassung

- „Künstliche Intelligenz“: Meist Insellösungen für spezifische Probleme
- Viele Methoden, oft schwer miteinander vergleichbar
- Gemeinsamkeit: Daten bilden Grundlage für Training

- Privatsphäreangriffe können sensible Daten aus Modellen extrahieren
  - Disclaimer: Aufwand ist oft erheblich, Vorwissen erforderlich
  - Modelle nicht unbedacht teilen

- Techniken zur Privatsphäresteigerung existieren
  - Individuelle Abwägung, keine Universallösungen
  - In Zukunft: Verstärkter Einsatz von Verschlüsselung im KI-Kontext möglich

Referenzen

- **[CP21]** Cabrero-Holgueras, J., & Pastrana, S. (2021). SoK: Privacy-Preserving Computation Techniques for Deep Learning. *Proceedings on Privacy Enhancing Technologies*, *2021*(4), 139-162.
- **[DS19]** Davaslioglu, K., & Y.E. Sagduyu. (2019). Trojan attacks on wireless signal classification with adversarial machine learning. *IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*.
- **[Ga+18]** Ganju, K., et al. (2018). Property inference attacks on fully connected neural networks using permutation invariant representations." *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*.
- **[OSF19]** Orekondy T., Schiele B., and Fritz M. (2019). Knockoff nets: Stealing functionality of black-box models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.
- **[SS+17]** Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 3-18). IEEE.
- **[VB22]** https://venturebeat.com/ai/google-releases-differential-privacy-tools-to-commemorate-data-privacy-day/
- **[ZH20]** Zhu, L., & Han, S. (2020). Deep leakage from gradients. In *Federated learning* (pp. 17-31). Springer, Cham.