RECHTSWISSENSCHAFTLICHE FAKULTÄT
UNIVERSITÄT DES SAARLANDES

Lehrstuhl für Rechtsinformatik
Universität des Saarlandes

Technischer Datenschutz für KI-Anwendungen: Herausforderungen und Lösungsansätze

Christoph Sorge
Lehrstuhl für Rechtsinformatik
Universität des Saarlandes

Der Lehrstuhl in Kürze

- Drittmittelstarker und großer Lehrstuhl, besetzt mit einem Informatiker
- Schwerpunkte
  - Datenschutz durch Technik (Privacy Enhancing Technologies)
  - Informationsrechtliche Fragestellungen an der Schnittstelle zur IT-Sicherheit
  - IT für Justiz und Verwaltung
  - IT-Forensik
  - Maschinelles Lernen auf juristischen Texten

Der Lehrstuhl: Einbettung und Kooperationen

Künstliche Intelligenz: Grundlagen

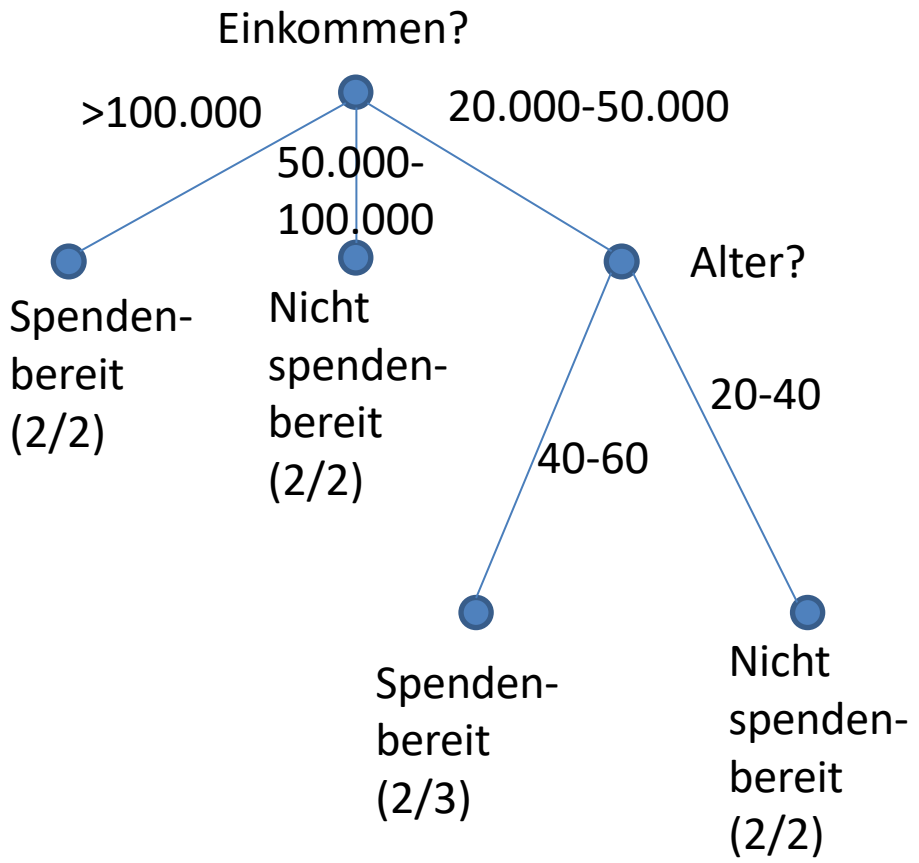
- Artificial Intelligence / Künstliche Intelligenz
  - Eigentlich unscharfer Begriff
  - Große Breite an Forschungsaktivitäten, z.B. Systeme zum automatischen Schlussfolgern anhand vorgegebener Regelwerke, Wissensrepräsentation, Wahrnehmung der Umwelt, Sprachverarbeitung, …

- Große Fortschritte in den letzten Jahren insbesondere im Bereich des Maschinellen Lernens durch Einsatz statistischer Verfahren

- Deshalb hier der Schwerpunkt

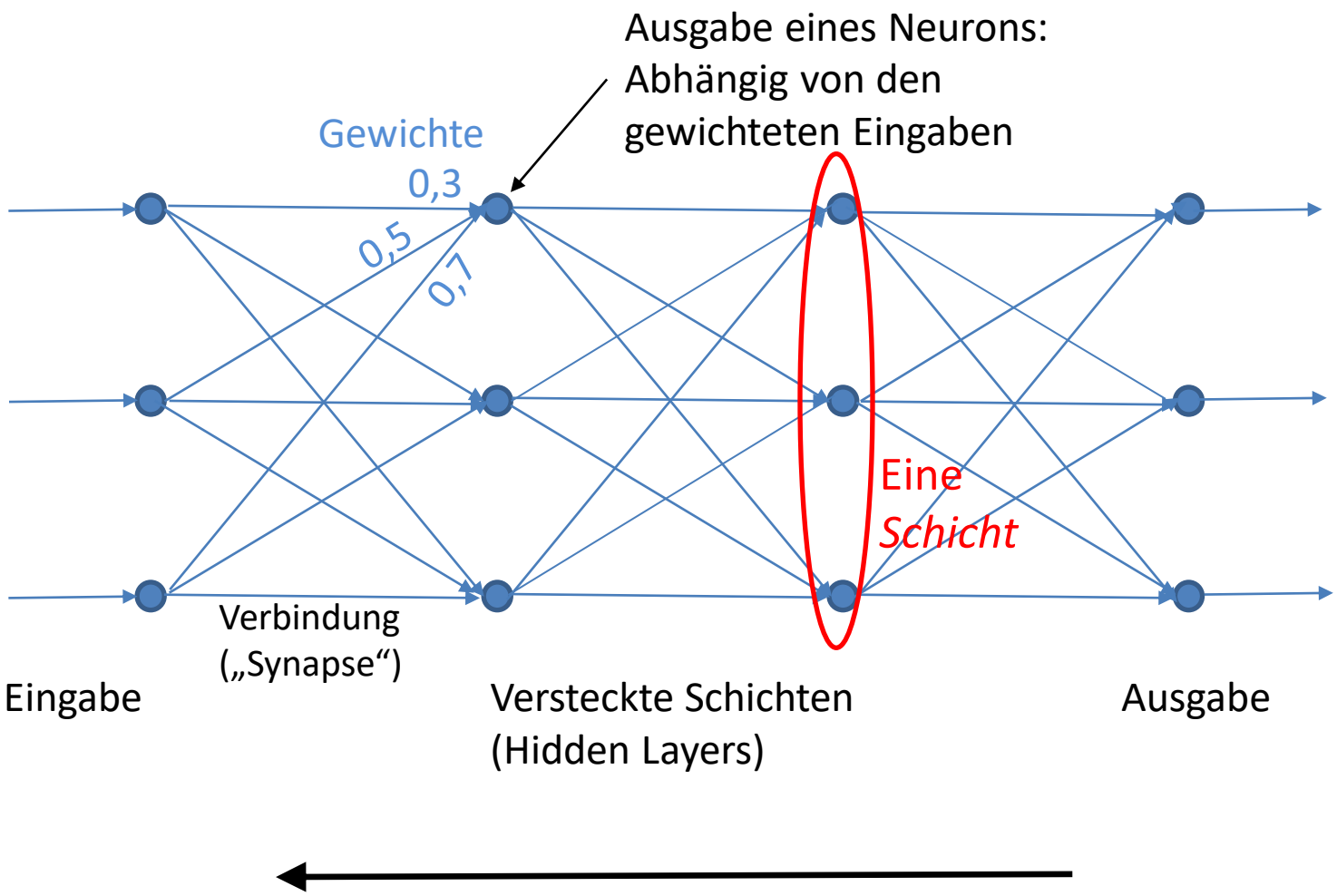
Beispiel: Entscheidungsbaum

Einkommen	Alter	Spenden-bereitschaft
20.000 bis 50.000	20-40	Nein
20.000 bis 50.000	20-40	Nein
20.000 bis 50.000	40-60	Ja
20.000 bis 50.000	40-60	Ja
20.000 bis 50.000	40-60	Nein
50.000 bis 100.000	20-40	Nein
50.000 bis 100.000	40-60	Nein
> 100.000	40-60	Ja
> 100.000	20-40	Ja



Ziel: Spendenbereitschaft von Menschen vorhersagen
→ Aufbau eines Modells anhand von Trainingsdaten, dann Anwendung des Modells (hier: Entscheidungsbaum)

Beispiel: Neuronale Netze



Lernprozess: Aktualisierung der Gewichte anhand von Rückmeldungen über die Ausgabe

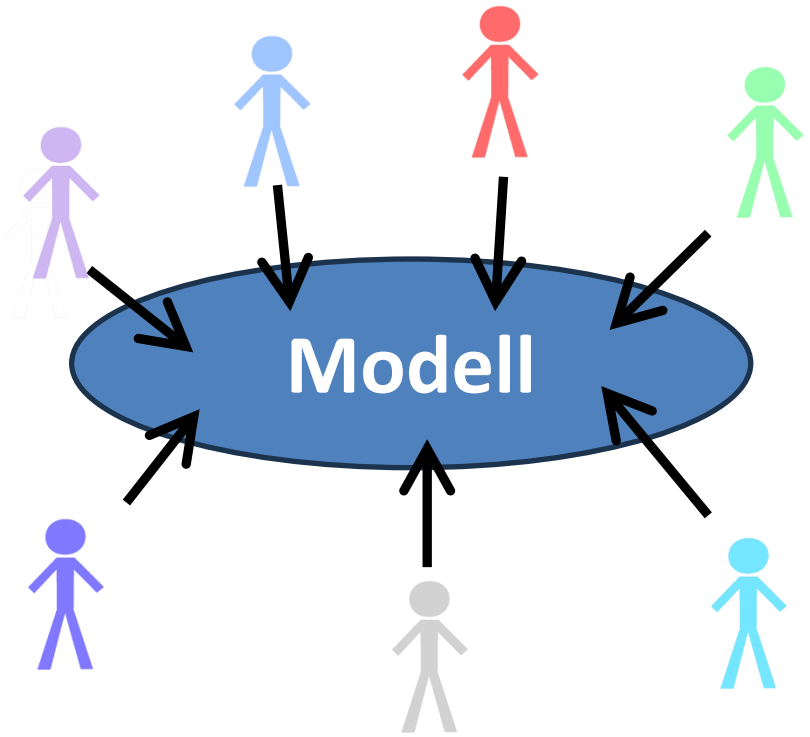
Heutige Neuronale Netze:

- Oft zwei- bis dreistellige Anzahl Schichten, aktuell auch schon > 1000 Schichten („Deep Learning“)
- In einigen Fällen hunderttausende bis Millionen Neuronen

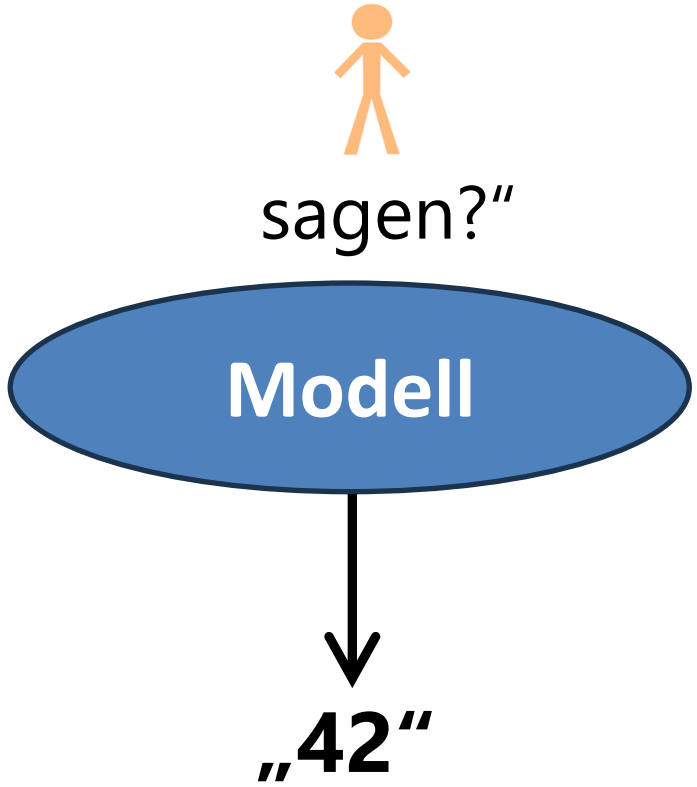
Auch hier: Aufbau eines Modells anhand von Trainingsdaten

Modelle – abstrakter

- Künstliche Intelligenz

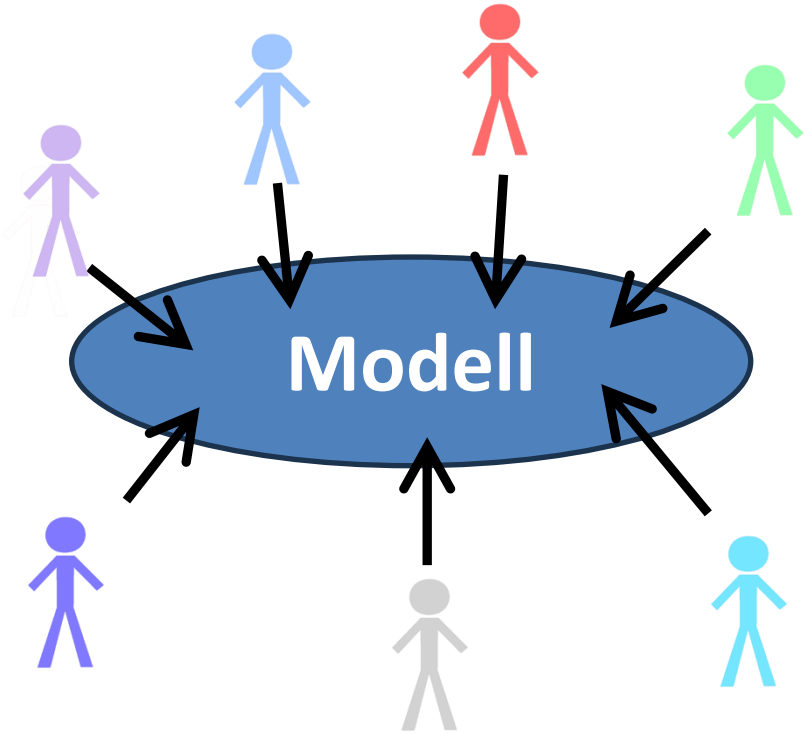


„Was kannst du mir über

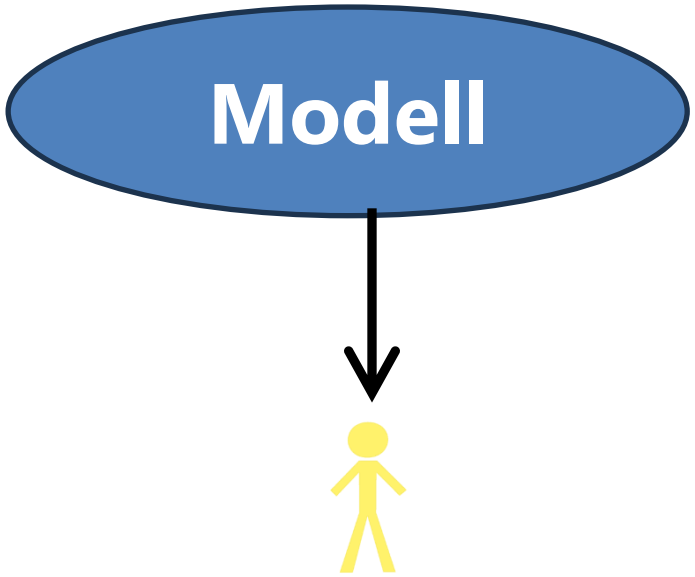


Modelle – abstrakter

- Künstliche Intelligenz



„Gib mir mehr!“



Leistungsfähigkeit maschineller Lernverfahren

- Durchbruch in der Leistungsfähigkeit maschineller Lernverfahren in den letzten Jahren

- Beispiel: Im Wettbewerb „ImageNet Large Scale Visual Recognition Challenge“ erreichen maschinelle Lernverfahren seit 2015 bessere Klassifikationsergebnisse als Menschen

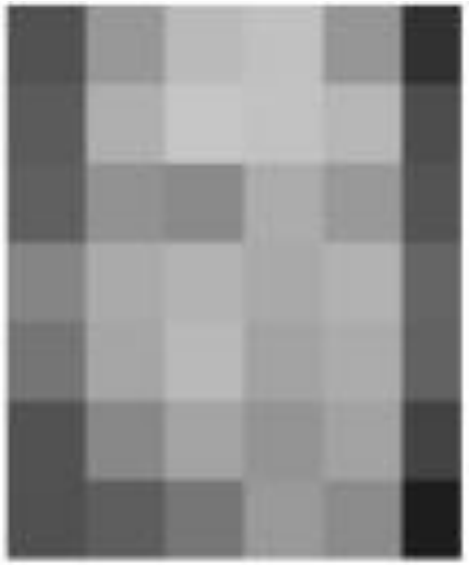
Bildquelle: https://cs.stanford.edu/people/karpathy/cnnembed/

Potentielle Probleme aus Sicht des Datenschutzes

- Aktuelle Versprechungen: Künstliche Intelligenz erlaubt
  - das Erlernen von Zusammenhängen und Mustern in großen Datenmengen
  - auch über die Grenzen einzelner Anwendungsdomänen hinweg
  - sowie das Abbilden der Zusammenhänge in einem Modell
  - selbst wenn diese Zusammenhänge für Menschen nicht ohne weiteres nachvollziehbar sind,
  - und in vielen Fällen mit besserer Treffsicherheit als Menschen, so dass Entscheidungen direkt durch die Maschine getroffen werden können

Problem 1: Erkennen von Mustern

- Problem: Erkennen von Mustern
  - Beispiel: McPherson et al 2016, „Defeating Image Obfuscation with Deep Learning“
  - Wiedererkennung einer Person anhand des „verpixelten“ Bilds (rechts)
  - In einem Datensatz aus 400 Bildern von 40 Personen: Erkennung der Person mit einer Genauigkeit von > 95%



→Potentielle De-Anonymisierung momentan noch als anonym geltender Daten

- Anwendung z.B. auch bei Videoüberwachung

Frage: Wie mit einem sich dynamisch ändernden Personenbezug rechtlich umgehen?

Problem 2: Erkennen von Mustern in großen Datenmengen und die ökonomischen Folgen

- Im Grundsatz: Erkennung von Mustern im Verhalten von Personen erfolgversprechender, …
  - wenn möglichst viele Daten über eine Person gesammelt und zusammengeführt werden
  - wenn Daten über möglichst viele Personen gesammelt und zusammengeführt werden

- Praxisbeispiel: Spracherkennung durch Google (und Amazon etc.): in der Regel beim Anbieter, nicht auf dem jeweiligen Endgerät

Frage: Wie die durch KI resultierende Monopolisierungstendenz regulatorisch einhegen?

Problem 3: Domänenüberschreitendes Zusammenführen von Daten

- Zusammenführen von Daten aus verschiedenen Anwendungsdomänen verspricht neue Erkenntnisse in der Forschung und ggf. übergreifende Anwendungen
- Beispiel medizinische Forschung
  - Bereits bekannt: Möglichkeit, bestimmte Herzerkrankungen anhand der Daten von Fitness-Trackern zu erkennen
  - Neue Erkenntnisse denkbar durch Einbeziehung von
    - Suchanfragen
    - Postings in den Social Media
    - Terminkalender, Flugbuchungen, …
    - Online-Bestellungen (→ Ernährungsgewohnheiten)
    - Arbeitszeiterfassung

Problem 3: Domänenüberschreitendes Zusammenführen von Daten

- Zusammenführen von Daten aus verschiedenen Anwendungsdomänen verspricht neue Erkenntnisse in der Forschung und ggf. übergreifende Anwendungen
- Anwendungen:

PETs (Privacy Enhancing Technologies) gut verstanden, Anonymisierung meist möglich

	Einzelne Nutzer	Mehrere Nutzer
Eindimensionale Daten	Berichte (z.B. Gewicht im Zeitverlauf) u.a. für Gesundheitszwecke	Wettbewerb
Mehrdimensionale Daten	Korrelationen (z.B. Auswirkungen der Ernährung auf die Fitness des Einzelnen)	Forschung (Einschließlich: Finden neuer unerwarteter Zusammenhänge)

Besonders großer Nutzen durch KI

Problem 3: Domänenüberschreitendes Zusammenführen von Daten

- Zusammenführen von Daten aus verschiedenen Anwendungsdomänen verspricht neue Erkenntnisse in der Forschung und ggf. übergreifende Anwendungen

- Spannungsverhältnis zu…
  - Zweckbindungsgrundsatz
  - Datenschutz durch Technikgestaltung und durch datenschutzfreundliche Voreinstellungen (→ Welche Daten sind erforderlich für eine bestimmte Verarbeitung? Wie stark muss ein Attribut mit dem Zielattribut korreliert sein, wenn es für Maschinelles Lernen eingesetzt wird?)
  - Einwilligungserfordernissen – kann es noch eine informierte Einwilligung geben? (Insbesondere: Verarbeitung besonderer Kategorien personenbezogener Daten)

Problem 4: Erkennen von Mustern in großen zusammengesuchten Datenmengen aus dem Web

- Beispiel große Sprachmodelle (Large Language Models, LLMs): Trainingsdaten oft im World Wide Web gesammelt
  - Datensammlung für Nutzer in der Regel nicht transparent
  - Rechtsgrundlage der Verarbeitung?

- Umgang mit Betroffenenrechten
  - Wie Betroffene identifizieren?
  - Betroffene ggf. „identifizierbar genug“ für Anwendbarkeit der DSGVO – aber nicht „einfach genug identifizierbar“ für automatische Identitätsprüfung

Problem 5: Abbilden von Zusammenhängen in einem Modell

- Oft getroffene, implizite Annahme: Auch beim Lernen aus personenbezogenen Datensätzen stellt das Modell eine Aggregation dar, die keine personenbezogenen Daten mehr enthält

- Aktuelle Informatik-Forschung zeigt: Diese Annahme stimmt oft nicht

- Beispiel: Shokri et al. 2017, „Membership Inference Attacks Against Machine Learning Models"

  - Gängige maschinelle Lernverfahren ermöglichen Rückschlüsse darauf, ob ein bestimmter Datensatz (→ in der Anwendung: Daten über eine bestimmte Person) beim Trainieren des Modells einbezogen wurde

Problem 5: Abbilden von Zusammenhängen in einem Modell

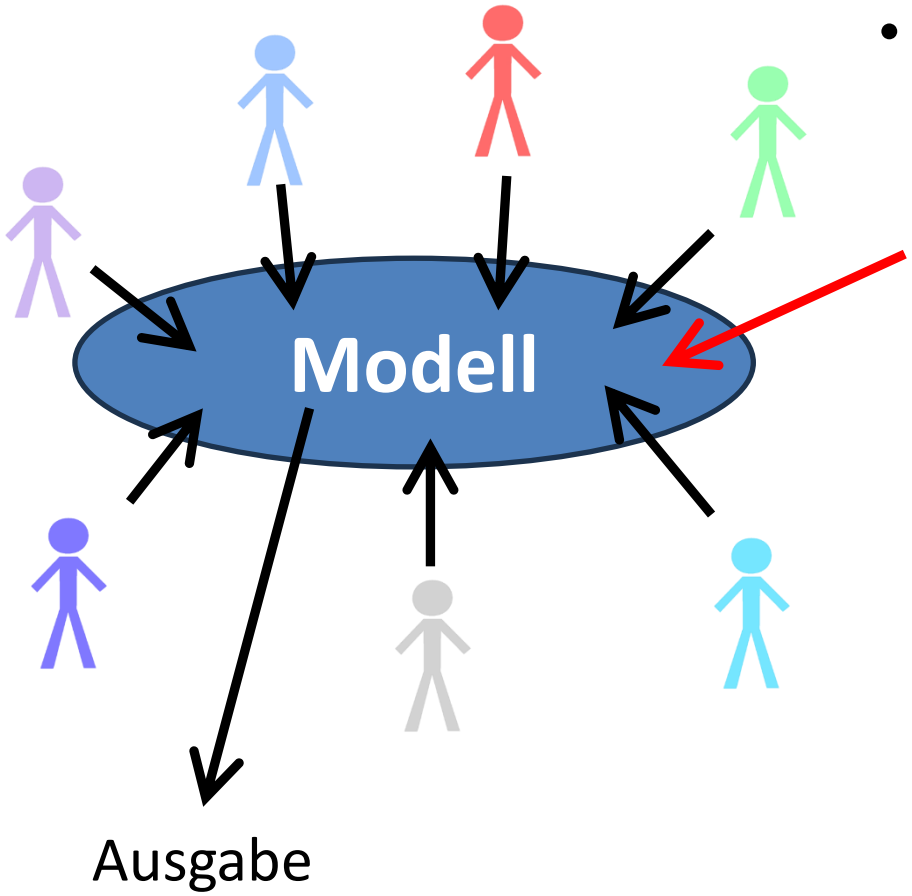
- Oft getroffene, implizite Annahme: Auch beim Lernen aus personenbezogenen Datensätzen stellt das Modell eine Aggregation dar, die keine personenbezogenen Daten mehr enthält

- De facto: Gängige Evaluationen von maschinellen Lernverfahren prüfen die Aussagekraft und Genauigkeit des Modells, aber nicht
  - ob „zu viele“ Daten im Modell enthalten sind
  - ob eigentlich irrelevante Attribute im Modell mit berücksichtigt werden
  - ob Rückschlüsse auf einzelne Datensätze / Personen in den Trainingsdaten möglich sind

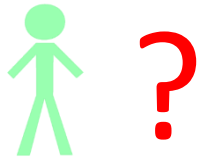
- Problem beispielsweise bei der Autovervollständigung in Smartphone-Tastatur-Apps
- Folgen für den Löschungsanspruch / Recht auf Vergessenwerden?

Problem 5: Abbilden von Zusammenhängen in einem Modell

- Klassifikation von „Privacy-Angriffen“ auf ML-Modelle nach Rigaki/Garcia (ACM Computing Surveys 2023)

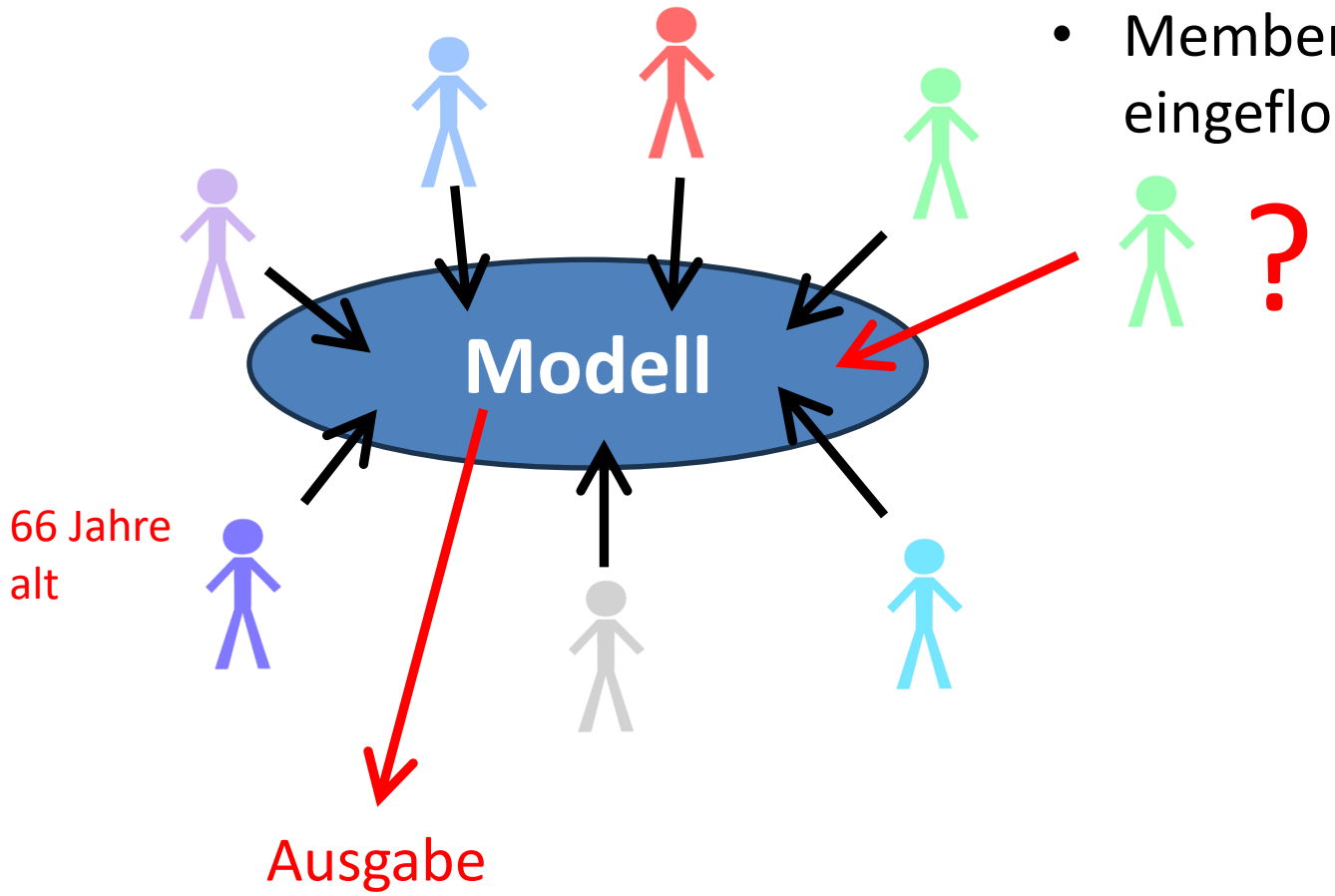


- Membership Inference: Ist in das Modell eingeflossen?



Problem 5: Abbilden von Zusammenhängen in einem Modell

- Klassifikation von „Privacy-Angriffen“ auf ML-Modelle nach Rigaki/Garcia (ACM Computing Surveys 2023)

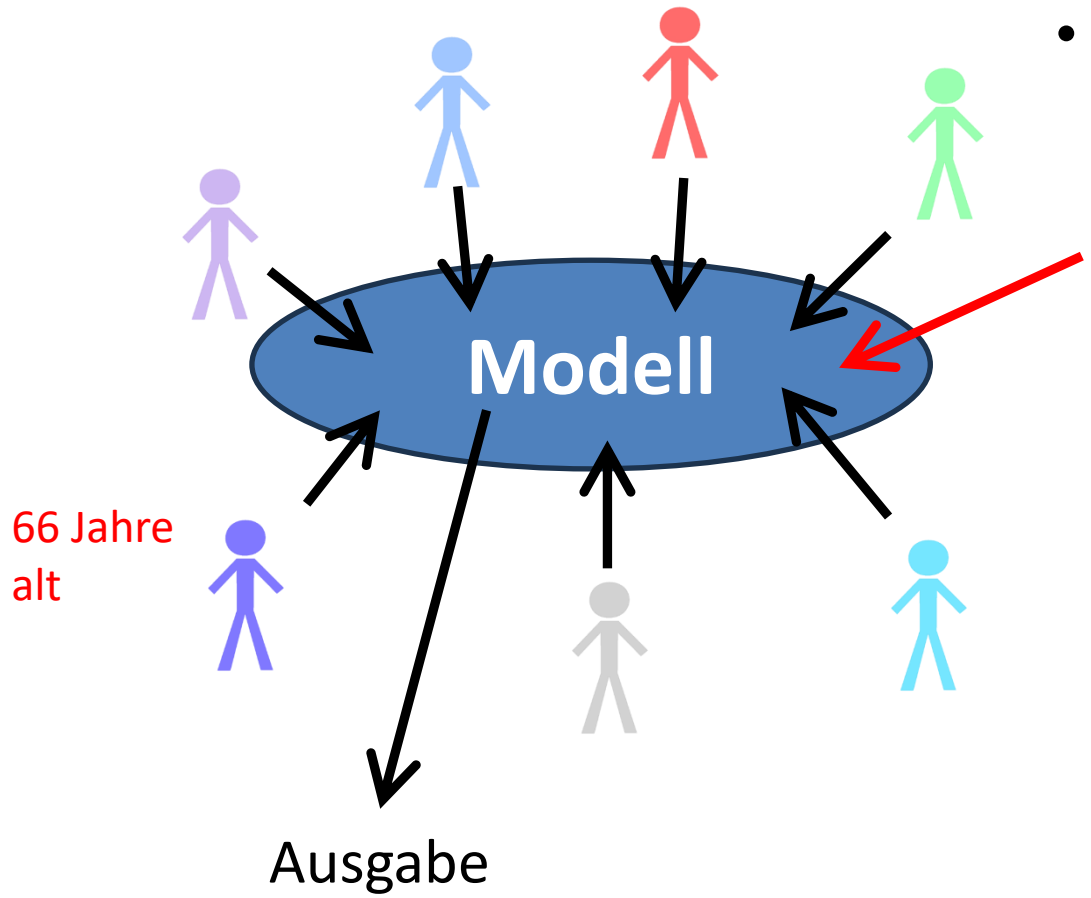


- Membership Inference: Ist in das Modell eingeflossen?

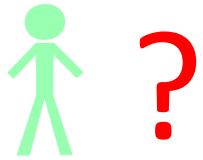
- Reconstruction: Rückschluss insb. aus Ausgaben und einigen Attributen von  auf weitere/alle ihre Attribute aus den Trainingsdaten

Problem 5: Abbilden von Zusammenhängen in einem Modell

- Klassifikation von „Privacy-Angriffen“ auf ML-Modelle nach Rigaki/Garcia (ACM Computing Surveys 2023)



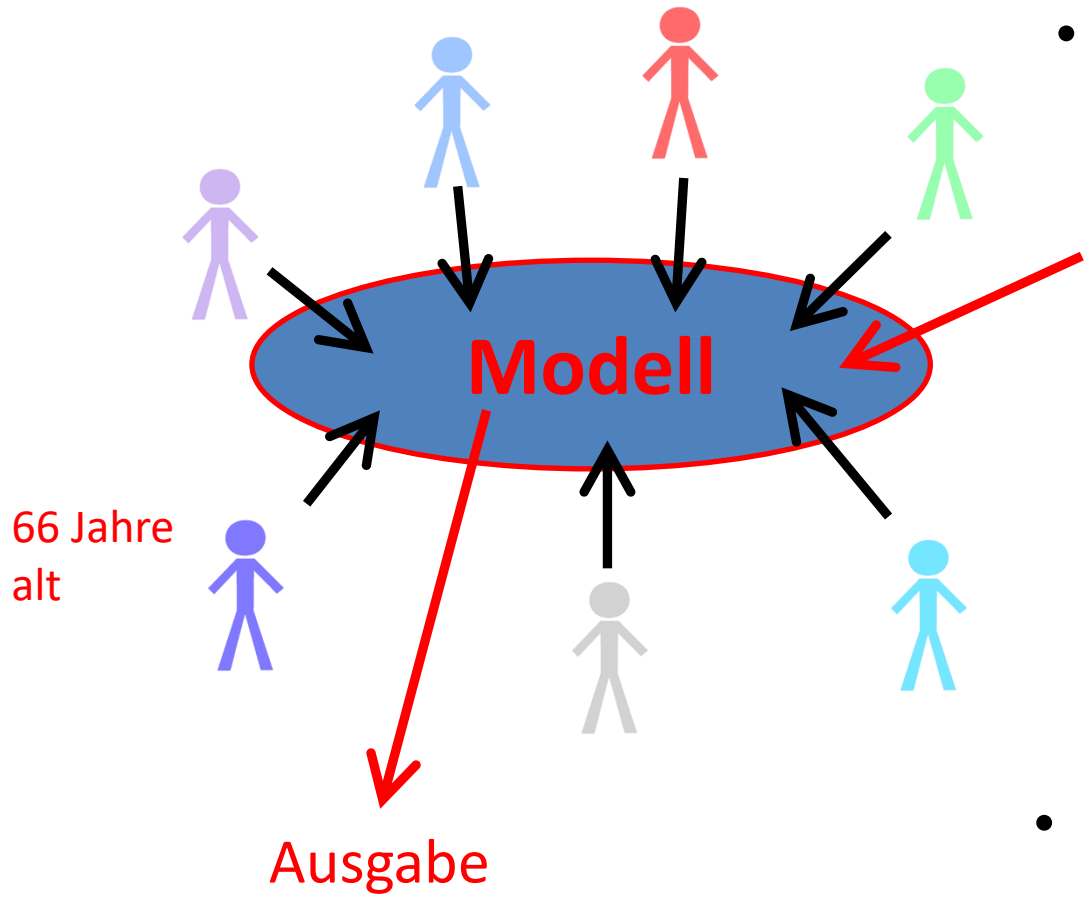
- Membership Inference: Ist in das Modell eingeflossen?



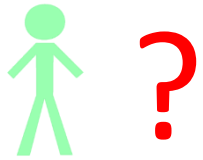
- Reconstruction: Rückschluss insb. aus Ausgaben und einigen Attributen von  auf weitere/alle ihre Attribute aus den Trainingsdaten
- Property Inference: Rückschluss auf Attribute von  , die nicht in den Trainingsdaten stecken

Problem 5: Abbilden von Zusammenhängen in einem Modell

- Klassifikation von „Privacy-Angriffen“ auf ML-Modelle nach Rigaki/Garcia (ACM Computing Surveys 2023)



- Membership Inference: Ist in das Modell eingeflossen?



- Reconstruction: Rückschluss insb. aus Ausgaben und einigen Attributen von  auf weitere/alle ihre Attribute aus den Trainingsdaten
- Property Inference: Rückschluss auf Attribute von  , die nicht in den Trainingsdaten stecken
- Model extraction: (Weitgehende) Rekonstruktion des Modells aus den Ausgaben

Problem 6: Nachvollziehbarkeit

- Einige Verfahren (z.B. Entscheidungsbäume): Im Prinzip leicht nachvollziehbar, nur ggf. zu umfangreich für schnellen Überblick
- Andere Verfahren (z.B. Neuronale Netze): In einigen Anwendungen drastisch höhere Leistungsfähigkeit, aber Nachvollziehbarkeit allenfalls für Experten
- Aktuelle Forschung arbeitet an der Erklärbarkeit der Ergebnisse Neuronaler Netze und anderer Verfahren
  - Fraglich, ob dies in allen Einzelfällen zufriedenstellend gelingt
- Vereinbarkeit mit Transparenz-Anforderungen des Datenschutzrechts? (→ Art. 5 Abs. 1 lit. a DSGVO)
- Frage: Vorrang für Transparenz oder für bestmögliche Ergebnisse?

Problem 7: Automatisierte Entscheidungen durch KI

- Automatisierte Entscheidungen auf Grundlage maschineller Lernverfahren
  - Im Datenschutzrecht: Aktuell geregelt in Art. 22 DSGVO – „Verbot automatisierter Einzelentscheidungen“

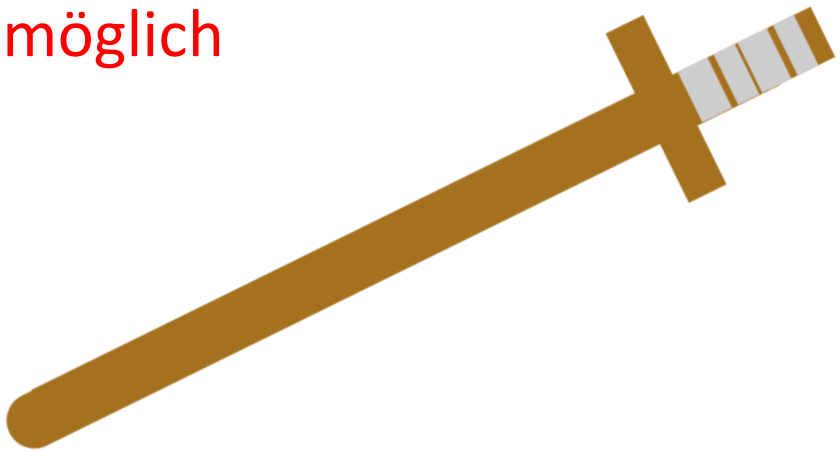
- Zulässigkeit automatisierter Einzelentscheidungen nach Art. 22 Abs. 2, wenn
  - sie „für den Abschluss oder die Erfüllung eines Vertrags zwischen der betroffenen Person und dem Verantwortlichen erforderlich ist“
  - der Betroffene ausdrücklich eingewilligt hat
  - (ggf. aufgrund Rechtsvorschrift)

Problem 6: Automatisierte Entscheidungen durch KI

- Automatisierte Entscheidungen auf Grundlage maschineller Lernverfahren
  - Im Fall der Zulässigkeit: Schutzmaßnahmen, insbesondere „mindestens das Recht auf Erwirkung des Eingreifens einer Person seitens des Verantwortlichen, auf Darlegung des eigenen Standpunkts und auf Anfechtung der Entscheidung“
  - Im Ergebnis also: Entweder die Automatisierung dient nur der Entscheidungsvorbereitung (→ Entscheidung durch den Menschen), oder sie kann durch einen Menschen nachgeprüft werden

- De facto: Reines „Abnicken“ von Entscheidungen möglich

→ Art. 22 als „stumpfes Schwert“?



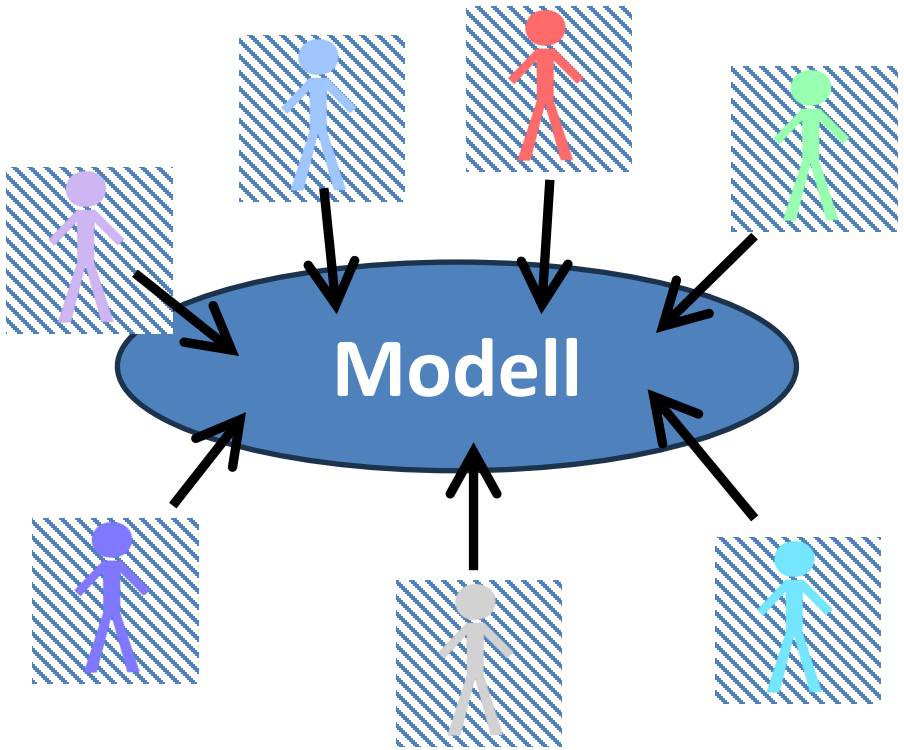
https://openclipart.org/detail/288547/wooden-sword

Problem 6: Automatisierte Entscheidungen durch KI

- Bemerkungen
  - Weitere Herausforderungen für maschinelle Entscheidungen auch außerhalb des Datenschutzrechts (Persönlichkeitsschutz nach § 823 I BGB, Diskriminierungsschutz aus AGG, …)
  - Aber: Auch potentielle Vorteile für die Betroffenen
    - Beispiel: Diskriminierung (!)
  - Technische Möglichkeit des Testens und zumindest Ansätze der Nachvollziehbarkeit (im Gegensatz zu menschlichen Entscheidungen)
  - Rechtsprobleme werden dadurch ggf. erst sichtbar

Privacy-Preserving Machine Learning

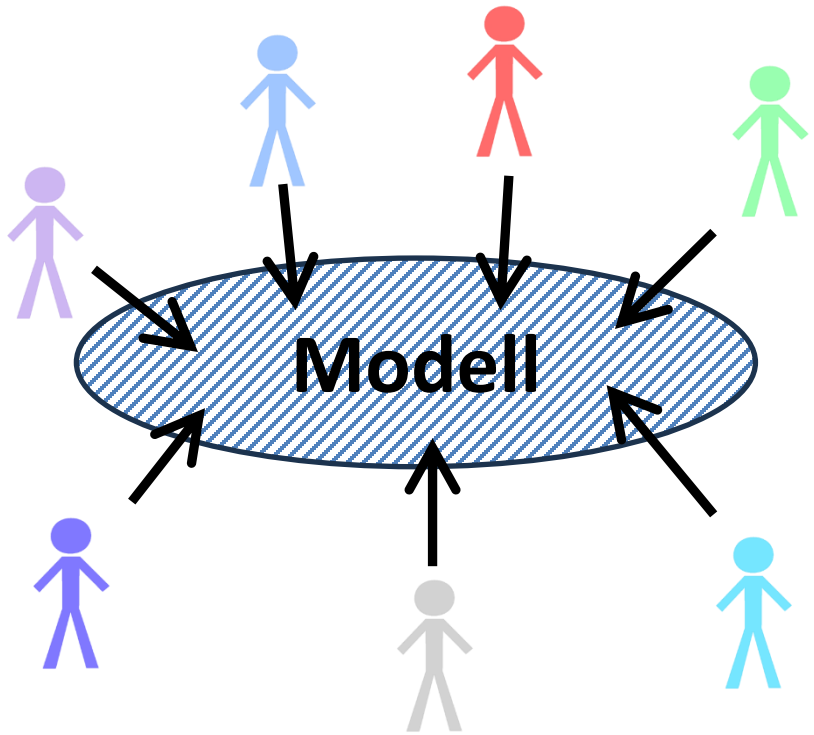
- Grundlegende Ansätze



- Anonymisierung der Trainingsdaten vor dem Training (z.B. Differential Privacy)
  - Zu beachten: Anonymisierung ist nicht trivial, auch bei Differential Privacy Abhängigkeit von gewählten Parametern

Privacy-Preserving Machine Learning

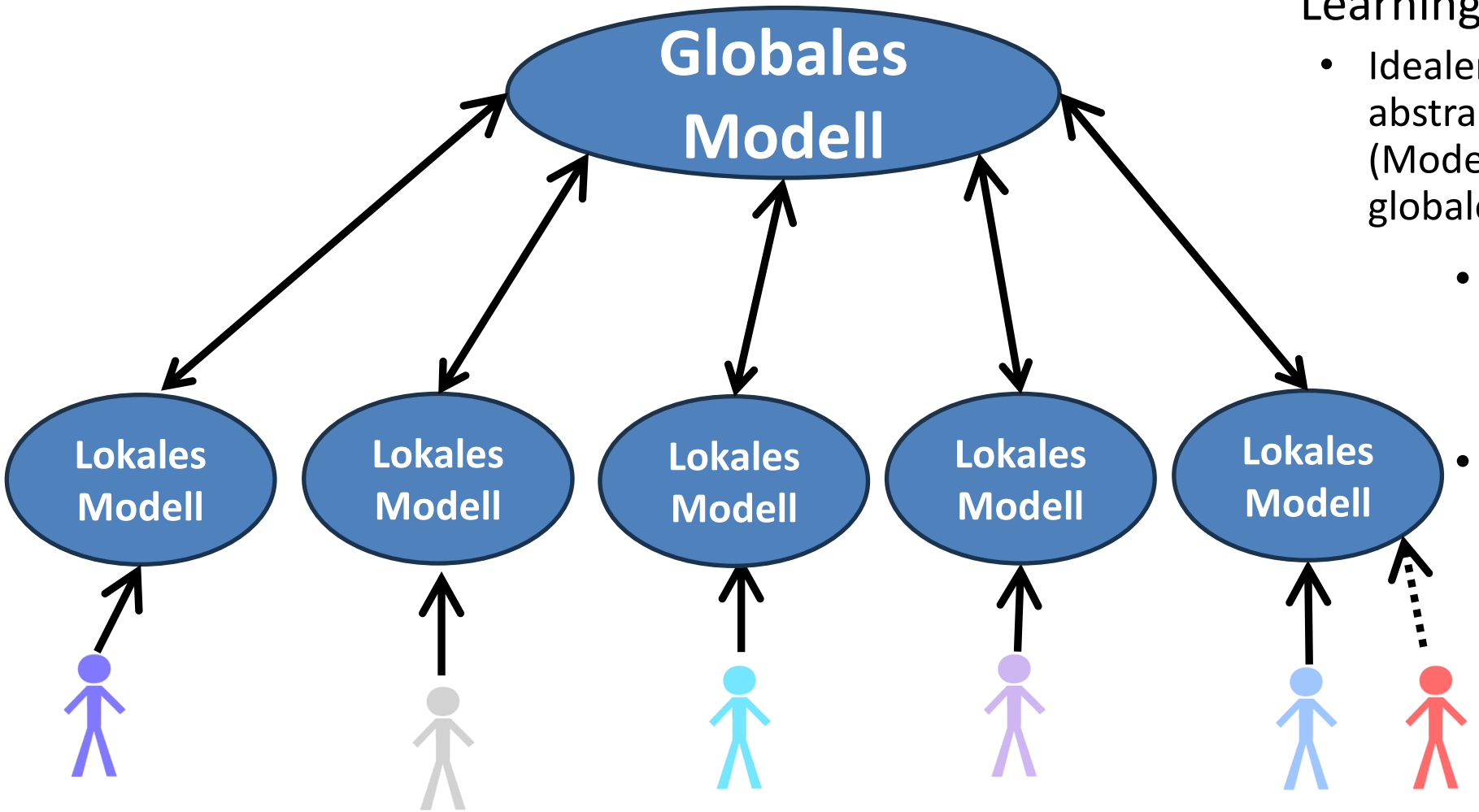
- Grundlegende Ansätze



- Kombination von Training und Anonymisierung

Privacy-Preserving Machine Learning

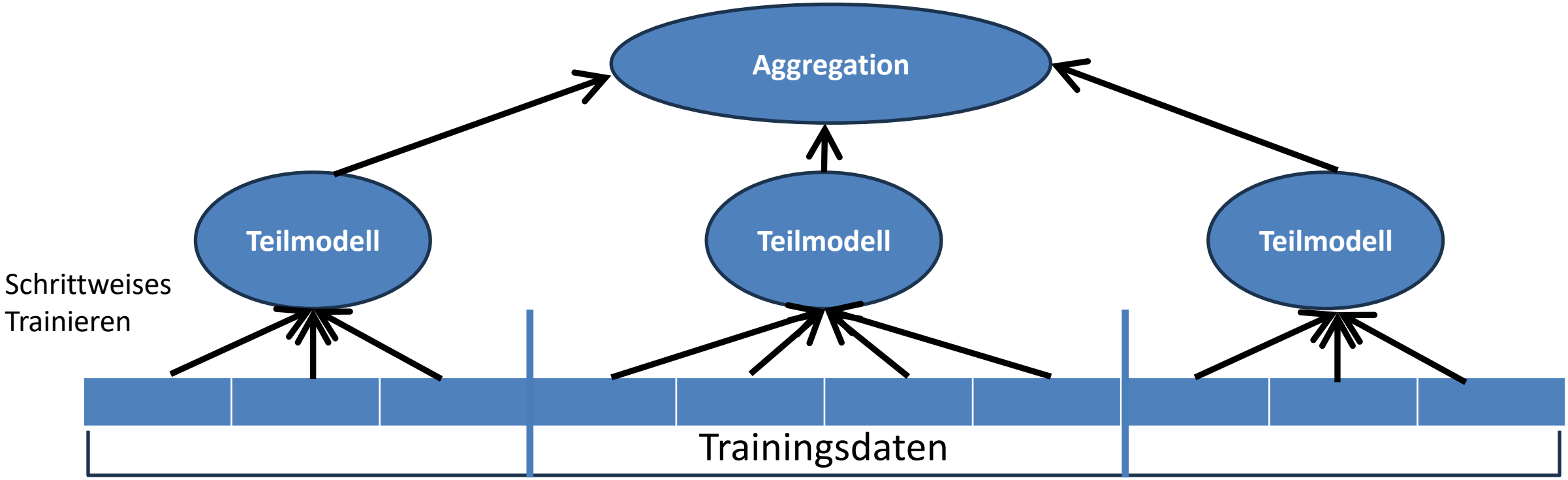
- Grundlegende Ansätze



- Föderiertes Lernen (Federated Learning)
  - Idealerweise: Nur bereits abstrahierte Daten (Modellparameter-Updates) an das globale Modell übermittelt
    - De facto: Auch hier ggf. noch Angriffe möglich (→Personenbezug?)
    - Eignung nur für bestimmte Anwendungsszenarien

Machine Unlearning

- Personenbezug eines Modells führt zu bisher nicht vertieftem Problem: Umsetzung des Rechts auf Löschung
  - Im Grundsatz: Neues Training ohne den zu löschenden Datensatz – allerdings: hohe Kosten, da ggf. tagelange Berechnungen notwendig
  - Lösungsansätze u.a. unter dem Stichwort „Machine Unlearning“ diskutiert
  - Beispiel SISA (Bourtoule et al., 42nd IEEE Symposium of Security and Privacy)



Anwendungsbeispiel IT-Sicherheit

- Maschinelles Lernen beispielsweise für Intrusion Detection im weitesten Sinne einsetzbar
  - Erkennen ungewöhnlichen Systemverhaltens („Anomalien“) und idealerweise Einordnung als Ergebnis eines Angriffs oder zufällige Fehlfunktion
  - Auch hier: Bislang eher wenige Datenquellen und eher „mäßig intelligente“ Systeme – aber: technischer Fortschritt ermöglicht Zusammenführung (potentiell immer mehr) verschiedener Datenquellen und deren kombinierte („ganzheitliche“) Auswertung

- Probleme aus Sicht des Datenschutzes?

Rechtlicher Ausblick – konkret

- Mögliche Erleichterungen des maschinellen Lernens im neuen europäischen Datenrecht

- Beispiel: Art. 54 der geplanten KI-Verordnung
  - Zweckänderung personenbezogener Daten zur Entwicklung und Erprobung „bestimmter innovativer KI-Systeme im Reallabor“ – u.a. für Zwecke der öffentlichen Sicherheit und öffentlichen Gesundheit

- Beispiel: Art. 34 Abs. 1 lit. g des geplanten European Health Data Space
  - Ausnahme von Art. 9 DSGVO für „Training, Erprobung und Bewertung von Algorithmen […] die zur öffentlichen Gesundheit oder sozialen Sicherheit beitragen“ → nur noch Rechtsgrundlage nach Art. 6 DSGVO notwendig

Rechtlicher Ausblick – abstrakt

- Gängiges Vorurteil: Informatik ist binär, Recht ist unscharf
- Aber: Anwendbarkeit des (Datenschutz)Rechts ist tatsächlich eine binäre Entscheidung
  - Verarbeitung personenbezogener Daten → Anwendbarkeit DSGVO
- Umgang mit „ein bisschen Personenbezug”?
  - Jemand errechnet eine Wahrscheinlichkeit von 17%, dass meine personenbezogenen Daten in ein Modell aus einer klinischen Studie eingeflossen sind
- Rechtlicher Umgang damit?

Bildquelle: https://thispersondoesnotexist.com

Fazit

- Hier nur ein Ausschnitt der Forschung aus dem Bereich der Künstlichen Intelligenz mit ihrem Bezug zum Datenschutz dargestellt

- Ähnlich zu bisherigen Forschungsergebnissen zur (De-)Anonymisierung:
  - Annahmen hinterfragen
  - Datenschutzrisiken frühzeitig problematisieren

- Interdisziplinäre Forschung weiterhin notwendig, um Ergebnisse
  - der KI-Forschung
  - der technischen Datenschutz-Forschung
  - der juristischen Datenschutz-Forschung

  zusammenzuführen

Fazit (2)

- Für dir Praxis anstehende Fragen
  - (Wie) geht Anonymisierung? Müssen Anforderungen an Anonymisierung neu definiert werden? Welcher Personenbezug steckt in Modellen?
  - Woher kommen Trainingsdaten – insbesondere angesichts des Zweckbindungsgrundsatzes? Führt ein hohes Datenschutzniveau zu Wettbewerbsnachteilen durch mangelnde Datenverfügbarkeit – in Forschung und Praxis? Und/oder schießt das neue Datenrecht über das Ziel hinaus?
  - Wie können Rahmenbedingungen auch für innovative KMU und Startups geschaffen werden, um Datenschutz zu gewährleisten, aber keine Rechtsabteilungen zu erfordern?
  - Passen KI und Datenschutz überhaupt zusammen?