

Genauigkeit, Robustheit und Cybersicherheit in der KI-VO



Christoph Sorge
Lehrstuhl für Rechtsinformatik
Universität des Saarlandes

Der Lehrstuhl in Kürze



Kooptierung
Assoziierung



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

Der Lehrstuhl: Einbettung und Kooperationen



Kooptierung
Assoziierung



Normtexte

Artikel 15 Abs. 1 KI-VO

Hochrisiko-KI-Systeme werden so konzipiert und entwickelt, dass sie ein angemessenes Maß an **Genauigkeit, Robustheit und Cybersicherheit** erreichen und in dieser Hinsicht während ihres gesamten Lebenszyklus beständig funktionieren.

Artikel 13 Abs. 3 lit. a sublit. ii KI-VO

Die Betriebsanleitungen enthalten [...] die Merkmale, Fähigkeiten und Leistungsgrenzen des Hochrisiko-KI-Systems, einschließlich [...] des Maßes an Genauigkeit – einschließlich diesbezüglicher Metriken –, Robustheit und Cybersicherheit gemäß Artikel 15, für das das Hochrisiko-KI-System getestet und validiert wurde und das zu erwarten ist, sowie aller bekannten und vorhersehbaren Umstände, die sich auf das erwartete Maß an Genauigkeit, Robustheit und Cybersicherheit auswirken können;

Genauigkeit („accuracy“)

- Begriff „Genauigkeit“ klingt einfach und wird in der Informatik auch (bei Klassifikationsverfahren) in einer einfachen Bedeutung verwendet – als **Anteil der richtig klassifizierten** an allen klassifizierten Elementen
- Aber: Bewertung eines Klassifikationsverfahrens (und eines KI-Systems im Allgemeinen) nicht allein mit dieser einfachen Metrik machbar
 - Beispiel: Vorhersage des Ergebnisses einer Verfassungsbeschwerde mit 98% „Genauigkeit“
- Zahlreiche **weitere Metriken gängig und sinnvoll** – auch die KI-VO geht (in Artikel 13 Abs. 3 lit. a sublit. ii) davon aus und meint daher mit „Genauigkeit“ offenbar nicht nur die einzelne, oft so bezeichnete Metrik

Evaluation maschineller Lernverfahren – einfacher Fall

- Annahme: Binäre Klassifikation (z.B. missbräuchliche oder nicht missbräuchliche Kreditkartentransaktion) – missbräuchlich = positiver Fall
- Unterscheidung in
 - true positive (tp, korrekt als **positiv = missbräuchlich** erkannt)
 - false positive (fp, **fälschlich** als **positiv = missbräuchlich** erkannt)
 - true negative (tn, korrekt als **negativ = unproblematisch** erkannt)
 - false negative (fn, **fälschlich** als **negativ = unproblematisch** erkannt)

Welcher Anteil wurde korrekt klassifiziert?

$$Genauigkeit = \frac{tp + tn}{tp + fp + fn + tn}$$

Welcher Anteil der echten X wurde als X klassifiziert?

$$Recall = \frac{tp}{tp + fn}$$

Welcher Anteil der als X klassifizierten Instanzen ist tatsächlich X?

$$Precision = \frac{tp}{tp + fp}$$

Evaluation maschineller Lernverfahren – einfacher Fall

	Missbrauch	Kein Missbrauch
Klassifiziert als Missbrauch	20 (tp)	80 (fp)
Klassifiziert als kein Missbrauch	20 (fn)	880 (tn)

Genauigkeit: $(tp+tn) / (tp+fp+fn+tn) = (20+880)/1000 = 900/1000 = 90\%$

Fehlerrate: 10%

Precision: $P = tp / (tp+fp) = 20 / (20+80) = 20/100 = 20\%$

Recall: $R = tp / (tp+fn) = 20 / (20+20) = 20/40 = 50\%$

Robustheit

- Hinweise zum Begriff der Robustheit in EG 27 (zum Grundsatz der „Technischen Robustheit und Sicherheit [safety]“) und insb. EG 75
- Demnach (EG 75 S. 2) wird Robustheit gesehen als Widerstandsfähigkeit „in Bezug auf schädliches oder anderweitig unerwünschtes Verhalten sein, das sich aus Einschränkungen innerhalb der Systeme oder der Umgebung, in der die Systeme betrieben werden, ergeben kann“

Robustheit in der technischen Literatur

- Begriffsverwendung dürfte der in der Softwaretechnik entsprechen, z.B.

„Robustness is the degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions.“

(Petke/Clark/Langdon, Software Robustness: A Survey, a Theory, and Prospects, in: Proceedings of the 29th ACM Joint Meeting, European Software Engineering Conference and Symposium on the Foundations of Software Engineering S. 1476.)

- “ML Model robustness denotes the capacity of a model to sustain stable predictive performance in the face of variations and changes in the input data.”

(Ben Braiek/Khomh, <https://arxiv.org/pdf/2404.00897>, in Anlehnung an Freiesleben/Grote, Synthese vol. 202, art. 109)

- Im Detail aber auch in der technischen Literatur keine scharfe Abgrenzung

Robustheit: Beispiel

- Trainingsdaten



Apfel



Apfel



Apfel



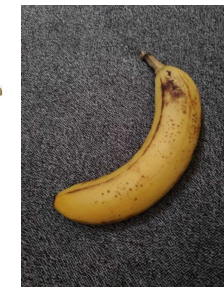
Apfel



Banane



Banane

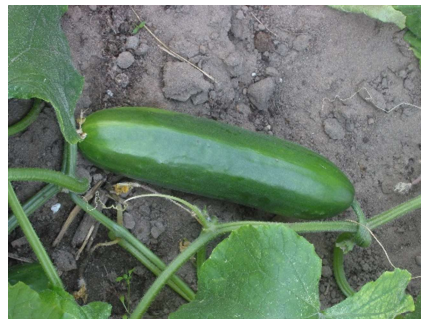


Banane



Banane

- Eingabe



→ Ausgabe: Klassifikation als Banane?

→ Unterbrechung des Betriebs bei Anomalien oder Betrieb außerhalb vorher festgelegter Parameter als möglicher Mechanismus für Robustheit (EG 75 S. 4 KI-VO)

Apfel 1: Justus Blümer, Apfel, CC BY 2.0, Link: <https://www.flickr.com/photos/justusbluemer/6042777587>
 Apfel 2: "© Superbass / CC-BY-SA-4.0 (via Wikimedia Commons)" Link: <https://commons.wikimedia.org/wiki/File:Apfel-Wellant.jpg>
 Apfel 3: "© Superbass / CC-BY-SA-4.0 (via Wikimedia Commons)" Link: <https://commons.wikimedia.org/wiki/File:Apfel-Jonagold.jpg>
 Apfel 4: "© Superbass / CC-BY-SA-4.0 (via Wikimedia Commons)" Link: <https://commons.wikimedia.org/wiki/File:Apfel-Berlepsch.jpg>
 Banane 1: Creative Commons Attribution-Share Alike 2.5 Generic license Link: https://commons.wikimedia.org/wiki/File:Banane_%C3%A0_45%C2%B0.jpg
 Banane 2: Filo gèn, Creative Commons Attribution-Share Alike 4.0 International, 3.0 Unported, 2.5 Generic, 2.0 Generic and 1.0 Generic license Link: https://commons.wikimedia.org/wiki/File:Banana_on_whitebackground.jpg

Banane 3: Gaurav Dhawj Khadka, CC BY-SA 4.0 Link: https://upload.wikimedia.org/wikipedia/commons/0/0e/Banana_9.jpg
 Banane 4: SMART Servier Medical Art, Creative Commons Attribution-Share Alike 3.0 Unported Link: https://commons.wikimedia.org/wiki/File:Banana_clipart.png
 Gurke: Rasbak, Creative Commons Attribution-Share Alike 3.0 Unported license Link: https://commons.wikimedia.org/wiki/File:Komkommer_Cucumis_sativus_'Melita'.jpg

Cybersicherheit

- Mögliche Angriffe auf KI-Systeme in zahllosen Variationen denkbar
 - „Klassisch“: Angriffe auf Infrastruktur (z.B. durch Schadsoftware, Phishing-Angriffe), dadurch Veränderung von Trainingsdaten, Modellen etc.
 - Aber auch KI-spezifische Angriffe (z.B. Membership Inference, in EG 76 S. 2 KI-VO schön übersetzt als „Inferenzangriffe auf Mitglieder Daten“)
- Begriff der Cybersicherheit offenbar auch in der KI-VO so breit zu verstehen, vgl. EG 76 KI-VO

Cybersicherheit

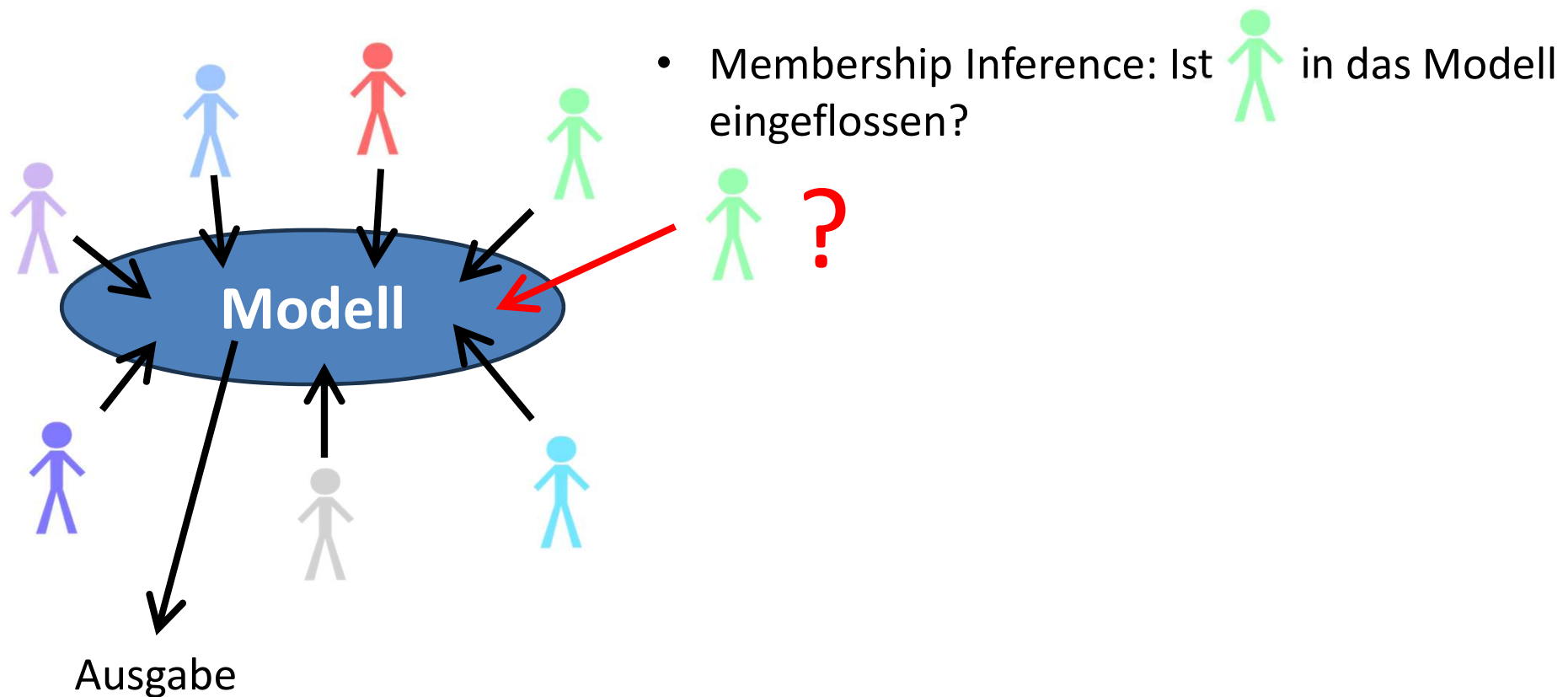
- Beispiele an der Schnittstelle Robustheit / Cybersicherheit
 - Vorgehen: Analyse eines gelernten Modells eines maschinellen Lernverfahrens und Konstruktion von Eingaben, bei denen das Lernverfahren ungewünschte Ergebnisse produziert – oder Manipulation des Modells selbst
 - Beispiel: Eykholt et al. (2018), „Robust Physical-World Attacks on Deep Learning Visual Classification“ – Aufkleber auf Verkehrsschildern wie im Beispiel rechts führen unter Laborbedingungen zu **100% Fehlklassifikation**
 - Beispiel: Bagdasaryan et al. (2018), „How To Backdoor Federated Learning“ – unter bestimmten Umständen: **100% Wahrscheinlichkeit**, dass eine ins Modell eingeführte Hintertür bei einer bestimmten Eingabe ein vom Angreifer gewähltes Ergebnis ausgibt



Bild: Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song. Robust Physical-World Attacks on Deep Learning Visual Classification, Computer Vision and Pattern Recognition (CVPR 2018).

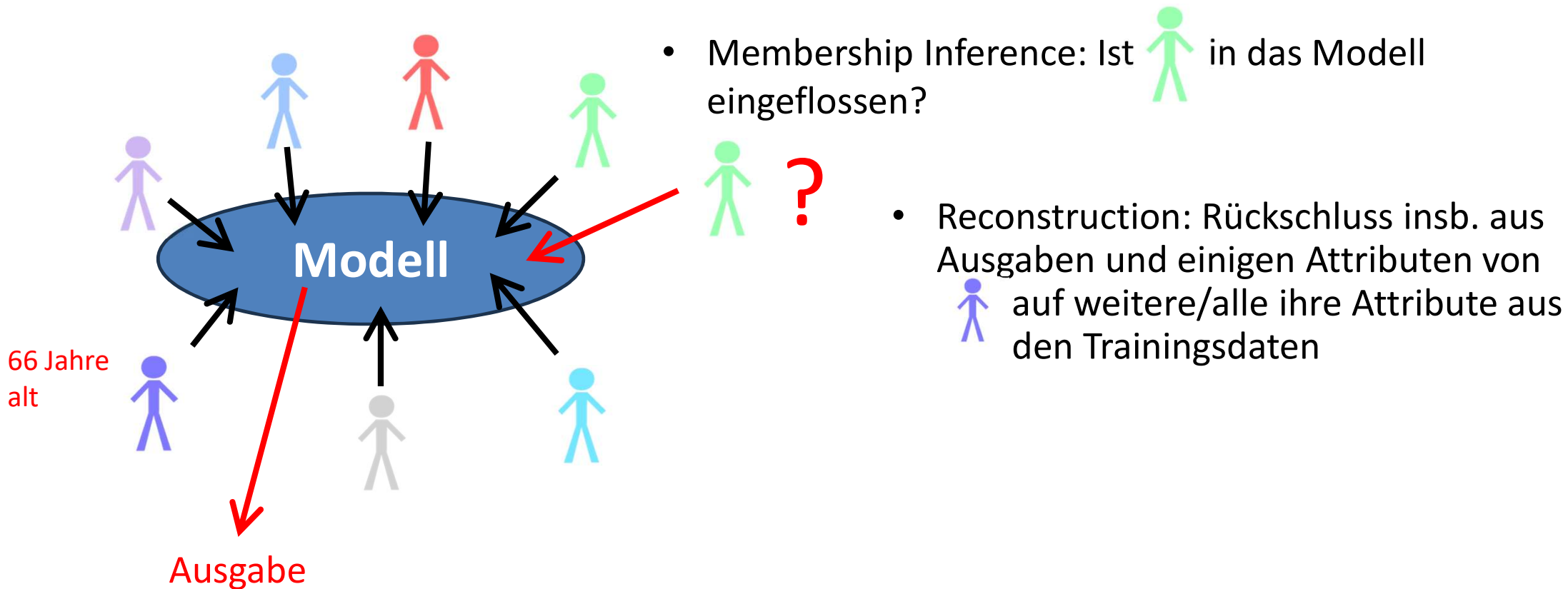
Privacy-Angriffe

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



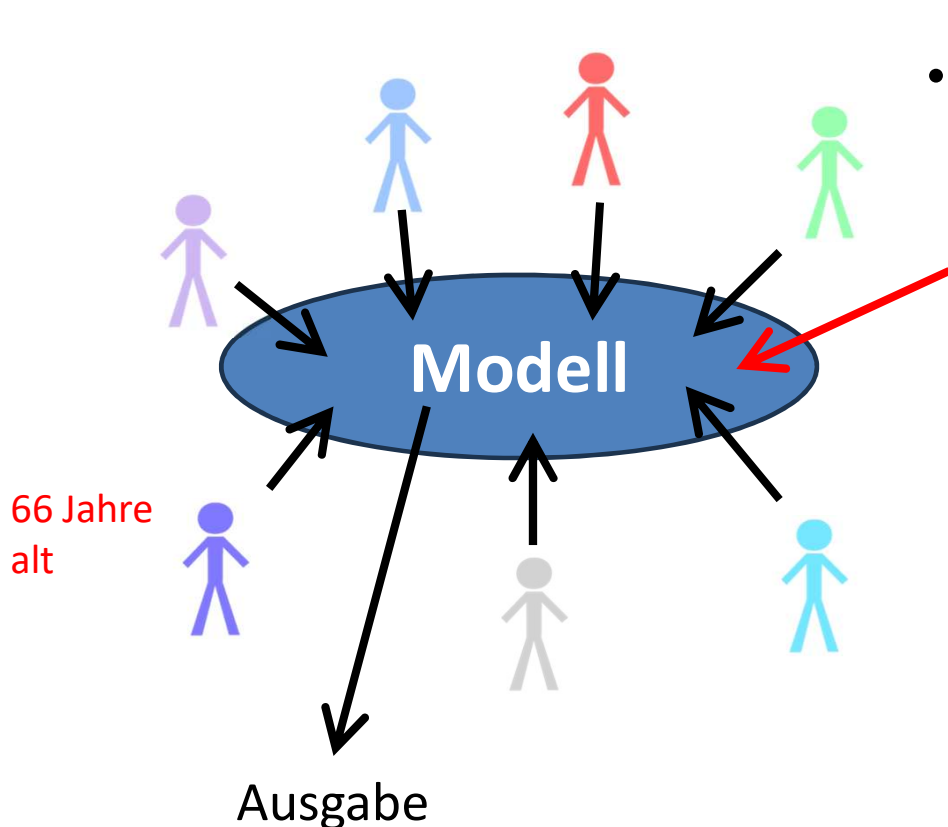
Privacy-Angriffe


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





Privacy-Angriffe

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



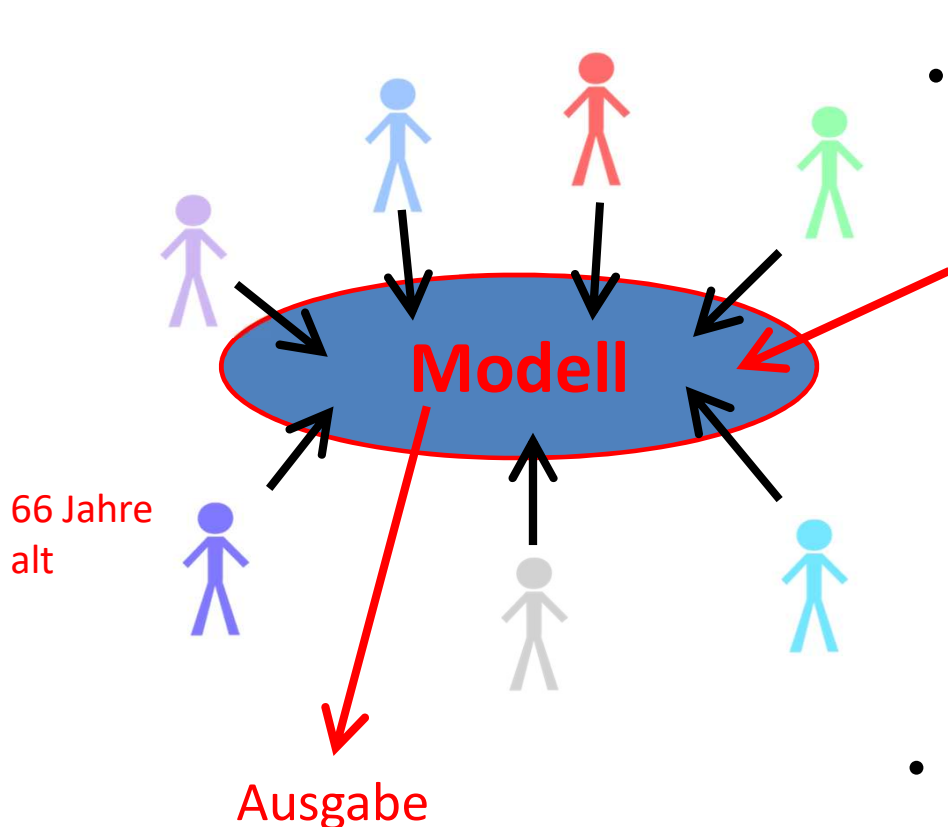
- Membership Inference: Ist  in das Modell eingeflossen?






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

Privacy-Angriffe

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



66 Jahre
alt

- 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

Fazit

- Begriffe Genauigkeit, Robustheit und Cybersicherheit finden direkte Anknüpfungspunkte in der wissenschaftlichen Literatur der Informatik und sind auch zu Recht miteinander verknüpft
 - Besondere Herausforderungen:
 - Ggf. unintuitive Evaluationsmetriken
 - Eigenschaften der Daten führen ggf. zu Limitierung erreichbarer Genauigkeit
 - Umfangreiche Bedrohungen der Cybersicherheit
 - Umfangreiche Möglichkeiten, die Robustheit von KI-Systemen bewusst oder unbewusst zu verletzen
- } keine allgemeingültigen Messwerte, bei denen ein Modell „gut“ oder „schlecht“ ist