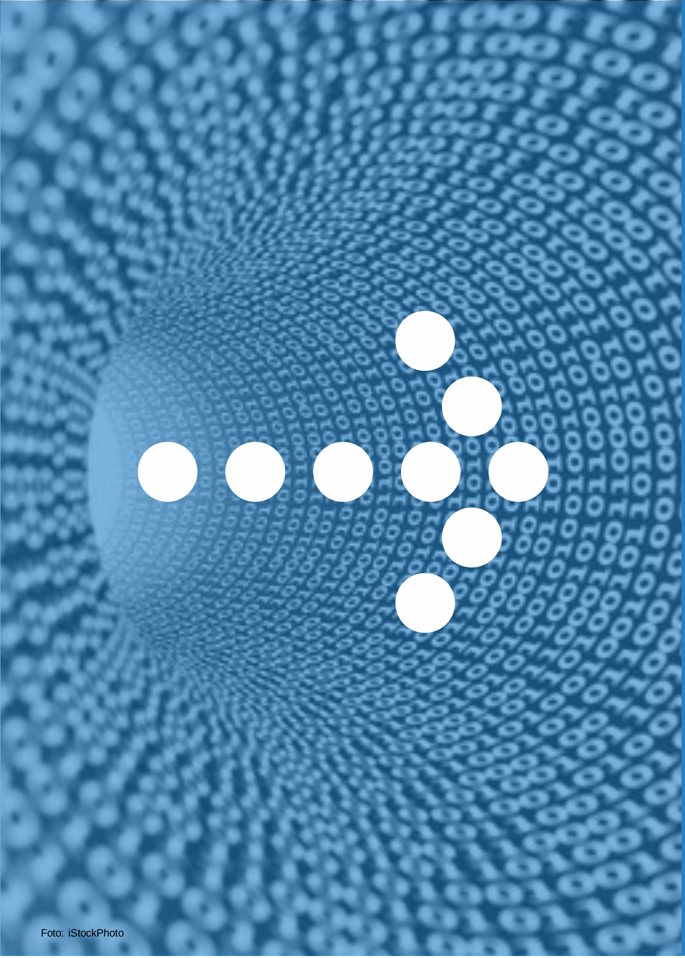


Foto: iStockPhoto

Säker delning av AI-modeller

Sinna Lindquist

2025 05 08

FOI

En presentation om

- Tvärvetenskap
- Teknik
- Juridik
- Hur förklarar man AI?

FOI

Att dela en AI-modell

Risker

- Antagonister som vill skada
- Modellen kan ”läcka” data
    - Avslöja arbetsmetoder
    - Läcka personuppgifter

FOI

Tekniken först

Test i flera steg

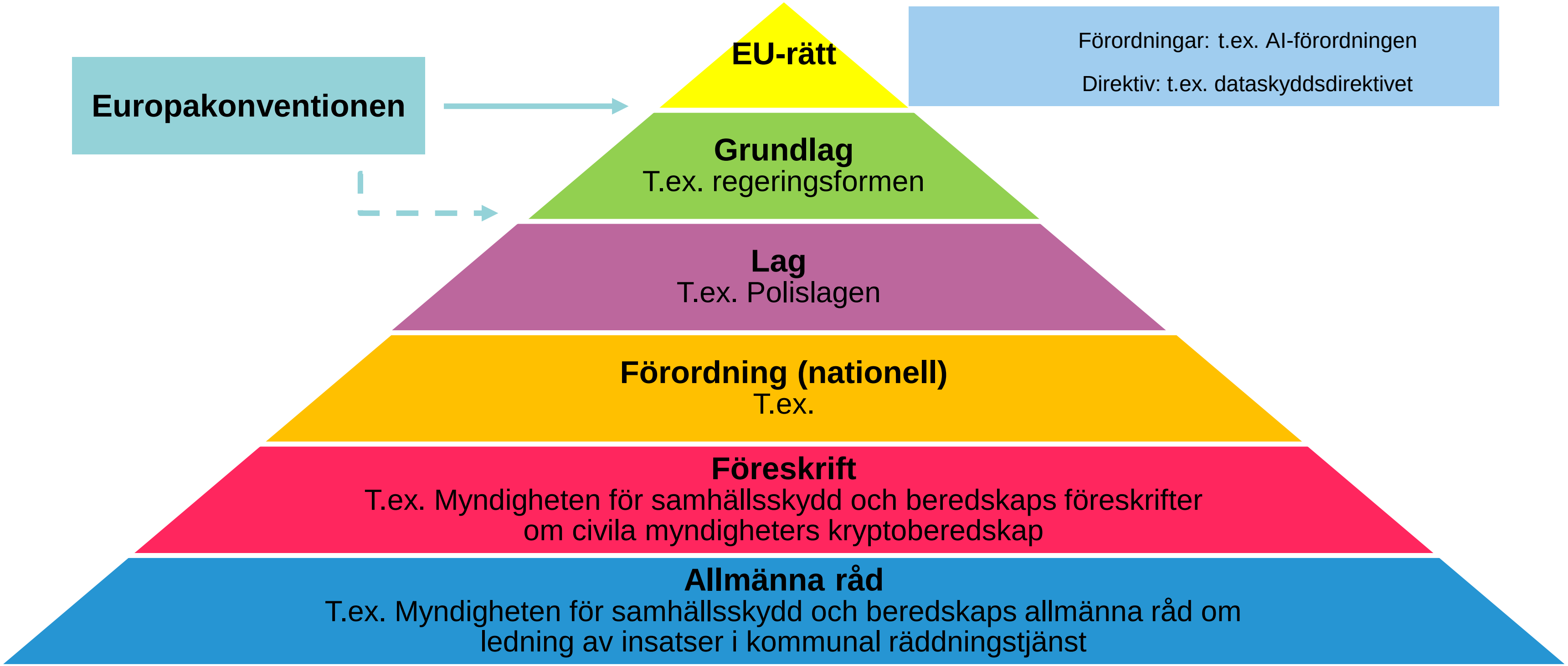
- Träna språkmodell
- Attackera (*Membership Inference Attack, MIA*)
- Skydda (*Differencial Privacy, DP*)

- Balansgång mellan skydd och prestanda

Typfall – Modelldelning från Polisen till SÄPO

- Modellen är tränad på öppen data
- Träningsdatamängden innehåller personuppgifter
- Modellen som tränas är en NER-modell
  (Named Entity Recognition)
- Modellens användningsområde är inte beskrivet
- Ett fåtal användare, inte för alla

Lagarnas inbördes rangordning



AI-förordningen

- Påminner om att alla berörda parter uppmanas att på lämpligt sätt beakta de 7 etiska principerna för utveckling av frivillig bästa praxis och standarder

- AI-system som klassas som hög risk ska särskilt beakta effekterna för den personliga integriteten, samt beakta barns särskilda rättighetsskydd

- Biometriska personuppgifter lyfts fram som särskilt problematiska, kan bl.a. leda till bias, särskild förbudsreglering i brottsbekämpande verksamhet, art. 5 g och h

- Rätten till integritet och skydd av personuppgifter måste garanteras under AI-systemets hela livscykel

- Förordningen påverkar inte förordning (EU) 2016/680 (genomförd gm brottsdatalagen), art. 2(7)

- Teknisk robusthet och säkerhet
- Integritet och dataförvaltning

Exempel på central lagstiftning

Nationell rätt

Polislagen

Lagen om säkerhetspolisens behandling av personuppgifter

Polisdatalagen

Brottsdatalagen (EU-direktiv/gäller ej nationell säkerhet)

EU-rätt (gäller ej nationell säkerhet)

GDPR (undantag i kap 1, art. 8, p. g)

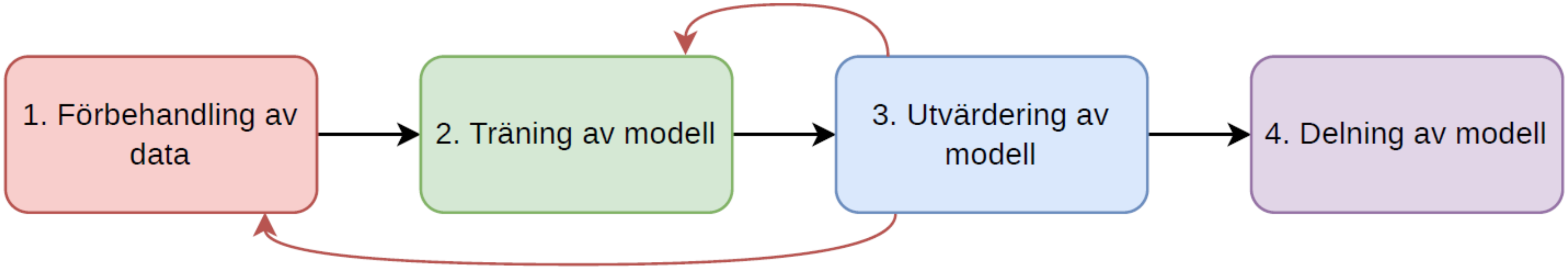
Dataskyddsdirektivet

AI-förordningen (delvis, fr.o.m. feb-25)

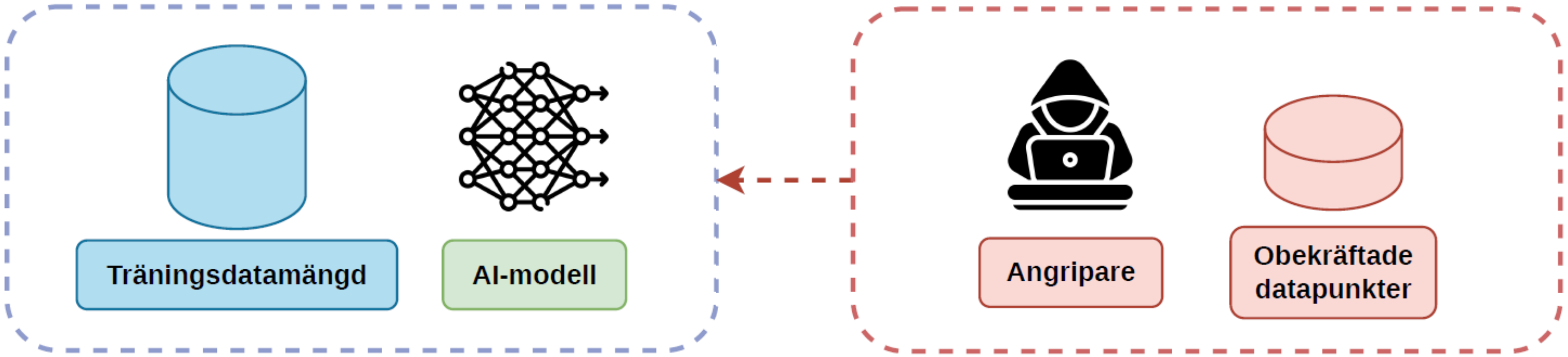
EU-domstolens rättspraxis, bl.a. *Tele2 Sverige AB & Watson och Digital Rights Ireland*

FOI

Modelldelningsprocess



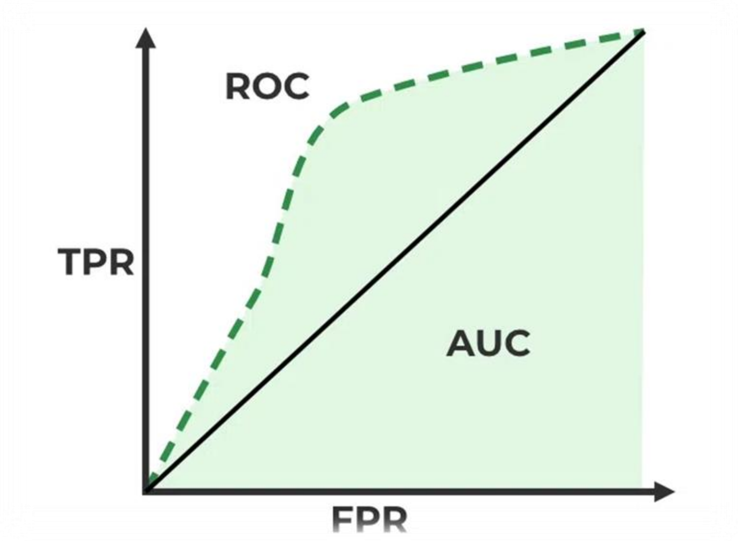
Hur fungerar en attack?



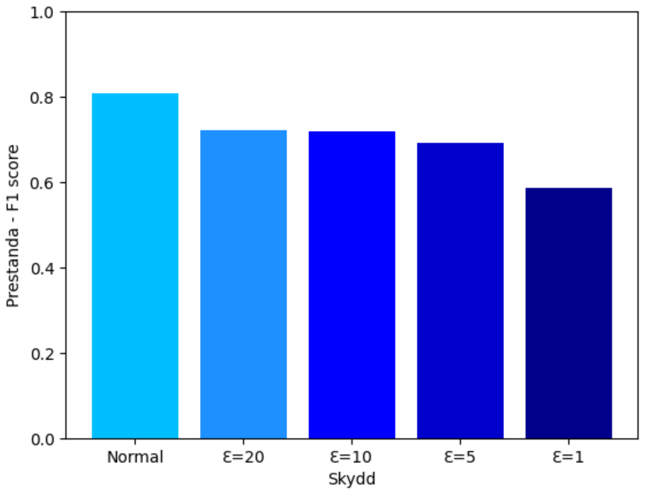
Vad är en läcka?

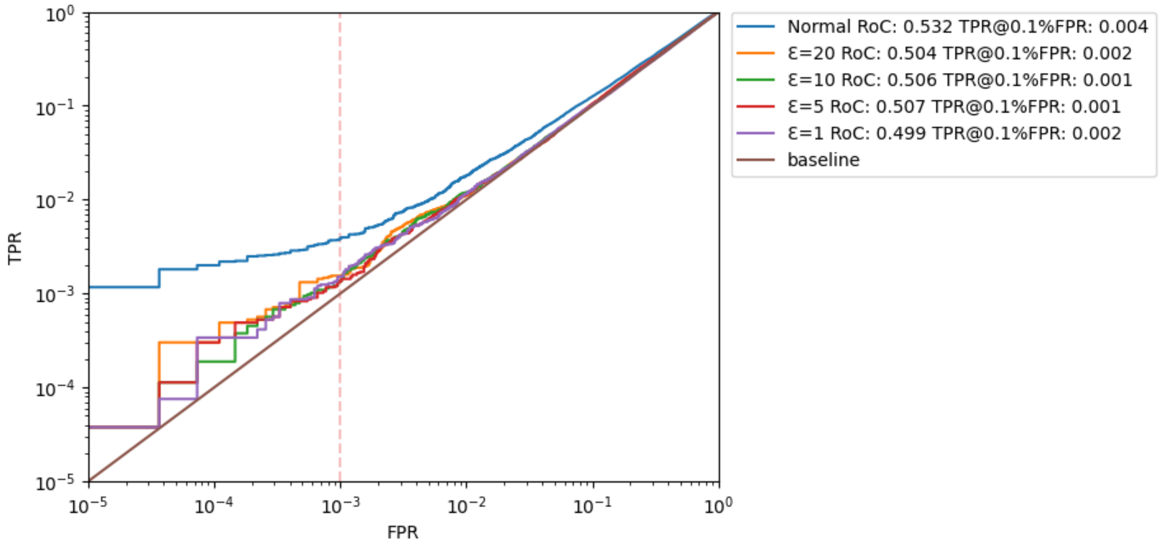
$$\mathrm{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\mathrm{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$



Resultat och Slutsatser





Differential Privacy

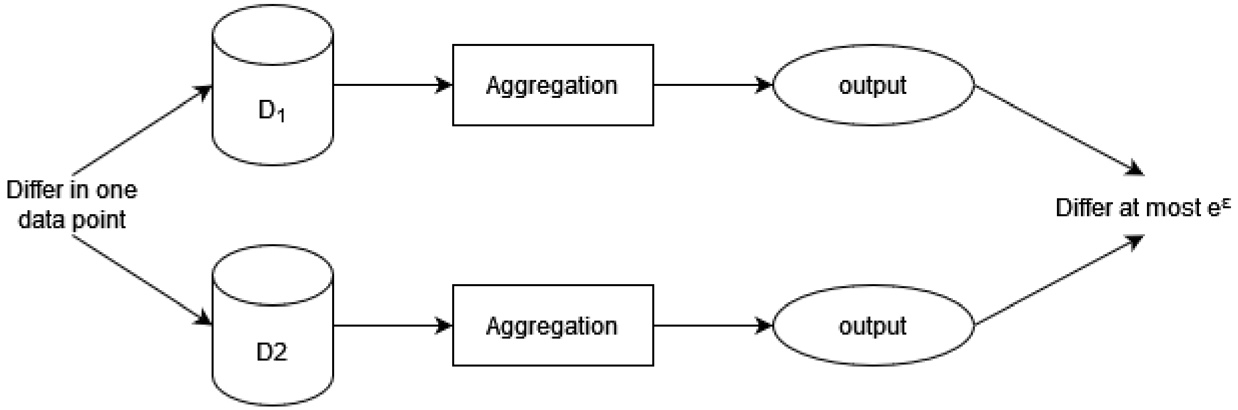
- Definition

$$P[M(D_1) \in S] \leq \exp(\epsilon) \cdot P[M(D_2) \in S] + \delta$$

Sannolikheten att garantin inte håller

Mått på garanterad säkerhet

- Garanti



Resultat av workshopen

”Kan innehålla spår av data”

- Användningen av modellen spelar roll
- Riskanalys – hur ska den göras?
- Förklaringsmodell
  - Vad måste förklaras?’
  - Vad måste fördjupas?

FOI

Nästa steg

- Förtydliga förklaringsmodellen
- Problematisera *Säker* ur ett juridiskt perspektiv
- Träna, attackera och skydda en NER-modell
- Verkligt fall: Transkriberingsmodell

Vi som jobbar med det här

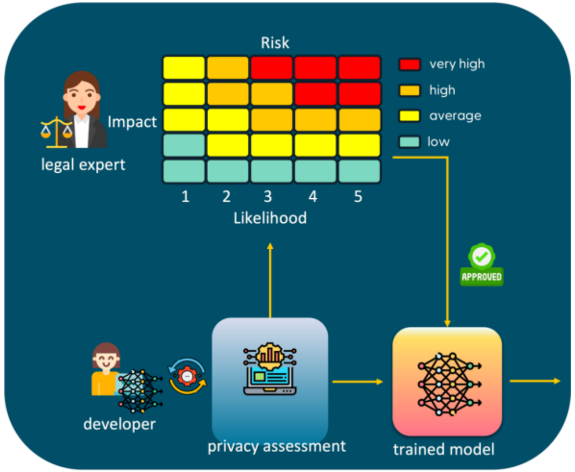
AI/ML: linus.kanestad, axel.borg, björn.pelzer

Juridik: anna.wetter.ryde

MDI: sinna.lindquist @foi.se

TACK!

Andra initiativ



- LeakPro (AI Sweden)
  – Öppet verktyg
  – Test av informationsläckage
  – Samla attackmetoder
- Next Genereation Machine Learning Model Exchange
  – NATO IST RTG