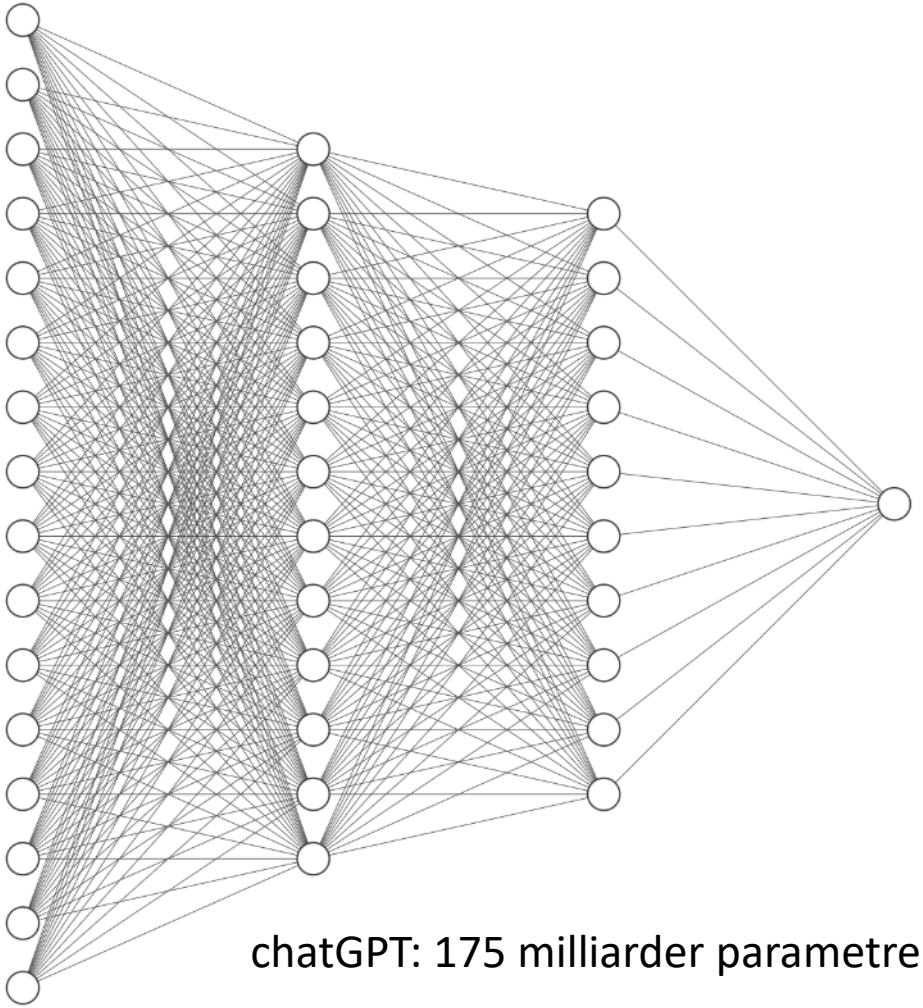
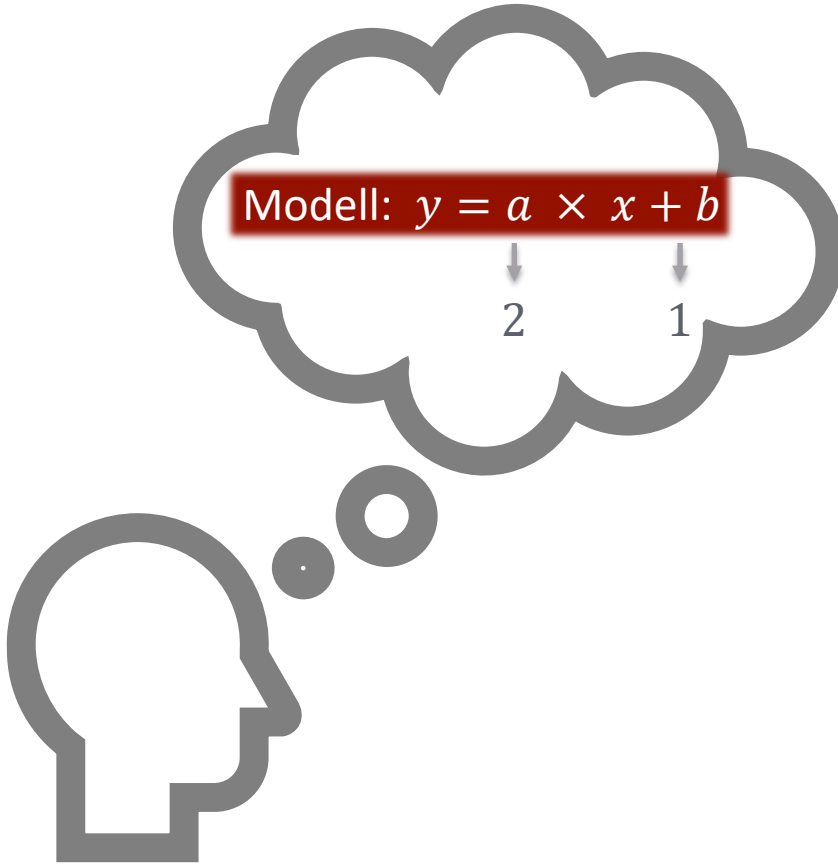
Rettferdighet i algoritmens tidsalder

Matematikk og etikk
IKT-rettskurset 24. mars 2023

Thale C. G. Gjerdsbakk, BULL
Robindra Prabhu, NAV IT

Hvordan lager man en maskinlæringsmodell?

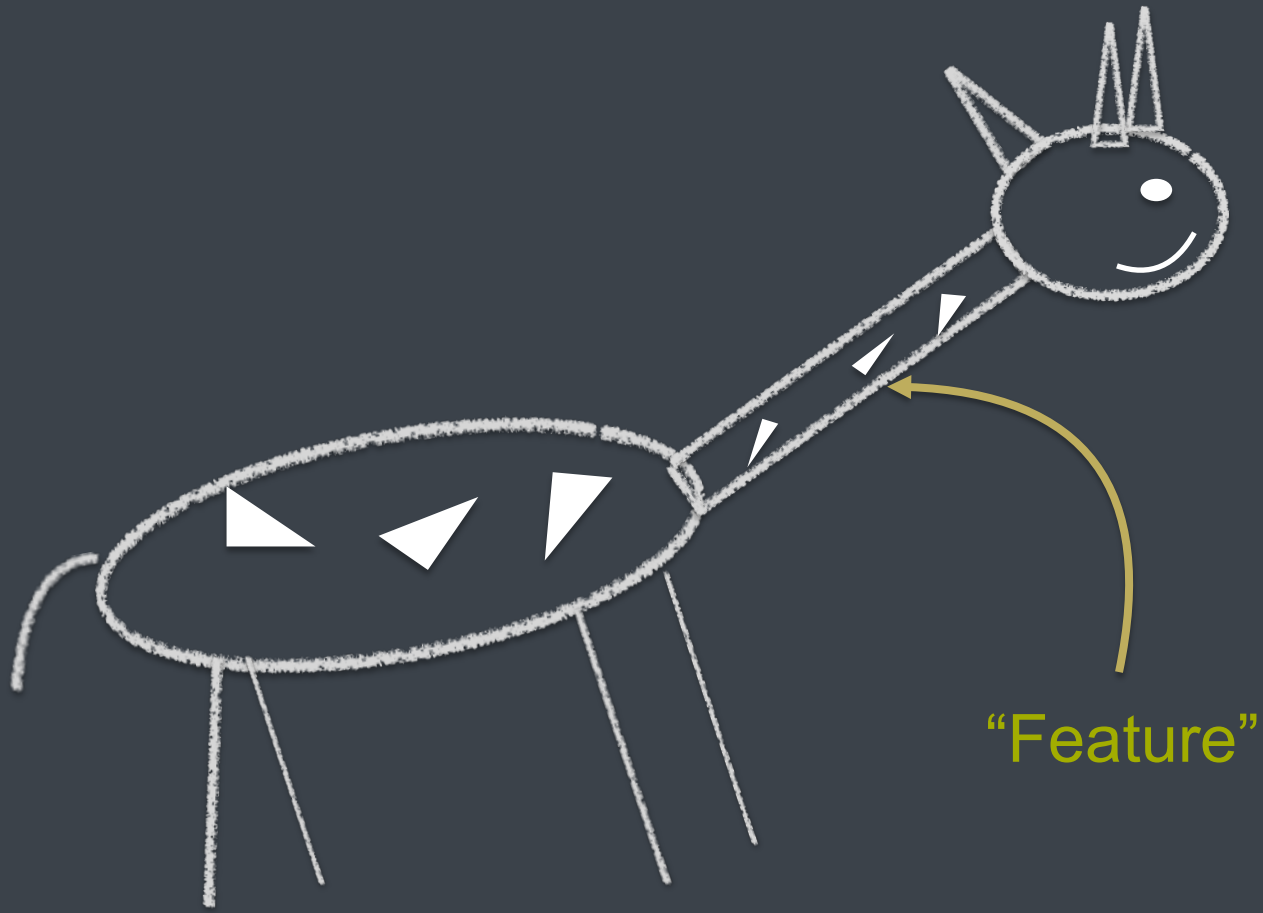




chatGPT: 175 milliarder parametre!

Input Layer ∈ $\mathbb{R}^{16}$ Hidden Layer ∈ $\mathbb{R}^{12}$ Hidden Layer ∈ $\mathbb{R}^{10}$ Output Layer ∈ $\mathbb{R}^{1}$

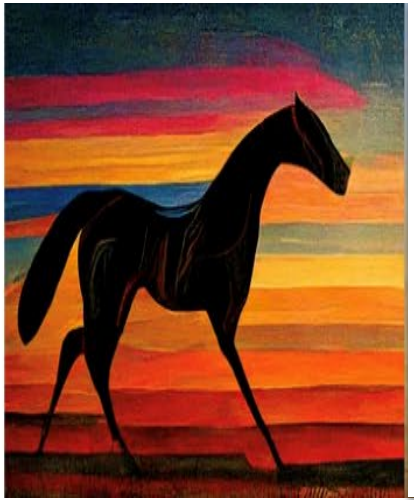
BULL


“Feature”

Kilde: https://shap.readthedocs.io/en/latest/#

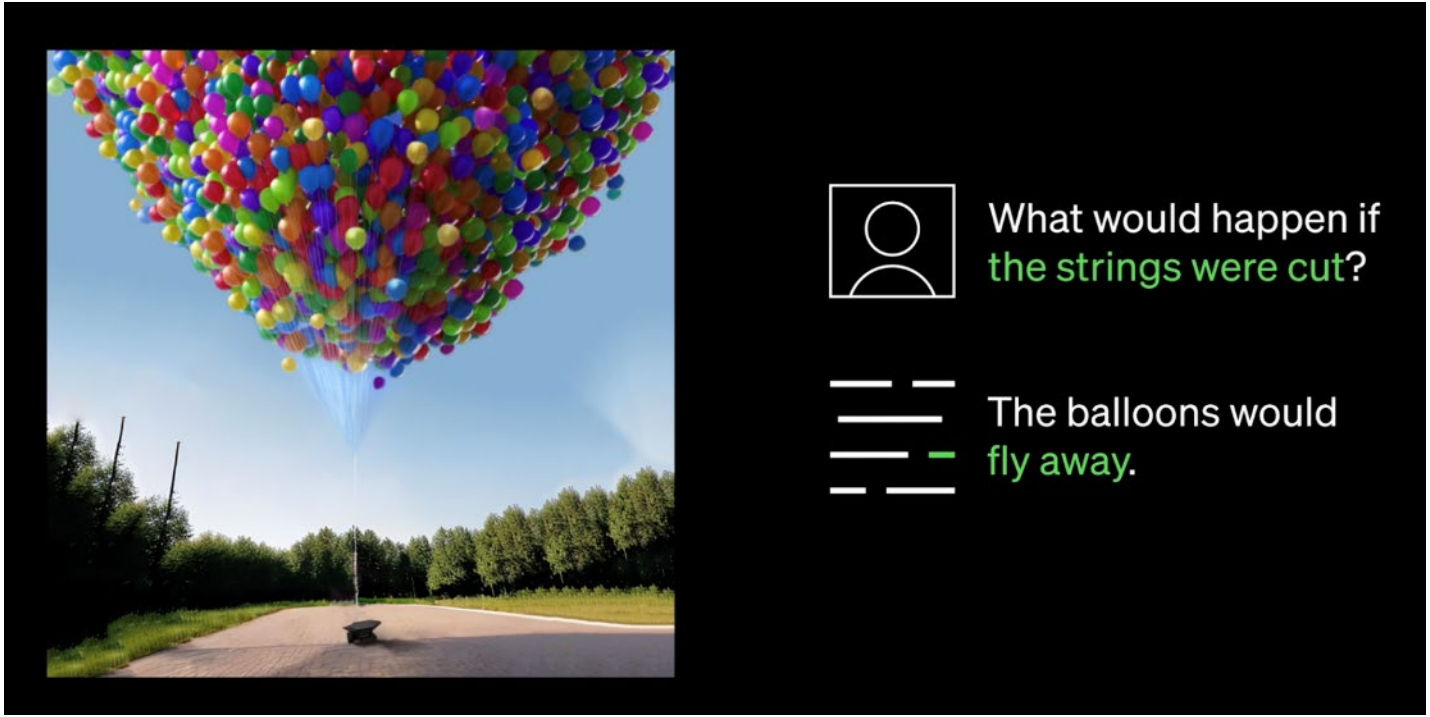
Hei, mitt navn er Ola Nordmann PER og mitt personnummer er 15044216652. FNR Jeg kommer til å ringe 01.01.2021 DTM for å snakke om min 50 prosent AMOUNT stilling. Du kan nå meg på 99 99 99 99 TLF hvis du vil ha tak i meg. Jeg bor i Trondheim LOC .

Generative modeller



Kilde: https://medium.com/mlearning-ai/an-advanced-guide-to-writing-prompts-for-midjourney-text-to-image-aa12a1e33b6

GPT-4



Kilde: https://openai.com/product/gpt-4

Maskinlæring har fått veldig gode vekstvilkår

1

Data, data og mere data!

2

Regnekraft

Hva kan gå galt?

MIT News
ON CAMPUS AND AROUND THE WORLD
SUBSCRIBE

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

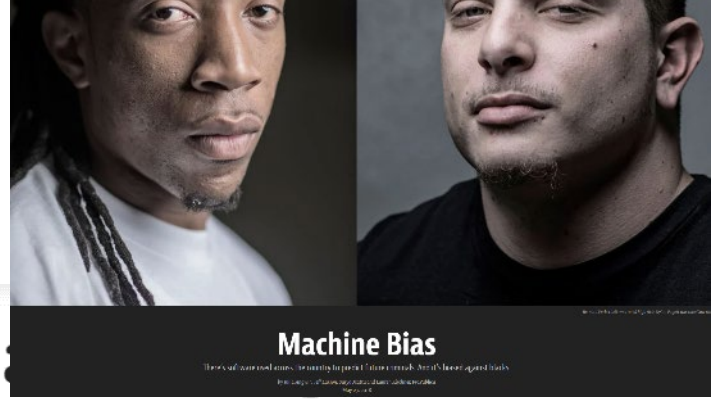
Watch Video

Larry Hardesty | MIT News Office
February 11, 2018

MOTHERBOARD
TECH BY VICE

The Story of How the Australian Government Screwed Its Most Vulnerable People

In 2016, the Australian government began automating welfare back payments. What happened next was an absolute disaster.



scandal serves as a
rope over risks of using
algorithms

Welfare surveillance system
human rights, Dutch court ru

England A-level downgrades hit pupils from disadvantaged areas hardest

Analysis also shows pupils at private schools benefited most from algorithm

A-level results - live updates

children of childcare benefit
n into care

Robot sette karakteren til Amanda - resultatet overraska

Etter mykje skepsis til skriverobotar i skulen, testar no lærarar ut robot til å vurdere oppgåver og setje karakterar.

▲ Students at Newham Collegiate sixth form centre in east London react after receiv
Photograph: Tolga Akmen/AFP/Getty Images

Pupils from disadvantaged backgrounds have been worst
controversial standardisation process used to award A-le
England this year, while pupils at private schools benefit

Private schools increased the proportion of students achi
A* and A - twice as much as pupils at comprehensives, of

tory Algorithm Wrongly
nds of Families of Fraud

Unfair, inaccurate and obscure

Grades constitute personal data, and pursuant to the GDPR, personal data needs to be accurate and processed fairly and transparently. The Norwegian Data Protection Authority considers that the IB grades are inaccurate because they do not reflect the students' individual academic level, which is the purpose of grading. Instead, the grades appear to be a prediction of what the students' exam grades might have been had the exams not been cancelled, but this is not possible to accurately predict.

The Norwegian Data Protection Authority also considers that it is unfair to base grades on how other students at the same school have performed previously.

- This could lead to discrimination as the grading model differentiates between students attending different schools. In addition, there has been a lack of transparency regarding the grading model, says Judin.

Gov
by p

When Algorithms Give Real Student Imaginary Grades

In-person final exams were canceled for thousands of student this spring, so computers stepped in — to disastrous effect.

BULL

Hva kan gå galt?

Hvordan blir data til?

Historisk skjevhet

Populasjon?

Representasjon/ utvalg

Forklaringsvariable?

Måleskjevhet

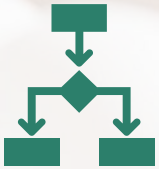
Tar modellen snarveier?

Aggregeringsskjevhet

Hvordan vurderer vi modellen?

Evaluering

Utrulling og bruk



Ankringsskjevhet

Bias og kvantitativ analyse innen velferd - opphav til skjevheter og relasjon til utfallsrettferdighet, Tidsskrift for velferdsforskning, **Vol 25 • Number 3 • 22 November 2022, Pages:** 1–24

Skjevheter virker uungåelige…

- så hva gjør vi da?

Rettferdighetsprinsippet

- Personvernets grunnlov
- Dynamisk
  - Tid
  - Sted
  - Kontekst
- Hva er rettferdighet?
  - Likebehandling
  - Ulik behandlig for likt resultat

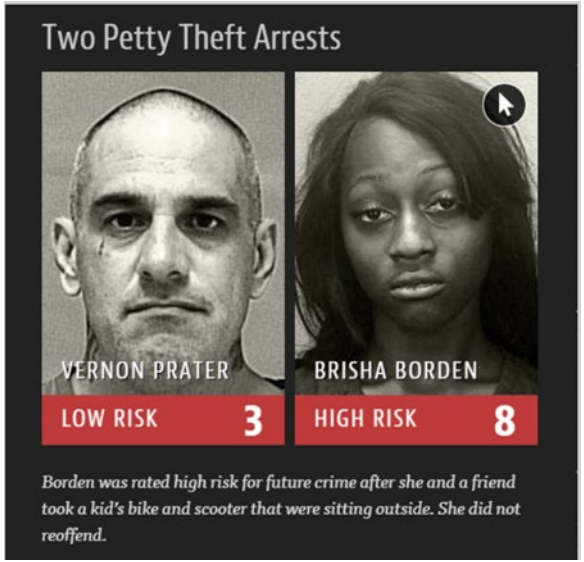
Rettferdighetsprinsippet

- Behandlingsansvarlig må definere

- Samfunnets oppfatning
- Rimelige forventninger
- Ujevn maktbalanse
- Større perspektiv
- Mulige negative konsekvenser

Risikoen for diskriminering

- Algoritmeskjevhet
- COMPAS
  - Risikovurderingsverktøy i domstoler
  - Forskjeller I feilvurderingene



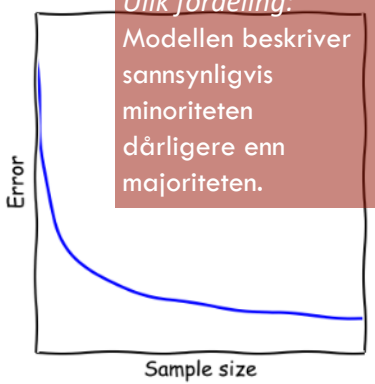
Kilde: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Prediction Fails Differently for Black Defendants

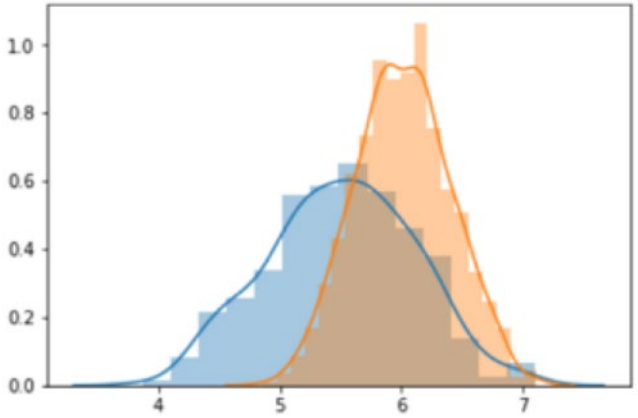
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

...selv uten bias

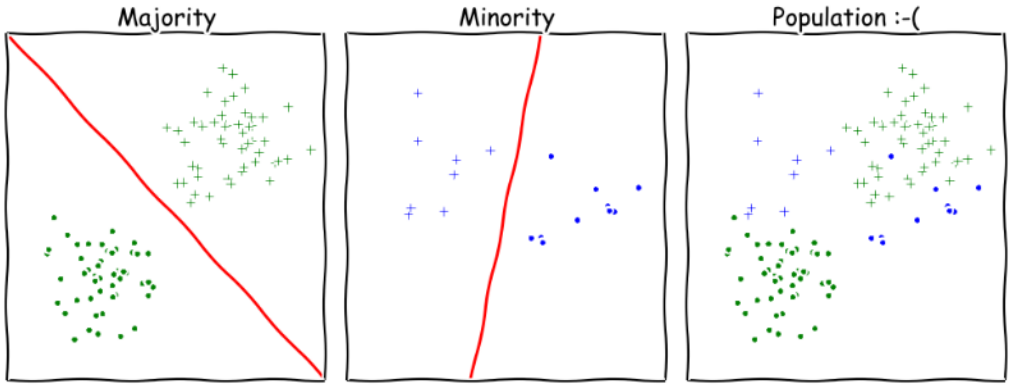


Ulik fordeling:
Modellen beskriver sannsynligvis minoriteten dårligere enn majoriteten.



Ulik oppførsel:
Mønstre som gjelder for majoriteten, må ikke gjelde for minoriteten.

Kompleksitet:
Kan være vanskelig å finne en kombo som gjelder for alle



https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

Bias

≠

Urettferdighet/
urettmessig
forskjellsbehandling

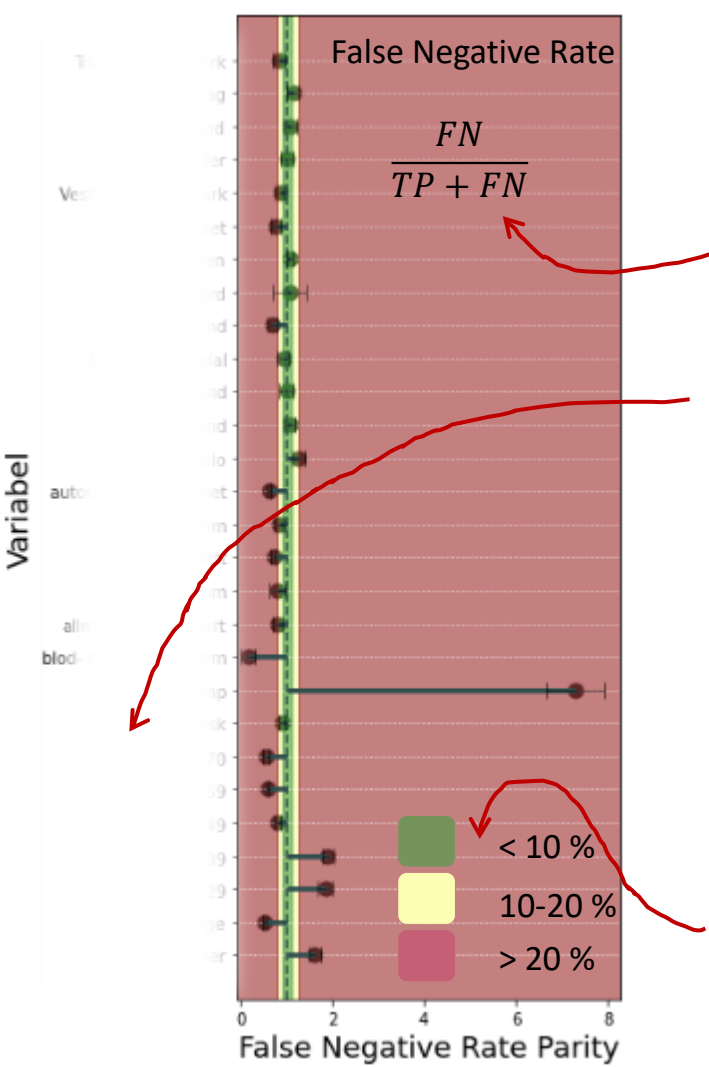
≠

Diskriminering

Dette er ulike begreper, men vi de ***påvirker*** hverandre!

Hvor mye urettferdighet tåler rettferdighetsprinsippet?

- Ulike rettferdighetsmål
- Mange rettferdighetsparametre
- Nøyaktig tallfesting
- Litt urettferdighet er alltid en risiko
- Akseptabel→uakseptabel
- Ikke en engangsøvelse



1

Hva sier loven? Hvordan operasjonalisere rettferdighets- og likebehandlingsprinsippene i modellen? Hvilke verdier bygger vi inn? Hva går vi glipp av?

2

Hvilke (diskriminerings-)grunnlag skal vi teste mot? Hva med informasjon vi ikke har?

Et teknisk plot…

fullt av normative and juridiske vurderinger

3

Hvor mye urettferdighet tåler vi før det blir «for mye»?

BULL

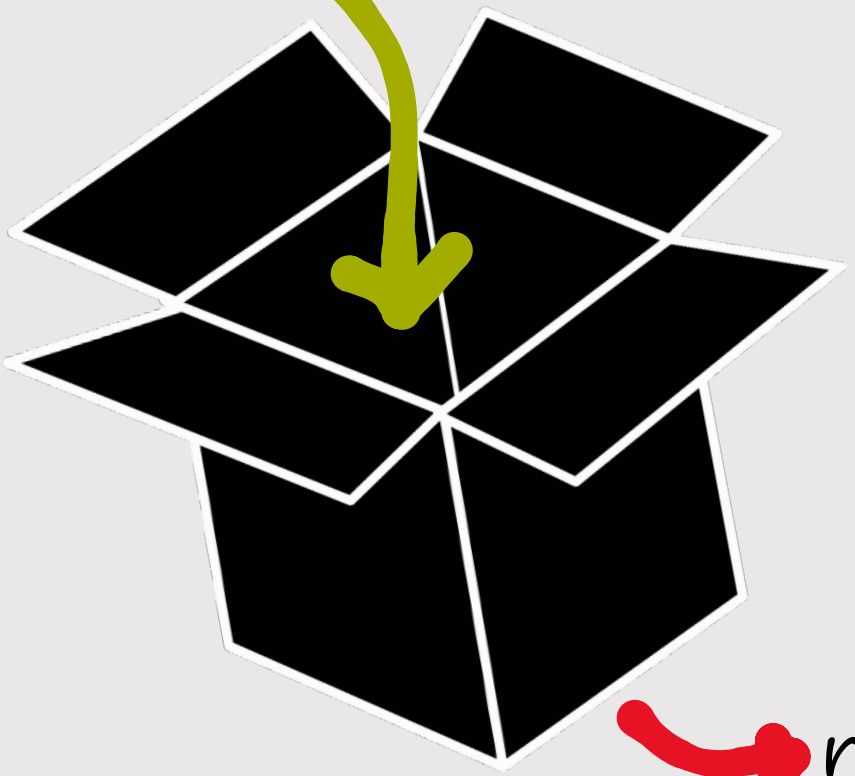
Psykologi og kunnskap

- Menneskelig innblanding
- Ankringseffekten
- Opplæring og informasjon
  - Bruksområde
  - Treffsikkerhet
  - Menneskelige svakheter
  - Nok ressurser

Åpenhet

- Forutsetning for rettferdighet
  - Hvor stor grad av åpenhet?
  - Må selv vurdere tiltak

data



resultat

Forskningen svarer på utfordringen
- metoder for å forklare har kommet på løpende bånd

Partial dependence plots

LIME

Shapley values

Integrated gradients

Counter-factuals

Feature importance

Anchors

dowitcher red-backed sandpiper

0 = Relationship

13 = Education-Num

4 = Marital Status

39 = Country

2174 = Capital Gain

7 = Workclass

0 = Capital Loss

1 = Occupation

40 = Hours per week

3 other features

−0.93

+0.49

−0.42

−0.34

−0.27

−0.26

−0.18

−0.13

−0.12

+0.01

−2.0 −1.5 −1.0 −0.5 0.0 0.5

−0.006 −0.004 −0.002 0.000 0.002 0.004 0.006

SHAP value

$E[f(X)] = 0.243$

BULL

modellutvikleren

juristen

System

Tolkbarhet (modellorientert)

Kan vi forstå og forutsi hvordan modellen vil oppføre seg i ulike situasjoner?

VS

Rettmessighet (justifiability)

Er et modellutfall eller en modelloppførsel saklig og riktig?

Individ

Forklarbarhet (utfallsorientert)

Hva vektla modellen for å komme frem til dette resultatet?

VS

Forklaring (funksjonell)

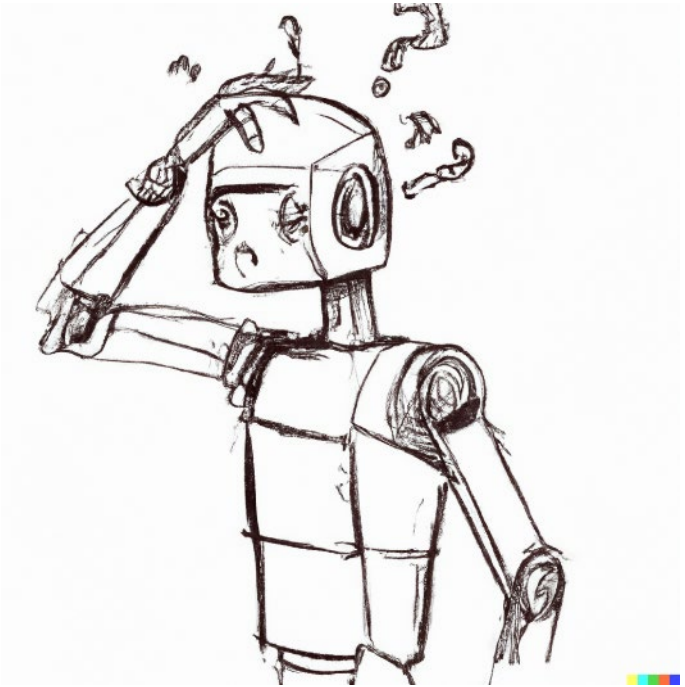
Har bruker nødvendig informasjon for å forstå, ivareta og utøve sine rettigheter / sin funksjon?

Åpenhet

- Formålet
- Forklarbarhet? Er det mulig?
- Forståelig
- Definere mottaker
- Grunnleggende informasjon
- Mulige konsekvenser

Hva nå?

- Hva er rettferdighet? Kontekstualiser!
  - I tråd med rimelige forventninger
  - Ujevn maktbalanse
  - Større perspektiv
- Finnes uakseptable skjevheter i KIen?
- Riktig kompetanse
- Opplæring av brukerne
- Informasjon
- Kontinuerlig kontroll



Hva kan dere be utviklere om?

1. Bias er vanskelig å unngå, men *utfallskjevheter* både kan og bør man vurdere

   @data scientist: anta at modellen er urettferdig og bevis heller det motsatte…

2. Fairness-metrikker er tretten på dusinet – ingen universell løsning. Å velge riktig mål er en ***etisk*** og ***juridisk*** øvelse som må gjøres i ***kontekst*** – ikke en teknisk øvelse som bør overlates utviklere!

   @data scientist: ikke gjør dette alene…

3. Fairness-metrikker er ingen silver-bullet, men kan likevel være nyttige verktøy for både utviklere og tilsyn.

   @produktutviklere/tilsyn: Se dere ikke blind på treffsikkerhet…be om rettferdighetsmål

nav

BULL

Takk for oss!

Thale C. G. Gjerdsbakk
tcgg@bull.no
95 45 82 67

Robindra Prabhu
robindra.prabhu@nav.no
95 05 80 12