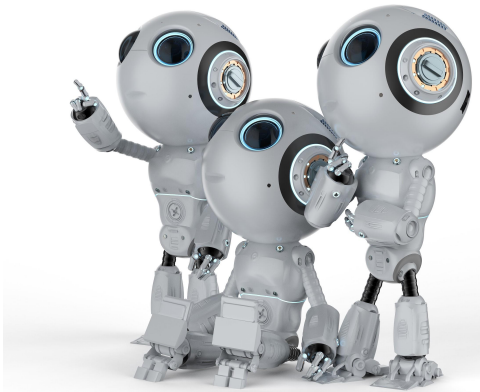
Ansvarlig kunstig intelligens

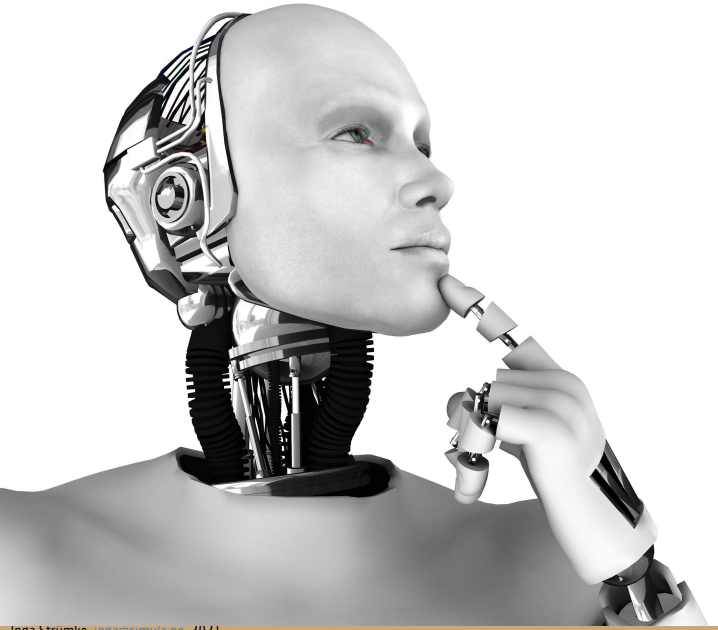
hvorfor er det så vanskelig?

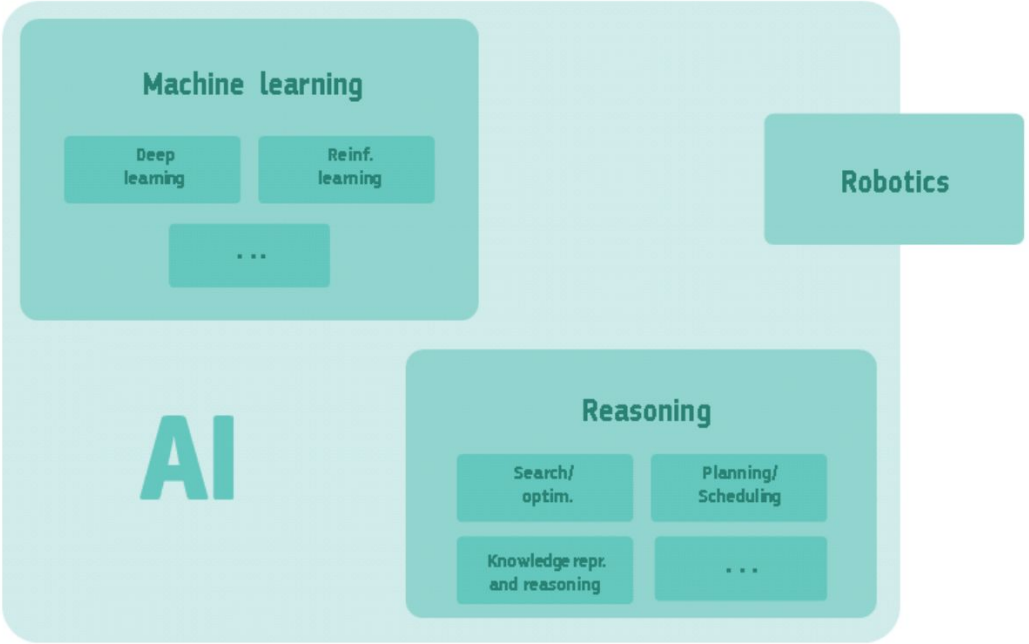


Inga Strümke, 2021
inga@simula.no

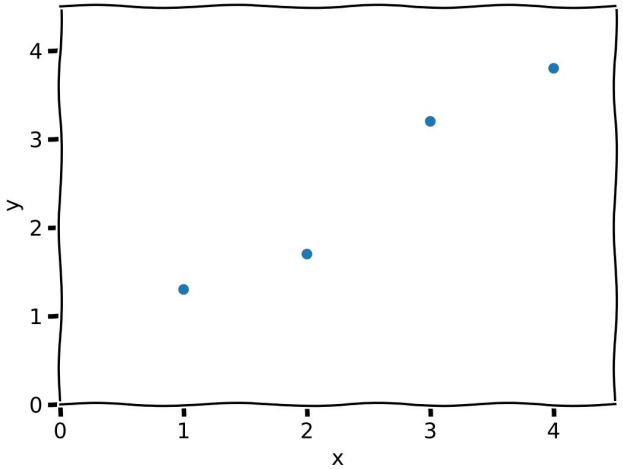
Hva er det?

1. Salgspitch
2. Det uoppnåelige
3. Tidenes samlebetegnelse



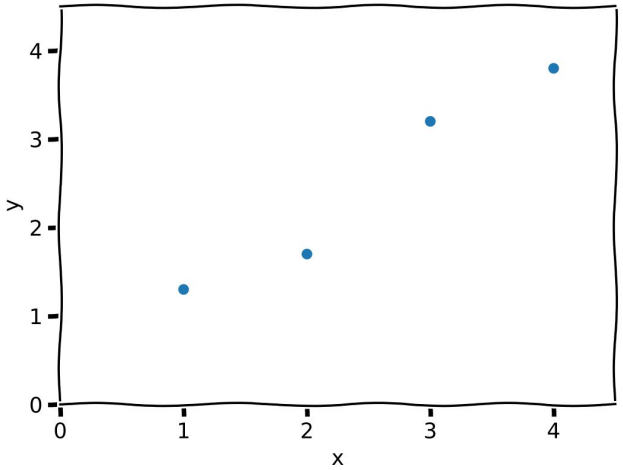


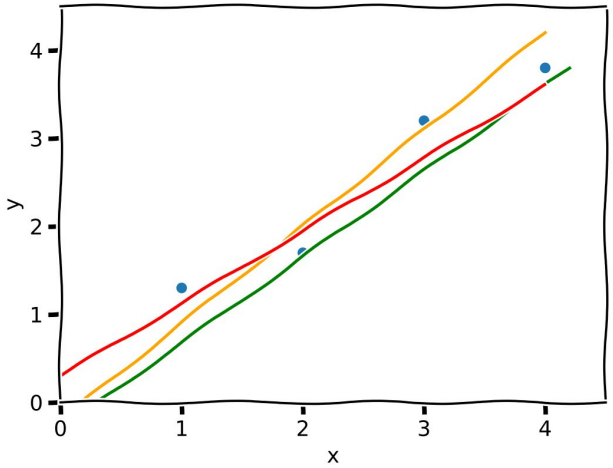
Maskinlæring på 30 sekunder



Inga Strümke, inga@simula.no, 2021

Maskinlæring på 30 sekunder





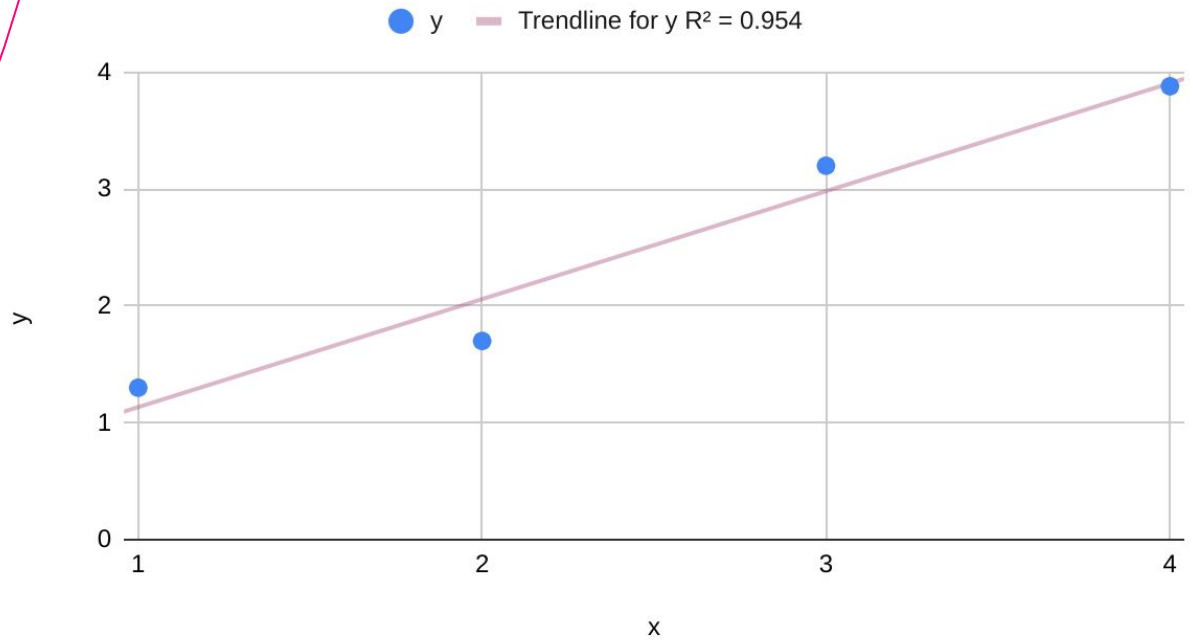
Inga Strümke, inga@simula.no, 2021

Maskinlæring … i Excel

1. Data
2. Mål
3. Regnekraft

x	y
1	1.3
2	1.7
3	3.2
4	3.88

$$R^2 = \left(\frac{1}{n-1} \frac{\sum (x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y} \right)^2$$



Inga Strümke, inga@simula.no, 2021

Maskin*læring*

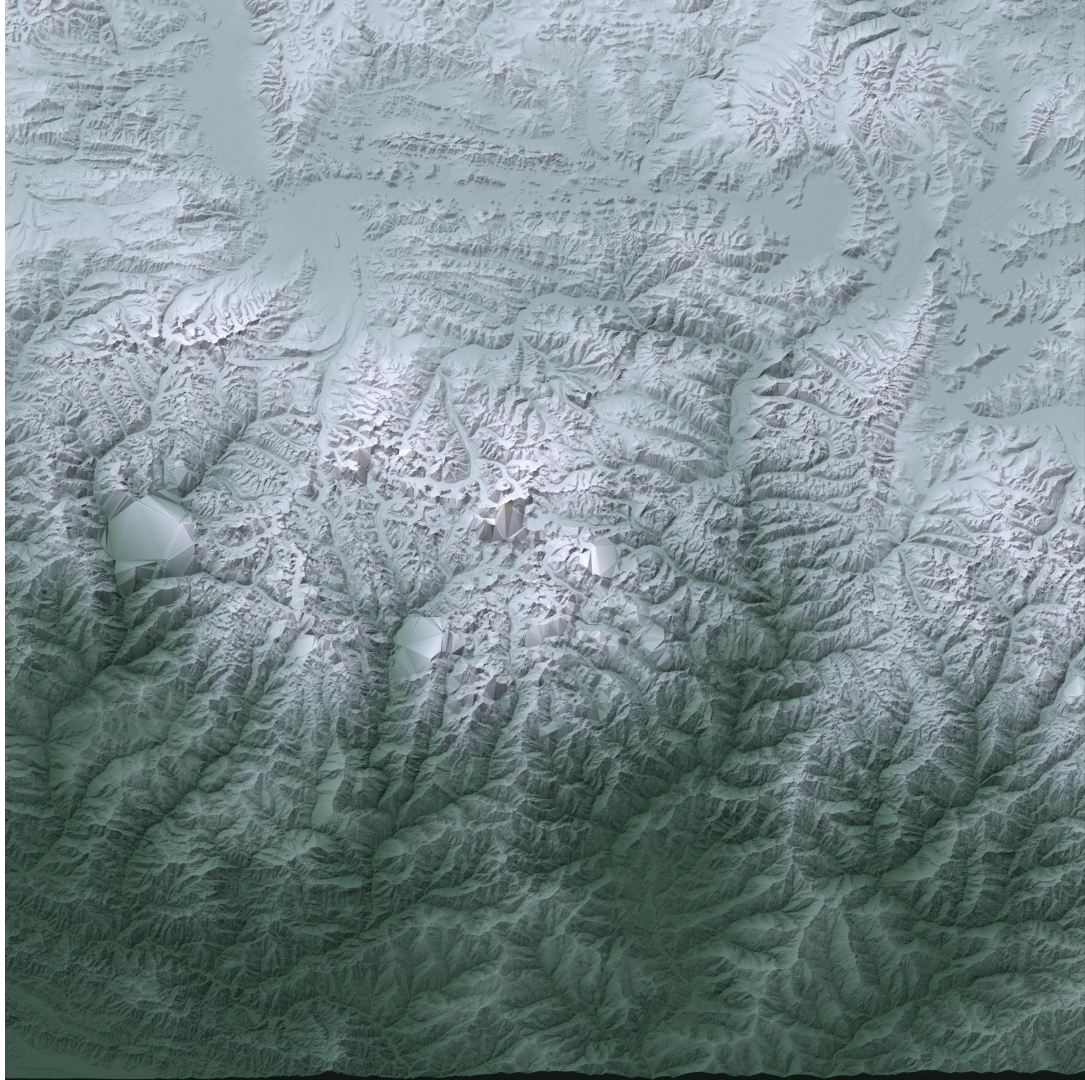
Bruke data til å **prøve seg frem** til den løsningen som oppnår målet best.

≈ Løpe rundt i Himalaya med bind for øynene og prøve å finne høyeste punkt.

Everest? Anapurna?

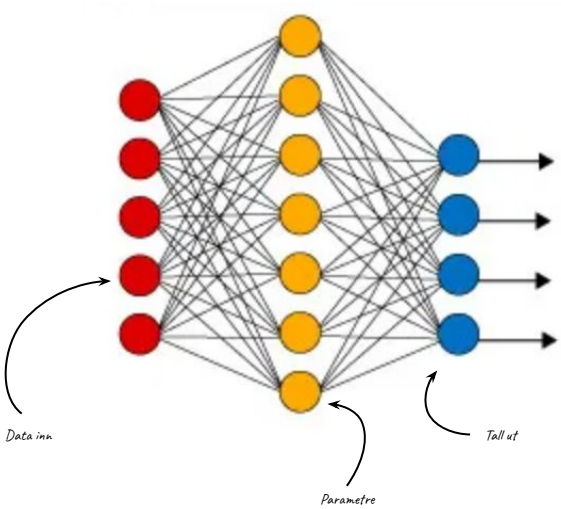
Finner en topp!

Mange mulige ruter i samme terreng!



Maskinlæring!!!

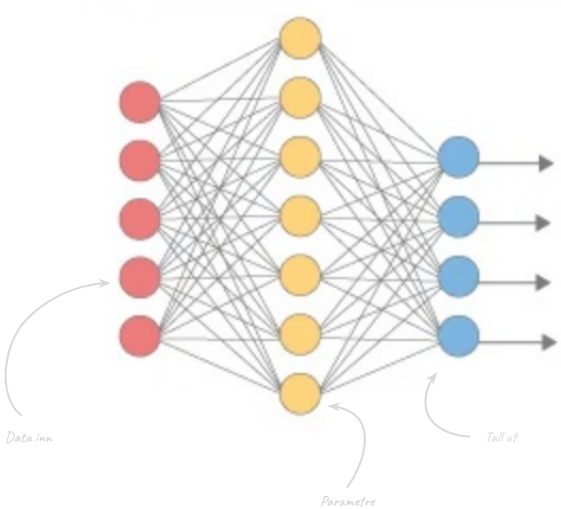
*Nevralt nettverk: Mange parametre som kan **tilpasses** og **kombineres** for å oppnå målet*

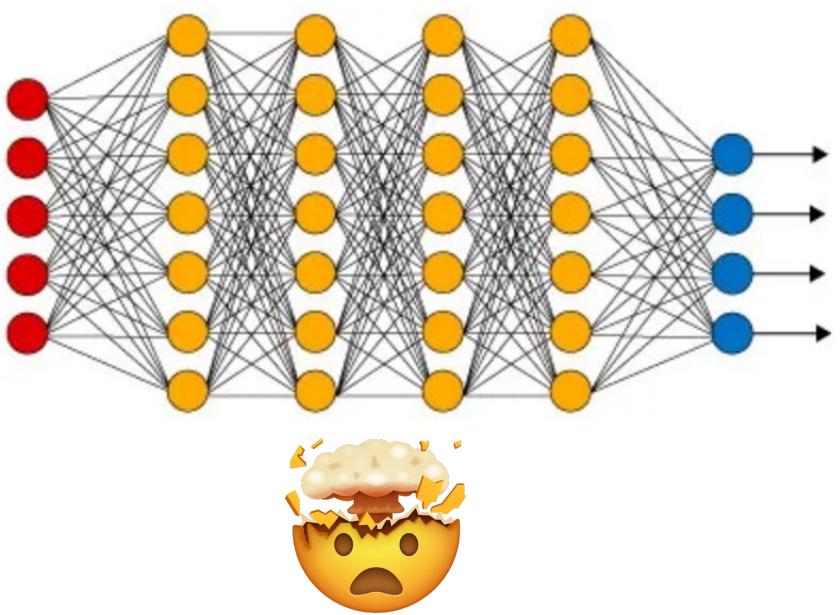


https://thedatascientist.com/what-deep-learning-is-and-isnt/

Maskinlæring!!!

Dypt nevralt nettverk: MANGE parametre som kan tilpasses og kombineres

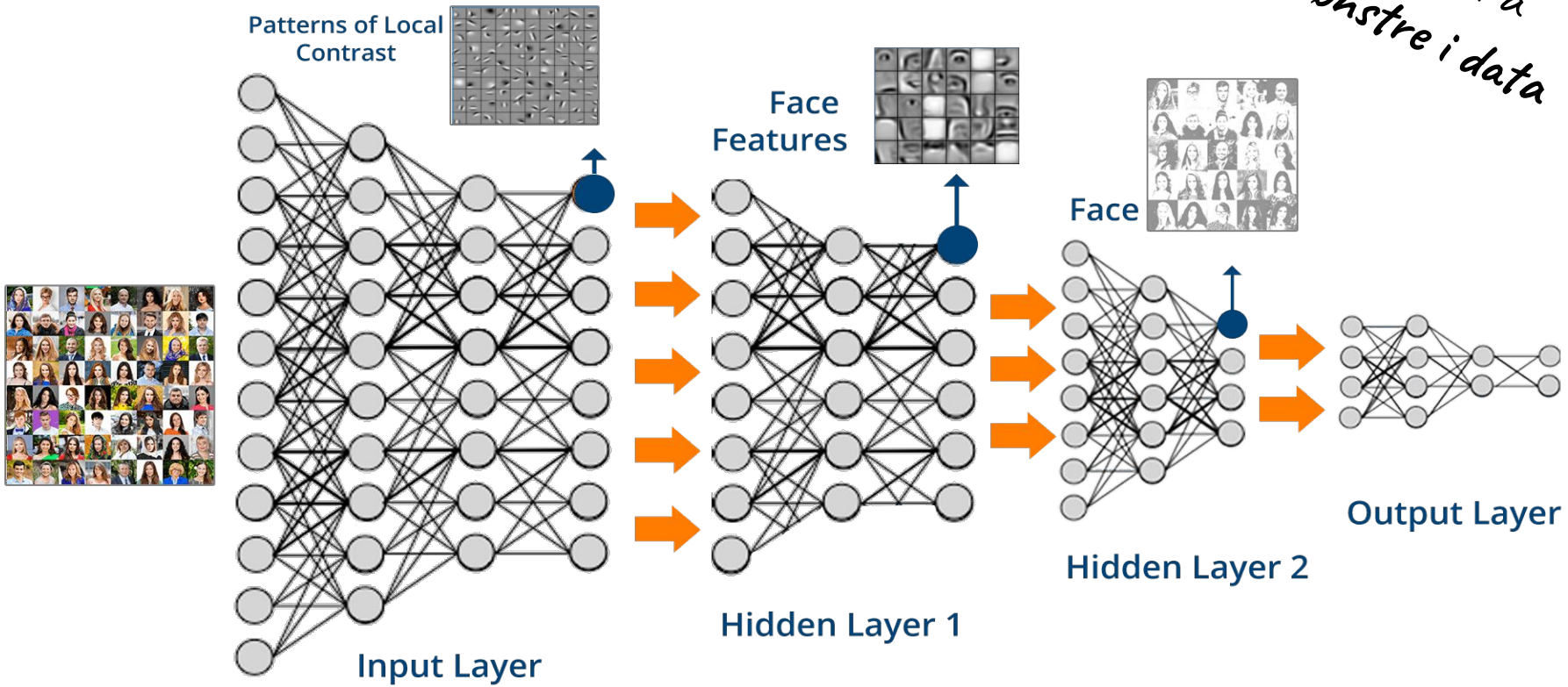




https://thedatascientist.com/what-deep-learning-is-and-isnt/

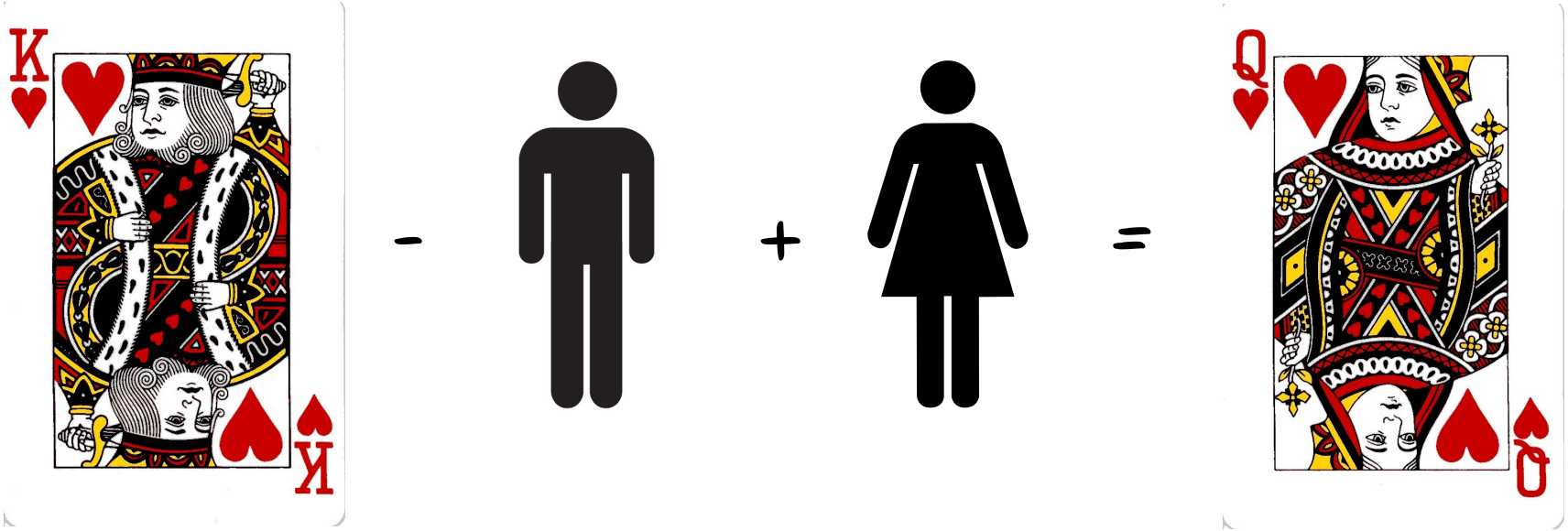
Ansiktsgjenkjenning

Alt koker ned til å finne mønstre i data

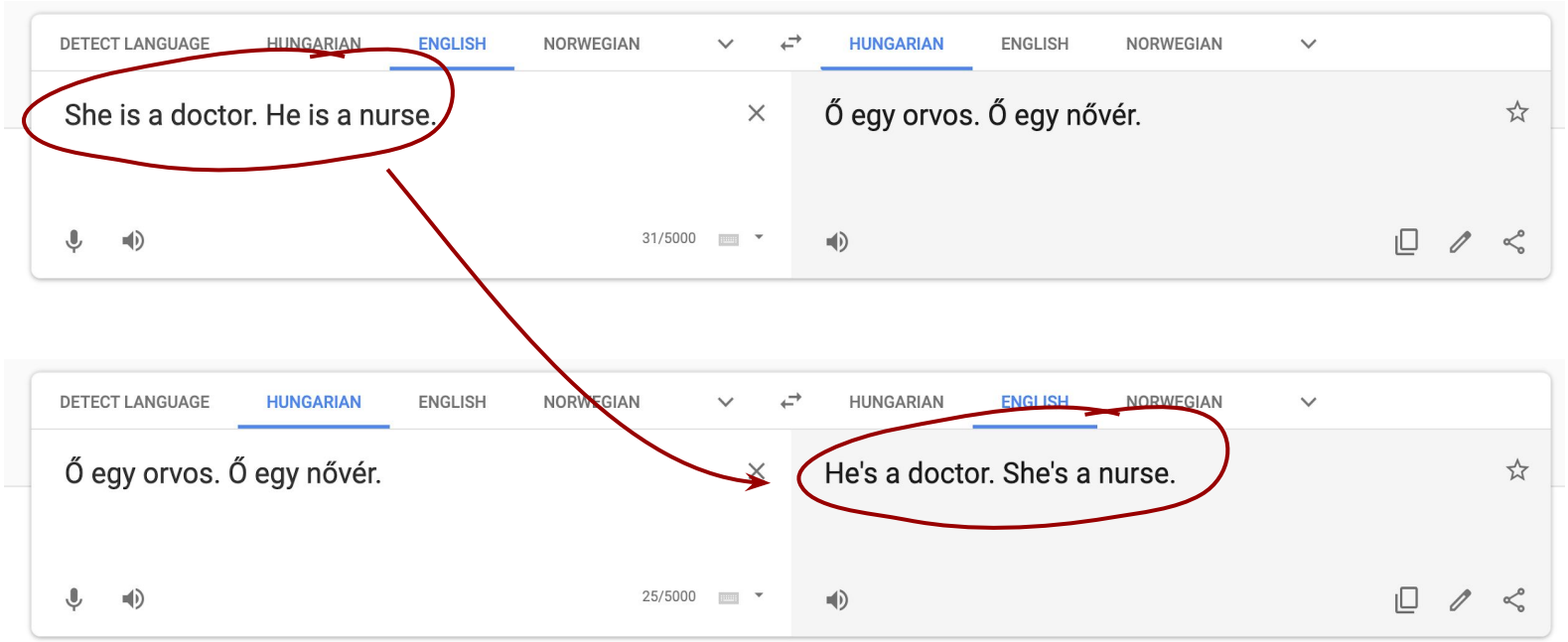


Språkforståelse

Representasjoner av konsepter og relasjoner



Bias 👻


DETECT LANGUAGE
HUNGARIAN
ENGLISH
NORWEGIAN
HUNGARIAN
ENGLISH
NORWEGIAN
She is a doctor. He is a nurse.
Ő egy orvos. Ő egy nővér.
31/5000
DETECT LANGUAGE
HUNGARIAN
ENGLISH
NORWEGIAN
HUNGARIAN
ENGLISH
NORWEGIAN
Ő egy orvos. Ő egy nővér.
He's a doctor. She's a nurse.
25/5000

Bias ≠ underrepresenterte grupper

Hvilke pasienter bør behandles først?
De som vil koste helsevesenet mest!

Sykdom er ikke eneste driver for kostnad...

⇒ *Diskriminering ikke pga ulik representasjon av gruppene i dataene, men* ***målet*** *i den kulturelle og historiske konteksten*

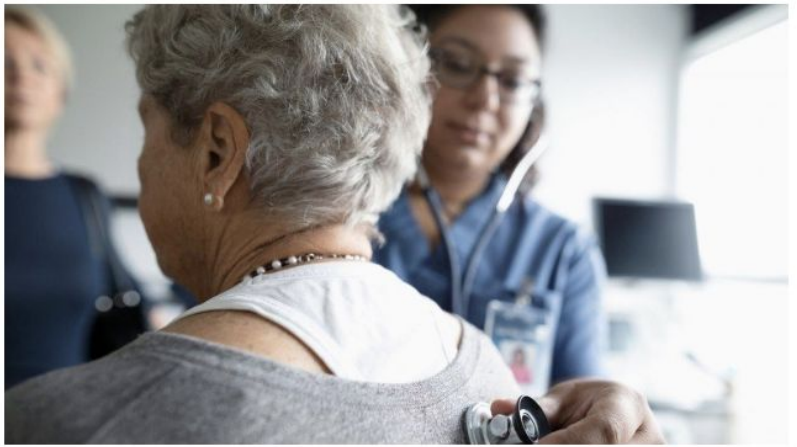
Easy fix: *Diskriminering reduseres >80% ved å predikere også forventet tilbakefall i sykdomsforløp*

Bias kan korrigeres - hvis det oppdages

MIT Technology Review

Artificial Intelligence Oct 25

A biased medical algorithm favored white people for health-care programs



https://www.technologyreview.com/f/614626/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs

Bias ≠ underrepresenterte grupper

Det finnes over 20(?) matematiske definisjoner på rettferdighet. Når vi velger én, bryter vi som regel med en annen.

Statistisk paritet

Forskjellen i raten heldige utfall for den priviligerte og upriviligerte gruppen

Ulik påvirkning

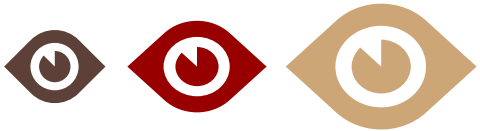
Forholdet mellom raten heldige utfall for den priviligerte og den upriviligerte gruppen

Like muligheter

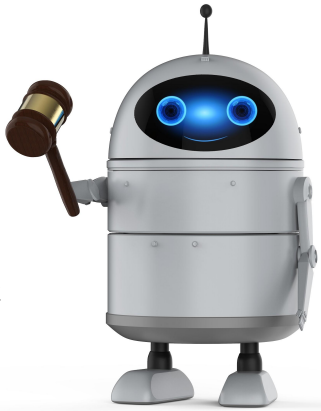
Differansen mellom sann positiv rate mellom gruppene

Gjennomsnittlige odds

Gjennomsnittlig forskjell mellom falsk positiv rate og sann positiv rate mellom gruppene



Compas: AI-dommeren



Oppgave: Beregn risiko for gjentakelse → hvorvidt siktede bør fengsles i påvente av rettssak

Formålet: Gjøre rettssystemet mer rettferdig. Erstatte dommeres forutinntatthet med et objektivt, testbart verktøy.

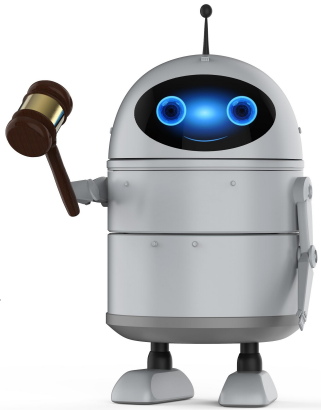
Modellen har ikke tilgang til beskyttede egenskaper, som etnisitet, alder, kjønn og funksjonsgrad

Modellen er **ikke trent på informasjon om etnisitet.**

Oppfører den seg rasistisk?

MIT Technology Review: courtroom algorithm game

Compas: AI-dommeren



Oppgave: Beregn risiko for gjentakelse → hvorvidt siktede bør fengsles i påvente av rettssak

Formålet: Gjøre rettssystemet mer rettferdig. Erstatte dommeres forutinntatthet med et objektivt, testbart verktøy.

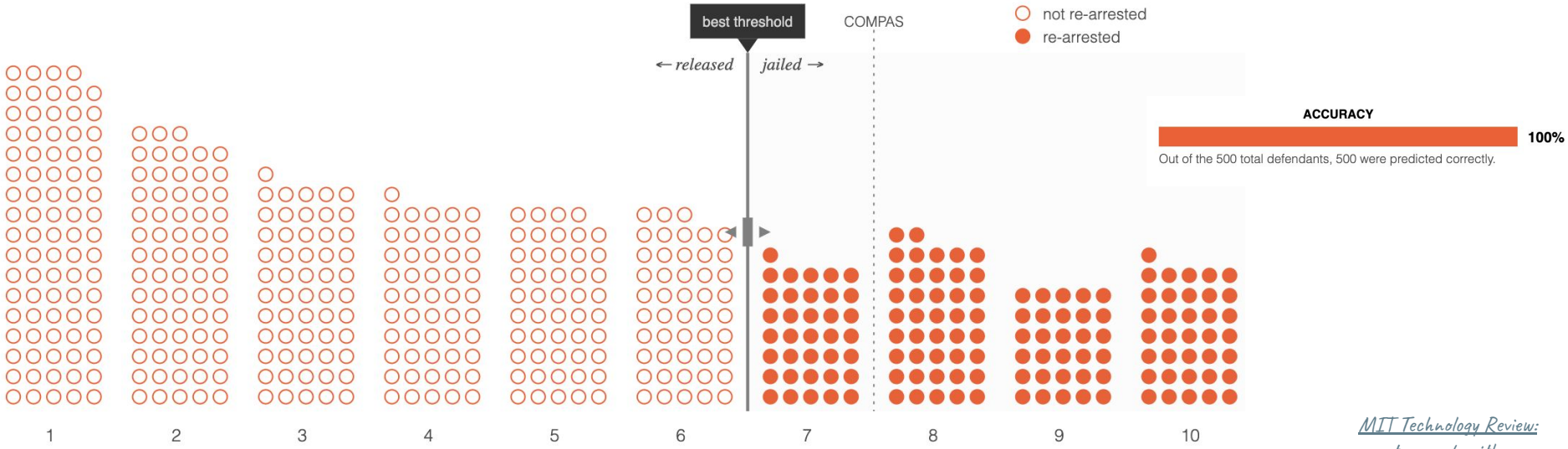
Modellen har ikke tilgang til beskyttede egenskaper, som etnisitet, alder, kjønn og funksjonsgrad

Modellen er **ikke trent på informasjon om etnisitet**, men oppfører seg rasistisk (demonstrert av ProPublica)

Hvorfor?

MIT Technology Review: courtroom algorithm game

Compas: La oss gjøre den rettferdig

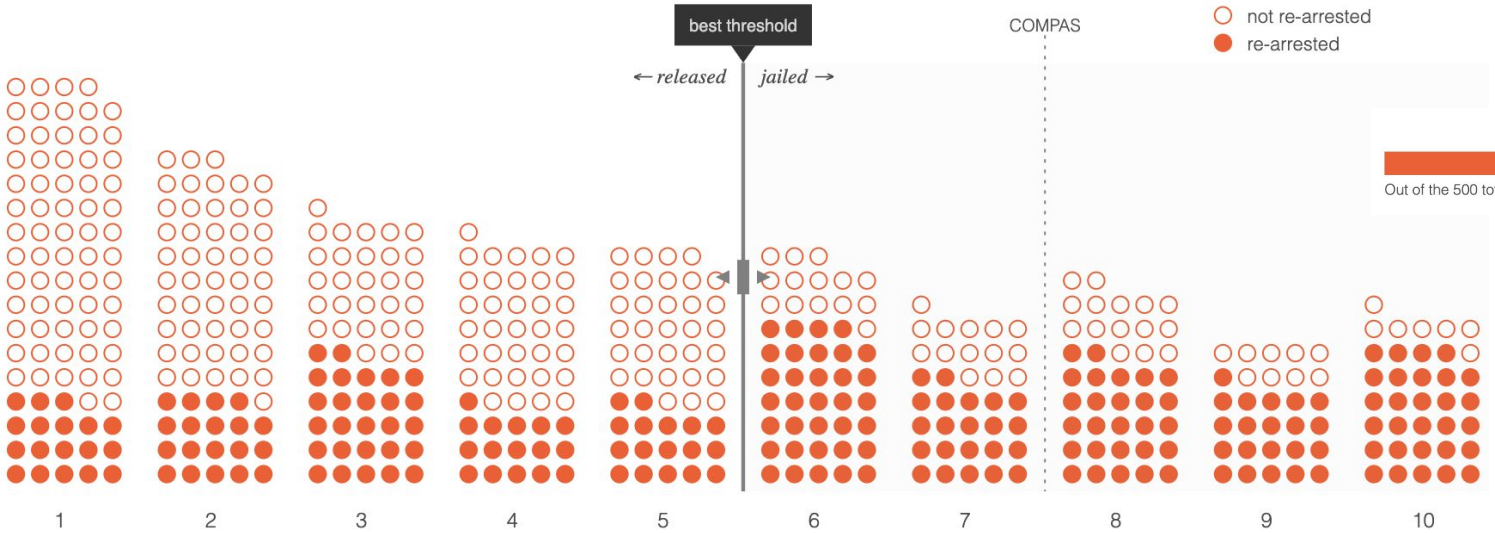


MIT Technology Review: courtroom algorithm game

Compas: La oss gjøre den rettferdig

FEILRATE: Hvor mange mennesker som ikke ble fengslet men gjorde noe straffbart

Hvordan et menneske kommer til å oppføre seg kan ikke forutsies med 100% treffsikkerhet



MIT Technology Review: courtroom algorithm game

Inga Strümke, inga@simula.no, 2021

Compas: La oss gjøre den rettferdig

Komplikasjon: Ulike etnisiteter arresteres med ulik frekvens. Årsaker? Historisk diskriminering?

Konsekvens: Det oppstår to grupper i modellen.

HVA ER RETTFERDIG?

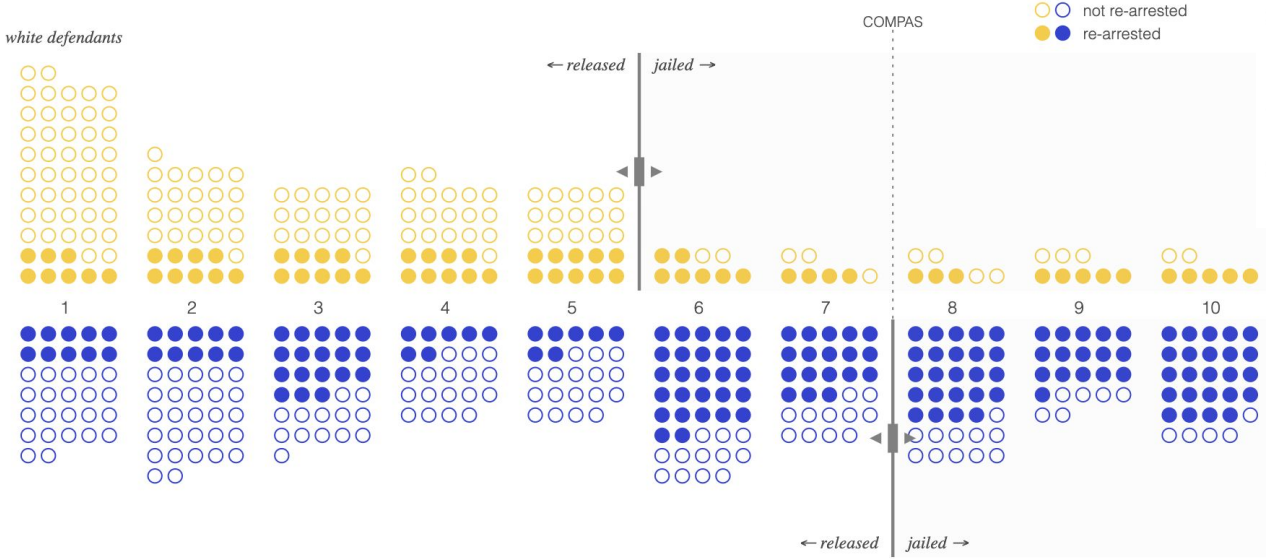
1. Hold feilraten lik mellom gruppene (like mye feilaktig fengsling av mørkhudede og hvite)
2. Behandle mennesker med samme risiko likt

MIT Technology Review: courtroom algorithm game

Inga Strümke, inga@simula.no, 2021

Rettferdighet: Like grupper

PROBLEM: *To mennesker med samme risiko behandles ulikt basert på hudfarge*

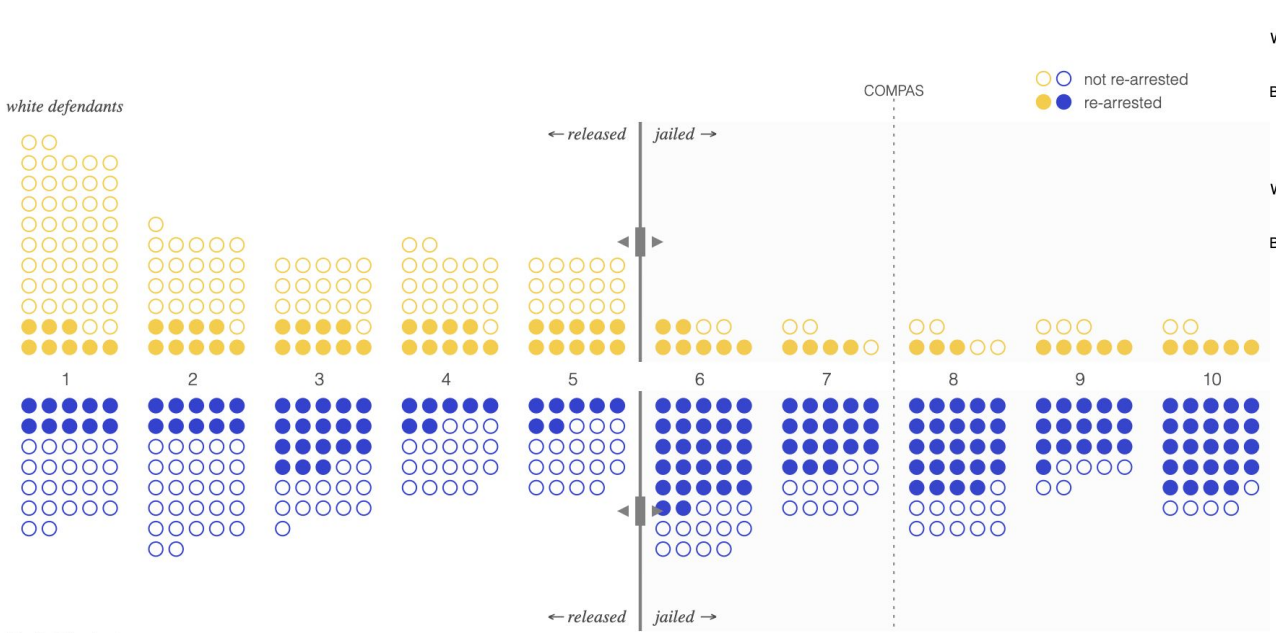


MIT Technology Review: courtroom algorithm game

Inga Strümke, inga@simula.no, 2021

Rettferdighet: Like muligheter

PROBLEM: Feilratene er ulike - flere mørkhudede enn hvite fengsles feilaktig!



MIT Technology Review: courtroom algorithm game

"...Det er et grunnleggende prinsipp i likestillingsjussen at hver enkelt kvinne og mann har krav på å bli vurdert ut fra sine individuelle egenskaper. Dette gjelder selv om kjønn bare er en liten del av totalvurderingen og selv om statistikken er korrekt..."

Likestillingsombudets brev av 25.03.03 til Forsikringsselskapene og FNH

European Commission

EU rules on gender-neutral pricing in insurar industry enter into force

Forsikringer må prises likt

En ny EU-dom om kjønnsnøytral prising av forsikringer, vil kaste om på forsikringspremiene. Også i Norge.

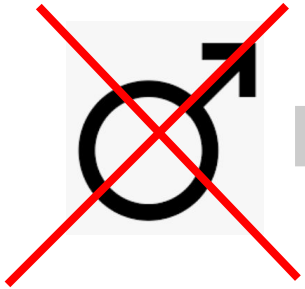
Kjønnsnøytrale forsikringspremier og ytelser

Finansdepartementet har foreslått regler om kjønnsnøytrale premier og ytelser i private, frivillige forsikringer utenfor arbeidsforhold. Forslaget er en direkte konsekvens av rettsutviklingen i EU. De nye reglene skal gjelde for nye forsikringsavtaler etter 1. januar 2015.

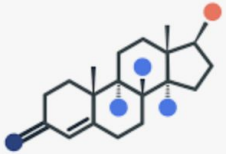
Nye regler kan gi dyrere forsikring

Fra desember neste år blir det forbudt å tilby ulik pris på livsforsikring for kvinner og menn. Resultatet kan bli dyrere forsikring. Les rådene for hvordan du bør tilpasse deg her.

Loven forbyr forsikringsselskaper å bruke kjønn som faktor for beregning av pris

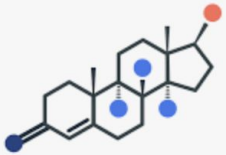


?



?

?



?

Likevel har prisforskjellene økt mellom kvinner og menn

Vanskelig 1:

Juss, etikk, samfunnsviten

Avklaringer og rettspraksis

Vaghet som løsning på hastighet

Løsningen er IKKE å senke standarden!

CENTER FOR DATA INNOVATION

What the Evidence Shows About the Impact of the GDPR After One Year

Specifically, the evidence shows that the GDPR:

1. Negatively affects the EU economy and businesses
2. Drains company resources
3. Hurts European tech startups
4. Reduces competition in digital advertising
5. Is too complicated for businesses to implement
6. Fails to increase trust among users
7. Negatively impacts users' online access
8. Is too complicated for consumers to understand
9. Is not consistently implemented across member states
10. Strains resources of regulators

Vanskelig 1:

Juss, etikk, samfunnsviten

Avklaringer og rettspraksis

Vaghet som løsning på hastighet

Løsningen er IKKE å senke standarden!

Åpenhet.

Den første industrielle revolusjonen

~ 1800: Erstattende

Det tok arbeiderne tre generasjoner å få tilbake den kjøpekraften de hadde før første fabrikk

Den andre industrielle revolusjonen

~ 1850: Forsterkende

Stål til elektrisitet. Arbeiderne ble mer produktive gjennom masseproduksjon og fikk en bedre forhandlingsposisjon.

Inga Strümke, inga@simula.no, 2021

Den fjerde industrielle revolusjonen

Erstattende?

Automatisering: Produksjon, kommunikasjon og ***beslutninger.***

Maskiner ser sammenhenger og tar beslutninger uavhengig av og bedre enn mennesker.

Den fjerde industrielle revolusjonen

Endringer i samfunnet pga automatisering ✓

Ta beslutninger bedre enn mennesker ✓

Maskin-til-maskin kommunikasjon ✓

Selvdiagnostisering og -reparasjon ✓

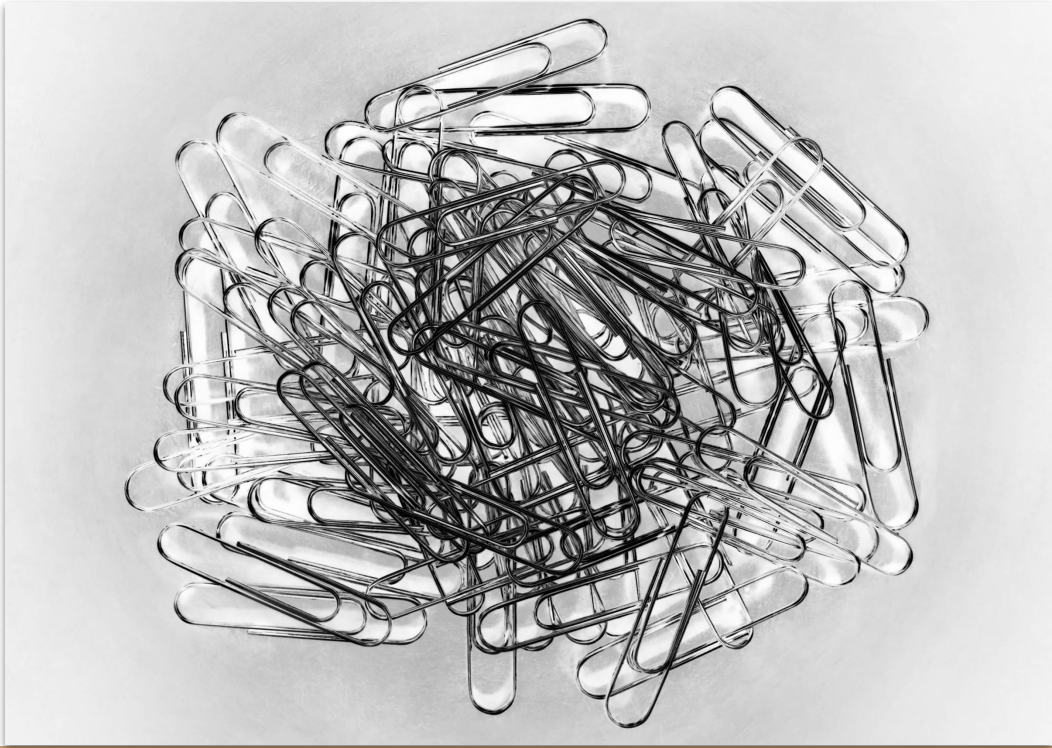
Hva kommer automatisering til å gjøre med samfunnet?

Inga Strümke, inga@simula.no, 2021

Vanskelig 2:

Hva er det som skjer?

Hvilket mål vil vi faktisk oppnå?

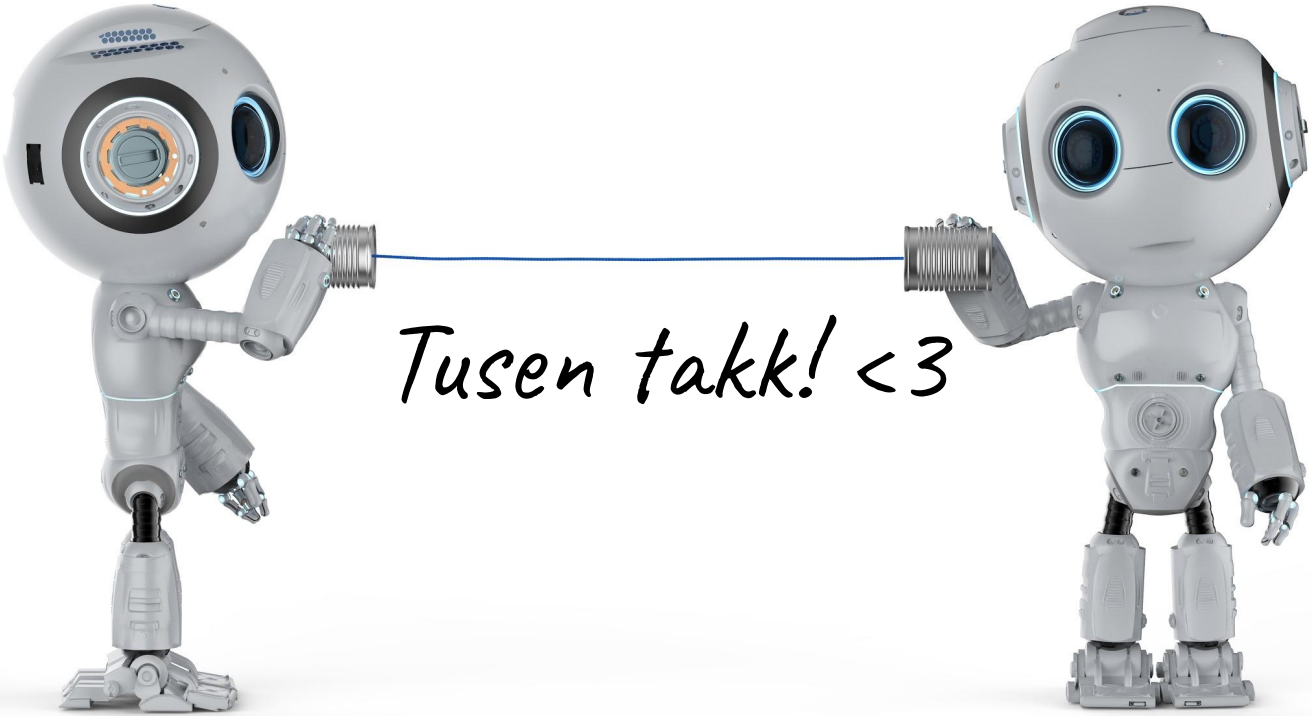


Vanskelig 2:

Hva er det som skjer?

Hvilket mål vil vi faktisk oppnå?

Mål.



Inga Strümke, 2021
inga@simula.no