

Ansvarlig KI

- hvorfor er det så vanskelig

Presentation by Inga Strümke
October 2018





- Jeg synes det er skummelt

Forbrukerombudet ble overrasket over at ansiktsanalyse ble brukt på reklameskilt i Oslo. Og folket er skeptisk.



SKEPTISK: Seniorrådgiver Stian D. Kringelbotn i Datatilsynet mener Peppes Pizza-reklamen er problematisk på flere måter. Han er heller ikke sikker på at den er lovlig, uten at han vil konkludere overfor Dinside. Foto: Datatilsynet

1

Hva er ansvarlig
kunstig intelligens

Ansvarlig kunstig intelligens

2. Kunstig intelligens

Samlebetegnelse: Automatisering, maskinlæring, robotisering, ekspertregler, ...

- Automatiserte beslutninger eller beslutningsstøtte for mennesker
- Maskinlæring: Innsikt fra *historiske* data og en målfunksjon



Modenhet

Samtidig som nye bruksområder for kunstig intelligens dukker opp hele tiden, blir verktøyene mer komplekse. Diskusjonen blir tverrfaglig mellom tekniske og sosiale fagfelt



Tradisjonell programmering

Logikkbaserte systemer definer av ekspertregler laget av mennesker eller basert på data.



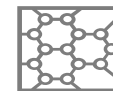
Statistiske modeller

Enkle modeller - f.eks. lineær regresjon - som er basert på data, men er enkle å tolke og til tider intuitive



Maskinlæring, inkludert dyp læring

Avanserte modeller som kan lære mer komplekse sammenhenger fra data, muliggjør automatisering og støtter mer komplekse beslutninger. Disse modellene trenger betraktelig mer data enn tradisjonelle modeller.



Tilpasningsdyktive og kontinuerlig lærende systemer

Adaptive systemer, som sanntidsprising, er vanskeligere å teste og validere, samt forutsi oppførselen til.

Ansvarlig kunstig intelligens

2. Kunstig intelligens

Samlebetegnelse: Automatisering, maskinlæring, robotisering, ekspertregler, ...

- Automatiserte beslutninger eller beslutningsstøtte for mennesker
- Maskinlæring: Innsikt fra *historiske* data og en målfunksjon

1. Ansvarlig

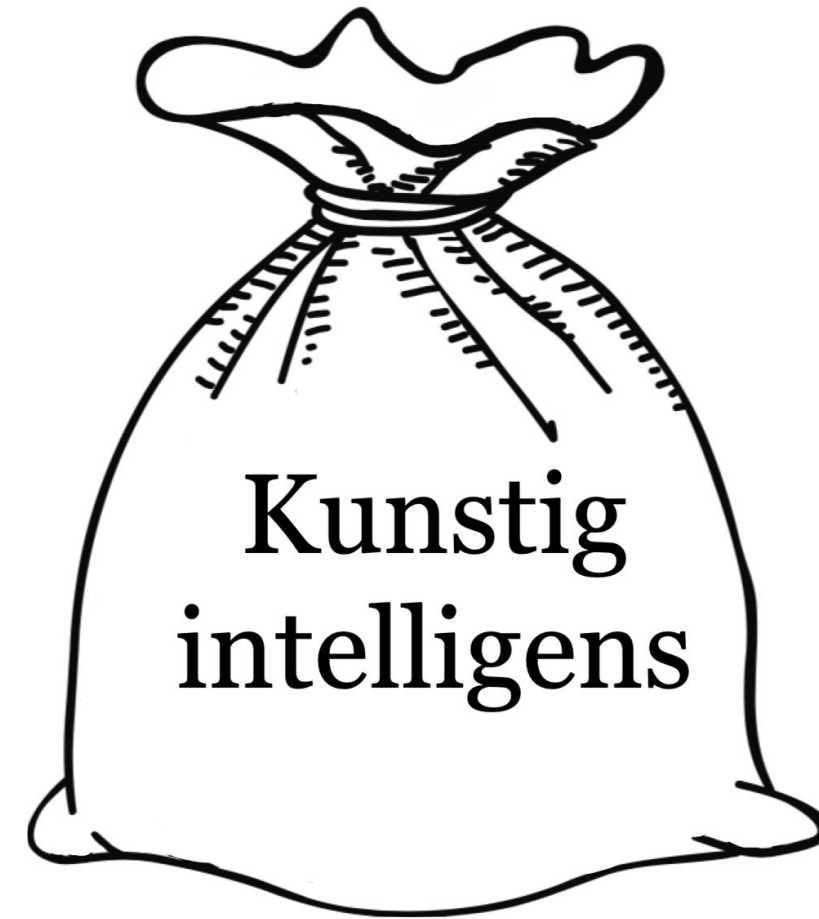
Fagkontekst

Juridisk kontekst

Kulturell kontekst

Teknisk kontekst

Før vi kan kvalitetssikre at en datamaskin tar riktige valg, må vi ha en felles enighet om hva det innebærer



Ansvarlig kunstig intelligens: Bestanddelene

De gode gamle

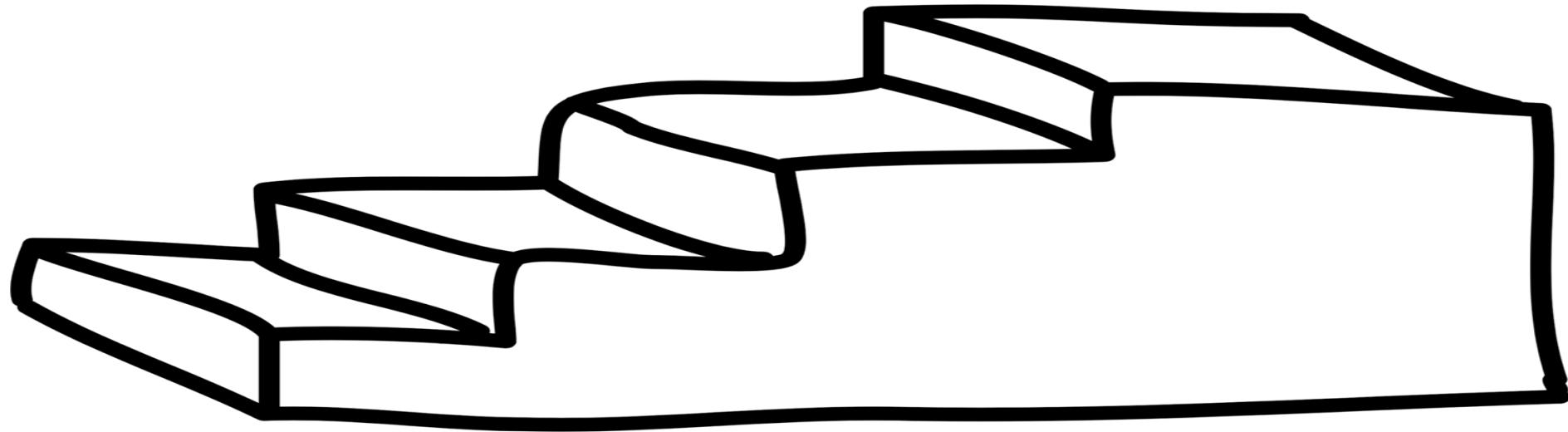
- Verdiskaping
- Regulatorisk og juridisk etterlevelse
- Robust og sikkert
- Personvern
- Skalerbar
- Bærekraftig

De nye

- Rettferdighet
- Forutinntatthet (bias)
- Gjennomsiktighet
- Forklarbarhet
- Forståelse

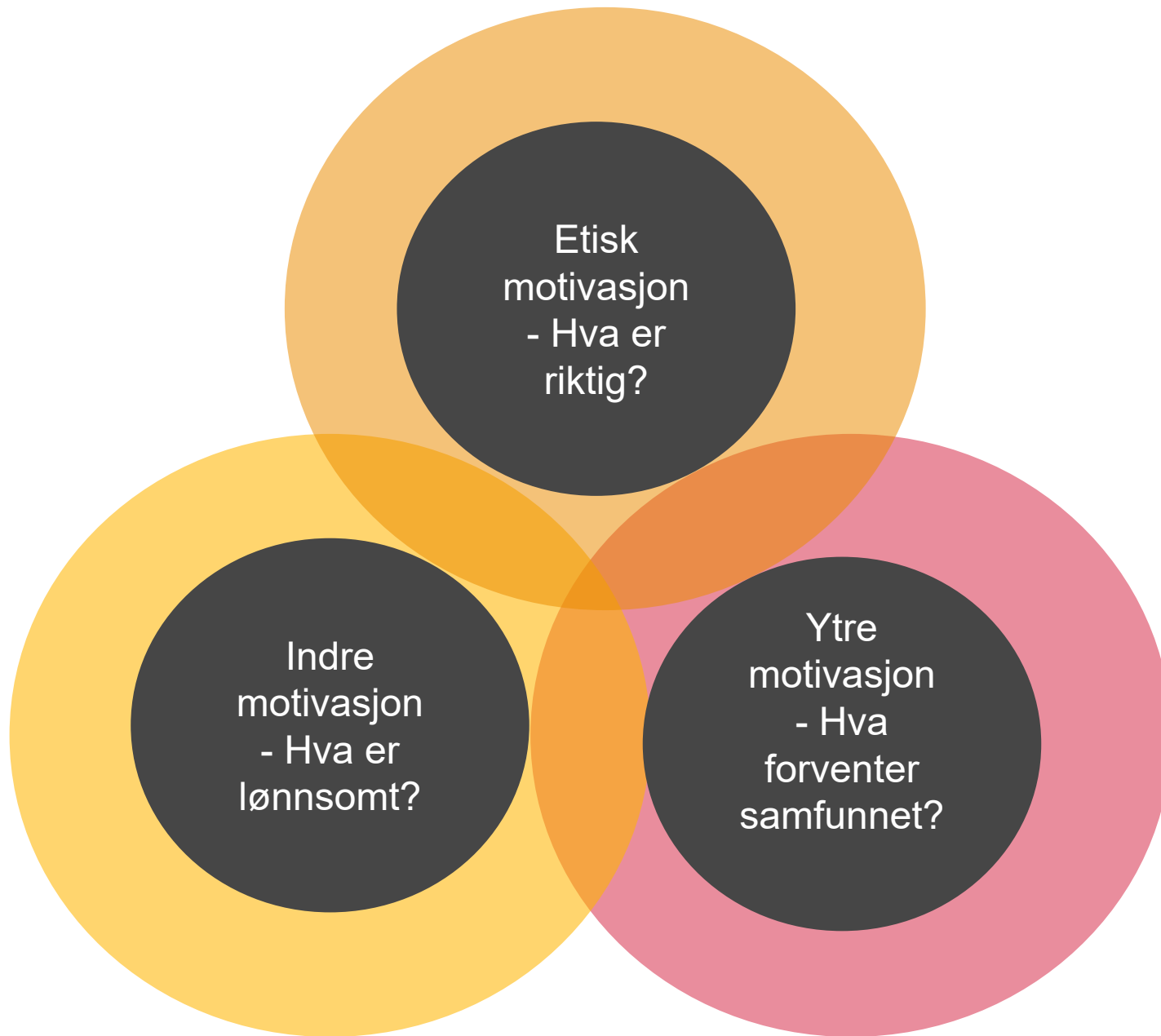
De underliggende

- Tillitsverdig
- Åpen
- Angripelig
- Ansvarsfordelt



2

Hvorfor er ansvarlig
kunstig intelligens
viktig?



Tillitsbrudd ⇒

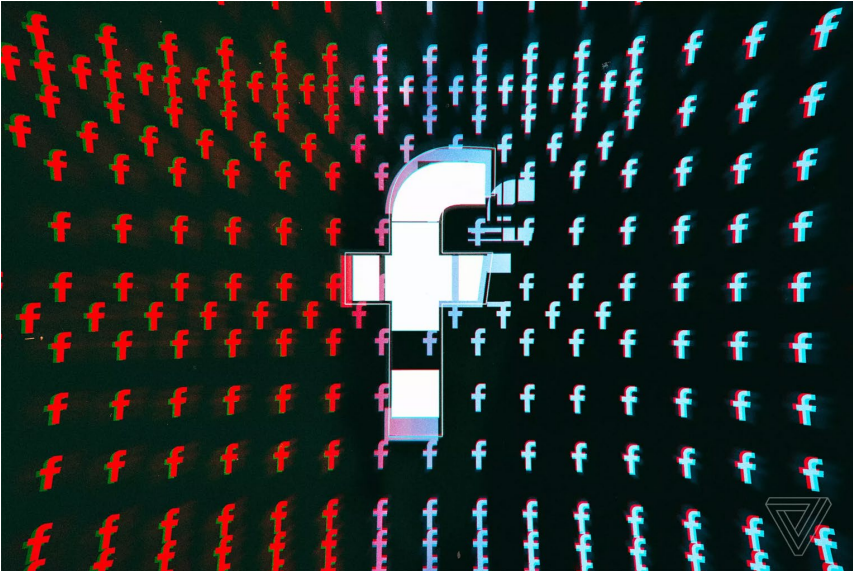
- Mindre muligheter for norske aktører
- Færre løsninger for norske brukere utviklet i henhold til norsk regelverk
- En norsk AI-vinter?

Globalt er det flere land som fokuserer på etiske prinsipper for kunstig intelligens

42 land signerte OECD sine den 22. Mai 2019

- Inkluderende vekst, bærekraftig utvikling og velferd
- Menneskevennlige verdier og rettferdighet
- Gjennomsiktighet og forklarbarhet
- Robusthet og sikkerhet
- Ansvarlighet

... i fall det må skremmel til:



**“Decisions are made to benefit Republicans
because they are paranoid about their
reputation among conservative Republicans,
particularly Trump.”**

- Tidligere Facebook-ansatt forteller nyhetsbrevet
Popular Information at plattformen fatter høyre-
vennlige beslutninger. 17. Oktober 2019.

Trump elected president in Norway

By INGA STRUMKE

Donald Trump gained an unlikely ally this week, when the Norwegian government officially congratulated him on winning the US presidency. Their reasons? They have no choice.

Norway is acutely aware of the US's importance to the European Union. Trump's policies, including his current face-off with the European Union, for instance, certainly not helped.

But government comments on Trump's victory, published on Wednesday, revealed the pressure the small Nordic country faces in trying to remain in the EU.

“Norway, like most EU member states, has no alternative but to honour its ties to the EU and to support the values the bloc stands for, including human rights,” a government statement said.

European Union European Council President Donald Tusk also praised Trump for “once again showing respect for the EU”, and said that the US president “is not just a figurehead or a leader for the United States”.

“He is someone for whom we have a deep and personal relationship,” Tusk said in the European Council's official Twitter account.



3

Hvorfor er ansvarlig
kunstig intelligens
vanskelig?

Bias / forutinntatthet

The image consists of two screenshots of the Google Translate web interface, illustrating a gender bias in the Hungarian translation of a sentence. A red arrow points from the first screenshot to the second, showing the result of switching the source language.

Top Screenshot: The interface shows the source language as **ENGLISH** and the target language as **HUNGARIAN**. The input text is "She is a doctor. He is a nurse." (highlighted in a red oval). The output text is "Ő egy orvos. Ő egy nővér." (highlighted in a light gray box). The interface also shows a microphone icon, a speaker icon, a character count of 31/5000, and icons for copy, edit, and share.

Bottom Screenshot: The interface shows the source language as **HUNGARIAN** and the target language as **ENGLISH**. The input text is "Ő egy orvos. Ő egy nővér." (highlighted in a light gray box). The output text is "He's a doctor. She's a nurse." (highlighted in a red oval). The interface also shows a microphone icon, a speaker icon, a character count of 25/5000, and icons for copy, edit, and share.

Selv om kilden til bias ligger i dataene, er det ikke så enkelt som under- og overrepresenterte grupper

- **Hvite pasienter får høyere risiko av algoritmen**, og har dermed større mulighet til å få behandling enn mørkhudede som er like syke

Hvorfor?

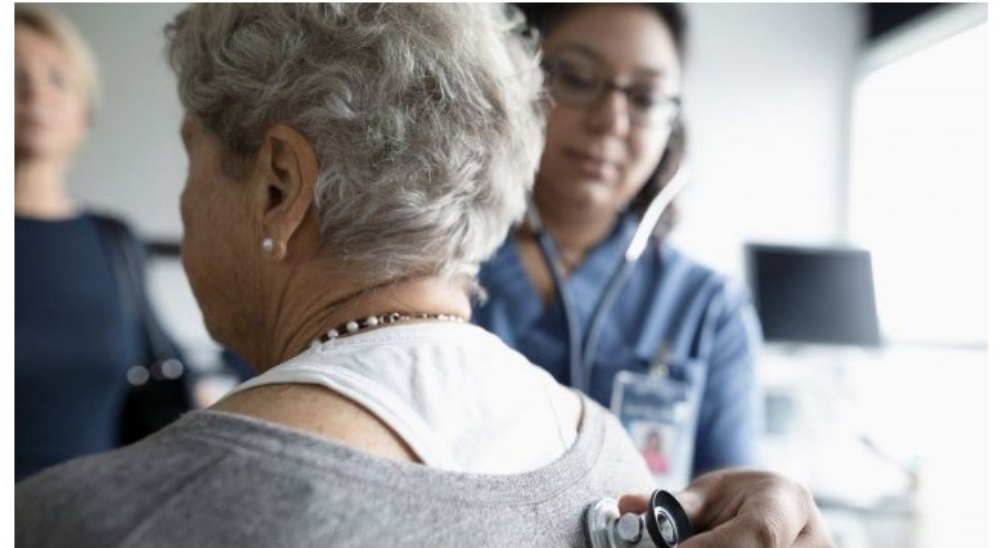
- Medisinske forløp brukes til å predikere hvor mye ulike mennesker kommer til å koste helsevesenet
- Av sosioøkonomiske årsaker har mørkhudede pasienter påført helsevesenet lavere kostnader enn hvite pasienter med samme tilstander

⇒ Grunnen er ikke ulik representasjon av gruppene i dataene, men evalueringsfunksjonen i den kulturelle og historiske konteksten

**MIT
Technology
Review**

Artificial Intelligence Oct 25

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



Rettferdighet

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 privilegerte og upriviligerte gruppen



Perspektiv: Gruppene må behandles likt

Ulik påvirkning

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

Like muligheter

Differansen mellom sann positiv rate mellom den privilegerte og den upriviligerte gruppen

Gjennomsnittlige odds

Gjennomsnittlig forskjell mellom falsk positiv rate og sann positiv rate mellom privilegert og upriviligert gruppe



Perspektiv: Observasjonene reflekterer ferdighet

COMPAS

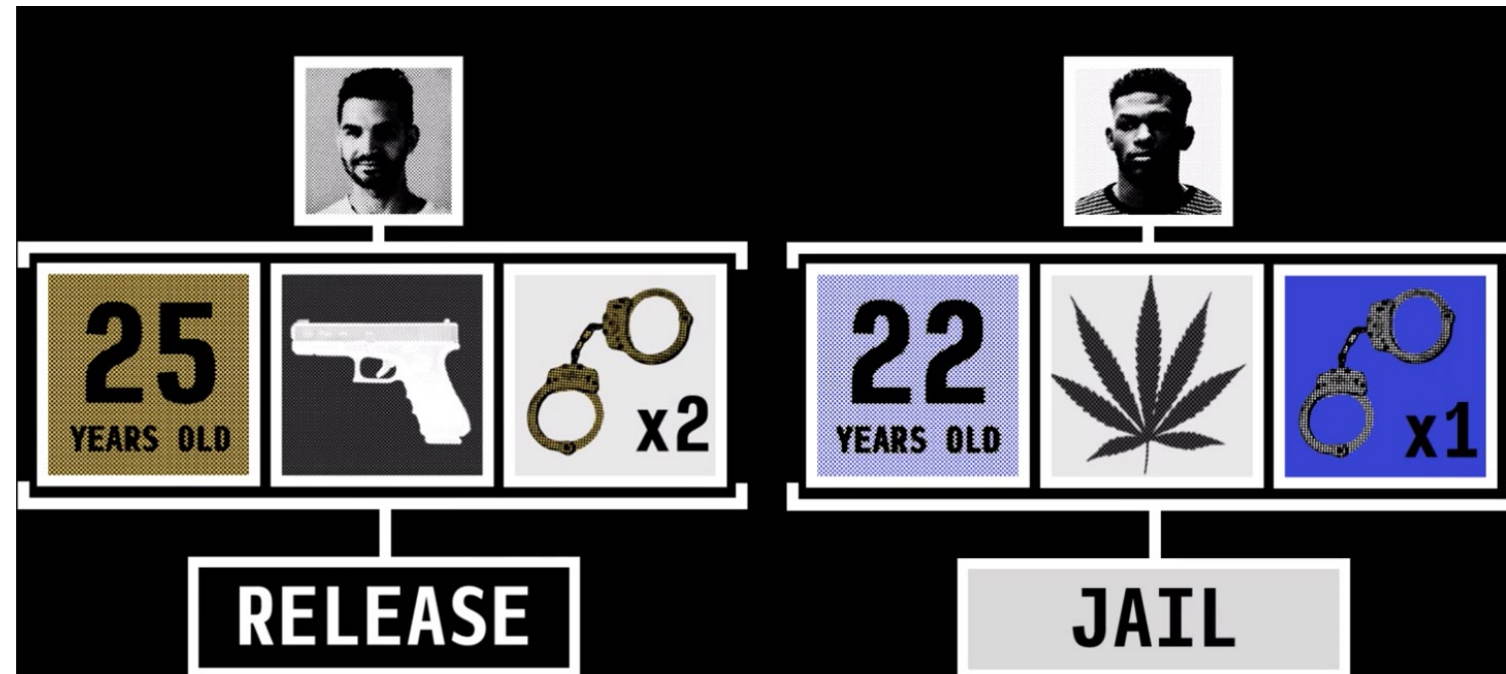
Oppgave: Beregn risiko for gjentakelse → hvorvidt den siktede bør fengsles i påvente av rettssak

Formålet er å gjøre rettssystemet mer rettferdig, ved å erstatte dommers forutinntatthet og intuisjon med et objektive, testbart verktøy.

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

Algoritmen er **ikke trent på informasjon om etnisitet**, men har den rasistisk oppførsel?

Hvorfor?



COMPAS - urettferdig?

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

Formålet er å gjøre rettssystemet mer rettferdig, ved å erstatte dommers forutinntatthet og intuisjon med et objektive, testbare verktøy.

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

Algorithm er **ikke trent på informasjon om etnisitet**, men har likevel rasistisk oppførsel, som [demonstrert i 2016](#) av ProPublica.

Hvorfor?

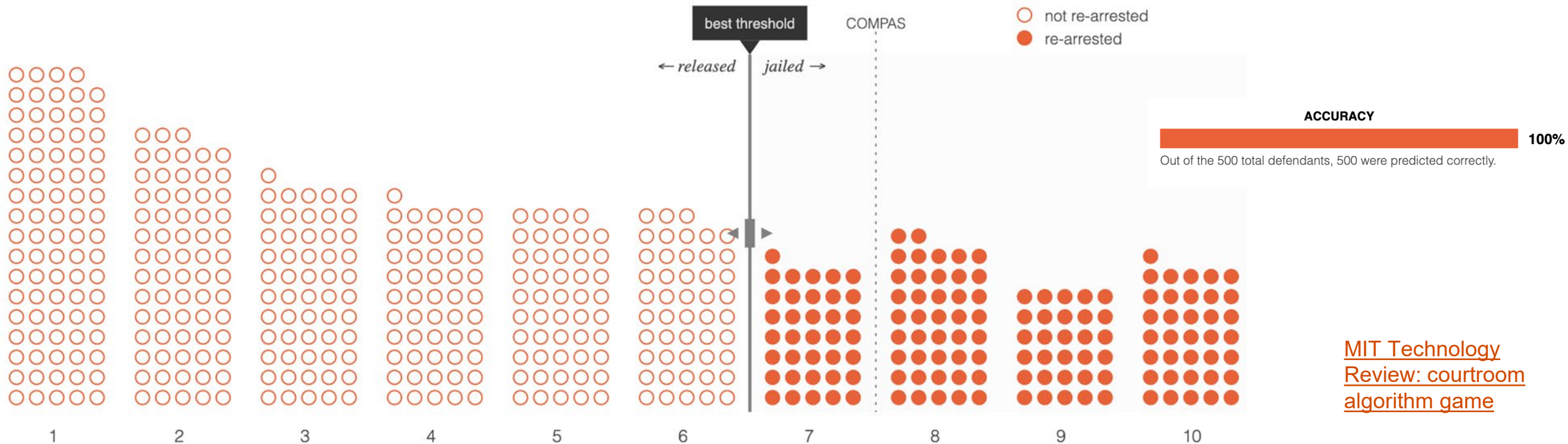
Fordi det finnes prediktorer for etnisitet i datasettet - altså dataene kunne like godt blitt brukt til å lage en etnisitetsprediktor.



COMPAS - la oss gjøre den rettferdig!

Definisjon

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



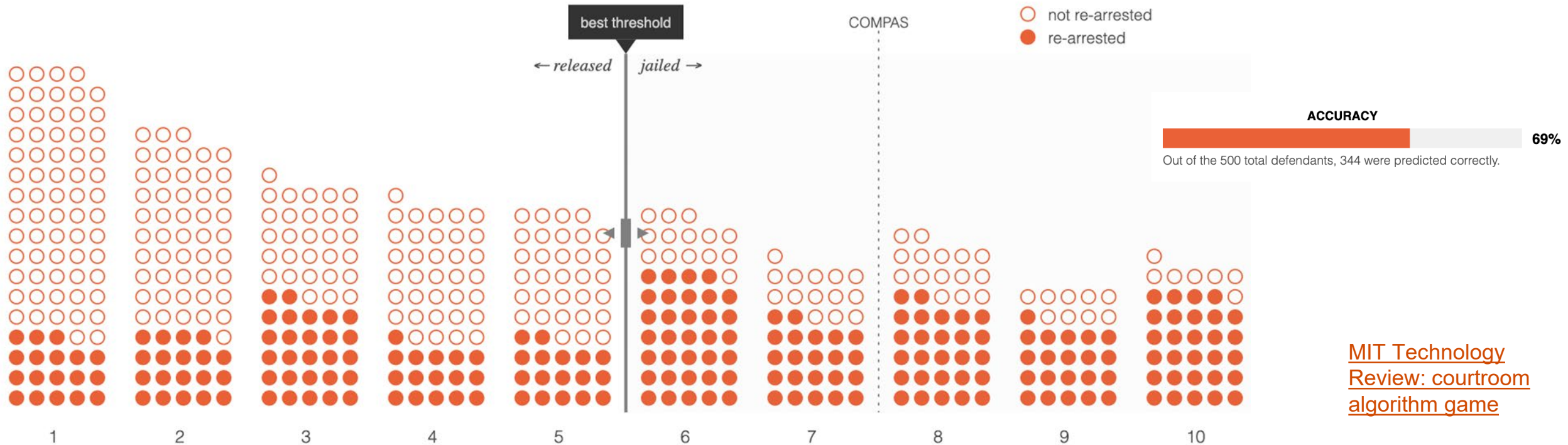
[MIT Technology Review: courtroom algorithm game](#)

COMPAS - la oss gjøre den rettferdig!

Definisjon

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](#)

COMPAS - la oss gjøre den rettferdig!

Definisjon

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

Komplikasjon: Hvite og mørkhudede arresteres med ulik frekvens. Årsaker? Historisk diskriminering?

Konsekvens: Det oppstår to grupper i modellen. Vi skulle ønske det ikke var sånn, men vi må forholde oss til det.

HVA ER RETTFERDIG?

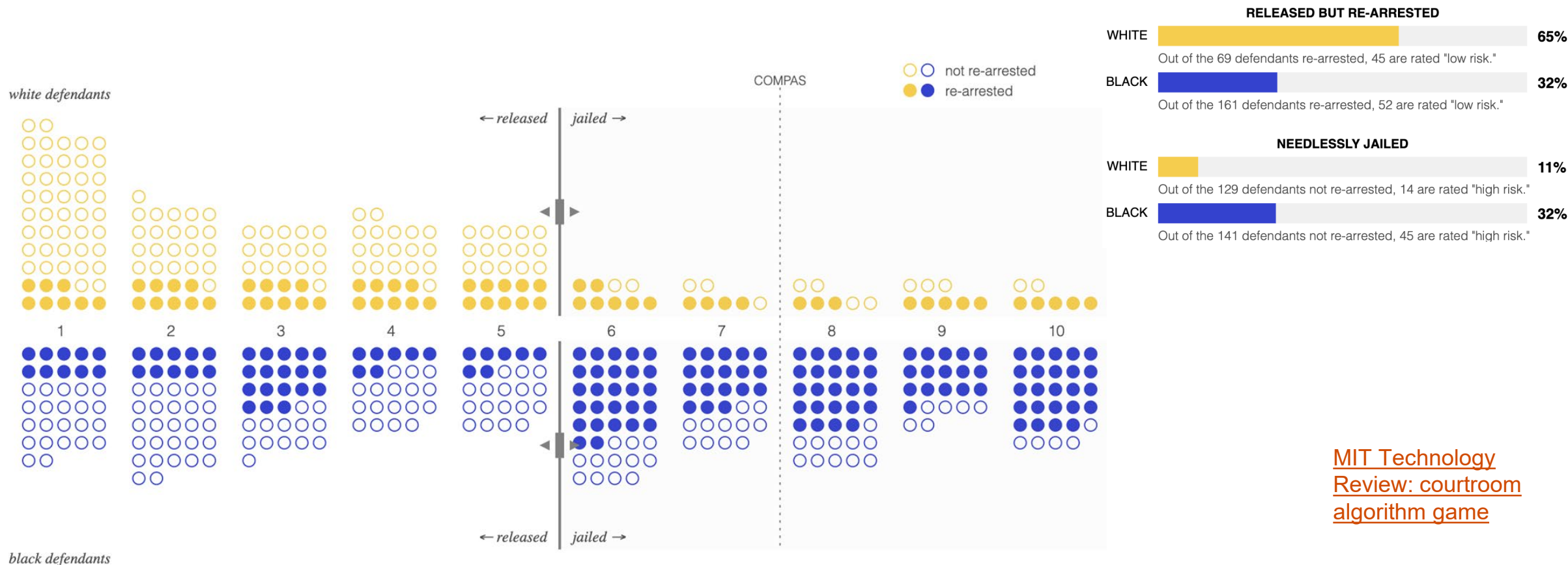
1. Hold feilraten lik mellom ulike grupper (like mye feilaktig fengsling av mørkhudede og hvite)
2. Behandle mennesker med samme risiko likt

COMPAS - la oss gjøre den rettferdig!

RETTFERDIGHET:

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

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



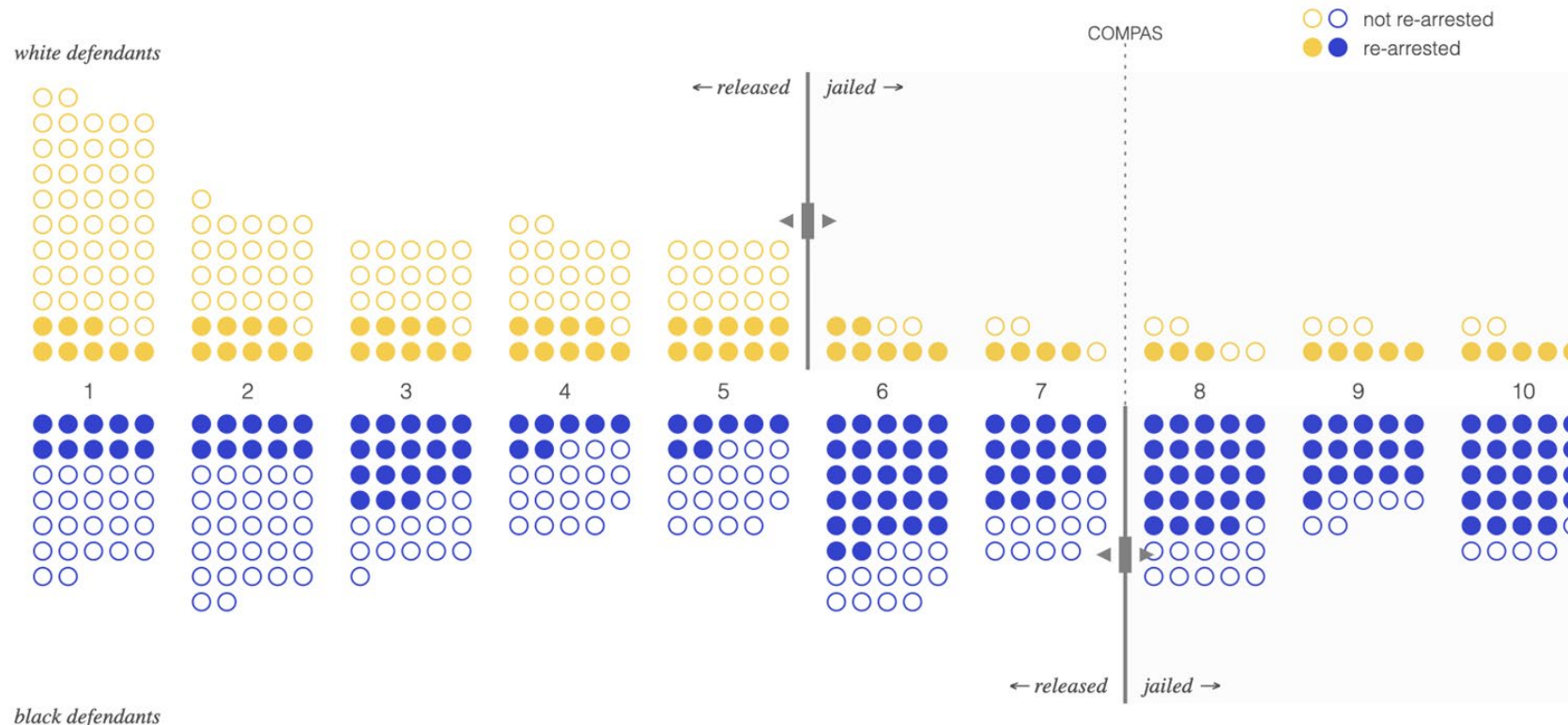
[MIT Technology Review: courtroom algorithm game](#)

COMPAS - la oss gjøre den rettferdig!

RETTFERDIGHET:

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

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



[MIT Technology Review: courtroom algorithm game](#)

“

Mens **bias** er et veldefinert konsept som kan kontrolleres, er **rettferdighet** en menneskeskapt idé som må forstås i kulturell kontekst og der ulike interesser må veies opp mot hverandre

Forklarbarhet

Elefanten i algoritmerommet - og noe vi bare må bli enige om



Hva betyr det?

- Oppskrift på hvordan ting funker
 - Introspeksjon (den underliggende logikken)
 - Rettferdiggjøring (hva er beslutningen basert på)
- Ulike ting for ulike roller

Hva trenger vi det til?

- **Modellbyggere:** Finne ut om modellen har plukket opp de riktige sammenhengene
- **Produkteieren:** Finne ut om modellens atferd er kompatibel med domenet - marked, kunder, reguleringer
- **Sluttbrukere:** Forstå avgjørelsene som rammet dem og hvordan de bør forholde seg til dem



For dypdykkerne: Diskusjon om forklarbarhet mellom Morten Goodwin og meg
<https://soundcloud.com/user-123248803/inga-strumke-episode-1-fra-arendalsuka>



Vi må bli enige om forventningene til forklaring

Disse vil være ulike mellom bransjene, og bør utarbeides av ekspertgrupper med både representanter fra bransjene, og uavhengige fagpersoner fra statistikk og maskinlæring

Eksempel: Prediksjonsmodell for hyttepriser

Forklaring gjennom Shapley-verdier



Location	Trysil
Sqm	130 m ²
Elevation	898 m
Built	1976
Distance to ski	6.2 km
Distance to road	18 meters
Distance to sea	175 km
Distance to lake	10 km
Neighbours	776 w/1000 m
Near neighbour	31 m
Population	76 w/5 km



Lowverk og data

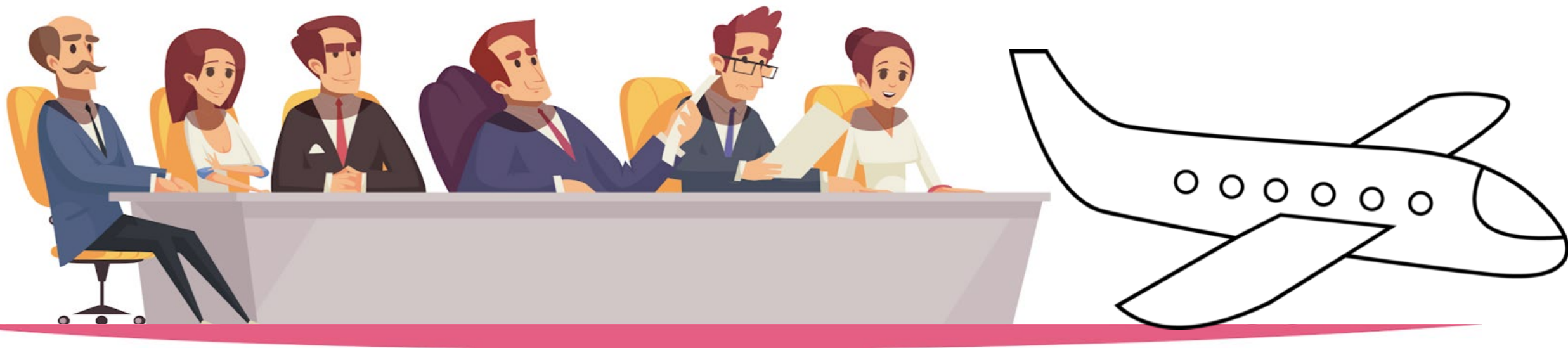


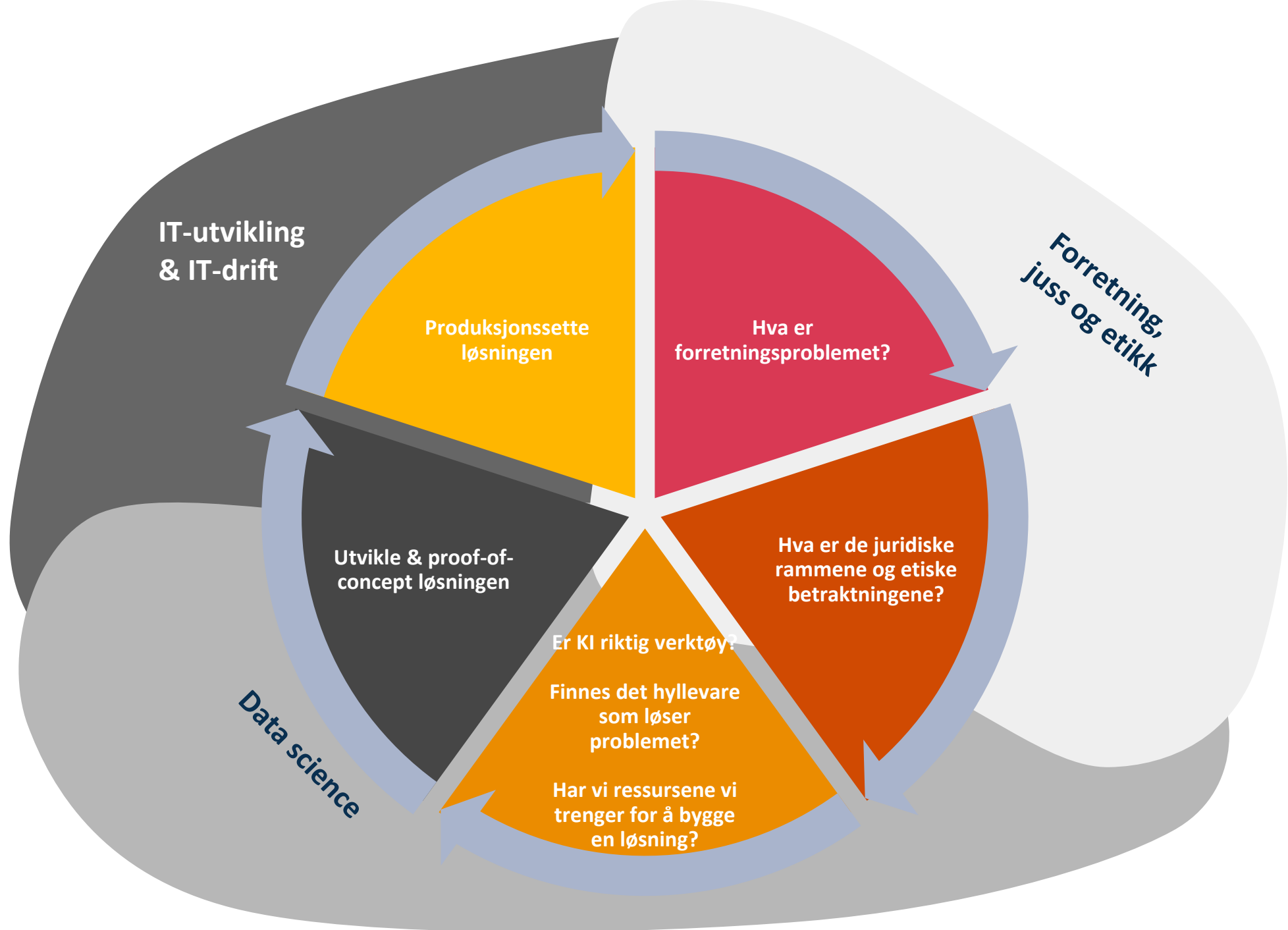
Spørsmålet i dag er ikke “*hva* vil du bruke dataene til?”, men “*hvilke* data vil du ha?”

Google og Facebook er flinkere til å be om data enn mange andre - men er det de som gagnar oss mest om får dataene våre?

4

Hvordan skal vi få til
ansvarlig kunstig intelligens?





Tips for å utvikle ansvarlig kunstig intelligens

Prosjektgangen

1. Innse at dette er **tverrfaglig** og komplekst
2. Få oversikt over relevant **lovverk**, **bransjespesifikke** krav og rammer, tilgjengelig **data**
3. Involver noen med **domenekunnskap**
4. Involver en **jurist**
5. Hvis løsningen skal brukes på mennesker: Involver en **etiker** eller **samfunnsviter**
6. Involver en **statistiker** eller noen som har erfaring med modellbygging og -evaluering

Takk!

inga.struemke@pwc.com

pwc.no



© 2019 PwC. Med enerett. I denne sammenheng refererer "PwC" seg til PricewaterhouseCoopers AS, Advokatfirmaet PricewaterhouseCoopers AS, PricewaterhouseCoopers Accounting AS og PwC Tax Services AS som alle er separate juridiske enheter og uavhengige medlemsfirmaer i PricewaterhouseCoopers International Limited.