Prediksjon av skattekortet

Trygve Bertelsen Wiig

William Peer Berg

BEKK

Hvor mange kjenner til denne?

Restskatt? 

Triggerspørsmål:
Kan problemet reduseres ved å bruke maskinlæring?

Dagens løsning

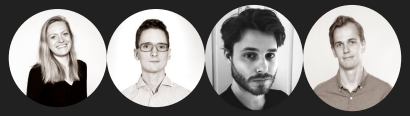
```
if person.inntekt_forrige_år > 0:
    inntekt = person.inntekt_forrige_år*LONNSOKNING_SATS
    if person.nav.arbeidsgiver == None:
        inntekt = 0
(...)
```

↓ 20 år senere...

Mål:

2 år gammel skattedata

\+ Maskinlæring

× 

= Skattekort for alle nordmenn

Hva er en PoC?

- Grovt, men fungerende produkt
- Vill og målløs eksperimentering

Hva er en PoC?

- ~~Grovt, men fungerende produkt~~
- ~~Vill og målløs eksperimentering~~
- Kontrollert eksperimentering for å finne ut om en idé har livets rett

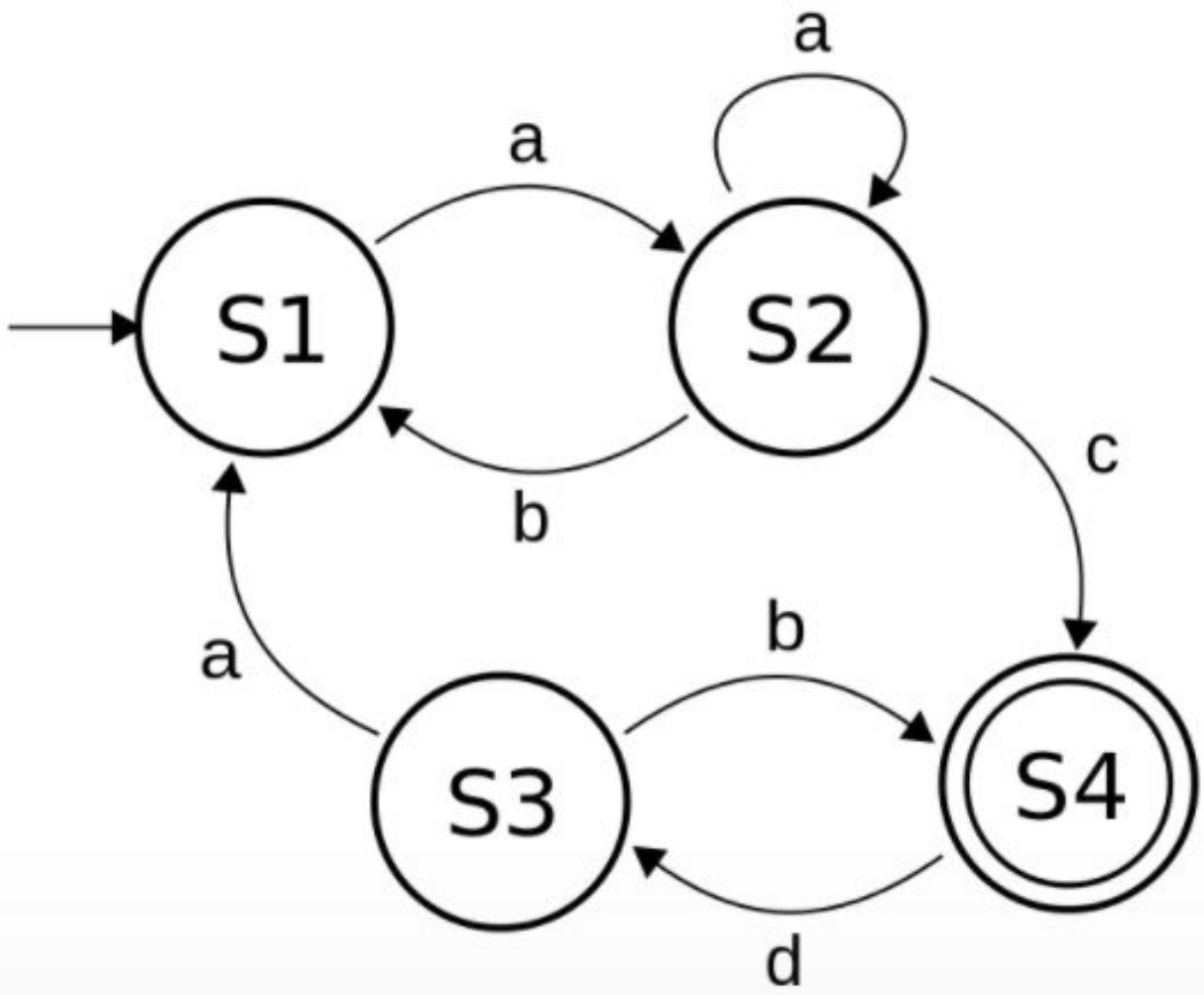
Parprogrammering
(Svært) smidig
VDI

Arbeidsform

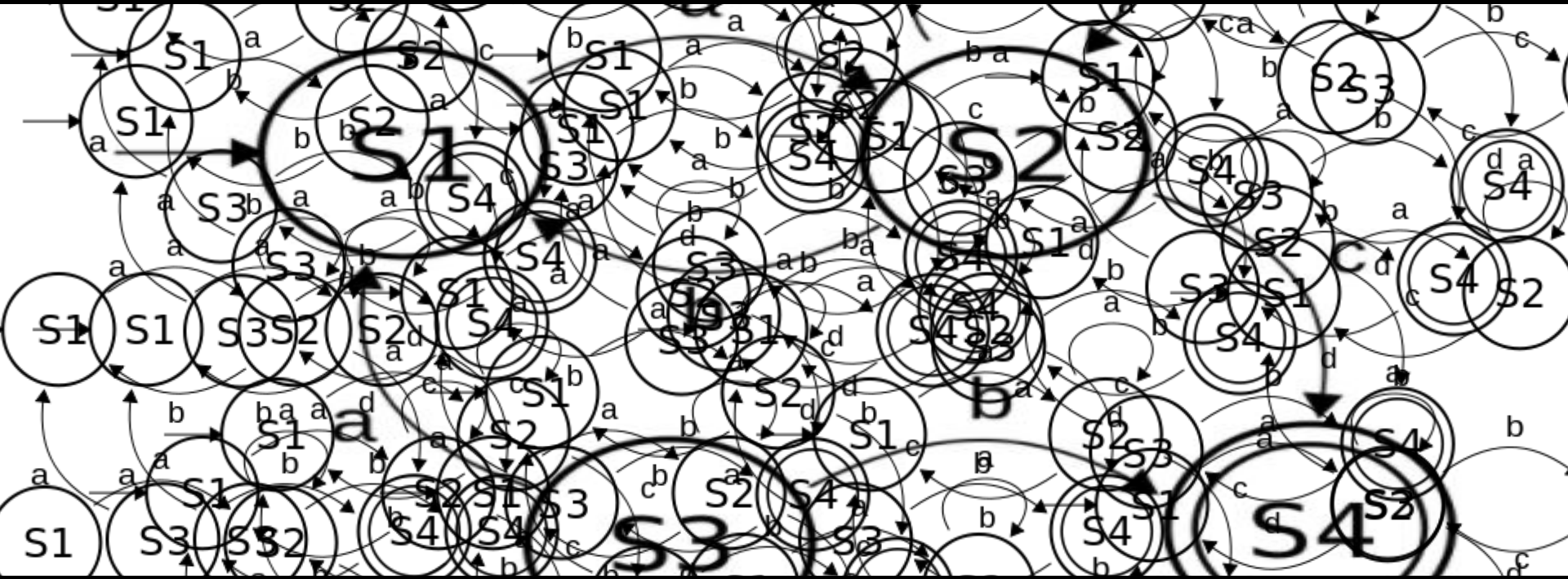
Kanban-board
Daglige standups
Individuelle initiativ

```
VDI$ ./tren_modell.py
Trener modell...
Status etter 0 timer og 52 minutter:   modell 4% ferdig trent
Status etter 2 timer og 18 minutter:   modell 13% ferdig trent
Status etter 4 timer og 11 minutter:   modell 22% ferdig trent
Status etter 6 timer og 56 minutter:   modell 30% ferdig trent
Status etter 8 timer og 0 minutter:   modell 36% ferdig trent
Status etter 10 timer og 55 minutter:   modell 41% ferdig trent
Status etter 12 timer og 23 minutter:   modell 49% ferdig trent
```

State of the art



I virkeligheten....

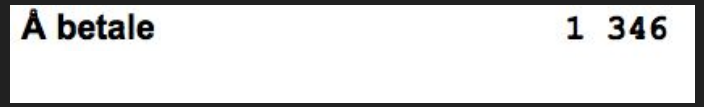


...men det funker... stort sett!

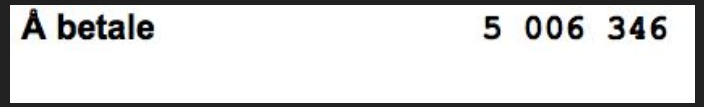
Problemet...

Dyrt...

Dyrere...



Veldig dyrt!

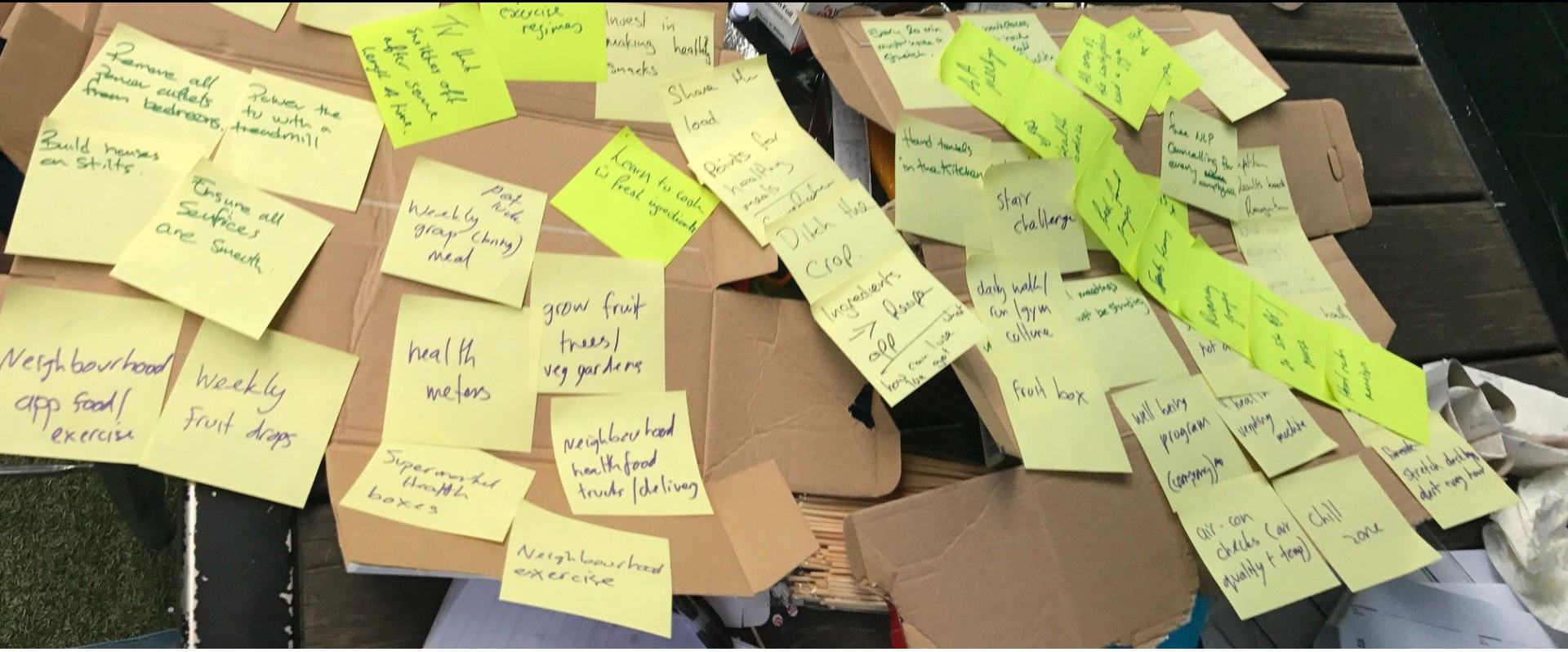


Ting folk gjør med maskinlæring:

Generere forstyrrende bilder...

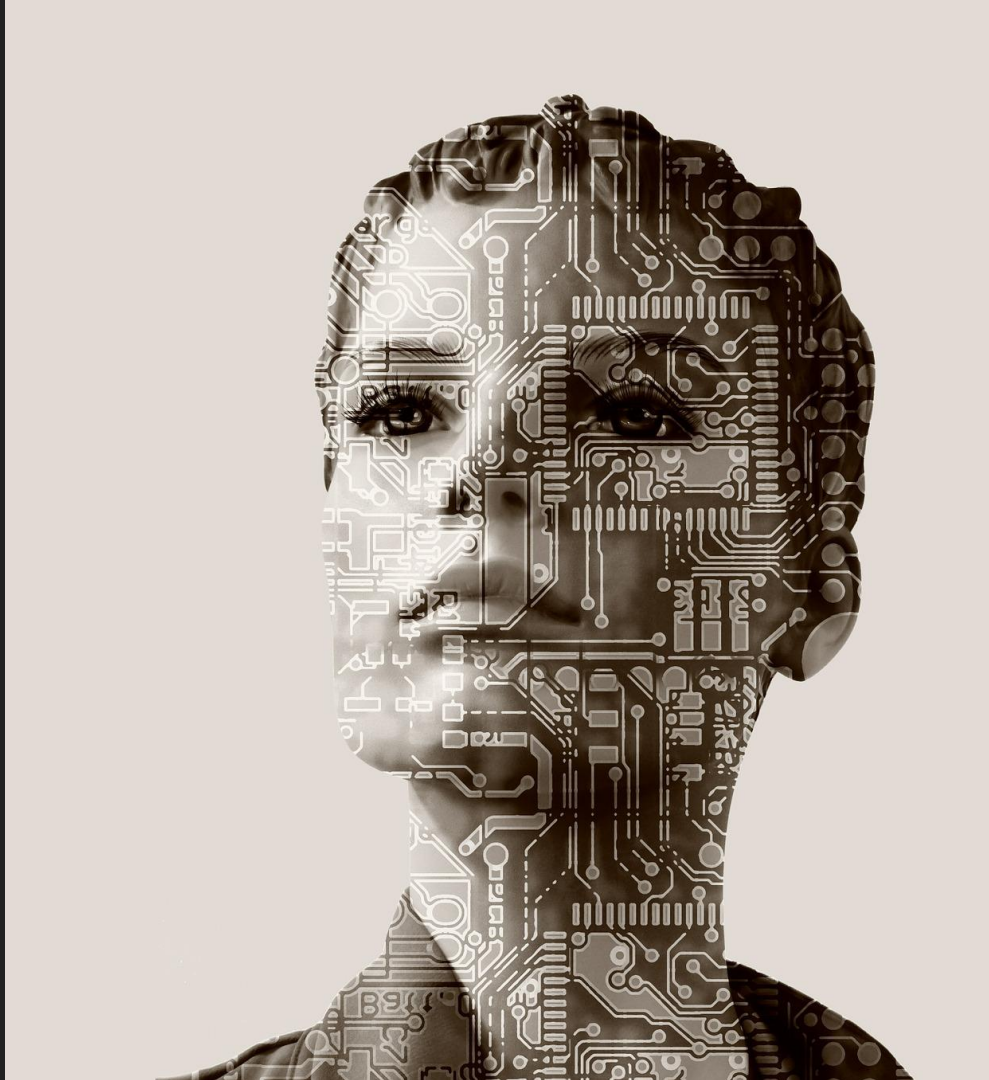
PayPal detekterer svindel basert på transaksjonsdata

Elkjøp oppsummerer kundefeedback ved bruk av sentimentanalyse



Maskinlæring ser sammenhenger mennesker ikke ser…

…eller som vi bruker årevis på å finne.



Kan man også forutsi skattekortet?

Skatteetaten

Skattetrekksmelding:

Informasjon om skattekortet 2017

Dette brevet er kun til din informasjon. Det skal ikke leveres videre til arbeidsgiver/pensjonsutbetaler.

- Dine arbeidsgivere/pensjonsutbetalere henter skattekortet elektronisk, uten at du må foreta deg noe.

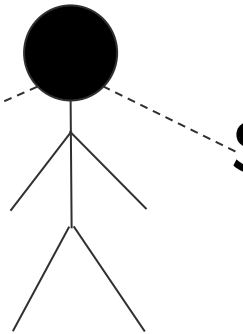
Ditt skattetrekk for 2017

- **31 %** skattetrekk av dine inntekter

Kan man også forutsi skattekortet?

Skatteetaten

Avsetning
reineierfond: kr 0



Skatt: ????

Skattetrekksmelding:

Informasjon om skattekortet 2017

Dette brevet er kun til din informasjon. Det skal ikke leveres videre til arbeidsgiver/pensjonsutbetaler.

- Dine arbeidsgivere/pensjonsutbetalere henter skattekortet elektronisk, uten at du må foreta deg noe.

Ditt skattetrekk for 2017

- **31 %** skattetrekk av dine inntekter

Hypotese:

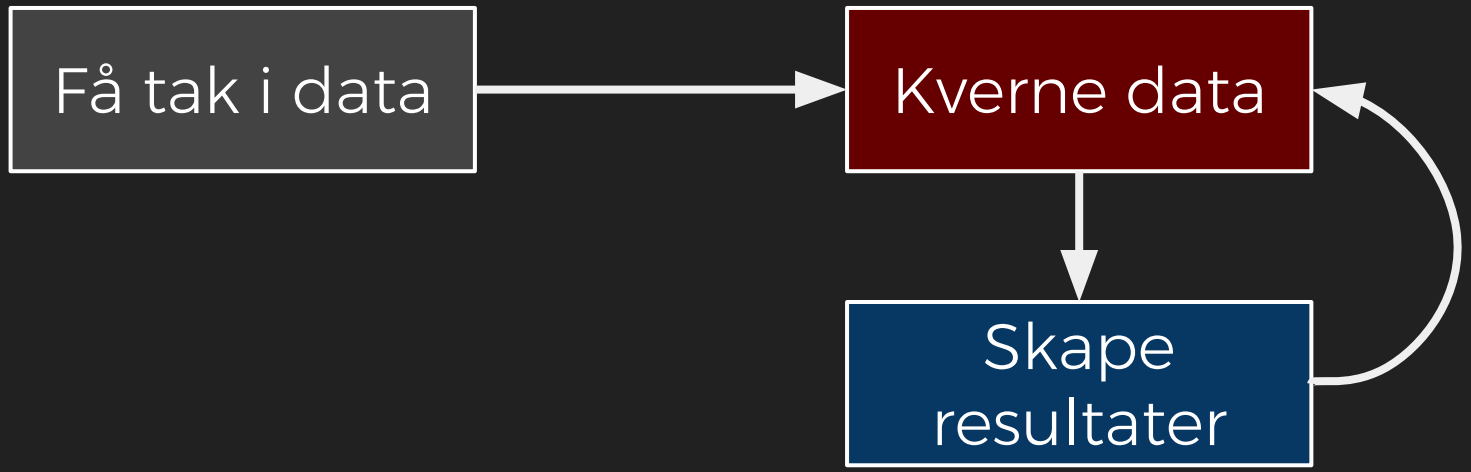
Maskinlæring kan gi en bedre prognose av skattekortet – slik at vi sparer Skatteetaten for ressurser.

Gjennomføring

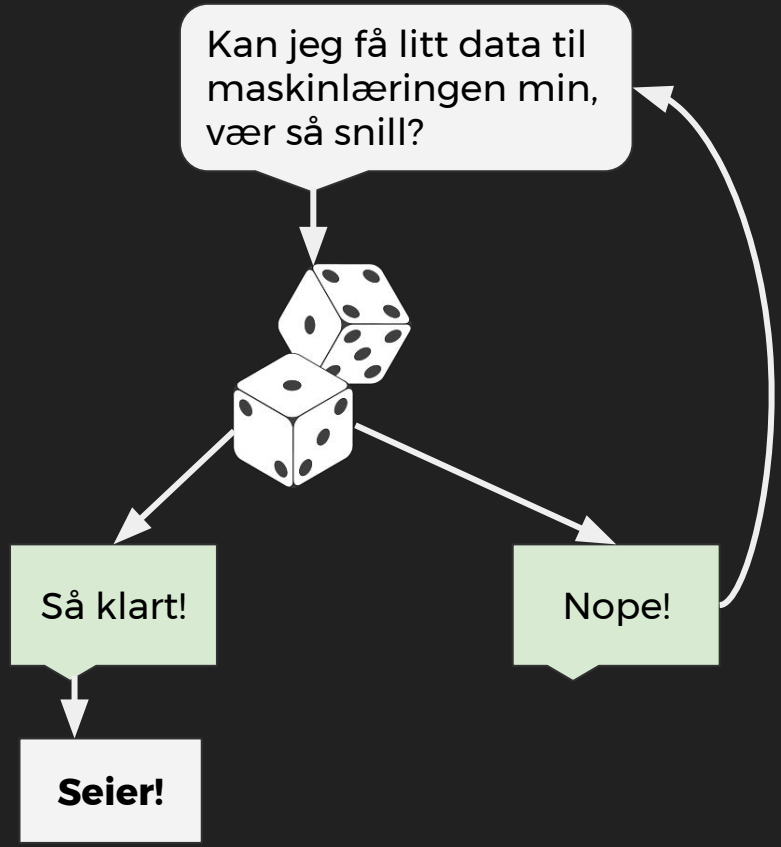
Prosessen

1. Få tak i data
2. Kverne data
3. Skape resultater

Prosessen – et litt riktigere bilde



Få tak i data (for oss)



Kverne data

Fødselsnr.	Fødselsår	Inntekt	Fradrag reinsdyr	Sivilstand
01019112345	1991	500 000	0	0
31121998765	1919	2000	0	4
01010114689	2001	300	0	0
04110492843	2004	0	0	0

Gjøre attributter mer tolkbare

Fødselsnr.	Alder	Inntekt	Fradrag reinsdyr	Sivilstand
01019112345	26	500 000	0	0
31121998765	98	2000	0	4
01010114689	16	300	0	0
04110492843	13	0	0	0

Normalisere flyttall

Fødselsnr.	Alder	Inntekt	Fradrag reinsdyr	Sivilstand
01019112345	26	1.5	0	0
31121998765	98	-1.7	0	4
01010114689	16	-0.5	0	0
04110492843	13	-3	0	0

Fjerne informasjonsløs data

Fødselsnr.	Alder	Inntekt	Sivilstand
01019112345	26	1.5	0
31121998765	98	-1.7	4
01010114689	16	-0.5	0
04110492843	13	-3	0

Binærkod kategoriske variabler

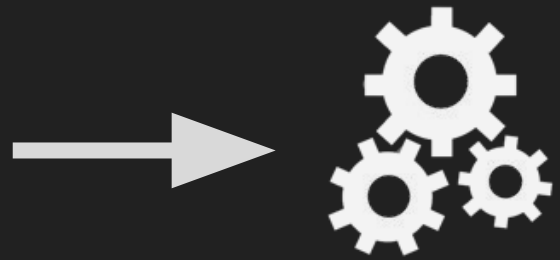
Fødselsnr.	Alder	Inntekt	Sivilstand=0	Sivilstand=4
01019112345	26	1.5	1	0
31121998765	98	-1.7	0	1
01010114689	16	-0.5	1	0
04110492843	13	-3	1	0

Anonymisere fødselsnummer

Fødselsnr.	Alder	Inntekt	Sivilstand=0	Sivilstand=4
k9ej3289f4	26	1.5	1	0
kf83idjmsc	98	-1.7	0	1
8274hngkd	16	-0.5	1	0
92infnkisdsj	13	-3	1	0

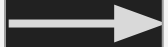
Modellen trenes på en viss andel av befolkningen...

Fødselsnr.	Inntekt 2013	Inntekt 2015	...
01019112345	500 000	600 000	
31121998765	2000	120 000	
01010114689	300	600	
04110492843	0	0	



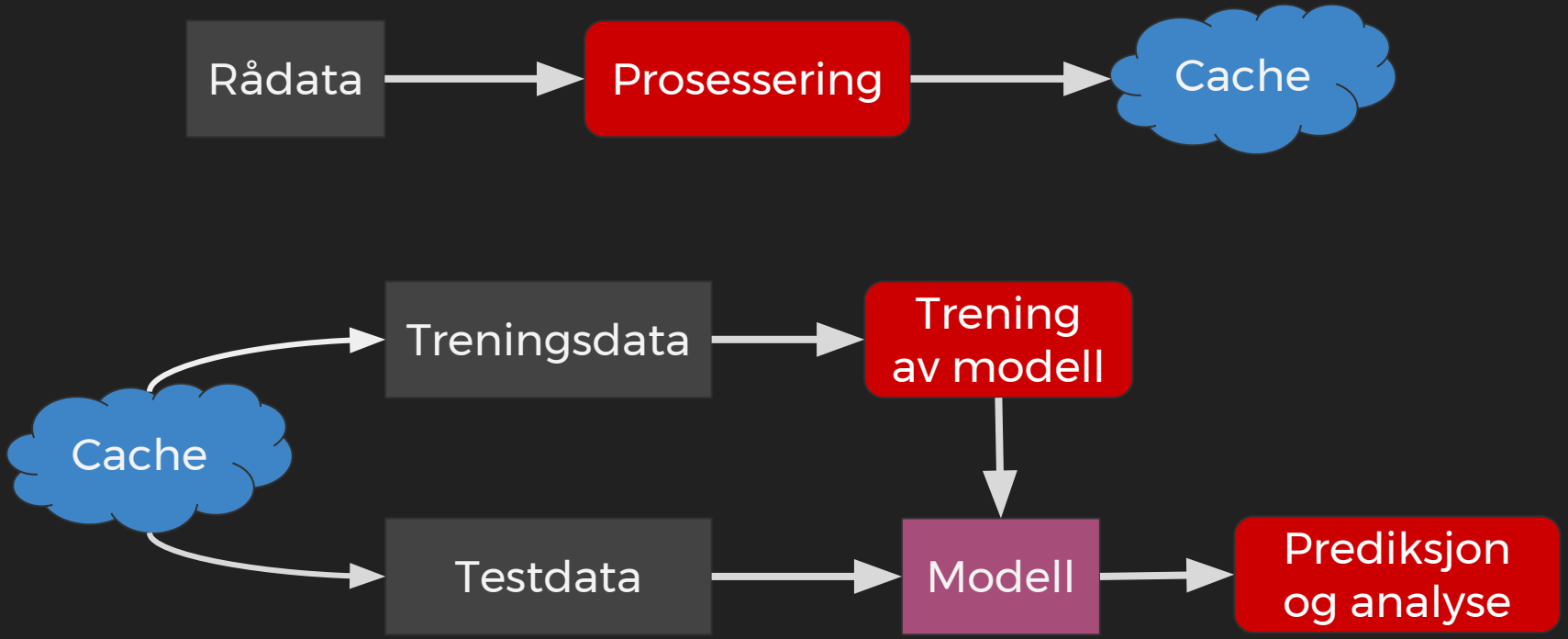
...og evalueres på en usett del av befolkningen

Inntekt 2013	**Gjeld 2013**	**...**
535 000	400 000	
379 000	14 000 000	
800 000	5	



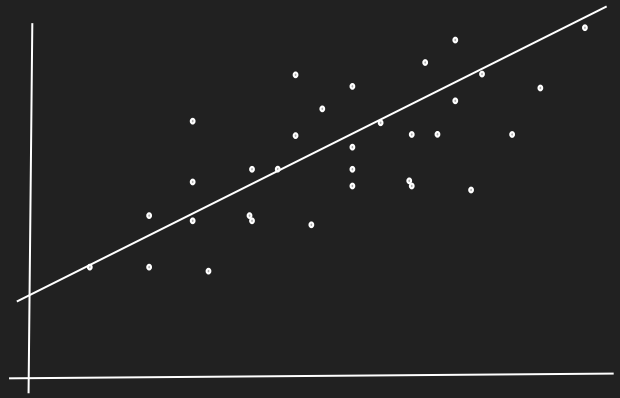
Inntekt 2015	**Gjeld 2015**	**...**
600 000	3 500 000	
???	???	
???	???	


scikit
learn


Rådata
Prosessering
Cache
Treningsdata
Trening
av modell
Cache
Testdata
Modell
Prediksjon
og analyse

Modeller

Lineær regresjon



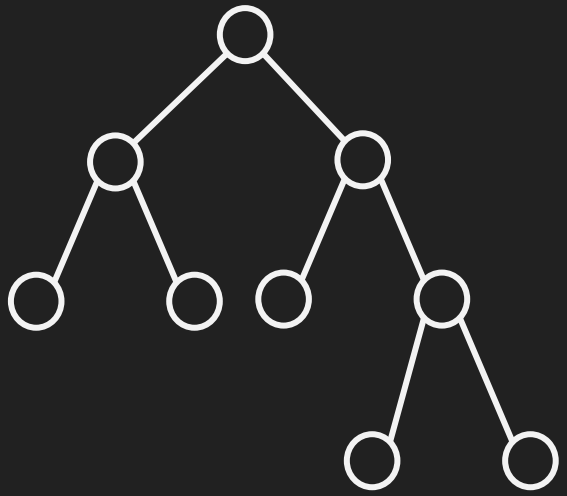
Fordeler:

- Selve definisjonen på god tolkbarhet
- Rask og enkel å bruke

Ulemper:

- Gav resultater som var lite tilfredsstillende

Random forest



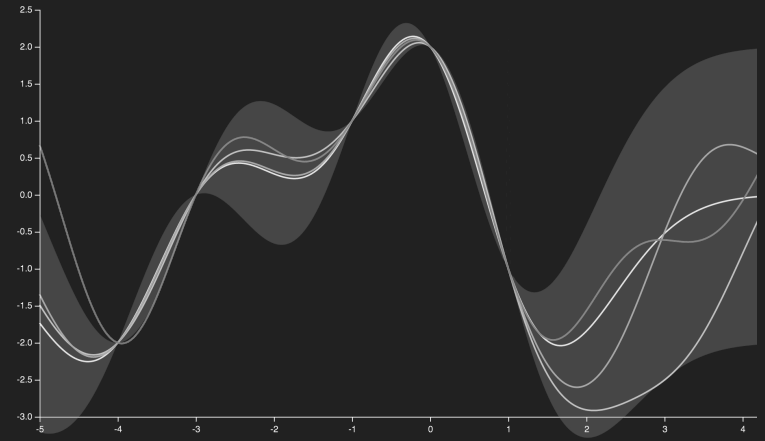
Fordeler:

- Større muligheter for å finne komplekse, ulineære sammenhenger
- Rask og få parametre å skru på
- Relativt tolkbar

Ulemper:

- Gav ikke optimale resultater, særlig på felt med mange 0-verdier

Gaussiske prosesser



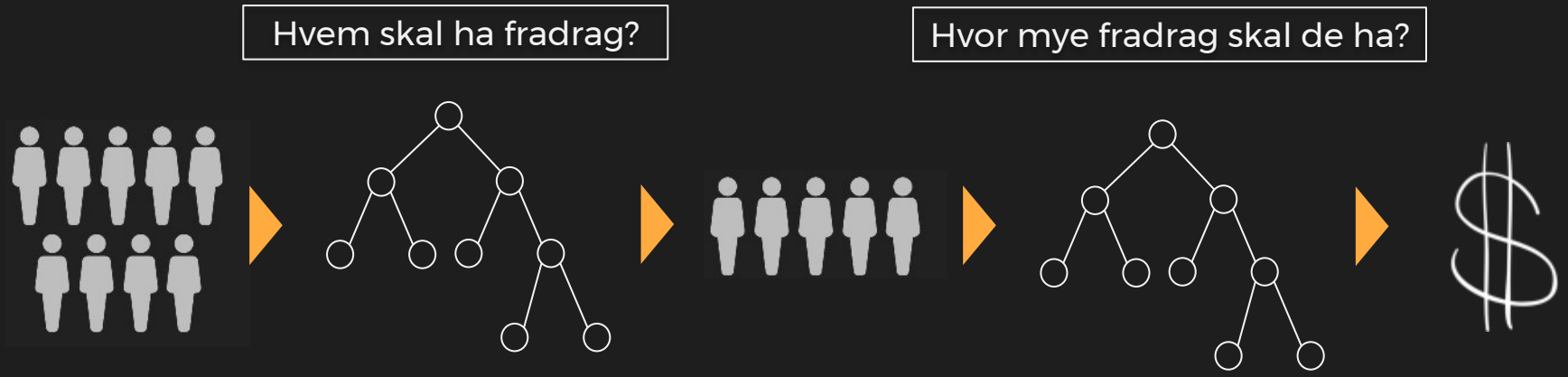
Fordeler:

- Stort matematisk rammeverk for å analysere oppførsel
- Gav mindre avvik på fradragsfelt med mange 0-verdier

Ulemper:

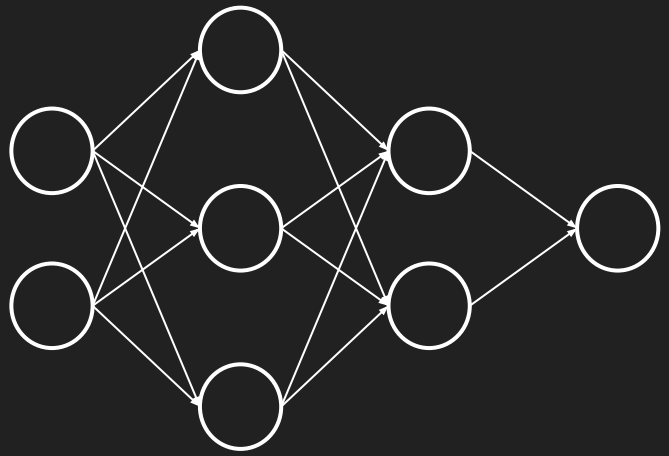
- Tung både teoretisk og i ytelse – mindre “plug and play” enn Random Forest

To-stegs-modeller



Gav mindre avvik for fradragsfelt som ofte er 0

Nevrale nettverk



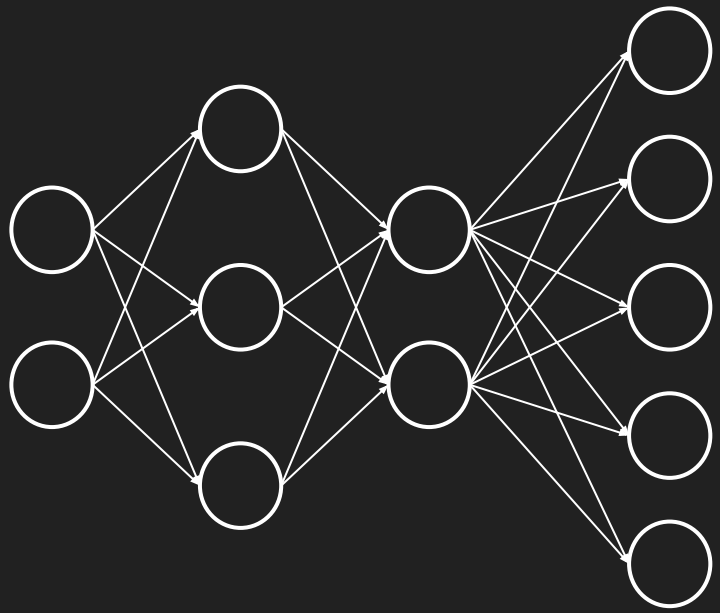
Fordeler:

- Raskt å komme i gang med
- Mest lovende teknikken i fagfeltet

Ulemper:

- Prediksjoner begrunnes ikke (“black box”)
- Krever noe erfaring for enkelte datasett

Modellvelgemodellen



Nevralt nettverk

Random forest

Gaussiske prosesser

Tostegs random forest

Sofus

Resultater

Fokuserte på et utvalg felter

1. Inntekt

2. Gjeld

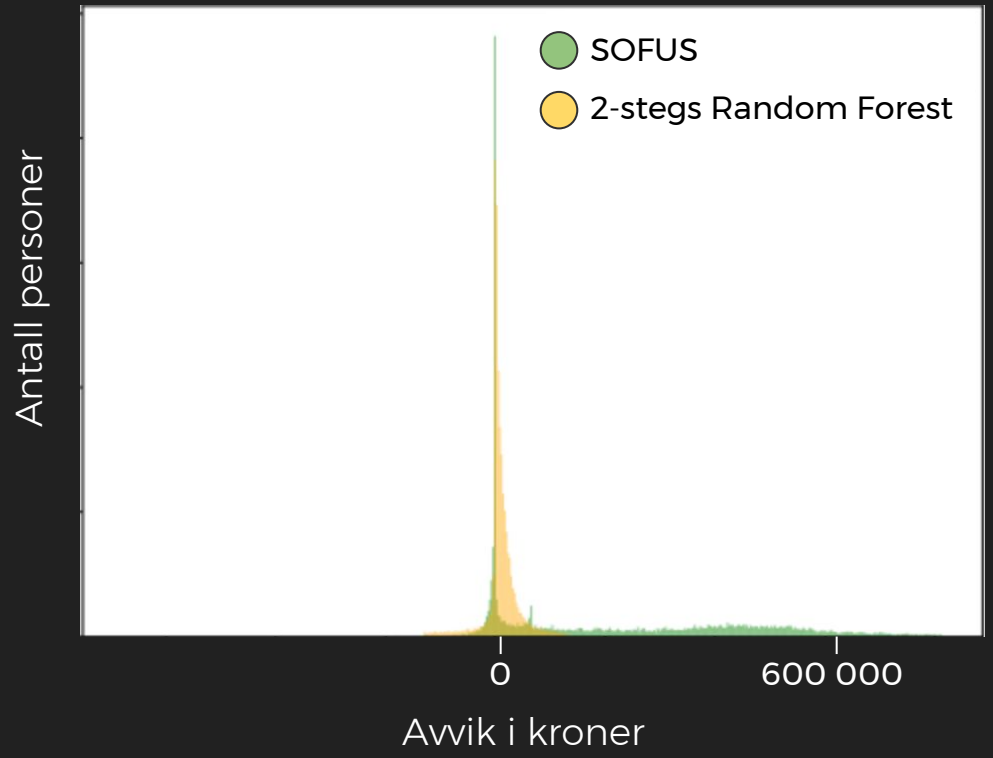
3. Rentefradrag

4. Inntektsfradrag pensjonsinnskudd og fagforeningskontigent

Sammenlignet med SOFUS

- ✓ Hvor stor del av befolkningen lager maskinlæringsmodellen en mer nøyaktig prognose enn Sofus?
- ✓ Hva er den relative reduseringen/økningen i avvik mellom faktisk skattegrunnlag og prognose?

Mål: gjette inntekt



Resultater per variabel

Variabel	Metode	% bedre enn Sofus	% likt
Inntekt	NN, 2RF	86%	0%
Rentefradrag	MVM	76%	0%
Fagfor. kont.	GP	33%	30%
Gjeld	GP	30%	20%

Prosessen, re-visited

1. Få tak i data (kaste terning)
2. Kverne data (bygge modeller)
3. Skape resultater (evaluere modellene)

Erfaringer

Ønsket tidsplan...

= 1 dagsverk

Forberede data

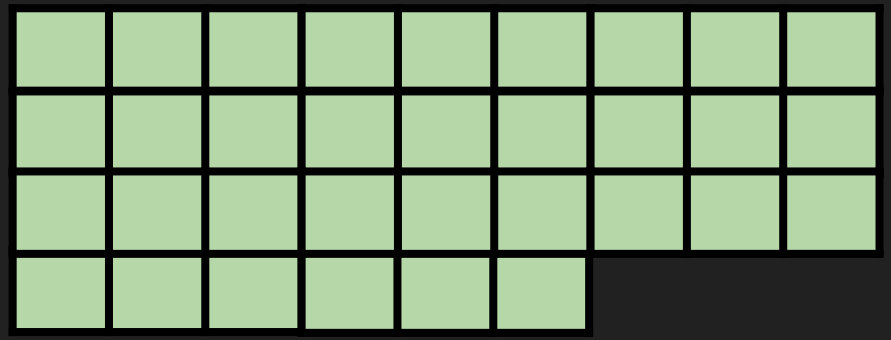
Kverne data

Skape resultater

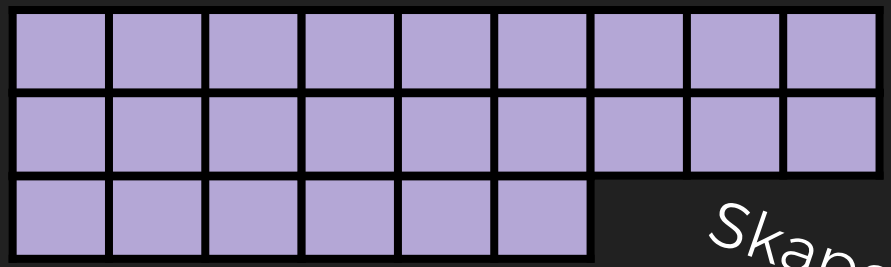
...faktisk tidsbruk

= 1 dagsverk

Forberede data



Kverne data



Skape resultater

Maskinlæring fordrer en tilpasset utviklingsprosess

Veien videre

Takk for oss.

Spørsmål?