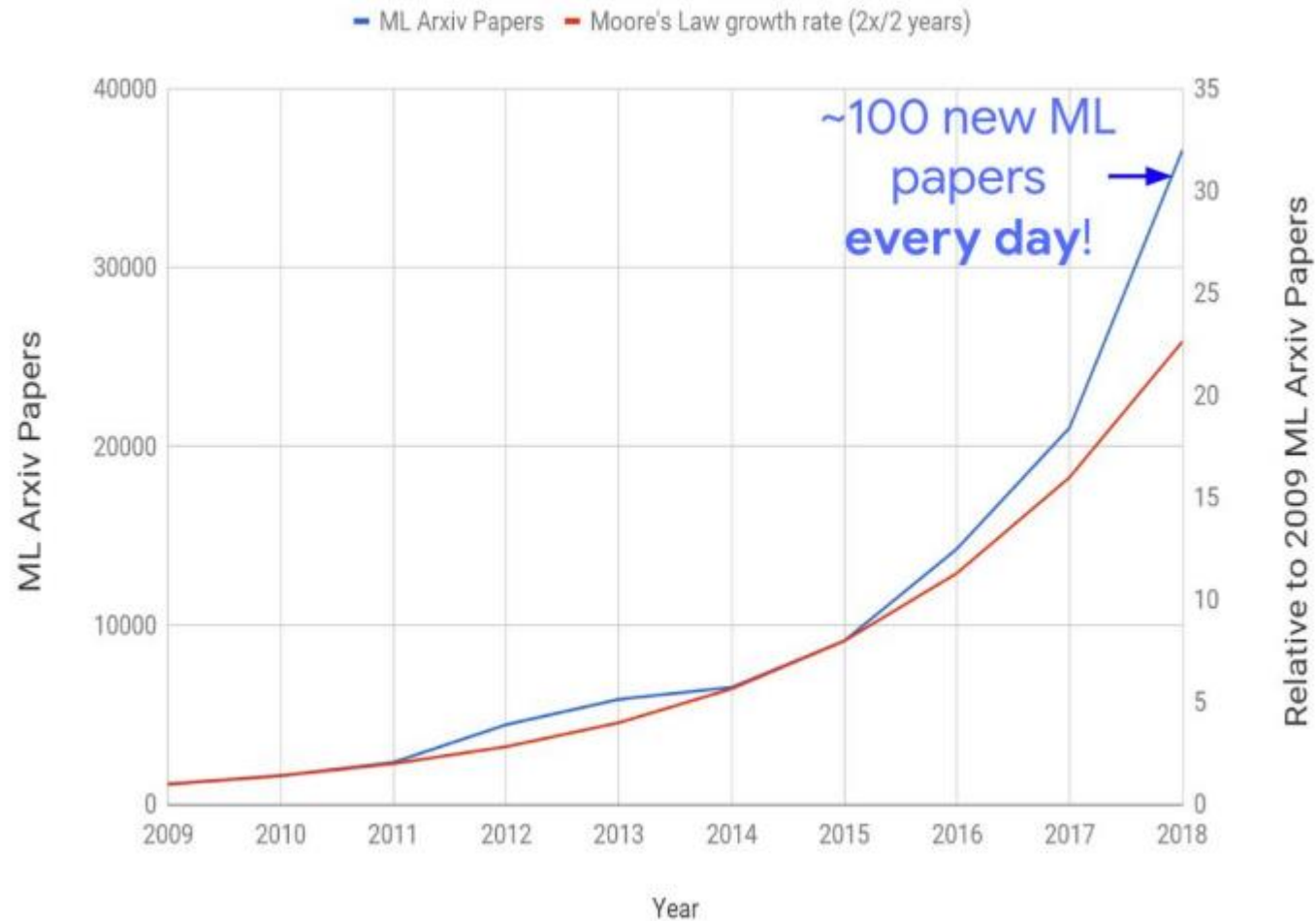


Minikurs i maskinlæring og AI

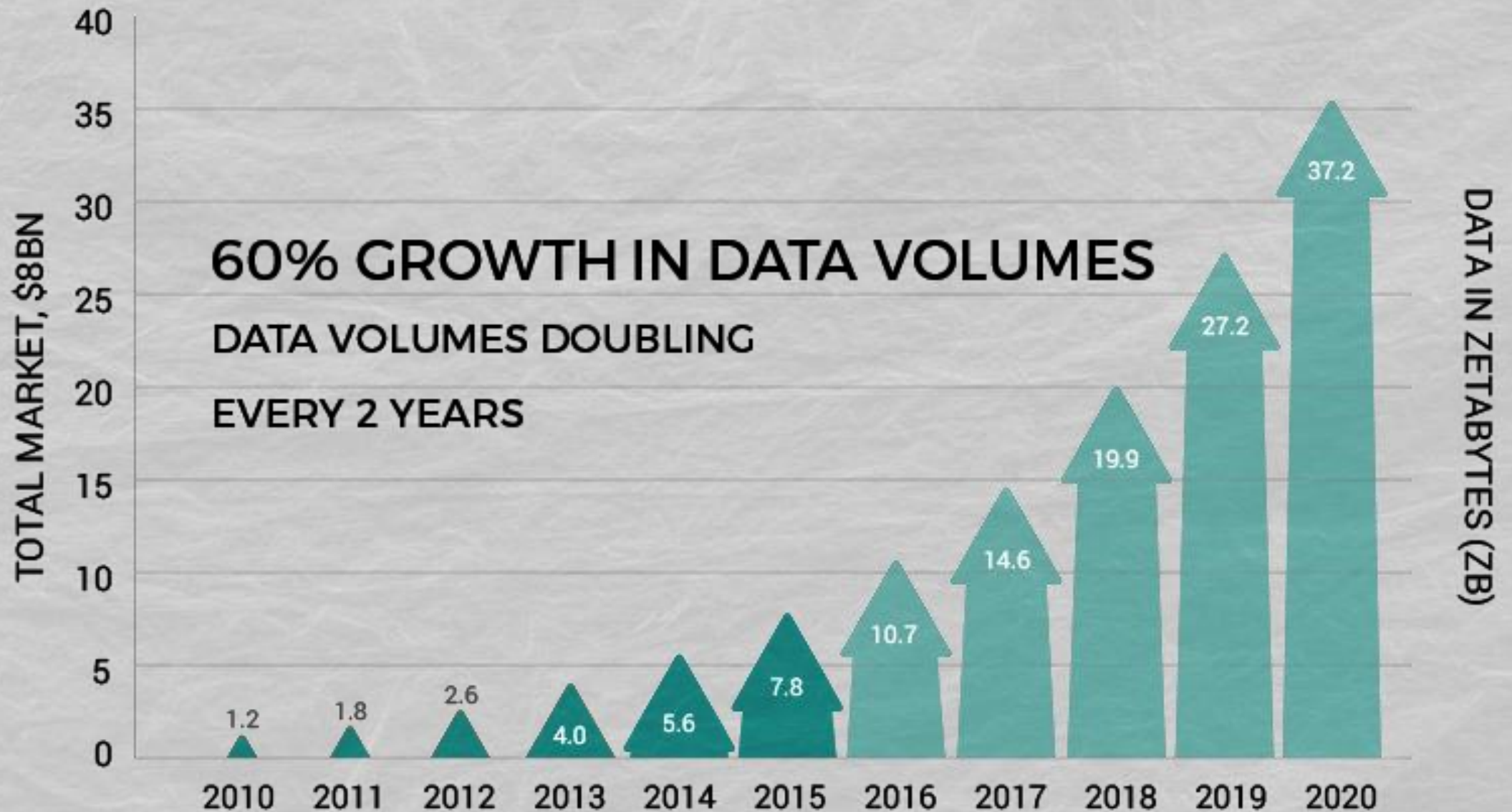
Kick-off 2022, Vegard Bjørgan & Andreas Thyholt Henriksen



Forskning innen maskinlæring

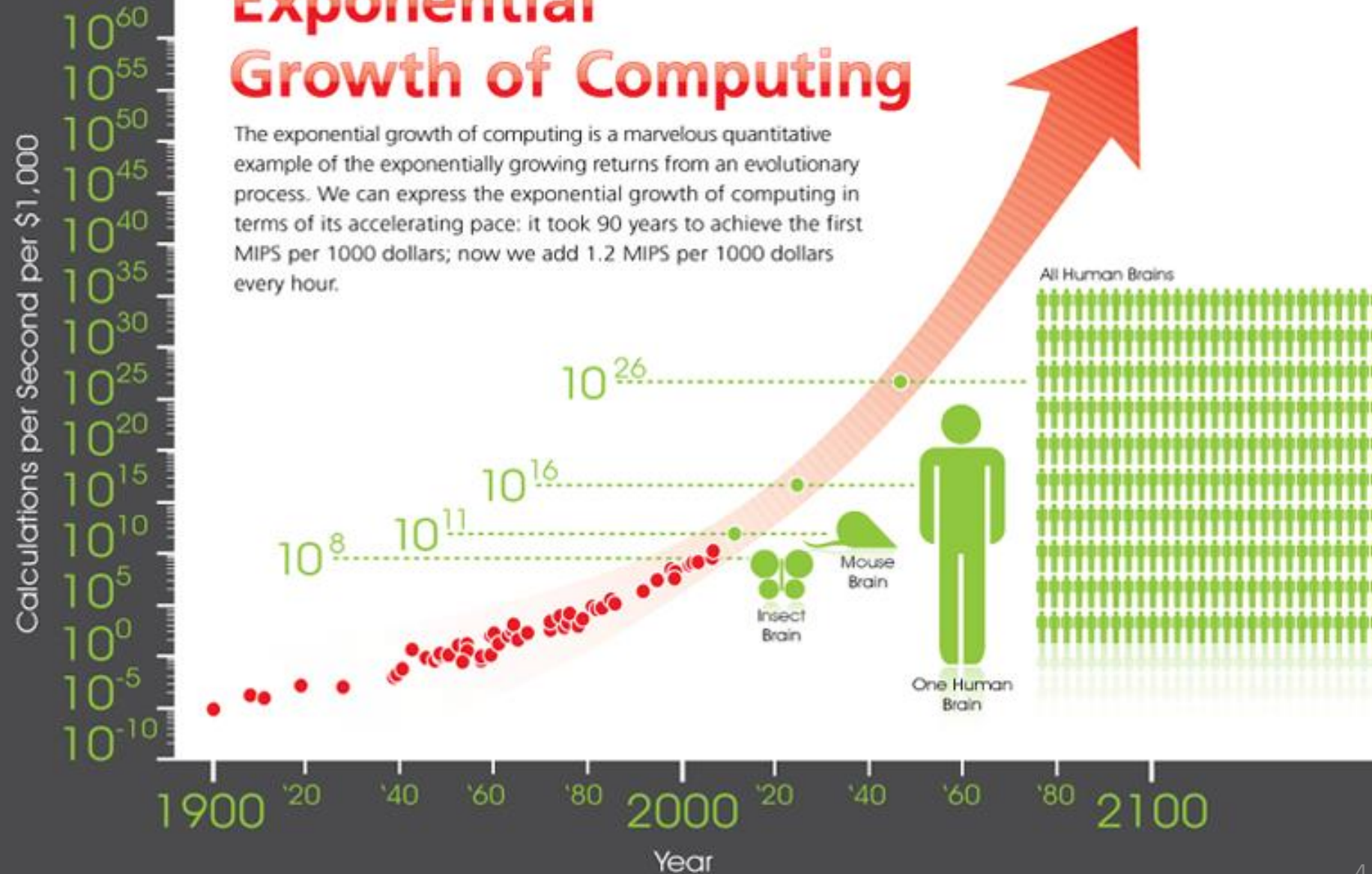


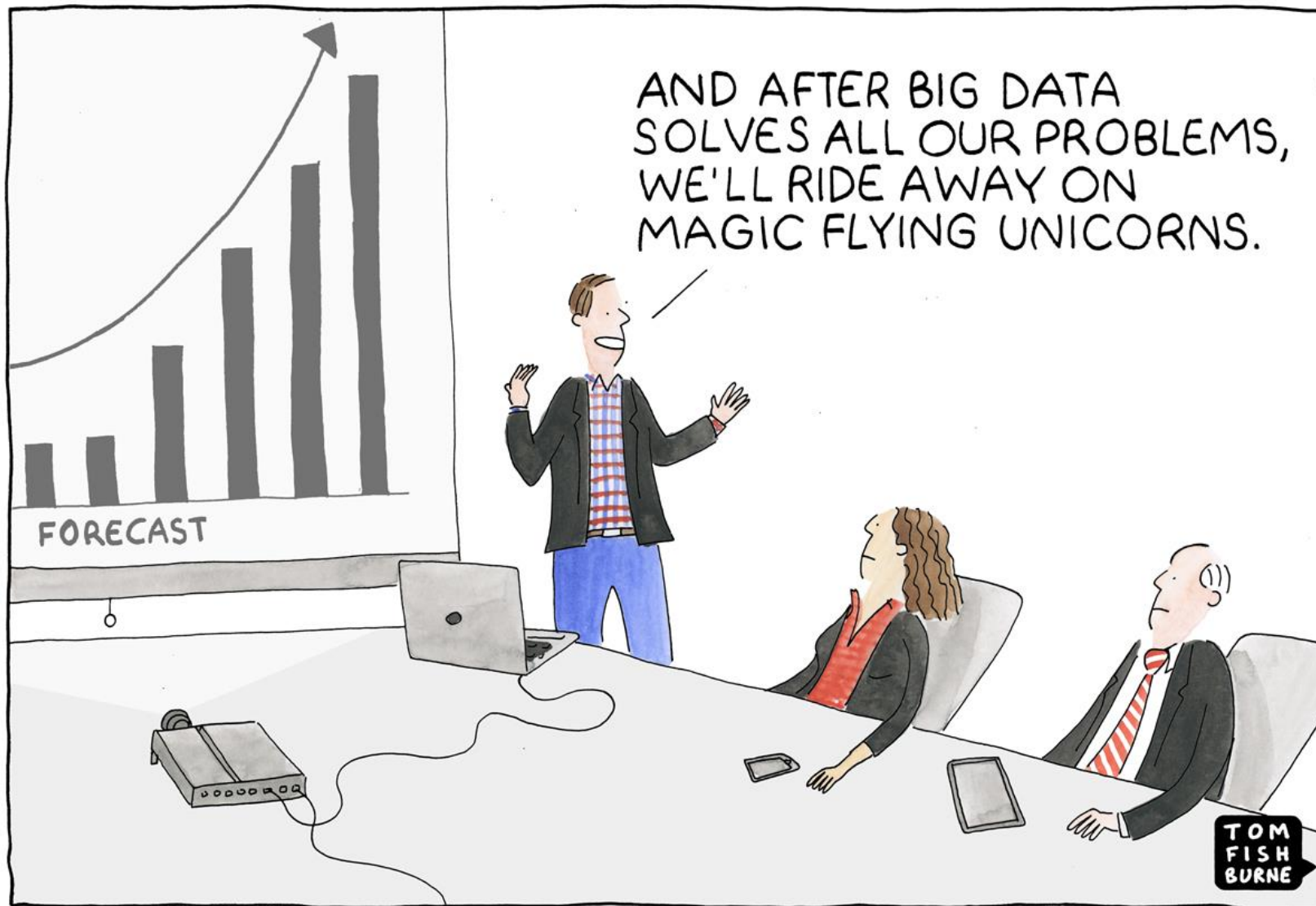
Data – en fornybar ressurs



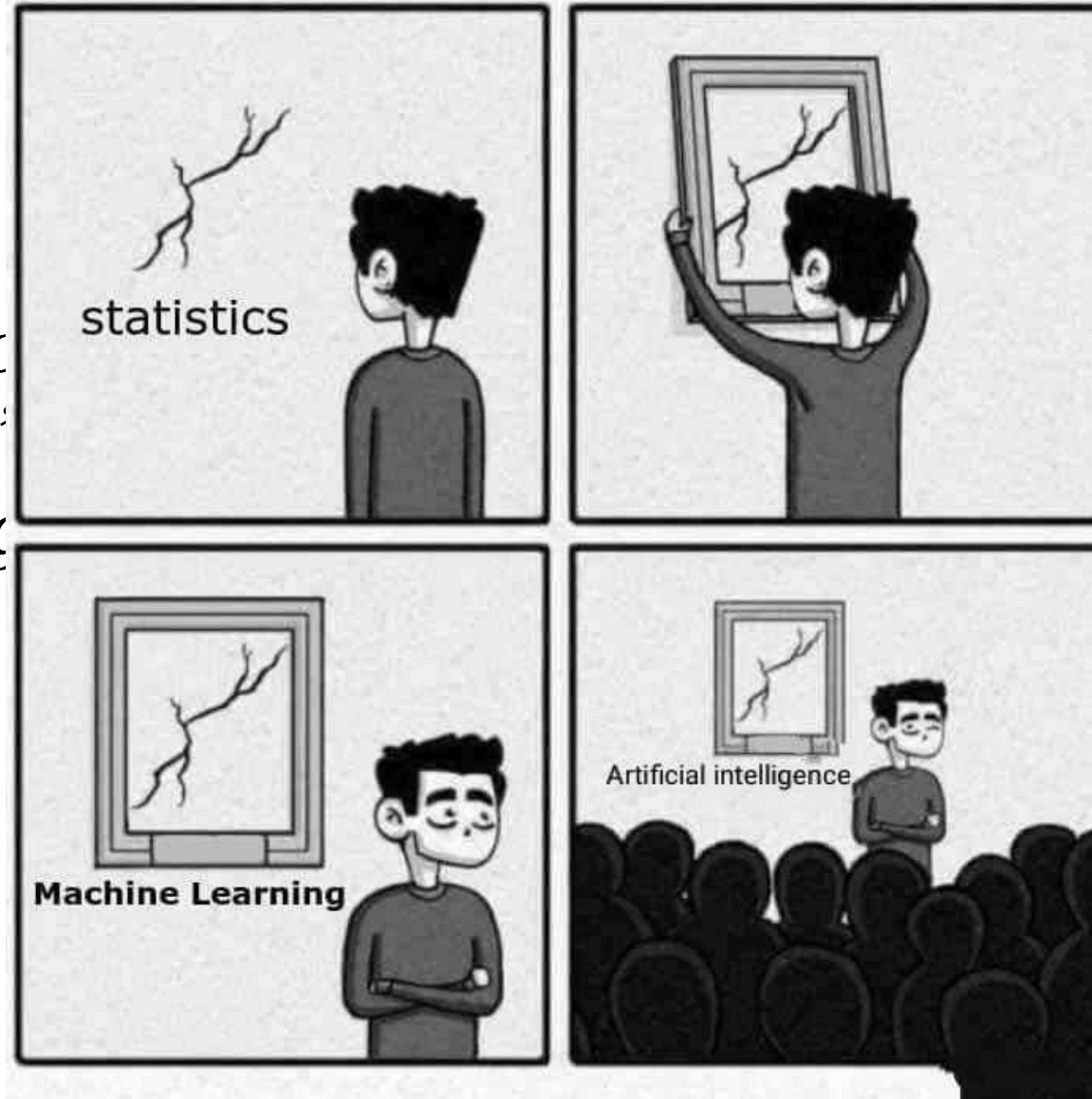
Exponential Growth of Computing

The exponential growth of computing is a marvelous quantitative example of the exponentially growing returns from an evolutionary process. We can express the exponential growth of computing in terms of its accelerating pace: it took 90 years to achieve the first MIPS per 1000 dollars; now we add 1.2 MIPS per 1000 dollars every hour.





“When you’re hiring, it’s not linear regression.”



you’re not hiring, it’s not linear regression.”

Hva er maskinlæring *egentlig*? Begreper

Traditional Programming

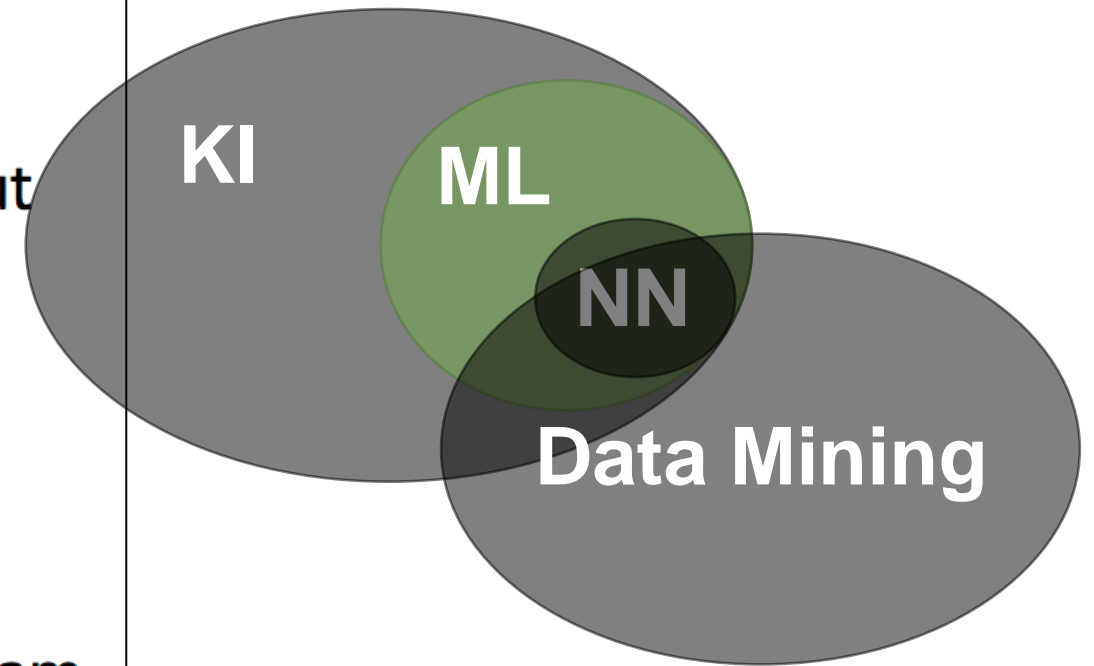


Machine Learning



Et dataprogram som blir bedre på en oppgave ved å trene på den

"Lær med eksempler, ikke med regler"



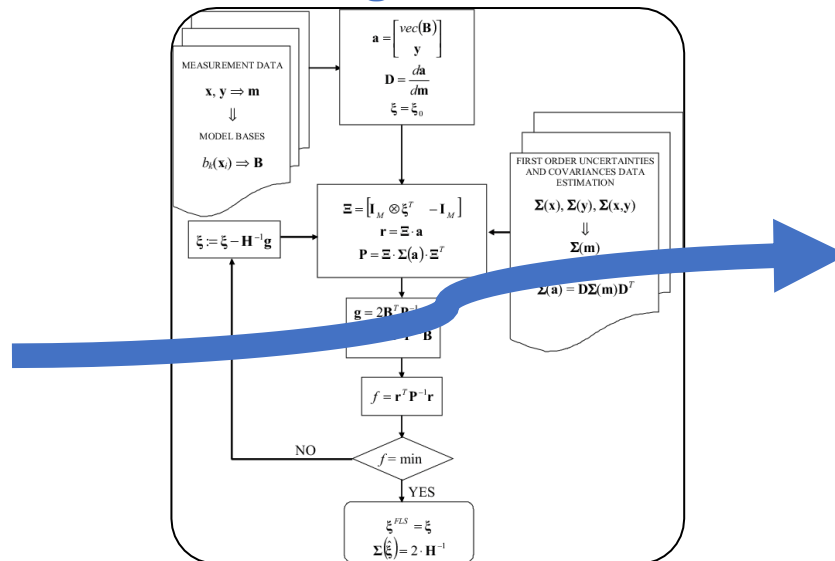
Modell

- En funksjon som en datamaskin kan bruke for å omgjøre data til en beslutning eller prediksjon
- En algoritme trener opp modellen på eksisterende data

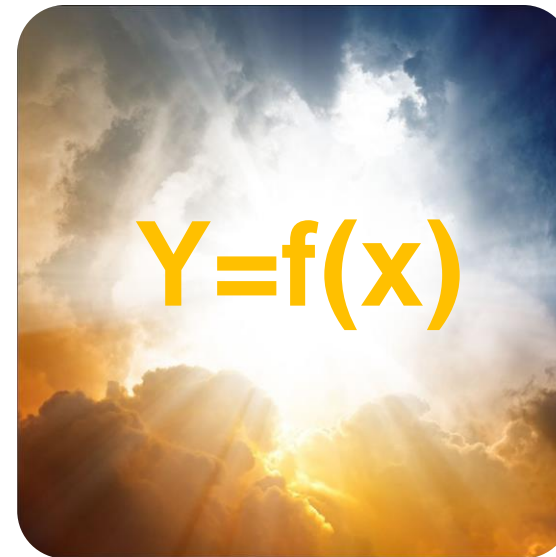
Eksisterende data



Algoritme

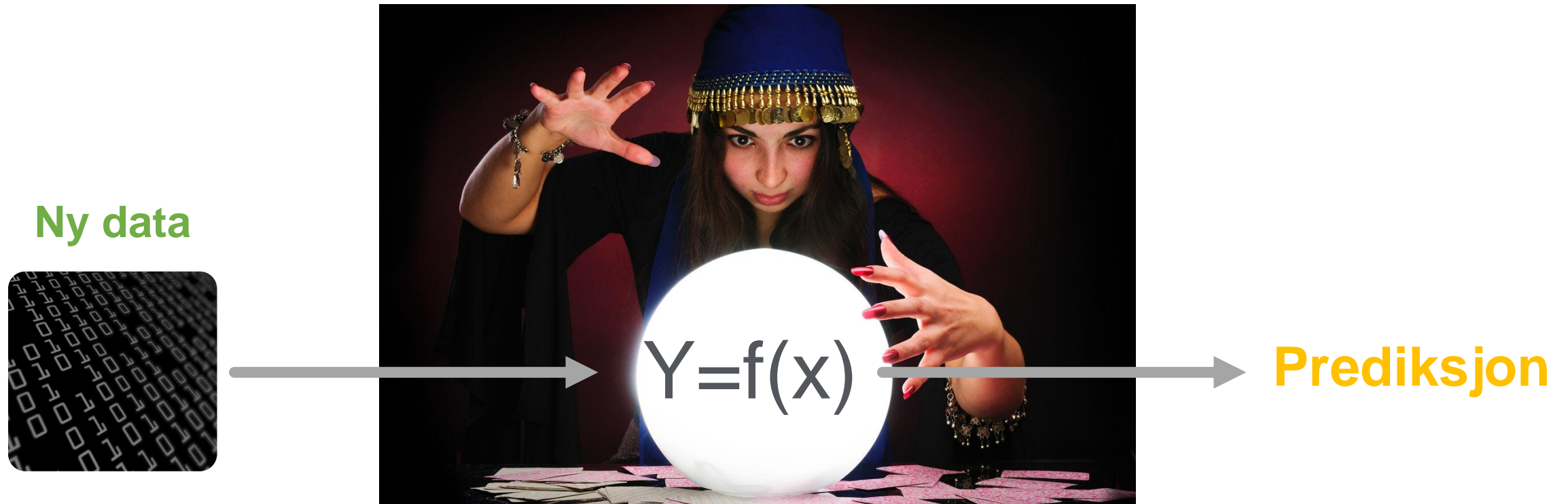


Modell

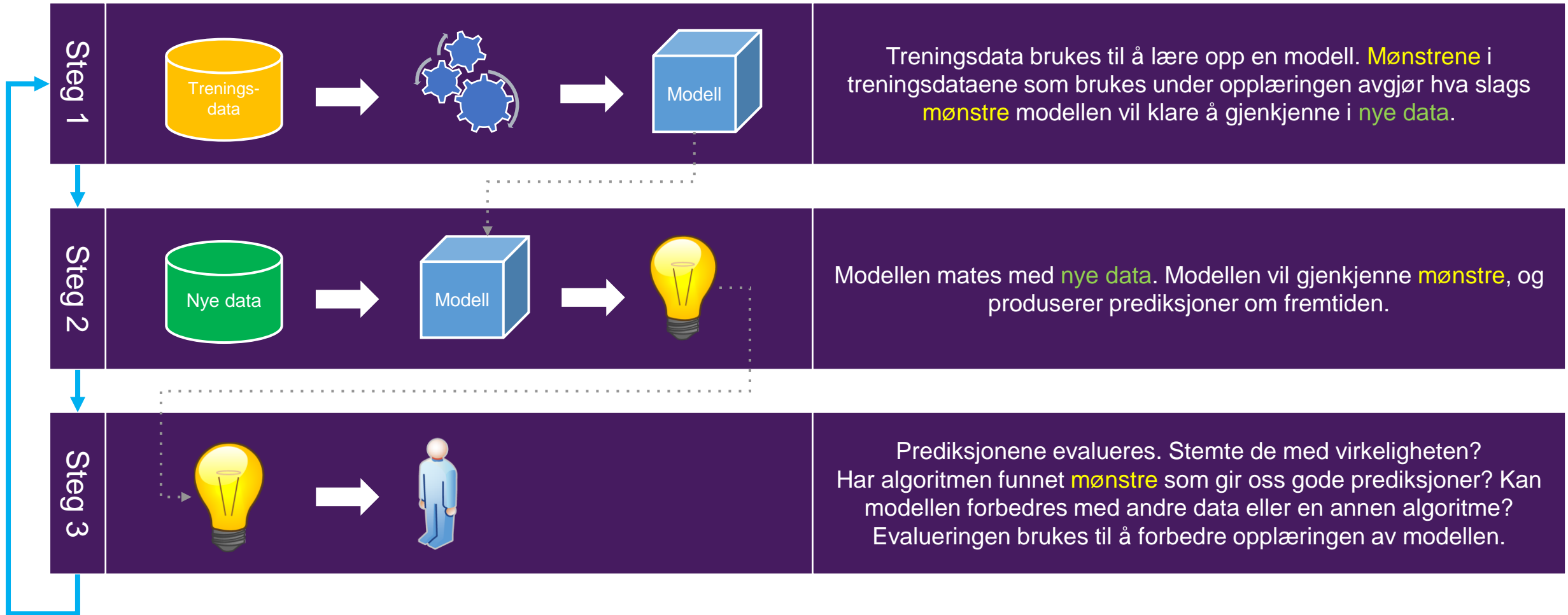


Prediksjon (eng: Prediction)

«Spådommen» til en modell (kvalifisert gjetting?)



Hvordan fungerer maskinlæring i praksis?



Eksempler

Observasjoner: radene i en tabell

StudID	Fornavn	Etternavn	Kjønn	Fødselsår	Karaktersnit t	Antall studiepoeng	Fikk jobb etter endt studieperiode
19203	Ola	Nordmann	M	1990	4.8	180	Nei
73729	Kari	Nordkvinne	K	1995	5.4	180	Ja
43923	Andreas	Kråkestad	M	1982	6.0	360	Ja
32423	Mari	Lie	K	1989	4.1	60	Nei
...

Egenskaper (Feature)

Egenskap: data vi vet om *observasjonen*, aka “uavhengig variable”

StudID	Fornavn	Etternavn	Kjønn	Fødselsår	Karaktersnit t	Antall studiepoeng	Fikk jobb etter endt studieperiode
19203	Ola	Nordmann	M	1990	4.8	180	Nei
73729	Kari	Nordkvinne	K	1995	5.4	180	Ja
43923	Andreas	Kråkestad	M	1982	6.0	360	Ja
32423	Mari	Lie	K	1989	4.1	60	Nei
...

Choosing the right thing to measure and getting the right metrics in place are extremely critical to succeeding with machine learning systems.
(Scott Clark, the founder and CEO of SigOpt)

Label ("output")

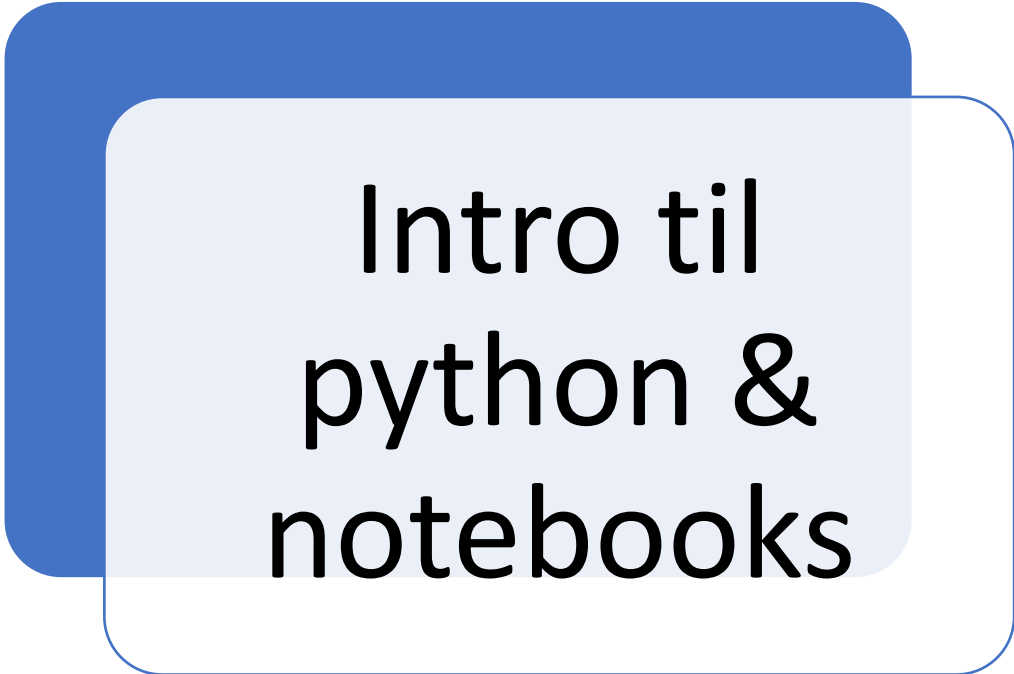
Fasit: den ønskelige outputen til modellen, aka "avhengig variable"

StudID	Fornavn	Etternavn	Kjønn	Fødselsår	Karaktersnit t	Antall studiepoeng	Fikk jobb etter endt studieperiode
19203	Ola	Nordmann	M	1990	4.8	180	Nei
73729	Kari	Nordkvinne	K	1995	5.4	180	Ja
43923	Andreas	Kråkestad	M	1982	6.0	360	Ja
32423	Mari	Lie	K	1989	4.1	60	Nei
...

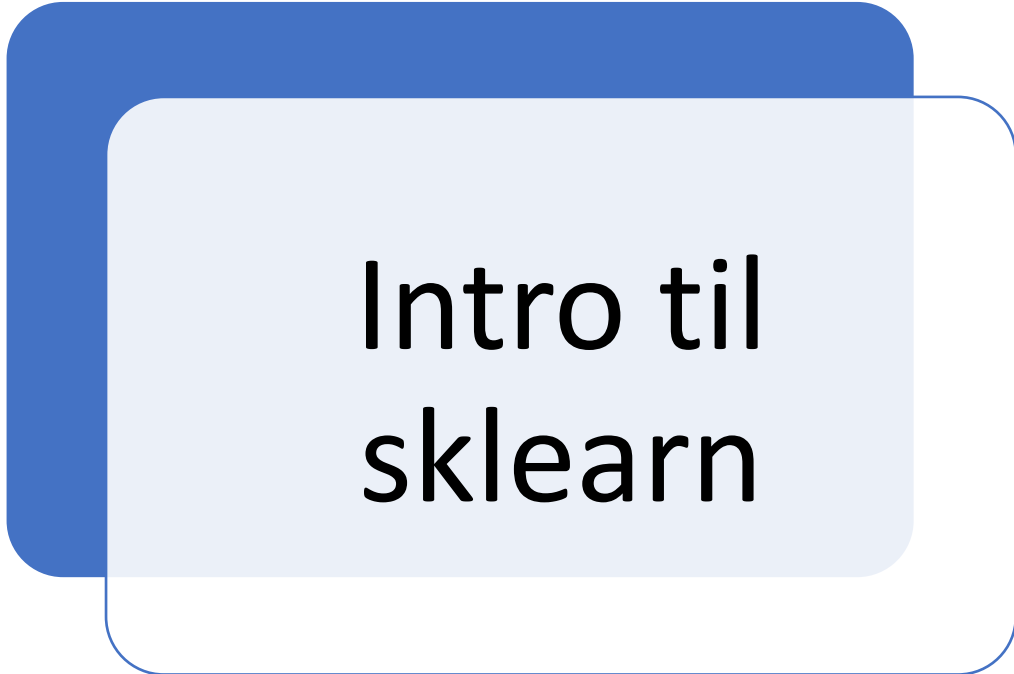
The key thing is that ML/AI is not magic and it doesn't solve every problem. It's a thing-labeler and it's up to you to figure out what you need labeled.
(Cassie Kozyrkov, Chief Decision Intelligence Engineer, Google.)



Eksempel notebook 1



Intro til
python &
notebooks



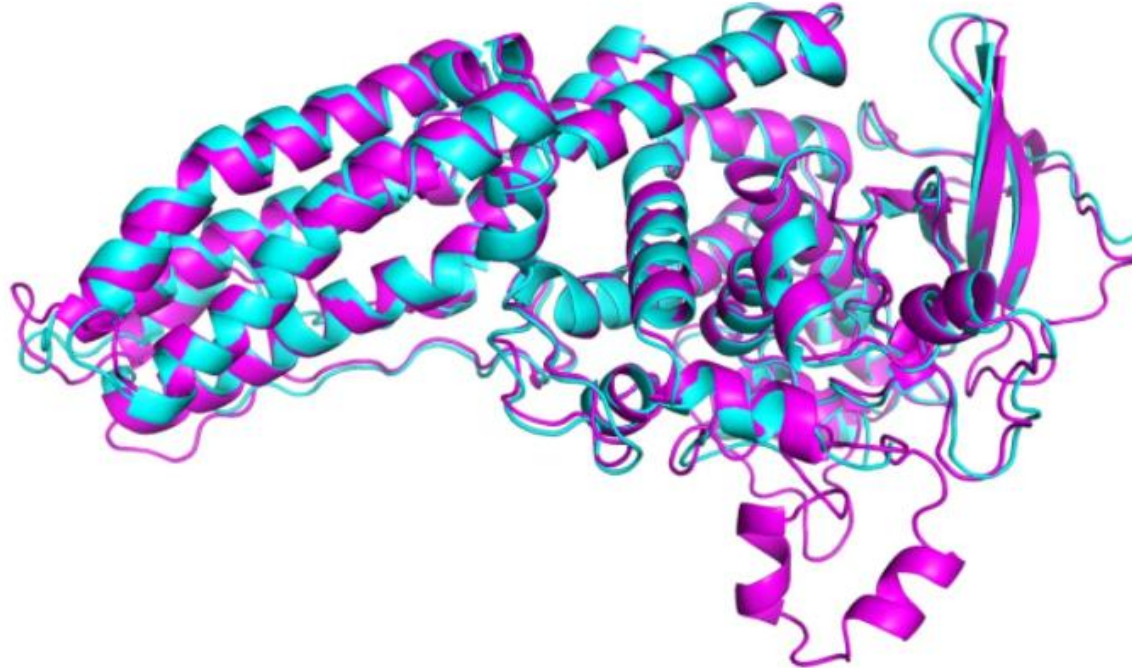
Intro til
sklearn

Anvendelser

Predikere proteinstruktur

DeepMind's protein-folding AI has solved a 50-year-old grand challenge of biology

AlphaFold can predict the shape of proteins to within the width of an atom. The breakthrough will help scientists design drugs and understand disease.



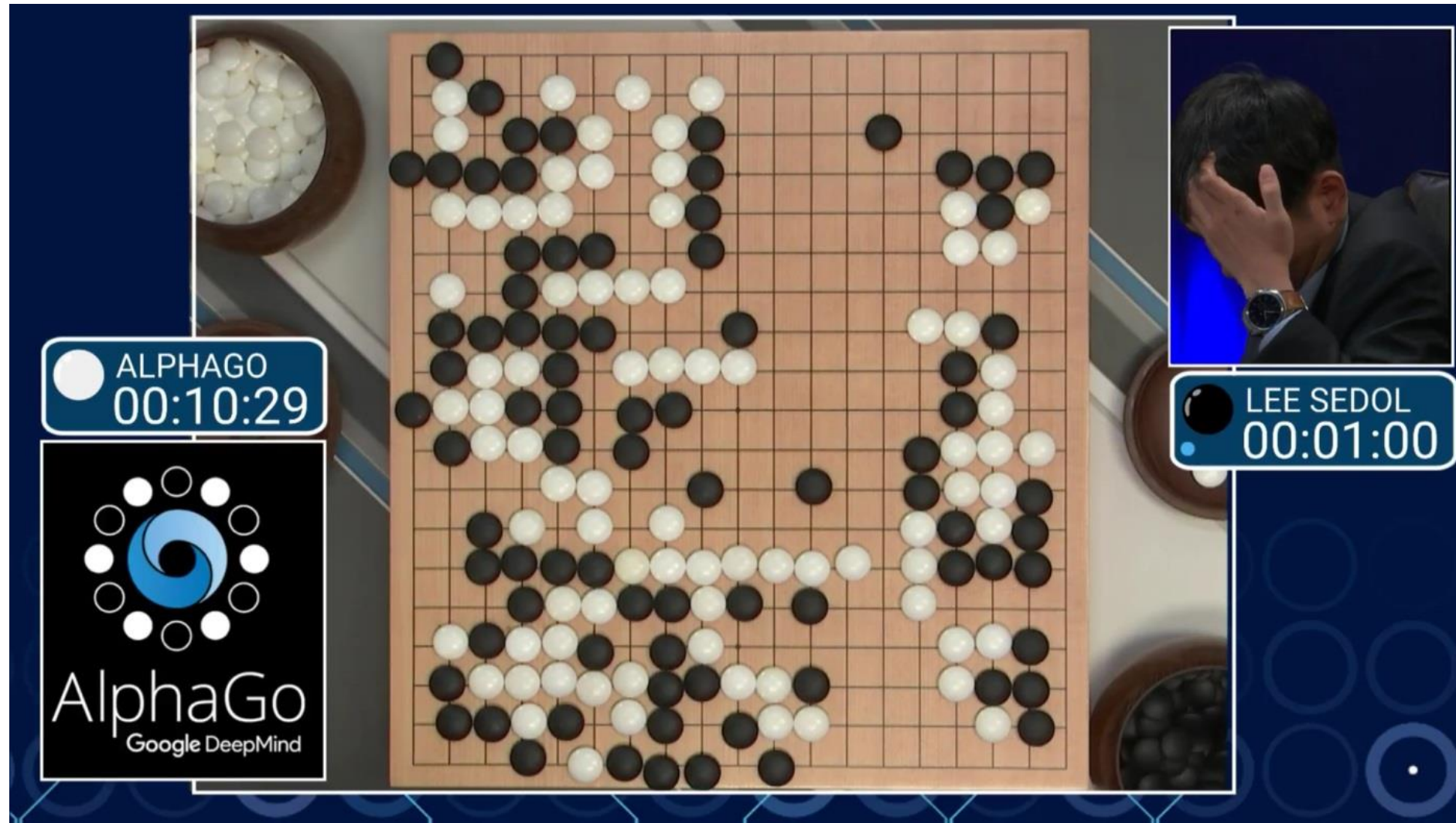
Selvkjørende biler



Naturlig språkprosessering



Slå verdens beste mennesker i Go



Slå verdens beste mennesker i Dota2

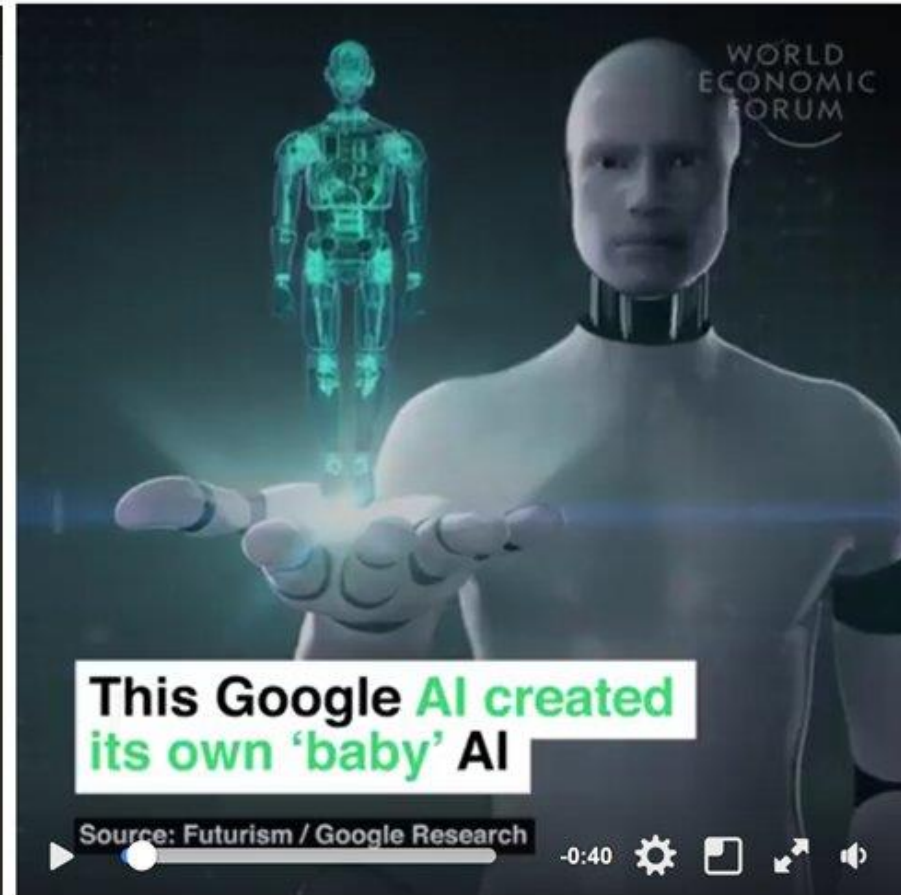


Hype?

Google Intern :

```
grid_search.py
1 from keras.layers import *
2 from keras.models import *
3 from .data import load_data
4
5 x, y, x_test, y_test = load_data()
6
7 def get_model(num_layers):
8     model = Sequential()
9     for _ in range(num_layers):
10         model.add(Dense(100, activation='sigmoid'))
11     model.compile(loss='mse', optimizer='sgd')
12     return model
13
14 best_model = None
15 best_loss = None
16
17 for i in range(1, 10):
18     model = get_model(i)
19     model.fit(x, y)
20     loss = model.evaluate(x_test, y_test)
21     if best_loss is None or loss < best_loss:
22         best_loss = loss
23         best_model = model
24
```

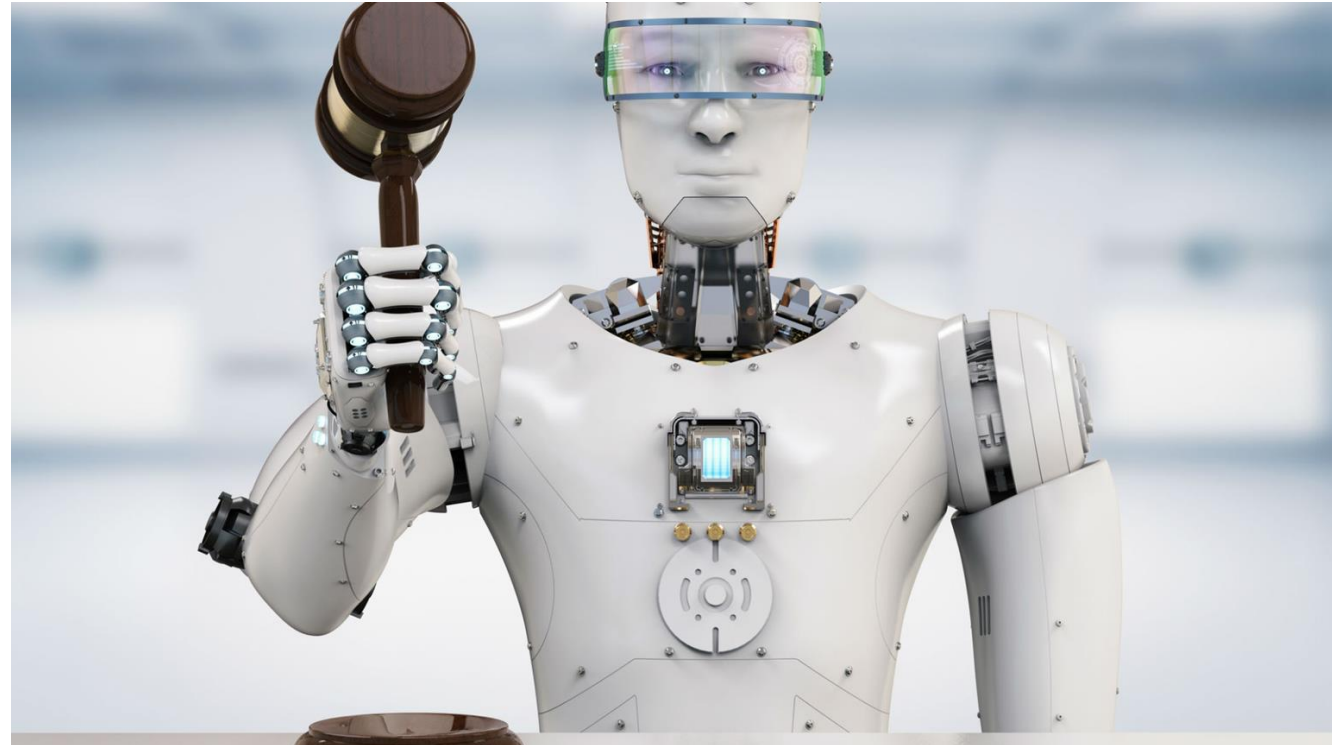
Media :



Domstoladministrasjonen

Hypoteser om anvendelser:

- Rettstolk: sanntidsoversettelse til/fra norsk
- Dommerassistent: produsere utkast til dom basert på tidligere avgjørelser
- Robotutreder: rettskildesøk, sammenstilling av momenter etc.
- Berammingsrobot: planlegge rettsmøter (personer, rom, etc)
- Dommerrobot: avsi avgjørelser automatisk
- Effektiviseringsrobot: oppdage saker som står i fare for å bli forsinket
- Chatbot for publikum
- Chatbot for selv-prosederende parter og profesjons-utøvere



Pågående AI-initiativer i NAV

	Oppfølging av arbeidssøkende		Klassifisering av næringskoder og yrkeskoder
	Innsikt og prediksjoner knyttet til sykefravær		Maskinlæring for produksjon av syntetiske testdata
	Analyse av uføreforløp		«Process mining» for prosessinnsikt
	Klassifisering og beriking av arbeidsmarkedsdata		Identifisering av feilutbetalinger

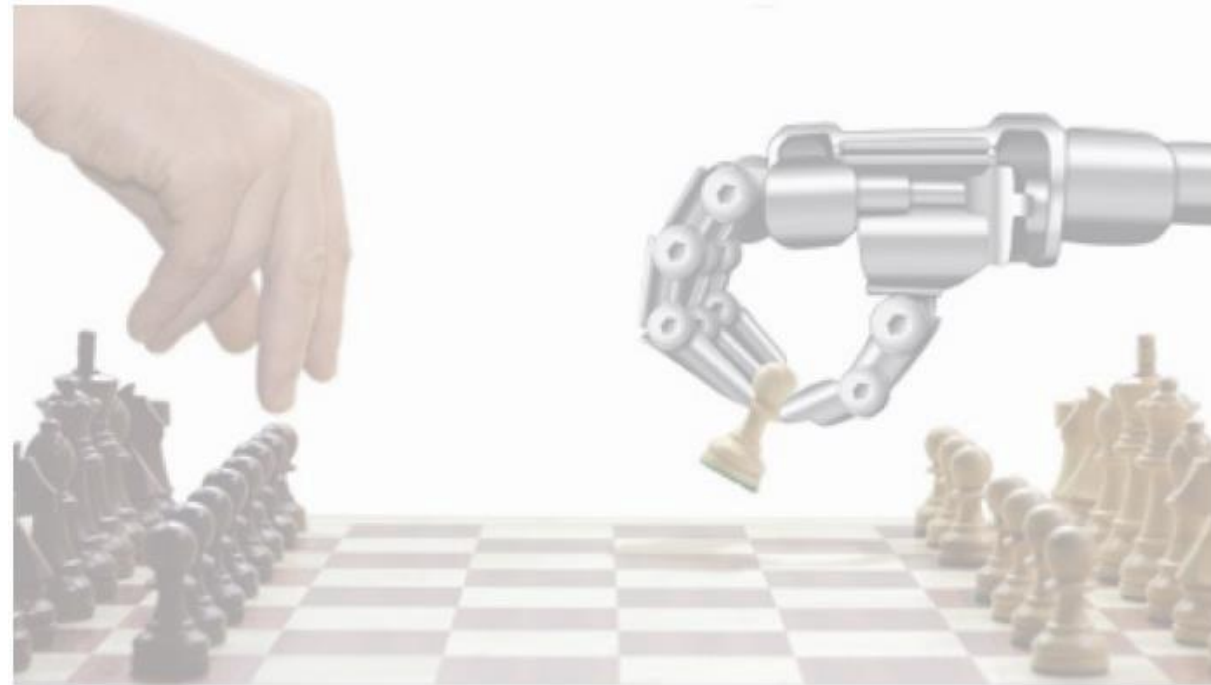
Stordatalab 2.0

Eksempler fra kundeprosjekter

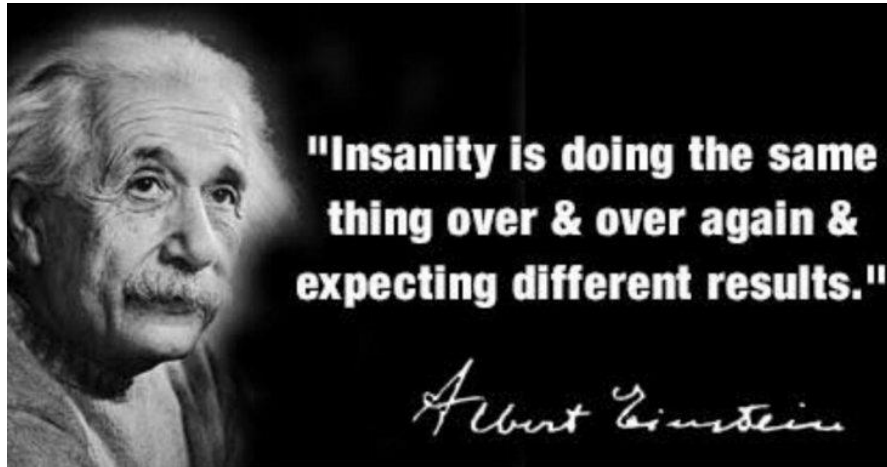


AI og samspill med mennesker

- Utfylle hverandre
- Langt igjen til AGI
- Bedre på noen ting
- Enkelte ting er det langt igjen før maskiner kan gjøre like godt som mennesker



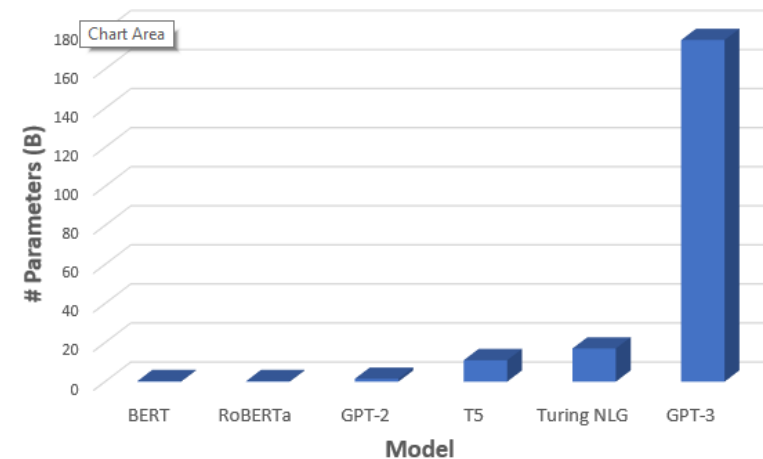
State of the art – moar data



*MACHINE LEARNING:



GPT-3 is great milestone in the artificial intelligence community. GPT-3 scraped almost every text data on the internet. "One API to rule many"



Human brain cells in a dish learn to play Pong faster than an AI

Hundreds of thousands of brain cells in a dish are being taught to play *Pong* by responding to pulses of electricity – and can improve their performance more quickly than an AI can



MIND 17 December 2021

By Michael Le Page

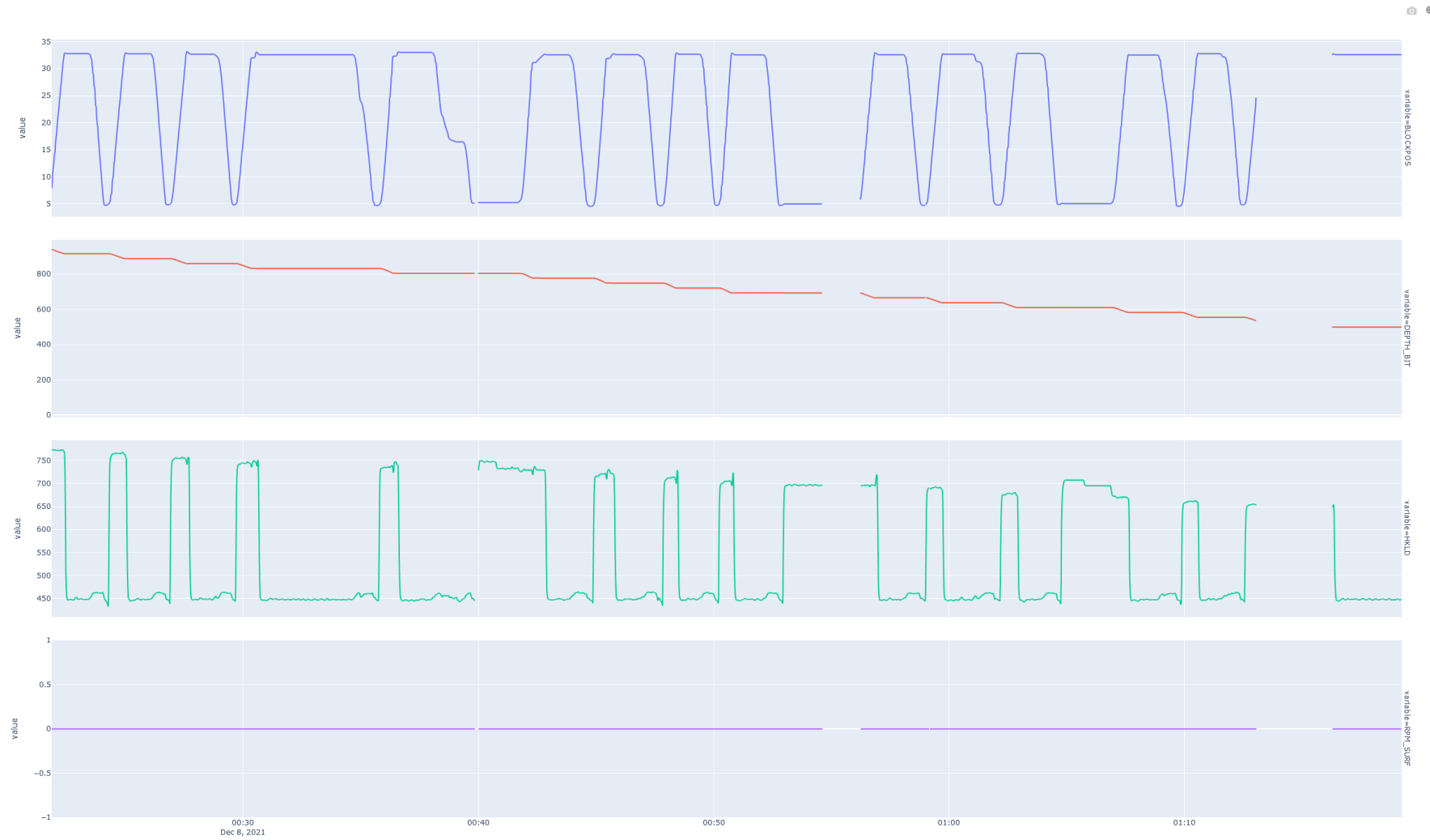


Datatypes

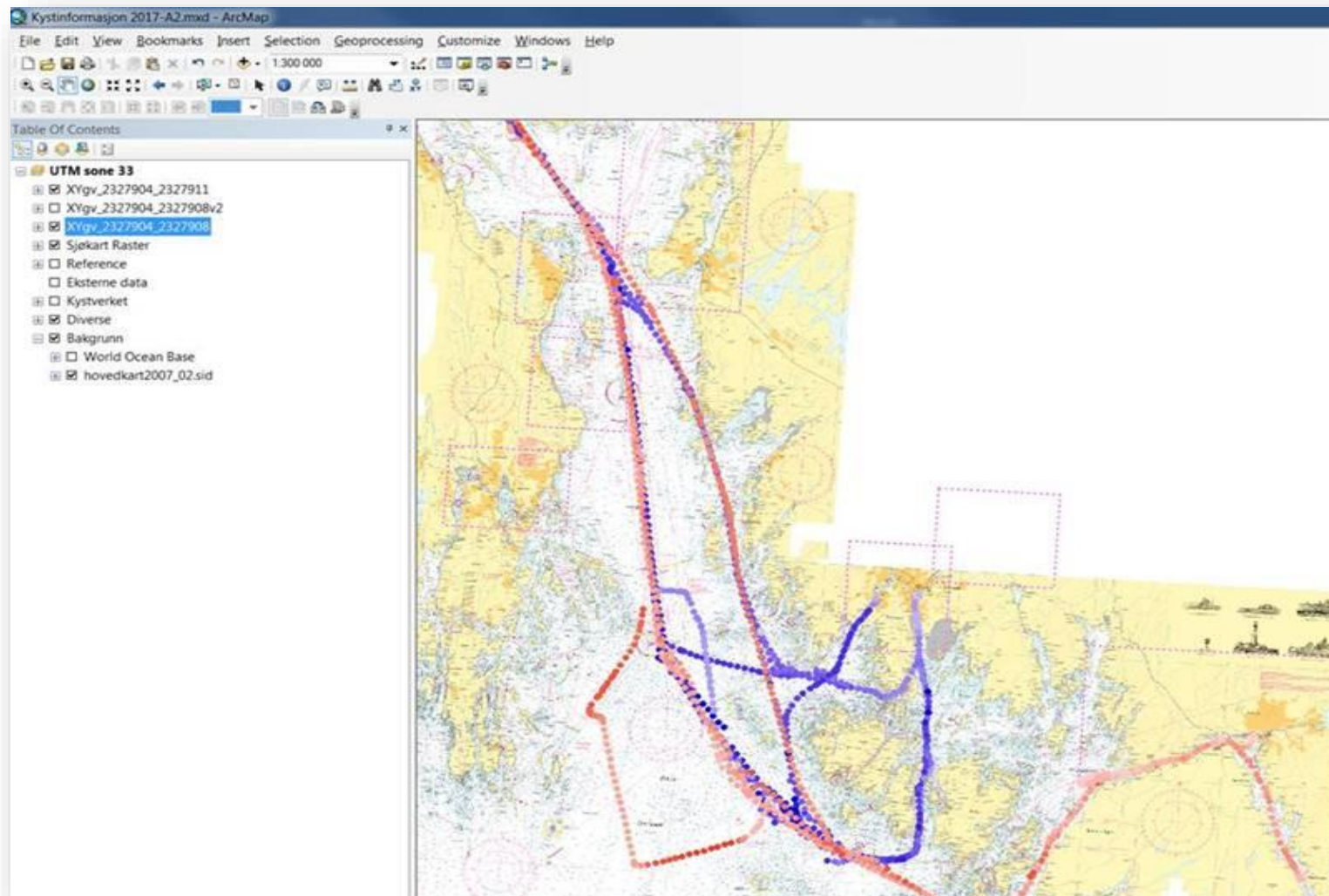
Tabelldata

StudID	Fornavn	Etternavn	Kjønn	Fødselsår	Karaktersnit t	Antall studiepoeng	Fikk jobb etter endt studieperiode
19203	Ola	Nordmann	M	1990	4.8	180	Nei
73729	Kari	Nordkvinne	K	1995	5.4	180	Ja
43923	Andreas	Kråkestad	M	1982	6.0	360	Ja
32423	Mari	Lie	K	1989	4.1	60	Nei
...

Tidsseriedata



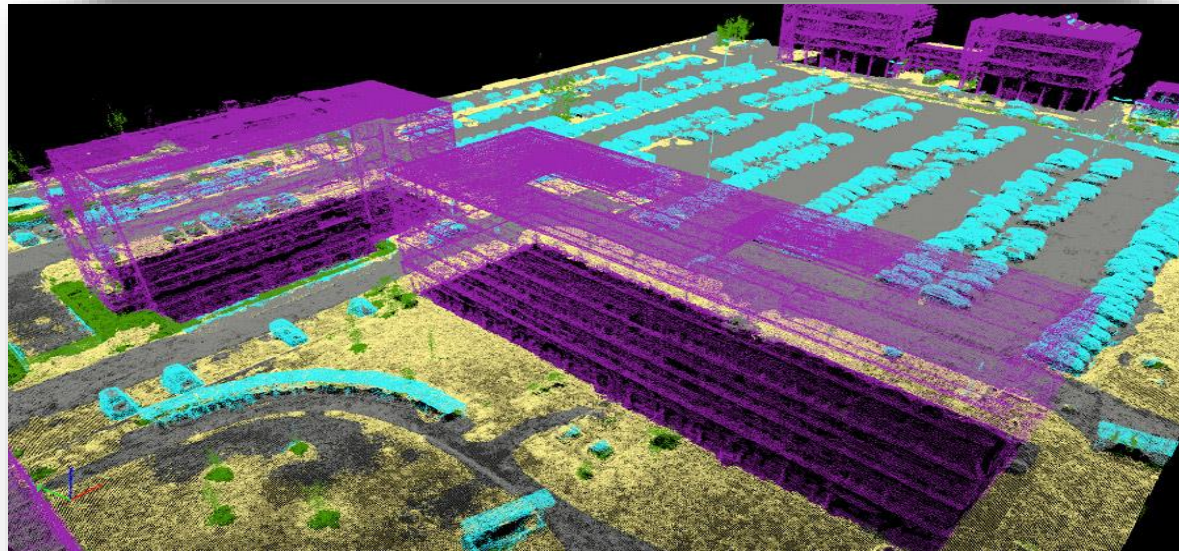
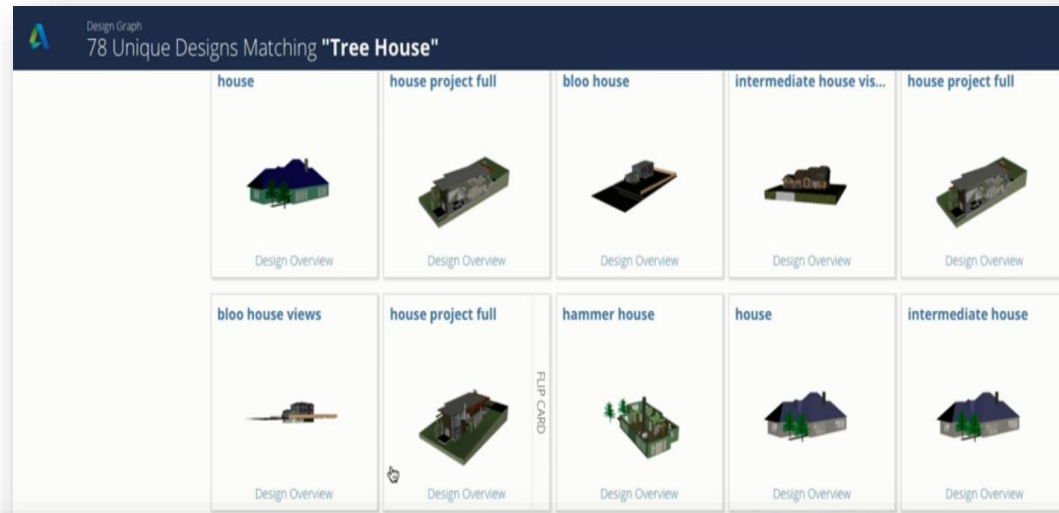
Geodata



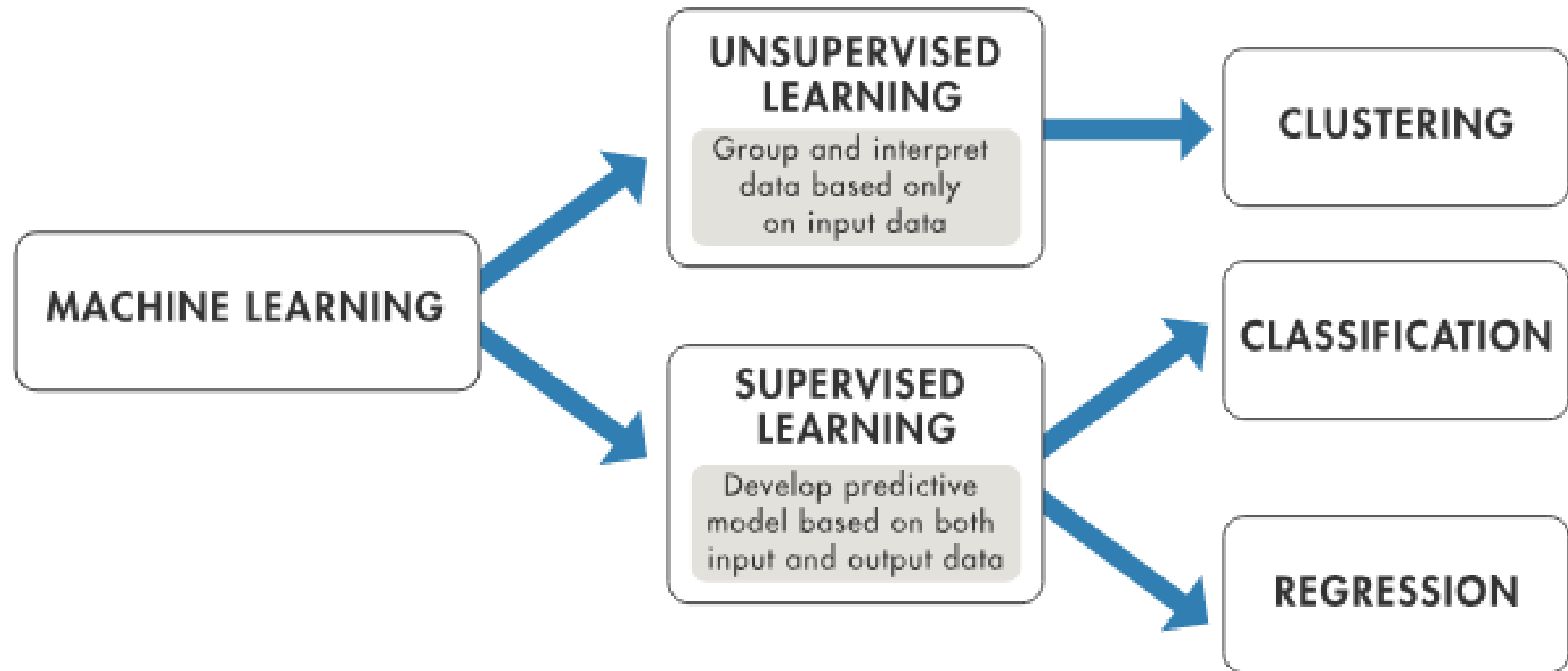
Tekstdata

```
{
  "_id": "5a4b2d21931f106509512881",
  "index": 0,
  "guid": "ceb73450-31a6-44ae-a51a-ea9aa07a5ca4",
  "isActive": false,
  "registered": "2015-02-09T02:33:43 -01:00",
  "text": "Hei, kan Norconsult hjelpe meg mitt firma med rådgiving ang. et nytt byggeprosjekt?"
},
{
  "_id": "5a4b2d21931f106453456436",
  "index": 0,
  "guid": "ceb76550-31a6-44ae-a51a-ea9aa07a6ca2",
  "isActive": false,
  "registered": "2016-02-09T02:33:43 -01:00",
  "text": "Jeg trenger umiddelbar hjelp med ISYPlant. Dette haster!!!"
},
```

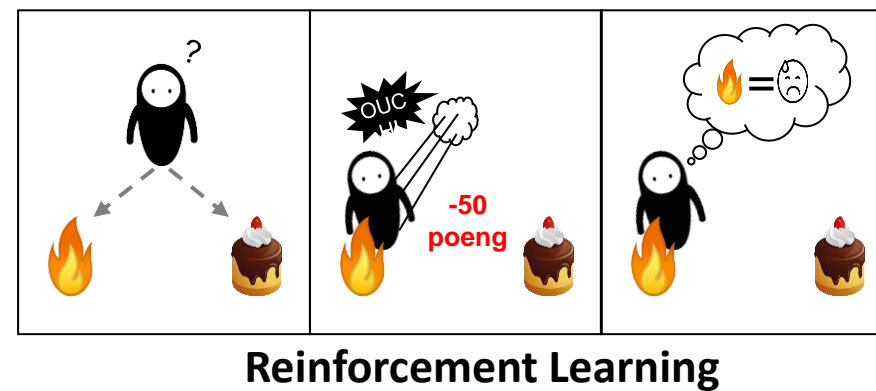
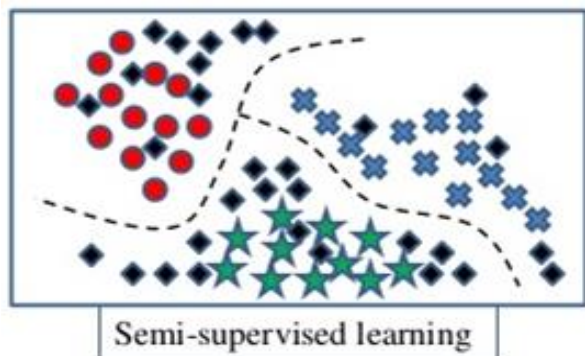
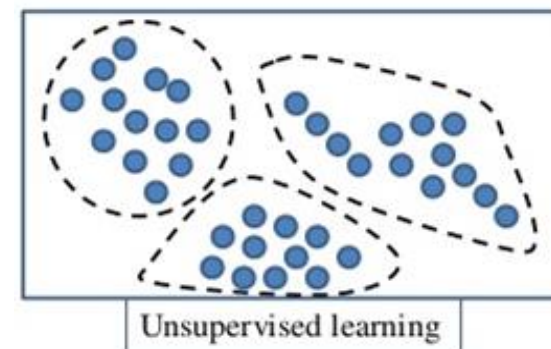
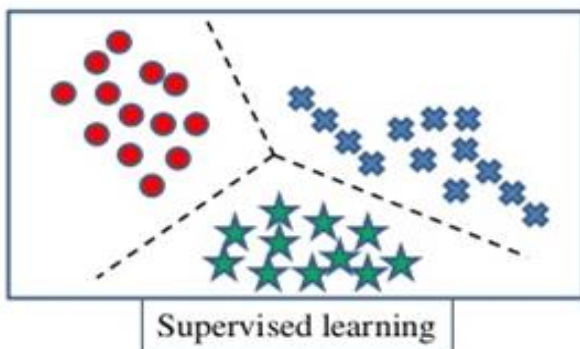
Multimedia-data (bilder, video, 3D)



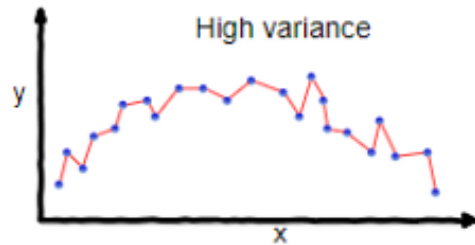
Maskinlæringsmetoder



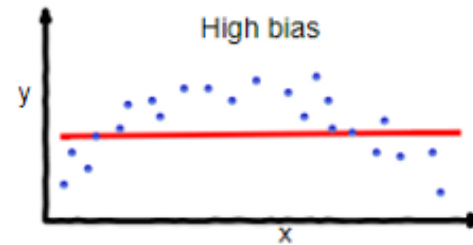
Maskinlæringsmetoder



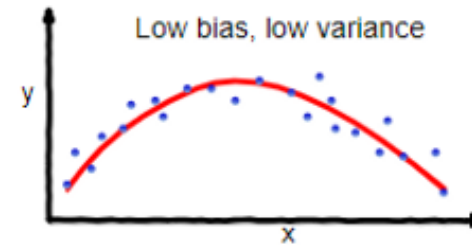
Varians og bias i modellen



overfitting



underfitting



Good balance

Bias

- Avstanden mellom gjennomsnittet til prediksjonene og gjennomsnittet til fasit/observasjonene
- Høy bias tyder på en feil i modellen
- Underfitting

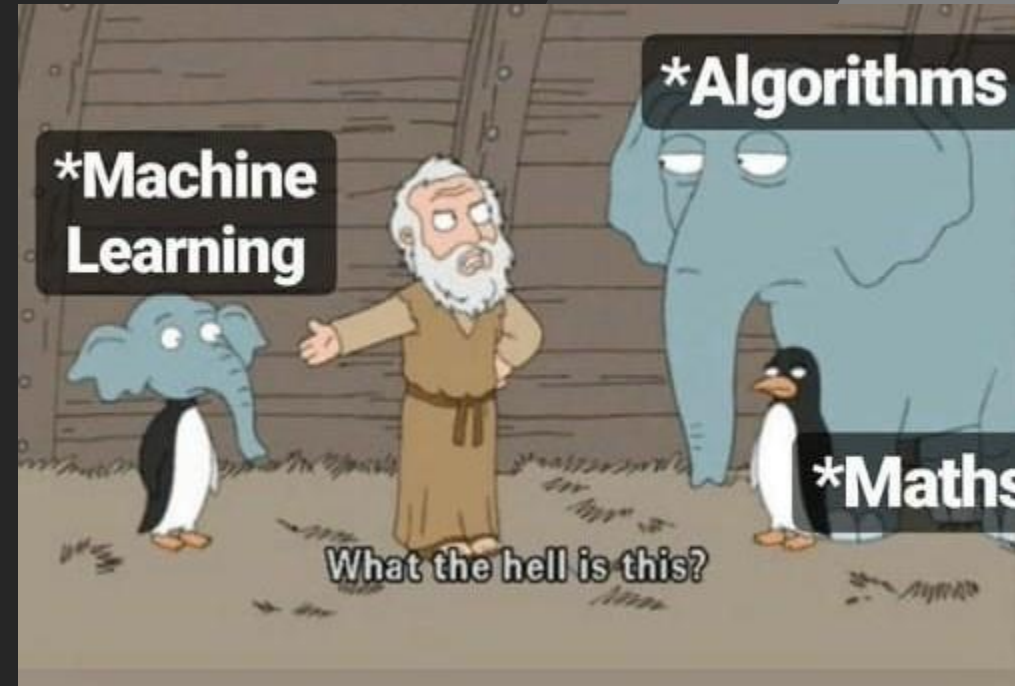
Varians

- Spredningen til prediksjonene
- Høy varians tyder på at modellen er sensitiv for støy i inputdataen
- Overfitting

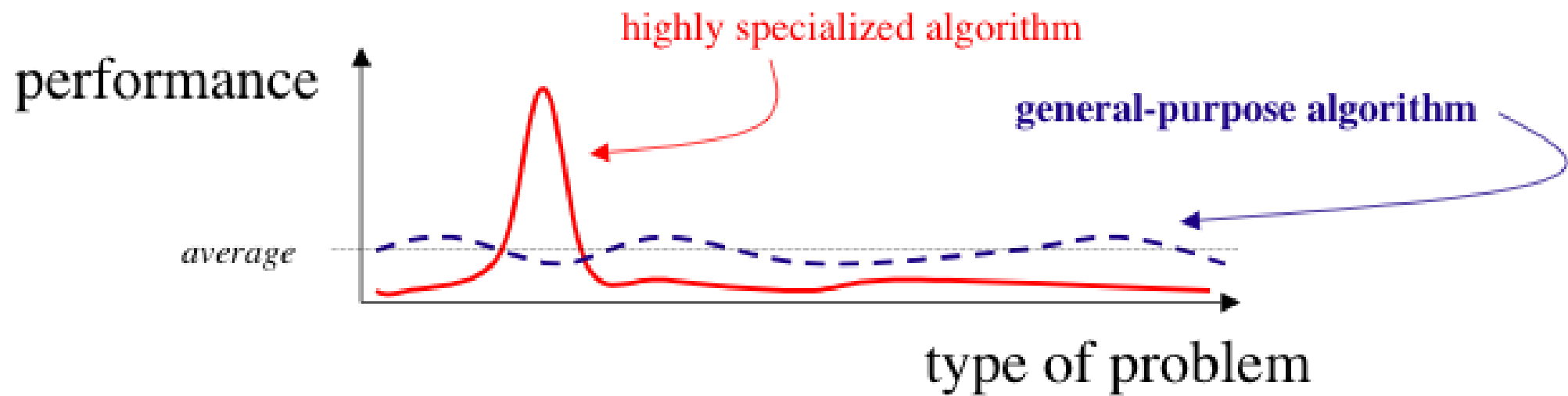
Eksempel notebook 2

- Forskjellige datatyper
- Supervised learning
- Unsupervised learning

Modeller og modellvalg

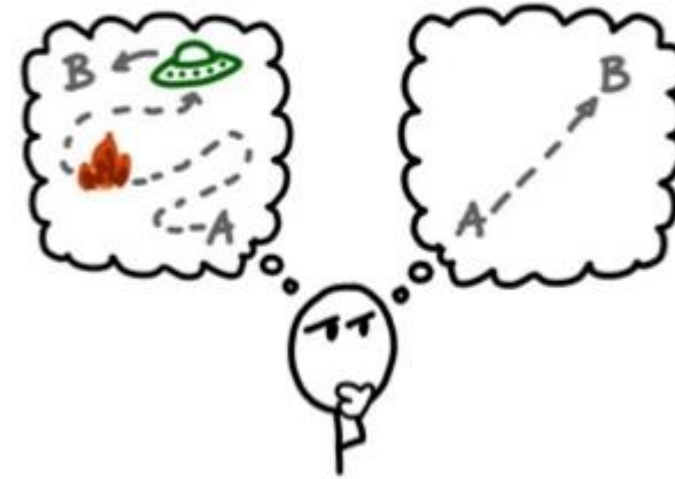


- «No Free Lunch»



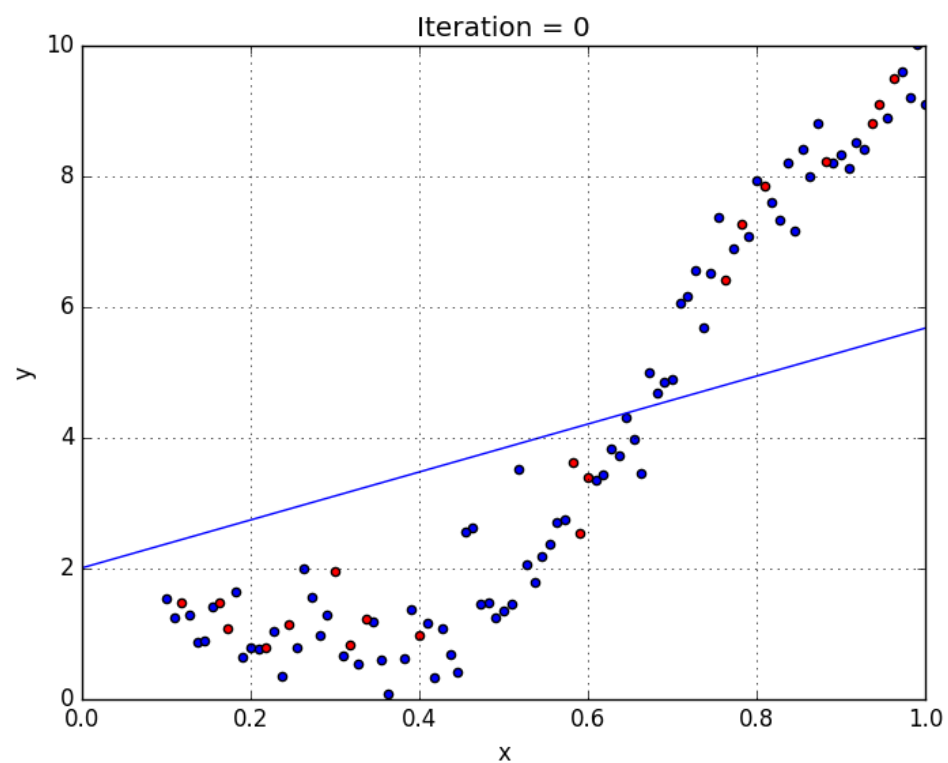
Occam's Razor

- Velg den enkleste hypotesen som kan forklare problemet
- I praksis: Et beslutningstre kan være å foretrekke fremfor en mer komplisert modell

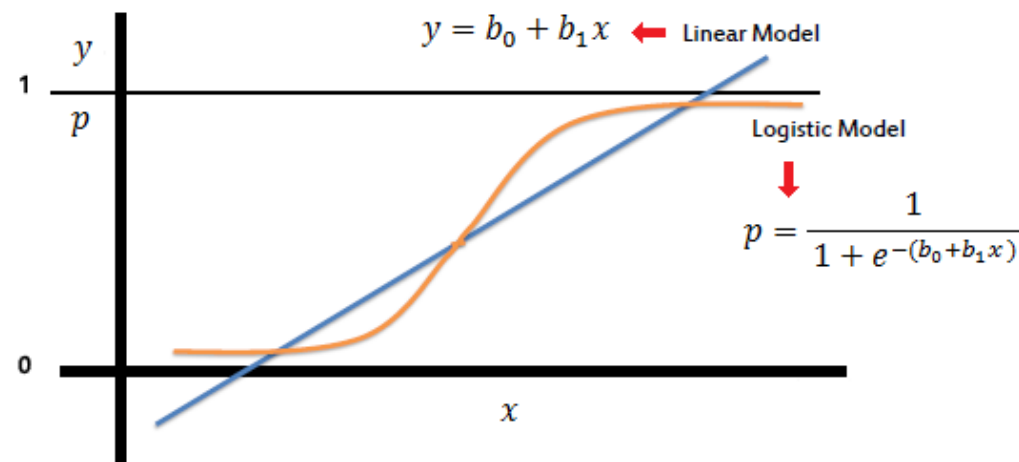


"When faced with two equally good hypotheses, always choose the simpler."

Linear Regression



Logistic regression



K-Means Eksempel

k-means be like:

Gruppering
(Clustering)

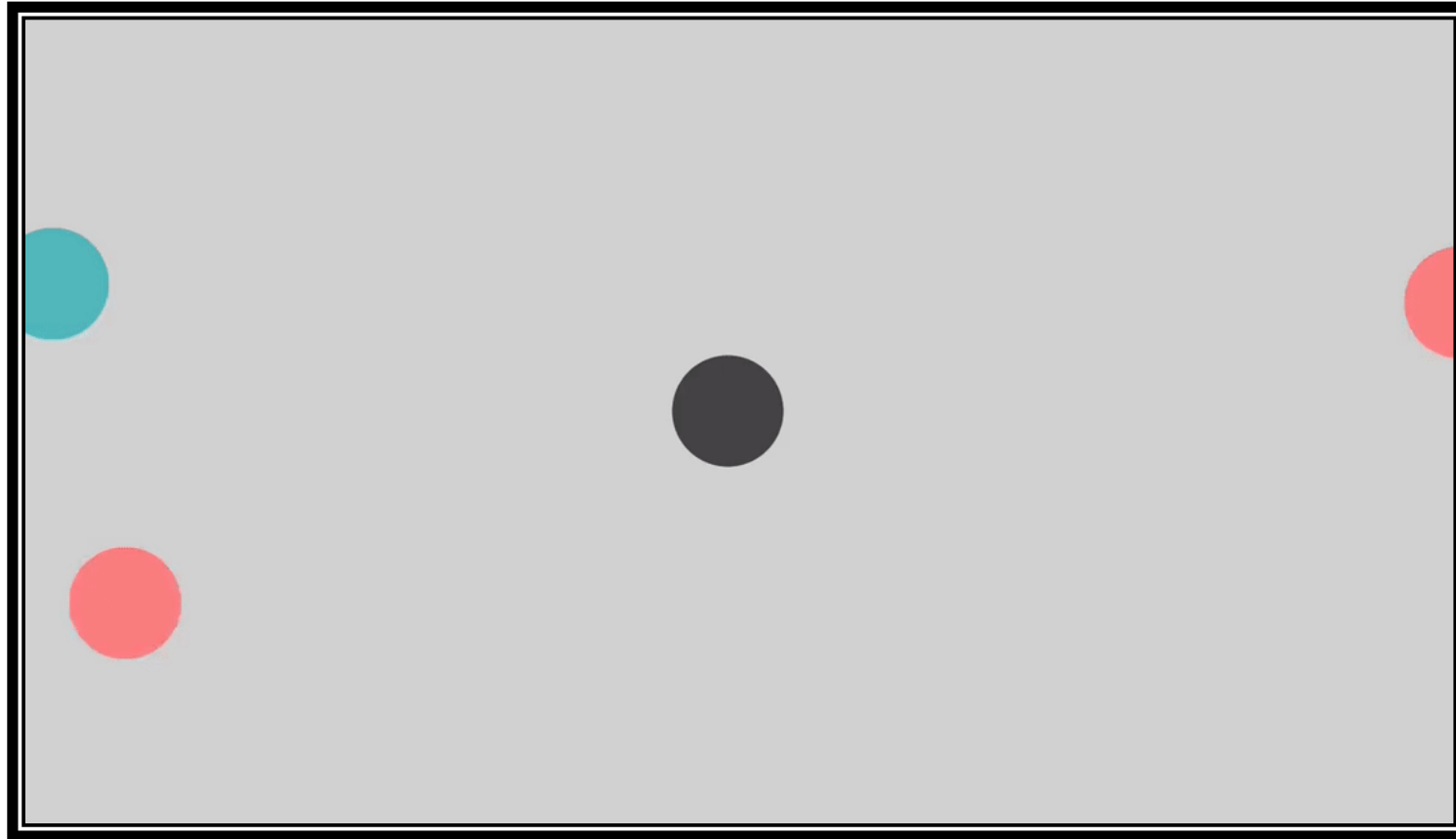


K-Nearest Neighbour (KNN)

Eksempel

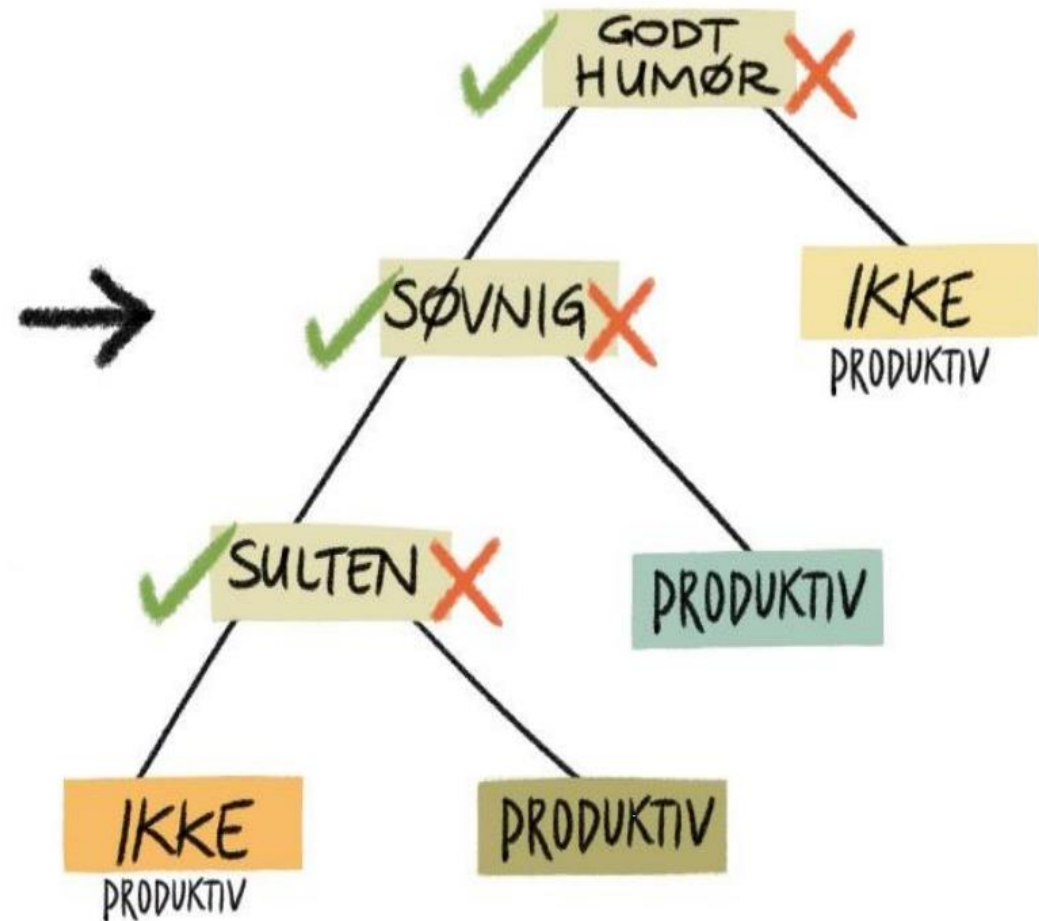
Klassifisering
(Classification)

Regresjon
(Regression)

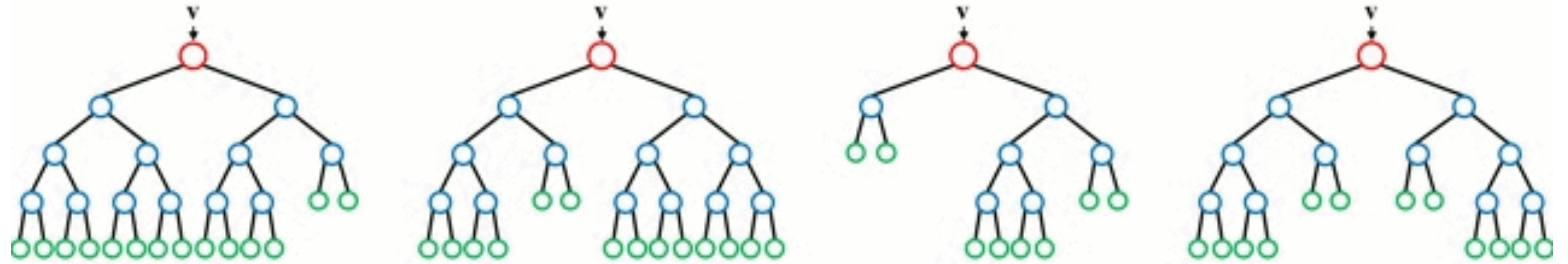


Decision Tree

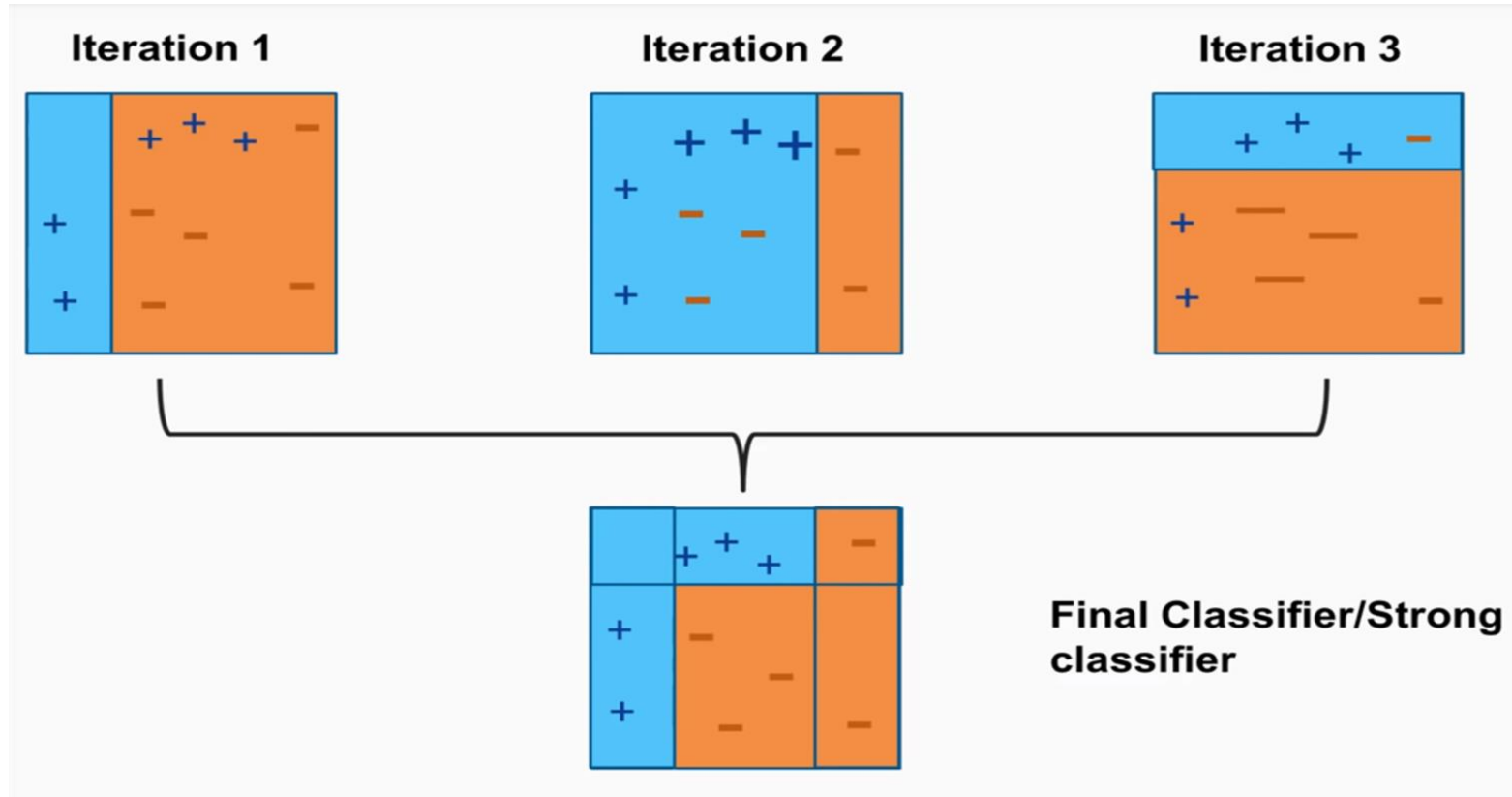
SØVNIG	SULTEN	GODT HUMØR	PRODUKTIV
×	×	×	×
×	×	✓	✓
×	✓	×	×
×	✓	✓	✓
✓	×	×	×
✓	×	✓	✓
✓	✓	×	×
✓	✓	✓	×

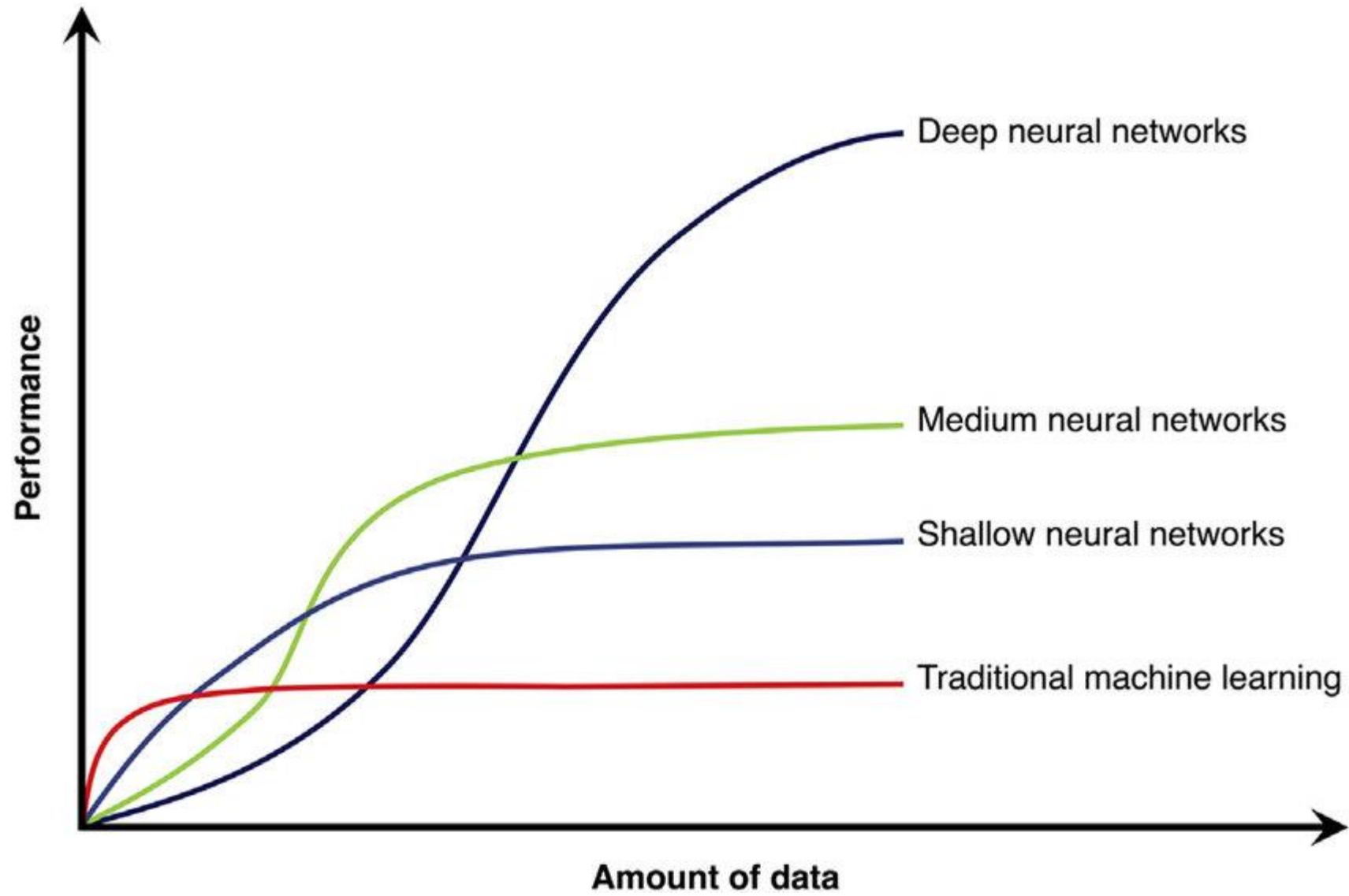


Random Forest (Bagging)



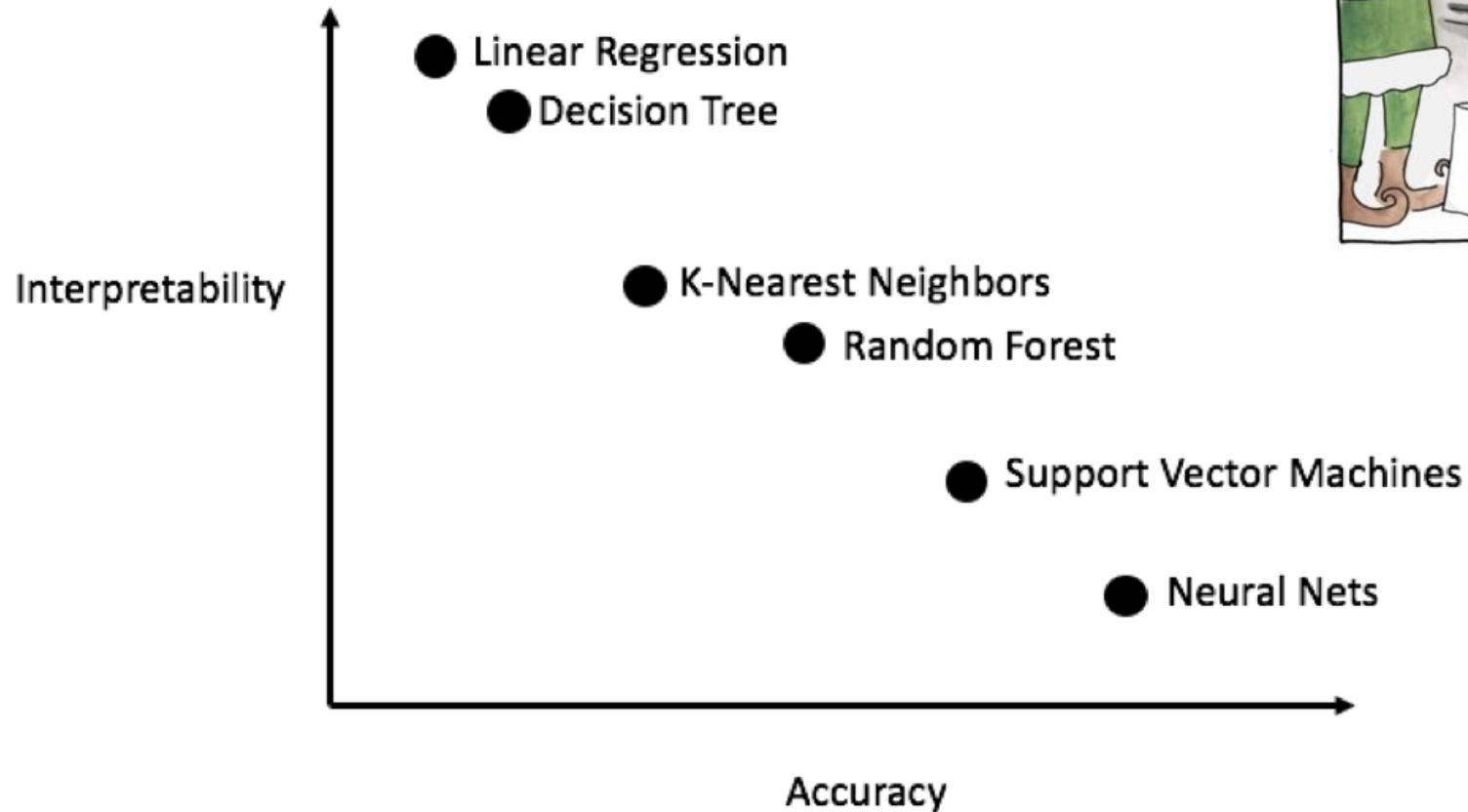
AdaBoost (Adaptive Boosting)





Tolkning

Kan vi gi svar på hvorfor modellen har tatt akkurat denne avgjørelsen?



© marketoonist.com

Har jeg et
maskinlæringsproblem?

Data – fundamentet for å kunne gjøre maskinlæring

Metode - Klassifisering, regresjon eller clustering

Reflekterer dataen hva vi prøver å predikere

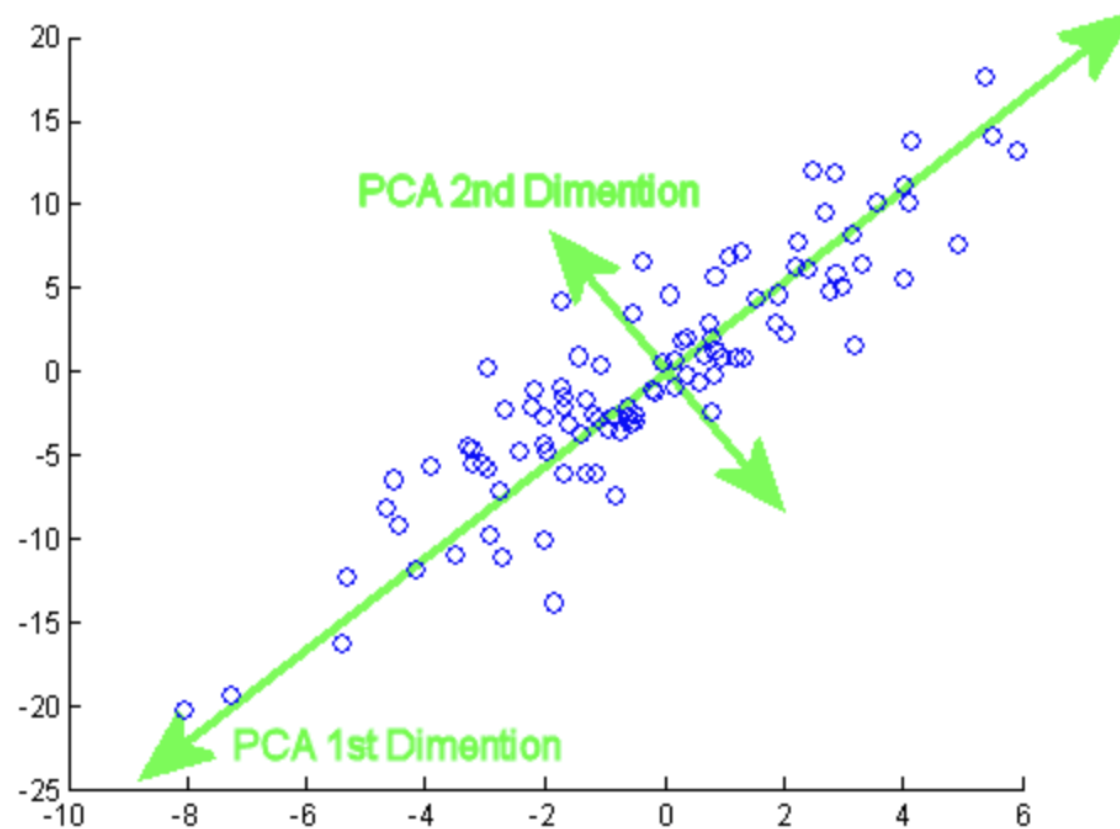
Krav og begrensninger til data og tolkning

Hvilke type modeller kunne egnet seg og hvordan reflekterer det på datamengden vi har?

Eksempel notebook 3

- Trene mer avanserte modeller

Principal Component Analysis (PCA)



Preprocessing

Fjerne & fylle verdier

Standardisere data

Fjerne data

Label-encoding

One-hot-encoding

Rescaling
Data

Standardizing
Data

Binarizing
Data

One Hot
Ending

Label
Encoding

Prestasjonsmål

Mean squared error

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n e_t^2$$

Root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$$

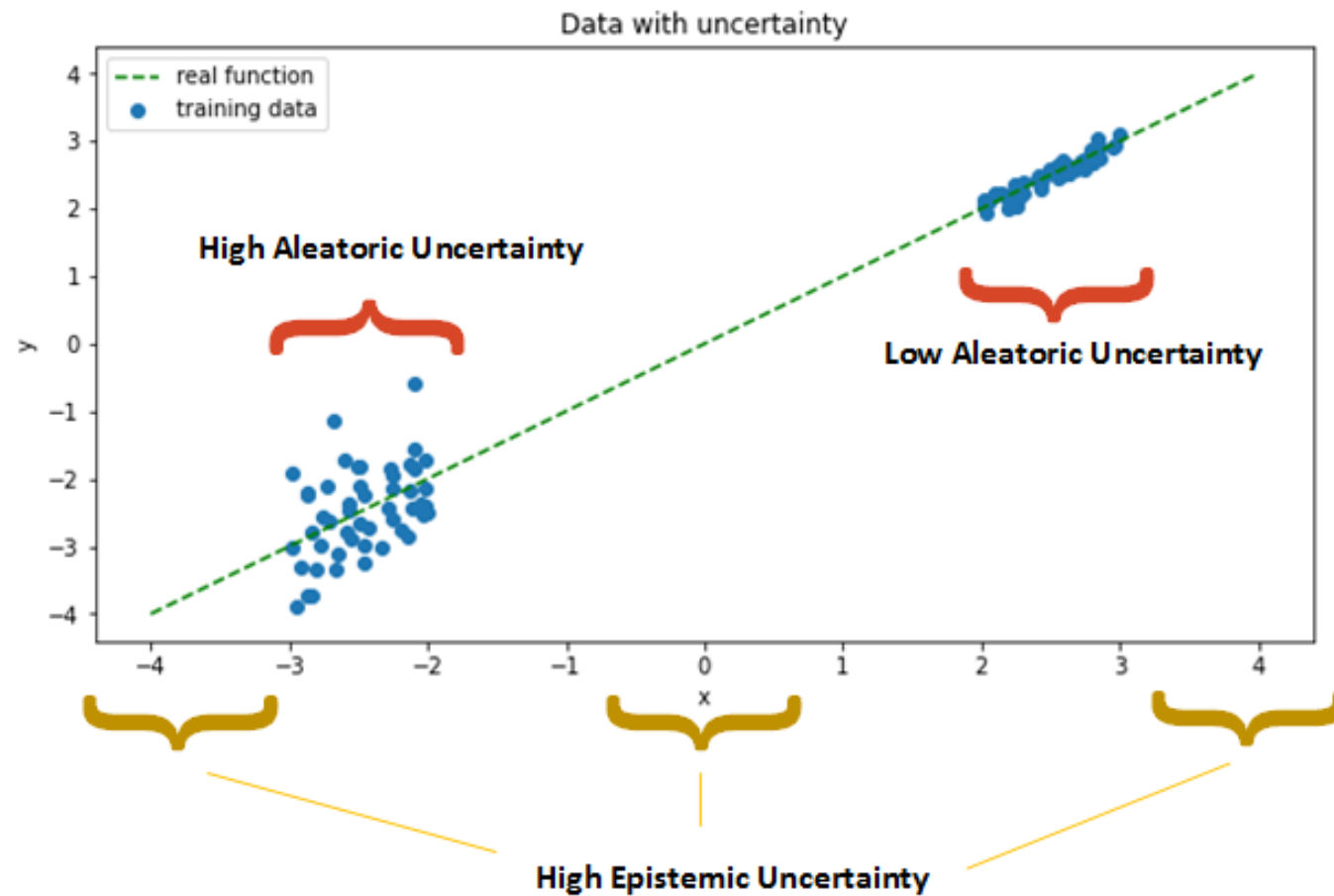
Mean absolute error

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |e_t|$$

Feature engineering

- Legge til domenekunnskap til eksisterende data
- Eksempelvis:
 - Benytte seg av GPS koordinater til å regne avstand mellom punkter
 - Endre på datoer til å gi variabler for hendelser som «is_weekend» etc.
 - Legge til innkjøpspris på produkter til å avregne profitt

Usikkerhet



Eksempel notebook 4

- Preprocessering
- Pipelines