Hvordan forklarer vi kunstig intelligens?

TADAgen NAV, 24.01.24

Martin Jullum, jullum@nr.no
Seniorforsker, Norsk Regnesentral

NORSK REGNESENTRAL
NORWEGIAN COMPUTING CENTER

Viktigheten av å forklare

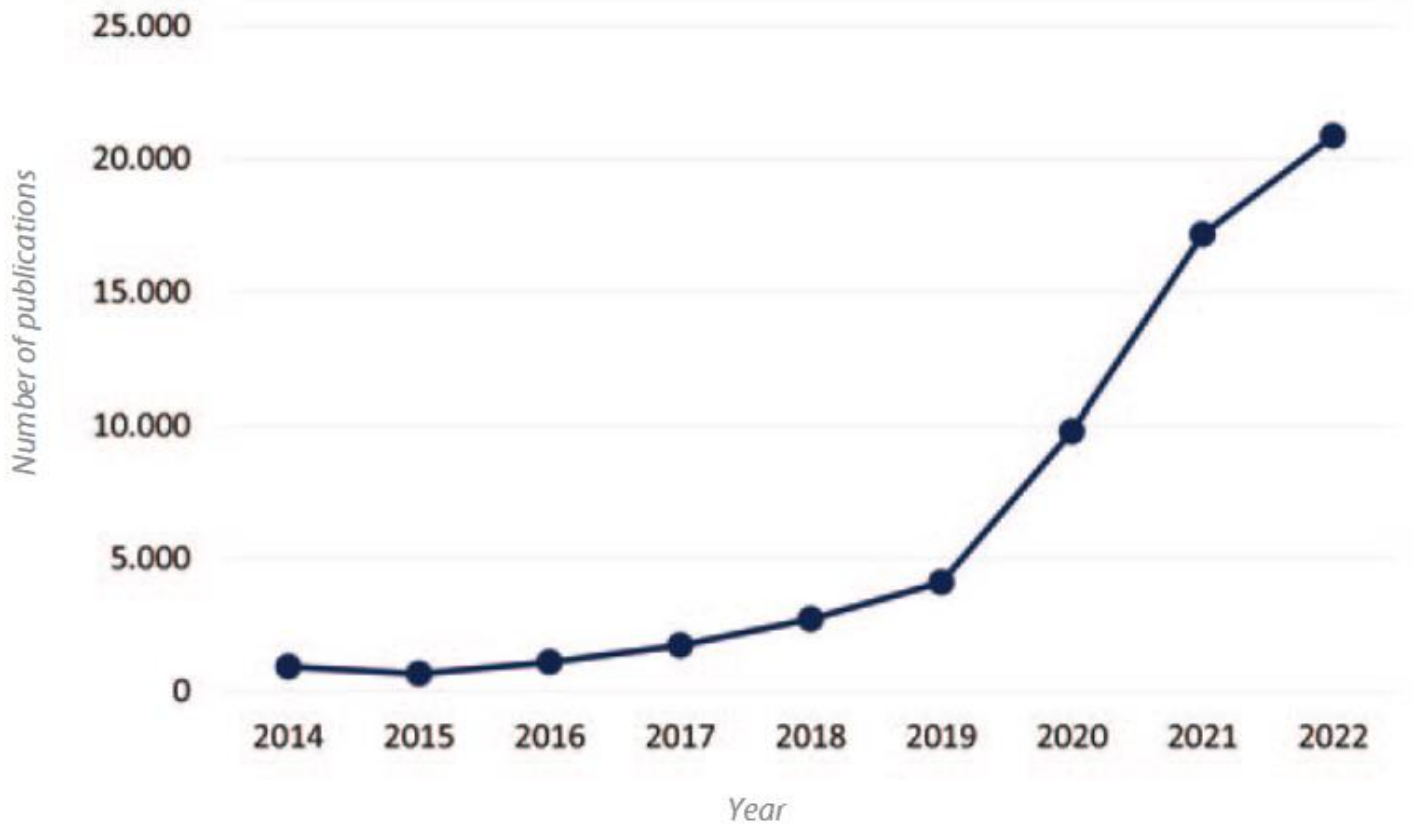
- Trygghet og tillit blant utviklere, ansvarlige og brukere
- Sette brukeren i stand til å utfordre en automatisk beslutning
- Sikre ansvarlig bruk av data/modell (personvern og diskriminering)
- Lovverk: AI ACT?, forvaltningsloven?, GDPR?
- Hjelpe utvikler å forbedre modellen/KI-systemet ved å oppdage feil/uønsket atferd

Forskningsfeltet Explainable AI (XAI)

- 1. Prøve å forstå hva kompliserte sort-boks-modeller gjør
- 2. Utvikle modeller som er forklarbare i seg selv

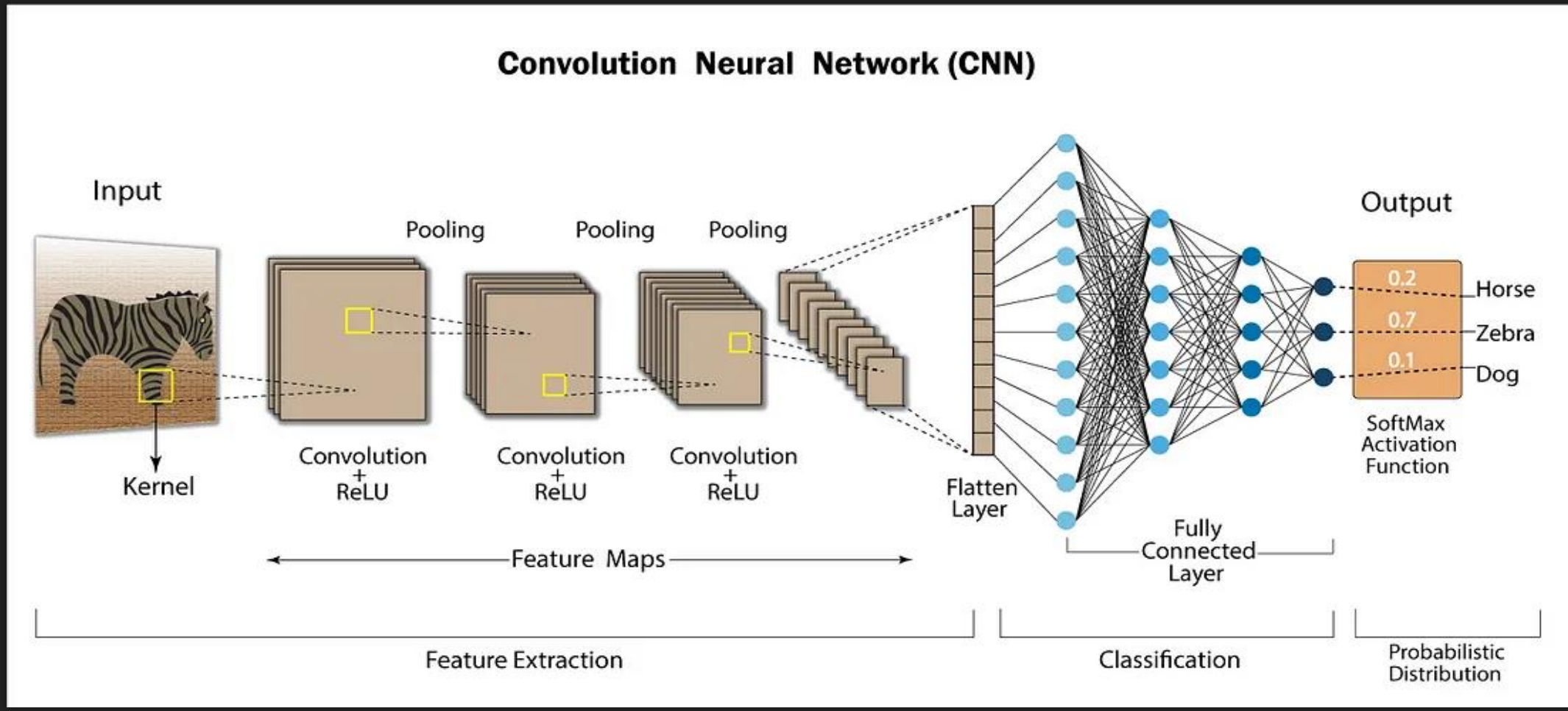
- Målet: At beslutninger basert på slike modeller blir mer transparente, forståelige og tolkbare for mennesker

Figure 2. Number of scientific publications per year on Explainable Artificial Intelligence (XAI).

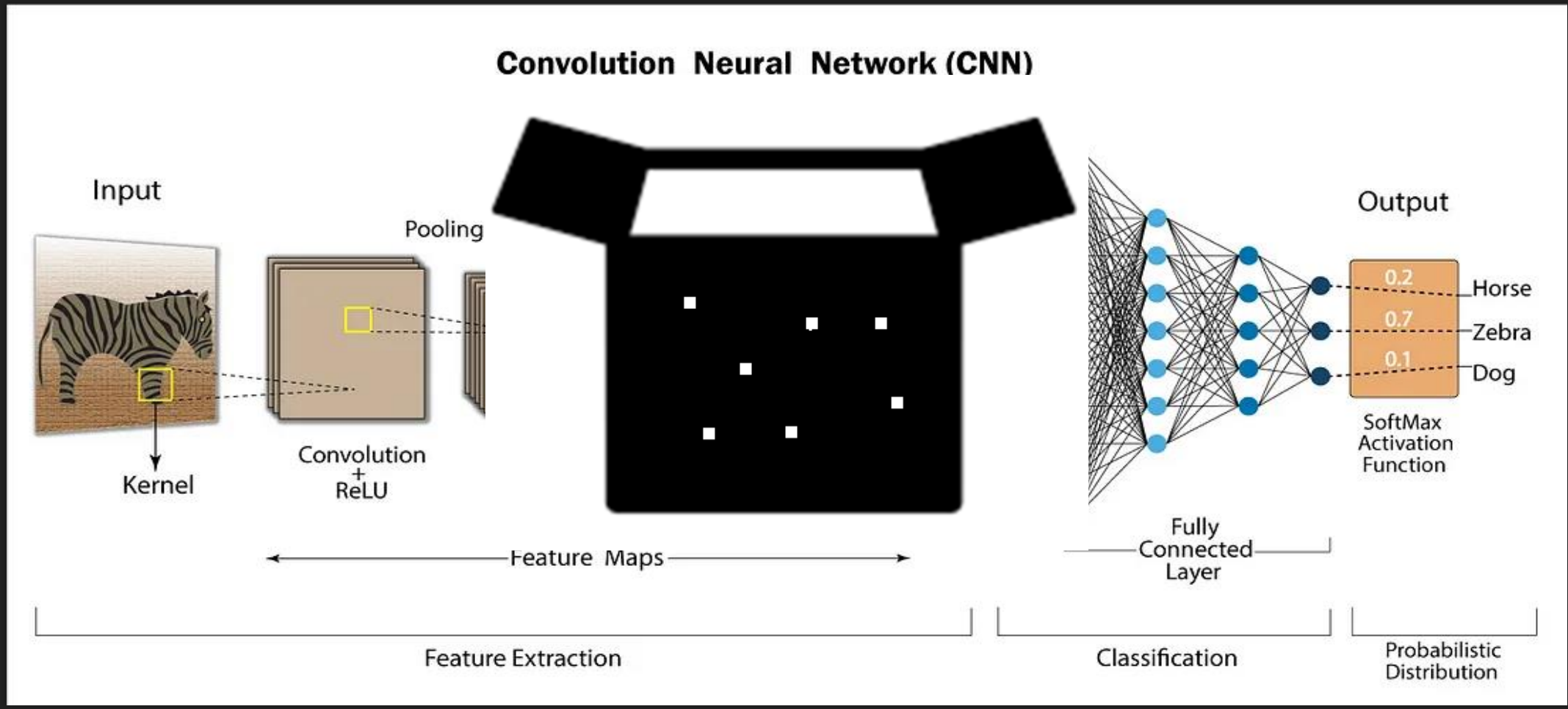


Figur fra Management solutions rapport:
"Explainable artificial intelligence (XAI) – Challenges of model interpretability" (2023)

Kompliserte modeller ER vanskelige å forstå



Kompliserte modeller ER vanskelige å forstå



Enkle modeller er ikke alltid enkle…

- Enkel lineær modell med normalfordelte variabler x_1 og x_2:

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2$$

- **Enkel forklaring**: y øker med b_1 når x_1 øker med 1, og tilsvarende for x_2
- **Forutsetning**: x_1 og x_2 er uavhengige -> Sjelden tilfellet i praksis

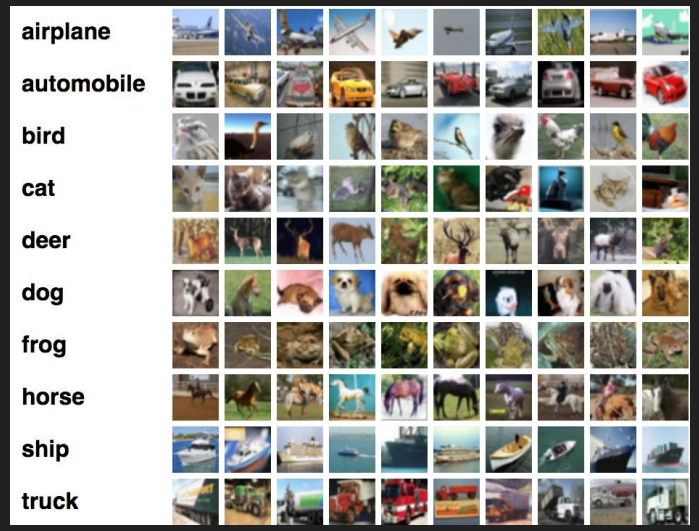
- F.eks ved $\text{corr}(x_1, x_2) \approx 1$, $E[x_1] \approx E[x_2]$,:
  y øker med b_1+b_2 når x_1 øker med 1 (fordi da øker typisk også x_2 med 1).

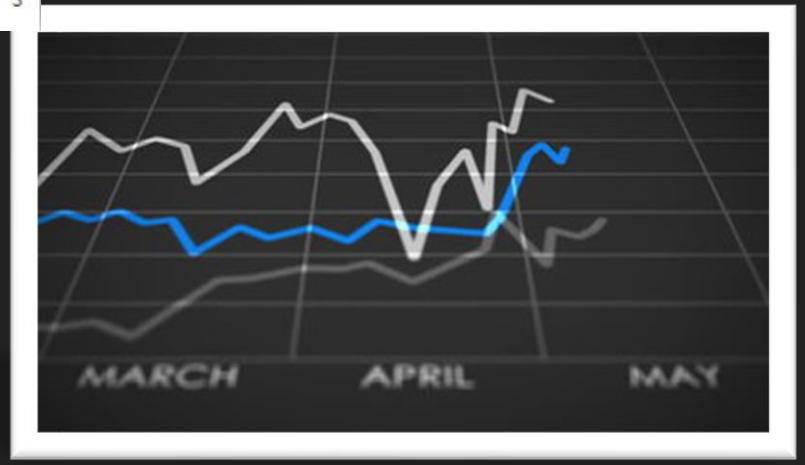
- Mer komplisert når avhengigheten er middels sterk/ikke-lineær/lokalt varierende, det er flere variabler, eller modellen er ikke-lineær

Ulike datatyper krever ulik forklaringsmetodikk

Survived	Pclass						Embarked
1	3	female	30	0	1	3.26	S
	2	male	12	0	1	21.77	
0	1	male	9	0	2	8.86	S
0	3	male	13	0	0	16.07	S
0	2	male	40	2	0	-0.09	S
...	...	...	...	...	...	...	...
0	3	female	31	0	2	40.78	
0	2	female	30	1	0	12.36	
1	3	female	32	1	0	-0.88	S
0	3	male	42	0	0	5.7	S
		male	13	0		.49	S

PREDIKSJONS-MODELLER FOR TABELLDATA

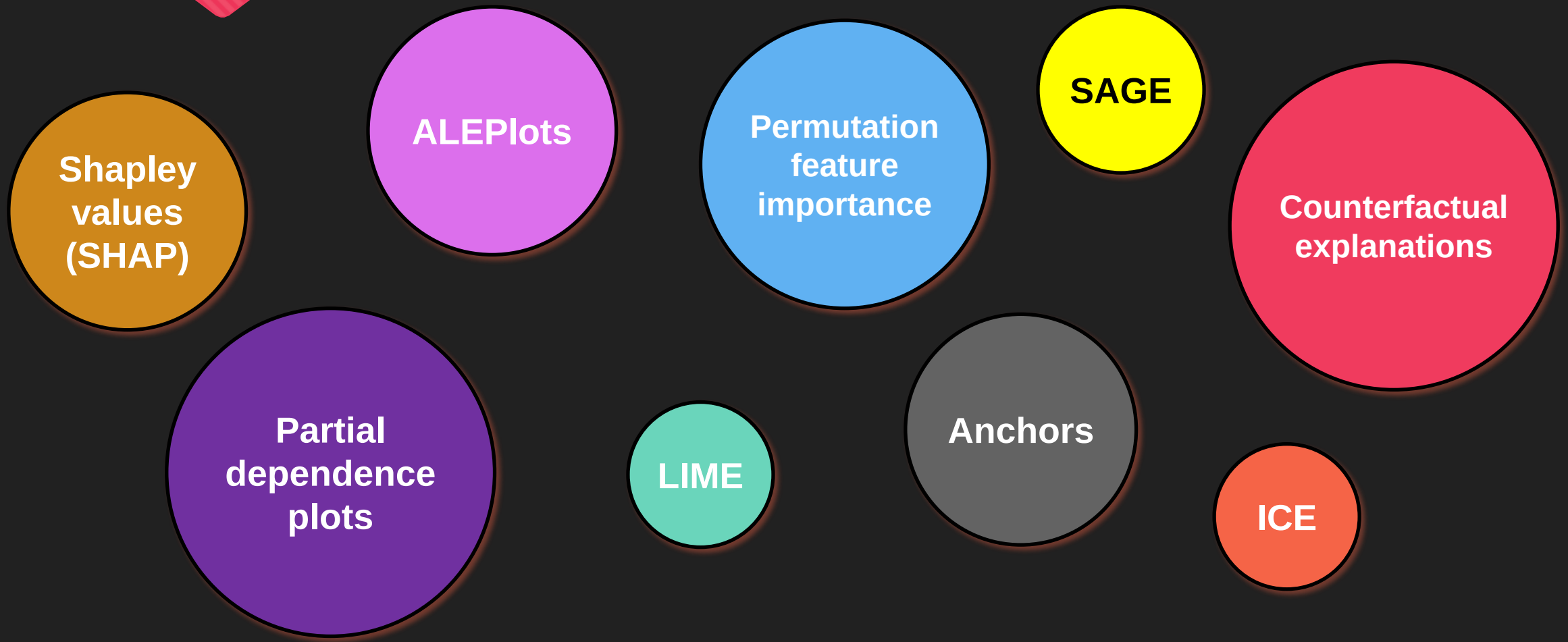




Large language models

Graph

MANGE FORKLARINGSMETODER



PRESENTASJONSFORMER

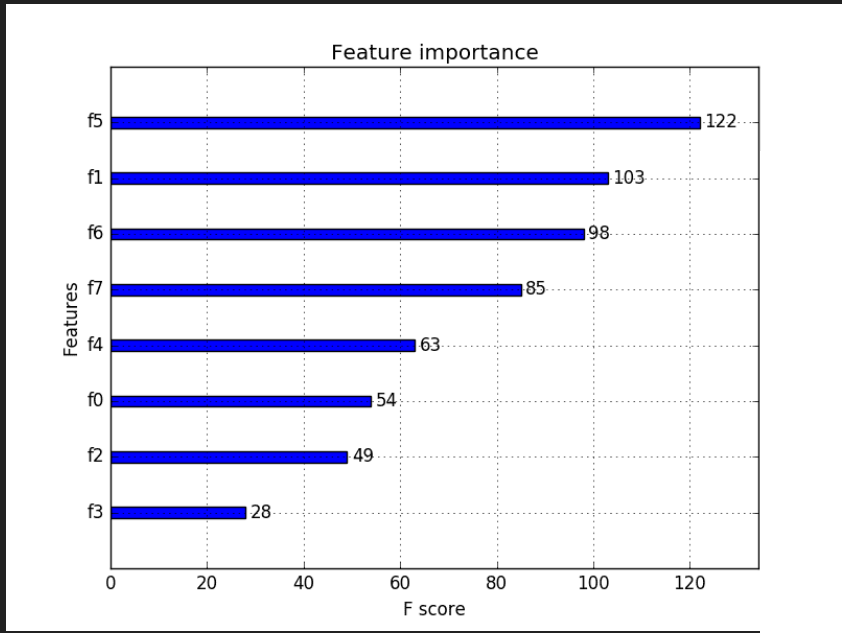
Shapley values (SHAP)

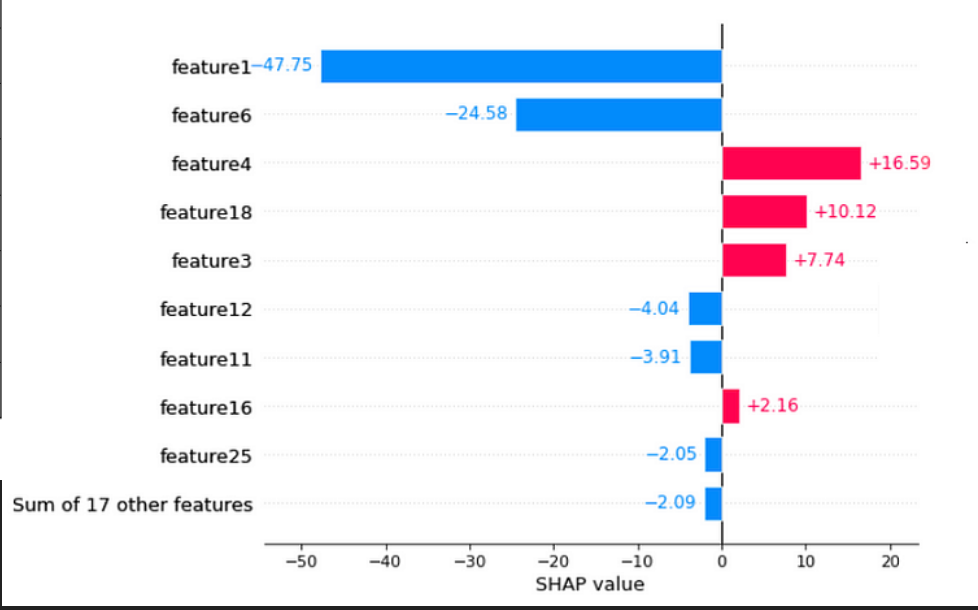
SAGE

LIME

Permutation feature importance

VARIABELVIKTIGHET/ VARIABELBIDRAG



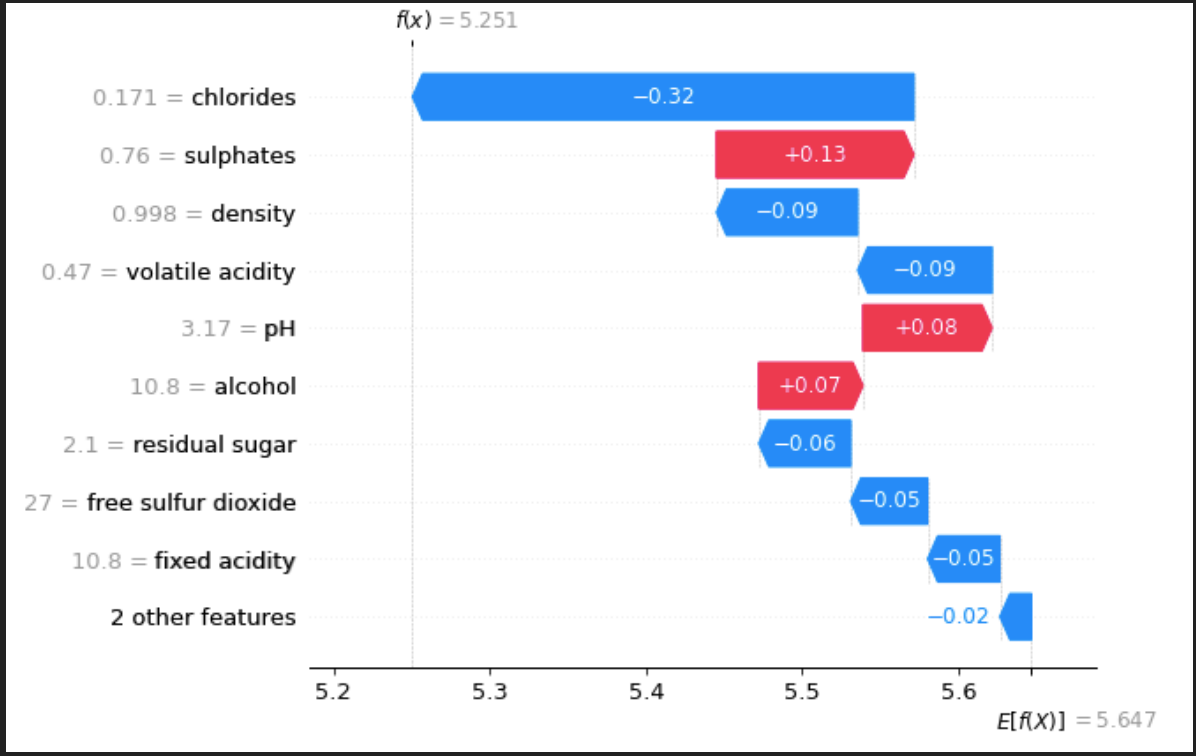


PRESENTASJONSFORMER

Shapley values (SHAP)

SAGE

DEKOMPONERING



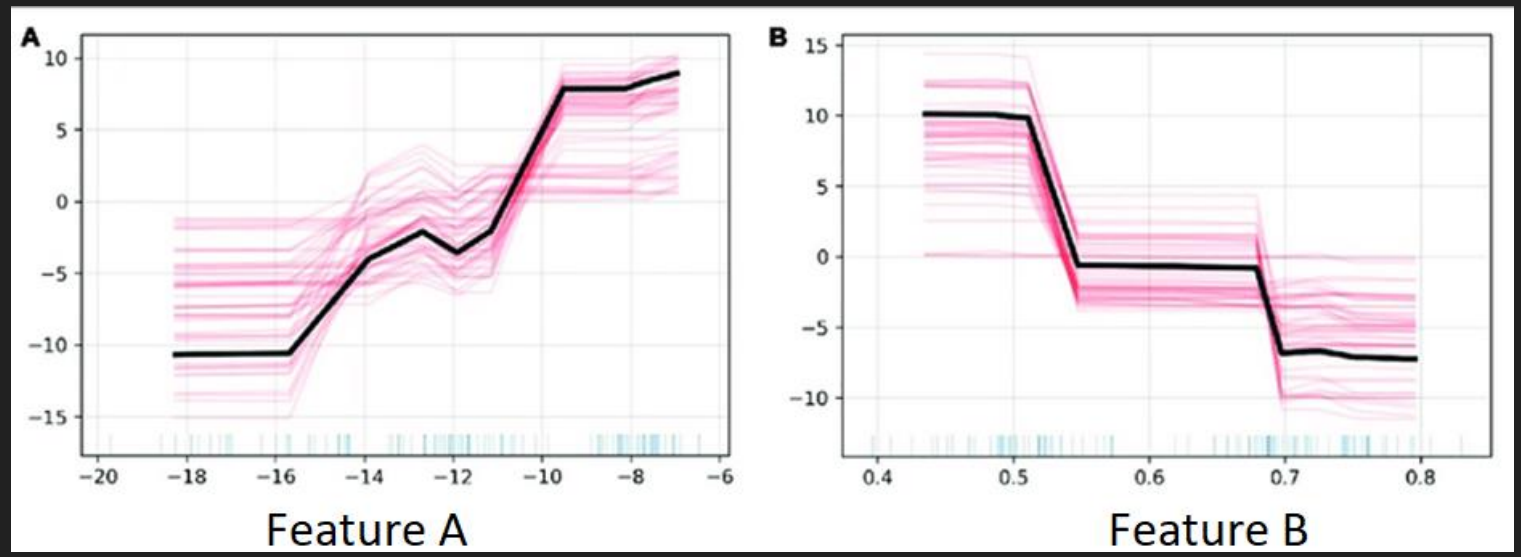
PRESENTASJONSFORMER

ALEPlots

ICE

Partial dependence plots

VARIABELEFFEKT-PLOTT



PRESENTASJONSFORMER

Counterfactual explanations

EKSEMPLER

Observations to explain

ID	Features				*f(x)*	*Decision*
	Age	Sex	Salary	Def. last year		
1	30	F	$ 6000	yes	0.18	0
2	25	M	$ 4500	no	0.30	0

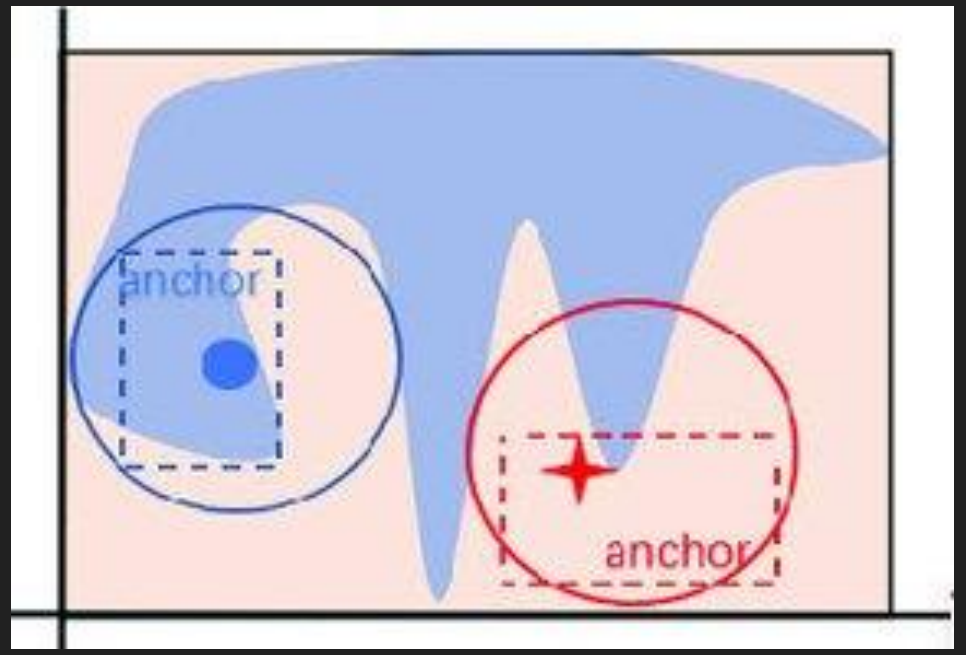
Final counterfactual explanations

Explain ID	Age	Sex	Decision	Salary	Def. last year
1	30	F	1	$ 6000	no
2	25	M	1	$ 4800	no

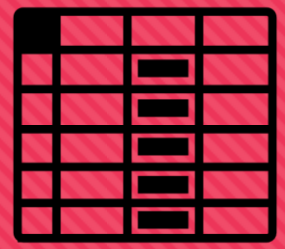
PRESENTASJONSFORMER

Anchors

REGLER



TYPER FORKLARINGER



$f(x)$

Forklaringsskala	
Global	Lokal

Effekten av å	
Observere variabler	Endre variabel-verdier

Hensyntas datafordeling	
Ja	Nei

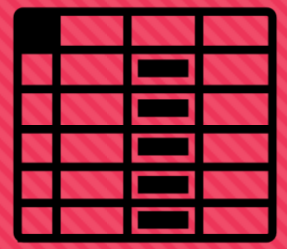
Anvendelighet	
Modell-spesifikk	
Modell-agnostisk	

Prediksjon direkte	Avledet beslutning
Forklaring av	

Input-variabler	Trenings-obs.
Forklaring av	

En og en	Fellesbidrag /interaksjon
Forklare variabelbidrag	

TYPER FORKLARINGER



$f(x)$

Forklaringsskala
- Global
- Lokal

Effekten av å
- Observere variabler
- Endre variabel-verdier

Hensyntas datafordeling
- Ja
- Nei

Anvendelighet
- Modell-spesifikk
- Modell-agnostisk

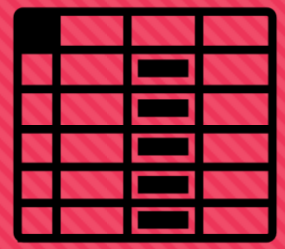
(Conditional) Shapley values

Forklaring av
- Prediksjon direkte
- Avledet beslutning

Forklaring av
- Input-variabler
- Trenings-obs.

Forklare variabelbidrag
- En og en
- Fellesbidrag /interaksjon

TYPER FORKLARINGER



$f(x)$

Forklaringsskala
- Global
- Lokal

Effekten av å
- Observere variabler
- Endre variabel-verdier

Hensynta datafordeling
- Ja
- Nei

Anvendelighet
- Modell-spesifikk
- Modell-agnostisk

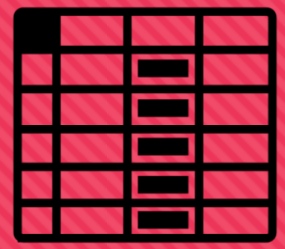
Permutation feature importance

Forklaring av
- Prediksjon direkte
- Avledet beslutning

Forklaring av
- Input-variabler
- Trenings-obs.

Forklare variabelbidrag
- En og en
- Fellesbidrag /interaksjon

TYPER FORKLARINGER



$f(x)$

Forklaringsskala
- Global
- Lokal

Effekten av å
- Observere variabler
- Endre variabel-verdier

Hensynta datafordeling
- Ja
- Nei

Anvendelighet
- Modell-spesifikk
- Modell-agnostisk

Counterfactual explanations

Forklaring av
- Prediksjon direkte
- Avledet beslutning

Forklaring av
- Input-variabler
- Trenings-obs.

Forklare variabelbidrag
- En og en
- Fellesbidrag /interaksjon

UTFORDRINGER OG FALLGRUVER

- Vet ikke hva slags forklaring man vil ha
- Forstår ikke hva slags forklaring forklaringsmetoden(e) gir
- Valgt forklaringsmetode er ikke den brukeren ønsker seg
- Forklaringene er upresise/feil:
  - Bruker grove approksimasjoner fordi det tar for lang tid å generere presise forklaringer
  - Ignorerer avhengighet mellom variablene
  - Ekstrapolering utenfor dataområdet
  - Metodeantagelser ikke oppfylt

HVORDAN NAVIGERE I XAI-JUNGELEN?

eXplego

An XAI-method selection tool by

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

nav

BigInsight

- Interaktivt beslutningstre som hjelper deg å velge forklaringsmetode

www.explego.nr.no