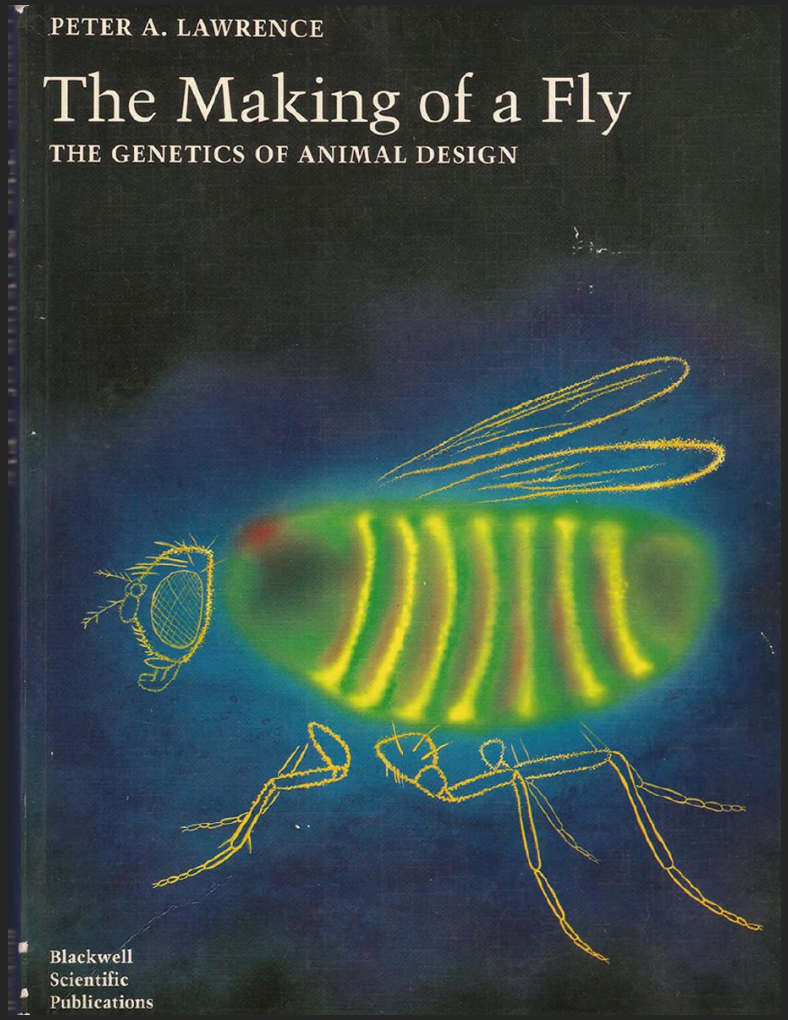
Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

BigInsight

Forklarbar kunstig intelligens:

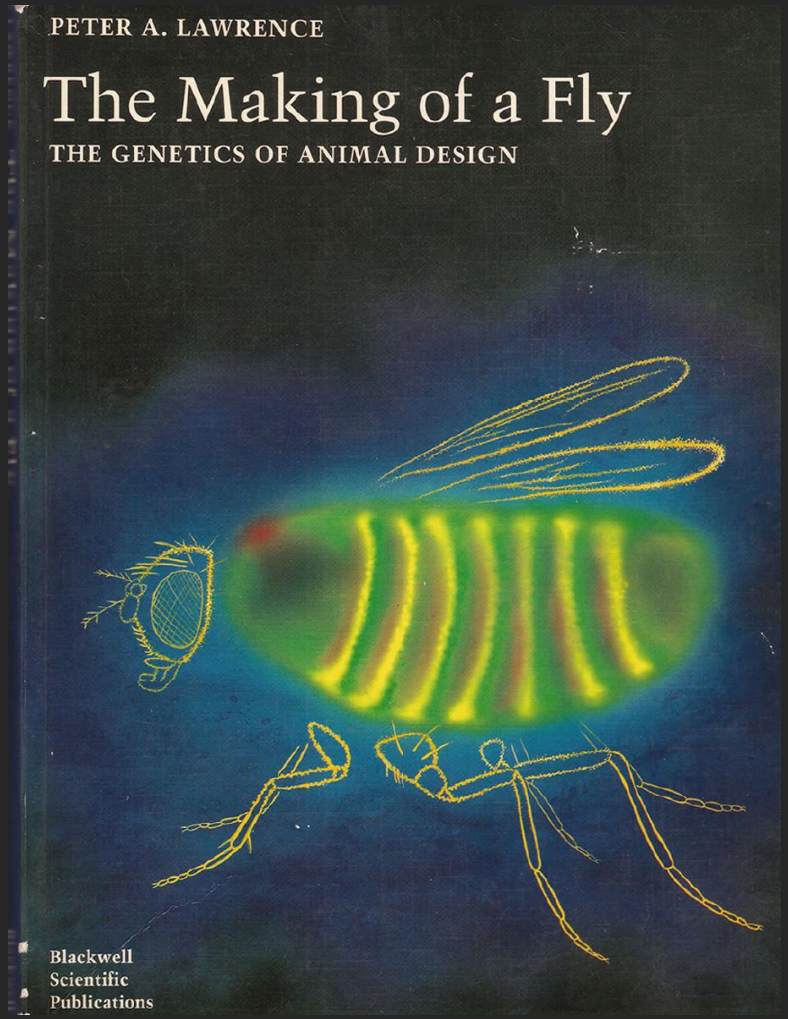
Metodegjennomgang – utvalgte forklaringsmetoder

Anders Løland



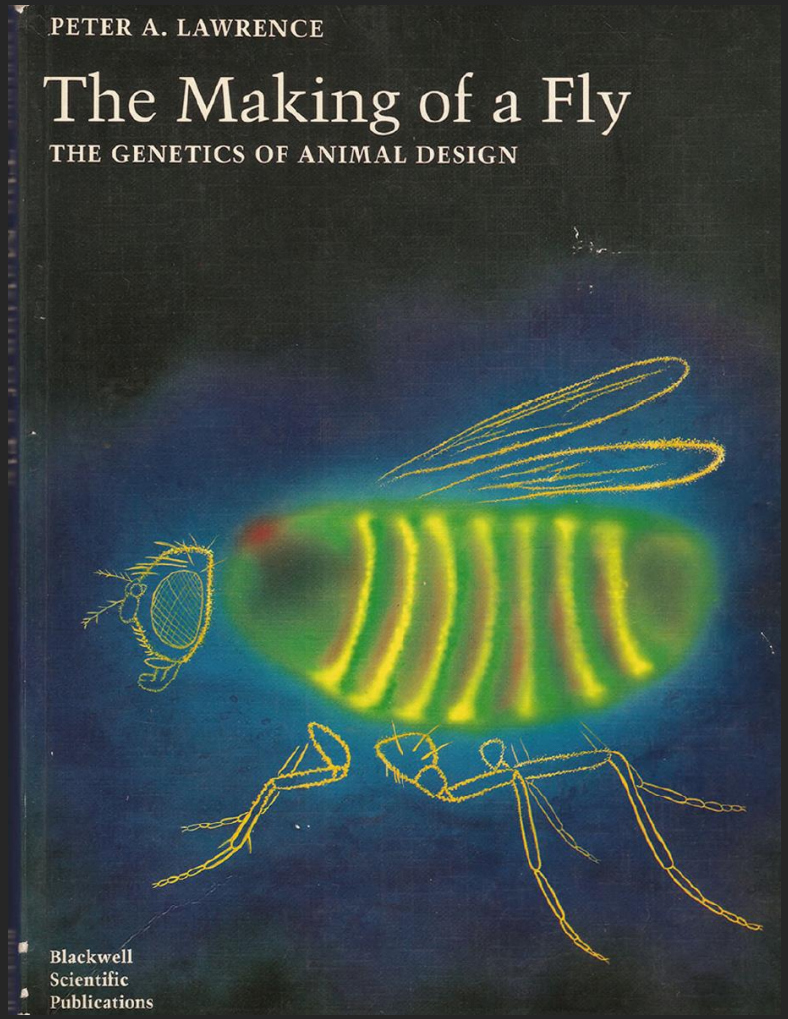
amazon

$106,23



Selger 1 =
0,9983 x Selger 2

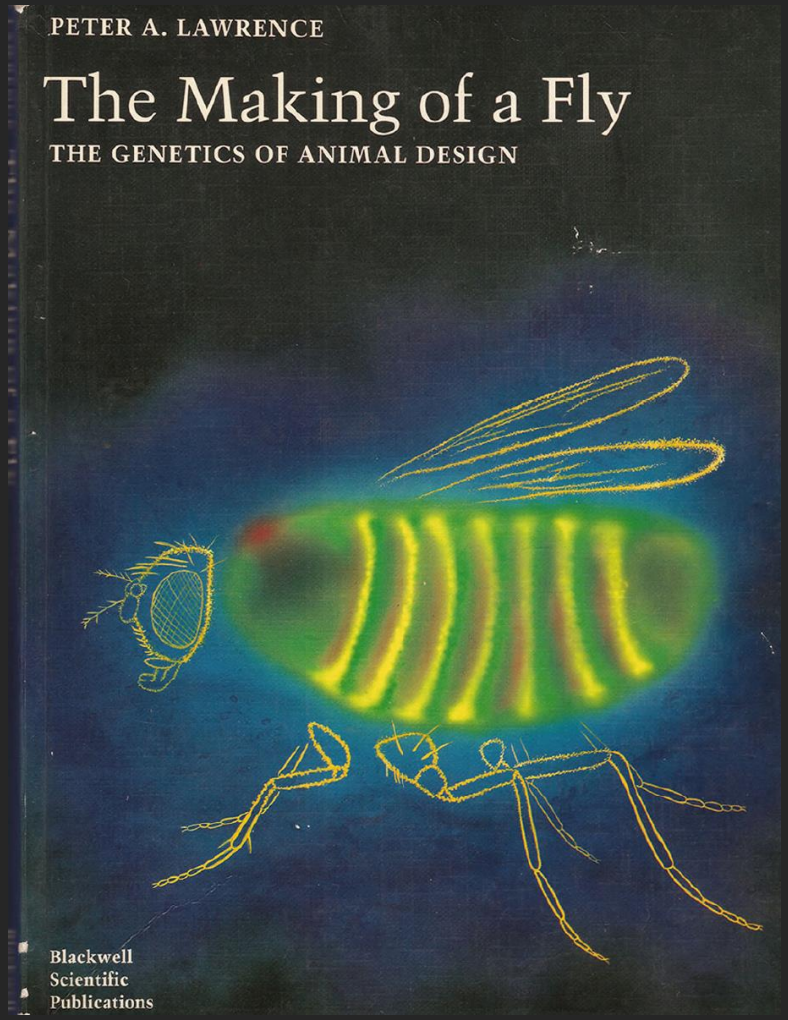
Selger 2 =
1,270589 x Selger 1



amazon

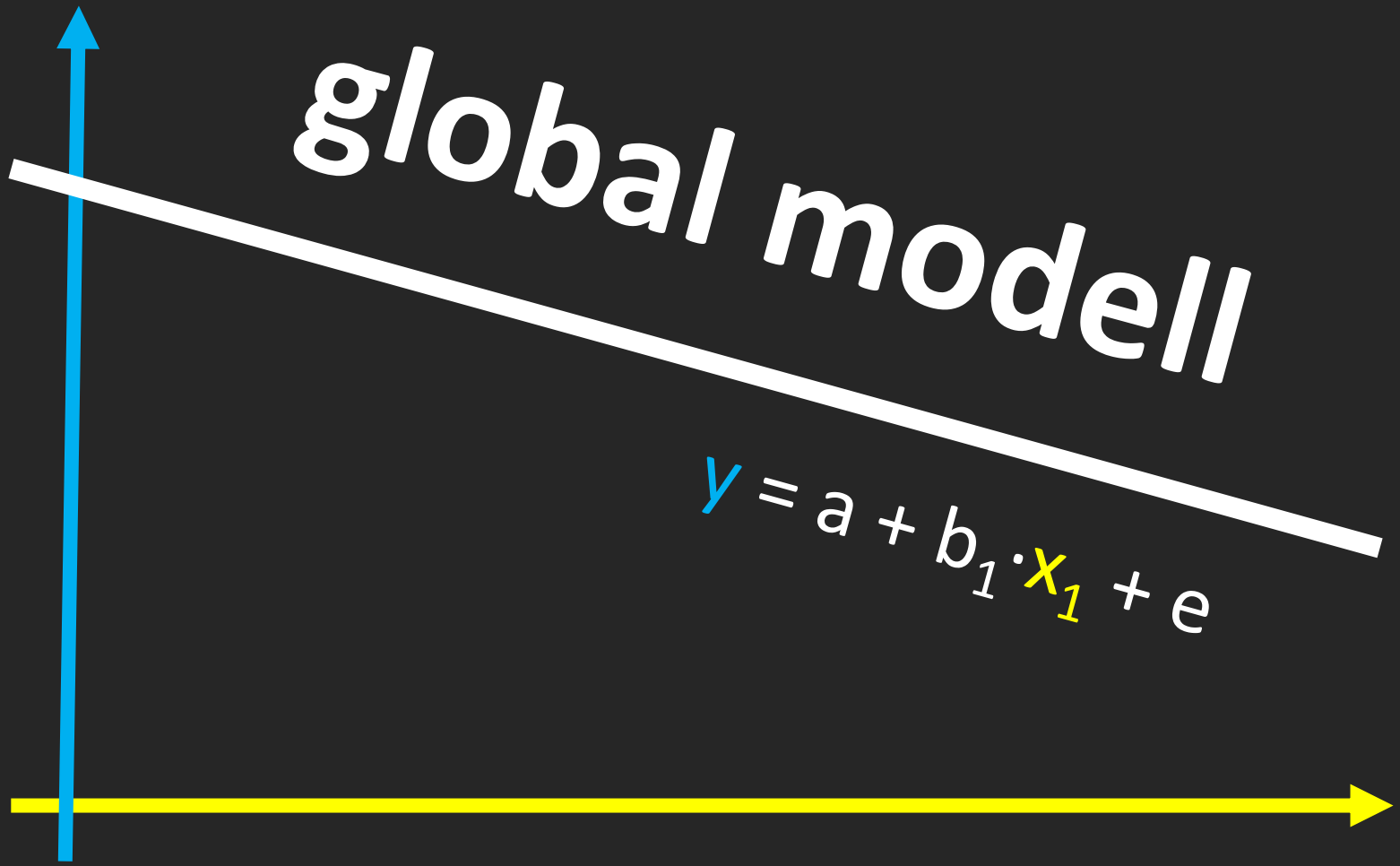
$23 698 655,93

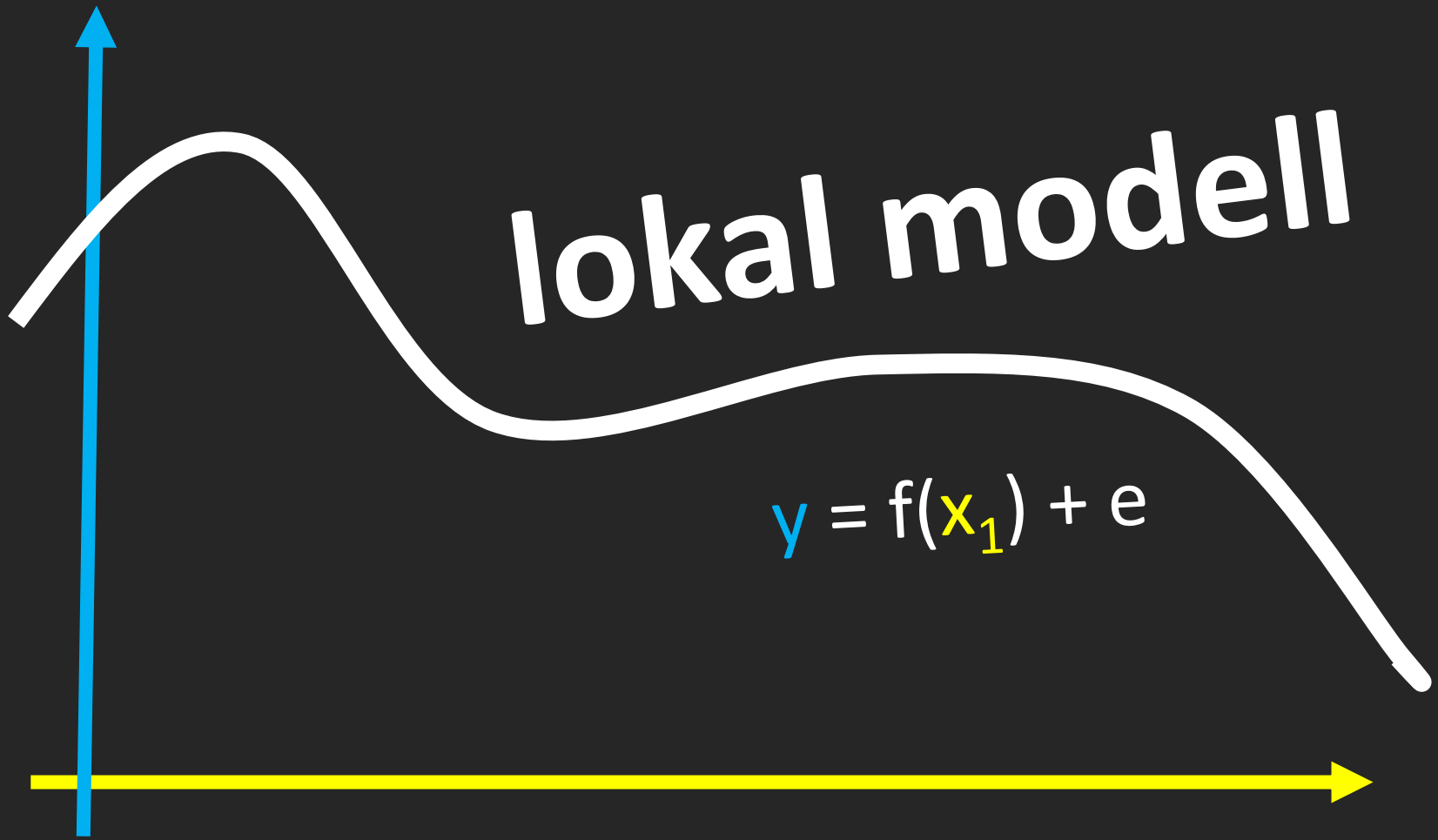
https://www.wired.com/2011/04/amazon-flies-24-million/



enkle algoritmer

enkle å forklare
hvis vi får titte
(litt) inn i dem


global modell
$y = a + b_1 \cdot x_1 + e$


lokal modell
y = f(x1) + e

modell*spesifikk* forklaring

enklere, men ikke like nyttig(?)

modell*agnostisk* – modelluavhengig – forklaring

vanskeligere (og nyttigere!)

alder

$$y = a + b_1 \cdot x_1 + e$$

kredittpoeng

forklaring: b_1 (og a)

alder

$$y = a + b_1 \cdot x_1 + e$$

kredittpoeng

model*spesifikk* forklaring
= model*agnostisk forklaring*

alder

$$y = a + b_1 \cdot x_1 + e$$

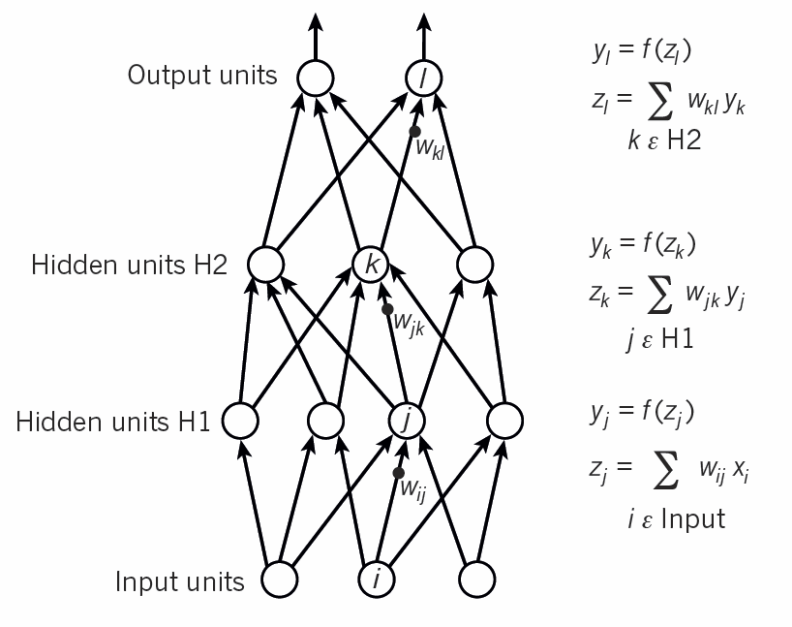
kredittpoeng

lokal forklaring = *global* forklaring

enkel, forklarbar modell?

$$y = a + b_1 \cdot x_1 + e$$

eller



LeCun et al: «Deep learning», Nature (2015)

bedre prediksjoner, utilgjengelig modell?

alder

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

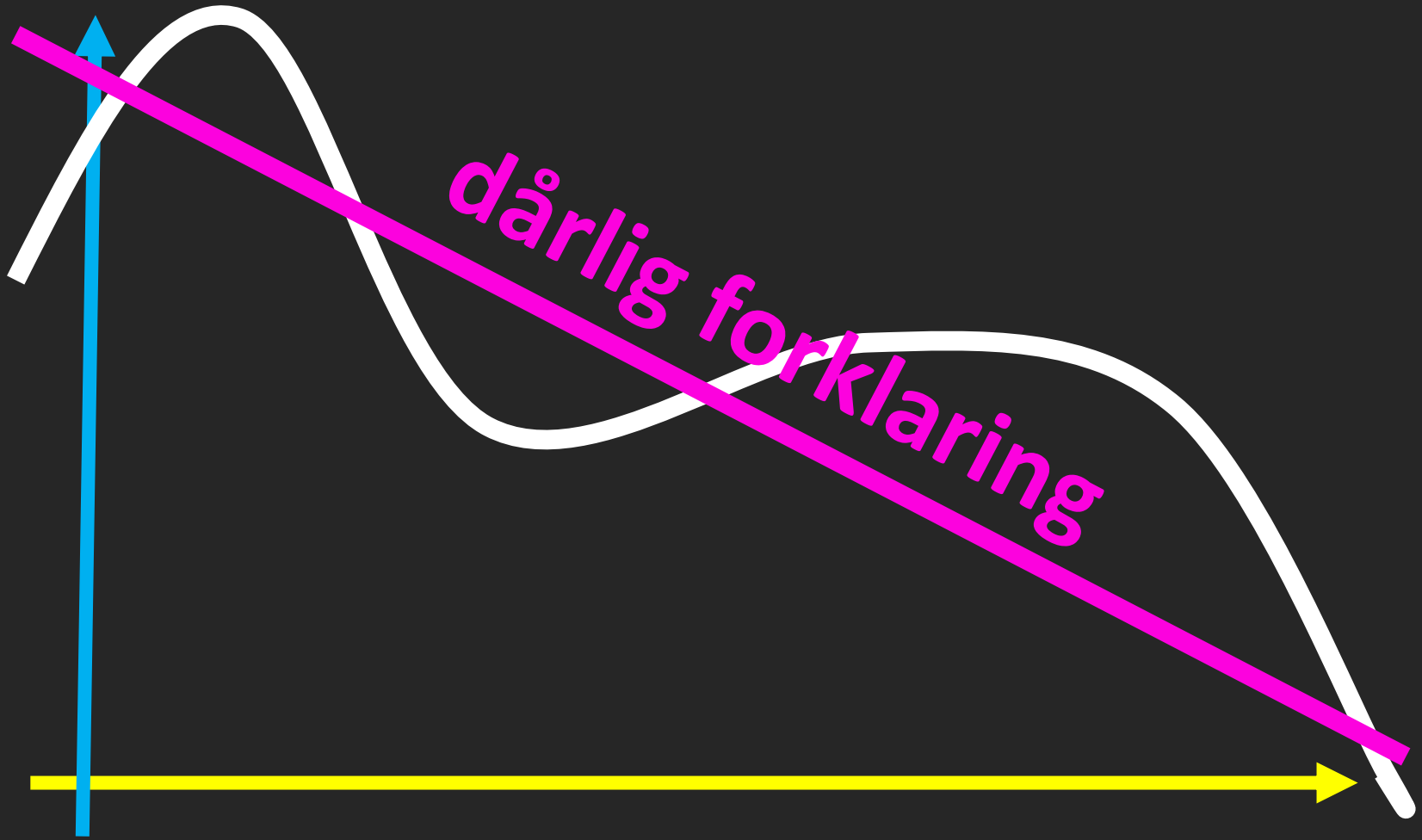
kredittpoeng

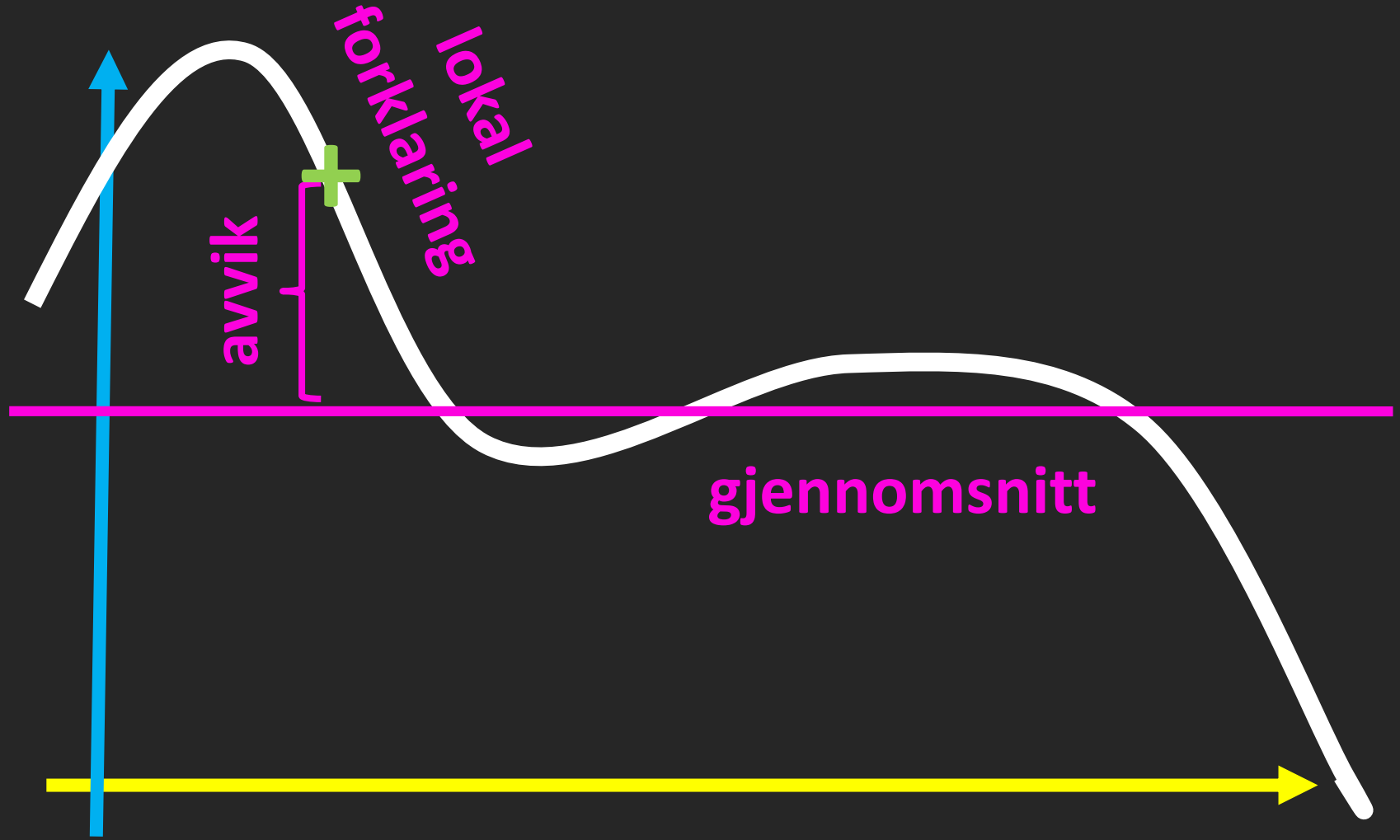
studiepoeng

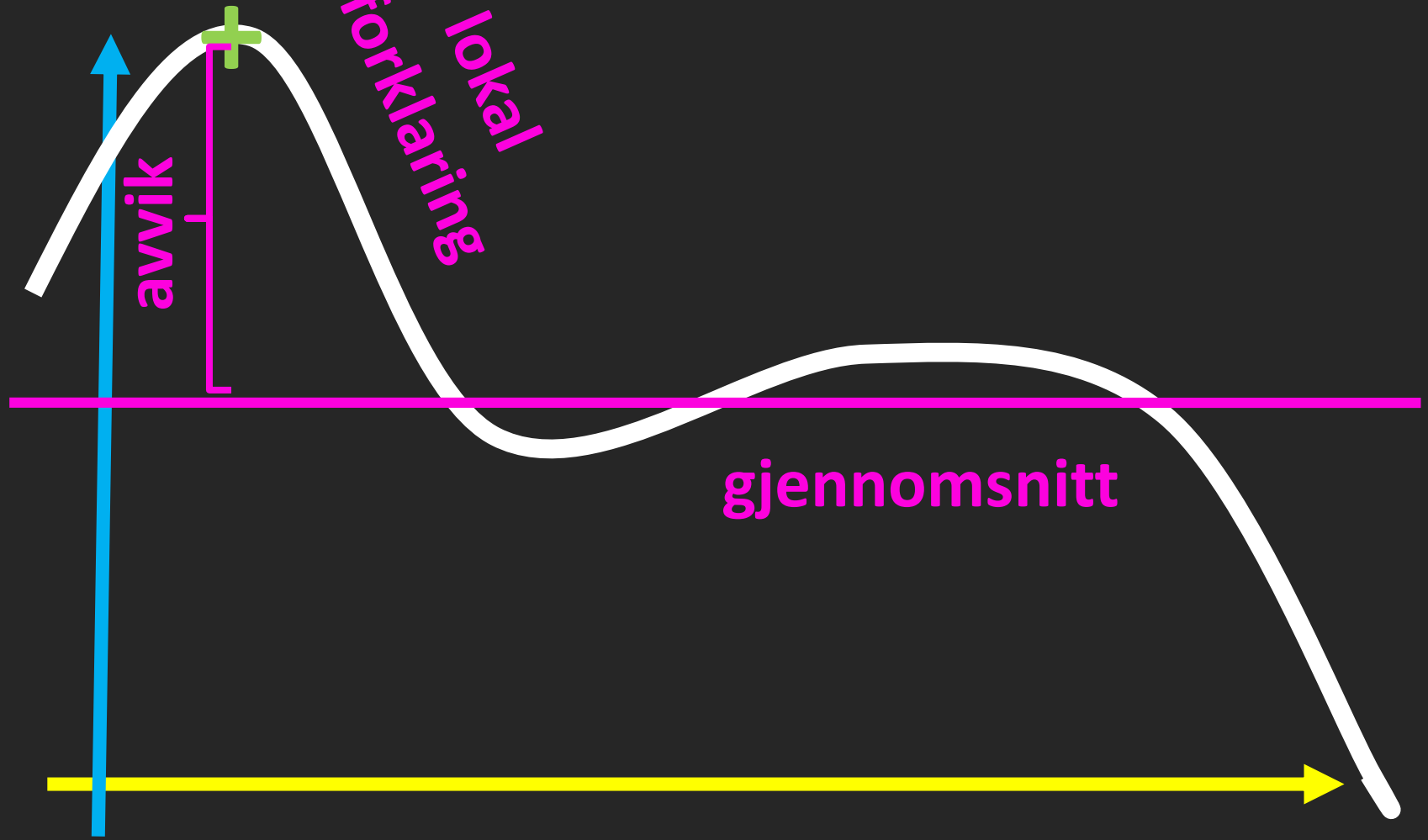
forklaring: b_1 og b_2 (og a)

hvis alder og studiepoeng er uavhengige...

forklaring av lokale modeller


dårlig forklaring


lokal forklaring
avvik
gjennomsnitt


lokal
forklaring
avvik
gjennomsnitt

modell*agnostiske* forklaringer:

Shapley-verdier

(LIME – Local Interpretable Model-agnostic Explanations)

maskinlæringsmodell:

47,75 % sannsynlighet

for at du ikke klarer å

betjene lånet ditt

maskinlæringsbanken:
du får ikke lån
(må under 20 %)

hvorfor 47,75 %?

3 inkassokrav

bor på bygda

yrke: bussjåfør

alder: 38 år

sivilstand: gift

Lloyd Shapley

"för teorin om stabila allokeringar och för utformning av marknadsinstitutioner i praktiken"

spillteoretisk/matematisk fundament

rettferdig fordeling etter innsats

rettferdig fordeling av bidrag til prediksjon fra maskinlærings-modell etter innsats fra egenskaper

forklarer avvik fra gjennomsnittet

summen av Shapley-verdiene = prediksjonen

to perfekt avhengige variable får hver ½ av Shapley-verdien

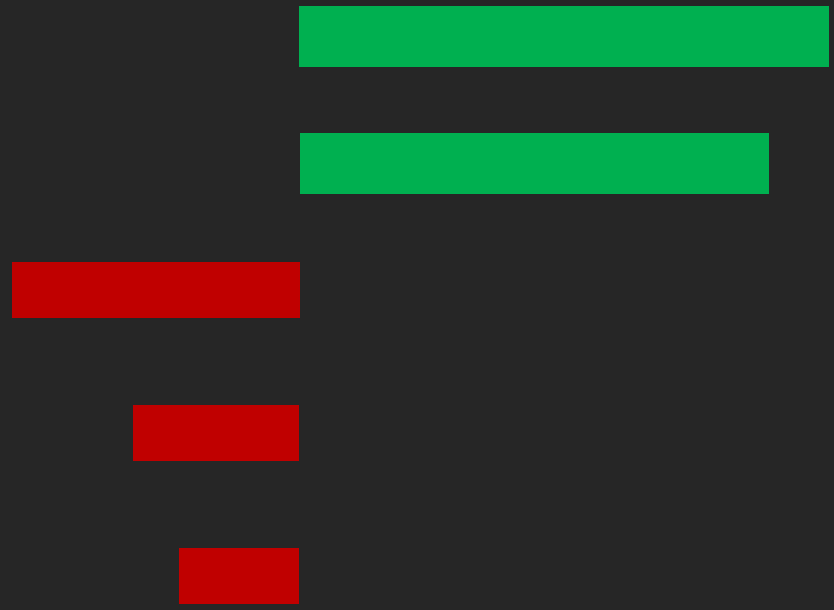
3 inkassokrav

bor på bygda

yrke: bussjåfør

alder: 38 år

sivilstand: gift



hvis du ikke hadde fått inkassokrav, ville du fått lån

kontrafaktisk forklaring

hvis du hadde bodd i byen og vært professor, ville du fått lån

3 inkassokrav

bor på bygda

yrke: professor

alder: 17 år

sivilstand: skilt

antall inkassokrav

bosted

yrke

alder

sivilstand

antall inkassokrav

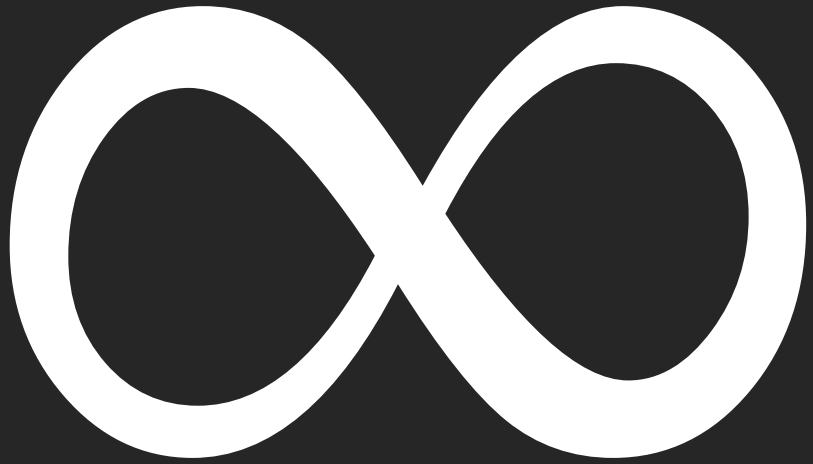
~~bosted~~

~~yrke~~

~~alder~~

~~sivilstand~~


du
lån ikke innvilget
lån innvilget
alder
antall inkassokrav

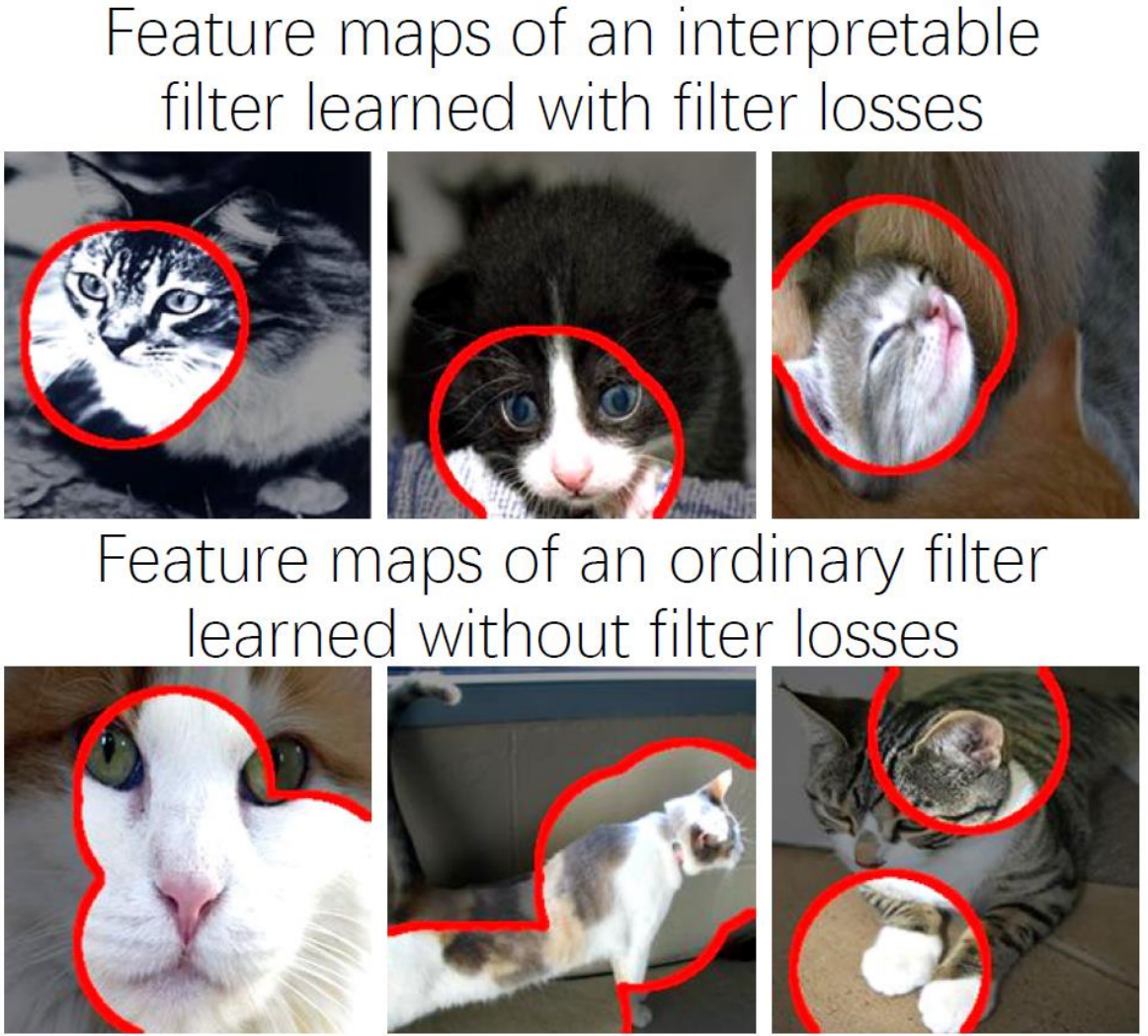


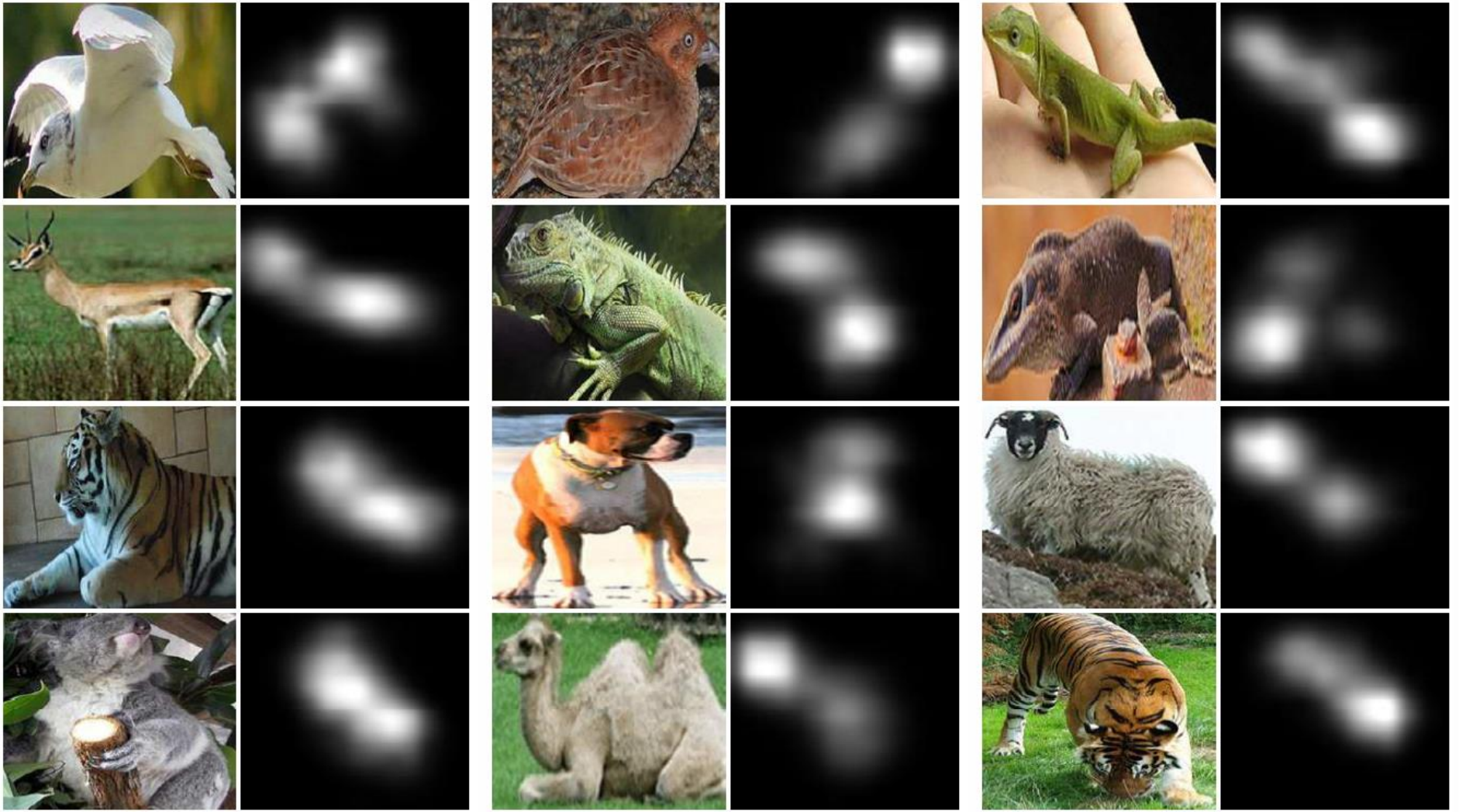
kontrafaktiske forklaringer

forklaring av modeller for **bilder** er både vanskeligere og enklere

modell-spesifikk forklaring

Zhang, Wu and Zhu (2018)
Interpretable CNNs
The IEEE Conference on Computer Vision and Pattern Recognition





til slutt: *hvilken forklaringsmetode er best?*