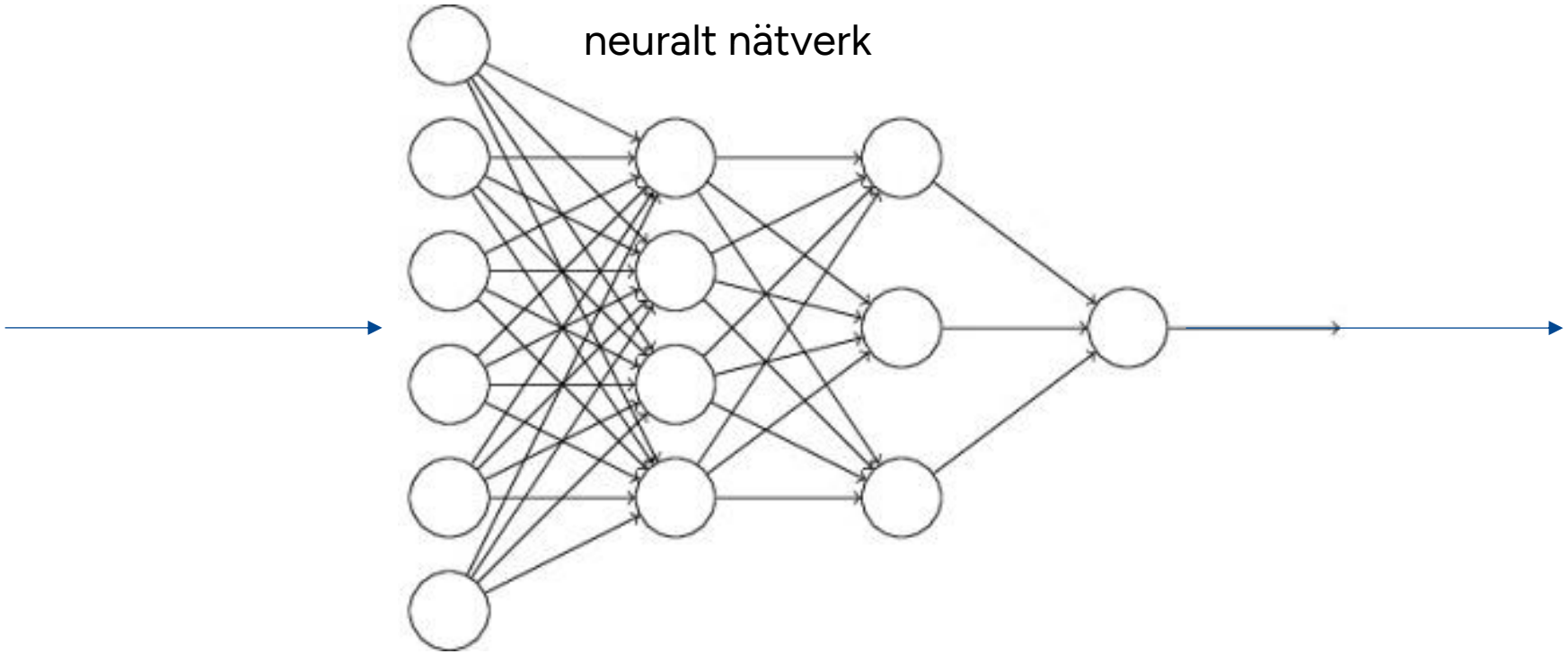
Varför funkar AI (inte)?

en föreläsning om geometrin bakom tekniken

Kathlén Kohn

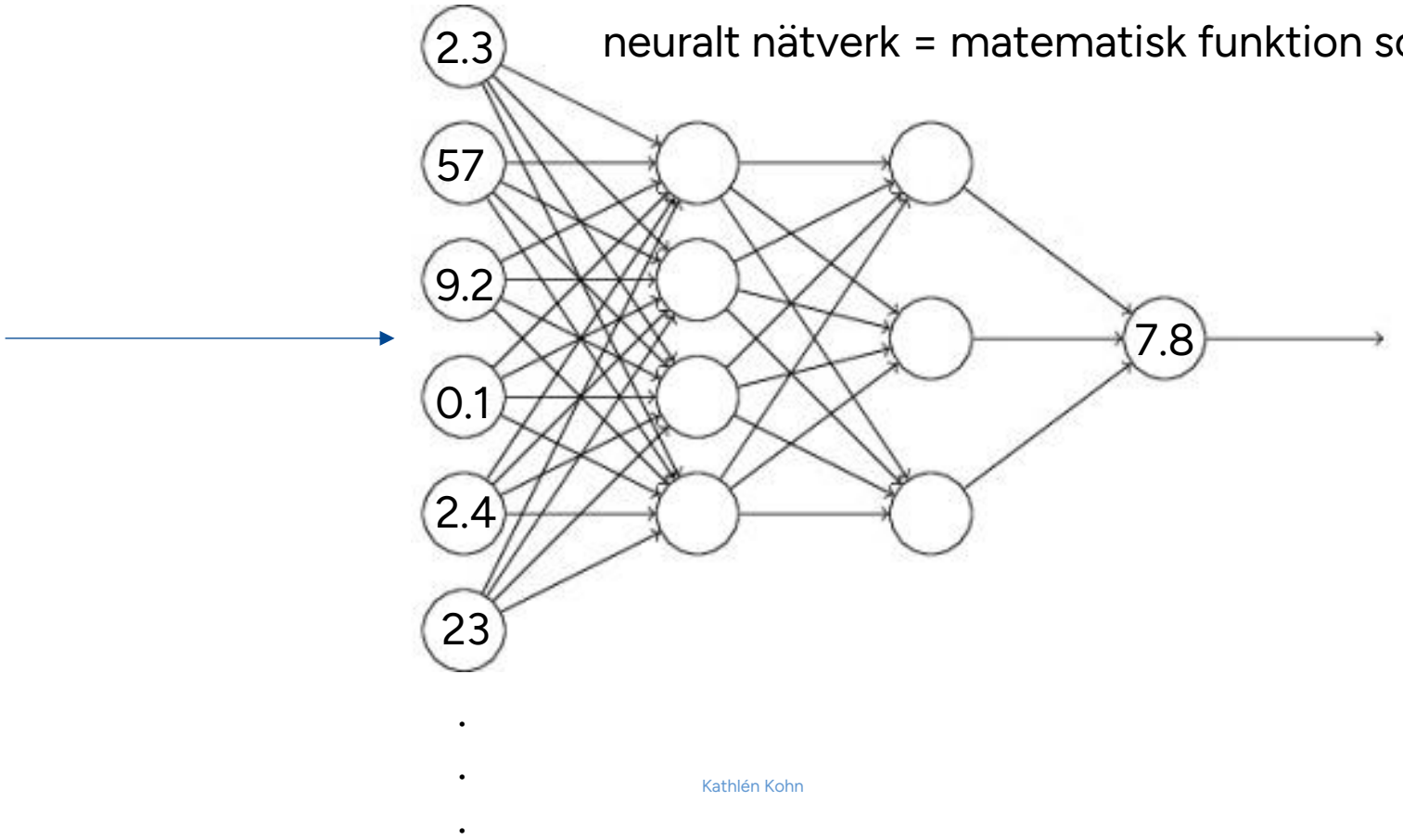
Vad är djupinlärning egentligen?



katt

hund

Vad är djupinlärning egentligen?

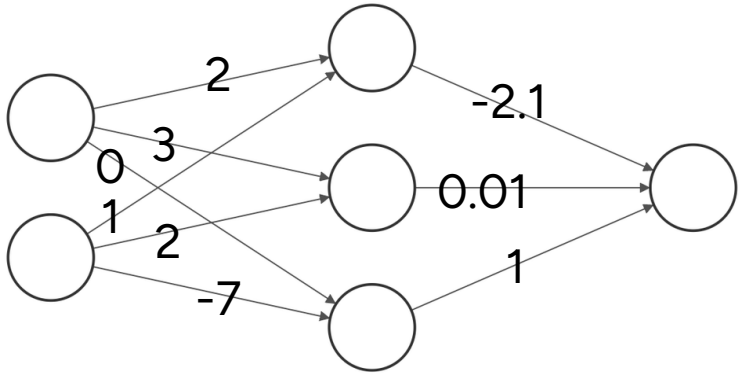


neuralt nätverk = matematisk funktion som gör nånting

om positiv: katt

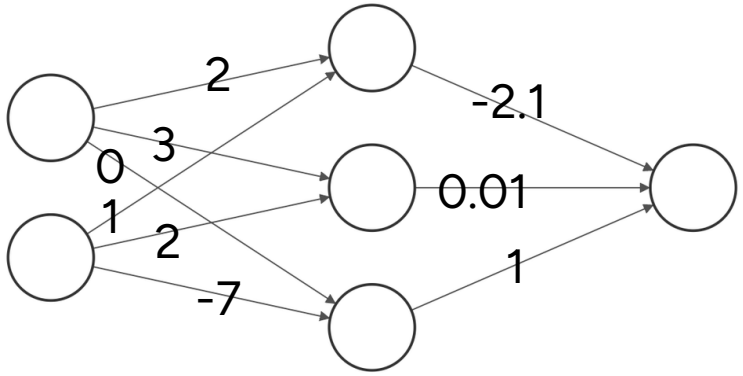
om negativ: hund

Ett mindre exempel

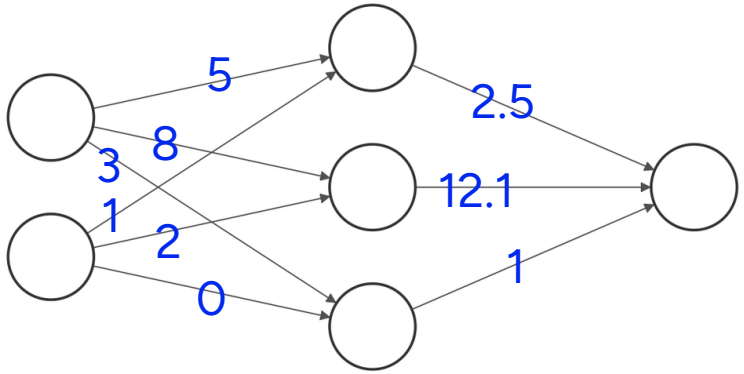


en matematisk funktion

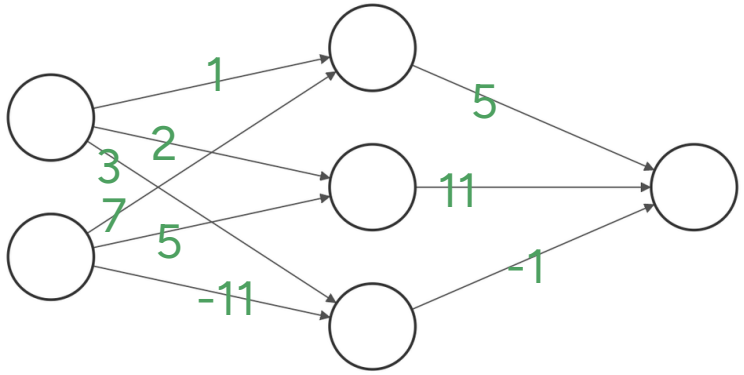
Ett mindre exempel



en matematisk funktion

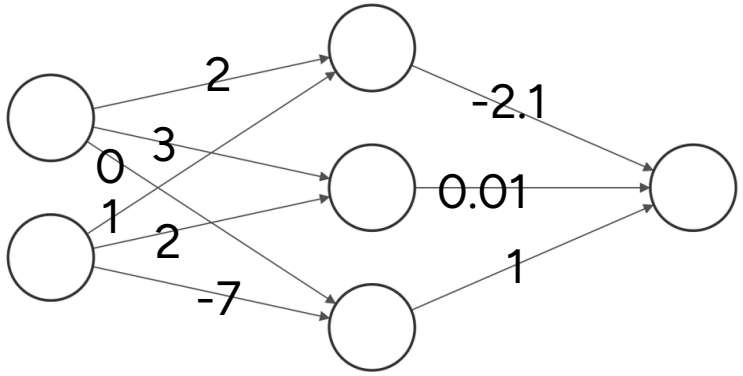


en annan matematisk funktion

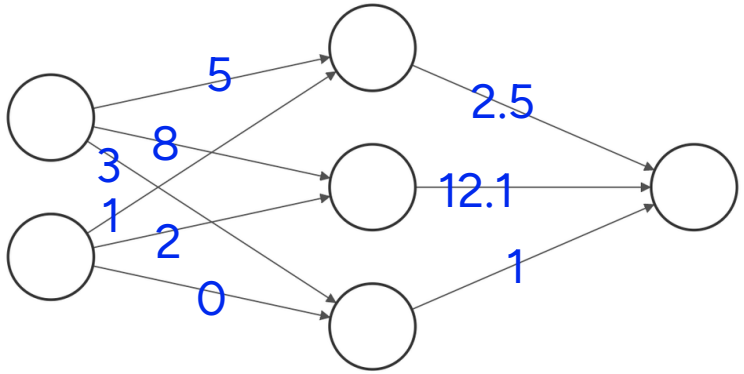


en till matematisk funktion

Ett mindre exempel

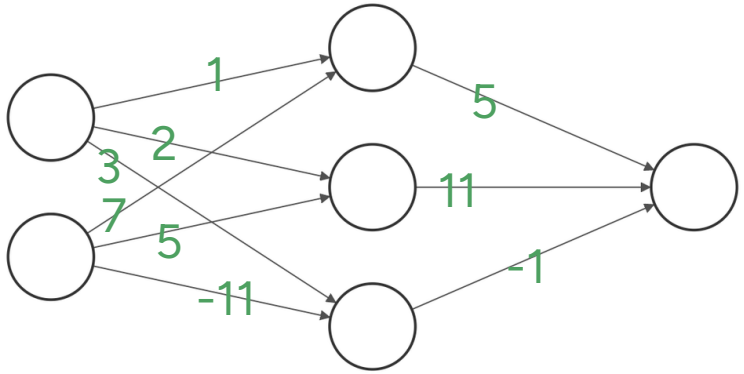


en matematisk funktion



en annan matematisk funktion

Djupinlärning försöker hitta den **bästa** matematiska funktionen som löser problemet, t.ex. klassificera katter och hundar.



en till matematisk funktion

Ibland verkar det funka, t.ex.

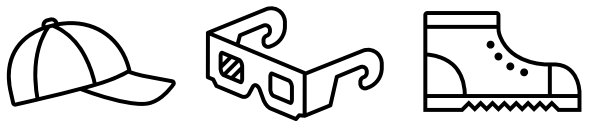
- känna igen föremål på bilder

katt

- slutföra meningar

”Katten som äter lasagne är… orange”

- rekommendera shoppingartiklar



Ibland verkar det funka, t.ex.

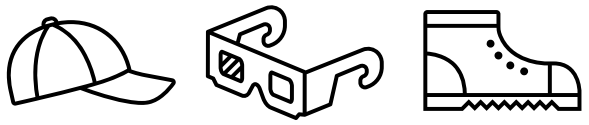
- känna igen föremål på bilder

katt

- slutföra meningar

”Katten som äter lasagne är… orange”

- rekommendera shoppingartiklar



ibland inte

- skilja jämna och udda tal

2 365 1 4 14 42 77 1255559

Ibland verkar det funka, t.ex.

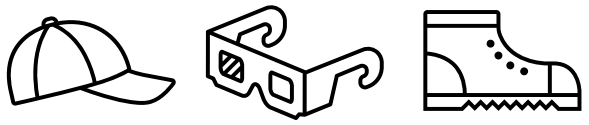
- känna igen föremål på bilder

katt

- slutföra meningar

”Katten som äter lasagne är… orange”

- rekommendera shoppingartiklar



ibland inte

- skilja jämna och udda tal

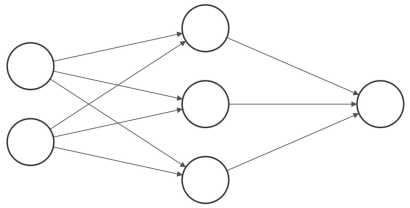
2 365 1 4 14 42 77 1255559

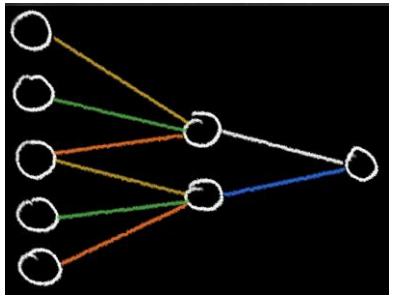
- hallucinationer

”katten fångade en fågel”

Vi har ingen teori som verkligen förklarar varför djupinlärning fungerar så bra som den gör!
(vi har bara pusselbitar)

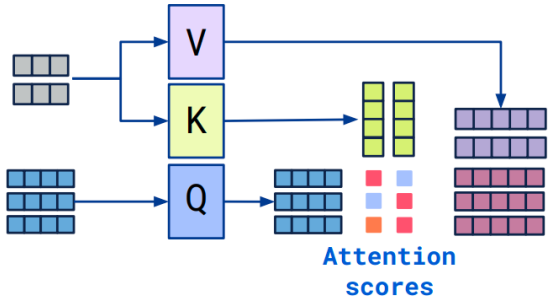
Vi har ingen teori som verkligen förklarar varför djupinlärning fungerar så bra som den gör!
(vi har bara pusselbitar)



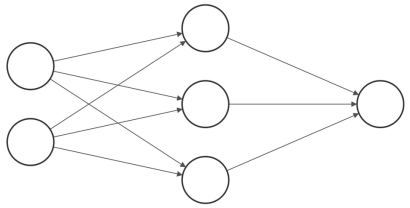


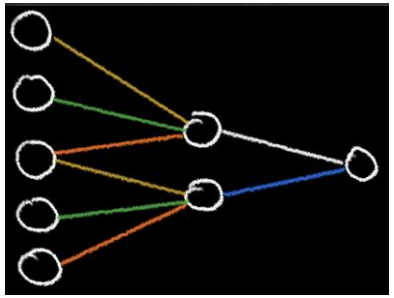
Praktiska skäl:

- det finns ett zoo av nätverksarkitekturer; moderna nätverk använder en komplicerad mix



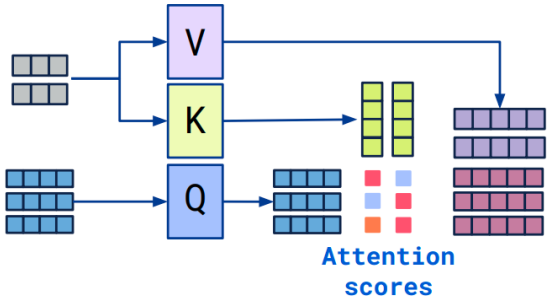
Vi har ingen teori som verkligen förklarar varför djupinlärning fungerar så bra som den gör!
(vi har bara pusselbitar)



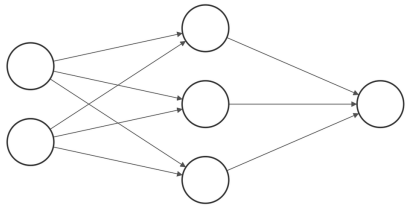


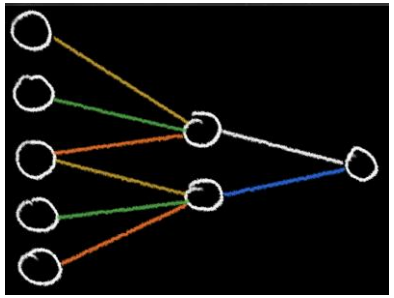
Praktiska skäl:

- det finns ett zoo av nätverksarkitekturer; moderna nätverk använder en komplicerad mix
- massa med ingenjörshack i praktiken



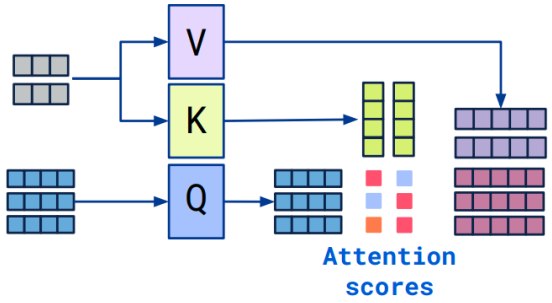
Vi har ingen teori som verkligen förklarar varför djupinlärning fungerar så bra som den gör!
(vi har bara pusselbitar)



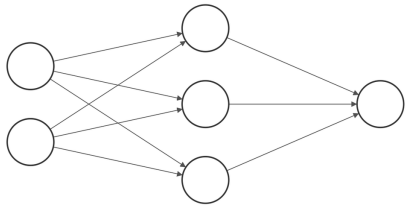


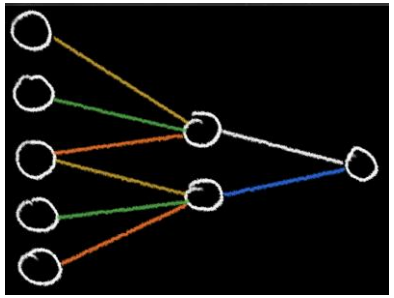
Praktiska skäl:

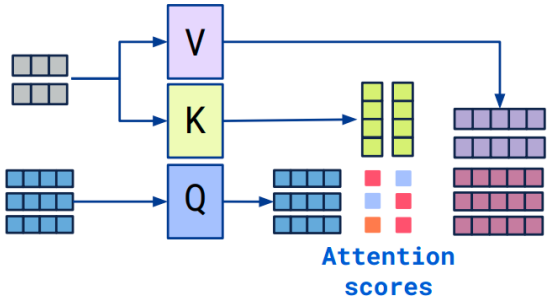
- det finns ett zoo av nätverksarkitekturer; moderna nätverk använder en komplicerad mix
- massa med ingenjörshack i praktiken
- extremt snabb praktisk utveckling



Vi har ingen teori som verkligen förklarar varför djupinlärning fungerar så bra som den gör!
(vi har bara pusselbitar)







Praktiska skäl:

- det finns ett zoo av nätverksarkitekturer; moderna nätverk använder en komplicerad mix
- massa med ingenjörshack i praktiken
- extremt snabb praktisk utveckling

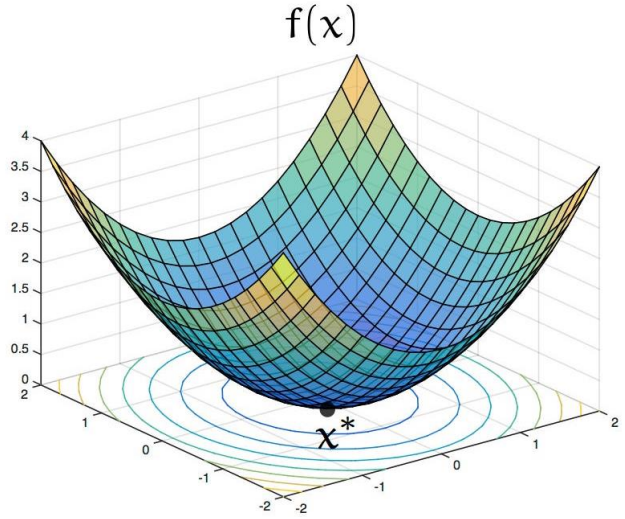
Oklara frågor: Är nätverksbeteendet en egenskap hos den använda arkitekturen? Orsakas det av några ingenjörshack? Kommer det från data man använde för träning? .....

Teoretisk förståelse tar mycket längre tid än praktiska experiment!

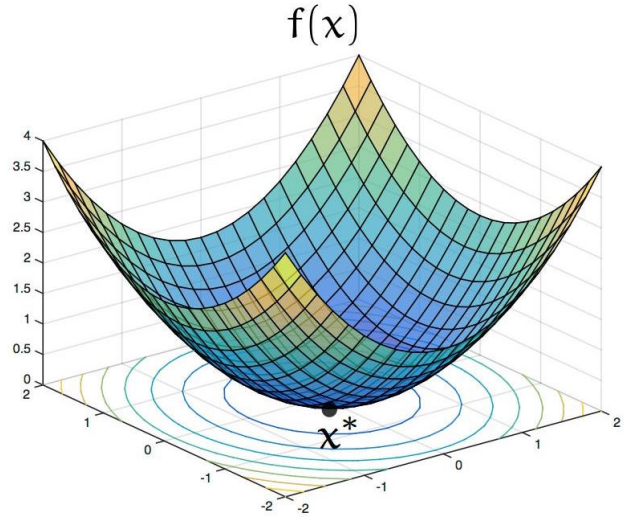
Vi har ingen teori som verkligen förklarar varför djupinlärning fungerar så bra som den gör!
(vi har bara pusselbitar)

Teoretiska skäl:

- Vi har en bra matematisk förståelse av ’konvex optimering’



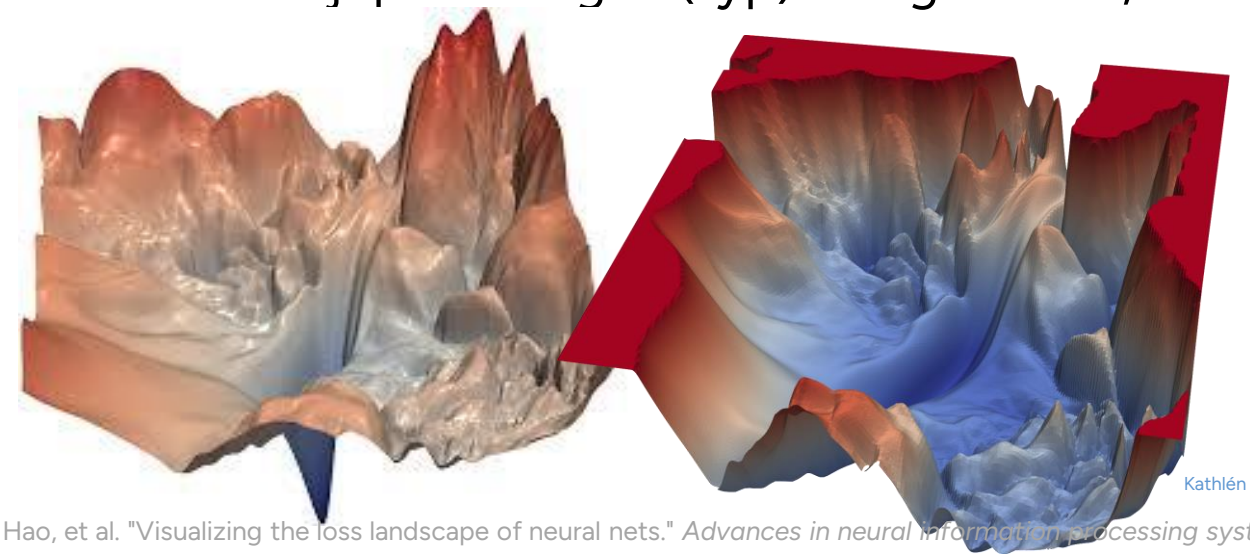
Vi har ingen teori som verkligen förklarar varför djupinlärning fungerar så bra som den gör!
(vi har bara pusselbitar)



Teoretiska skäl:

- Vi har en bra matematisk förståelse av 'konvex optimering'

- MEN: djupinlärning är (typ) aldrig konvex, ...



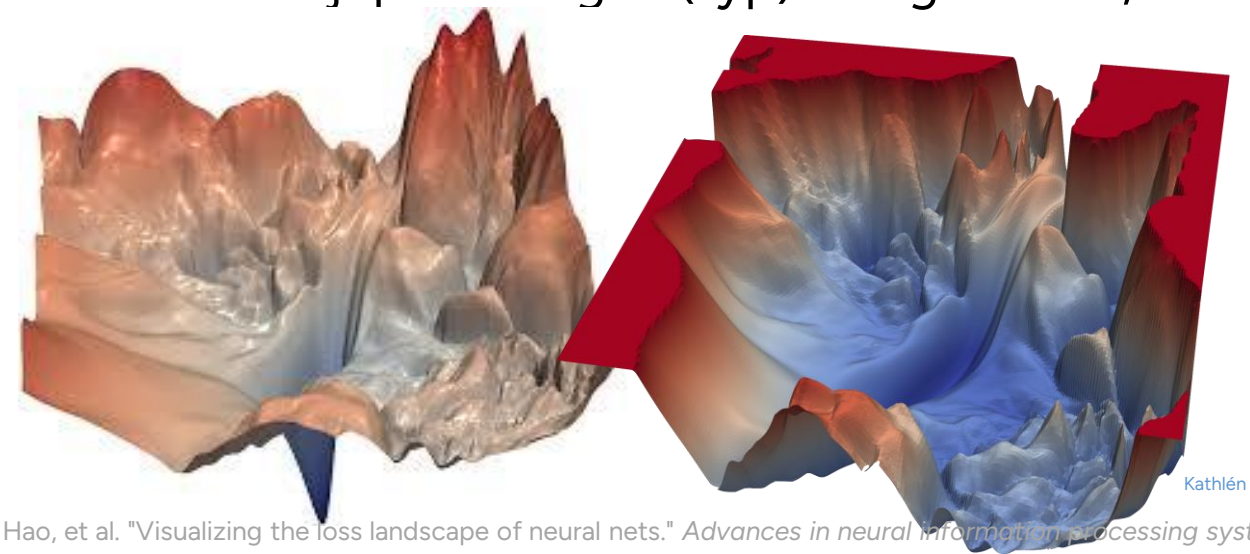
Li, Hao, et al. "Visualizing the loss landscape of neural nets." *Advances in neural information processing systems* 31 (2018).

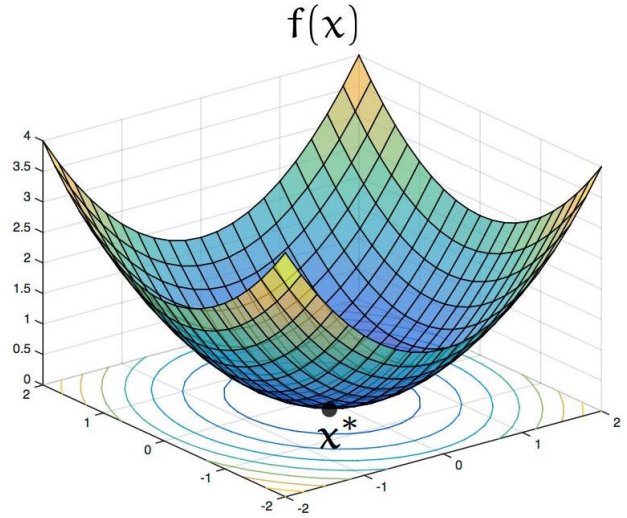
Vi har ingen teori som verkligen förklarar varför djupinlärning fungerar så bra som den gör!
(vi har bara pusselbitar)

Teoretiska skäl:

- Vi har en bra matematisk förståelse av ’konvex optimering’

- MEN: djupinlärning är (typ) aldrig konvex, …





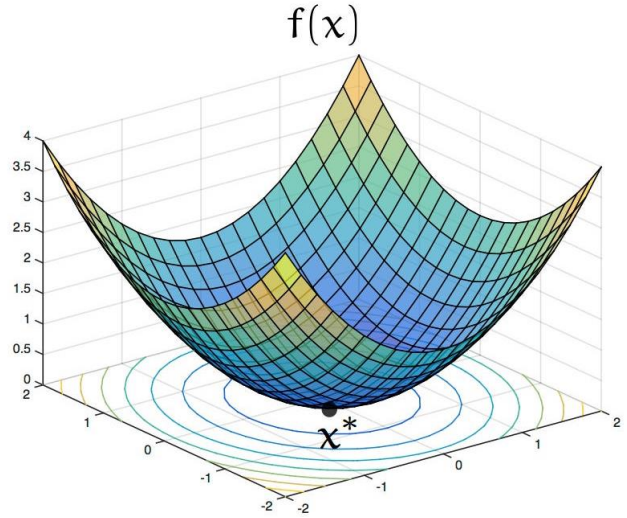
… inte ens förenklade leksaksnätverk,
så kallade ’linjära nät’,
som i princip aldrig används i praktiken.

Li, Hao, et al. "Visualizing the loss landscape of neural nets." *Advances in neural information processing systems* 31 (2018).

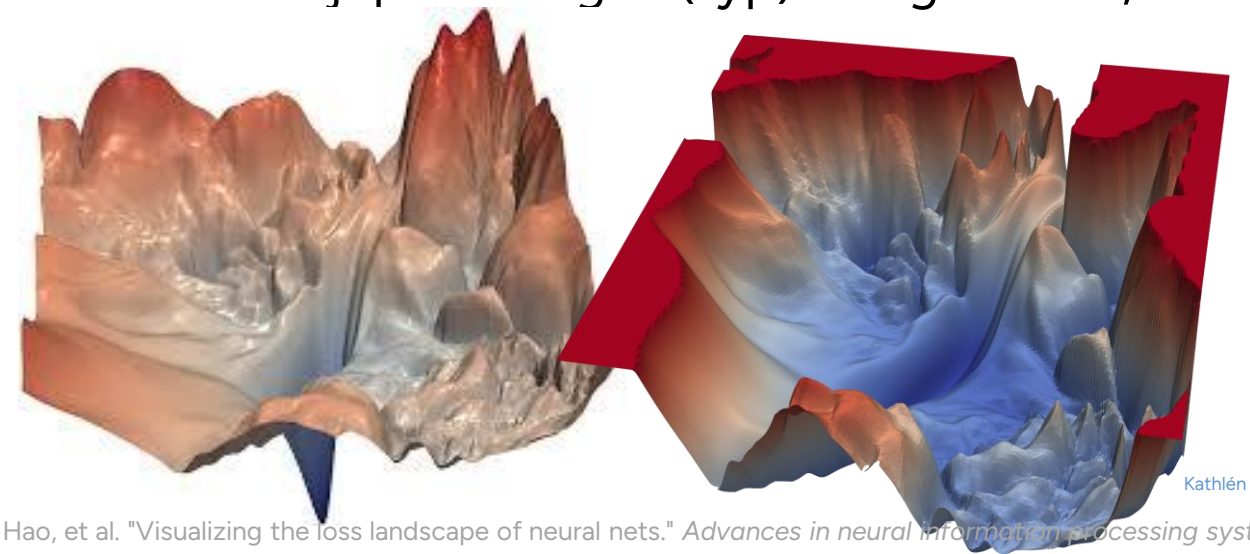
Vi har ingen teori som verkligen förklarar varför djupinlärning fungerar så bra som den gör!
(vi har bara pusselbitar)

Teoretiska skäl:

- Vi har en bra matematisk förståelse av 'konvex optimering'



- MEN: djupinlärning är (typ) aldrig konvex, ...



... inte ens förenklade leksaksnätverk,
så kallade 'linjära nät',
som i princip aldrig används i praktiken.

➢ egen teori utvecklad för linjära nät,
men den funkar sällan för praktiska nätverk

Li, Hao, et al. "Visualizing the loss landscape of neural nets." *Advances in neural information processing systems* 31 (2018).

Vi har endast experimentella belägg, ingen allmän teori, för följande frågor:

1. Varför fungerar djupinlärning så förvånansvärt bra?

2. Vilka är de exakta orsakerna till att den ibland misslyckas / hallucinerar?

3. Hur bör man designa en nätverksarkitektur för en given uppgift?

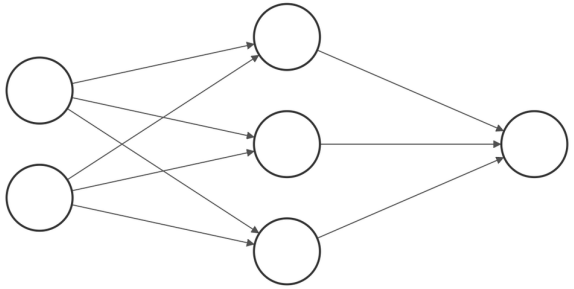
   ➢ Detta är mer konst än vetenskap!

klassificera katter och hundar, väderprognos, analysera dokument, osv.

på väg mot en allmän teori...

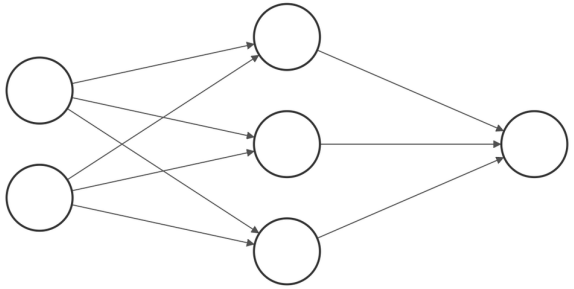
på väg mot en allmän teori...

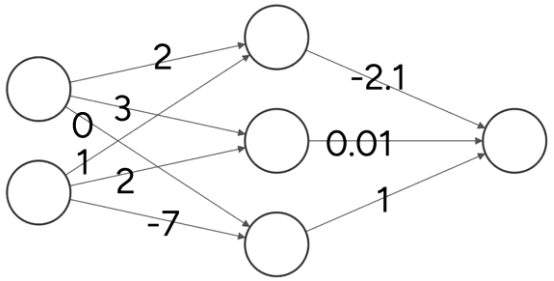
Kom ihåg: denna nätverksarkitektur ger många matematiska funktioner:



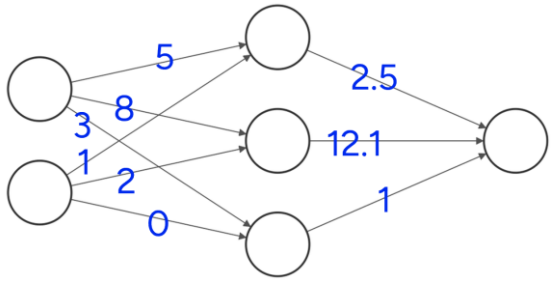
på väg mot en allmän teori...

Kom ihåg: denna nätverksarkitektur ger många matematiska funktioner:

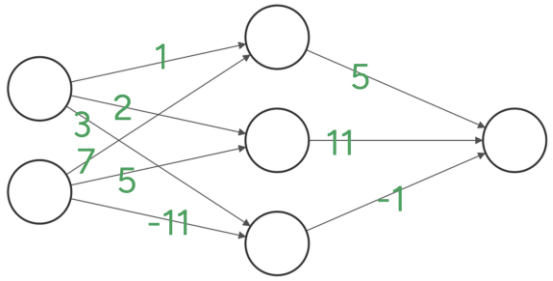




en matematisk funktion

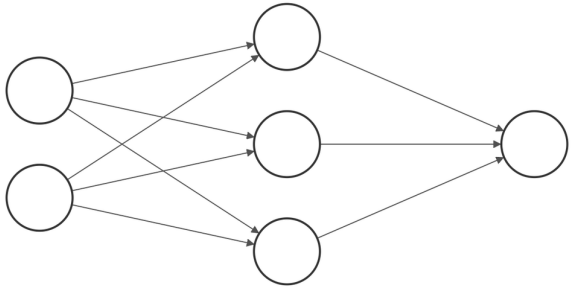


en annan matematisk funktion

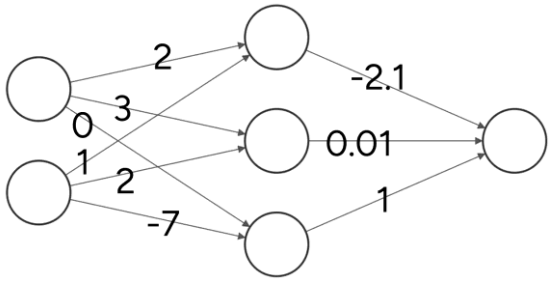


en till matematisk funktion

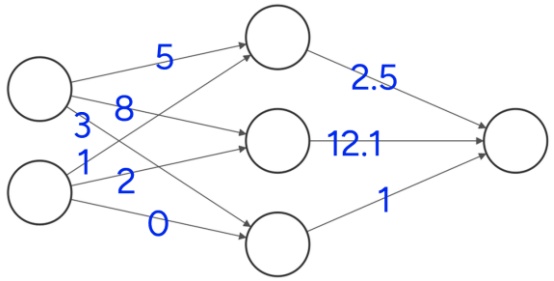
på väg mot en allmän teori...



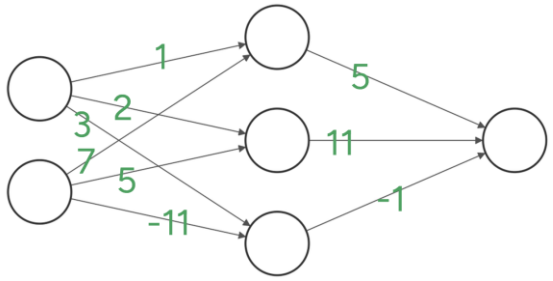
Kom ihåg: denna nätverksarkitektur ger många matematiska funktioner:



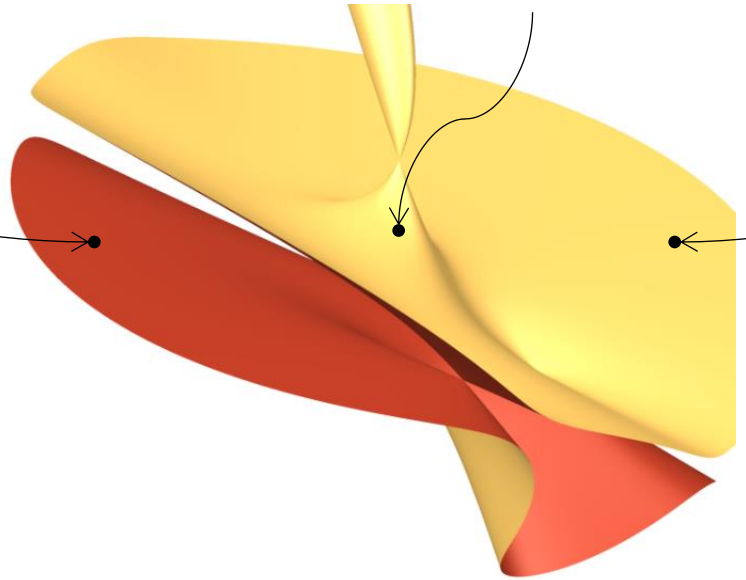
en matematisk funktion



en annan matematisk funktion

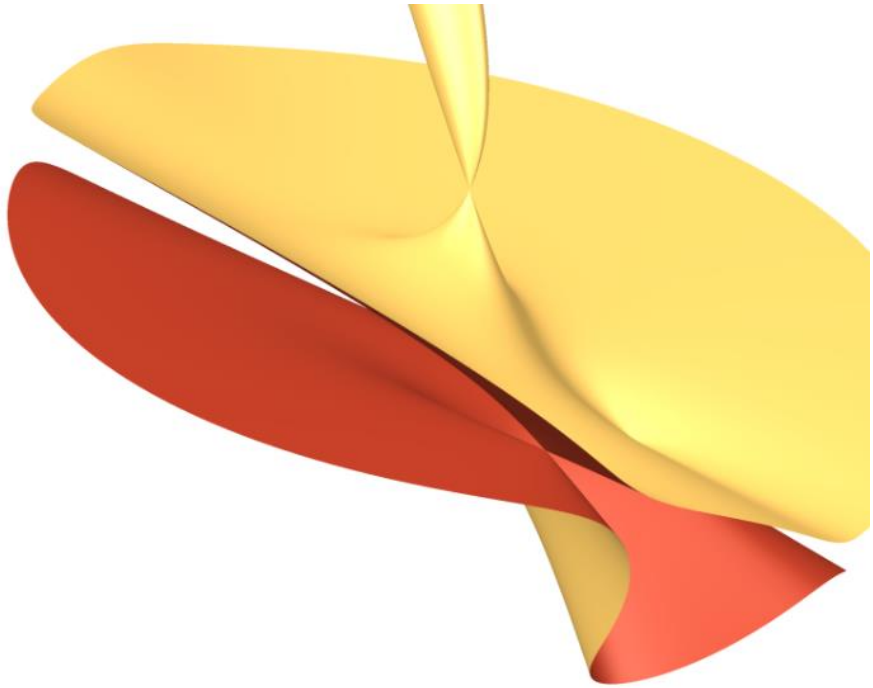


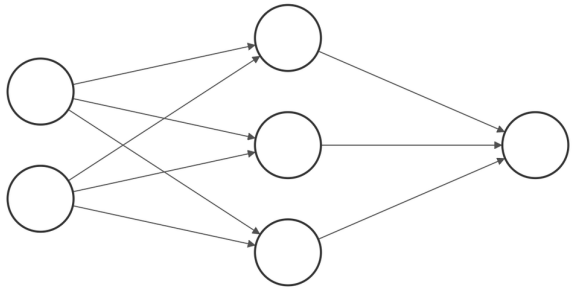
en till matematisk funktion

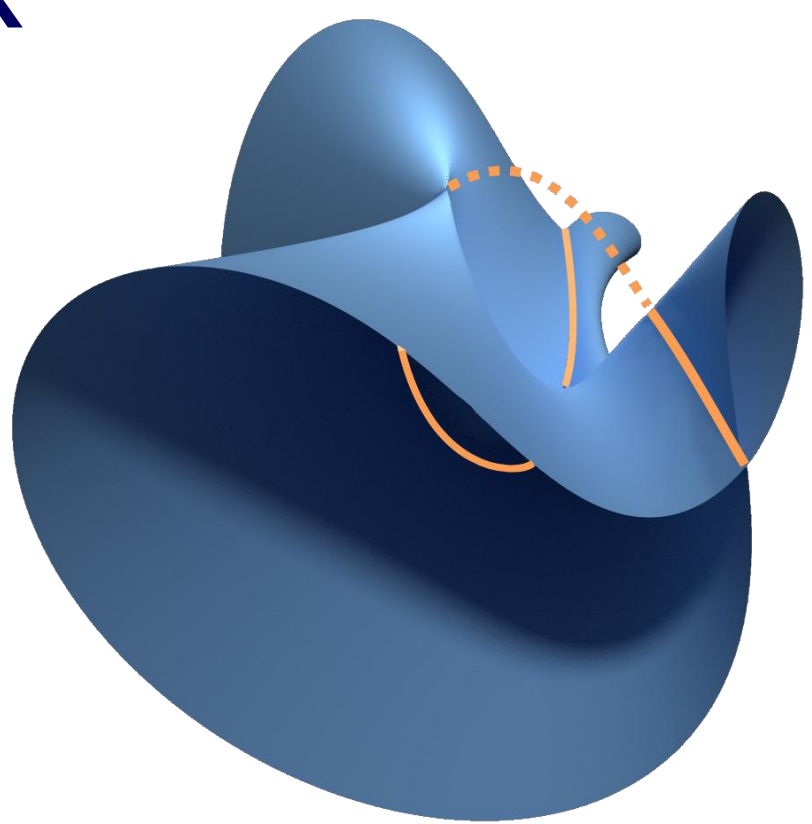


= rummet av alla funktioner som kommer från denna nätverksarkitektur

geometrin av neurala nätverk

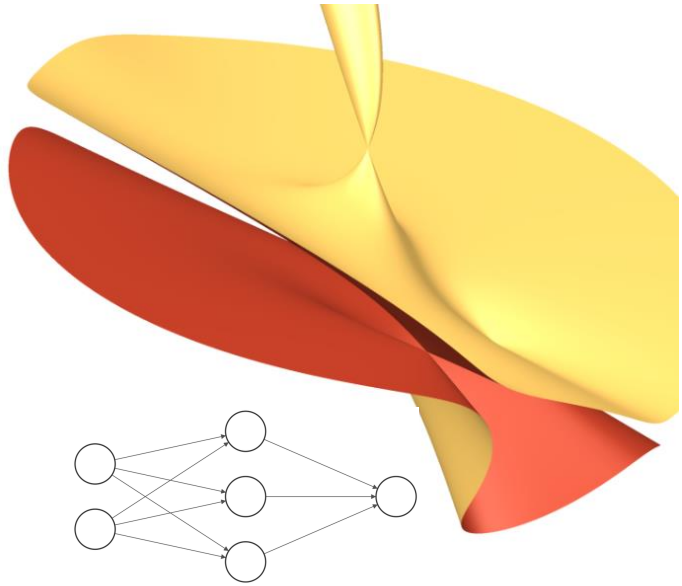






Baby-ChatGPT

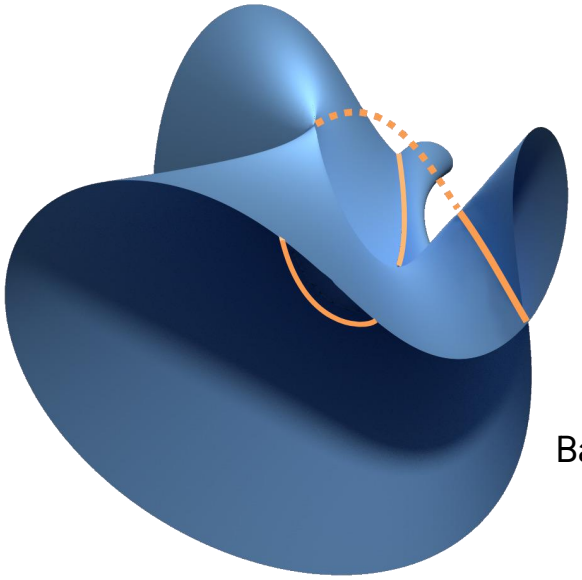
geometrin av neurala nätverk



1. Vilka geometriska egenskaper har funktionsrummen av neurala nätverk?

2. Hur påverkar den geometriska formen den praktiska djupinlärningen?

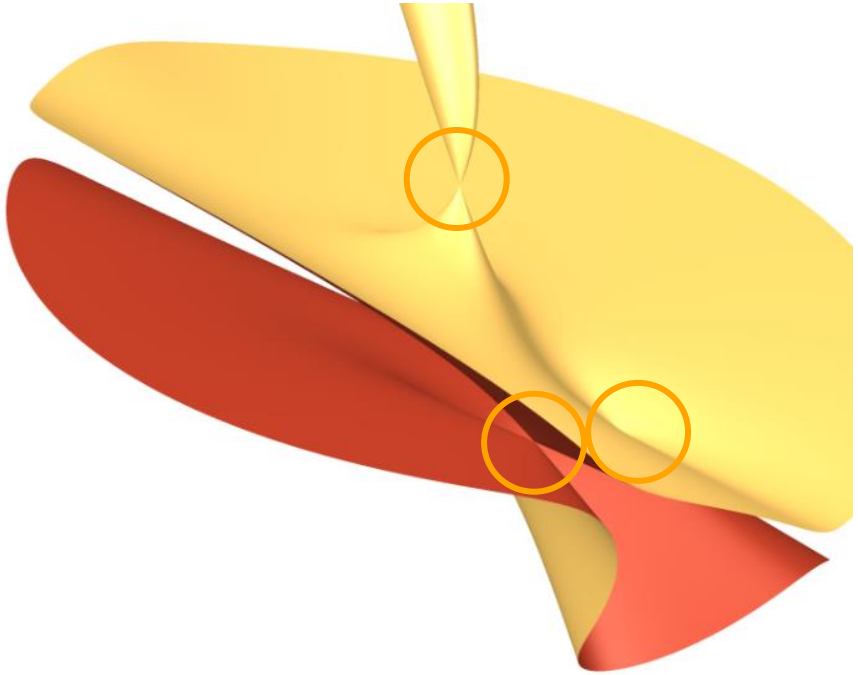
3. Jämför formen hos olika nätverksarkitekturer !



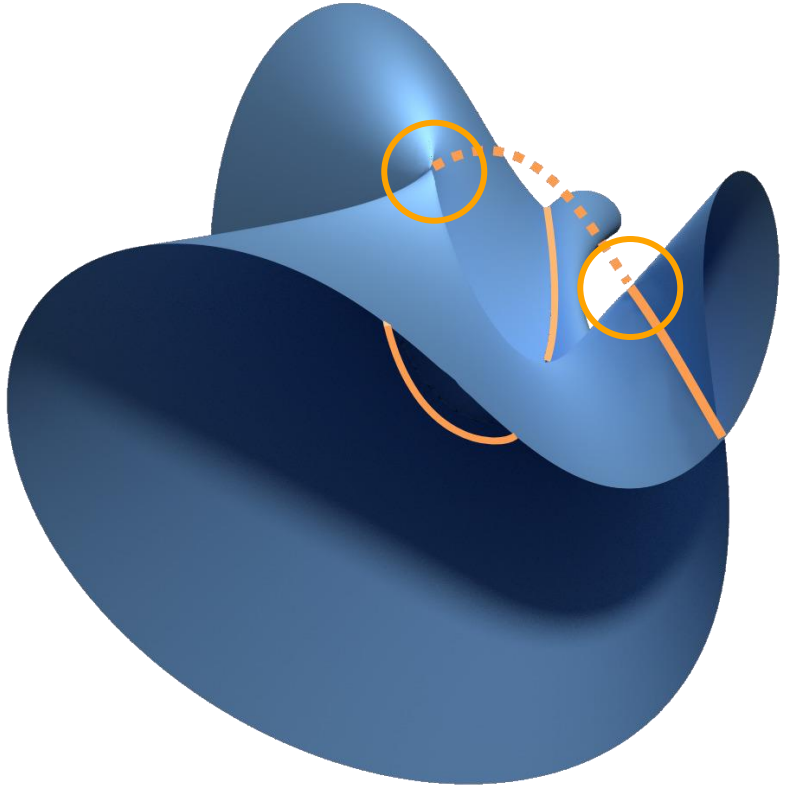
Baby-ChatGPT

ett exempel

1. Vilka geometriska egenskaper har funktionsrummen av neurala nätverk?



singularitäter

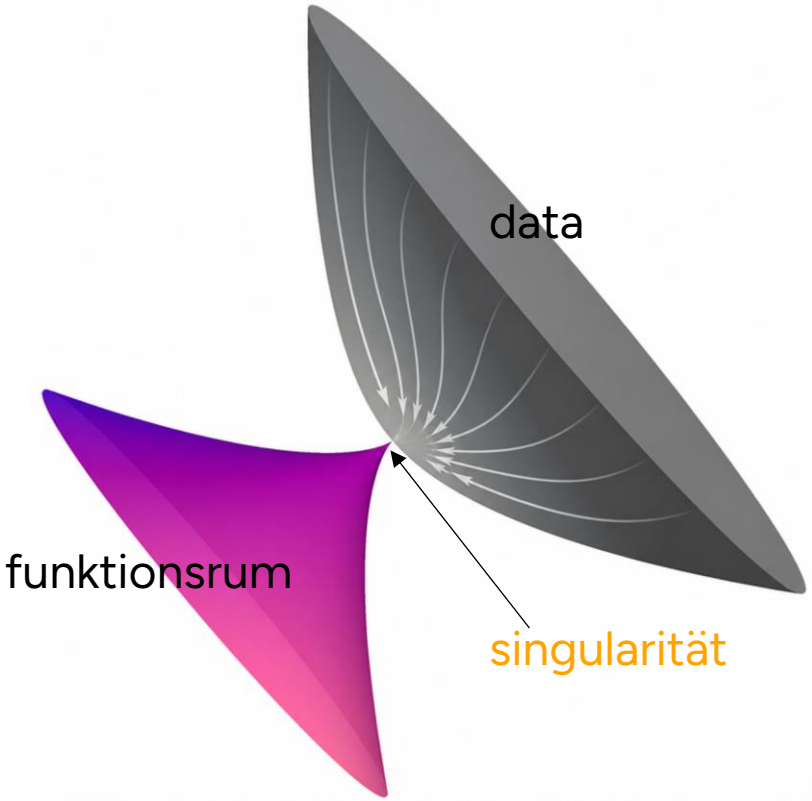


ett exempel

2. Hur påverkar den geometriska formen den praktiska djupinlärningen?

ett exempel

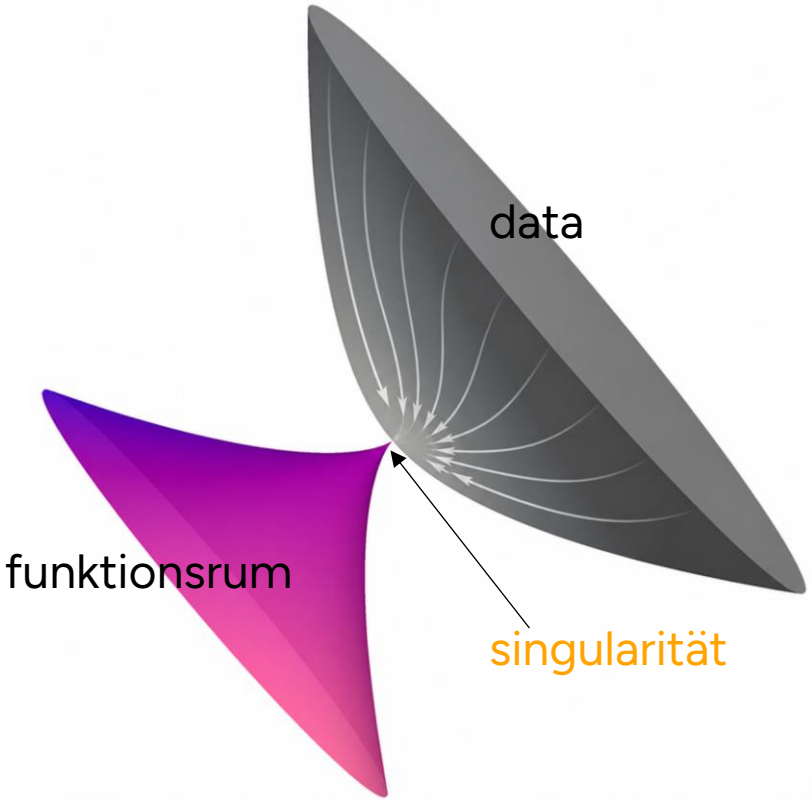
2. Hur påverkar den geometriska formen den praktiska djupinlärningen?



Nätverksarkitekturen föredrar dessa singulära funktioner framför andra (”implicit bias”).

ett exempel

2. Hur påverkar den geometriska formen den praktiska djupinlärningen?

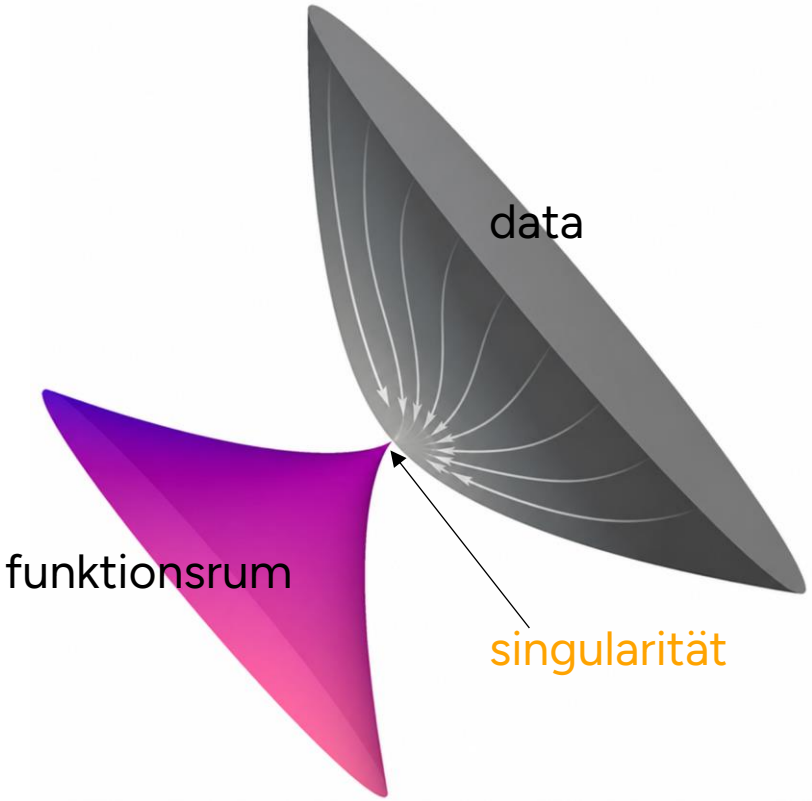


Nätverksarkitekturen föredrar dessa singulära funktioner framför andra (”implicit bias”).

➢ De gör att nätverket funkar bra på nya indata! ☺

ett exempel

2. Hur påverkar den geometriska formen den praktiska djupinlärningen?



Nätverksarkitekturen föredrar dessa singulära funktioner framför andra (”implicit bias”).

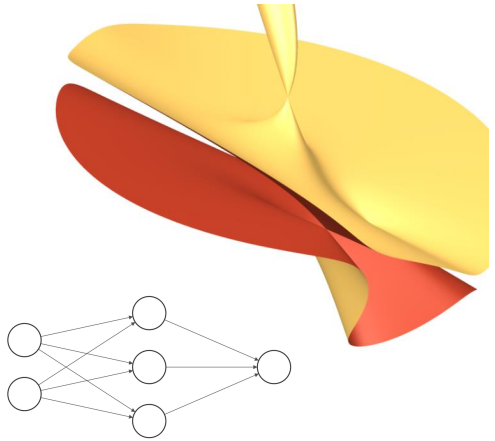
➢ De gör att nätverket funkar bra på nya indata! ☺

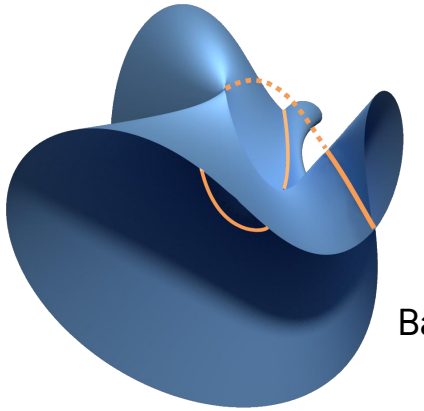
Det är matematiskt svårt att hitta dessa singulära funktioner när man tränar ett nätverk med standard gradientmetoder.

➢ Den geometriska bilden förklarar varför man behöver optimeringshacks som ”pruning”.

ett exempel

3. Jämför formen hos olika nätverksarkitekturer !





Baby-ChatGPT

Förhoppning:
utveckla matematiska designprinciper för neurala nätverksarkitekturer, beroende på uppgiften

(just nu: jätte experimentellt; som sagt: mer konst än vetenskap!)

klassificera katter och hundar,
väderprognos, analysera dokument, osv.

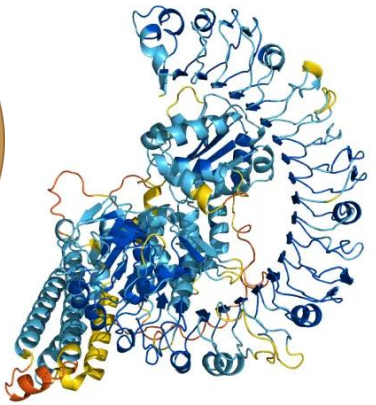
ett

3. Jä

Förho
utvec

(just nu

AlphaFold-storyn



- 2024 Nobelpris i kemi
- Djupinlärning för att förutsäga proteinstrukturer
- AlphaFold 2 har vacker matematisk design (inbyggda symmetrier)
- AlphaFold 3 har till största delen tagit bort det

väderprognos, analysera dokument, osv.

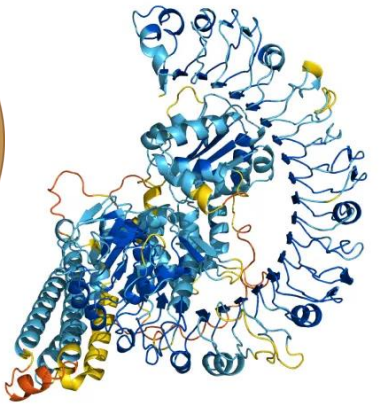
ett

3. Jä

Förho
utvec

(just nu

AlphaFold-storyn

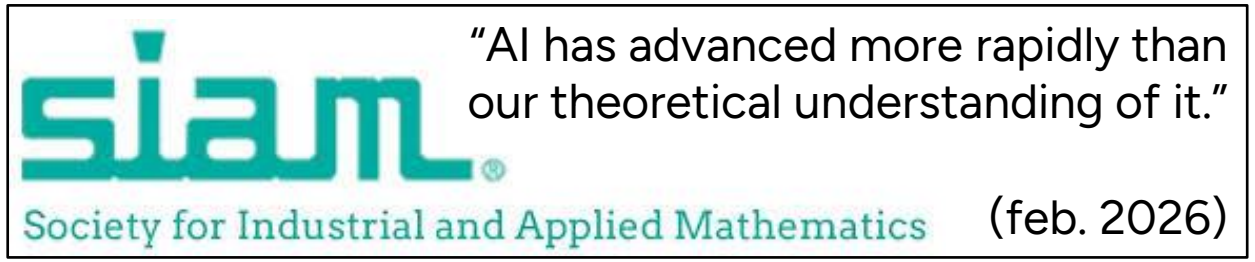


- 2024 Nobelpris i kemi
- Djupinlärning för att förutsäga proteinstrukturer
- AlphaFold 2 har vacker matematisk design (inbyggda symmetrier)
- AlphaFold 3 har till största delen tagit bort det

Frågor: Är AlphaFold 3 verkligen en fundamentalt bättre design än AlphaFold 2? Eller beror dess praktiska framgång delvis på ingenjörsval, implementeringsdetaljer, träningsupplägg och hårdvarukompatibilitet?

klassificera katter och hundar, väderprognos, analysera dokument, osv.

sammanfattning



Vi har ingen holistisk teori som förklarar framgångarna och misslyckandena hos moderna AI-system, eller deras design för olika uppgifter.

sammanfattning

> “AI has advanced more rapidly than our theoretical understanding of it.”
>
> Society for Industrial and Applied Mathematics (feb. 2026)

Vi har ingen holistisk teori som förklarar framgångarna och misslyckandena hos moderna AI-system, eller deras design för olika uppgifter.

Långsiktigt: Jag tror att vi kommer att ha en sådan teori.

sammanfattning

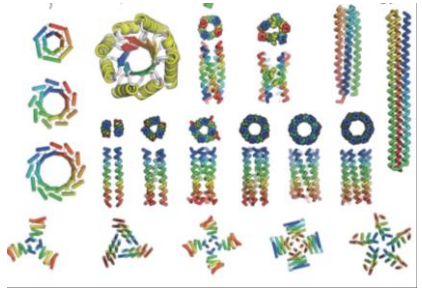
> “AI has advanced more rapidly than our theoretical understanding of it.”
>
> Society for Industrial and Applied Mathematics (feb. 2026)

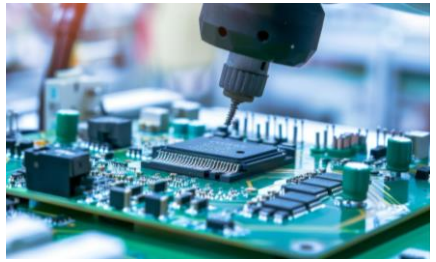
Vi har ingen holistisk teori som förklarar framgångarna och misslyckandena hos moderna AI-system, eller deras design för olika uppgifter.

Långsiktigt: Jag tror att vi kommer att ha en sådan teori.

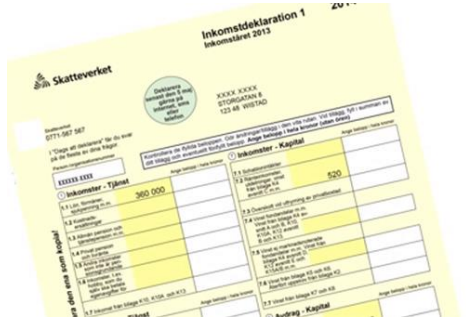
Kortsiktigt: Använd AI-system för att hitta potentiella lösningar på problem, och använd sedan **formell verifiering** för att kontrollera att lösningen är korrekt.

```
for i in people.data.users:
    response = client.api.statuses.user_timeline.get(screen_name=i.scre
    print 'Got', len(response.data), 'tweets from', i.screen_name
    if len(response.data) != 0:
        ltdate = response.data[0]['created_at']
        ltdate2 = datetime.strptime(ltdate,'%a %b %d %H:%M:%S +0000 %Y'
        today = datetime.now()
        howlong = (today-ltdate2).days
        if howlong < daywindow:
            print i.screen_name, 'has tweeted in the past' , daywindow,
            totaltweets += len(response.data)
            for j in response.data:
                if j.entities.urls:
                    for k in j.entities.urls:
                        newurl = k['expanded_url']
                        urlset.add((newurl, j.user.screen_name))
        else:
            print i.screen_name, 'has not tweeted in the past', daywind
```





Kathlén Kohn



Matthew Trager
Amazon Web Services AI Labs, NY

Guido Montúfar
University of California Los Angeles &
Max-Planck Institute for Mathematics in the Sciences Leipzig

Joan Bruna
Courant Institute
New York University