AI för naturligt språk

Vad är naturligt språk-behandling?

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Vad är naturligt språk-behandling?

- **Naturligt språk-behandling** utvecklar metoder för att förstå och producera mänskligt språk med hjälp av datorer.

  eng. *natural language processing (NLP)*

- Några välkända tillämpningar av naturligt språk-behandling är smarta sökmotorer, maskinöversättning och digitala assistenter.

- I den här kursen kommer vi att fokusera på behandling av naturligt språk i textform (snarare än talat språk).

naturligt språk-behandling

Private · Research · Teaching · LiU

Google

naturligt språk-behandling

All · Images · Videos · Shopping · News · More · Settings · Tools

About 4 820 000 results (0,57 seconds)

Inom Natural Language Processing (NLP) är fokus på semantik och tekniska system med förmåga att förstå mänskligt/**naturligt språk**, vilket inkluderar automatisk tolkning av innebörd genom analys av beroenden, identifiering av referenser och temporal ordning.
6 Nov 2015

cs.lth.se › forskning › naturligt-spraakbehandling

Naturligt språkbehandling | Datavetenskap

About featured snippets • Feedback

medium.com › den-växande-markna... · Translate this page

Den växande marknaden för naturlig språkbehandling | by ...

Med hjälp av **naturlig språkbehandling**, vilket är drivet av AI-tekniken bakom maskin- och djupinlärning, så blir det enklare för kunder att göra sig förstådda för ...

dator8.info › 2012/10 › naturligt-spr... · Translate this page

naturligt språk behandling verktyg - Dator

Dator > **naturligt språk behandling** verktyg ... lättare interagera med datorn , och många anser att interagera med en dator via naturligt tal att vara mer naturlig .

ai-competence.se › topic › behandlin... · Translate this page

Behandling av naturliga språk – AI Competence for Sweden

Behandling av **naturliga språk** (eng. Natural Language Processing) handlar om samband mellan datorer och mänskliga, **naturliga språk**, i synnerhet hur datorer ...

sv.qaz.wiki › wiki › Natural_languag... · Translate this page

Naturlig språkbehandling - Natural language processing - qaz ...

17 Feb 2021 — **Naturlig språkbehandling** (NLP) är ett underfält av lingvistik , datavetenskap ... Utmaningar vid bearbetning av naturligt språk involverar ofta ...

techworld.idg.se › python-bibliotek-... · Translate this page

8 bra Pythonbibliotek för NLP – behandling av naturliga språk ...

15 Aug 2020 — Det finns mycket att hämta från Pythons bibliotek när det kommer till **behandling** av **naturliga språk**, eller det som vi till vardags kallar NLP: ...

en.glosbe.com › Swedish-Swedish dictionary

Naturligt språkbehandling - Swedish definition, grammar ...

Learn the definition of '**Naturligt språkbehandling**'. Check out the pronunciation, synonyms and grammar. Browse the use examples '**Naturligt språkbehandling**' ...

www.dumay.info › Article › Naturlig... · Translate this page

Naturliga språkbehandlingsverktyg - dumay

Naturliga språkbehandlingsverktyg: Postad av:Jeanette Morales. **Naturlig språkbehandling** är hur datorer och människor interagerar med naturligt mänskligt tal ...

Natural language processing (Naturligt språk-behandling)

Natural language processing is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. Wikipedia

Model

Technology

Programs

Textbook

People also search for

View 15+ more


Machine learning


Artificial intelligence

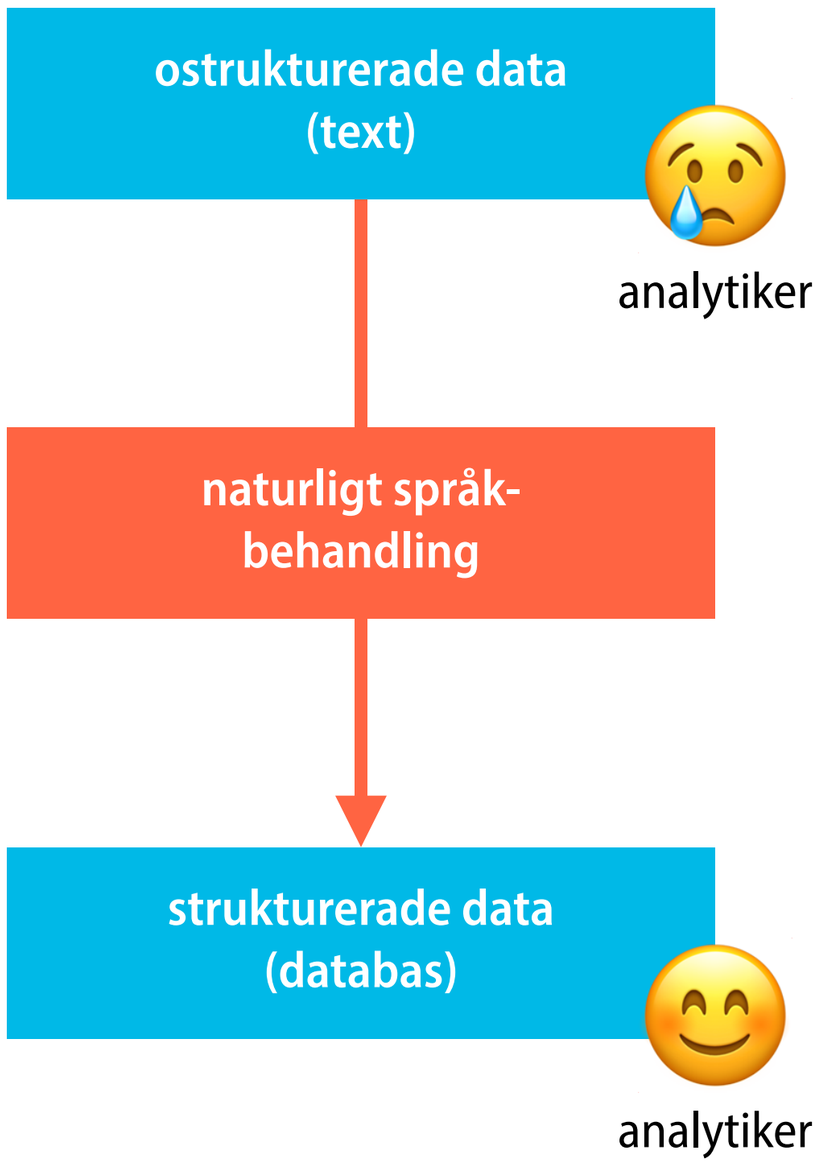

Computer vision


Artificial neural network


Internet of things

Feedback

Kunskapsglappet



Jeopardy!

Hon är känd för sitt arbete inom naturligt språk-behandling och föddes i Huddersfield.

University of Cambridge, CC BY 2.0, via Wikimedia Commons

SPARQL-sökfråga mot DBPedia

```
SELECT DISTINCT ?x WHERE {
  ?x dbo:knownFor dbr:Natural_language_processing.
  ?x dbo:birthPlace dbr:Huddersfield.
}
```

AI för naturligt språk

Två paradigmer

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Jeopardy!

Hon är känd för sitt arbete inom naturligt språk-behandling och föddes i Huddersfield.

University of Cambridge, CC BY 2.0, via Wikimedia Commons

SPARQL-sökfråga mot DBPedia

```
SELECT DISTINCT ?x WHERE {
  ?x dbo:knownFor dbr:Natural_language_processing.
  ?x dbo:birthPlace dbr:Huddersfield.
}
```

Två paradigmer

Eisenstein (2019), § 1.2.1

- **Lingvistisk kunskap**

  Bygga kedjor av systemkomponenter som producerar lingvistiskt motiverade representationer.

  ordklasser, syntaxträd, semantiska representationer

- **Djupinlärning**

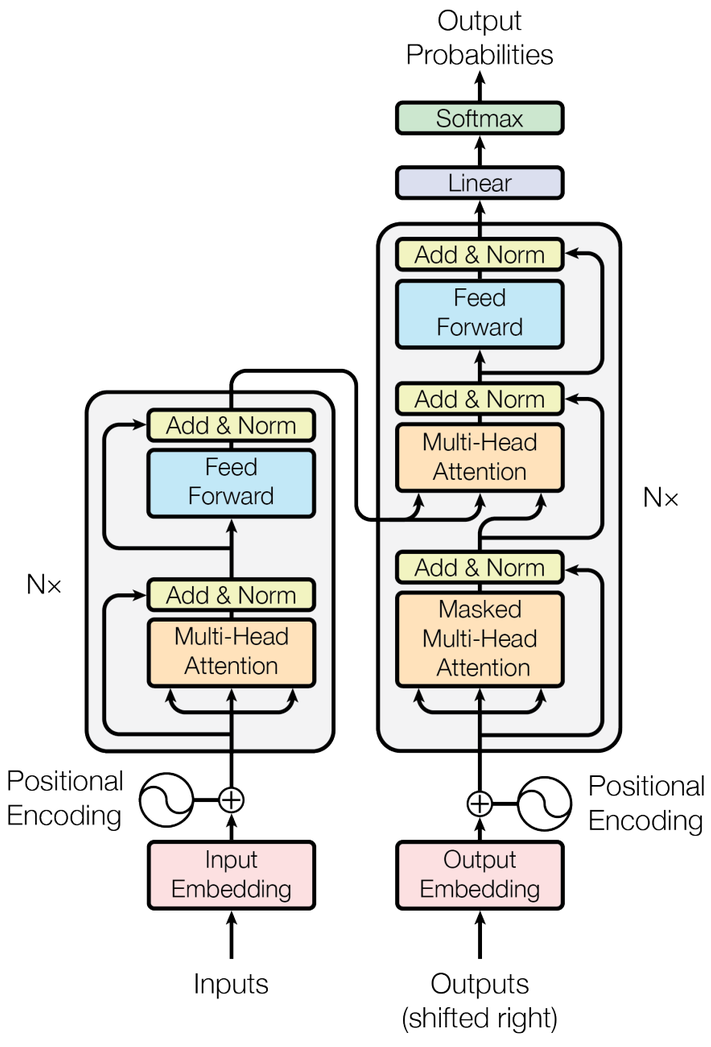
  Träna djupa neuronnät som direkt transformerar rå text till den struktur som tillämpningen behöver.

Lingvistiska representationer

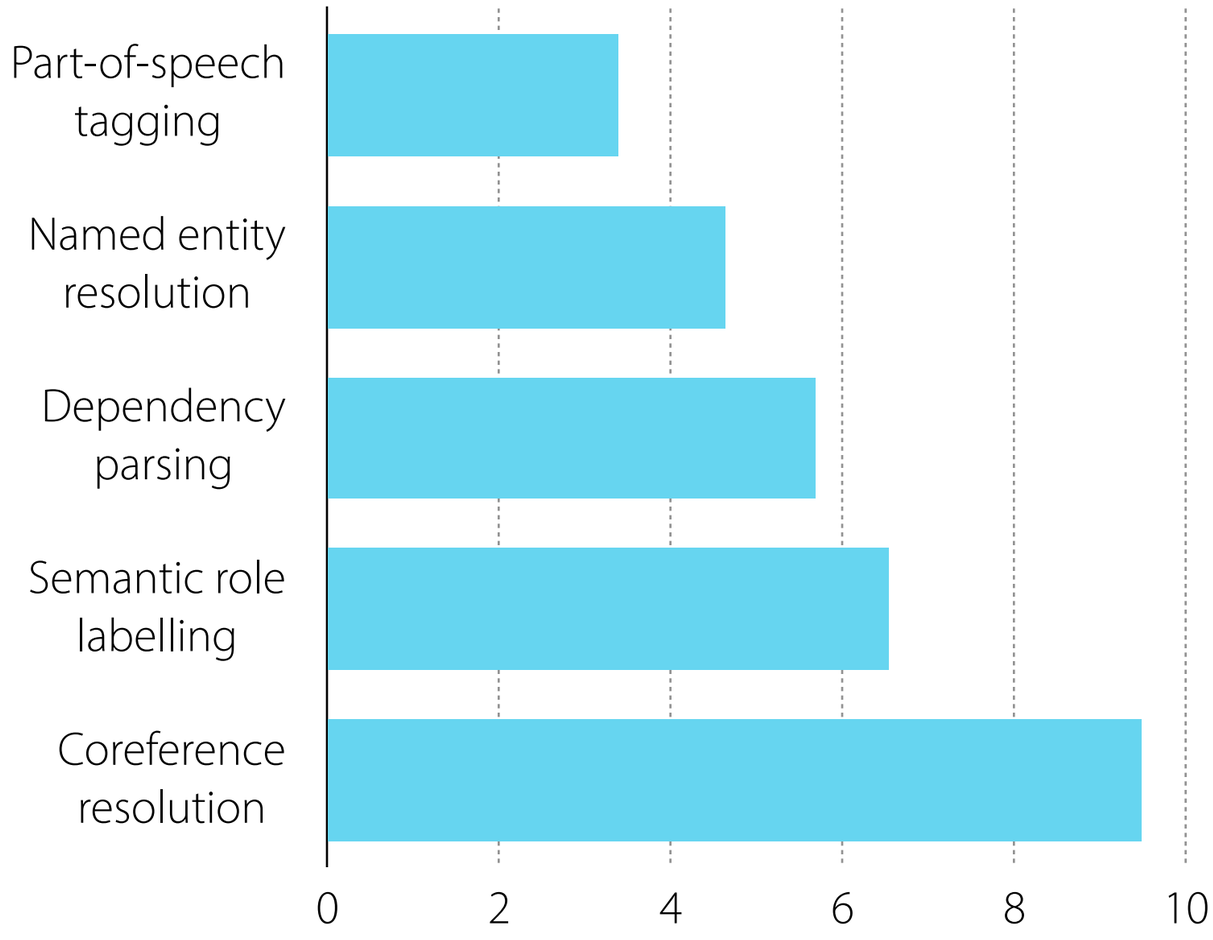
```
dbr:Karen_Sparck_Jones dbo:birthPlace dbr:Huddersfield
```

Lingvistiska representationer

‘Natural language processing from scratch’



Vaswani et al. (2017)



Tenney et al. (2019)

AI för naturligt språk

Sökning och inlärning

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Sökning och inlärning

översättning till engelska

predicerat utvärde

invärde

mening på svenska

$$\hat{y} = \underset{y}{\operatorname{argmax}}\ \operatorname{score}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$$

möjligt utvärde

mening på engelska

modell-parametrar

sannolikheter, vikter

Eisenstein (2019), § 1.2.2

Sökning och inlärning

Eisenstein (2019), § 1.2.2

- **Sökkomponenten**

  Sökkomponenten ansvarar för att hitta det utvärde $\boldsymbol{y}$ som har högst poäng relativt till invärdet $\boldsymbol{x}$ och modellparametrarna $\boldsymbol{\theta}$.

  kräver effektiva algoritmer

- **Inlärningskomponenten**

  Inlärningskomponenten ansvarar för att hitta de modellparametrar $\boldsymbol{\theta}$ som maximerar systemets kvalité.

  ofta någon form av övervakad maskininlärning

Sökning och inlärning

Jag	bad	om	en	kort	bit
PRON	VERB	ADP	DET	ADJ	NOUN
NOUN	NOUN	SCONJ	NUM	ADV	VERB
		ADV	PRON	NOUN	
			NOUN		

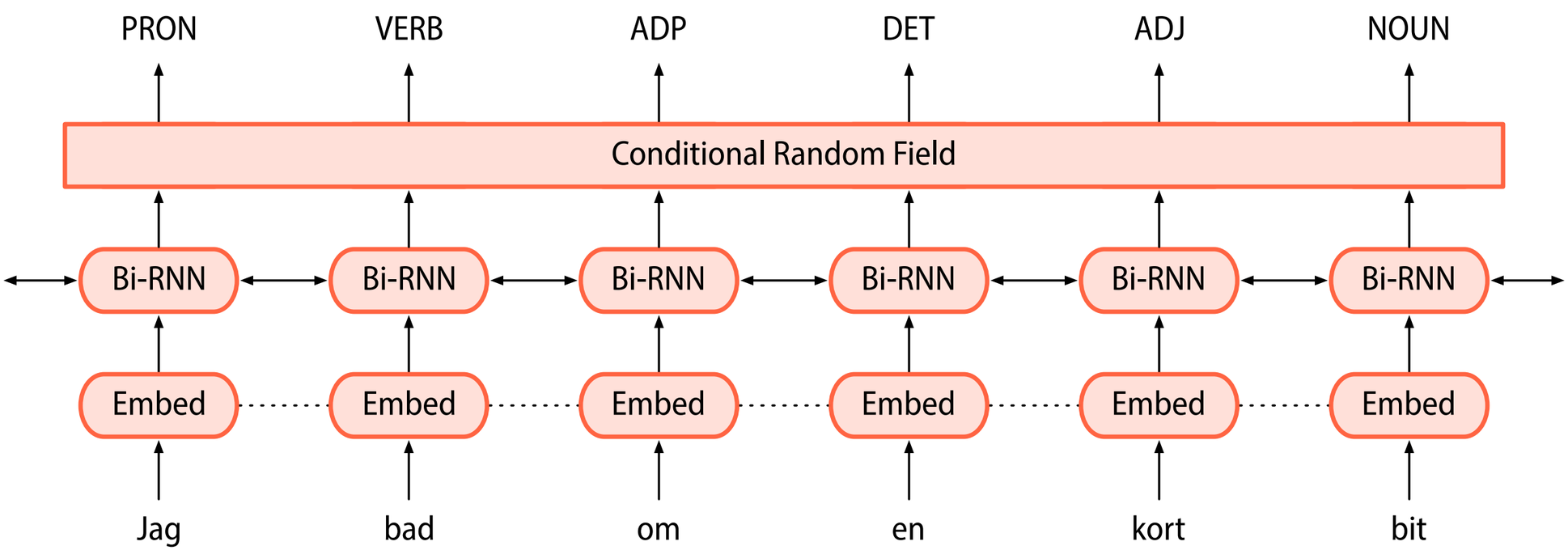
Exempel från Joakim Nivre

Sökning och inlärning

Jag	bad	om	en	kort	bit
PRON	VERB	ADP	DET	ADJ	NOUN
95,83%	50,00%	63,16%	92,52%	66,67%	100,00%
NOUN	NOUN	SCONJ	NUM	ADV	VERB
4,17%	50,00%	35,93%	4,64%	33,33%	0,00%
		ADV	PRON	NOUN	
		0,92%	2,85%	0,00%	
			NOUN		
			0,00%		

Data: UD Swedish Talbanken (training data)

Sökning och inlärning



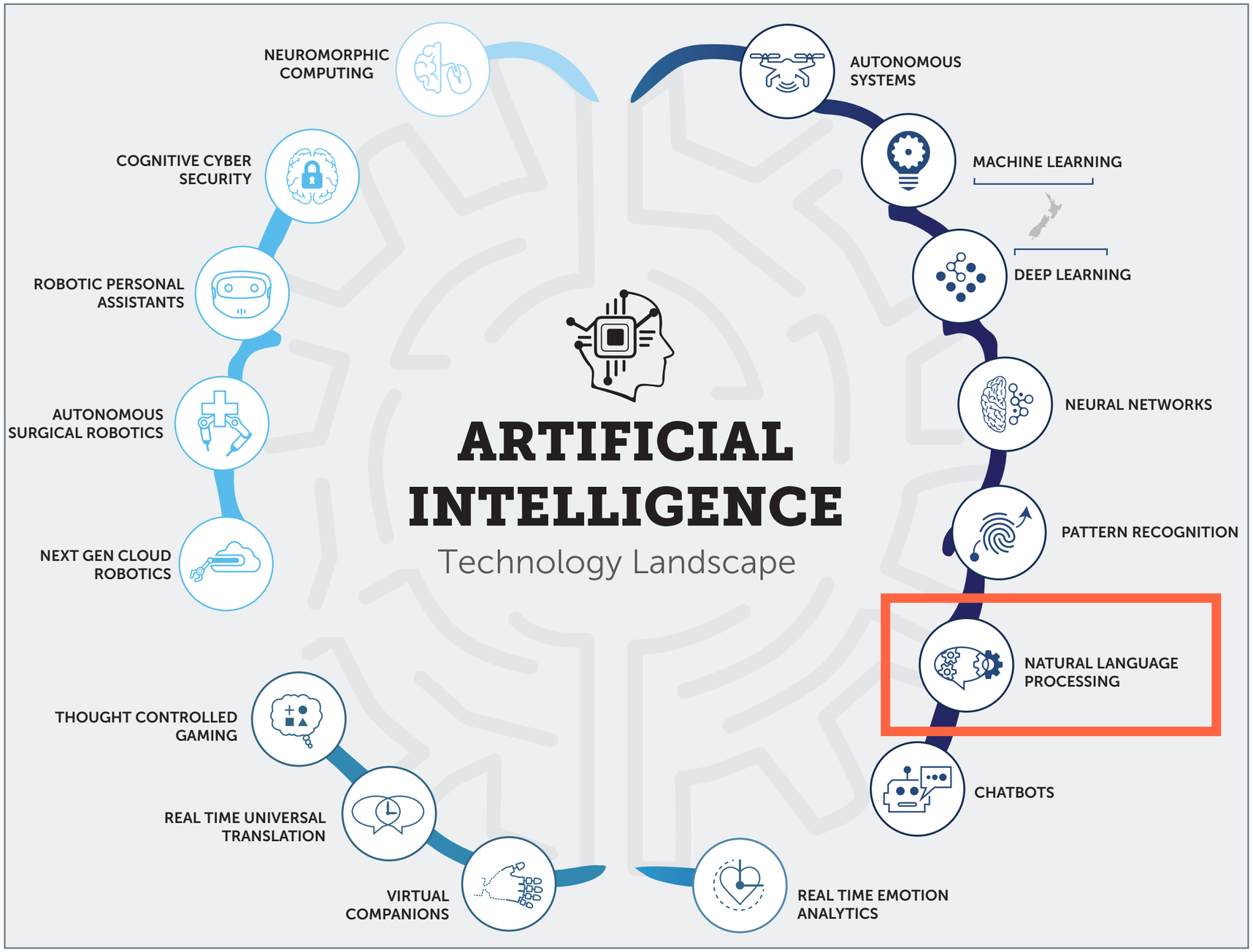
Vikterna i neuronnätet sätts med hjälp av en inlärningsalgoritm.

AI för naturligt språk

NLP i ett bredare perspektiv

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET



Callaghan Innovation: AI Demystified (2017)

Kommersiellt intresse

Bloomberg

Google

amazon | science

Megagon Labs

Baidu 百度

DeepMind

IBM

FACEBOOK AI

Kommersiellt intresse

Andra AP-fonden • Bonnier • Consid • Doctrin

Ericsson • Etteplan • Findwise

Fodina Language Technology • Gavagai • IamIP

iMatrics • Modular Finance • Opera Software

Redeye • Saab • Schibsted • Sectra • Spotify

Storytel • Visma

Etiska frågeställningar

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

▲ The AI wrote a new passage of fiction set in China after being fed the opening line of Nineteen Eighty-Four by George Orwell (pictured). Photograph: Mondadori/Getty Images

The creators of a revolutionary AI system that can write news stories and works of fiction - dubbed “deepfakes for text” - have taken the unusual step of not releasing their research publicly, for fear of potential misuse.

[The Guardian (2019-02-14)]

Miljö och hållbarhet

Consumption	**CO_2e (lbs)**
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Strubell et al. (2019)

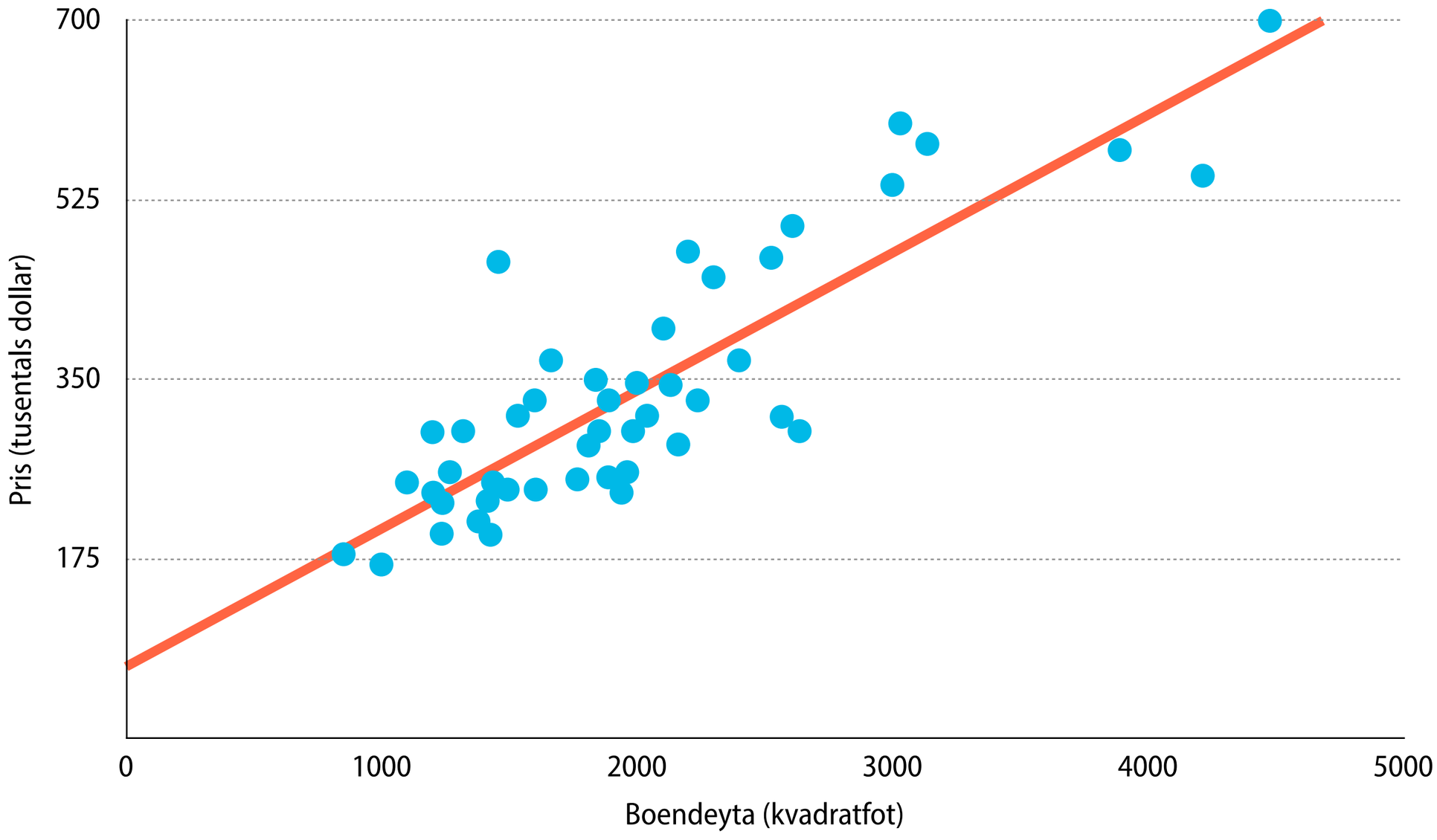
AI för naturligt språk

Linjär regression

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Linjär regression



Linjär regression

- **Modell**

  Sambandet mellan prediktorn (boendeyta) och den predicerade variabeln (pris) kan beskrivas som en linjär funktion.

  räta linjens ekvation

- **Optimering**

  Hitta den mest anpassade linjära funktionen – den funktion som minimerar modellens fel, relativt till data.

  felfunktion: kvadratiskt medelfel

Linjär modell – räta linjens ekvation

prediktor

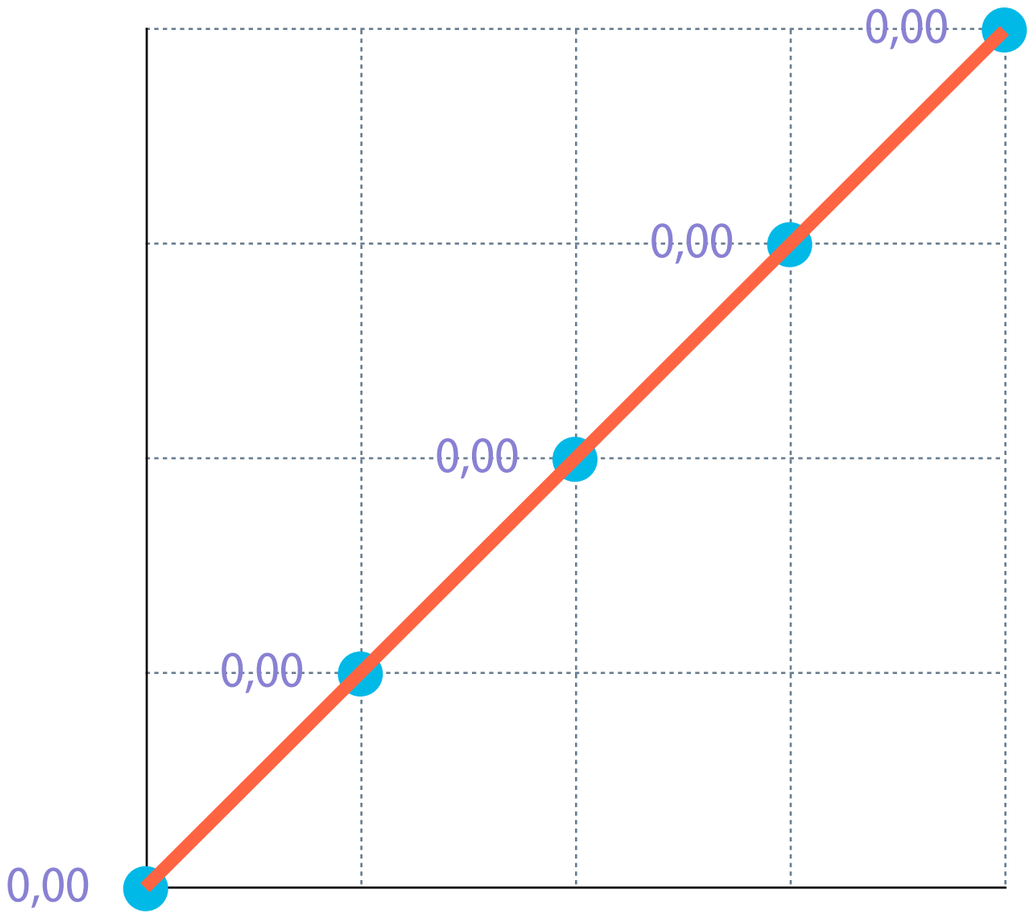
predicerad variabel — $y = xw + b$ — baskonstant

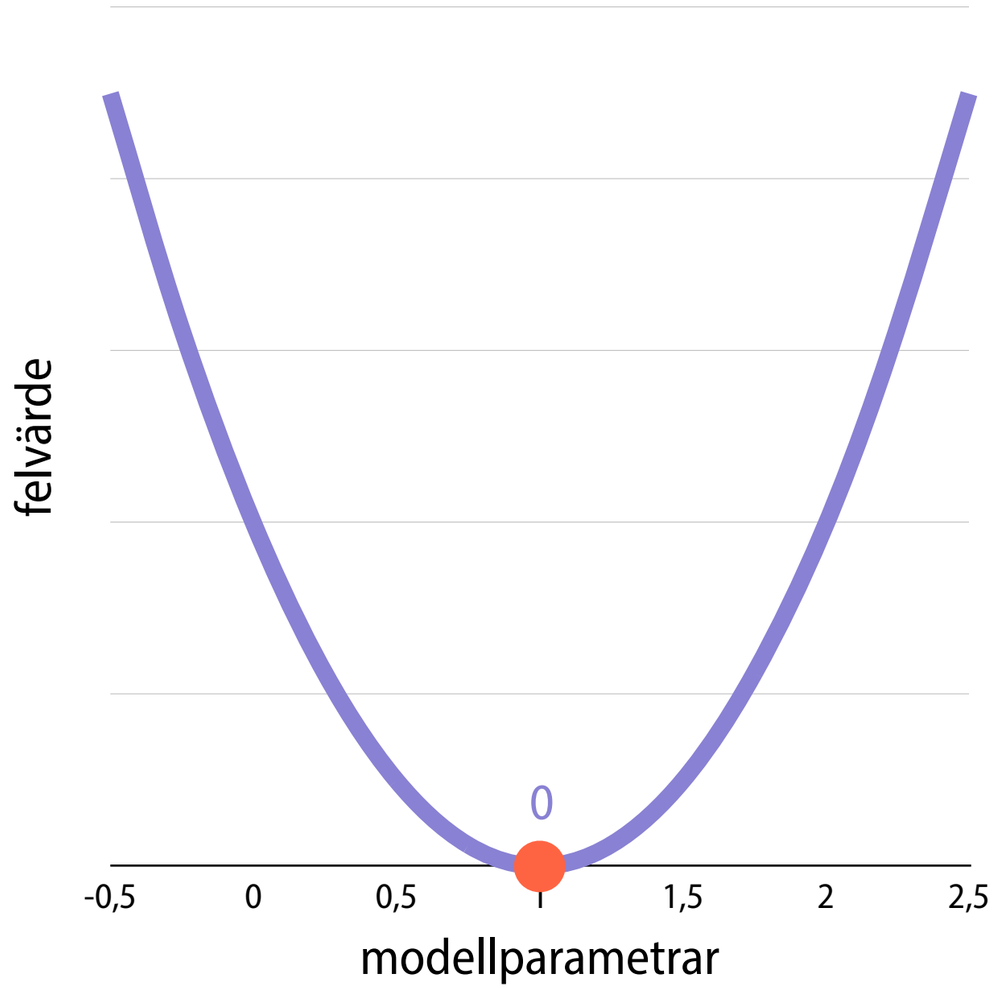
eng. *bias*

lutningskoefficient

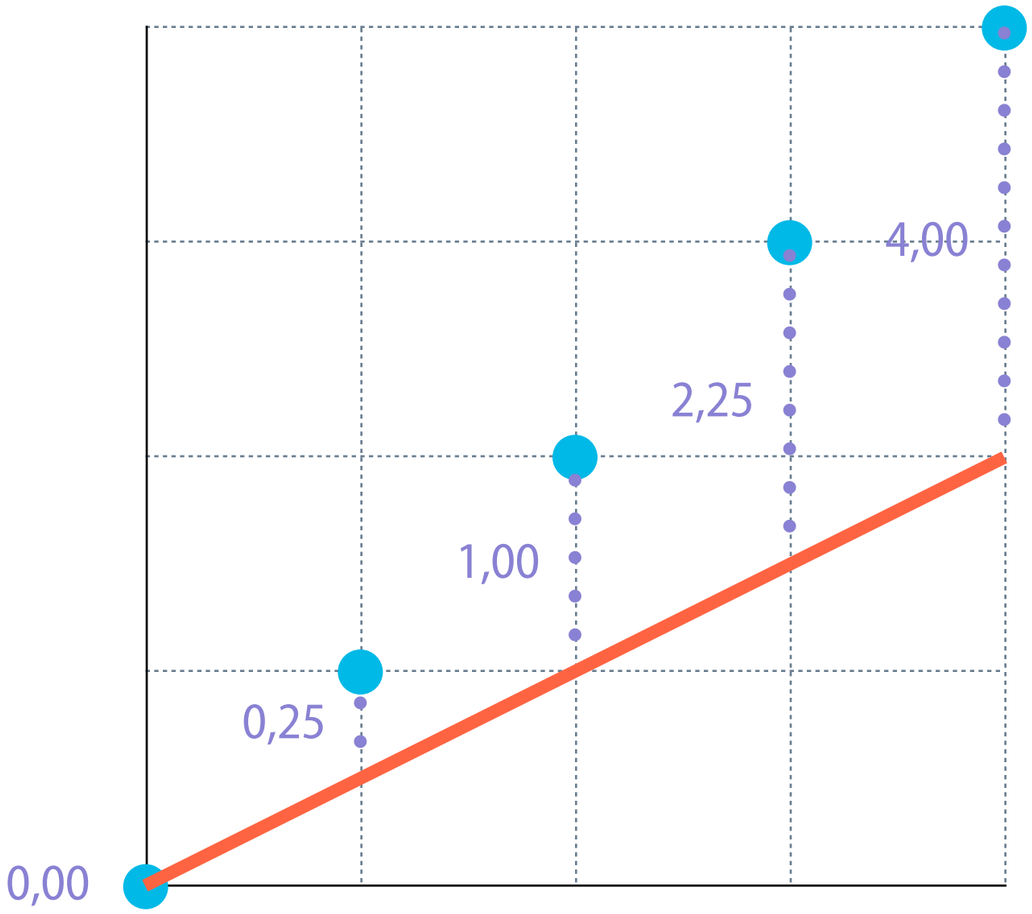
vikt – eng. *weight*

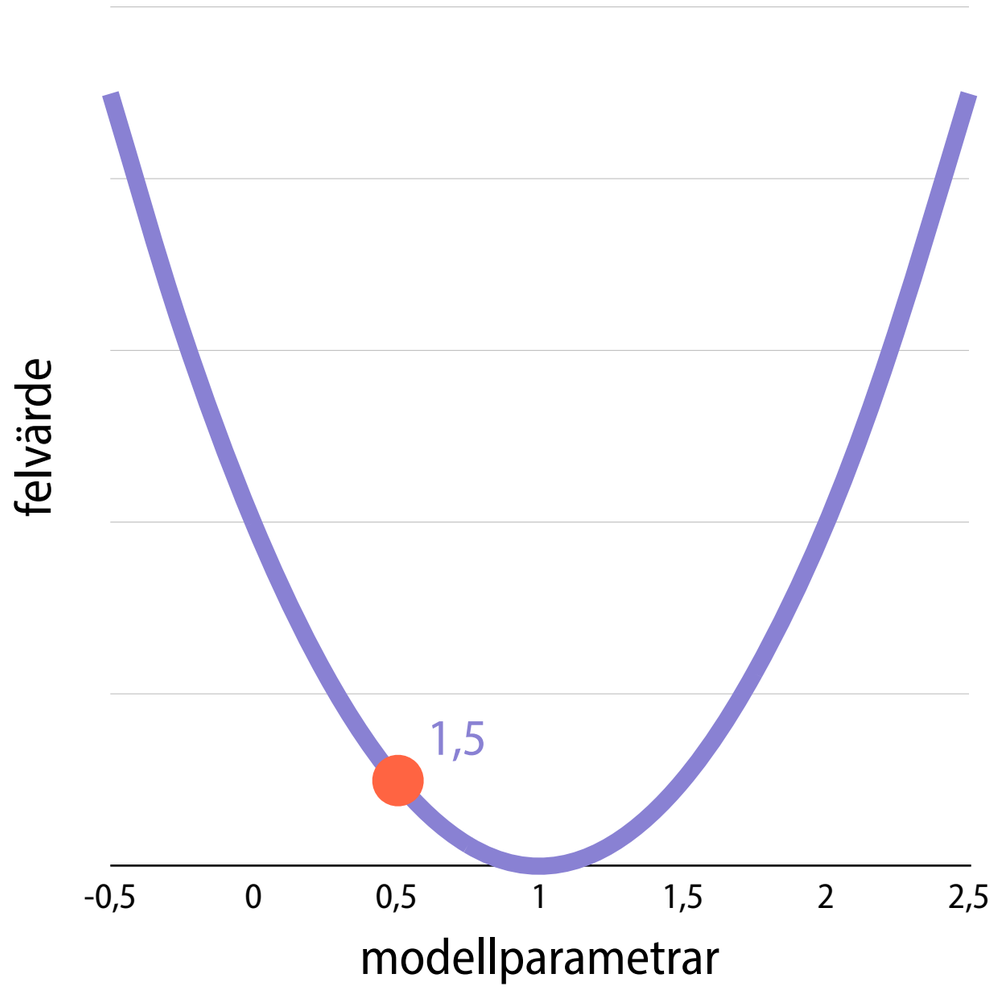
Felfunktion – kvadratiskt medelfel



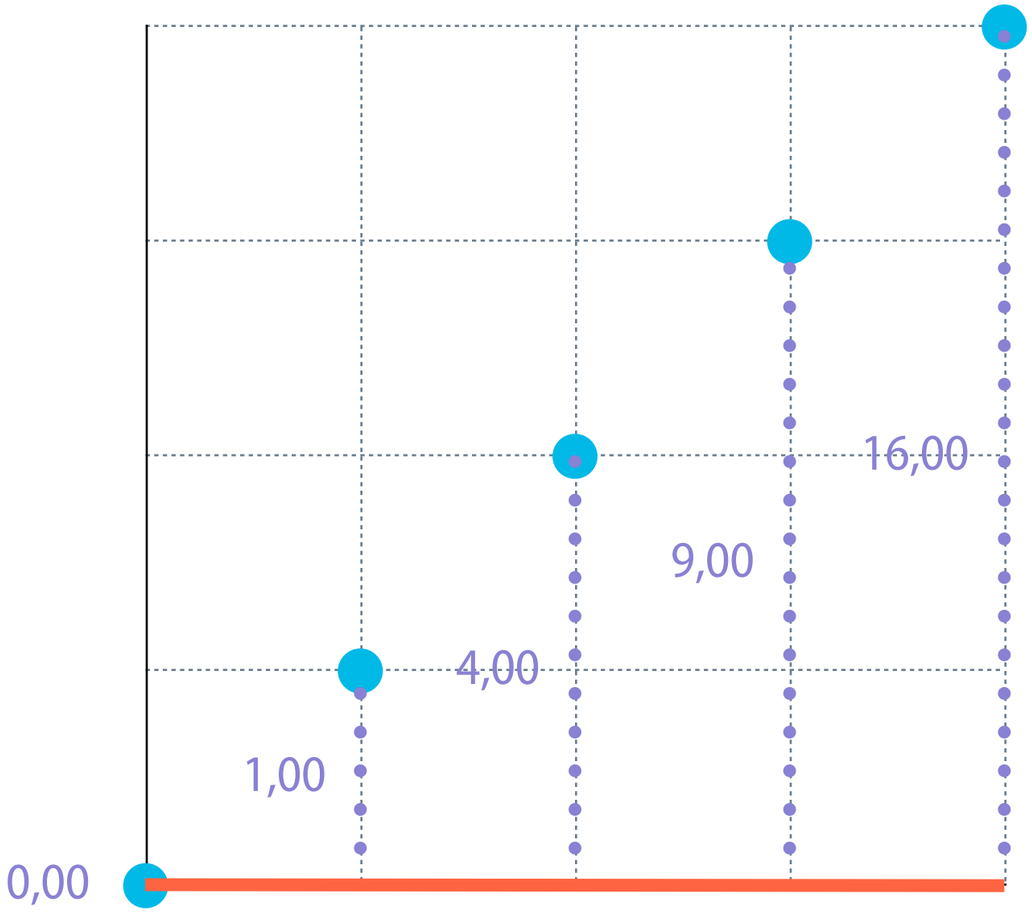


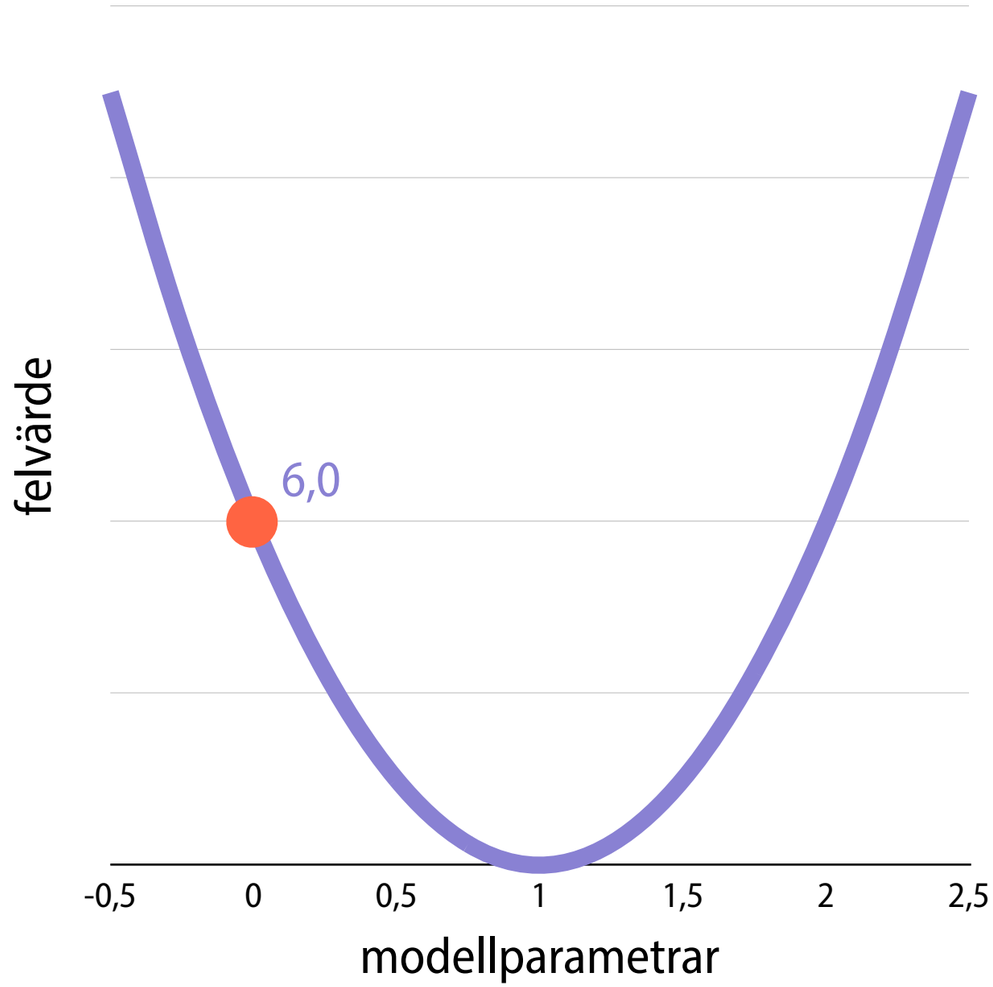
Felfunktion – kvadratiskt medelfel





Felfunktion – kvadratiskt medelfel





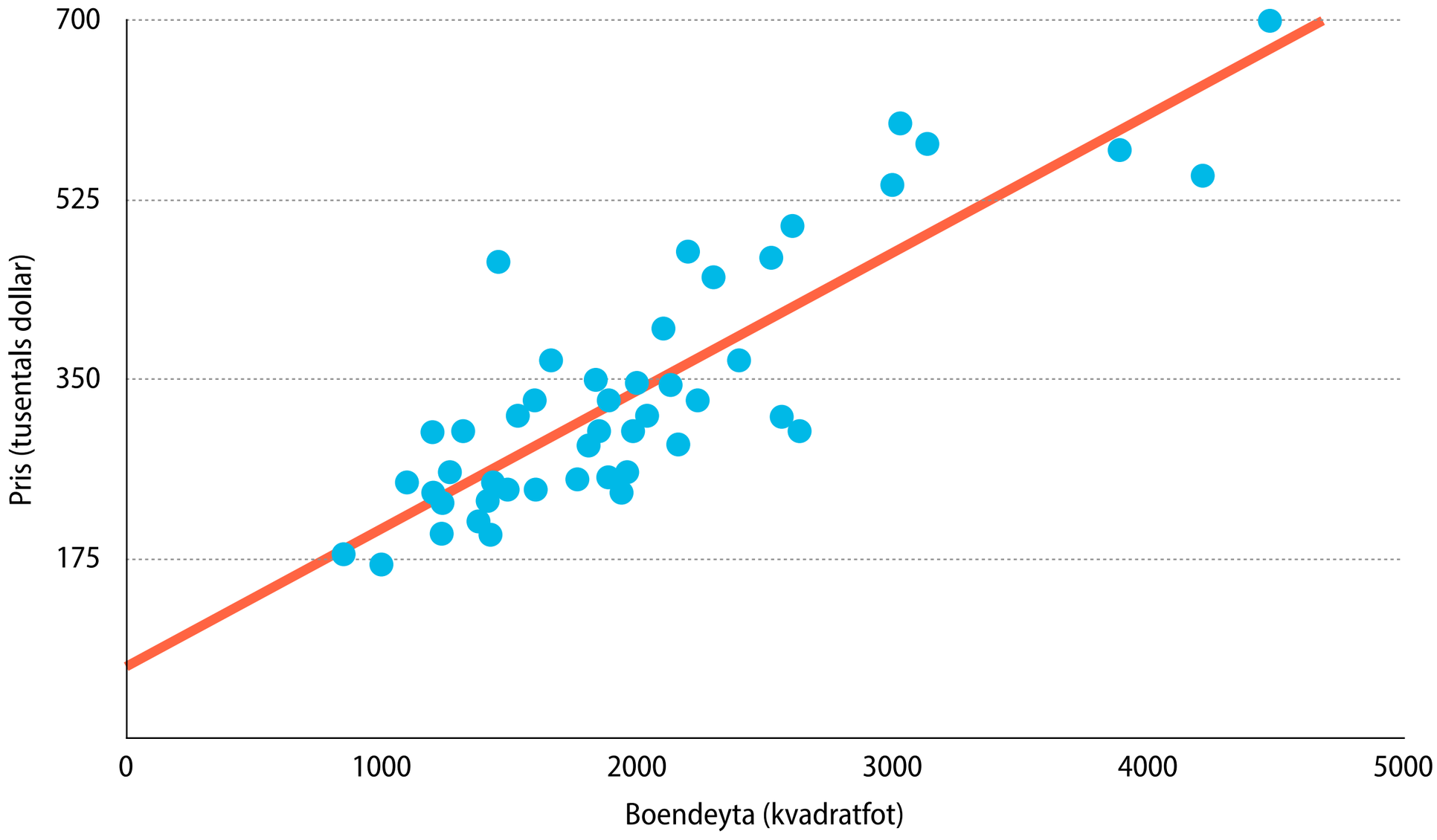
AI för naturligt språk

Generaliserad linjär regression

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Linjär regression



Linjär regression

Boendeyta (x)	Pris (y)
2 104	399 900
1 600	329 900
2 400	369 000
1 416	232 000
…	…

Linjär regression med flera prediktorer

	Boendeyta (x_1)	Antal sovrum (x_2)	Pris (y)
$\boldsymbol{x}$	2 104	3	399 900
	1 600	3	329 900
	2 400	3	369 000
	1 416	2	232 000
	…		…

$$\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} 2104 & 3 \end{bmatrix}$$

Linjär regression med flera prediktorer

- **Modell**

  Sambandet mellan prediktorerna (boendeyta, antal sovrum) och den predicerade variabeln (pris) är linjärt.

  allmän linjär funktion

- **Optimering**

  Hitta den mest anpassade linjära funktionen – den funktion som minimerar modellens fel, relativt till data.

  felfunktion: genomsnittligt kvadratfel

Linjär regression med flera prediktorer

prediktorer

predicerad variabel

$$y = \boldsymbol{x}\boldsymbol{w} + b$$

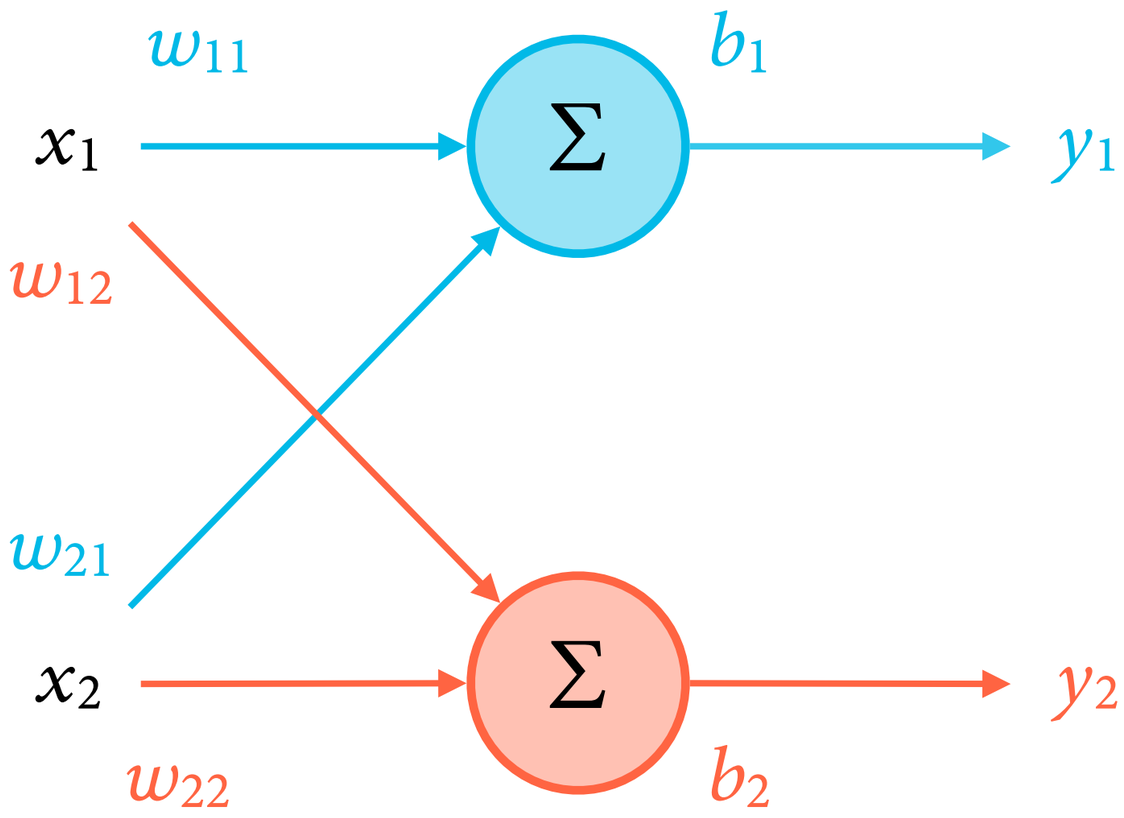
baskonstant

eng. *bias*

lutningskoefficienter

vikter – eng. *weight*

Den generaliserade linjära modellen



$$y_1 = x_1 w_{11} + x_2 w_{21} + b_1 \qquad y_2 = x_1 w_{12} + x_2 w_{22} + b_2$$

Den generaliserade linjära modellen

prediktorer

predicerade variabler — $\boldsymbol{y} = \boldsymbol{x}\boldsymbol{W} + \boldsymbol{b}$ — baskonstanter

eng. *biases*

lutningskoefficienter

vikter – eng. *weights*

Räkneexempel

$$\boldsymbol{x} = \begin{bmatrix} 1 & 4 \end{bmatrix} \qquad \boldsymbol{W} = \begin{bmatrix} 2 & 5 & 2 \\ 7 & 3 & 5 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

$$\boldsymbol{xW} = \begin{bmatrix} 1 \cdot 2 + 4 \cdot 7 & 1 \cdot 5 + 4 \cdot 3 & 1 \cdot 2 + 4 \cdot 5 \end{bmatrix} = \begin{bmatrix} 30 & 17 & 22 \end{bmatrix}$$

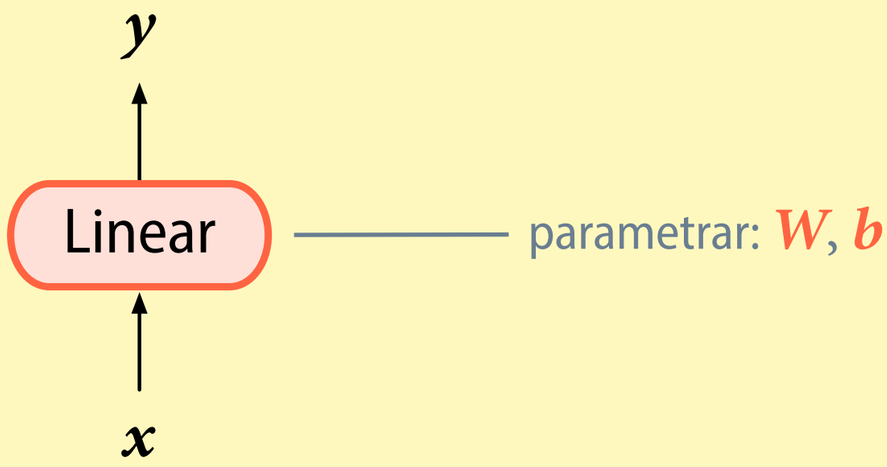
$$\boldsymbol{xW} + \boldsymbol{b} = \begin{bmatrix} 30 & 18 & 22 \end{bmatrix}$$

Linjära neuronnät

matematisk modell

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{W} + \boldsymbol{b}$$

grafisk notation



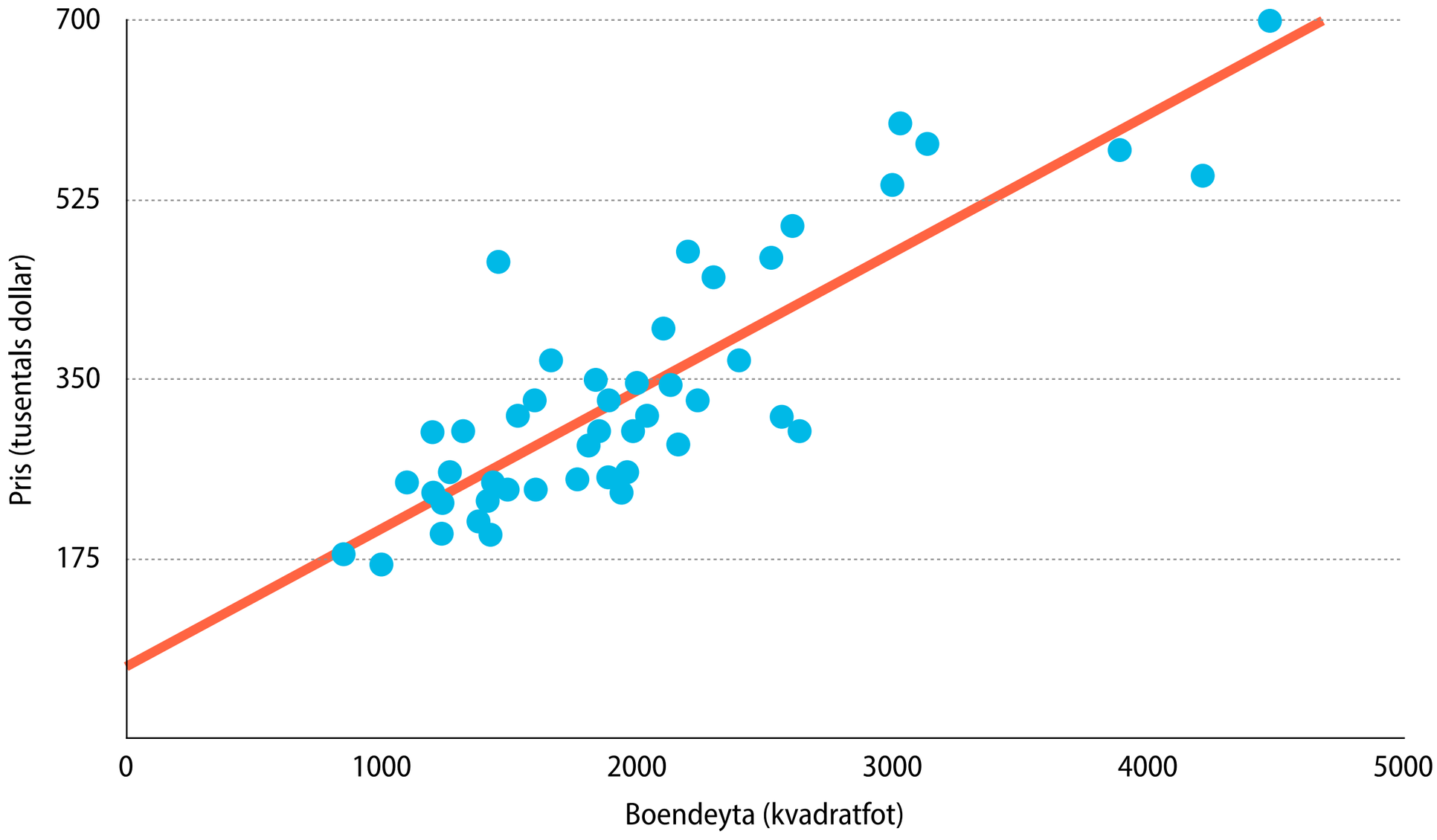
AI för naturligt språk

Gradientsökning

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Linjär regression



Hur hitta den bästa linjära modellen för datat?

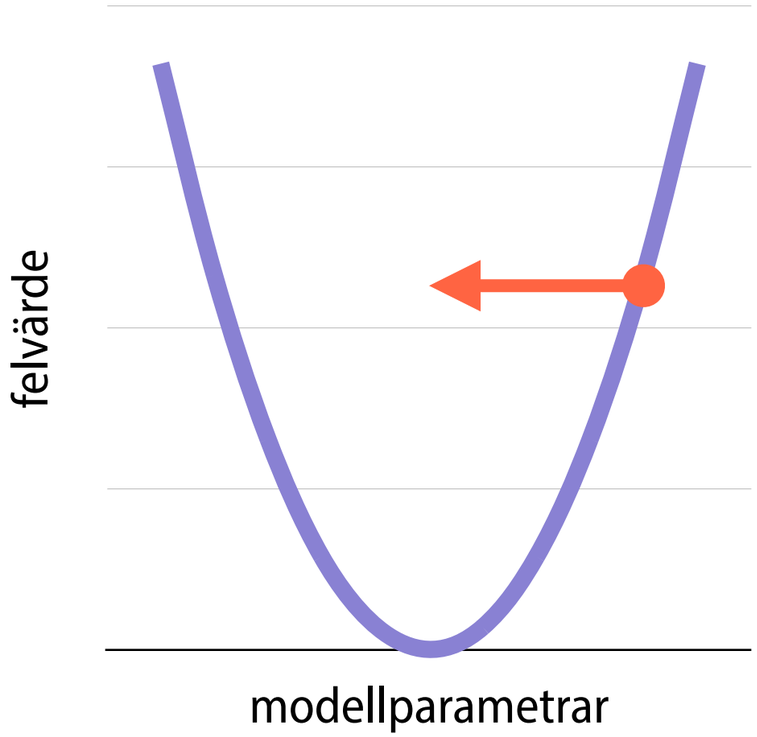
- För att få en ungefärlig lösning kan vi för hand rita in en rät linje och sedan beräkna dess lutning och baskonstant.

- För att få en exakt lösning kan vi använda minsta kvadrat-metoden, som finns bl.a. på många grafritande miniräknare.

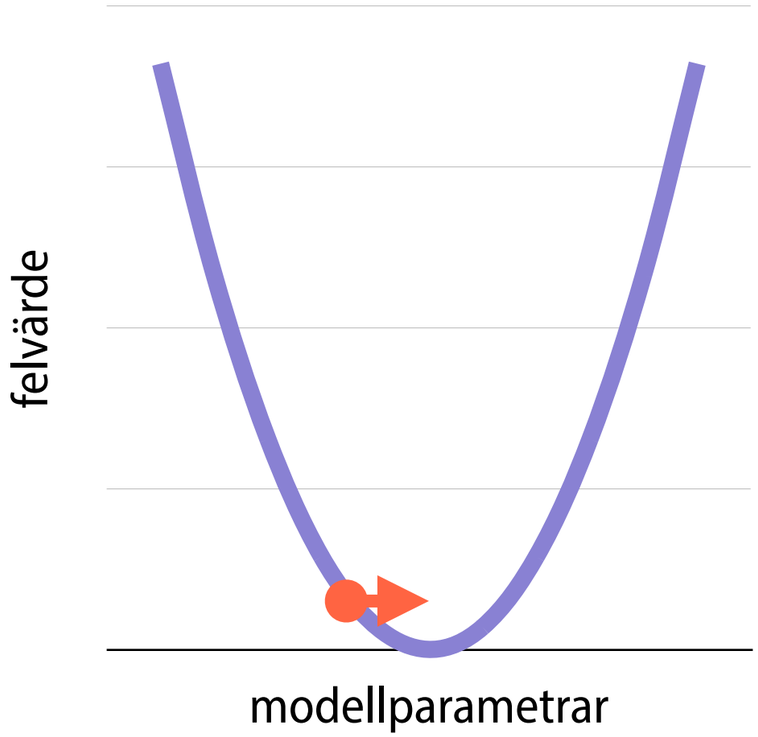
- Ett annat sätt att lösa uppgiften är att använda sig av en approximativ, numerisk metod: **gradientsökning**.

  gradient = generalisering av derivata till fler än en oberoende variabel

Gradientsökning – intuition

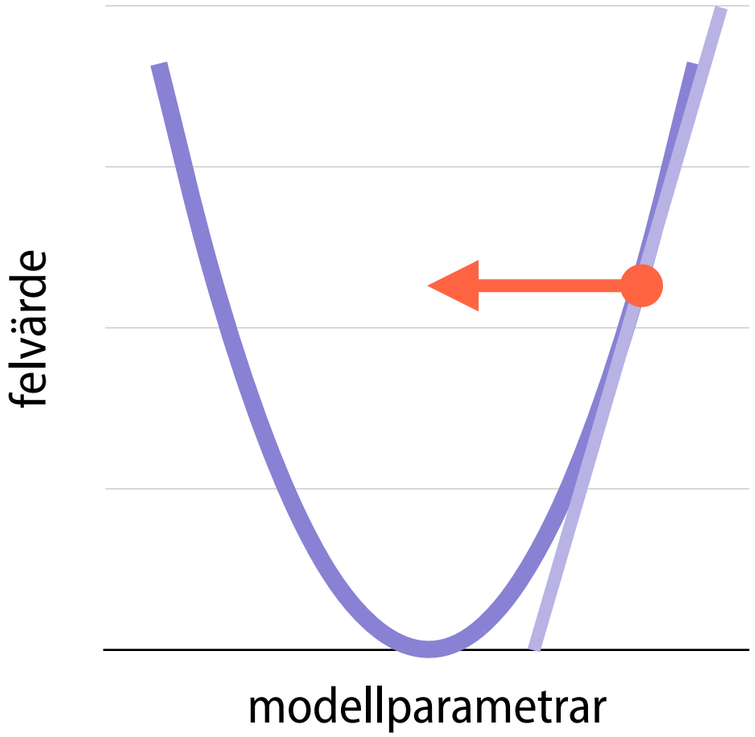


subtrahera
ett relativt stort värde

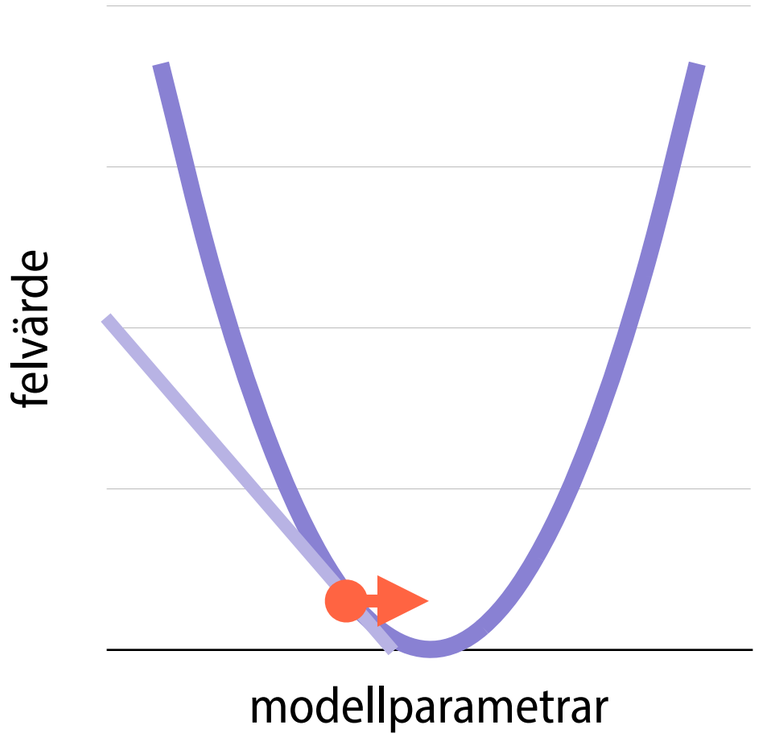


addera
ett relativt litet värde

Gradientsökning – intuition



subtrahera
felfunktionens gradient



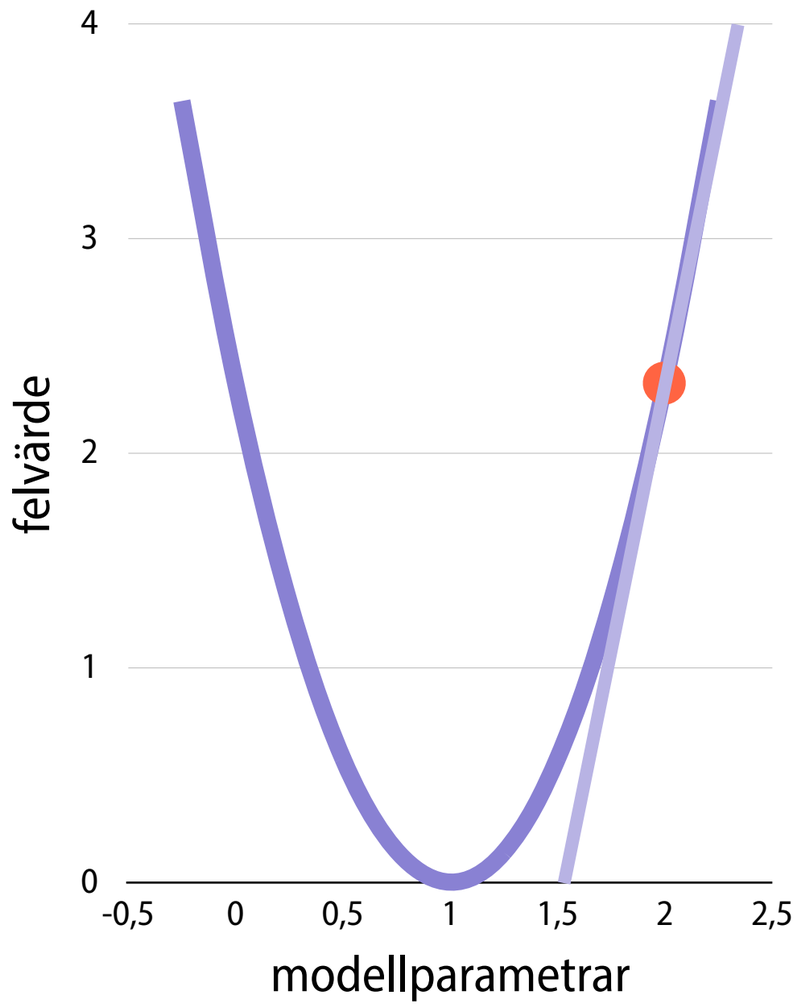
subtrahera
felfunktionens gradient

Gradientsökning

1. Börja med framslumpade modellparametrar
2. Beräkna felfunktionens gradient för den aktuella uppsättningen modellparametrar, relativt till datapunkterna

   förutsätter att felfunktionen är deriverbar
3. Uppdatera modellparametrarna genom att subtrahera gradienten
4. Upprepa steg 2–3 tills felvärdet är tillräckligt lågt

Gradientsökning

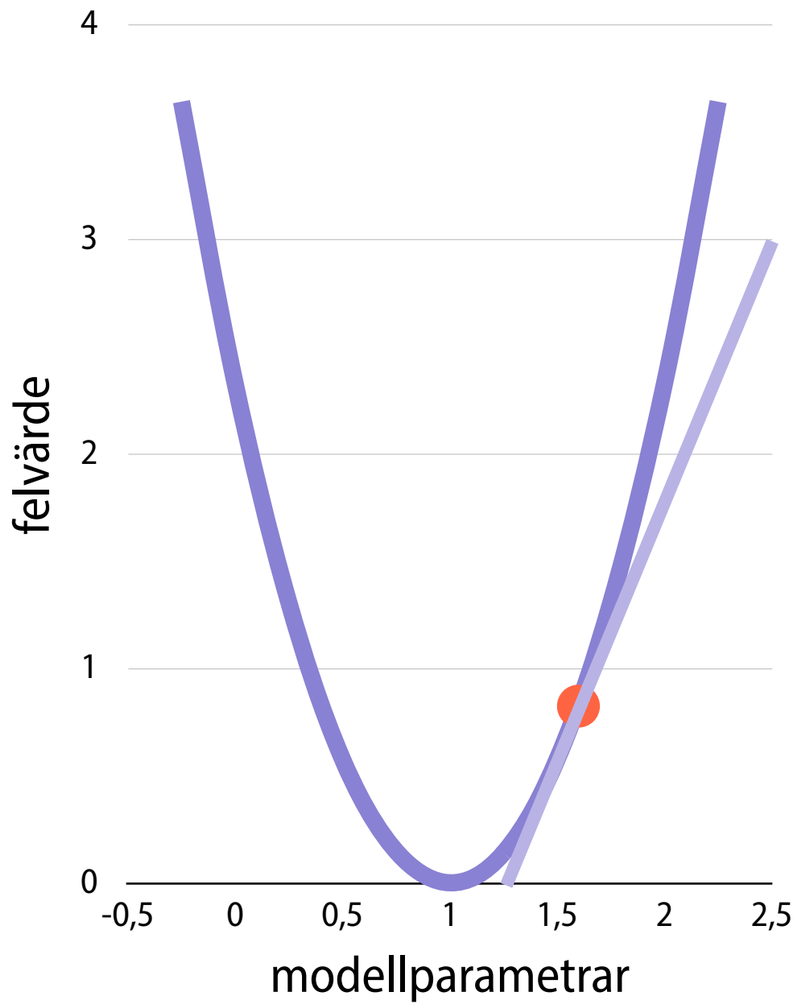


modell	felvärde	gradient
2,000	2,33	4,67

steglängdsfaktor = 0,1

subtrahera 0,467

Gradientsökning

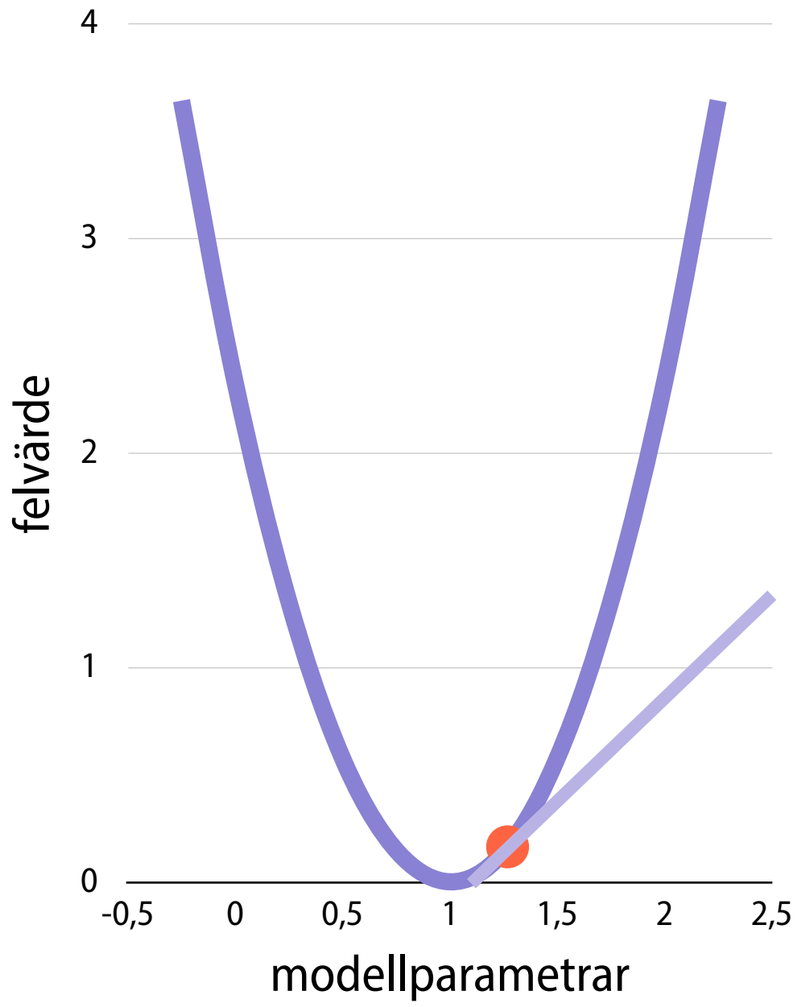


modell	felvärde	gradient
2,000	2,33	4,67
1,533	0,66	2,49

steglängdsfaktor = 0,1

subtrahera 0,249

Gradientsökning

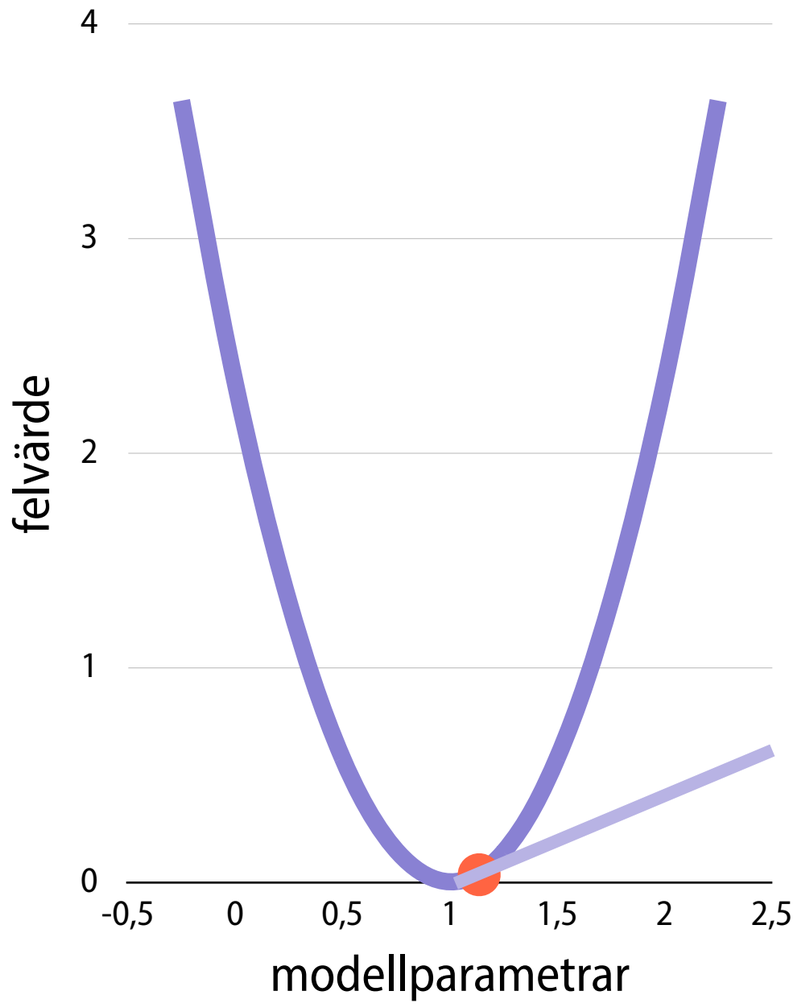


modell	felvärde	gradient
2,000	2,33	4,67
1,533	0,66	2,49
1,284	0,19	1,33

steglängdsfaktor = 0,1

subtrahera 0,133

Gradientsökning



modell	felvärde	gradient
2,000	2,33	4,67
1,533	0,66	2,49
1,284	0,19	1,33
1,151	0,05	0,71

steglängdsfaktor = 0,1

subtrahera 0,071

Den osannolika effektiviteten av gradientsökning

- Om felfunktionen är konvex kommer gradientsökningen hitta en optimal uppsättning parametrar.

  Kvadratiskt medelfel är en sådan konvex felfunktion.

- De flesta felfunktioner är *inte* konvexa, och det finns ingen garanti att gradientsökning kommer hitta en optimal lösning.

- I praktiken är det dock så att gradientsökning fungerar förvånansvärt bra när den används för att träna neuronnät.

  många tekniska tricks