

Statistik och Dataanalys I

Statistik - Vetenskapen om Data

Mattias Villani



Statistiska institutionen
Stockholms universitet



mattiasvillani.com



@matvil

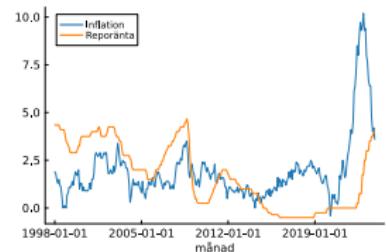


[mattiasvillani](https://github.com/mattiasvillani)

Statistik inom ekonomi

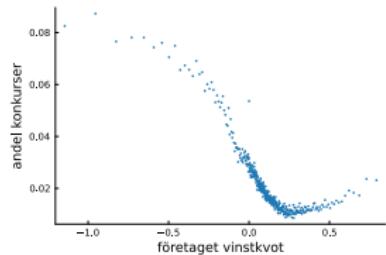
■ Riksbankens räntesättning

- ▶ Riksbankens mål: 2% inflation per år.
- ▶ Hur påverkar reporäntan inflationen?
- ▶ Prognoser över framtida inflation.
- ▶ Beslut under osäkerhet



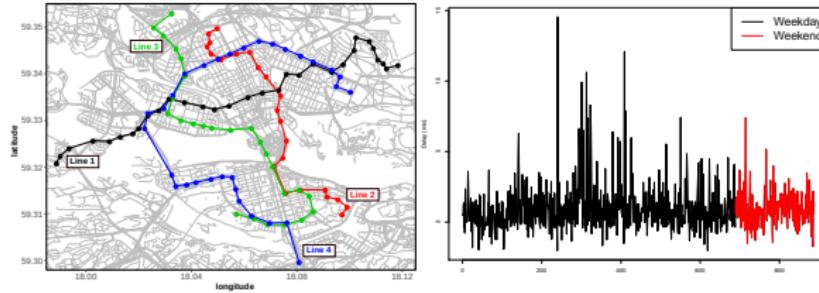
■ Företagskonkurser

- ▶ Data på alla svenska aktiebolag
 - målvariabel: konkurs/ej konkurs
 - orsakssvariabler: vinst, tillgångar, anmärkningar, ålder, makro.
- ▶ Vilka variabler förutsäger en konkurs?
- ▶ Prediktion av ekonomins konkursrisk.



Förseningar i lokaltrafiken

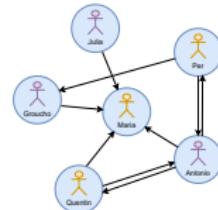
- Mål1: **förutsäga förseningar** för stadsbussar.
- Mål2: **säkerheten** i prediktionen: **5 min, 5 min, 5 min**
- Data: alla förseningar för alla busslinjer i Sthlm under 1 år.
- Mål: förutsäga förseningen för 12.15-bussen till Tegnérsgatan.
- Förklarande variabler:
 - ▶ försening för 12.15-bussen vid hållplatser innan Tegnérsgatan.
 - ▶ förseningar för tidigare bussar vid hållplats Tegnérsgatan.
 - ▶ tid på dagen
 - ▶ rusningstid?



Nätverksdata

- Socialt **nätverk**: individer och deras **relationer**.
- Data: noder** (personer) och **länkar** (relationer).

	Julia	Per	Antonio	Quentin	Groucho	Maria
Julia	0	0	0	0	0	1
Per	0	0	1	0	1	0
Antonio	0	1	0	1	0	1
Quentin	0	0	1	0	0	1
Groucho	0	0	0	0	0	1
Maria	0	0	0	0	0	0

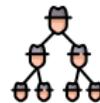


- Sociala nätverk (Twitter, Facebook etc)



- Kriminella nätverk

- Noder: personer.
- Länkar: har gjort brott tillsammans?



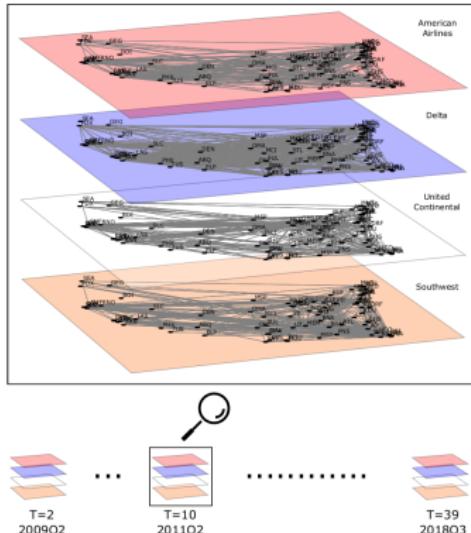
- Kulturella nätverk

- Noder: Skådespelare.
- Länkar: Spelat i samma pjäs eller film.



Amerikansk flygplanstrafik

- **Noder:** flygplatser. **Länkar:** flygrutter.
- **Dynamiska nätverk** vars länkar förändras över tid.
- **Multipla lager:** en graf för varje flygbolag.
- Data: 80 flygplatser för 4 flygbolag över 10 års tid.
- Delmål: **förutsäga nätverkets utveckling.**

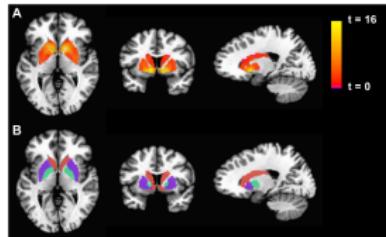


Var i hjärnan skapas vårt språk?

- Person i MR scanner **pratar-knyter handen-pratar** osv.

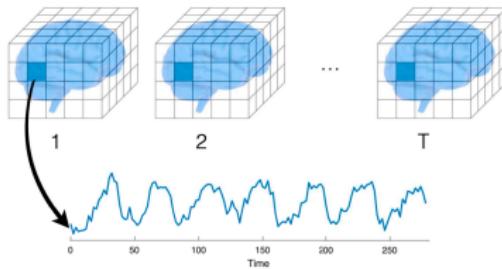


Lars Kruse, AU Kommunikation, CC license



[Source](#), CC license

- Mäter** mängden syresatt blod på tusentals ställen i hjärnan.



- Vilka hjärnregioner aktiveras** när man pratar? Språkcentra.

Optimala kunskapsprov och intelligens

- Mäta elevers kunskaper: Nationella prov, PISA etc.
- **Statistisk modell:**

Provsvar (data) \implies elevens sanna kunskapsnivå (inferens)

- **Designa optimala prov** för att mäta kunskapsnivå.
- **Adaptiva prov**: vid datorbaserade prov kan man välja optimal fråga för varje student baserat på tidigare svar under provet.
- Pågående forskningsprojekt vid statistiska institutionen.
- **Psykologi**: vad är **intelligens**, och hur mäter man det?
En eller fler-dimensionellt? **Statistisk faktorenanalys**.



Klimatförändringar - blekning av korallrev

- Korallreven - **havets regnskogar.**
- Hem till 25% av alla arter.
- Minskning med 14% under perioden 2009-2018.
- **Blekning** bl a pga ökad havstemperatur.
- **Regression:**
 - ▶ målvariabel: andel blekt korall
 - ▶ förklarande variabler: extrema havstemperaturer
 - ▶ kontrollvariabler: havsdjup, solstrålning etc.



Källa : Wikipedia (vänster: J. Roff. Höger: Holobionics. Licenser: CC BY-SA 3.0).

Statistiker får jobb som data scientists



Andreea Taylor · 1st
Staff Machine Learning Engineer at Voi



Sebastian Ankargren · 1st
Data Scientist at Spotify



Qurat Anwar · 1st
Artificial Intelligence|DiversifAI|AI Product management



Emelie Wahlström · 1st
Program/Project Manager & Data Scientist at Combient Mix · PhD, Data Scientist - Ericsson



Parfait Munezero · 1st
PhD, Data Scientist - Ericsson



Leif Jonsson · 1st
Ph.D., Expert AI & Machine Learning - Ericsson

Artificiell intelligens och maskininlärning



- Statistik är grunden för modern AI.



... the reader should have some knowledge of basic statistics, including variance, correlation, simple linear regression, and basic hypothesis testing (e.g. p-values and test statistics).

- **Deep Learning Book:** Kapitel 3:

Sannolikheter, slumpvariabler, sannolikhetsfördelningar, väntevärde, varians, kovarians, korrelation, regression, Bayes sats, Normalfördelning, osv.

Bilder, text och ljud är data

- Bilder är data. En pixels färg = numerisk tal.

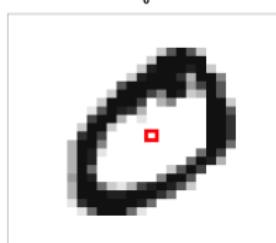
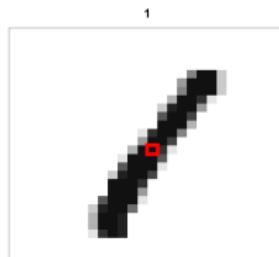


- Mål: få en maskin att känna igen handskrivna siffor.
- Data: 60000 handskrivna siffror mellan 0-9.
- Varje bild har 28×28 pixlar med värde mellan 0 och 255:

0 = svart

128 = mellangrå

255 =

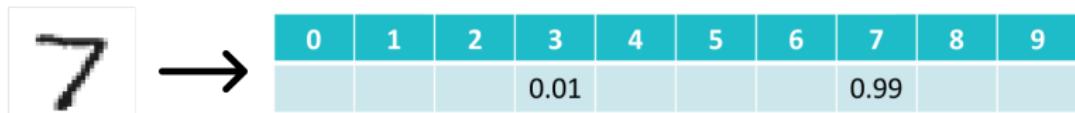


Prediktion av handskrivna siffror

- AI-maskin:



- Men **hur säker** är AI-maskinen?
- **Statistisk prognosmodell** som ger *sannolikhetsfördelningar*:



- Djupa neurala nätverk (**deep learning**) bygger på statistik.

RGB data

- Red-Green-Blue tripplar.

Röd pixel (217,7,0)



Gul pixel (219,222,4)

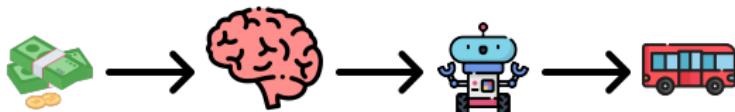


Grå pixel (135,137,144)



Statistik - a love story 😍

- **Data/information** finns numera **överallt**.
Internet, smartphones, sensorer, betalkort, läsplattor
- **Data är det nya guldet**. Facebook, Google etc lever på datainsamling och analys av data.
- Statistiker arbetar inom alla fält. Frihet att byta fält.

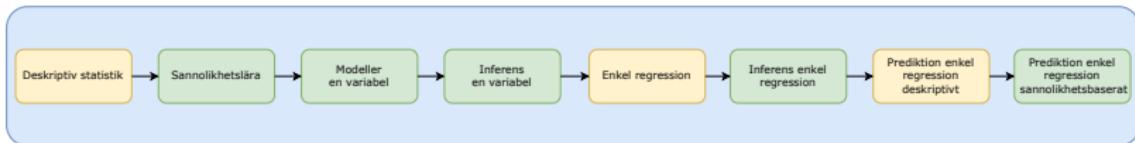


- Annat ämne + mycket **statistik gör dig unik**.
- **Empiriska bevis inom vetenskap** avgörs av statistik.
 - ▶ Är Covid-vaccin effektiva?
 - ▶ Fungerar kognitiv beteendeterapi?
 - ▶ Har inkomstskillnaderna i Sverige ökat?
- Statistik ⇒ informerad medborgare. **Förstå och tolka data**. **Kritiskt ifrågasätta data**. **Samla in** bättre data.

SDA1 - en modern kurs

- Fokus på **dataanalys i R** och **datorbaserat arbetssätt**.
- **Sambandsanalys** 😍 tidigt för motivation.
- Större fokus på **prediktion** (även för att välja modell).
- **Sannolikhetslära senare**, när man insett varför det behövs.
- **Fokus på grundidéer**. Färre varianter av metoder.

"Traditionell" kurs



Statistik och dataanalys I

