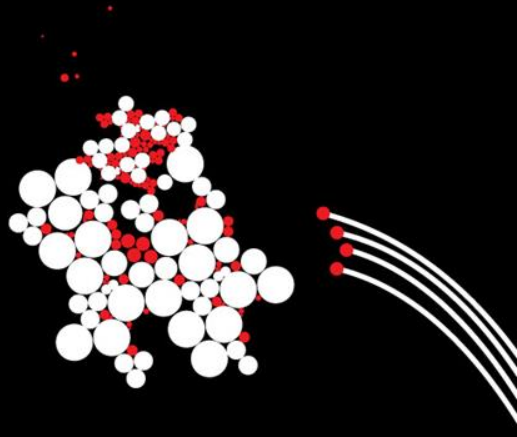
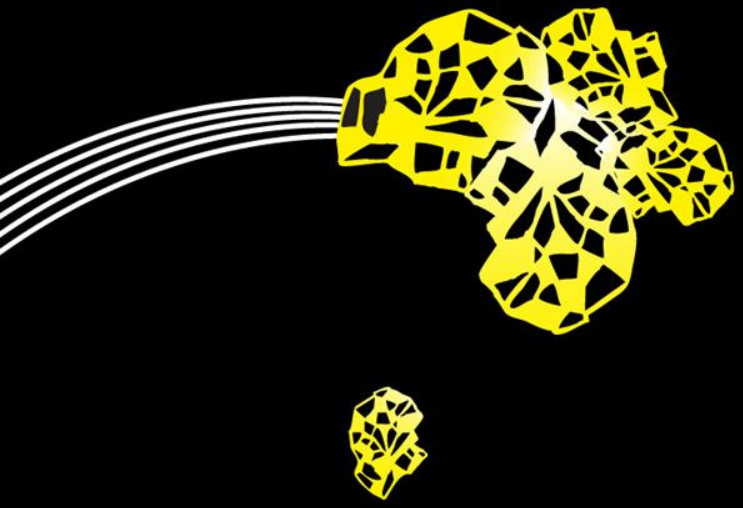
UNIVERSITEIT TWENTE.



BIG DATA EN DE WISKUNDE ACHTER HAAR SUCCES

MAURICE VAN KEULEN



WAT IS BIG DATA?

Wat nou “BIG”?

Sinds 1975 VLDB conferentie: Very Large DataBases

Officiële uitleg:
de 4 V’s

- Volume
- Velocity
- Variety
- Veracity

Veel oude wijn in nieuwe zakken

Mijn uitleg

when magic happens!

“**Big**”
de hoeveelheid data overschrijdt een grens waar intelligent semantisch gedrag uit de data oprijst

VOORBEELD “MAGIE” IN BIG DATA: SCENE COMPLETION

Hays, J., Efros, A. 2007. Scene Completion Using Millions of Photographs. ACM Trans. Graph. 26, 3, Article 4 (July 2007), 7 pages. http://doi.acm.org/10.1145/1239451.1239455.

VOORBEELD “MAGIE” IN BIG DATA: GOOGLE TRANSLATE

90 talen

Geen grammatica

Wat dan wel?

documenten van de Verenigde Naties (6 talen)

Statistische analyse met

1. Tweetalige collectie van meer dan miljoen woorden

én

2. twee enkeltalige collecties van meer dan een miljard woorden

VOORBEELD “MAGIE” IN BIG DATA: IBM WATSON

Watson: Kunstmatig intelligent systeem dat vragen kan beantwoorden die gesteld zijn in natuurlijke taal

Wint Jeopardy quiz show in 2011 tegen voormalig winnaars Brad Rutter en Ken Jennings

WEL BIG DATA, MAAR GEEN MAGIE

- Data analytics
  - Business analytics / business intelligence
  - Data warehousing en OLAP
  - e-Science
- Mining
  - Data mining
  - Text mining

Is onderzoek naar het Higgs-deeltje ‘big data’?

Natuurlijk!

WAT IS BIG DATA?

Waar komt de magie vandaan?

Uit de data?!?

Welke wiskunde kan toveren met data?

Kansrekening!!!

KANSREKENING: DE ESSENTIE

Hoe weet ik of en hoe een dobbelsteen ‘oneerlijk’ is?

Heel vaak dobbelen!

WET VAN DE GROTE GETALLEN

Stelling

- over het resultaat van het heel vaak uitvoeren van hetzelfde experiment
- het gemiddelde convergeert naar de verwachtingswaarde
- hoe vaker, hoe dichterbij

$$\overline{X}_n=(X_1+\ldots+X_n)/n$$

$$\lim_{n\to\infty} \overline{X}_n = \mu \qquad (\mu \text{ is de verwachtingswaarde})$$

TAALMODELLEN: $P(T_1,\ldots,T_N)$

ENGELS: LANGUAGE MODELS

OCR spraak typefouten

Taalmodel:

- “een stuk text” is een meer waarschijnlijke lijst met woorden in het Nederlands dan “ccn stk toksl”
- P([een,stuk,tekst]) > P([ccn,stk,toksl])

Aanpak en P([…]) te bepalen

- Gegeven een **grote** collectie teksten
- Wijs blind naar 3 opeenvolgende woorden (of 3x woord pakken)
- Doe dit **vaak** … heel vaak ➔ N
- Tel hoe vaak je [een,stuk,tekst] hebt aangewezen ➔ w
- P([een,stuk,tekst]) = w/N

big data

tri-gram

simultane kansverdeling

SCIENTIFIC PAPER GENERATOR

HTTP://PDOS.CSAIL.MIT.EDU/SCIGEN/

- $P(T_{n+1}|T_1,\ldots,T_n)$: de kans op woord/token T_{n+1} gegeven voorafgaande woorden/tokens $T_1,\ldots,T_n$
- Benader deze simultane kansverdeling obv collectie wetenschappelijke artikelen
- Gebruik het om automatisch wetenschappelijke teksten te genereren, voeg wat plaatjes en grafieken toe …
- … en we hebben de wetenschap weggeautomatiseerd!

- Gebruikt voor aantonen wantoestanden reviewsysteem
- Gegenereerde teksten lijken heel echt, maar zijn natuurlijk complete onzin

HOE EEN INFORMATICUS NAAR EEN KANSVERDELING KIJKT

Okay, we kunnen $P(T_n|T_1,\ldots,T_{n-1})$ en $P(T_1,\ldots,T_n)$ uitrekenen …

… we kunnen die ook opslaan! als een grote opzoektabel!

- Aantal woorden in Engelse taal
  - Global Language Monitor (2014): 1.025.109,8
  - Merriam Webster's dictionary (1993; incl appendix): ±470.000; Oxford English Dict 2nd ed: similar number
- Naieve berekening van grootte:
  - Filter minst relevante woorden weg tot, zeg, 100.000
  - Grootte = 10^{5n} * 'grootte-van-getal' (±4 bytes)
    n=3 ➔ $4*10^{15}$ = 4 TB (kB=10^3, MB=10^6, GB=10^{12}, TB=10^{15})

Voorbeeld: Microsoft Web N-gram Services
http://weblm.research.microsoft.com/

GROTE ÉN KLEINE TOEPASSINGEN

Voorbeeld kleine toepassing: zoekterm-aanvuller:

- Je begint te typen … en je krijgt gelijk suggesties …
- … op basis van top-*k* van $P(T_n \mid T_1,\ldots,T_{n-1})$
- … berekend op basis van zoektermen andere gebruikers

Deze kennen we natuurlijk al: Zoekmachines

- Werking: stel vraag, vergelijk met documenten (via index), sorteer passende documenten naar **relevantie**
  - Relevantie(D_i) = kans dat de gebruiker die de vraag stelde in feite op zoek was naar document D_i
  - maw, Relevantie(D_i) = $P(D_i \mid T_1,\ldots,T_n)$
  - ➢ Google: Combinatie taalmodel en Pagerank

GOOGLE PAGE RANK (±1998)

Volgens Google:

- “PageRank works by counting the **number** and **quality** of links to a page to determine a **rough estimate of how important** the website is. The underlying assumption is that more important websites are **likely to receive more links** from other websites.”

Het algoritme is gebaseerd op:

- “PageRank is a **probability distribution** used to represent the *likelihood that a person randomly clicking on links* will arrive at any particular page”

- Aha, ook simpelweg meer kansrekening

GOOGLE PAGE RANK (±1998)

- Stel een miljoen apen surfen over het web door willekeurig op links te klikken en URLs in te typen
- Op elk moment, naar verwachting welk percentage apen kijkt naar pagina D?

Gegeven document D, haar pagerank in stap *n* is:

- $P_n(D) = (1-\lambda)P_0(D) + \lambda(\Sigma_{A \text{ linkt naar } D} P_{n-1}(D')P(D|A))$

waarbij

- P(D|A): kans dat de app D bereikt via A = 1/ aantal uitgaande links dan A
- λ: kans dat de aap op een link klikt
- 1-λ: kans dat de aap een URL intypt

Kansverdeling over webpagina's. Dat zijn er veel, maar niet onmogelijk veel

BIG DATA: WAAROM NU?

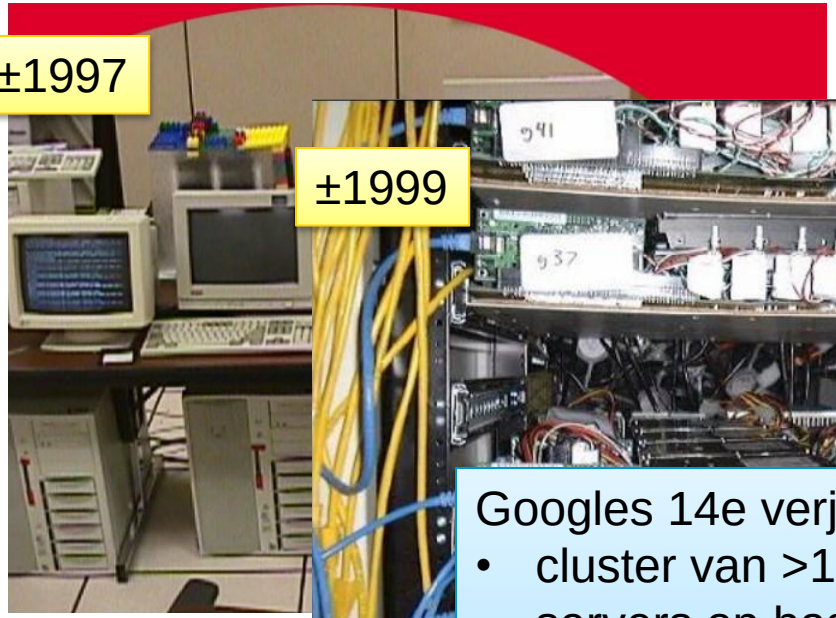
Kansrekening voor dergelijke toepassingen is

- Modelleren
- Tellen, optellen, vermenigvuldigen, sorteren
- Voor heel heel heel heel heel veel teksten
  … om de wet van te grote getallen op te laten gaan

Wat heeft Google en consorten ons gebra

- Niet zozeer zoektechnologie …
- … maar technologie die het mogelijk maakt woorden te tellen e.d. voor **voldoende grote** collecties teksten
- Googles “index” is een simultane kansverdeling!

GOOGLE



Googles 14e verjaardag:

- cluster van >100,000 servers op basis van doorsnee hardware
- >20 miljard web pagina's geïndiceerd / vindbaar

COMPUTERS LEREN LEZEN

Eén van de big data-beloftes: computers te leren lezen … ja echt ***begrijpend*** lezen

- IBM Watson kan dit (tot op zekere hoogte)

Taal is vreselijk **ambigu**

Wat is er eigenlijk zo moeilijk aan lezen?

- Voorbeeld-tweet:
  - **Lady Gaga** - **Speechless** live @ **Helsinki** 10/13/2010 http://www.youtube.com/watch?v=yREociHyijk . . . @ladygaga also talks about her Grampa who died recently
- Nog eentje: "**Paris Hilton** stayed in the **Paris Hilton**"

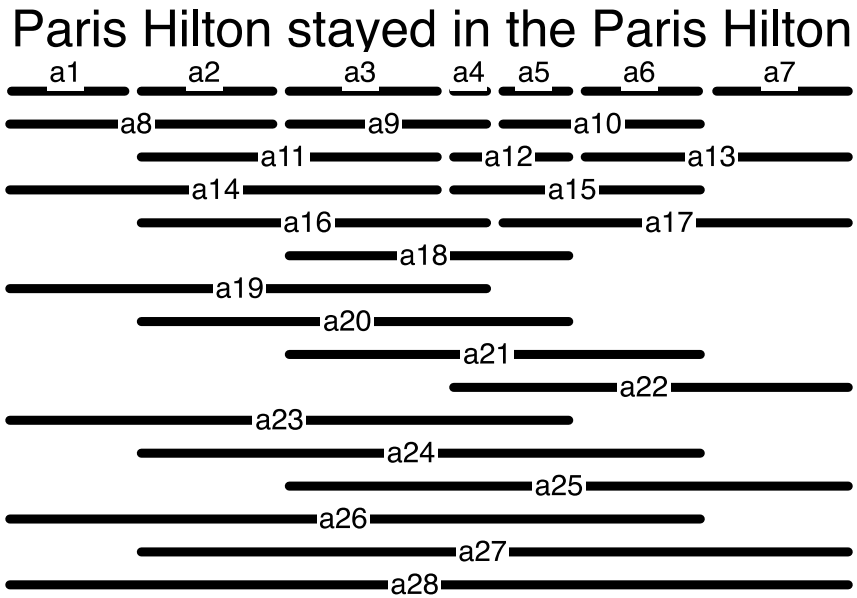
INTERPRETEREN = ANNOTEREN MET BETEKENIS

Annotatie = verwijzing naar een entiteit in een kennisbank

Ambiguïteit: elke annotatie meerdere mogelijke kandidaten

Die combinatie annotaties die het waarschijnlijkst is, dwz het beste bij elkaar past,
is de meest waarschijnlijke interpretatie van de zin

Sherlock Holmes-style:
“when you have eliminated the impossible, whatever remains, however improbable, must be the truth”

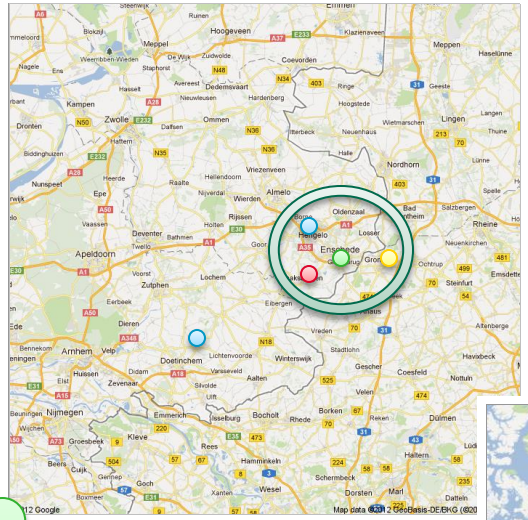


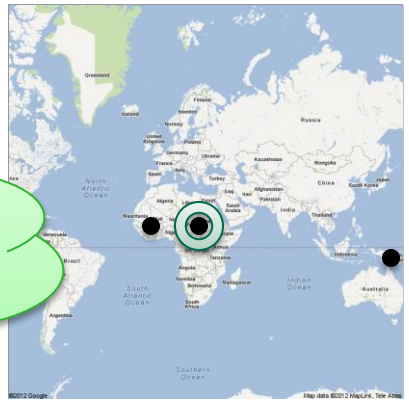
VOORBEELD MET TOPONIEMEN

TOPONIEM = WOORD(EN) DIE VERWIJZEN NAAR EEN LOCATIE

The cottage is in **Usselo**. **You** can shop in the nearby towns of **Enschede**, **Hengelo** and **Gronau**. Cool boat rides on the river **Dinkel**.

- Usselo: 1 (NL)
- Enschede: 1 (NL)
- Hengelo: 2 (NL, NL)
- Gronau: veel (DE)
- You: 4 (Burkina Faso, Papua New Guinea, Chad, Chad)





Bij elkaar passen: NL/NL/NL/DE/Chad ➔ NL

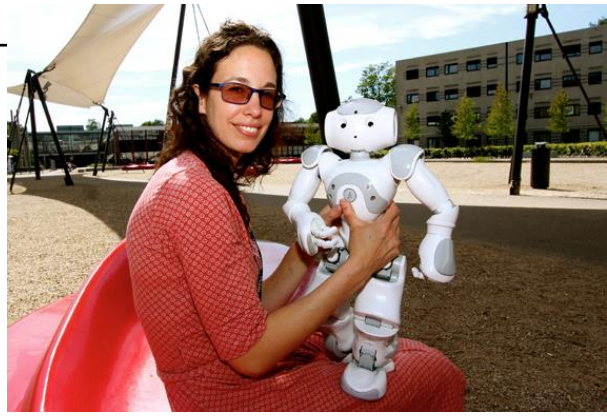
Heel vaak past “You” niet, dwz ligt ‘t ver van de rest ➔ geen toponiem

WAT LIGT ER NOG MEER IN HET VERSCHIET?

EEN SELECTIE

Social robotics / Kunstmatige intelligentie

- begrijpen van taal
- begrijpen van emoties
- begrijpen van non-verbale communicatie
- begrijpen van humor!?!



Prof. Vanessa Evers

Commercie: inzicht in klanten, producten, kansen (vb: micro-targeting, onderhoud)

Zorg en voeding: inzicht in cellen, stoffen, reacties, lichaam, effectiviteit (vb: medicijnen, behandelingen, personalisatie)

Maatschappij: milieuonderzoek, forensics / intelligence (bv: fraude)

BIG DATA HEEFT VALKUILEN

- Mythe: meer data is beter
- Mythe: meer data scientists is beter

En de standaard valkuilen van kansrekening / statistiek:

- **Bias**, met bijvoorbeeld als gevolg
  - Discriminatie
  - Onjuiste inzichten en beslissingen / overgeneralisatie
- We zien **correlaties** geen oorzakelijke verbanden
  - Bijvoorbeeld Google Flu

CONCLUSIE (1)

Welke wiskunde kan **toveren met data**?

- **Kansrekening**

We kunnen tegenwoordig (simultane) kansverdelingen

- heel dicht benaderen op basis van voldoende data
- volledig opslaan, ook de hele hele grote
- deze gebruiken, voor hele grote, maar ook voor de meest kleine toepassingen

CONCLUSIE (2)

Pas echt big data … “when **magic** happens”

- De hoeveelheid data overschrijdt een grens waar intelligent semantisch gedrag uit de data oprijst

Voorbeelden:

- Scene completion, Google Translate, IBM Watson

Grote beloftes voor kunstmatige intelligentie

- Eén nader bekeken: Natuurlijke taalverwerking
- Aantrekkelijke andere onderwerpen: social robotics, (fraud) forensics / intelligence, milieu, zorg & voeding