Big Data, Big Biases?

Auke Hunneman
Associate Professor
BI Norwegian Business School

Kampanjeskolen: Fremtidens handel
October 16th, 2018

Personal Information

Auke Hunneman
Associate Professor @ BI Oslo
Associate Dean Msc in Business Analytics @BI Oslo
Adjunct Associate Professor @ UiS

Contact Information
Phone: +47 4641 0573
Email: auke.hunneman@bi.no
Office: C4-087

Big Data in Retailing

- Walmart collects around 2.5 petabytes (= 2,500,000 gigabytes) of information every hour about transactions, customer behavior, location, and devices (McAfee et al. 2012).

- Five dimensions of big data:
  - *Customer*-level data (transaction data, geodemographics, SoMe, UGC);
  - *Product* data (at more granular SKU level and for more attributes and levels);
  - Real-*time* data which allows for continuous measurement of customer behavior);
  - Customer's geo-spatial *location*, which enables hypertargeting.
  - Omni-*channel* data.

Possible Benefits

- Supply chain management (predictive analytics);
- Understanding customer in-store movements;
- Micro targeting (e.g. geo-location based);
- Insights into customer journeys across touchpoints;
- Evaluating the profit impact of each channel;
- Customization of product offering and communications towards consumers.
- Web shop optimization (A/B testing)

Magma Article

- Hunneman, A., (2018) “Store skjevheter ved bruk av stordata?,” Magma, 18 (4), 68-71.
- Hunneman, A., (2018) “Big Data, Big Biases?,” Medium, https://medium.com/@aukehunneman/big-data-big-biases-a2a3046217be.

STORE SKJEVHETER VED BRUK AV STORDATA?

AUKE HUNNEMAN er førsteamanuensis ved Institutt for markedsføring på Handelshøyskolen BI. Han har en ph.d i økonomi fra Universitetet i Groningen. Han har publisert i flere forskningstidsskrift som Journal of Retailing og Journal of Business Research. Hans forskningsinteresser er innen detaljhandel, sosiale nettverk, markedsføringsmodeller og *marketing accountability*.

SAMMENDRAG

Skaper bruk av stordata store skjevheter? Til tross for at tilgangen av stordata har bidratt til store endringer i hvordan vi designer forretningsmodeller, og fortsatt vil være en viktig driver i denne utviklingen, har flere den siste tiden kommet opp med en del kritiske merknader. Målet med denne artikkelen er å nyansere de ofte over-optimistiske forventningene til de positive effektene av stordata som verktøy. Artikkelen identifiserer ulike mulige begrensninger ved bruken av stordata. Disse begrensningene er knyttet til utfordringer både med selve analyseverktøyet og de slutningsfeil som skjer som en følge av menneskelig fortolkning (og samspillet mellom disse to). Ved å anerkjenne de utfordringene som kan oppstå ved utarbeidelse av prediksjoner og forecasting, er det mitt mål å sikre at stordata ikke kun blir en variant av Keiserens nye klær.

Stordata er i vinden! Et enkelt Google-søk på ordene «Big Data» genererer i skrivende stund (per 26. januar 2018) over 88 millioner treff. Også i næringslivspressen er det stor entusiasme. McKinsey omtaler stordata som «den mest revolusjonerende muligheten innen markedsføring og salg siden internett ble allemannseie», og *Harvard Business Review* utpekte *data scientist* (datainge niør) til det 21. århundrets mest sexy stillingstittel (Davenport & Patil, 2012; McKinsey Chief Marketing and Sales Officer Forum, 2013). Det later ikke til å være tvil om at stordata har snudd opp ned på måten vi driver forretninger på, og vil fortsette med å gjøre det. Samtidig har det også kommet noen innvendinger i den senere tid. Stordata er en global industri til en verdi av 130 milliarder dollar, men likevel viser det seg at over 73 prosent av stordataprosjektene er ulønnsomme (Wang, 2018). Eksperter advarer dessuten bedriftene mot å satse for mye på innsamling og lagring av data uten å vite hvordan man faktisk kan generere verdier (både for kundene og for bedriften selv) av disse dataene.

Jeg mener det er på sin plass å nyansere de tidvis altfor optimistiske forventningene til de store datasettenes fortreffelighet. Misforstå meg ikke: Jeg er ikke

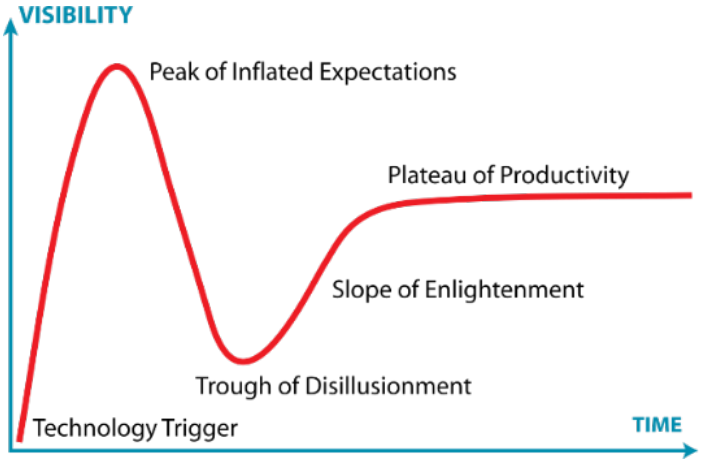
Big Data Are Everywhere…

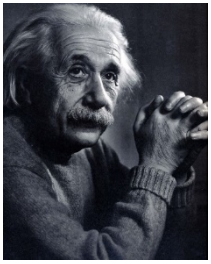
- **Large enthusiasm for Big Data.**
  - McKinsey: “the biggest game-changing opportunity for marketing and sales since the Internet went mainstream”.
  - Harvard Business Review: “Data scientist is the sexiest job of the 21st century”.
- **However:**
  - Too large focus on data capture and storage, too little on value creation.
  - 73% of big data projects are not profitable (Wang 2018).

- **Hence: I call for a cautious and a little skeptical approach.**

Amara's Law

"We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run." (Brooks 2017)





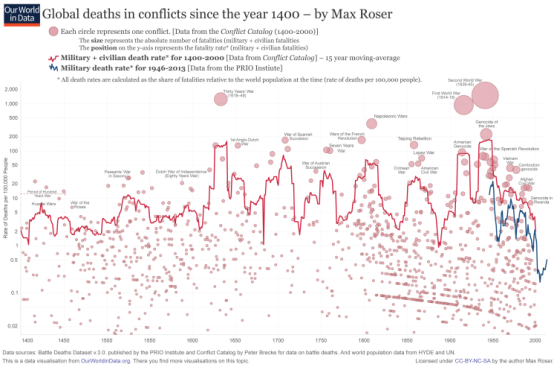
Not Everything Counts

P&G Cuts More Than $100 Million in 'Largely Ineffective' Digital Ads

Consumer product giant steers clear of 'bot' traffic and objectionable content

- **Humans prefer the measurable over the immeasurable: quantification bias.**
  - Are digital media so popular among advertisers because we can measure their effectiveness? → impressions, clicks, etc. are not purchases.
- **We may miss out on the context of behavior (culture, emotions, etc.).**
  - Netflix only discovered the phenomenon "binge watching" after a netnographer spent considerable time observing their customers.

Signal vs Noise



- **High-frequency data are by definition noisy.**
- **How can we make sure we focus on the signal and not the noise?**
- **People respond stronger to losses than to similar gains (prospect theory).**
- **Important considerations: whether and when to take action?**

Needle In A Haystack

- **Datasets with a large number of variables contain many spurious relationships.**
  - Hence: Data need to be approached from the perspective of a “theory” in order to find meaningful patterns.
  - What we see depends on our expectations and questions (Felin 2018): https://www.youtube.com/watch?v=vJG698U2Mvo&feature=youtu.be

- **Status quo:**
  - Practice: small stats on big data → biases
  - Academia: big stats on small data → overfitting

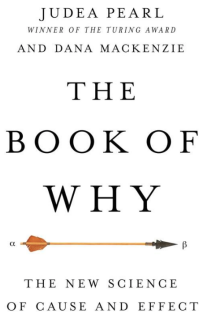
  - Solution: big stats on big data! (large samples and increased computing power).

Beer And Diaper Sales Are Correlated

What exactly does that mean?:

- When fathers are sent out on an errand to buy diapers, they often purchase a six-pack of their favorite beer as a reward.
- Or: beer and diapers are located in close proximity within the store.
- Or: both.

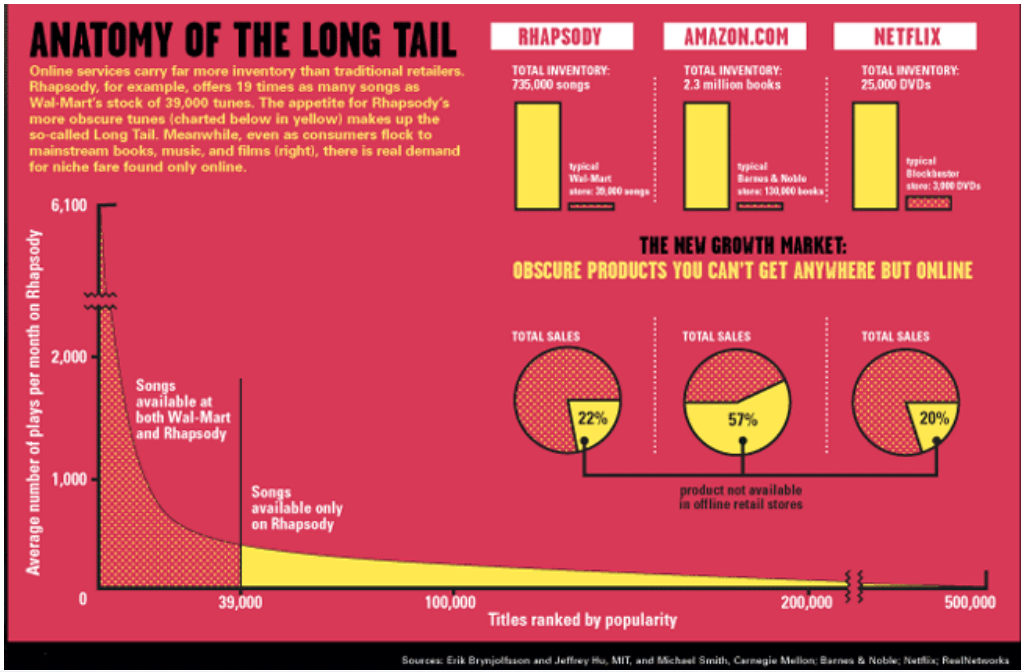
- Hence, we need to know the WHY?



The Winner Takes It All

- **Online products are scalable!**
- **Social influence is amplified.**
- **Long tail distributions are very common.**
  - Google “owns” 90% of search advertising, Facebook earns 80% of mobile social traffic, and Amazon about 75% of the e-book sales.
- **Prediction becomes nearly impossible: A market functions as a complex system. We need extremely large samples!**

A Long Tail Distribution



The Degrees of Freedom Problem

- **The same consequences can be caused by several antecedents.**
  - If you see a puddle of water on the floor, it's hard to tell where it came from. A particularly shaped ice cube or something else?
- **Sociologists call this the micro-macro problem.**
- **Combined with our *narrative fallacy*, this leads to conspiracy theories in times of data overload (Harari 2015).**

Birds Of Same Feather Flock Together

- **We tend to connect to similar people (homophily).**
- **As a result, we can decide to only listen to information that «suits» us (confirmation bias).**
- **Access to new information is limited. Detrimental to innovation!**
- **We need to leverage the strength of weak ties again (Granovetter 1977)**