Westfälische Wilhelms-Universität Münster

We SBE Combining the best.

Past, Present, and Future of Big Data

Gottfried Vossen
University of Münster, Germany

Prof. Dr. Gottfried Vossen
DBIS Group,
Dept. of Information Systems
WWU Münster
Germany

ERCIS

One step ahead with big data?

One step ahead with big data?

The New York Times

Desperate for Workers, Restaurants Turn to Robots

They can make French fries, mix drinks and even clean toilets, and they never ask for a raise. But they also break down.

https://www.m/watch?v=McTJlQHrHuE

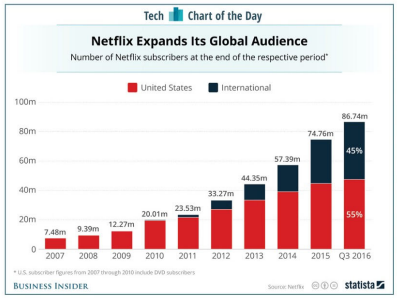
Contents

- BD Past
  - Digitalization
  - BI
- BD Present
  - Data as a resource
  - AI on the raise
- BD Future
  - No more talk of “big”
  - Data as a currency
  - Risky issues

Contents

- **BD Past**
  - **Digitalization**
  - **BI**
- BD Present
  - Data as a resource
  - AI on the raise
- BD Future
  - No more talk of “big”
  - Data as a currency
  - Risky issues

The Netflix Evolution



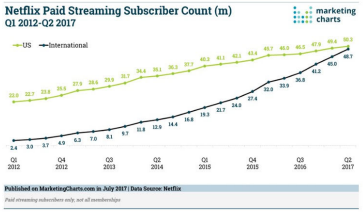
1997 — **NetFlix Founded**

1998 — **Offers DVD Rentals**
Website launches and offers 925 rentals.

1999 — **Subscription Model Begins**
Users are able to subscribe to a pay-by-month service to have DVD's delivered to their house.

2000 — **Studio Deals**
Agreements signed with Warner Brothers and Columbia film studios

2003 — **1 million subscribers**
Netflix had clearly made an impact in the DVD delivery market and was gaining a following (the chart below shows the subscriber count from 2012-2017. Netflix now approaching 100 million subscribers.)



2007 — **Begins Offering Video Streaming**
Probably its biggest and most impactful directional shift. Also, a month after streaming launched, they delivered their 1 billionth DVD.

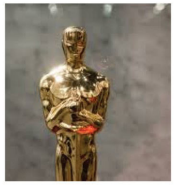
2010 — **International Company**
Launched in Canada

2011 — **Qwikster Fails in One Month**
Netflix launches Qwikster to handle DVD rentals, while Netflix became only video streaming.

2013 — **House of Card's Launches**
Its first original series earns 9 Emmy nominations and proved that TV will now be controlled by streaming platforms.

2016 — **160 Countries Go Live**
At the 2016 CES, Netflix tripled its scope from availability in 60 countries to 230.

2017 — **Wins First Oscar**
The White Helmets wins for Best Documentary Short Subject. Netflix reaches 100 million members globally.



The Beginnings

Term made popular by Tim O'Reilly in 2005

- Blogs
- Wikis
- Audio
- Video
- Social software
- .....

By Markus Angermeier - http://kosmar.de/archives/2005/11/11/the-huge-cloud-lens-bubble-map-web20/

Early Implication: Recommendation

Forbes / Tech

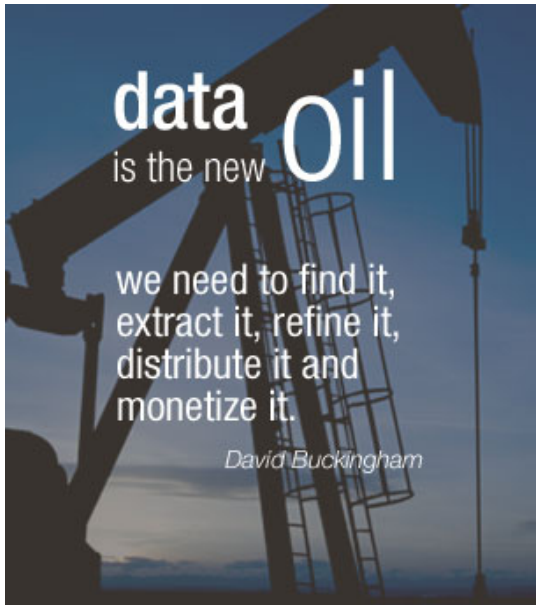
Top

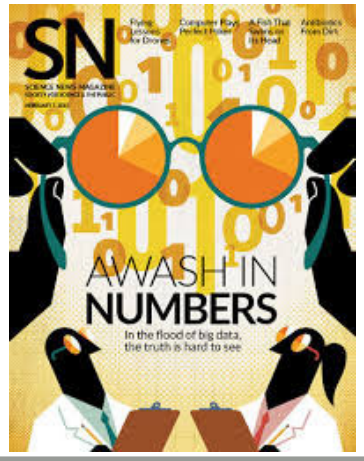
FEB 16, 2012 @ 11:02 AM **2,913,914** VIEWS

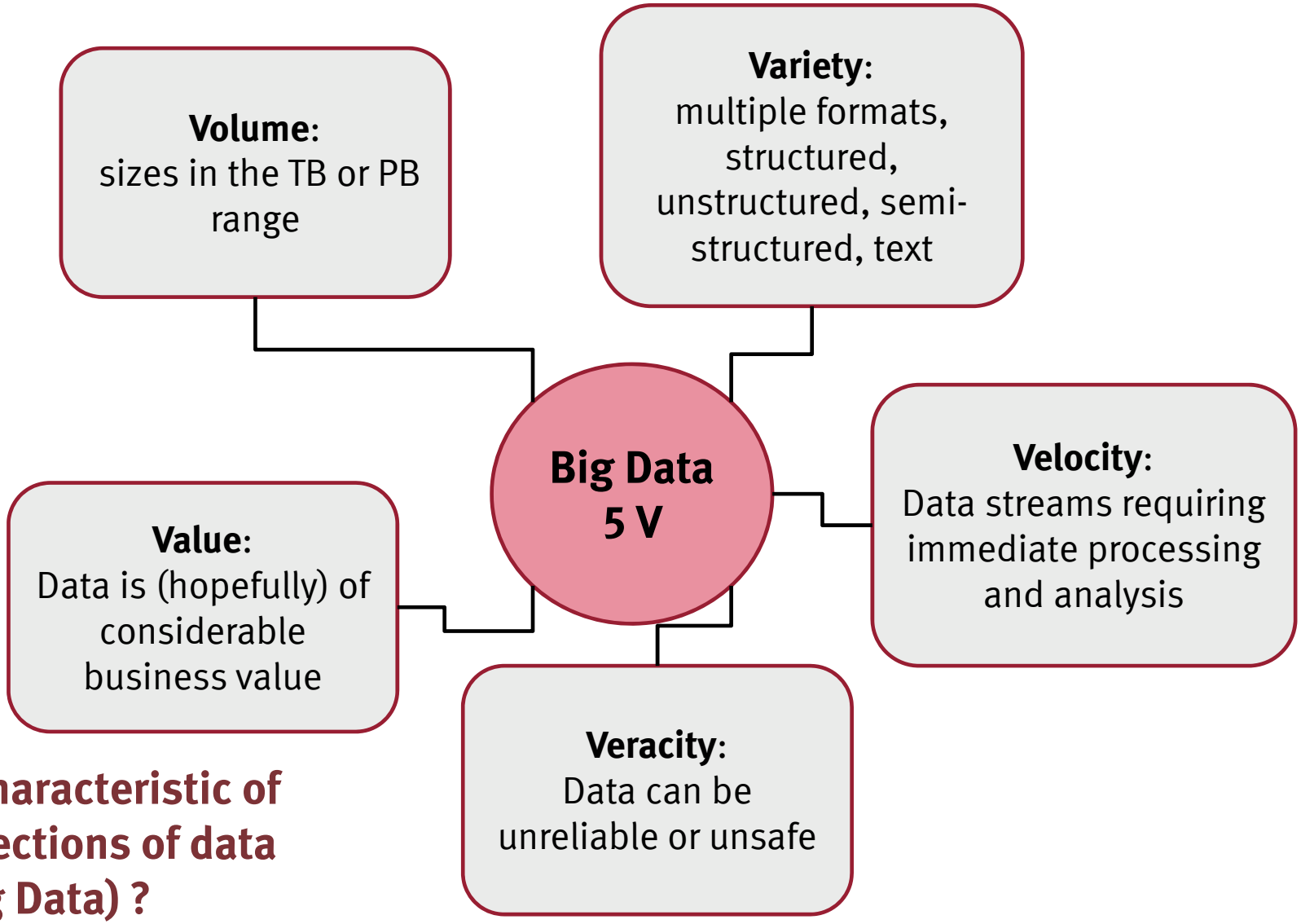
How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Then it took off …







What is characteristic of large collections of data (a.k.a. Big Data) ?

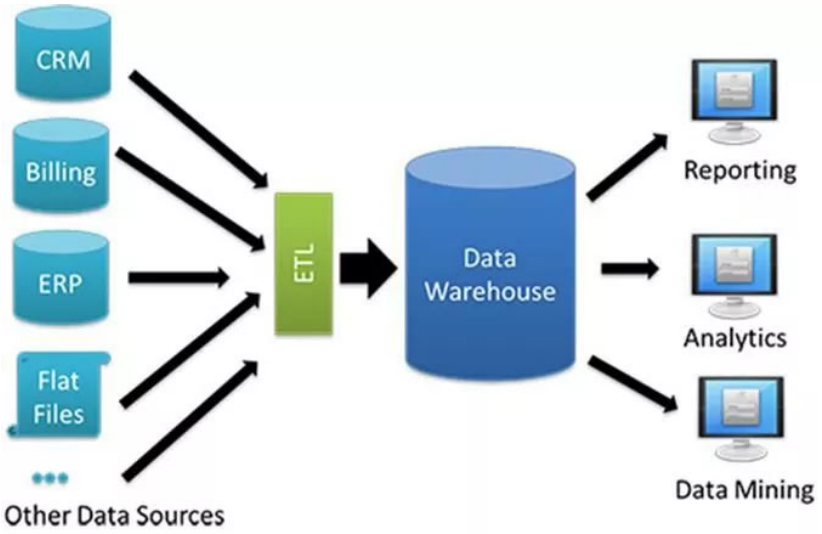

2021 This Is What Happens In An Internet Minute
60 SECONDS
facebook
1.4 Million Scrolling
21.1 Million Texts Sent
YouTube
500 Hours Content Uploaded
414,764 Apps Downloaded
695,000 Stories Shared
200,000 People Tweeting
2 Million Swipes
197.6 Million Emails Sent
2 Million Views
5,000 Downloads
932 Smart Audio Devices Shipped
3 Million Images Viewed
69 Million Messages Sent
3.4 Million Snaps Created
$1.6 Million Spent Online
28,000 Subscribers Watching
9,132 Connections Made
Linked in
NETFLIX
imgur
amazon echo
Google Home
Tik Tok
twitch
tinder
Google play
App Store
Created By:
@LoriLewis
@OfficiallyChadd


2013
Google
Spotify
YouTube
facebook
DOMAINS
WordPress
skype
flickr
WEBSITES
twitter
ebay
snapchat
tumblr.
amazon
E-MAIL
Linked in
Pinterest
Instagram
100 Stunden Videos hochgeladen
2 Millionen Suchanfragen gestellt
14 neue Songs hinzugefügt
70 neu registriert
15 Tausend Songs von iTunes heruntergeladen
347 neue Blogs entstehen
571 neu ins Netz gestellt
105 Tausend Euro Handelsvolumen erbracht
20 Tausend Fotos hochgeladen
204 Millionen E-Mails versendet
3600 Fotos veröffentlicht
11 Tausend Nutzer surfen hier
11 Tausend Geschäftsanfragen gestellt
93 Tausend Euro Warenumsatz getätigt
104 Tausend Fotos geladen
278 Tausend Tweets gepostet
20 Millionen Fotos betrachtet
1,4 Millionen Gespräche geführt
1,8 Millionen Mal „like" geklickt
41 Tausend Kommentare gepostet
FOCUS 33/2013

The idea behind 'Big Data' is that everything we do leaves a digital trace (i.e., data) which we (and others) can analyze and use. Big Data thus refers to this data collecting and our ability to derive benefit from it.

Furthermore the derivation of benefit is increasingly automated via artificial intelligence and in particular machine learning. Big Data helps to train, but also to verify and apply usages in a variety of areas.

The Arrival of BI



Contents

- BD Past
  - Digitalization
  - BI
- **BD Present**
  - **Data as a resource**
  - **AI on the raise**
- BD Future
  - No more talk of “big”
  - Data as a currency
  - Risky issues

Computer Science Resources

- Time
  - How much **time** does an algorithm need to complete a computation (measured relative to the size of the input)?
- Space
  - How much **memory** does an algorithm need to complete a computation?
- Classical distinction: efficient (i.e., polynomial) vs. inefficient (i.e., exponential) algorithms

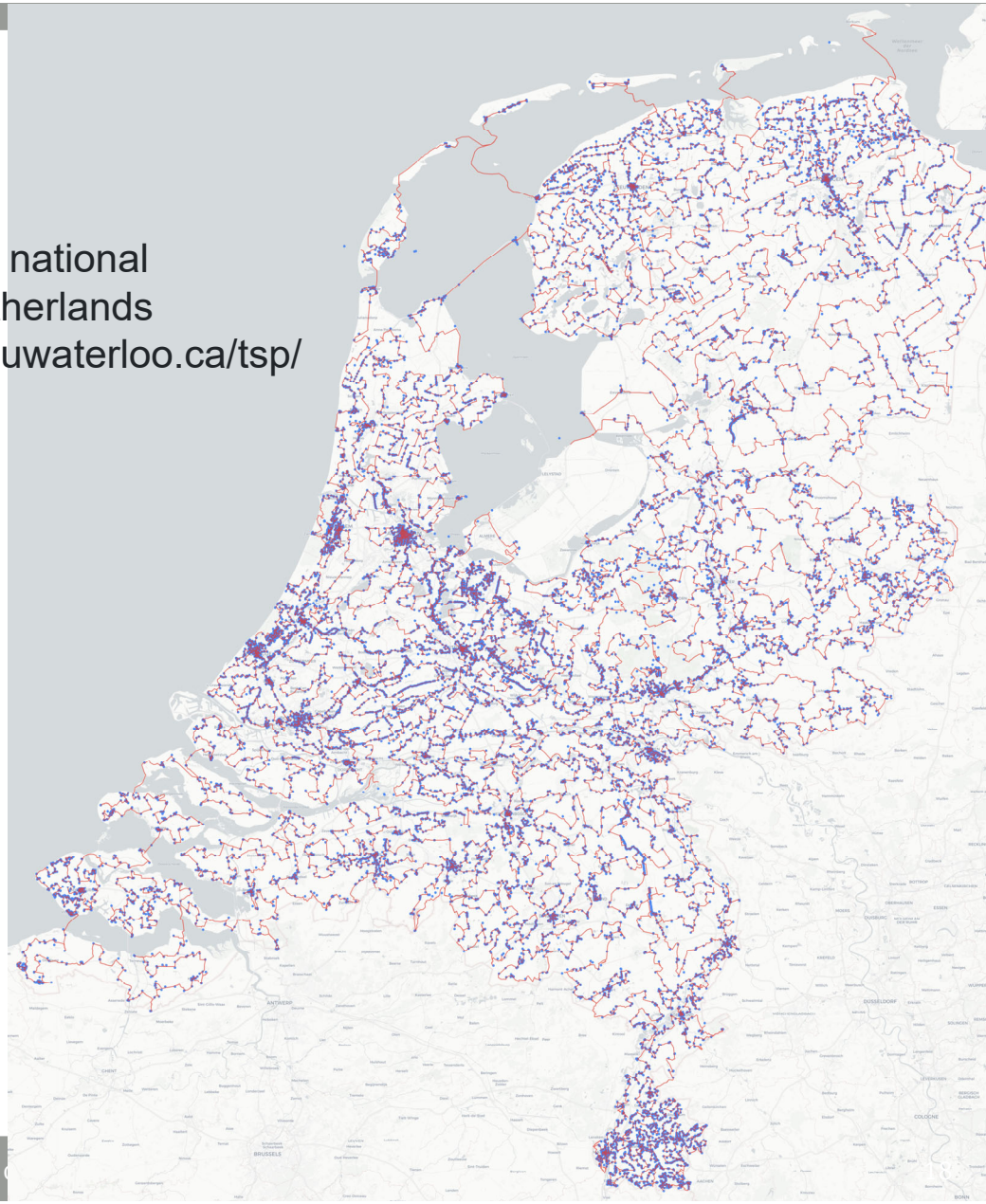
Example: Sorting n numbers

- Time
  - O(n log n)
- Space
  - O(n)
- Conclusion: problem can efficiently be solved

Example: TSP

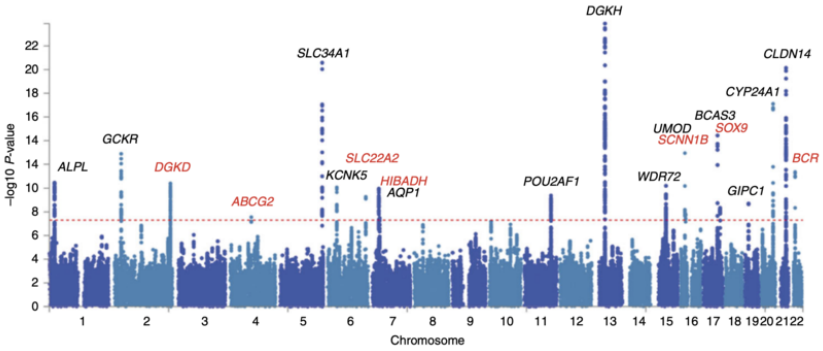
- Time
  - $O(2^n)$
- Space
  - polynomial
- Conclusion: problem cannot efficiently be solved

NL57912
Cycling tour to 57,912 national monuments in the Netherlands
see https://www.math.uwaterloo.ca/tsp/



Example: Testing n Items for Similarity

- Time
  - $O(n^2)$ (compare each item to every other item)
- Space
  - O(n) (factors are ignored)

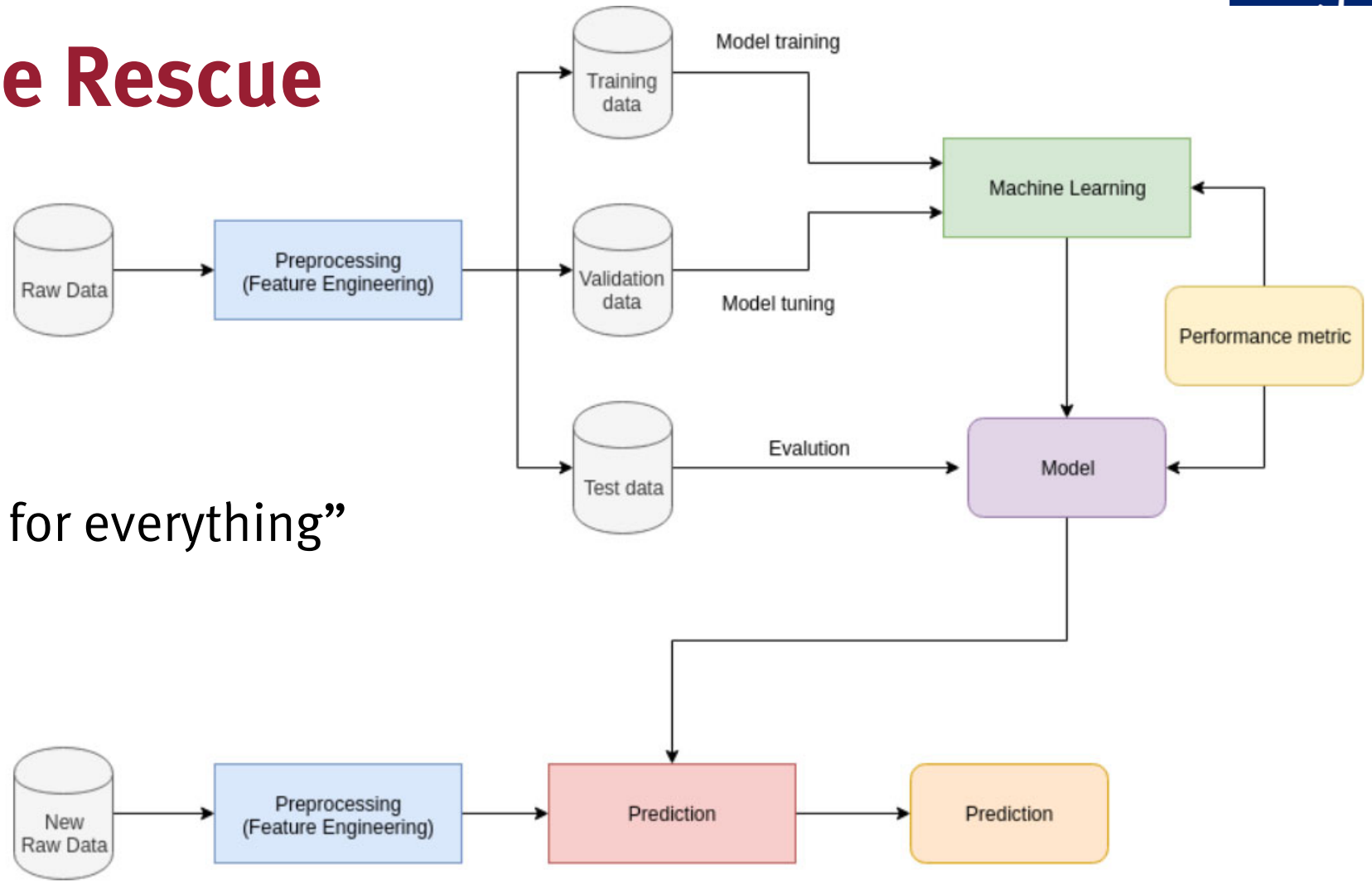


- Seems efficiently solvable, BUT: n can be extremely large, e.g., plagiarism finder, entity resolution, news aggregation, recommendation

Data as a Resource

- If $n = 10^{30}$ or larger, $O(n^2)$ is no longer efficient, but demands new algorithmic solutions, e.g., locality-sensitive hashing
- LSH hashes similar input items into the same "buckets" with high probability
- Conclusion: big data requires new approaches to old problems; not everything solvable by the KIWI principle (although hardware can do a lot)

ML to the Rescue



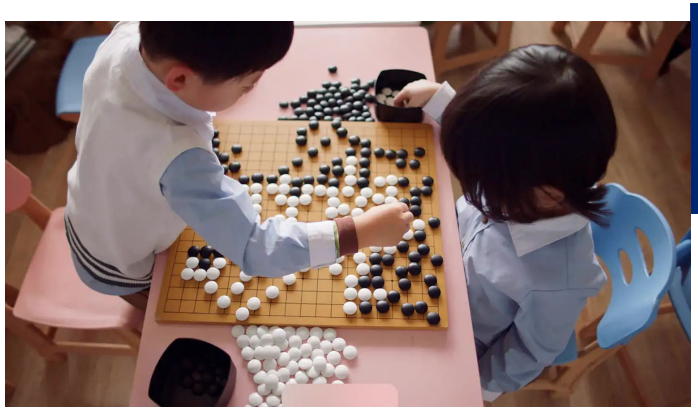
"one algorithm for everything"

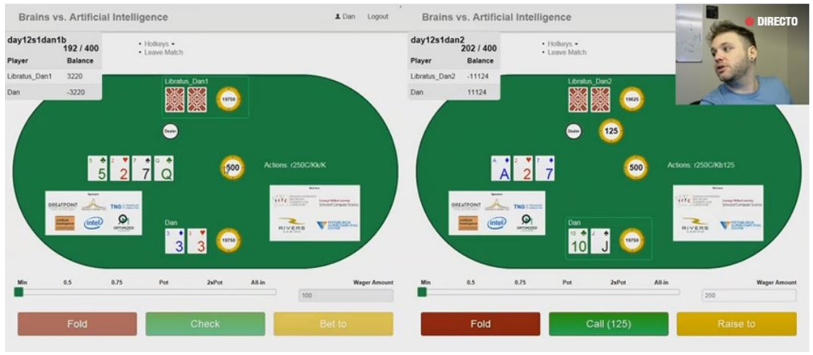
Data-Driven Algorithm Design

- Idea: use learning and data to design an algorithm
- Can overcome major shortcomings of classical design by adapting the algorithm to the domain at hand, in particular when it is applied repeatedly
- **Goal:** given a family of algorithms **F**, a sample of typical instances from domain D (w/ unknown distribution), find algorithm that performs well on new instances from D.
- Examples:
  - Clustering of news articles
  - Pricing
  - Auction design

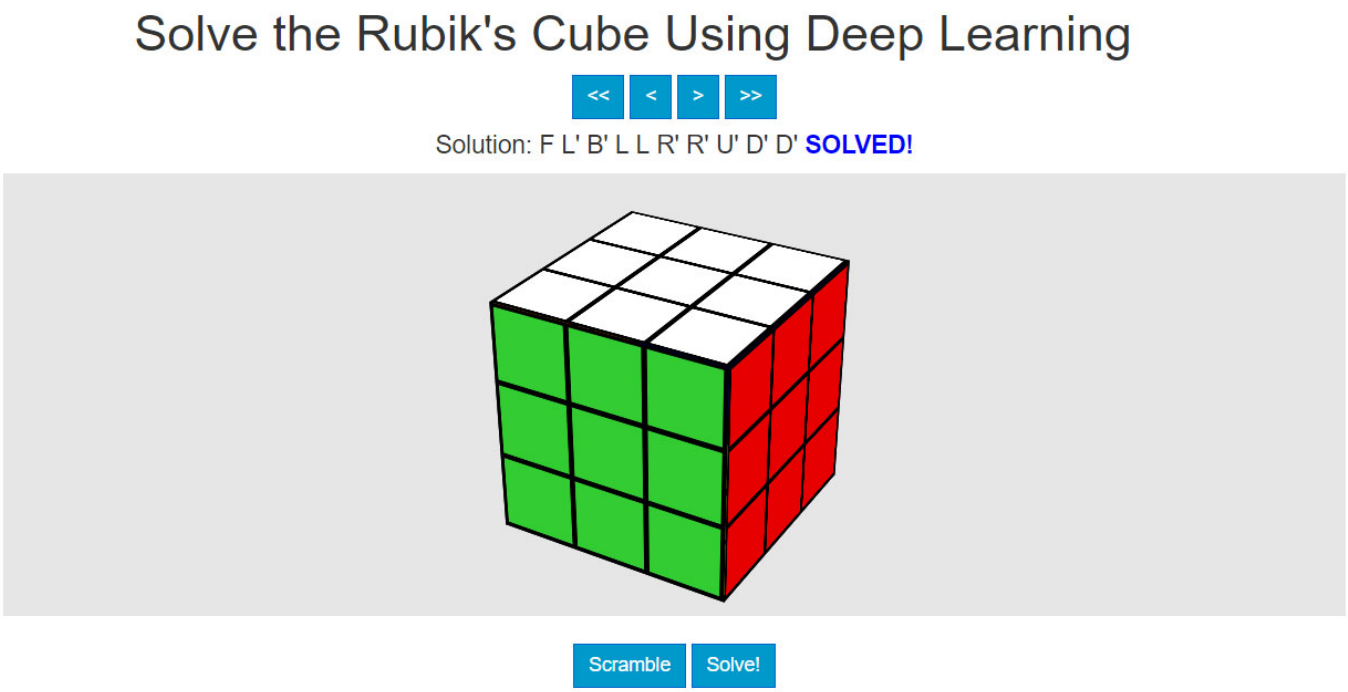
Progress in Gaming

- DeepMind‘s AlphaGo 2015/6
- CMU Libratus 2017
- Rubik‘s Cube 2019
- CMU + FB‘s Pluribus AI 2019





DeepCubeA


Solve the Rubik's Cube Using Deep Learning
<< < > >>
Solution: F L' B' L L R' R' U' D' D' SOLVED!
Scramble
Solve!
Use the mouse to turn the cube.
Turn the faces with the U/D/L/R/B/F keys. Hold shift to turn faces couter-clockwise.
Press scramble to randomly scramble the cube. Press solve to solve the cube using deep learning!

Basketball

Computers Are the New Basketball Coaches

'Today's players will not argue with a computer.' The latest shot-tracking technology in basketball is the latest sign of a profound shift in the making of professional athletes.

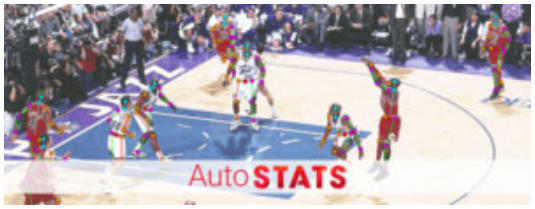
By *Ben Cohen*
Updated July 19, 2019 3:06 pm ET

Admiral Schofield was in the middle of the most important workout of his life at the NBA draft combine a few months ago when he introduced himself to a man in a polo shirt with a logo he recognized. It was the least he could do. Schofield might not be a professional basketball player if not for this person he'd never met.

https://www.wsj.com/articles/nba-technology-coaches-are-computers-11563478009

Further reading:
https://www.sciencemag.org/news/2019/09/watch-ai-help-basketball-coaches-outmaneuver-opposing-team

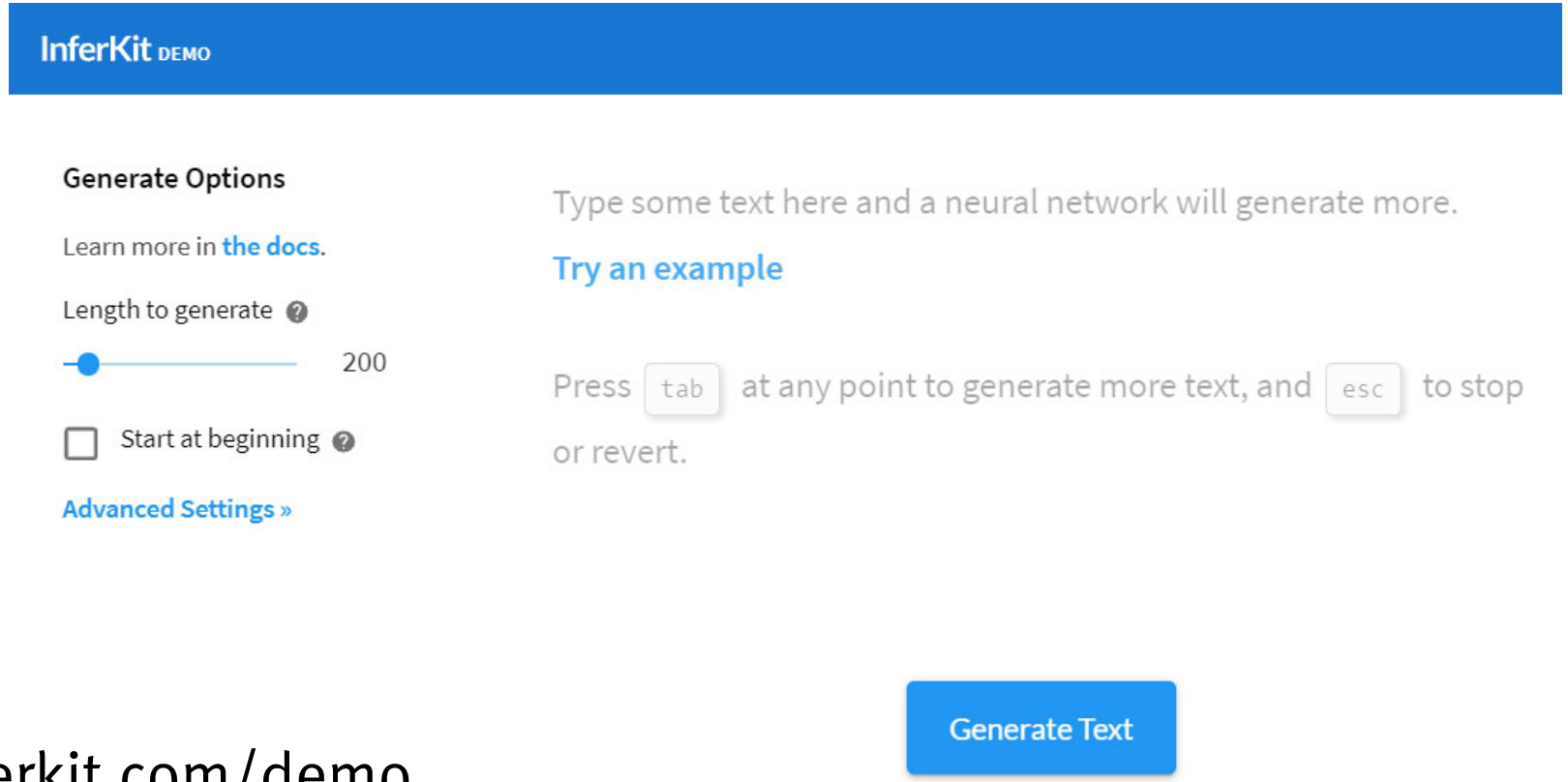
Moneyball, currently in 3.0



- **1.0**: Oakland Athletics' General Manager Billy Beane adopts Sabermetrics with his 2002 and 2003 teams
- **2.0**: commenced around 2010, when STATS rolled out its SportVU program in the NBA, a camera system that uses computer vision to track the movement of players on the screen
- **3.0**: Advanced player tracking data can now be gleaned from the TV broadcasts on CBS Sports and others. Stats Perform makes this data available through its AutoStats program, for detecting things that couldn't previously be detected using deep learning

- Source: https://www.datanami.com/2021/10/05/were-in-the-moneyball-3-0-era-heres-what-it-means-for-live-sports

https://talktotransformer.com/



app.inferkit.com/demo

State of AI in 2021: www.stateof.ai

State of AI Report 2021

The **State of AI Report** analyses the most interesting developments in AI. We aim to trigger an informed conversation about the state of AI and its implication for the future. The Report is produced by AI investors **Nathan Benaich** and **Ian Hogarth**.

10 most evolving BD tech in 2022

1. Elasticsearch: a free open search & distributed analytics engine
2. Hadoop: popular open-source framework
3. MongoDB: a distributed document database
4. Tableau: a visualization tool
5. Cassandra: an open-source column store & distributed NoSQL DBMS
6. RapidMiner: a widely used data analytics platform
7. Qlik: a real-time data integration and analytics cloud platform
8. KNIME: Konstanz Information Miner , an open-source reporting, data analytics, and integration platform
9. Splunk: a platform to transform machine-generated data into times series events
10. R: a PL for statistical computing

https://www.datasciencecentral.com/profiles/blogs/10-most-evolving-big-data-technologies-to-catch-up-on-in-2022

Big Data Present Age

- Data helps
  - to support arguments,
  - to win games,
  - to make more precise decisions,
  - to design algorithms,
  - to develop new business models,
  - to control robots,
  - to steer cars safely (thanks to GPS),
  - to fly planes (thanks to autopilot),
  - to continuously improve search engines and recommenders,
  - to let the stock market crash from time to time.

Contents

- BD Past
  - Digitalization
  - BI
- BD Present
  - Data as a resource
  - AI on the raise
- **BD Future**
  - **No more talk of “big”**
  - **Data as a currency**
  - **Risky issues**

General Observations

- "Big Data" is no longer used as a term, since any data is just big these days
- AI is now applied to anything, everywhere
- The idea of "one algorithm" instead of many seems intriguing, as does that of "general" AI
- Technically using AI is getting easier every week; many things can now be done in your browser, e.g.,
  https://www.dlology.com/blog/top-10-deep-learning-experiences-run-on-your-browser/
  https://gizmodo.com/5-awesome-ai-experiences-you-can-test-out-in-your-brows-1833489624

Goals we are used to

- Data analysis in real-time
- Precise predictions instead analysis of the past
- Person-, situation- and time-individual digital offers (goods, media, services, temperature, light, etc.)
- DIY principle – the customer is my (unpaid) employee!
- On the other hand: request for “explainable AI” and interest in ethical considerations regarding AI and its applications, e.g., dotdata.com, is in its infancy

Future Data Usage Scenarios

- Data as a manufacturing input
  - Well underway in the context of Industry 4.0
- Data as a manufacturing *output*
  - Data is all you get when you buy a product, e.g., a cellphone, a cabinet, a bicycle, even a house

Example: Your Next Car

https://localmotors.com/

- Comes as a dataset which comprises
  - Basic car features
  - Security features you are interested in
  - Accessories as you desire

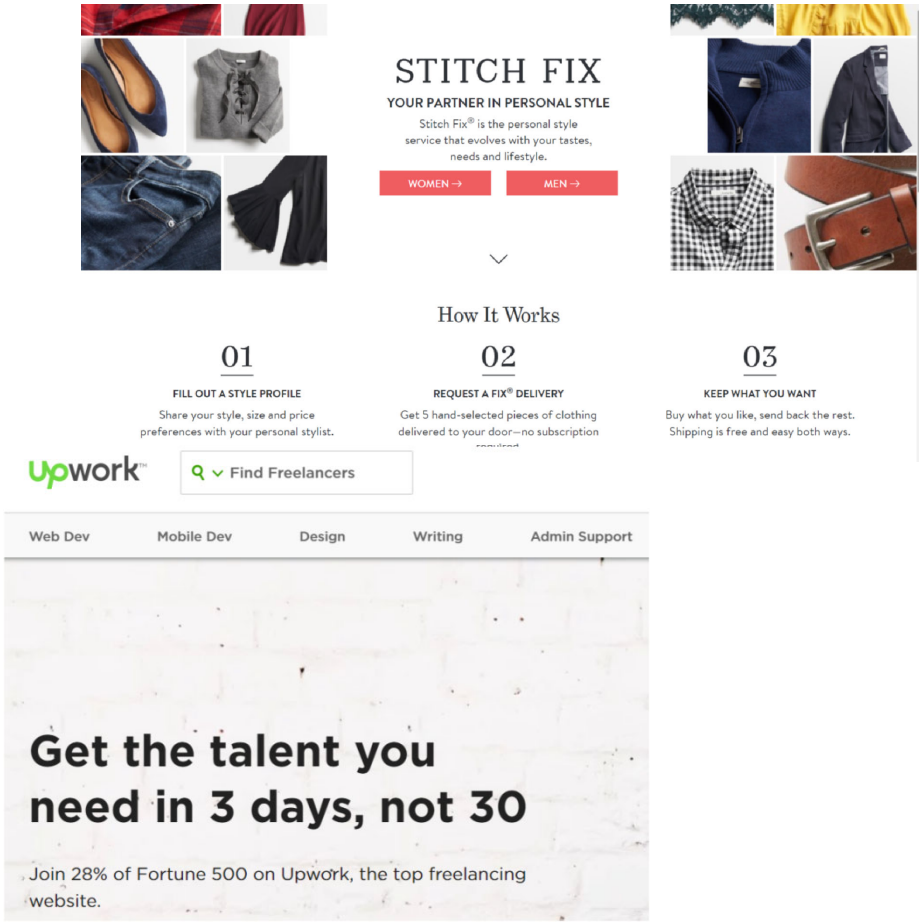
  → mandatory and optional stuff

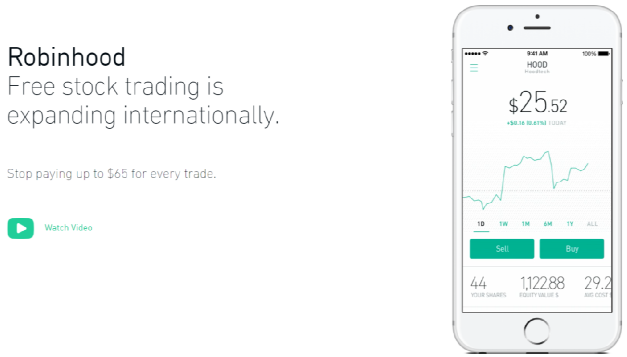
- You buy (but not necessarily get) the data, it is processed at a print shop, from where your new car can be picked up.

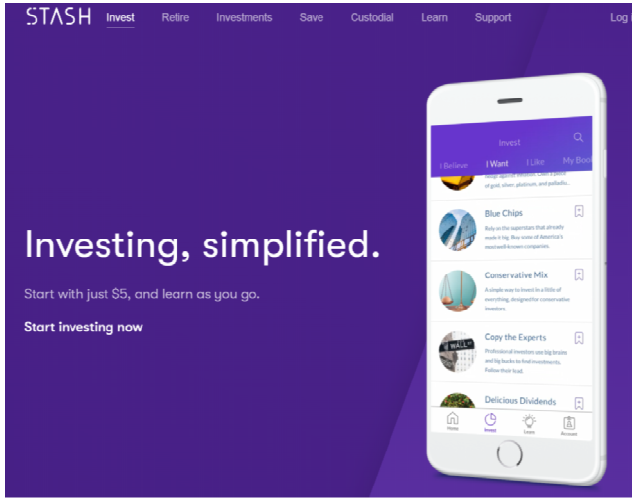
Towards Data as a Currency

- BlaBlaCar idea: market place for passengers
- Brings together millions of drivers and passengers having the same destination every month
- *Most notable: the price is no longer of central importance*; passengers can see
  - how much the driver talks,
  - which music he listens to,
  - whether you can bring your pet …

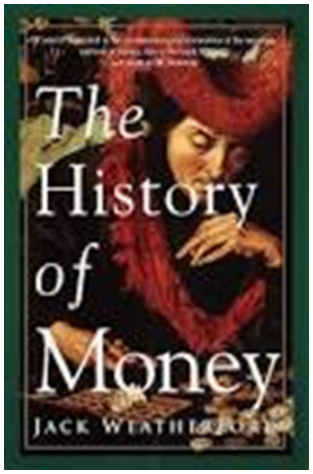
Similar Data-Rich Approaches







What is happening here?



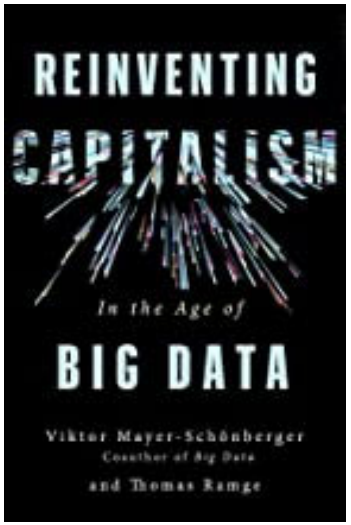
- Prices have been good enough for
  - assessment of goods, products, services
  - comparison of products serving a similar purpose
  - communication between buyer and seller

  for 3,000 years.

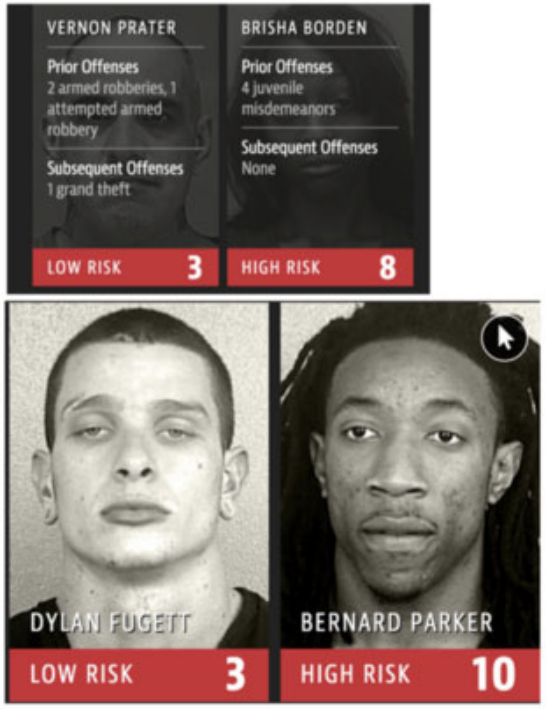
- But: the price of a product or service is a singular number which aggregates all the information that is available about that product or service!

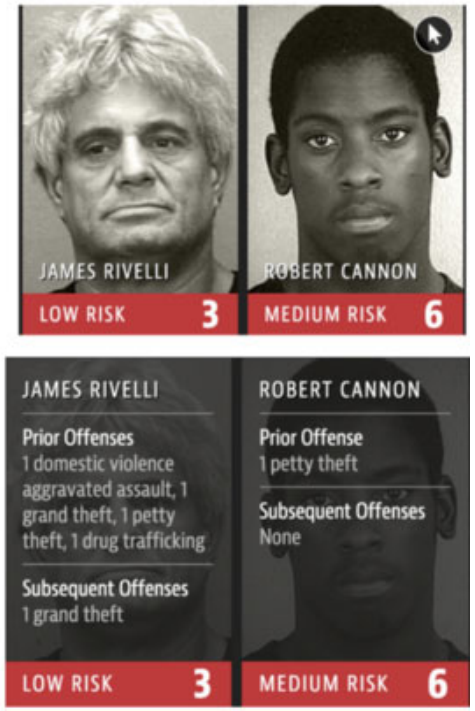
The will change anytime soon!

- Many new market places emphasize data richness and make use of it.
- In many areas the notion of price will thus lose its dominant role for assessment.
- In the near future many single-digit prices might hence get replaced by data.



Finally: Risky Issues





Racial bias in the COMPAS software, discovered by ProPublica

More recently: gender bias

https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1

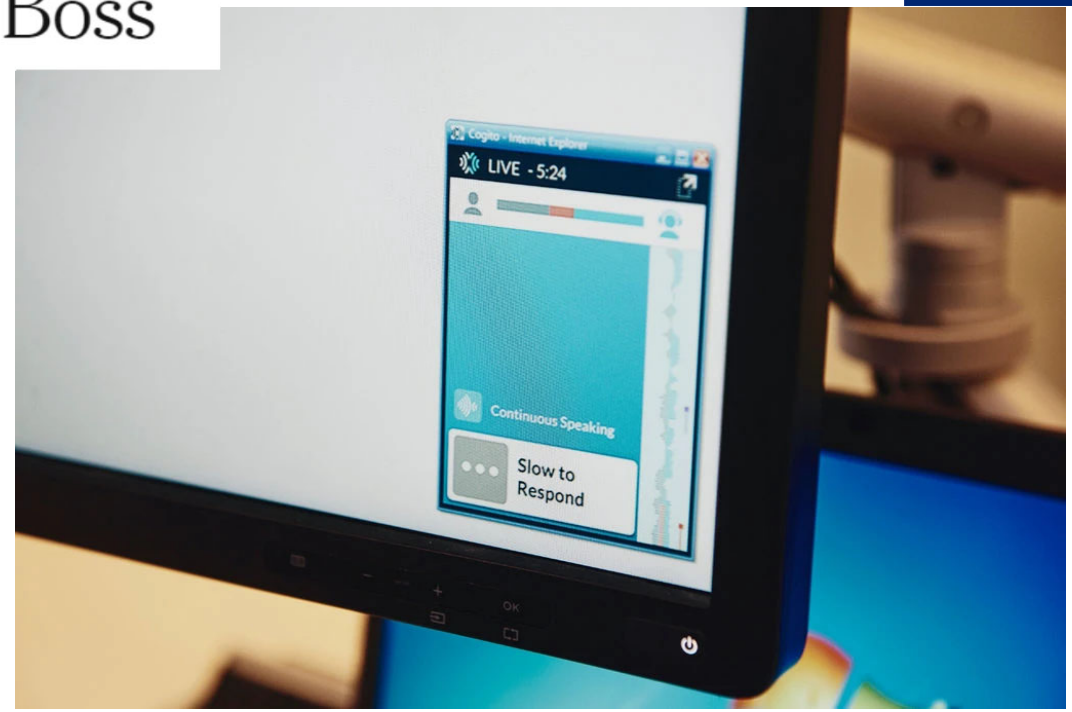
Job Performance Measurement

Wearables can determine whether you're a productive employee.

Employer Tools

https://www.washingtonpost.com/technology/2019/06/28/wearable-technology-started-by-tracking-steps-soon-it-may-allow-your-boss-track-your-performance/

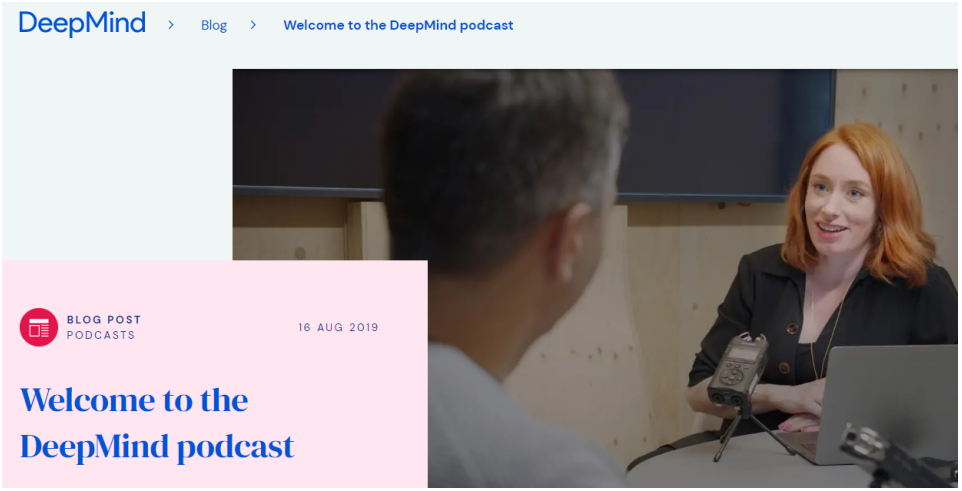
A Machine May Not Take Your Job, but One Could Become Your Boss

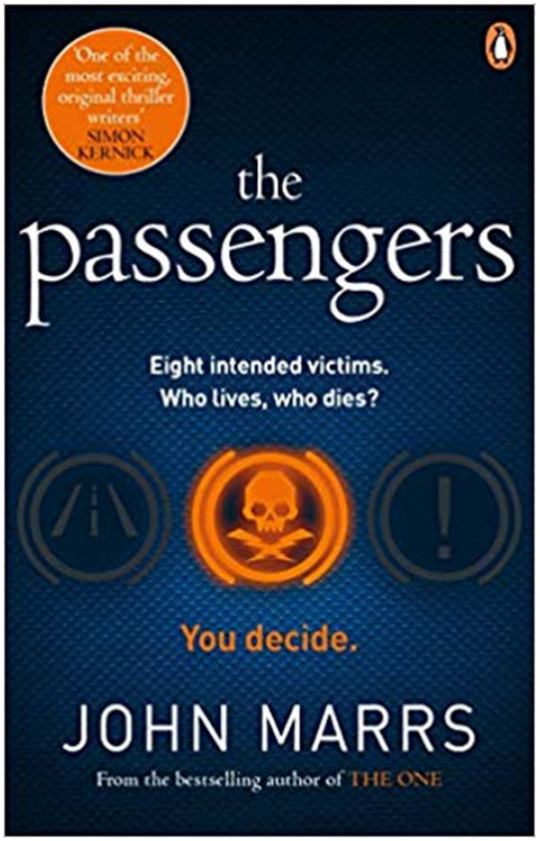


https://www.nytimes.com/2019/06/23/technology/artificial-intelligence-ai-workplace.html

Employer Tools

Something for Your Eyes and Ears





Conclusions: Past, Present and Future of Big Data

- Drivers:
  - Web 2.0, Digitalization
  - Ever increasing computer power
  - Shift in AI from rule-based to statistics-based
- Implications:
  - Data drives almost everything technical these days
  - Protection of personal data a growing challenge
- What to expect:
  - More robots
  - Restructuring of work
  - Towards an on-demand, convenience society

The End