Introduction to Big Data

Big Data – Philosophical perspective

What is more valuable, if you had to pick one?

- experience or intelligence?

- Traditional (computer) science: **logic!** [intelligence]
  - understand the problem, build model / algorithm
  - answer question from implementation of model
- New science: **statistics!** [experience]
  - collect data
  - answer question from data (what did others do?)

Questions and (some) answers

- Find a spouse?
- Should Adam bite into the apple?
- 1 + 1?
- Cure for cancer?
- How to treat a cough?
- *Should I give Donald a loan?*
- Premium for fire insurance?
- When should my son come home?
- Which book should I read next?
- Translate from German to English.

Questions and (some) answers

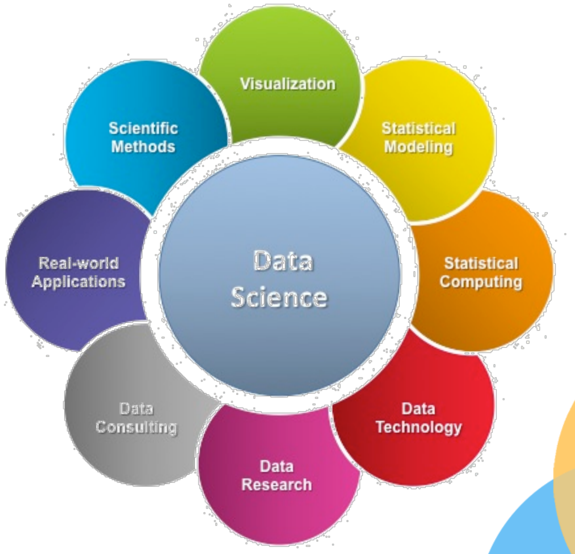
- Find a spouse? *I do not want to know!*
- Should Adam bite into the apple? *If you believe…*
- 1 + 1? *Definition*
- Cure for cancer? *I do not know. Maybe.*
- How to treat a cough? *Yes. (Google Insight)*
- *Should I give Donald a loan? Yes.(e.g.,Schufa)*
- Premium for fire insurance? *Yes.(e.g., …)*
- When should my son come home? *No! But…*
- Which book should I read next? *Yes. (Amazon)*
- Translate from German to English.*Yes.(GoogleTransl.)*

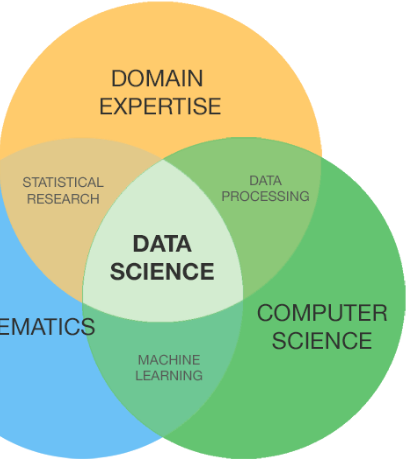
Data Science

- New approach to do science
  - Step 1: Collect data
  - Step 2: Generate Hypotheses
  - Step 3: Validate Hypotheses
  - Step4: (Goto Step 1 or 2)

- Why is this a good approach?
  - Automated: no thinking, less error
- Why is this a bad approach?
  - How to debug without a ground truth?

- More generally, **interdisciplinary** emerging field (see images)





Statistic
Visualisation
Pattern recognit.
DATA science
Data Mining
Machine Learning
Neurocomputing
& Database Data process

“Big” data - Pros & Cons

- Pros
  - tolerate errors
  - discover the long tail and corner cases – machine learning works much better

- Cons
  - More data, more error (e.g., semantic heterogeneity)
  - With enough data you can prove anything
  - still need humans to ask right questions

Big Data Success Story

- Google Translate
  - You collect snippets of translations
  - You match sentences to snippets
  - You continuously debug your system
- Why does it work?
  - There are tons of snippets on the Web
  - There is a ground truth that helps to debug system

Google Translate is based on something called "statistical machine translation". This means that they gather ***as much text as they can find*** *that seems to be parallel between two languages, and then they crunch their data to find the likelihood that something in Language A corresponds to something in Language B. This method* ***works to some extent for language pairs where a lot of more-or-less parallel data*** *is available, for example English-Spanish. […] (quora.com)*

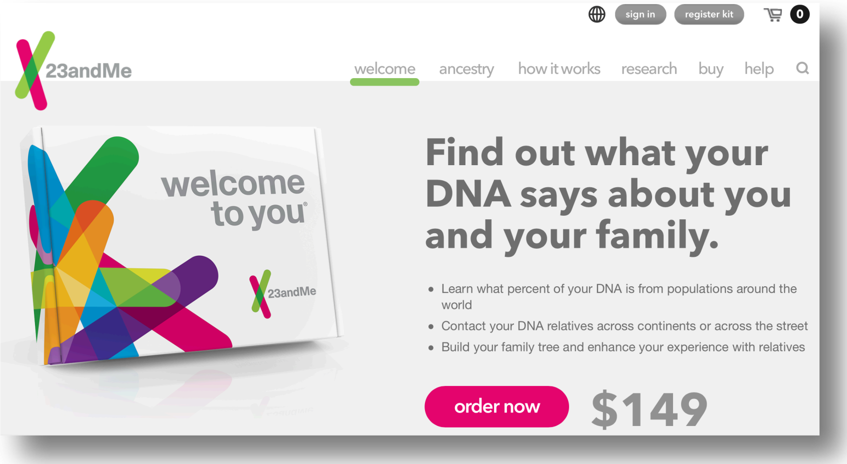
Big Data – Business perspective

Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, June 2011

It is a new business model

- People pay with data, e.g. Facebook, Google, Twitter:
  - use service, give data
  - Google sells your data to advertisers
  - you pay advertisers indirectly

- 23andMe, Amazon:
  - pay service + give data
  - sells data and
  - uses data to improve service



Find out what your DNA says about you and your family.

- Learn what percent of your DNA is from populations around the world
- Contact your DNA relatives across continents or across the street
- Build your family tree and enhance your experience with relatives

$149

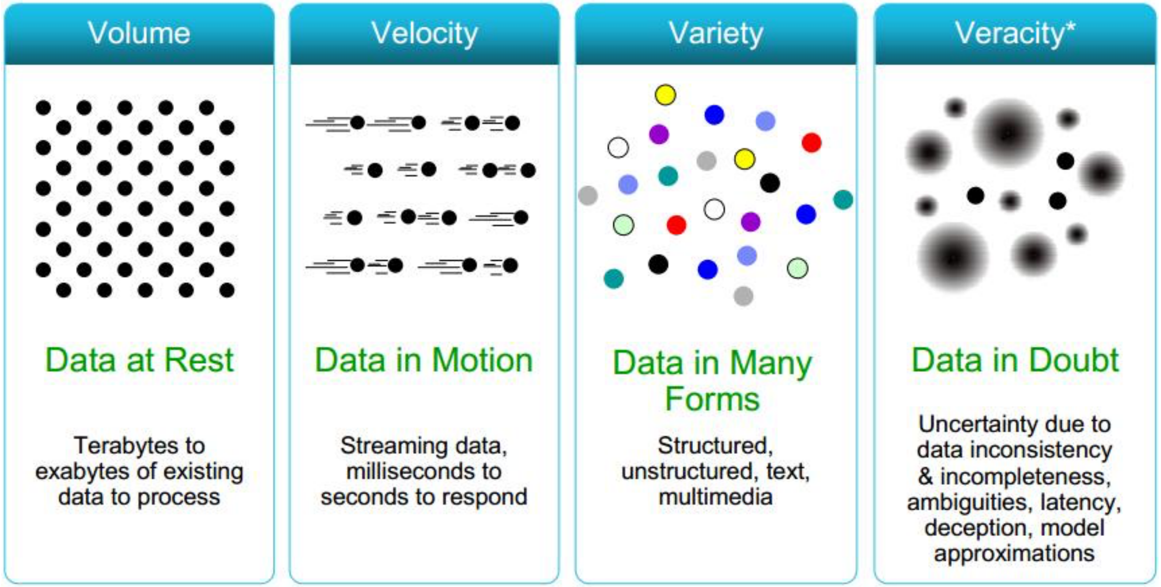
Big Data – Technical perspective

- You collect all data
  - the more the better -> statistical relevance,
  - keeping all is cheaper than deciding what to keep

- You decide independently what to do with data
  - run experiments on data when question arises

- Huge difference to traditional information systems
  - Design upfront what data to keep and why!!!
    (e.g., waterfall model of software engineering!)

Consequences


Volume
Data at Rest
Terabytes to exabytes of existing data to process
Velocity
Data in Motion
Streaming data, milliseconds to seconds to respond
Variety
Data in Many Forms
Structured, unstructured, text, multimedia
Veracity*
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

- **Volume**: data at rest
  - it is going to be a lot of data
- **Velocity** (Speed): data in motion
  - it is going to arrive fast
- **Variety** (Diversity): data in many formats
  - Different shapes (e.g., different versions, different sources)
- **Veracity**: data in doubt
  - do you know what you have?

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

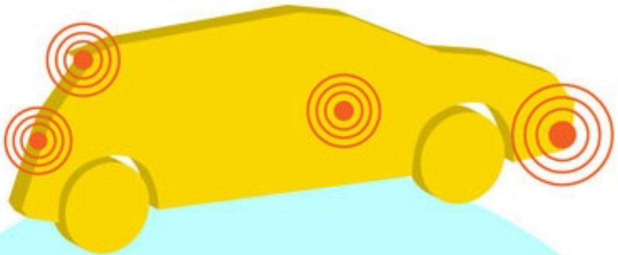
Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session



Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

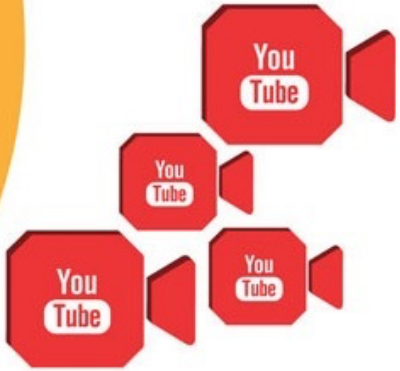
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

Variety

DIFFERENT FORMS OF DATA

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month



30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate