# Big Data Recap

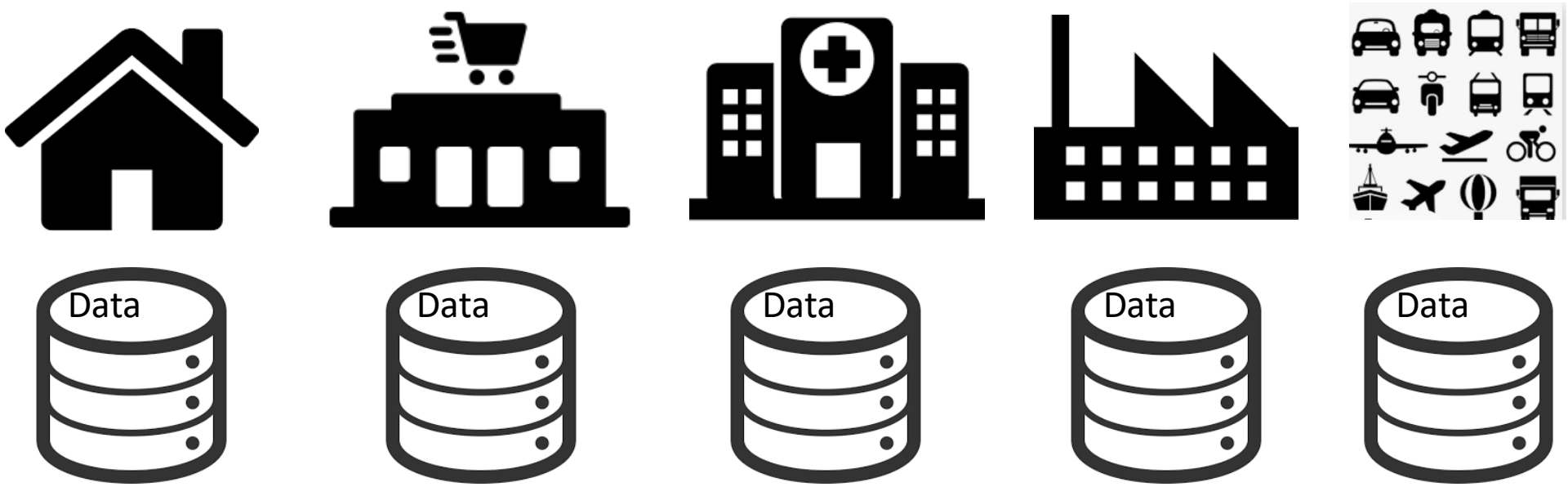
# Big data: ambiguous and relative concept

- 1663- John Graunt first to use statistical data analysis to deal with huge amounts of data in order to study bubonic plague
- 1800s- statistics include data collection and analysis
- 1880- US Census Bureau estimated 8 years to process data
- 1880- Herman Hollerith (US CB) created Hollerith tabulating machine (data processed in 3 months)
- 1927- Fritz Pfleumer invented magnetic tape data storage
- 1943- British decoding machine (5000 char / s)
- 1945- Von Neumann – EDVAC
- 1965- US gov. store millions of fingerprints and taxes
- 1989- Tim Berners-Lee create WWW
- 1999- IoT invented
- 2013- IoT generalized
- 2005- Roger Magoulas coins the **term big data** where BI soft could not be used
- 2005- creation of Hadoop (based on Nutch and merged with Google MapReduce)

# An era of Data

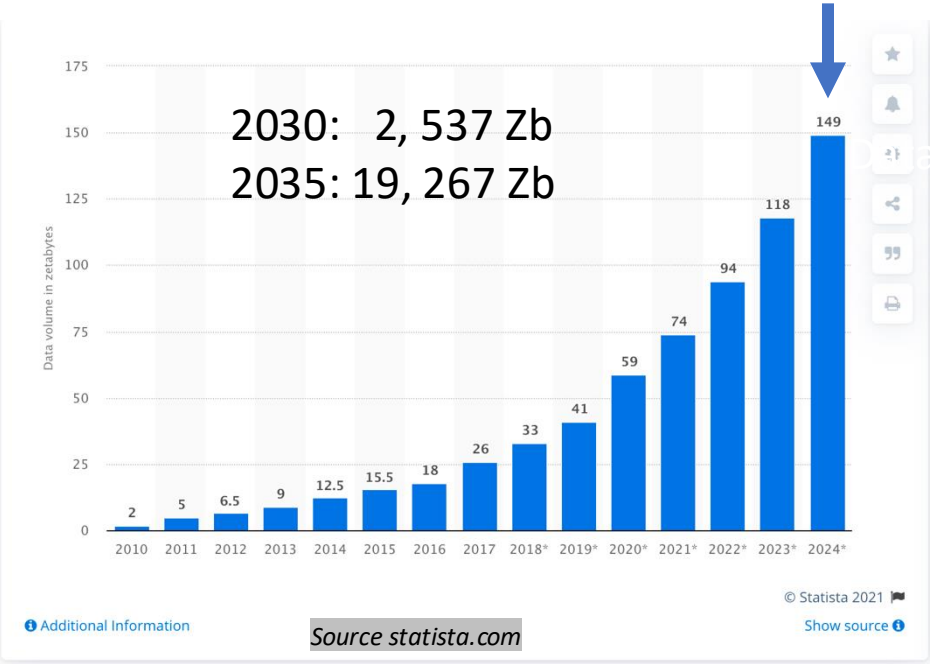
**2013: 90% of the world's data was created in the previous 2 years (source SINTEF)**

- **Mobile Sensors** and **trackers** popular and widely accepted
- **IoT** in all sectors retail, health, industry, tourism
- **Systematic data collection**



# Data production & value

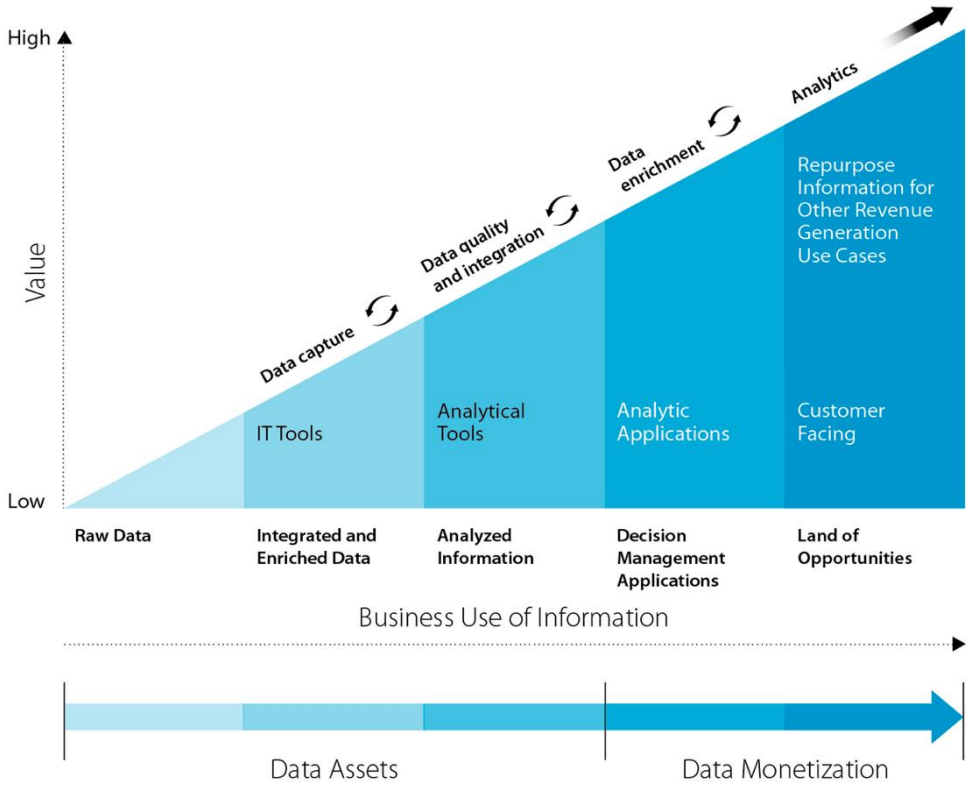
How much data is created every day?




*Source statista.com*

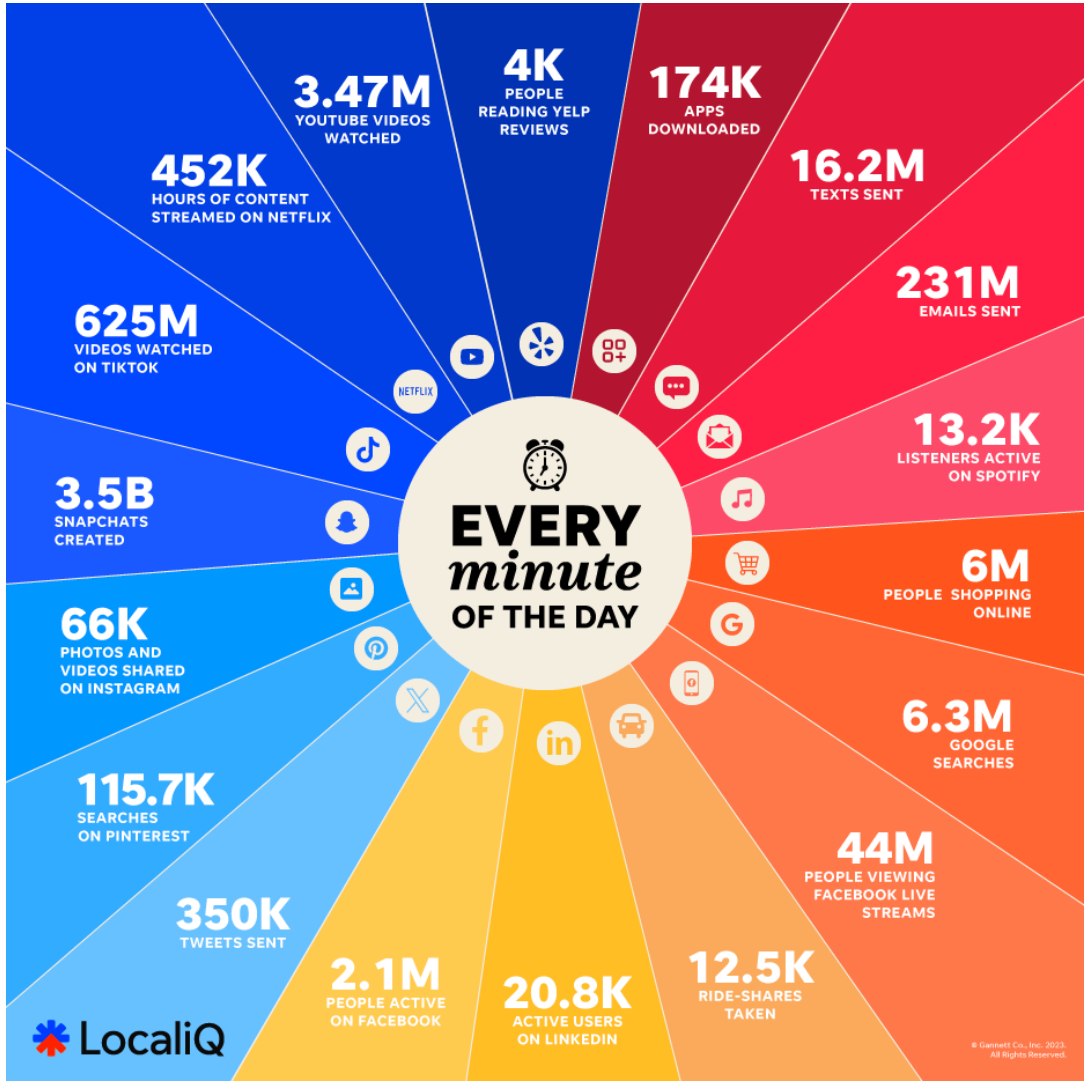
| Value | Metric |
|---|---|
| 1000 | kB kilobyte |
| $1000^2$ | MB megabyte |
| $1000^3$ | GB gigabyte |
| $1000^4$ | TB terabyte |
| $1000^5$ | PB petabyte |
| $1000^6$ | EB exabyte |
| $1000^7$ | ZB zettabyte |
| $1000^8$ | YB yottabyte |

Data has become a strategic asset with a huge economic value



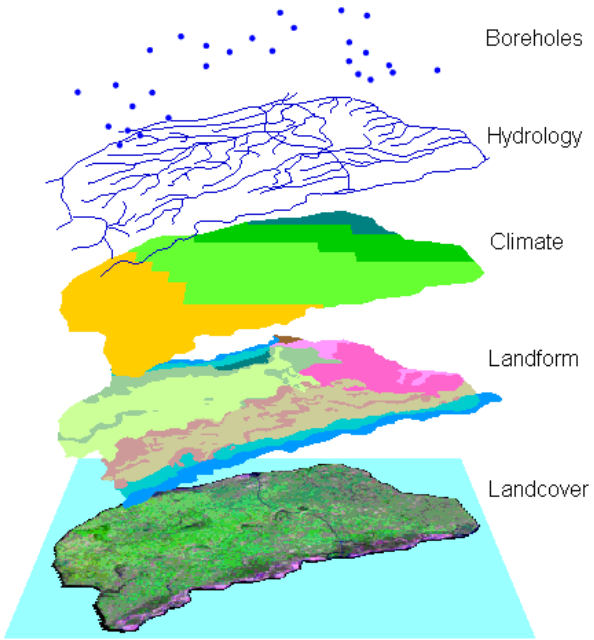

*Source idevnews.com*

# Media usage in 1 minute

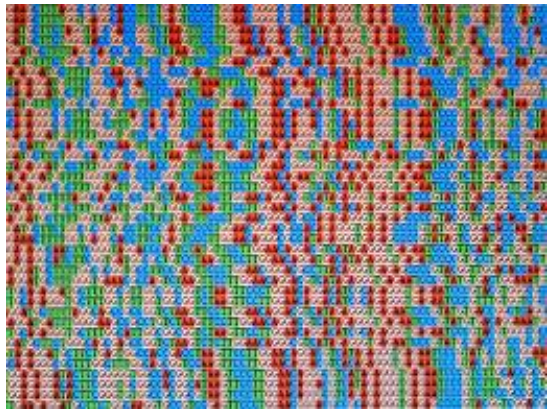


# Data driven research

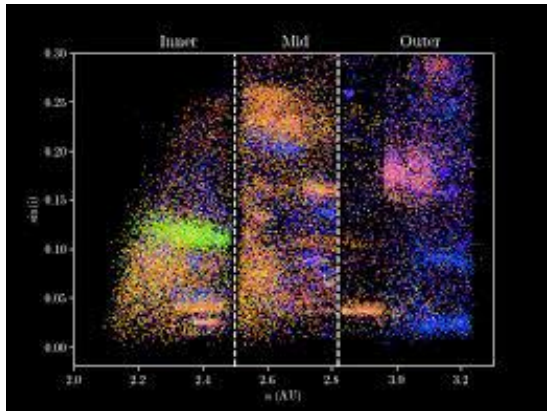
1. Experiment
2. Acquire data
3. Analyse data
4. Elaborate models/theories

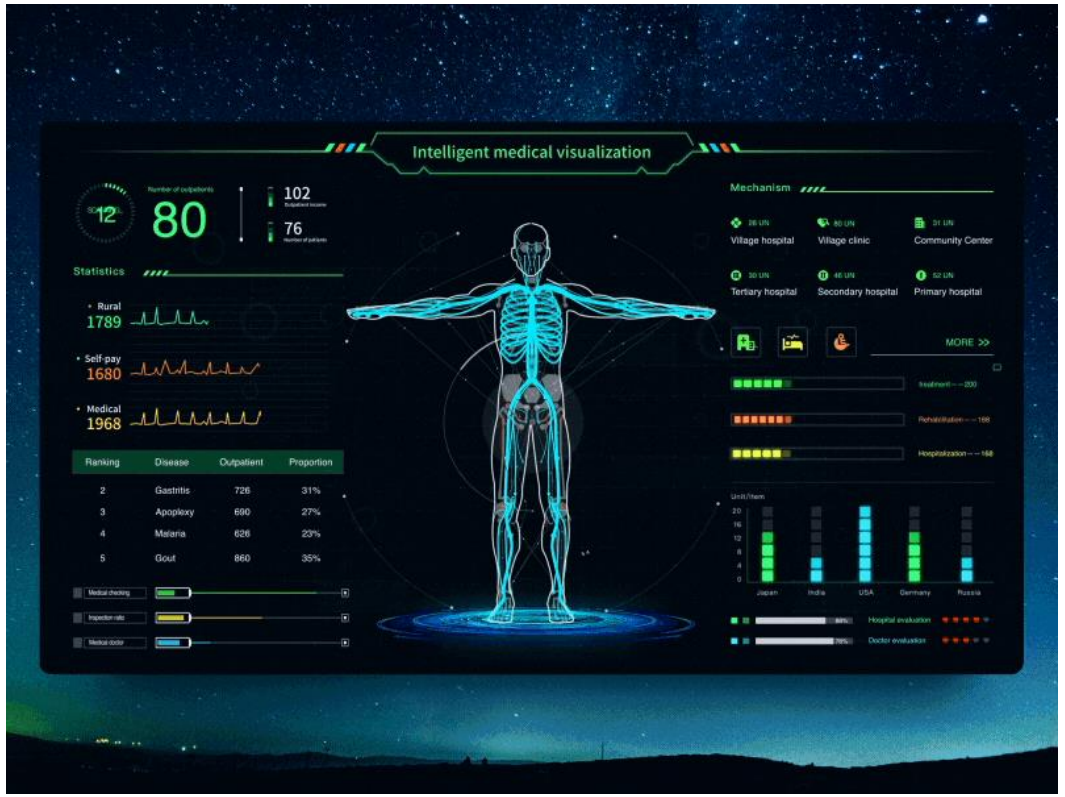



genes



astroML

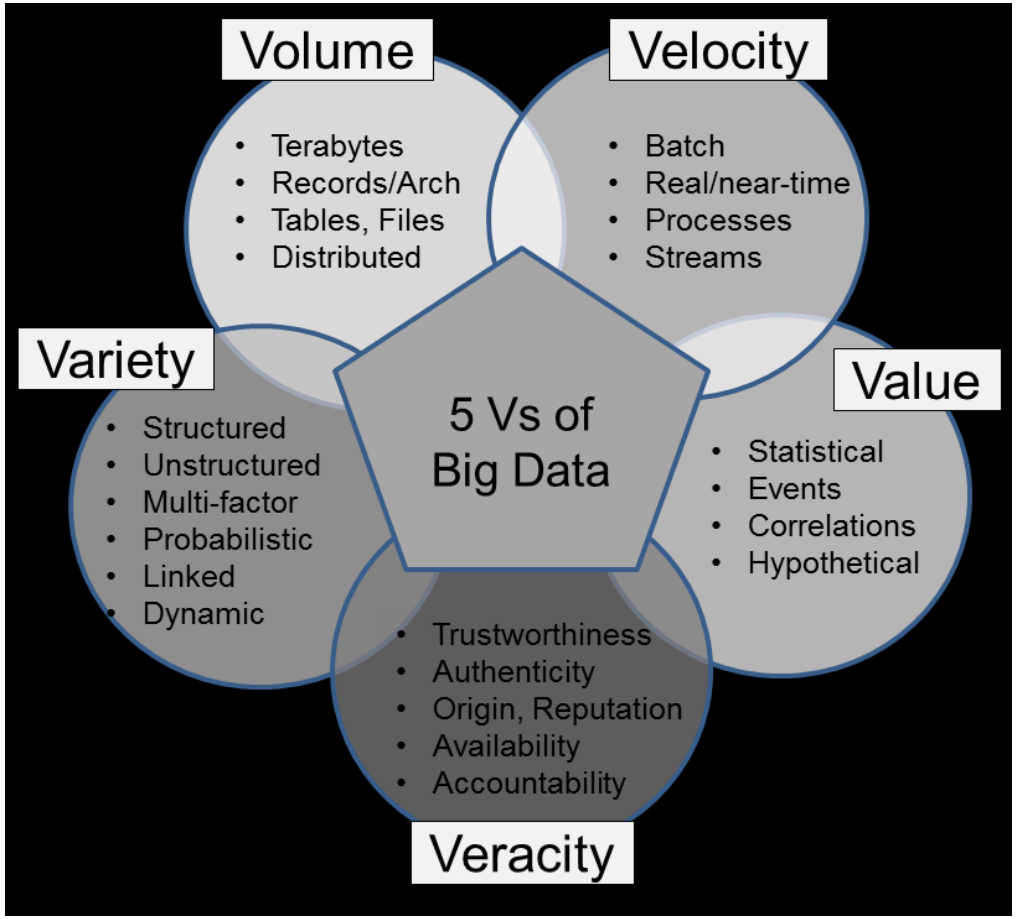


# Engineering sources

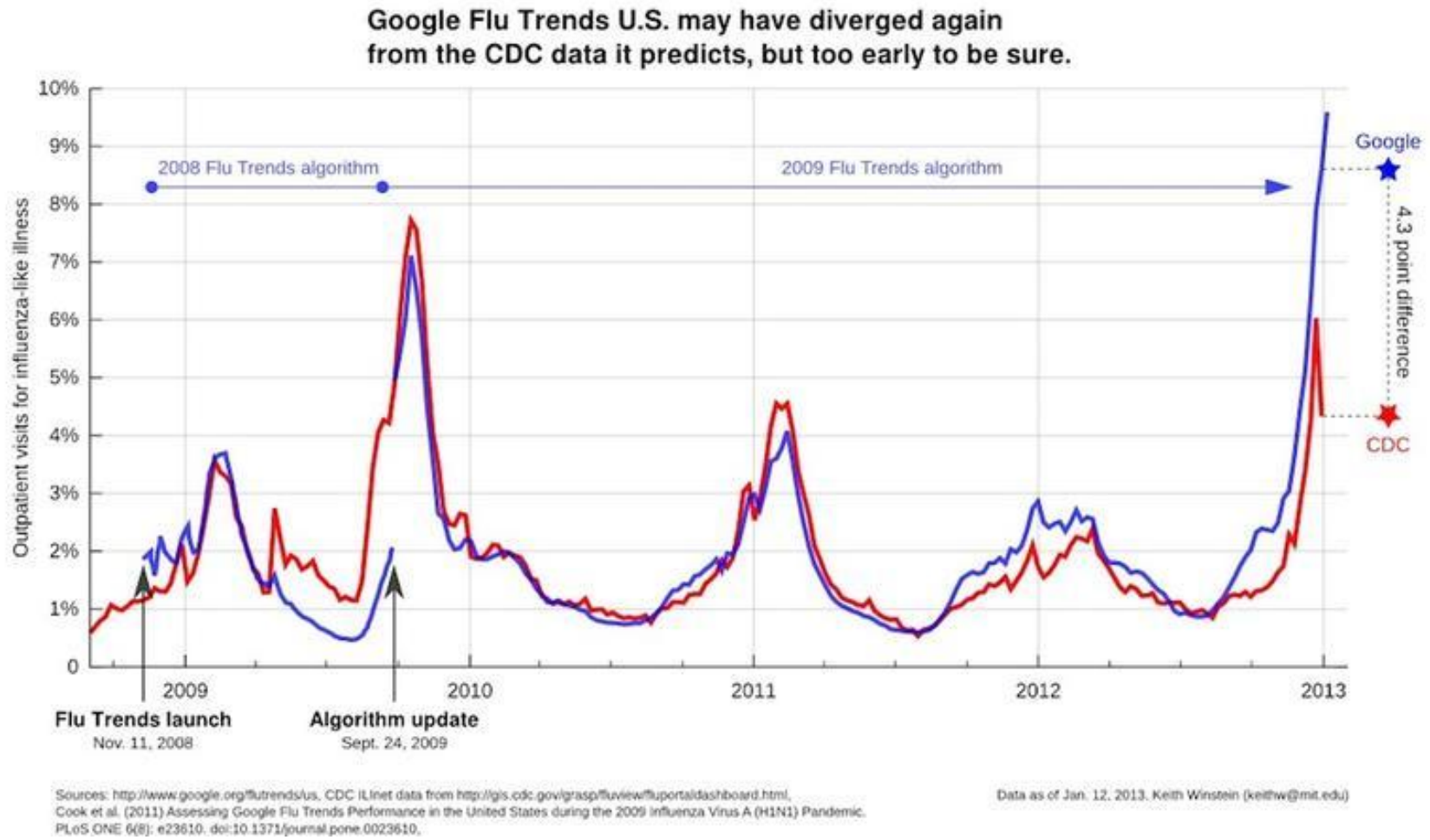
# Properties of big data

- **Volume**: large volume of data to process
- **Variety**: different formats
- **Velocity**: high rate of data accumulation
- Veracity: consistency, noise, uncertainty, completeness, timeliness of data
- Value: meaning, insights



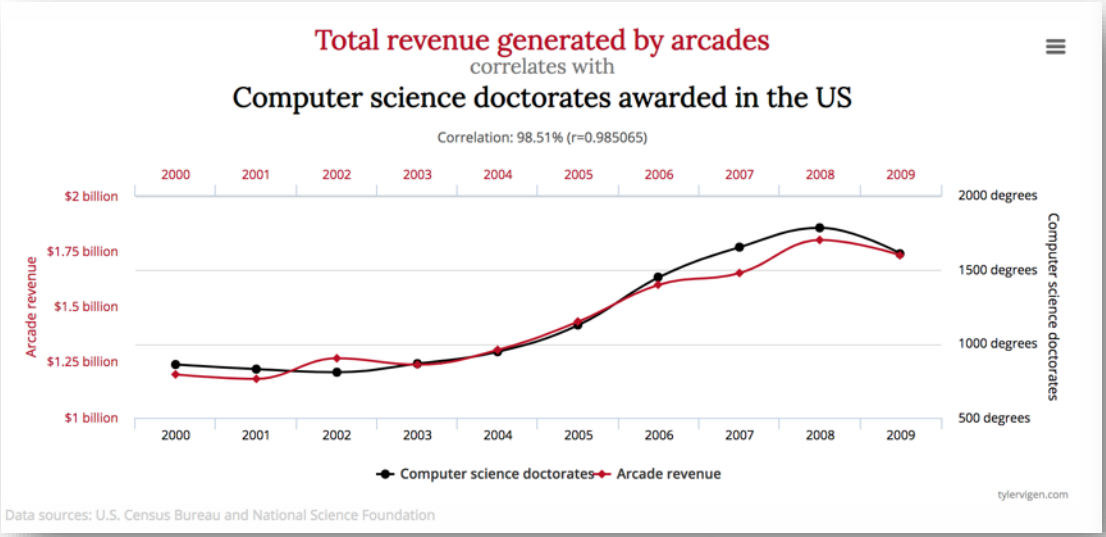
# Google Flu predictions

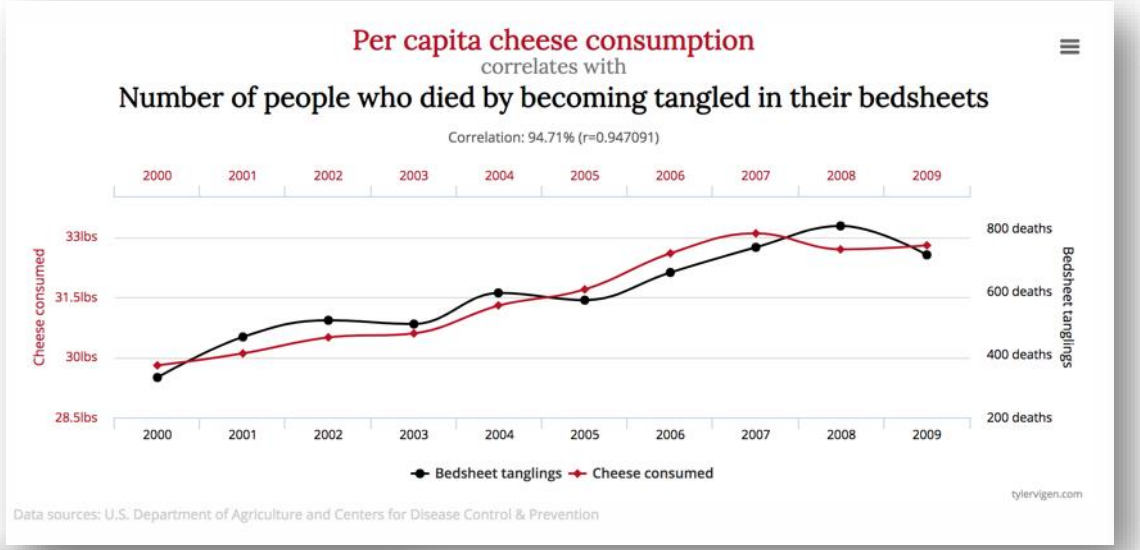
## large scale application of big data

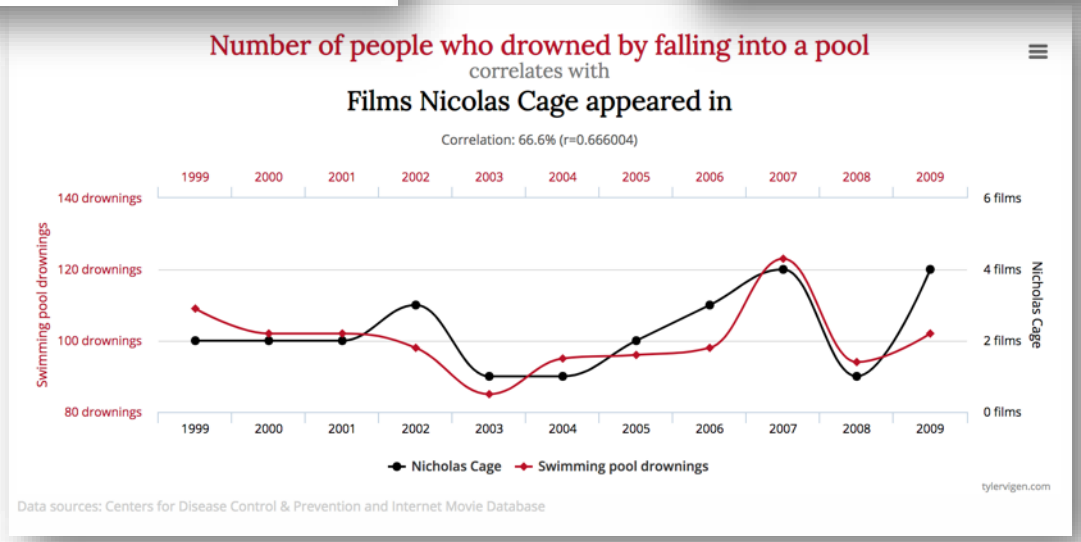


# Warning!

Correlation != Causation




http://www.tylervigen.com/spurious-correlations

# Big data challenges

- Requirement for
  - increased scalability: storage, querying, processing
  - Fast analytics
- Scalability
  - Vertical (scale up): more resources (CPU, memory…)
  - Horizontal (Scale out): multiple computers, distributed processing
- But how to distributed database storage?