

# Big Data Recap

# Big data: ambiguous and relative concept

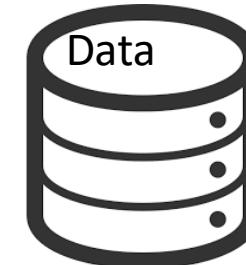
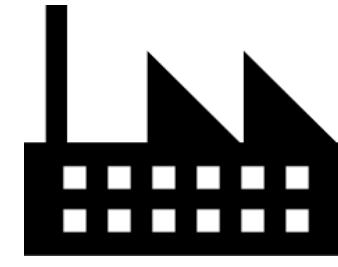
- 1663- John Graunt first to use statistical data analysis to deal with **huge amounts of data** in order to study bubonic plague
- 1800s- statistics include data collection and analysis
- 1880- US Census Bureau estimated **8 years to process data**
- 1880- Herman Hollerith (US CB) created Hollerith 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

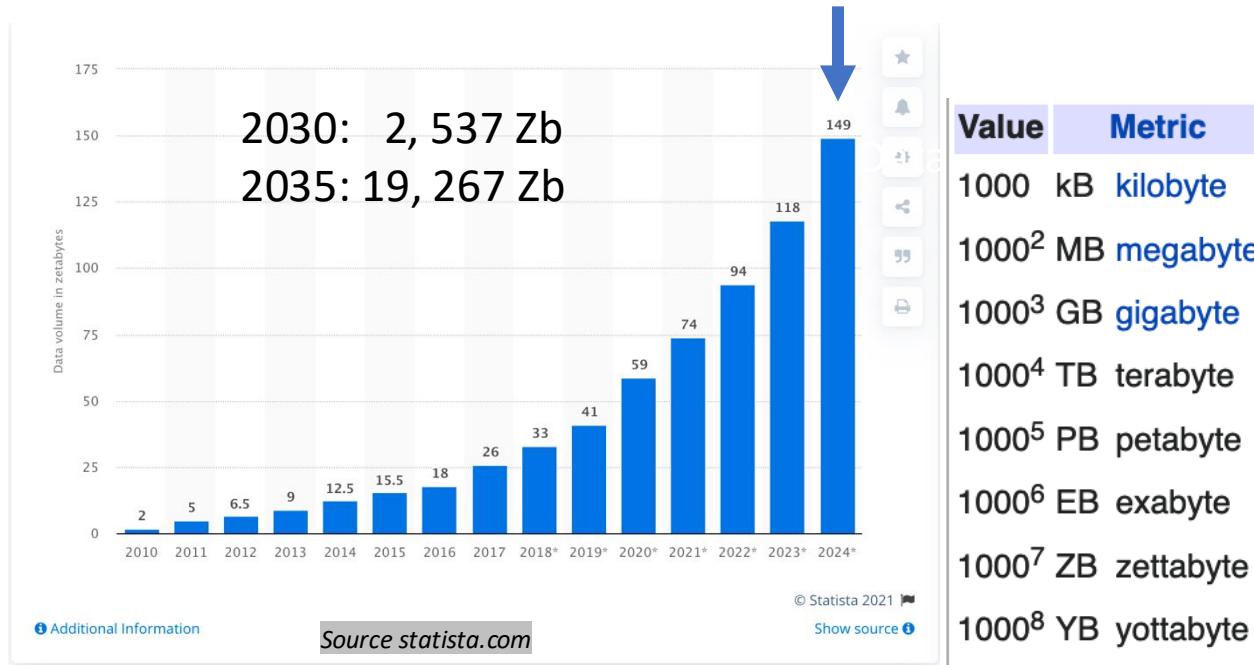
2013: 90% of the world's data was created in the previous 2 years (source SINTEF)

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

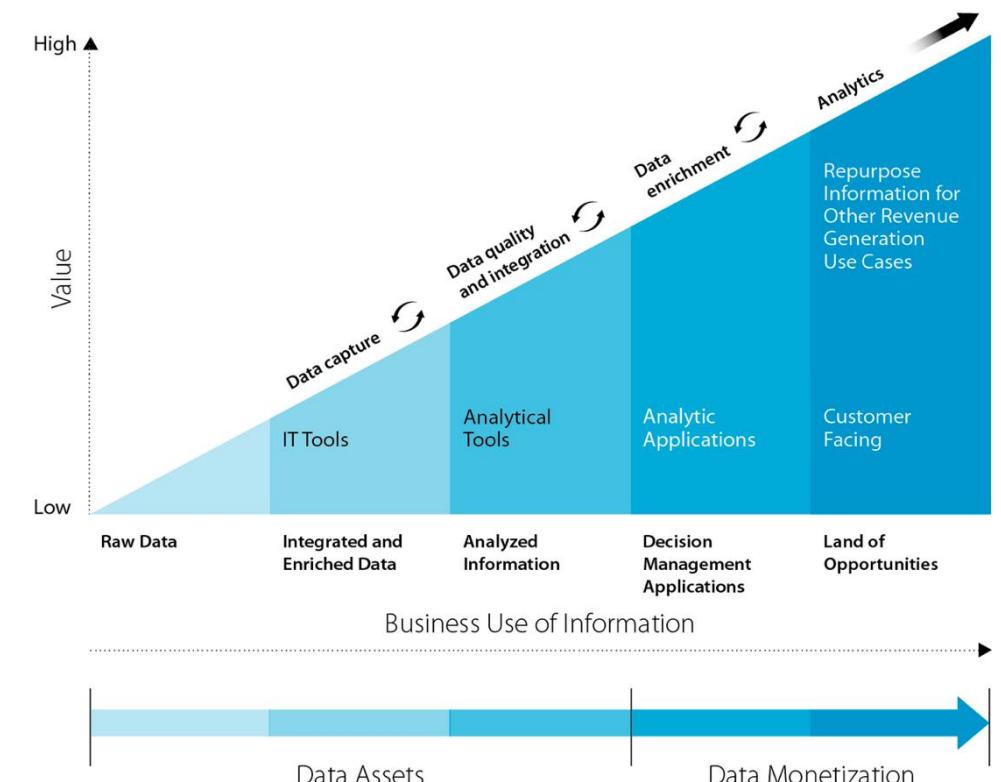


# Data production & value

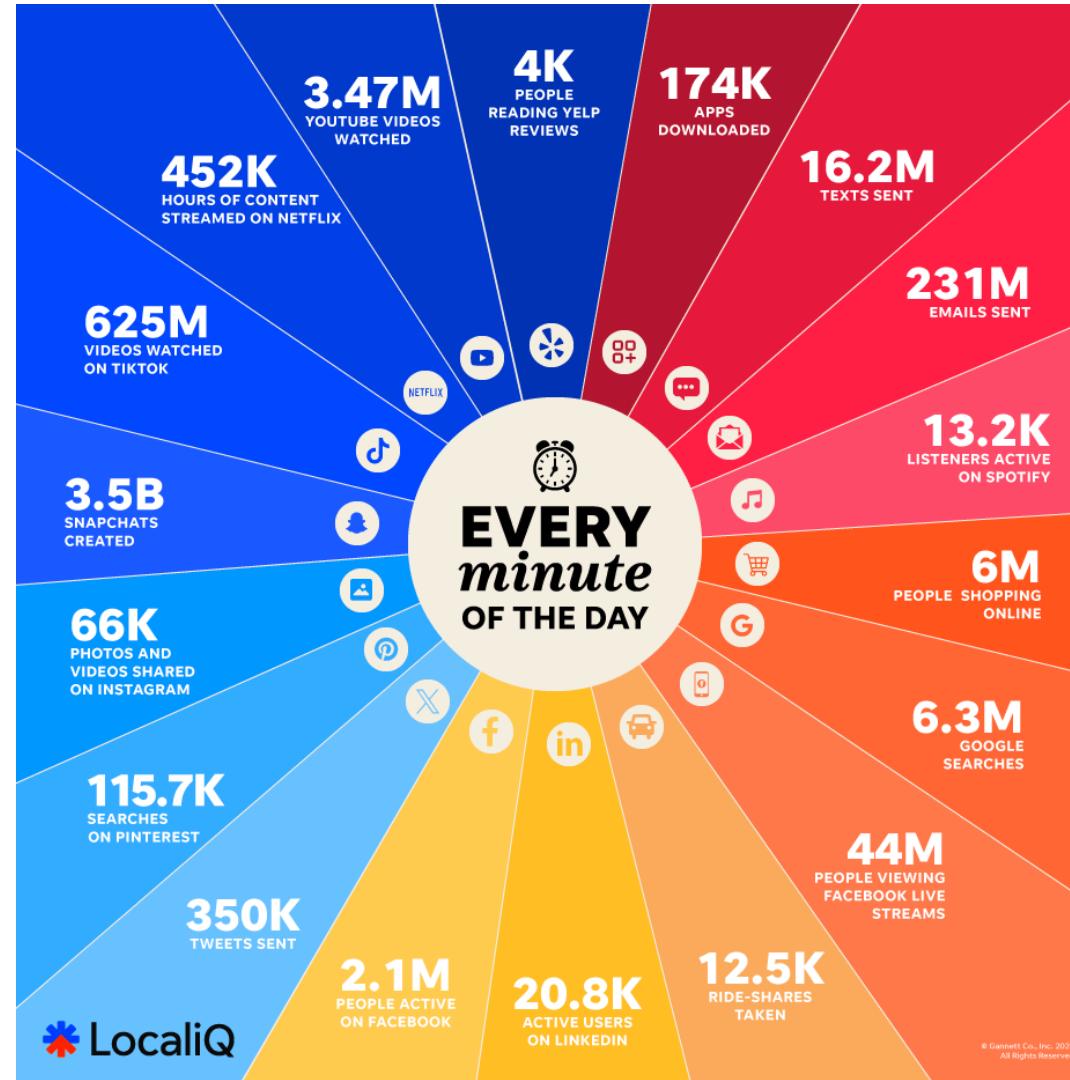
How much data is created every day?



Data has become a strategic asset with a huge economic value

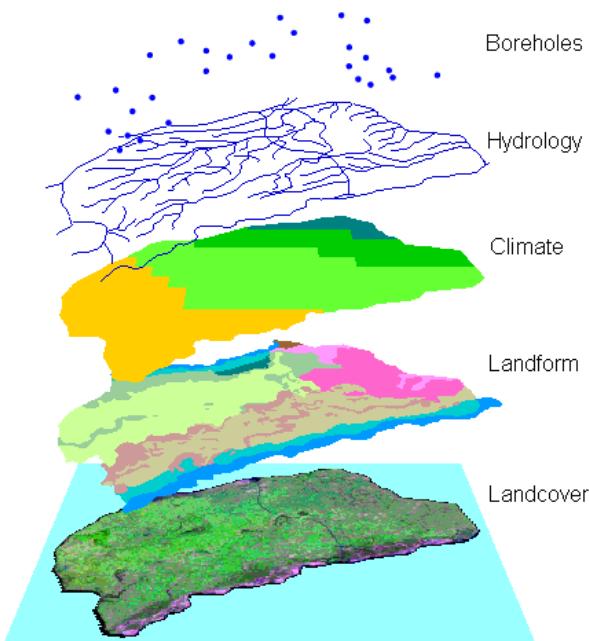


# Media usage in 1 minute

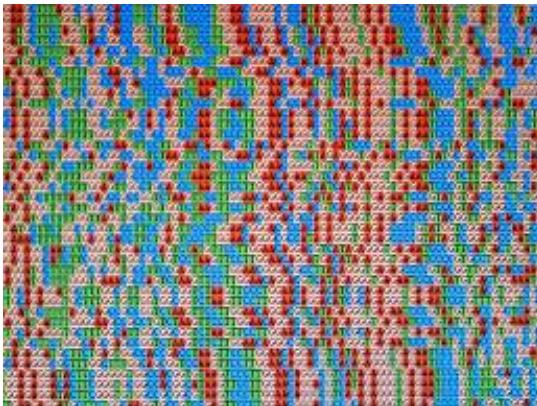


# Data driven research

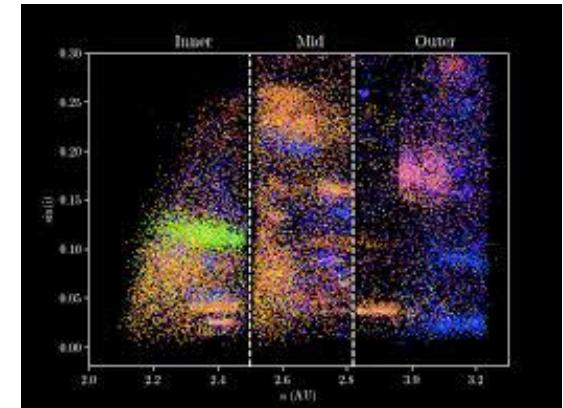
1. Experiment
  2. Acquire data
  3. Analyse data
  4. Elaborate models/theories



## genes



astroML



**Intelligent medical visualization**

**Statistics**

- Rural: 1789
- Self-pay: 1680
- Medical: 1968

**Number of consultations**: 80

**Reduced income**: 102

**Number of patients**: 76

**Mechanism**

- 36 LN Village hospital
- 40 LN Village clinic
- 21 LN Community Center
- 30 LN Tertiary hospital
- 46 LN Secondary hospital
- 52 LN Primary hospital

**Ranking Disease Outpatient Proportion**

Ranking	Disease	Outpatient	Proportion
2	Gastritis	726	31%
3	Apoplexy	680	27%
4	Malaria	626	23%
5	Gout	860	35%

**Medical checking**

**Inductor rate**

**Medical doctor**

**Unit/item**

**200M** India USA Germany Russia

**Hospital evaluation**

**Doctor evaluation**

A central feature is a 3D wireframe model of a human skeleton standing on a circular base.

# Engineering sources

**Connected Plane**  
40 TB per day (0.1% transmitted)

**Connected Factory**  
1 PB per day (0.2% transmitted)

**Public Safety**  
50 PB per day (<0.1% transmitted)

**Weather Sensors**  
10 MB per day (5% transmitted)



**Intelligent Building**  
275 GB per day (1% transmitted)

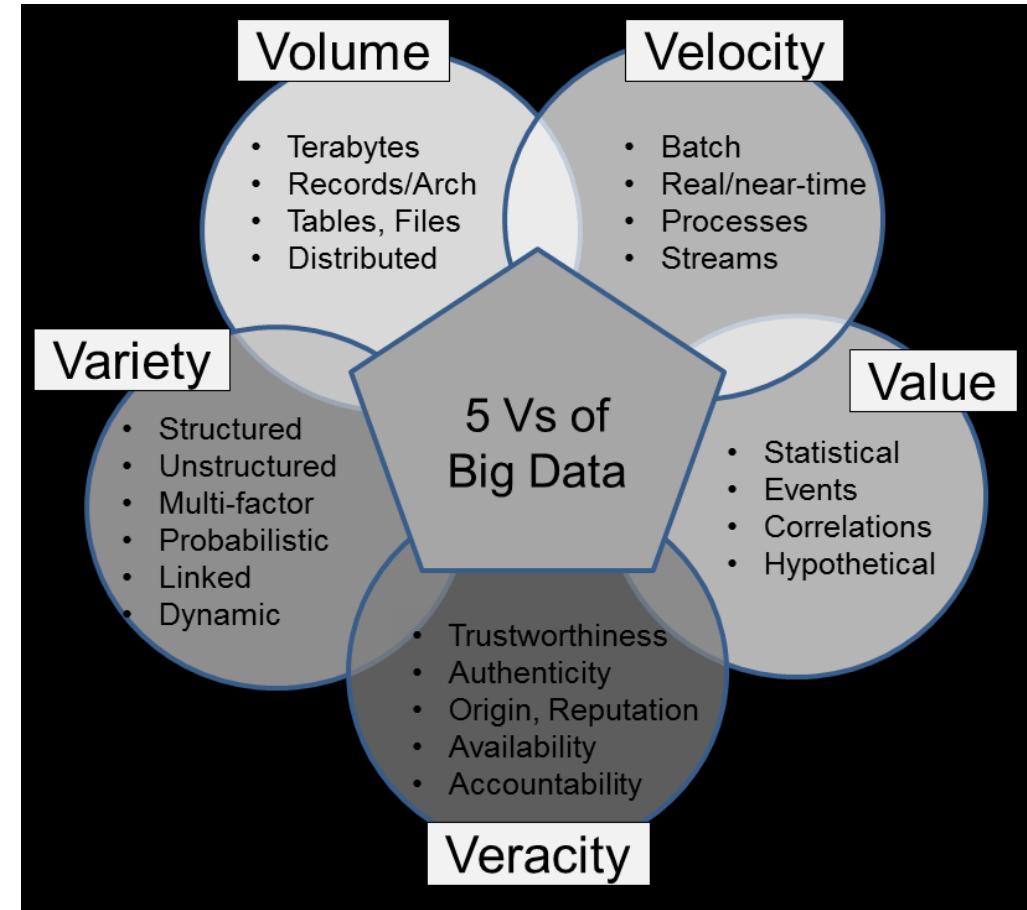
**Smart Hospital**  
5 TB per day (0.1% transmitted)

**Smart Car**  
70 GB per day (0.1% transmitted)

**Smart Grid**  
5 GB per day (1% transmitted)

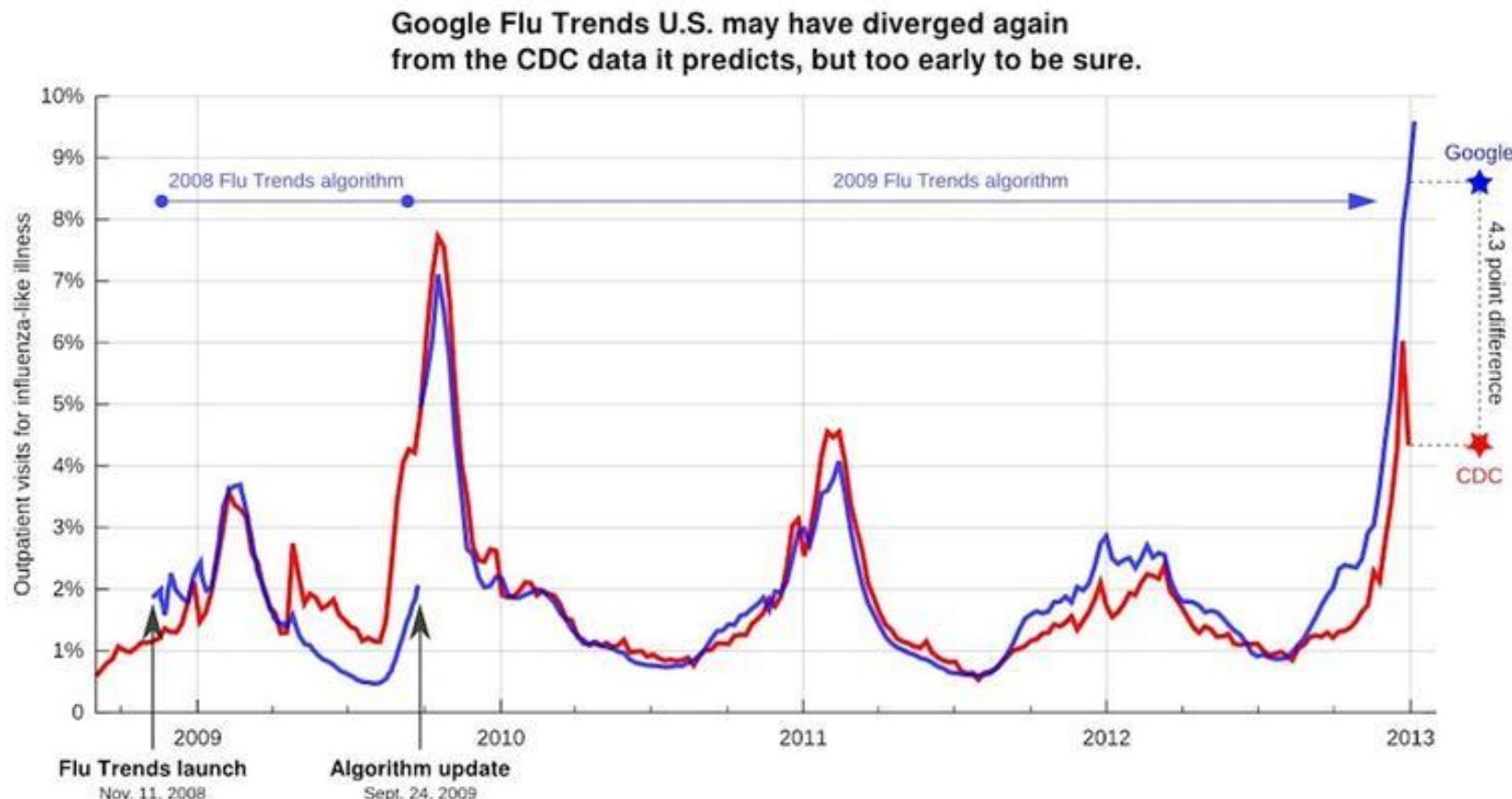
# Properties of big data

- **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

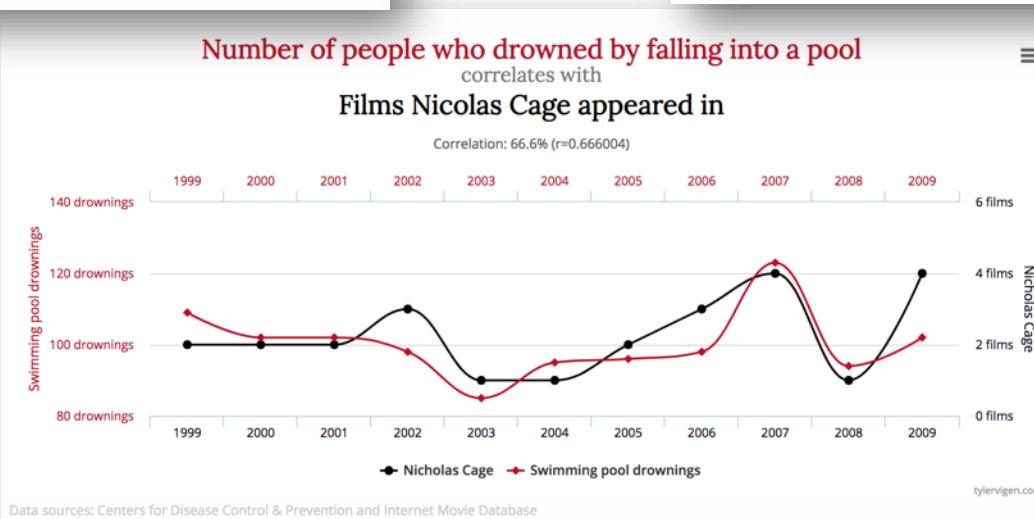
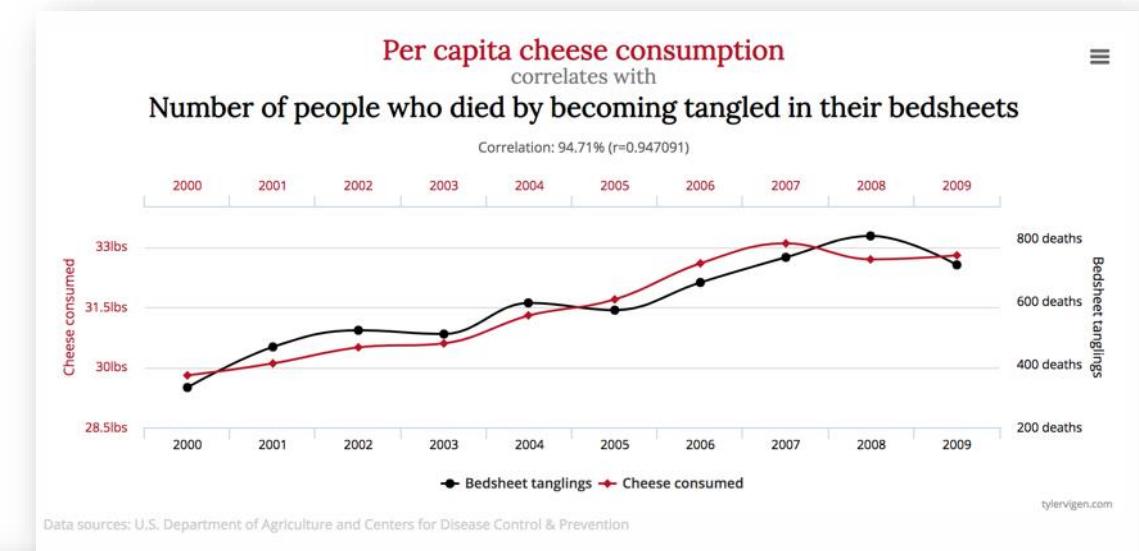
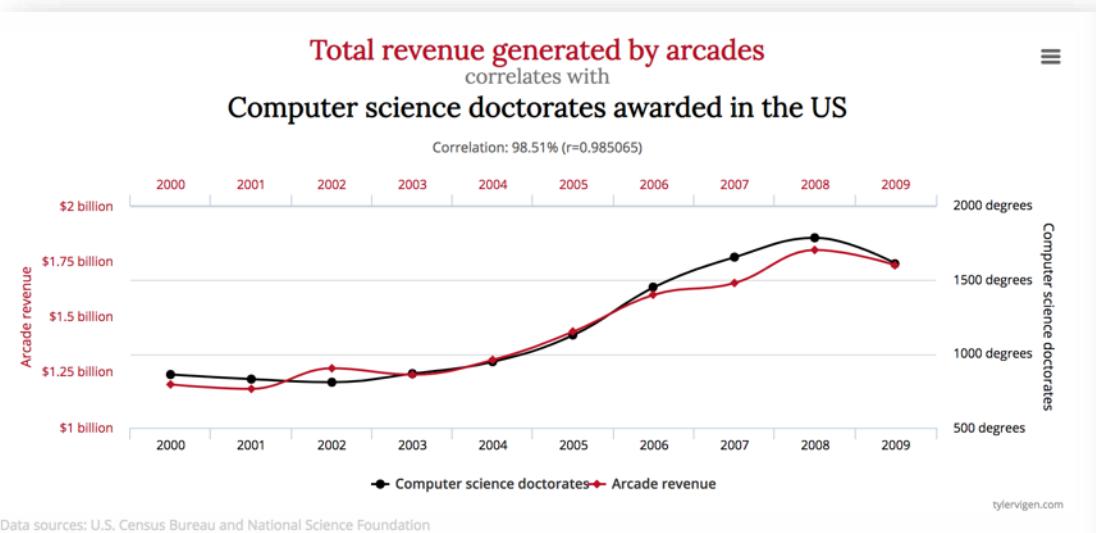


Sources: <http://www.google.org/flutrends/us>, CDC ILINet data from <http://gis.cdc.gov/grasp/fluvview/fluportal/dashboard.html>, Cook et al. (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. PLoS ONE 6(8): e23610. doi:10.1371/journal.pone.0023610.

Data as of Jan. 12, 2013. Keith Winstein (keithw@mit.edu)

# Warning!

## Correlation != Causation



# 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?

