哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

Technology and Application of Big Data

Qing LIAO(廖清)

School of Computer Science and Technology

HIT

Course Details

- Instructor:
  - Qing LIAO, liaoqing@hit.edu.cn
  - Rm. 303B, Building C
  - Office hours: by appointment
- Course web site:
  - liaoqing.me
- Reference books/materials:
  - Big data courses from University of California
  - Book: BIG DATA: A Revolution That Will Transform How We Live, Work, and Think
  - Papers
- Grading Scheme:
  - Paper Report 30%
  - Final Exam 70%

What You Learnt: Overview

- Topics:
  1) Introduction of Big Data
  2) Characterizes of Big Data
  3) How to Get Value from Big Data
  4) Technologies of Big Data
  5) Applications of Big Data
- Prerequisites
  - Statistics and Probability would help
    - But not necessary
  - Machine Learning would help
    - But not necessary

Previous Section

- What Launched the Big Data?

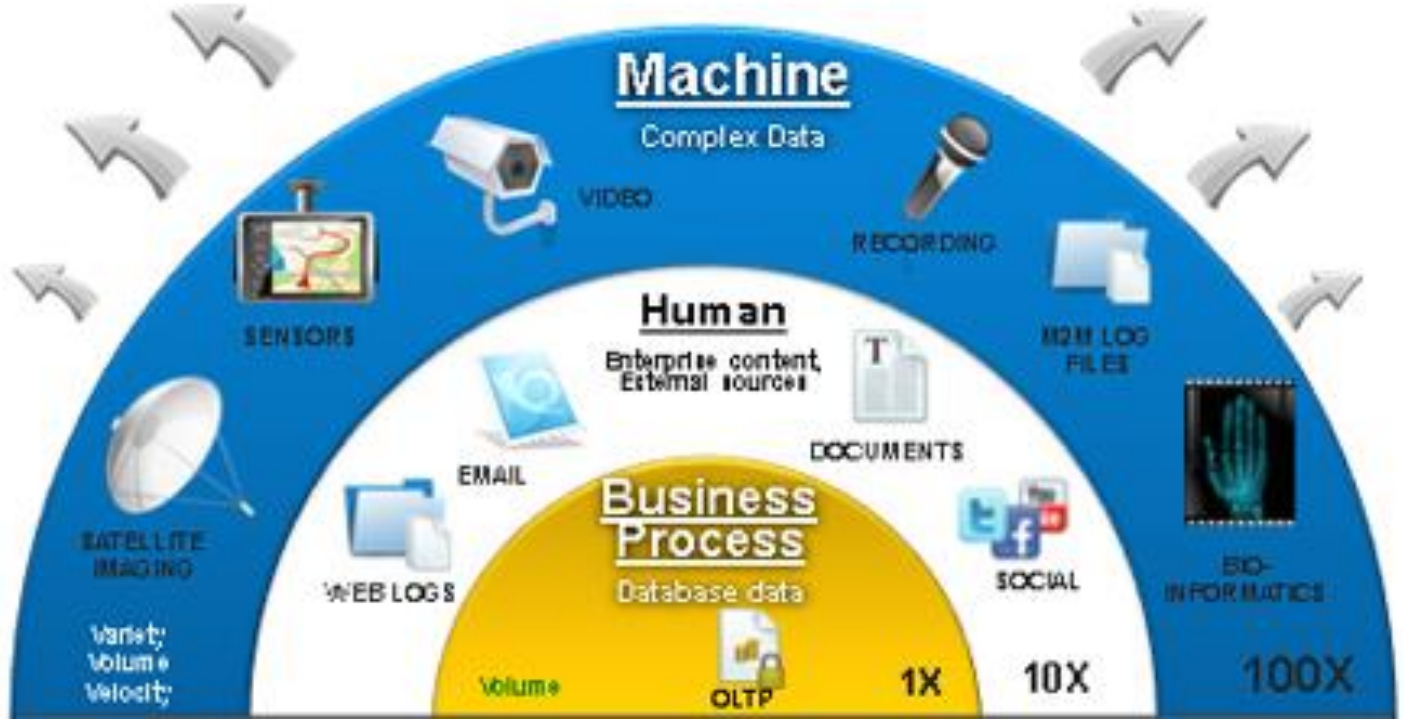
First Opportunity: Data Torrent

Second Opportunity:
Computing Anytime, Anywhere

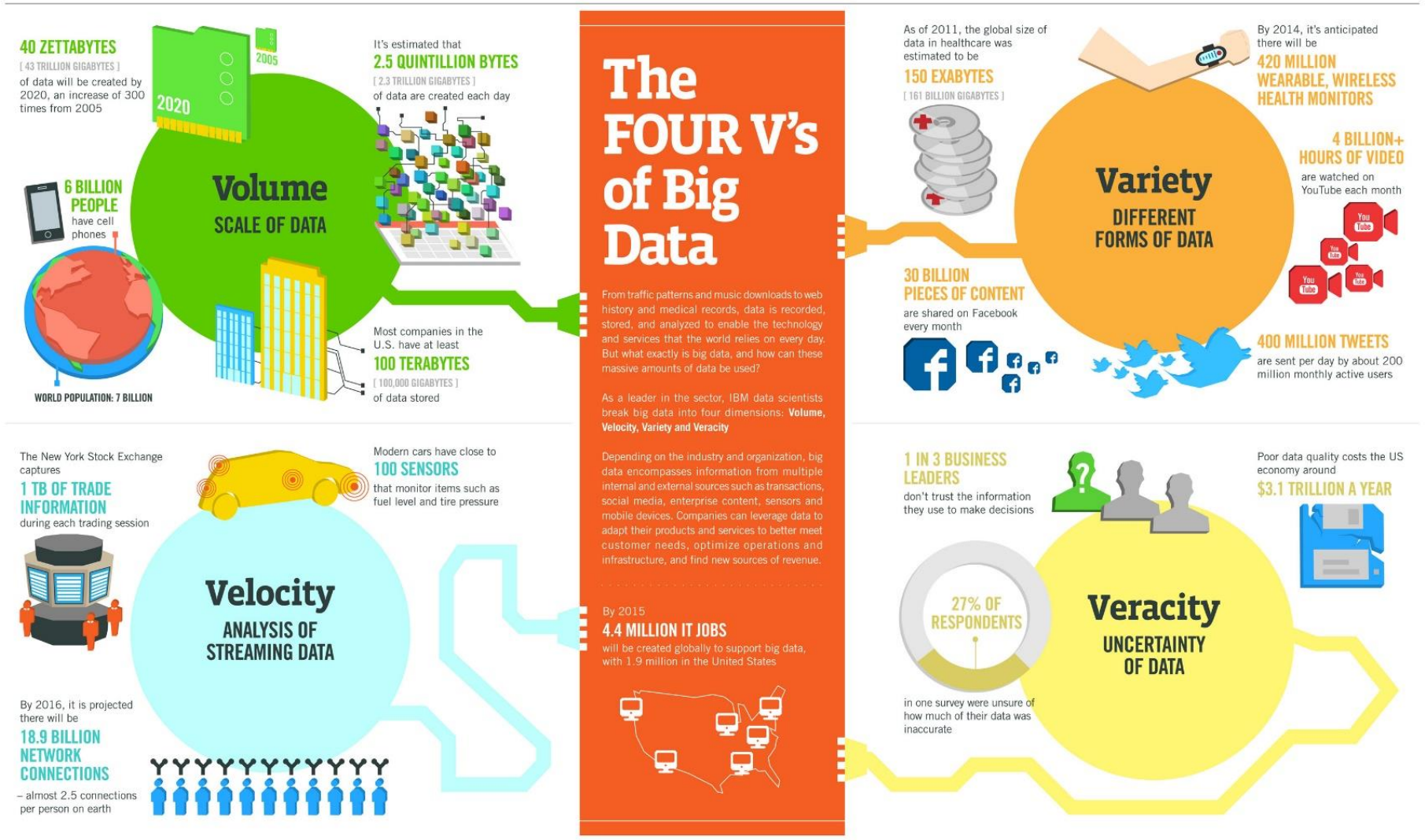
Launched the Big Data Era

Previous Section

- Where Does Big Data Come From?

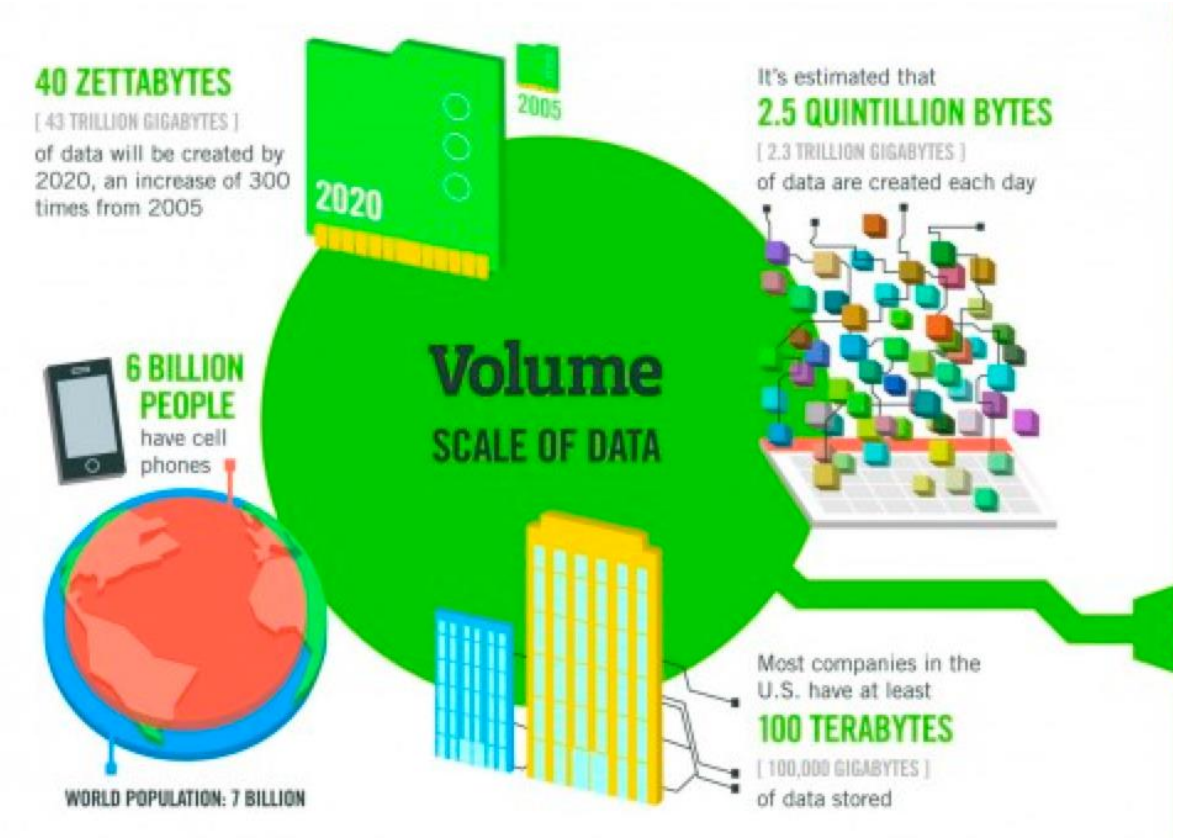


Characterizes of Big Data



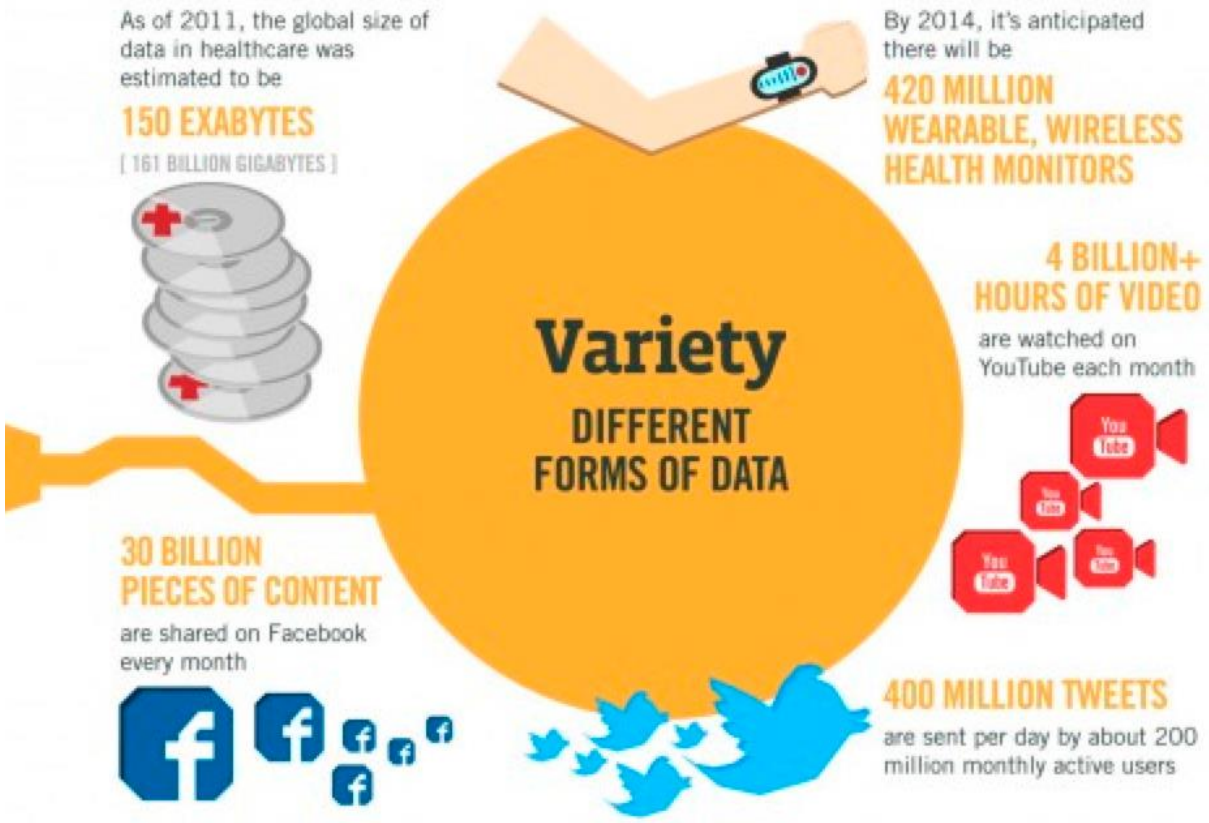
Characterizes of Big Data

- This refers to the vast amounts of data that is generated every second/minute/hour/day in our digitized world.



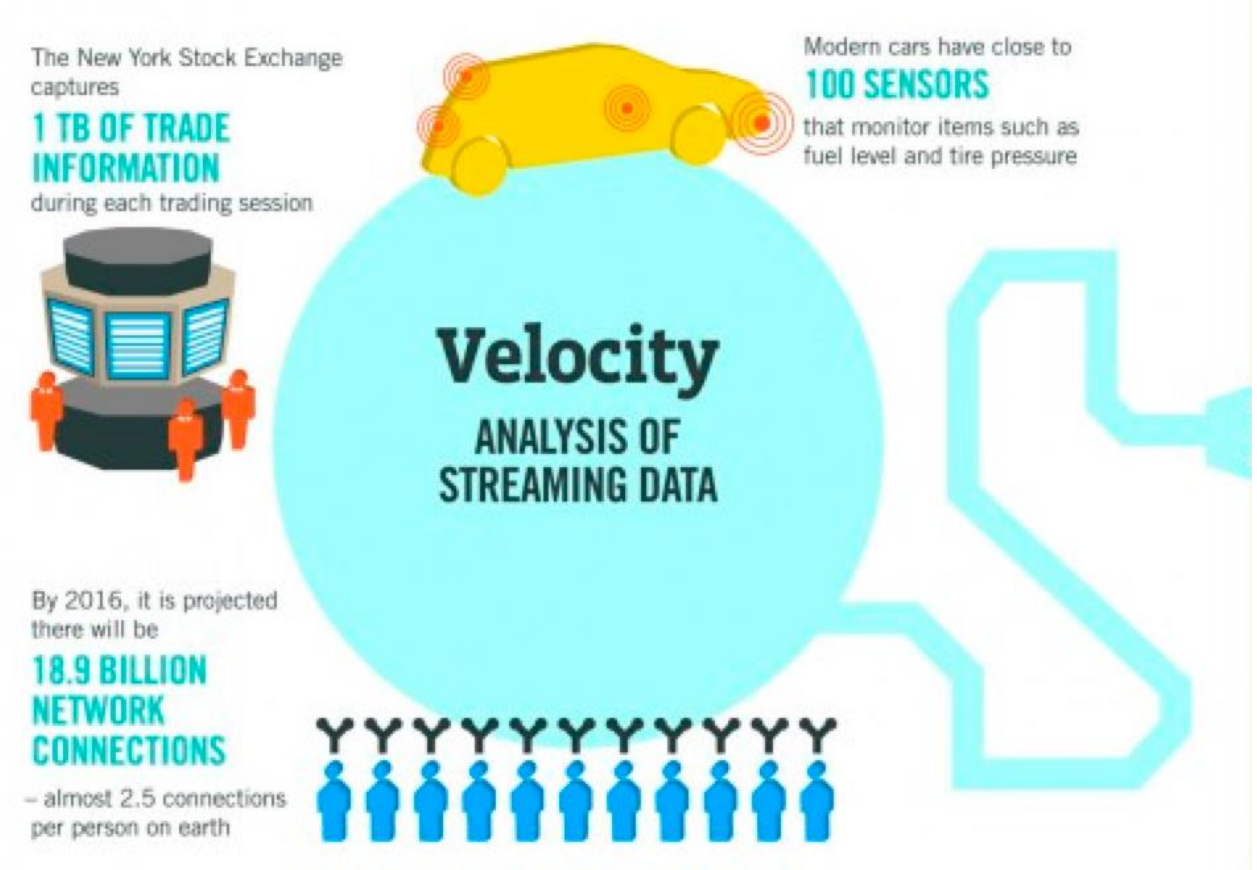
Characterizes of Big Data

- This refers to the ever-increasing different forms that data can come in, e.g., text, images, voice, geospatial.



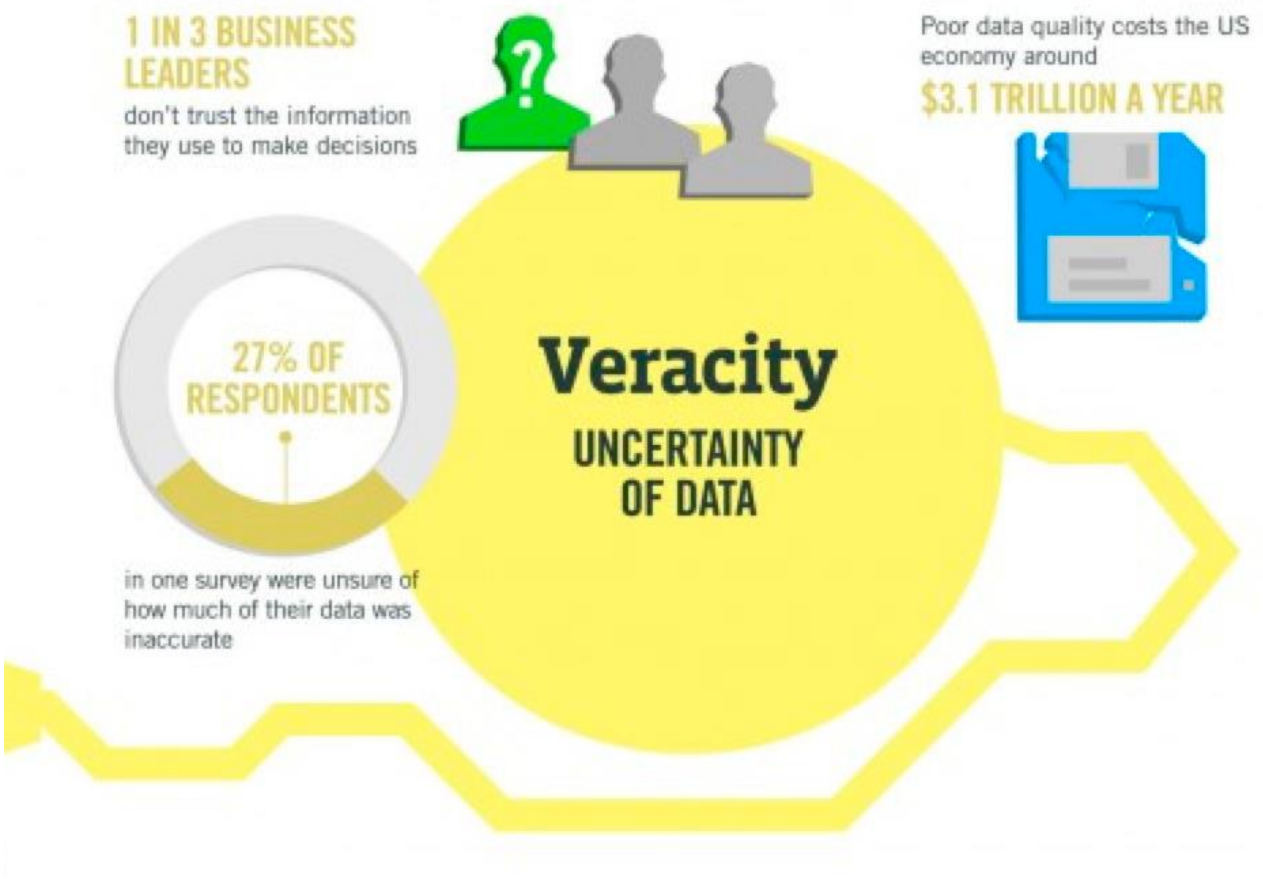
Characterizes of Big Data

- This refers to the speed at which data is being generated and the pace at which data moves from one point to the next.



Characterizes of Big Data

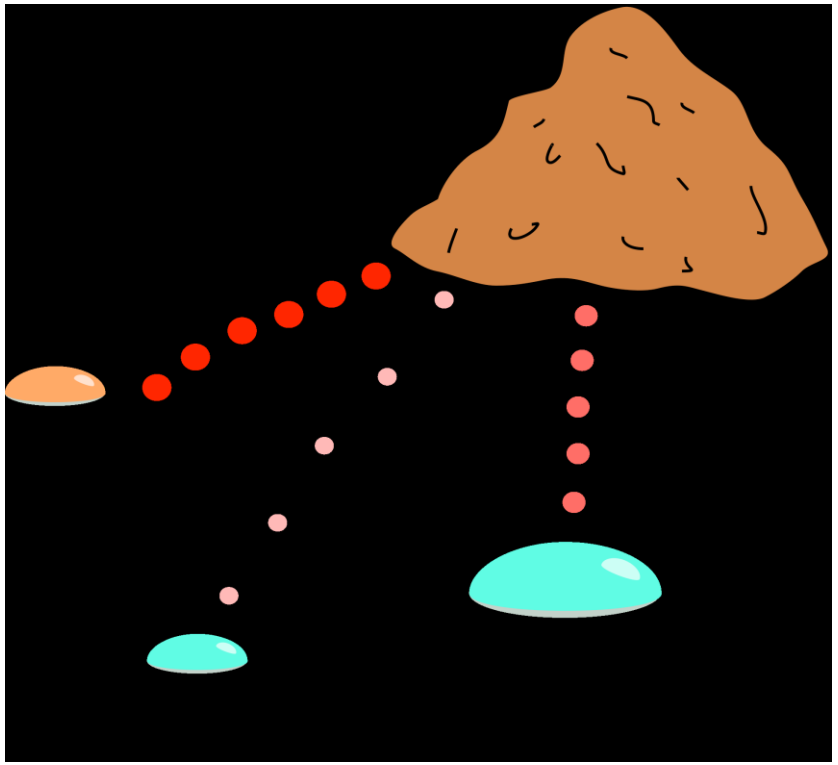
- This refers to the quality of the data, which can vary greatly.



Characterizes of Big Data

- Volume

Volume = Size



Characterizes of Big Data

- Volume

  Every minute…

204 Million emails

facebook

200,000 photos

1.8 Million likes

YouTube

1.3 Million video views

72 hours of video uploads

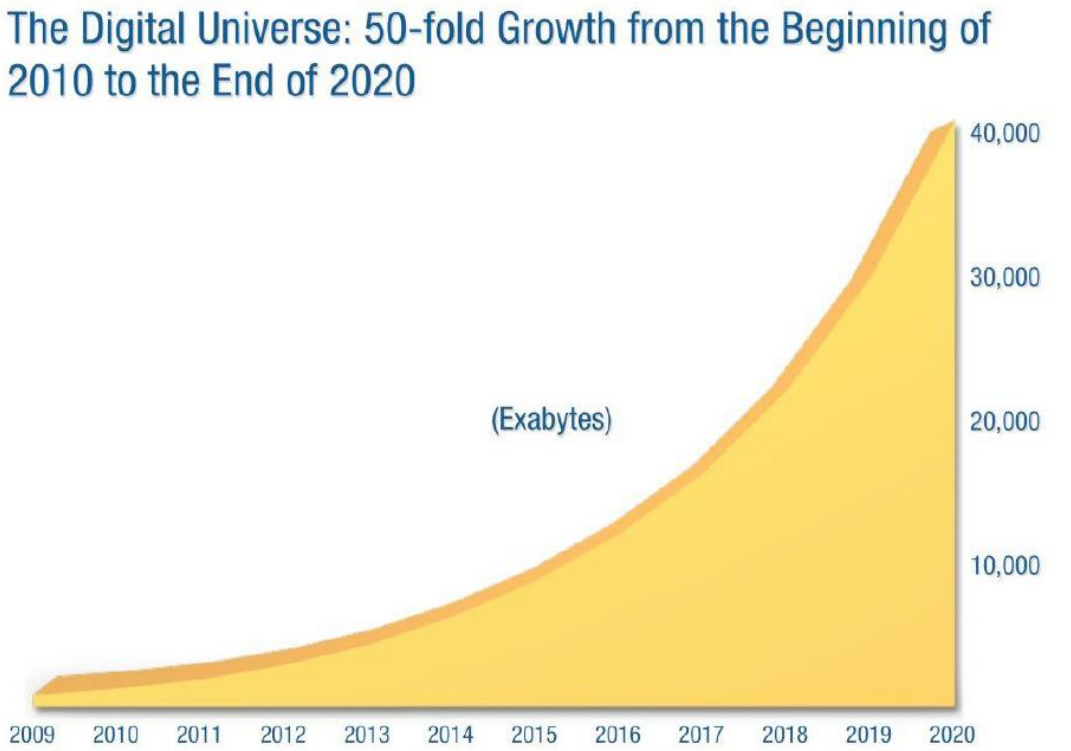
100 MBs ~= couple of volumes of Encyclopedias

A DVD ~= 5 GBs

1 TB ~= 300 hours of good quality video

Characterizes of Big Data

- Volume



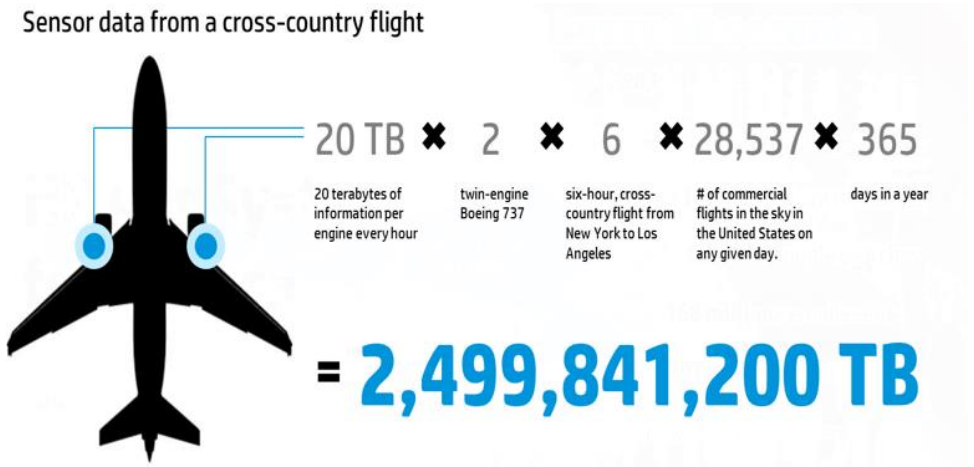
This IDC graph predicts exponential growth of data from around 3 zettabytes in 2013 to approximately 40 zettabytes by 2020. An exabyte equals 1,000,000,000,000,000,000 bytes and 1,000 exabytes equals one zettabyte. Source: IDC's Digital Universe Study, December 2012, http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

Exponential data growth!

Characterizes of Big Data

- Volume

More data = Better safety

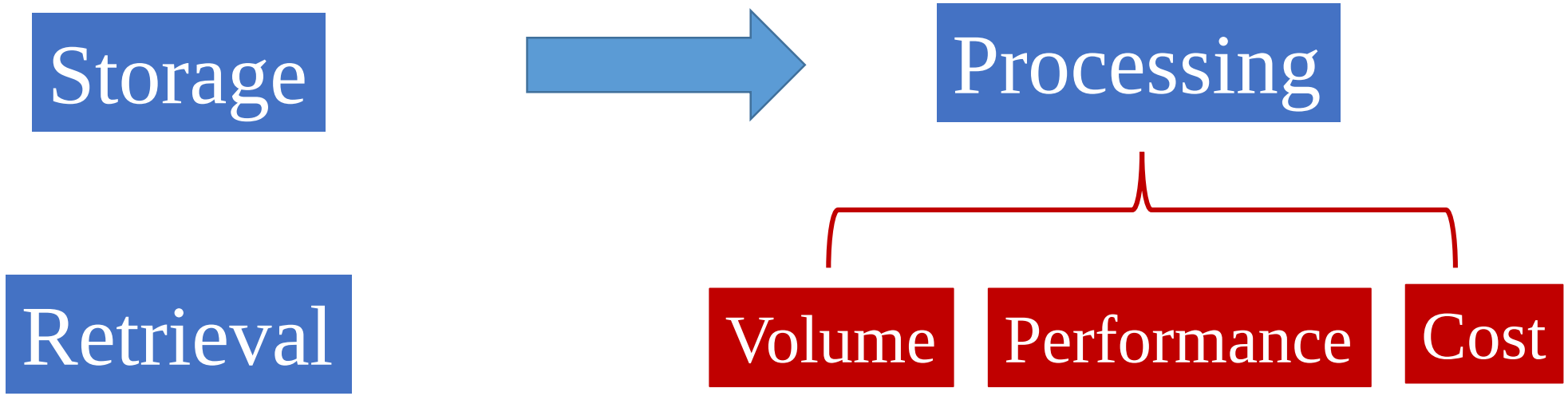


Volume

Business Insight

Characterizes of Big Data

- Volume：Challenges

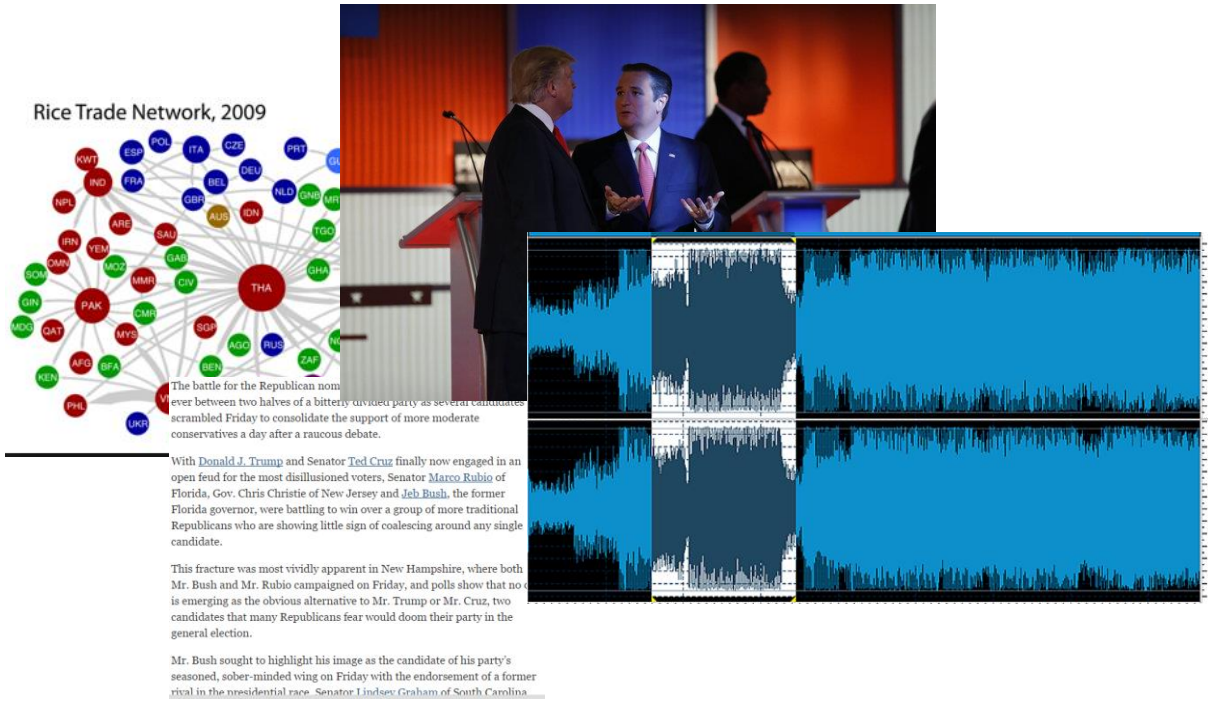


Characterizes of Big Data

- Variety

Variety = Complexity

Cars marketplace				
vendor	Model	Price	Mileage	VIN Code
Chevrolet	Corvette	17226	25965.0	ILLAKAWAZDZ
Chevrolet	Corvette	34229	46429.0	RCPNSRYGXO
Chevrolet	Corvette	27982	50209.0	NWLGCEVEHG
Chevrolet	Corvette	51825	72998.0	NGVZSCIZGSM
Chevrolet	Corvette	52845	34364.0	PSDRUYYOIJG
Chevrolet	Malibu	37874	37273.0	VLFPQPWNEFD
Chevrolet	Malibu	15600	71441.0	EXLJGDWOZSA
Chevrolet	Malibu	52447	46700.0	NLMGJZAKBRD
Chevrolet	Malibu	27129	36254.0	OIPFUIENLEHSX
Chevrolet	Malibu	28846	77162.0	WRCOOFREZLL
Chevrolet	Malibu	46165	60590.0	HUFTTHQHSFJ
Chevrolet	Malibu	18263	37790.0	JLMHNAESHVD

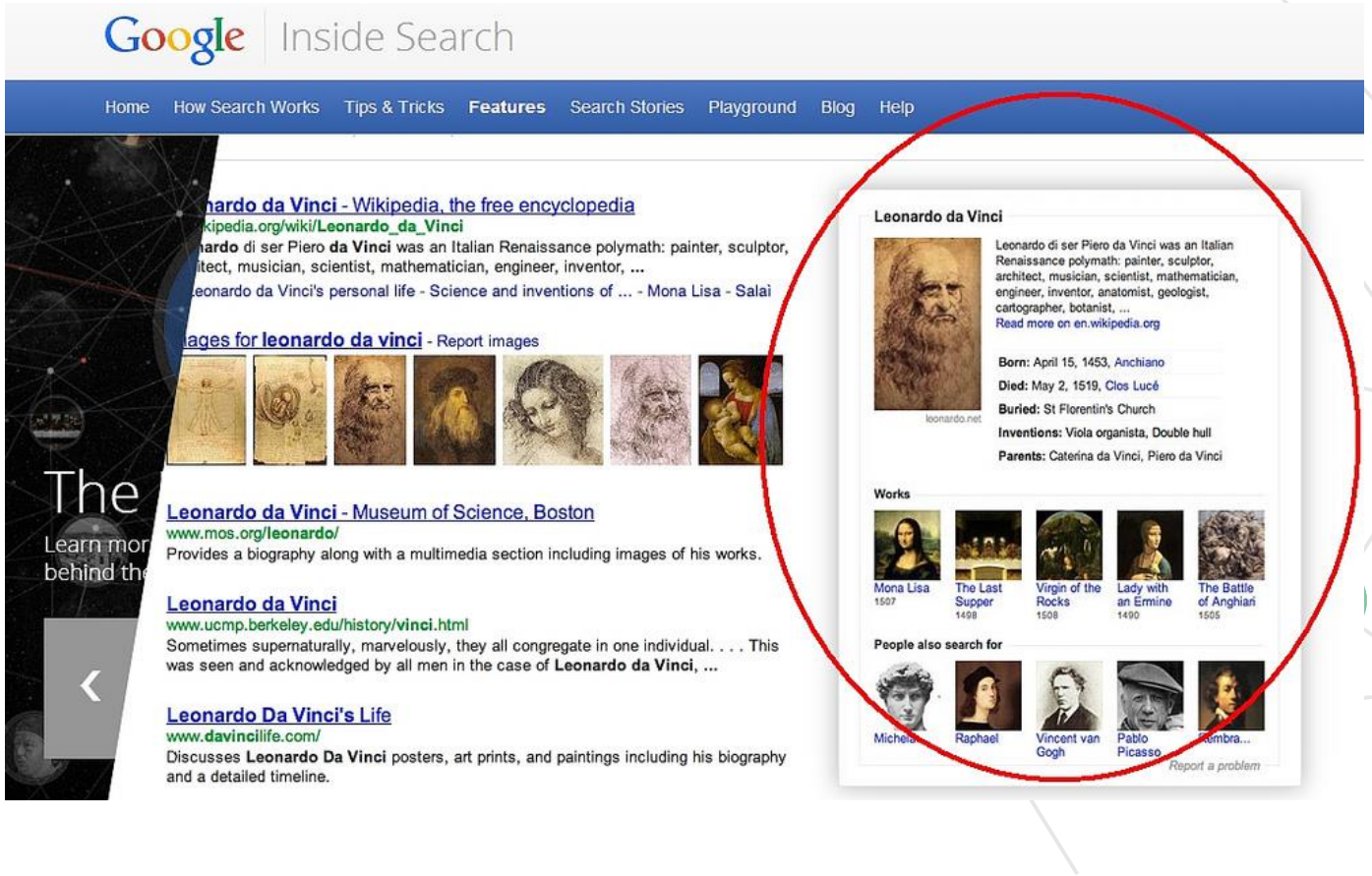


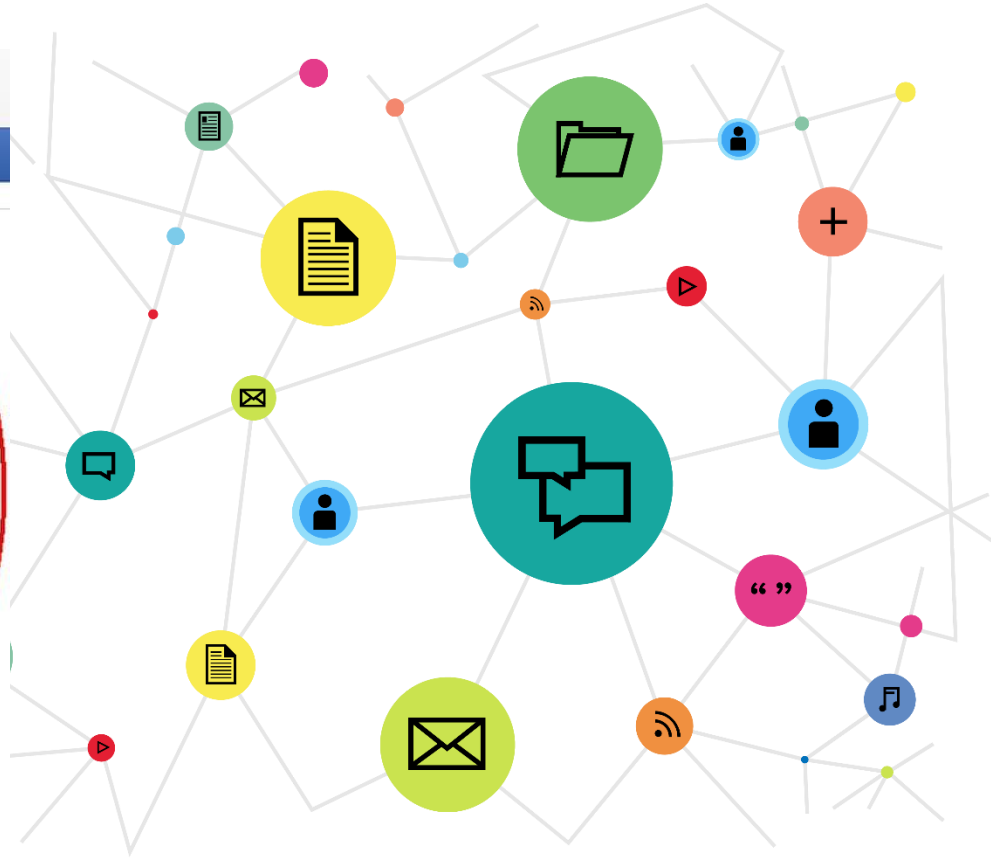
Data were confined only to tables

Today, Data are more heterogeneous

Characterizes of Big Data

- Variety: Knowledge Graph





Characterizes of Big Data

- Velocity

Velocity = Speed

Speed of creating data

Speed of storing data

Speed of analyzing data

Characterizes of Big Data

- Velocity

Late decisions

Missing opportunities

Characterizes of Big Data

- Velocity

Real-time Processing

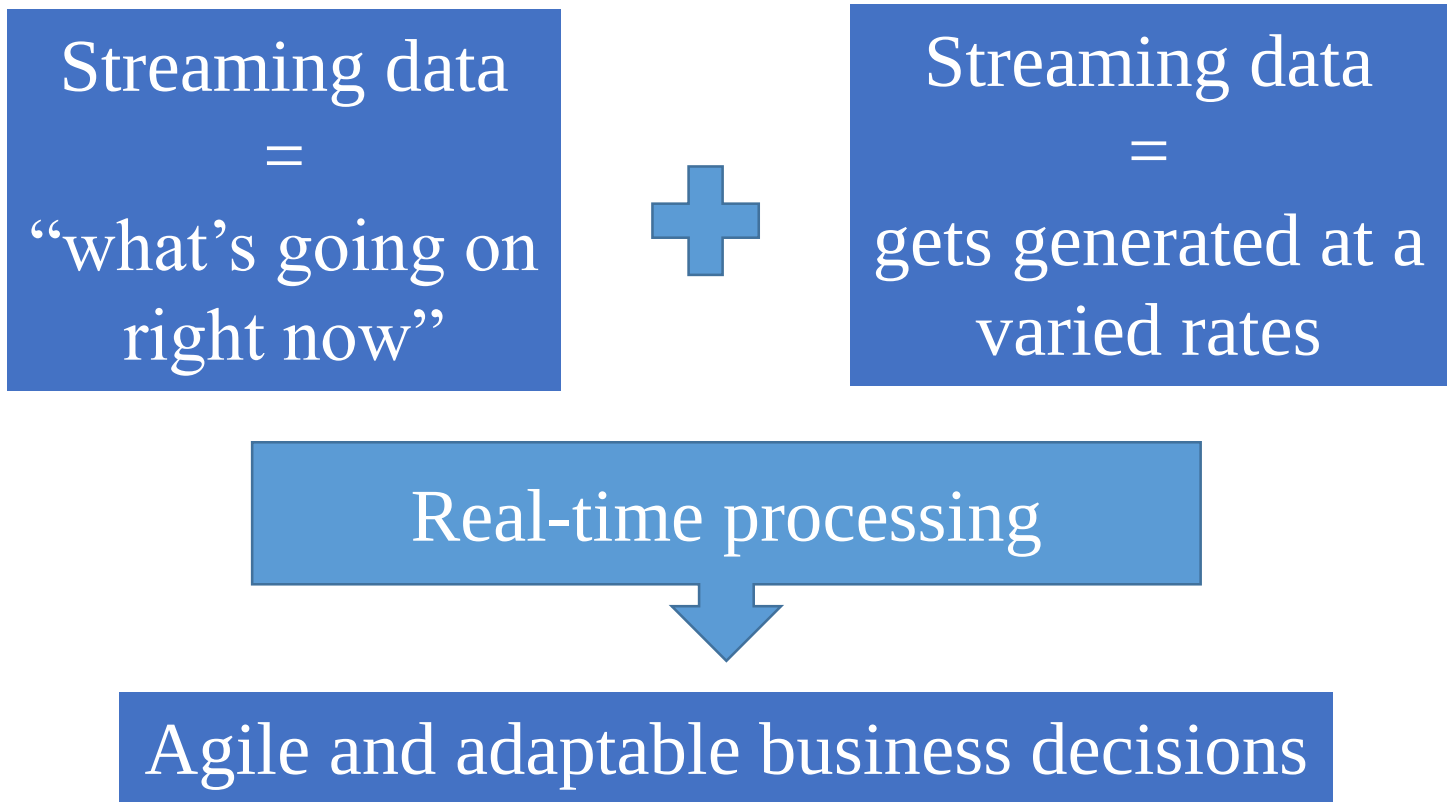
Instantly capture streaming data → Feed real time to machines → Process Act Real Time → Act

Batch Processing

Collect Data → Clean Wait Data → Feed in Chunks → Wait → Act

Characterizes of Big Data

- Velocity



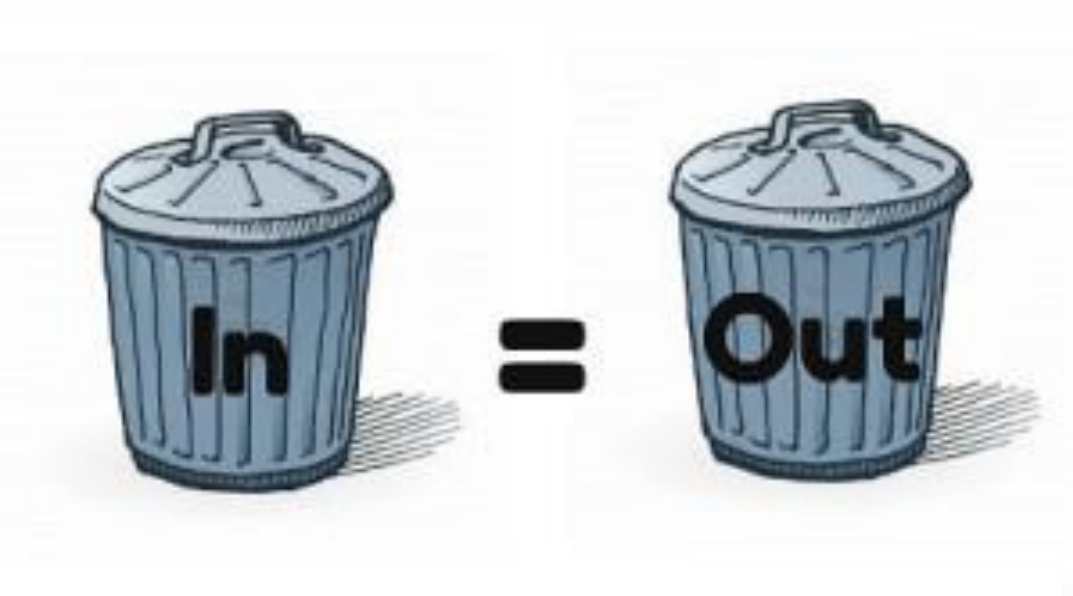
Characterizes of Big Data

- Veracity

Veracity = Quality

Validity

Volatility

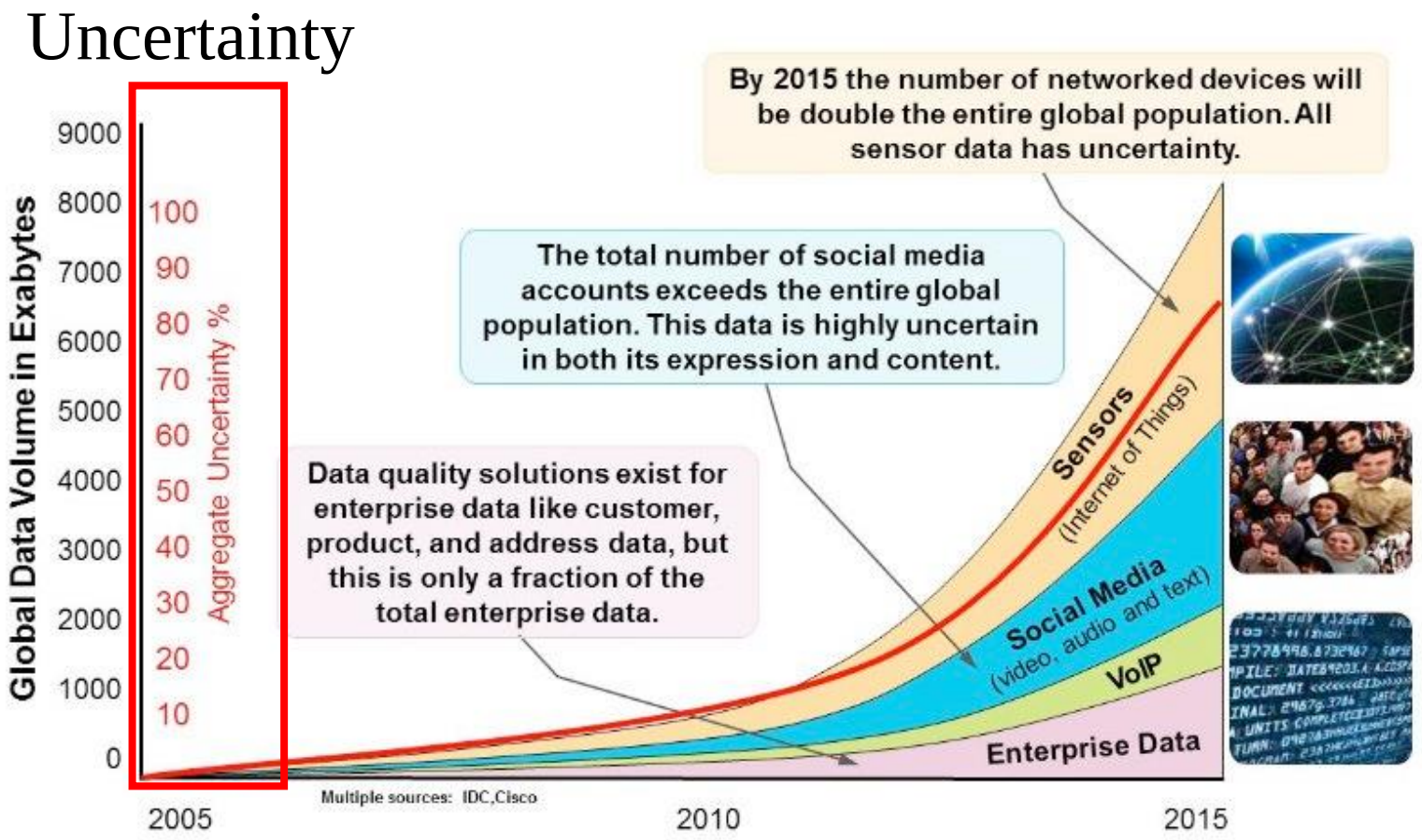


Accuracy of data

Reliability of the data source

Characterizes of Big Data

- Veracity

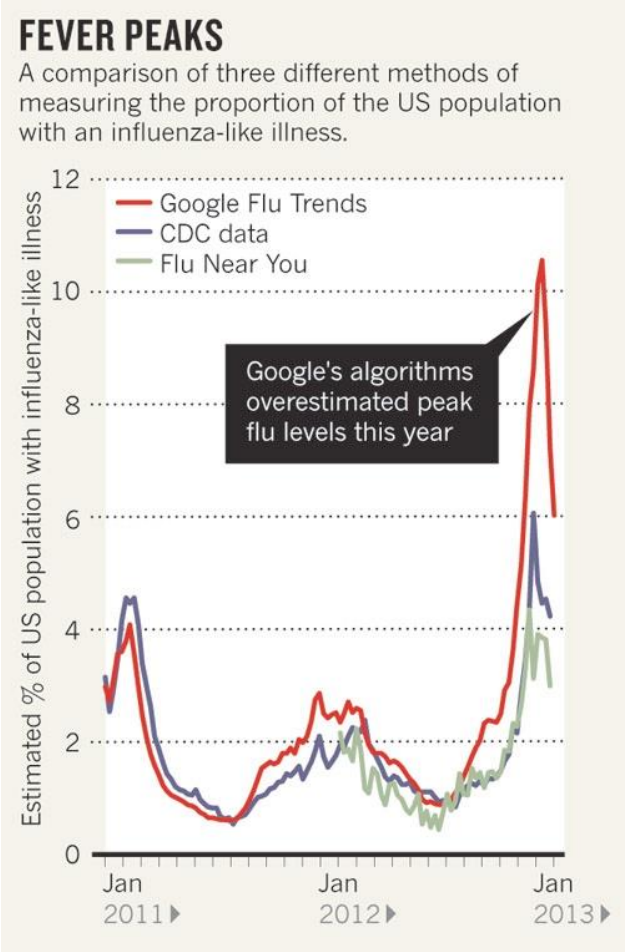


Characterizes of Big Data

- Veracity

Google Flu Trends

Uncertainty



Accuracy of data

Reliability of the data source

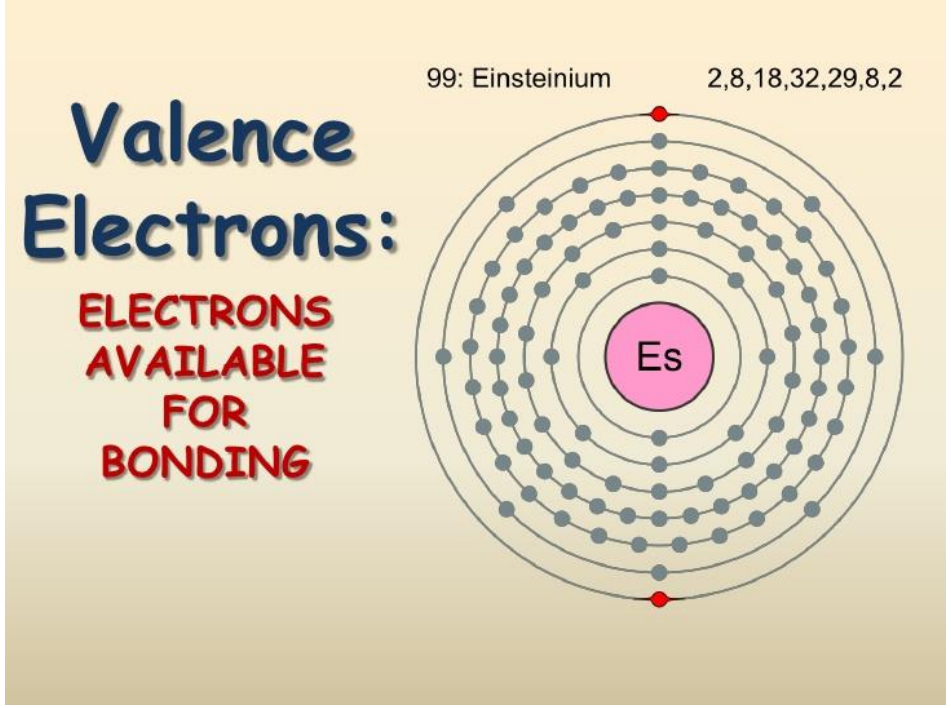
Context within analysis

Veracity

Characterizes of Big Data

- Valence

Valence = Connectedness



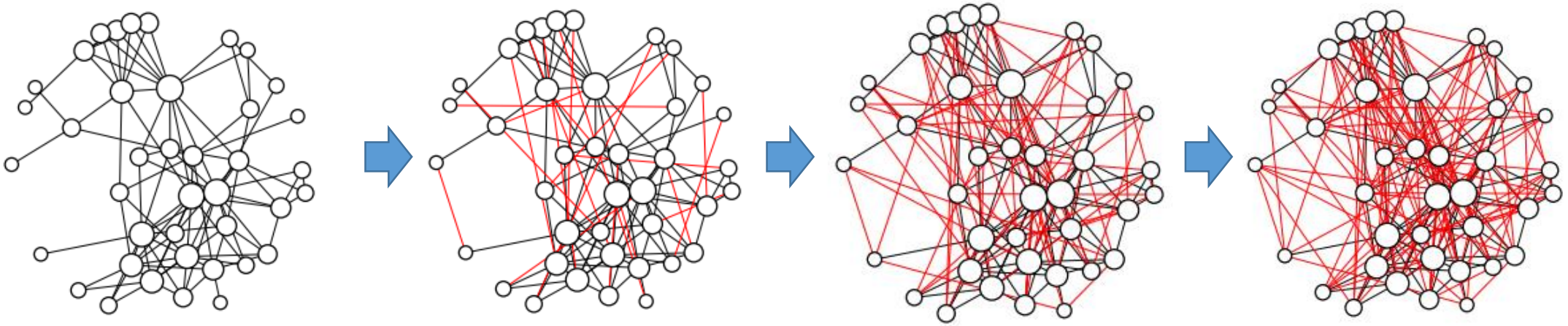
Data Connectivity

Two data items are connected when they are related to each other

Characterizes of Big Data

- Valence

Why worry about Valence?



Valence increases over time → Makes the data connections denser → Organizational Behavior

Characterizes of Big Data

- Valence

Challenges

More complex data exploration algorithms → Inefficient

Modeling and prediction of valence changes → Change with time

Characterizes of Big Data

- Summary