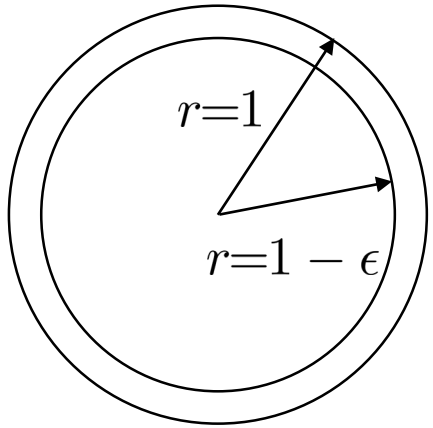
CS3319 数据科学基础
Foundations of Data Science

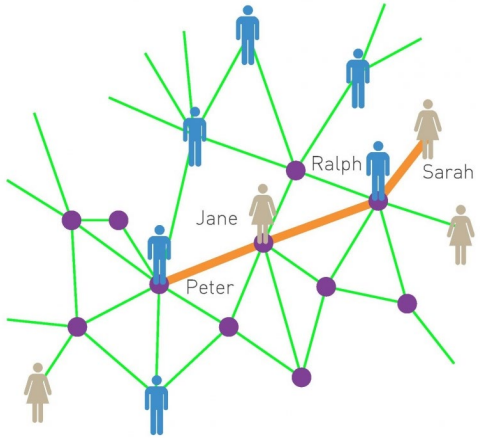
丁家昕

John Hopcroft Center

- 一个质量为1，分布均匀，维度趋近于无穷的球体，我们取其表面接近于0的薄薄一层球壳，这个球壳的质量是多少？0？很小的非0数字？1？



- 发一条消息给一个完全没有接触过的外国人，你通过把这个消息发给自己的朋友，再由朋友用相同的方式转发，大约需要多少轮对方会收到？100？10？6？3？



- 一个人每天产生多少数据？如何估计？

任课教师简介——丁家昕

- 上海交通大学John Hopcroft Center，长聘教轨副教授，IIOT智能物联网中心
- 研究方向：时空数据挖掘，强化学习, 物联网
- 主页：http://jhc.sjtu.edu.cn/~jiaxinding

- 教育背景
  - 2019年University of California, Davis博士后
  - 2018年State University of New York at Stony Brook计算机博士
  - 2012年北京大学信息科学技术学院学士

课程简介

- 本课程推荐教科书:
  - Jure Leskovec, Anand Rajaraman, Jeff Ullman, “**The Mining of Massive Datasets**”, Cambridge University Press.
    http://www.mmds.org/
  - Avrim Blum, John Hopcroft, Eavindran Kannan, “**Foundations of Data Science**”, Cambridge University Press.

课程简介

- 中英双语
- 上课时间：每周二9-10节，9-16周周四3-4节
- 要求：Python编程基础，数学基础(线性代数，概率，算法)
- 助教：吴凯龙，谢予乐
- Office：电院1-203
- Office Hour：微信群答疑，现场答疑提前预约

成绩组成

- 无考试
- 平时成绩（10%）
  - 出勤，随堂小测
- 作业（45%）
  - 4次作业+Lab
- 大作业（2-3人一组）（45%）
  - 期中提交提案(英文，描述问题，相关工作，预期结果)
  - 提交程序撰写报告（英文，论文形式）
  - 期末展示（第十六周，短报告）
  - 评价标准：算法性能排名，期末展示同学评价，报告老师助教评分

学术规范

- 在规定的时间之前提交作业
- 鼓励交流讨论，但是作业要自己完成
- 如果答案、程序或者方法是从其他途径（书、论文、论坛）获得，请自觉引用

课程目标

- 建立数据科学的知识体系
- 探索大数据前沿
- 动手实践数据科学算法，体验科研

Data Science Outline

- Fundamentals
  - Statistics
  - MapReduce
- High dimensional data
  - Locality sensitive hashing
  - Dimensional reduction
  - SVD
- Graph data
  - Link analysis
  - Representation learning
  - Social network
- Streaming data
  - Counting sketch
  - Bloom filter
- Applications
  - Recommender system
  - Privacy
  - Spatio-temporal data

Syllabus

- Introduction (2)
- Data Science Basics (2)
- MapReduce (2)
- Locality Sensitive Hashing(4)
- Graph Data (8)
  - Basics of graphs
  - Representation learning
  - Advanced topics
- Social networks (4)
- SVD and Recommender Systems (6)
- Streaming Data (6)
- Data Privacy (2)
- Spatio-Temporal Data (2)
- Short Presentations(2)

- 4 Labs Friday Week 9, 11, 13, 15 (Tentative)

CS 3319 Foundations of Data Science

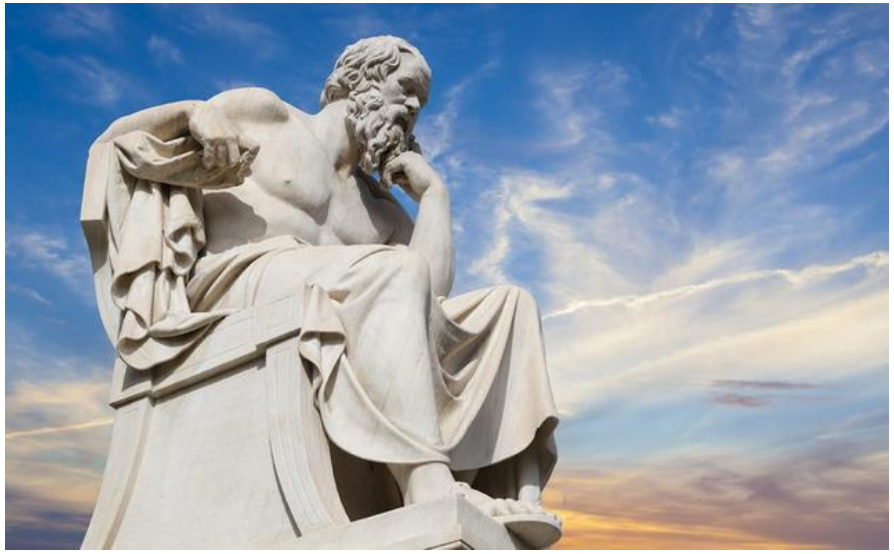
1.Introduction

Jiaxin Ding

John Hopcroft Center

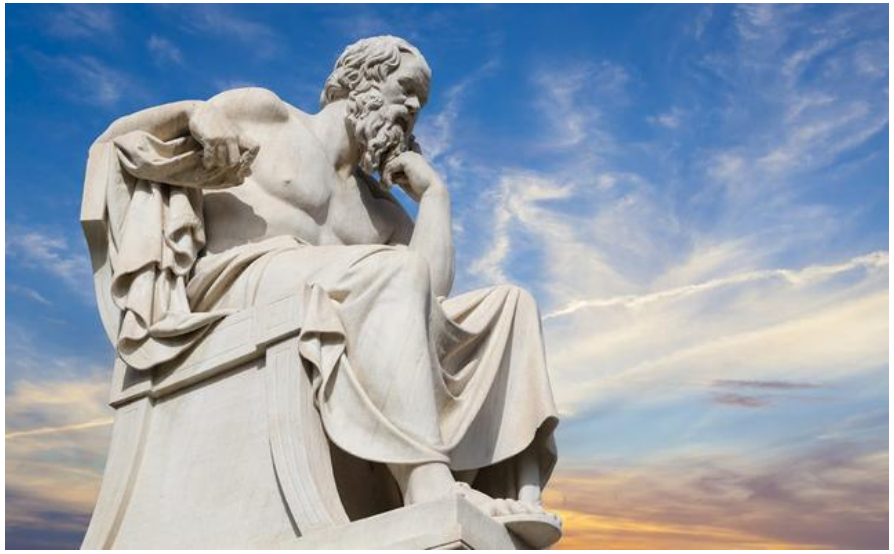
Three Questions

- What is Data Science?
- Where does data come from?
- What does data science do?



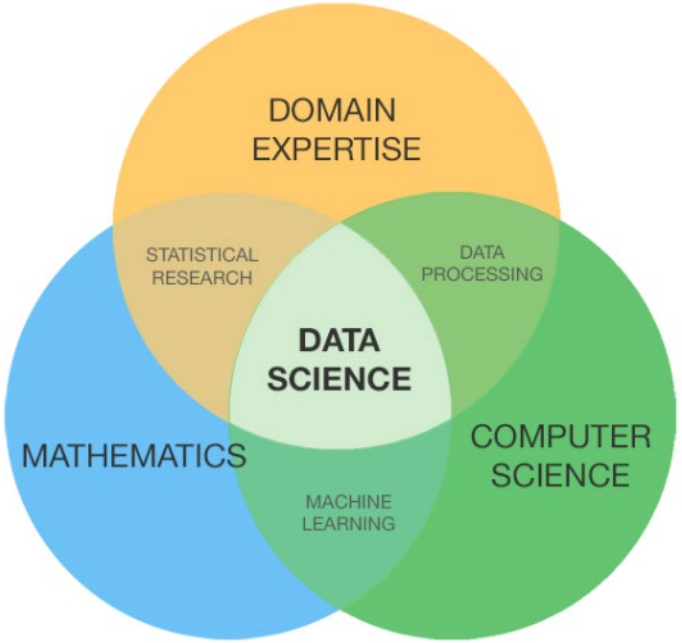
Three Questions

- What is Data Science?
- Where does data come from?
- What does data science do?



Data Science

- Wikipedia: **Data science** is an **interdisciplinary** field that uses scientific methods, processes, algorithms and systems to extract **knowledge** and insights from noisy, **structured** and **unstructured** data, and apply knowledge and actionable insights from data across a broad range of **application** domains.
  - Data science is related to **data mining**, **machine learning** and **big data**.



Evolution of Sciences: Data Science

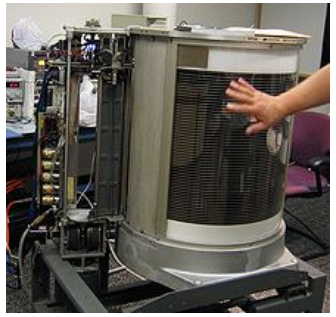
- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Computational Science traditionally meant *simulation*. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes.

$$\mathcal{F}\ (X) \longrightarrow Y$$

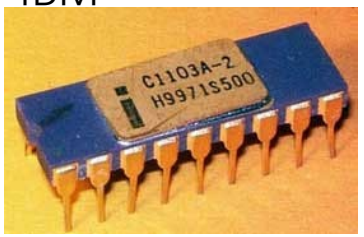
Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Evolution of Data Science: Storage

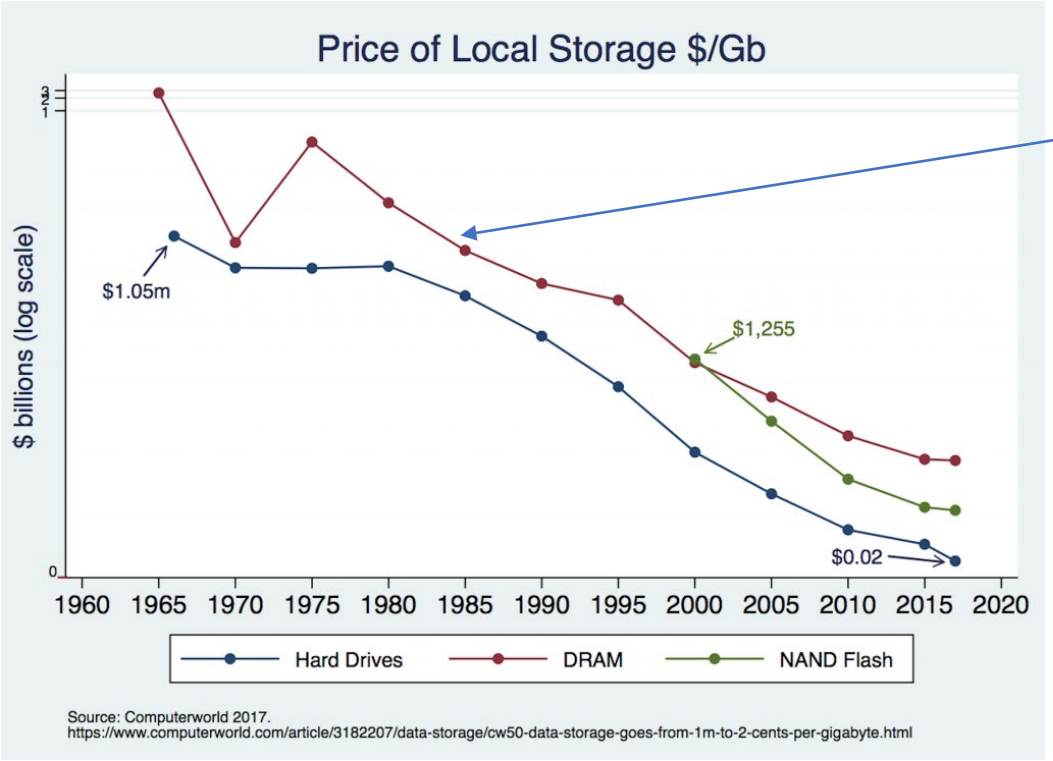
- The storage capability goes up and the unit price goes down.



3.7M Hard Drive
IBM



1kb Memory
IBM

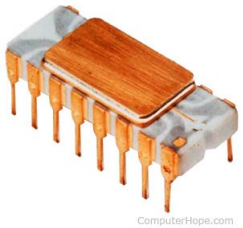


640K is more memory than anyone will ever need on a computer.
Bill Gates?

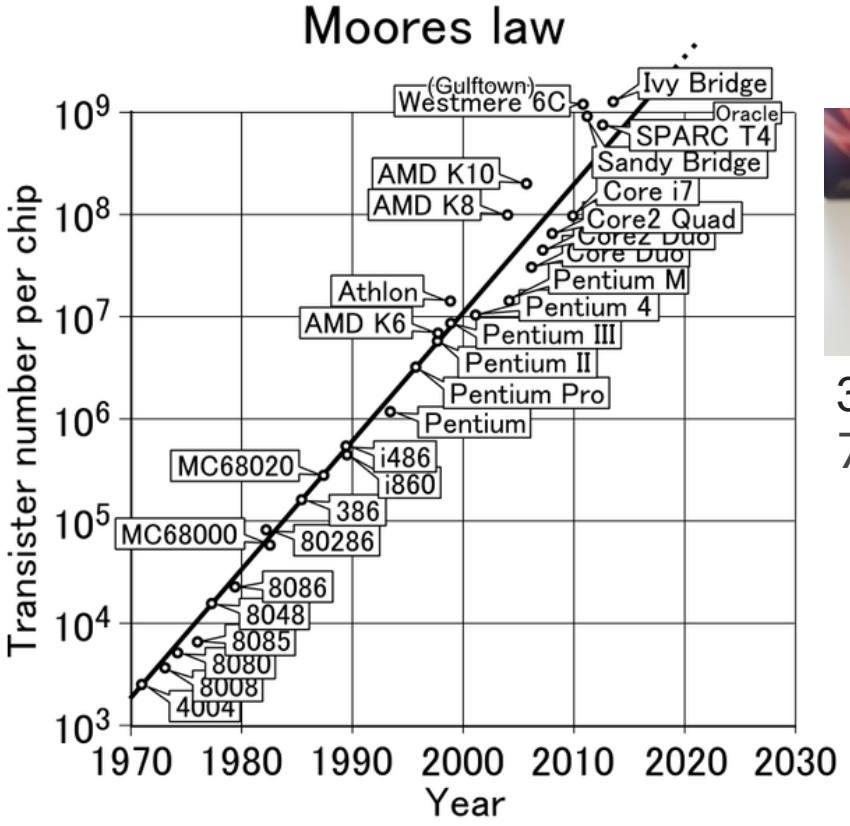
Evolution of Data Science : Computation

- **Moore's law：** the number of transistors on a microchip doubles every two years, though the cost of computers is halved.

Intel 4004



2,300 晶体管,
60,000 OPS,
1971



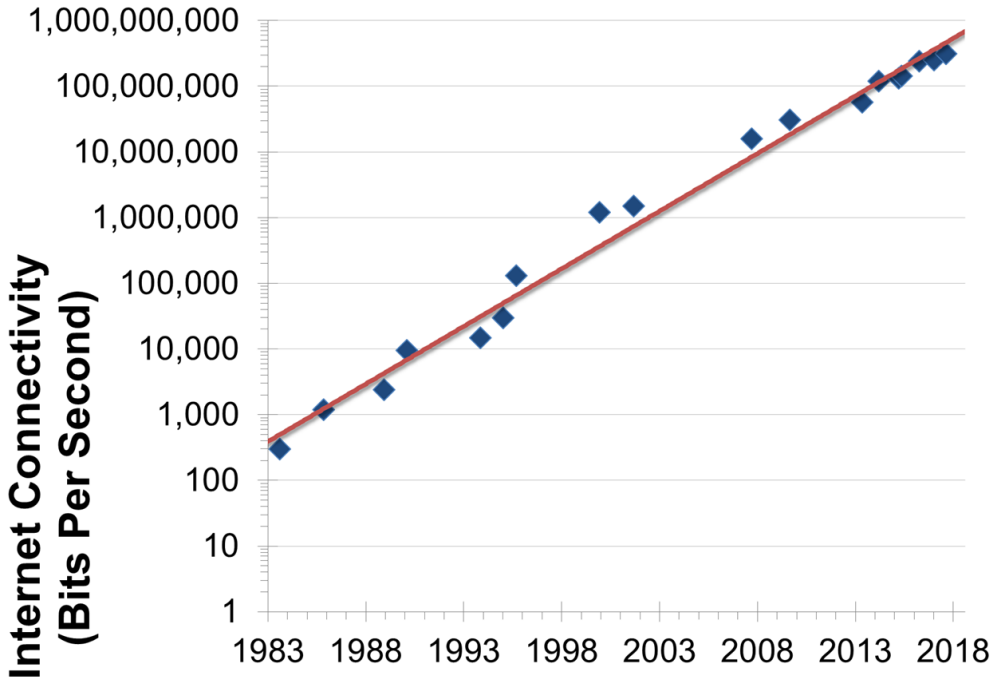
CPU晶体管数目随时间变化情况

395亿 晶体管,
7nm, 3GHz, 2019

GPU, 8000亿 晶体管,
7nm, 4000TOPS, 2022

Evolution of Data Science : Bandwidth

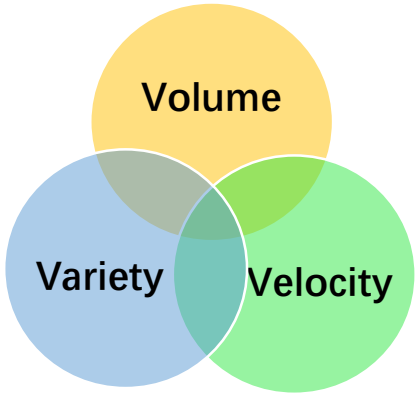
- **Nielsen's Law of Internet bandwidth：** A high-end user's connection speed grows by 50% per year.



网络带宽随时间变化情况

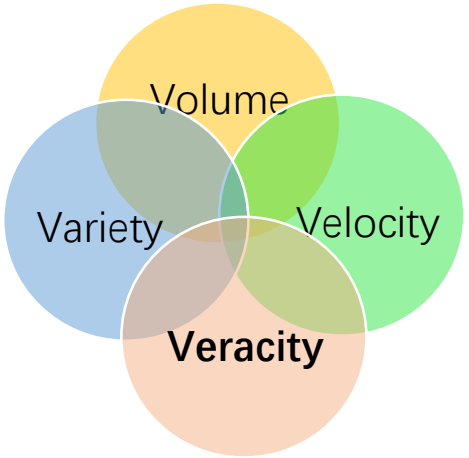
6V of Big Data

Gartner's 3V, 2001



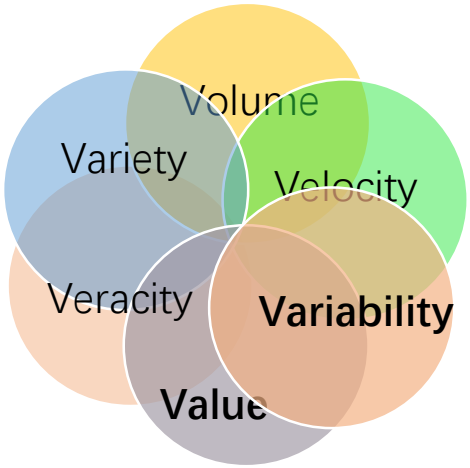
Volume: 数据量
Velocity: 数据输入输出速度
Variety: 数据形式

IBM's 4Vs



Veracity：准确性

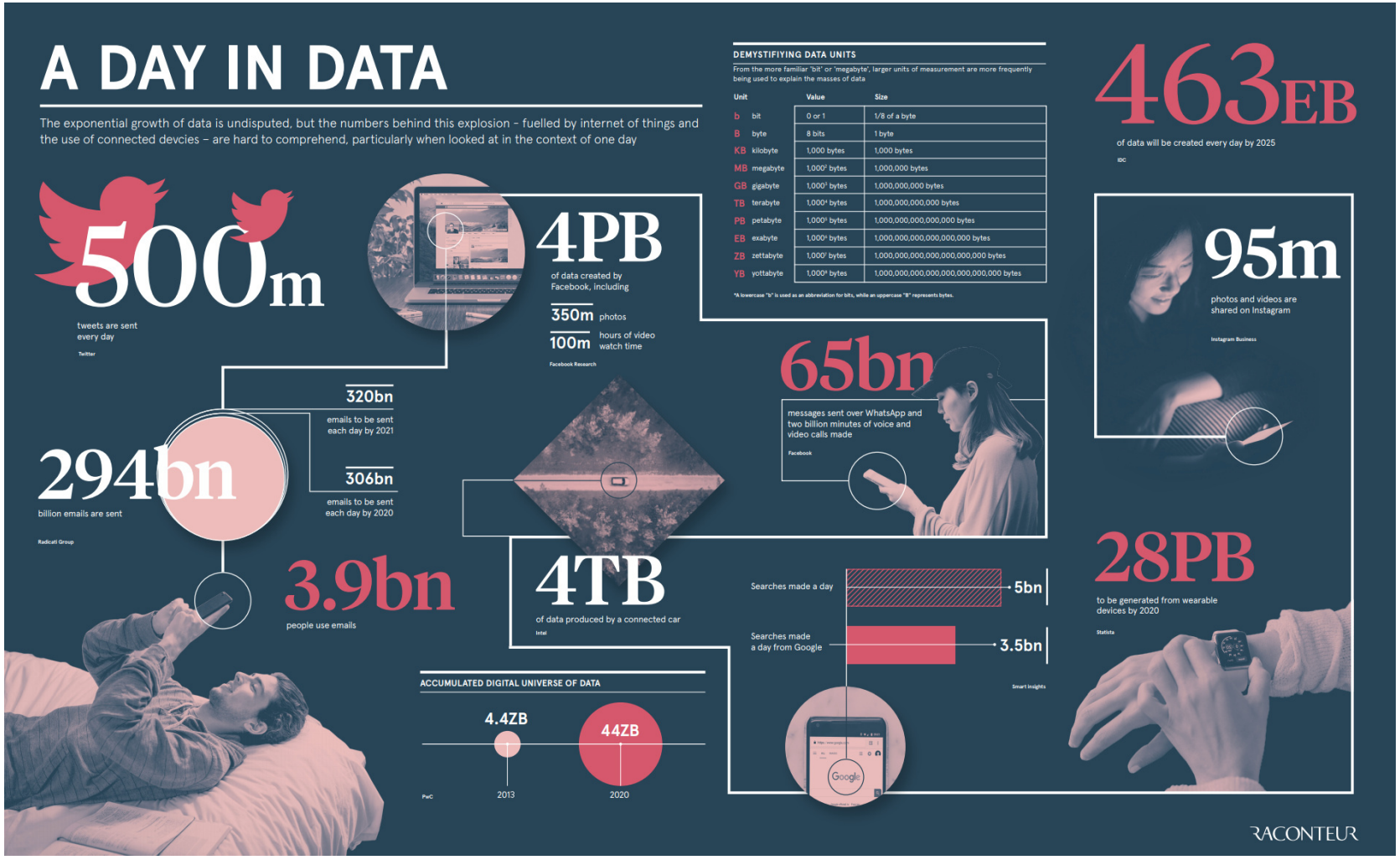
Now 6Vs

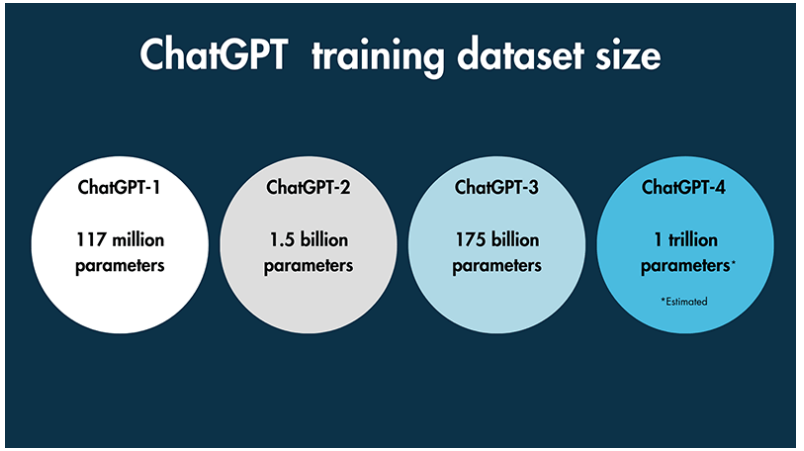


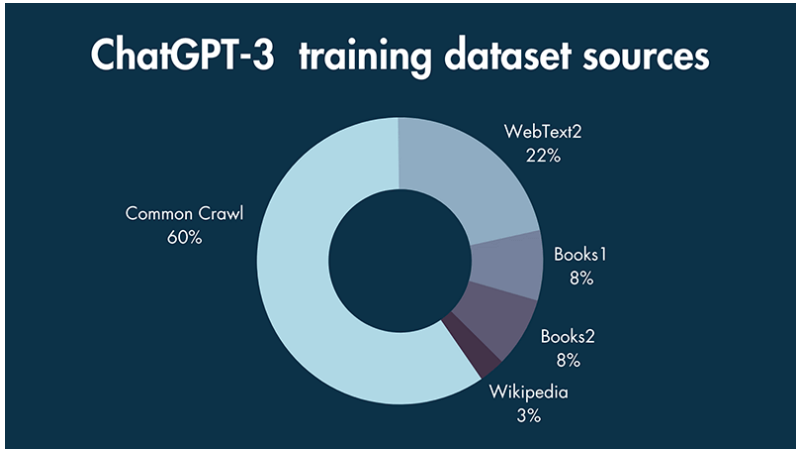
Value：价值
Variablity：变动性

Volume

- Ingest, process and store very **large** datasets.







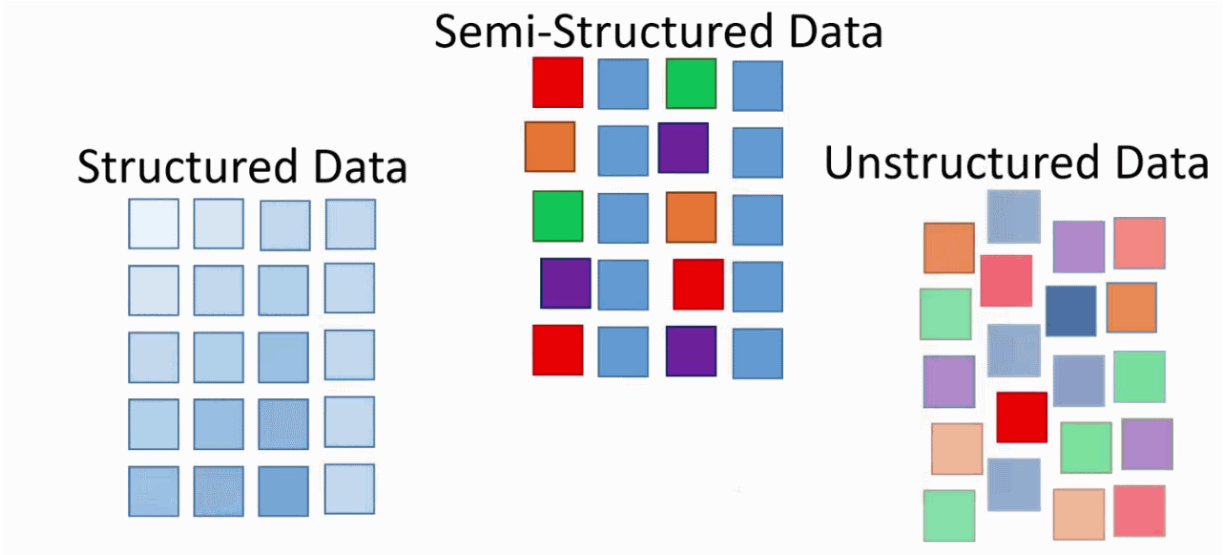
Source: datacamp

Velocity

- The **data flow** is massive and continuous which is valuable to researchers as well as business for decision making for strategic competitive advantages and ROI.
- For processing data with high velocity, **streaming** analytics were introduced.

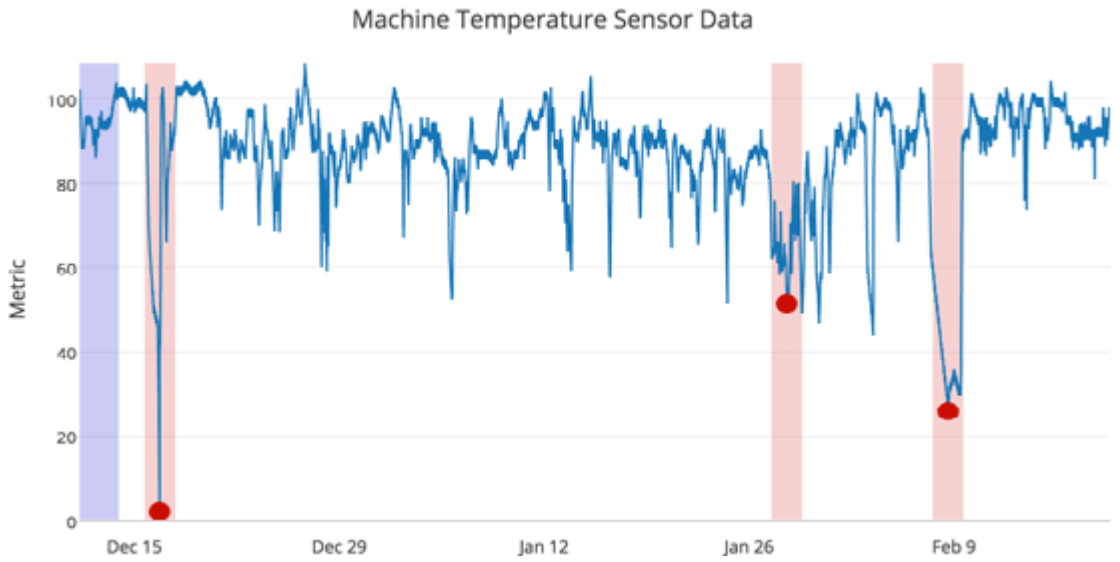
Variety

- Data from different sources and types which may be **structured** or **unstructured**.
  - Structured data： data in the relational databases, ~10%
  - Unstructured data： data without predefined manner (e.g. text), ~ 80%
  - Semi-structured data： data suitable for relational databases but with tags (e.g. XML, JSON)



Veracity

- **Biases**, **noises** and **abnormality** in data.
  - **Identify** the relevance of data and ensure data cleansing is done to only store valuable data.
  - **Verify** that the data is suitable for its intended purpose and usable within the analytic model.



Variability

- Variability refers to data whose **meaning** is constantly **changing**.
  - Particularly the case when gathering data relies on language processing.

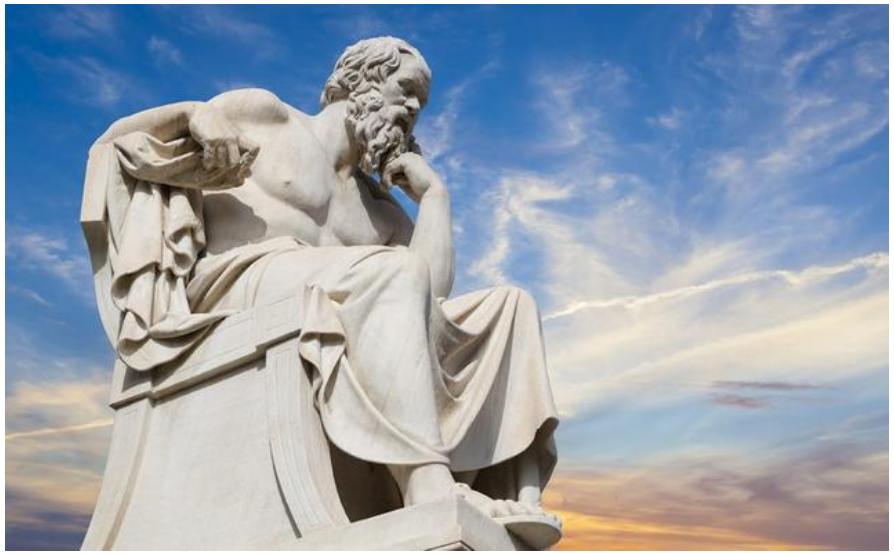
问：今天下雨，去不去上课？
答：我去，我不去。

Value

- The potential **value** of Big Data is huge but the **density** is low.

Three Questions

- What is Data Science?
- Where does data come from?
- What does data science do?



Revolution of Data Generation

1.Traditional Databases

Records are collected in the relational databases.

Relational Database

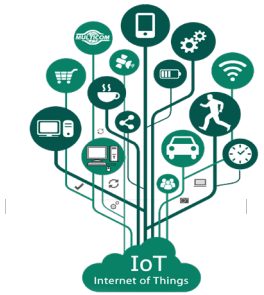
2.Content based system

First data explosion in the era of Web 2.0. Users can generate their own **contents**. The widespread of mobile devices also accelerate the content generating.

Web 2.0
Mobile Networks

3.Sensing based systems

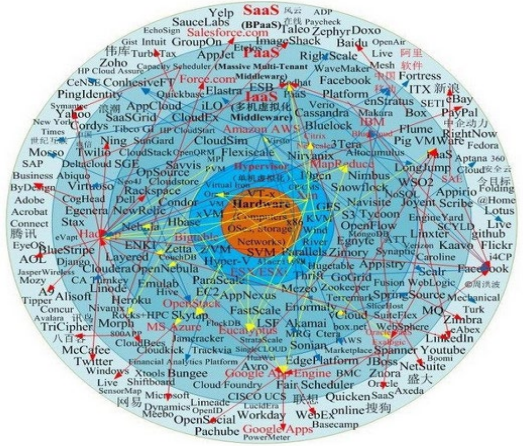
Second wave of data explosion in the era of IoT. Densely deployed connected sensing devices are sensing the world everyday. Huge volume of sensing data is generated.



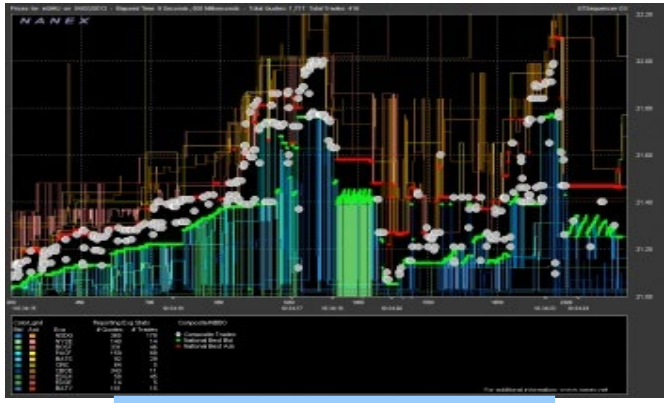
Internet of Things

Big Data in Different Areas

Multi-Media



Network

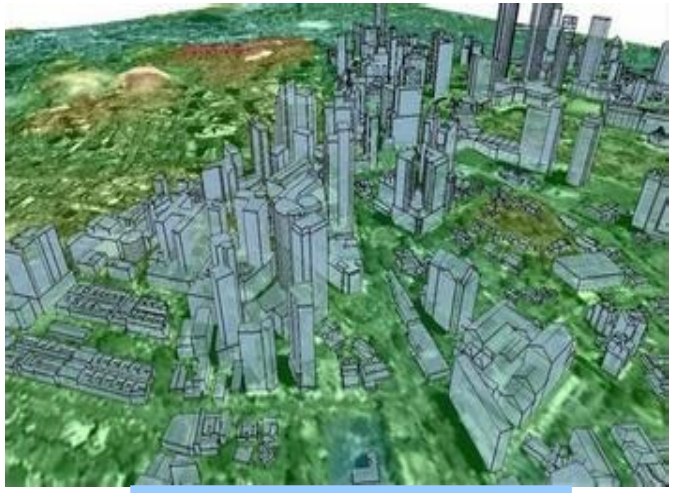


Finance

Big Data in Different Areas

Urban

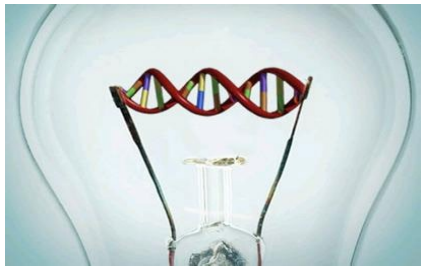
Biology Medical

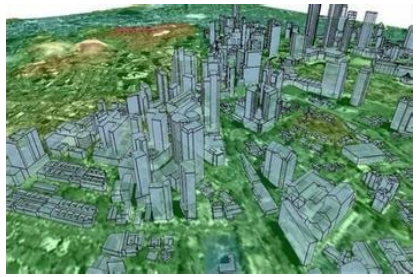


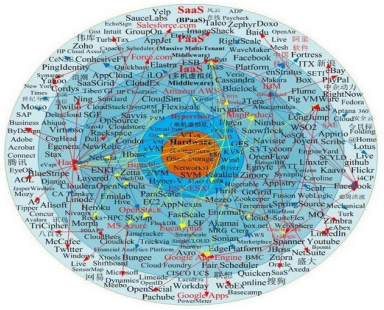
Geography

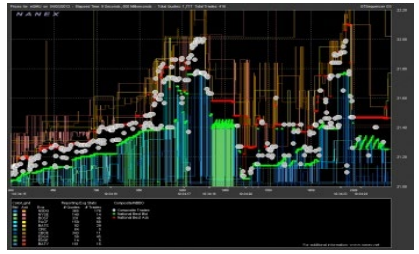
Different types of Big Data

- High dimensional data
- Graph data
- Stream data



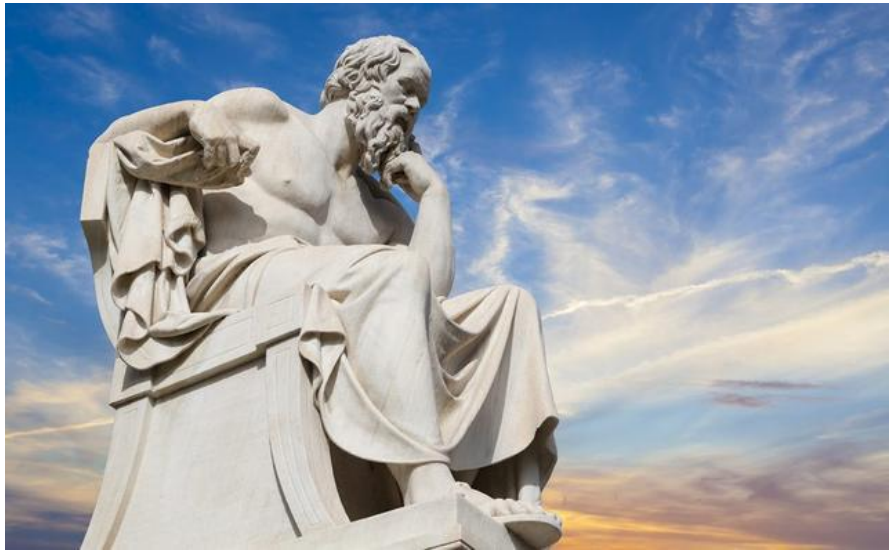






Three Questions

- What is Data Science?
- Where does data come from?
- What does data science do?



Data Science Processes

- **Problem Statement:**
  - ask good questions!
- **Data collections:**
  - search of the data
- **Data cleaning:**
  - deal with missing values, invalid entries, range errors, etc.
- **Data analysis and exploration:**
  - find patterns, statistics, features, embedding, etc.
- **Data modeling and deployment:**
  - Knowledge discovery, prediction.

Practice to Ask Good Questions

- Data scientists are encouraged to ask:
  - What **exciting** things might you be able to learn from a given data set?
  - What things do you/your people really want to know?
  - What **data sets** might get you there?

Douban: Movie Data

正在热映 全部正在热映» 即将上映» 1 / 7

热辣滚烫 ★★★★☆ 7.9 选座购票 | 第二十条 ★★★★☆ 7.7 选座购票 | 飞驰人生2 ★★★★☆ 7.7 选座购票 | 破战 ★★☆☆☆ 3.3 选座购票 | 年会不能停！... ★★★★☆ 8.1 选座购票

热辣滚烫的短评 · · · · · · (全部 245183 条) 我要写短评

热门 / 最新 / 好友

比邻心理 看过 ★★★★☆ 2024-02-10 11:57:43 河北 43292 有用

本片最大的成功是女主练了拳击后没有跟她的搭子在一起。

吴吴 看过 ★★★★☆ 2024-02-10 15:54:19 湖北 36808 有用

她拍这部片子的目的根本不是说她瘦了这个事，是告诉你，无论你这个人活的有多烂、多失败、多差劲，只要你找到一件事，哪怕只是很小的一件事，一直做、专注做、拼了命地做下去，你会赢的，哪怕就一次。

非诚勿扰3 5.4 | 年会不能停！ 8.1 | 三大队 7.7 | 瞒天过海 5.8 | 照明商店 6.2

热辣滚烫 (2024)

导演: 贾玲
编剧: 贾玲 / 孙集斌 / 刘宏禄 / 郭宇鹏 / 卜钰
主演: 贾玲 / 雷佳音 / 张小斐 / 杨紫 / 沙溢 / 更多...
类型: 剧情 / 喜剧
制片国家/地区: 中国大陆
语言: 汉语普通话
上映日期: 2024-02-10(中国大陆)
片长: 129分钟
又名: YOLO / 中国版百元之恋
IMDb: tt28151876

豆瓣评分

7.9 ★★★★☆ 483058人评价

5星 32.6%
4星 39.0%
3星 21.9%
2星 4.4%
1星 2.0%

好于 94% 喜剧片
好于 91% 剧情片

最近的5部作品（已上映） · · · · · · (全部)

2024 2023 2022

热辣滚烫 7.9 | 热烈 7.2 | 热气腾腾欢笑吧 | 三个少年 | 王牌少年加载中

合作2次以上的影人 · · · · · · (全部)

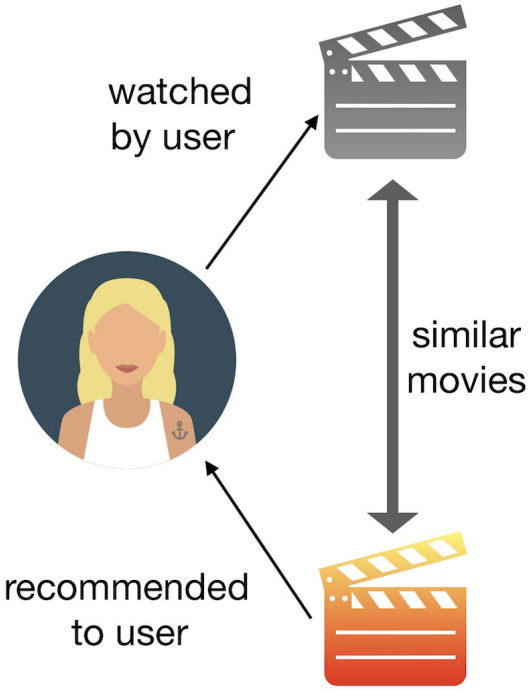
沈腾 合作作品 (28) | 大张伟 合作作品 (20) | 潘斌龙 合作作品 (19) | 张小斐 合作作品 (18) | 白凯南 合作作品 (17) | 瞿颖 合作作品 (17)

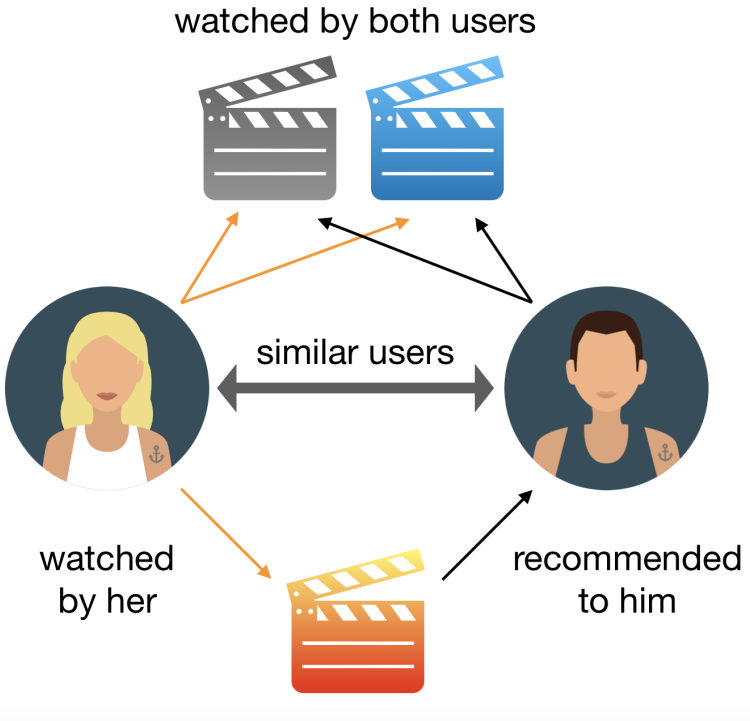
Questions

- Can we predict how well people will like a movie? What about its gross?
- Is there anything special of this actor?
- What does the social network of actors look like?
- Can you recommend a movie to me?
- Can you predict the identity of a reviewer?
- What is the spatio-temporal distribution of the reviewers?

Application: Recommender System

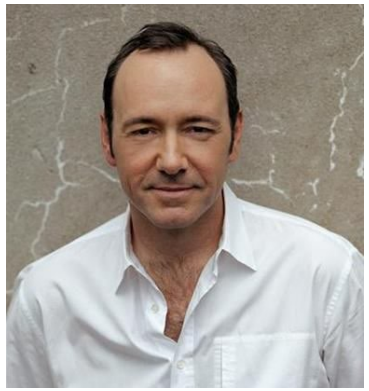
- How to recommend new movie to a user?
- Recommend by content
- Recommend by similar user





Application: Prediction of Drama Success

Users like

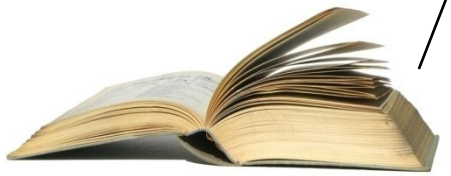


Kevin Spacey

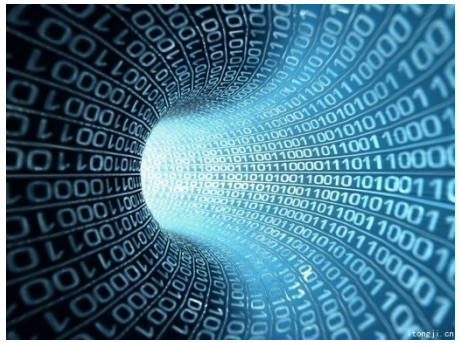
Users like

David Fincher

Users like



BBC drama, <<House of Cards>> (1990 version)

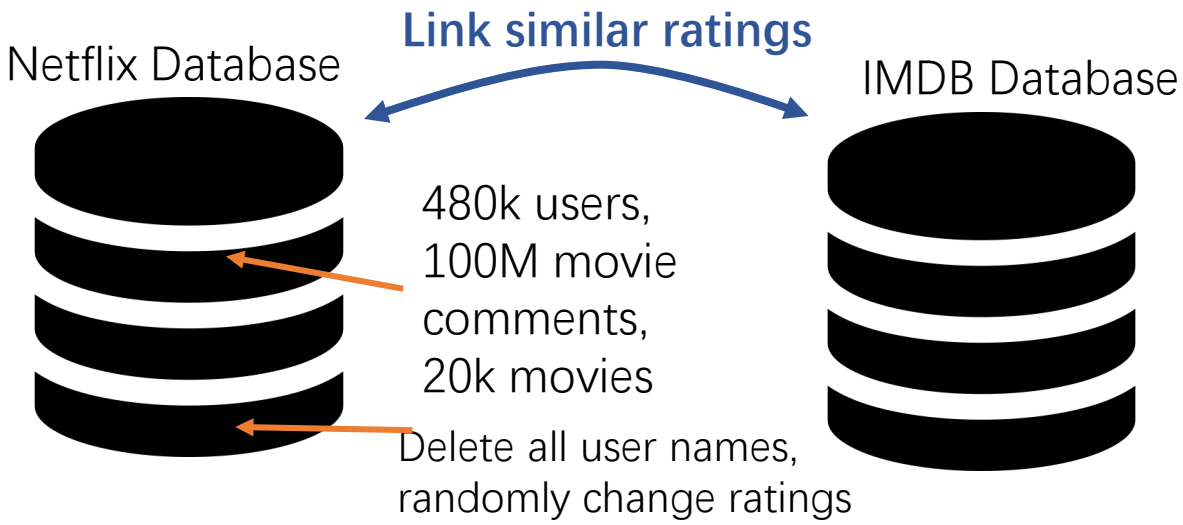


Highly associated,

Success of the Netflix Drama series

Application: Data Privacy

- Dilemma of Data Anonymization
  - Netflix prize **De-anonymization** attack[1]



[1] Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets." *2008, IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 2008.

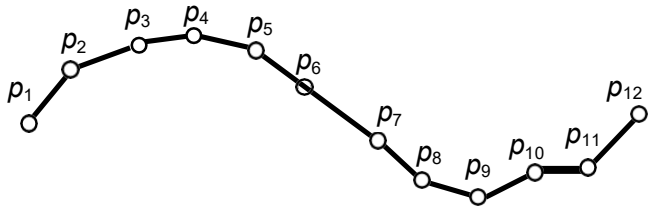
Application: Social Networks

- Community detection / node classification
- Information diffusion modeling
- Friends/Tweets/Job Candidates suggestion

Application: Spatio-Temporal Data Mining

- A spatio-temporal trajectory

$$p_i = (x; y; t)$$



- Behavior modeling of humans and vehicles in the cities
- Prediction of human / vehicles / environment in a certain spatio-temporal point
- Optimization including car route scheduling, lane design, factory relocation

Three Questions

- What is data science?
  - Interdisciplinary:
    - data mining + machine learning + big data techniques
  - On unstructured and structured data
  - 6V big data
- Where does data come from?
  - Web 2.0, IoT
  - Streaming data, graph, high dimensional data
- Where does data science do?
  - Ask questions, collect and clean data, analyze data, and make decision
  - Applications