生物学大数据系列之一

第一讲 大数据时代

The Age of Big Data

作者：维克托·迈尔-舍恩伯格
Viktor Mayer-Schonberger

BBC的纪录片“地平线”专题：大数据时代

Outlines:

1. Definition--------大数据的概念
2. Characteristics---大数据的特征
3. Application------大数据的应用
4. Challenge---------人类面临的挑战

1. Definition--大数据的概念

- No single standard definition…

"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

1. 大数据的概念

- 大数据是指数据集的大小超过了现有典型的数据库软件和工具处理能力；

- 捕捉、存储、聚合、管理大数据以及对数据深度分析的新技术和新能力，正在快速增长，正像预测芯片增长的摩尔定律一样。

---McKinsey Global Institute

Cases:

- 100年前，一个内科医生能知道医学的全面知识；
- 今天，一个基层医生需要知道10000种疾病、3000种药物和1100多种实验室检查才能跟上发展步伐.

Cases:

- 一个典型的完整基因组序列包含120～150GB压缩数据，需要500GB存储空间；
- EMBL数据库收集到数据的速度每年递增200%；
- 在PubMed中已经有1800万医学文章，现在每年增加接近百万篇.

数据的增长速度

"Every 2 Days We Create As Much Information As We Did Up To 2003."

--Eric Schmidt , Google

Zettabytes

Exabytes

Petabytes

Terabytes

2000 2005 2010 2015

44x

未来十年…

数据的增长速度

今天在世界上有90%的数据是在过去的两年里创造出来的！

Zettabytes

Exabytes

Petabytes

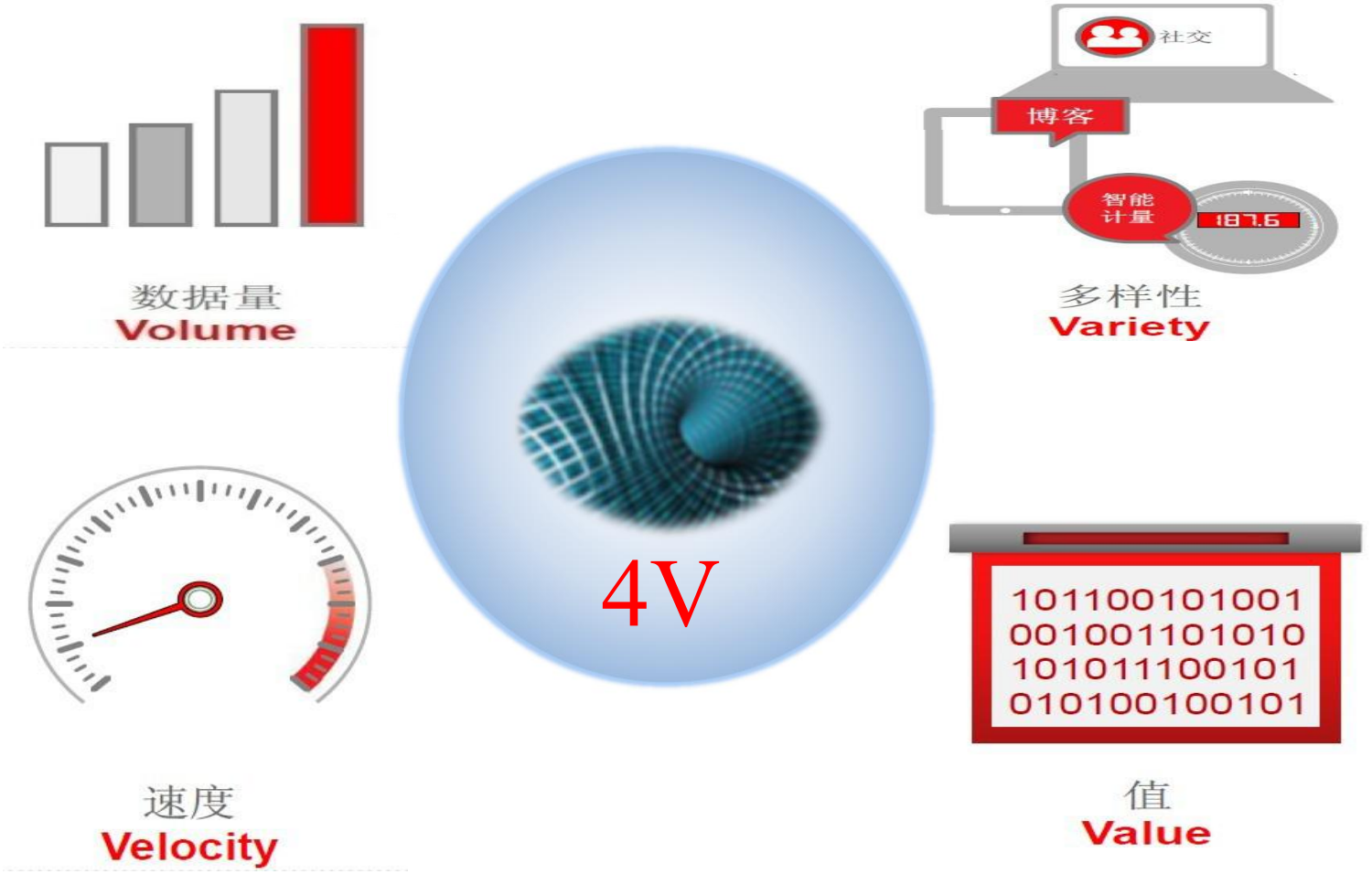
Terabytes

44x

未来十年…

2000 2005 2010 2015

2. Characteristics --大数据的特征



2.1 Volume--数据量

2000年，全球

一年

产生

TWO EXABYTES

数据

1 EB = 1,024 PB = 1,048,576 TB = 1,152,921,504,606,846,976 Bytes

1 TB = 1,024 GB = 1,048,576 MB = 1,099,511,627,776 Bytes

2.1 Volume--数据量

2011

2000年，全球

一年 每天

产生

TWO EXABYTES

数据

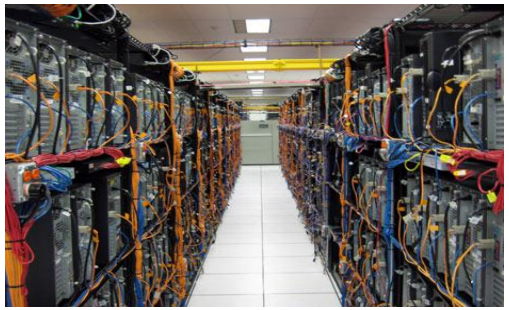
1 EB = 1,024 PB = 1,048,576 TB = 1,152,921,504,606,846,976 Bytes

1 TB = 1,024 GB = 1,048,576 MB = 1,099,511,627,776 Bytes

Cases:

Volume of Big Data

- 每天，发出邮件3000亿封。
- 每天，推特产生2亿条信息，12 TB的数据。
- 每天，脸谱上传2亿张照片，300TB的数据。
- 每分钟，60小时的视频上传到Youtube上。



微软集装箱数据中心 和 Google的服务器农场

云计算和服务器技术使存储和计算不再是问题

2.2 Variety--多样性

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data.

Cases:

医疗影像

MERCK City of Hope

基因测序

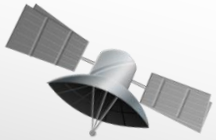
BROAD INSTITUTE illumina

地震探查

PetroChina arcis SEISMIC SOLUTIONS

媒体和娱乐

SONY NBC Sports



卫星图像

NASA NAVTEQ

产品开发

Pratt & Whitney Une société de United Technologies ARM

more...

2.3 Velocity--速度

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

2.3 Velocity--速度

- 1秒定律：实时获取需要的信息.

Velocity of Big Data

- 每秒，亚马逊要处理80笔订单。
- 每小时，淘宝交易500万笔（双11），2天200亿。
- 实时，大量的物联网传感节点、路况信息等。

2.4 Value --价值

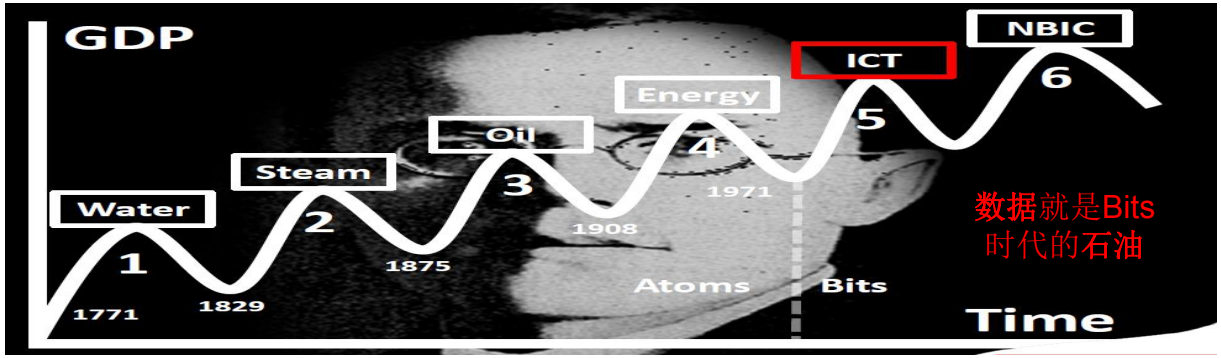
作者：维克托·迈尔-舍恩伯格
Viktor Mayer-Schonberger

Cases:

传统企业天然缺失“用户行为”数据，难以获取消费者的真实需求！

互联网公司完整追踪“用户行为”数据，用大数据技术形成“用户档案”，甚至可以洞悉消费者潜在需求。

用户行为暴露其真实的需求，保存与散失，差别犹如云泥。



3. Application--大数据的应用

- **Data-driven World--** Fish and Oceans of Data
- **What we do with these amount of data?**

3.1 国家层面

大数据成为一种国家战略

2012年三月，奥巴马政府发布《大数据研究和发展倡议》，作为国家战略，计划投入2亿美元的政府研究经费，以解决国家最为迫切的挑战。

2012年五月，联合国发布白皮书《大数据促发展：挑战与机遇》，指出：大数据对于联合国和各国政府来说是一个历史性的机遇。

Cases:

奥巴马在2012年的总统大选中之所以最后胜出，借用了大数据的方法。

Cases:

US 2012 Election

- Drive traffic to other campaign sites
  Facebook page (33 million "likes")
  YouTube channel (240,000 subscribers and 246 million page views).
- Every single night, the team ran 66,000 computer simulations, Reddit!!!

- Data mining for individualized ad targeting
- YouTube channel (23,700 subscribers and 26 million page views)

3.2 商业应用

会员为亚马逊贡献了三分之一的运营收入，而究其原因，大概与亚马逊精准的“推荐系统”有关。曾在亚马逊的网站上消费过的朋友可能注意过，当你选择一种商品的时候，他总会很贴心地为你推荐相关的产品。拿国内流行的一句话说：“他比你更了解你自己”。

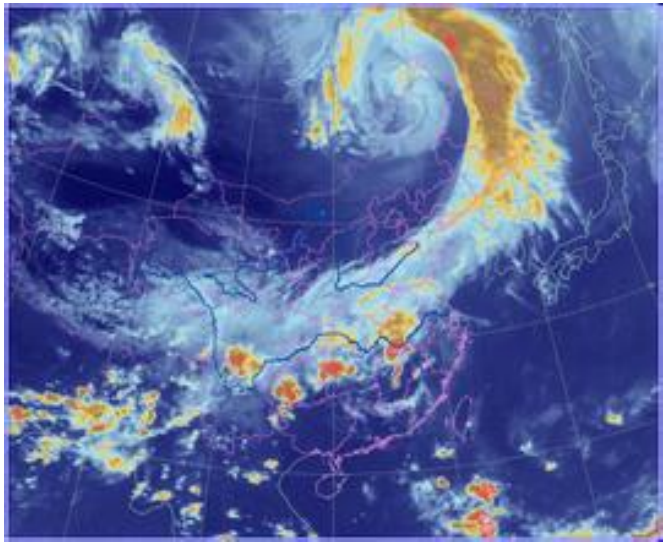
Cases:

股价预测

Twitter消息成对冲基金经理预测股价走势利器：专家们发现Twitter消息由于具有直接性的特点，因而可以更准确地测量人们的情绪。2010年10月发布的一项研究中，利用社交网站来预测纽约道琼斯指数的走势，结果准确率可以达到了87.6%。

3.3 科学应用

一家德国的航空公司，在飞机上安装了许多监测设备，在执行日常的飞行任务时，获取大量气象数据（如：气温、气压等），通过采集大量的数据并将其反馈给当地的气象部门，他们惊喜地发现，天气预报的准确率提高了7个百分点。这实在是非常了不起。



3.4 医疗应用

加拿大多伦多的一家医院，针对早产婴儿，每秒钟有超过3000次的数据读取。通过这些数据分析，医院能够提前知道哪些早产儿出现问题并且有针对性地采取措施，避免早产婴儿夭折。而研究表明，那些由于早产不幸夭折的孩子们在“特定时期”并不会有剧烈的生命体征变化，而通过大数据分析，只要及时进行医疗干预，这些灾难完全可以避免。

启示：大数据的魔力在于不仅仅是事后的分析评估，而是能够在某种程度上“预知未来”。

4. Challenges—人类所面临的挑战

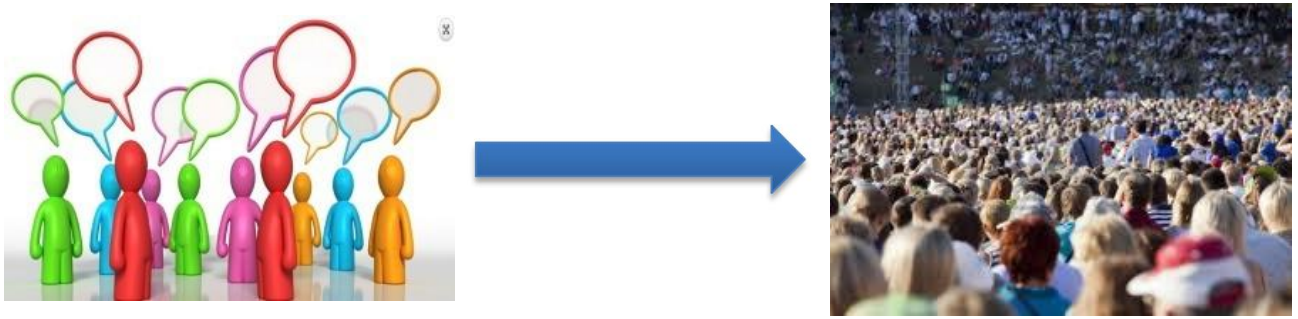
这是一场革命，庞大的数据资源使得各个领域开始量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。

—哈佛大学社会学教授加里•金

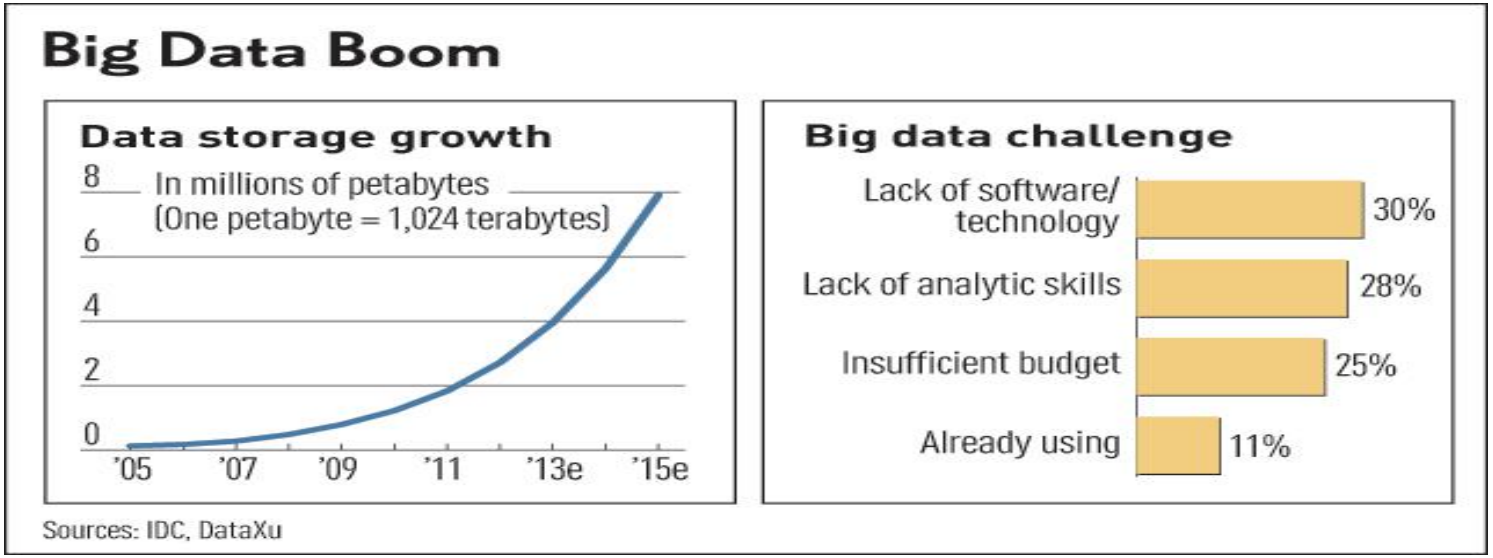
改变了信息产生和消费的模式：

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data



4.1 Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

4.2 大数据是把双刃剑

这是一个最好的时代，这是一个最坏的时代

不仅仅是隐私的泄露，还有被预知的可能性

大数据能预测我们可能生病、拖欠还款和犯罪的算法，会让我们无法购买保险、无法贷款、甚至在实施犯罪前就被预先逮捕。统计把大数据放在了首位，但即便如此，个人意志是否应该凌驾于大数据之上呢？

Cases:

- 美国折扣零售巨头Target公司因网络安全漏洞发生数据外泄事件，7000多人受到影响；
- 棱镜门事件则让更多人深刻认识到，大数据是把双刃剑.

Thanks for your attentions!