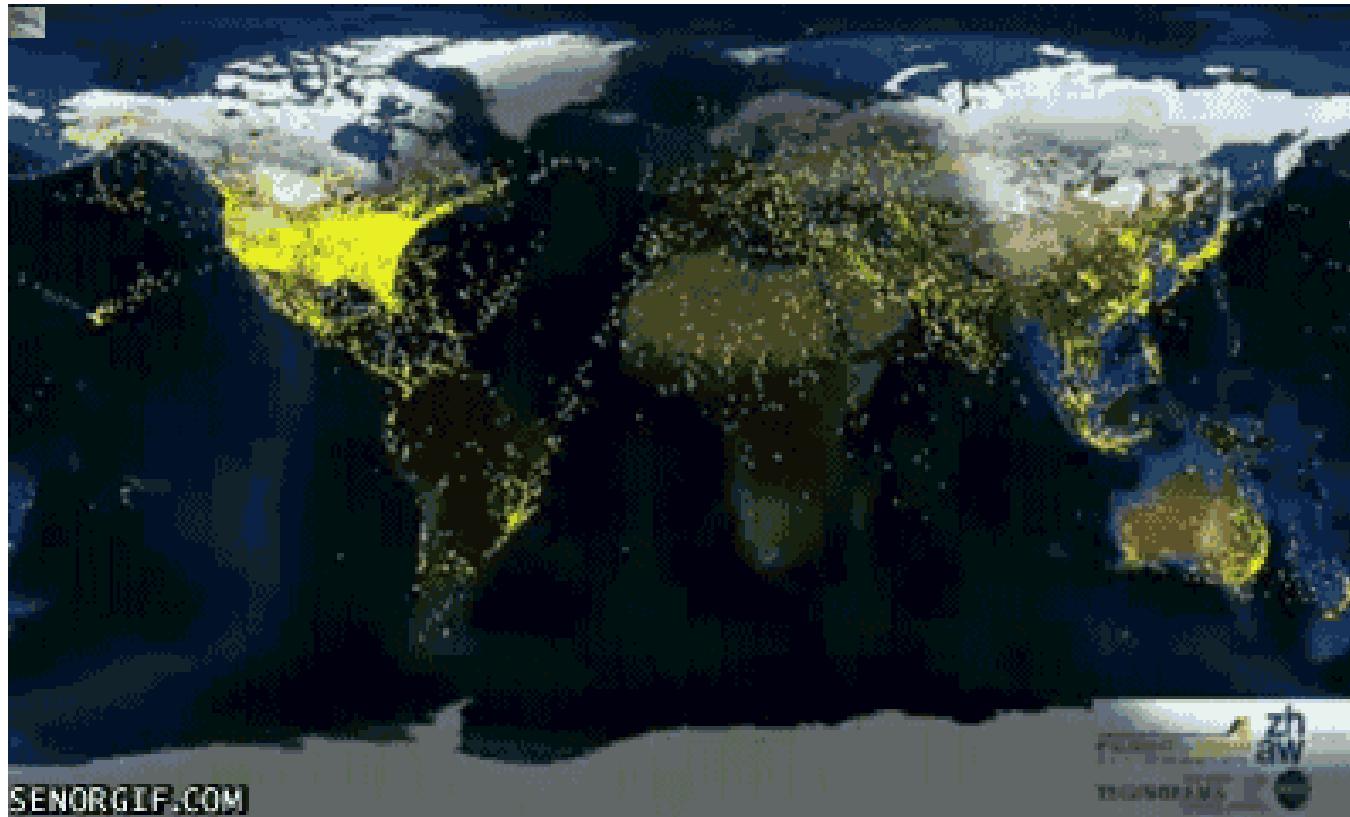




数据科学与工程导论

Introduction to Data Science and Engineering

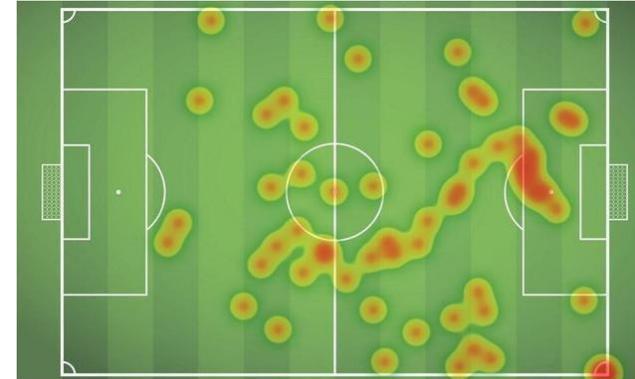


洞见：

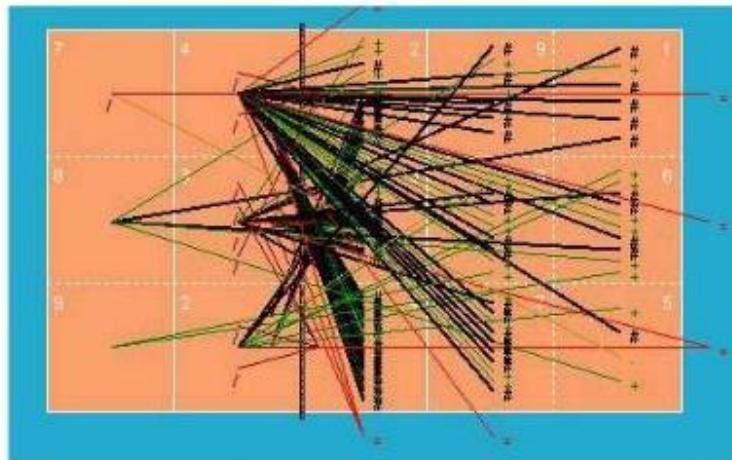
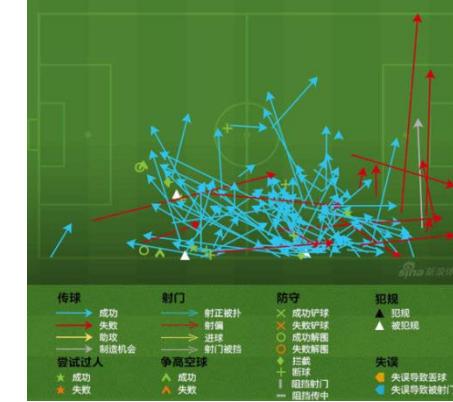
航空业的时空规律 | 世界人流的规律 | 经济活力

全球航班路线轨迹

开篇实例



法国VS德国 格列兹曼活动位置热力图



中国队决赛对俄罗斯进攻线路图



洞见：

敌我优势 | 身体机能 | 战略部署

体育赛事策略规划

第1章 绪论

1
2
3
4
5

信息文明与数据简史

数据科学的基本内涵

第四范式：数据密集型科学

数据科学的应用

实践：以Git和Python为中心



比较重要的四个人种



智人迁徙至全世界的路线示意
(仅用了两千余年)

人类的祖先



- 生产力水平底下
- 以简单自然农业为生

原始文明



农业文明

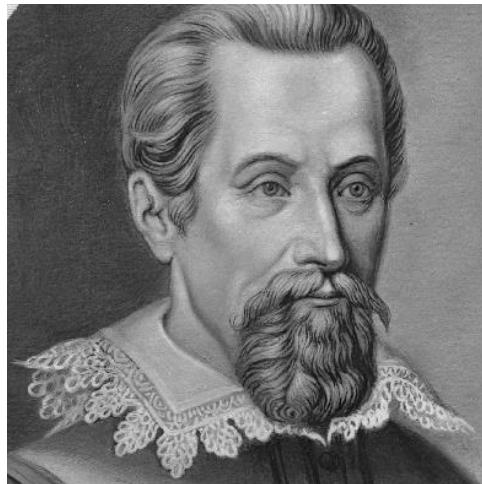
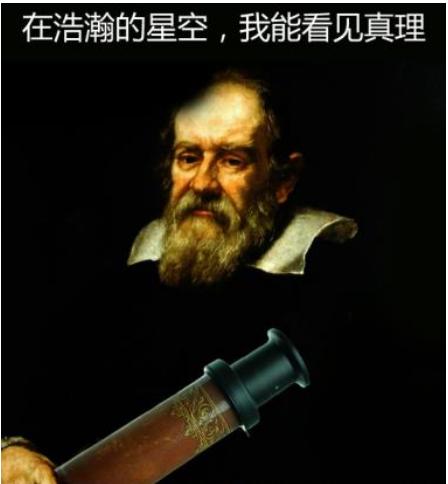
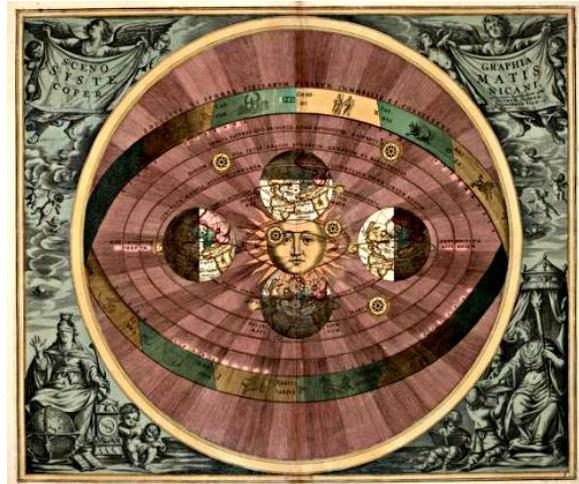
- 自然经济和农业生产占主导地位
- 利用可再生能源（畜力、水力等）

工业文明



人类文明的启蒙与机械思维

大胆假设 小心求证



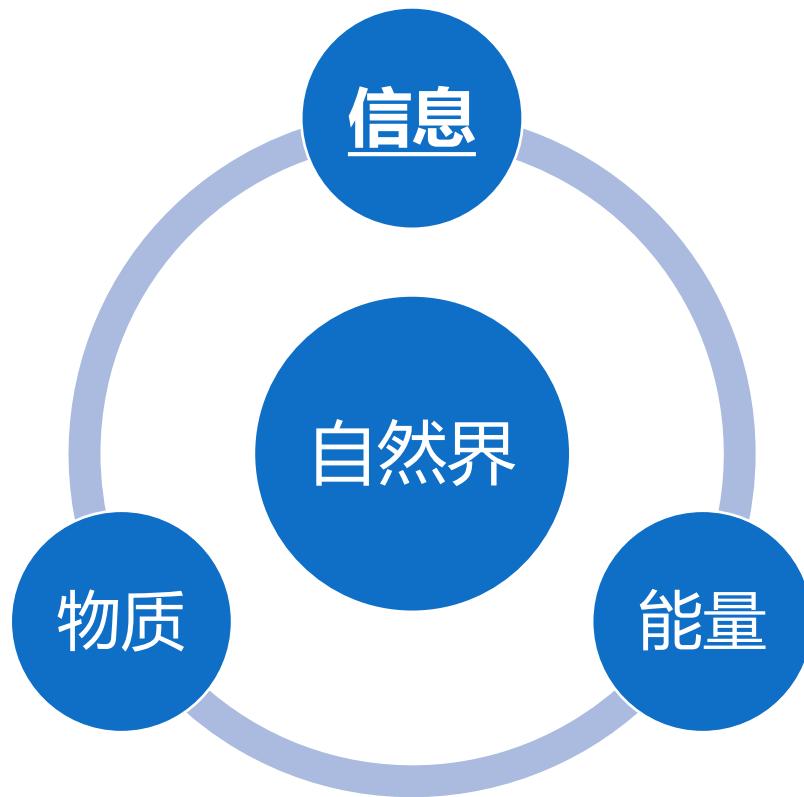
开普勒发现三定律

积累足够的观测数据

提出一个先验的世界模型

调整模型的参数直至拟合数据

机械思维

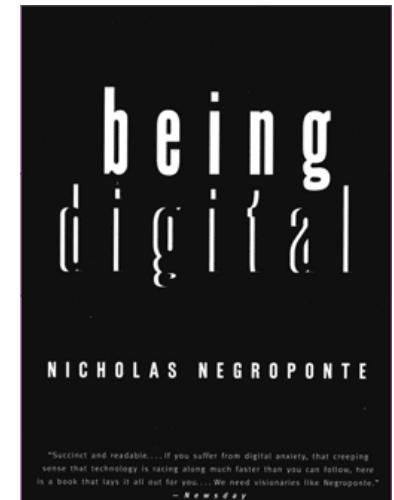


信息革命	技术
第一次信息革命	语言
第二次信息革命	文字
第三次信息革命	造纸术及印刷术
第四次信息革命	电报、电话和电视
第五次信息革命	计算机、互联网和物联网
第六次信息革命	云计算、大数据与人工智能

人类文明的进击与数据思维

今天的大数据是我们信息化到一定阶段之后，必然出现的一个现象（自然现象），这种史无前例的变化有几个主要的驱动力：

- 摩尔定律所驱动的指数增长模式；（**比特化**）
- 技术低成本化驱动的万物的数字化；（**信息化**）
- 宽带移动泛在互联驱动的人机物广联连接，以及最后大规模的汇聚（**网化、物化、云化**）



尼葛洛庞帝,《数字化生存》,1995

人类文明的进击与数据思维

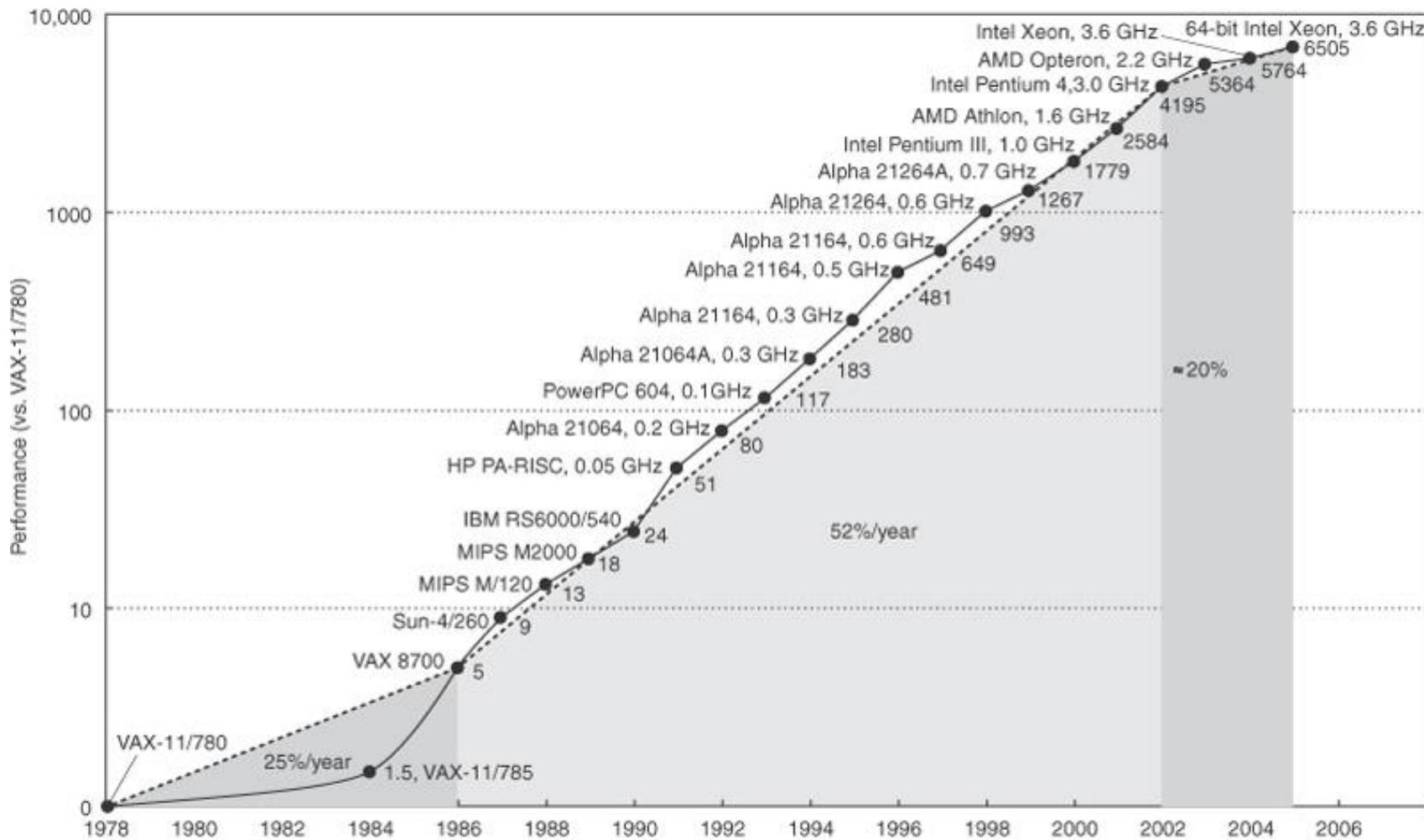
1965年微芯片上的元件数增加了1倍，Gordon Moore于是预言这一趋势近期内将继续。1975年他修改为每两年翻一翻，后来又说是18个月，或者说按指数律增长，每年46%。这就是著名的**摩尔定律**。

- 美国的主粮玉米从1950年以后平均产量每年增长2%；
- 蒸汽涡轮发电机把热能转换为电能效率在20世纪年增长率为1.5%；
- 1881-2014室内灯光有效性年平均增长2.6%，而室外为3.1%；
- 洲际旅行远洋客轮效率平均每年提高5.6%；
- 1973-2014汽车的燃油的换能效率年平均提高2.5%。



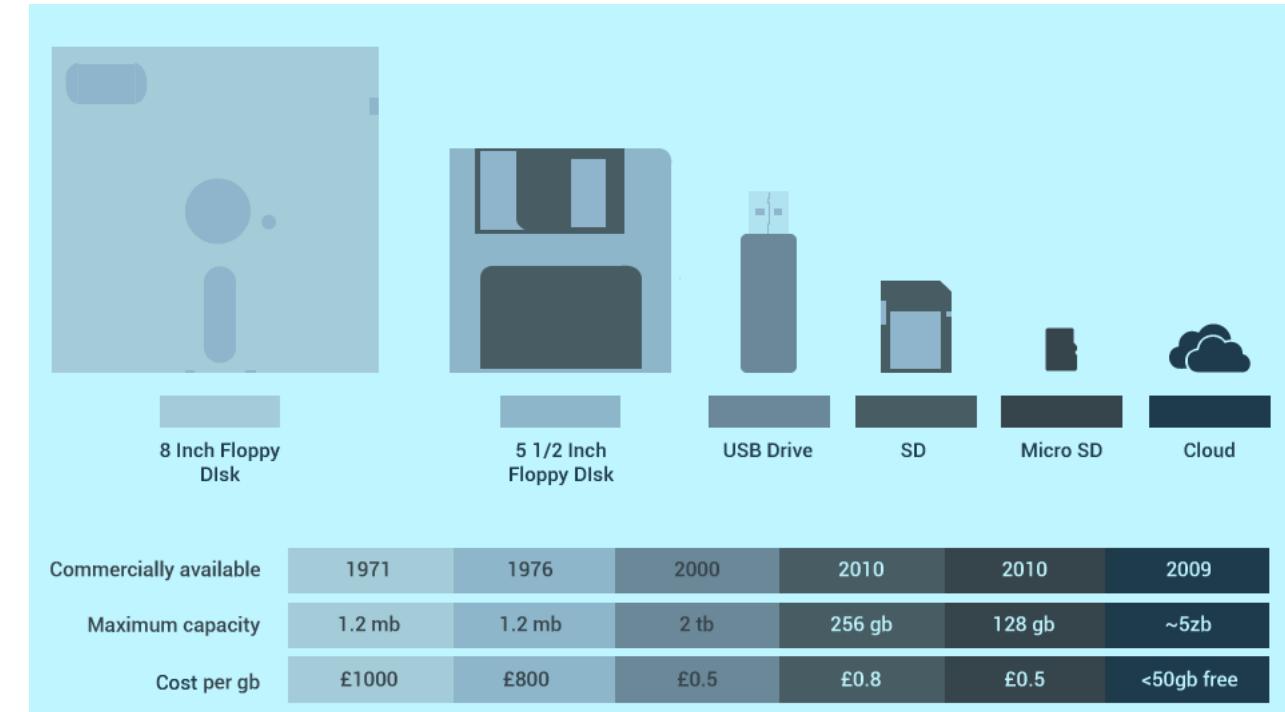
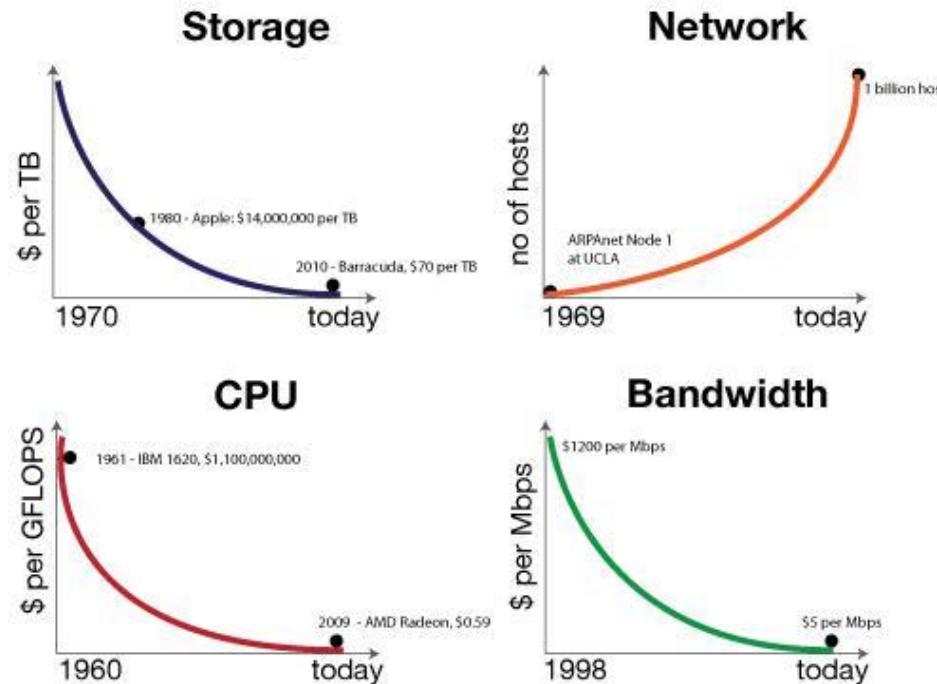
摩尔定律

1.1 信息文明与数据简史



摩尔定律

比特化的基础：计算、存储、网络



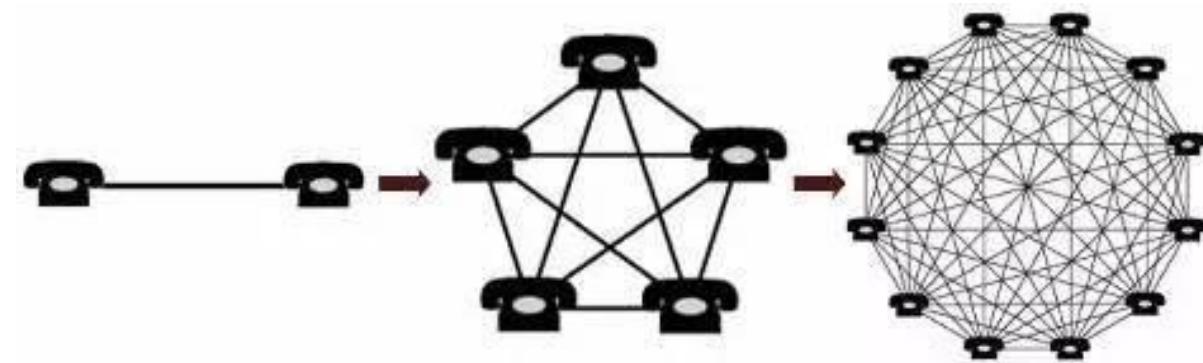
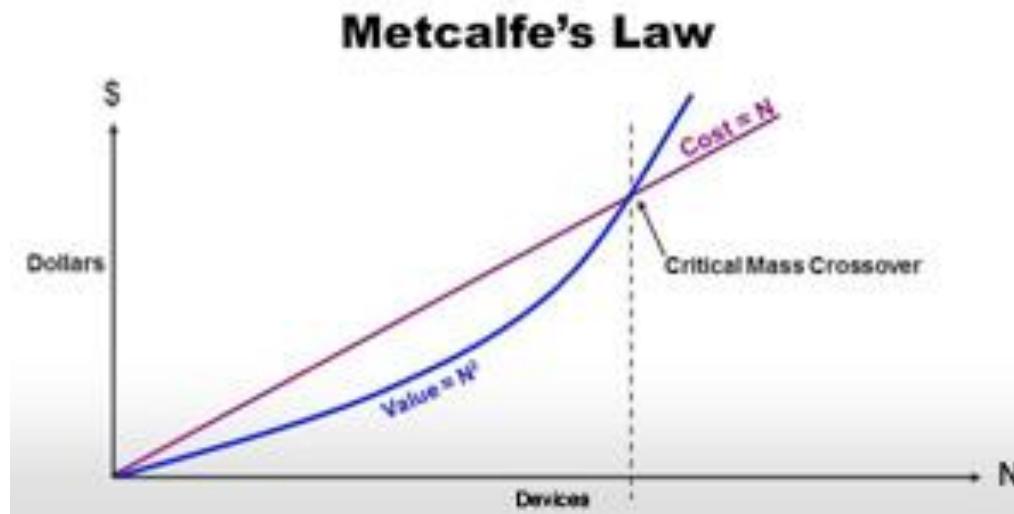
摩尔定律



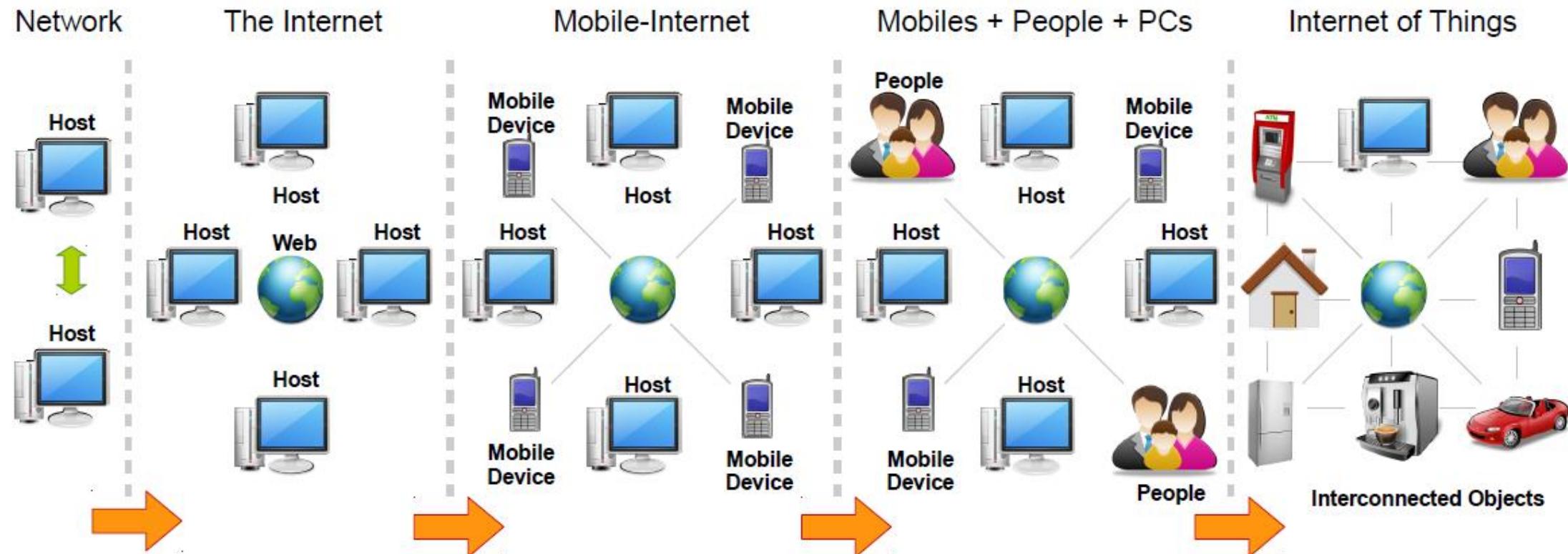
- 元件越来越小、越来越密、越来越快、越来越便宜
- 数据/信息的采集、存储、分析、展示越来越方便

摩尔定律的结果

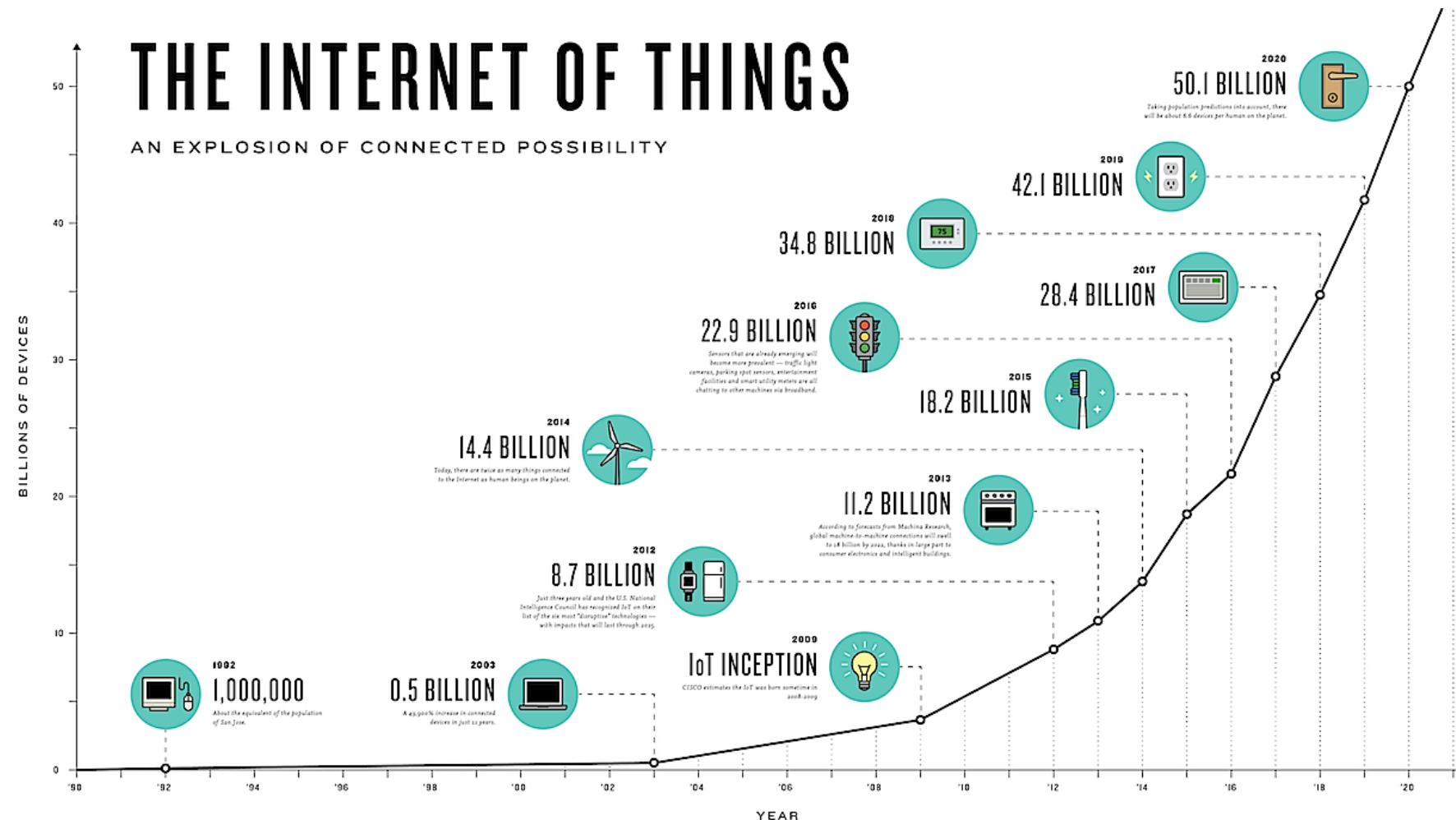
- **梅特卡夫定律：**网络的价值随着用户数量的平方数增加而增加。
- 联网的用户越多，网络的价值越大，联网的需求也就越大。



梅特卡夫定律



物联网



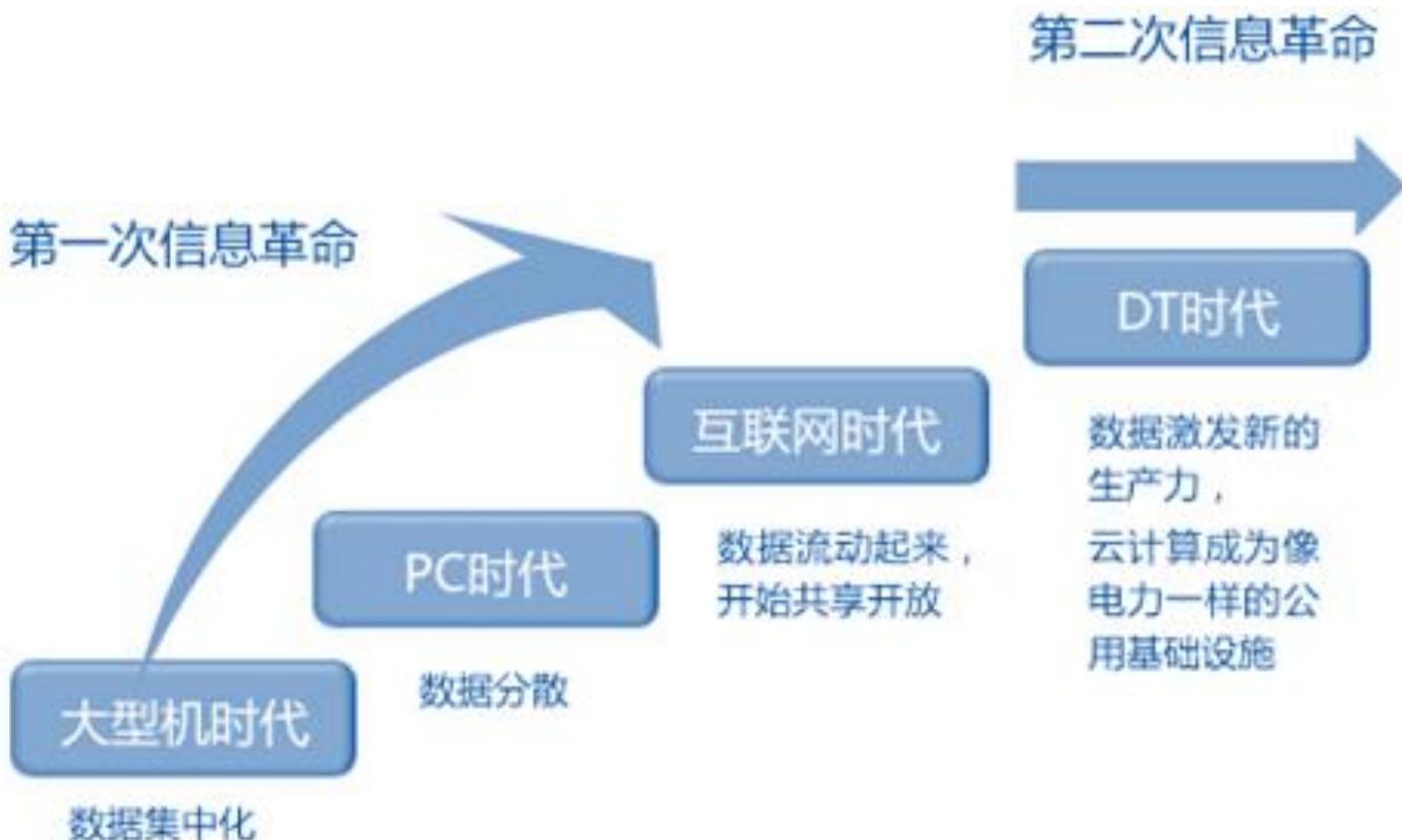
物联网

1.1 信息文明与数据简史

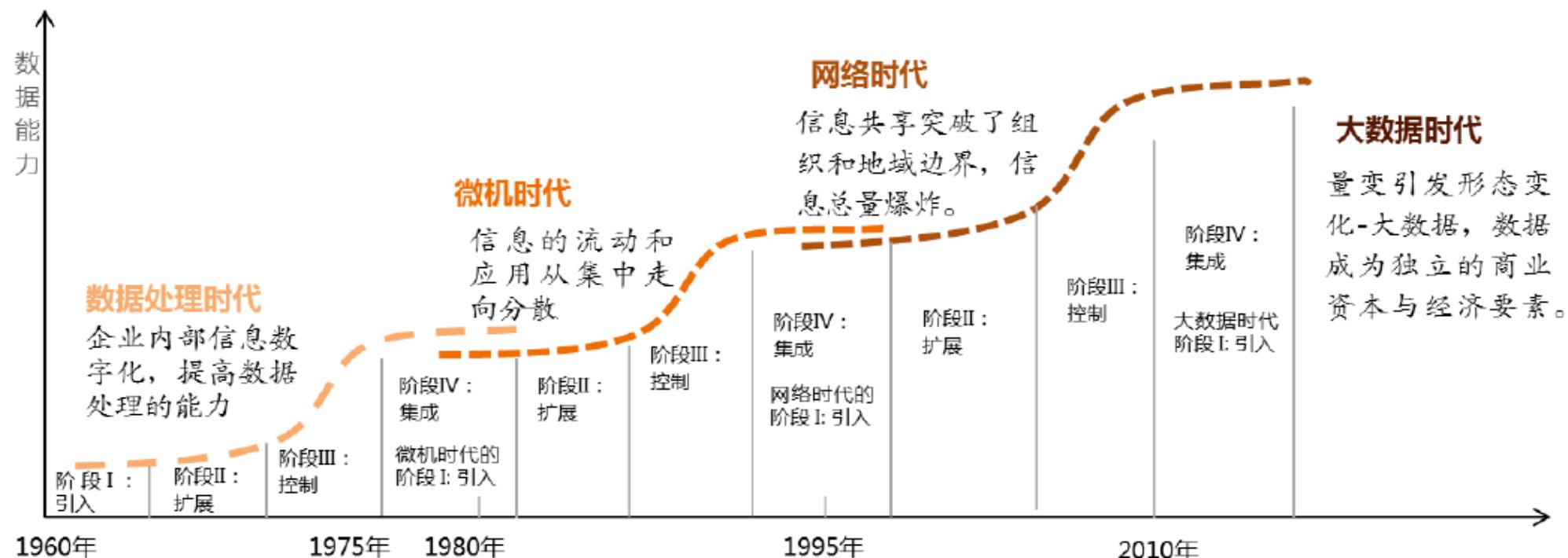
原始文明	农业文明	工业文明	信息文明	人类未来
农业革命 (解放体力)	工业革命 (解放脑力)	信息革命 (解放脑力)	智能革命 (超越脑力)	
采集时代	农耕时代	机械时代	数字时代	智慧时代
人之力	物之力	能之力	算之力	智之力



人类文明的升华与智慧时代



从IT时代到DT时代



由数据驱动的世界观

- 大数据重新定义商业模式
- 大数据重新定义研发新路径
- 大数据重新定义企业新思维

从IT时代到DT时代

1.1 信息文明与数据简史



1944

The acknowledgement of big data was identified by Fremont Rider, Wesleyan University Librarian, who estimated that libraries in American universities would expand to over 200,000,000 volumes by 2040. Today, Yale Library alone has approximately 12.5 million volumes across 20 buildings on campus.



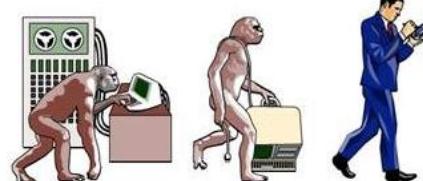
1971

Arthur R. Miller, author of the book 'The Assault on Privacy,' identified that "too many information handlers seem to measure a man by the number of bits of storage capacity his dossier will occupy."



1949

Claude Shannon, known as the "Father of Information," carried out research on big storage capacity on items such as punch cards and photographic data. One of the largest items on Shannon's list was the Library of Congress, measuring over 100 trillion bits of data.



1961

Derek Price's research on scientific knowledge concluded that scientific journals had doubled every 15 years. This is now better known as "law of exponential increase".



1981

The Hungarian Central Statistics Office carried out a research project that is still on-going today. This involved accounting for the country's information for industries via measuring data volume in bits.

1983

Author Ithiel de Sola Pool looked at the growth trends in 17 major communications media from 1960 to 1977, and concluded that the flow of information had exponentially grown by 2.9% throughout that period due to broadcasting and media.



1996

Digital storage became more cost-effective for storing data than paper.

从IT时代到DT时代

1.1 信息文明与数据简史

1997

The term “big data” was used for the first time when researchers M. Cox and D. Ellsworth wrote an article identifying that the rise of data was becoming an issue for current computer systems. In other words, the “problem of big data.”



2000

Peter Lyman and Hal R. Varian published the first study that quantified, in computer storage terms, the total amount of new and original information created in the world annually. The study concluded in 1999, a year in which the world had produced approximately 1.5 exabytes of unique information.



2001

Doug Laney published a research note titled “3D Data Management: Controlling Data Volume, Velocity, and Variety.” A decade later, these “3 V’s” became the defining dimensions of big data.

2005

“What is Web 2.0?” was published by writer Tim O'Reilly, in which he stated that “data is the next Intel inside and SQL is the new HTML. Database management is a core competency of Web 2.0 companies, so much so that they have sometimes referred to these applications as ‘infoware’ rather than merely software.”



2008

Economists Bret Swanson and writer George Gilder published “Estimating the Exaflood” which stated U.S. IP traffic could reach one zettabyte by 2015, and that the U.S. Internet of 2015 would be at least 50 times larger than it was in 2006.

2011

Martin Hilbert and Priscila Lopez, estimated that the world's information storage capacity grew at a compound annual growth rate of 25% per year between 1986 and 2007. They also estimated that in 1986, 99.2% of all storage capacity was analog, but by 2007, 94% of all storage capacity was digital.



2009

Researchers Roger E. Bohn and James E. Short found that modern Americans consumed information for an average of almost 12 hours per day. Consumption totalled 3.6 zettabytes and 10,845 trillion words, corresponding to 100,500 words and 34 gigabytes for an average person on an average day.

Businesses are beginning to implement new in-memory technology such as SAP HANA to analyse and optimize mass quantities of data. Companies are becoming ever more reliant on utilizing data as a business asset to gain competitive advantage, with big data leading the charge as arguably the most important new technology to understand and make use of in order to remain relevant in today's rapidly changing market.

从IT时代到DT时代

第1章 绪论

1
2
3
4
5

信息文明与数据简史

数据科学的基本内涵

第四范式：数据密集型科学

数据科学的应用

实践：以Git和Python为中心

数据学

- Dataology
- 用科学的方法研究数据

数据科学

- Data Science
- 用数据的方法研究科学

数据学科

数据工程

- Data Engineering
- 数据科学的工程实现

数据道德与职业行为准则

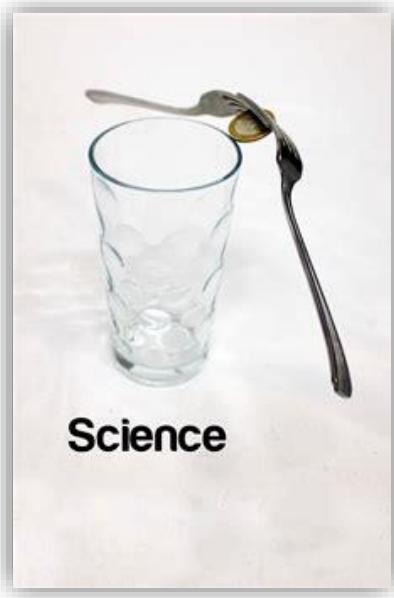
- Data of Ethics & Professional Conduct

什么是数据学科（专业）

- **数据**: 存在于赛博空间
- **信息**: 自然界、人类社会及人类思维活动中存在和发生的现象
- **知识**: 人们在实践中所获得的认识和经验
- 现实自然界 vs 数据自然界

数据学 (Dataology)

数据科学就是以数据为中心的，利用计算思维与数据思维来开展



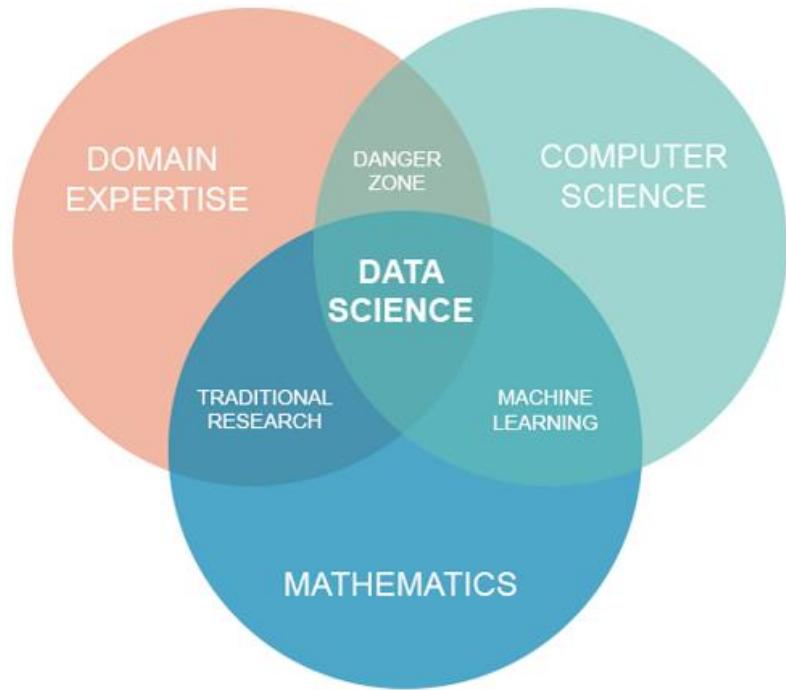
理解世界
_____&
科学方面 问题求解

工程方面



数据科学 (Data Science)

Data science is *interdisciplinary*

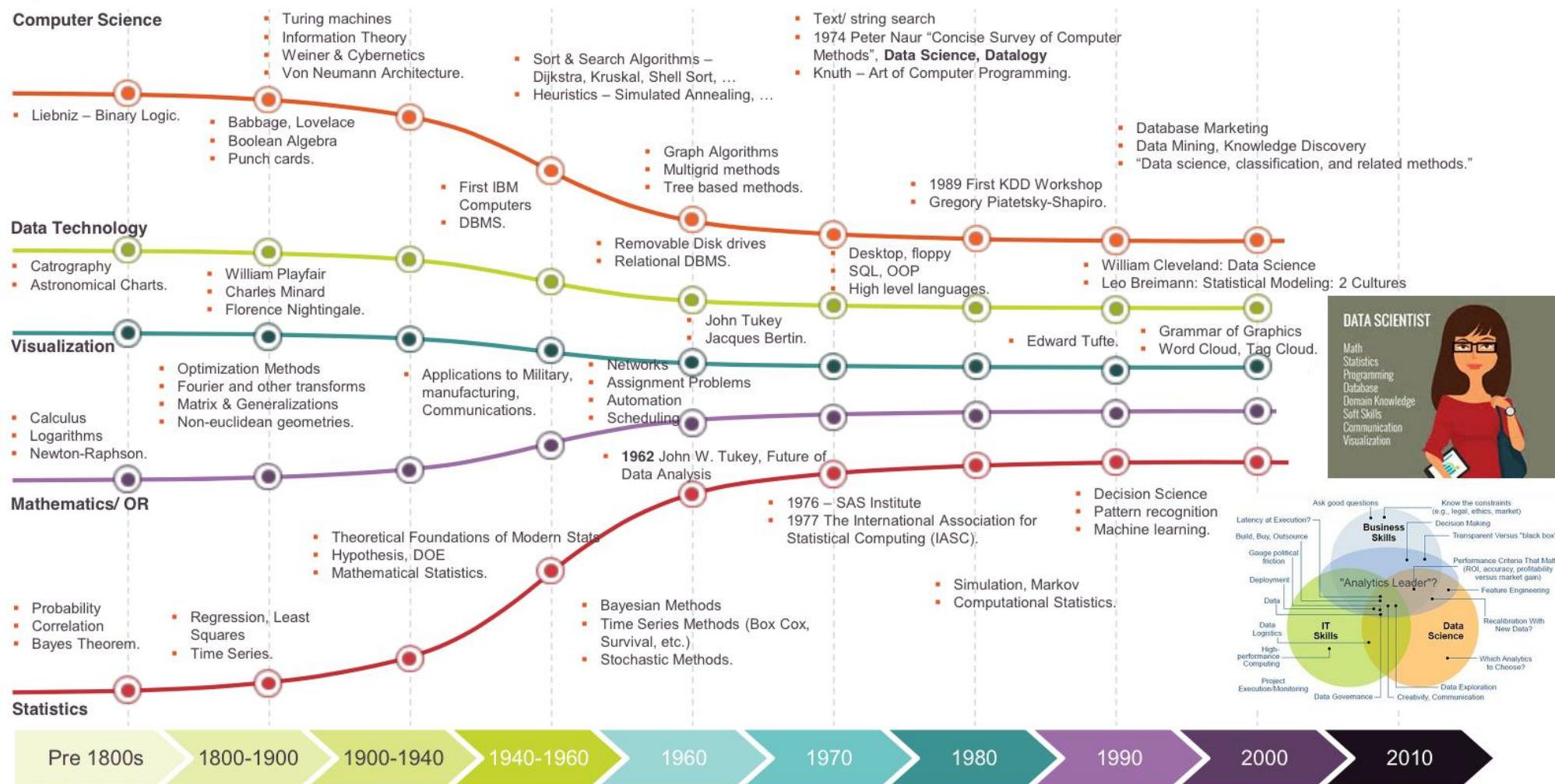


Drew Conway's Venn Diagram of Data Science

More **Union** than
Intersection

数据科学 (Data Science)

1.2 数据科学与工程的基本内涵

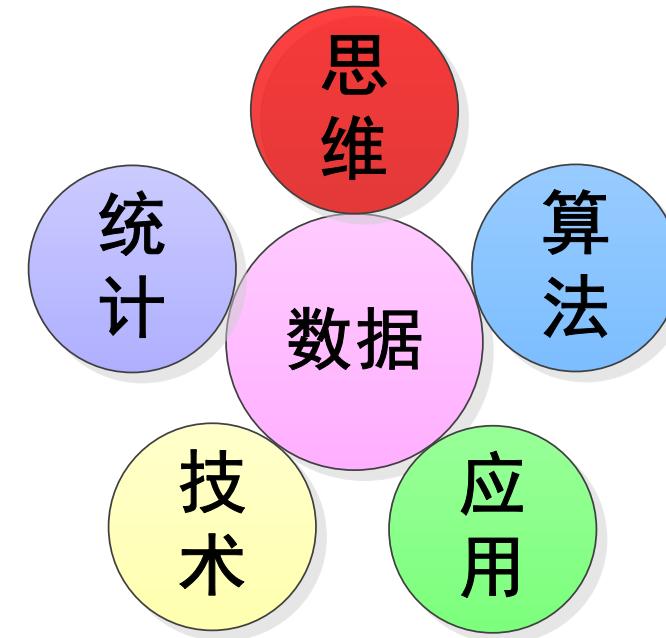
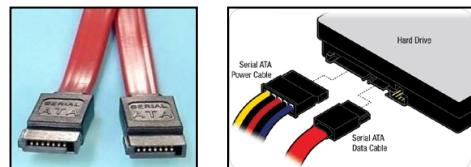


Created by Mamatha Upadhyaya, modified by Mr. Wang

数据科学 (Data Science)

• D-SATA模型

- 数据思维 (Data Thinking)
- 统计模型 (Statistical Model)
- 算法计算 (Algorithmic Computing)
- 数据技术 (Data Technology)
- 综合应用 (Application)



数据科学的五大要素

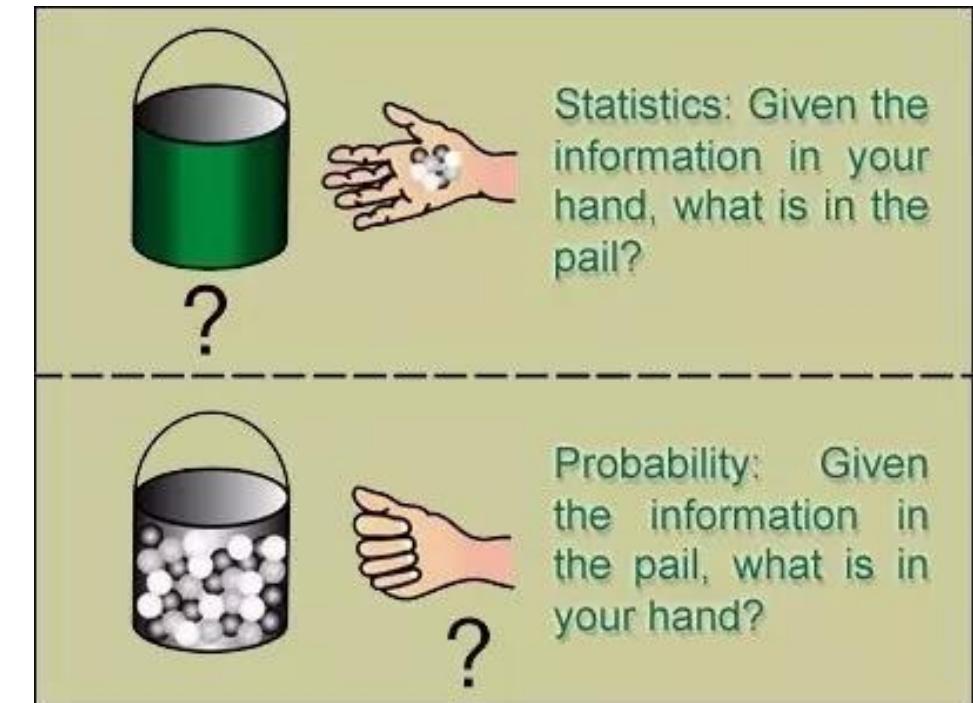
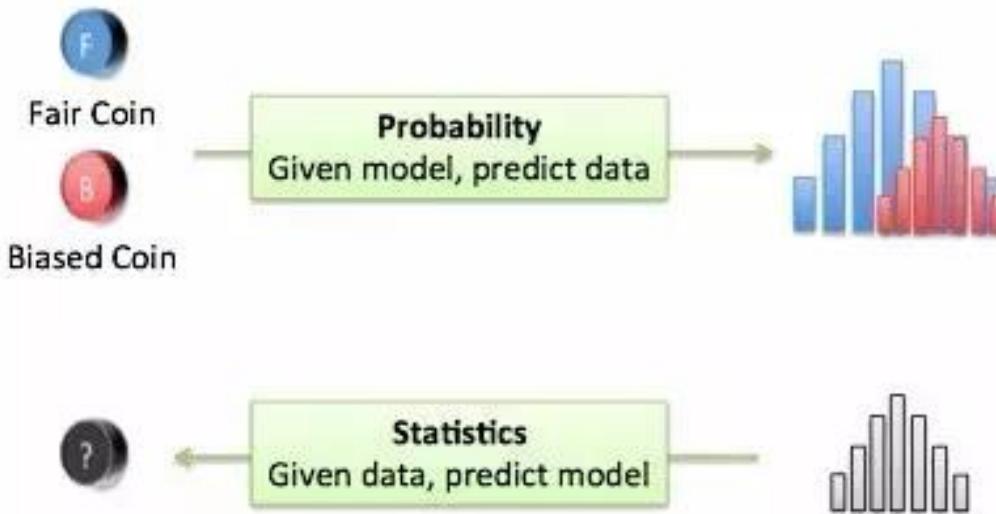
数据思维

- **数据思维** (Data Thinking)
 - 简单来说就是：不是我觉得，而是数据证明；
 - “我觉得”是一种直觉化经验化的思维，数据证明则是数据分析的最直接体现，它依托于数据导向型的思维，而不是技巧，前者是指导，后者只是应用；
 - 以数据为中心的问题求解。
- 在数据科学中的具体体现：
 - **计算思维** (Computational thinking)
 - **统计思维** (Statistical thinking)
 - **设计思维** (Design thinking)

数据科学的五大要素

统计模型

Probability & Statistics



数据科学的五大要素

算法计算

程序 = 数据结构 + 算法

- 数据结构是用来干什么的？装数据的，使得算法能在上面做有效的计算。
- **数据是信息之源泉**，没有数据就没有信息，没有大量数据就没有有效信息。
- 而**算法**是把信息从数据中提取出来的手段，没有算法的提取，数据还是数据，永远不会变成有用的信息。
- 因此，数据和算法的关系应该是相辅相成，融为一体的。

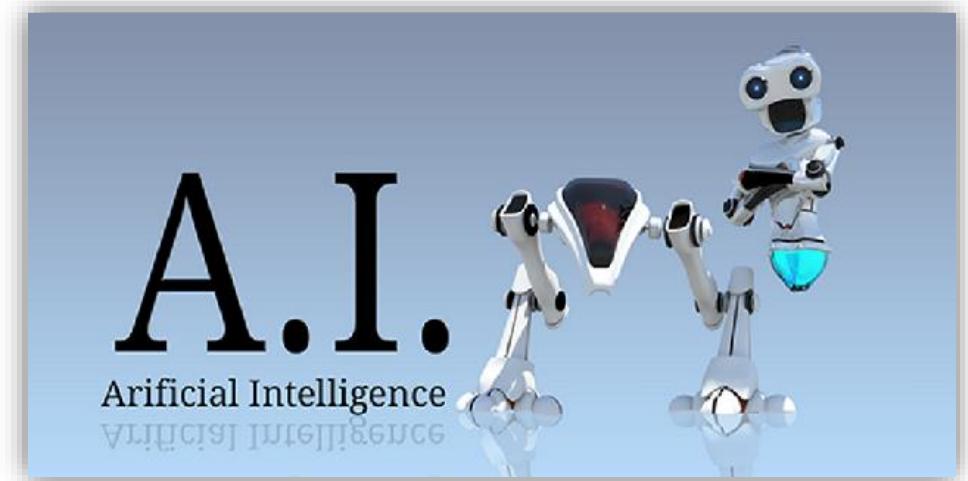
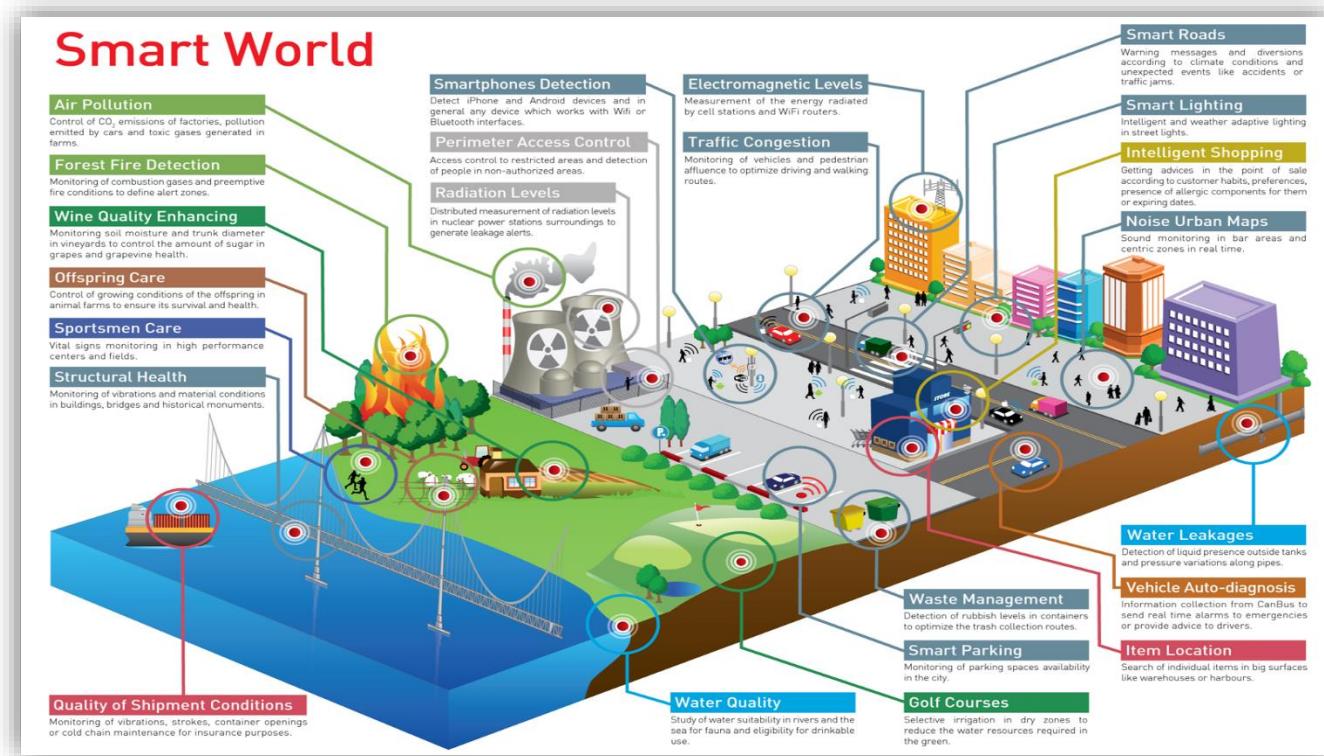
数据科学的五大要素

数据技术

ETL 数据装载工具	Workflow 工作流开发工具	数据质量管理工具	可视化 报表工具	机器学习 建模工具	统计挖掘 开发工具	资源 管理工具	分析管理工具
SQL批处理 Batch Processing	交互式分析 OLAP Analysis	实时数据库 OLTP Transactional Processing	数据挖掘 机器学习 算法库 / 框架 Machine Learning	深度学习 Deep Learning	图分析引擎 Graph Analysis	流处理引擎 Streaming Processing	应用级引擎
批处理框架 Map/Reduce, Tez		高性能处理框架 Spark			向量处理框架 TensorFlow		
短时任务资源管理框架 YARN		长时任务资源管理框架 Mesos			资源隔离 / 调度 / 管理框架 Kubernetes		
分布式文件系统 HDFS	分布式大表 HBase	搜索引擎 Elastic Search		分布式缓存 Redis	消息队列 Kafka	分布式协作服务 Zookeeper	分布式存储引擎

数据科学的五大要素

综合应用



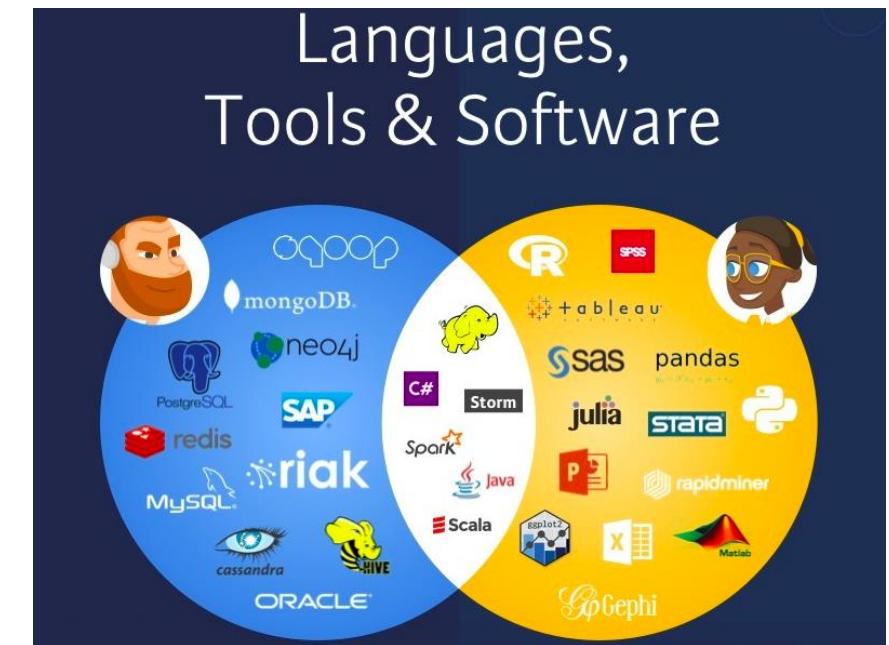
数据科学的五大要素

- **数据思维**: 包括计算思维、统计思维、设计思维等;
- **数学基础**: 微积分、线性代数、概率统计、离散数学等;
- **数据建模与评估**: 统计模型、回归模型、模型评估等;
- **算法实现**: 问题求解能力和算法设计;
- **数据管理**: 设计数据的整个生命周期, 包括感知、存储、计算、分析、可视化等;
- **知识转化**: 沟通交流, 道德规范等。

数据科学的核心知识点

支持数据学和数据科学的研究和活动的工程实现，包括：

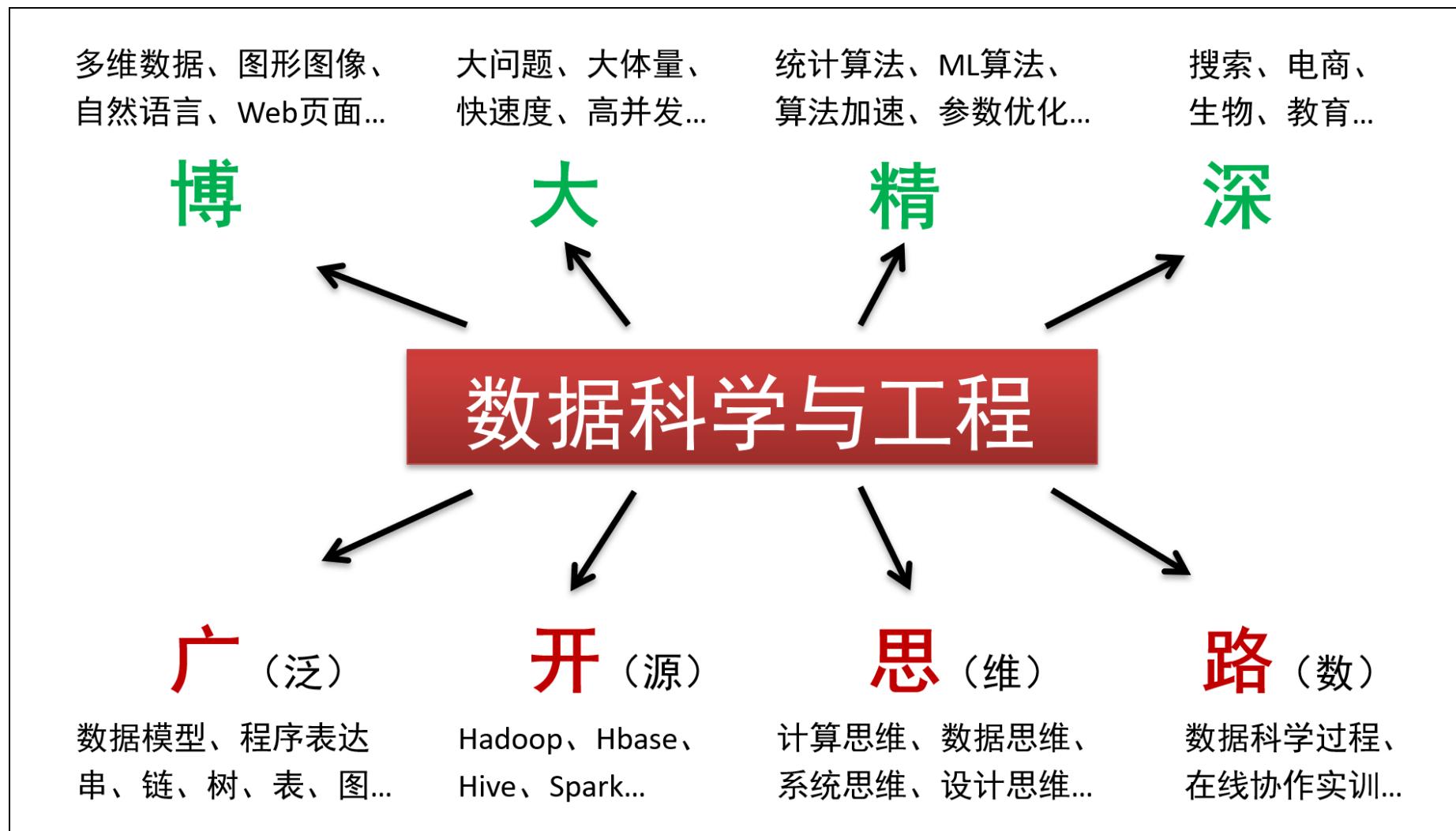
- 数据基础设施
- 数据全生命周期管理过程
- 数据科学过程方法论和工具
- 数据处理与分析系统
- 数据分析编程语言
- 可视化工具
-



数据工程 (Data Engineering)

- 数据隐私与安全
- 数据道德与数据伦理
- 开放数据
- 数据关联的社会问题
- 数据相关的职业规划
-

数据道德与职业行为准则



数据科学与工程的挑战与应对

第1章 绪论

1
2
3
4
5

信息文明与数据简史

数据科学与工程的基本内涵

第四范式：数据密集型科学

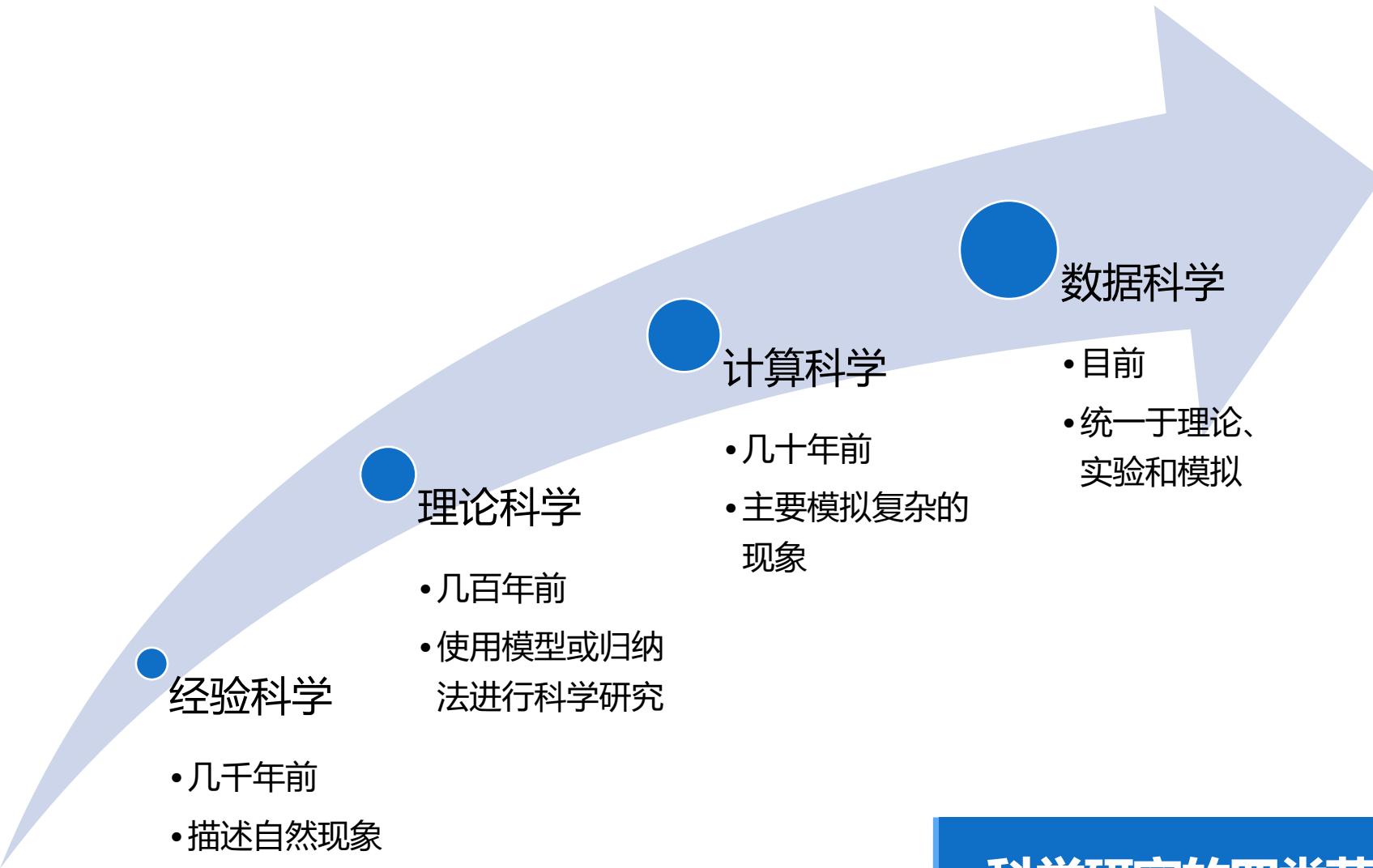
数据科学与工程的应用

实践：以Git和Python为中心

- “范式”（paradigm）这一概念最初由美国著名科学哲学家 Thomas Samuel Kuhn于1962年在《科学革命的结构》中提出来。
- 指的是常规科学所赖以运作的理论基础和实践规范，是从事某一科学的科学家群体所共同遵从的世界观和行为方式。

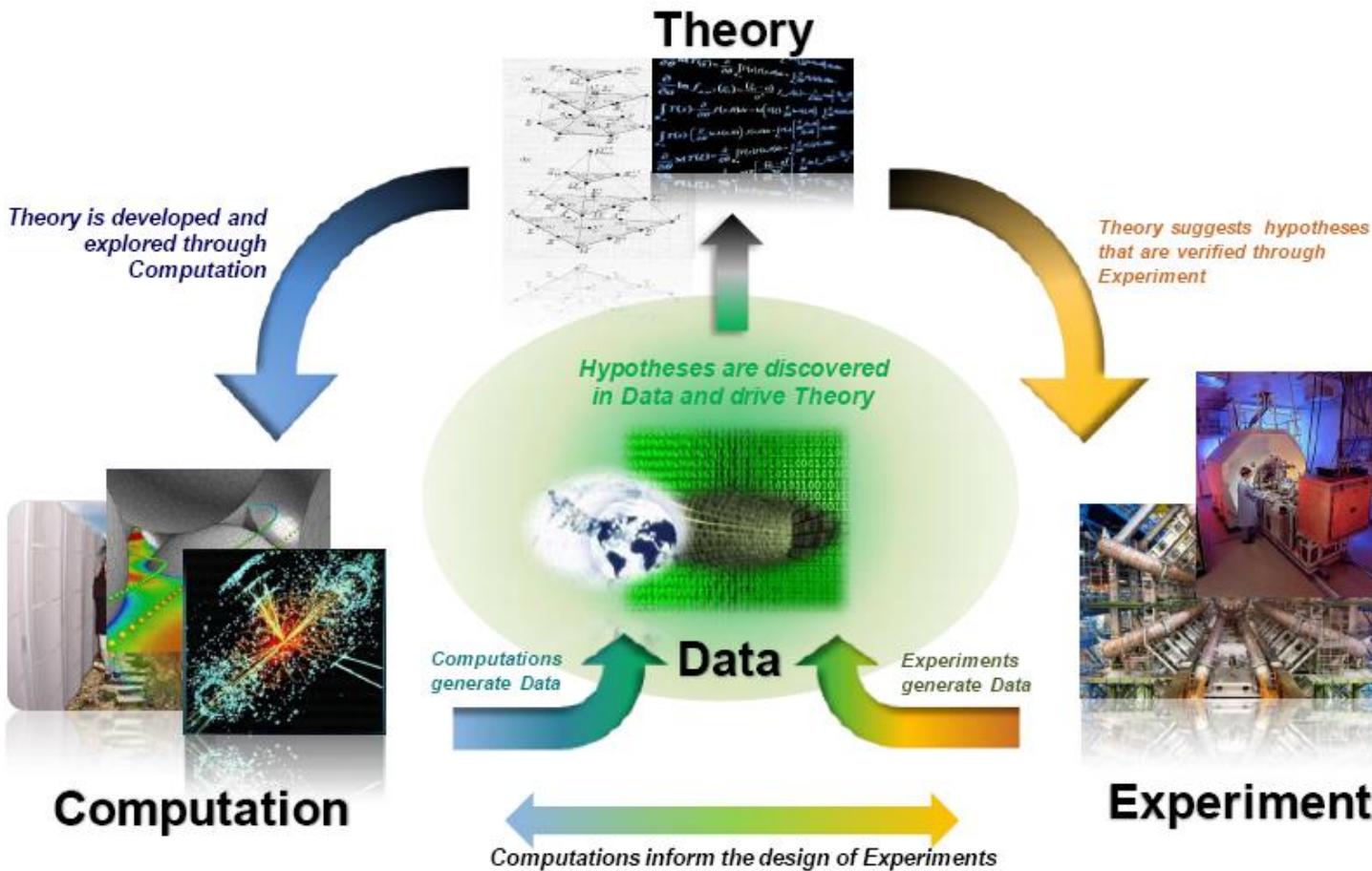
什么是科学范式

1.3 第四范式：数据密集型科学



科学研究的四类范式

1.3 第四范式：数据密集型科学



科学研究的四类范式

1.3 第四范式：数据密集型科学



科学研究的四类范式

第1章 绪论

1
2
3
4
5

信息文明与数据简史

数据科学的基本内涵

第四范式：数据密集型科学

数据科学的应用

实践：以Git和Python为中心

1.4 数据科学与工程的应用



医疗



宇宙



娱乐



市民公用服务



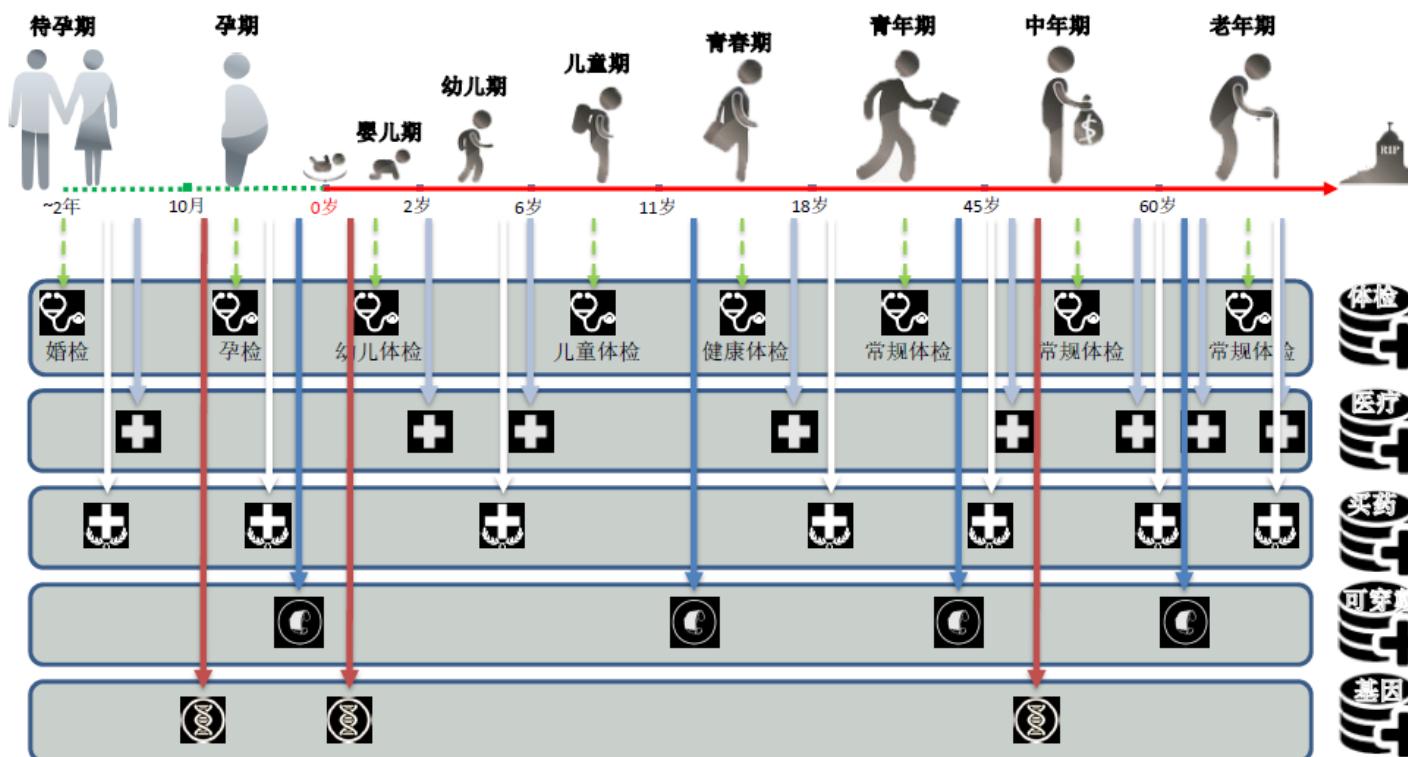
商业



网络空间安全

典型应用领域

- 人们可以收集包括病例、治疗方案、病人基本特征在内的针对疾病特点的数据来建立数据库。协助医生更快捷准确的制定医疗方案，帮助更多人即使进行治疗。同时也有利于医药行业开发出更加有效的药物和医疗器械。



数据医疗

数据交通，平安畅行无阻

- 大数据使智能交通的潜在价值得到有效挖掘，通过对交通信息的感知和收集，对各个管理系统中海量数据的共享运用、有效分析，对交通态势的研判预测等，能大大提高智能交通的智能化。



智慧城市

数据食品，餐桌上的安全

- 通过大数据管理将海量数据聚合在一起，将离散的数据需求集合能形成长尾，从而满足传统中难以实现的需求。



智慧城市

数据调控，对症才好下药

- 未来大数据将会从各个方面帮助政府实施高效和精细化管理。政府运作效率的提升，决策的科学客观，财政支出合理透明都将大大提升国家整体实力，成为国家竞争优势。大数据带给国家和社会的益处将会极大的想象空间。



智慧城市

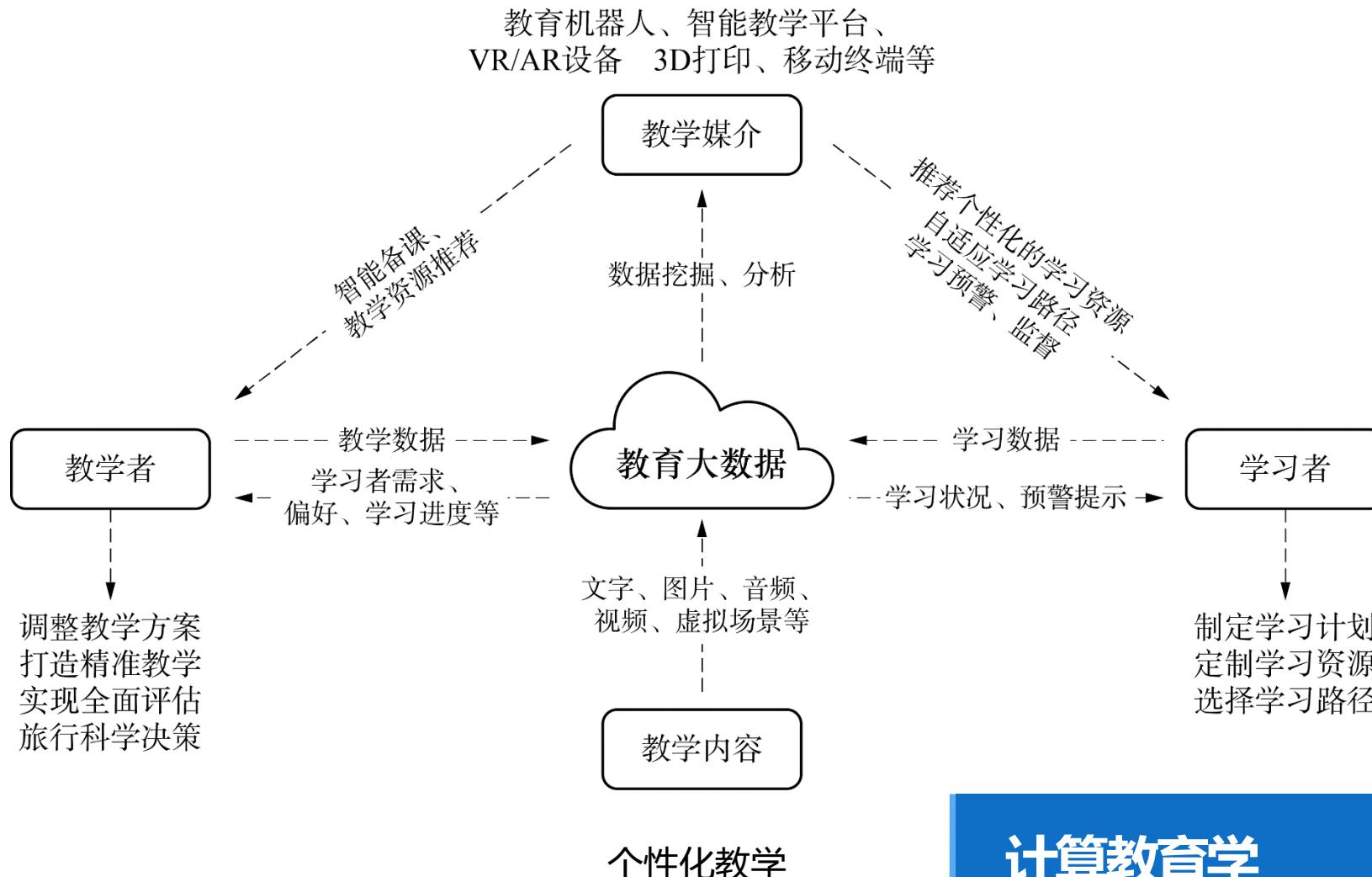
数据舆情， 请叫我上帝

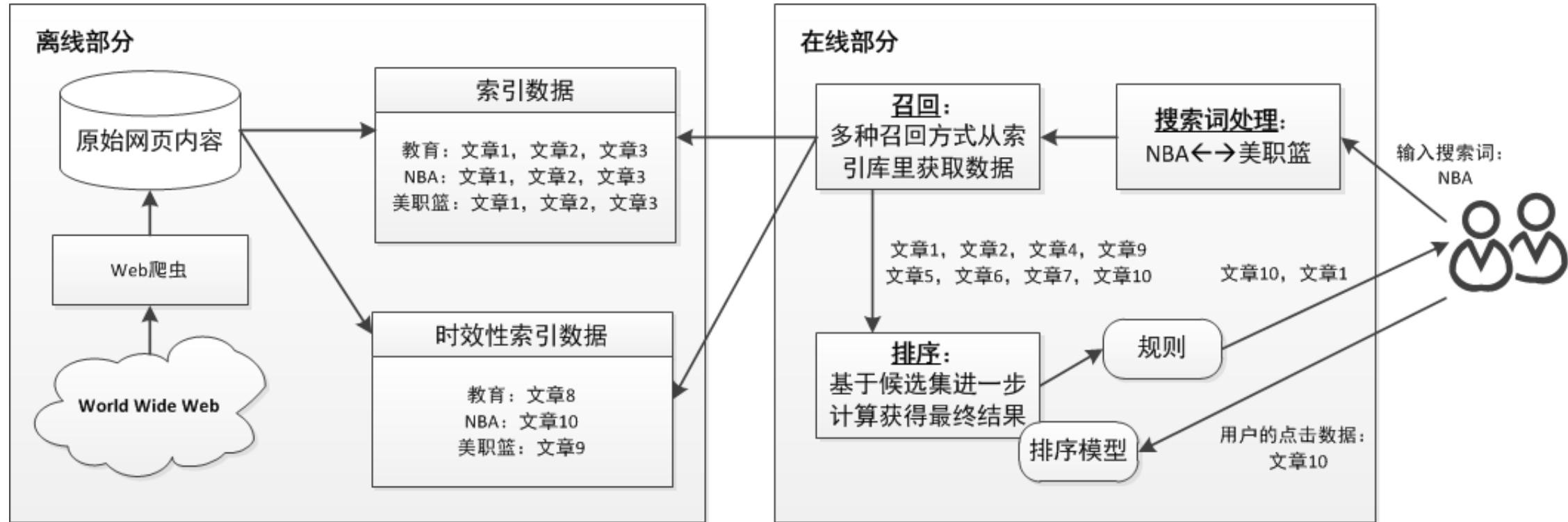
- 监管部门可以通过大量的多类型数据， 创建一张犯罪高发地区热点图。同时， 还将相邻片区等各种因素加入到数据模型中，并根据历史罪犯记录的地点统计不断修正所得出的预测数据。



智慧城市

1.4 数据科学与工程的应用





搜索引擎

■ 推荐系统成为连接数据特征与用户需求的桥梁

- 普通用户难以直接从大数据中获取所需信息
- 推荐系统将大数据从单纯的数据层面转化到用户可以理解的信息层面，满足客户的需求

■ 推荐系统为企业带来**巨大价值**

- 2016年“双十一”淘宝交易额破千万
- 今日头条app人均日使用时间超40分钟
- UC浏览器个性化推荐月活跃用户超过3.3亿

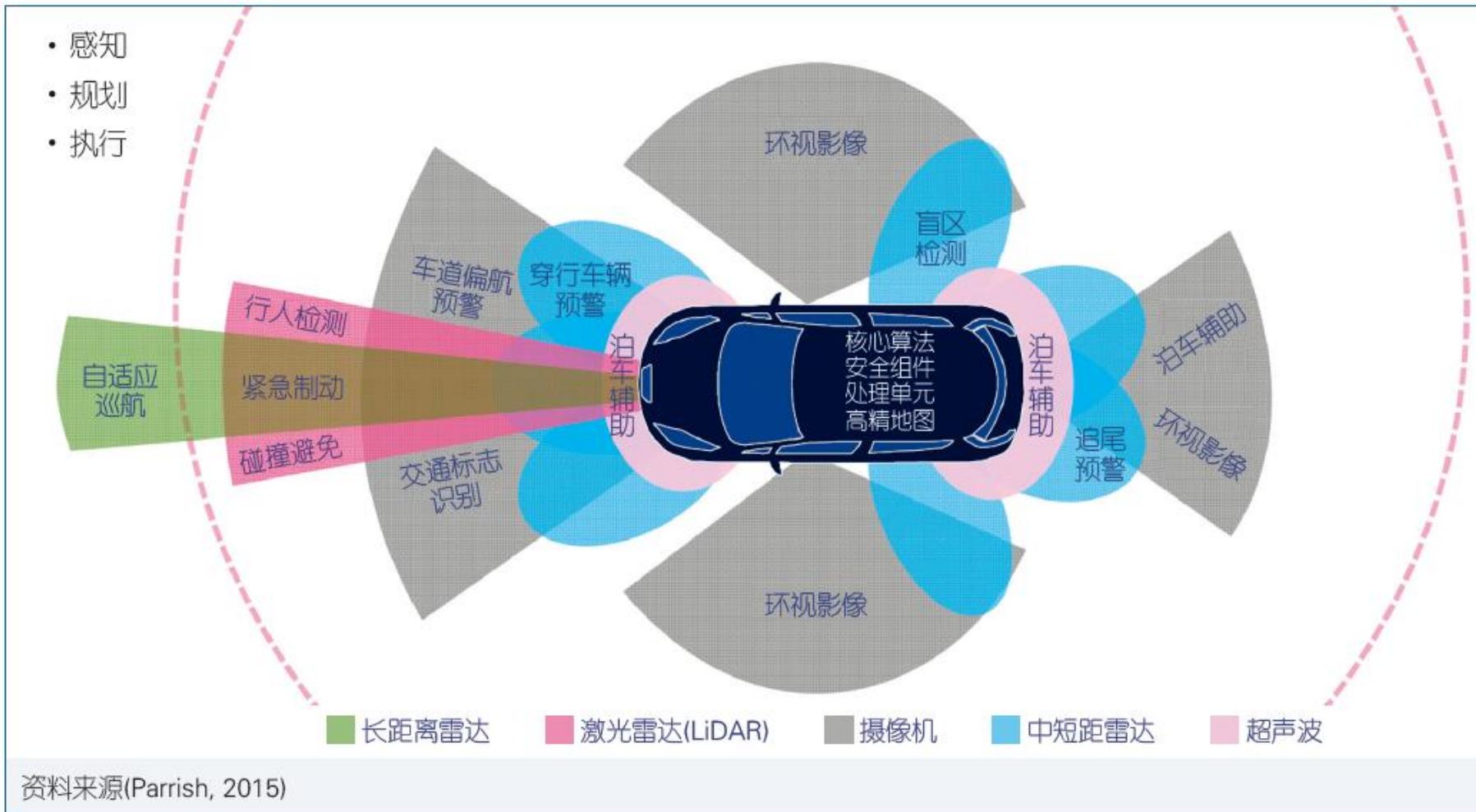
我们正在离开信息的时代，进入推荐的时代。

——Chris Anderson in *The Long Tail*



推荐系统

1.4 数据科学与工程的应用



无人驾驶

第1章 绪论

1
2
3
4
5

信息文明与数据简史

数据科学的基本内涵

第四范式：数据密集型科学

数据科学的应用

实践：以Git和Python为中心

- Git（发音：/git/）是一个分布式版本控制软件，最初由林纳斯·托瓦兹创作，于2005年以GPL发布。最初目的是为更好地管理Linux内核开发而设计。
- **Tool:** Distributed Version Control System
- **Directory:** Content Management System
- **Tree:** history storage system
- Stupid content tracker
- **Git is SUPER cool**



Git 简介

- GitHub是一个面向开源及私有软件项目的托管平台，因为支持Git作为唯一的版本库格式进行托管，故名GitHub。
- GitHub于2008年4月10日正式上线，除了Git代码仓库托管及基本的Web管理界面以外，还提供了订阅、讨论组、文本渲染、在线文件编辑器、协作图谱（报表）、代码片段分享（Gist）等功能。
- 目前，其注册用户已经超过百万，托管版本数量也非常多，其中不乏知名开源项目Ruby on Rails、jQuery等。



GitHub 简介

GitHub流行的原因

面向协作，提供良好的分享和协作体验，极大降低了参与协作的门槛

1. 标准的沟通方式
2. 标准的代码管理方式
3. 标准的 Web 社交方式
4. 开源非常酷



GitHub 简介

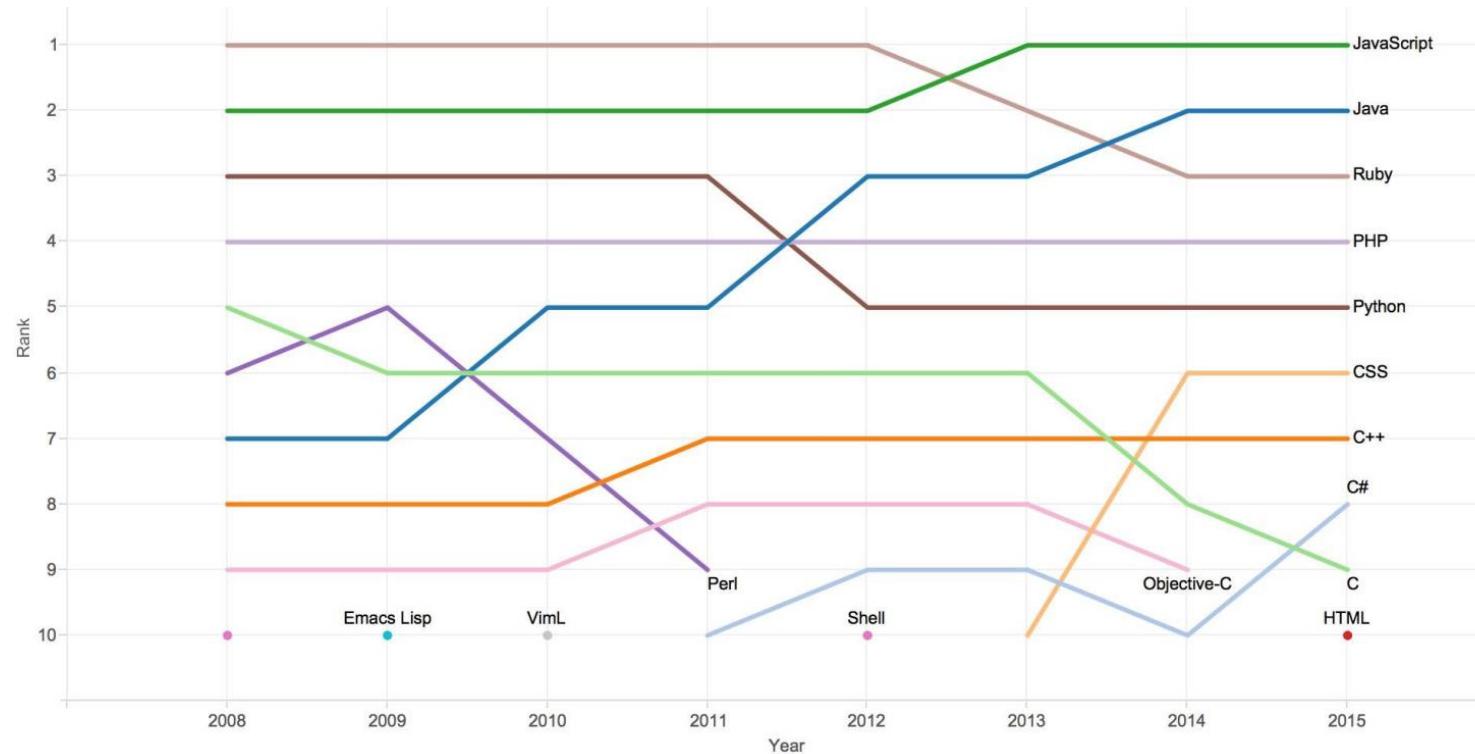


- Python（发音： /'paɪθən/ ）是一种广泛使用的解释型、高级编程、通用型编程语言，由吉多·范罗苏姆创造，第一版发布于1991年。
- Python是一种面向对象、解释型的计算机程序语言。
- 它的语法简单，并且包含了一组功能完备的标准库，能够轻松完成很多常见的任务。

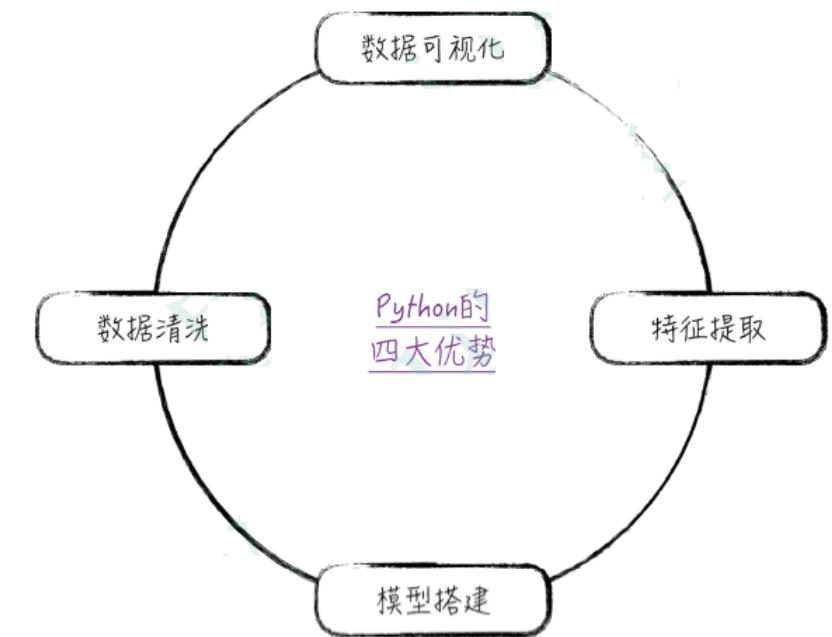


Python 简介

1.5 实践：以Git与Python为中心



编程语言流行排行榜



Python在数据科学中的地位

Python 简介

Thanks!

立即体验

