第一章 绪论

数据

1.数据的定义：数据通常是为了**分析**和**决策**而收集的统计资料和观测值。

2.数据的类型：数据可以是**定性的**（例如描述性的，如“味道好闻”）或**定量的**（例如具体的数值，如体重70kg）。

3.数据的用途：数据为**推理**和**计算**提供基础。这表明数据是**分析**、**逻辑推理**和**数学计算**的基本要素。

4.数据的来源和普遍性：数据可以来源于多种渠道，如**日常交易**、**传感器记录**、**互联网活动**等，而且它们在我们的日常生活中无处不在。

尽管通过增加服务器存储容量来存储这些庞大的数据相对容易实现，但对这些大量数据进行**有效管理**和**深入分析**却是一个复杂且具有挑战性的任务。

为了系统地研究和分析数据，深入理解其含义，并将这些信息作为决策制定和问题解决的有效工具，**数据科学**应运而生。

数据科学

- 数据科学是一个**综合性学科**，它融合了**统计学**、**计算机科学**、**信息科学**以及相关学科的技术和理论，专注于从**结构化**和**非结构化数据**中提取知识。它的目标是通过一系列过程，包括数据收集、清洗、探索、建模和分析，来发现数据中的模式和关联，并将这些发现应用于预测分析和决策支持。此外，数据科学还强调数据的呈现，确保分析结果能够被理解和利用。

Data acquisition

Data cleaning

Data preprocessing

Data representation

Data evaluation

Data analysis

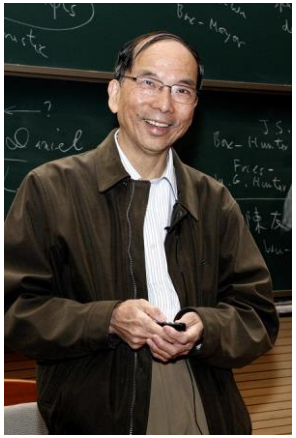
数据科学的历史

- 1962年，约翰·图基首次描述了一个在方法论和应用上与现代数据科学相似的领域，他称之为“数据分析”。
- 1974年，“数据科学”这个术语首次出现在彼得·诺尔的工作中，他提出将其作为计算机科学的替代术语。
- 1997年，C.F. Jeff Wu在一次演讲中提出了将统计学重命名为数据科学的想法，这体现了统计学与新兴数据处理技术之间的融合。
- 2001年，威廉·S. 克利夫兰提出了数据科学作为独立学科的概念，这个学科强调数据分析、计算和信息技术的交叉。
- 2008年，DJ Patil和Jeff Hammerbacher被认为是首次为自己的职位使用“数据科学家”这一头衔，标志着这一角色在工业界的正式诞生。

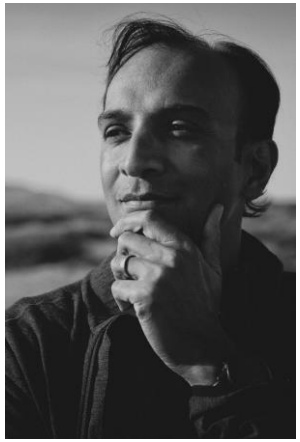
约翰·图基

彼得·诺尔

威廉·S. 克利夫兰



Jeff Wu



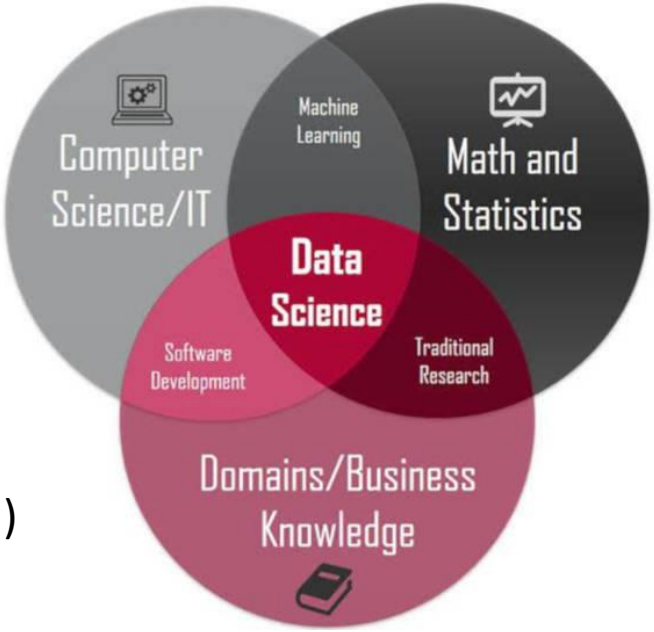
DJ Patil

数据科学家

- 数据科学家1）在统计学方面的能力超越了普通软件工程师，2）在软件工程的技能上也超出了一般统计学家。这样的跨学科专长要求数据科学家不仅**精通数据分析所需的统计技术**，同时也能够**通过编程实践这些技术**。此外，3）他们还需对其服务的行业有深入的理解。

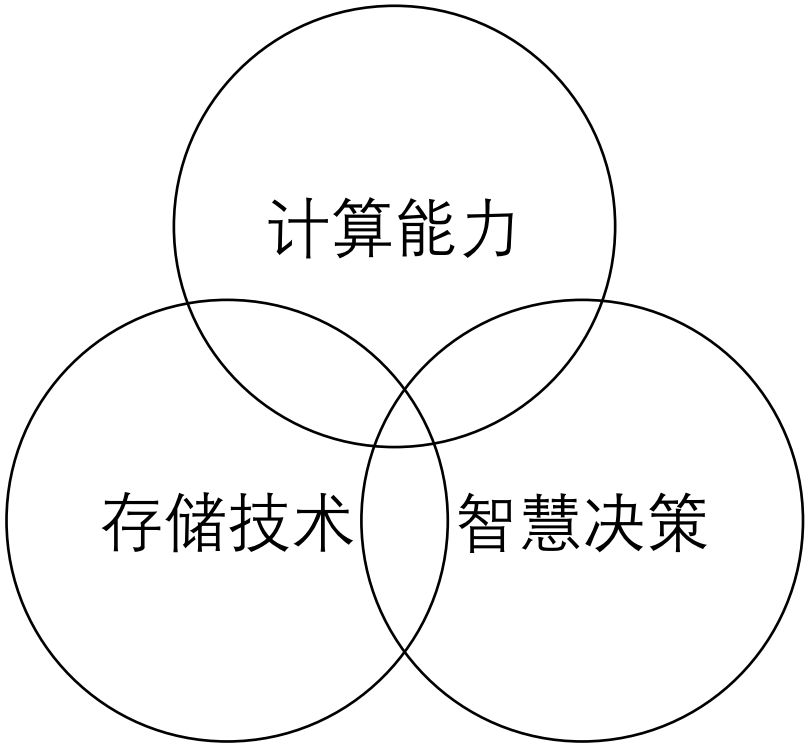
数据科学家的工作（来自一位知乎网友）：

- 商业逻辑理解与思考（占10%时间）
- 数据检查与清洗（占25%时间）
- 特征工程（占20%的时间）
- 数据建模（占5%时间）
- 与客户,同事，或者上级沟通（占20%时间)
- 写模型文档，数据分析文档等。（占15%时间)



从数据到大数据

- 大数据指的是分布在多个系统上的大规模、非集中化的原始数据集。这些数据以高速度从各种来源产生，并且呈现出多样的格式和结构。

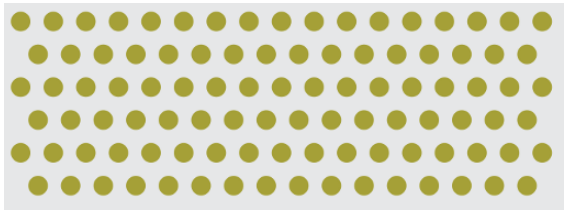


- 存储成本的下降
- 运行、计算速度越来越快
- 对人工智能的追求
- 传感器的发展
- 移动互联
- 商业需要
- ...

大数据的四个维度

- 四个维度（4”V”： Volume, Variety, Velocity, Veracity）

体量（Volume）

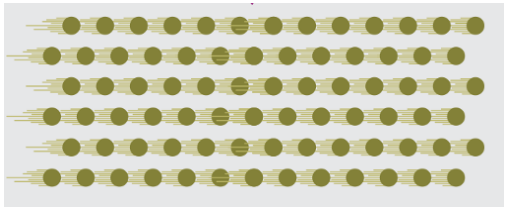


数据量庞大，TB至PB级数据

多样性（Variety）

结构化，非结构化，
文本，多媒体

速度（Velocity）



数据生成和处理的速度

真实性（Veracity）

数据的质量和可信度。

大数据的四个维度

- **Volume（体量）**

体量是大数据的一个核心特征，它指的是**巨量的数据。**然而，定义'巨大的数据量'是相对的，这在不同的行业、领域和地区之间会有所不同。随着技术的进步，今天被视为巨大的数据量在未来可能会显得更加普遍。

- 伴随着各种随身设备、物联网和云计算、云存储等技术的发展，人和物的所有轨迹都可以被记录，数据因此被大量生产出来。
- 移动互联网的核心网络节点是人，不再是网页，人人都成为数据制造者，短信、微博、照片、录像都是其数据产品。
- 数据来自无数自动化传感器、自动记录设施、生产监测、环境监测、交通监测、安防监测等。
- 数据来自自动流程记录，刷卡机、收款机、电子不停车收费系统，互联网点击、电话拨号等设施以及各种办事流程登记等。
- 大量自动或人工产生的数据通过互联网聚集到特定地点，包括电信运营商、互联网运营商、政府、银行、商场、企业、交通枢纽等机构，形成了大数据之海。

大数据的四个维度

- **Variety（多样性）**

大数据的多样性指的是大数据具有各种不同类型的数据，包括结构化数据（如数据库中的数据）、半结构化数据（如电子邮件）和非结构化数据（如视频和图片）。

- 北京市交通智能化分析平台的数据不仅来自于路网的摄像头和传感器，还包括公交、轨道交通、出租车等多种交通工具的运营数据。例如，4万辆浮动车每天产生2000万条记录，而交通卡的刷卡记录、手机定位数据、出租车运营记录以及电子停车收费系统数据都在每天产生数以百万计的数据点。这些数据种类繁多，覆盖交通领域的各个方面。
- 语音助理可以处理的数据类型包括语音、文字、用户位置、搜索偏好等，还会根据用户的默认家庭地址或当前位置来过滤搜索结果。
- ChatGPT可以接受的数据类型包括文字，语音，图片。所以这是一个**多模态**的大模型

大数据的四个维度

- **Velocity（速度）**

数据的创建、处理和分析速度要足够“快”。“快”主要是由于数据的实时生成和业务流程及决策中的迫切需求。速度的提升能够缩短数据的时延。对于那些时间敏感的业务至关重要，如实时欺诈监测或高频交易。

- 汽车导航系统能够**实时**响应交通状况变化，快速重新规划行车路线，确保驾驶者能够最有效率地抵达目的地。
- 航空公司根据航班的搜索量和余票情况，能够**实时**调整票价，以优化利润。
- 酒店通过**即时更新**多个在线营销渠道（如携程、飞猪、去哪儿）上的房间信息，确保各平台显示的剩余房间数量保持一致。
- 银行利用**实时监测**系统来迅速识别并响应信用卡盗刷行为，保障客户资金安全。
- 公安机关通过**实时监测**系统，快速识别并处理电信诈骗活动，保护公众不受欺诈。

大数据的四个维度

- **Veracity（真实性）**

  数据真实性。大数据的真实性（Veracity）指的是数据的准确性、可信度和可靠性。这一特征强调的是数据本身的质量，包括数据的来源、完整性和上下文的正确性。在大数据环境中，因为数据量大且来源多样，数据的质量可能参差不齐。

- 电商平台聚集了成千上万的消费者评论和评分。这些数据是**非结构化**的，包括文本评论和星级评分。大数据的真实性挑战在于如何确保这些评论和评分是真实可靠的，而不是虚假或误导性的。例如，一些评论可能来自真正的消费者，而另一些可能是由制造商或竞争对手伪造的。

- 在社交媒体平台上，用户发布的内容、点赞、分享和评论等构成了海量的数据。市场研究人员和品牌经理利用这些数据来分析消费者行为和市场趋势。然而，数据真实性的挑战在于确保所分析的数据能够代表真实的用户意见和行为，而不是由虚假账户或自动化的文本生成工具产生的误导信息。（这在文本大模型发展的今天尤其重要）

大数据的来源

- **社交数据**来源包括微信聊天记录、抖音短视频、各种评论、搜索记录以及通过社交媒体平台上传和分享的一般媒体内容。
- **机器数据**是指由工业设备、安装在机械上的传感器以及追踪用户行为的网络日志所生成的信息。这些数据的来源包括医疗设备、智能穿戴设备、交通摄像头、卫星、汽车、智能家居、物联网设备等。
- **交易数据**是指在线上和线下发生的所有日常交易所生成的数据。这些数据包括发票、付款订单、存储记录、交货收据等。
- **网络数据**主要来源于各种静态网页，
- **数据库数据**包括医疗数据库、学术数据库等。
- **政府和公共数据**：来自政府部门和公共机构的数据，如人口普查数据、经济统计、交通流量数据、公共健康数据等。
- **卫星和遥感数据**：这些数据主要来自于卫星、无人机或其他远程感测技术，广泛用于地理信息系统、环境监测、农业、气候变化研究等领域。
- **生物医学和基因组数据**：医疗保健领域产生的大量数据，包括临床试验数据、病人健康记录、基因组数据等。

大数据涉及到的技术

数据管理

- 数据清洗
- 数据描述
- 数据评估
- 数据发布
- 数据存取，使用和安全

数据流

- 数据收集
- 数据存储
- 数据访问

数据分析

- 统计分析
- 模拟和建模
- 可视化技术
- 具体应用方法

大数据的关键处理步骤

2011年麦肯锡全球研究所的报告描述了大数据的主要组成部分和生态系统，主要包括：：

1.数据分析技术：如机器学习、自然语言处理等用于解析数据的高级技术。

2.大数据技术：包括云计算、数据库等支撑大数据处理和存储的技术。

3.数据可视化：通过图表、图形及其他数据展示方式，直观地呈现数据分析结果。

大数据涉及到的技术

数据管理

- 数据清洗
- 数据描述
- 数据评估
- 数据发布
- 数据存取，使用和安全

数据流

- 数据收集
- 数据存储
- 数据访问

数据分析

- 统计分析
- 模拟和建模
- 可视化技术
- 具体应用方法

大数据的关键处理步骤

2011年麦肯锡全球研究所的报告描述了大数据的主要组成部分和生态系统，主要包括：：

1.数据分析技术：如机器学习、自然语言处理等用于解析数据的高级技术。

2.大数据技术：包括云计算、数据库等支撑大数据处理和存储的技术。

3.数据可视化：通过图表、图形及其他数据展示方式，直观地呈现数据分析结果。

大数据的应用

1. **理解客户、满足客户服务需求**

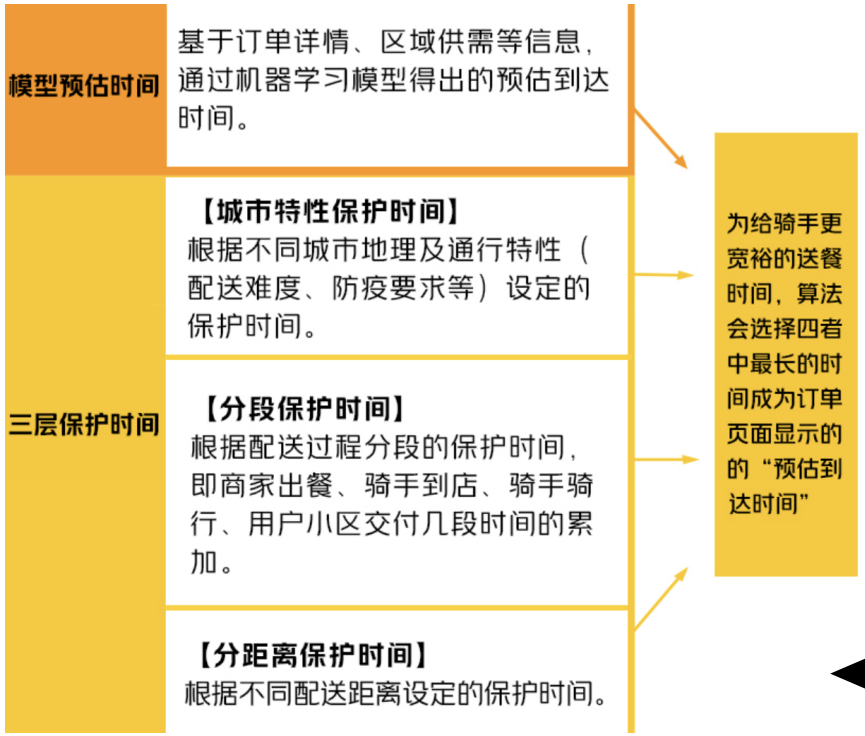
大数据的应用可以极大地增强企业对客户行为的理解和预测能力，进而显著提升客户体验。

- 微信聊天记录来推荐商品
- 通过淘宝搜索记录做个性化推荐
- **大数据杀熟**（基于用户的历史数据来对不同用户展示不同价格的）

大数据的应用

2. 优化业务流程

大数据的应用可以显著优化业务流程。例如，在供应链管理中，大数据可以预测需求波动，实现库存的精准调配；而在物流配送上，可以根据实时交通状况和天气变化调整路线，提高配送效率和减少成本。



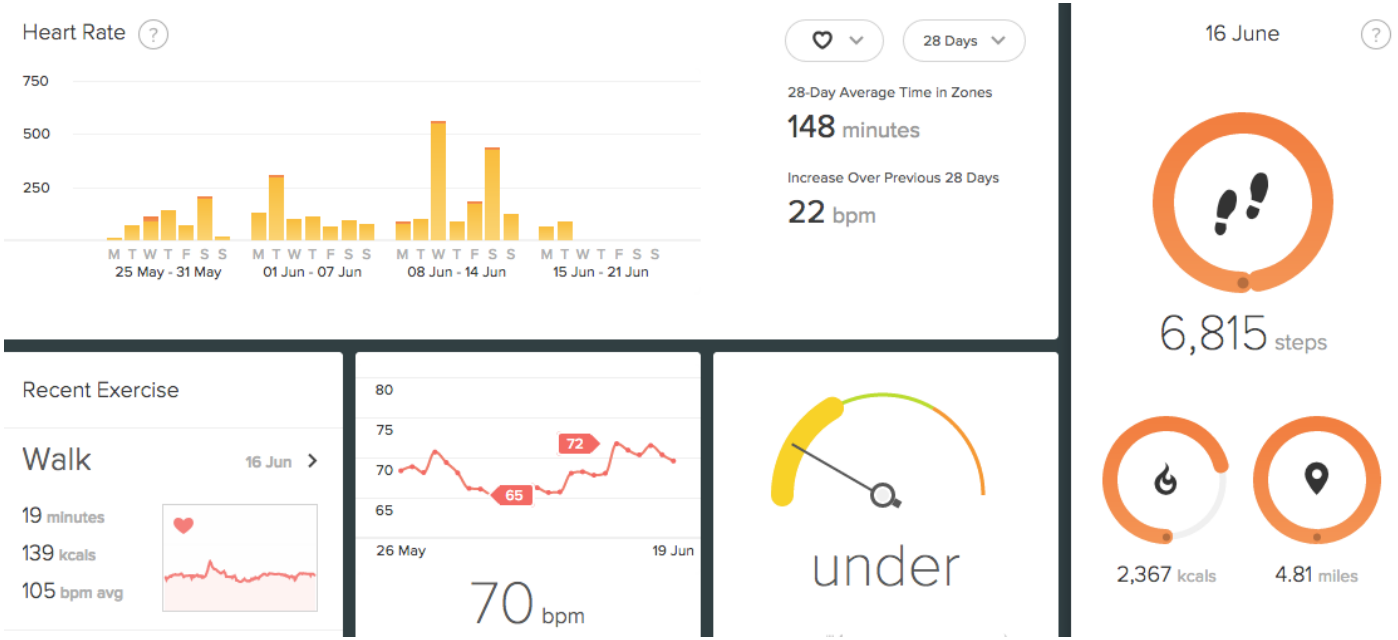
- 美团配送时间预测
- 导航目的地路线选择
- 滴滴预估到达时间

◀ 美团配送时间计算模型

大数据的应用

3. 改善生活

大数据同样深入到我们日常生活的每个角落。比如，智能手表和手环等设备持续生成数据，来监测健康指标；交友平台依靠算法分析个人偏好、行为习惯以及交互模式，帮助用户找到合适的伙伴。



某手环记录的数据（通过数据仪表板来展示）

大数据的应用

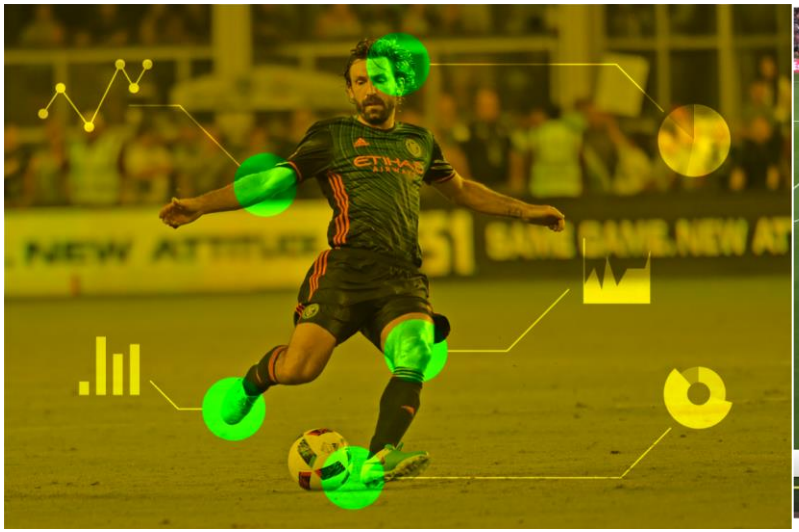
4. 提高医疗和研发

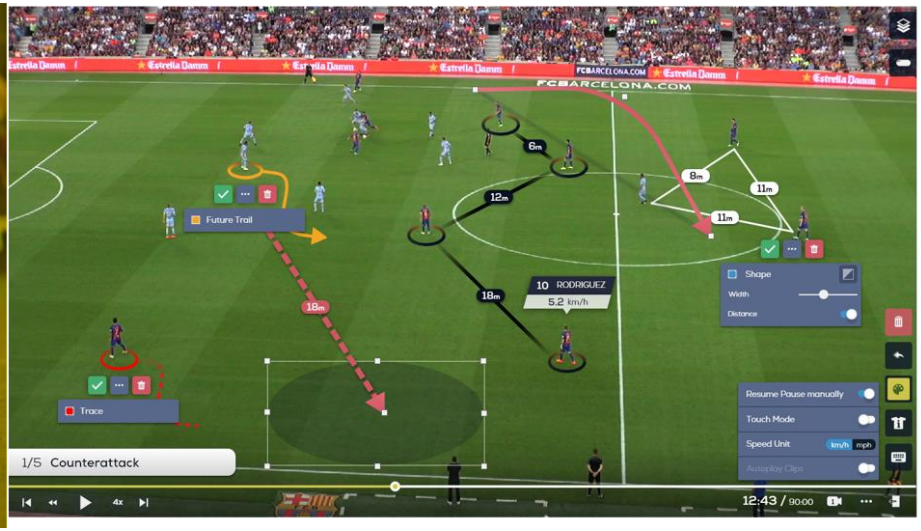
通过分析健康监测数据和临床数据（医疗大数据），医生能够制定个性化的治疗方案，及时预测疾病风险。例如，医院通过监控早产儿的生命体征，提前发现并处理潜在的健康问题，从而提高医疗的准确性和救治的成功率。

大数据的应用

5. 提高体育成绩

大数据分析技术能够提升训练效果和竞技表现。例如，通过视频分析技术，我们能够精确追踪足球比赛中球员的运动轨迹，活动热区以及战术表现。同时，装配在运动器材上的传感器（22年的世界杯的足球内部就装有传感器）能够提供详细的比赛性能数据，帮助裁判做出正确的判罚，也能够帮助教练员分析如何提高技巧和战略。





大数据的应用

6. 优化机器和设备性能

大数据分析能提升机器和设备的智能化。电车在行驶过程中实时记录加速度、刹车力度、电池充电状态和定位信息。即使车辆停止，关键信息如胎压，电池状态，周边状况也会持续传输到手机端，以便进行故障预警和性能监控。这些数据能够帮助汽车厂家分析故障原因，了解用户的驾驶习惯，并进一步开发自动驾驶功能等。

大数据的应用

7. 改善安全和执法

大数据的运用能够改善安全措施和执法工作。1）企业通过大数据技术分析网络流量模式，有效预防和响应网络攻击。2）执法机构利用大数据工具分析犯罪模式和行为趋势，更迅速地追踪并捕捉罪犯。3）信用卡公司则运用大数据算法实时监控交易活动，以便快速识别和阻止欺诈性交易。

大数据的应用

8. 改善我们的城市

大数据在城市管理和优化方面也发挥着重要的作用。利用实时交通流数据，交通管理者能够优化交通信号、减少拥堵，并提高公共交通的效率。社交媒体和天气数据的分析帮助应急管理部门在自然灾害发生时迅速做出反应，同时指导居民避开危险区域。

大数据的应用

9. 金融领域

1.风险控制： 中国人民银行的征信系统使得金融机构能够通过征信系统来评估和控制贷款风险，从而准确预测借贷违约的可能性，并作出更明智的放贷决策。

2.保险定价： 在保险行业，特别是车险领域，大数据分析允许保险公司根据车主的事故历史、职业、年龄、性别等多种因素来定制个性化的保险产品。

3.高频交易（HFT）:在股票市场，高频交易利用大数据算法在极短的时间内分析市场数据和外部信息，来作出交易决策。这些算法能够捕捉到微小的价格变动，并在秒级别内自动执行买卖订单，利用市场波动赚取利润。

课堂思考

1. 写出大数据的四个特点
2. 为什么会实现“数据”到“大数据”的跨越（写出至少两个）