



概论



覃雄派

提纲



概论

- 什么是数据科学(家)
- 生产、生活的数字化，数据量的增长
- 数据科学的兴起
 - 数据比以往任何时候都更容易产生与获取
 - 人们的决策比以往任何时候都更基于数据驱动
 - 人们处理数据的能力比以往任何时候都强大
- 数据科学家应该具备什么能力
 - 不同类型的数据
 - 数据科学的任务
 - 数据科学家的能力要求
- 教学计划与考核要求



概论

- About Teacher
- 任课教师：覃雄派
 - 信息学院副教授
 - qxp1990@ruc.edu.cn
 - 研究方向：高性能数据库，大数据分析，信息检索
 - 请同学们及时进入老师建立的班群





概论

- About Student

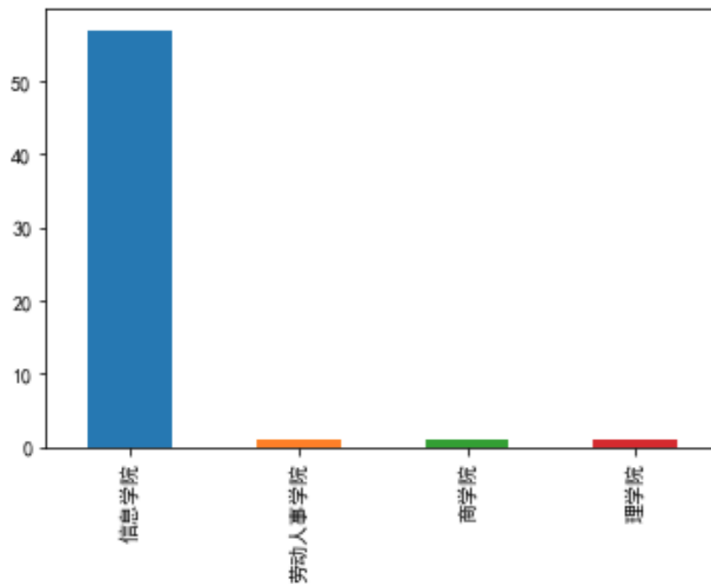
```
import pandas as pd
import matplotlib.pyplot as plt
stu = pd.read_excel('./数据科学导论范举老师班名单.xlsx')
stu.head()
```

	序号	学号	姓名	院系	专业	班级	年级
0	1	2016202xxx	李晓桐	信息学院	数学与应用数学	2016级理科试验班4班	2016
1	2	2017201xxx	杜凌志	商学院	管理科学	2017级本科管理科学班	2017
2	3	2018200xxx	张昊博	信息学院	理科试验班 (信息与数学)	2018级理科试验班3班	2018
3	4	2018200xxx	张龔然	信息学院	理科试验班 (信息与数学)	2018级理科试验班3班	2018
4	5	2018201xxx	杨晓彤	信息学院	理科试验班 (信息与数学)	2018级理科试验班3班	2018

概论

- About Student

```
counts = pd.value_counts(stu['院系'])  
plt.rcParams['font.sans-serif'] = ['SimHei']  
counts.plot.bar()  
plt.show()
```



概论





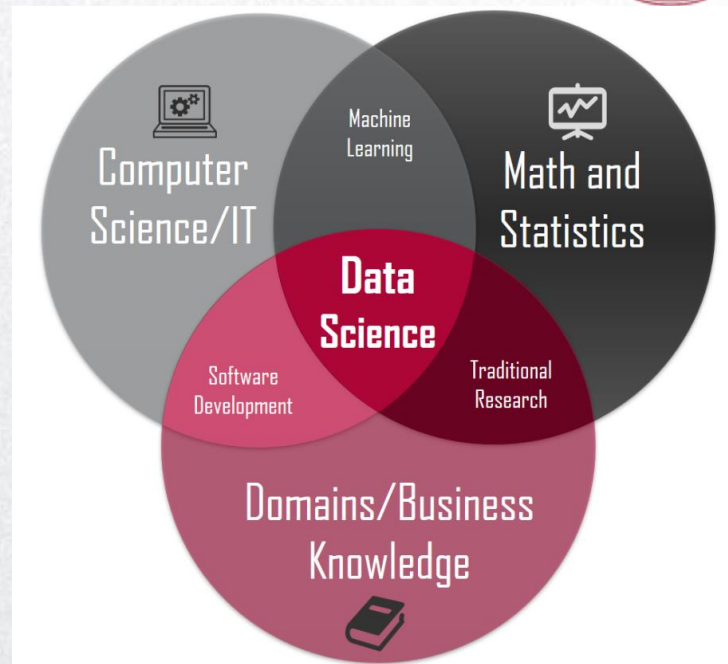
概论

- 什么是数据科学(家)?
 - 说什么的都有
 - A data scientist is a data analyst who lives in California
 - A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician
 - Data Science is statistics on a Mac
 - 没有人真的知道什么是数据科学(家).....
 - 数据科学方兴未艾，概念内涵在不断地变化发展
 - 人们对数据科学还没有明确地形成统一的定义

概论

- 一个公认却很宽泛的定义

Data science is an **inter-disciplinary** field that uses scientific methods, processes, algorithms and systems to extract **knowledge** and **insights** from many structural and unstructured data.

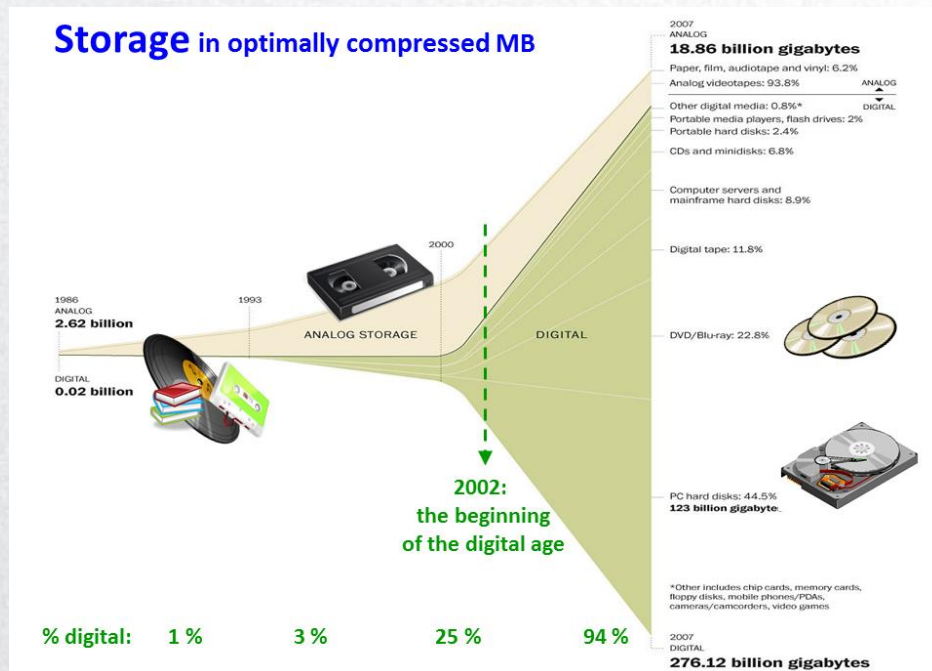


概论



概论

- 数据以史无前例的速度在增长
 - 右图：1986年到2007年的数据变化
 - 模拟数据
 - 数字数据
 - 推动数字数据增长的重要动力
 - Web → 搜索引擎
 - 社交媒体
 - 云计算
 - 大数据时代！



概论

- 数据以史无前例的速度在增长
- 数字信息量每2.5年翻一番
 - 2007年: 276.12 billion GB = $2.76 * 10^{14}$ MB



到**2020**年，数据相比
2007年大约翻了**5**番，
请问数据量有多大？

A. $1.4 * 10^{15}$ MB

B. $8.8 * 10^{15}$ MB



概论

- 数据以史无前例的速度在增长
- 数字信息量每2.5年翻一番

– 2007年: 276.12 billion GB = 2.76×10^{14} MB



到**2020**年，数据相比
2007年大约翻了**5**番，
请问数据量有多大？

A. 1.4×10^{15} MB

B. 8.8×10^{15} MB

地球年龄：46亿年

- 假设一首MP3歌曲时长4分钟，占用空间的大小是10MB。如果2020年的数字信息量全部用来存储歌曲，够一个人不眠不休听多久？



请计算上题：一个人
可以不眠不休听多久？

A. 67亿年 $\frac{8.8 \times 10^{15}}{\frac{10}{4} \times 60 \times 24 \times 365} = 6.7 \times 10^9$

B. 67万年

数据总量/每年可以听的数据量

概论

- 大数据的三个V
- Volume
 - GB→TB→ZB→EB
- Velocity
 - 数据快速处理分析
- Variety
 - 企业内部/外部
 - 结构化/非结构化
 - 文本、图片、视频.....



概论

- 人们的行为发生了深刻变化.....



教宗本笃十六世

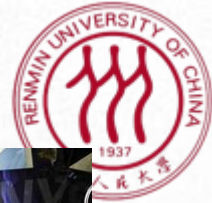


教宗方济各

数字痕迹

Digital Traces

概论



- 数字痕迹(Digital Traces)
 - 纽约证交所每个交易日生成 1TB 的交易数据
 - Facebook每天大概接收用户500+T的社交数据，主要是照片、视频、文本消息、评论等
 - 2019年春晚全球观众参与百度APP互动次数达到208亿次



趋势1：数据比以往任何时候都更容易产生与获取







概论



概论

- 这个世界变得越来越“数据驱动”
- 智能推荐
 - 信息的极度爆炸，使得人们找到他们需要的信息变得越来越难
 - 面对海量的数据，用户需要更加智能的、更加了解他们需求、口味和喜好的信息发现机制，于是推荐系统应运而生
 - 推荐算法能够根据用户的喜好，向用户推荐商品或者服务。在电子商务 (E-commerce, 比如Amazon等) 网站、音乐、电影、和图书分享网站，推荐引擎取得了巨大的成功。

Customers who viewed this item also viewed these products

			
Dualit Food XL1500 Processor	Kenwood kMix Manual Espresso Machine	Weber One Touch Gold Premium Charcoal Grill-57cm	NoMU Salt Pepper and Spice Grinders
\$560	★★★★★ \$250	\$225	\$3
Add to cart	Select options	Add to cart	View options

amazon.com

Recommended for You

Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.

		
Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop	Google Apps Administrator Guide: A Private-Label Web Workspace	Googlepedia: The Ultimate Google Resource (3rd Edition)

概论

- 这个世界变得越来越“数据驱动”
- 智能推荐离不开对海量用户行为数据的分析
 - 基于内容的推荐: 发现物品或者内容的相关性, 进行推荐
 - 协同过滤: 它根据用户对物品或者信息的偏好, 进行推荐

	西虹市首富	疯狂动物城	我不是药神	雨中曲	飞屋环游记	霸王别姬	虎口脱险
u1							
u2							
u3							
你		?	?			?	

概论

- 这个世界变得越来越“数据驱动”
- 信息流广告(feeds流广告)
 - 微信朋友圈广告是典型的feeds流广告
 - feeds广告就是与内容混排在一起的广告
 - 最不像广告的广告，长得最像内容的广告
 - Feeds广告操作简单，打扰性低，已经成为移动互联网时代主流的广告形式。
 - 建立在用户行为记录和大数据分析基础上，个性化推荐



概论

- 这个世界变得越来越“数据驱动”



下面哪些是信息流广告(多选)?



A



B



C

概论

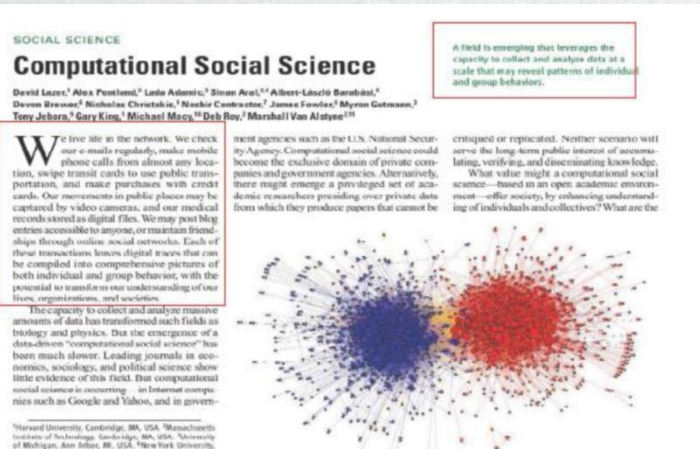
- “数据驱动” 的科学研究
- 2007年, 图灵奖得主Jim Gray提出数据密集型科学为科学的第四范式
 - Empirical(实验科学, 实验归纳)- 钻木取火
 - Theoretical (理论科学, 归纳总结)- 牛顿三定律
 - Computational (计算科学, 计算机仿真)- 模拟核试验, 天气预报
 - And now data-driven (**data-Intensive**)



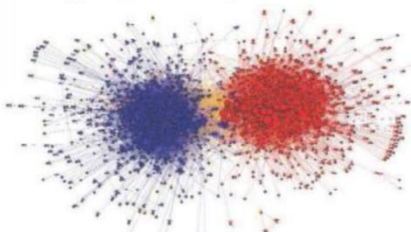
Data-driven science is the "fourth paradigm" of science that uses the computational analysis of large data as primary scientific method and "to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other".

概论

- 计算社会科学
- 数据驱动的社会科学 Data-Driven Social Science
 - 从海量数据挖掘中理解社会现象的有价值知识
 - 综合应用社会科学和计算技术解释社会现象

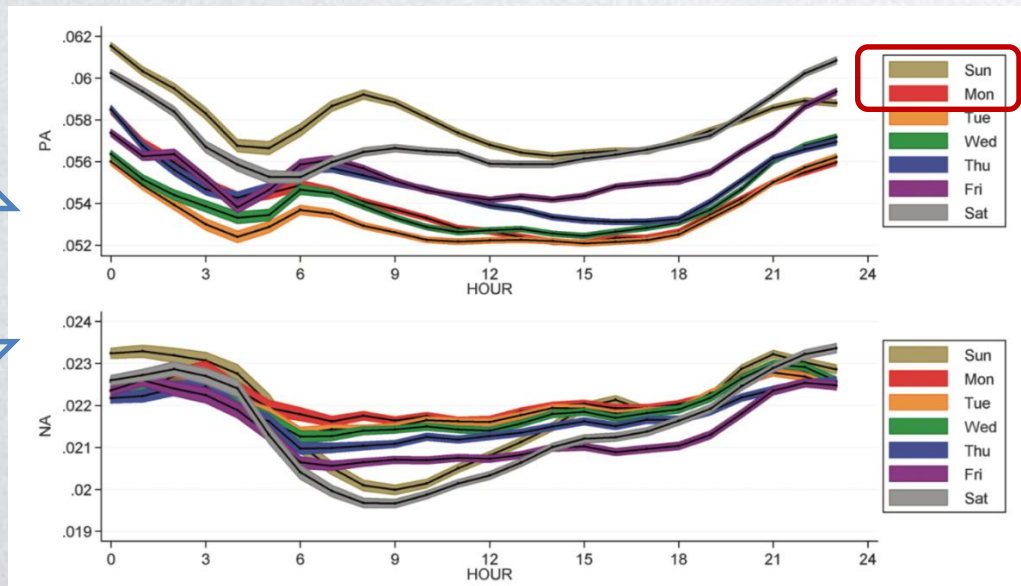


左图是2004年美国大选前，政治博客之间的链接网络。
你能发现什么特点？
注：红色代表支持民主党，而蓝色代表支持共和党



概论

- 大数据驱动的群体情绪变化
- 针对用户每天发布的Tweets进行文本分析



人的情绪在一天的不同时刻是会发生变化的

- 正、负面情绪并非此消彼长
- 周末我最high
- 周一不开心……

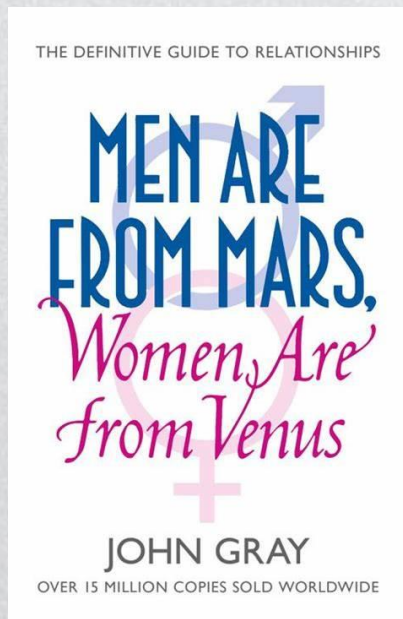
Science. 2011 Sep 30;333(6051):1878-81. doi: 10.1126/science.1202775.

Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures

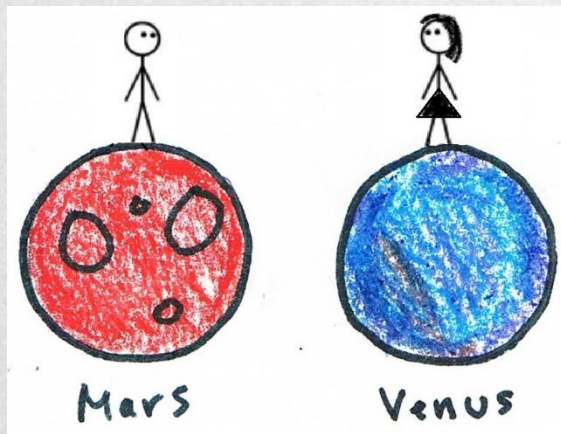
Scott A Golder , Michael W Macy

概论

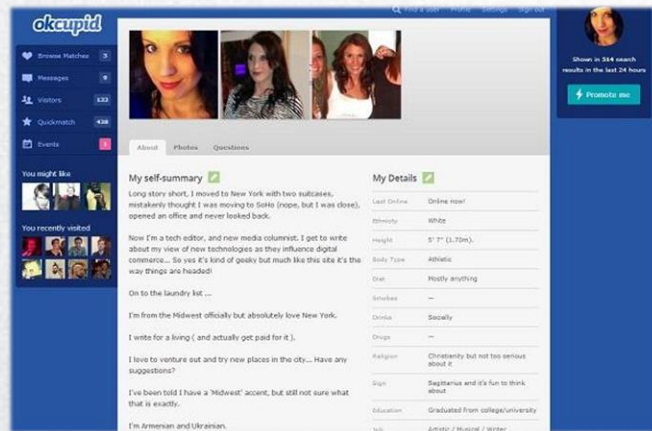
- 数据驱动的性别研究
- 性别研究(Gender Study)
 - 不同性别之间对于配偶的期待是否存在着显著差异?



我的年龄 vs. 配偶年龄

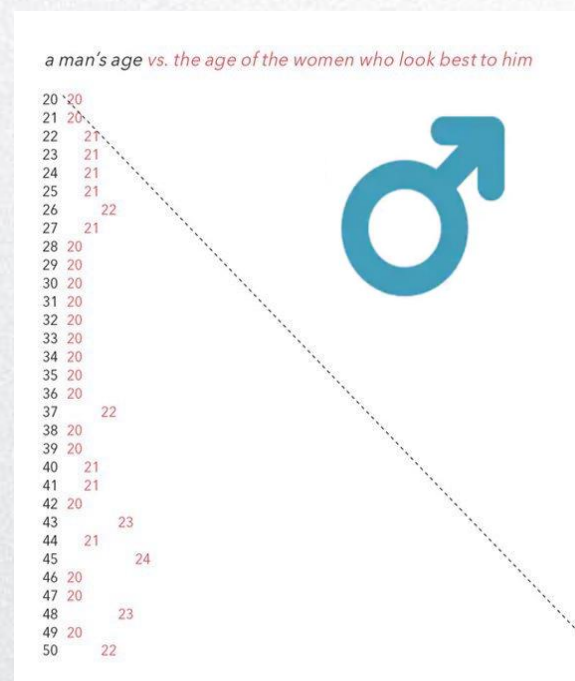
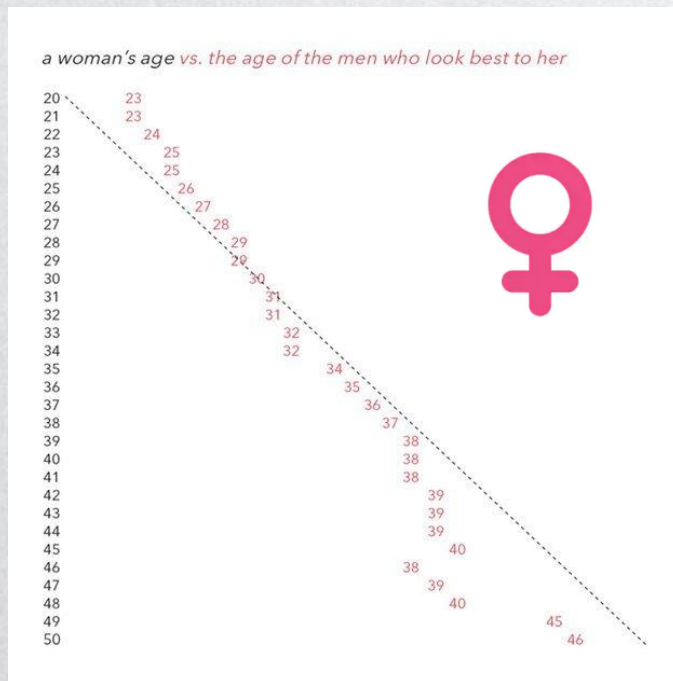


收集在线交友网站的数据.....



概论

- 数据驱动的性别研究



概论

- 数据驱动的决策支持



传统的企业
数据只有数据库管
理员或CIO关心



今天的企业
数据是企业的核心，所有
事情都越来越数据驱动

趋势2：人们的决策比以往任何
时候都更基于数据驱动

概论



概论

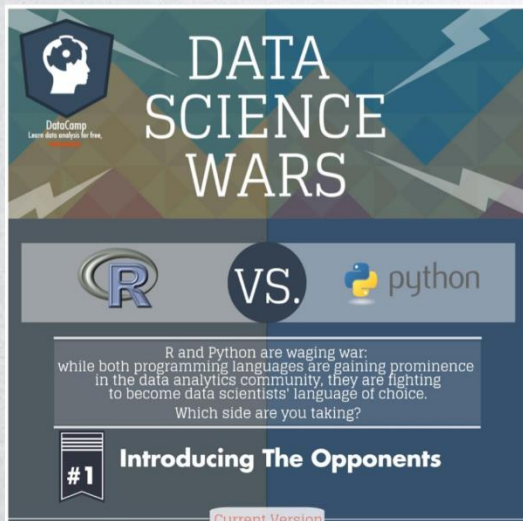
- 人们处理数据的能力与日俱增



从计算机的视野看
数据科学+大数据

概论

- 编程语言

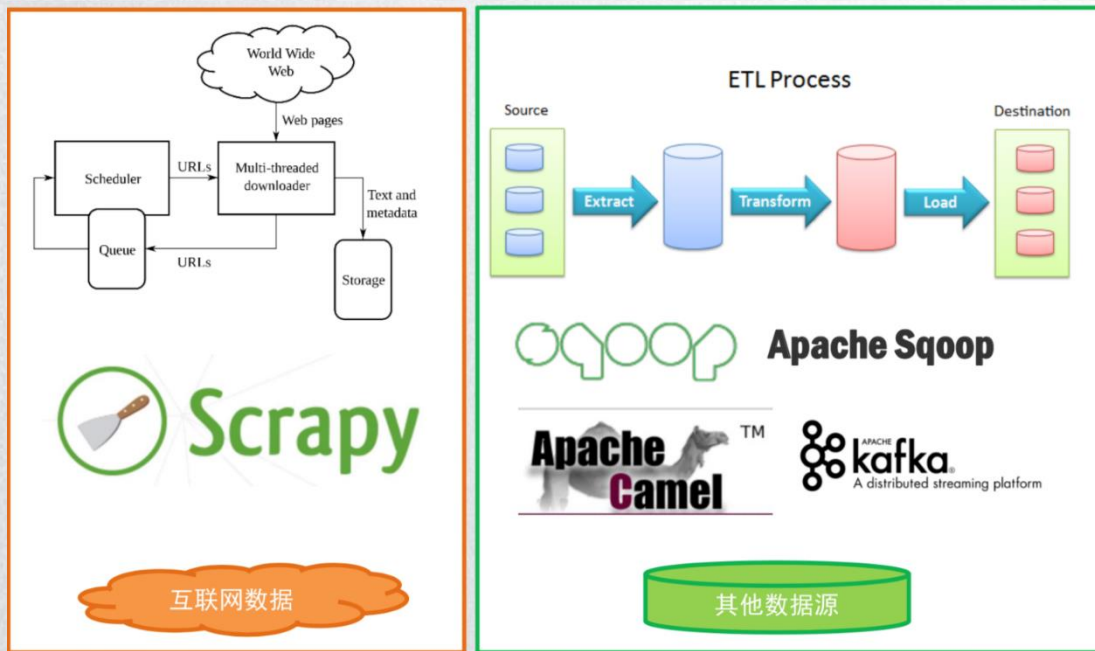


<https://www.tiobe.com/tiobe-index/>

Feb 2021	Feb 2020	Change	Programming Language	Ratings	Change
1	2	▲	C	16.34%	-0.43%
2	1	▼	Java	11.29%	-6.07%
3	3		Python	10.86%	+1.52%
4	4		C++	6.88%	+0.71%
5	5		C#	4.44%	-1.48%
6	6		Visual Basic	4.33%	-1.53%
7	7		JavaScript	2.27%	+0.21%
8	8		PHP	1.75%	-0.27%
9	9		SQL	1.72%	+0.20%
10	12	▲	Assembly language	1.65%	+0.54%
11	13	▲	R	1.56%	+0.55%
12	26	▲	Groovy	1.50%	+1.08%
13	11	▼	Go	1.28%	+0.15%
14	15	▲	Ruby	1.23%	+0.39%
15	10		Swift	1.13%	-0.33%

概论

- 数据采集与治理



概论

- 数据存储与计算平台



关系型数据



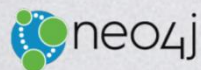
数据仓库
近实时分析



文档数据



分析引擎



图数据



流数据处理



HDFS

非结构化数据



MapReduce

批处理分析



全文索引

概论

- 分析平台与工具



分布式机器学习



机器学习工具集



图挖掘



自然语言处理



深度学习框架

概论

- 数据可视化



Apache ECharts

一个基于 JavaScript 的开源可视化图表库

快速入门

所有示例



趋势3：人们处理数据的能力比以往任何时候都强大

概论

- 计算机科学提供了大量方法
- 使数据处理更高效(Efficient)与更可扩展(Scalable)
 - 更优的算法: $O(N^2) \rightarrow O(N \log N)$
 - 并行与分布式处理
 - 异构硬件: GPU + CPU

$$r = r_{xy} = \frac{\text{Cov}(x, y)}{S_x \times S_y}$$

- 思考
 - 为什么数据科学 ≠ 统计学?



按照定义计算上述皮尔森相关系数的时间复杂度是多少?

A. $O(N)$

B. $O(N^2)$



概论

- 回顾：数据科学缘何兴起
 - 近年来的三个重要趋势
 - 数据比以往任何时候都更容易产生与获取
 - 人们的决策比以往任何时候都更基于数据驱动
 - 人们处理数据的能力比以往任何时候都强大
 - Data is new Oil!
 - 人们迫切地需要收集更多的数据、处理数据，从数据中洞察知识
 - 因此，数据科学应运而生
 - 数据科学的定义可能会随着时间而改变
 - 但它解决日益增长的数据规模与人们希望从数据中挖掘真知之间的矛盾这一点，不会改变

概论



概论

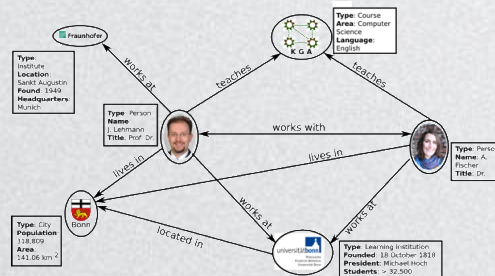
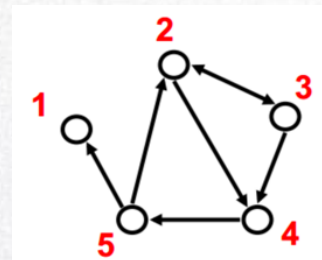
- 管理与处理不同类型的数据
- Variety: 数据的种类繁多
 - 数组、矩阵
 - 键值对
 - 实体-关系表
 - 时序数据、流数据
 - 图数据
 - 文本数据
 - 多媒体数据
 - ...

	item ₁	item ₂	item ₃	...	item _n
user ₁		5	2		1
user ₂	3				
user ₃	1		3		
...					
user _{m-1}	5		4		2
user _m		4			3

Primary Key		Products		
Partition Key	Sort Key	Attributes		
		Schema is defined per item		
Product ID	Type			
1	Book ID	Odyssey	Homer	1871
2	Album ID	6 Partitas	Bach	
2	Album ID	Partita No. 1		
3	Movie ID	The Kid	Drama, Comedy	Chaplin

Manufacturer					
ID	Name	Contact			
M-01	Hello World Tech.	534-55-7478			
M-02	ABC Technologies	283-92-8511			

Product					
ID	ManufacturerID	Name			
PDT-0001	M-01	Tiger T7 Bluetooth Headphones			
PDT-0002	M-01	DD-027 In-Ear Headphones, Black			
PDT-0003	M-02	Mr. 1022 Deep Bass Earbuds			



来源: 科技日报

据《新科学家》网站最新发布的信息, 超过40%的昆虫物种可能在未来几十年内灭绝, 其中蝴蝶、蜜蜂和苍蝇受到的影响最大, 主要原因是栖息地的丧失。这是对过去40年来所有昆虫长期调查得出的令人震惊的结论。

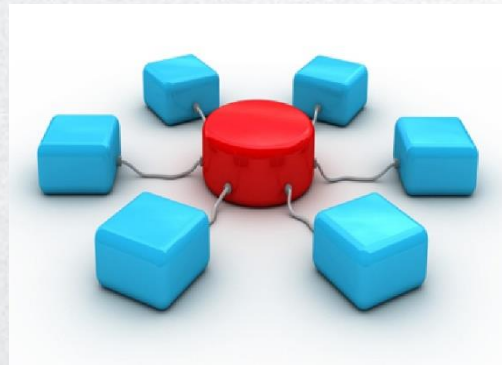
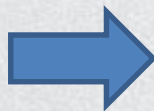
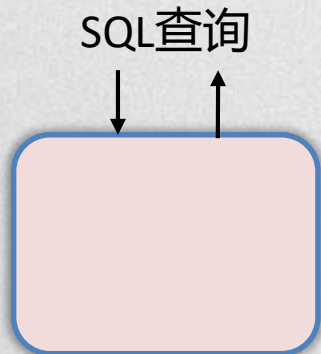
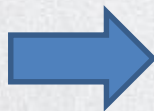
“这种影响对地球生态系统将是灾难性的, 因为昆虫是世界上许多生态系统的基石。”论文作者说, 他们来自澳大利亚悉尼大学和中国农业科学院。

研究发现, 昆虫减少的最大原因是栖息地丧失; 其次, 寄生虫和疾病也起着重要作用, 例如, 瓦螨的蔓延导致蜜蜂种群的衰退; 最后, 气候变化似乎也有影响, 热带地区的昆虫可能对温度变化的耐受性较差, 其数量可能已经因全球变暖而有所下降。



概论

- 管理与处理不同类型的数据
- Variety: 数据的种类繁多, 多数据源 → 思维方式的改变
- 封闭世界假设
 - Everything stated by the database is **true**;
 - Everything else is **false**
- 开放世界假设
 - A statement **may be true** irrespective of whether or not it is *known* to be true



数据库系统

汇集更多数据 → 更好的分析结果

概论

- 为什么要处理**不同类型**的数据
- 考虑一个场景：请你基于数据分析原因

中国iPhone销量下滑速度是整个市场的两倍

2019年02月11日 23:03 3558 次阅读 稿源：威锋网  4 条评论

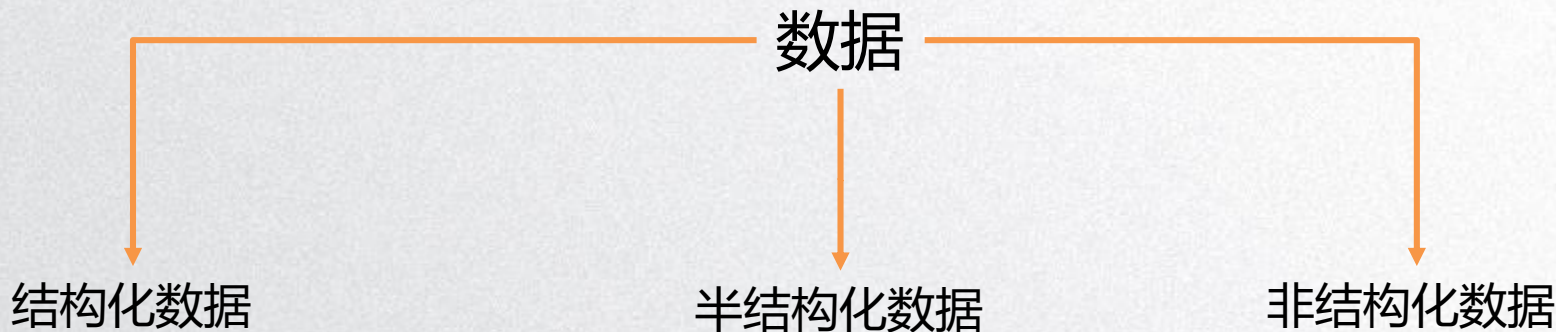
苹果在其假日季度财报电话会议上透露，iPhone 在中国的糟糕销售是导致该公司季度收入达不到预期的主要原因。市场分析公司 IDC 本周对 iPhone 在中国市场的糟糕程度进行了估计，在中国，iPhone 销量的下滑速度是智能手机市场整体下滑速度的两倍。



- 你要采集哪些数据来支撑你的分析？

概论

- 为什么要处理不同类型的数据



China Smartphone Shipment Market Share (%)	Q2 2017	Q2 2018	YoY Growth
HUAWEI	20%	26%	22%
OPPO	19%	19%	-9%
vivo	17%	18%	-1%
Xiaomi	13%	13%	-10%
Apple	8%	9%	0%
Others	23%	16%	-37%
TOTAL	100%	100%	-7%

Counterpoint
Technology Market Research



6***m

PLUS会员

★★★★☆

手机还行。信号不怎么好。

银色 公开版 256GB 2018-12-03 10:43



旧***替

★★★★☆

用起来还好。还是很相信京东的!

深空灰色 公开版 256GB 2018-12-04 17:18



O***b

★★★★☆

商品很好。信号很差

深空灰色 公开版 64GB 2018-12-14 18:45



h***8

PLUS会员

★★★★☆

物流速度快。信号是有点问题!

深空灰色 公开版 256GB 2018-10-03 07:00

全球各地的评论媒体对 iPhone Xs 和 iPhone Xs Max 进行了测试。下面是他们做出的一些评论:

Mashable

“再度改进的摄像头硬件结合了新的‘智能 HDR’自动技术，由神经网络引擎和 A12 仿生的图像信号处理器再添动力，意味着你可以充分享用先进的摄像头光学技术和计算摄影技术带来的益处。”

TechCrunch

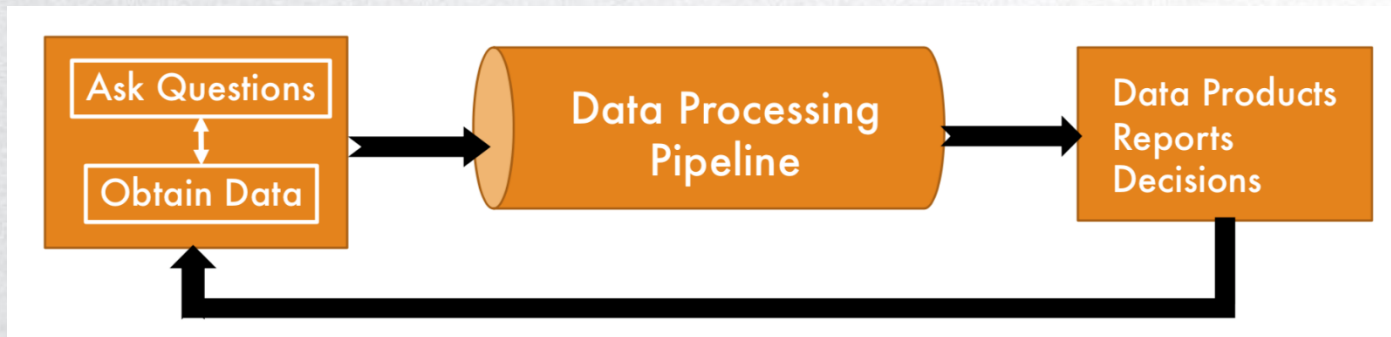
“谈到中央处理器性能，这款开创性的规模化 7 纳米架构已带来显著成效。iPhone Xs 拥有可媲美笔记本电脑的运行速度和远超 iPhone X 的处理性能，其架构的成效由此可见一斑。”

Daring Fireball

“iPhone 镜头和感光元件的品质无法与体积更大的专业相机相比，甚至相差较远。这是由于物理定律的限制。但是，传统的相机企业在定制化芯片和软件方面却逊色于 Apple，他们的相机无法像 iPhone 一样便于随身携带，也无法随时连接互联网进行分享。从长期考虑，明智的投资应当用于芯片和软件。”

概论

- 数据科学的工作流程



循环迭代式的工作流程

- 先提出问题，再收集与分析相关的数据
- 先收集数据，再分析可以回答哪些问题

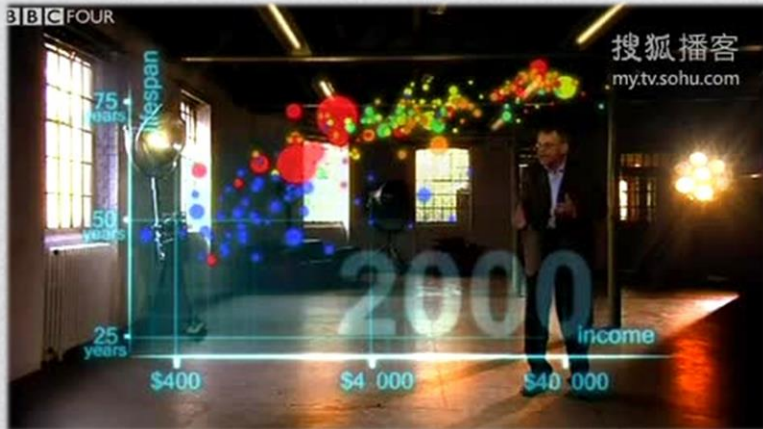
任务1：从数据中洞见真知

任务2：数据驱动决策支持

概论

- 核心任务1：从数据中洞见真知
- Raw Data → Insights
 - Hans Rosling's *200 Countries, 200 Years, 4 Minutes*
 - 视频链接: http://www.iqiyi.com/w_19s0nozzyd.html

播放



回答以下问题：

- 图表中的横纵坐标、圆点的大小分别表示什么含义
- 图表的左下角和右上角分别表示什么含义
- 图表中圆点聚在一起和比较分散分别表示什么含义
- 放眼世界，你能从可视化中看出什么趋势
- 与19世纪初相比，当今世界国家之间发展是否更不均衡？
- 聚焦中国，你能从可视化中观察到哪些趋势或特点

概论

- 核心任务2：基于数据驱动进行决策支持



狐狸



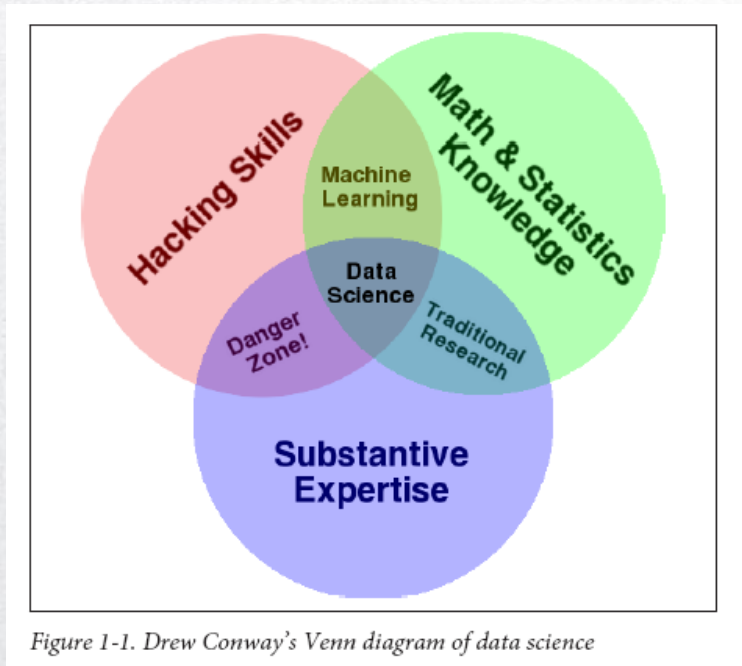
狗



狐狸 or 狗?

概论

- 需要掌握数据分析的技能与工具
- 编程语言
 - Python/R数据分析生态以实用工具
- 与数据分析相关的技术
 - 数据库系统
 - 数理统计
 - 机器学习
 - 线性代数/最优化
 - 数据可视化
 -
- 需要了解领域知识





概论

- 需要能够处理大规模的数据
- One thousand data instances
- One million data instances
- One billion data instances
- One trillion data instances
- Those are not different **numbers**, those are different **mindsets**



概论

- 需要能够处理大规模的数据
 - 上千级别(thousand)数据样本
 - 无需编程自动化处理
 - 百万级别(million)数据样本
 - 自动处理
 - 低于 $O(n^2)$ 的算法
 - 并行化后的 $O(n^2)$ 算法
 - 十亿级别(billion)数据样本(Web-scale)
 - 如何存储开始成为问题(分布式存储)
 - 万亿级别(trillion)数据样本
 - 数据很难存储在同一个物理地址
 - 分布式处理, 但是又要考虑容错等机制, 复杂性进一步上升
 - 几乎无法得知数据的全貌(难以实现有效采样)
 - 数据隐私/不一致/数据分布偏斜都成为问题



概论

- 需要了解数据伦理问题
- Target的“神”预测带来的隐私担忧
 - 2012年，明尼苏达州一家Target门店被客户投诉，一位中年男子指控Target将婴儿产品优惠券，寄给他的女儿，而他的女儿只是一个高中生，实在不可理喻
 - 但是没有过多久，他却给Target来电道歉，因为经他逼问，他女儿后承认自己真的怀孕了
 - 这位高中生没有告诉过父亲她怀孕了，也没有在Target调查问卷上留下过类似的记录。
- Target的数据分析师开发了怀孕预测模型
 - 通过分析这位女孩购买无味湿纸巾和补镁药品就预测到她可能怀孕了

概论

- 需要了解数据伦理问题
- 信息茧房：智能推荐的危机？
 - “我们只听我们选择的东西和愉悦我们的东西”
 - 在一个封闭的信息环境里，团队成员互相强化已有的观点



别被算法困在“信息茧房”



概论

- 需要了解数据伦理问题
- 信息茧房：智能推荐的危机？



曾经的女神





概论

- 回顾：数据科学家应具备什么能力？
 - 管理和处理各种类型的数据
 - 解决数据科学的两个核心任务
 - 从数据中洞见真知
 - 基于数据驱动进行决策支持
 - 掌握数据分析的技能与工具
 - 能够处理大规模的数据
 - 了解数据伦理问题

概论





概论

- 课程定位：数据科学系列课程的先导课
- 课程目标：
 - 让同学们对数据科学有一个整体的认识
 - 针对文本、图等不同类型的数据进行深入讲解
 - 了解数据分析的基本技术
 - 训练同学们使用Python的生态工具从头到尾地完成1-2个项目
- 数据科学导论课能让你成为数据科学家吗？
 - 不能.....
 - 但我们希望这是一个好的开端！



概论

- 课程覆盖的内容
 - 管理和处理各种类型的数据
 - 文本、图、关系、Web、时间序列、流数据.....
 - 解决数据科学的两个核心任务
 - 从数据中洞见真知: raw data → Insights
 - 基于数据驱动进行决策支持: 文本分析、图数据分析.....
 - 掌握数据分析的技能与工具
 - Python及其数据分析工具
 - 机器学习初步
 - 数据库系统、统计、最优化.....
 - 能够处理大规模的数据
 - 初步介绍一些分布式数据处理工具

红字内容是课程覆盖



概论

- 课程不会深入的内容
 - 数据库系统与技术(如SQL查询)
 - 实际上, 数据科学家需要非常熟练的掌握数据库技术
 - 我们把这部分知识留给后续的数据库相关课程
 - Python程序设计与数据分析编程实践
 - 实际上, 这部分对成为一个数据科学家来讲非常重要
 - 我们认为你们已经掌握Python, 或者能够通过自学+上机课掌握基本的技能
 - 复杂的机器学习与深度学习模型
 - 实际上, 机器学习与深度学习正变得越来越重要
 - 我们会讲解机器学习的基本思想与最简单的模型, 把更复杂的知识留给后续的课程
 - 数据伦理问题
 - 目前学术界与工业界对数据伦理的研究尚待深入



概论

- 课程大纲
 - (1)Intro: 数据科学概论
 - (2)EDA: 探索式数据分析
 - Python基础、常用工具pandas/numpy、数据探索、预处理
 - (3)ML: 机器学习初步与实践
 - 分类(KNN)、聚类(K-means)与回归(linear regression), 以及Python机器学习实践(如scikit-learn)
 - (4)Text: 文本数据分析与处理
 - 文本的预处理(如中文分词)、文本的分类、文本的检索...
 - (5)Graph: 图数据分析与处理
 - 图的基本概念、图的构建与可视化、图的中心度分析、图的社区检测、影响力分析
 - (6)System: 大数据分布式处理

概论





概论

- 课程考核
 - 期末成绩：40%
 - 期末考试(笔试)
 - 平时成绩：60%
 - 期中考试/平时上机：15%
 - 平时作业与课程项目：80%
 - 课堂表现：5%



概论

- 自学内容：Python语言基础
 - 编程编程自学网址：https://www.sololearn.com
 - 个人电脑访问该网站，需要用Google或者Facebook账号进行注册，如果没有账号需要梯子才能注册
 - 或者可以先用app注册，然后登录用网站或者app都行
 - 学 习 其 中 Python 3 Tutorial 教 程 (<https://www.sololearn.com/Play/Python/>)。
 - 总共学习9个modules
 - 在OBE上提交证书



概论

