

# 绪论：数据科学介绍

DSCI6001P 数据科学基础

主讲人：何向南

12 Sep 2024

[hexn@ustc.edu.cn](mailto:hexn@ustc.edu.cn)



- 何向南 中国科学技术大学，教授、博导
  - 个人主页：<http://staff.ustc.edu.cn/~hexn/>
  - 人工智能与数据科学学院副院长，数据空间研究院（合肥综合性国家科学中心）副院长
  - 获奖：阿里青橙奖，Elsevier高被引学者，安徽五四青年奖章，SIGIR最佳论文奖项 etc.
- 学术基本情况
  - 研究兴趣：推荐系统、数据挖掘、因果推理、机器学习、大模型 etc.
  - 论文主要发表地：SIGIR、KDD、WWW、IJCAI、NeurIPS、TKDE、TOIS etc.
  - 谷歌学术引用次数4万余次，h-index=97
- 创建USTC Lab for Data Science (<http://data-science.ustc.edu.cn/>)
  - 教授3人，副研究员/博士后4人，博士20余名，硕士40余名
- 教育经历
  - 2008.9 – 2011.6，华东师范大学，软件学院，本科
  - 2011.7 – 2019.2，新加坡国立大学，计算机学院，博士、博士后
  - 2019.3 – 至今，中国科学技术大学，教授、博导

# 课程考核

- 平时作业: 50%
  - 5次作业 (笔试)
- 实践项目: 50%
  - 1~2次实践项目 (编程)

教师:



何向南

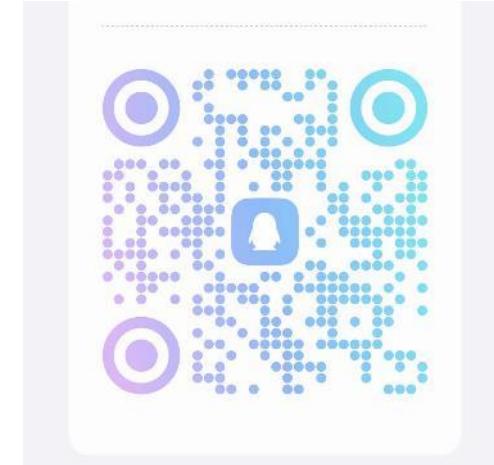
hexn@ustc.edu.cn

助教:



汪远博一  
wy1001@  
mail.ustc.edu.cn

课程QQ群 (群号: 547810062)



方羿研二  
peterfang@  
mail.ustc.edu.cn

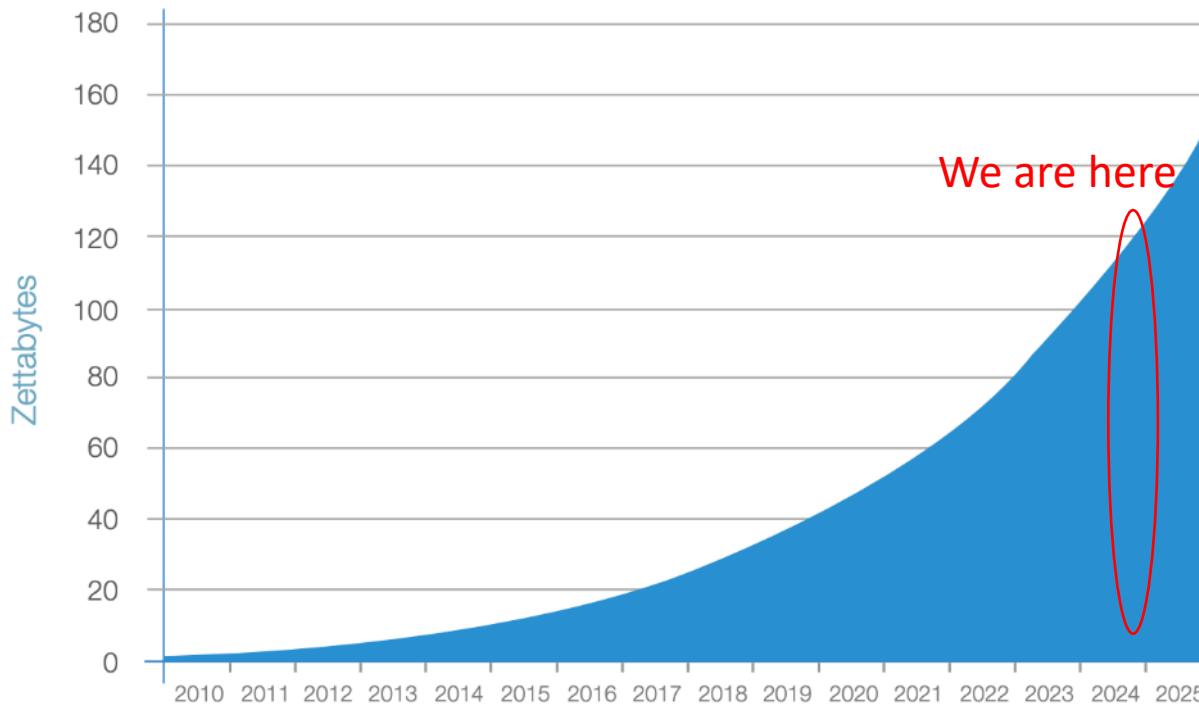
- 数据科学概述

- 数据科学的兴起
- 数据科学家应该具备什么样的能力
- 教学计划和考核要求

- 数据科学基本内容简介

- 机器学习
- 关联规则挖掘
- 自然语言处理
- 图和社交网络分析
- 分布式计算

- IDC发布的报告《数据时代2025》中预测，到2025年，全球数据圈将扩展至163ZB（1ZB等于1万亿GB），相当于2016年所产生16.1ZB数据的十倍。



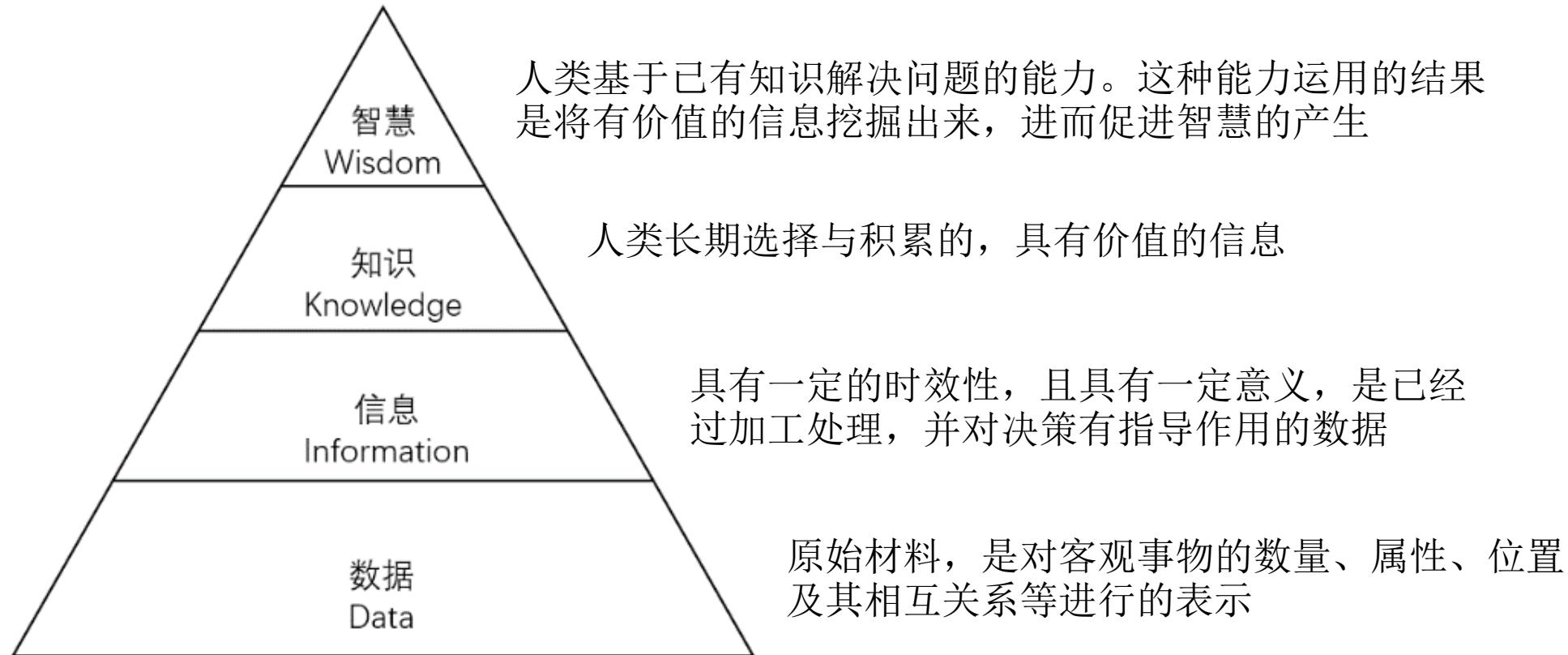
资料来源：IDC“数据时代2025”研究，希捷赞助，2017年3月

- 推动数字数据增长的重要动力
  - Web1.0 → 搜索引擎
  - Web 2.0 → 推荐系统、社交媒体
  - IoT → 终端设备
  - AIGC → 大模型
  - 数字化、智能化

- 数据中蕴藏着价值，如何挖掘出这些价值，是对传统科学技术（计算机科学、统计学、数学）的巨大挑战。

- 不同领域定义不同：
  - **统计学**：为了找出问题背后的规律而需要的，与问题相关的**变量的观测值**，是对客观现象进行计量的结果。
  - **计算机科学**：所有能输入到计算机，并被程序处理的符号总称，是用于输入计算机进行处理，具有一定意义的**数字、字母、符号和模拟量等**的通称。
  - **数据科学**：在一定背景下**有意义的**对于现实世界中的事物**定性或定量的记录**
- 数据的类型
  - **依据结构分类**：结构化数据、非结构化数据。
    - 数字、字符、日期等属于结构化；文字、图片、视频、音频属于非结构化
  - **依据形式分类**：文本数据、属性数据、声音数据、图片数据、视频数据等等
  - **依据来源分类**：观测数据、实验数据。
  - 如何分类，取决于我们想要用数据解决什么样的问题。

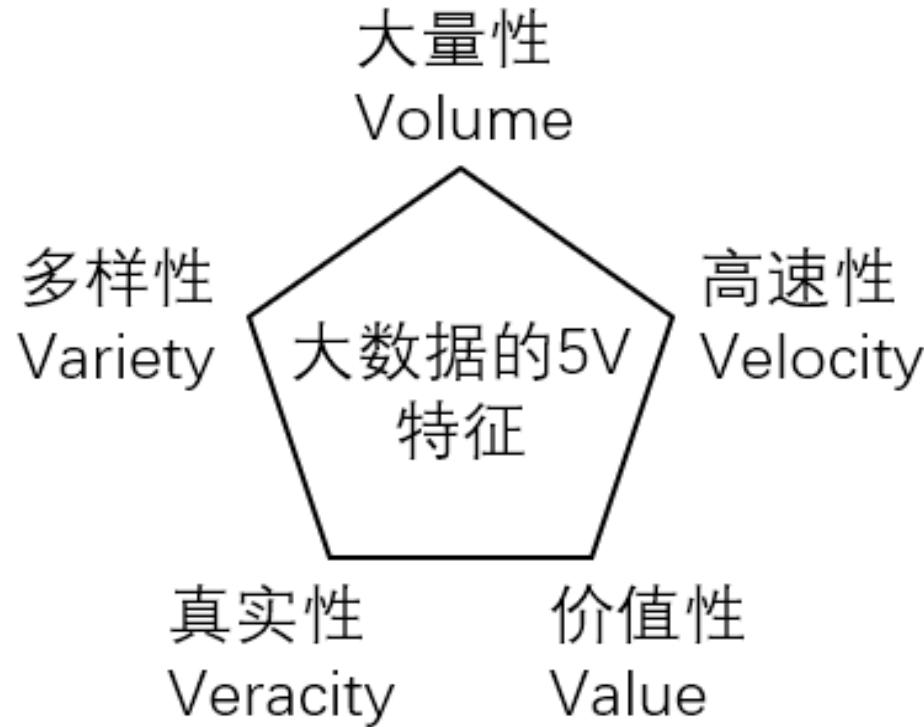
## • 数据、信息、知识、智慧 (DIKW)



- 数据科学的任务：以数据为研究对象，提炼出数据中蕴含的对决策有益的信息和知识。

## • 一些早期对“大数据”的定义

- 数据科学家**John Rauser**: 任何超过了一台计算机处理能力的数据。
  - 咨询公司**麦肯锡**: 无法在一定时间内用传统数据库软件工具对其进行抓取、管理和处理的数据集合。
  - 咨询公司**高德纳**: 大量、高速、或多变的信息资产，它需要新型的处理方式去促成更强的决策能力、洞察力与最优化处理。
- 
- 以上的定义都指出：
    - 大数据依旧是数据，或数据相关的过程；
    - 大数据的规模并非一定要达到某一确切的数值，关键在于，是否超过了实际情况下的数据存储能力和数据计算能力。



- **Volume 大量性**
- **Velocity 高速性**
  - 数据增长速度快，要求处理的实时性
  - E.g., 搜索引擎、推荐系统
- **Variety 多样性**
  - 企业、政府、社会...
  - 文本、图片、视频...
  - 交易、传感、移动轨迹、社交媒体 ...
- **Veracity 真实性**
  - 确保采集数据的真实性、客观性
  - 通过数据分析，还原和预测事物的本来面目
- **Value 价值性**
  - 不管数据多么大、高速、多样，能发挥其价值才是王道

## • 一些历史定义

- 最早提出数据科学概念的Peter Naur: “研究处理数据的科学”。
- 美国计算机科学家William S. Cleveland: 随着计算机科学的发展而扩展的，统计学中数据分析的技术领域，叫做数据科学。
- Journal of Data Science: 所有与数据有关的东西，如数据收集、处理、建模、分析等，但其重点应该为数据应用。
- 李国杰院士：数据科学是数学（统计、代数、拓扑等）、计算机科学、基础科学和各种应用科学融合的科学，类似钱学森先生提出的‘大成智慧学’。
- 复旦大学数据科学研究中心：关于数据的科学，用来研究探索Cyberspace中数据奥秘的理论、方法和技术。

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.

数据科学通过一系列科学的流程，研究现实世界方方面面产生的数据，从而完成从数据中抽取出**信息和知识的任务**，发现事物背后隐藏的规律，最终使数据的集成度更高，价值密度更大。

The Words Keep Changing:

- 2000: Data Mining, Knowledge Discovery from Databases
- 2010: Big Data
- 2020: Data Science future .....

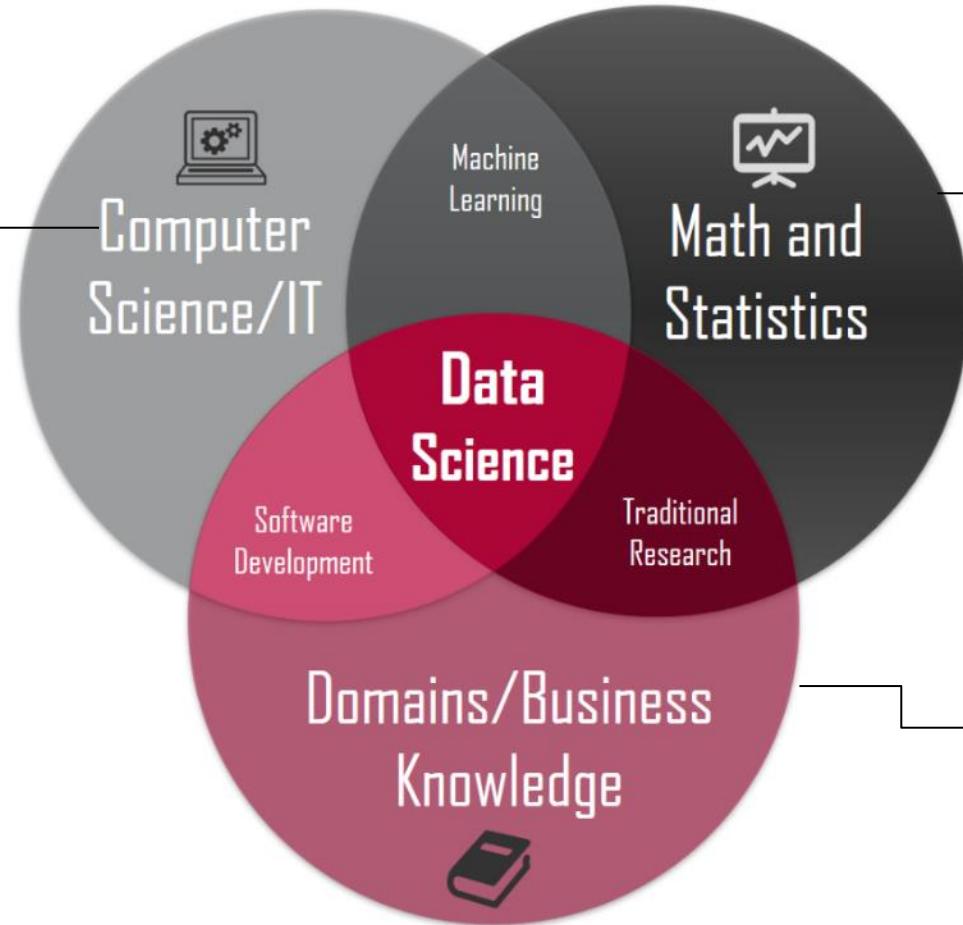
Actually, the essence is the same:

- Put the fastest hardware, systems, and algorithms together to solve problems in the commercial, scientific, and real worlds.

# 数据科学（学科）的韦恩图

数据科学是作为支撑数据研究与应用的新兴交叉学科。2010年9月，Drew Conway 使用韦恩图定义了数据科学的理论体系。

原图是Hacking skills，在收集数据，清理数据，处理数据，分析数据等一系列流程中，需要用到的计算机科学，人工智能等方面的方法与技术。



在对数据进行分析处理的过程中，需要用到的数学和统计学方法理论。

数据科学工作中涉及到实质性领域知识(Substantive Expertise)，领域知识对与发现和解决实际问题至关重要。

# 趋势1：数据比以往任何时候都更容易产生与获取

- 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



人们的行为发生了深刻变化.....



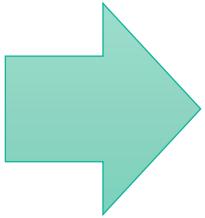
教宗本笃十六世



教宗方济各

数字痕迹  
Digital Traces

## 趋势2：人们的决策比以往都更基于数据驱动



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

- 智能推荐

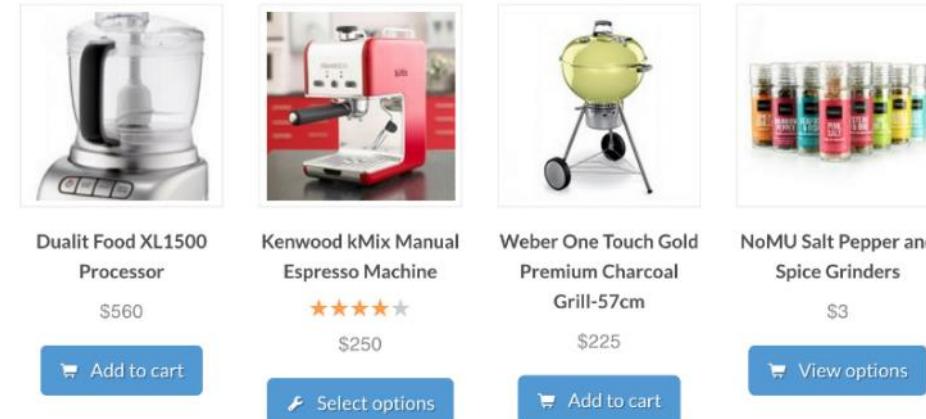
- 信息的极度爆炸，使得人们找到他们需要的信息变得越来越难
- 面对海量的数据，用户需要更加智能的、更加了解他们需求、口味和喜好的信息发现机制，于是推荐系统应运而生

- 信息流广告

- 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**操作性简单，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

Customers who viewed this item also viewed these products



# “数据驱动”的科学研究

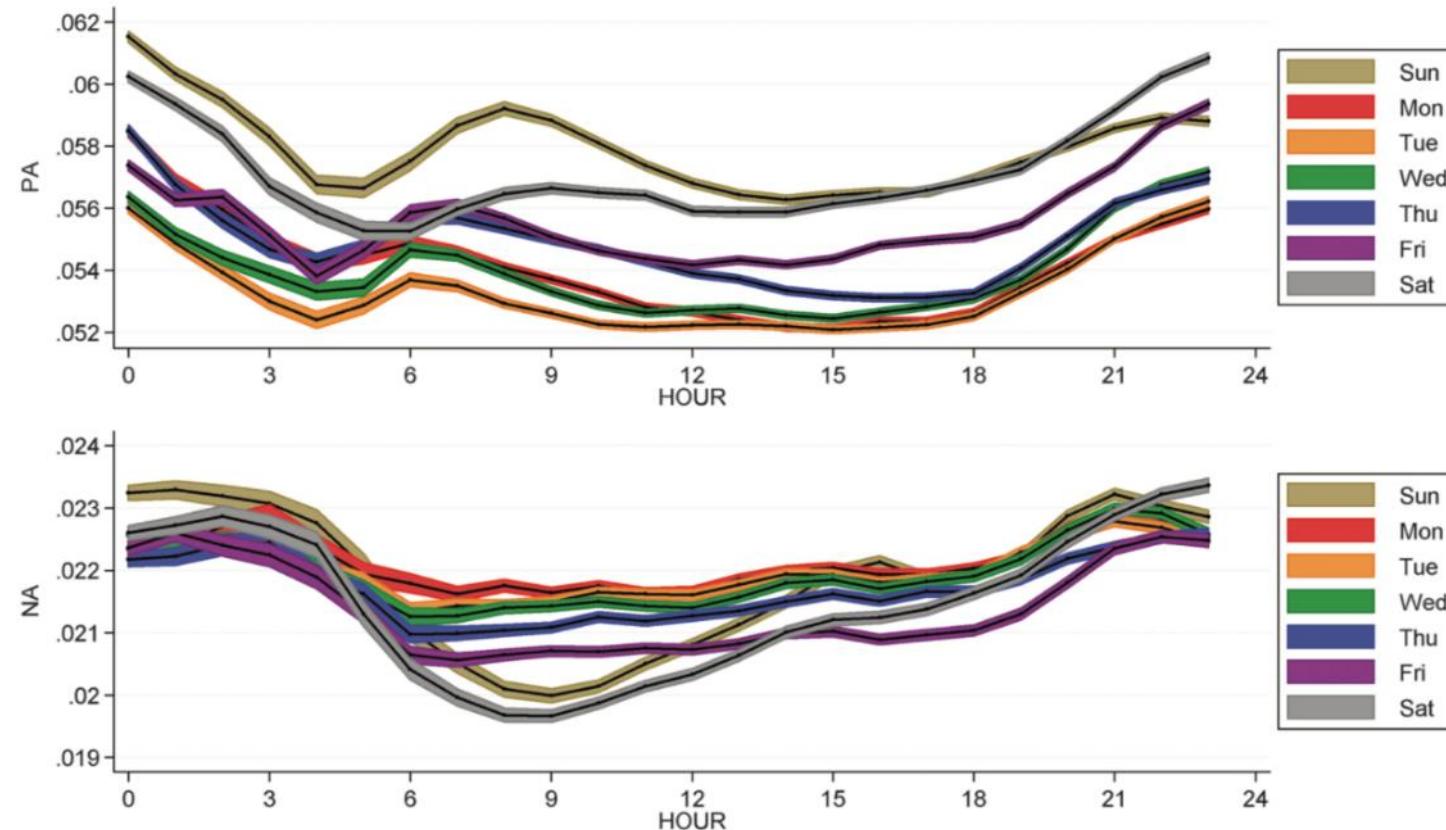
- 2007年，图灵奖得主Jim Gray提出数据密集型科学为科学的第四范式



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 ". ---- Jim Gray

# 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



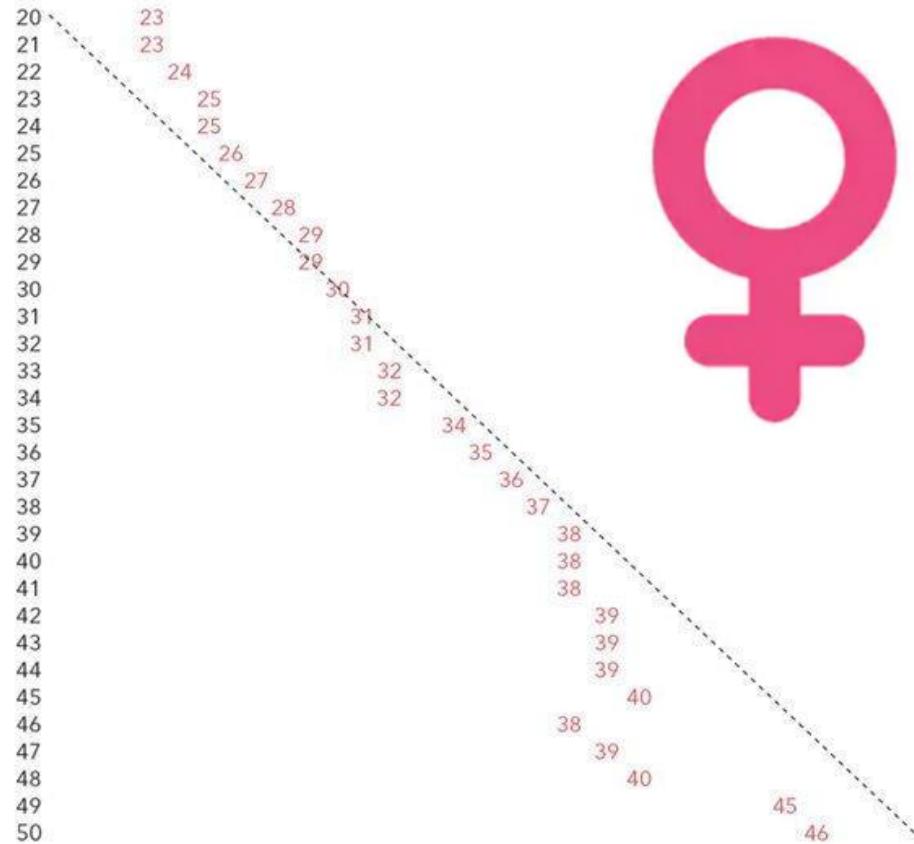
Source: Golder and Macy: Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. Science, VOL 333, 2011

人的情绪在一天的不同时刻是会发生变化的

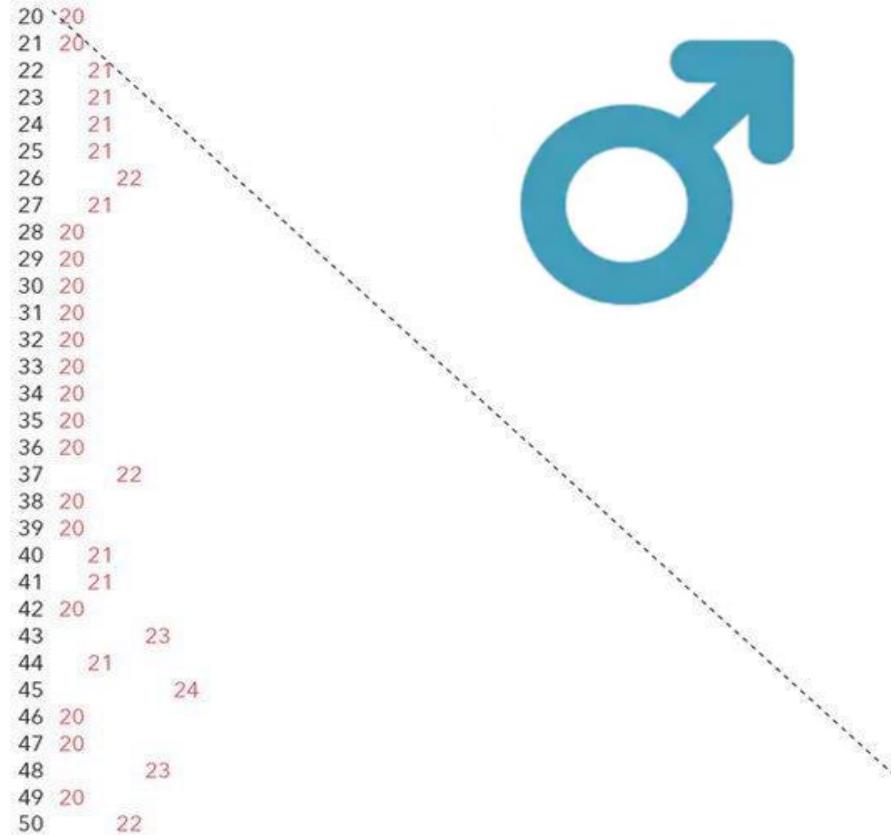
- 正、负面情绪并非此消彼长
- 周末我最high
- 周一不开心……

- 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?

a woman's age vs. the age of the men who look best to her



a man's age vs. the age of the women who look best to him



# 数据驱动的推荐系统

猜你喜欢



关注 分享

## 机器学习

击败AlphaGo的武林秘籍，赢得人机大战的必由之路：人工智能大牛周志华教授巨著，全面揭开机器学习的奥秘

周志华 著

京东价 **¥68.40** [7.8折] [定价 ¥88.00] (降价通知)

累计评价  
10万+

促销信息 **换购** 购买1件可优惠换购热销商品 立即换购 >

**加价购** 满10元另加26.90元，或满12元另加16.90元，或满15元另加9.90元，

即可在购物车换购热销商品 详情 >

以上促销可在购物车任选其一

增值业务 **助力环保，传递知识，旧书换新**

排名 自营 计算机与互联网销量榜 第 4 位

配送至 安徽合肥市蜀山区笔架山街道 ✓ 有货

由 京东 发货，并提供售后服务。23:10前下单，预计明天(09月14日)送达

重量 0.92kg

猜你喜欢

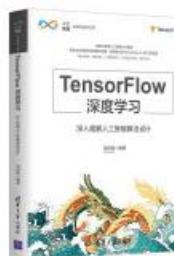
1/4



¥56.80



¥118.00



¥84.60



¥56.60



¥66.30



¥122.90

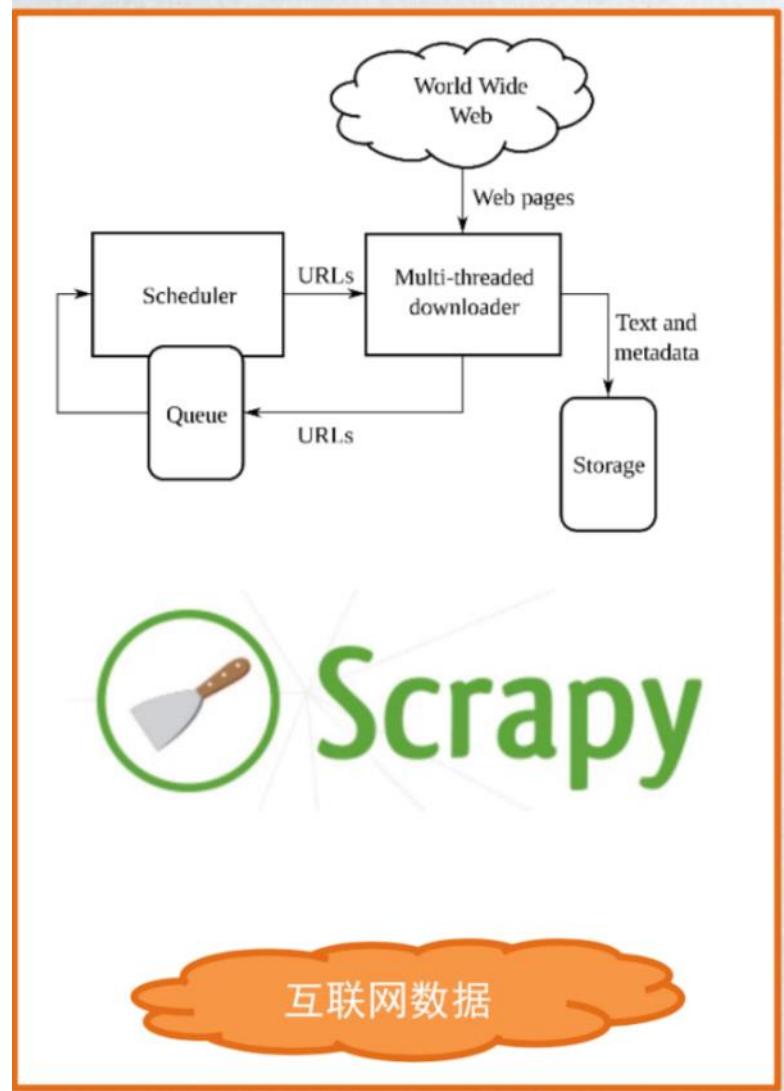
# 趋势3：人们处理数据的能力比以往任何时候都强大



从计算机的视野看数  
据科学+大数据

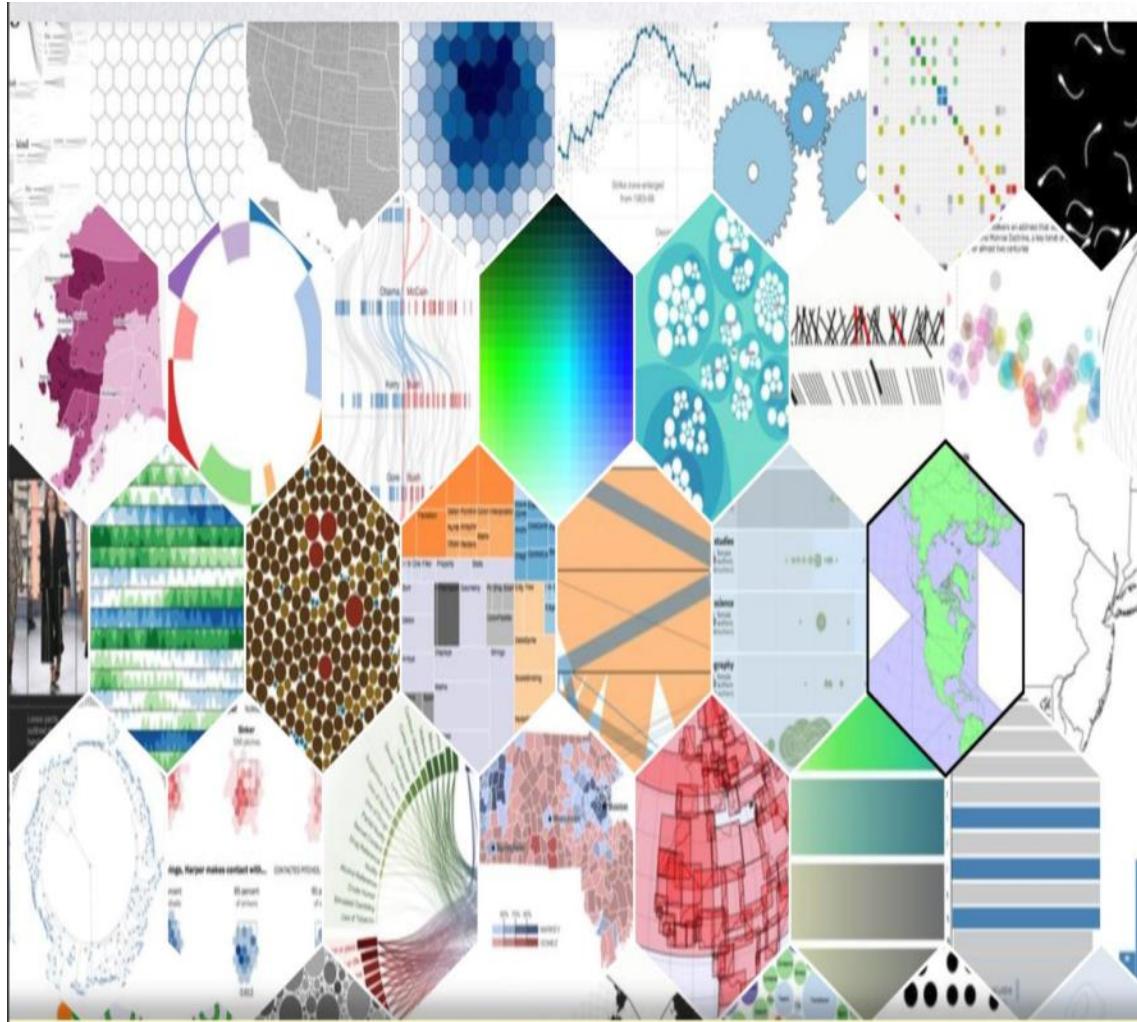


如果你对Python还不熟，强烈建议花一周的时间学习《CS221 Python Review Tutorial》：  
[https://colab.research.google.com/drive/1-9Z\\_dLRJBWZdKaMNLqBMF9TrXc1553IK?usp=sharing](https://colab.research.google.com/drive/1-9Z_dLRJBWZdKaMNLqBMF9TrXc1553IK?usp=sharing)





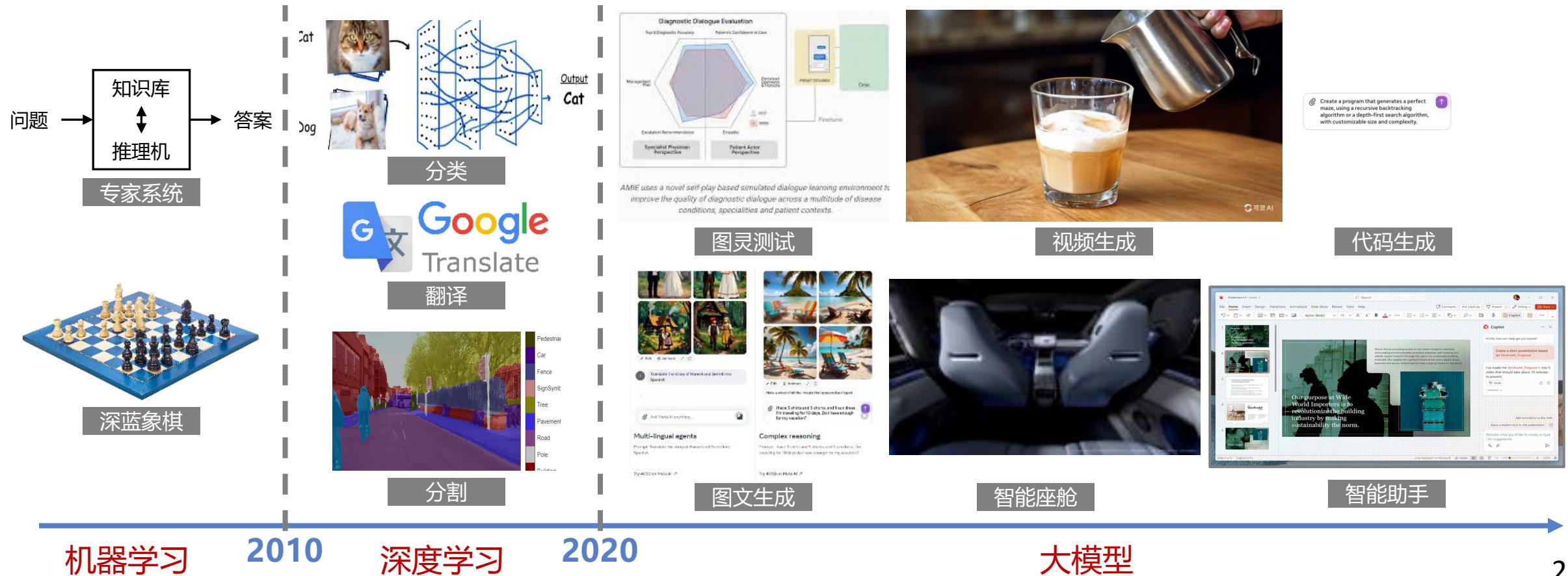
# 数据可视化



开源工具	商业工具
Processing	Tableau
Many Eyes	Spotfire (TIBCO)
D3.js	QlikView
R (基础包、lattice、ggplot2)	FineBI
ECharts	SAS Visual Analytics
Google Charts	Adobe Illustrator
Flot	HighCharts
Gephi	iCharts
Envision.js	Joliccharts
Prefuse	Dundas Chart
Arbor.js	LightningChart
Chart.js	Microsoft Excel
Paper.js	
NodeBox	

## 趋势4：大模型正在成为新的数据处理引擎

- 以ChatGPT为代表的大模型开启了通用人工智能时代：Instruction Following, In-context Learning, World Knowledge, Reasoning, Planning ...



- 比尔盖茨：ChatGPT是生成式人工智能的“iPhone Moment”

鸡兔同笼，头共10，足共28，鸡兔各几只？

首先，我们设鸡为x只，兔为y只。

由题意得：

1. 鸡和兔的头数之和是10，即：  
 $x + y = 10 \dots\dots(1)$
2. 鸡有2只脚，兔有4只脚。总脚数为28，即：  
 $2x + 4y = 28 \dots\dots(2)$

从(1)式，我们可以解出 $y = 10 - x$

将这个值代入(2)式中，得：  
 $2x + 4(10 - x) = 28$   
 $2x + 40 - 4x = 28$   
 $-2x = -12$   
 $x = 6$

所以，鸡有6只。

再代入(1)式，得：  
 $y = 10 - 6 = 4$

所以，兔有4只。

答案：鸡有6只，兔有4只。

## 2023年上海高考作文题

24.一个人乐意去探索陌生世界，仅仅是因为好奇心吗？请写一篇文章，谈谈你对这个问题的认识和思考。

要求：(1) 自拟题目；(2) 不少于 800 字。

**人之探陌，岂惟好奇？**此一问也，可谓深思。夫好奇者，人之所以探陌之大动也，其能激发人之求知之欲，追寻未知之境。然则人之探陌之动，非一而足，或有多重，或有相悖。

**首焉，好奇固为人之探陌之要因。**如大哲爱氏言：“吾无他才，惟好奇耳。”好奇者，一种强欲也，使人寻求所问，解开所谜。试思之，闻有新奇之地、未知之文、神秘之事，岂不心怀好奇？欲知其地何状、其文何异、其事何因。是时也，好奇之妙用大显，令人起探陌之念。

**次焉，探陌亦与人之成长发展相关。**探陌途中，必遭艰难险阻，需不断学习适应。此种挑战促人成长、发展技能、提升自我。由探陌而得，能广其见识，增其理解包容，养其开明灵活之思维。此种成长发展之机遇，亦为人之探陌之重动也。

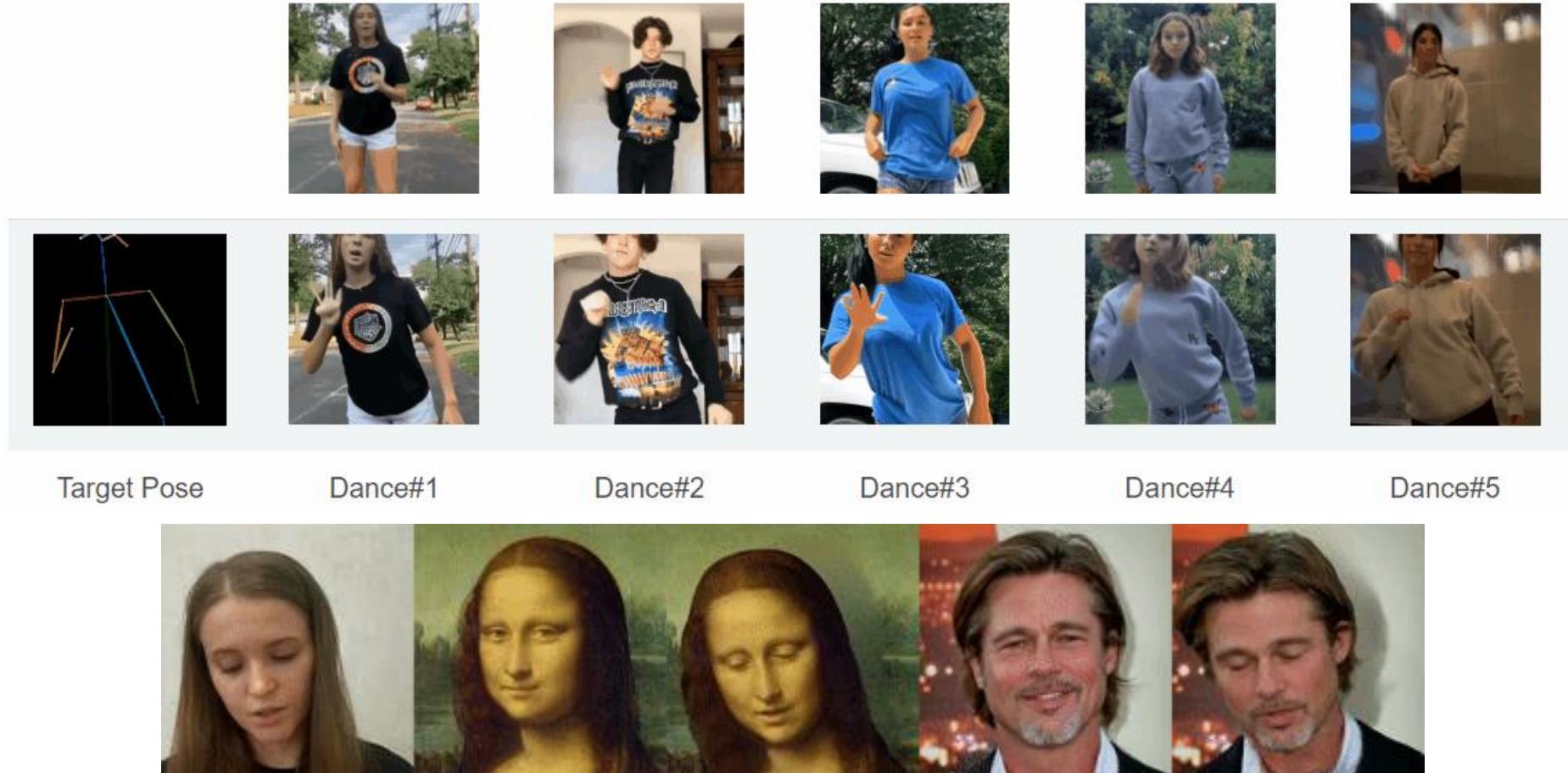
**又焉，探陌亦与人情社交有关。**在探陌中，常识新友，与其交流分享经历见解。此种社交互动可以丰富人之交游网，建立更广泛之人脉。且夫与异文异俗之人交流，能增进对他文他俗之理解尊重，促进跨文跨俗之沟通合作。故人亦因此社交因素而乐于探陌。

**末焉，探陌亦或出于消极被动之因。**或有人以不满逃避而探虚拟陌或他领<sup>2</sup>；或有人以恐惧无奈或迫于压力而探异质环；或有人以受外诱误导而探危殆不义之陌。此等探陌之动非善，甚或致不良之果。

综上所述，人之探陌岂惟好奇？**好奇诚为重要推动力也，而探陌之乐、成长发展以及人情社交亦起关键作用也。**同时亦当警惕消极被动之探陌动力，勿走向谬途。探陌乃一种全面体验也，涉及人心理需求、成长发展以及社交关系等多方面也。故**无论好奇力何如强盛也，皆可知人之探陌原因远不止于此。**

## 趋势4：大模型正在成为新的数据处理引擎

- 多模态大模型近年来发展迅速



# 趋势4：大模型正在成为新的数据处理引擎

- 业界愈发依赖大模型完成数据标注、预处理、生成、智能分析等任务

## 学习工作：

- Microsoft 365 Copilot: work, excel, ppt
- Github Copilot : 代码补全，自动写代码
- Chatpdf : 自动分析解读pdf文件
- 文章润色、文章内容总结

## 日常生活：

- 智能客服：电商平台的客服机器人
- 生活智能助手：查询天气、美食、服装
- 内容创作：生成笑话、文案、剧本、歌曲

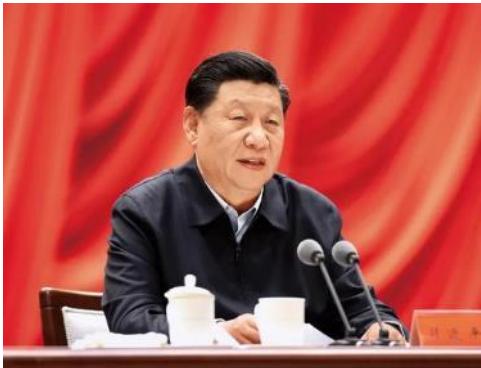
## 医疗健康：

- 心理辅导：提供心理咨询，情感支持
- 辅助医生诊断，提供建议



## 趋势5：数据正在成为重要的生产要素

- 我国正在大力推进数据要素市场的发展：数据->要素化->流通交易->释放价值



2019年10月31日  
中共中央  
十九届四中全会

数据可作为生产要素  
按贡献参与分配

2022年12月19日  
中共中央 国务院  
《关于构建数据基础制度  
更好发挥数据要素作用的意见》

数据要素里程碑  
《数据二十条》正式印发

- 数据要素的概念
  - 数据要素是指以电子形式存在的、通过计算的方式参与到生产经营活动并发挥重要价值的数据资源。

### 生产要素发展历程

十七世纪

农业（土地）是一切  
财务的源泉

十八世纪  
70年代

“资本和劳动对国民财富的  
影响” 亚当·斯密《国富论》

二十世纪  
20年代

技术是生产要素

二十一世纪

数据是生产要素

## 趋势5：数据正在成为重要的生产要素

- 我国正在大力推进数据要素市场的发展：数据->要素化->流通交易->释放价值

“数据二十条” 中强调数据流通与收益分配的重要性

### 流通交易 相关政策

“  
（八）支持探索多样化、符合数据要素特性的**定价模式和价格形成机制**，推动...信息数据市场自主定价。  
”，

“  
（九）统筹构建规范高效的**数据交易场所**，建立健全数据交易规则。促进**区域性数据交易场所**和**行业性数据交易平台**与**国家级数据交易场所**互联互通。  
”，

### 四项制度

促进数据合规高效流通使用、赋能实体经济

数据产权

流通交易

收益分配

安全治理

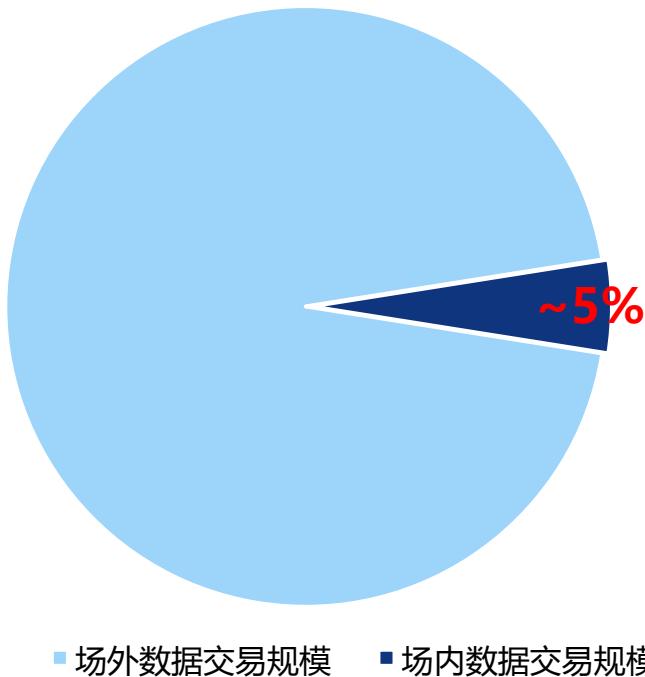
### 收益分配 相关政策

“  
（十二）健全数据要素**市场评价贡献机制**。强化基于数据价值创造和价值实现的激励导向。  
”，

“  
（十三）建立保障公平的**数据要素收益分配体制机制**，更加关注公共利益和相对弱势群体。加大政府引导调节力度，探索建立公共数据资源开放收益合理分享机制。  
”，

## 趋势5：数据正在成为重要的生产要素

- 但是，当前数据要素市场并不活跃，面临着众多挑战：政策、法规、技术 etc.



我国场内数据交易规模仅占数据交易行业总额的约**5%**  
——《数据要素市场生态体系研究报告（2023年）》



以“关键词搜索”模式为主

数据交易场所交易不活跃的原因：

- 数据权属体系构建不完善
- 数据质量与合规性难保证
- 数据供需匹配技术不成熟
- 数据定价与竞价机制不完善

... ...

- 近年来的五个重要趋势
  1. 数据比以往任何时候都更容易产生与获取
  2. 人们的决策比以往任何时候都更基于数据驱动
  3. 人们处理数据的能力比以往任何时候都强大
  4. 大模型正在成为新的数据处理引擎
  5. 数据正在成为重要的生产要素
- 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，但我们认为，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洞见之间的矛盾这一点，不会改变。

- 数据科学概述

- 数据科学是如何兴起的
- 数据科学家应该具备什么样的能力
- 教学计划和考核要求

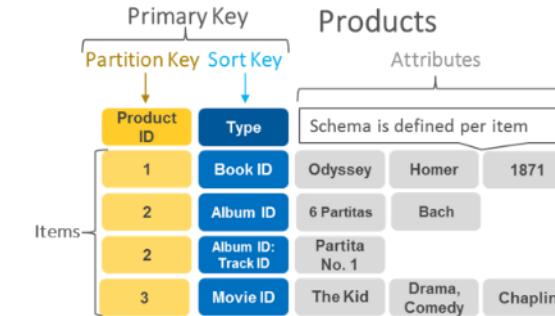
- 数据科学基本内容简介

- 机器学习
- 关联规则挖掘
- 自然语言处理
- 图和社交网络分析
- 分布式计算

# 管理与处理各种类型的数据

- 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
  - ...

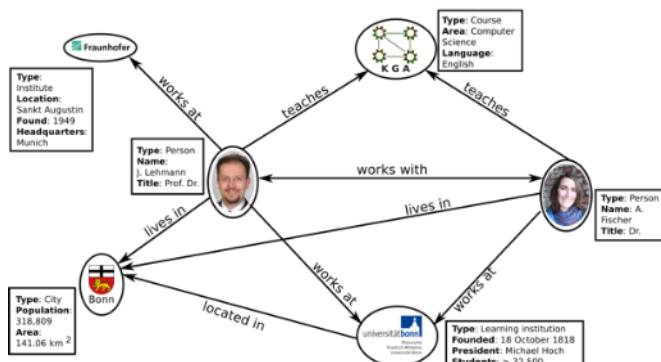
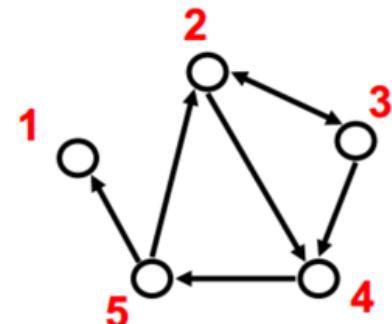
	$item_1$	$item_2$	$item_3$	...	$item_n$
$user_1$	5	2			1
$user_2$	3				
$user_3$	1		3		
.					
.					
$user_{m-1}$	5		4		2
$user_m$		4			3



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

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》网站最新发布的消息，超过40%的昆虫物种可能在未来几十年内灭绝，其中蝴蝶、蜜蜂和蜣螂受到的影响最大，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础。”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。



## 西班牙电信：数据变现



### Smart Steps

2012年成立大数据部门：Telefonica Dynamic Insights。推出了名为“Smart Steps”的产品，通过脱敏的用户位置数据，可以对某个时段、某个地点人流量的关键影响因素进行分析。“Smart Steps”为零售商新店设计和选址、设计促销方式、与客户反馈等提供决策支撑，从而帮助零售商更好地理解和满足客户需求、降低成本；也可帮助市政委员会统计、预测各种场景下的人流量。

### 运输模型

利用实时数据，预测人流量，出行模式以及分析交通网络的变化。

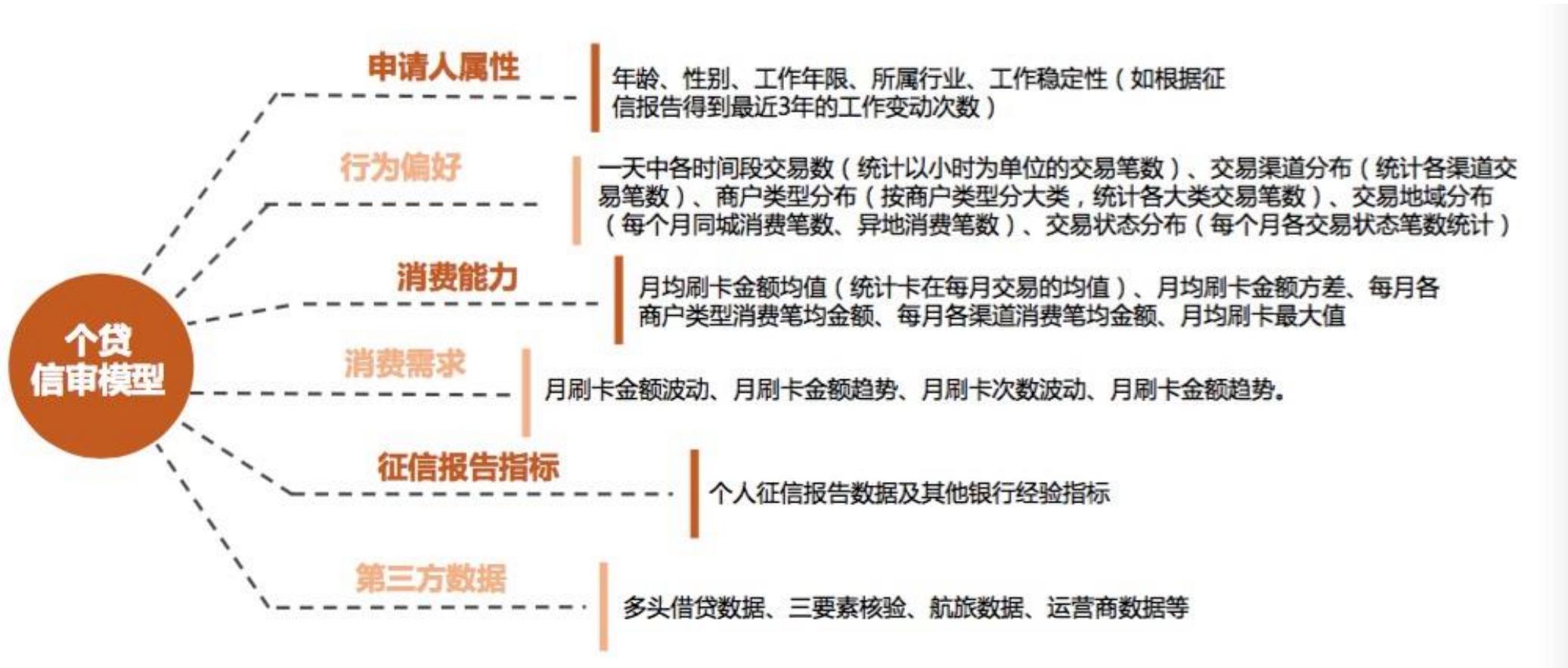
### 交通建设规划

提供旅途时间，起始、目的地的人流热力图等。

### 衡量经济发展

通过人流量辅助分析经济运行状况。

## 消费金融：用户画像和信贷评估

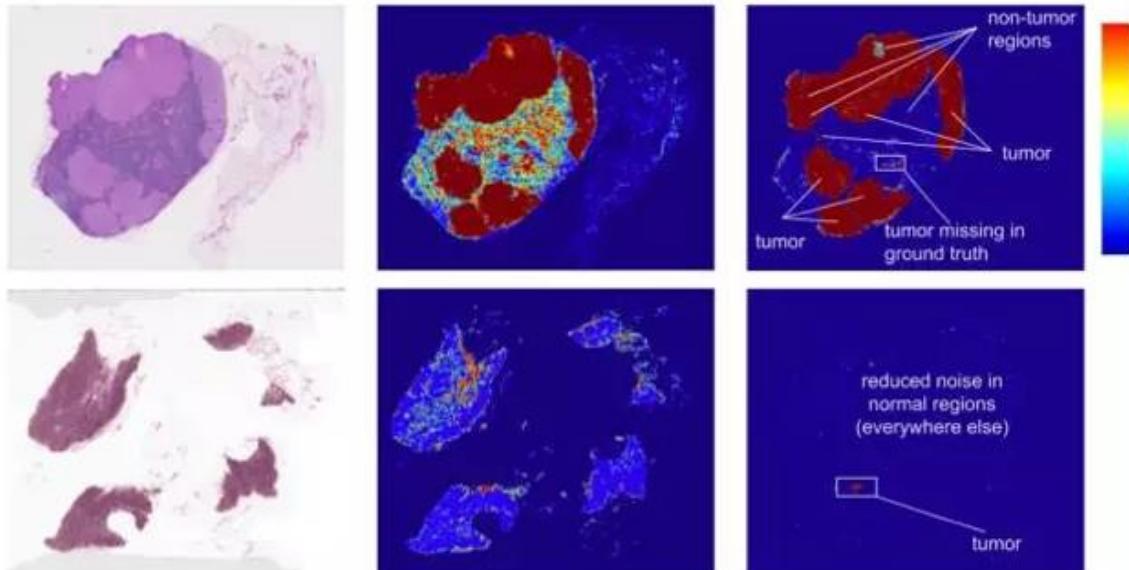


模型评估后决定是否给予信贷

普林科技：开源金融银行卡信用评估

## 健康医疗：皮肤癌诊断

- 大量患者切片的数据，以及何处病患的标记，训练集充足
- 切片一般都是高清晰度的，一张切片有上千万甚至上亿像素，不便直接训练，专家们将照片切割成了 $128 \times 128$ 像素的标准大小



Andre Esteva, et al. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* (2017)

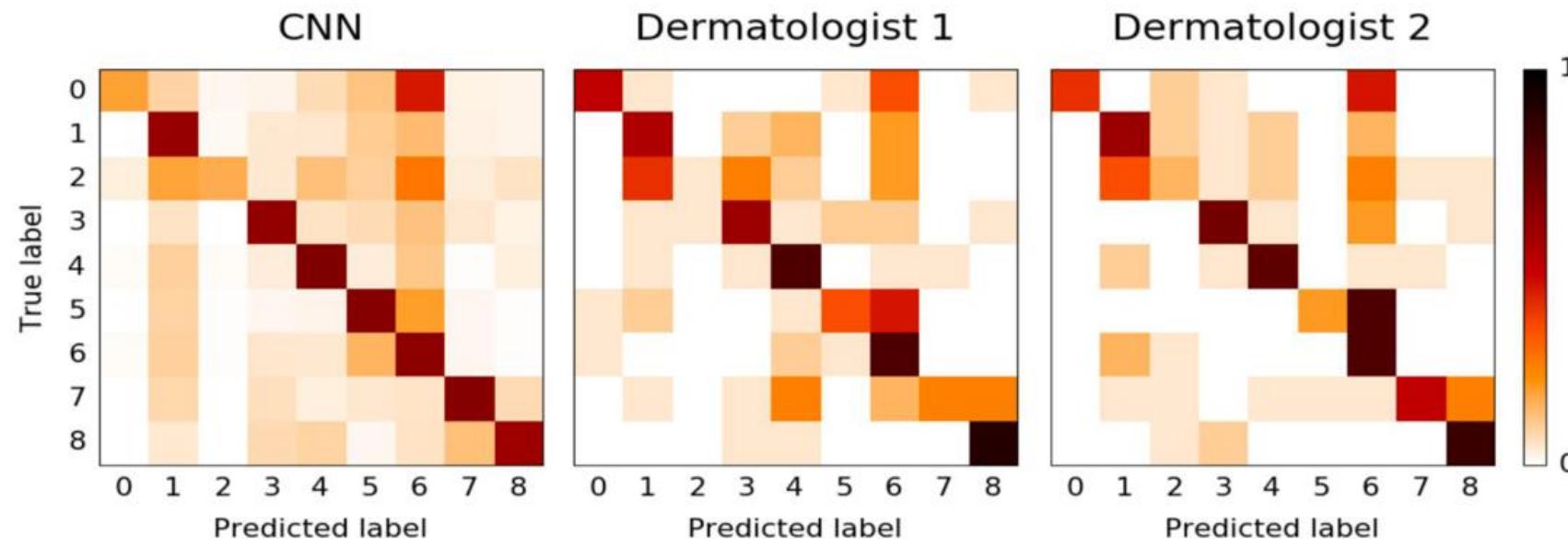
## 健康医疗：皮肤癌诊断

深度学习: 129,450 临床图片, 757 种疾病

分类准确率: 72.1%

人类病理学家准确率: 66%左右。

基于CNN的方法诊断准确率超过了人类病理学家的平均准确率。



- 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

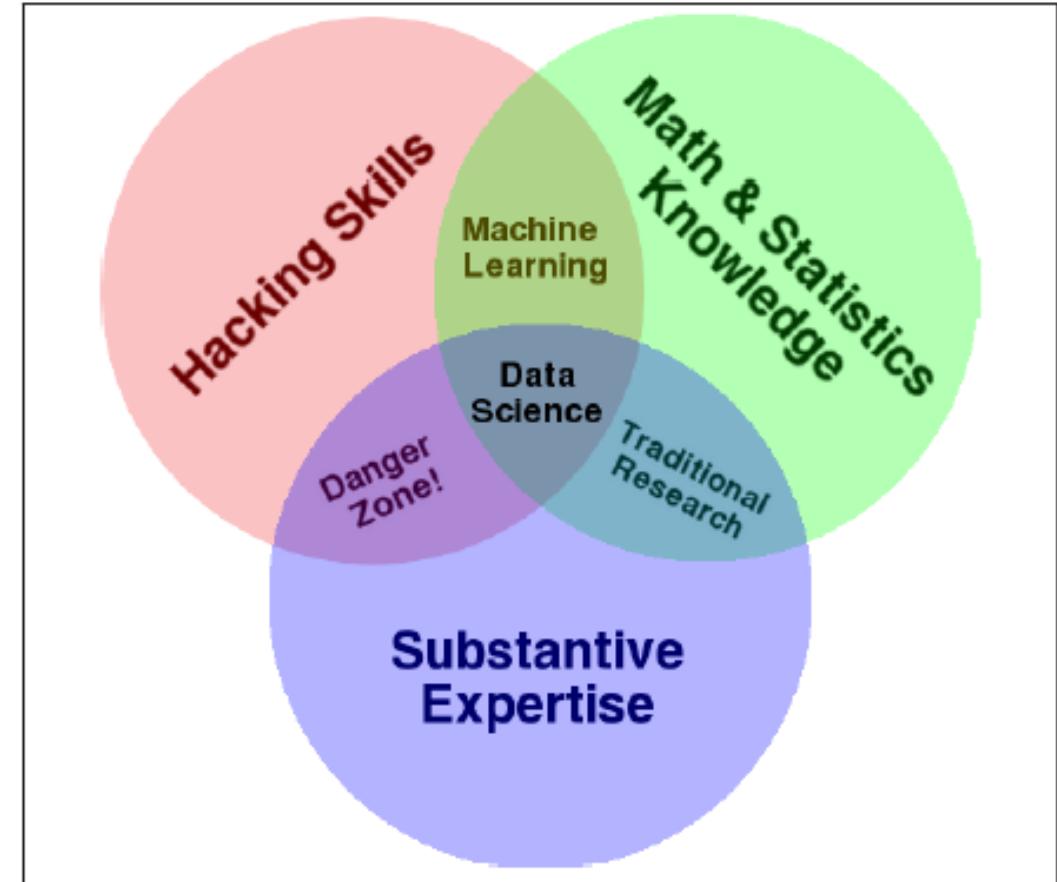


Figure 1-1. Drew Conway's Venn diagram of data science

- 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**

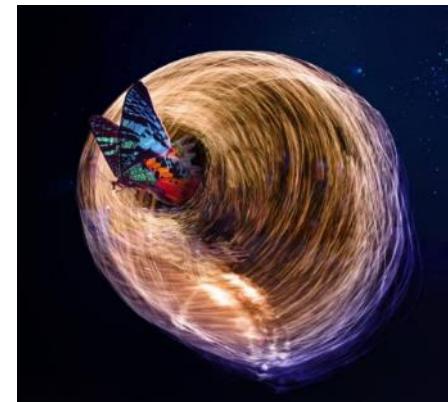
- 千级别(thousand)数据样本
  - 无需编程自动化处理
- 百万级别(million)数据样本
  - 自动处理
    - 低于 $O(n^2)$ 的算法
    - 并行化后的 $O(n^2)$ 算法
- 十亿级别(billion)数据样本 (Web-scale)
  - 如何存储开始成为问题 (分布式存储)
- 万亿级别(trillion)数据样本 (大模型的数据训练体量)
  - 数据很难存储在同一个物理地址
  - 分布式处理, 但是又要考虑容错等机制, 复杂性进一步上升
  - 几乎无法得知数据的全貌(难以实现有效采样)
  - 数据隐私/不一致/数据分布偏斜都成为问题

- 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

## • 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。



观点频道 人民网评 图解 原创快评 刺激专栏 网友来论 报系言论 每日新评 观点1+1 学习知新 治国理政 投稿信箱

人民日报评论 人民日报社论 任仲平 评论员 今日谈 人民视点 人民论坛 人民时评 望海楼 国纪平 睿评 检索 刺激·国情

### 人民网三评算法推荐：警惕算法走向创新的反面

人民网二评算法推荐：别被算法困在“信息茧房”

- 外卖环境成本如何控制
- 全方位管住塑料垃圾污染
- 实名制打击婚骗关键在落实
- 用强制性国标遏止月饼市场乱象
- 规避试用期行为须打防并举
- 高楼烂尾，管理责任不能烂尾

人民网一评算法推荐：不能让算法决定内容

- 把恶心当有趣，“共享女友”行之不远
- 红牛之争启示现代企业重视无形资产
- 车牌拥有冷暖不均，摇号政策还需完善

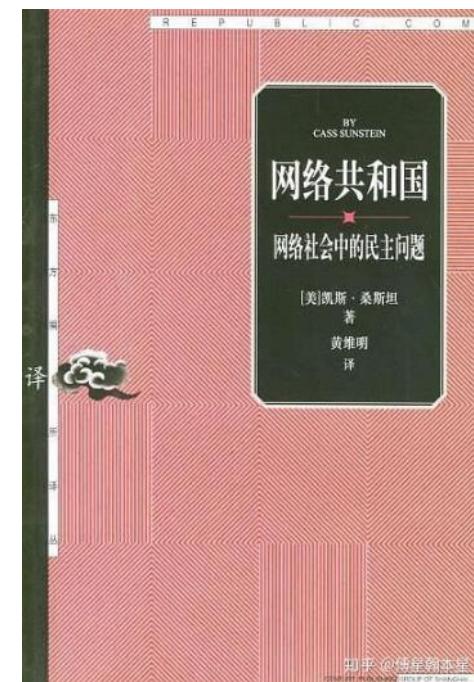
2017党报评论融合发展论坛

2017党报评论融合发展论坛在京举行

人民网评 人民日报要论

人民网评：老支书看升旗落泪告诉我们什么

文化产业新闻



- 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

- 数据科学概述

- 数据科学是如何兴起的
- 数据科学家应该具备什么样的能力
- 教学计划和考核要求

- 数据科学基本内容简介

- 机器学习
- 关联规则挖掘
- 自然语言处理
- 图和社交网络分析
- 分布式计算

- 课程定位：数据科学系列课程的先导/基础课
- 课程目标
  - 让同学们对数据科学有一个整体的认识
  - 训练同学们使用Python的生态工具从头到尾地完成一个项目
- 数据科学基础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！

- 管理和处理各种类型的数据
  - 文本、图、Web、用户行为、关系、流数据、时间序列.....
- 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：文本分类、图中心性分析.....
- 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.....
- 了解一些数据科学领域的学术前沿
  - 知识图谱、因果、大模型、区块链、大数据存储

**红字内容是  
课程覆盖**

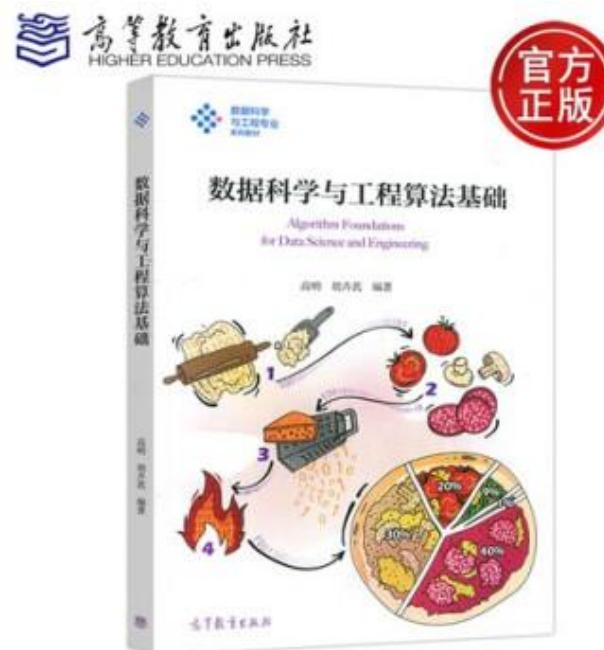
- 数据库系统与技术 (如SQL查询)
  - 实际上，数据科学家需要非常熟练的掌握数据库技术
  - 我们把这部分知识留给后续的数据库相关课程
- Python程序设计与数据分析编程实践
  - 实际上，这部分对成为一个数据科学家来讲非常重要
  - 我们认为你们能够通过自学+项目实践掌握基本的技能
- 复杂的机器学习与深度学习模型
  - 实际上，机器学习与深度学习正变得越来越重要
  - 我们会讲解机器学习的经典方法和基本思想，把更复杂的知识留给后续的课程
- 数据伦理问题
  - 目前学术界与工业界对数据伦理的研究尚待深入

## 参考教材：



<<数据科学导引>>

欧高炎 等编著  
高等教育出版社  
北京大学



<<数据科学与工程算法基础>>

高明 等编著  
高等教育出版社  
华东师范大学

## 课程计划：

1. 绪论&数据预处理
2. 回归&Python实践基础
3. 分类
4. 集成
5. 聚类
6. 关联规则&贝叶斯
7. 信息检索
8. 社会网络
9. 深度学习
10. 哈希&流数据
11. 知识图谱
12. 因果
13. 前沿-区块链/存储/大模型
14. 实践项目汇报I
15. 实践项目汇报II

# 课程考核

- 平时作业: 50%
  - 5次作业 (笔试)
- 实践项目: 50%
  - 1~2次实践项目 (编程)

教师:



何向南

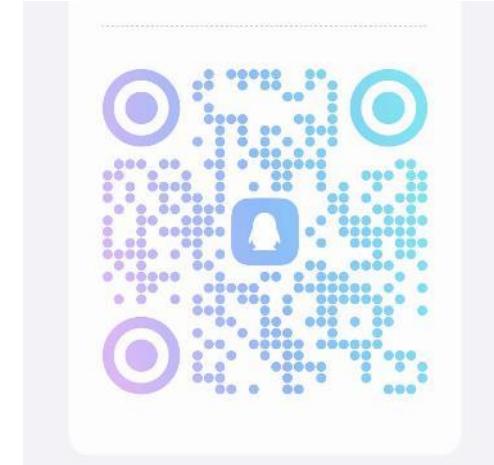
hexn@ustc.edu.cn

助教:



汪远博一  
wy1001@mail.ustc.edu.cn

课程QQ群 (群号: 547810062)



方羿研二  
peterfang@mail.ustc.edu.cn

- 数据科学概述

- 数据科学是如何兴起的
- 数据科学家应该具备什么样的能力
- 教学计划和考核要求

- 数据科学基本内容简介

- 人工智能与机器学习
- 关联规则挖掘
- 自然语言处理
- 图和社交网络分析
- 分布式计算

- 用数据的方法研究科学 (AI4Science)
  - 生命科学、物质科学、天体信息学等
- 用科学的方法研究数据
  - 统计学、机器学习、数据挖掘、人工智能



开普勒：分析数据产生价值



表 1 太阳系八大行星绕太阳运动的数据

行星	周期 (年)	平均距离	周期 <sup>2</sup> /距离 <sup>3</sup>
水星	0.241	0.39	0.98
金星	0.615	0.72	1.01
地球	1.00	1.00	1.00
火星	1.88	1.52	1.01
木星	11.8	5.20	0.99
土星	29.5	9.54	1.00
天王星	84.0	19.18	1.00
海王星	165	30.06	1.00

开普勒第三定律：  
行星绕太阳运行的周期的平方和行星离太阳的平均距离的立方成正比

## 主要包括：数据采集、存储、和分析

- 常见的数据类型

- 表格：最为经典的数据，e.g., 行代表样本、列代表特征
- 点集：很多数据都可以看成是某种空间的点的集合
- 时间序列：文本、通话和DNA序列等都可以看成是时间序列
- 图像：可以看成两个变量的函数
- 网页和报纸：每篇文章都可以看成是时间序列，整个网页或报纸又具有空间结构
- 网络数据：网络数据本质上是图，由节点和联系节点的边构成

数据分析的基本假设：观察到的数据都是由背后的一个模型产生

数据类型	模型
表格	有监督学习模型
点集	概率分布
时间序列	随机过程（如隐式马氏过程等）
网络	图模型、贝叶斯模型

- 数据分析的主要困难
  - **数据量大**
  - **维数高 (核心困难)**
    - 维数灾难：模型复杂度和计算量随着维数的增加而指数增长
    - 如何克服？
      - 将模型限制在一个极小的特殊类里面，如线性模型。
      - 利用数据可能有的特殊结构(例如稀疏性，低维或低秩，光滑性等)，通过正则化和降维来实现。
      - 先验假设：文本大模型 -> predict next word
  - **类型复杂**：表格、图像、文本、视频
  - **噪音大**：数据在生成、采集、传输和处理等流程均可能引入噪音

- 算法的重要性
  - 与模型相辅相成的是算法，以及算法在计算机上实现
  - 从算法角度看，处理大数据有两条思路
    - 降低算法的复杂度：
      - 如随机梯度下降等
    - 分布式计算：
      - 把大问题分解成小问题，然后分而治之. 如著名的MapReduce框架

# 1. 人工智能

- 人工智能的目标：让机器（像人一样）能够完成复杂任务

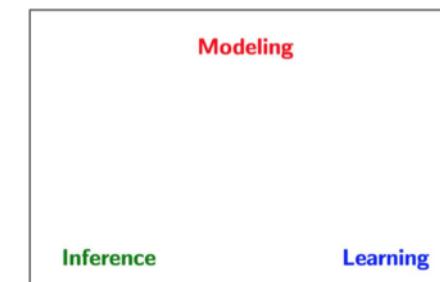


Artificial Intelligence is a field of study that enables machines to **mimic** human "cognitive" functions, such as "learning" and "problem solving".



人工智能是**计算机**科学的一个分支，它企图了解智能的实质，并生产出一种新的能以**人类智能**相似的方式做出反应的智能机器，

- **Modeling - Inference - Learning paradigm**



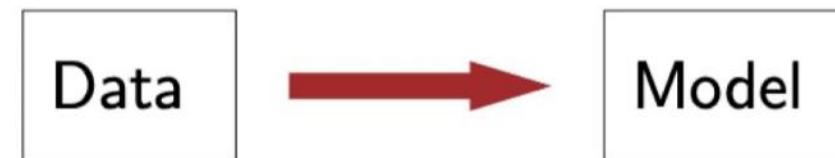
"Learning is any process by which a system improves performance from experience."

(学习是系统从经验中提高性能的任何过程)

——Herbert Simon (赫伯特·西蒙/司马贺)  
图灵奖、诺贝尔奖得主，计算机科学家

We define *machine learning* as a set of methods that can automatically **detect patterns in data**, and then use the uncovered patterns to **predict future data**, or to **perform other kinds of decision making under uncertainty** (such as planning how to win a game).

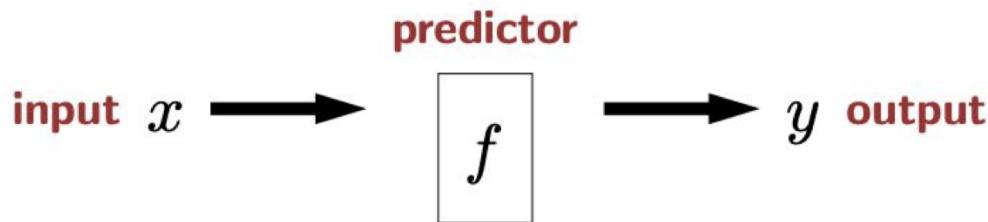
— 《Machine Learning: A probabilistic perspective》



## Learning from Data

- 人工智能近期成功的主要驱动力
- 将复杂性从“代码”转移到“数据”
- 需要泛化能力

- **Given (input)**
  - 样本的收集 (文档, 测量, 基因序列 ...)
  - 样本的编码 (字符串, 图片, 数据库记录, ...)
- **Hypothesis (model)**
  - 模型应当描述样本自身和样本之间的关系
  - 模型应当刻画来自同一来源/分布的样本
  - 模型可以做什么?
    - 图像分类: 细胞图像 → 是否为癌症
    - 文本分类: 文档 → 是否描述相似主题
    - 自动驾驶: 周围环境的激光雷达图像 → 汽车的路径规划
    - ...

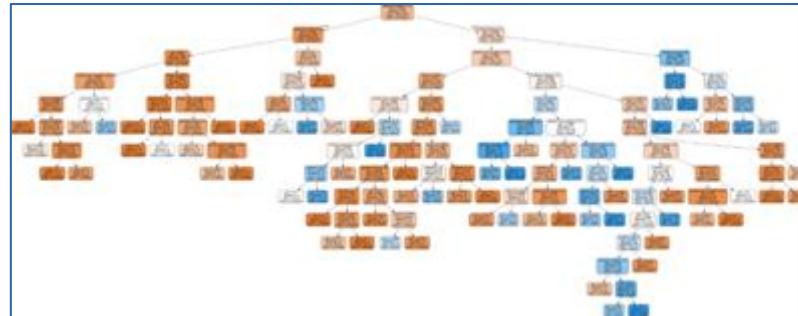


- **Inference (estimate, predict)**
  - 使用样本输入学习 (估计) 模型的参数
  - 根据 (新) 样本输入预测输出

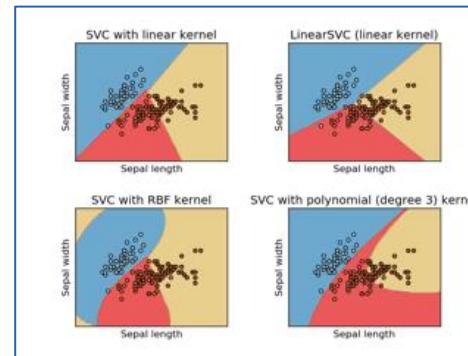
## 预测、决策

数据

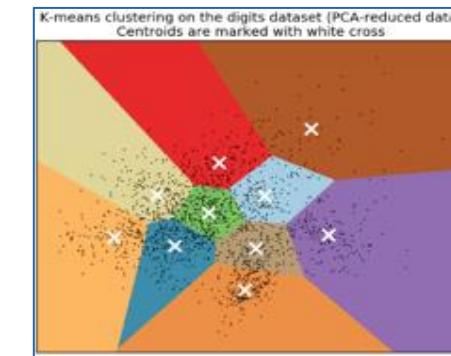
mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	origin	carname
18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11	70	1	plymouth satellite
16	8	205	150	3229	18	70	1	amc rebel st
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10	70	1	ford galaxie 500
14	8	454	220	4354	9	70	1	chevrolet impala



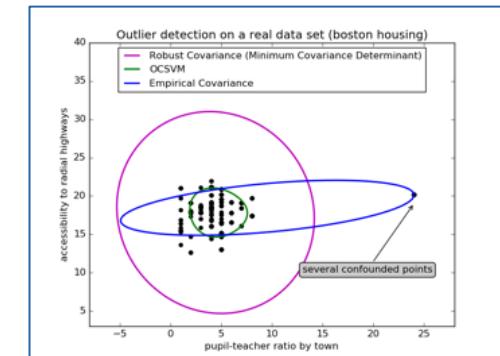
分类



聚类



异常点检测



- 基本概念

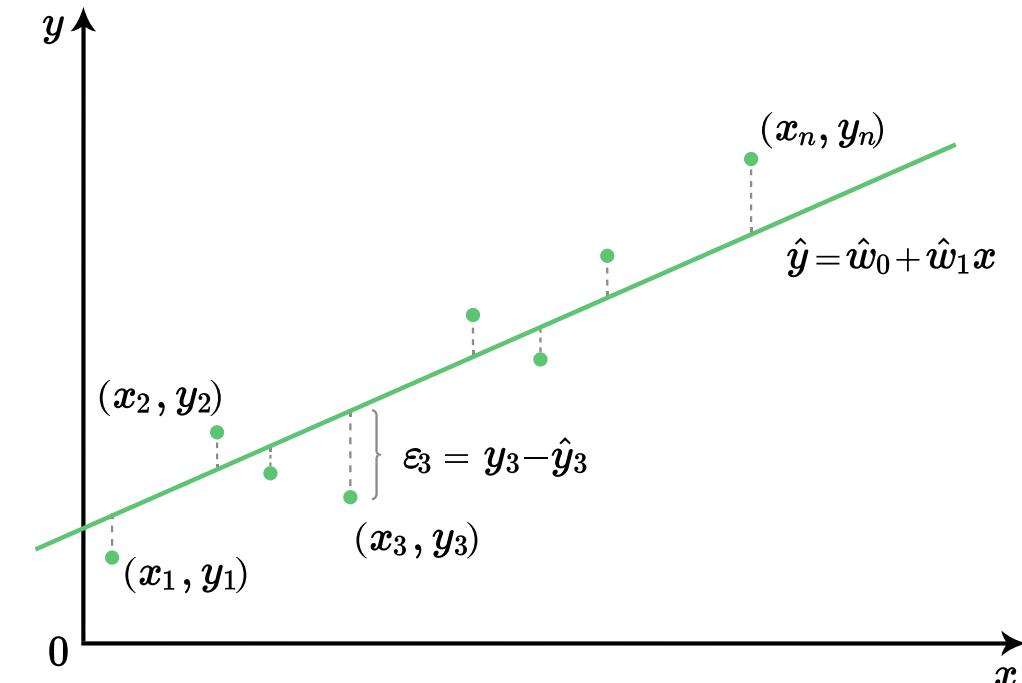
- 训练集：用于训练模型的数据集
- 测试集：用于测试模型的数据集
- 模型：建立数据的输入 $x$ 和输出 $y$ 之间的映射关系  $y = f(x)$
- 损失函数： $L(f(\mathbf{x}_i), y_i) = (f(\mathbf{x}_i) - y_i)^2$
- 优化目标：

$$\min_{f \in F} \frac{1}{n} \sum_{i=1}^n L(y_i, f(\mathbf{x}_i))$$

- **有监督学习** (supervised learning)
  - 数据集中的样本带有标签，有明确目标
  - 回归和分类
- **无监督学习** (unsupervised learning)
  - 数据集中的样本没有标签，没有明确目标
  - 聚类、降维、密度估计、关联规则挖掘
- **强化学习** (reinforcement learning)
  - 智慧决策的过程，通过过程模拟和观察来不断学习、提高决策能力
  - 例如：AlphaGo（蒙特卡洛树搜索）
- **自监督学习** (self-supervised learning)

- 数据集中的样本带有标签
- 目标：找到样本到标签的最佳映射
- 典型方法
  - **回归模型**：线性回归、岭回归、LASSO和回归样条等
  - **分类模型**：逻辑回归、K近邻、决策树、支持向量机等

- 典型的有监督任务，样本的标签为连续型，如收入、销量等
- 应用场景：
  - 流行病学：吸烟对死亡率和发病率影响的早期证据来自采用了回归分析的观察性证据
  - 金融：资本资产定价模型利用线性回归以及Beta系数的概念分析和计算投资的系统风险
  - 经济学：预测消费支出，固定投资支出，存货投资，一国出口产品购买，劳动力需求，劳动力供给



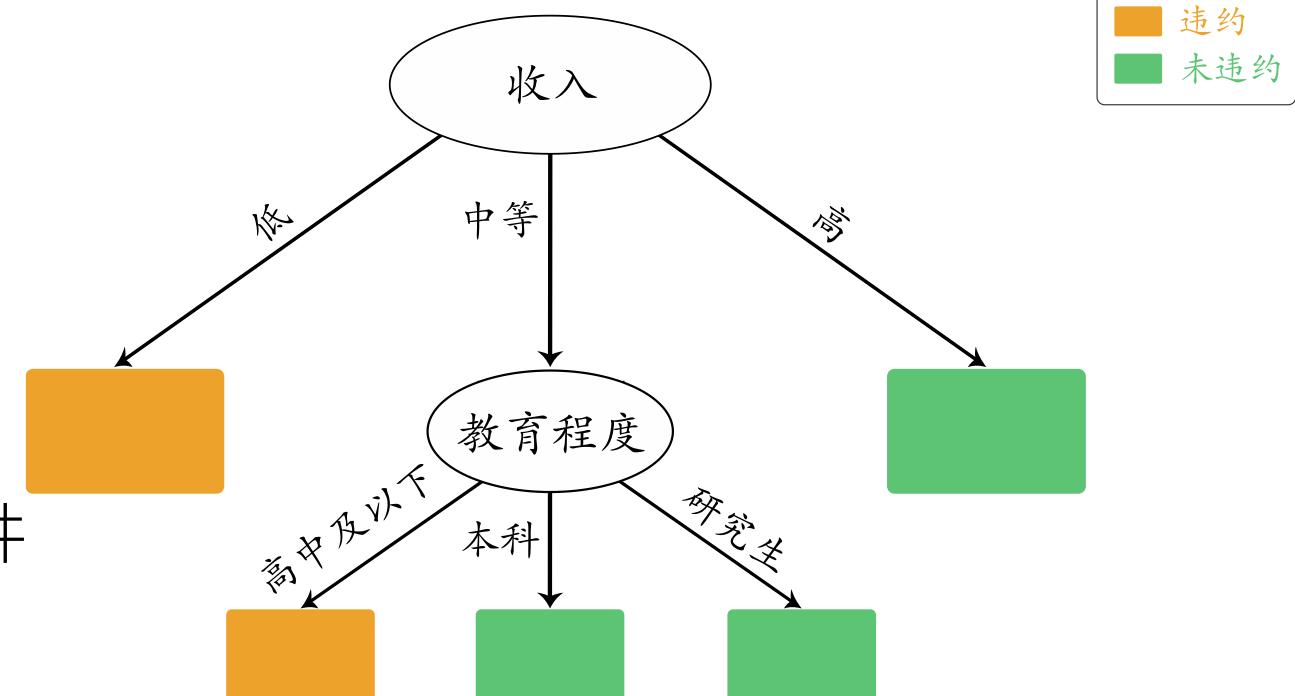
- 典型的有监督学习任务，样本标签为离散型。

包括二分类和多分类问题。

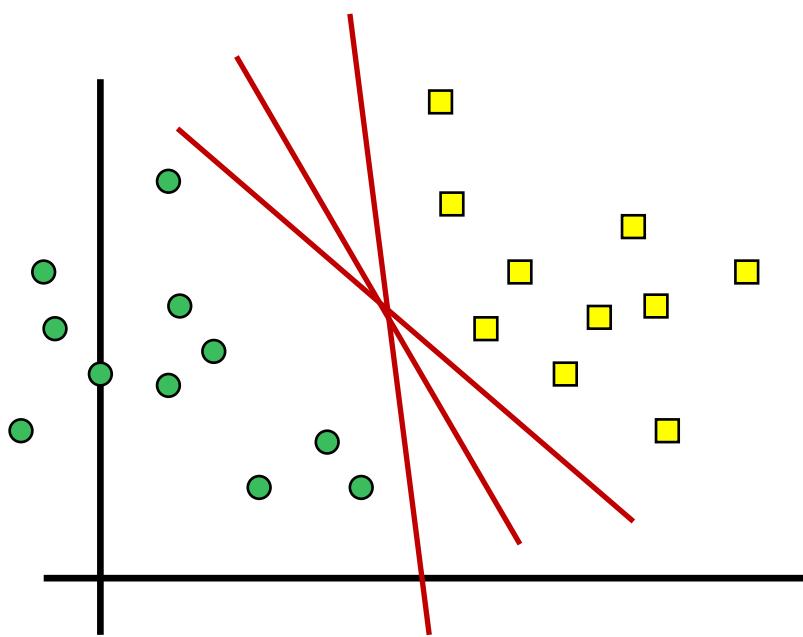
- 应用场景：

- 信用风险评估
- 预测肿瘤细胞是良性还是恶性
- 邮件的分类：正常邮件/垃圾邮件
- 客户流失预测

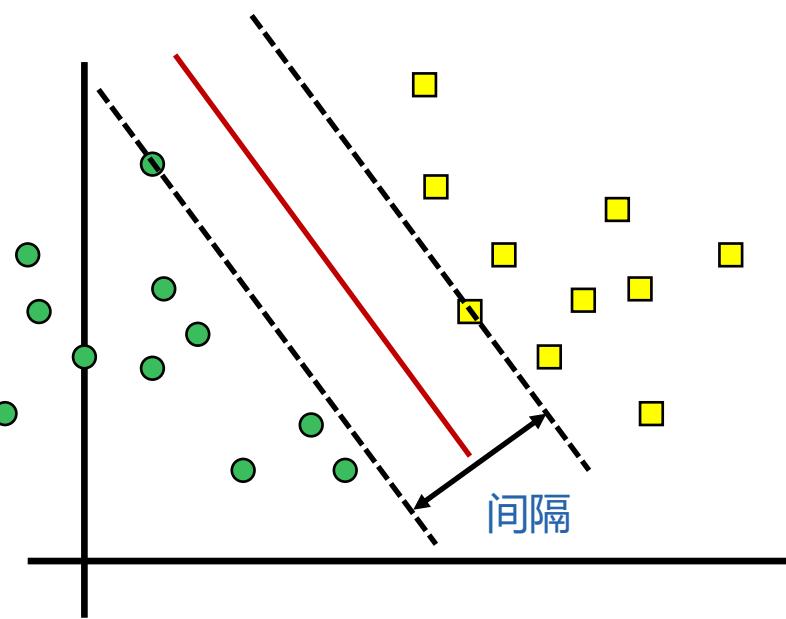
决策树——贷款违约预测



线性可分数据

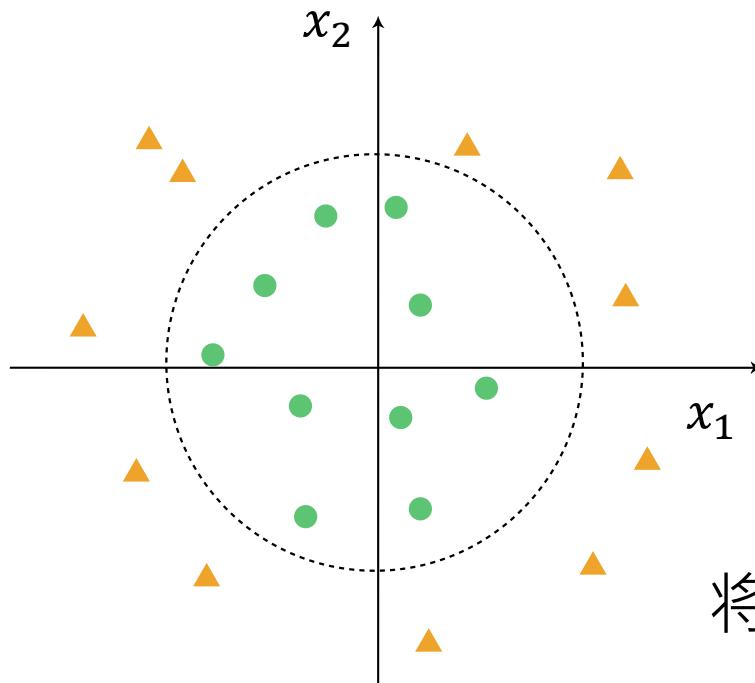


哪一条分割线更好?



具有最大间隔的分割线是最好的!

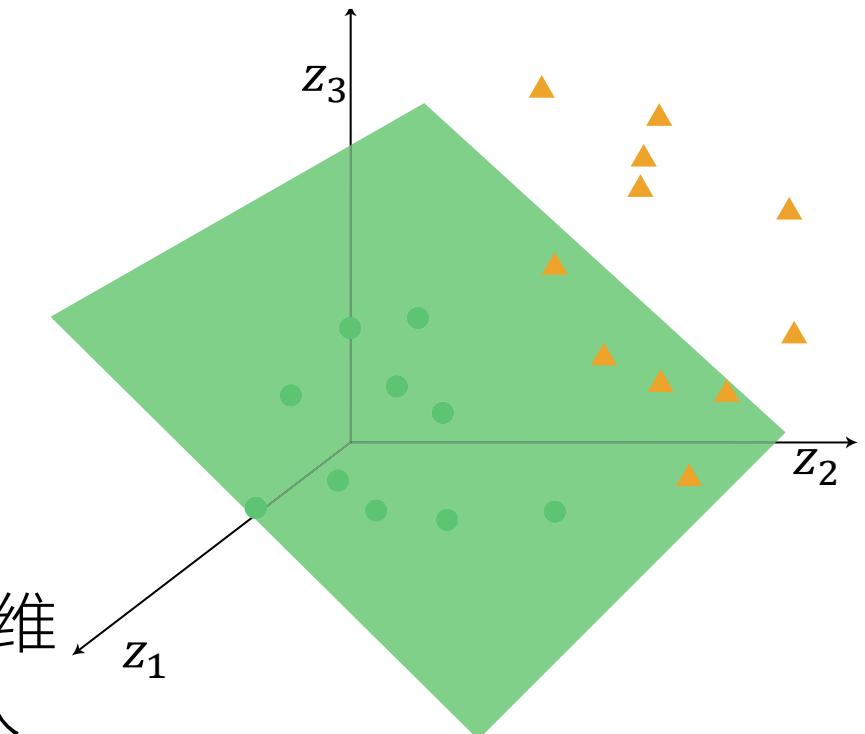
如果数据并不是线性可分的？



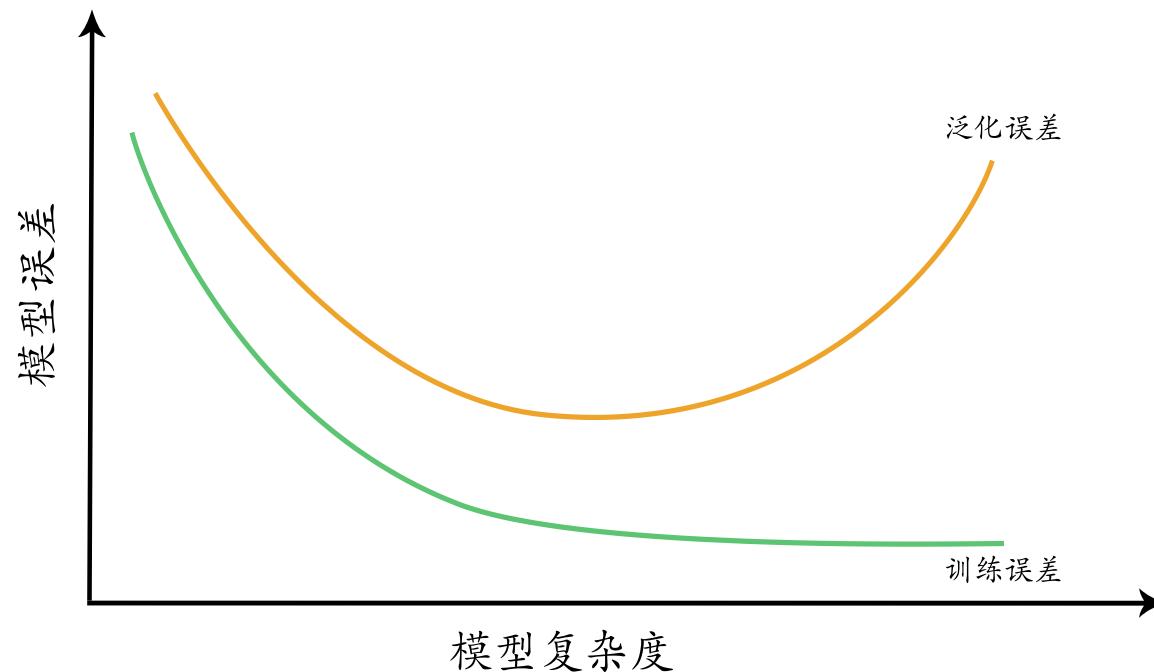
映射trick!



将数据点从2维空间映射到3维  
空间中，使得数据线性可分



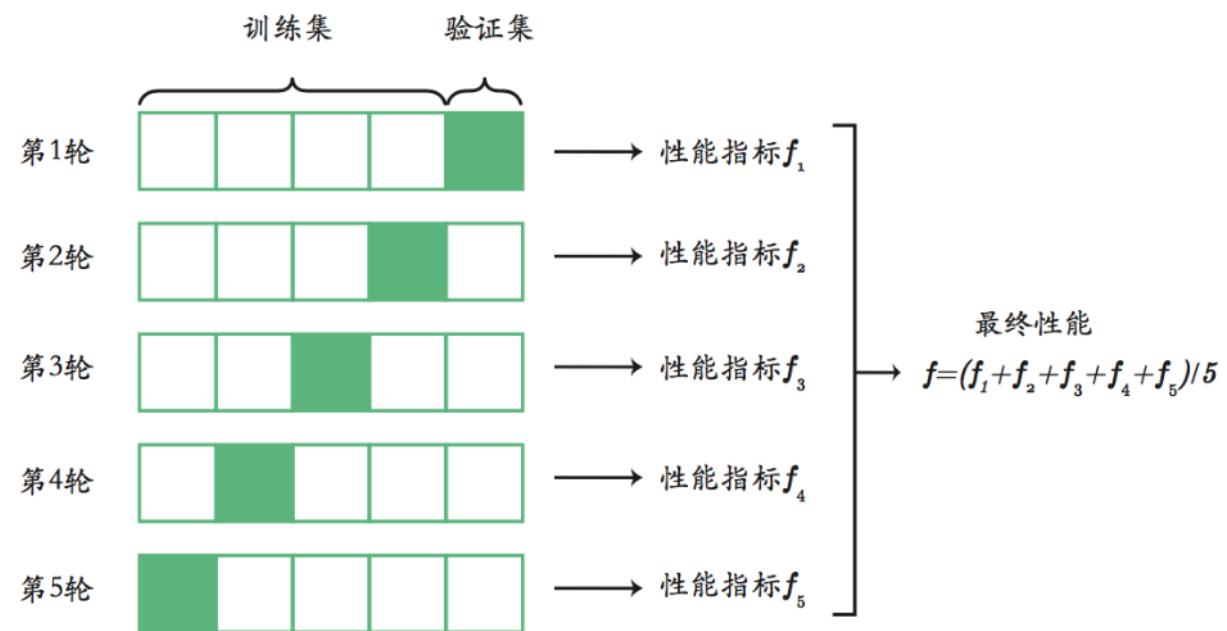
- 模型过于复杂(例如参数过多), 导致所选模型对已知数据预测得很好, 但对未知数据预测很差。



- 正则化:  $\min_{f \in F} \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)) + \lambda J(f)$
- 正则化是模型选择的典型方法
- 在误差函数上加一个正则项, 正则项通常为参数向量的范数
- 在训练误差和模型复杂度之间的权衡

- **交叉验证**: 基本想法是重复地使用数据。将数据集随机切分，将切分的数据集组合为训练集和测试集，在此基础上反复进行训练，测试和模型选择。
- **K折交叉验证 (k-fold cross validation)**

- 随机地将数据切分为 $k$ 个互不相同大小相同的子集；
- 每次利用 $k-1$ 个子集的数据训练模型，余下的数据测试模型；
- 最后选择在 $k$ 次测评中平均性能最好的模型。

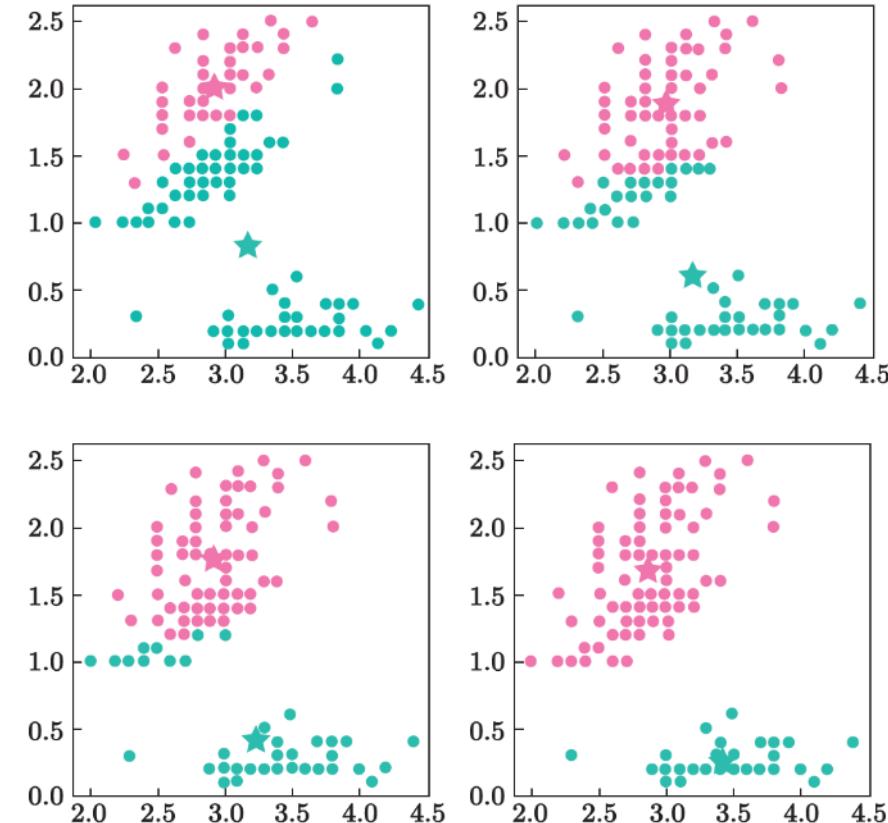


- 可以处理没有标签的数据
- 根据数据本身的分布特点，挖掘反映数据的内在特性
- 典型方法
  - 聚类、降维、关联规则挖掘等

- 目的：将数据集中相似的样本进行分组，使得：
  - 同一组对象之间尽可能相似；
  - 不同组对象之间尽可能不相似。
- 应用场景：
  - **基因表达水平聚类**：根据不同基因表达的时序特征进行聚类，得到基因表达处于信号通路上游还是下游的信息
  - **篮球运动员划分**：根据球员相关数据，将其划分到不同类型（或者不同等级）的运动员阵营中
  - **客户分析**：把客户细分成不同客户群，每个客户群有相似行为，做到精准营销

## K-Means聚类

- 1. 选择K个点作为初始质心
- 2. Repeat:
  - 将每个点指派到最近的质心，形成K个簇
  - 重新计算每个簇的质心
- 3. 直到质心不发生变化



## 2. 关联规则挖掘

- 目的：分析特征之间的关联关系。
- 应用场景：
  - 购物分析：用于促销、货架管理和存货管理
  - 气象预测：基于关联规则对灾害天气的预测
  - 医疗信息：发现与某种疾病关联的并发症
  - 推荐系统：找出商品之间的购买关系，从而进行商品推荐

年轻的父亲的购物篮子  
(啤酒-尿不湿案例)

TID	项集
1	{面包, 牛奶}
2	{面包, 尿布, 啤酒, 鸡蛋}
3	{牛奶, 尿布, 啤酒, 可乐}
4	{面包, 牛奶, 尿布, 啤酒}
5	{面包, 牛奶, 尿布, 可乐}

### 3. 自然语言处理和文本分析

- 从非结构化的文本数据中提取有用的信息和知识
- 主要问题：分词与词性标注、命名实体识别、句法分析、语义消歧、文本分类和聚类、和情感分析等
- 应用场景：
  1. 网络舆情分析。商品评论：好评/差评；投诉数据分析、法院判决文本
  2. 新闻分类、摘要。新闻类别分类：体育、社会、法制、经济等
  3. 机器翻译。不同语言之间的翻译：中英翻译等

只是我的手机屏幕上有一小块刮痕，不过平时也不太会注意到，就算了，懒得再申请换货了。

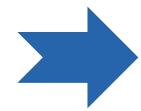
手机不错的，除了一个屎黄色的手机套我不喜欢，其他一切都很完美；用了几天，感觉挺好的

试了几天再来评价，，功能强大，不论是上网还是游戏运行速度都很快，屏幕也很清晰。手感很不错

超好用，N个贊！

手机挺好的，刚打开用了一天 没有啥毛病，然而我说了能不能给绿色的壳，客服说他们是随机发的，发一个一样颜色的壳这么难吗！！！ 顺丰4天，就这样了.....

说什么呢，高大上。颜值爆表，美的无法形容。使用了几天，很流畅，舒服，内存占用有点高，不过不影响，依然很快，无卡顿。非常满意！稍后上图



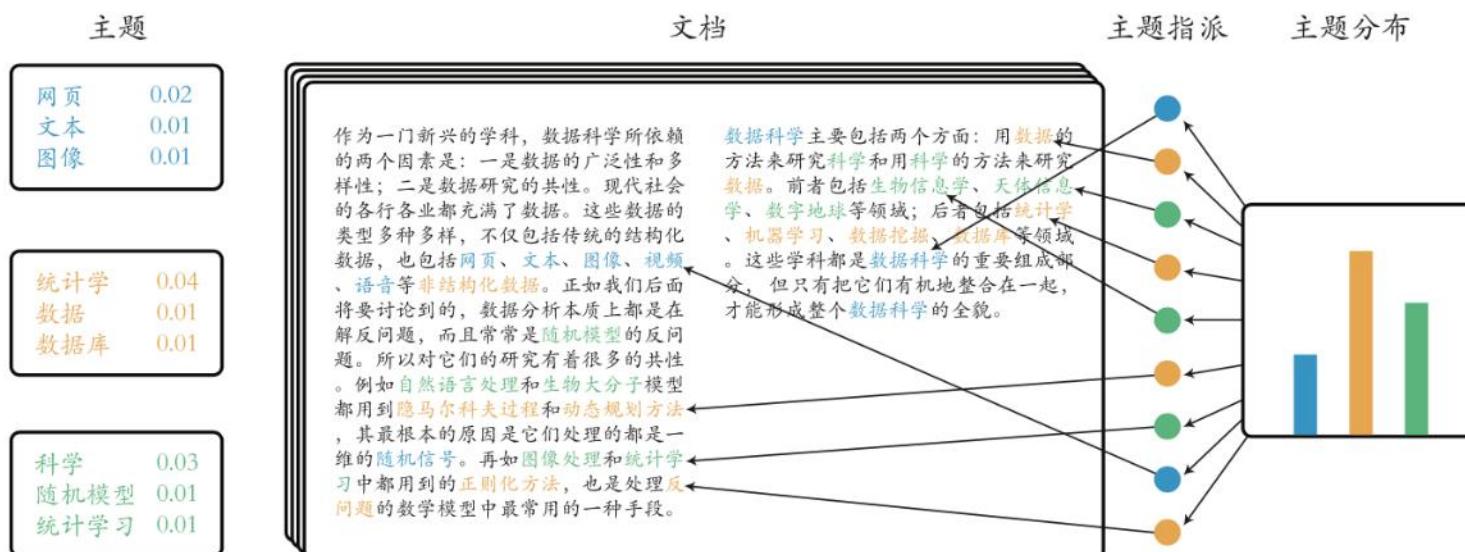
评价对象	对象描述词	观点表达式
速度	速度、反应速度	速度快、速度好、速度太慢、速度够快、速度惊人、速度良好、速度真快、
性能	系统、性能、功能	系统流畅、基本流畅、系统稳定、系统简洁、系统快、系统简单、性能优异、性能强大、性能强劲、功能强悍、机子流畅
物流	物流、物流业、发货	物流不错、物流快、物流蛮快、物流很棒、物流超快、物流慢、物流太慢、发货快、发货迅速、发货太慢
客服	客服、服务态度、态度	客服不错、客服小气、客服耐心、客服完美、服务态度差、服务态度很棒、服务态度蛮好、
颜色	颜色	颜色好、颜色漂亮、颜色不错、颜色较暗
屏幕	屏幕、画面、像素	屏幕碎、画面清晰、屏幕清晰、屏幕不灵、屏幕细腻、屏幕亮丽、屏幕小、屏幕很大、屏幕窄、像素好、
手感	手感、质感	手感不错、质感不错、手感好、手感细腻、手感极佳

## • 主题分析

- 挖掘海量文本集合中的主题
- 分析单个文本的主题分布
- 将文本从词典（十万级）降维到主题（几百）

## • 情感分析

- 政治选举
- 股票市场
- 舆情事件
- 电影票房
- 用户心情



- 主要问题

- 中心度、链接分析、社区发现、影响力分析等

- 应用场景：

- 1. 专家/网页/用户重要度评估

根据网络结构评估节点重要性，节点：专家、用户、网页等

- 2. 舆论领袖挖掘：根据信息传播网络，发现舆论领袖、关键人物

- 3. 欺诈团伙检测

电信欺诈、交易欺诈、信用卡申请欺诈

- 4. 供应链安全：根据税务数据、进出口数据，构造企业关系网络，分析供应链的脆弱性、依赖性等

绘制国家供应链网络，预警下一次系统性风险. <https://mp.weixin.qq.com/s/Qbc3CmfDfpQSCBXeTqQviQ>

# 5. 分布式计算

- 如何对大规模数据进行处理和分析
- **主要问题：**
  - 单机环境下大数据数据处理
  - 集群环境下的大规模数据处理
  - 大规模数据下的建模分析（分布式机器学习）
- **应用场景：**
  - 1. 大规模数据处理
    - 并行计算、Hadoop/MapReduce平台
  - 2. 大规模数据下的模型构建
    - 并行算法、硬件加速（GPU和深度学习）、Spark等分布式架构
  - 3. 算法的并行化、数据的并行化



- 数据科学概述

- 数据科学是如何兴起的
- 数据科学家应该具备什么样的能力
- 教学计划和考核要求

- 数据科学基本内容简介

- 机器学习
- 关联规则挖掘
- 自然语言处理
- 图和社交网络分析
- 分布式计算

课程QQ群（群号：547810062）

