数据科学导论

Introduction to Data Science

第一章 数据科学基础

陈恩红，黄振亚
Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2025.html

助教：于峻浩，程程
data_science_2025@163.com

9/22/2025

授课教师

课程负责人：陈恩红 教授

- 信息与智能学部副部长
- 认知智能全国重点实验室副主任
- 国家杰出青年基金获得者
- “大数据分析及应用”科技部重点领域创新团队负责人

课程团队成员：黄振亚 副教授

- 中国科大计算机学院
- 2021全球AI华人新星百强
- 中国科学青年人才托举工程

课程历史沿革

1998 首次开设面向研究生的数据挖掘课程，持续至今

2012 大数据时代的到来，“大数据驱动科学发现”

2013 邀请 AAAS/IEEE 会士熊辉教授、加拿大双院院士裴健教授讲授“龙星课程”

2014 在实验室开设面向本科生的数据挖掘与机器学习研讨班

2016 开始组建课程组 广泛收集课程资源

2017 首次开设本科生《数据科学导论》通识课

2024 信智学部大三大四本科生搬迁高新校区，课程属性修改为 计算机专业课

课程目标

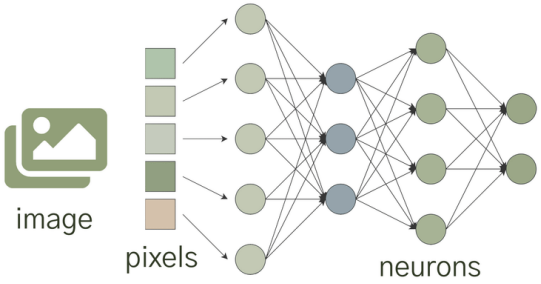
- 全面了解数据科学的基础知识
  - 包括数据分析的常用技术、发展前沿和应用案例
  - 了解数据的“能”与“不能”
- 树立数据科学的基本思路
- 初步掌握使用数据分析手段解决实际应用问题的能力

用科学的方法研究和应用数据

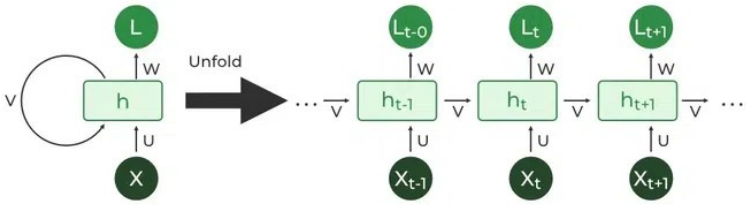
选修数据科学导论课程的同学将来可能从事不同领域的科学研究或者技术开发，希望这门课程带给你们的是终身受用的数据思维和创新能力。

数据科学基础

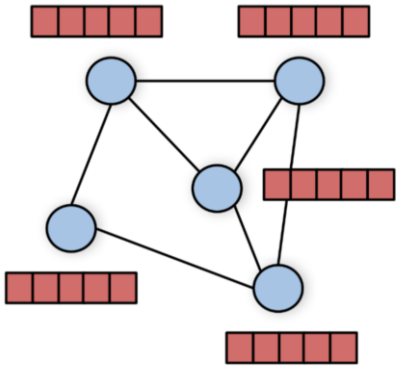
- 以模型为中心(model-centric)的数据科学技术
  - 围绕目标任务特性，设计不同模型结构



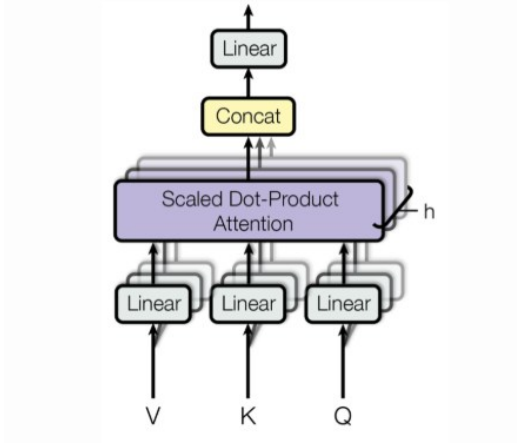
卷积神经网络CNN



循环神经网络RNN



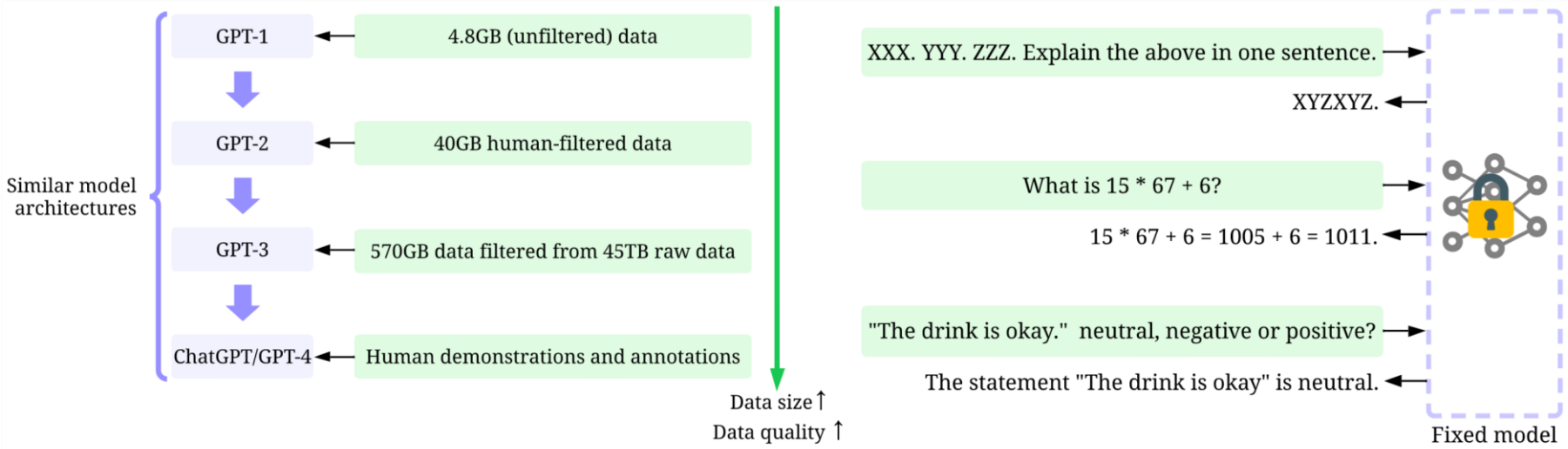
图神经网络GNN



自注意力神经网络Transformer

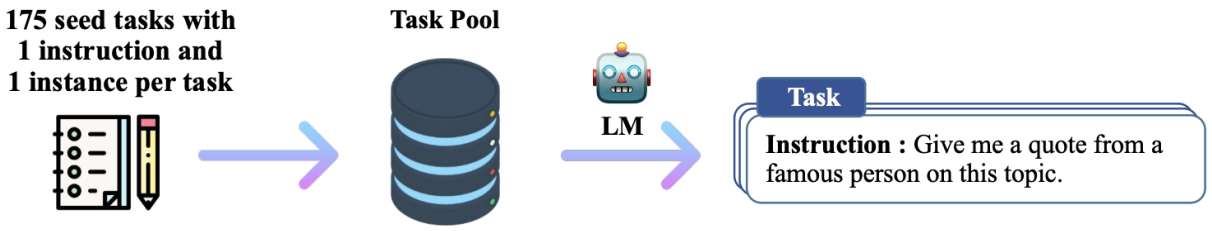
数据科学基础

- 人工智能逐渐从以模型为中心过渡到以数据为中心
  - **GPT成功的数据基石：**GPT进化中，模型结构保持相似，训练数据的规模、质量得到极大提升
  - **数据导向的模型应用：**当模型足够强大，仅仅需要修改推理数据（提示工程）便可完成目标任务



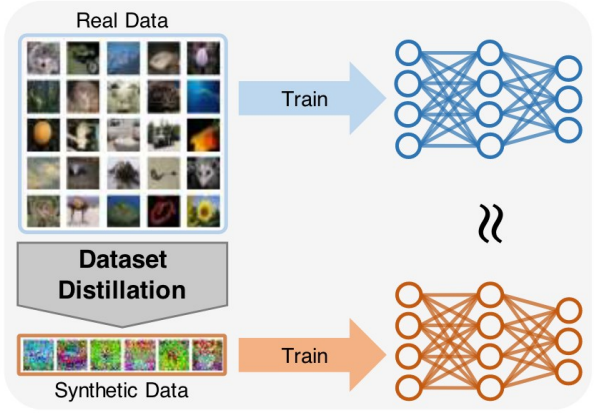
数据科学基础

- 以数据为中心(data-centric)的数据科学技术
  - 增加数据数量



数据生成

  - 改善数据质量

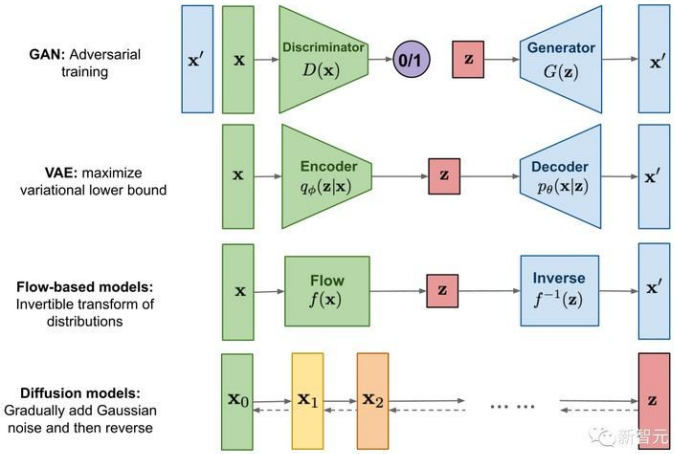


数据蒸馏

数据选择

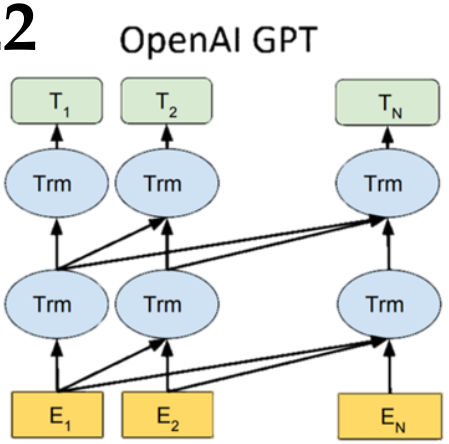
数据科学基础

- 大数据催生人工智能新浪潮—扩散模型-2022
  - 任务：AI图像生成
  - 应用数据集：LAION-5B
    - 80TB量级
    - 58.5亿个图像-文本对
  - 图像数据集规模变化：
    - Cifar-10：6万张
    - ImageNet：1400万张
    - LAION-5B：58.5亿张



数据科学基础

- 大数据催生人工智能新浪潮- ChatGPT-2022
  - 任务：文本对话
  - 数据量：5GB增加到45TB
    - 96%以上是英文，其它20个语种不到4%
  - 参数量：1.17亿增加到1750亿
  - 文本数据规模变化：



GPT

无监督预训练，有监督微调

5G文本数据 | 1.17亿模型参数

在9/12任务上最优，包括问答、语义相似度、文本分类

2018

GPT-2

多任务、零样本学习（zero-shot）

40G文本数据 | 15亿模型参数

在7/8任务上最优，包括阅读理解、翻译、问答

2019

GPT-3

小样本学习（few-shot）

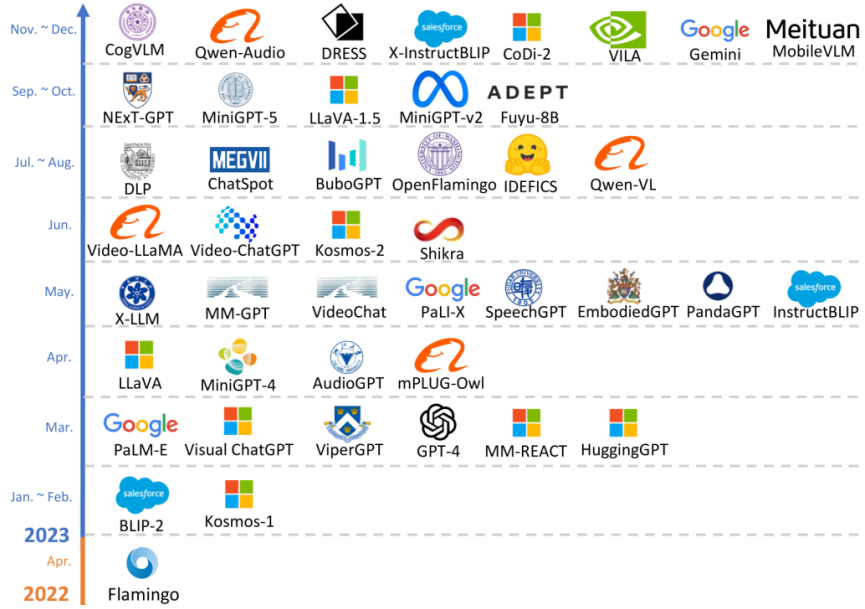
45T文本数据 | 1750亿模型参数

在阅读理解任务上超越当时所有zero-shot模型

2020

数据科学基础

- 大数据催生人工智能新浪潮- 多模态大模型（GPT-4(o)、Sora等等）-2023至今
  - 任务：多模态对话、多模态内容生成
  - 数据量：GPT-4：45TB文本数据增加到1PB多模态数据
  - 参数量：GPT-4：1750亿增加到1.76万亿参数



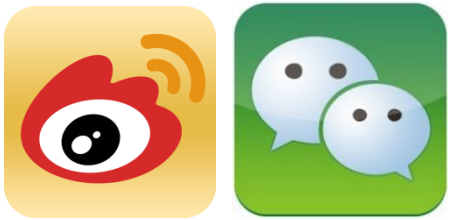
数据科学基础

- 人工智能的发展离不开大数据
  - 大数据是新时代的生产要素（十四五）
  - 我国已进入以大数据为核心资源的数字经济时代（二十大）

- **数据是什么？**
  - 从计算机科学的角度，所有能够输入到计算机并被计算机程序处理的符号的总称

文字数据

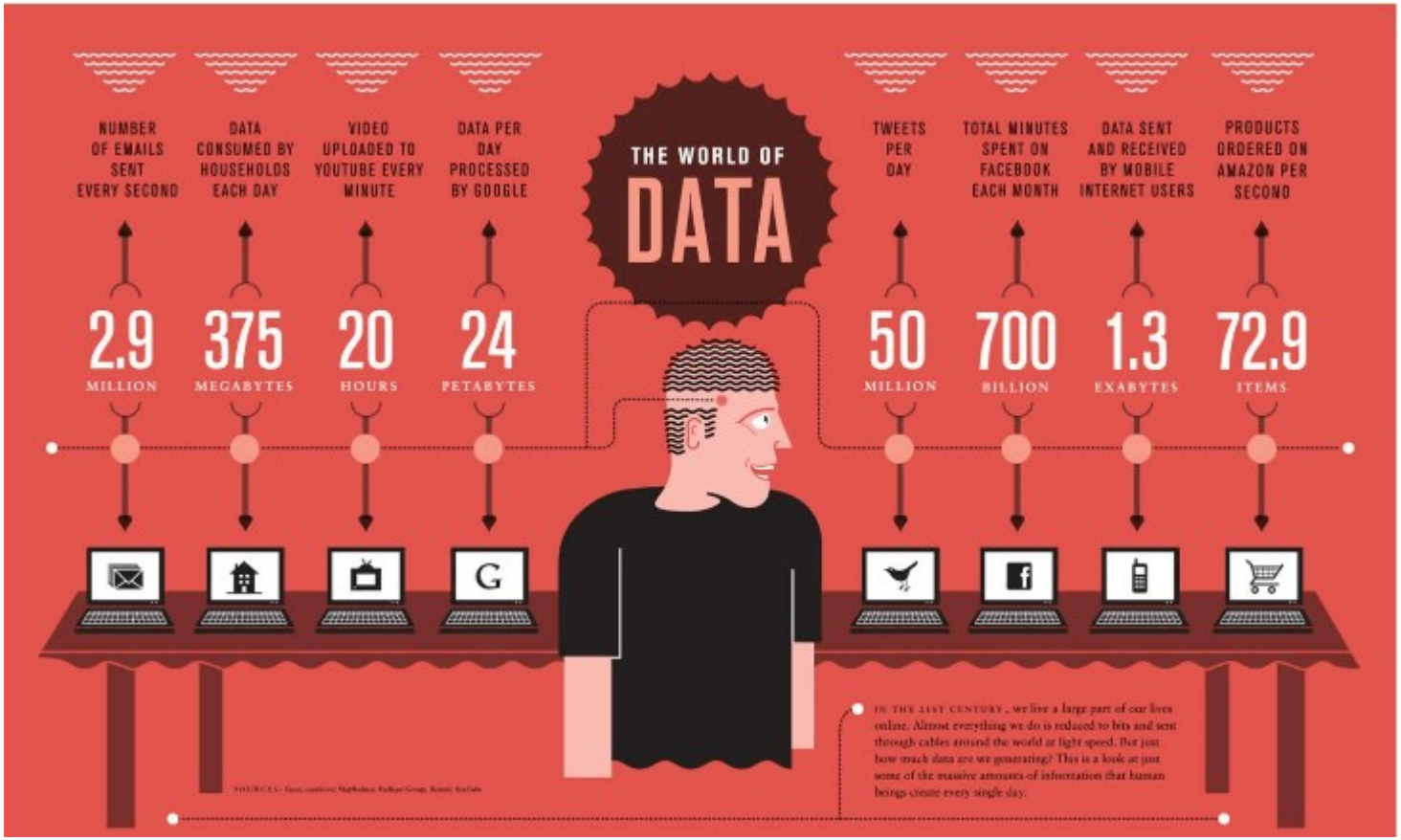
方位数据



沟通数据

数据科学基础

- **数据从何而来？**
  - 我们生活在数据中，所有人都在制造和分享数据



数据科学基础

- 大数据概念的提出

从2008年9月，《Nature》杂志首次出版一期大数据专刊，科学家们提出“大数据真正重要的是新用途和新见解，而非数据本身”

数据科学基础

- 郭华东院士：大数据是实实在在的一个方向、一门学科或者一项技术，大数据的出现才带动了数字经济的发展。

Q: 大数据与人工智能的关系是怎样的?

A: 大数据和AI是一对孪生兄弟

Q: 大数据是如何赋能数字经济发展的?

A: 数字经济所涵盖的就不仅仅是数据，大数据以及其他数字技术促进高质量发展

大数据、数字技术促进高质量发展，数字经济使人类走上高质量发展的轨道，这些都是目前人们正在实践的活动，且需要进一步往前发展的方向。大数据为实现高质量发展奠定了基础，数字经济的实践使人类高质量发展得到了更高程度的保证

数据科学基础

大数据有多大?

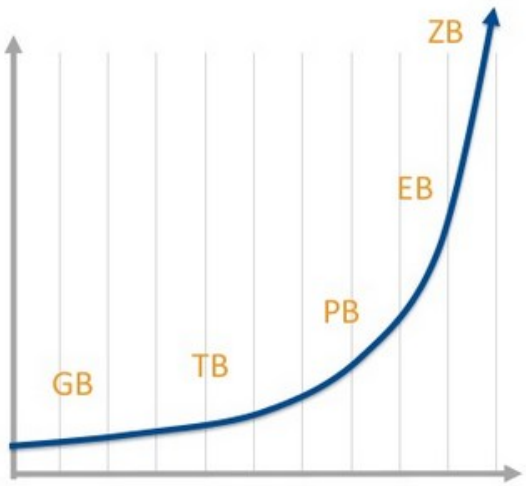
PB是大数据层次的临界点

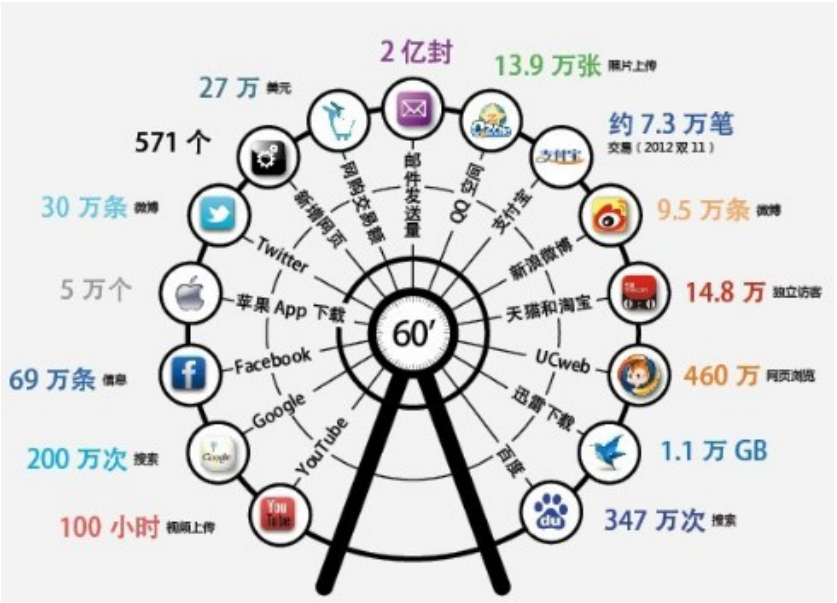
- 数据量已到ZB等级

KB->MB->GB->TB->PB->EB->ZB->YB->NB->DB

PB以上级别的数据，最有效的传输方式是空运，而不是网络

- 大数据不仅仅只是量大!

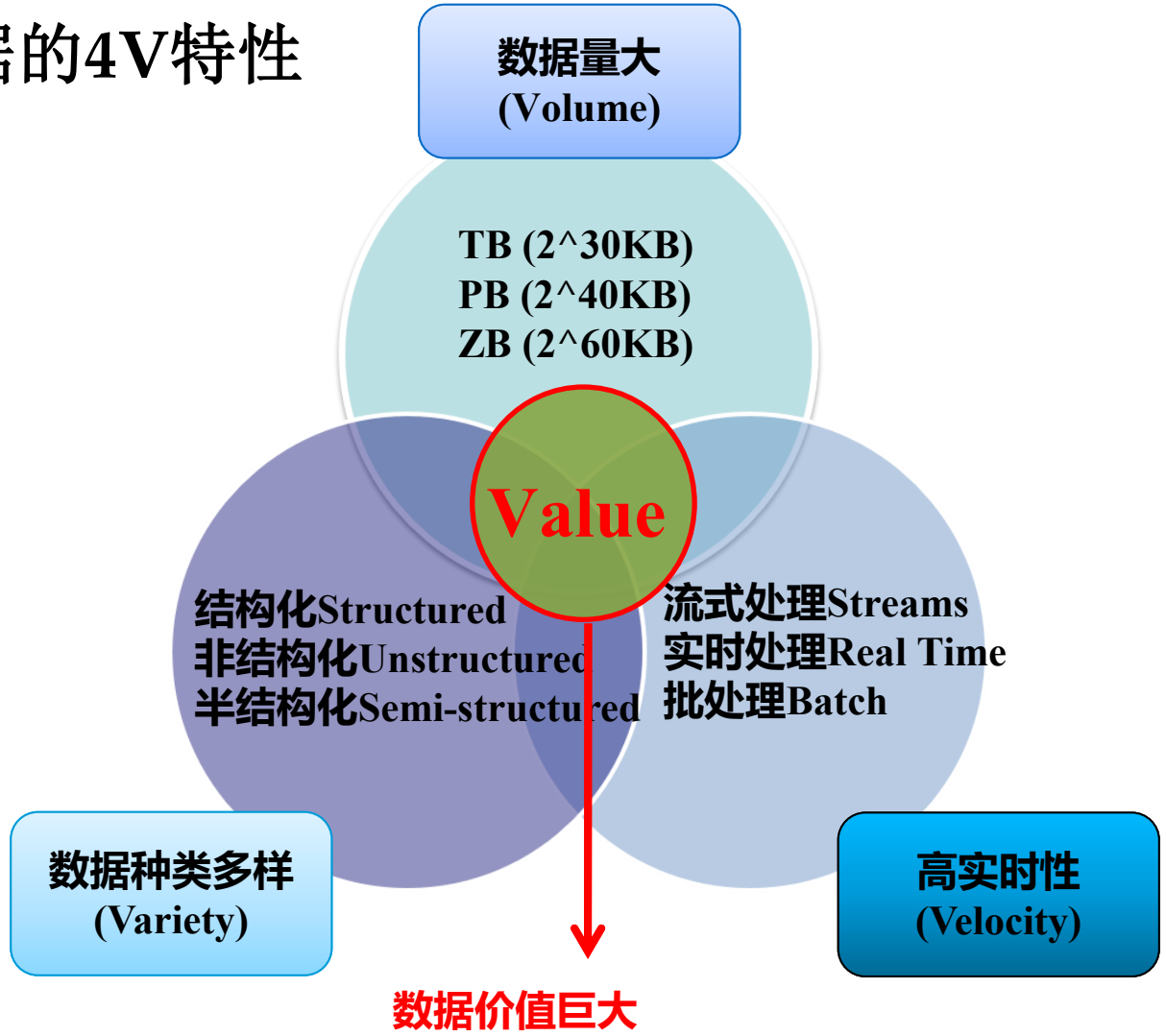




60秒，我们能产生多少数据?

数据科学基础

- 大数据的4V特性



大数据---Volume(数据量巨大)

阿里所保有的、经过清洗的历史数据已超过100PB。

——阿里数据仓库负责人七公（汪海）

百度现在的数据规模已经到了EB级，每天处理的数据量到了上百PB。

——百度大数据部总监薛正华

全球数据总量在2020年达到60ZB，2023年达到129ZB，预计2027年达到291ZB。

——IDC互联网数据中心

$$1\ ZB = 2^{10}\ EB = 2^{20}\ PB = 2^{30}\ TB = 2^{40}\ GB$$

- 1 ZB = 地球上沙粒的总量，1 EB = 4000个美国国会图书馆的藏书

数据科学基础

- 大数据--- Variety(数据类型多)

数据形式的多样:

- 结构化数据，半结构化数据，非结构化数据
- 关系数据库数据、xml/JASON文档、音视频数据

数据来源的多样性:

- 不同的IT应用系统
- 各种设备（手机、手环）
- 互联网、物联网
- 其它

时空数据

图像数据

文本数据

事务数据

视频数据

音频数据

- 大数据--- Velocity(高实时性)

1秒定律：对于大数据应用而言，必须要在1秒钟内形成答案，否则这些结果可能就是过时的、没有意义的

实时金融交易监控

实时欺诈检测

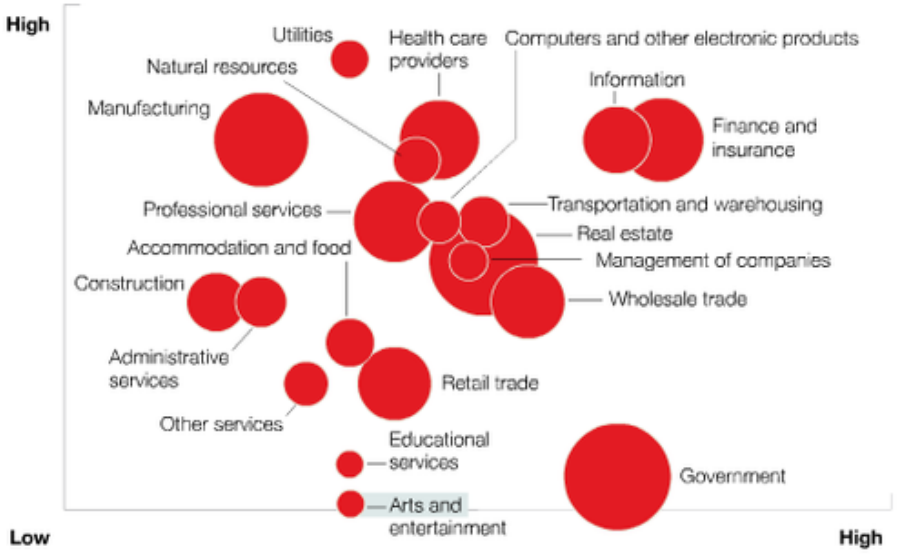
自动驾驶汽车

网络入侵检测

- **大数据--- Value(价值巨大但价值密度低)**

挖掘大数据中的价值类似**沙里淘金**，需要从海量数据中挖掘稀疏但珍贵的信息

所有产业都可以应用大数据产生价值



各产业GDP占比
(以美国经济为例)

图：麦肯锡对各个行业从大数据中获得价值难易程度的分析

数据科学基础

- 国际战略布局与定位

美国的大数据布局起步 – 上升为国家战略

- ✓ 2012年3月29日，美国联邦政府整合6个部门宣布2亿美元的“Big Data Research and Development Initiative”
- ✓ 目标：国家安全、新兴产业、科学发现与新型学科

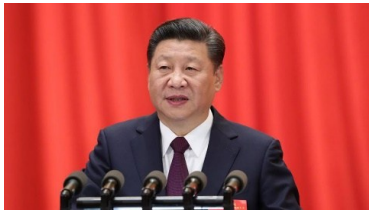
美国公布《数据战略实施计划》–突破核心科学与技术挑战

- ✓ 2022年，美国防信息系统局发布《数据战略实施计划》
- ✓ 目标：改善数据集成和利用、信息技术和网络能力，提高该机构“将数据用作战略资产”的能力。

欧盟的大数据规划 - 数据基础设施为先导

- ✓ 地平线(Horizon 2020) - The Framework Program for Research and Innovation
- ✓ GRDI 2020 - Global Research Data Infrastructures
- ✓ FP7 Call 8 Intelligent Information Management - Big Data

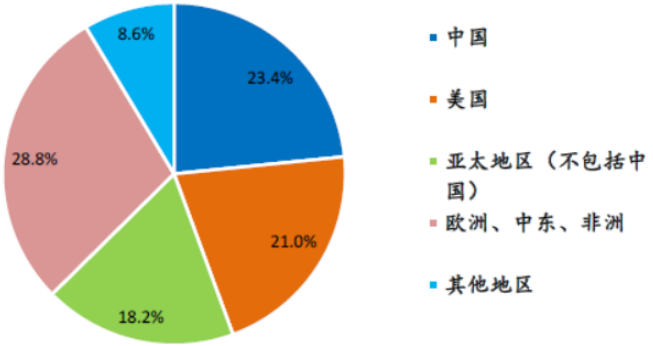
- **我国的大数据战略**
  - 2015年，十八届五中全会首次提出“**国家大数据战略**”，标志着大数据战略正式上升为国家战略。
  - 2018年，在**十三届全国人大一次会议**中，国务院总理李克强在作政府工作报告时，三次提到大数据
  - 2020年，《关于构建更加完善的要素市场化配置体制机制的意见》将大数据正式**列为新型生产要素**
  - 2021年，**《“十四五”大数据产业发展规划》：**明确了大数据发展的四大任务
  - 2024年，十七部门关于印发《**“数据要素×”**三年行动计划（2024—2026年）》的通知：构建**以数据为关键要素**的数字经济。



2022年，习近平总书记在《二十大报告》中强调：加快发展数字经济，促进数字经济和实体经济深度融合，打造具有国际竞争力的数字产业集群

数据科学基础

- **我国是数据产生和应用最大的国家**
  - 大数据是推动**数字经济**发展的关键生产要素
    - 2022年我国数字经济规模占GDP比重达到41.5%(50万亿元)
  - 大数据是重塑国家**竞争优势**的重大发展机遇
    - 2022年中国数据产量占全球10.5%(世界第二)；预计2025年成为最大数据圈
  - 大数据是实现**治理能力现代化**的重要创新工具
    - 2021年我国数字政府行业市场规模有望达到5000亿元
  - 大数据是建设**数字中国**的关键创新动力
    - 2021年全国工业互联网产业增加规模预计突破4万亿元



数据科学基础

十四五规划中的“大数据”

中共中央关于制定国民经济和社会发展第十四个五年规划的建议

（2020年10月29日中国共产党第十九届中央委员会第五次全体会议）

系统布局新型基础设施，加快第五代移动通信、工业互联网、大数据中心等建设。

推进土地、劳动力、资本、技术、数据等要素市场化改革。

21世纪的“数据”相当于　20世纪的“石油”国家基础战略资源

数据科学基础

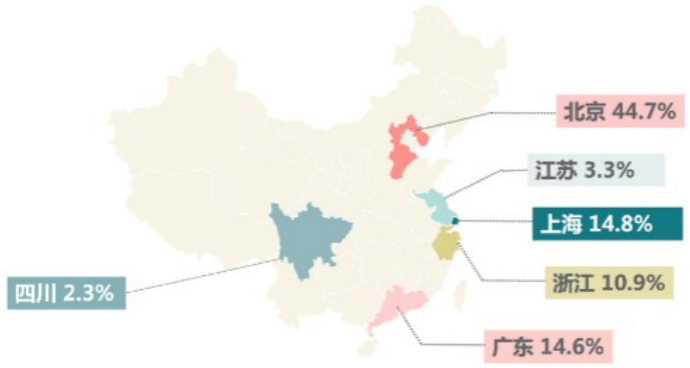
- 二十大中的“大数据”
  - 要加快发展数字经济，促进**数字经济**和实体经济深度融合，打造具有国际竞争力的**数字产业集群**。
  - **数字经济：**
    - **人工智能**
    - **大数据**
    - 电子信息
    - 5G

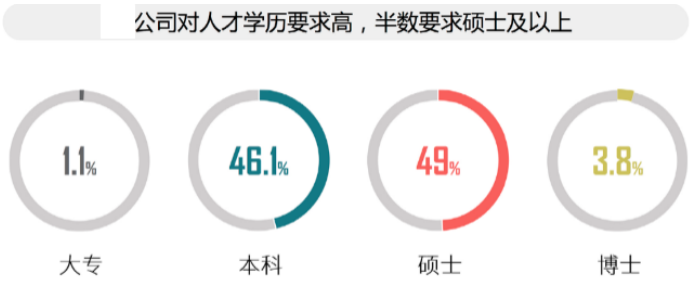
数据科学基础

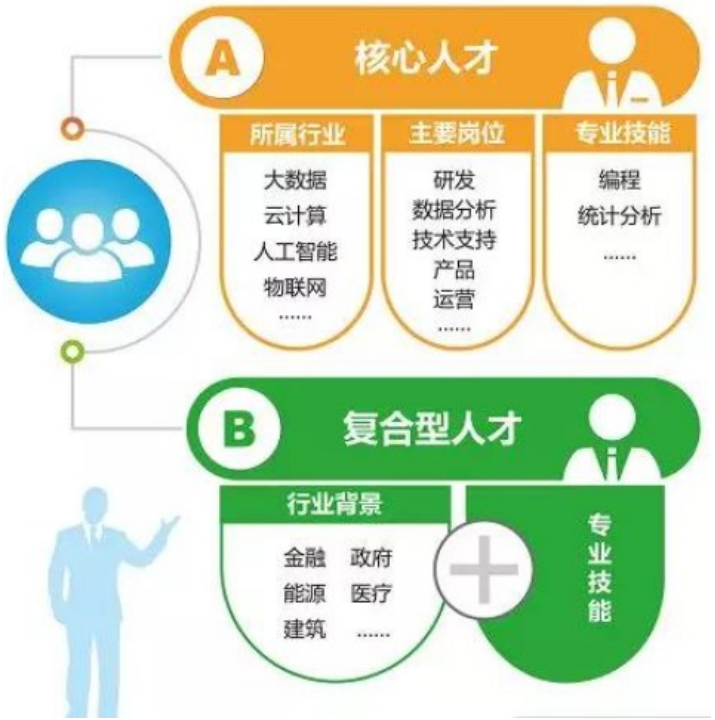
- 大数据人才缺口—国家需求
  - 2023年“两会”过后，国务院组建国家数据局(数据化国家队)
  - 急需具备大数据技能的新工科人才：理论基础+工程实践经验

数据科学基础

- 大数据人才缺口—市场需求
  - 市场对大数据人才的需求日益增加，供求关系不成正比，2025年人才缺口超过200万人
  - 产业发展对大数据人才提出更高要求







数据科学基础

- 大数据新工科人才需要具备以下素质

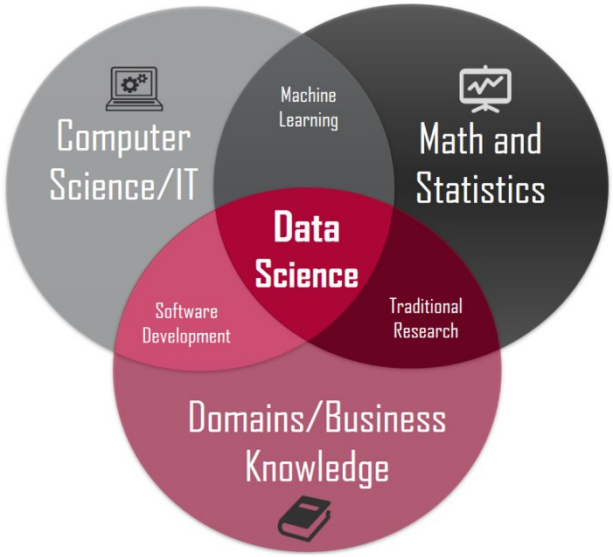
理论基础扎实，能理解运用数据科学中的理论模型

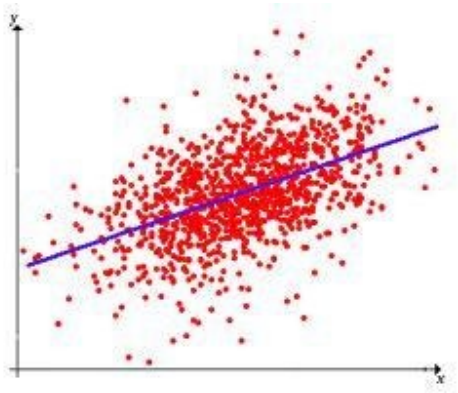
实践能力强，具有处理大数据的能力

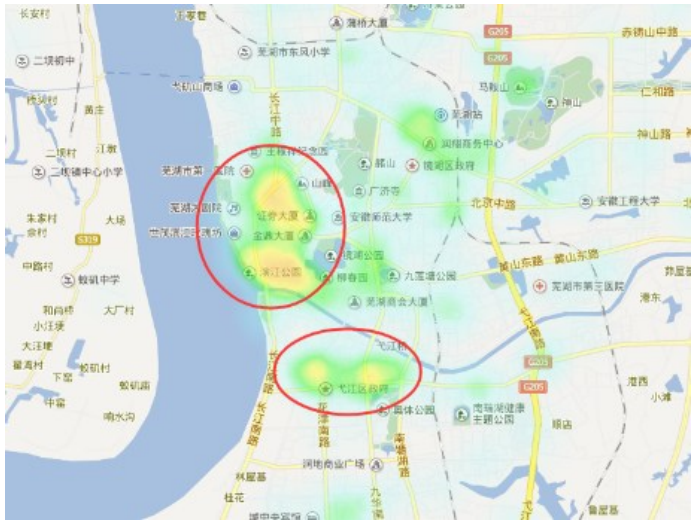
跨界能力强，能够解决特定行业的大数据应用问题

数据科学基础

- **大数据新工科人才需要具备以下素质**
  - 学习理论知识：数学（基础）+ 计算机科学 + 交叉学科知识
  - 锻炼实践能力：编程、数据分析、数据可视化等
  - 培养跨界能力：应用场景、领域知识







数据科学基础

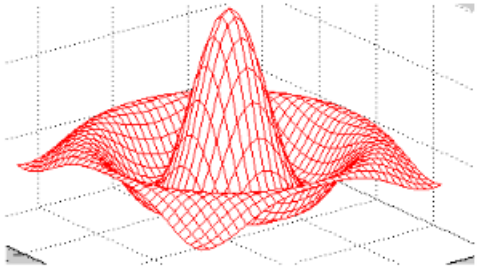
- **1. 理论基础扎实，能理解运用数据科学中的理论模型**
- 数学是学习数据科学的基础
  - 数学与优化：数学分析的应用
    - 梯度下降
    - 搜索方向：负梯度方向、牛顿方向
    - 算法收敛性
  - 数学与聚类：线性代数的应用
    - 社交网络聚类的问题形式化
    - 线性代数知识求解
  - 数学与图卷积网络：傅里叶变换的应用
    - 图表征学习
    - 图上的傅里叶变换与卷积
  - 。 。 。

数据科学基础

- 1. 理论基础扎实，能理解运用数据科学中的理论模型
  - 数学是学习数据科学的基础
    - 例如，数学与优化：数学分析的应用

模型学习（机器学习）

找到合适的 w ，使 f(w, x) 最接近 D



例如，线性回归损失函数

$$L(w)=\sum_{d_i \in D} f(w,x_i)-y_i$$

$$w=argmin_w L(w)$$

优化方法

常见问题： $\min_{x \in R^n} f(x)$

- 梯度下降
- 牛顿法/拟牛顿法
- 。。。

数据科学基础

- **1. 理论基础扎实，能理解运用数据科学中的理论模型**
  - 数学是学习数据科学的基础
    - 例如，数学与聚类：线性代数的应用

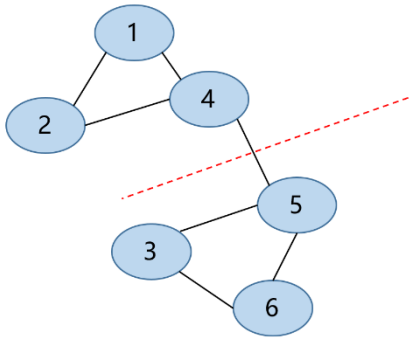
社交网络划分：物以类聚，人以群分

- ✓ 将全校学生划分为不同班级?
- ✓ 将员工划分为不同公司?
- ✓ 将用户划分为不同追星圈?

关键点：利用不同个体之间的联系

问题转化

无向图分割问题



$$W=\begin{pmatrix}0&1&0&1&0&0\\1&0&0&1&0&0\\0&0&0&0&1&1\\1&1&0&0&1&0\\0&0&1&1&0&1\\0&0&1&0&1&0\end{pmatrix}$$

$$D=\begin{pmatrix}2&0&0&0&0&0\\0&2&0&0&0&0\\0&0&2&0&0&0\\0&0&0&3&0&0\\0&0&0&0&3&0\\0&0&0&0&0&2\end{pmatrix}$$

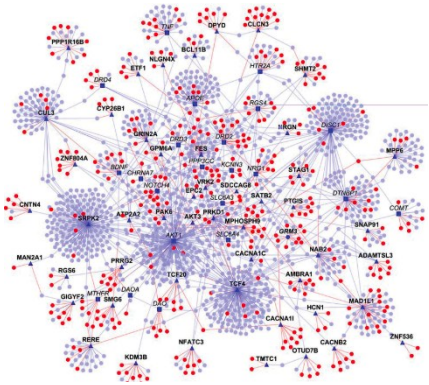
常用知识:

- ●特征值分解
- ●奇异值分解
- ●QR分解
- ●矩阵求逆相关定理

数据科学基础

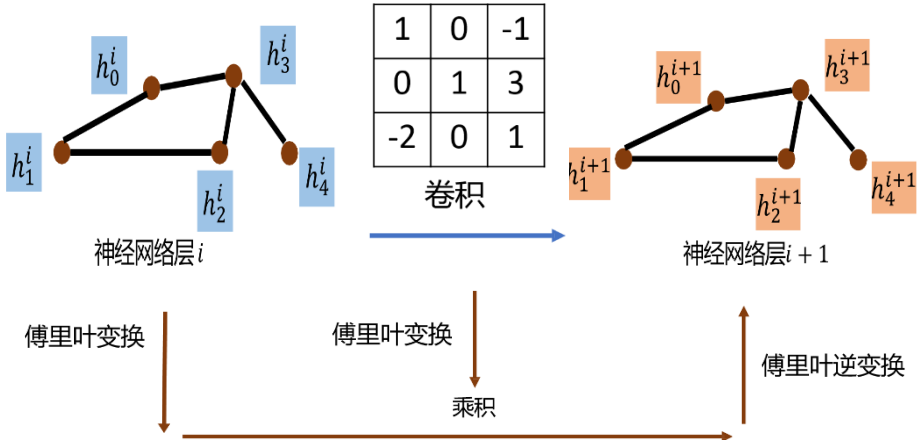
- **1. 理论基础扎实，能理解运用数据科学中的理论模型**
  - **数学是学习数据科学的基础**
    - **例如，数学与图卷积网络：傅里叶变换的应用**

图数据：分子图、社交网络等



- **典型任务**
  - 节点分类，关系（边）预测等
  - 图分类，图属性预测，图生成

图卷积网络：一类典型方法



- **Idea：卷积定理：函数卷积的傅里叶变换是函数傅立叶变换的乘积**
- **一般傅里叶变换　至　图上傅里叶变换**

数据科学基础

- **2. 实践能力强，具有处理大数据的能力**
  - Python等编程技术，Web技术、数据库技术、可视化技术等
  - 常用工具使用：如**大模型工具**

```
def SumOfKsubArray(arr, n, k): SumOfKsubArray(arr, n, k):
  Sum = 0Sum = 0
  S = deque()= deque()
  G = deque()= deque()
  for i in range(k):for i in range(k):
    while (len(S) > 0 and arr[S[-1]] >= arr[i]):while (len(S) > 0 and arr[S[-1]] >= arr[i]):
      S.pop().pop()
    while (len(G) > 0 and arr[G[-1]] <= arr[i]):while (len(G) > 0 and arr[G[-1]] <= arr[i]):
      G.pop().pop()
    G.append(i).append(i)
    S.append(i).append(i)
  for i in range(k, n):for i in range(k, n):
    Sum += arr[S[0]] + arr[G[0]]Sum += arr[S[0]] + arr[G[0]]
    while (len(S) > 0 and S[0] <= i - k):while (len(S) > 0 and S[0] <= i - k):
      S.popleft().popleft()
    while (len(G) > 0 and G[0] <= i - k):while (len(G) > 0 and G[0] <= i - k):
      G.popleft().popleft()
    while (len(S) > 0 and arr[S[-1]] >= arr[i]):while (len(S) > 0 and arr[S[-1]] >= arr[i]):
      S.pop().pop()
    while (len(G) > 0 and arr[G[-1]] <= arr[i]):while (len(G) > 0 and arr[G[-1]] <= arr[i]):
      G.pop().pop()
    G.append(i).append(i)
    S.append(i).append(i)
  Sum += arr[S[0]] + arr[G[0]]Sum += arr[S[0]] + arr[G[0]]
  return Sumreturn Sum
```

数据科学基础

- 3. 跨界能力强，能够解决特定行业的大数据应用问题

改变这个世界的第四种力量—大数据的应用

暴力

世界著名未来学家托夫勒
《第三次浪潮》作者

知识

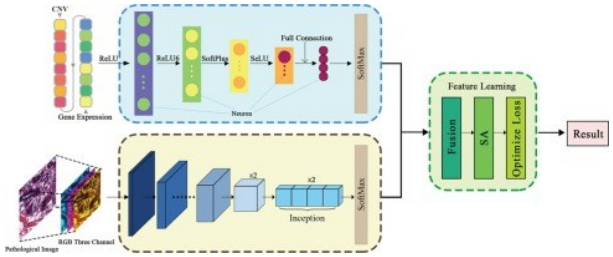
金钱

大数据

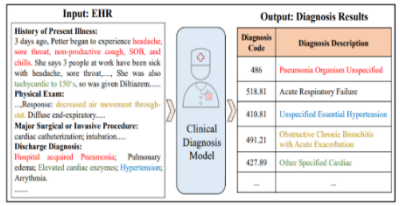
数据科学基础

- 数据蕴含着巨大的价值—智慧医疗
  - 通过对患者建立AI电子病历
  - 整合患者的全时段、多模态的健康数据（病例文本、检查影像等）
  - 实现对患者的疾病诊断、病灶识别、药物推荐等

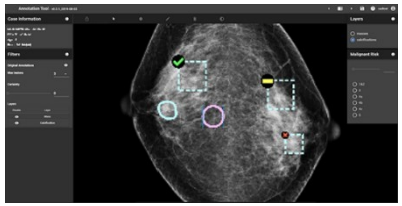
AI电子病历



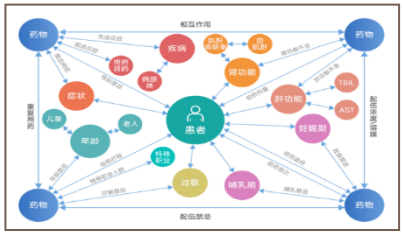
多模态医疗数据挖掘模型



疾病诊断



病灶识别



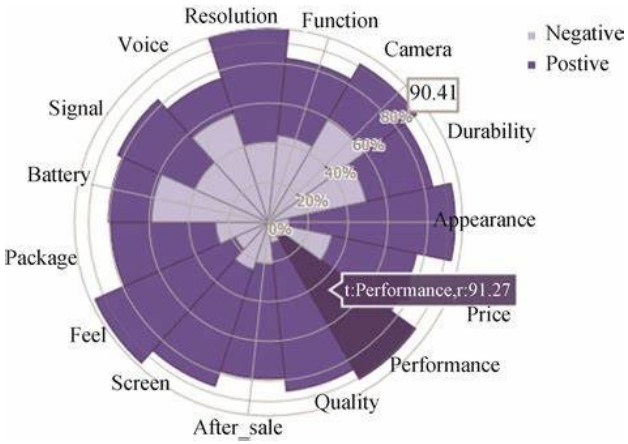
药物推荐

- 数据蕴含着巨大的价值—安防领域
  - 公安监控智能分析

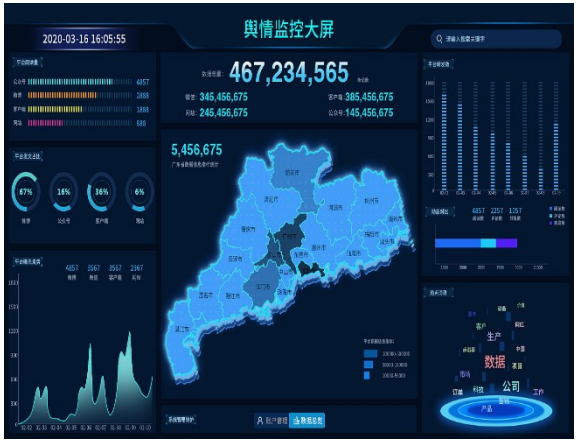
区间超速判定

“天眼”追凶

- 数据蕴含着巨大的价值—安防领域
  - 舆情监测
    - 通过对新闻网站、论坛、博客等的文章和评论进行文本挖掘和情感分析，安防系统能够实时捕捉公众对特定话题的反应。
    - 通过大数据分析识别并跟踪网络上虚假信息的传播路径，并帮助相关机构及时干预，防止谣言扩散引发社会恐慌或动乱。



舆情情感分析



传播途径监测