



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

1

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/DS2026.html>

助教：罗胤玻，李宁

ds_intro2026@163.com



课程要求与考核方式

3

- 课程目标：用科学的方法研究和应用数据
- 课程要求，具体要求之后布置
 - 课堂出勤与作业
 - 文献调研报告
 - 实验与报告（需要编程）
 - 课程交流与课程汇报
- 考核方式
 - 课堂（30%）+调研（30%）+实验（40%）



课程说明

4

- 《数据科学导论》，CS4023
- 上课时间：1-18周，周二下午8、9节：15:35 - 17:30
- 上课地点：高新校区 GT-B212
- 同类型课程：学分不可重复认定
 - 人工智能与数据科学学院《数据分析与实践》DS3001
 - 人文学院《新媒体大数据分析》NNM2012
- 几点说明
 - 课堂与平时重要，大家保证出勤
 - 东区西区的同学需跨校区上课，信智学部可考虑大三大四再选
 - 非信智学部的同学，可考虑《新媒体大数据分析》
 - 不同专业中的学分认定，需跟相关专业教秘确认
 - 需要做实验，只想听课堂课程的同学谨慎选择



数据科学基础

5

□ “大数据”相关专业是当前的热门专业

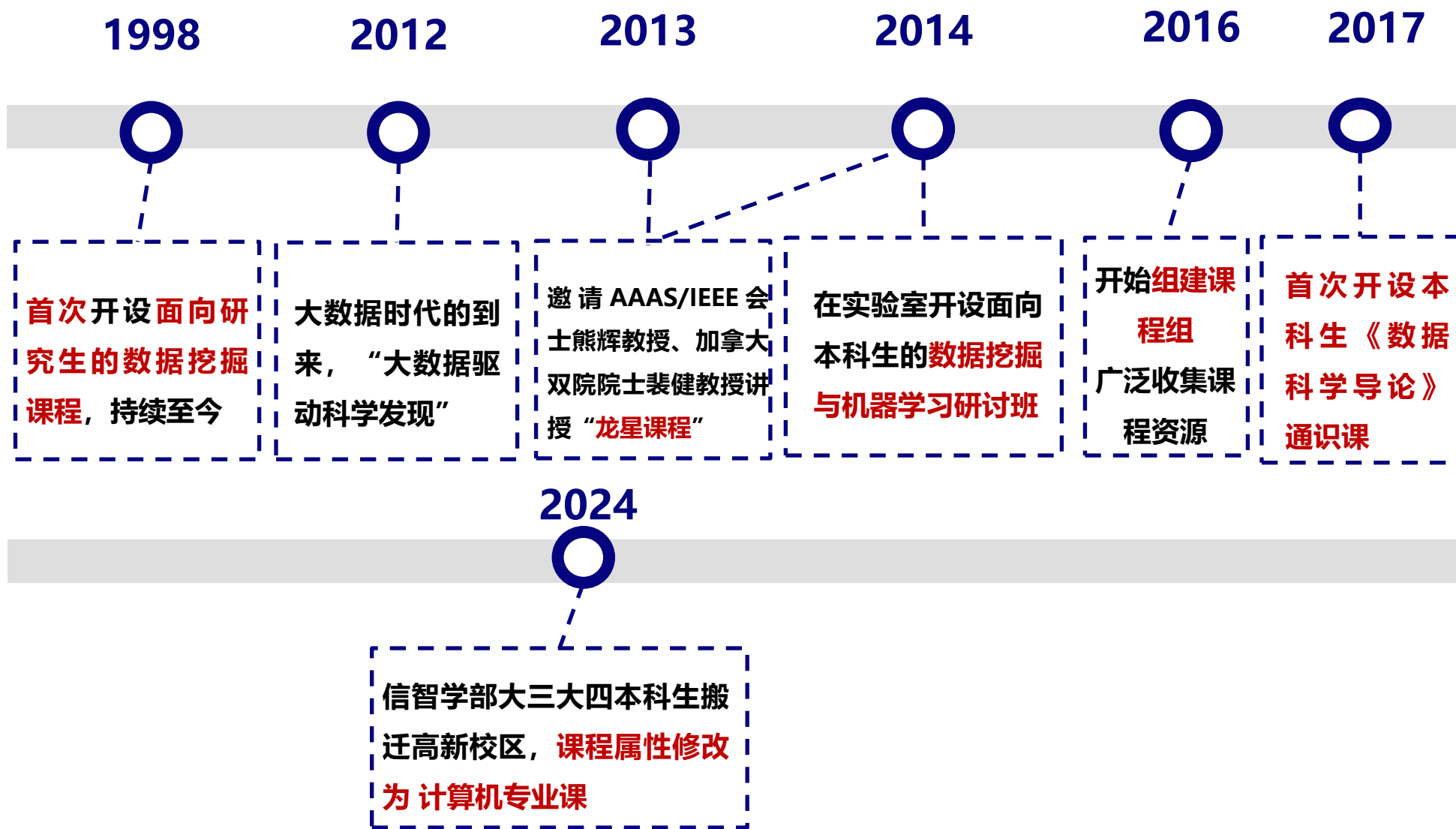
- 教育部《普通高等学校本科专业备案和审批结果》显示，**大数据技术相关专业**是近年中高校新增数量最多的专业之一
 - **数据科学与大数据技术**自2017年来有**700+**所高校新增该专业
 - **人工智能**自2018年来有**700+**所高校新增该专业
- 高校与大数据企业对大数据**高端人才**和**复合人才**需求旺盛
- 大数据相关专业将朝着**精细化、融合化**的方向发展

新增备案本科专业	2017	2018	2019	2020	2021	2022	2023	2024
人工智能	-	35	180	130	95	59	38	
数据科学/大数据技术	250	203	137	62	40	30	31	
大数据管理与应用	5	25	51	59	42	38	32	



课程历史沿革

6





联系方式

7

- 授课教师：陈恩红，黄振亚
- 课程主页：<http://staff.ustc.edu.cn/~huangzhy/Course/DS2026.html>
- QQ群：1050506077
- 联系方式
 - 助教：罗胤玻，李宁
 - ds_intro2026@163.com
 - 教师：黄振亚，陈恩红
 - huangzhy@ustc.edu.cn





本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

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/DS2026.html>

助教：罗胤玻，李宁

ds_intro2026@163.com



课程目标

9

- 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

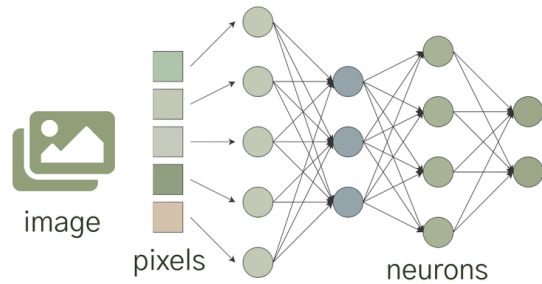
用科学的方法研究和应用数据

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，

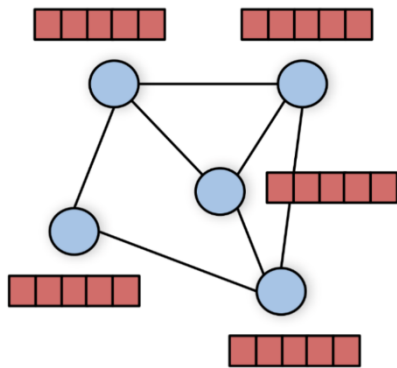
希望这门课程带给你们的是终身受用的数据思维和创新能力。



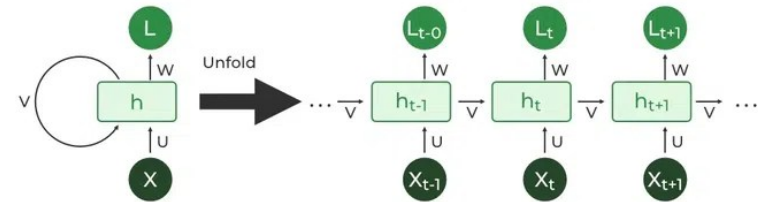
- 以模型为中心(model-centric)的数据科学技术
 - 围绕目标任务特性, 设计不同模型结构



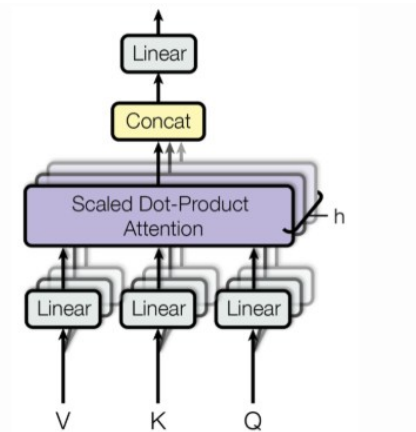
卷积神经网络CNN



图神经网络GNN



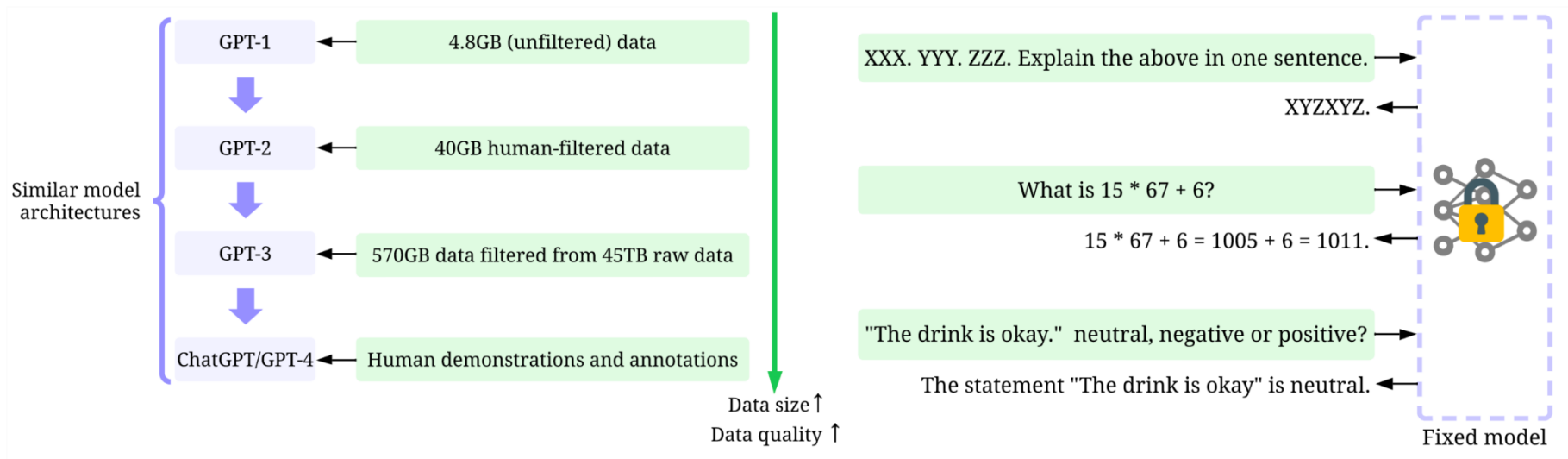
循环神经网络RNN



自注意力神经网络Transformer



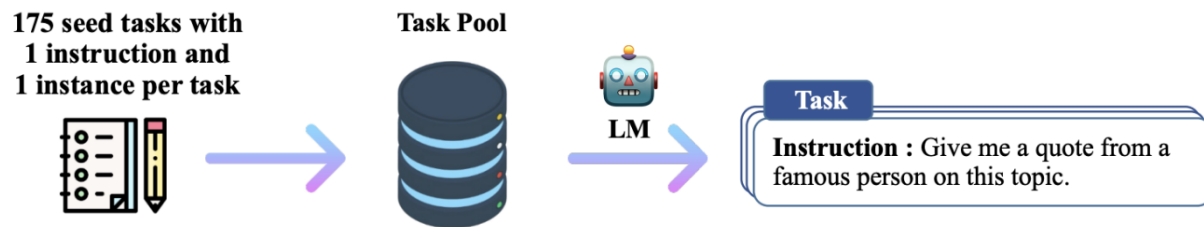
- 人工智能逐渐从以模型为中心过渡到以数据为中心
 - GPT成功的数据基石：** GPT进化中，模型结构保持相似，训练数据的规模、质量得到极大提升
 - 数据导向的模型应用：** 当模型足够强大，仅仅需要修改推理数据（提示工程）便可完成目标任务





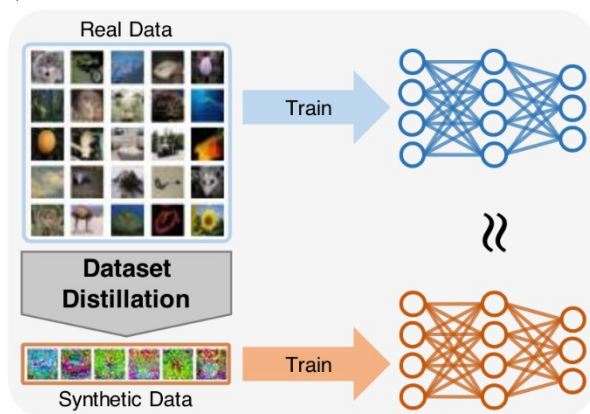
□ 以数据为中心(data-centric)的数据科学技术

□ 增加数据数量



数据生成

□ 改善数据质量



数据蒸馏



数据选择

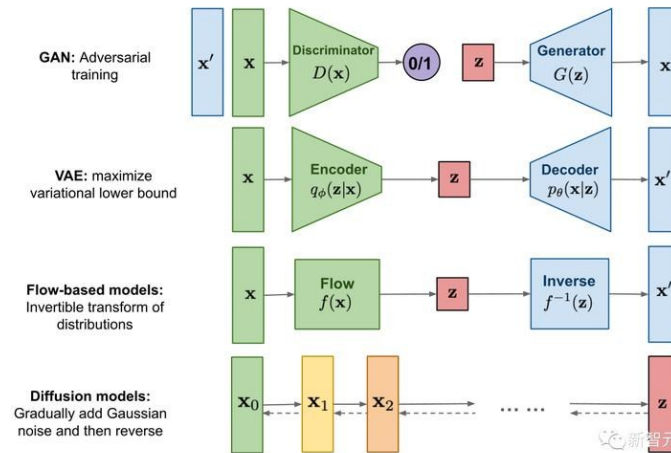
□ 大数据催生人工智能新浪潮—扩散模型-2022

- 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：

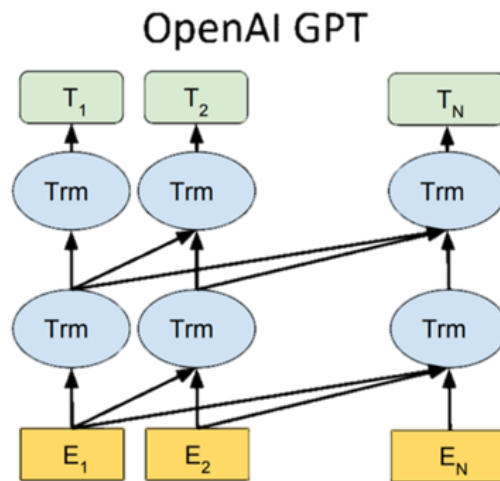
- Cifar-10: 6万张
- ImageNet: 1400万张
- LAION-5B: 58.5亿张





□ 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)

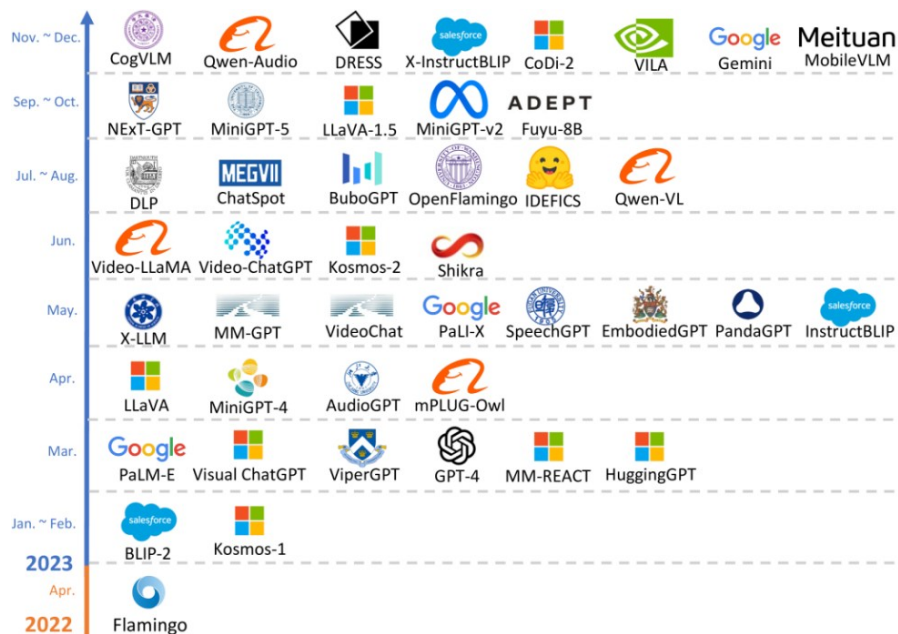
45T文本数据 | 1750亿模型参数

在阅读理解任务上超越当时所有zero-shot模型

2020



- 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？

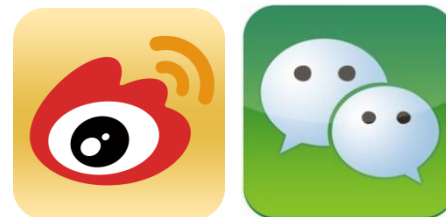
- 从计算机科学的角度，所有能够输入到计算机并被计算机程序处理的符号的总称
- 数据是人工智能技术与产业发展的生产要素
- 我国已经进入了以大数据为核心资源的数字经济时代



文字数据



方位数据



沟通数据



数据科学基础

17

□ 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



150亿个设备连接到互联网



全球每秒钟发送
290万封email



每天有 **2.88 万**小时
视频上传Youtube



全球数据量发展

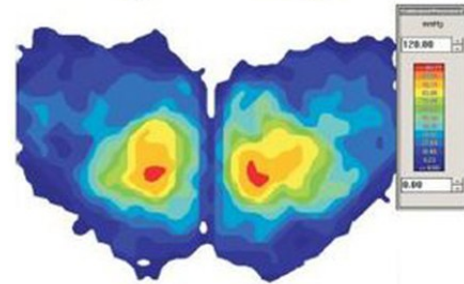
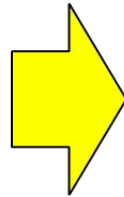
信息爆炸



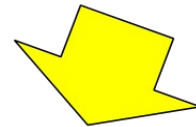
□ 例子：最不可能的地方获得数据

□ 当一个人坐着的时候，他的身形、姿势和重量分布都可以量化和数据化。

在汽车座椅下部安装
360个压力传感器



测量人对椅子施加的压力，
用0~256的数值量化



- 把人体屁股特征转化成了数据，产生独属于每个乘坐者的精确数据资料。
- 汽车可以准确的识别乘坐者的身份：汽车防盗系统



数据科学基础

19

□ 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

20

- 李德毅院士：大数据本身，既不是科学也不是技术，它反映的是网络时代的一种客观存在

你们说的大数据到底是啥？**大数据的输入和输出是？**

我不认为数据等同于价值，**哪些数据才有价值？**

大数据到底是**噱头+忽悠**，还是**真金白银**啊？

我没看清楚大数据的价值，但很清楚**大数据的大成本**，真能赚回来吗？

未来真的**不会大数据就不能赢了吗？**

我用SQL Server用的好好的，一定要**现在就转大数据吗？**

所谓的大数据牛的公司，**到底牛在哪？**



大数据就是数据，没什么可神秘的。它是一种原材料，数据库、数据挖掘、云计算、高性能计算、机器学习等都可以看作是对这种原材料进行存储烹饪加工等的手段和技术，目的就是做出各种美食（例如，AlphaGo打败李世石）



数据科学基础

21

- 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程度的保证





大数据有多大？

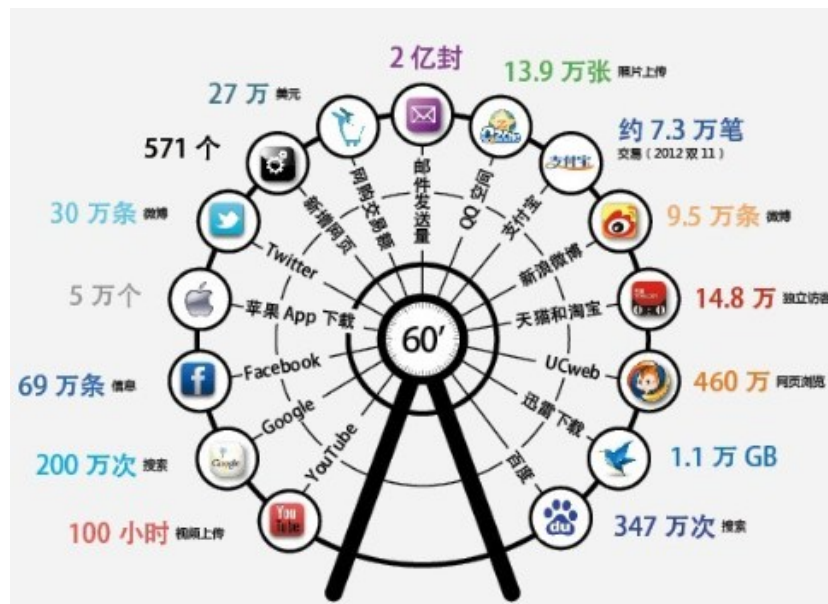
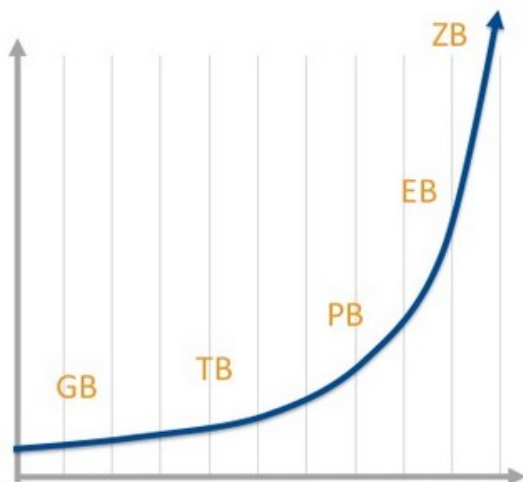
◆ 数据量已到ZB等级

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

PB是大数据层次的临界点

PB以上级别的数据，最有效的传输方式是空运，而不是网络

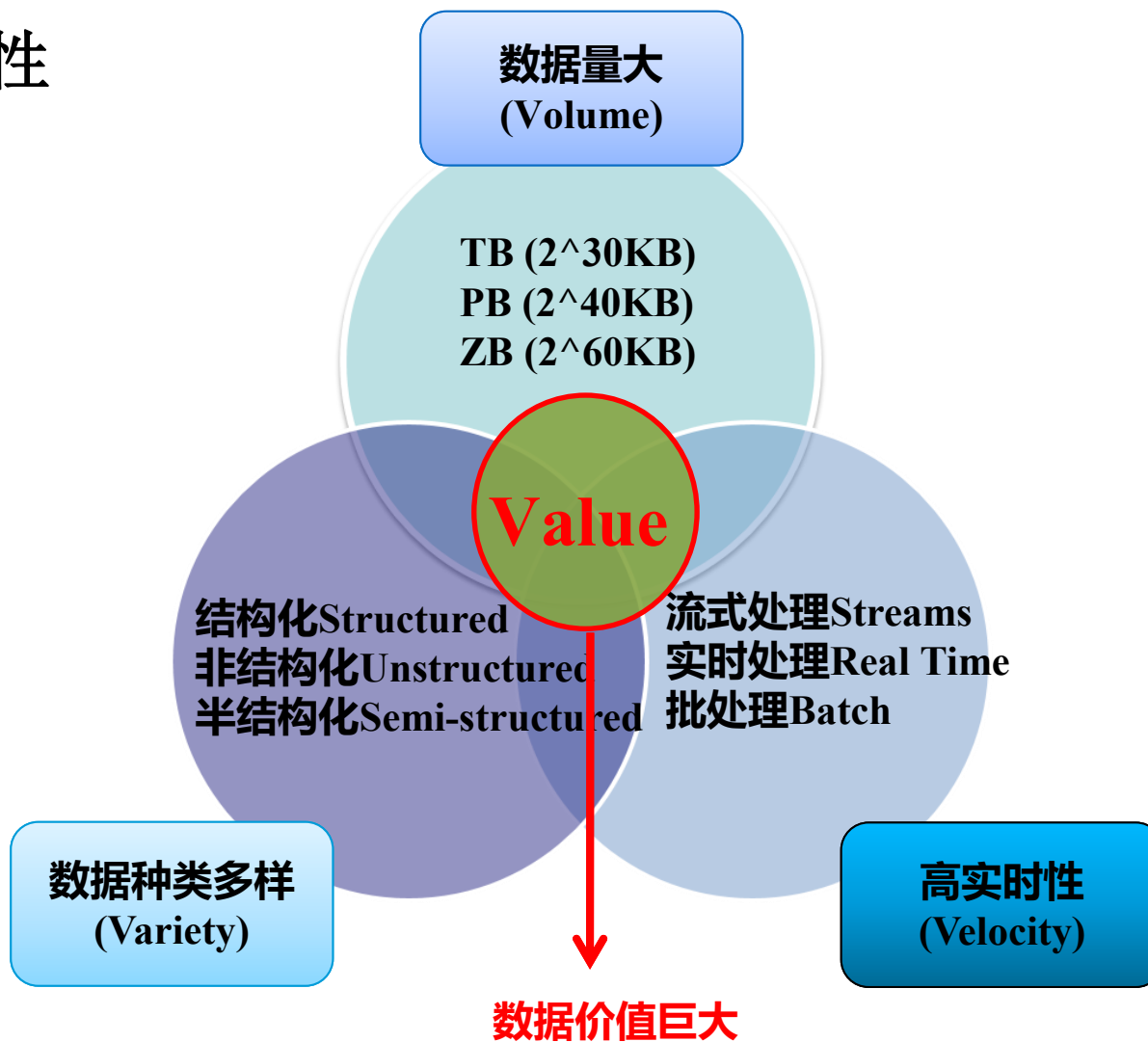
◆ 大数据不仅仅只是量大！



60秒，我们能产生多少数据？



□ 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 100 PB。
- **394 ZB**: 全球数据量预计从 2024 年的 149 ZB 增至 2028 年的 394 ZB 以上。
- 2026—2032 年: 按当前趋势, 公开的人类生成文本可能在这一时期被大模型训练充分利用。

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

- 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

25

□ 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



文本数据



图像数据



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



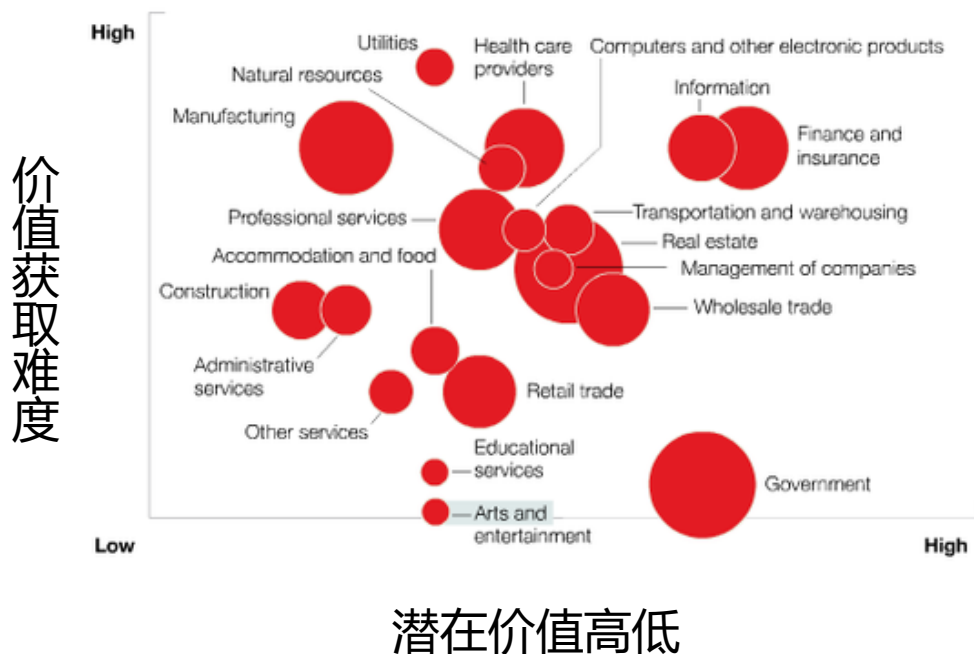
网络入侵检测



□ 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



□ 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。



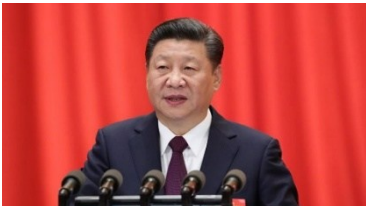
美国人工智能战略 - 数据基础设施为先导

- ✓ 2025.07：《美国人工智能行动计划》提出 90 余项联邦行动，聚焦“加速创新、建设人工智能基础设施、国际领导力”。
- ✓ 2026.02：启动 AI Agent 标准计划；通过行政令和国家安全备忘录强化智能体互操作、关键基础设施安全和国家安全领域应用。



□ 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。
- 2026年，“**十五五**”规划设立数字中国专篇，提出强化算力、算法、数据供给，深化“人工智能+”。



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

30

□ 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
 - 2024年我国数字经济规模达到**59.2万亿元**，占GDP比重**43.8%**。
 - 2025年我国数字经济核心产业**增加值**占GDP比重达到**10.5%以上**
- 大数据是重塑国家**竞争优势**的重大发展机遇
 - 2025年我国数据生产总量达**52.26ZB**，占全球**27.44%**
- 大数据是实现**治理能力现代化**的重要创新工具
 - 2025年政务数据共享枢纽平台累计支撑调用超过**5400亿次**
- 大数据是建设**数字中国**的关键创新动力
 - 2025年我国人工智能核心产业规模超过**1.2万亿元**





数据科学基础

31

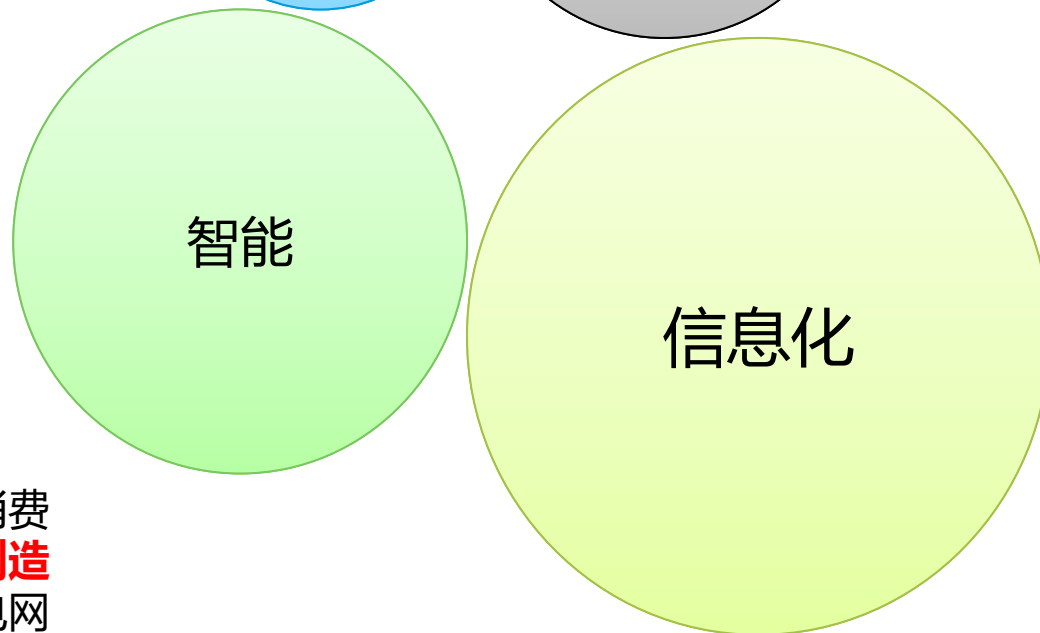
□ 十三五规划建议中的“大数据”

关键词

- **国家大数据战略**
- 大数据技术



- **互联网+**
- 互联网金融
- 下一代互联网



- 智能消费
- **智能制造**
- 智能电网

- 社会信息化
- 流通信息化
- 农业信息化
- 教育信息化
- 信息化战争
- 社会治理信息化
- 信息基础设施
- 信息技术



□ 十四五规划中的“大数据”

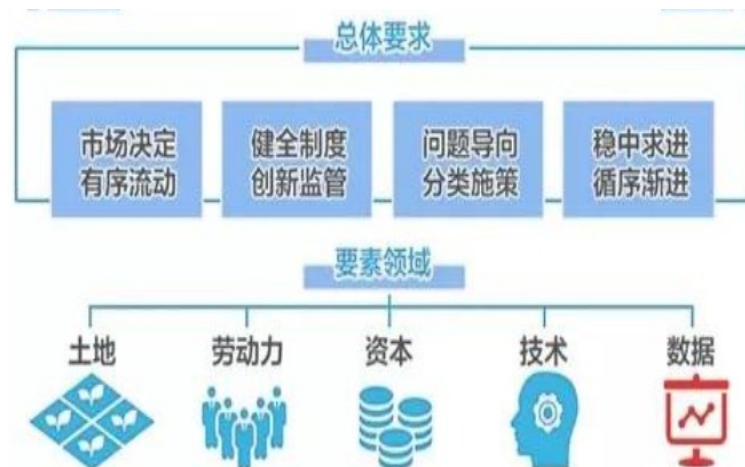
中共中央关于制定国民经济和社会发展第十四个五年规划的建议
(2020年10月29日中国共产党第十九届中央委员会第五次全体会议)

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。



数据是新时代生产要素

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



- “十五五”：建设现代化基础设施“六张网”
 - 水网：保障水资源配置与防洪安全
 - 新型电网：促进绿色能源生产、调度和消纳
 - 算力网：连接数据中心、超算中心和智算中心，统一调度计算资源
 - 新一代通信网：通过5G-A、6G、光纤和卫星互联网传输数据
 - 城市地下管网：支撑供水、排水、燃气等城市生命线
 - 物流网：连接生产、流通和消费



截至2026年7月底，全国智能算力规模达到245万PFL0PS，八大国家算力枢纽和三个算电协同发展区占全国智算规模的85%以上。



数据科学基础

34

□ 大数据人才缺口—国家需求

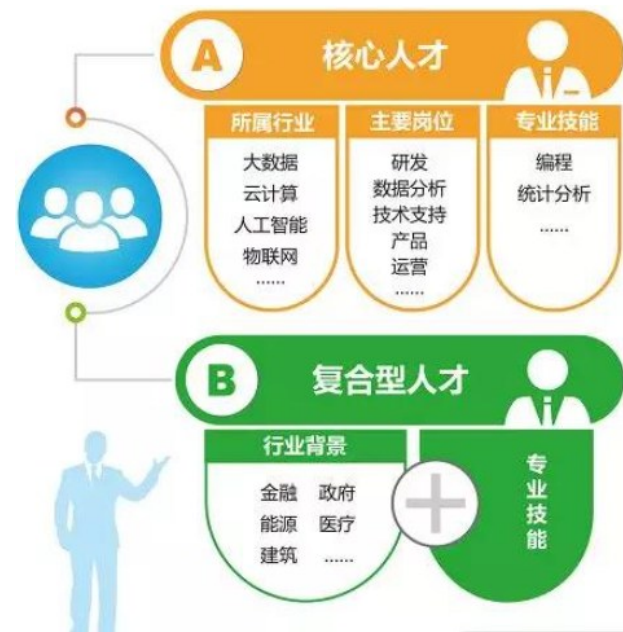
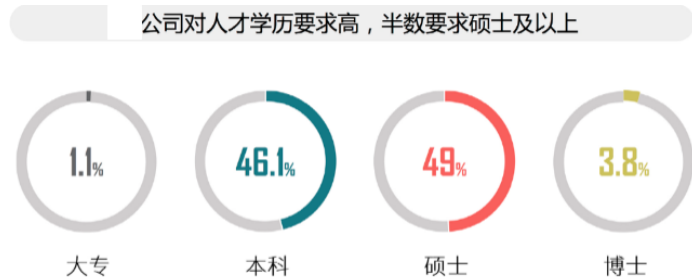
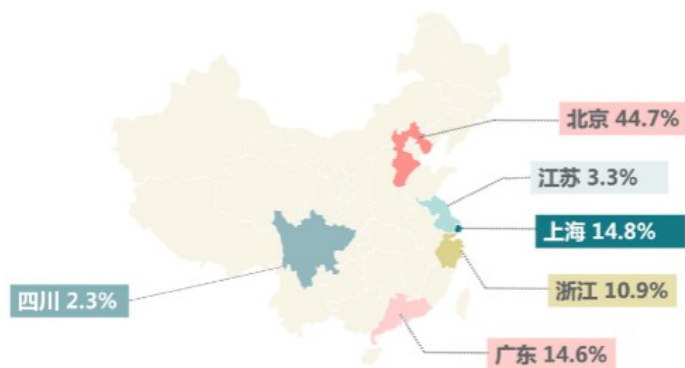
- 2023年“两会”过后，国务院组建**国家数据局(数据化国家队)**
- “十五五”提出深化“**人工智能+**”、提升全民数字技能、强化人工智能就业创造效应
- 2026 年**高质量数据集建设行动**新增对数据治理、专家标注、AI-Ready 数据和行业复合人才的需求。
- 急需具备大数据技能的新工科人才：**理论基础+工程实践经验**





□ 大数据人才缺口—市场需求

- 大数据专家、AI与机器学习专家位列 2025—2030 年全球增长最快的岗位。
- 到 2030 年，接近 40% 的岗位技能预计发生变化，AI与大数据是增长最快的技能方向。
- 中国人工智能研究人员由 2015 年约 4400 人增至 2024 年约 5.2 万人；需求集中在算法、机器视觉、智能驾驶和行业应用。

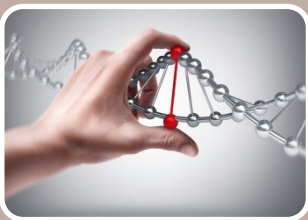




数据科学基础

36

□ 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



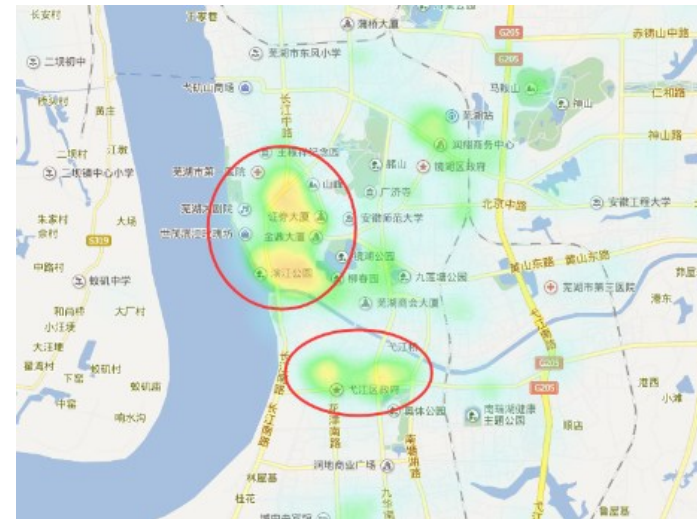
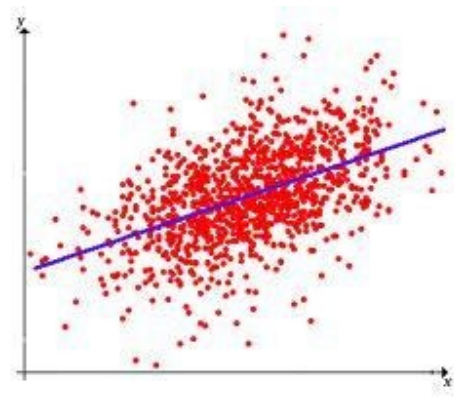
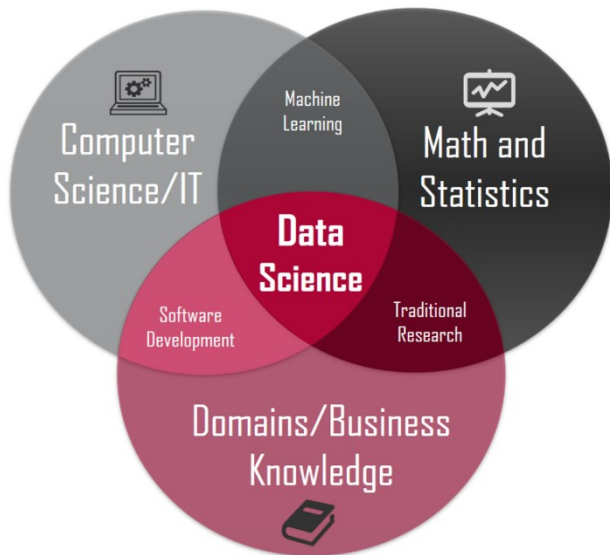
跨界能力强，能够解决特定行业的应用问题



数据科学基础

37

- 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

38

- 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
 - 。 。 。

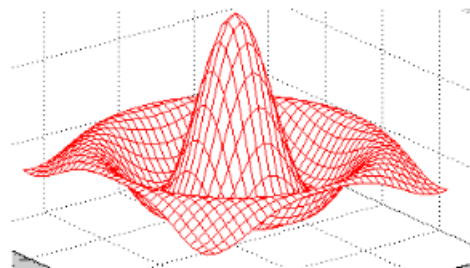


- 1. 理论基础扎实，能理解运用数据科学中的理论模型
 - 数学是学习数据科学的基础
 - 例如，数学与优化：数学分析的应用



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

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

$$w = \operatorname{argmin}_w L(w)$$

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
- . . .



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：线性代数中的应用

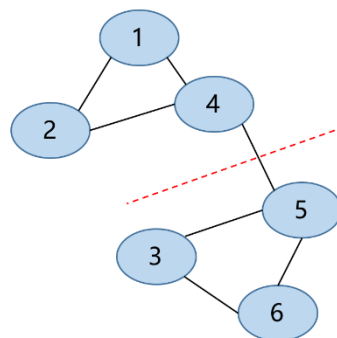
社交网络划分：物以类聚，人以群分



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

- ✓ 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



数据科学基础

41

- 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
 - 大模型：算法，工具，平台等

```
1 def SumOfKSubArray(arr, n, k): SumOfKSubArray(arr, n, k):
2     Sum = 0
3     S = deque()
4     G = deque()
5     for i in range(k):
6         while (len(S) > 0 and arr[S[-1]] >= arr[i]):
7             S.pop()
8         while (len(G) > 0 and arr[G[-1]] <= arr[i]):
9             G.pop()
10        S.append(i)
11        G.append(i)
12    for i in range(k, n):
13        Sum += arr[S[0]] + arr[G[0]]
14        while (len(S) > 0 and S[0] <= i - k):
15            S.popleft()
16        while (len(G) > 0 and G[0] <= i - k):
17            G.popleft()
18        while (len(S) > 0 and arr[S[-1]] >= arr[i]):
19            S.pop()
20        while (len(G) > 0 and arr[G[-1]] <= arr[i]):
21            G.pop()
22        S.append(i)
23        G.append(i)
24        Sum += arr[S[0]] + arr[G[0]]
25    return Sum
```



知乎 @邢建







数据科学基础

43

改变这个世界的第四种力量—大数据的应用

暴力



金钱



世界著名未来学家托夫勒
《第三次浪潮》作者

知识



大数据



数据科学基础

44

- 数据蕴含着巨大的价值—智慧教育

社会科学研究



自然科学分析

理论+实证研究:

经验/实证/量表分析等

教师A:
注重学生的
创新能力和
批判性思维



学生: 表现活跃,
经常主动发言, 但
会打断他人发言



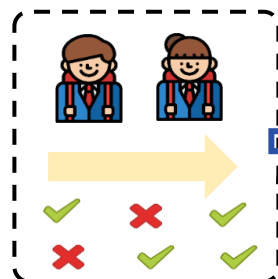
教师B:
重视纪律性和
团队协作能力



理论+实证+大数据研究:

理论假设、数据建模、数据分析、数据验证等

学生的学习过程

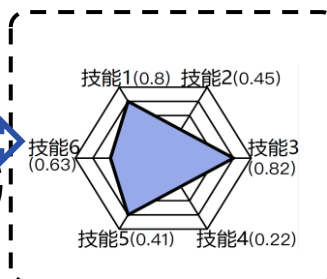


认知
心理学

心理测
量学

计算机
科学

知识掌握状态





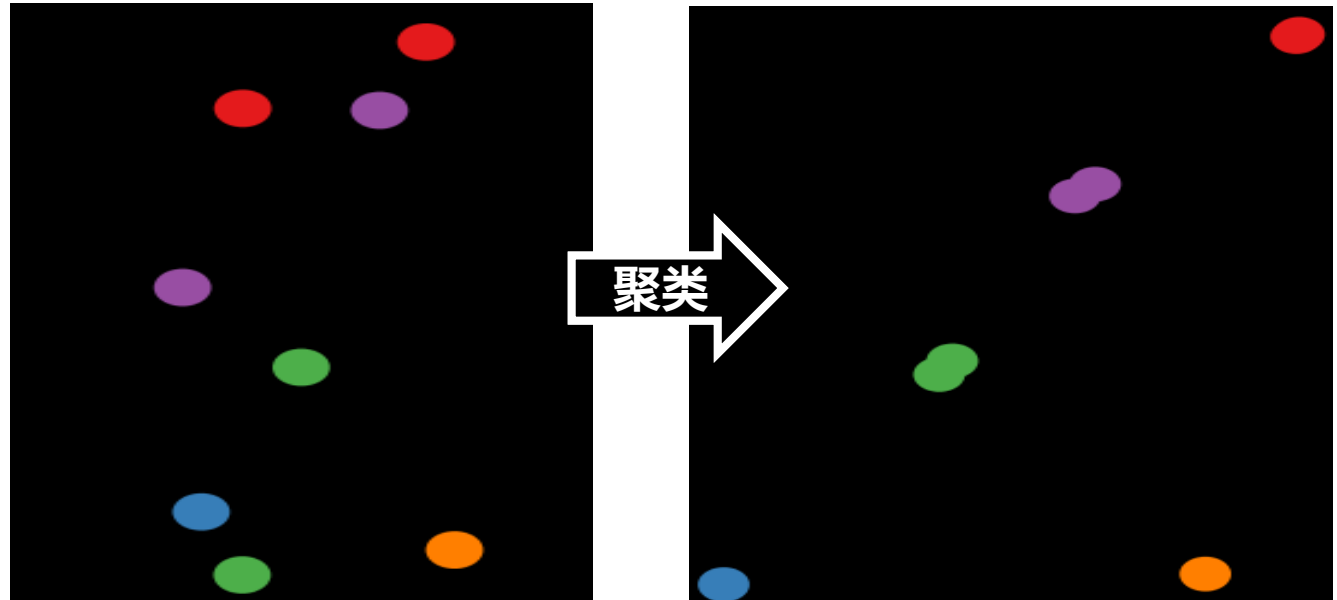
数据科学基础

45

- 数据蕴含着巨大的价值
 - 优化教师教学



根据考试数据对班级进行简单聚类，根据聚类结果，发现**70%**的类里，两个班集是同位授课教师





数据科学基础

46

- 数据蕴含着巨大的价值—智慧教育
 - 考试试题质量与公平性



关于英语科目考试成绩的说明

[发布时间:2018-11-27 阅读量:1570]

浙江省高考英语科目一年安排2次考试，考生可报考2次，选用其中较高1次的成绩。在2018年11月刚结束的英语科目考试中，根据答卷试评情况，发现部分试题与去年同期相比难度较大。为保证不同次考试之间的试题难度大体相当，浙江省招委组织专家研究论证，在制订评分细则时，决定面向所有考生，对难度较大的第二部分（阅读理解）、第三部分（语言运用）的部分试题进行难度系数调整，实施加权赋分。其他试题未作调整。

浙江省教育考试院
2018年11月27日



数据蕴含着巨大的价值—智慧教育

首页 > AI实验室 > 图片转写

关于函数 $f(x) = 4\sin\left(2x + \frac{\pi}{3}\right) (x \in \mathbb{R})$ ，有下列命题：

① $y=f(x)$ 的表达式可改写为 $y = 4\cos\left(2x - \frac{\pi}{6}\right)$ ；

② $y=f(x)$ 是以 2π 为最小正周期的周期函数；

③ $y=f(x)$ 的图象关于点 $\left(-\frac{\pi}{6}, 0\right)$ 对称；

④ $y=f(x)$ 的图象关于直线 $x = -\frac{\pi}{6}$ 对称。

其中正确的命题的序号是_____。（注：把你认为正确的命题的序号都填上）

识别结果

序号	内容
1	关于函数 $f(x)=4\sin\left(2x+\frac{\pi}{3}\right) (x\in\mathbb{R})$ ，有下列命题：关于函数 $f(x)=4\sin\left(2x+\frac{\pi}{3}\right) (x\in\mathbb{R})$ ，有下列命题：
2	① $y=f(x)$ 的表达式可改写为 $y=4\cos\left(2x-\frac{\pi}{6}\right)$ ；
3	② $y=f(x)$ 是以 2π 为最小正周期的周期函数；
4	③ $y=f(x)$ 的图象关于点 $\left(-\frac{\pi}{6}, 0\right)$ 对称；
5	④ $y=f(x)$ 的图象关于直线 $x=-\frac{\pi}{6}$ 对称。
6	其中正确的命题的序号是_____。（注：把你认为正确的命题的序号都填上）

- (1) 图文识别
- (2) 公式转写
- (3) 试题切分
- (4) 属性标注
- (5) 难度预估
- (6) 自动组卷
- (7) 变形题生成
- (8) 相似题检索

.....



数据科学基础

48

数据蕴含着巨大的价值——智慧教育

- 试题表征
- 知识点预测
- 试题属性预估
- ...



大模型插件技术+知识增强

2023年高考试卷分析基于大模型插件技术，自研知识增强的解题分析模型，围绕高考改革和命题要求进行高考试卷和试题的分析，展现新时代高考改革的新面貌。

[立即前往 →](#)



区分度+信度+多学科分析

2022年高考试卷分析基于人工智能+大数据技术与实际结合增加了区分度、信度等多方面的分析，并且支持更多学科分析，带来更全面的高考解读

[立即前往 →](#)

多维度、深层次分析历年高考试卷

2021年高考试卷是怎样的？中科大利用人工智能带你再开“天眼视角”

中国科大本科招生 2021-06-25 19:39 发表于安徽

以下文章来源于中国科大大数据分析省重点实验室，作者BDAA-LUNA团队



中国科大大数据分析省重点实验室

中国科学技术大学大数据分析与应用安徽省重点实验室，依托科大计算机学院和大数...



在2020年，中国科学技术大学智慧教育研究团队首次利用人工智能+大数据的分析方式针对2020年的高考数学（点击回顾）进行了解读，带大家一种观察高考的全新视角。今年我们如约而至并且**全新升级**，由数学单科的分析，升级为全科分析，依然利用强大的人工智能+大数据分析方式，为大家带来今年高考试卷的全方位解读，为了让大能更方便快速的感受高考的魅力，本文以**语数英**为例，抽取报告的部分**精彩内容**给大家展示。

想要获取完整版报告，关注“中国科大大数据分析省重点实验室”公众号，在后台回复“2021报告”即可获得。

“数学分析”

首先我们一起来看看今年数学学科高考试题的内容。数学学科的特点是具有**高度抽象性**和**严密的逻辑性**，在去年分析的基础上，我们优化了各个分析的模型，力求能全面、准确的展示今年高考数学的特点。

数形结合百般好



全新升级 | 2022高考试卷分析完整版新鲜出炉！更丰富的学科！更全面的分析！

原创 BDAA-LUNA团队 中国科大大数据分析省重点实验室 2022-06-30 20:30 发表于安徽

[点击蓝字 关注我们](#)

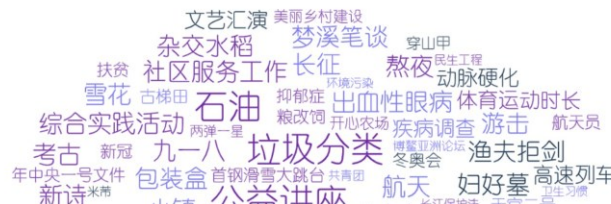
在2020年、2021年，中国科学技术大学智慧教育研究团队均利用人工智能+大数据的分析方式对高考试卷试题进行了解读，带大家一种观察高考的全新视角。新的一年有新的变化，今年教育部发布了《关于做好2022年普通高校招生工作的通知》，其中强调要全力保障高考安全稳定、进一步深化高考改革。我们也全新升级，增加了**区分度、信度**等多方面的分析，并且支持更多学科分析，带来更全面的高考解读。

想要获取完整版报告，关注“中国科大大数据分析省重点实验室”公众号，在后台回复“2022报告”即可获得。

01 总体分析

我们先对各学科的整体情况进行了**综合分析**。我们综合了2022年全国卷与新高考试卷的数据，从情境关键词、字数等方向进行解析统计，并展示了各省新高考的改革情况，对今年高考试卷的整体情况做一个总结。

情境关键词





数据蕴含着巨大的价值——智慧教育



请给出对于2023年全国卷高考语文的四道作文题的评析

全国乙卷作文题以习近平总书记的讲话为引子，让考生思考个体与集体的关系以及多样性的重要性。这个题目涉及到人际交往、社会发展和文化多元化等问题，考生可以从不同的角度进行论述，展示自己的思维深度和独立见解。

新高考一卷作文题以故事的力量为主题，引导考生思考故事在传递信息、表达情感和改变命运方面的作用。这个题目可以引发考生对于文学艺术、传统文化和故事叙述的思考，同时也能让他们关注到故事在现实生活中的实际应用。

全国甲卷作文题围绕技术发展对时间的影响展开，引发考生对于科技进步

2023年高考语文作文评析



今年全国乙卷文科数学的选择中最好玩的是哪道题，为什么？

从娱乐性的角度来看，我认为第3题“网格纸上绘制的一个零件的三视图，网格小正方形的边长为1，求零件的表面积”是最有趣的题目。

这道题目之所以有趣，是因为它需要学生通过观察图形，结合几何知识计算一个零件的表面积。通过观察图形，学生需要注意到零件有三个视图，即正视图、侧视图和俯视图，并且要根据这些视图来计算表面积。

这道题目涉及到几何图形的认知和计算，需要学生在综合运用知识的同时，进行空间想象和几何判断。对于喜欢解决几何问题、喜欢思考图形和空间关系的学生来说，这道题目可能会更具趣味性。

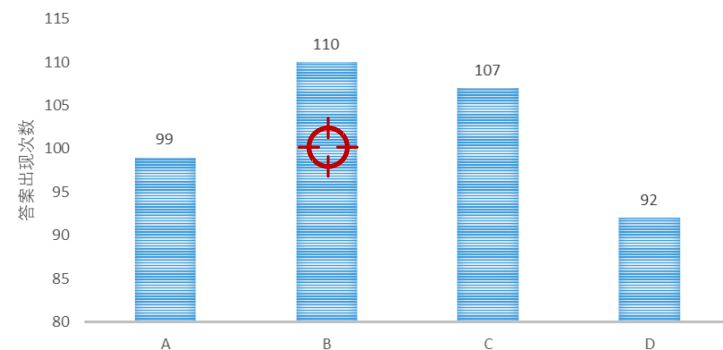
2023年全国乙卷文科数学“最好玩”的题

知识增强大模型评高考



对18年至22年的高考数学试卷中的人名进行了识别，上图中展现了不同人名所对应的题目难度。“小明”位居榜首。

2016-2020选择题选项答案出现次数



选择题金C银B靠谱吗？

高考“冷知识”

数据蕴含着巨大的价值—智慧教育

任务驱动：小模型独立负责多种任务

EduNLP

教育资源理解算法库

python 3.6 | 3.7 | 3.8 | 3.9 pypi v0.0.9 test passing codecov 98% docs passing
downloads 40/month license Apache-2.0 DOI 10.5281/zenodo.7318103

EduNLP is a library for advanced Natural Language Processing in Python and is one of the projects of EduX plan of BDAA. It's built on the very latest research, and was designed from day one to be used in real educational products.

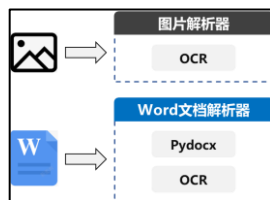
成分分解

语法解析

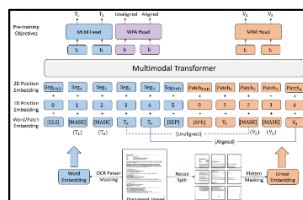
令牌化

向量化

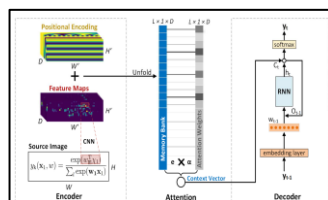
下游模型



文档解析

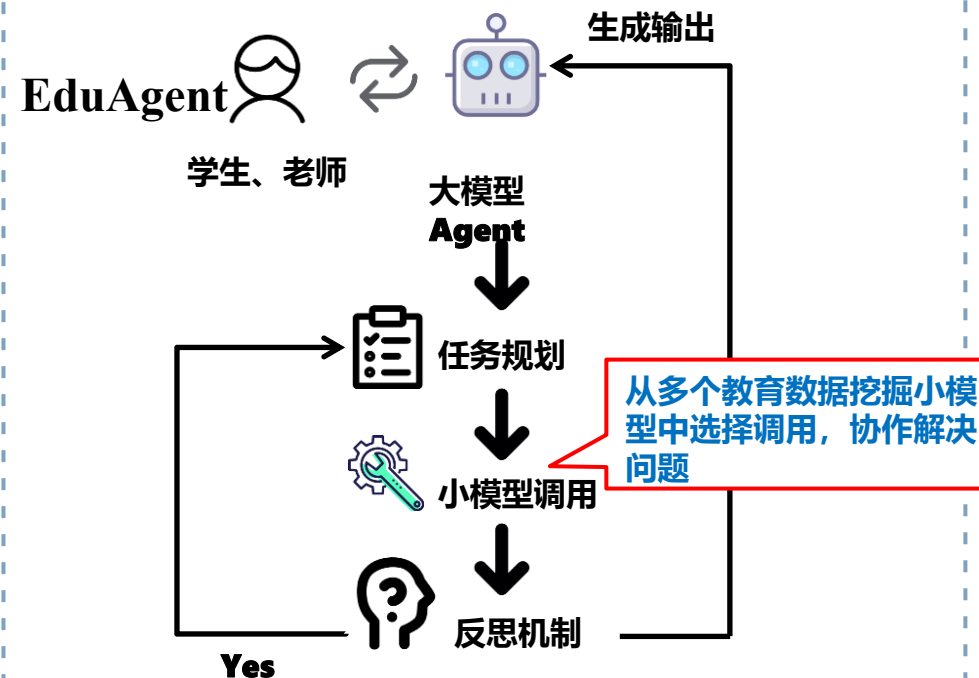


公式识别



属性理解

智能体驱动：大模型与小模型协作



EduNLP: 面向学科资源理解的小模型算法库

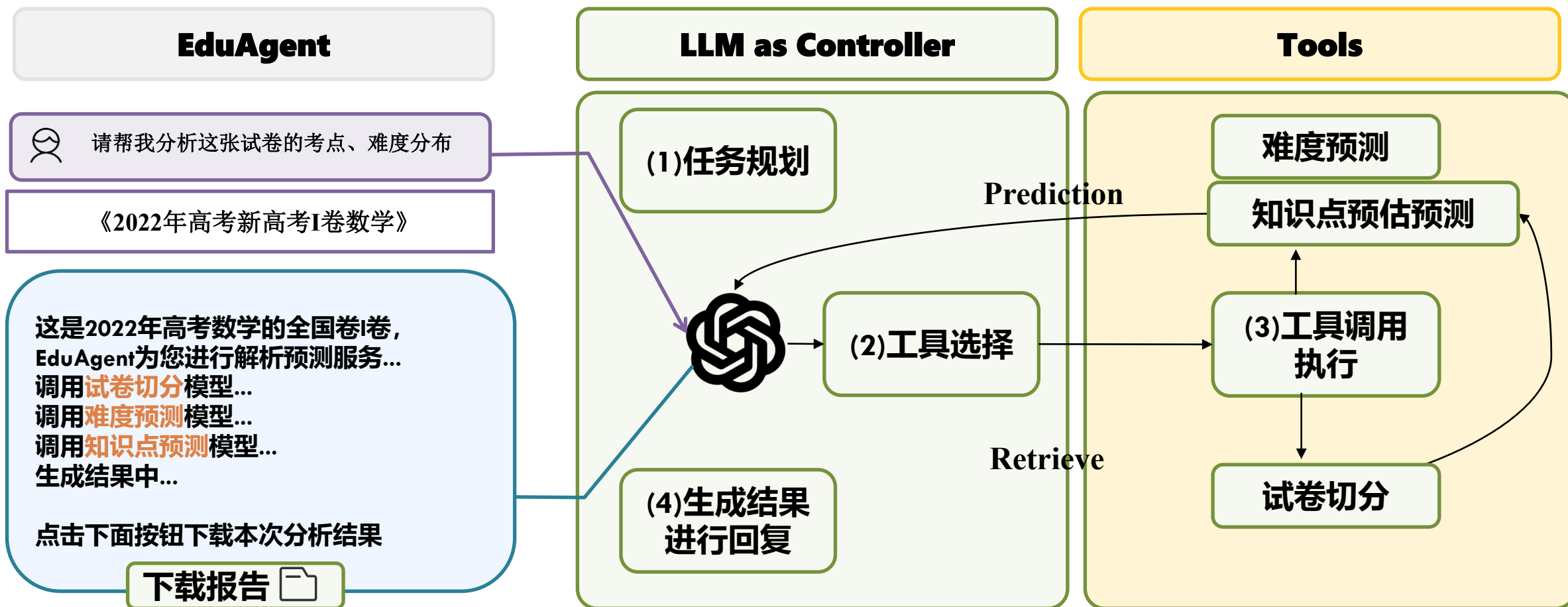
EduAgent: 学科资源复杂分析的教育大模型智能体



数据科学基础

51

数据蕴含着巨大的价值——智慧教育





数据科学基础

52

- 数据蕴含着巨大的价值—智慧教育
 - 大模型驱动的智慧教育

学：智慧助教

教：数字教师

研：科研助手

管：以智助管

大模型学习助手

什么是布斯算法？已知二进制数 $X=0.1010$, $Y=-0.1101$ ，利用布斯算法求 $[X*Y]$ 的补码。

布斯算法是一种补码乘法算法，用于计算两个用补码表示的二进制数的乘积。在布斯算法中，被乘数和乘数均用补码表示，连同符号位一起参与运算。布斯算法的运算过程包括将两个数的补码相乘，并根据乘数相邻两位的情况进行特定操作，最终得到乘积的补码。

根据已知的二进制数 $X=0.1010$, $Y=-0.1101$ ，可以进行布斯算法求解 $[X*Y]$ 的补码：

1. 将 X 和 Y 表示为补码： $[X]_{补} = 00.1010$
 $[Y]_{补} = 11.0011$ $[-X]_{补} = 11.0110$

2. 根据布斯算法的运算过程，进行乘法运算，得到乘积的补码： $[X*Y]_{补} = 1.01111100$

因此，根据布斯算法， $[X*Y]$ 的补码为1.01111100。



星火科研助手

星火科研助手致力于创建高效、可靠的科技文献服务平台，为学术研究人员、企业专业人士和其它相关用户提供服务，助力科技发展创新。

成果调研 论文研读 辅助写作



关注订阅号
及时了解科研服务前沿信息



关注服务号
随时随地了解科研助手动态





数据科学基础

53

- 数据蕴含着巨大的价值——社会科学
 - ◆ 社交媒体 比 问卷调查 提供了更有代表性的结果
 - ◆ 智能引导社会成员的行为



15万名奥巴马支持者在Facebook安装了“奥巴马2012”应用，而通过这个程序，总统竞选团队可以间接得到这些支持者数百万的Facebook好友信息。



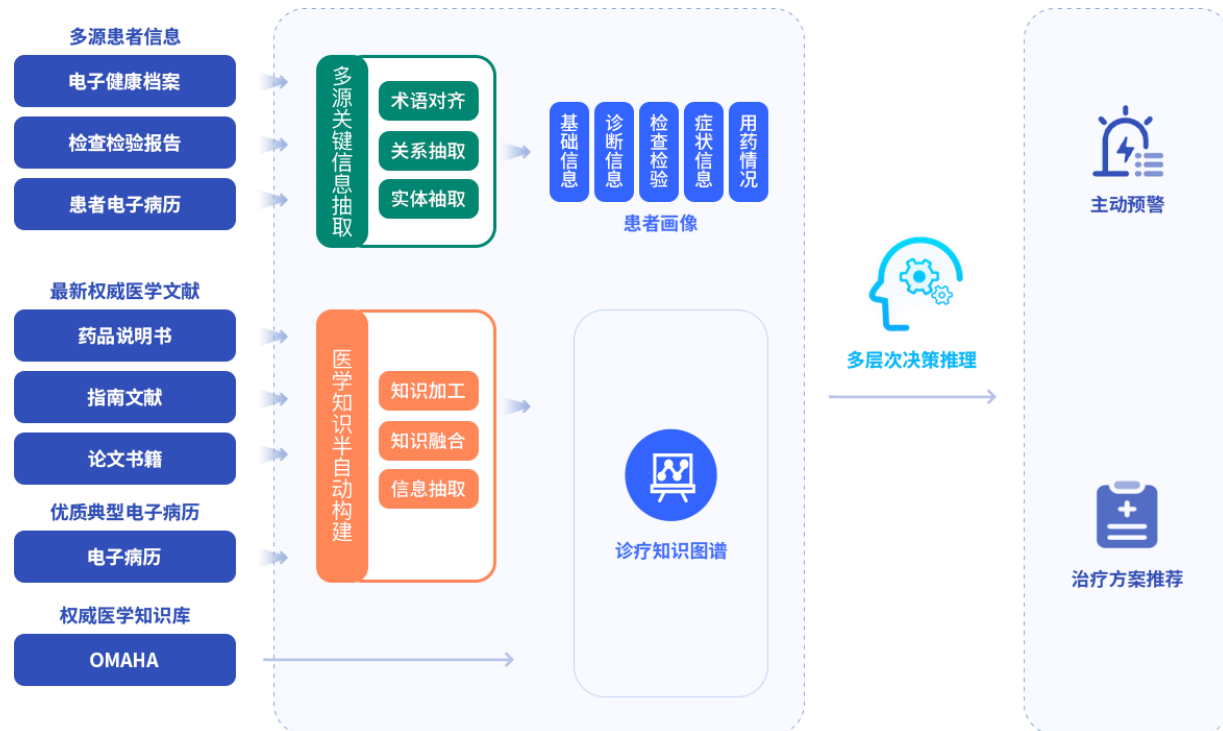
有一种说法称，特朗普的团队聘用数据分析公司，做了精准的广告投放，影响了那些徘徊不定的选民，拿下了决定性的关键州选举人票





数据蕴含着巨大的价值——智慧医疗

- 通过整合患者的**全时段**、**多模态**的健康数据（病例文本、检查影像等）与多源医疗数据
- 通过智能推理分析患者病情，为医生提供**疾病风险预警**、**诊疗方案推荐**、**合理用药审核**和**个性化治疗建议**。

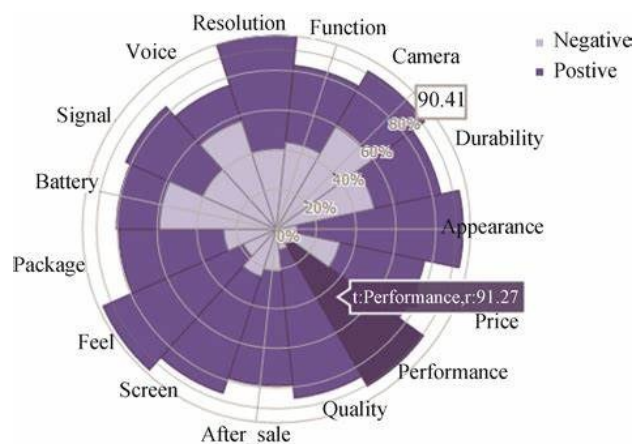




数据蕴含着巨大的价值—安防领域

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

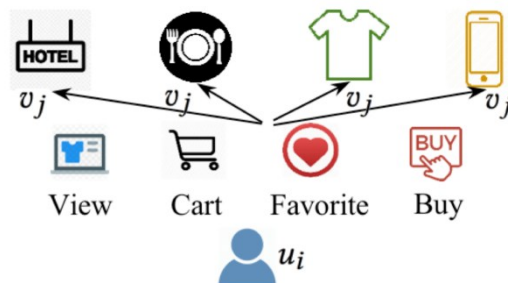


数据科学基础

- 数据蕴含着巨大的价值—电子商务
 - 计算广告



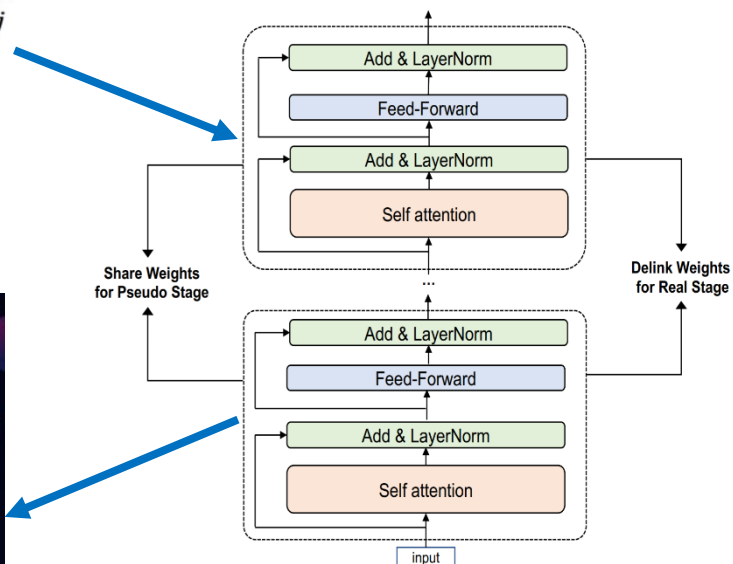
电商平台



海量用户多样交互行为



促进用户消费、提升平台收益



达摩院10万亿参数
M6-10T模型

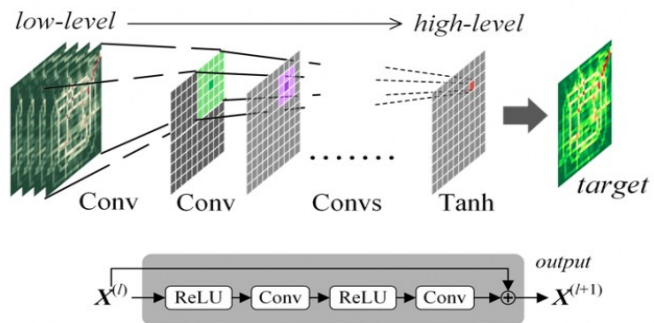


- 数据蕴含着巨大的价值 — 智慧城市
 - 基于运营商基站数据、交通路口和车辆移动轨迹数据等
 - 提升城市**交通管控**、**交通服务**和**规划水平**，实现**用户未来行程规划**、**交通路口信号灯调控**、**合理规划道路建设**等

车辆
移动
数据



交通
流预
测模
型



行程规划



信号灯调控



道路规划



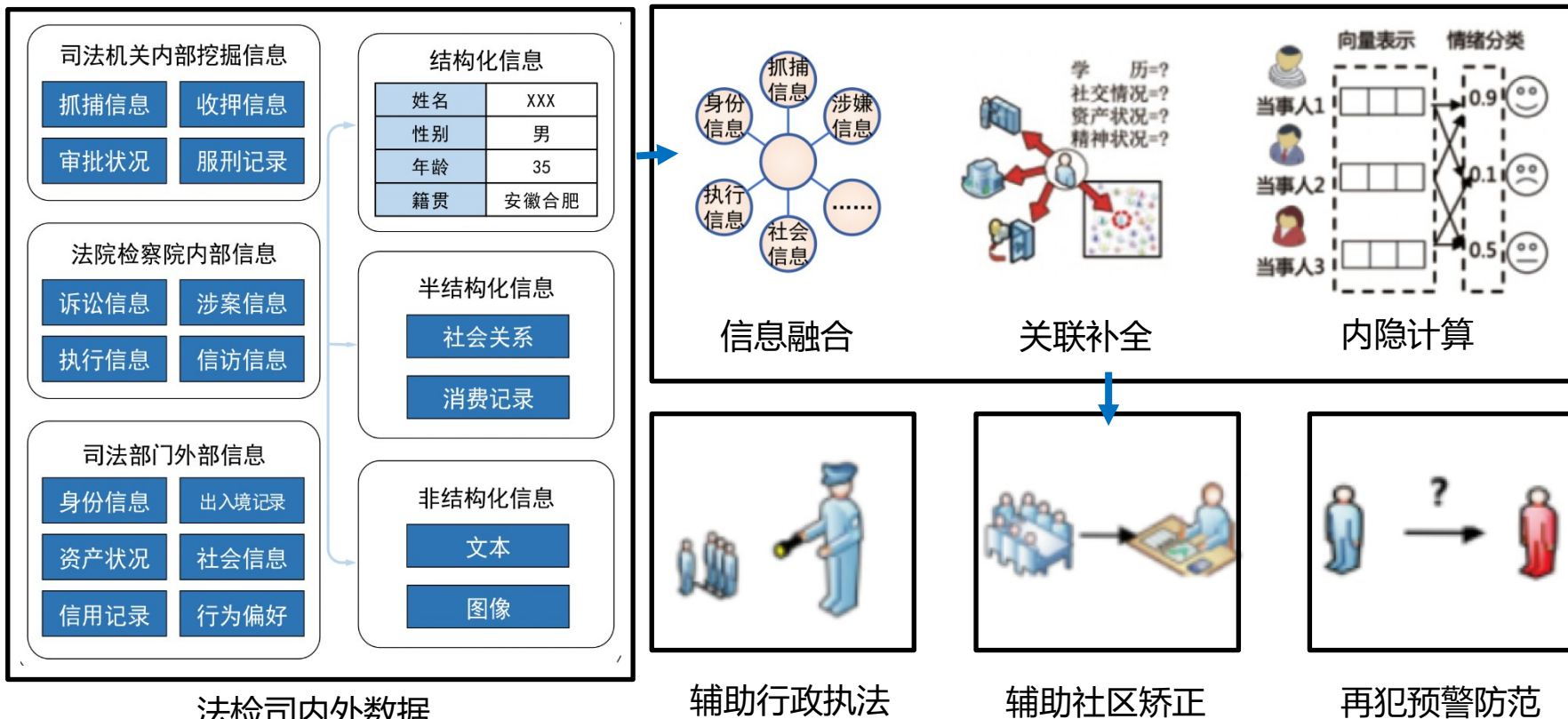


数据科学基础

58

数据蕴含着巨大的价值——智慧司法

- 基于法、检、司等部门关于涉案当事人的内外部数据，构建涉案当事人画像，辅助**行政执法**、**社区矫正**、进行**再犯预警防范**等，**提升司法管理服务的质量和效率**
- 对起诉状、证据材料、庭审记录和裁判文书等电子卷宗数据进行分析处理，**辅助法官判案**。

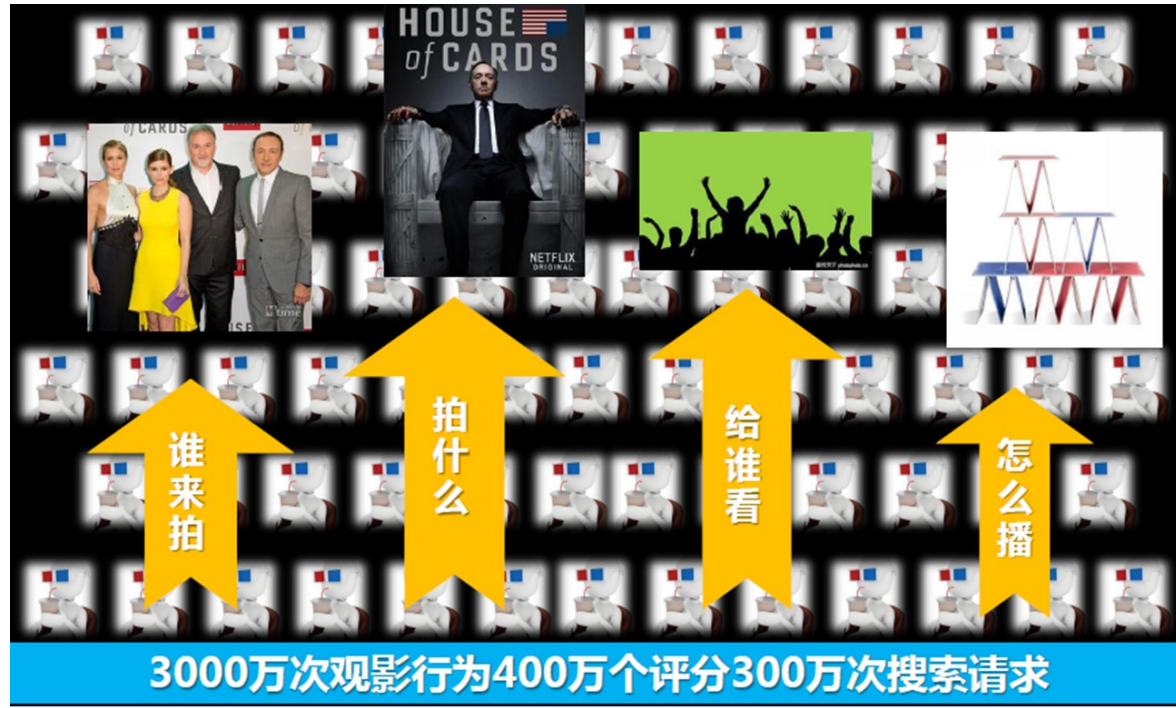




数据科学基础

59

- 数据蕴含着巨大的价值—文化娱乐
 - 纸牌屋效应：数据决定影视剧的内容



大卫·芬奇
凯文·史派西

老版《纸牌屋》

喜欢老版纸牌屋
及同类剧的用户

13集同时上线



- 数据蕴含着巨大的价值—企业与管理
 - 公司管理方面：大数据优化人力资源管理

