



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

新媒体大数据分析

New Media Big Data Analysis

第一章 数据科学基础

黄振亚, 朱孟潇, 张凯, 唐秀秀

课程主页:

<http://staff.ustc.edu.cn/~huangzhy/Course/NM2026.html>

助教: 谭施越 朱嘉骏

bigdata_20262026@163.com

8/31/2026



数据科学基础

3

- 国内外优秀高校信息与传播学院开展“数据分析”等专业课程，以培养科教融合、协同育人人才培养模式的新模式

地区	学校	院系	课程	学分	类别	年级
中国	中国人民大学	新闻学院	统计数据分析	2	通识教育	大一、大二
中国	复旦大学	新闻学院	数据分析与信息可视化	3	专业基础课	大三
中国	山东大学	新闻传播学院	大数据推断基础（双语）	2	专业选修课	大一
中国	华中科技大学	新闻与信息传播学院	人工智能与数据科学	2	学科基础课	大二
中国	中国科学技术大学	计算机科学与技术学院	数据科学导论	2	核心通识	大二及以上
美国	University of Illinois at Urbana, Champaign	Journalism	Data Science Discovery	4	专业课	undergraduate
美国	Stanford University	Humanities & Sciences	Data Science 101	5	专业课	undergraduate
美国	The University of Chicago	Humanities	Data Analysis I: Introduction to Statistics	100-units	专业课	undergraduate
美国	Carnegie Mellon University	Humanities and Arts	Methods for Statistical & Data Science	9units	专业课	undergraduate
美国	University of Florida	Journalism and Communications	Advanced Data Journalism	3	专业课	undergraduate



课程基本情况

4

- 科技传播系（25系）新增开设“**网络与新媒体**”专业
 - 以原有的传播学专业为基础，借助理工科培养的优势，初步形成了“文理管交叉、产学研一体”的复合型人才培养模式
 - 致力于培养具有深厚的理工知识背景、扎实的网络传播和新媒体应用技能、卓越的新媒体策划与运营理念、出色的中英文表达能力及国际化视野的高级专门人才
- 开展《**新媒体大数据分析**》课程建设，更好地适应该专业人才培养目标
 - **面向科技传播系高年级本科学生开放**，将通过理论教学、案例分析与课题实践相结合的方式，全景介绍大数据科学分析的基础概念、基本方法、应用实践与前沿概况
 - 使学生树立运用大数据分析和相关算法工具解决实际问题的基本思路，并初步掌握新媒体大数据分析中解决实际问题应用问题的能力。



课程基本情况

5

通过开展《新媒体大数据分析》课程建设，希望达到以下本科教学培养目标：

知识目标

- 了解数据科学的基本概念、思路
- 初步掌握数据处理、数据分析、结果可视化等入门级知识
- 熟悉新媒体大数据分析前沿概况和经典应用场景

能力目标

- 树立运用数据和相关算法工具解决实际问题的基本思路
- 初步掌握用于数据分析手段解决实际问题能力，即如何使用科学的方法研究和应用数据（包括用数据的方法研究科学）

素质目标

- 培养正确的“**数据价值观**”，增强从数据科学的角度认识社会、理解社会、服务社会的意识



课程基本情况

6

《新媒体大数据分析》专业基础课程共**40学时理论+20学时实验**：

□ 第一部分：数据科学基础

- 重点介绍**课程定位**，概述数据科学与大数据分析的发展历程，和重点案例

□ 第二部分：大数据分析基础

- 介绍大数据分析的生命周期流程，培养学生认识数据的特点，**掌握数据分析基本原理和方法**，掌握数据特征的提取和评价方法等。内容包括：
 1. 数据采集与存储
 2. 数据预处理与特征工程
 3. 数据统计分析



建设内容

7

□ 第三部分：大数据分析算法

- 介绍**大数据分析的基本算法**。培养学生掌握分类、聚类、关联规则挖掘等基本分析算法：
 1. 大数据分析 with 分类
 2. 大数据分析 with 聚类
 3. 大数据分析 with 关联规则挖掘

□ 第四部分：新媒体大数据分析应用

- 围绕新媒体数据，介绍大数据分析成功案例。培养学生掌握对不同类型的**新媒体数据****进行实际分析和应用的能力**，内容拟包括：
 1. 文本大数据分析案例
 2. 图片与视频大数据分析案例
 3. 网络大数据分析案例

□ 实验介绍：

- Python基础、分类与聚类实验等



课程几点说明

9

- 2026年课程（NNM2012）
 - 主要面向科技传播系“网络与新媒体专业”高年级本科生
 - 以及 感兴趣的工科方向学生

- 同类型课程： **学分不重复认定**
 - 大数据学院：《数据分析与实践》 DS3001
 - 计算机学院：《数据科学导论》 CS4023

- 关于其他专业的学分认定请咨询本专业教秘



课程几点说明

10

- 2026年课程（NNM2012）
 - 面向科技传播系“网络与新媒体专业”高年级本科生
- 时间: (1-18周)周一下午6,7节: 14:00-15:35
- 地点: 东区2303



课程要求与考核方式

11

- 课程目标：用科学的方法研究和应用数据
- 课程要求
 - 1. 课堂出勤，日常作业
 - 2. 文献调研报告（独立完成）
 - 3. 实验任务（独立完成，需要编程），之后布置给大家
 - 4. 课程交流与课程汇报
- 考核方式
 - 课堂+作业（30%）+调研报告（30%）+实验报告（40%）



联系方式

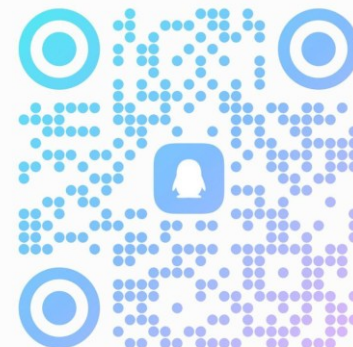
12

- 课程主页：
<http://staff.ustc.edu.cn/~huangzhy/Course/NM2026.html>
- 授课教师：黄振亚，朱孟潇，张凯，唐秀秀
- 助教：
 - 谭施越 朱嘉骏
 - 课程邮箱：bigdata_20262026@163.com
- QQ群：1071511308



2026秋-新媒体...

群号: 1071511308



扫一扫二维码，加入群聊



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

13

新媒体大数据分析

New Media Big Data Analysis

第一章 数据科学基础

黄振亚, 朱孟潇, 张凯, 唐秀秀

课程主页:

<http://staff.ustc.edu.cn/~huangzhy/Course/NM2026.html>

助教: 谭施越 朱嘉骏

bigdata_20262026@163.com

8/31/2026



课程目标

14

- 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

选修新媒体大数据分析课程的同学将来可能从事不同行业的科学研究、技术开发、产品管理等，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。

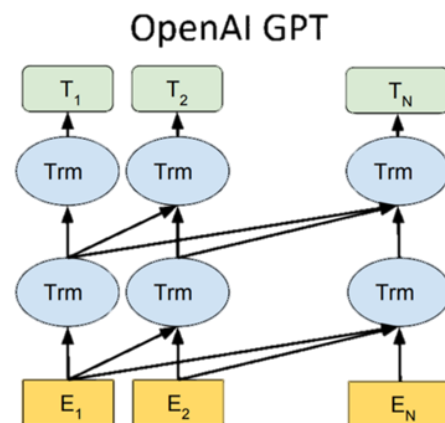


数据科学基础

15

□ ChatGPT：大数据催生人工智能新浪潮

- 参数量从1.17亿增加到1750亿
- 数据量从5GB增加到45TB
 - 96%以上是英文，其它20个语种不到4%



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

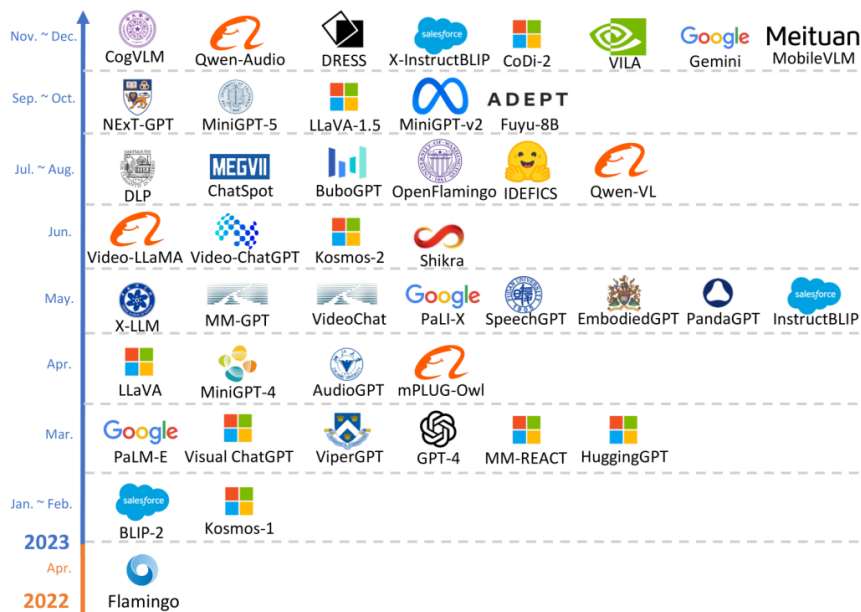


数据科学基础

16

□ 大数据催生人工智能新浪潮- 多模态模型（GPT-5、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

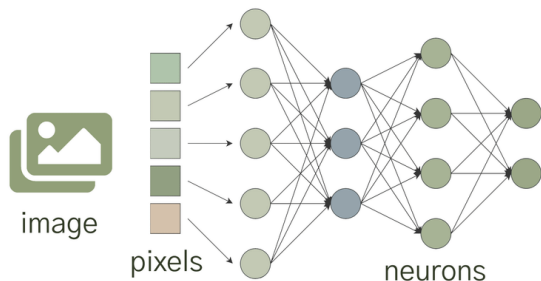




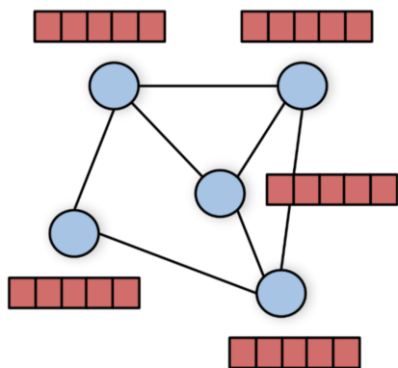
数据科学基础

17

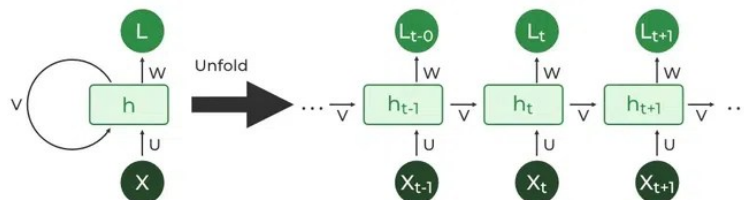
- 以模型为中心(model-centric)的数据科学技术
 - 围绕目标任务特性，设计不同模型结构



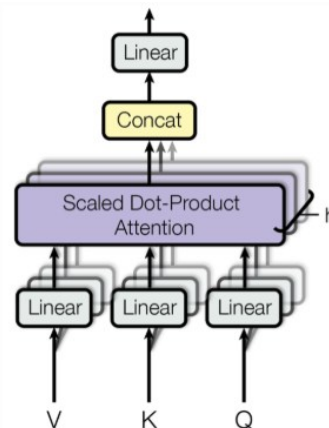
卷积神经网络CNN



图神经网络GNN



循环神经网络RNN



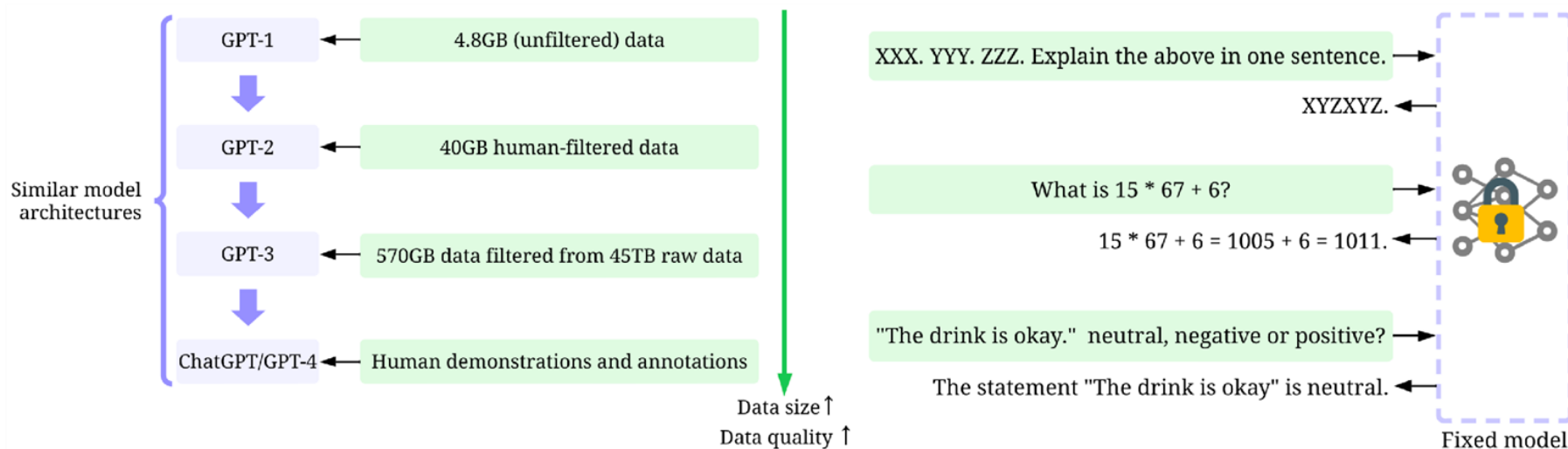
自注意力神经网络Transformer



数据科学基础

18

- 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



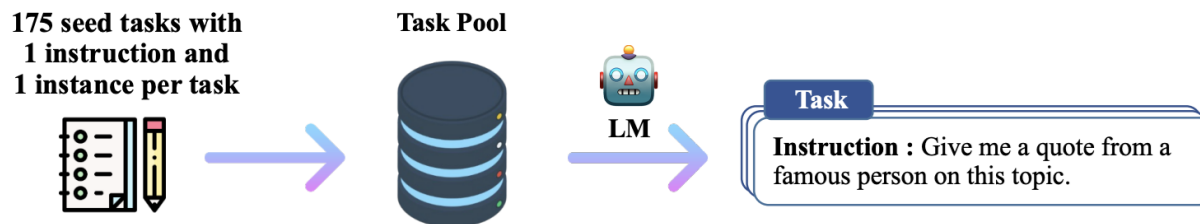


数据科学基础

19

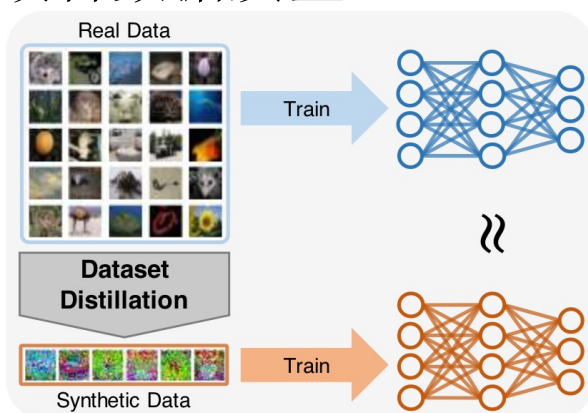
□ 以数据为中心(data-centric)的数据科学技术

□ 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

20

□ 数据

- 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
- 新时代的生产要素（十四五）

□ “人-机-物”三元融合，世界已经成为数据化的世界



当文字成为数据



当方位成为数据

一切事物的数据化



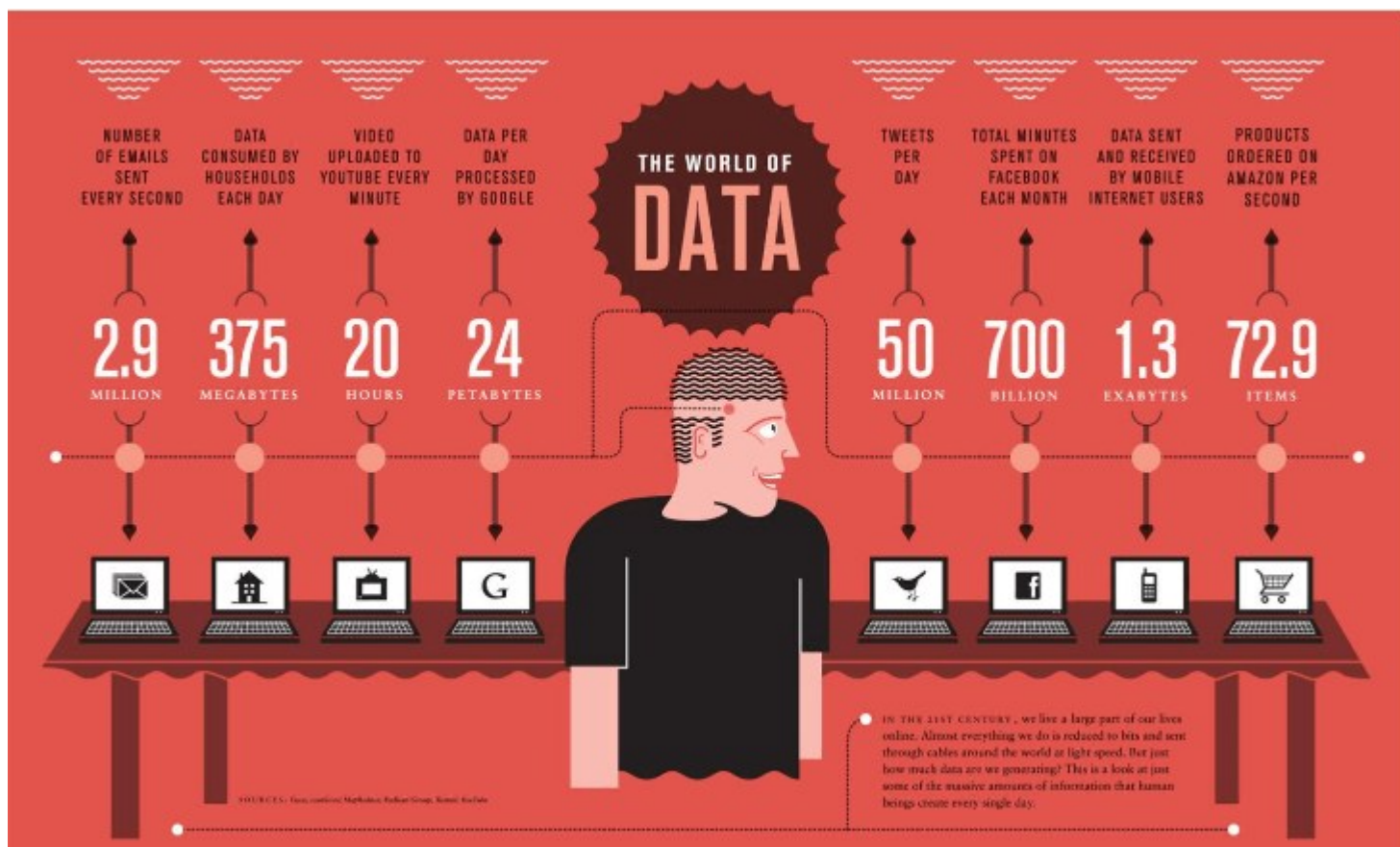
当沟通成为数据



数据科学基础

21

□ 我们生活在数据中，所有人都在制造和分享数据





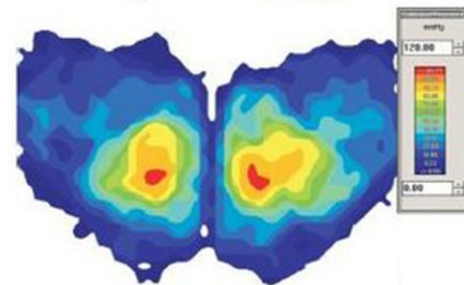
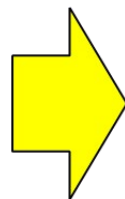
数据科学基础

22

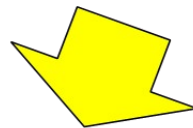
□ 例子：最不可能的地方获得数据

- 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

23

□ 大数据的提出



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

24

□ 铺天盖地的“大数据”

Baidu

大数据

大数据新闻1亿篇

Q 网页 图片 视频 资讯 贴吧 文库 知道 地图 采购 更多

百度为您找到相关结果约100,000,000个

搜索工具

搜索智能精选

大数据_百度百科



*大数据(big data), 或称巨量资料, 指的是所涉及的资料量规模巨大到无法透过主流软件工具, 在合理时间内达到撷取、管理、处理, 并整理成为帮助企业经营决策更积极目的资讯。在维克托·迈尔-舍恩伯格及肯尼

百度百科

大数据 - 官方网站 官方



大数据“完全占领”了互联网和IT领域之后，开始进入各行各业，形成了政府大数据、教育大数据、医疗大数据、交通大数据、金融大数据、保险大数据、公安大数据、法院大数据、旅游大数据、.....



数据科学基础

25

- 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

26

- 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

27

大数据有多大?

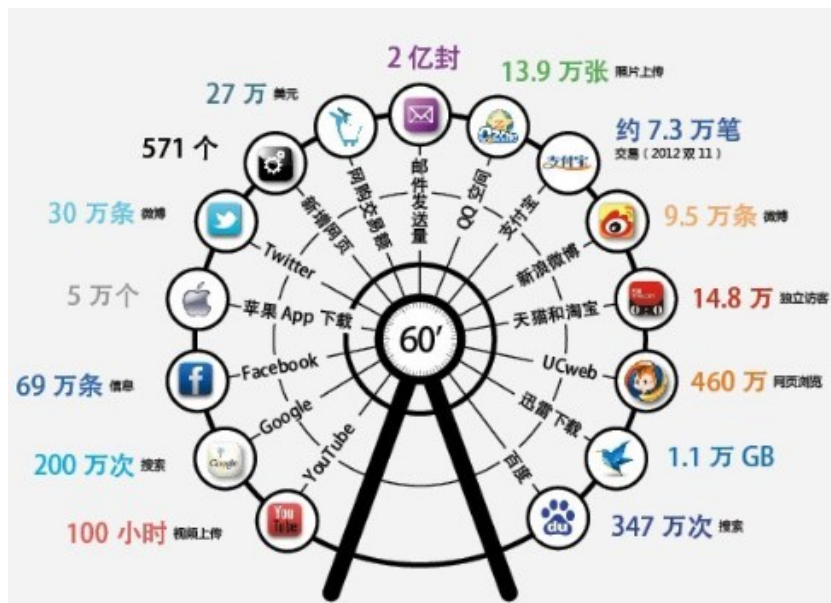
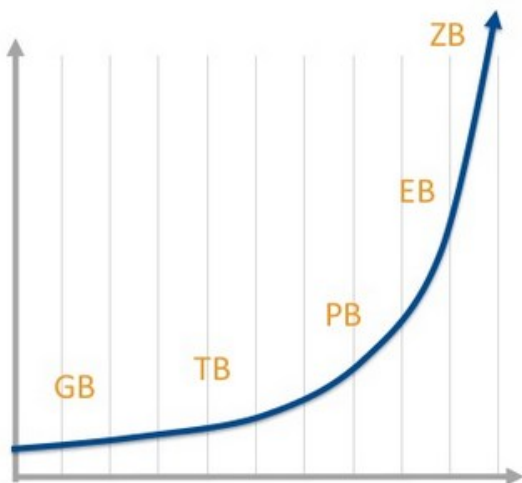
◆ 数据量已到ZB等级

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!



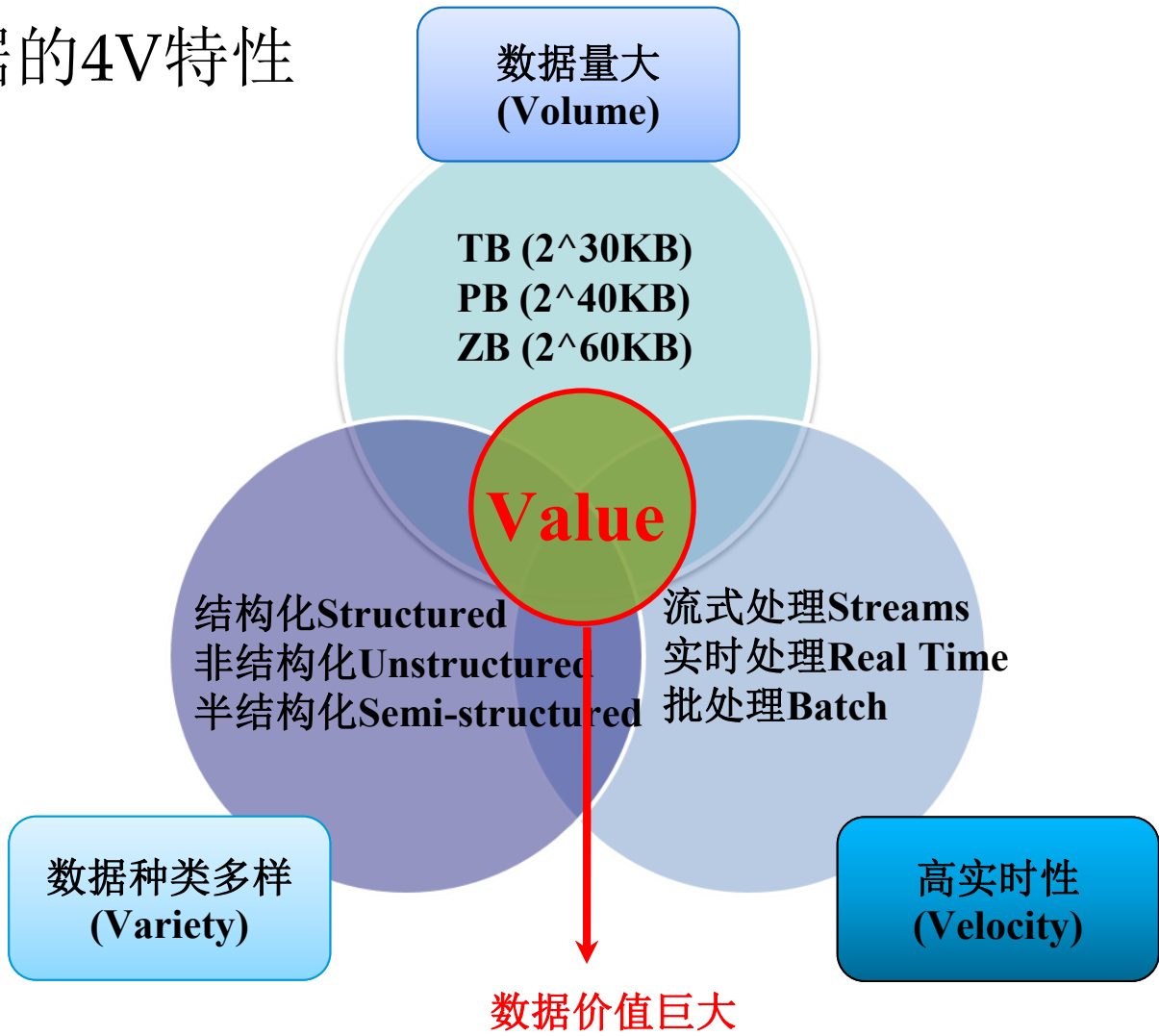
60秒，我们能产生多少数据?



数据科学基础

28

□ 大数据的4V特性





数据科学基础

29

□ 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，预计2025年达到175ZB，我国数据占18%。

——IDC互联网数据中心

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

- 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

30

□ 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



文本数



事务数



视频数



音频数



数据科学基础

31

□ 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

在百度输入关键字：
“汽车维修”、“挖掘机 学习”

某在线电影网站

Baidu 百度 新闻 网页 贴吧 知道

汽车维修

汽车修理培训//汽车修理培训 www.ahslqx.com

汽车修理培训我选三联,专业的师资队伍,一流的校园环境

● 三联汽修学院 ● 初中成才计划 ● 高中成才计划

某IT业界资讯网站



点击查看

汽修第一品牌

咨询电话: 0551-6420 6420

咨询QQ: 8008008099



用心成就未来
挖掘机、装载机、吊车、叉车、叉车技术

0551-63431467

- 挖掘机学员
- 单反相机相片

例如用户在合肥某台PC上，打开百度输入关键字片刻之后，再打开其它网站，就会看到相关的广告，并且所推荐的是地理位置信息相关的（合肥、安徽）



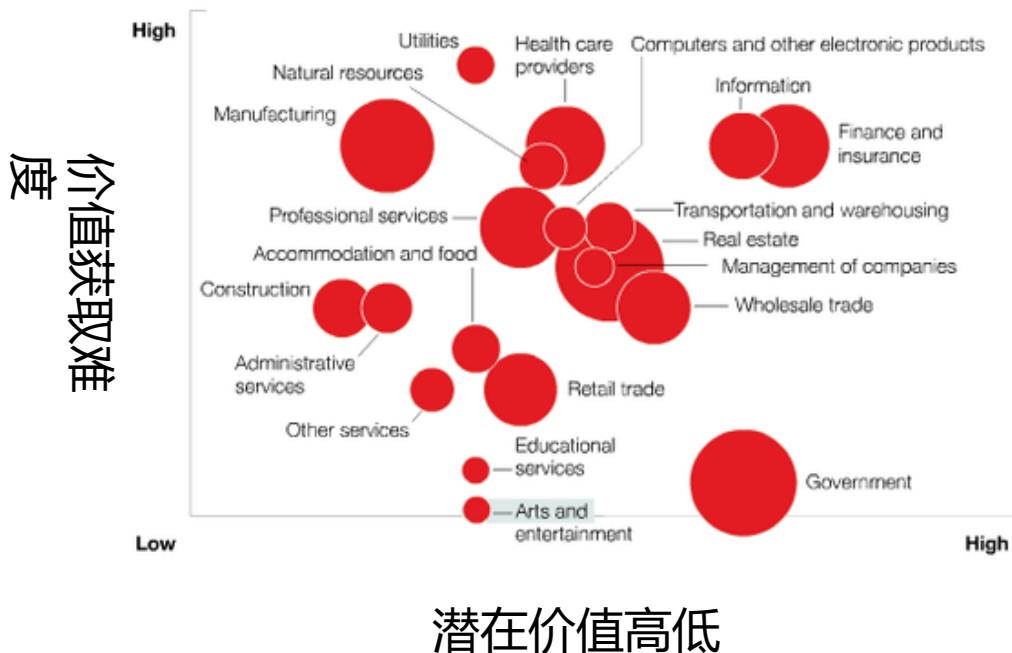
数据科学基础

32

□ 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

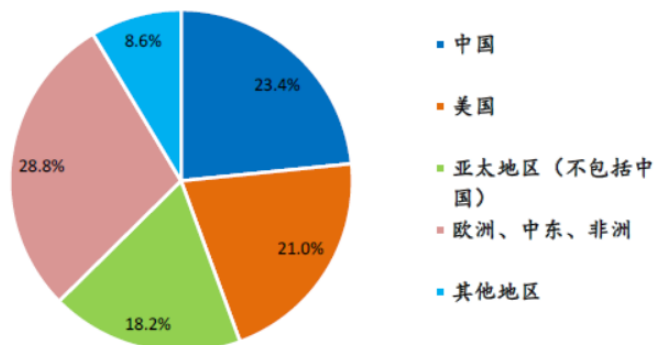


数据科学基础

33

□ 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
 - 2020年我国数字经济规模占GDP比重达到**38.6%(39.2万亿元)**
- 大数据是重塑国家**竞争优势**的重大发展机遇
 - 2018年，中国数据圈占全球**23.4%(7.6ZB)**；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

34

□ 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。
- 2017年，《大数据产业发展规划(2016-2020年)》实施
- 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。



数据科学基础

35

□ 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)



- 加快构建全国一体化**大数据中心体系**，强化算力统筹智能调度，建设若干国家枢纽节点和**大数据中心集群**，建设E级和10E级超级计算中心
- 培育壮大人工智能、**大数据**、区块链、云计算、网络安全等新兴数字产业
- 推动5G、**大数据中心**等新兴领域能效提升
- 完善适用于**大数据**环境下的数据分类分级保护制度（大数据安全）



数据科学基础

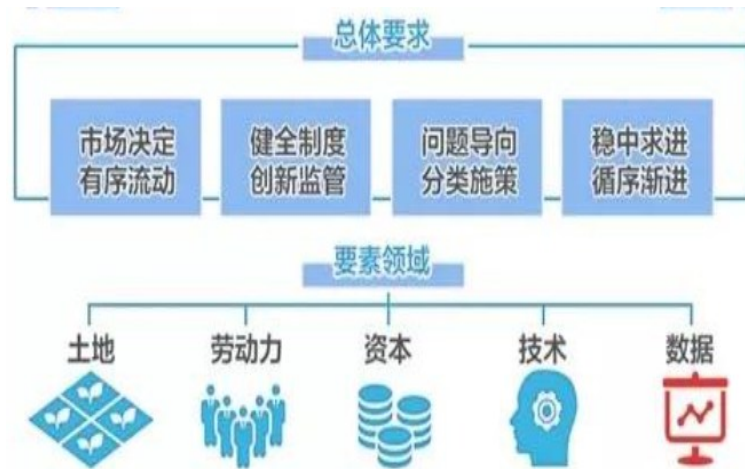
36

□ 十四五规划中的“大数据”

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。





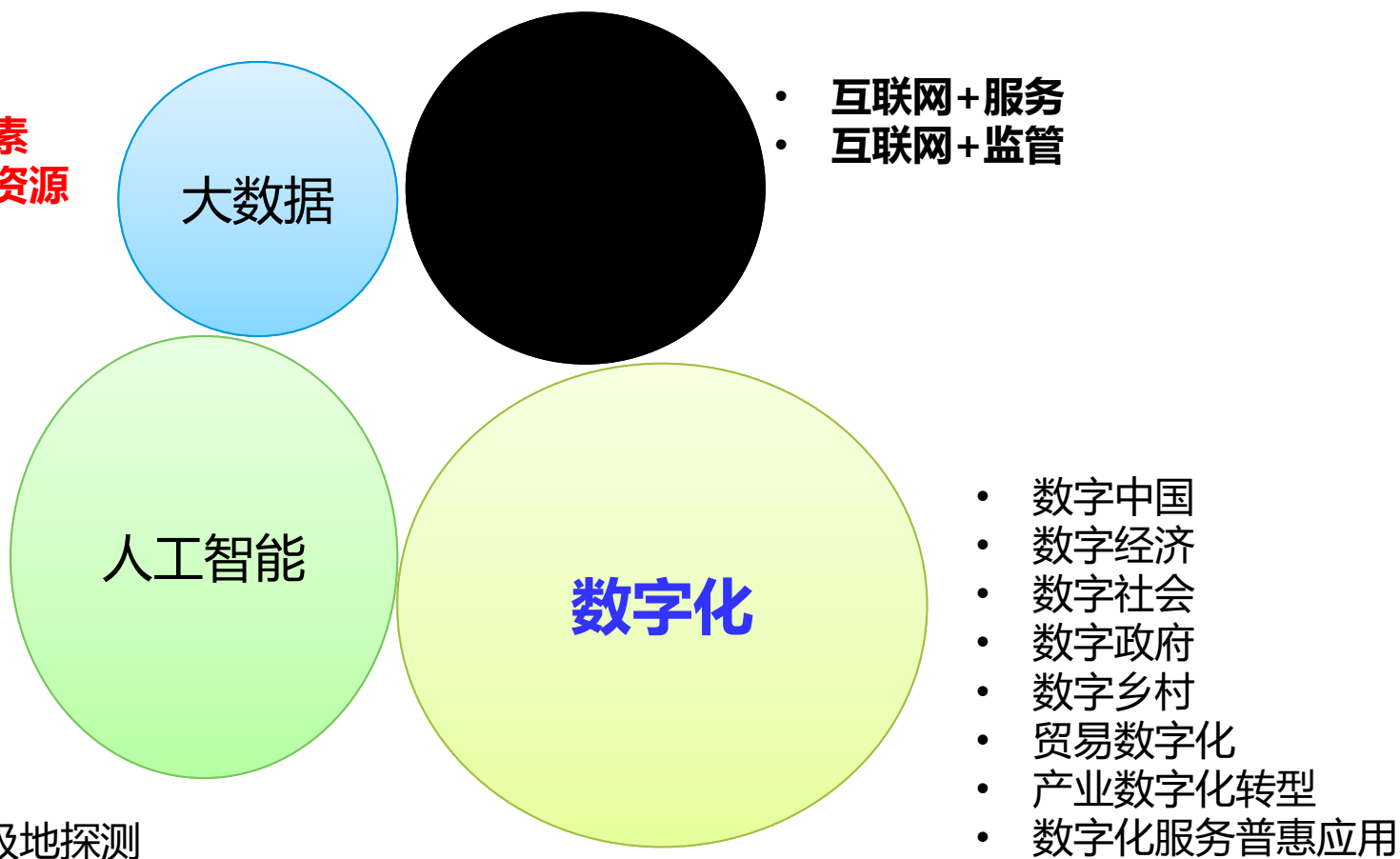
数据科学基础

37

□ 十四五规划中的“大数据”

关键

- **新时代生产要素**
- **国家基础战略资源**
- 大数据技术

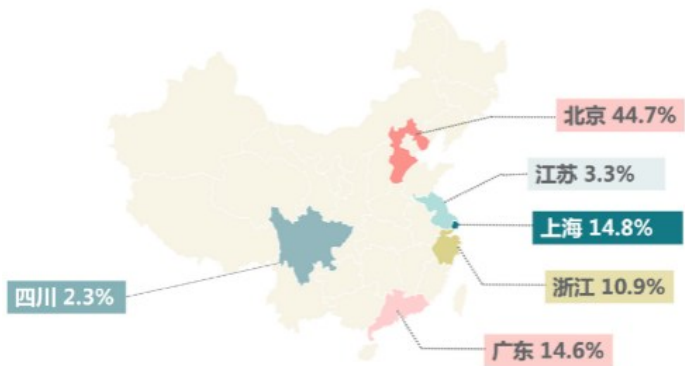


- 新一代人工智能
- 量子信息
- 集成电路
- 脑科学
- 基因生物技术
- 深空深地深海和极地探测

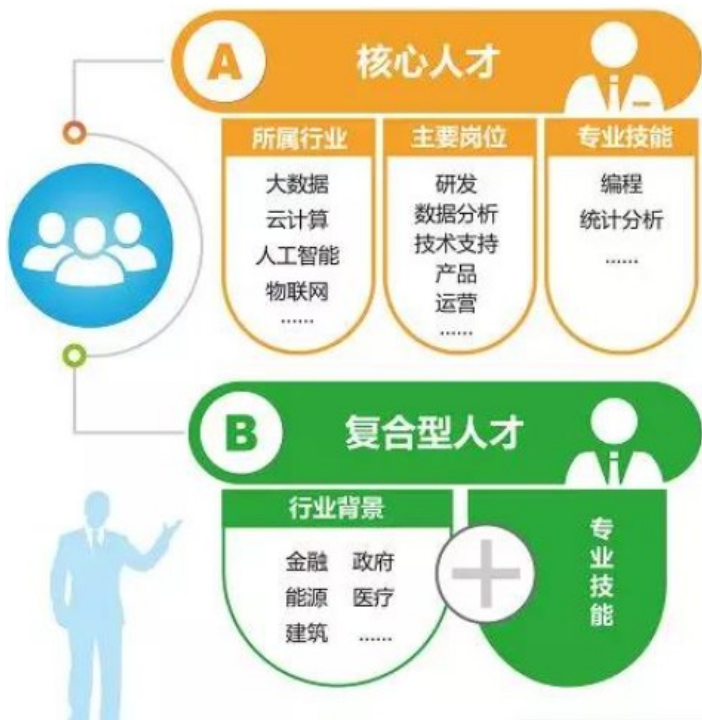
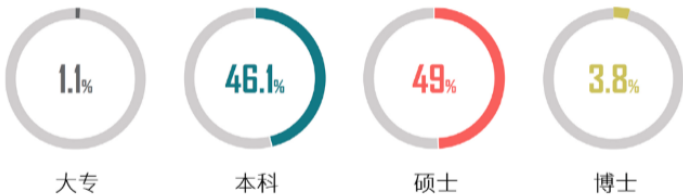


大数据人才缺口—市场需求

- 市场对大数据人才的需求日益增加，供求关系不成正比，2025年人才缺口达到230万
- 行业发展对大数据人才提出更高要求



公司对人才学历要求高，半数要求硕士及以上





数据科学基础

39

□ 大数据人才缺口—专业需求

- **传统的工科人才**培养模式难以满足新兴产业的发展要求，亟需探索大数据工科人才培养的新途径和新模式
- **大数据工科人才**培养面临的挑战

校内外都存在针对大数据技术的广泛需求，但**专业划分偏窄、偏细**，使得现有大数据相关的技术体系和积累较为零散，给跨学科交叉创新与技术应用带了诸多限制

新工科时代涌现出大量的新兴经济行业（如共享经济、数字经济），使得**传统的专业设置略显陈旧**，培养的人才知识体系不能有效应对新兴行业的大数据人才需求

很多工科学生的理论知识扎实，但实验实训、人文基础不够，创新性受限，**向专业数据科学研究人才的转化率较低**



数据科学基础

40

- 国内外优秀高校信息与传播学院开展“数据分析”等专业课程，以培养科教融合、协同育人人才培养模式的新模式

地区	学校	院系	课程	学分	类别	年级
中国	中国人民大学	新闻学院	统计数据分析	2	通识教育	大一、大二
中国	复旦大学	新闻学院	数据分析与信息可视化	3	专业基础课	大三
中国	山东大学	新闻传播学院	大数据推断基础（双语）	2	专业选修课	大一
中国	华中科技大学	新闻与信息传播学院	人工智能与数据科学	2	学科基础课	大二
中国	中国科学技术大学	计算机科学与技术学院	数据科学导论	2	核心通识	大二及以上
美国	University of Illinois at Urbana, Champaign	Journalism	Data Science Discovery	4	专业课	undergraduate
美国	Stanford University	Humanities & Sciences	Data Science 101	5	专业课	undergraduate
美国	The University of Chicago	Humanities	Data Analysis I: Introduction to Statistics	100-units	专业课	undergraduate
美国	Carnegie Mellon University	Humanities and Arts	Methods for Statistical & Data Science	9units	专业课	undergraduate
美国	University of Florida	Journalism and Communications	Advanced Data Journalism	3	专业课	undergraduate



数据科学基础

41

- 科技传播系致力于培养的“科技媒体和科技传播”英才应具备以下素质



专业基础扎实，有卓越的数理基础、创意设计、管理能力



实践能力强，具有处理多媒体大数据的能力



跨界能力强，能够解决新媒体行业的大数据应用问题

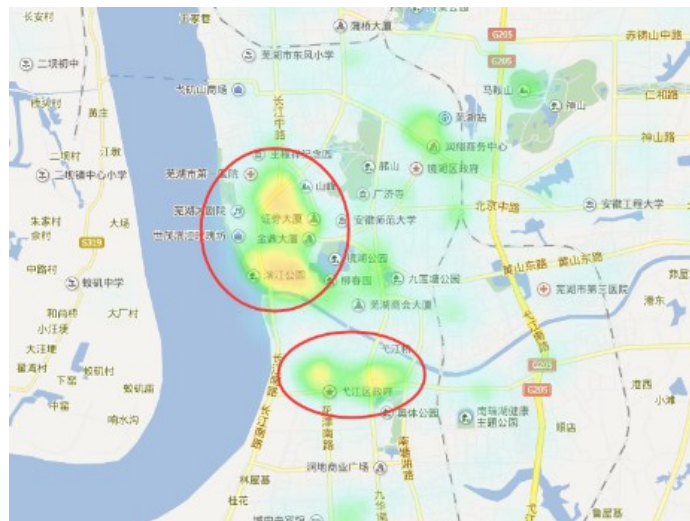
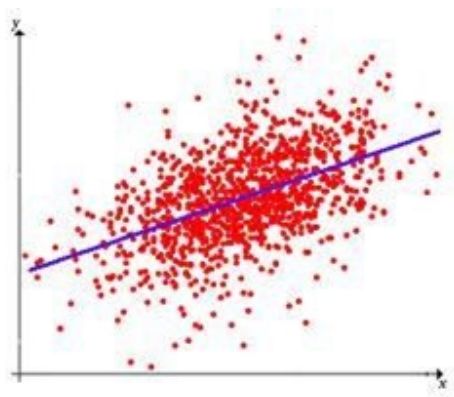
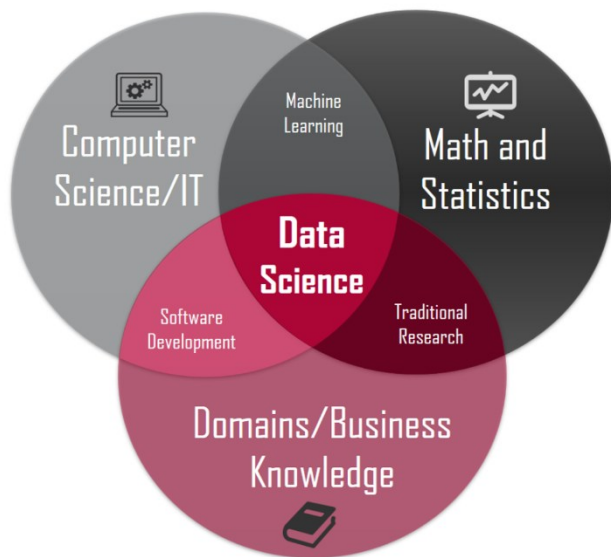


数据科学基础

42

□ 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

43

- 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

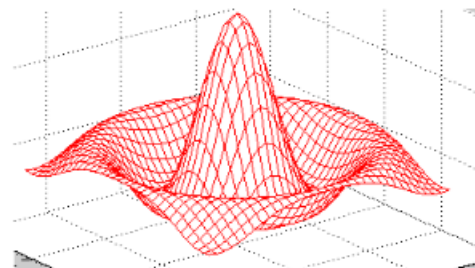
44

- 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



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
- . . .



数据科学基础

45

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

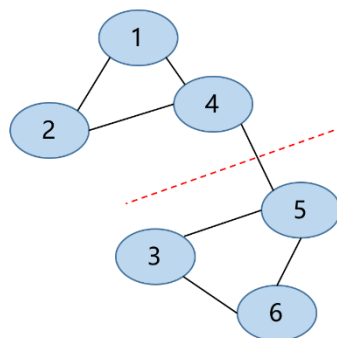
社交网络划分：物以类聚，人以群分



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

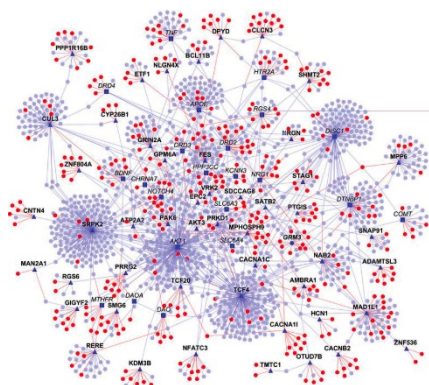


数据科学基础

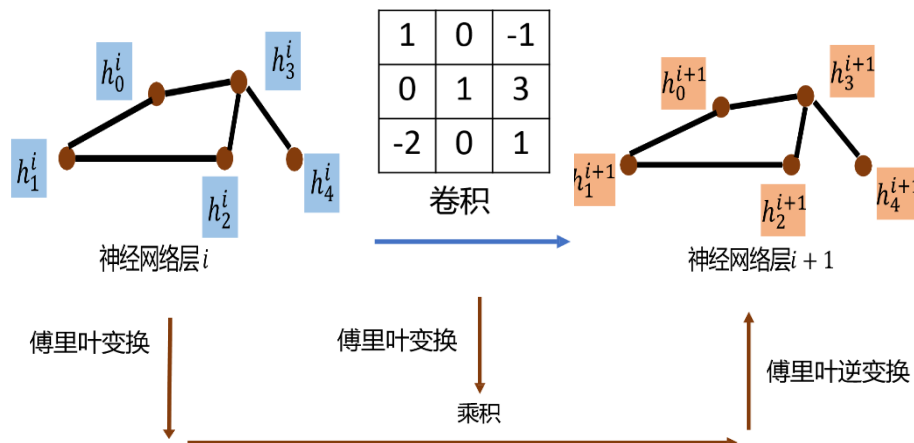
46

- 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
 - 数学与图卷积网络：傅里叶变换的应用

图数据：分子图、社交网络等



图卷积网络：一类典型方法



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

- 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

47

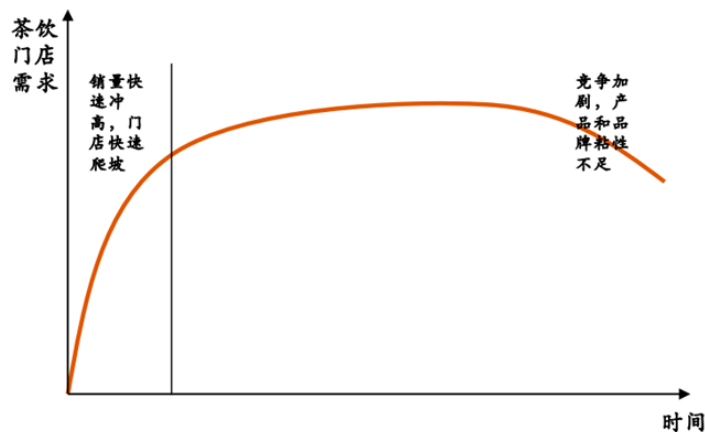
1. 创意设计与管理能力

- 运用数据分析结论帮助产品设计与营销管理

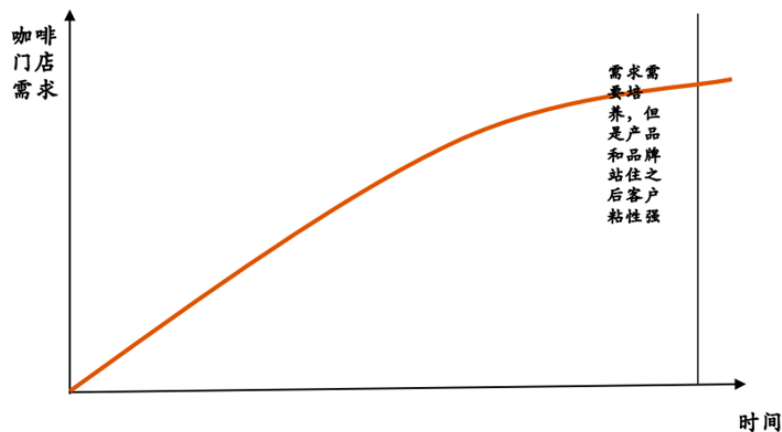
- 例：瑞幸咖啡

- 竞品分析，价格分析，用户分析，地理分析等
- 新媒体营销等

茶饮需求相对成熟，门店能够快速形成赚钱效应



咖啡需求需要培养，单店爬坡周期更长





数据科学基础

48

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**，可视化工具

```
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        G.append(i)
11        S.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        G.append(i)
23        S.append(i)
24        Sum += arr[S[0]] - arr[G[0]]
25    return Sum
26
```





数据科学基础

49

3. 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

50

改变这个世界的四种力量

暴力



知识



大数据



世界著名未来学家托夫勒
《第三次浪潮》作者



金钱



数据科学基础

51



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

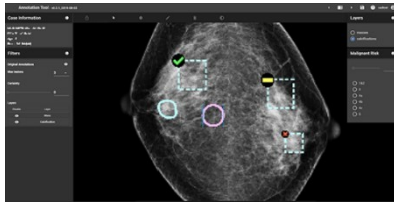


AI电子病历

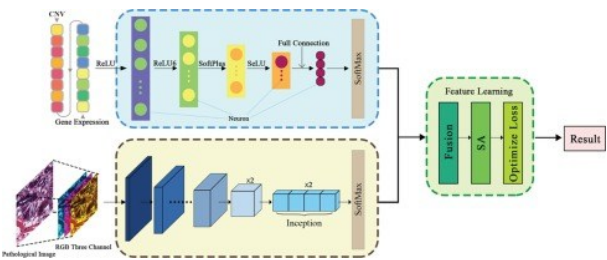


Input: EHR		Output: Diagnosis Results															
History of Present Illness: 3 days ago, Peter began to experience headache, cough, chest, non-productive cough, SOB, and chills. She says 3 people at work have been sick with headache, sore throat.... She was also reluctant to 150%, so was given Dilantin.... Physical Exam: ...Response: decreased air movement through... ...Diffuse end-expiratory... Major Surgical or Invasive Procedure: cardiac catheterization; intubation... Discharge Diagnosis: Pneumonia; Pneumonia; Pulmonary edema; Elevated cardiac enzymes; Hypertension; Atrial fibrillation.		 Clinical Diagnosis Model	<table><tr><th>Diagnosis Code</th><th>Diagnosis Description</th></tr><tr><td>486</td><td>Pneumonia Organism Unspecified</td></tr><tr><td>518.81</td><td>Acute Respiratory Failure</td></tr><tr><td>410.81</td><td>Unspecified Essential Hypertension</td></tr><tr><td>491.21</td><td>Obstructive Chronic Bronchitis with Acute Exacerbation</td></tr><tr><td>427.89</td><td>Other Specified Cardiac</td></tr><tr><td>...</td><td>...</td></tr></table>	Diagnosis Code	Diagnosis Description	486	Pneumonia Organism Unspecified	518.81	Acute Respiratory Failure	410.81	Unspecified Essential Hypertension	491.21	Obstructive Chronic Bronchitis with Acute Exacerbation	427.89	Other Specified Cardiac	...	...
Diagnosis Code	Diagnosis Description																
486	Pneumonia Organism Unspecified																
518.81	Acute Respiratory Failure																
410.81	Unspecified Essential Hypertension																
491.21	Obstructive Chronic Bronchitis with Acute Exacerbation																
427.89	Other Specified Cardiac																
...	...																

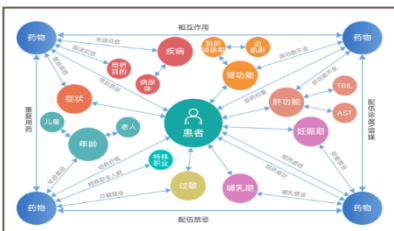
疾病诊断



病灶识别



多模态医疗数据挖掘模型



药物推荐



数据科学基础

52

- 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

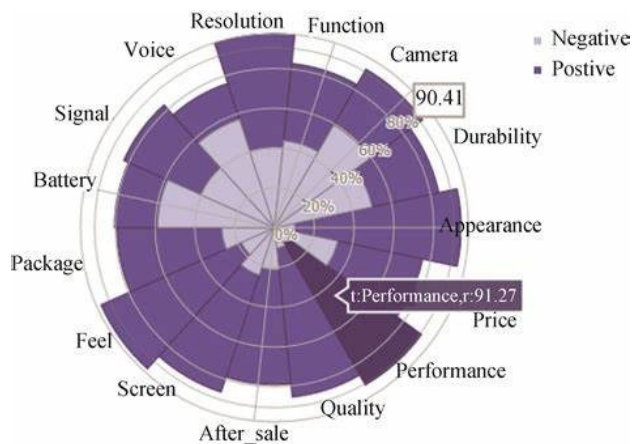


数据科学基础

53

数据蕴含着巨大的价值—安防领域

舆情监测



舆情情感分析



传播途径监测



数据科学基础

54



数据蕴含着巨大的价值——智慧教育
因材施教

学生的学习行为数据

1	[A]	[C]	[D]
2	[A]	[B]	[D]
3	[B]	[C]	[D]
4	[A]	[B]	[C]
5	[A]	[B]	[D]

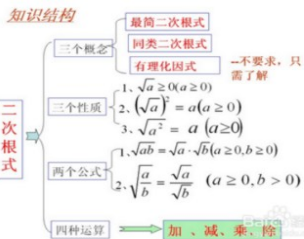


大数据
分析

学生认知水
平画像

试题难度等
特征的预测

个性化学习
推荐

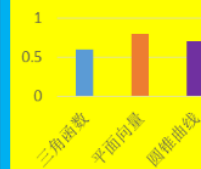


试题-知识点

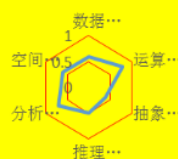


姓名	张三
学号	9527
平均正确率	85%
综合水平	90.562

考点掌握情况



能力分布图谱



$$9 - 3 \div \frac{1}{3} + 1 = ?$$

$$\frac{4}{7} \div 8 = ?$$

易





数据科学基础

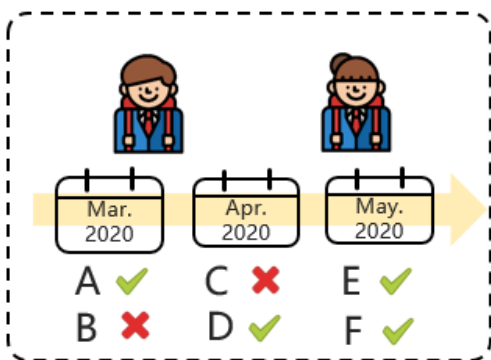
55



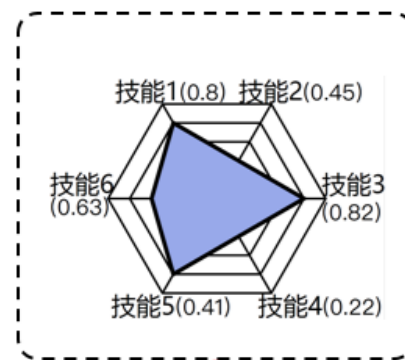
数据蕴含着巨大的价值—智慧教育

- 学习者能力分析
- 量表范式 到 数据驱动的模式

学习过程相关数据



知识能力及其发展变化



反馈调整教学策略



数据科学基础

56



数据蕴含着巨大的价值—智慧教育

■ 考试试题质量与公平性



浙江省教育考试院
ZHEJIANG EDUCATION EXAMINATIONS AUTHORITY

组织机构 信息公开 政策法规 政策解读 2018年11月27日 星期二 11:11:39 请输入关键字

普通高考 | 学考选考 | 研究生考试 | 成人高考 | 自学考试 | 社会考试 | 教师资格考试 | 海外考试

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



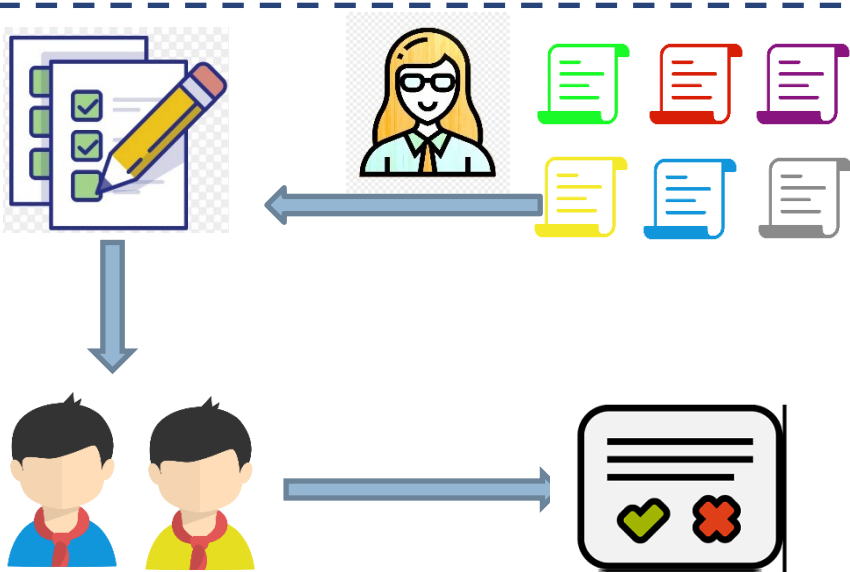
数据科学基础

57

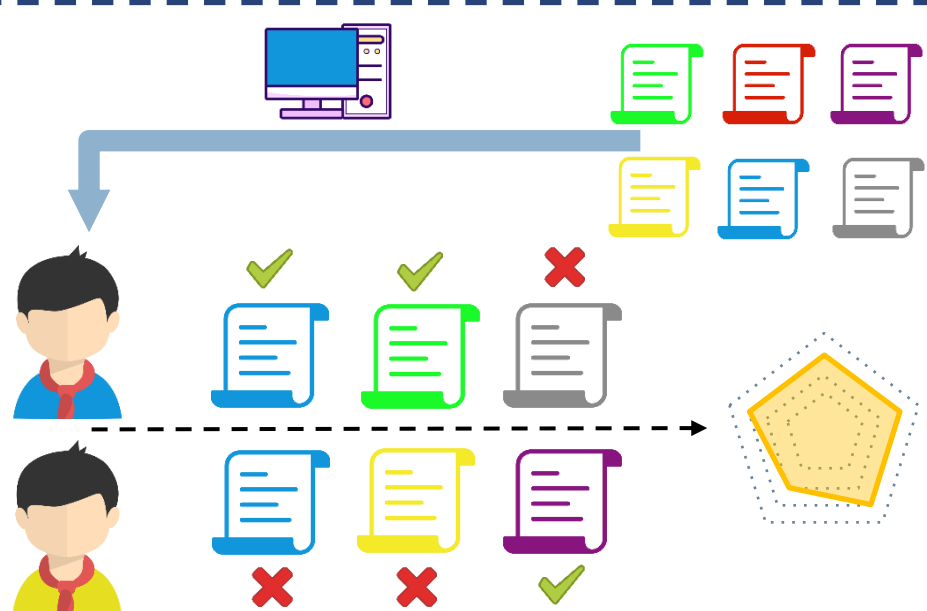
- 数据蕴含着巨大的价值——智慧教育
- 数据驱动的自适应测试 (CAT)

GRE

纸笔考试 ← 考试 → 计算机自适应测试



Selecting a uniform set of questions



Selecting a tailored sequence of questions



数据科学基础

58

?

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

59

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。



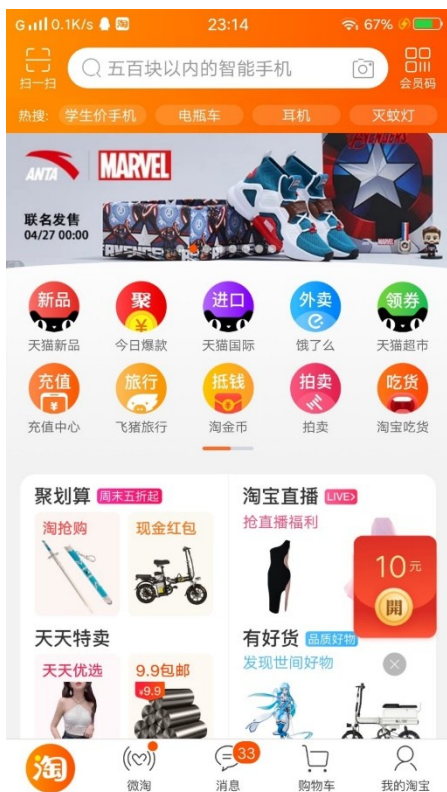
有一种说法称，特朗普的团队聘用数据分析公司，做了精准的广告投放，影响了那些徘徊不定的选民，拿下了决定性的关键州选举人票



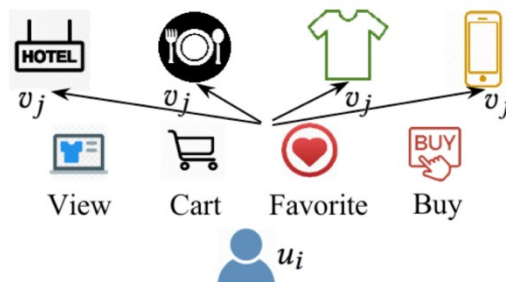


数据科学基础

数据蕴含着巨大的价值—电子商务 计算广告



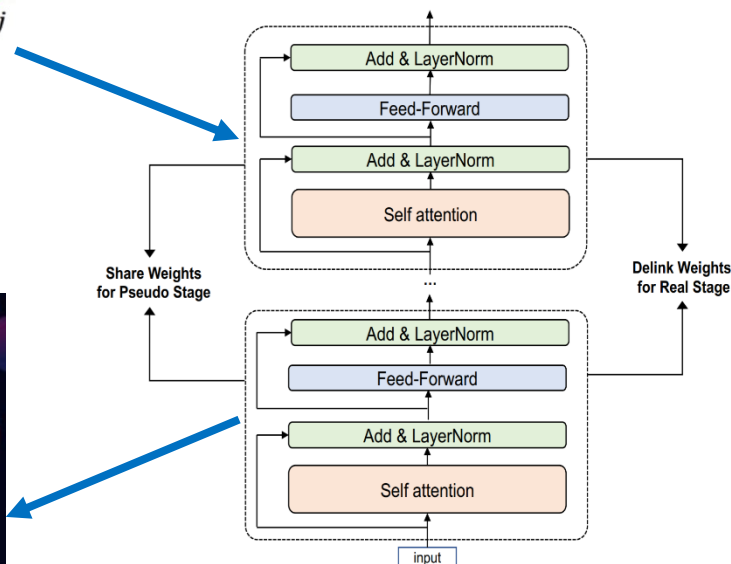
电商平台



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



数据科学基础

- 数据蕴含着巨大的价值—电子商务
 - 精准搜索、个性化消费推荐



天猫双11：破亿交易时间

年份	10亿	100亿	500亿
2014	3分	38分28秒	21时12分
2015	1分12秒	12分28秒	9时52分22秒
2016	52秒	6分58秒	2时30分20秒
2017	28秒	3分01秒	40分12秒
2018	21秒	2分05秒	26分03秒
2019	14秒	1分36秒	12分49秒



数据科学基础

62



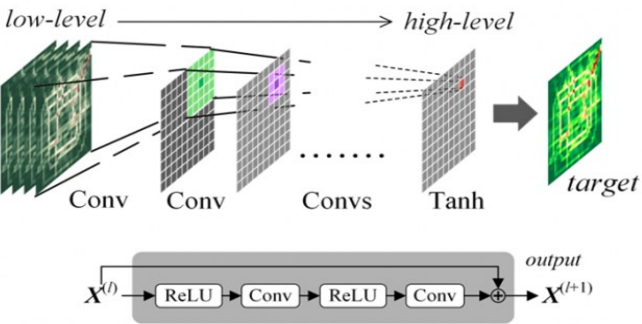
数据蕴含着巨大的价值 — 智慧城市

- 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



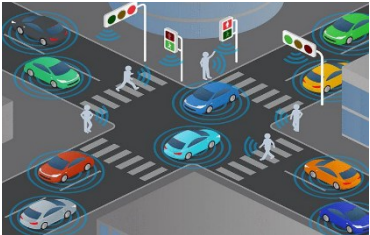
交通
流预
测模
型



行程规划



信号灯调控



道路规划





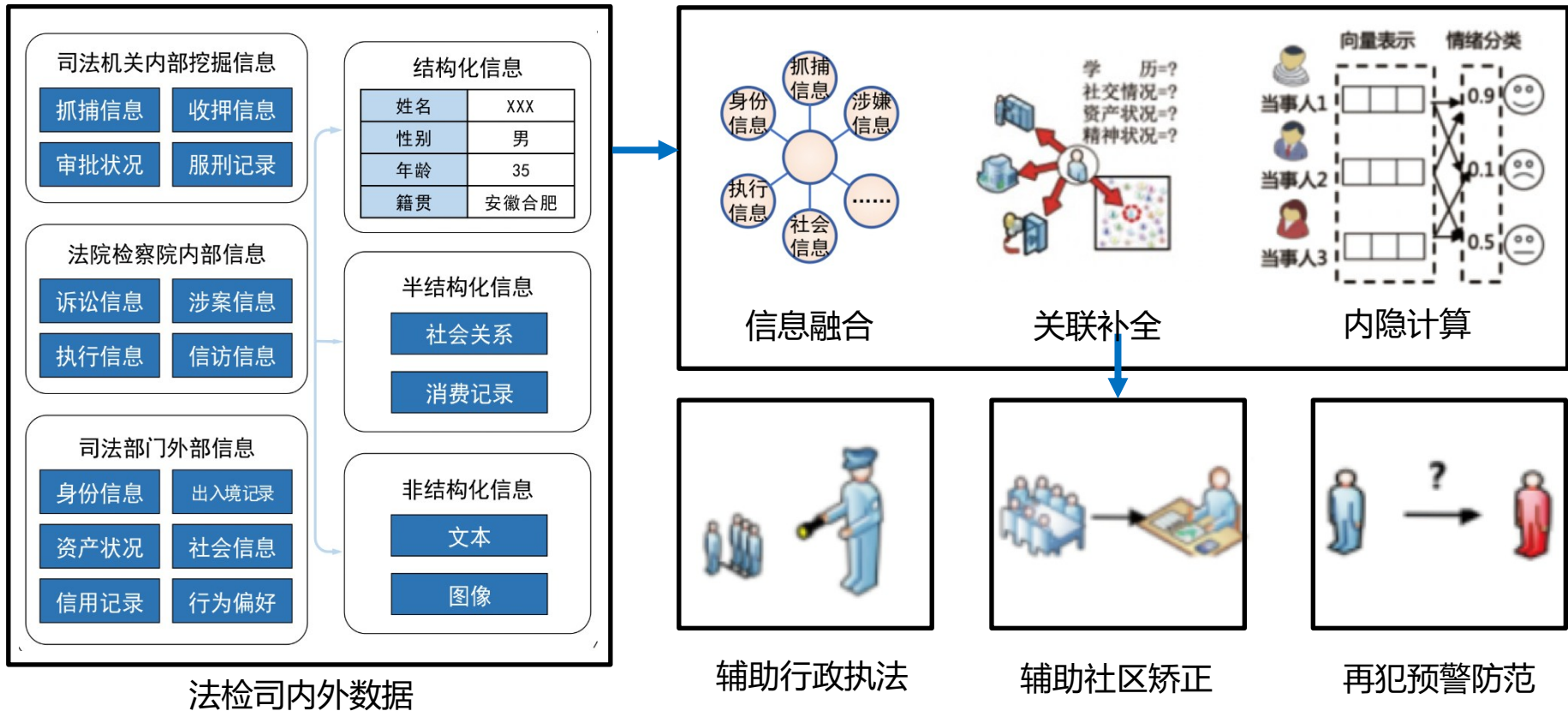
数据科学基础

63



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部数据与外部数据等
- 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**





数据科学基础

64



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



数据科学基础

65

数据蕴含着巨大的价值——文化娱乐 ■ 机器作诗



用户写作意图



江、船、秋风



春风、杨柳



松、竹、山、牧童



诗歌自动生成系统



对应诗词

江北江南万顷秋，
船头人去水悠悠。
一帆一棹秋风急，
又有离人万里愁。

杨柳千条拂地垂，
一川春水浸桃花。
游人不识湖中路，
游遍人间野水涯。

乔松古木两三间，
松竹阴中一径斜。
白鸟不知山路远，
牧童踏过野人家。



数据科学基础

66

□ 机器作诗 PK 古代诗人



哪首诗是人写的？

秋夕湖上

一夜秋凉雨湿衣，
西窗独坐对夕晖。
湖波荡漾千山色，
山鸟徘徊万籁微。

机器

秋夕湖上

荻花风里桂花浮，
恨竹生云翠欲流。
谁拂半湖新镜面，
飞来烟雨暮天愁。

宋代诗人葛绍体



数据科学基础

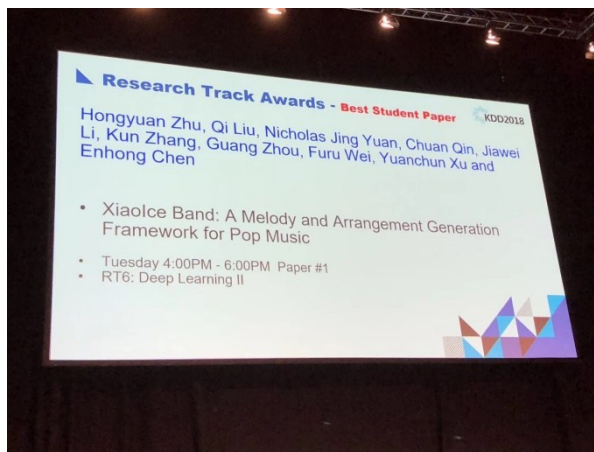
67



数据蕴含着巨大的价值—文化娱乐

- 流行音乐的旋律与编曲生成
- 机智过人：

<http://tv.cctv.com/2017/11/24/VIDEo7JWp0u0oWRmPbM4uC Bt171124.shtml>



【KDD18最佳论文揭晓】中科大等斩获最佳学生论文，刘兵获创新奖，清华大学唐杰任副主席

● 首页 ● 新闻博览

我校获数据挖掘领域顶级国际会议KDD 2018最佳学生论文奖



数据科学基础

68

数据蕴含着巨大的价值—文化娱乐

流行音乐的旋律与编曲生成

$\text{♩} = 90$

长笛

鼓组

无品电贝司

乐队小提琴

古典吉他

4

Fl.

D. Set

Frd. El. B.

Vlns.

Guit.

ApowerREC

8

Fl.

D. Set

Frd. El. B.

Vlns.

Guit.

12

Fl.

D. Set

Frd. El. B.

Vlns.

Guit.



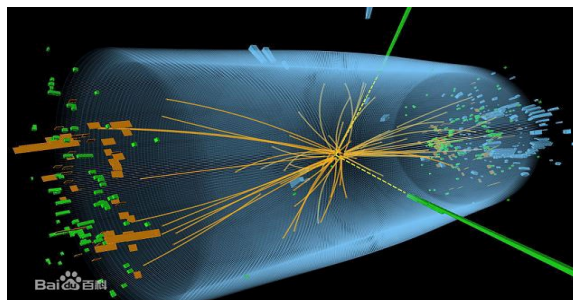
数据科学基础

69

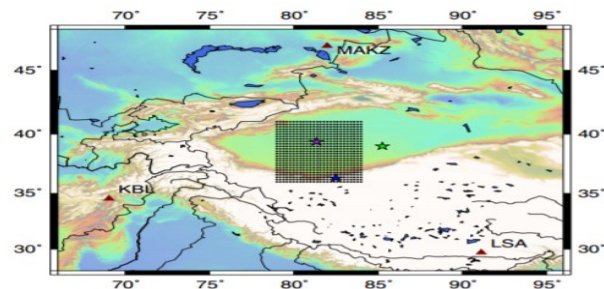
- ❑ 数据蕴含着巨大的价值—科学技术
- ◆ 大数据推动科学新技术发现



天文大数据搜索新星



物理大数据预测分子属性



大数据地震速报、余震预测



生物大数据改良基因

科学
大数据



数据挖掘保护知识产权



数据科学基础

70

- **科技大数据来自于物理世界**
 - 科学实验数据或传感数据
 - 技术描述型数据—专利、论文
- **集多种特点于一身**
 - 采集的高代价性
 - 复杂性
 - 超高维度
 - 高度计算复杂性
 - 高度的不确定性
 - 学科知识壁垒
 - 信息与通信技术高度集成性

单一学科



数据驱动

多学科交叉



关系型数据库的鼻祖Jim Gray（右）





数据科学基础

71

■ 2007年，Jim Gray总结出了四个科学范式





数据科学基础

72

□ 新媒体

- 利用数字技术、网络技术和移动通信技术，通过互联网、宽带局域网、无线通信网和卫星等渠道，以电视、电脑和手机为主要输出终端，向用户提供视频、音频、语音数据服务、连线游戏、远程教育等集成信息和娱乐服务的所有新的传播手段或传播形式的总称，包括“新兴媒体”，也包括“新型媒体”

□ 新媒体大数据

- 新媒体服务场景中收集获取的数据
- 按数据的模态类型：语音，图片，视频，网络，文本



数据科学基础

73

□ 多媒体大数据研究热点—语音大数据

□ 语音识别

- 微软英语语音识别实现词错率5.9%的突破，第一次超越人类
- 科大讯飞语音识别词错率3%左右。中文领域突出





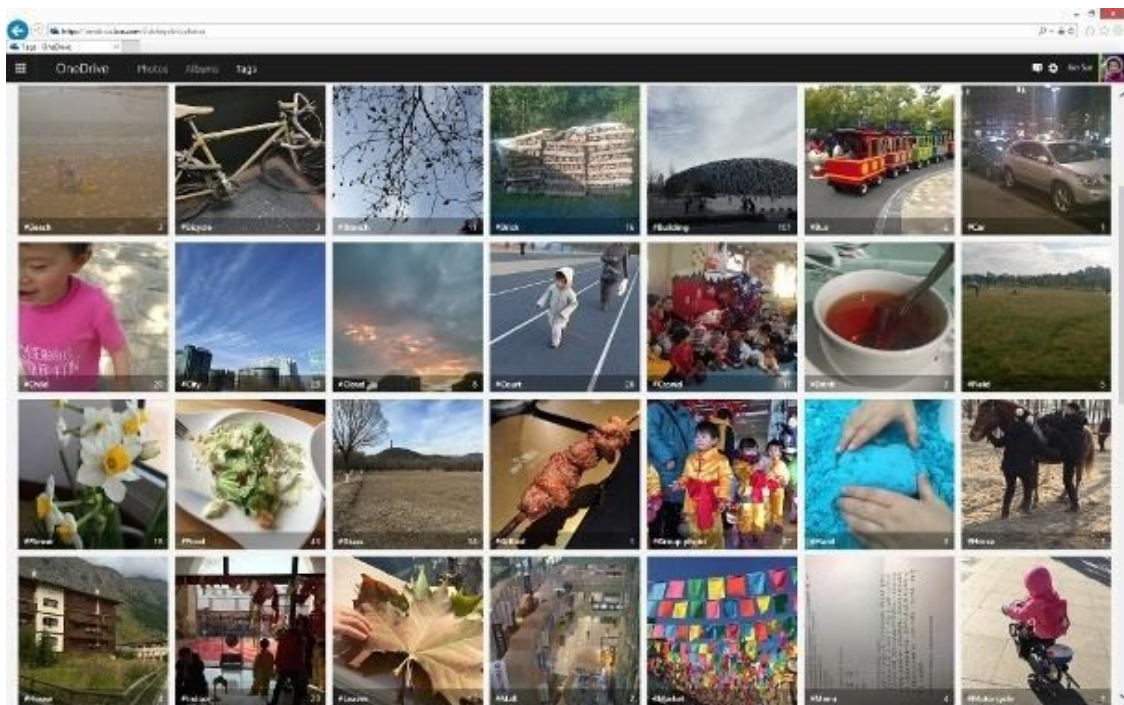
数据科学基础

74

□ 多媒体大数据研究热点—图片大数据

□ 图像识别

- ImageNet图像数据库上，人工智能已达到2.99%的错误率（公安部三所），低于人类5.1%的错误率



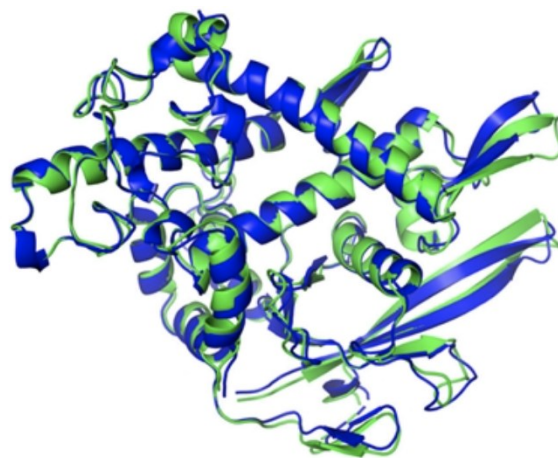
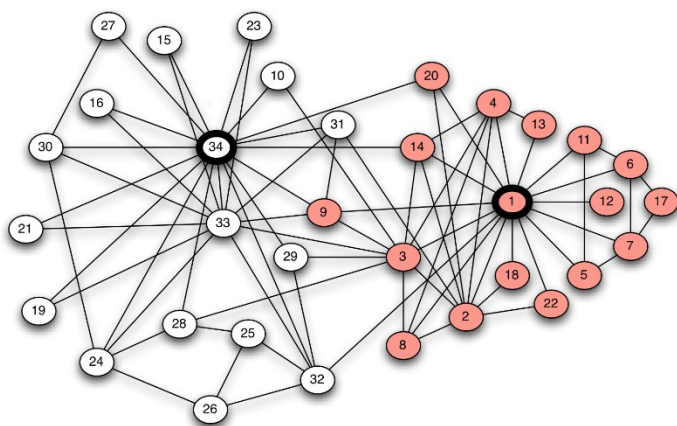
李飞飞
斯坦福大学、谷歌Ai前任首席科学家



数据科学基础

75

- 多媒体大数据研究热点—网络大数据
 - 社会网络分析
 - 蛋白质结构图
 - 知识图谱

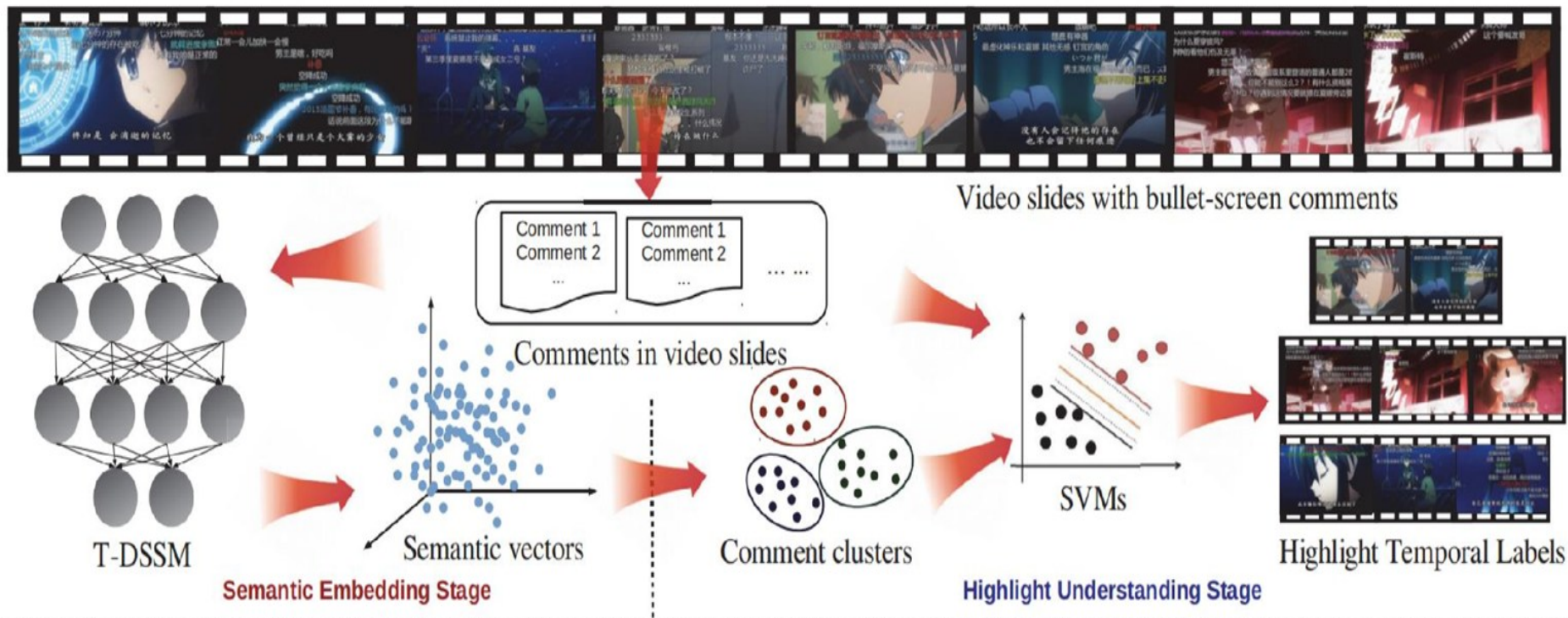




数据科学基础

76

- 多媒体大数据研究热点—视频大数据
 - 社会网络分析





数据科学基础

77

- 多媒体大数据研究热点—文本大数据
 - 自然语言处理
 - 机器阅读，机器翻译，文本推理，知识图谱等

通用语言理解评估（GLUE）基准

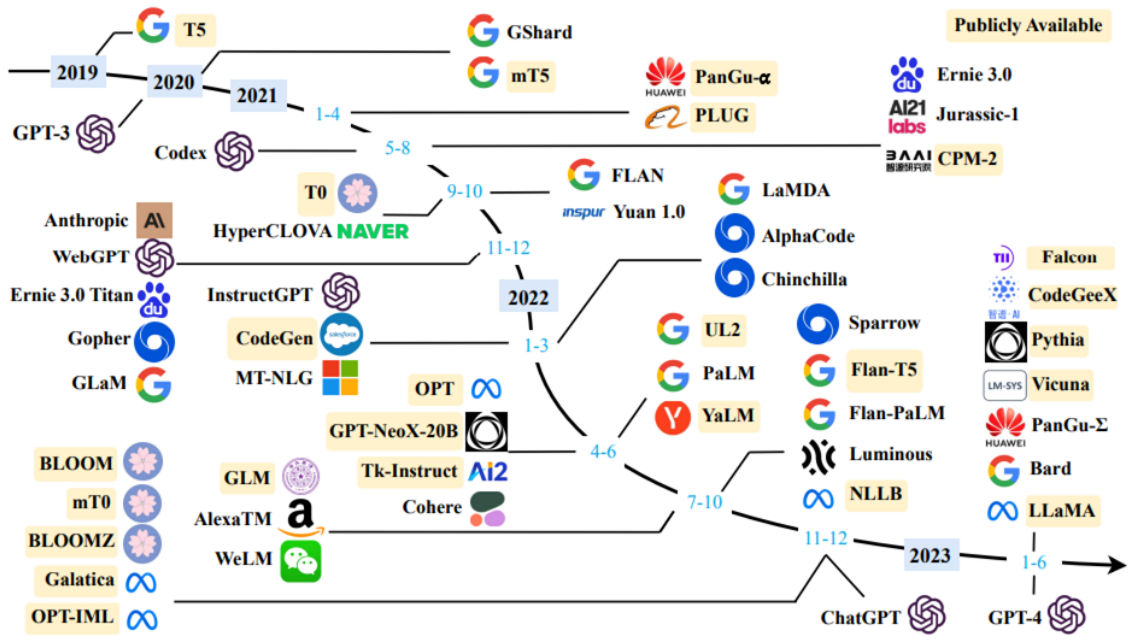
Rank Name		Model		URL	Score	CoLA	SST-2	MRPC	STS-B
+	1	Alibaba DAMO NLP	StructBERT		90.3	75.3	97.1	93.9/91.9	93.0/92.5
	2	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8
	3	ERNIE Team - Baidu	ERNIE		90.1	72.8	97.5	93.2/91.0	92.9/92.5
	4	Microsoft D365 AI & MSR AI & GATECHMT-DNN-SMART			89.9	69.5	97.5	93.7/91.6	92.9/92.5
+	5	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5
+	6	Huawei Noah's Ark Lab	NEZHA-Large		88.7	67.4	97.2	93.2/91.0	92.2/91.6
+	7	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1
	8	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0
	9	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9
+	10	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7
	11	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6



数据科学基础

78

- 多媒体大数据研究热点—文本大数据
 - 大语言模型——多项技术融合
 - 如今大语言模型层出不穷，例如 ChatGPT, ChatGLM, LLaMa等，具备大量知识，能够以类人的方式和人类进行多轮对话，在多种自然语言理解相关任务上表现良好。



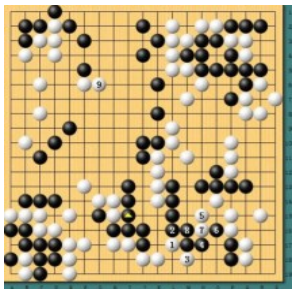


数据科学基础

□ 大数据的研究—数据与知识融合，让人工智能更“聪明”

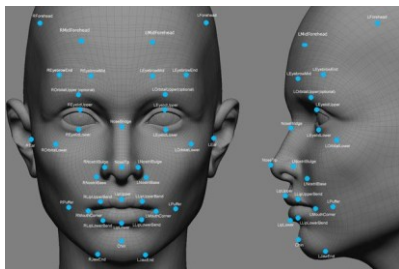
计算智能

- 规则明确、特定领域



感知智能

- 语音、图像、视频



认知智能

- 理解、推理、解释





数据科学基础

□ 大数据的研究—从数据中的相关性到世界的因果推断

逻辑关系

- 归纳法、数理逻辑、布尔代数系统

$$(a \vee b) \vee c = a \vee (b \vee c)$$
$$(a \wedge b) \wedge c = a \wedge (b \wedge c)$$

重推理

相关关系

- 贝叶斯网络、机器学习、深度学习



重分析
(学习)

因果关系

- 因果关系是有方向的、存在时序先后性

万有引力



数据分析+
逻辑推理



数据科学基础

➤ L1: 基础对话者

- 对话与交互能力：当前(如GPT-4),能流畅的对话
- 特点：以对话为核心，提供信息、解答问题、辅助创作等

➤ L2: 推理者

- 推理与问题解决能力：OpenAI-O1接近L2。PhD-Level。
- 特点：像人类一样分析问题、制定方案，甚至在某些领域超越人类

➤ L3: 智能行动者

- 自主决策能力：AI能独立思考并根据复杂情境采取行动(接近人类)
- 特点：不仅能够思考，还能在真实世界中执行决策，实现人机协同

➤ L4: 创新者

- 创造能力：AI能进行创造性思维和协助人类进行发明和创造，推动科技进步
- 特点：具备创新思维和创造力，能够提出新的想法和解决方案

➤ L5: 组织者

- AGI: AI能执行和组织人类所有工作，标志着真正的人工通用智能的实现
- 特点：具备人类智能和组织能力，能管理和优化整个社会的运行。

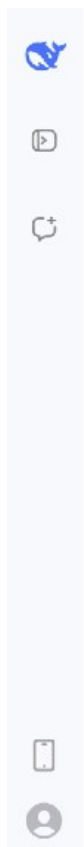
OpenAI's 5 Step to AGI

- | | |
|---------|---|
| Level 1 | Chatbots, AI with conversational language |
| Level 2 | Reasoners, human-level problem solving |
| Level 3 | Agents, systems that can take actions |
| Level 4 | Innovators, AI that can aid in invention |
| Level 5 | Organizations, AI that can do the work of an organization |



数据科学基础

□ 大数据的研究—推理大模型



我是 DeepSeek，很高兴见到你！

我可以帮你写代码、读文件、写作各种创意内容，请把任务交给我吧~

给 DeepSeek 发送消息



DeepSeek



联网搜索

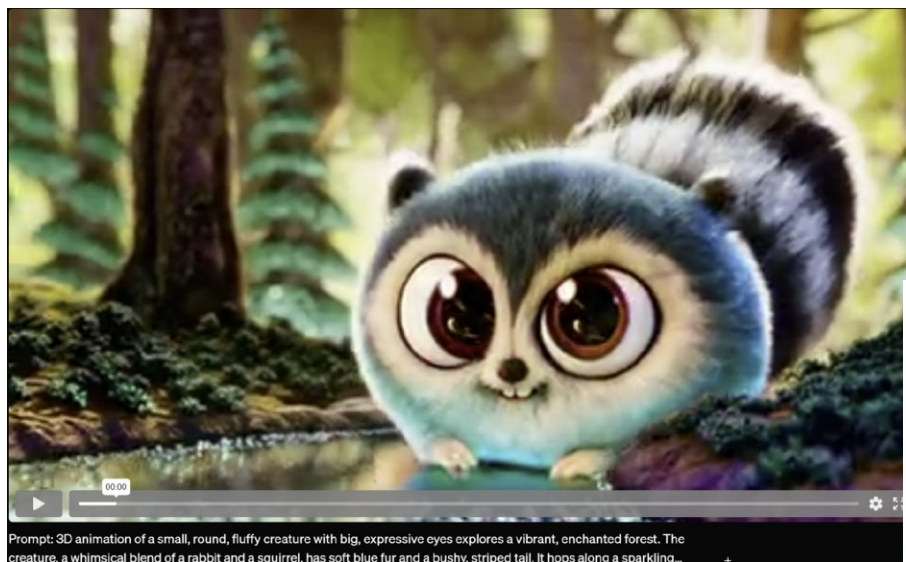




数据科学基础

□ 大数据的研究—多模态大模型

- SORA: <https://openai.com/index/sora/>
- 智象未来: <https://hidreamai.com/home>





数据科学基础

84

□ 大数据与人工智能

- ABC当前AI的技术体系

Big data

数据

大数据是人工智能发展的**基石**，人工智能的核心在于数据支持。

机器学习算法是人工智能的**核心**，是今天引领人工智能发展潮流的一大类算法

算法

Algorithm

AI

算力

Computation

人工智能算法的实现需要强大的计算能力**支撑**，特别是深度学习算法的大规模使用，对计算能力提出了更高的要求。



数据科学基础

85

□ 大数据与人工智能

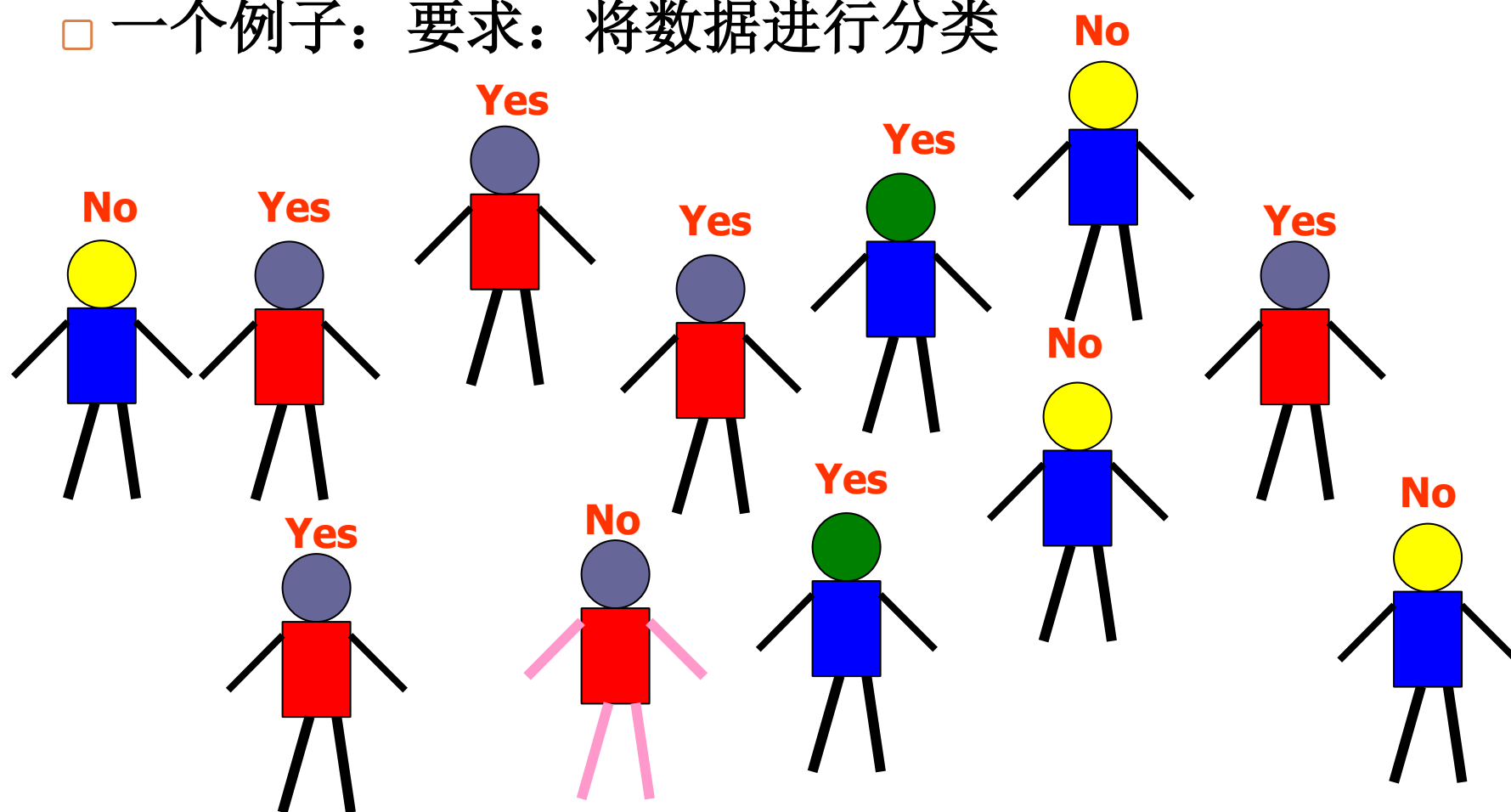
□ 现阶段，人工智能的核心是对大数据进行的**特征抽取**与**机器学习算法**





数据+分类学习的方法

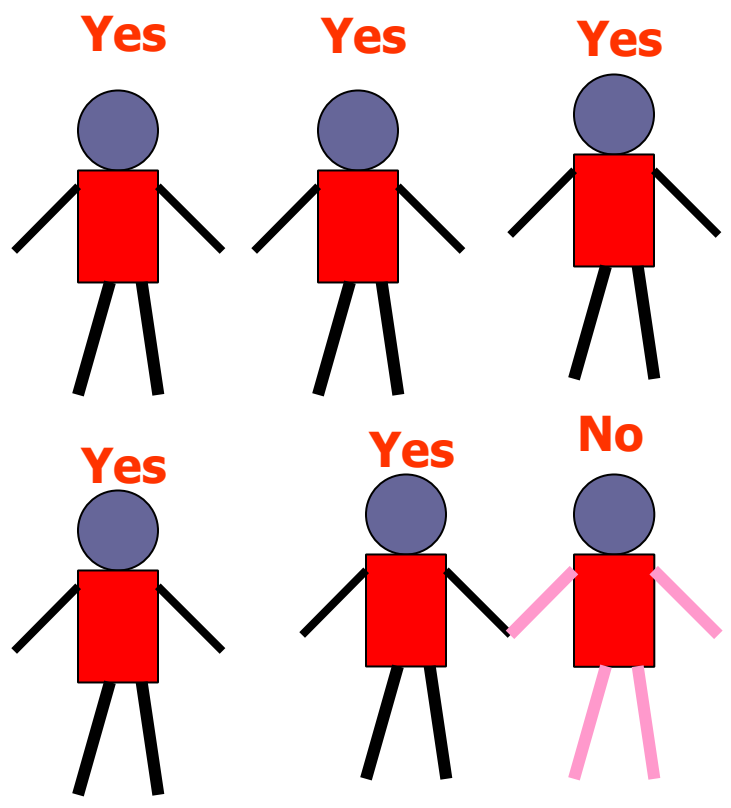
□ 一个例子：要求：将数据进行分类



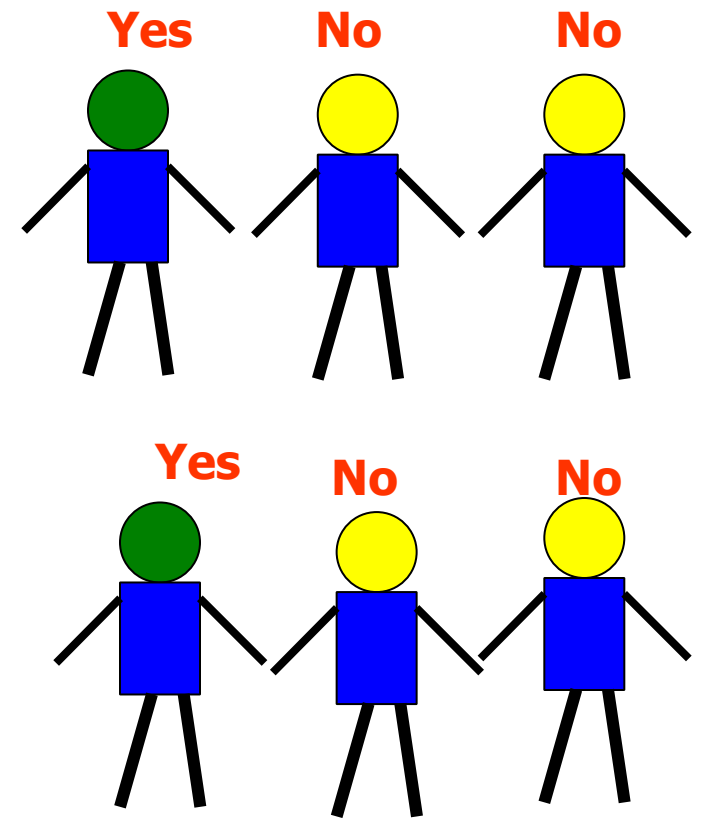


数据+分类学习的方法

躯干：红色



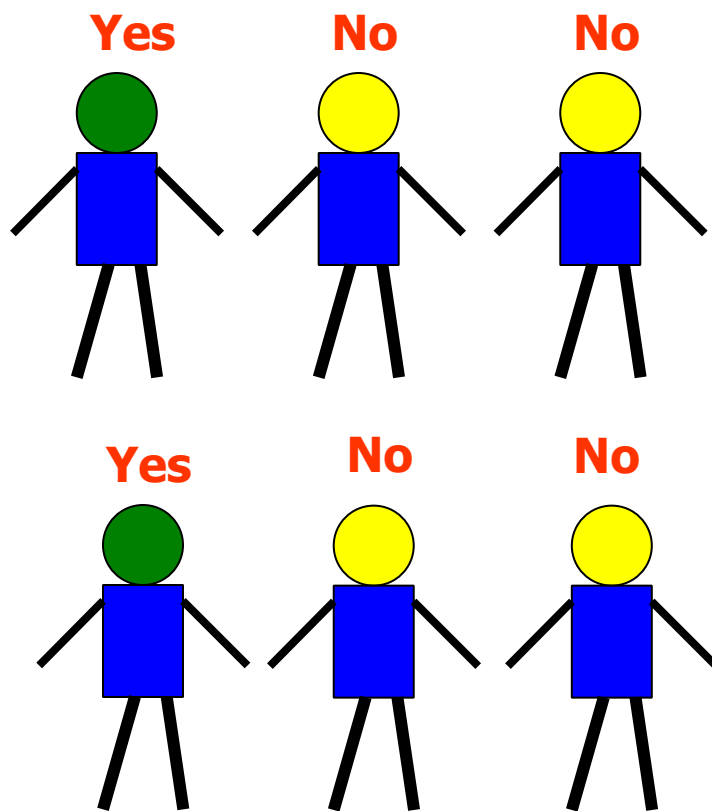
躯干：蓝色





数据+分类学习的方法

躯干：蓝色



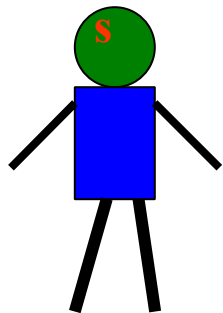


数据+分类学习的方法

躯干：蓝色

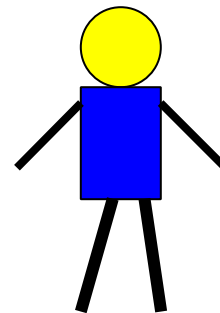
头：绿色

Ye

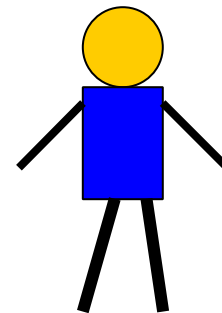


头：黄色

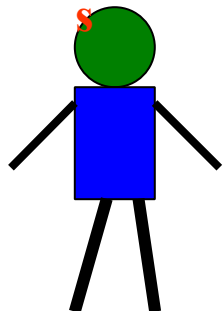
No



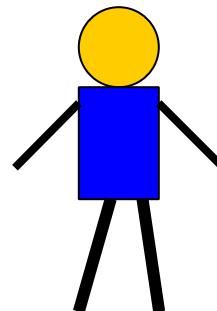
No



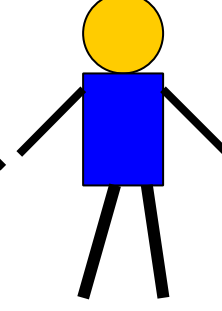
Ye



No



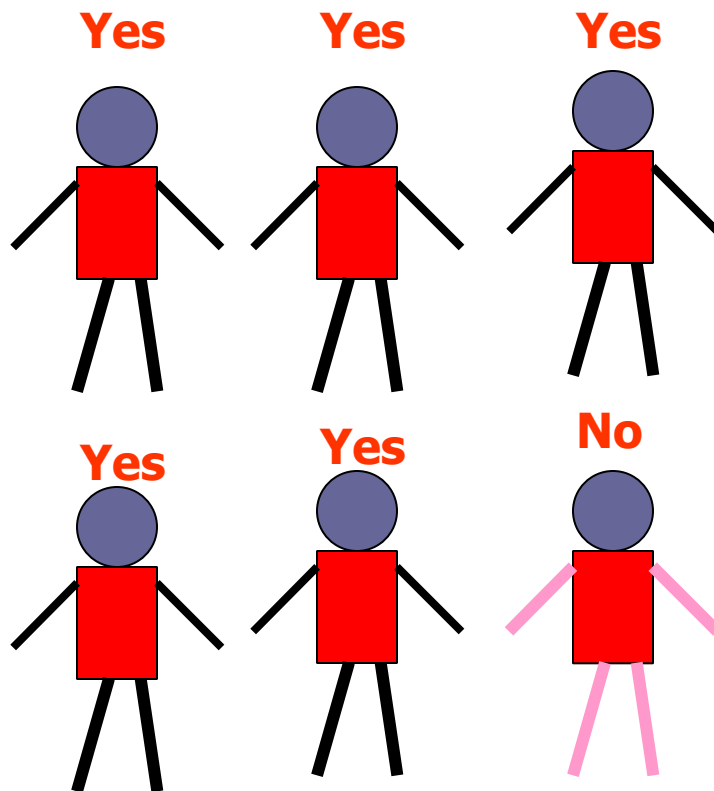
No





数据+分类学习的方法

躯干：红色



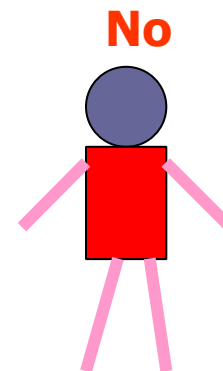
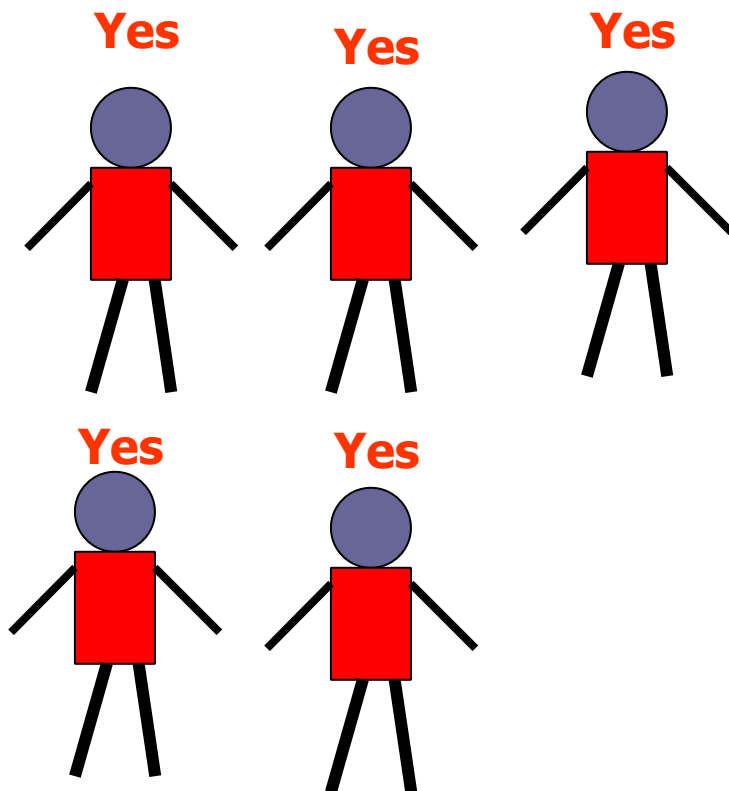


数据+分类学习的方法

躯干：红色

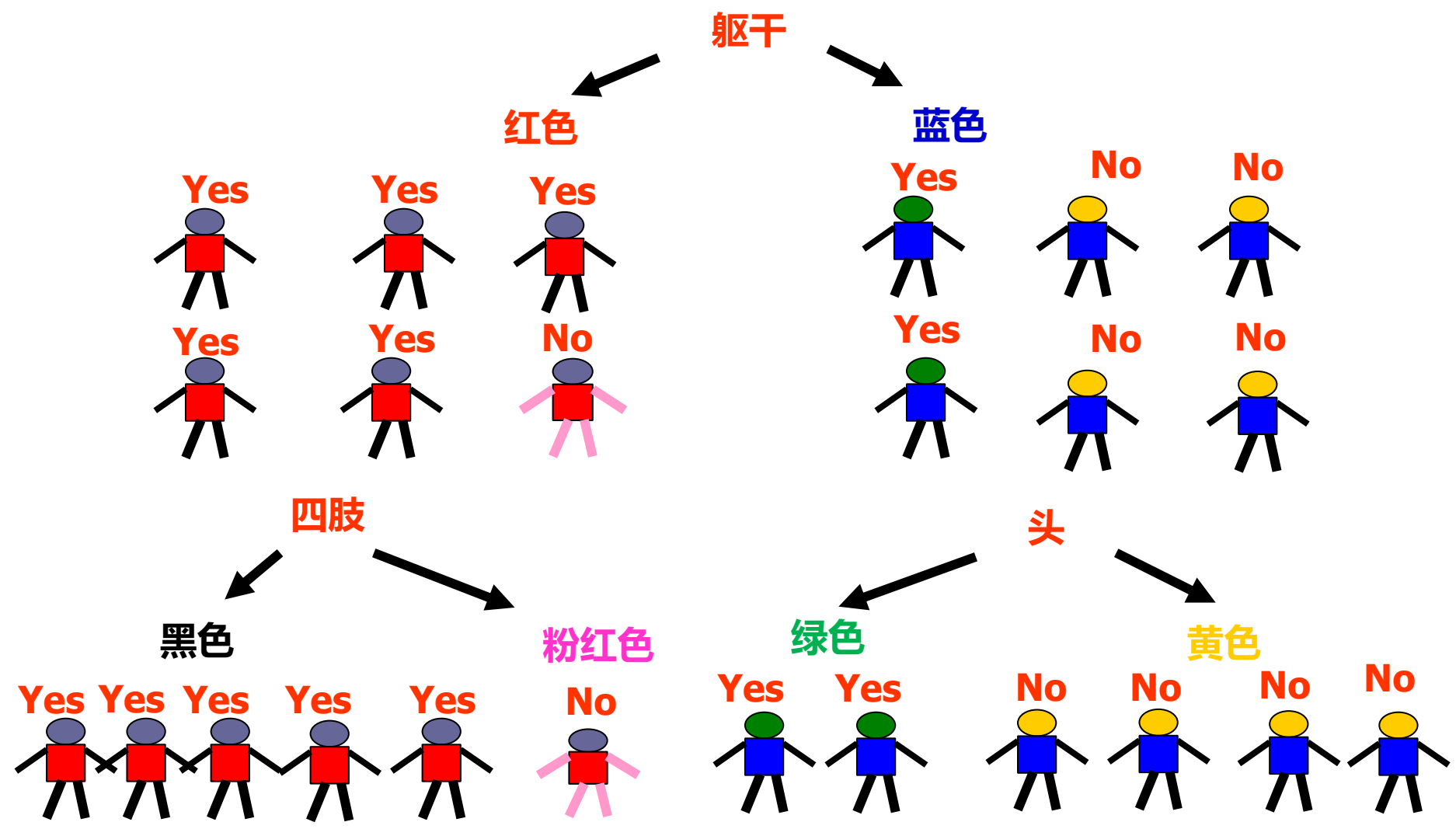
四肢：黑色

四肢：粉红色





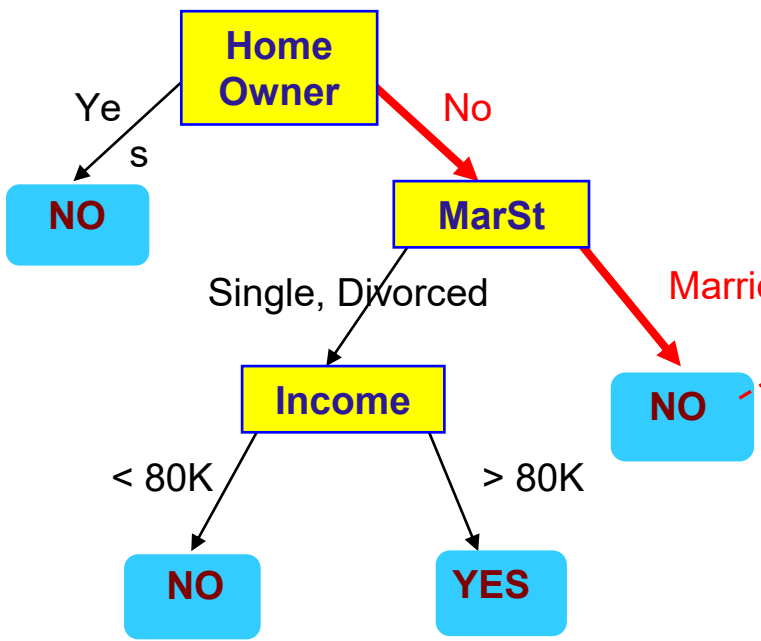
数据+分类学习的方法





数据+分类学习的方法

□ 决策树——使用模型对测试数据分类



Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

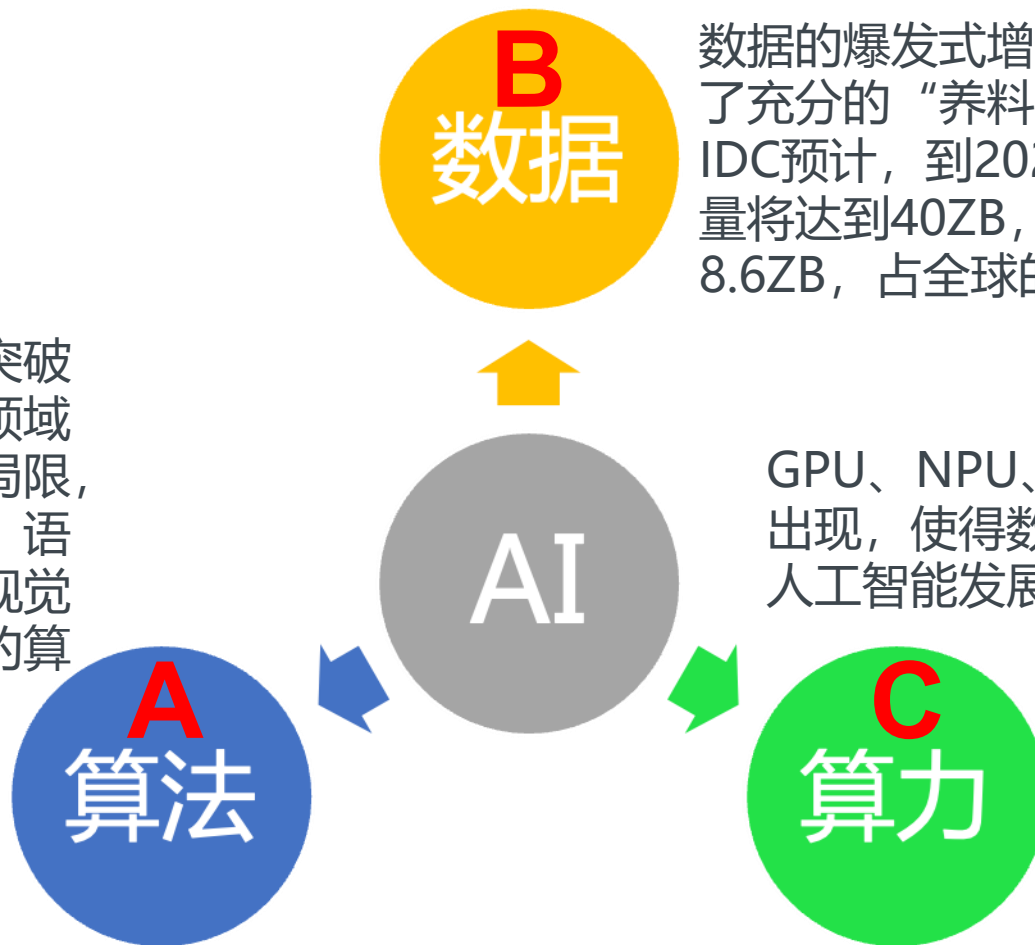
Assign Defaulted to "No"



数据科学基础

□ 大数据的研究—数据驱动人工智能成熟与商业化

深度学习的出现突破了过去机器学习领域浅层学习算法的局限，颠覆了语音识别、语义理解、计算机视觉等基础应用领域的算法设计思路



数据的爆发式增长为人工智能提供了充分的“养料”，市场调研机构IDC预计，到2020年，全球数据总量将达到40ZB，我国数据量将达到8.6ZB，占全球的21%左右。

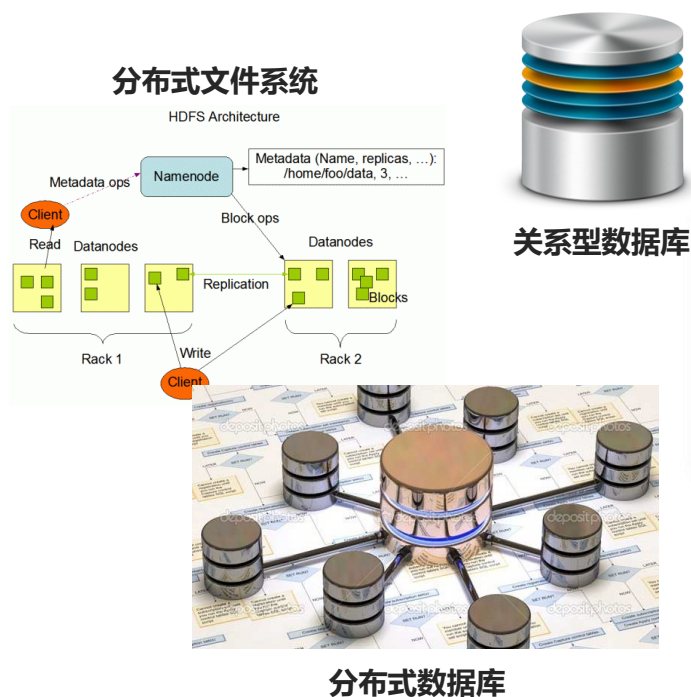
GPU、NPU、FPGA等专用芯片的出现，使得数据处理速度不再成为人工智能发展的瓶颈



数据科学基础

95

- 包括高效的CPU/GPU、云计算、AI芯片、多机集群并行化处理等技术手段
 - 数据处理和智能计算任务的多元化促使相关软件的多样化



9/22/2025



数据科学基础

96

- 包括高效的CPU/GPU、云计算、AI芯片、多机集群并行化处理等技术手段



- 云计算: EPYC (霄龙) 处理器; Project 47服务器



- 新一代处理器架构**VOLTA**: 新一代NVIDIA NVLink高速互联技术
- 云计算: Tesla V100 GPU 加速器; DGX-2 全球最大GPU; GPU云平台
- 机器人: Jetson Xavier 机器人专用AI芯片
- GPU工作站: 基于Volta架构的GV100
- 自动驾驶: Drive Xavier 首个自动驾驶处理器



- 云计算: 至强可扩展处理Purley; Xeon+FPGA (云端/设备端低功耗性能计算); Xeon Phi+Nervana (云端高性能计算)
- 无人驾驶: EyeQ 4/EyeQ 5 SoC
- 边缘计算: Myriad X VPU. Movidius Myriad X 视觉处理单元 (VPU), 是全球首个配备专用神经网络计算引擎的SoC
- 自学习神经元芯片: Loihi



- 云计算: 可重配置加速堆栈 (FPGA-Accelerator Stack)
- 设备端: reVISION 加速堆栈



- CPU架构: Cortex-A76
- GPU架构: Mali G76



- 云计算: TPU 3.0; Cloud TPU
- 移动端: Pixel VisualCore



- 移动端: 骁龙855处理器
- 智能驾驶: C-V2X芯片组



- 移动端: A12 芯片



- 移动端: 麒麟980芯片



- 跨界处理器: i.MX RT1060



数据科学基础

97

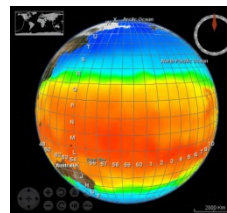


存储（如硬盘、数据库）



$$\min_X f(X) + \lambda \cdot \text{rank}(X)$$

分析、挖掘和学习



可视化



收集、传输



数据安全与个人隐私



生产、记录



基本程序与算法

.....



计算（平台与架构等）