



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/DS2024.html>

助教：肖桐  
ds\_intro2024@163.com

9/5/2024



# 课程要求与考核方式

2

- 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
  - 文献调研报告
  - 实验报告（需要编程）
  - 课程交流与课程汇报
- 考核方式
  - 课堂（30%）+调研报告（30%）+实验报告（40%）



# 课程说明

3

- 上课时间：1-18周，周二下午8、9节：15:35 - 17:30
- 上课地点：高新区 GT-B105
- 同类型课程：**学分不可重复认定**
  - 大数据学院《数据分析与实践》DS3001
  - 人文学院《新媒体大数据分析》NNM2011
- 几点说明
  - 课堂与平时重要，大家保证出勤
    - 不许叠课
    - 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
    - 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



# 联系方式

4

- 授课教师：黃振亚，陈恩红
- 课程主页：  
<http://staff.ustc.edu.cn/~huangzhy/Course/DS2024.html>
- QQ群：483794507
- 联系方式
  - 教师：
    - 黃振亚，[huangzhy@ustc.edu.cn](mailto:huangzhy@ustc.edu.cn)
  - 助教：
    - 肖桐
    - [ds\\_intro2024@163.com](mailto:ds_intro2024@163.com)



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

5

# 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/DS2024.html>

助教：肖桐  
ds\_intro2024@163.com

9/5/2024



# 课程历史沿革

7

1998

2012

2013

2014

2016

2017

首次开设面向研究生的数据挖掘课程，持续至今

大数据时代的到来，“大数据驱动科学发现”

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”

在实验室开设面向本科生的数据挖掘与机器学习研讨班

开始组建课程组  
广泛收集课程资源

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  
迁高新区，课程属性修改  
为计算机专业课



# 课程目标

8

- 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

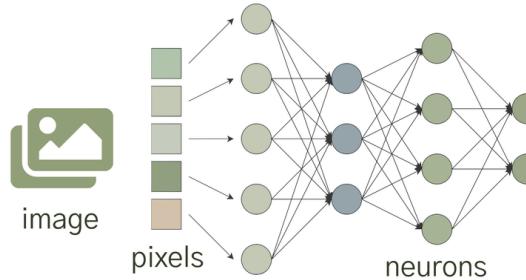
选修数据科学导论课程的同学将来可能从事不同领域的科学研究或者技术开发，希望这门课程带给你们的是终身受用的数据思维和创新能力。



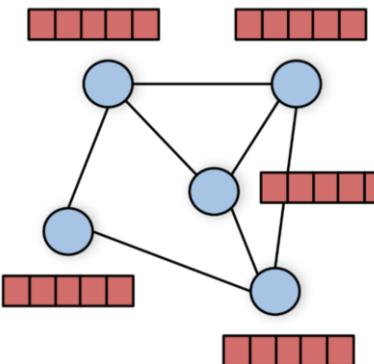
# 数据科学基础

9

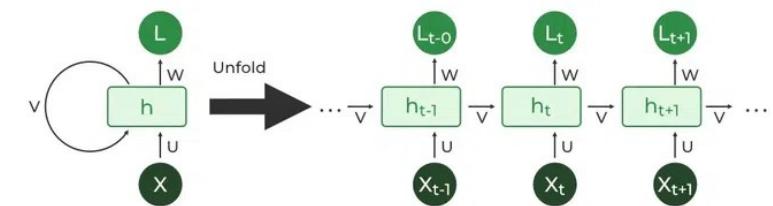
- 以模型为中心(model-centric)的数据科学技术
- 围绕目标任务特性，设计不同模型结构



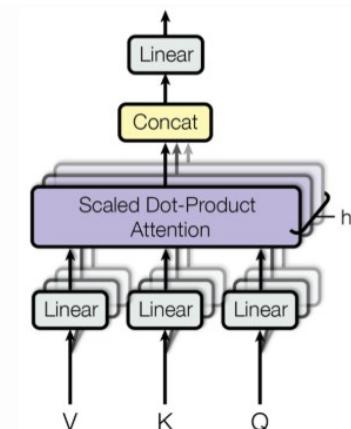
卷积神经网络CNN



图神经网络GNN



循环神经网络RNN



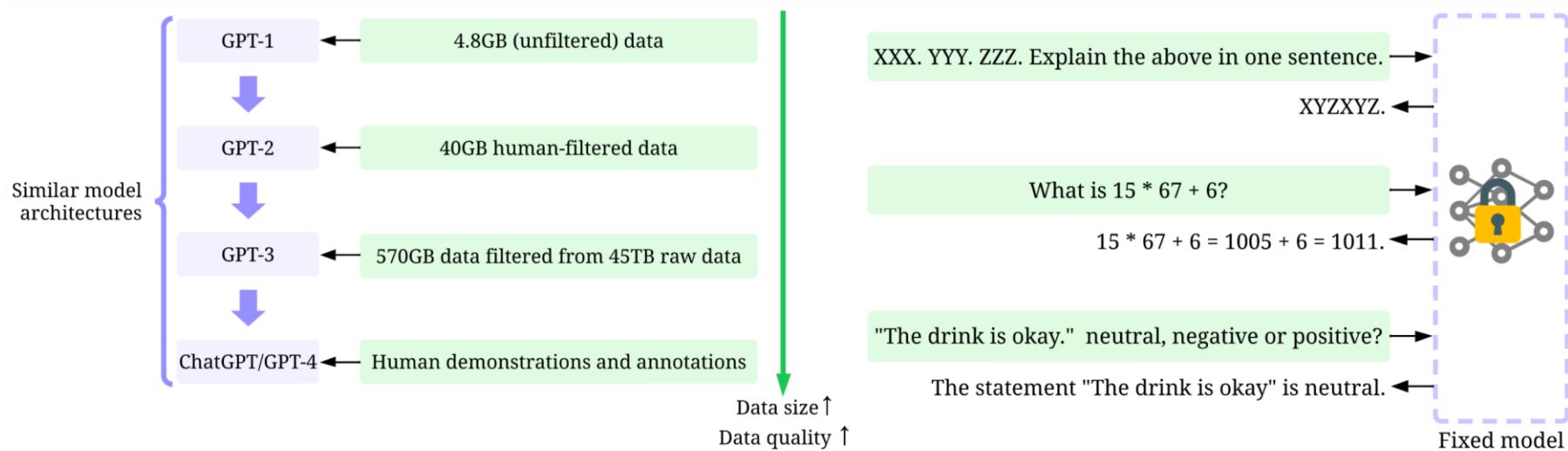
自注意力神经网络Transformer



# 数据科学基础

10

- 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

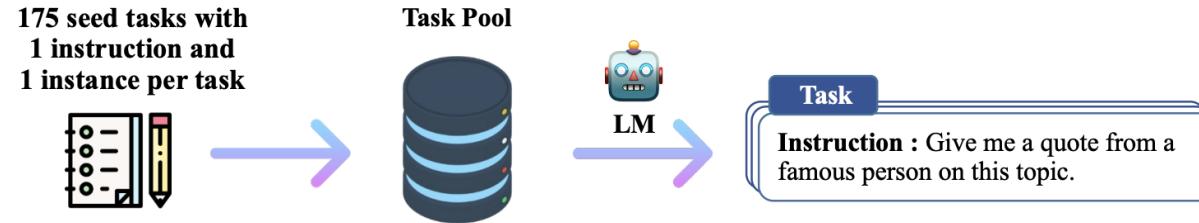




# 数据科学基础

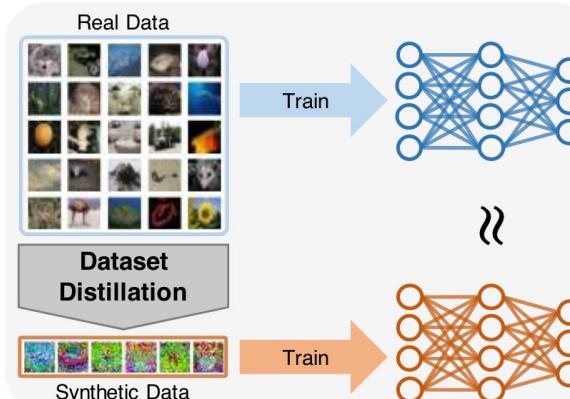
11

- 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

12

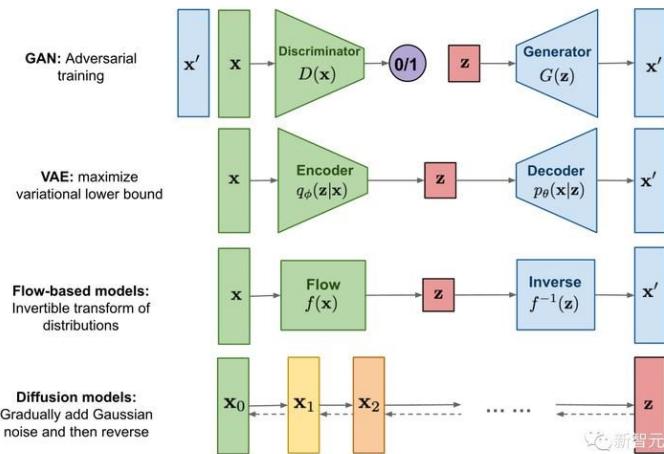
## □ 大数据催生人工智能新浪潮—扩散模型-2022

- 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



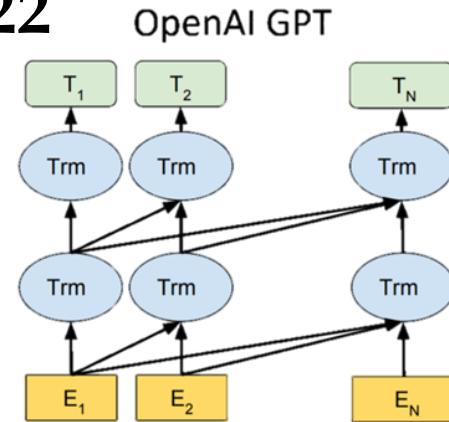


# 数据科学基础

13

## □ 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

### 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

### 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

2018

2019

2020

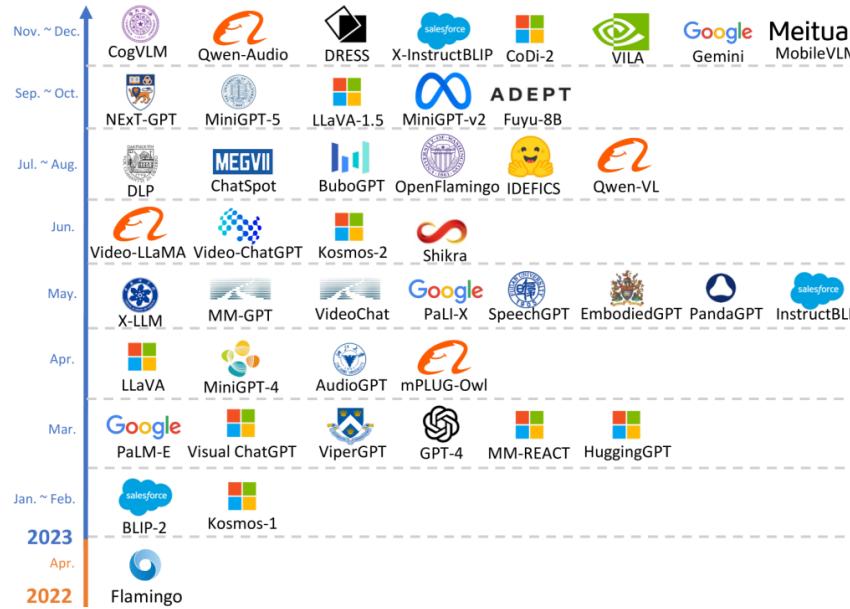


# 数据科学基础

14

## □ 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

15

- 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



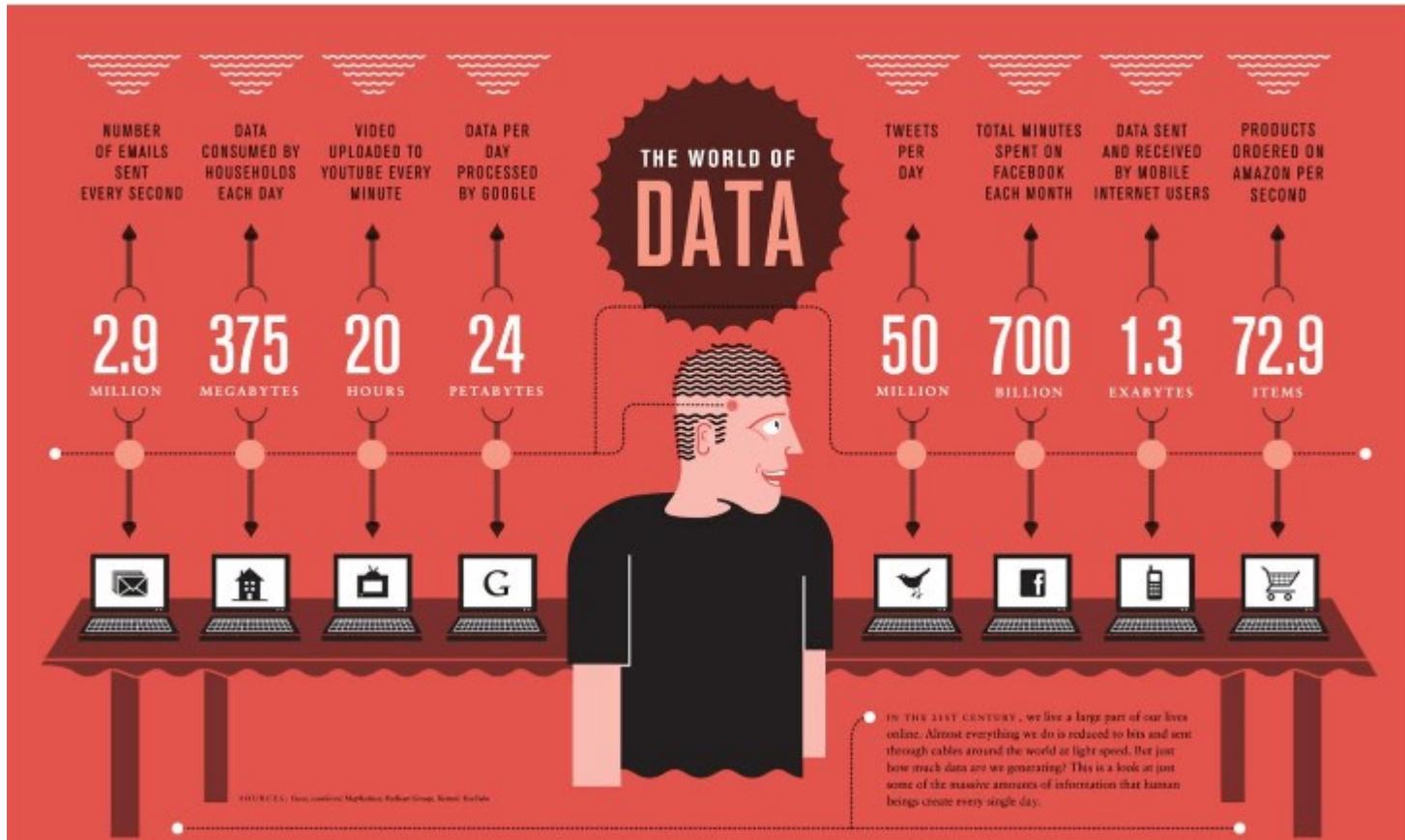
沟通数据



# 数据科学基础

16

- 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



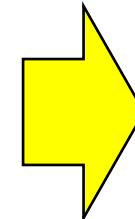


# 数据科学基础

17

## □ 例子：最不可能的地方获得数据

- 在现代农业中，使用传感器网络来监测和收集有关农田环境的数据。



土壤湿度监测  
气象数据采集  
光照强度监测  
.....

- 收集空气温度、光照强度、土壤湿度等数据，产生每个农田的全天候精确环境数据
- 帮助提高农业生产效率、减少资源浪费，有助于实现精准农业



# 数据科学基础

18

## □ 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

## 大数据有多大?

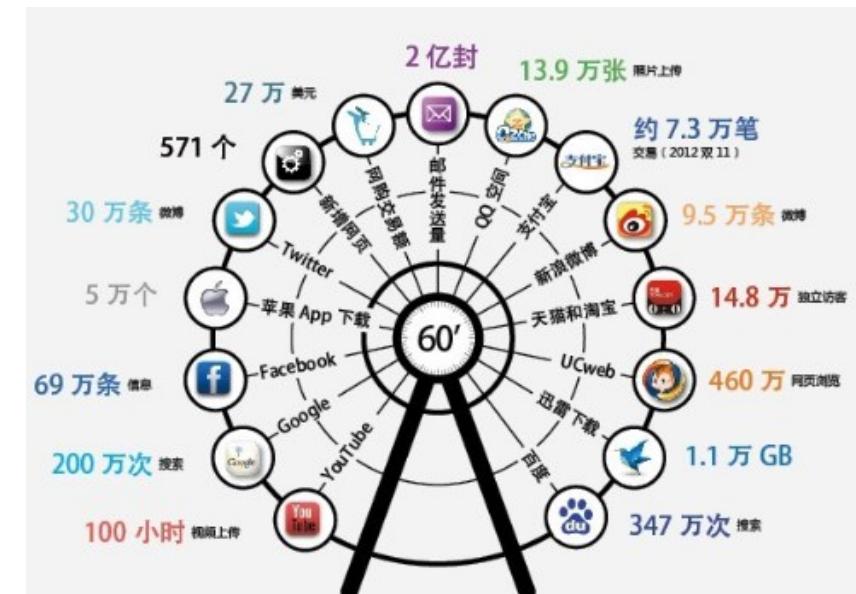
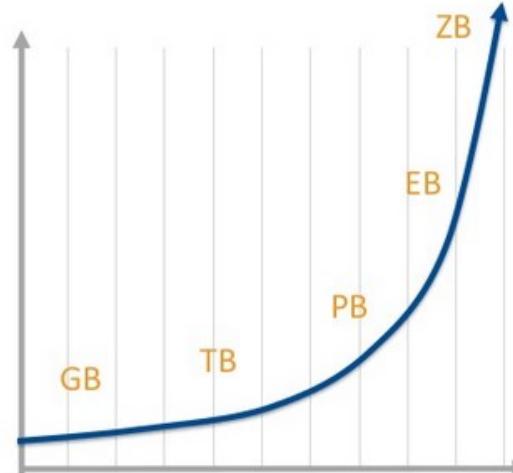
PB是大数据层次的临界点

### ◆ 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！

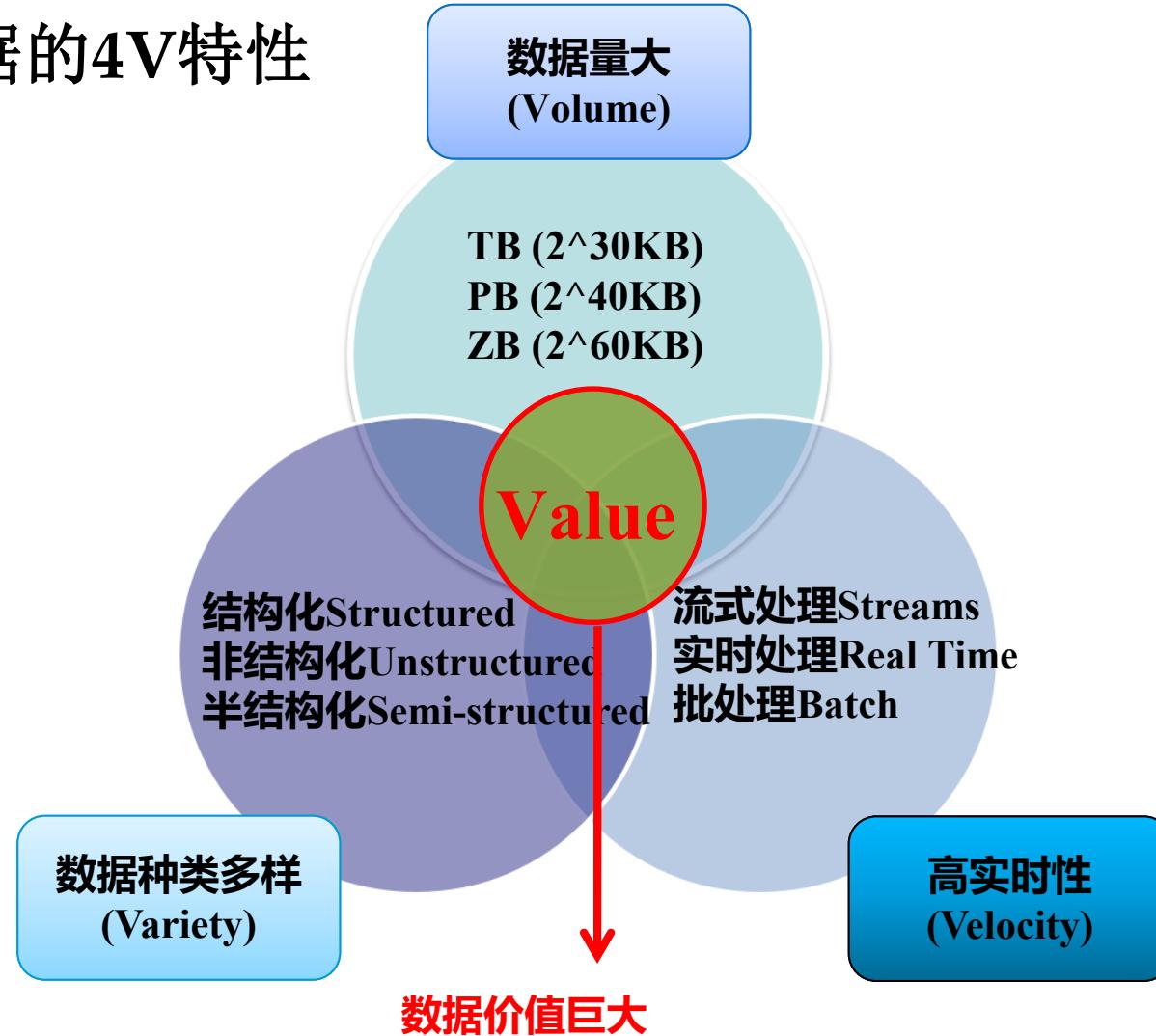




# 数据科学基础

21

## □ 大数据的4V特性





# 数据科学基础

22

## □ 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

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

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



# 数据科学基础

23

## □ 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



事务数据



视频数据



音频数据



时空数据



文本数据



图像数据



# 数据科学基础

24

## □ 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



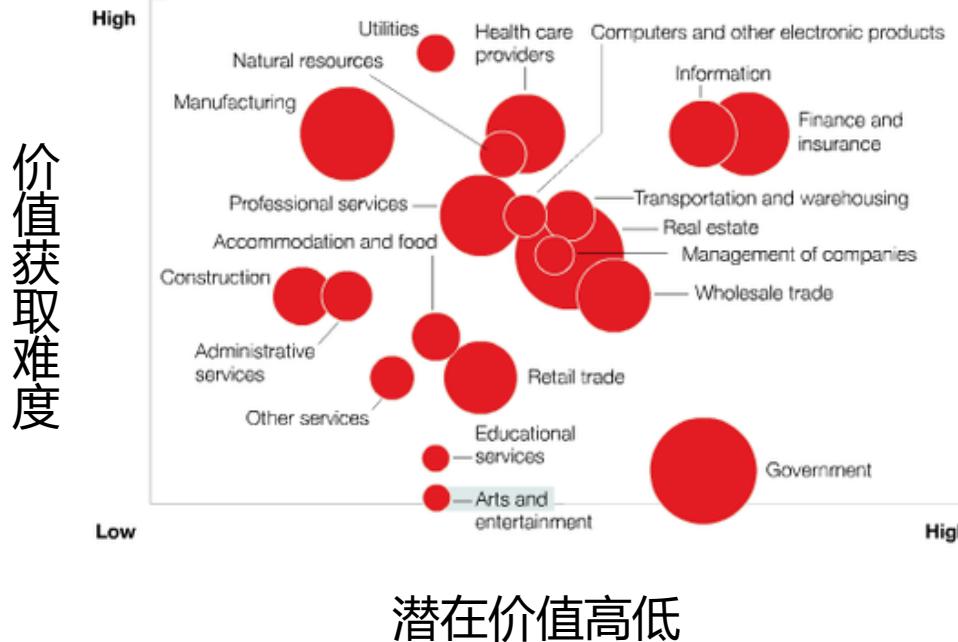


# 数据科学基础

25

## □ 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

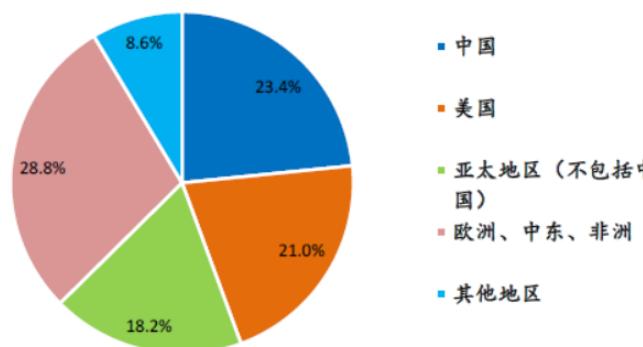


# 数据科学基础

28

## □ 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

32

## □ 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



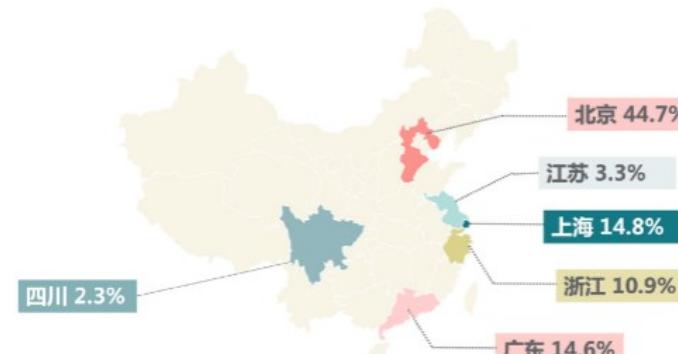


# 数据科学基础

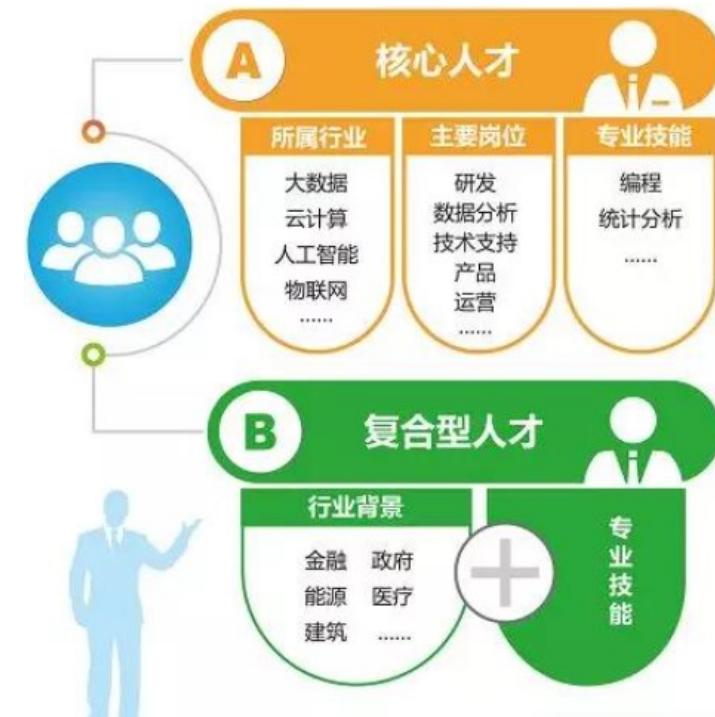
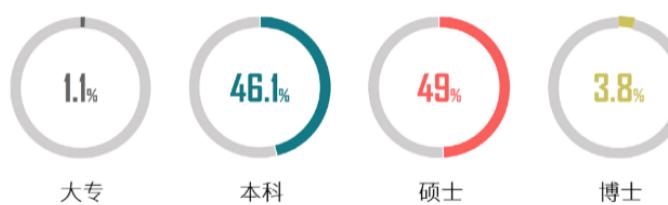
33

## □ 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



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

34

## □ 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

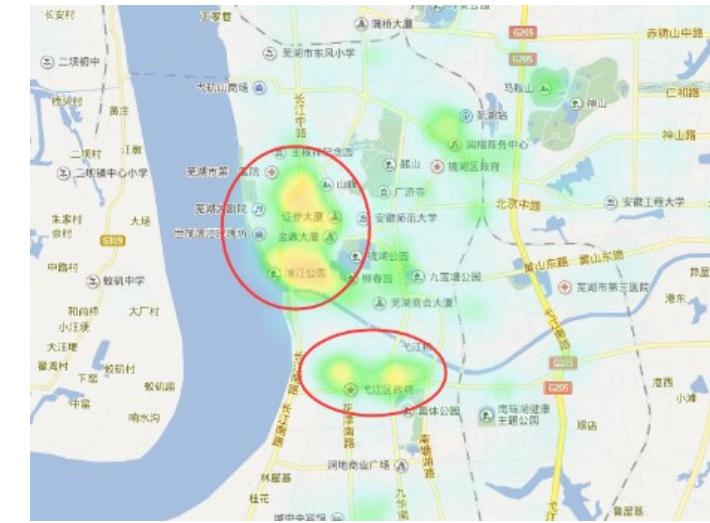
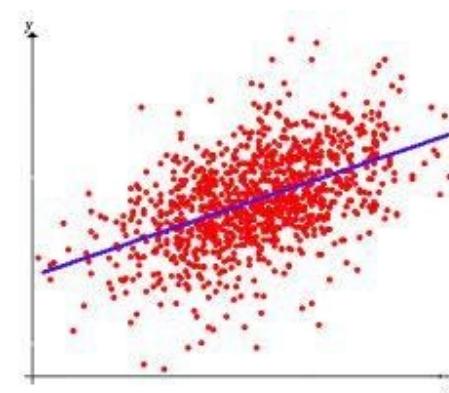
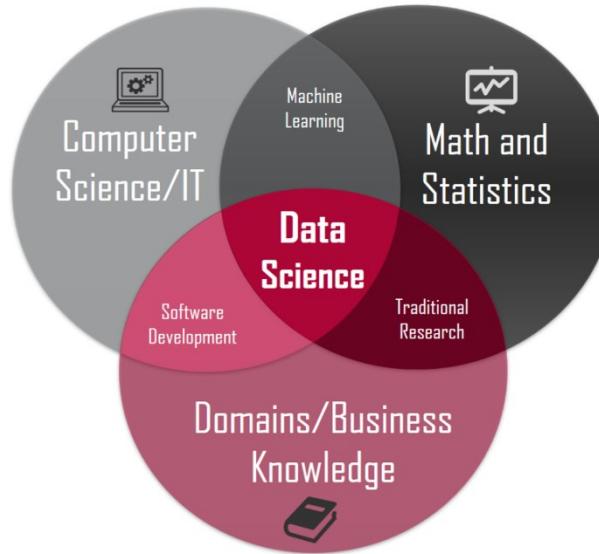


# 数据科学基础

35

## □ 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

36

- 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
  - . . .



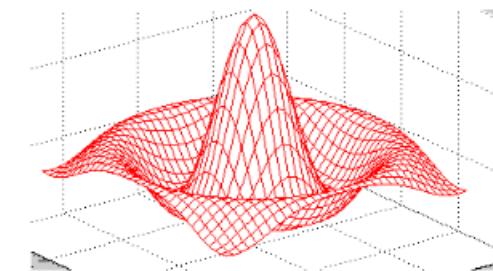
# 数据科学基础

37

- 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

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

优化方法



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

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

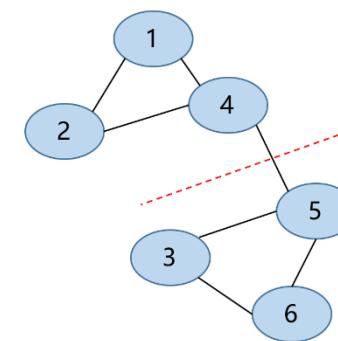
38

- 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
    - 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



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

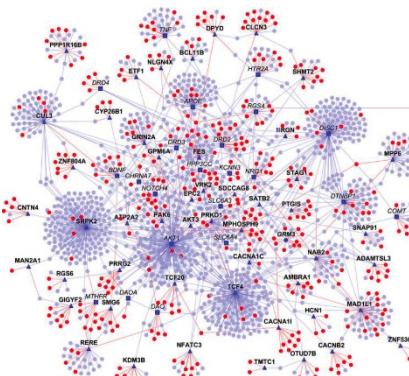


# 数据科学基础

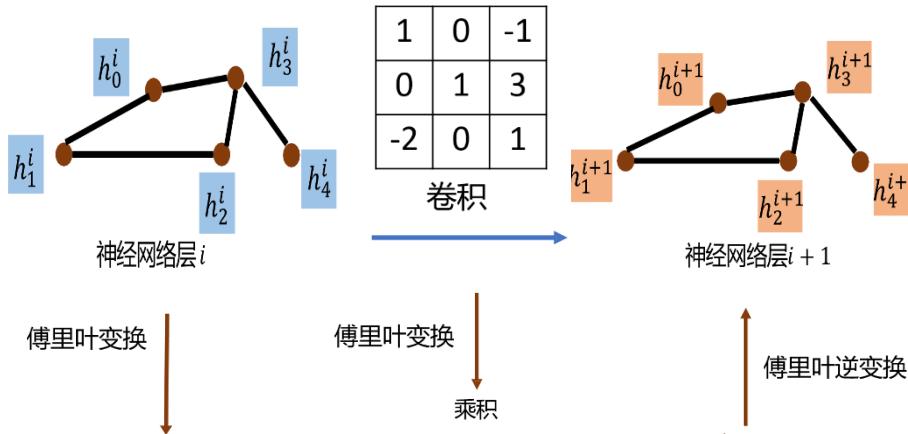
39

- 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



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

40

## □ 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**

```
1 def SumOfKsubArray(arr, n, k):
2     Sum = 0
3     S = deque() = deque()
4     G = deque() = deque()
5     for i in range(k):for i in range(n):
6         while (len(S) > 0 and arr[S[-1]] >= arr[i]):while (len(S) > 0 and arr[S[-1]] >= arr[i]):
7             S.pop()
8         while (len(G) > 0 and arr[G[-1]] <= arr[i]):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):for i in range(k, n):
13        Sum += arr[S[0]] + arr[G[0]]Sum += arr[S[0]] + arr[G[0]]
14        while (len(S) > 0 and S[0] <= i - k):while (len(S) > 0 and S[0] <= i - k):
15            S.popleft()
16        while (len(G) > 0 and G[0] <= i - k):while (len(G) > 0 and G[0] <= i - k):
17            G.popleft()
18        while (len(S) > 0 and arr[S[-1]] >= arr[i]):while (len(S) > 0 and arr[S[-1]] >= arr[i]):
19            S.pop()
20        while (len(G) > 0 and arr[G[-1]] <= arr[i]):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]]Sum += arr[S[0]] + arr[G[0]]
25    return Sumreturn Sum
26
27
```



知乎 @邢建





# 数据科学基础

41

## □ 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

42

## □ “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**753所高校**新增该专业
  - 人工智能自2018年来有**537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
人工智能	-	35	180	130	95	59	38
数据科学与大数据技术	250	203	137	62	40	30	31
大数据管理与应用	5	25	51	59	42	38	32

2017年以来新增大数据相关专业的高校数量



# 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



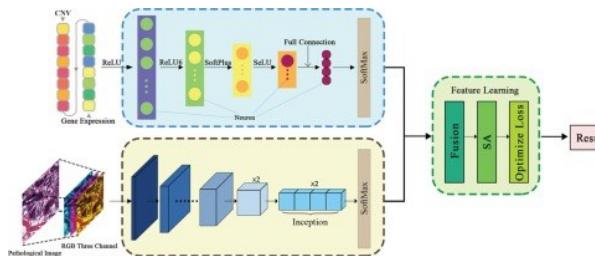
# 数据科学基础

44

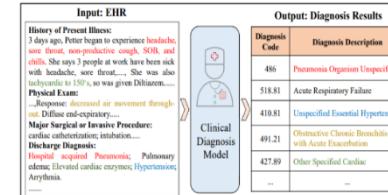
- 数据蕴含着巨大的价值—智慧医疗
  - 通过对患者建立AI电子病历
  - 整合患者的全时段、多模态的健康数据（病例文本、检查影像等）
  - 实现对患者的疾病诊断、病灶识别、药物推荐等



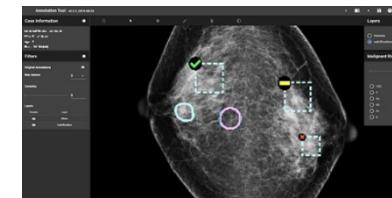
AI电子病历



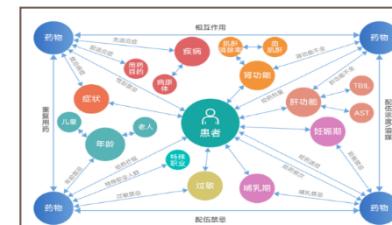
多模态医疗数据挖掘模型



疾病诊断



病灶识别



药物推荐



# 数据科学基础

45

- 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



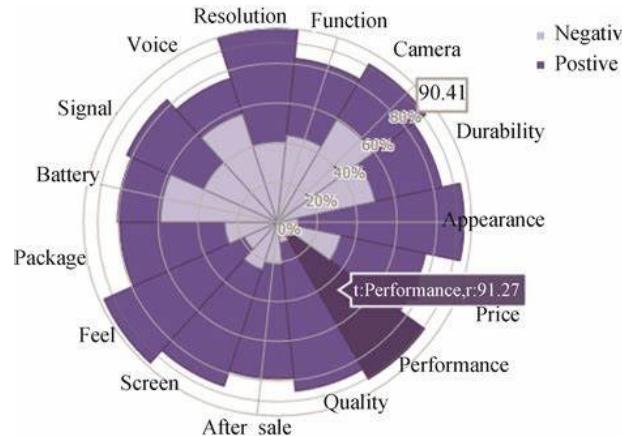
# 数据科学基础

46

## □ 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

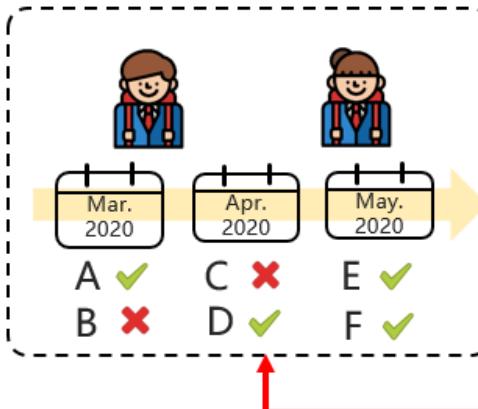


# 数据科学基础

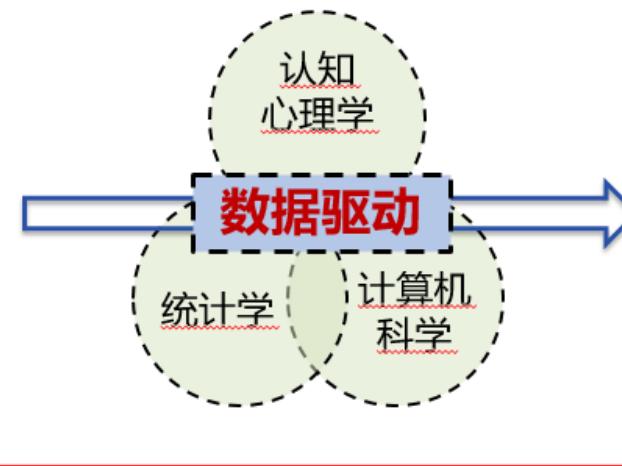
47

- 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 数据驱动的范式

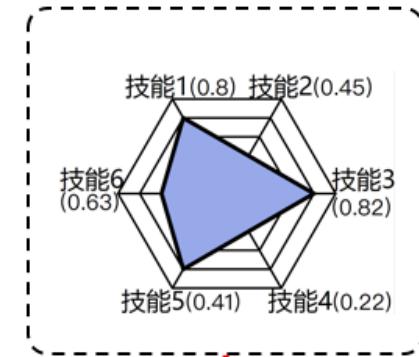
学习过程相关数据



知识能力及其发展变化



反馈调整教学策略

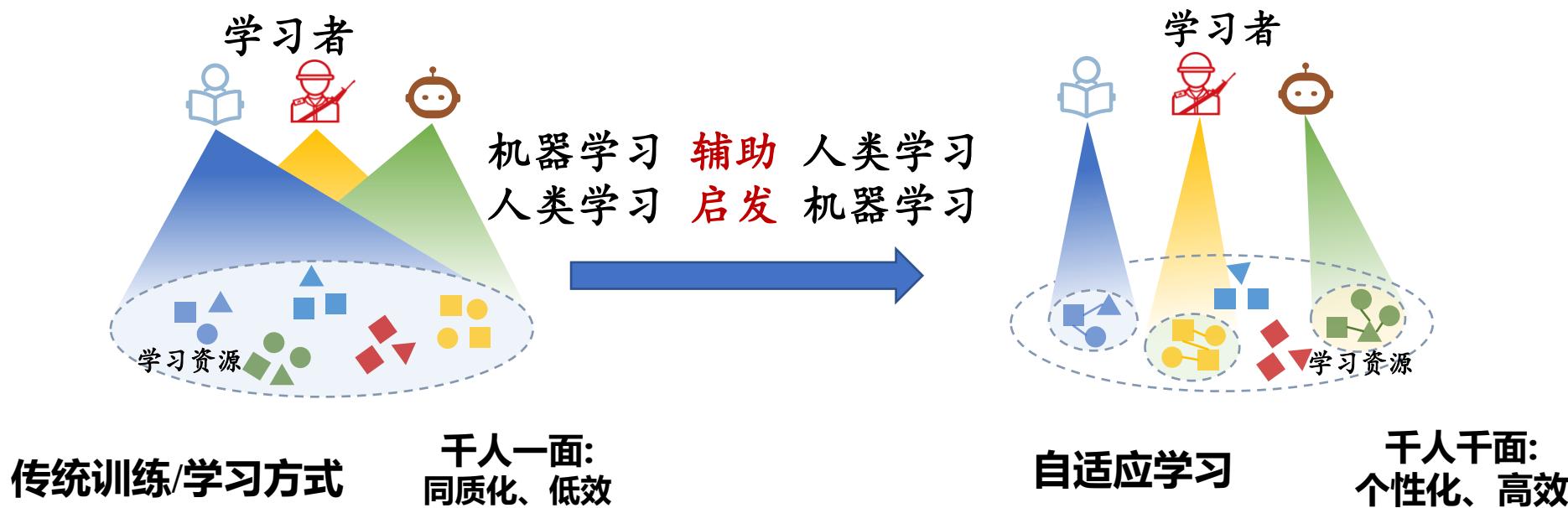




# 数据科学基础

48

- 数据蕴含着巨大的价值—智慧教育
  - 规模化因材施教：千人千面的自适应学习



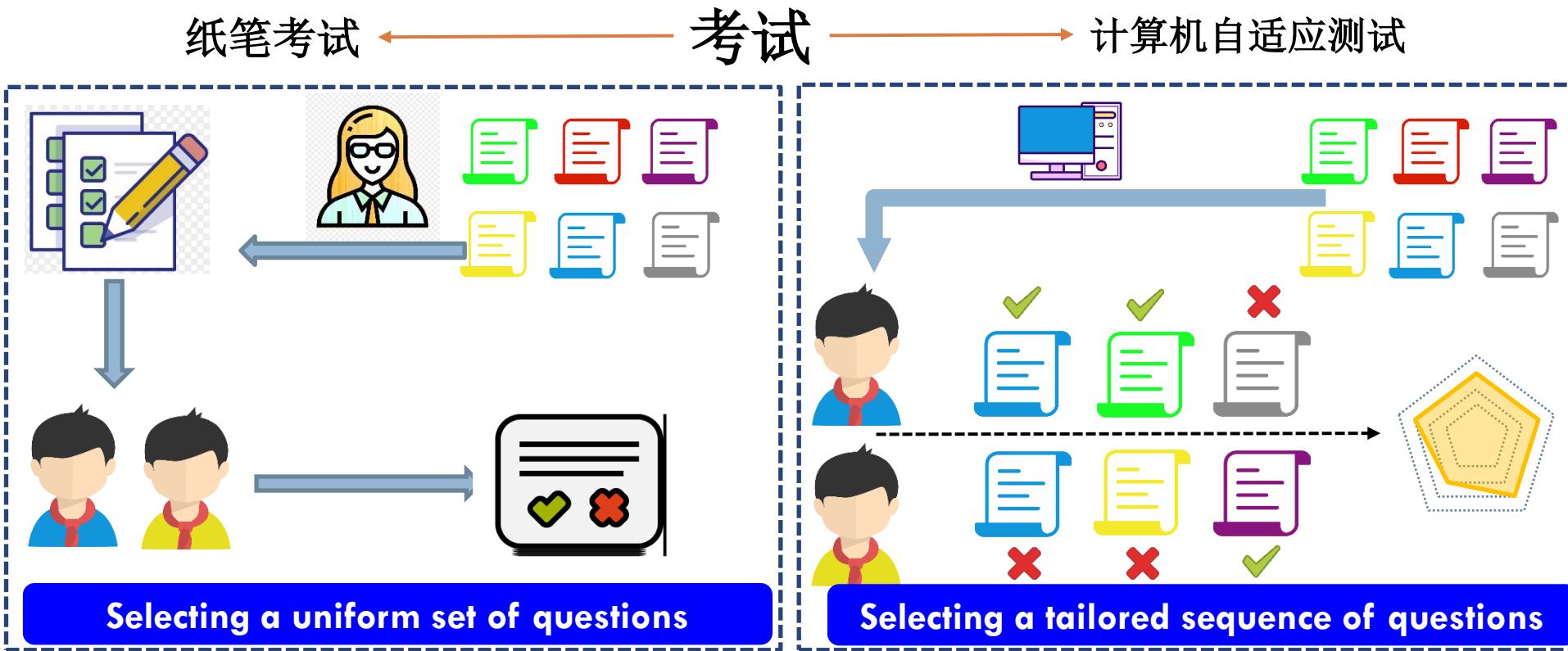


# 数据科学基础

49

- 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®





# 数据科学基础

50

- 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。



# 数据科学基础

51

- 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 $[X \cdot Y]$ 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 $[X \cdot Y]$ 的补码：

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

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： $[X \cdot Y]_{\text{补}} = 1.01111100$

因此，根据布斯算法， $[X \cdot Y]$ 的补码为  
 $1.01111100$ 。



数字化智能体系统

## 星火科研助手

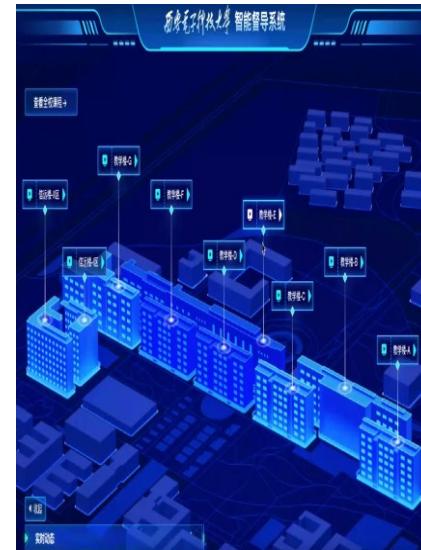
星火科研助手致力于创建高效、可靠的科技文献服务平台，为学术研究人员、企业专业人士和其它相关用户提供服务，助力科技发展创新。

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

52

## □ 数据蕴含着巨大的价值—智慧教育

高校编程平台(CODIA): <https://code.bdaa.pro>

- 平台特色：利用人工智能技术辅助学生编程学习
  - 面向学生：提供学生诊断、习题推荐、智能助教等个性化服务
  - 面向老师：支撑教师自主出题，测试情况跟踪等
- 平台应用：已在中国科大本科生课程实践教学中使用
  - 2021-2024年，计算机学院《程序设计II》、《数据结构》课程实验教学
  - 2021-2024年，实验室保研/考研机试



最近更新

C语言初学者看过来，这里有属于你的一份惊喜  
知识路径模块上线,迎来多重更新学以致练，能力进阶，知识路径C语言模块与配套习题重磅上线，整合智能三模块，打通学练全流程

NICE/WC  
迎接新学员上线  
立即查看 立即体验

14459人 用户数量 | 371259次 提交次数 | 1717天 安全运行

用户 ovu6f6kox 刚刚提交了题目 的结果 用时 29 ms

智慧助教 用户调用：  
21825 times 28574341 tokens

编程到来，智能将来

掌握编程，是对智能时代的准备，亦是提升竞争力的重要途径。在这里，你可以获得：

- 历年考研及保研机试题库
- 数据分析及算法等课程相关题库
- 个性化能力评估和诊断

现在加入

数据结构

- 算法基本操作
- 字符串处理
- 二叉树遍历
- 堆栈和队列
- 递归和剪枝
- 动态规划



# 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。



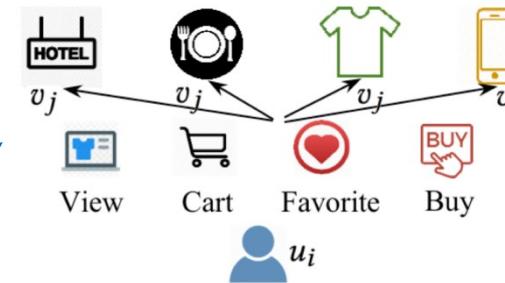
有一种说法称，特朗普的团队聘用数据分析公司，做了精准的广告投放，影响了那些徘徊不定的选民，拿下了决定性的关键州选举人票。





# 数据科学基础

- 数据蕴含着巨大的价值—电子商务
  - 计算广告

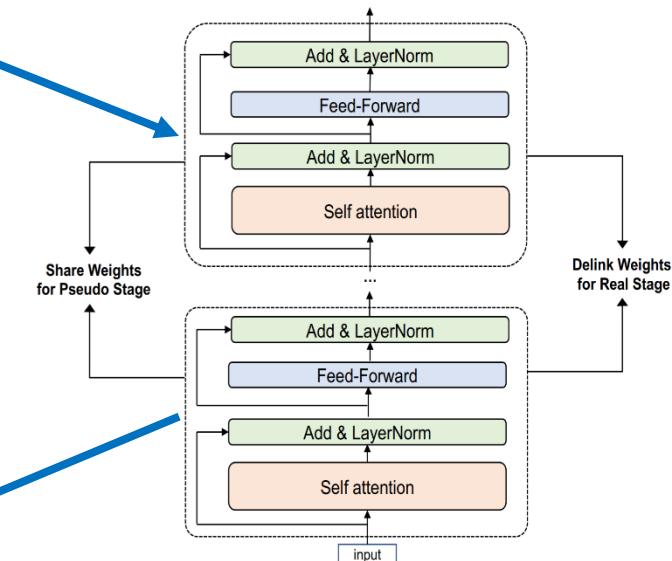


海量用户多样交互行为



电商平台

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

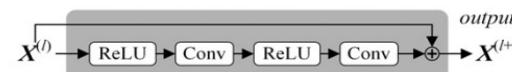
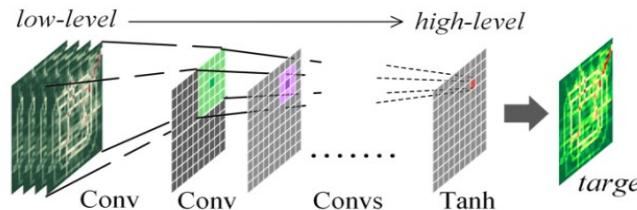
55

- 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

56

## □ 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**

