華東師範大學
EAST CHINA NORMAL UNIVERSITY

数据科学与工程算法基础

Algorithm Foundations of Data Science and Engineering

第一章 绪论

$$(1+x)^n = 1 + \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} + \cdots$$

课程提纲

Content

课程提纲

Content

新时代的新生产要素

01

农耕时代

政治经济学之父**威廉•配第**说：“**土地**为财富之母，而**劳动**则为财富之父”

02

工业时代

马歇尔的划时代著作**《经济学原理》**出版，提出了生产要素四元论——**土地、劳动、资本**和**技术**

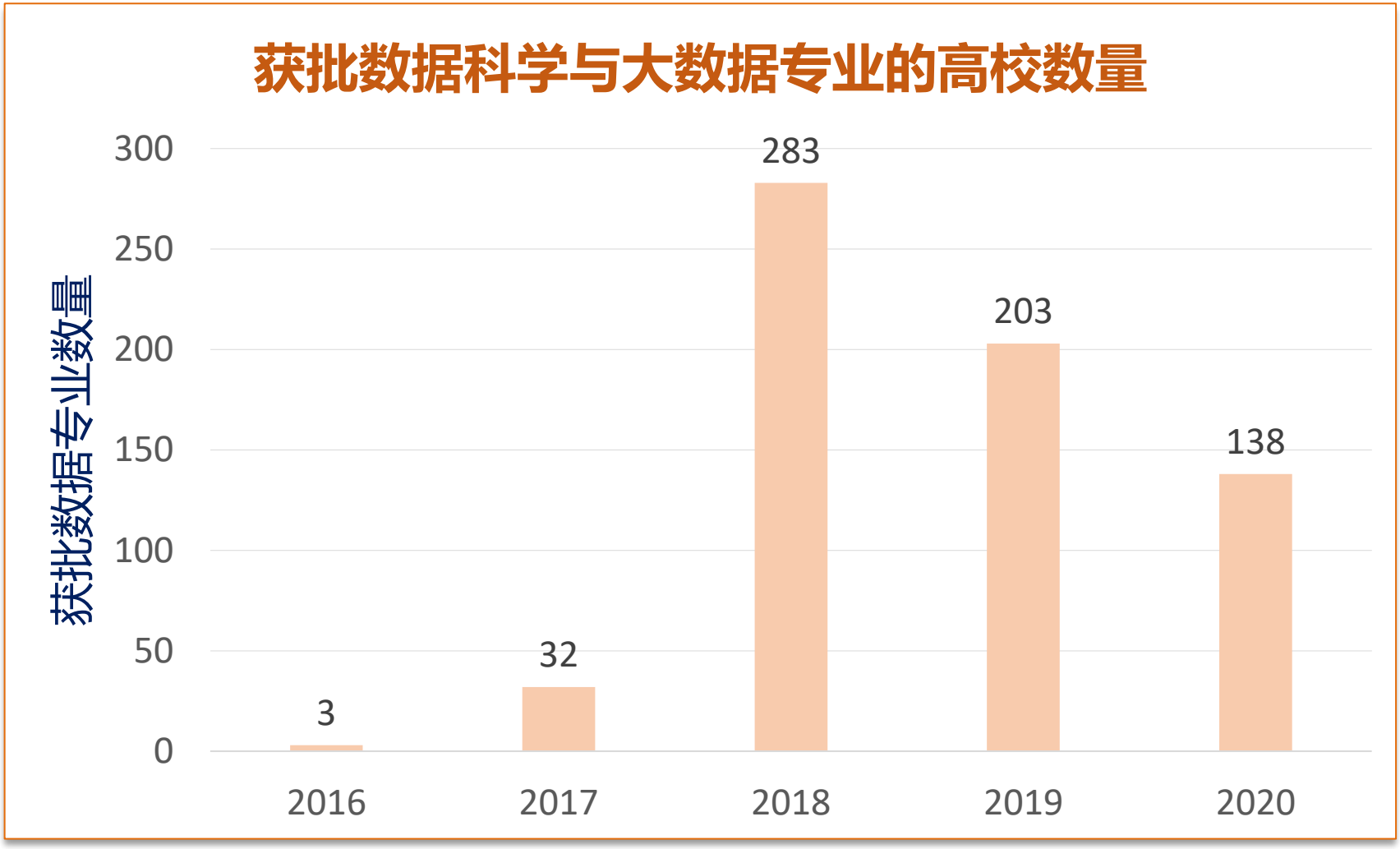
03

数字时代

2020年4月，国务院将**数据**作为与土地、劳动力、资本、技术并列的生产要素

Data is Power!

数据科学与大数据技术人才培养



全国**超过 700 家**高校开始培养数据科学与大数据技术人才

数据赋能

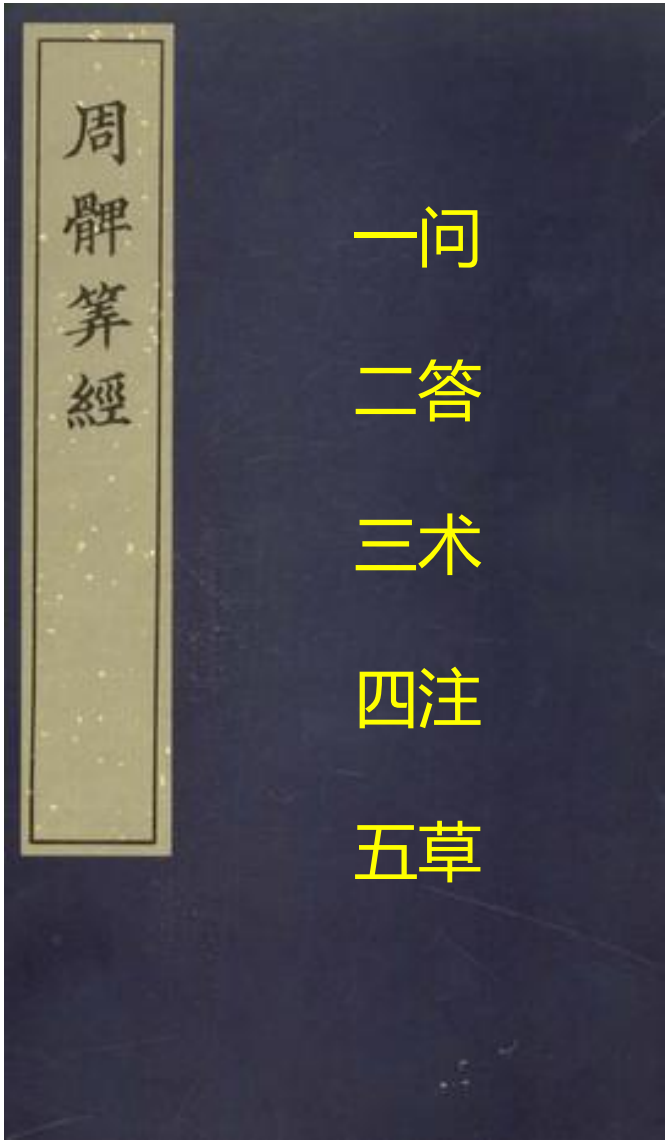
数据 **算力** **算法**

- 数据本身不具有价值，**赋能业务**才能实现其价值
- 数据赋能 = 数据 + 算力 + 算法
  - 互联网和物联网应用的发展，致使各行各业积累了海量数据
  - **摩尔定律**为数据赋能提供了硬件支撑
  - **大数据生态**的不断壮大为数据赋能提供了平台支撑
  - 万事俱备，只欠东风 ------ **算法**

课程提纲

Content

中国古代数学



- “算法”即演算法，其名称出自**《周髀算经》**
- 吴文俊：中国古代数学就是一部**算法大全**
  - 不同于希腊数学的公理化论证
  - 注重通用方法，而不是特殊技巧
  - 中国古代数学其实只有一个关键字：**术**
  - 相当于现代计算机科学中的算法
    - ✓辗转相除法（欧几里德算法）
    - ✓割圆术
    - ✓秦九韶算法
    - ✓……

经典算法

- 算法是完成一个任务的具体步骤和方法
  - 有穷指令
  - 无二义性
  - 可行性
  - 输入/输出
- 常用算法
  - 穷举法
  - 递归法
  - 动态规划
  - 贪心法
  - 回溯法
  - 分支限界
  - 分治法
  - ……
- 算法复杂度
  - 时间复杂度
  - 空间复杂度
  - 网络传输复杂度
- 除了时空复杂度，实际应用中的算法还有很多的考量

与经典算法不同

- 关注点不同
  - 数据处理的全生命周期
  - 性能指标：时间、空间和网络 + 算法精度
  - 数据特点：结构化、半结构化和非结构化
  - 问题特点：综合运用线性代数、概率统计和优化等数学知识对问题进行建模

- 经典算法主要介绍高效算法设计和分析技巧
  - 经典算法是**前序课程**
  - 实际应用中的算法还需关注**数据特点**和**求解问题本身**

数据全生命周期

除了时空指标，数据处理因阶段的不同
算法设计的考量因素也不同

数据服务

- 数据可视化
- 可解释性

数据分析挖掘

- 预测精度
- 信息损失

数据存储与查询

- 数据一致性
- 存储效率
- 查询效率

数据采集汇聚

- 数据完整性
- 数据可靠性
- 数据质量

数据科学Top-25算法

N	Algorithm	2016	2011	Domain
1	Regression	67%	58%	Statistics
2	Clustering	57%	52%	Data Mining/Statistics
3	Decision Trees/Rules	55%	60%	Data Mining
4	Visualization	49%	38%	Visualization
5	K-nearest neighbors	46%	-	Data Mining
6	PCA	43%	-	Statistics
7	Statistics	43%	48%	Statistics
8	Random Forests	38%	-	Data Mining
9	Sequence analysis	37%	30%	Data Mining
10	Text Mining	36%	28%	NLP
11	Ensemble methods	34%	28%	Machine Learning
12	SVM	34%	29%	Machine Learning
13	Boosting	33%	23%	Machine Learning

数据科学Top-25算法

N	Algorithm	2016	2011	Domain
14	Neural networks	24%	27%	Machine Learning
15	Optimization	24%	-	Optimization
16	Naive Bayes	24%	22%	Machine Learning
17	Data Integration	22%	20%	Data Management
18	Anomaly detection	20%	16%	Data Mining
19	Deep Learning	19%	-	Machine Learning
20	SVD	16%	-	Algebraic
21	Association rules	15%	29%	Data Mining
22	Graph Mining	15%	14%	Data Mining
23	Bayesian networks	13%	-	Machine Learning
24	Genetic algorithms	8.8%	9.3%	Machine Learning
25	Survival Analysis	7.9%	9.3%	Statistics
26	EM	6.6%	-	Statistics

常用算法总结

N	研究领域	数量
1	数据挖掘	9
2	机器学习	8
3	统计学	4
4	可视化	1
5	自然语言处理	1
6	数据管理	1
7	优化	1
8	代数	1

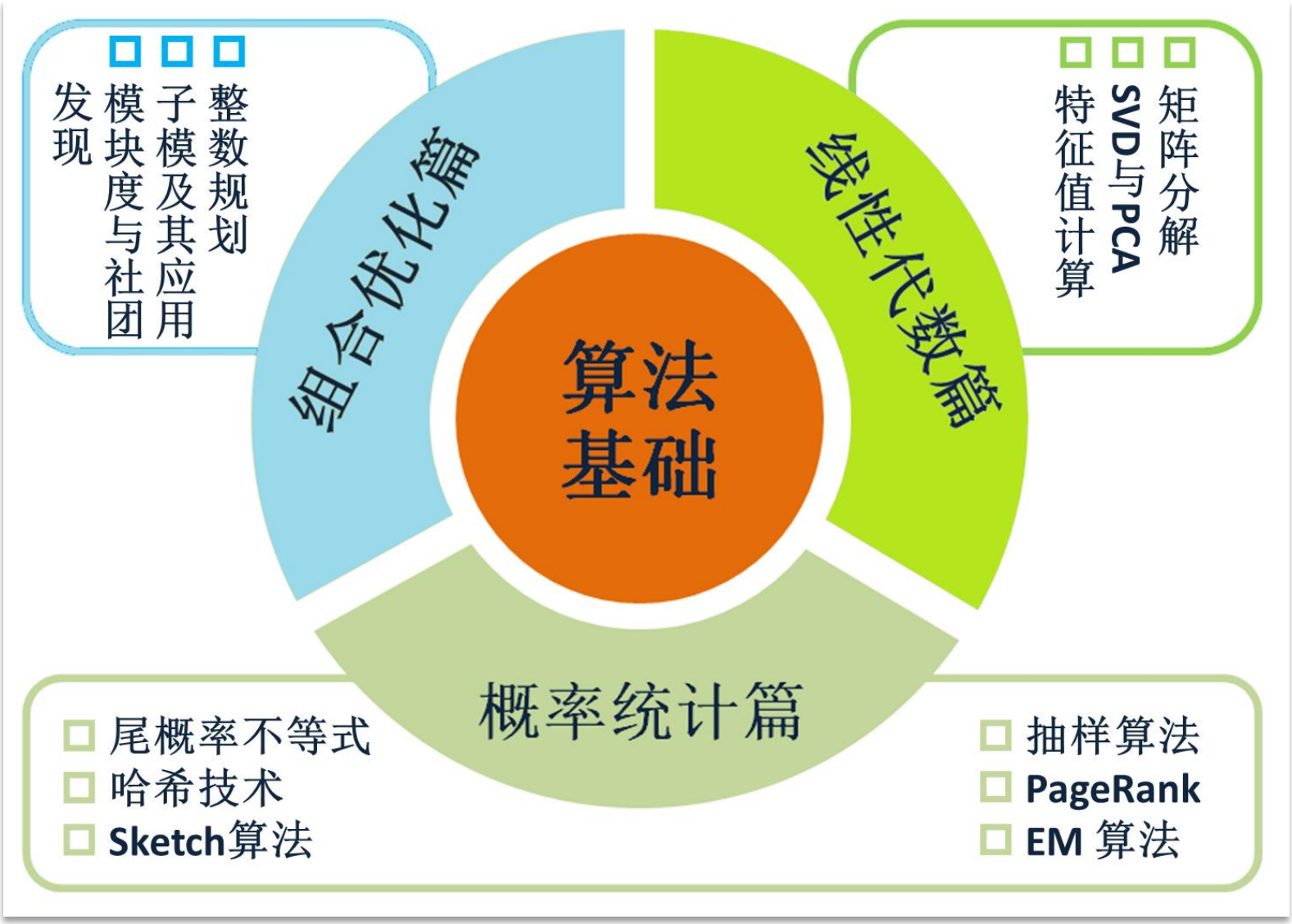
N	数据模型	数据类型
1	关系数据	结构化数据
2	时间序列	半结构化
3	图数据	半结构化
4	文本数据	非结构化
5	图片	非结构化
6	视频	非结构化
7	音频	非结构化

- 数据科学与工程算法涉及多个不同学科
- 处理数据类型囊括结构化、半结构化和非结构化数据
- 算法涉及**数据处理的全生命周期**

数据全生命周期与算法

阶段	关注点	典型算法
数据采集	抽样可靠性、广泛性和代表性	抽样算法、哈希技术
数据预处理	数据质量和数据可用性	EM算法、哈希技术
数据存储与管理	数据一致性、查询与存储效率	尾概率不等式 哈希技术 Sketch算法
数据分析挖掘	预测精度、信息损失	PageRank、SCD与PCA 矩阵分解 社区发现 EM算法 整数规划 子模函数优化
数据可视化	可解释性、无损性	特征值计算 SCD与PCA 矩阵分解 社区发现

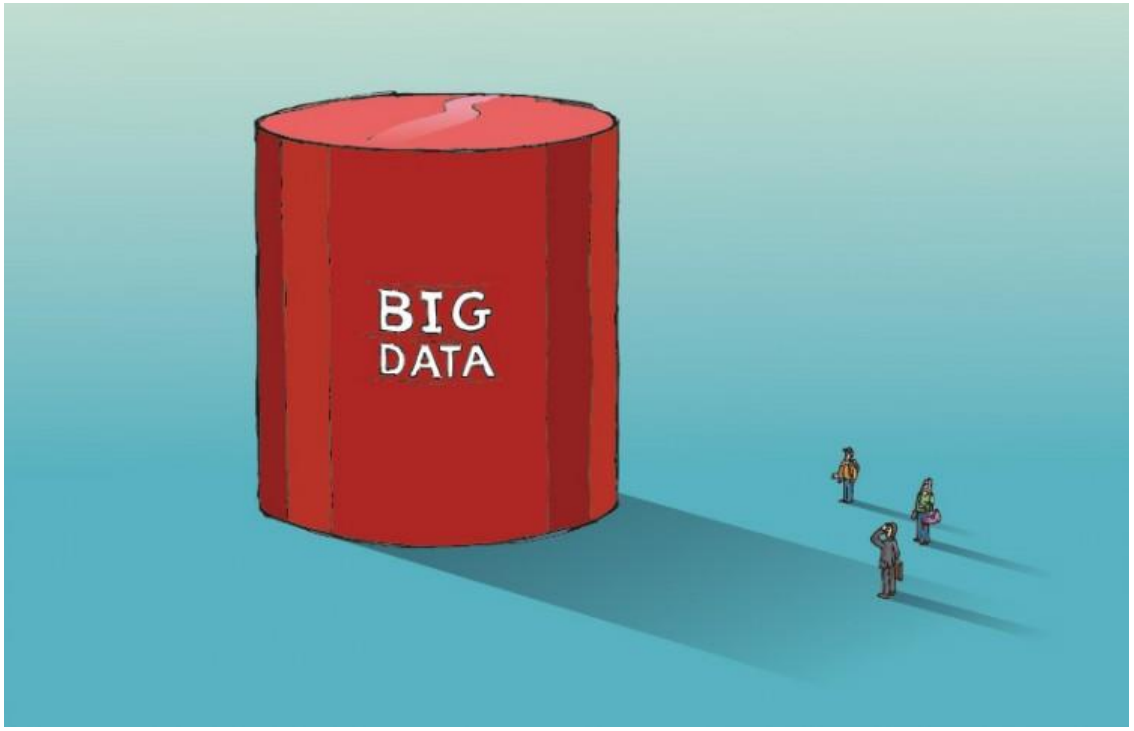
教材建设内容


组合优化篇
整数规划
子模及其应用
模块度与社团发现
线性代数篇
矩阵分解
SVD与PCA
特征值计算
算法基础
概率统计篇
尾概率不等式
哈希技术
Sketch算法
抽样算法
PageRank
EM 算法

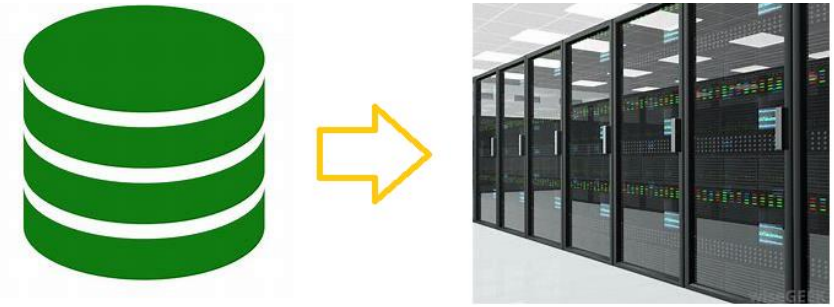
课程提纲

Content

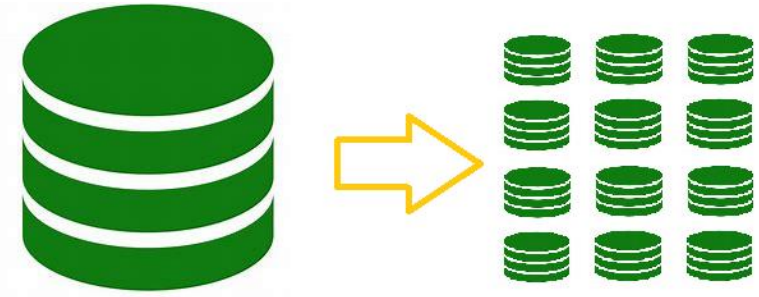
数据特点 — 数据规模



- 数据规模持续快速增加
- PB 级数据已很常见
  - 互联网企业
  - 大型科学设备
  - 物联网应用

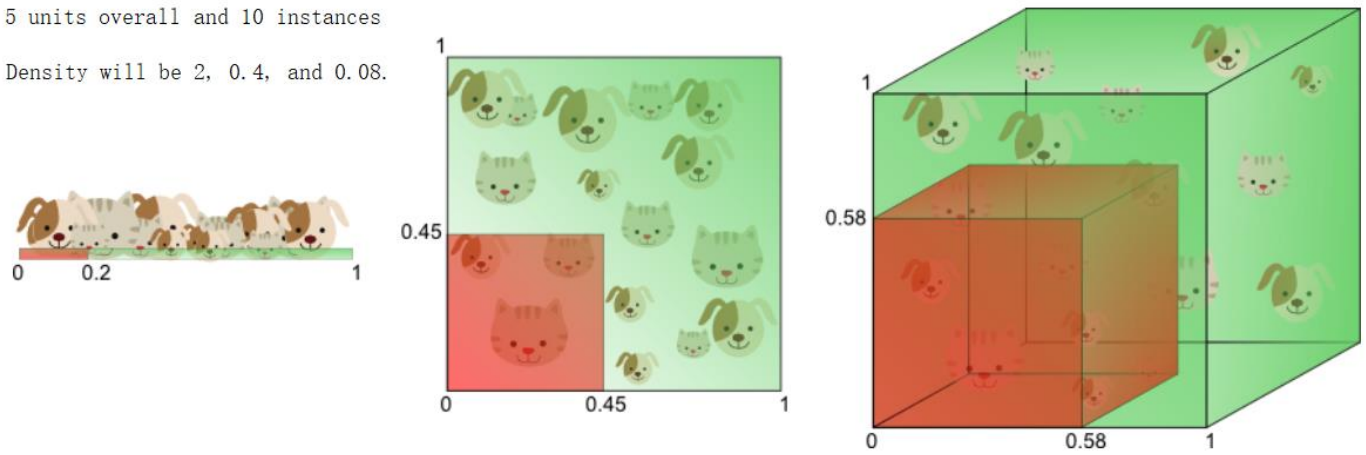


硬件扩展



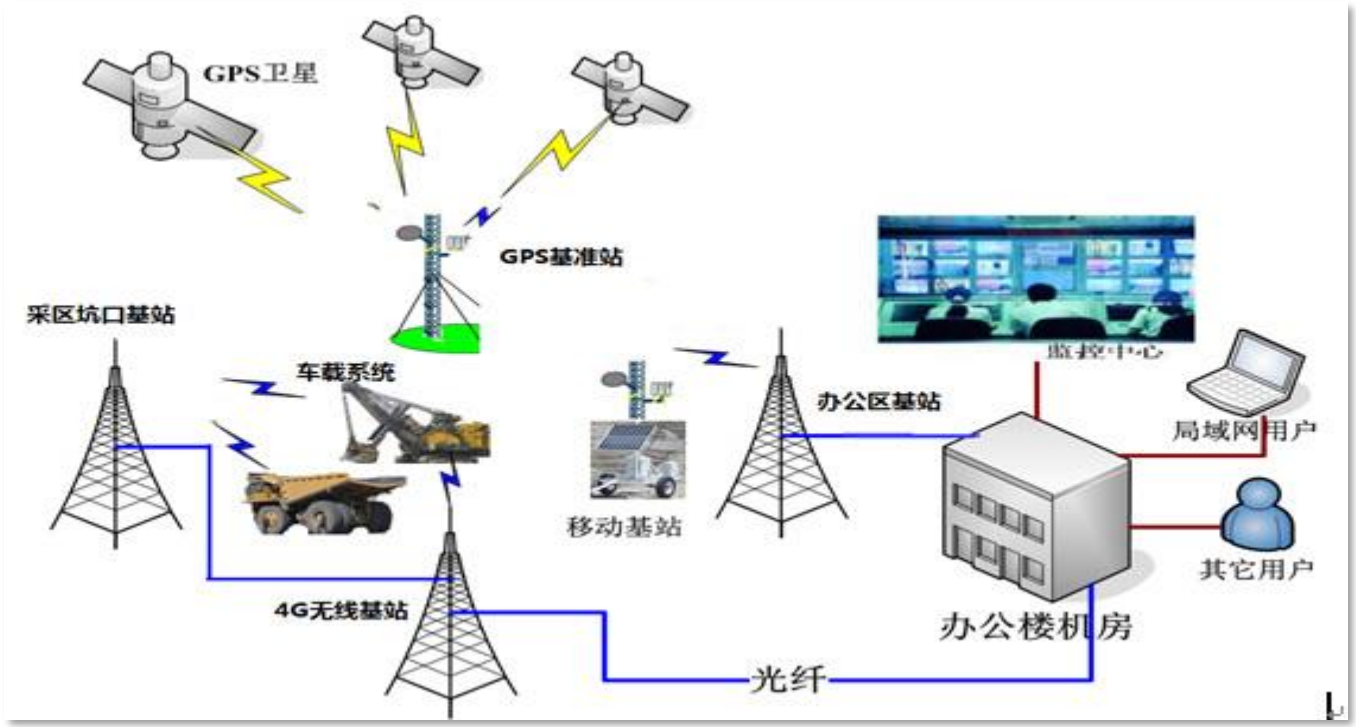
抽样

数据特点 — 数据维度



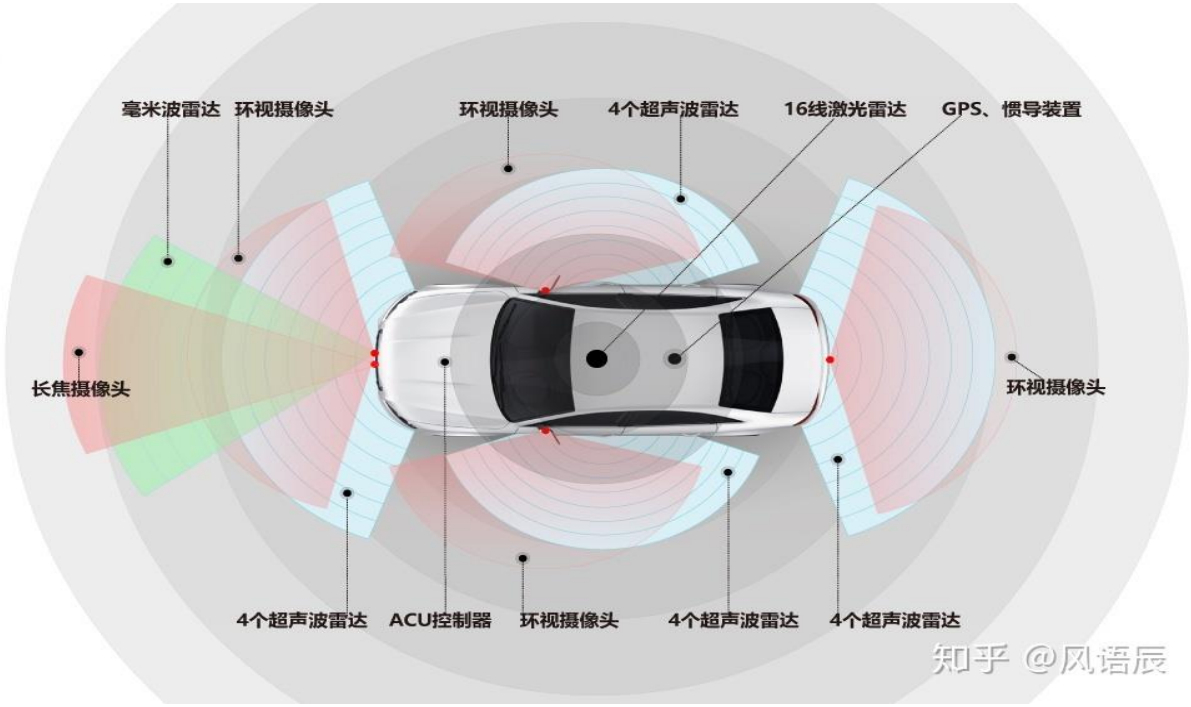
- 二维特征空间中，为覆盖 20% 的特征空间，需要获取 45% 的样本数量，在三维情况下，这种问题变得更糟，比例将会增至 58% ($0.58^3=0.2$)
- 如果维度继续增加，则训练数据规模需要指数级增加才能避免过拟合
- 文本、图和图像等数据类型都是高维数据
- 高维数据会导致“维度灾难”，应对方法
  - 降维
  - 提高模型复杂度
  - 增加学习样本数据规模

数据特点 — 数据到达速度



- 数据到达和服务响应速度
  - 随着物联网技术的发展，数据达到的速度越来越快
  - 很多 To C 应用需要实时响应用户需求
  - 应对方法
    - ✓降低算法复杂度
    - ✓放弃精确的计算结果

数据特点 — 数据异构程度



- 自动驾驶是典型的多模态数据应用场景
  - 各类传感设备：雷达、摄像头、超声波、GPS等
  - 如何综合多模态数据进行决策
    - ✓数据清洗与去噪
    - ✓异构数据对齐、融合
    - ✓提升数据质量

算法评价

指标类型	角度	指标	含义
效率指标	时间	Elapsed time	时间开销
	空间	Storage	空间开销
	网络	Communication	网络传输
精度指标	分类问题	Precision	准确率
		Recall	召回率
		F1	F1 值
		AUC	ROC 曲线下面积
	回归问题	MAE	平均绝对误差
		MSE	均方误差
		RMSE	均方根误差
	排序问题	MAP	平均精度均值
		NDCG	归一化折损累计增益
		MRR	排序倒数均值

效率指标

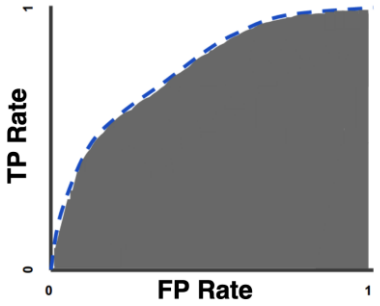
- 时间开销
  - 在线服务
  - 数据库查询
- 空间开销
  - 传感器设备
  - 移动设备
- 网络传输开销
  - 分布式数据处理平台

精度指标：分类问题

以二分类为例

预测结果	真实结果	
	正例	反例
正例	True Positive (TP)	False Positive (FP)
反例	False Negative (FN)	True Negative (TN)

- 召回率（Recall/TPR）：$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$
- 准确率（Precision/Accuracy）：$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$
- F_β值：$\mathrm{F}_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R}$
- ROC曲线下方面积（AUC）
  - ✓ $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$



分类问题：例子

- 以二分类为例

预测结果	真实结果	
	正例	反例
正例	30	20
反例	10	40

- 召回率（Recall/TPR）：$\text{Recall} = \frac{30}{30+10} = 0.75$
- 准确率（Precision/Accuracy）：$\text{Precision} = \frac{30}{30+20} = 0.6$
- F_1值：$\text{F}_1 = \frac{2PR}{P+R} = \frac{2*0.6*0.75}{0.6+0.75} = \frac{0.9}{1.35} = 0.667$

精度指标：回归问题

- 总样例 $[1,2,\ldots,N]$
- 样例$\boldsymbol{x}_i$ 预测值$\hat{y}_i$ 真实值y_i
- 平均绝对误差（MAE）
  - $\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$
- 均方误差（MSE）
  - $\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$
- 均方根误差（RMSE）
  - $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$

回归问题：例子

样例编号	1	2	3	4	5
x_i	2	4	5	7	8
y_i	6	7	10	14	17

- 假设线性回归函数为 $\hat{y} = 2x + 1$
- 则计算回归预测值 $\hat{y}_i$ 分别为 $5, 9, 11, 15, 17$

- $\mathrm{MAE} = \frac{1}{5} \cdot (1 + 2 + 1 + 1 + 0) = 1.0$
- $\mathrm{MSE} = \frac{1}{5} \cdot (1 + 4 + 1 + 1 + 0) = 1.4$
- $\mathrm{RMSE} = \sqrt{1.4} \approx 1.18$

排序问题

- 平均精度（MAP）
  - $\mathrm{AveP}(q_i) = \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{j}{r_{ij}}$
  - $\mathrm{MAP} = \frac{1}{Q}\sum_{i=1}^{Q}\mathrm{AveP}(q_i)$

第 i 个查询的第 j 个相关结果

第 j 个相关结果在第 i 个查询结果中的排名

- 归一化折损累计增益（NDCG）
  - $\mathrm{DCG}_K = \sum_{i=1}^{K}\frac{2^{rel_i}-1}{\log_2(i+1)}$
  - $\mathrm{NDCG}_K = \frac{\mathrm{DCG_K}}{\mathrm{IDCG_K}}$
- 排序倒数均值（MRR）
  - $\mathrm{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}$

第 i 个查询的第一个相关结果在其查询结果中的排名

排序问题：例子

- 对给定算法和查询 q_1, q_2, q_3
- 其中查询 q_1 有三个相关结果排名为 $2, 5, 6$；查询 q_2 有三个相关结果排名为 $1, 2, 4$；查询 q_3 有三个相关结果排名为 $3, 4, 7$
- 在该例子中， q_1 的平均精度 $\text{AveP}(q_1) = \frac{1}{3} \cdot \left(\frac{1}{2} + \frac{2}{5} + \frac{3}{6}\right) = \frac{1.4}{3} = 0.467$
- 类似的，$\text{AveP}(q_2) = \frac{1}{3} \cdot \left(\frac{1}{1} + \frac{2}{2} + \frac{3}{4}\right) = 0.917$， $\text{AveP}(q_3) = 0.42$
- $\text{MAP} = \frac{1}{3} \cdot (0.467 + 0.917 + 0.42) = 0.601$
- 在该例子中，排序倒数均值 $\text{MRR} = \frac{1}{3} \cdot \left(\frac{1}{2} + \frac{1}{1} + \frac{1}{3}\right) = 0.611$

排序问题：例子

- 对给定算法和查询 q
- 查询 q 三个最相关结果的相似度为 $1.0, 0.9, 0.8$；算法返回的 Top-3 结果相似度分别为 $0.8, 0.7, 0.9$

- 查询结果的折损累计增益 $\mathrm{DCG}_3 = \frac{2^{0.8}-1}{\log_2 2} + \frac{2^{0.7}-1}{\log_2 3} + \frac{2^{0.9}-1}{\log_2 4} = 1.568$
- 理想折损累计增益 $\mathrm{IDCG}_3 = \frac{2^{1}-1}{\log_2 2} + \frac{2^{0.9}-1}{\log_2 3} + \frac{2^{0.8}-1}{\log_2 4} = 1.917$
- 归一化折损累计增益 $\mathrm{NDCG}_3 = \frac{1.568}{1.917} = 0.818$

算法设计原则

- 提高算法效率，增强可扩展性
  - 为应对数据规模的不断增加，算法效率是一个重要因素
  - 提升算法效率可以**节约资源**、**提升用户体验**
- 避免维度灾难
  - 数据维度越高，算法的泛化能力可能越弱
  - 设计的算法需要能够**避免维度灾难**
- 提高算法处理异构数据的能力
  - 多模态数据提升模型预测能力
  - 同时也**带来数据质量问题**
- 提高算法的鲁棒性和精确性
  - 鲁棒性刻画数据扰动对算法性能的影响，如噪声、缺失值和不一致等质量的变化
  - 鲁棒性高意味着算法具有**更好的实用价值**

本章小结

- 算法涉及**数据处理全生命周期**
  - 数据采集与汇聚
  - 数据存储与管理
  - 数据分析与挖掘
  - 数据服务

- 算法设计需要综合考虑**数据处理阶段**和**数据特点**
  - 数据不同处理阶段的关注点不同
  - 数据不同特点影响算法的设计