清华大学数据库技术与应用

数据准备 I

授课教师：计算机系王健楠

授课学期：2026年（春季）

清華大學 Tsinghua University

目录

数据准备仍然是瓶颈

清華大學 Tsinghua University

2014

The New York Times

50%-80%

数据准备时间占比

“对于数据科学家来说，‘清洁工’般的繁琐工作 (Janitor Work) 是**发现数据价值的最大绊脚石**。”

2020

45%

数据准备时间占比

尽管工具在不断进步，但调查显示
数据清洗与加载依然是**耗时最多**的环节。

Reference:
[1] The New York Times (2014). For Big Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights.
[2] Anaconda (2020). State of Data Science 2020: Moving from Hype to Maturity.

为什么数据准备困难?

清華大學 Tsinghua University

数据收集

(就像去菜场买菜)

数据清洗

(就像摘菜洗菜)

数据集成

(就像切菜配盘)

数据分析

(最终的烹饪环节)

准备环节占用了绝大部分时间!

为什么数据准备困难?

数据收集

(就像去菜场买菜)

数据清洗

(就像摘菜洗菜)

数据集成

(就像切菜配盘)

数据分析

(最终的烹饪环节)

准备环节占用了绝大部分时间!

1. **问题琐碎且繁多** 比如:日期格式不统一、地址重复录入、拼写错误等"脏活累活"。
2. **人员专业能力参差不齐** 不同人员的数据科学素养和编程能力差异巨大,难以统一标准。
3. **领域特定性强** 金融、医疗、社科等不同领域的"行话"和数据规则壁垒很高。

人机协同的数据准备：三大方向

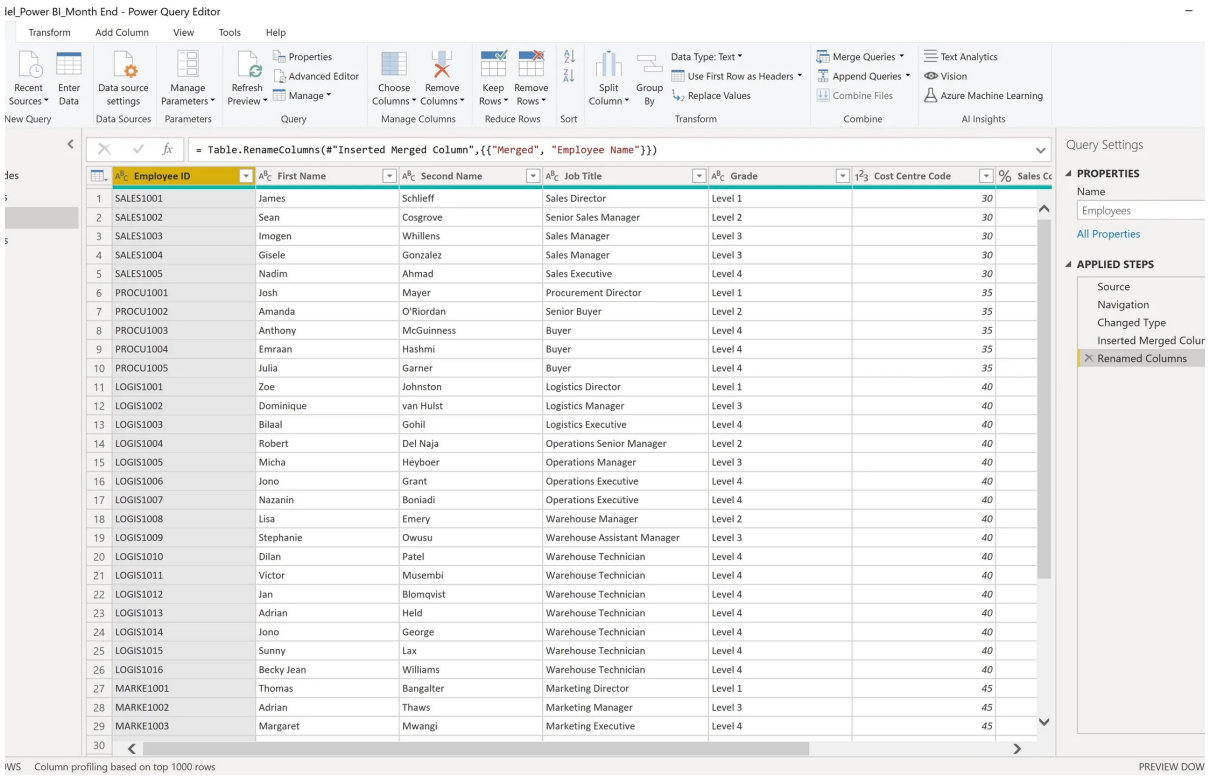
清華大學 Tsinghua University

电子表格 GUI

工作流 GUI

笔记本 GUI

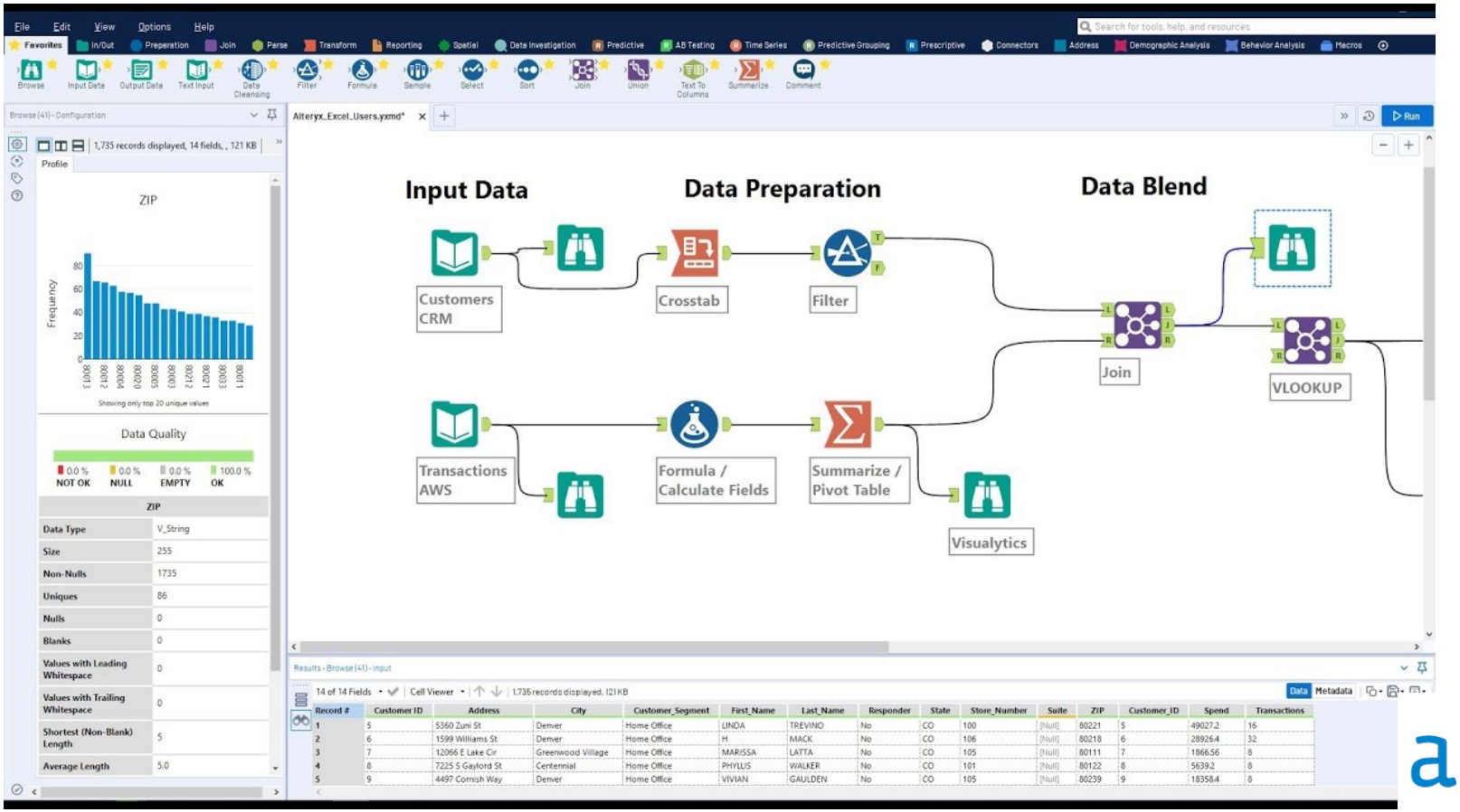
电子表格 GUI



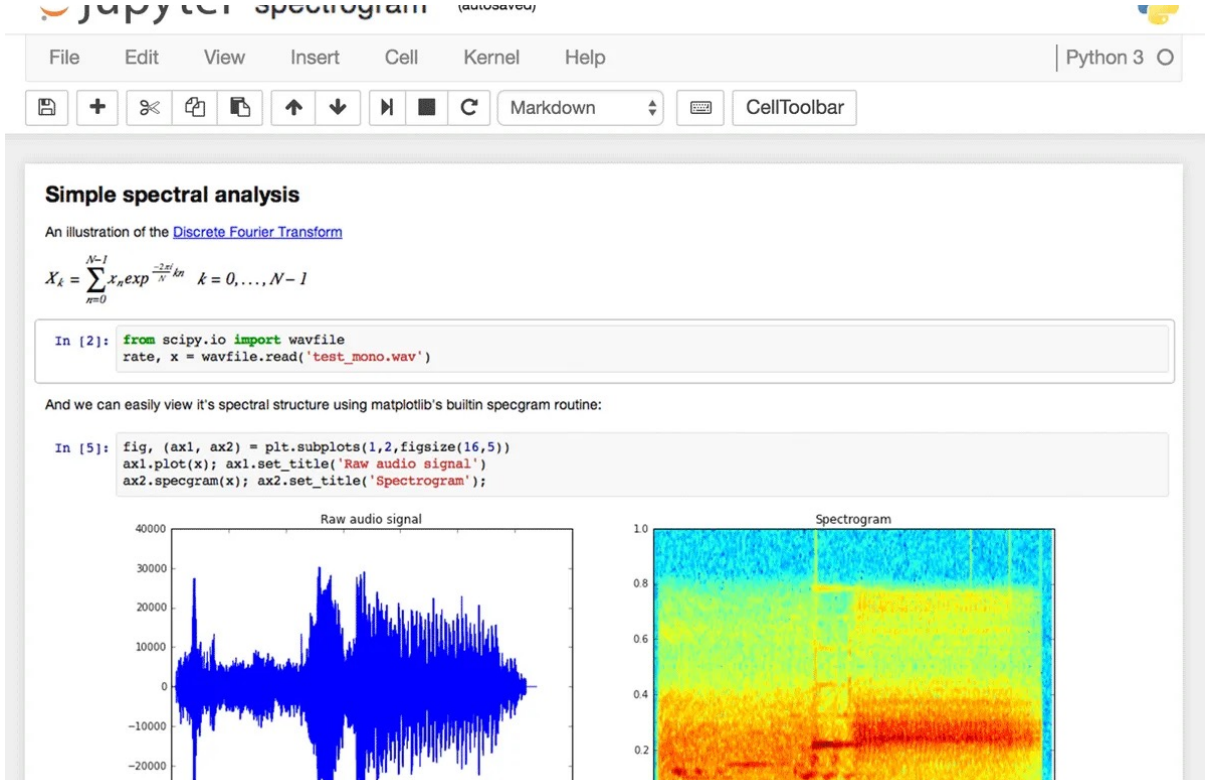
	Employee ID	First Name	Second Name	Job Title	Grade	Cost Centre Code	Sales Cc
1	SALES1001	James	Schlieff	Sales Director	Level 1	30	
2	SALES1002	Sean	Cosgrove	Senior Sales Manager	Level 2	30	
3	SALES1003	Imogen	Whillens	Sales Manager	Level 3	30	
4	SALES1004	Gisele	Gonzalez	Sales Manager	Level 3	30	
5	SALES1005	Nadim	Ahmad	Sales Executive	Level 4	30	
6	PROCU1001	Josh	Mayer	Procurement Director	Level 1	35	
7	PROCU1002	Amanda	O'Riordan	Senior Buyer	Level 2	35	
8	PROCU1003	Anthony	McGuinness	Buyer	Level 4	35	
9	PROCU1004	Emraan	Hashmi	Buyer	Level 4	35	
10	PROCU1005	Julia	Garner	Buyer	Level 4	35	
11	LOGIS1001	Zoe	Johnston	Logistics Director	Level 1	40	
12	LOGIS1002	Dominique	van Hulst	Logistics Manager	Level 3	40	
13	LOGIS1003	Bilaal	Gohil	Logistics Executive	Level 4	40	
14	LOGIS1004	Robert	Del Naja	Operations Senior Manager	Level 2	40	
15	LOGIS1005	Micha	Heyboer	Operations Manager	Level 3	40	
16	LOGIS1006	Jono	Grant	Operations Executive	Level 4	40	
17	LOGIS1007	Nazanin	Boniadi	Operations Executive	Level 4	40	
18	LOGIS1008	Lisa	Emery	Warehouse Manager	Level 2	40	
19	LOGIS1009	Stephanie	Owusu	Warehouse Assistant Manager	Level 3	40	
20	LOGIS1010	Dilan	Patel	Warehouse Technician	Level 4	40	
21	LOGIS1011	Victor	Musembi	Warehouse Technician	Level 4	40	
22	LOGIS1012	Jan	Blomqvist	Warehouse Technician	Level 4	40	
23	LOGIS1013	Adrian	Held	Warehouse Technician	Level 4	40	
24	LOGIS1014	Jono	George	Warehouse Technician	Level 4	40	
25	LOGIS1015	Sunny	Lax	Warehouse Technician	Level 4	40	
26	LOGIS1016	Becky Jean	Williams	Warehouse Technician	Level 4	40	
27	MARKE1001	Thomas	Bangalter	Marketing Director	Level 1	45	
28	MARKE1002	Adrian	Thaws	Marketing Manager	Level 3	45	
29	MARKE1003	Margaret	Mwangi	Marketing Executive	Level 4	45	
30							

图示：使用 Excel (Power Query) 进行自动化数据清洗与整理

工作流 GUI


Input Data
Data Preparation
Data Blend
Customers CRM
Crosstab
Filter
Join
VLOOKUP
Transactions AWS
Formula / Calculate Fields
Summarize / Pivot Table
Visualytics
ZIP
Data Quality

笔记本 GUI



图示：Jupyter Notebook 交互式编程环境，展示代码与可视化结果的结合

该走哪个方向?

清華大學 Tsinghua University

> *"Data Prep 市场规模在 2019 年估值为**32.9 亿美元**，预计到 2027 年将达到**181.1 亿美元**。"*

25.64%

年均复合增长率 (CAGR)

Source: Verified Market Research

三个方向	目标用户
电子表格 GUI	面向非程序员
工作流 GUI	面向非程序员
笔记本 GUI	面向数据科学家

大模型出来之后呢?

目录

数据准备任务

清華大學 Tsinghua University

1 **数据收集**

从哪里收集?

怎么收集?

2 清洗与 EDA

3 正则表达式

4 数据集成

从哪里收集数据？ — 内部数据

企业内部数据通常存储在防火墙内，是数据准备工作中最基础且最可控的来源。

- **1. 数据仓库**

  存储结构化业务数据，通常为表格形式。

- **2. 系统日志**

  记录系统运行状态，通常为文本文件。

- **3. 办公文档**

  包含丰富信息的非结构化文档 (Word / Excel / PDF)。

- **4. 多媒体数据**

  非结构化的视听资料 (视频 / 音频 / 图像)。

从哪里收集数据？ — 外部数据

外部数据源能够极大地丰富内部数据的维度，提供更广阔的视角。

- **1. 网页**

  互联网公开数据源，通过爬虫技术获取。

- **2. 应用程序接口**

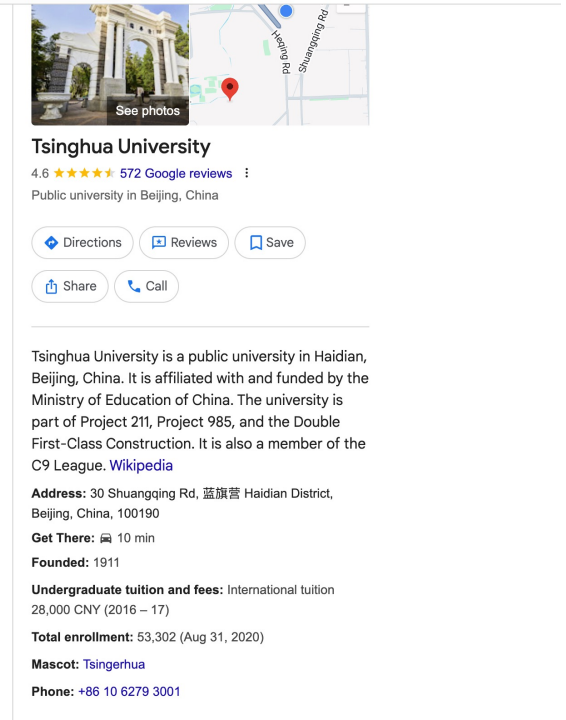
  提供规范的结构化数据访问通道，比爬虫更稳定。

- **3. 开放数据**

  政府、科研机构发布的公共数据集，格式规范。

- **4. 知识图谱**

  结构化的语义知识库，用于增强数据的关联性与解释性。

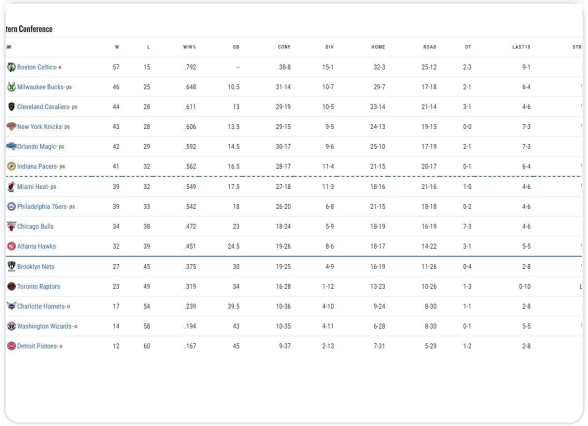


数据分类

Tsinghua University

数据

结构化数据



tern Conference

	W	L	WIN%	GB	CONF	DIV	HOME	ROAD	OT	LAST10	STR
Boston Celtics- e	57	15	.792	–	38-8	15-1	32-3	25-12	2-3	9-1	
Milwaukee Bucks- pe	46	25	.648	10.5	31-14	10-7	29-7	17-18	2-1	6-4	
Cleveland Cavaliers- pe	44	28	.611	13	29-19	10-5	23-14	21-14	3-1	4-6	
New York Knicks- pe	43	28	.606	13.5	29-15	9-5	24-13	19-15	0-0	7-3	
Orlando Magic- pe	42	29	.592	14.5	30-17	9-6	25-10	17-19	2-1	7-3	
Indiana Pacers- pe	41	32	.562	16.5	28-17	11-4	21-15	20-17	0-1	6-4	
Miami Heat- pe	39	32	.549	17.5	27-18	11-3	18-16	21-16	1-0	4-6	
Philadelphia 76ers- pe	39	33	.542	18	26-20	6-8	21-15	18-18	0-2	4-6	
Chicago Bulls	34	38	.472	23	18-24	5-9	18-19	16-19	7-3	4-6	
Atlanta Hawks	32	39	.451	24.5	19-26	8-6	18-17	14-22	3-1	5-5	
Brooklyn Nets	27	45	.375	30	19-25	4-9	16-19	11-26	0-4	2-8	
Toronto Raptors	23	49	.319	34	16-28	1-12	13-23	10-26	1-3	0-10	L
Charlotte Hornets- o	17	54	.239	39.5	10-36	4-10	9-24	8-30	1-1	2-8	
Washington Wizards- o	14	58	.194	43	10-35	4-11	6-28	8-30	0-1	5-5	
Detroit Pistons- e	12	60	.167	45	9-37	2-13	7-31	5-29	1-2	2-8	

NBA 球队排名表 (Table)

半结构化数据


```
{
"game_id": "0022300567",
"home_team": "LAL",
"away_team": "GSW",
"play_by_play": [
{
"clock": "10:45",
"event": "JUMP_BALL",
"player": "A. Davis",
"coordinates": {"x": 50, "y": 50}
},
```


比赛数据日志 (JSON)

非结构化数据

新闻报道文章 (Text/Web)

数据收集的挑战

清華大學 Tsinghua University

- **1. 数据发现**

  问题： 如何找到相关数据?

  方案： 需要掌握领域业务知识，并运用高级检索技能。

- **2. 数据隐私**

  问题： 如何保护用户隐私?

  方案： 采用数据脱敏处理，结合差分隐私技术防止反推。

- **3. 安全性**

  问题： 如何避免数据泄露?

  方案： 通过严格的访问控制权限与数据加密技术保障安全。

获取数据

清華大學 Tsinghua University

- **1. 从CSV文件** 通用的表格数据格式，易于交换与读取。
- **2. 从JSON文件** 半结构化嵌套数据，网络传输标准。
- **3. 从网页** 通过爬虫获取公开互联网信息。
- **4. 从应用程序接口** 标准化的数据交互。
- **5. 从数据库** 关系型或非关系型数据库。
- **6. 从分布式文件系统** 大数据存储基础。
- **7. 从云存储** 对象存储，海量数据湖。

从CSV文件加载数据

清華大學 Tsinghua University

- **1. 什么是 CSV?**

  即逗号分隔值文件。用逗号分隔列，换行分隔行。简单轻量，所有工具都支持。

- **2. 注意事项**
  - **编码问题**：需指定正确编码（如 UTF-8）。
  - **分隔符**：有时需调整参数适配分号或制表符。
  - **缺失值**：可自动识别空值或自定义标记。
  - **数据类型**：指定类型可优化内存占用。

read_csv.py

```python
import pandas as pd

# 1. 基本读取方法
df=pd.read_csv('data.csv')

# 2. 处理常见情况：指定头部、分隔符、编码
df=pd.read_csv(
    'data_raw.txt',
    header=0,  # 以此行为列名
    sep=',',  # 指定分隔符
    encoding='utf-8' # 防止中文乱码
)
```

从 JSON 文件加载数据

清華大學 Tsinghua University

1

什么是 JSON?

JSON (JavaScript Object Notation) 是一种轻量级的数据交换格式，易于人阅读和编写，同时也易于机器解析和生成。常用于存储嵌套数据（如数组、字典）。

2

结构特点

具有自描述性，元数据与记录通常包含在同一文件中。支持复杂的嵌套结构，如列表中的字典。

3

注意事项

非矩形结构 (Non-rectangular) 与 字段不一致问题。直接读取可能无法得到平铺的表格，需要进行数据展平 (Flattening) 处理。

```
{ "name": "姚明", "info": { "height": 2.26,
"team": "火箭" } }
```

```python
importpandasaspd
importjson

# 方法 1: 直接读取标准 JSON 文件
df=pd.read_json('data.json')

# 方法 2: 处理嵌套结构（先加载后转换）
withopen('data.json','r')asf:
data=json.load(f)

# 将嵌套字典转换为 DataFrame
df=pd.DataFrame(data)

# 如果结构非常复杂，可以使用 json_normalize
frompandasimportjson_normalize
df_flat=json_normalize(data)

# 输出结果示例
print(df_flat.head())
# Output:
# name info.height info.team
# 0 姚明 2.26 火箭
```

网页爬取工具概览

清華大學 Tsinghua University

01. 发起请求 (Requests)

目标： 向网站发送请求获取网页内容。

工具： requests (简单易用，Python首选)

02. 解析网页 (Parsing)

目标： 从 HTML 文件中提取需要的数据。

工具： Beautiful Soup (容错性高)

03. 集成框架 (Framework)

目标： 完整的爬虫解决方案(请求+解析+存储)。

工具： Scrapy (功能强大，适合大规模爬取)

scraper.py

```
importrequests
frombs4importBeautifulSoup

# 1. 发起请求：获取网页源代码
url='https://news.tsinghua.edu.cn'
resp=requests.get(url)
resp.encoding='utf-8'

# 2. 解析网页：提取数据
soup=BeautifulSoup(resp.text,'html.parser')

# 提取所有 h3 标签的新闻标题
titles=soup.find_all('h3',class_='news-title')

# 3. 数据处理：打印结果
fortitleintitles:
print(title.text.strip())
```

```
> Output:
清华大学举行2025年开学典礼...
计算机系团队获得国际大奖...
```

开始爬取前

清華大學 Tsinghua University

- **1. 寻找结构化文件**

  检查： 页面是否提供 CSV / JSON / XML 版本的下载链接。

  例子： 国家统计局官网直接提供 Excel 格式的统计数据下载，无需爬取 HTML 页面。

- **2. 检查现成工具库**

  检查： 是否有现成的 Python 库提供数据访问接口。

  例子： 获取股票数据可以使用 yfinance 库，获取天气数据可用专门 API，比自己写爬虫更稳定。

- **3. 确认抓取权限**

  检查： 务必检查 robots.txt 文件和服务条款 (Terms of Service)。

  例子： 访问淘宝网的 /robots.txt 可看到哪些路径允许爬取，违反规则可能面临法律风险。

爬取时的注意事项

清華大學 Tsinghua University

- **1. 控制请求频率**

  说明： 避免给目标网站造成过载，防止被封禁。

  例子： 每次请求后等待 1-2 秒，模拟人类访问行为，切勿短时间高频请求。

- **2. 小规模测试**

  说明： 先在少量请求上测试代码逻辑，确保解析规则正确。

  例子： 先爬取 5 条数据验证代码无误，再扩大到全部数据范围。

- **3. 保存中间结果**

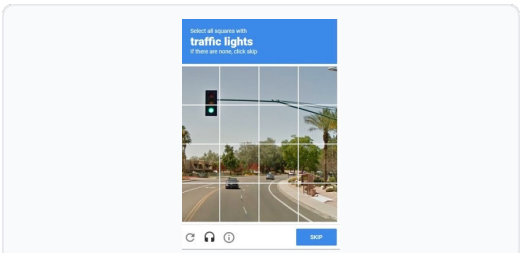
  说明： 本地保存每次请求的结果，避免程序中断后从头开始。

  例子： 爬取 1000 页数据时，每爬 100 页就自动保存一次文件。

- **4. 处理验证码**

  说明： 留意验证码等反爬机制。

  例子： 如右图所示，遇到此类 CAPTCHA 验证时，程序通常无法自动通过，需要人工干预或调整爬取策略。



数据准备任务

清華大學 Tsinghua University

1 数据收集

2 **清洗与 EDA**

3 正则表达式

4 数据集成

探索性数据分析 (EDA)

清華大學 Tsinghua University

EDA 就是对数据进行的拆箱体验！

理解全貌 → **发现洞见** → **支撑决策**

案例分析：学生成绩数据

1. 基础概况检查

查看数据集的规模与结构：

- 共有多少名学生?
- 包含多少个科目?
- 成绩分布范围是否合理（如 0-100 分）？

2. 模式与异常识别

深入挖掘数据特征：

- 发现数学科目平均分明显低于其他科目。
- 存在个别异常高分或低分（离群值）。

3. 提出行动建议

基于数据制定策略：

- 建议加强数学教学资源的投入。
- 调查异常成绩的具体原因（如是否存在作弊或录入错误）。

EDA：数据的开箱时刻

清華大學 Tsinghua University

买了新 iPhone

第一件事：开箱 (Unboxing)

- 检查外观是否完好
- 测试功能是否正常
- 确认配件是否齐全

=

拿到新数据

第一件事：EDA

- 查看数据结构和字段
- 检查数据质量和完整性
- 发现缺失值和异常值

EDA中的关键数据属性

清華大學 Tsinghua University

1. 结构

数据文件的“形状”

2. 粒度

每一条记录的粗细程度

3. 时序性

数据在时间维度上的表示

4. 真实性

数据捕捉现实的程度

矩形数据

表格 / 数据框

每一列 = 字段 / 属性 / 特征

姓名	年龄	城市	工资
张三	25	北京	15000
李四	30	上海	18500
王五	28	深圳	16200

每一行 = 记录 / 观察

具名列，多类型 (Mixed Types)

每一列包含不同类型的数据（如：字符串、整数、浮点数、日期等）。

适用于数据清洗、探索性分析 (EDA) 与业务报表。

矩阵

$$\begin{bmatrix} 1.0 & 0.5 & 0.2 \\ 0.5 & 1.0 & 0.8 \\ 0.2 & 0.8 & 1.0 \\ 0.1 & 0.3 & 0.9 \end{bmatrix}$$

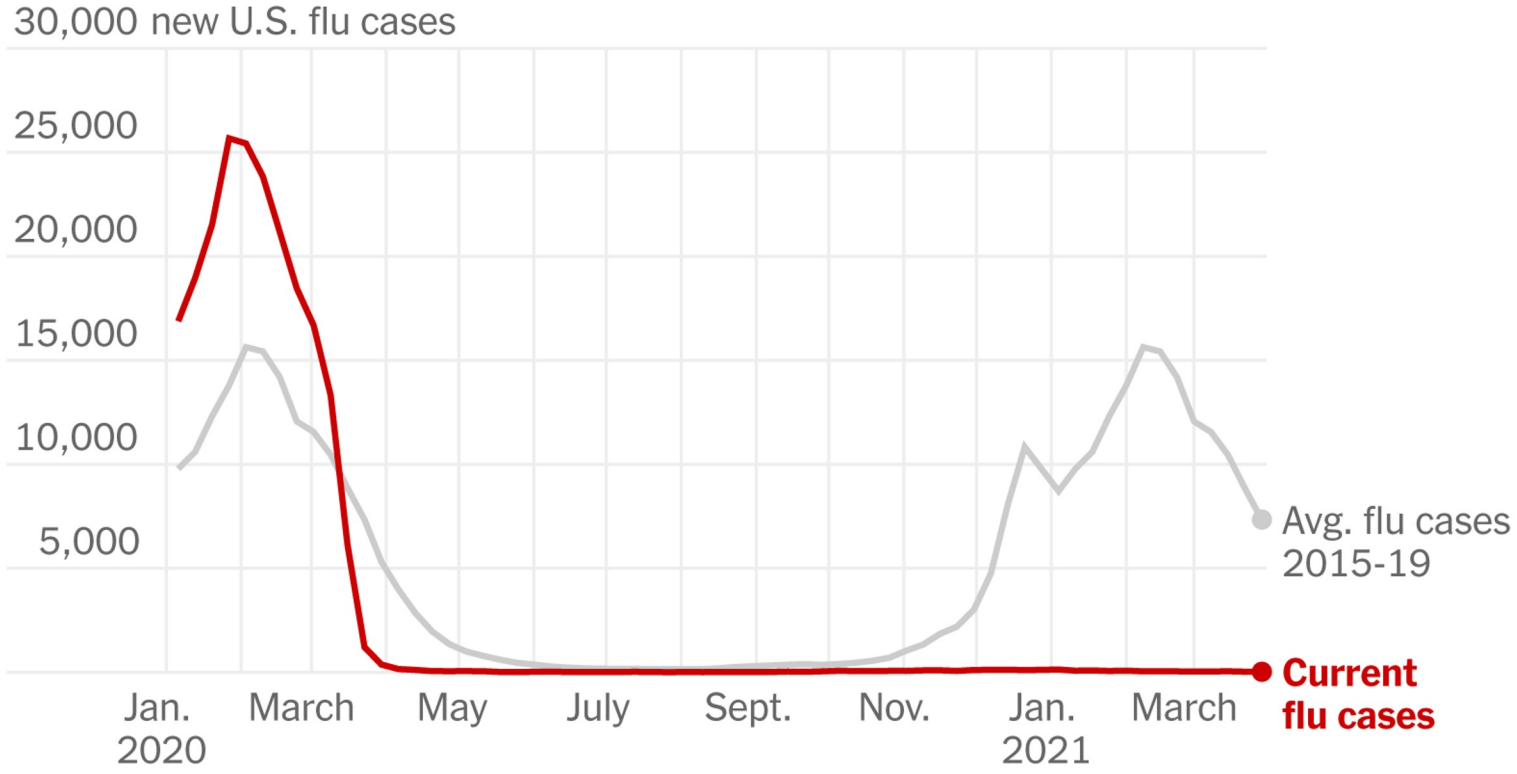
同质数值类型 (Homogeneous)

所有元素必须是相同的数据类型（通常全部为数值 float/int）。

主要用于线性代数计算与建模，计算速度极快，但缺乏语义。

COVID-19疫情期间的其他疾病

为什么流感在 2021 年 3 月份消失了?



Source: New York Times

CDC分析示例：数据长什么样?

角色设定： 你是 CDC（疾控中心）的数据分析师，任务是计算美国各州的 TB（结核病）发病率。

TB incidence†		
2019	2020	2021
2.71	2.16	2.37

$$\text{TB 发病率} = \frac{\text{病例数}}{\text{人口数}} \times 100{,}000$$

计算发病率需要的两类数据

清華大學 Tsinghua University

TB 病例数

```
[7]: tb_df = pd.read_csv("data/cdc_tuberculosis.csv",)
     tb_df
```

[7]:

	Unnamed: 0	No. of TB cases	Unnamed: 2	Unnamed: 3
0	U.S. jurisdiction	2019	2020	2021
1	Total	8,900	7,173	7,860
2	Alabama	87	72	92
3	Alaska	58	58	58
4	Arizona	183	136	129
...	...	...	...	...
48	Virginia	191	169	161
49	Washington	221	163	199

来源：CDC 数据源

包含各州、各年份的确诊病例统计。

人口数据

州名 (State)	年份 (Year)	人口数 (Pop)
Alabama	2019	4,903,185
Alaska	2019	731,545
Arizona	2019	7,278,717
Arkansas	2019	3,017,804
California	2019	39,512,223

来源：美国人口普查局

提供各州对应年份的人口普查数据。

其他数据格式

除表格外，现代数据科学还需要处理大量非结构化数据。这些数据在原始形式上差异巨大，但都可以通过特征提取转化为矩阵或表格形式。

- **1. 图像**

  应用场景：医学影像诊断、自动驾驶视觉感知。

- **2. 音频**

  应用场景：语音识别、情感分析与声纹识别。

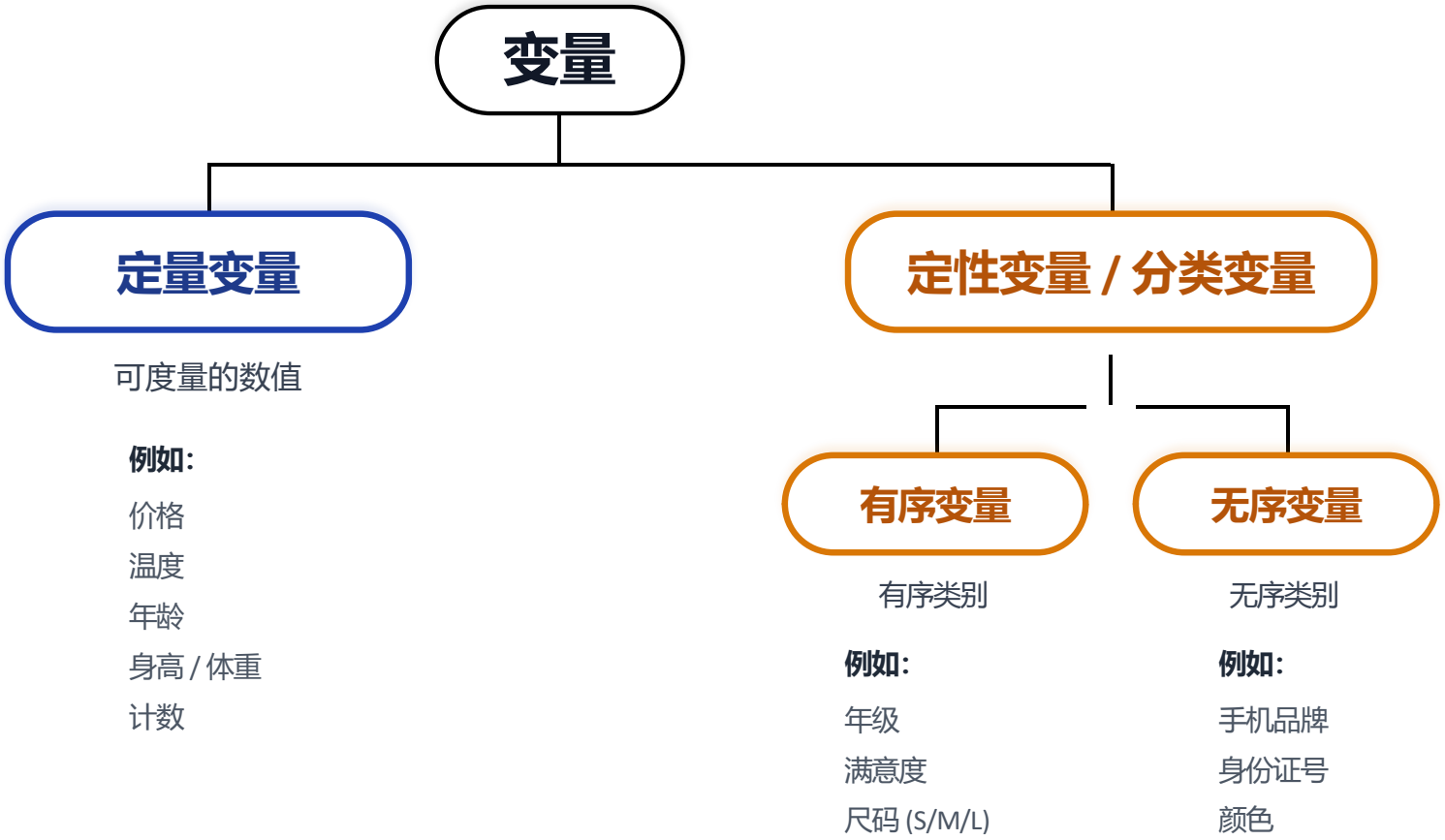
- **3. 视频**

  应用场景：目标跟踪与行为分析、人脸识别与安防监控。

- **4. 文本**

  应用场景：大语言模型、法律文档审查与信息抽取。

变量特征类型


变量
定量变量
可度量的数值
例如:
价格
温度
年龄
身高 / 体重
计数
定性变量 / 分类变量
有序变量
有序类别
例如:
年级
满意度
尺码 (S/M/L)
无序变量
无序类别
例如:
手机品牌
身份证号
颜色

为什么要识别变量类型?

清華大學 Tsinghua University

识别变量类型是选择正确的分析方法、避免得出错误结论的前提。

1

决定可视化方法

不同类型的变量需要不同的图表来展示其分布特征。
定量变量 → 直方图 / 散点图；定性变量 → 条形图 / 饼图。

变量	类型	推荐图表
年龄	定量	直方图 (Histogram)
性别	定性	饼图 (Pie Chart)

2

决定统计分析方法

变量类型决定了我们可以进行哪些数学运算。
定量变量 → 均值 / 标准差；定性变量 → 频数 / 众数。

变量	类型	适用统计量
工资	定量	平均值、涨幅
学历	有序	人数占比、众数

3

避免错误分析

误把定性变量当做定量变量处理，会产生毫无意义的计算结果。

变量	类型	错误操作示例
身份证号	定性	计算"平均身份证号"
邮政编码	定性	对邮编求和

变量类型识别练习

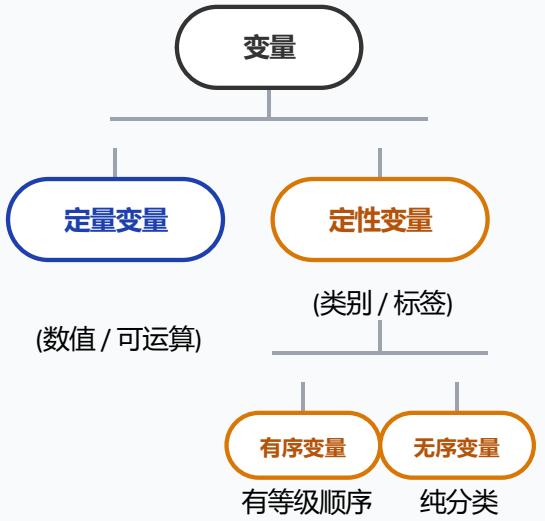
清華大學 Tsinghua University

请对以下变量进行分类：

思考它们是定量还是定性？若为定性，是否有序？

1. **CO2 浓度 (ppm)**
   例如：400ppm, 420.5ppm
2. **收入分档**
   例如：低收入、中等收入、高收入
3. **种族**
   例如：汉族、回族、维吾尔族
4. **政党**
   例如：民主党、共和党
5. **年份**
   例如：2020, 2021, 2022
6. **GPA 绩点**
   例如：3.5, 3.8, 4.0

参考逻辑树



小贴士：

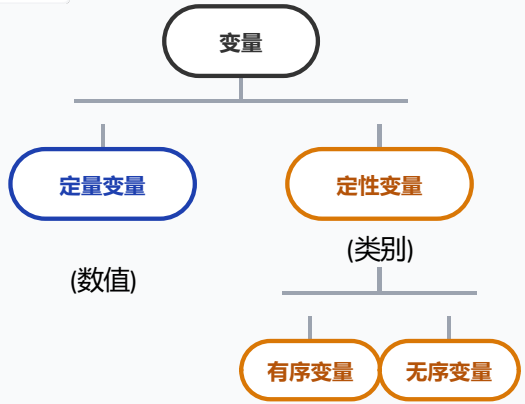
判断关键在于：计算平均值是否有意义？是否有等级顺序？

变量类型识别练习 - 答案解析

清華大學 Tsinghua University

1. **CO2 浓度 (ppm)**
   400ppm, 420.5ppm — 定量变量
2. **收入分档**
   低 / 中 / 高收入 — 有序变量
3. **种族**
   汉族、回族... — 无序变量
4. **政党**
   民主党、共和党 — 无序变量
5. **年份**
   2020, 2021... — 定量 或 有序 (情境依赖)
6. **GPA 绩点**
   3.5, 3.8... — 定量变量
7. **日期时间**
   2024-01-15... — 定量 或 有序 (情境依赖)

参考逻辑树



重点提示：
“年份”和“日期”根据分析目的不同（计算间隔 vs 分组趋势），可作为定量或有序变量处理。

EDA中的关键数据属性

清華大學 Tsinghua University

1. **结构**
数据文件的“形状”

2. 粒度

每一条记录的粗细程度

3. **时序性**
数据的时间维度特性

4. **真实性**
数据是否准确反映现实

粒度对比说明：细粒度 vs 粗粒度

细粒度

包含每个学生的详细成绩，数据未被折叠

班级	姓名	科目	分数
一班	张三	数学	85
一班	张三	英语	92
二班	李四	数学	78
二班	李四	英语	88

粗粒度

按班级维度进行聚合，计算平均分

班级	平均分
一班	88.5
二班	83.0

注：这是对左侧数据的汇总

TB (结核病) 数据的粒度分析

清華大學 Tsinghua University

思考：每一行代表什么？

并非所有行都具有相同的粒度 (Granularity)。这种混合层级结构在公共卫生数据中非常常见。

示例数据: df_tb_cases.csv

#	State	County	Cases	粒度
0	**US Total**	*NaN*	**8,900**	COARSE
1	**Alabama**	**Total**	**87**	MEDIUM
2	Alabama	Autauga County	12	FINE
3	Alabama	Baldwin County	5	FINE
4	**Alaska**	**Total**	**58**	MEDIUM

国家级汇总　州级汇总　县级明细

EDA中的关键数据属性

1. 结构 数据文件的“形状”

2. 粒度 每一条记录的粗细程度

3. 时序性

数据的时间维度特性

4. 真实性 数据是否准确反映现实

高效存储日期时间

我们通常习惯将日期时间写成字符串形式，例如：**01/01/2025 3:30pm**

这个字符串包含了 **17 个字符**。

如果是字符串存储...

假设有一个包含 10 亿条记录的日期时间列

~ 170 亿个字符

17GB

存储占用

⬇️

如果改用整数存储?

同样 10 亿条记录，存储为整数格式

~ 40 亿字节

4 GB

大幅节省空间!

时序性：Unix / POSIX 时间

清華大學 Tsinghua University

- **1. 定义与原点**

  日期时间以 **秒** 为单位进行度量。
  起始时间（Epoch）为 **1970年1月1日 UTC**（协调世界时）。

- **2. 转换示例**

  2025年6月30日 11:00 → 对应 **1751306400** 秒

  1950年6月30日 11:00 → 对应 **-615535200** 秒 (负数表示1970之前)

- **3. 数值表示的优势**

  可以直接进行高效的数学运算。
  例如：通过减法和除法计算两个日期之间相隔的天数。

```
( t2 - t1 ) / 86400
```

Pandas 日期时间转换

清華大學 Tsinghua University

1 **核心功能**

使用 **pd.to_datetime()** 将字符串转换为 Pandas 的 Timestamp 对象，这是时间序列分析的第一步。

2 **关键参数**

format：指定输入字符串的日期格式（如 '%Y-%m-%d'）
errors：处理解析错误（'coerce' 将无法解析的值设为 NaT）

3 **应用场景**

电商订单数据中，'order_time' 通常以字符串形式存储（如 "2025-01-12 14:30"），需转换为 datetime 类型以便进行时间运算。

```
importpandasaspd

# 1. 创建示例电商订单数据

df=pd.DataFrame({
'order_id':[101,102,103],
'order_time':['2025-01-12 14:30',
'2025-01-12 15:45',
'2025-01-13 09:15']
})

# 2. 转换前：查看数据类型
print(df['order_time'].dtype)# object (string)

# 3. 核心转换：字符串 -> datetime
df['order_time']=pd.to_datetime(df['order_time'])

# 4. 转换后：再次查看数据类型
print(df['order_time'].dtype)
```

Pandas 日期时间组件提取

1. **dt.date**
   提取具体的日期部分（年-月-日），去除时间信息。

2. **dt.dayofweek**
   提取星期几的数值表示（0 代表周一，6 代表周日）。

3. **dt.hour**
   提取具体的小时数（0-23），常用于分析日内趋势。

4. **dt.month** / **dt.year**
   分别提取月份和年份，用于按月或按年聚合分析。

extract_components.py

```python
# 使用 dt 访问器提取组件
df['date']=df['order_time'].dt.date
df['weekday']=df['order_time'].dt.dayofweek
df['hour']=df['order_time'].dt.hour

# 输出结果示例
print(df.head(3))

# Output:
# order_time date weekday hour
# 0 2025-01-12 14:30 2025-01-12 6 14
# 1 2025-01-12 09:15 2025-01-12 6 9
# 2 2025-01-13 18:45 2025-01-13 0 18
```

Pandas 日期时间运算与格式化

清華大學 Tsinghua University

1 **时间差计算**

两个日期时间对象相减，会自动得到 timedelta 类型，表示时间间隔。

2 **时间加减运算**

可以使用 pd.Timedelta 对日期时间进行加减操作，如增加天数、小时等。

3 **格式化输出**

使用 dt.strftime() 方法可以将日期时间转换为指定格式的字符串，支持自定义格式（如中文日期）。

```
importpandasaspd
curr_time=pd.Timestamp.now()

# 1. 计算时间差：处理时长
#（当前时间 - 订单时间）
df['duration']=curr_time-df['order_time']

# 2. 时间加减：预计送达时间（订单时间 + 3天）
df['estimated']=df['order_time']+
pd.Timedelta(days=3)

# 3. 格式化输出：中文日期格式
df['del_str']=df['estimated'].dt.strftime(
'%Y年%m月%d日')

# 输出结果示例:
# 0 2023-11-11 10:00:00 12 days 04:30:15
# 2023年11月14日
```

EDA中的关键数据属性

清華大學 Tsinghua University

1. **结构** 数据文件的“形状”

2. **粒度** 每一条记录的粗细程度

3. **时序性** 数据的时间维度特性

4. 真实性

数据是否准确反映现实

课堂互动：这个数据集有哪些潜在问题？

清華大學 Tsinghua University

电商交易数据样本

ID	CATEGORY	STATE	LOCATION	DEVICE	PURCHASED
0	Shoes	CA	CA	1	1
1	Socks	NM	NM	1	0
2	Socks	XY	XY	1	0
3	Shirts	NY	NY	1	NA
4	Shoes	FL	FL	1	0
4	Shoes	FL	FL	1	0
5	Shirts	CA	CA	1	0
6	Pnts	TX	TX	1	1
7	Hats	CA	CA	1	-1

请观察左侧数据

这是一个看似普通的销售记录表，但其中包含了一些常见的“脏数据”问题。请尝试找出它们。

- ? 是否有不合法的类别或拼写错误？
- ? 是否存在重复的记录？
- ? 是否有缺失值或无意义的数值？

数据质量问题答案

清華大學 Tsinghua University

1 **拼写错误 (Typo)** 第 6 行 · Category 字段

值为 "Pnts"，应为 "Pants"。

影响：会导致分类统计不准确，将同一类商品视为两类。

2 **重复记录 (Duplicate)** 第 4 行 & 第 5 行

ID 均为 4，且所有字段内容完全相同。

影响：会导致销售额、订单量等统计指标虚高（Double Counting）。

3 **缺失值 (Missing Value)** 第 3 行 · Purchased 字段

值为 NA，表示数据缺失。

影响：无法判断该用户是否购买，需决定删除该行或填充默认值（如0）。

4 **异常值 (Outlier)** 第 7 行 · Purchased 字段

值为 -1，不符合业务逻辑（购买状态通常为 0 或 1）。

影响：可能是数据录入错误或系统故障，需修正或剔除。

缺失数据处理方法

清華大學 Tsinghua University

- **1. 保留空值**

  最安全、最透明的处理方式，避免引入人为偏差。

- **2. 删除记录**

  需谨慎使用，直接丢弃数据会丢失信息，仅在随机缺失且数据量大时考虑。

- **3. 插补 / 填值**

  使用推荐值填补缺失，需基于合理的业务假设。常用方法包括：

  - 简单填充（均值 / 中位数 / 众数）
  - 模型预测（使用回归等预测缺失值）
  - 热卡填充（从相似记录中随机抽取）
  - 多重插补（多次随机插补并汇总）

数据清洗工具

清華大學 Tsinghua University

Python

- 缺失数据处理
- 去重

OpenRefine

- 开源软件
- OpenRefine 服务 (**RefinePro**)

Data Wrangler

- 斯坦福/伯克利研究项目
- 商业化 (**Trifacta**)

dataprep

http://dataprep.ai

```
import pandas as pd
from dataprep.clean import clean_country

df = pd.DataFrame({"country": ["USA", "country: Canada", " France ", "233", " tr "]})

clean_country(df, "country")
```

	country	country_clean
0	USA	United States
1	country: Canada	Canada
2	France	France
3	233	Estonia
4	tr	Turkey

Dataprep.clean 自动清洗效果示例

异常值检测

美国公司员工年龄数据案例

清華大學 Tsinghua University

原始数据集 (员工年龄):

1 20 21 21 22 26 33 35 36 37 39 42 45 47 54 57 61 62

均值 (Mean)

$\mu = \frac{1}{n} \sum x_i = 37$

样本标准差 (Sample Stdev)

$s = \sqrt{\frac{1}{n-1} \sum (x_i - \mu)^2} \approx 16$

异常值判定区间 : [μ - 2s, μ + 2s] = [37 - 32, 37 + 32] = **[5, 69]**

异常值检测

美国公司员工年龄数据案例 (包含极端值)

清華大學 Tsinghua University

原始数据集 (员工年龄):

1 20 21 21 22 26 33 35 36 37 39 42 45 47 54 57 61 62 400

均值 (Mean)

$\mu = 1/n \sum x_i = \mathbf{56}$

样本标准差 (Sample Stdev)

$s = \sqrt{(1/(n-1) \sum (x_i - \mu)^2)} = \mathbf{83}$

异常值判定区间 : [μ - 2s, μ + 2s] = [56 - 166, 56 + 166] = **[-110, 222]**

异常值检测

中位数 (Median) 与 绝对中位差 (MAD)

清華大學 Tsinghua University

原始数据集 (含极端值):

1 20 21 21 22 26 33 35 36 37 39 42 45 47 54 57 61 62 400

中位数 (Median)

$$\text{Median} = \text{median}(x_i) = \mathbf{37}$$

不受极端值 (1, 400) 影响

绝对中位差 (MAD)

$$\text{MAD} = \text{median}(|x_i - \text{Median}|) = \mathbf{15}$$

鲁棒的标准差估计量

异常值判定区间 : [Median - 2 × MAD, Median + 2 × MAD] = [7, 67]

结论: 极端值 1 和 400 均落在区间外，被正确识别为异常值

数据准备任务

清華大學 Tsinghua University

1. 数据采集
2. 清洗与 EDA
3. 正则表达式
4. **数据集成**

数据集成：三个步骤

模式映射

- 创建全局模式
- 将局部模式映射到全局模式

实体解析

- 稍后将详细学习

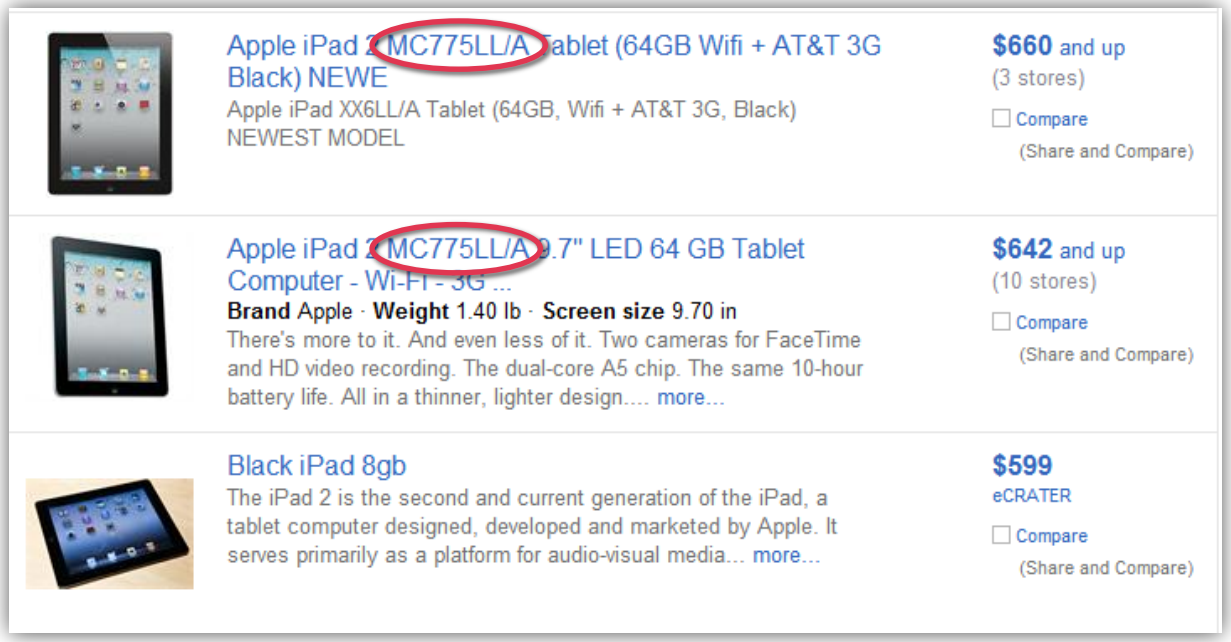
数据融合

- 基于置信度分数解决冲突

想了解更多?

- Anhai Doan, Alon Y. Halevy, Zachary Ives. Principles of Data Integration. Morgan Kaufmann Publishers, 2012.

实体解析


Apple iPad 2 MC775LL/A Tablet (64GB Wifi + AT&T 3G Black) NEWE
Apple iPad XX6LL/A Tablet (64GB, Wifi + AT&T 3G, Black) NEWEST MODEL
$660 and up
(3 stores)
Compare
(Share and Compare)
Apple iPad 2 MC775LL/A 9.7" LED 64 GB Tablet Computer - Wi-Fi - 3G ...
Brand Apple · Weight 1.40 lb · Screen size 9.70 in
There's more to it. And even less of it. Two cameras for FaceTime and HD video recording. The dual-core A5 chip. The same 10-hour battery life. All in a thinner, lighter design.... more...
$642 and up
(10 stores)
Compare
(Share and Compare)
Black iPad 8gb
The iPad 2 is the second and current generation of the iPad, a tablet computer designed, developed and marketed by Apple. It serves primarily as a platform for audio-visual media... more...
$599
eCRATER
Compare
(Share and Compare)

实体解析的输出

ID	产品名称	价格
r_1	iPad Two 16GB WiFi White	$490
r_2	iPad 2nd generation 16GB WiFi White	$469
r_3	iPhone 4th generation White 16GB	$545
r_4	Apple iPhone 3rd generation Black 16GB	$375
r_5	Apple iPhone 4 16GB White	$520

(r_1, r_2) , (r_3, r_5)

实体解析技术

基于相似函数的方式

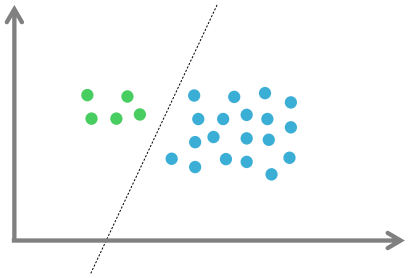
- 相似函数 (e.g., $\text{Jaccard}(r, s) = |\frac{r \cap s}{r \cup s}|$)
- 阈值 (e.g., 0.8)

Jaccard(r1, r2) = 0.9 ≥ 0.8　匹配

Jaccard(r4, r8) = 0.1 < 0.8　不匹配

基于学习的方式

- 把一对记录表示成一个特征向量



基于相似度的方法

假设相似度函数为 Jaccard。

问题定义

给定一个表 T 和一个阈值，问题的目标是找到所有满足以下条件的记录对$\theta(r,s) \in T \times T Jaccard(r,s) \geq \theta$

Naïve解法需要n^2

过滤与验证

步骤一：过滤

- 去除明显不相似的记录对

步骤二：验证

- 仅对保留下来的记录对计算 Jaccard 相似度

过滤是如何工作的?

什么是"明显不相似的记录对"?

- 如果两条记录不共享任何词，则它们是明显不相似的。
- 在这种情况下，它们的 Jaccard 相似度为零，因此不会作为结果返回，可以安全地过滤掉。

如何高效地返回至少共享一个词的记录对?

- 为了帮助大家理解解决方案，我们先考虑问题的简化版本，即假设每条记录只包含一个词

简化版本

假设每条记录只有一个词。编写一个 pandas查询来完成过滤。

r_1	Apple
r_2	Apple
r_3	Banana
r_4	Orange
r_5	Banana

输出： (r1, r2), (r3, r5)

```
df.merge(df.rename(columns={'id':'id2'}), on='word').query('id < id2')
```

需要 n^2 比较吗？

一般情况

假设每条记录可以有多个词。

r_1	Apple, Orange
r_2	Apple
r_3	Banana
r_4	Orange, Apple
r_5	Banana

r_1	Apple
r_1	Orange
r_2	Apple
r_3	Banana
r_4	Orange
r_4	Apple
r_5	Banana

1. 这个新表可以看作是旧表的 倒排索引。
2. 在新表上运行之前的 SQL 并去除冗余的记录对。

对效率不满意?

探索更强的过滤条件

- 过滤共享零个词元的记录对
- 过滤共享 一 个词元的记录对
- ....
- 过滤共享 k 个词元的记录对

挑战

- 如何为这些更强的条件开发高效的过滤算法?

Jiannan Wang, Guoliang Li, Jianhua Feng.
Can We Beat The Prefix Filtering? An Adaptive Framework for Similarity Join and Search.
SIGMOD 2012:85-96.

对结果质量不满意?

TF-IDF

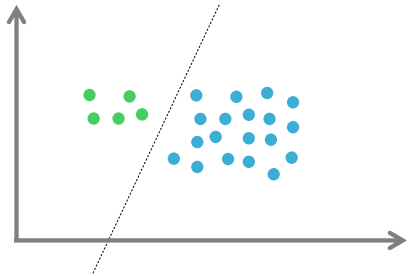
- 使用加权 Jaccard：$\mathrm{WJaccard}(r,s)=\frac{\mathrm{wt}(r\cap s)}{\mathrm{wt}(r\cup s)}$

众包

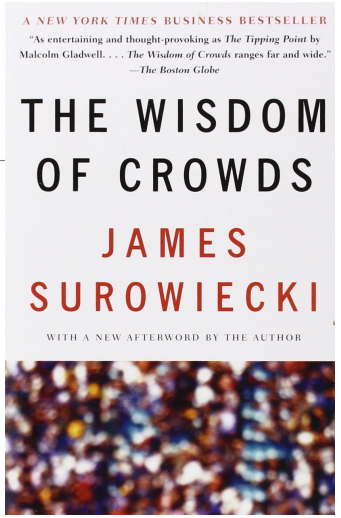
- 让人来判断两条记录是否匹配

基于学习的方法

- 将实体解析建模为分类问题



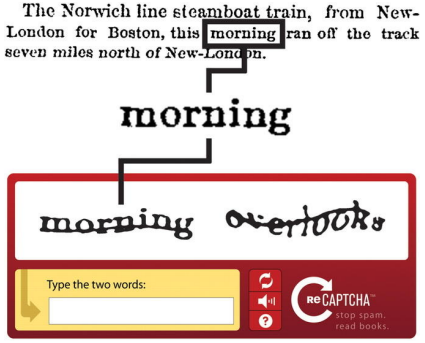
群体智慧



这意味着什么?

- 三个臭皮匠，赛过诸葛亮

一些著名的例子

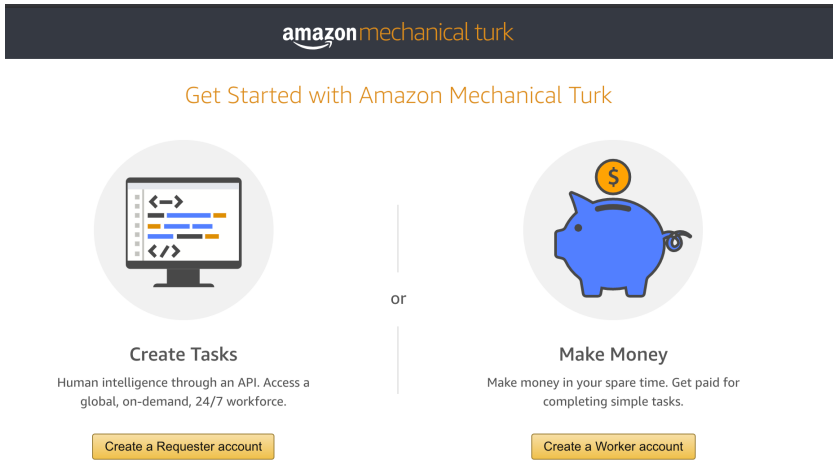


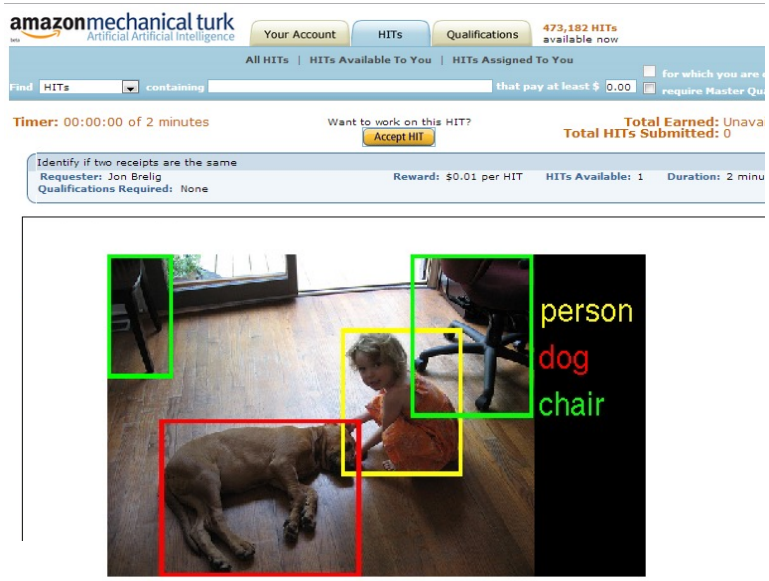
行业调研

Company	Team	Persona
Amazon	Product classification	Largely single-case user
Captricity	Focus of large part of company	Largely single-case user
Dropbox	Single person consulting several teams	Multi-case user / Internal provider
Facebook	Entities team	Multi-case user
Flipora	Startup CTO	Multi-case user
GoDaddy	Small business data extraction	Multi-case user
Groupon	Merchant data team	Multi-case user
Google	Internal crowdsourcing team	Internal provider
Google	Web knowledge discovery team	Multi-case user
LinkedIn	Single person consulting several teams	Multi-case user / Internal provider
Microsoft	Internal crowdsouricng team	Internal provider
Microsoft	Search relevance team	Multi-case user
Youtube	Crowdsourcing team	Largely single-case user

Amazon Mechanical Turk

50万+ 工人*





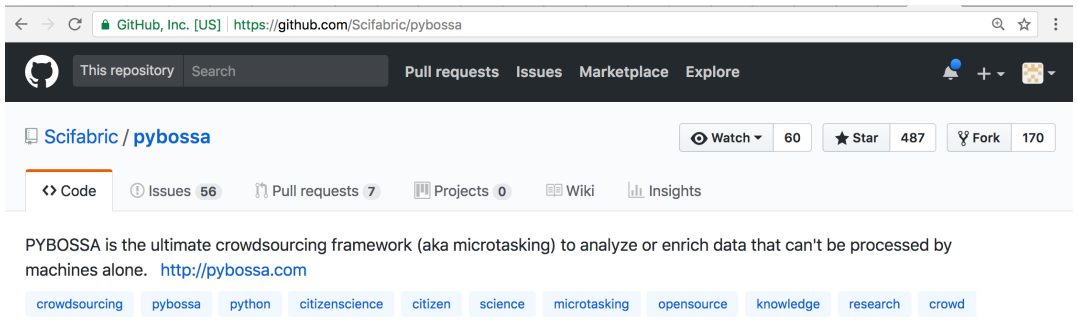
* https://requester.mturk.com/tour

众包可能不适用 ☹

如果你的数据是机密的怎么办?

- 例如：医疗数据、客户数据

内部众包平台



众包可能不适用 ☹

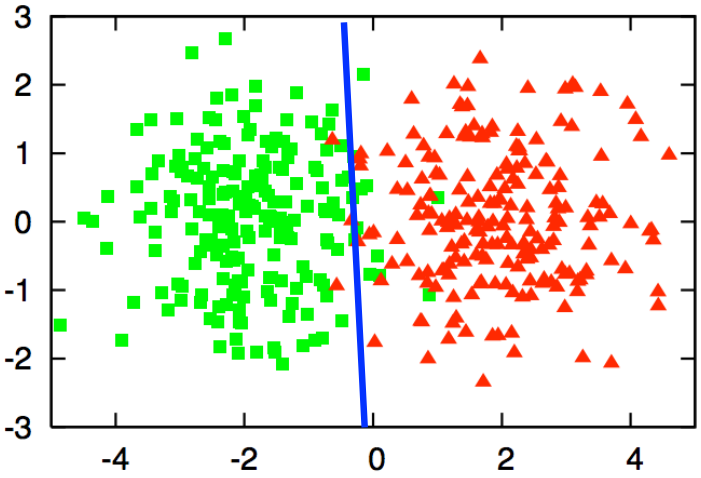
如果你的数据量非常大怎么办？

- 例如：标注 1000 万张图片

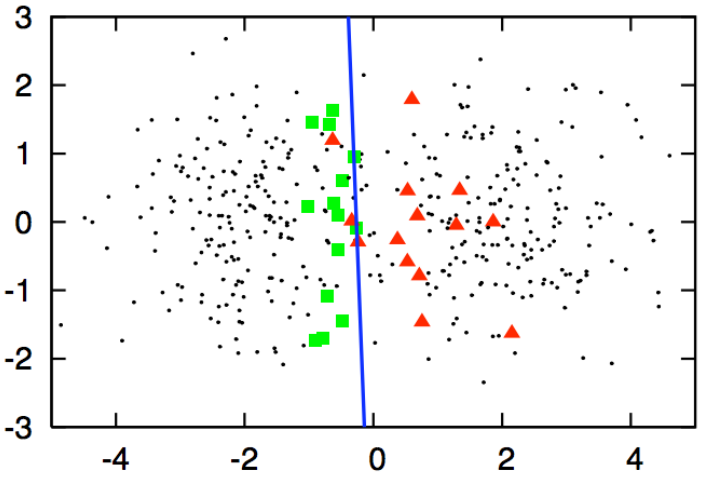
主动学习

主动学习

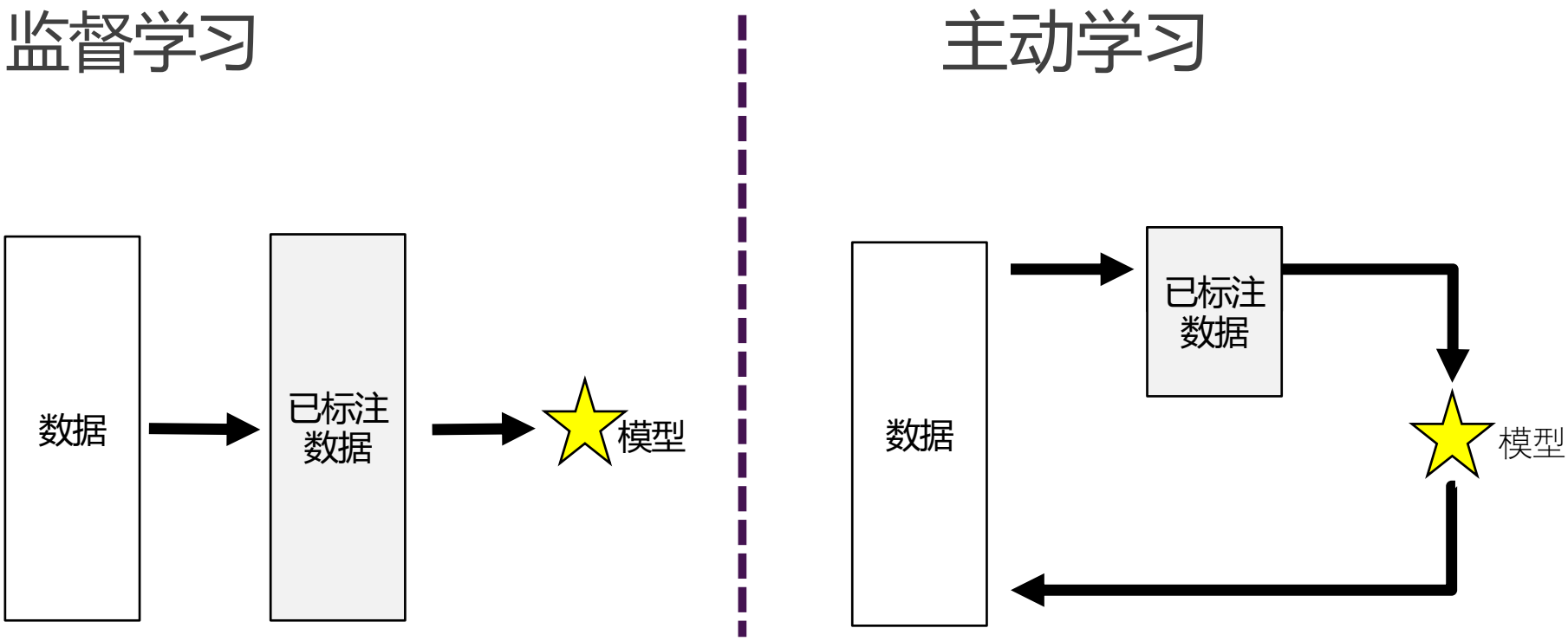
监督学习



主动学习



工作流程



查询策略

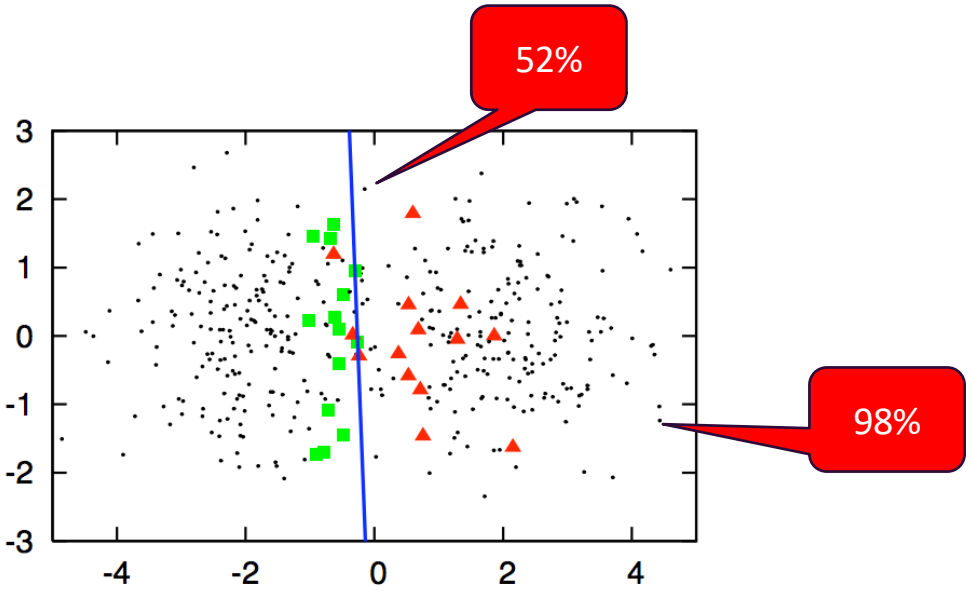
应该标注哪些数据点?

- 不确定性采样
- 委员会查询
- 期望误差减少
- 期望模型变化
- 方差减少
- 密度加权方法

Settles, Burr. "Active learning literature survey." University of Wisconsin, Madison 52.55-66 (2010): 11.

不确定性采样

选取最不确定的数据点进行标注



逻辑回归

- predict_proba(X)

总结

Preppin' Data
每周挑战，帮助你学习数据准备和使用 Tableau Prep
https://preppindata.blogspot.com/

数据集成

- 模式映射、实体解析、数据融合

实体解析

- 基于相似度、众包、主动学习