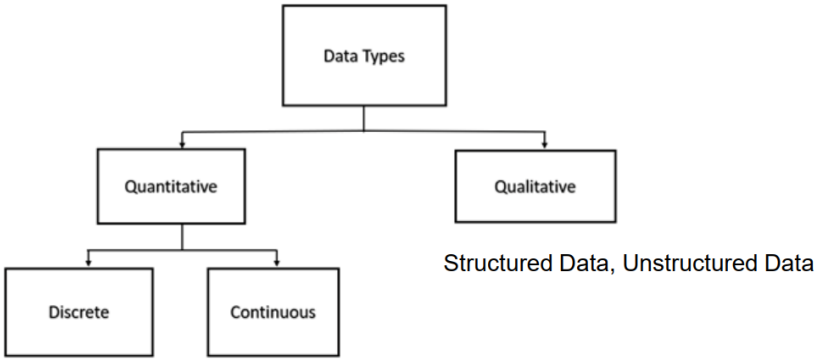
数据科学&工具

数据处理周期：Data Collection → Data Processing → Exploratory Analysis & Data visualization → Analysis Hypothesis Testing & ML → Insight & Policy Decision

首先要对问题有好的定义 (**Well-defined Problem**)

1. **数据收集**：从外部、从公司现有数据集、自我创造的数据收集
    - Primary一手数据收集：通常用于处理当前没有相关数据情况，收集方式包含 Surveys, Interviews, Monitors
    - Secondary二手数据收集：有相关数据，收集方式包含直接拿取数据集，文章，统计等
2. **数据处理**：清洗以保证数据质量
3. **数据探索和可视化**：思考数据中有什么，*'It's important to understand what you CAN DO before you learn to measure how WELL you seem'*，可视化可以帮助我们了解数据之间的规律
4. **假设检验&机器学习模型**
5. **解释**：给出相关政策，使利益相关者（但通常不是技术人员）信服

Data：可以是Observations, facts, numbers, quantities, information, graphs, measurement



Data种类：

- Tabular Data表格：行与列的结构化数据
- Text Data：人类可读的文字
- Graph Data：通过结点和边表达实体之间的联系
- 非结构化数据：视频、图像、音频、生物信息、触觉

Data形式：

- CSV：通过commas分割的数据
- TSV：通过tabular分割的数据
- Image：像素矩阵。可被压缩成
  - Lossy有损压缩：牺牲数据减少大小，e.g. JPEG
  - Lossless无损压缩：保留数据品质同时压缩，e.g. ZIP
- Database：存储数据的结构，高效存储，操作数据，包含DBMS和Storage两个部分
- JSON：文本形式，包含Keys和Values，易读性高，可兼容多种数据类型
- XML\HTML：markup标记语言（使用tag），HTML常用于制作网站，XML更倾向于数据存储。BeautifulSoup在Python中常用于提取HTML和XML数据；Restful API常用于交互

Pandas：

- Seris存储一整列的数据
- DataFrame存储一整个数据集

```
import pandas as pd
```

```
df = pd.read_csv('dataset.csv')                                  # Lo
print(df['column_name'])                                         # Ac
df['new_column'] = df['column1'] + df['column2']                 # Cr
df['column_name'].sum()                                          # Su
df[‘column name’][boolean condition]                             # Fi
df[boolean condition]                                            # Fi

df[df['condition']][column_name].statistics_function()
df.groupby('condition')[column_name].statistics_function()

df.to_csv('filename.csv')                                        # Sa
```

Database&SQL：

- Inner Join：找出两个数据库中所有满足连接微词join-predicate的元素

```
SELECT columns_from_both_tables
FROM table1
INNER JOIN table2
ON table1.column1 = table2.column2
```

Experimental Design：通过计划和实验来验证一个**假设**

- Planning and Conducting 实验来收集有用的数据
- Independent Variable：影响结果的因素、原因

- Dependent Variable：结果
- Dependent & Independent Variable：很多独立变量可能会互相影响，实验需要最小化相关联变量的出现
- 假设Hypothesis：一个解决问题的实验性声明
- 干扰因子Confounder：外来的变量，会影响独立和非独立变量
    - 三种方法：控制、随机化、重复
    - **控制**：建立控制组Control Group（不含干扰因子）和治疗组Treatment Group（含干扰因子），比较干扰因子在两种组别的效果
    - **随机**：将对象随机分配到控制组和治疗组，这将减少系统性干扰，且每个对象都要分配到两个组别
    - **重复**：重复实验，保证实验的一致性和可靠性

收集数据：

- 观察研究Observational Studies：不加干预和操纵
    - 横断面研究cross sectional studies：在同一时间从不同对象上收集数据
    - 回顾性研究Retrospective studies：通过过去的事件研究来验证研究对象关联性
    - 前瞻性研究Prospective studies：研究人员在一段时间，跟踪一组研究对象，收集对象的相关数据
- 问卷调查Surveys：从有结构的问卷或访谈中收集信息
- 实验Experiments
    - 安慰剂效应Placebo Effect：人们相信治疗能够影响他们，尽管没有任何的真实治疗
    - 盲Blinding：控制组和治疗组都不知道自己是否得到了真正的治疗
        - 单盲：实验参与者或研究者不知道
        - 双盲：实验参与者和研究者都不知道
- 模拟仿真Simulations：创建人工环境为真实世界问题建模，通常用于真实情况难以实现的情况

采样方法：抽样方法是从总体中选择单个项目/事件(观察单位)纳入样本的过程，常见方法有

- 随机采样Random Sampling：从整个群体中随机采样

- 分层采样Stratified Sampling：将群体根据特征分成不同的组，再从不同组中随机采样
- 整群抽样Cluster Sampling：将群体根据特征分成不同的组，挑选组并将其所以个体作为样本
- 系统抽样Systematic Sampling：在整个群体中，每第k个人挑选作为样本
- 便利抽样Convenience Sampling：挑选对研究者最便利的群体作为样本