MM5425 商业分析

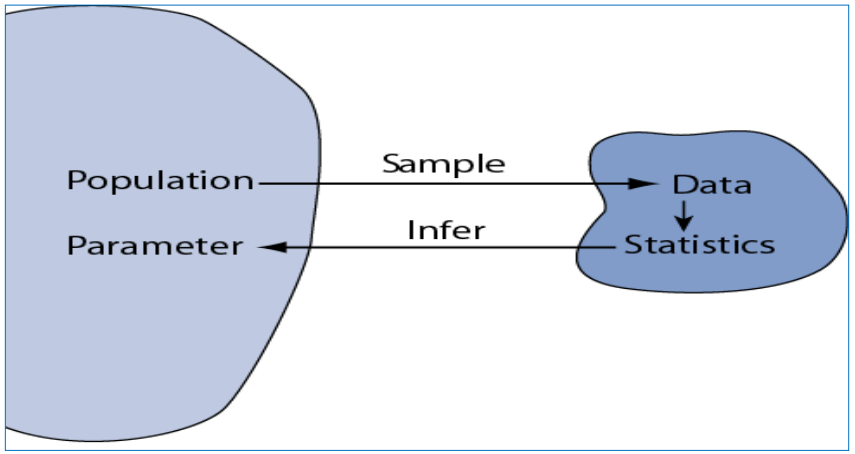
WEEK 3 LECTURE – HYPOTHESIS TEST, DATA MINING PROCESS

第三课 内容

- 假设检验 Hypothesis Testing
  - 原假设和备择假设 Null and alternative hypothesis
  - 统计检验Test statistic
  - P值 P-value
  - 显著性水平 Significance level

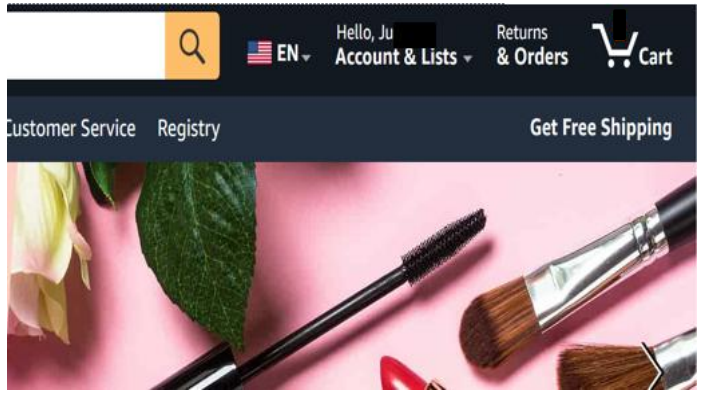
- **跨行业数据挖掘流程** (CRISP-DM):
  I. 业务理解 Business Understanding
  II. 数据理解 Data Understanding
  III. 数据准备 Data Preparation
  IV. 建模 Modelling
  V. 评估 Evaluation
  VI. Deployment

假设检验
Hypothesis Testing

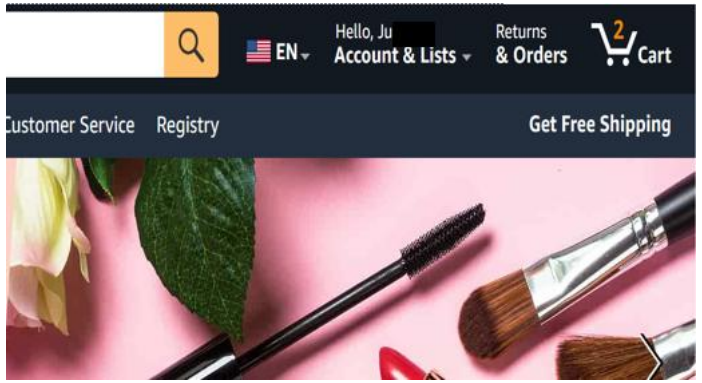


INFERENCE FROM SAMPLE

例子: 网站优化



A

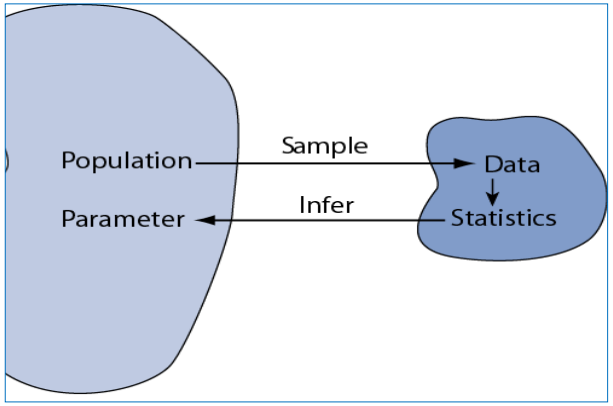


B

网页设计A和网页实际B，哪一个更优?

假设检验简介

- 假设检验是一种统计方法来证明某事是否正确.
- 例子:
  - 新网页比旧网页好吗?
  - 新的营销活动比其他营销活动更好吗?
- 当从总体中抽取随机样本时，所获得的信息可用于对总体特征进行推断



假设检验步骤Hypothesis Testing Steps

1. 制定原假设和备择假设。
2. 选择合适的检验统计量，并确定拒绝或保留原假设的标准。
3. 计算检验统计量的 p 值，即在原假设成立时观察到该统计量值的条件概率。
4. 决策：如果统计量的值落在临界区，则拒绝原假设，否则接受原假设。

1: 原假设和备择假设

原假设 (H0) 是当前持有的信念

备择假设 (H1) 认为 “H0 为假”

在检验假设时，我们会确定它是单尾检验还是双尾检验.

双尾Two-tailed: 参数发生变化

$$H_0 : \mu = 6.7$$
$$H_1 : \mu \neq 6.7$$

单尾One-tailed: 参数变大或者变小

$$H_0 : \mu = 6.7 \quad H_1 : \mu > 6.7$$

or

$$H_0 : \mu = 6.7 \quad H_1 : \mu < 6.7$$

例子 1: 单尾检验 One-sided Testing

过去五年，该店平均顾客满意度为6.7。店铺改造后，经理有理由相信该店顾客满意度会更高.

原假设 Null hypothesis $H_{0:}\ \mu = 6.7$ (“没有分别”)

备择假设The **alternative hypothesis** $H_{1:}\ \mu > 6.7$ (“客户更加满意”)

例子 2: 双尾检验 two-sided Testing

陈述——男生和女生毕业时的平均薪资是不同的

$$H_0: \mu_m = \mu_f$$
$$H_A: \mu_m \neq \mu_f$$

其中，μ_m 和 μ_x 分别表示男生和女生在毕业时的平均薪资。

2: 检验统计量 Test Statistic

Z检验：如果总体方差已知。通常用于样本量大于30的情况:

$$\text{z-statistic} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

T检验：如果总体方差未知。通常用于样本量小于30的情况

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

3: 计算P值 *p*-value

p 值回答的问题是：在原假设（**H_0**）**为真**的情况下，观察到当前检验统计量或更极端结果的概率是多少？

因此，p 值是支持原假设的证据。p 值越小，反对原假设的证据就越强。

所以，p 值很小 ⇒ 有强有力的证据**拒绝**原假设（H_0）。

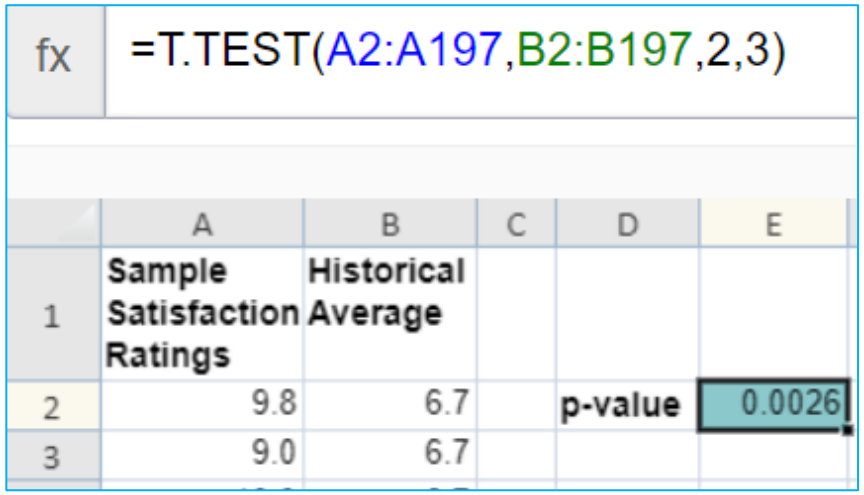
3: 计算p值 *p*-value

单样本检验示例

$H_{0:}$ μ = 6.7 (“没有不同”)

$H_{1:}$ μ ≠ 6.7 (“客户满意度有不同”)

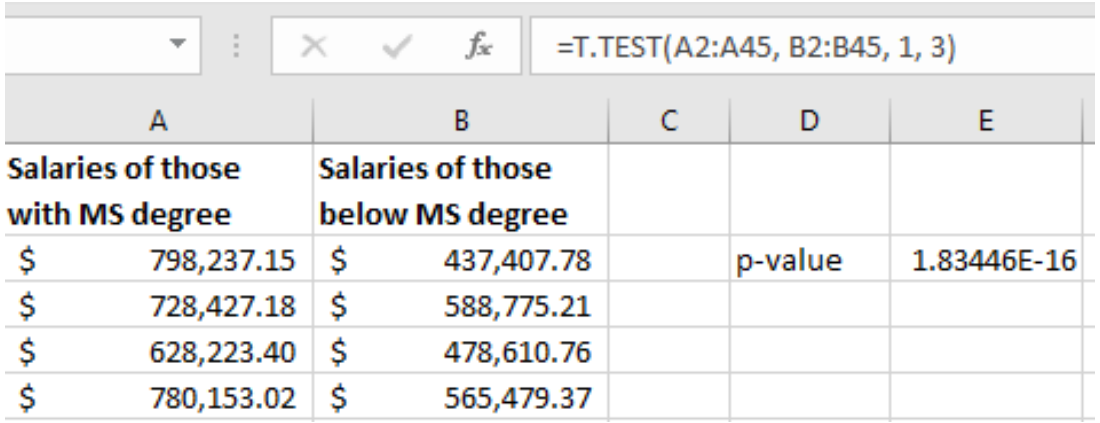
假设历史平均满意度评级为 6.7，而 196 名新随机样本的平均满意度评级为 7.3，标准差为 2.8.



3: 计算p值 p-value

双样本检验示例
硕士毕业生的平均年薪高于非硕士毕业生.

$H_0 : \mu_{MS} = \mu_{not\ MS}$
$H_1 : \mu_{MS} > \mu_{not\ MS}$



A	B	C	D	E
Salaries of those with MS degree	Salaries of those below MS degree			
$ 798,237.15	$ 437,407.78		p-value	1.83446E-16
$ 728,427.18	$ 588,775.21			
$ 628,223.40	$ 478,610.76			
$ 780,153.02	$ 565,479.37			

4: 决策标准——显著性水平

- 显著性水平，通常用 α 表示，是根据计算得到的 p 值来决定是否拒绝原假设的判断标准.
- 显著性水平也定义了置信水平。

  补充说明：

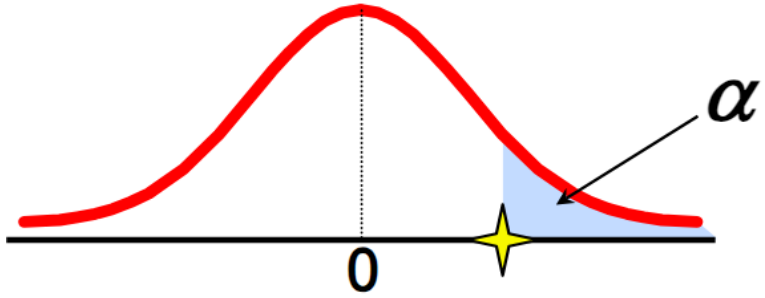
  置信水平 = 1 – 显著性水平（α）。

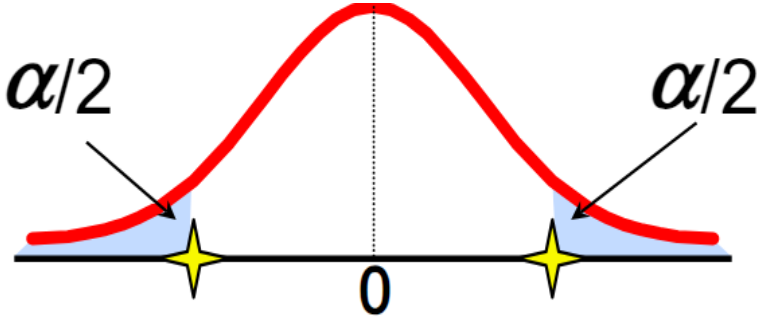
  例如，显著性水平为 0.05 时，置信水平为 95%。：

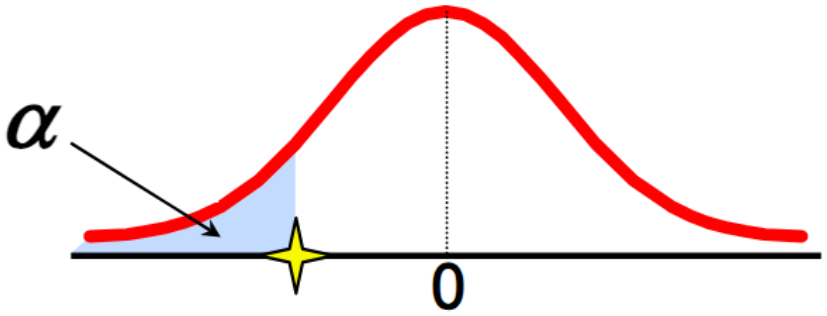
4: 决策标准

- 显著性值 α 是 p 值的最大阈值

  - Reject H_0 when $p \leq \alpha$
  - Retain H_0 when $p > \alpha$







网站 A/B 测试示例

下图显示了网站A和B的销售额.

取 $\alpha = .05$, $p = 0.0339$, $p < \alpha \Rightarrow$ 拒绝 reject H_0

结果显示网站 B 的销量比网站 A 有显著提高。

Mean	127.7358	128.1242
Variance	12,533.4069	12,620.1956
Observations	750,706	749,294
Hypothesized Mean Difference	0.0000	
Mean Difference	0.3884	
% Mean Difference	0.0030	
P-value	0.0339	

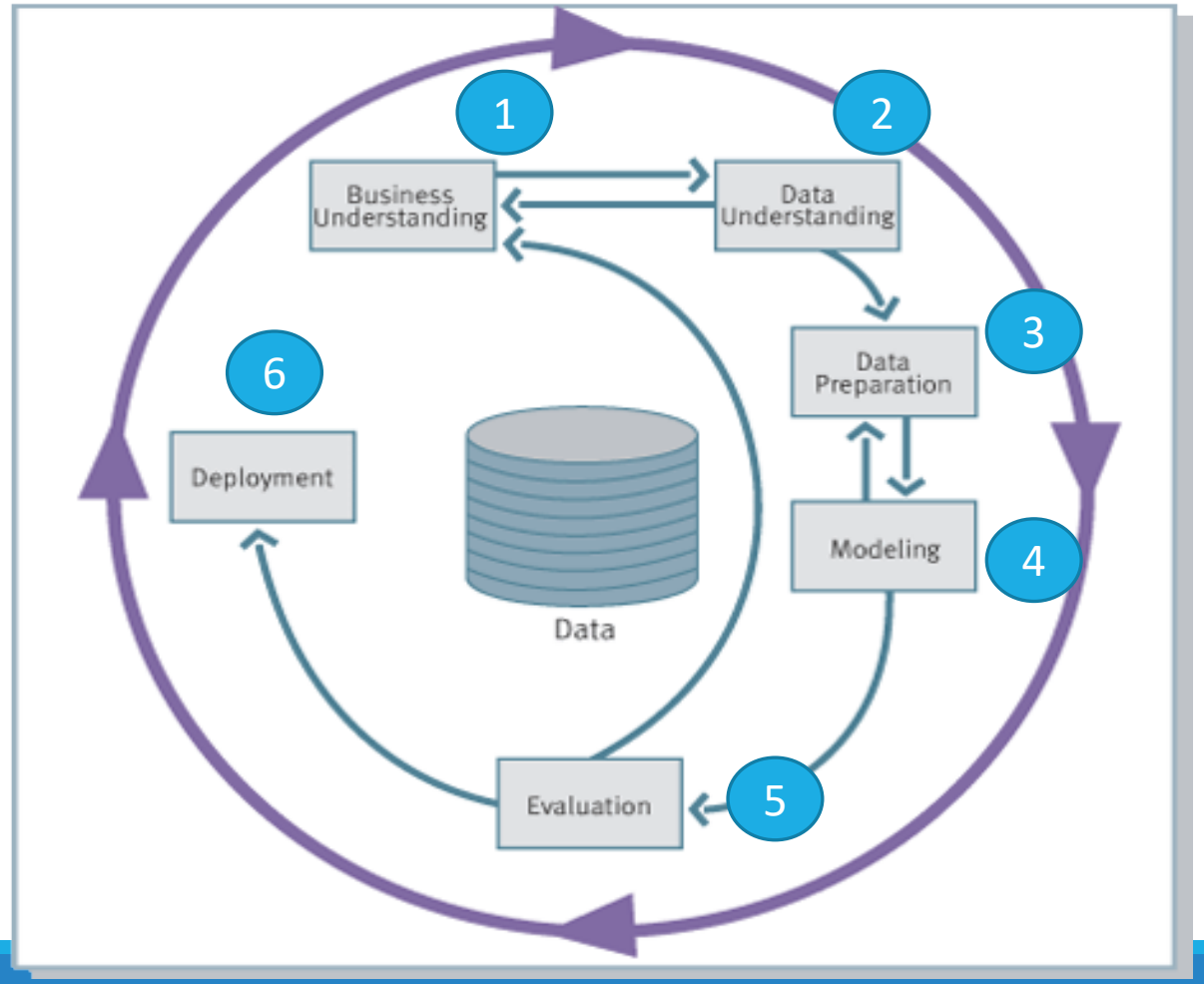
跨行业数据挖掘标准流程(CRISP-DM）

DATA MINING PROCESS

商业分析的基本原则

- 人类的决策和行为**并非随机**。
- 基于过往行为（历史数据）分析模式和关系。
- 依靠这些模式和关系来改进决策过程和业务成果

跨行业数据挖掘标准流程(CRISP-DM)



Step 1: 商业/业务理解

商业/业务理解包括:

- 明确业务目标 Determine business objectives
- 状态评估 Assess situation
- 决定数据挖掘目标 Determine data mining goals
- 制作项目计划 Produce project plan

Determine business objectives

?

商业/业务理解

明确业务目标

场景评估

硬件/
软件

大规模数据

人员

大规模数据

数据

客户浏览/
购买数据

商业/业务理解

明确业务目标

场景评估

数据挖掘目标

?

制作项目计划

项目目标示例

Phase	Time	Resources	Risks
Business understanding	1 week	All analysts	Economic change
Data understanding	3 weeks	All analysts	Data problems, technology problems
Data preparation	5 weeks	Data mining consultant, some database analyst time	Data problems, technology problems
Modeling	2 weeks	Data mining consultant, some database analyst time	Technology problems, inability to find adequate model
Evaluation	1 week	All analysts	Economic change, inability to implement results
Deployment	1 week	Data mining consultant, some database analyst time	Economic change, inability to implement results

Step 2: 数据理解

对象及其属性的集合

属性是对象的属性或特性

- 例如：人的眼睛颜色、温度等。
- 属性也称为变量、字段、特性、维度或特征

属性集合是用来描述一个对象

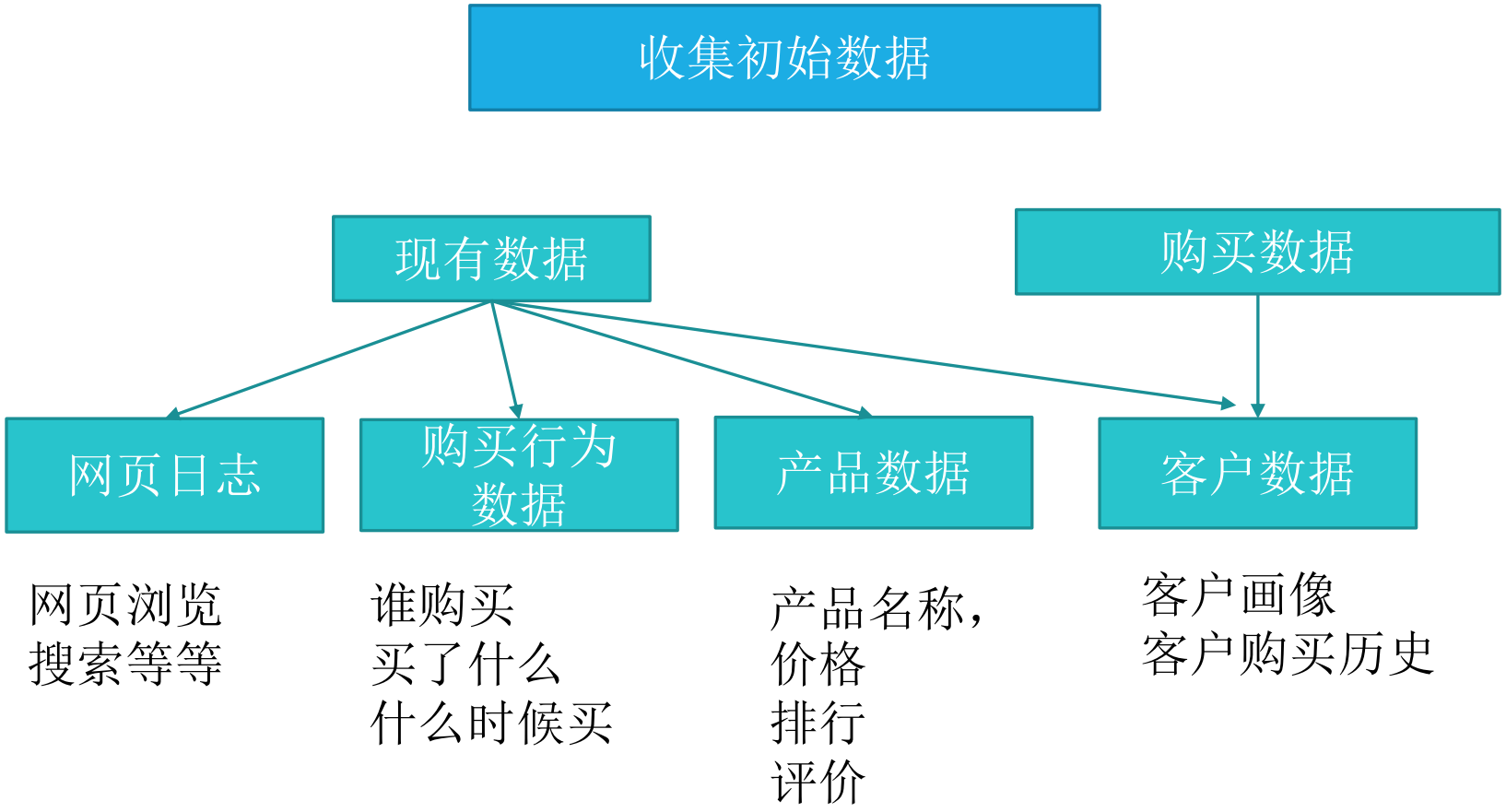
- 对象也称为记录、点、案例、样本、实体或实例

Attributes

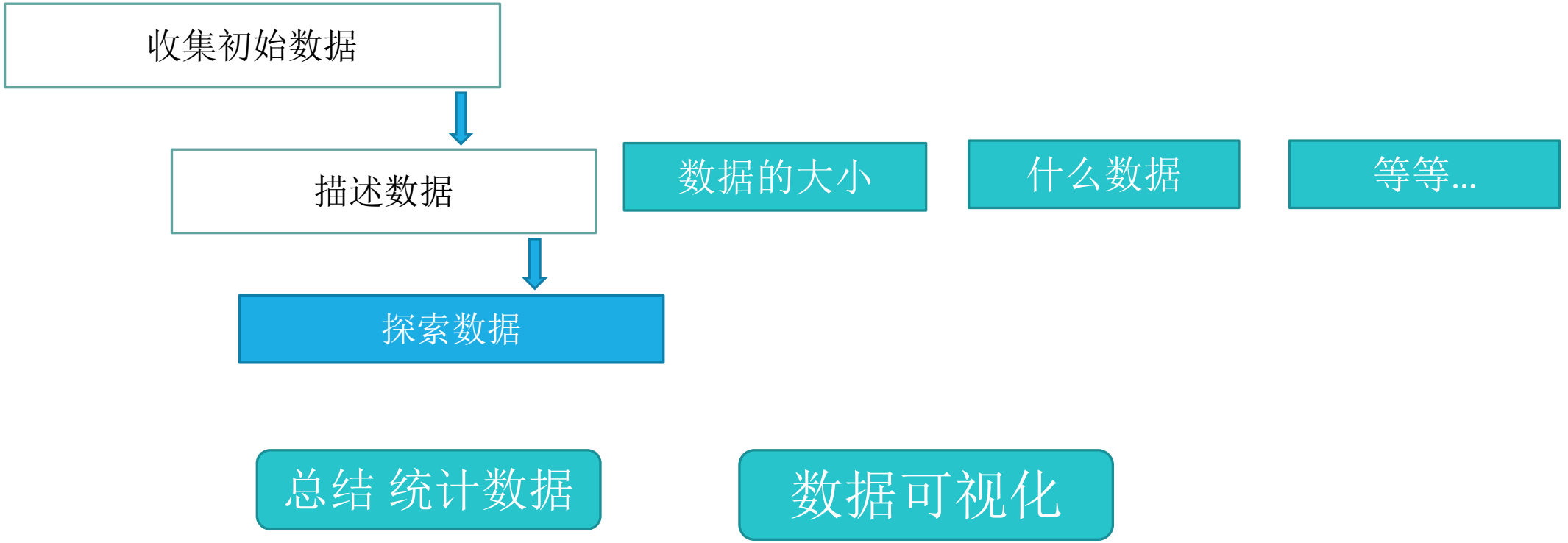
Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

数据理解: 收集数据 Collect Data



数据理解: 探索数据



Step 3: 数据准备 – 选择数据

选择观测值

- 抽样 Sampling

选择感兴趣的属性

- 选择正确的数据列：特征工程
- 要包含哪些产品/客户属性

数据准备 – 数据清理

- 数据质量差会对许多数据处理工作产生负面影响
- 数据质量问题实际例子:
  - 噪音和异常值 Noise and outliers
  - 错误数据 Wrong data
  - 假数据 Fake data
  - 缺失数据 Missing values
  - 重复数据 Duplicate data

数据准备 – 整合数据

网页日志	购买行为数据	产品数据	客户数据
网页浏览 网页搜索	收购买 买了什么 什么时候购买e	产品名称 产品价格 产品排名 产品评价	客户画像 客户购买记录

Go down deep enough into anything and you will find Mathematics.

— *Dean Schlicter*

Step 4: 建模 Model Building

- 模型Model

$$Y = f(X)$$

数据挖掘目标 Data mining goal: 了解 f 功能
函数是什么 learn what *f* function is

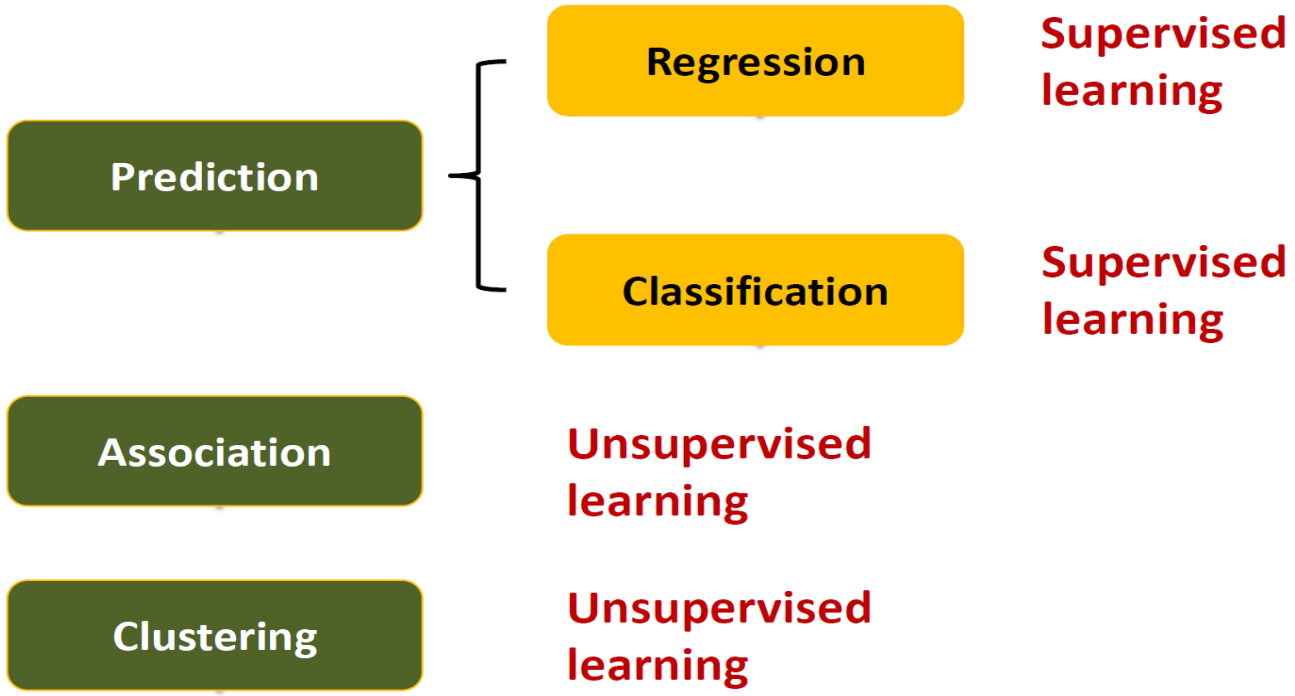
- 例子:

$$需求demand = f(价格price)$$

$$成绩grades = f(学习小时hours\ studied)$$

数据挖掘模型

模型任务:


Prediction
Regression
Classification
Association
Clustering
Supervised learning
Supervised learning
Unsupervised learning
Unsupervised learning

Step 5: 模型评估

预测 **Prediction (监督学习 supervised learning)**

回归分析 Regression

$$Y \; vs \; \hat{Y} \; = \; f(X)$$

Metrics	Formula
Mean absolute error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Mean squared error (MSE)	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Root mean squared error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

哪一个结果最佳?

Regression task	**MSE**
Model 1	0.04
Model 2	0.27
Model 3	0.13
Model 4	0.08

模型评估

二元分类的混淆矩阵 **Confusion matrix for binary classification**

	Actual 0	Actual 1
Predicted 0	True Negatives (TN)	False Negatives (FN)
Predicted 1	False Positives (FP)	True Positives (TP)

准确度 Accuracy = (TN + TP) / (TN + FN + FP + TP)

Classification task	Accuracy
Model 1	0.84
Model 2	0.92
Model 3	**0.94**
Model 4	0.89

Step 6: 部署 Deployment

- 在部署阶段，数据挖掘的结果和数据挖掘技术被实际应用起来。

  例如，将预测模型应用于某些信息系统或业务流程：

- 向被预测为特别有风险的客户发送特别优惠；

  在管理信息系统中使用欺诈检测模型，监控账户并创建“案例”进行审查