

2022-2023秋季课程:数据科学与大数据导论

Introduction to Data Science and Big data


# Chapter 2: Data Science Fundamentals

曹劲舟 助理教授

深圳技术大学 大数据与互联网学院

caojinzhou@sztu.edu.cn

2022年9月



# Outline

---

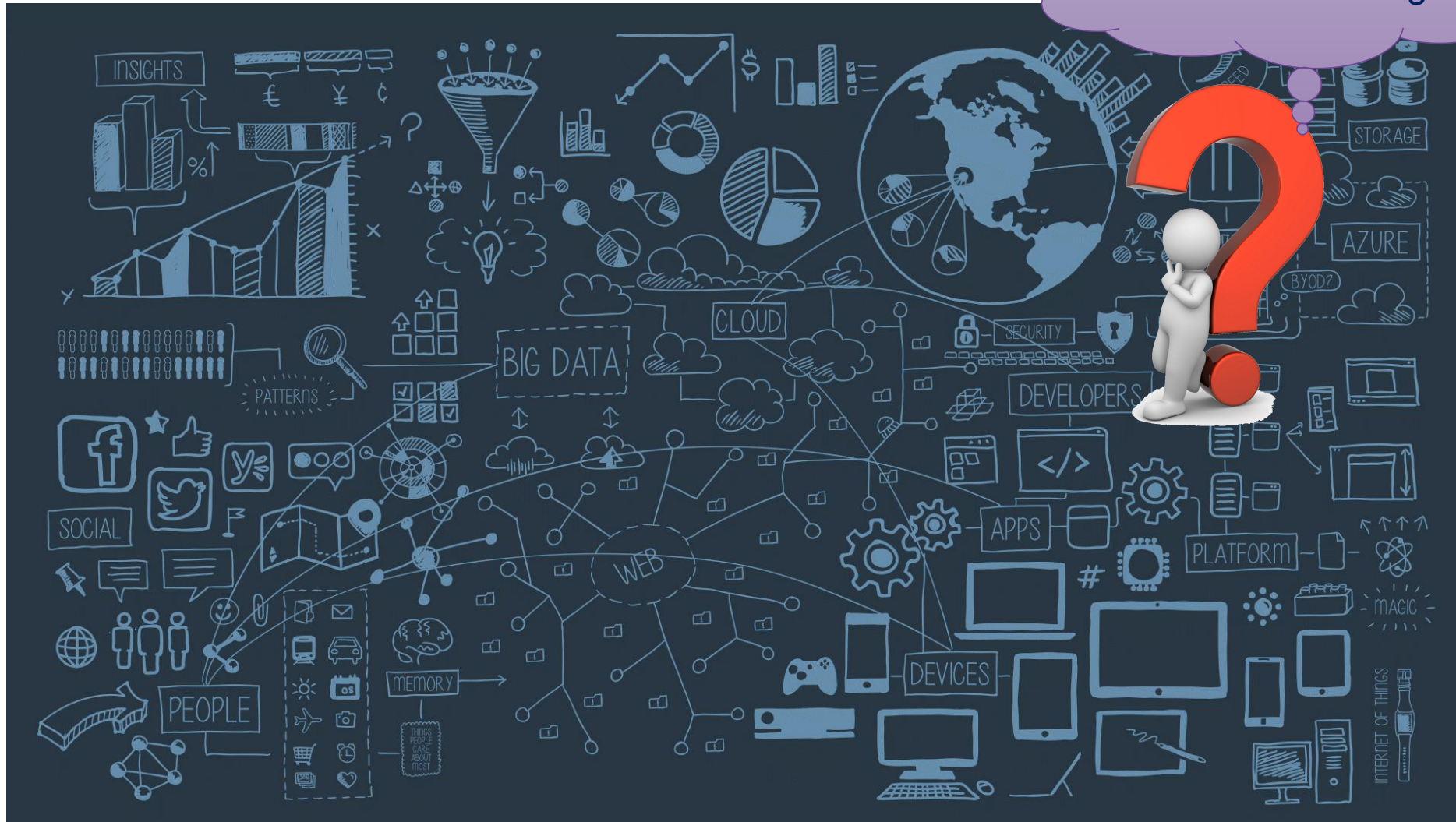
- ❑ What is Data Science?
- ❑ The history of Data Science
- ❑ Doing Data Science
- ❑ The Skillset of Data Scientists
- ❑ Summary

# **01 Data Science – What IS IT? 什么是数据科学？**



# What is “Data Science” ?!?

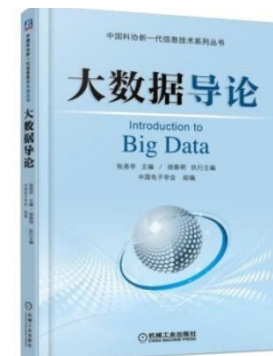
Tools? Big Data?  
Machine Learning?





# Defining Data Science

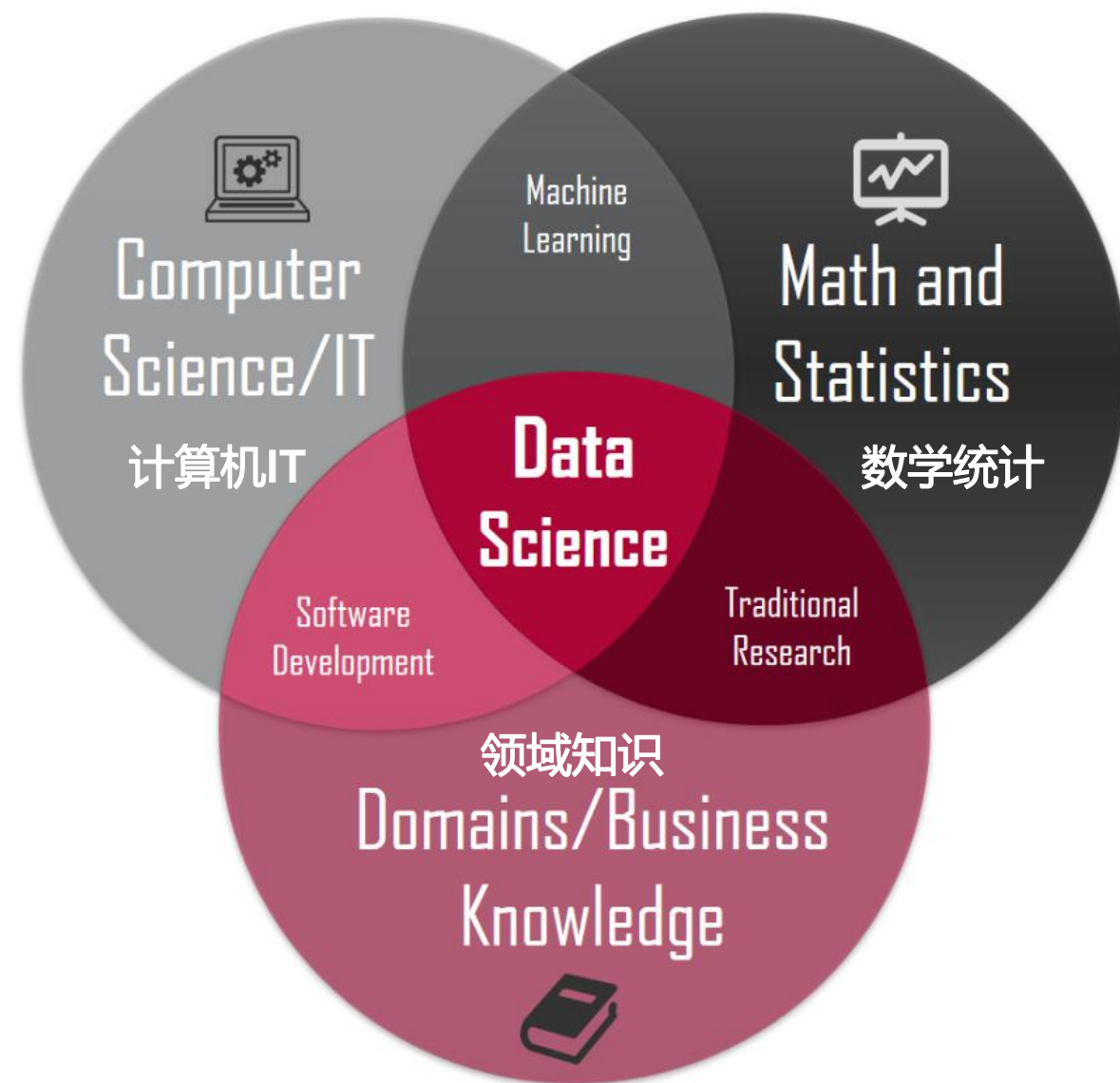
- 什么是数据科学？
  - 基于传统的数学、统计学的理论和方法，运用计算机技术进行大规模数据计算、分析和应用的一门学科（摘录自《大数据导论》）
  - 用数据的方法研究科学，用科学的方法研究数据（摘录自鄂维南院士“数据科学的基本内容”）



# Defining Data Science

□ 一个公认却很宽泛的定义

Data science is an **inter-disciplinary** field that uses scientific methods, processes, algorithms and systems to extract **knowledge** and **insights** from many structural and unstructured data.



# Contrast: Data Mining 数据挖掘

---

- 数据科学包括数据分析（统计学和机器学习）、计算机科学以及领域知识，以从数据中提取价值为目的。可以看作是对数据的商业加工，其不仅仅可以将数据转化为信息，还可以转化为产品（个性化推荐、实时竞价、精准营销）。
- 数据挖掘(Data Mining)旨在从大量的、不完全的、有噪声的、模糊的、随机的数据中,提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识。
- 数据挖掘是数据科学的组成部分，用来挖掘潜在的信息；
- 数据科学得出的结论是人的智力活动结果，而数据挖掘得出的结论是机器从学习集（或训练集、样本集）发现的知识规则。

# Contrast: Big Data 大数

The world is generating data at a higher rate, and so the need of "Data Science" & "Data Analytics" tools increases to analyze and manage this "Big Data".



WHAT IS DATA SCIENCE?	WHAT IS DATA ANALYTICS?	WHAT IS BIG DATA?
<b>Data Science</b> is a field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data.	<b>Data Analytics (DA)</b> is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems & software.	<b>Big Data</b> refers to voluminous amounts of structured or unstructured data that organizations can potentially mine & analyze for business gains.
APPLICATION AREAS		
1. Digital advertisements 2. Internet Research 3. Recommender System 4. Image/Speech Recognition	1. Gaming 2. Travel 3. Energy Management 4. Healthcare	1. Communication 2. Retail 3. Financial services 4. Education
TOOLS & LANGUAGES		
1. Python 2. SAS 3. SQL	1. R 2. Tableau Public 3. Apache Spark	1. Hadoop 2. NoSQL 3. Hive
ANNUAL SALARY		
Data Scientist \$130,323	Big Data Specialist \$69,845	Data Analyst \$62,066



# 03 Doing Data SciENCE

## 数据科学处理流程



# The process of data science

## □数据科学的工作流程



循环迭代式的工作流程

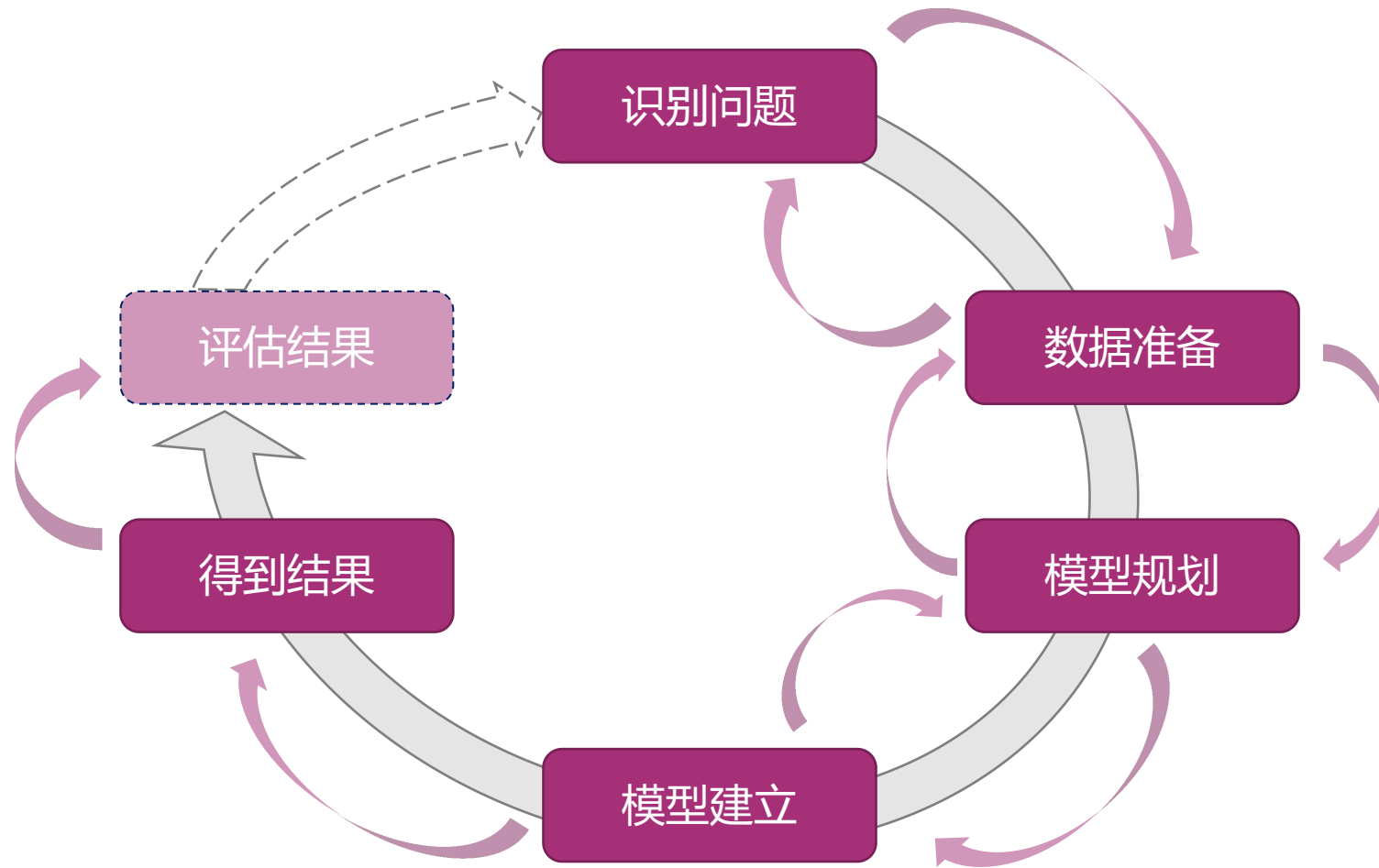
- 先提出问题，再收集与分析相关的数据
- 先收集数据，再分析可以回答哪些问题

任务1：从数据中洞见真知

任务2：数据驱动决策支持

# Process of Data Science Projects

---



# 识别问题 Identify problems

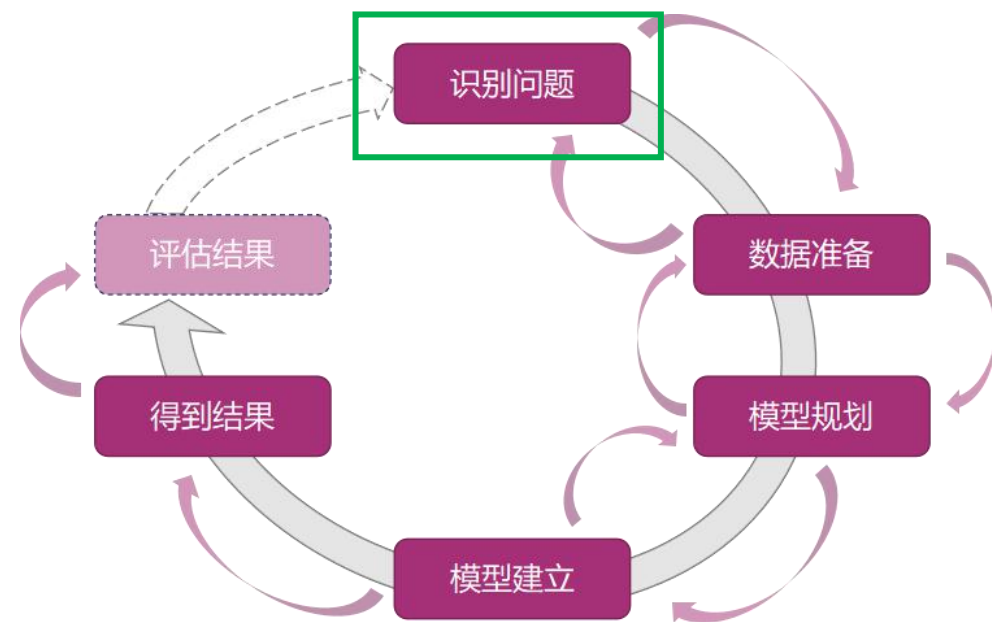
## □步骤第一步

## □学习领域知识

- 了解项目数据和用例的知识
- 用于解释结果的知识

## □学习以往经验

- 参考过去在类似问题上的项目
  - 差异，失败原因，项目的不足之处

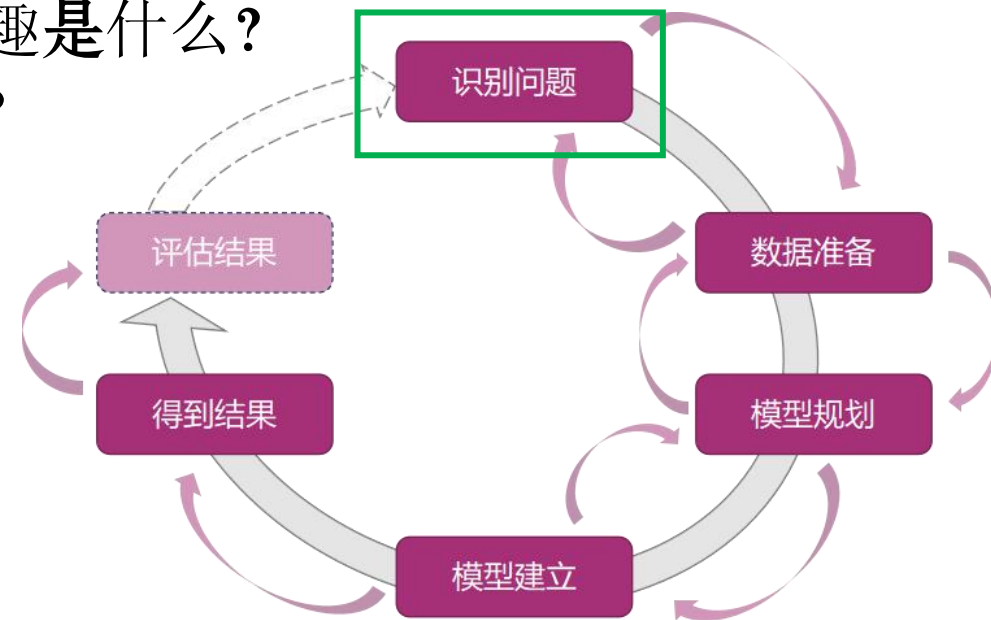




# 识别问题Identify problems

## □Frame the problem框定问题

- 框架是陈述要解决的数据分析问题的过程
- 为什么这个问题很重要?
- 谁是主要的利益相关者，他们在项目中的兴趣是什么?
- 目前的情况是什么，激发项目的痛点是什么?
- 项目的目标是什么?
  - 业务需求
  - 研究目标
- 为实现目标需要做什么?
- 项目的成功标准是什么?
- 项目有哪些风险?



# 识别问题 Identify problems

## □分析可用资源

### ■ Technologies 技术资源

- 计算和存储资源
- 分析框架的许可证

### ■ Data 数据资源

- 可用数据是否足以满足使用的需要?
- 是否需要其他数据? 是否可以在项目范围内收集其他数据?

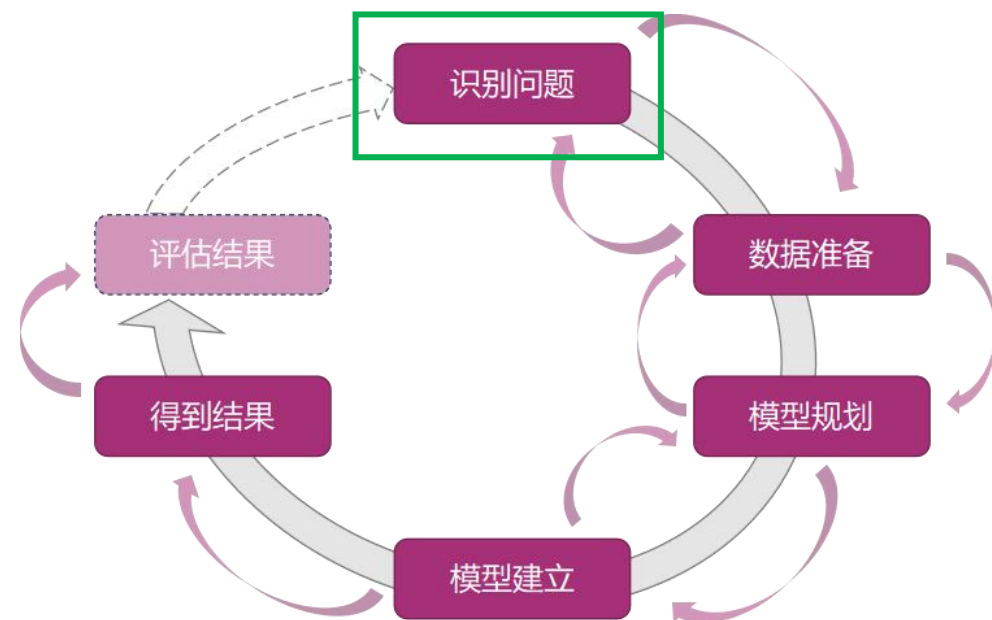
### ■ Timeframe 大体时间

- 项目时间和人力工作时间

### ■ Human resources 人力资源

- 谁可以参与该项目?
- 技能组合是否适合项目的任务?

→ 资源充足才开始项目!



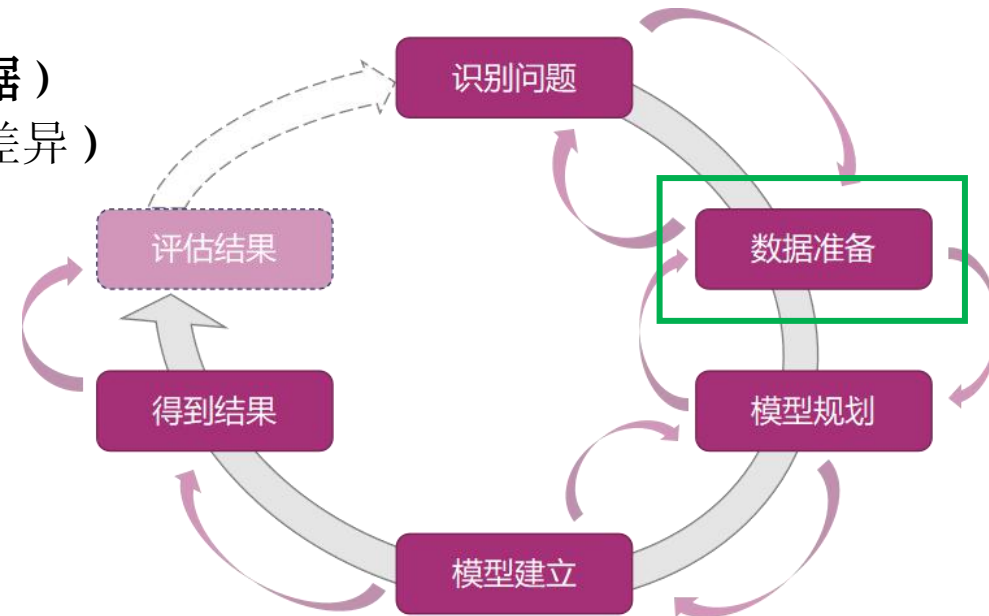
# 数据准备 Data Preparation

## □ 为项目创建数据存储设施

- 数据仓库/csv文件/...  $\leftrightarrow$  分布式存储
- 依数据量大小而定

## □ 提取Extract - 转换Transform - 加载Load (ETL) the data

- 定义如何查询现有数据库以提取所需数据
- 确定原始数据所需的转换
  - 数据质量检查（例如，过滤缺失的数据、不可信的数据）
  - 数据结构化（例如，对于非结构化数据，数据结构的差异）
  - 数据转换（例如时间戳、字符编码）
- 将数据加载到分析环境中



# 数据准备 Data Preparation

## □深入了解数据

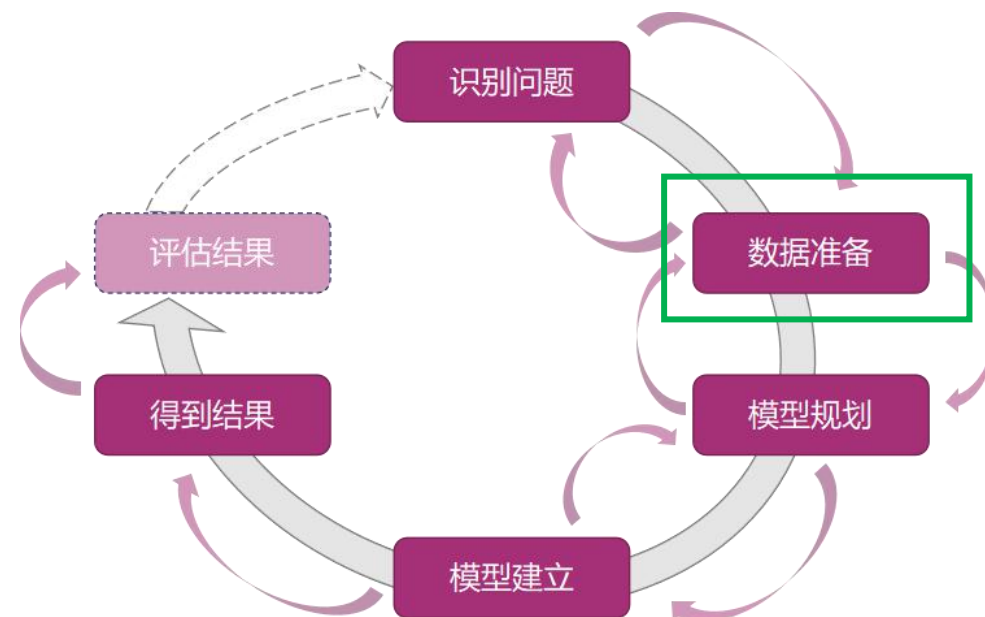
- 了解所有可用的数据源
- 例如，关系数据库中的每一列字段包含什么？
- 如何将结构强加于半/准/非结构化数据？

## □初步分析和可视化数据

- 描述性统计
- 相关性分析
- 直方图、密度图、成对图等可视化。

## □清理和规范化（归一化）数据

- 丢弃不需要的数据
- 归一化以消除比例效应





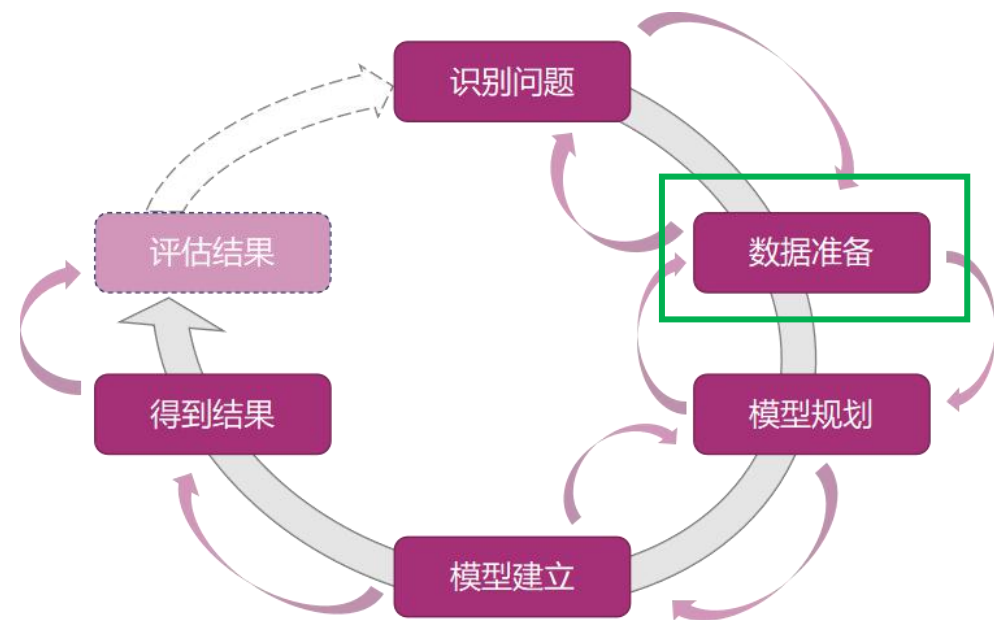
# 数据准备 Data Preparation

## □ 清理数据

- 丢弃不需要的数据
- 区分复杂的基础架构和单台机器进行分析

### ■ 举例:

- 1 亿次测量
  - 每次测量 10 个浮点特征 → 每次测量 80 字节
  - 3 个有用的特征 ≈ 每次测量 24 字节
  - 7.45 GB 包含所有特征, 2.23 GB 仅包含有用特征
- 可以毫无问题地使用笔记本电脑来清理数据



# 模型规划 Model Planning

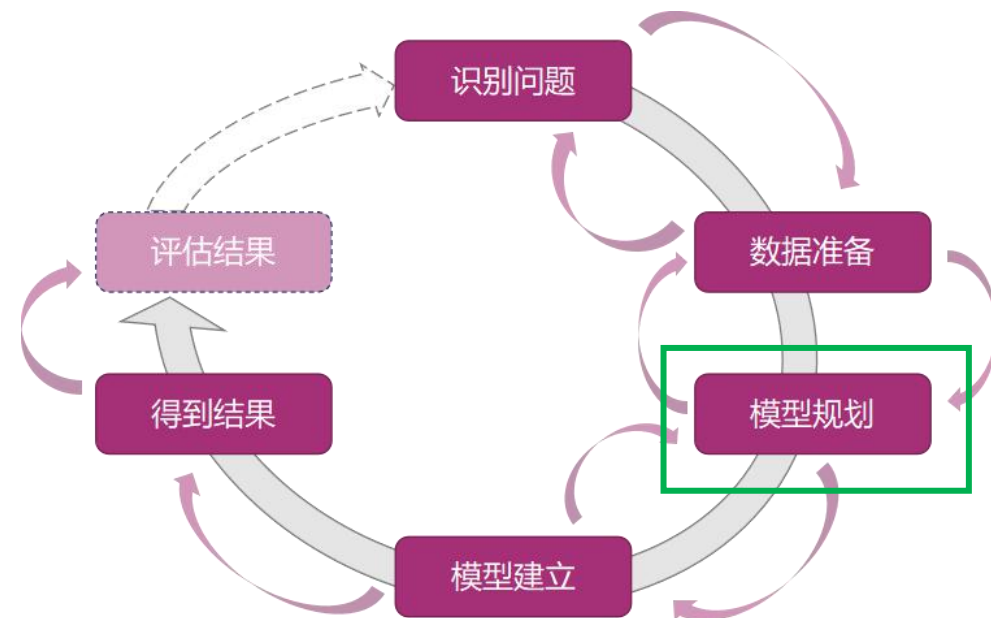
## □确定数据分析方法

## □要适合实现目标

- 决定方法的类型
  - 分类、回归、聚类、关联挖掘……
- 其他因素也会限制可用的方法
  - 例如，如果洞察力很重要，则不能使用“黑盒”方法

## □要适合可用的数据

- 数据大小、结构、……



黑盒方法是一种仅获取结果但并不真正理解为什么以这种方式计算输出的方法。  
白盒方法会解释为什么输出是这样的。

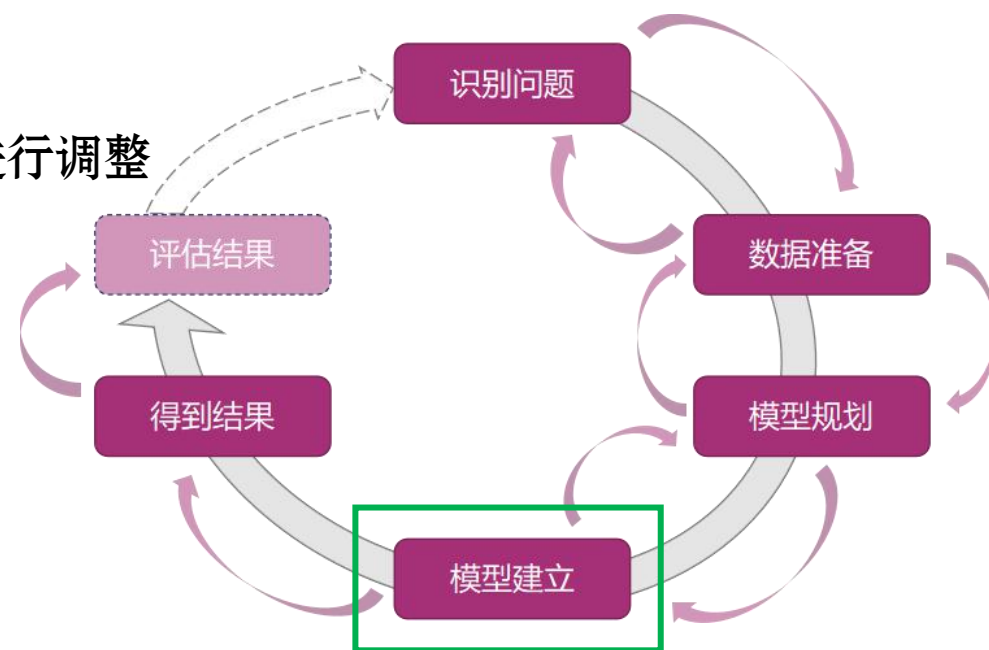
# 模型建立 Model Building

## □使用确定的方法执行分析

- 经常迭代的过程!

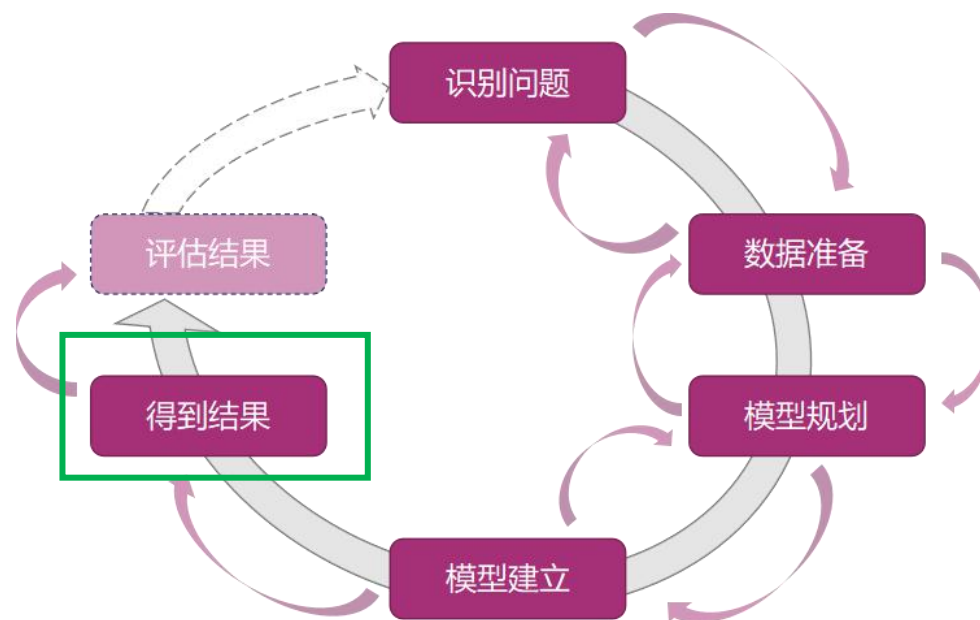
## □每个阶段分别执行, 因为这可能非常耗时

- 使用小规模数据集示例进行模型规划
- 在模型构建期间使用具有潜在大量超参数的真实大数据集进行调整



# 获取结果 Obtain Results

- 主要问题：达到要求了吗？
- 将结果与提出问题阶段的假设进行比较
- 确定主要发现
- 尝试量化结果的价值
  - 商业价值，例如预期投资回报率 (ROI)
  - 先进技术的进步
- 总结调查结果

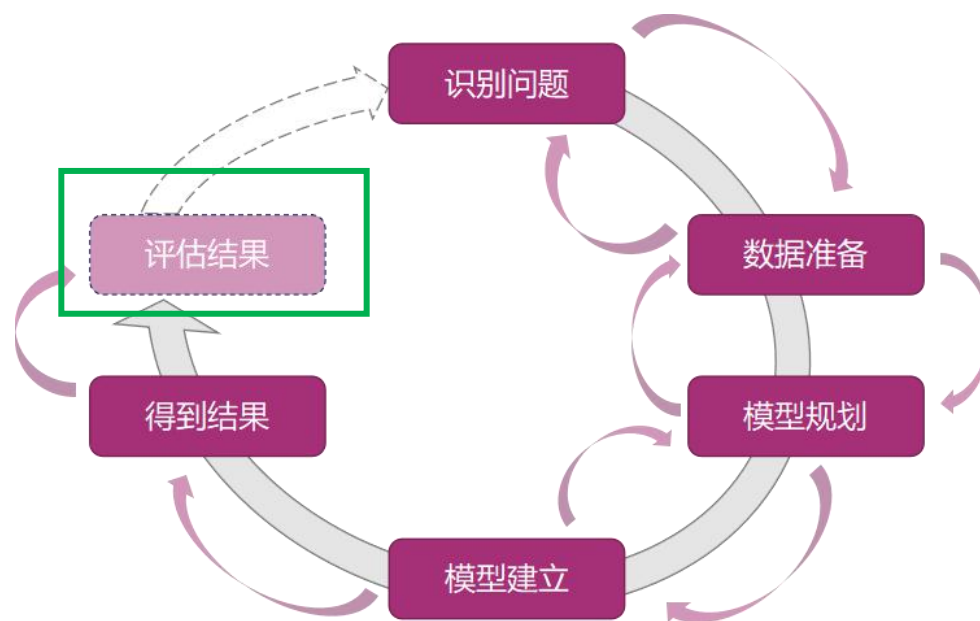




# 评估结果

## □ 定义更新和重新训练模型的过程

- 数据变旧，模型过时
- 数据驱动模型应定期更新
- 流程是必需的



# **04 The Skillset of Data Scientists**

## **数据科学家应该具备什么能力**

