数据科学与计算机学院
School of Data and Computer Science

自然语言处理

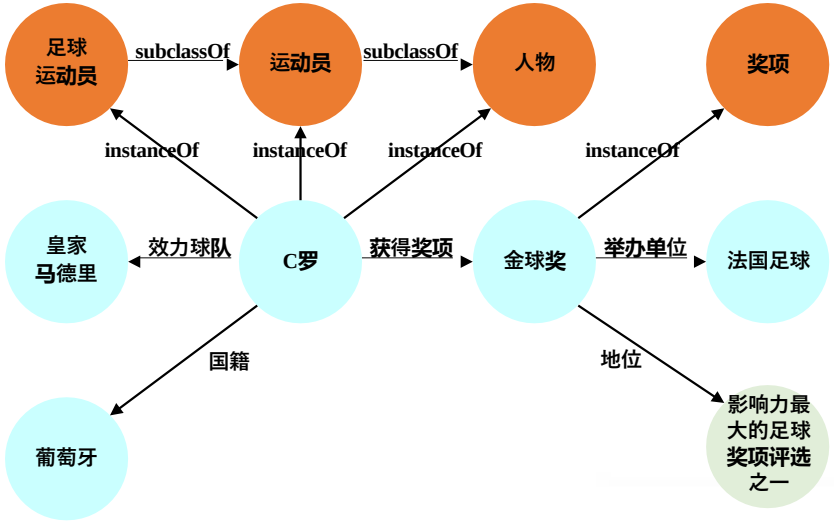
Natural Language Processing

权小军 教授
中山大学数据科学与计算机学院
quanxj3@mail.sysu.edu.cn

课程回顾

知识图谱

- 知识图谱(Knowledge Graph)以结构化的方式描述客观世界中概念、实体及其之间的关系；
- 本质上是一种**大规模语义网络**(semantic network)



知识图谱

- 知识图谱通过对错综复杂的文档的数据进行有效的加工、处理、整合，转化为简单、清晰的“实体-关系-实体”的三元组，最后聚合大量知识;

实体	关系	实体

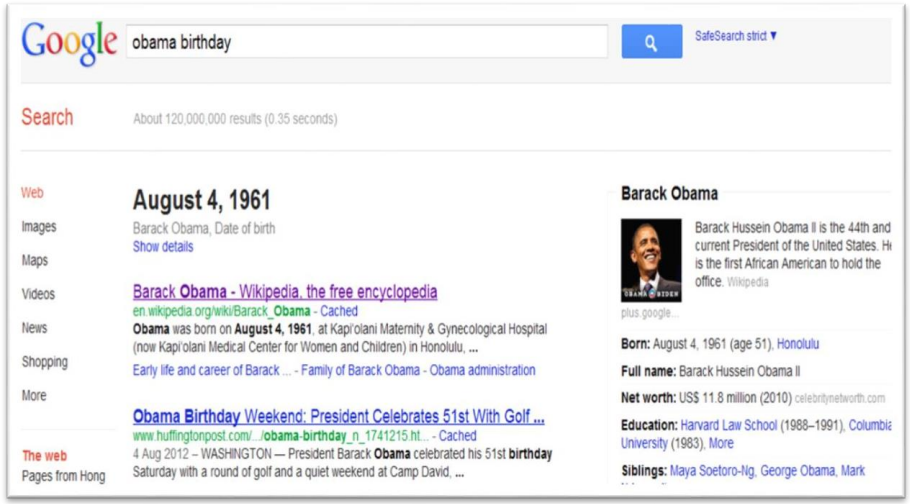
实体	属性	值

诞生标志

- 2012年5月，Google收购Metaweb公司，并正式发布知识图谱

- 搜索核心需求：让搜索通往答案
  - 无法理解搜索关键词
  - 无法精准回答

- 根本问题
  - 缺乏大规模背景知识
  - 传统知识表示难以满足需求



KG优势1： large scale

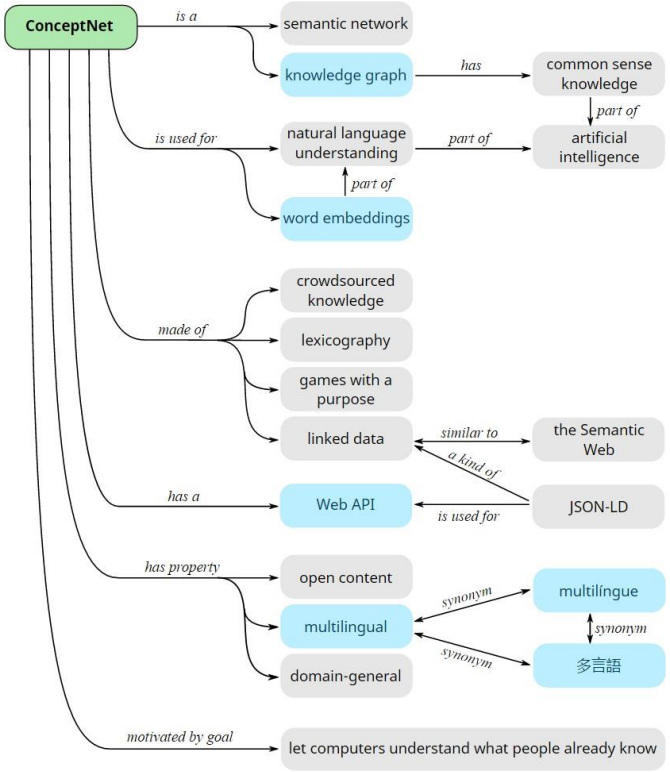
- Higher coverage over entities and concepts

KGs	# of Entities/Concepts	# of Relations
YAGO	10 Million	120 Million
DBpedia	28 Million	9.5 **Billion**
Probase	2.7 Million	70 **Billion**
BabelNet	14 Million	**5 Billion**
CN-DBpedia	17 Million	200 Million

KG优势2: semantically rich

- Higher coverage over numerous semantic relationships

KGs	# of Relations
DBpedia	1,650
YAGO1	14
YAGO3	74
CN-DBpedia	100 Thousands



KG优势3：high quality

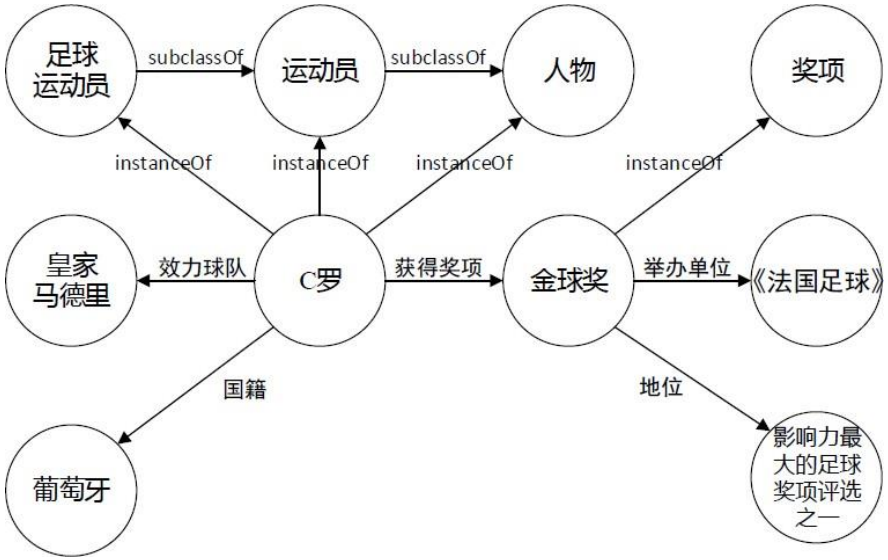
❑ High quality

- Big data: Cross validation by multiple sources
- Crowd sourcing: quality guarantee

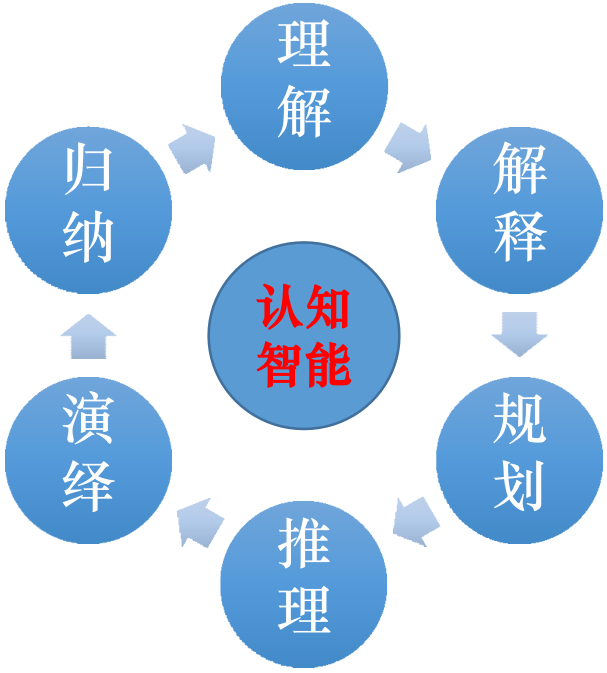
专职院士	中国工程院院士5人
专职院士	中国科学院院士15人
专职院士	国家重大科学研究计划首席科学家9人
中文名	中山大学
主管部门	中华人民共和国教育部
创办人	孙中山
创办时间	1924年
博士后	科研流动站41个

KG优势4: friendly structure

- Structured organization
  - By RDF
  - By graph



认知智能是智能化的关键



Can machine **think like humans**?

- 理解与解释是后深度学习时代人工智能的核心使命之一

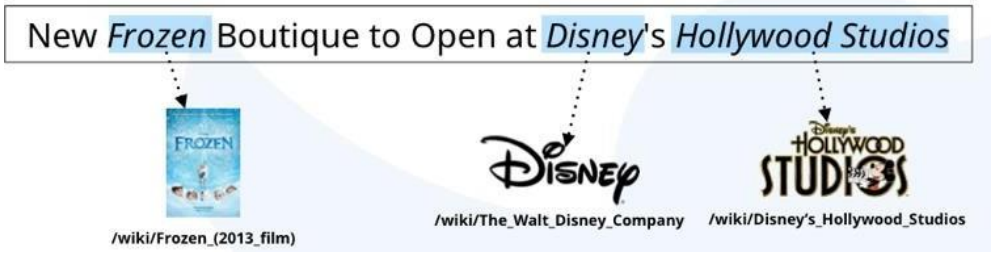
机器语言理解需要背景知识

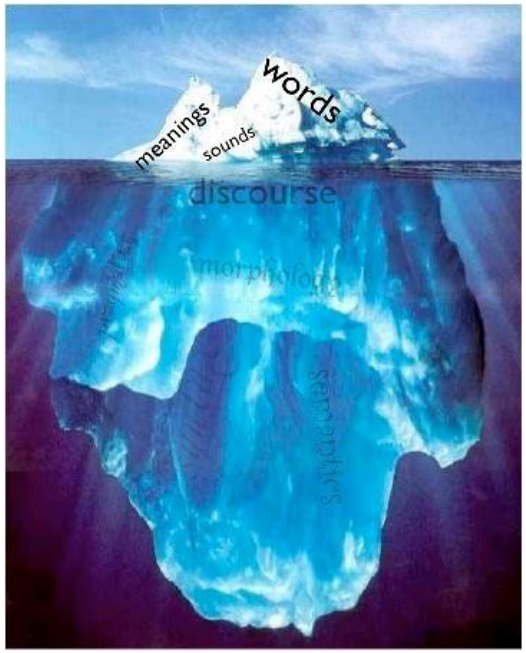
❑ Language is complicated

- **Ambiguous**, **contextual** and **implicit**
- Seemingly **infinite** number of ways to express the same meaning

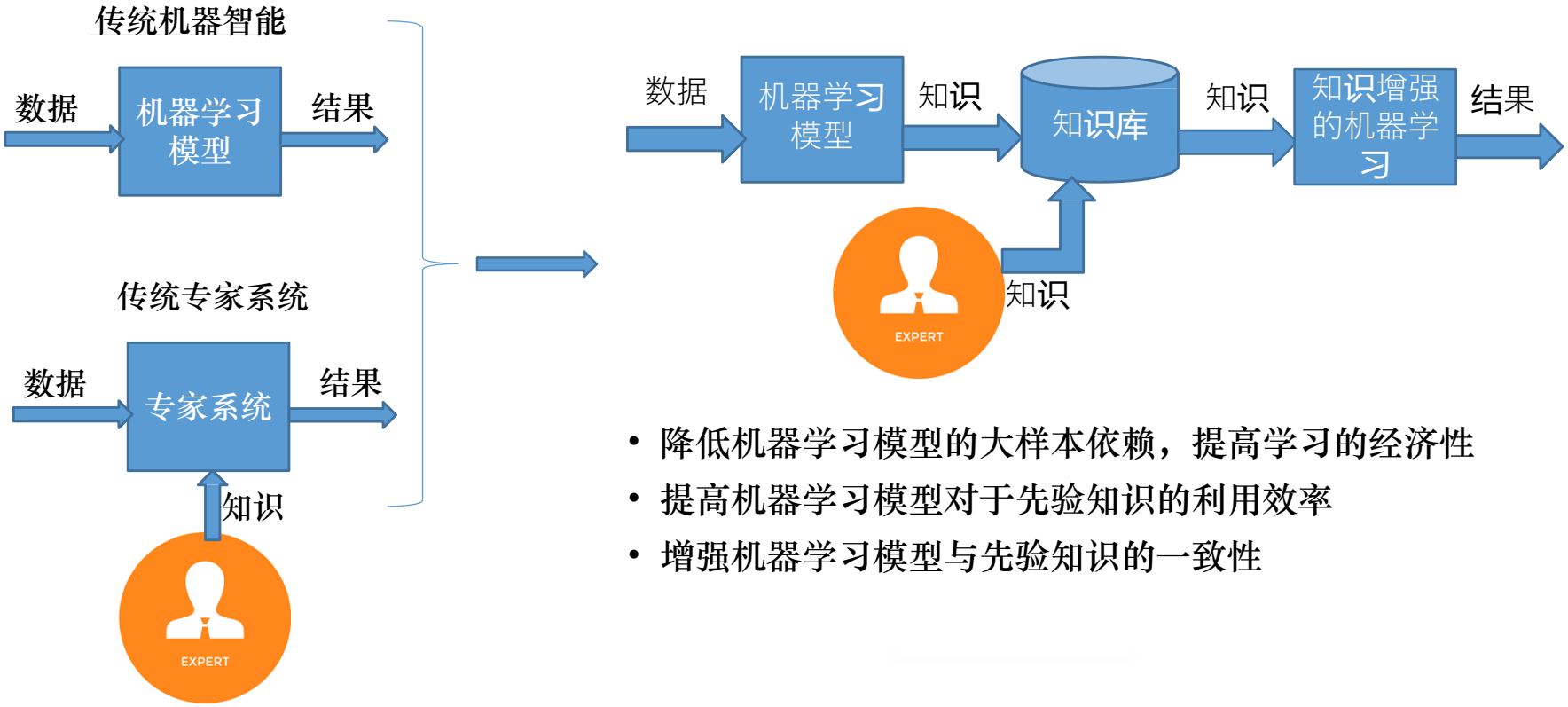
❑ Language understanding is difficult

- Grounded only in **human cognition**
- Needs significant **background knowledge**





知识增强机器学习能力



- 降低机器学习模型的大样本依赖，提高学习的经济性
- 提高机器学习模型对于先验知识的利用效率
- 增强机器学习模型与先验知识的一致性

知识图谱的关键技术

知识图谱的关键技术

- 知识图谱的架构
- 知识图谱的构建
- 知识图谱的管理

知识图谱的关键技术

- 知识图谱的架构
- 知识图谱的构建
- 知识图谱的管理

知识图谱的架构

知识图谱在逻辑上可分为**模式层**与**数据层**两个层次。

模式层：

- **模式层**构建在数据层之上，是知识图谱的核心，通常采用本体库来管理知识图谱的模式层
- 本体是结构化知识库的概念模板，通过本体库而形成的知识库不仅层次结构较强，并且冗余程度较小

模式层：实体-关系-实体，实体-属性-值

知识图谱的架构

知识图谱在逻辑上可分为**模式层**与**数据层**两个层次。

数据层：

- **数据层**主要是由一系列的事实组成，而知识将以事实为单位进行存储

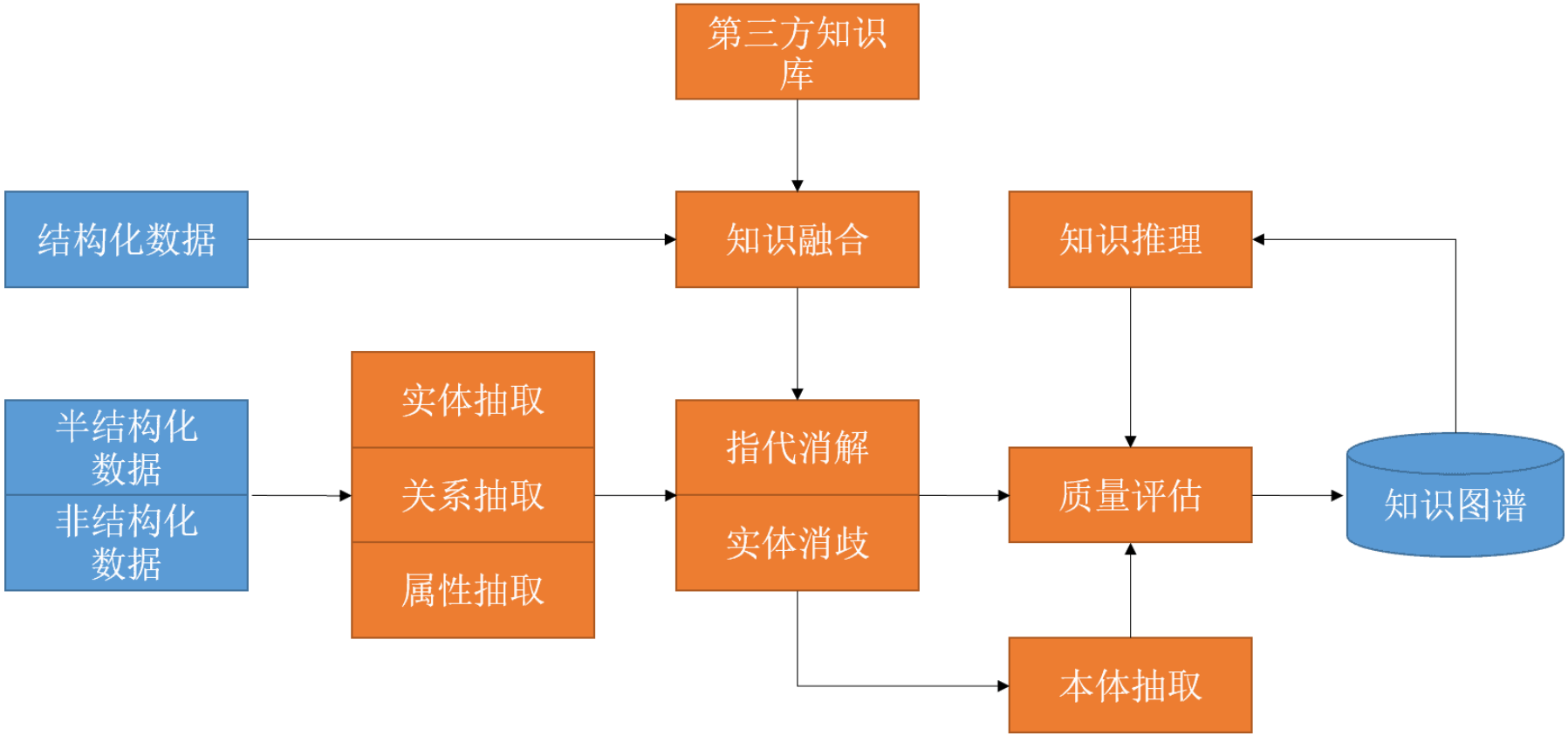
数据层：比尔盖茨-妻子-梅琳达盖茨，比尔盖茨-总裁-微软

知识图谱的关键技术

- 知识图谱的架构
- **知识图谱的构建**
- 知识图谱的管理

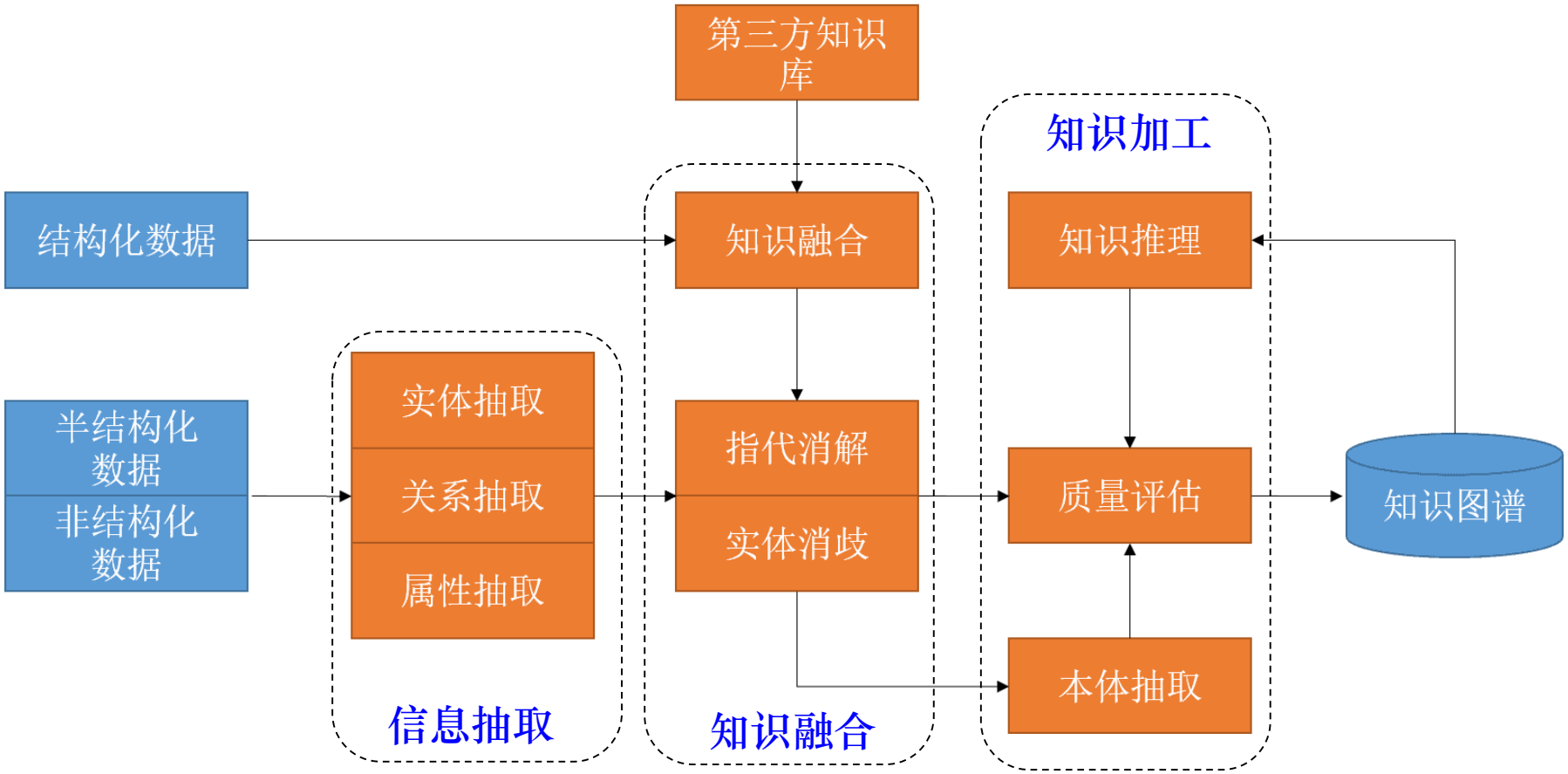
知识图谱的构建

知识图谱的技术架构：



知识图谱的构建

知识图谱的技术架构：



知识图谱的构建

知识图谱的技术架构：

•**信息抽取**：从各种类型的数据源中提取出实体、属性以及实体间的相互关系，在此基础上形成本体化的知识表达；

•**知识融合**：在获得新知识之后，需要对其进行整合，以消除矛盾和歧义，比如某些实体可能有多种表达，某个特定称谓也许对应于多个不同的实体等；

•**知识加工**：对于经过融合的新知识，需要经过质量评估之后（部分需要人工参与甄别），才能将合格的部分加入到知识库中，以确保知识库的质量。

知识图谱的构建：信息抽取

- 信息抽取（information extraction）是一种自动化地从半结构化和无结构数据中抽取实体、关系以及实体属性等结构化信息的技术
- 信息抽取是知识图谱构建的第一步，关键问题是：如何从异构数据源中自动抽取信息得到候选单元
- 涉及的关键技术包括：实体抽取、关系抽取和属性抽取

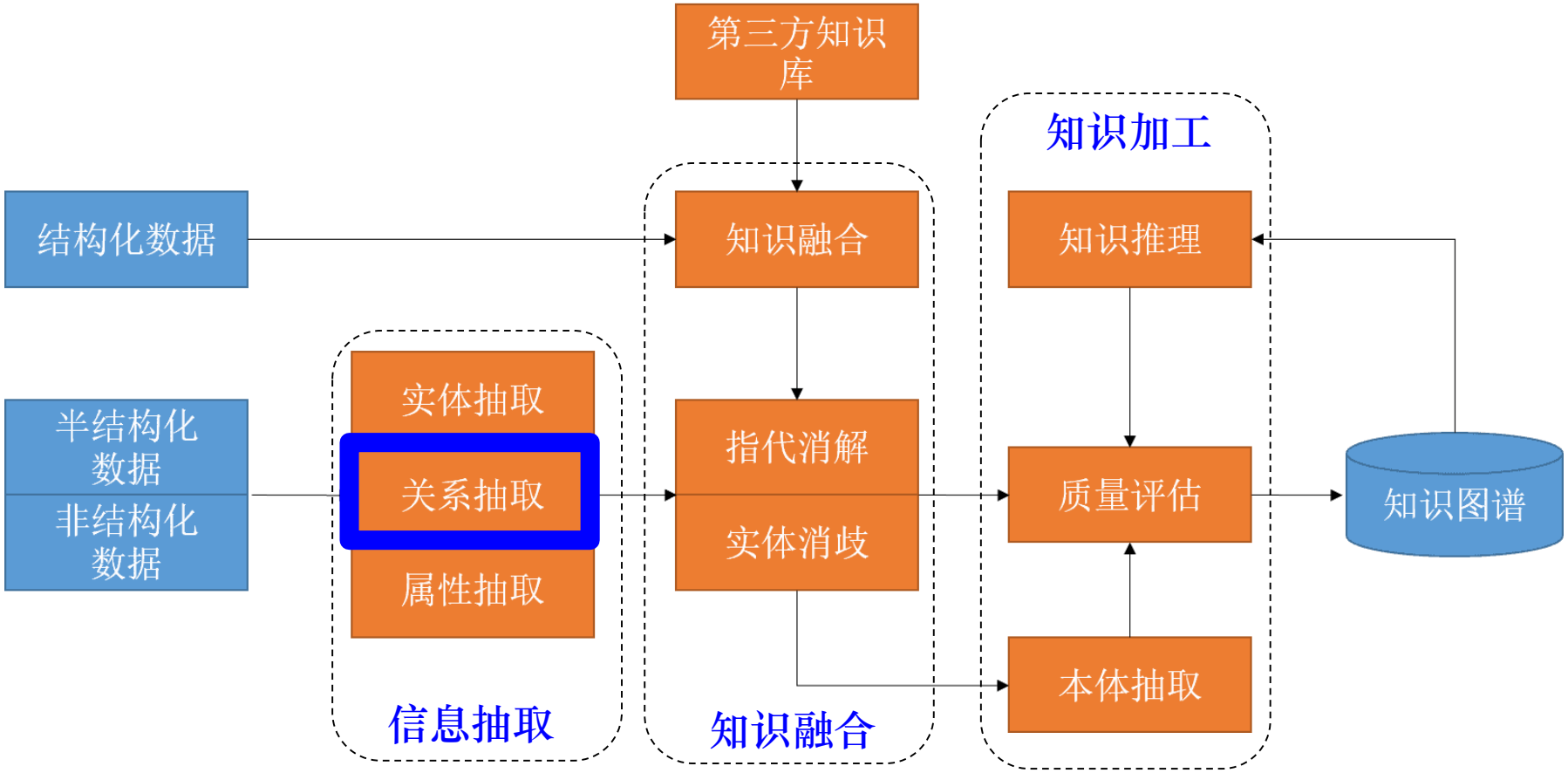
信息抽取：实体抽取

- 实体抽取又称为命名实体识别（named entity recognition，NER），是指从文本数据集中自动识别出命名实体
- 实体抽取的质量（准确率和召回率）对后续的知识获取效率和质量影响极大，因此是信息抽取中最为基础和关键的部分

可以根据已知的实体实例进行特征建模，利用该模型处理海量数据集得到新的命名实体列表，然后针对新实体建模，迭代地生成实体标注语料库

知识图谱的构建

知识图谱的技术架构：



信息抽取：关系抽取

- 文本语料经过实体抽取，得到的是离散的命名实体，为了得到语义信息，还需要从相关的语料中提取出实体之间的关联关系，通过关联关系将实体（概念）联系起来，形成网状的知识结构，研究关系抽取技术的目的，就是解决如何从文本语料中抽取实体间的关系这一基本问题

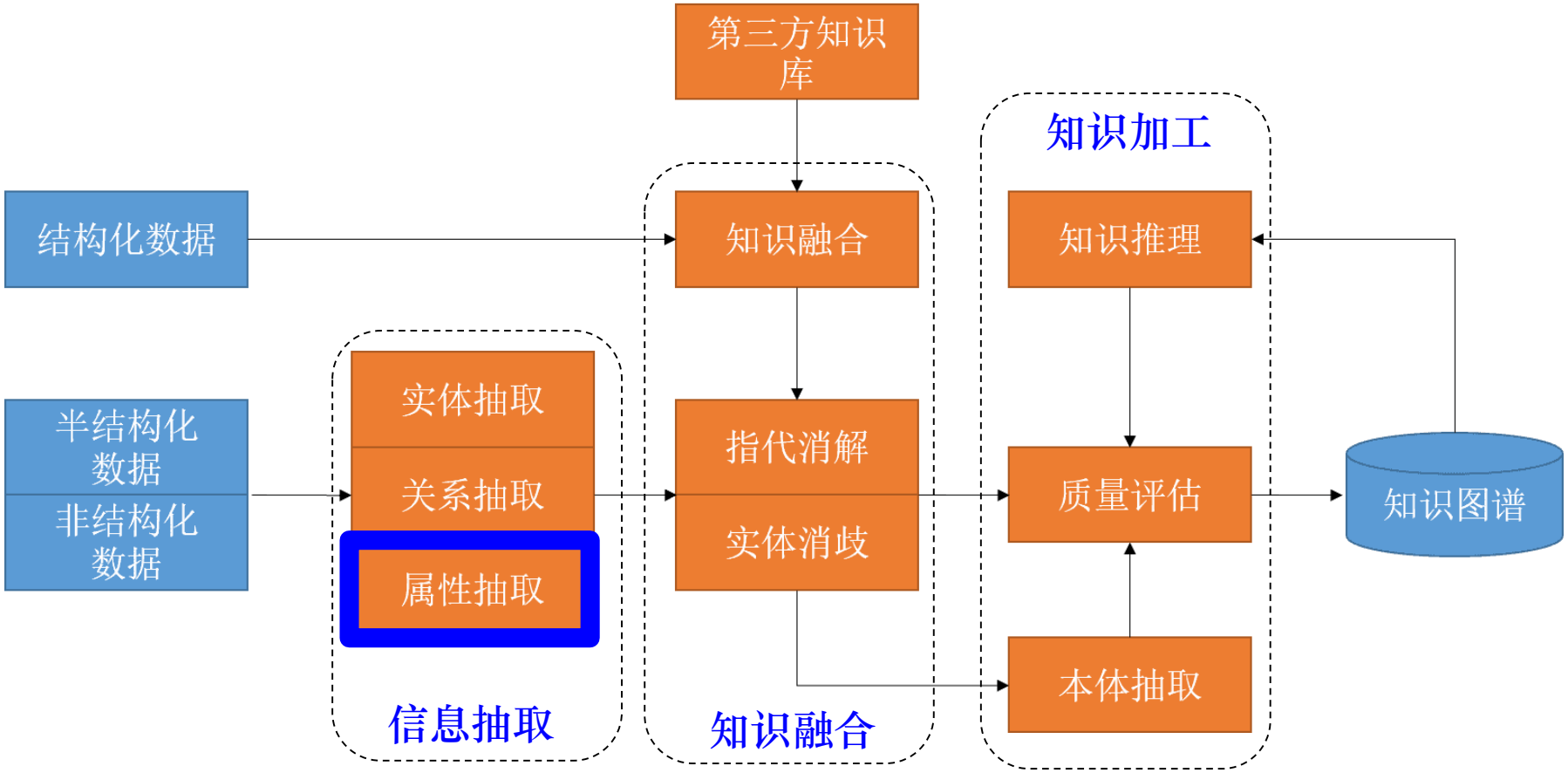
信息抽取：关系抽取

关系抽取技术:

1. 人工构造语法和语义规则（模式匹配）
2. 统计机器学习方法
3. 基于特征向量或核函数的有监督学习方法
4. 研究重点转向半监督和无监督
5. 开始研究面向开放域的信息抽取方法
6. 将面向开放域的信息抽取方法和面向封闭领域的传统方法结合

知识图谱的构建

知识图谱的技术架构：



信息抽取：属性抽取

- 属性抽取(Attribute Extraction)的目标是从不同信息源中采集特定实体的属性信息
- 例如针对某个公众人物，可以从网络公开信息中得到其昵称、生日、国籍、教育背景等信息

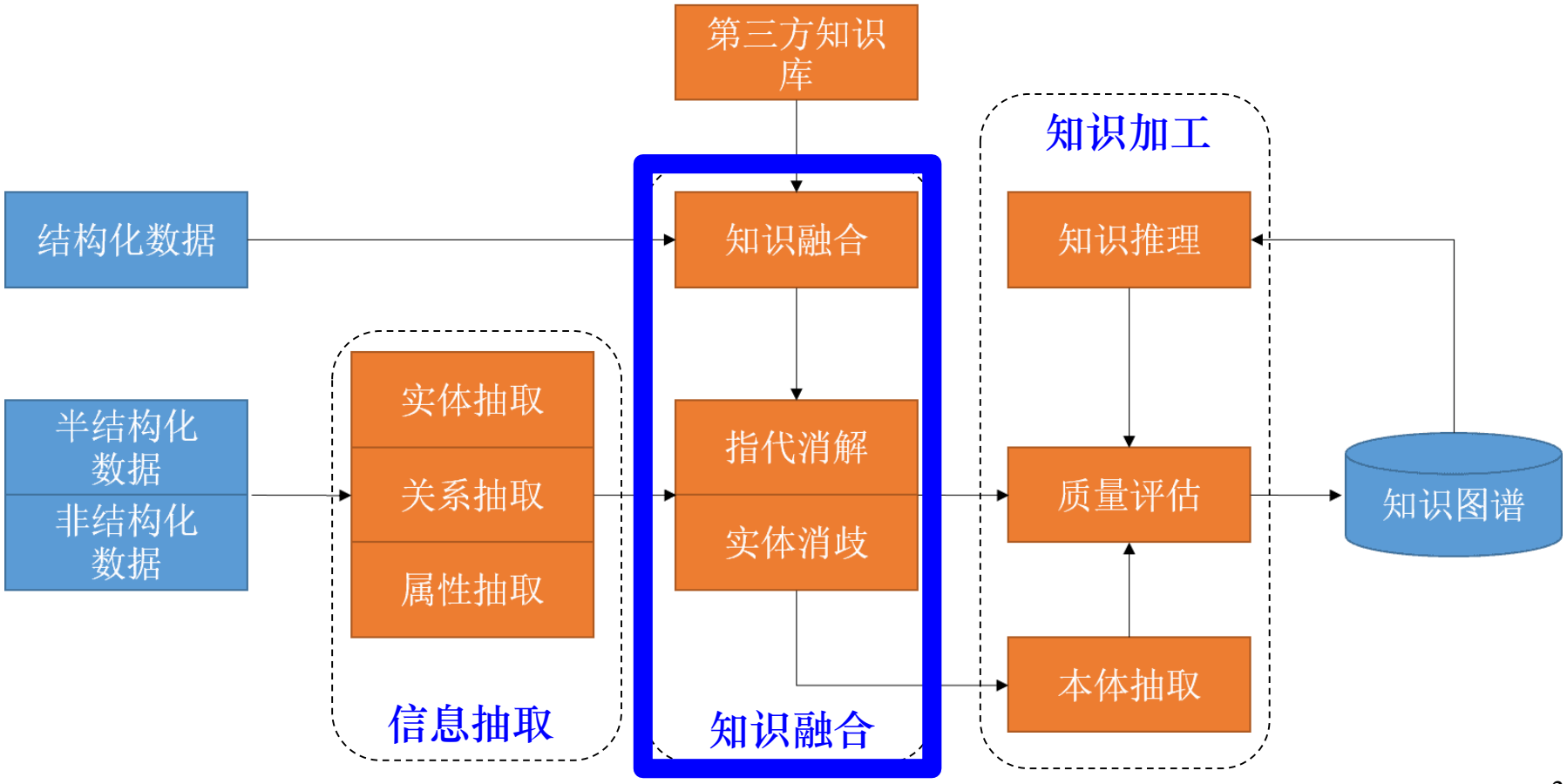
信息抽取：属性抽取

❑属性抽取(Attribute Extraction):

1. 将实体的属性视作实体与属性值之间的一种名词性关系，将属性抽取任务转化为关系抽取任务
2. 基于规则和启发式算法，抽取结构化数据
3. 基于半结构化数据，通过自动抽取生成训练语料，用于训练实体属性标注模型，然后将其应用于对非结构化数据的实体属性抽取
4. 采用数据挖掘的方法直接从文本中挖掘实体属性和属性值之间的关系模式，据此实现对属性名和属性值在文本中的定位。

知识图谱的构建

知识图谱的技术架构：



知识图谱的构建：知识融合

- 通过信息抽取，我们就从原始的非结构化和半结构化数据中获取到了实体、关系以及实体的属性信息
- 如果我们将接下来的过程比喻成拼图的话，那么这些信息就是拼图碎片，散乱无章，甚至还有从其他拼图里跑来的碎片、本身就是用来干扰我们拼图的错误碎片
- 拼图碎片（信息）之间的关系是扁平化的，缺乏层次性和逻辑性；拼图（知识）中还存在大量冗杂和错误的拼图碎片

知识图谱的构建：知识融合

- 知识融合包括：实体链接和知识合并
  - **实体链接**（entity linking）：是指对于从文本中抽取得到的实体对象，将其链接到知识库中对应的正确实体对象的操作
  - 其基本思想是首先根据给定的实体指称项，从知识库中选出一组候选实体对象，然后通过**相似度计算**将指称项链接到正确的实体对象

知识图谱的构建：知识融合

❑实体链接的流程：

1. 从文本中通过实体抽取得到实体指称项
2. 进行**实体消歧**和**共指消解**，判断知识库中的同名实体与之是否代表不同的含义，以及是否有其他实体与之表示相同的含义
3. 在确认知识库中对应的正确实体对象之后，将该实体指称项链接到知识库中对应实体
4. **实体消歧**：专门用于解决同名实体产生歧义问题的技术，通过实体消歧，就可以根据当前的语境，准确建立实体链接，实体消歧主要采用聚类法
5. **共指消解**：**主要用于解决多个指称对应同一实体对象的问题**。在一次会话中，多个指称可能指向的是同一实体对象。利用共指消解技术，可以将这些指称项关联到正确的实体对象

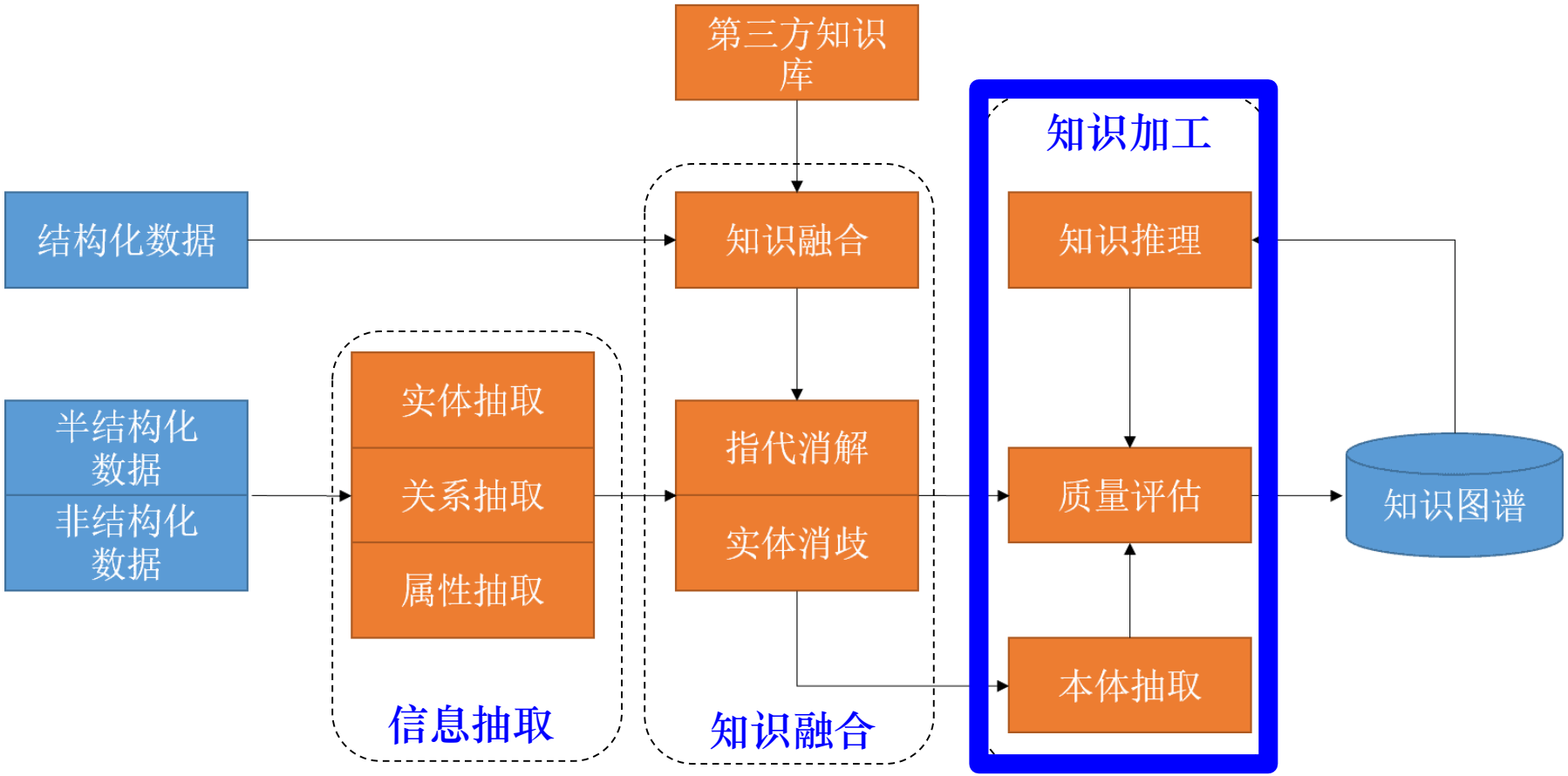
知识图谱的构建：知识融合

❑知识合并

- 构建知识图谱时，可从第三方知识库或结构化数据获取输入
- 将外部知识库融合到本地知识库需要处理两个层面的问题：
  - 数据层的融合，包括实体的指称、属性、关系以及所属类别等，主要的问题是如何避免实例以及关系的冲突问题，造成不必要的冗余
  - 通过模式层的融合，将新得到的本体融入已有的本体库中
- 然后是合并关系数据库，在知识图谱构建过程中，一个重要的高质量知识来源是企业或者机构自己的关系数据库。为了将这些结构化的历史数据融入到知识图谱中，可以采用资源描述框架（RDF）作为数据模型，其实质就是将关系数据库的数据换成RDF的三元组数据

知识图谱的构建

知识图谱的技术架构：



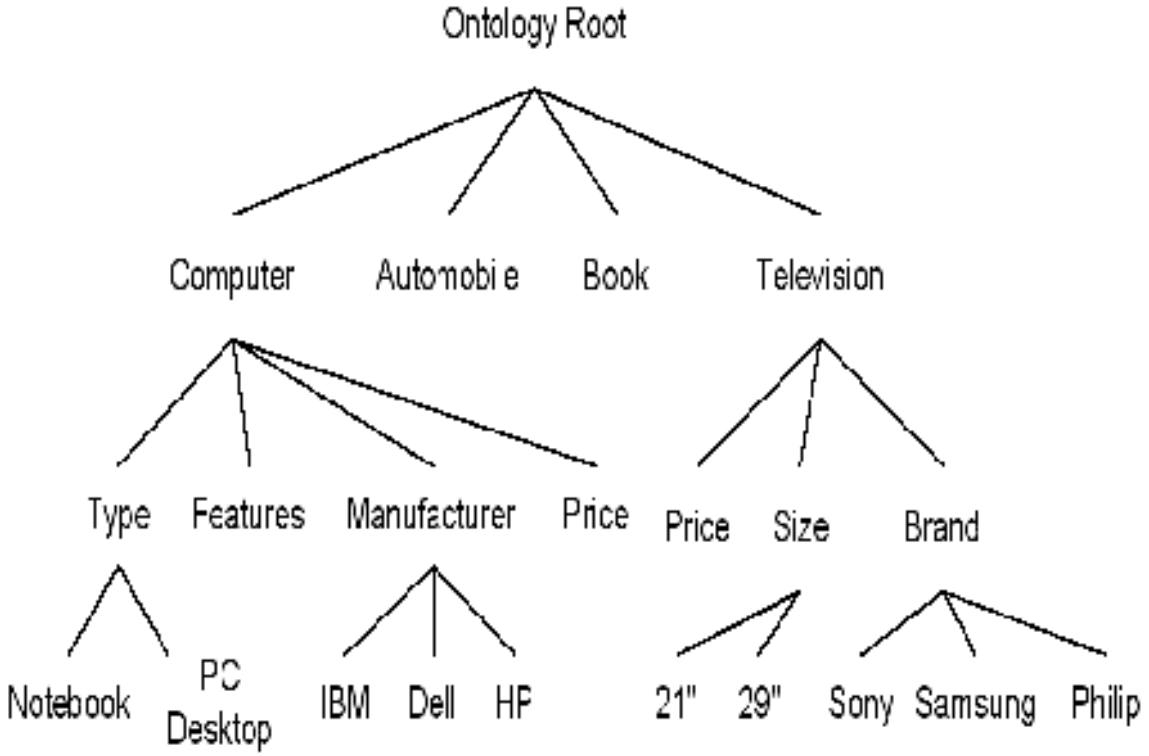
知识图谱的构建：知识加工

- 通过信息抽取，从原始语料中提取出了实体、关系与属性等知识要素；经过知识融合，消除实体指称项与实体对象之间的歧义，得到一系列基本的事实表达
- 然而事实本身并不等于知识。要想最终获得结构化、网络化的知识体系，还需要经历知识加工的过程
- 知识加工主要包括：**本体构建**、**知识推理**和**质量评估**

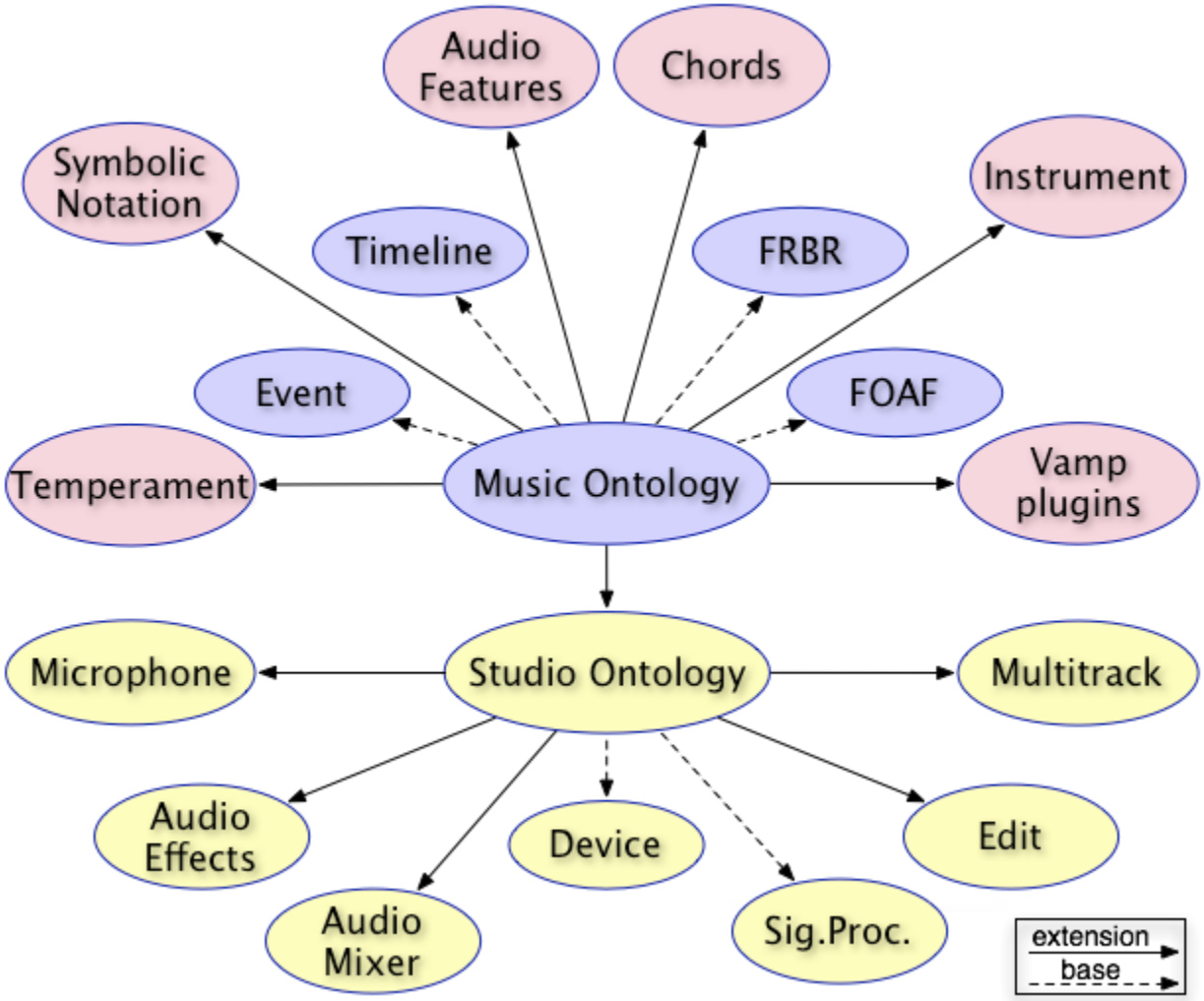
知识加工：本体构建

- 本体（ontology）是指人工的概念集合、概念框架，如“人”、“事”、“物”等
- 本体可以采用人工编辑的方式手动构建，也可以以数据驱动的自动化方式构建本体。因为人工方式工作量巨大，且很难找到符合要求的专家，因此当前主流的全局本体库产品，都是从一些面向特定领域的现有本体库出发，采用自动构建技术逐步扩展得到的
- 自动化本体构建过程包含三个阶段：
  - 实体并列关系相似度计算
  - 实体上下位关系抽取
  - 本体的生成

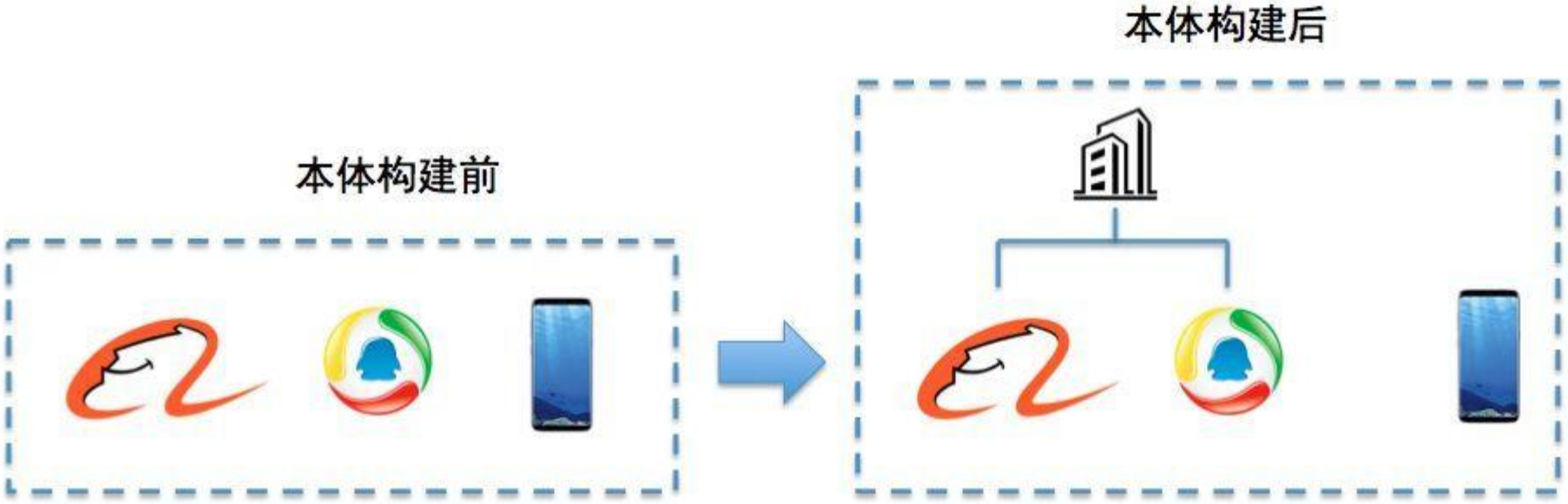
知识加工：本体构建


Ontology Root
Computer
Automobile
Book
Television
Type
Features
Manufacturer
Price
Price
Size
Brand
Notebook
PC
Desktop
IBM
Dell
HP
21"
29"
Sony
Samsung
Philip

知识加工：本体构建

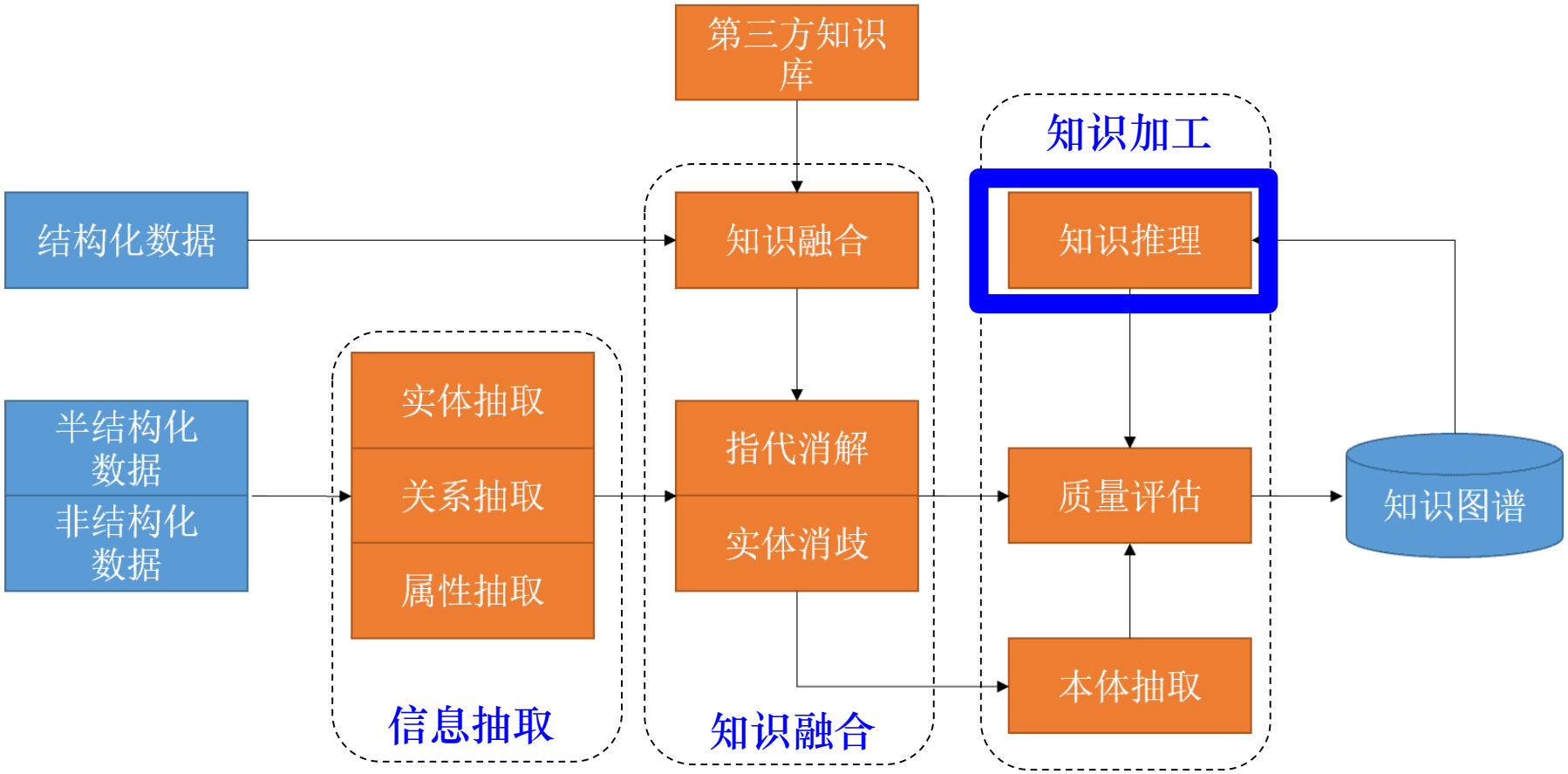

Audio Features
Chords
Symbolic Notation
Instrument
Timeline
FRBR
Event
FOAF
Temperament
Music Ontology
Vamp plugins
Microphone
Studio Ontology
Multitrack
Audio Effects
Device
Edit
Audio Mixer
Sig.Proc.
extension
base

知识加工：本体构建



知识图谱的构建

知识图谱的技术架构：



知识加工：知识推理

- 本体构建之后，一个知识图谱的雏形便已经搭建好了。但此时知识图谱之间的关系大多是残缺的，可以使用知识推理技术完成知识发现
- 我们可以发现：如果A是B的配偶，B是C的主席，C坐落于D，那么我们就可以认为，A生活在D这个城市
- 当然知识推理的对象也并不局限于实体间的关系，也可以是实体的属性值，本体的概念层次关系等
- 推理属性值：已知某实体的生日属性，可以通过推理得到该实体的年龄属性
- 推理概念：已知(老虎，科，猫科)和（猫科，目，食肉目）可以推出（老虎，目，食肉目）

知识加工：质量评估

- 质量评估也是知识库构建技术的重要组成部分，这一部分存在的意义在于：可以对知识的可信度进行量化，通过舍弃置信度较低的知识来保障知识库的质量

知识图谱的关键技术

- 知识图谱的架构
- 知识图谱的构建
- **知识图谱的管理**

知识图谱管理

知识图谱的管理，主要是知识库的更新，包括**概念层的更新**和**数据层的更新**:

- 概念层的更新是指新增数据后获得了新的概念，需要自动将新的概念添加到知识库的概念层中
- 数据层的更新主要是新增或更新实体、关系、属性值，对数据层进行更新需要考虑数据源的可靠性、数据的一致性（是否存在矛盾或冗杂等问题）等可靠数据源，并选择在各数据源中出现频率高的事实和属性加入知识库

知识图谱管理

知识图谱的内容更新有两种方式：

- **全面更新**：指以更新后的全部数据为输入，从零开始构建知识图谱。这种方法比较简单，但资源消耗大，而且需要耗费大量人力资源进行系统维护
- **增量更新**：以当前新增数据为输入，向现有知识图谱中添加新增知识。这种方式资源消耗小，但目前仍需要大量人工干预（定义规则等），因此实施起来十分困难

Thank you!

权小军 中山大学数据科学与计算机学院