AI科技大本营 达观数据 DATA GRAND

达观数据
知识图谱关键技术与应用

全球领先的文本智能处理专家

达观数据
DATA GRAND

目录 CONTENTS

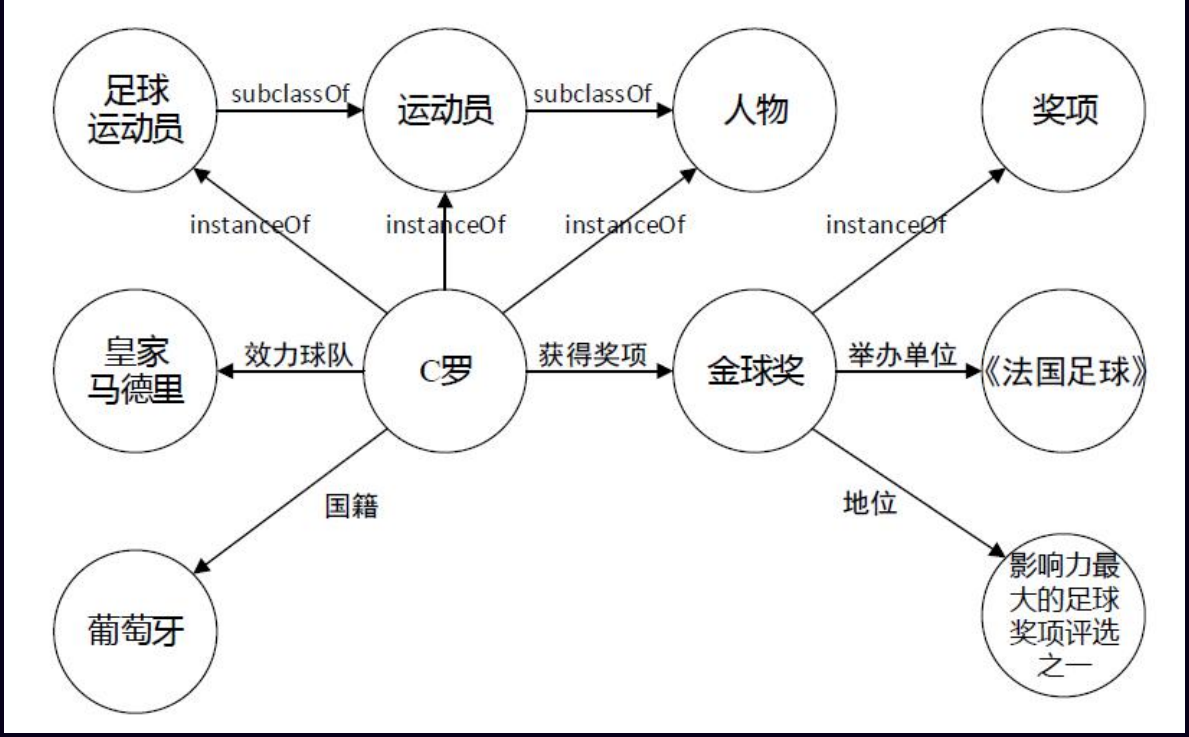
达观数据
DATA GRAND

1

THE FIRST

知识图谱概述

知识图谱的直观展示



- 知识图谱本质上是一种语义网络，将客观的经验沉淀在巨大的网络中
- 结点代表实体（entity）或者概念（concept）
- 边（edge）代表实体/概念之间的语义关系

知识图谱的表示方法

达观数据
DATA GRAND

构成知识图谱的核心三元组

- 三元组 实体 属性 关系， Entity， Attribute， Relation
- 抽取为 <实体1，关系，实体2> 和 <实体1， 属性1， 属性值1>
- 例如 <达观数据， is-a， 人工智能公司>；<人工智能公司， subclass，高科技公司>；<达观数据，start-time， 2015年>

基于已有的三元组，可以推导出新的关系

- 通过关系可以推导出新的关系。例如 <人工智能公司 subclass 高科技公司> <Google is-a 人工智能公司> -→ 可以推导出 <Google is-a 高科技公司>， 因为subclass的实例可以有继承关系.
- 又例如 <翅膀 part-of 鸟> <麻雀 kind-of 鸟> -→ 推导出 <翅膀 part-of 麻雀>

为什么要使用三元组来描述知识图谱

- 三元组是一种简单的易于人类解读的结构
- 三元组方便编写计算机程序来进行抽取和加工处理

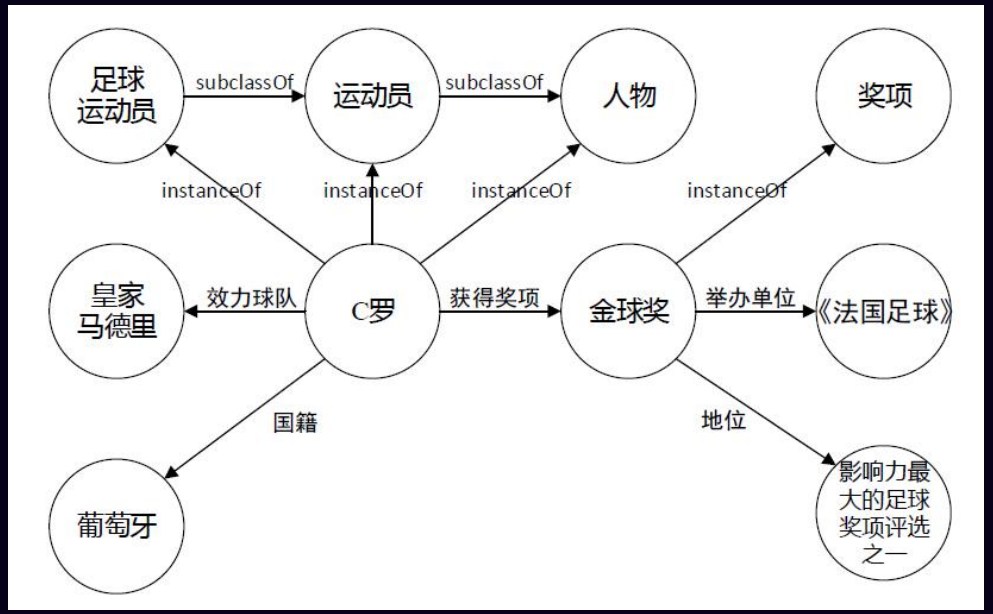
AI为什么需要知识图谱

- AI需要从感知智能迈向认知智能，认知的建立需要思考的能力，而思考是建立在知识基础上的

- 知识图谱富含实体、概念、属性、事件、关系等信息，基于一定的知识推理为可解释性AI提供了全新的视角和机遇

C罗为啥辣么牛？

- 知识图谱有助于消除自然语言和深度学习黑盒之间的语义鸿沟



通过知识图谱中的概念理解自然语言

达观数据
DATA GRAND

在问答研究中，自然语言问题的理解或者语义表示是一个难题。同样语义的问题表达方式往往是多样的。

- how many people are there in Shanghai?
- what is the population of Shanghai?
- 如果用同样的句式问及北京、南京，甚至任何一个城市人口呢?
- How many people are there in $City ?
- shanghai is_a city，beijing is_a city

通用知识图谱 vs 行业知识图谱

- 面向通用领域
- 以常识性知识为主
- “结构化的百科知识”
- 强调知识的广度
- 使用者是普通用户

- 面向某一特定领域
- 基于行业数据构建
- “基于语义技术的行业知识库”
- 强调知识的深度
- 使用者是行业人员

达观数据
DATA GRAND

2
THE SECOND

知识图谱典型行业应用介绍

知识图谱助力多个行业的人工智能应用

达观数据
DATA GRAND

- Google
- Bing
- 百度

- 微软小冰
- 公子小白

聊天机器人

- IBM Watson Health

决策支持

- Siri
- Google Now
- 微软小娜
- 百度度秘

私人助理

- Apple Watch
- Ticwatch

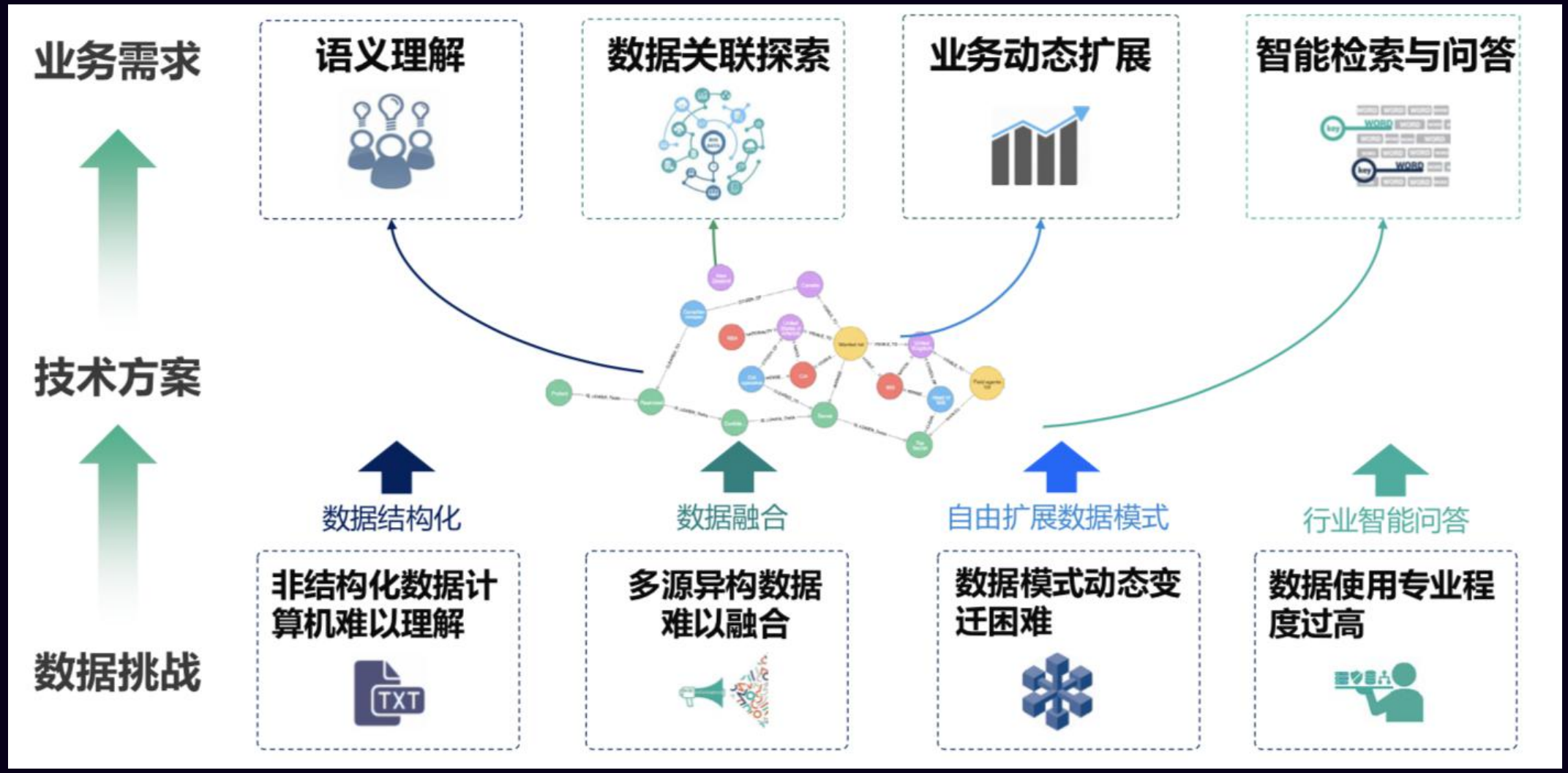
智能硬件

- 智能家居
- 智能厨房

智能硬件

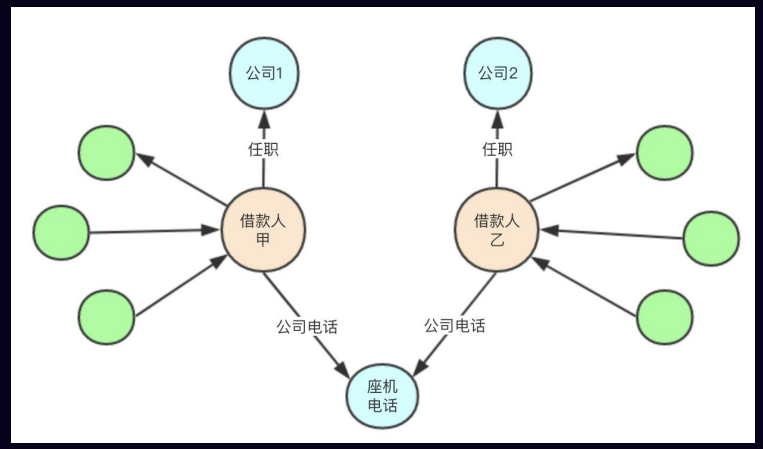
知识图谱助力企业实现商业智能

达观数据
DATA GRAND

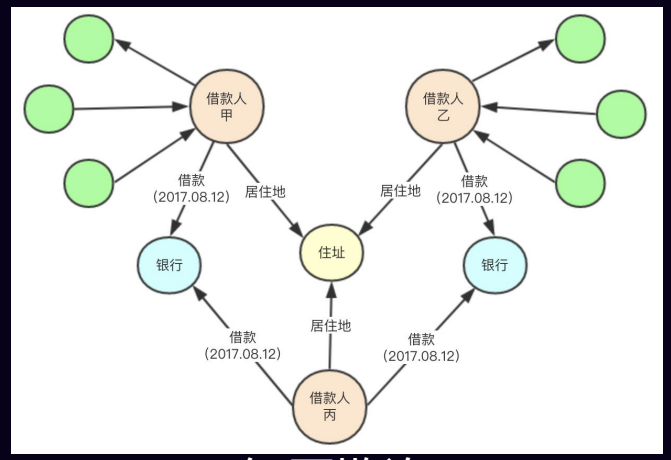


知识图谱金融行业应用1：风控反欺诈

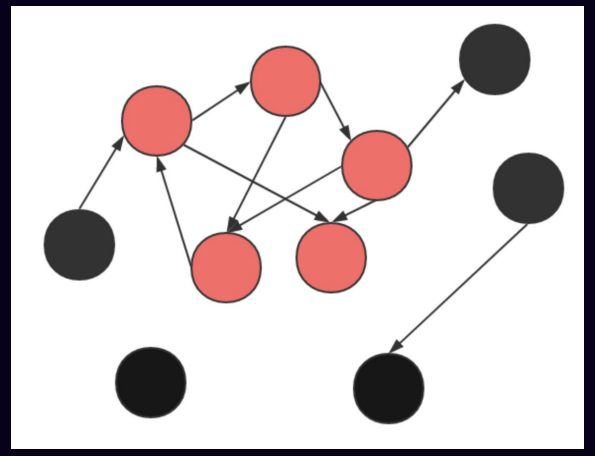
达观数据
DATA GRAND



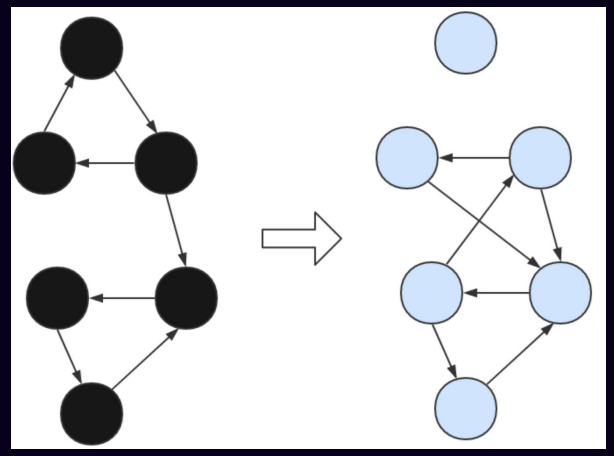
不一致性验证



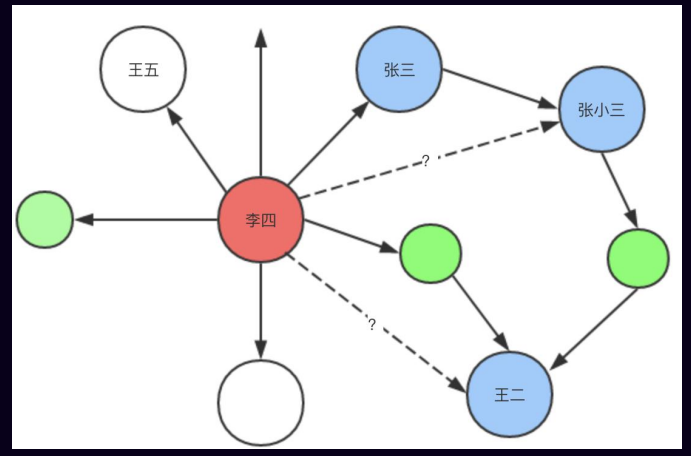
组团欺诈



静态异常检测



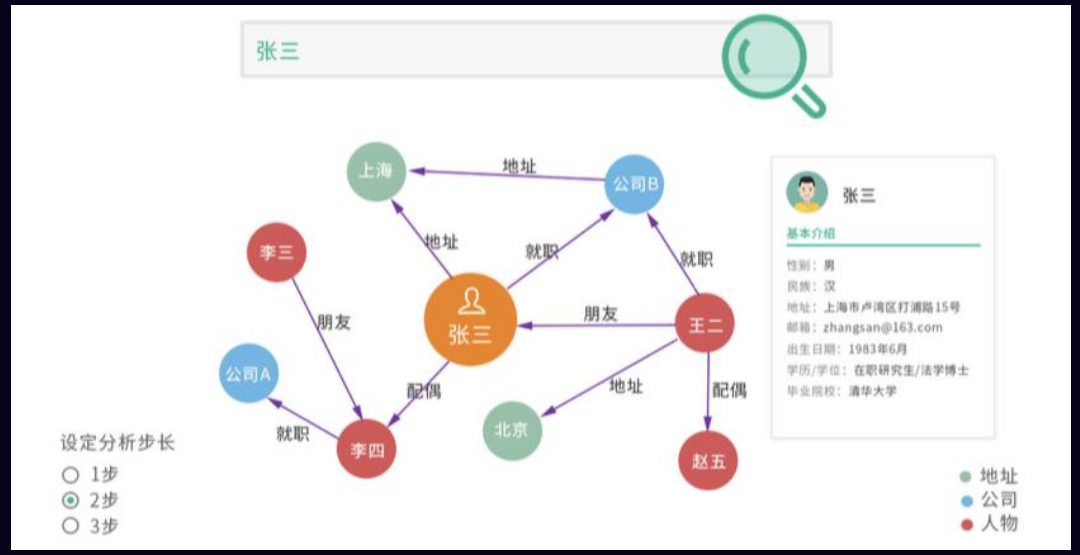
动态异常检测

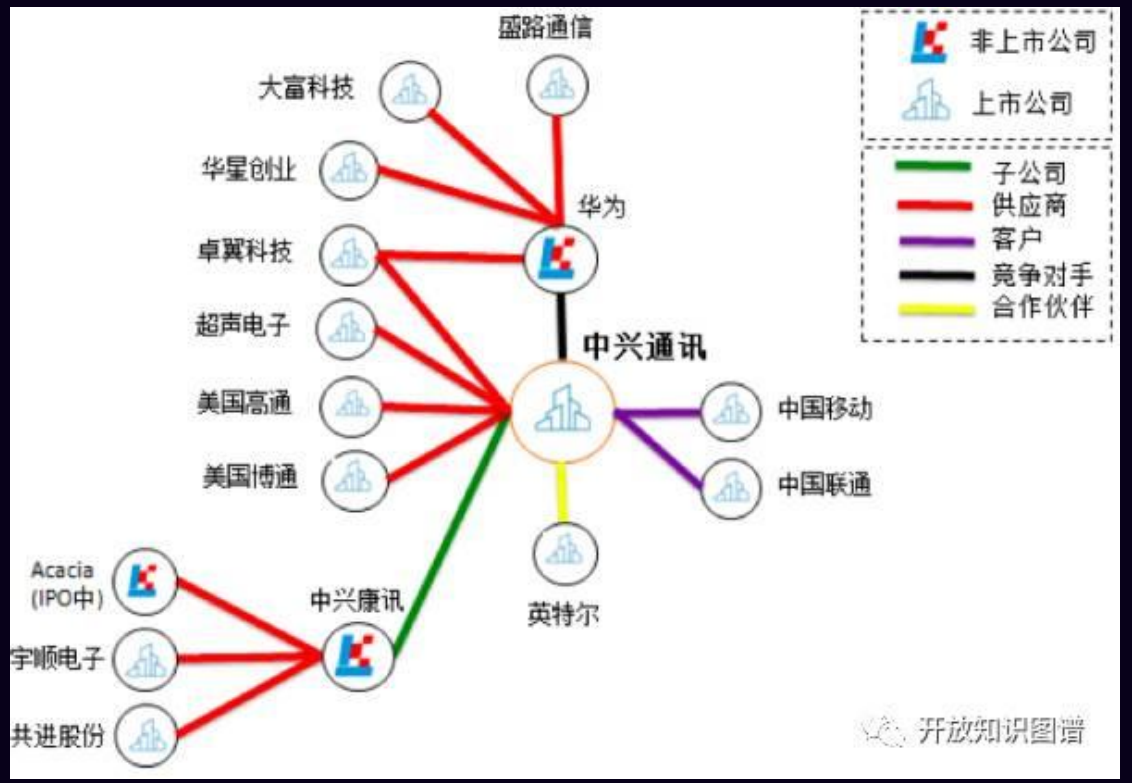


失联客户管理

知识图谱行业应用2：辅助信贷审核投研分析

达观数据
DATA GRAND





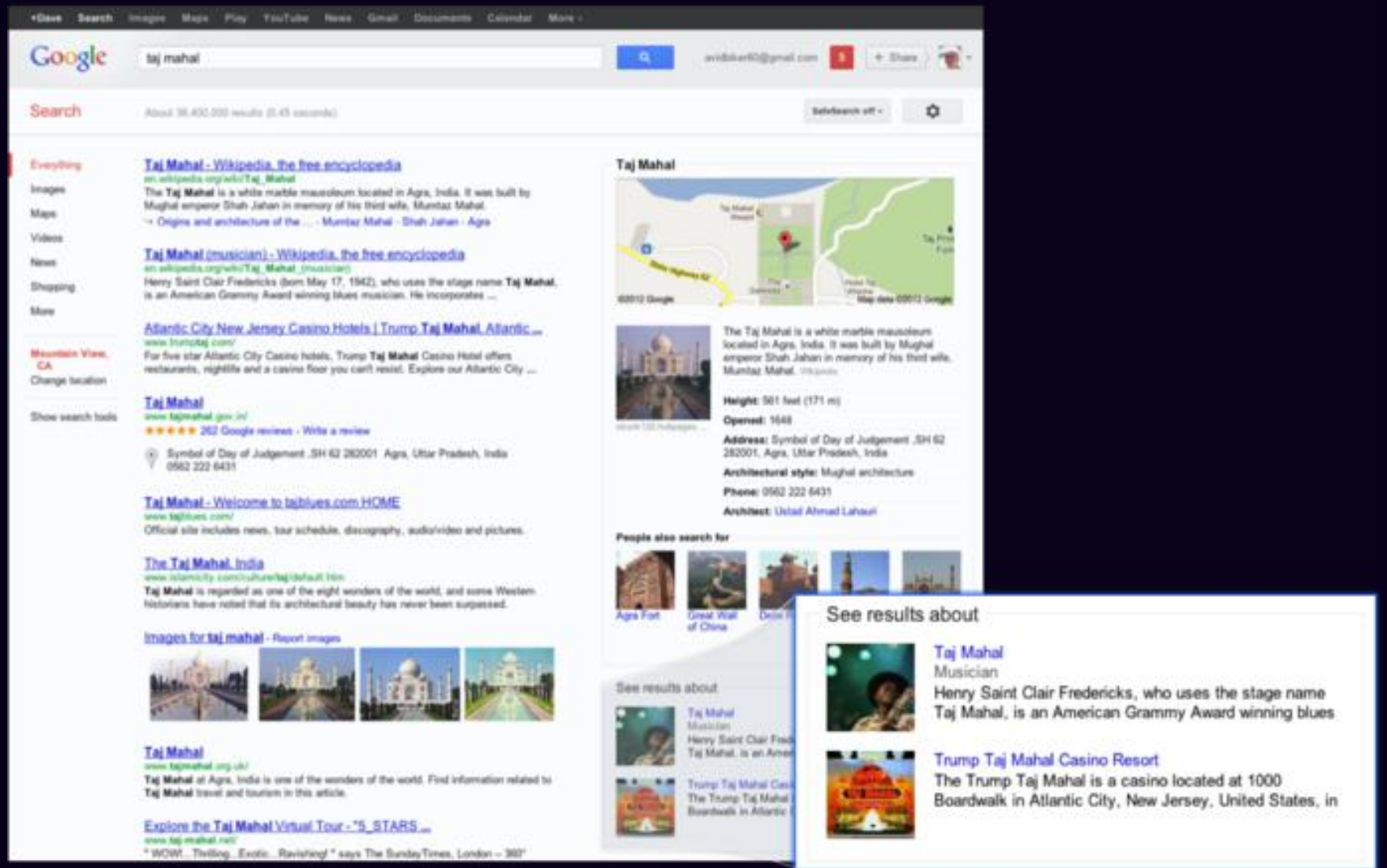
知识图谱行业应用4：精准营销

达观数据
DATA GRAND

“A knowledge graph allows you to take core information about your customer—their name, where they reside, how to contact them—and relate it to who else they know, how they interact on the web, and more”-- Michele Goetz, a Principal Analyst at Forrester Research

知识图谱在搜索引擎中的应用：谷歌知识图谱

达观数据
DATA GRAND



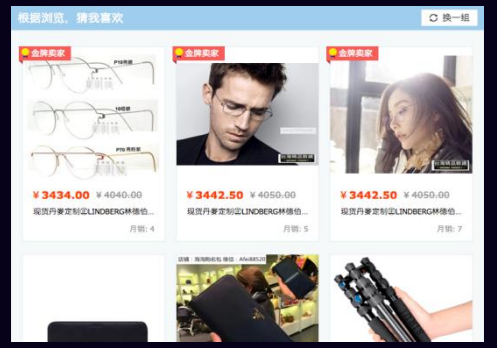
1. Find the right thing

2. Get the best summary

3. Go deeper and broade

利用知识图谱来提供个性化推荐

- 场景化推荐 （沙滩鞋->游泳衣、防晒霜、海岛度假产品）
- 任务型推荐 （牛肉卷、羊肉卷->火锅底料、电磁炉？ ；螺丝、螺钉->多功能螺丝刀）
- 冷启动环境下推荐 （语义标签：摄影VS旅游；相同导演或相同主演的电影；）
- 跨领域推荐 （微博如何推荐淘宝商品？用户经常晒九寨沟、黄山、泰山的照片->淘宝登山装备 ）
- 知识型推荐 （清华大学、北京大学->复旦大学（985名校);阿里、百度->腾讯（互联网BAT等）



- 精准感知任务与场景，想用户之未想
- 从基于行为的推荐发展到行为与语义融合的智能推荐

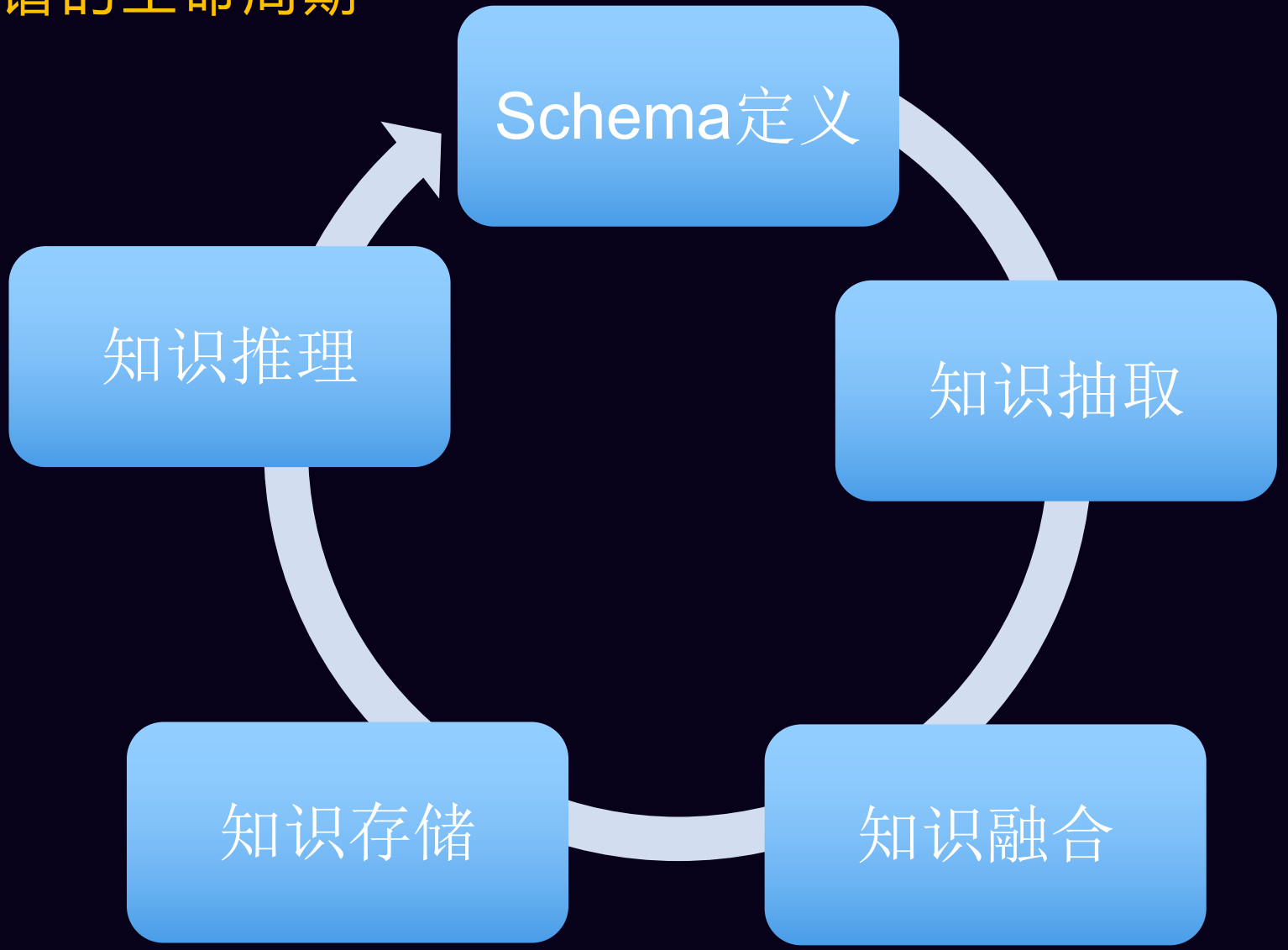
达观数据
DATA GRAND

3

THE THIRD

如何构建知识图谱？

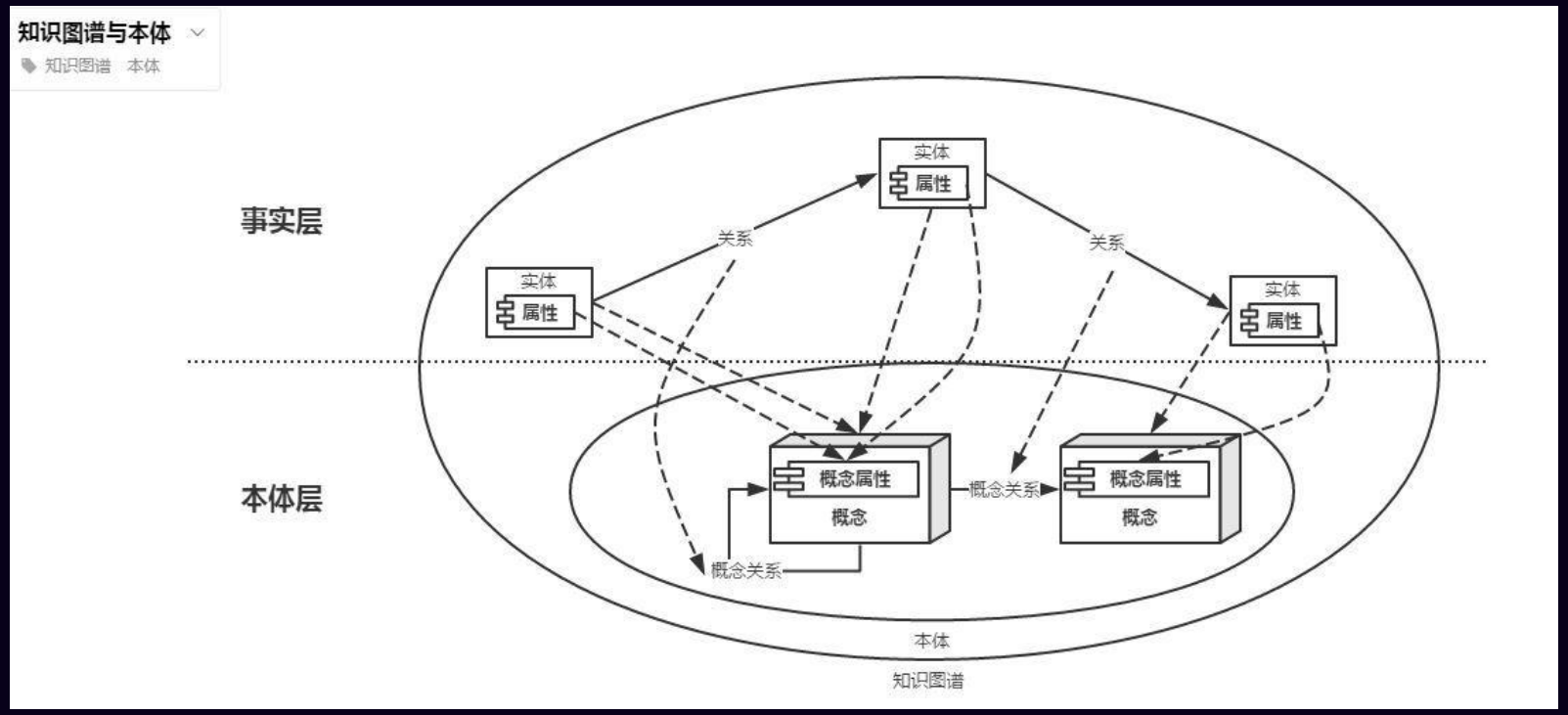
知识图谱的生命周期



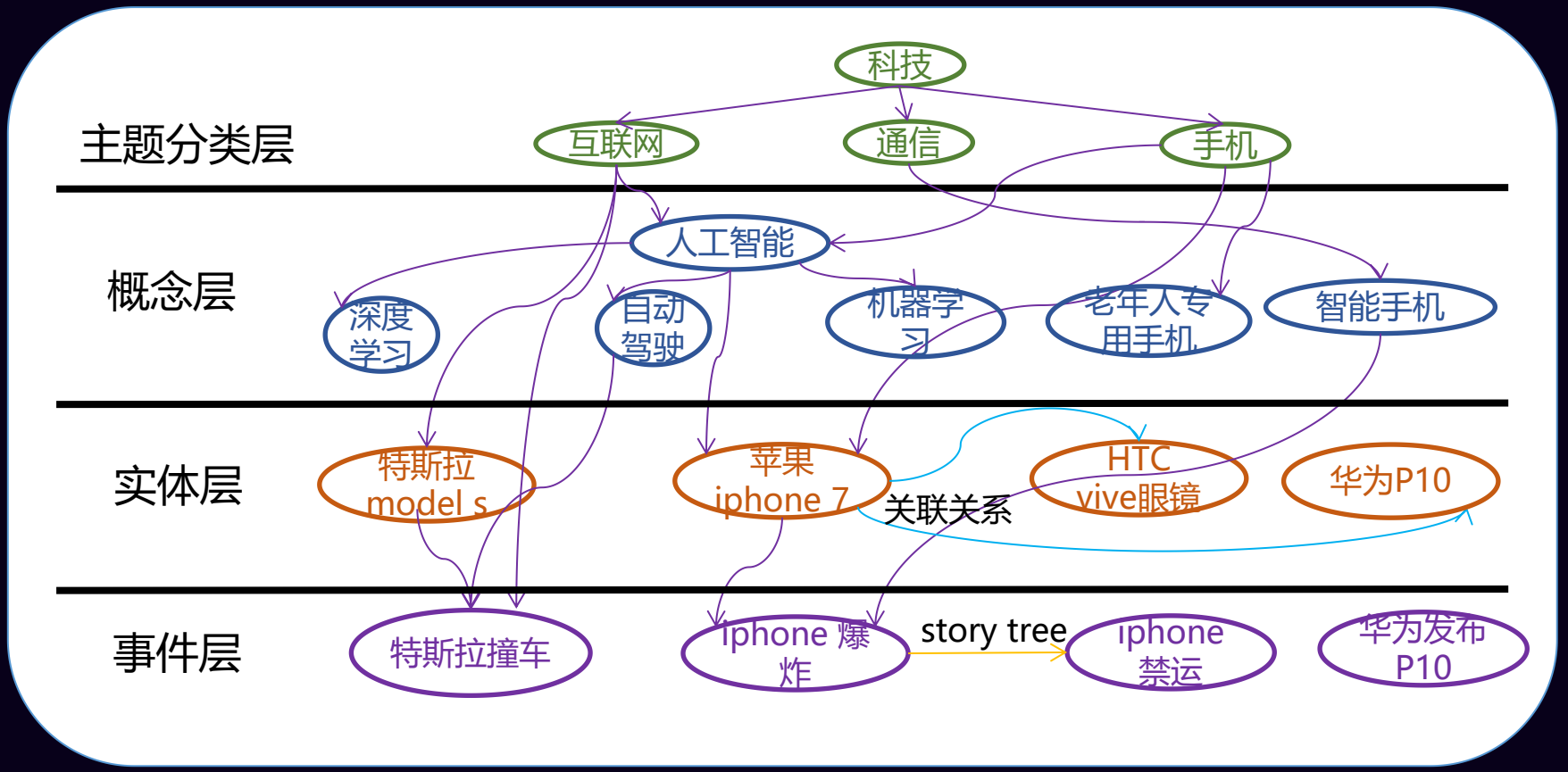
本体的概念

本体（Ontology），在维基百科的定义是：

> In computer science and information science, an ontology is a **formal naming and definition of the types, properties, and interrelationships of the entities** that really or fundamentally exist for **a particular domain** of discourse. It is thus a practical application of philosophical ontology, with a taxonomy.



本体层（模式，Schema）的定义的栗子



主题分类： 汽车; 美系汽车
概念： 美系豪华车；
驾驶感出众的车
实体： 凯迪拉克XT5
事件： 凯迪拉克XT5发售


凯迪拉克XT5轻混驾到，是雪中送炭还是锦上添花?

爱车兵团
09-04 18:58

+关注

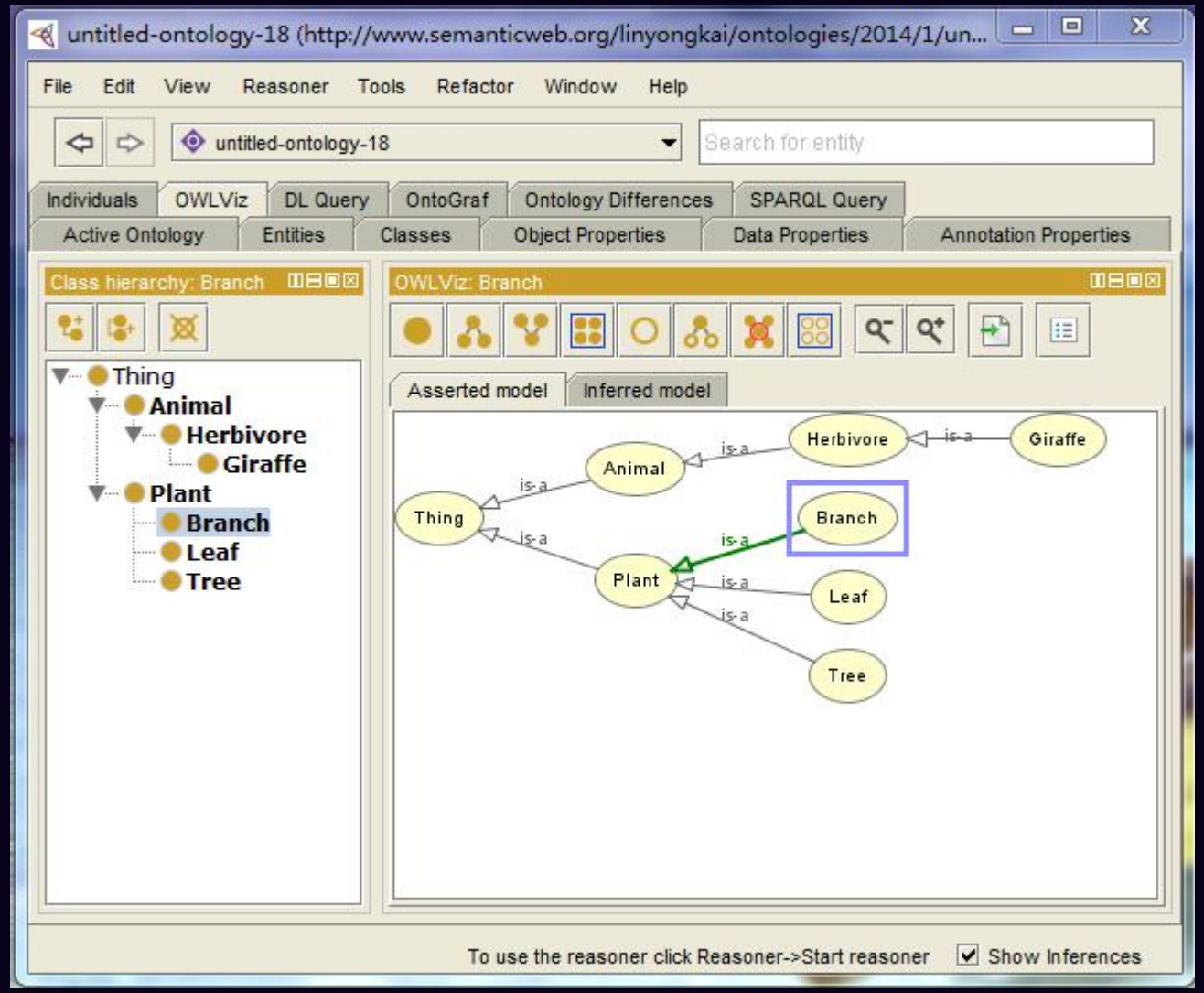
凯迪拉克XT5作为一款美式都市豪华SUV，它有着不随大流，车身线条独特的外观设计，以及科技感十足的内饰设计，再加上那个神奇的尾箱容量，就已经征服不少拥有多重身份的成功人士。

而这次兵哥所试驾的XT5轻混动版，是在原

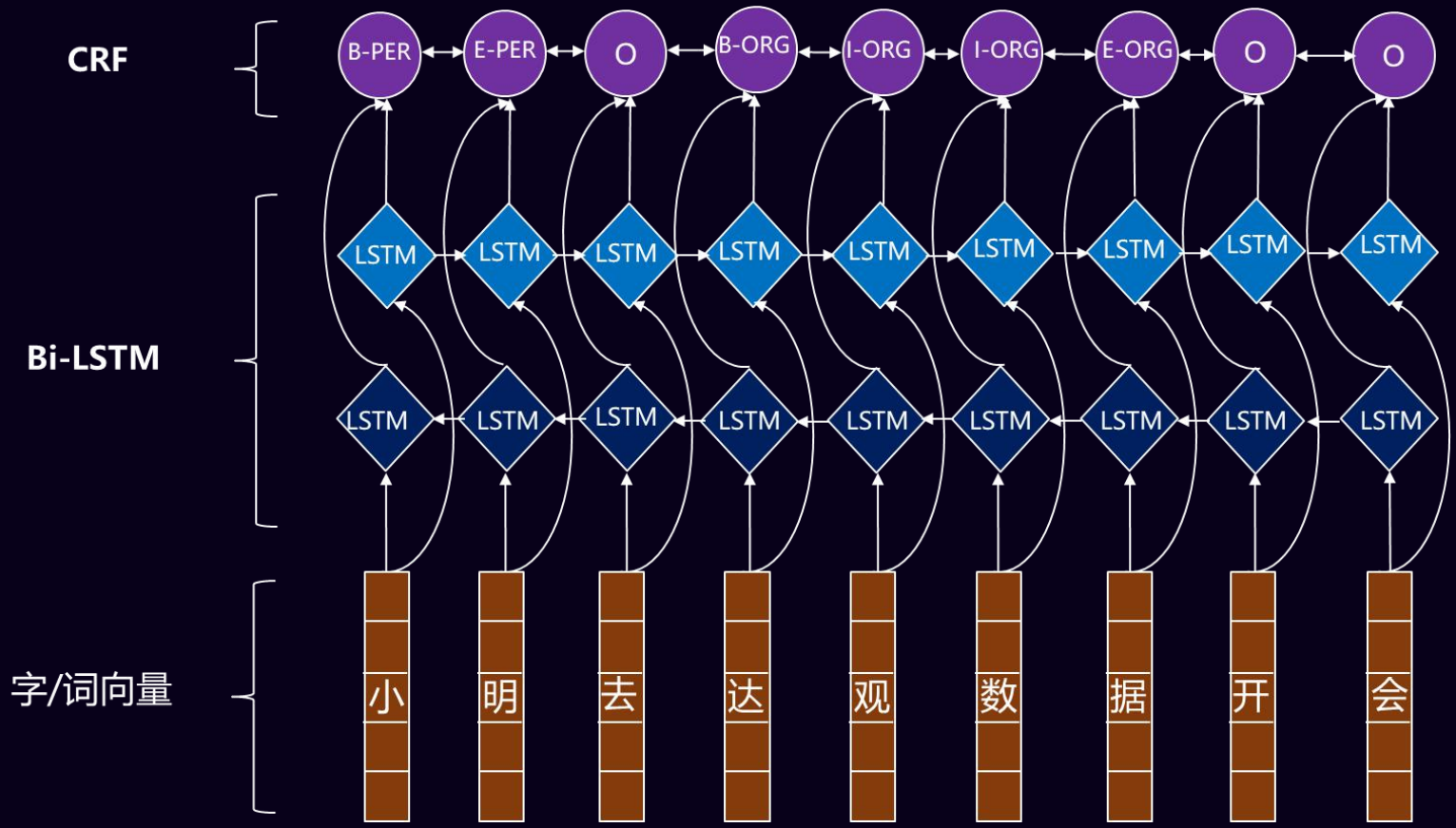
我来说两句

开源的本体编辑工具：Protégé

- 斯坦福大学医学院生物信息研究中心基于Java语言开发的本体编辑器
- 屏蔽了具体的本体描述语言，用户只需在概念层次上进行领域本体模型的构建
- 基于RDF(S),OWL等语义规范
- 对中文支持不好



知识抽取：实体抽取（NER）



- Bi-LSTM双向网络分别从前往后和从后往前进行序列信号的记忆和传递是常见做法

  华为发布了新一代的麒麟处理X
  X鲜和美国签订了新一轮的谅解备忘录

- CRF等经典方法结果可控性好，在序列标注时，在顶层用CRF对Bi-LSTM的结果进行二次操作可得到更好的结果

- 信号输入层，对中文进行embedding能起到非常好的效果

- 对英文先进行卷积CNN操作往往能抽取出单词的前后缀等信息，对提升效果有帮助

知识抽取：关系抽取技术

达观数据
DATA GRAND

- 有监督的学习方法
- 半监督的学习方法
- 无监督的学习方法

- 有监督学习法因为能够抽取并有效利用特征，在获得高准确率和高召回率方面更有优势，是目前业界应用最广泛的一类方法

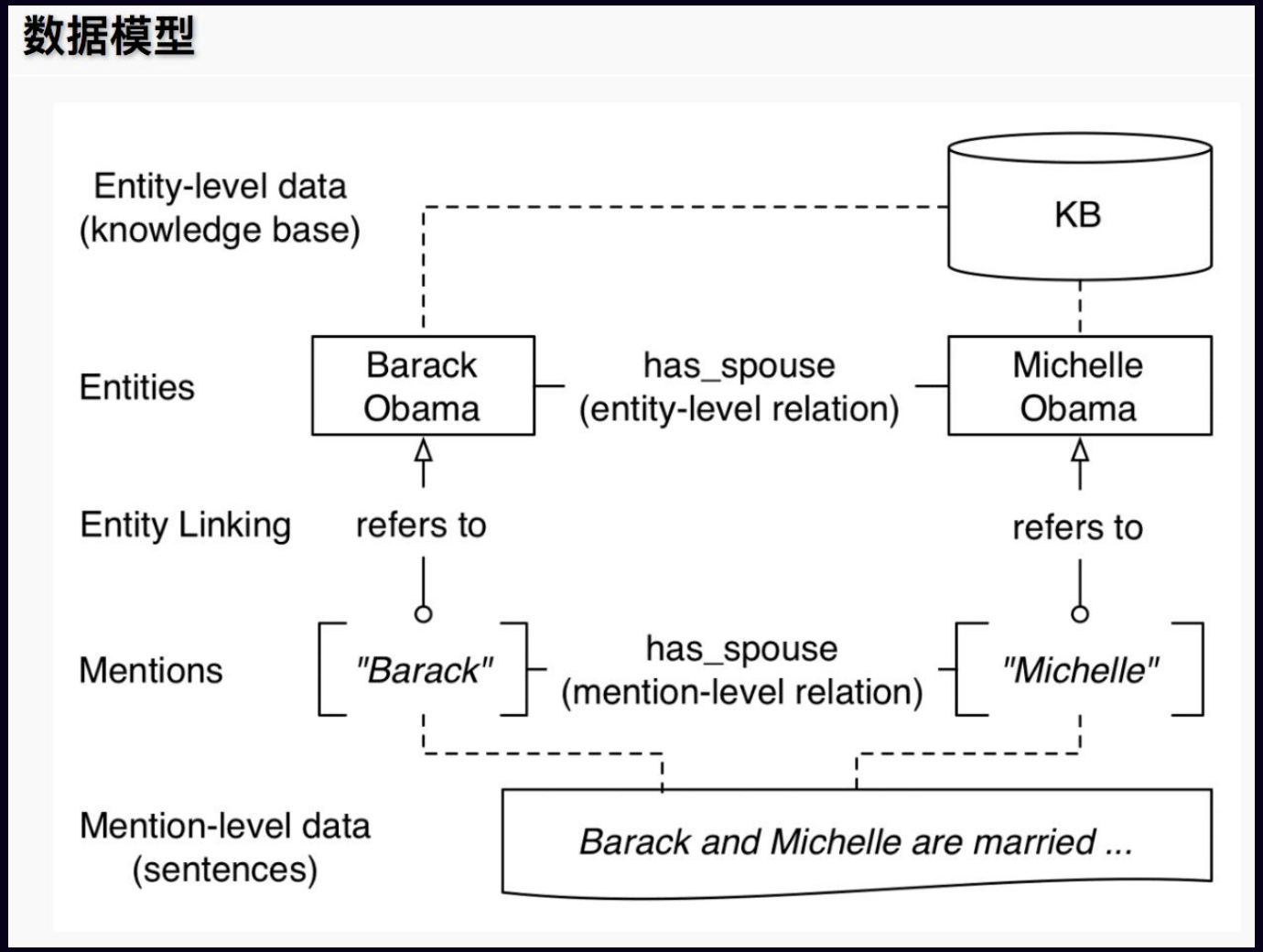
- 远程监督的学习方法

Deepdive：知识抽取框架

达观数据
DATA GRAND

- Deepdive是由斯坦福大学InfoLab实验室开发的一个开源知识抽取系统。它通过弱监督学习，从非结构化的文本中抽取结构化的关系数据，可以判断两个实体间是否存在指定关系。具有较强的灵活性，可以自己训练模型。
- DeepDive要求开发者思考特征而不是算法
- 可以通过使用已有的领域知识指导推理，接受用户反馈，提高预测的质量
- 使用Distant supervision技术，只需少量/甚至不需要训练数据

Deepdive：知识抽取框架

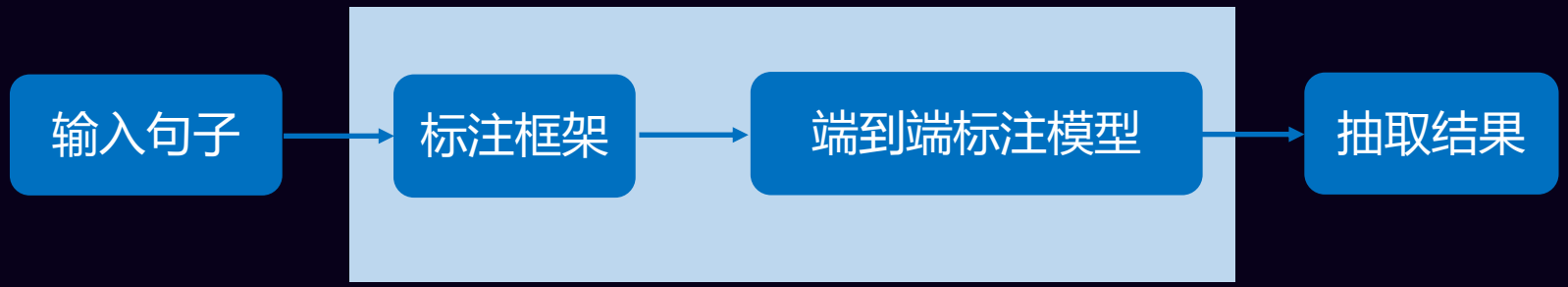


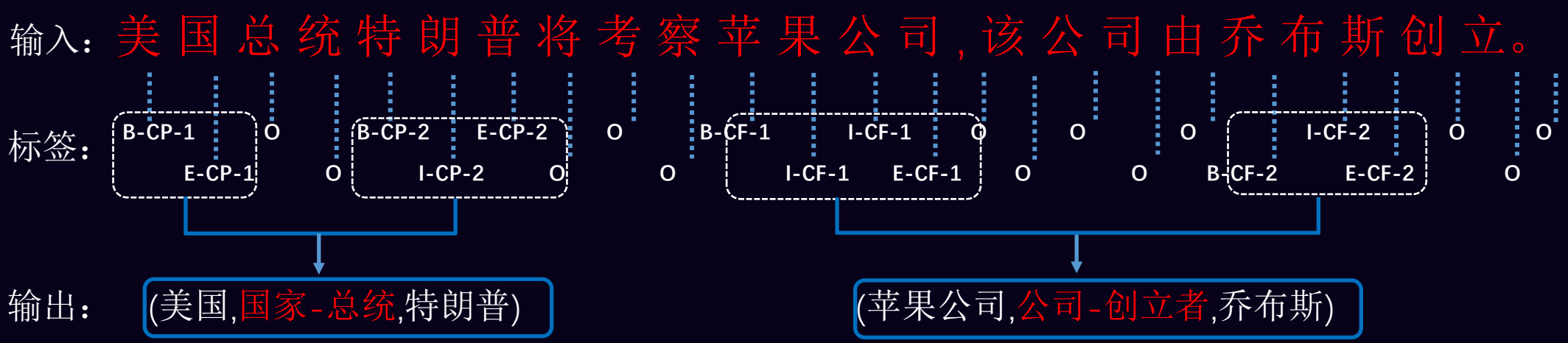
- 实体：现实中存在的事物，如奥巴马
- Mention：对实体的一个引用，如“奥巴马”三个字
- 实体级关系：实体间的关系
- Mention级关系：Mention间的关系
- 实体级数据：实体级关系的集合，如Freebase（知识库）中的关系
- Mention级数据：包含Mention的数据，如”Barack and Michelle are married”这个句子
- 实体耦合：Mention与实体的映射

知识图谱关系抽取：基于深度学习端到端的联合标注

达观数据
DATA GRAND

- 将抽取问题转换成标注任务，训练一个端到端标注模型来抽取关系
- 根据标签序列，将同样关系类型的实体合并成一个三元组作为最后的结果





知识融合：实体链接和实体合并

达观数据
DATA GRAND

- 实体对齐：旨在发现具有不同标识实体但却代表真实世界中同一对象的那些实体，并将这些实体归并为一个具有全局唯一标识的实体对象添加到知识图谱中。

- 实体对齐主要通过计算实例之间相似度：具有相同描述的实体可能代表同一实体（字符相似）；具有相同属性-值的实体可能代表相同对象（属性相似）；具有相同邻居的实体可能指向同一个对象

- 充分考虑数据源的可靠性以及不同信息在各个数据源中出现的频度等因素来决定最终选用哪个类别或哪个属性值。

- 利用来自如LOD (linked open data)中已有的人工对齐标注数据（使用owl:sameAs关联两个实体）可以作为训练数据学习发现更多相同的实体对。

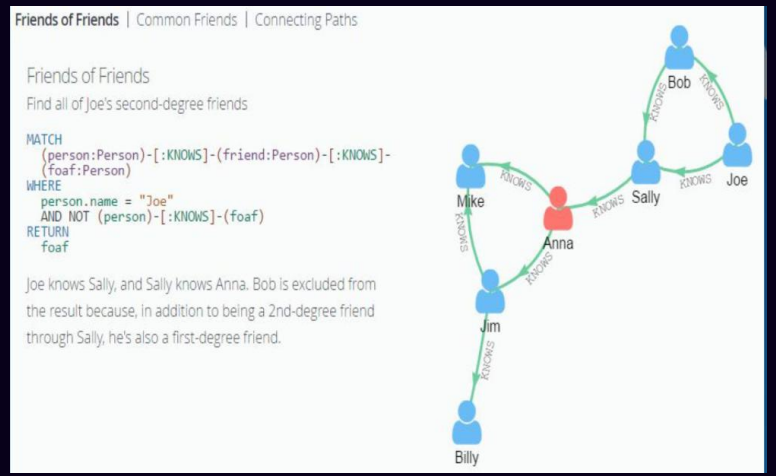
- 无论何种自动化方法都无法保证100%的准确率，这些方法的产出结果将作为候选供人工进一步审核和过滤。

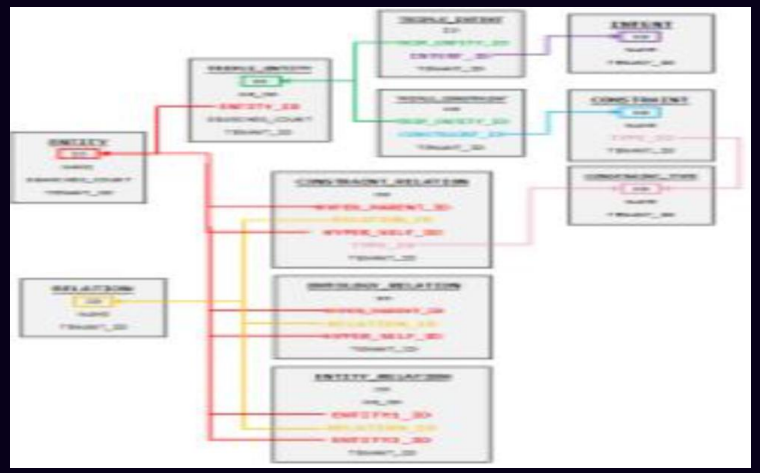
知识存储：数据库选择

达观数据
DATA GRAND

数据库层面的选择有：图数据库、NoSQL数据库、关系数据库

- 若KG结构复杂，且关系复杂，连接多，建议使用图数据库，如Neo4J等
- 若KG侧重节点知识，关系简单，连接少，可以使用传统关系数据库
- 若考虑KG的性能、扩展性和分布式等，可以使用NoSQL数据库
- 根据实际情况，也可以多种数据库融合使用（ES和Neo4J）





知识存储：数据库选择

达观数据
DATA GRAND

	TiTan	Graph Engine	Neo4J
是否开源	是	是	是
License	Apache License 2.0	MIT	GPL（开源）、AGPL（商业）
平台	Linux	Windows	Windows/Linux/Mac OS
数据量级	千亿	百亿	百亿
查询语言	Gremlin	LINQ	Cypher
API	Java	C#	Java/Python/Ruby/JS/Go/Php/.Net/C++/Spring等
Java版本	1.8以上	不支持	1.8以上
存储后端	Cassandra/Hbase/Berkeley DB	RAM	嵌入式、基于磁盘的专有文件系统
分布式	支持	支持	支持，但较弱

知识推理：基于符号的推理

达观数据
DATA GRAND

RDF idea

- Use (directed) graphs as data model

- "Resource Description Framework"

RDFS:Class and Instance

- Classes: sets of instance
- Example： 人工智能公司
- Classes can have hierarchy
  - Example: 人工智能公司是高科技公司

人工智能公司 subclass 高科技公司

RDFS:Reasoning

- From

Google RDF:type 人工智能公司

and

人工智能公司 subclass 高科技公司

we can infer

Google RDF:type 高科技公司

知识推理：基于OWL本体的推理

达观数据
DATA GRAND

OWL

- Ontology Web Language
- Has description logics as its logical underpinning
- W3C standard ontology language
- Expressive logical language
  - Negation: Car ⊑ ¬Train （Disjontness(Car,Train)）
  - Existential restriction:

Heart is a muscular organ that is part of the circulatory system

→ Heart⊑MuscularOrgan⊓∃part-of.CirculatorySystem

分类的例子

苹果由富达和黑石投资。

$Apple \sqsubseteq \exists beInvestedBy.(Fidelity \sqcap BlackStone)$

借助富达融资的公司都是创新企业。

$\exists beFundedBy.Fidelity \sqsubseteq InnovativeCcompanies$

借助黑石融资的公司都是创新企业。

$\exists beFundedBy.BlackStone \sqsubseteq InnovativeCcompanies$

$beInvestedBy \sqsubseteq beFundedBy$ 投资即是帮助融资。

$Apple \sqsubseteq \exists beInvestedBy.Fidelity$ 苹果由富达投资。

$Apple \sqsubseteq \exists beFundedBy.Fidelity$ 苹果由黑石投资。

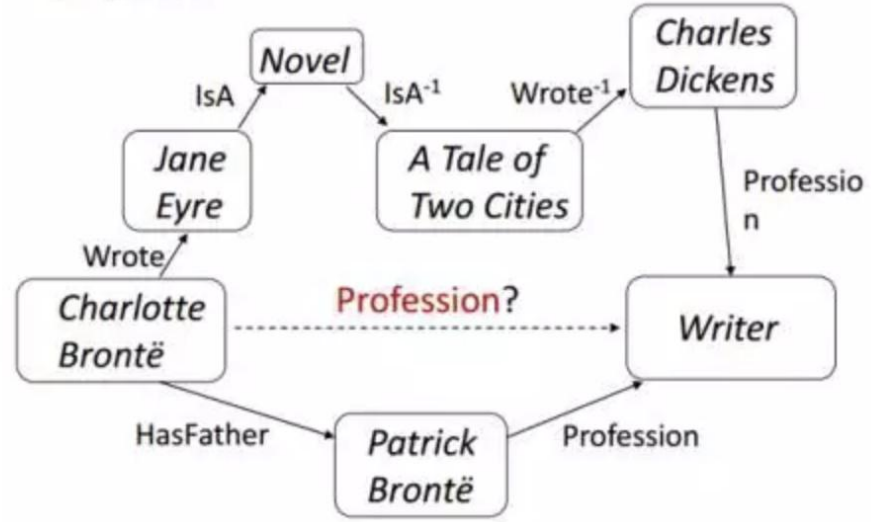
$Apple \sqsubseteq InnovativeCompanies$ 苹果是创新企业。

知识推理：基于图的方法（PRA算法）

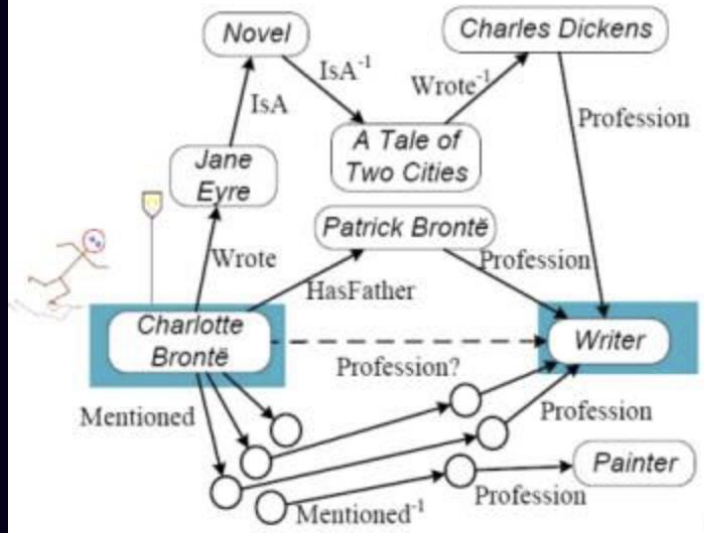
达观数据
DATA GRAND

Graph-based method

- 基本思想
  - 将连接两个实体的路径作为特征来预测其之间可能存在的关系



Path Ranking Algorithm (PRA)



Lao et.al.

G=(N,E, R)

- N: nodes (instances or concepts)
- E: edges
- R: edge types

Note: r^{-1}: reverse of edge type r

Path type π: $<r_1, r_2, \ldots, r_n>$

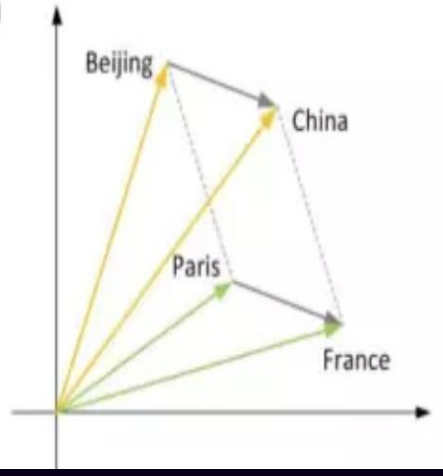
e.g. <HasFather, Profession>

知识推理：基于分布式知识语义表示的方法（Trans系列模型）

达观数据
DATA GRAND

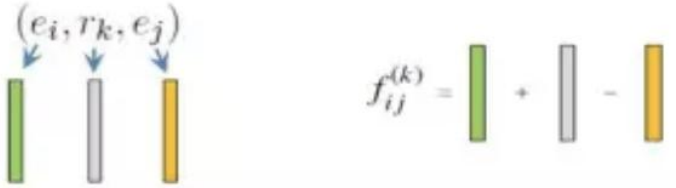
TransE Model

- Motivation
  - China – Beijing = France – Paris = <capital-of>
  - Beijing + <capital-of> = China
  - Paris + <capital-of> = France



TransE Model

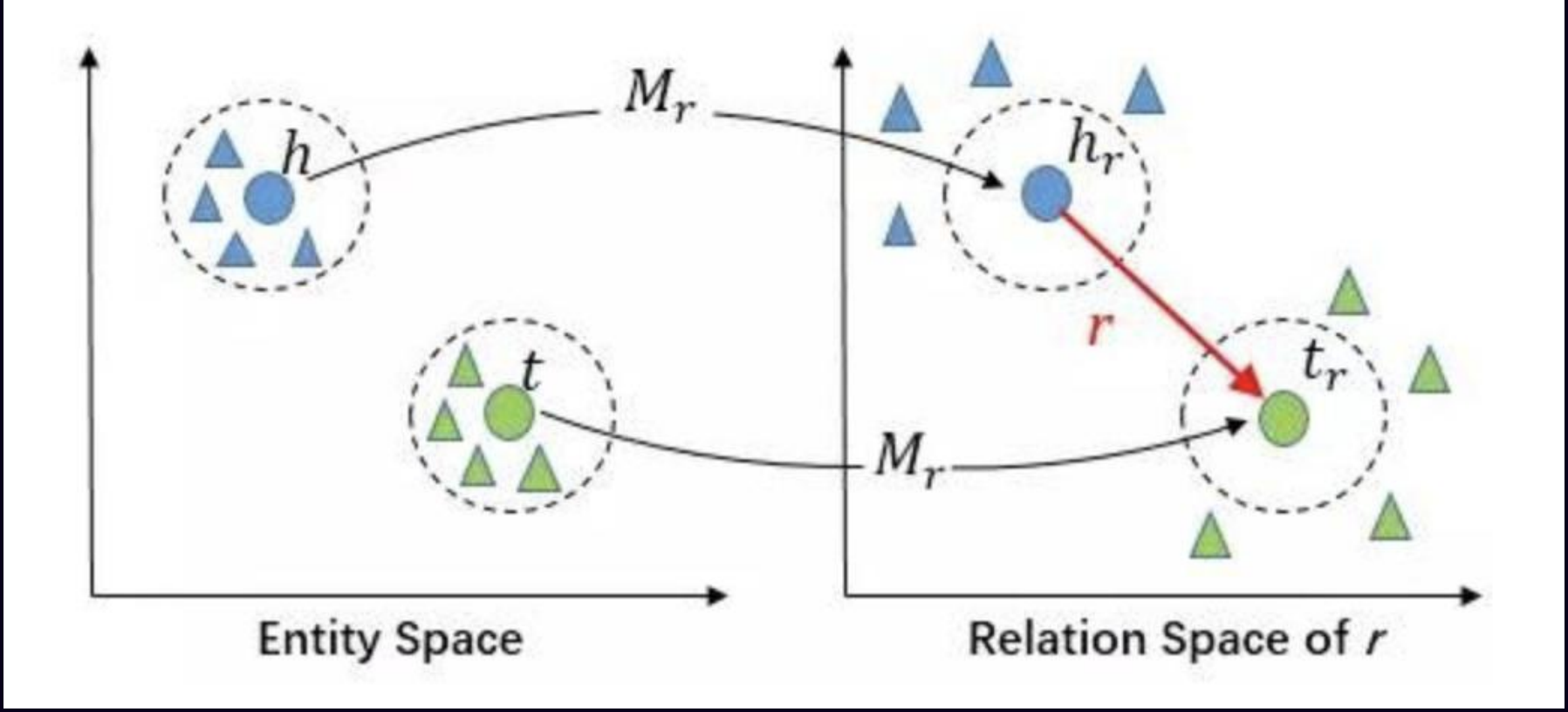
- Entity/Relation representation
  - Entities as vectors + relations as vectors
- Scoring function definition
  - Distance function: $f(e_i, r_k, e_j) = \|\mathbf{e}_i + \mathbf{r}_k - \mathbf{e}_j\|_1$



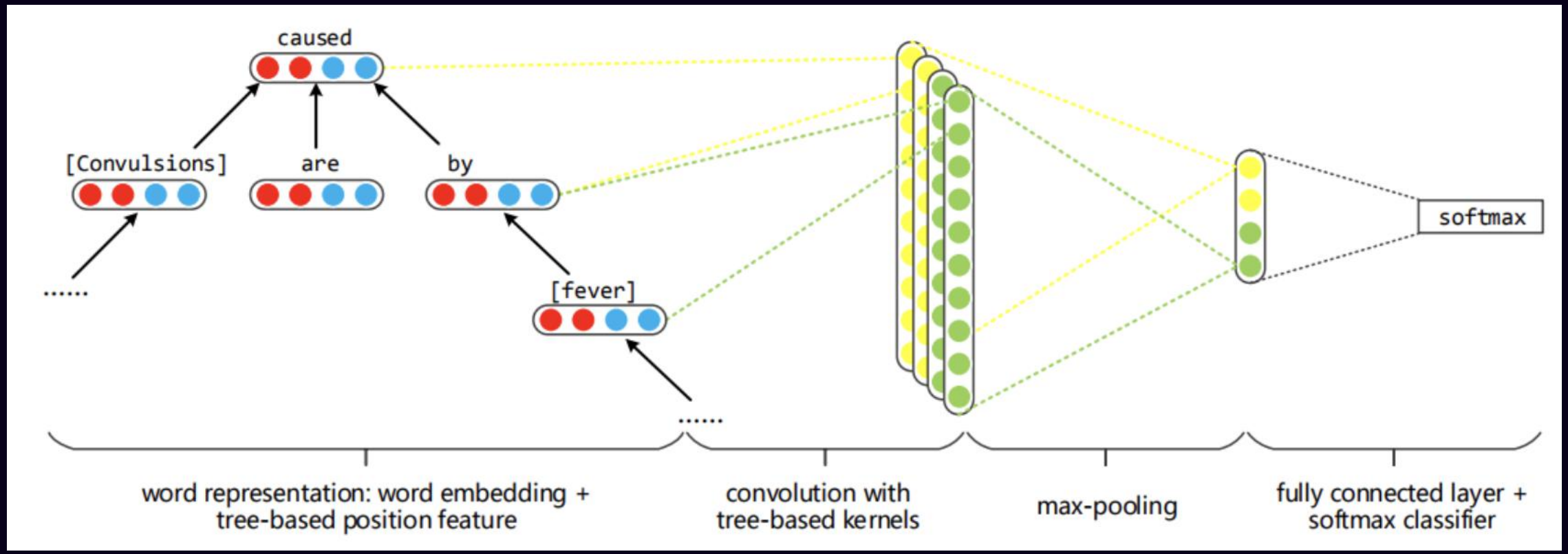
- Parameter estimation
  - Pairwise ranking loss: $\min_{\{\mathbf{e}_i\},\{\mathbf{r}_k\}} \sum_{l^+ \in O} \sum_{l^- \in N_{l^+}} \left[\gamma + f(e_i, r_k, e_j) - f(e_i', r_k, e_j')\right]_+$

将实体和关系映射到一个低维的embedding空间中，基于知识的语义表达进行推理建模

知识推理：TransR 模型



知识推理：基于深度学习的推理



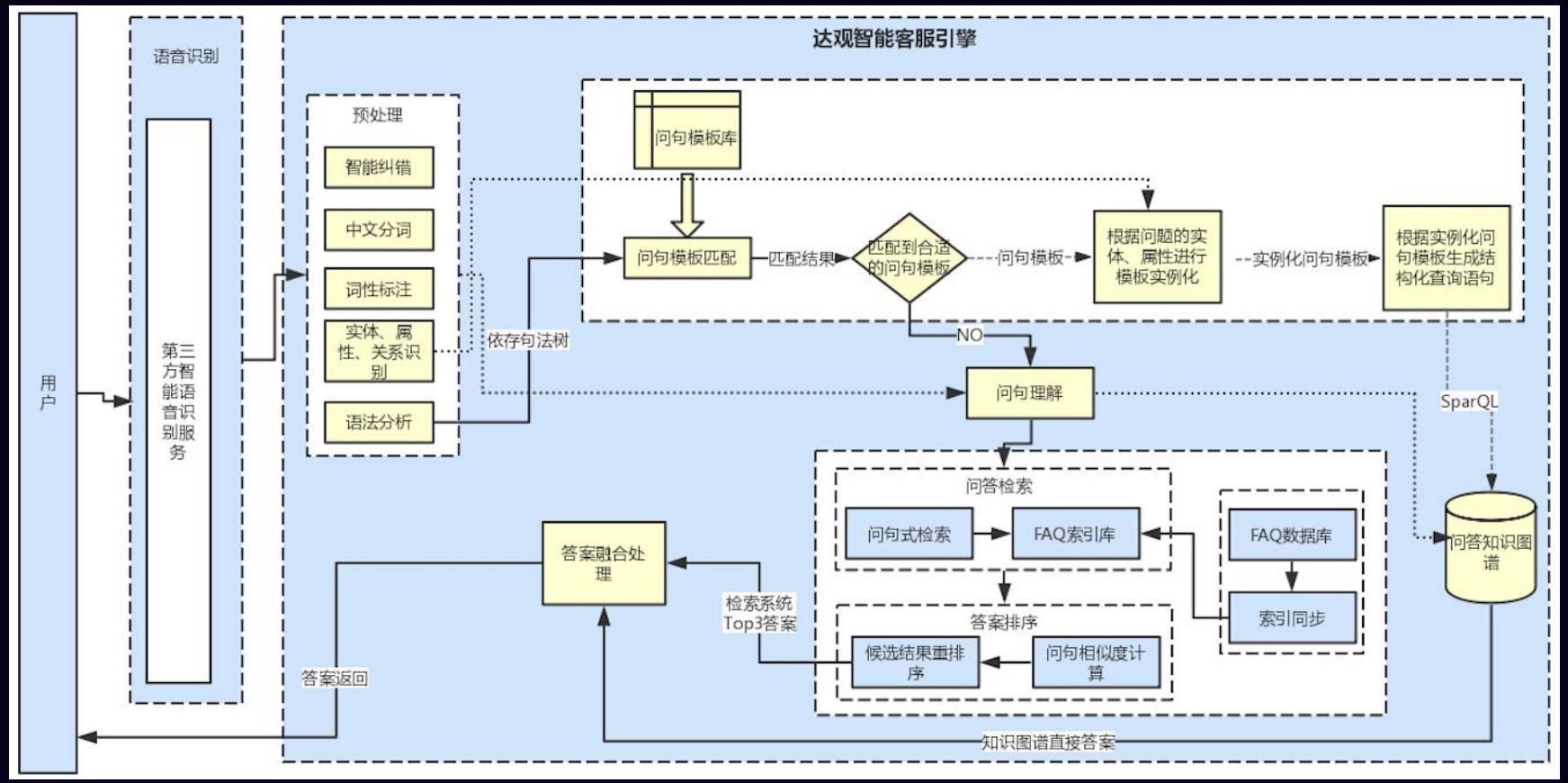
达观数据
DATA GRAND

4

THE THIRD

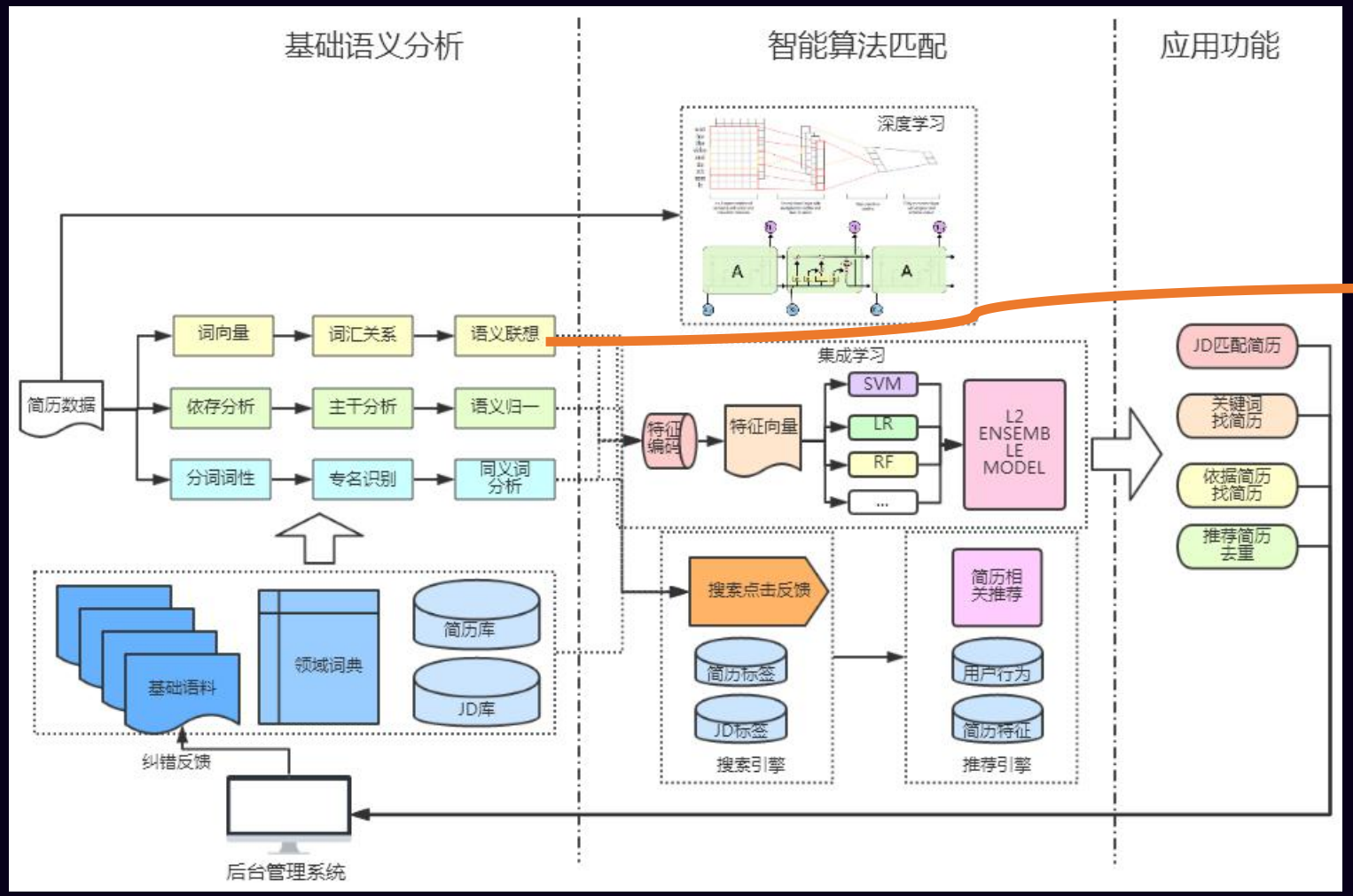
达观经验与案例

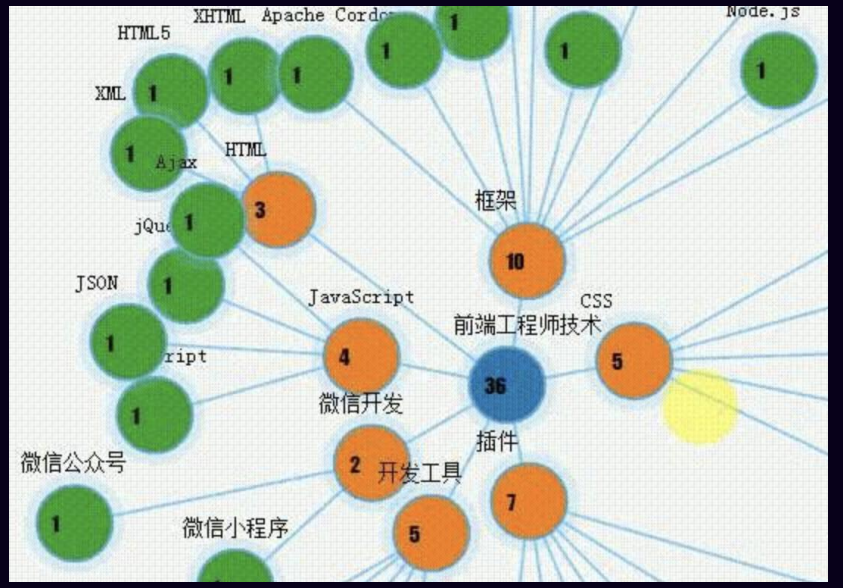
知识图谱在达观问答系统中的应用



知识图谱在HR人岗精准匹配场景中的应用

达观数据
DATA GRAND





领域知识图谱的建设的几点经验

达观数据
DATA GRAND

- 界定好范围，明确的场景和问题定义：不能为了知识图谱而知识图谱
- 知识定义比较关键：根据问题场景进行相关的领域模式schema定义，定义出领域概念层次结构、概念之间的关系类型定义
- 数据是基础，根据问题和场景梳理领域相关数据，结构化、半结构化、无结构化行业语料、百科相关数据、行业词典、专家规则
- 不要重复造轮子，合理利用百科的数据和开放知识图谱的数据
- 必须要有验证与反馈机制，确保知识的精确性
- 是一个持续迭代的系统工程，人机交互，持续运营，不断丰富完善

中文开放知识图谱

达观数据
DATA GRAND

现有的公开的知识图谱数据

达观数据
DATA GRAND

- 英文
- Freebase 6800万实体，10亿关系
- DBpedia 364万实体（来自wikipedia）
- Yago （德国，6.4亿实体）
- Wolframalpha 10亿实体

- 中文
- 复旦知识工场 1700w实体，27万概念，3300万 is-a关系
- WikiData中文 2700w中文条目
- Zhishi.me包含常见的百科词条

达观数据
DATA GRAND

达观数据 技术服务伴随客户共同成长

全球领先的智能文本处理专家