

《知识图谱: 概念与技术》

知识图谱概述

参考教材

- 知识图谱：概念与技术
- 作者: 肖仰华 等
- 出版社: 电子工业出版社
- 考核形式:
小组项目 (1-3人一组)
期末考试
- 网站: <https://bnu05pp.github.io/KGGraduateCourse/2022.html>



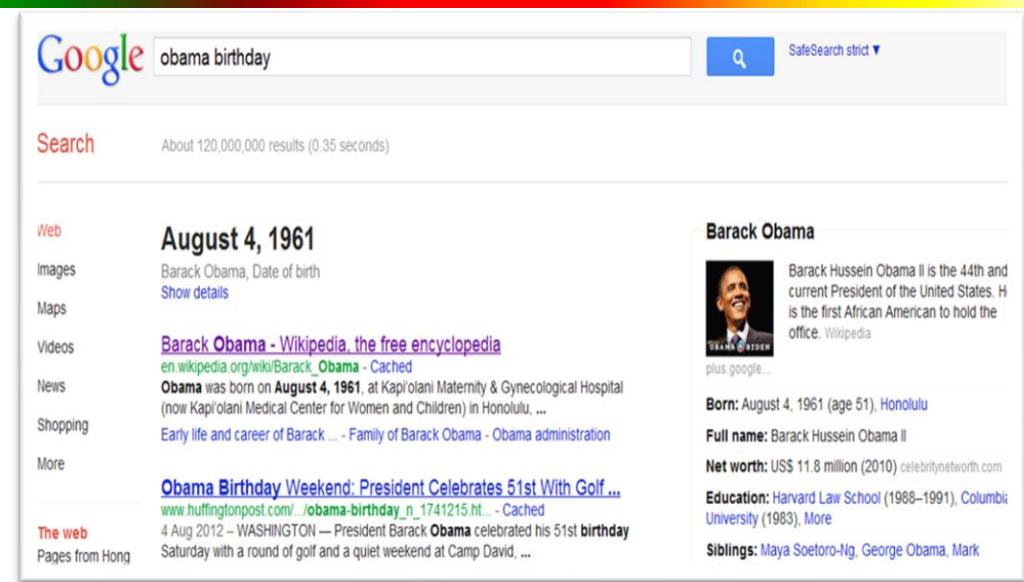
本章大纲

- 基本概念
- 历史沿革
- 研究意义
- 应用价值
- 知识图谱分类

知识图谱的基本概念

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



<https://www.fastcompany.com/1671024/google-buys-metaweb-one-company-could-revolutionize-google-search>

知识图谱的狭义概念

- 知识图谱(Knowledge Graph)本质上是一种**大规模语义网络**(semantic network)
- 富含**实体(entity)**、**概念(concepts)**及其之间的各种语义关系(semantic relationships)

作为一种**语义网络**，是大数据时代**知识表示**的重要方式之一

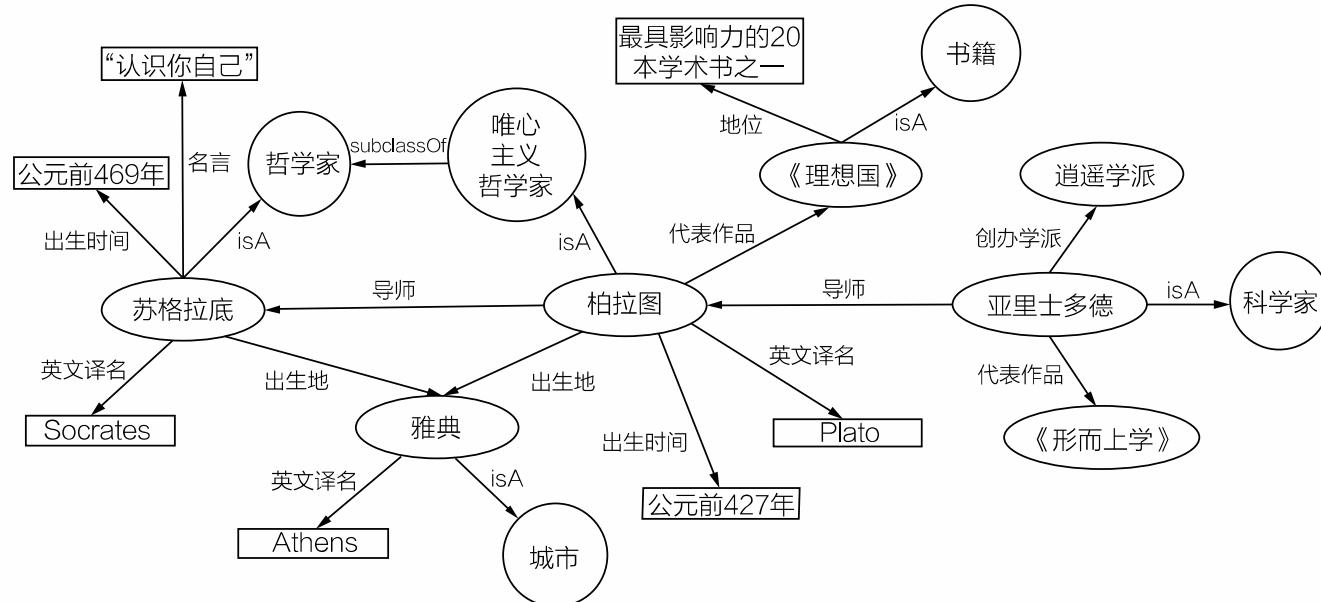


图 1-1 关于古希腊三大哲学家的知识图谱片段

语义网络

- 语义网络是一种以图形化的(Graphic)形式通过点和边表达知识的方式[1]，其基本组成元素是点和边

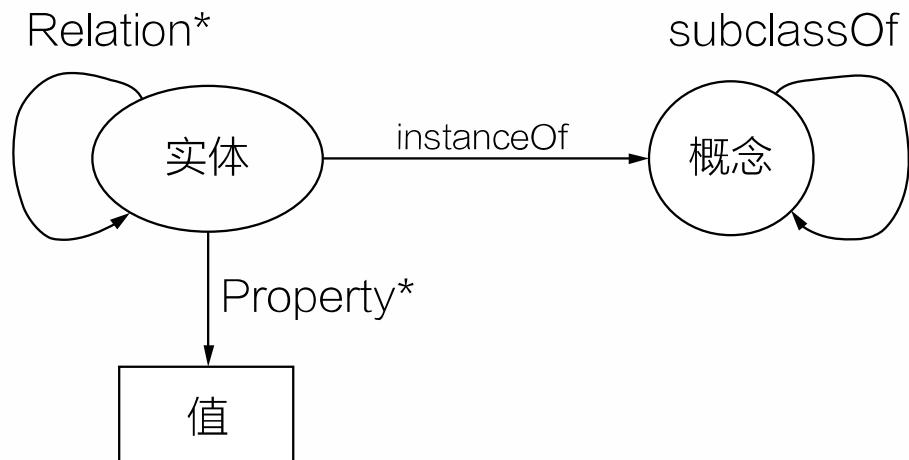


图 1-2 语义网络的组成（图中星号表示可能存在多个不同的属性或者关系）

KG组成- Node-Entity

- Entity/Objects/Instances
 - Wikipedia: An **entity** is something that exists as itself, as a subject or as an object, actually or potentially, concretely or abstractly, physically or not.
 - 黑格尔《小逻辑》：能够独立存在的，作为一切属性的基础和万物本原的东西



KG组成- Node-Concept

- Concept
 - In [metaphysics](#), and especially [ontology](#), a concept is a fundamental [category of existence](#).
 - (mental) representations of categories
- Category
 - Groups of entities which have something in common;
- Type/class
 - WIKITIONARY: A grouping based on shared characteristics; a [class](#).

CATEGORIZATION:

- 1、the process of formation of categories;
- 2、the process of identifying X as a member of a particular category Y;

[owl:Thing](#)
◦ [Activity](#) (edit)
▪ [Game](#) (edit)
▪ [BoardGame](#) (edit)
▪ [CardGame](#) (edit)
▪ [Sales](#) (edit)
▪ [Sport](#) (edit)
▪ [Athletics](#) (edit)
▪ [Boxing](#) (edit)

DBpedia Types

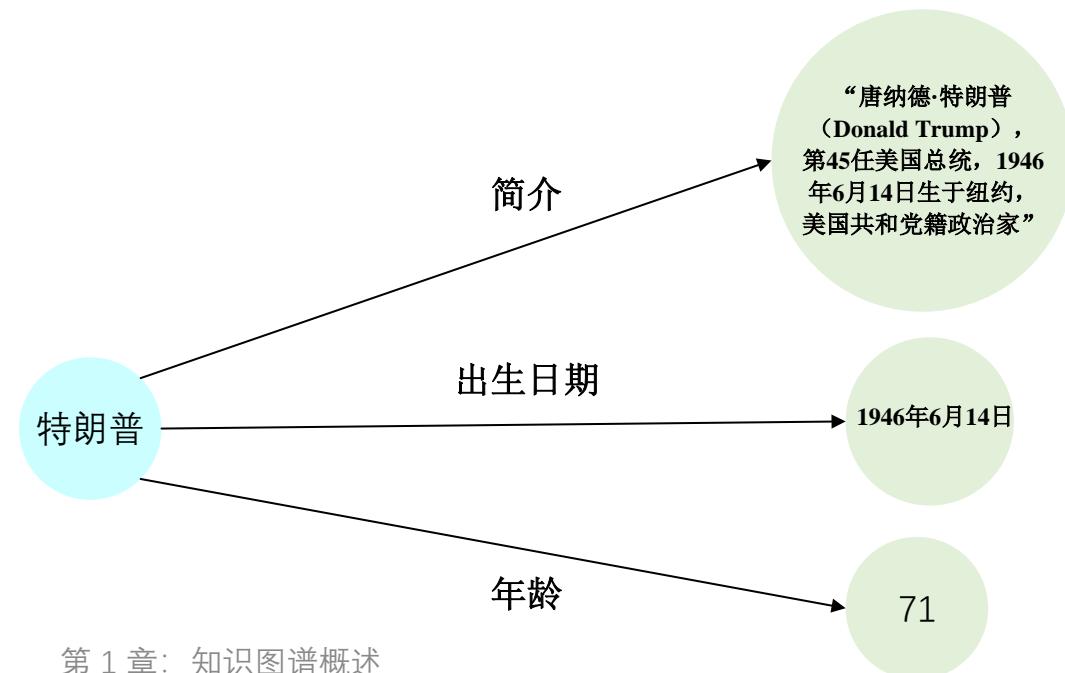
microsoft

- company
- software company (Basic-level concept)
- largest OS vendor

Probase
Categories

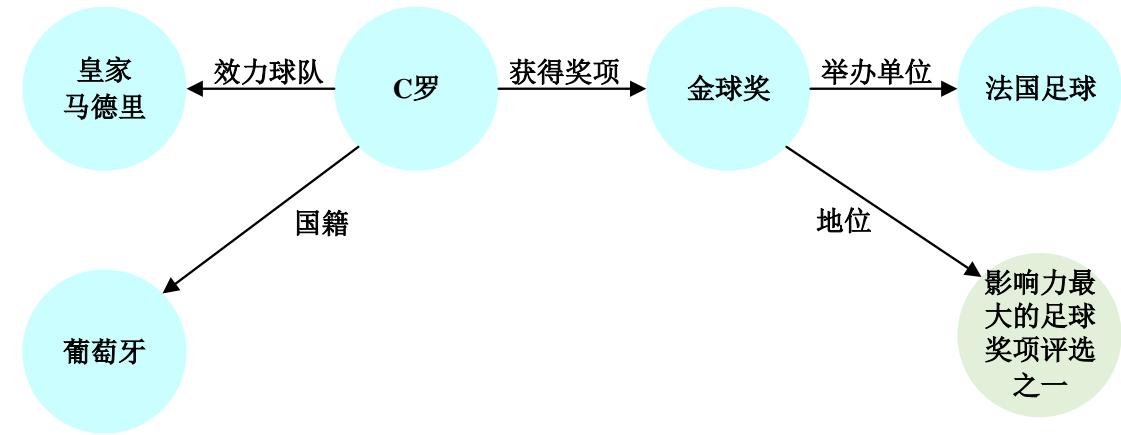
KG组成- Node-Value

- Date
 - 特朗普 出生日期 1946年6月14日
- String
 - 特朗普 简介 “唐纳德·特朗普 (Donald Trump) , 第45任美国总统, 1946年6月14日生于纽约, 美国共和党籍政治家”
- Numeric
 - 特朗普 年龄 71



KG组成- 边

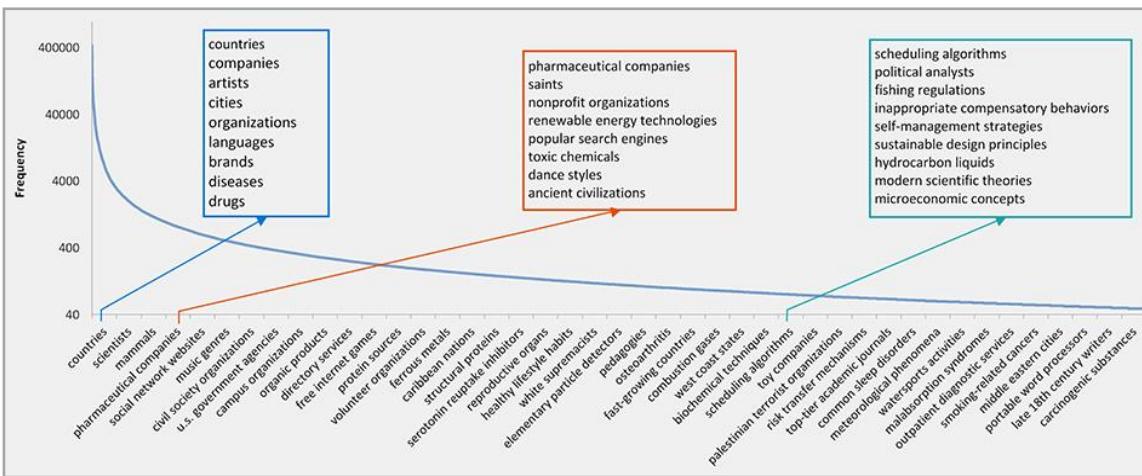
- Relation
 - 侧重实体(individual)之间的关系
 - Examples:
 - Sitting-On: An apple sitting on a table
 - Taller-than: [Washington Monument](#) is taller than the [White House](#)
- Property/Attribute/Quality
 - A characteristic/quality that describes an object
 - Examples:
 - size, color, weight, composition, and so forth, of an object



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

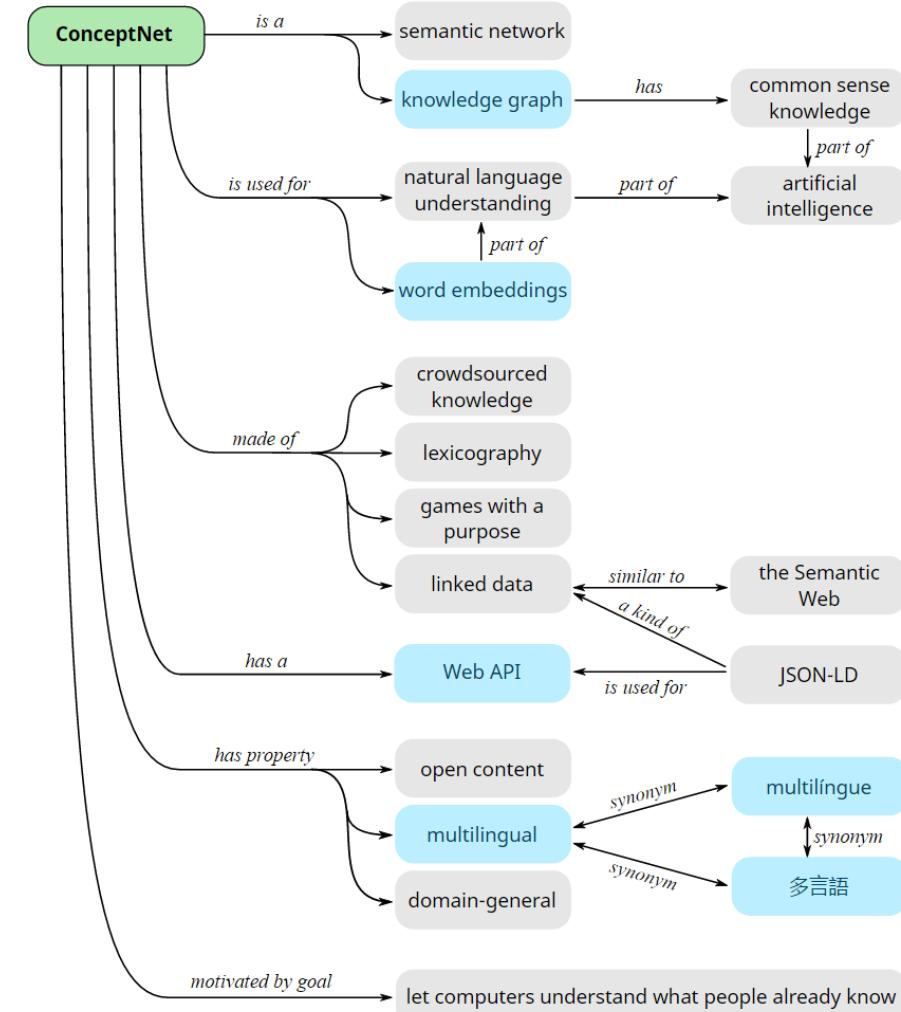


Existing Taxonomies	Number of Concepts
Freebase [5]	1,450
WordNet [13]	25,229
WikiTaxonomy [26]	111,654
YAGO [35]	352,297
DBPedia [1]	259
ResearchCyc [18]	$\approx 120,000$
KnowItAll [12]	N/A
TextRunner [2]	N/A
OMCS [31]	N/A
NELL [7]	123
Probase	2,653,872

KG优势2: semantically rich

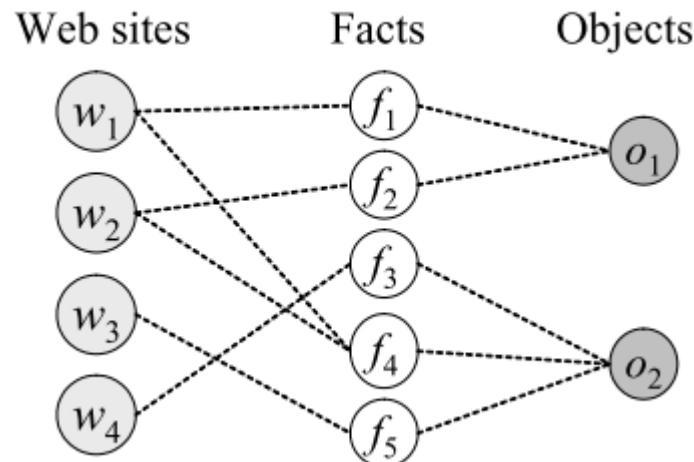
- Higher coverage over numerous semantic relationships

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

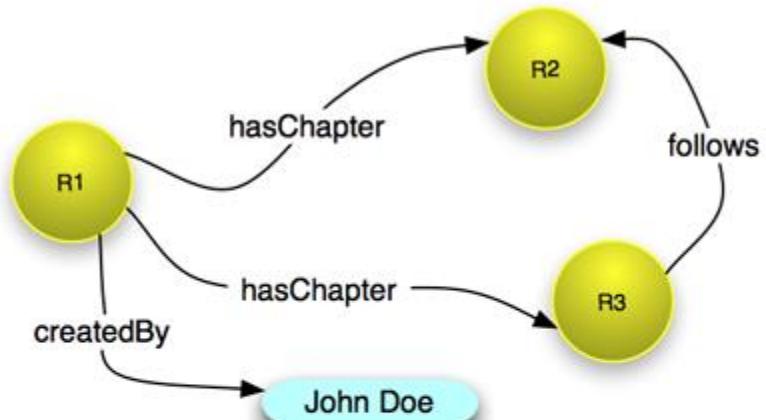


[Yin, et al. 2017]

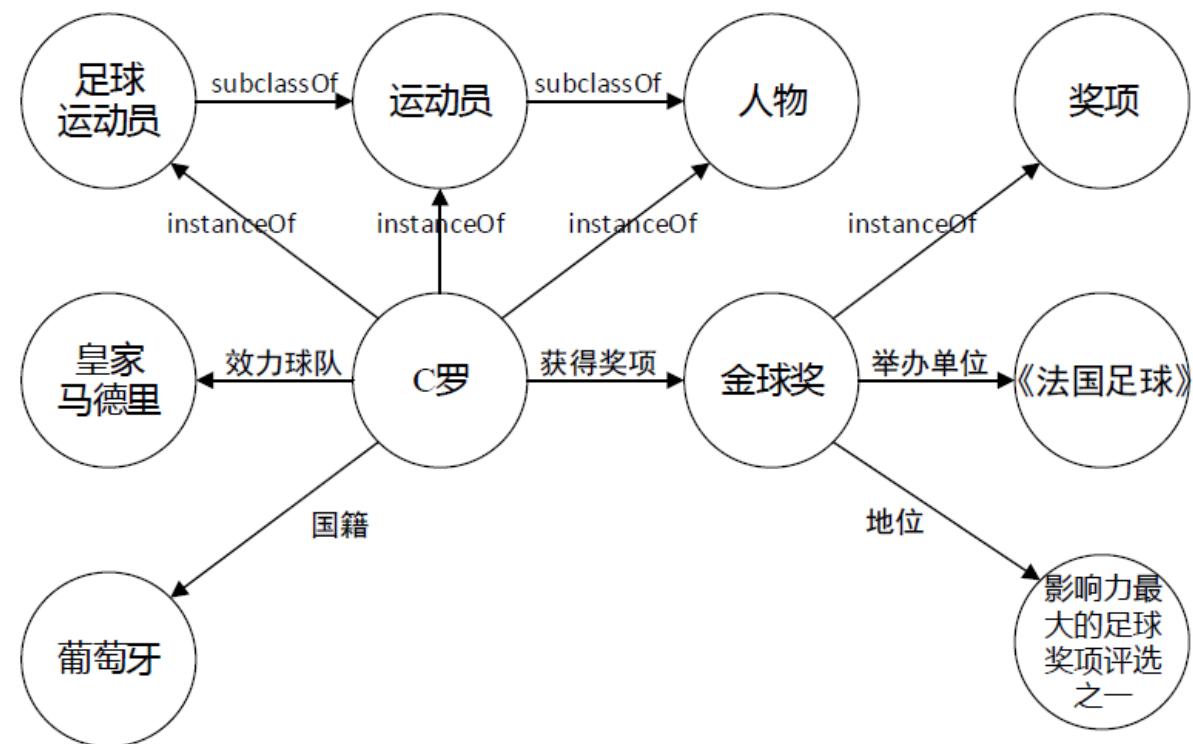


KG优势4: friendly structure

- Structured organization
 - By RDF
 - By graph



Subject	Predicate	Object
R1	hasChapter	R2
R1	hasChapter	R3
R3	follows	R2
R1	createdBy	"John Doe"



KG的不足1：高质量模式缺失



- 提升知识图谱的规模往往付出质量方面的代价
 - 可以预先定义人的“身高”取值范围为0.5m ~ 2.3m，但可能存在某个人，其身高达到2.31m
 - “妻子”作为一条关系通常只有单一取值，不可以是多值的，但是古代人未必如此，当今世界的某个偏远部落也未必如此
- 知识图谱在设计模式时通常会采取一种“经济、务实”的做法：也就是允许模式（Schema）定义不完善，甚至缺失

模式定义不完善或缺失对知识图谱中的数据语义理解以及数据质量控制提出了挑战

KG不足2：封闭世界假设不再成立

- 传统数据库与知识库的应用通常建立在封闭世界假设（CWA）基础之上。CWA 是假定数据库或知识库中不存在（或未观察到）的事实即为不成立的事实
- 大多数开放性应用不遵守这一假设。也就是说，在这些应用中缺失的事实或知识未必为假
 - 很难保证知识图谱中关于柏拉图的信息完整，很可能会缺失柏拉图父母的信息。但常识告诉我们柏拉图一定有父母。

不遵守CWA 给知识图谱上的应用带来了巨大的挑战

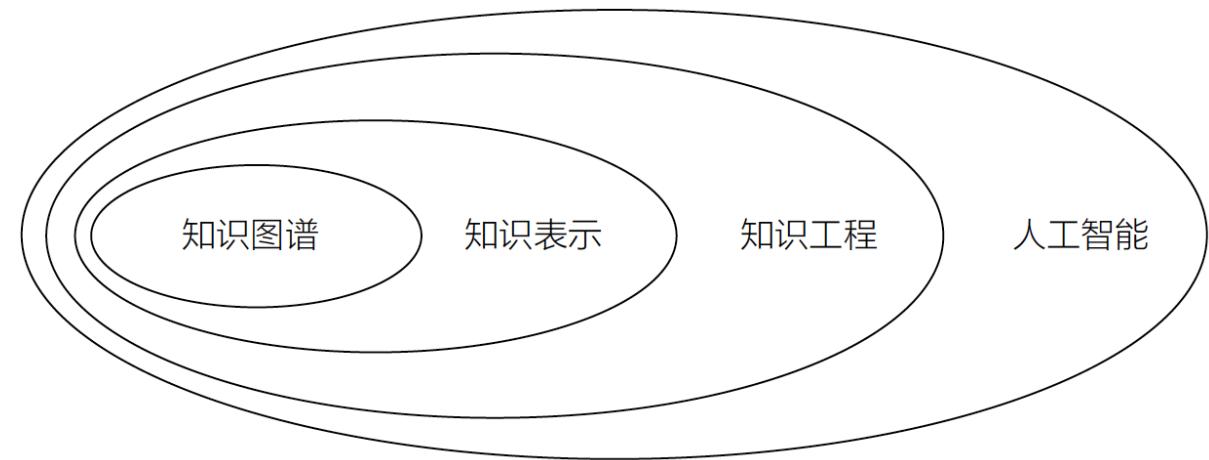
KG不足3：大规模自动化知识获取成为前提



- 传统知识工程依赖专家完成知识获取，这一方式难以实现大规模知识获取，难以满足知识图谱的规模要求
- 大规模自动化知识获取是知识图谱与传统语义网络的根本区别
- 大规模自动化知识获取的方式是多样的
 - 从文本中自动抽取
 - 基于大规模众包平台的知识标注
 - 多种方式混合

知识图谱的广义内涵

- 作为一种技术体系，是大数据时代知识工程的代表性进展
- 作为一门学科，知识图谱属于人工智能范畴
- 知识表示是发展知识工程最关键的问题之一。而知识表示的一个重要方式就是知识图谱



知识图谱的学科地位

学科地位

人工智能

知识工程

知识表示

知识图谱

AI (*Artificial Intelligence*): Think, act, humanly or rationally

"The exciting new effort to make computers think ... *machines with minds*, in the full and literal sense."

(Haugeland, 1985)

"AI ... is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

KE (*Knowledge engineering*) is an engineering discipline that involves integrating knowledge into computer systems in order to solve complex problems normally requiring a high level of human expertise

KR (*Knowledge representation*) is dedicated to representing information about the world in a form that a computer system can utilize to solve complex tasks such as diagnosing a medical condition or having a dialog in a natural language.

KG (*Knowledge graph*) is a large scale *semantic network* consisting of entities/concepts as well as the semantic relationships among them

知识图谱的历史沿革

知识工程 (KE) 的源起- Symbolism

- 符号主义的主要观点

- 认知即计算
- 知识是信息的一种形式,是构成智能的基础
- 知识表示、知识推理、知识运用是人工智能的核心



Newell

Simon

- Physical Symbol System

- A physical symbol system has the necessary and sufficient means of general intelligent action
- The mind can be viewed as a device operating on bits of information according to formal rules.

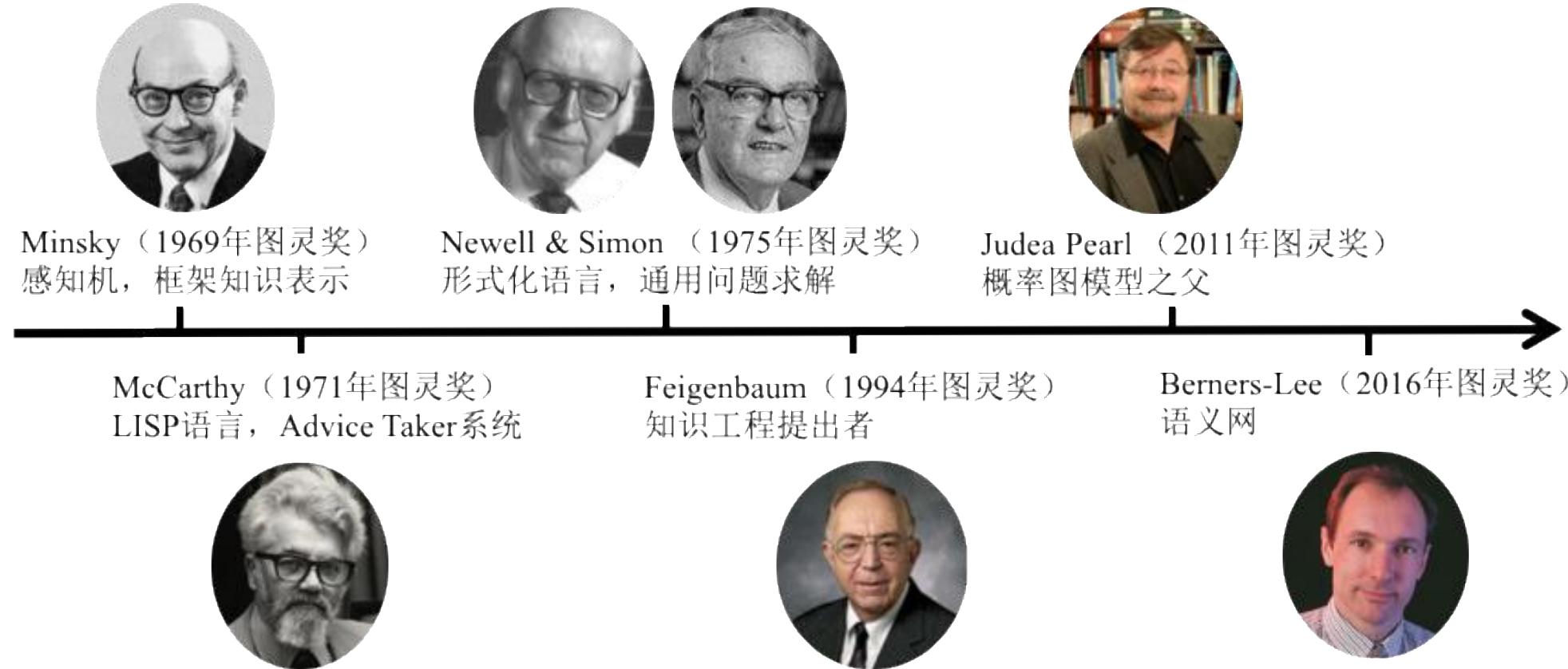
AI System=Knowledge + Reasoning

- GOFAI (“good old fashioned artificial intelligence”, proposed by John Haugeland)

- Focused on these kind of high level symbols, such as <dog> and <tail>

[Newell, Allen et al. 1976], [Dreyfus, Hubert 1979]

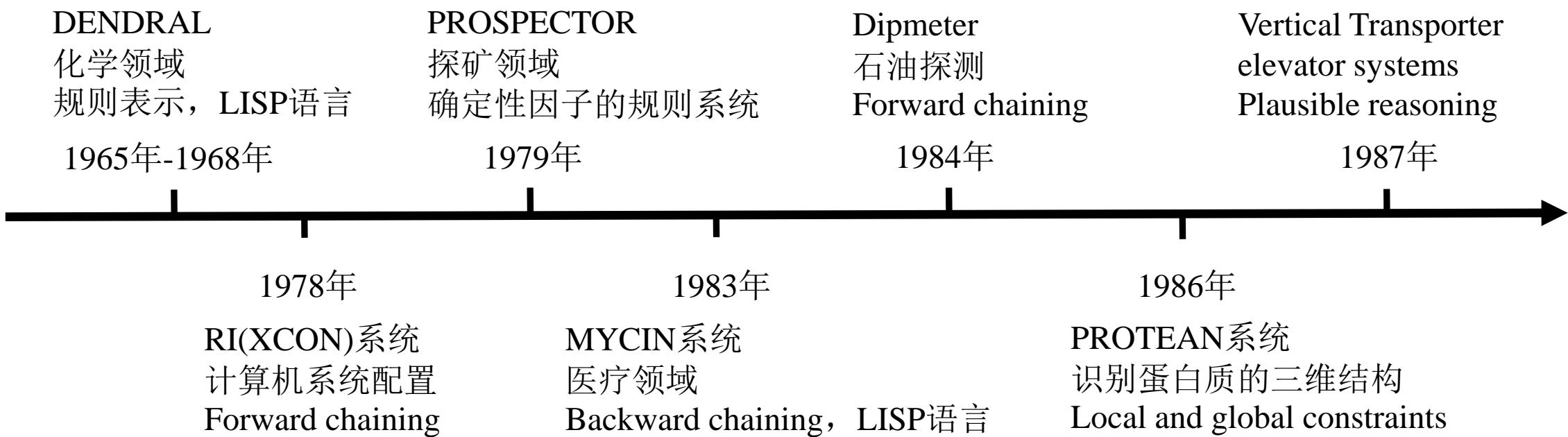
传统KE-代表性人物与成就



KE (Knowledge engineering) is an engineering discipline that involves integrating knowledge into computer systems in order to solve complex problems normally requiring a high level of human expertise. Ref Wikipedia

知识工程是以知识为处理对象，研究知识系统的知识表示、处理和应用的方法和开发工具的学科

传统KE-代表性系统



■ 传统知识工程在规则明确、边界清晰、应用封闭的应用场景取得了巨大成功

传统KE的基本特点

- 自上而下: 严重依赖专家和人的干预

- 规模有限
- 质量存疑

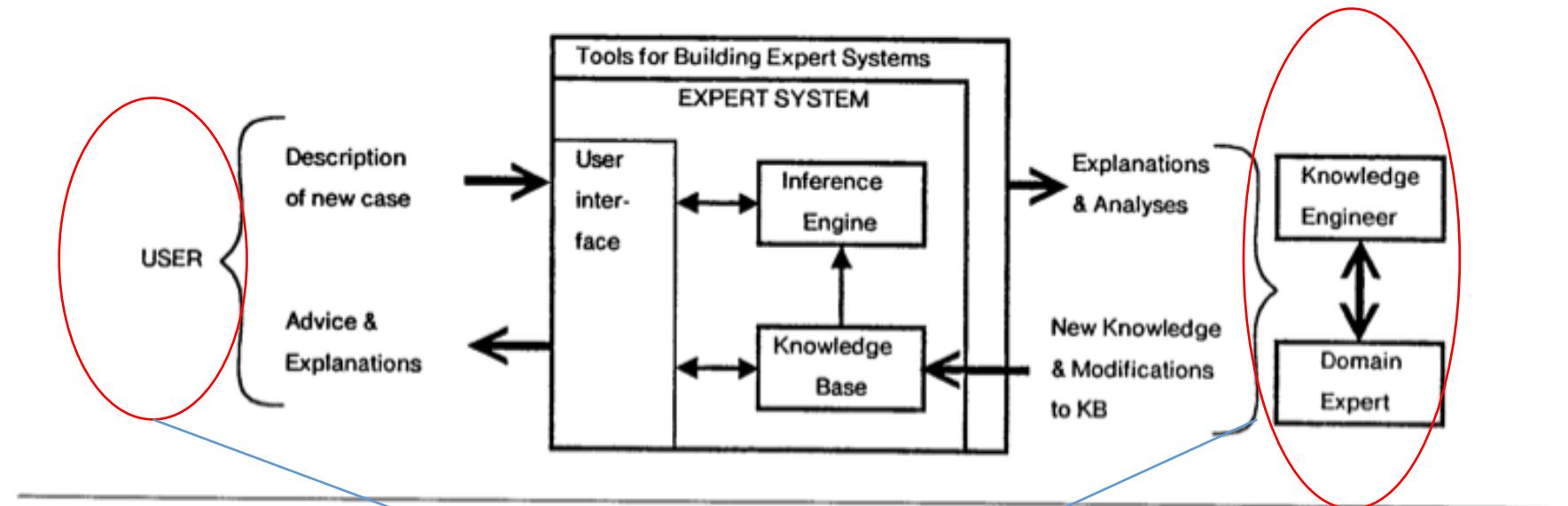


FIGURE 1-2 Interaction of a knowledge engineer and domain expert with software tools that aid in building an expert system. Arrows indicate information flow.

MYCIN专家系统中的人工参与部分

传统KE的主要挑战：知识获取困难

- 隐性知识、过程知识等难以表达
 - 如何表达做蛋炒饭的知识？
 - 老中医看病用到了哪些知识？
- 领域知识的形式化表达较为困难
- 专家知识不可避免地存在**主观性**
- 不同专家之间知识可能存在**不一致性**
- 知识表达**难以完备**，缺漏是常态

例 1：如图，在四边形 HIJK 中，M、N、O、P 分别是边 HI、IJ、KJ、HK 的中点，连接 MN、NO、OP、MP，求证四边形 MNOP 是平行四边形。

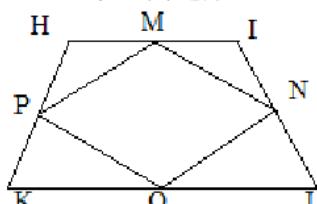


图 5-1 例 1 的图形

2022/2/25

(Point O)
(Point P)
(PointCollinearRelation P M)
(PointCollinearRelation P O)
(PointCollinearRelation O N)
(PointCollinearRelation M N) ←
(PointCollinearRelation H P K)
(PointCollinearRelation K O J)
(PointCollinearRelation I N J)
(PointCollinearRelation H M I)
结论：
(ParallelogramRelation (Quadrangle M N O P))

(Quadrangle H I J K)
(MidpointRelation M (Segment H I))
(MidpointRelation N (Segment I J))
(MidpointRelation O (Segment J K))
(MidpointRelation P (Segment H K))
初始图形信息：

(Point H)
(Point I)
(Point J)
(Point K)
(Point M)
(Point N)

基于规则系统的高中
几何自动解题过程

```
rule "ParallelogramToLineParallel"  
when  
    $pr : ParallelogramRelation(  
        $A : quadrangle.getVertexA(),  
        $B : quadrangle.getVertexB(),  
        $C : quadrangle.getVertexC(),  
        $D : quadrangle.getVertexD())  
then  
    Line AB = new Line($A, $B);  
    Line BC = new Line($B, $C);  
    Line CD = new Line($C, $D);  
    Line DA = new Line($D, $A);  
    LineParallelRelation lpr1 = new LineParallelRelation(AB, CD);  
    LineParallelRelation lpr2 = new LineParallelRelation(BC, DA);  
    Condition condition = new Condition($pr);  
    Conclusion conclusion = new Conclusion(lpr1, lpr2);  
    GEOMETRY_REASONING.buildNetwork(drools.getRule().getName(),  
    condition, conclusion);  
    GEOMETRY_REASONING.addFact(lpr1, lpr2);  
end
```

第 1 章：知识图谱概述

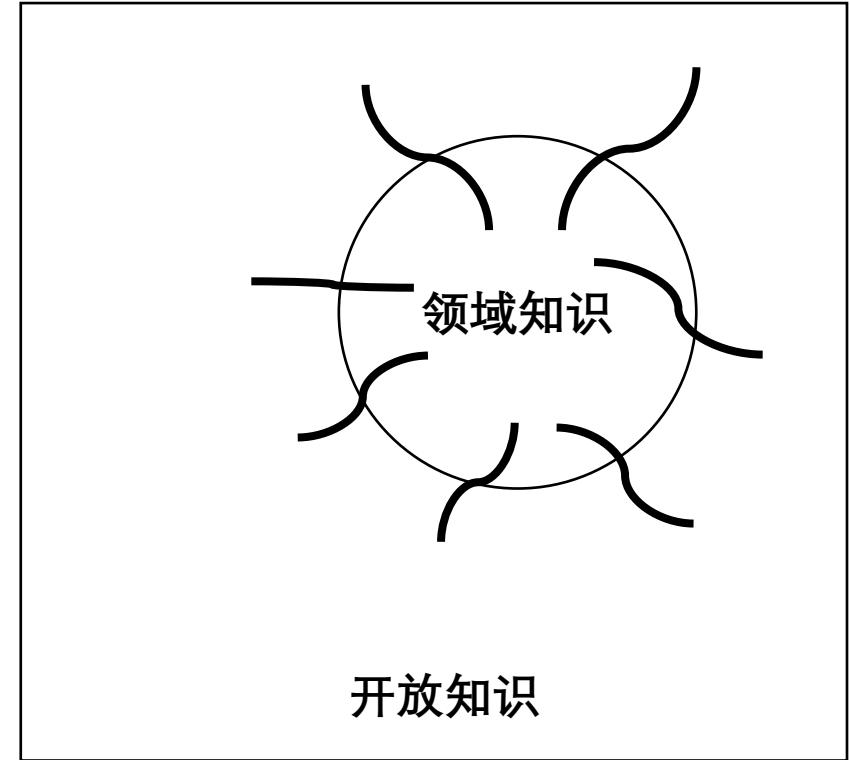
26

传统KE的主要挑战： 知识应用困难

- 应用易于超出预先设定的知识边界
- 很多应用需要常识的支撑
- 难以处理异常情况
- 难以处理不确定性推理
- 知识更新困难

Can pig fly?

Rule: if x is a bird then x can fly
How about ostrich?



行业应用中的知识需求难以封闭于预设的领域知识边界内

互联网应用催生大数据时代知识工程（BigKE）

- 大规模开放性应用
 - 永远不知道用户下一个搜索关键字是什么
 - “创造101”、“吃鸡”、“纸片人”、“蛙儿子”
- 精度要求不高
 - 搜索引擎从来不需要保证每个搜索的理解和检索都是正确的
- 应用/推理简单
 - 大部分搜索理解与回答只需要实现简单的推理
 - 简单推理：“姚明的身高是多少”
 - 复杂推理：“姚明老婆的婆婆的儿子有多高”
- 互联网时代的大规模开放性应用需要全新的知识表示，谷歌知识图谱诞生，知识工程迈入大数据时代

热搜		更多 >
< 热门搜索	世说新词	>
排名	关键词	搜索指数
1	青春同学会	2492416 ↑
2	明日之子	2361543 ↑
3	世界杯	1010255 ↑
4	糖果翻译手机	689751 ↑
5	韩庚	609600 ↑
6	陈瑶	564731 ↓
7	跨界歌王	559905 ↑
8	流星花园	521628 ↑
9	大街网	508397 ↑

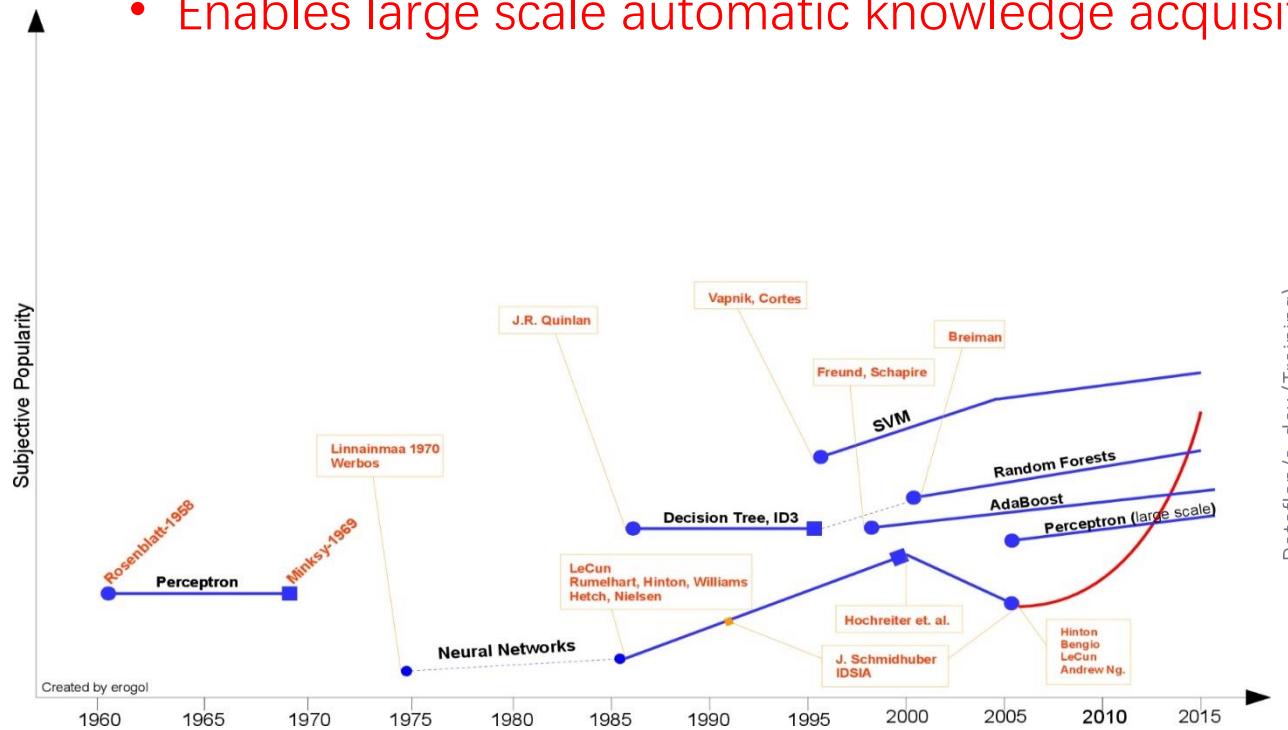
互联网上的搜索关键字具有开放性、规模巨大等特点



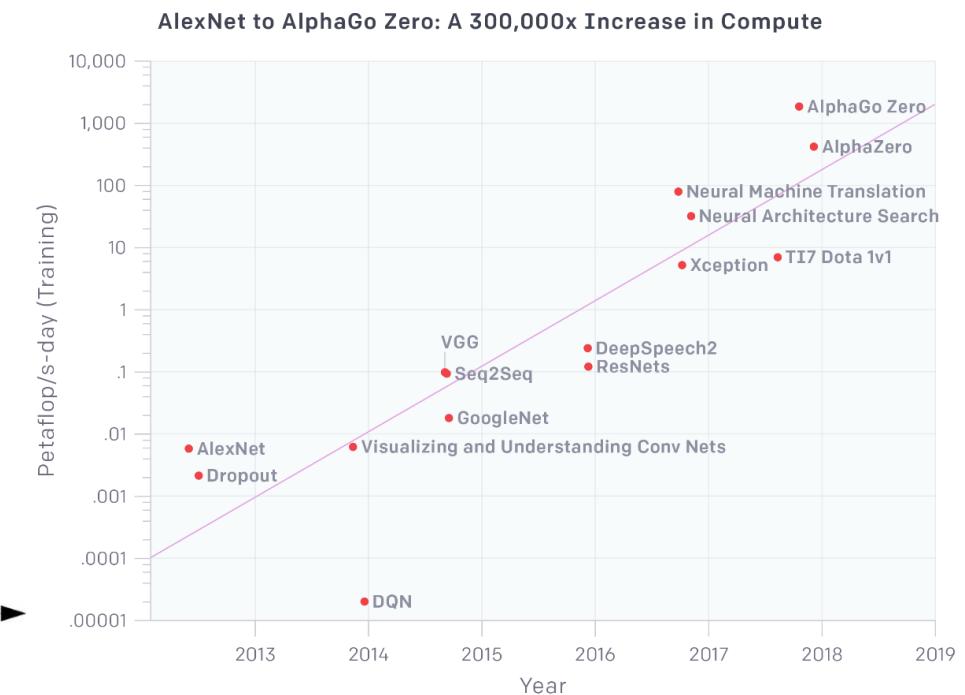
2012年，谷歌推出其知识图谱
已满足搜索中知识应用需求

大数据时代的机遇—大规模自动知识获取

- Big Data + Machine Learning+ Powerful Computation
 - ▲ • Enables large scale automatic knowledge acquisition



<http://www.erogol.com/brief-history-machine-learning/>



<https://blog.openai.com/ai-and-compute/>

数据驱动的大规模自动化知识获取

- 自下而上: 网页文本、搜索日志、购买记录……

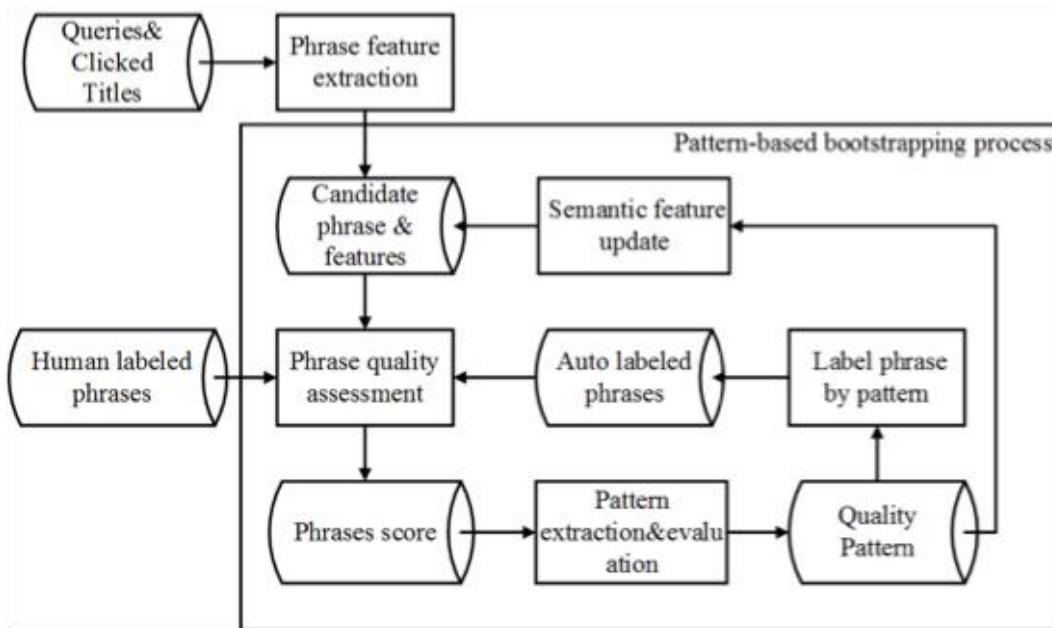


Fig. 1. Pattern-based bootstrapping framework.

基于搜索日志的消费场景知识挖掘

Hearst pattern

NP such as NP, NP, ..., and/or NP such as NP,* or|and NP
NP, NP*, or other NP
NP, NP*, and other NP
NP, including NP,* or | and NP NP, especially NP,* or|and NP

面向文本的基于规则isA知识抽取

办公用品: 中性笔||订书机||别针/回形针||胶带/胶纸/胶条

养猫必备: 猫砂||逗猫棒||猫主粮||猫抓板

洗簌用品: 衣物用刷||皂盒||脸盆||洗漱杯

基于购物记录的消费场景知识挖掘

大数据时代的机遇—众包技术

- 众包与群智成为大规模知识获取的一条新路径

案例1: 基于知识问答验证码的知识获取

- 复旦大学知识工场实验室提供知识验证码服务，通过众包的方式对现有知识进行验证

请通过验证

请点击下文中该问题答案的任意部分: 下大坪村的面积是多少? 太难了, 换一个

下大坪村隶属于云南省大理鹤庆县黄坪镇均华村委会, 该村国土面积0.92平方公里, 海拔1500米, 年平均气温20 °C, 年降水量700毫米, 农民收入主要以种植业为主。

登录!

<http://kw.fudan.edu.cn/ddemos/vcode/>

案例2: 基于众包的Taxonomy构建

- DBpedia通过众包方式构建了DBpedia Ontology

Mapping en:Infobox_book

Template Mapping (help)	
map to class	Book
Mappings	
Property Mapping (help)	
template property	author
ontology property	
ontology property	author
Property Mapping (help)	
template property	illustrator
ontology property	
ontology property	illustrator

Mapping el:Βιβλίο

Template Mapping (help)	
map to class	Book
Mappings	
Property Mapping (help)	
template property	συγγραφέας
ontology property	author
Property Mapping (help)	
template property	εικονογράφηση
ontology property	illustrator

{}{Infobox_book

```
| author =  
| title_orig =  
| translator =  
| illustrator =  
| subject =  
| genre =  
}}
```

{}{Βιβλίο

```
| συγγραφέας =  
| ειδος =  
| εκδότης =  
| πρώτη_έκδοση =  
| ISBN =  
| εικονογράφηση =  
}}
```

大数据时代的机遇—高质量UGC

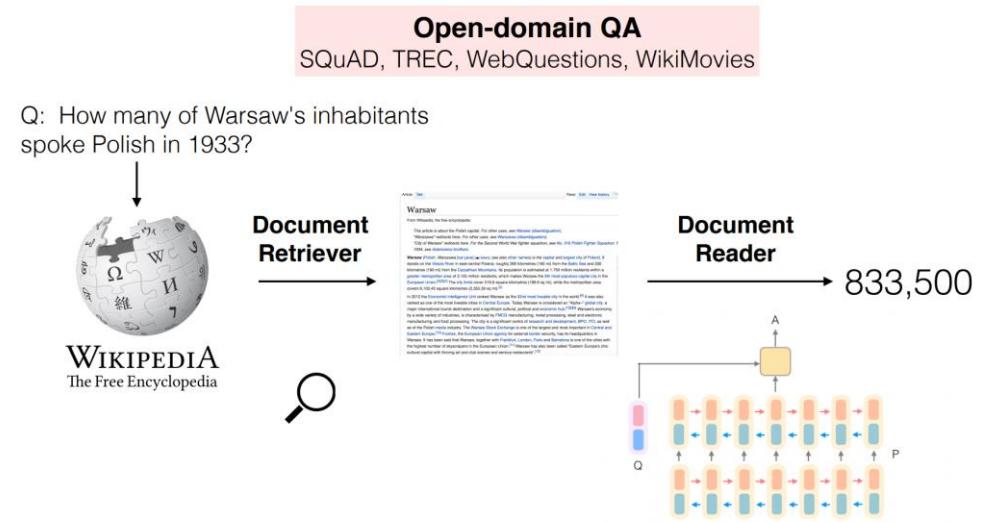
- Web2.0时代到来，产生大量的高质量UGC(User Generated Content)
 - 提供获得广大用户一致认可的高质量数据源
 - Wikipedia, 百度百科
 - 为自动挖掘知识提供了高质量数据源
 - 为构建抽取模型提供了高质量样本

周杰伦					
版本对比	更新时间	全部版本	贡献者	修改原因	区块链信息
<input type="checkbox"/>	2018-06-06 03:36	查看	w_ou	内链修复	查看
<input type="checkbox"/>	2018-03-11 16:14	查看	海渊~ ~ ~天控	内容扩充 内链	查看
<input type="checkbox"/>	2018-03-01 20:20	查看	爱锦瑟的年华	图片	查看
<input type="checkbox"/>	2018-02-28 18:59	查看	爱锦瑟的年华	内容扩充 参考资料	查看
<input type="checkbox"/>	2018-02-15 08:05	查看	紫雪510	内容扩充 参考资料	查看
<input type="checkbox"/>	2018-02-11 20:54	查看	5ssax	更正错误 图片	查看
<input type="checkbox"/>	2018-02-10 11:31	查看	Mini小北1992	完善作品信息	查看

Wiki和百科的编辑机制保证了UGC内容的质量

2022/2/25

第1章：知识图谱概述



Ref: Danqi Chen, etc.. Reading Wikipedia to Answer Open-Domain Questions

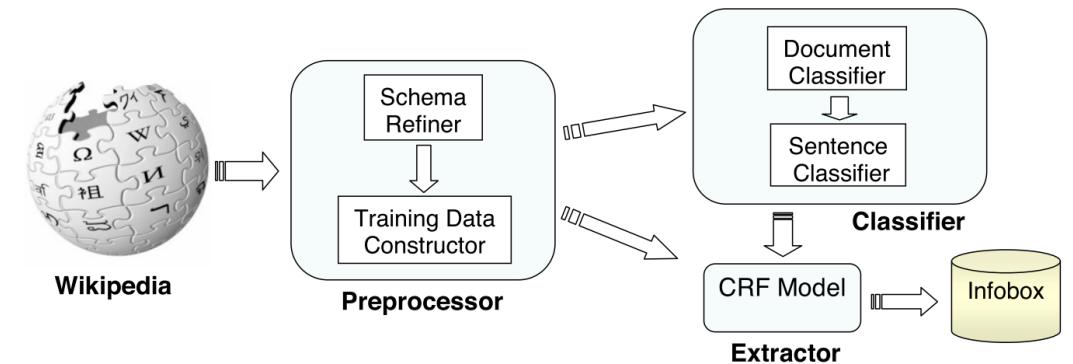


Figure 3: Architecture of KYLIN's infobox generator.

Ref: Fei Wu, etc.. Autonomously Semantifying Wikipedia

知识图谱的研究意义

未来已至：人类已经进入智能时代

- 大数据的日益积累、计算能力的快速增长为人类进入智能时代奠定了基础
- 大数据为智能技术的发展带来了前所未有的数据红利
- 机器计算智能、感知智能达到甚至超越人类

2012年，在图像识别的国际大赛ILSVRC(大型视觉辨识挑战竞赛)中，加拿大多伦多大学的研究团队基于深度卷积神经网络的模型[1]夺冠，把TOP5错误率降到15.3%，领先第二名超过十个百分比，震惊学术圈。

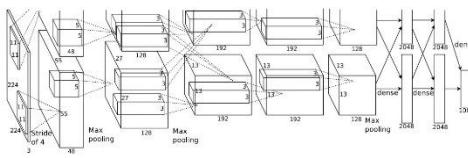


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440-186,624-64,896-43,264-4096-4096-1000.
<http://yisong.csail.mit.edu/21.110/day2.pdf>

2016年，Google全资收购的DeepMind推出名为AlphaGo的围棋程序[2]，以4:1的总比分击败世界顶级职业围棋选手李世石，让全世界开始关注人工智能技术巨大的应用前景。

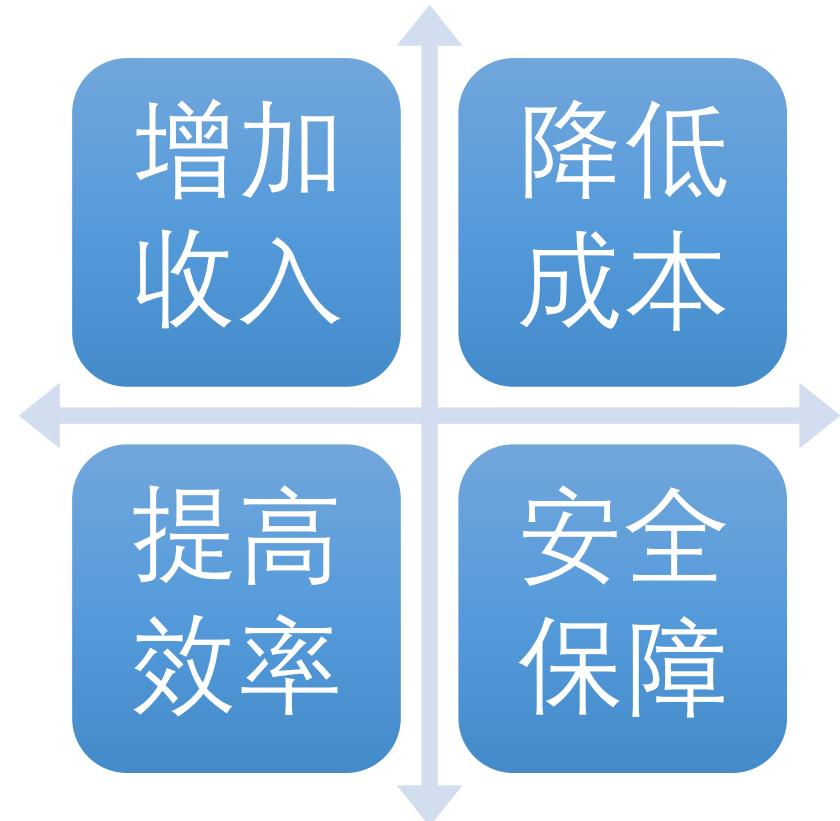


2017年，DeepMind联合游戏公司暴雪，宣布共同开发可以在“星际争霸2”中与人类玩家对抗的人工智能，并且发布了旨在加速即时战略游戏的人工智能应用的工具集[3]。



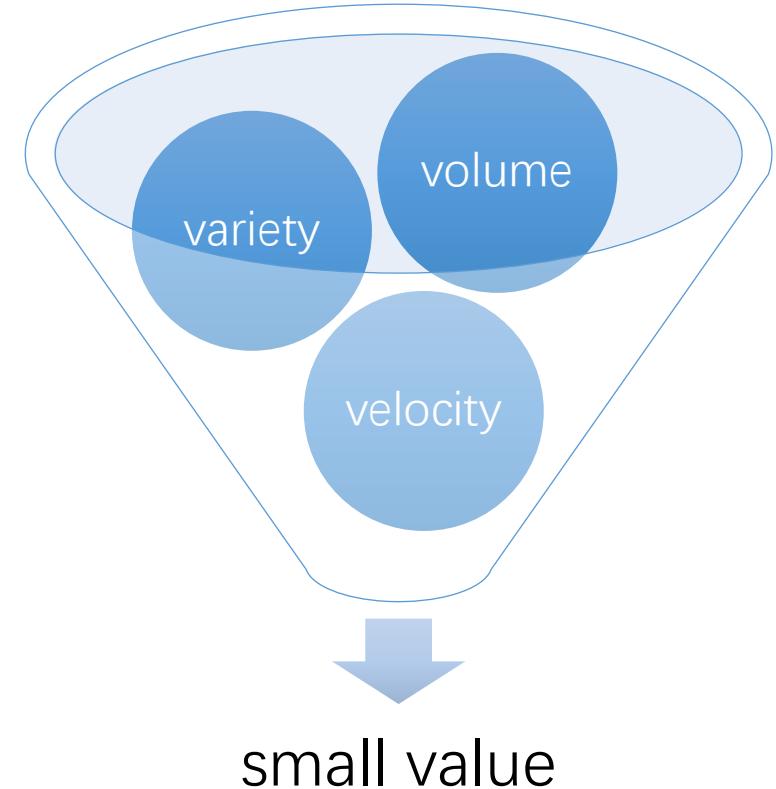
各行业智能化升级与转型

- 宏观形势
 - 人口红利消失
 - 专家成本高昂
 - 实体经济结构转型
 - 传统行业发展内涵升级
- 技术发展态势
 - 数据丰富
 - 场景丰富
 - 丰富AI技术积累



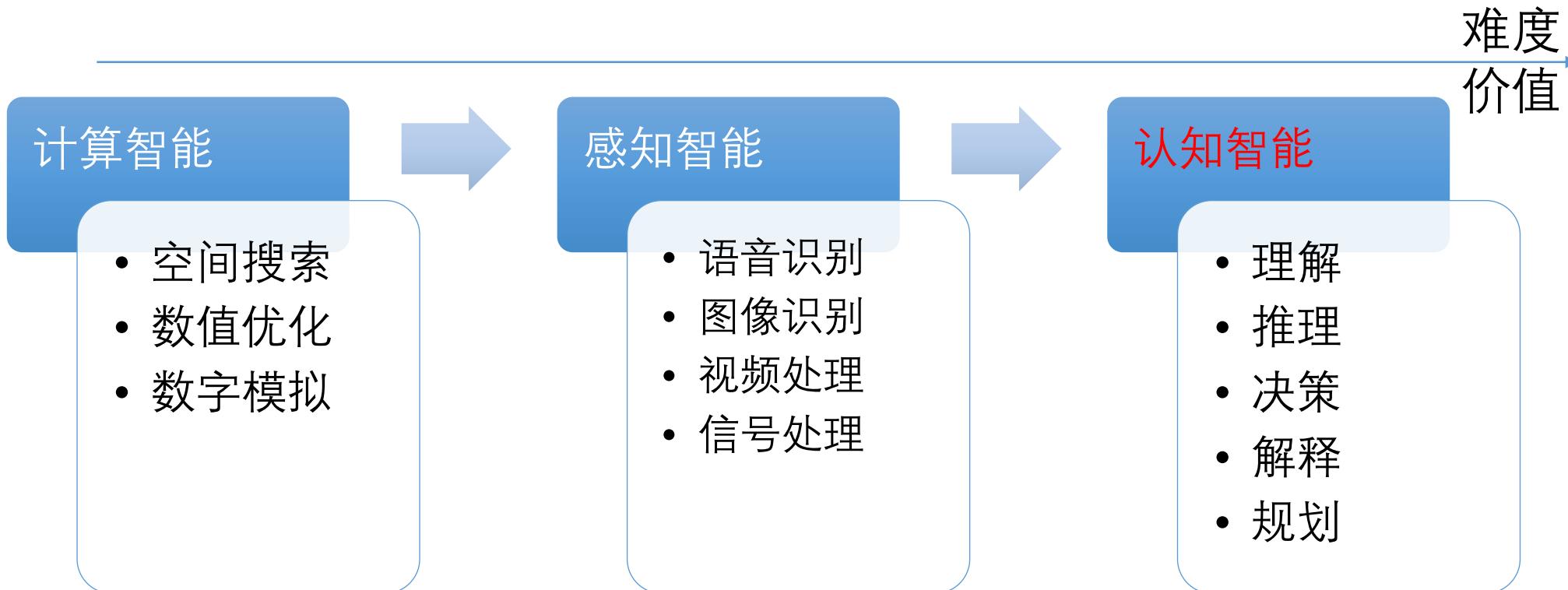
大数据价值变现困难倒逼智能化转型

- 《中国经济周刊》
 - “**投上百亿建大数据中心 内部称产出十分微小”
- 英特尔中国研究院院长吴甘沙
 - “鉴于大数据信息密度低，大数据是贫矿，投入产出比不见得好。”
- 李国杰院士
 - “实际上，大数据的价值，主要体现在它的驱动效应上，大数据对经济的贡献，并不完全反映在大数据公司的直接收入上，应考虑对其他行业效率和质量提高的贡献。”



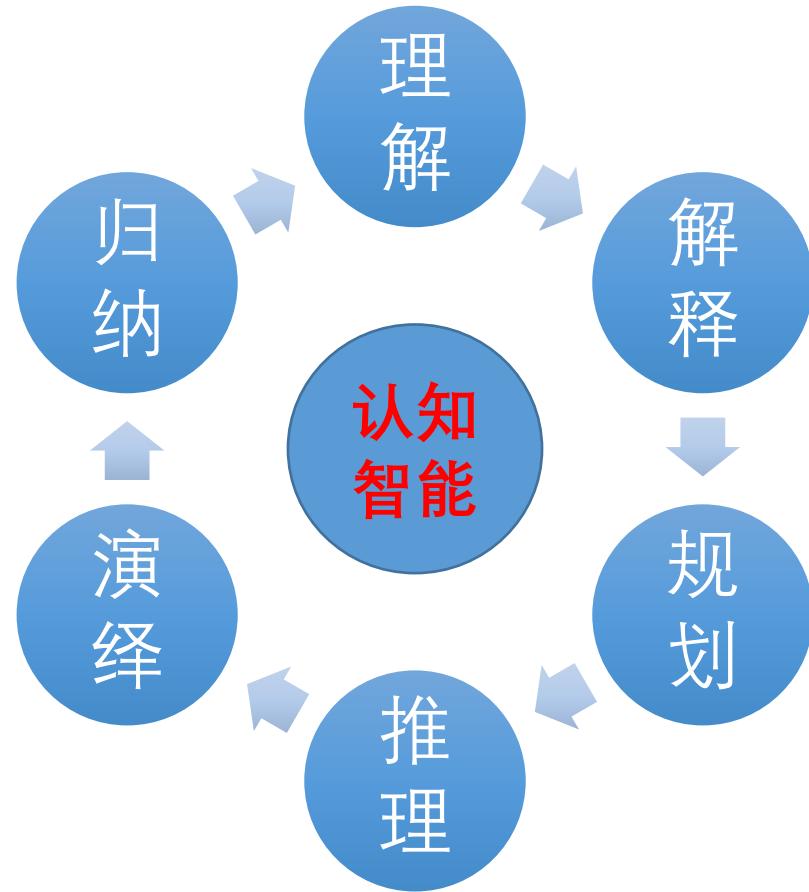
大数据价值变现的尴尬现状：**高射炮打蚊子，大材小用**

智能化需要机器智能，特别是认知智能



- 随着数据红利消耗殆尽，以深度学习为代表的感知智能遇到天花板
- 认知智能将是未来一段时期内AI发展的焦点，是进一步释放AI产能的关键

认知智能是智能化的关键



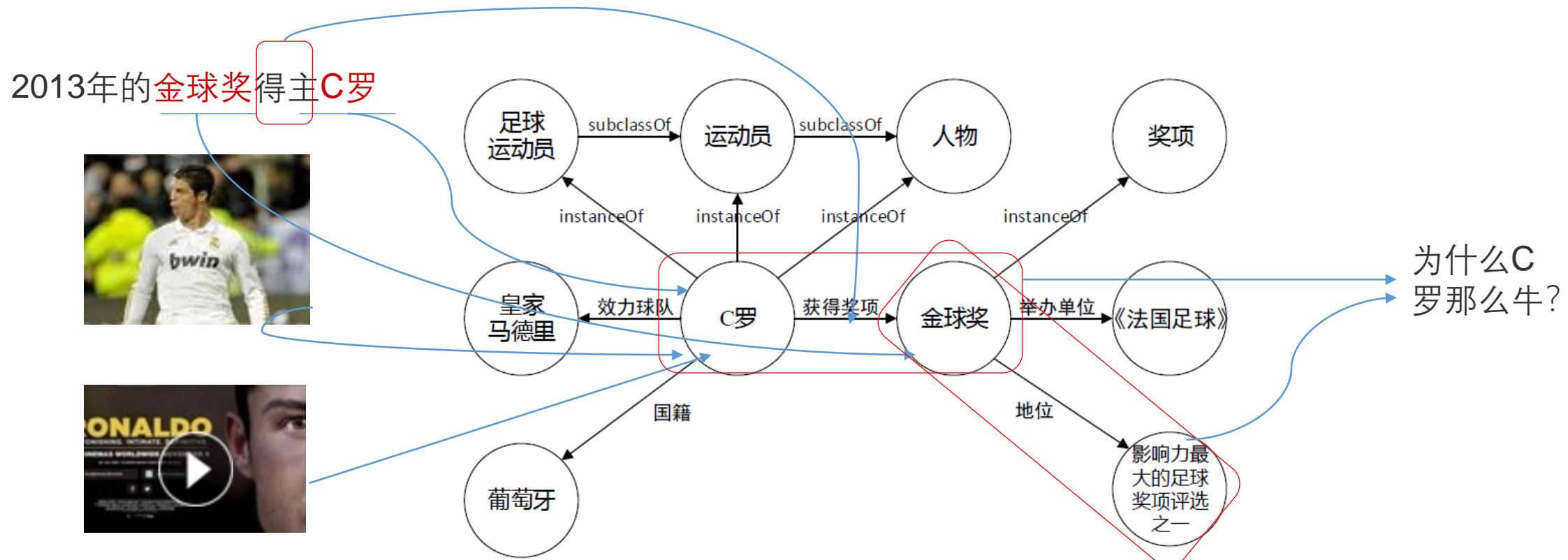
Can machine *think like humans?*



■ 理解与解释是后深度学习时代人工智能的核心使命之一

知识图谱使能认知智能

- 机器理解数据的本质：建立从数据到知识库中实体、概念、关系的映射
- 机器解释现象的本质：利用知识库中实体、概念、关系解释现象的过程



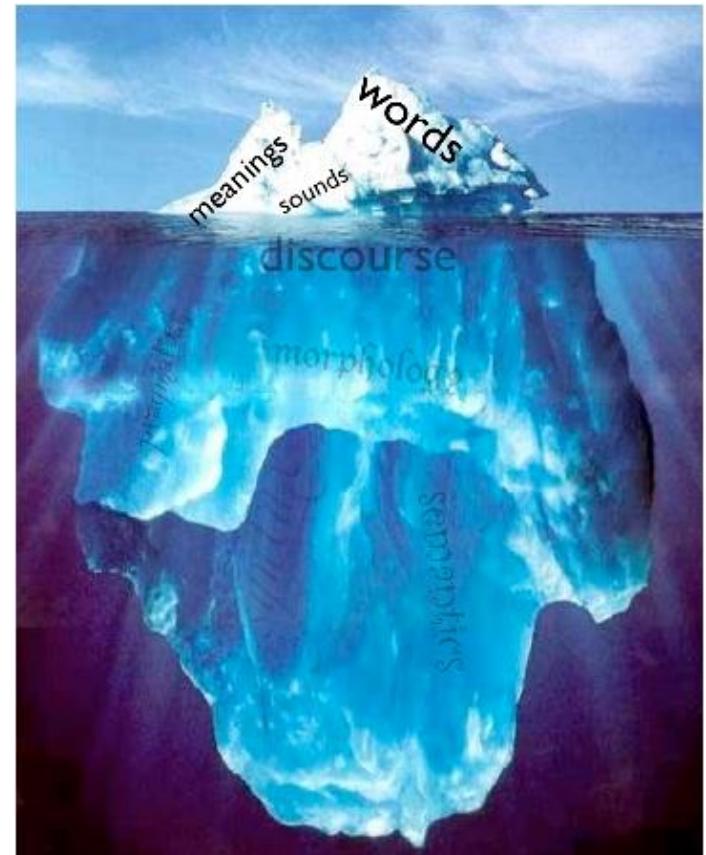
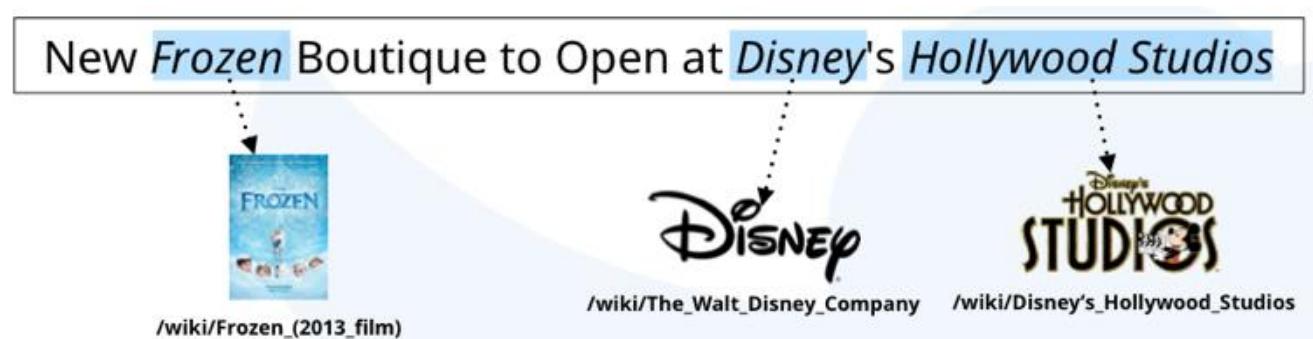
机器语言理解需要背景知识

Language is complicated

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

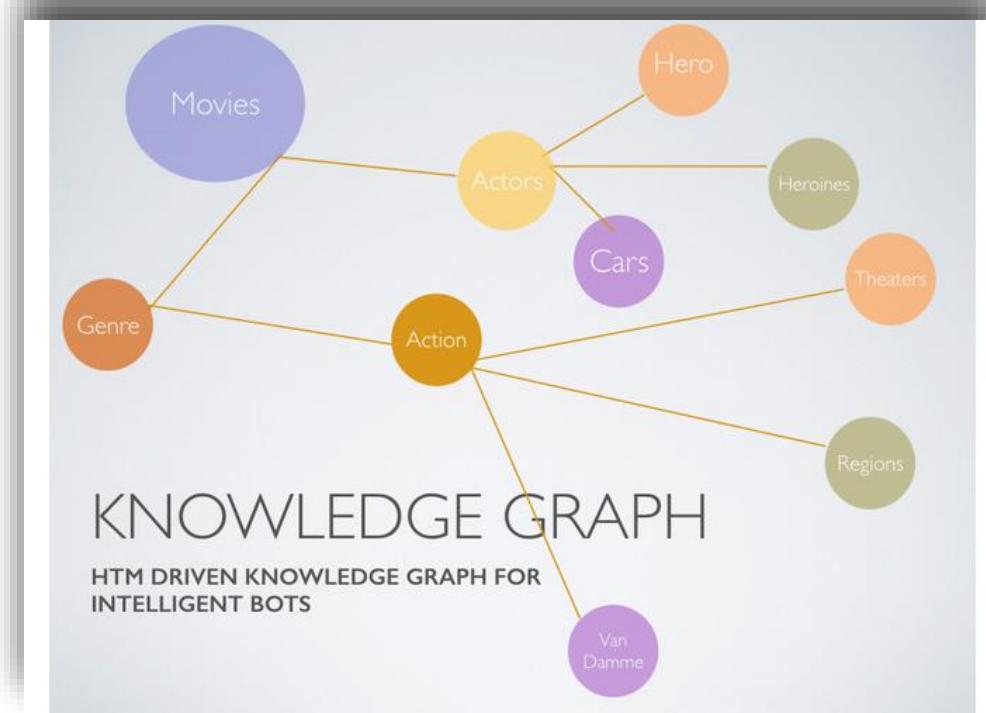
Language understanding is difficult

- Grounded only in human cognition
- Needs significant background knowledge



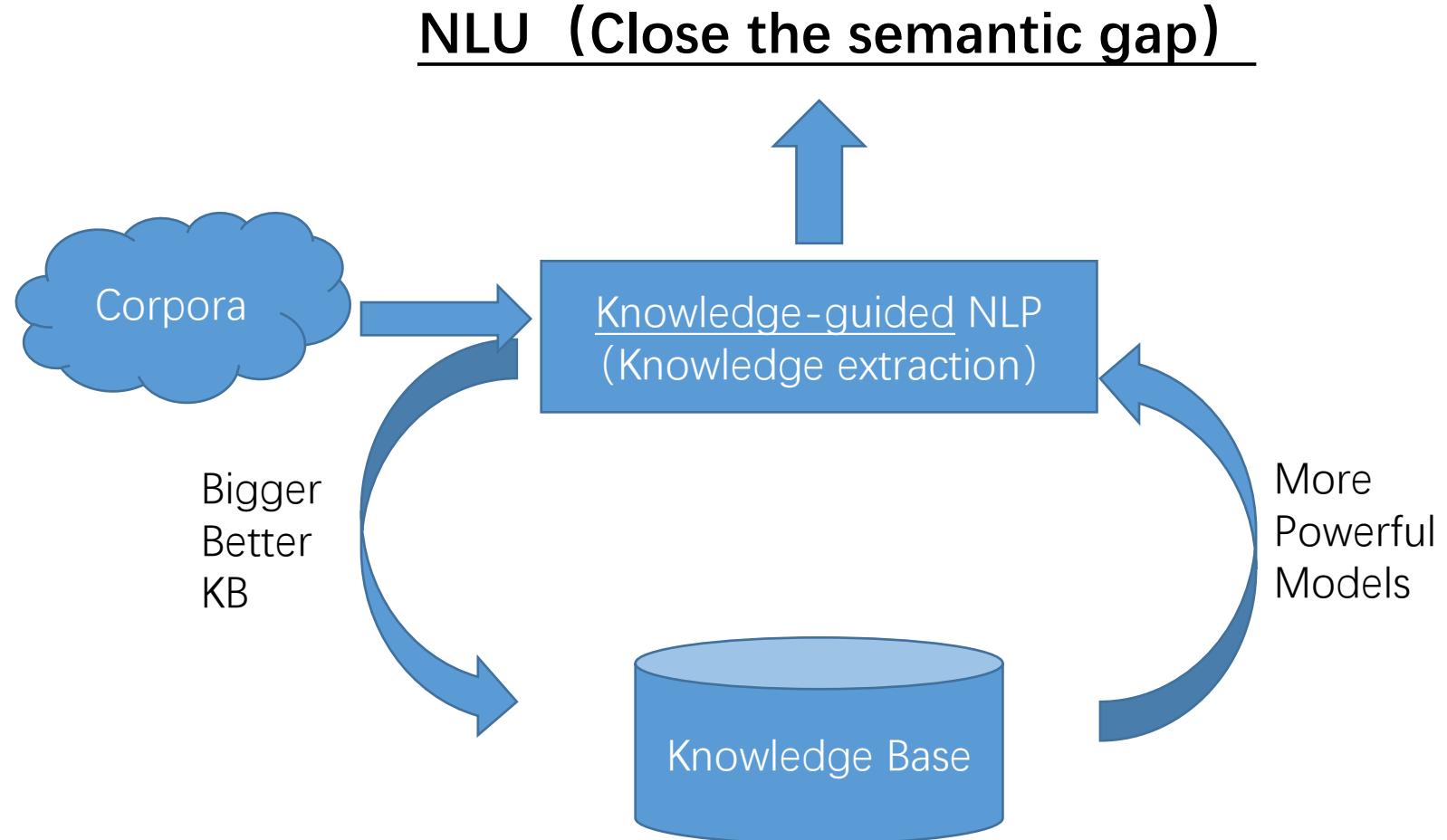
知识图谱使能(Enable)机器语言认知

- Language understanding of machines needs knowledge bases
 - Large scale
 - Semantically rich
 - Friendly structure
 - High quality
- Traditional knowledge representations can not satisfy these requirements, but KG can
 - Ontology
 - Semantic network / frame
 - Texts



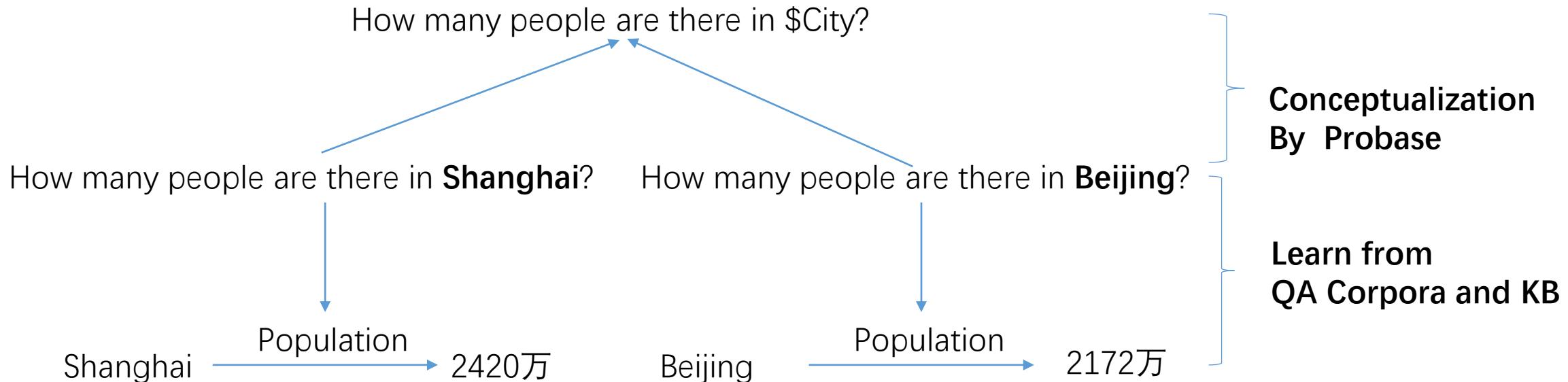
■ **NLP+KB= NLU**, NLP=Natural language processing, NLU=natural language understanding

The roadmap of knowledge-guided NLP



Example: Using concepts to understand a natural language?

- Representation: **concept based templates**.
 - Questions are asking about **entities**. The semantic of the question is reflected by its corresponding concept.
 - Advantage: Interpretable, user-controllable
- **Learn templates from QA corpus, instead of manfully construction.**



知识图谱使能可解释人工智能

鲨鱼为什么那么可怕?
因为它们是食肉动物

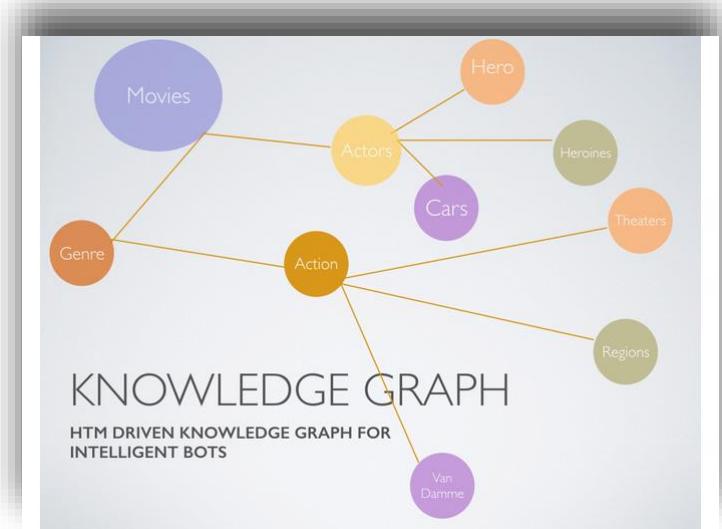
鸟儿为何能够飞翔?
因为它们有翅膀

鹿晗关晓彤最近为何刷屏?
因为关晓彤是鹿晗女朋友

概念

属性

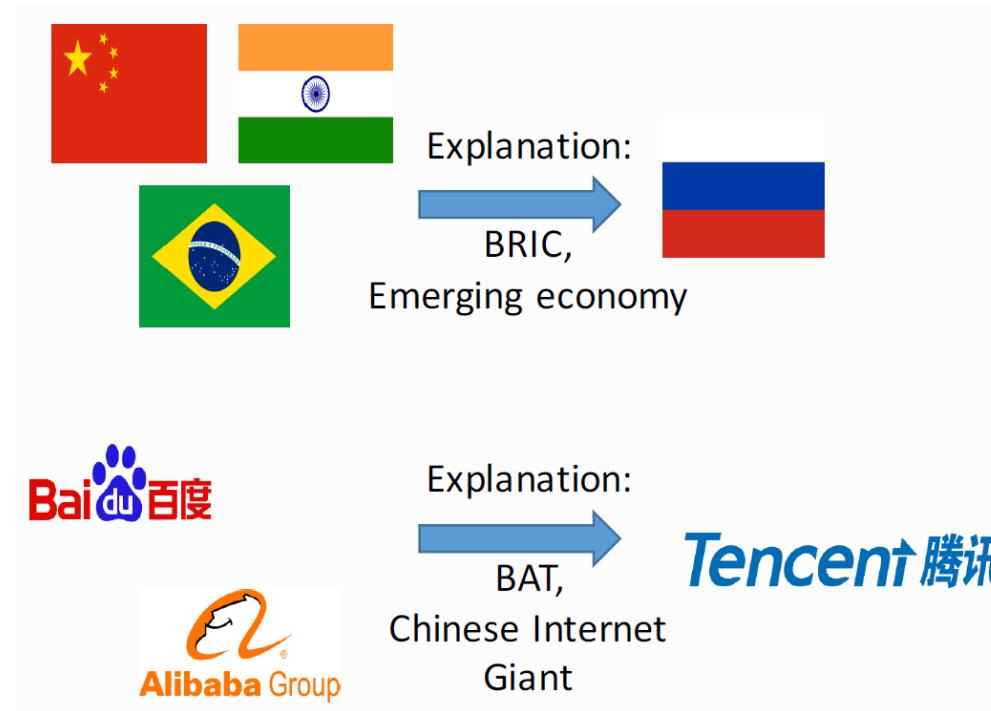
关系



解释取决于人类认知的基本框架;
概念、属性、关系是认知的基石

"Concepts are the glue that holds our mental world together"
--Gregory Murphy

Example 1: Explainable entity recommendation using taxonomy

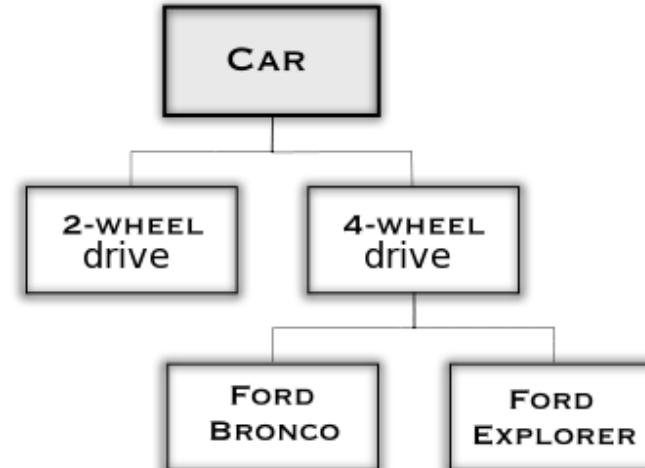


Problem:

Given a set of entities, can we understand its concept and recommend a most related entity?

Applications:

E-commerce: if users are searching samsung s6, and iPhone 6, what should we recommend and why?



Taxonomy

[Yi Zhang, et al, 2017]

Example 2: Explain a Concept/Category using Properties

Problem:

How do we understand a concept/category?

Example:

How to understand “Bachelor”

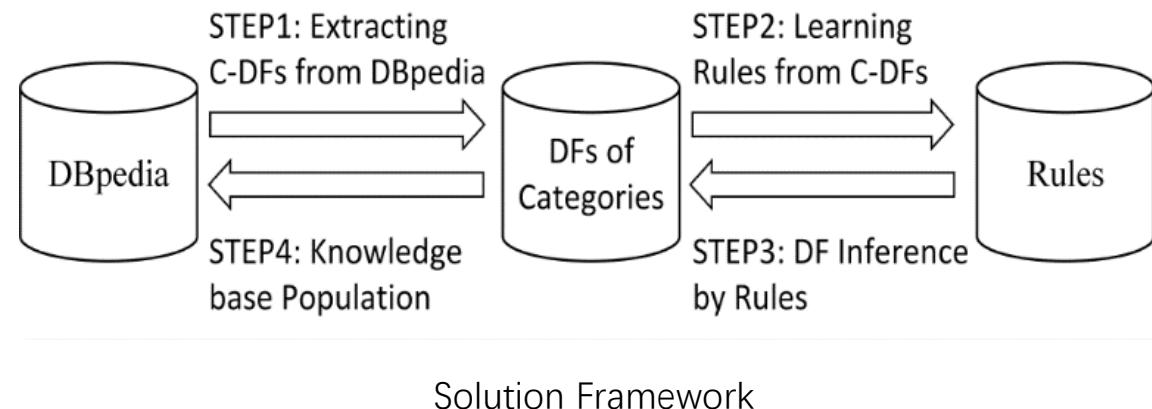
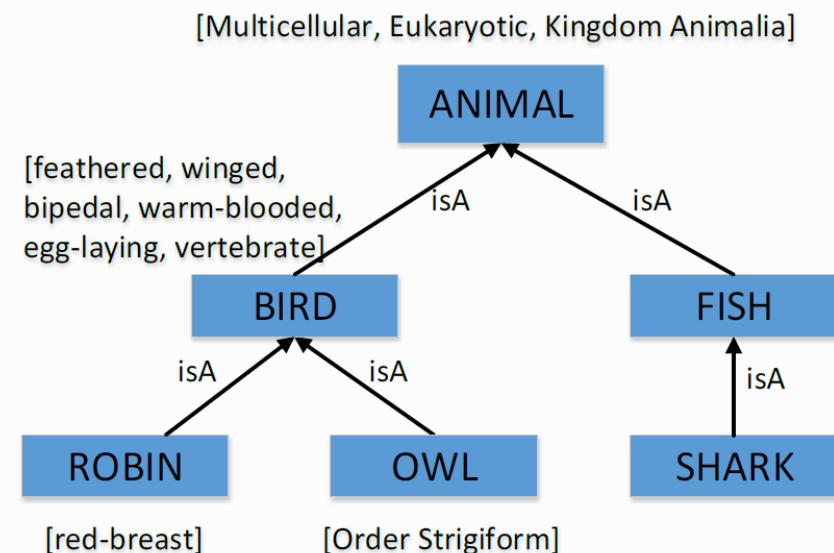
=> (Sex=man, Marriage status=unmarried)

Basic Idea:

Mining Dbpedia, using properties to explain a category

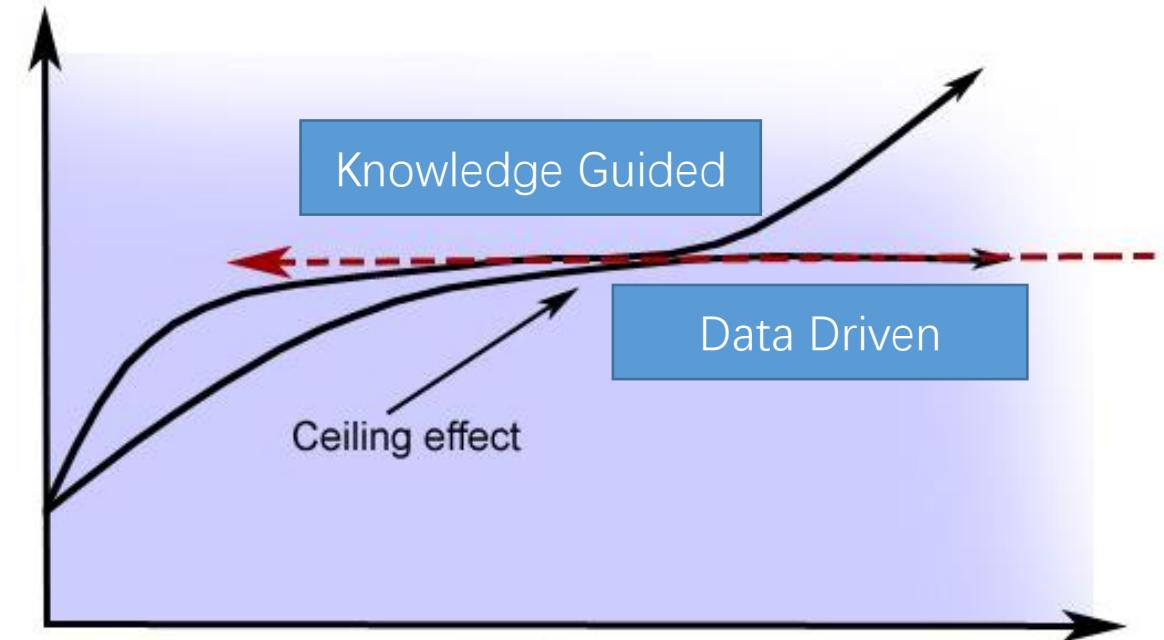
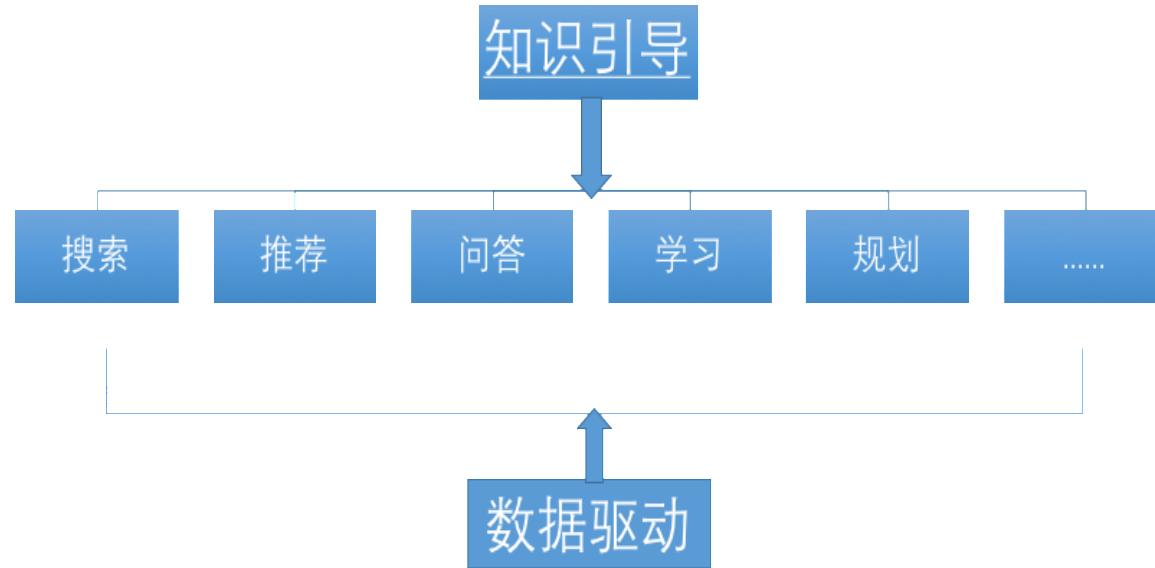
Model:

Mining **Defining Features** from DBpedia



[Bo Xu, et al, 2016]

知识引导将成为解决问题的主要方式



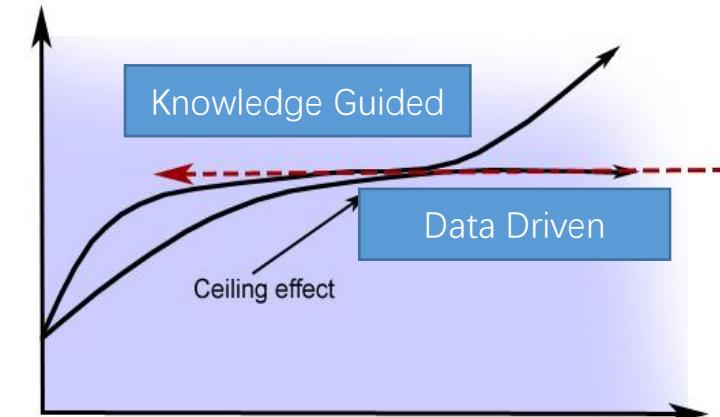
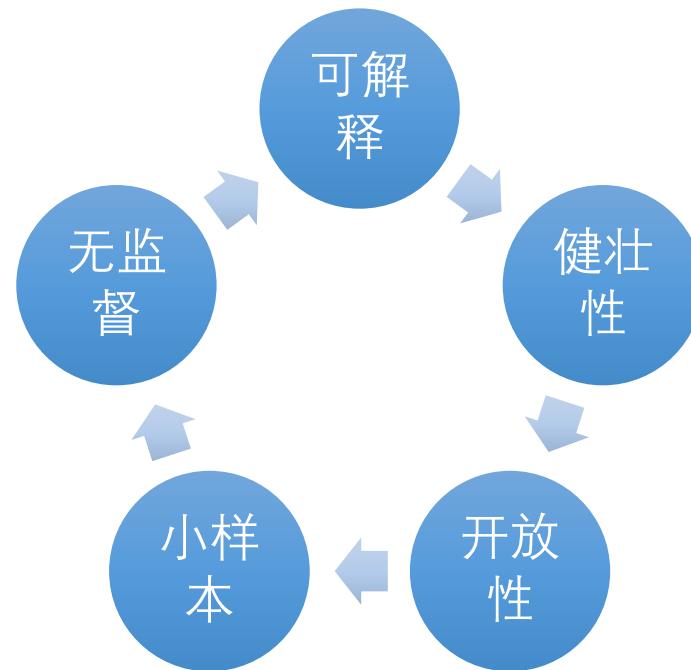
- “数据驱动”利用统计模式解决问题
- 单纯依赖统计模式难以有效解决很多实际问题

张三把李四打了，他进医院了

张三把李四打了，他进监狱了

知识引导突破统计学习的天花板

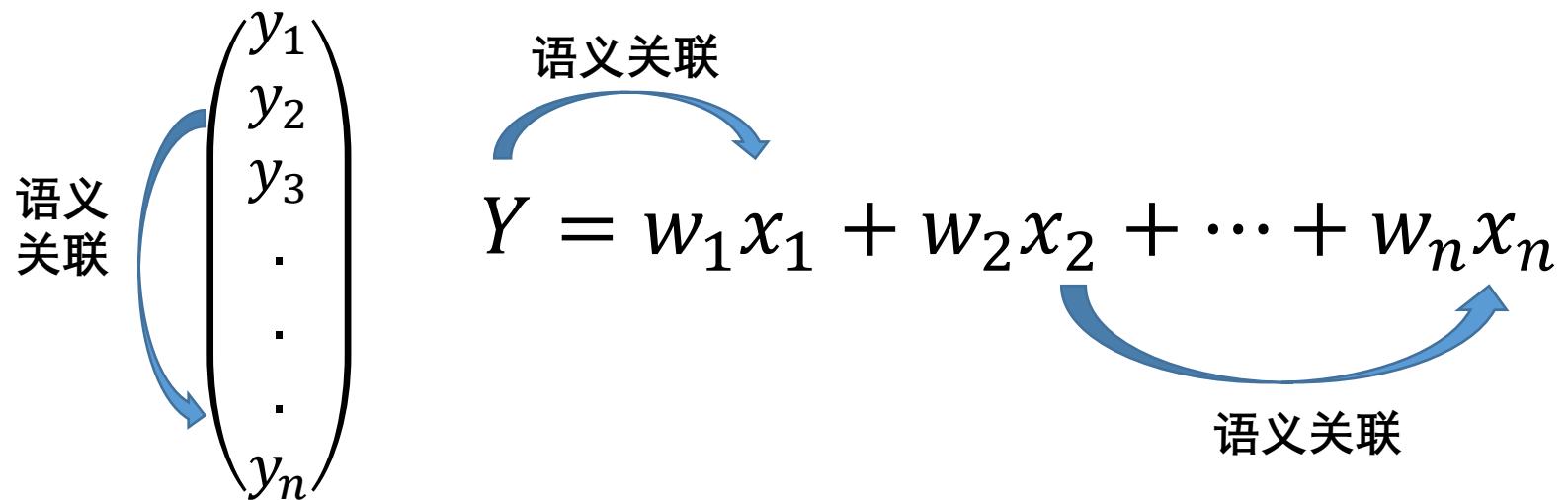
- 依赖数据驱动的统计学习基于统计模式解决问题
- 单纯统计模式日益面临性能的天花板



张三把李四打了，他进医院了
张三把李四打了，他进监狱了

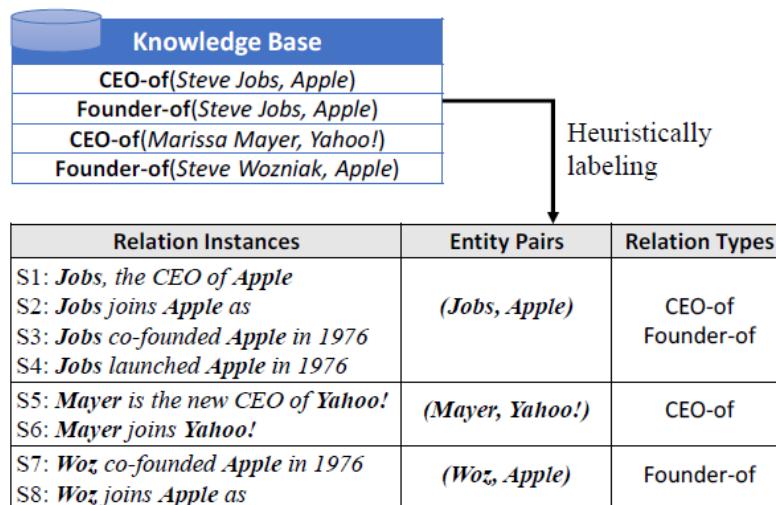
开放性：提升机器学习的开放性

- 难以处理开放性问题
 - Zero-shot learning: Unknown Labels
 - One-shot learning: Rare labels
- 忽略解释变量、响应变量及其之间的各类语义关联



无监督：知识库给机器学习提供丰富的样本

- 远程监督 (Distant Supervision) 提供大规模自动化弱标注样本
- 领域专家构建的知识库质量精良，提供高质量的种子样本

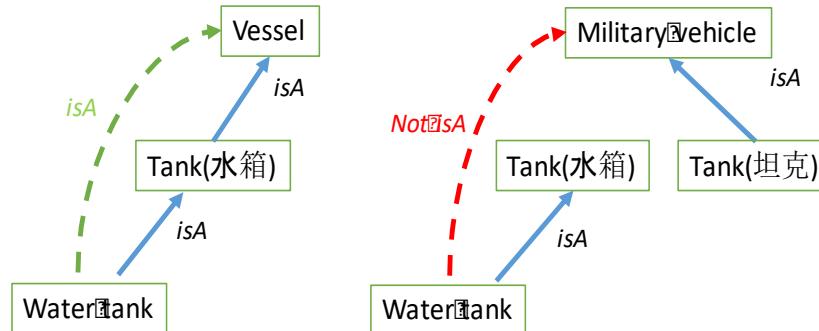


Idea: 通过结构化知识库与文本比对，完成大规模弱标注，广泛应用于实体识别、关系抽取等任务

Ref: Distant supervision for relation extraction without labeled data. ACL09

Lexical Taxonomy isA传递性判定问题

Example: Einstein isA Physicist, Physicist isA Job, is Einstein a Job?



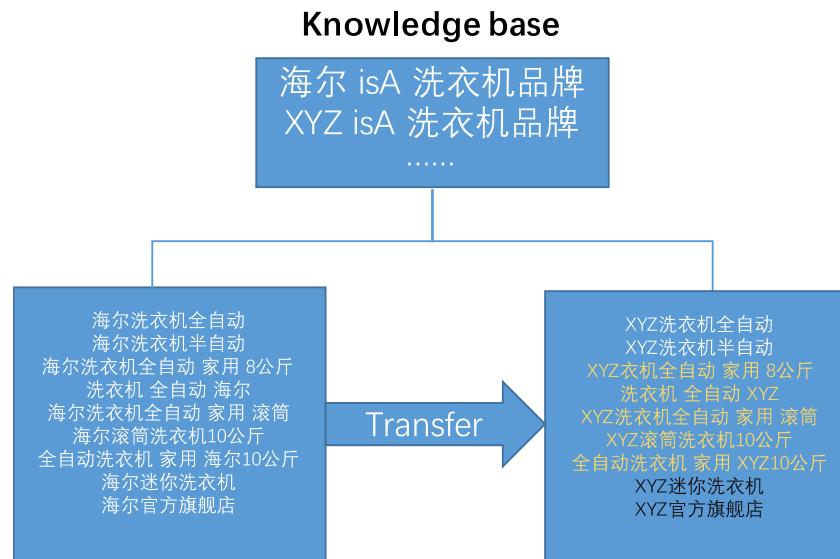
正例: water tank - tank - vessel 负例: water tank - tank - military vehicle

Idea: 使用专家构建的WordNet，自动化构造判断isA传递性的标注样本

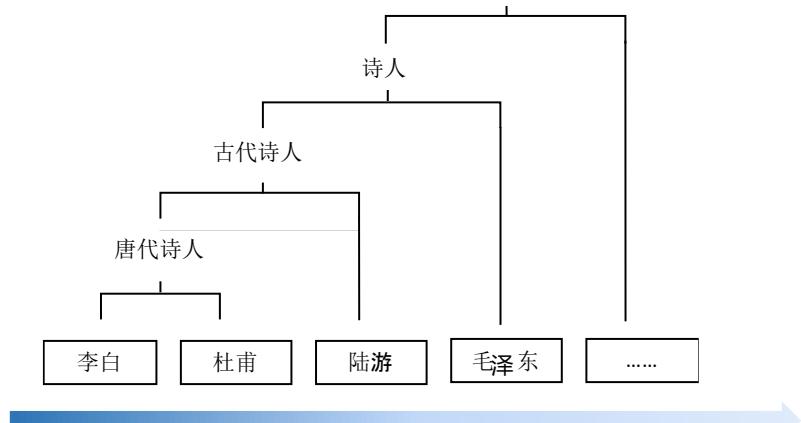
Ref: On the Transitivity of Hypernym-hyponym Relations in Data-Driven Lexical Taxonomies, (AAAI 2017)

小样本：知识引导下的样本增强

- 符号知识可以广泛用于指导样本的生成、选择、增强、优化
 - 边界样本选择、负样本选择、代表性样本选择、Unknown Unknowns识别
- 有效应对样本稀缺、有效处理长尾对象



Idea: 使用Taxonomy指导样本生成，提升长尾品牌词的embedding效果



相似 Positive Sample:
Q: 李白被誉为什么? A: 诗仙
Negative Samples:
Q: 杜甫的中文名是?
Q: 陆游经典作品有哪些?
.....

Idea: 使用Taxonomy指导负样本生成，提升问答准确性；落地在知识工场知识问答系统（不倒翁问答）在通用领域做到85%的准确率

健壮性：符号知识优化机器学习模型

- 符号知识被广泛用于，提升机器学习对于有偏样本的健壮性
 - 构建正则项、约束、事后检验、注意力机制

$$\text{Maximize} \sum_{t \in T} \left(\max_{m \in M_e} P(t|m) - \theta \right) \times x_{e,t}$$

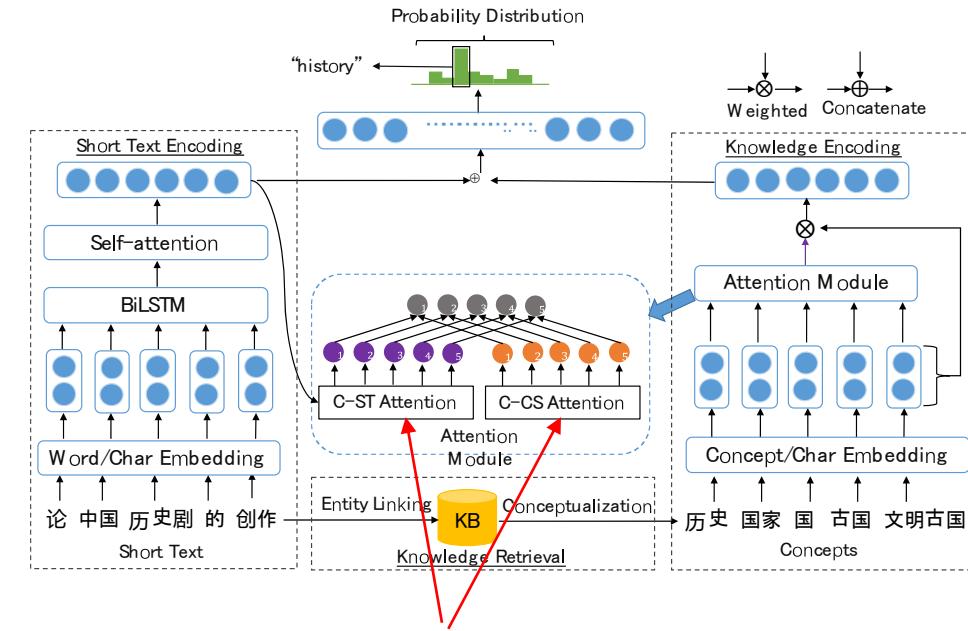
Subject to

$$\forall_{ME(t_1, t_2)} x_{e,t_1} + x_{e,t_2} \leq 1$$

$$\forall_{ISA(t_1, t_2)} x_{e,t_1} - x_{e,t_2} \leq 0$$

Type Disjointness
Constraint

Type Hierarchy
Constraint



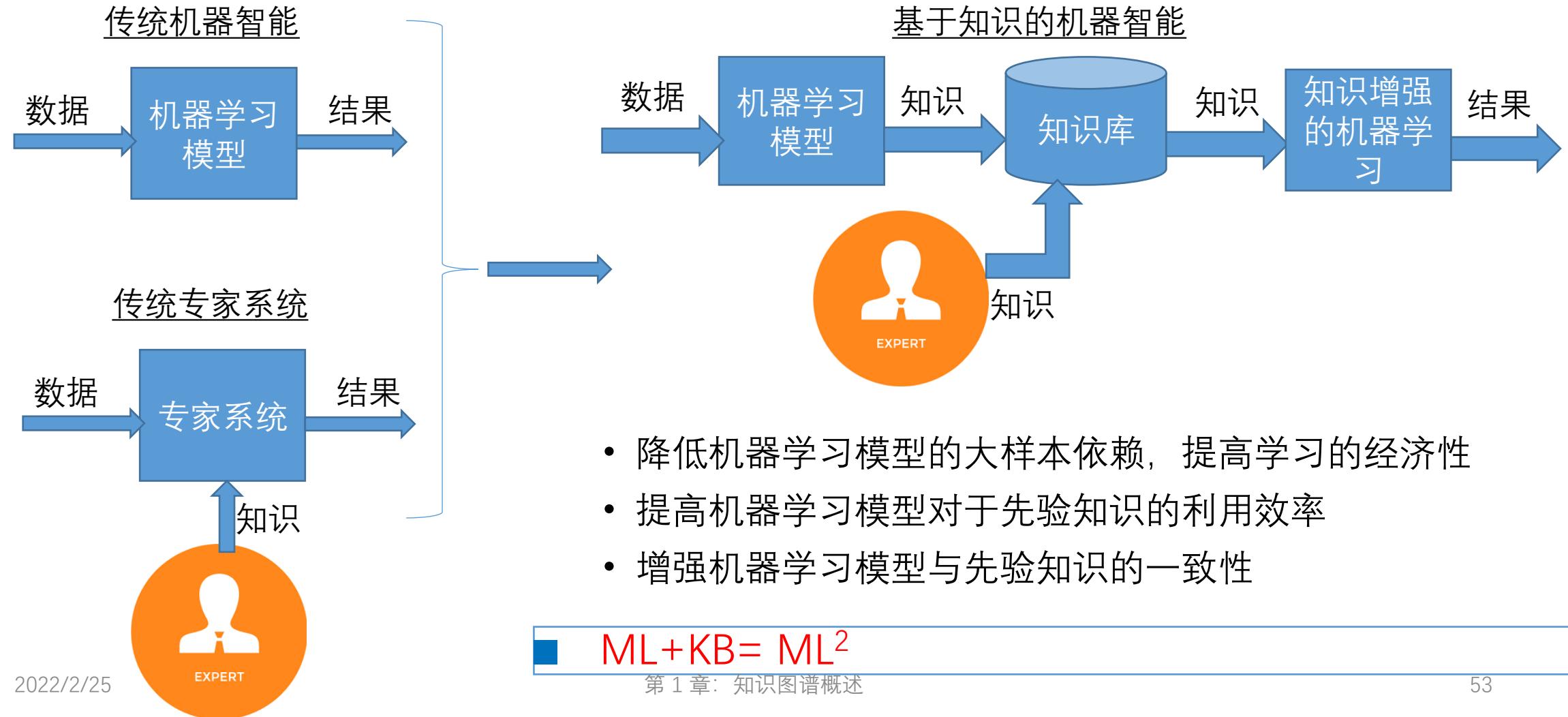
Idea: 使用Type之间的语义约束对于结果进行筛选

Ref, METIC: Multi-Instance Entity Typing from Corpus, CIKM
2018

Idea: 使用CN-Probase中的概念关系构建Attention,
优化端文本分类模型

Ref; Deep Short Text Classification with Knowledge Powered Attentions, (AAAI 2019)

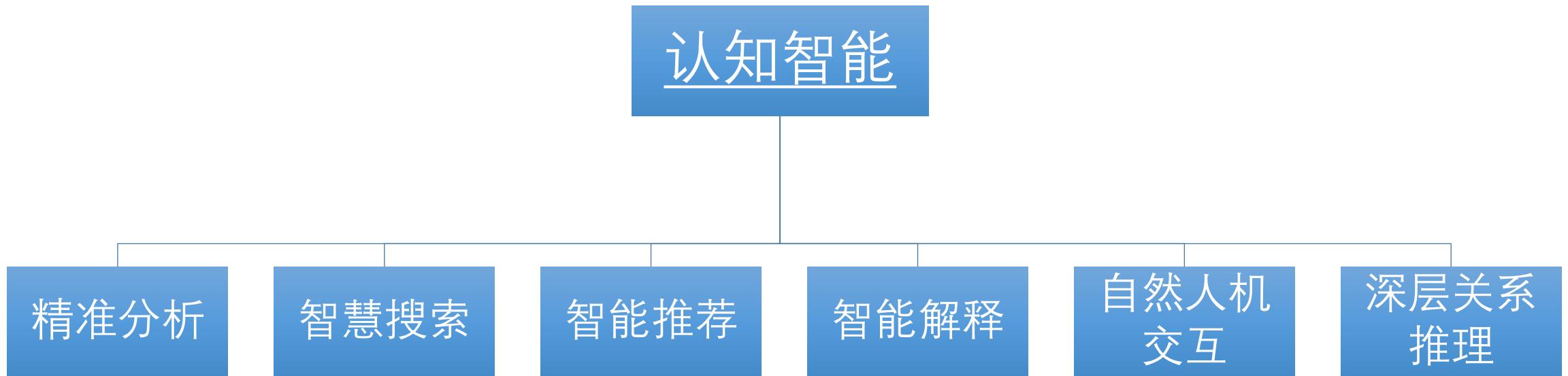
知识将显著增强机器学习能力



知识图谱的应用价值

知识图谱应用

- 认知智能应用需求广泛多样，需要对传统信息化手段的**全面而彻底**的革新
- 认知智能：人类脑力解放，机器生产力显著提高



精准分析

- 精准化数据分析
 - 舆情分析
 - 热点统计
 - 军事情报分析
 - 商业情报分析
- 精细化数据分析
 - 酒店评论抽取
 - 个性化制造

[深扒王宝强离婚内幕 最大祸根源于谁_百山探索](#)

[深度解析宝宝离婚闹剧事件 细说婚姻幸福真谛!_央广网](#)

[宝强离婚最新动态,DNA结果公布马蓉原形毕露_新闻频道_中华网](#)

.....宝宝不知道宝宝的宝宝是不是宝宝亲生的宝宝，宝宝现在担心的是宝宝的宝宝不是宝宝的宝宝如果宝宝的宝宝真的不是宝宝的宝宝那就吓死宝宝了宝宝的宝宝为什么要这样对待宝宝，宝宝很难过，如果宝宝和宝宝的宝宝因为宝宝的宝宝打起来了，你们到底支持宝宝还是宝宝的宝宝！【宝宝心里苦，但是宝宝不说】

[军民融合南海掀波 陆渔船舰队近逼菲中业岛](#)

→ 菲律宾 相关

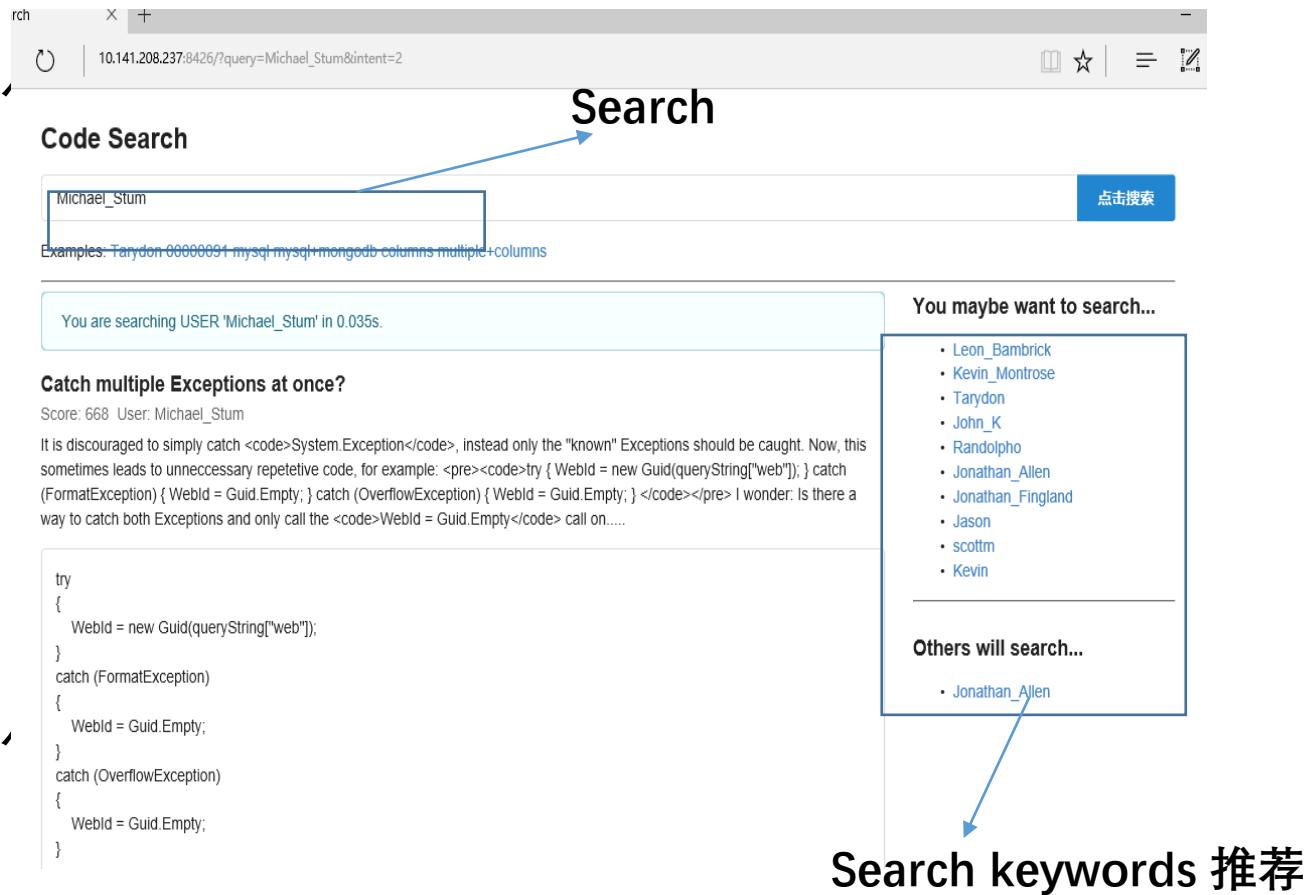
[意大利华人捐古版中国地图 证明钓鱼岛为中国领土](#)

→ 日本 相关

■ 大数据的精准、精细分析需要智能化技术支撑

智慧搜索

- 精准搜索意图理解
 - 精准分类、语义理解、个性化推荐
- 复杂多元对象搜索
 - 表格、文本、图片、视频
 - 文案、素材、代码、专家
- 多粒度搜索
 - 篇章级、段落级、语句级
- 跨媒体搜索
 - 不同媒体数据联合完成搜索

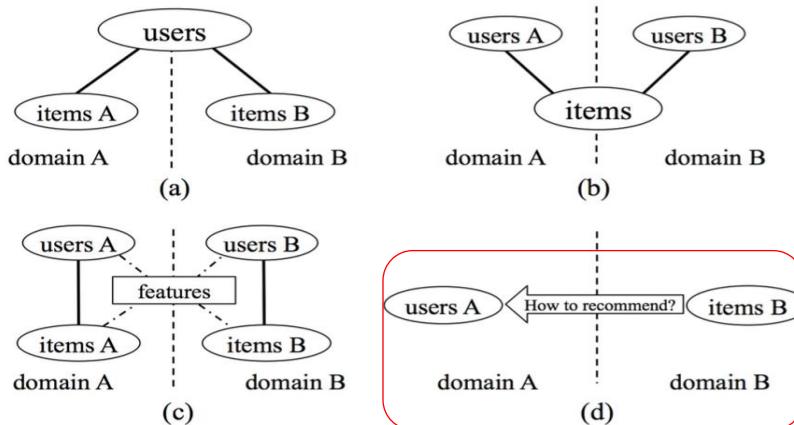


一切皆可搜索，搜索必达

智能推荐

- 场景化推荐
- 任务型推荐
- 冷启动环境下的推荐
- 跨领域推荐
- 知识型推荐

电商领域的
场景化推荐



跨领域推荐，比如给微博用户推荐taobao商品，存在巨大的vocabulary gap

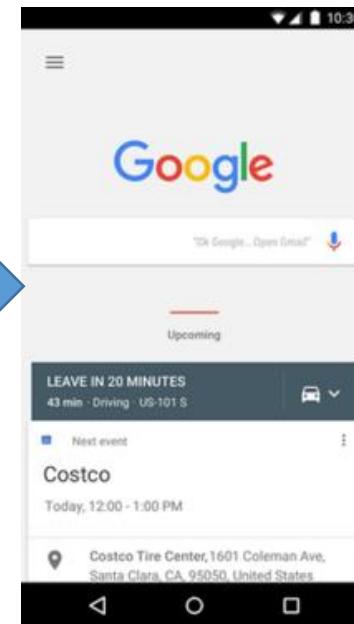


■ 精准感知任务与场景，想用户之未想
■ 从基于行为的推荐发展到行为与语义融合的智能推荐

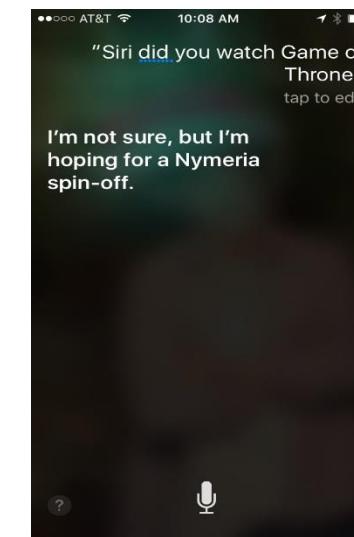
自然人机交互



Google Now



Apple Siri



Amazon Alexa



KW Xiao Cui



Question Answering (QA) systems in industries and academics

- 人机交互方式将更加自然，对话式交互取代关键词搜索成为主流交互方式
- 一切皆可问答：图片问答、新闻问答、百科问答

决策支持：深层关系发现 / 推理



Why baoqiang select Qizhun Zhang as his lawyer?



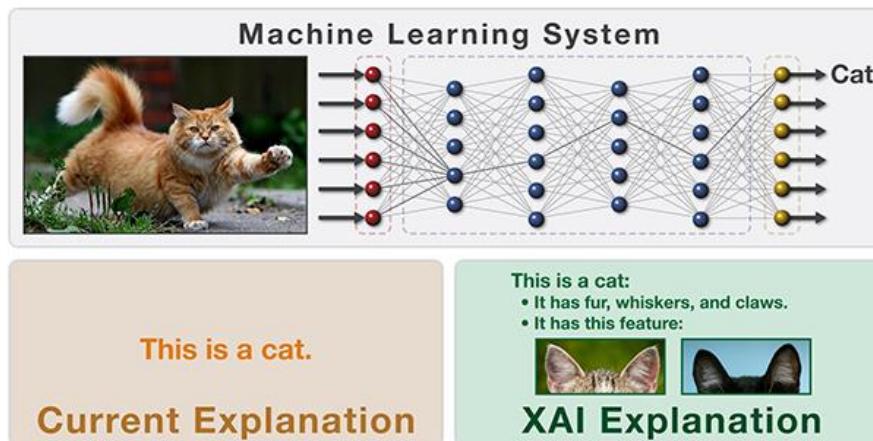
Why A invests B?

■ 隐式关系发现、深层关系推理将成为智能的主要体现之一

智能解释

- 事实解释
- 关系解释
- 过程解释
- 结果解释

解释机器学习过程



找到约 3,400,000 条结果 (用时 0.80 秒)

Things you didn't know about Donald Trump's wife - Nicki Swift

www.nickiswift.com/7996/things-didnt-know-donald-trumps-wife/ 翻译此页

Although Donald Trump has made quite a show for himself, his current wife, Melania, has mostly ... Melania called him after she returned from a photo shoot in the Caribbean. ... No matter who you are married to, you still need to lead your life.

解释事实

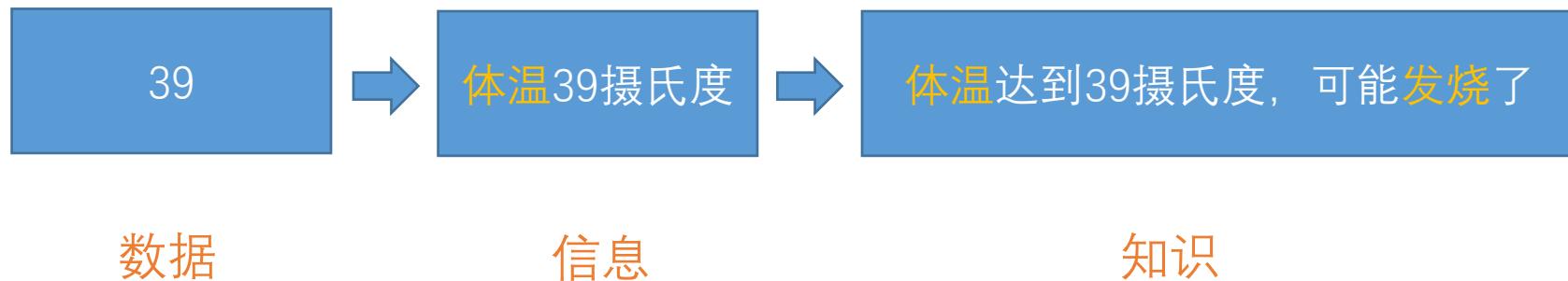
■ 解释是智能的重要体现之一，将是人们对于智能系统的普遍期望

■ 可解释是智能系统决策结果被采信的前提

知识图谱的分类

数据、信息与知识

- 数据：对客观世界的符号化记录
- 信息：被赋予意义的数据
- 知识：信息之间有意义的关联



知识分类-事实知识

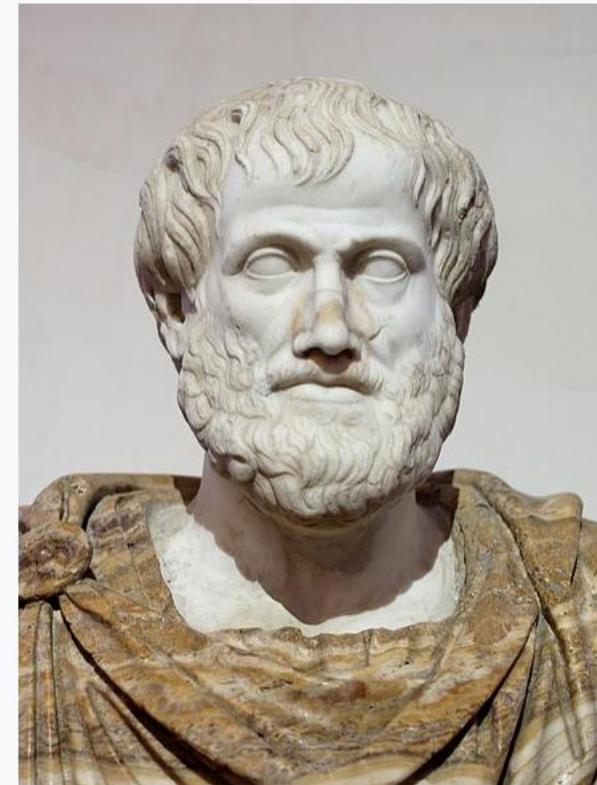
- **事实知识 (Factual Knowledge)** : 关于某个特定实体的基本事实

- 实体的特定属性或关系
- 复杂的文本描述

西方古典哲学
的集大成者

出生日期属性

亚里士多德
Αριστοτέλης

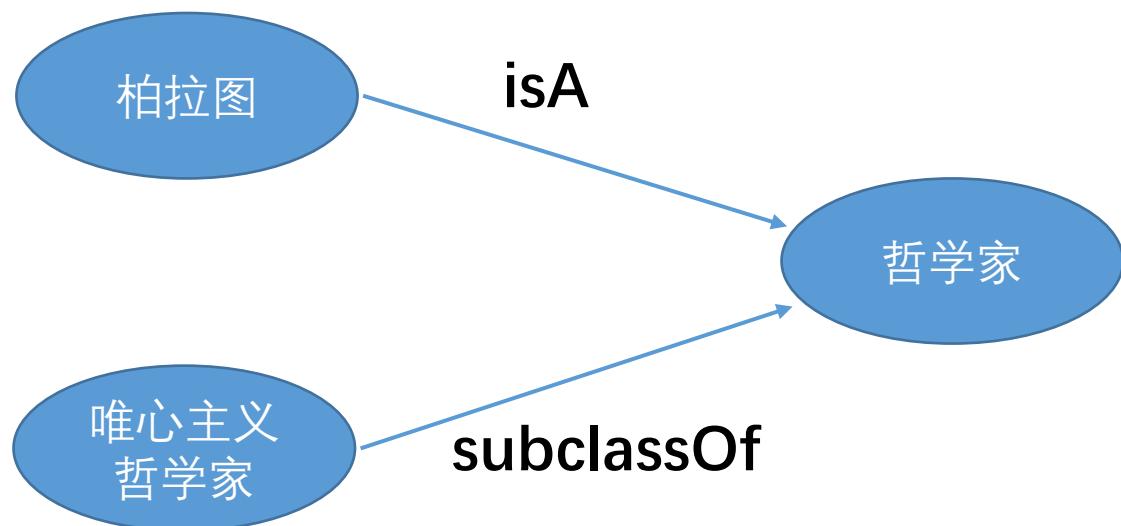


亚里士多德

出生 公元前384年6月19日

知识分类-概念知识

- 概念知识 (Taxonomy Knowledge)
- 实体与概念之间的类属关系 (isA关系)
- 子概念与父概念之间的子类关系 (subclassOf)

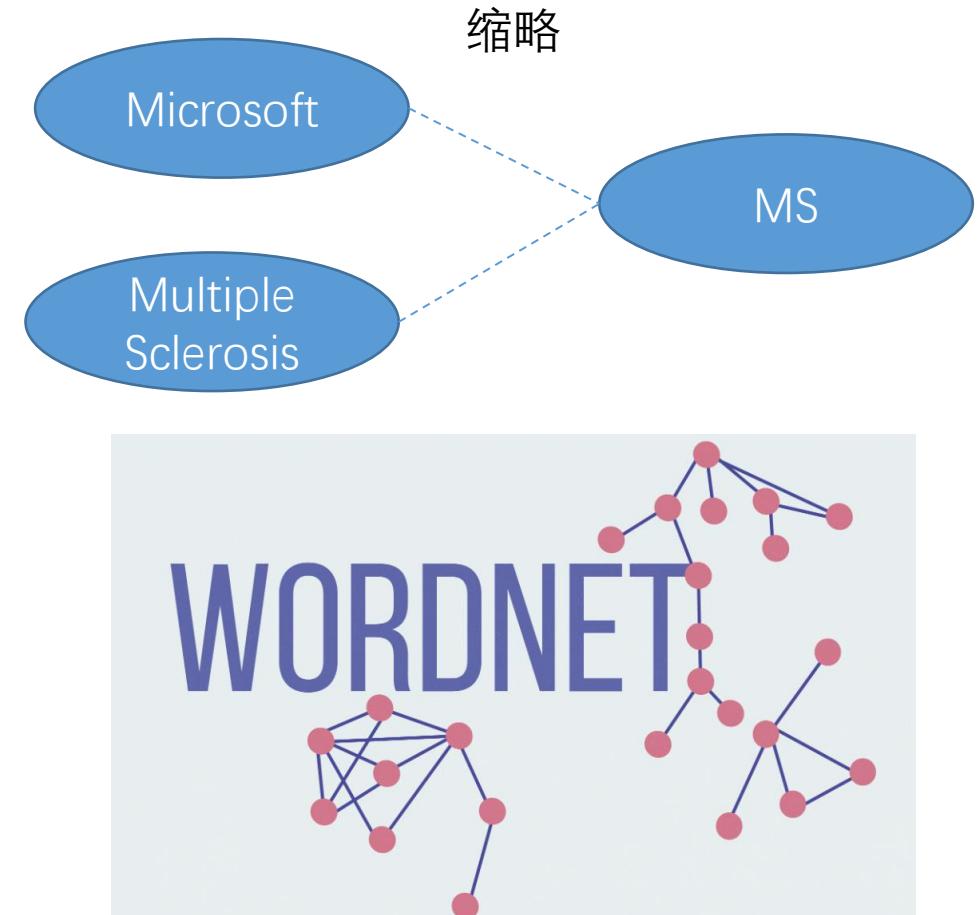
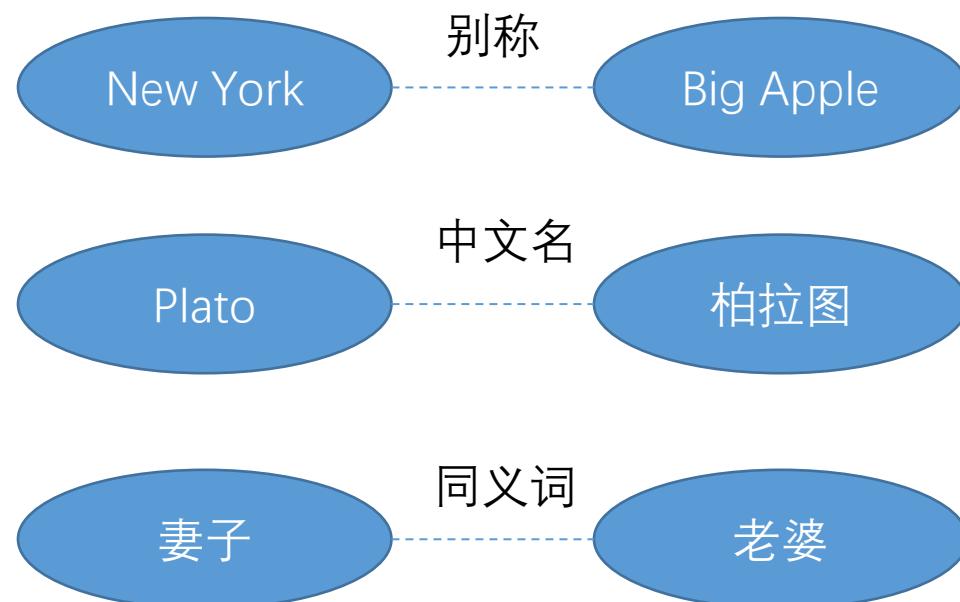


ProBase

典型概念知识图谱

知识分类-词汇知识

- 词汇知识 (Lexical Knowledge)
- 实体与词汇
- 词汇与词汇



知识类别-常识知识

- common-sense knowledge (properties):

- hasAbility (Fish, swim), hasAbility (Human, write),
- hasShape (Apple, round), hasProperty (Apple, juicy),
- hasMaxHeight (Human, 2.5 m)

- common-sense knowledge (rules):

- $\forall x: \text{human}(x) \Rightarrow \text{male}(x) \vee \text{female}(x)$
- $\forall x: (\text{male}(x) \Rightarrow \neg \text{female}(x)) \wedge (\text{female}(x) \Rightarrow \neg \text{male}(x))$
- $\forall x: \text{human}(x) \Rightarrow (\exists y: \text{mother}(x,y) \wedge \exists z: \text{father}(x,z))$

知识图谱的领域特性

- 通用知识图谱
(General-purpose Knowledge Graph)
- 领域知识图谱(Domain-specific Knowledge Graph), 有时又称为行业知识图谱
- 企业知识图谱
(enterprise knowledge graph)

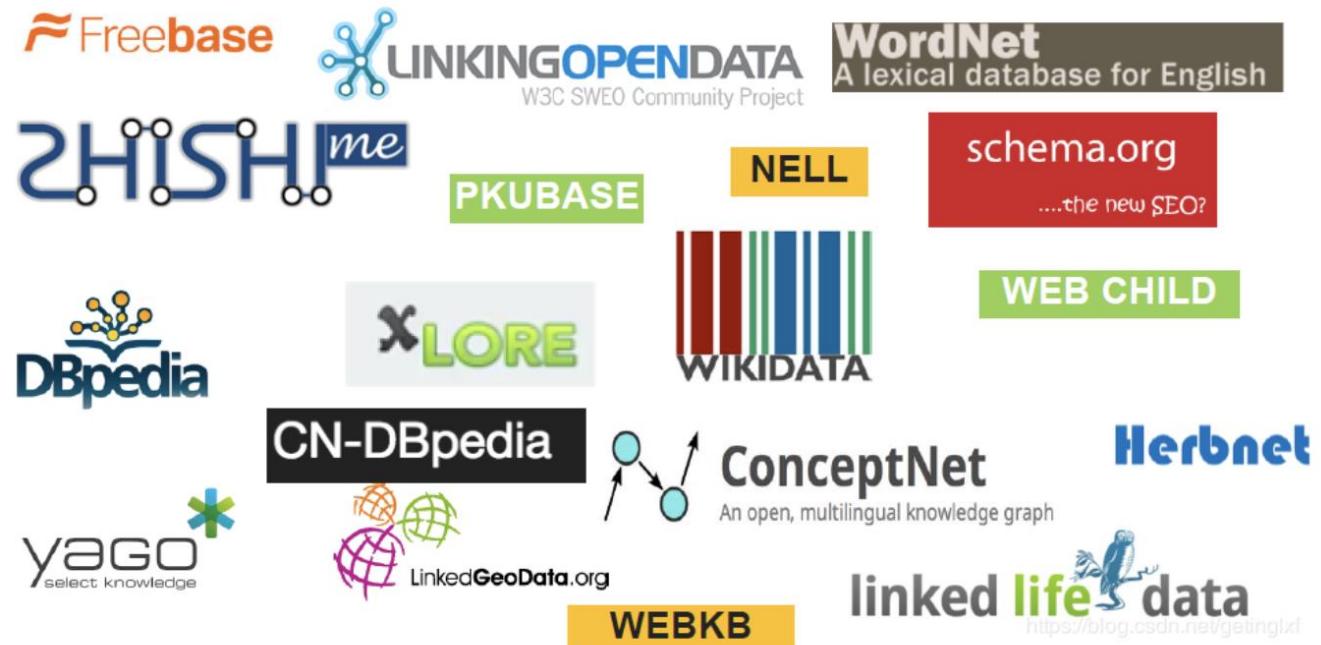
表 1-1 DKG 与 GKG 的区别

		DKG	GKG
知识表示	广度	窄	宽
	深度	深	浅
	粒度	细	粗
知识获取	质量要求	苛刻	高
	专家参与	重度	轻度
	自动化程度	低	高
知识应用	推理链条	长	短
	应用复杂性	复杂	简单

通用知识图谱 (GKG)

- 通用知识图谱 (General-purpose Knowledge Graph)

- 涵盖广、层级浅
- 质量要求并非苛刻



常见通用知识图谱

领域知识图谱的表示

- 广度（更窄）
- 深度（更深）
 - 电商领域：“韩版夏装连衣裙”vs.“连衣裙”
 - 娱乐领域：“内地鼻子长得帅的男明星”vs.“男明星”
- 粒度（更细）
 - 企业图谱：文档级别
 - 司法领域：条款级别
 - 教育领域：知识点级别

领域知识图谱的获取

- 对质量要求苛刻
- 专家参与度高
 - 自动化构建
 - 人工知识验证

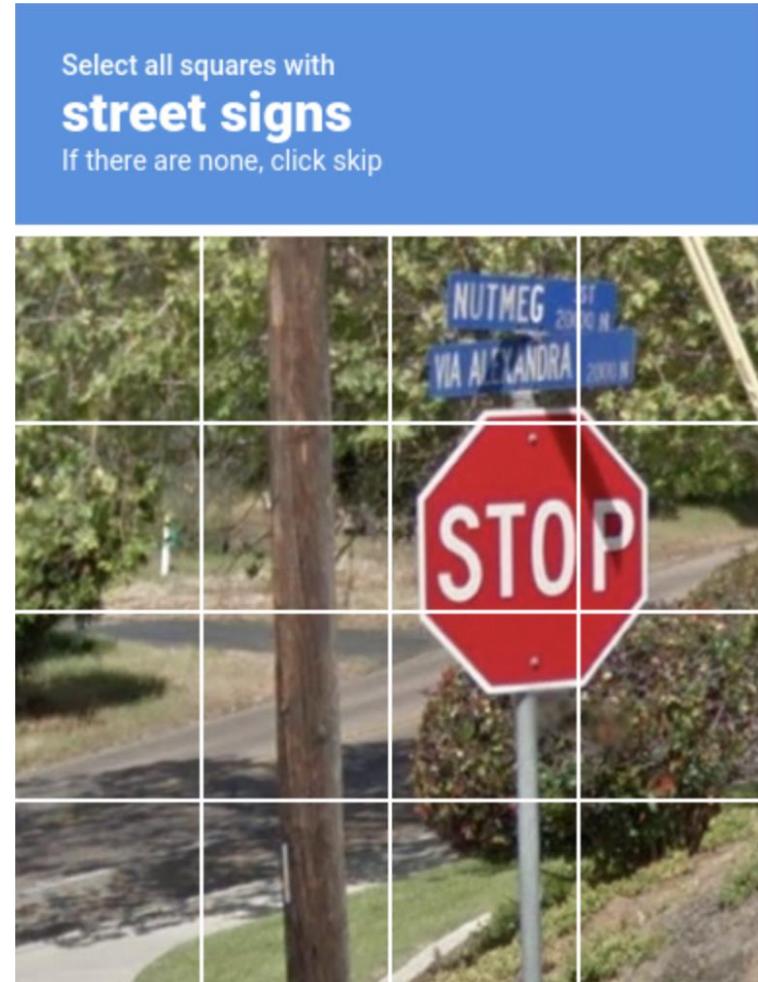
案例1: 基于知识问答验证码的知识获取

- 复旦大学知识工场实验室提供知识验证码服务,通过众包的方式对现有知识进行验证

请通过验证

请点击下文中该问题答案的任意部分: 下大坪村的面积是多少? 太难了, 换一个
下大坪村隶属于云南省大理鹤庆县黄坪镇均华村委会, 该村国土面积0.92平方公里, 海拔1500米, 年平均气温20℃, 年降水量700毫米, 农民收入主要以种植业为主。

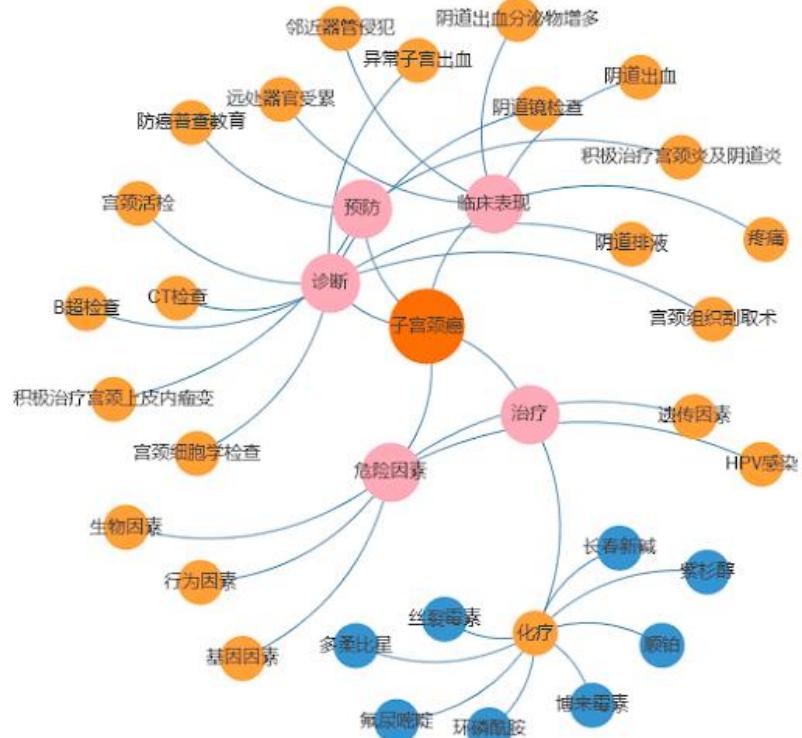
登录!



众包验证码

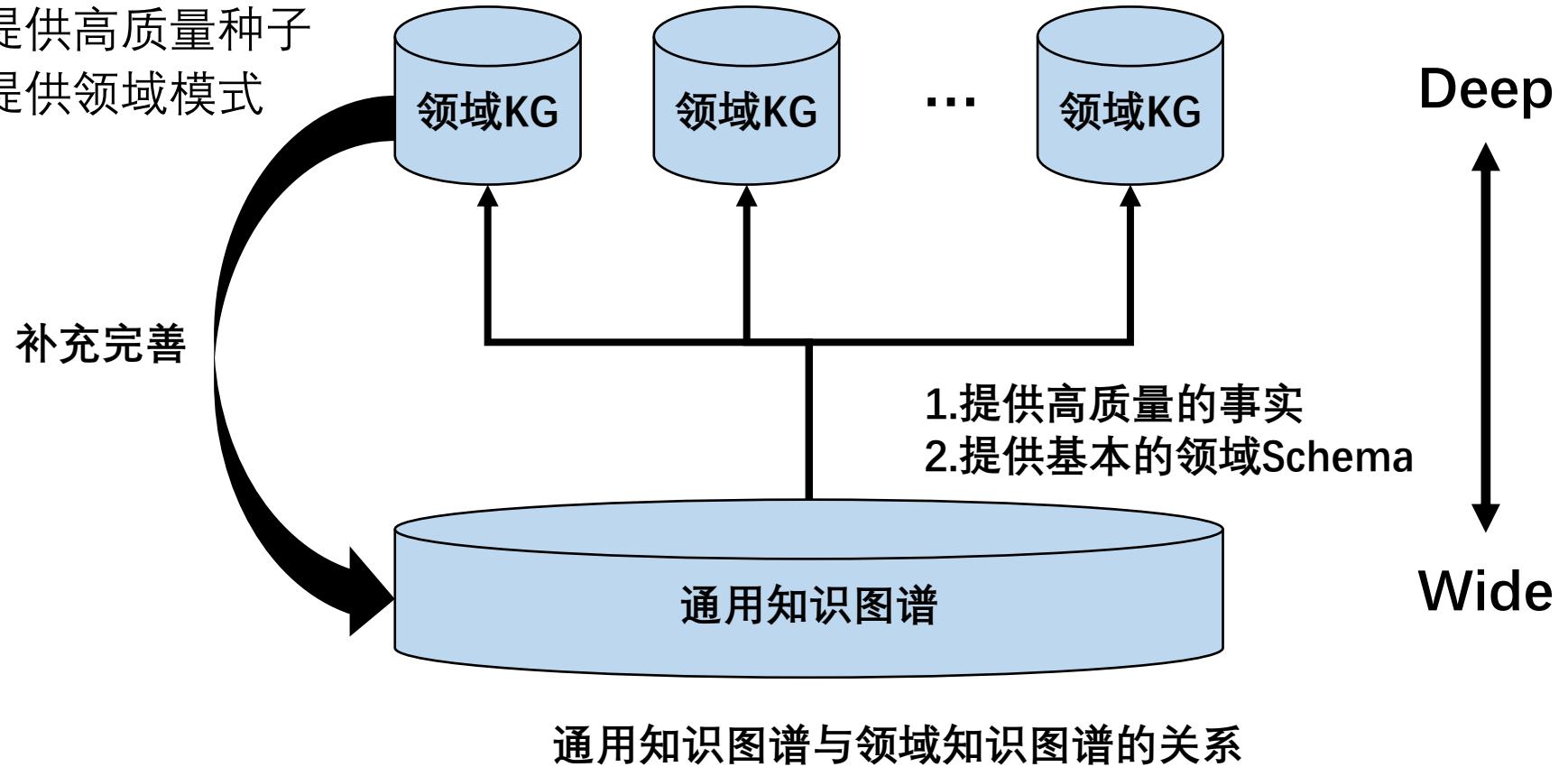
领域知识图谱的应用

- 深度推理
 - 相对稠密的DKG上，长距离推理结果仍有意义
 - 稀疏的GKG上，多步推理易导致语义漂移（Semantic Drift）
- 复杂查询
 - 公共安全领域对重点监控人群查询稠密子图
 - GKG中多为一到两步的邻居查询



DKG与GKG之间的联系

- GKG与DKG相互支撑
 - GKG为DKG提供高质量种子
 - GKG为DKG提供领域模式



知识图谱分类

- 自动化程度
- 数据来源结构化程度
- 跨语言
- 通用/specific

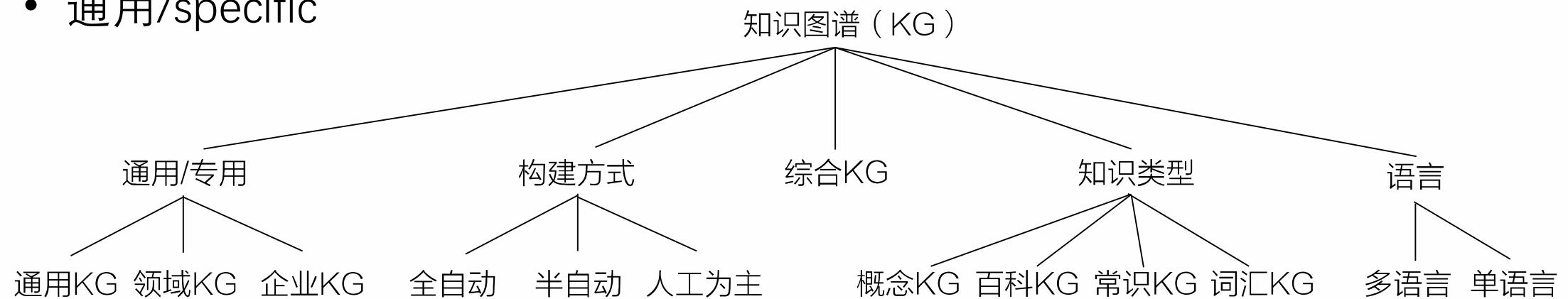
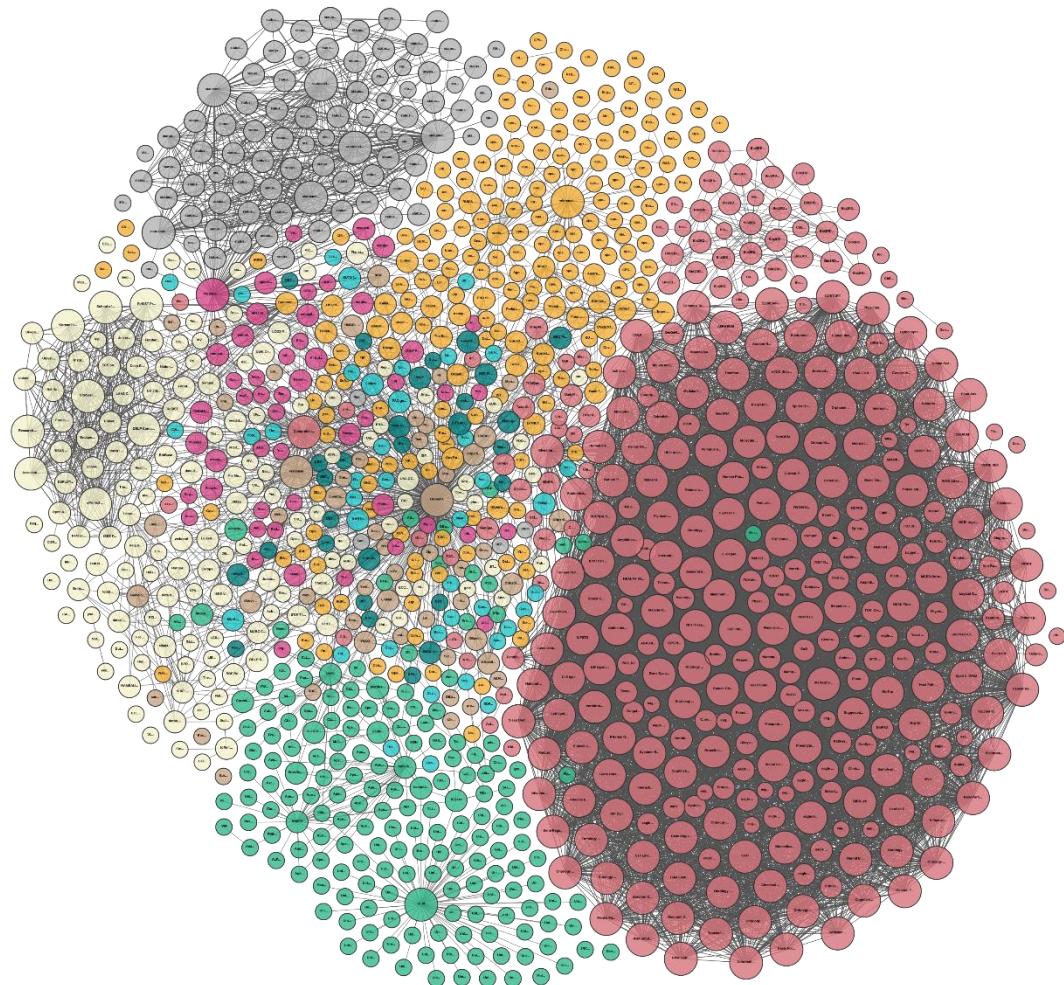


图 1-9 知识图谱的分类体系

典型知识图谱-越来越多的知识图谱应运而生

Yago, WordNet, FreeBase, Probbase, NELL, CYC, DBpedia…

时间	知识图谱数量
2017-03-16	1,139
2014-08-30	570
2011-09-19	295
2010-09-22	203
2009-07-14	95
2008-09-18	45
2007-11-07	28
2007-05-01	12



"Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul
Buitelaar, Anja Jentzsch and Richard Cyganiak. <http://lod-cloud.net/>"

2022/2/25

第1章：知识图谱概述

75

Cyc

- 简介
 - 常识知识图谱
- 样例
 - (#\$isa #\$BillClinton #\$UnitedStatesPresident)
 - "Bill Clinton belongs to the collection of U.S. presidents"
- 特点
 - 通过人工方法将上百万条人类常识编码成机器可用的形式，用以进行智能推断
- 规模
 - 目前ResearchCyc知识图谱中包含了700 万条断言（事实和规则），涉及63 万个概念，38000 种关系

<http://www.cyc.com/>

WordNet

- 简介
 - 基于认知语言学的英语词典
- 样例
 - S: (n) **car**, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) "*he needs a car to get to work*"
- 特点
 - 以同义词集合（synset）作为一个基本单元
- 规模

POS	<i>Unique Strings</i>	<i>Synsets</i>	<i>Total</i>
			<i>Word-Sense Pairs</i>
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

[George A Miller. 1995]

<https://wordnet.princeton.edu/>

ConceptNet

- 简介
 - 大型的多语言常识知识库
- 样例
 - “刘德华”
- 特点
 - 知识来源丰富
 - 众包(Crowd-Sourcing)
 - 资源 (例如Wiktionary 和Open Mind Common Sense)
 - 带目的的游戏 (如Verbosity 和nadya.jp)
 - 专家创建的资源(如WordNet 和JMDict)

The screenshot displays the ConceptNet API endpoint for the Chinese term '劉德華' (Andy Lau). The top navigation bar includes links for Documentation, FAQ, Chat, and Blog. The main content area is organized into several sections:

- 劉德華** (zh) **劉德華**
A Chinese term in ConceptNet 5.6
Sources: the PTT Pet Game and CC-CEDICT 2017-10
View this term in the API
- 劉德華 wants...**
 - zh 唱歌 →
 - zh 有小孩 →
 - zh 演唱會 →
 - zh 賺錢 →
 - zh 跳舞 →
 - zh 養小孩 →
- 劉德華 doesn't want...**
 - zh 狗仔 →
 - zh 不紅 →
 - zh 失眠 →
 - zh 蒼蠅 →
 - zh 被罵 →
- Things with 劉德華**
 - zh 手 →
 - zh 老婆 →
 - zh 腳 →
 - zh 臉蛋 →
 - zh 錢 →
- Subevents of 劉德華**
 - zh 帥大叔 →
 - zh 演員 →
 - zh 香港偶像 →
- 劉德華 is a type of...**
 - zh 四大天王 →
 - zh 天王 →
 - zh 華人 →
- Effects of 劉德華**
 - zh 如痴如醉 →
 - zh 歌曲 →
 - zh 歡樂 →
- Synonyms**
 - en andy lau →
 - zh 刘德华 →
- Location of 劉德華**
 - zh 一線男星 →

<http://conceptnet.io/>

2022/2/25

第 1 章：知识图谱概述

[Robert Speer et al. 2012]

78

GeoNames

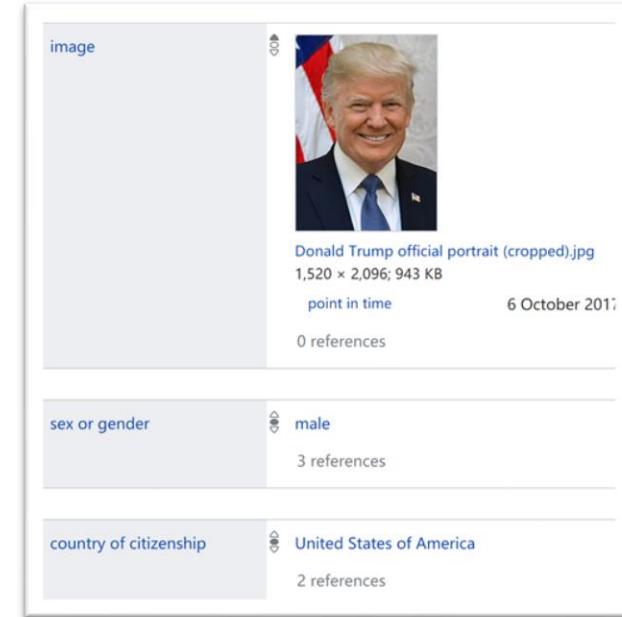
- 简介
 - 全球地理数据库
- 样例
 - “中国”
- 特点
 - 多语言地理位置信息
- 统计
 - 它包含了将近200 种语言的1000 万个地理信息，包括位置的经纬度、行政区划、邮政编码、人口、海拔和时区等信息



<http://www.geonames.org/>

Freebase/Wikidata

- 简介
 - Freebase 所有知识采用结构化的表示形式，可由机器和人编辑
 - Wikidata是维基百科的姐妹工程，同样可由机器和人自由编辑
 - 2016年8月31日，Freebase宣布关闭，所有数据汇入Wikidata
- 样例
 - “Donald Trump”
- 特点
 - 众包构建
 - 结构化三元组
- 统计
 - Wikidata目前包含49,915,906个实体



[Bollacker et al. 2008]

DBpedia

- 简介
 - 从维基百科页面中自动抽取出结构化的知识，构建而成的大型通用百科图谱
- 样例
 - “A”

```
<http://dbpedia.org/resource/A> <http://dbpedia.org/property/name> "Latin Capital Letter A"@en .  
<http://dbpedia.org/resource/A> <http://dbpedia.org/property/name> "Latin Small Letter A"@en .  
<http://dbpedia.org/resource/A> <http://dbpedia.org/property/map> "ASCII 1"@en .
```
- 特点
 - 多语言
 - 自动构建
- 统计
 - 共收录有127 种不同语言共计2800万实体
 - 其中英文实体数量最大，为467 万

[Jens Lehmann et al., 2015]

<http://wiki.dbpedia.org/>

YAGO

- 简介
 - 采用自动的方式构建，数据来源于维基百科、WordNet 以及 GeoNames
- 样例
 - <Albert_Einstein> <isMarriedTo> <Elsa_Einstein>
- 特点
 - 每类关系的准确率都经过人工评估，达到95% 以上
 - 融合了WordNet的纯层次结构以及维基百科的标签分类体系
 - 部分事实增加了时间和空间两种维度
 - 多语言融合
- 统计
 - 1千万实体， 1.2亿事实 [Fabian, M. S. et al. 2007]

<http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/#c10444>

Open IE

- 简介
 - 互联网开放关系抽取系统，主要从句子中抽取开放关系
- 样例
 - From: “The U.S. president Barack Obama gave his speech on Tuesday and Wednesday to thousands of people.”
 - To:
 - (Barack Obama, is the president of, United States)
 - (Barack Obama, gave his speech, on Tuesday)
- 特点
 - 开放关系抽取，Never-Ending
- 统计 [Banko et al. 2007], [Etzioniet al. 2011]
 - 目前已经从十亿的互联网页面中抽取出了50 亿条关系

<http://openie.allenai.org/>

BabelNet

- 简介
 - 多语言知识图谱
- 样例
 - “周杰伦”
- 特点
 - 271 种语言
 - 自动融合
- 统计
 - 最新版为BabelNet 3.7, 共包含1400 万个实体



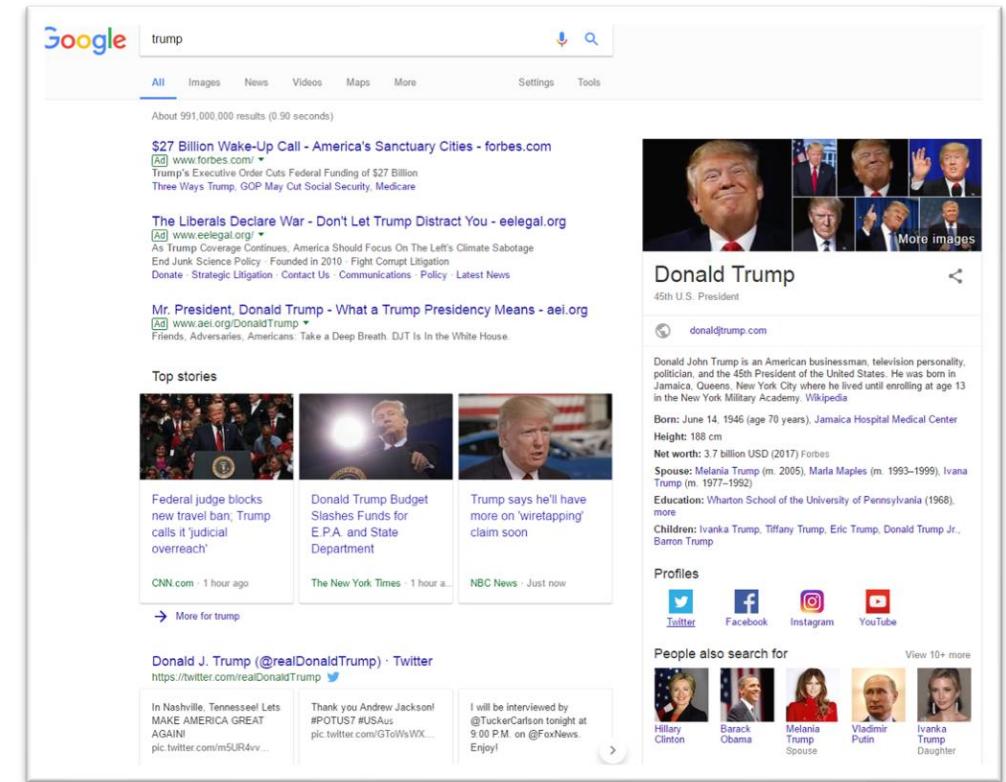
The screenshot shows the BabelNet entity page for Jay Chou (周杰伦). At the top, there is a language selector bar with Chinese, Arabic, English, French, German, Greek, Hebrew, Hindi, Italian, Japanese, and a plus sign for all preferred languages. Below the bar, there is a small icon of a person singing and the ID bn:03342151n, followed by NOUN, Named Entity, and categories: 1979年出生, 世界音乐奖获得者, 全球华语歌曲排行榜最受欢迎男歌手, 十大中文金曲奖全国最受欢迎男歌手... The main title is "ZH 周杰伦" with three pronunciation icons. Below the title, a brief description states: 周杰伦 (英语: Jay Chou; 1979年1月18日 -) 台湾的流行歌曲男歌手、音乐家、唱片制片人、演员、导演、电竞团队队长兼老板。 There are links to Wikipedia and more definitions. To the right, there is a grid of entity properties and values, such as IS A (人), COUNTRY OF CITIZENSHIP (台湾), DISCOGRAPHY (周杰伦音乐作品列表), GIVEN NAME (杰), INSTRUMENT (钢琴), OCCUPATION (演员, 作者, 电影导演), PLACE OF BIRTH (台北市), SPOUSE (昆凌), and VOICE TYPE (男高音). A "Less relations" link is also present.

<http://babelnet.org/>

[Roberto Navigli et. al., 2012]

Google KG

- 简介
 - 谷歌知识图谱于2012 年发布，被认为搜索引擎的一次重大革新
- 样例
 - “Donald Trump”
- 特点
 - 规模巨大
 - 用于增强搜索引擎的搜索能力
- 统计
 - 5700万实体， 180亿关系



Probase

- 简介
 - 概念图谱，数据源来自微软搜索引擎Bing 的网页，主要利用Hearst Pattern 从文本中抽取IsA 关系
- 样例
 - From: “... in tropical countries such as Singapore, Malaysia, ...”
 - To:
 - (Singapore, isA, tropical countries)
 - (Malaysia, isA, tropical countries)
- 特点
 - 概念规模最大
 - 自动构建
- 统计
 - 1200万实体， 540万概念

ID	Pattern
1	$NP \text{ such as } \{NP,\}^* \{(or and)\} NP$
2	$\text{such } NP \text{ as } \{NP,\}^* \{(or and)\} NP$
3	$NP\{,\} \text{ including } \{NP,\}^* \{(or and)\} NP$
4	$NP\{,NP\}^* \{,\} \text{ and other } NP$
5	$NP\{,NP\}^* \{,\} \text{ or other } NP$
6	$NP\{,\} \text{ especially } \{NP,\}^* \{(or and)\} NP$

[Wu et al. 2012]

搜狗知立方/百度知心

• 搜狗知立方

• 简介

- 中文知识图谱，应用于搜狗搜索引擎

• 特点

- 侧重于娱乐领域

范冰冰的身高

范冰冰身高
168cm

范冰冰，1981年9月16日生于山东青岛，电影演员、歌手，毕业于上海师范大学谢晋影视艺术学院。1996年参演电视剧《女强人》。1998年主演电视剧《还... 详情>>

男友 李晨 180cm 前男友 王学兵 180cm 纬闻 陆毅 182cm 荧幕情侣 李治廷 175cm 搭档 林心如 167cm

• 百度知心

• 简介

- 中文知识图谱，应用于百度搜索引擎

• 特点

- 融合百度百科知识

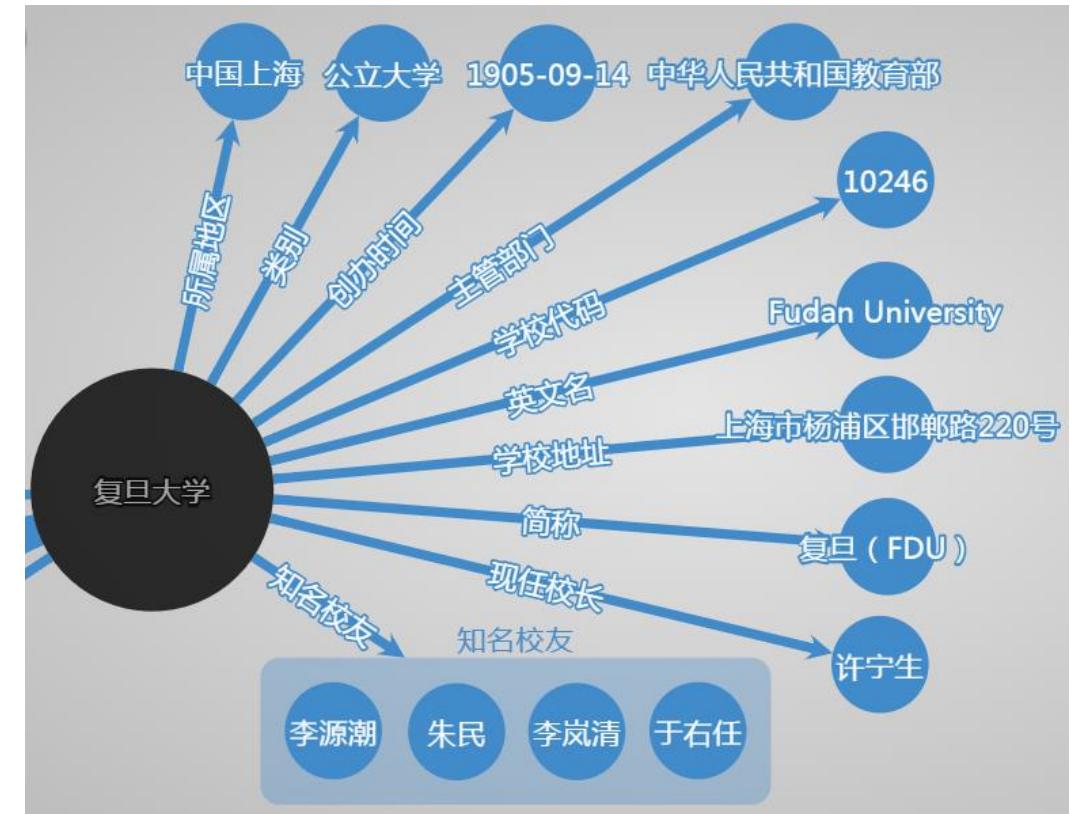
刘德华的出生日期

刘德华生日：
1961年9月27日(天秤座)

刘德华 (Andy Lau)，1961年9月27日出生于中国香港，演员、歌手、作词人、制片人。1981年出演电影处女作《彩云曲》。1983年主演的武侠剧《神雕侠侣》在香港获得62点的收... 详情>>

CN-DBpedia

- 简介
 - 由复旦大学知识工场实验室构建
 - 融合通用百科和领域百科数据
- 样例
 - “复旦大学”
- 特点
 - 实时更新
 - 完整的数据/服务接口
- 统计
 - 1600万实体，2亿关系



[Bo Xu et al., 2017]

reference

- [George A Miller. 1995] Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41, 1995.
- [Robert Speer et al. 2012] Representing general relational knowledge in conceptnet 5. In LREC, pages 3679–3686, 2012.
- [Jens Lehmann et al., 2015] DBpedia: A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia.
- [Fabian, M. S. et al. 2007] Yago: A core of semantic knowledge unifying wordnet and wikipedia
- [Bo Xu et al., 2017] CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System
- [Roberto Navigli et. al., 2012] BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network
- [Etzioniet al. 2011] "Open information extraction: The second generation." IJCAI. Vol. 11. 2011.

- [Wu et al. 2012] "Probase: A probabilistic taxonomy for text understanding." Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012.
- [Banko et al. 2007] "Open information extraction from the web." IJCAI. Vol. 7. 2007.
- [Newell, Allen et al. 1976] "Computer Science as Empirical Inquiry: Symbols and Search", Communications of the ACM, 19 (3)
- [Dreyfus, Hubert 1979] What Computers Still Can't Do, New York: MIT Press.
- [陈文伟 et. AI] 知识工程与知识管理
- [Yin, et al. 2017] Truth Discovery with Multiple Conflicting Information Providers on the Web, kdd07
- [Wanyun Cui et al. 2017] KBQA: Learning Question Answering over QA Corpora and Knowledge Bases, (VLDB 2017)
- [Yi Zhang, et al, 2017] Entity suggestion with conceptual explanation, (IJCAI 2017)
- [Bo Xu, et al, 2016] Learning Defining Features for Categories. (**IJCAI 2016**)