《知识图谱: 概念与技术》

第 1 讲
知识图谱概述

肖仰华

复旦大学

shawyh@fudan.edu.cn

本章大纲

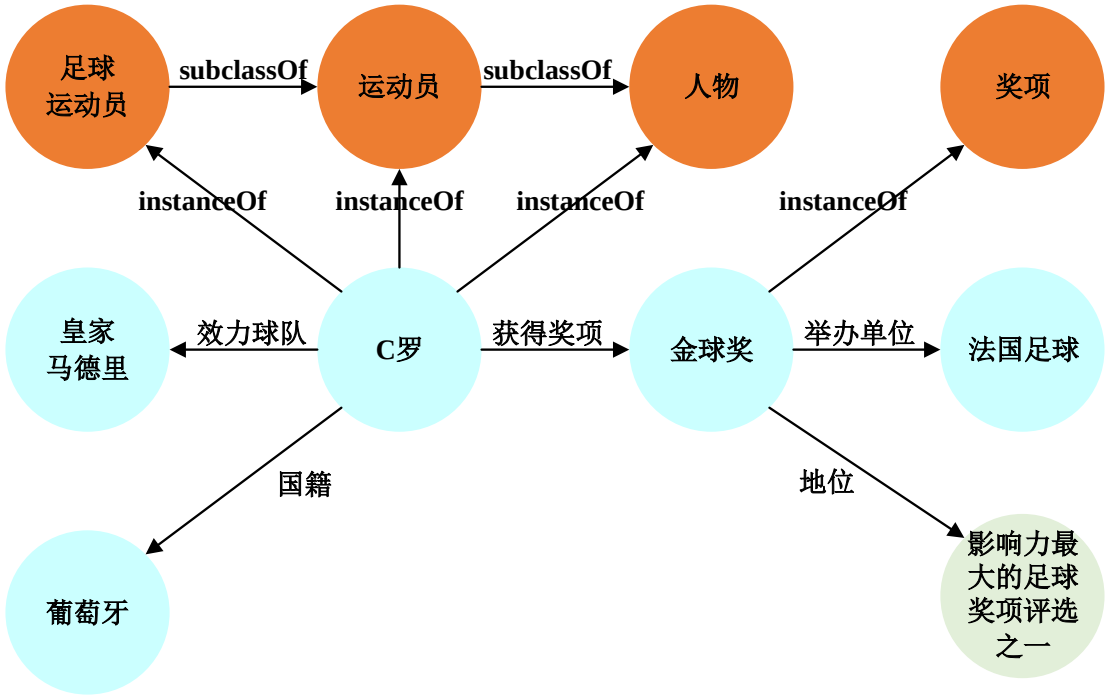
- 知识图谱概念
- 知识图谱内涵
- 知识图谱源起
- 知识图谱优势
- 知识分类
- 典型知识图谱
- 知识图谱价值
- 知识图谱应用

知识图谱概念

知识图谱

- 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
- 作为一种技术体系，是大数据时代**知识工程**的代表性进展



知识图谱示例子。知识图谱富含实体、概念、属性、关系等信息

领域知识图谱

- 领域（行业）知识图谱 (Domain-specific Knowledge Graph)
  - 聚焦于特定领域或者行业的知识图谱
- 企业知识图谱(Enterprise knowledge graph)
  - 贯穿企业各业务部门的知识图谱

各类领域知识图谱

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.

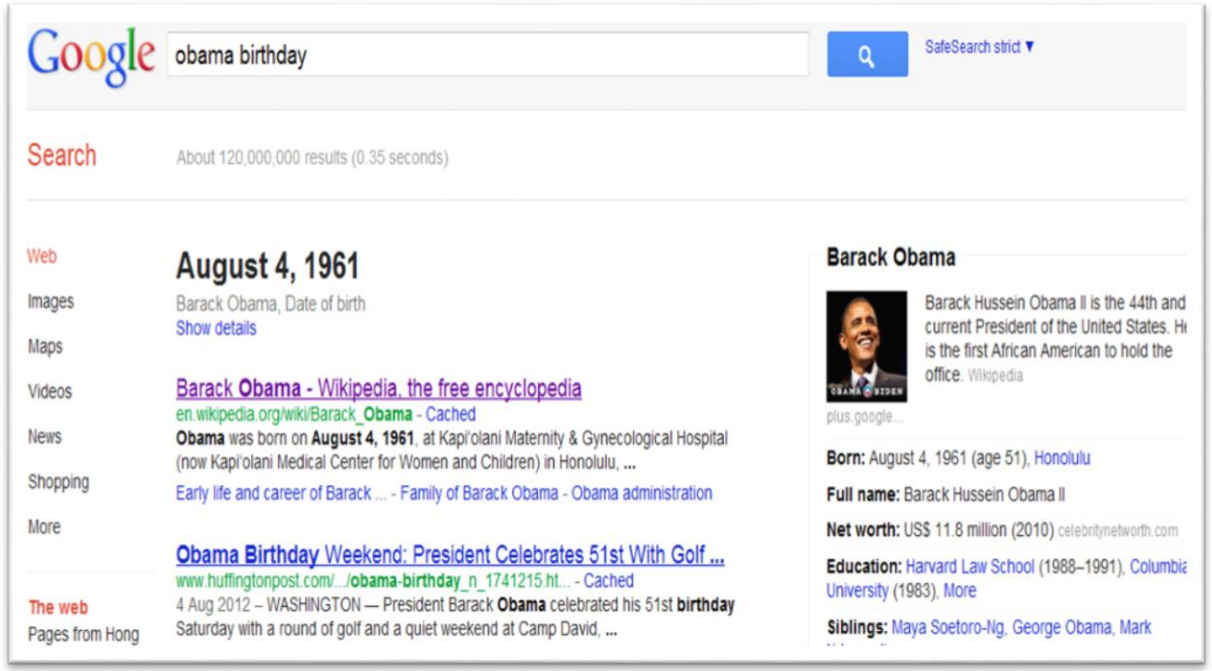
KG (Knowledge graph) is a large scale ***semantic network*** consisting of entities/concepts as well as the semantic relationships among them

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

知识图谱内涵

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

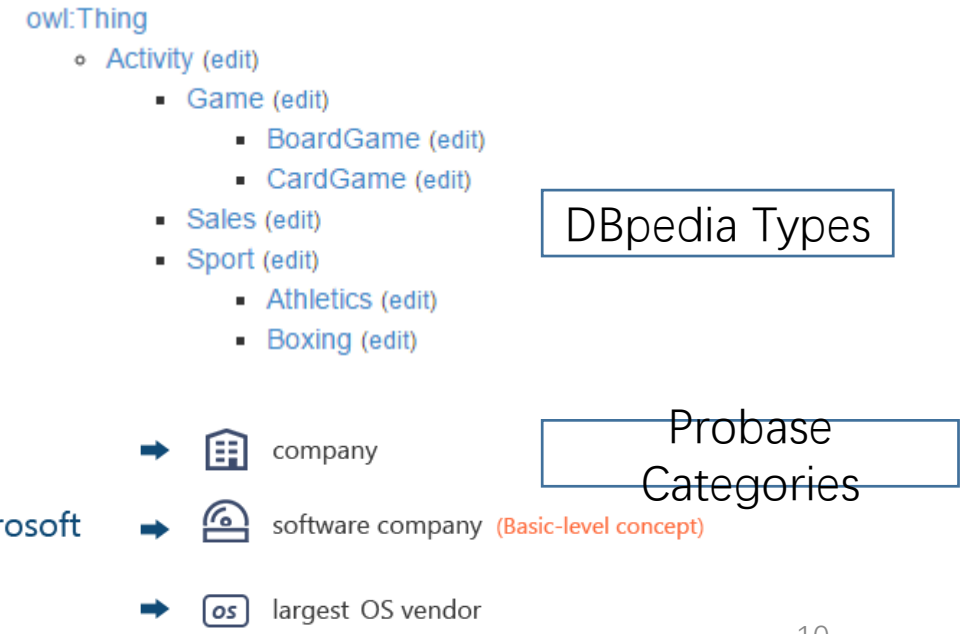
KG组成- Node-Concept

- 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:

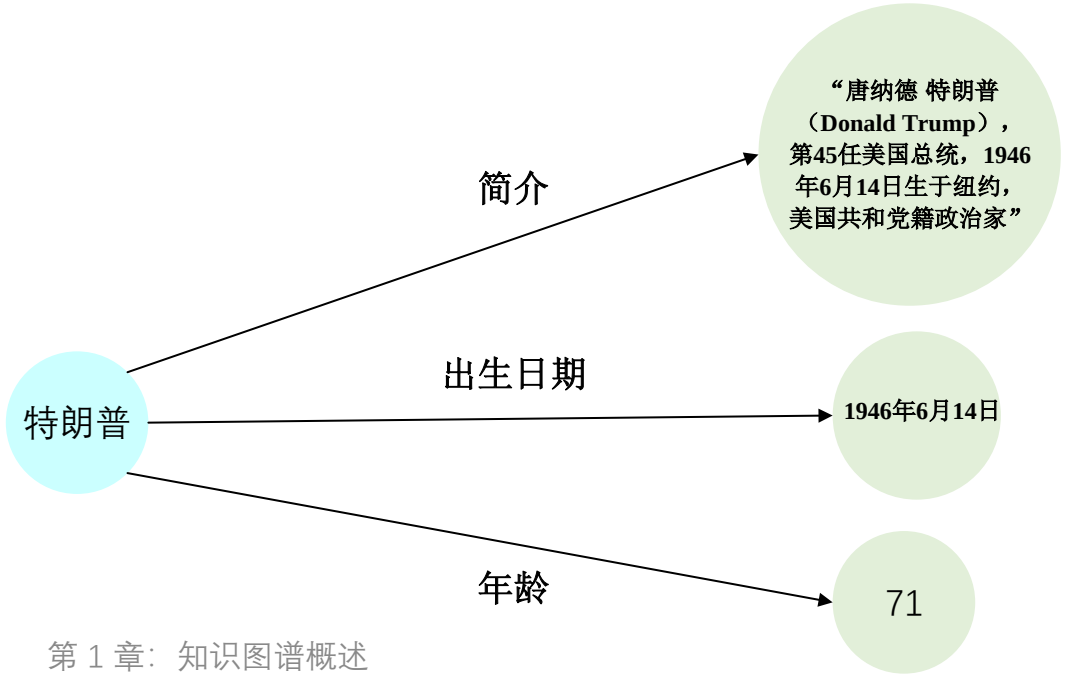
1、the process of formation of categories;

2、the process of identifying X as a member of a particular category Y;



KG组成- Node-Value

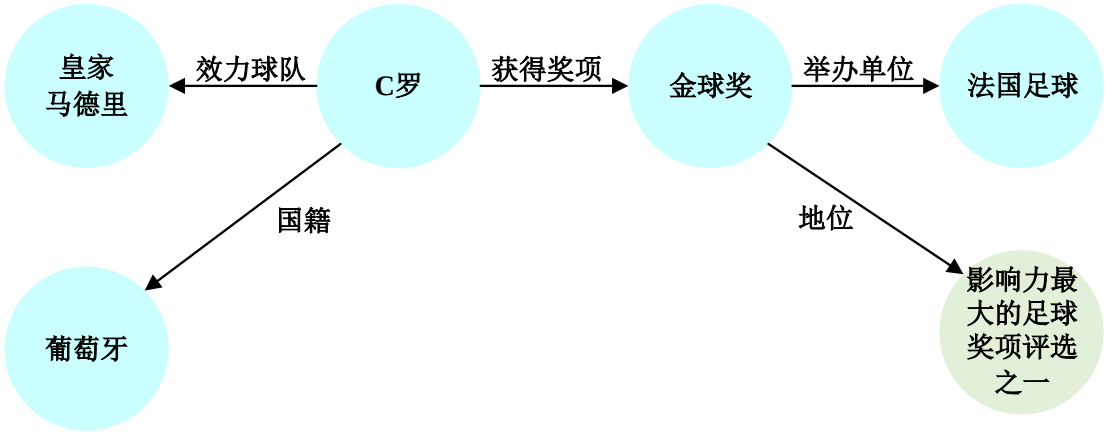
- Date
  - 特朗普 出生日期 1946年6月14日
- String
  - 特朗普 简介 “唐纳德·特朗普（Donald Trump），第45任美国总统，1946年6月14日生于纽约，美国共和党籍政治家”
- Numeric
  - 特朗普 年龄 71



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



知识图谱源起

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
- 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.
- GOFAI（“good old fashioned artificial intelligence”, proposed by John Haugeland）
  - Focused on these kind of high level symbols，such as <dog> and <tail>

Newell

Simon

AI System=Knowledge + Reasoning

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

传统KE-代表性人物与成就

Minsky（1969年图灵奖）
感知机，框架知识表示

Newell & Simon （1975年图灵奖）
形式化语言，通用问题求解

Judea Pearl （2011年图灵奖）
概率图模型之父

McCarthy（1971年图灵奖）
LISP语言，Advice Taker系统

Feigenbaum（1994年图灵奖）
知识工程提出者

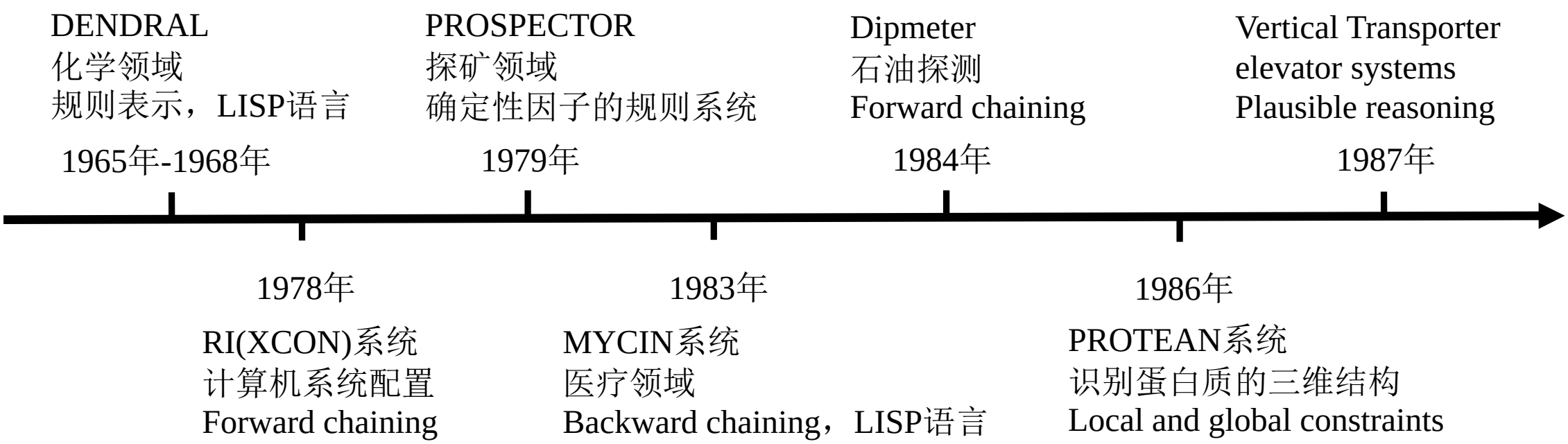
Berners-Lee（2016年图灵奖）
语义网

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

[陈文伟 et. Al]

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

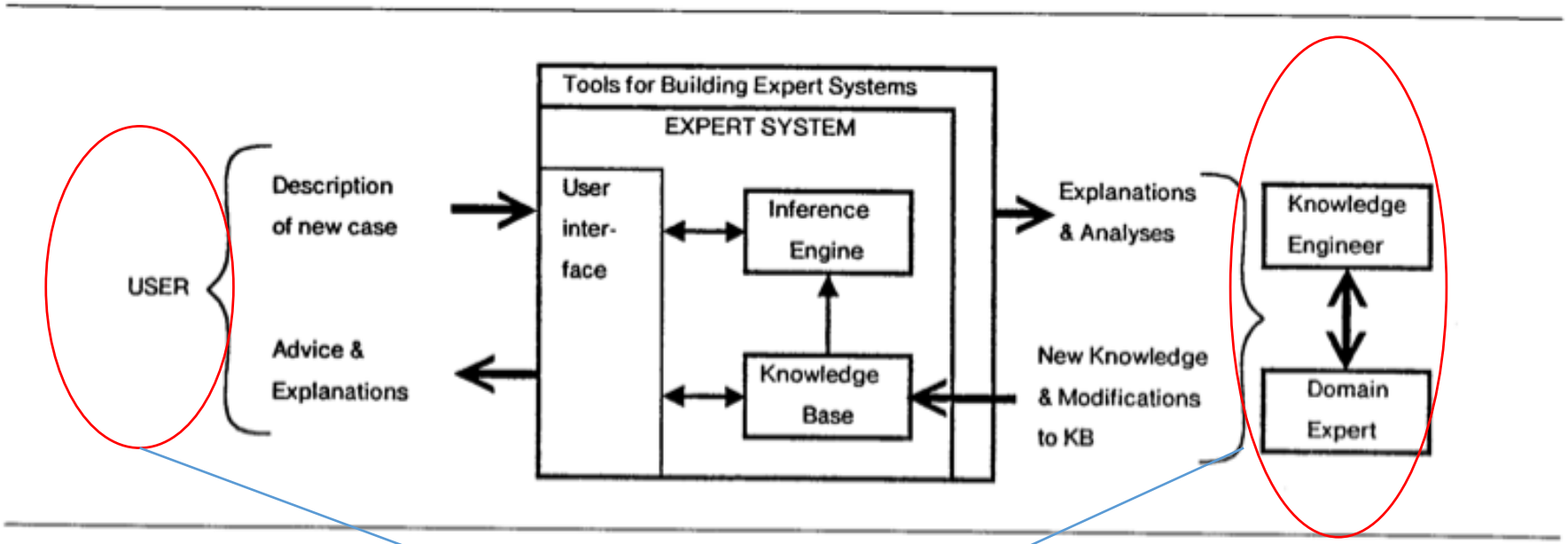


FIGURE 1-2 Interaction of a knowledge engineer and domain expert with software tools that aid in building an expert system. Arrows indicate information flow.

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。

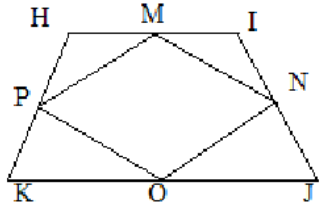


图 5-1 例 1 的图形

(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))

基于规则系统的高中几何自动解题过程

(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)

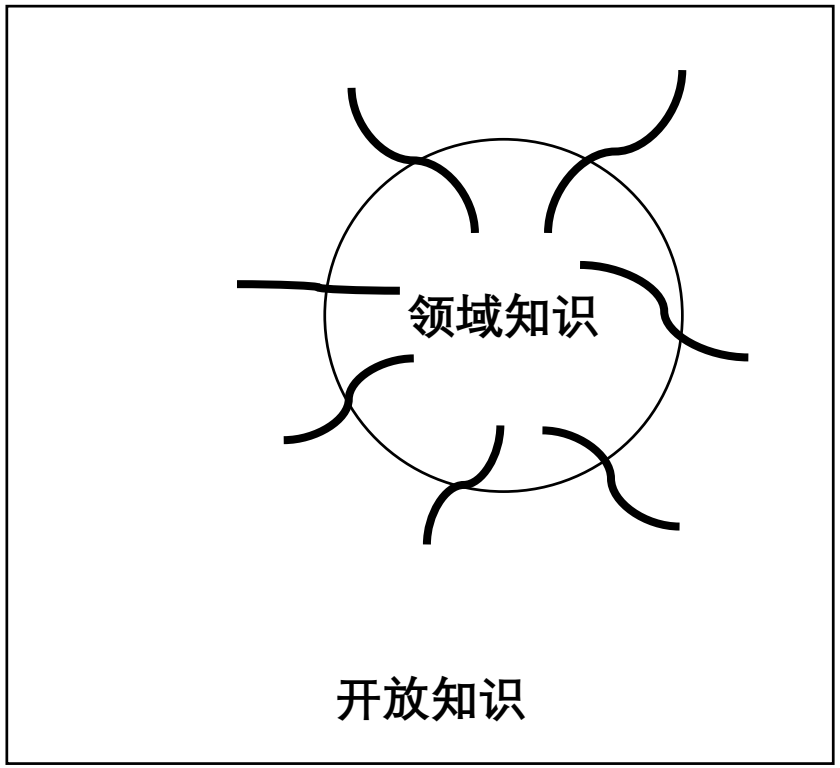
```
rule "ParallelogramToLineParallel"
when
    $pr : ParallelogramRelation(
        $A : quadrangle.getVertexA(),
        $B : quadrangle.getVertexB(),
        $C : quadrangle.getVertexC(),
        $D : quadrangle.getVertexD())
then
    Line AB = new Line($A, $B);
    Line BC = new Line($B, $C);
    Line CD = new Line($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
```

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”

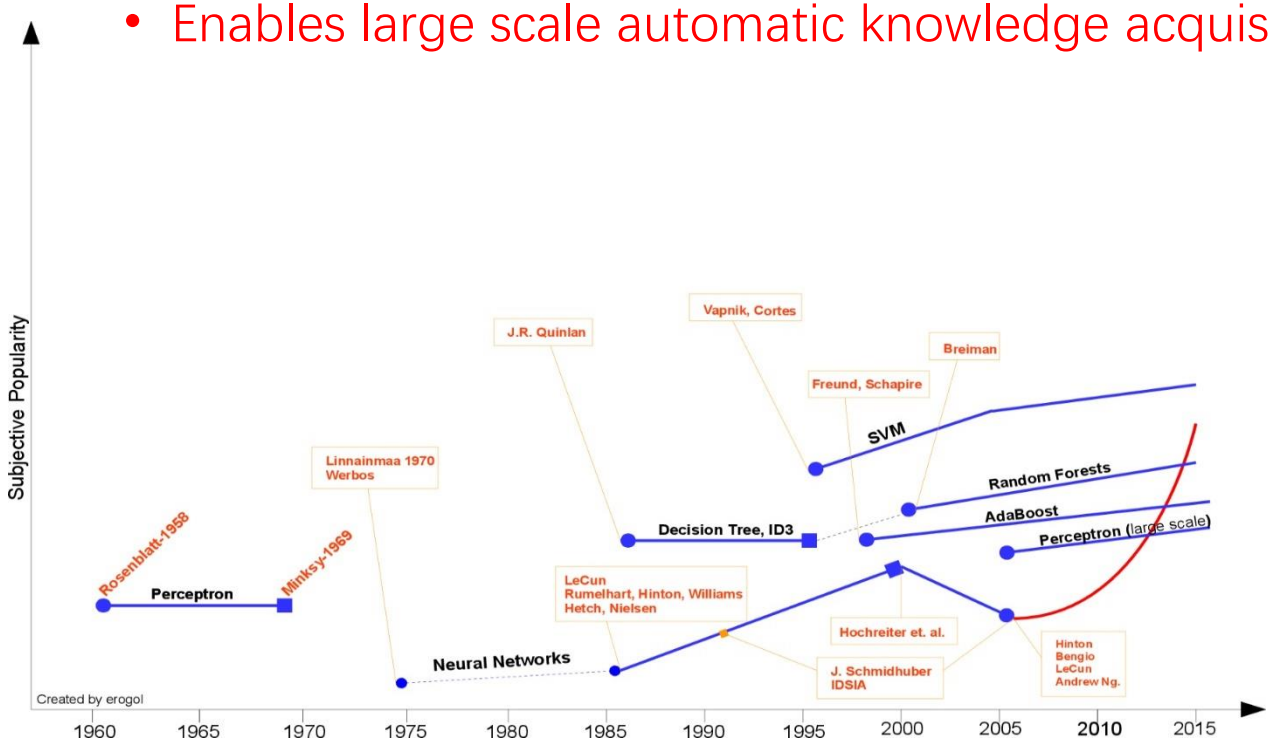
互联网上的搜索关键字具有开放性、规模巨大等特点

- **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**

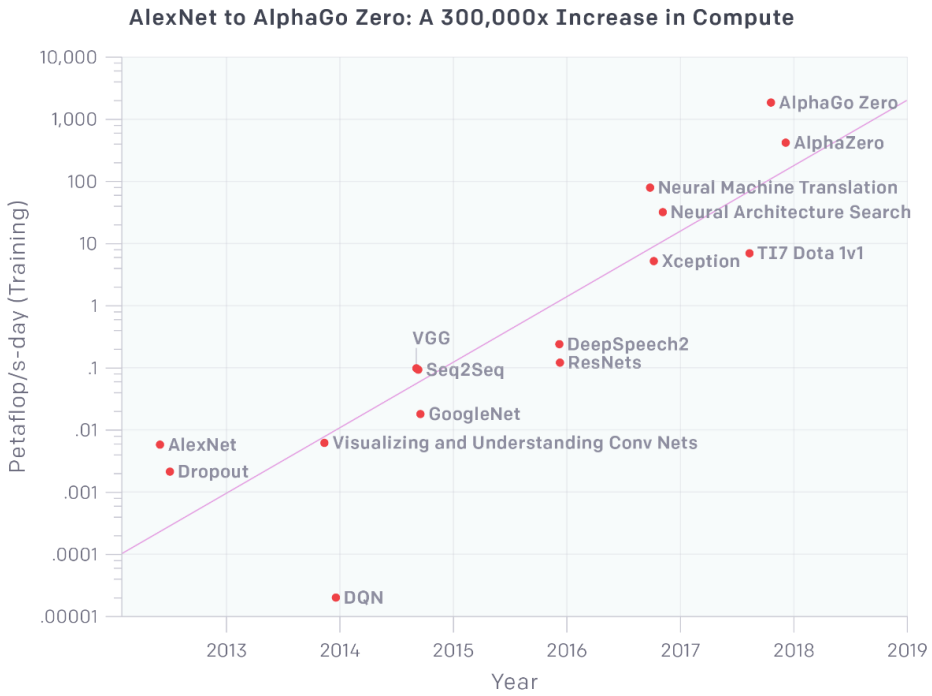
2012年，谷歌推出其知识图谱已满足搜索中知识应用需求

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……

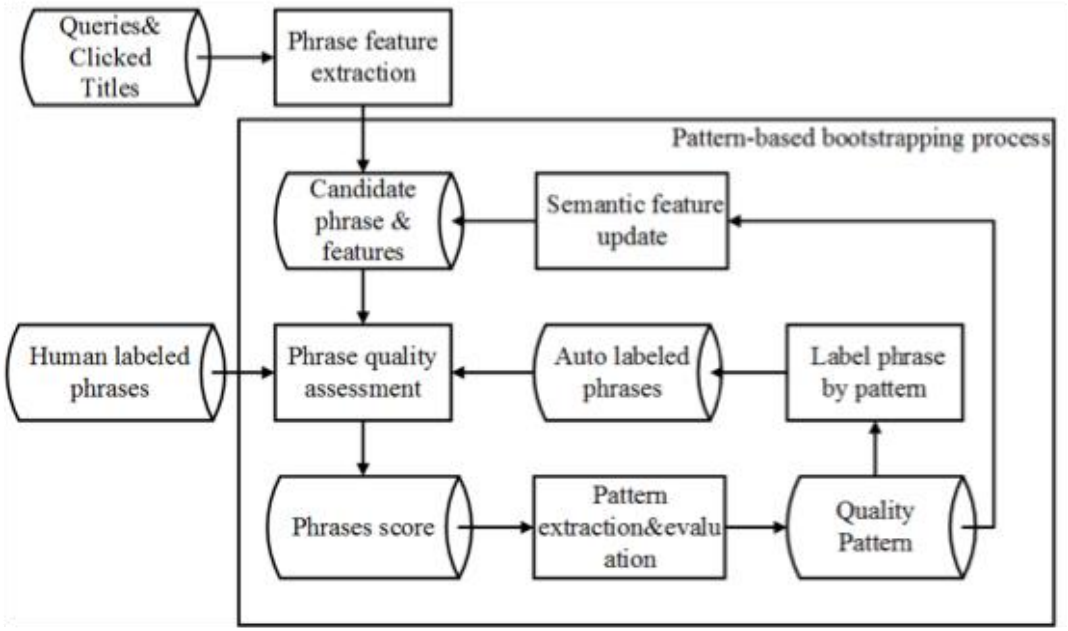


Fig. 1. Pattern-based bootstrapping framework.

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 NP 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

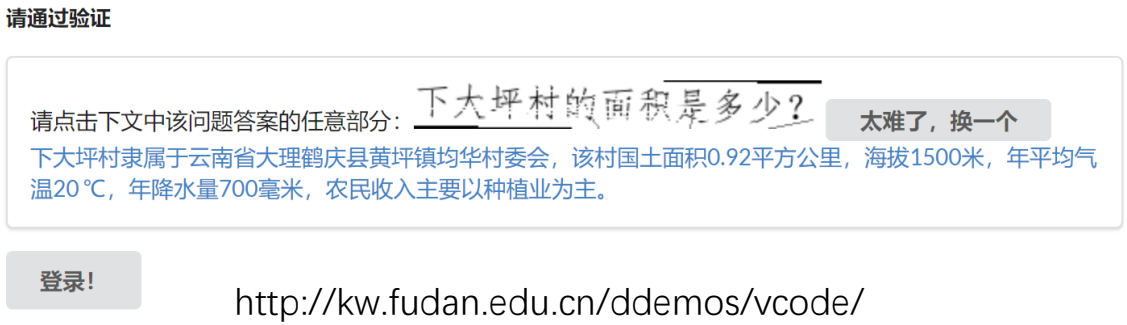
Ref: Mining High-quality Phrase from Query logs in E-commerce, Under review

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

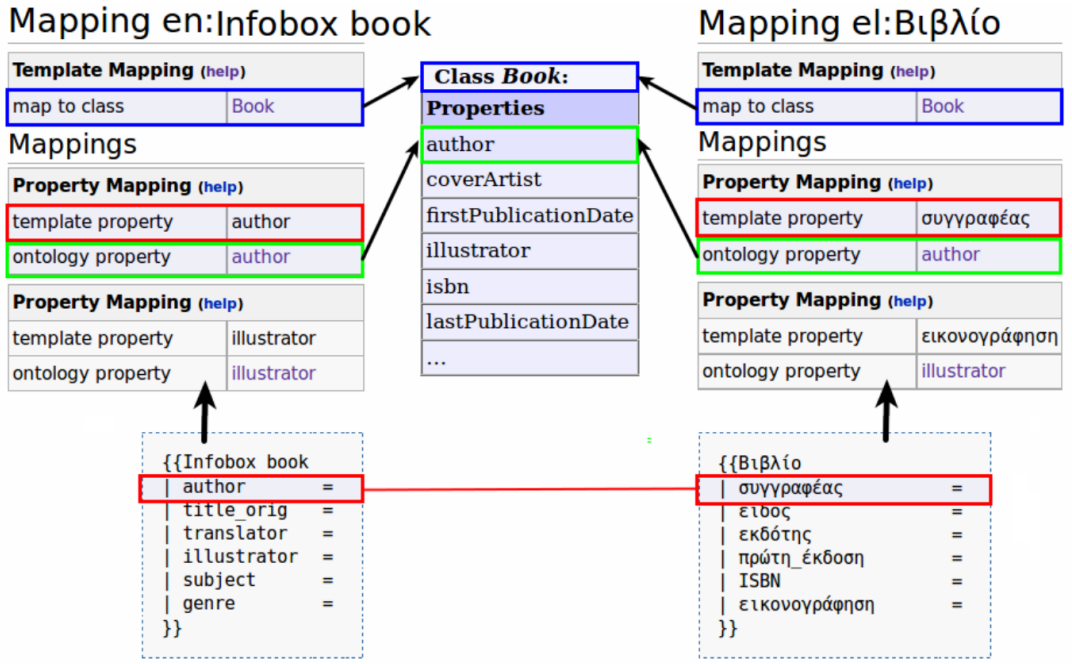
- 复旦大学知识工场实验室提供知识验证码服务，通过众包的方式对现有知识进行验证



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

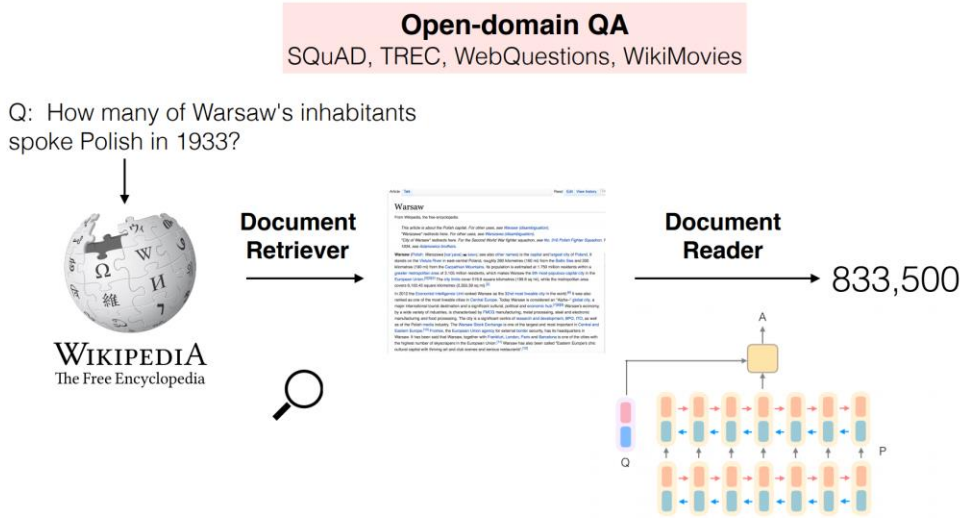
案例2:基于众包的Taxonomy构建

- DBpedia通过众包方式构建了DBpedia Ontology



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



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

周杰伦 共被编辑4812次

版本对比	更新时间	全部版本	贡献者	修改原因	区块链信息
	2018-06-06 03:36	查看	w_ou	内链修复	查看
	2018-03-11 16:14	查看	海澜～～～天倥	内容扩充 内链	查看
	2018-03-01 20:20	查看	爱锦瑟的年华	图片	查看
	2018-02-28 18:59	查看	爱锦瑟的年华	内容扩充 参考资料	查看
	2018-02-15 08:05	查看	紫雪510	内容扩充 参考资料	查看
	2018-02-11 20:54	查看	5ssax	更正错误 图片	查看
	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

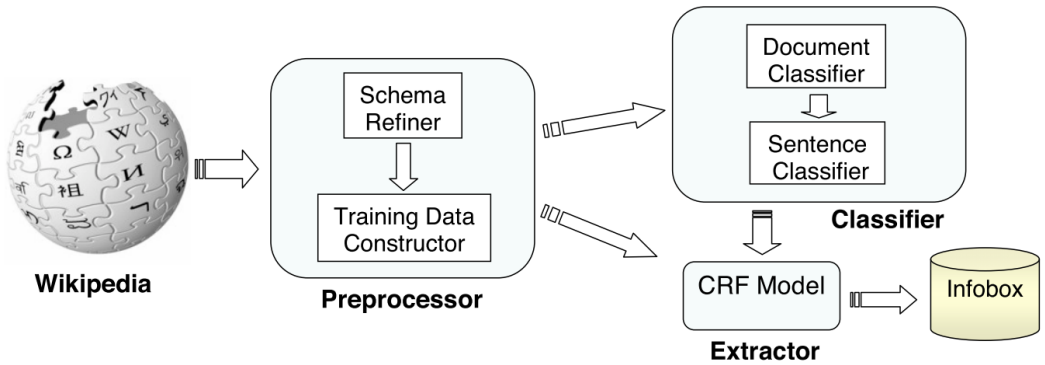


Figure 3: Architecture of KYLIN's infobox generator.

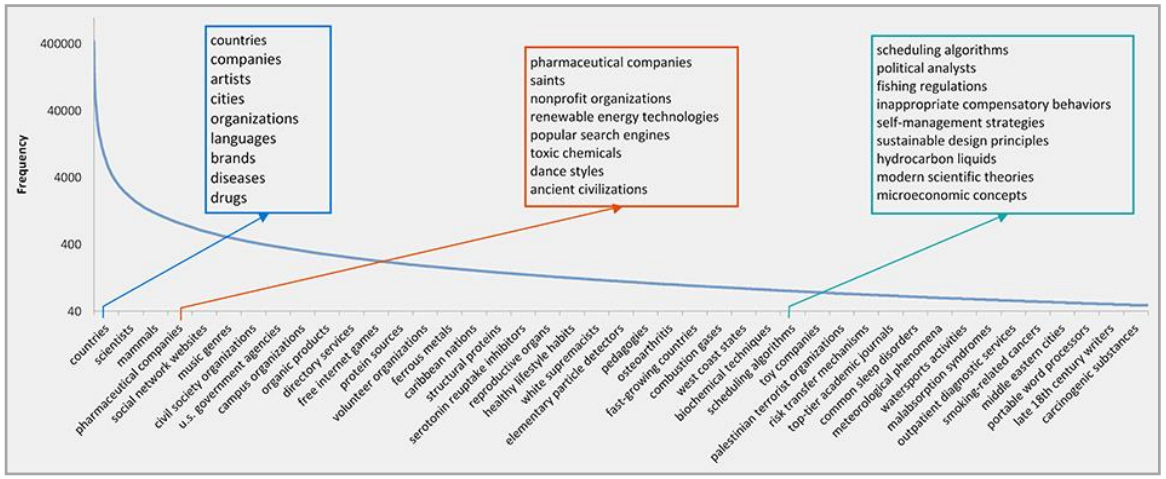
Ref: Fei Wu, etc.. Autonomously Semantifying Wikipedia

知识图谱优势

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

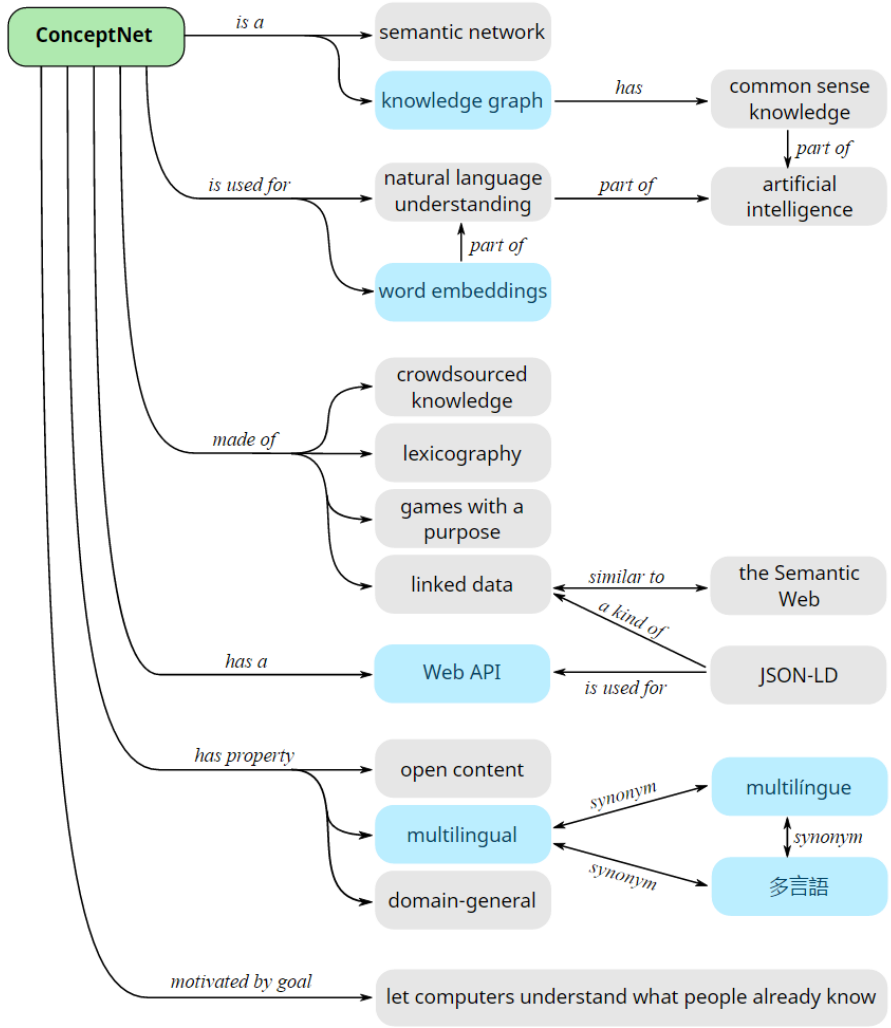


Existing Taxonomies	Number of Concepts
Freebase [5]	1,450
WordNet [13]	25,229
WikiTaxonomy [26]	111,654
YAGO [35]	352,297
DBPedia [1]	259
ResearchCyc [18]	≈ 120,000
KnowItAll [12]	N/A
TextRunner [2]	N/A
OMCS [31]	N/A
NELL [7]	123
Probase	**2,653,872**

KG优势2: semantically rich

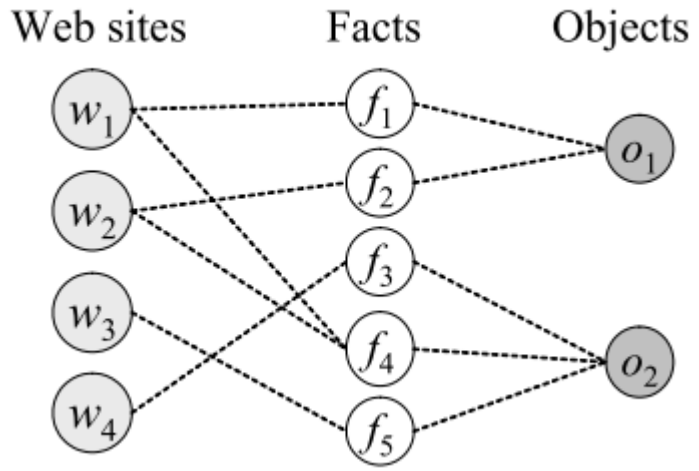
- Higher coverage over numerous semantic relationships

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



KG优势3： high quality

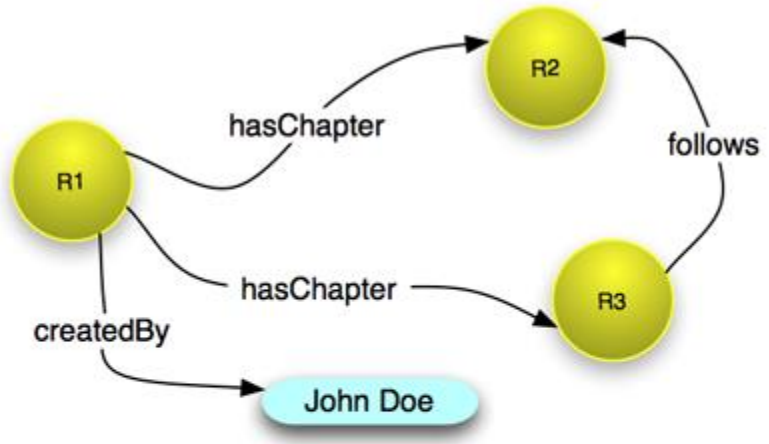
- High quality
  - Big data: Cross validation by multiple sources
  - Crowd sourcing: quality guarantee



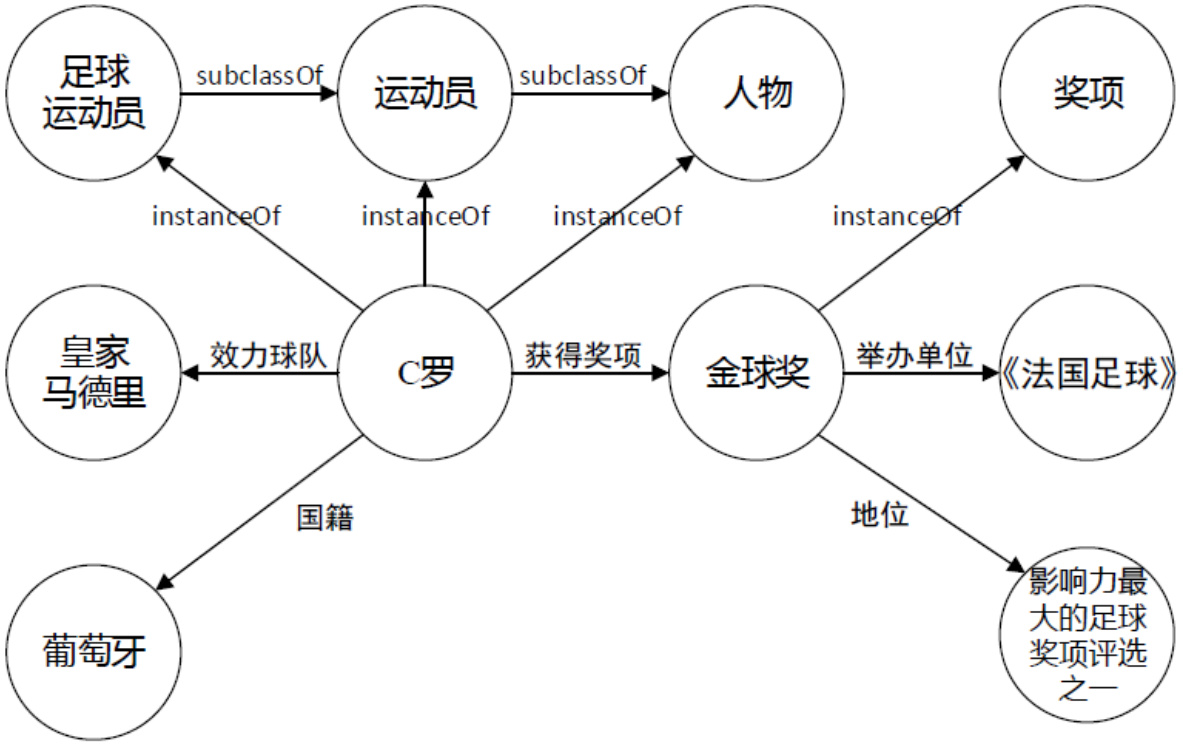
[Yin, et al. 2017]

KG优势4: friendly structure

- Structured organization
  - By RDF
  - By graph



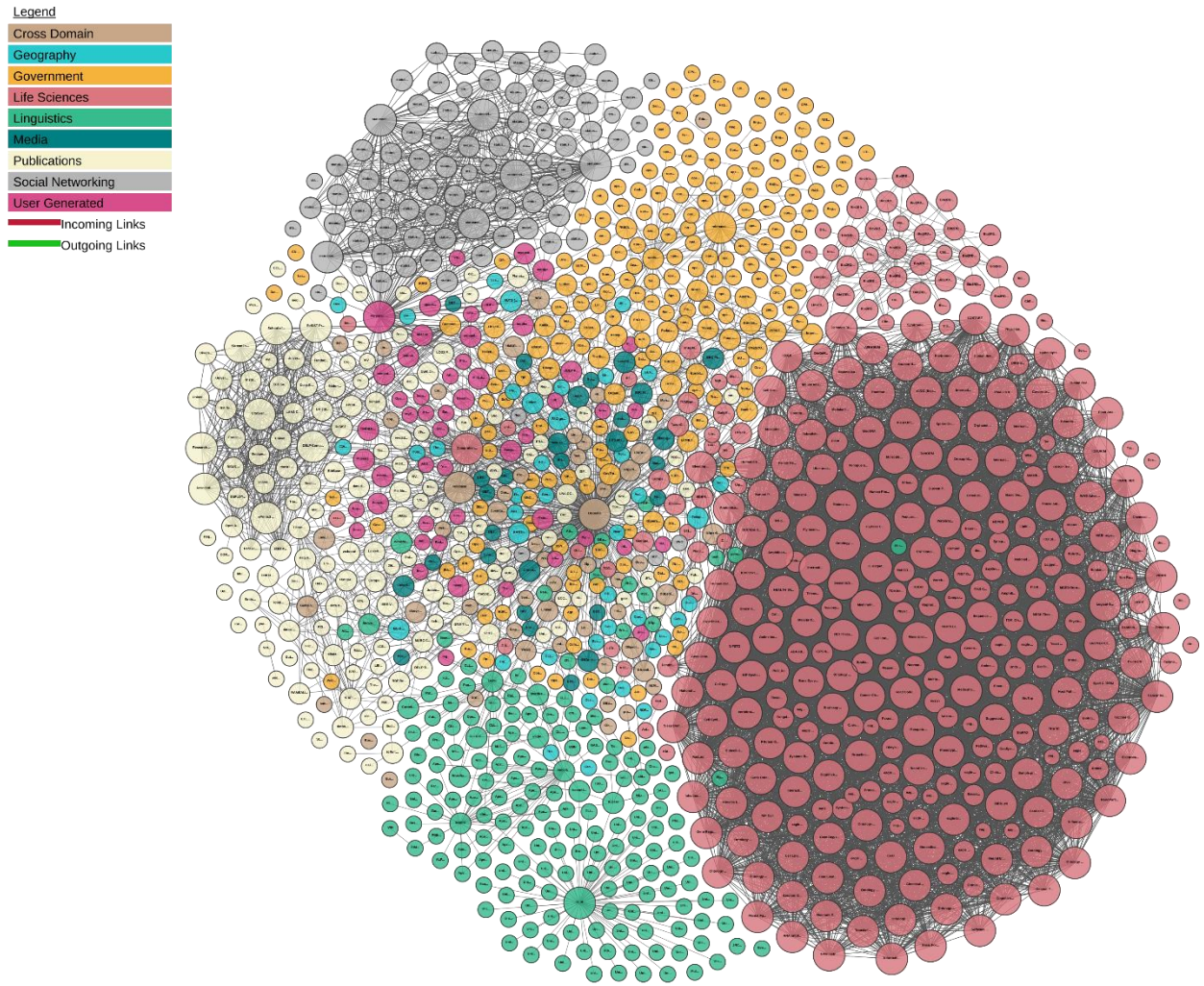
Subject	Predicate	Object
R1	hasChapter	R2
R1	hasChapter	R3
R3	follows	R2
R1	createdBy	"John Doe"



越来越多的知识图谱应运而生

Yago， WordNet, FreeBase, Probase, 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/"

知识图谱价值

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。

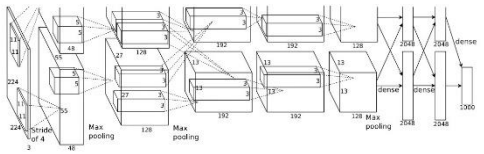
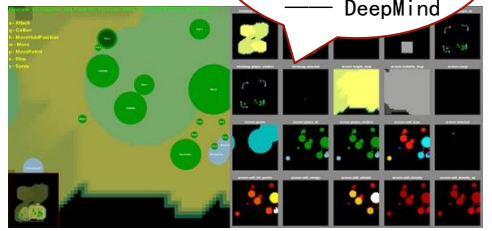


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input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

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]。

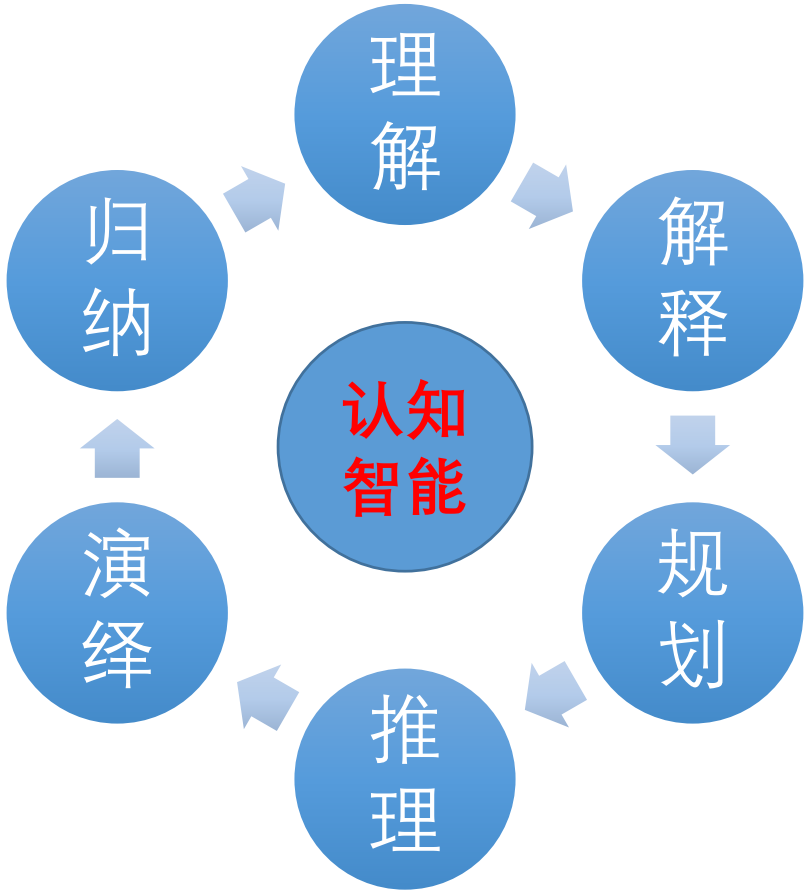


智能化升级与转型

- 智能化升级与转型已经成为各行各业的普遍诉求
- 从信息化走向智能化是必然趋势
- AI+成为AI赋能传统行业的基本模式

- 战略意义
  - 全方位、深度渗透到各行各业、各个环节
  - 颠覆性影响，重塑行业形态，甚至社会形态

认知智能是智能化的关键

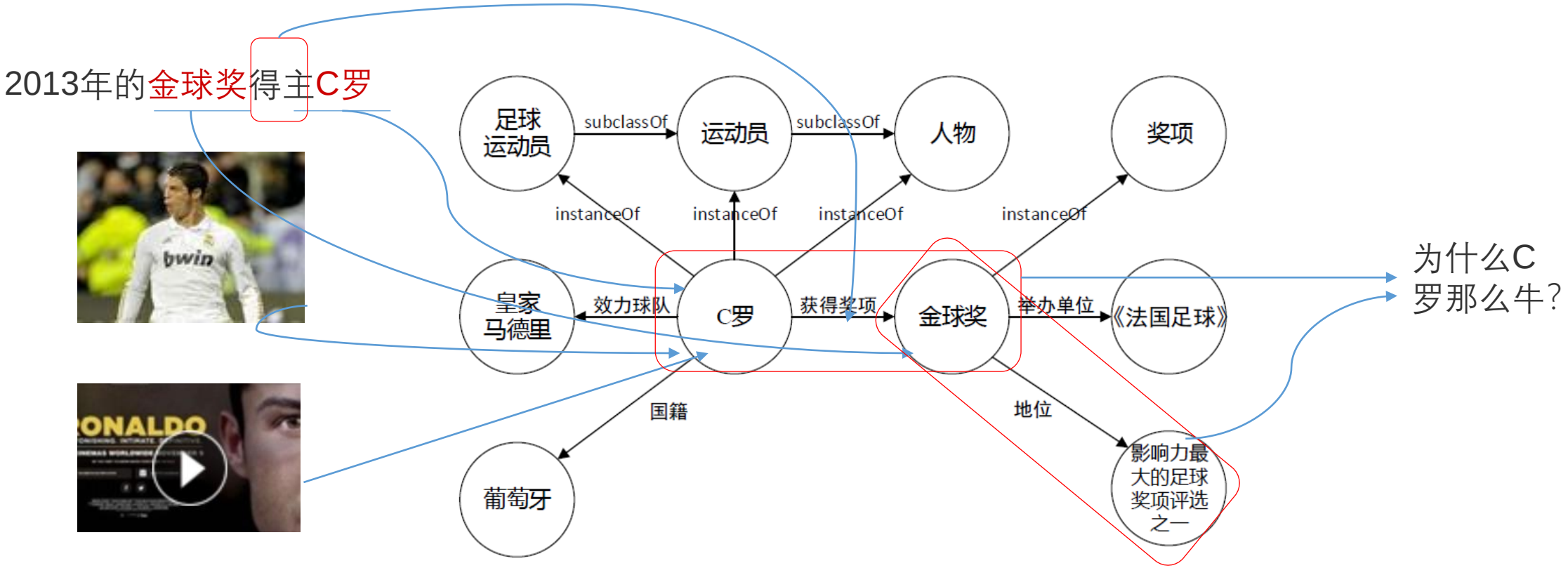


Can machine ***think like humans***?

- 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



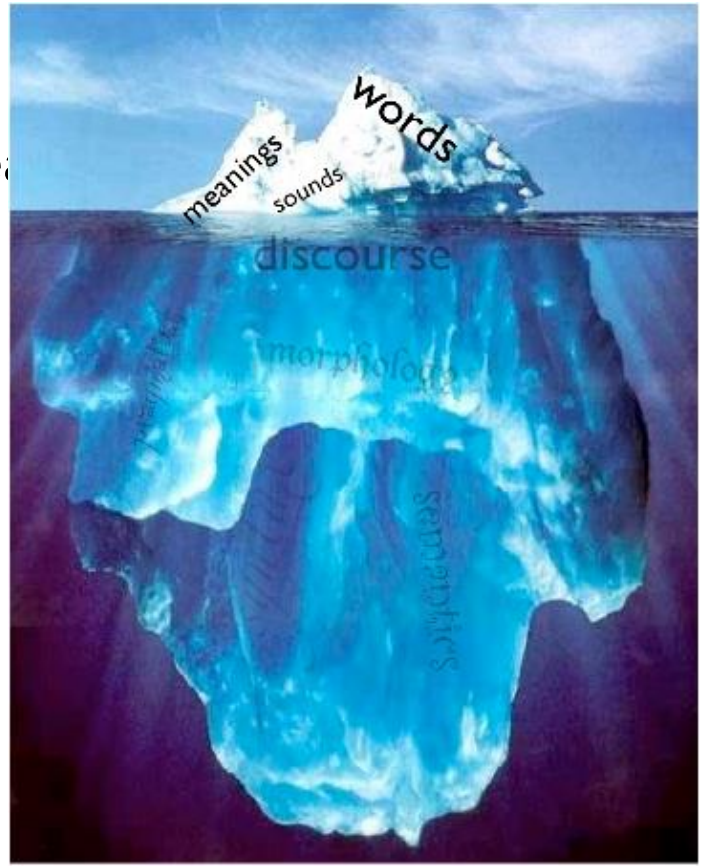
机器语言理解需要背景知识

Language is complicated

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

Language understanding is difficult

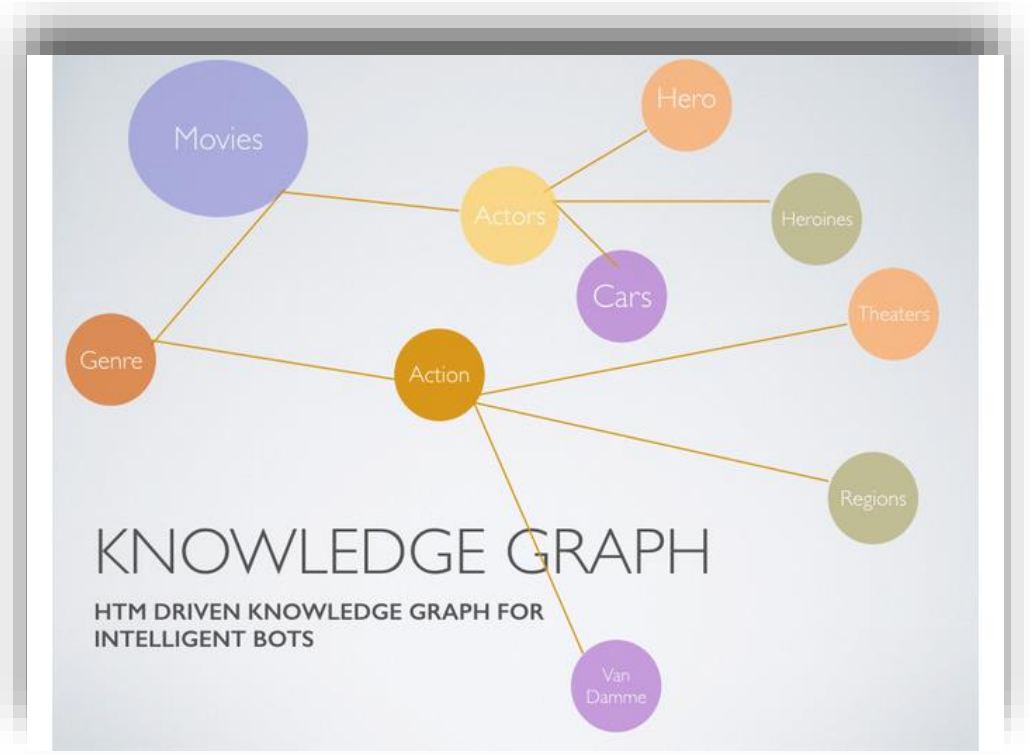
- Grounded only in **human cognition**
- Needs significant **background knowledge**



知识图谱*使能*(Enable)机器语言认知

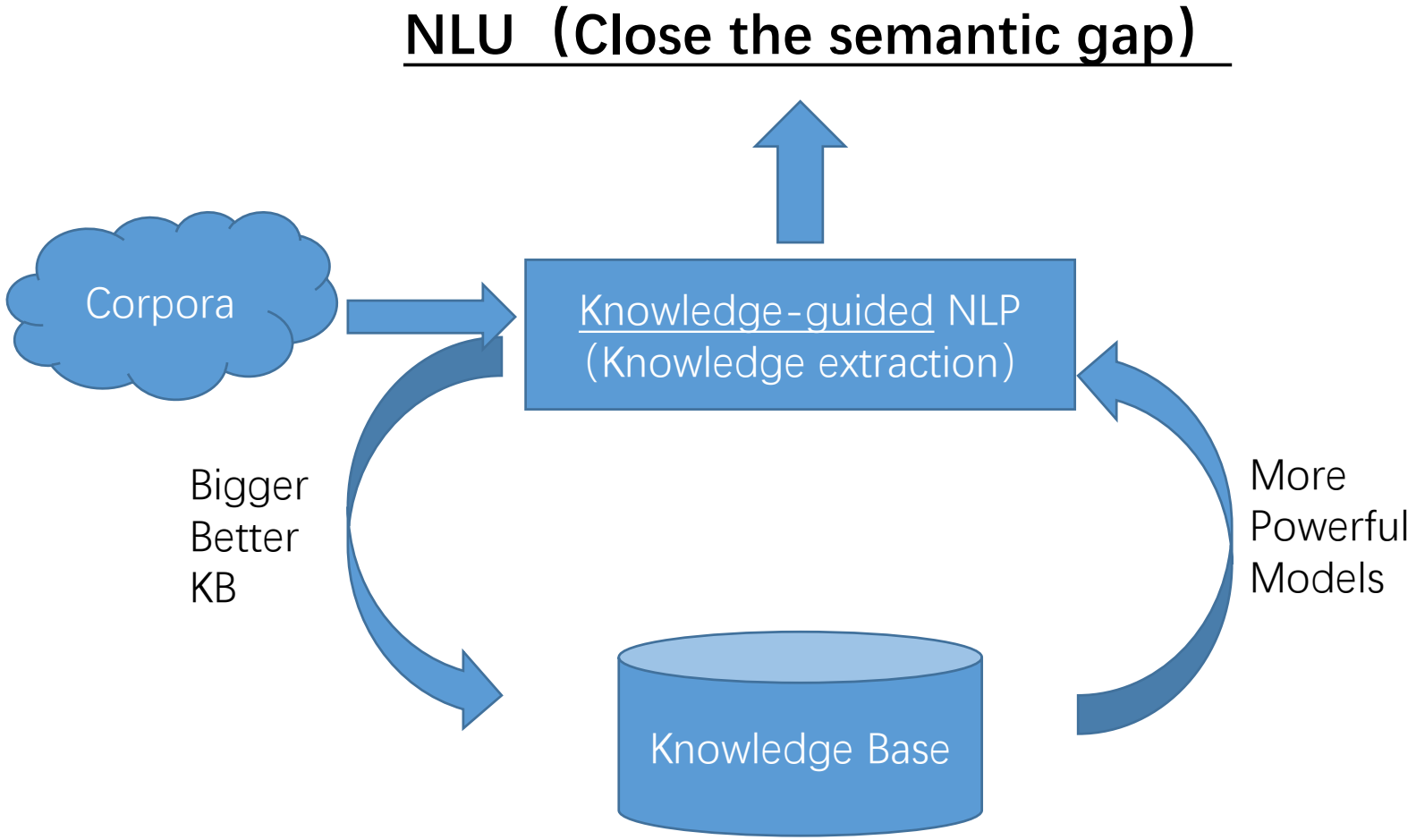
- Language understanding of machines needs knowledge bases
  - Large scale
  - Semantically rich
  - Friendly structure
  - High quality

- Traditional knowledge representations can not satisfy these requirements, but KG can
  - Ontology
  - Semantic network / frame
  - Texts



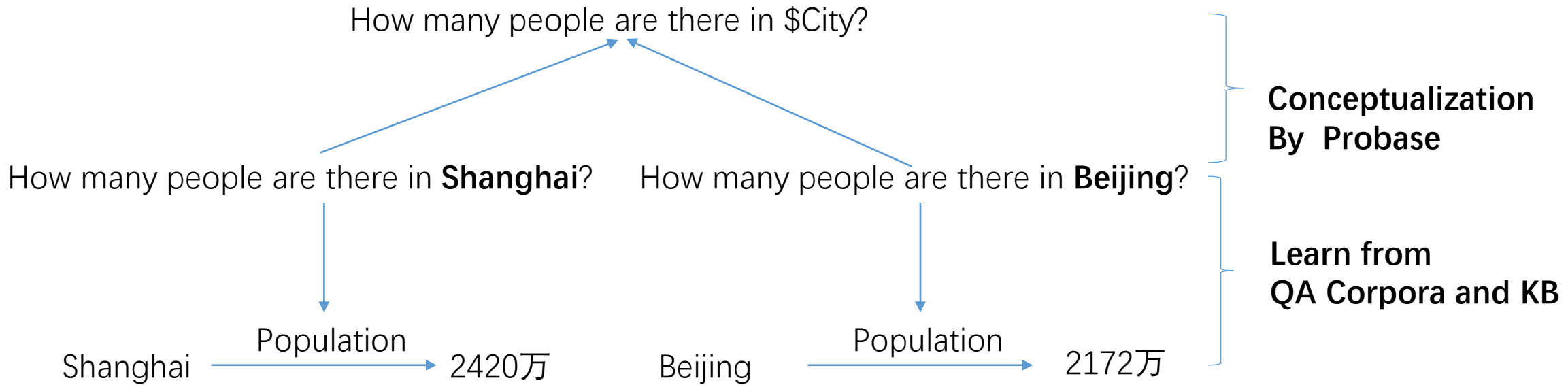
NLP+KB= NLU， NLP=Natural language processing， NLU=natural language understanding

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.



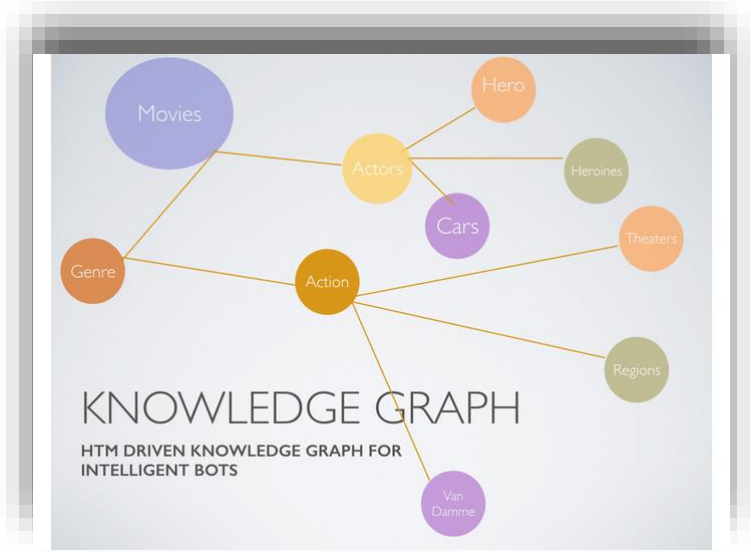
[Wanyun Cui et al. 2017]

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 → 概念

鸟儿为何能够飞翔?
因为它们有翅膀 → 属性

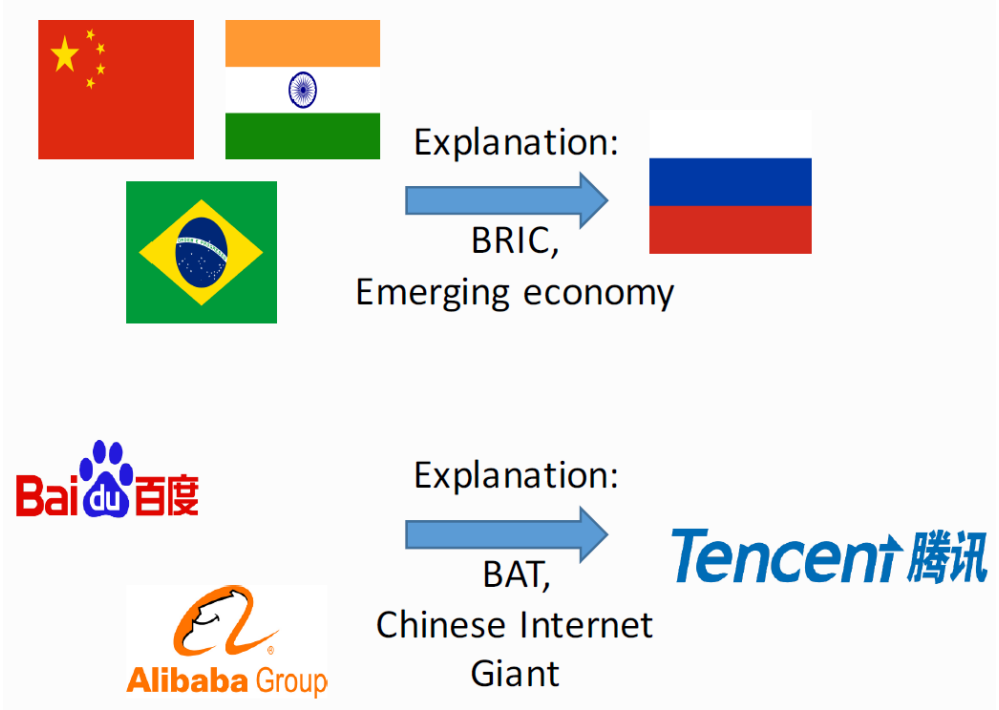
鹿晗关晓彤最近为何刷屏?
因为关晓彤是鹿晗女朋友 → 关系



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

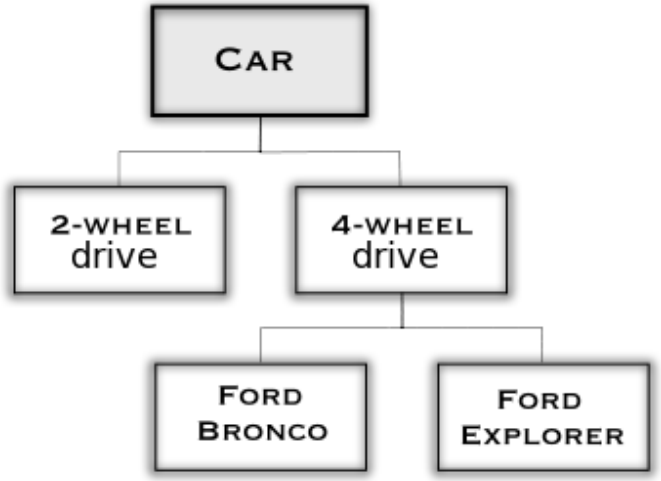
"Concepts are the glue that holds our mental world together"
--Gregory Murphy

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”

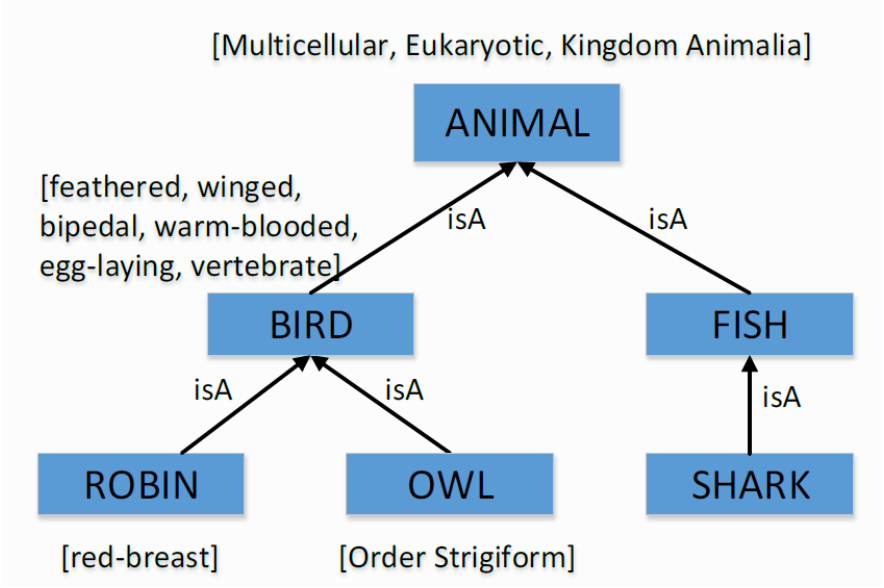
=> (Sex=man, Marriage status=unmarried)

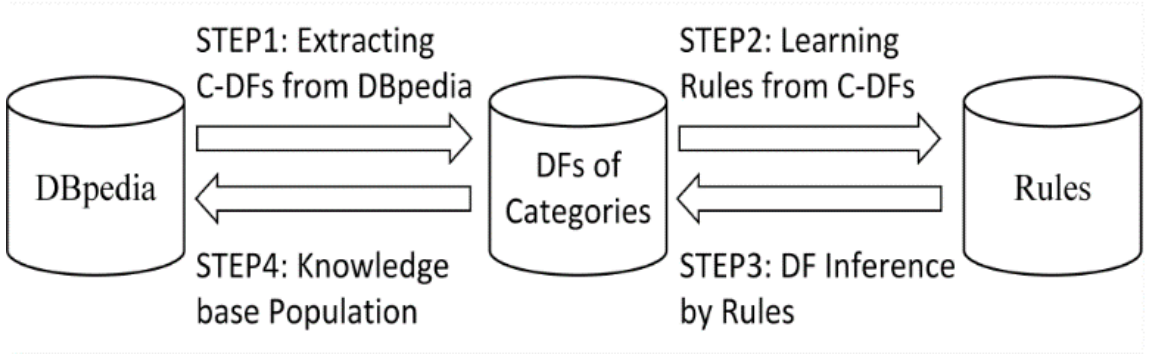
Basic Idea:

Ming Dbpedia, using properties to explain a category

Model:

Mining **Defining Features** from DBpeida

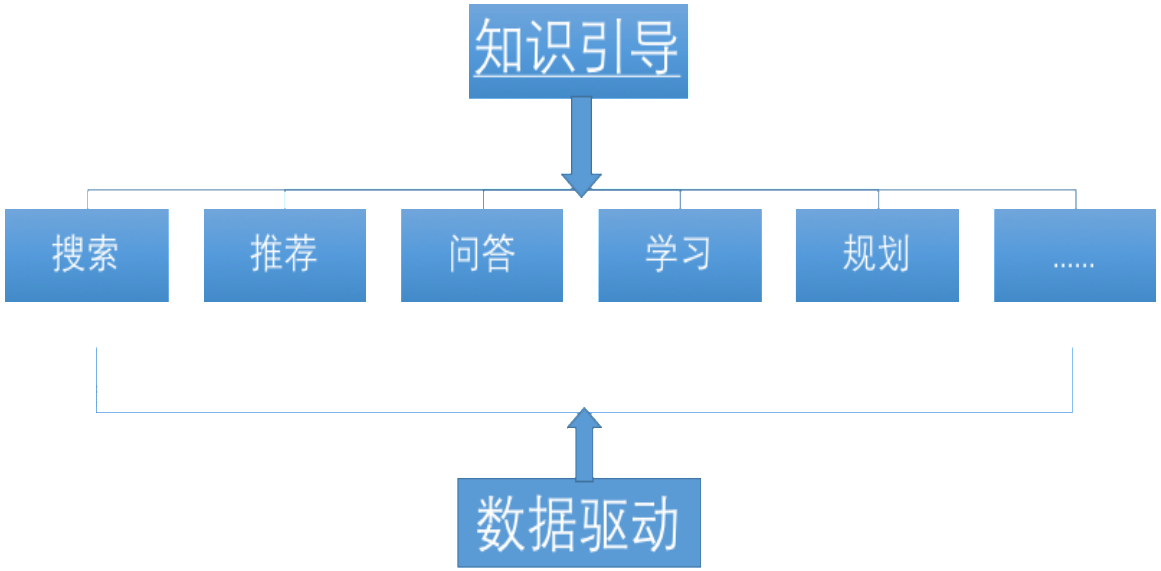


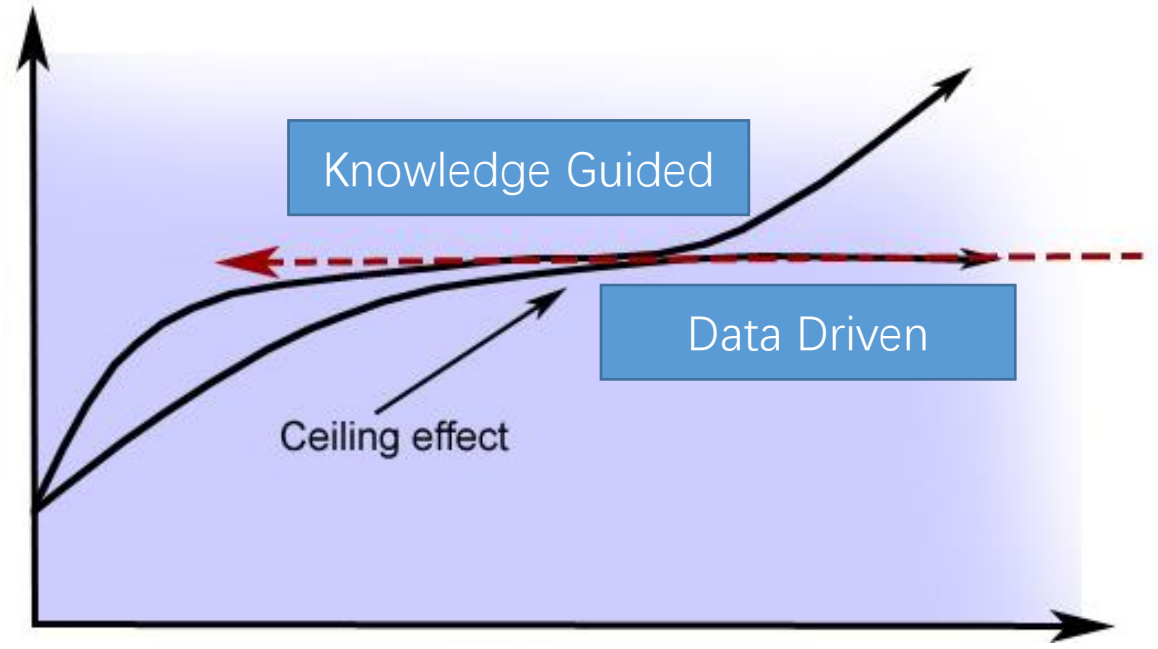


Solution Framework

[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

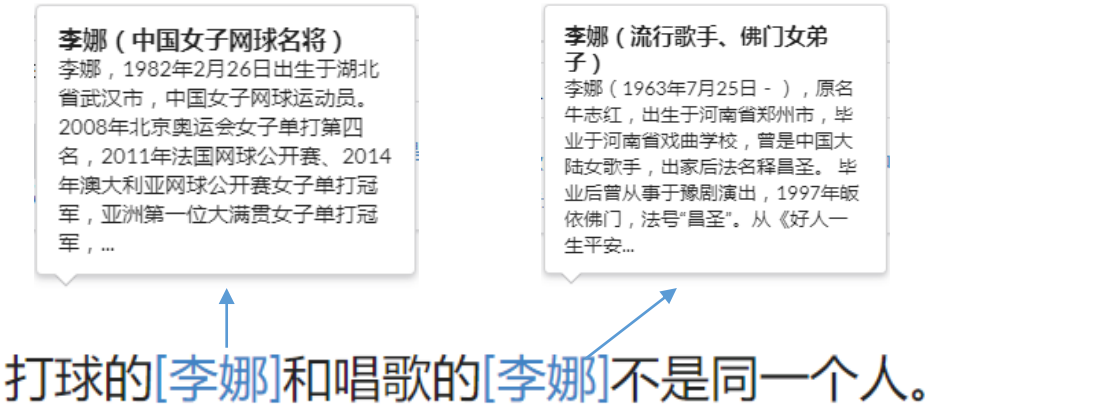
Example 1: Use Concepts for Chinese Entity Linking

- Entity linking： P(e|C),
  - where C is context and e is candidate entity
- Basic idea: using concepts (t) in knowledge base

$$P(e_i|C) = \sum_t P(e_i|t) \times P(t|C)$$

Typicality of an entity within a concept

The probability to observe an entity of t given context C



李娜（中国女子网球名将）：人物、体育人物、运动员、名将

李娜（流行歌手、佛门女弟子）：人物、演员、歌手、弟子

	** Entity Annotation API	Our Method
Precision	56.7%	86.1%
Recall	67.8%	84.5%
F1	61.7%	85.3%

Example 2: Using knowledge to prevent semantic drift in pattern based IE

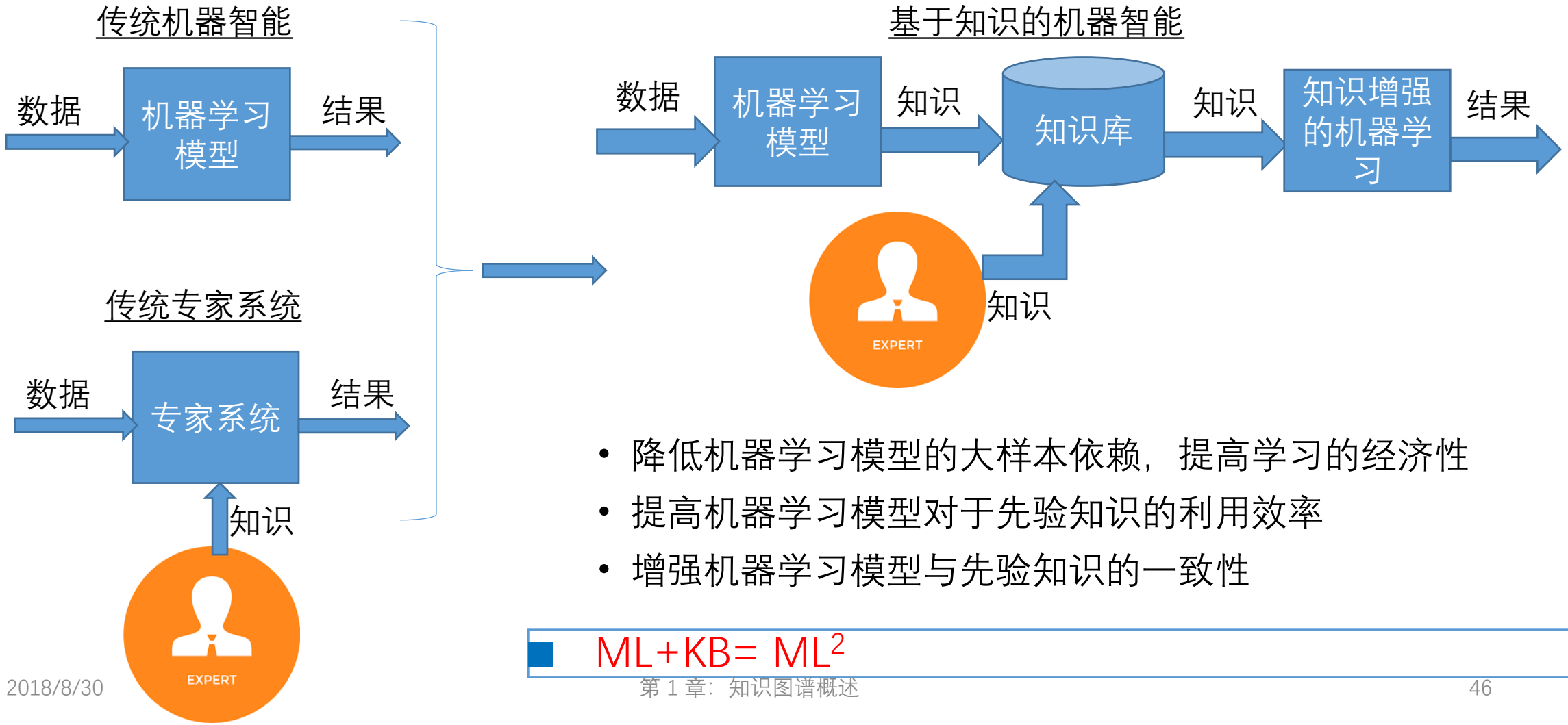
- Pattern based bootstrapping is popular
- Problem: semantic drift
  - <China isA country> =>
  - ‘occupation of $’, =>
  - ‘occupation of Planet earth’=>
  - <Planet Earch isA coutntry>
- Principles: no bad patterns, only wrong applications
- Our idea
  - Run a pattern on the text for an appropriate entity
  - Using knowledge to guide the execution of the learned pattern
  - 95%+ accuracy

#复旦大学（Fudan University），简称“复旦”，位于中国上海，由中华人民共和国教育部直属，中央直管副部级建制，位列211工程、985工程、双一流A类，入选“珠峰计划”、“111计划”、“2011计划”、“卓越医生教育培养计划”，为“九校联盟”成员、中国大学校长联谊会成员、东亚研究型大学协会成员、环太平洋大学协会成员，是一所世界知名、国内顶尖的综合性研究型的全国重点大学。\n复旦大学创建于1905年，原名复旦公学，是中国人自主创办的第一所高等院校，创始人为中国近代知名教育家马相伯，首任校董

<复旦大学 - 简称 - 复旦>
<复旦大学 - 创始人 - 马相伯>
……

鹿晗	外文名称	LU HAN
鹿晗	出生日期	1990年4月20日
刘诗诗	职业	影视出品人
张艺兴	外文名称	LAY
张艺兴	出生日期	1991年10月7日
angelababy	出生日期	1989年2月28日
赵丽颖	出生地	河北省廊坊市
杨幂	出生日期	1986年9月12日
郑爽（中国内地90后女演员）	出生地	辽宁省沈阳市
宋茜	外文名称	Victoria
宋茜	出生日期	1987年2月2日
宋茜	职业	广告模特在亚洲地区正式开始演艺活动
刘德华（中国香港男演员、歌手、词作人）	出生日期	1961年9月27日
刘德华（中国香港男演员、歌手、词作人）	代表作品	只知道此刻爱你
李易峰	出生日期	1987年5月4日
李易峰	出生地	四川成都
李易峰	代表作品	小先生
周杰伦（华语流行男歌手）	出生日期	1979年1月18日
周杰伦（华语流行男歌手）	主要成就	台湾电影金马奖年度台湾杰出电影
周杰伦（华语流行男歌手）	代表作品	Jay

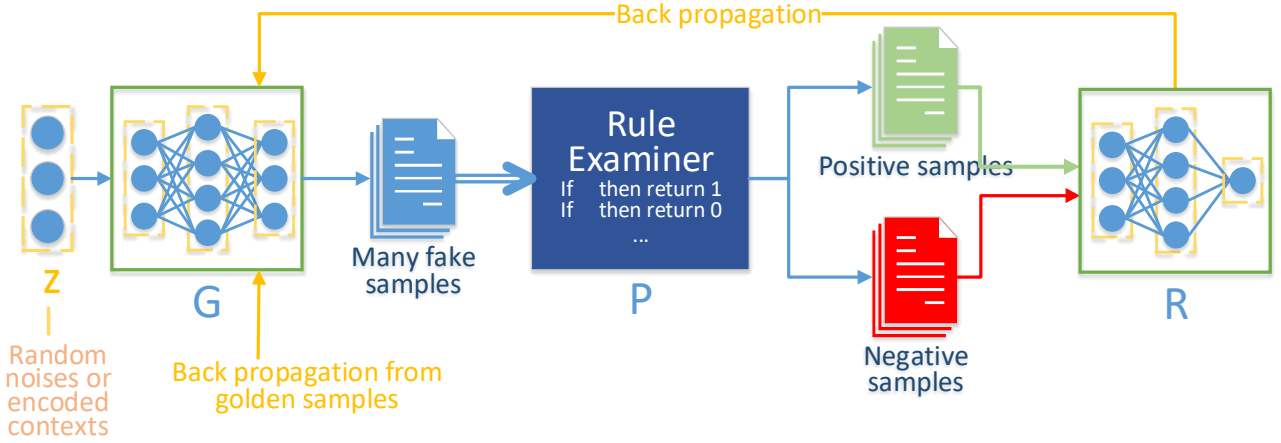
知识将显著增强机器学习能力

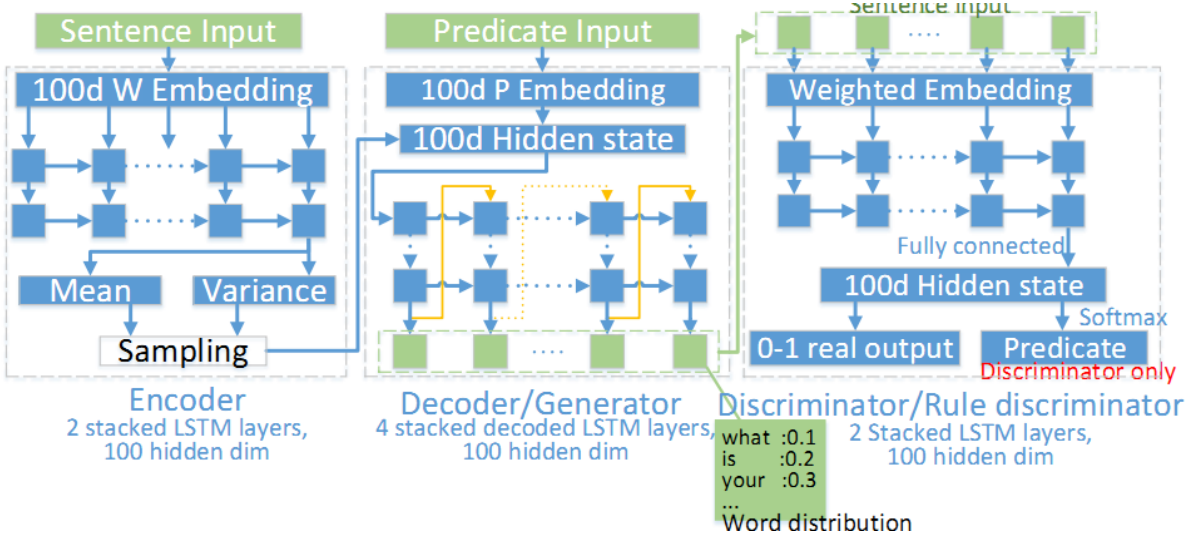


- 降低机器学习模型的大样本依赖，提高学习的经济性
- 提高机器学习模型对于先验知识的利用效率
- 增强机器学习模型与先验知识的一致性

$ML+KB= ML^2$

Example 1: Deep language generation with prior knowledge





Incorporating Complicated Rules in Deep Generative Models, under review

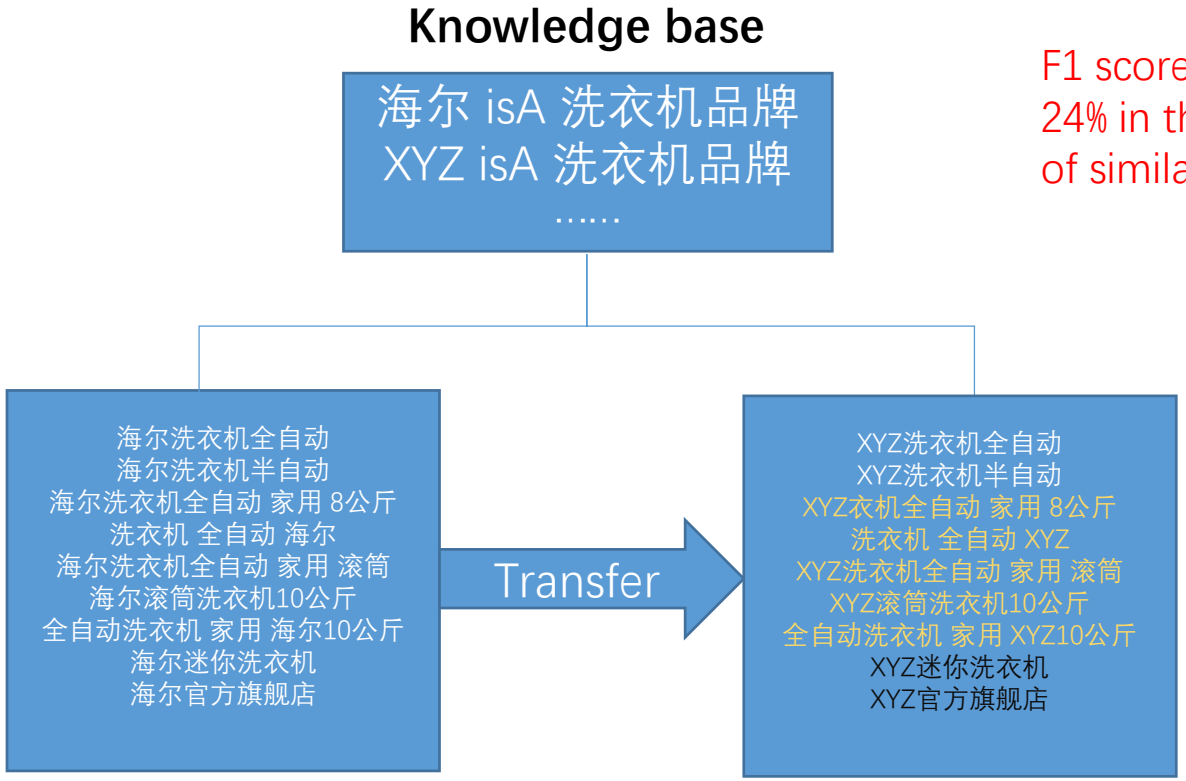
Rules for Chinese question generation
1 Sentences should end with '#' (a special character).
2 The subject should appear only once in a sentence.
3 There are no continuously repeated characters in the sentence.
4 The length of sentences should be more than 4 characters. *(A Chinese question should not be too short.)*
5 The number of low frequency words should less than half of sentence length.

在超级验证码中的应用

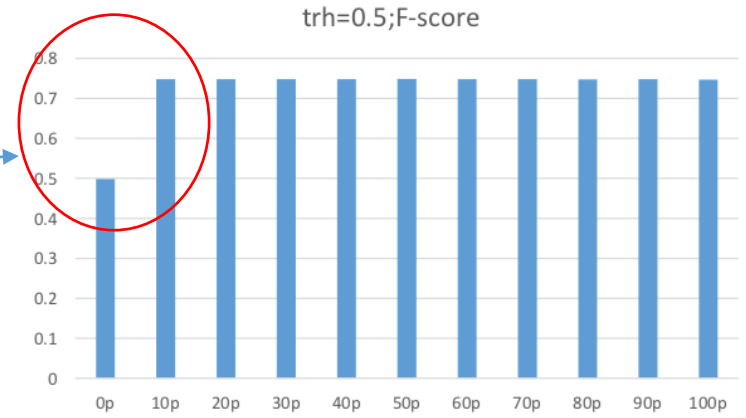
Demo地址：http://kw.fudan.edu.cn/ddemos/vcode/
API地址：http://kw.fudan.edu.cn/apis/supervcode/

Example 2: Long-tailed query term embedding guided by knowledge

- In Deep IR, its hard to train effective word embedding for long tailed query terms



F1 score increases by 24% in the evaluation of similar queries



raw

	日式 -1
0	日式 1.0
1	剑林 0.9090448594528984
2	山田烧 0.9076092872105608
3	摩登主妇 0.9041964983257302
4	lototo 0.903521871998948
5	朵颐 0.9018902344911408
6	川岛屋 0.8992372679673781
7	一人食 0.8990165065859497
8	手绘碗 0.8966848604326612
9	二人食 0.8946378874188308

new

	日式 -1
0	日式 1.0
1	纯白色 0.9288789768835618
2	韩式 0.9282034998043911
3	法式 0.927806029695622
4	风格 0.9275196253763414
5	田园 0.9249904565619058
6	禅意 0.9235103898321229
7	素雅 0.9199717204810847
8	欧式 0.9188282247364457
9	田园风 0.9182616342150595

知识将成为比数据更为重要的资产

- 大数据时代是得“数据者” 得天下
- 人工智能时代是得“知识者”得天下

- 数据是石油，知识就是石油的萃取物

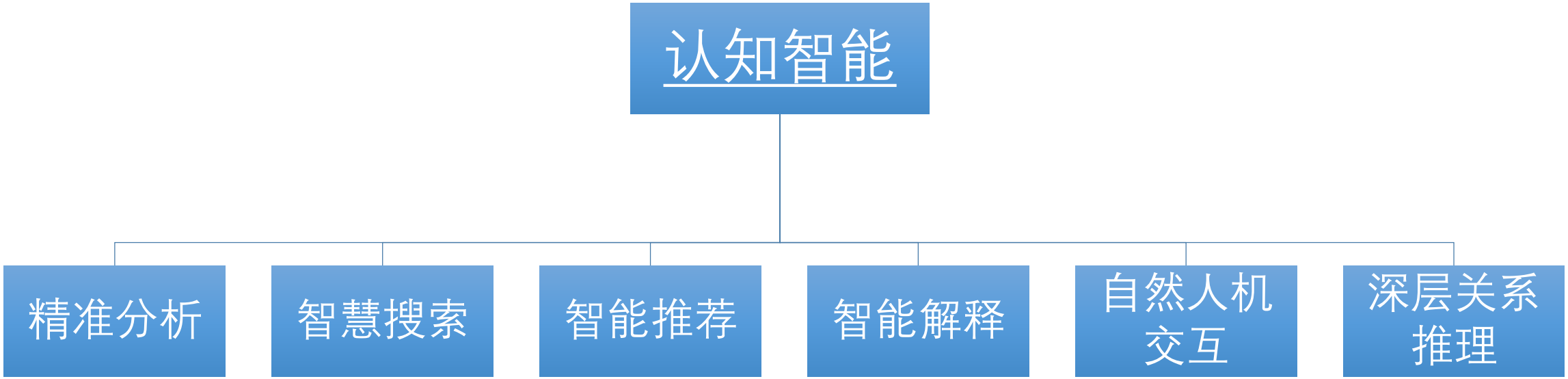
知识加工与石油萃取

“Knowledge is power in AI”，　Edward Feigenbaum

知识图谱应用

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】

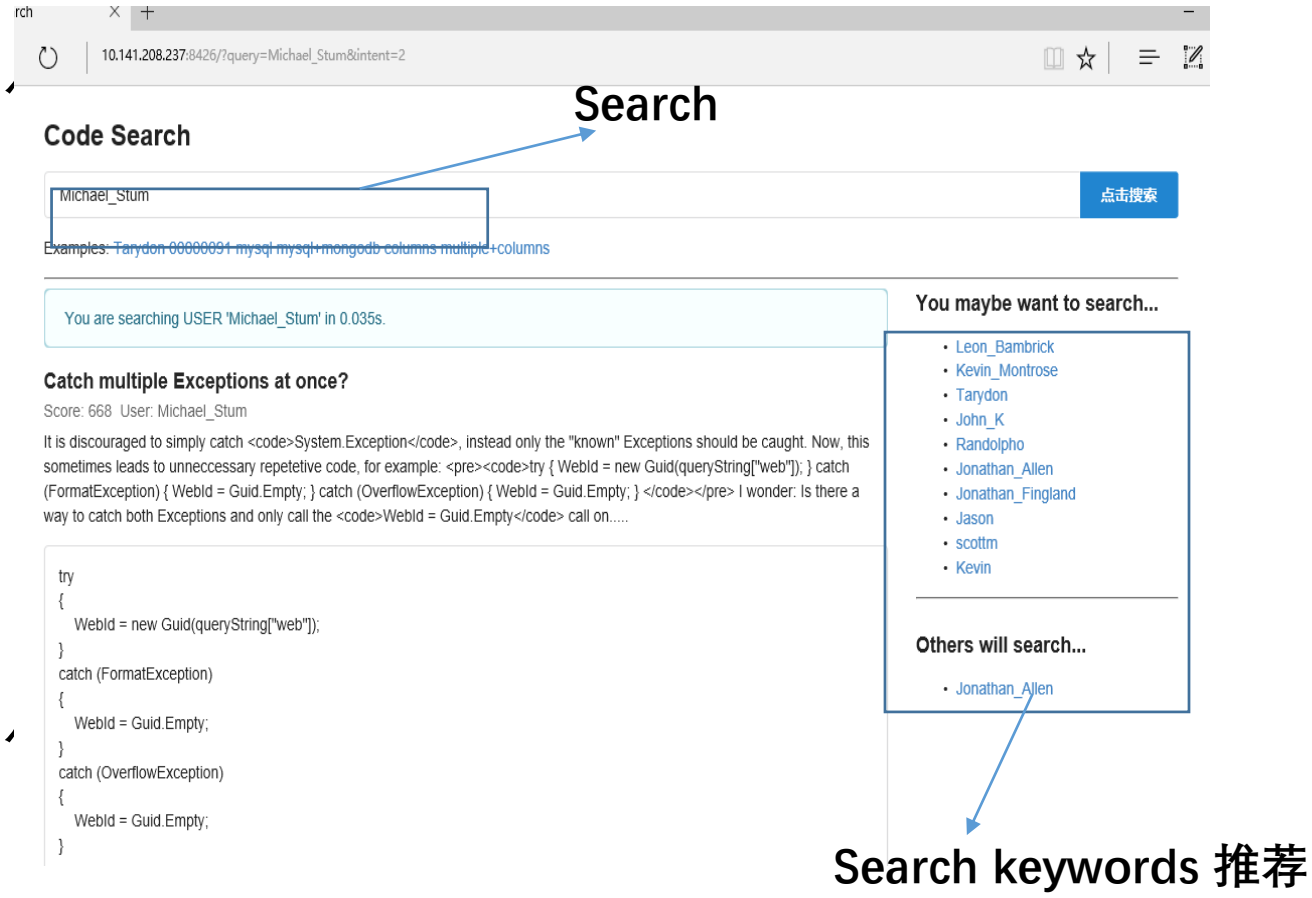
军民融合南海掀波 陆渔船舰队近逼菲中业岛 → 菲律宾 相关

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

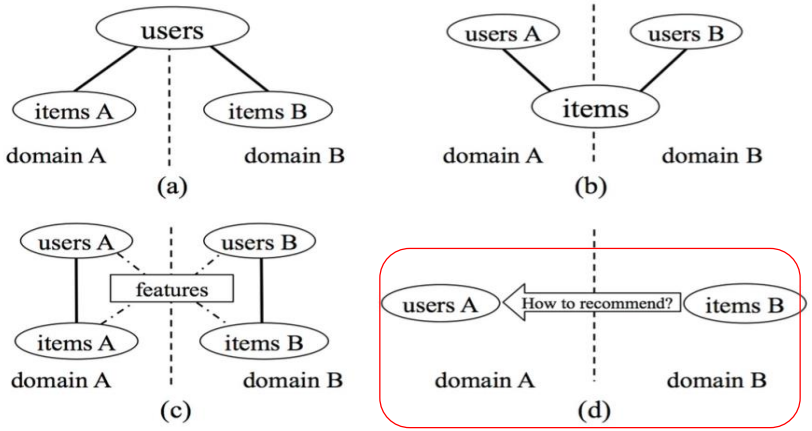
- 精准搜索意图理解
  - 精准分类、语义理解、个性
- 复杂多元对象搜索
  - 表格、文本、图片、视频
  - 文案、素材、代码、专家
- 多粒度搜索
  - 篇章级、段落级、语句级
- 跨媒体搜索
  - 不同媒体数据联合完成搜索



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



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

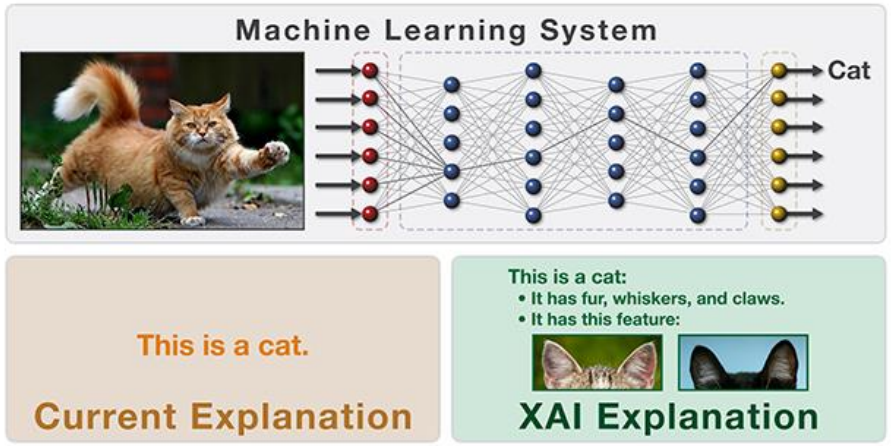
电商领域的场景化推荐

- 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

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

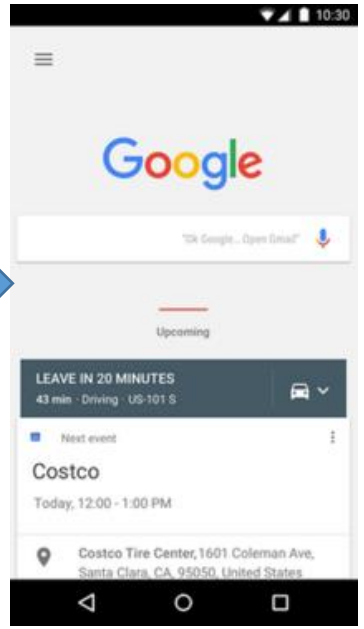


解释事实

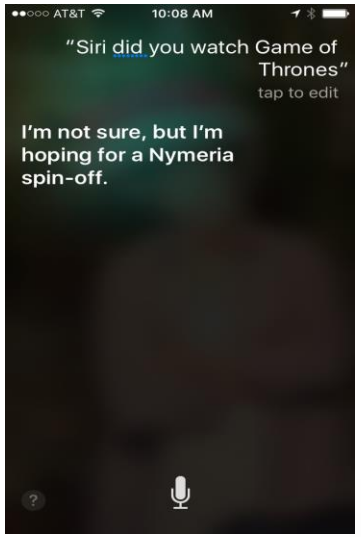
- 解释是智能的重要体现之一，将是人们对于智能系统的普遍期望
- 可解释是智能系统决策结果被采信的前提

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

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

知识分类

知识类别

- factual knowledge
  - bornIn (SteveJobs, SanFrancisco), hasFounded (SteveJobs, Pixar),
  - hasWon (SteveJobs, NationalMedalOfTechnology), livedIn (SteveJobs, PaloAlto)

- taxonomic knowledge (ontology):
  - instanceOf (SteveJobs, computerArchitects), instanceOf(SteveJobs, CEOs)
  - subclassOf (computerArchitects, engineers), subclassOf(CEOs, businesspeople)

知识类别

- lexical knowledge (terminology):
  - means (“Big Apple“, NewYorkCity), means (“Apple“, AppleComputerCorp)
  - means (“MS“, Microsoft) , means (“MS“, MultipleSclerosis)
- contextual knowledge (entity occurrences, entity-name disambiguation)
  - maps (“Gates and Allen founded the Evil Empire“, BillGates, PaulAllen, MicrosoftCorp)
- linked knowledge (entity equivalence, entity resolution):
  - sameAs (Apple, AppleCorp), sameAs (hasFounded, isFounderOf)

知识类别

- multi-lingual knowledge:
  - meansInChinese („乔戈里峰“, K2), meansInUrdu („کے ٹو“, K2)
  - meansInFr („école“, school (institution)), meansInFr („banc“, school (of fish))
- temporal knowledge (fluents):
  - hasWon (SteveJobs, NationalMedalOfTechnology)@1985
  - marriedTo (AlbertEinstein, MilevaMaric)@[6-Jan-1903, 14-Feb-1919]
  - presidentOf (NicolasSarkozy, France)@[16-May-2007, 15-May-2012]
- spatial knowledge:
  - locatedIn (YumbillaFalls, Peru), instanceOf (YumbillaFalls, TieredWaterfalls)
  - hasCoordinates (YumbillaFalls, 5°55‘11.64“S 77°54‘04.32“W),
  - closestTown (YumbillaFalls, Cuispes), reachedBy (YumbillaFalls, RentALama)

知识类别

- 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: human(x) $\Rightarrow$ male(x) $\vee$ female(x)
  - $\forall$ x: (male(x) $\Rightarrow \neg$ female(x)) $\wedge$ (female(x)) $\Rightarrow \neg$ male(x))
  - $\forall$ x: human(x) $\Rightarrow$ ($\exists$ y: mother(x,y) $\wedge \exists$ z: father(x,z))
  - $\forall$ x: animal(x) $\Rightarrow$ (hasLegs(x) $\Rightarrow$ isEven(numberOfLegs(x))

知识类别

- emerging knowledge (open IE):
  - hasWon (MerylStreep, AcademyAward)
  - occurs („Meryl Streep“, „celebrated for“, „Oscar for Best Actress“)
  - occurs („Quentin“, „nominated for“, „Oscar“)
- multimodal knowledge (photos, videos):
  - JimGray
  - JamesBruceFalls

- social knowledge (opinions):
  - admires (maleTeen, LadyGaga), supports (AngelaMerkel, HelpForGreece)
- epistemic knowledge ((un-)trusted beliefs):
  - believe(Ptolemy,hasCenter(world,earth)), believe(Copernicus,hasCenter(world,sun))
  - believe (peopleFromTexas, bornIn(BarackObama,Kenya))

典型知识图谱

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

ID	知识图谱	构建方式	数据来源	语言	范围
1	Cyc	人工	——	英文	通用
2	WordNet	人工	——	英文	通用
3	ConceptNet	自动	知识图谱	多语言	通用
4	GeoNames	半自动	百科	多语言	领域
5	Freebase	半自动	百科	英文	通用
6	YAGO	自动	百科	多语言	通用
7	DBpedia	半自动	百科	多语言	通用
8	Open IE	自动	纯文本	英文	通用
9	BabelNet	自动	知识图谱	多语言	通用
10	Google KG	自动	混合	多语言	通用
11	Probase	自动	纯文本	英文	通用
12	搜狗知立方	自动	百科	中文	通用
13	百度知心	自动	百科	中文	通用
14	CN-DBpedia	自动	百科	中文	通用

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	***Unique Strings***	***Synsets***	***Total Word-Sense Pairs***
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)

[Robert Speer et al. 2012]

http://conceptnet.io/

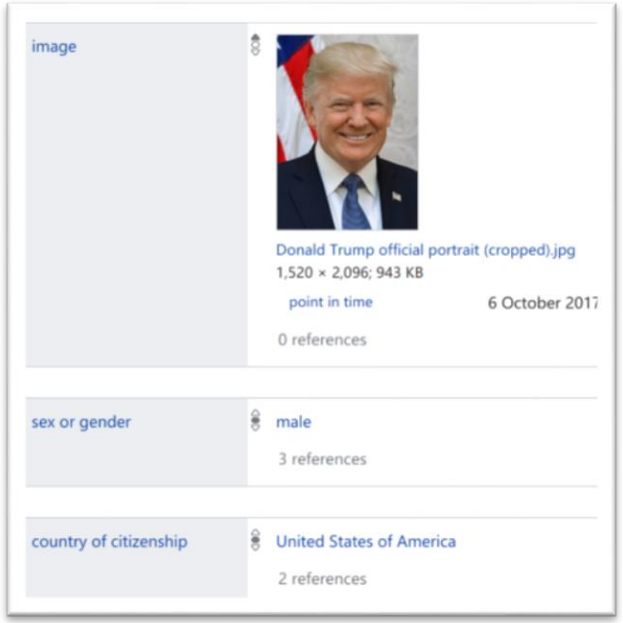
GeoNames

- 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



[Bollackeret 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”
- 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

```
<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 .
```

[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
  - 目前已经从十亿的互联网页面中抽取出了50 亿条关系

[Banko et al. 2007], [Etzioniet al. 2011]

http://openie.allenai.org/

BabelNet

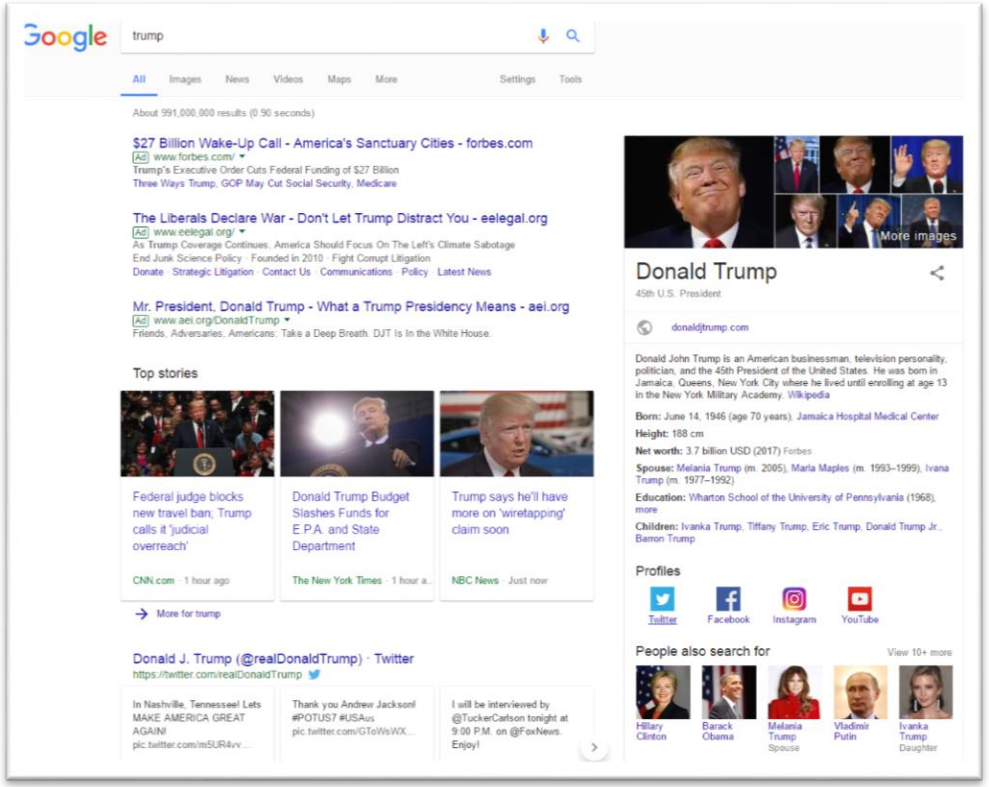
- 简介
  - 多语言知识图谱
- 样例
  - “周杰伦”
- 特点
  - 271 种语言
  - 自动融合
- 统计
  - 最新版为BabelNet 3.7，共包含1400 万个实体

[Roberto Navigli et. al., 2012]

http://babelnet.org/

Google KG

- 简介
  - 谷歌知识图谱于2012 年发布，被认为是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* such as {*NP*,}*{(or | and)} *NP*
2	such *NP* as {*NP*,}*{(or | and)} *NP*
3	*NP*{,} including {*NP*,}*{(or | and)} *NP*
4	*NP*{,*NP*}*{,} and other *NP*
5	*NP*{,*NP*}*{,} or other *NP*
6	*NP*{,} especially {*NP*,}*{(or | and)} *NP*

[Wu et al. 2012]

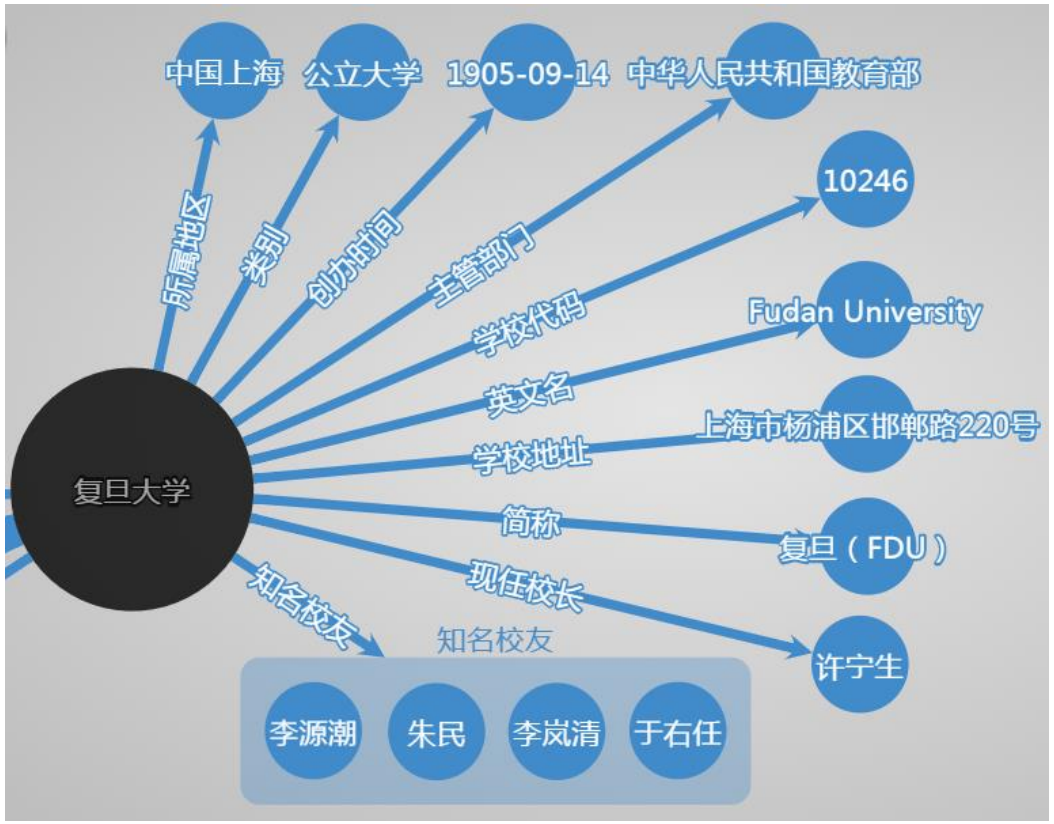
搜狗知立方/百度知心

- 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

- 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

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. Al] 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)**