知识图谱与认知智能

肖仰华

复旦大学知识工场实验室

Kw.fudan.edu.cn

shawyh@fudan.edu.cn

提纲

智能时代

催生

智能应用

需要

认知智能

依赖

知识图谱

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。

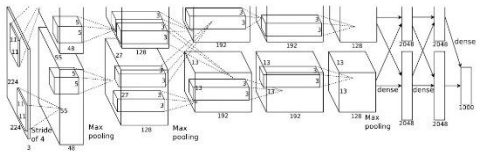
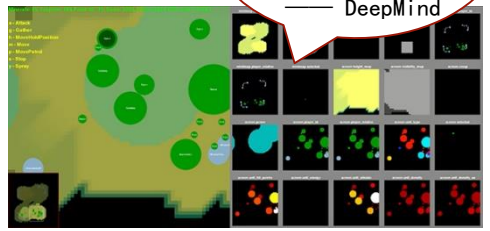


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input 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

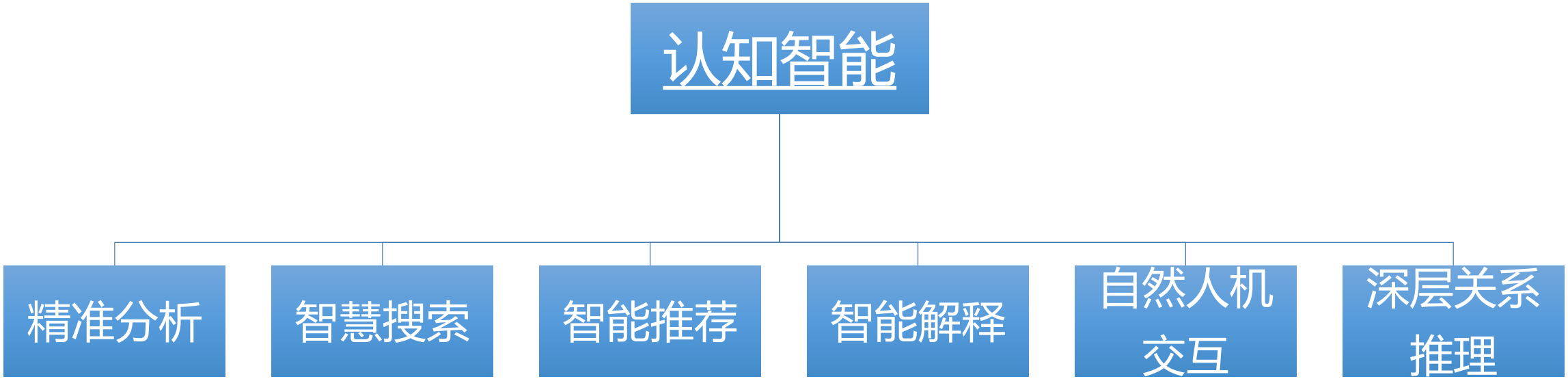
- 战略意义
  - 全方位、深度渗透到各行各业、各个环节
  - 颠覆性影响，重塑行业形态，甚至社会形态

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

认知智能的应用需求

- 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】

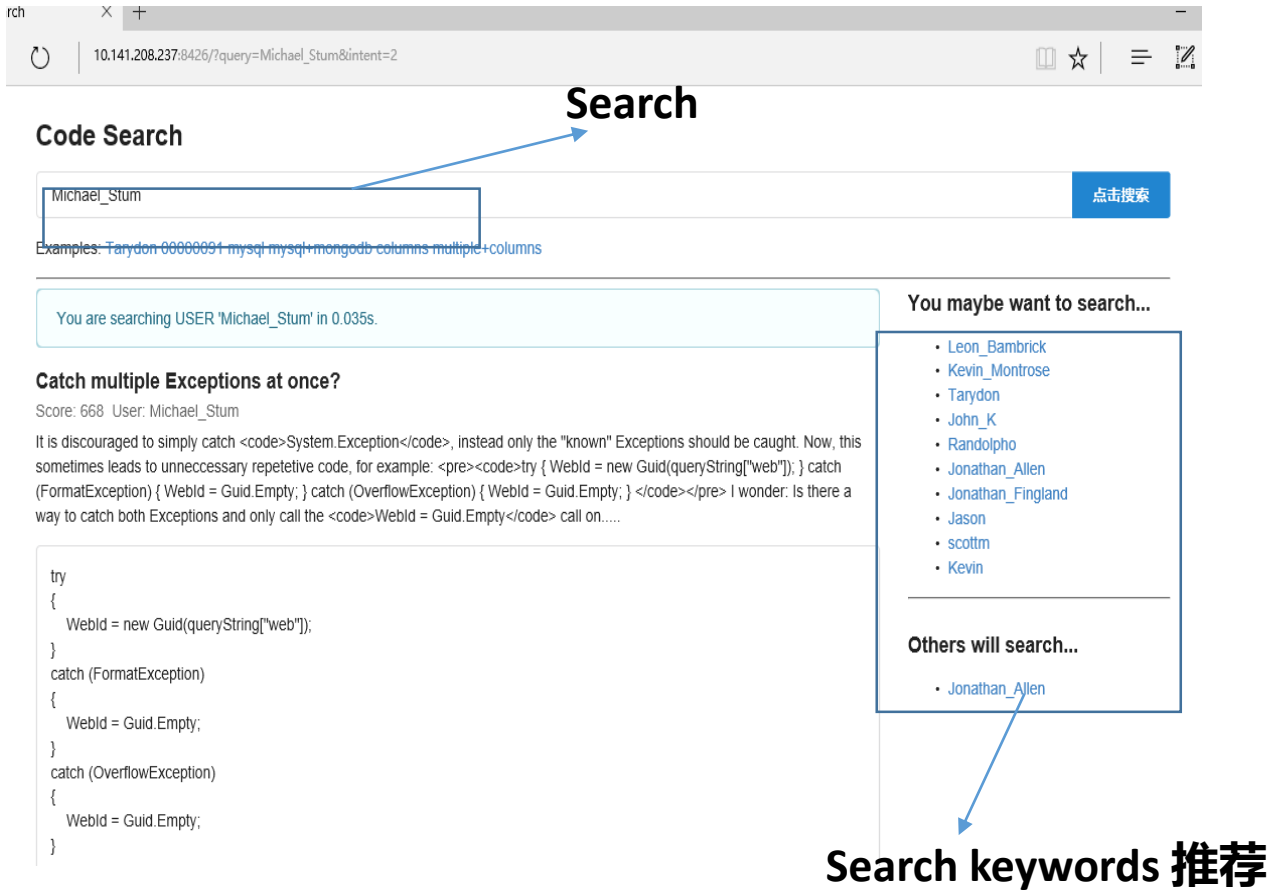
军民融合南海掀波 陆渔船舰队近逼菲中业岛 → 菲律宾 相关

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

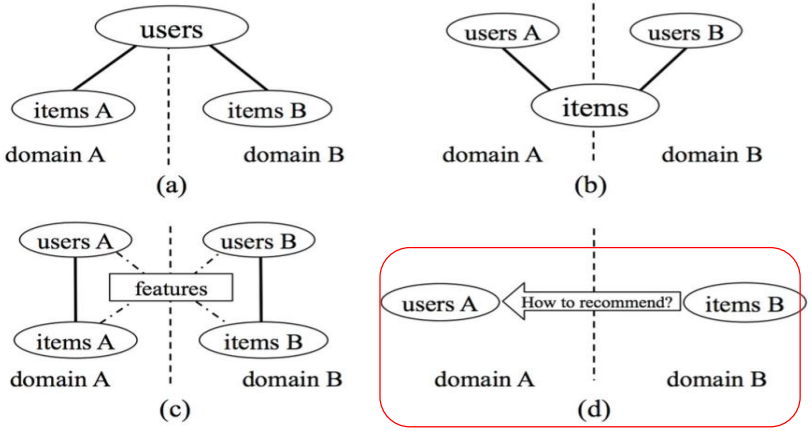
- 精准搜索意图理解
  - 精准分类、语义理解、个性化
- 复杂多元对象搜索
  - 表格、文本、图片、视频
  - 文案、素材、代码、专家
- 多粒度搜索
  - 篇章级、段落级、语句级
- 跨媒体搜索
  - 不同媒体数据联合完成搜索任务



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

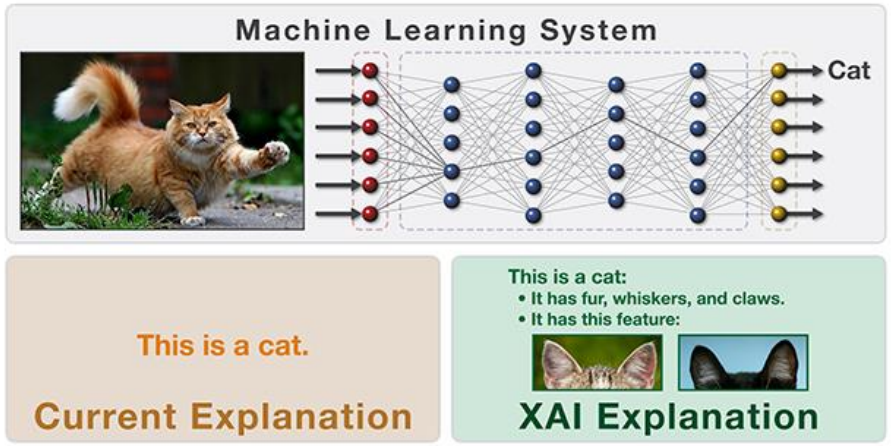
电商领域的场景化推荐

- 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

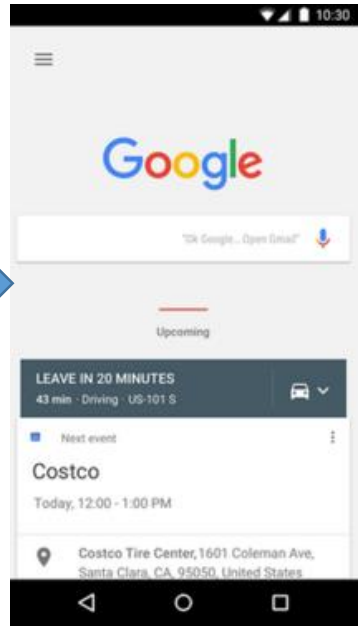


解释事实

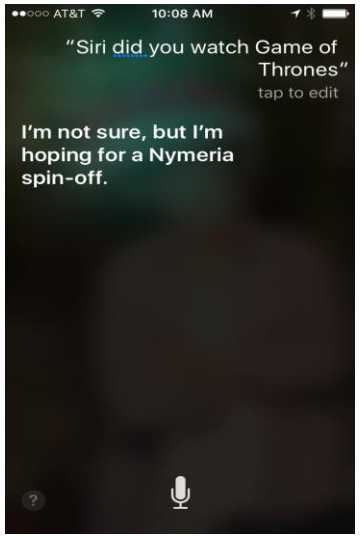
- 解释是智能的重要体现之一，将是人们对于智能系统的普遍期望
- 可解释是智能系统决策结果被采信的前提

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

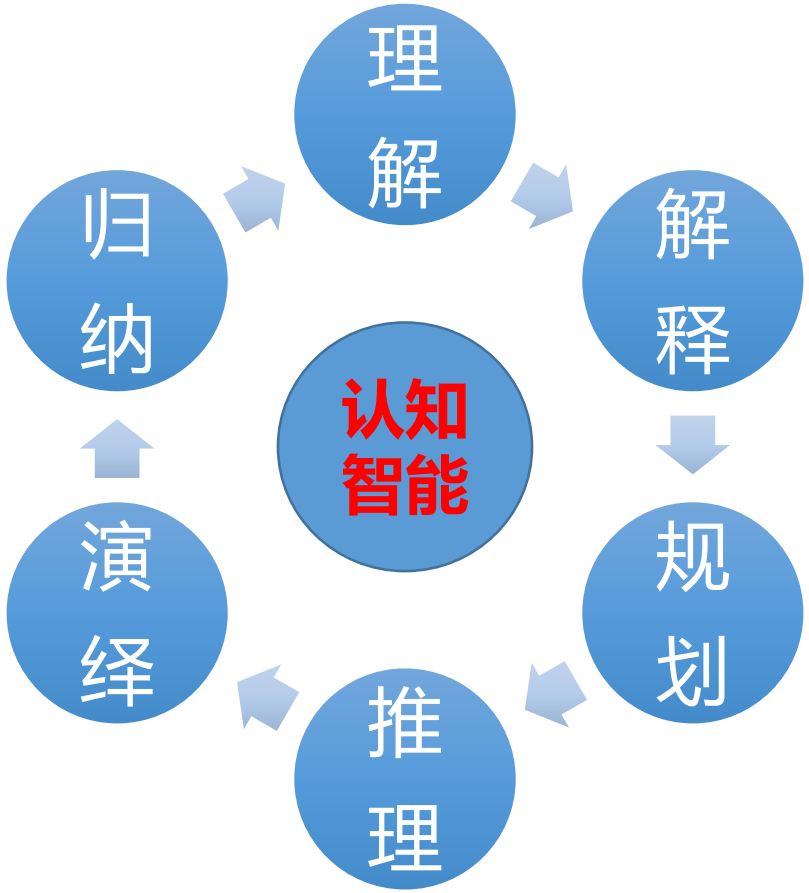
深层关系发现 / 推理

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

认知智能是智能化的关键



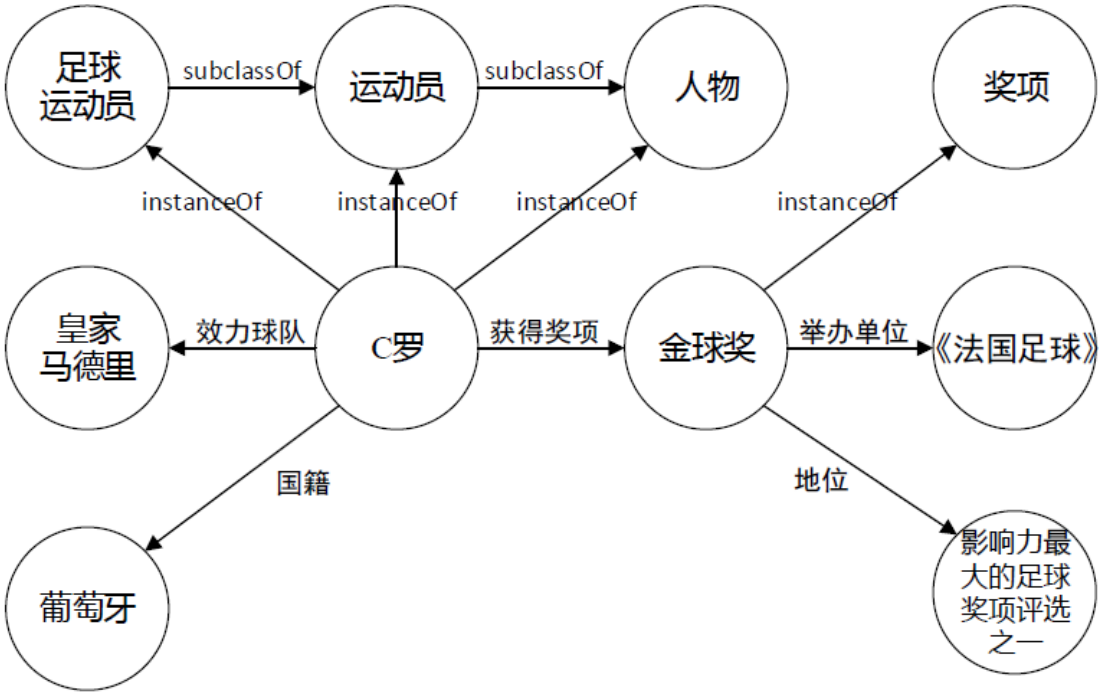
Can machine ***think like humans***?

- 理解与解释是后深度学习时代人工智能的核心使命之一

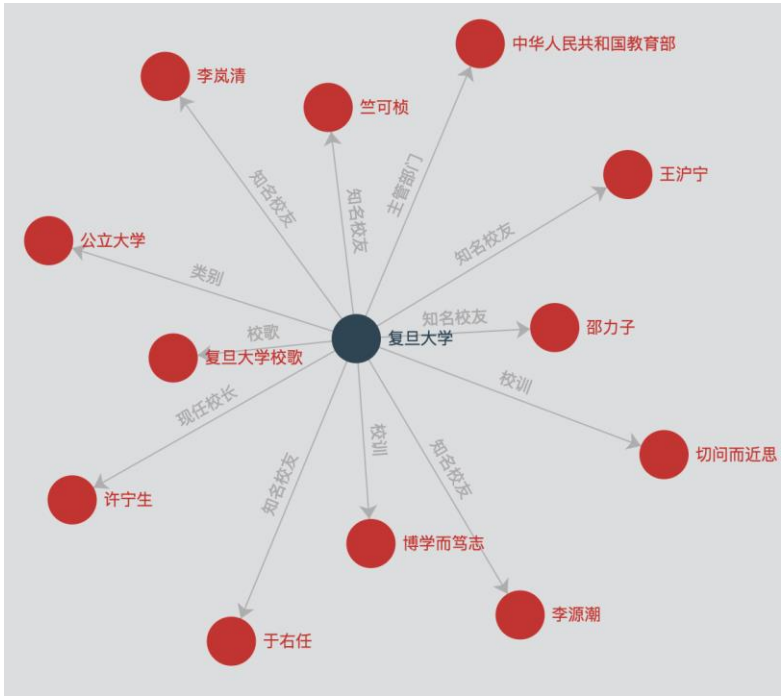
知识图谱是实现认知智能的关键技术，
是实现机器认知智能的使能器（Enabler）

知识图谱

- Knowledge Graph is a **large scale** semantic network
  - Consisting of entities/concepts as well as the semantic relationships among them



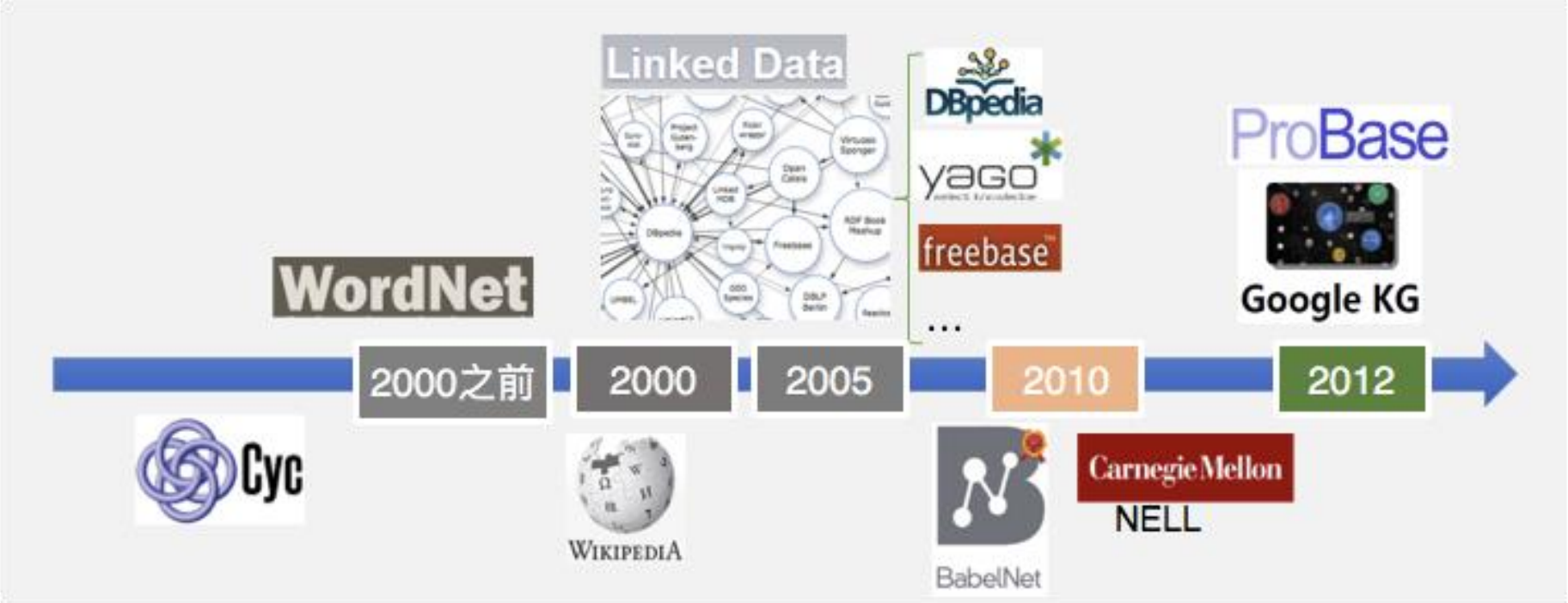
知识图谱富含实体、概念、属性、关系等信息，使得机器理解与解释成为可能



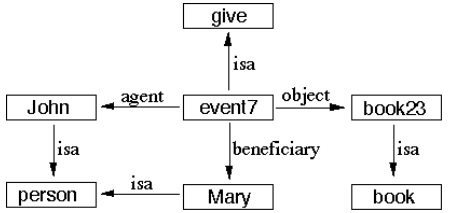
http://knowledgeworks.cn/cndbpedia/search

知识图谱的广义内涵

- 知识图谱作为一种语义网络，是大数据时代知识表示的重要方式之一
- 知识图谱作为一种技术体系，是大数据时代知识工程的代表性进展



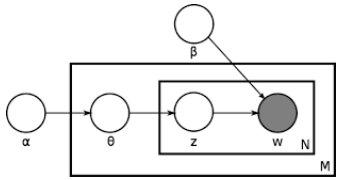
从“小”知识到“大”知识



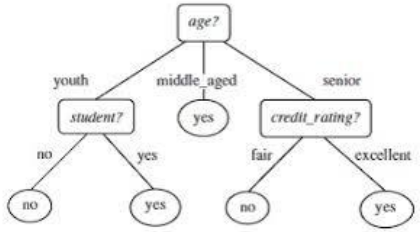
Ontology,Frame

$$\forall x \forall y (P(f(x)) \rightarrow \neg (P(x) \rightarrow Q(f(y), x, z)))$$

Logic rules



Bayesian network



Decision tree

- 传统知识工程，专家构建，代价高昂，规模有限；知识边界易于突破，难以适应大数据时代开放应用到规模化需求
- 大规模开放应用需要“大”知识（大规模知识库）

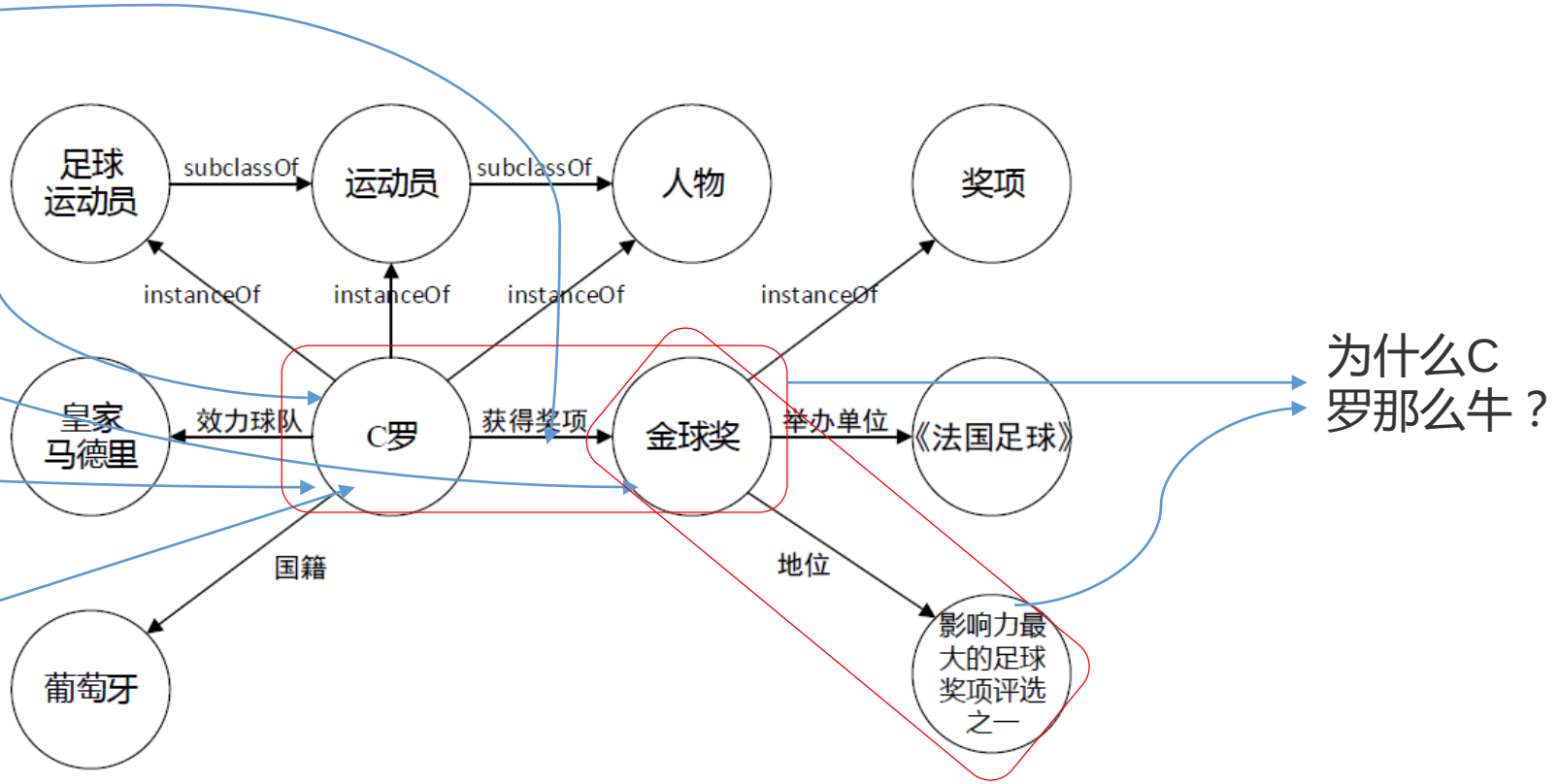
Big Knowledge

- Small knowledge+ big data= big knowledge
- 知识图谱引领知识工程复兴

基于知识图谱的认知智能

- 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

2013年的金球奖得主C罗



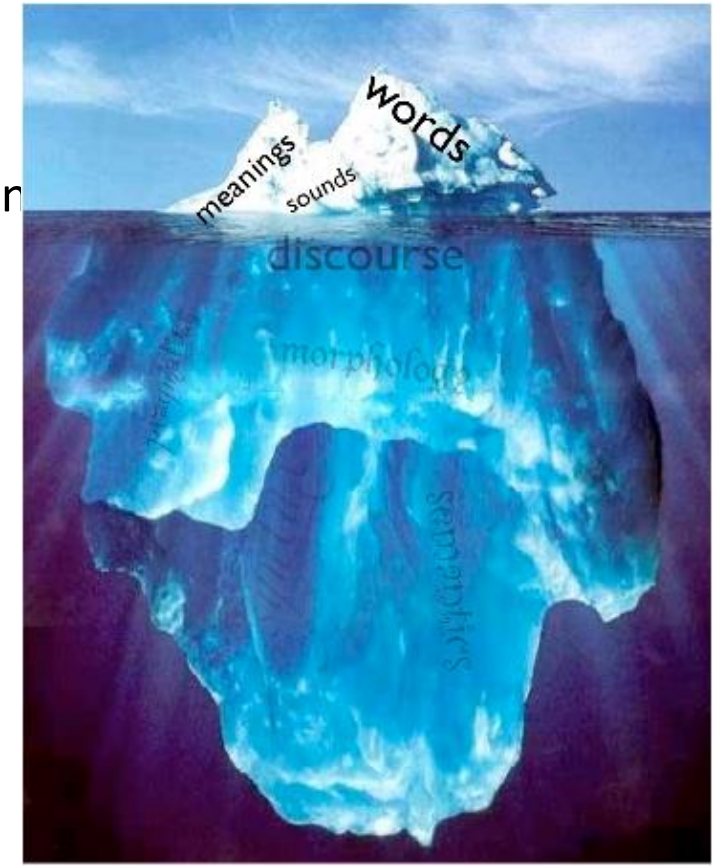
机器语言理解需要背景知识

Language is complicated

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

Language understanding is difficult

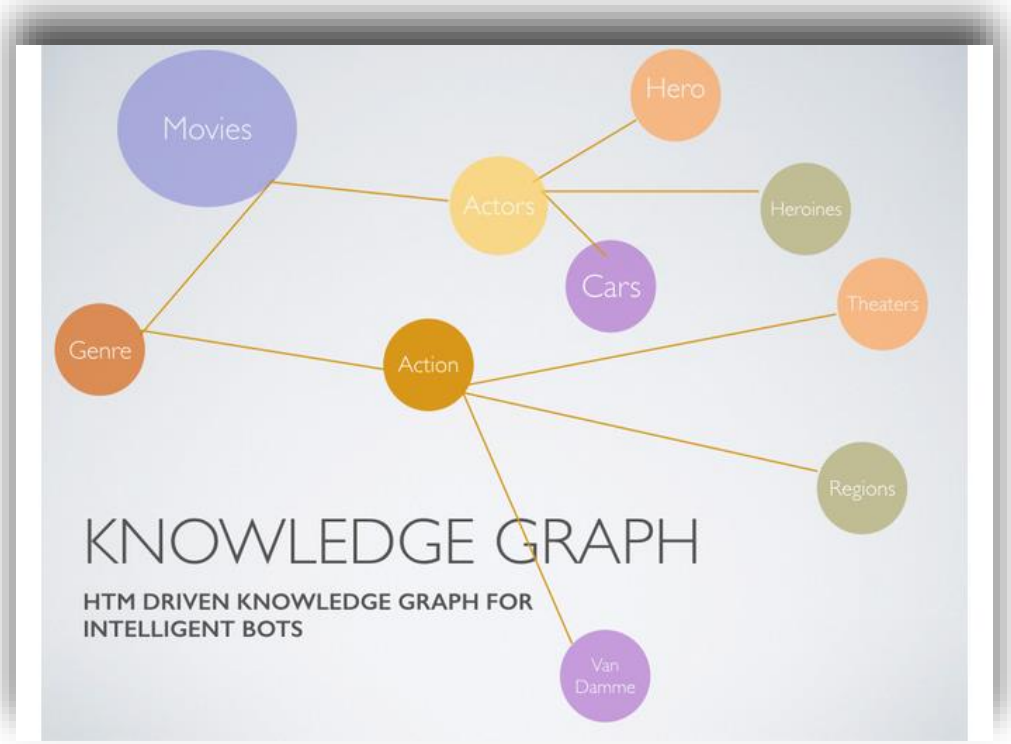
- Grounded only in **human cognition**
- Needs significant **background knowledge**



知识图谱*使能*(Enable)机器语言认知

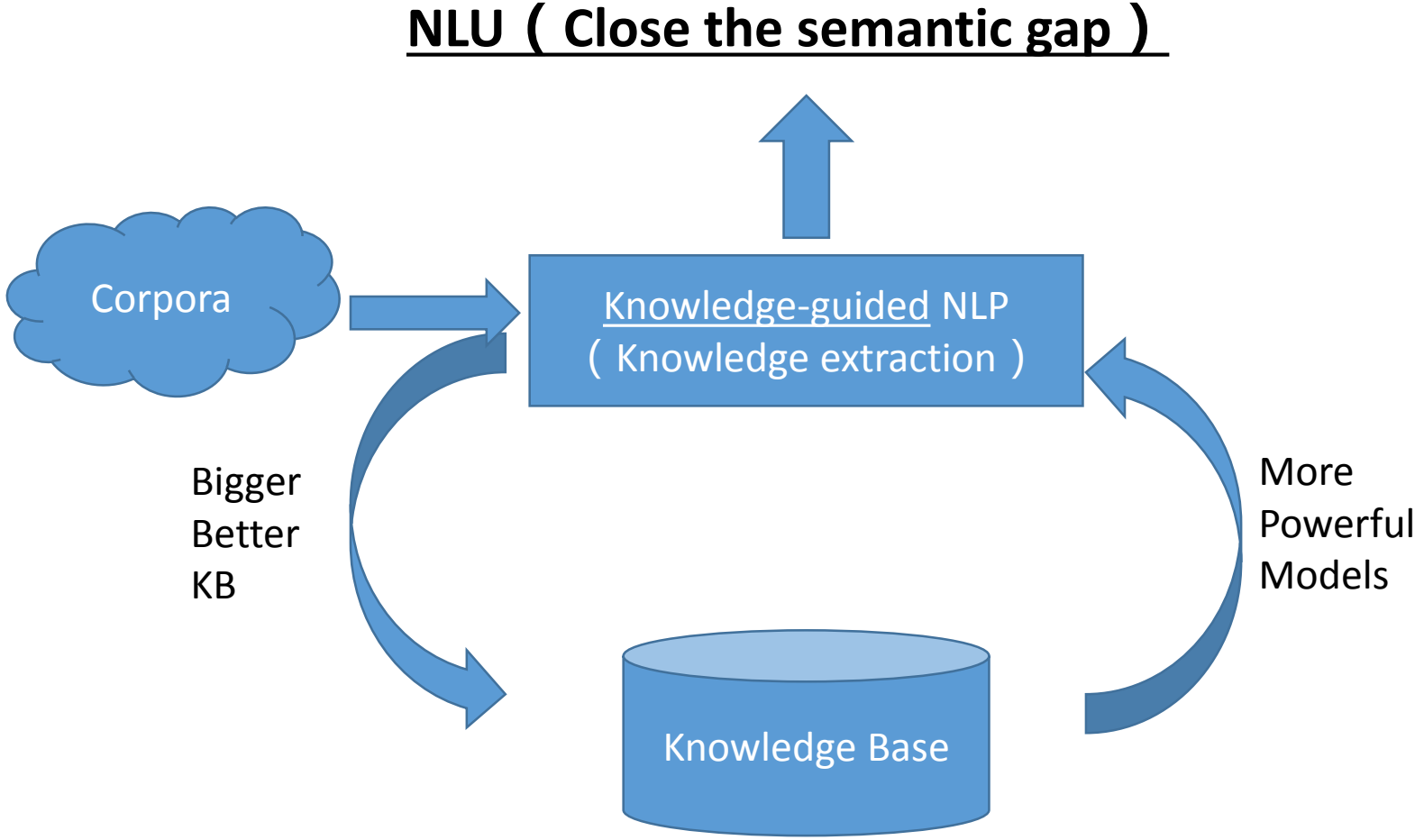
- Language understanding of machines needs knowledge bases
  - Large scale
  - Semantically rich
  - Friendly structure
  - High quality

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



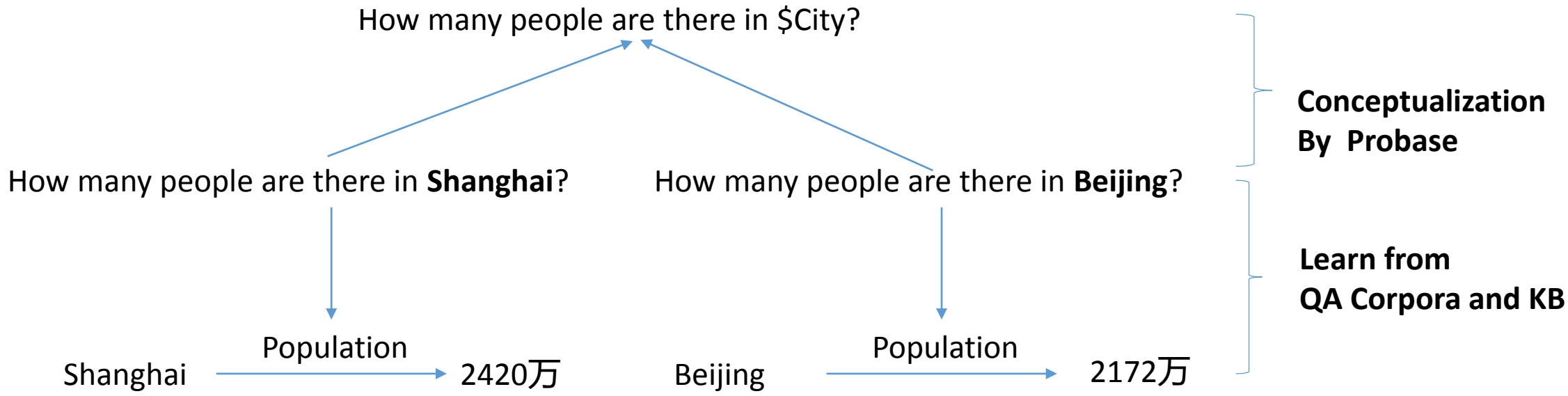
NLP+KB= NLU， NLP=Natural language processing， NLU=natural language understanding

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, Yanghua Xiao*, Haixun Wang, Yangqiu Song, Seung-won Hwang, Wei Wang, KBQA: Learning Question Answering over QA Corpora and Knowledge Bases, (**VLDB 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

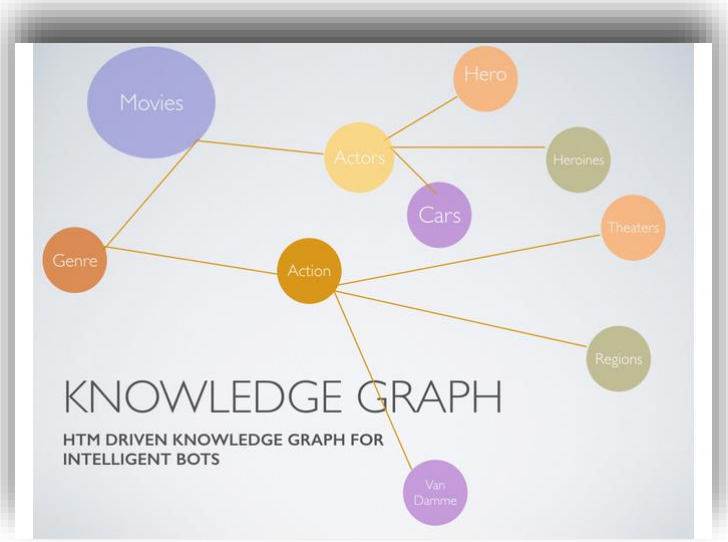
鸟儿为何能够飞翔？
因为它们有翅膀

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

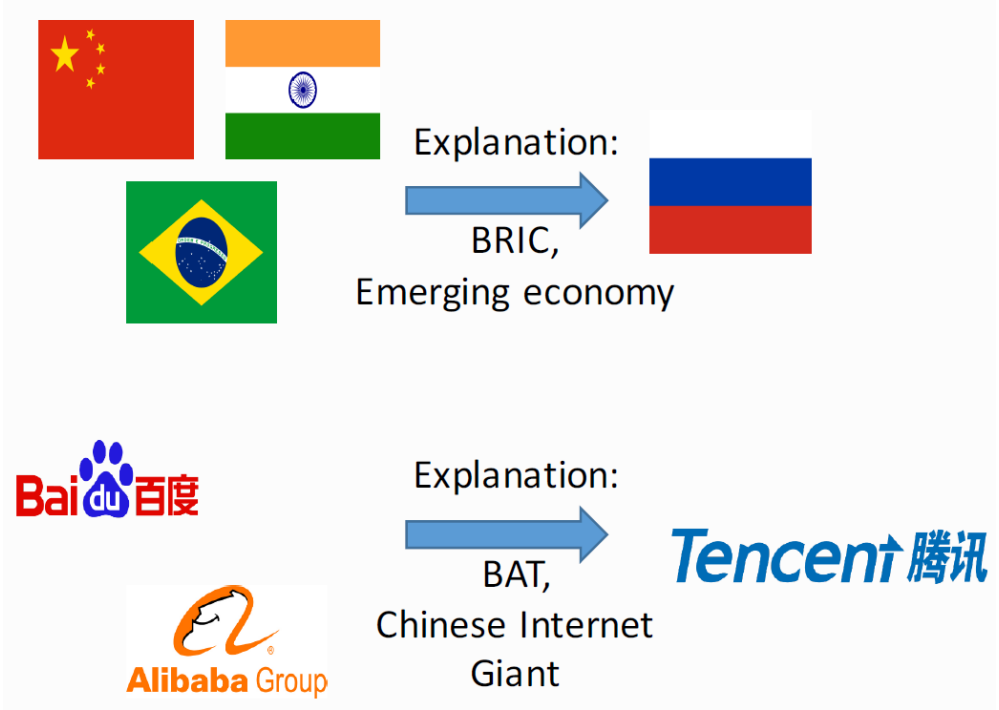
关系



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

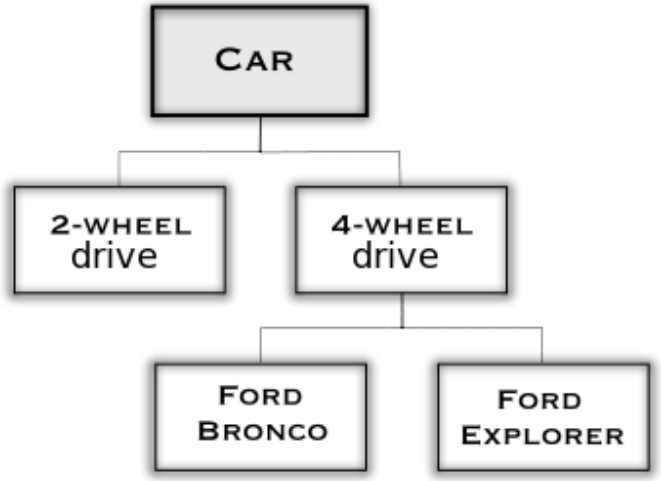
“Concepts are the glue that holds our mental world together”
--Gregory Murphy

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, Entity suggestion with conceptual explanation, (IJCAI 2017)

Example 2: Explain a Concept/Category using Properties

Problem:

How do we understand a concept/category?

Example：

How to understand “Bachelor”

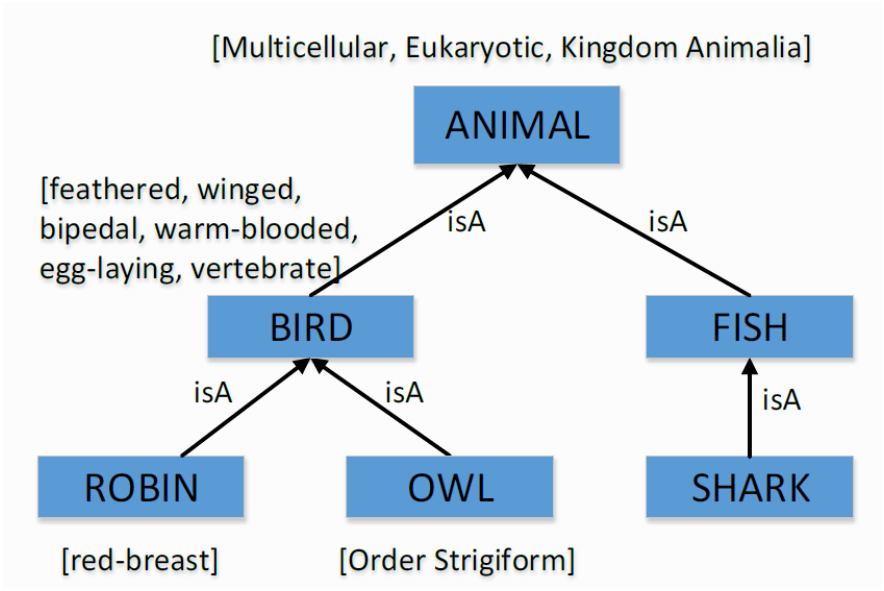
=> (Sex=man, Marriage status=unmarried)

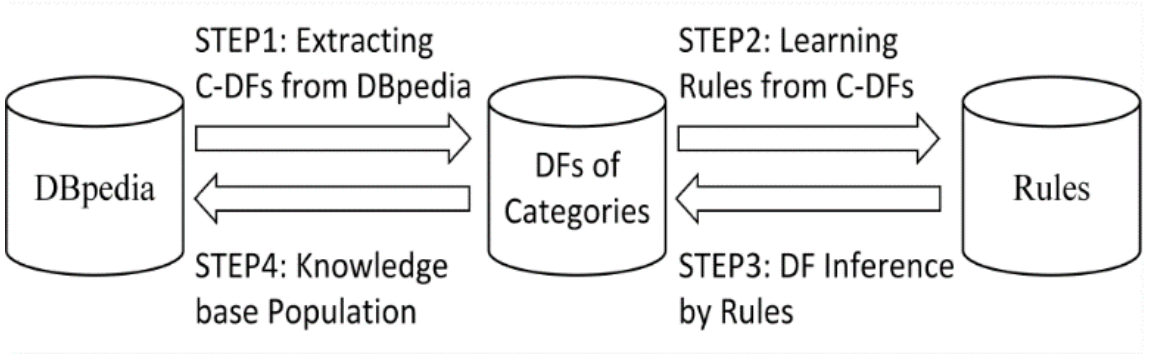
Basic Idea:

Ming Dbpedia, using properties to explain a category

Model:

Mining **Defining Features** from DBpeida

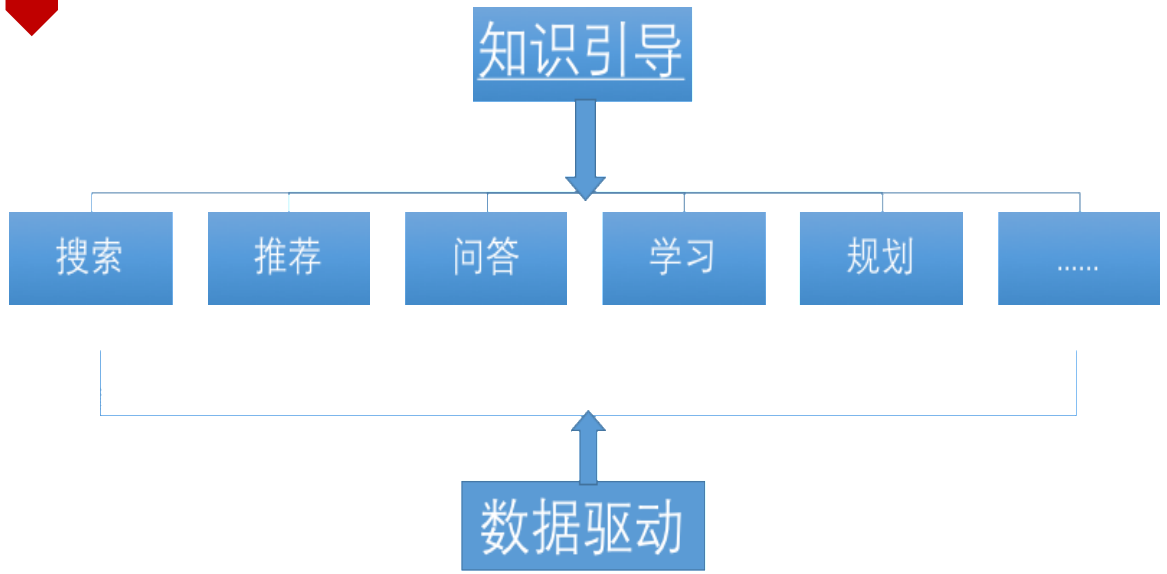


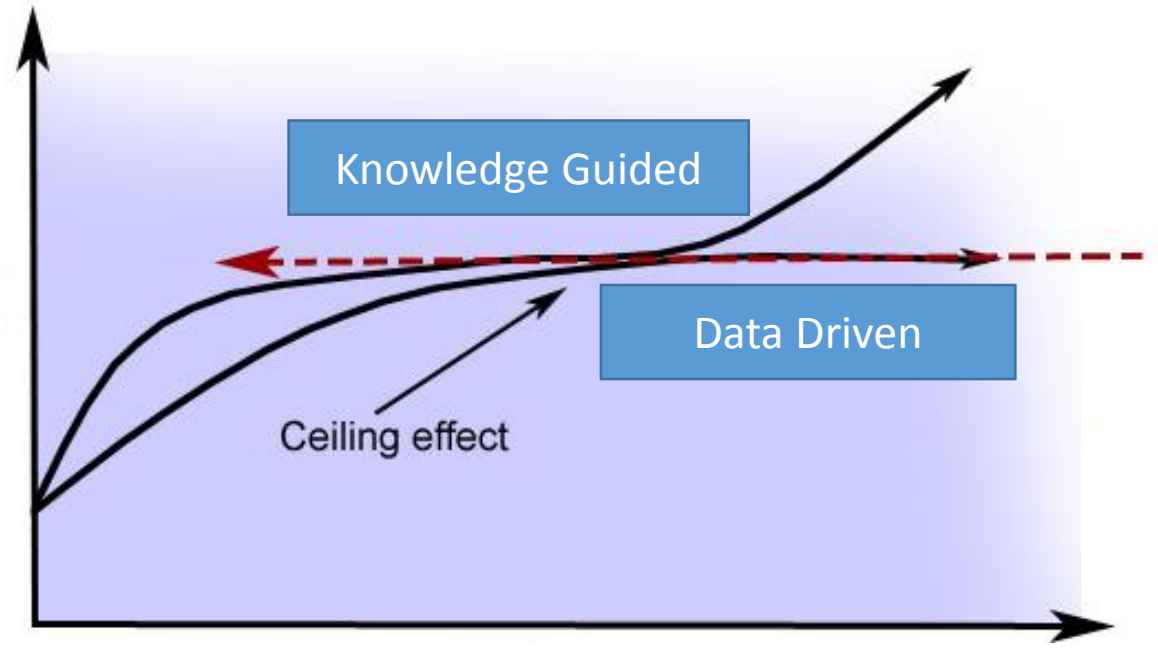


Solution Framework

Bo Xu, et al, Learning Defining Features for Categories. **(IJCAI 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

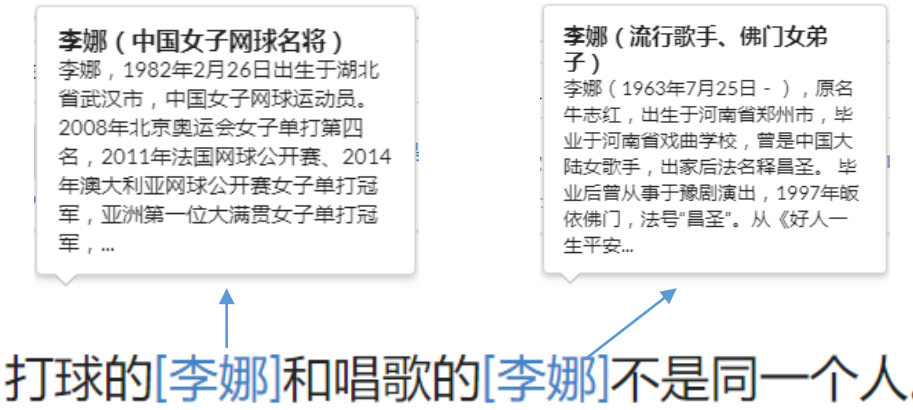
Example 1: Use Concepts for Chinese Entity Linking

- 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

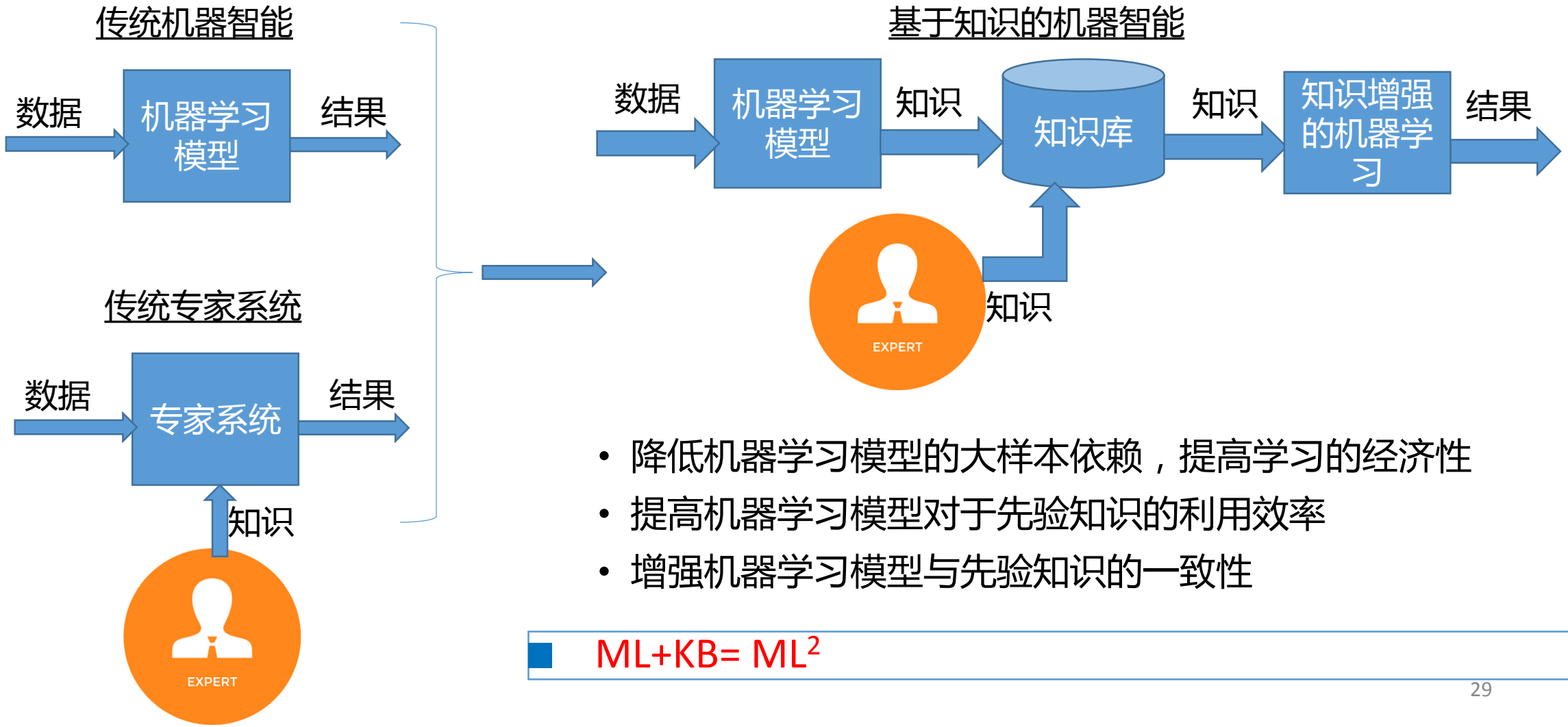
- Pattern based bootstrapping is popular
- Problem: semantic drift
  - <China isA country> =>
  - ‘occupation of $’, =>
  - ‘occupation of Planet earth’=>
  - <Planet Earch isA coutntry>
- Principles: no bad patterns, only wrong applications
- Our idea
  - Run a pattern on the text for an appropriate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

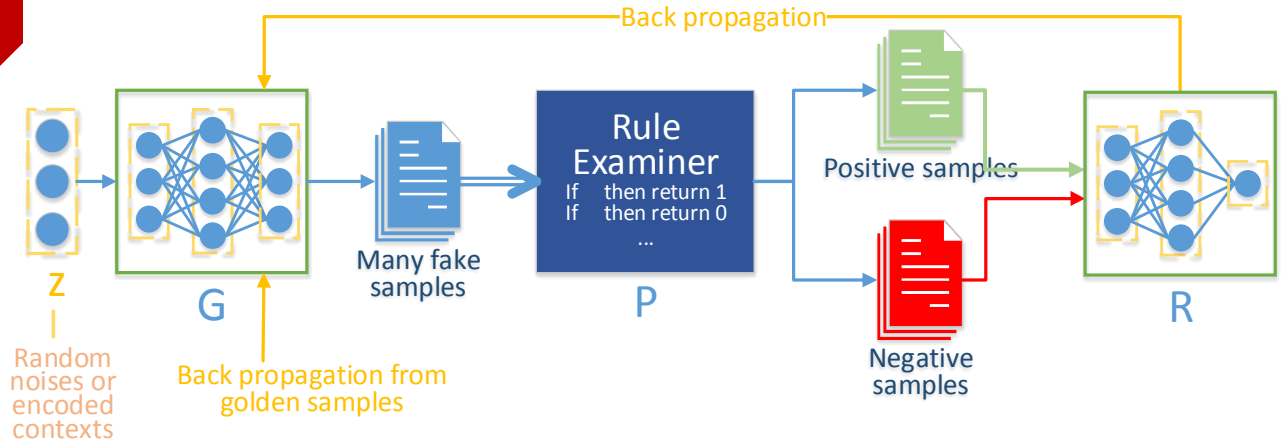
知识将显著增强机器学习能力



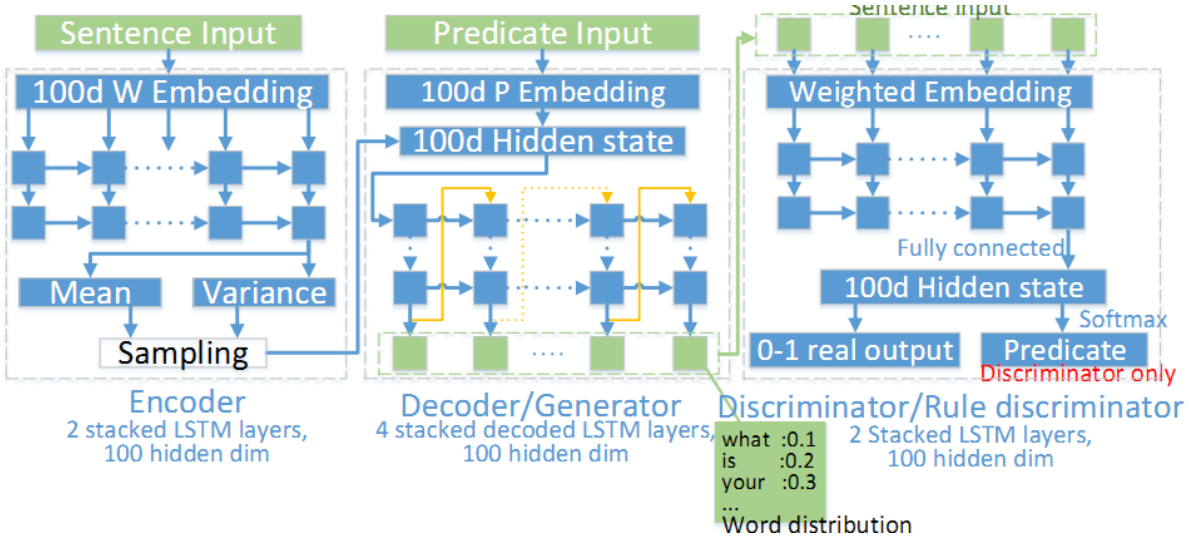
- 降低机器学习模型的大样本依赖，提高学习的经济性
- 提高机器学习模型对于先验知识的利用效率
- 增强机器学习模型与先验知识的一致性

$ML+KB= ML^2$

Example 1: Deep language generation with prior knowledge



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.



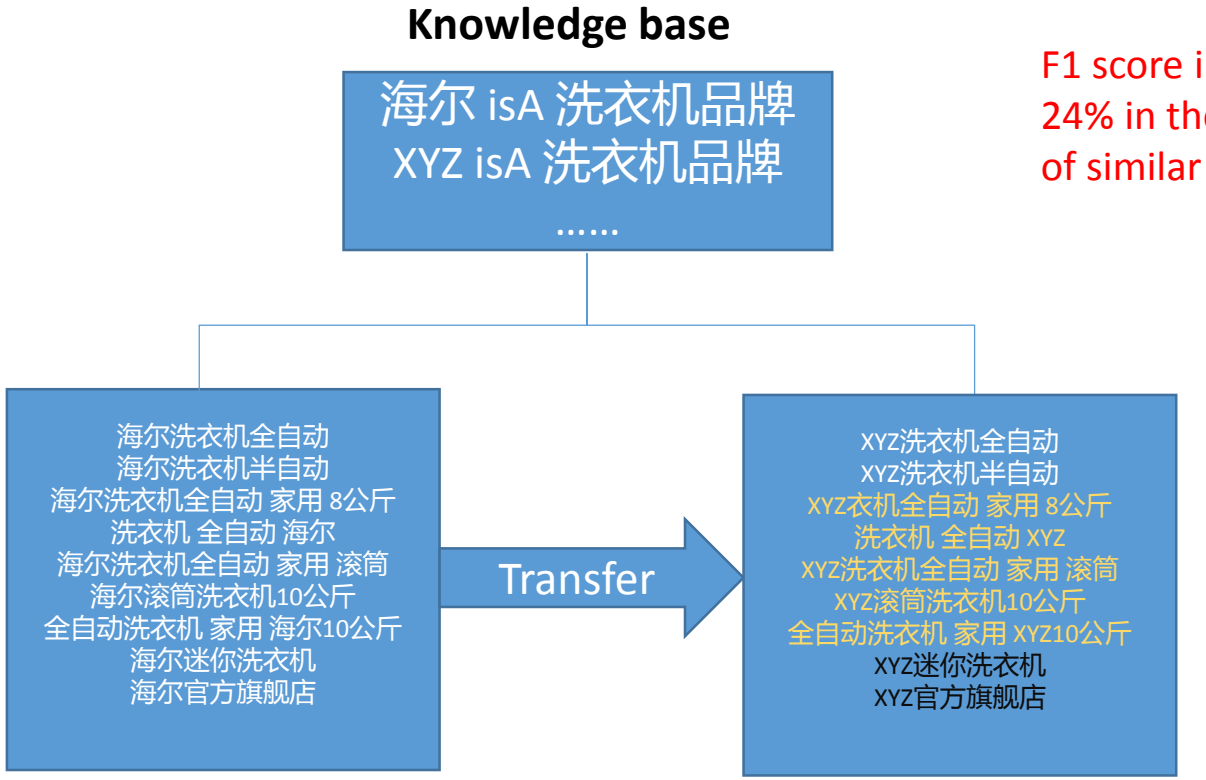
Incorporating Complicated Rules in Deep Generative Models, under review

在超级验证码中的应用

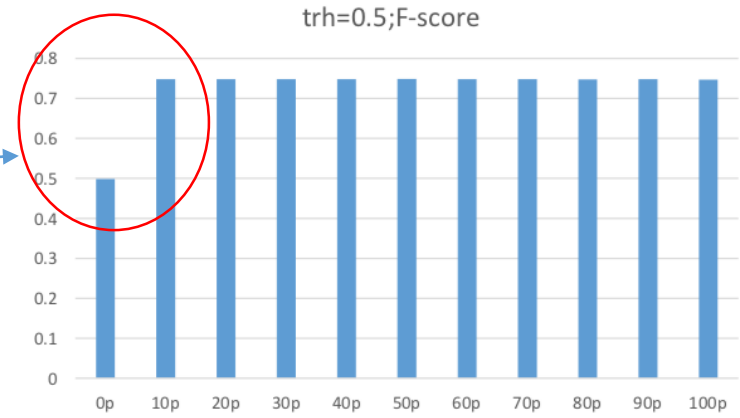
Demo地址：http://kw.fudan.edu.cn/ddemos/vcode/
API地址：http://kw.fudan.edu.cn/apis/supervcode/

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
- 经过知识沉淀的机器智能使得知识工作自动化成为可能
- 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

总结-1

- 人类社会已经进入智能时代
- 智能时代催生大量智能化应用需求
- 认知智能是实现行业智能化的关键技术
- 知识图谱技术是实现机器认知智能的核心技术
  - 知识图谱使能机器语言认知
  - 知识图谱使能可解释人工智能
  - 知识引导成为问题求解方式之一
  - 知识将显著增强机器学习能力

总结-2

- 定理1: NLP+KB=NLU
- 定理2: Small knowledge+ big data= big knowledge
- 定理3: ML+KB= ML^2

总结-3

知识的沉淀与传承，
铸就了人类文明的辉煌，
也将成为机器智能持续提升必经道路。

References

- Yi Zhang, Yanghua Xiao*,Seuongwon Hwang, Haixun Wang, X.Sean Wang, Entity Suggestion with Conceptual Explanation, (IJCAI2017)
- Jiaqing Liang, Sheng Zhang, Yanghua Xiao*, How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source, (IJCAI2017)
- Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui and Yanghua Xiao*, CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System, (IEA/AIE2017)
- Jiaqing Liang, Yi Zhang, Yanghua Xiao*, Haixun Wang, Wei Wang, Probase+: Inferring Missing Links in Conceptual Taxonomies, Transactions on Knowledge and Data Engineering (TKDE2017)
- Wanyun Cui, Yanghua Xiao*, Haixun Wang, Yangqiu Song, Seung-won Hwang, Wei Wang, KBQA: Learning Question Answering over QA Corpora and Knowledge Bases, (VLDB 2017)
- Jiaqing Liang, Yi Zhang, Yanghua Xiao*, Haixun Wang, Wei Wang and Pinpin Zhu, On the Transitivity of Hypernym-hyponym Relations in Data-Driven Lexical Taxonomies, (AAAI 2017)
- Jiaqing Liang, Yanghua Xiao*, Yi Zhang, Seung-Won Hwang and Haixun Wang, Graph-based Wrong IsA Relation Detection in a Large-scale Lexical Taxonomy, (AAAI 2017)
- Xiangyan Sun, Yanghua Xiao*, Haixun Wang, Wei Wang, On Conceptual Labeling of a Bag of Words, (IJCAI 2015)
- Xiangyan Sun, Haixun Wang, Yanghua Xiao*, Zhongyuan Wang, Syntactic Parsing of Web Queries, (EMNLP 2016)

References

- Bo Xu, Chenhao Xie, Yi Zhang, Yanghua Xiao*, Haixun Wang and Wei Wang, Learning Defining Features for Categories, (IJCAI 2016)
- Wanyun Cui, Yanghua Xiao*, Wei Wang, KBQA: An Online Template Based Question Answering System over Freebase, (IJCAI 2016)
- Wanyun Cui, Xiyou Zhou, Hangyu Lin,Yanghua Xiao*,Seungwon Hwang,Haixun Wang and Wei Wang, Verb Pattern: A Probabilistic Semantic Representation on Verbs, (AAAI 2016)
- Yi Zhang, Yanghua Xiao*,Seuongwon Hwang, Haixun Wang, X.Sean Wang, Entity Suggestion with Conceptual Expanation, (IJCAI2017)
- Deqing Yang, Yanghua Xiao*,Yangqiu Song, Wei Wang, Semantic-based Recommendation Across Heterogeneous Domains, (ICDM 2015)
- Deqing Yang, Yanghua Xiao*, Hanghang Tong, Wanyun Cui and Wei Wang, Towards Topic Following in Heterogeneous Information Networks, (ASONAM 2015)
- Xiangyan Sun, Yanghua Xiao*, Haixun Wang, Wei Wang, On Conceptual Labeling of a Bag of Words, (IJCAI 2015)
- Deqing Yang, Yanghua Xiao*, An Integrated Tag Recommendation Algorithm Towards Weibo User Profiling, (DASFAA 2015)
- Deqing Yang, Jingrui He, Huazhen Qin, Yanghua Xiao and Wei Wang, A Graph-based Recommendation across Heterogeneous Domains, (CIKM 2015)
- Jiaqing Liang, Yanghua Xiao, et,al. Tag inference with knowledge graph, (under review), Technique Report of KW@Fudan