



知识图谱研究的回顾与展望

肖仰华

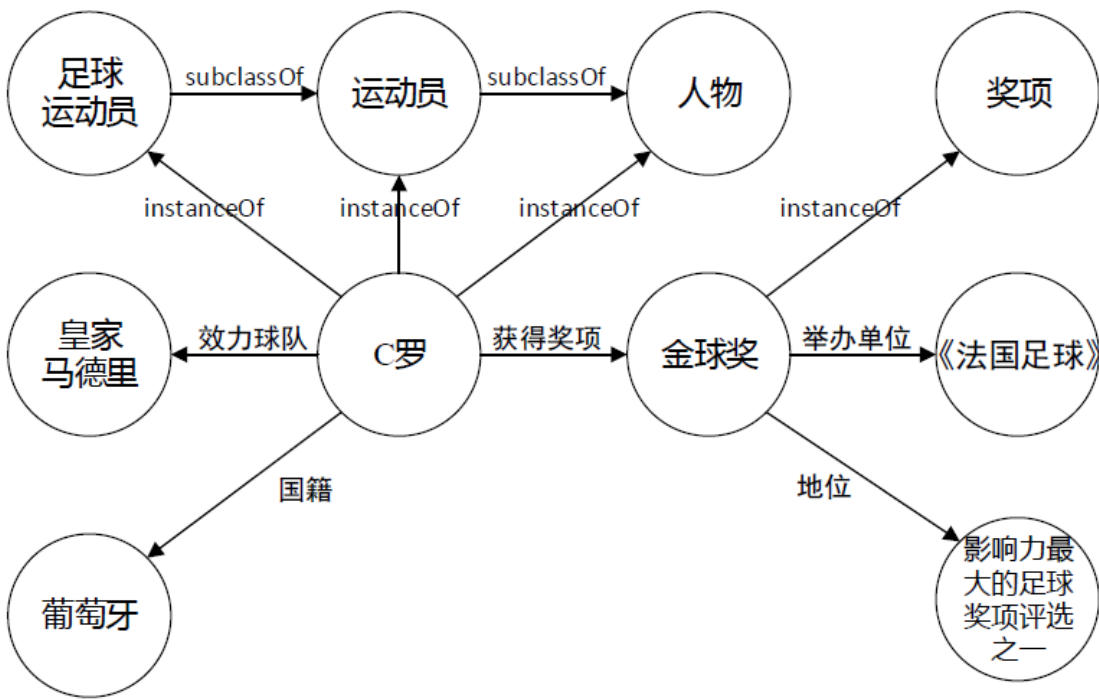
复旦大学知识工场实验室

shawyh@fudan.edu.cn

2017-09-23

Knowledge Graph

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



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

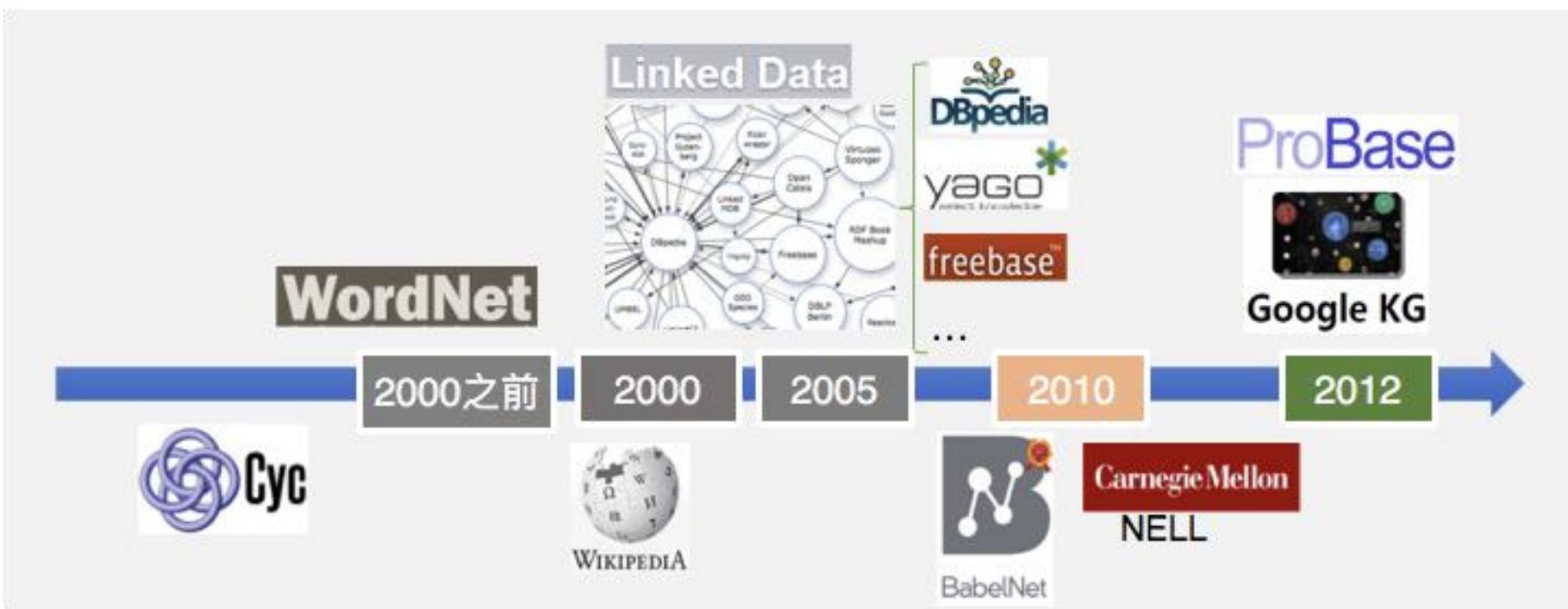
representation) is dedicated to ***representing information about the world*** in a form that a computer system can utilize to solve complex tasks such as diagnosing a medical condition or having a dialog in a ***natural language***.

KG (Knowledge

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

历史沿革

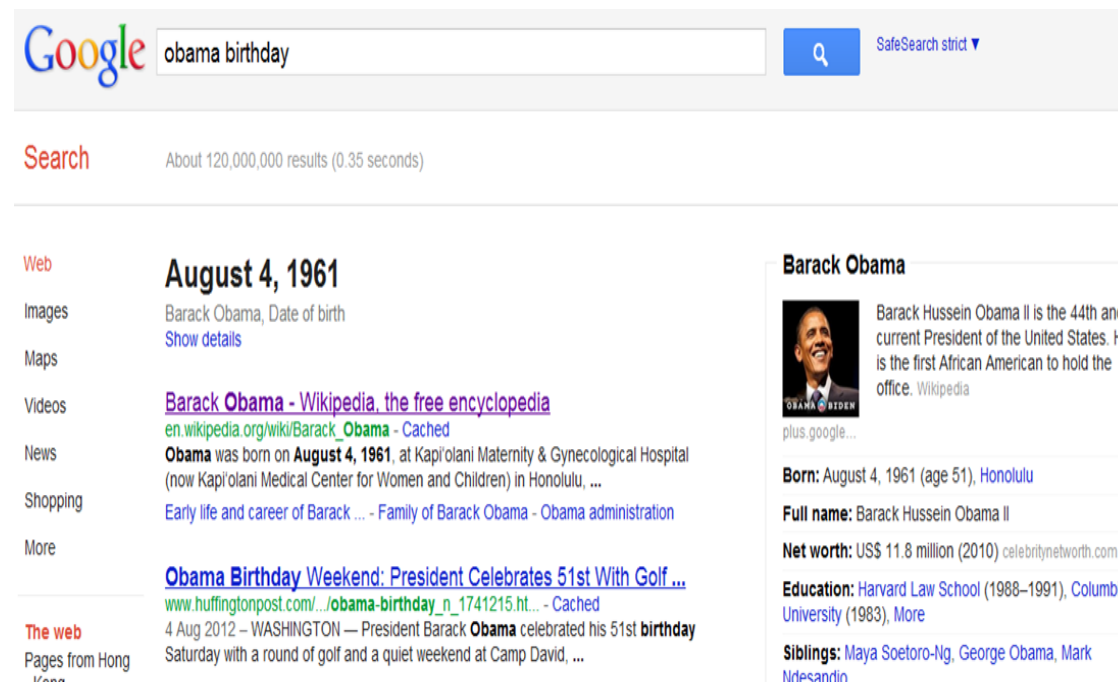
- 知识图谱作为一种语义网络，是大数据时代知识表示的重要方式之一
- 知识图谱作为一种技术体系，是大数据时代知识工程的代表性进展



诞生背景



- 2012年5月，Google正式发布知识图谱
- 搜索核心诉求：让搜索通往答案
 - 无法理解关键词
 - 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



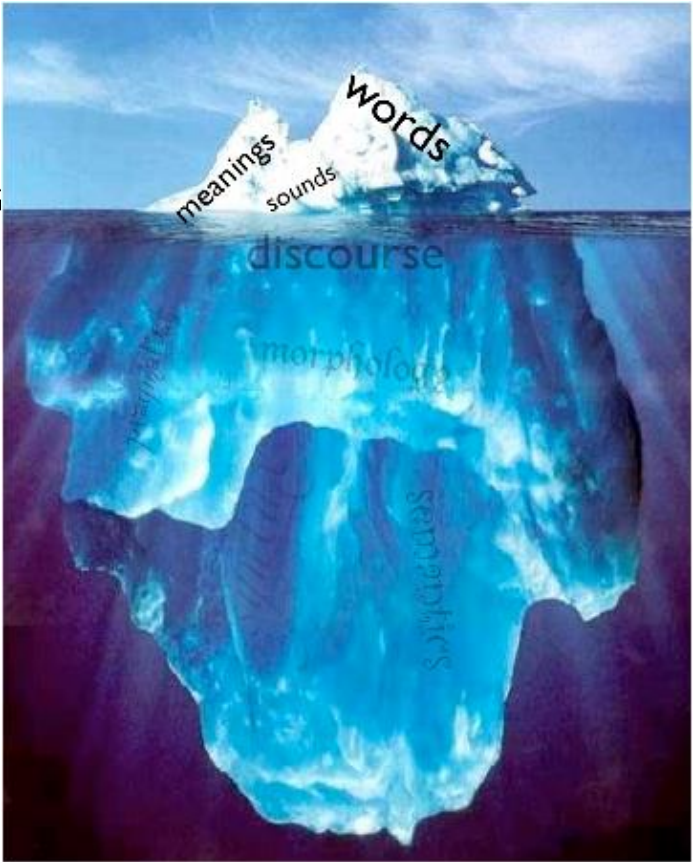
语言理解需要背景知识

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**



New *Frozen* Boutique to Open at *Disney's Hollywood Studios*



[/wiki/Frozen_\(2013_film\)](#)



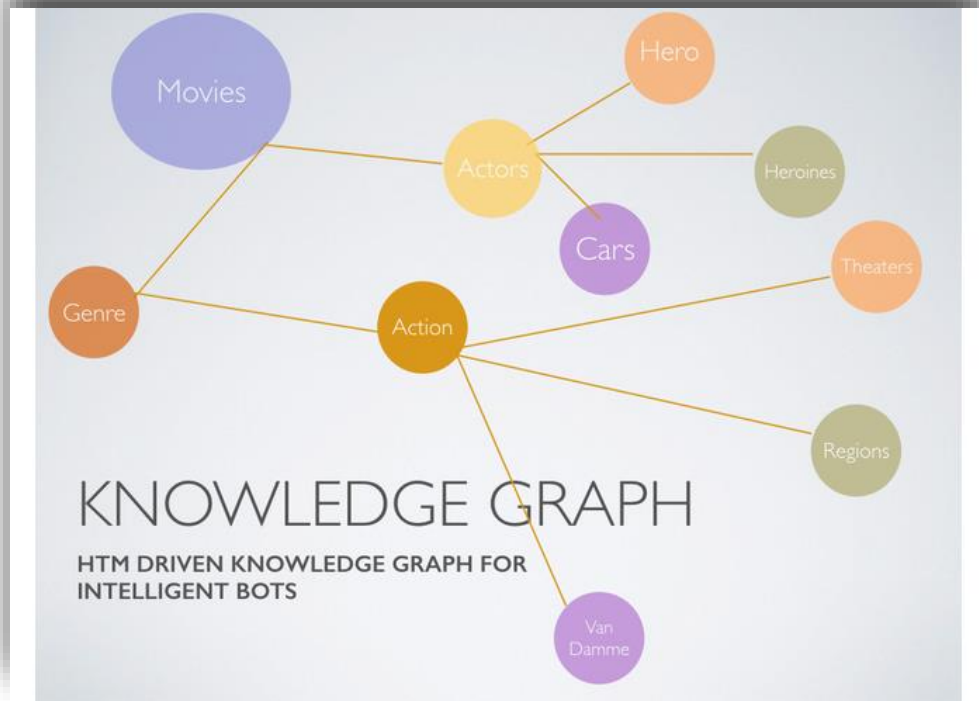
[/wiki/The_Walt_Disney_Company](#)



[/wiki/Disney's_Hollywood_Studios](#)

知识图谱 使能(Enable)机器语言认知

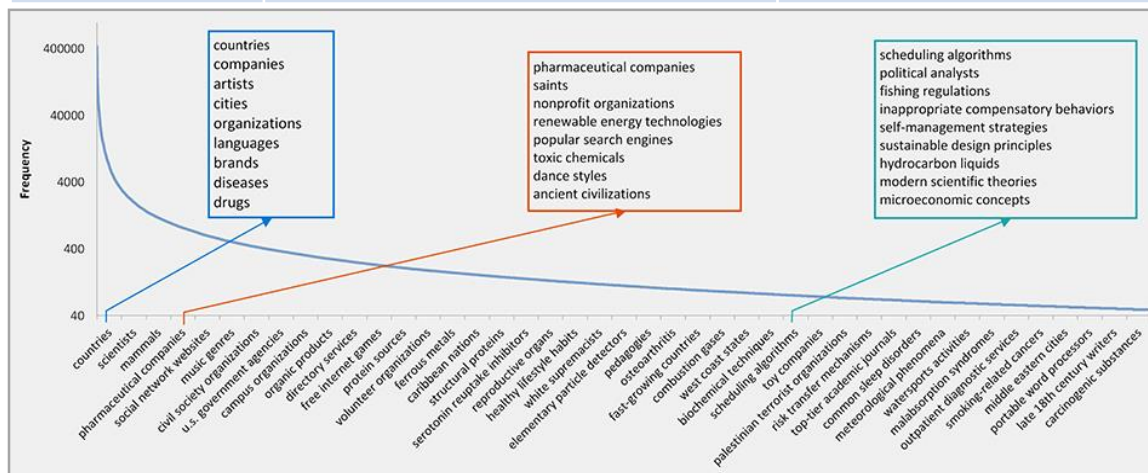
- Language understanding of machines needs knowledge bases
 - Large scale
 - Semantically rich
 - Friendly structure
 - High quality
- Traditional knowledge representations can not satisfy these requirements, but KG can
 - Ontology
 - Semantic network
 - Texts



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

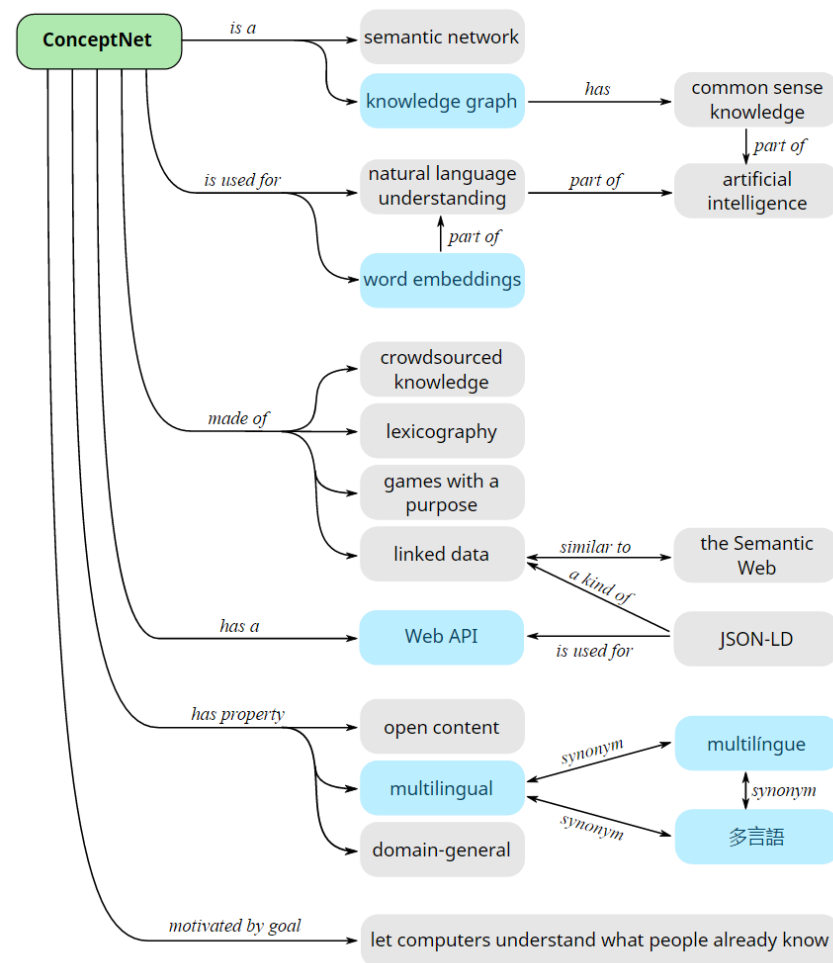


Existing Taxonomies	Number of Concepts
Freebase [5]	1,450
WordNet [13]	25,229
WikiTaxonomy [26]	111,654
YAGO [35]	352,297
DBPedia [1]	259
ResearchCyc [18]	≈ 120,000
KnowItAll [12]	N/A
TextRunner [2]	N/A
OMCS [31]	N/A
NELL [7]	123
Probase	2,653,872

KG优势2 : semantically rich

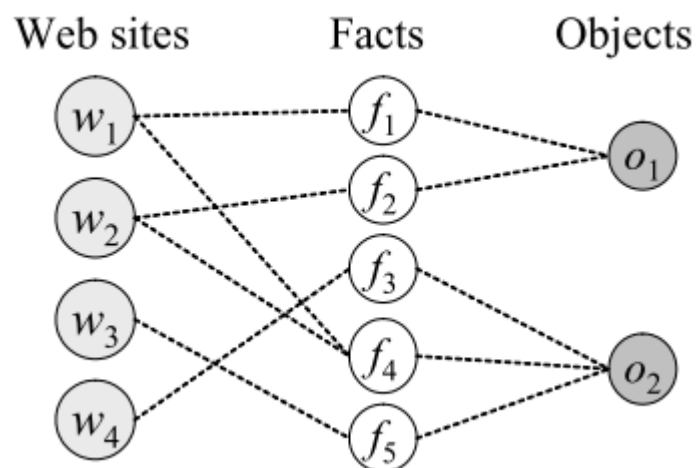
- Higher coverage over numerous semantic relationships

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, etc., Truth Discovery with Multiple Conflicting Information Providers on the Web, kdd07]



CN-DBpedia

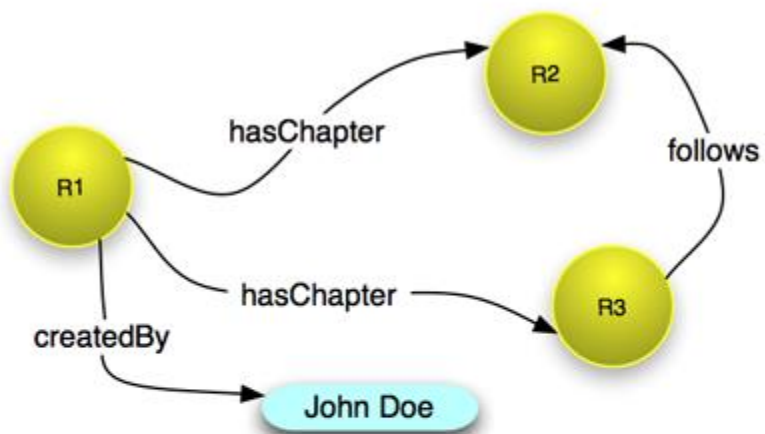
Q InfoBox

专职院士	25人	👍	👎
中文名	复旦大学	👍	👎
主管部门	中华人民共和国教育部	👍	👎
主要奖项	SCI论文单篇被引用次数全国第一	👍	👎
主要奖项	诺贝尔奖得主名誉教授10位	👍	👎

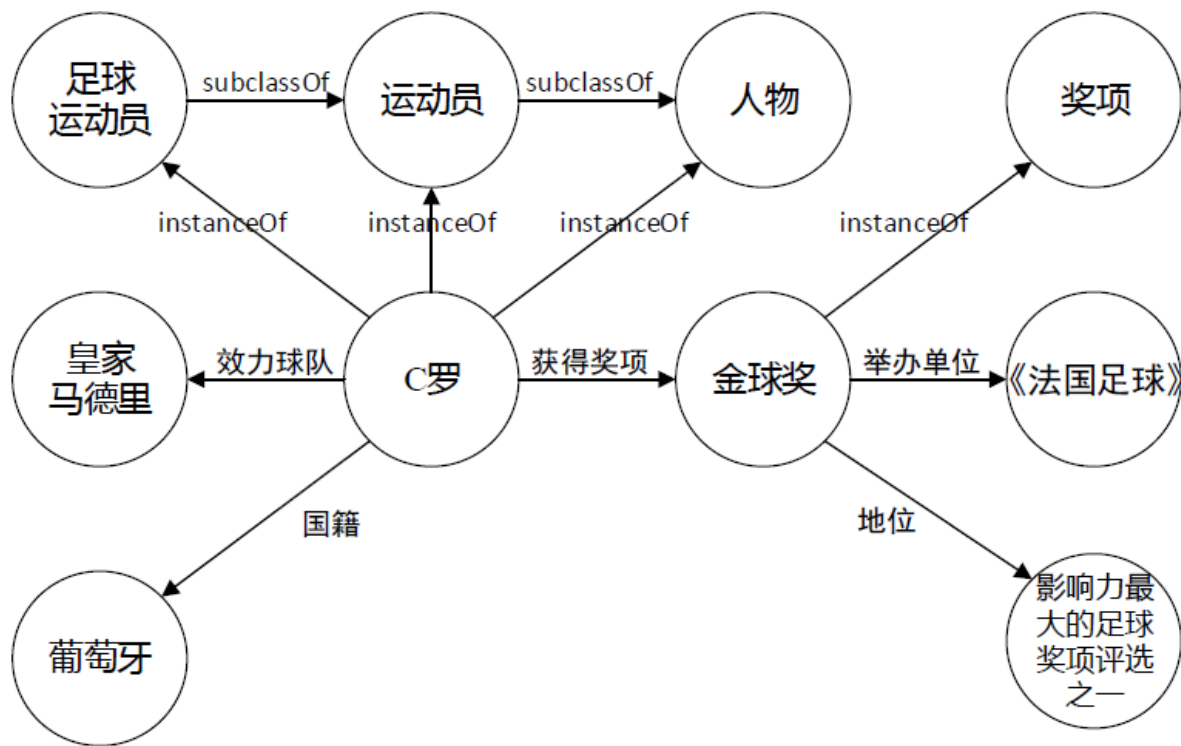
KG优势4：friendly structure

- Structured organization

- By RDF
- By graph



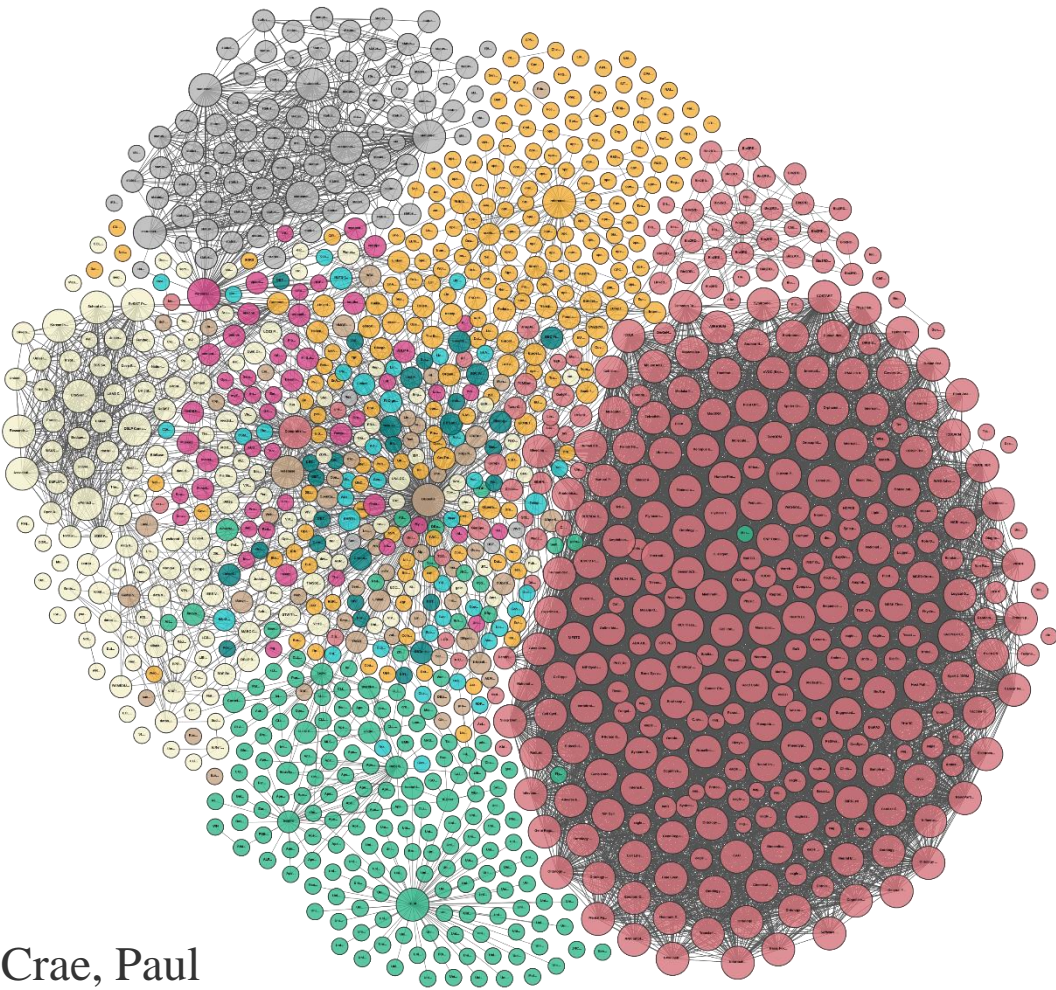
Subject	Predicate	Object
R1	hasChapter	R2
R1	hasChapter	R3
R3	follows	R2
R1	createdBy	"John Doe"



越来越多的知识图谱诞生

- 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/>"

变革

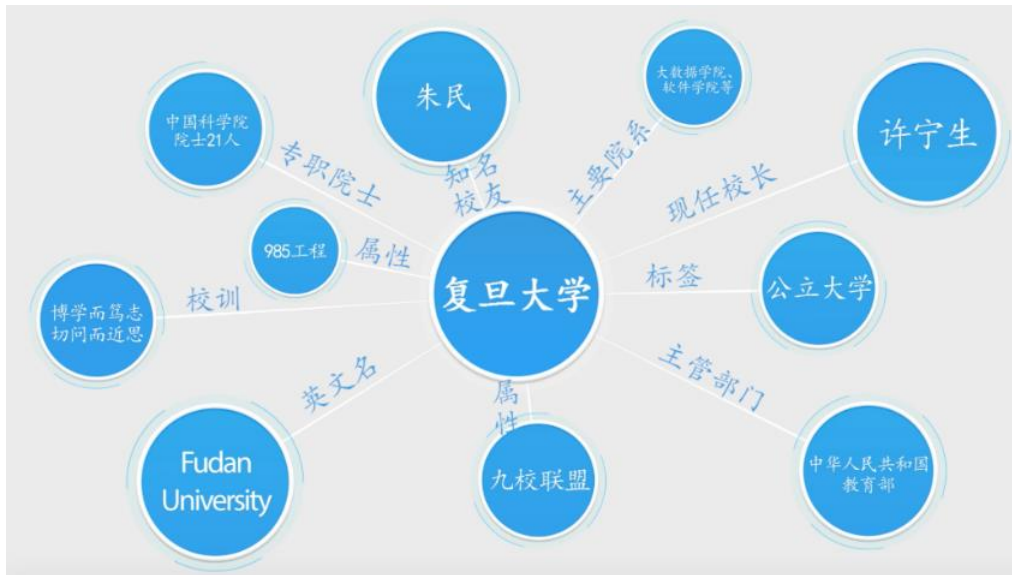
应用场景

- 从通用 vs 领域 / 行业应用
- 从搜索延伸至推荐、问答等复杂任务
- 从简单关系发现到深层关系推理
- 从回答what问题到回答why问题
- 从关键词交互到更自然的人机对话式交互

技术生态

- 机器学习
- 自然语言处理
- 知识图谱

从通用到领域 / 行业应用

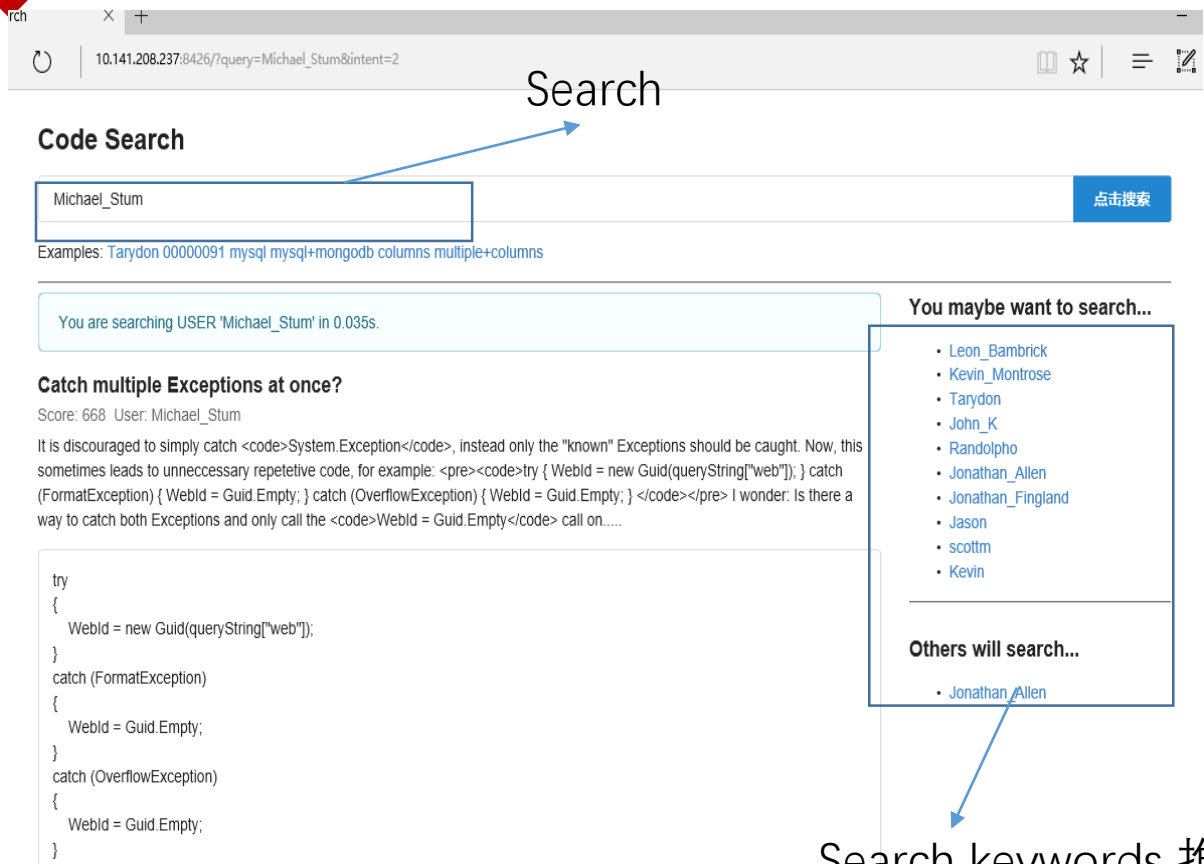


通用百科知识图谱：CN-DBpedia
累计API调用近3.7亿次



观点1：通用知识库在通用人工智能中扮演重要角色，是未来竞争的战略制高点，不容忽视
观点2：自动化大规模通用KG构建、样本稀疏环境下的自动化领域KG构建仍是瓶颈问题

从搜索延伸至推荐、问答等复杂任务



Search

Code Search

Michael_Stum

Examples: Tarydon 0000091 mysql mysql+mongodb columns multiple+columns

You are searching USER 'Michael_Stum' in 0.035s.

Catch multiple Exceptions at once?
Score: 668 User: Michael_Stum
It is discouraged to simply catch `System.Exception`, instead only the "known" Exceptions should be caught. Now, this sometimes leads to unnecessary repetitive code, for example:

```
<pre><code>try { WebId = new Guid(queryString["web"]); } catch (FormatException) { WebId = Guid.Empty; } catch (OverflowException) { WebId = Guid.Empty; } </code></pre> I wonder: Is there a way to catch both Exceptions and only call the WebId = Guid.Empty call on.....
```

You maybe want to search...

- Leon_Bambrick
- Kevin_Montrose
- Tarydon
- John_K
- Randolpho
- Jonathan_Allen
- Jonathan_Fingland
- Jason
- scottm
- Kevin

Others will search...

- Jonathan_Allen

Search keywords 推荐



小Cui问答

15:37

复旦大学在哪里

上海市杨浦区邯郸路220号

fudan的校训是什么

切问而近思, 博学而笃志

复旦校长是谁

许宁生

复旦是哪一年成立的

1905年09月14日

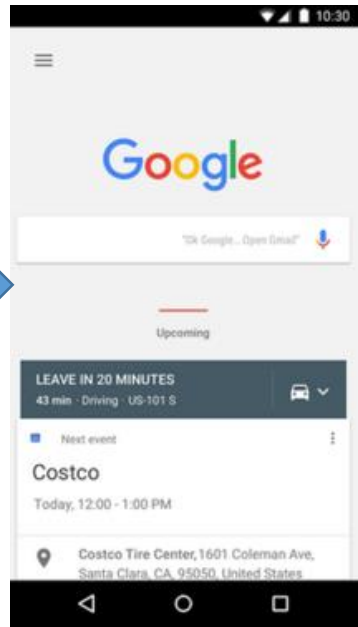
知识问答 (小Cui问答)

观点：KG将在更多复杂、多元任务中发挥重要作用

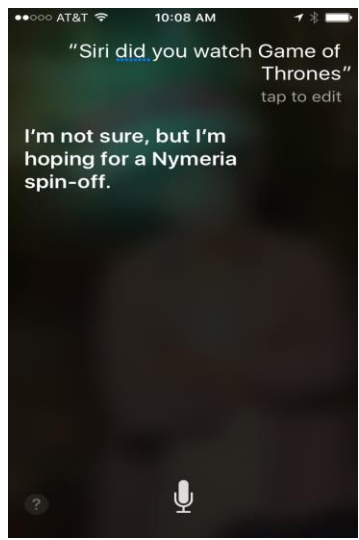
从关键词交互到更自然的人机对话式交互



Google Now



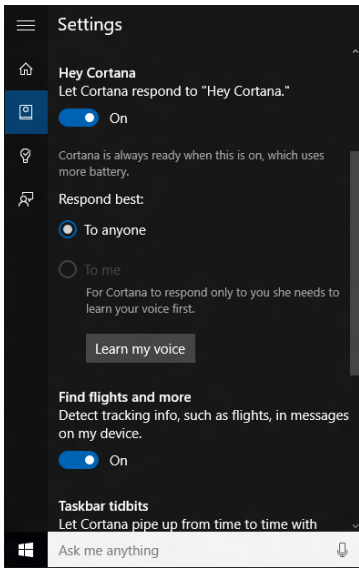
Apple Siri



Amazon Alexa



Microsoft Cortana



Question Answering (QA) systems in industries

观点：自然语言交互将成为人机交互的主流方式，将对机器自然语言认知提出更高要求

从回答what问题到回答why问题



What, Who, When



How, Why

观点：可解释是未来人工智能发展的核心诉求之一，是人机互信的前提

从简单关系发现到深层关系推理



Why baoqiang select Qizhun Zhang as his lawyer?

Why A invests B?

观点： 隐式关系发现、深层关系推理将成为智能的主要体现之一

技术生态



机器学习

- 深度学习飞速发展，在数据丰富的场景与任务下取得较好效果

机器学习

- 小样本学习、无监督学习手段有限；现有模型难以有效利用大量先验知识

自然语言处理

- 自然语言处理（NLP）的模型与算法，在深度学习模型推动下，发展迅速

自然语言处理

- 总体上离实际应用需求还很远；NLU（自然语言理解）还很初步

知识图谱

- 英文图谱积累迅速，大量高质量手工构建图谱已在应用中发挥作用

知识图谱

- 其他语种图谱缺乏、常识缺乏；有效利用大规模数据驱动的知识图谱手段缺乏

- 知识图谱构建的有效策略与方案
 - 如何充分利用知识的跨语言特性，如何区别对待数据来源的不同结构成程度，基于概念模板的迭代式抽取；基于语义与语法混合模式的抽取
 - 《Data Driven Approaches for Large Scale Knowledge Graph Construction》
- 大规模常识的获取与理解
 - 当前人工智能普遍缺乏常识理解能力，常识缺乏是人工智能研究的重大制约瓶颈。人工智能的发展必须尽快突破这一瓶颈。常识理解是通用人工智能的核心问题，是人工智能发展的战略制高点。为此，必须尽快开展大规模常识获取与理解的研究。
 - 《Large Scale Commonsense Knowledge Acquisition and Understanding》
- 样本稀疏环境下的领域知识获取
 - 有着丰富样本的高频知识的获取能够通过常规统计学习模型有效解决。但是出现频次较低的长尾知识的获取一直是知识工程领域的困难问题。如何将高频知识的获取模型有效迁移到样本稀疏的低频知识，是当前知识获取领域面临的重大难题。
 - 《电商智能化与知识图谱》

- 数据驱动与知识引导深度融合的新型机器学习模型
 - 如何将符号化知识有机融入基于数据的统计学习模型，是当前人工智能的重大问题，是降低深度学习模型的样本依赖，突破机器学习模型效果的天花板的关键所在。
 - 《当知识图谱遇见深度学习》，中国人工智能通信
- 基于知识图谱的可解释人工智能
 - 符号化的知识图谱具有形象直观的特性，为弥补深度学习在解释性方面的缺陷提供了可能。如何利用知识图谱解释各类机器学习，特别深度学习，以及高层次决策模型的结果，是当前人工智能领域的重大科学问题。
 - 《Explainable AI using Knowledge Graphs》
- 知识获取中的人机协作机制与方法
 - 当前人工智能的发展还需要人类的有效指导。人在环中是人工智能发展的基本模式。然而如何有效利用人力，如何有效切分人机合作的边界仍是知识获取过程中人机协作机制的核心问题。
 - 《未来人机区分》

- 知识驱动的机器语言理解

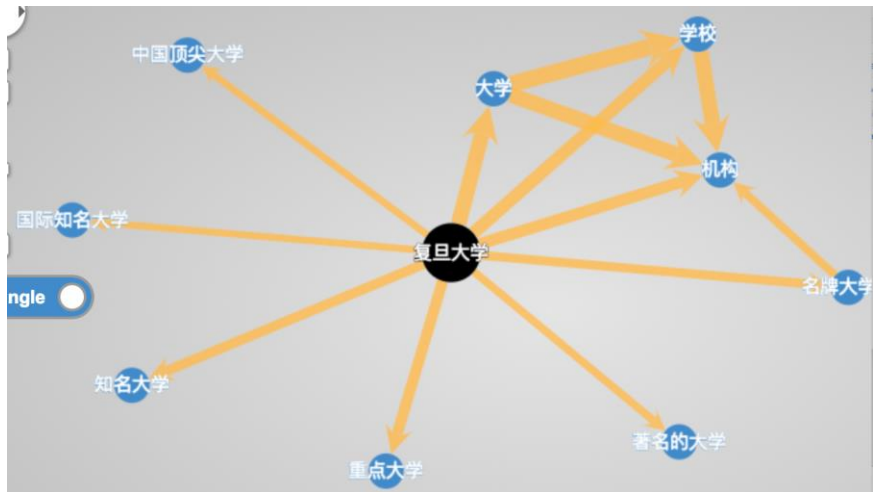
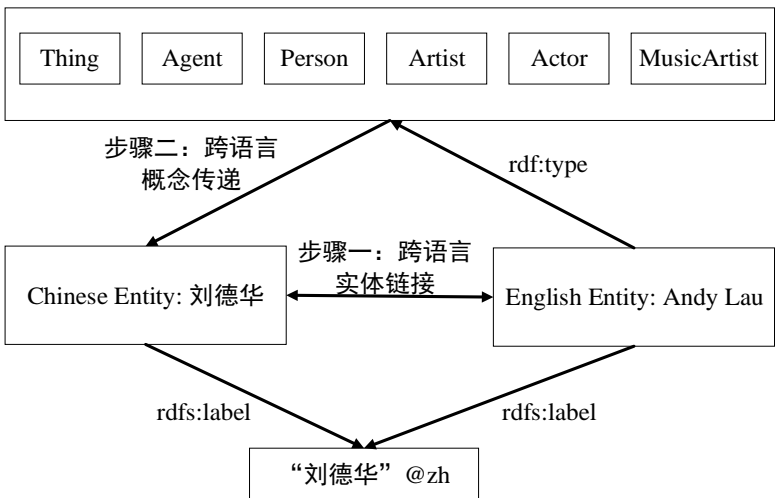
- 机器语言理解构建在机器认知能力基础之上的。大量知识库的涌现为知识驱动下的机器认知能力的实现提供了新的机遇，为机器语言理解提供了大量背景知识，使得机器理解人类语言成为可能。为此，必须深入开展知识驱动的机器语言理解模型与方法的研究
- 《Enabling Machines to Understand Human Language》

- 知识驱动的搜索与推荐

- 在应用层面，大量知识库的出现为互联网应用的两个核心问题搜索与推荐带来全新的机遇。如何有效利用符号知识理解搜索意图，改进检索效果，提升推荐精准性是当前互联网应用的核心问题。
- 《Knowledgeable Search and Recommendation》
- 《User Understanding with Knowledge Graph》

知识图谱构建的有效策略与方案

- 充分利用知识的跨语言特性，但也要注意知识的本地文化特性



Q Type

rdf:type	<http://dbpedia.org/ontology/Organisation>
rdf:type	<http://dbpedia.org/ontology/EducationalInstitution>
rdf:type	<http://dbpedia.org/ontology/University>
rdf:type	<http://dbpedia.org/ontology/Agent>

Principle : knowledge is independent on languages, but its expression in a specific language is usually biased

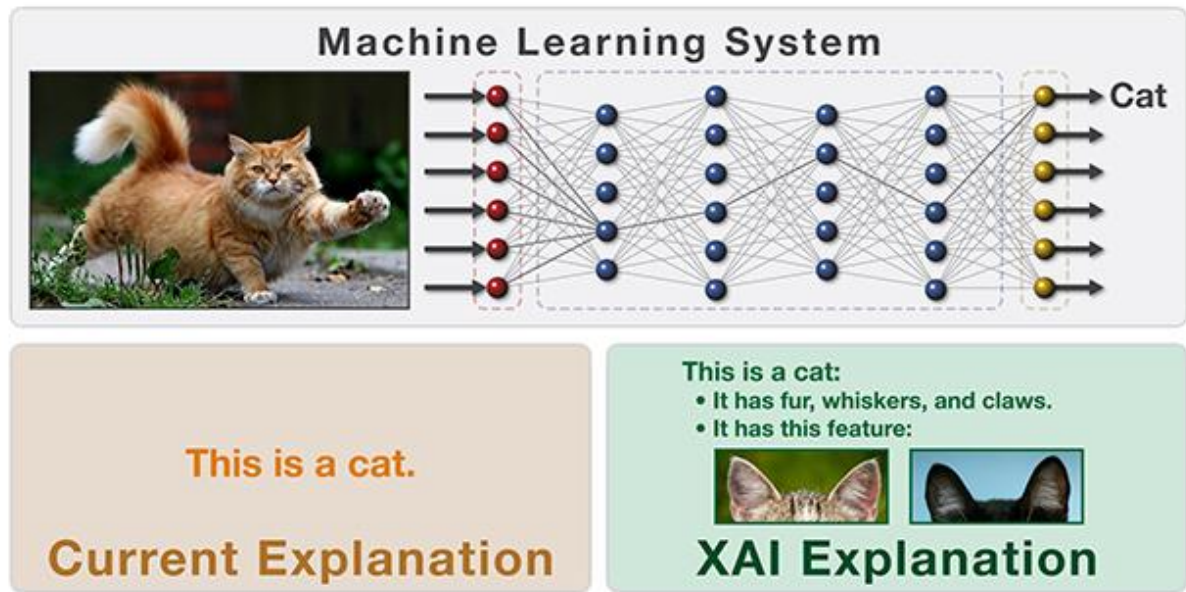
知识图谱构建的有效策略与方案

- 先易后难
 - 结构化-> 半结构化->非结构化
- 避免从零开始
 - 以通用图谱中的领域图谱作为种子
 - 问题：如何有效发现领域实体与关系？
- 跨领域迁移
 - 从邻近领域迁移
 - 问题：如何迁移具有共性的知识？



基于知识图谱的可解释人工智能

- 增强AI模型的透明(transparency)是人机互信的前提
- Mr. David Gunning
 - New machine-learning systems will have the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future.



<https://www.darpa.mil/program/explainable-artificial-intelligence>

基于知识图谱的可解释人工智能

• 解释概念

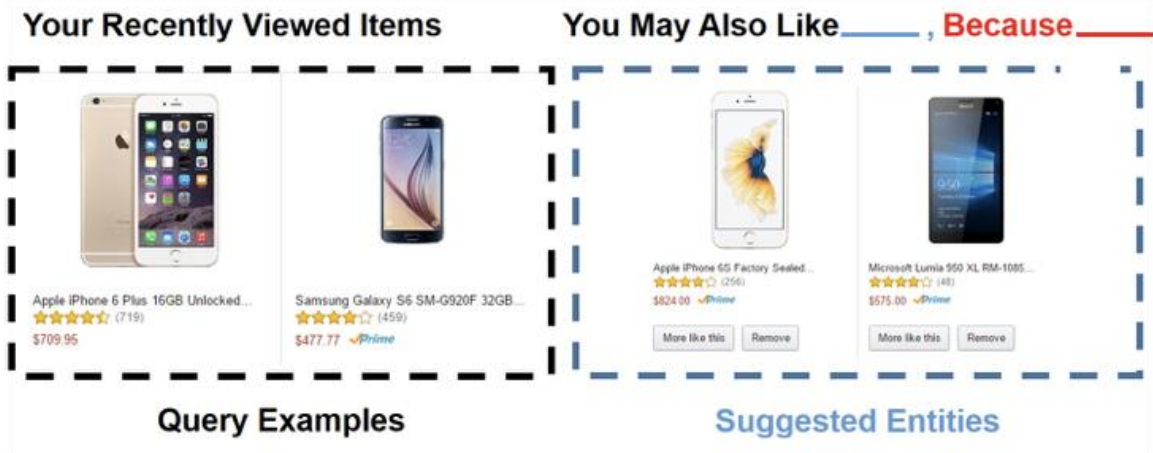
- Bachelor -> A man, unmarried (in IJCAI 2016)

• 解释实体集

- China, Japan, India -> Asian Country (in IJCAI 2015)

• 解释推荐

- Samsung S6, Iphone 6 -> Huawei P9



$$\operatorname{argmin}_{e \in E-q} KL(P(C|q), P(C|q, e))$$

$$KL(P(C|q), P(C|q, e)) = \sum_{i=1}^n P(c_i|q) * \log\left(\frac{P(c_i|q)}{P(c_i|q, e)}\right)$$

解释推荐, in IJCAI2017



Edward Feigenbaum

Knowledge is Power in AI



卡尔·雅斯贝斯

即将到来的的是一个终点，还是一个起点？它会不会是一个起点，其重要性相当于人最初成为人的时候，所不同的只是人现在拥有大量新获得的工具以及在一个新的、更高的水准上的经验能力？

——《时代的精神状况》

Thank YOU !



- Our LAB: Knowledge Works at Fudan University
 - <http://kw.fudan.edu.cn>