数据科学与计算机学院
School of Data and Computer Science

自然语言处理

Natural Language Processing

权小军 教授
中山大学数据科学与计算机学院
quanxj3@mail.sysu.edu.cn

知识图谱概述

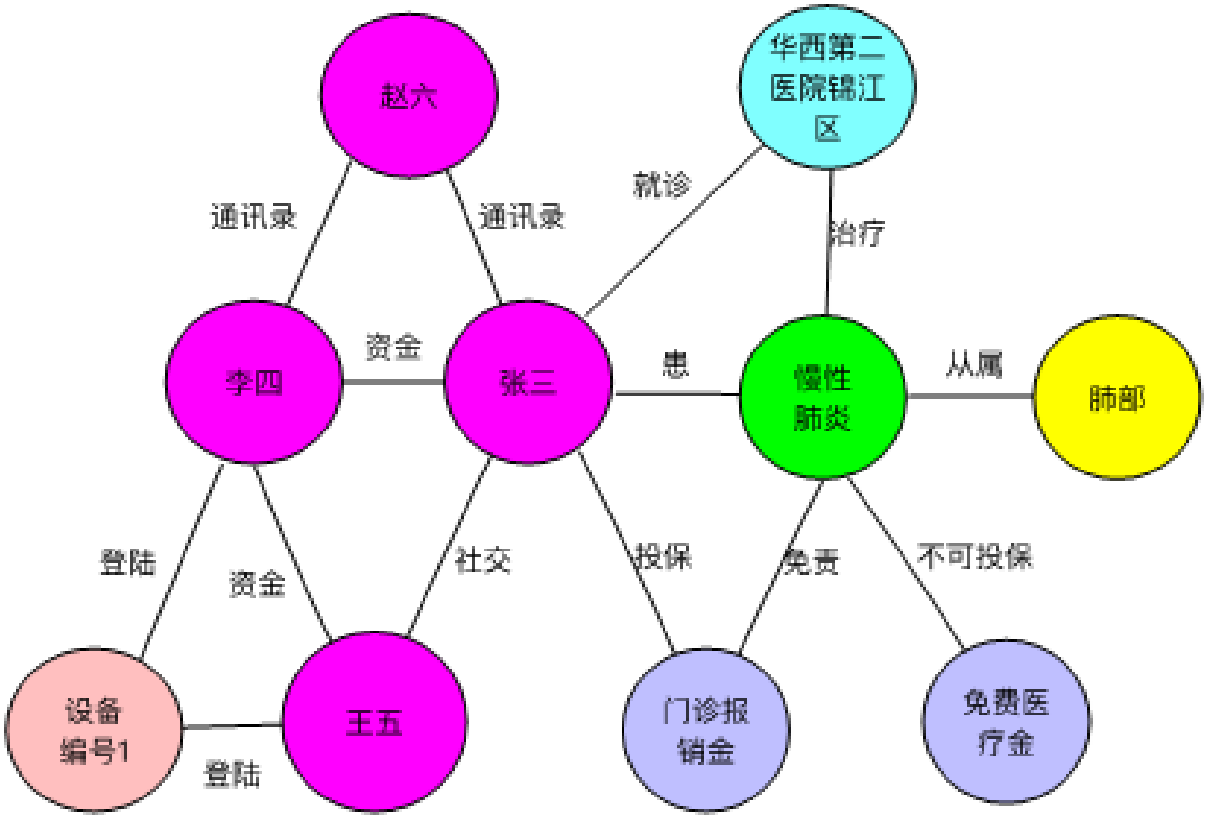
本章大纲

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

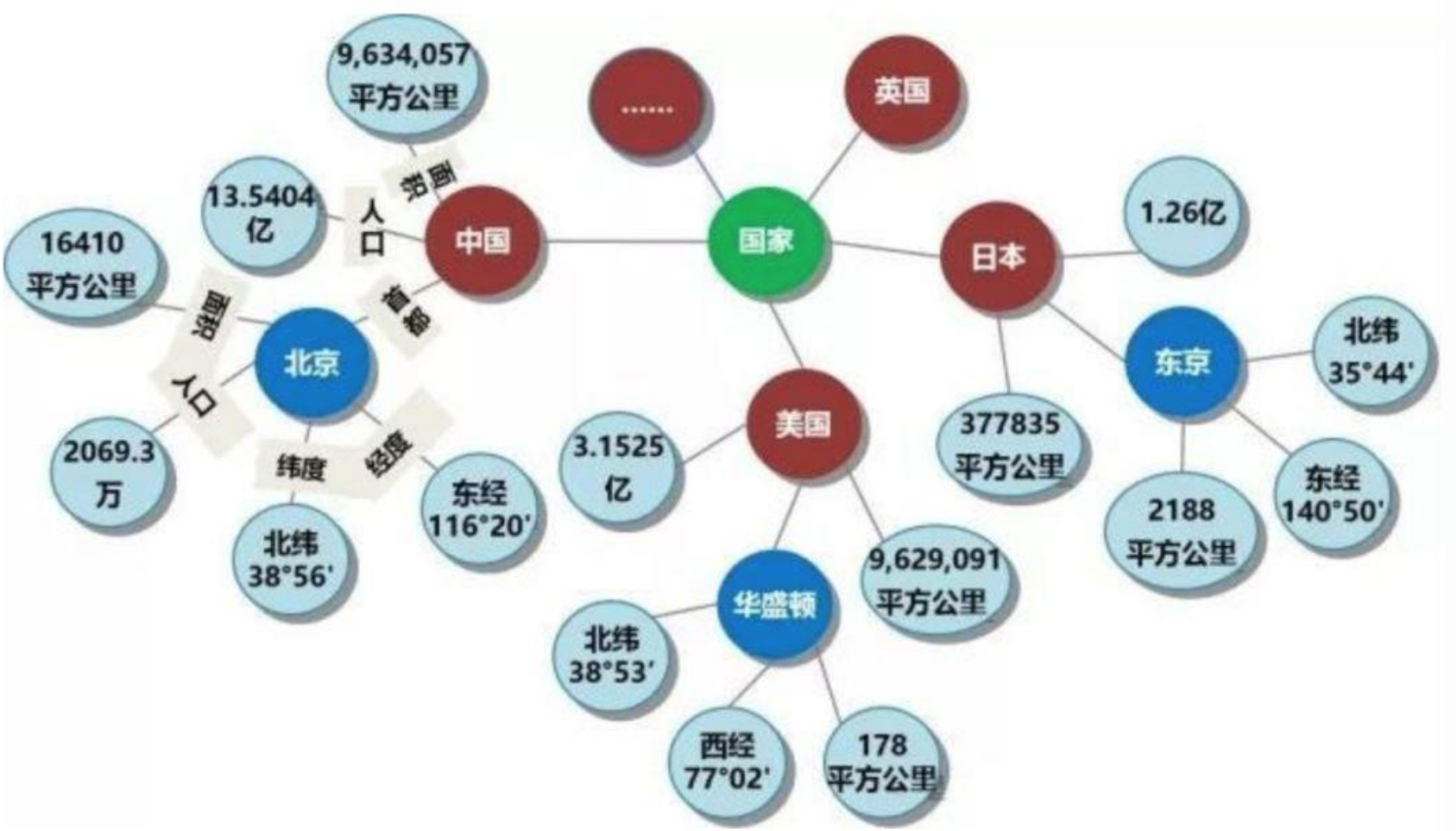
本章大纲

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

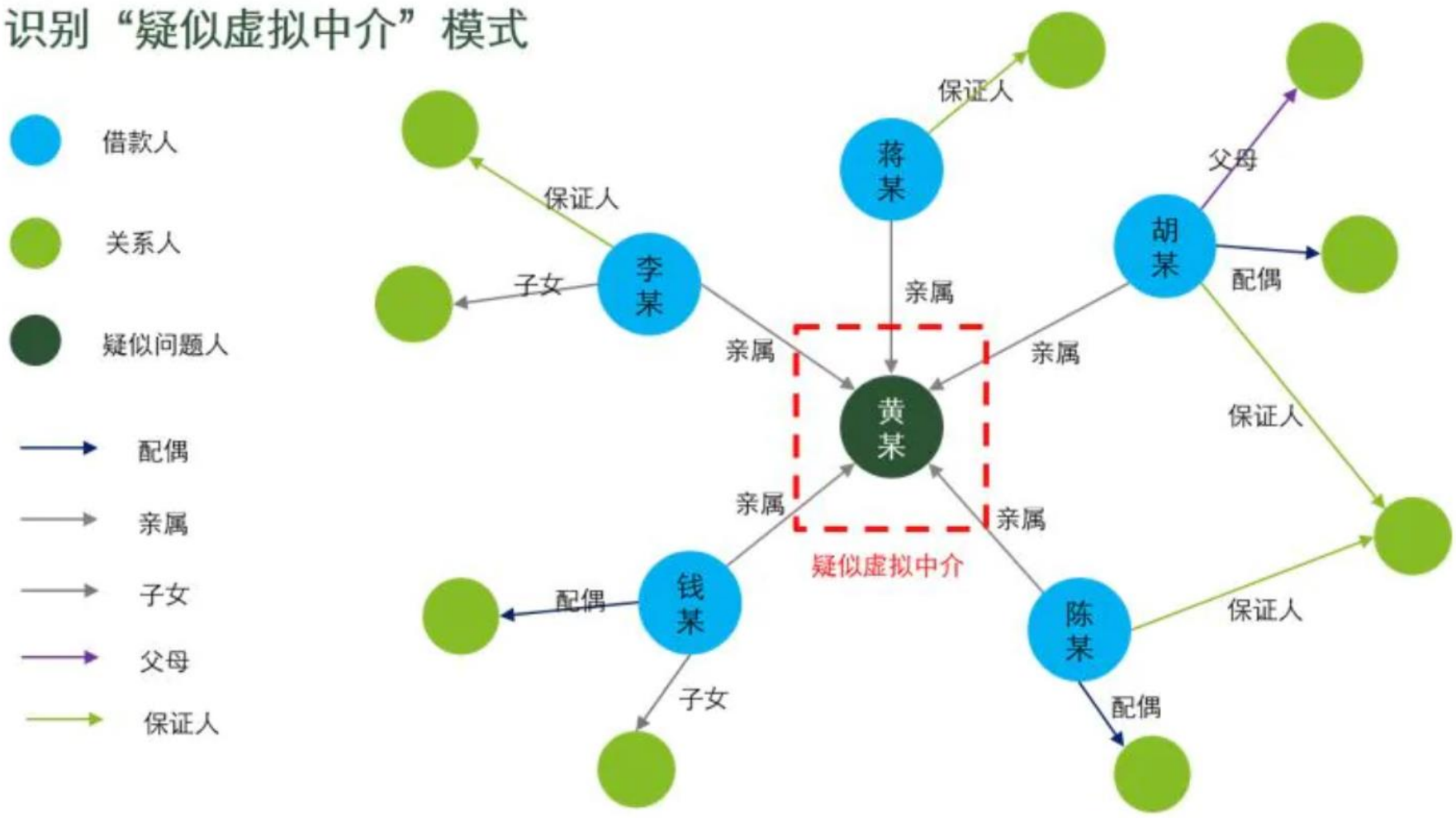
知识图谱


赵六
华西第二医院锦江区
通讯录
通讯录
就诊
治疗
李四
资金
张三
患
慢性肺炎
从属
肺部
登陆
资金
社交
投保
免责
不可投保
设备编号1
登陆
王五
门诊报销金
免费医疗金

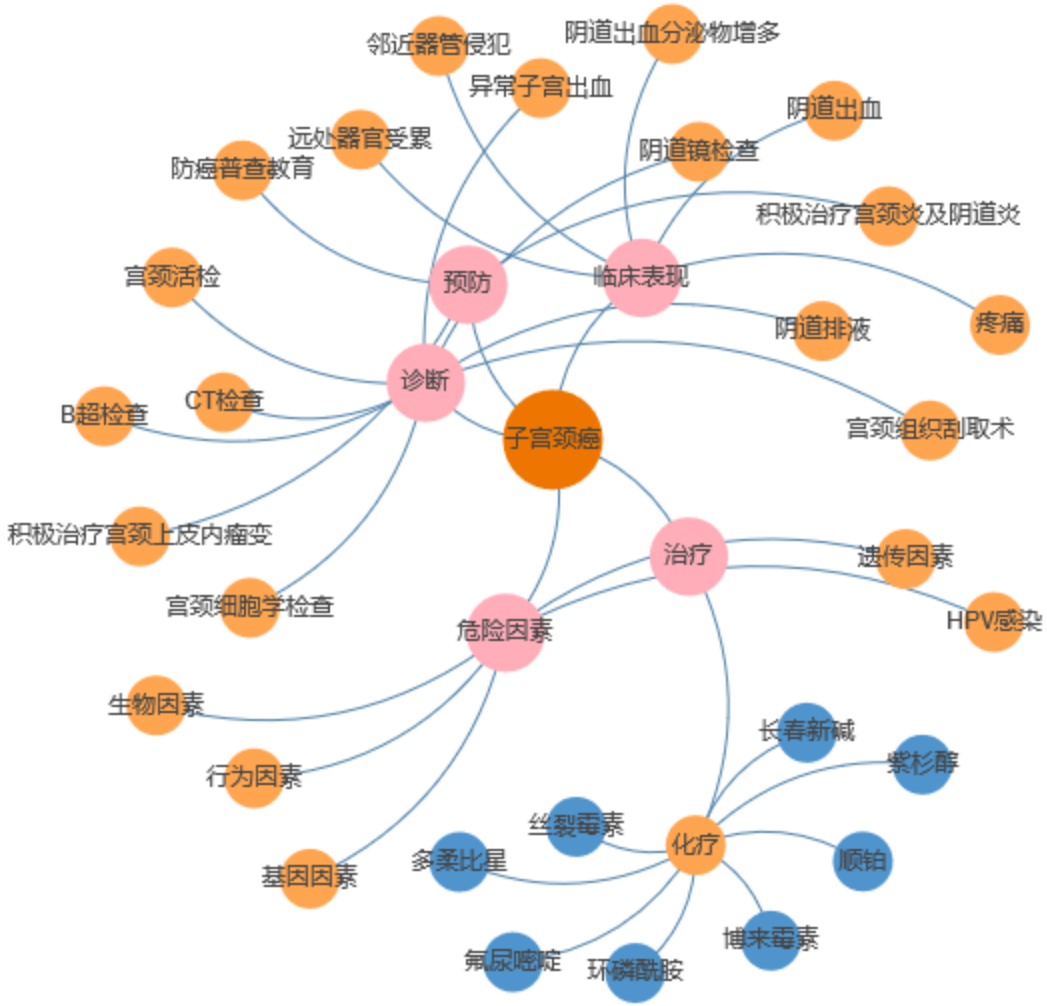
知识图谱


9,634,057
平方公里
面积
13.5404
亿
人口
中国
16410
平方公里
面积
北京
人口
2069.3
万
首都
纬度
经度
北纬
38°56'
东经
116°20'
.....
英国
国家
日本
1.26亿
美国
3.1525
亿
377835
平方公里
东京
北纬
35°44'
2188
平方公里
东经
140°50'
9,629,091
平方公里
华盛顿
北纬
38°53'
西经
77°02'
178
平方公里

知识图谱

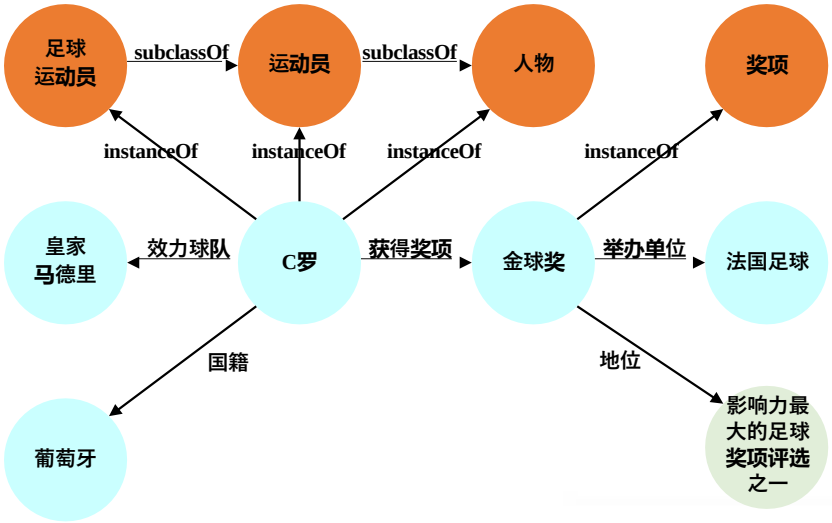

识别“疑似虚拟中介”模式
借款人
关系人
疑似问题人
配偶
亲属
子女
父母
保证人
蒋某
李某
胡某
黄某
钱某
陈某
疑似虚拟中介

知识图谱


邻近器官侵犯
阴道出血分泌物增多
异常子宫出血
阴道出血
远处器官受累
阴道镜检查
防癌普查教育
积极治疗宫颈炎及阴道炎
宫颈活检
预防
临床表现
阴道排液
疼痛
诊断
B超检查
CT检查
子宫颈癌
宫颈组织刮取术
积极治疗宫颈上皮内瘤变
治疗
遗传因素
宫颈细胞学检查
危险因素
HPV感染
生物因素
长春新碱
紫杉醇
行为因素
丝裂霉素
化疗
基因因素
多柔比星
顺铂
博来霉素
氟尿嘧啶
环磷酰胺

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)



知识图谱示例子。知识图谱富含实体、概念、属性、关系等信息

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
- 作为一种技术体系，是大数据时代**知识工程**的代表性进展

领域知识图谱

- 领域知识图谱 (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

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

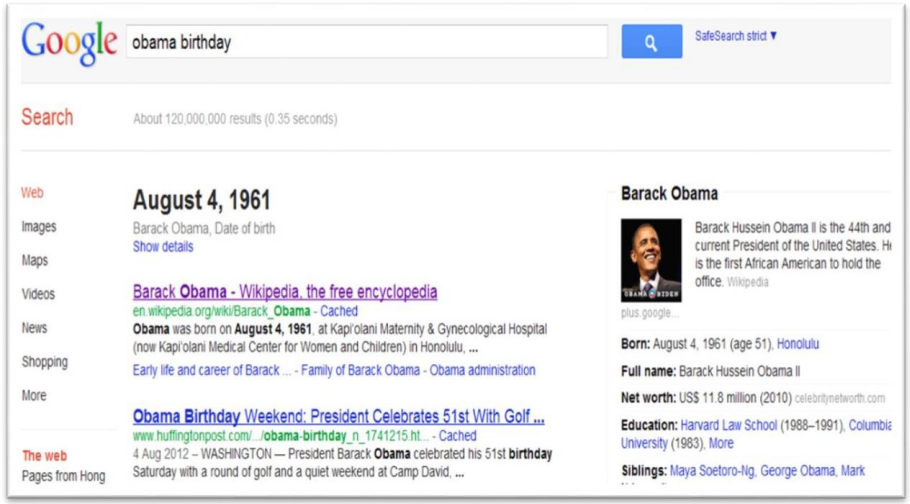
KR (Knowledge representation) is dedicated to **representing information about the world** in a form that a computer system can utilize to solve complex tasks such as diagnosing a medical condition or having a dialog in a natural language.

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

本章大纲

- 知识图谱概念
- **知识图谱内涵**
- 知识图谱优势
- 知识图谱价值
- 知识图谱应用
- 典型知识图谱

KG组成- Node-Entity

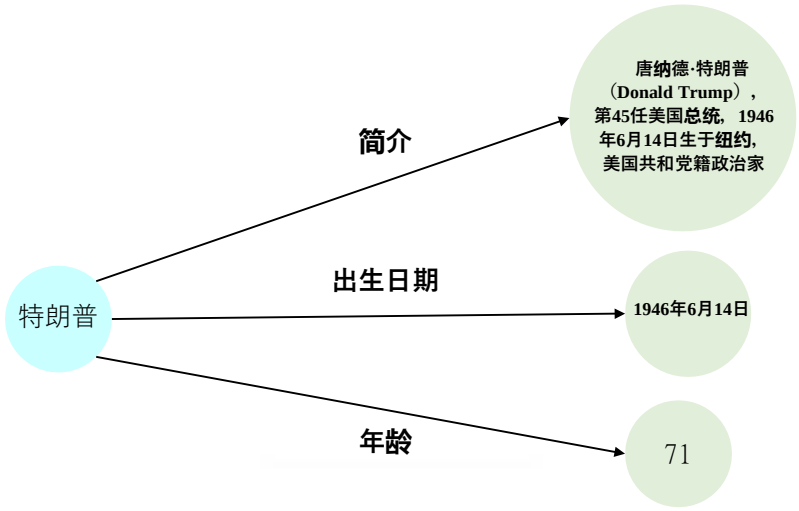
- **Entity/Objects/Instances**
  - Wikipedia: An **entity** is something that exists as itself, as a subject or as an object, actually or potentially, concretely or abstractly, physically or not.
  - 百度百科：客观存在并可相互区别的事物，学生、教师、课程都是实体

KG组成- Node-Concept

- Concept
  - A concept is a fundamental category of existence.
  - (mental) representations of categories
- Category
  - Groups of entities which have something in common;

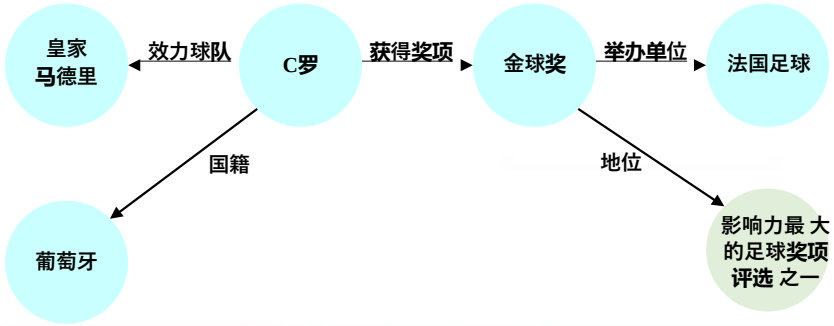
KG组成- Node-Value

- 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



本章大纲

- 知识图谱概念
- 知识图谱内涵
- **知识图谱优势**
- 知识图谱价值
- 知识图谱应用
- 典型知识图谱

KG优势1： large scale

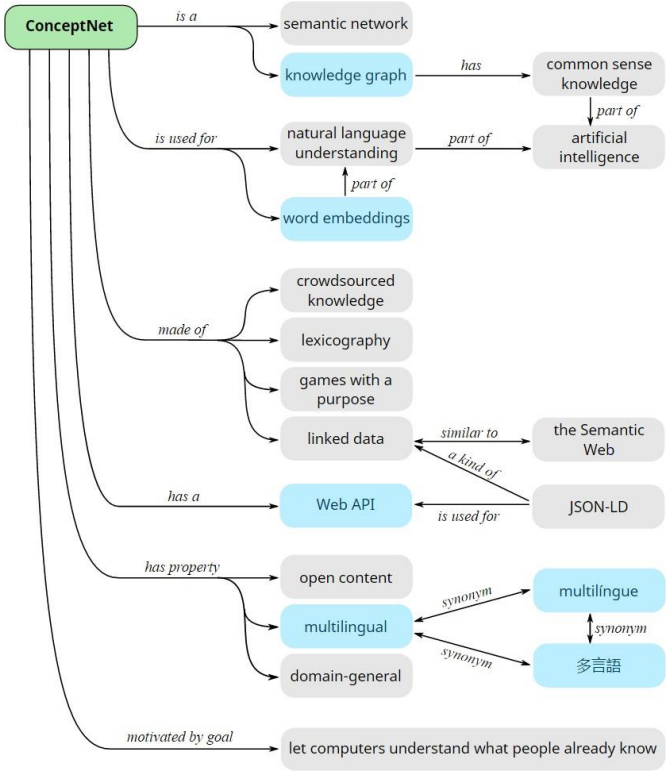
- Higher coverage over entities and concepts

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

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

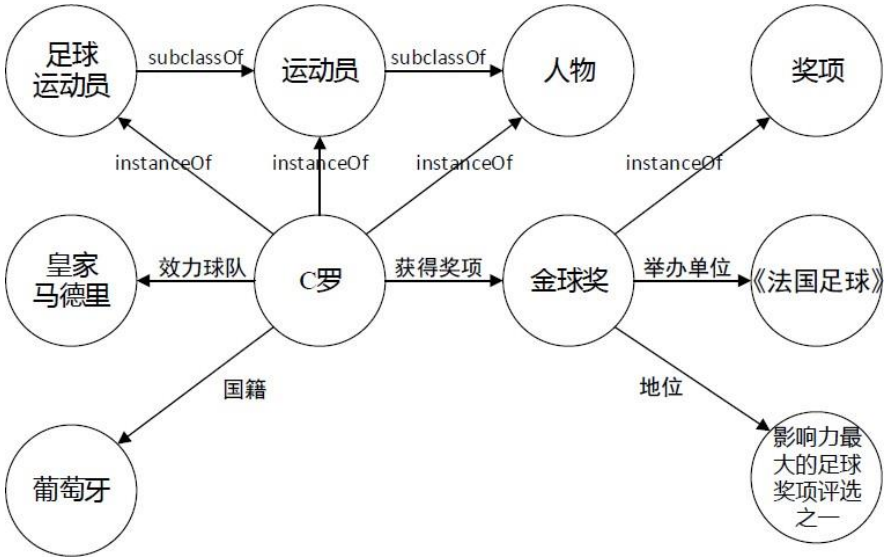
❑ High quality

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

专职院士	中国工程院院士5人
专职院士	中国科学院院士15人
专职院士	国家重大科学研究计划首席科学家9人
中文名	中山大学
主管部门	中华人民共和国教育部
创办人	孙中山
创办时间	1924年
博士后	科研流动站41个

KG优势4: friendly structure

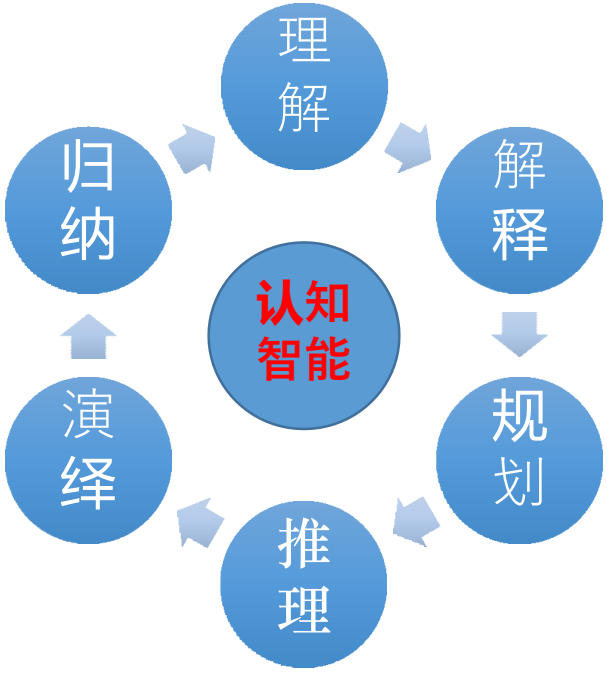
- Structured organization
  - By RDF
  - By graph



本章大纲

- 知识图谱概念
- 知识图谱内涵
- 知识图谱优势
- **知识图谱价值**
- 知识图谱应用
- 典型知识图谱

认知智能是智能化的关键

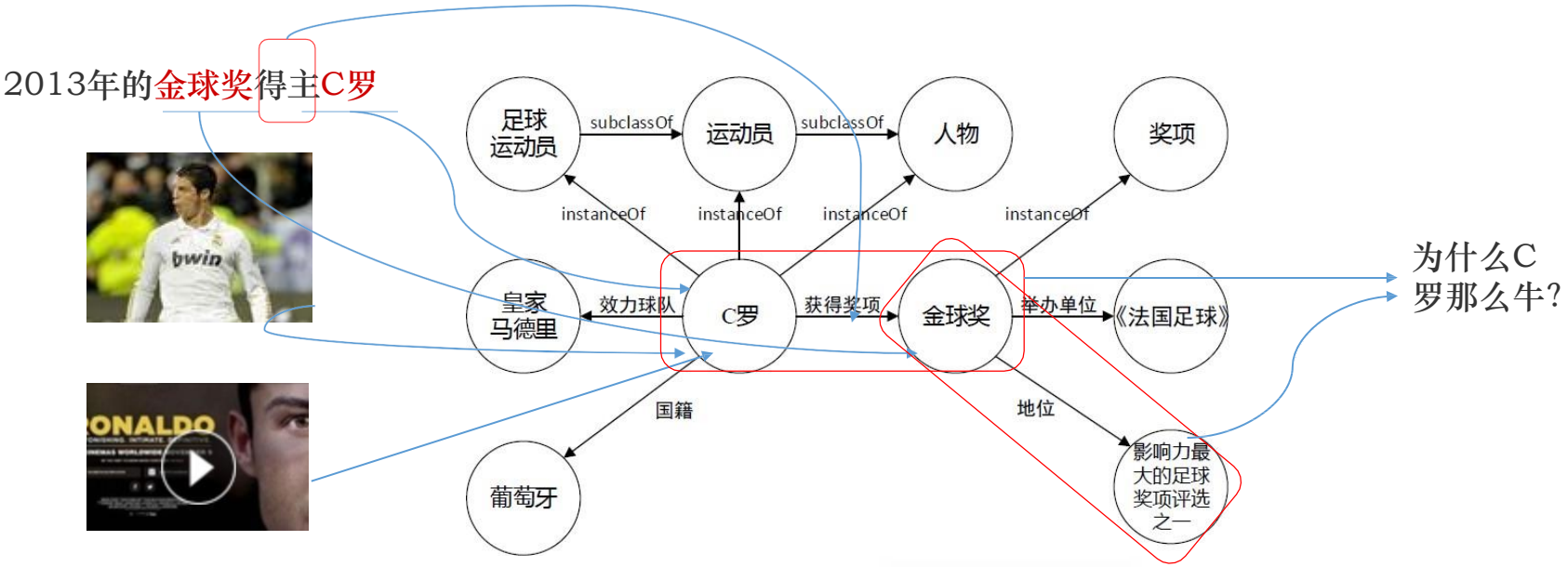


Can machine **think like humans**?

- 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



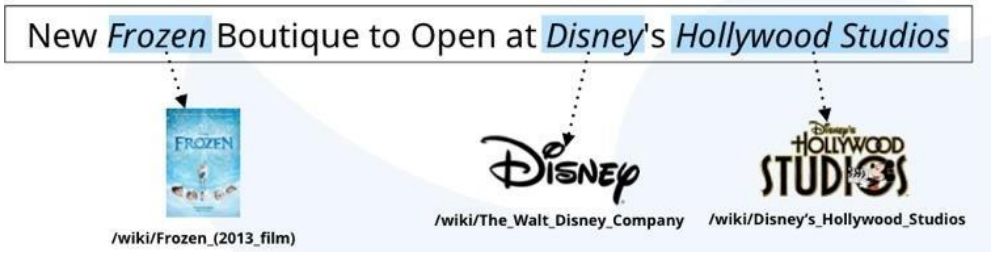
机器语言理解需要背景知识

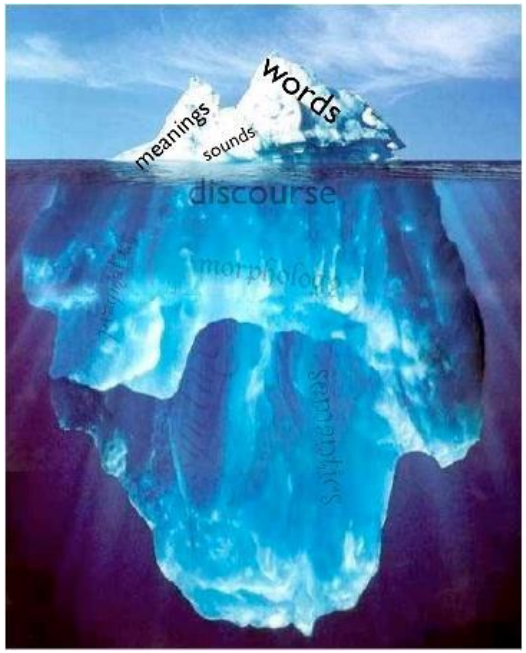
❑ Language is complicated

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

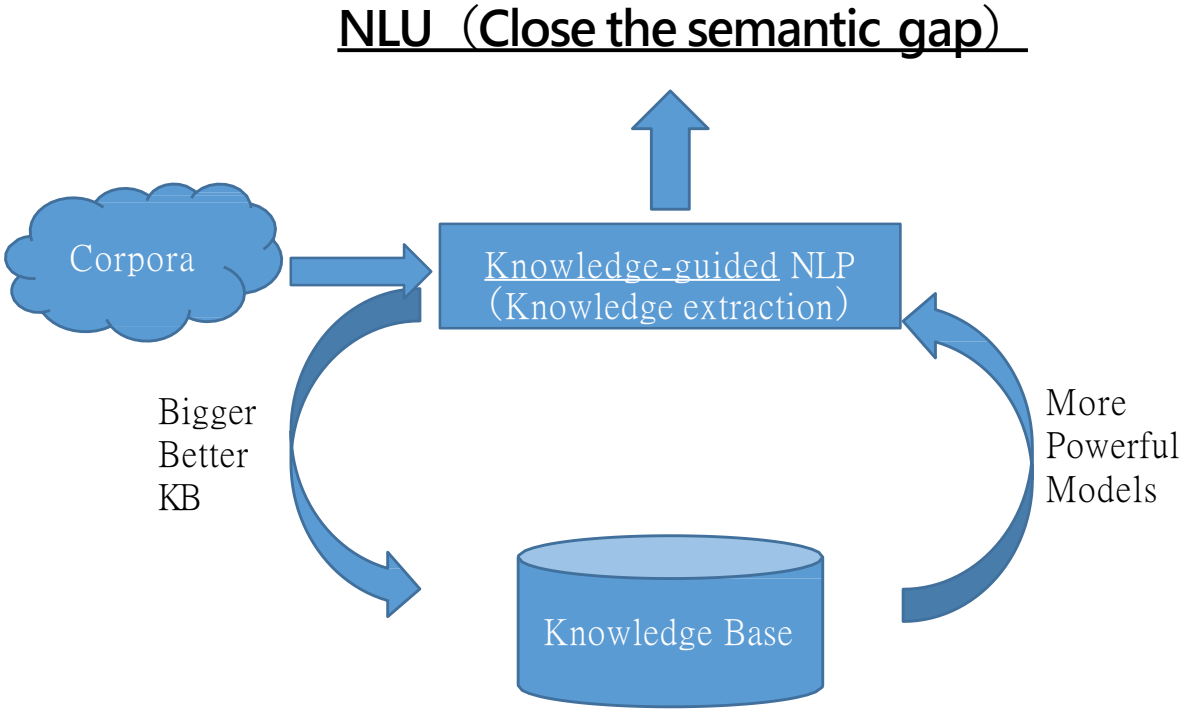
❑ Language understanding is difficult

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

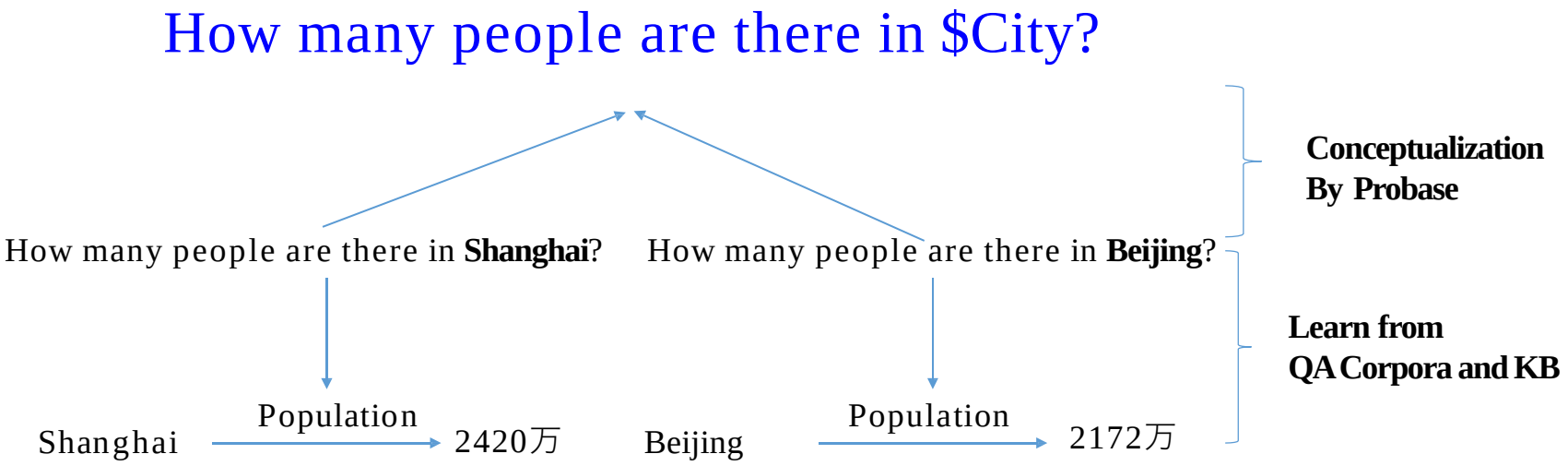




The roadmap of knowledge-guided NLP



Example: Using concepts to understand a natural language?



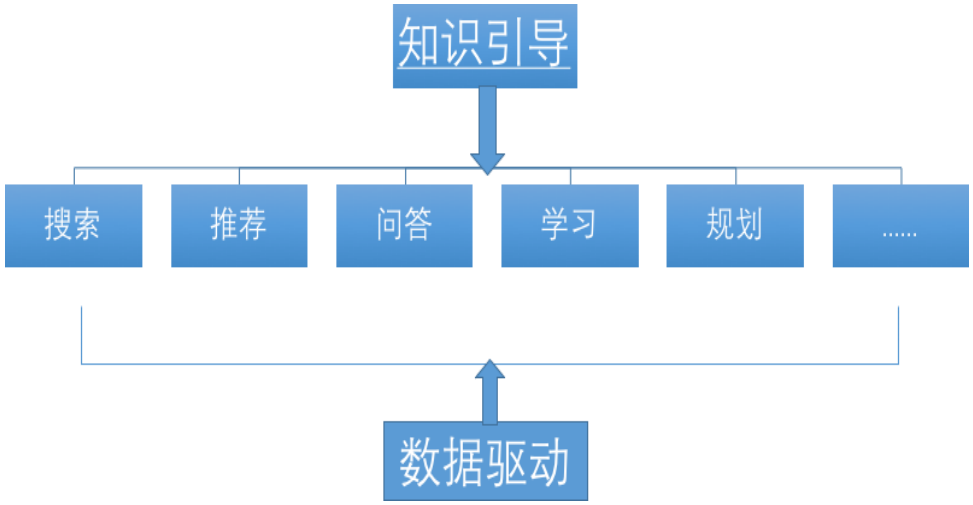
[Wanyun Cui et al. 2017]

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

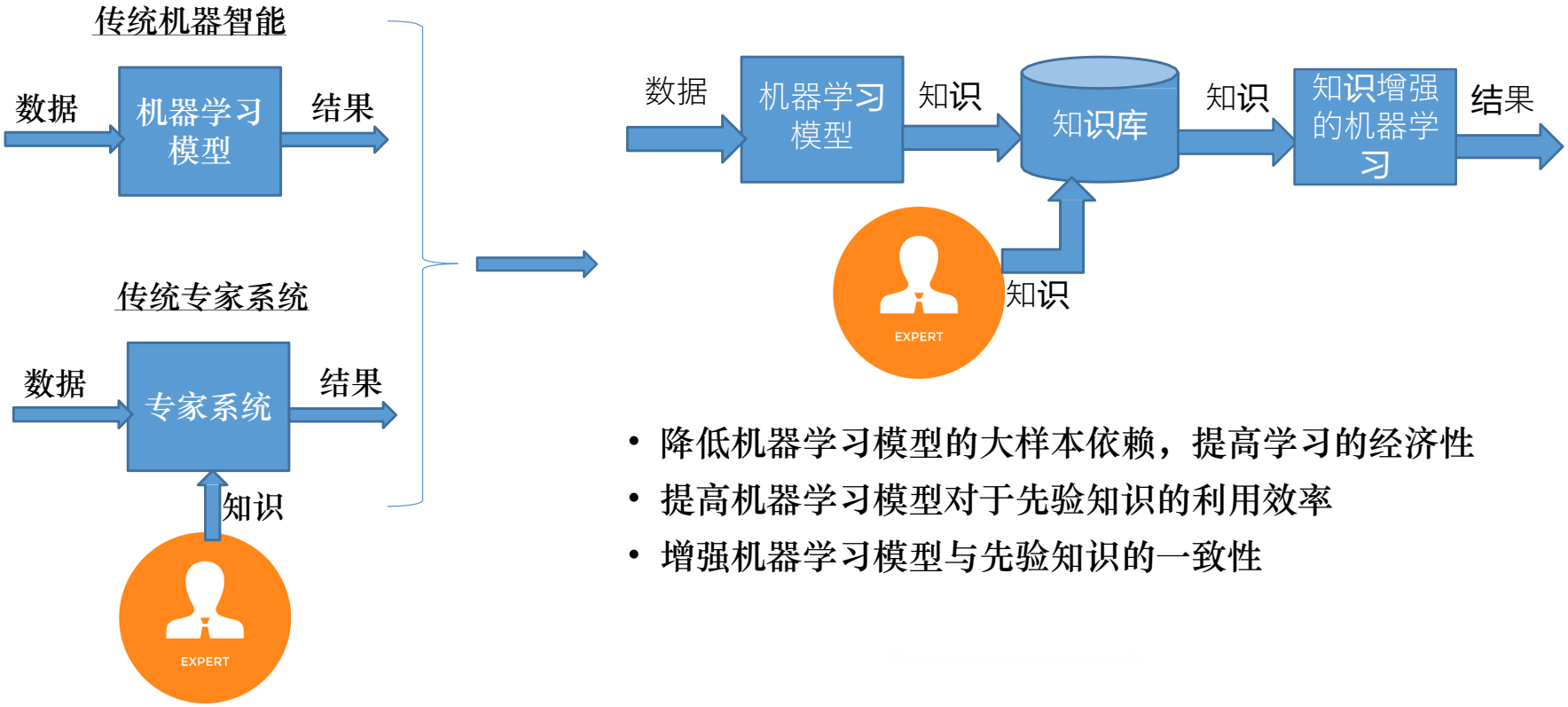
鸟儿为何能够飞翔?
因为它们有翅膀 → 属性

知识引导将成为解决问题的新方式



- “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

知识增强机器学习能力



- 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

本章大纲

- 知识图谱概念
- 知识图谱内涵
- 知识图谱优势
- 知识图谱价值
- **知识图谱应用**
- 典型知识图谱

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

深扒王宝强离婚内幕 最大祸根源于谁_百山探索

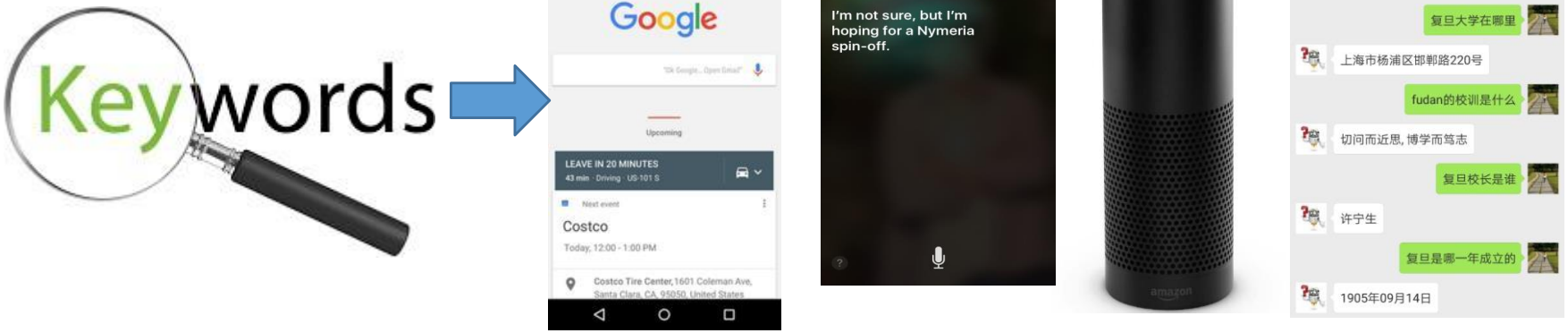
深度解析宝宝离婚闹剧事件 细说婚姻幸福真谛!_央广网

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

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

自然人机交互



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?

本章大纲

- 知识图谱概念
- 知识图谱内涵
- 知识图谱优势
- 知识图谱价值
- 知识图谱应用
- 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

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

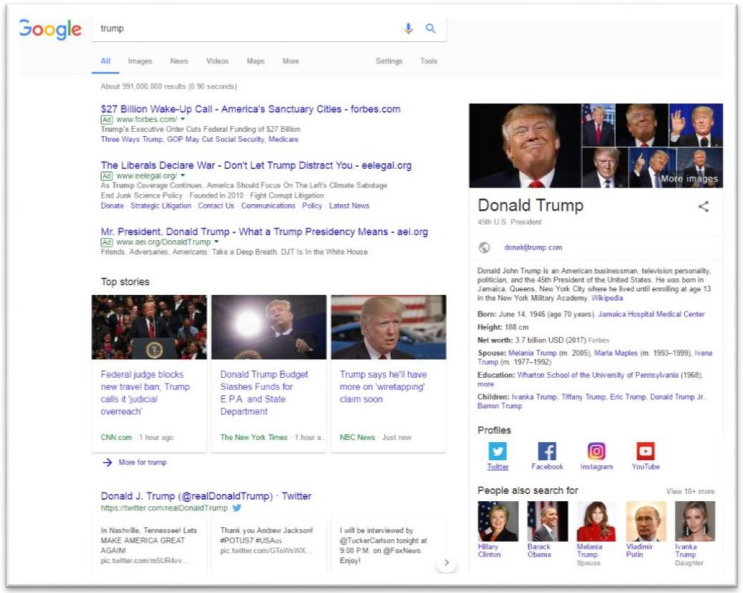
- 简介
  - 大型的多语言常识知识库
- 样例
  - 刘德华
- 特点
  - 知识来源丰富
    - 众包(Crowd-Sourcing)
    - 资源（例如Wiktionary 和Open Mind Common Sense）
    - 带目的的游戏（如Verbosity 和 nadya.jp）
    - 专家创建的资源(如WordNet 和 JMDict)

http://conceptnet.io/

[Robert Speer et al. 2012]

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



Thank you!

权小军 中山大学数据科学与计算机学院