
中国中文信息学会前沿技术讲习班
CIPS ATT7 & CCKS2017 Tutorial


# 知识图谱导论

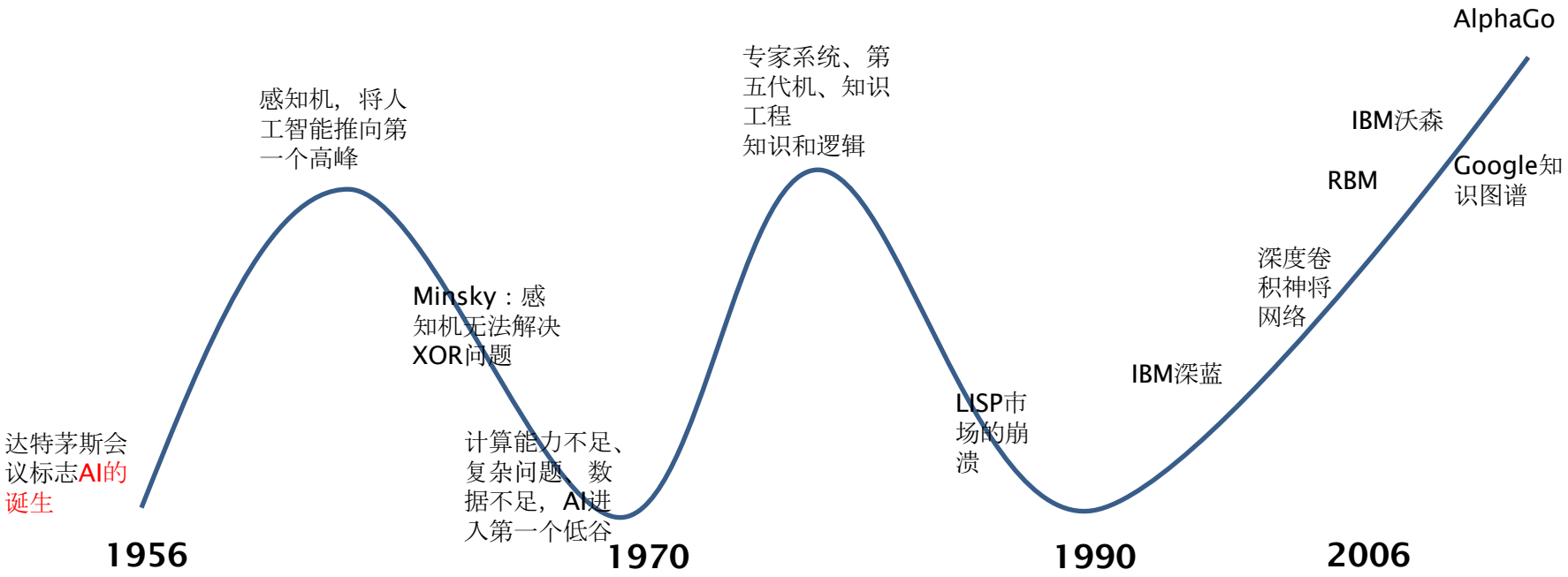
**刘康**　中国科学院自动化研究所
**韩先培**　中国科学院软件研究所

成都 2017-8-26

# 目录

# 目录

# 人工智能的发展历史

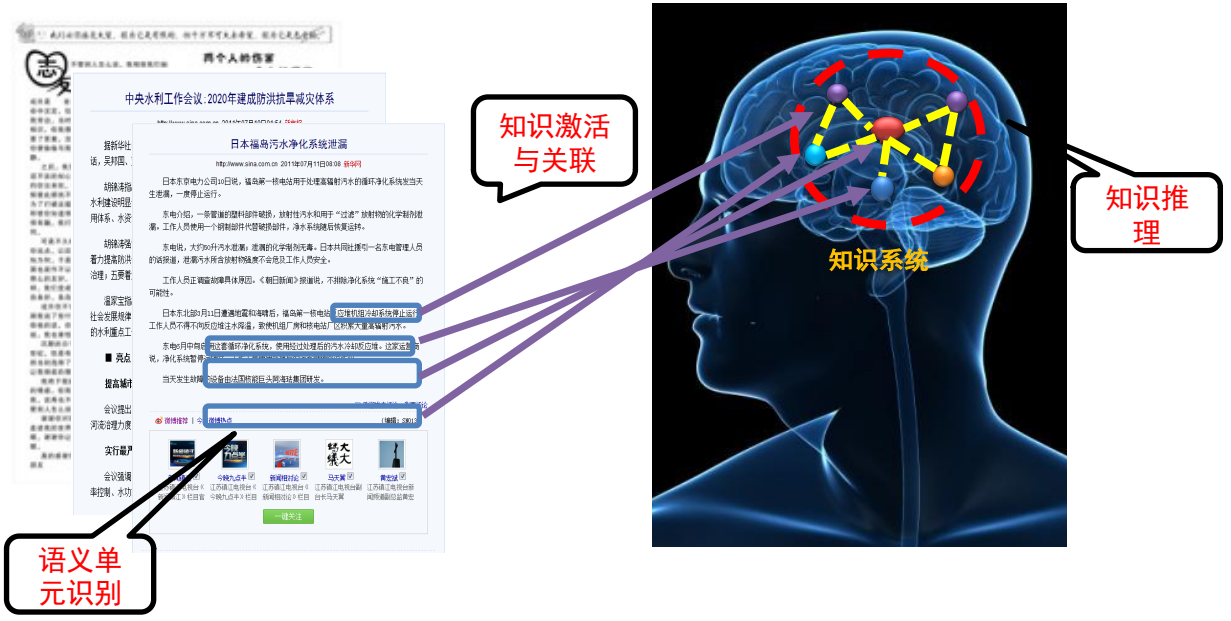

AlphaGo
感知机，将人工智能推向第一个高峰
专家系统、第五代机、知识工程
知识和逻辑
IBM沃森
RBM
Google知识图谱
深度卷积神将网络
Minsky：感知机无法解决XOR问题
IBM深蓝
LISP市场的崩溃
达特茅斯会议标志AI的诞生
计算能力不足、复杂问题、数据不足，AI进入第一个低谷
1956
1970
1990
2006

# 知识就是力量

*Knowledge is power, and the computer is an amplifier of that power. We are now at the dawn of a new computer revolution... Knowledge itself is to become the new wealth of nations.*

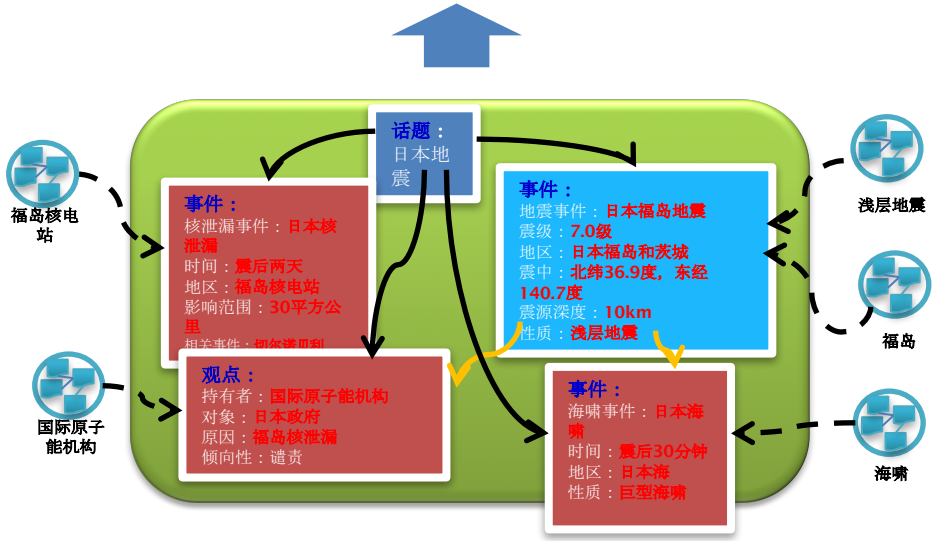
-Edward Feigenbaum (专家系统之父，1994年图灵奖)

# 深度自然语言理解需要知识的支撑



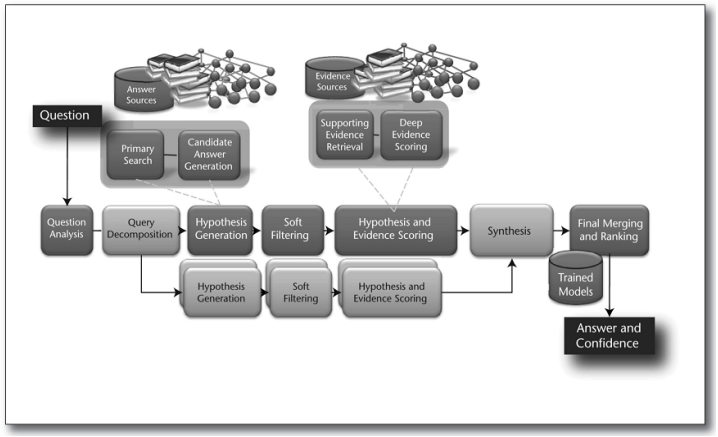
# 深度自然语言理解需要知识的支撑

2011年4月11日17点16分，日本东北部的福岛和茨城地区发生里氏7.0级强烈地震（震中北纬36.9度、东经140.7度，即福岛西南30公里左右的地方，震源深度10公里，属于浅层地震）。当局已经发布海啸预警。震后约30分钟后在日本海地区发生巨型海啸，同时造成福岛核电站出现核泄漏。震后第十天，国际原子能机构对于日本政府反应迟钝进行了谴责。



# IBM Watson

- 沃森(Watson)：2011年，IBM研发的超级计算机“沃森”在美国知识竞赛节目《危险边缘Jeopardy!》中上演“人机问答大战”，战胜人类选手Ken和Brad

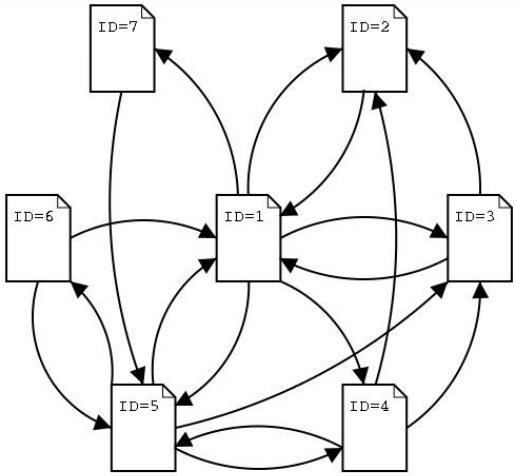

辅助医疗

金融辅助决策

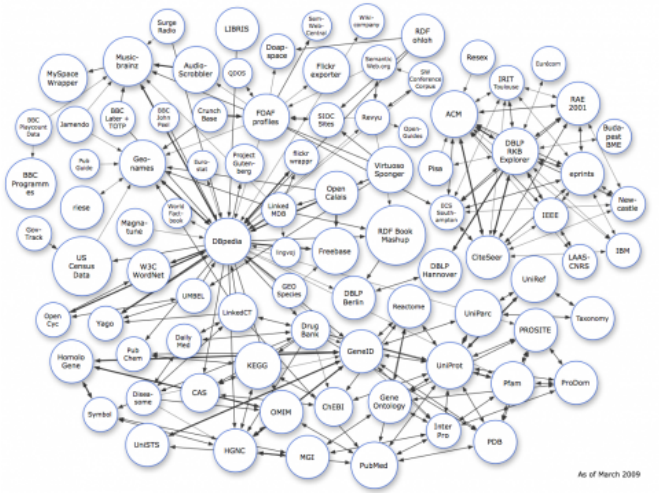
企业服务

# Semantic Web

- Tim Berners-Lee 1998年提出语义网的概念
  - 通过给全球信息网上的文档（如：标准通用标记语言下的一个应用HTML）添加能够被计算器所理解的语义“元数据”（Meta data），从而使整个互联网成为一个通用的信息交换媒介




Page/Document web



Data web

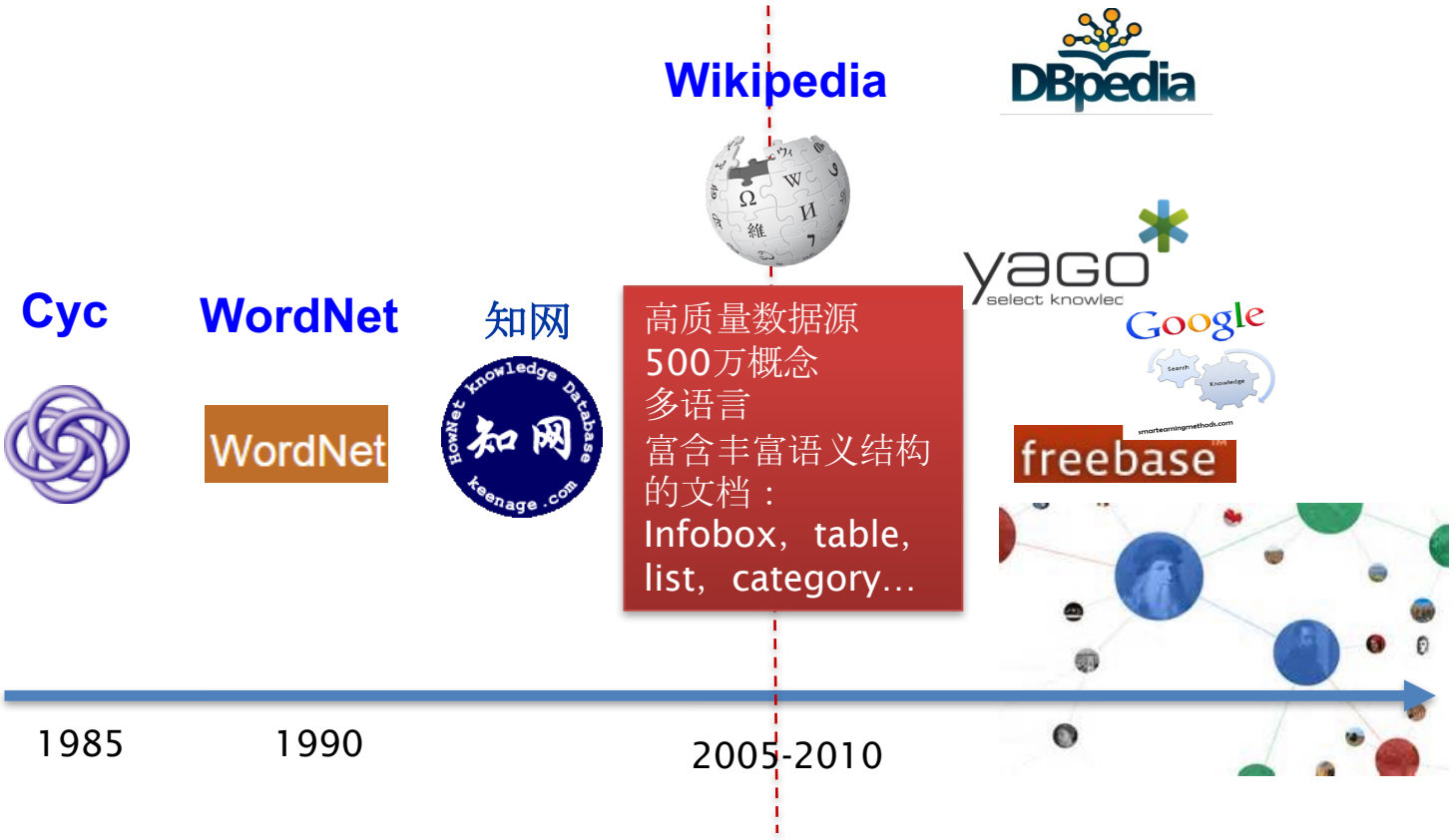
# 已有的知识图谱

- 语言知识图谱
  - WordNet：155, 327个单词，同义词集117,597个，同义词集之间由22种关系连接
- 事实性知识图谱
  - OpenCyc： 23.9万个实体，1.5万个关系属性，209.3万个事实三元组
  - Freebase：4000多万实体，上万个属性关系，24多亿个事实三元组
  - DBpedia：400多万实体，48,293种属性关系，10亿个事实三元组
  - YAGO2 ：980万实体，超过100个属性关系， 1亿多个事实三元组
  - 百度百科：词条数1000万个
  - 互动百科：800万词条，5万个分类，68亿文字

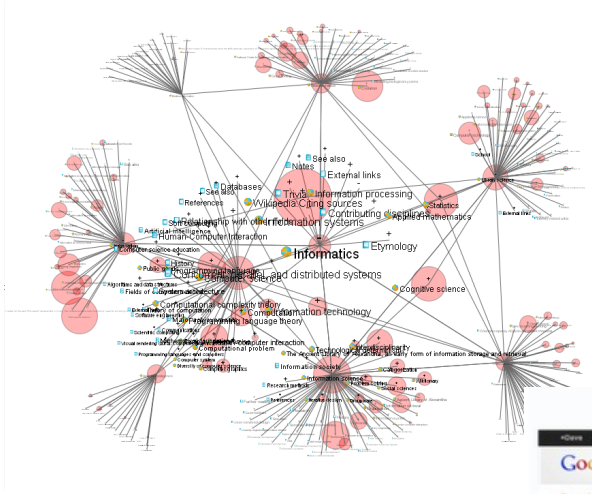
# 已有的知识图谱

- 领域知识图谱
  - Kinships：描述人物之间的亲属关系，104个实体，26种关系，10,800个三元组
  - UMLS：医学领域，描述医学概念之间的联系，135个实体，49种关系，6,800个三元组。
  - Cora：2,497个实体，7种关系，39,255个三元组
- 机器自动构建的知识图谱
  - NELL： 519万实体，306种关系， 5亿候选三元组
  - Knowledge Vault: 4500万实体，4469种关系，2.7亿三元组

# 知识图谱历史

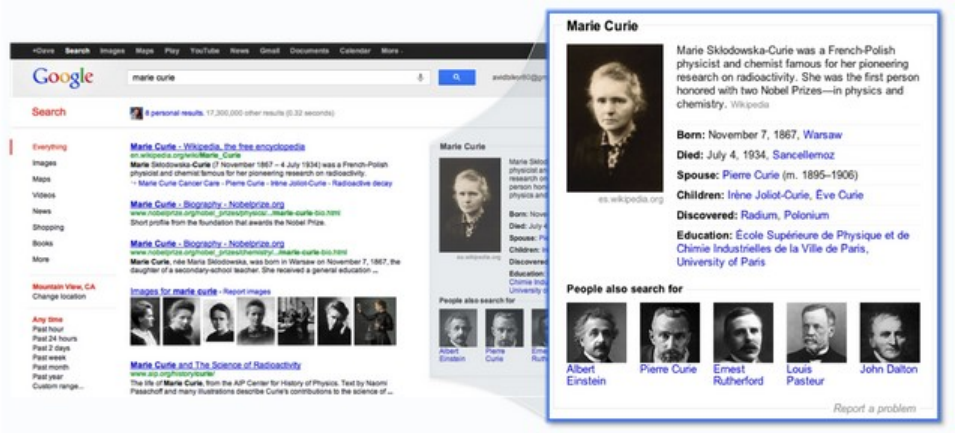

Wikipedia
DBpedia
Cyc
WordNet
知网
yago
Google
freebase
高质量数据源
500万概念
多语言
富含丰富语义结构的文档：
Infobox，table，list，category…
1985
1990
2005-2010

# Knowledge Graph





超过5亿实体
超过35亿条关系

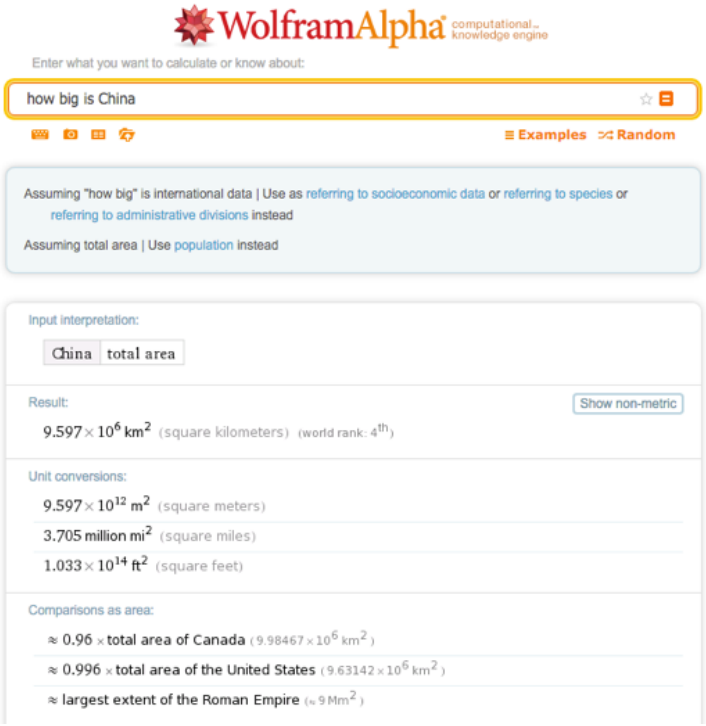



ProBase

百度知心

搜狗知立方

# 问答


WolframAlpha computational knowledge engine
Enter what you want to calculate or know about:
how big is China
Examples Random
Assuming "how big" is international data | Use as referring to socioeconomic data or referring to species or referring to administrative divisions instead
Assuming total area | Use population instead
Input interpretation:
China total area
Result:
Show non-metric
9.597 × 10^6 km^2 (square kilometers) (world rank: 4th)
Unit conversions:
9.597 × 10^12 m^2 (square meters)
3.705 million mi^2 (square miles)
1.033 × 10^14 ft^2 (square feet)
Comparisons as area:
≈ 0.96 × total area of Canada (9.98467 × 10^6 km^2)
≈ 0.996 × total area of the United States (9.63142 × 10^6 km^2)
≈ largest extent of the Roman Empire (≈ 9 Mm^2)


姚明个子有多少
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约3,030,000个
搜索工具
姚明身高:
226cm
姚明，1980年生于上海市徐汇区，祖籍吴江震泽。中国篮球运动员。1998年4月，他入选王非执教的国家队，开始篮球生涯。2002年，他以状元秀身份被NBA的休斯敦火箭队选中。20... 详情>>
来自百度百科 | 报错

# 精准搜索


奥巴马
百度一下
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约86,500,000个
搜索工具
奥巴马_百度百科
姓名：贝拉克·侯赛因·奥巴马
生日：1961年8月4日 职业：政治家、律师、总统
简介：贝拉克·侯赛因·奥巴马（Barack Hussein Obama），1961年8月4日出生，美国民主党籍政治家，第44任美国总统，为美...
人物经历 执政表现 主要作品 家庭生活 人物评价 更多>>
查看“奥巴马”全部4个含义>>
baike.baidu.com/
奥巴马的最新相关信息
奥巴马卸任后当NBA球队老板? 白宫发言人:在讨论
ESPN消息,美国总统奥巴马即将卸任,而他未来要做的事情似乎已经确定好了,那就是当一支NBA球队的老板。 奥巴马卸任后当NBA球队老板? 白宫发言人:在讨...
网易体育 49分钟前
奥巴马欲成NBA球队老板 总统助奇才抢杜兰... 腾讯体育 1小时前
奥巴马的算盘: 英国留欧利于美国外交 搜狐财经 2小时前
奥巴马称朝鲜仍是特别威胁 决定对朝制裁... 网易军事 18小时前
作为最亲密的盟友 奥巴马对英国“脱欧”... 新浪财经 18小时前
奥巴马_百度图片
历年时代周刊年度风云人物
展开
普京 俄罗斯铁腕总统
克林顿 曾任美国总统
小布什 美国第43任总统
比尔·盖茨 微软公司创始人前首富
肯尼迪 美国第35任总统
罗斯福 美国蝉联4届的总统
斯大林 苏联共产党中央总书记
里根 演员出身的美国总统
美国民主党员
展开
希拉里 美国第67任国务卿
拜登 美国现任副总统
卡罗琳·肯尼迪 被美国人称全国的宝贝
克里 美国第68任国务卿

# 精准搜索

王健林 下载报告

介绍： 王健林，男，1954年10月24日出生于四川省广元市，1989年起担任大连万达集团股份有限公司董事长。1970年入伍，1986年毕业于辽宁大学，同年7月进入大连市西岗区人民政府任办公室主任，1989年进入房地产行业，1993年起担任大连万达集团股份有限公司董事长、总裁。是中共十七大代表、第十一届全国政协常委、第十一届全国工商联副主席，兼任... 详细

他的所有商业角色　他的所有企业

| 角色 | 企业 | 省份地区 | 开业日期 | 注册资本 | 经营状态 |
|---|---|---|---|---|---|
| | 大连万达集团股份有限公司 | 辽宁 | 1992-09-28 | 100000 万人民币 | 存续（在营、开业、在册） |
| | 北京中住联华科技发展有限公司 | 北京 | 2001-04-25 | 120 万元 | 吊销 |
| | 北京环球高尔夫俱乐部有限公司 | 北京 | 1993-04-10 | 1120万元美元 | 吊销，未注销 |
| | 大连万达商业开发有限公司 | 辽宁 | 1994-12-12 | 0 万 | 吊销，未注销 |
| | 大连合兴投资有限公司 | 辽宁 | 2007-04-27 | 7860 万人民币 | 存续（在营、开业、在册） |
| | 大连万达通信有限公司 | 辽宁 | 1998-01-23 | 15000 万 | 吊销，未注销 |
| | 大连万达酒业有限公司 | 辽宁 | 1998-03-30 | 240 万 | 其他 |
| | 大连万达中心大厦有限公司 | 辽宁 | 1995-04-29 | 10000 万人民币 | 吊销，未注销 |
| | 四川万达房地产有限公司（转成都市锦江区） | 四川 | 1999-07-22 | 2000万人民币 | 存续（在营、开业、在册） |
| | 大连宏大物业管理有限公司 | 辽宁 | 1995-04-11 | 1000 万人民币 | 吊销，未注销 |
| | 大连天兴大酒店有限公司 | 辽宁 | 1997-01-20 | 100 万 | 吊销，未注销 |

# 关系搜索


beta
人立方
关系搜索
renlifang.msra.cn
六度秀
分享
帮助
周杰伦
奥巴马
六度搜索
周杰伦
迪克
切尼
奥巴马
点击上方人物，自定义关系链
希拉里
赖斯
汤姆
侯佩岑
布什
切尼
拉姆斯菲尔德
范特西
周杰伦
迪克
奥巴马
墨菲
爱德华兹
蔡依林
考夫曼
希拉里·克林顿

# 分类浏览


Baidu音乐
请输入歌名、歌词、歌手或专辑
薛之谦 林忆莲《我不.. 李宗盛 五月天 梁静茹 莫文蔚 陈梓童 孙盛希《Betw.. 周杰伦 张杰
首页 歌单 动态 NEW 歌手 分类 榜单 MV 演出
标签
热门
新歌 热歌 中国好声音 经典老歌 电视剧 广场舞 热
欧美 热 轻音乐 热 DJ 舞曲 80后 网络歌曲 热 劲爆
儿歌 纯音乐 粤语 民歌 钢琴曲 萨克斯
古典音乐 对唱 佛教音乐 成名曲 草原歌曲
心情
伤感 热 激情 安静 舒服 热 甜蜜 励志
寂寞 想念 浪漫 怀念 喜悦 深情
美好 怀旧 轻松
风格
小清新 热 DJ 舞曲 纯净 唯美 轻音乐 舒缓
劲爆 热 慢摇 民歌 热 青春 好听 交响乐
乐播
笑话段子 相声曲艺 热 脱口秀 母婴儿童 小说读物 综艺娱乐
都市情感 商业财经 教育 健康 新闻时事 科技
生活 社会文化 英语播客


音乐分类 > 纯净
纯净 共1000首歌
收听该分类电台
收听音乐人热播电台
相关无损大碟推荐
换一换
柴科夫斯基作品精选
浪漫主义钢琴诗人·肖邦盘点
Beethoven: Overtures
Mozart: Le nozze di Figaro - Highlights
全部 播放选中歌曲 加入播放列表 收藏 批量下载
01 心有灵犀 熊天平 《雪候鸟》
02 心有灵犀 熊天平 《雪候鸟》
03 有一个人 齐豫 《齐豫个人中文..》
04 心有灵犀 熊天平 《熊心万丈》
05 夏天的风 元卫觉醒 《艾回20年200曲..》
06 可不可以爱 何炅 《可以爱》
07 Eternity Kelly Sweet 《We Are One》
08 Dreamer Sophie Zelmani 《Time To Kill》
09 Ayo Technology Milow 《Milow》
10 A Little Love 冯曦妤 《A Little Love》
11 我不想忘记你 郭静 《我不想忘记你》
审批文号：WJ214775
12 陪你到世界的终结 棉花糖 《小飞行》
13 小飞行 棉花糖 《小飞行》
14 多余的解释 许嵩 《自定义》
15 想念 许哲佩 《气球》

# 推荐


战狼2
百度一下
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约5,830,000个
搜索工具
战狼2_百度百科
关注点 TA说：深度解读功与过
《战狼 II》是吴京执导的动作军事电影，由吴京、弗兰克·格里罗、吴刚、张翰、卢靖姗、淳于珊珊、丁海峰等主演。该片于2017年7月27日在中国内地上映 。影片讲述了脱下军装的冷锋被卷入了一场非洲国家的叛乱，本来能够安全撤离的他无法忘记军...
剧情简介 演职员表 角色介绍 音乐原声 幕后花絮 更多>>
baike.baidu.com/
战狼2的最新相关信息
战狼2依然无敌 其他片即便是"炮灰"也各有特色
上周五，《心理罪》《侠盗联盟》《鲛珠传》集体进入电影市场,对票房强势的《战狼2》形成围剿。但姜还是老的辣!从上周末的市场走向来看,《战狼2》依然...
新浪 17分钟前
最全票房信息都在这里!战狼2票房将超6亿... 人民网重庆站 37分钟前
《战狼2》助力国产片票房逆袭 大批国片火... 电影网 45分钟前
专访《战狼2》航拍主飞手:辛苦与压力并存 华夏经纬 4小时前
外媒:《战狼2》触动中国观众的民族情结.... 观察者网 6小时前
战狼2吧_百度贴吧
关注用户：10万人 累计发贴：52万
精品贴(21)
★【战狼2】??【BD1280】[全]完 整出了。在线观... 点击：2402 回复：9
战狼2是我贡献过票房的电影 点击：1334 回复：5
战狼2一出，我就是元老了，哈哈 点击：488 回复：57
查看更多战狼2吧的内容>>
相关影视作品
收起
叶问4
甄子丹主演的动作电影
印囧
徐峥编导公路喜剧片
火蓝刀锋2之雄鹰展翅
杨志刚主演当代军旅剧
使徒行者2
苏万聪执导警匪电视剧
僵尸世界大战2
大闹东海
复仇者联盟3
敢死队4
环太平洋2
金刚狼3
杀破狼2
极限特工3
孤狼特别突击队
利刃出击
拆弹专家
第一滴血5

# 推理

梁启超的儿子的老婆的情人的父亲　搜狗搜索

网页　论坛　知识　新闻　博客　百科　更多　什么是分类搜索　找到约 173,657 条结果

**梁启超的儿子的老婆的情人的父亲：**

徐申如

**推理说明：**梁启超的儿子是梁思成、梁思成的妻子是林徽因、林徽因的情人是徐志摩、徐志摩的父亲是徐申如

梁启超的儿子的老婆的情人的老婆-读书-DoNews.COM
2004年6月15日... 作者 帖子主题： 知道是谁不？ 作者 帖子主题： RE：梁启超的儿子的老婆的情人的老婆 【（shengfang）回复（cool）的大作】陆小慢 作者 帖子主题： RE:...
donewsIT门户 - home.donews.com/.../477746.html - 2004-6-15 - 快照 - 预览

梁启超的儿子的老婆的情人的父亲 最佳答案 搜狗知识搜索
梁启超的儿子的老婆的情人的老婆是谁 - 已解决 搜搜问问 2007-11-25
答：梁启超的儿子呢是中国的著名建筑师梁思成 梁思成的老婆呢叫林薇茵看过《人间四月天》的人应该知道啊 那么林薇茵的情人呢就是大名鼎鼎的徐志摩啦 那么徐志摩的老婆...
梁启超的儿子的老婆的情人的老婆是谁？？？ - 已解决 搜搜问问 2011-3-4
梁启超的儿子的太太的情人的太太分别是谁 - 已解决 搜搜问问 2007-12-9
梁启超的儿子的妻子的情人的老婆是谁？ 百度知道 2006-10-4

**梁启超**

梁启超(1873.2.23—1929.1.19)，生于广东新会。1894年，梁启超提倡变法，并于上海主撰《时务报》，著《变法通议》，刊布报端，启发国人之革新思想。与谭嗣同..
相关阅读

**出生：**1873-02-23 / 广东新会
**逝世：**1929-01-19
**妻子：**李蕙仙 (正室) / 王桂荃 (老婆)
**人物关系：**梁宝瑛 (父亲) / 梁思达 (儿子) / 梁思忠 (儿子) / 梁思懿 (女儿) / 梁思成 (儿子)
**个人名言：**享受工作的同时享受生活

**著作**　更多>>


中国近三百年学..

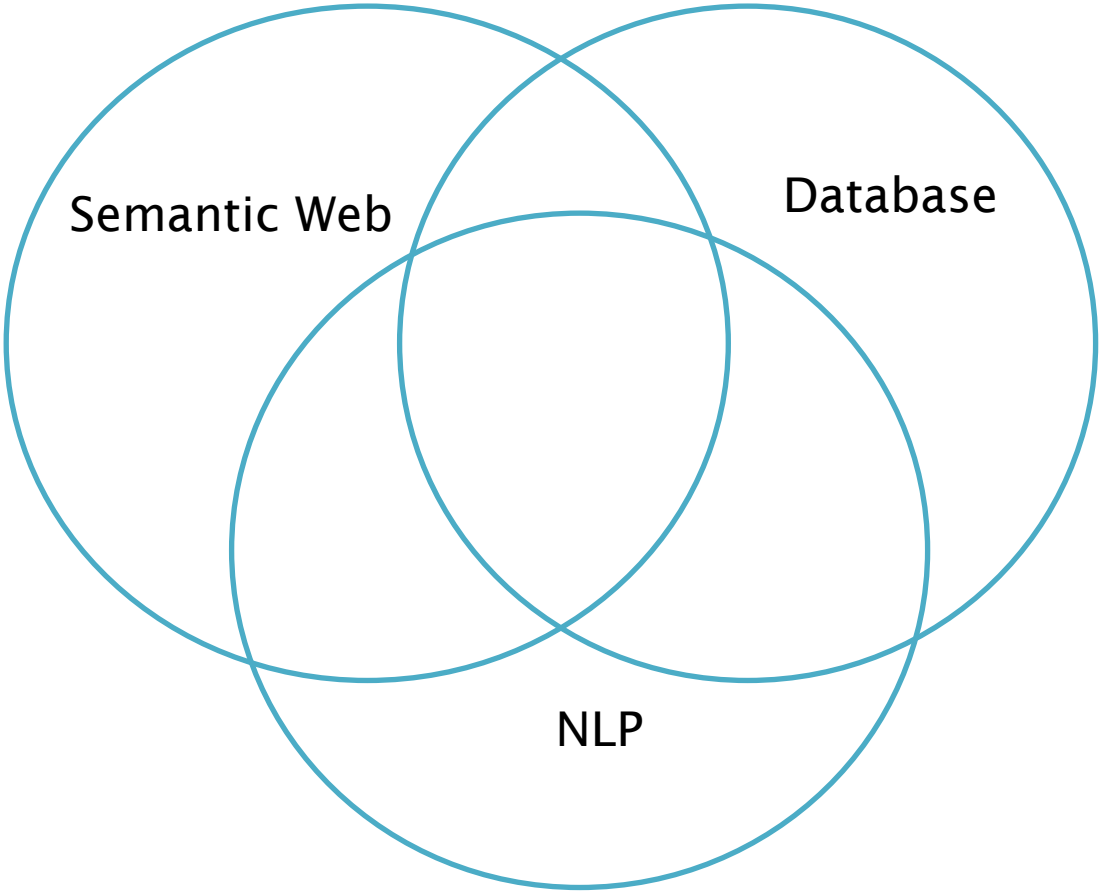

中国历史研究法


新大陆游记

李鸿章传


清代学术概论

# Knowledge Graph 涉及的领域

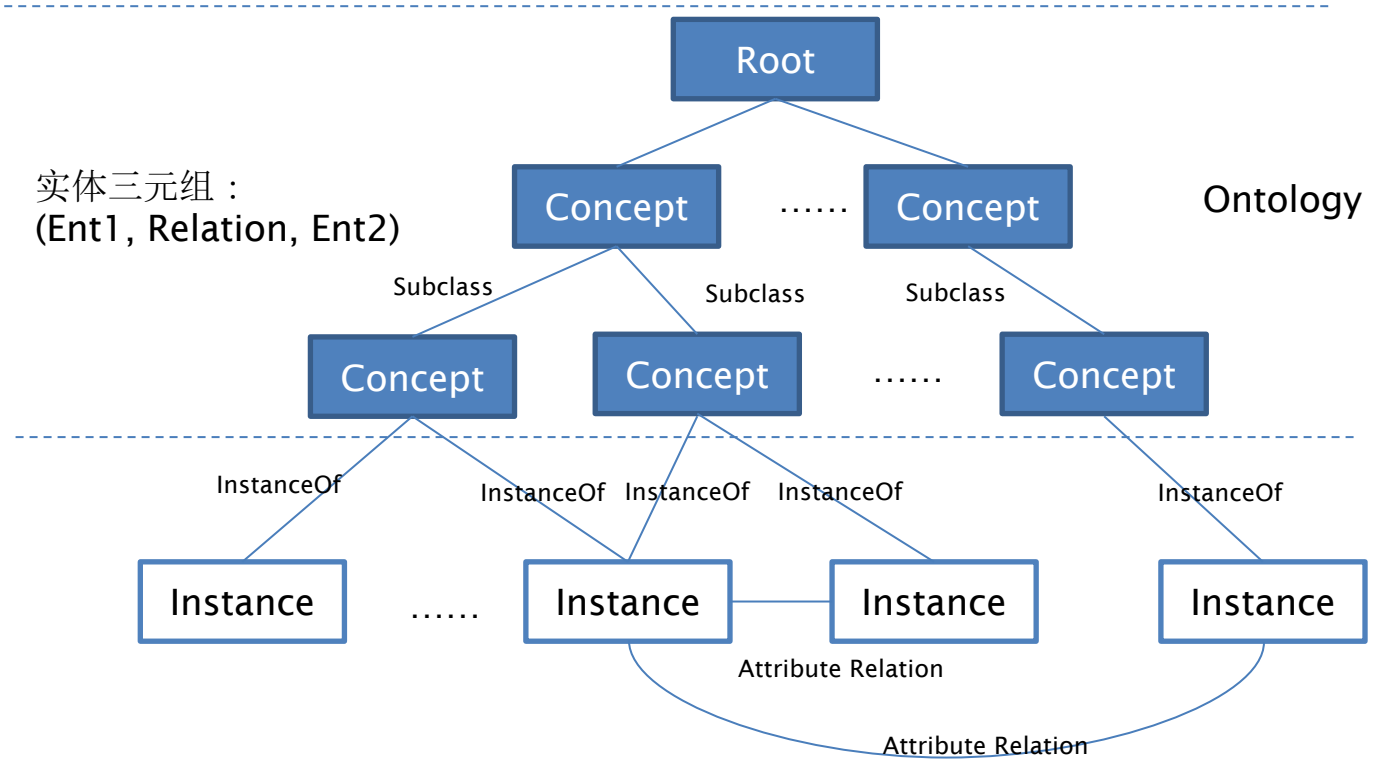


# 什么是知识图谱

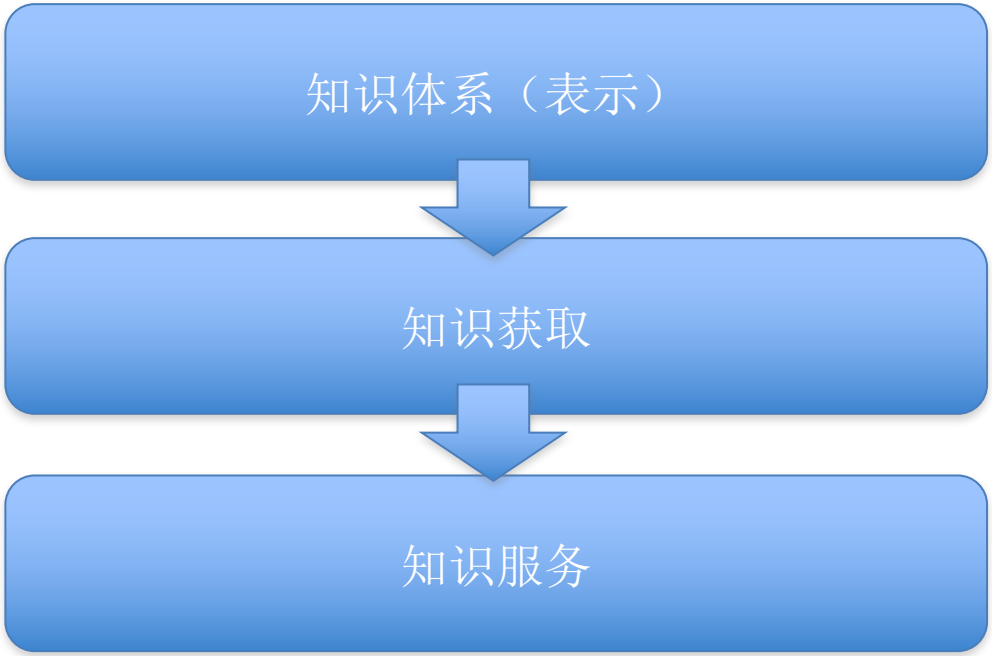
- The Knowledge Graph is a system that understands facts about people, places and things and how these entities are all connected.

- 知识图谱本质上是一种语义网络。其结点代表实体（entity）或者概念（concept），边代表实体/概念之间的各种语义关系

# 知识图谱包含哪些内容

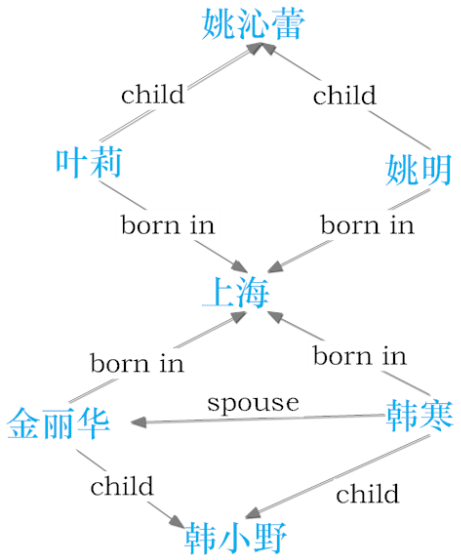

实体三元组：
(Ent1, Relation, Ent2)
Root
Concept
……
Concept
Ontology
Subclass
Subclass
Subclass
Concept
Concept
……
Concept
InstanceOf
InstanceOf
InstanceOf
InstanceOf
InstanceOf
Instance
……
Instance
Instance
Instance
Attribute Relation
Attribute Relation

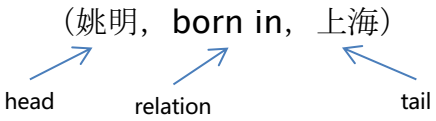
# 三个层面问题



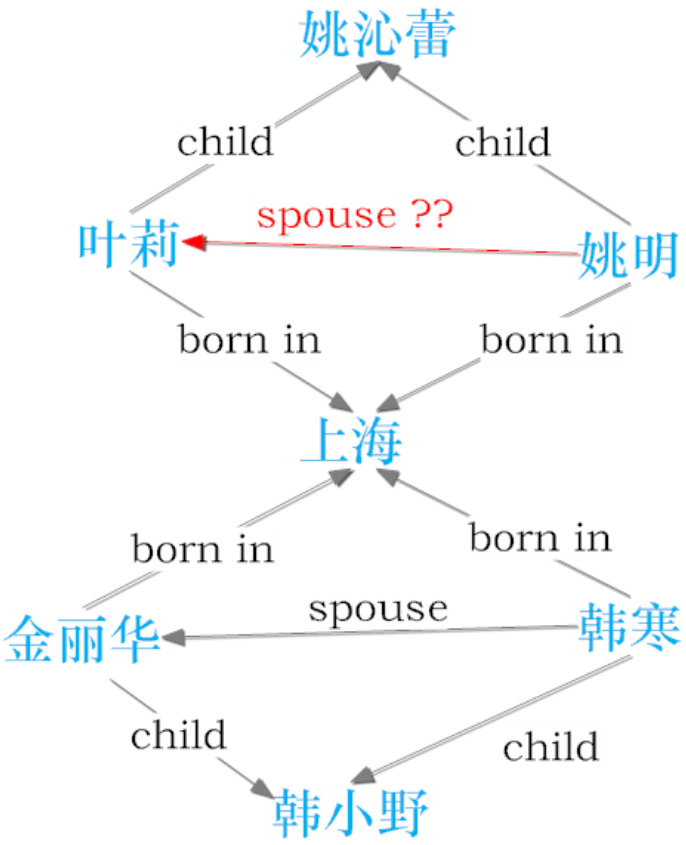
# 知识图谱概览（基于符号的表示）

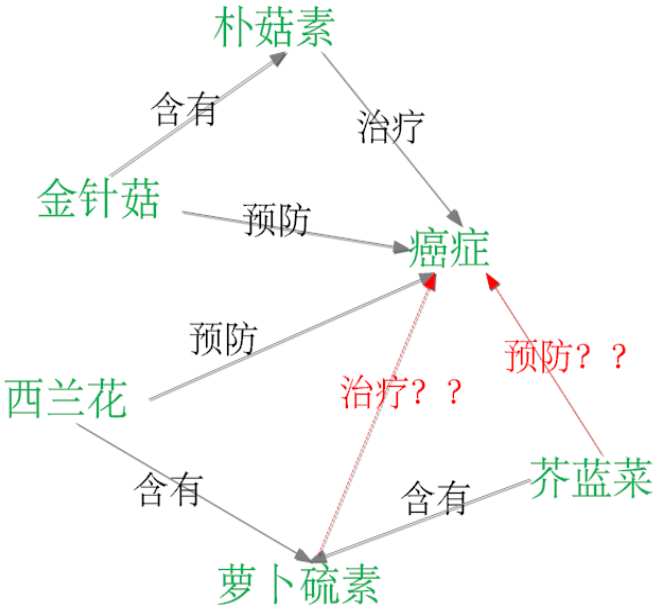
- 知识库是一个有向图
  - 多关系数据(multi-relational data)
  - 节点：实体/概念
  - 边：关系/属性
  - 关系事实 = *(head, relation, tail)*
    - head：头部实体
    - relation：关系/属性
    - tail：尾部实体



# 知识图谱概览（基于符号的表示）

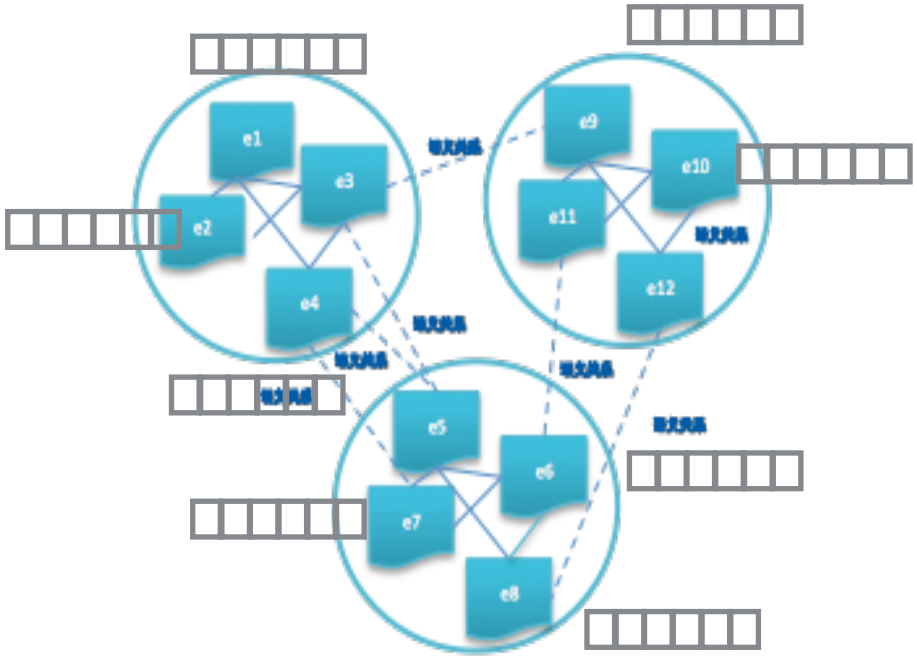



$\text{child}(A,B) \wedge \text{child}(A,C) \Rightarrow \text{spouse}(B,C)$

$\text{含有}(A,B) \wedge \text{治疗}(B,C) \Rightarrow \text{预防}(A,C)$

# 知识图谱概览（分布式表示）



# 知识体系组织形式

- Ontology vs. Knowledge Base
  - Ontology：共享概念化的规范，涉及概念、关系和公理三个要素
  - Knowledge：服从于ontology 控制的知识单元的载体
  - Ontology是蛋糕的模具，Knowledge Base是蛋糕

- 公理：Formal Ontology vs. Lightweight Ontology
  - Formal Ontology: 大量使用公理
  - Lightweight Ontology: 不用或很少使用公理

# 知识体系组织形式

- Ontology
  - 树状结构，不同层节点之间是严格的IsA关系
  - 优点：可以适用于知识的推理
  - 缺点：无法表示概念的二义性（运动员：体育？人物？）
- Taxonomy
  - 树状结构，上下位节点之间非严格的IsA关系
  - 优点：可以表示概念的二义性（体育→运动员）
  - 缺点：不适用于推理，无法避免概念冗余（餐厅：美食？机构？地点？）
- Folksonomy
  - 类别标签，更加开放
  - 优点：能够涵盖更多的概念
  - 缺点：如何进行标签管理？

# 知识体系组织形式

- 目前的知识资源多是采用Folksonomy与Taxonomy相结合的组织形式
  - 但是能够覆盖的类别还很少

| *全部* | 含有开放分类(Folksonomy)的页面数比例 |
| --- | --- |
| 互动百科 | 70.19% |
| 百度百科 | 64.38% |


  - 这些开放式类别标签存在冗余、不规范的问题，标签之间也缺乏关联
    - 体育、人物
    - 1980年、购房、房产、房地产……

# 知识体系组织形式

- 类别属性定义不统一
  - 已有的体系框架
    - GeoNames
    - DBpedia Ontology
    - TexonConcept Ontology
    - KOS
    - Schema.org
      - 1）面对站长，而不是面对知识
      - 2）体系覆盖度不足，局限于英文
      - 3）细致化不足

- Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries ...
- Embedded non-text objects: AudioObject, ImageObject, VideoObject
- Event
- Health and medical types: notes on the health and medical types under MedicalEntity.
- Organization
- Person
- Place, LocalBusiness, Restaurant ...
- Product, Offer, AggregateOffer
- Review, AggregateRating

Schema.org

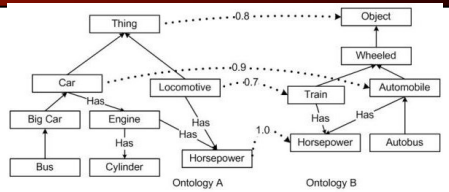
| | | | |
|---|---|---|---|
| 中文名: | 李娜 | 籍贯: | 武汉市 |
| 性别: | 女 | 民族: | 汉族 |
| 国籍: | 中国 | 出生年月: | 1982年2月26日 |
| 星座: | 双鱼座 | 职业: | 运动员 女子网球选手 |
| 毕业院校: | 华中科技大学 | 身高: | 172厘米 |

互动百科

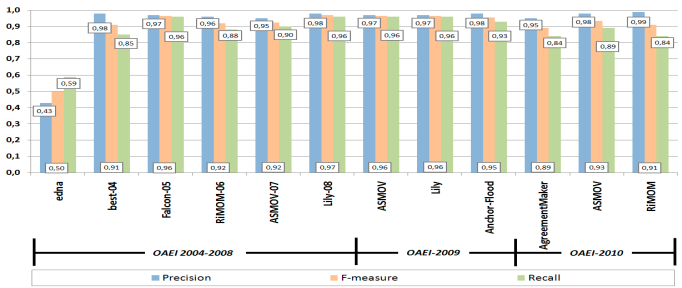
| | | | |
|---|---|---|---|
| 中文名 | 李娜 | 主要奖项 | |
| 外文名 | Li Na | 重要事件 | |
| 别　名 | 娜姐 | | |
| 国　籍 | 中国 | | |
| 民　族 | 汉 | 启蒙教练 | |
| 出生地 | 湖北省武汉市江岸区 | 训练地 | |
| 出生日期 | 1982年2月26日 | 教　练 | |
| 毕业院校 | 华中科技大学（新闻学专业） | 丈　夫 | |

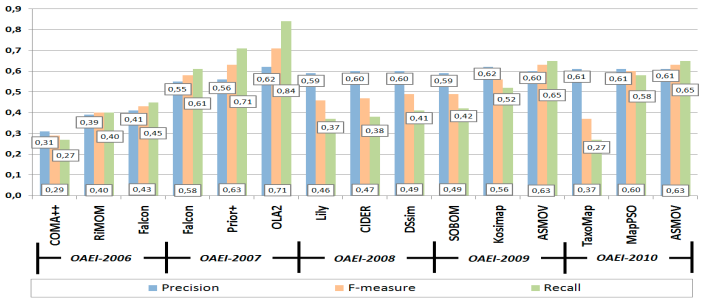
百度百科

# Ontology Matching




- 建立体系间的Alignment
  - 挖掘概念之间SameAs关系
  - 评测：Ontology Alignment Evaluation Initiative
    - 2004-2013
      - Benchmarks (bibliographic references), Web directories, Anatomy (biomedical)
    - 关键：概念之间的相似度计算
    - 挑战
      - Large-scale ontology matching and evaluation
      - Matching with background knowledge (Increase recall but hurt precision)
      - Multiple matchers and selection
      - Incorporating social information

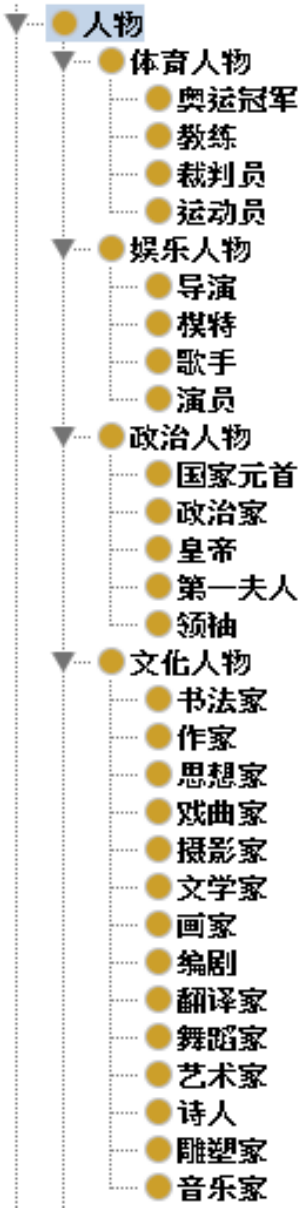


# KG基本概念

- Node：概念（Concept）



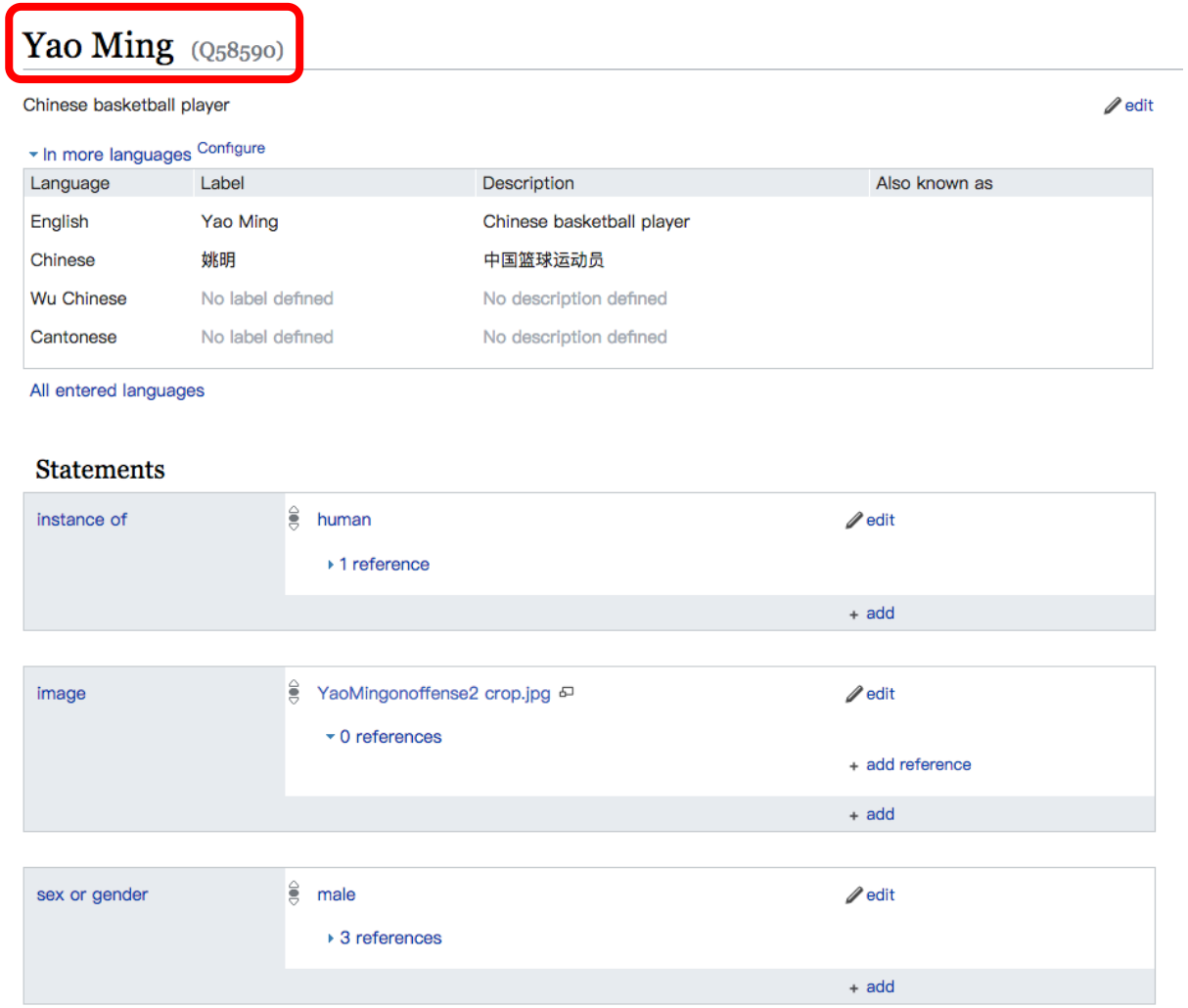


# KG基本概念

- Node：领域（Domain/Topic）


# KG基本概念

- Node：实例/实体（Entity/Objects/Instance）



**Yao Ming** (Q58590)

Chinese basketball player — edit

In more languages Configure

| Language | Label | Description | Also known as |
|---|---|---|---|
| English | Yao Ming | Chinese basketball player | |
| Chinese | 姚明 | 中国篮球运动员 | |
| Wu Chinese | No label defined | No description defined | |
| Cantonese | No label defined | No description defined | |

All entered languages

**Statements**

| | | |
|---|---|---|
| instance of | human<br>1 reference | edit<br>+ add |
| image | YaoMingonoffense2 crop.jpg<br>0 references | edit<br>+ add reference<br>+ add |
| sex or gender | male<br>3 references | edit<br>+ add |

# KG基本概念

- Node：值（Value）
  - 实体（Entity）
    - （姚明，出生地，上海市）
  - 字符串（String）
    - （北京大学，学术传统，兼容并包、思想自由）
  - 数字（Number）
    - 平方公里：（北京市，面积， 1.641万）
    - 公斤：（姚明，体重，140公斤）
    - 米：（姚明，身高， 2.29米）
    - …
  - 时间（Date）
    - （姚明，出生年份，1981年）
  - 枚举（Enumerate）
    - （姚明，性别，男）
  - ……

# KG基本概念

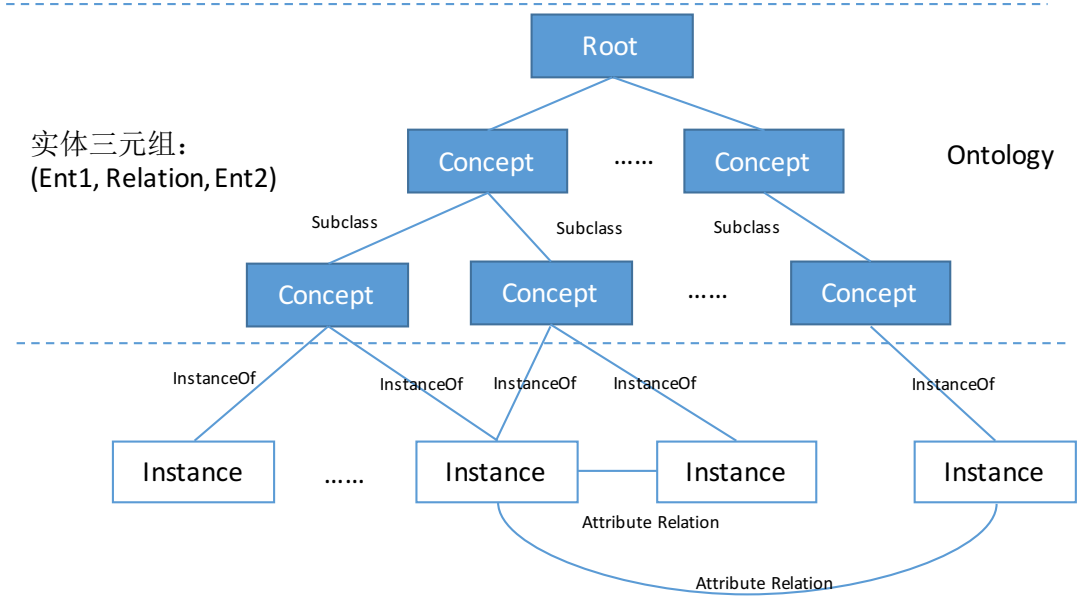
- 边：关系
  - Subclass
  - Type
  - Relation
  - Property、Attribute

(狗， Is-A， 哺乳动物)

(旺财， Is-A， 狗)

(旺财， 朋友， 小白)

(旺财， 颜色， 黄色)



# KG基本概念

- 关系：Taxonomic Relation vs. Non-taxonomic Relation
  - Taxonomic Relation： is-a/Hypernym-Hyponym
  - Non-taxonomic Relation：概念之间的相互作用
    - Meoronymy 部分整体
    - Thematic roles 论旨角色
    - Attribute 属性
    - Possession 领属
    - Casuality 因果
    - ……

# KG基本概念

- Node：高阶三元组
  - 与时间、地点相关
    - （（美国，总统，特朗普），开始时间，2017）
  - 事件
    - Compound Value Type

A Compound Value Type is a Type within Freebase which is used to represent data where each entry consists of multiple fields. Compound value types, or CVT's are used in Freebase to represent complex data.

# Example of CVT

**Donald Trump** (Q22686)

spouse

Ivana Trump
- start time: 7 April 1977
- end time: 22 March 1992

1 reference
- reference URL: http://hollywoodlife.com/celeb/ivana-trump/
- quote: Ivana married Donald Trump on April 7, 1977. (English)

Melania Trump
- start time: 22 January 2005
- place of marriage: Mar-a-Lago

1 reference
- reference URL: http://www.hollywoodreporter.com/features/trumps-wedding-melania-bill-hill-880088
- quote: on Jan. 22, 2005, it was a different story. Trump married model Melania Knauss (English)

Marla Maples
- end time: 8 June 1999
- start time: 19 December 1993

0 references

# Example of CVT

Michael Jordan (Q41421)

member of sports team

North Carolina Tar Heels men's basketball — edit

| | |
|---|---|
| end time | 1984 |
| start time | 1981 |
| position played on team / speciality | shooting guard |
| | swingman |
| sport number | 23 |
| captain | *no value* |

0 references

+ add reference

Chicago Bulls — edit

| | |
|---|---|
| end time | 1993 |
| start time | 1984 |
| position played on team / speciality | shooting guard |
| sport number | 23 |
| captain | yes |

0 references

+ add reference

Scottsdale Scorpions — edit

| | |
|---|---|
| end time | 1994 |
| start time | 1994 |
| position played on team / speciality | outfielder |
| sport number | 35 |

0 references

+ add reference

# 知识分类

- 百科知识

标注系统 管理中心 导入体系 导入数据 浏览体系 共指列表 高级管理 你好：admin 修改日志 Log out

-root
-创造性工作 [外文名,中文名]
+绘画 [别名,作者,尺寸,类型,年代,收藏单位]
+雕塑 [材料,高度,别名,作者,寓意,类型,年代]
+电视剧 [集数,导演,颜色,语言,编剧,地区,主演,别名,发行时间,获得荣誉,上映时间,制片人,类型,出品公司]
+音乐 [语言,地区,曲长,别名,发行时间,作曲者,类型]
+书籍 [isbn,语言,装帧,页数,开本,出版社,发行时间,作者,别名,字数,类型,价格]
+电影 [导演,颜色,语言,编剧,片长,主演,别名,发行公司,发行时间,获得荣誉,上映时间,imdb编码,制片人,分级,类型,出品公司,地区]
+软件应用 [语言,发行时间,开发者,编程语言,操作系统]
-组织 [外文名,所在地,中文名,地址,别名,创建时间]
+教育组织 [类型]
+运动队 [主教练,代表队员,获得荣誉,所在联赛,运动项目]
+非政府组织 [创始人]
+政府机构 []
+公司 [注册资本,上市代码,公司口号,经营范围,法人代表,上市市场,证券简称,年盈利,总部所在地,宗旨理念,员工数,产品,创始人,总资产,行业,净利润,性质]
-人物 [出生地,外文名,毕业院校,政党,职业,籍贯,去世日期,别名,信仰,国籍,中文名,体重,血型,星座,出生日期,性别,身高,相关事件,民族]
+网络人物 [代表作品,艺名,获得荣誉,主要成就]
+文化人物 [代表作品,获得荣誉,主要成就]
+娱乐人物 [艺名,获得荣誉,主要成就,经纪公司]
+政治人物 [获得荣誉,主要成就]
+虚拟人物 []
+体育人物 [运动项目,获得荣誉,主要成就]
+经济人物 [获得荣誉,主要成就]
+社会科学人物 [获得荣誉,主要成就]
+自然科学人物 [获得荣誉,主要成就]
-地点 [外文名,所在地,面积,中文名,别名,位置]
+公共设施 []
+旅游景点 [主要景点,开园时间,闭园时间,邮政区码,地址,电话区码,门票,海拔,创建时间,分级,竣工日期]
+行政区域 [GDP,主要景点,民族,现任领导人,著名高校,下辖地区,方言,邮政区码,知名企业,车牌代码,火车站,电话区码,创建时间,机场,时区,人口,政府驻地,知名产业,特产,名人]
+地形地貌 [气候]

# 知识分类

- 领域知识

股票： tags［行业，地区，板块，股票种类］

{

基本属性：{

股票代码（ID）[String] (六位整形数字):

股票种类 [String] (A B H N S股):

股票简称 [String]

股票英文简称 [String]:

上市日期 [Date] (1980.01.01-now):

上市地点 [String]:

上市板 [String]:

交易币种 [String]:

股票面值 [double] (>=0):

摘牌日期 [Date] (1980.01.01-now):

}

发行属性：{

公司名称: [公司实体 Id]

成立日期 [Date]:

上市日期 [Date]:

发行数量(万股) [Double]

发行价格(元) [Double]:

发行市盈率 [Double]:

预计募资(万元) [Double]:

实际募资(万元) [Double]:

主承销商 [String]:

上市保荐人 [String]:

}

# 知识分类

- 事实性知识
  - (乔布斯，CEO，苹果）
  - (中华人民共和国，首都，北京）

- 主观性知识

“我今年天让入手诺基亚5800，把玩不到24小时，目前感觉5800屏幕很好，操作也很方便，通话质量也不错，但是外形有些偏女性化，不适合男生。这些都是小问题，最主要的问题是电池不耐用，只能坚持一天，反正我觉得对不起这个价格。”

- 外形
- 电池

- 屏幕
- 操作
- 通话质量

# 知识分类

- 场景知识
  - 打人，打篮球
  - MJ出版了三部专辑→Michael Jackson
  - MJ获得了NBA总冠军→Michael Jordan
  - 订机票的步骤，红烧肉的做法
- 语言知识
  - (乔丹，SameAs， 佐敦)
  - (乔丹，SameAs, Jordan)
  - (Microsoft， SameAs, MS)
  - (hasFounded, SameAs, isFounderof)
- 常识知识 (Common-sense Knowledge)
  - hasAbility(鸟，飞）， hasAbility(人，说话）
  - hasProperties(水，透明）， hasShape(球，圆的）
  - moreHeavy(大象，小狗）
  - Mother(x,y) ^ Brother(z,x) ➔ Uncle(z,y)

# 目录

# 知识图谱系统的架构

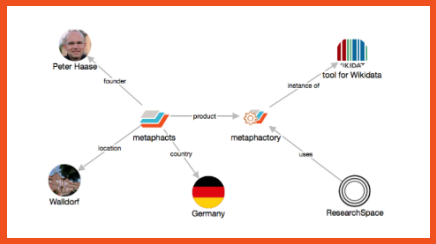
**Applications**

- Semantic Search
- Analytics
- Knowledge Sharing
- Question Answering
- Dashboards
- Knowledge Management

**Algorithms**

- Inferencing
- Entity Recognition
- Text Understanding
- Machine Learning
- Disambiguation
- Recommendations

**Knowledge Graph**




- Entities
- Relationships
- Semantic Descriptions

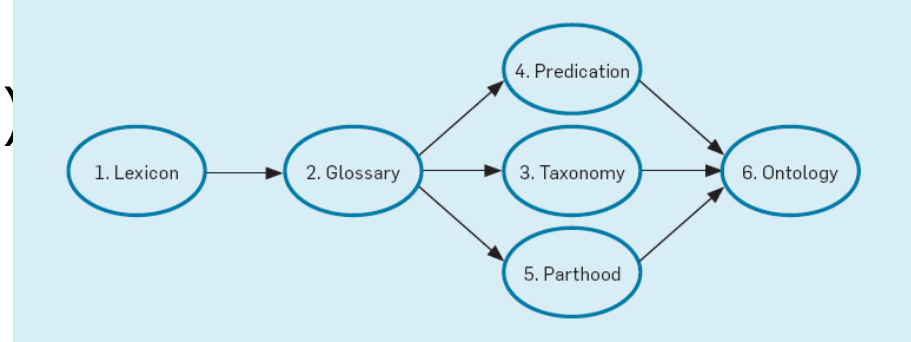
**Data Sources**

Data Transformation, Integration
Natural Language Processing

XML SQL CSV

# 生命周期—领域知识建模

- **输入：**
  - 目标领域 (医疗、金融...)
  - 应用场景




- **输出：**领域知识本体
  - 领域实体类别体系
  - 实体属性
  - 领域语义关系
  - 语义关系之间的关系

- **关键技术：**
  - Ontology Engineering

Ontology engineering

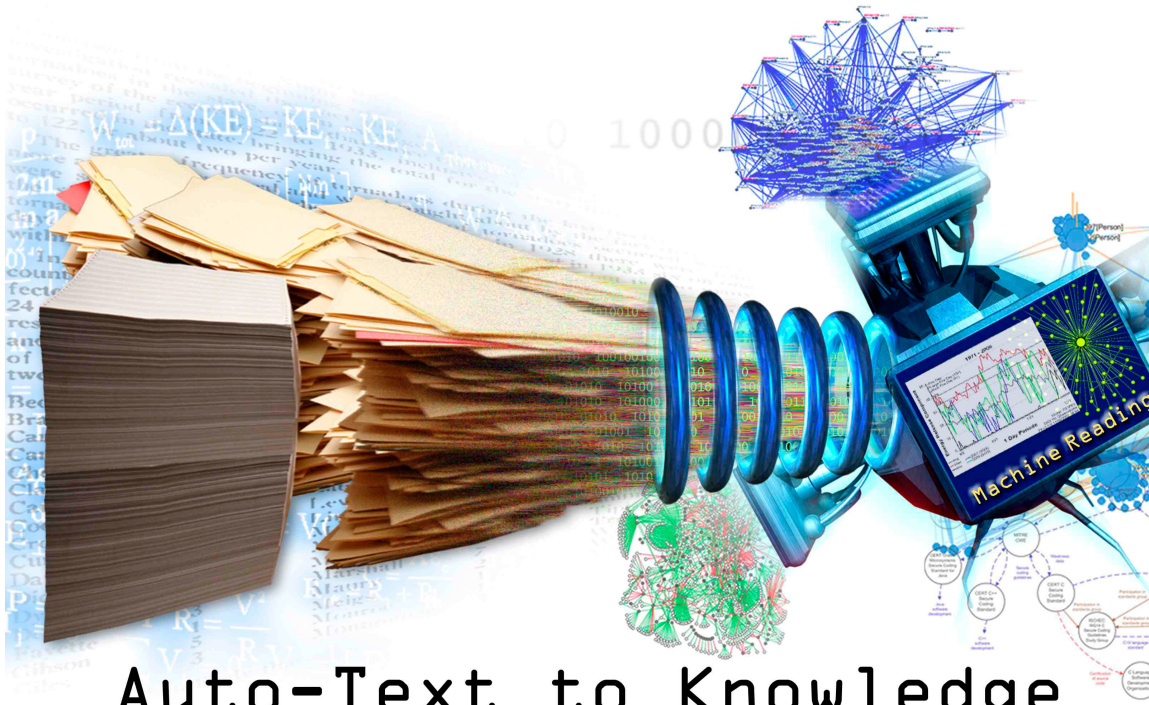
File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **Term** | **Synonyms** | **Kind** | **Description [source]** |
| 2 | Delivery address | Shipping address | Complex property | Location to which goods are to be sent [1]. |
| 3 | Invoice | Bill | Object | Itemized list of goods shipped, usually specifying the price and the terms of sale [2]. |
| 4 | Postal address | Address | Complex property | Information that locates and identifies a specific address, as defined by the postal service [3]. |
| 5 | Purchasing conditions | Purchase terms and conditions | Object | Conditions related to the transaction and the trade [4]. |
| 6 | Purchase order | PO | Object | Commercial document issued by a buyer to a seller, indicating types, quantities, and agreed prices for products or services the seller will provide to the buyer [5]. |
| 7 | Customer | Client | Actor | One who purchases a commodity or service [2]. |
| 8 | Invoicing | Issuing invoice | Process | Making or issuing an invoice for goods or services [6]. |
| 9 | Purchasing | Buying | Process | Acquisition of something for payment [6]. |

1. Lexicon | 2. Glossary | 3. Taxonomy | 4. Predication | 5. Parthood

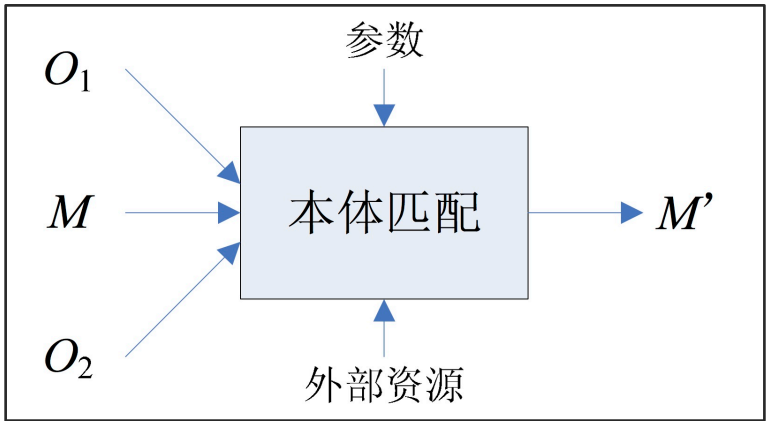
# 生命周期—知识获取

- **输入：**
  - 领域知识本体
  - 海量数据：文本、垂直站点、百科
- **输出：**领域知识
  - 实体集合
  - 实体关系/属性
- **主要技术：**
  - 信息抽取
  - 文本挖掘



# 生命周期—知识集成

- **输入：**
  - 抽取出来的知识
  - 现有知识库
  - 知识本体

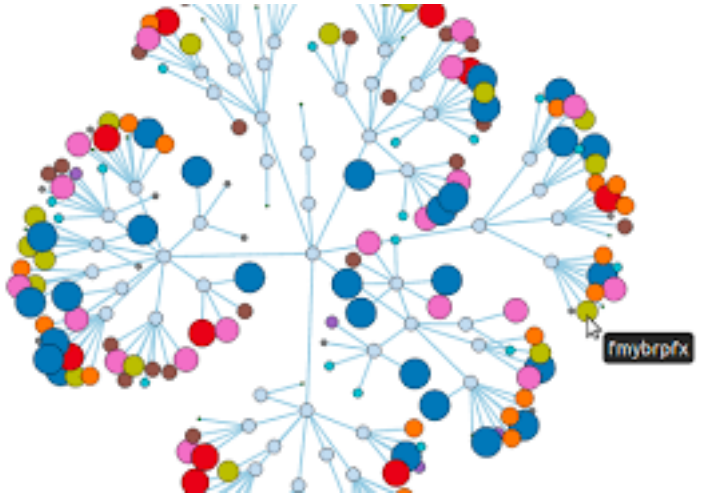



- **输出：**
  - 知识置信度
  - 统一知识库


- **关键技术：**
  - Ontology Matching
  - Entity Linking

# 生命周期—知识存储/查询/推理

- **输入：**
  - 大规模知识库知识
- **输出：**
  - 知识库存储/查询/推理服务
- **主要技术：**
  - 知识表示
  - 知识查询语言
  - 存储/检索引擎
  - 推理引擎



# 知识图谱的生命周期

- **知识建模**
  - 建模领域知识结构
- **知识获取**
  - 获取领域内的事实知识
- **知识集成**
  - 估计知识的可信度，将碎片知识组装成知识网络
- **知识存储**
  - 提供高性能知识服务

# 目录

# 代表性知识图谱

- **人工构建知识图谱**
  - WordNet
  - CYC
- **基于Wikipedia的知识图谱**
  - Yago
  - DBPedia
  - Freebase
- **文本抽取知识图谱**
  - NELL

# 代表性知识图谱

- **人工构建知识图谱**
  - WordNet
  - CYC
- **基于Wikipedia的知识图谱**
  - Yago
  - DBPedia
  - Freebase
- **文本抽取知识图谱**
  - NELL

# WordNet

WordNet
A lexical database for English

- **WordNet是什么?**
  - 一部在线词典数据库系统，采用了与传统词典不同的方式，即按照词义而不是词形来组织
  - 词语被聚类成词义簇(synset)，词义之间通过语义关系连接成大的概念网络
- 由普林斯顿大学认知科学实验室在1985年建立

# WordNet包含的知识

**描述的对象**

- compound（复合词）
- phrasal verb（短语动词）
- collocation（搭配词）
- idiomatic phrase（成语）
- word（单词）

**对象之间的语义关系**

- 同义反义关系（synonymy，antonymy）
- 上下位关系（hyponymy，hypernym，troponymy）
- 部分整体关系（entailment，meronymy）

**部分句法信息**

- 简单的动词基本句式信息（Verb Sentence Frames）

# WordNet的核心概念

- **Synset：** WordNet 将英语的名词、动词、形容词、和副词组织为Synsets，每一个Synset表示一个基本的词汇概念

- **概念关系**
  - 同义关系
  - 反义关系
  - 上位关系
  - 下位关系
  - 整体关系（名词）
  - 部分关系（名词）
  - 蕴含关系（动词）
  - 因果关系（动词）
  - 近似关系（形容词）

**newspaper词义的上位synsets**

newspaper, paper
=> press, public press
=> print media
=> medium
=> instrumentality
=> artifact, artefact
=> whole, unit
=> object, physical object
=> physical entity
=> entity

# WordNet组织示例

Concept hierarchy

{equipment}

Concept ("synset")

Synonymous words for the concept

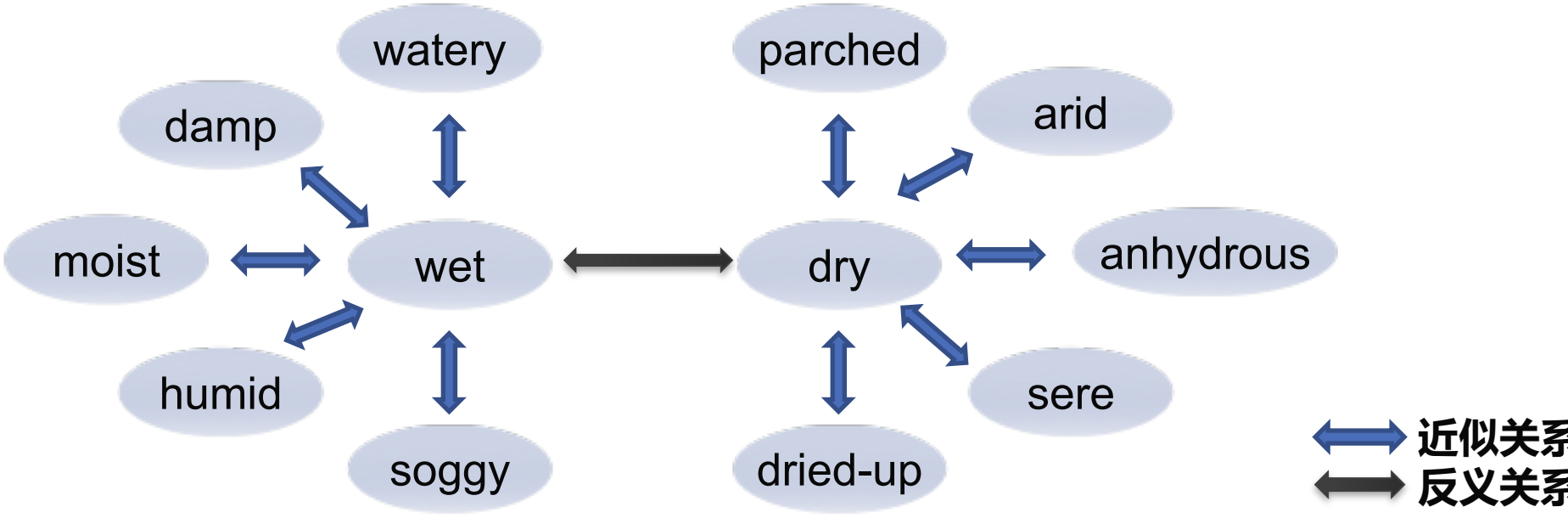
{ camera, photographic camera}

{ camera, television camera}

Gloss describing the concept

"television equipment consisting of a lens system that focuses an image on a photosensitive mosaic"

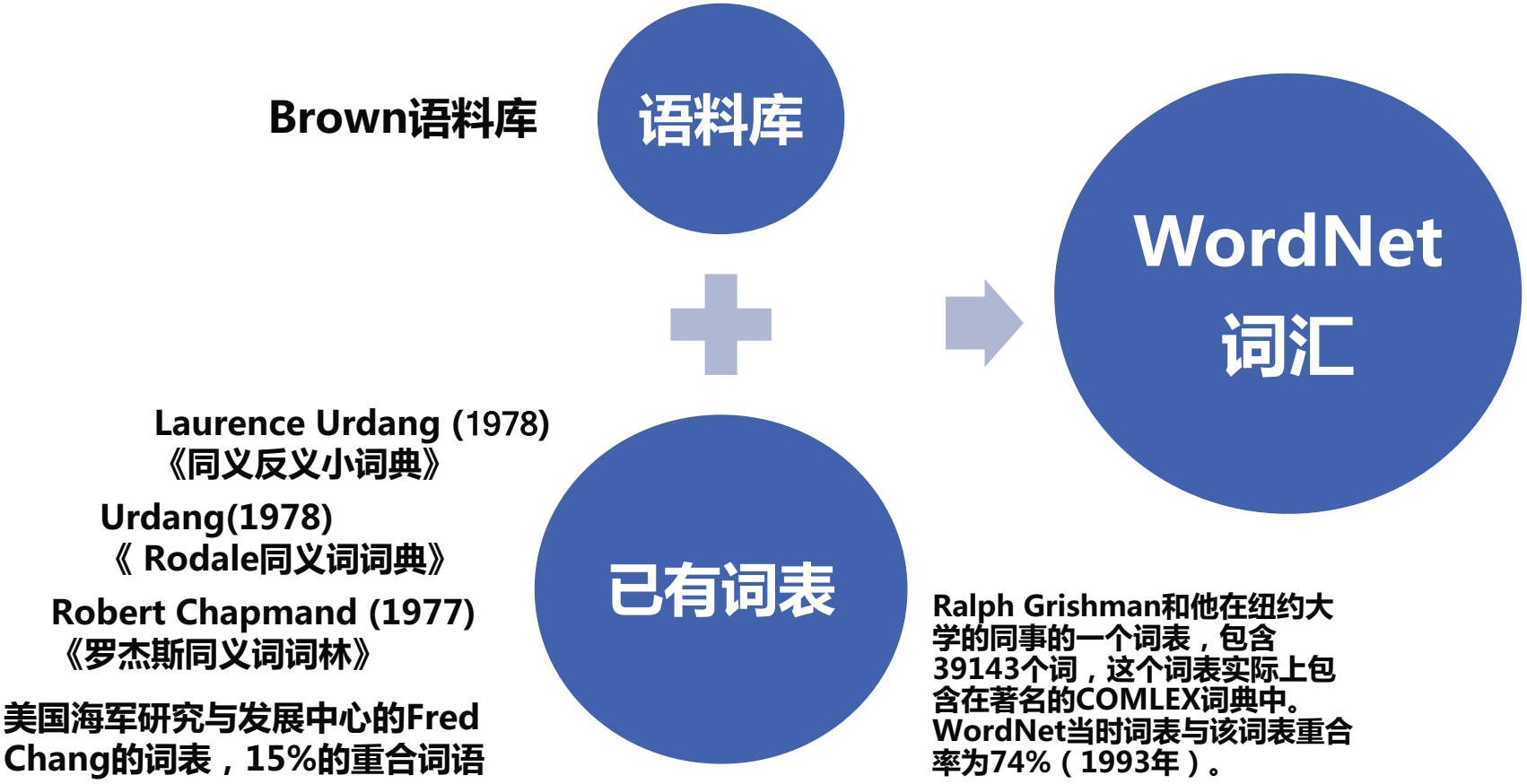
# WordNet组织示例




**基于反义、近义组织的形容词synset**

# WordNet的构建方法

- **人工构建+机器辅助（后续有很多自动构建技术研究）**



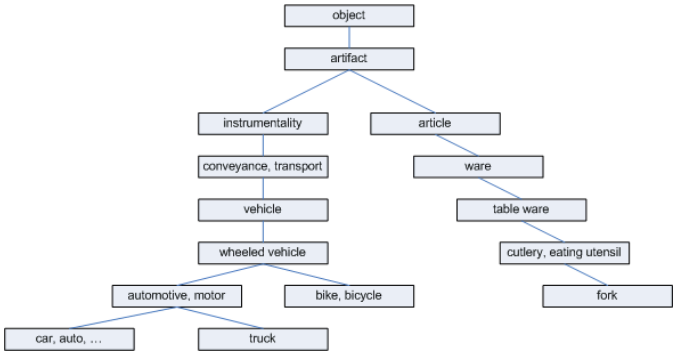
# WordNet规模

| POS | Unique Terms | Synsets | Total Word-Sense Pairs |
|---|---|---|---|
| Noun | 109195 | 75804 | 134716 |
| Verb | 11088 | 13214 | 24169 |
| Adjective | 21460 | 18576 | 31184 |
| Adverb | 4607 | 3629 | 5748 |
| Totals | 146350 | 111223 | 195817 |

# WordNet的应用

- **WordNet在自然语言处理中被广泛应用**
  - 作为词义消歧的目标知识库
  - 作为高质量的Taxonomy
  - 用于计算语义相似度

| Word 1 | Word 2 | lin | wup | path | remarks |
|---|---|---|---|---|---|
| genuineness | genuine | 0 | 0 | 0 | needs explanation |
| Valid | reality | 0 | 0 | 0 | needs explanation |
| Painter | paint | 0.15 | 0.62 | 0.09 | Good Enough |
| Really | fact | 0 | 0 | 0 | needs explanation |
| Real | reality | 0.1 | 0.3 | 0.09 | Good Enough |
| really | reality | 0 | 0 | 0 | needs explanation |
| paint | painting | 0.3 | 0.8 | 0.12 | Good Enough |




WordNet: a lexical database for English

GA Miller - Communications of the ACM, 1995 - dl.acm.org

Abstract Because meaningful sentences are composed of meaningful words, any system that hopes to process natural languages as people do must have information about words and their meanings. This information is traditionally provided through dictionaries, and

被引用次数：9323　相关文章　所有 34 个版本　引用　保存　更多

# WordNet综述

| | |
|---|---|
| Content | Adjectives, verbs, nouns and adverbs of the English language |
| Format | Visualization tool<br>data downloadable in Prolog-like format |
| Main strength | High quality lexicon for English |
| Technique | Manual |
| Size | Words: 155k |
| | Senses: 117k |
| | Word-sense pairs: 207k |
| License | Proprietary, free use |
| Reference | [Miller, Comm ACM 1995] |
| URL | http://wordnet.princeton.edu |

一个高质量英文电子词典和本体

# 代表性知识图谱

- **人工构建知识图谱**
  - WordNet
  - **CYC**
- **基于Wikipedia的知识图谱**
  - Yago
  - DBPedia
  - Freebase
- **文本抽取知识图谱**
  - NELL

# Cyc


- **Cyc** 是一个由Douglas Lenat 在1984年启动的人工智能项目，其目的是构建**一个完整的、机器可使用的本体体系和人类常识知识库**
  - 500万条知识
  - 50万概念
- **OpenCyc** 是其开放出来免费供大众使用的一部分知识
  - 24万概念
  - 200万条知识
- **ResearchCyc**是提供Research Liscence供研究使用的完整版

# Cyc构建方法--人手工构建

- 1986年, Doug Lenat估计整个Cyc需要包括**25万条规则，耗费350人年**
  - 这是一个明显的低估

- 近年来，也开始使用自动构建方法，从自然语言中抽取知识

- 2008年开始，Cyc开始将其资源与Wikipedia、DBpedia和Freebase等资源开始建立Link

# 知识表示语言--CycL

- **CycL:** 自己设计的知识表示语言
  - 基于First Order Predicate Logic (FOPL)
  - 采用类似于LISP语言的语法
- **Constants:** 概念名字
  - *Cyc, DougLenat, BaseKB, EnglishWord*
- **Variables**
  - 以?开头
  - *?TYPE*
- **Predicates**
  - *isa, genls, comment*
- **Functions**
  - *(FruitFn AppleTree) → Apple*

# 知识表示语言-CycL

- **Formulas**
  - 形式为(*predicate arg1 arg2 ...*)的表达式
  - 5种真值: {*true, default true, false, default false, unknown*}
  - 示例:
    - *(isa Dog BiologicalSpecies)*
    - *(genls Dog Carnivore)*
    - *(skillCapableOf LinusVanPelt PlayingAMusicalInstrument performedBy)*
- **Logical connectors**
  - *not, and, or, implies*
- **Quantifiers**
  - *任一, 存在*

# Cyc中包含的知识

- Cyc中包含的大部分知识都是**常识**
  - *Every tree is a plant(所有树都都是植物)*
  - *Plants die eventually（植物都会死）*

- 核心概念：
  - **Individuals**: #$BillClinton, #$France
  - **Collections**:
    - #$Tree-ThePlant（包含了所有的树）
    - #$EquivalenceRelation(包括所有的等同关系)

# Cyc中包含的知识-Sentences

- **事实通过CycL sentences 来表示**
  - *(#$isa #$BillClinton #$UnitedStatesPresident)*
  - *(#$genls #$Tree-ThePlant #$Plant):"All trees are plants".*
- **Rules:** 包含变量(以?开头)的句子
  - 如果一个对象是SUBSET的成员，同时SUBSET是SUPERSET的子类，那么OBJ是SUPERSET的成员

```
(#$implies
   (#$and
     (#$isa ?OBJ ?SUBSET)
     (#$genls ?SUBSET ?SUPERSET))
   (#$isa ?OBJ ?SUPERSET))
```

# Predicate usage in Upper Cyc

| *Freq.* | *Predicate* | *Description* |
|---|---|---|
| 4503 | isa | 实例类别 |
| 2695 | comment | 描述term的用法 |
| 2565 | genls | 类别上下位关系 |
| 920 | arg1Isa | argument 1 constraint |
| 836 | arg2Isa | argument 2 constraint |
| 525 | genlPreds | 谓词上下位关系 |
| 301 | not | 逻辑非连接符 |
| 243 | resultIsa | 函数的输出类别 |
| 120 | arg3Isa | argument 3 constraint |
| 107 | implies | 逻辑蕴含关系 |

# The Cyc Ontology and Content

Tens of thousands of Predicates

**between, owns, isa, beliefs**

Hundreds of thousands of Concepts

**CityOfLjubljanaSlovenia, Witbrock, Slovenia, SlovenianLanguage**

Cyc contains:

| | |
|---|---|
| >15,000 | Predicates |
| >300,000 | Concepts |
| >3,500,000 | Assertions |

# 推理引擎与应用

- Cyc提供了非常多的推理引擎，支持演绎推理和归纳推理；同时也提供了扩展推理机制的模块
- 支持自然语言的解析

| | |
|---|---|
| English Words | 18,796 |
| Syntactic Frame Links | 23,336 |
| Single-word Denotation Mappings | 27,681 |
| Multi-word Phrase Denotation Mappings | 44,298 |
| Verbal Semantic Frame Links | 3,701 |
| Noun Semantic Frame Links | 2,578 |
| WordNet 2.0 Links | 11,322 |
| Names<br>(Includes chemical symbols, person/place/organization names, acronyms, etc.) | 100,811 |
| Predicate-based Phrasal Links<br>(genTemplates for paraphrase) | 9,637 |

# 基于Cyc的自然语言解析示例

As of Feb (#$February). 24 (24), Air Force (#$UnitedStatesAirForce) officials (#$PublicOfficial #$OrganizationRepresentative) reported (#$RegisteringAComplaint #$Reporting) that personnel (#$Employee) in the area (#$Area 0 #$FieldOfStudy #$Region-Underspecified) numbered (#$Counting) close to 8,000 (8000). The 100 (100) aircraft (#$AirTransportationDevice) based (#$Base-Support #$MilitaryBase-Grounds #$BaseOfLandProtrusion #$NitrogenBase #$ChemicallyBasicSubstance) in Saudi Arabia (#$SaudiArabia) for patrols (#$Patrolling) over southern Iraq ((#$SouthernRegionFn #$Iraq)) has (#$possesses) seen (#$VisualPerception #$MeetingSomeone #$sees) the addition (#$DoingAddition) of two (2) dozen (12) F-15 (#$FighterPlane-F15) and F-16 fighter jets (#$FighterPlane-F16) to Bahrain (#$Bahrain-TheIsland #$Bahrain (#$CityNamedFn Bahrain #$Bahrain)). The Air Force (#$UnitedStatesAirForce) has (#$possesses) also authorized (#$GrantingPermission) the dispatch (#$SendingSomething) of 12 (12) F-117 (#$FighterPlane-F117) stealth (#$DodgeStealthCar) fighter jets (#$JetOfFluid #$JetPropelledAircraft) to Kuwait (#$CityOfKuwaitKuwait (#$ProperSubcollectionNamedFn-Ternary kuwait #$Individual 34057665-f4ed-11d9-9bea-0002b3a85b0b) #$Kuwait), three (3) B-1 bombers (#$B-1-Bomber) to Bahrain (#$Bahrain-TheIsland #$Bahrain (#$CityNamedFn Bahrain #$Bahrain)) and 14 (14) B-52 (#$B-52-Bomber) bombers (#$SubmarineSandwich #$BomberPlane #$Bomber) to the island (#$Island) of Diego Garcia. It also has (#$possesses) diverted (#$AmusingSomeone #$DivertingSomething) dozens (#$Dozens-Quant 12) of support (#$SupportingSomething #$ShowingSupportForSomeone (#$SubcollectionOfWithRelationFromTypeFn #$PartiallyTangible #$supportingObject #$SupportingSomething)) aircraft (#$AirTransportationDevice) to the region (#$TheRegion) for refueling (#$Refueling (#$MakingAvailableFn #$CombustibleFuelSubstance)),

# Cyc综述

一个人工撰写的常识知识库

| | Cyc |
|---|---|
| Content | Common sense knowledge, axioms |
| Main strength | Huge ontology, with tools |
| Technique | Manual |
| License | proprietary, OpenCyc is Apache License V2.0 |
| Entities | 500k |
| Assertions | 5m |
| Relations | 15k |
| Tools | Reasoner, NL tool |
| URL | http://cyc.com |
| References | [Lenat, Comm. ACM 1995] |

# 代表性知识图谱

- **人工构建知识图谱**
  - WordNet
  - CYC
- **基于Wikipedia的知识图谱**
  - Yago
  - DBPedia
  - Freebase
- **文本抽取知识图谱**
  - NELL

# Wikipedia

- 免费的在线百科全书
  - 2001年开始
  - crowdsource的方式构建
  - 目标：构建全世界最大的百科全书


主要特点

高质量数据源

500万概念

多语言

富含丰富语义结构的文档：Infobox，table，list，category...

# Wikipedia: 文档结构

标题=概念

唐太宗

唐太宗李世民（598年1月28日－649年7月10日[1][2][3]），中国唐朝第二任皇帝。祖籍陇西郡成纪县（今甘肃省天水市秦安县北），生于陕西武功县，626年至649年在位。父亲是唐高祖李渊，母亲是窦皇后

概念文本描述

Infobox:以(属性，值)对形式呈现的信息表格

每个页面有多个类别，类别组成Taxonomy

- 分类：598年出生，649年逝世　唐朝皇帝

唐太宗

概要

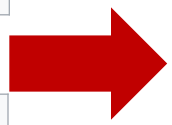
姓名 李世民
庙号 太宗
谥号 文皇帝（649年初谥）
文武圣皇帝（674年加谥）
文武大圣皇帝（749年加谥）
文武大圣大广孝皇帝（754年加谥）
陵墓 昭陵
政权 唐朝
在世 598年1月28日－649年7月10日（52岁）
在位 626年9月4日－649年7月10日
年号 贞观：627年－649年

# 出发点

- 基于Wikipedia的知识库都基于几乎相同的思路：
  - **从Wikipedia丰富的半结构化信息中挖掘知识**
  - 包括：Infobox，Category，超链接, Table，List...
- 不同之处在于
  - 如何处理有歧义的属性映射
  - 如何构建知识库的Taxonomy

| | |
|---|---|
| 在世 | 598年1月28日 - 649年7月10日（52岁） |

| | |
|---|---|
| 出生 | 隋文帝仁寿四年604年 |



BirthDate（李世民，598年）

BirthDate（李元吉，604年）

# 这些知识库具有相同的数据模型

- **一个知识库包含一个集合的实体**
  - 猫王、李世民、赵高、唐朝、分封制...
- **实体被划分到不同的类别中**
  - 歌手(猫王)，皇帝(李世民), 朝代(唐朝), 制度(分封制)
- **类别通过上下文关系等关系相互关联**
  - SubClassOf(歌手，人)， SubClassOf(皇帝，人)
- **实体和类别都通过属性和相互之间的关系来描述**
  - BirthDate（李世民，598年），Has(歌手，歌曲)
- **关系可以通过蕴含关系来进行推理**
  - 歌曲→作品， 收购 → 持有

# DBPedia

- 2007年开始，其主要目标是构建一个社区，通过社区成员来**定义和撰写准确的抽取模板，从维基百科中抽取结构信息，并将其发布到Web上**

- 社区通过人工的方式构建了Taxonomy
  - 280个类别
  - 覆盖约50%的维基百科实体

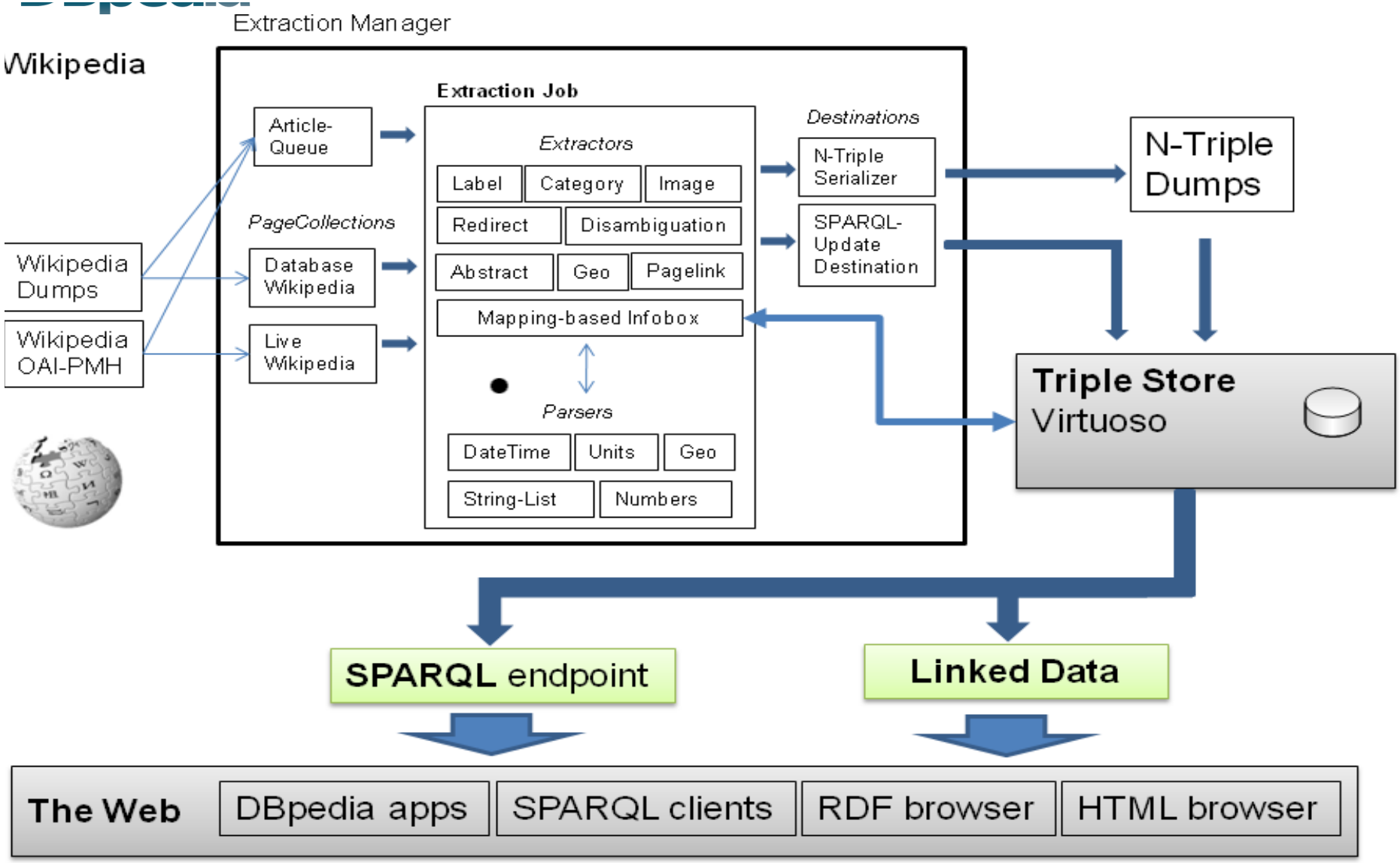
DBpedia

# 抽取方法

- **DIEF - DBpedia Information Extraction Framework**
  - 目标：抽取WIkipedia中的结构化信息
  - 方法：基于属性mapping的Infobox抽取，Raw Infobox Extraction, Feature Extraction, Statistical Extraction
  - 编程语言：Scala & Java

- DBPediaLive： 持续保持与Wikipedia的同步
  - 2013年六月，英语维基百科有将近330万次编辑(每分钟越77次)

# 抽取框架图


Extraction Manager
Wikipedia
Extraction Job
Article-Queue
PageCollections
Database Wikipedia
Live Wikipedia
Wikipedia Dumps
Wikipedia OAI-PMH
Extractors
Label
Category
Image
Redirect
Disambiguation
Abstract
Geo
Pagelink
Mapping-based Infobox
Parsers
DateTime
Units
Geo
String-List
Numbers
Destinations
N-Triple Serializer
SPARQL-Update Destination
N-Triple Dumps
Triple Store Virtuoso
SPARQL endpoint
Linked Data
The Web
DBpedia apps
SPARQL clients
RDF browser
HTML browser

# 抽取知识示例

## About: Harry Froboess

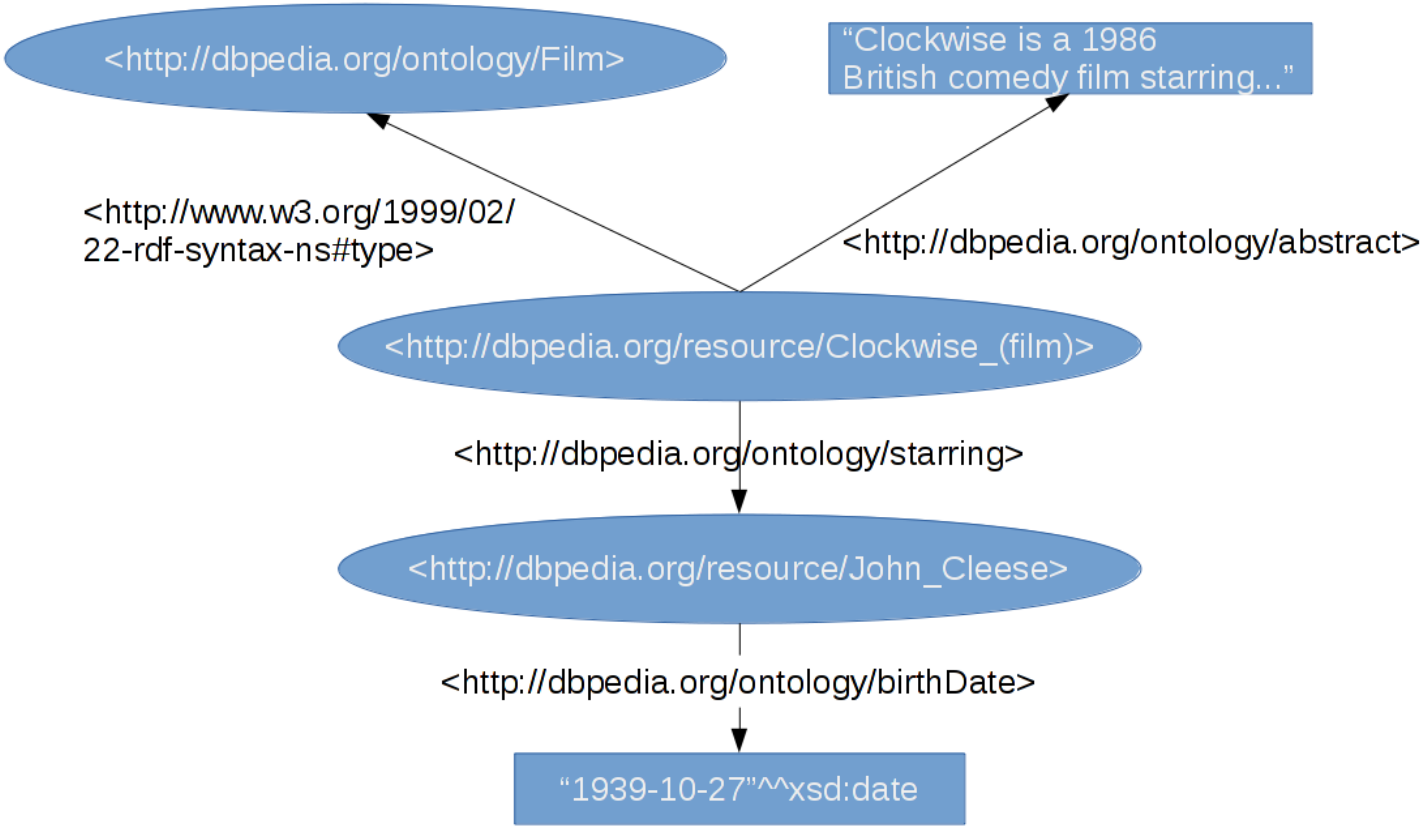
An Entity of Type : agent, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Friedrich Harald August Froboess (1899–1985) born on 10.23.1899 in Dresden, Germany, was a German stunt diver, and high diver.

| Property | Value |
|---|---|
| dbpedia-owl:abstract | ▪ Friedrich Harald August Froboess (1899–1985) born on 10.23.1899 in Dresden, Germany, was a German stunt diver |
| dbpedia-owl:alias | ▪ Froboess, Harry Arias |
| dbpedia-owl:birthDate | ▪ 1899-01-01 (xsd:date) |
| dbpedia-owl:birthName | ▪ Harry Arias Froboess |
| dbpedia-owl:birthPlace | ▪ dbpedia:Bern<br>▪ dbpedia:Dresden<br>▪ dbpedia:Germany<br>▪ dbpedia:Switzerland<br>▪ dbpedia:United_States |
| dbpedia-owl:birthYear | ▪ 0023-01-01 (xsd:date)<br>▪ 1899-01-01 (xsd:date) |
| dbpedia-owl:deathDate | ▪ 1899-01-01 (xsd:date) |
| dbpedia-owl:deathPlace | ▪ dbpedia:Switzerland |
| dbpedia-owl:deathYear | ▪ 0012-01-01 (xsd:date)<br>▪ 1899-01-01 (xsd:date) |
| dbpedia-owl:occupation | ▪ dbpedia:Diving<br>▪ dbpedia:Stunt_performer<br>▪ dbpedia:Harry_Froboess__1 |
| dbpedia-owl:spouse | ▪ dbpedia:Berlin<br>▪ dbpedia:Switzerland |

# 使用RDF来表示抽取出来的知识

- 支持复杂的结构化query，SPASQL语言查询
- 支持与Web上其他数据集的链接和集成



# DBPedia综述

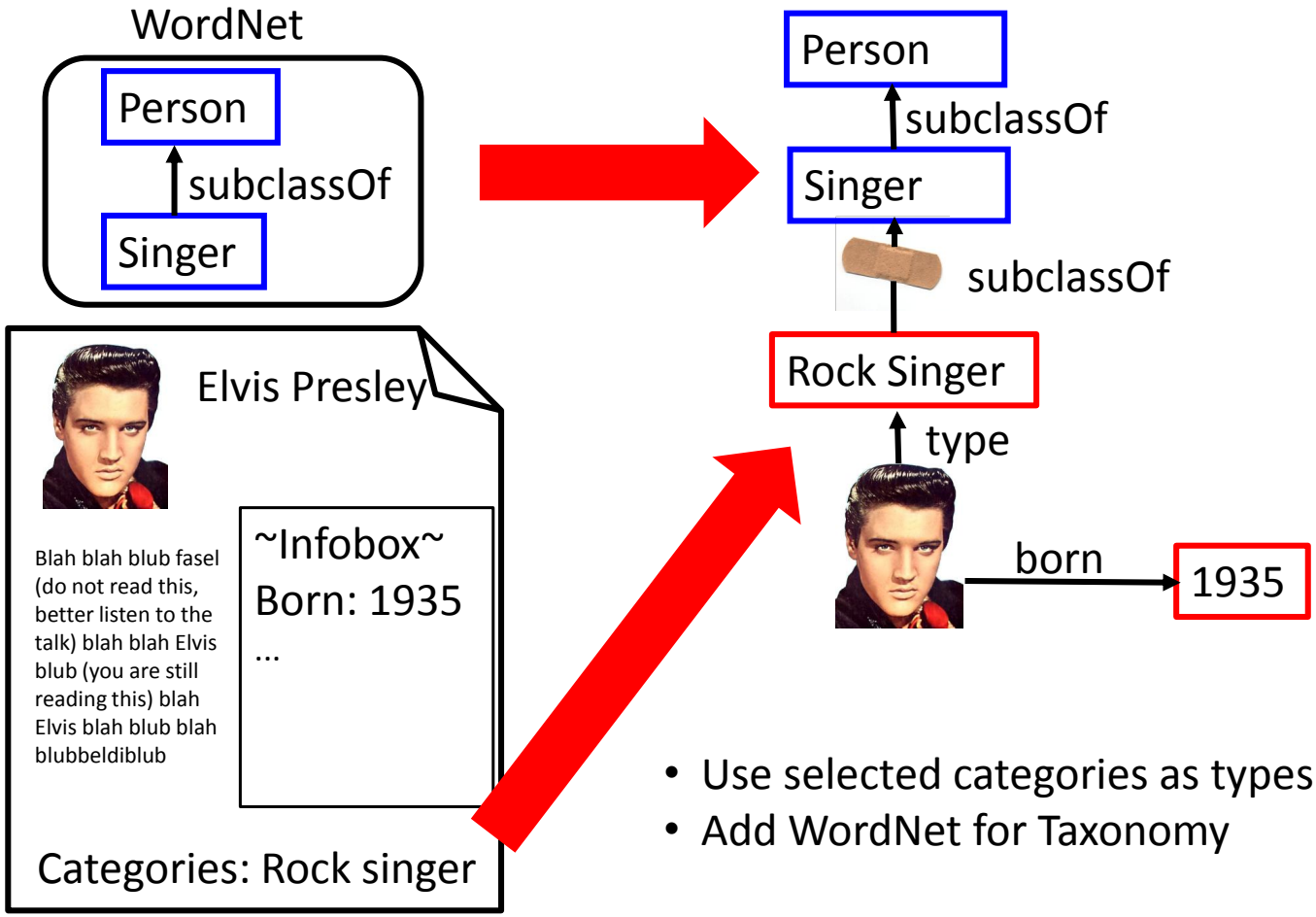
| | |
|---|---|
| Content | Entities of public interest |
| Format | RDF, API, SPARQL |
| Sources | Wikipedia, YAGO/WordNet |
| Main strengths | Focus on coverage,<br>interlinking with other data sets |
| Technique | Extraction from Wikipedia + manual supervision by the community |
| Size | Entities: 3.5m (in manual taxonomy: 1.7m)<br>Facts: 670m<br>Attributes: 9k (manually defined: 1k)<br>Manual Classes: 280 |
| License | CC-BY-SA & GNU FDL |
| URL | http://dbpedia.org |
| Reference | [Auer, ISWC 2007], [Bizer09, JWS 2009] |

# Yago(Yet Another Great Ontology )

- 德国马普研究所从2007年开始的一个项目
- 融合WordNet和Wikipedia
  - 从Wikipedia的结构中抽取信息：Infoboxes,类别...
  - 包括时间和地点标注
  - 人工采样评估
  - >1亿事实和100种关系

# Yago Taxonomy构建

- 使用WordNet的Taxonomy作为基础
- 将Wikipedia中的类别加入到WordNet中



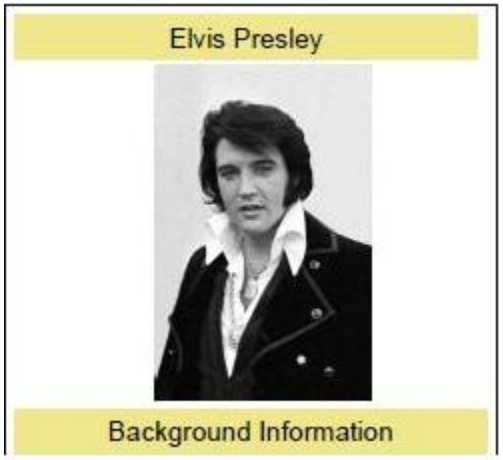
# Yago语义关系

- **人工定义了100多种语义关系**
  - wasBornOnDate，locatedIn， hasPopulation

- **抽取方法：主要采用手写的规则抽取**
  - Infobox Harvesting:信息框
  - Word-Level Techniques:重定向页
  - Category Harvesting:类别信息抽取
  - Type Extraction: 维基类别,WordNet类别

# 抽取示例

Infobox




Died: August 16, 1977

Attribute Map

| Attribute | Relation | Inverse | Manifold | Indirect |
|---|---|---|---|---|
| | … | | | |
| Died | diedOnDate | | | |
| | … | | | |

Relation Map

| Relation | Domain | Range |
|---|---|---|
| | … | |
| diedOnDate | person | yagoDate |
| | … | |

**Elvis Presley** **diedOnDate** **August 16, 1977**

# 抽取准确率

| Relation | Total Facts | Evaluated | Accuracy |
|---|---|---|---|
| actedIn | 12636 | 69 | 97.36% +- 2.64% |
| created | 225563 | 94 | 98.04%+-1.96% |
| graduatedFrom | 15583 | 57 | 96.84%+-3.16% |
| type | 8414398 | 208 | 97.68%+-1.83% |
| subClassOf | 367040 | 339 | 93.42%+-2.67% |
| diedIn | 28834 | 88 | 97.91%+-2.09% |

# Yago综述

| | |
|---|---|
| Content | Entities of public interest |
| Format | TSV, RDF, XML, N3, Web Interface |
| Sources | Wikipedia, WordNet, Geonames |
| Main strength | Focus on precision, geotemporal annotations, multilingual |
| Precision | 95% |
| Technique | Extraction from Wikipedia + matching with WordNet & Geonames + consistency checks |
| Size | Entities: 3 m (+ geonames -> 10m)<br>Facts: 120m (+geonames -> 460m)<br>Relations: 100, Classes: 200k, Languages: 200 |
| License | Creative Commons BY-SA |
| URL | http://mpii.de/yago |
| References | [Suchanek, WWW 2007] [Hoffart, WWW 2011] [deMelo, CIKM 2010] |

# Freebase

- Metaweb公司2000年开始构建，2010年被Google收购


- 从Wikipedia和其他数据源（如IMDB、MusicBrainz）中导入知识

- 核心想法：
  - 在Wikipedia中，人们编辑文章
  - **在Freebase中，人们编辑结构化知识**

# 用户是Freebase知识构建的核心

- **编辑实体**
  - 创建实体
  - 将实体分到类别
  - 增加/修改属性/关系
  - 上传图片
- **编辑Schema**
  - 定义新类别
  - 定义类别的属性
- **Review**
  - 验证知识准确性
  - 投票
  - 删除错误知识
- **DataGame**
  - 寻找别名
  - 抽取事件日期
  - 使用Yahoo图片搜索加入图片

**用户在Freebase中的作用**

# Freebase综述

| | |
|---|---|
| Content | Entities with public information |
| Format | API, RDF |
| Construction | by the community<br>data import from public sources |
| Sources | Wikipedia, Libraries, WordNet, MusicBrainz… |
| Main strength | free and large |
| Size | Facts: several millions |
| | Entities: 20 m |
| License | Creative Commons Attribution (CC-BY) |
| URL | http://download.freebase.com |

一个大规模协同构建知识库，目前归Google所有

# 代表性知识图谱

- **人工构建知识图谱**
  - WordNet
  - CYC
- **基于Wikipedia的知识图谱**
  - Yago
  - DBPedia
  - Freebase
- **文本抽取知识图谱**
  - NELL

# NELL

**Read the Web**
Research Project at Carnegie Mellon University

- 2009年开始的CMU项目
- **输入：**
  - 初始本体（~800类别和关系）
  - 每个谓词的一些实例（~10-20个种子实例）
  - web(~10亿页面, ClueWeb)
  - 间歇性的人工干预
- **任务：**
  - 24 x 7 持续运行(从2010年开始)
  - 每天
    - 抽取更多知识来补充给定本体
    - 学习如何更好的构建抽取模型
- **结果：**超过9千万实例（不同置信度）

# 抽取结果示例

## Recently-Learned Facts

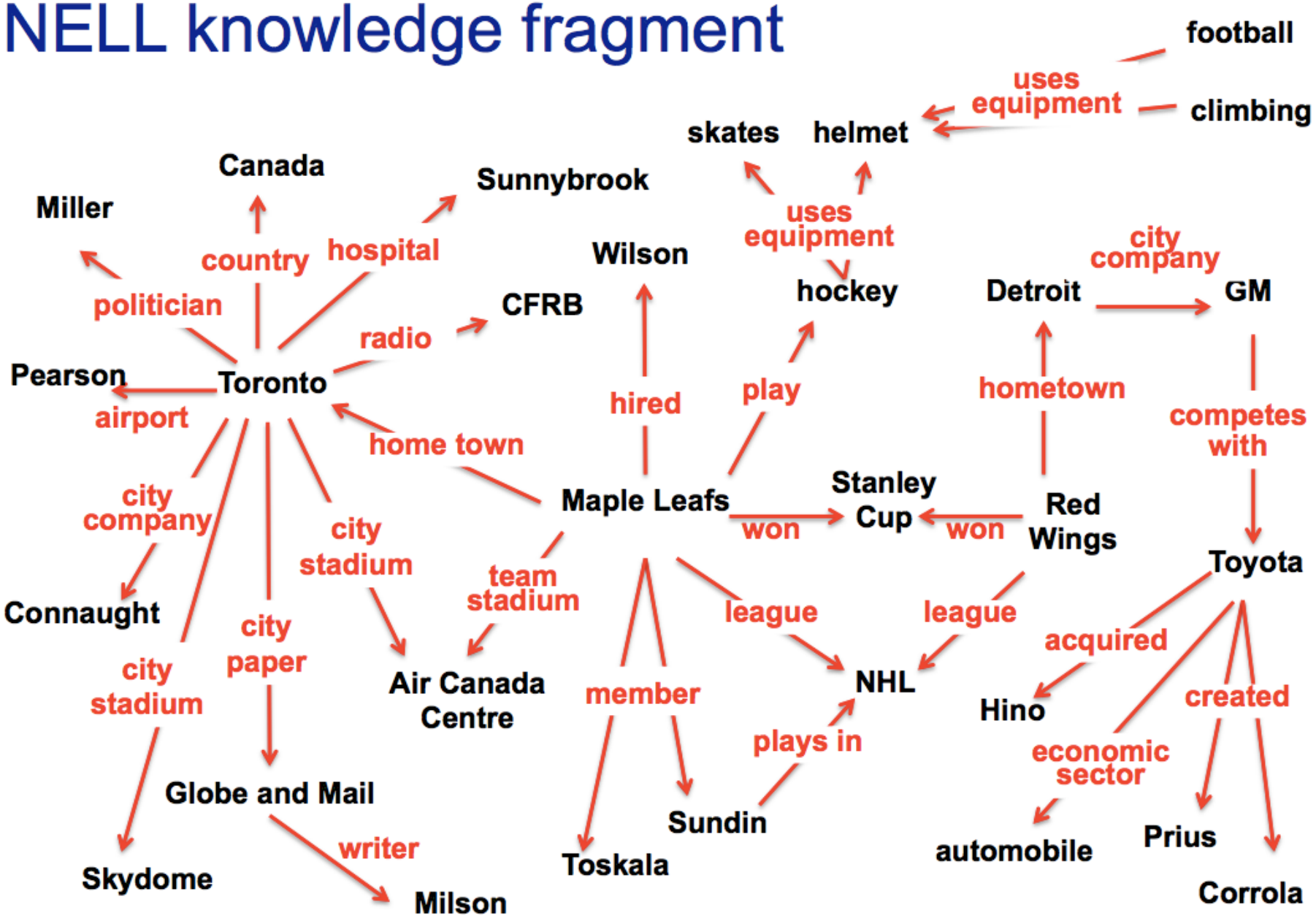
twitter

Refresh

| instance | iteration | date learned | confidence |
|---|---|---|---|
| rillito_river is a river | 1064 | 26-jun-2017 | 99.9 |
| james_dexter is a CEO | 1064 | 26-jun-2017 | 99.7 |
| richard_thompson is a Mexican person | 1064 | 26-jun-2017 | 100.0 |
| htc_droid_eris is a consumer electronic device | 1064 | 26-jun-2017 | 92.8 |
| radiant_snow is a weather phenomenon | 1066 | 15-jul-2017 | 99.8 |
| the_london_eye001 is a tourist attraction in the city london | 1069 | 03-aug-2017 | 100.0 |
| state_university is a sports team also known as michigan_state_university | 1067 | 21-jul-2017 | 100.0 |
| state_university is an organization also known as clemson | 1066 | 15-jul-2017 | 96.9 |
| irene_kirkaldy belongs to the religion seven_day_adventist | 1069 | 03-aug-2017 | 100.0 |
| kcal is a TV station in the city los_banos | 1069 | 03-aug-2017 | 100.0 |

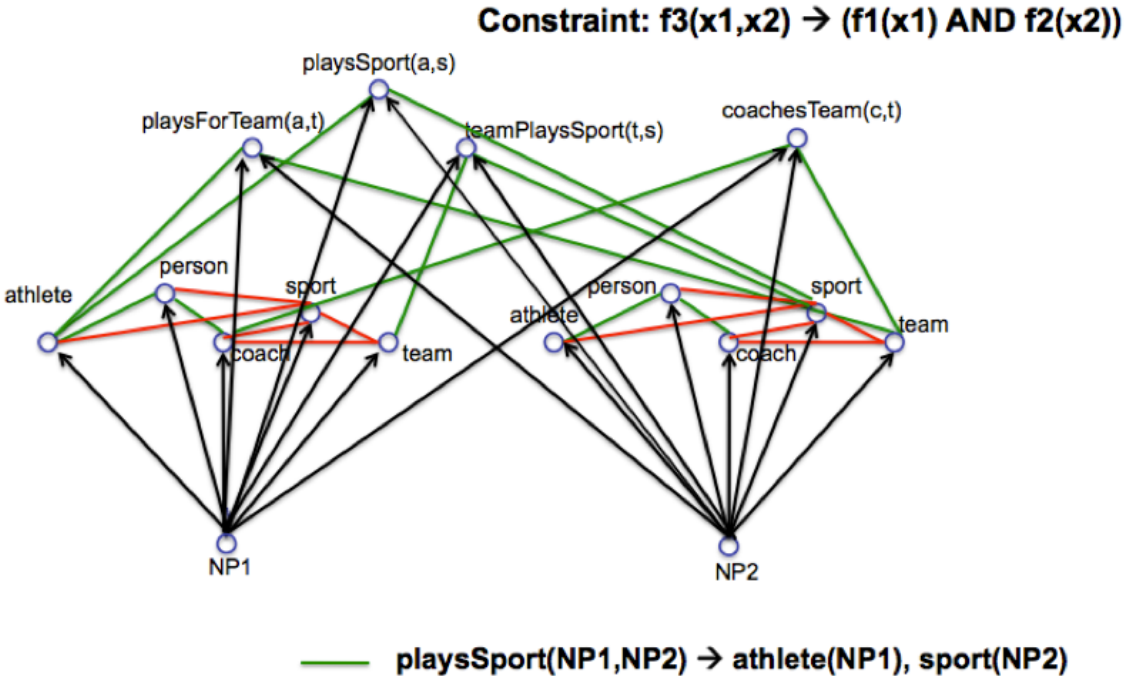
http://rtw.ml.cmu.edu

# NELL knowledge fragment



# 抽取步骤

- **1. 把名词短语划分到给定类别**
  - Entity Set Expansion: 基于Pattern的Bootstrapping
- **2. 分类名词短语之间的语义关系**
  - Coupling Learning



# 抽取步骤

- **3. 识别新的推理规则，用于发现新的关系实例**
  - PathRank

0.95 athletePlaysSport(?x,basketball) :- athleteInLeague(?x,NBA)

0.93 athletePlaysSport(?x,?y) :- athletePlaysForTeam(?x,?z)
teamPlaysSport(?z,?y)

0.91 teamPlaysInLeague(?x,NHL) :- teamWonTrophy(?x,Stanley_Cup)

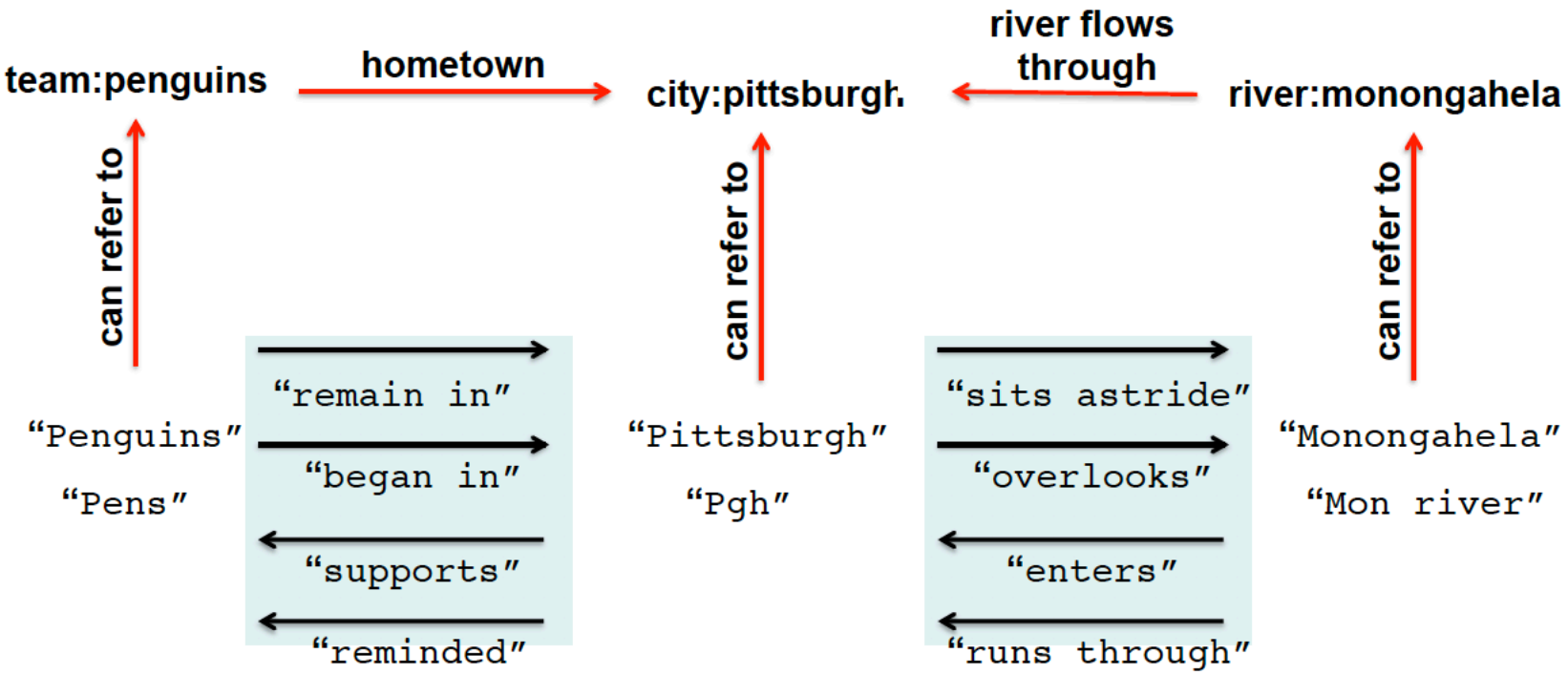
0.90 athleteInLeague(?x,?y):- athletePlaysForTeam(?x,?z),
teamPlaysInLeague(?z,?y)

0.88 cityInState(?x,?y) :- cityCapitalOfState(?x,?y),
cityInCountry(?y,USA)

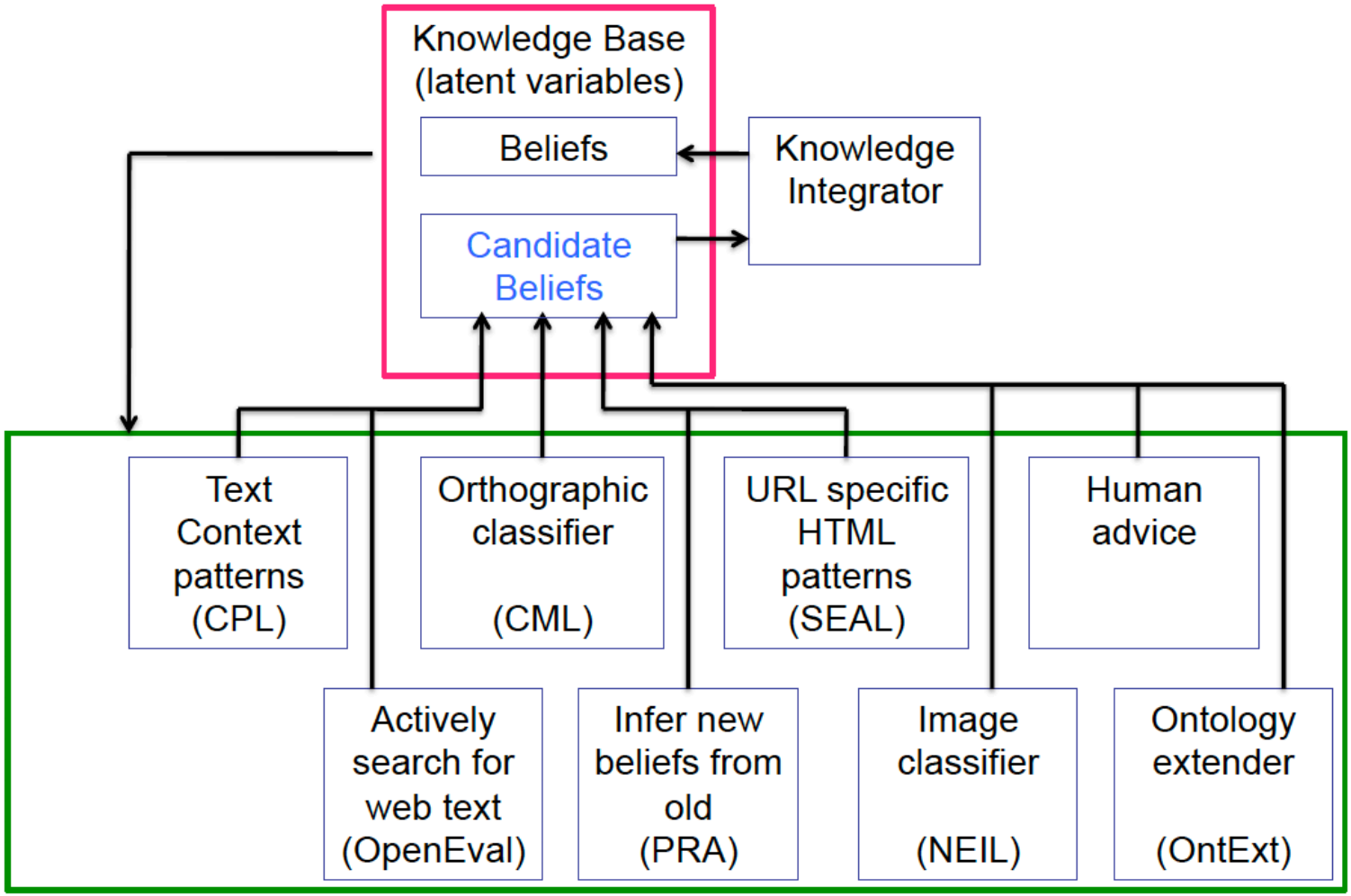
0.62* newspaperInCity(?x,New_York) :- companyEconomicSector(?x,media),
generalizations(?x,blog)

# 抽取步骤

- **4. 名词短语被映射到概念，动词短语被映射到关系**
  - Entity Linking, Entity Resolution



# 完整的抽取框架


Knowledge Base
(latent variables)
Beliefs
Candidate
Beliefs
Knowledge
Integrator
Text
Context
patterns
(CPL)
Orthographic
classifier
(CML)
URL specific
HTML
patterns
(SEAL)
Human
advice
Actively
search for
web text
(OpenEval)
Infer new
beliefs from
old
(PRA)
Image
classifier
(NEIL)
Ontology
extender
(OntExt)

# NELL综述

| | |
|---|---|
| Content | Entities mentioned on Web pages |
| Format | TSV |
| Construction | by a perpetual extractor |
| Sources | The Web |
| Main strength | Not limited to a specific source |
| Size | Facts: 800k |
| | Categories & relations: 633 |
| Reference | [Carlson, AAAI 2010] |
| URL | http://rtw.ml.cmu.edu/ |

一个基于文本信息抽取技术持续不断更新的知识库

# 代表性知识图谱总结

# 知识图谱

- **实体及其之间关系的语义描述**
  - 使用形式化知识表示（如RDF，RDFS，OWL)
- **Entities**
  - 真实世界对象 (things, places, people) 和抽象概念 (genres, religions, professions)
- **Relationships**
  - 将实体按语义关系链接成一张大网
- **Semantic descriptions**
  - 类别和属性
- 有时包含支持推理的公理知识 (如规则)

# 代表性知识图谱综述

- **人工构建知识图谱**
  - **WordNet**： 英文电子词典
  - **CYC**：常识知识库
- **基于Wikipedia的知识图谱**
  - **Yago**： Wikipedia+WordNet
  - **DBPedia**：基于社区抽取Wikipedia结构化信息
  - **Freebase**：知识编辑社区协同构建
- **文本抽取知识图谱**
  - **NELL**：从文本中持续自动抽取海量知识
- **还有许多其他知识库没有覆盖**：Wikidata、BabelNet、OpenMind、Probase...

# 当前KG的限制和不足

- **领域限制**
  - 一些知识库侧重于语言: WordNet,BabelNet
  - 一些知识库侧重于Schema： Cyc, UMBEL
  - 一些知识库侧重于Fact: DBPedia, Yago
- **对时空属性的建模**
  - 对动态性的实体，如Event建模不足
  - Yago 3在一定程度上考虑时间和地理属性
- **自动构建**
  - 自动构建是维护和保持KG质量和覆盖的核心技术
- **与LOD的集成**
  - 缺乏Schema之间的alignment
  - 往往只用到底层的表达能力，OWL的高级功能很少涉及

# KG展望

- **新的知识表示模型**
  - Ontology engineering已经被用了超过15年
- **新类型的知识图谱**
  - 不再围绕实体和关系的存储
  - 如Event-centric KG
- **知识图谱自动构建技术**
  - 在Freebase中，71%的人没有出生日期
  - 新技术：Distant Supervision, KG embedding, 知识集成（如Google的Knowledge Vault）

# Coffee Break

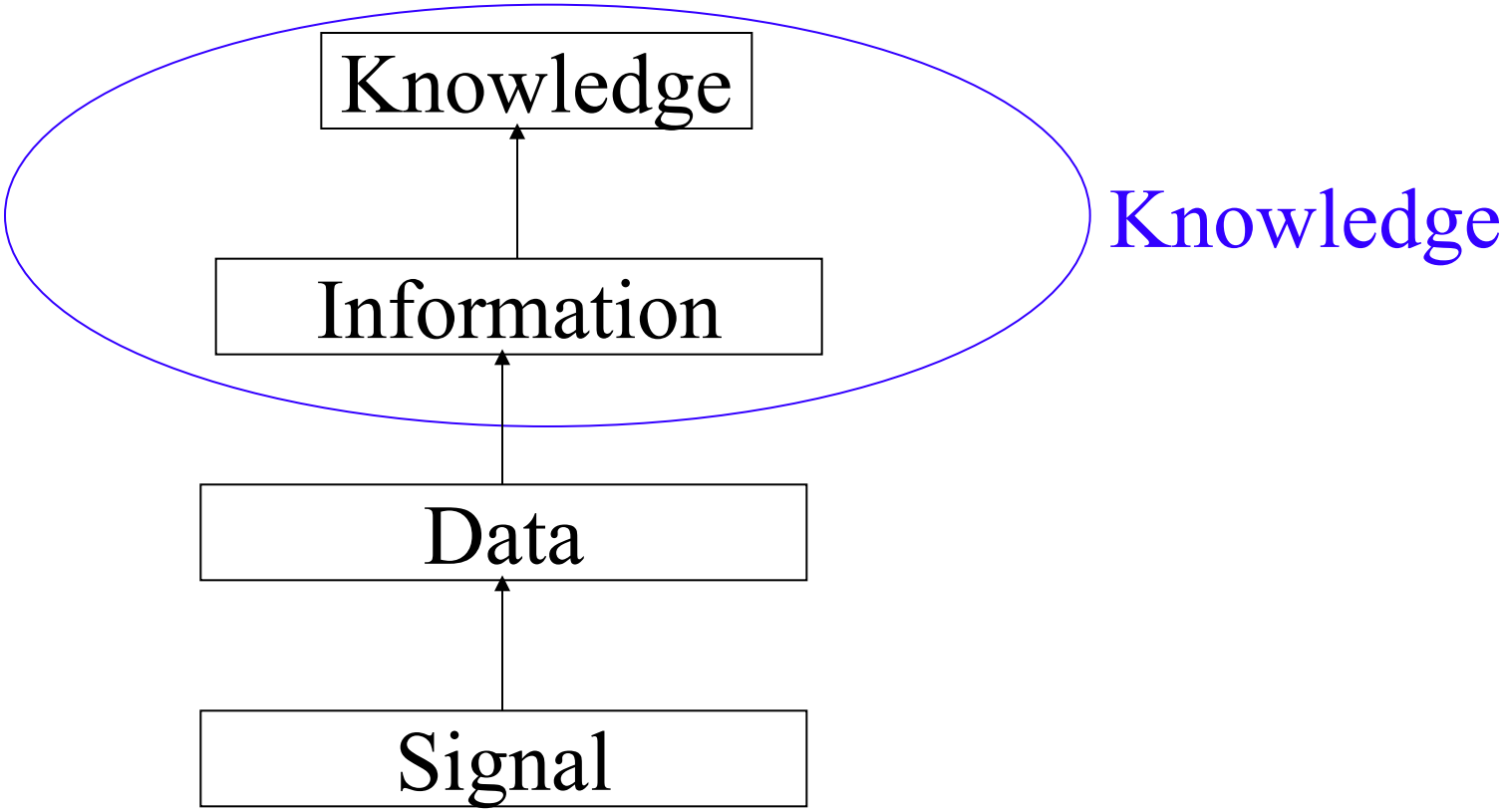
# 目录

# 基于符号的知识表示与推理

- 知识及知识表示

- 符号表示知识方法及实现
  - Logic
  - Semantic Net
  - Frame
  - Script
  - 语义网知识表示语言体系

# 什么是知识？




Knowledge = Facts＋Rules＋Control Strategy +(有时) Faiths

# 知识的类别体系

- Facts: 陈述性知识(declarative knowledge)
  - John 是一个 贼，约翰 喜欢 酒
  - 1+1 = 2, 2是偶数

知识图谱主要包含事实知识

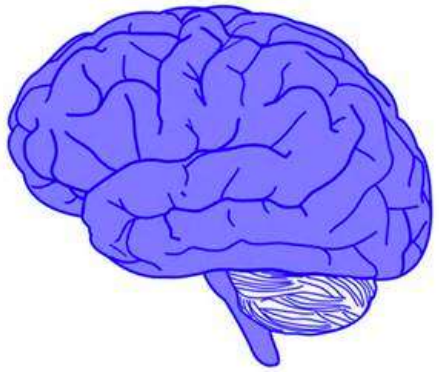
- Rules: 程序性知识(procedural knowledge)
  - 解决一个问题的系列步骤： 菜谱中的炒菜步骤
  - 规则： 一个贼 喜欢 一个东西 → 这个贼很有可能去 偷取 它

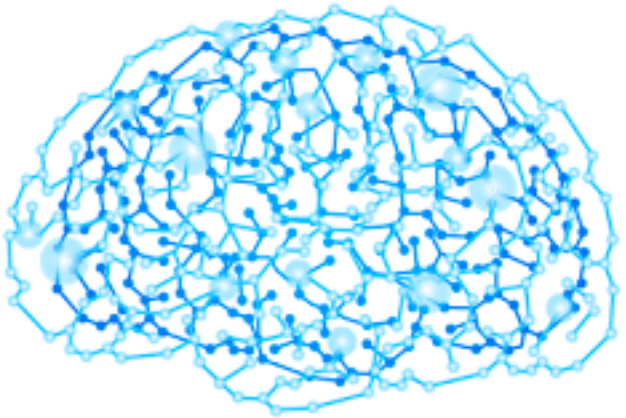
- Control Strategy: 元知识(meta, super knowledge)
  - 推理策略(reasoning strategy)
  - 搜索策略(search strategy)
  - ...

# 知识表示

# 知识表示(Knowledge Representation)

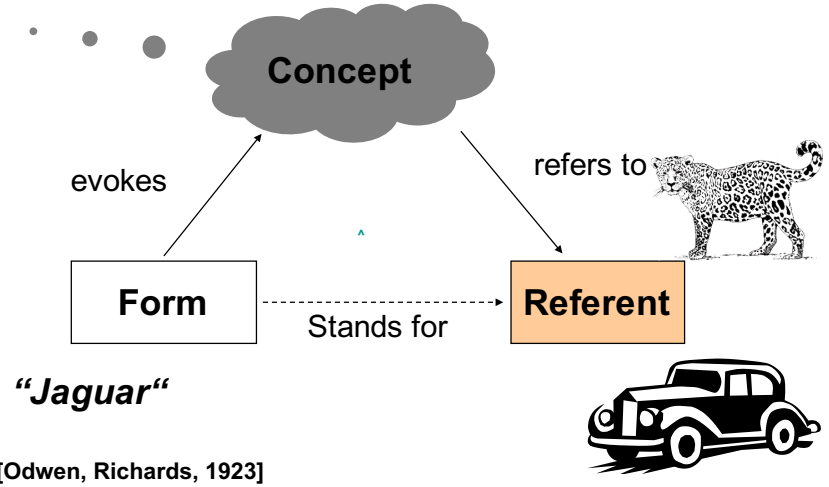
- **知识表示**同时是认知科学和AI中的科学问题
  - 在**认知科学**中，KR主要**关注人如何存储和处理信息**
  - 在**人工智能**中，KR的主要关注**如何表示关于世界的信息，并通过常识和事实来得出结论**，这样计算机程序就可以通过这些知识来模拟人类智能
  - AI科学家大量的借用了认知科学的表示理论(原型、框架、神经网络).





# 知识表示

- 一个知识表示是**事物本身的一个替代**，使得我们可以通过思考而不是行动来确定事物的后果
  - *吃饭会饱（不需要每次都吃来确认），冰岛是一个国家(虽然没去过)*
- 一个知识表示是一个**本体约定(ontological commitment)集合**，它回答如下问题：**我们该用什么术语（terms）来思考这个世界？**




[Odwen, Richards, 1923]

**What Is a Knowledge Representation? Davis et al., 1993**

# 知识表示的一些核心问题

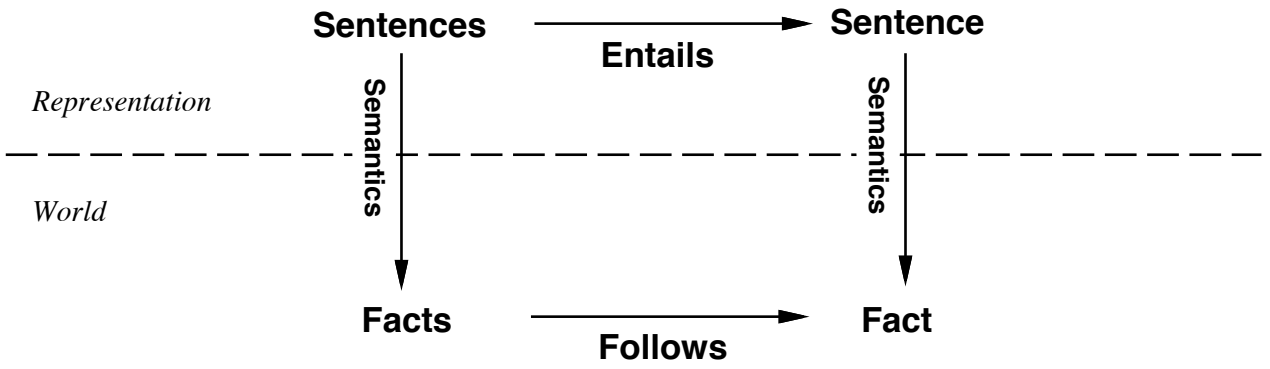
- 人如何表示知识？
- 知识的本质是什么？我们如何表示它?
- 我们如何使用符号结构(symbol structure)来表示知识？
- 一个表示schema应该处理特定领域还是要处理通用领域？
- 一个表示schema的表示能力如何？
- Schema应该是陈述式的还是过程式的？
- …

# 知识表示的主要方法

- 自然语言(对人而言最方便的表示和传播手段)
- **符号表示方法**
  - **Logic**
  - **Semantic Net**
  - **Frame**
  - **Script**
  - **语义网知识表示体系**
- 许多其他的方法，如分布式表示方法（第二部分）

# 知识表示语言的组件

- **Syntax:** 知识表示语言中使用到的原子符号(atomic symbols)，原子符号如何组合成合法的语句
- **Semantics**: 知识表示中的一个句子在世界上对应的事实，用于决定一个句子的真值
- **Inference**: 如何从已有知识中得到新知识的机制



# PREDICATE LOGIC

# 谓词逻辑

- 关于对象的逻辑,用于表示关于实体对象的知识
  - **objects**(terms)
  - **properties**(unary predicates on terms)
  - **relations**(n-ary predicates on terms)
  - **functions**(mapping from terms to other terms)

- 相比命题逻辑，谓词逻辑提供了一套更为灵活且紧凑的知识表示方式，它提供了表示和推理对象属性及不同对象间关系的机制

# 一阶谓词逻辑的Syntax

Sentence → AtomicSentence
| Sentence Connective Sentence
| Quantifier Variable Sentence
| ¬Sentence
| (Sentence)

AtomicSentence → Predicate(Term, Term, ...)
| Term=Term

Term → Function(Term,Term,...)
| Constant
| Variable

Connective → ∨ | ∧ | ⇒ | ⇔

Quanitfier → ∃ | ∀

Constant → A | John | Car1

Variable → x | y | z |...

Predicate → Brother | Owns | ...

Function → father-of | plus | ...

# 对象的表示

- 使用terms来表示对象：
  - **Constants**: Block1, 小明, John
  - **Function symbols**: father-of, successor, plus
  - **n-ary function**： 将一个包含n个term的tuple映射为一个新term: father-of(John), succesor(0), plus(plus(1,1), 2)

- Terms是对象的名字，Logical function让我们可以用有限的term来紧凑的表示无限的对象

# 命题(Proposition)

- **Predicate**: 谓词是一个动词词组，用于描述对象的属性，或是不同对象之间的关系

- 命题是谓词加应用于该谓词的一个term元组，表示一个属性或objects之间的关系
  - Brother(John, Fred), Left-of(Square1, Square2)
  - GreaterThan(plus(1,1), plus(0,1))
  - 其语义是在特定interpretation中的真假值

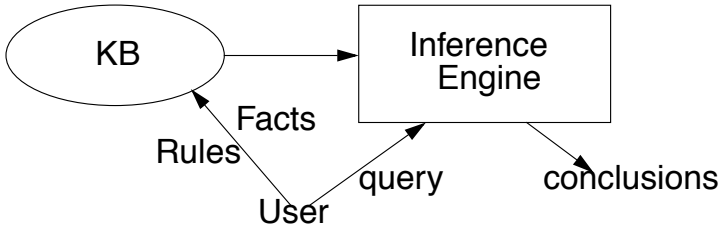
- 复杂命题可以通过逻辑连词($\vee$ $\neg$ $\wedge$ $\Rightarrow$ $\Leftrightarrow$ )来构建
  - Owns(John,Car1) $\vee$ Owns(Fred, Car1)
  - Sold(John,Car1,Fred) $\Rightarrow$ $\neg$Owns(John, Car1)

# 量词

- 通过量词机制，允许声明关于一个集合的对象的知识，而不需要一个个枚举它们
- Universal quantifier: $\forall x$
  - $\forall x$ Loves(x, FOPC)
  - $\forall x$ Whale(x) $\Rightarrow$ Mammal(x)
- Existential quantifier: $\exists$
  - $\exists x$ Loves(x, FOPC)
  - $\exists x$(Cat(x) $\wedge$ Color(x,Black) $\wedge$ Owns(Mary,x))

# Logical KB

- **一个逻辑知识库包含**：
  - 用于描述谓词之间关系的公理
    - $\forall x,y$ Bachelor(x) $\Leftrightarrow$ Male(x) $\wedge$ Adult(x) $\wedge$ $\neg \exists y$Married(x,y)
  - 谓词的定义
    - $\forall x$ Adult(x) $\Leftrightarrow$ Person(x) $\wedge$ Age(x) >=18
  - 事实集合
    - Male(Bob), Age(Bob)=21, Married(Bob, Mary)
- **用户可以查询特定的知识库**
  - Adult(Bob) ?
  - Bachelor(Bob) ?



# 推理机制

- **相等变换**：如P ∧ True |– True
- **蕴含推理**：如P |– P ∧ Q
- **假言推理（三段论）**： 如果P |– Q 且P 则 Q
- **Universal Elimination:** ∀x Loves(x, FOPC) |– Loves(Ray, FOPC)
- **Existential Elimination:** ∃x (Owns(Mary,x) ∧ Cat(x)) |– Owns(Mary,MarysCat) ∧ Cat(MarysCat)
- **Existential Introduction:**Loves(Ray, FOPC) |– ∃x Loves(x, FOPC)
- **…**

# 逻辑作为知识表示

- **优点**
  - 有语义
  - 表达能力强

- **缺点**
  - 不高效
  - 不可判定性
  - 无法表达过程知识
  - 无法做缺省推理

# SEMANTIC NET

# Semantic Net

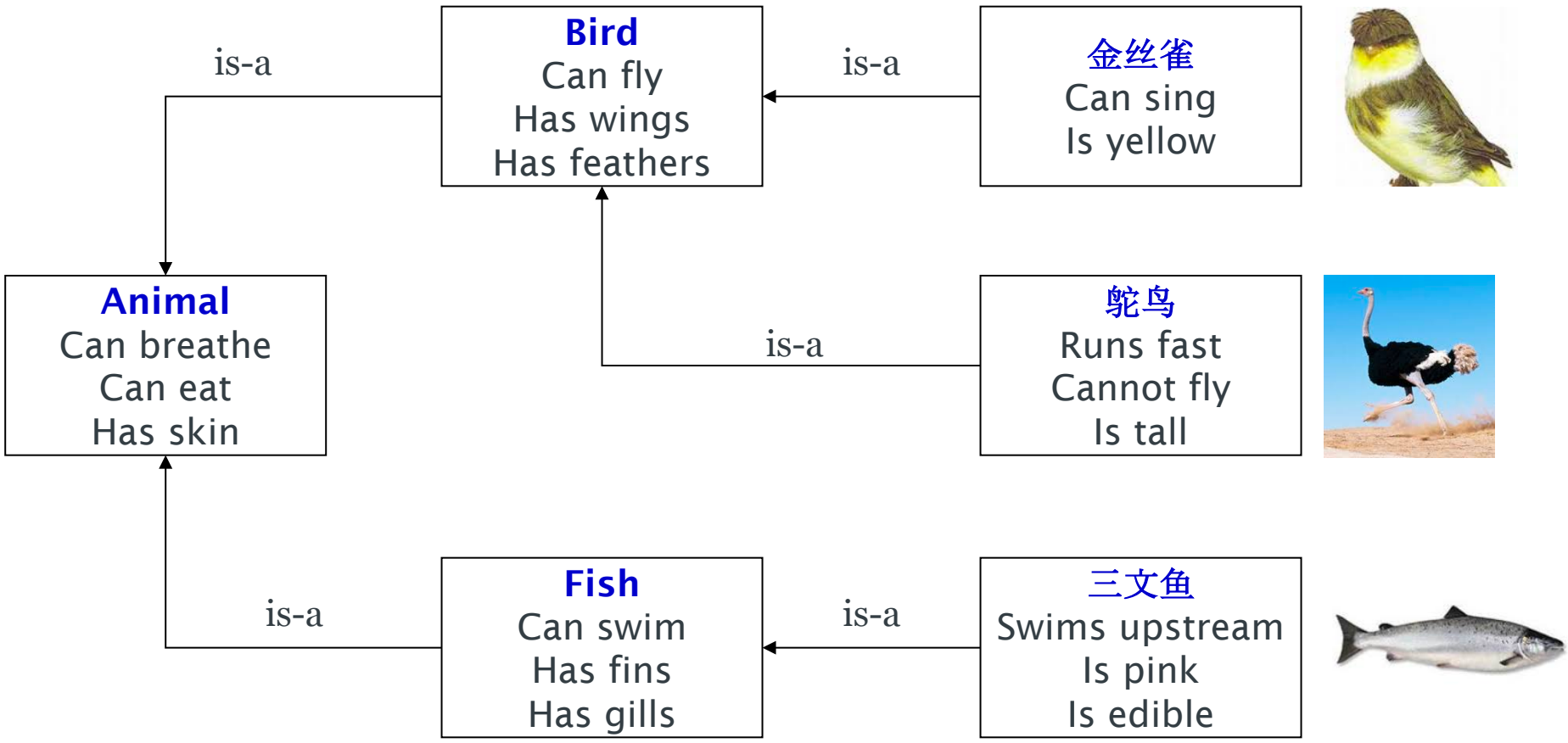
- 启发IDEA:
  - **人脑记忆的一个重要特征是人脑中不同信息片段之间高度连接**
  - **高度相关的概念能够比不太相关的概念更快的回忆起来**
- Semantic Net是**一个通过语义关系连接的概念网络**
- Semantic Net将知识表示为相互连接的点和边模式
  - *Nodes表示实体,属性,事件,值等*
  - *Links表示对象之间的语义关系*

M.Quillian, (1968). Semantic Memory, in M. Minsky (ed.), Semantic Information Processing, pp 227-270, MIT Press
J. F. Sowa (1987). "Semantic Networks". in Stuart C Shapiro. Encyclopedia of Artificial Intelligence. Retrieved 2008-04-29.

# 通常使用的语义关系

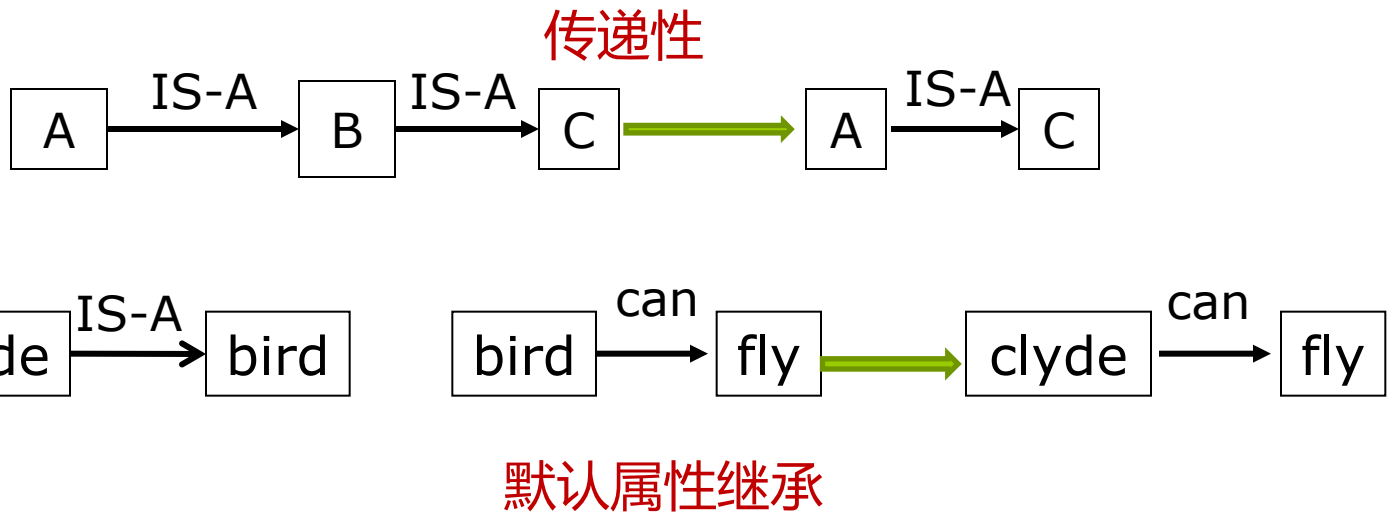
- **IS-A**
- **PART-OF**
- **MODIFILES:** on, down, up, bottom, moveto,…
- **领域特定的Link类型**
  - 医疗：症状、治疗、病因…
  - 金融：收购、持有、母公司…

# 示例


Animal
Can breathe
Can eat
Has skin
Bird
Can fly
Has wings
Has feathers
Fish
Can swim
Has fins
Has gills
金丝雀
Can sing
Is yellow
鸵鸟
Runs fast
Cannot fly
Is tall
三文鱼
Swims upstream
Is pink
Is edible
is-a
is-a
is-a
is-a
is-a

# Semantic Net中的推理(1)

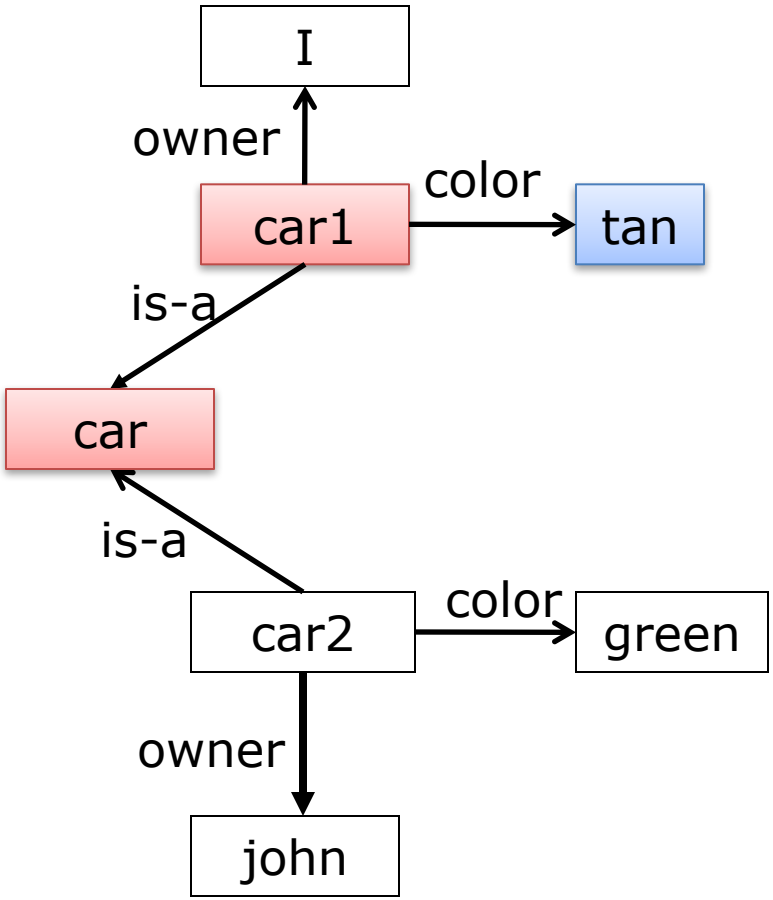
- **Inheritance（继承）**
  - the *is-a* and *instance-of* representation provide a mechanism to implement this
  - Inheritance also provides a means of dealing with *default reasoning*

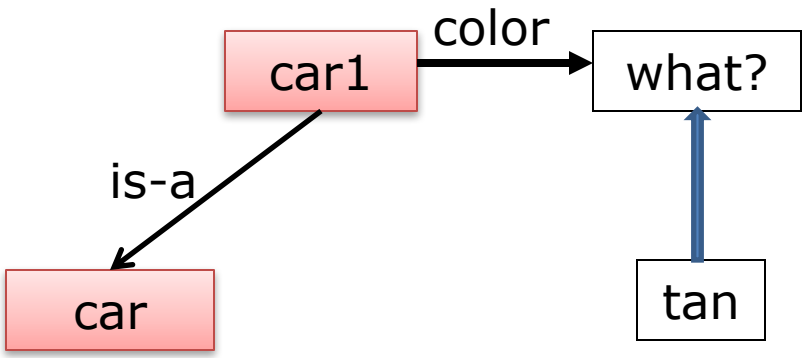


# Semantic Net中的推理(2)

- **Intersection search(交集搜索)**
  - The notion that *spreading activation* out of two nodes and finding their intersection finds relationships among objects.
  - 两个SemanticNet之间的匹配交集
  - Many advantages including entity-based organization and fast parallel implementation.
  - However very structured questions need highly structured networks

# Intersection Search示例




What color is the car1?

# FRAME/框架

# 框架理论

- 框架理论认为人们对现实世界中各种事物的认识都是以一种类似于框架的结构存储在记忆当中的，当面临一个新事物时，就从记忆中找出一个适合的框架，并根据实际情况对其细节加以修改补充，从而形成对当前事物的认识

Proposed in 1968 by Marvin Minsky http://web.media.mit.edu/~minsky

# Frame表示

- 知识通过Frame的形式来表示, 每一个Frame表示一种典型原型化场景(a stereotypical situation)
  - Frame-Based KR类似于面向对象编程，区别在于编码的对象不同
  - 一个Frame类似于数据库中的数据记录结构或数据库记录

| 商业交易 | |
|---|---|
| 买家 | ? |
| 卖家 | ? |
| 物品 | ? |
| 价格 | ? |
| 交易时间 | ? |

# Frame表示

- **Frame包含slot names和slot fillers**
  - 一个Frame的slot集合能够表示与该框架相关的对象
  - Slot可以指向其他的Frame，Procedure，Slot
- **两类Frame**
  - Class Frame：类似于面向对象编程里面的Class
  - Individual or Instance Frame：类似于面向对象编程里面的Object
  - Slots 类似于OO里面的variables/methods
- **不同的Frame通常被组织成一个层次体系结构**
  - Instance Frame - *instance_of->* Class Frame
  - Class Frame - *subclass_of->* Class Frame

# Frame表示例子

**DOG**

Fixed
- legs: 4

Default
- diet: 肉食
- sound: bark

Variable
- size:
- colour:

**COLLIE（牧羊犬）**

Fixed
- breed of: DOG
- type: sheepdog

Default
- size: 65cm

Variable
- colour:

子类可以从父类继承属性和默认属性值

# Frame表示上的推理

- 能够推理类别之间和类别与实例之间的ISA关系
  - $\forall x\ \text{elephant}(x) \Rightarrow \text{mammal}(x)$
- 能够使用slots和slot values来推理属性知识
  - $\forall x\ \text{mammal}(x) \Rightarrow \text{has\_part}(x, \text{head})$
- 对象可以继承所有父类的属性
  - $\text{elephant}(\text{clyde}) \Rightarrow \text{has\_part}(\text{clyde}, \text{head}),$ $\text{color}(\text{clyde}, \text{gray})$
- 可以继承原型属性值，同时也可以覆盖原型属性值
  - mammal通常有furry,但是elephant没有

# Frame表示上的推理示例

- $\forall$ x mammal(x) $\Rightarrow$ has_part(x, head)
- $\forall$ x elephant(x) $\Rightarrow$ mammal(x)

- elephant(clyde)
  $\therefore$
  mammal(clyde)
  has_part(clyde, head)

# 基于继承的紧凑表示

- 继承所有必须属性
- 只有在默认属性不对时显式覆盖父属性值(*标识默认属性)

**MAMMAL:**
subclass: ANIMAL
has_part: head
*furry: yes

**ELEPHANT**
subclass: MAMMAL
has_trunk: yes
*colour: grey
*size: large
*furry: no

**Clyde**
instance: ELEPHANT
colour: pink
owner: Fred

**Nellie**
instance: ELEPHANT
size: small

# Frame表示的优缺点

- **优点**
  - 直接表示领域知识模型
  - 支持默认推理
  - 高效
  - 支持过程知识（slot filler可以是一个过程）

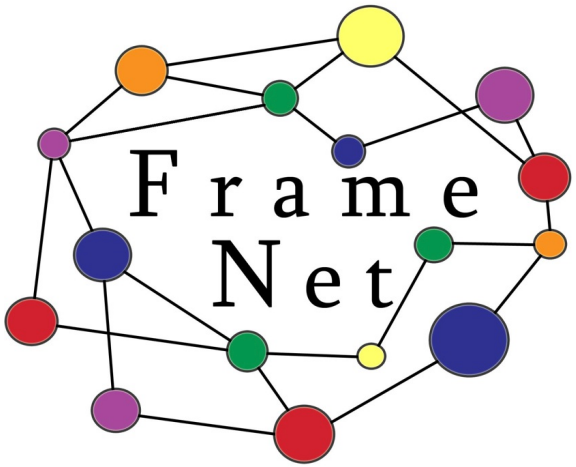
- **缺点**
  - 表达能力受限制
  - 缺乏标准（slot-filler values）
  - 更像是一种方法论而不是一种特定表示
  - 没有直接与reasoning/inference机制关联

# FrameNet--Frame知识库

- FrameNet [Baker/Fillmore/Lowe, 1998]
  - 围绕框架构建,论元标签在不同的框架之间共享
  - 共包含4000多个英文谓词
  - 200,000人工标注的句子（语义角色标注）
- 包括框架、词元、框架关系、例句及篇章

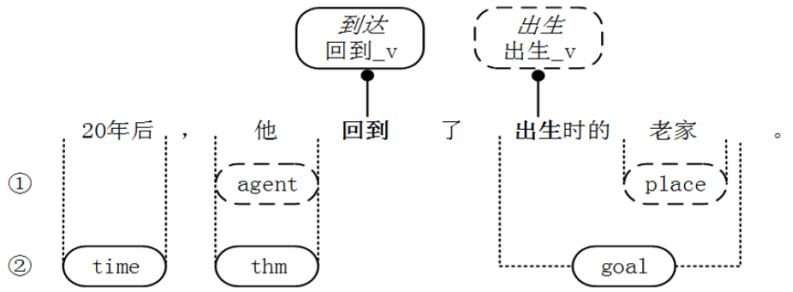
| 商业交易 | |
|---|---|
| 买家 | ? |
| 卖家 | ? |
| 物品 | ? |
| 价格 | ? |
| 交易时间 | ? |

可以触发该框架的词汇单元:
auction.n, auction.v
retail.v, retailer.n sale.n,
sell.v, seller.n vend.v,
vendor.n



# 汉语框架网Chinese FrameNet

- 山西大学汉语框架网与语义计算研究室 (http://sccfn.sxu.edu.cn/)
- 框架323个
- 词元3947个
- 例句20000条
- 全文标注200篇




| 框架名 | 位移 Motion | |
|---|---|---|
| 定义 | 转移体从起点出发，于终点结束，两个地点之间经过的是路径。从一般的[位移]框架继承的那些框架是在这个简单思想上加入了一些细节，如凸显终点的[到达]，凸显起点的[出发]，凸显路径的[穿越]。 | |
| 核心框架元素 | 转移体<br>Theme [Thm] | 转移体是改变地点的实体。<br>我今天去了体育馆。 |
| | 方向<br>Direction [Dir] | 移动的方向。<br>向校长走过去。 |
| | 终点<br>Goal [Goal] | 终点是转移体终止的地方。<br>车驶入了漫道。 |
| | 路径<br>Path [Path] | 路径是指转移体行进在其上的路线。<br>他绕过爸爸进了客厅。 |
| | 起点<br>Source [Src] | 起点是转移体在位置改变之前的处所。<br>警察从门口走开了。 |
| | 区域<br>Area [Area] | 区域是在没有明确路径的情况下，转移体位移的背景。<br>他在屋子里不安地走动。 |
| 非核心框架元素 | 载体<br>Carrier [Car] | 载体是运输转移体的工具。 |
| | 形容<br>Depictive [Dep] | 描述位移发生时转移体的状态。 |
| | 距离<br>Distance [Dist] | 距离表示位移的长度。<br>小嫩枝在水上漂流了大约 100 码。 |
| | 动作时量<br>Duration_of_action [Dur_action] | 位移发生的时间数量。<br>气球漂流了数小时。 |
| | 方式<br>Manner [Manr] | 位移发生的方式。<br>一艘海军飞艇在暴风雨中疯狂地漂流。 |
| | 速度<br>Speed [Spd] | 转移体位移的速度。<br>灰尘能够以每小时 25 公里的速度漂移。 |
| | 时间<br>Time [Time] | 位移发生的时间。<br>在对核工厂开火后，放射性云能够扩散着跨越英国。 |
| 框架关系 | 父框架：无<br>子框架：[集体位移/Mass_motion], [有向位移/Motion_directional]<br>总框架：[位移情境/Motion_scenario]<br>分框架：无<br>总　域：无<br>分　域：[到达/Arriving], [肢体运动/Body_movement], [带来/Bringing], [伴随/Cotheme], [离开/Departing], [散发/Emanating], [逃避/Evading], [分泌/Excreting], [液体运动/Fluidic_motion], [光运动/Light_movement], [运动噪音/Motion_noise], [运动情境/Motion_scenario], [交通工具操作/Operate_vehicle], [出售/Placing], [改道/Redirecting], [消除/Removing], [自动/Self_motion], [射击/Shoot_projectiles], [传播/ Travel]<br>参　照：无<br>后续过程：无<br>结果状态：无 | |
| 词元 | *吹v，漂流v，漂浮v，飞v，滑行v，走v，移动v，滚动v，滑动v，滑翔v，漂移v，漂v* | |

# SCRIPT/脚本

# Script表示

- 脚本与框架类似，由一组槽组成，用来表示特定领域内一组事件的发生序列
- 类似于Frame表示
  - 使用继承和slots
  - 描述原型知识(stereotypical knowledge)，但是关注事件知识
- 基于Conceptual Dependency Theory构建

R.C. Schank and R. Abelson (1977) Scripts, Plans, Goals and Understandingå

# Script定义

- 一个脚本是一个事件序列，包含了**一组紧密相关的动作及改变状态的框架** [Winston, 1992]

- 一个脚本是一个描述**特定上下文中的原型事件序列**的结构化表示 [Luger, Stubblefield,1998]

# Script的组成元素

- **进入条件**
  - 给出在脚本中所描述事件的前提条件。
- **角色**
  - 是一些用来表示在脚本所描述事件中可能出现的有关人物的槽。
- **道具**
  - 是一些用来表示在脚本所描述事件中可能出现的有关物体的槽。
- **场景**
  - 用来描述事件发生的真实顺序。一个事件可以由多个场景组成，而每个场景又可以是其他的脚本。
- **结局**
  - 给出在脚本所描述事件发生以后所产生的结果。

# Script例子

下面以夏克的“餐厅”脚本为例来说明各个部分的组成。

**(1) 进入条件**：① 顾客饿了，需要进餐；② 顾客有足够的钱。

**(2) 角色**：顾客，服务员，厨师，老板。

**(3) 道具**：食品，桌子，菜单，钱。

**(4) 场景**：

场景1：进入—— ① 顾客进入餐厅；② 寻找桌子；③ 在桌子旁坐下。

场景2：点菜—— ① 服务员给顾客菜单；② 顾客点菜；
③ 顾客把菜单还给服务员；④ 顾客等待服务员送菜。

场景3：等待—— ① 服务员告诉厨师顾客所点的菜；②厨师做菜，顾客等待。

场景4：吃饭—— ① 厨师把做好的荣给服务员；② 服务员把菜送给顾客；
③ 顾客吃菜。

场景5：离开—— ① 服务员拿来账单；② 顾客付钱给服务员；
③ 顾客离开餐厅。

**(5) 结果：**① 顾客吃了饭，不饿了；② 顾客花了钱；③ 老板赚了钱；④ 餐厅食品少了。

# Script的推理

- 基于脚本事件因果链的推理
  - 只有符合特定条件脚本才会发生
  - 只有符合结束条件脚本才会结束
  - 事件和事件直接的因果链顺序推理：*进入→ 点菜→ 等待→ 吃饭 → 付钱→ 离开*
  - 预测未知事件：如果已知该脚本适合于所给定的事件，则对一些在脚本中没有明显提出的事件，可以通过脚本进行预测

# Script的特点

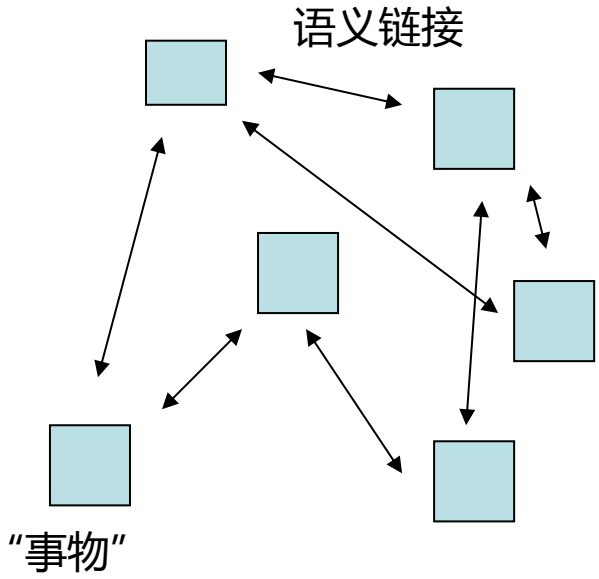
- 缺点
  - 脚本结构与框架结构相比表达能力更受约束
  - 表示范围更窄

- 优点：
  - 适合于表达预先构思好的特定知识或顺序性动作及事件，如理解故事情节等
  - 适合于自然语言理解中的阅读理解等应用

# SEMANTIC WEB知识表示语言

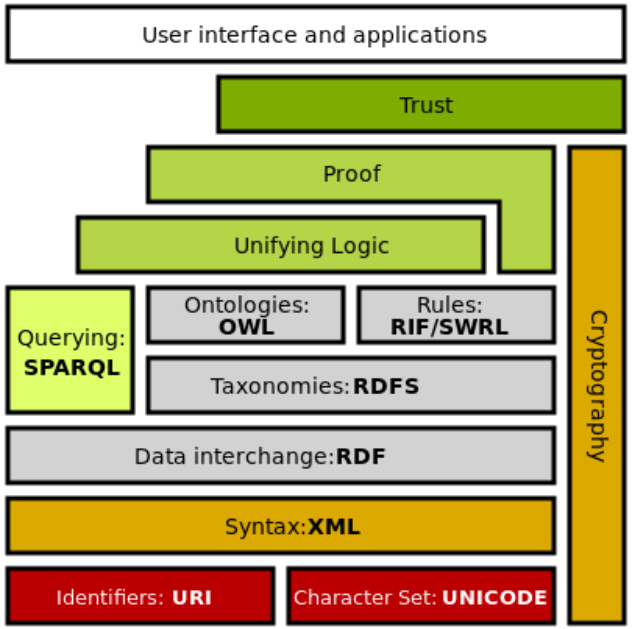
# 数据万维网Web of Data

- 全球开发的知识共享平台
- 使用语义网技术
  - 在Web上发布结构化数据
  - 在不同数据源中的数据之间建立连接
- 特征
  - Web上的事物拥有唯一的URI
  - 事物之间由链接关联(如人物、地点、事件、建筑物)
  - 事物之间链接显式存在并拥有类型
  - Web上数据的结构显式存在



# 语义网信息描述语言

- 语义网提供了一套为**描述数据而设计的表示语言**和工具，**用于形式化的描述一个知识领域内的概念、术语和关系**




- HTML描述文档和文档之间的链接

- RDF, RDFS, OWL和XML**能够描述事物和事物之间的关系，如人、会议、飞机和飞机组件**
  - Use URIs as names for things
  - When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
  - Include links to other URIs, so

# 主要组件

- 包括一系列的W3C标准和工具：
  - Resource Description Framework (RDF)
  - RDF Schema (RDFS)
  - Web Ontology Language (OWL)
  - SPARQL, an RDF query language
  - …

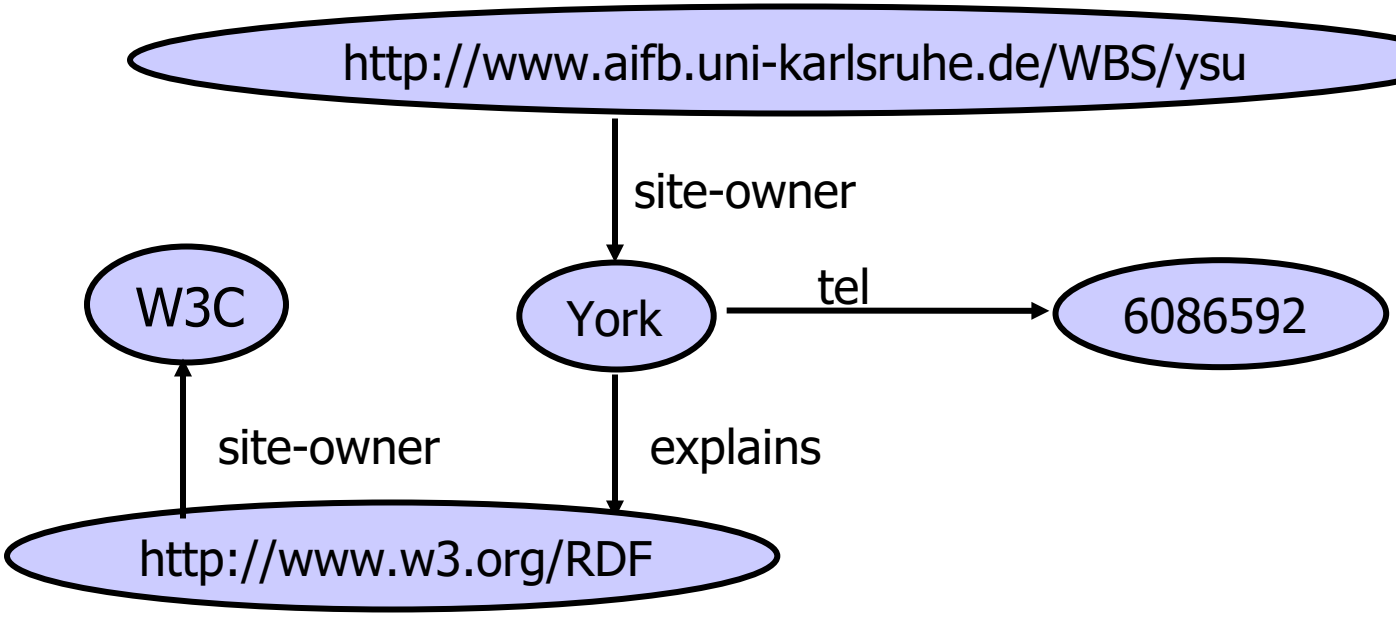
# RDF

- RDF是一种表述对象（web resources）和对象之间关系的简单语言
- 使用**(subject, predicate, object)**三元组的形式来陈述关于对象(使用URI标识的resources)的知识，也就是两个对象之间的带类别链接
- RDF是一个通用模型，可以用各种不同的格式表示，如XML、N-Triples、N3、JSON-LD等


主语　谓语　宾语

```
(<http://…isbn…6682>, <http://…/original>, <http://…isbn…409X>)
```

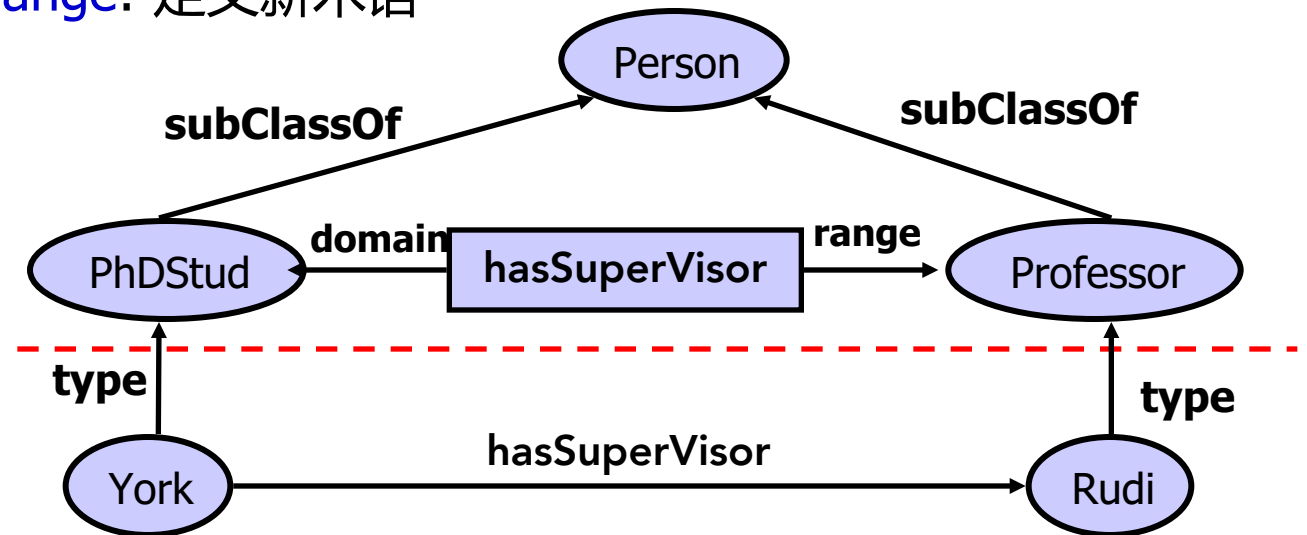
# RDF描述示例




```
<rdf:Description rdf:about="#York">
   <tel>6086592</tel>
</rdf:Description>
```

# RDF Schema

- RDFS 是RDF的一个扩展，提供了一个用于描述RDF resources的属性(properties)和类别(classes)的术语表(**vocabulary**)
- 上述词表被组织成一个带类别的层次体系结构(typed hierarchy)
  - Class, subClassOf, type : 描述类别子类别
  - Property, subPropertyOf: 属性层次体系结构
  - domain, range: 定义新术语



# RDF(S)术语表

| RDF | RDFS |
| --- | --- |
| rdf:type<br>rdf:Property | rdfs:domain<br>rdfs:range<br>rdfs:Resource<br>rdfs:Class<br>rdfs:subClassOf<br>rdfs:subPropertyOf |
| … others (rectification, annotation, literal, collection, container) | |

# RDF Schema syntax in XML

```
<rdf:Description ID="MotorVehicle">
    <rdf:type resource="http://www.w3.org/...#Class"/>
    <rdfs:subClassOf rdf:resource="http://www.w3.org/...#Resource"/>
   </rdf:Description>

<rdf:Description ID="Truck">
    <rdf:type resource="http://www.w3.org/...#Class"/>
    <rdfs:subClassOf rdf:resource="#MotorVehicle"/>
</rdf:Description>

<rdf:Description ID="registeredTo">
    <rdf:type resource="http://www.w3.org/...#Property"/>
    <rdfs:domain rdf:resource="#MotorVehicle"/>
    <rdfs:range rdf:resource="#Person"/>
</rdf:Description>

<rdf:Description ID=”ownedBy">
    <rdf:type resource="http://www.w3.org/...#Property"/>
    <rdfs:subPropertyOf rdf:resource="#registeredTo"/>
</rdf:Description>
```

# 术语表

- RDFS提供了定义术语表(*vocabularies*)的能力:
  - 属性集合和类别集合
  - 与其它术语表中的术语的关系
- 术语表例子:
  - Dublin Core terms: creator, date, …
  - FOAF terms: characterization of persons
  - Good Relations: eCommerce terms
  - Creative Commons: copyright classes, license relations, …
  - **schema.org terms: events, organizations, places, reviews, …**
  - …

# RDF和RDFS

- RDF(S)提供了
  - (很小的)**本体约定(ontological commitment)** 来建模primitives
  - 一个用于知识表示的词汇表(subClassOf、subPropertyOf、domain、range...)
  - 可以用来定义术语表(**vocabulary)**
- 但是:
  - 不能准确描述语义
  - 缺少推理模型

Web Ontology Language = OWL

# 本体Ontology

"People can't **share knowledge** if they do not speak a **common language**." [Davenport & Prusak, 1998]

"An ontology is an **explicit specification** of a **conceptualization**." [Gruber, 1993]

- 本体提供了人和机器之间的更好的交流(communicate)机制
- 本体通过概念**标准化(standardize)**和**形式化(formalize)**词语的意义

# 本体的五元组表示$O = \{C, R, F, A, I\}$

- **C – 概念集合**，通常以Taxonomy形式组织
  - 球星，清华校友
- **R – 关系**，描述概念或者实例之间语义关系的集合
  - subClassOf，birthplace
- **F – 函数**, 一组特殊的关系，关系中第n个元素的值由其他n-1个元素的值确定
  - Price-of-a-used-car 由 the car-model, manufacturing data 和 kilometers确定
- **A – 公理**
  - 如果A是B的子女，B是C的子女，则A是C的子孙
- **I –具体个体**
  - 如:Peter是概念学生的实例

# Web Ontology Language = OWL

- **OWL进一步提供了更多的术语来描述属性和类别**
  - 类别之间的关系 (e.g. disjointness)
  - 基数cardinality (e.g. "exactly one")
  - equality
  - richer typing of properties
  - characteristics of properties (e.g. symmetry)
  - 枚举类
  - ...

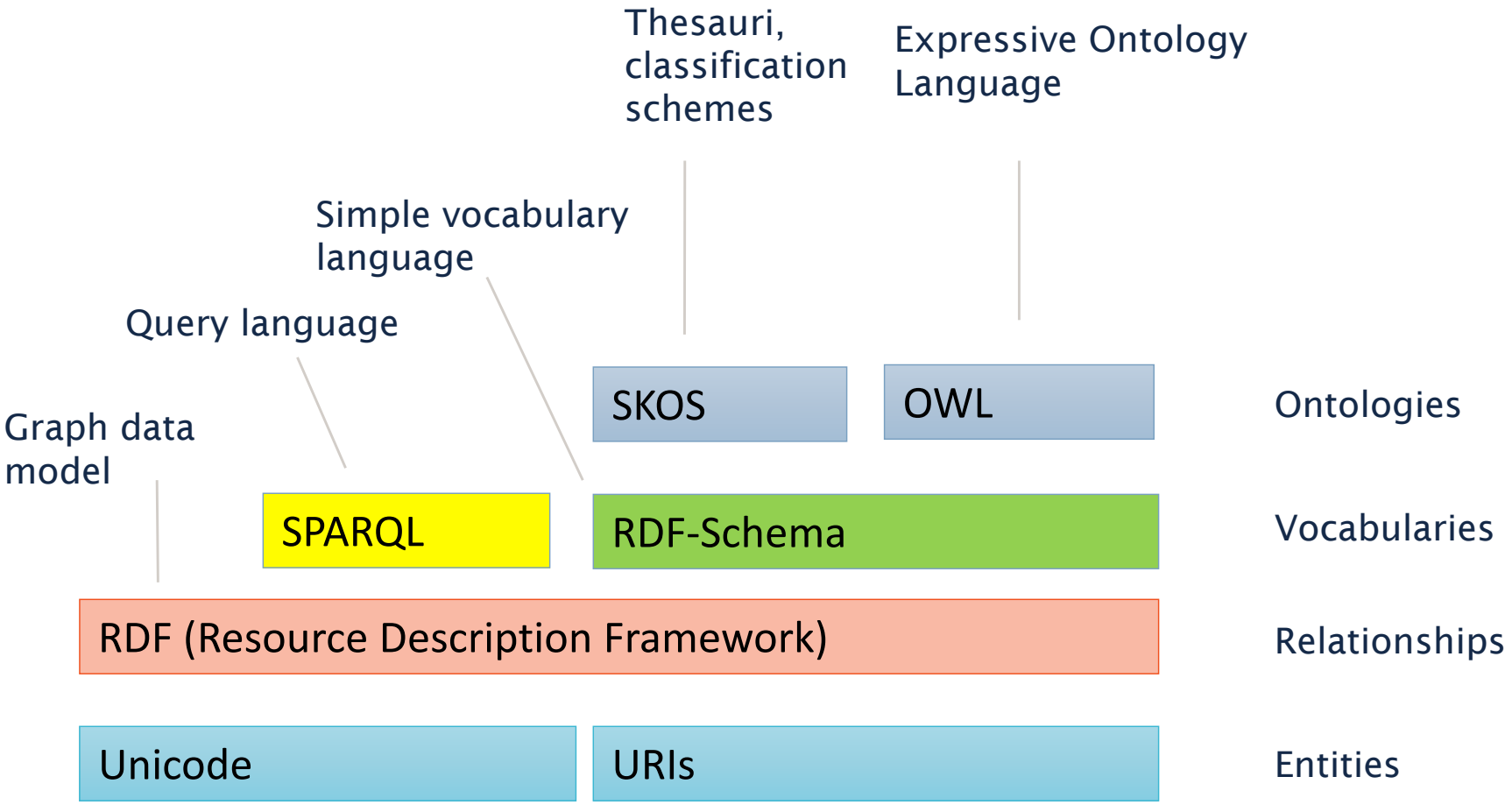
# 相比于RDFS的扩展

- **构建类别**
  - 如：与，或，非，存在，任一
- **构造属性**
  - 如： inverseOf
- **属性特征**
  - 如: transitive, functional, symmetric
  - email属性是inverse functional, i.e., two different subjects cannot have identical objects
- **属性和类别间关系**
  - Equality, non-equality(between classes, properties, ind.)
  - wl:equivalentClass: two classes have the same individuals
  - owl:disjointWith: no individuals in common

# OWL词汇表

| Classes | Class Construction | |
|---|---|---|
| owl:Class<br>owl:Thing<br>owl:Nothing | owl:complementOf<br>owl:intersectionOf<br>owl:unionOf | Boolean |
| owl:Restriction<br>owl:onProperty | owl:allValuesFrom<br>owl:someValuesFrom<br>owl:hasValue | qualification |
| **Non-equality** | | |
| owl:differentFrom<br>owl:disjointWith<br>owl:AllDifferent<br>owl:distinctMembers | owl:cardinality<br>owl:minCardinality<br>owl:maxCardinality | cardinality |
| | owl:oneOf | |
| **+ RDFS Plus vocabulary** | | |

# 语义网知识描述语言体系

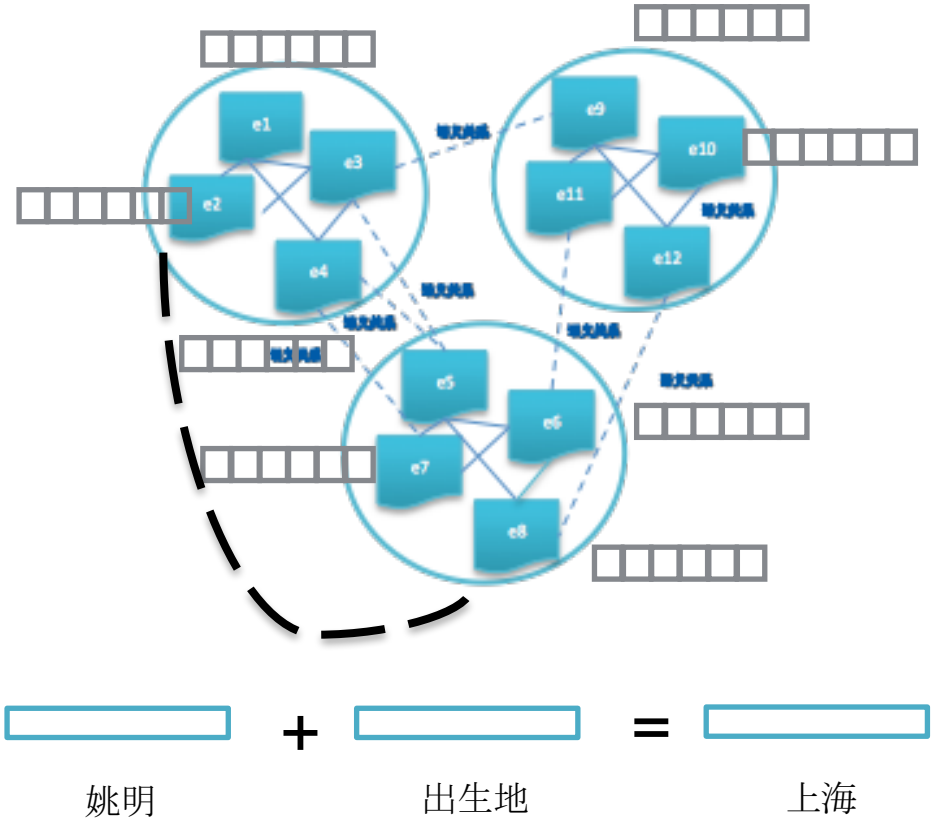


# 知识表示小结

- **一阶谓词逻辑**
  - 一套灵活且紧凑的对象知识表示方式
  - 完善的推理能力和表达能力支持
- **Semantic Net**
  - 一个通过语义关系连接的概念网络
  - 表达能力受限，直观但是缺乏语义支持
- **Frame**
  - 直接表示领域知识模型,支持默认推理
  - 表达能力受限，缺乏语义支持
- **Script**
  - 表达事件知识，非常受限， 对符合条件的场景非常适合
- **语义网知识表示语言体系**
  - 面向Web of Data,完善的一整套体系支持
  - 有的时候杀鸡用牛刀

# 目录

# 基于分布式的知识表示和推理

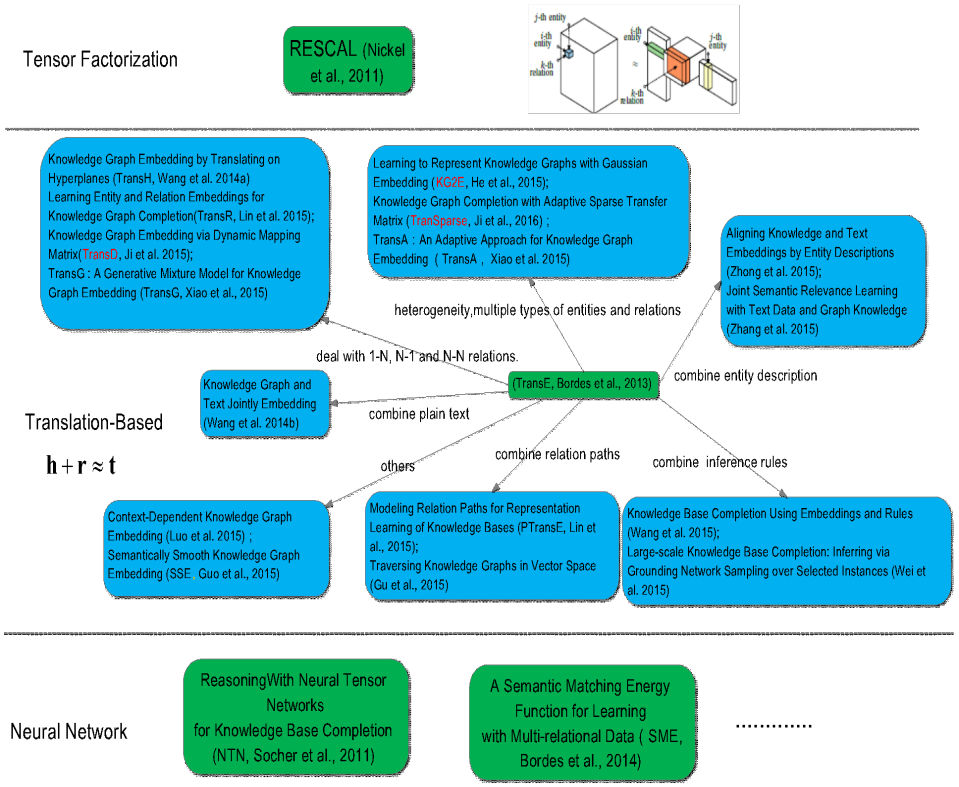


# 知识图谱表示学习方法分类

张量分解

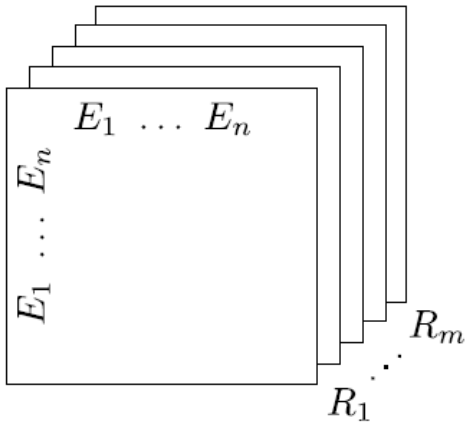
基于翻译的模型

神经网络模型



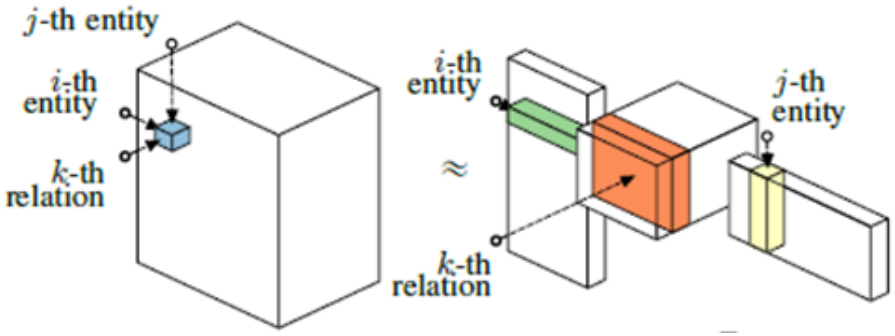
# 用张量表示知识图谱

- 知识图谱中三元组的结构是（头部实体$h$，关系$r$，尾部实体$t$），其中$r$连接头尾实体。以$E_1, E_2, \cdots E_n$表示知识图谱中的实体，以$R_1, R_2, \cdots R_n$表示知识图谱中的关系，则可以使用一个三维矩阵（张量）表示知识图谱




Nickel et al. (2011). A three-way model for collective learning on multi-relational data. In Proceedings of the 28th international conference on machine learning (ICML-11).

# 张量分解得到实体、关系表示



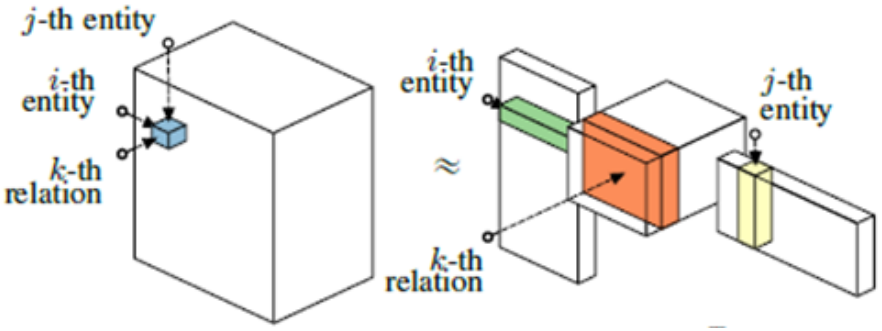
# 分解的目标函数

- 表示知识图谱的张量记为$\hat{Y}$，其第$k$个矩阵记为$Y_k$，则有

$$\mathrm{Y_k} = \mathrm{AR_kA^T} \qquad k = 1,2,\cdots m$$

- 其中$A \in R^{n\times r}$，$Y_k \in R^{n\times n}$，$R_k \in R^{r\times r}$.

- 这是一个低秩分解，$r$表示矩阵$A$的秩。 $A$的每一行表示一个实体的向量，转置后其每一列表示一个实体的向量，矩阵$Y_k$是第$k$种关系的矩阵，表示该种关系在向量空间中与头尾部实体相互作用。



# 分解的目标函数

- 由上述内容可知，$\mathbf{A}$和$\mathbf{R}_k$均是待求解的变量。因此目标函数是：

$$\min_{\mathbf{A},\mathbf{R}_k} f\left(\mathbf{A},\mathbf{R}_k\right)+g\left(\mathbf{A},\mathbf{R}_k\right)$$

其中$f\left(\mathbf{A},\mathbf{R}_k\right)$是目标函数

$$f\left(\mathbf{A},\mathbf{R}_k\right)=\frac{1}{2}\left(\sum_k\left\|\mathbf{Y}_k-\mathbf{A}\mathbf{R}_k\mathbf{A}^T\right\|_F^2\right)$$

$g\left(\mathbf{A},\mathbf{R}_k\right)$是正则化项：

$$g\left(\mathbf{A},\mathbf{R}_k\right)=\frac{1}{2}\lambda\left(\left\|\mathbf{A}\right\|_F^2+\sum_k\left\|\mathbf{R}_k\right\|_F^2\right)$$

# 分解的目标函数

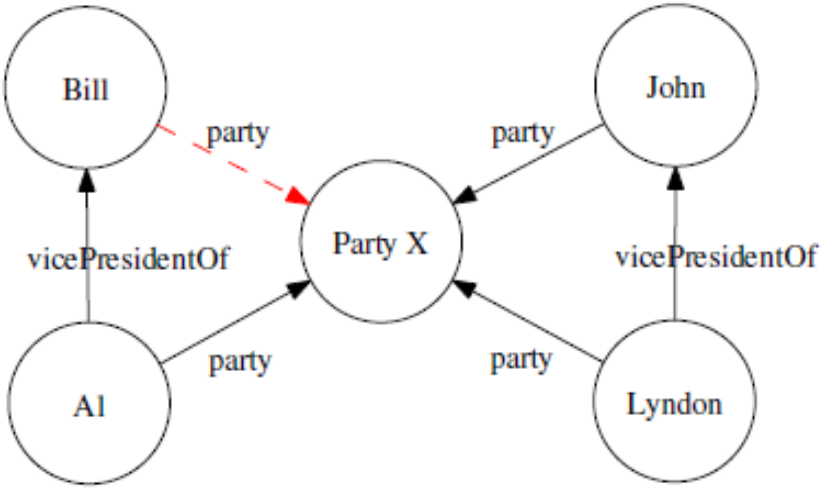
- 将目标函数写成分量形式

$$f(\mathbf{A},\mathbf{R}_k)=\frac{1}{2}\left(\sum_k\left\|\mathbf{Y}_k-\mathbf{A}\mathbf{R}_k\mathbf{A}^T\right\|_F^2\right)\Rightarrow f(\mathbf{A},\mathbf{R}_k)=\frac{1}{2}\sum_{i,j,k}\left(y_{ijk}-\mathbf{a}_i^T\mathbf{R}_k\mathbf{a}_j\right)^2$$

其中 $y_{ijk}$ 是张量中的一个元素，$\mathbf{a}_i$ 表示 $\mathbf{A}$ 的第 $i$ 行，即：

$$[\mathbf{a}_1,\mathbf{a}_2,.....,\mathbf{a}_n]=\mathbf{A}$$

# 模型的解释




$$\mathbf{a}_{\text{A1}}^{T}\mathbf{R}_{\text{party}}\mathbf{a}_{\text{partyX}} \approx \mathbf{a}_{\text{Lyndon}}^{T}\mathbf{R}_{\text{party}}\mathbf{a}_{\text{partyX}} \Rightarrow \mathbf{a}_{\text{A1}}^{T} \approx \mathbf{a}_{\text{Lyndon}}^{T}$$

$$\mathbf{a}_{\text{A1}}^{T}\mathbf{R}_{\text{vicePresidentOf}}\mathbf{a}_{\text{Bill}} \approx \mathbf{a}_{\text{Lyndon}}^{T}\mathbf{R}_{\text{vicePresidentOf}}\mathbf{a}_{\text{John}} \Rightarrow \mathbf{a}_{\text{Bill}} \approx \mathbf{a}_{\text{John}}$$

# 分解的计算方法

- 更新 $\mathbf{A}$

$$\mathbf{A} \leftarrow \left[\sum_{k=1}^{m} \mathbf{Y}_k \mathbf{A} \mathbf{R}_k^T + \mathbf{Y}_k^T \mathbf{A} \mathbf{R}_k\right]\left[\sum_{k=1}^{m} \mathbf{B}_k + \mathbf{C}_k + \lambda \mathbf{I}\right]^{-1}$$

$$\mathbf{B}_k = \mathbf{R}_k \mathbf{A}^T \mathbf{A} \mathbf{R}_k^T, \quad \mathbf{C}_k = \mathbf{R}_k^T \mathbf{A}^T \mathbf{A} \mathbf{R}_k$$

- 更新 $\mathbf{R}_k$

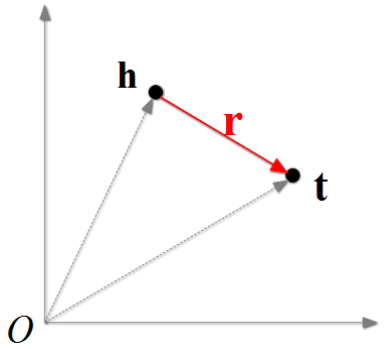
$$\mathbf{R}_k \leftarrow \left(\mathbf{Z}^T \mathbf{Z} + \lambda \mathbf{I}\right)^{-1} \mathbf{Z}\ \mathrm{vec}\left(\mathbf{Y}_k\right)$$

$$\mathbf{Z} = \mathbf{A} \otimes \mathbf{A}$$

- 迭代直到 $f\left(\mathbf{A}, \mathbf{R}_k\right) / \left\|\hat{\mathbf{Y}}\right\|_F^2 \le \varepsilon$ 或者超过最大迭代次数。$\varepsilon$ 是设定的一个很小的数

# 基于翻译的模型：TransE

- 用向量表示实体和关系。关系事实 = (head, relation, tail) 简写为（$h, r, t$），其对应的向量表示为（$\mathrm{h}, \mathrm{r}, \mathrm{t}$）。




$$\mathbf{h} + \mathbf{r} = \mathbf{t}$$

中国+首都=北京

法国+首都=巴黎

俄罗斯+首都=莫斯科

Bordes, et al. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems, 2013* (pp. 2787-2795).

# 翻译模型的学习

- 势能函数
  - 对于真实事实的三元组（$h,r,t$），要求$h+r=t$；而对于错误的三元组则不满足该条件

$$f\left(h,r,t\right)=\left\|\mathbf{h}+\mathbf{r}-\mathbf{t}\right\|_2$$

$f$(姚明　出生于　北京)　　>　　$f$ (姚明　　出生于　上海)

# 翻译模型的学习

- 目标函数：

$$\sum_{(h,r,t)\in\Delta}\sum_{(h',r,t')\in\Delta'}[\gamma+f(h,r,t)-f(h',r,t')]_+$$

其中 $[x]_+=\max(0,x)$， $\Delta$ 表示知识库中三元组的集合，$\Delta'$ 表示三元组的负样本集合。

约束条件： $(h,r,t)$ $\|\mathbf{h}\|\le 1,\|\mathbf{r}\|\le 1,\|\mathbf{t}\|\le 1$

# 生成负样本的方法

- 负样本生成策略

1. 在实体集合中随机选择实体$h'(t')$，替换$(h,r,t)$中的 $h(t)$，生成负样本$(h',r,t)$或者$(h,r,t')$。

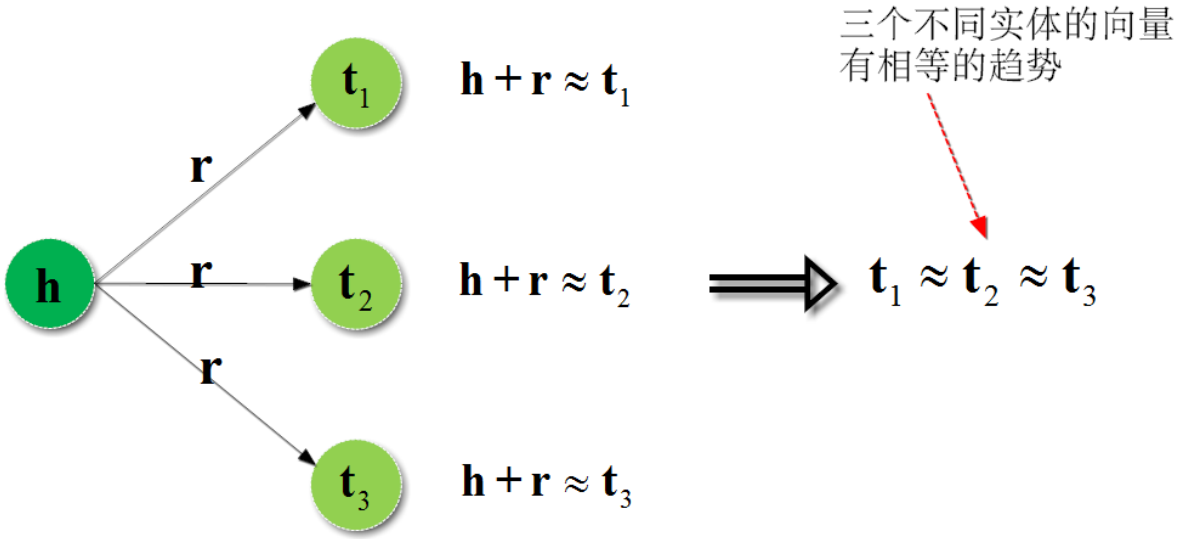
2. 在选择替换实体的时候，不是完全随机在实体集合中选择，而是在适合关系 $r$ 关系的实体集合中随机选取。例如：

   对进行尾部实体替换时，只是用其他的地名替换“上海”，如“成都”，而不会使用人名进行替换。

(姚明　出生于　上海)

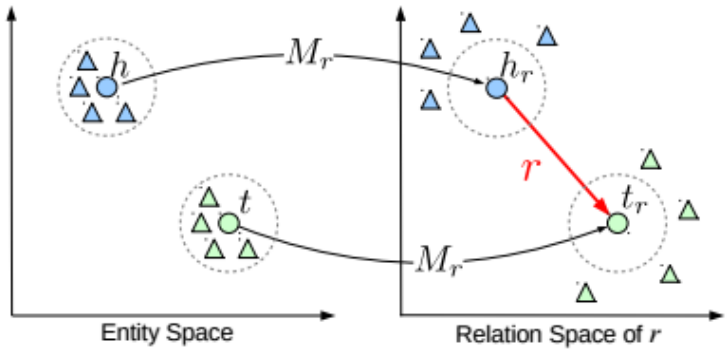
# 知识图谱数据问题

- 知识图谱中关系有“1-1”、“1-N”、“N-1”、“N-N”多种类型



# 解决方案: TransR

- **TranH、TransR、TransD**




TransH

$$\mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r \qquad \mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r$$

$$f_r(\mathbf{h},\mathbf{t}) = \|(\mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r)\|_2^2$$

TransR

$$\mathbf{h}_r = \mathbf{h}\mathbf{M}_r \quad \mathbf{t}_r = \mathbf{t}\mathbf{M}_r$$

$$f_r(h,t) = \|\mathbf{h}_r + \mathbf{r} - \mathbf{t}_r\|_2^2$$

TransD

$$\mathbf{M}_{rh} = \mathbf{r}_p\mathbf{h}_p^\top + \mathbf{I}^{m\times n}$$
$$\mathbf{M}_{rt} = \mathbf{r}_p\mathbf{t}_p^\top + \mathbf{I}^{m\times n}$$

$$\mathbf{h}_\perp = \mathbf{M}_{rh}\mathbf{h}, \quad \mathbf{t}_\perp = \mathbf{M}_{rt}\mathbf{t}$$

$$f_r(\mathbf{h},\mathbf{t}) = -\|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_2^2$$

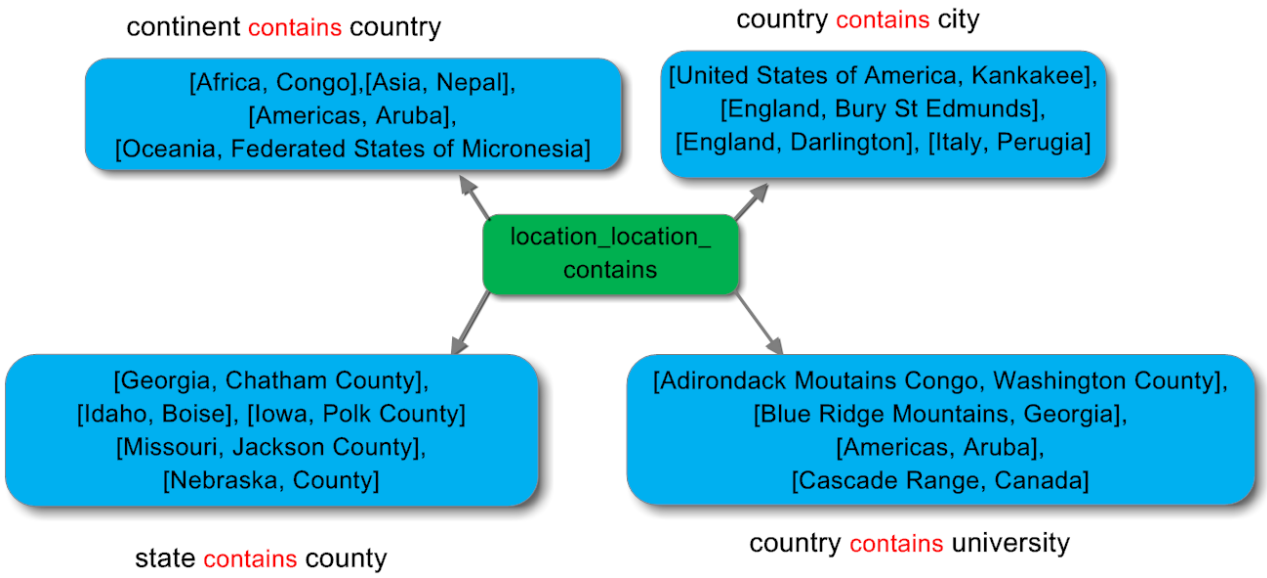

Wang, et al. Knowledge Graph Embedding by Translating on Hyperplanes. In Proceedings of *AAAI* 2014

Lin, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In Proceedings of *AAAI* 2015

Ji, et al. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of ACL* 2015

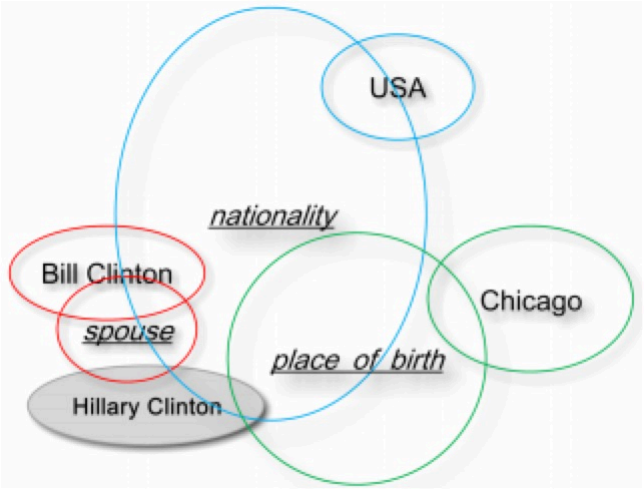
# 知识图谱数据问题

- 实体和关系通常会出现在多个不同的三元组中，类似于一词多义，实体和关系在不同的三元组中常呈现出不同的含义。



# 利用协方差描述关系的不确定性

- 多维高斯分布表示符号
  - 均值向量表示该符号的位置(含义)
  - 协方差矩阵表示该符号的多样性(不确定性)
    - 包含事实越多，该实体语义越明确
    - 关系越复杂，该关系确定性越弱




He, et al. Learning to represent knowledge graphs with Gaussian embedding. In *Proceedings of CIKM 2015*

# KG2E

$$P_e = H - T \sim N(u_h - u_t, \Sigma_h + \Sigma_t) \qquad P_r = R \sim N(u_r, \Sigma_r)$$

$$\begin{aligned}
\mathcal{E}(h,r,t) &= \mathcal{E}(\mathcal{P}_e, \mathcal{P}_r) = \mathcal{D}_{\mathcal{KL}}(\mathcal{P}_e, \mathcal{P}_r) \\
&= \int_{x \in \mathcal{R}^{k_e}} \mathcal{N}(x; \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r) \log \frac{\mathcal{N}(x; \boldsymbol{\mu}_e, \boldsymbol{\Sigma}_e)}{\mathcal{N}(x; \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)} dx \\
&= \frac{1}{2} \Big\{ tr(\boldsymbol{\Sigma}_r^{-1} \boldsymbol{\Sigma}_e) + (\boldsymbol{\mu}_r - \boldsymbol{\mu}_e)^T \boldsymbol{\Sigma}_r^{-1} (\boldsymbol{\mu}_r - \boldsymbol{\mu}_e) \\
&\qquad - \log \frac{det(\boldsymbol{\Sigma}_e)}{det(\boldsymbol{\Sigma}_r)} - k_e \Big\}
\end{aligned}$$

势函数 $\mathcal{E}(h,r,t) = \frac{1}{2}(\mathcal{D}_{\mathcal{KL}}(\mathcal{P}_e, \mathcal{P}_r) + \mathcal{D}_{\mathcal{KL}}(\mathcal{P}_r, \mathcal{P}_e))$

目标函数 $\mathcal{L} = \sum_{(h,r,t) \in \Gamma} \sum_{(h',r',t') \in \Gamma'_{(h,r,t)}} [\mathcal{E}(h,r,t) + \gamma - \mathcal{E}(h',r',t')]_+$

# 神经网络方法

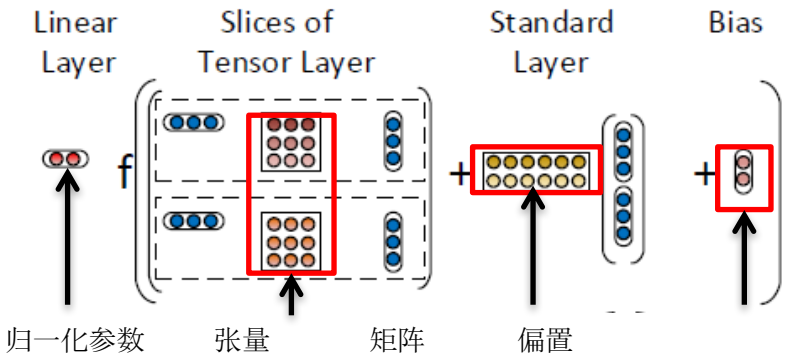
- 神经网络模型
  - Neural Tensor Network
  - Semantic Matching Energy Network

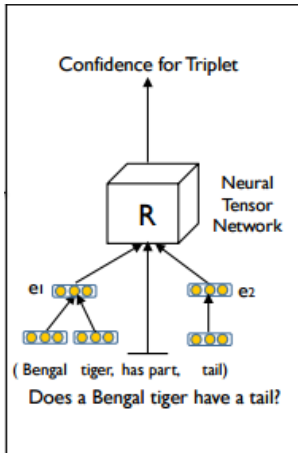
神经网络模型的核心思想：使用神经网络为三元组定义势能函数，在训练目标中，要求正确的三元组具有较高的能量，错误的三元组具有较低的能量。通过惩罚错误的三元组完成学习过程，使得正确的三元组和错误三元组的能量有一个明显的分界线。

$$\mathcal{L} = \sum_{(h,r,t)\in\Gamma} \sum_{(h',r',t')\in\Gamma'_{(h,r,t)}} [\mathcal{E}(h,r,t) + \gamma - \mathcal{E}(h',r',t')]_+$$

# Neural Tensor Network

- 关系表示



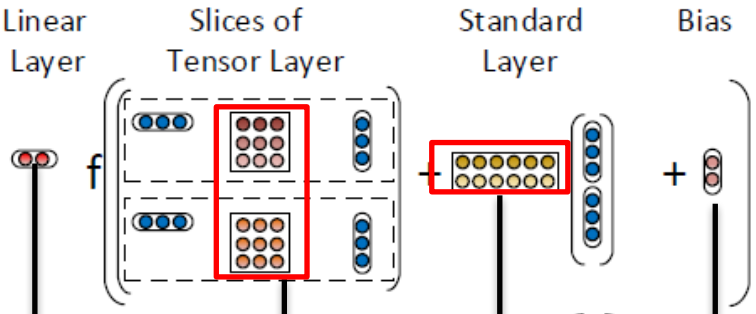




g(h,r,t)

Socher, et al. Reasoning with neural tensor networks for knowledge base completion. In Advances in Neural Information Processing Systems 2013

# Neural Tensor Network

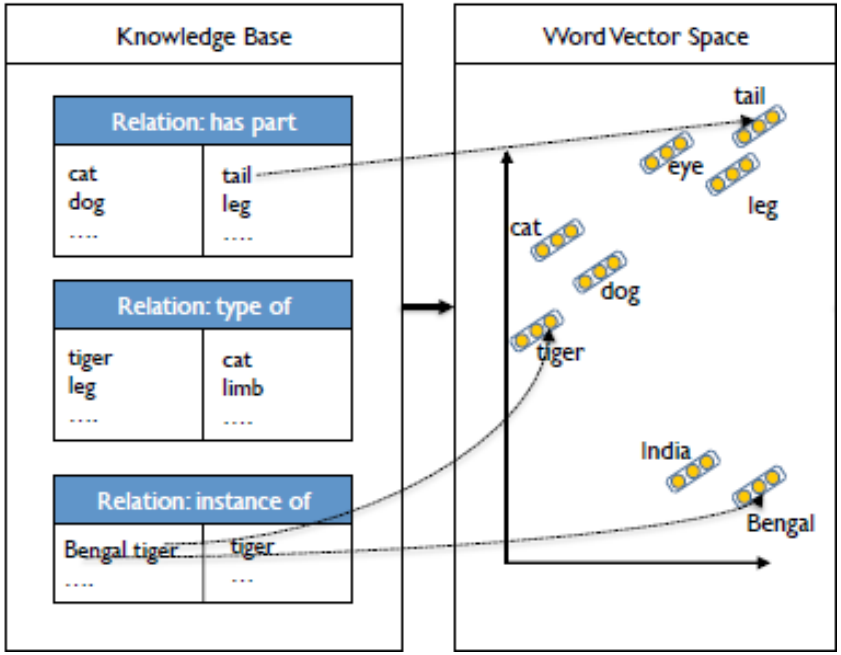
- 势能函数表示




$$g(h,r,t) = \mathbf{u}_r^T f\left(\mathbf{h}^T \mathbf{W}_r^{[1:k]} \mathbf{t} + \mathbf{V}_r \begin{bmatrix} \mathbf{h} \\ \mathbf{t} \end{bmatrix} + \mathbf{b}_r\right)$$

# Neural Tensor Network

- 实体表示




BankOfChina = （Bank + Of + China）/ 3

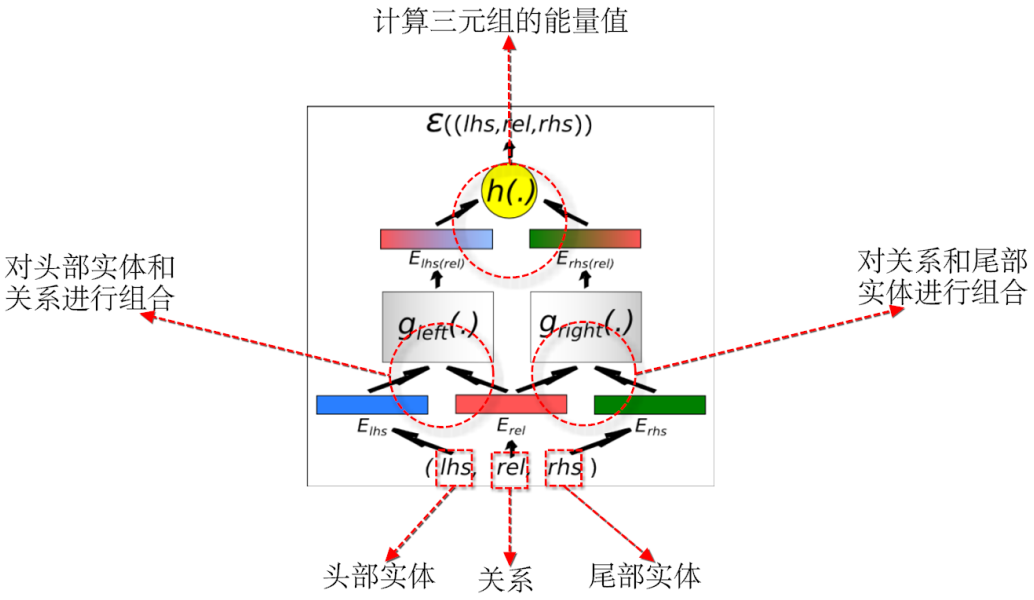
# Neural Tensor Network

- 训练目标和方法

$$J(\mathbf{\Omega}) = \sum_{i=1}^{N}\sum_{c=1}^{C}\max\left(0, 1 - g\left(T^{(i)}\right) + g\left(T_c^{(i)}\right)\right) + \lambda\left\|\mathbf{\Omega}\right\|_2^2$$

- 训练集中正样本：$T^{(i)} = \left(h^{(i)}, r^{(i)}, t^{(i)}\right)$
- 负样本：$T_c^{(i)} = \left(h^{(i)}, r^{(i)}, t_c\right)$，随机选择实体替换头部或者尾部实体
- 参数：$\mathbf{\Omega}$ 表示所有待学习的参数，包括实体向量，关系的张量，矩阵等。
- 优化方法：L-BFGS

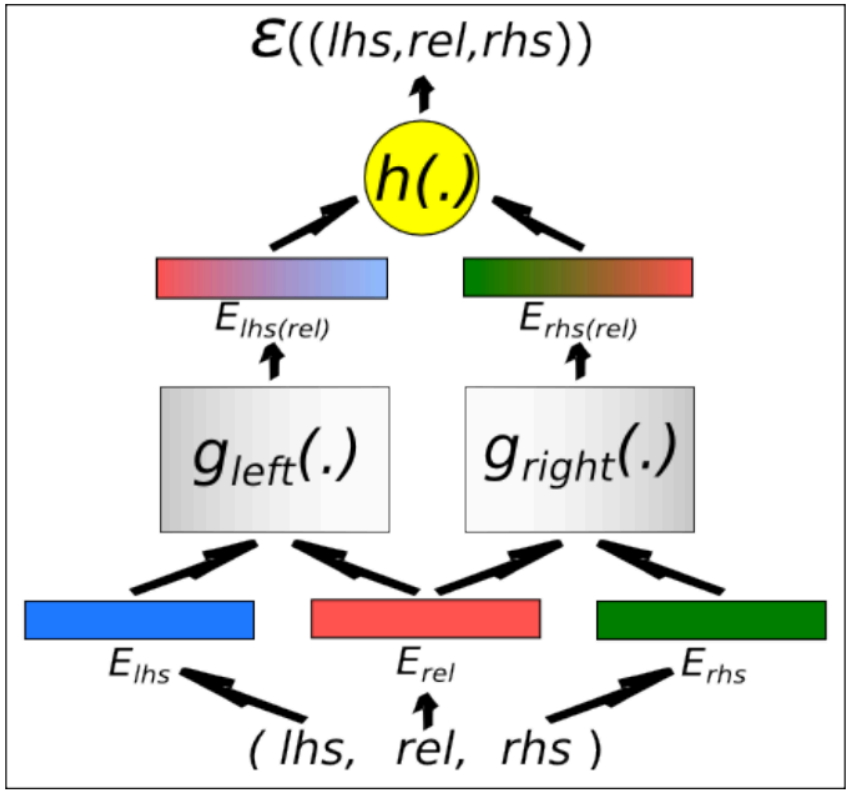
# Semantic Matching Energy Network




$$\mathcal{L} = \sum_{(h,r,t)\in\Gamma} \sum_{(h',r',t')\in\Gamma'_{(h,r,t)}} [\mathcal{E}(h,r,t) + \gamma - \mathcal{E}(h',r',t')]_+$$

Bordes, et al. A semantic matching energy function for learning with multi-relational data. Machine Learning, 94(2), 233-259, 2014.

# Semantic Matching Energy Network




Linear Form

$$E_{lhs}(rel) = g_{left}(E_{lhs}, E_{rel}) = W_{l1}E_{lhs}^{\intercal} + W_{l2}E_{rel}^{\intercal} + b_l^{\intercal}.$$
$$E_{rhs}(rel) = g_{right}(E_{rhs}, E_{rel}) = W_{r1}E_{rhs}^{\intercal} + W_{r2}E_{rel}^{\intercal} + b_r^{\intercal}.$$
$$\mathcal{E}((lhs, rel, rhs)) = \left(W_{l1}E_{lhs}^{\intercal} + W_{l2}E_{rel}^{\intercal} + b_l^{\intercal}\right)^{\intercal}\left(W_{r1}E_{rhs}^{\intercal} + W_{r2}E_{rel}^{\intercal} + b_r^{\intercal}\right)$$

Bi-linear Form

$$E_{lhs}(rel) = g_{left}(E_{lhs}, E_{rel}) = \left(W_l \bar{\times}_3 E_{rel}^{\intercal}\right) E_{lhs}^{\intercal} + b_l^{\intercal}.$$
$$E_{rhs}(rel) = g_{right}(E_{rhs}, E_{rel}) = \left(W_r \bar{\times}_3 E_{rel}^{\intercal}\right) E_{rhs}^{\intercal} + b_r^{\intercal}.$$
$$\mathcal{E}((lhs, rel, rhs)) = \left(\left(W_l \bar{\times}_3 E_{rel}^{\intercal}\right) E_{lhs}^{\intercal} + b_l^{\intercal}\right)^{\intercal}\left(\left(W_r \bar{\times}_3 E_{rel}^{\intercal}\right) E_{rhs}^{\intercal} + b_r^{\intercal}\right)$$

# 评测任务与数据集

- 三元组分类
  - 任务描述

    判断给定的三元组是否是正确的。例如：

    （姚明 出生于 上海）√　　（姚明 出生于 成都）×

  - 评测标准

    这是一个二分类任务，以分类准确率为评测指标。

  - 数据集

    常用的数据集有 WN11， FB13， FB15k。

- **WN11**：训练集112,581三元组，验证集2,609三元组，测试集10,544三元组，包含实体38,696个，关系11种。
- **FB13**：训练集316,232三元组，验证集5,908三元组，测试集23,733三元组。包含实体75,043个，关系13种。
- **FB15k**：训练集483,142三元组，验证集50,000三元组，测试集50,000三元组。包含实体14,951个，关系1,345种。

# 评测任务与数据集

- 链接预测
  - 任务描述

    挖去三元组中的实体或者关系，然后在实体（关系）集中选择实体（关系）将其补全：

    （姚明 出生于 上海）　　（姚明 出生于 ？）成都 北京 上海 .....

    （姚明 出生于 上海）　　（姚明 ？上海）飞往 居住地 出生于 .....

  - 评测标准

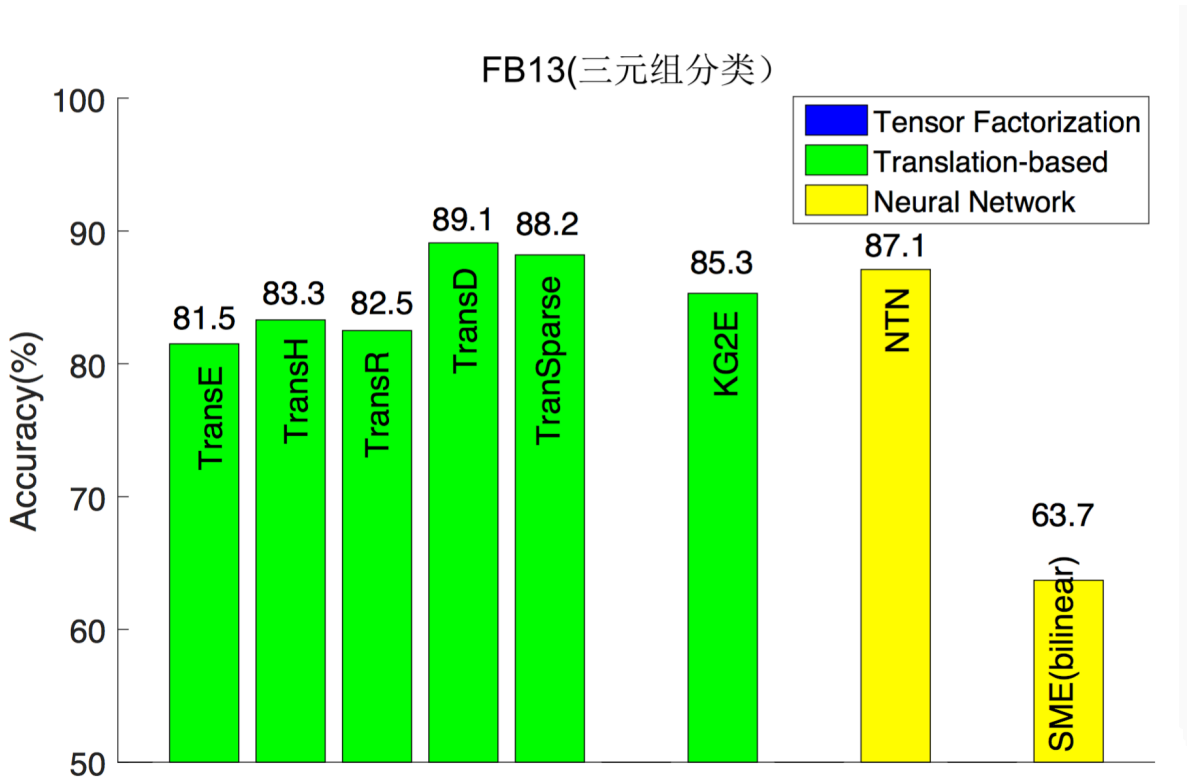
    计算正确实体的排名，排名越靠前，模型越优。计算测试集所有三元组头尾部实体的平均排名 mean rank和排名在前10的比例 @Hits10。

  - 数据集

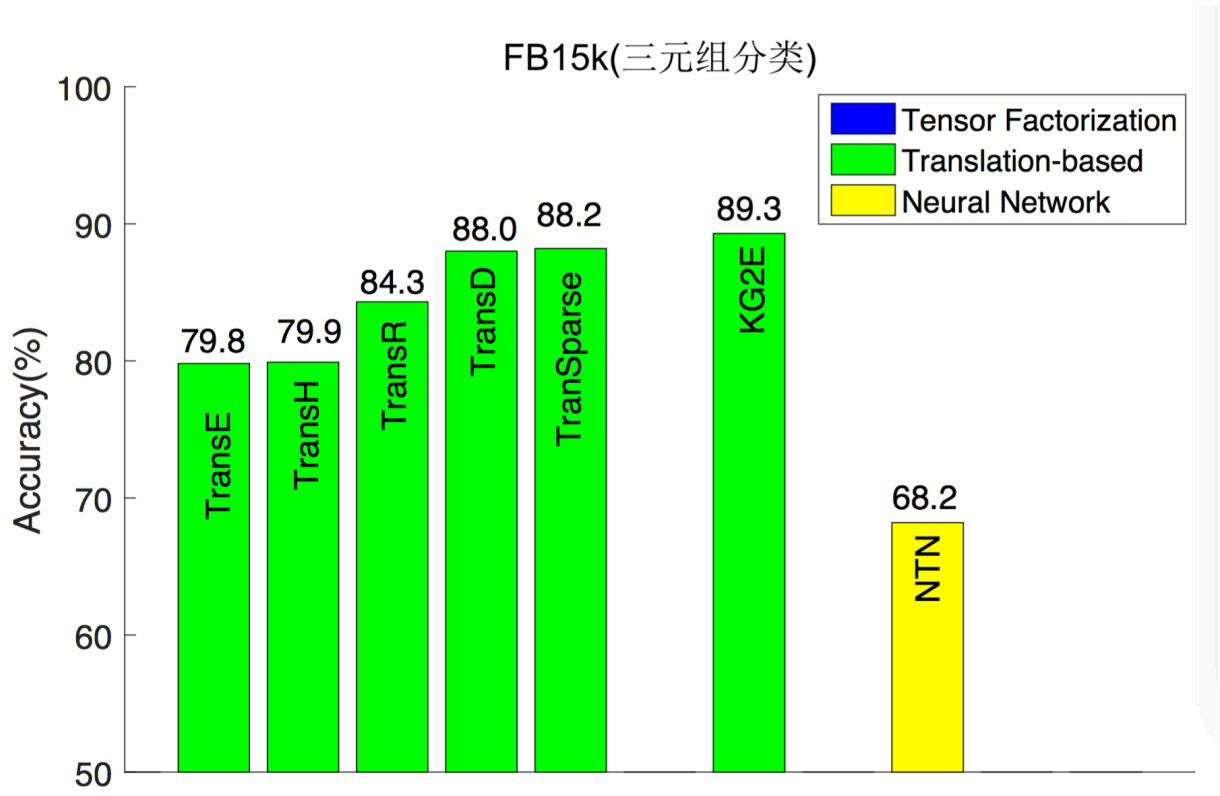
    常用的数据集有 WN18， FB15k。

- **WN18**：训练集141,442三元组，验证集5,000三元组，测试集5,000三元组，包含实体40,943个，关系18种。
- **FB15k**：训练集483,142三元组，验证集50,000三元组，测试集50,000三元组。包含实体14,951个，关系1,345种。

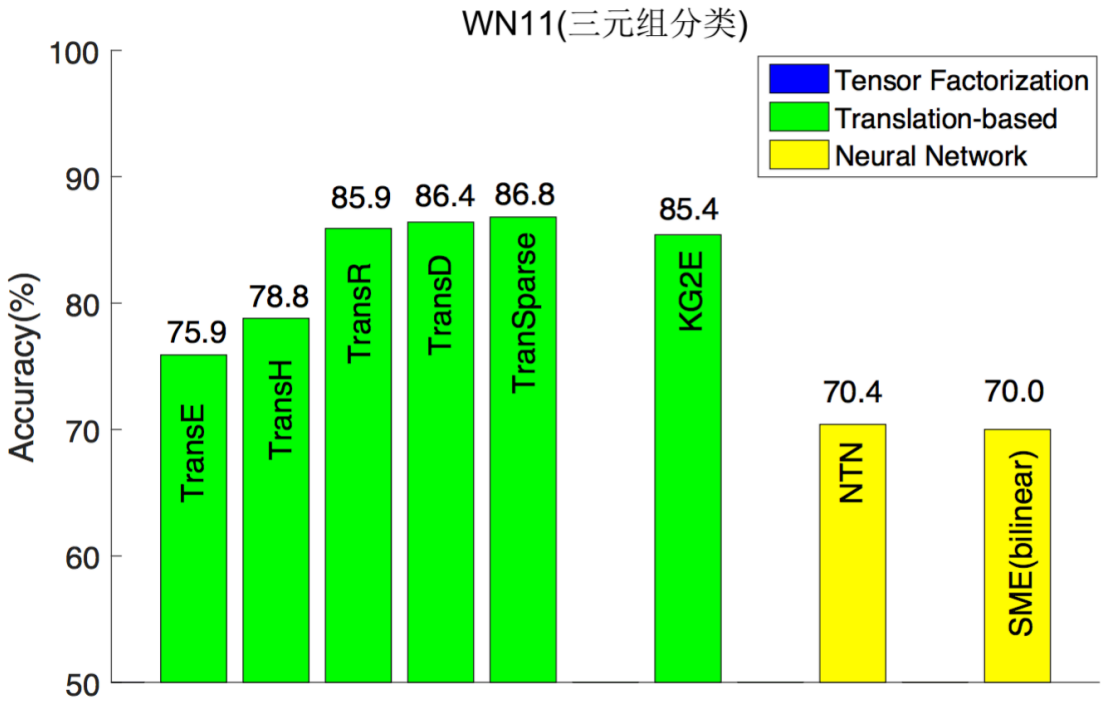
# 三元组分类


FB13(三元组分类）
Accuracy(%)
100
90
80
70
60
50
Tensor Factorization
Translation-based
Neural Network
81.5
TransE
83.3
TransH
82.5
TransR
89.1
TransD
88.2
TranSparse
85.3
KG2E
87.1
NTN
63.7
SME(bilinear)

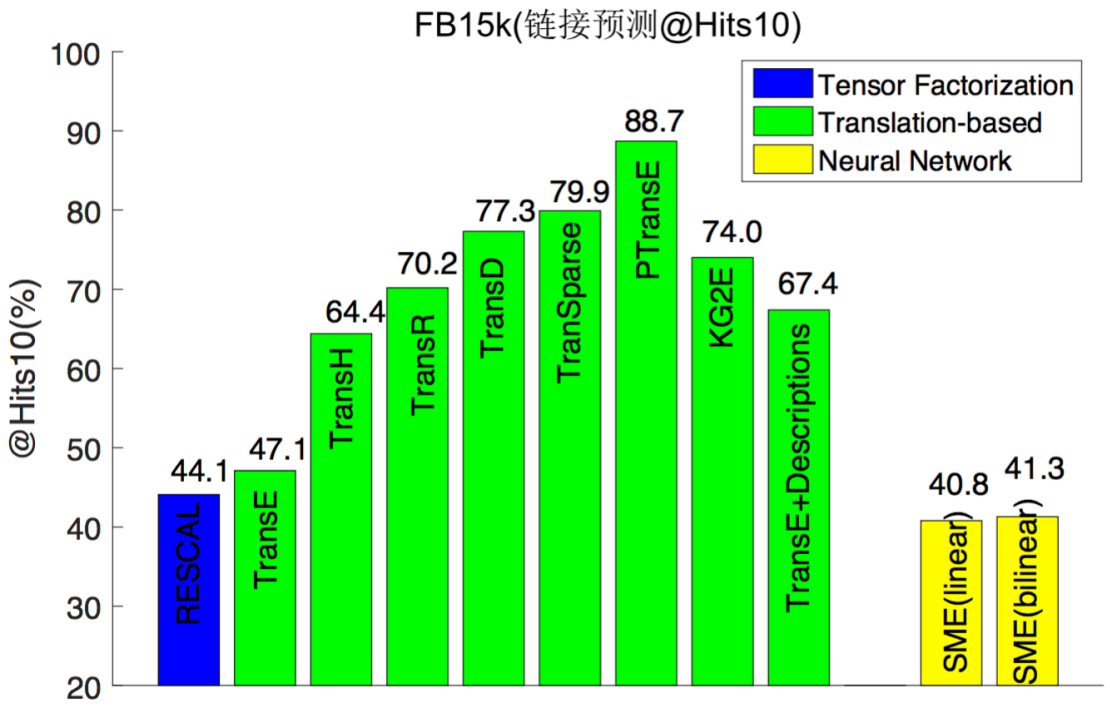
# 三元组分类


FB15k(三元组分类)
Tensor Factorization
Translation-based
Neural Network
Accuracy(%)
100
90
80
70
60
50
79.8
TransE
79.9
TransH
84.3
TransR
88.0
TransD
88.2
TranSparse
89.3
KG2E
68.2
NTN

# 三元组分类


WN11(三元组分类)
Tensor Factorization
Translation-based
Neural Network
Accuracy(%)
100
90
80
70
60
50
75.9
78.8
85.9
86.4
86.8
85.4
70.4
70.0
TransE
TransH
TransR
TransD
TranSparse
KG2E
NTN
SME(bilinear)

# 链接预测（知识库补全）


FB15k(链接预测@Hits10)
@Hits10(%)
100
90
80
70
60
50
40
30
20
Tensor Factorization
Translation-based
Neural Network
RESCAL
TransE
TransH
TransR
TransD
TranSparse
PTransE
KG2E
TransE+Descriptions
SME(linear)
SME(bilinear)
44.1
47.1
64.4
70.2
77.3
79.9
88.7
74.0
67.4
40.8
41.3

# 小结

- 当前知识图谱表示学习的三大类方法：
  - 基于重构的张量分解方法：通过矩阵分解获得关系、实体的表示
  - 基于翻译模型的方法
  - 基于神经网络的方法

# 总结

- 知识图谱是人工智能技术的基础部件
- 知识图谱的基本概念和可用的代表性知识图谱
- 知识图谱!==知识
- 基于符号的知识表示
  - 可解释、显式推理
  - 难以大规模、语义鸿沟
- 基于分布式的知识推理
  - 可学习、可计算，适合大规模、开放域
  - 不可解释、隐式推理

# Q&A

刘康 kliu@nlpr.ia.ac.cn
韩先培 xphan@iscas.ac.cn