知识图谱

孙栩
信息科学技术学院
xusun@pku.edu.cn

NLP如何解决人类语言问题?

知识图谱

自然语言理解

- 词法
- 词义
- 句法
- 语义
- 情感
- …

自然语言生成

- 翻译
- 基于数据生成
- 问答
- 对话
- …

知识图谱

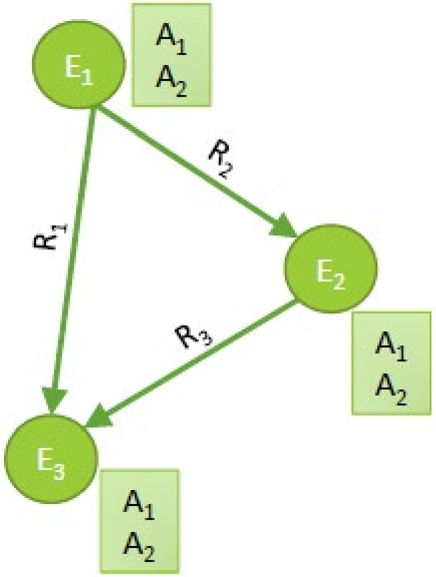
- **知识图谱**
  - 简介
- **知识抽取**
  - 基于NLP的IE
- **图谱构建**
  - 概率图模型、随机游走
  - 嵌入(embedding)方法
- **总结**

知识图谱 – 定义

- 图形式的知识

- 表示实体、属性和关系

- 实体：点
- 属性：点的标签
- 关系：带标签的边

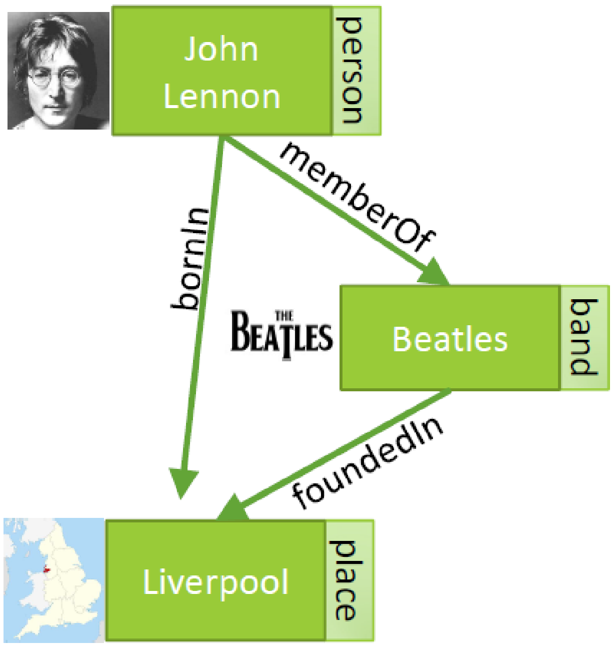


知识图谱 – 定义

- 图形式的知识

- 表示实体、属性和关系

- 实体：点
- 属性：点的标签
- 关系：带标签的边



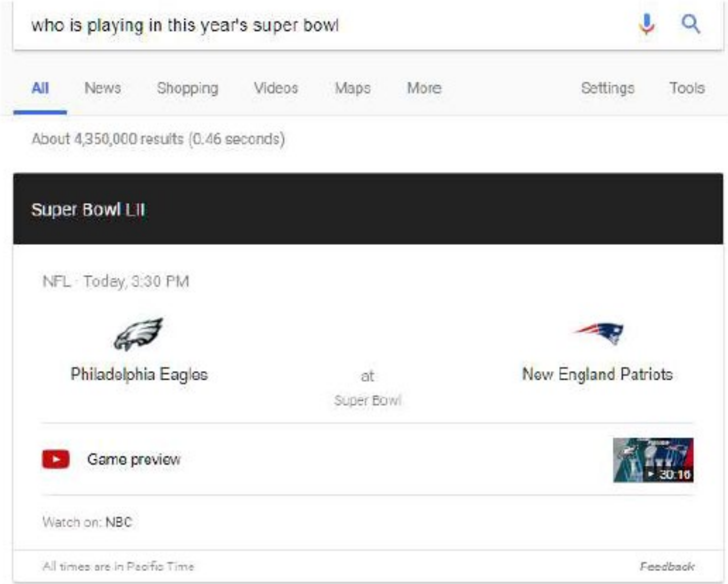
知识图谱 – 意义

- **对于人**
  - 高密度的信息流
  - 更直观的结构表示
  - 知识驱动任务的工具

- **对于人工智能**
  - 多种任务的关键组件
  - 数据与人类语义间的桥梁
  - 可以利用“图分析的大量成果”，比如图的关系的矩阵表示和运算

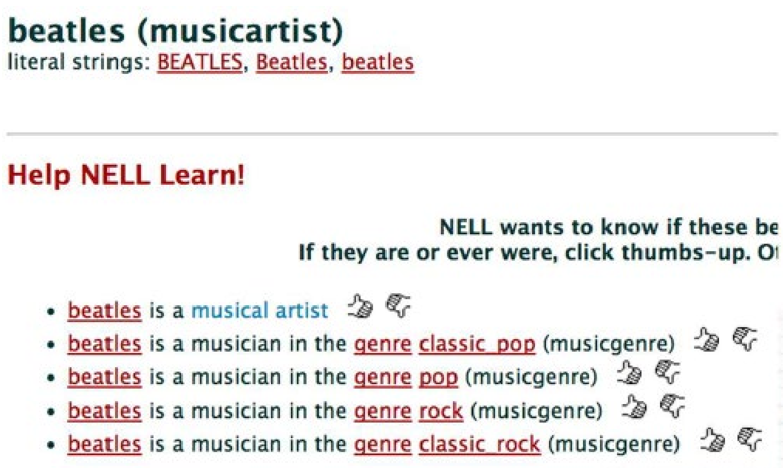
知识图谱 – 意义

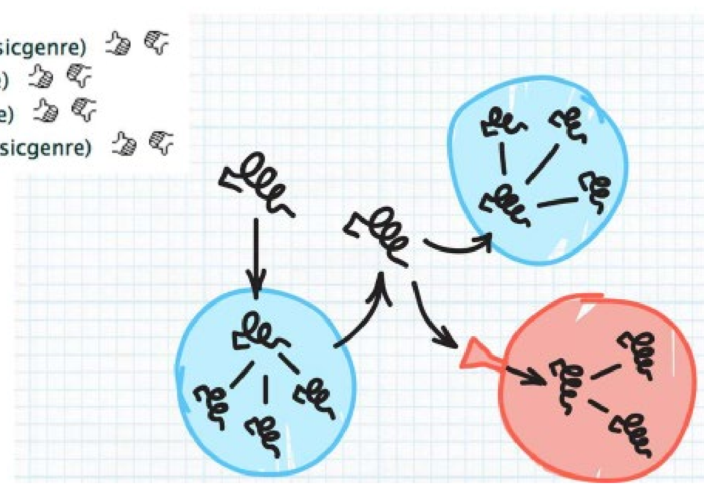
- 应用：问答/智能体



知识图谱 – 意义

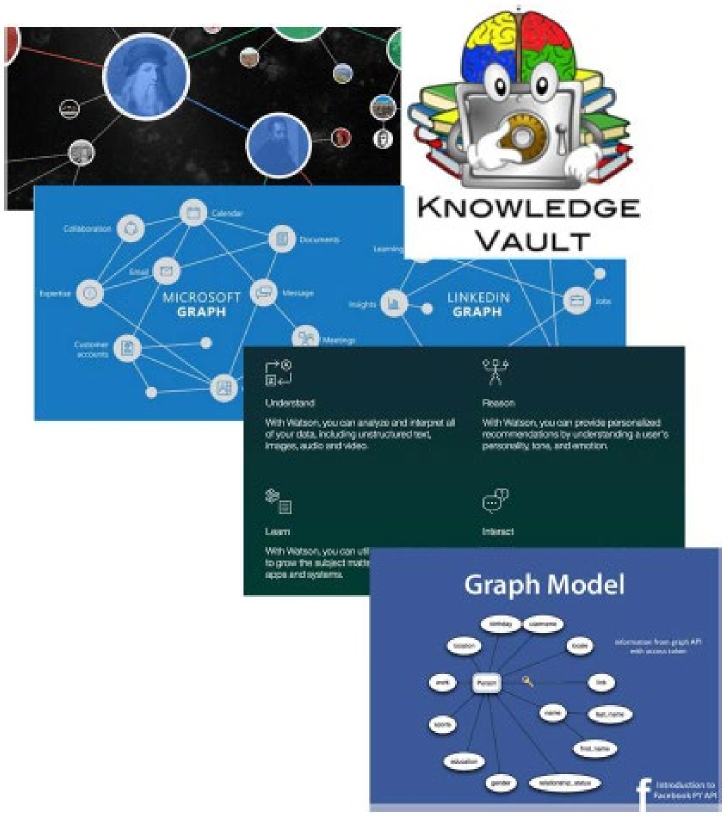
- 应用：辅助知识发现（通过知识图谱自动构建出一些语句，让人判断对不对，比如“药物合成”）





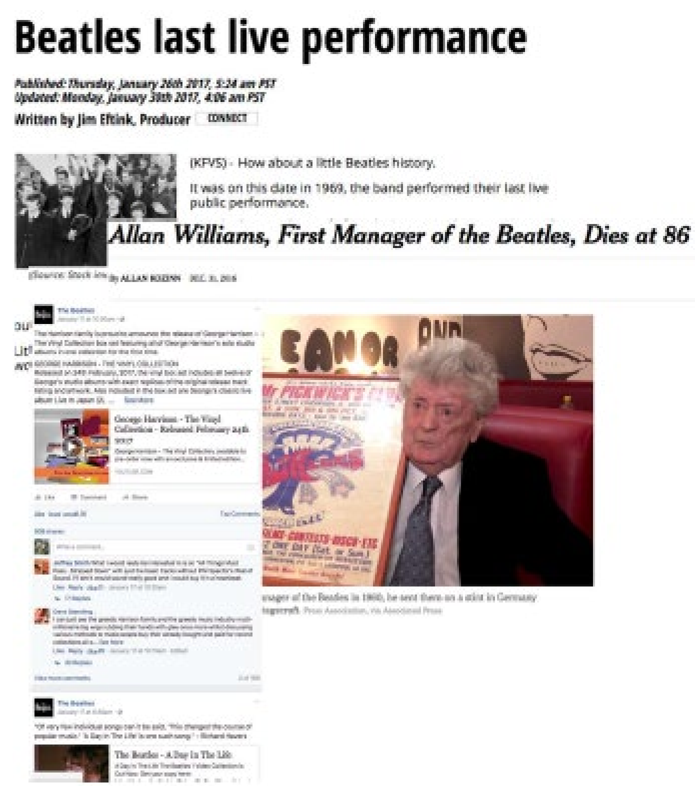
知识图谱 – 意义

- **很多大学、公司都在做这个方向**
  - 有很多产品推出
- **Google KG**
  - Knowledge vault
- **Amazon Product Graph**
- **Facebook Graph API**
- **IBM Watson**
- **Microsoft Satori**
  - Project Hanover/Literome
- **LinkedIn KG**
- **…**



知识图谱 – 来源

- 结构化文本
  - Wikipedia infoboxes
  - Tables, Databases
  - Social Nets
- 非结构化文本
  - 互联网、新闻、社交媒体
  - 右边例子是非结构化的文本

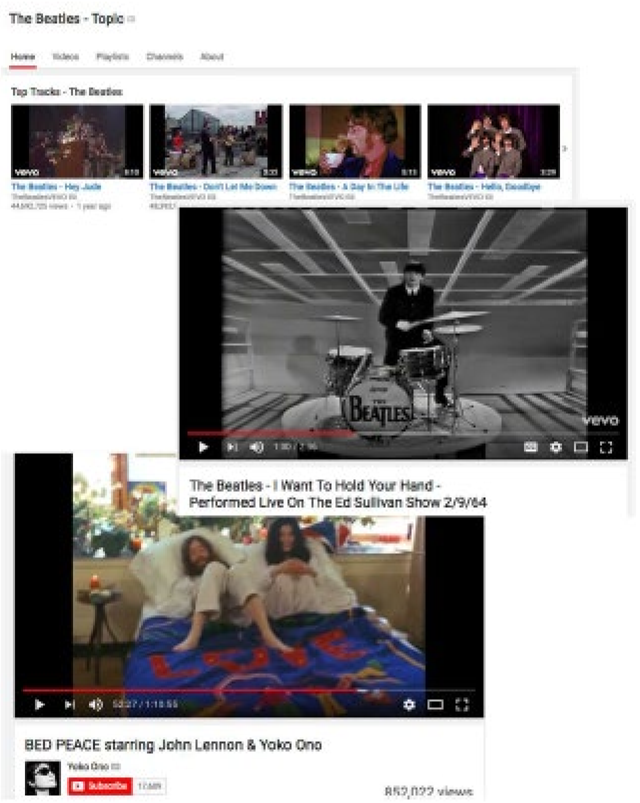


知识图谱 – 来源

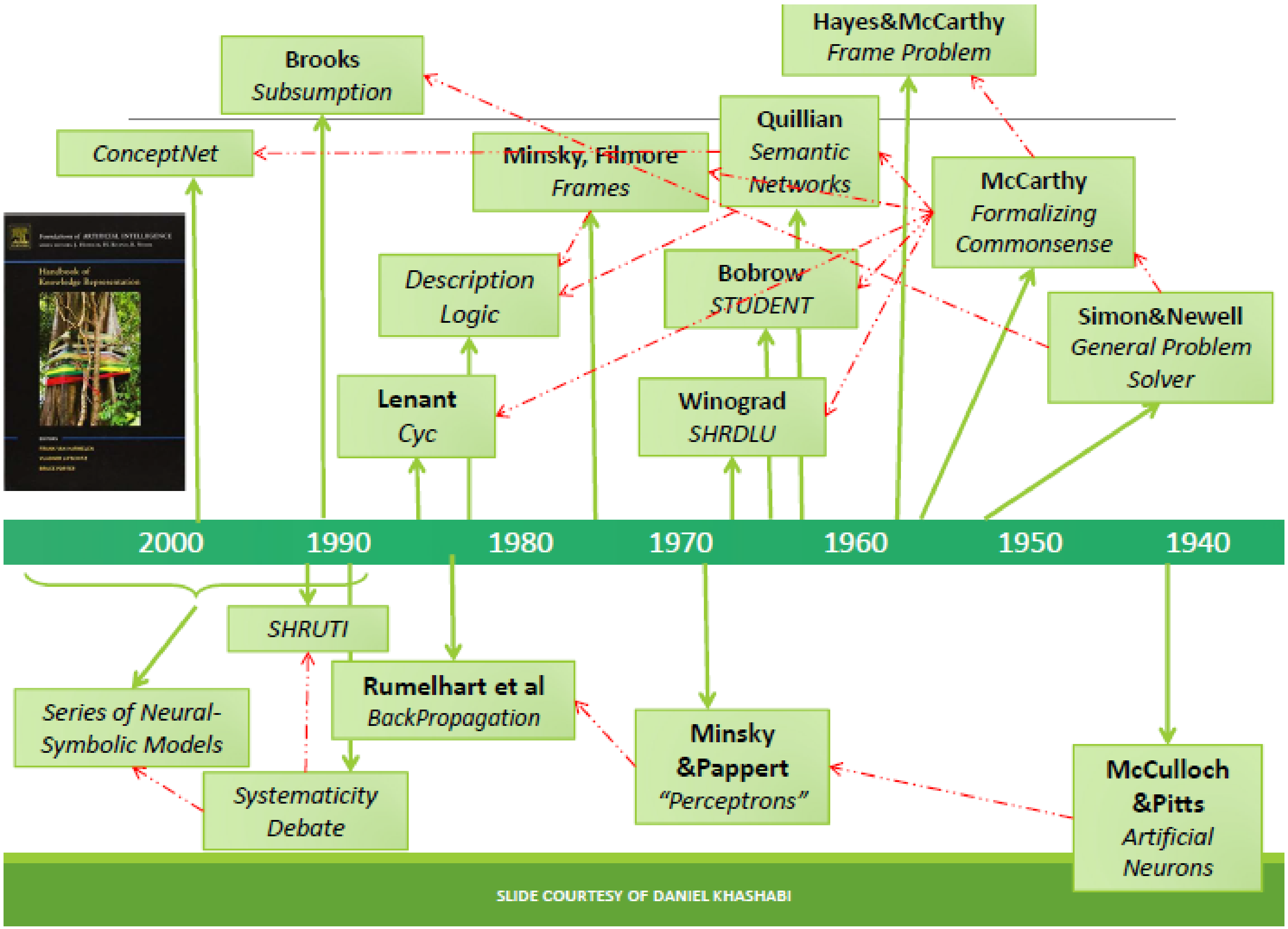
- **结构化文本**
  - Wikipedia infoboxes
  - Tables, Databases
  - Social Nets
- **非结构化文本**
  - 互联网、新闻、社交媒体
- **图片（理论上可以，实际上难度较大）**

知识图谱 – 来源

- **结构化文本**
  - Wikipedia infoboxes
  - Tables, Databases
  - Social Nets
- **非结构化文本**
  - 互联网、新闻、社交媒体
- **图片**
- **视频（更加困难）**



知识图谱 – 理论、系统的历史演化

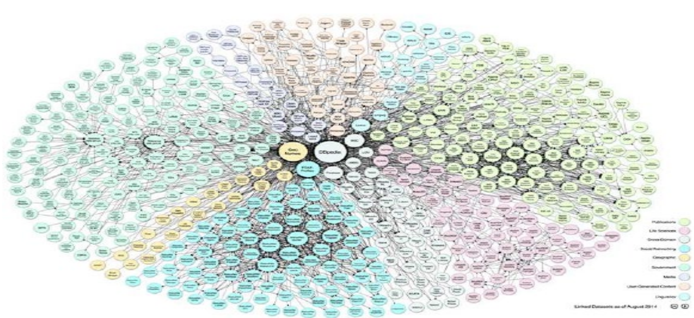


知识图谱 – 表示方法

- **数十年研究**
- **大多数知识图谱使用RDF(relation description framework)三元组（主，谓，宾）**
- **理想化的知识图谱基本上采用以下三原则：**
  - 1，Temporal scoping：时间范围原则：任何概念都是有时间范围的
  - Reification：具体化：一定要具体到具体的事物，比如obama应该表示为一个具体的全名…
  - Skolemization：形式化、仪式化、庄严化：比如RDF的三元组是个典型例子
- **ABox (assertions) vs Tbox (terminology)**
  - 知识是用自然语言的声明？还是专业术语来声明？比如freeBase是基于自然语言的(is.located.in)，conceptNet/Yago是基于术语的（关系是严格人进行定义的，相当于对关系进行了过滤）

知识图谱 – 跟语义网的区别

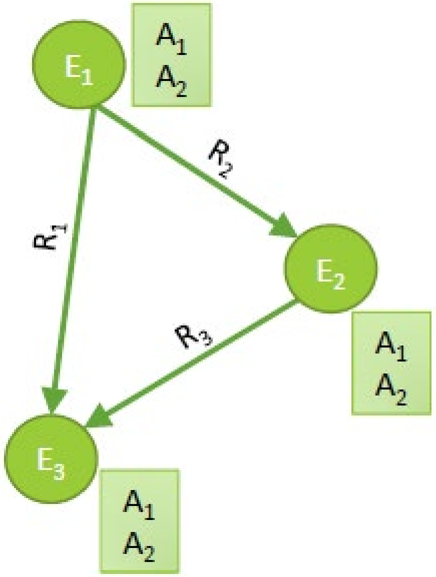
- **语义网 (semantic web)**
  - 定义和交换知识的标准
  - 跟知识图谱是相交的关系，有的知识图谱用到了语义标准，如RDF（通过语义网把dbPedia和Yago联系在一起）
- **语义网想表示整个世界的所有知识，起初是想把非结构化/非语义话的WWW数据表示为结构化/语义化的数据，构建语义世界，想做成WWW一样的Universal的系统**
- **后来遇到障碍，但是目前也有一定的用途，比如它把所有不同的知识图谱融合在一起，定义了知识图谱之间的信息交换标准**



知识图谱 – 研究内容

给了大规模的数据，确定以下3点：

- **哪些是实体**
  - 节点
- **什么是属性**
  - 标签
- **关系如何**
  - 边



两个步骤

步骤1：知识抽取

- **实体**
  - NER
  - 通过共指消解确定实体
- **属性**
  - 通过NER确定的具体类别：比如人、地名
- **关系**
  - 关系抽取RE
  - 语义角色标注SRL

步骤2：图谱构建（已经有了部分的图谱，怎么进行补全/概念合并）

- **怎么补全“点”**
  - Entity linking用于点的合并
- **标签**
  - Collective classification：通过这点周围的环境来确认其标签
- **边**
  - Link prediction用于补全边

- **知识图谱**
  - 简介
- **知识抽取**
  - 基于NLP的IE
- **图谱构建**
  - 概率图模型、随机游走
  - 嵌入(embedding)方法
- **总结**

知识抽取

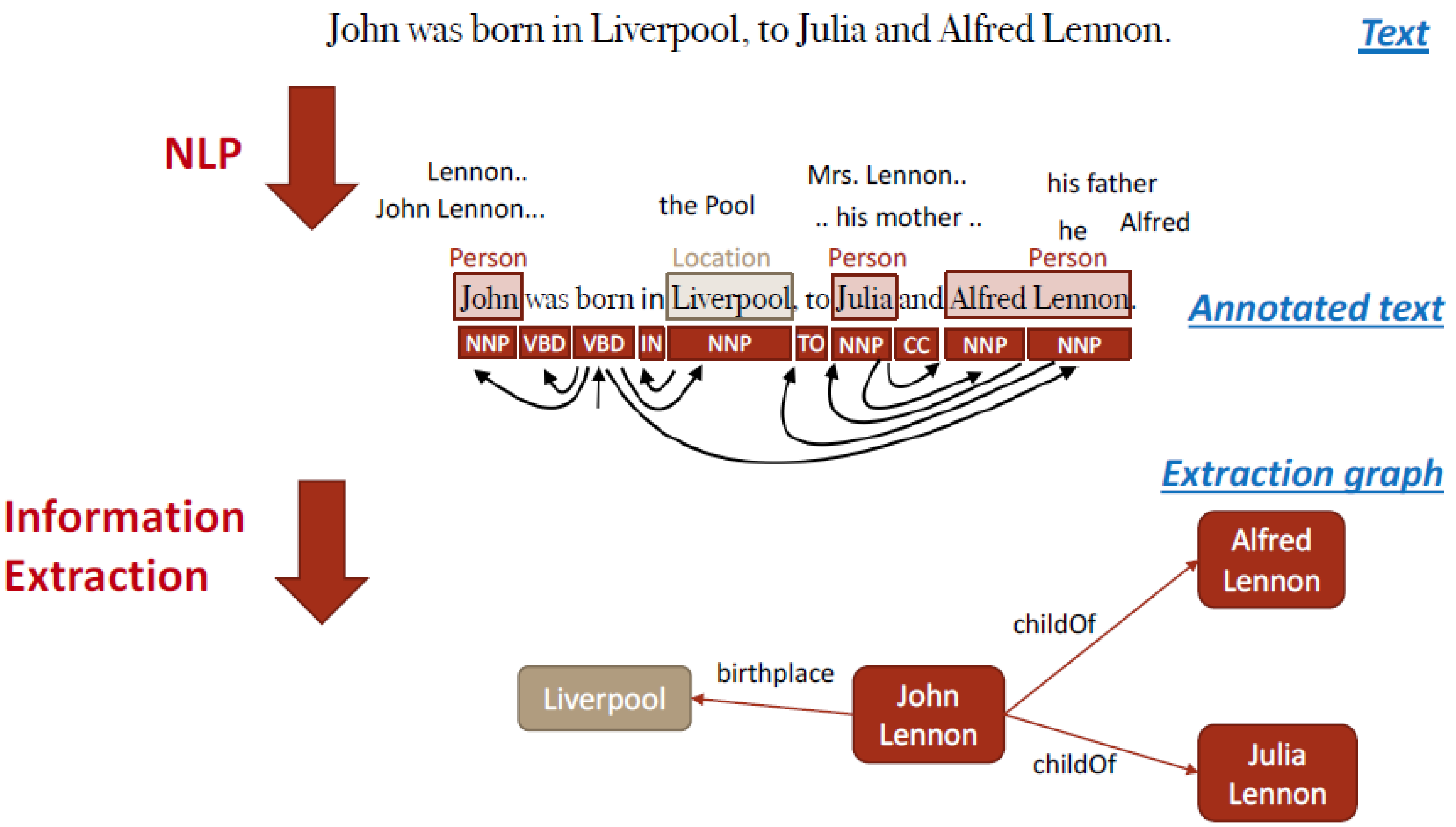
Unstructured
Ambiguous
Lots and lots of it!

Structured
Precise, Actionable
Specific to the task

Humans can read them, but
... very slowly
... can't remember all
... can't answer questions

Can be used for downstream applications, such as creating Knowledge Graphs!

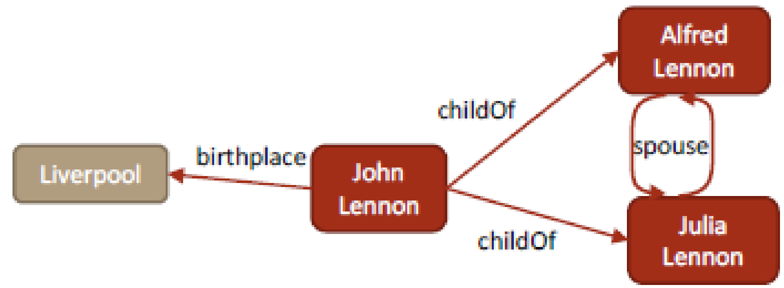
知识抽取


John was born in Liverpool, to Julia and Alfred Lennon.
Text
NLP
Lennon..
John Lennon...
the Pool
Mrs. Lennon..
.. his mother ..
his father
he
Alfred
Person
Location
Person
Person
John was born in Liverpool, to Julia and Alfred Lennon.
NNP VBD VBD IN NNP TO NNP CC NNP NNP
Annotated text
Information Extraction
Extraction graph
Alfred Lennon
childOf
birthplace
Liverpool
John Lennon
childOf
Julia Lennon

知识抽取 - 分解

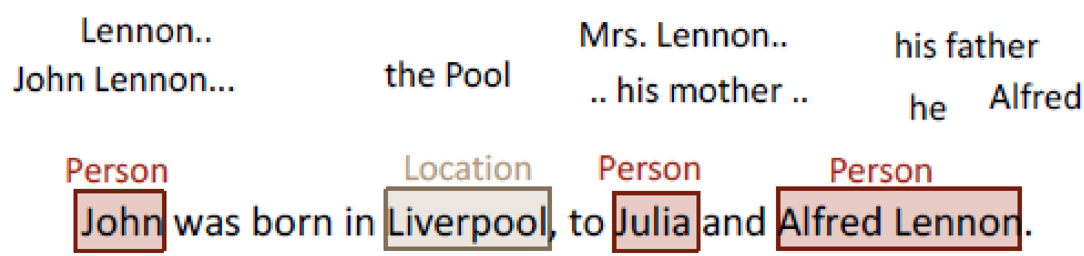
Information Extraction

Entity resolution,
Entity linking,
Relation extraction…



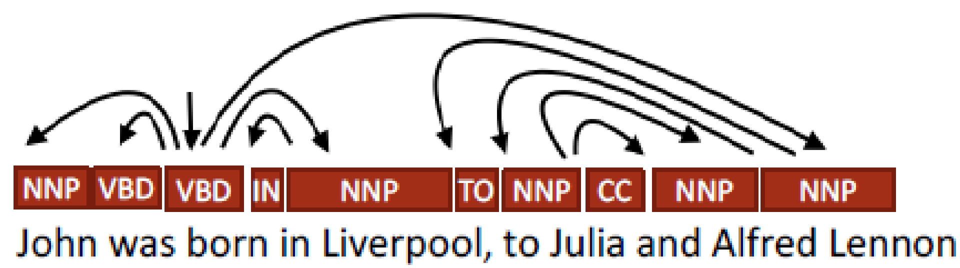
Document

Coreference Resolution…



Sentence

Dependency Parsing,
Part of speech tagging,
Named entity recognition…



知识抽取 – Entity Resolution & Linking

如何确定一个实体?

…during the late 60's and early 70's, **Kevin Smith** worked with several local…

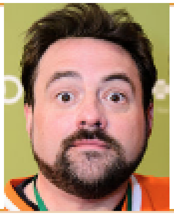
…the term hip-hop is attributed to **Lovebug Starski**. What does it actually mean…

Like Back in 2008, the Lions drafted **Kevin Smith**, even though Smith was badly…

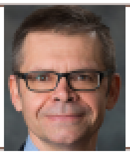
… backfield in the wake of **Kevin Smith**'s knee injury, and the addition of Haynesworth…

The filmmaker **Kevin Smith** returns to the role of Silent Bob…

Nothing could be more irrelevant to **Kevin Smith**'s audacious ''Dogma'' than ticking off…



… The Physiological Basis of Politics," by **Kevin Smith**, Douglas Oxley, Matthew Hibbing…



知识抽取 – Entity Resolution & Linking

同名实体

- 同类型实体
  - Kevin Smith
- 互相命名
  - Clinton, Amazon
- 部分引用
  - First names
  - Location instead of full name

异名实体

- 昵称
  - Drumpf
- 错字、笔误
  - Baarak, Barak, Barrack
- 不一致的引用
  - MSFT, APPL, GOOG

知识抽取 – Entity Resolution & Linking

Washington drops 10 points after game with UCLA Bruins.

Candidate Generation

Washington DC, George Washington, Washington state, Lake Washington, Washington Huskies, Denzel Washington, University of Washington, Washington High School, ...

Entity Types LOC/ORG

Washington DC, ~~George Washington~~, Washington state, Lake Washington, Washington Huskies, ~~Denzel Washington~~, University of Washington, Washington High School, ...

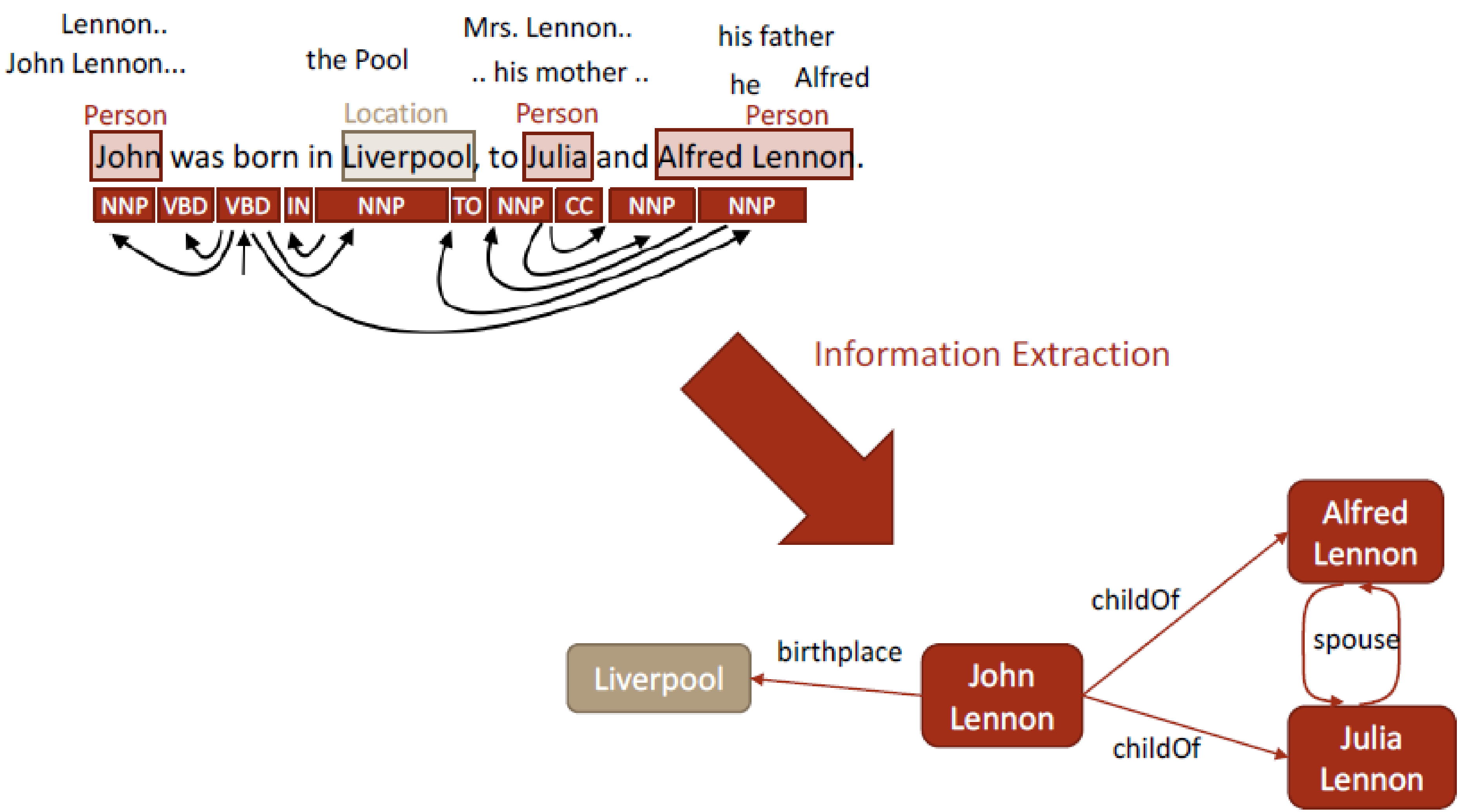
Coreference UWashington, Huskies

~~Washington DC~~, ~~George Washington~~, ~~Washington state~~, ~~Lake Washington~~, Washington Huskies, ~~Denzel Washington~~, University of Washington, ~~Washington High School~~, ...

Coherence UCLA Bruins, USC Trojans

~~Washington DC~~, ~~George Washington~~, ~~Washington state~~, ~~Lake Washington~~, Washington Huskies, ~~Denzel Washington~~, ~~University of Washington~~, ~~Washington High School~~, ...

知识抽取 – Information Extraction



知识抽取 – Information Extraction

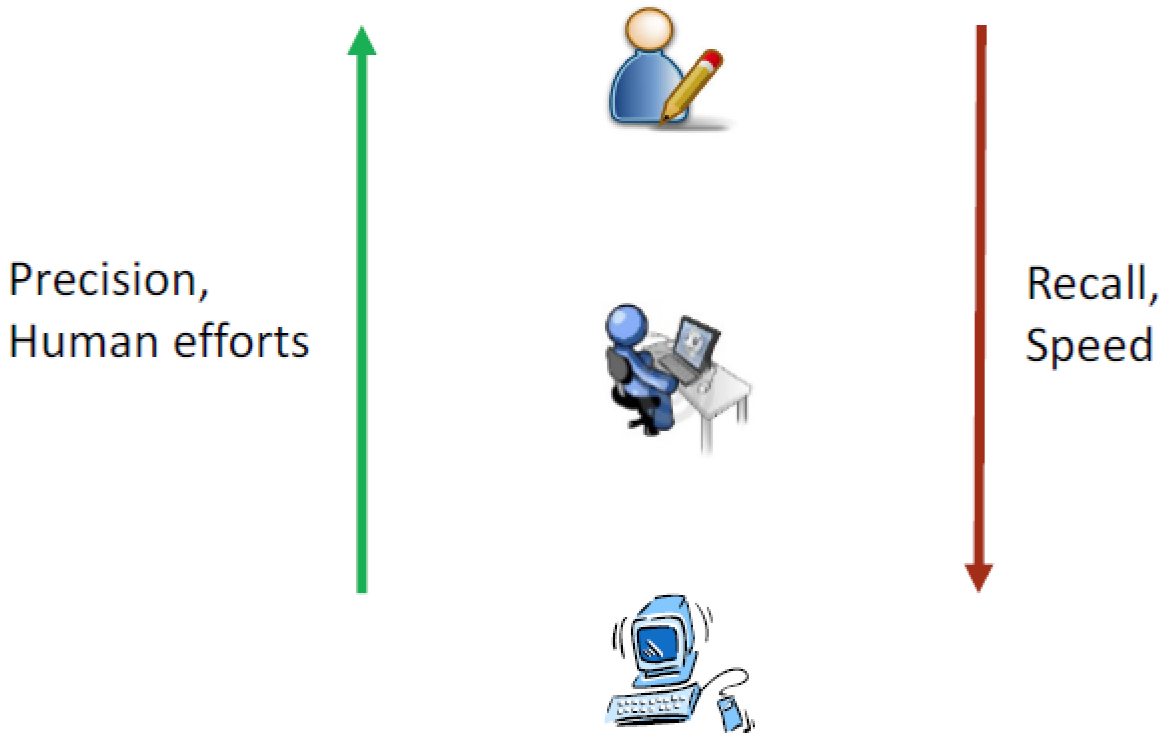
三个子问题

- 确定领域
- 学习抽取
- 事实评分

三类方法

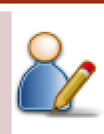
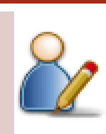
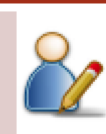
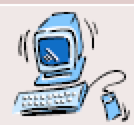
- 监督
- 半监督
- 无监督

知识抽取 – Information Extraction



知识抽取 – Information Extraction

- 实际中

	Defining domain	Learning extractors	Scoring candidate facts	Fusing extractors
ConceptNet				
NELL				Heuristic rules
Knowledge Vault				Classifier
OpenIE				

知识图谱

- **知识图谱**
  - 简介
- **知识抽取**
  - 基于NLP的IE
- **图谱构建**
  - 概率图模型、随机游走
  - 嵌入(embedding)方法
- **总结**

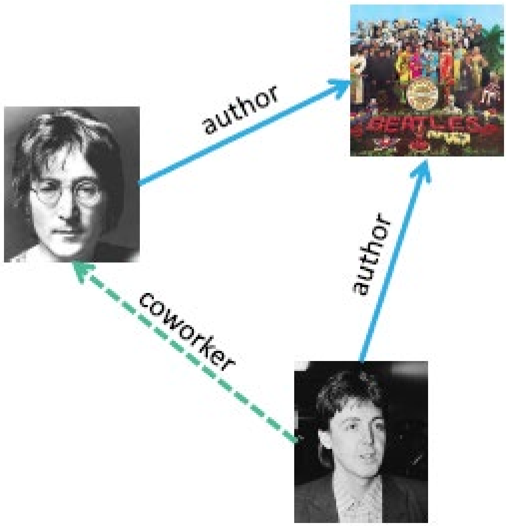
图谱构建

- **主要问题：抽取的图谱**

- **歧义**

图谱构建

- 主要问题：抽取的图谱
- 歧义
- 不完整



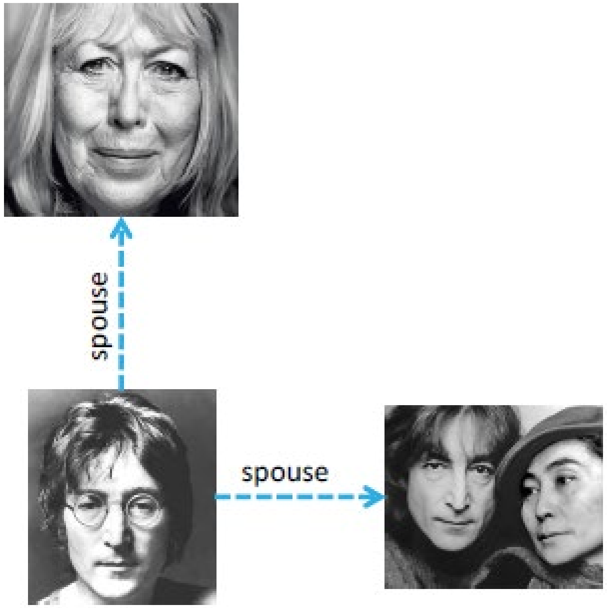
图谱构建

- 主要问题：抽取的图谱

- 歧义

- 不完整

- 不一致



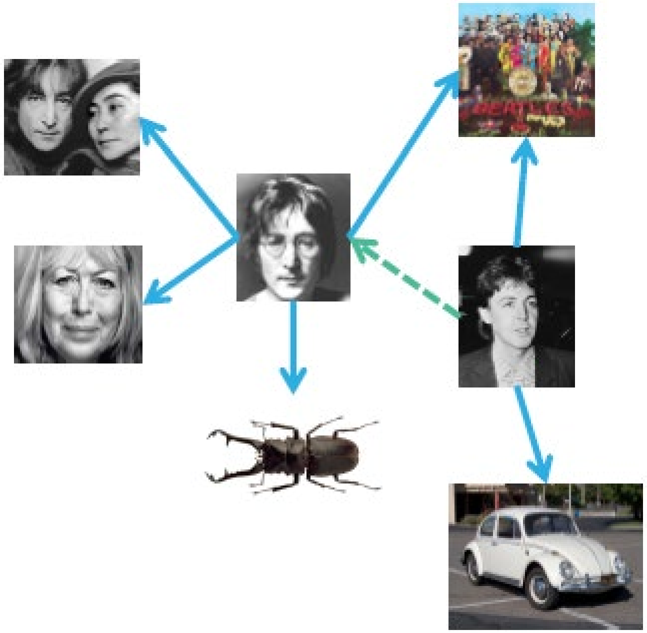
图谱构建

- 主要问题：抽取的图谱

- 歧义

- 不完整

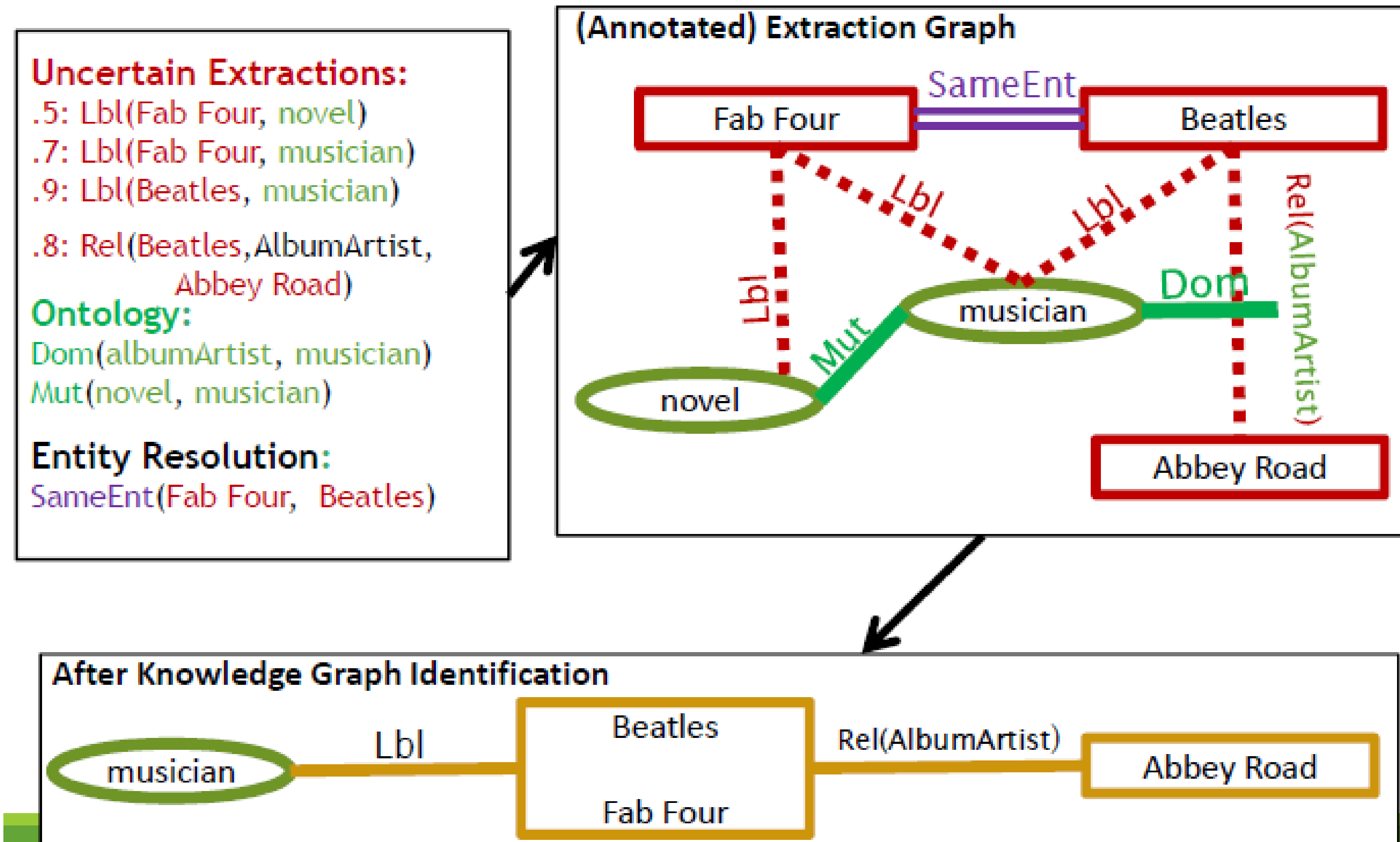
- 不一致



图谱构建

- **目标**
  - 过滤清理
  - 补充完全
- **融合本体约束和关系模式**
- **发掘统计规律**
- **方法：**
  - 概率模型：概率图模型、随机游走
  - 嵌入模型

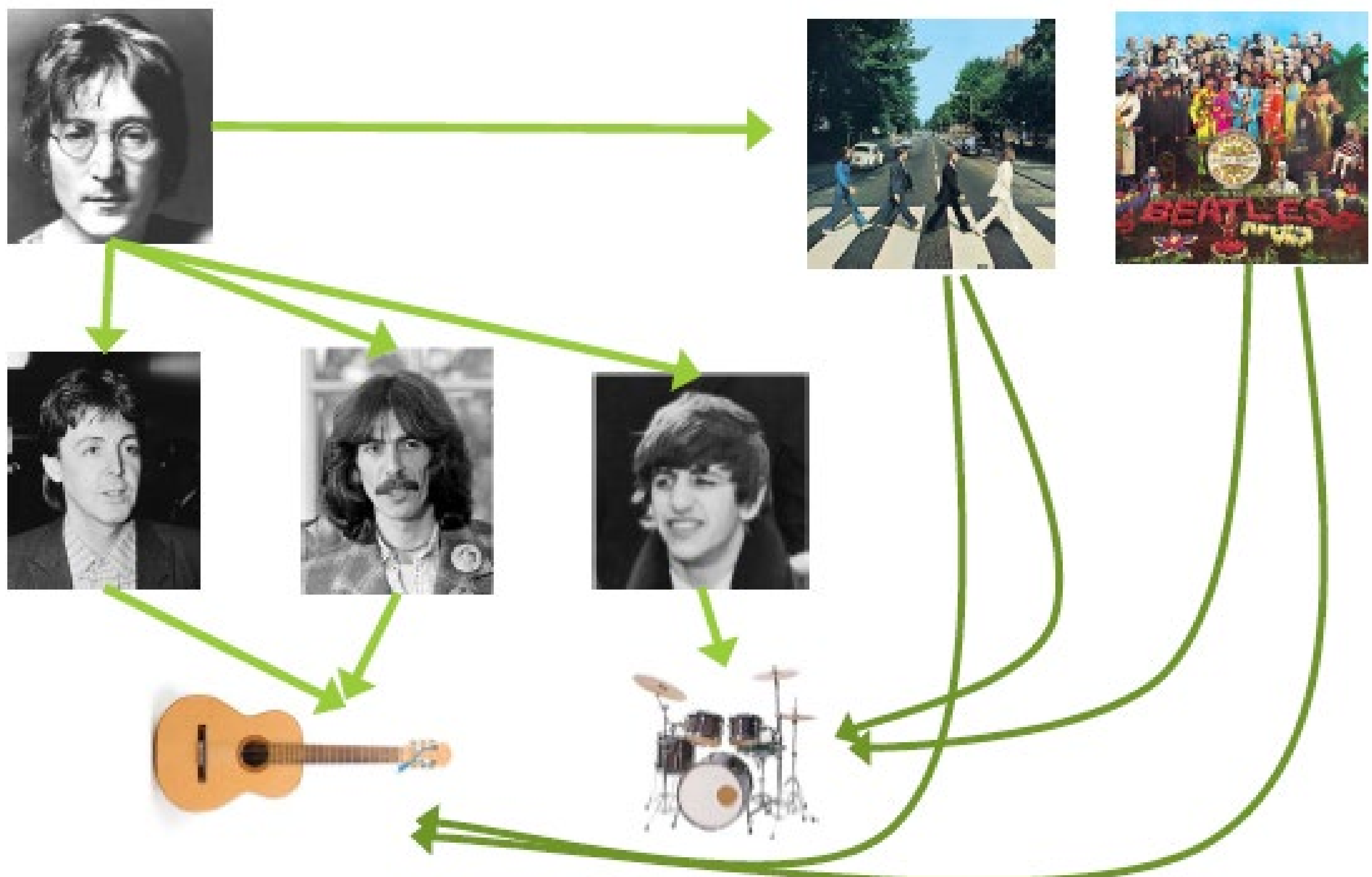
图谱构建 – 概率图模型



PUJARA+ISWC13; PUJARA+AIMAG15

图谱构建 – 随机游走

Query: R(Lennon, PlaysInstrument, ?)



图谱构建 – 嵌入模型

概率模型的限制

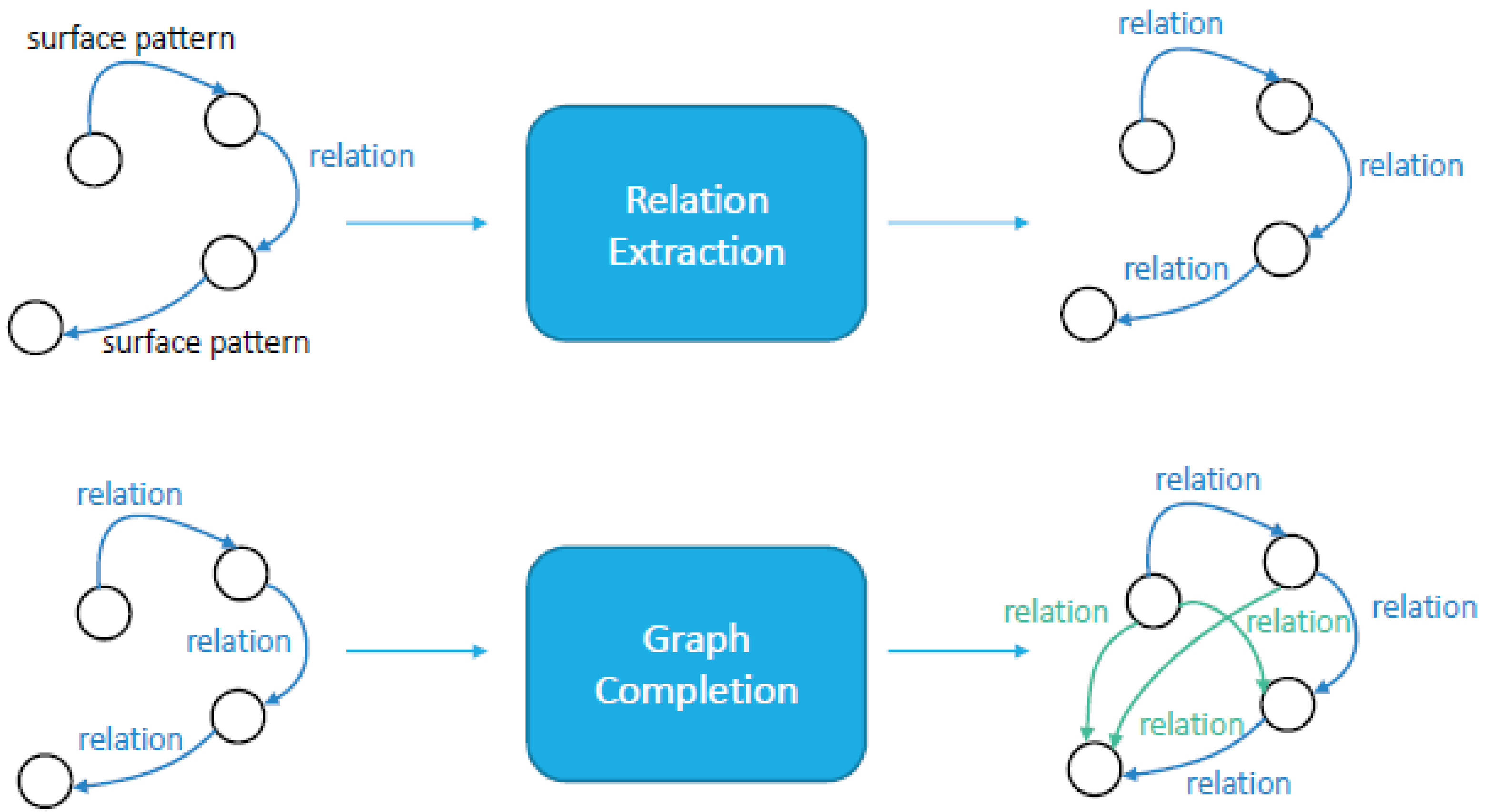
- **使用逻辑关系**
  - 人工设计
  - 泛化困难
- **复杂度**
  - 规则越多约复杂
  - 经常NP-Hard
  - 查询有时是NP-Hard
  - 难以并行

嵌入模型

- **使用分布式语义**
  - 一切皆为向量
  - 捕捉多种关系
  - 数据中习得
- **复杂度**
  - 依赖于隐含维度
  - SGD
  - 查询代价低
  - GPU并行

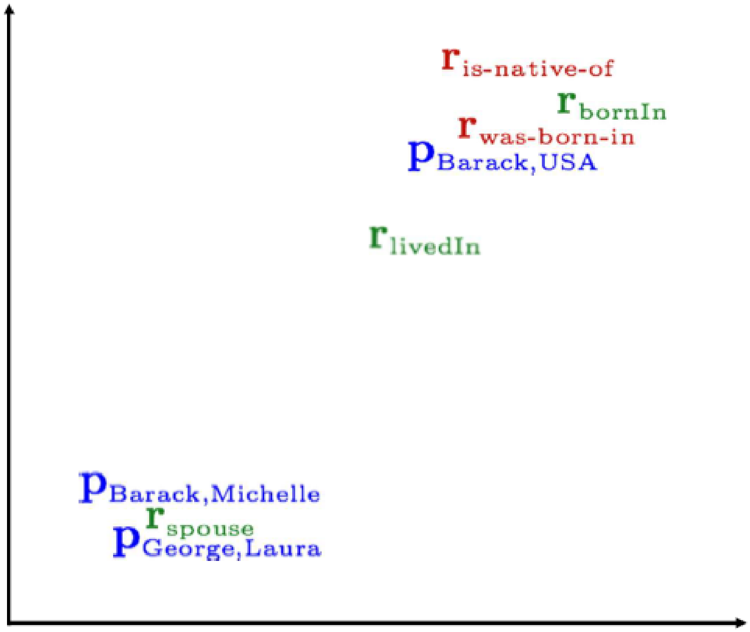
图谱构建 – 嵌入模型

相关任务



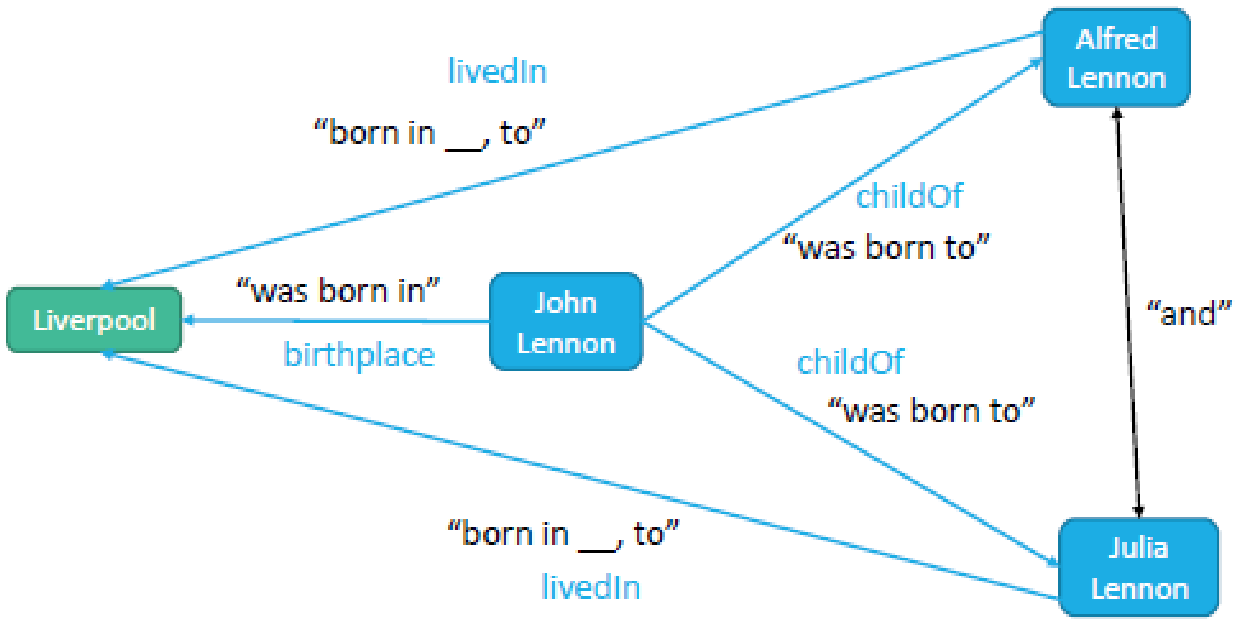
图谱构建 – 关系抽取

- relation embedding
- 相似关系、相似向量
  - 关系过滤、消解
- 推断
  - 关系补全、预测



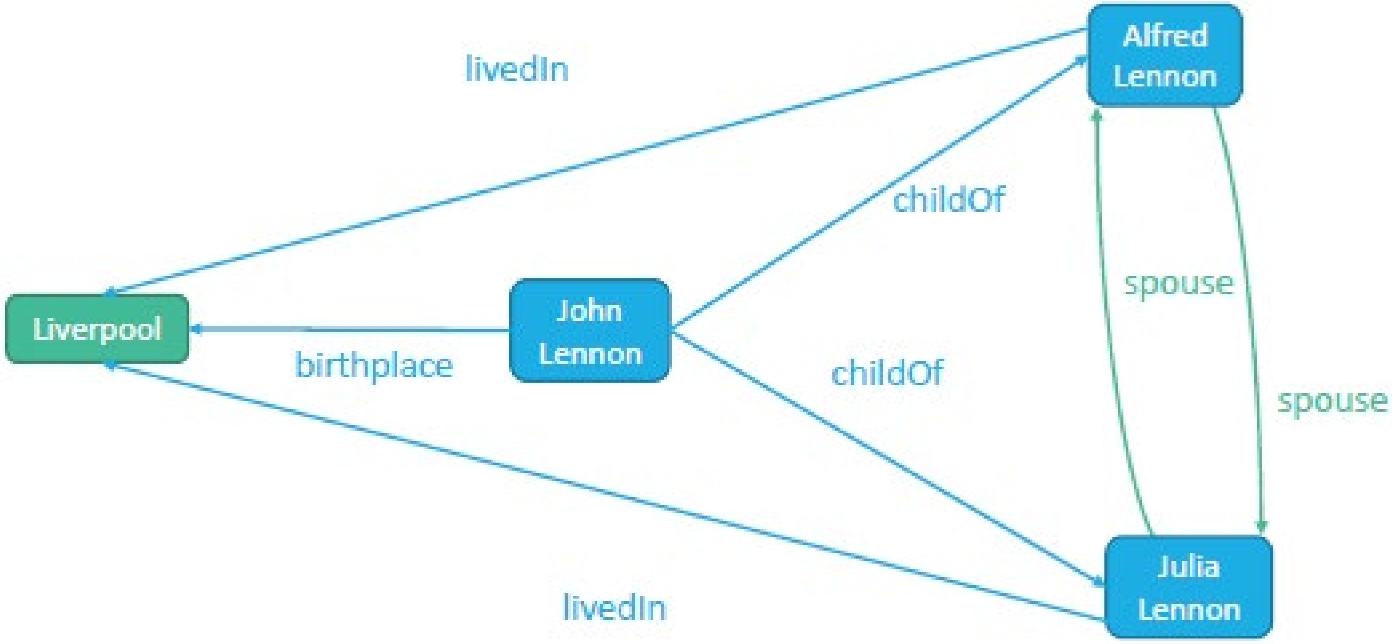
图谱构建 – 图谱补全

已有信息



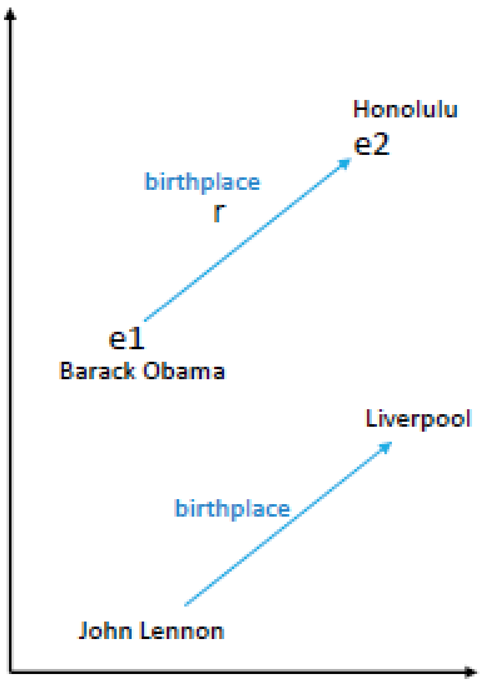
图谱构建 – 图谱补全

补全结果



图谱构建 – 图谱补全

张量建模: 翻译/平移向量



TransE

$$S\left(r(a,b)\right) = -\|\mathbf{e}_a + \mathbf{R}_r - \mathbf{e}_b\|_2^2$$

TransH

$$S\left(r(a,b)\right) = -\|\mathbf{e}_a^{\perp} + \mathbf{R}_r - \mathbf{e}_b^{\perp}\|_2^2$$

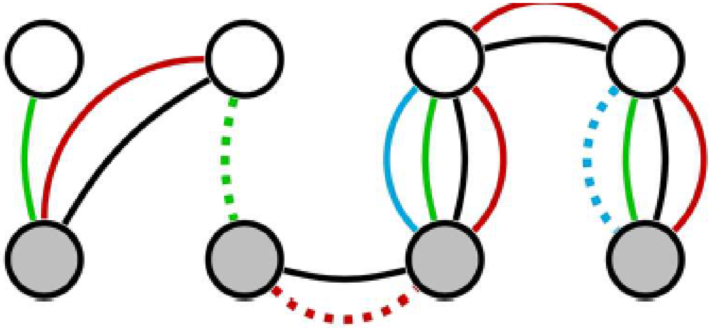
$$\mathbf{e}_a^{\perp} = \mathbf{e}_a - \mathbf{w}_r^T \mathbf{e}_a \mathbf{w}_r$$

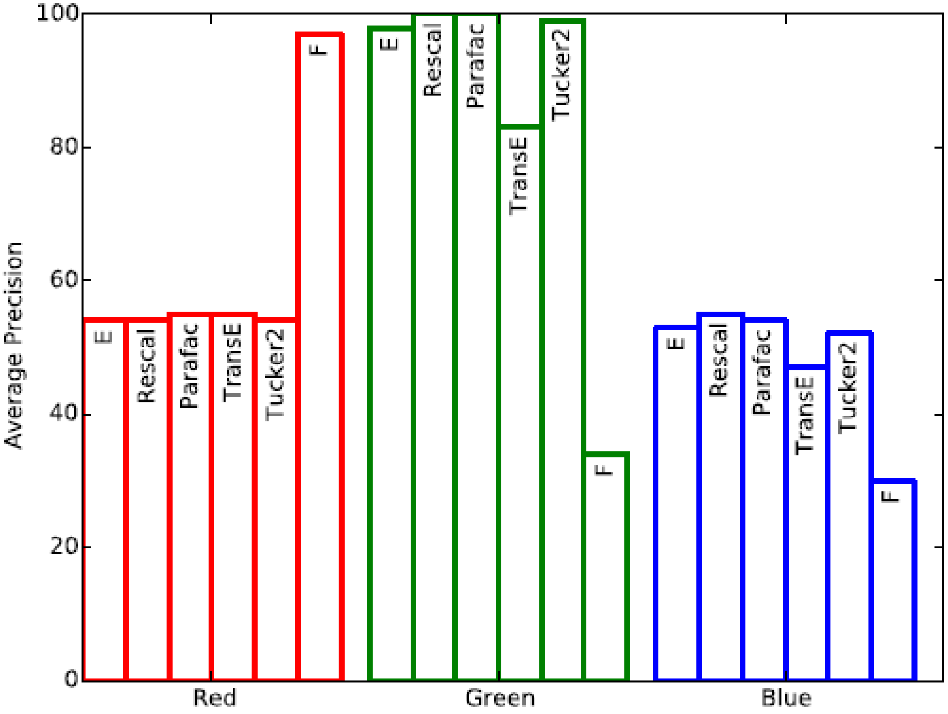
TransR

$$S\left(r(a,b)\right) = -\|\mathbf{e}_a \mathbf{M}_r + \mathbf{R}_r - \mathbf{e}_b \mathbf{M}_r\|_2^2$$

图谱构建 – 图谱补全

局限性





- **Red**: deterministically implied by **Black**
  - needs *pair-specific* embedding
  - Only **F** is able to generalize
- **Green**: needs to estimate entity types
  - needs *entity-specific* embedding
  - Tensor factorization generalizes, **F** doesn't
- **Blue**: implied by **Red** and **Green**
  - Nothing works much better than random

总结

□ 成功案例

Open Information Extraction

NELL Knowledge Base Browser
CMU Read the Web Project

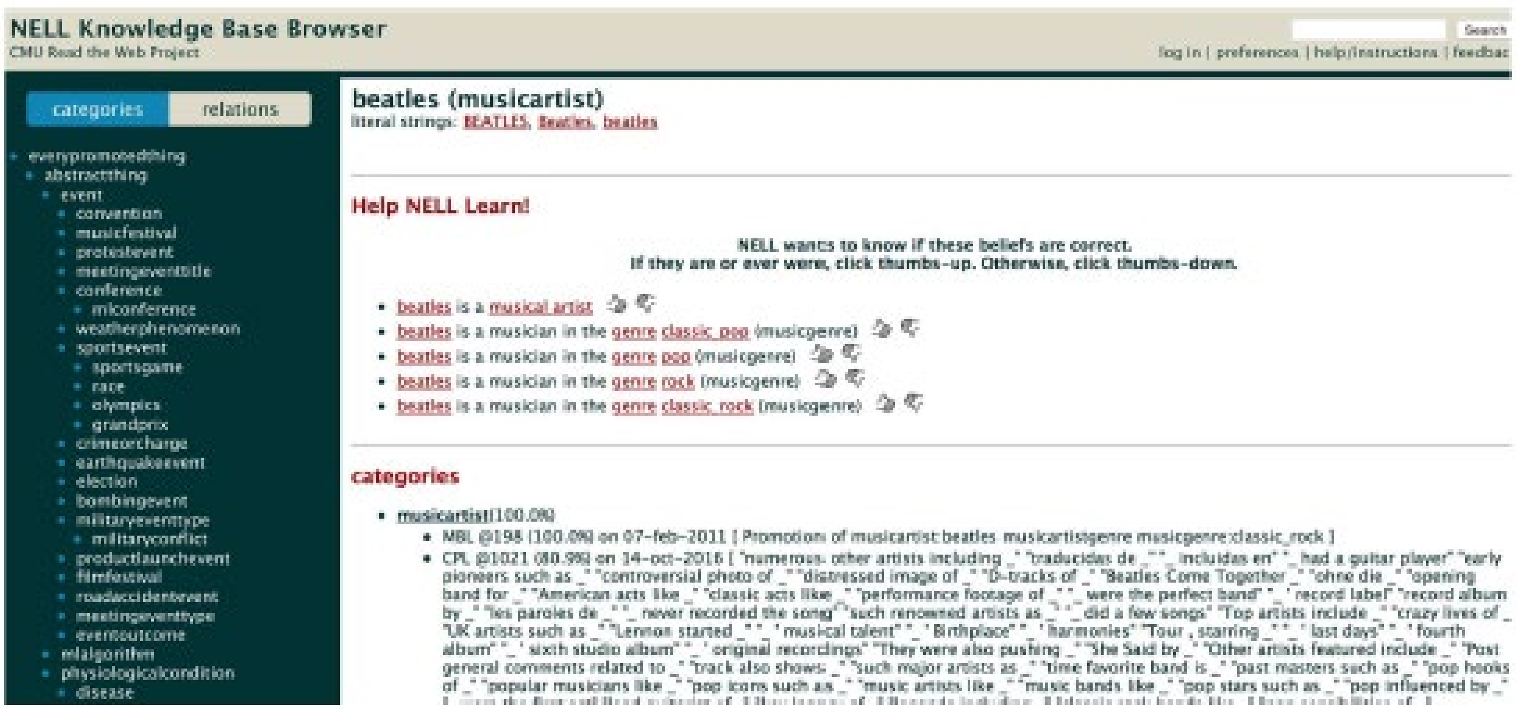
ConceptNet
An open, multilingual knowledge graph

YAGO

DeepDive v0.8.0
Think about **features**, not **algorithms**.

总结

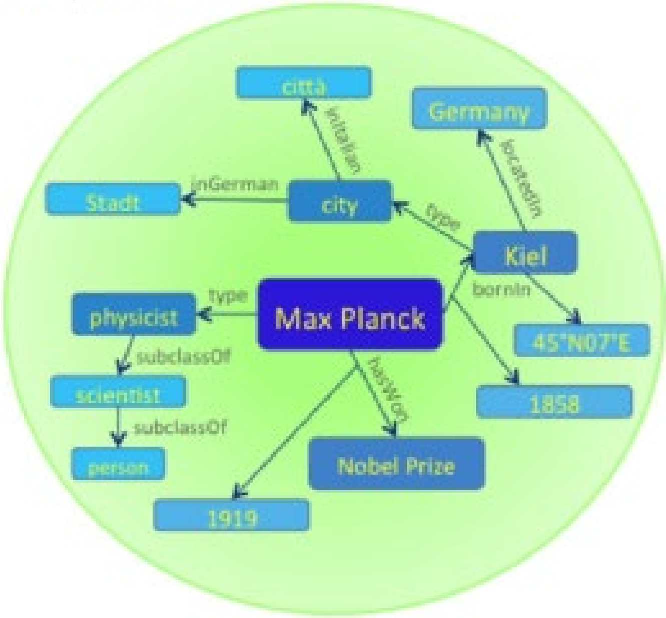
成功案例 - NELL


NELL Knowledge Base Browser
CMU Read the Web Project
Search
log in | preferences | help/instructions | feedbac
categories
relations
beatles (musicartist)
literal strings: BEATLES, Beatles, beatles
Help NELL Learn!
NELL wants to know if these beliefs are correct.
If they are or ever were, click thumbs-up. Otherwise, click thumbs-down.
beatles is a musical artist
beatles is a musician in the genre classic pop (musicgenre)
beatles is a musician in the genre pop (musicgenre)
beatles is a musician in the genre rock (musicgenre)
beatles is a musician in the genre classic rock (musicgenre)
categories
musicartist (100.0%)
everypromotedthing
abstractthing
event
convention
musicfestival
protestevent
meetingeventtitle
conference
mlconference
weatherphenomenon
sportsevent
sportsgame
race
olympics
grandprix
crimeorcharge
earthquakeevent
election
bombingevent
militaryeventtype
militaryconflict
productlaunchevent
filmfestival
roadaccidentevent
meetingeventtype
eventoutcome
mlalgorithm
physiologicalcondition
disease

总结

成功案例 - YAGO

- **Input:** Wikipedia infoboxes, WordNet and GeoNames
- **Output:** KG with 350K entity types, 10M entities, 120M facts
- Temporal and spatial information



总结

◻ 成功案例 - ConceptNet

Thanks