

# 《知识图谱: 概念与技术》

## 第 7 讲 知识图谱质量控制

---

李直旭  
苏州大学

# 背景 - 数据质量问题无处不在

例如

关系名: 学生

字段: 学号, 姓名, 性别, 专业代号

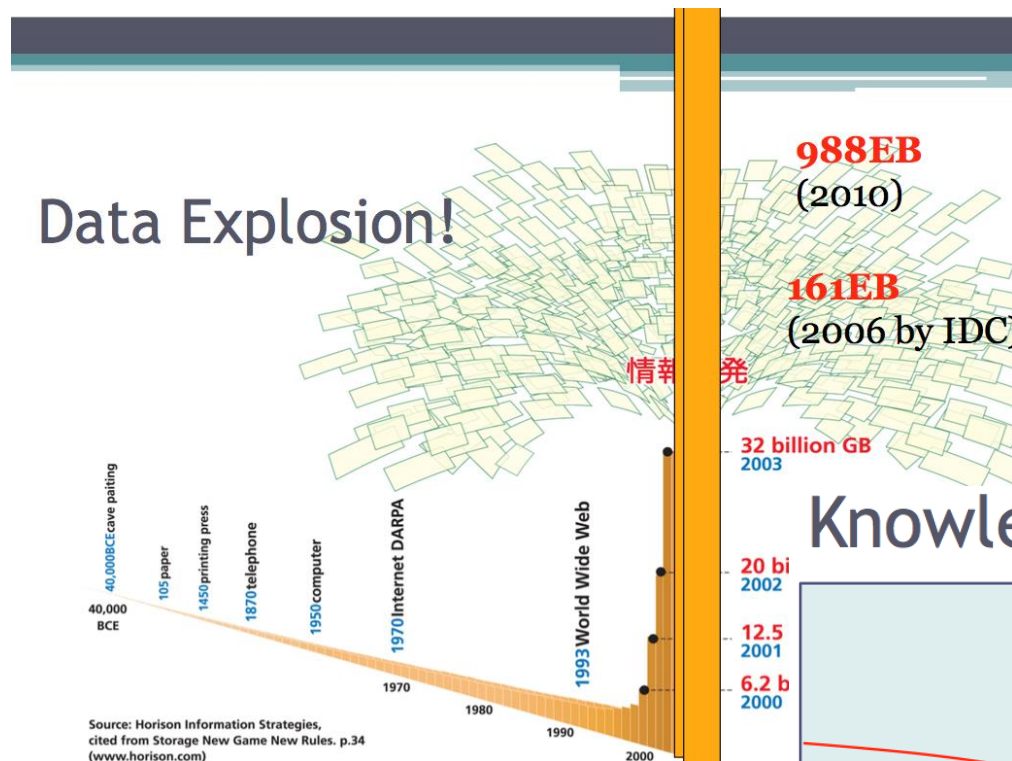
关系 (二维表)

学号	姓名	性别	专业代号
990101	章三	男	102001
990102	李辉	男	102001
990103	黄化	女	102002

记录

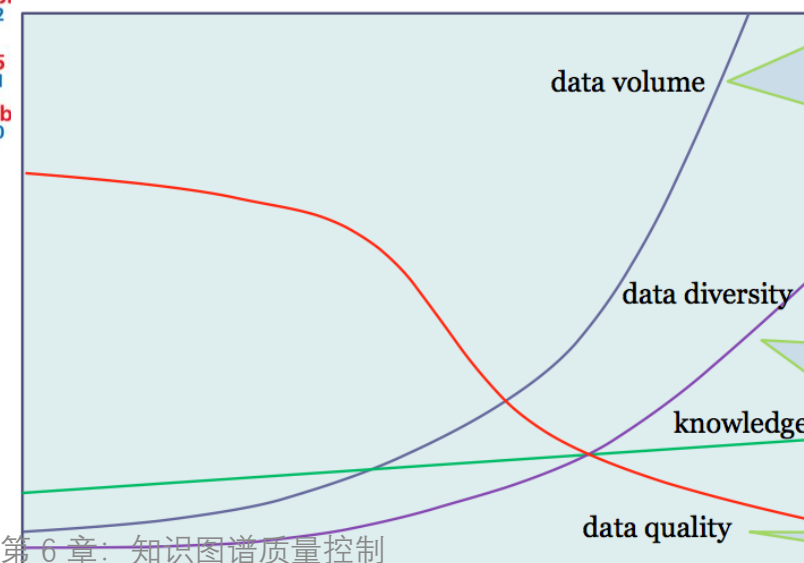


# 背景 - 甚至日趋严重



**数据量剧增  
伴随数据质量剧降**

## Knowledge Explosion?

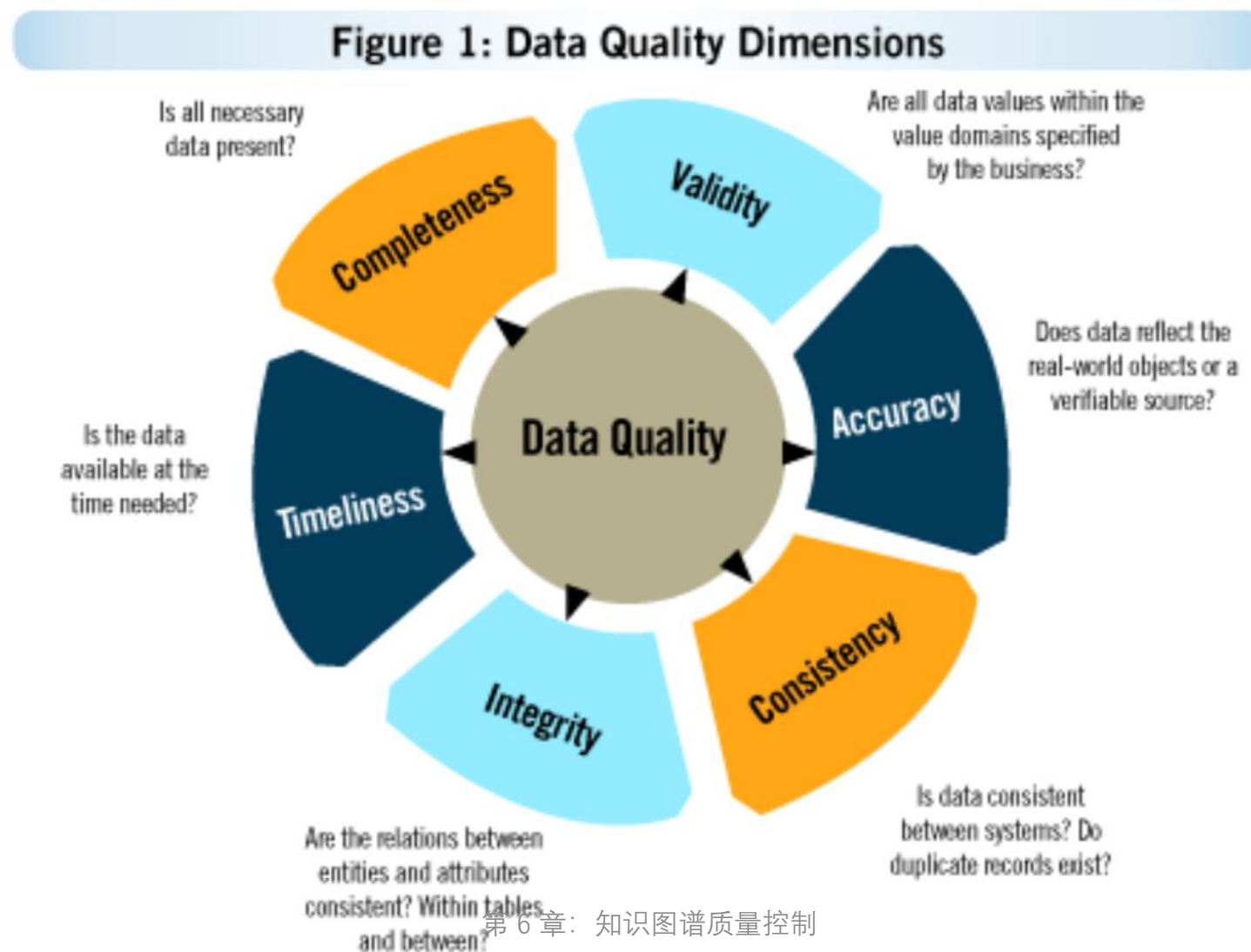


Indexing, search, data mining, query optimization, data streams, web scale data management, information extraction and understanding, data-driven research, parallel processing ...

Data integration, federated databases, domain-specific query processing, ontology, semantic web, dataspace ...

?

# 关系数据库的六个数据质量维度





# 举例：关系型数据库

CID↕	Name↕	Address↕	City↕	Sex↕
11↕	张三↕	邯郸路 220 号计算机楼 527 室↕	上海↕	0↕
24↕	李四↕	<u>鄞奉路</u> 978 号 7 号楼 702 室↕	宁波↕	1↕

CNO↕	Name↕	Gender↕	Address↕	Phone/Fax↕
24↕	王五↕	F↕	杭州市朝晖二区 555 号 2-308 室 310012↕	0571-88480666/↕ 0571-87074789↕
493↕	李四↕	M↕	宁波市 <u>鄞奉路</u> 978 号 7 号楼 702 室 315012↕	0574-87074789↕

NO↕	Name↕	Gender↕	Address↕	<u>city</u> ↕	<u>zip</u> ↕	Pone↕	Fax↕	CID↕	<u>Cno</u> ↕
1↕	张三↕	F↕	邯郸路 220 号 计算机楼 527 室↕	上 海↕	↕	↕	↕	11↕	↕
2↕	李四↕	M↕	<u>鄞奉路</u> 978 号 7702 室↕	宁 波↕	315012↕	0574-87074789↕	↕	24↕	493↕
3↕	王五↕	F↕	<u>朝二区</u> 555 号 2-308 室↕	杭 州↕	310012↕	1571-88480666↕	0571-↕ 88480667↕	↕	24↕

- **Different Schemas:** e.g., “Sex”-“Gender”, “Phone/Fax”-“Phone”+“Fax”
- **Inconsistency values:** e.g., “0/1”-“F/M”

# 举例: DBLP

- **Polyseme**: 10+ different “Wei Wang”
- **Synonyms**: “Pei Lee” and “Pei Li”

Wei Wang:	16
Tao Wang:	18
Jun Zhang:	21
Wei Li:	27
Lei Wang:	30
Michael Wagner:	5
Jim Smith:	3

The screenshot displays the DBLP website interface. On the left, the 'Search dblp' section shows 'Author results' for 'Wei Wang'. A red circle highlights the 'Exact matches' section, which lists several entries for 'Wei Wang' from different institutions, including National University of Singapore, University at Albany / Purdue University, and Fudan University. Below this, 'Likely matches' are also listed. On the right, two panels show search results for 'Pei Li' and 'Pei Lee'. The 'Pei Li' panel shows 'Journal Articles' from 2015 and 2014. The 'Pei Lee' panel shows 'Conference and Workshop Papers' from 2014, 2013, and 2012. The interface includes navigation links like 'Home > Persons' and 'Other persons with a similar name'.

dblp  
computer science bibliography

[+] Search dblp  
[-]

> Home

[+] Author results

Exact matches

- Wei Wang
- Wei Wang 0001  
National University of Singapore
- Wei Wang 0002  
College of Nanoscale Science, University at Albany / Purdue University
- Wei Wang 0003  
School of Life Science, Fudan University, China
- Wei Wang 0004  
Center for Engineering and Scientific Computation, Zhejiang University
- show all

Likely matches

- Wei Wang 0010  
UCLA / University of North Carolina at Chapel Hill
- Wei Wang 0009  
Fudan University, Shanghai, China
- Weidong Wang
- Wei-Fan Wang  
aka: Welfan Wang

show all 351 matches

[+] Pei Li  
[-]

> Home > Persons

[+] Other persons with a similar name

[+] Journal Articles

2015

- [j19] Teng Li, Jian Mao, 'Rating cloud stor
- [j18] Jinxin Zhang, Chac 3-D simulation st Reliability 55(8): 1
- [j17] Haibin Duan, Pei L Interactive Learr (2015)

2014

- [j16] Yingwen Chen, Mi Empirical study c 2014: 180 (2014)
- [j15] Pei Li, Yunchuan S Modeling and pei Communication S

[+] Pei Lee  
[-]

> Home > Persons

[+] Other persons with a similar name

[+] Conference and Workshop Papers

2014

- [c5] Pei Lee, Laks V. S. Lakshmanan, Evangelos E. Milios: CAST: A Context-Aware Story-Teller for Streaming Social Co
- [c4] Pei Lee, Laks V. S. Lakshmanan, Evangelos E. Milios: Incremental cluster evolution tracking from highly dynam
- [c3] Pei Lee, Laks V. S. Lakshmanan, Mitul Tiwari, Sam Shah: Modeling impression discounting in large-scale recommen

2013

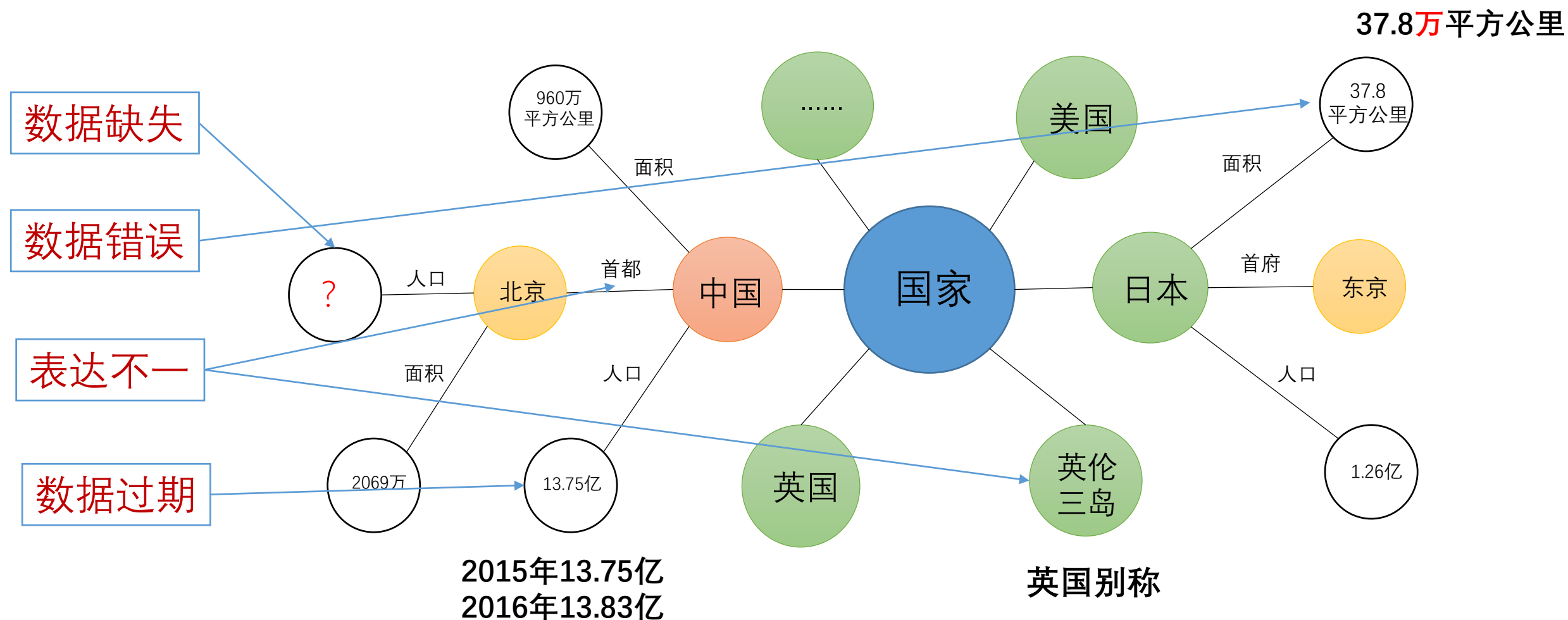
- [c2] Pei Lee, Laks V. S. Lakshmanan, Evangelos E. Milios: KeySee: supporting keyword search on evolving events in s

2012

- [c1] Pei Lee, Laks V. S. Lakshmanan, Jeffrey Xu Yu: On Top-k Structural Similarity Search. ICDE 2012: 774-785

# 知识图谱同样存在质量问题!

# 知识图谱中的质量问题



# 如何应对？

做好质量控制 - 知识图谱构建中**必不可少**的一环

- **高质量的知识图谱**
  - **构建层面**：首先是我们追求的**构建目标**。
  - **应用层面**：决定了最终应用的**落地效果**。



# 第 6 章 知识图谱质量控制

- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
- 知识图谱数据补全的质量控制
- 知识图谱数据更新的质量控制



# 知识图谱质量评估与控制概述

---

# 本节大纲

- 知识图谱质量评估与控制概述
  - 知识图谱质量评估概述
  - 知识图谱质量控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
- 知识图谱数据补全的质量控制
- 知识图谱数据更新的质量控制

# 何谓“高质量”知识图谱？

## • 从知识构建角度

- 仅关注构建层面
- 衡量数据和知识本身的质量
- 有一些通用的衡量标准
- 准确、一致、时效、完整…

## • 从最终应用角度

- 关注KG应用层面
- 没有通用衡量标准，case by case
- 只看KG是否满足应用需求
- KG结构、表达方式…

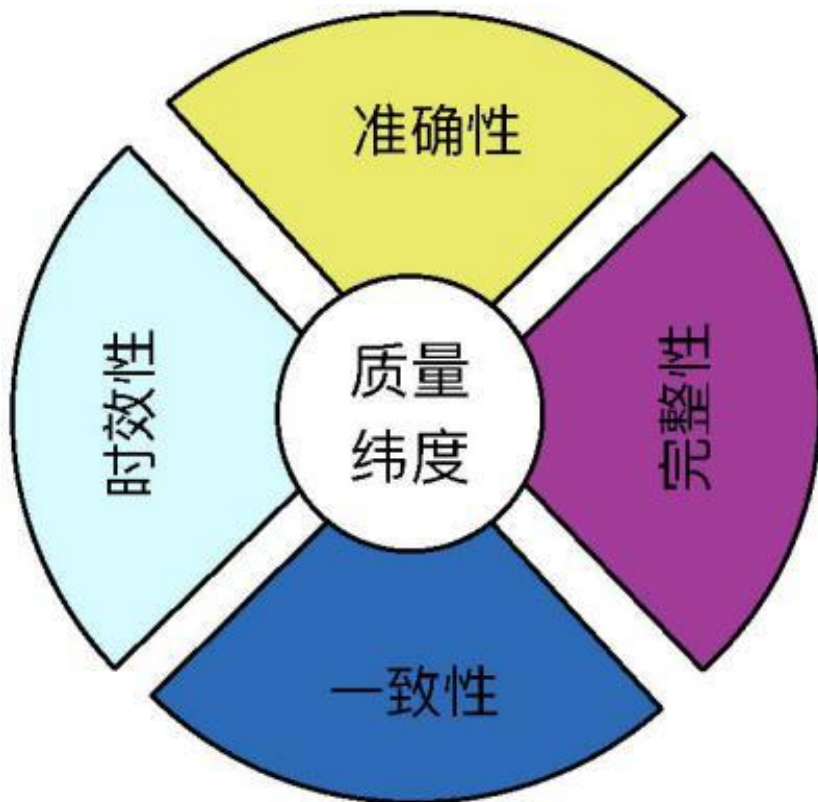
本章关注点



暂不关注



# 知识图谱质量的评估维度



知识图谱数据质量的四个维度

- **一致性**：考察图谱中的知识是否统一、一致。
- **准确性**：考察图谱中各类知识的正确程度。
- **时效性**：准确性的一个子维度，但其主要强调图谱中知识是否是当下最新知识。
- **完整性**：考察图谱中的知识对某相关领域的覆盖程度。



# 知识图谱质量的评估维度

知识图谱的质量评估细分表

<div>质量维度</div> <div>知识维度</div>	一致性	准确性	时效性	完整性
概念	√	√	√	√
实体	√	√	√	√
属性	√	√	√	√
关系	√	√	√	√
属性值	√	√	√	√

# 知识图谱质量的检测与评估

## 知识图谱质量检测与评估一览表

质量维度 知识维度	一致性	准确性	时效性	完整性
概念	专家人工	专家人工（抽样）检测	OR	专家人工 OR 外部数据
实体	冗余实体检测			
属性	一致性检测 (冲突检测)	借助外部领域数据 对比评估	OR	完整度评估
关系				
属性值				

# 一致性评估

## • 概念：量小，人工更靠谱

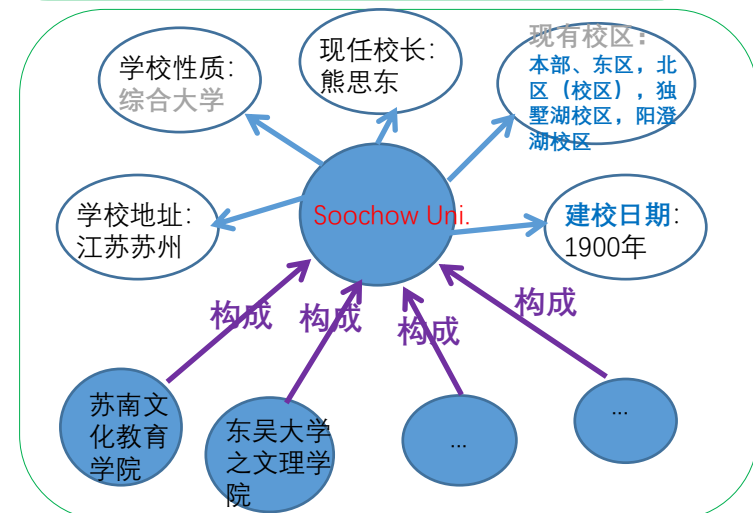
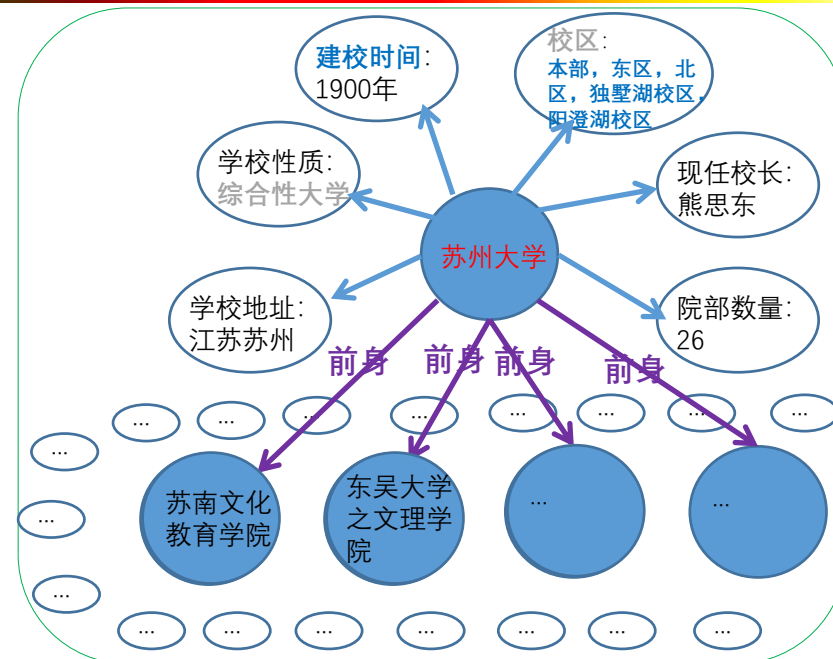
- 抽样：抽取部分样本数据
- 评估：人工评判质量

## • 实体：冗余实体检测

- 检测发现知识图谱中的冗余实体
- 核心：实体匹配算法
  - 实体相似度计算模型

## • 属性、关系、属性值：一致性检测


- 发现表达不一致的数据。
- 核心：多源融合算法
  - 属性匹配；实体匹配；属性值归一化



# 完整性评估

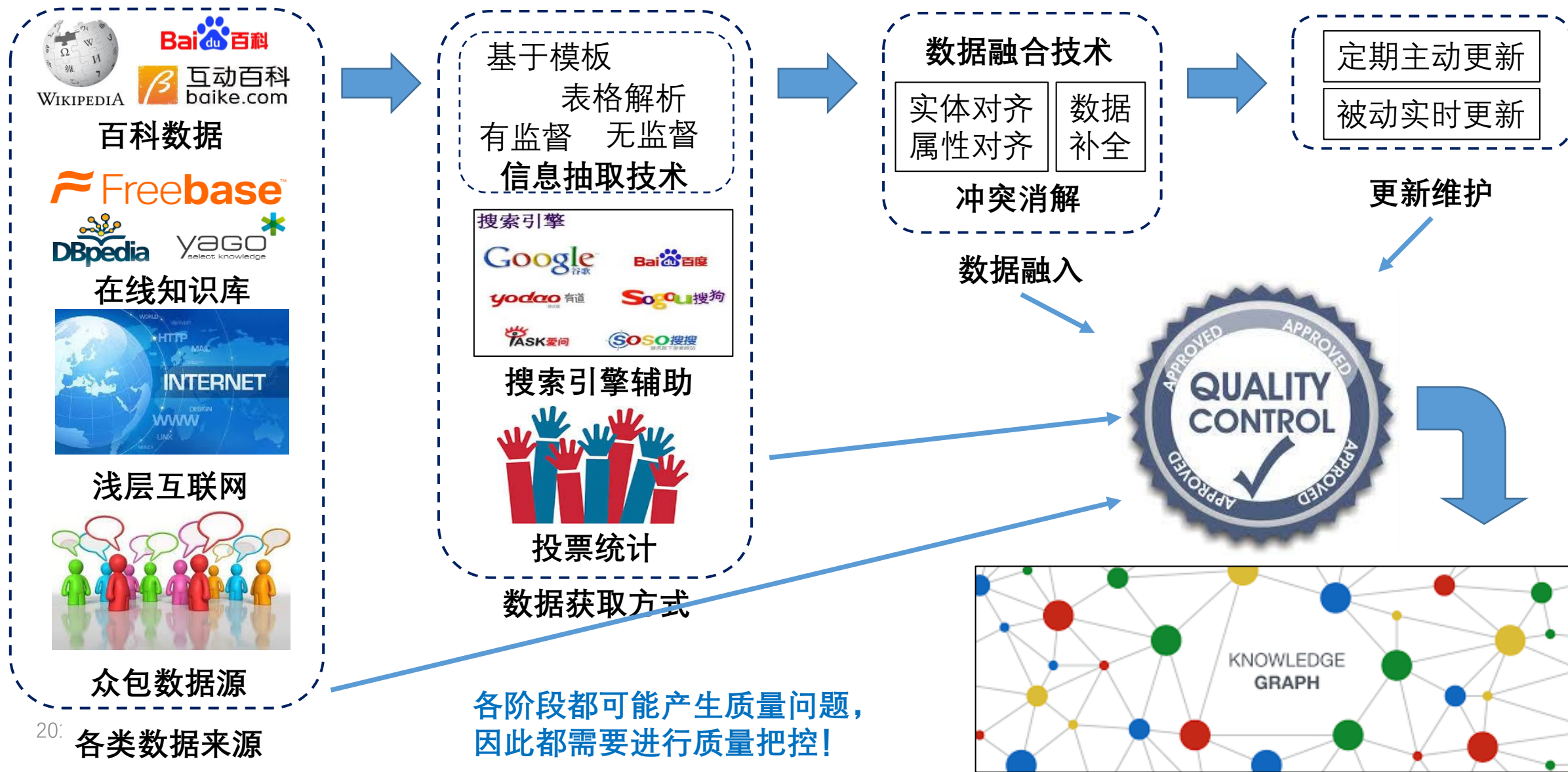
- 概念、实体：人工抽样评估法
  - 抽样式获取样本数据，检查数据的缺失率
- 关系、属性、属性值：完整度评估
  - 基于分布的完整性评估（以基于三大百科构建的KG为例）
    - 计算三大百科在概念C下的属性A的总体完整度和独立完整度
    - 计算概念C下缺失属性A的实体需要补全的概率
    - 使用实体需要补全的概率衡量实体所在概念c需要补全属性A的概率
    - 用所有概念的完整度的加权平均计算KG的完整度

# 本节大纲

- 知识图谱质量评估与控制概述
  - 知识图谱质量评估概述
  - 知识图谱质量控制概述 
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
- 知识图谱数据补全的质量控制
- 知识图谱数据更新的质量控制



# KG质量问题源自何处？



# 知识图谱质量控制概述

- 知识图谱数据来源的质量控制
  - 各类数据源的质量控制方法
- 知识图谱数据获取的质量控制
  - 搜索质量控制，信息抽取质量控制
- 知识图谱数据融入的质量控制
  - 知识的融合统一，知识的链接融入
- 知识图谱数据补全的质量控制
  - 实体类型补全、关系补全、属性值补全
- 知识图谱数据更新的质量控制
  - 错误数据清洗，过期数据更新

# 知识图谱质量控制概述

- 知识图谱数据来源的质量控制
  - 各类数据源的质量控制方法
- 知识图谱数据获取的质量控制
  - 搜索质量控制，信息抽取质量控制
- 知识图谱数据融入的质量控制
  - 知识的融合统一，知识的链接融入
- 知识图谱数据补全的质量控制
  - 实体类型补全、关系补全、属性值补全
- 知识图谱数据更新的质量控制
  - 错误数据清洗，过期数据更新

# 知识图谱质量控制概述 – 数据来源



百科数据



在线知识库



浅层互联网



众包数据源

- 常见的知识来源质量评估
  - 互联网获取数据的质量评估
  - 众包获取数据的质量评估

# 知识图谱质量控制概述 – 数据来源

## • 互联网数据的质量评估

- 基于网站权威性的数据可信度评估，  
如.mil>int>.gov>.org>.edu>.com>.net

- 基于关联规则的数据可信度评估

通过发现评论者对图书是否存在偏见，进而间接体现评论的可信度





# 知识图谱质量控制概述 – 数据来源

## • 互联网数据的质量评估

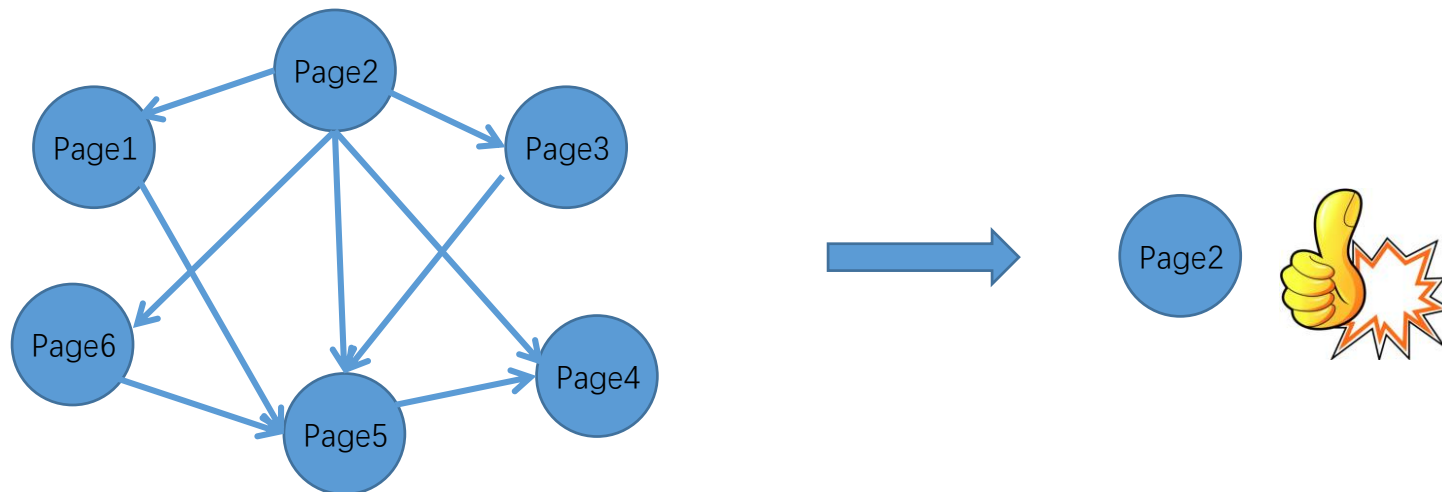
### • 信任值传播机制

#### • 评价机制：

- 某条信息发布网站的可信度越高，这条信息的可信度就越高；
- 某条信息被转载的次数越多，这条信息的可信度就越高

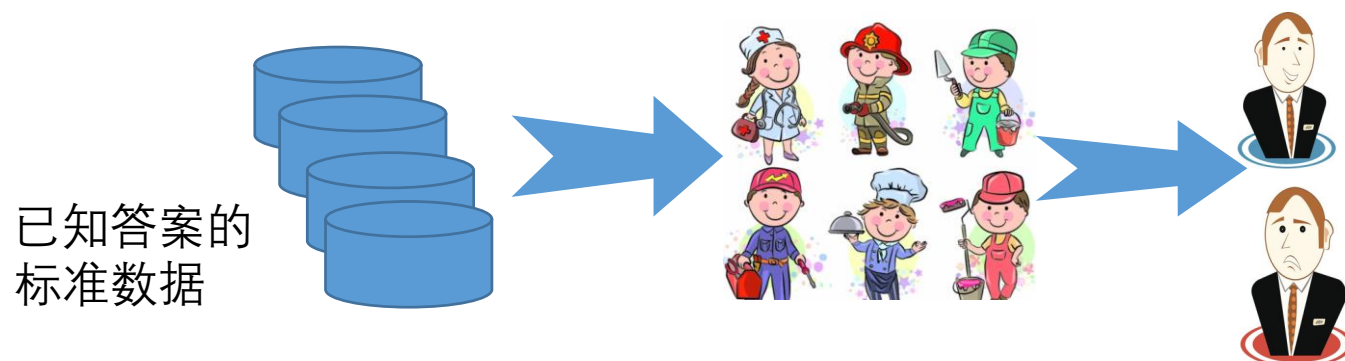
#### • 预处理：

- 有无重复记录，若有需去重

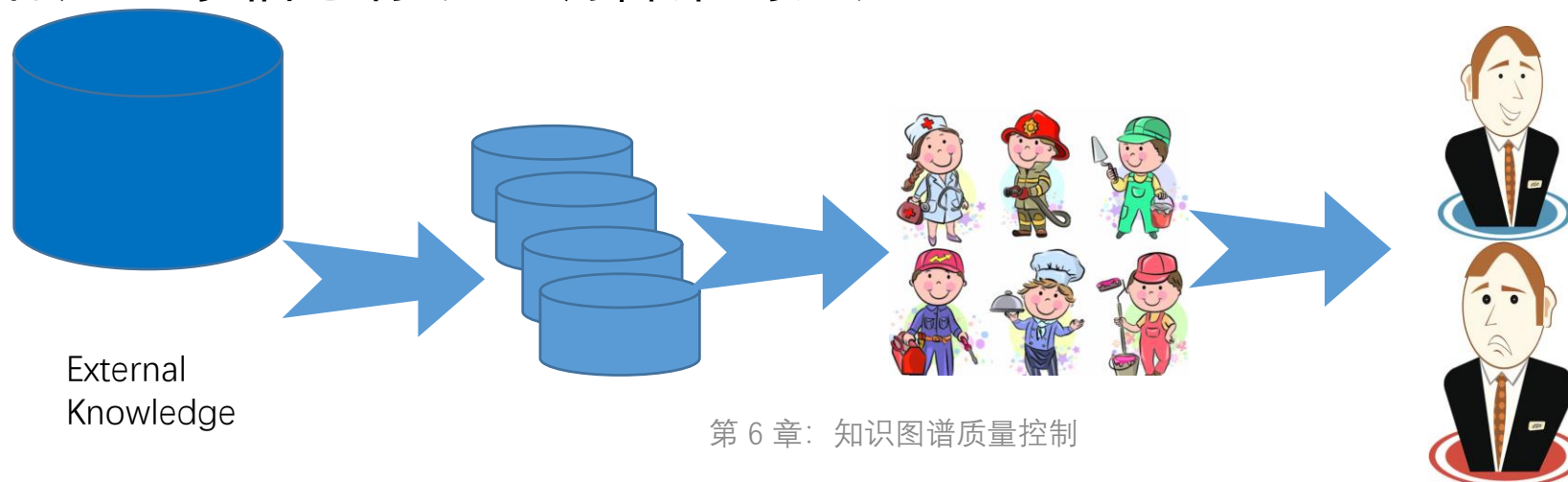


# 知识图谱质量控制概述 – 数据来源

- 众包数据的质量评估（各类其他数据类似）
  - 使用黄金标准数据(Golden standard data)评估



- 利用冗余信息标识正确答案的方法



# 知识图谱质量控制概述

- 知识图谱数据来源的质量控制
  - 各类数据源的质量控制方法
- 知识图谱数据获取的质量控制
  - 搜索质量控制，信息抽取质量控制
- 知识图谱数据融入的质量控制
  - 知识的融合统一，知识的链接融入
- 知识图谱数据补全的质量控制
  - 实体类型补全、关系补全、属性值补全
- 知识图谱数据更新的质量控制
  - 错误数据清洗，过期数据更新

# 知识图谱质量控制概述 – 数据获取

## • 从浅层互联网获取知识

- 构建查询词：通过搜索引擎，搜索到相关网页
- 信息抽取：再搜索到的网页中，抽取获得相关知识

## • 从自然文本中获取知识

- 基于pattern的方法：syntax-based自增迭代和 semantic-based自增迭代
- 基于模型的方法：从文本中抽取大量的和给定语法 pattern匹配的句子，然后借助语义分析工具，从句子中抽取需要的信息

数据获取的方式主要依托：WSE+IE  
常用质量评估方法：黄金数据集评估  
如何控制质量？。。。主要是IE的质量

基于模板  
表格解析  
有监督 无监督  
信息抽取技术



搜索引擎辅助



投票统计

# 知识图谱质量控制概述 – 数据获取

- 如何管理迭代式IE中出现的错误和纠正错误？（语义漂移问题）
  - 依托非结构化文本构建知识图谱主要依赖信息抽取技术
  - 信息抽取的主流是自增迭代式（bootstrapping）信息抽取技术
  - 自增迭代式信息抽取的一大问题是语义漂移问题
  - 语义漂移问题？
    - 自增迭代式的抽取最终都会倾向于抽取到一些含义模糊的实例或者与目标语义类相关性较弱的上下文模式，导致开放式自动信息抽取(IE)系统的抽取质量的降低

# 知识图谱质量控制概述

- 知识图谱数据来源的质量控制
  - 各类数据源的质量控制方法
- 知识图谱数据获取的质量控制
  - 搜索质量控制，信息抽取质量控制
- 知识图谱数据融入的质量控制
  - 知识的融合统一，知识的链接融入
- 知识图谱数据补全的质量控制
  - 实体类型补全、关系补全、属性值补全
- 知识图谱数据更新的质量控制
  - 错误数据清洗，过期数据更新

# 知识图谱质量控制概述 – 数据融入

- **知识的融合统一：**融合多源知识库(图谱)中的数据
  - 概念对齐与融合
    - 包括概念合并、概念上下位关系合并以及概念的属性定义合并
  - 实体对齐
    - 判断相同或不同数据集中的两个实体是否指向真实世界同一对象
  - 属性对齐
    - 识别来自单一或多个数据源的属性之间存在的对应关系
  - 属性值归一化
    - 规范同一类型的属性值的表现形式
- **知识链接与融入：**将获取的各类知识“链接”到知识图谱
  - 概念链接（量少，人工融入最准确）
  - 实体链接（**关键问题，研究热点：刚需、海量、歧义性大**）
  - 属性链接（实体链接正确了，属性链接相对简单些）

# 知识图谱质量控制概述

- 知识图谱数据来源的质量控制
  - 各类数据源的质量控制方法
- 知识图谱数据获取的质量控制
  - 搜索质量控制，信息抽取质量控制
- 知识图谱数据融入的质量控制
  - 知识的融合统一，知识的链接融入
- 知识图谱数据补全的质量控制
  - 实体类型补全、关系补全、属性值补全
- 知识图谱数据更新的质量控制
  - 错误数据清洗，过期数据更新



# 知识图谱质量控制概述 – 数据补全

- 实体类型补全
  - 又称：实体分类，或 类型断言 (Type Assertions)
  - 旨在给出实体缺失的上位概念
- 实体间关系补全
  - 又称：关系预测 (Relation Prediction)
  - 旨在补全图谱中缺失的实体间的一些关系
- 实体属性值补全
  - 与数据库领域研究的数据补全 (data imputation) 相近似
  - 旨在补全图谱中实体缺失的属性值

# 知识图谱质量控制概述

- 知识图谱数据来源的质量控制
  - 各类数据源的质量控制方法
- 知识图谱数据获取的质量控制
  - 搜索质量控制，信息抽取质量控制
- 知识图谱数据融入的质量控制
  - 知识的融合统一，知识的链接融入
- 知识图谱数据补全的质量控制
  - 实体类型补全、关系补全、属性值补全
- 知识图谱数据更新的质量控制
  - 错误数据清洗，过期数据更新

# 知识图谱质量控制概述 – 数据更新

- 错误数据清洗

- 关系数据库：找出并修正关系数据库中的错误属性值
- 知识图谱中：找出并修正图谱中的错误属性值或实体间关系

- 过期数据更新

- 随着时间的推移，数据是变动的
  - 一直在变的：人口，年龄，职位，作品数量，美国总统。。。
  - 不断新增的：新人，新公司，新词，。。。
- 旨在保持知识图谱中数据的“新鲜度”

# 本节大纲

- 在接下来的章节中，我们将对以下几方面的研究工作展开介绍

知识图谱数据获取的质量控制

搜索质量控制，信息抽取质量控制

知识图谱数据融入的质量控制

知识的融合统一，知识的链接融入

知识图谱数据补全的质量控制

实体类型补全、关系补全、属性值补全

知识图谱数据更新的质量控制


错误数据清洗，过期数据更新

# 知识图谱数据获取的质量控制

---

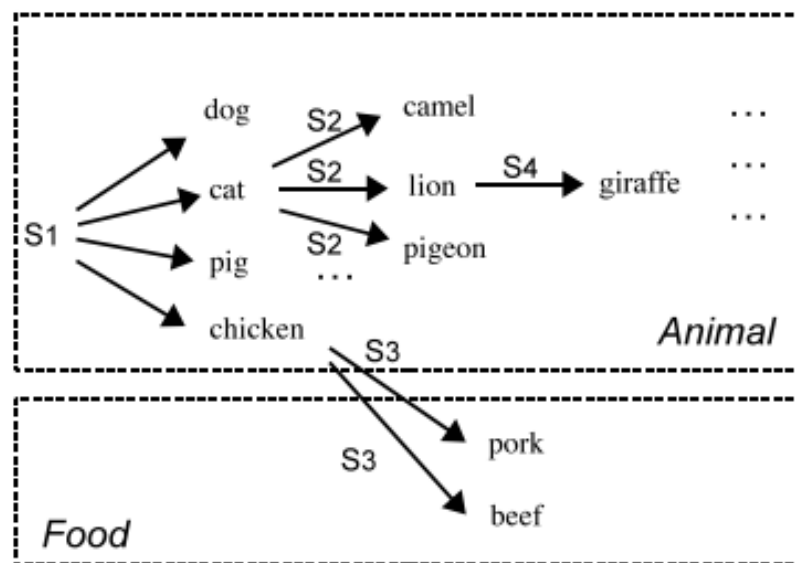
# 本节大纲

---

- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
  - 语义漂移问题的处理技术 
- 知识图谱数据融入的质量控制
- 知识图谱数据补全的质量控制
- 知识图谱数据更新的质量控制

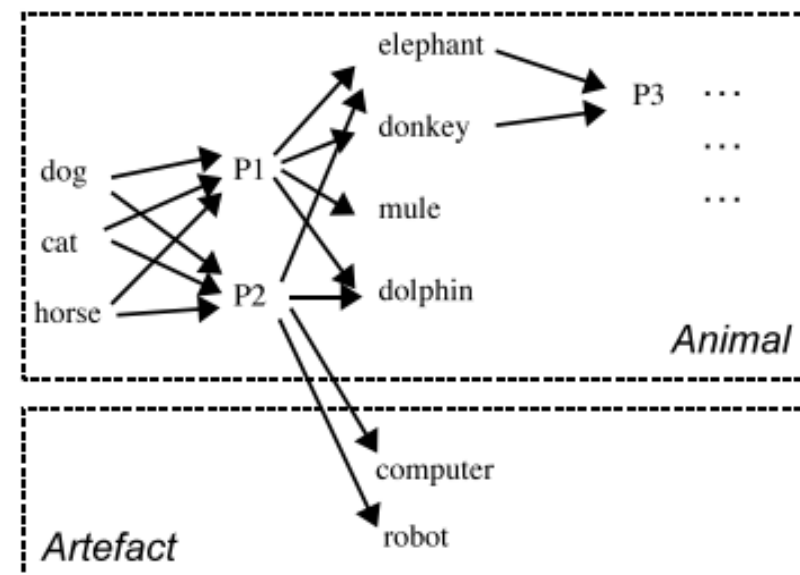
# 知识图谱质量控制概述 – 数据获取

- 什么是语义漂移？ ---- 两种自增迭代式IE中的语义漂移问题示例



S1="Animals **such as** dog, cat, pig and chicken, grow fast."  
 S2="Yoga Postures are named after animals **such as** camel, pigeon, lion and cat."  
 S3="Common food from animals **such as** pork, beef and chicken."  
 S4="Animals from African countries **such as** Giraffe and Lion."

(a) Semantic-based bootstrapping mechanism



P1: "... X is a kind of mammal ..."  
 P2: "Sometime, X is as clever as human beings"

(b) Syntax-based bootstrapping mechanism

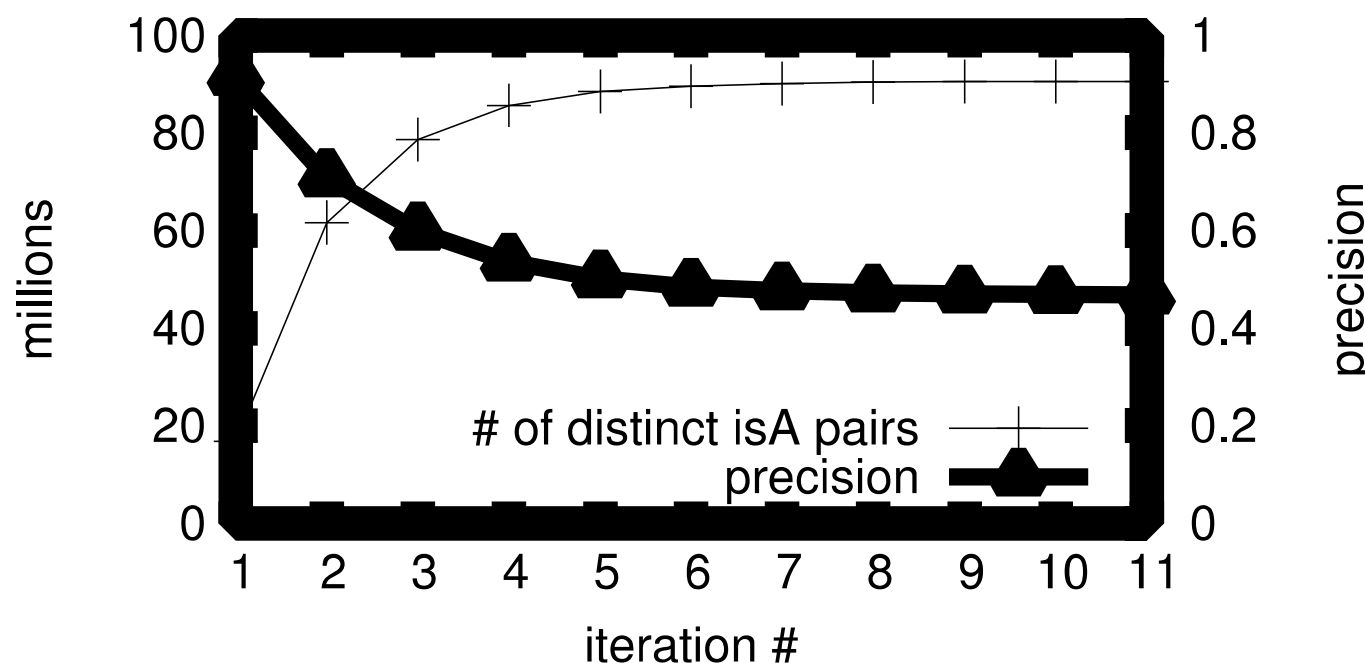
# 知识图谱质量控制概述 – 数据获取

## • 语义漂移造成的危害

### • 常见的迭代抽取式系统

• e.g.: *KnowItAll*, *SnowBall*, *ProBase* ...

• 在几轮之后，准确度急剧下降。。。





# 知识图谱质量控制概述 – 数据获取

## • 语义漂移问题处理主流方法

- Mutual Exclusion Bootstrapping (PACLING'07)
  - Drop those instances belonging to mutually exclusive classes
- Type Checking (WSDM'10)
  - Check the type of an entity for correctness
- Random Walk Ranking (ICDM'06)
  - Construct a graph, do random walk ranking
- Pattern-Relation Duality Ranking (WSDM'11)
  - The quality of a pattern (tuple) can be determined by the tuples (patterns) it extracts.
- A Model based on Detected Drifting Points (EDBT'14, TKDE'17)

# 语义漂移问题处理技术

- Mutual Exclusion Bootstrapping
  - **Pros and Cons:** High Precision, Low Recall

## Positives:

Canada

Egypt

France

...

war with ×  
ambassador to ×  
war in ×  
occupation of ×

Planet Earth  
Freetown  
North Africa

## Negatives:

Asia

Europe

London

Florida

...

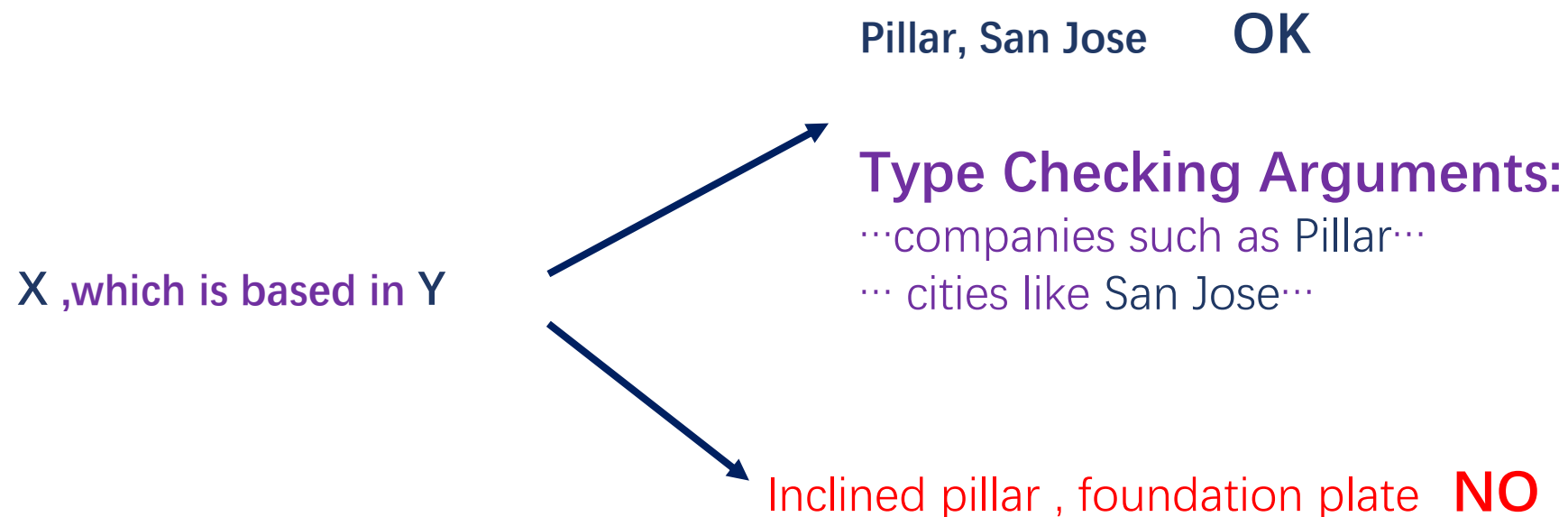
nations like ×  
countries other than ×  
country like ×

Pakistan  
Sri Lanka  
Greece  
Russia

# 语义漂移问题处理技术

## • Type Checking

- Checking types of relevant entities
- **Pros and Cons:** High Precision, Low Recall



# 语义漂移问题处理技术

## • 主流方法

- Mutual Exclusion Bootstrapping (PACLING'07)
  - Drop those instances belonging to mutually exclusive classes
- Type Checking (WSDM'10)
  - Check the type of an entity for correctness
- Random Walk Ranking (ICDM'06)
  - Construct a graph, do random walk ranking
- Pattern-Relation Duality Ranking (WSDM'11)
  - The quality of a pattern (tuple) can be determined by the tuples (patterns) it extracts.
- A Model based on Detected Drifting Points (EDBT'14, TKDE'17)

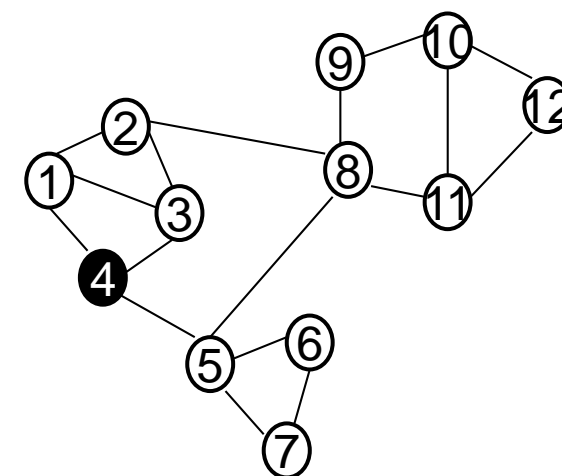
# 语义漂移问题处理技术

- Random Walk based Cleaning

$$\vec{r}_i = c\tilde{W}\vec{r}_i + (1-c)\vec{e}_i$$

Ranking vector      Adjacent matrix      Restart p      Starting vector

$$\begin{pmatrix} 0.13 \\ 0.10 \\ 0.13 \\ 0.22 \\ 0.13 \\ 0.05 \\ 0.05 \\ 0.08 \\ 0.04 \\ 0.03 \\ 0.04 \\ 0.02 \end{pmatrix} = 0.9 \times \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 1/2 & 1/2 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 0 & 1/4 & 0 & 0 & 0 & 1/2 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/3 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/3 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.13 \\ 0.10 \\ 0.13 \\ 0.22 \\ 0.13 \\ 0.05 \\ 0.05 \\ 0.08 \\ 0.04 \\ 0.03 \\ 0.04 \\ 0.02 \end{pmatrix} + 0.1 \times \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$



# 语义漂移问题处理技术

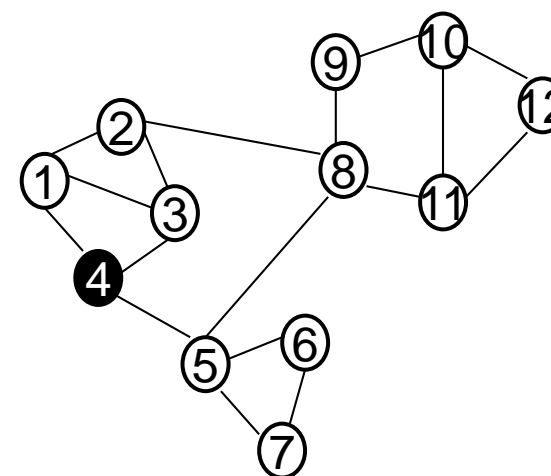
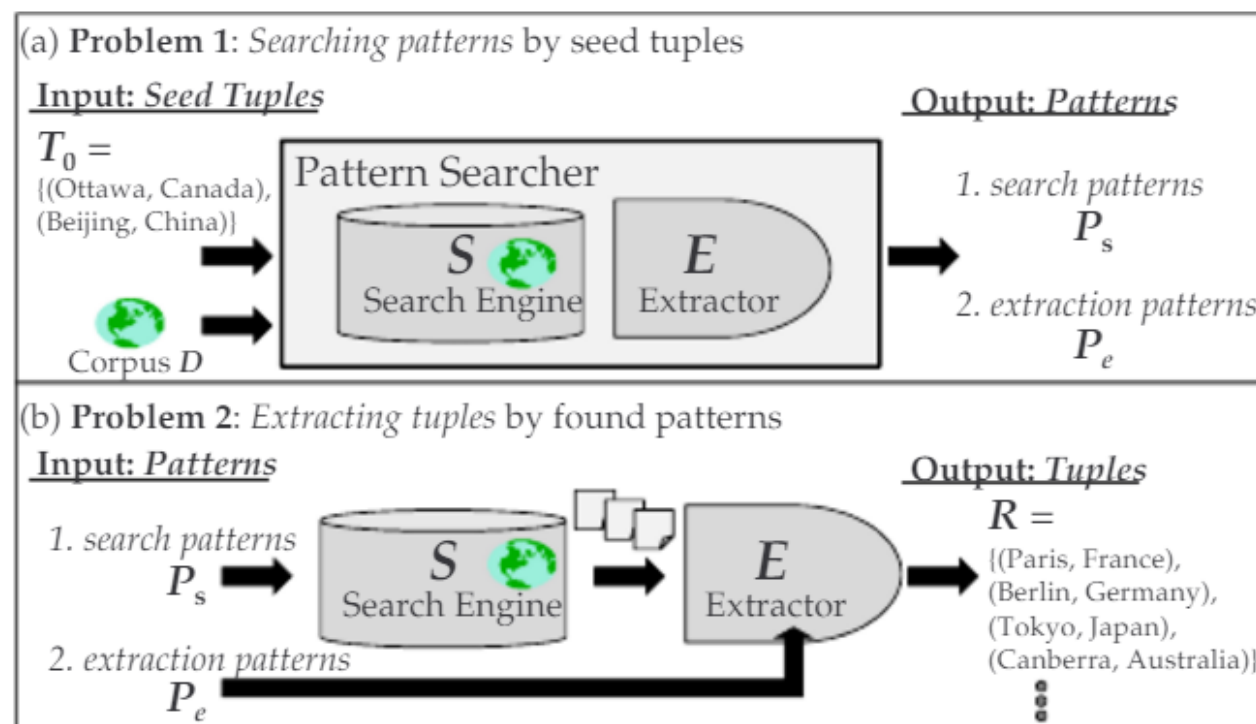
## • 主流方法

- Mutual Exclusion Bootstrapping (PACLING'07)
  - Drop those instances belonging to mutually exclusive classes
- Type Checking (WSDM'10)
  - Check the type of an entity for correctness
- Random Walk Ranking (ICDM'06)
  - Construct a graph, do random walk ranking
- Pattern-Relation Duality Ranking (WSDM'11)
  - The quality of a pattern (tuple) can be determined by the tuples (patterns) it extracts.
- A Model based on Detected Drifting Points (EDBT'14, TKDE'17)

# 语义漂移问题处理技术

## • Pattern-Relation Duality

- **Idea:** The quality of a pattern (tuple) can be determined by the tuples (patterns) it extracts.
- **Cons:** still can not reach high precision and recall



RW on Precision  
 RW on Recall  
 $F\text{-Score} = \text{Precision} + \text{Recall}$   
 Ranking with F-Score

# 语义漂移问题处理技术

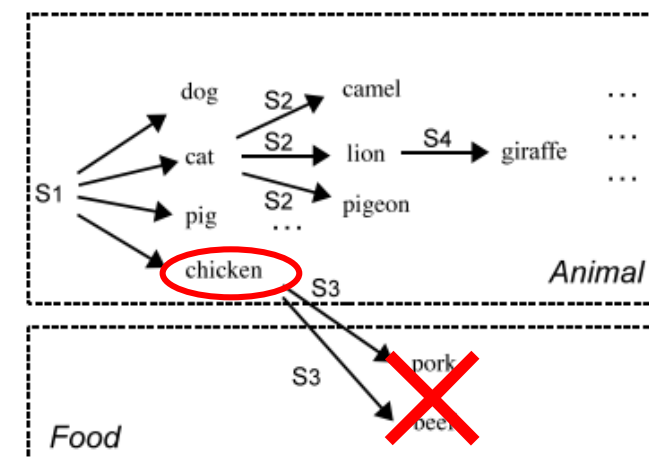
## • 主流方法

- Mutual Exclusion Bootstrapping (PACLING'07)
  - Drop those instances belonging to mutually exclusive classes
- Type Checking (WSDM'10)
  - Check the type of an entity for correctness
- Random Walk Ranking (ICDM'06)
  - Construct a graph, do random walk ranking
- Pattern-Relation Duality Ranking (WSDM'11)
  - The quality of a pattern (tuple) can be determined by the tuples (patterns) it extracts.
- A Model based on Detected Drifting Points (EDBT'14, TKDE'17)



# 语义漂移问题处理技术

- 一个基于漂移点 (**Drifting Points**) 检测的预测模型
  - **Intuition:** Drifting Points (DPs) are the reasons of Semantic Drift.
- 两种DPs:
  - Intentional DPs
    - Synonyms such as Chicken
  - Accidental DPs
    - Errors by themselves
    - E.g., ... Countries such as France, Germany, Japan and New York.

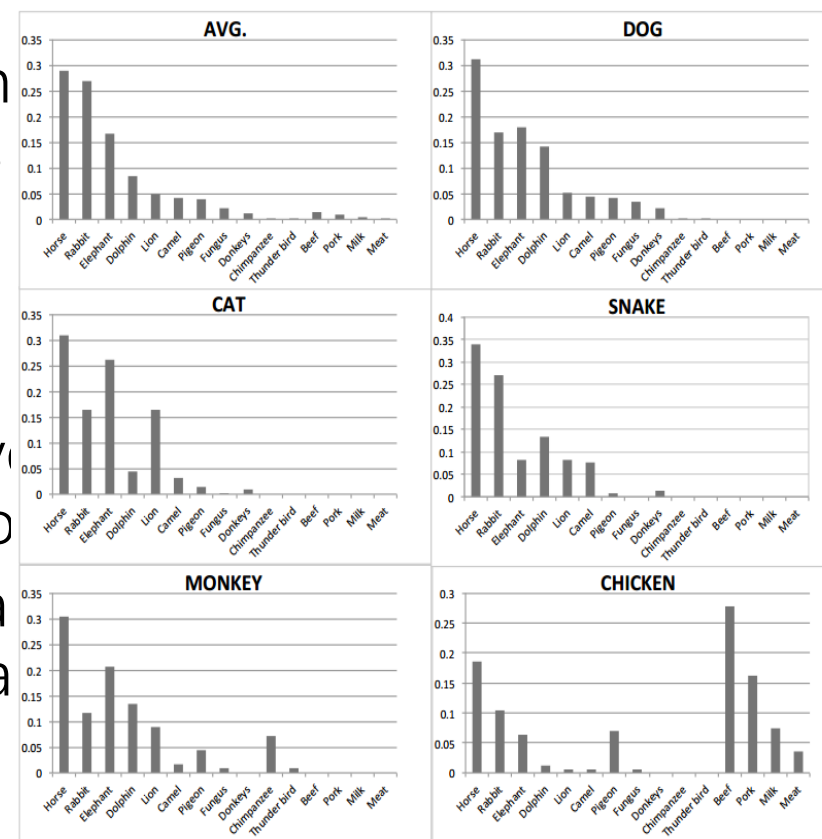


# 语义漂移问题处理技术

## • 漂移点 (DPs) 一些特征:

- For a target class, the distribution of instances different from the distribution of instances class.
- If classes  $C_1$  and  $C_2$  are mutually exclusive, likely an Intentional DP.
- An accidental DP is usually supported by very few instances is derived from very few (mostly one) instance.
- An error extraction (e isA C) triggered by a weak evidence, since the extraction is usual instances of C.

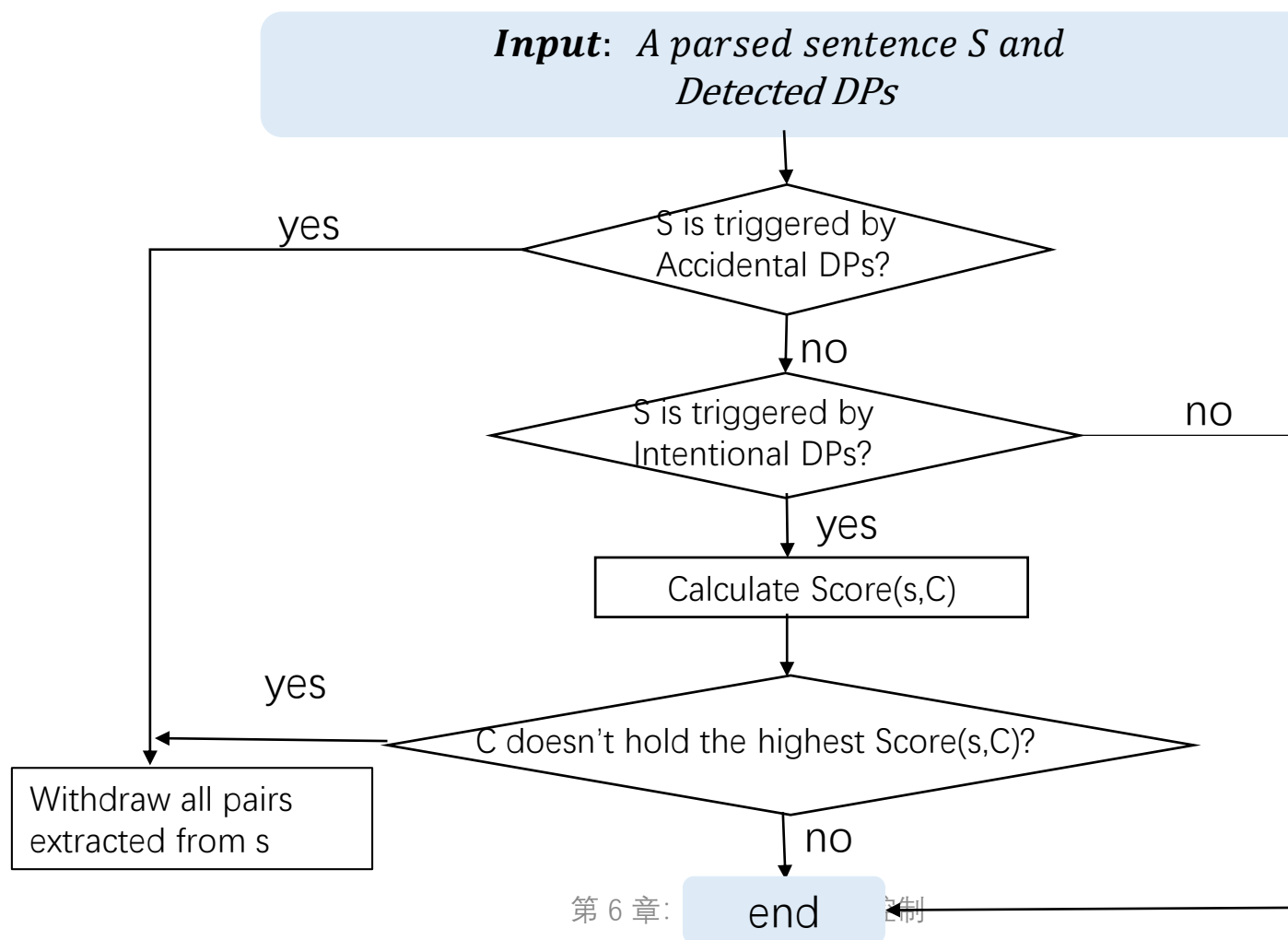
## • 利用以上特征, 构建DP检测模型



Distributions of instances triggered by DPs and non-DPs

# 语义漂移问题处理技术

- 检测到漂移点之后，再根据检测的漂移点，发现抽取中的错误。



# 语义漂移问题处理技术

Cleaning Method	$p_{error}$	$r_{error}$	$p_{correct}$	$r_{correct}$
Before Cleaning	-	-	0.4305	1.0
MEx	0.9119	0.1570	0.4592	<b>0.9832</b>
TCh	0.9423	0.1451	0.4789	0.9724
RW-Rank	0.5753	0.5831	0.5636	0.6509
PRDual-Rank	0.5621	0.6545	0.5812	0.6940
DP Cleaning	<b>0.9696</b>	<b>0.9145</b>	<b>0.8921</b>	0.9393

- (1) $p_{error}$ : percentage of removed errors in all the removed instances;
- (2) $r_{error}$ : percentage of removed errors in all the errors under each concept;
- (3) $p_{correct}$ : percentage of remained correct instances in all the remained instance;
- (4) $r_{correct}$ : percentage of remained correct instances in all the correct instances under each concept

# 知识图谱数据融入的质量控制

---

# 本节大纲

- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
  - 关系数据库中的数据融合统一
  - 知识图谱中的知识融合统一
  - 知识图谱中的知识链接融入
- 知识图谱数据补全的质量控制
- 知识图谱数据更新的质量控制



# 关系数据库中的数据融合与统一

- 数据融合一般包括如下两个关键步骤：

模式匹配  
(Schema Mapping)

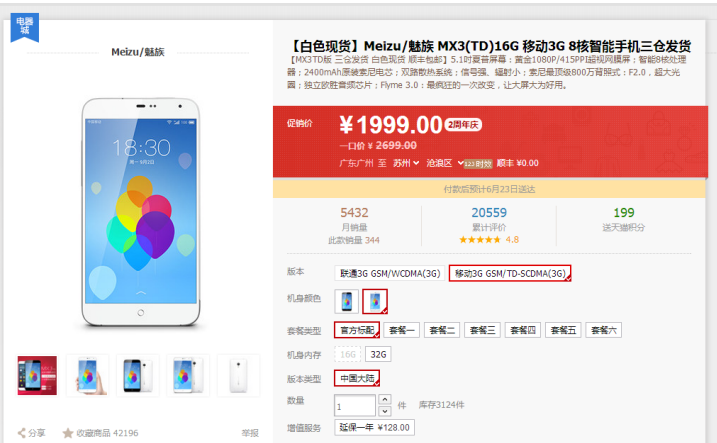
型号	品牌	京东价	屏幕尺寸	GPU	...
MX3	魅族	1799	5.1英寸	八核	...

型号	品牌	促价价	尺寸	...
MX3	MeiZu	1999	5.1英寸	...

记录匹配  
(Record Matching)

型号	品牌	京东价	屏幕尺寸	GPU	...
MX3	魅族	1799	5.1英寸	八核	...

型号	品牌	促价价	尺寸	...
MX3	MeiZu	1999	5.1英寸	...



型号	品牌	尺寸	GPU	京东价	天猫价	...
MX3	魅族	5.1英寸	八核	1799	1999	

# 关系数据库中的数据融合与统一

## • 模式匹配（Schema Mapping）

- 基于属性名字的字符串相似度的方法
  - 如基于编辑距离，Jaccard距离，欧式距离等等
- 基于实例下面的属性值的方法
  - 如基于统计的方法：抽样，众数等等
  - 基于共同的实例数量的方法

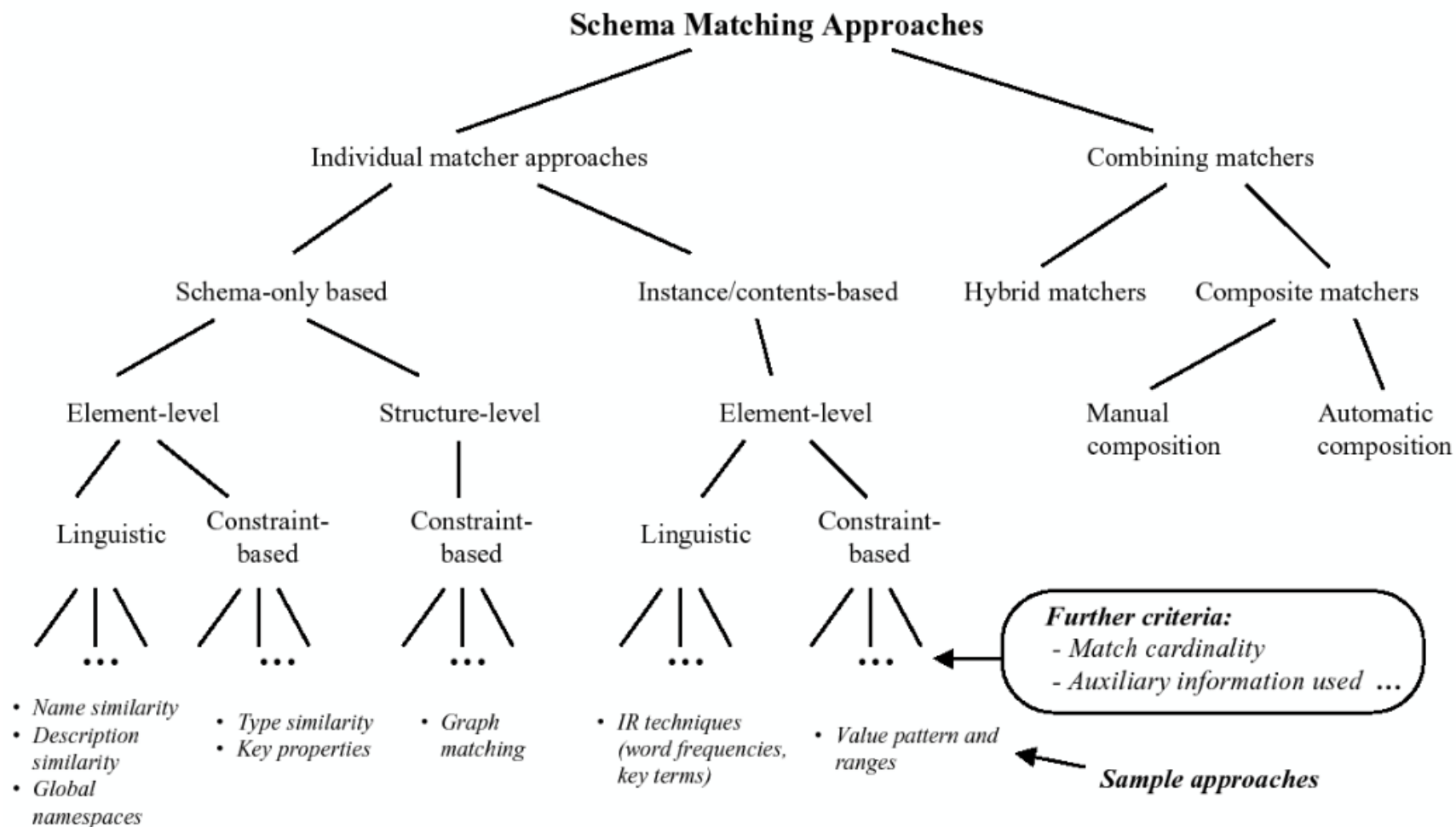
Extend Storage
32G
16G
8G
16G
—
32G

ROM
4G
8G
4G
16G
8G
8G
—



# 关系数据库中的数据融合与统一

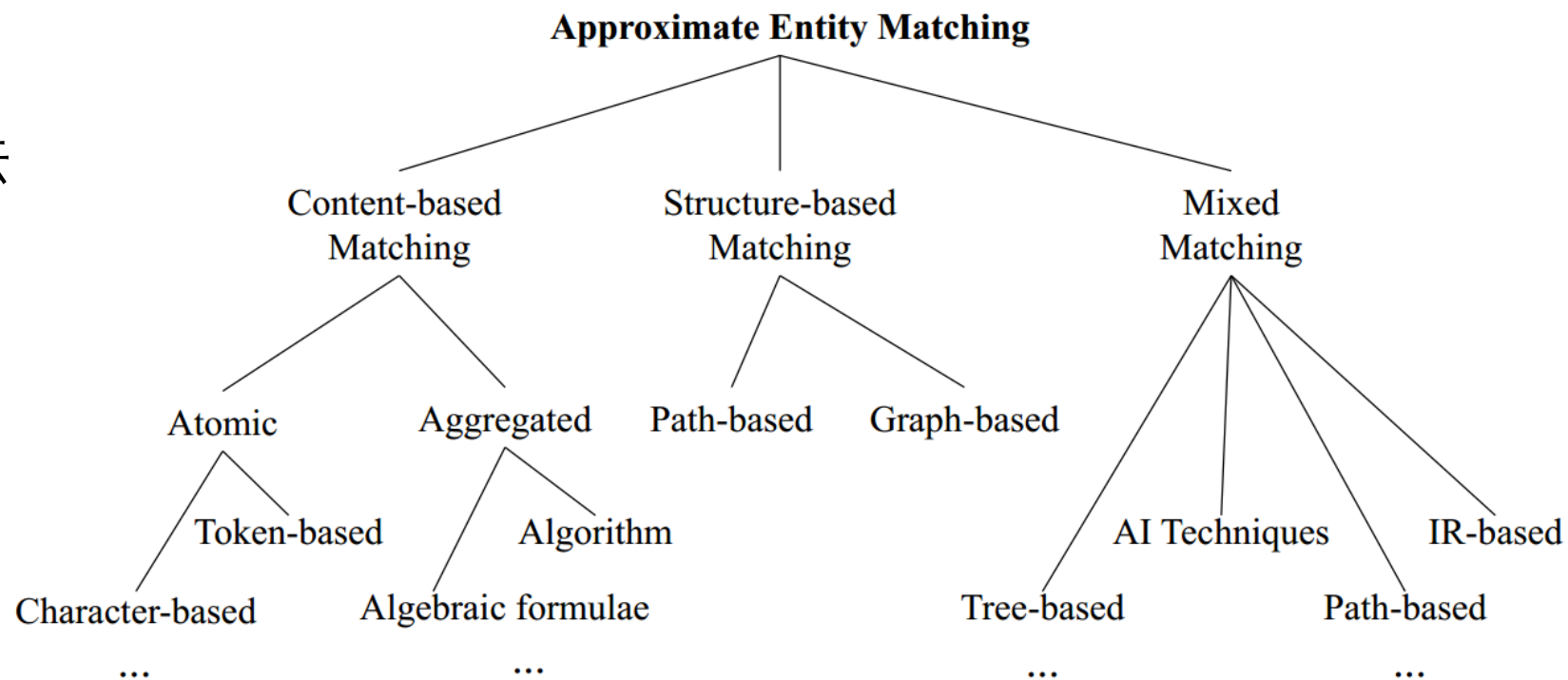
## • 模式匹配 (Schema Mapping)



# 关系数据库中的数据融合与统一

## • 记录匹配 (**Record Matching**)

- 基于内容的方法
- 基于结构的方法
- 基于混合模式的方法



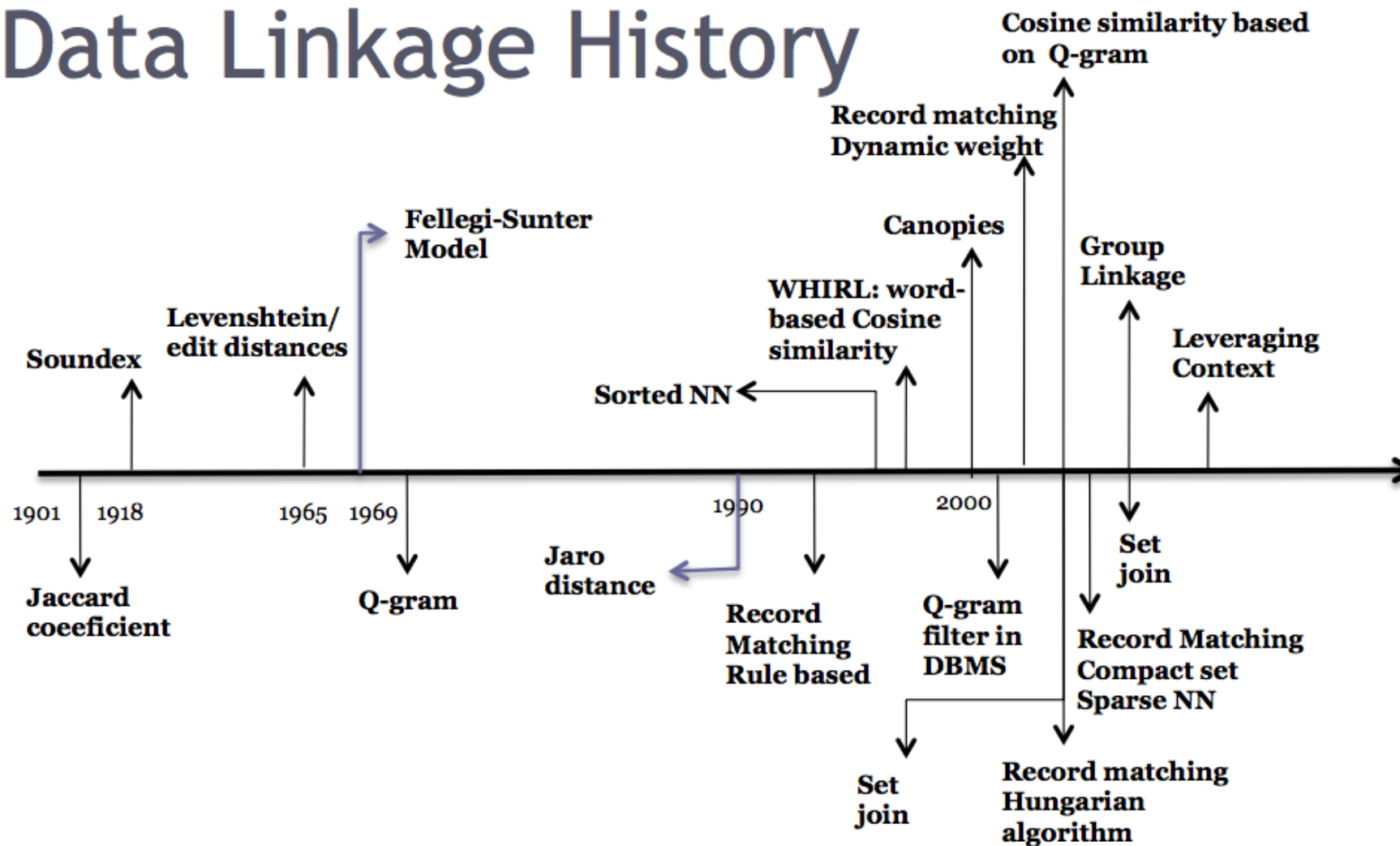
# 关系数据库中的数据融合与统一

## • 记录匹配（Record Matching）

- 结构化主属性：编辑距离，Jaccard，Q-gram等
  - ✓ 优点：可以唯一决定一个实体
  - ✓ 缺点：易受表达方式多样化的影响
- 结构化非主属性：匹配树，基于实例驱动的匹配方法等
  - ✓ 优点：选择决策能力高的非主属性参与匹配
  - ✓ 缺点：没有考虑树中不同层次节点的重要性的不同；易受缺失值影响
- 文本类型非主属性：基于阈值的匹配方法，基于无监督学习上下文的匹配方法等
  - ✓ 优点：同时考虑文本之间的字符串相似性和语义相似性
  - ✓ 缺点：依赖于WordNet，模型过于单一，健壮性差
- 借助外部资源的匹配方法：Crowdsourcing等
  - ✓ 优点：准确性高
  - ✓ 缺点：开销大，工人的准确性难以评估

# 关系数据库中的数据融合与统一

## Data Linkage History



# 本节大纲

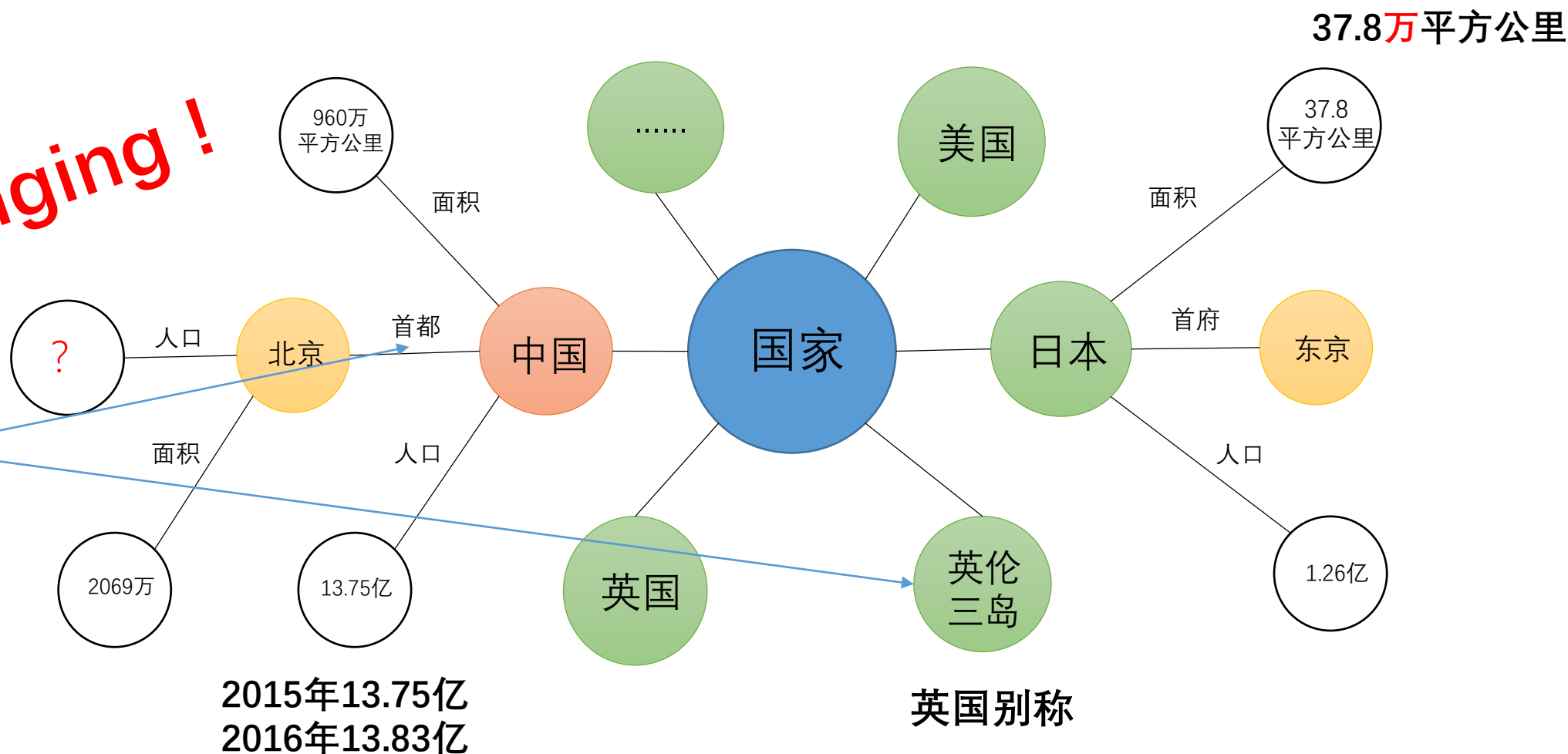
- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
  - 关系数据库中的数据融合统一
  - 知识图谱中的知识融合统一
  - 知识图谱中的知识链接融入
- 知识图谱数据补全的质量控制
- 知识图谱数据更新的质量控制



# 知识图谱中的知识融合与统一

More  
challenging!

表达不一



# 知识图谱中的知识融合与统一

- 概念对齐与融合
  - 包括概念合并、概念上下位关系合并以及概念的属性定义合并
- 实体对齐
  - 判断相同或不同数据集中的两个实体是否指向真实世界同一对象
- 属性对齐
  - 识别来自单一或多个数据源的属性之间存在的对应关系
- 属性值归一化
  - 规范同一类型的属性值的表现形式

# 概念对齐与融合

- 主流方法
  - 专家人工构建
  - 从可靠的结构化数据中映射生成
- 建模工具
  - Protégé
    - 开源软件
    - 基于RDF(S)、OWL等语义网
    - 图形化界面
    - 提供在线版本WebProtégé
  - PlantData
    - 商用软件
    - 屏蔽OWL, 可自定义本体语言





# 实体对齐

- 目标

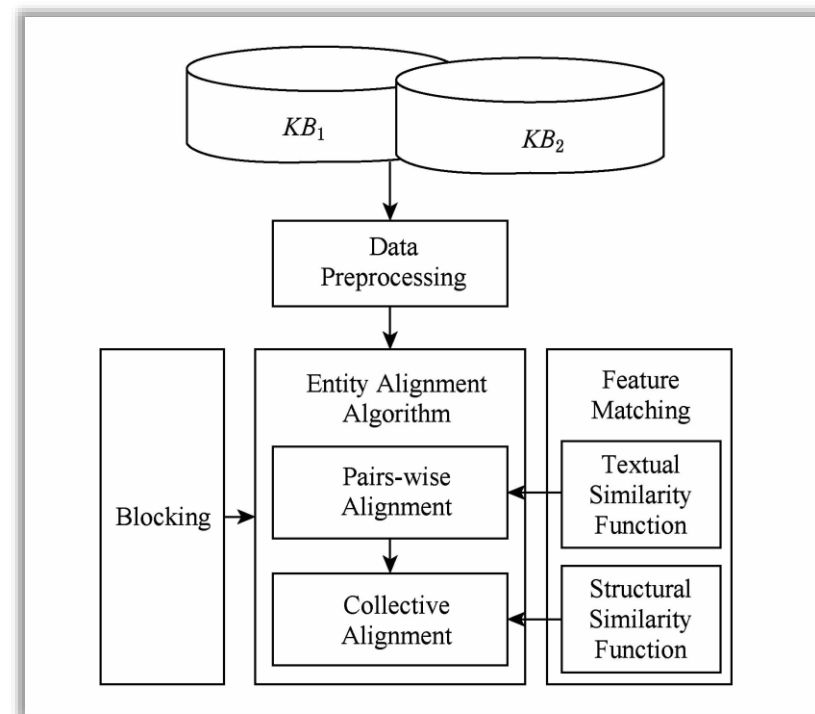
- 高质量链接多个现有知识库，并从顶层创建一个大规模的统一知识库，从而帮助机器理解底层数据

- 评价指标

- 质量：实体对齐的准确性和全面性
- 效率：大规模数据下的匹配时耗

- 问题与挑战

- 计算复杂度
- 数据质量
- 先验对齐数据的获取



# 实体对齐

- 主流方法
  - Property-based
    - 机器学习方法
      - Febrl – A Freely Available Record Linkage System with a Graphical User Interface(KDD 2008)
    - 基于概率
      - PARIS: Probabilistic Alignment of Relations, Instances, and Schema (VLDB 2012)
  - Relation-based
    - Embedding方法
      - Iterative Entity Alignment via Joint Knowledge Embeddings(IJCAI 2017)
  - Property & Relation-based
    - Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding(ISWC 2017)
  - Crowdsourcing-combined
    - Hike: A Hybrid Human-Machine Method for Entity Alignment(CIKM 2017)

# 实体对齐

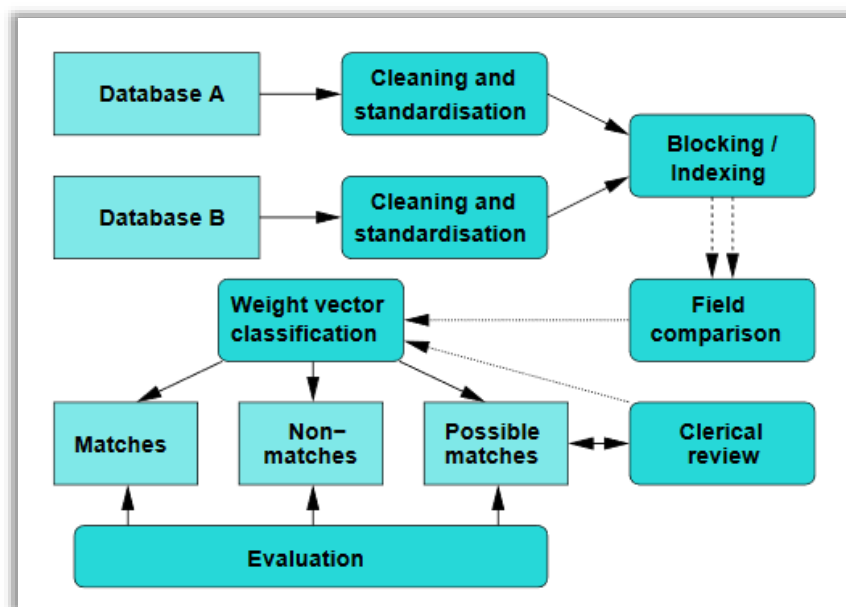
- 主流方法
  - Property-based
    - 机器学习方法
      - Febrl – A Freely Available Record Linkage System with a Graphical User Interface(KDD 2008)
    - 基于概率
      - PARIS: Probabilistic Alignment of Relations, Instances, and Schema (VLDB 2012)
  - Relation-based
    - Embedding方法
      - Iterative Entity Alignment via Joint Knowledge Embeddings(IJCAI 2017)
  - Property & Relation-based
    - Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding(ISWC 2017)
  - Crowdsourcing-combined
    - Hike: A Hybrid Human-Machine Method for Entity Alignment(CIKM 2017)

# Property-based 之 机器学习方法

## • Motivation

- 对于不同领域的实体，用于实体匹配的特征可能不同，为了能够学习到不同的匹配规则，提出利用已有的alignment seeds，通过机器学习的方法习得个性化的匹配规则，以解决领域之间的差异性问题。

## • Framework



# 实体对齐

## • 主流方法

- Property-based

- 机器学习方法

- Febrl – A Freely Available Record Linkage System with a Graphical User Interface(KDD 2008)

- 基于概率

- PARIS: Probabilistic Alignment of Relations, Instances, and Schema (VLDB 2012)

- Relation-based

- Embedding方法

- Iterative Entity Alignment via Joint Knowledge Embeddings(IJCAI 2017)

- Property & Relation-based

- Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding(ISWC 2017)

- Crowdsourcing-combined

- Hike: A Hybrid Human-Machine Method for Entity Alignment(CIKM 2017)

# Property-based 之 基于概率

## • Motivation

- 机器学习的方法往往需要人工标注alignment seeds作为训练数据，为了解决训练数据匮乏及参数调节困难的问题，提出基于概率去衡量两个实体的匹配程度。

## • Idea

- 两实体匹配的概率公式：
$$Pr_1(x \equiv x') = 1 - \prod_{\substack{r(x,y) \\ r(x',y')}} (1 - fun^{-1}(r) \times Pr(y \equiv y')).$$
- 两实体不匹配的概率公式：
$$Pr_2(x \equiv x') = 1 - \prod_{r(x,y)} (1 - fun^{-1}(r) \prod_{r(x',y')} (1 - Pr(y \equiv y'))).$$
- 最终的实体匹配概率公式：
$$Pr_3(x \equiv x') = Pr_1(x \equiv x') \times Pr_2(x \equiv x').$$

- 其中， $(x, r, y)$ 为三元组， $fun^{-1}(r)$ 为逆函数性，其大小表明事实三元组的同一关系中宾语相等对主语相等的决定能力，逆函数值越大，则在宾语相等的情况下，主语相等的可能性越大。

# 实体对齐

## • 主流方法

- Property-based

- 机器学习方法

- Febrl – A Freely Available Record Linkage System with a Graphical User Interface(KDD 2008)

- 基于概率

- PARIS: Probabilistic Alignment of Relations, Instances, and Schema (VLDB 2012)

- Relation-based

- Embedding方法

- Iterative Entity Alignment via Joint Knowledge Embeddings(IJCAI 2017)

- Property & Relation-based

- Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding(ISWC 2017)

- Crowdsourcing-combined

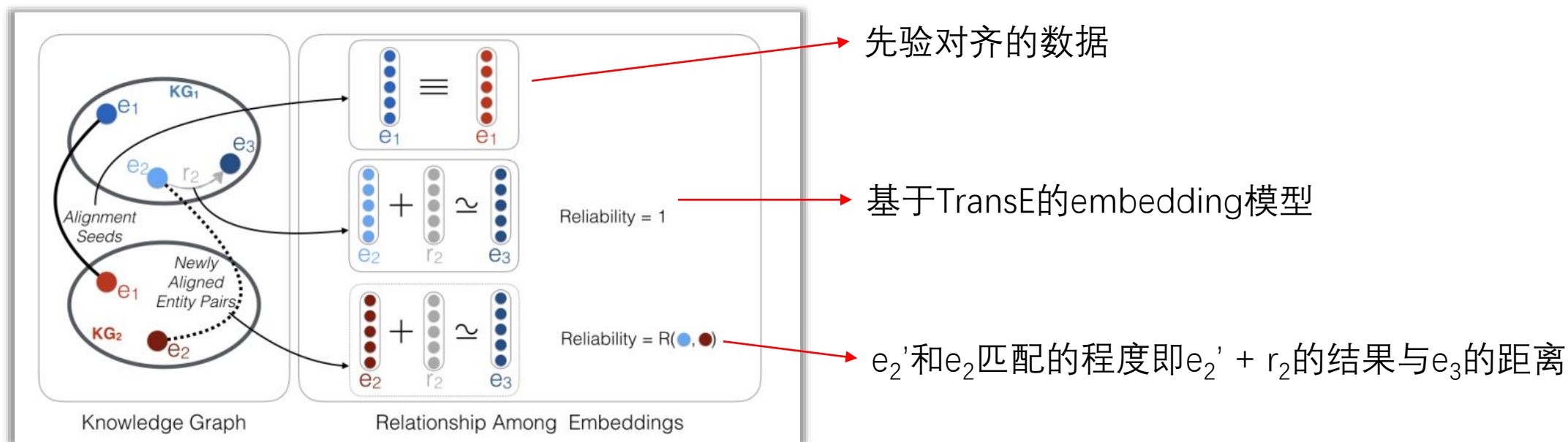
- Hike: A Hybrid Human-Machine Method for Entity Alignment(CIKM 2017)

# Relation-based

- Motivation

- Property-based方法的效果往往受限于属性表达方式多样性，而基于关系的structure embedding则很好地避开了这一障碍，通过将实体及实体间关系映射到同一向量空间中，以衡量实体间的距离。

- Framework





# 实体对齐

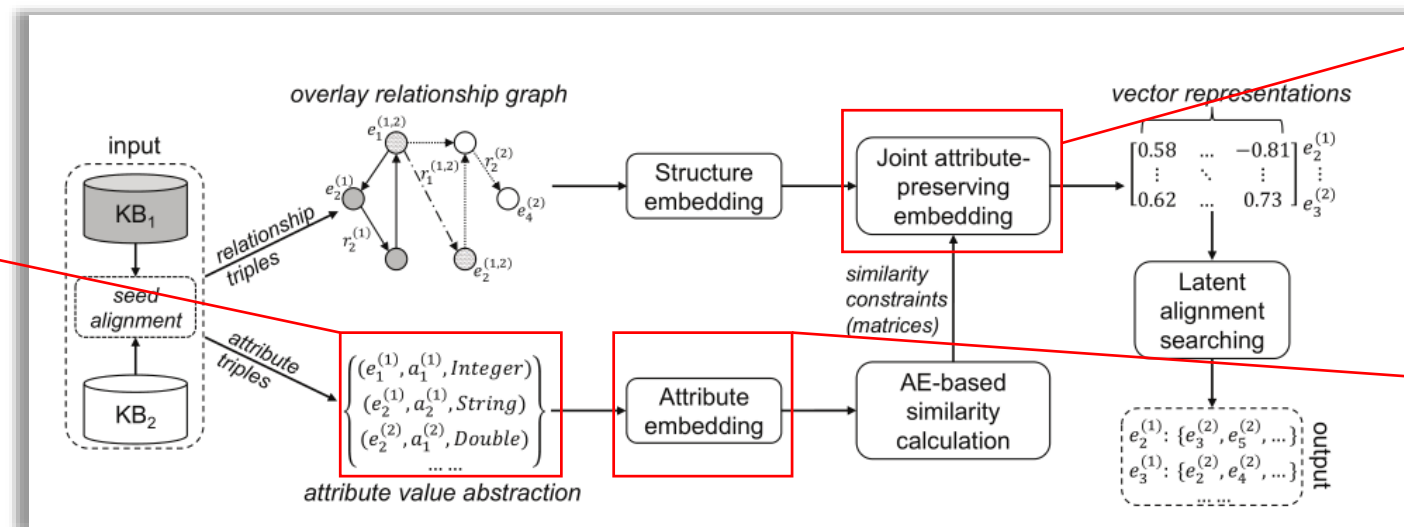
- 主流方法
  - Property-based
    - 机器学习方法
      - Febrl – A Freely Available Record Linkage System with a Graphical User Interface(KDD 2008)
    - 基于概率
      - PARIS: Probabilistic Alignment of Relations, Instances, and Schema (VLDB 2012)
  - Relation-based
    - Embedding方法
      - Iterative Entity Alignment via Joint Knowledge Embeddings(IJCAI 2017)
  - Property & Relation-based
    - Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding(ISWC 2017)
  - Crowdsourcing-combined
    - Hike: A Hybrid Human-Machine Method for Entity Alignment(CIKM 2017)

# Property & Relation-based

## • Motivation

- 主流的Embedding方法通常只考虑实体的relationship，而属性三元组 (attribute triple) 则被忽略，而对于那些relationship很稀疏的实体，只用structure embedding可能缺乏有效信息，通过attribute embedding模型引入属性信息，则更加有利于实体匹配。

## • Framework



最终的实体相似  
度为SE与AE分别  
得到的相似度的  
结合:  $\mathcal{O}_{joint} = \mathcal{O}_{SE} + \delta \mathcal{O}_S$

实体的embedding  
为该实体所有属性  
embedding的平均  
值。

将属性三元组(S,P,O)中的O  
简化成数据类型，如  
Integer、String、Double等，  
再对P进行embedding，目  
标是让那些经常为同一实  
体所拥有的属性之间的距  
离更近。

# 实体对齐

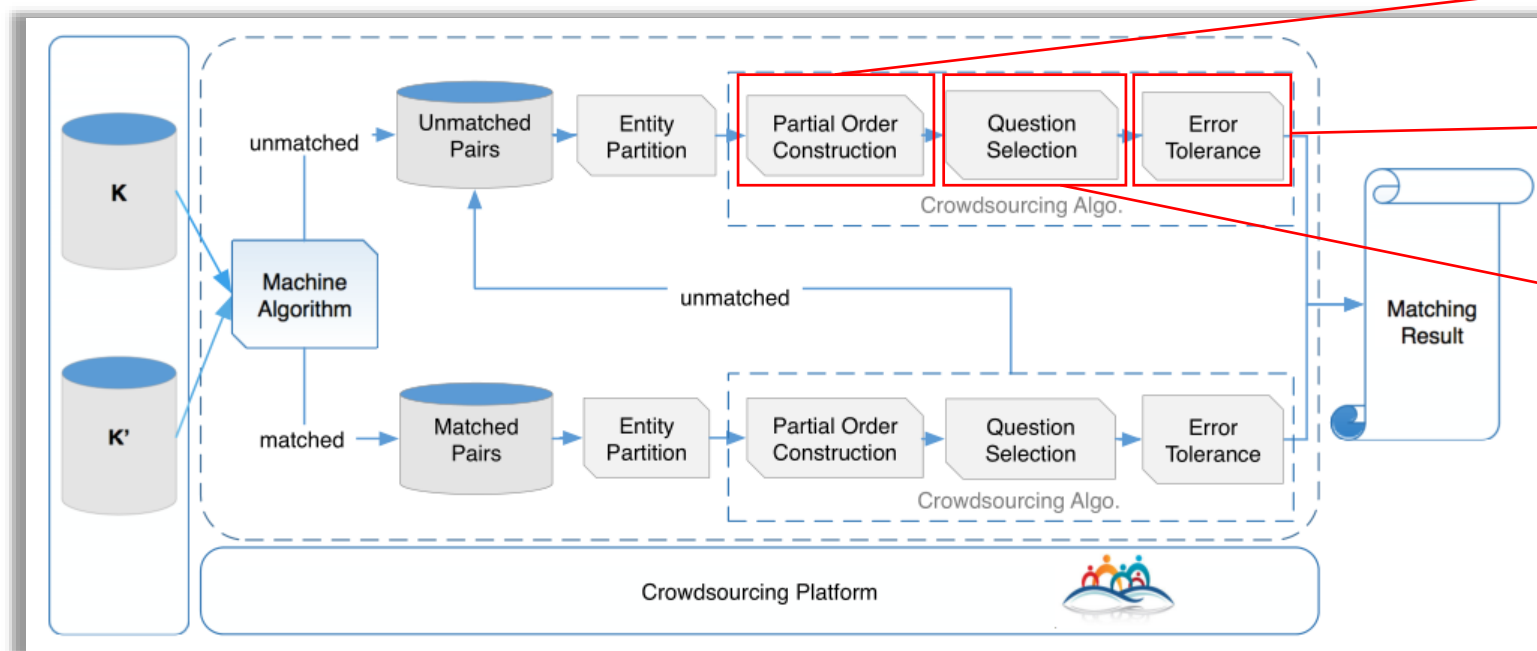
- 主流方法
  - Property-based
    - 机器学习方法
      - Febrl – A Freely Available Record Linkage System with a Graphical User Interface(KDD 2008)
    - 基于概率
      - PARIS: Probabilistic Alignment of Relations, Instances, and Schema (VLDB 2012)
  - Relation-based
    - Embedding方法
      - Iterative Entity Alignment via Joint Knowledge Embeddings(IJCAI 2017)
  - Property & Relation-based
    - Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding(ISWC 2017)
  - Crowdsourcing-combined
    - Hike: A Hybrid Human-Machine Method for Entity Alignment(CIKM 2017)

# Crowdsourcing-combined

- Motivation

- 为了弥补自动化实体对齐方法召回率低的缺点，本文提出可借助众包平台提升对齐效果

- Framework



找出最具有推理期望  
(Inference Expectation)  
的实体对

容错机制

利用匹配对之间的传递关系进行推理，以减少众包问题数量，使得众包代价最小化

# Crowdsourcing-combined

## • 偏序模型

- 根据以下规则建立右图所示偏序图

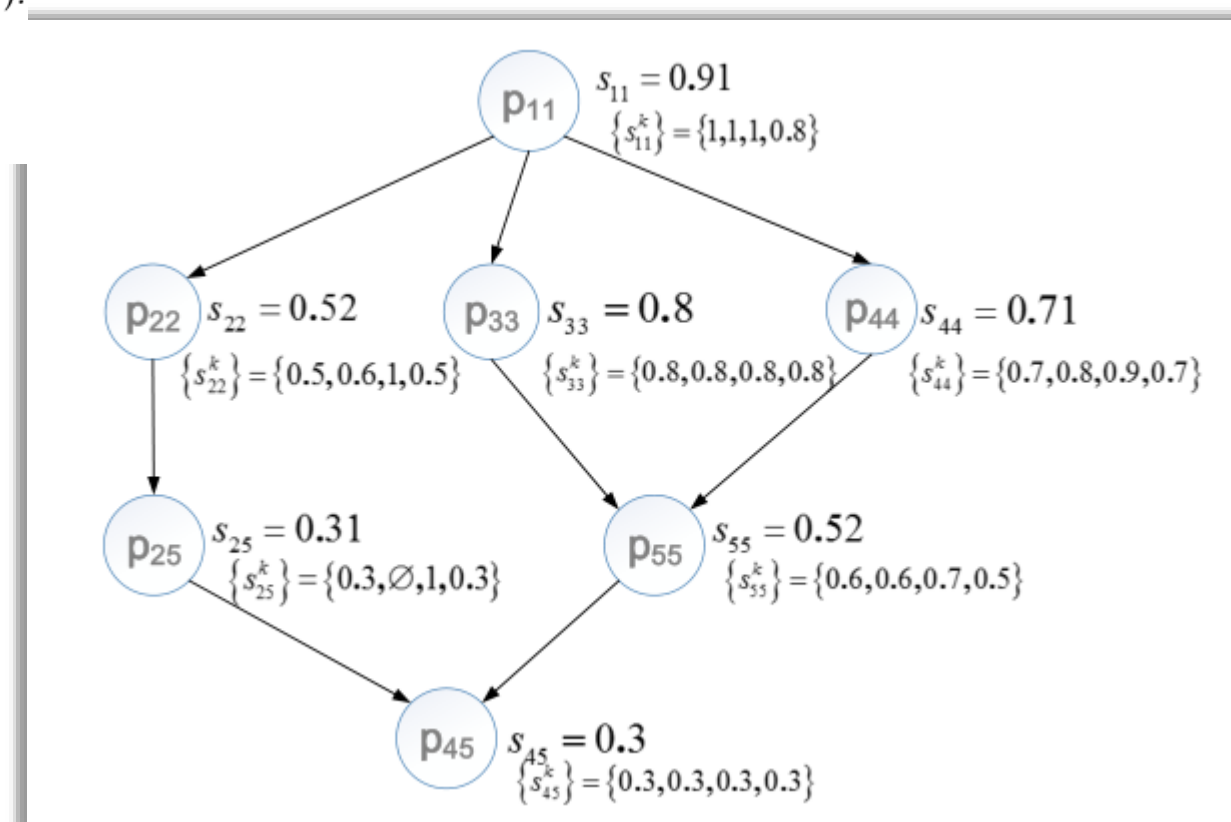
DEFINITION 5 (PARTIAL ORDER). The partial order  $<$  can be formally defined as follows: Given two pairs  $p_{ij} = (e_i, e_j), p_{i'j'} = (e_{i'}, e_{j'})$ . For each  $pp^k, s_{ij}^k \geq s_{i'j'}^k$ , and  $\sum_k(s_{ij}^k) > \sum_k(s_{i'j'}^k)$ , then  $p_{ij} > p_{i'j'}$ , we say  $p_{ij}$  and  $p_{i'j'}$  are comparable and  $p_{ij}$  precedes  $p_{i'j'}$ , or  $p_{i'j'}$  succeeds  $p_{ij}$ .

- 推理期望公式如下：

$$E(p_{ij}) = s_{ij} \cdot |\text{pre}(p_{ij})| + (1 - s_{ij}) \cdot |\text{suc}(p_{ij})|$$

- 其中，pre和suc分别表示前驱和后继结点， $s_{ij}^k$ 表示在第k个属性对匹配的情况下 $p_{ij}$ 的相似度，

$s_{ij} = \sum_{k=1}^m \omega_k \cdot s_{ij}^k$  表示实体相似度， $w_k$  为第k个属性对对于实体相似度计算的权重。



# 属性对齐

- 意义
  - 其结果可作为实体对齐及本体构建的基础
  - 完善的属性对应关系有利于提高语义检索、问答系统的召回率
- 挑战
  - Web信息的不完整、噪声多等特性
  - 中文表意的灵活性使得属性间的关系尤为复杂
- 主流方法
  - 人工建立属性映射表
  - 基于属性的扩展(extension), 针对关系型属性
    - 对于三元组(S, P, O), 若O为实体, 则称P为关系型属性, (S, O)即为P的扩展
    - 利用已有的实体匹配结果, 通过计算匹配数与共现数的比例确定同义属性
  - 基于属性值相似度, 针对非关系型属性
    - 对于三元组(S, P, O), 若O不是实体, 则称P为非关系型属性
    - 使用聚类法获取单一数据集内部的同义属性簇

# 属性值归一化

- 意义
  - 消除属性值不同量级以及不同表达方式的影响，使数据具有一致性
  - 属性值规范化后，更加有利于实体对齐和属性对齐
  - 更加能够满足问答系统的需求
- 挑战
  - 表达方式的多样化，无法100%覆盖所有形式进行归一
- 方法
  - 将属性值分类
    - 数据类型
    - 表达规律
  - 按类别制定统一表达规则
    - 以常用的表达习惯为标准
    - 半自动化抽取固定的Pattern
  - 按规则进行归一化

# 本节大纲

- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
  - 关系数据库中的数据融合统一
  - 知识图谱中的知识融合统一
  - 知识图谱中的知识链接融入
- 知识图谱数据补全的质量控制
- 知识图谱数据更新的质量控制





# 知识图谱中的知识链接与融入

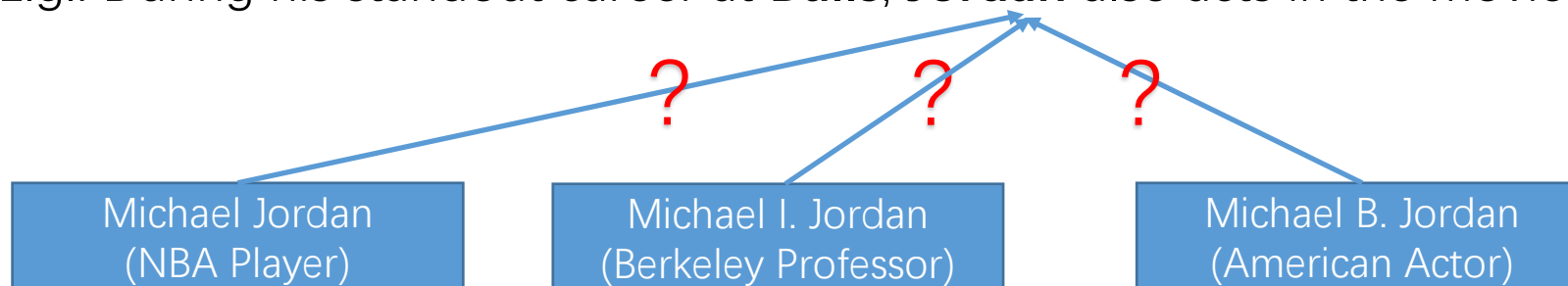
- **知识链接与融入：** 将获取的各类知识“链接”到知识图谱
  - 概念链接（量少，人工融入最准确）
  - 实体链接（**关键问题，研究热点：刚需、海量、歧义性大**）
  - 属性链接（实体链接正确了，属性链接相对简单些）
  
- **实体链接**
  - 实体链接是解决命名实体歧义问题的一种重要方法，该方法通过将具有歧义的实体指称项链接到给定的知识库中从而实现实体歧义的消除。
  - 难点：**一词多义，多词一义**

# 实体链接 (Entity Linking)

Also known as Entity Recognition and Disambiguation

## 1. Polysemy (一词多义)

E.g.: During his standout career at **Bulls**, **Jordan** also acts in the movies **Space Jam**.



## 2. Synonyms (多词一义)

- E.g.: Barack Hussein Obama(USA president)
  - m.02mjmr(Freebase)
  - Barack Obama(Dbpedia)
  - 贝拉克·侯赛因·奥巴马(CN-Dbpedia)

# 实体链接（Entity Linking） – Polysemy

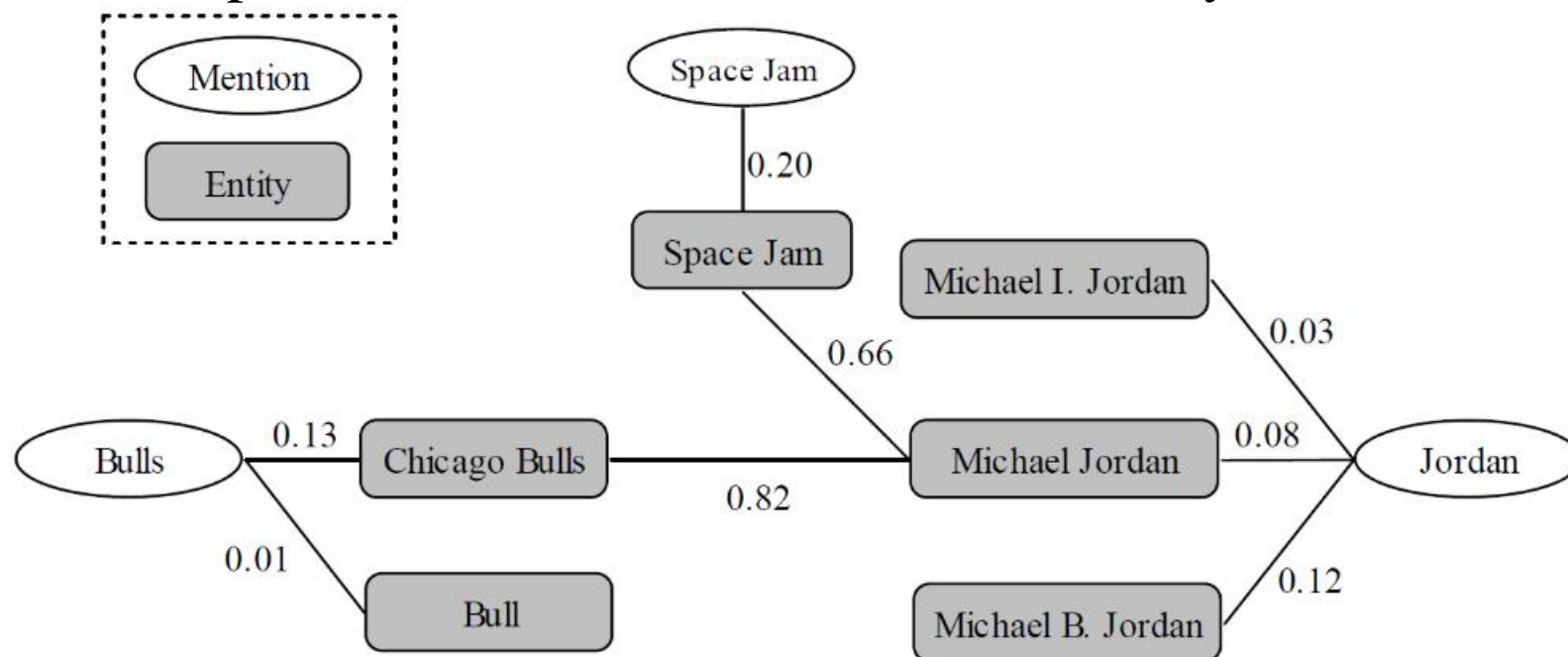
- 解决歧义性问题的主流方法
  - EL based on Local Compatibility (CIKM'07, EMNLP'07, IJCAI'09, COLING'10...)
  - EL Based on Simple Relations (CIKM'08, AAAI'08)
  - Pair-Wise Collective EL Approaches (ACL'10)
  - Graph-Based Collective EL Approaches (SIGIR'11, 14)
  - Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation(IJCAI'2015)
  - Deep Joint Entity Disambiguation with Local Neural Attention(EMNLP'17)
  - Neural Cross-Lingual Entity Linking(AAAI,18)

# 实体链接（Entity Linking） – Polysemy

- 解决歧义性问题的主流方法
  - EL based on Local Compatibility (CIKM'07, EMNLP'07, IJCAI'09, COLING'10...)
  - EL Based on Simple Relations (CIKM'08, AAAI'08)
  - Pair-Wise Collective EL Approaches (ACL'10)
  - Graph-Based Collective EL Approaches (SIGIR'11, 14)
  - Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation(IJCAI'2015)
  - Deep Joint Entity Disambiguation with Local Neural Attention(EMNLP'17)
  - Neural Cross-Lingual Entity Linking(AAAI'18)

# 实体链接 (Entity Linking) – Polysemy

- Local Compatibility Based Approaches (CIKM'07, EMNLP'07, IJCAI'09, COLING'10...)
  - Idea:** Extract the discriminative features of an entity from its textual description, such as “NBA”, “Basketball Player” to MJ.



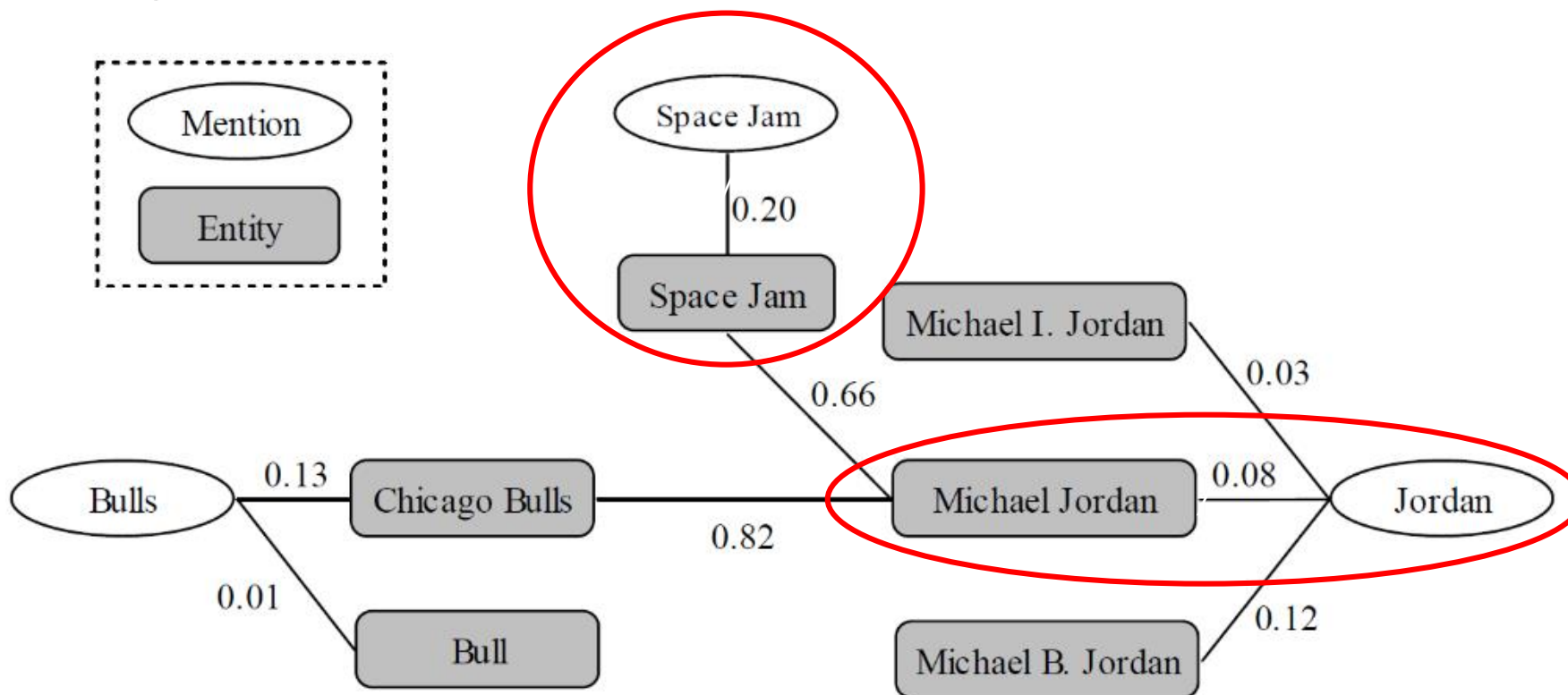
During his standout career at **Bulls**, **Jordan** also acts in the movies **Space Jam**.

# 实体链接（Entity Linking） – Polysemy

- 解决歧义性问题的主流方法
  - EL based on Local Compatibility (CIKM'07, EMNLP'07, IJCAI'09, COLING'10...)
  - EL Based on Simple Relations (CIKM'08, AAI'08)
  - Pair-Wise Collective EL Approaches (ACL'10)
  - Graph-Based Collective EL Approaches (SIGIR'11, 14)
  - Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation(IJCAI'2015)
  - Deep Joint Entity Disambiguation with Local Neural Attention(EMNLP'17)
  - Neural Cross-Lingual Entity Linking(AAI'18)

# 实体链接 (Entity Linking) – Polysemy

- Simple Relational Approaches (CIKM'08, AAAI'08)
  - **Idea:** the referent entity of a name mention should be coherent with its unambiguous contextual entities



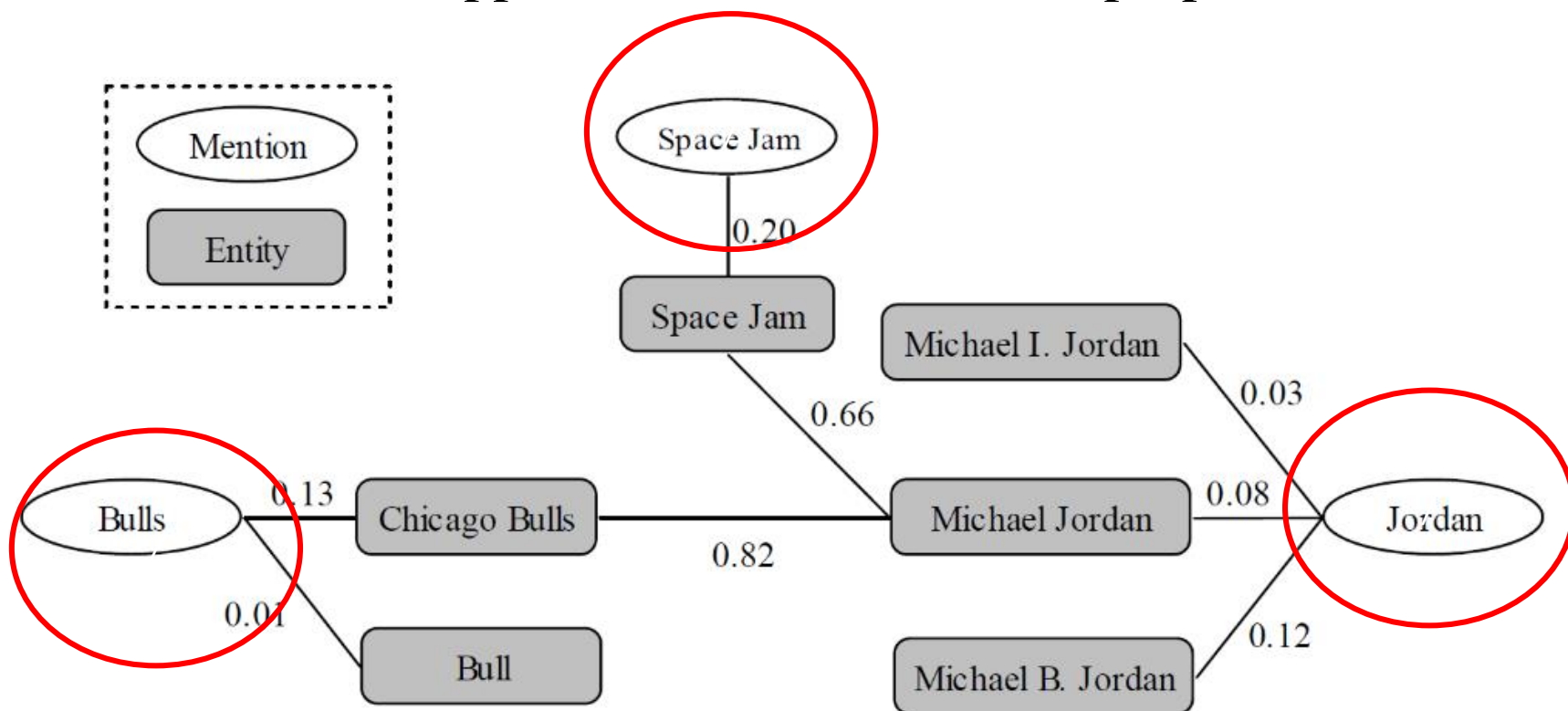
# 实体链接（Entity Linking） – Polysemy

- 解决歧义性问题的主流方法
  - EL based on Local Compatibility (CIKM'07, EMNLP'07, IJCAI'09, COLING'10...)
  - EL Based on Simple Relations (CIKM'08, AAAI'08)
  - **Pair-Wise Collective EL Approaches (ACL'10)**
  - Graph-Based Collective EL Approaches (SIGIR'11, 14)
  - Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation(IJCAI'2015)
  - Deep Joint Entity Disambiguation with Local Neural Attention(EMNLP'17)
  - Neural Cross-Lingual Entity Linking(AAAI'18)



# 实体链接（Entity Linking） – Polysemy

- Pair-Wise Collective Approaches (ACL'10)
  - **Idea:** Model and exploit the pair-wise interdependence between EL decisions (NP-HARD), and approximation solutions are proposed.

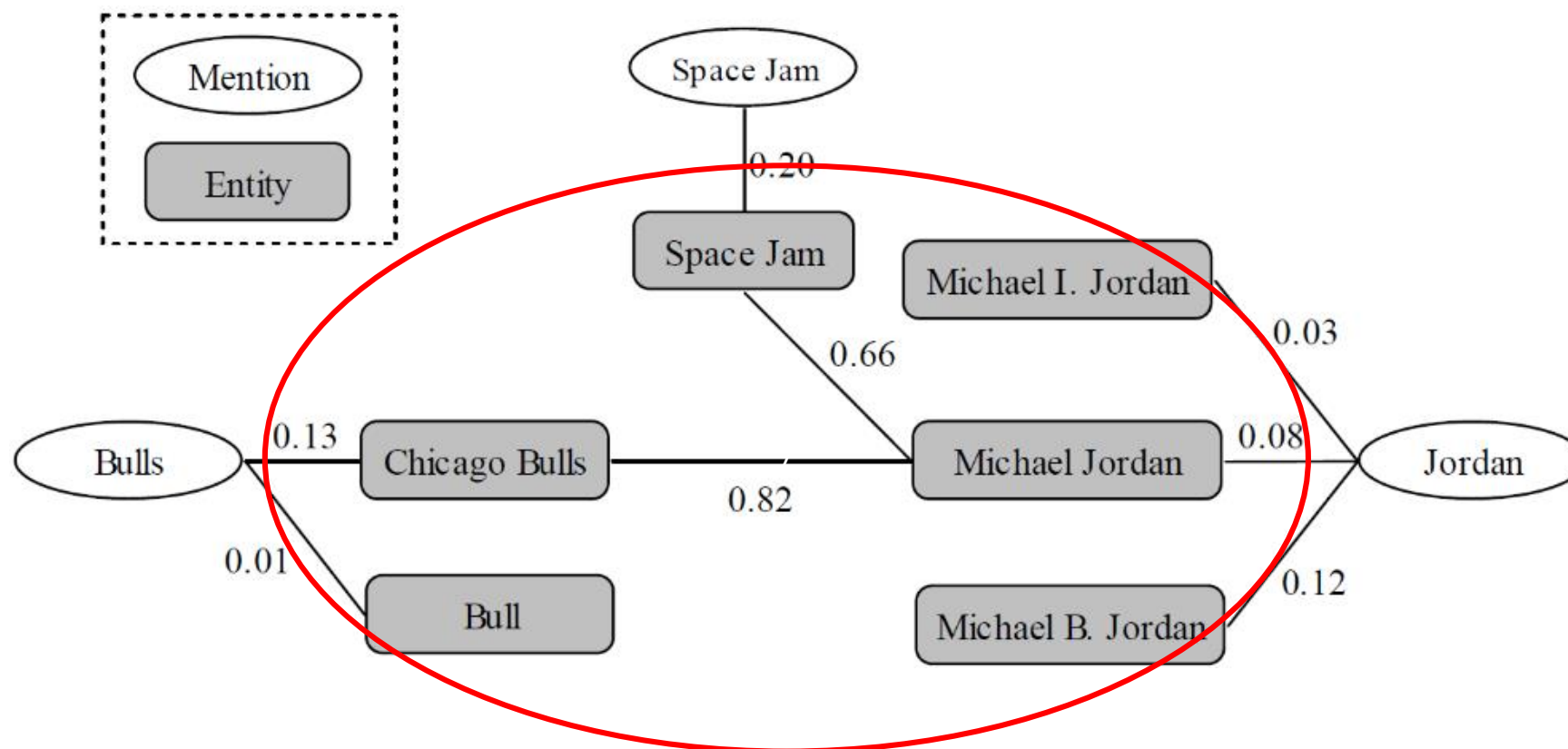


# 实体链接（Entity Linking） – Polysemy

- 解决歧义性问题的主流方法
  - EL based on Local Compatibility (CIKM'07, EMNLP'07, IJCAI'09, COLING'10...)
  - EL Based on Simple Relations (CIKM'08, AAAI'08)
  - Pair-Wise Collective EL Approaches (ACL'10)
  - Graph-Based Collective EL Approaches (SIGIR'11, 14)
  - Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation(IJCAI'2015)
  - Deep Joint Entity Disambiguation with Local Neural Attention(EMNLP'17)
  - Neural Cross-Lingual Entity Linking(AAAI'18)

# 实体链接 (Entity Linking) – Polysemy

- Graph-Based Collective Approaches(SIGIR 11,14)
  - Idea:** Model and exploit the global interdependence by graph-based collective EL method

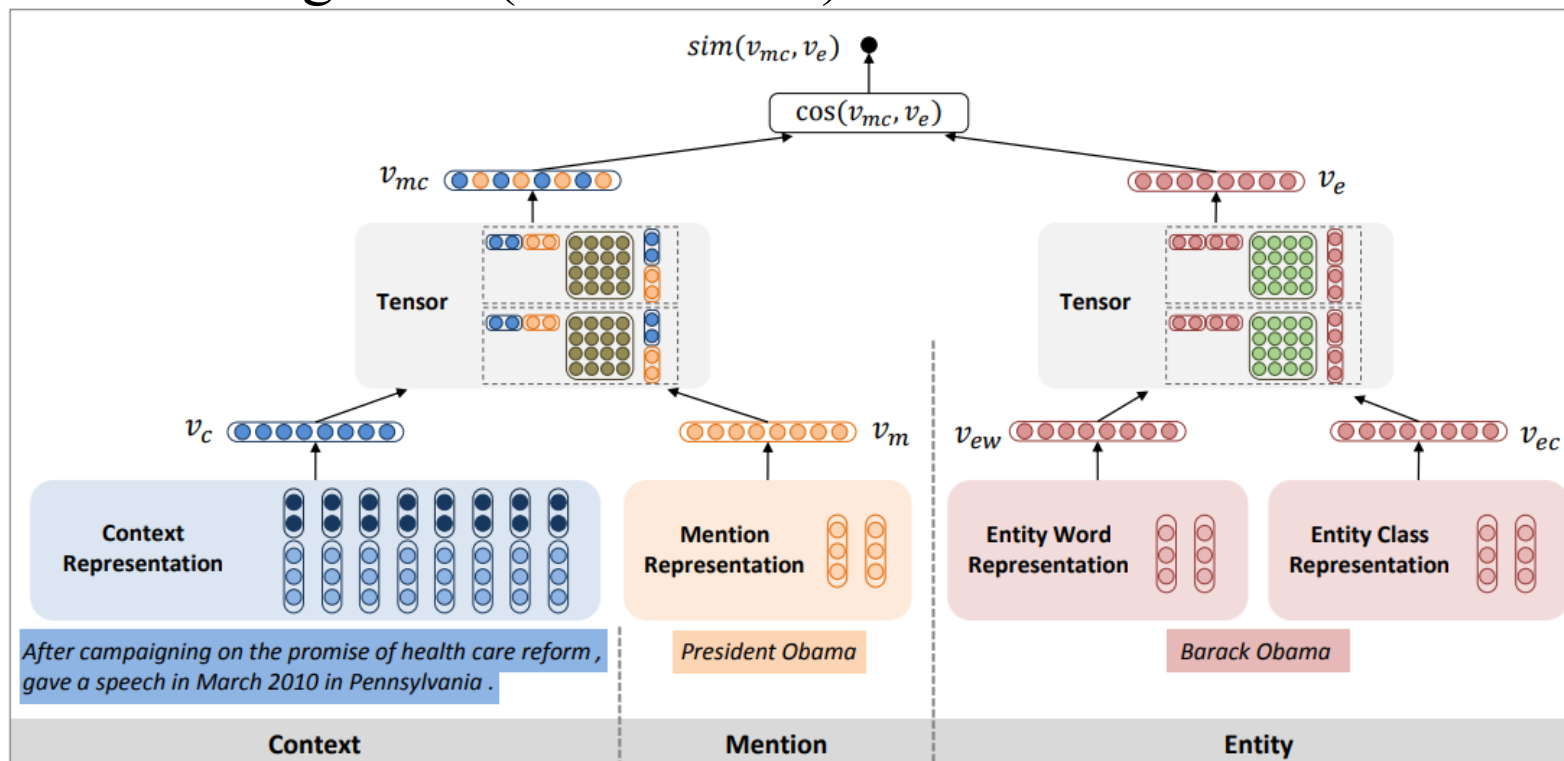


# 实体链接（Entity Linking） – Polysemy

- 解决歧义性问题的主流方法
  - EL based on Local Compatibility (CIKM'07, EMNLP'07, IJCAI'09, COLING'10...)
  - EL Based on Simple Relations (CIKM'08, AAAI'08)
  - Pair-Wise Collective EL Approaches (ACL'10)
  - Graph-Based Collective EL Approaches (SIGIR'11, 14)
  - Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation(IJCAI'2015)
  - Deep Joint Entity Disambiguation with Local Neural Attention(EMNLP'17)
  - Neural Cross-Lingual Entity Linking(AAAI'18)

# 实体链接 (Entity Linking) – Polysemy

- Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation(IJCAI'2015)



Ideas:

- ① embed mention, context and entity in continuous vector space to capture their semantic representations.
- ② The variable-sized context are modeled with convolutional neural networks.

Example:

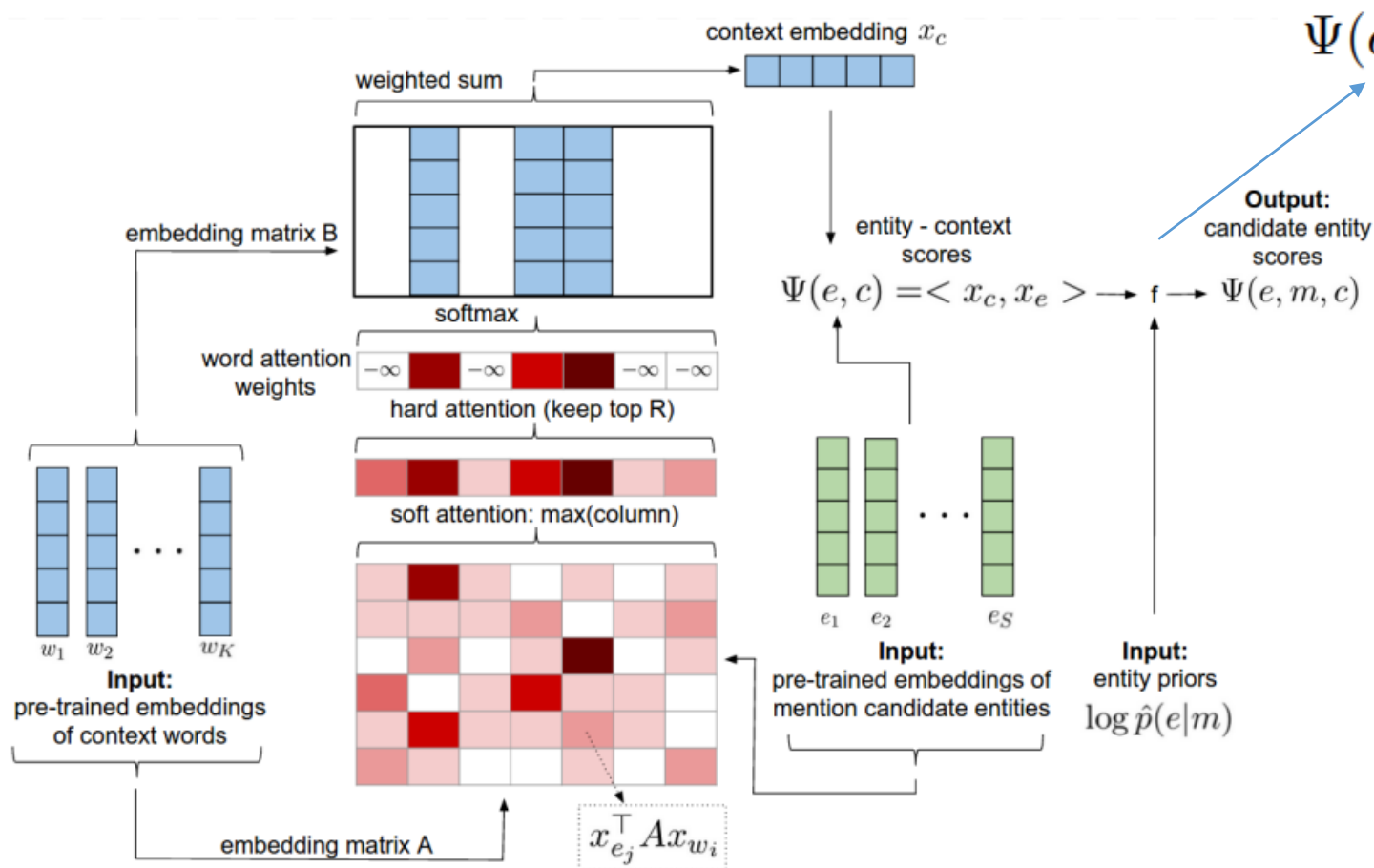
After campaigning on the promise of health care reform, **President Obama** gave a speech in March 2010 in Pennsylvania

# 实体链接（Entity Linking） – Polysemy

- 解决歧义性问题的主流方法
  - EL based on Local Compatibility (CIKM'07, EMNLP'07, IJCAI'09, COLING'10...)
  - EL Based on Simple Relations (CIKM'08, AAI'08)
  - Pair-Wise Collective EL Approaches (ACL'10)
  - Graph-Based Collective EL Approaches (SIGIR'11, 14)
  - Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation(IJCAI'2015)
  - Deep Joint Entity Disambiguation with Local Neural Attention(EMNLP'17)
  - Neural Cross-Lingual Entity Linking(AAI'18)

# 实体链接（Entity Linking）– Polysemy

- Deep Joint Entity Disambiguation with Local Neural Attention(EMNLP'17)



Ideas:

- ① entity embeddings
- ② a neural attention mechanism over local context windows,

Inputs:

- ① context word vectors
- ② candidate entity priors
- ③ candidate entity embeddings

Outputs:

- ① entity scores

# 实体链接（Entity Linking） – Polysemy

- 解决歧义性问题的主流方法
  - EL based on Local Compatibility (CIKM'07, EMNLP'07, IJCAI'09, COLING'10...)
  - EL Based on Simple Relations (CIKM'08, AAAI'08)
  - Pair-Wise Collective EL Approaches (ACL'10)
  - Graph-Based Collective EL Approaches (SIGIR'11, 14)
  - Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation(IJCAI'2015)
  - Deep Joint Entity Disambiguation with Local Neural Attention(EMNLP'17)
  - Neural Cross-Lingual Entity Linking(AAAI'18)

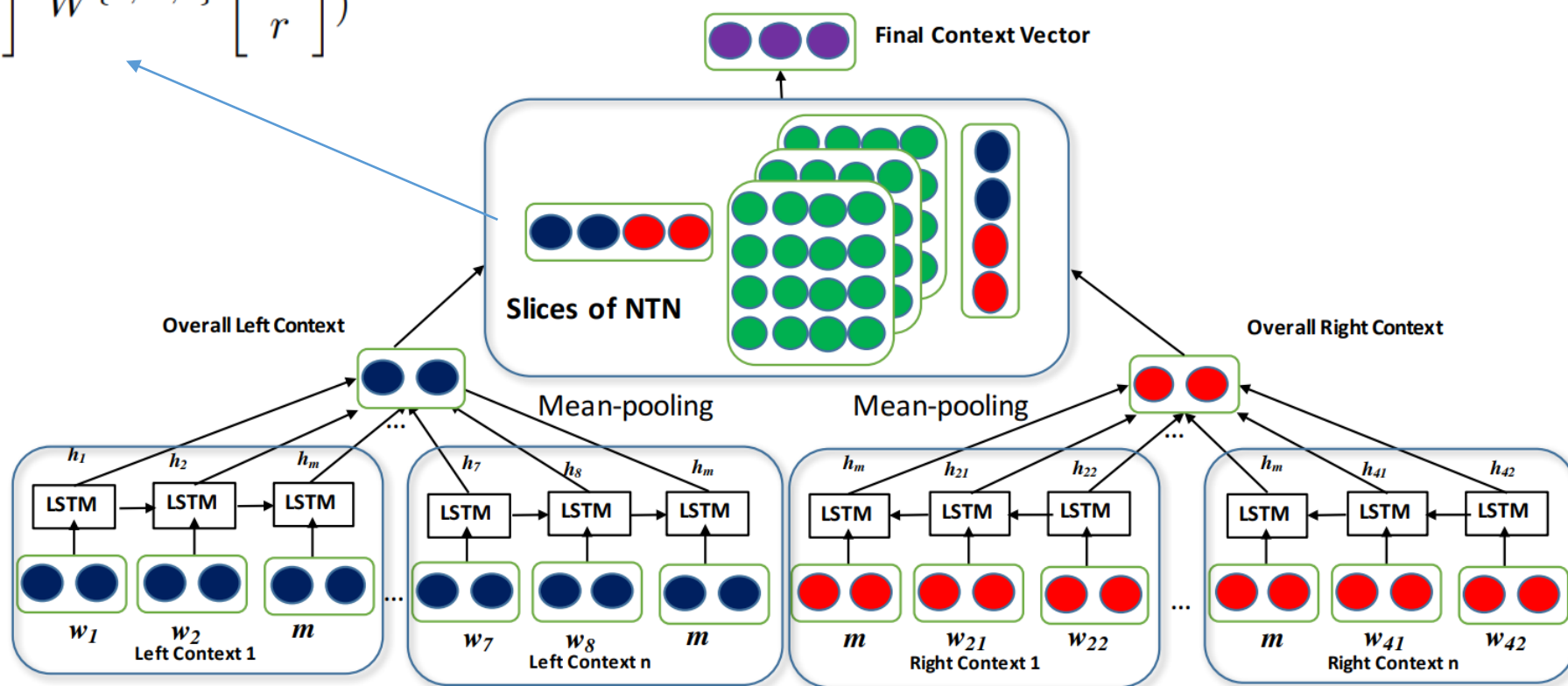


# 实体链接 (Entity Linking) – Polysemy

- Neural Cross-Lingual Entity Linking(AAAI,18)

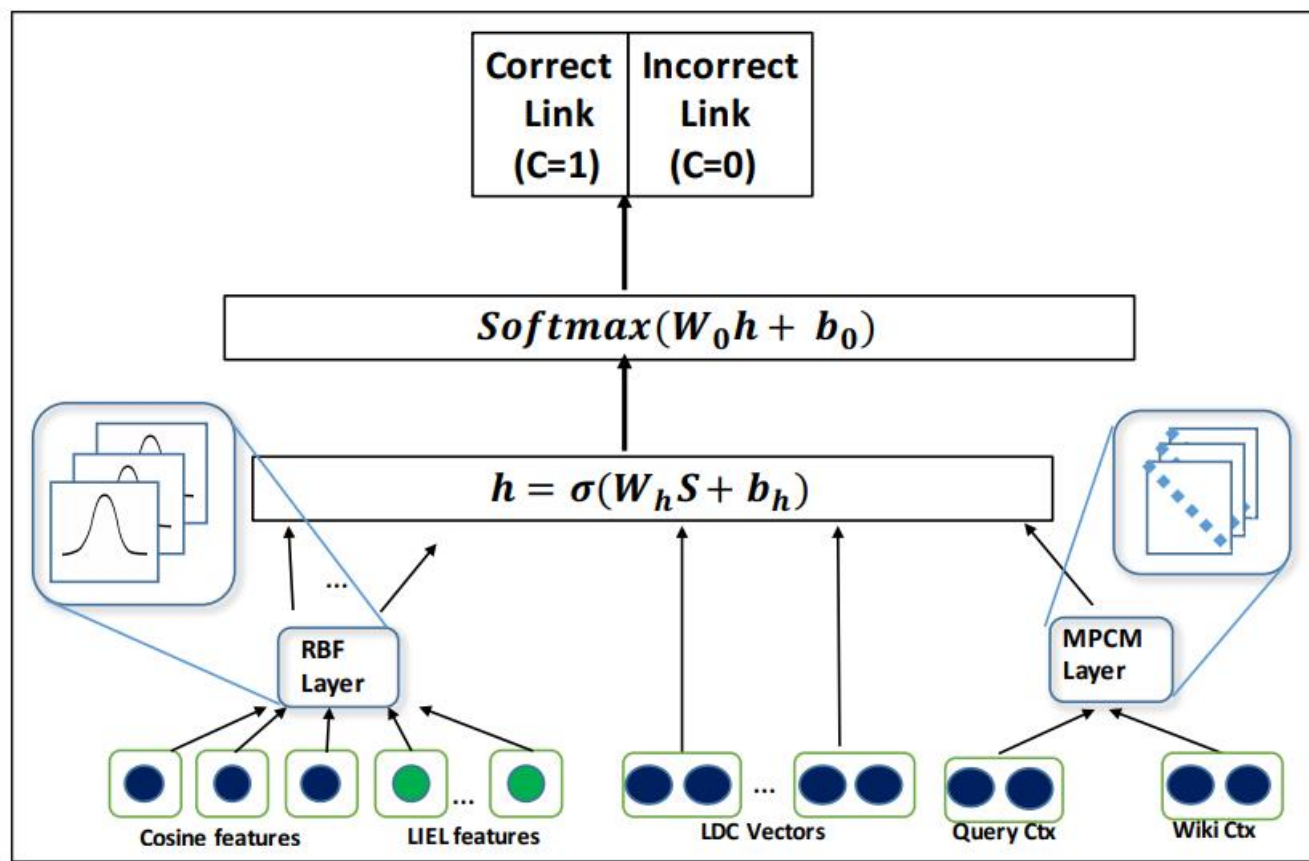
$$NTN(l, r; W) = f\left(\begin{bmatrix} l \\ r \end{bmatrix}^t W^{\{1, \dots, k\}} \begin{bmatrix} l \\ r \end{bmatrix}\right)$$

Ideas:  
consider context to be  
the words surrounding  
a mention within a  
window of length n



# 实体链接（Entity Linking） – Polysemy

- Neural Cross-Lingual Entity Linking(AAAI,18)



Ideas:

## LIEL

- ① “how many words overlap between the mention and Wikipedia title match?”
- ② “how many outlink names of the candidate Wikipedia title appear in the query document?”

## LDC

$S = [s_1, \dots, s_m]$ ,  $T = [t_1, \dots, t_n]$

S:source context, T:Wikipedia paragraph T

For each word  $s_i$  in S, finds a matching word  $\hat{s}_i$  from T, for each  $t_j$  in T, finds a matching word  $\hat{t}_j$  in S.

## MPCM

train weight vectors to re-weigh the dimensions of the input vectors and then compute the cosine similarity.

# 实体链接 (Entity Linking)

Also known as Entity Recognition and Disambiguation

## 1. Polysemy (一词多义)

E.g.: During his standout career at **Bulls**, **Jordan** also acts in the movies **Space Jam**.

## 2. Synonyms (多词一义)

E.g.: Barack Hussein Obama(USA president)

- m.02mjmr(Freebase)
- Barack\_Obama(Dbpedia)
- 贝拉克·侯赛因·奥巴马(CN-Dbpedia)

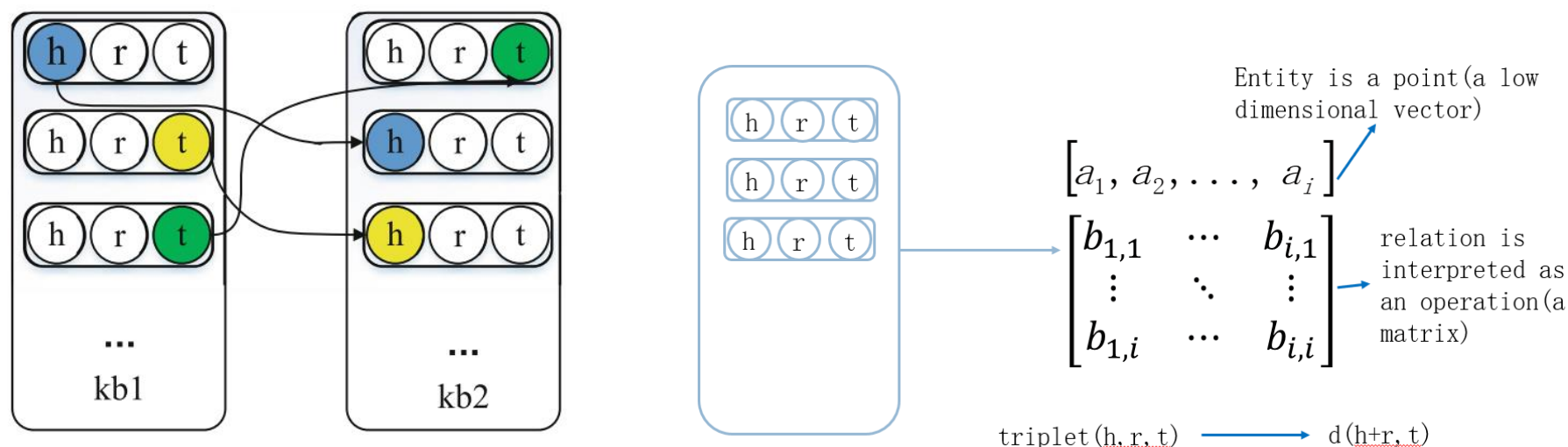
# 实体链接（Entity Linking）– Synonyms

- **Approaches for Solving Synonym Problems**

- String-matching based methods (CITISIA'09)
  - Edit Distance, Jaccard, Cosine, Hybrid Metrics...
- Collective alignment methods (VLDB'11, SIGKDD'13)
  - Use various information of entities such as *Properties*, *Relations*, *Instances* to construct a probabilistic matching model
- Based on structure similarity only (CCKS'16)
  - Whole Knowledge Base Embedding
- Top-k String Auto-Completion with Synonyms.(DASFAA'2017)
  - trie-based algorithms, auto-completion
- Automatic Induction of Synsets from a Graph of Synonyms. (ACL'2017)
  - graph-based approach that induces synsets using synonymy dictionaries and word embeddings.

# 实体链接 (Entity Linking) – Synonyms

- Based on structure similarity only(CCKS 16)
  - Idea:** (1)give some initial alignments(seed entity alignments); (2) learn the embedding of the two KBs in a uniform embedding vector space connected by the seed entities “bridge”



**Fig. 2.** Selecting seed entities in two KBs.

# 实体链接（Entity Linking） – Synonyms

- Top-k String Auto-Completion with Synonyms.(DASFAA’2017)
  - propose three data structures to support efficient top-k completion queries with synonyms for different space and time complexity trade-offs.

Search string:abmp

Iter.	$p_{ra}$	$p_{rr}$	Note
1	Pop first element from queue: $m = \varepsilon$ (root of $\mathcal{T}_D$ ), $p_r = abmp$		
1.1	$\varepsilon \checkmark$	$abmp \times$	$\varepsilon$ is found in $\mathcal{T}_D$ , but $abmp$ is not found in $\mathcal{T}_R$ .
1.2	$a \checkmark$	$bmp \times$	$a$ is found in $\mathcal{T}_D$ , but $bmp$ is not found in $\mathcal{T}_R$ .
1.3	$ab \checkmark$	$mp \checkmark$	$mp$ is found in $\mathcal{T}_R$ . The target of its links are $\underline{c}$ and $ab\underline{c}$ . $ab\underline{c}$ is the correct link target. Push it to queue.
1.4	$abm \times$		Break loop.
2	Pop first element from queue: $m = abc$ , $p_r = \emptyset$		
2.1	$abc \checkmark$	$\emptyset$	Node $ab\underline{c}$ is a leaf, so add it to result set. $p_{rr}$ is empty, so push all children of $ab\underline{c}$ to queue (but it has no child).
3	The queue is empty. Therefore the final result is “abc”.		

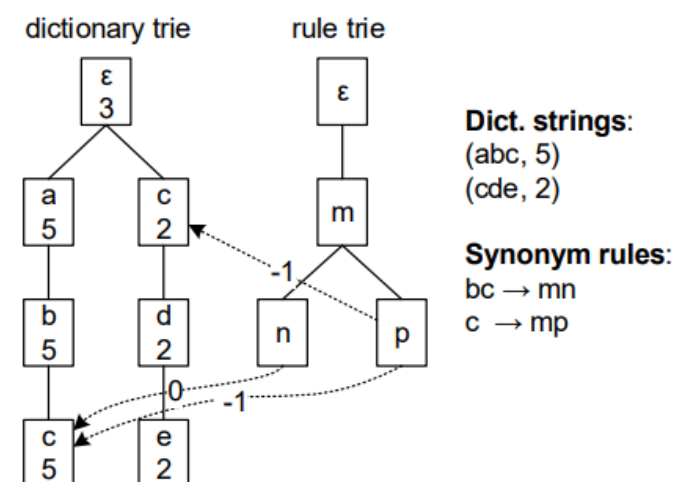




Fig. 2: TT example

# 实体链接（Entity Linking） – Synonyms

- Top-k String Auto-Completion with Synonyms.(DASFAA'2017)

Example:

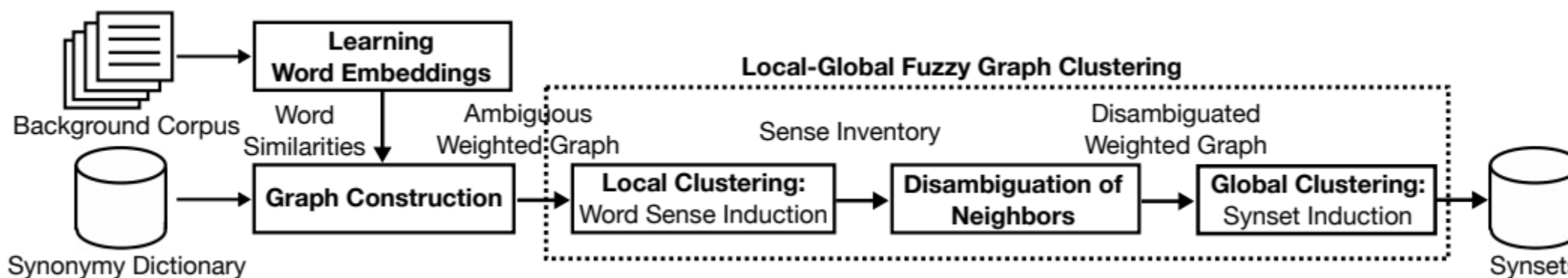
Andy Pa		IL2	
Andrew Pavlo		Interleukin-2	
Andrew Parker		Interleukin-2 biological activity	
Andrew Packard		Interleukin-2 and cancer	

Given three dictionary strings  $D$  including “Andrew Pavlo”, “Andrew Parker” and “Andrew Packard” and one synonym rule  $R = \{“Andy” \rightarrow “Andrew”\}$ . If a user enters “Andy Pa”. Then all three strings are returned as top-3 completions.



# 实体链接（Entity Linking）– Synonyms

- Automatic Induction of Synsets from a Graph of Synonyms. (ACL'2017)
  - ① build a weighted graph of synonyms extracted from commonly available resources. ② apply word sense induction to deal with ambiguous words. ③ cluster the disambiguated version of the ambiguous input graph into synsets.



**Figure 1:** Outline of the WATSET method for synset induction.



# 知识图谱数据补全的质量控制

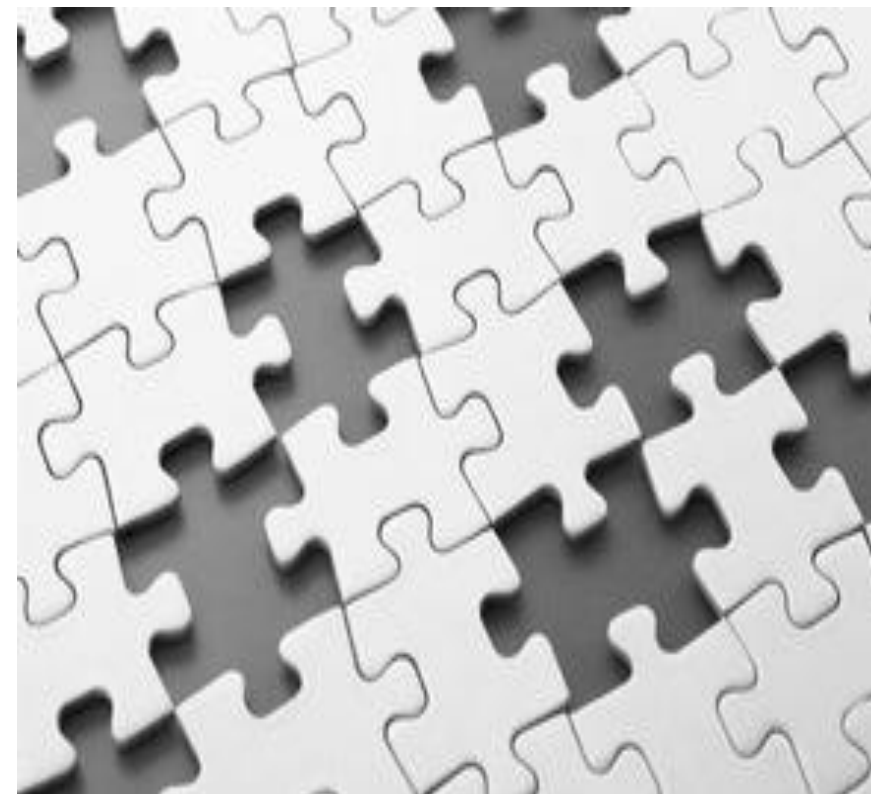
---

# 本节大纲

- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
- 知识图谱数据补全的质量控制
  - 关系数据库中的缺失数据补全技术
  - 知识图谱中的实体类型补全技术
  - 知识图谱中的实体关系补全技术
  - 知识图谱中的实体属性值补全技术
- 知识图谱数据更新的质量控制

# 什么是缺失数据 (Missing Data)

- 缺失数据：因各种原因应得而没有得到的数据
- 广义：各个维度的数据缺失
  - 数据库缺失
  - 数据表缺失
  - 属性缺失
  - 元组缺失
  - 属性值缺失
- 狭义：属性值缺失问题



# 数据表中的属性值缺失

Microsoft Excel - Summary\_ERC\_ProcRepresMethods.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF

A1 Subject

	A	B	C	D	E	F
	Subject	SexF1M2	Age	WorkExperience	EASGEN1C	EASGEN2C
1	1	1	24	2	6	5
2	2	1	24		5	5
3	3	2	34	15	6	6
4	4	2			6	5
5	5	2	44	25	6	8
6	6	2	24	1	6	6
7	7	1	29	8	3	5
8	8	1	23	2	7	6
9	9	2	26	2	6	5
10	10	2			6	5
11	11	2	25		6	6
12	12	2	26	1	5	5
13	13	1		15	6	5
14	14	2	29	6	5	5
15	15	1	27	5	6	6
16	16	1	28	3	6	5
17	17	2	28	6	6	6
18	18	1	36	12	3	3
19	19	2			6	5
20	20	2	23	3	6	5
21	21	2	23	1	6	5
22	22	2	31	1	6	6
23	23	2	28	5	6	5
24	24	2	28		6	6
25	25	1	27	2	6	7
26	26	2	20		5	7
27	27	2	22	4	6	5
28	28	2		2	1	2
29	29	2		3	1	2
30	30	2	22	2	5	5
31	31	1	22	3	6	5
32	32	1	23	7	5	3

QuantRaw / QualRaw / QuantNoMissVal / QuantNoMissValPLS / EFA\_Alphas / Qu

含有缺失数据的数据表（数据集）是困扰数据统计、数据挖掘、行为研究等领域的常见问题。

# 缺失数据实际案例









- UCI, 一个作为机器学习领域基准数据库的数据集, 超过40%的数据库都含有缺失数据



Browse Through: 394 Data Sets

Table View List View

Default Task
Classification (289)
Regression (74)
Clustering (67)
Other (54)
Attribute Type
Categorical (37)
Numerical (244)
Mixed (55)
Data Type
Multivariate (306)
Univariate (16)
Sequential (40)
Time-Series (75)
Text (37)
Domain-Theory (22)
Other (21)
Area
Life Sciences (89)
Physical Sciences (47)
CS / Engineering (129)
Social Sciences (23)
Business (25)
Game (10)
Other (67)
# Attributes
Less than 10 (90)
10 to 100 (182)
Greater than 100 (67)
# Instances
Less than 100 (19)
100 to 1000 (137)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
 <a href="#">3D Road Network (North Jutland, Denmark)</a>	Sequential, Text	Regression, Clustering	Real	434874	4	2013
 <a href="#">AAAI 2013 Accepted Papers</a>	Multivariate	Clustering		150	5	2014
 <a href="#">AAAI 2014 Accepted Papers</a>	Multivariate	Clustering		399	6	2014
 <a href="#">Abalone</a>	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
 <a href="#">Abscissic Acid Signaling Network</a>	Multivariate	Causal-Discovery	Integer	300	43	2008
 <a href="#">Activities of Daily Living (ADLs) Recognition Using Binary Sensors</a>	Multivariate, Sequential, Time-Series	Classification, Clustering		2747		2013
 <a href="#">Activity Recognition from Single Chest-Mounted Accelerometer</a>	Univariate, Sequential, Time-Series	Classification, Clustering	Real			2014
 <a href="#">Activity Recognition system based on Multisensor data fusion (AReM)</a>	Multivariate, Sequential, Time-Series	Classification	Real	42240	6	2016

# 如何应对数据缺失？

## • 统计学界的做法

- 目的：尽量降低缺失数据造成的影响
- 手段
  - 删除法：直接删除缺失数据所在行
  - 加权调整法：降低缺失数据权重
  - 简单替代法
    - 特殊值 (unknown) 替代
    - 平均值 (mean) 替代
    - 就近 (close-fit) 替代
    - .....
- 一般仅适用于可计量数据

## • 数据库界的做法

- 目的：尝试找回并填补缺失数据 (Data Imputation)
- 适用于可计量和不可计量数据 (或看做：字符串型数据)



# 关系数据库中的数据补全技术

---

- Data Imputation based on Local (Small) Data
  - Model-based, Rule-based
- Data Imputation based on External (Big) Data
  - Crowdsourcing, Web List/Tables, Surface Web
- Hybrid Data Imputation Approaches
  - Web + Rules, Web + Crowd

# Model-based Imputation Overview

- 回归补全法 (Regression-based Imputation)
  - 线性回归 (Linear Regression) (CSDA95)
  - K近邻回归 (K-Nearest Neighbors) (HIS02)
  - 决策树补全 (Decision Tree-based) (KDD96)
- 极大似然估计法 (Maximum Likelihood)
  - 期望最大化法则 (Expectation Maximization, EM) (RSS77)
- 多重补全法：将数据集填m次，将各个结果加以综合。
  - PMM法 (Predictive Mean Matching) (SAGE97)
  - 趋势得分法 (Propensity Score) (SAGE00)
  - 马尔科夫链蒙特卡罗法 (Markov Chain Monte Carlo, MCMC) (ASTA08)



# Rule-based Imputation with FD/CFDs

- 规则 – 属性间的关系
  - **Functional Dependencies (FDs)**
    - $\{\text{country-code}, \text{area-code}\} \rightarrow \{\text{city}\}$  means “attribute *city* depend on attribute *country-code* and *area-code*”
  - **Conditional Functional Dependencies (CFDs)**
    - $\{\text{country-code}=01, \text{zip}\} \rightarrow \{\text{street}\}$

Country-code	Area-code	Phone	Name	Street	City	ZIP
01	908	1111111	Mike	Tree Ave.	NYC	07974
01	908	1234567	Rick		NYC	07974
01	212	5675765	Joe	Elm Str.		01202
01	212	2345155	Jim	Elm Str.	CA	01202
44	131	2455325		Oak Ave.	EDI	EH4 1DT
44	131		Lan	High St.	EDI	EH4 1DT

# 关系数据库中的数据补全技术

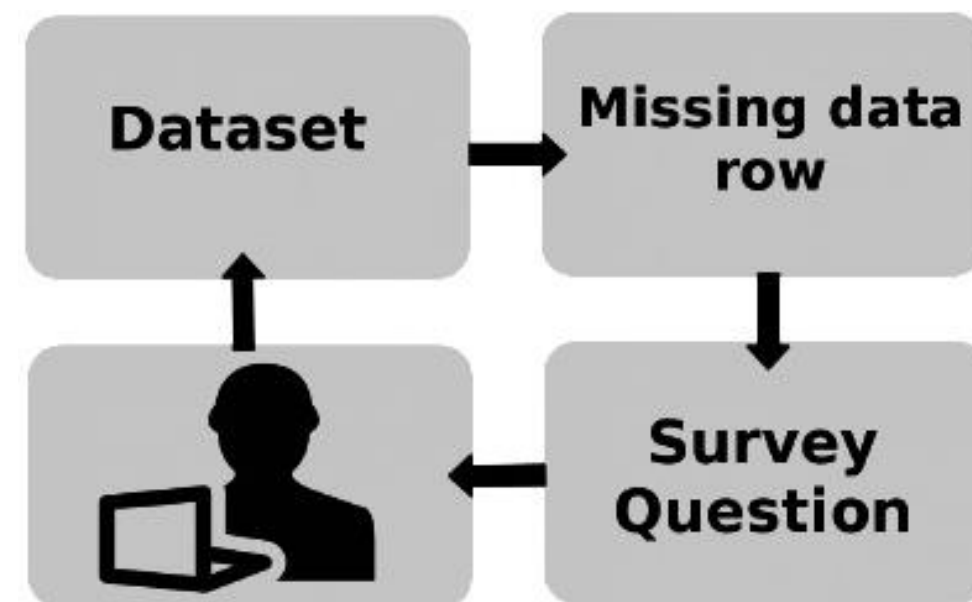
---

- Data Imputation based on Local (Small) Data
  - Model-based, Rule-based
- Data Imputation based on External (Big) Data
  - Crowdsourcing, Web List/Tables, Surface Web
- Hybrid Data Imputation Approaches
  - Web + Rules, Web + Crowd

# Crowdsourcing-based Approaches

- Crowd直接补全

- 将每个缺失数据及所在tuple等关键信息，形成单个众包任务
- 将所有填补任务交由众包平台分发给众包工人
- 根据相关算法（主动/被动）选择工人完成填补
- 缺点：
  - 容易引入众包工人的错误输入
  - 不可并发，众包填补的效率较低
  - 全部填补由众包工人完成，成本过高



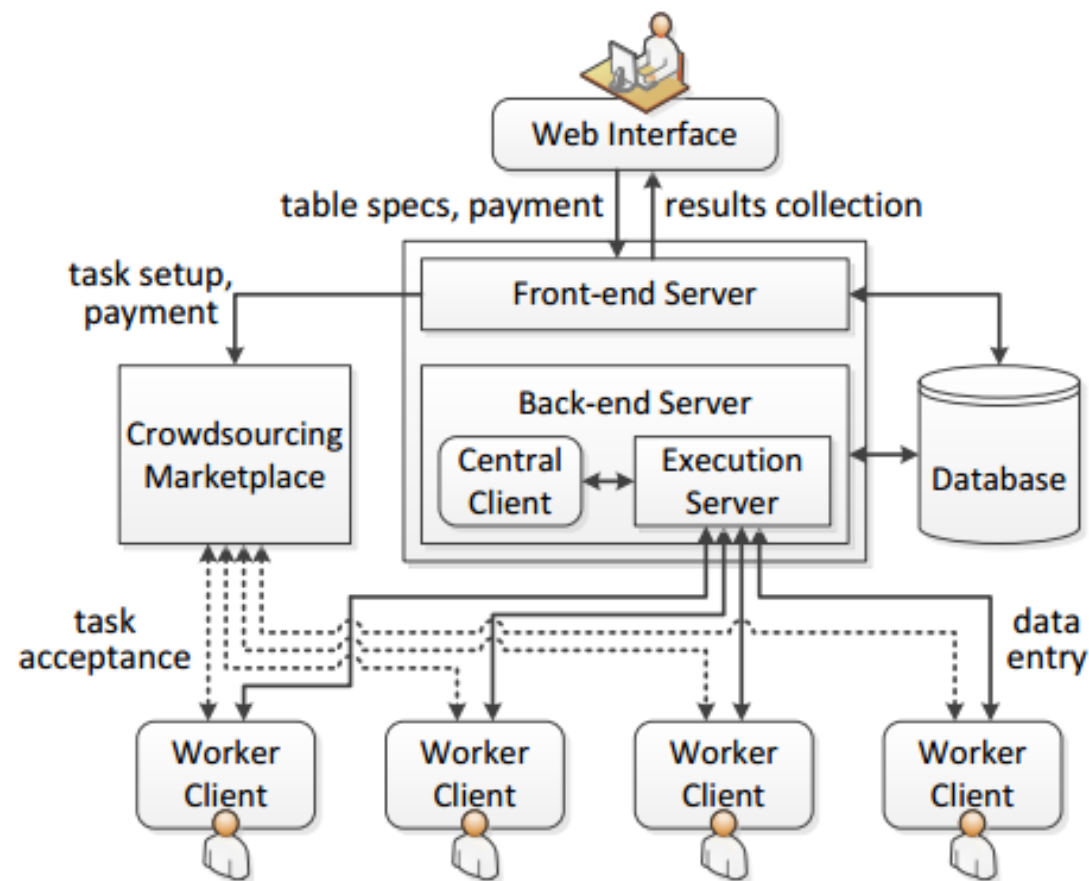
# Crowdsourcing-based Approaches

## • CrowdFill系统

- 不同于以往单个众包工人完成单个任务的模式
- CrowdFill给所有参与众包工人一个 partially-filled 表格
- 每个众包工人不仅可以完成填空任务，还可以对其他工人的填补输入选择进行：up voting 或者 down voting.
- 系统的并发机制可以让众多工作并发完成任务。

## • 优点：

- 提高了众包填补质量
- 提高了众包填补效率



# Crowdsourcing-based Approaches

## • Crowd辅助补全

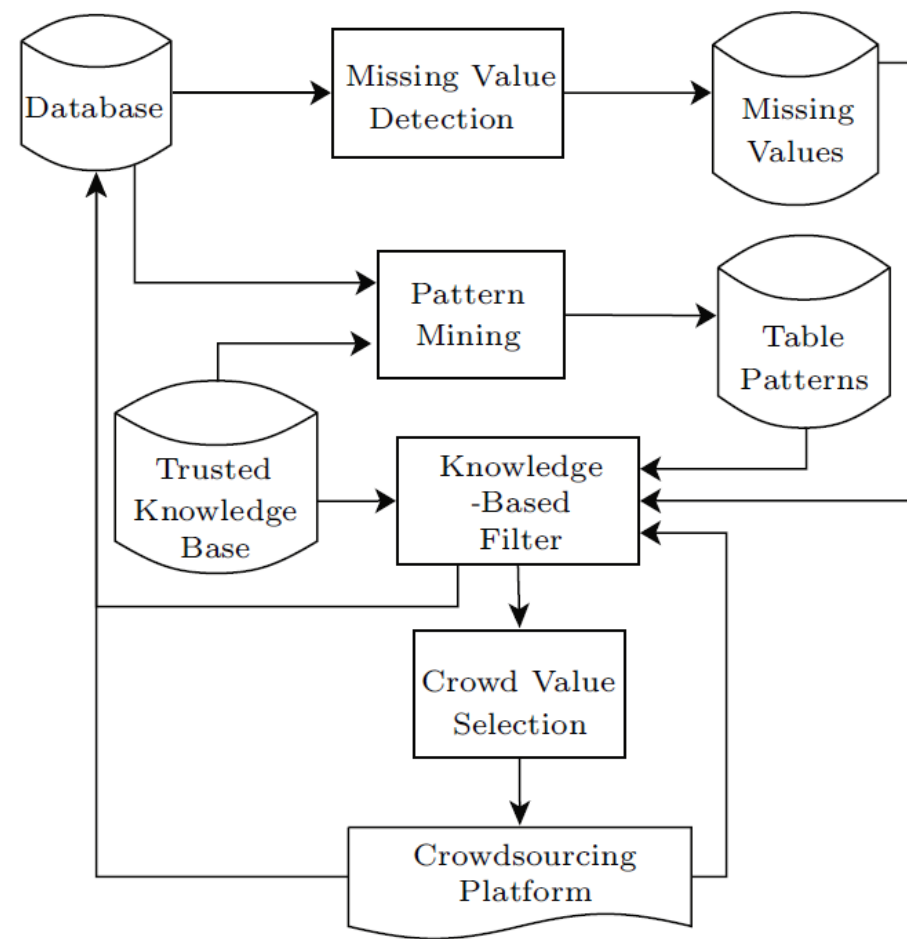
- Crowd不用于解决所有缺失问题。仅作为Knowledge Base (KB)填补的辅助完成填补。

## • 步骤：

- 1. 对比数据集和KB，得到在KB中获取缺失数据的pattern；
- 2. 利用patterns从KB中获取部分缺失数据；
- 3. KB无法完成获取的缺失数据交由众包平台完成。

## • 优点：

- 大大减少了众包成本



# Imputation From Web Lists/Tables

- 从Web结构化信息中抽取缺失数据

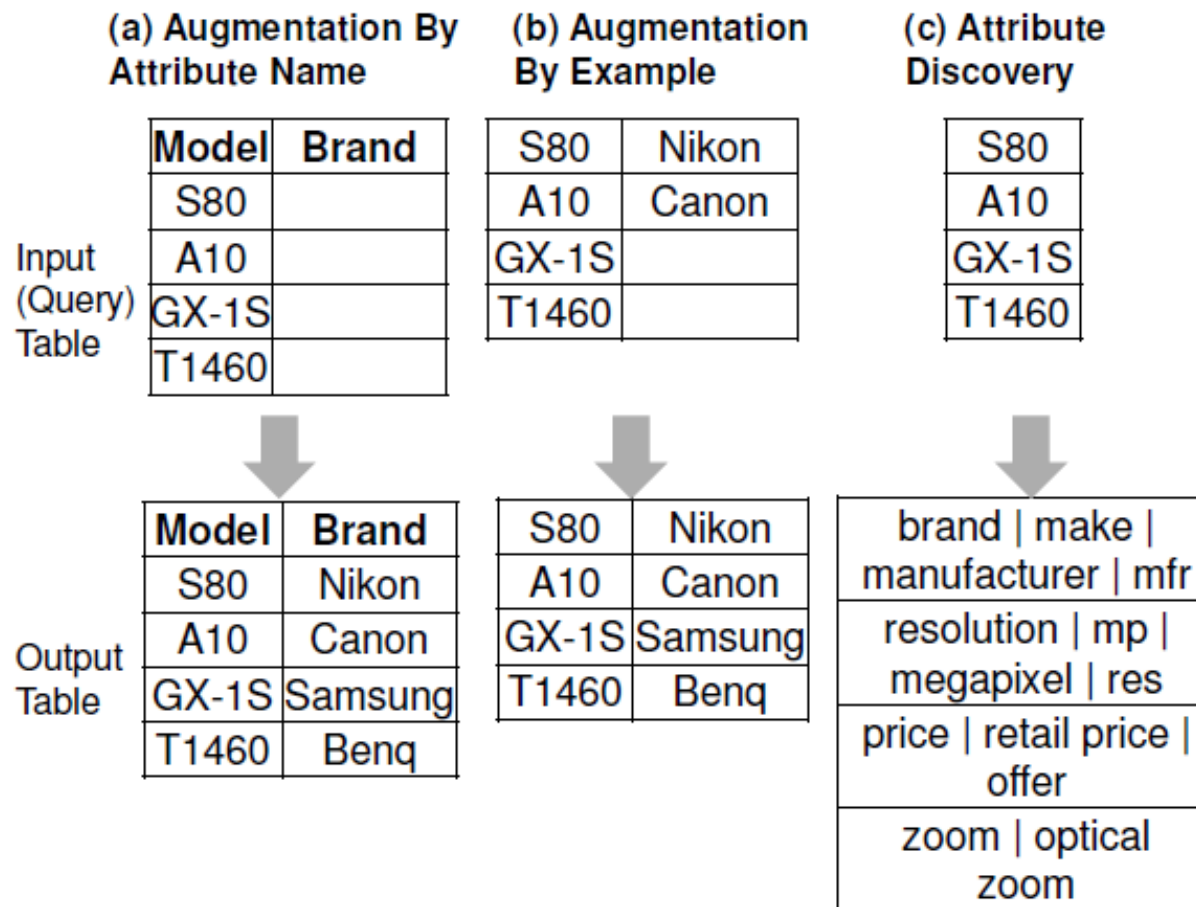
- 根据属性名扩展
- 根据实例扩展
- 发现新属性

- 优点

- 简单快捷
- 质量高

- 不足

- 数据量有限
- 更多信息存在于非结构化信息中



# Question ...

---

- Could we get missing data from the whole Surface Web?

**Quantity:** Data in the Surface Web >>  
Web List Data +  
Online KBs (e.g., freebase)

- What tools could be leveraged?
  - Web search engines: Google, Bing, Baidu
    - Finding the web pages containing the missing data
  - Information Extraction + NLP tools
    - Identifying the exact missing data from the web pages

# WebPut: Web-based Data Imputation

- **WebPut: Web-based Data ImPutation.**
  - Retrieve missing attribute values from the web
- The premise underlying WebPut:
  1. Missing data values are accessible on the Web
  2. All data values in a table are consistent

*Column  
Consistency*

Name	Email	Title	Uni.	State
Jack Davis	jdavis@mit.edu	Professor	MIT	MA
Tom Smith	tomsmith2@cs.cmu.edu		CMU	PA
Bill Wilson		Doctor	UIUC	IL
Bob Brown	bbrown7@yale.edu	A/Professor	Yale	NY
Ama Jones		Ms		CA
	lank@ucla.edu			
...	...	...	...	...

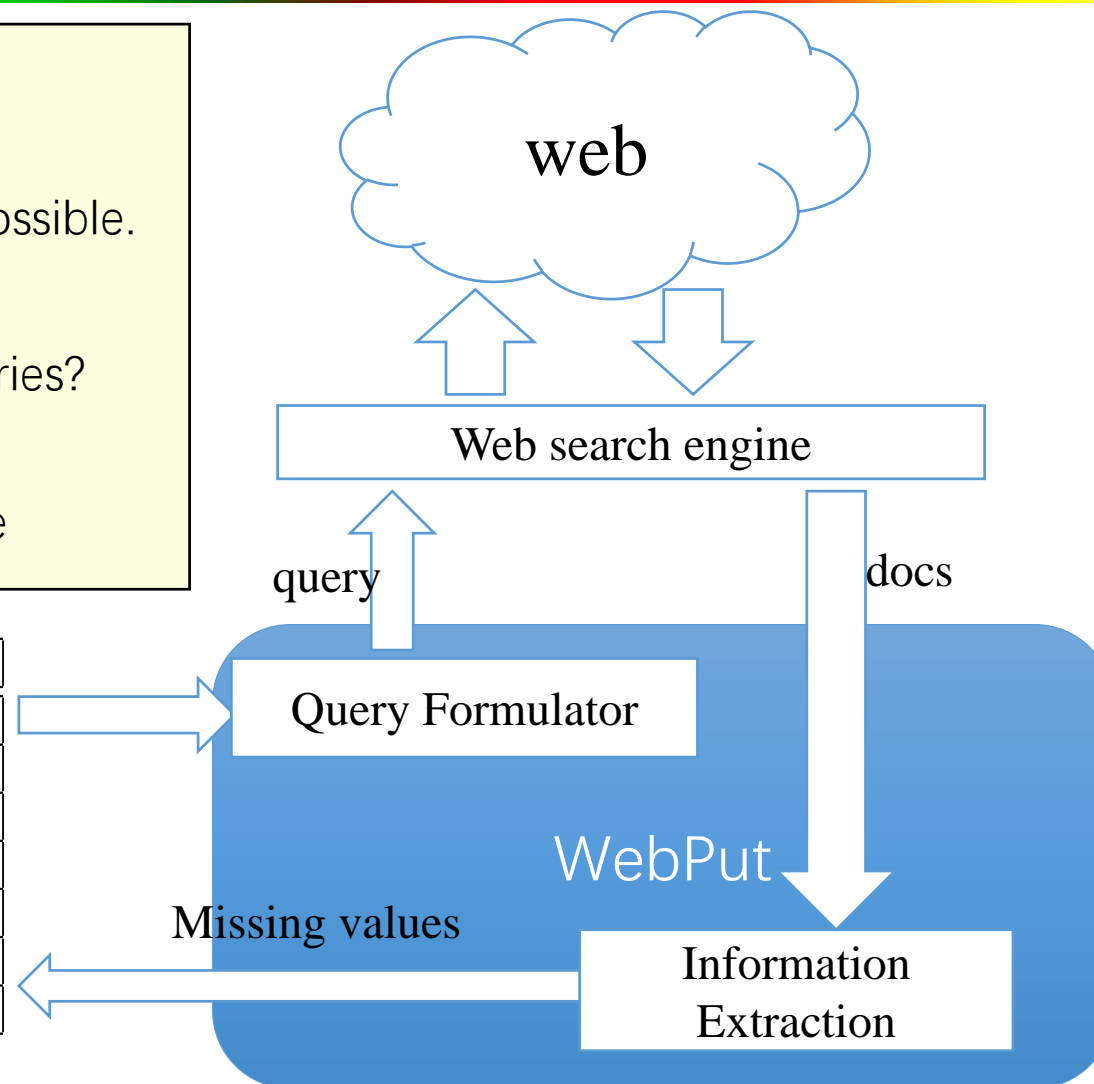
*Tuple  
Consistency*



# WebPut – The Big Picture

- **Goals:**
  - **Effectiveness** – high imputation accuracy;
  - **Efficiency** – as few imputation queries as possible.
- **Challenges:**
  - How to **formulate** effective imputation queries?
  - How to **choose** queries for one blank?
  - How to **schedule** the imputation of multiple blanks?

Name	Email	Title	Uni.	State
Jack Davis	jdavis@mit.edu	Professor	MIT	MA
Tom Smith	tomsmith2@cs.cmu.edu		CMU	PA
Bill Wilson		Doctor	UIUC	IL
Bob Brown	bbrown7@yale.edu	A/Professor	Yale	NY
Ama Jones		Ms		CA
	lank@ucla.edu			
...	...	...	...	...



# Query Formulation: Pattern Based

Name	Email
Jack Davis	jdavis@mit.edu
Tom Smith	tomsmith2@cs.cmu.edu
Bill Wilson	

**Learning Query:** "Jack M. Davis + jdavis@mit.edu"

**Learning Query:** "Tom Smith + tomsmith2@cs.cmu.edu"

... question, please  
send email to  
Jack M. Davis  
(Email: jdavis@mit.edu)

..., please feel free to  
send email to Tom  
Smith (Email:  
tomsmith2@cs.cmu.edu)

Sometimes, no pattern is learned!

Sometimes, the pattern is too **strict** to find missing values!

**Pattern:**

send email to [NAME] (Email: [EMAIL])

<Bill Wilson, ?>

Fill

billwilson@uiuc.edu

Imputation Query  
Formulation

"send email to Bill Wilson (Email: +)"

Extract

... send email to  
Bill Wilson (Email:  
billwilson@uiuc.edu)...

# Query Formulation: Co-occurrence Based

Name	Email	Title	Uni.
Jack Davis	jdavis@mit.edu	Professor	MIT
Tom Smith	tomsmith2@cs.cmu.edu		CMU
Bill Wilson		Doctor	UIUC
Bob Brown	bbrown7@yale.edu	A/Professor	Yale
Ama Jones		Ms	

**Learning Query: "Jack M. Davis + Professor + MIT"**

MIT → Department of Electrical Engineering and Computer Science → Faculty List:...  
Jack M. Davis  
Professor of Computer Science and Engineering in the.

**Learning Query: "Bob Brown + A/Prof. + Yale"**

Joined Yale Faculty 1982...  
Office location:...  
Telephone: ...  
Bob Brown is an A/Prof. of computer science department at Yale, ...

Co-occurrence Based can find **more** missing values

In WebPut, we use both methods

**Context Terms:**

Faculty, department

na Jones, Ms., ?> Imputation Query Formulation

"Ama Jones + Ms. + (Faculty OR department)"

Fill

UIUC

Extract

UIUC → Department of computer science...  
Faculty ...  
Prof. Ada Janes  
Ms. Ama Jones  
...

# 各种方法的优劣势对比分析

- 基于Rules的方法
  - 优势：recall比较低
  - 劣势：precision较高，且很高效
- 基于Web的imputation
  - 优势：充分利用web资源
  - 劣势：
    - 需要发出大量的查询，非常依赖搜索引擎
    - 可能引入Web中的噪音
- 基于Crowd的imputation
  - 优势：准确度较高
  - 劣势：crowd成本很高
- 可否取长补短，优化组合？

# Overview

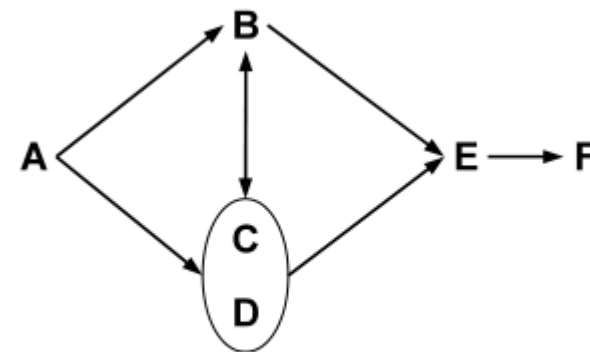
---

- Data Imputation based on Local (Small) Data
  - Model-based, Rule-based
- Data Imputation based on External (Big) Data
  - Crowdsourcing, Web List/Tables, Surface Web
- Hybrid Data Imputation Approaches
  - Web + Rules, Web + Crowd

# Hybrid Approach 1: Web + Rules

仅基于给定FD/CFDs的填补:

	A	B	C	D	E	F
T <sub>1</sub>	a <sub>1</sub>	b <sub>1</sub>	c <sub>1</sub>	d <sub>1</sub>	e <sub>1</sub>	f <sub>1</sub>
T <sub>2</sub>	a <sub>2</sub>	b <sub>1</sub>	c <sub>1</sub>	d <sub>1</sub>	e <sub>1</sub>	f <sub>1</sub>
T <sub>3</sub>	a <sub>3</sub>				e <sub>2</sub>	f <sub>2</sub>
T <sub>4</sub>	a <sub>4</sub>	b <sub>3</sub>				
T <sub>5</sub>	a <sub>5</sub>		c <sub>3</sub>	d <sub>3</sub>		



(b) Dependencies between Attributes

Perform inferring only: Fill 3 blanks at T<sub>1</sub>[E], T<sub>1</sub>[F], T<sub>2</sub>[B] only.

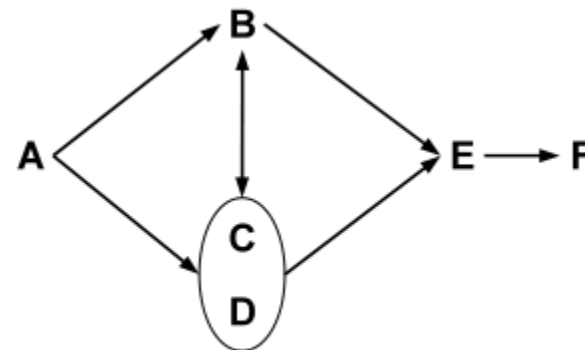
Perform retrieving only: Need 13 retrieving operations.

A simple hybrid way(inferring first, retrieving second): Still issue 10 retrieving operations.

# Hybrid Approach 1: Web + Rules

基于FD/CFDs+Web的填补:

	A	B	C	D	E	F
$T_1$	$a_1$	$b_1$	$c_1$	$d_1$	$e_1$	$f_1$
$T_2$	$a_2$	$b_1$	$c_1$	$d_1$	$e_1$	$f_1$
$T_3$	$a_3$	$b_2$	$c_2$	$d_2$	$e_2$	$f_2$
$T_4$	$a_4$	$b_3$	$c_3$	$d_3$	$e_2$	$f_2$
$T_5$	$a_5$	$b_3$	$c_3$	$d_3$	$e_2$	$f_2$



(b) Dependencies between Attributes

The 1<sup>st</sup> Inference:  $T_1[E]$ ,  $T_1[F]$ ,  $T_2[B]$

The 1<sup>st</sup> Retrieving:  $T_3[B]$ ,  $T_5[B]$

The 2<sup>nd</sup> Inference:  $T_4[C]$ ,  $T_4[D]$

The 2<sup>nd</sup> Retrieving:  $T_3[C]$ ,  $T_3[D]$ ,  $T_4[E]$

The 3<sup>rd</sup> Inference:  $T_4[F]$ ,  $T_5[E]$ ,  $T_5[F]$

Only require 5 retrieving operations

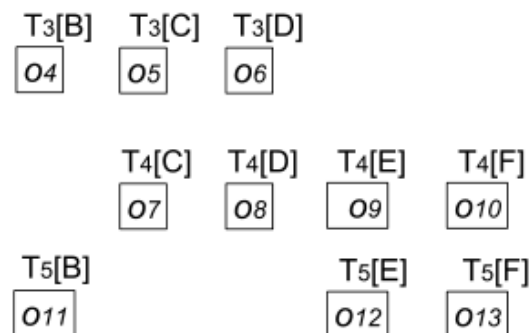
**Problem:**

How to Identify an **optimal** scheduling scheme?

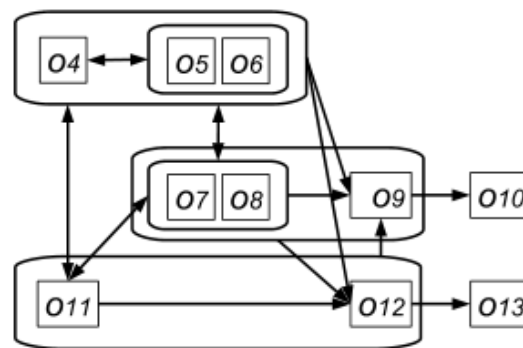
- Maximum Recall,
- Minimum Retrieving Times

# Hybrid Approach 1: Web + Rules

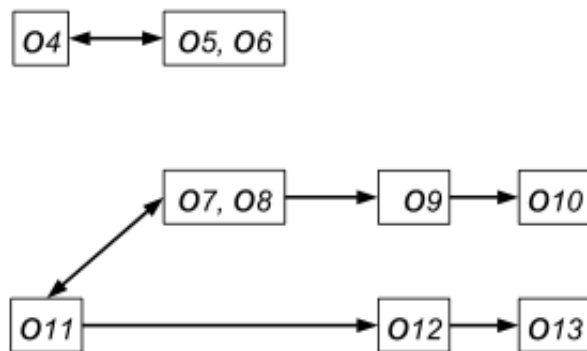
- We hope to retrieve those un-inferable values only!
  - Building **Inference Dependency Graph**
  - Identifying Nodes for Retrieving from the Graph



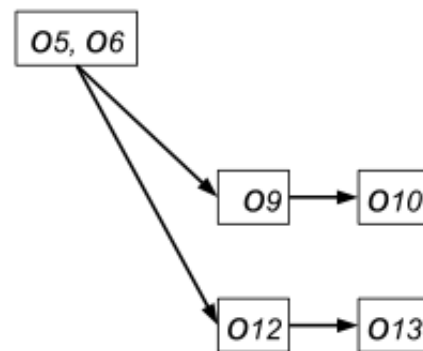
(a) Step 1. Denote Missing Values as Nodes



(b) Step 2: Put in all Possible Edges



(c) Step 3. Simplify the Graph



(d) Graph at the 2<sup>nd</sup> Retrieving Step



# Hybrid Approach 1: Web + Rules

- **Data Sets:** Two real datasets (PersonInfo and DBLP).

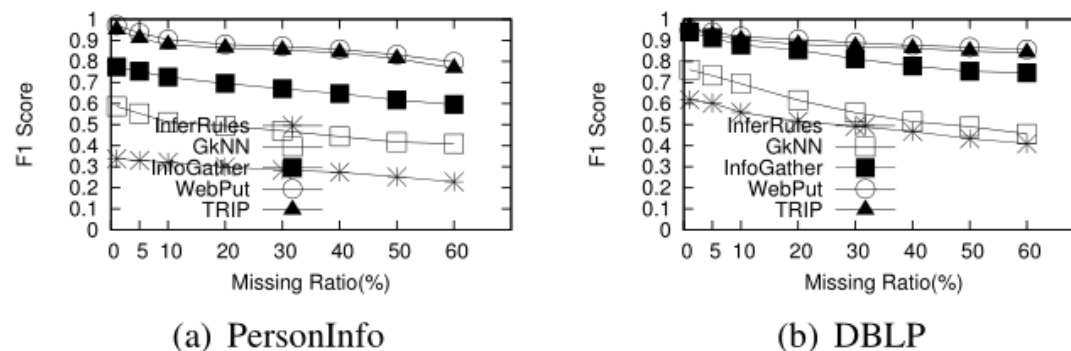


Fig. 3. Comparing the F1 Scores on Two Real Data Sets

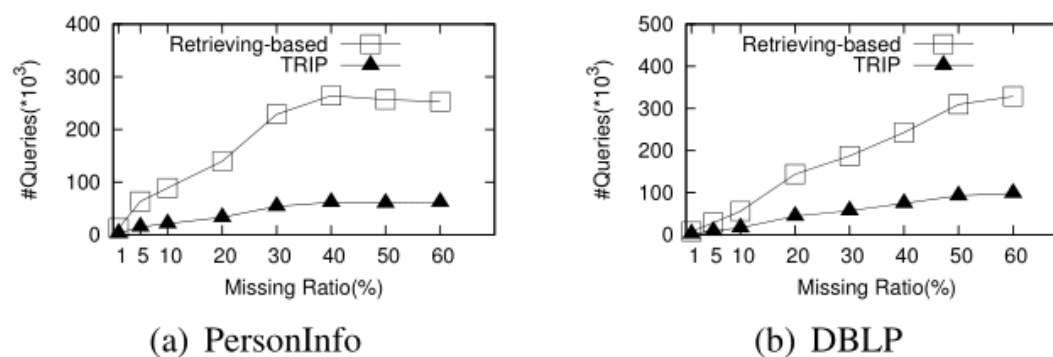
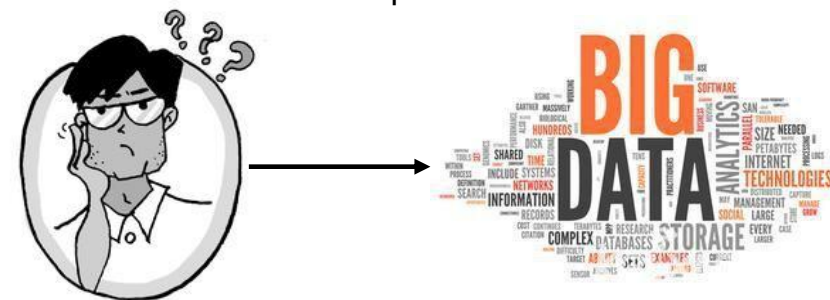


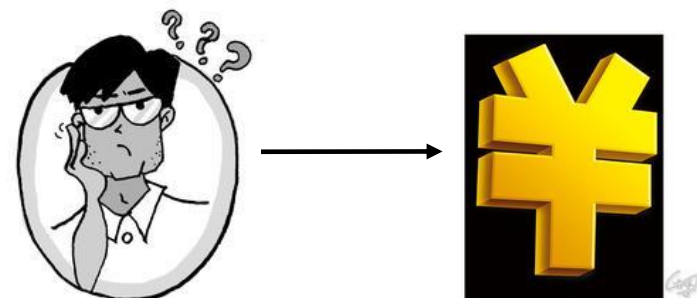
Fig. 4. Comparing the #Queries of the Retrieving Method and TRIP

# Hybrid Approach 2: Web + Crowd

- Web-based method
  - Advantage: almost free sources on the Web
  - Disadvantage: also some missing values we can not capture from the Web and need a certain amount of human knowledge precision and recall



- Crowdsourcing-based method
  - Advantage: higher precision and recall than the Web-based method
  - Disadvantage: need to pay the crowd worker high reward



# 本节大纲

- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
- 知识图谱数据补全的质量控制
  - 关系数据库中的缺失数据补全技术
  - 知识图谱中的实体类型补全技术
  - 知识图谱中的实体关系补全技术
  - 知识图谱中的实体属性值补全技术
- 知识图谱数据更新的质量控制



# 知识图谱中的实体类型补全

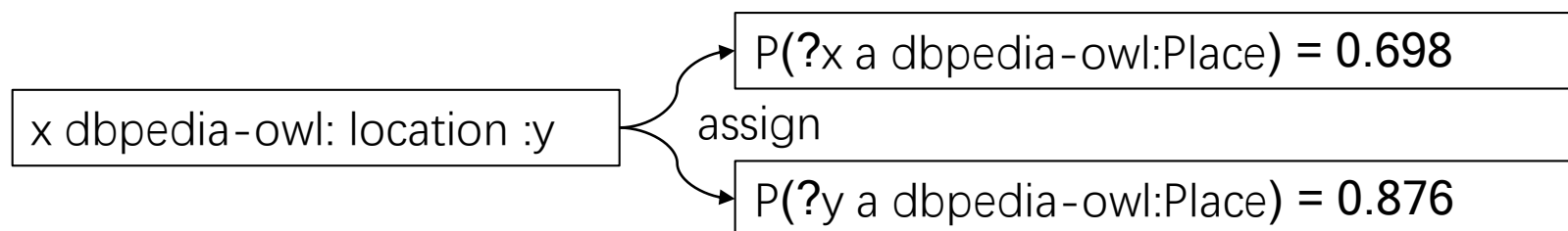
- 类型断言 (Type Assertions)
  - 基于内部知识 (Internal Knowledge-based)
    - SDType (ISWC'13);
    - Neural Joint Learning (PACLIC'16);
    - Path-CNN (WISE'18);
    - and some other methods.
  - 基于外部知识 (External Knowledge-based)
    - Tipola (ISWC'12);
    - Classifier based on Wiki Links (LDOW'12)
    - Crowdsourcing (APWeb'18)

# 基于内部知识 (Internal Knowledge-based)

- **SDDType**: using **Statistical Distribution of types** in the subject and object positions for predicting the instance's types.

Table 1. Type distribution of the property `dbpedia-owl:location` in DBpedia

Type	Subject (%)	Object (%)
<code>owl:Thing</code>	100.0	88.6
<code>dbpedia-owl:Place</code>	69.8	87.6
<code>dbpedia-owl:PopulatedPlace</code>	0.0	84.7
<code>dbpedia-owl:ArchitecturalStructure</code>	50.7	0.0
<code>dbpedia-owl:Settlement</code>	0.0	50.6
<code>dbpedia-owl:Building</code>	34.0	0.0
<code>dbpedia-owl:Organization</code>	29.1	0.0
<code>dbpedia-owl:City</code>	0.0	24.2
...	...	...



# 基于内部知识 (Internal Knowledge-based)

134

## Implementation

subject	predicate	object
dbpedia:Mannheim	dbpedia-owl:federalState	dbpedia:Baden-Württemberg
dbpedia:Steffi:Graf	dbpedia-owl:birthPlace	dbpedia:Mannheim
...	...	...

① Input data

resource	type
dbpedia:Mannheim	dbpedia-owl:Place
dbpedia:Mannheim	dbpedia-owl:Town
...	...

resource	predicate	frequency
dbpedia:Mannheim	dbpedia-owl:federalState	1
dbpedia:Mannheim	dbpedia-owl:birthPlace <sup>-1</sup>	140
...	...	...

② Compute basic distributions

type	apriori probability
dbpedia-owl:Place	0.3337534
dbpedia-owl:Town	0.0523772
...	...

predicate	weight
dbpedia-owl:federalState	0.3337534
dbpedia-owl:birthPlace <sup>-1</sup>	0.0523772
...	...

③ Compute weights and conditional probabilities

predicate	type	probability
dbpedia-owl:federalState	dbpedia-owl:Place	1.0000000
dbpedia-owl:birthPlace <sup>-1</sup>	dbpedia-owl:Town	0.1760390
...	...	...

④ Materialize missing types

resource	type	score
dbpedia:Heinsberg	dbpedia-owl:Place	0.8856929
dbpedia:Heinsberg	dbpedia-owl:PopulatedPlace	0.8110996
...	...	...

# 基于内部知识 (Internal Knowledge-based)

## • Neural Joint Learning

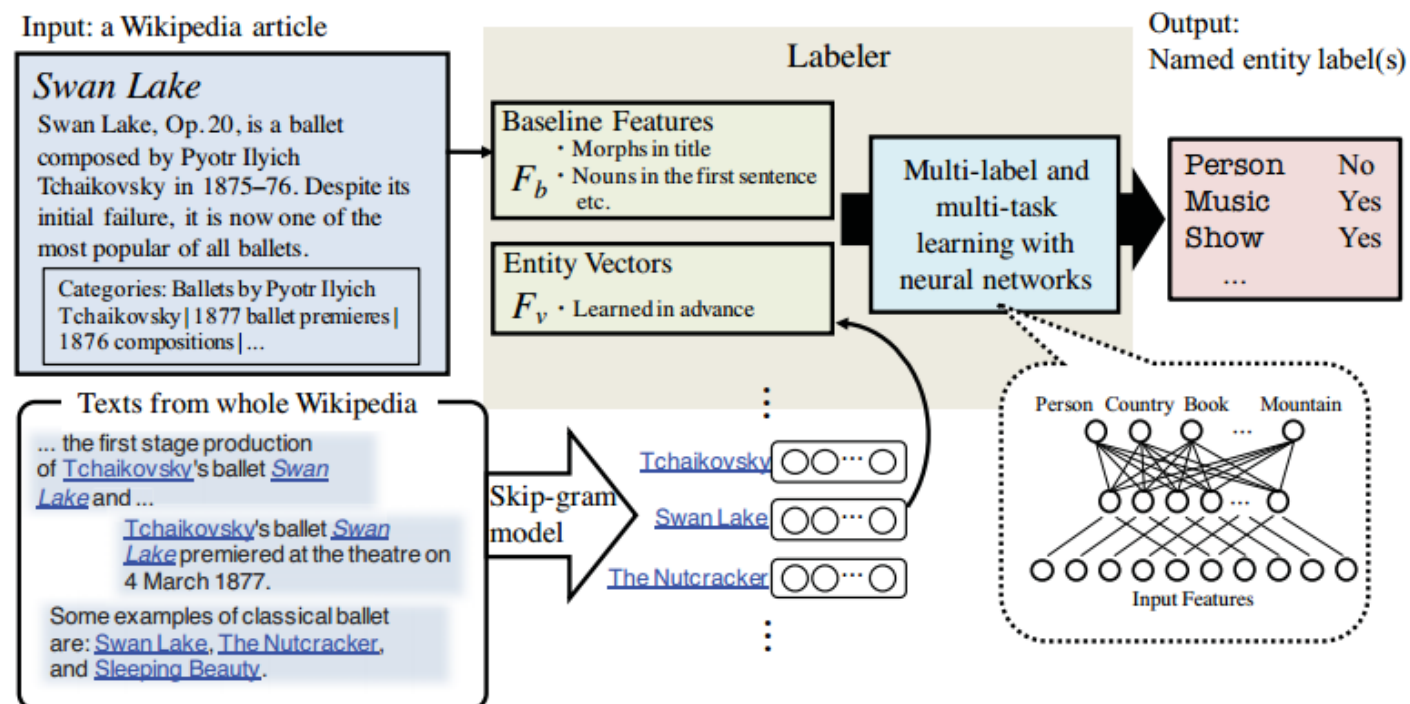


Figure 1: Automatic assignment of NE labels to Wikipedia articles based on multi-task learning and vector representation of articles

M. Suzuki et al. Neural Joint Learning for Classifying Wikipedia Articles into Fine Grained Named Entity Types , PACLIC'16

# 基于内部知识 (Internal Knowledge-based)

## • Path-CNN

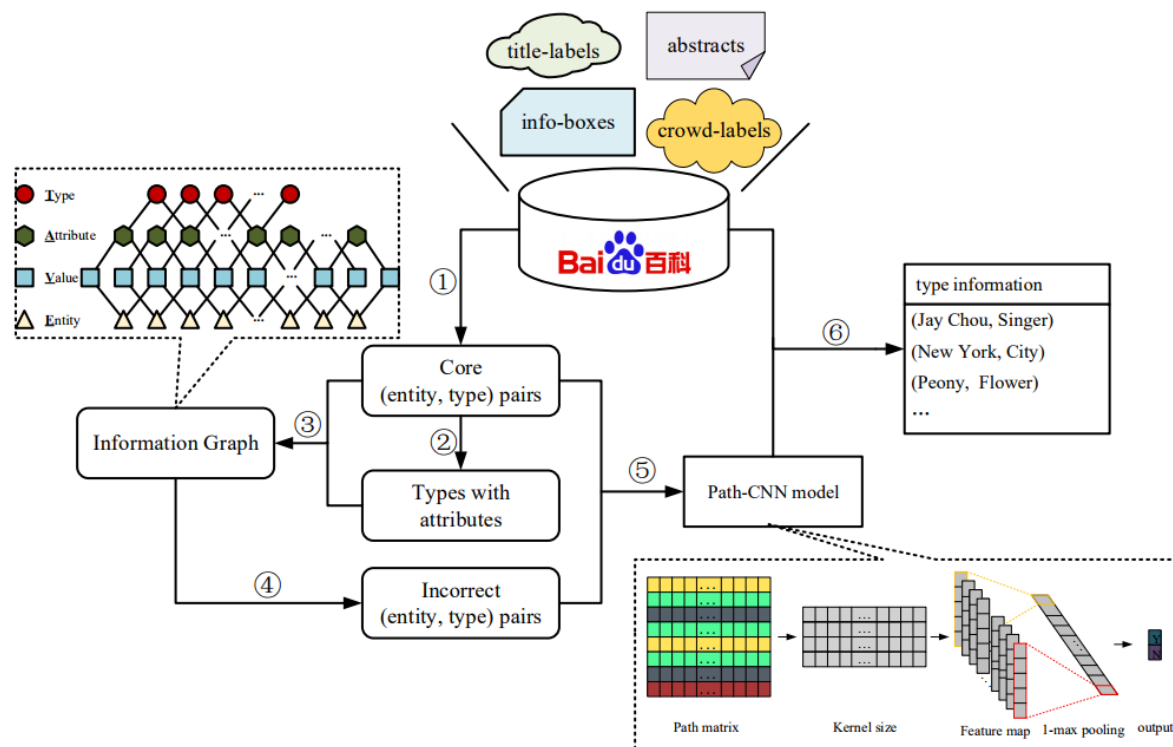


Fig. 1: Workflow.

M. Hao et al. Mining High-Quality Fine-Grained Type Information from Chinese Online Encyclopedias, WISE'18



# 基于内部知识 (Internal Knowledge-based)

## • Path-CNN

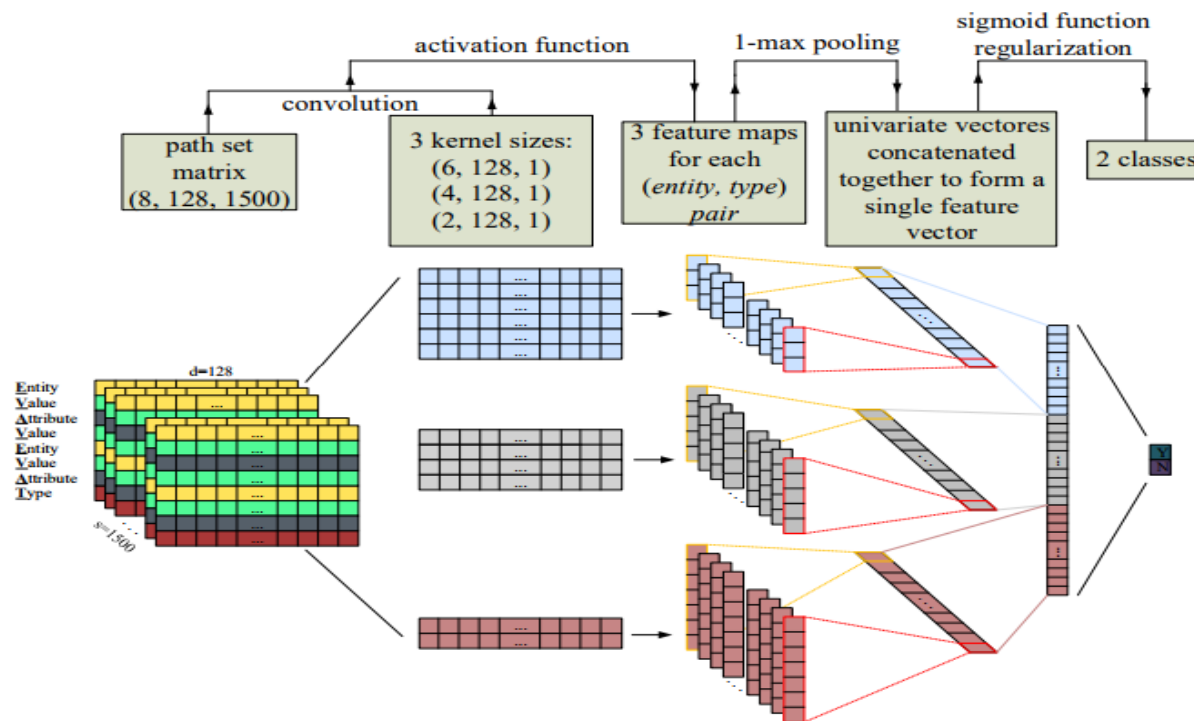


Fig. 3: The Architecture of the Path-CNN Model.

M. Hao et al. Mining High-Quality Fine-Grained Type Information from Chinese Online Encyclopedias, WISE'18

# 基于内部知识 (Internal Knowledge-based)

## • 其他方法

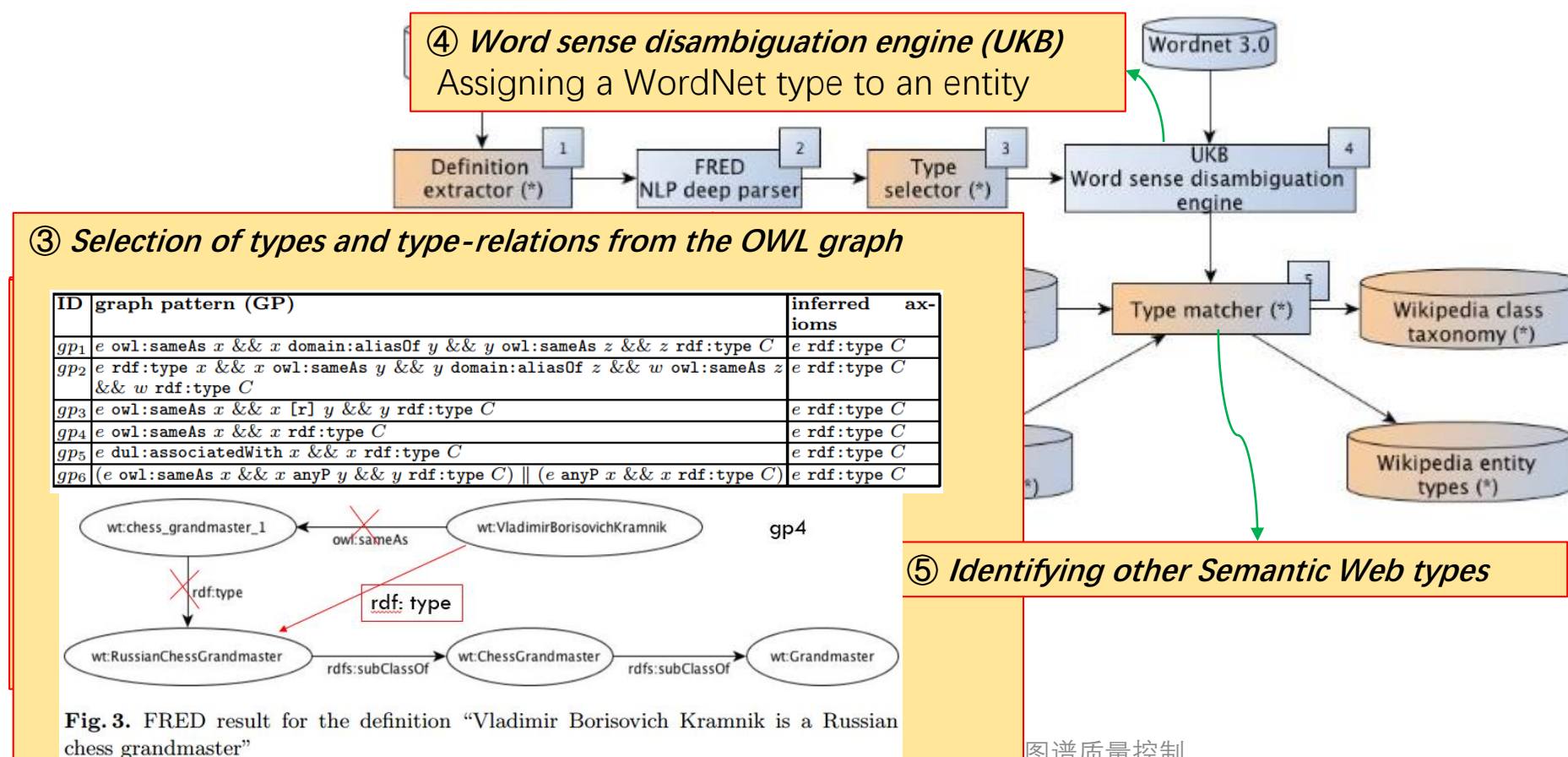
- Training a *Classification Model* (e.g., *SVMs*)
  - E.g., Exploiting interlinks between the knowledge graphs to classify instances in one knowledge graph based on properties present in the other.
- *Association Rule Mining* for predict missing information.
  - Exploit association rules to predict missing types in DBpedia based on such redundancies.
- Using *Topic Modeling* for type prediction
  - E.g., LDA is applied to find topics for documents of entities.

# 知识图谱中的实体类型补全

- 类型断言 (Type Assertions)
  - 基于内部知识 (Internal Knowledge-based)
    - SDType (ISWC'13);
    - Neural Joint Learning (PACLIC'16);
    - Path-CNN (WISE'18);
    - and some other methods.
  - 基于外部知识 (External Knowledge-based)
    - Tipola (ISWC'12);
    - Classifier based on Wiki Links (LDOW'12)
    - Crowdsourcing (APWeb'18)

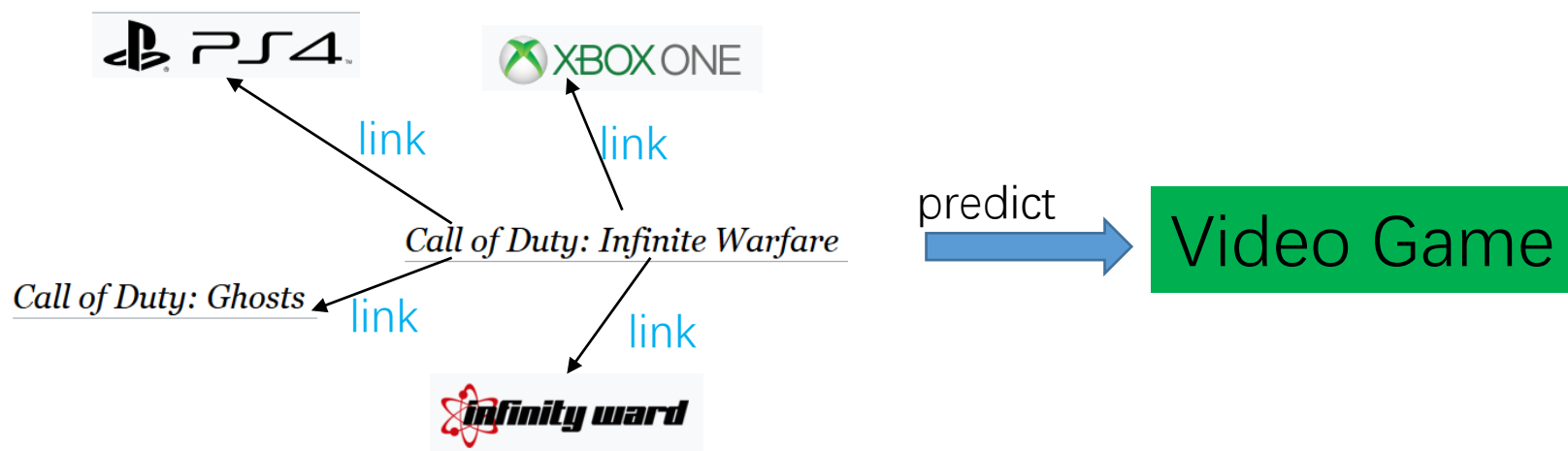
# 基于外部知识 (External Knowledge-based)

- **Tipalo Algorithm:** identifies the most appropriate types for an entity by interpreting its natural language definition.



# 基于外部知识 (External Knowledge-based)

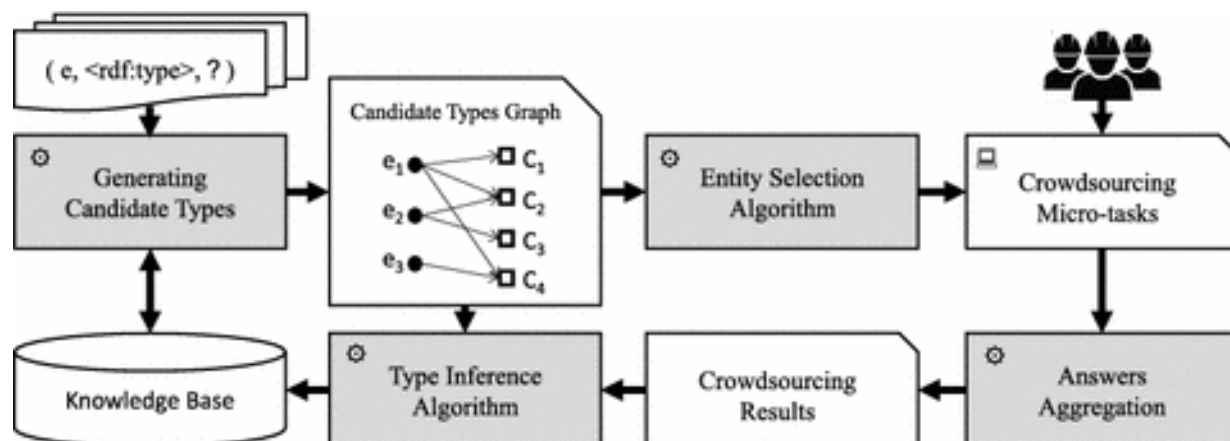
- Classifier based on wiki Links
  - using **Wikipedia link graph** to predict types in a KG
  - interlinks between Wikipedia pages are exploited to create feature vectors, e.g., based on the categories of the related pages.



Nuzzolese et al. Type inference through the analysis of Wikipedia links, LDOW'12

# 基于外部知识 (External Knowledge-based)

## • Crowdsourcing



步骤:

1. Generating Candidate Types: SDType
2. Selecting Entities for Crowdsourcing
  - a greedy-based algorithm based on the expected utilities
3. Inferring Types Using Crowdsourcing Results

Z. Dong et al. Using Crowdsourcing for Fine-Grained Entity Type Completion in Knowledge Bases, APWeb'18

# 本节大纲

- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
- 知识图谱数据补全的质量控制
  - 关系数据库中的缺失数据补全技术
  - 知识图谱中的实体类型补全技术
  - 知识图谱中的实体关系补全技术
  - 知识图谱中的实体属性值补全技术
- 知识图谱数据更新的质量控制



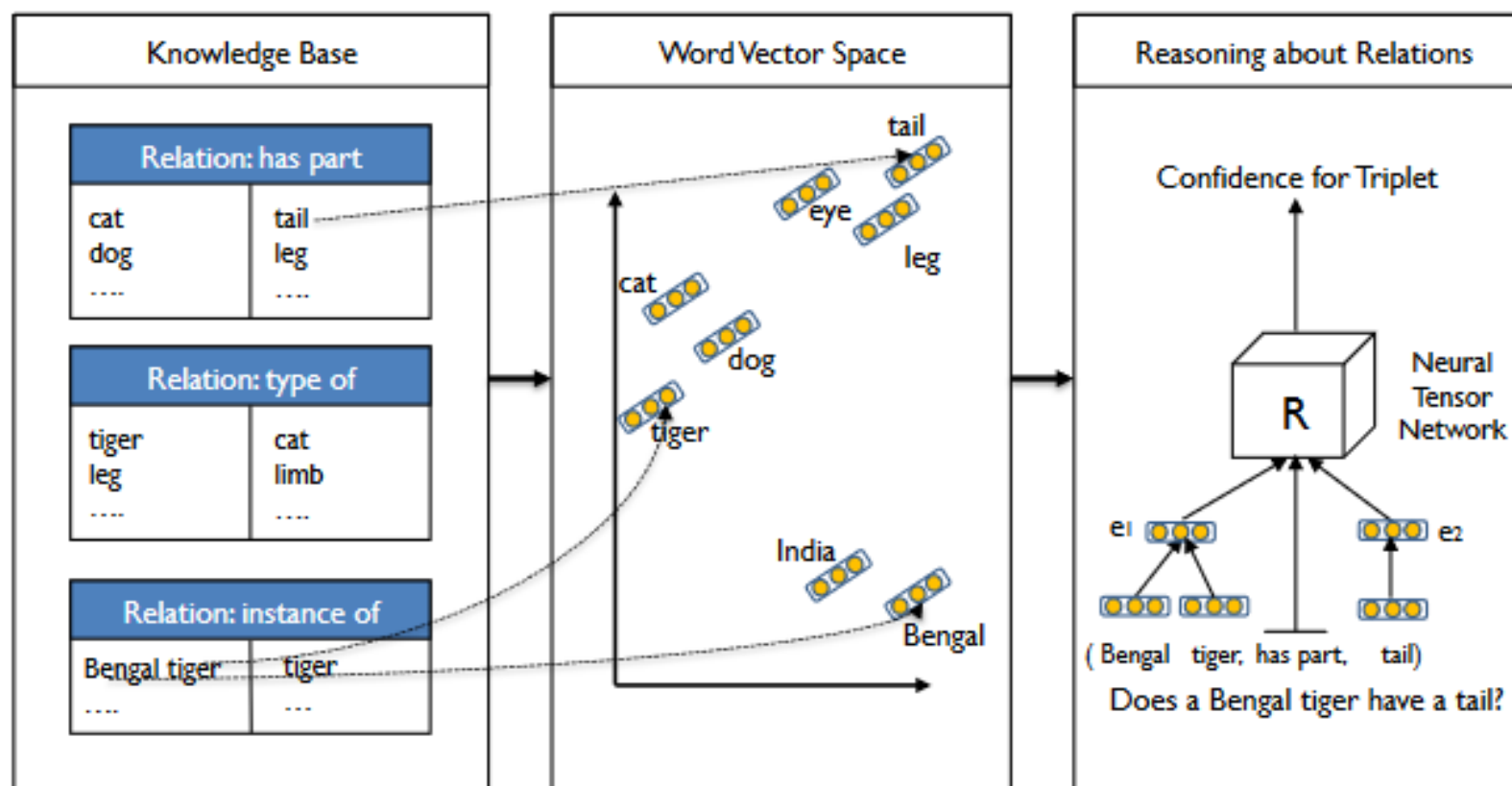
# 知识图谱中的实体关系补全技术

- 关系预测 (Relation Prediction)
  - 基于内部知识 (Internal Knowledge-based)
    - Neural Tensor Network (NIPS'13) ; Mining Association Rules(ISWC'15)
  - 基于外部知识 (External Knowledge-based)
    - Matching HTML Tables to DBpedia(WIMS'15); and some other methods



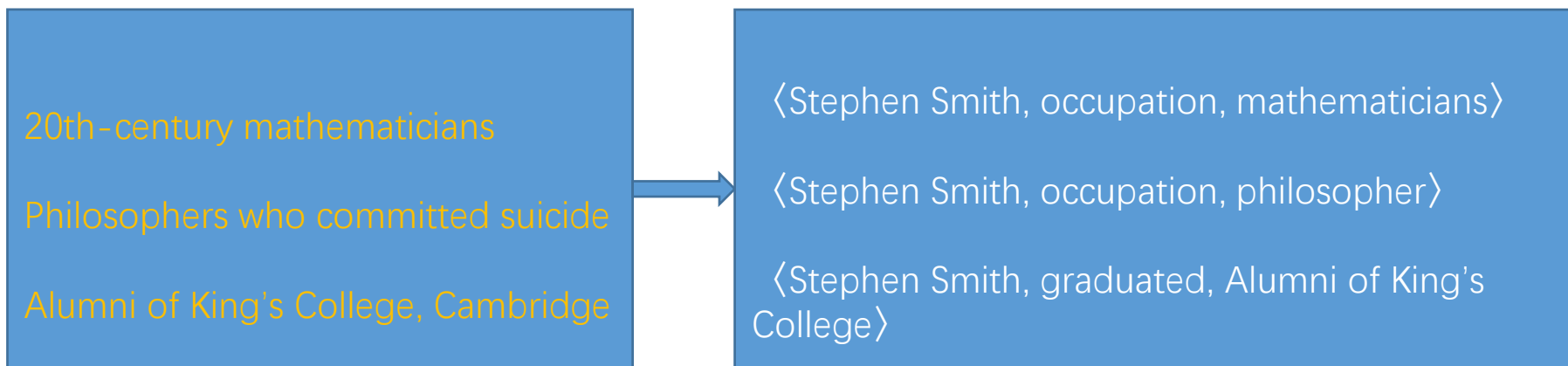
# 基于内部知识 (Internal Knowledge-based)

- Neural tensor network is suitable for reasoning over relationships between two entities.



# 基于内部知识 (Internal Knowledge-based)

- Mining *Association Rules* for predicting relations.
  - Mining of association rules which predict relations between entities in DBpedia from Wikipedia categories is proposed.



# 知识图谱中的实体关系补全技术

- 关系预测（Relation Prediction）
  - 基于内部知识（Internal Knowledge-based）
    - Neural Tensor Network (NIPS'13) ; Mining Association Rules(ISWC'15)
  - 基于外部知识（External Knowledge-based）
    - Matching HTML Tables to DBpedia(WIMS'15); and some other methods

# 知识图谱中的实体关系补全技术

- Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism, ACL18

Normal	S1: Chicago is located in the United States.	Chicago → country → United States
	{<Chicago, country, United States>}	
EPO	S2: News of the list's existence unnerved officials in Khartoum, Sudan's capital.	Sudan → capital → Khartoum Sudan → contains → Khartoum
	{<Sudan, capital, Khartoum>, <Sudan, contains, Khartoum>}	
SEO	S3: Aarhus airport serves the city of Aarhus who's leader is Jacob Bundsgaard.	Aarhus → leaderName → Jacob Bundsgaard Aarhus → cityServed → Aarhus Airport
	{<Aarhus, leaderName, Jacob Bundsgaard>, <Aarhus Airport, cityServed, Aarhus>}	

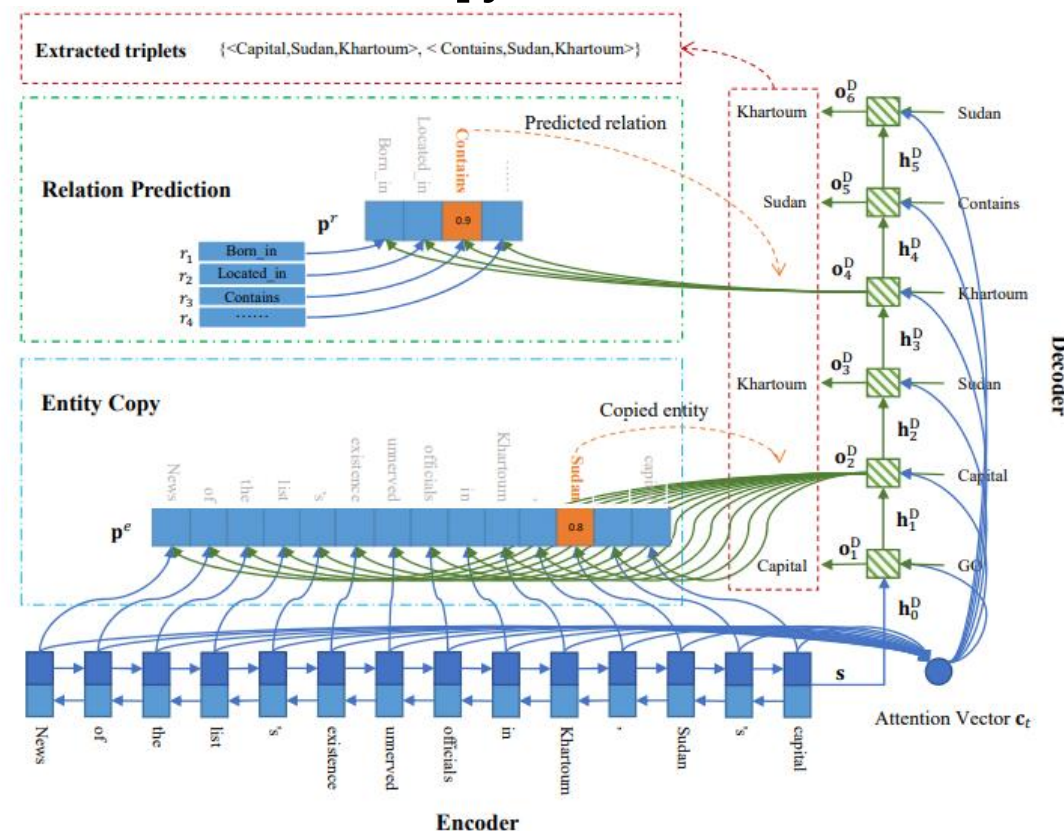


Figure 2: The overall structure of OneDecoder model. A bi-directional RNN is used to encode the source sentence and then a decoder is used to generate triples directly. The relation is predicted and the entity is copied from source sentence.

# 基于外部知识 (External Knowledge-based)

## • Matching HTML Tables to Dbpedia

### • Challenges:

- pairs of table columns have to be matched to properties in the DBpedia ontology
- rows in the table need to be matched to entities in Dbpedia

### • Solution:

- evaluated on a gold standard mapping for a sample of HTML tables from the WebDataCommons Web Table corpus

University	Present President
University of Oxford	Andrew D. Hamilton
University of Cambridge	Leszek Krzysztof Borysiewicz
University College London	Michael Arthur



```
<University of Oxford, present_president, Andrew D. Hamilton >
<University of Oxford, present_president, Andrew D. Hamilton >
<University of Oxford, present_president, Andrew D. Hamilton >
```

# 基于外部知识 (External Knowledge-based)

- **Distant supervision** with a large text corpora;
  - Step 1: **Seed Entities** in the knowledge graph are linked to the text corpus by means of Named Entity Recognition
  - Step 2: Seek for **text pattern** which correspond to relation types
  - Step 3: Apply those patterns to find **additional relations** in the text corpus
  - *A Bootstrapping way with starting seeds in KG.*
- Based on **web search engines**:
  - Discover **frequent context terms** for relations
  - Use those **frequent context terms** to formulate search engine queries for filling missing relation values.
- Based on **another KG**
  - Using Interlinks between KGs to fill gaps and do knowledge transfer

# 本节大纲

- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
- 知识图谱数据补全的质量控制
  - 关系数据库中的缺失数据补全技术
  - 知识图谱中的实体类型补全技术
  - 知识图谱中的实体关系补全技术
  - 知识图谱中的实体属性值补全技术
- 知识图谱数据更新的质量控制



# 实体属性值补全技术对比

	关系数据库的属性值补全	知识图谱的属性值补全
相同点	<p>都是属性值补全问题</p> <p>很多补全技术可以通用</p>	
不同点	补全对象一般为单一关系表内的缺失属性值，结构较为统一简单。	补全对象为不同概念下不同实体的确实属性值，每个概念下的实体的相关属性不尽相同，可看做是很多个小关系表的缺失属性值补全。



# 知识图谱中的实体属性值补全方法

## Are All People Married? Determining Obligatory Attributes in Knowledge Bases, WWW18

**研究问题：**找到知识图谱中概念的必有属性

**解决方案：**

- 提出基于概念的层次结构来推断概念的必有属性
- 基本假设
  - 假设知识库的不完全性在知识库的所有类中都是均匀分布的
  - 如果一个属性在某的概念中分布很稀疏，而在其它概念中分布较密，则可判断它一定不是分布较为稀疏的概念的必有属性

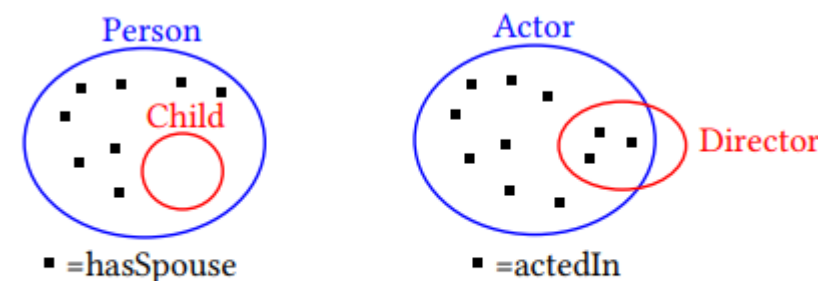


Figure 1: Examples of attributes and classes.

# 知识图谱中的实体属性值补全方法

## Are All People Married? Determining Obligatory Attributes in Knowledge Bases, WWW18

Generalization Rules:

A generalization rule for a KB  $K$  is a formula of the form  $A \subseteq B$ , where  $A$  and  $B$  are classes of  $K$ , subject sets of  $K$ , or intersections thereof.

e.g.

$$President_K \subseteq presidentOf_K$$

$$conf(A \subseteq B) = \frac{|A \cap B|}{|A|}$$

Confidence Ratio :

$$s_p^K(c, c') = \frac{conf(c \setminus c' \subseteq p_K)}{conf(c \cap c' \subseteq p_K)}$$

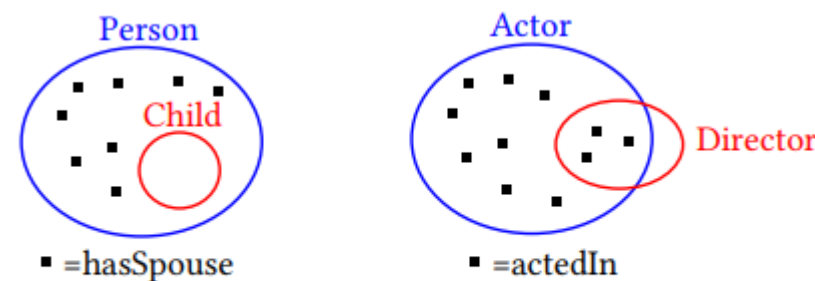


Figure 1: Examples of attributes and classes.

# 知识图谱中的实体属性值补全方法

Are All People Married? Determining Obligatory Attributes in Knowledge Bases, WWW18

---

## Algorithm 1: ObligatoryAttribute

---

**Input:** KB  $K$ , class  $c$ , property  $p$ , threshold  $\theta$ ,  
threshold  $\theta' = 100$

**Output:** true if  $c \subseteq p\mathcal{W}$  is predicted

```

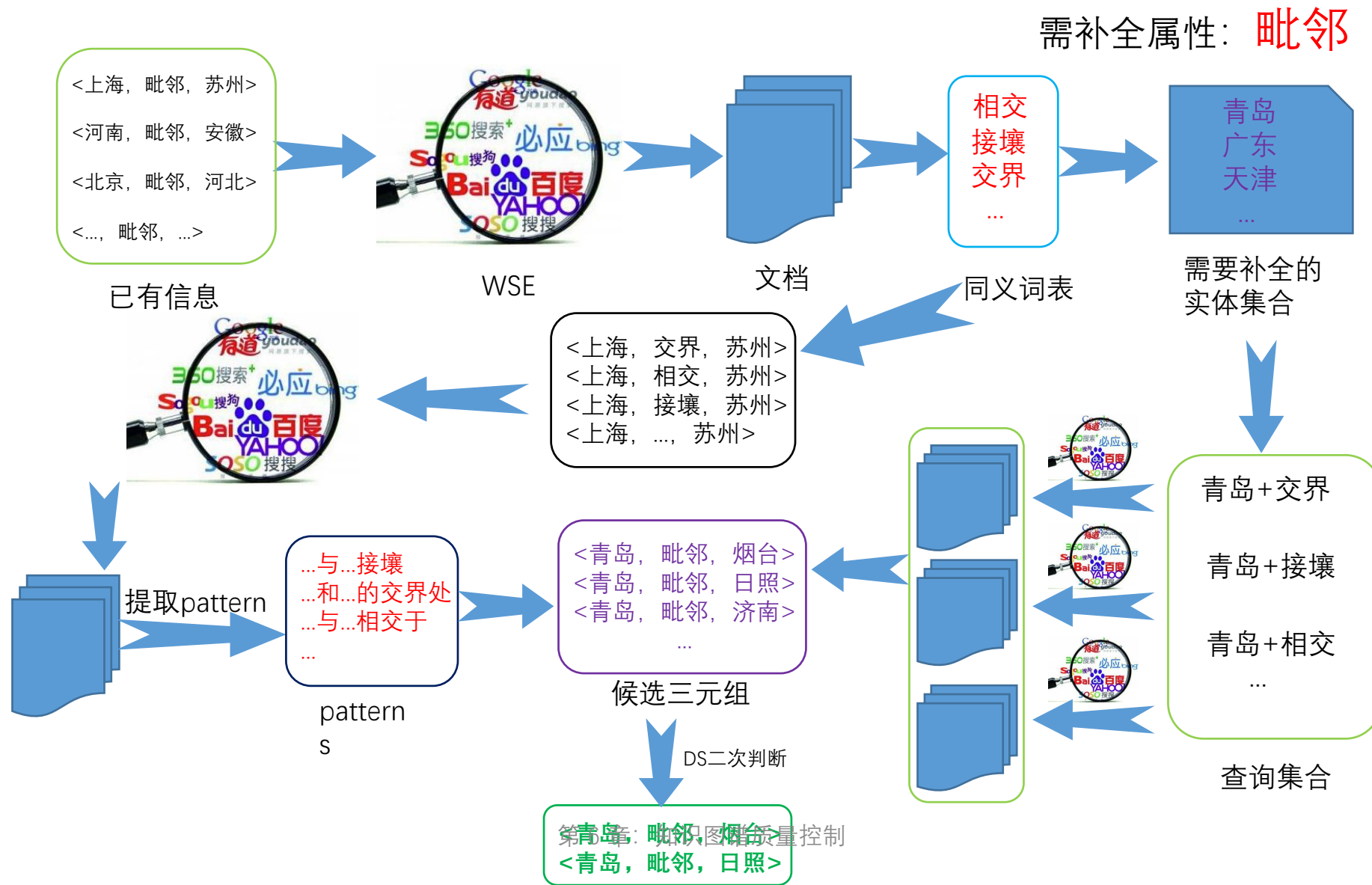
1 if  $|c \cap p_K| < \theta'$  then
2   return false
3 for stable class  $c'$  do
4   if  $|\log(s_p^K(c, c'))| > \log(\theta)$  then
5     return false
6   if  $\log(s_p^K(c', c)) > \log(\theta)$  then
7     return false
8 return true
  
```

---

警告:

只能尽可能将那些不是必有的属性排除。当一个class中没有必有属性的时候, 该算法的性能就很差。

# 知识图谱中的实体属性值补全方法举例




# 知识图谱数据更新的质量控制

---

# 本节大纲

---

- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
- 知识图谱数据补全的质量控制
- 知识图谱数据更新的质量控制
  - 错误数据清洗 
  - 过期数据更新

# 错误数据清洗

## • 什么是数据清洗？

- 关系数据库：找出并修正关系数据库中的错误属性值
- 知识图谱中：找出并修正图谱中的错误属性值或实体间关系

Name	University	City	Country
San Zhang	Soochow Uni.	Soochow	China
Si Li	Soochow Uni.	Hefei	China
Er Wang	Soochow Uni.	New York	USA

# 错误数据清洗

## 数据清洗常见方法

- 基于规则的方法 Quality Rules/Constraints:
  - FDs/CFDs: *e.g.*, “(University -> City, 0.98)”
  - Identify inconsistent data in Conflicts, choose a minimum change
  - User-Defined Quality Rules: *e.g.*, *change all capital into lowercase*
  - Inconsistency Data != Erroneous Data
  - Minimum Change != Correct Change
- 基于模型的方法 Machine Learning Models:
  - Learn models with the existing data for data cleaning
  - Can NOT guarantee correctness
- 基于众包的方法 Crowdsourcing:
  - Let the Crowd help clean the data: can reach a much higher precision/recall
  - *e.g.*, *find errors, fill blanks, make choices between conflicts*
  - Expensive!
- 混合方法 Hybrid: Rule-based+Crowd, Model-based+Crowd, ...



# 关系数据库中的错误数据清洗

## Example

- Schema: Cust(country, area-code, phone, street, city, zip)
- Instance:

country	area-code	phone	street	city	zip
44	131	1234567	Mayfield	NYC	EH4 8LE
44	131	3456789	Crichton	NYC	EH4 8LE
01	908	3456789	Mountain Ave	NYC	07974

- ✓ functional dependencies (FDs):

cust[country, area-code, phone]  $\rightarrow$  cust[street, city, zip]

cust[country, area-code]  $\rightarrow$  cust[city]

The database satisfies the FDs. **Is the data consistent?**

The Challenge

# 关系数据库中的错误数据清洗

## Example

- $\text{cust}([\text{country} = 44, \text{zip}] \rightarrow [\text{street}])$ 
  - In the UK, zip code uniquely determines the street
  - The constraint may not hold for other countries
- It expresses a fundamental part of the semantics of the data
- It can NOT be expressed as a traditional FD

country	area-code	phone	street	city	zip
44	131	1234567	Mayfield	NYC	EH4 8LE
44	131	3456789	Crichton	NYC	EH4 8LE
01	908	3456789	Mountain Ave	NYC	07974

# 关系数据库中的错误数据清洗

## Example

- $\text{cust}([\text{country} = 44, \text{area-code} = 131, \text{phone}] \rightarrow [\text{street}, \text{zip}, \text{city} = \text{EDI}])$
- $\text{cust}([\text{country} = 01, \text{area-code} = 908, \text{phone}] \rightarrow [\text{street}, \text{zip}, \text{city} = \text{MH}])$ 
  - In the UK, if the area code is 131, then the city has to be EDI
  - In the US, if the area code is 908, then the city has to be MH
- $t_1, t_2$  and  $t_3$  violate these constraints
  - refining  $\text{cust}([\text{country}, \text{area-code}, \text{phone}] \rightarrow [\text{street}, \text{city}, \text{zip}])$
  - combining constants and variables

id	country	Area-code	phone	street	city	zip
t1	44	131	1234567	Mayfield	NYC	EH4 8LE
t2	44	131	3456789	Crichton	NYC	EH4 8LE
t3	01	908	3456789	Mountain Ave	NYC	07974

# 知识图谱中的错误数据清洗

- 如何清洗知识图谱中的错误数据？
  - 关系数据库中的数据清洗方法都可以借鉴使用。
  - 在知识图谱内部做知识推理，是一种可能的方案。
    - 知识推理技术还比较初级
    - 推理只可发现有限的错误
  - 借助互联网等外部数据对Kg数据进行数据验证，是可行的方案。
    - 需要有可用外部数据支撑
  - 借助众包来做数据清洗，也是比较靠谱的渠道。
    - 众包优化的代价需要优化，否则代价较高

# 本节大纲

- 知识图谱质量评估与控制概述
- 知识图谱数据来源的质量控制
- 知识图谱数据获取的质量控制
- 知识图谱数据融入的质量控制
- 知识图谱数据补全的质量控制
- 知识图谱数据更新的质量控制
  - 错误数据清洗
  - 过期数据更新



# 过期数据更新

- 随着时间的推移，数据是变动的
  - 一直在变的：人口，年龄，职位，作品数量，美国总统。。。。
  - 不断新增的：新人，新公司，新词，。。。。
- 如何保持知识图谱中数据的“新鲜度”？
  - 数据更新
- 数据的更新机制
  - 定期全局更新机制（缺点：耗时耗力）
  - 基于更新频率预测的更新机制（缺点：耗时、不准确、无法加入新词）
  - 基于热点词发现的更新机制（IJCAI 2017）

# 过期数据更新 - 基于热点词发现的更新

## • 基于热点事件发现的更新机制

- Basic Idea: 对互联网上的热词进行监控。一个实体之所以变成热词，会有两个原因。一个是新词，比如即将发布的iPhone8。另一个是旧词，但知识发生了变化，比如说特朗普变成美国总统了。

## • 整体框架



- We find hot entities from Web.

- We synchronize (update or insert) these seed entities.

- We find more related entities via the hyperlinks in the latest pages.

- We synchronize these expanded entities according to their priority, which is provided by a predictor.

(Liang et. al., How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source, IJCAI2017)

Thanks!

