Data Quality Issues in Constructing Knowledge Graph
知识图谱构建中的质量控制

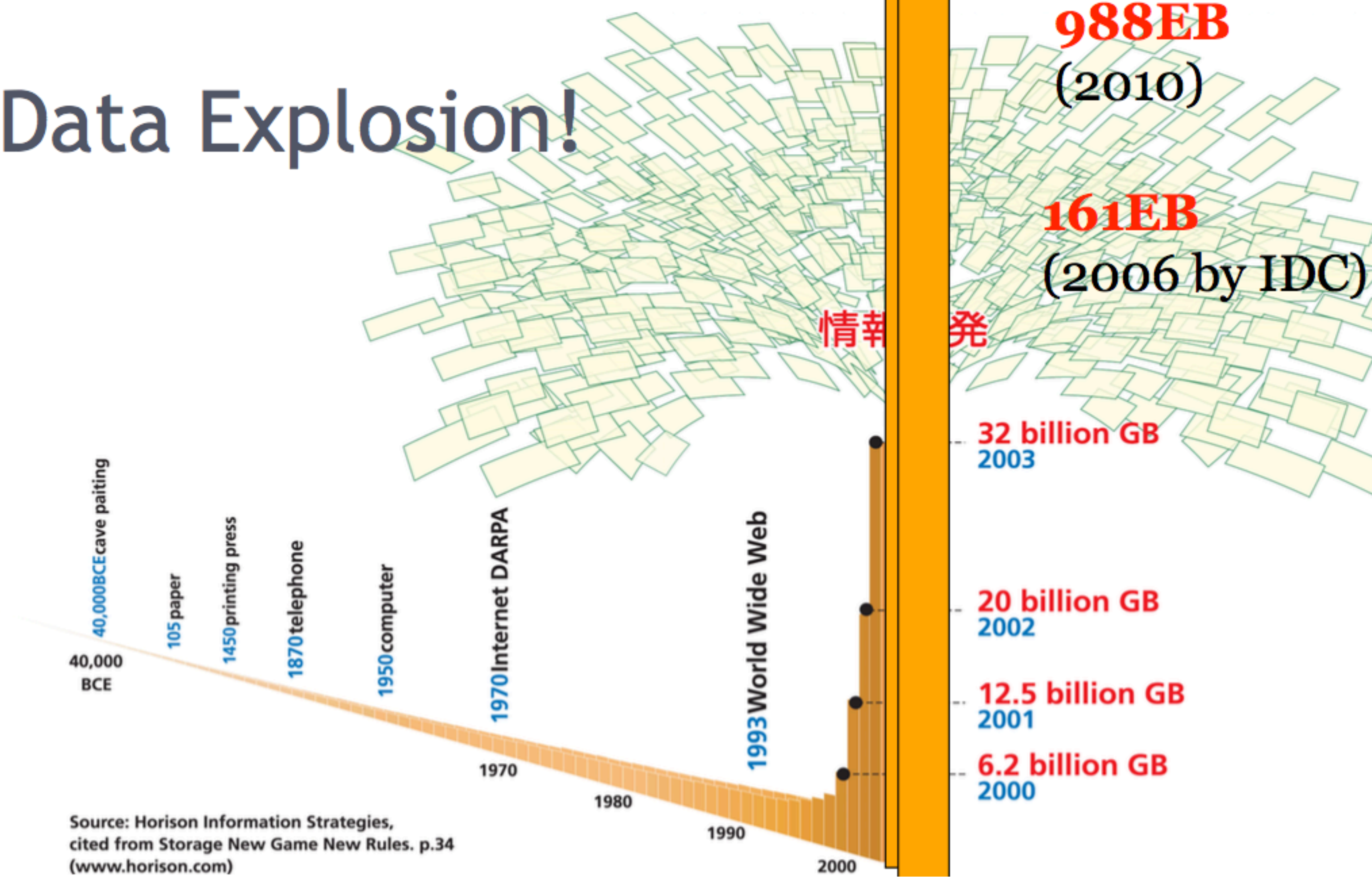
2017. 07. 13
华东师范大学

苏州大学 先进数据分析研究中心 李直旭

Outline

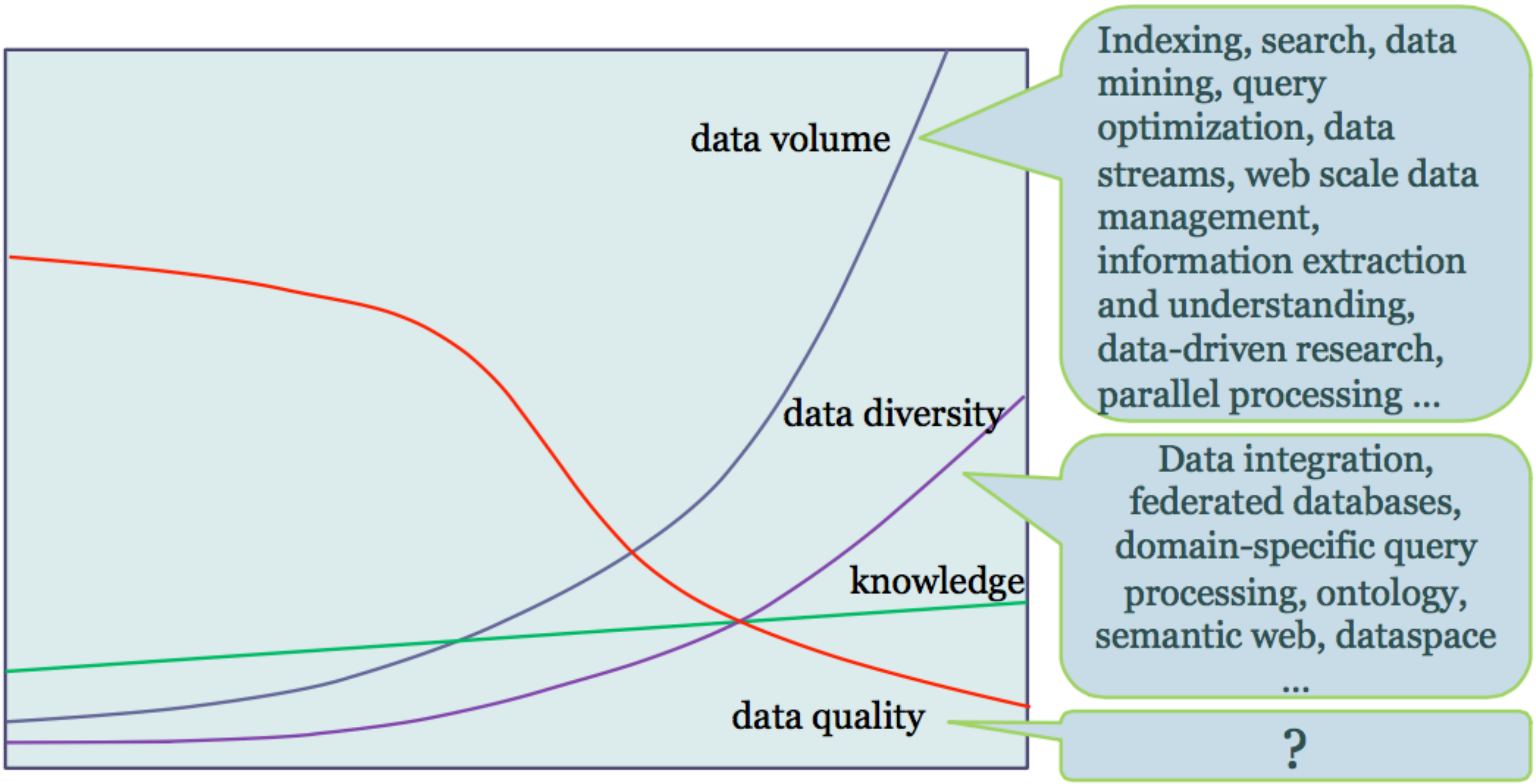
- Introduction to DQ
- Computational DQ Problems
- Data Quality Issues in Constructing KG
  - Data Cleaning in KG
  - Entity Linking in KG
  - Data Imputation in KG
- Conclusions

Data Explosion!



Source: Horison Information Strategies,
cited from Storage New Game New Rules. p.34
(www.horison.com)

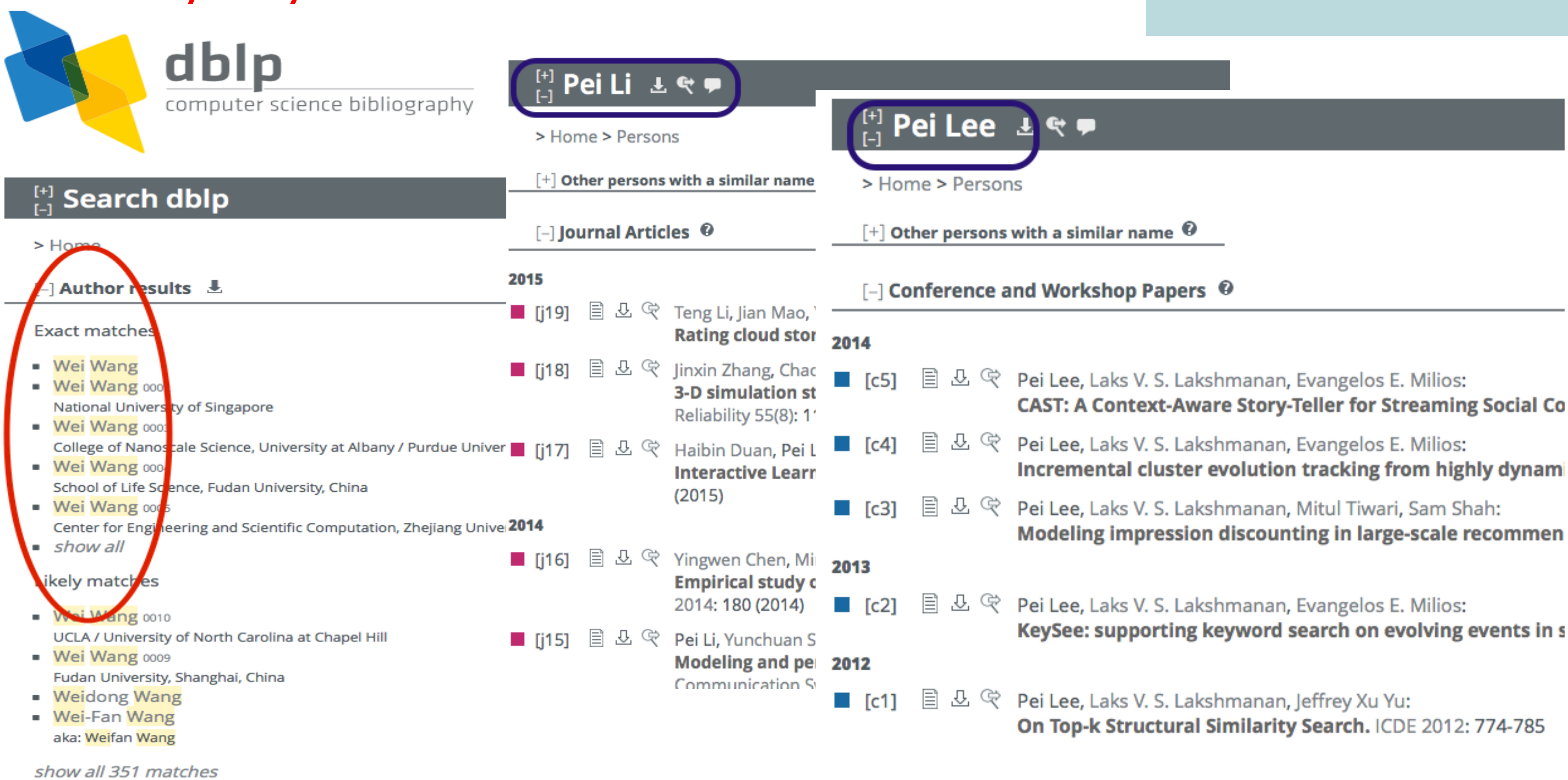
Knowledge Explosion?



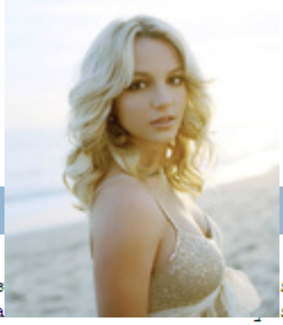
DQ Problems in DBLP

- 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



Difficult Names in Google Search



488941 britney spears	29 britent spears	9 brinttany spears	5 brney spears	3 britiy spears	spears
40134 brittany spears	29 brittnany spears	9 britanay spears	5 broitney spears	3 britmeny spea	spears
36315 brittney spears	29 britttany spears	9 britinany spears	5 brotny spears	3 britneeey spears	2 brirttany spears
24342 britany spears	29 btiney spears	9 britn spears	5 bruteny spears	3 britnehy spears	2 brirttney spears
7331 britny spears	26 birttney spears	9 britnew spears	5 btiyney spears	3 britnely spears	2 britain spears
6633 briteny spears	26 breitney spears	9 britneyn spears	5 btrittney spears	3 britnesy spears	2 britane spears
2696 britteny spears	26 brinity spears	9 britrney spears	5 gritney spears	3 britnetty spears	2 britaneny spears
1807 briney spears	26 britenay spears	9 brtiny spears	5 spritney spears	3 britnex spears	2 britania spears
1635 brittny spears	26 britneyt spears	9 brtittney spears	4 bittny spears	3 britneyxxx spears	2 britann spears
1479 brintey spears	26 brittan spears	9 brtny spears	4 bnritney spears	3 britnity spears	2 britanna spears
1479 britanny spears	26 brittne spears	9 brytny spears	4 brandy spears	3 britntey spears	2 britannie spears
1338 britiny spears	26 btittany spears	9 rbitney spears	4 brbritney spears	3 britnyey spears	2 britannt spears
1211 britnet spears	24 beitney spears	8 birtiny spears	4 breatiny spears	3 britterny spears	2 britannu spears
1096 britiney spears	24 birteny spears	8 bithney spears	4 breetney spears	3 brittneey spears	2 britanyl spears
991 britaney spears	24 brightney spears	8 brattany spears	4 bretiney spears	3 brittnney spears	2 britanyt spears
991 britnay spears	24 brintiny spears	8 breitny spears	4 brfitney spears	3 brittnyey spears	2 briteeny spears
811 brithney spears	24 britanty spears	8 breteny spears	4 briattany spears	3 brityen spears	2 britenany spears
811 brtiney spears	24 britenny spears	8 brightny spears	4 brieteny spears	3 briytney spears	2 britenet spears
664 birtney spears	24 britini spears	8 brintay spears	4 briety spears	3 brltney spears	2 briteniy spears
664 brintney spears	24 britnwy spears	8 brinttey spears	4 briitny spears	3 broteny spears	2 britenys spears
664 briteney spears	24 brittni spears	8 briotney spears	4 briittany spears	3 brtaney spears	2 britianey spears
601 bitney spears	24 brittnie spears	8 britanys spears	4 brinie spears	3 brtiiany spears	2 britin spears
601 brinty spears	21 biritney spears	8 britley spears	4 brinteney spears	3 brtinay spears	2 britinary spears
544 brittaney spears	21 birtany spears	8 britneyb spears	4 brintne spears	3 brtinney spears	2 britmy spears
544 brittnay spears	21 biteny spears	8 britnrey spears	4 britaby spears	3 brtitany spears	2 britnaney spears
364 britey spears	21 bratney spears	8 britnty spears	4 britaey spears	3 brtiteny spears	2 britnat spears
364 brittiny spears	21 britani spears	8 brittner spears	4 britainey spears	3 brtnet spears	2 britnbey spears
329 brtney spears	21 britanie spears	8 brottany spears	4 britinie spears	3 brytiny spears	2 britndy spears
269 bretney spears	21 briteany spears	7 baritney spears	4 britinney spears	3 btney spears	2 britneh spears
269 britneys spears	21 brittay spears	7 birntey spears	4 britmney spears	3 drittney spears	2 britneney spears
244 britne spears	21 brittinay spears	7 biteney spears	4 britnear spears	3 pretney spears	2 britney6 spears
244 brytney spears	21 brtany spears	7 bitiny spears	4 britnel spears	3 rbritney spears	2 britneye spears
220 breatney spears	21 brtiany spears	7 breateny spears	4 britneuy spears	2 barittany spears	2 britneyh spears
220 britiany spears	19 birney spears	7 brianty spears	4 britnewy spears	2 bbbritney spears	2 britneym spears
199 britnney spears	19 birrtney spears	7 brintye spears	4 britnmey spears	2 bbitney spears	2 britneyyy spears
163 britnry spears	19 britnaey spears	7 britianny spears	4 brittaby spears	2 bbritny spears	2 britnhey spears
147 breatny spears	19 britnee spears	7 britly spears	4 brittery spears	2 bbrittany spears	2 britnjey spears
147 brittiney spears	19 britony spears	7 britnej spears	4 britthey spears	2 beitany spears	2 britnne spears
147 britty spears	19 brittanty spears	7 britneyu spears	4 brittnaey spears	2 beitny spears	2 britnu spears
147 brotney spears	19 britttney spears	7 britniey spears	4 brittnat spears	2 bertney spears	2 britoney spears
147 brutney spears	17 birtny spears	7 britnnay spears	4 brittneny spears	2 bertny spears	2 britrany spears
133 britteney spears	17 brieny spears	7 brittian spears	4 brittnye spears	2 betney spears	2 britreny spears
133 briyney spears	17 brintty spears	7 briyny spears	4 britttteny spears	2 betny spears	2 britry spears
121 bittany spears	17 brithy spears	7 brrittany spears	4 briutney spears	2 bhriney spears	2 britsany spears

Another Example with KBs

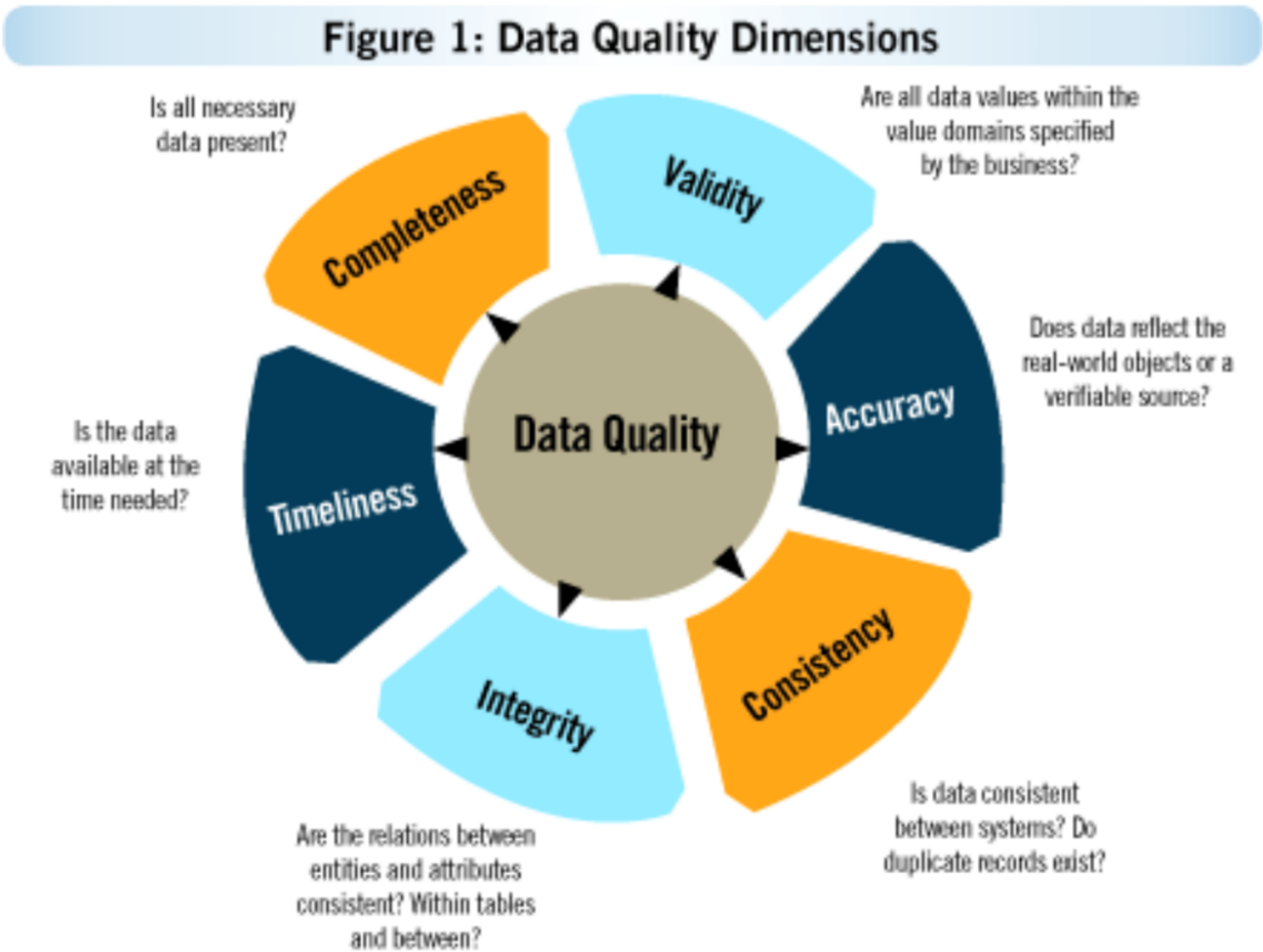
CID	Name	Address	City	Sex
11	张三	邯郸路 220 号计算机楼 527 室	上海	0
24	李四	鄞奉路 978 号 7 号楼 702 室	宁波	1

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	宁波市鄞奉路 978 号 7 号楼 702 室 315012	0574-87074789

NO	Name	Gender	Address	city	zip	Pone	Fax	CID	Cno
1	张三	F	邯郸路 220 号计算机楼 527 室	上海				11	
2	李四	M	鄞奉路 978 号 7702 室	宁波	315012	0574-87074789		24	493
3	王五	F	朝二区 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”
- **Missing values**

Six DQ Dimensions



The Taxonomy of DQ Problems

- Data quality Problems 数据质量问题
  - Single Data Source 单数据源问题
    - Schema 模式相关
      （缺乏完整性约束，粗劣的设计模式）
      - — 唯一值 Uniqueness
      - — 参考完整性 Referential Integrity
    - Instance 实例相关
      （数据输入错误）
      - — 拼写错误 Typos
      - — 冗余 / 重复 Redundancy
      - — 矛盾的数据 Conflicts
  - Multiple Data Source 多数据源问题
    - Schema 模式相关
      （不同的数据模型和模式设计）
      - — 命名冲突 Naming Conflicts
      - — 结构冲突 Structural Conflicts
    - Instance 实例相关
      （矛盾的或不一致的数据）
      - — 不一致的聚集层级 Inconsistency Aggregation Hierarchies
      - — 不一致的时间点 Inconsistency Time Points

Computational Data Quality Problems

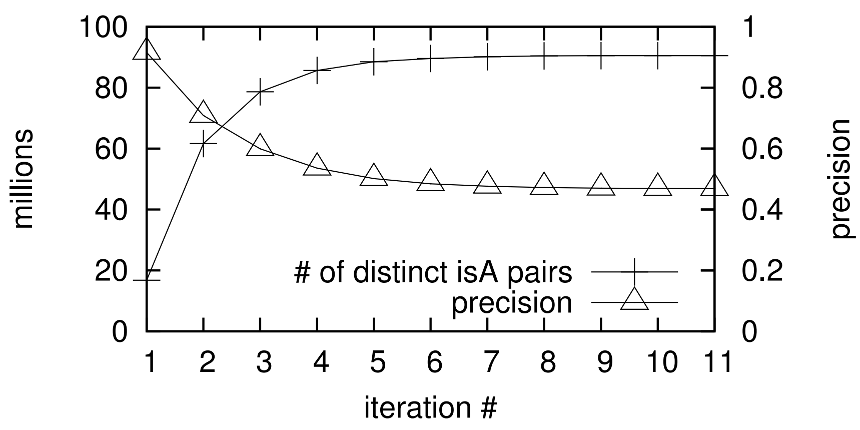
- Data Integration
    - Schema Mapping
    - Record Matching
- Data Cleaning
- Data Imputation
- Data Provenance
- Data Uncertainty
- Data Constraints

Outline

- Introduction to DQ
- Computational DQ Problems
- Data Quality Issues in Constructing KG
  - Data Cleaning in KG
  - Entity Linking in KG
  - Data Imputation in KG
- Conclusions

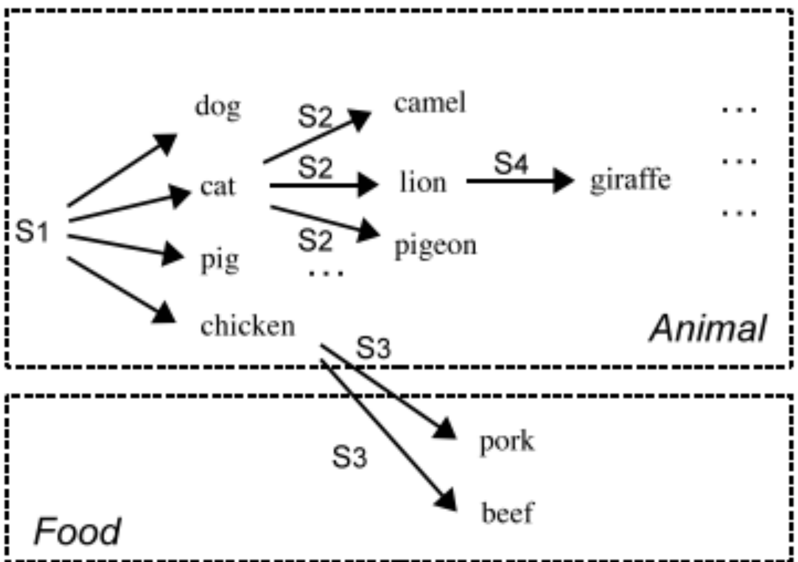
Data Cleaning in Constructing KG

- Open IE -> Knowledge Graph
- Bootstrapping Mechanisms
  - e.g.: *KnowItAll*, *SnowBall*, *ProBase* …
- However, the *accuracy decreases sharply* after several iterations.



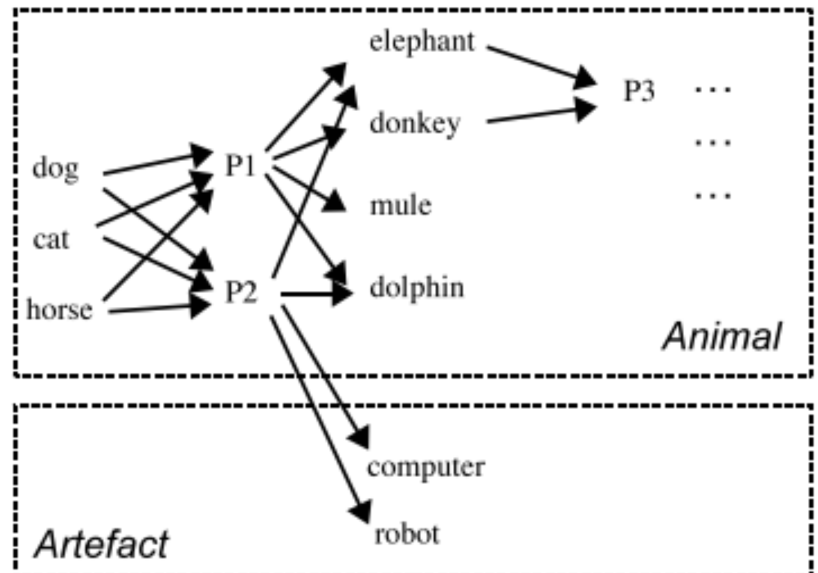
Data Cleaning in KG

- A Major Reason - **Semantic drift** happens



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

Data Cleaning in KG

- **Mainstream approaches**
  - 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)

Data Cleaning in KG

- **Mainstream approaches**
  - 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)

Data Cleaning in KG

- 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

Data Cleaning in KG

- **Mainstream approaches**
  - 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)

Data Cleaning in KG

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

X ,which is based in Y

Pillar, San Jose OK

Type Checking Arguments:

…companies such as Pillar…

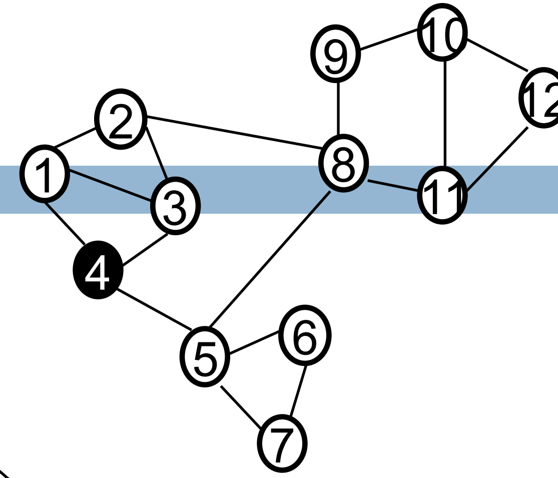
… cities like San Jose…

Inclined pillar , foundation plate **NO**

Data Cleaning in KG

- **Mainstream approaches**
  - 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)

Data Cleaning in KG



- 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

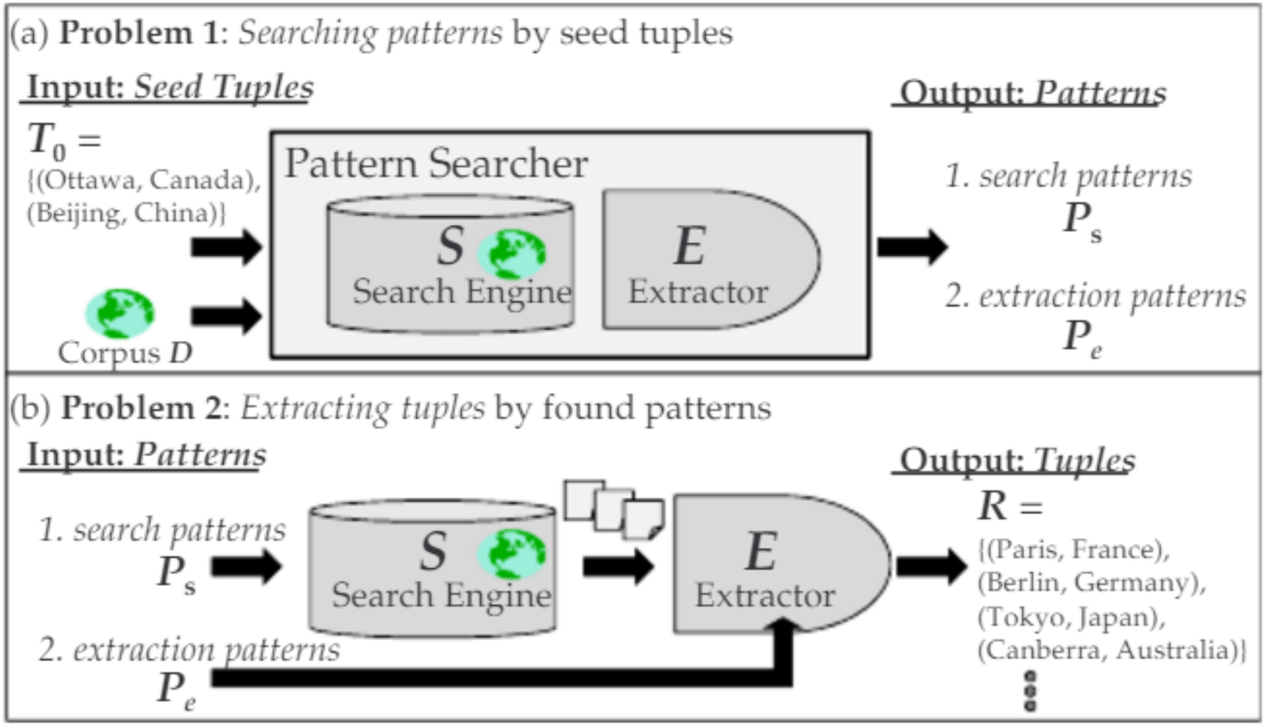
$$\begin{pmatrix}0.13\\0.10\\0.13\\0.22\\0.13\\0.05\\0.05\\0.08\\0.04\\0.03\\0.04\\0.02\end{pmatrix} = 0.9\times\begin{pmatrix}
0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 1/4 & 0 & 0 & 0 & 0\\
1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
1/3 & 0 & 1/3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

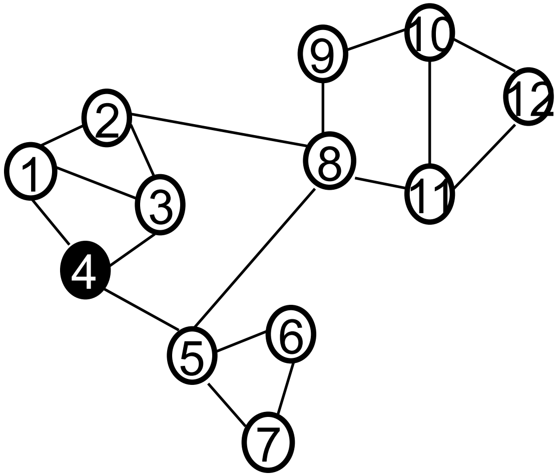
Data Cleaning in KG

- **Mainstream approaches**
  - 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)

Data Cleaning in KG

- **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-Score = Precision+Recall

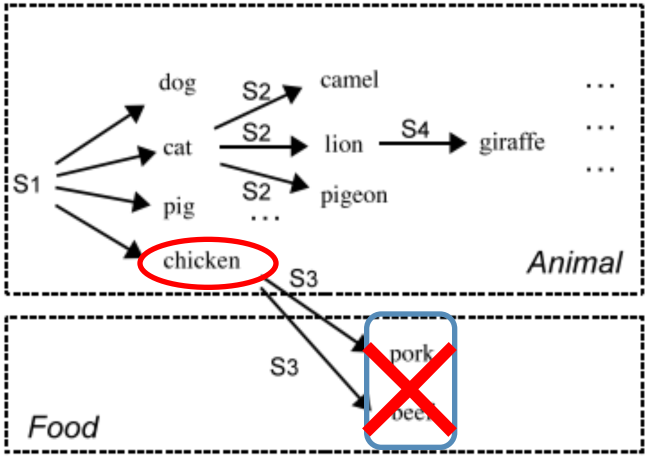
Ranking with F-Score

Data Cleaning in KG

- **Mainstream approaches**
  - 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)

Data Cleaning in KG

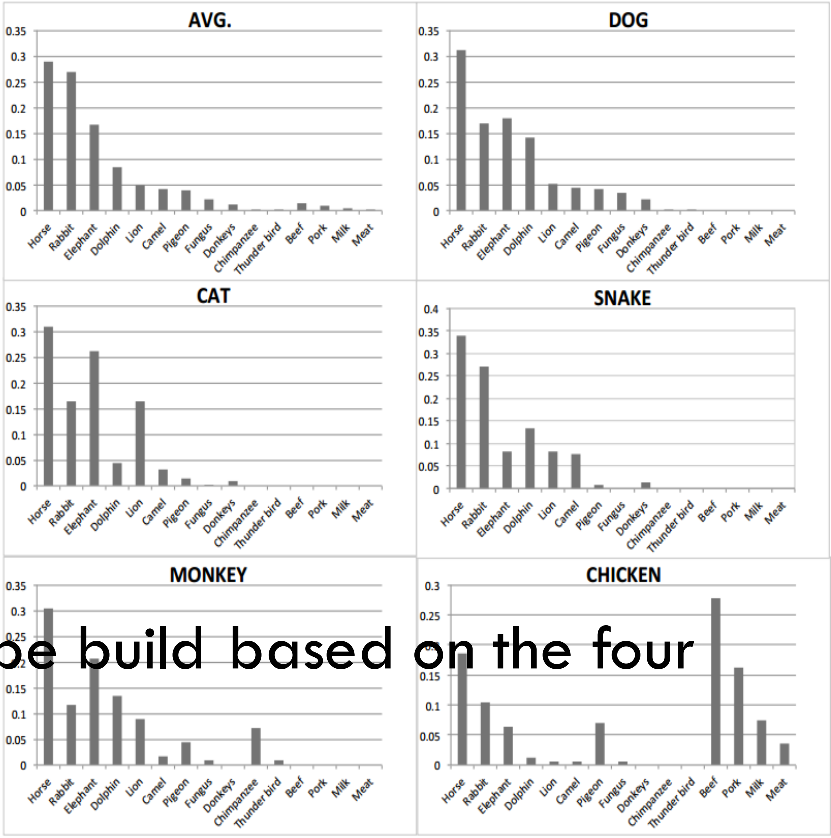
- Cleaning Model based on Detected **Drifting Points**
  - **Intuition:** **Drifting Points (DPs)** are the reasons of Semantic Drift.
  - Two kinds of DPs
    - Intentional DPs
      - Synonyms such as Chicken
    - Accidental DPs
      - Errors by themselves
      - E.g., … Countries such as France, Germany, Japan and New York.



Z. Li et al., Overcoming Semantic Drift in Information Extraction, EDBT'14

Data Cleaning in KG

- Properties of DPs
  - For a target class, the distribution of instances triggered by a DP is different from the distribution of instances that truly belong to the target class.



  - If classes C_1 and C_2 are mutually … an Intentional DP.
  - An accidental DP is usually suppor… instance is derived from very few (…
  - An error extraction (e isA C) trigg… evidence, since the extraction is usu…

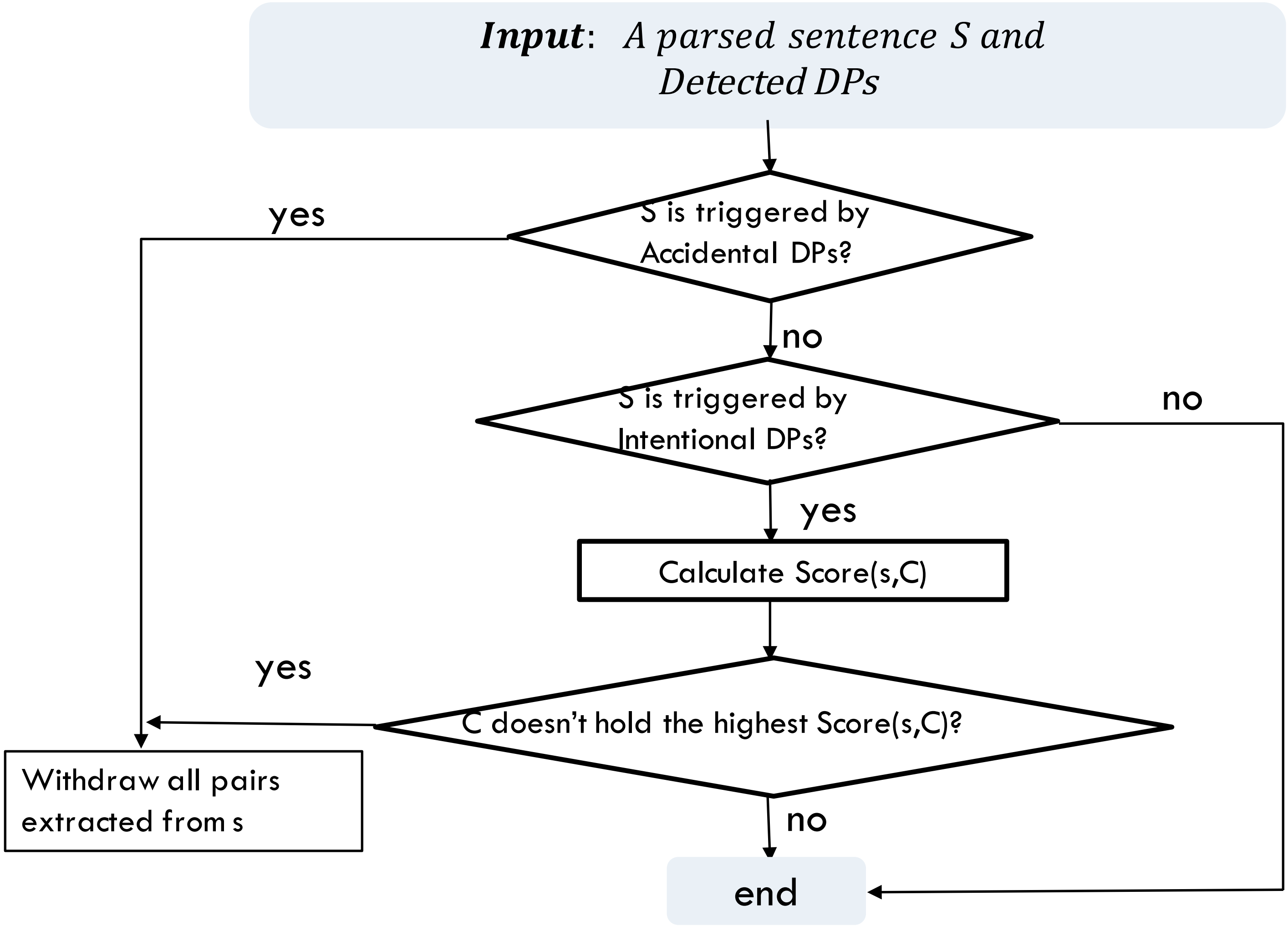
- A DP Detection Model can be build based on the four properties of DPs.

Z. Li et al., Overcoming Semantic Drift in Information Extraction, EDBT'14

Distributions of instances triggered by DPs and non-DPs

Data Cleaning in KG

- Finding Errors based on detected DPs



Z. Li et al., Overcoming Semantic Drift in Information Extraction, EDBT'14

Data Cleaning in KG – Experiments

Cleaning Method	p_{error}	r_{error}	$p_{correct}$	$r_{correct}$
Before Cleaning	-	-	0.4305	1.0
MEx	0.9119	0.1570	0.4592	**0.9832**
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	**0.9696**	**0.9145**	**0.8921**	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

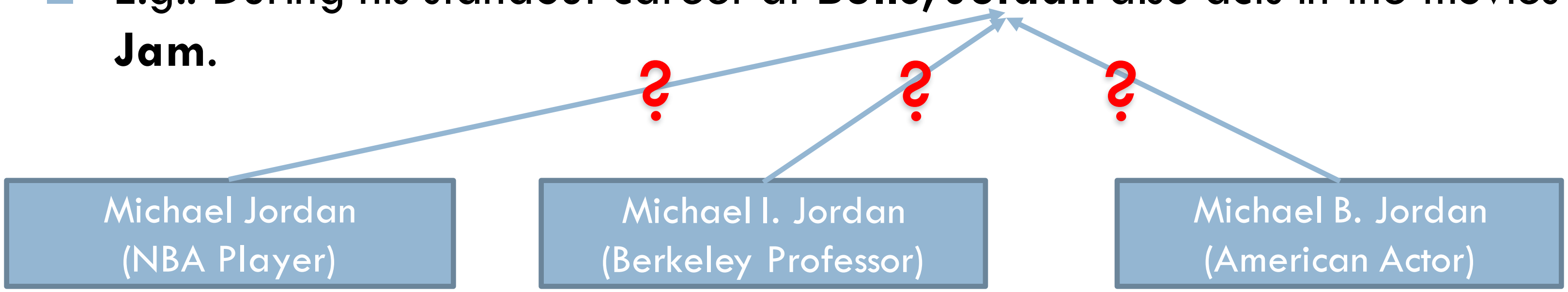
Z. Li et al., Overcoming Semantic Drift in Information Extraction, EDBT'14

Outline

- Conventional Data Quality Problems
  - Introduction to DQ
  - Computational DQ Problems and Solutions
- Data Quality Issues in Knowledge Graph
  - Data Cleaning in KG
  - Entity Linking in KG
  - Data Imputation in KG
- Conclusions

Entity Linking in KG

- 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)

Entity Linking in KG – Polysemy

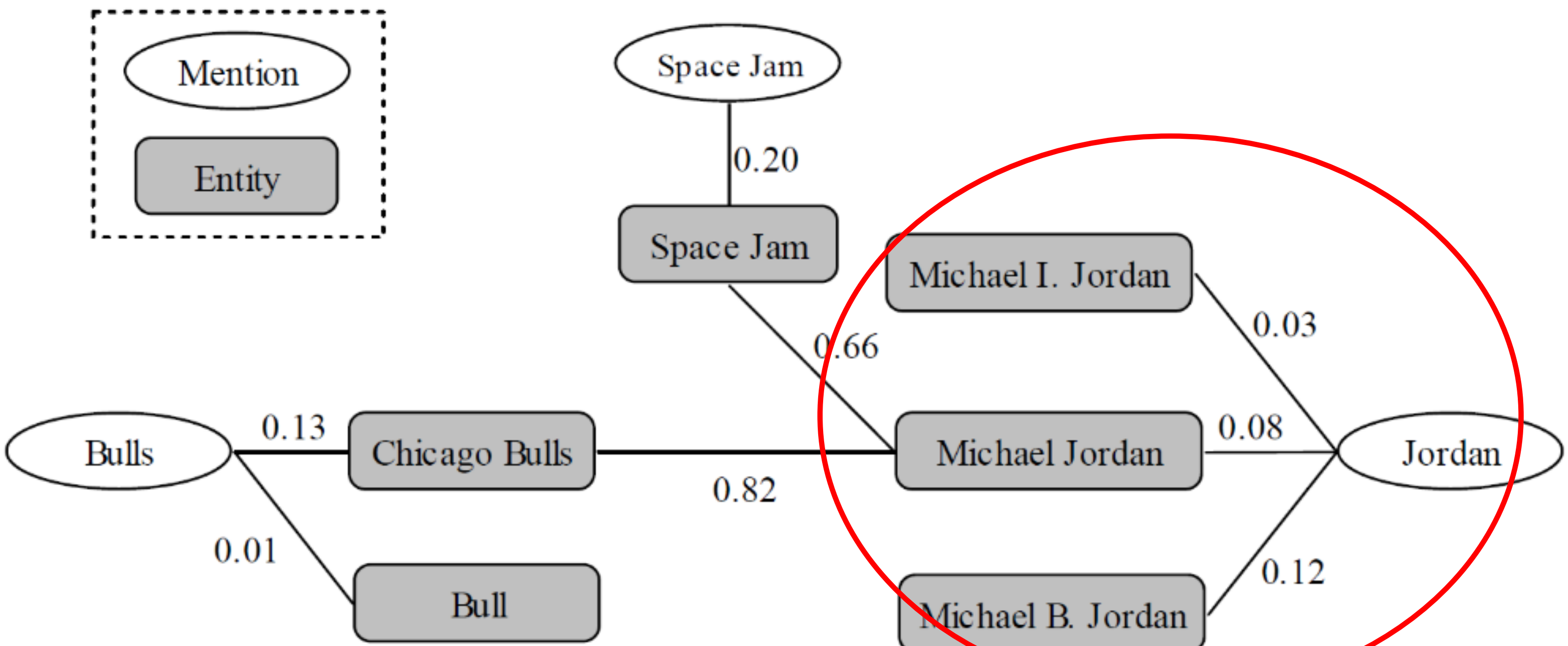
- **Main Approaches for Solving Polysemy**
  - EL based on Local Compatibility (CIKM'07, EMNLP'07, IJCAI'09, COLING'10…)
  - EL Based on Simple Relations (CIKM'08, AAAI'08)
  - Pair-Wise Collective EL Approaches (ACL'10)
  - Graph-Based Collective EL Approaches (SIGIR'11, 14)

Entity Linking in KG – Polysemy

- **Main Approaches for Solving Polysemy**
  - 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)

Entity Linking in KG – 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.



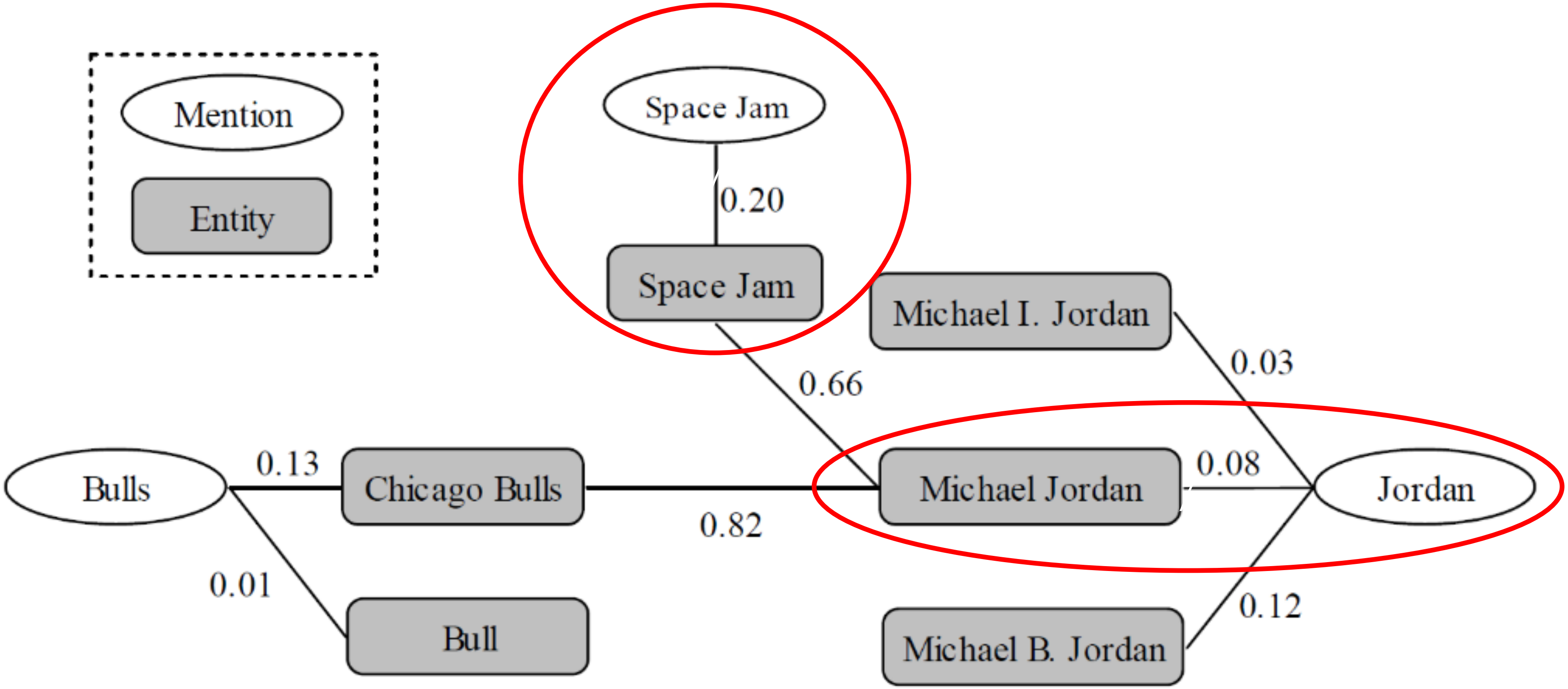
During his standout career at **Bulls**, **Jordan** also acts in the movies **Space Jam**.

Entity Linking in KG – Polysemy

- **Main Approaches for Solving Polysemy**
  - 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)

Entity Linking in KG – Polysemy

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

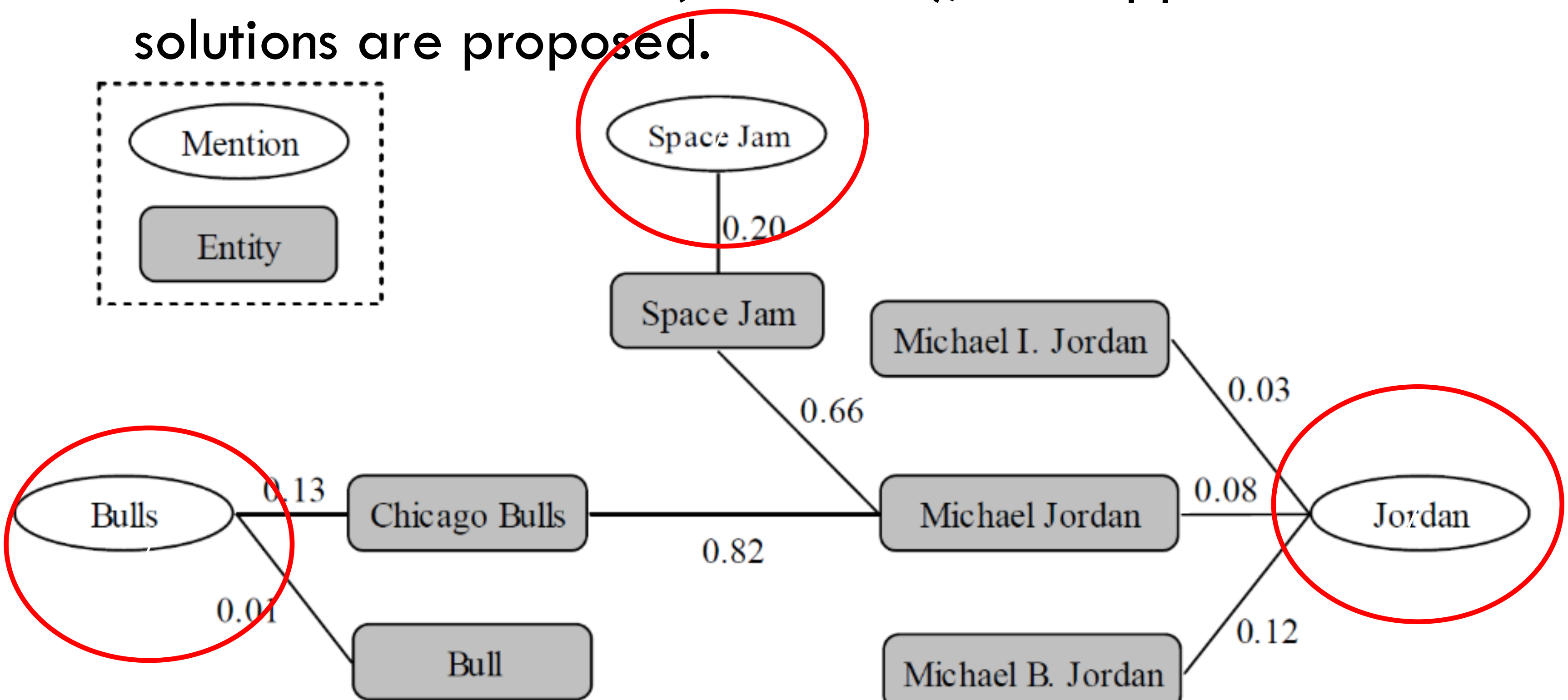


Entity Linking in KG – Polysemy

- **Main Approaches for Solving Polysemy**
  - 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)

Entity Linking in KG – 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.



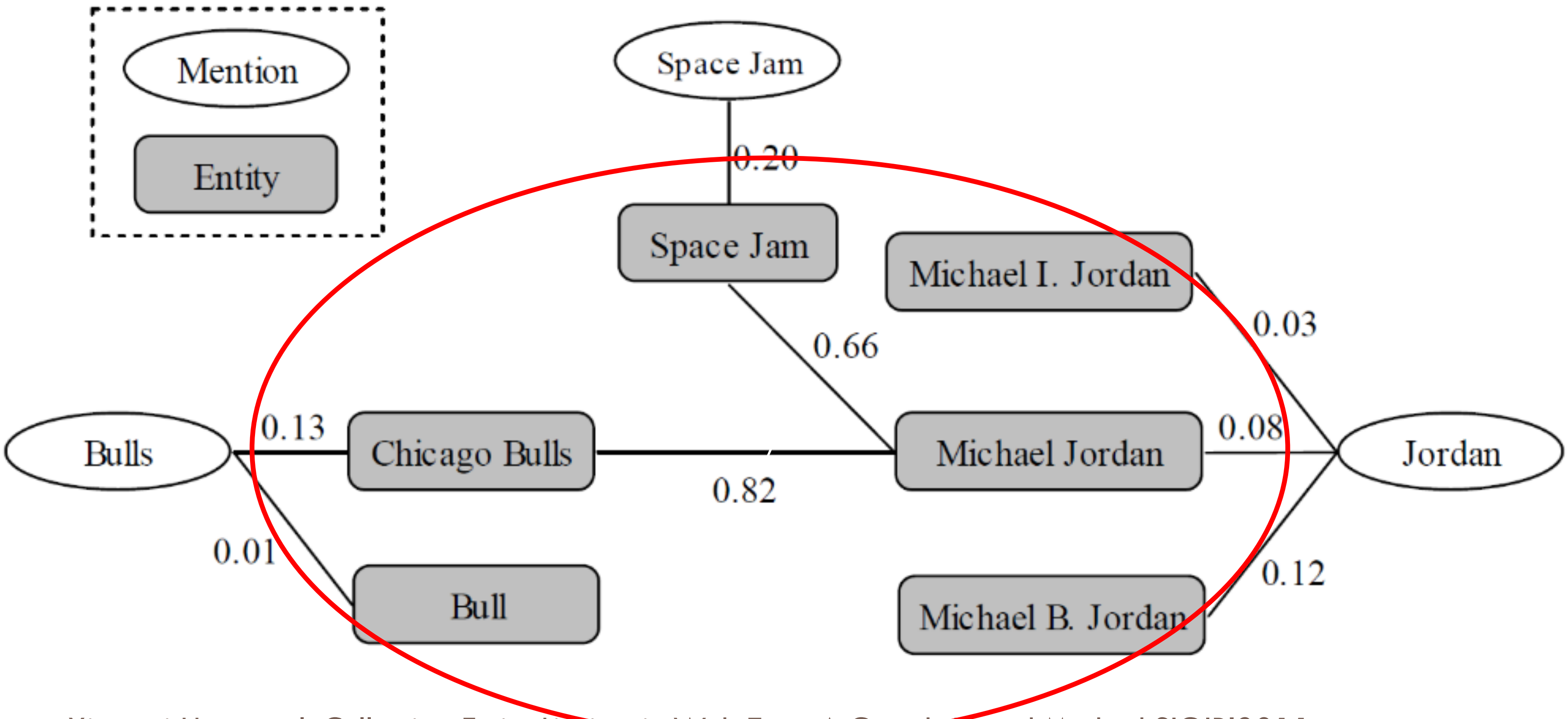
Kulkarni, S. at al, *Collective annotation of Wikipedia entities in web text*. SIGKD'09

Entity Linking in KG – Polysemy

- **Main Approaches for Solving Polysemy**
  - 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)

Entity Linking in KG – Polysemy

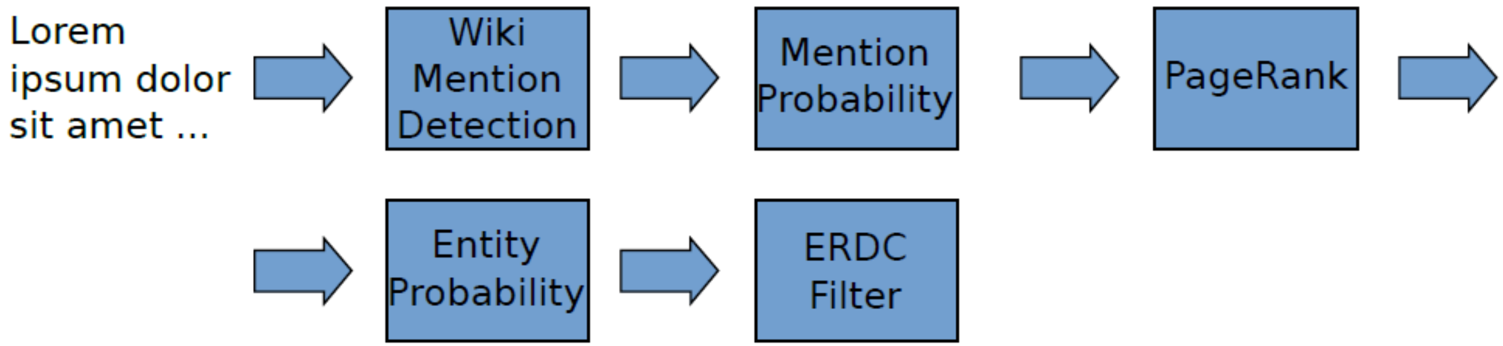
- Graph-Based Collective Approaches(SIGIR 11,14)
  - **Idea**: Model and exploit the global interdependence by graph-based collective EL method



Xianpei Han at al. Collective Entity Linking in Web Text: A Graph-Based Method.SIGIR'2011

Entity Linking in KG – Polysemy

- Graph-Based Collective Approaches(SIGIR 14)



Entity Linking in KG

- 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.**

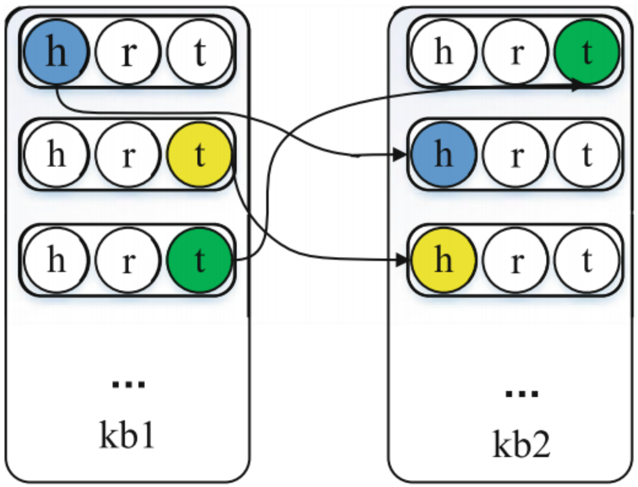
- **2. Synonyms（多词一义）**
- E.g.: Barack Hussein Obama(USA president)
  - m.02mjmr(Freebase)
  - Barack_Obama(Dbpedia)
  - 贝拉克·侯赛因·奥巴马(CN-Dbpedia)

Entity Linking in KG – 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 <u>probabilistic matching model</u>
  - Based on structure similarity only (CCKS'16)
    - Whole Knowledge Base Embedding

Entity Linking in KG – 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"



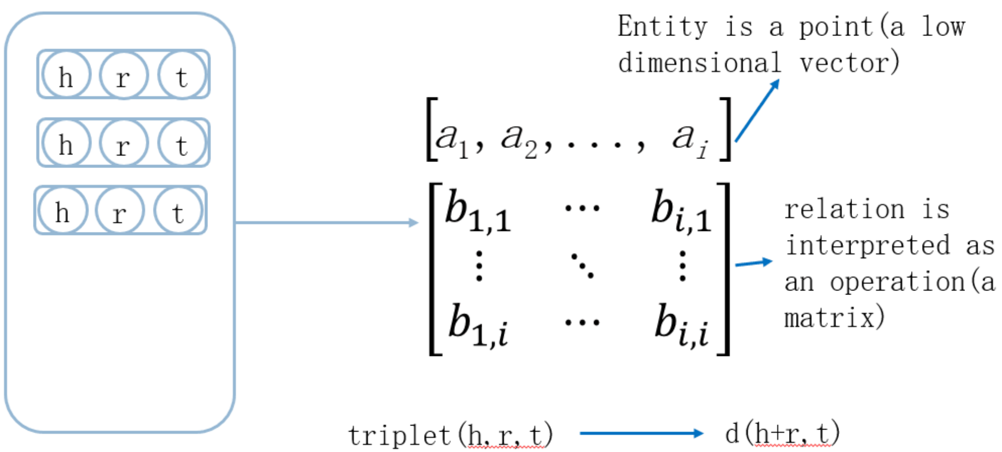


Fig. 2. Selecting seed entities in two KBs.

Yanchao Hao at al. A Joint Embedding Method for Entity Alignment of Knowledge Bases. CCKS 2016

Outline

- Conventional Data Quality Problems
  - Introduction to DQ
  - Computational DQ Problems and Solutions
- Data Quality Issues in Knowledge Graph
  - Data Cleaning in KG
  - Entity Linking in KG
  - Data Imputation in KG
- Conclusions

Data Imputation in KG

- Data Imputation in KG aims at increasing the coverage of KG

- Tasks
  - Missing entities
  - Missing types for entities (known as classification)
  - Missing relations that hold between entities

Data Imputation in KG – Approaches

- Type Assertions
  - Internal Knowledge-based
    - SDType (ISWC'13); and some other methods
  - External Knowledge-based
    - Tipola (ISWC'12); Classifier based on Wiki Links (LDOW'12)

- Relation Prediction
  - Internal Knowledge-based
    - Neural Tensor Network (NIPS'13) ; Mining Association Rules(ISWC'15)
  - External Knowledge-based
    - Matching HTML Tables to DBpedia(WIMS'15); and some other methods

Data Imputation in KG – Approaches

- Type Assertions
  - Internal Knowledge-based
    - SDType (ISWC'13); and some other methods
  - External Knowledge-based
    - Tipola (ISWC'12); Classifier based on Wiki Links (LDOW'12)

- Relation Prediction
  - Internal Knowledge-based
    - Neural Tensor Network (NIPS'13) ; Mining Association Rules(ISWC'15)
  - External Knowledge-based
    - Matching HTML Tables to DBpedia(WIMS'15); and some other methods

Internal Methods for Type Assertions

- **SDType:** 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 (%)
owl:Thing	100.0	88.6
dbpedia-owl:Place	69.8	87.6
dbpedia-owl:PopulatedPlace	0.0	84.7
dbpedia-owl:ArchitecturalStructure	50.7	0.0
dbpedia-owl:Settlement	0.0	50.6
dbpedia-owl:Building	34.0	0.0
dbpedia-owl:Organization	29.1	0.0
dbpedia-owl:City	0.0	24.2
...	...	...

x dbpedia-owl: location :y

assign

P(?x a dbpedia-owl:Place) = 0.698

P(?y a dbpedia-owl:Place) = 0.876

H. Paulheim et al. Type Inference on Noisy RDF Data, ISWC'2013

Internal Methods for Type Assertions

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^{-1}	140
...	...	...

② Compute basic distributions

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

③ Compute weights and conditional probabilities

predicate	weight
dpbedia-owl: federalState	0.3337534
dbpedia-owl: birthPlace^{-1}	0.0523772
...	...

predicate	type	probability
dbpedia-owl: federalState	dbpedia-owl: Place	1.0000000
dbpedia-owl: birthPlace^{-1}	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
...	...	...

Internal Methods for Type Assertions

- **Other Internal methods**
  - 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.

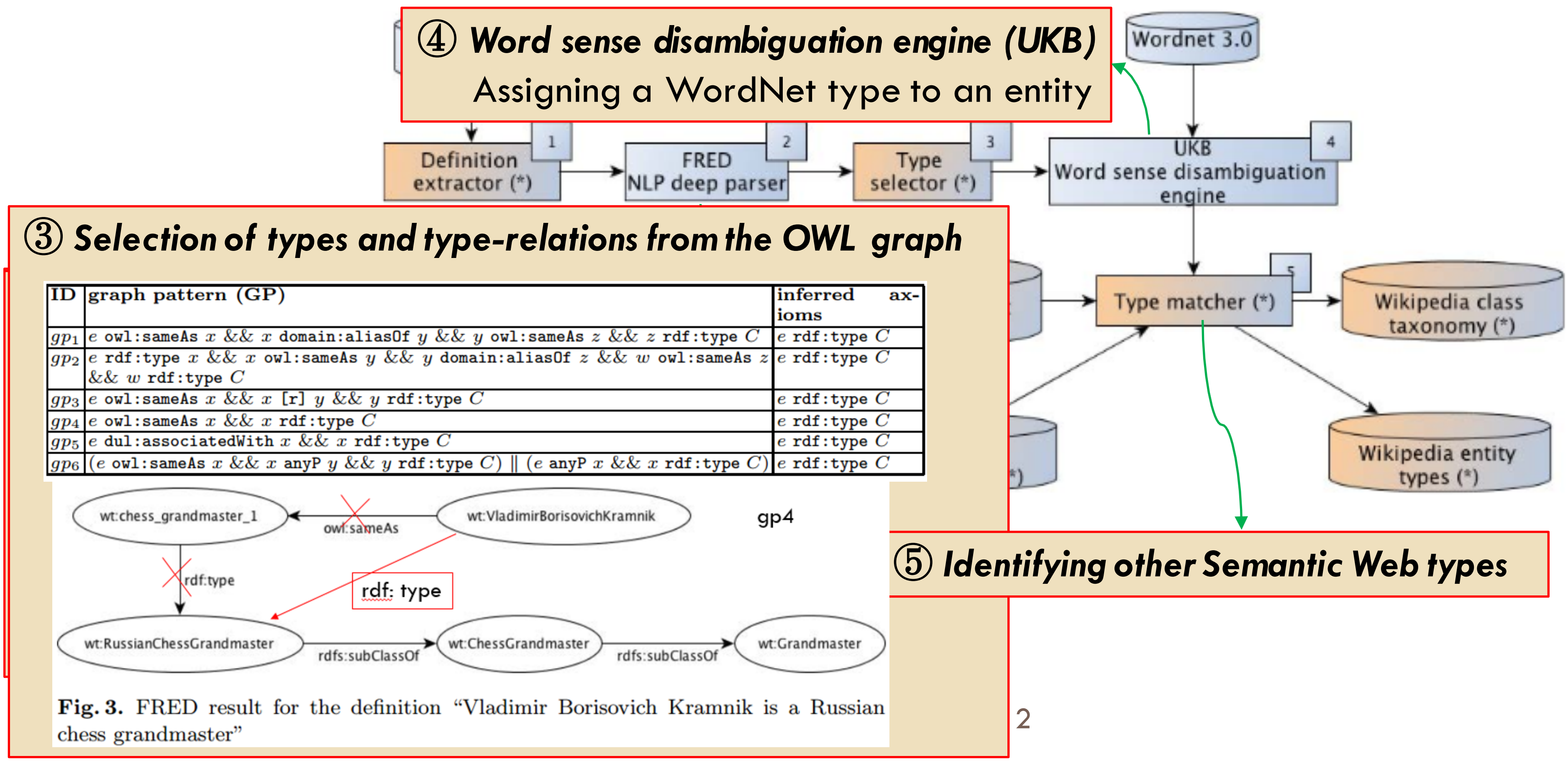
Data Imputation in KG – Approaches

- Type Assertions
  - Internal Knowledge-based
    - SDType (ISWC'13); and some other methods
  - External Knowledge-based
    - Tipola (ISWC'12); Classifier based on Wiki Links(LDOW'12)

- Relation Prediction
  - Internal Knowledge-based
    - Neural Tensor Network (NIPS'13) ; Mining Association Rules(ISWC'15)
  - External Knowledge-based
    - Matching HTML Tables to DBpedia(WIMS'15); and some other methods

External Methods for Type Assertions

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

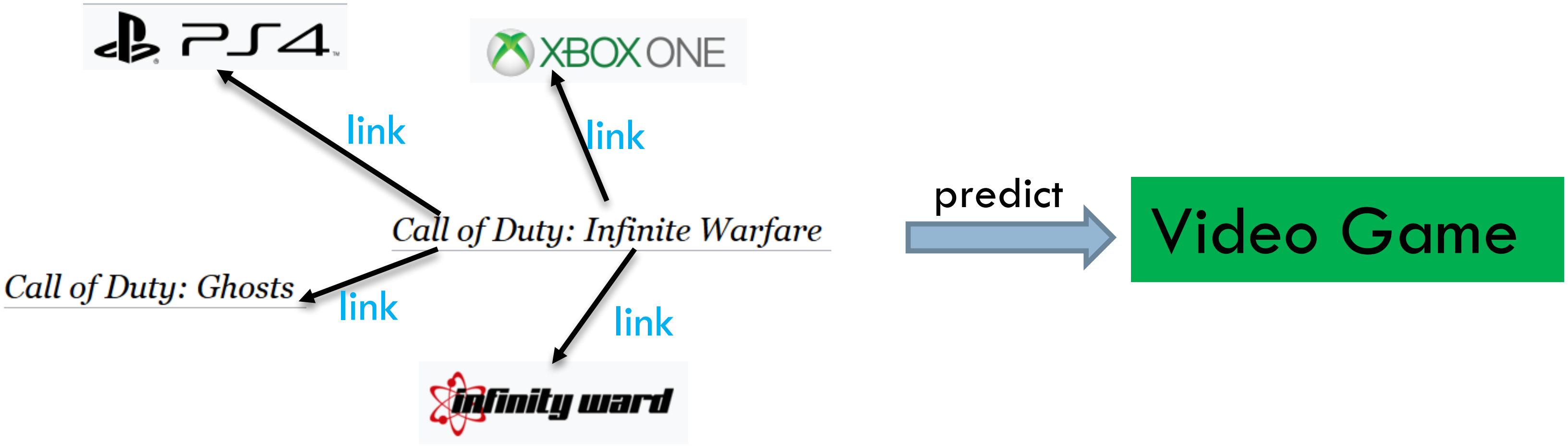


ID	graph pattern (GP)	inferred axioms
gp_1	e owl:sameAs x && x domain:aliasOf y && y owl:sameAs z && z rdf:type C	e rdf:type C
gp_2	e rdf:type x && x owl:sameAs y && y domain:aliasOf z && w owl:sameAs z && w rdf:type C	e rdf:type C
gp_3	e owl:sameAs x && x [r] y && y rdf:type C	e rdf:type C
gp_4	e owl:sameAs x && x rdf:type C	e rdf:type C
gp_5	e dul:associatedWith x && x rdf:type C	e rdf:type C
gp_6	(e owl:sameAs x && x anyP y && y rdf:type C) $\parallel$ (e anyP x && x rdf:type C)	e rdf:type C

Fig. 3. FRED result for the definition "Vladimir Borisovich Kramnik is a Russian chess grandmaster"

External Methods for Type Assertions

- 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

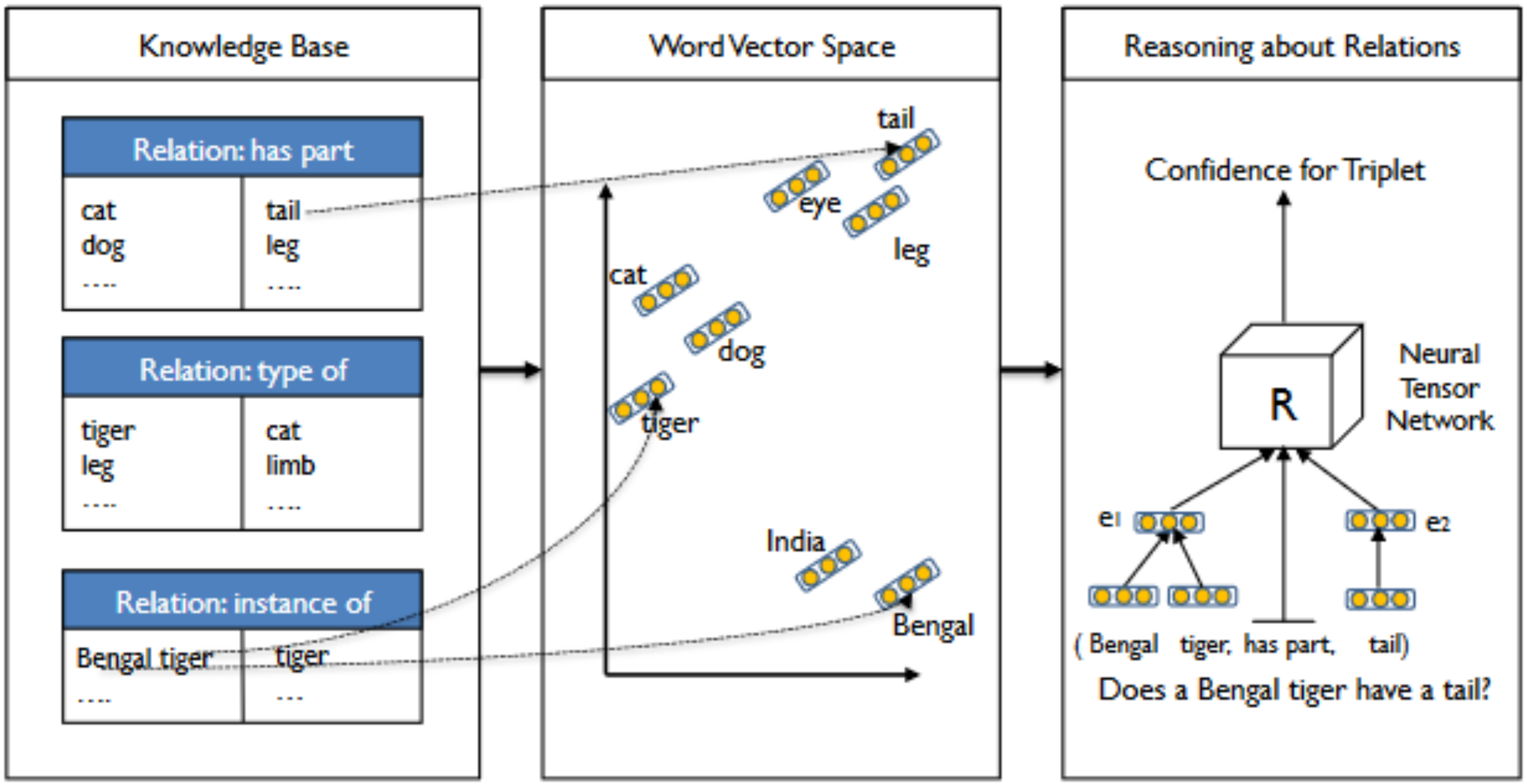
Data Imputation in KG – Approaches

- Type Assertions
  - Internal Knowledge-based
    - SDType (ISWC'13); and some other methods
  - External Knowledge-based
    - Tipola (ISWC'12); Classifier based on Wiki Links(LDOW'12)

- Relation Prediction
  - Internal Knowledge-based
    - Neural Tensor Network (NIPS'13) ; Mining Association Rules(ISWC'15)
  - External Knowledge-based
    - Matching HTML Tables to DBpedia(WIMS'15); and some other methods

Internal Methods for Relation Prediction

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



R. Socher et al. Reasoning with neural tensor networks for knowledge base completion, NIPS'13

Internal Methods for Relation Prediction

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

Stephen Smith’s Wikipedia categories

20th-century mathematicians

Philosophers who committed suicide

Alumni of King’s College, Cambridge

Stephen Smith’s relations

〈Stephen Smith, occupation, mathematicians〉

〈Stephen Smith, occupation, philosopher〉

〈Stephen Smith, graduated, Alumni of King’s College〉

Kim et al. The association rule mining system for acquiring knowledge of DBpedia from Wikipedia categories, ISWC’15

Data Imputation in KG – Approaches

- Type Assertions
  - Internal Knowledge-based
    - SDType (ISWC'13); and some other methods
  - External Knowledge-based
    - Tipola (ISWC'12); Classifier based on Wiki Links(LDOW'12)

- Relation Prediction
  - Internal Knowledge-based
    - Neural Tensor Network (NIPS'13) ; Mining Association Rules(ISWC'15)
  - External Knowledge-based
    - Matching HTML Tables to DBpedia(WIMS'15); and some other methods

External Methods for Relation Prediction

- **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 >

Ritze et al. Matching HTML Tables to Dbpedia, WIMS'15

External Methods for Relation Prediction

- 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

Mintz et al. Distant supervision for relation extraction without labeled data, ACL'09

Outline

- Conventional Data Quality Problems
  - Introduction to DQ
  - Computational DQ Problems and Solutions
- Data Quality Issues in Knowledge Graph
  - Data Cleaning in KG
  - Entity Linking in KG
  - Data Imputation in KG
- Conclusions

Conclusions

- Big Data -> Big Dirty Data
  - More Challenges …
  - More Opportunities…
- What can we do?
  - Use the rich knowledge
  - Better Precision and Recall
  - Pay Attention to Efficiency
  - Pay Attention to Cost

Thanks!