ML-Based Knowledge Graph Curation: Current Solutions and Challenges

Laure Berti-Equille

IRD Montpellier
Aix-Marseille University, CNRS, LIS, DIAMS
France
laure.berti@ird.fr
http://pageperso.lif.univ-mrs.fr/~laure.berti/

MEPDAW'19
5th Workshop on Managing the Evolution and Preservation of the Data Web

Data Quality Problems in KBs

What can go wrong ?

In DL:

- Invalid ABox: Class (concept), Property (role), Constant (individual)
- Invalid TBox: Set of axioms (Bad ontology design defining relationships: hierarchies, domains, ranges, etc.)

In RDF:

Invalid Triple:

<subject, property, object>

In KG:

Invalid Fact:

< head , relation , tail >

Invalid Reference to Extra-Information

- Mismatch of entity description
- Ambiguities in context mention

DATA QUALITY PROBLEMS

TYPE
Missing data
Anomalous data
Duplicate data
Inconsistent data
Obsolete data
Incorrect data

CARDINALITY
Single-Point
Collection

DETECTION/CORRECTION MODE

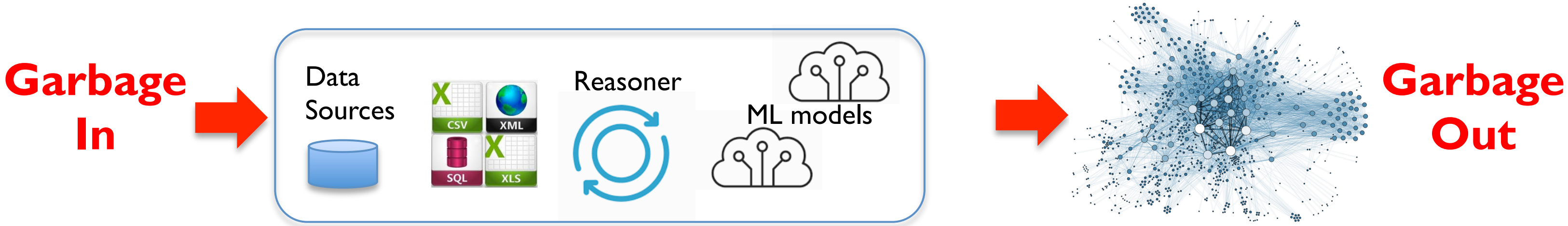
Manual Inspection:

- Expert and Human-In-the-Loop
- Find-Fix-Verify Crowdsourcing

Semi- or unsupervised techniques:

- Constraints, Rules, and Patterns
- Descriptive Statistics
- Model Inference and Machine Learning

Sources of errors in KB Construction/Population



Data Extraction	• Errors in unsupervised knowledge extraction from unstructured texts in open domain • Multi-lingual and cultural difficulties in information extraction • Identity problem due to context/description mismatch • Obsolescence
Entity Linking	• Accuracy of automatic data linking approaches and large-scale entity disambiguation
Knowledge Inference	• Inadequate knowledge representations (information loss) • Inadequacy of KG semantic embedding techniques for 1-N, N-1, and N-N relations
Knowledge Publishing	• Lack of automated large-scale knowledge verification and curation • Lack of KG completion explainability (provenance), comprehensiveness, and interpretability

Profiling and Assess KB Quality

Quality = Fitness for Use

User-defined Multidimensional Concept

Accuracy, Consistency, Freshness, Completeness, Uniqueness

Precision, Timeliness, Conciseness, Interpretability, Accessibility, Objectivity, Security, Relevance, Source Reputation, Understandability, Believability, Ease of use […]

Up to 179 dimensions for Data Quality[1]

only18 applicable to LOD[2] with a dedicated ontology[3]

[1] *Wang, Storey, Firth. A Framework for Analysis of Data Quality Research, IEEE Trans. Knowl. Data Eng., 7(4), p.623-640, 1995*
http://mitiq.mit.edu/documents/publications/TDQMpub/SURVEYIEEEKDEAug95.pdf

[2] *Acosta et al. Detecting Linked Data Quality Issues via Crowdsourcing: A DBpedia Study, Semantic Web, 2016*

[3] Debattista, Lange, Auer - daQ, an Ontology for Dataset Quality Information LDOW2014

Research Context

1. *Designing ML-based solutions for Data and Knowledge engineering is a very hot topic in DB community*
2. *Tsunami of Deep NN architectures and applications*

[ICDE'18 Tutorial]

[SIGMOD Blog, Feb. 2018]

ACM SIGMOD Blog

COURTING ML: WITNESSING THE MARR OF RELATIONAL & WEB DATA SYSTEMS MACHINE LEARNING

Azza Abouzied and Paolo Papotti

FEBRUARY 14, 2018

Big Data , Databases , Machine Learning No Comment

The web is an ever-evolving source of information, with data and knowled from it powering a great range of modern applications. Accompanying wealth of information, web data also introduces numerous challenges due diversity, volatility, inaccuracy, and contradictions. This year's WebDB 20 emphasizes the challenges and opportunities that arise at the intersecti

[SIGMOD'15 Panel]

Data Management in Machine Learning: Challenges, Techniques, and Systems

Arun Kumar, UC San Diego, La Jolla, CA, USA
Matthias Boehm, IBM Research – Almaden, San Jose, CA, USA
Jun Yang, Duke University, Durham, NC, USA

[SIGMOD'17 Tutorial]

Database Meets Deep Learning: Challenges and Opportunities

Wei Wang, Meihui Zhang, Gang Chen, H. V. Jagadish, Beng Chin Ooi, Kian-Lee Tan

[VLDB'17 Keynote]

Deep Learning (m)eats Databases

(shortened)

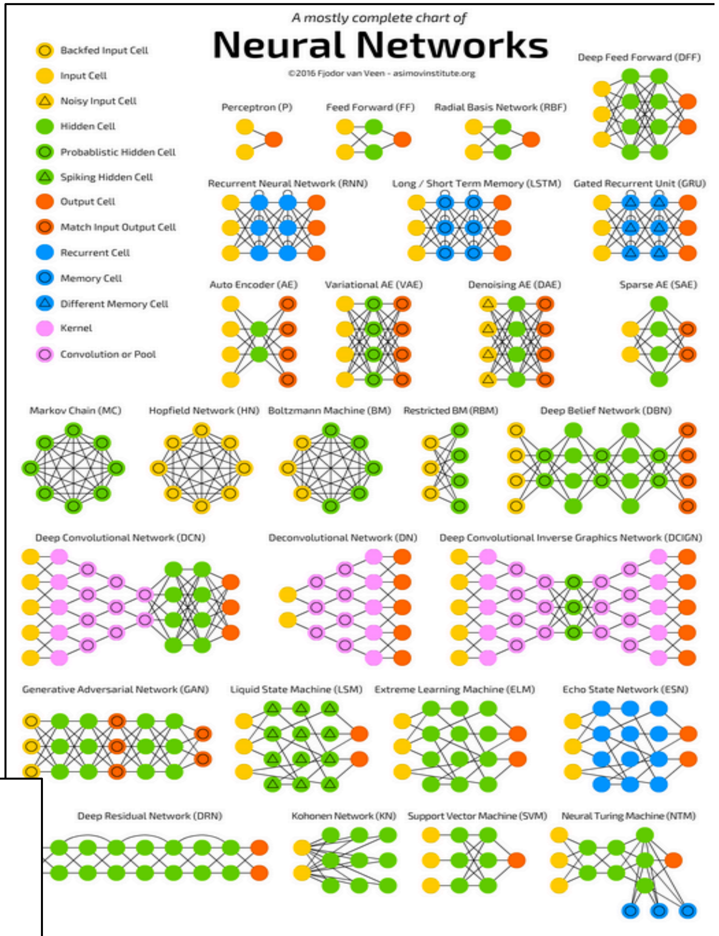
[SIGMOD Record 2016]

Machine Learning and Databases: The Sound of Things to Come or a Cacophony of Hype?

Divy Agrawal, QCRI, dagrawal@qf.org.qa
Magdalena Balazinska, University of Washington, magda@cs.washington.edu
Michael Cafarella, University of Michigan, michjc@michigan.edu
Michael Jordan, UC Berkeley, jordan@cs.berkeley.edu
Tim Kraska, Brown University, tim_kraska@brown.edu
Raghu Ramakrishnan, Microsoft, raghu@microsoft.com
Christopher Ré, Stanford, chrismre@cs.stanford.edu

DEEM
2nd Workshop on Data Management for End-to-End Machine Learning

[workshop@SIGMOD]



Outline

Introduction

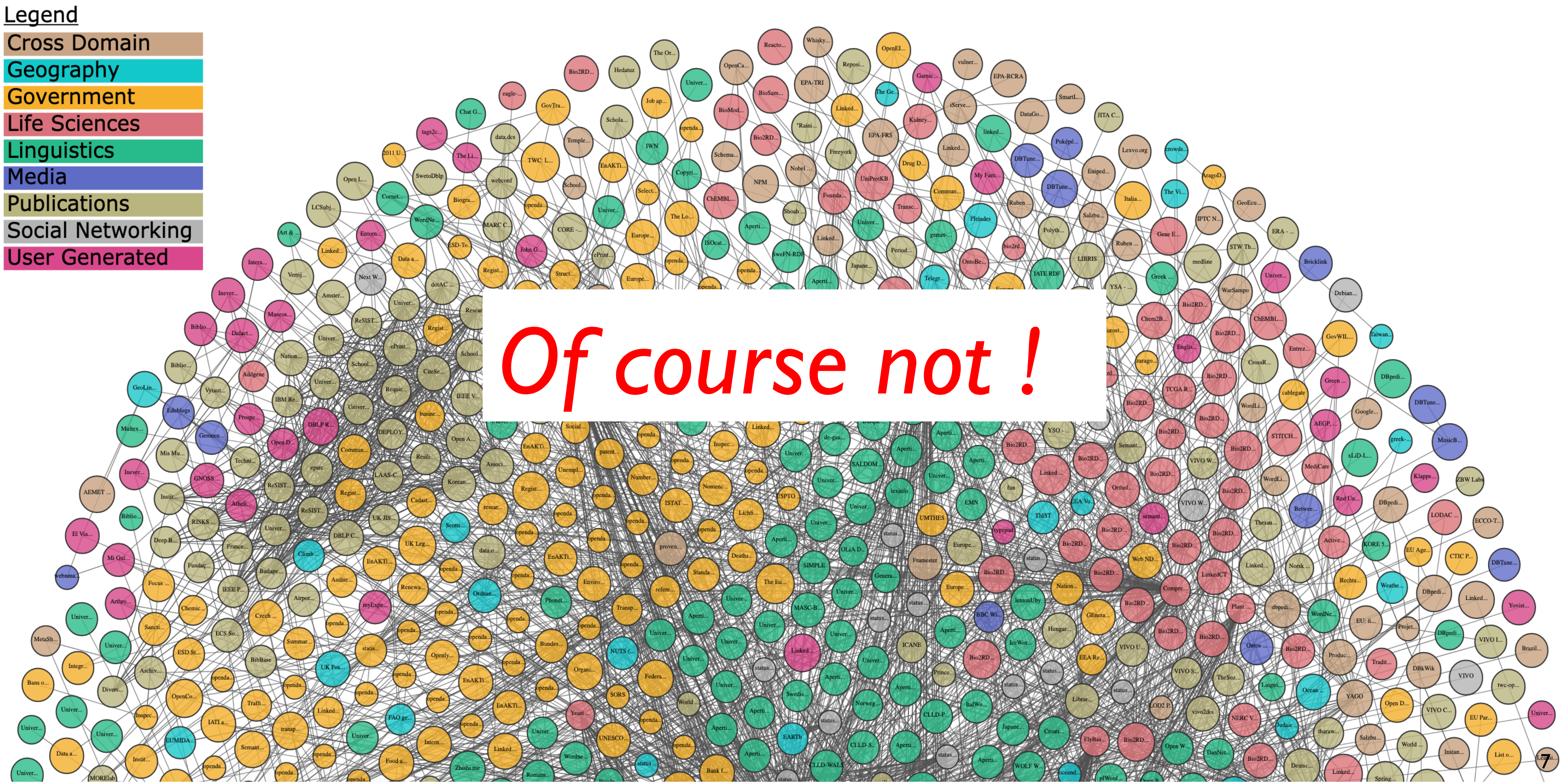
- Motivations
- Context
- Examples illustrating some relevant work

ML-based KG Curation

- KG refinement and ontology learning
- KG embedding
- KG completion
- Consistency checking and KG repairing

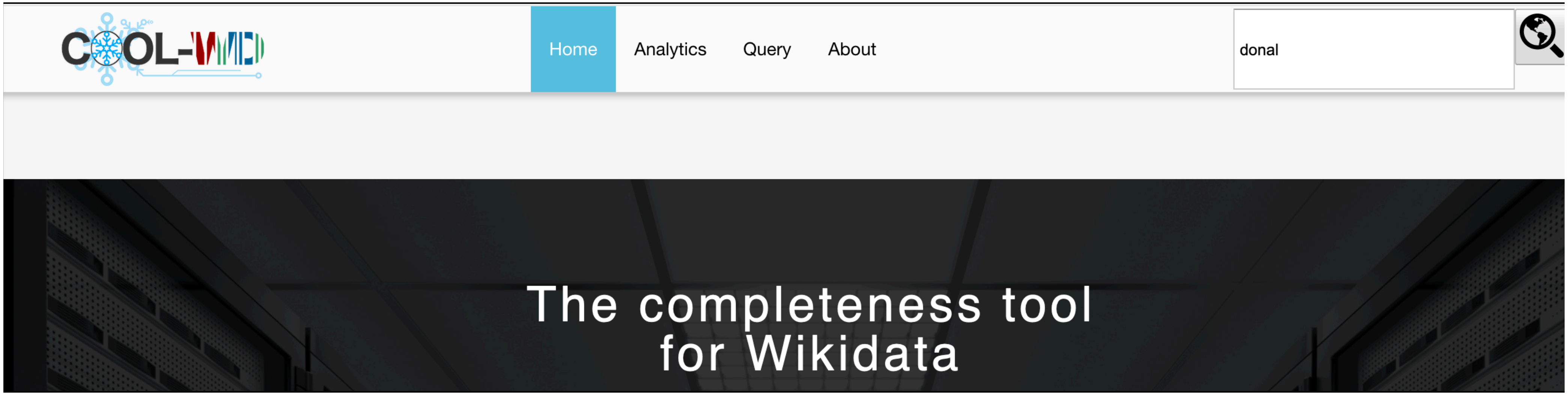
Concluding Remarks & Perspectives

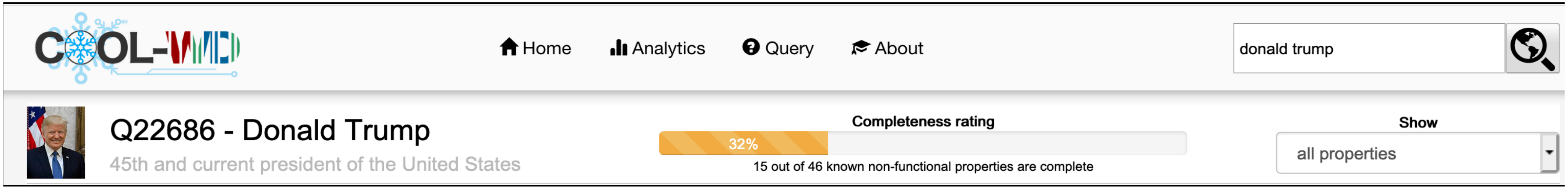
Are all resources and KBs equally complete, accurate, up-to-date, and trustworthy?



Of course not !

Example 1. Completeness



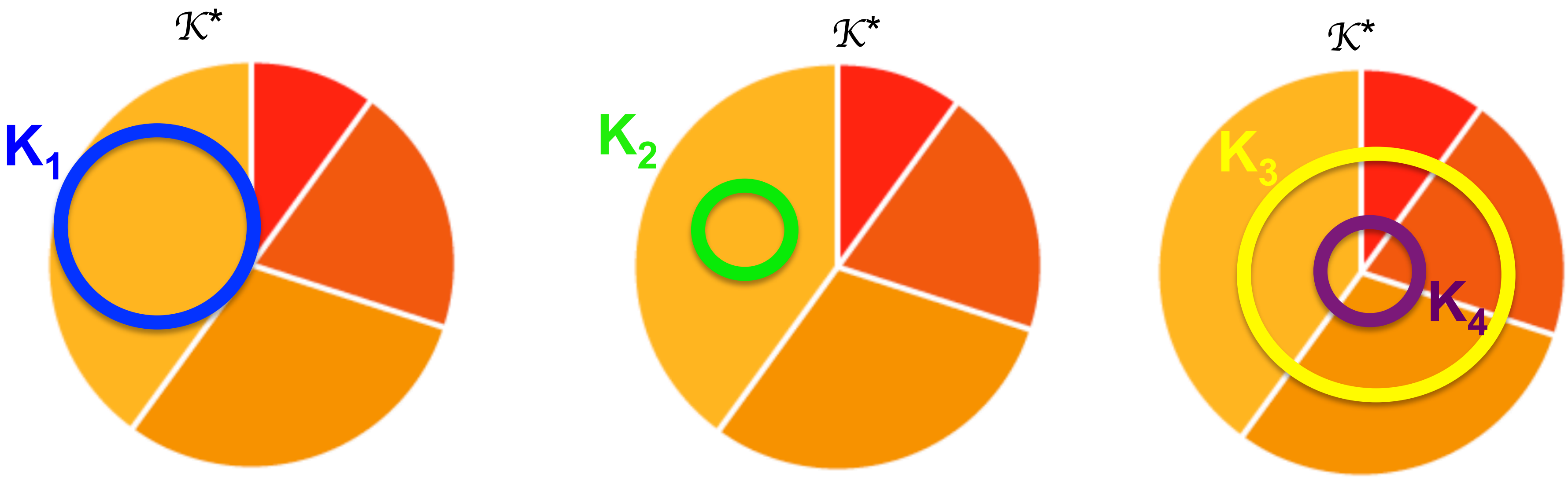


Class name	#Objects	#Properties	Class completeness
Cats	133	2	0.00%
Fictional donkeys and some fact about them	14	3	2.38%
US Presidents	79	5	3.54%
States of Austria	9	2	16.67%
Cantons of Switzerland	26	3	6.41%

F. Darari, R.E. Prasojo, S. Razniewski, W. Nutt. COOL-WD: A Completeness Tool for Wikidata. ISWC'17

Example 1 (Cont'ed). KB Representativeness and Bias

Suppose you have the accurate and complete knowledge of the world-wide populations per city grouped into 4 categories: e.g. (<100k, [100k,500k], [500k,1M], >1M) and 4 KBs.



K_1 is more complete than K_2 but both are somehow biased toward one category

K_1 and K_2 are not as representative as K_3 or K_4

- *Soulet, Giacometti, Markhoff, Suchanek: Representativeness of Knowledge Bases with the Generalized Benford's Law. International Semantic Web Conference (1) 2018: 374-390*
- *Wagner, Garcia, Jadidi, Strohmaier: It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. ICWSM. pp. 454–463 (2015)*
- *Callahan, Herring: Cultural bias in Wikipedia content on famous persons. J. of the Association for Information Science and Technology, 62(10), 1899–1915 (2011)*
- *Pitoura, Tsaparas, Flouris, Fundulaki, Papadakos, Abiteboul, Weikum. On Measuring Bias in Online Information. SIGMOD Record, Vol.46 No.4, December 2017*

Example 2. KB Correctness

Relational data quality problems

Nobel Laureates in Chemistry

Name	Institution	Institution_City	DoB
Skłodowska-Curie Marie	Institut Pasteur	Varsovie	07-11-1867
M. Curie	Pasteur Institute	Paris	1867-11-07
Melvin Calvin	UC Berkeley	Berkeley	1911-04-08
Marie Curien	Paris	Pasteur Institute	2007-11-07
Avram Hershko	NULL	Haifa	NULL
Ronald Hoffman		US	00000000

Misfielded Value

Representation

Duplicates

Typos

Inconsistencies

Incorrect Value

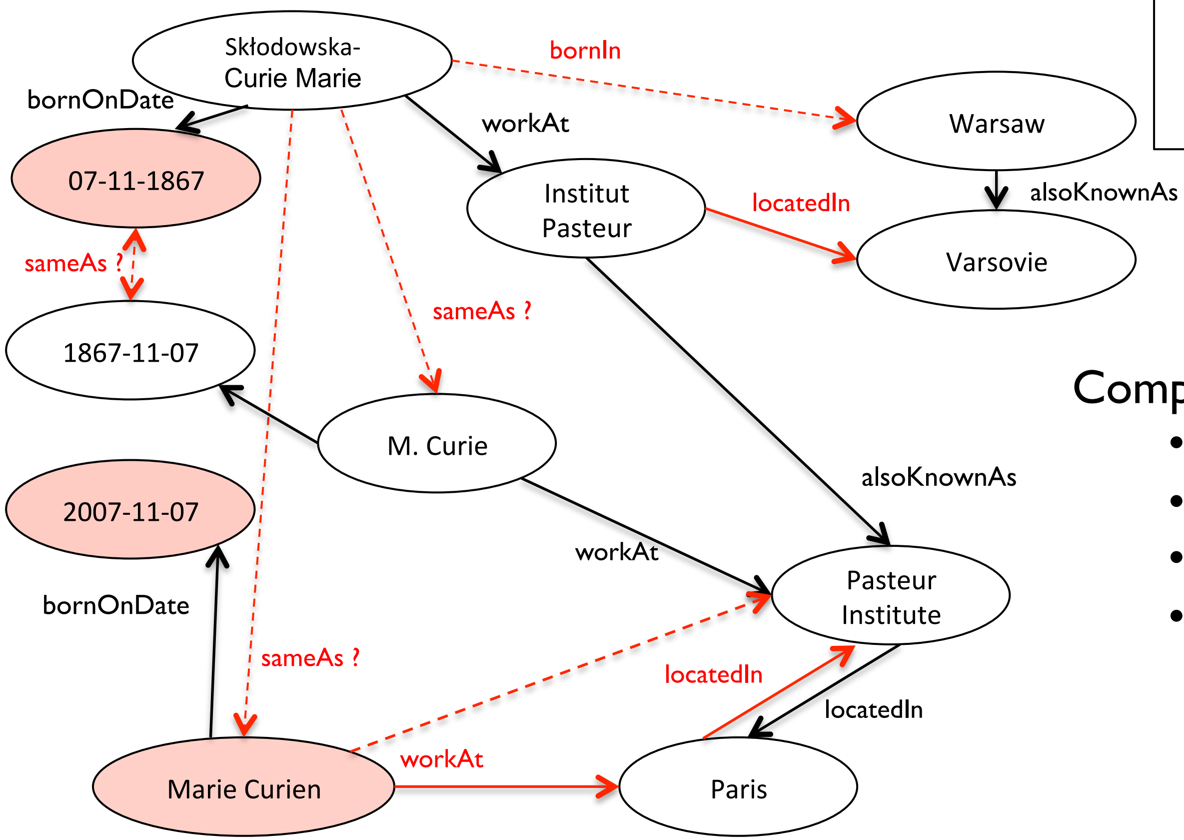
Incorrect Values

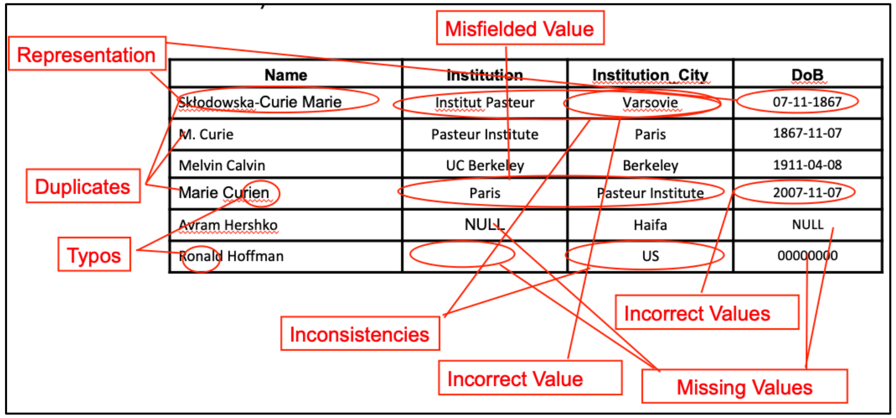
Missing Values

Example 2 (Cont'ed). KB Correctness

Knowledge Graph data problems

Nobel Laureates in Chemistry: Excerpt



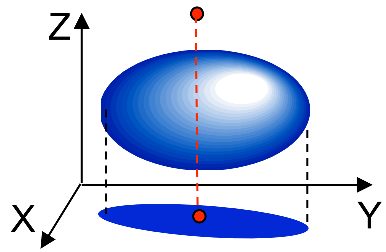


Name	Institution	Institution_City	DoB
Skłodowska-Curie Marie	Institut Pasteur	Varsovie	07-11-1867
M. Curie	Pasteur Institute	Paris	1867-11-07
Melvin Calvin	UC Berkeley	Berkeley	1911-04-08
Marie Curien	Paris	Pasteur Institute	2007-11-07
Avram Hershko	NULL	Haifa	NULL
Ronald Hoffman		US	00000000

Complex combination of:

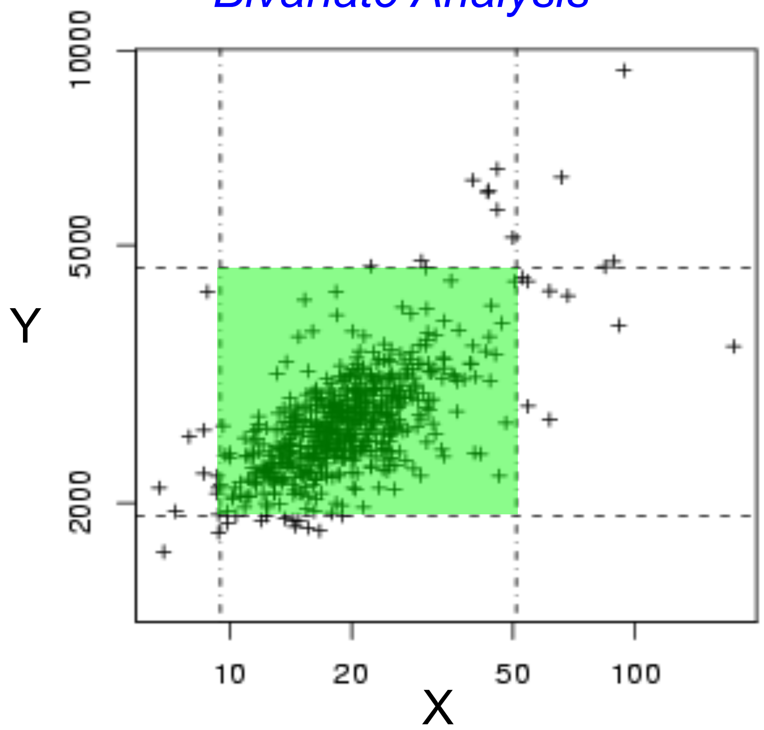
- Missing links and entities
- Spurious links : existence, type, direction
- Erroneous entity name
- Errors in literal values with various degrees of severity: formatting, up-to-dateness, veracity issues

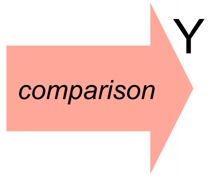
Example 3. Numerical Outliers



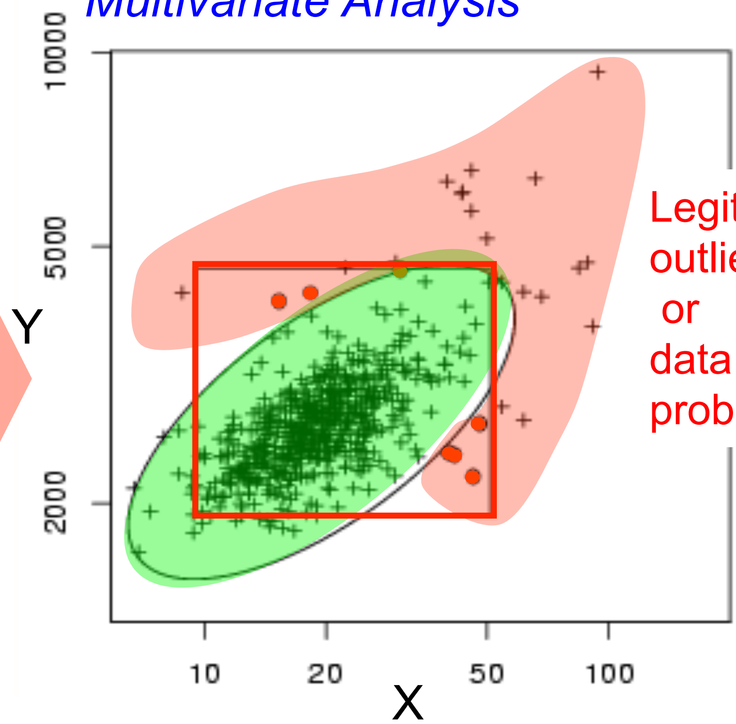
(Classical Setting)

Bivariate Analysis





Multivariate Analysis



Legitimate outliers or data quality problems?

Rejection area: Data space excluding the area defined between 2% and 98% quantiles for X and Y

Rejection area based on:
Mahalanobis_dist(cov(X,Y)) > $\chi^2(.98,2)$

Example 3 (Cont'ed). Numerical Outliers in KG

Need for more approaches leveraging ontology, constraints or dependencies to improve outlier detection

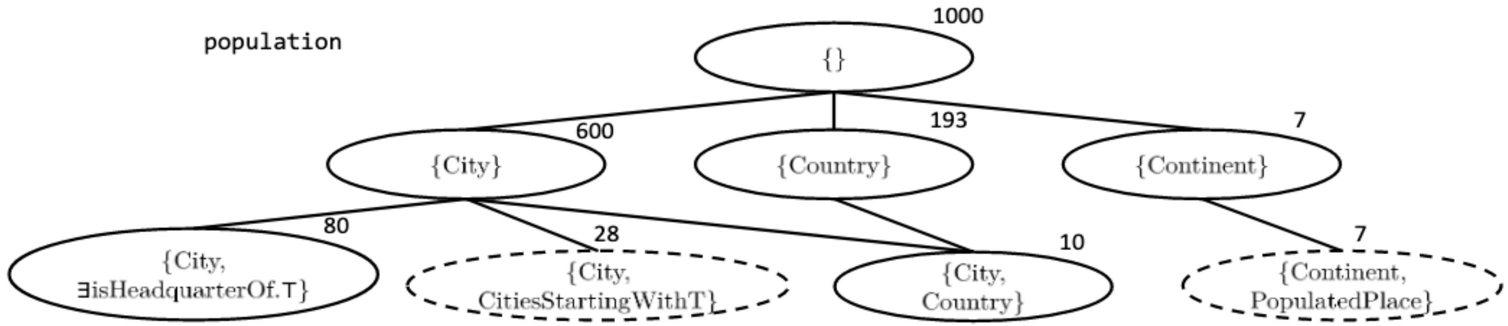


Fig. 1: Example for subpopulation lattice for property **population**. Numbers to the upper right of a node give the number of instances fulfilling the constraint set. Dashed nodes would be pruned, the left one for too low KL divergence, the right one for not reducing the instance set further.

Table 2: Area under the curve determined for the given samples and approaches

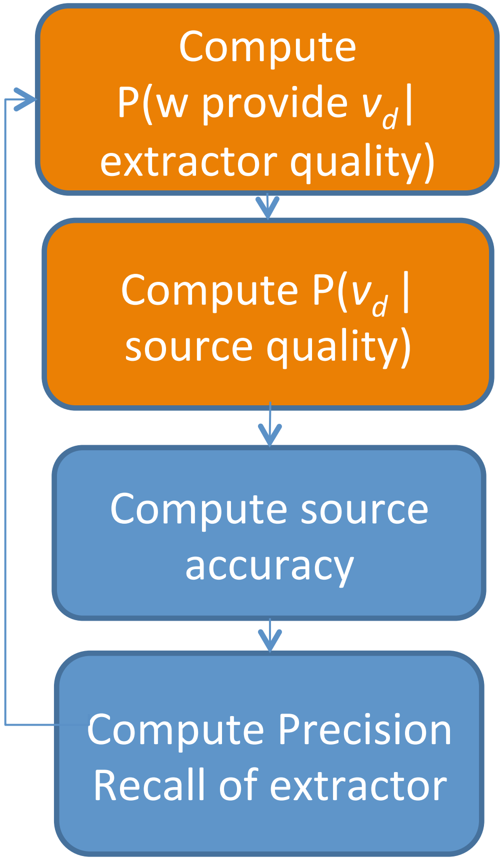
Approach	elevation	height	populationTotal
Outlier Detection	**0.872**	0.888	0.876
Cross-Checked Outlier Detection	0.861	**0.891**	**0.941**
Baseline	0.745	0.847	0.847
Multi-lingual Baseline	0.669	0.509	0.860

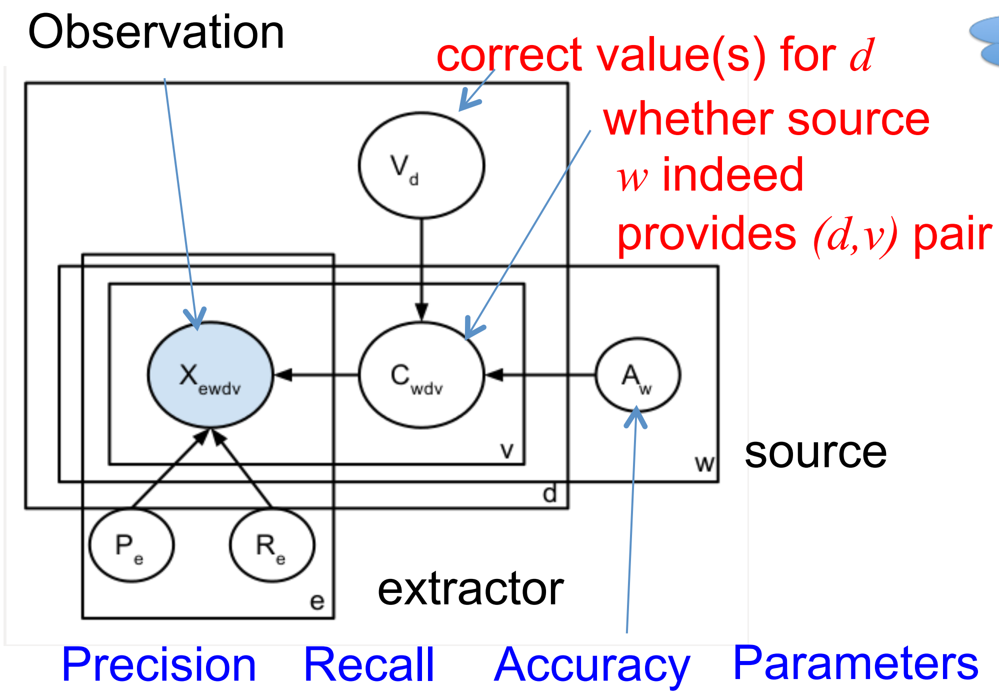
Fleischhacker, Paulheim, Bryl, Völker, and Bizer. Detecting Errors in Numerical Linked Data using Cross-Checked Outlier Detection. ISWC 2014
Debattista, Lange, Auer. A Preliminary Investigation Towards Improving Linked Data Quality Using Distance-based Outlier Detection, The Semantic Web, 2016.

Example 4. Veracity and Trustworthiness

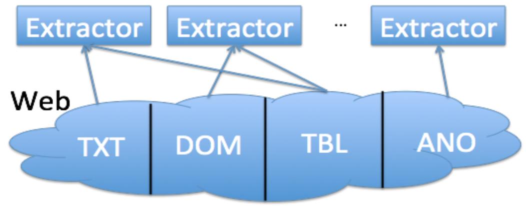
ML-based approach for knowledge-based trust:

- Multi-Layer Model based on EM and Bayesian inference
- Distinguish extractor errors from source errors



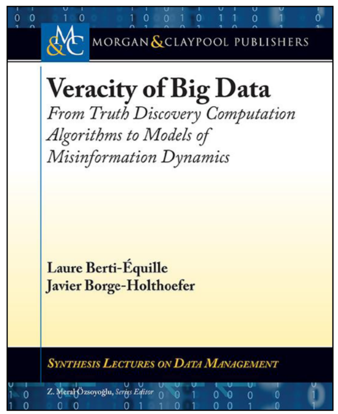


KNOWLEDGE VAULT



#Triples	3.0B (0.3B w. pr>=0.7)
#URLs	2.5B (28M Websites)
#Extractors	16

As of 2014



X. L. Dong, K. Murphy, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, W. Zhang. Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. VLDB 2015

Example 5: Up-to-dateness
Asynchronous Real World and KG evolution

Table 1. DBpedia - Classes and Properties

Version	OWL Class				RDF Property				Object Prop.			Datatype Prop.		
	#	Δ	(-)	(+)	#	Δ	(-)	(+)	#	(-)	(+)	#	(-)	(+)
3.2/3	174				720				384			336		
3.4	204	30	-2	32	2168	1448	-271	1719	1144	-139	899	1024	-132	820
3.5	255	51	-6	57	1274	-894	-1198	304	601	-673	130	673	-525	174
3.6	272	17	0	17	1335	61	-37	98	629	-26	54	706	-11	44
3.7	319	47	-1	48	1643	308	-17	325	750	-6	127	893	-11	198
3.8	359	40	-1	41	1775	132	-3	135	800	-1	51	975	-2	84
3.9	529	170	-1	171	2333	558	-8	566	927	-6	133	1406	-2	433
2014	683	154	-5	159	2795	462	-46	508	1079	-9	161	1716	-37	347
2015-04	735	52	-5	57	2819	24	-103	127	1098	-23	42	1721	-80	85
2015-10	739	4	-5	9	2833	14	-9	23	1099	-3	4	1734	-6	19
2016-04	754	15	0	15	2849	16	-2	18	1103	-1	5	1746	-1	13

Today's DBpedia Ontology: 685 classes described by 2,795 properties

Mihindukulasooriya, Poveda-Villalon, Garcia-Castro, Gomez-Perez. Collaborative Ontology Evolution and Data Quality -An Empirical Analysis, in OWL: Experiences and Directions – Reasoner Evaluation, Springer International Publishing, Cham, 2017, pp. 95–114.
https://www.w3.org/community/owled/files/2016/11/OWLED-ORE-2016_paper_9.pdf

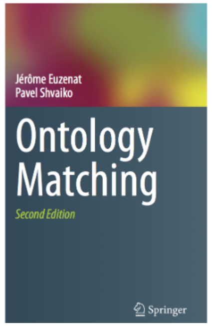
Outline

Introduction

- Motivations
- Context
- Examples illustrating some relevant work

ML-based KG Data Curation

ML-based Solutions for KG Curation



Knowledge Graph Refinement

Ontology Learning to learn a concept level description of a domain (e.g., Cities are Places)

Knowledge Extraction

Fact Extraction and Verification : Knowledge Fusion Methods

Completion of Knowledge Graphs

Knowledge Verification for Long-Tail Verticals

Furong Li[†] Xin Luna Dong[‡] Anno Langen[§] Yang Li[§]
[†]National University of Singapore [‡]Amazon [§]Google Inc.
furongli@comp.nus.edu.sg lunadong@amazon.com {arl, ngli}@google.com

- Learning Embeddings
- Methods for Entity Linking & Link Prediction : classification, rank, probabilistic graph models, deep (reinforcement) learning

Error Detection and Repair in Knowledge Graphs

- Rule learning for detecting/correcting erroneous type assertions, relations or literal values
- User-guided repair with updates

GLUE: Learning to find similar ontological concepts

- Glue applies ML technique to find, for each concept node in a taxonomy, the most similar concept in the other taxonomy
- It applies the multi-learning approach of LSD *(Learning Source Description)*

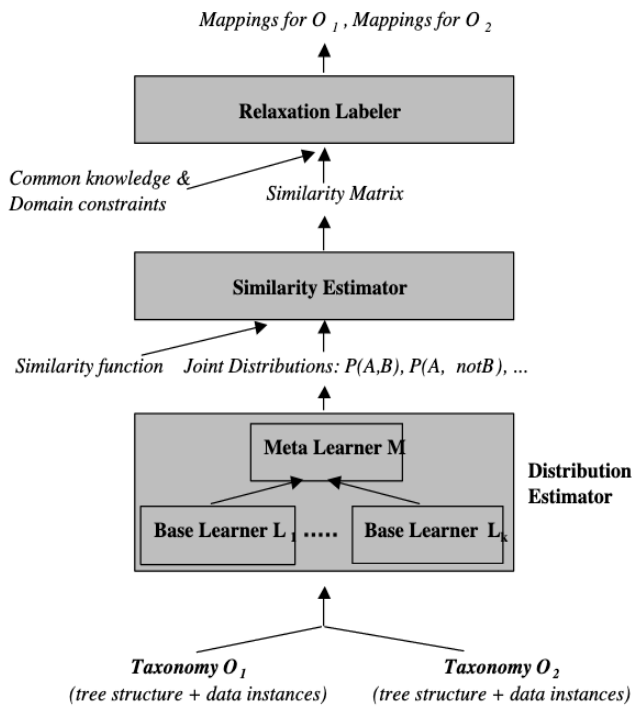


Fig. 2. The GLUE Architecture

Doan, Madhavan and Halevy. Ontology Matching: A Machine Learning Approach. Handbook on Ontologies in Information Systems (pp. 385-403), 2004

GLUE: Learning to find similar ontological concepts (2)

- It leverages the joint probability distribution:
  - P(A,B), P(A, not(B)),P(not(A),B), P(not(A),not(B))
- ML is used to infer whether P(A,B) can be approximated with P(A intersect B)
  - By defining a classifier for instances containing concept A (resp. B) and using it to classify instances of B (resp. A)

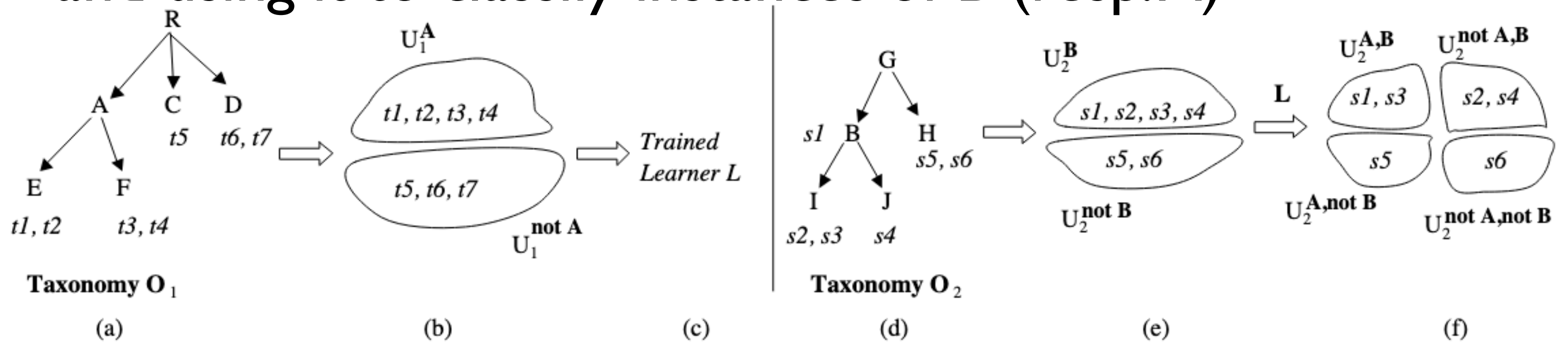


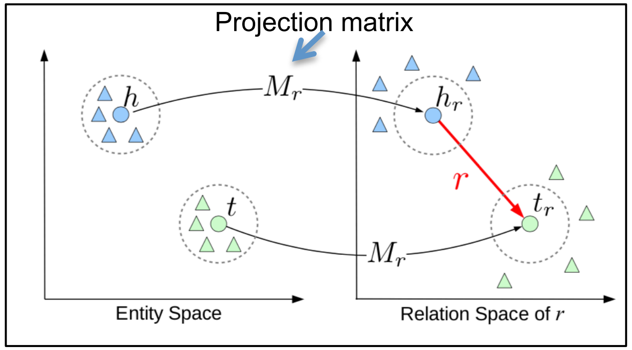
Fig. 3. Estimating the joint distribution of concepts A and B

Doan, Madhavan and Halevy. Ontology Matching: A Machine Learning Approach. Handbook on Ontologies in Information Systems (pp. 385-403), 2004

Learning distributed representations of entities and relations of KG

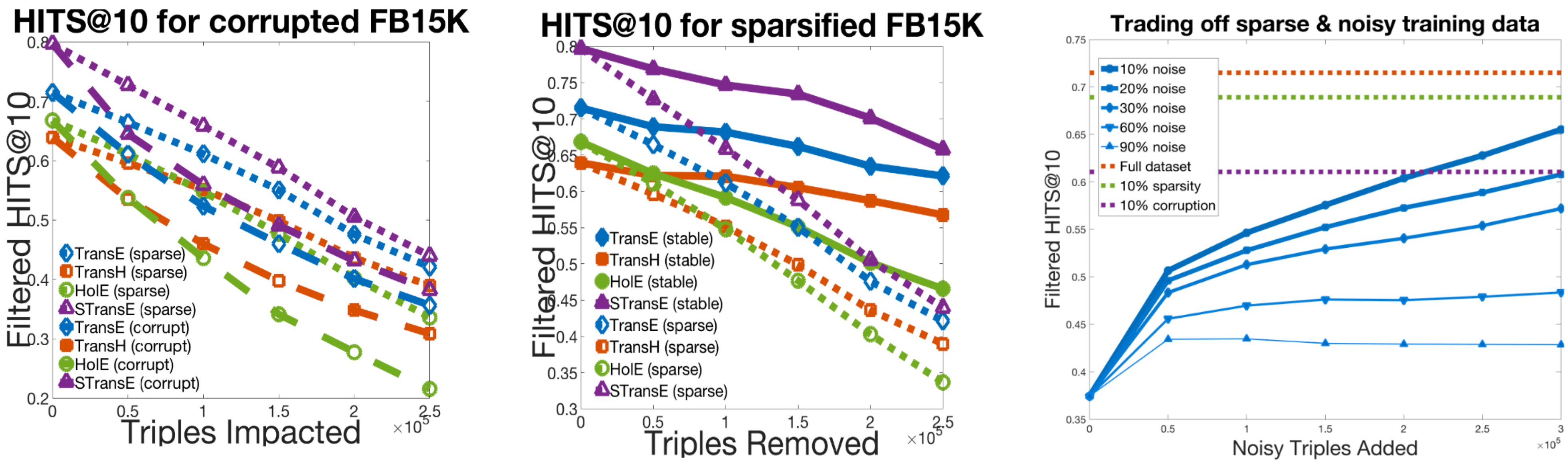
- Linear models
  - Translation-based : TransE, TransH, TransR, STransE, FTransE
  - Tensor product-based: RESCAL, DistMult, ComplEx, SimplE, TuckER
- Deep Learning or convolution
  - HypER, ConvE, ConKB, SLM, LFM, ER-MLP NTN

TransR



Model	Scoring Function	Relation Parameters	Space Complexity
RESCAL (Nickel et al., 2011)	$\mathbf{e}_s^\top \mathbf{W}_r \mathbf{e}_o$	$\mathbf{W}_r \in \mathbb{R}^{{d_e}^2}$	$\mathcal{O}(n_e d_e + n_r d_r^2)$
DistMult (Yang et al., 2015)	$\langle \mathbf{e}_s, \mathbf{w}_r, \mathbf{e}_o \rangle$	$\mathbf{w}_r \in \mathbb{R}^{d_e}$	$\mathcal{O}(n_e d_e + n_r d_e)$
ComplEx (Trouillon et al., 2016)	$\mathrm{Re}(\langle \mathbf{e}_s, \mathbf{w}_r, \bar{\mathbf{e}}_o \rangle)$	$\mathbf{w}_r \in \mathbb{C}^{d_e}$	$\mathcal{O}(n_e d_e + n_r d_e)$
ConvE (Dettmers et al., 2018)	$f(\mathrm{vec}(f([\underline{\mathbf{e}_s}; \underline{\mathbf{w}_r}] * w))\mathbf{W})\mathbf{e}_o$	$\mathbf{w}_r \in \mathbb{R}^{d_r}$	$\mathcal{O}(n_e d_e + n_r d_r)$
HypER (Balažević et al., 2018)	$f(\mathrm{vec}(\mathbf{e}_s * \mathrm{vec}^{-1}(\mathbf{w}_r \mathbf{H}))\mathbf{W})\mathbf{e}_o$	$\mathbf{w}_r \in \mathbb{R}^{d_r}$	$\mathcal{O}(n_e d_e + n_r d_r)$
SimplE (Kazemi & Poole, 2018)	$\frac{1}{2}(\langle \mathbf{h}_{e_s}, \mathbf{w}_r, \mathbf{t}_{e_o} \rangle + \langle \mathbf{h}_{e_o}, \mathbf{w}_{r^{-1}}, \mathbf{t}_{e_s} \rangle)$	$\mathbf{w}_r \in \mathbb{R}^{d_e}$	$\mathcal{O}(n_e d_e + n_r d_e)$
TuckER	$\mathcal{W} \times_1 \mathbf{e}_s \times_2 \mathbf{w}_r \times_3 \mathbf{e}_o$	$\mathbf{w}_r \in \mathbb{R}^{d_r}$	$\mathcal{O}(n_e d_e + n_r d_r)$

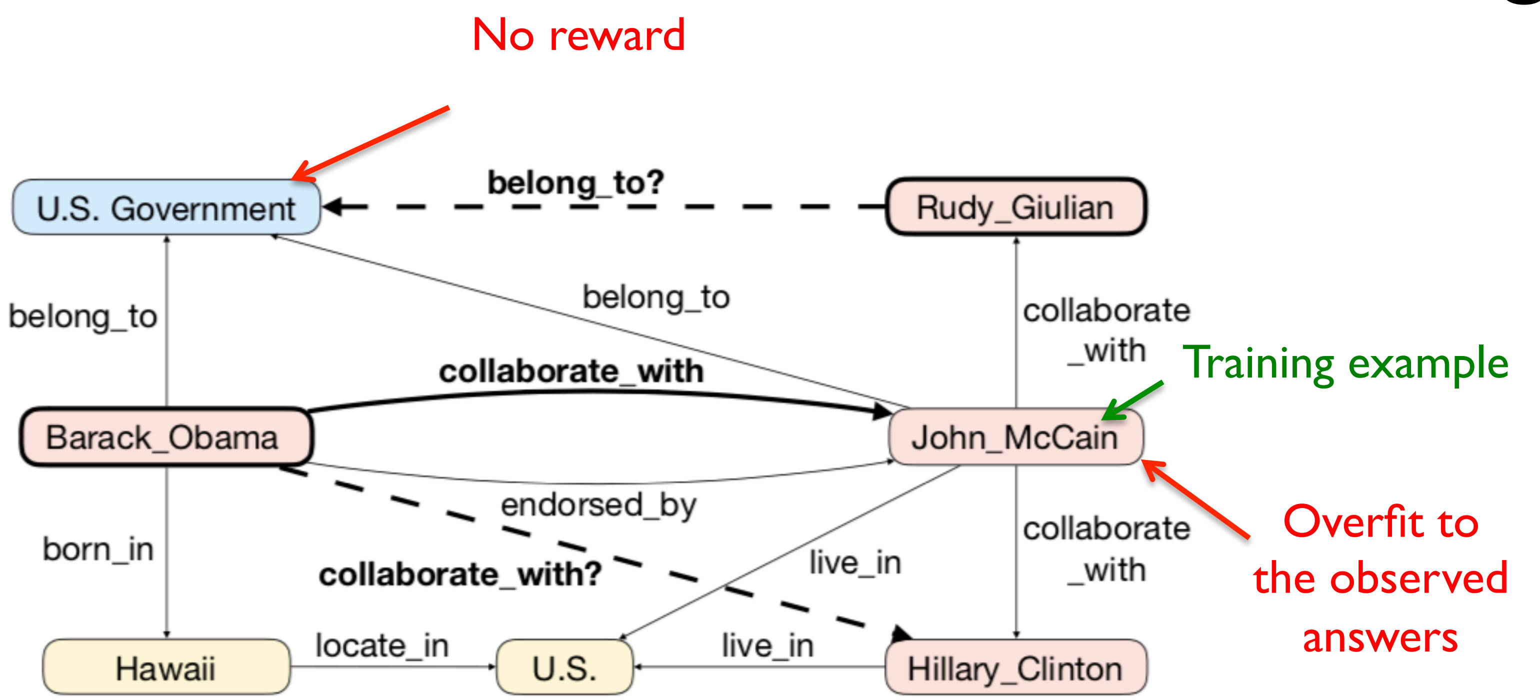
Impact of Noise and Sparsity in KG embeddings



A large, unreliable training dataset may be better than an extremely sparse, high-quality one.

Pujara, Augustine, Getoor. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. ACL 2017
https://www.github.com/linqs/pujara-emnlp17

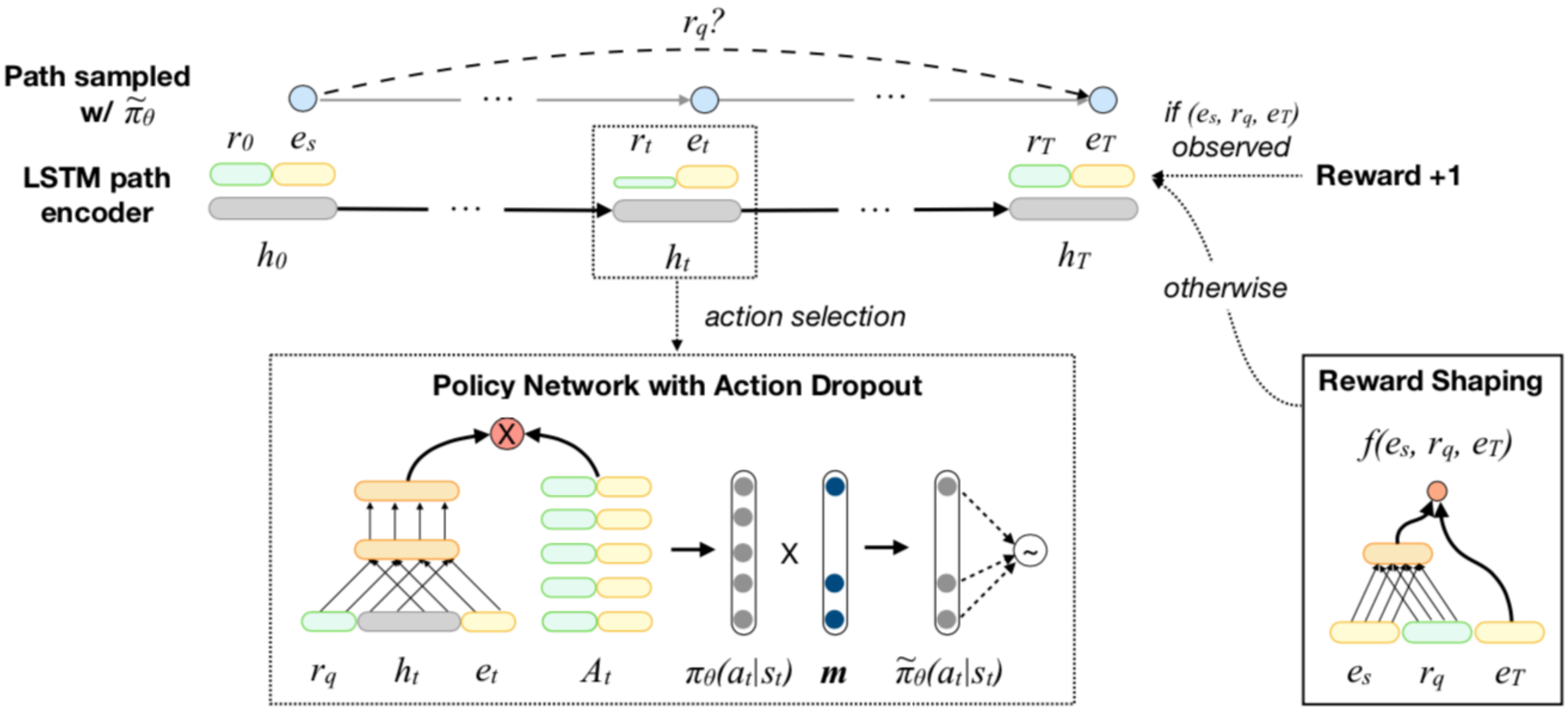
Link Prediction with Reinforcement Learning



- Leverage multi-hop KG query answering
- Use pre-trained model-based on-policy reinforcement learning
- New reward shaping and policy network with action dropout

Shaping Xi Victoria Lin, Socher, Caiming Xiong. Multi-Hop Knowledge Graph Reasoning with Reward. EMNLP 2018

Link Prediction with Reinforcement Learning



- Leverage multi-hop KG query answering
- Use pre-trained model-based on-policy reinforcement learning
- New reward shaping and policy network with action dropout

Shaping Xi Victoria Lin, Socher, Caiming Xiong. Multi-Hop Knowledge Graph Reasoning with Reward. EMNLP 2018

Fang et al., Joint Entity linking with Deep RL On Wednesday

On Wednesday

Joint Entity Linking with Deep Reinforcement Learning

WWW 2019, May 13-17, 2019, San Francisco, CA, USA Zheng Fang, Yanan Cao, Dongjie Zhang, Qian Li, Zhenyu Zhang, and Yanbing Liu

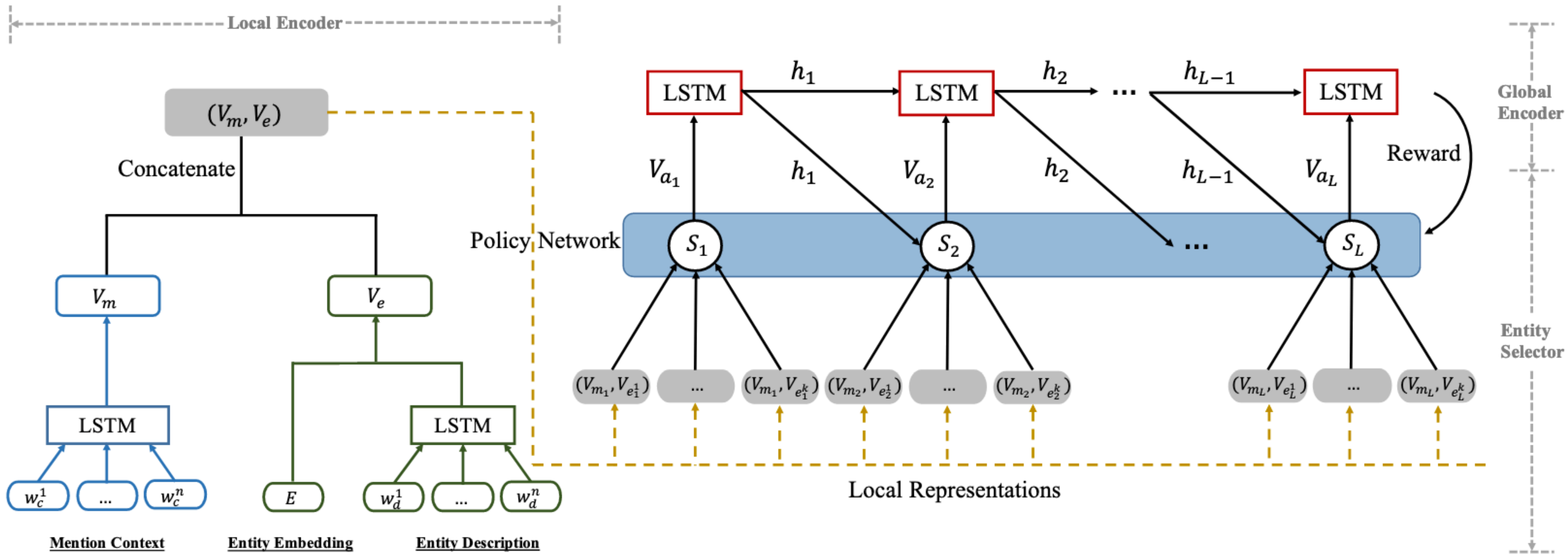
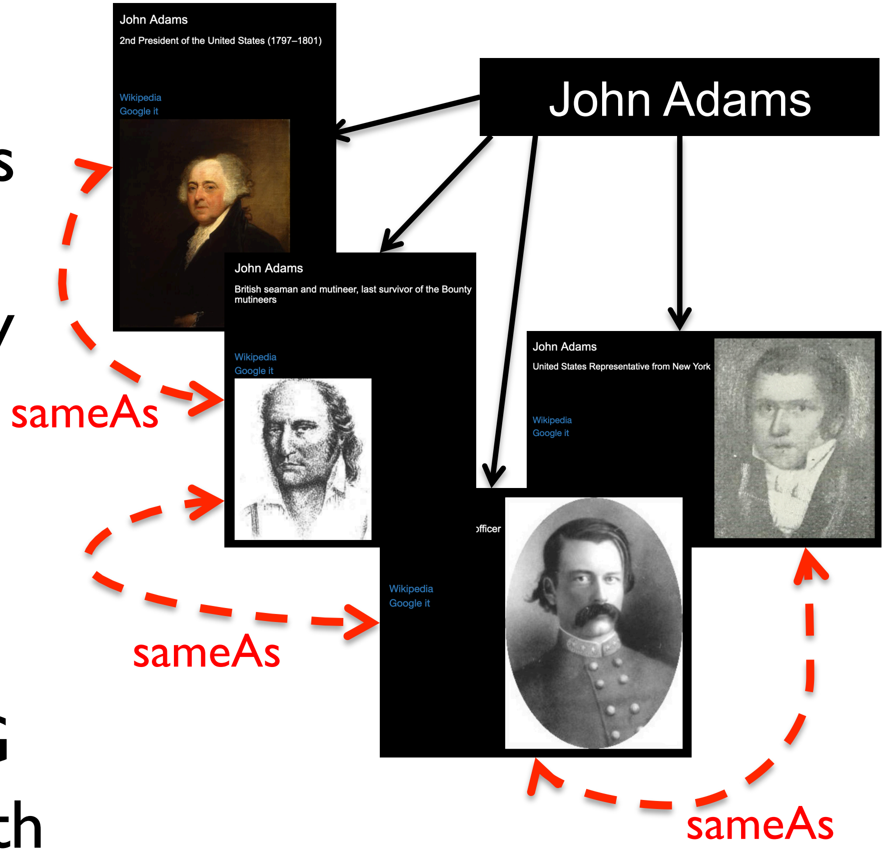


Figure 2: The overall structure of our RLEL model. It contains three parts: Local Encoder, Global Encoder and Entity Selector. In this framework, $(V_{m_t}, V_{e_t^k})$ denotes the concatenation of the mention context vector V_{m_t} and one candidate entity vector $V_{e_t^k}$. The policy network selects one entity from the candidate set, and V_{a_t} denotes the concatenation of the mention context vector V_{m_t} and the selected entity vector $V_{e_t^*}$. h_t represents the hidden status of V_{a_t}, and it will be input into S_{t+1}.

Identity Problem or Link Quality Problem ?

To assessing link quality:

- Network topology and link properties
- Link type, content, and context
- Ontology axioms and ontology quality
- Provenance: source and extractor reliability
- Accessibility, reachability
- Information gain
- Task-dependent properties: e.g., in KG completion: path predicting power, path diversity (to avoid overfitting due to spurious paths)



Error Detection and Repairing

- **Error detection**
  Probabilistic techniques [Ruckhaus et al. 2014, Debattista et al., 2015, Li et al. 2015]
- **Value imputation**
  Statistics: SDType [Paulheim, Bizer, 2014],
- **Pattern enforcement**
  - Syntactic patterns (date formatting)
  - Semantic patterns (name/address)
- **Consistency checks and value update** to satisfy
  - A set of rules, constraints, FDs, CFDs, Denial Constraints (DCs), Matching Dependencies (MDs) with minimal number of changes
  - Integrity, Cardinality, Range or String-based constraints using W3C Shape Constraints Language (SHACL) and Shape Expressions Language (ShEX) [Rashid et al. 2019] see http://github.com/AKSW/RDFUnit

Consistency analysis in evolving KB

Hypothesis(H)

H1: Dynamics features from periodic data profiling can help to identify completeness issues.
H2: Learning models can be used to predict correct integrity constraints using the outputs of the data profiling as features.

Learning Algorithm	Minimum Cardinality			Maximum Cardinality			Range		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Random Forest	**0.9890**	**0.9574**	**0.9730**	**0.9842**	**0.9920**	**0.9881**	**0.9457**	**0.9527**	**0.9594**
Least Squares SVM	0.9944	0.9468	0.9700	0.8491	0.9574	0.9000	0.8596	0.9231	0.8902
Multilayer Perceptron	0.9674	0.9468	0.9570	0.8167	0.9601	0.8826	0.8262	0.8657	0.8456
K-Nearest Neighbour	0.9511	0.9309	0.9409	0.8797	0.8750	0.8773	0.8361	0.8425	0.8393
Naive Bayes	0.9401	0.8351	0.8845	0.9065	0.7739	0.8350	0.8953	0.7951	0.8422

Rashida, Rizzo, Torchianoa, Mihindukulasooriyac, Corchoc, Garcia-Castroc. Completeness and Consistency Analysis for Evolving Knowledge Bases. Journal of Web Semantics. Volume 54, January 2019, Pages 48-71.

Rule discovery in KB

AMIE+: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/amie/

RuleN: http://web.informatik.uni-mannheim.de/RuleN/

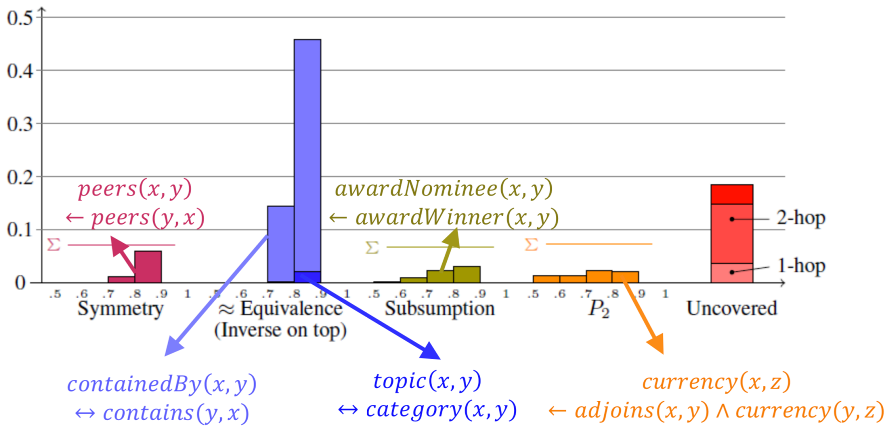
RUDIK: https://github.com/stefano-ortona/rudik

Pellissier, Tanon, Bourgaux Suchanek, Learning how to correct KB from Edit History
On Thursday

[1] *Galarraga, Teflioudi, Hose, Suchanek. Fast rule mining in ontological knowledge bases with AMIE+. The VLDB Journal, 24(6):707–730, 2015*
[2] *Meilicke et al. Fine-Grained Evaluation of Rule- and Embedding-Based Systems for Knowledge Graph Completion. ISWC 2018 (2018): 3–20.*
[3] *Ortona, Meduri, Papotti. Robust discovery of positive and negative rules in knowledge-bases. ICDE 2018.*

Fine-Grained Evaluation: Rule-based vs embedding-based approaches

Test Set Partitioning (FB15k)



	All (100%)		**Sym** (7.2%)		**Eq** (60%)		**Sub** (6.8%)		**P_2** (7.3%)		**UC** (18.4%)	
	h@1	h@10	h@1	h@10	h@1	h@10	h@1	h@10	h@1	h@10	h@1	h@10
AMIE	.647	.858	.906	.983	.766	.961	.720	.950	.451	**.736**	.205	**.486**
RuleN	**.772**	**.870**	**.992**	**1.0**	**.940**	**.982**	**.831**	**.954**	**.536**	.724	**.207**	.480
HolE	.366	.706	.046	.936	.484	.811	.505	.814	.179	.438	.127	.339
RESCAL	.267	.600	.126	.768	.308	.638	.333	.645	.288	.546	.158	.416
TransE	.031	.796	.000	.852	.039	.893	.024	.884	.019	.661	.027	.479

Meilicke et al. Fine-Grained Evaluation of Rule- and Embedding-Based Systems for Knowledge Graph Completion. ISWC 2018 (2018): 3–20.

Concluding Remarks

- ML provides a principled framework and efficient tools for automating and optimizing many KG management tasks (e.g., extraction, population, completion, consistency checking)
- Paradox: ML for KG curation need high quality training data
- Hybrid approaches combining **Humans-in-the-loop**, **AutoML techniques** and **distant supervision** are promising for KG curation

Perspectives for ML-Based KG Curation

- **Integrate the Human "in the Loop of ML-tools"**
  - "Taskify" and minimize the amount of interactions with the users while, at the same time, maximize the potential "ML benefit" for KG management tasks

- **Current efforts:**

  Crowdsourcing, active learning, user-guided repair
  - Detecting LoD Quality issues via Crowdsourcing (DBpedia) [Acosta et al. 2016]
  - Data cleaning with oracle crowds [Bergman et al., SIGMOD'15]
  - User-guided repair of KB [Arioua, Bonifati, EDBT 2018]

- **Direction**:
  - Orchestration of Humans and ML-tools for KG curation

Be inspired !

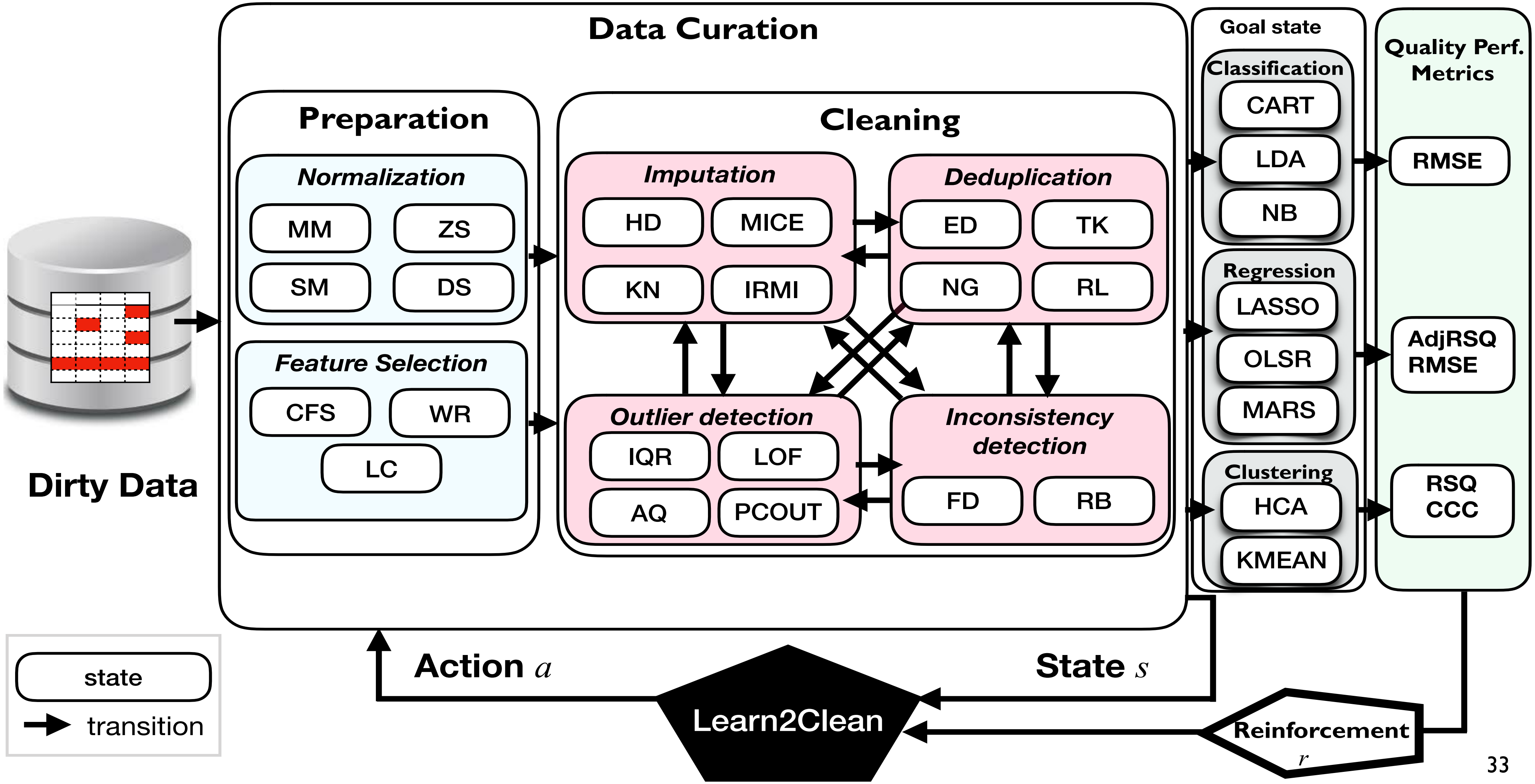
A Condensed View of ML-based curation solutions for structured data

Repair System	ML Approach	Goal
Febrl [Churches et al., 2002]	HMM and MLE	Standardizing loosely structured texts (e.g., name/address) based on the probabilistic model learnt from training data
SCARE [Yakout, Berti-Equille, Elmagarmid, SIGMOD'13]	Multiple ML models used to capture data dependencies across multiple data partitions	Find the candidate repair that maximizes the likelihood repair benefit under a cost threshold of the update
Continuous Cleaning [Volkovs et al., ICDE'14]	Logistic classifiers	Learning from past user repair preferences to recommend next more accurate repairs
Lens [Yang et al. ,VLDB'15]	Various ML models encoded in Domain Constraints	Declarative on-Demand ETL with prioritized curation tasks based on probabilistic query processing and PC-Tables
HoloClean [Rekatsinas et al.,VLDB 2017]	Probabilistic inference on factor graphs with SGD and Gibbs sampling	Mixing statistical and logical rules, DCs, MDs, etc. to infer candidate repairs in a scalable way with domain pruning and constraint relaxation
BoostClean [Krishnan et al., 2017]	Poster #1293 on Wednesday !	Mixing statistical and logical rules, domain constraints for detection and repair combinations to maximize the predictive accuracy over test data
Learn2Clean [Berti-Equille, TheWebConf2019]	Reinforcement Learning	Learn from trial-and-errors the sequence of data preprocessing tasks that maximizes the quality of a given ML model

Reinforcement learning for data cleaning

Learn2Clean: Optimizing the Sequence of Tasks for Data Preparation

[The Web Conference 2019]



Thanks!