KNOWLEDGE GRAPH CONSTRUCTION

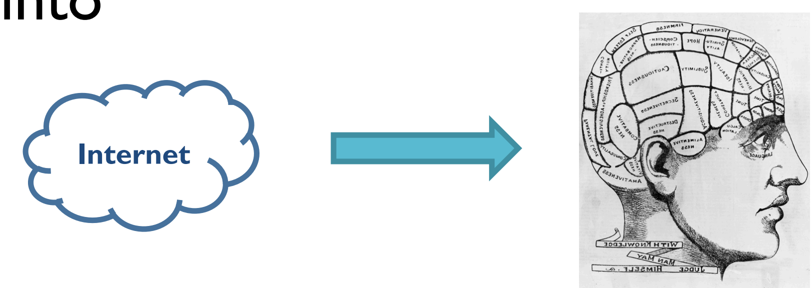
Jay Pujara

CMPS290C

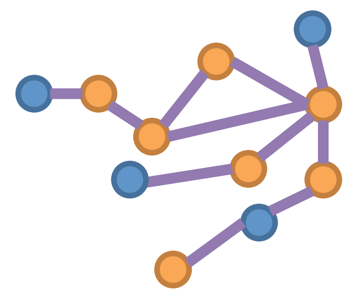
4/8/2014

Talk goals!

- Problem: converting noisy text into useful knowledge



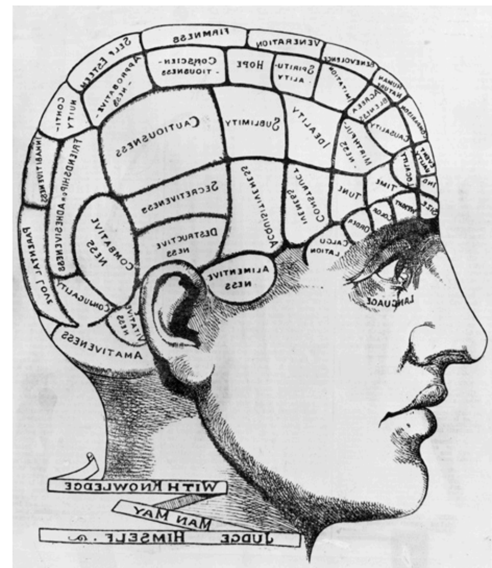
- Topics:
  - Current state-of-the-art in Information Extraction
  - Knowledge Graphs & SRL
  - PSL Models and demo
  - Tools & Datasets



Can Computers Create Knowledge?

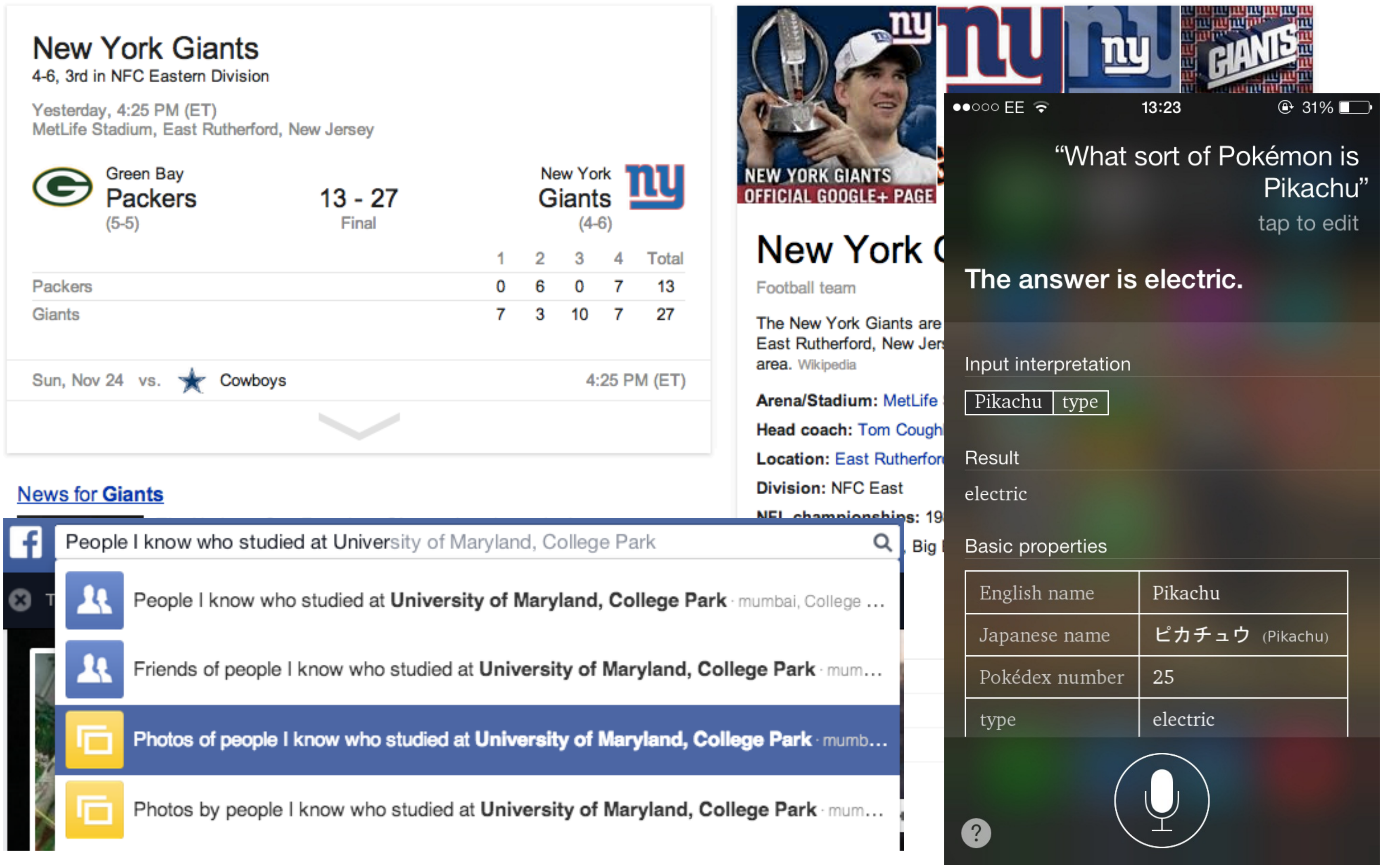
Internet

Massive source of
publicly available
information



Knowledge

Computers + Knowledge =



What does it mean to create knowledge?
What do we mean by knowledge?

Defining the Questions

- Extraction
- Representation
- Reasoning and Inference

Motivating Example

WASHINGTON (AP) — The head of the Internal Revenue Service told House Republicans on Wednesday that it would take years to provide all the documents they have subpoenaed in their probe of how the agency handled tea party groups' applications for tax-exempt status.

The comments by IRS chief John Koskinen drew a frosty response from Republicans who run the House Government Oversight and Reform Committee, one of several congressional panels investigating the controversy. The panel's chairman, Rep. Darrell Issa, R-Calif., warned him he should comply with the request "or potentially be held in contempt" of Congress, a sometimes threatened but seldom-used authority.

A Brief (Yet Helpful) Guide to Information Extraction

Extracting Entities: Named Entity Recognition

WASHINGTON (AP) — The head of the Internal Revenue Service told House Republicans on Wednesday that it would take years to provide all the documents they have subpoenaed in their probe of how the agency handled tea party groups' applications for tax-exempt status.

The comments by IRS chief John Koskinen drew a frosty response from Republicans who run the House Government Oversight and Reform Committee, one of several congressional panels investigating the controversy. The panel's chairman, Rep. Darrell Issa, R-Calif., warned him he should comply with the request "or potentially be held in contempt" of Congress, a sometimes threatened but seldom-used authority.

Extracting Entities: Named Entity Recognition

WASHINGTON (AP) — The head of the Internal Revenue Service told House Republicans on Wednesday that it would take years to provide all the documents they have subpoenaed in their probe of how the agency handled tea party groups' applications for tax-exempt status.

The comments by IRS chief John Koskinen drew a frosty response from Republicans who run the House Government Oversight and Reform Committee, one of several congressional panels investigating the controversy. The panel's chairman, Rep. Darrell Issa, R-Calif., warned him he should comply with the request "or potentially be held in contempt" of Congress, a sometimes threatened but seldom-used authority.

Understanding entities: Entity Resolution

head

Internal Revenue Service

House Republicans

Wednesday

the documents

the agency

tea party groups'

IRS chief

John Koskinen

Republicans

the House Government Oversight and Reform Committee,

congressional panels

the controversy.

The panel

chairman

Rep. Darrell Issa

him

he

the request

Congress

authority.

Understanding entities: Entity Resolution

head
IRS chief
John Koskinen
him
he

House Republicans
they
Republicans
the House Government Oversight and Reform Committee,
The panel

chairman
Rep. Darrell Issa

congressional panels

the controversy

the request

Congress

authority

Wednesday

the documents

the agency
Internal Revenue Service

tea party groups'

Understanding entities: Entity Linking

head of the Internal Revenue Service

IRS chief

John Koskinen

him

he

House Republicans

they

Republicans

the House Government Oversight and Reform Committee,

The panel

chairman

Rep. Darrell Issa

Koskinen

From Wikipedia, the free encyclopedia

Koskinen is a surname originating in Finland (in Finnish, it means "small rapids"), where it is the ninth most common[1] surname. It may also refer to:

- Aarno Yrjö-Koskinen (1885–1951), Finnish politician, ambassador and freiherr
- Harri Koskinen (born 1970), Finnish designer
- Jari Koskinen (born 1960), Finnish politician, Minister for Agriculture and Forestry of Finland
- Johannes Koskinen (born 1954), Finnish politician (M.P., Minister of Justice)
- John Koskinen, 2013 nominee for the position of US IRS commissioner and former president of the U.S. Soccer Foundation (2004–2008)
- Joonas Koskinen, Finnish ice hockey player
- Jukka Koskinen, Finnish musician (bassist for Norther, Wintersun)
- Jukka Koskinen (footballer) (born 1972), Finnish football (soccer) player
- Kalle Koskinen (born 1972), Finnish ice hockey player
- Kerkko Koskinen (born 1973), Finnish musician
- Lennart Koskinen (born 1944), clergyman in the Church of Sweden, serving as Bishop in Visby
- Mikko Koskinen (born 1988), Finnish hockey player for the Sound Tigers in AHL league
- Pasi Koskinen (born 1972), Finnish vocalist (Amorphis)
- Petri Koskinen (born 1983_, Finnish ice hockey player
- Rolf Koskinen (born 1939), Finnish orienteering competitor, European champion
- Sampo Koskinen (born 1979), Finnish football (soccer) player
- Sauli Koskinen (born 1985), a Finnish TV/radio personality and entertainment reporter
- Tapio Koskinen (born 1953), Finnish ice hockey player
- Yrjö Sakari Yrjö-Koskinen (1830–1903), Finnish politician (senator, Finnish Party), professor, historian

Darrell Issa

From Wikipedia, the free encyclopedia

Darrell Edward Issa (/ˈaɪsə/; born November 1, 1953) is the Republican U.S. Representative for California's 49th congressional district, serving since 2001. The district,numbered as the 48th District during his first term, covers the northern coastal areas of San Diego County, including cities such as Oceanside, Vista, Carlsbad and Encinitas, as well as a small portion of southern Orange County.[4]

He was formerly a CEO of Directed Electronics, a Vista, California-based manufacturer of automobile security and convenience products. The district was numbered as the 48th District during his first term and was renumbered the 49th after the 2000 Census. Since January 2011, he has served as Chairman of the House Oversight and Government Reform Committee.

As of 2013, Issa is a multi-millionaire with a net worth estimated at as much as $450 million, which, if accurate, makes him the wealthiest currently-serving member of Congress.[5][6][7]

Contents [hide]

Darrell Issa

Understanding entities: Entity Disambiguation

head of the Inte
IRS chief
John Koskinen
him
he

Koskinen

From Wikipedia, the free encyclopedia

Koskinen is a surname originating in Finland (in Finnish, it means "small rapids"), where it is the ninth most common[1] surname. It may also refer to:

- Aarno Yrjö-Koskinen (1885–1951), Finnish politician, ambassador and freiherr
- Harri Koskinen (born 1970), Finnish designer
- Jari Koskinen (born 1960), Finnish politician, Minister for Agriculture and Forestry of Finland
- Johannes Koskinen (born 1954), Finnish politician (M.P., Minister of Justice)
- John Koskinen, 2013 nominee for the position of US IRS commissioner and former president of the U.S. Soccer Foundation (2004–2008)
- Joonas Koskinen, Finnish ice hockey player
- Jukka Koskinen, Finnish musician (bassist for Norther, Wintersun)
- Jukka Koskinen (footballer) (born 1972), Finnish football (soccer) player
- Kalle Koskinen (born 1972), Finnish ice hockey player
- Kerkko Koskinen (born 1973), Finnish musician
- Lennart Koskinen (born 1944), clergyman in the Church of Sweden, serving as Bishop in Visby
- Mikko Koskinen (born 1988), Finnish hockey player for the Sound Tigers in AHL league
- Pasi Koskinen (born 1972), Finnish vocalist (Amorphis)
- Petri Koskinen (born 1983_, Finnish ice hockey player
- Rolf Koskinen (born 1939), Finnish orienteering competitor, European champion
- Sampo Koskinen (born 1979), Finnish football (soccer) player
- Sauli Koskinen (born 1985), a Finnish TV/radio personality and entertainment reporter
- Tapio Koskinen (born 1953), Finnish ice hockey player
- Yrjö Sakari Yrjö-Koskinen (1830–1903), Finnish politician (senator, Finnish Party), professor, historian

Extracting answers from text

WASHINGTON (AP) — The head of the Internal Revenue Service told House Republicans on Wednesday that it would take years to provide all the documents they have subpoenaed in their probe of how the agency handled tea party groups' applications for tax-exempt status.

The comments by IRS chief John Koskinen drew a frosty response from Republicans who run the House Government Oversight and Reform Committee, one of several congressional panels investigating the controversy. The panel's chairman, Rep. Darrell Issa, R-Calif., warned him he should comply with the request "or potentially be held in contempt" of Congress, a sometimes threatened but seldom-used authority.

Who is the head of the IRS?

Which Wednesday?

What is being subpoenaed by whom?

How do the House Republicans relate to Congress?

Who chairs the House Oversight & Reform Committee?

Which state does Darrell Issa represent?

How do the Republicans feel about the IRS chief?

Extracting answers from text: patterns

Leadership Patterns:
_ chief _
IRS chief John Koskinen
_ chairman _
The panel's chairman, Rep. Darrell Issa

Subset Patterns:
_ one of _
the House Government Oversight and Reform Committee, one of several congressional panels

Association Patterns:

_, _
Darrell Issa, R-Calif

Who is the head of the IRS?

Who chairs the House Oversight & Reform Committee?

How do the House Republicans relate to Congress?

Which state does Darrell Issa represent?

Representing knowledge from text

organizationleadbyperson(IRS, John Koskinen)

organizationleadbyperson(House Oversight & Reform Committee, Darrell Issa)

subpartoforganization(House Oversight & Reform Committee, Congress)

politicianmemberofpoliticsgroup(Darrell Issa, Republicans)

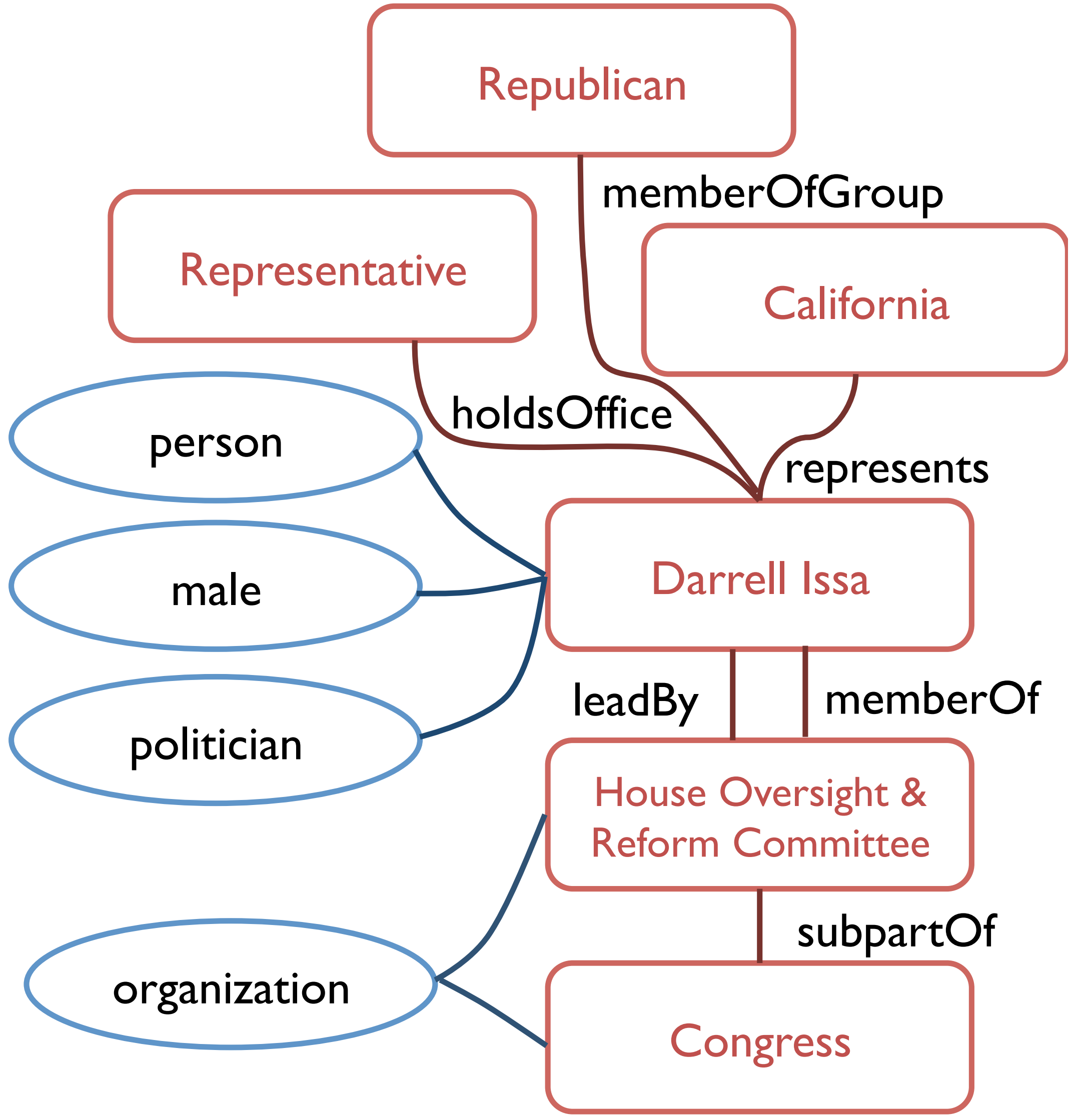
politicianholdsoffice(Darrell Issa, Representative)

locationrepresentedbypolitician(California, Darrell Issa)

Knowledge Graph representation

- Each entity is a node (red squares)
- Each node has attributes (blue circles)
- Edges between nodes represent relationships

This representation emphasizes the *relational structure* of knowlege



Real Systems & IE Resources

NLP Toolkits

The Stanford Natural Language Processing G

home · people · teaching · research · publications · software · events · l

The Stanford NLP Group makes parts of our Natural Language Processing software available to everyone. These are statistical NLP toolkits for various major computational linguistics problems. They can be incorporated into applications with human language technology needs.

All the software we distribute here is written in Java. All recent distributions require Oracle Java 6+ or OpenJDK 7+. Distribution packages include components for command-line invocation, jar files, a Java API, and source code. A number of helpful people have extended our work with bindings or translations for other languages. As a result, much of this software can also easily be used from Python (or Jython), Ruby, Perl, Javascript, and F# or other .NET languages.

Supported software distributions

This code is being developed, and we try to answer questions and fix bugs on a best-effort basis.

All these software distributions are open source, **licensed under the GNU General Public License** (v2 or later). Note that this is the *full* GPL, which allows many free uses, but *does not allow* its incorporation into any type of distributed proprietary software, even in part or in translation. **Commercial licensing** is also available; please contact us if you are interested.

Stanford CoreNLP
: An integrated suite of natural language processing tools for English and (mainland) Chinese in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference. See also: Stanford Deterministic Coreference Resolution, and the online CoreNLP demo, and the CoreNLP FAQ.

http://nlp.stanford.edu/software/

http://www.nltk.org/

http://opennlp.apache.org/

Named-entity recognition

Co-reference resolution

Parsing

Part-of-Speech Tagging

NLTK 3.0 documentation

NEXT | MODULES | INDEX

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. A new version with updates for Python 3 and NLTK 3 is in preparation.

openNLP™

The Apache Software Foundation
http://www.apache.org/

General

- Home
- Download
- Maven Dependency
- License
- Documentation
- News
- Mailing Lists
- Issue tracker
- Wiki

Welcome to Apache OpenNLP

The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.

It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. These tasks are usually required to build more advanced text processing services. OpenNLP also includes maximum entropy and perceptron based machine learning.

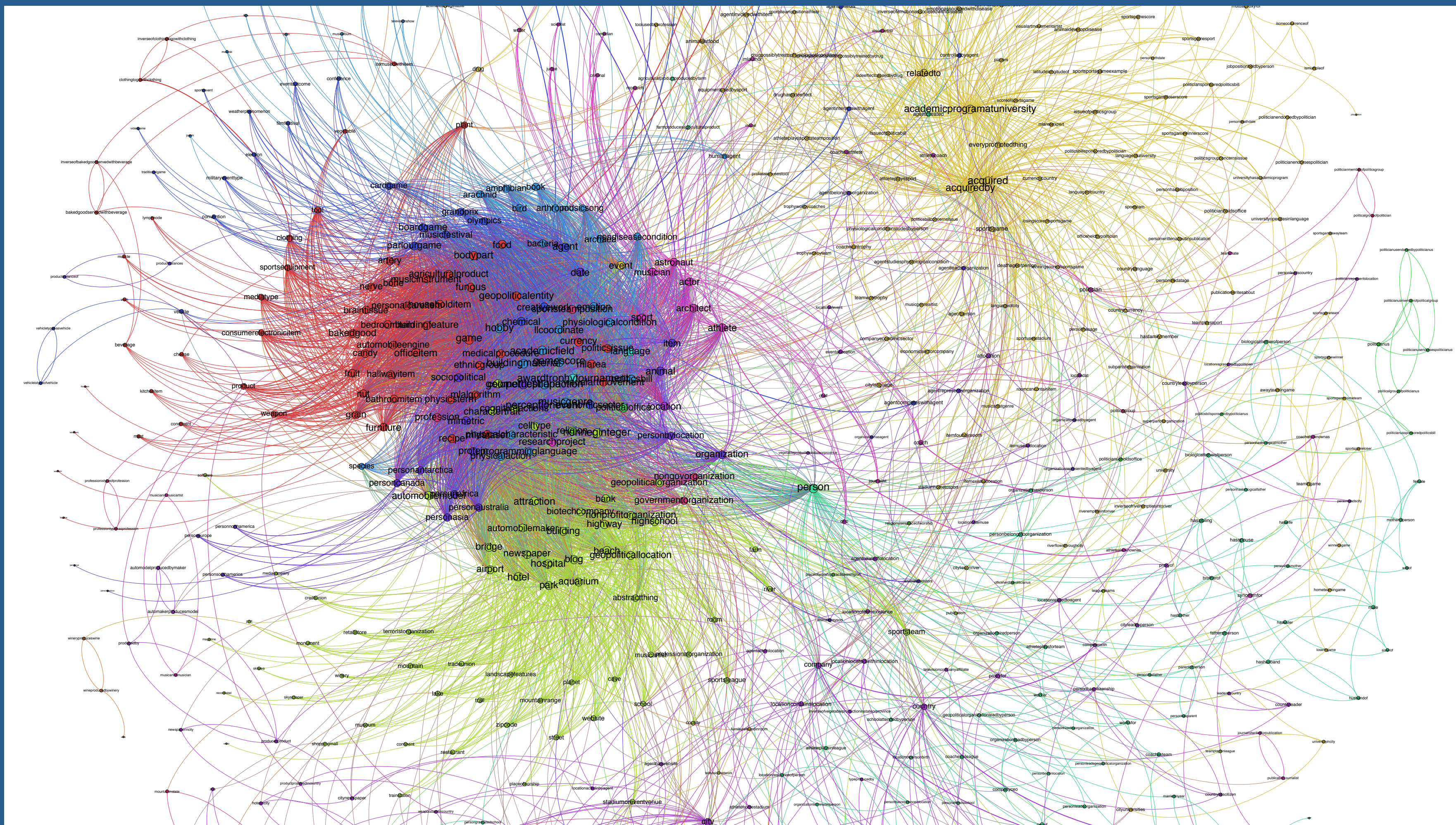
Information Extraction Systems (& KBs)

YAGO [120M]:
Extracts primarily from structured text (Wikipedia infoboxes), with a restrictive set of relations (100) and WordNet categories
http://www.mpi-inf.mpg.de/yago-naga/yago/

NELL [50M]:
Extracts from unstructured webpages (ClueWeb) with a broad set of predefined relations and categories (1000s) http://rtw.ml.cmu.edu/rtw/

OLLIE/KnowItAll [15M/5B]:
OpenIE - uses unstructured webpages (ClueWeb) with no predefined relations or categories
http://openie.cs.washington.edu/

Problem Solved?



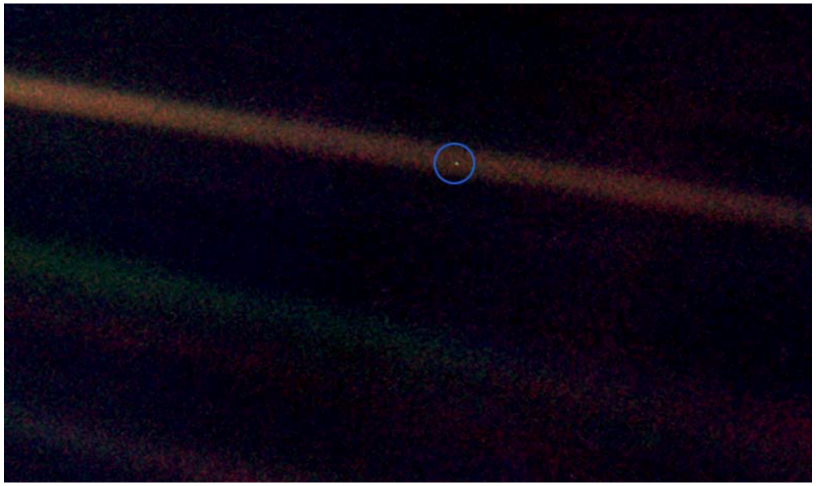
Each document is a "world" of information

- Many approaches are successful at resolving entities, and discovering relationships at the scope of a document

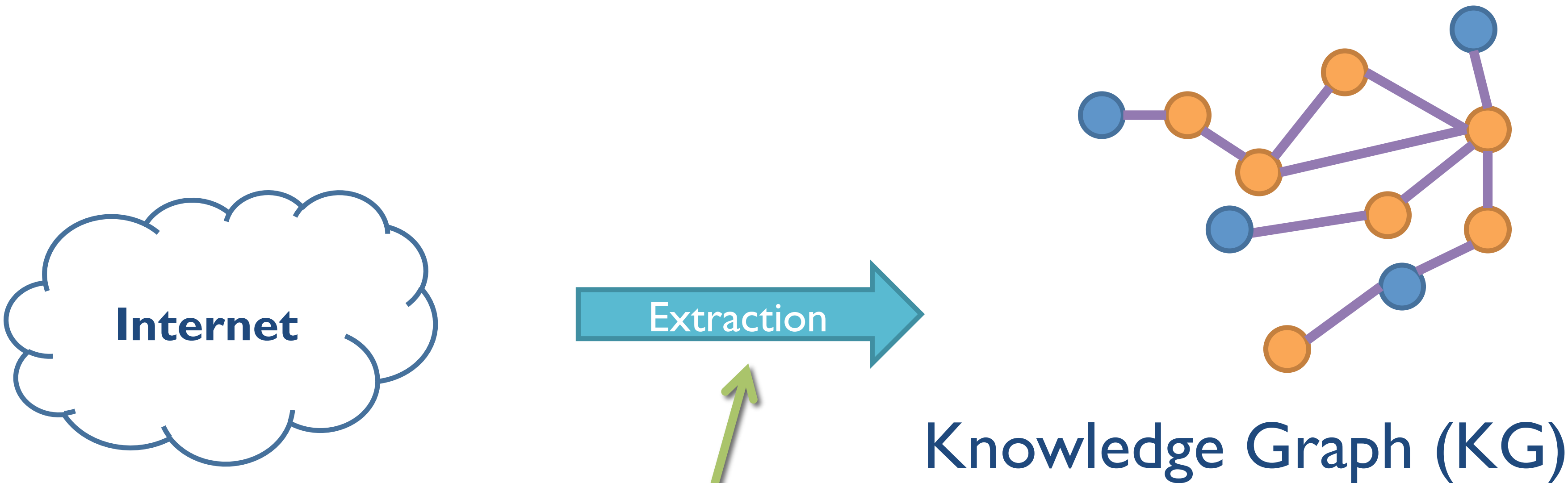
But what about the universe?

- Many approaches are successful at resolving entities, and discovering relationships at the scope of a document

- Building a knowledge base requires resolving entities and relationships across millions of documents



A Revised Knowledge-Creation Diagram

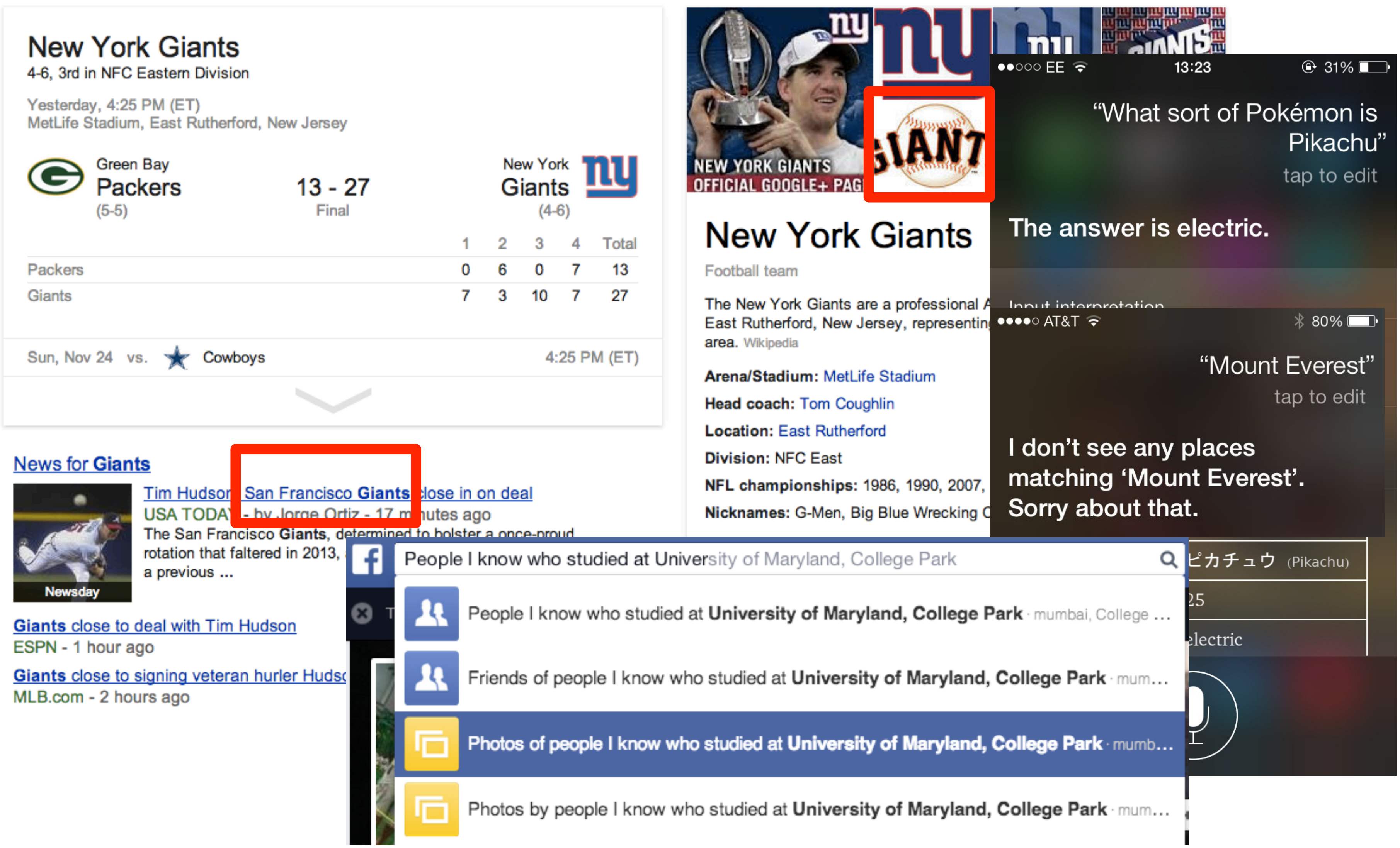


Massive source of publicly available information

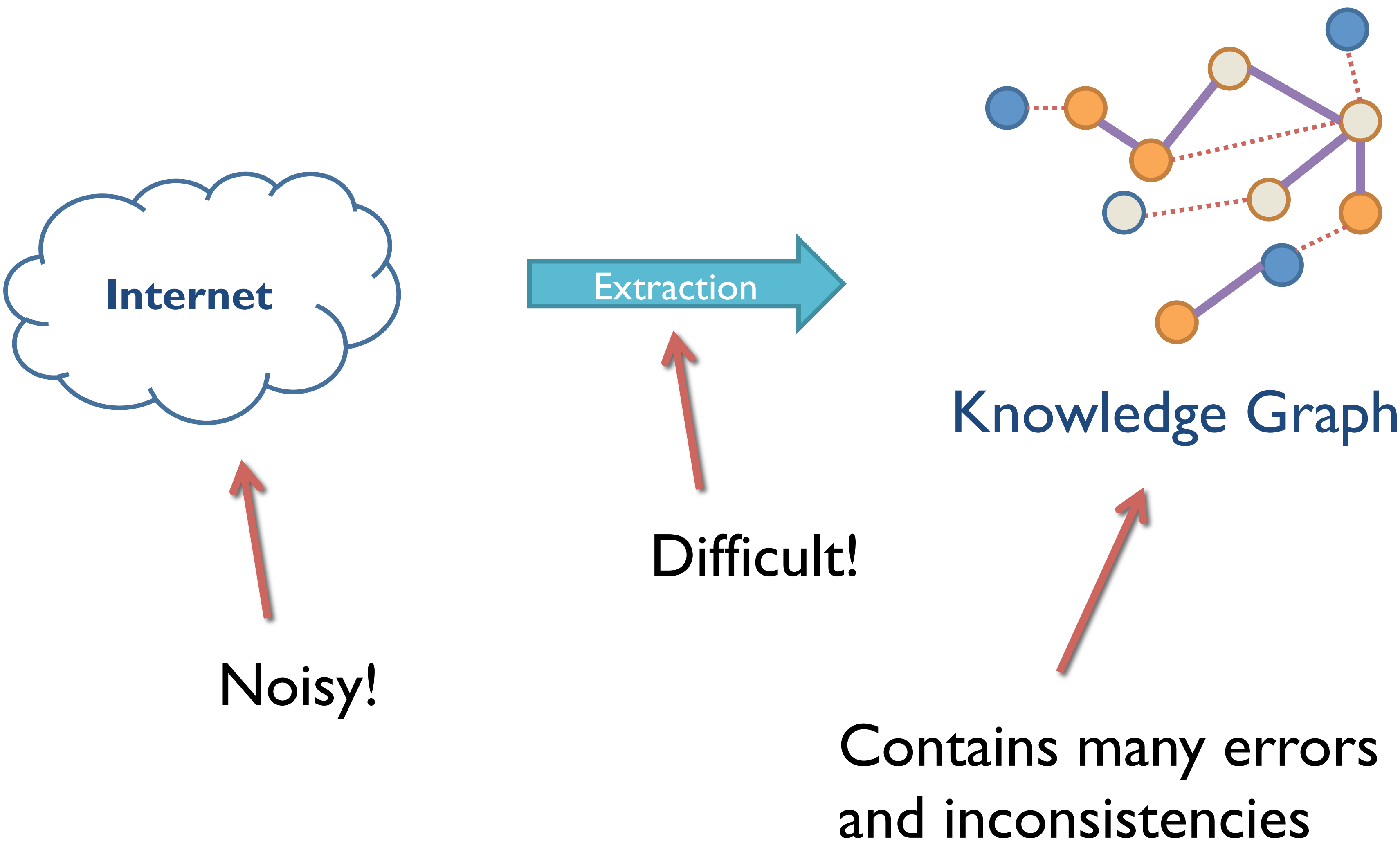
Cutting-edge IE methods

Structured representation of entities, their labels and the relationships between them

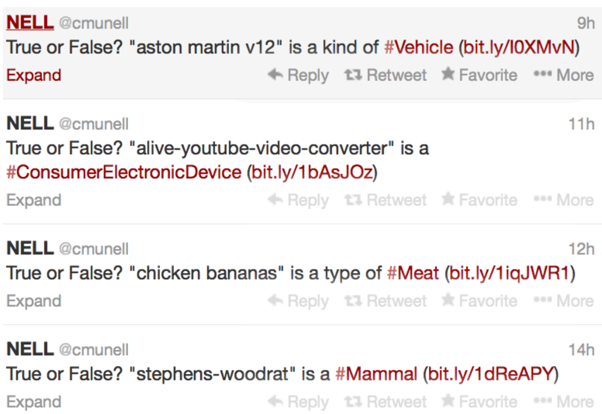
Knowledge Graphs in the wild



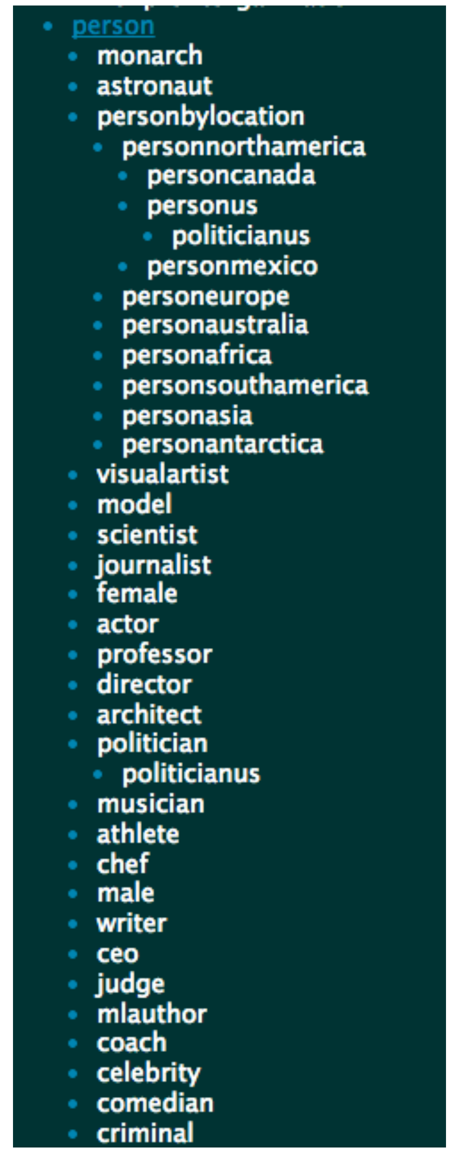
Motivating Problem: Real Challenges



NELL: The Never-Ending Language Learner



- Large-scale IE project (Carlson et al., AAAI10)
- Lifelong learning: aims to "read the web"
- Ontology of known labels and relations
- Knowledge base contains millions of facts



Examples of NELL errors

Entity co-reference errors

Kyrgyzstan has many variants:

- Kyrgystan
- Kyrgistan
- Kyrghyzstan
- Kyrgzstan
- Kyrgyz Republic

Saudi Cultural Days in the Kyrgyz Republic has concluded its activities in the capital Bishkek in the weekend in a special ceremony held on this occasion. The event was attended by Deputy Minister of Culture and Tourism of the Kyrgyz Republic Koulev Mirza; Kyrgyzstan's Ambassador to Saudi Arabia Jusupbek Sharipov; the Saudi Embassy Acting Chargé d'affaires to Kyrgyzstan, Mari bin Barakah Al-Derbas and members of the embassy staff, in the presence of a heavy turnout of Kyrgyz citizens.

The Days of Culture of Saudi Arabia in Kyrgyzstan will be held from 6 to 9 May.

Refugees are often from areas where conflict is historically embedded and marked in ideology and injustice. The Tsarnaev family emigrated from the Chechen diaspora in Kyrgzstan, a region Stalin deported the Chechens to in 1943. After the fall of the Berlin Wall in 1991, Chechens engaged in a battle for independence from Russia that led to the Tsarnaevs' petition for refugee status in the early

Home > Holiday Destinations > Kyrghyzstan > Bishkek > Climate Profile

Fast Forecast

Holiday Weather

Missing and spurious labels

Erik Kleyheeg has just returned from Lesvos with some new bird images. Included here are: Common Scops-Owl, Wood Warbler, Spanish Sparrow, Red-throated Pipit, Eurasian Chiff-chaff, and Cretzschmar's Bunting.

Anssi Kullberg has sent along some great trip reports to unusual places, including Kyrgyzstan, Pakistan,

Kyrgyzstan is labeled a bird and a country

Kyrgyzstan (/kɜrɡɪˈstɑːn/ *kur-gi-STAHN*;[5] Kyrgyz: Кыргызстан (IPA: [qɯrʁɯsˈstɑn]); Russian: Киргизия), officially the **Kyrgyz Republic** (Kyrgyz: Кыргыз Республикасы; Russian: Кыргызская Республика), is a country located in Central Asia.[6] Landlocked and mountainous, Kyrgyzstan is bordered by Kazakhstan to the north, Uzbekistan to the west, Tajikistan to the southwest and China to the east. Its capital and largest city is Bishkek.

Missing and spurious relations

Guidance

Kazakhstan / Kyrgyzstan – Consular Fees

Organisation: Foreign & Commonwealth Office
Page history: Published 4 April 2013

Kyrgyzstan's location is ambiguous – Kazakhstan, Russia and US are included in possible locations

Kyrgyzstan U.S. Air Base Future Unclear

A Central Asian country of incredible natural beauty and proud nomadic traditions, most of Kyrgyzstan was formally annexed to Russia in 1876. The Kyrgyz staged a major revolt against the Tsarist Empire in 1916 in which almost one-sixth of the Kyrgyz population was killed. Kyrgyzstan became a Soviet republic in 1936 and

Violations of ontological knowledge

- Equivalence of co-referent entities (sameAs)
  - SameEntity(Kyrgyzstan, Kyrgyz Republic)
- Mutual exclusion (disjointWith) of labels
  - MUT(bird, country)
- Selectional preferences (domain/range) of relations
  - RNG(countryLocation, continent)

Enforcing these constraints requires **jointly** considering multiple extractions *across* documents

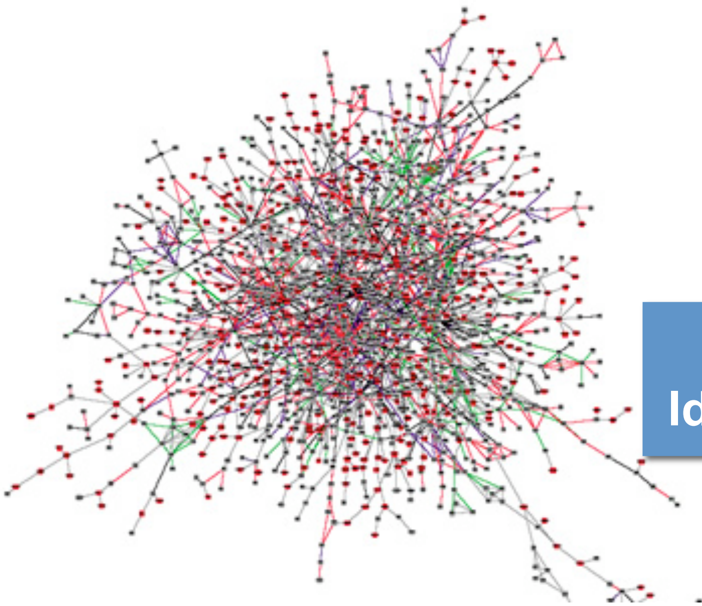
Examples where joint models have succeeded

- Information extraction
  - ER+Segmentation: Poon & Domingos, AAAI07
  - SRL: Srikumar & Roth, EMNLP11
  - Within-doc extraction: Singh et al., AKBC13

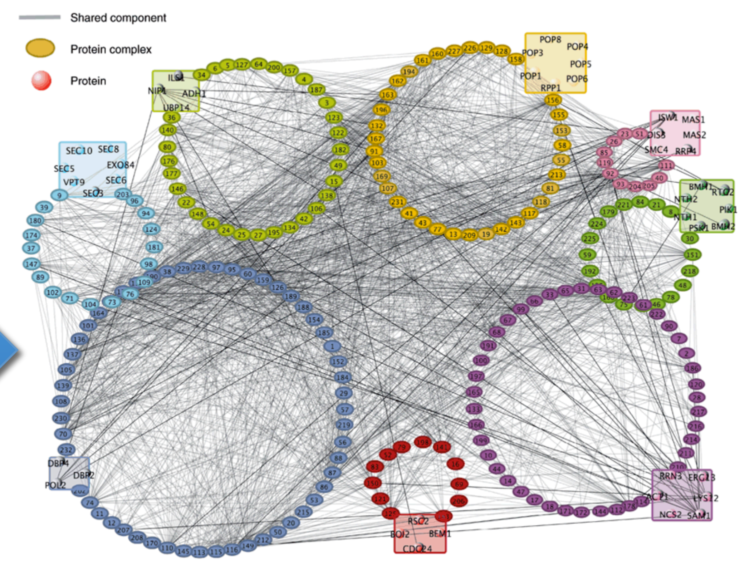
- Social and communication networks
  - Fusion: Eldardiry & Neville, MLG10
  - EMailActs: Carvalho & Cohen, SIGIR05
  - GraphID: Namata et al., KDD11

GRAPH IDENTIFICATION

Transformation



Graph Identification



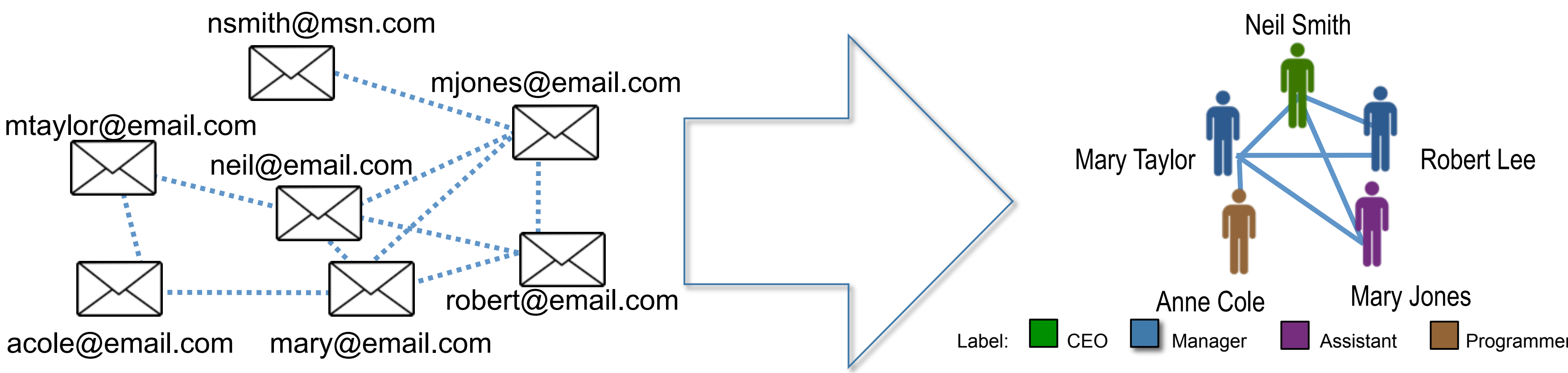
Input Graph

Output Graph

Available but inappropriate for analysis

Appropriate for further analysis

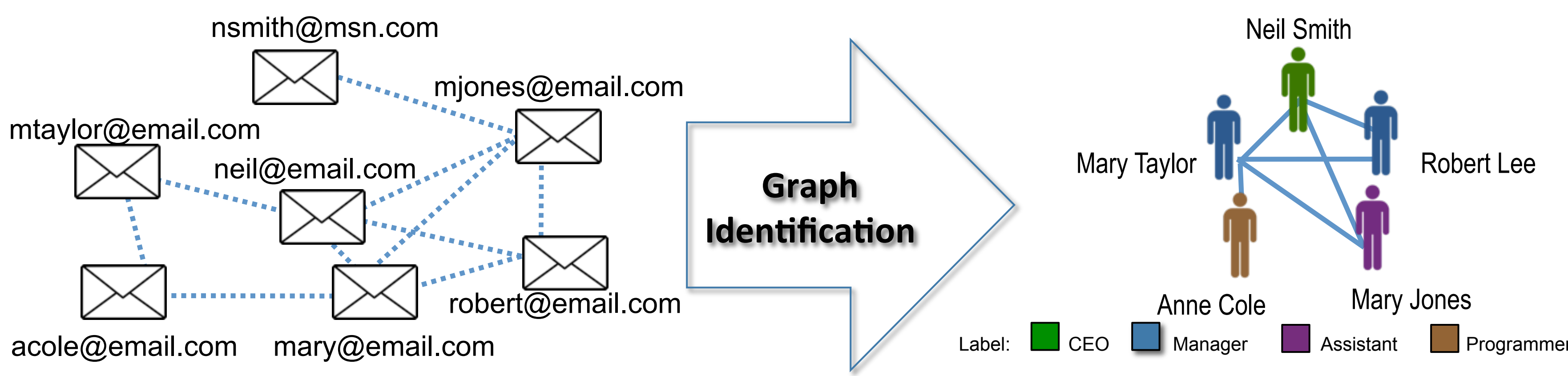
Motivation: Different Networks



<u>Communication Network</u>
Nodes: Email Address
Edges: Communication
Node Attributes: Words

<u>Organizational Network</u>
Nodes: Person
Edges: Manages
Node Labels: Title

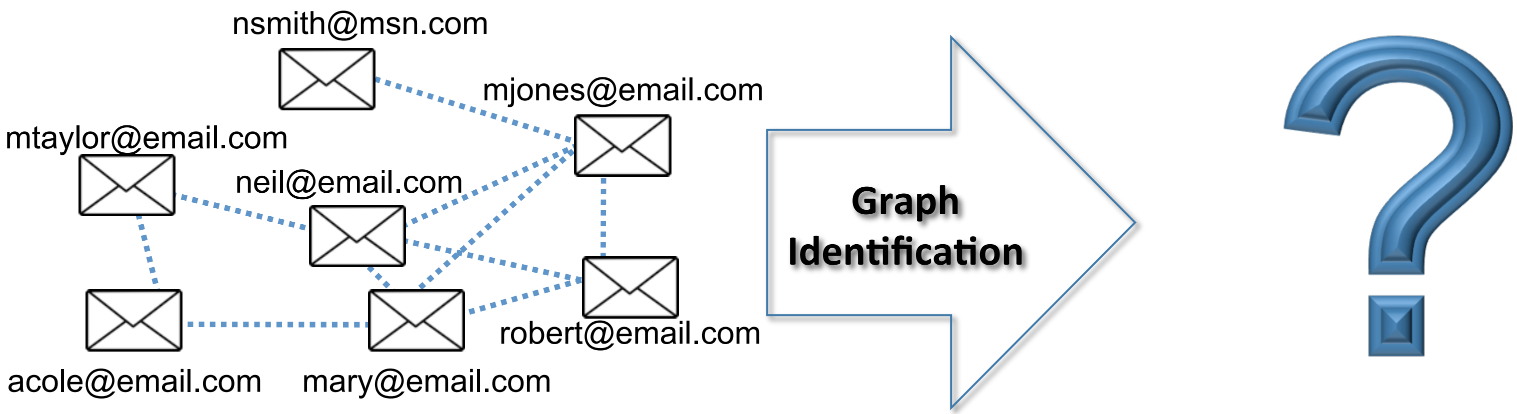
Graph Identification



Input Graph: Email Communication Network

Output Graph: Social Network

Graph Identification

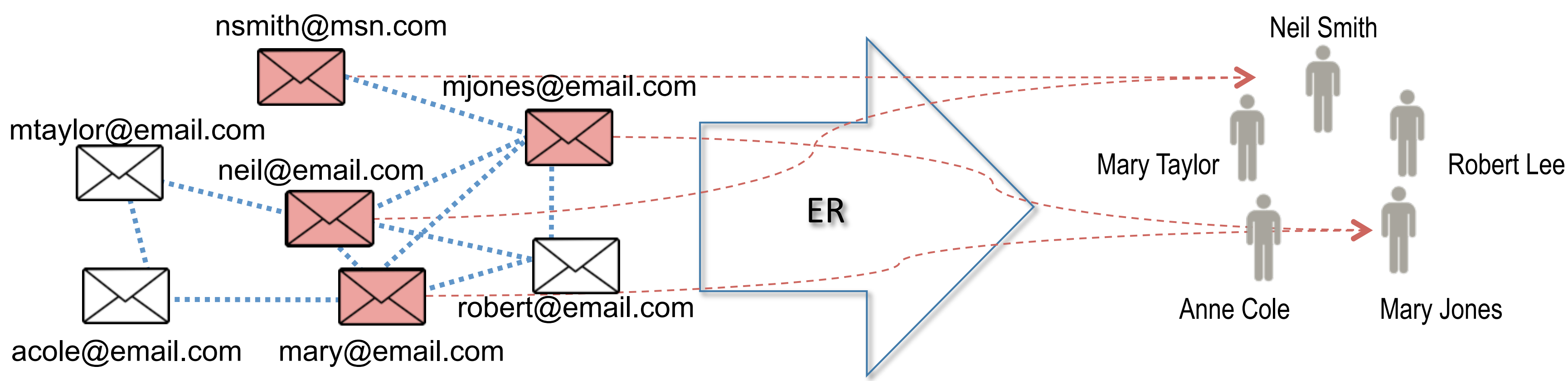


Input Graph: Email Communication Network

Output Graph: Social Network

- What’s involved?

Graph Identification

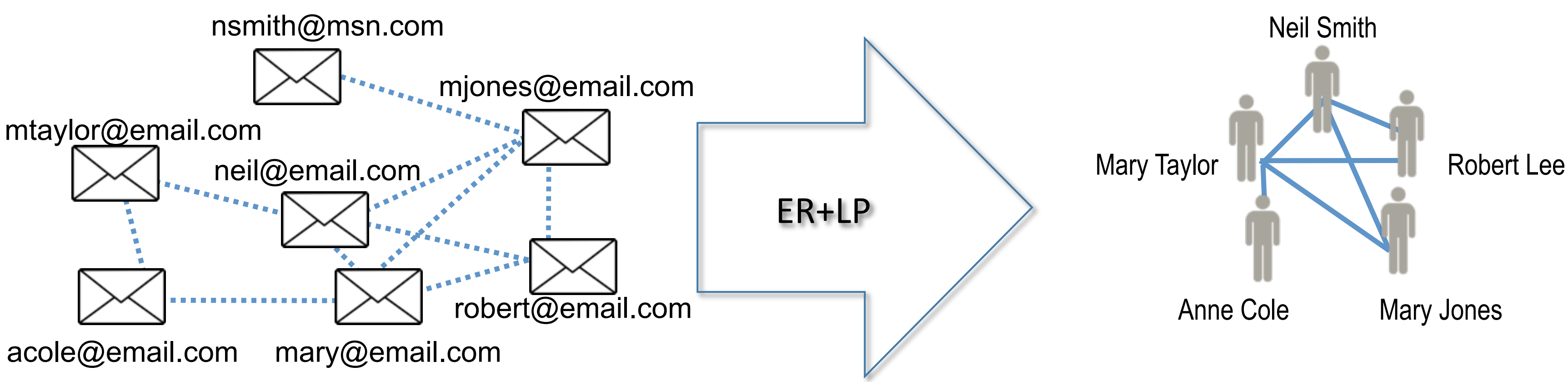


Input Graph: Email Communication Network

Output Graph: Social Network

- What's involved?
  - Entity Resolution (ER): Map input graph nodes to output graph nodes

Graph Identification

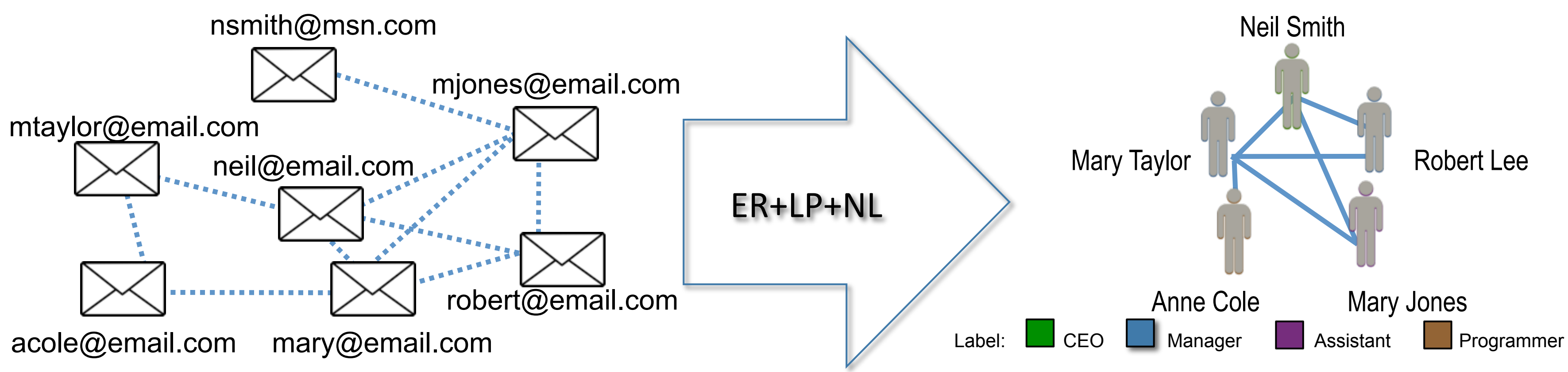


Input Graph: Email Communication Network

Output Graph: Social Network

- What's involved?
  - Entity Resolution (ER): Map input graph nodes to output graph nodes
  - Link Prediction (LP): Predict existence of edges in output graph

Graph Identification

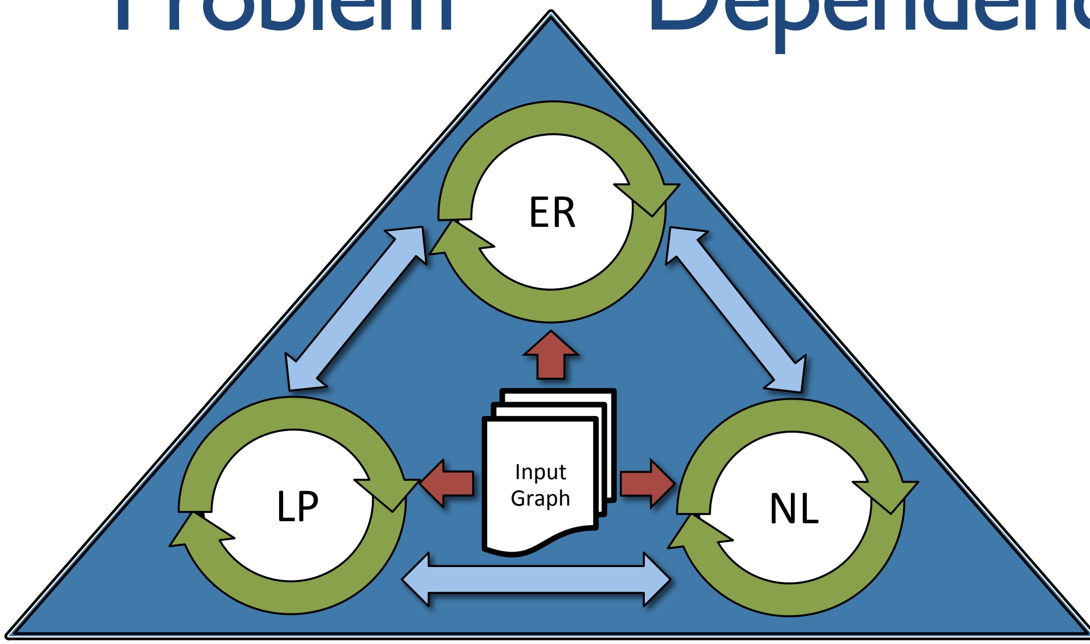


Input Graph: Email Communication Network

Output Graph: Social Network

- What's involved?
  - Entity Resolution (ER): Map input graph nodes to output graph nodes
  - Link Prediction (LP): Predict existence of edges in output graph
  - Node Labeling (NL): Infer the labels of nodes in the output graph

Problem Dependencies

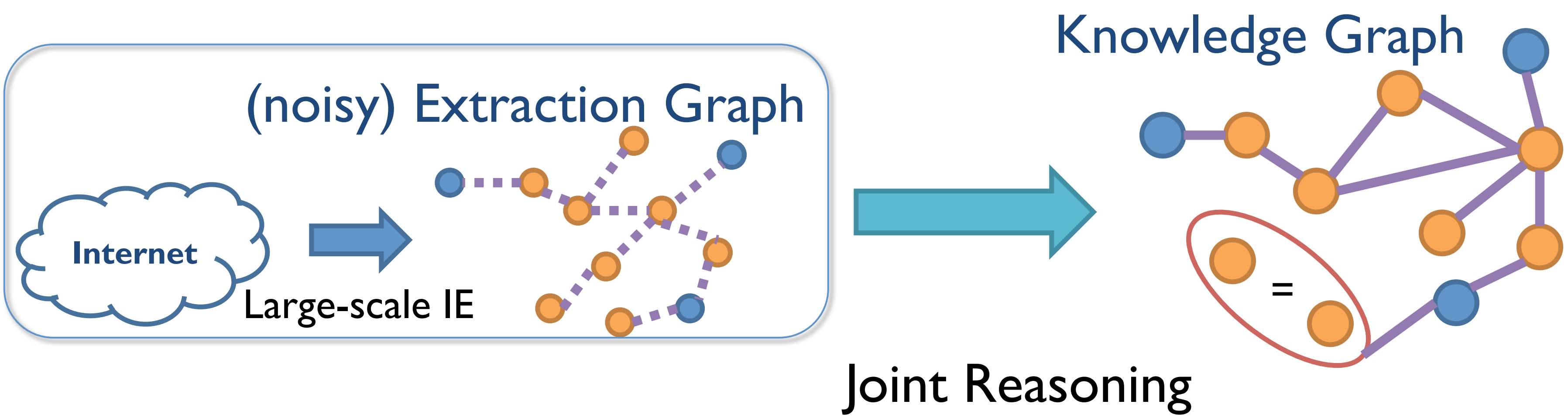


- Most work looks at these tasks in **<u>isolation</u>**
- In graph identification they are:
  - Evidence-Dependent – Inference depend on observed input graph
    e.g., ER depends on input graph
  - Intra-Dependent – Inference <u>within</u> tasks are dependent
    e.g., NL prediction depend on other NL predictions
  - Inter-Dependent – Inference <u>across</u> tasks are dependent
    e.g., LP depend on ER and NL predictions

KNOWLEDGE GRAPH IDENTIFICATION

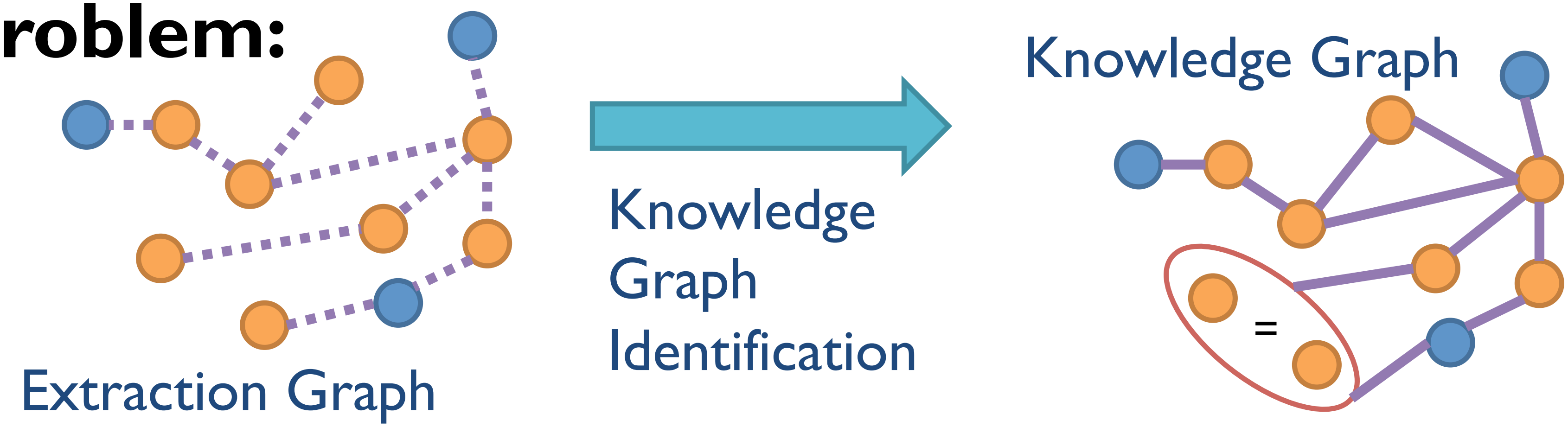
Pujara, Miao, Getoor, Cohen, ISWC 2013 (best student paper)

Motivating Problem (revised)



Knowledge Graph Identification

Problem:



Solution: *Knowledge Graph Identification* (KGI)

- Performs *graph identification*:
  - entity resolution
  - node labeling
  - link prediction
- Enforces *ontological constraints*
- Incorporates *multiple uncertain sources*

Illustration of KGI: Extractions

Uncertain Extractions:

.5: Lbl(Kyrgyzstan, bird)

.7: Lbl(Kyrgyzstan, country)

.9: Lbl(Kyrgyz Republic, country)

.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

Illustration of KGI: Ontology + ER

Uncertain Extractions:

.5: Lbl(Kyrgyzstan, bird)
.7: Lbl(Kyrgyzstan, country)
.9: Lbl(Kyrgyz Republic, country)
.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

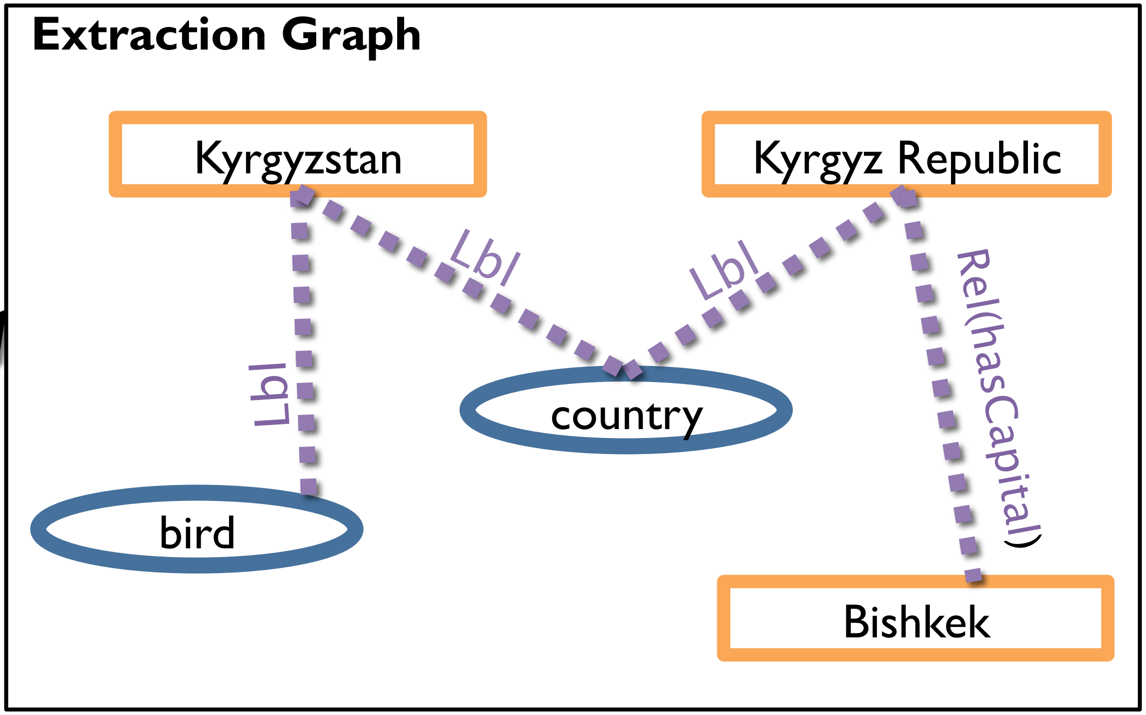


Illustration of KGI: Ontology + ER

Uncertain Extractions:
.5: Lbl(Kyrgyzstan, bird)
.7: Lbl(Kyrgyzstan, country)
.9: Lbl(Kyrgyz Republic, country)
.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

Ontology:
Dom(hasCapital, country)
Mut(country, bird)

Entity Resolution:
SameEnt(Kyrgyz Republic, Kyrgyzstan)

(Annotated) Extraction Graph

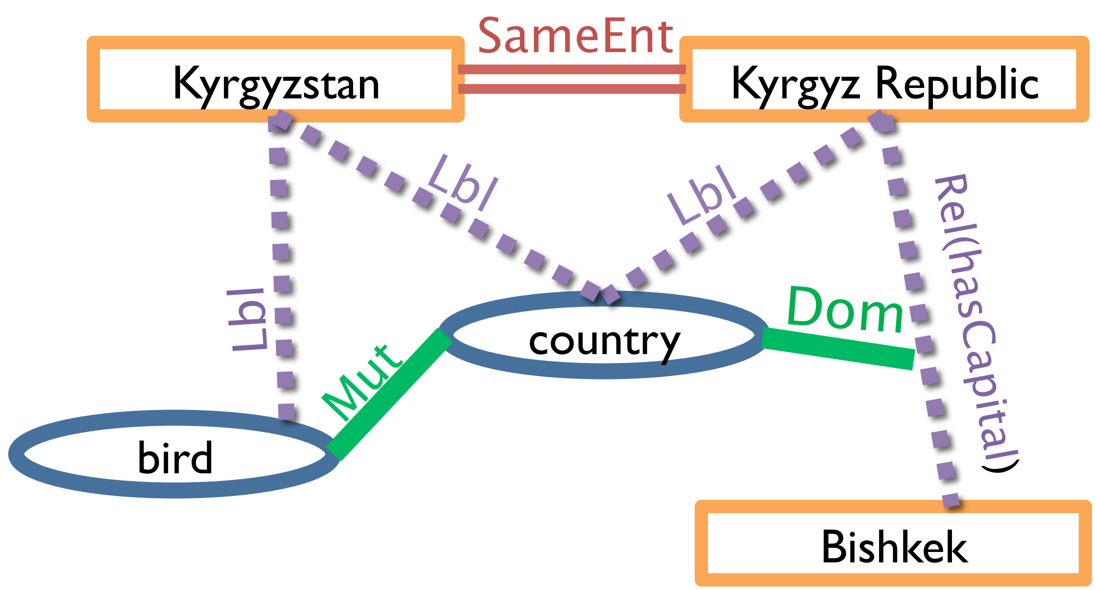
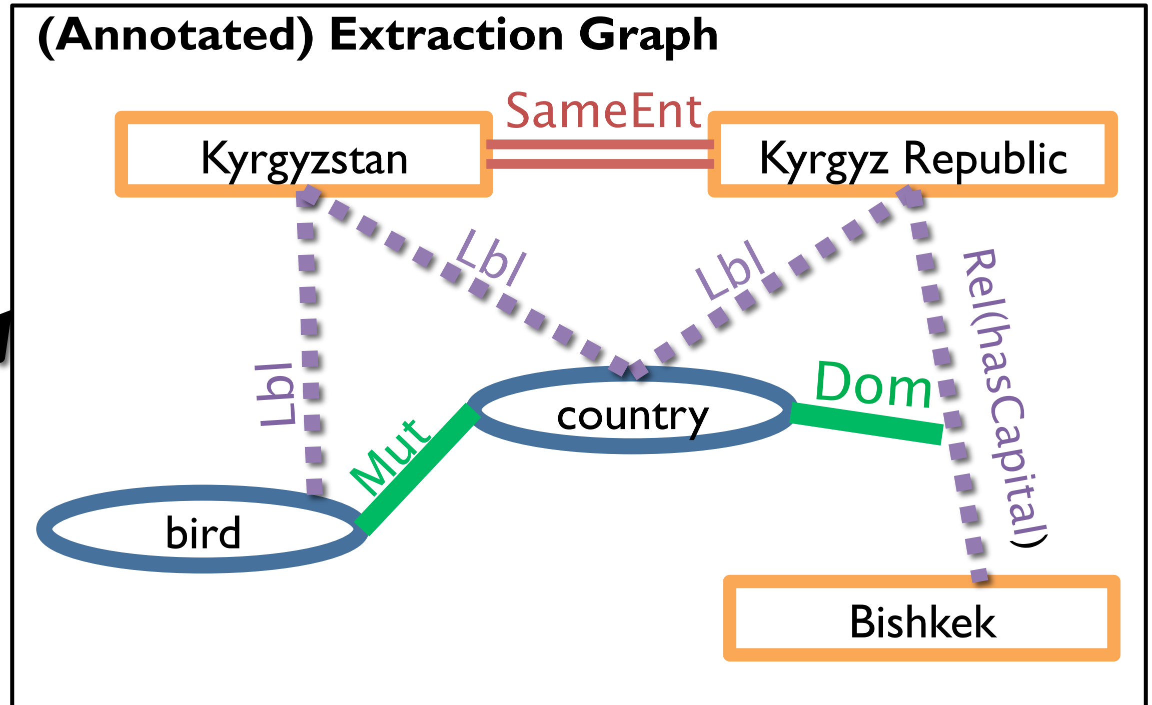


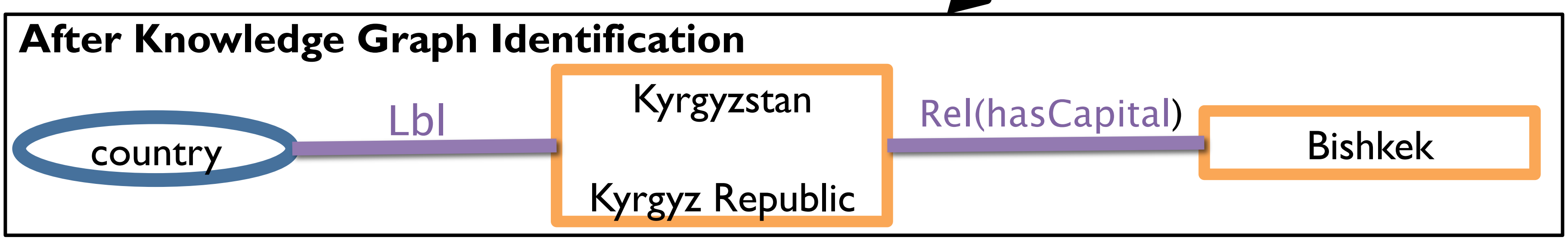
Illustration of KGI

Uncertain Extractions:
.5: Lbl(Kyrgyzstan, bird)
.7: Lbl(Kyrgyzstan, country)
.9: Lbl(Kyrgyz Republic, country)
.8: Rel(Kyrgyz Republic, Bishkek, hasCapital)

Ontology:
Dom(hasCapital, country)
Mut(country, bird)

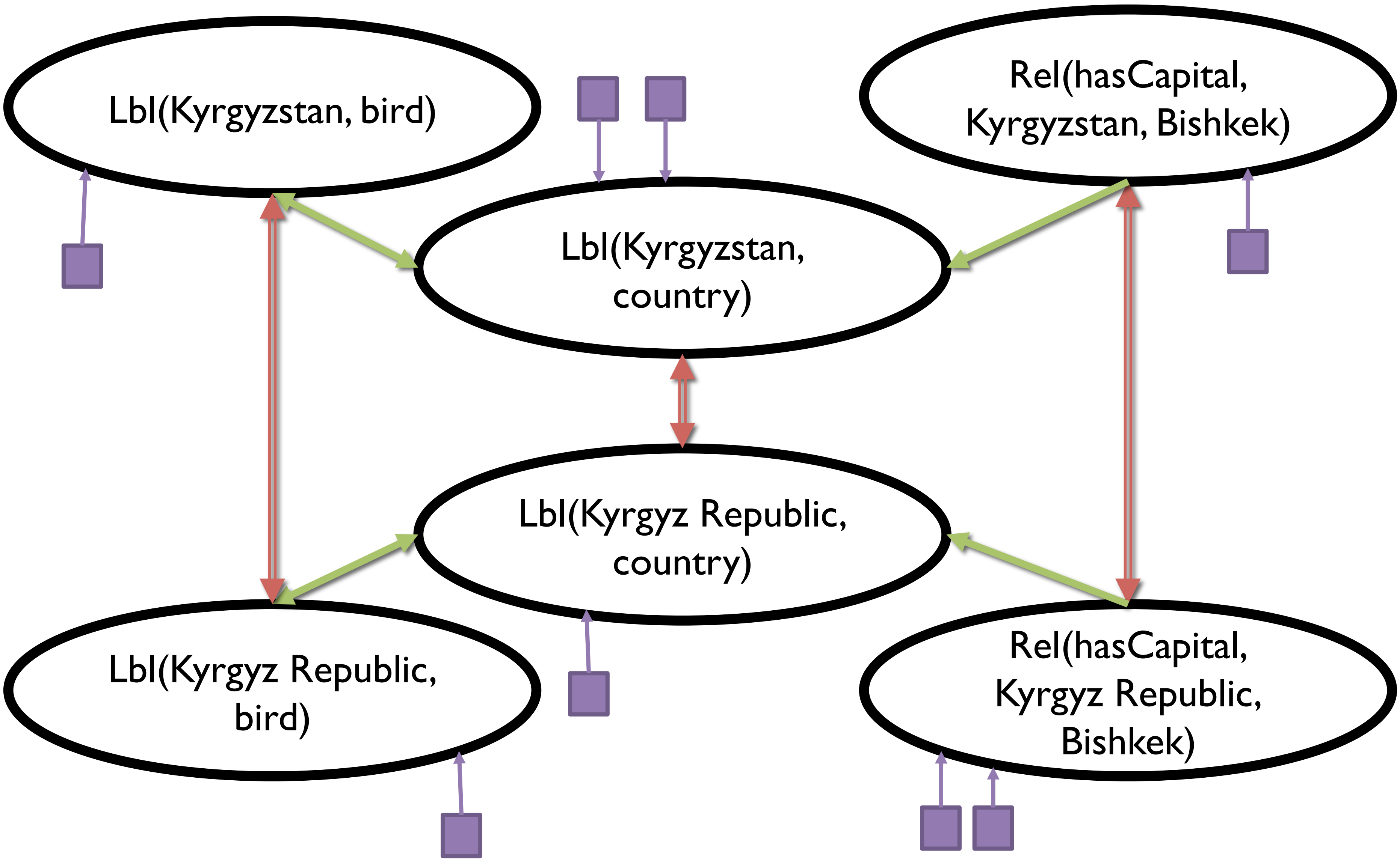
Entity Resolution:
SameEnt(Kyrgyz Republic, Kyrgyzstan)





Modeling Knowledge Graph Identification

Viewing KGI as a probabilistic graphical model



Background: Probabilistic Soft Logic (PSL)

(Broecheler et al., UAI10; Kimming et al., NIPS-ProbProg12)

- Templating language for hinge-loss MRFs, very scalable!
- Model specified as a collection of logical formulas

$$\textsc{SameEnt}(E_1, E_2)\ \tilde{\wedge}\ \textsc{Lbl}(E_1, L) \Rightarrow \textsc{Lbl}(E_2, L)$$

- Uses soft-logic formulation
  - Truth values of atoms relaxed to [0,1] interval
  - Truth values of formulas derived from Lukasiewicz t-norm

Background: PSL Rules to Distributions

- Rules are *grounded* by substituting literals into formulas

$$\mathbf{w_{EL}} : \textsc{SameEnt}(\text{Kyrgyzstan}, \text{Kyrygyz Republic}) \ \tilde{\wedge}$$
$$\textsc{Lbl}(\text{Kyrgyzstan}, \text{country}) \Rightarrow \textsc{Lbl}(\text{Kyrygyz Republic}, \text{country})$$

- Each ground $\boxed{\text{rule}}$ has a $\boxed{\text{weighted}}$ $\boxed{\textit{distance to satisfaction}}$ derived from the formula's truth value

$$\boxed{P(G \mid E)} = \frac{1}{Z}\exp\left[-\sum_{r \in R} w_r \varphi_r(G)\right]$$

- The PSL program can be interpreted as a joint probability distribution over all variables in knowledge graph, conditioned on the extractions

Background: Finding the best knowledge graph

- MPE inference solves $\max_G P(G)$ to find the best KG
- In PSL, inference solved by convex optimization
- Efficient: running time empirically scales with $O(|R|)$
  (Bach et al., NIPS12)

PSL Rules for KGI Model

PSL Rules: Uncertain Extractions

Weight for source T (relations)

Predicate representing uncertain relation extraction from extractor T

Relation in Knowledge Graph

$$\mathbf{w}_{\mathbf{CR}\text{-}T}: \quad \textsc{CandRel}_T(E_1, E_2, R) \qquad \Rightarrow \quad \textsc{Rel}(E_1, E_2, R)$$

$$\mathbf{w}_{\mathbf{CL}\text{-}T}: \quad \textsc{CandLbl}_T(E, L) \qquad \Rightarrow \quad \textsc{Lbl}(E, L)$$

Weight for source T (labels)

Predicate representing uncertain label extraction from extractor T

Label in Knowledge Graph

PSL Rules: Entity Resolution

$$\mathbf{w_{EL}} : \textsc{SameEnt}(E_1, E_2) \tilde{\wedge} \textsc{Lbl}(E_1, L) \Rightarrow \textsc{Lbl}(E_2, L)$$

$$\mathbf{w_{ER}} : \textsc{SameEnt}(E_1, E_2) \tilde{\wedge} \textsc{Rel}(E_1, E, R) \Rightarrow \textsc{Rel}(E_2, E, R)$$

$$\mathbf{w_{ER}} : \textsc{SameEnt}(E_1, E_2) \tilde{\wedge} \textsc{Rel}(E, E_1, R) \Rightarrow \textsc{Rel}(E, E_2, R)$$

SameEnt predicate captures confidence that entities are co-referent

- Rules require co-referent entities to have the same labels and relations

- Creates an *equivalence class* of co-referent entities

PSL Rules: Ontology

Inverse:

$$\mathbf{w_O}: \textsc{Inv}(R,S) \;\tilde{\wedge}\; \textsc{Rel}(E_1,E_2,R) \Rightarrow \textsc{Rel}(E_2,E_1,S)$$

Selectional Preference:

$$\mathbf{w_O}: \textsc{Dom}(R,L) \;\tilde{\wedge}\; \textsc{Rel}(E_1,E_2,R) \Rightarrow \textsc{Lbl}(E_1,L)$$
$$\mathbf{w_O}: \textsc{Rng}(R,L) \;\tilde{\wedge}\; \textsc{Rel}(E_1,E_2,R) \Rightarrow \textsc{Lbl}(E_2,L)$$

Subsumption:

$$\mathbf{w_O}: \textsc{Sub}(L,P) \;\tilde{\wedge}\; \textsc{Lbl}(E,L) \Rightarrow \textsc{Lbl}(E,P)$$
$$\mathbf{w_O}: \textsc{RSub}(R,S) \;\tilde{\wedge}\; \textsc{Rel}(E_1,E_2,R) \Rightarrow \textsc{Rel}(E_1,E_2,S)$$

Mutual Exclusion:

$$\mathbf{w_O}: \textsc{Mut}(L_1,L_2) \;\tilde{\wedge}\; \textsc{Lbl}(E,L_1) \Rightarrow \tilde{\neg}\textsc{Lbl}(E,L_2)$$
$$\mathbf{w_O}: \textsc{RMut}(R,S) \;\tilde{\wedge}\; \textsc{Rel}(E_1,E_2,R) \Rightarrow \tilde{\neg}\textsc{Rel}(E_1,E_2,S)$$

Adapted from Jiang et al., ICDM 2012

$[\phi_1]$ $\textsc{CandLbl}_{\text{struct}}(\texttt{Kyrgyzstan}, \texttt{bird})$

$\Rightarrow$ $\textsc{Lbl}(\texttt{Kyrgyzstan}, \texttt{bird})$

$[\phi_2]$ $\textsc{CandRel}_{\text{pat}}(\texttt{Kyrgyz Rep.}, \texttt{Asia}, \texttt{locatedIn})$

$\Rightarrow$ $\textsc{Rel}(\texttt{Kyrgyz Rep.}, \texttt{Asia}, \texttt{locatedIn})$

$[\phi_3]$ $\textsc{SameEnt}(\texttt{Kyrgyz Rep.}, \texttt{Kyrgyzstan})$

$\wedge\, \textsc{Lbl}(\texttt{Kyrgyz Rep.}, \texttt{country})$

$\Rightarrow \textsc{Lbl}(\texttt{Kyrgyzstan}, \texttt{country})$

$[\phi_4]$ $\textsc{Dom}(\texttt{locatedIn}, \texttt{country})$

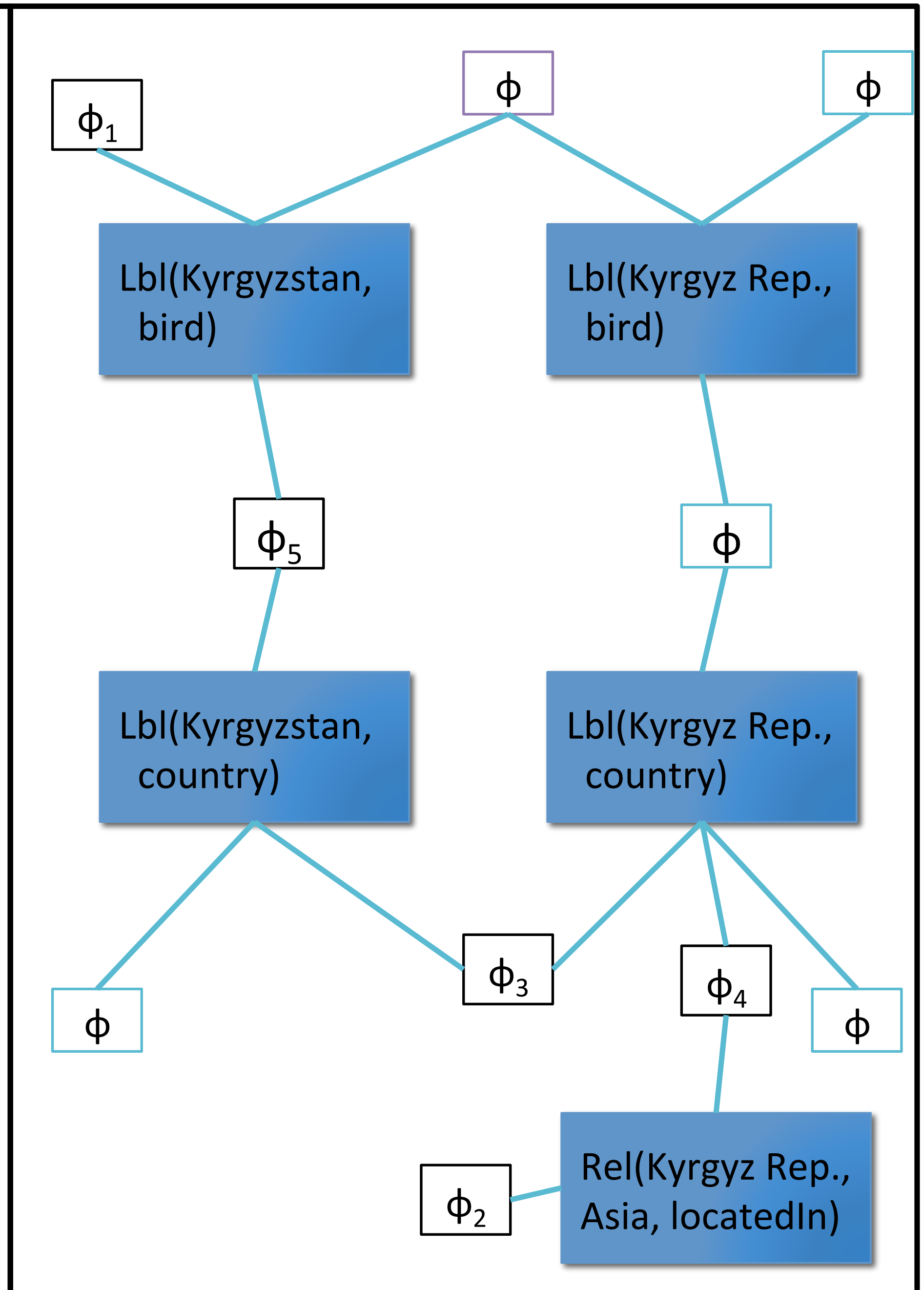
$\wedge\, \textsc{Rel}(\texttt{Kyrgyz Rep.}, \texttt{Asia}, \texttt{locatedIn})$

$\Rightarrow \textsc{Lbl}(\texttt{Kyrgyz Rep.}, \texttt{country})$

$[\phi_5]$ $\textsc{Mut}(\texttt{country}, \texttt{bird})$

$\wedge\, \textsc{Lbl}(\texttt{Kyrgyzstan}, \texttt{country})$

$\Rightarrow \neg\textsc{Lbl}(\texttt{Kyrgyzstan}, \texttt{bird})$



Probability Distribution over KGs

$$P(G \mid E) = \frac{1}{Z}\exp\left[-\sum_{r \in R} w_r \, \varphi_r(G)\right]$$

$$\left\{\begin{array}{ll} \textsc{CandLbl}_T(\texttt{kyrgyzstan}, \texttt{bird}) & \Rightarrow \textsc{Lbl}(\texttt{kyrgyzstan}, \texttt{bird}) \\ \textsc{Mut}(\texttt{bird}, \texttt{country}) & \tilde{\wedge} \; \textsc{Lbl}(\texttt{kyrgyzstan}, \texttt{bird}) \\ & \Rightarrow \tilde{\neg}\textsc{Lbl}(\texttt{kyrgyzstan}, \texttt{country}) \\ \textsc{SameEnt}(\texttt{kyrgz republic}, \texttt{kyrgyzstan}) & \tilde{\wedge} \; \textsc{Lbl}(\texttt{kyrgz republic}, \texttt{country}) \\ & \Rightarrow \textsc{Lbl}(\texttt{kyrgyzstan}, \texttt{country}) \end{array}\right.$$

Evaluation

Two Evaluation Datasets

	LinkedBrainz	NELL
Description	Community-supplied data about musical artists, labels, and creative works	Real-world IE system extracting general facts from the WWW
Noise	Realistic synthetic noise	Imperfect extractors and ambiguous web pages
Candidate Facts	810K	1.3M
Unique Labels and Relations	27	456
Ontological Constraints	49	67.9K

LinkedBrainz

MusicBrainz

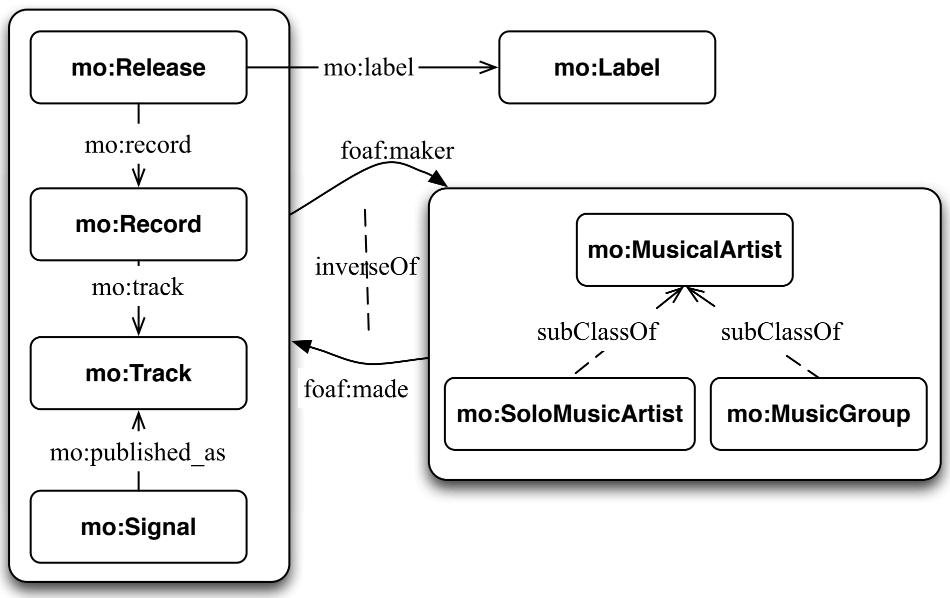
- Open source community-driven structured database of music metadata
- Uses proprietary schema to represent data

music ontology

- Built on popular ontologies such as FOAF and FRBR
- Widely used for music data (e.g. BBC Music Site)

LinkedBrainz project provides an RDF mapping from MusicBrainz data to Music Ontology using the D2RQ tool

LinkedBrainz dataset for KGI



Mapping to FRBR/FOAF ontology	
DOM	rdfs:domain
RNG	rdfs:range
INV	owl:inverseOf
SUB	rdfs:subClassOf
RSUB	rdfs:subPropertyOf
MUT	owl:disjointWith

LinkedBrainz experiments

Comparisons:

Baseline Use noisy truth values as fact scores

PSL-EROnly Only apply rules for **E**ntity **R**esolution

PSL-OntOnly Only apply rules for **Ont**ological reasoning

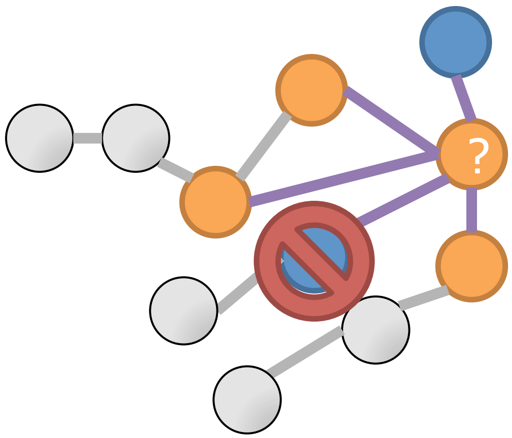
PSL-KGI Apply **K**nowledge **G**raph **I**dentification model

	AUC	Precision	Recall	F1 at .5	Max F1
Baseline	0.672	0.946	0.477	0.634	0.788
PSL-EROnly	0.797	0.953	0.558	0.703	0.831
PSL-OntOnly	0.753	0.964	0.605	0.743	0.832
PSL-KGI	0.901	0.970	0.714	0.823	0.919

NELL Evaluation: two settings

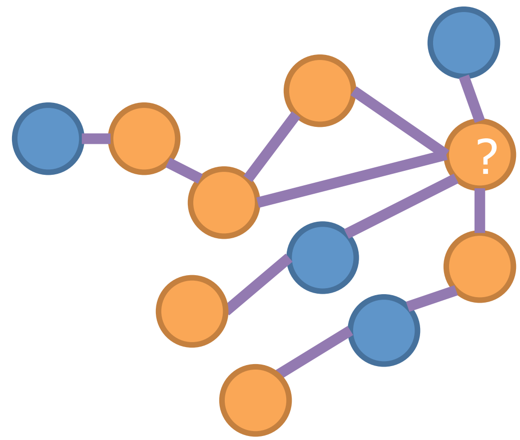
Target Set: restrict to a subset of KG

(Jiang, ICDM12)



- Closed-world model
- Uses a target set: subset of KG
- Derived from 2-hop neighborhood
- Excludes trivially satisfied variables

Complete: Infer full knowledge graph



- Open-world model
- All possible entities, relations, labels
- Inference assigns truth value to each variable

NELL experiments: Target Set

Task: Compute truth values of a target set derived from the evaluation data

Comparisons:

Baseline Average confidences of extractors for each fact in the NELL candidates
NELL Evaluate NELL's promotions (on the full knowledge graph)
MLN Method of (Jiang, ICDM12) – estimates marginal probabilities with MC-SAT
PSL-KGI Apply full Knowledge Graph Identification model

Running Time: Inference completes in 10 seconds, values for 25K facts

	AUC	F1
Baseline	.873	.828
NELL	.765	.673
MLN (Jiang, 12)	.899	.836
PSL-KGI	.904	.853

NELL experiments: Complete knowledge graph

Task: Compute a full knowledge graph from uncertain extractions

Comparisons:

NELL NELL's strategy: ensure ontological consistency with existing KB

PSL-KGI Apply full Knowledge Graph Identification model

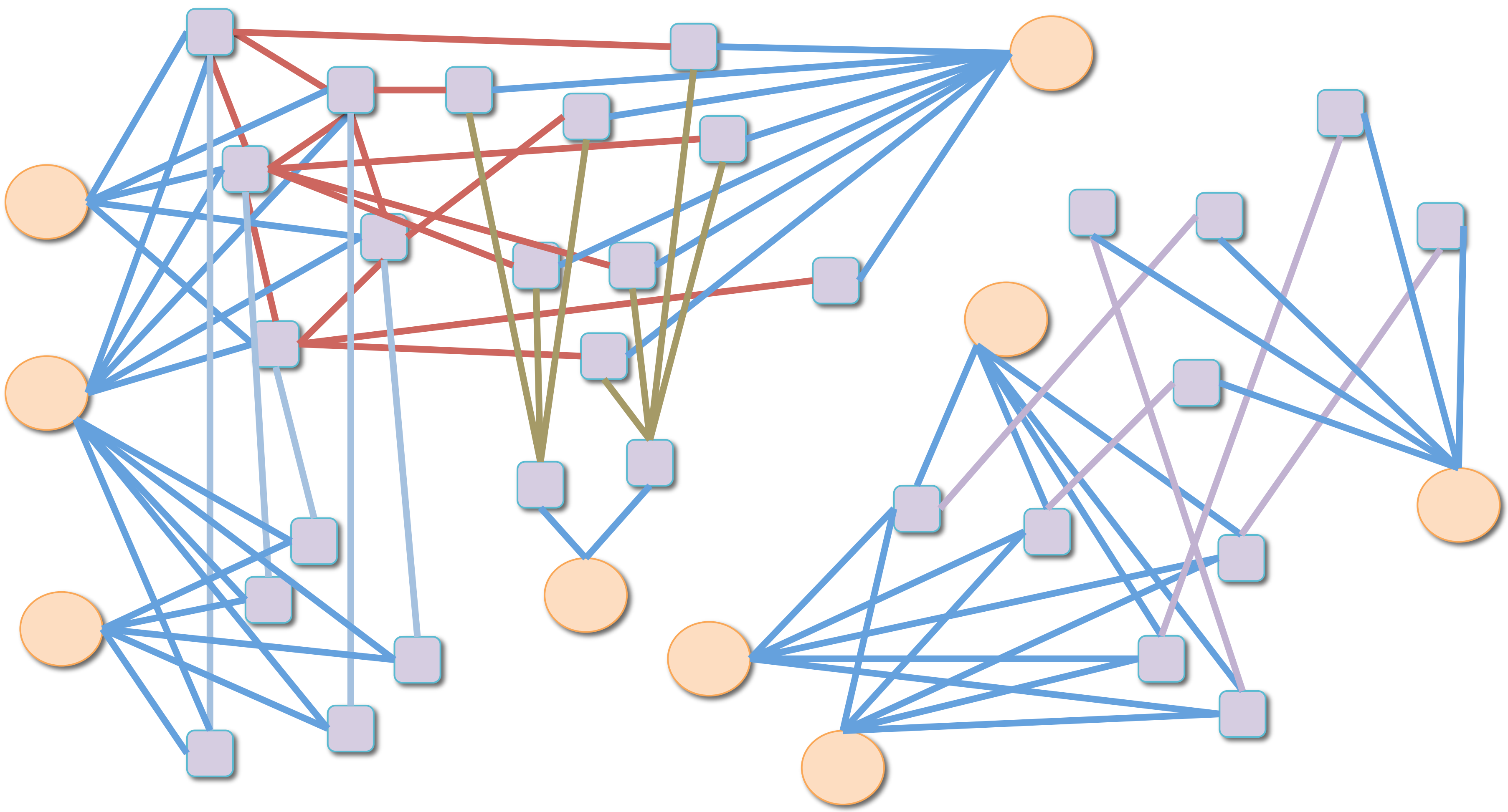
Running Time: Inference completes in 130 minutes, producing 4.3M facts

	AUC	Precision	Recall	F1
NELL	0.765	0.801	0.477	0.634
PSL-KGI	0.892	0.826	0.871	0.848

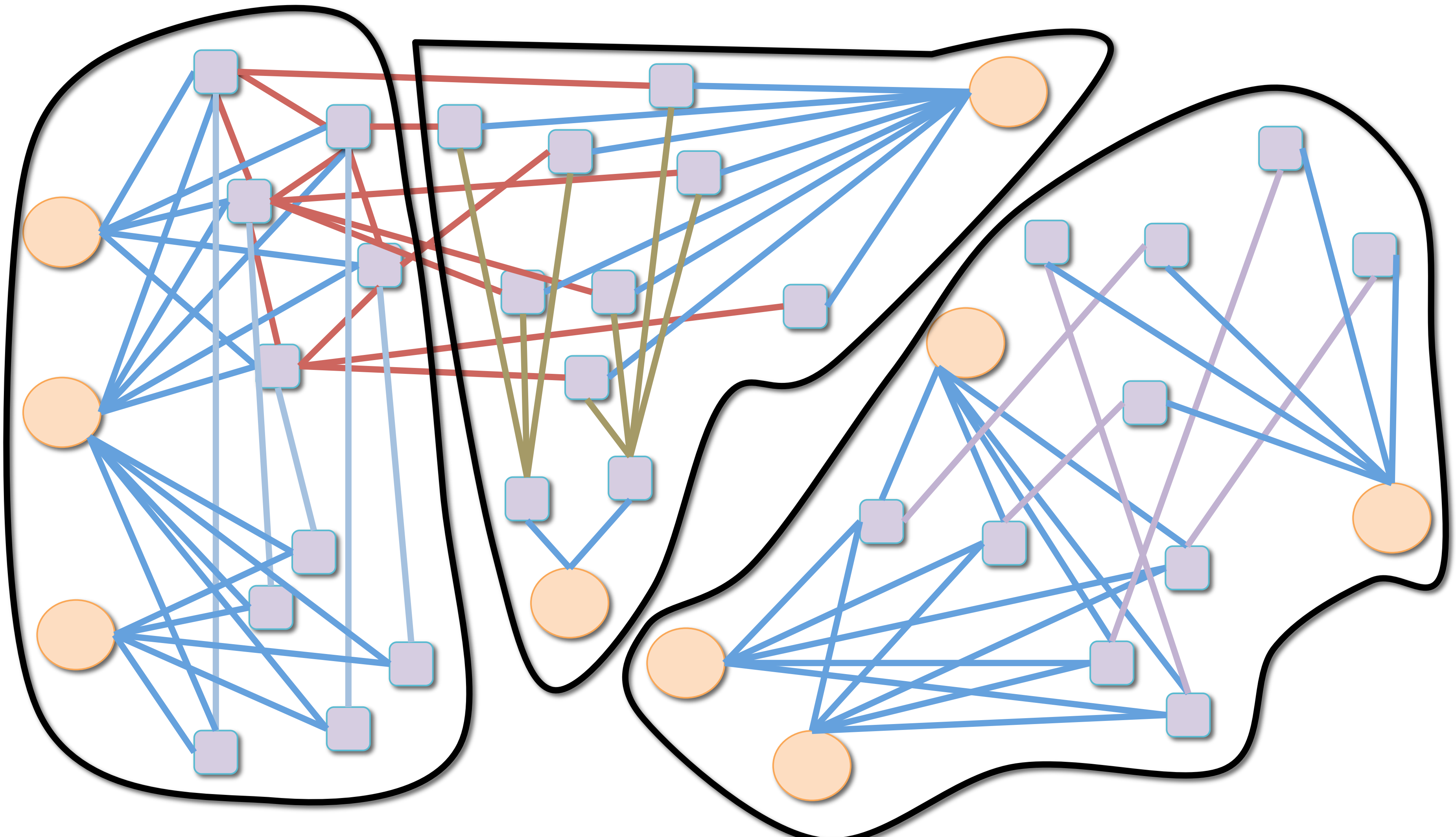
RESEARCH IDEAS

Scalability

Problem: Knowledge Graphs are HUGE



Solution: Partition the Knowledge Graph

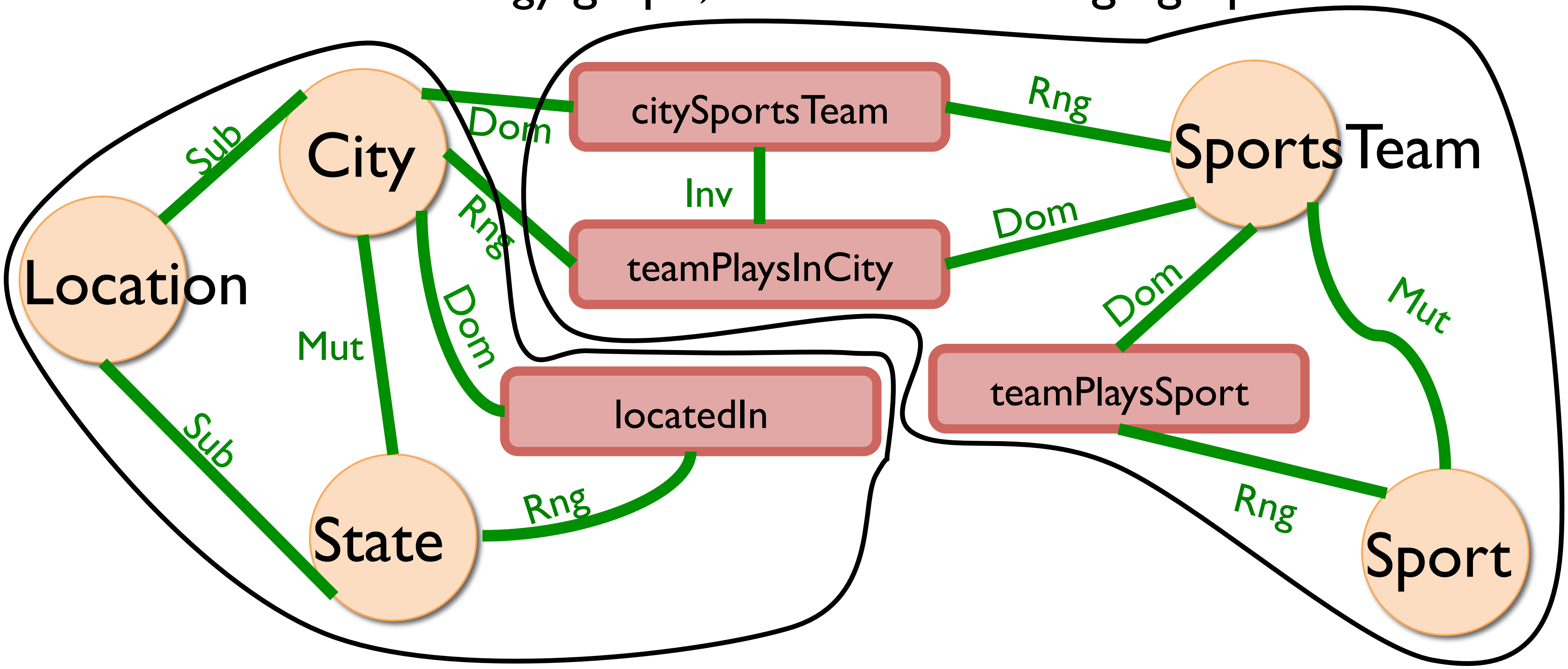


Partitioning: advantages and drawbacks

- Advantages
  - Smaller problems
  - Parallel Inference
  - Speed / Quality Tradeoff
- Drawbacks
  - Partitioning large graph time-consuming
  - Key dependencies may be lost
  - New facts require re-partitioning

Key idea: Ontology-aware partitioning

- Partition the *ontology* graph, not the knowledge graph



- Induce a partitioning of the knowledge graph based on the ontology partition

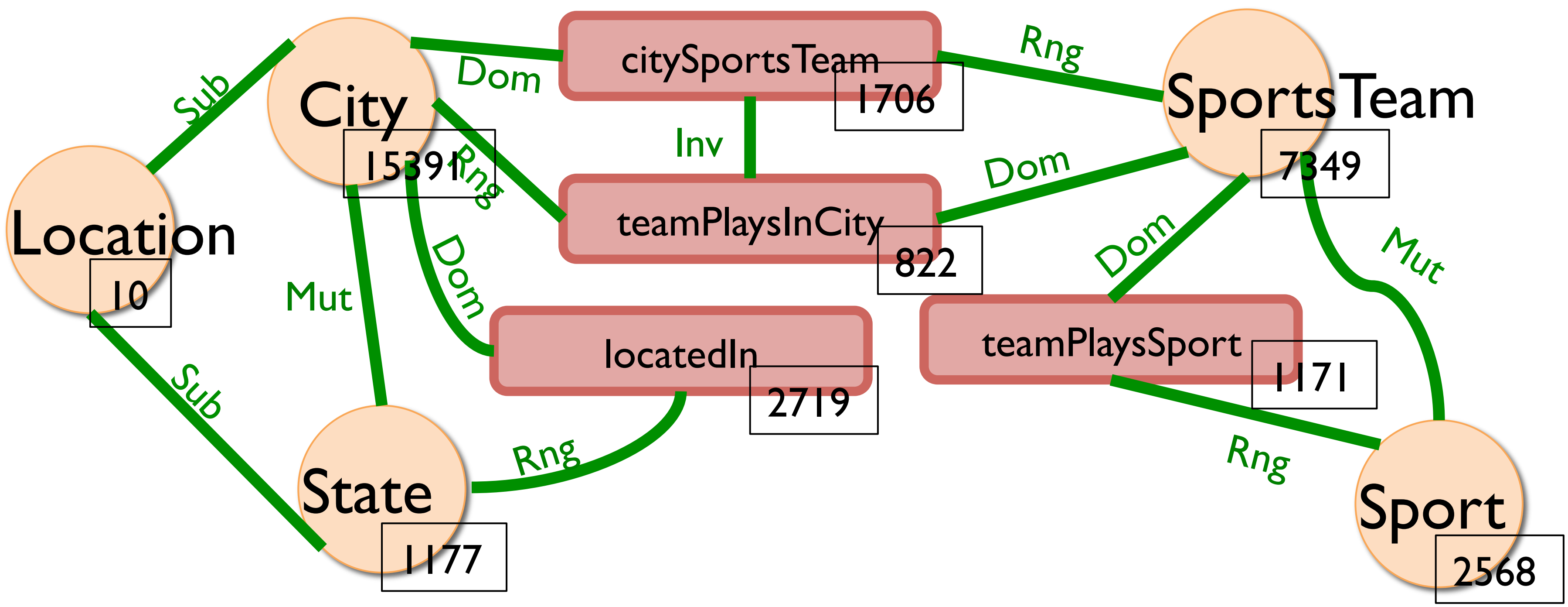
Considerations: Ontology-aware Partitions

- Advantages:
  - Ontology is a smaller graph
  - Ontology coupled with dependencies
  - New facts can reuse partitions

- Disadvantages:
  - Insensitive to data distribution
  - All dependencies treated equally

Refinement: include data frequency

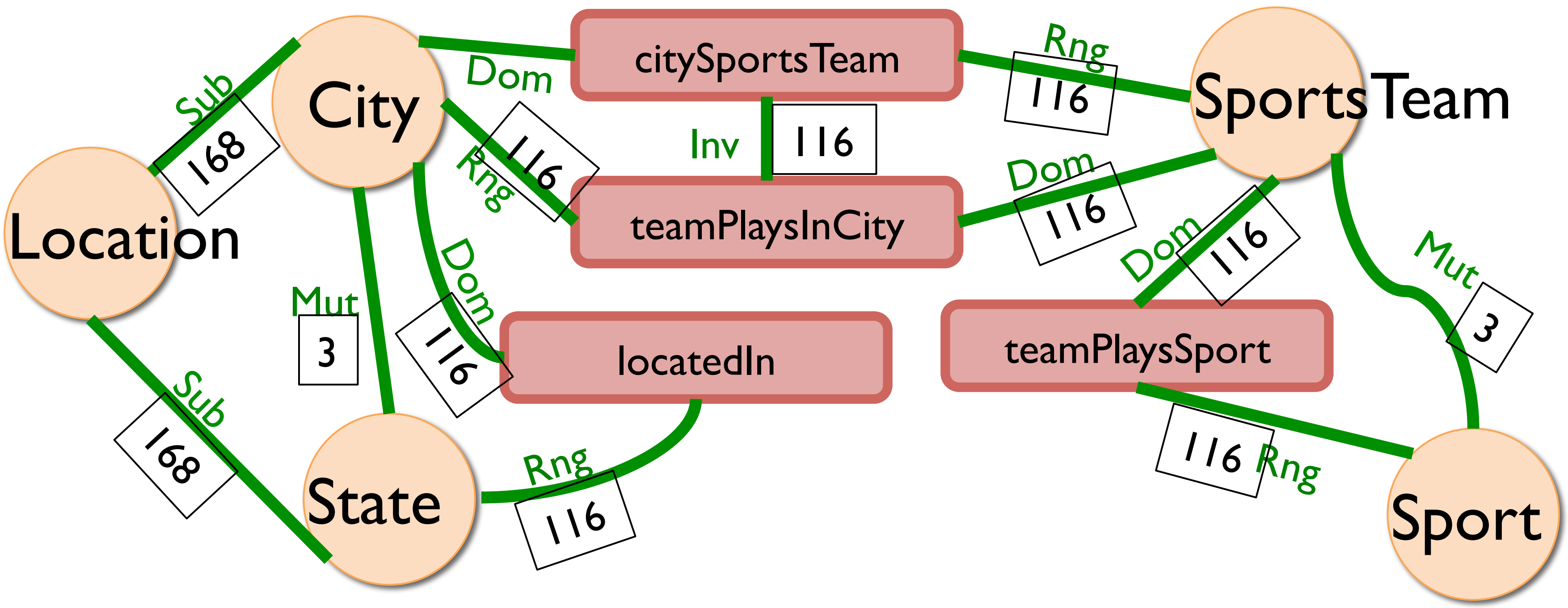
- Annotate each ontological element with its frequency



- Partition ontology with constraint of equal vertex weights

Refinement: weight edges by type

- Weight edges by their ontological importance



Experiments: Partitioning Approaches

Comparisons (6 partitions):

NELL	Default promotion strategy, no KGI
KGI	No partitioning, full knowledge graph model
baseline	KGI, Randomly assign extractions to partition
Ontology	KGI, Edge min-cut of ontology graph
O+Vertex	KGI, Weight ontology vertices by frequency
O+V+Edge	KGI, Weight ontology edges by inv. frequency

	AUPRC	**Running Time** (min)	**Opt. Terms**
NELL	0.765	-	
KGI	**0.794**	97	10.9M
baseline	0.780	**31**	3.0M
Ontology	0.788	42	4.2M
O+Vertex	0.791	**31**	3.7M
O+V+Edge	0.790	**31**	3.7M

Richer Models

Can we add more complex rules?

- The knowledge graph can have very intricate relationships between facts:

$$\textsc{CandRel}(A, T, \texttt{AthletePlaysForTeam})\ \tilde{\wedge}$$
$$\textsc{CandRel}(T, L, \texttt{TeamPlaysInLeague})$$
$$\Rightarrow \textsc{CandRel}(A, L, \texttt{AthletePlaysInLeague})$$

Can we formalize these relationships?

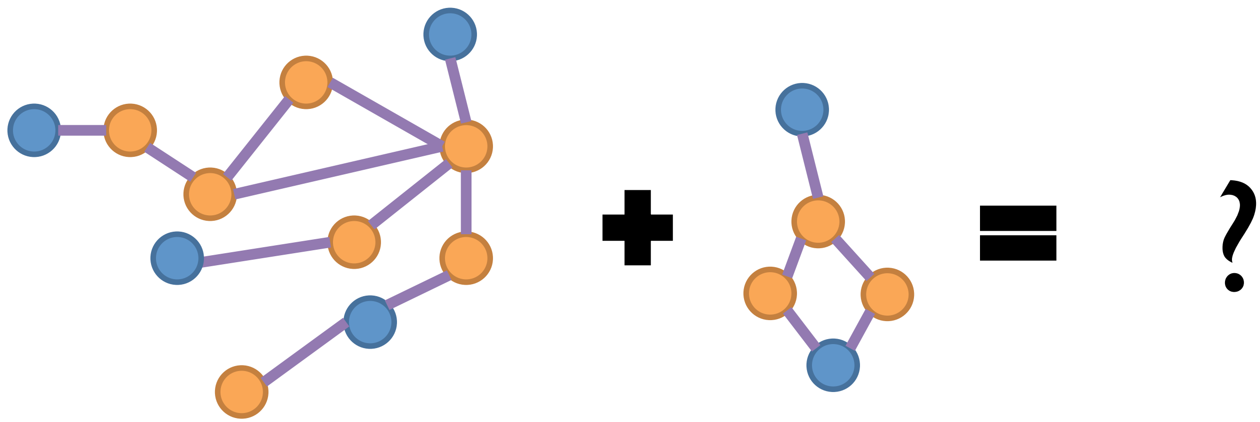
See:

"Learning First-Order Horn Clauses from Web Text" Schoenmackers, Etzioni, Weld, and Davis, EMNLP10

"Toward an Architecture for Never-Ending Language Learning" Carlson, Betteridge, Kisiel, Settles, Hruschka, and Mitchell. AAAI10.

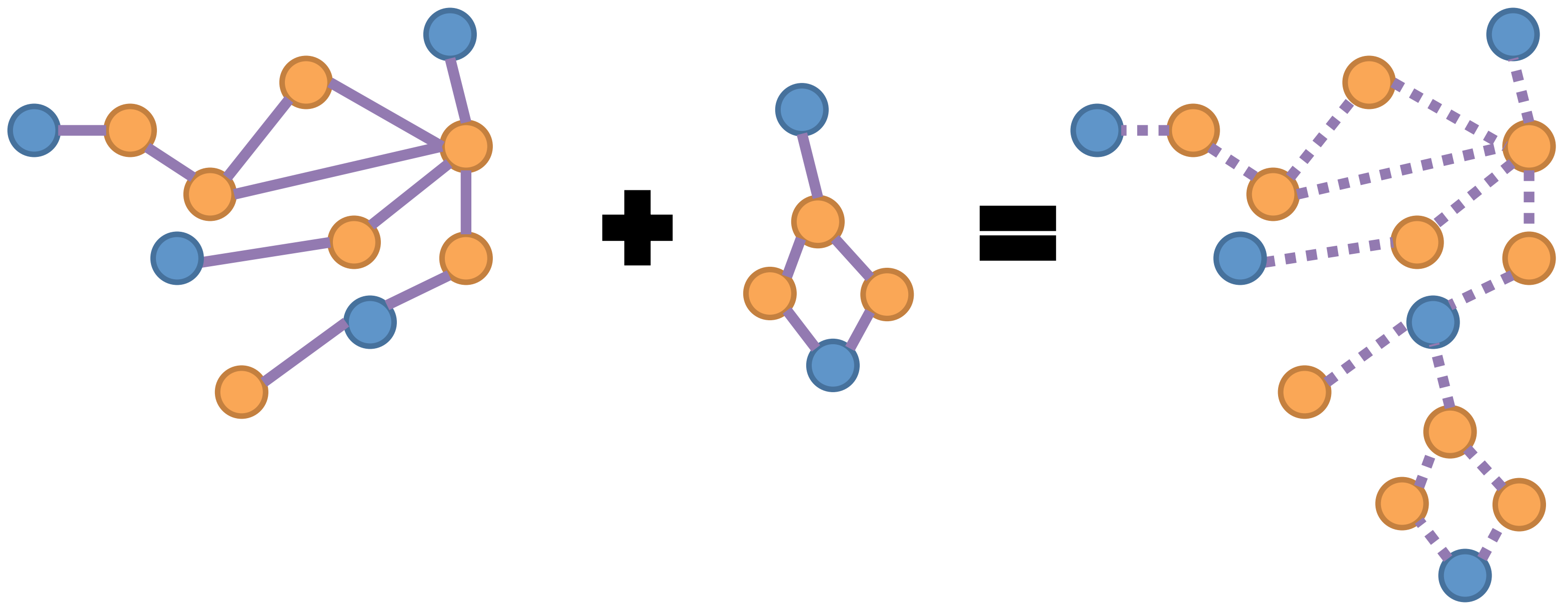
Evolving Models

Problem: Incremental Updates to KG

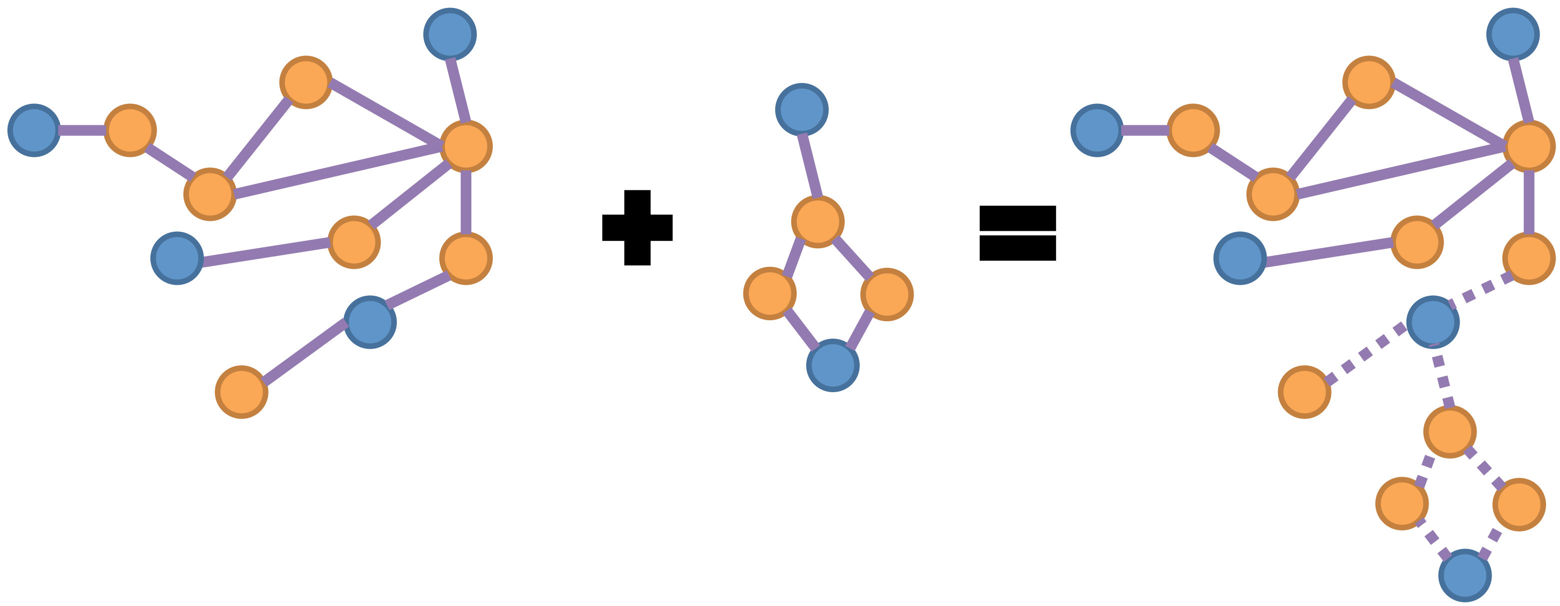


How do we add new extractions to the Knowledge Graph?

Naïve Approach: Full KGI over extractions



Approximation: KGI over subset of graph



Conclusion

- Knowledge Graph Identification is a powerful technique for producing knowledge graphs from noisy IE system output
- Using PSL we are able to enforce global ontological constraints and capture uncertainty in our model
- Unlike previous work, our approach infers complete knowledge graphs for datasets with millions of extractions

Code available on GitHub:

https://github.com/linqs/KnowledgeGraphIdentification