Constructing and Mining Web-scale Knowledge Graphs

Antoine Bordes (Facebook)
abordes@fb.com

Evgeniy Gabrilovich (Google)
gabr@google.com

The opinions expressed herein are the sole responsibility of the tutorial instructors and do not necessarily reflect the opinion of Facebook Inc. or Google Inc.

Technologies described might or might not be in actual use.

Acknowledgements

Thanks to Luna Dong, Matthew Gardner, Ni Lao, Kevin Murphy, Nicolas Usunier, and Jason Weston for fruitful discussions that helped improve this tutorial.

Special thanks to Philip Bohannon and Rahul Gupta for letting us use their slides on entity deduplication and relation extraction.

Outline of the tutorial

PART 1: Knowledge graphs

1. Applications of knowledge graphs
2. Freebase as an example of a large scale knowledge repository
3. Research challenges
4. Knowledge acquisition from text

PART 2: Methods and techniques

1. Relation extraction
2. Entity resolution
3. Link prediction

PART 1: KNOWLEDGE GRAPHS

The role of knowledge

- "Knowledge is Power" Hypothesis (the Knowledge Principle): "If a program is to perform a complex task well, **it must know a great deal about the world** in which it operates."

- The Breadth Hypothesis: "To behave intelligently in unexpected situations, an agent must be capable of falling back on **increasingly general knowledge**."

Lenat & Feigenbaum
Artificial Intelligence 47 (1991)
"On the Threshold of Knowledge"

Why (knowledge) graphs?

- We're surrounded by **entities**, which are connected by **relations**
- We need to store them somehow, e.g., using a **DB** or a **graph**
- **Graphs** can be processed **efficiently** and offer a convenient **abstraction**

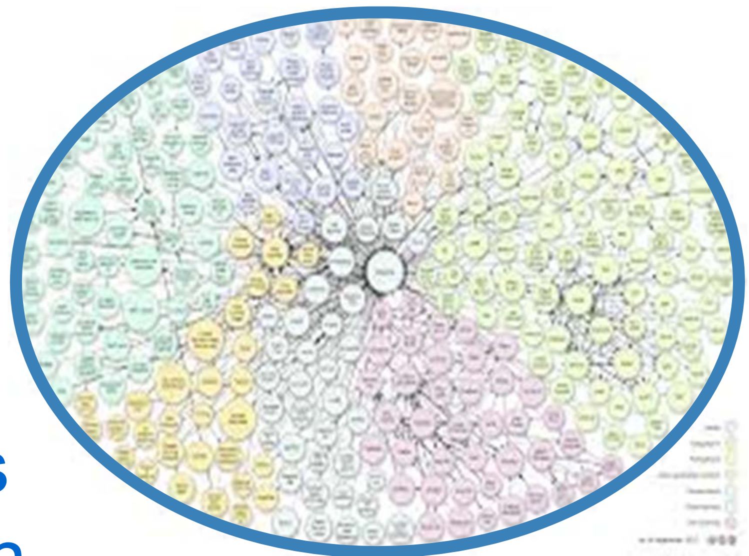
Knowledge graphs

Freebase

yago
select knowledge

NELL: Never-Ending Language Learning

DBpedia



OpenIE
(Reverb, OLLIE)

Facebook's
Entity Graph

Microsoft's
Satori

Google's
Knowledge Graph

A sampler of research problems

- **Growth:** knowledge graphs are incomplete!
  - *Link prediction*: add relations
  - *Ontology matching*: connect graphs
  - *Knowledge extraction*: extract new entities and relations from web/text
- **Validation:** knowledge graphs are not always correct!
  - *Entity resolution*: merge duplicate entities, split wrongly merged ones
  - *Error detection*: remove false assertions
- **Interface:** how to make it easier to access knowledge?
  - *Semantic parsing*: interpret the meaning of queries
  - *Question answering*: compute answers using the knowledge graph
- **Intelligence:** can AI emerge from knowledge graphs?
  - *Automatic reasoning* and planning
  - Generalization and abstraction

A sampler of research problems

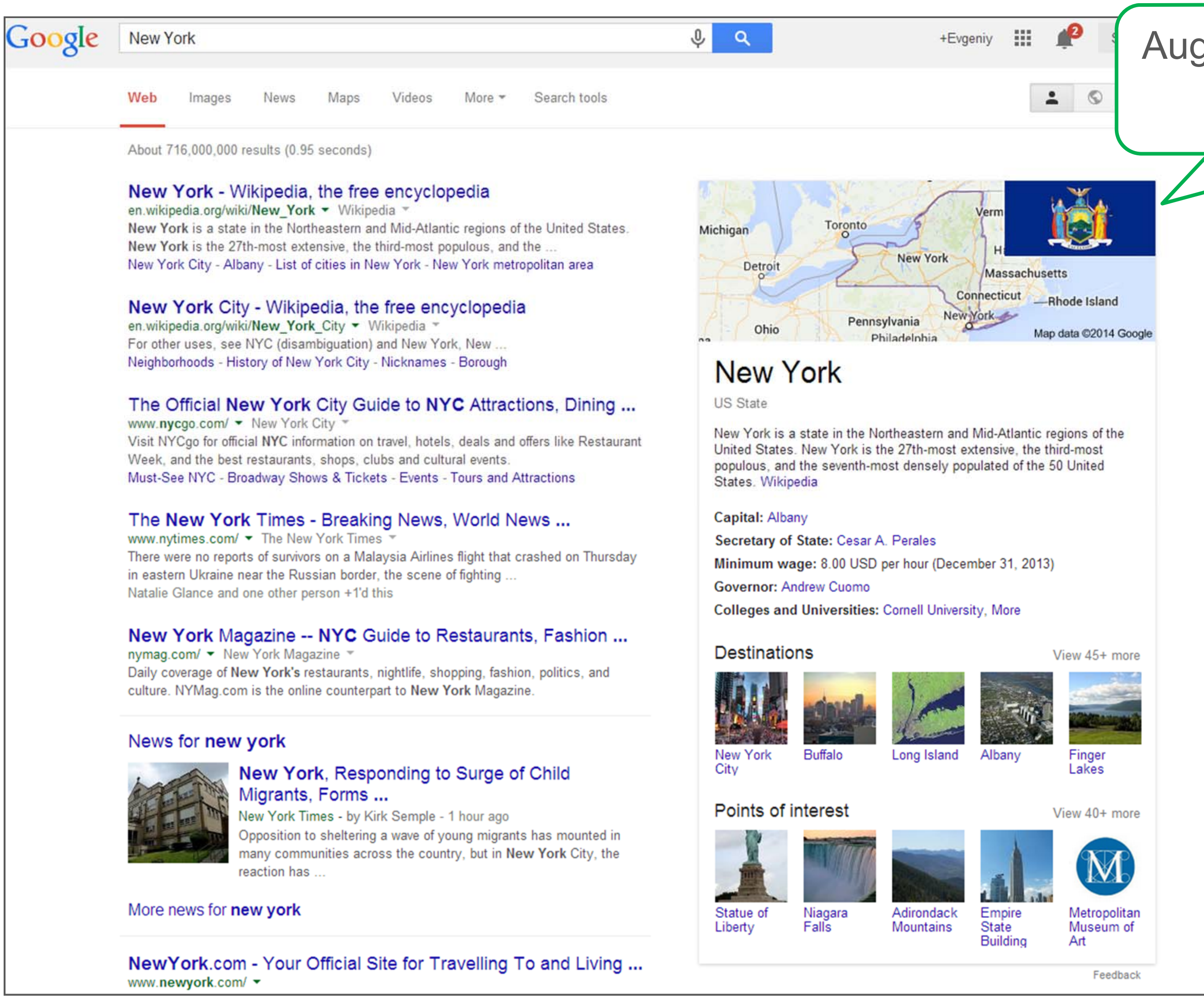
- **Growth:** knowledge graphs are incomplete!
  - *Link prediction*: add relations
  - *Ontology matching*: connect graphs
  - *Knowledge extraction*: extract new entities and relations from web/text
- **Validation:** knowledge graphs are not always correct!
  - *Entity resolution*: merge duplicate entities, split wrongly merged ones
  - *Error detection*: remove false assertions
- **Interface:** how to make it easier to access knowledge?
  - *Semantic parsing*: interpret the meaning of queries
  - *Question answering*: compute answers using the knowledge graph
- **Intelligence:** can AI emerge from knowledge graphs?
  - *Automatic reasoning* and planning
  - Generalization and abstraction

Connections to related fields

- Information retrieval
- Natural language processing
- Databases
- Machine learning
- Artificial intelligence

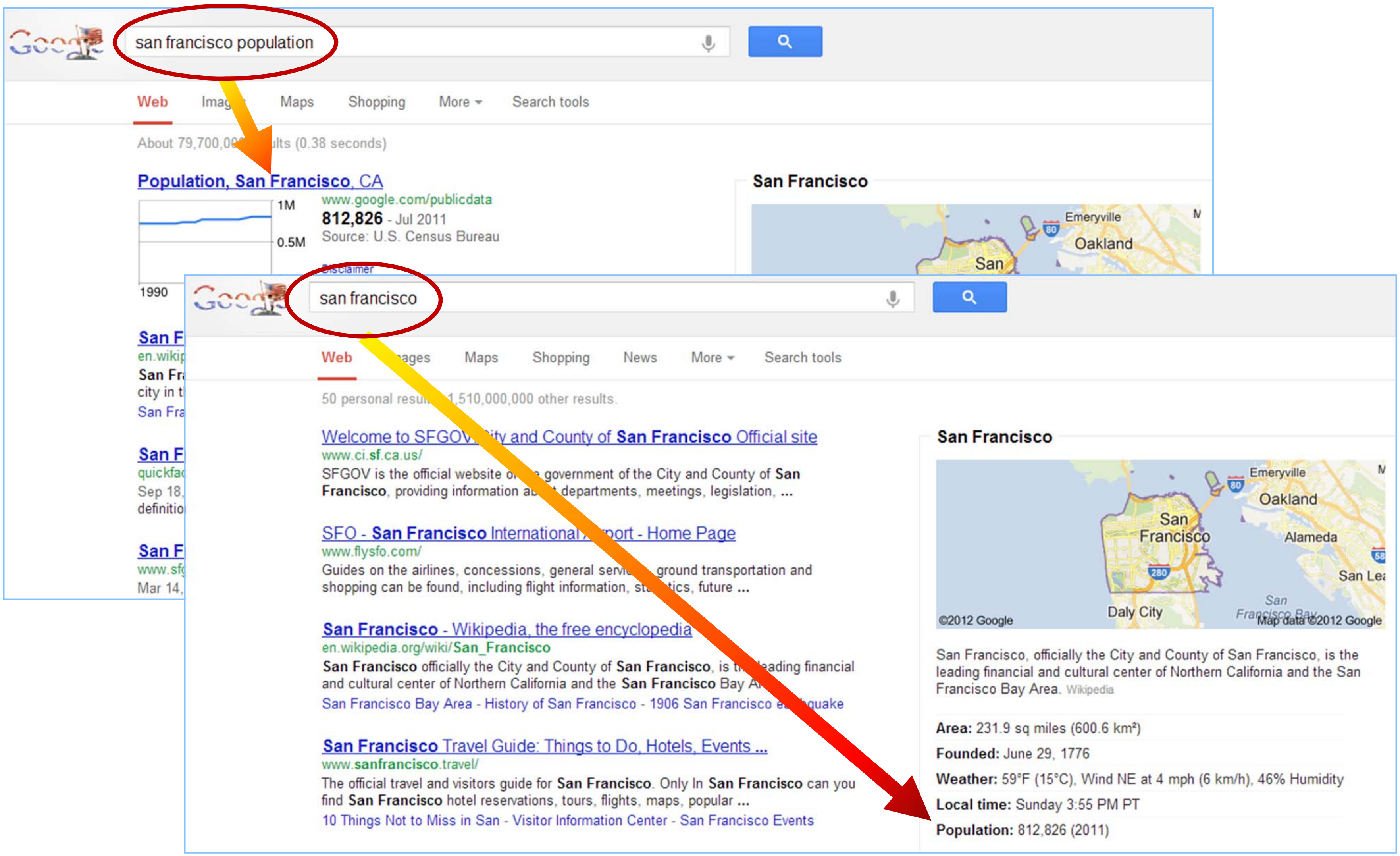
A SAMPLER OF APPLICATIONS OF KNOWLEDGE GRAPHS

Surfacing structured results in web search

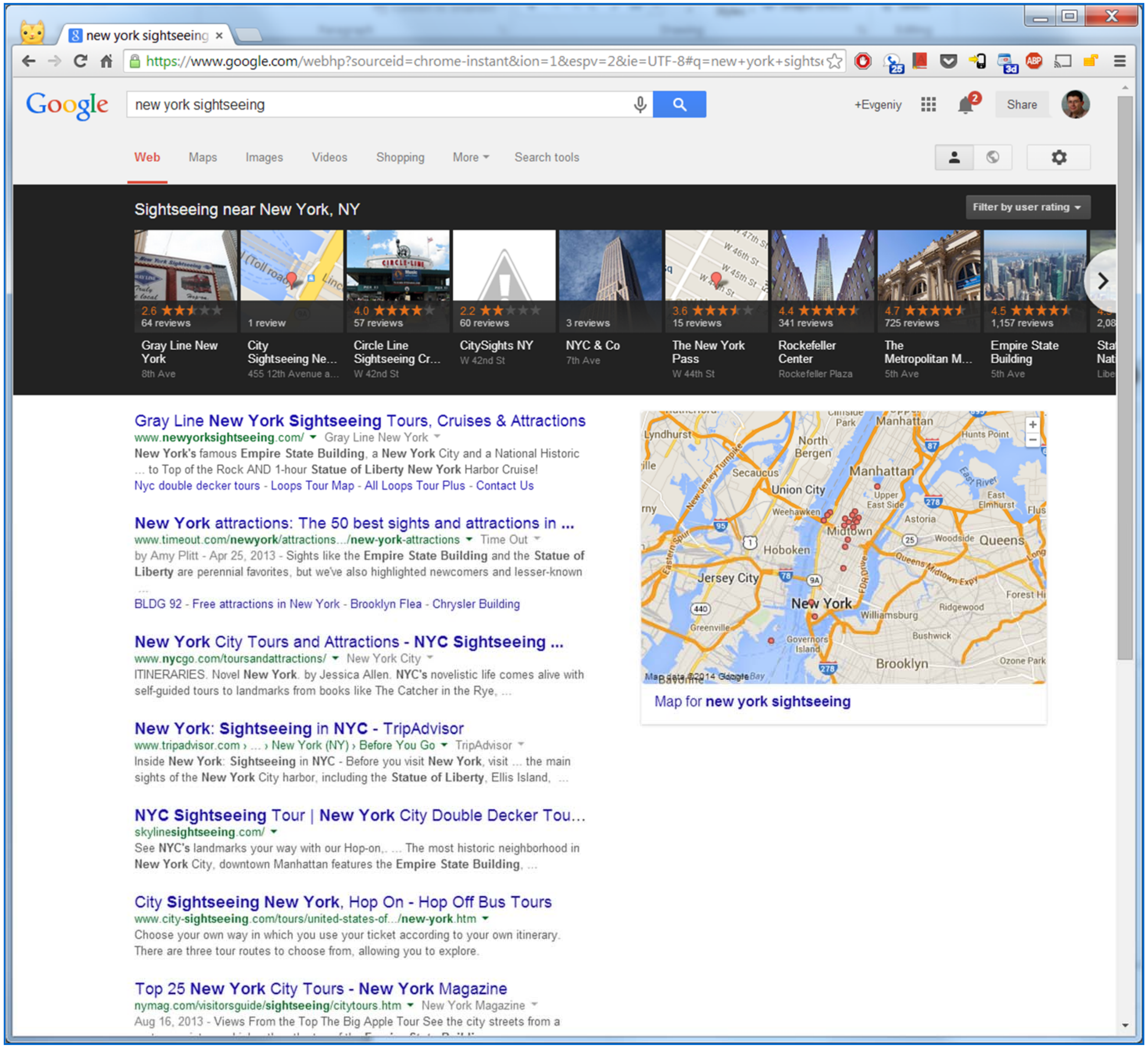


Augmenting the presentation with relevant facts

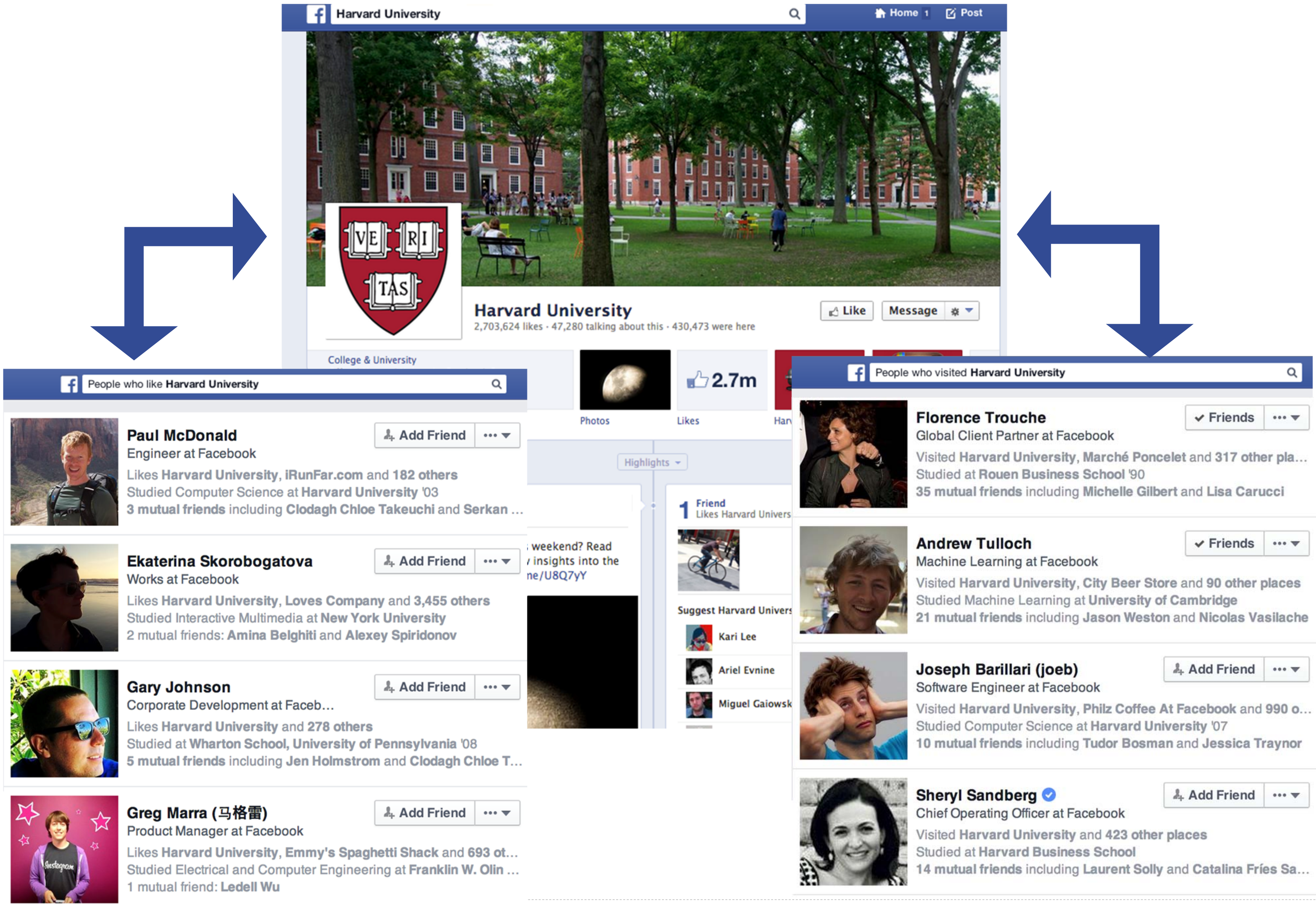
Surfacing facts proactively



Exploratory search



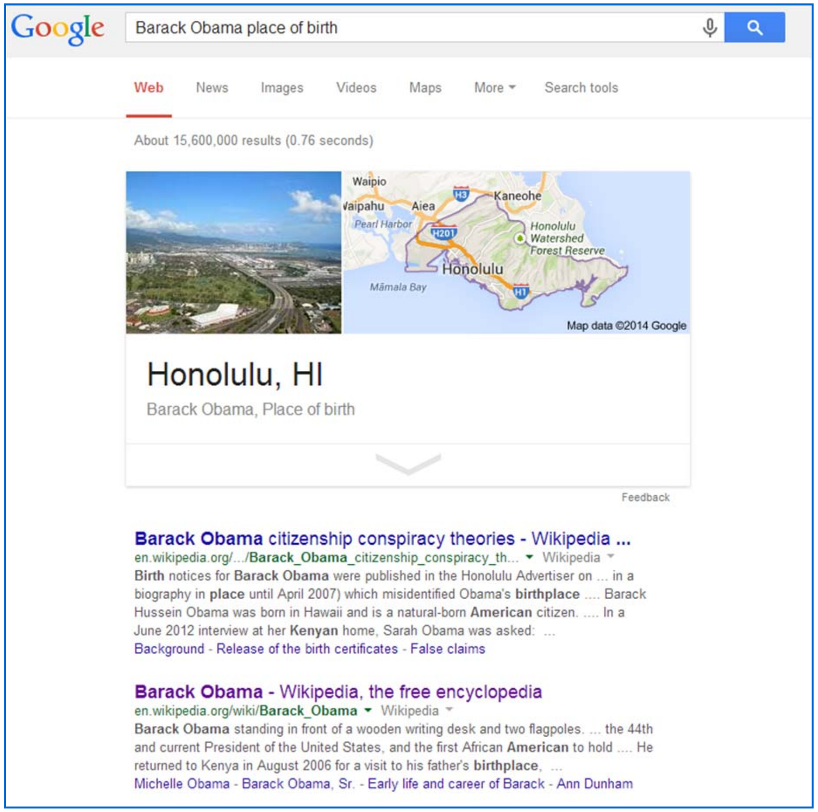
Connecting people, places and things



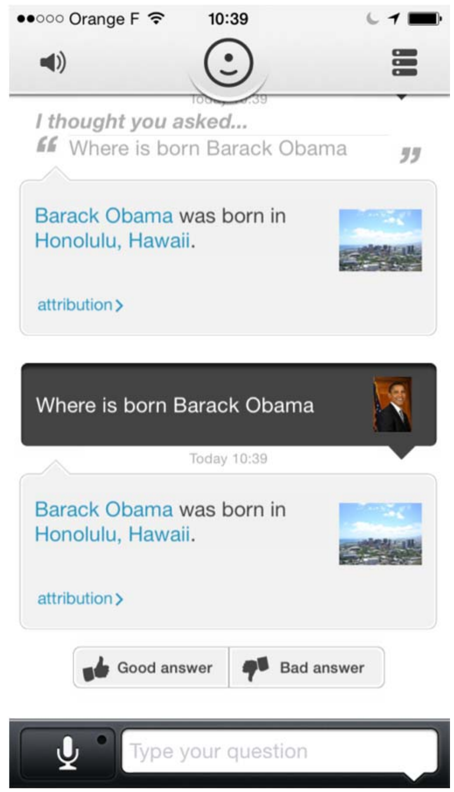
Connecting people, places and things

Structured search within the graph

Question answering



Google



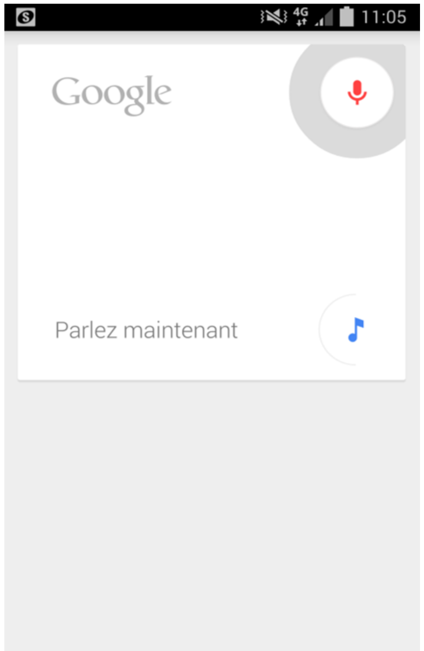
EVI

(Amazon)

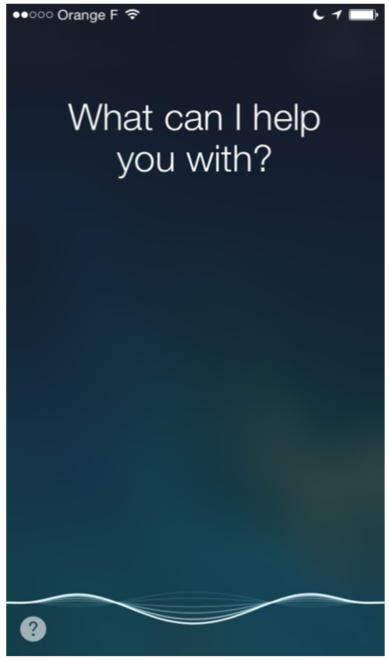
Siri

(Apple)

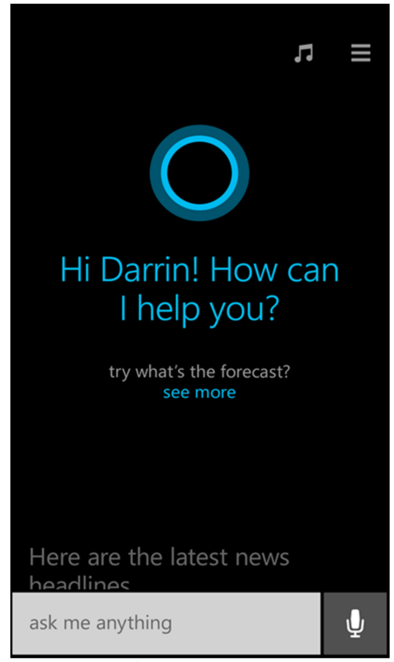
Towards a knowledge-powered digital assistant



OK Google



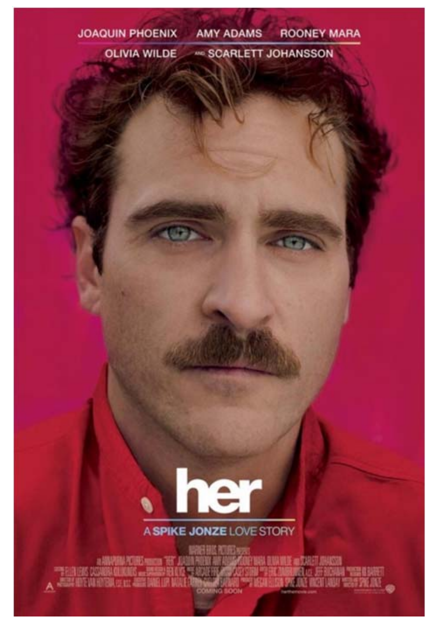
Siri (Apple)



Cortana (Microsoft)

- Natural way of accessing/storing knowledge
- Dialogue system
- Personalization
- Emotion

Interface revolution →



FREEBASE AS AN EXAMPLE OF A LARGE SCALE KNOWLEDGE REPOSITORY

Different approaches to knowledge representation

- Structured (e.g., Freebase or YAGO)
  - Both entities and relations come from a fixed lexicon
- Semi-structured
  - Predicates come from a fixed lexicon, but entities are strings
    - NELL used to be in this category, but is now structured (creating new entities as needed)
- Unstructured (Open IE)

Freebase

- **Freebase** is an open, Creative Commons licensed repository of structured data
- **Typed entities** rather than **strings**

Person Type

Key: /people/person Includes: Topic

A person is a human being (man, woman or child) known to have actually existed. Living persons, celebrities and politicians are persons.

Table Diagram

Properties

Property	ID	Expected Type
Date of birth	/people/person/date_of_birth	/type/datetime
Place of birth	/people/person/place_of_birth	/location/location
Country of nationality	/people/person/nationality	/location/country
Gender	/people/person/gender	/people/gender *enumerated*
Profession	/people/person/profession	/people/profession
Religion	/people/person/religion	/religion/religion
Ethnicity	/people/person/ethnicity	/people/ethnicity
Parents	/people/person/parents	/people/person
Children	/people/person/children	/people/person
Siblings	/people/person/sibling_s	/people/sibling_relationship
Spouse (or domestic partner)	/people/person/spouse_s	/people/marriage *mediatc*
Employment history	/people/person/employment_history	/business/employment_tenu
Education	/people/person/education	/education/education *med*

Relations are typed too!

The world changes, but we don't retract facts

We just add more facts!

Marriage Mediator Type

Key: /people/marriage

'Marriage' defines a relationship between two people. The person type uses it to store the two people in the relationship as well as a beginning and end date

More

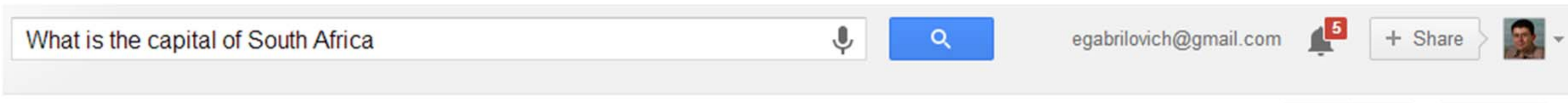
Table Diagram

Properties

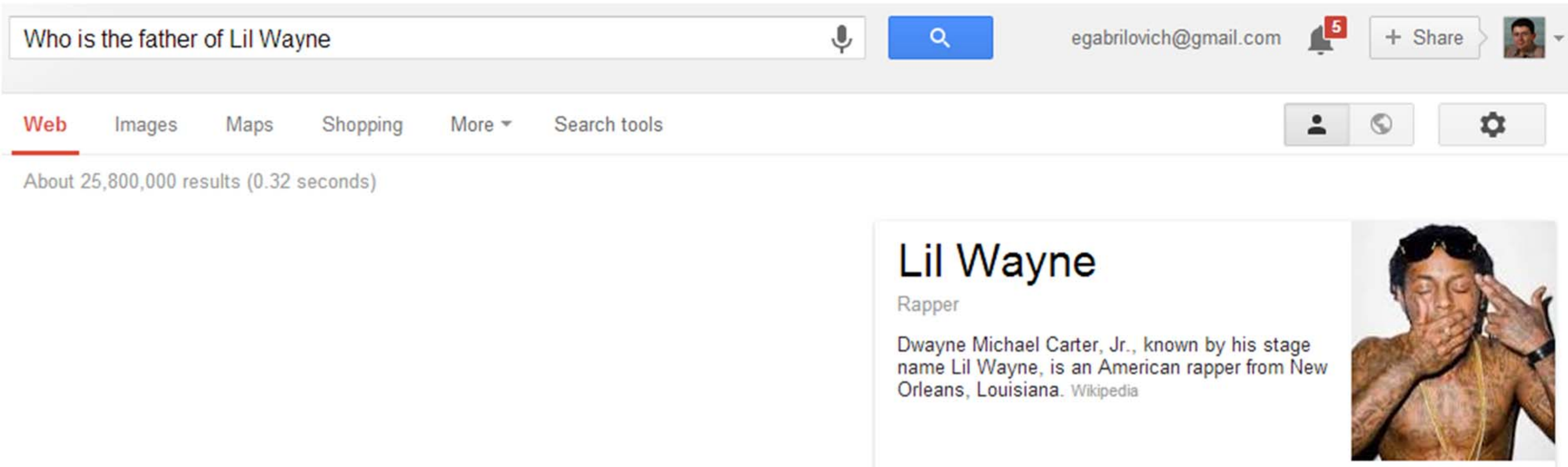
Property	ID	Expected Type
Spouse	/people/marriage/spouse	/people/person
From	/people/marriage/from	/type/datetime
To	/people/marriage/to	/type/datetime
Type of union	/people/marriage/type_of_union	/people/marriage_union_type *enumerated*
Location of ceremony	/people/marriage/location_of_ceremony	/location/location

A graph of inter-related objects

Schema limitations

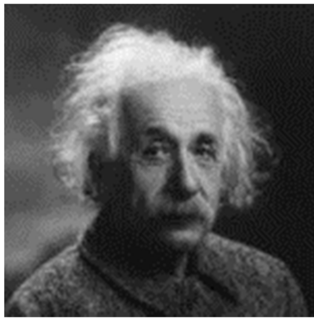


Schema limitations (cont'd)



Subject-Predicate-Object (SPO) triples

</m/0jcx, /m/04m8, /m/019xz9>



/en/albert_einstein

Albert Einstein

/en/ulm

Ulm

/people/person/place_of_birth

Place of birth

YAGO2 uses SPOTL tuples (SPO + Time and Location)

RESEARCH CHALLENGES

Challenging research questions

- How many facts are there ? How many of them can we represent ?
- How much the boundaries of our current knowledge limit what we can learn ?
- How many facts can be potentially extracted from text ?

Limits of automatic extraction

- Freebase: **637M** (non-redundant) facts
- Knowledge Vault (automatically extracted):

  302M confident facts with **Prob(true)> 0.9**
  - Of those, 223M are in Freebase (**~ 35%**)

Relations that are rarely expressed in text

Relation	% entity pairs not found	Notes
/people/person/gender	30%	Pronouns
/people/person/profession	18%	
/people/person/children and /people/person/parents	36%	
/medicine/drug_formulation/ manufactured_forms	99.9%	**Sample object:** "Biaxin 250 film coated tablet" (/m/0jxc5vb)
/medicine/manufactured_drug _form/available_in	99.4%	**Sample subject:** "Fluocinolone Acetonide 0.25 cream" (/m/0jxlbx9)
/book/author/works_written and /book/written_work/author	37%	**Sample book title:** "The birth day: a brief narrative of Eliza Reynolds, who died on Sunday, Oct 19, 1834" (/m/0ydpbtq)

Relations that are rarely expressed in text

Relation
/people/person/gender
/people/person/profession
/people/person/children and /people/person/parents
/medicine/drug_formulation/ manufactured_forms
/medicine/manufactured_drug _form/available_in
/book/author/works_written and /book/written_work/author

Albert Einstein College of Medicine
OF YESHIVA UNIVERSITY

Communications & Public Affairs
Newsroom
News Releases
Social Media Hub
Einstein in the Media
Features
Multimedia
Publications

Home > Newsroom > News Releases > Connectomics:

Connectomics

Connectomics: Mapping the Neural Network Governing Male Roundworm Mating

print | subscribe

July 26, 2012 – (BRONX, NY) – In a study published today online in *Science*, researchers at Albert Einstein College of Medicine of Yeshiva University have determined the complete wiring diagram for the part of the nervous system controlling mating in the male roundworm *Caenorhabditis elegans*, an animal model intensively studied by scientists worldwide.

Scott Emmons, Ph.D.

The study represents a major contribution to the new field of connectomics – the effort to map the myriad neural connections in a brain, brain region or nervous system to find the specific nerve connections responsible for particular behaviors. A long-term goal of connectomics is to map the human "connectome" – all the nerve connections within the human brain.

Because *C. elegans* is such a tiny animal – adults are one millimeter long and consist of just 959 cells – its simple nervous system totaling 302 neurons make it one of the best animal models for understanding the millions-of-times-more-complex human brain.

The Einstein scientists solved the structure of the male worm's neural mating circuits by developing software that they used to analyze serial electron micrographs that other scientists had taken of the region. They found that male mating requires 144 neurons – nearly half the worm's total number – and their paper describes the connections between those 144 neurons and 64 muscles involving some 8,000 synapses. A synapse is the junction at which one neuron (nerve cell) passes an electrical or chemical signal to another neuron.

Relations that are rarely expressed in text

Relation	% entity pairs not found	Notes
/people/person/gender	30%	Pronouns
/people/person/profession	18%	
/people/person/children and /people/person/parents	36%	
/medicine/drug_formulation/ manufactured_forms	99.9%	**Sample object:** "Biaxin 250 film coated tablet" (/m/0jxc5vb)
/medicine/manufactured_drug _form/available_in	99.4%	**Sample subject:** “Fluocinolone Acetonide 0.25 cream” (/m/0jxlbx9)
/book/author/works_written and /book/written_work/author	37%	**Sample book title:** “The birth day: a brief narrative of Eliza Reynolds, who died on Sunday, Oct 19, 1834” (/m/0ydpbtq)

Implicitly stated information

People /people

Person /people/person

Date of birth /people/person/date_of_birth

4004 BCE

Place of birth /people/person/place_of_birth

Garden of Eden

Country of nationality /people/person/nationality

This property has been flagged as having values, but those values are unknown. Remove this flag to add specific values.

Gender /people/person/gender

Male

Profession /people/person/profession

-

Religion /people/person/religion

-

Ethnicity /people/person/ethnicity

-

Parents /people/person/parents

This property has been flagged as having no values. Remove this flag to add new values.

Children /people/person/children

Children

Azura

Seth

Cain

Abel

Awan

Siblings /people/person/sibling_s

Sibling

This property has been flagged as having no values. Remove this flag to add new values.

Spouse (or domestic partner) /peop

Spouse **From**

Eve

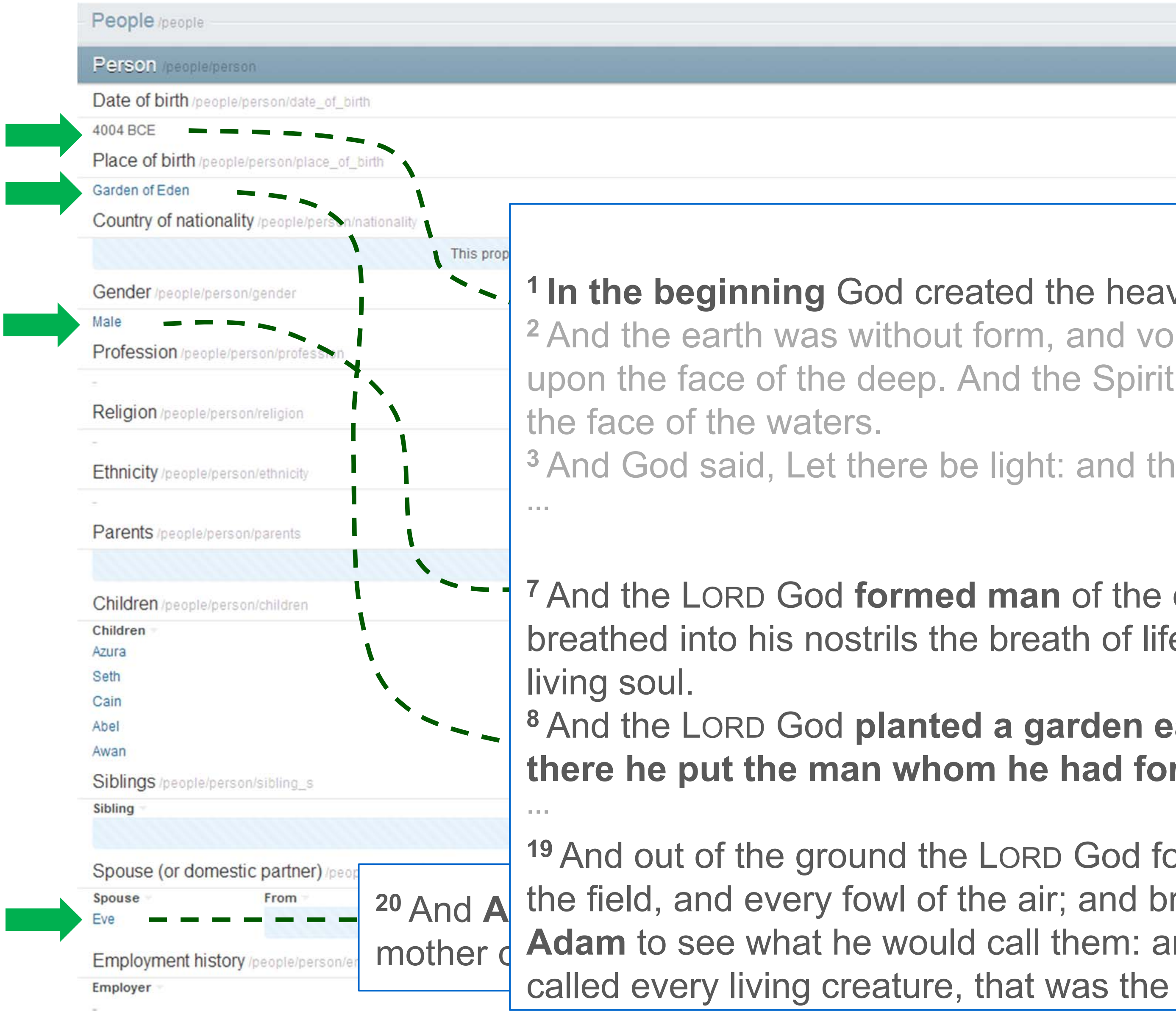
Employment history /people/person/e

Employer

-

[20] And **Adam called his wife's name Eve**; because she was the mother of all living. ***(Genesis 3:20)***

Implicitly stated information



20 And A
mother c

(Genesis 1)

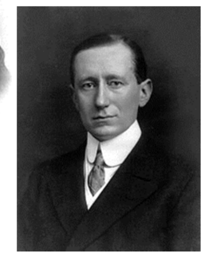
1 **In the beginning** God created the heaven and the earth.
2 And the earth was without form, and void; and darkness was
upon the face of the deep. And the Spirit of God moved upon
the face of the waters.
3 And God said, Let there be light: and there was light.
…

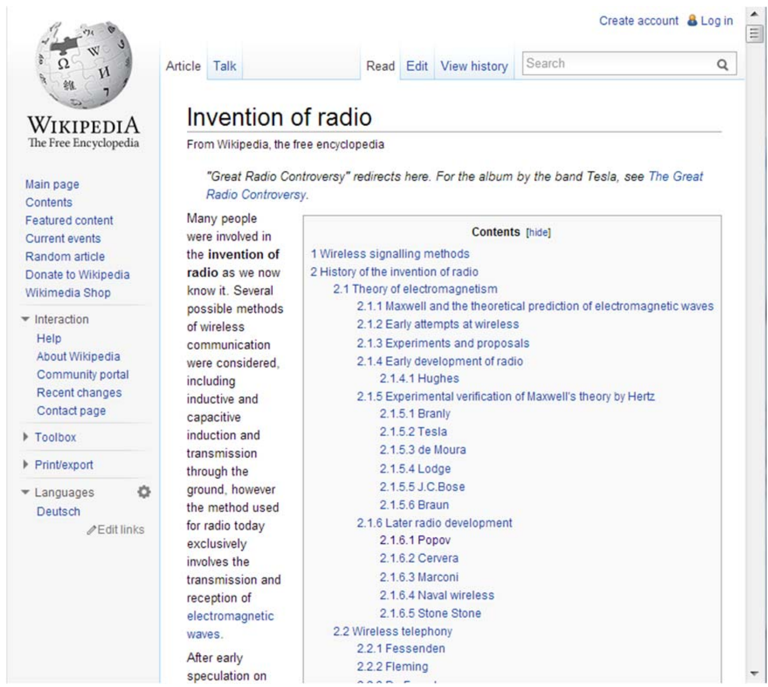
(Genesis 2)

7 And the LORD God **formed man** of the dust of the ground, and
breathed into his nostrils the breath of life; and man became a
living soul.
8 And the LORD God **planted a garden eastward in Eden; and**
there he put the man whom he had formed.
…
19 And out of the ground the LORD God formed every beast of
the field, and every fowl of the air; and brought them unto
Adam to see what he would call them: and whatsoever **Adam**
called every living creature, that was the name thereof.

Knowledge discovery: the long tail of challenges

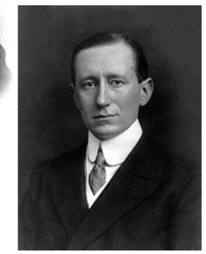
- Errors in extraction (e.g., parsing errors, overly general patterns)
- Noisy / unreliable / conflicting information
- Disparity of opinion (*Who invented the radio ?*)
- Quantifying completeness of coverage

Knowledge discovery: the long tail of challenges

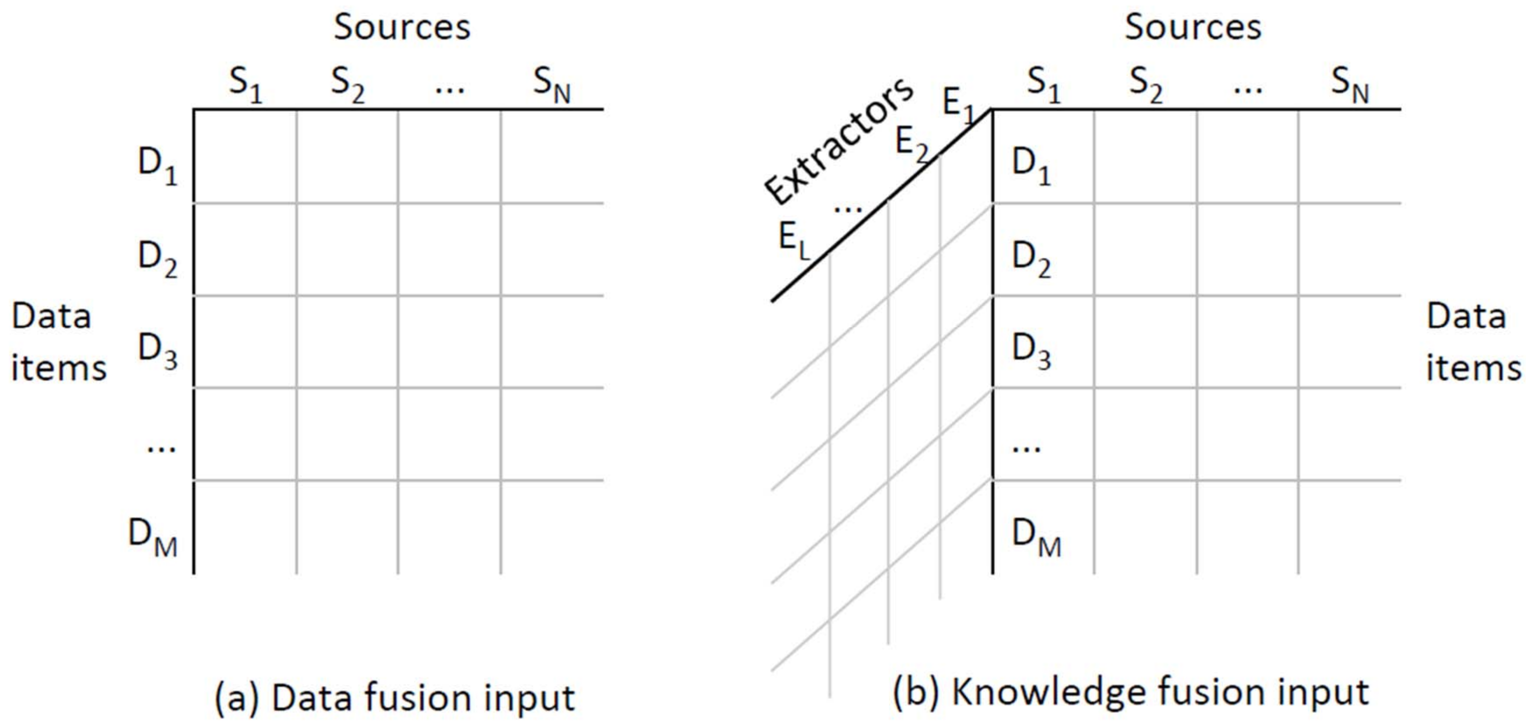
- Errors in extraction (e.g., parsing errors, overly general patterns)
- Noisy / unreliable / conflicting information
- Disparity of opinion (*Who invented the radio ?*)
- Quantifying completeness of coverage


- Fictional contexts
  - **</en/abraham_lincoln,**
    /people/person/profession,
    /en/vampire_hunter> ?

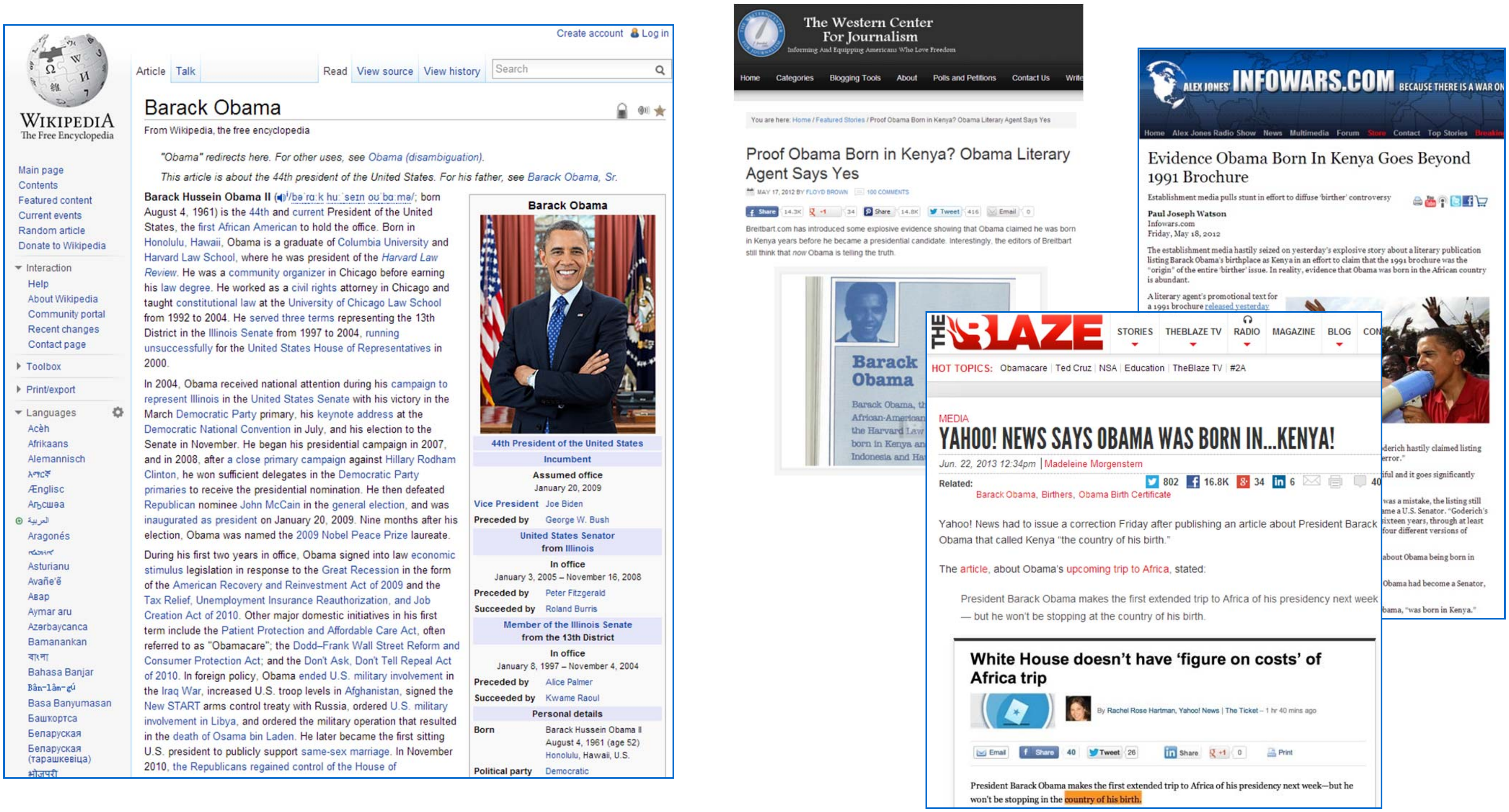
- Outright spam

Data fusion vs. knowledge fusion



[Dong et al., VLDB '14]

Should we trust all sources equally ?



Challenge: negative examples

- We already know a lot … but those are only **positive** examples!
- Many ways to get negative examples … none of them perfect ☹
  - Deleted assertions in Freebase
    - *Was the deletion justified ?*

Released ! See goo.gl/MJb3A

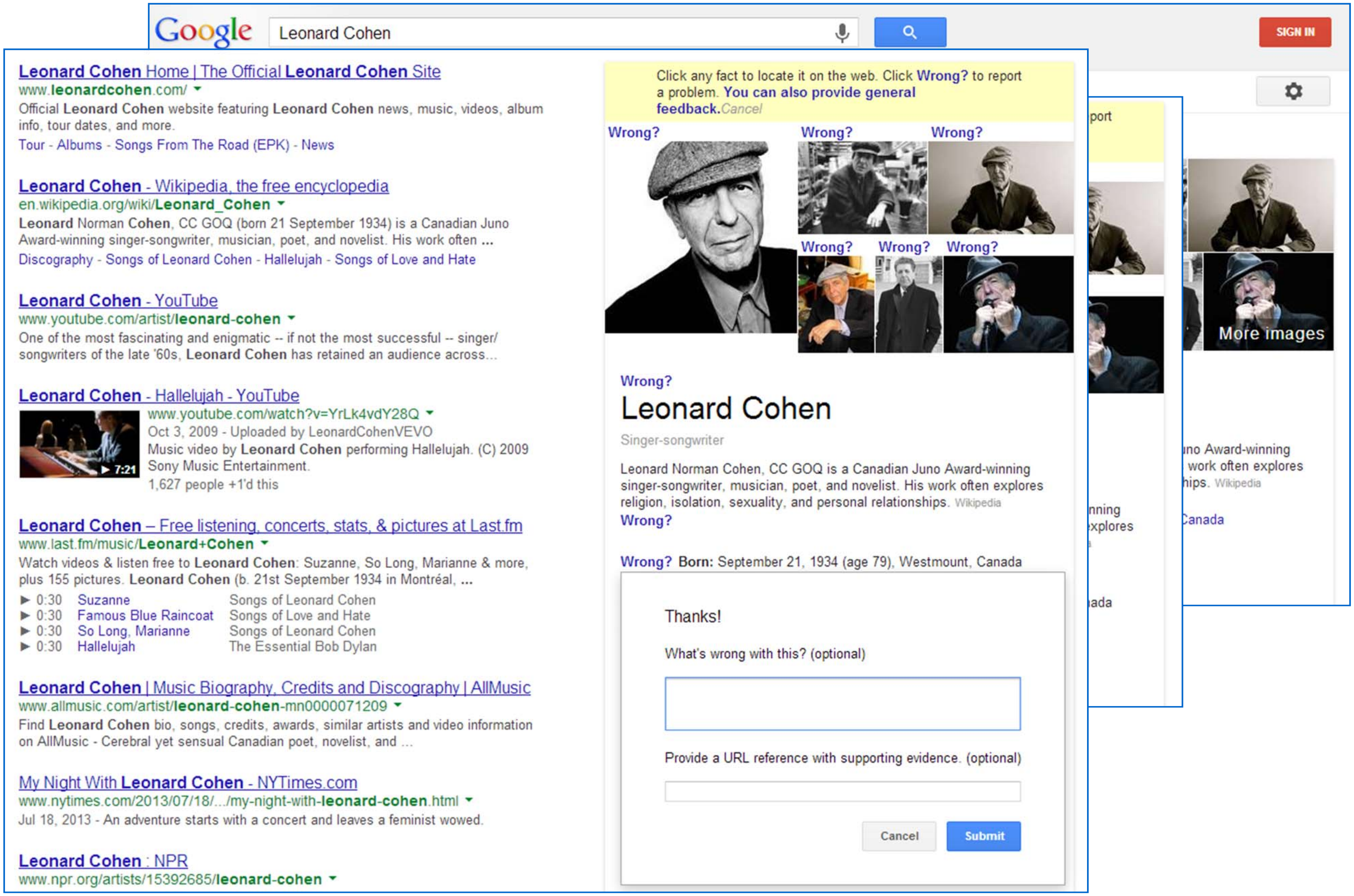
  - Inconsistencies identified with manually specified rules
    - *Poor coverage*
  - Examples judged by humans
    - *Optimized for accuracy on the positive class*

Released ! See goo.gl/MJb3A

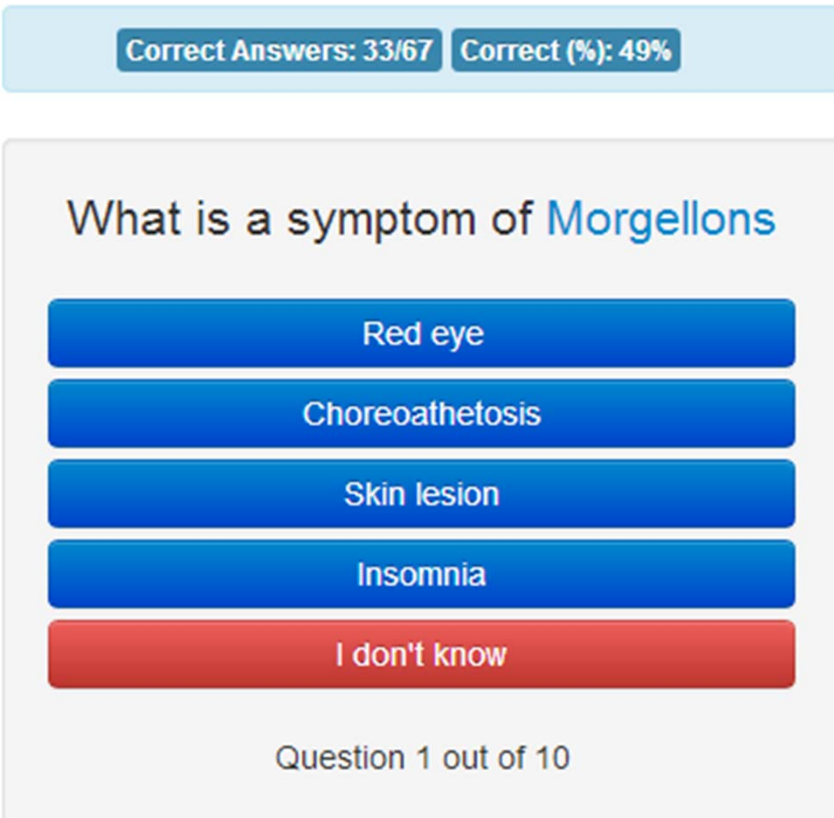
  - Automatically create negative examples using the closed world assumption
    - *Noisy, unless applied to functional relations*
  - Feedback from Web users
    - *Difficult to judge automatically*

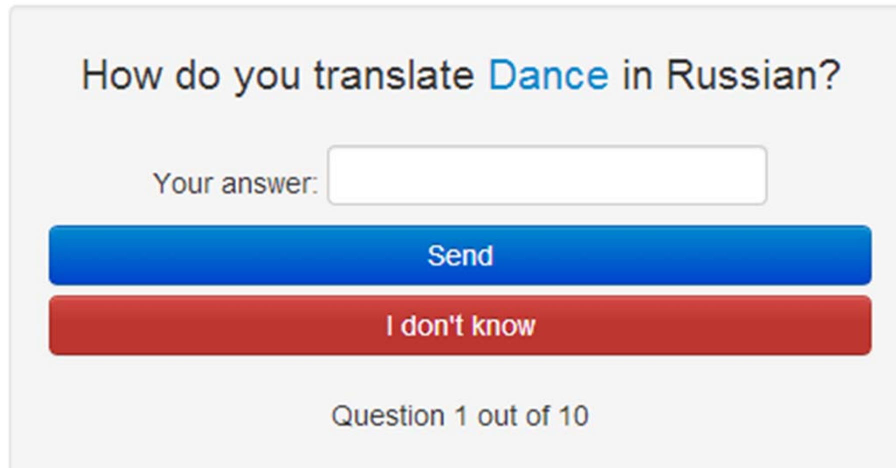
Crowdsourcing

Negative examples (cont'd): feedback from Web users



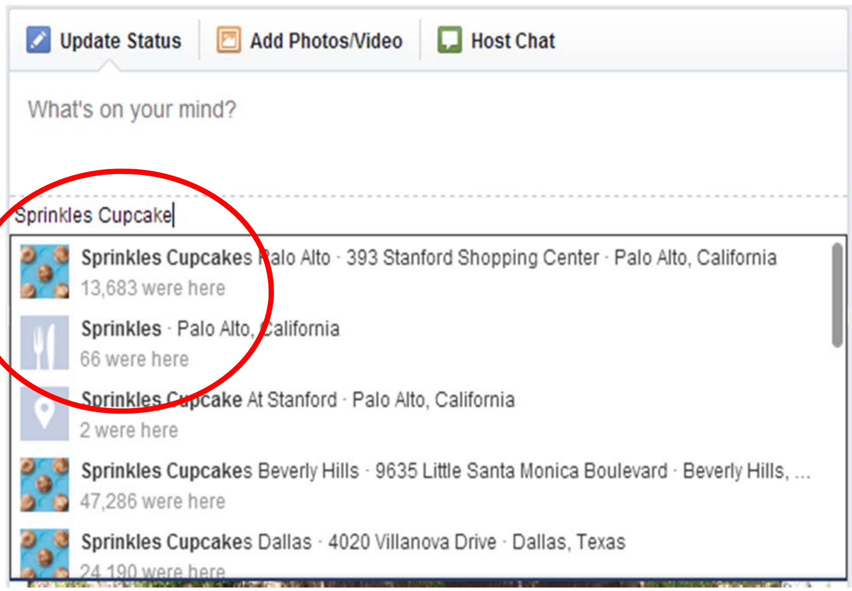
Quizz.us [Ipeirotis & Gabrilovich, WWW 2014]





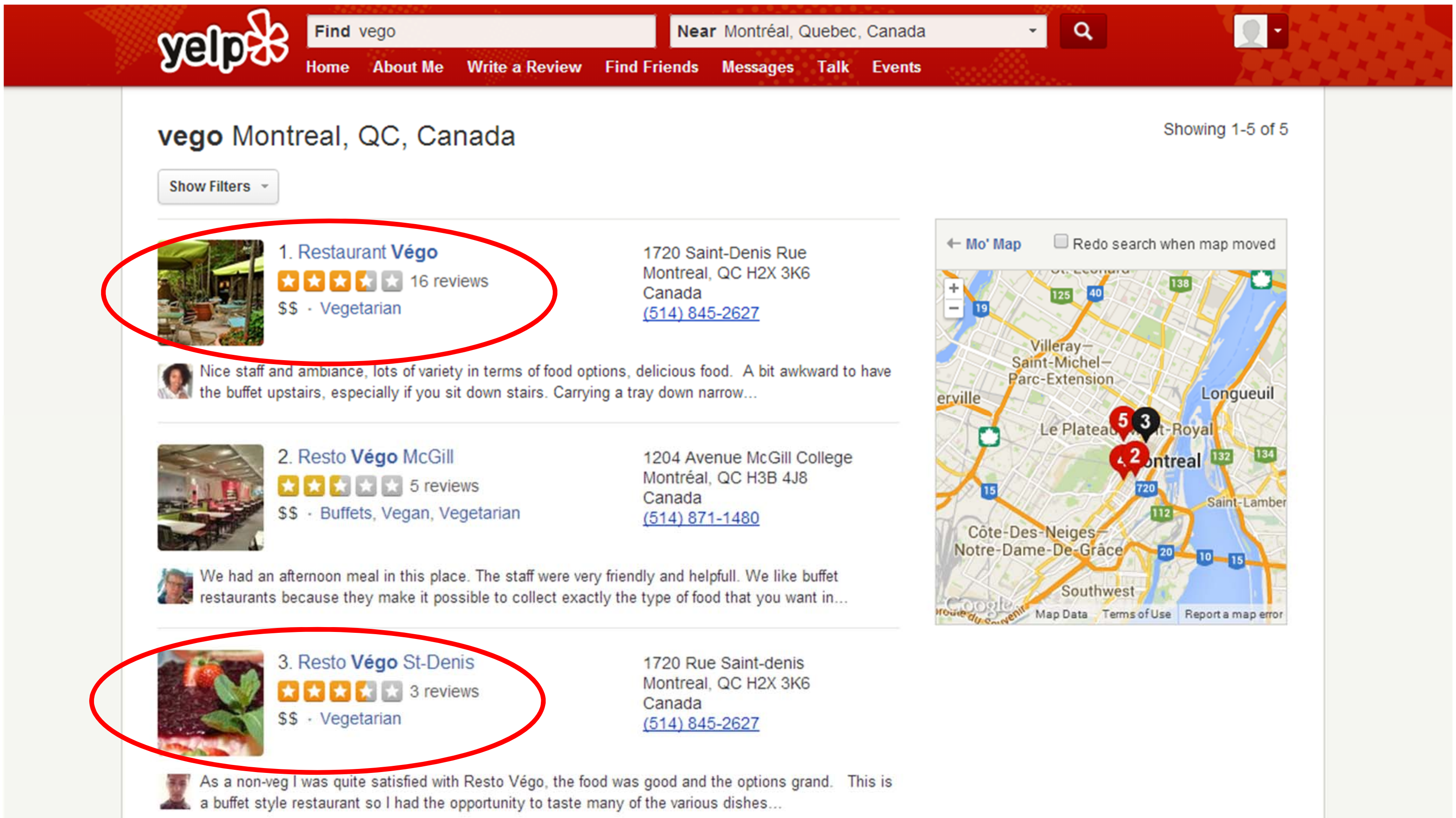
Entity resolution / deduplication

- Multiple mentions of the same entity is wrong and confusing.



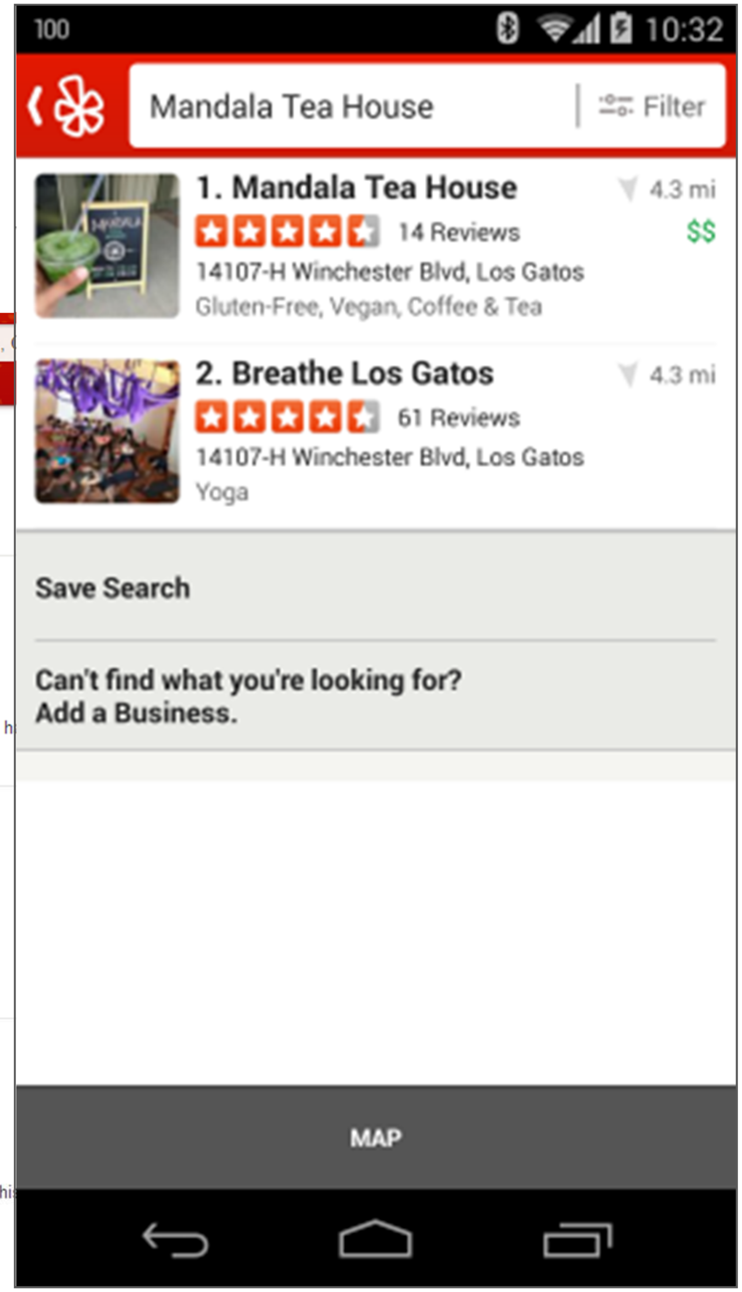
Entity resolution / deduplication

- Multiple mentions of the same entity is wrong and confusing.



Entity resolution / deduplication

- Multiple mentions of the same entity is

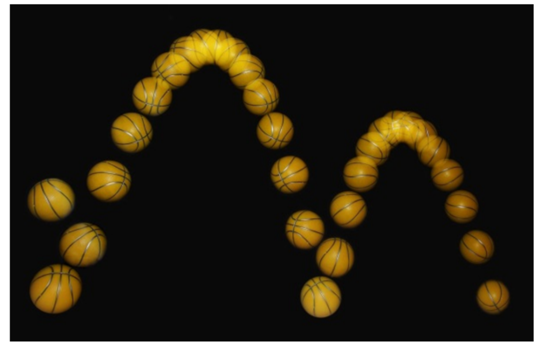


Commonsense knowledge

“Bananas are yellow.”

“Jasmine flowers smell good.”

“Balls bounce.”

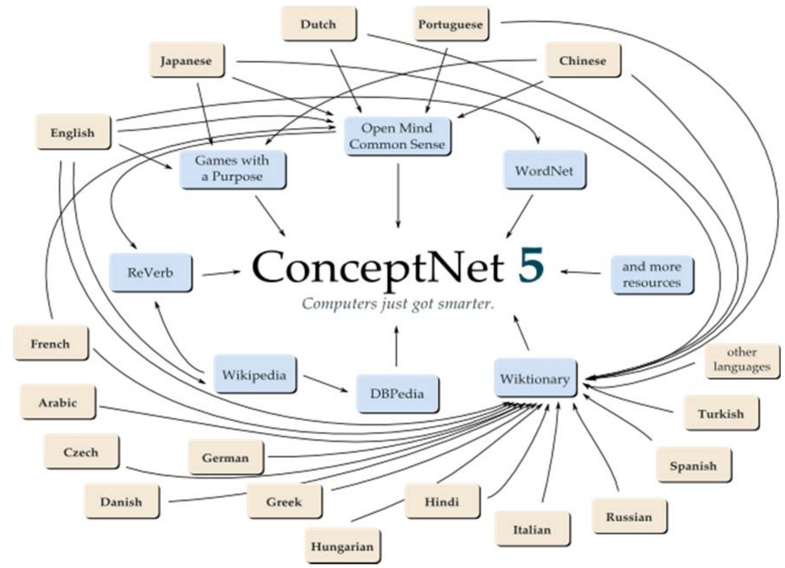


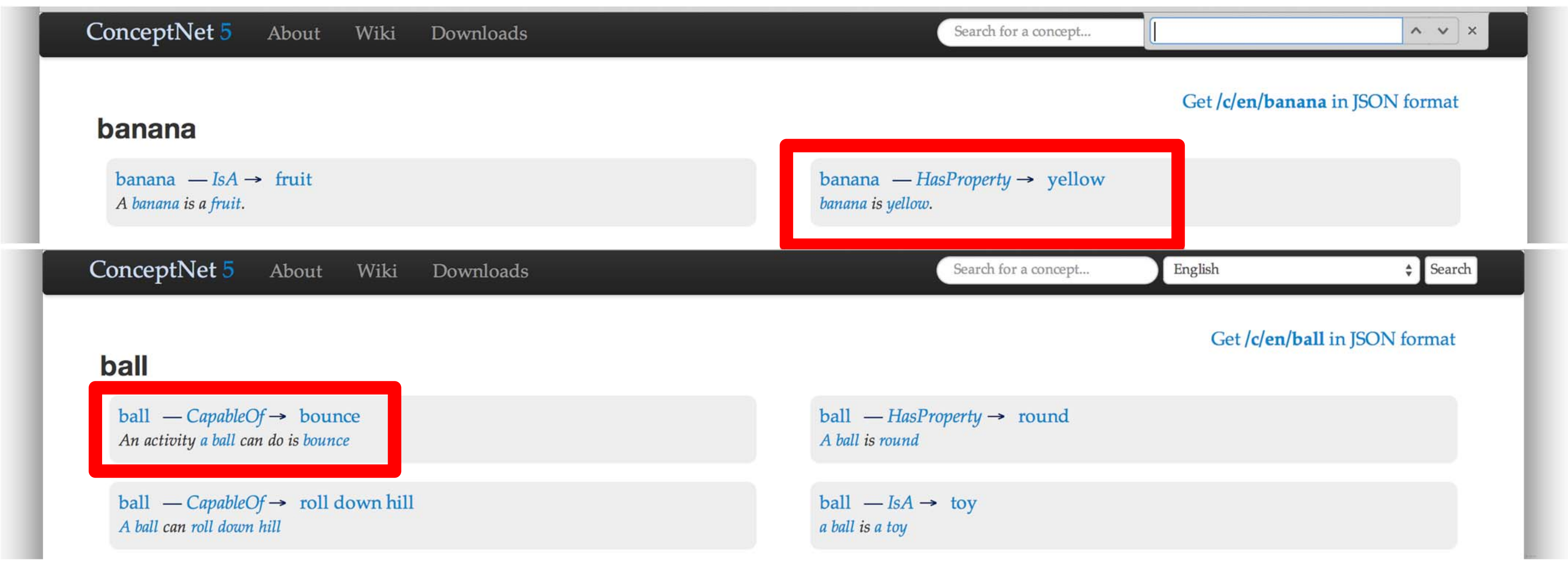
- Commonsense information is hard to collect (*too obvious*)
- Yet commonsense reasoning is often crucial

Commonsense knowledge

ConceptNet

- Nodes represent concepts (words or short NL phrases)
- Labeled relationships connecting them
  `saxophone → UsedFor → jazz`
  `learn → MotivatedByGoal → knowledge`





ConceptNet (cont'd)

- ConceptNet is a **(hyper)graph**
  - Edges about the edges
- Each statement has **justifications**
  - Provenance + reliability assessment
- The graph is **ID-less**
  - Every node has all the information necessary to identify it
  - Multiple branches can be developed in parallel and later merged
    - Take the union of the nodes and edges
    - No reconciliation

[Havasi et al., RANLP '07; Speer and Havasi, LREC '12]

http://conceptnet5.media.mit.edu/

Commonsense knowledge in YAGO

- WebChild [Tandon et al., WSDM '14]

  `(strawberry, hasTaste, sweet),(apple, hasColor, green)`

  - Acquired from the web using semi-supervised learning
  - Uses WordNet senses and web statistics to construct seeds

- Acquiring comparative commonsense knowledge from the web [Tandon et al., AAAI '14]

  `(car, faster, bike), (lemon, more-sour, apple)`

  - Uses Open IE

- Earlier work: [Tandon et al., AAAI '11]

  `CapableOf(dog, bark), PartOf(roof, house)`

  - Uses web n-gram data with seeds from ConceptNet

CYC

[Guha et al., CACM ’90] + http://www.cyc.com/publications

- OpenCYC
  - 239K terms, 2M triples
- ResearchCYC
  - 500K concepts, 5M assertions, 26K relations

Multiple modalities

Video

Text

How to jointly acquire knowledge from all these sources?

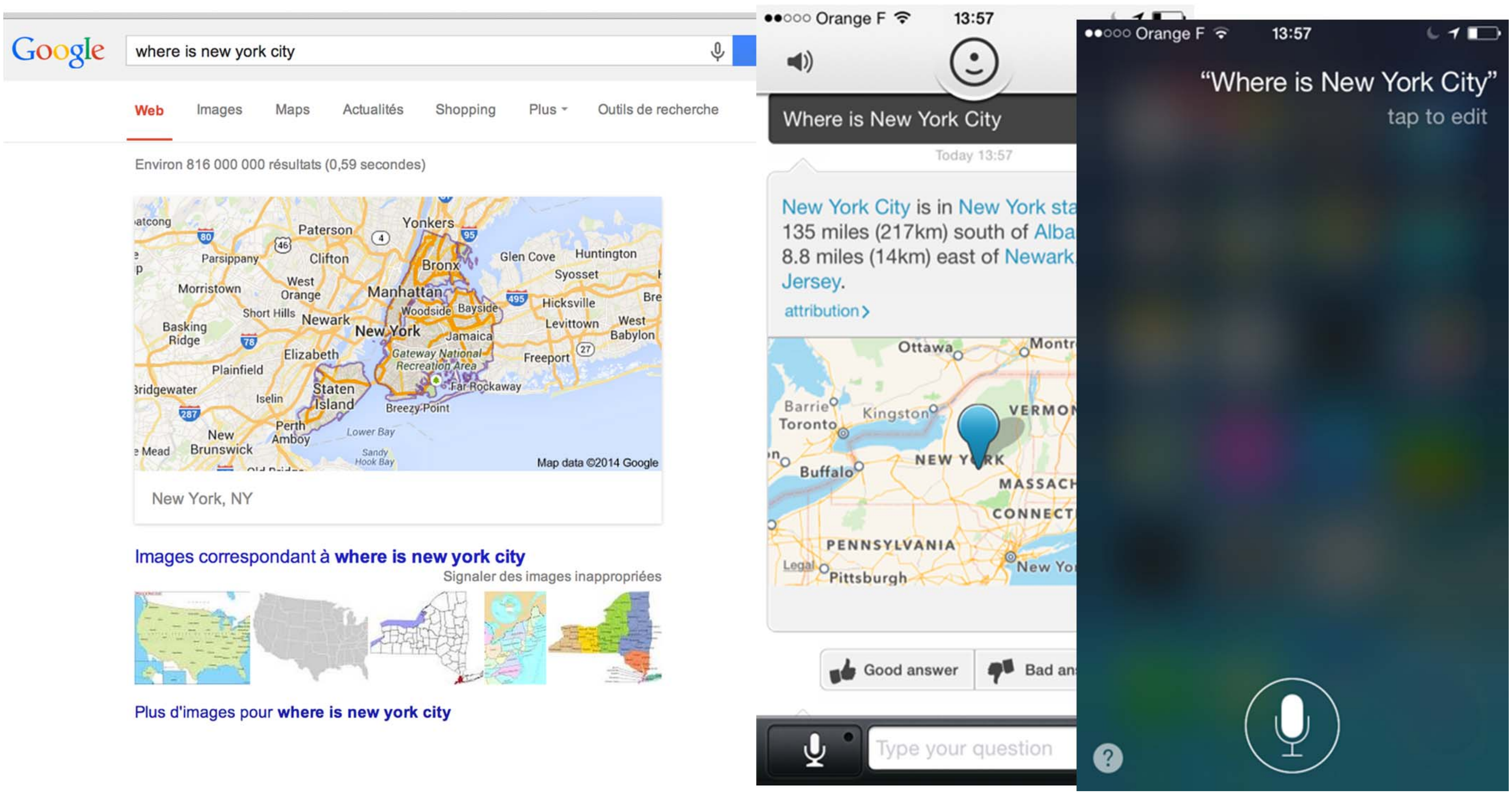
Images

Speech/sounds

Artificial worlds?

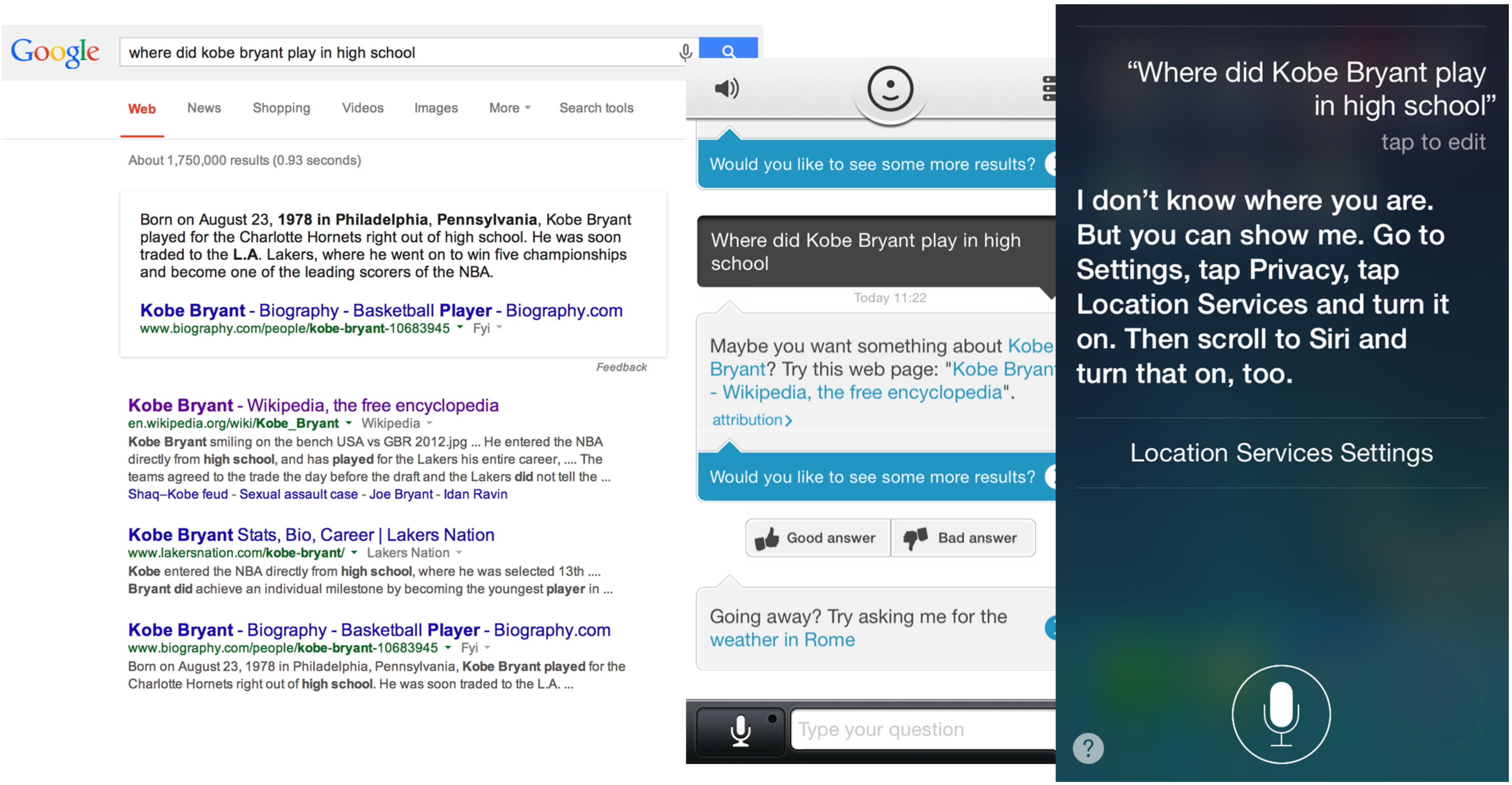
Natural interfaces to knowledge

"Where is New York City?"



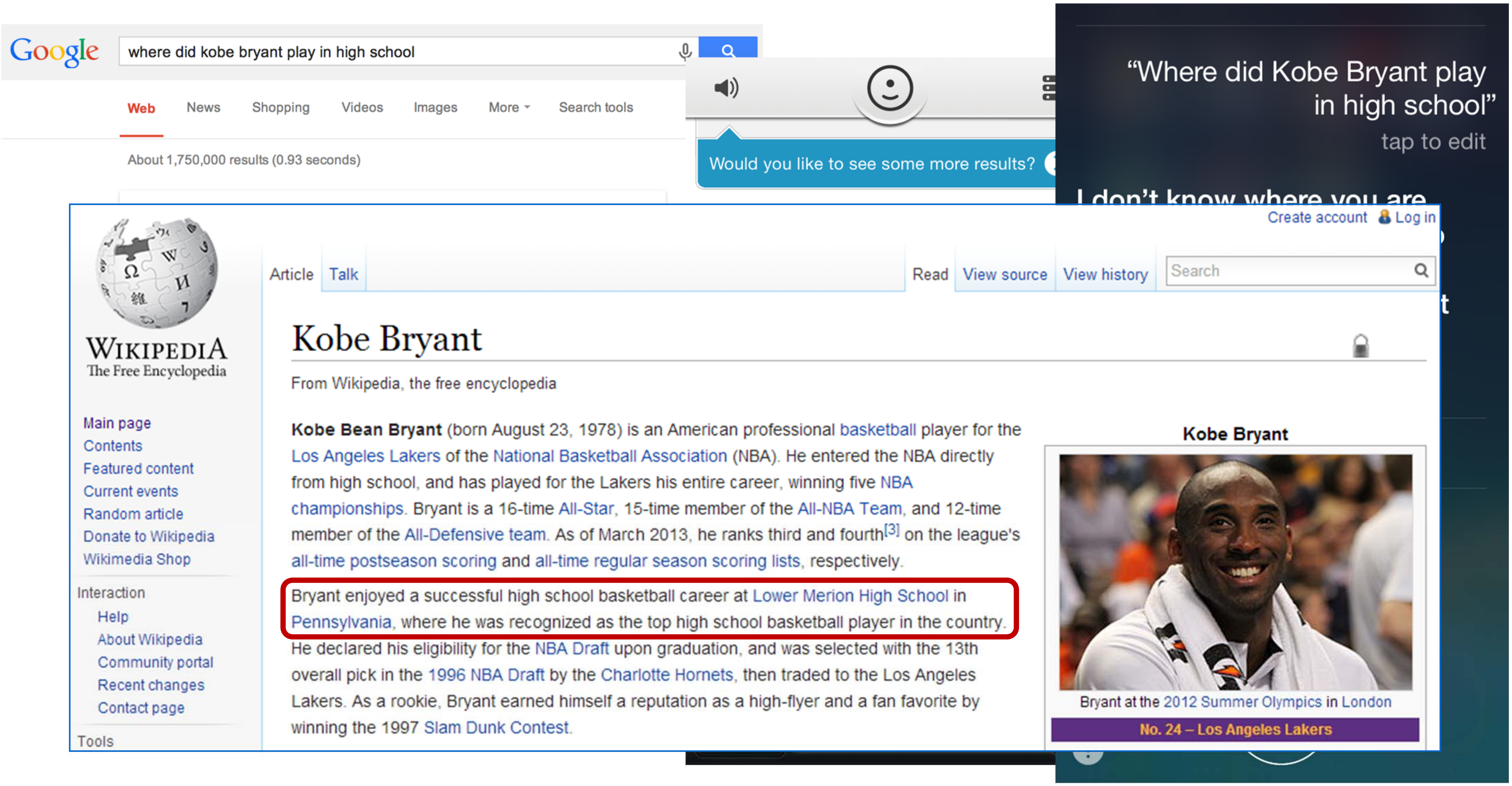
Natural interfaces to knowledge

"Where did Kobe Bryant play in high school?"



Natural interfaces to knowledge

"Where did Kobe Bryant play in high school?"



KNOWLEDGE ACQUISITION FROM TEXT

External sources of knowledge

- Text
  - Unstructured (NL text) or semi-structured (tables or pages with regular structure)
  - Relevant tasks: **entity linking, relation extraction**
- Structured knowledge bases (e.g., IMDB)
  - Relevant task: **entity resolution**

Possible approaches to knowledge acquisition from the Web

- **Unfocused**
  - Start from a collection of Web pages

  ➔ Non-targeted (blanket) extraction

- **Focused**
  - Formulate specific questions or queries, looking for missing data
  - Identify (a small set of) relevant Web pages

  ➔ Targeted extraction

Open IE – extracting **unstructured** facts from **unstructured** sources (text)

- **TextRunner** [Banko et al., IJCAI ’07], **WOE** [Wu & Weld, ACL ‘10]
- Limitations
  1. Incoherent extractions – the system makes independent decisions whether to include each word in the relation phrase, possibly gluing together unrelated words
  2. Uninformative extractions – those omitting critical information (e.g., “has” instead of “has a population of” or “has a Ph.D. in”)
- **ReVerb** [Fader et al., EMNLP ‘11] solves these problems by adding syntactic constraints
  - Every multi-word relation phrase must begin with a verb, end with a preposition and be a contiguous sequence of words)
  - Relation phrases should not omit nouns
  - Minimal number of distinct argument pairs in a large corpus

OLLIE: Open Language Learning for Information Extraction

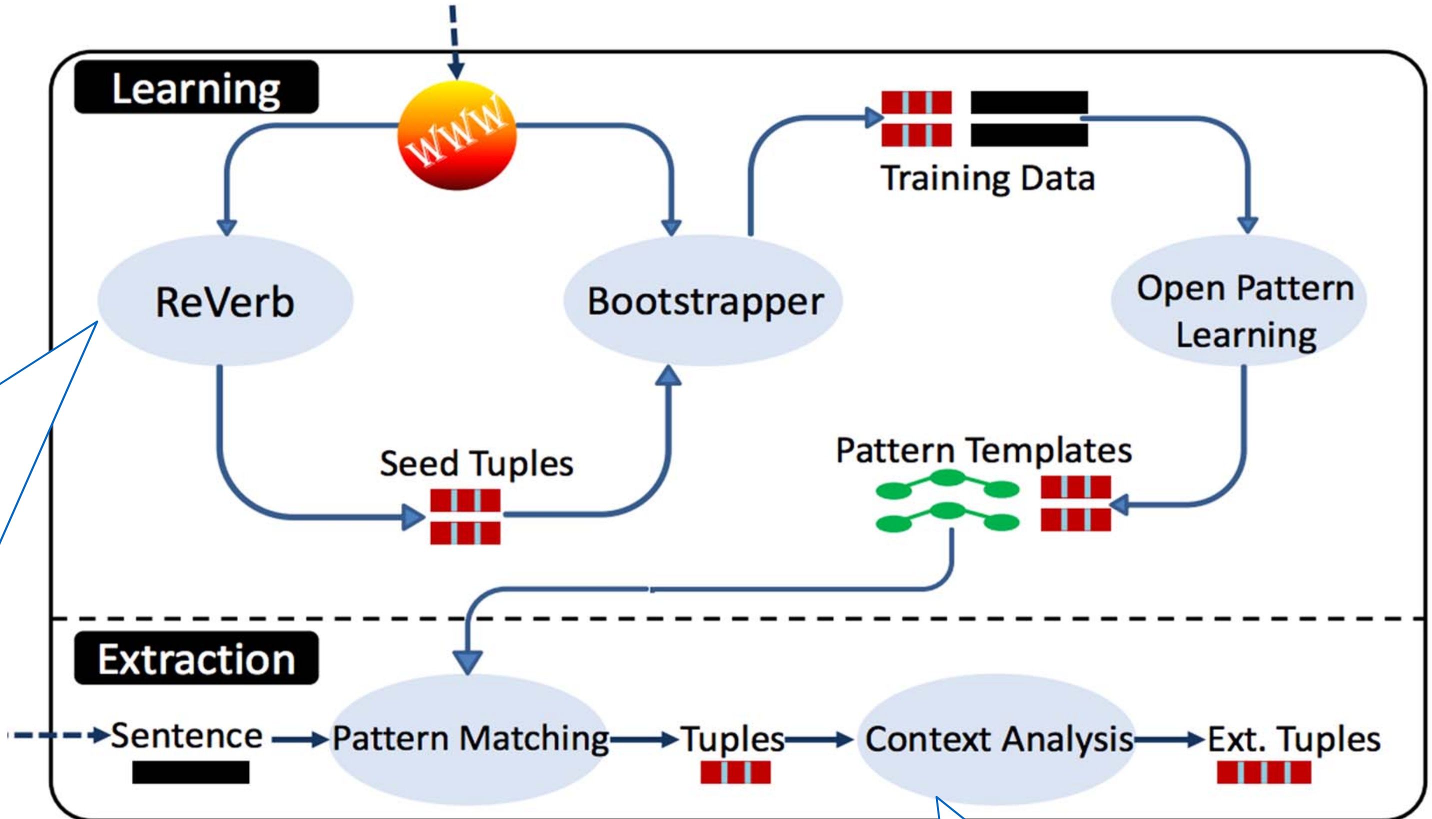
[Mausam et al., EMNLP '12]

Limitations of ReVerb

- Only extracts relations mediated by verbs
- Ignores context, potentially extracting facts that are not asserted

1. "After winning the Superbowl, the Saints are now the top dogs of the NFL."
   O: (the Saints; win; the Superbowl)
2. "There are plenty of taxis available at Bali airport."
   O: (taxis; be available at; Bali airport)
3. "Microsoft co-founder Bill Gates spoke at ..."
   O: (Bill Gates; be co-founder of; Microsoft)
4. "Early astronomers believed that the earth is the center of the universe."
   R: (the earth; be the center of; the universe)
   W: (the earth; be; the center of the universe)
   O: ((the earth; be the center of; the universe)
   AttributedTo believe; Early astronomers)
5. "If he wins five key states, Romney will be elected President."
   R,W: (Romney; will be elected; President)
   O: ((Romney; will be elected; President)
   ClausalModifier if; he wins five key states)

OLLIE (cont'd)



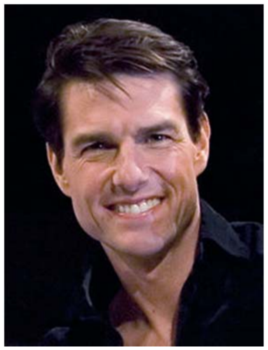
Uses ReVerb to build a seed set.

<u>Assumption:</u> every relation can also be expressed via a verb-based expression

Uses dependency parse structure

Extracting **structured** facts from **unstructured** sources (text)

/en/**tom_cruise**



/people/person/
date_of_birth

Thomas Cruise Mapother IV (born July 3, 1962),

/common/topic/
alias

widely known as Tom Cruise,

/people/person/
nationality

/people/person/
profession

is an American film actor and producer.

/en/**actor**

/en/**film_producer**

/en/**united_states**

Knowledge discovery

- Relying on humans
  - Volunteer contributions at Freebase.com
  - Import of large datasets (e.g., IMDB)
  - **Head + torso**
- Automatic extraction
  - Extraction from web pages
  - **The long tail**
  - Learning patterns using known facts

“… **jumped from** X **into** Y …”

</en/tower_bridge,

/transportation/bridge/body_of_water_spanned,

/en/river_thames>

http://www.flickr.com/photos/sandreli/4691045841/

Knowledge Fusion

Freebase

Webmaster annotations

Tables

NL text

Page structure

Deep neural network

Path Ranking Algorithm (PRA)

Extractors

Priors

Fusion

[Dong et al., KDD 2014]

Research 3: Statistical techniques for big data

Mon, 10:30-12, Empire West

KNOWLEDGE VAULT

Knowledge Fusion

Freebase

[Cafarella et al., VLDB '08; Venetis et al., VLDB '12]

[Franz et al., ISWC '09; Bordes et al., AI/Stats '12; Drumond et al., SAC '12; Socher et al., NIPS '13]

[Lao et al., EMNLP '11] (PRA)

(http://) Schema.org

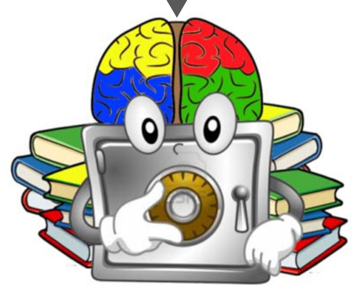
[Mintz et al., RANLP '09]

[Cafarella et al., CACM '11]

Extractors

Priors

[Dong et al., KDD 2014]

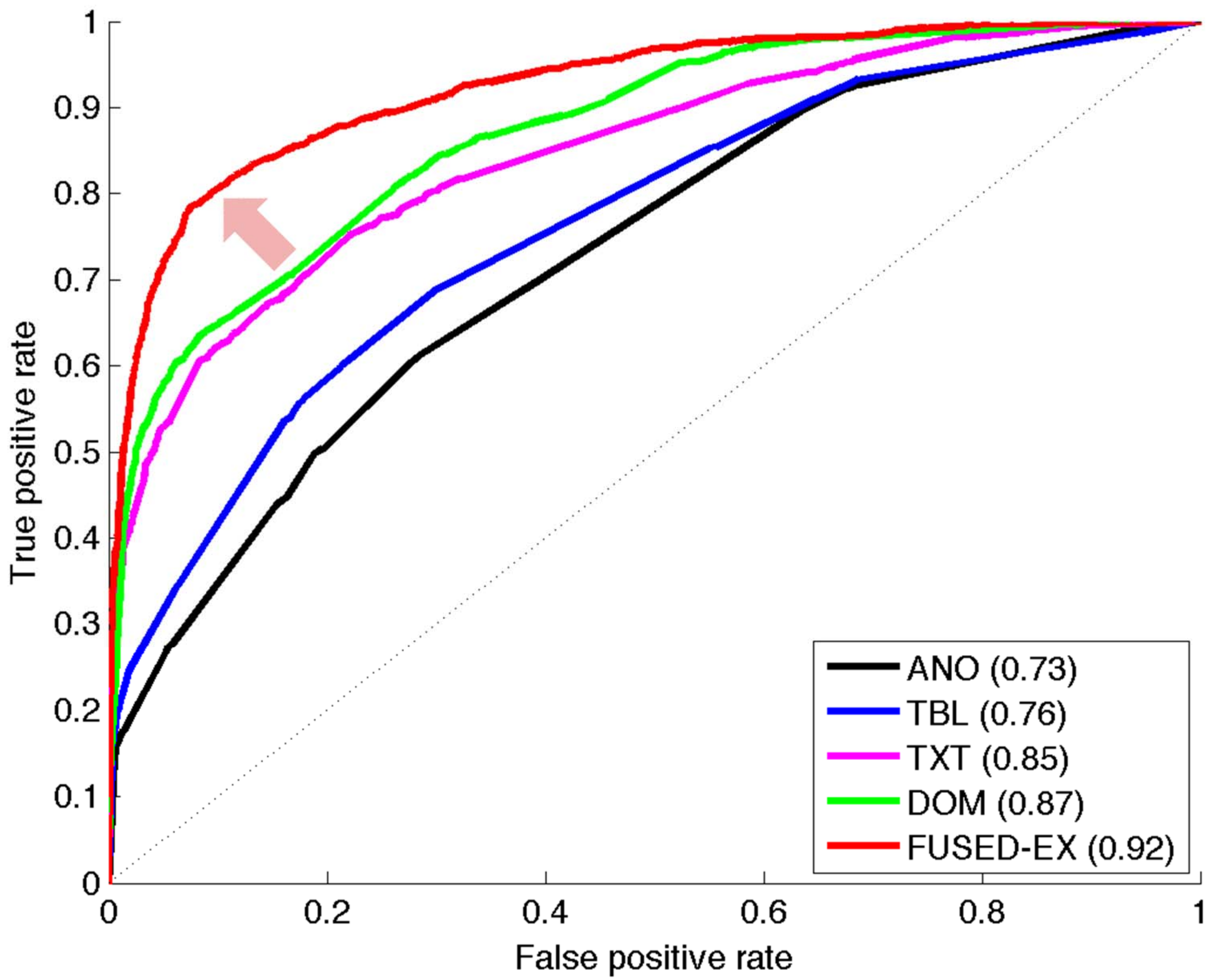


KNOWLEDGE VAULT

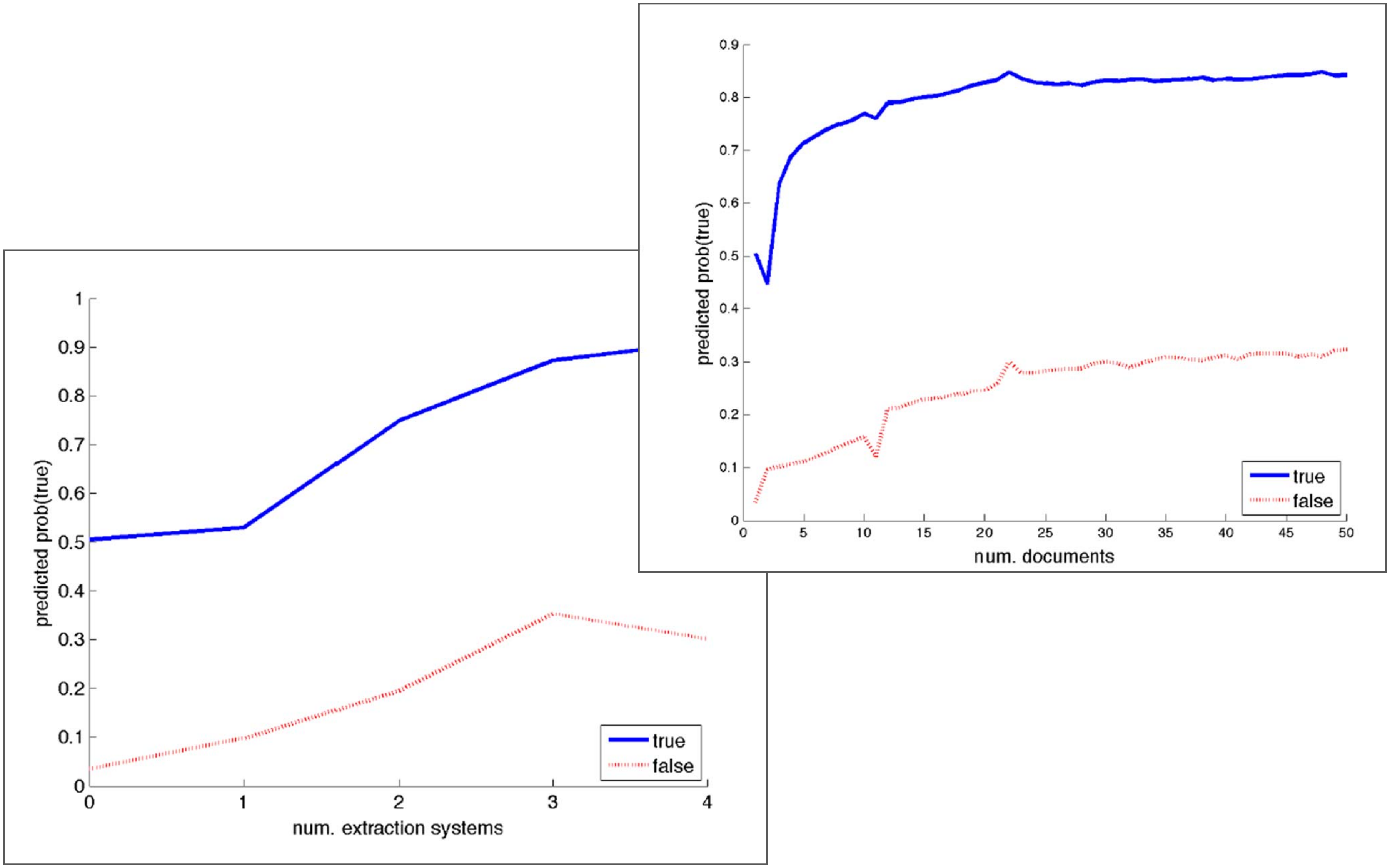
Research 3: Statistical techniques for big data

Mon, 10:30-12, Empire West

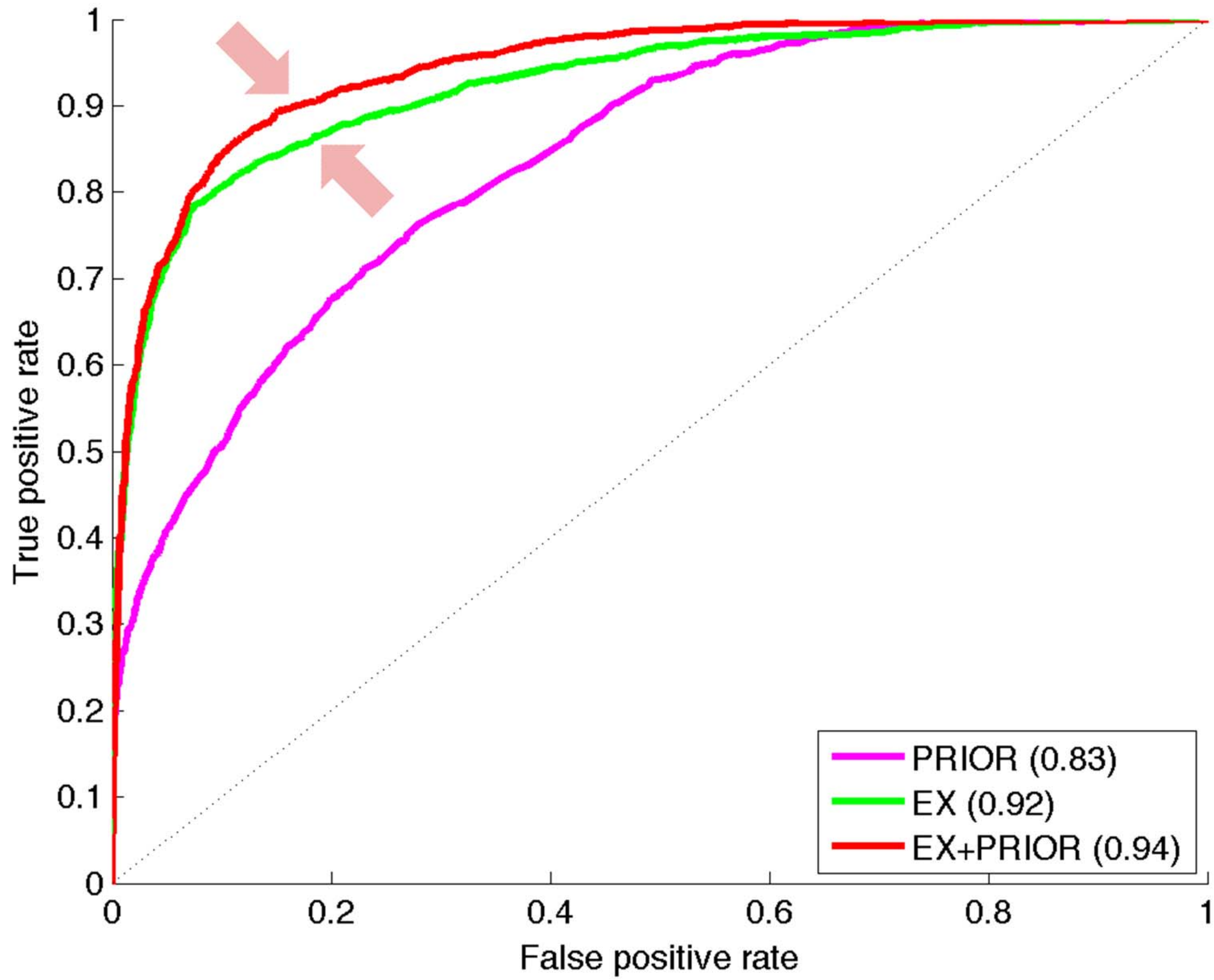
Fusing multiple extractors



The importance of adding more evidence



Fusing extractors with priors



Example: (Barry Richter, studiedAt, UW-Madison)

"In the fall of 1989, Richter accepted a scholarship to the University of Wisconsin, where he played for four years and earned numerous individual accolades ..."

"The Polar Caps' cause has been helped by the impact of knowledgeable coaches such as Andringa, Byce and former UW teammates Chris Tancill and Barry Richter."

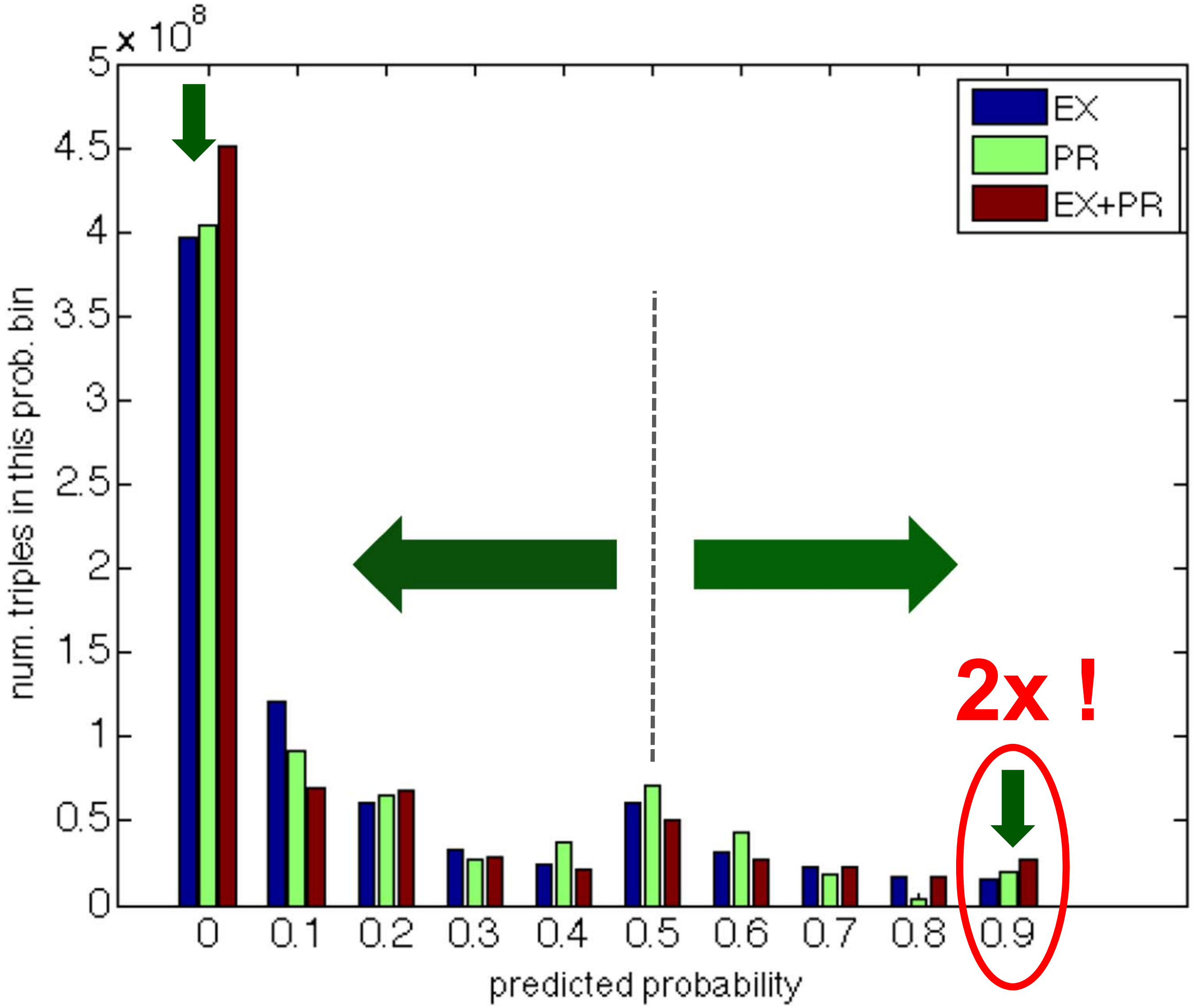
➔ Fused extraction confidence: **0.14**

Freebase

<Barry Richter, born in, Madison>
<Barry Richter, lived in, Madison>

➔ Final belief (fused with prior): **0.61**

The importance of prior modeling



Comparison of knowledge repositories

Total # facts in KNOWLEDGE VAULT > 2.5B

Name	# Entity types	# Entity instances	# Relation types	# Confident facts (relation instances)
Knowledge Vault (KV)	1100	45M	4469	302M
DeepDive [32]	4	2.7M	34	7M[a]
NELL [8]	271	5.19M	306	0.435M[b]
PROSPERA [30]	11	N/A	14	0.1M
YAGO2 [19]	350,000	9.8M	100	4M[c]
Freebase [4]	1,500	40M	35,000	637M[d]
Knowledge Graph (KG)	1,500	570M	35,000	18,000M[e]

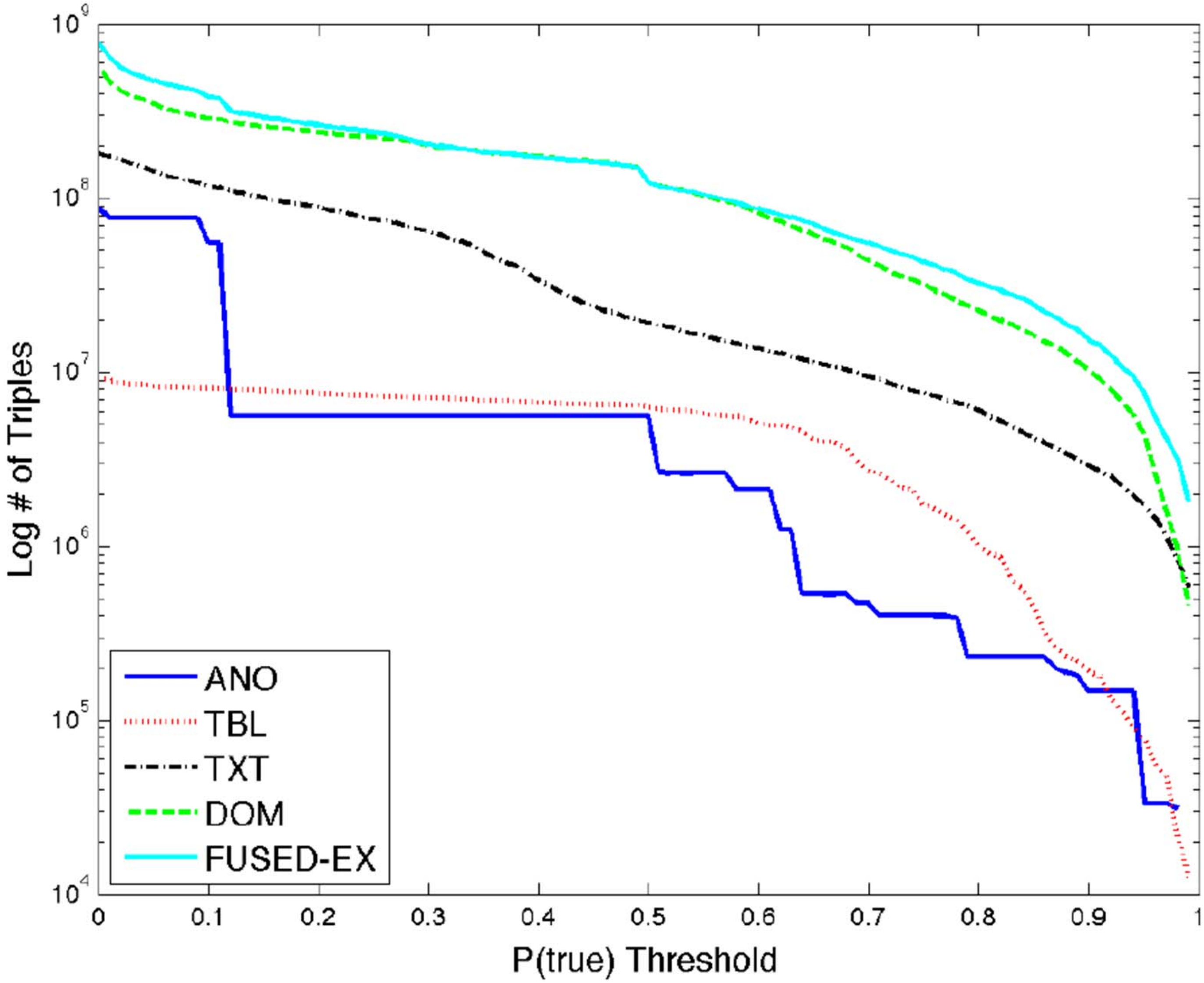
302M with Prob > 0.9

381M with Prob > 0.7

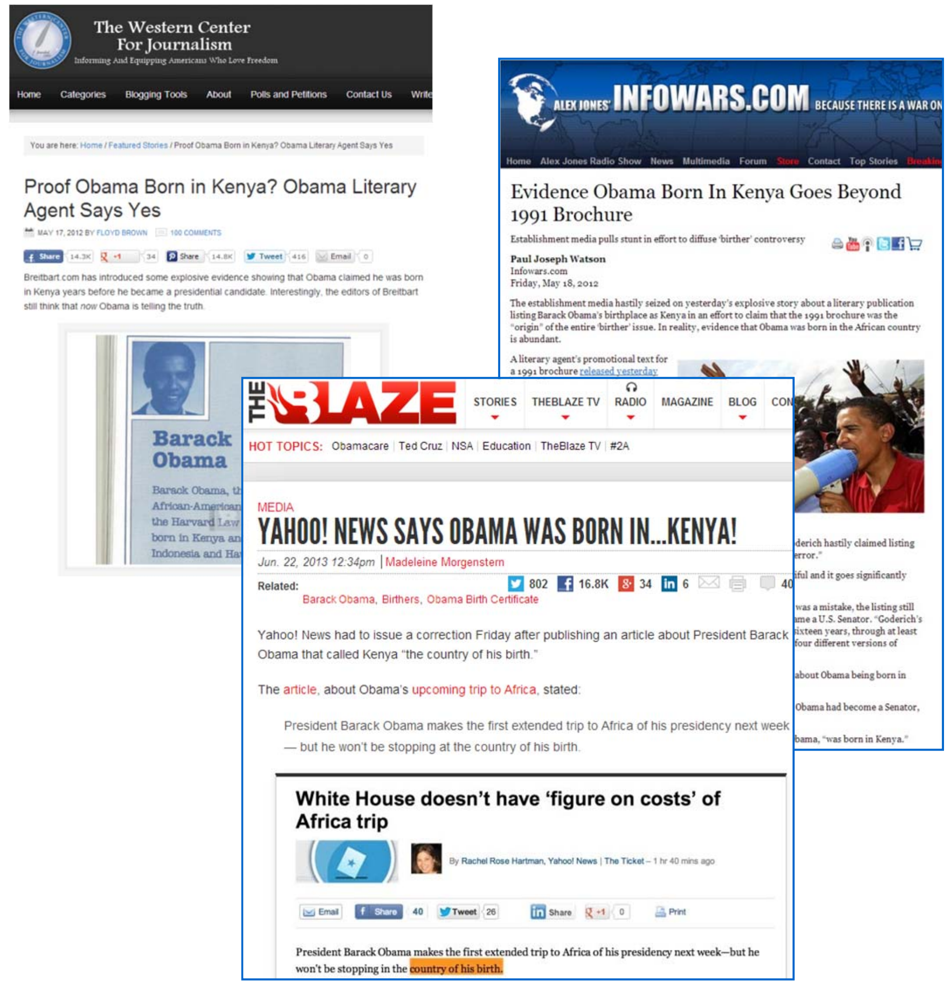
Open IE (e.g., Mausam et al., 2012)

5B assertions (Mausam, Michael Schmitz, personal communication, October 2013)

The yield from different extraction systems

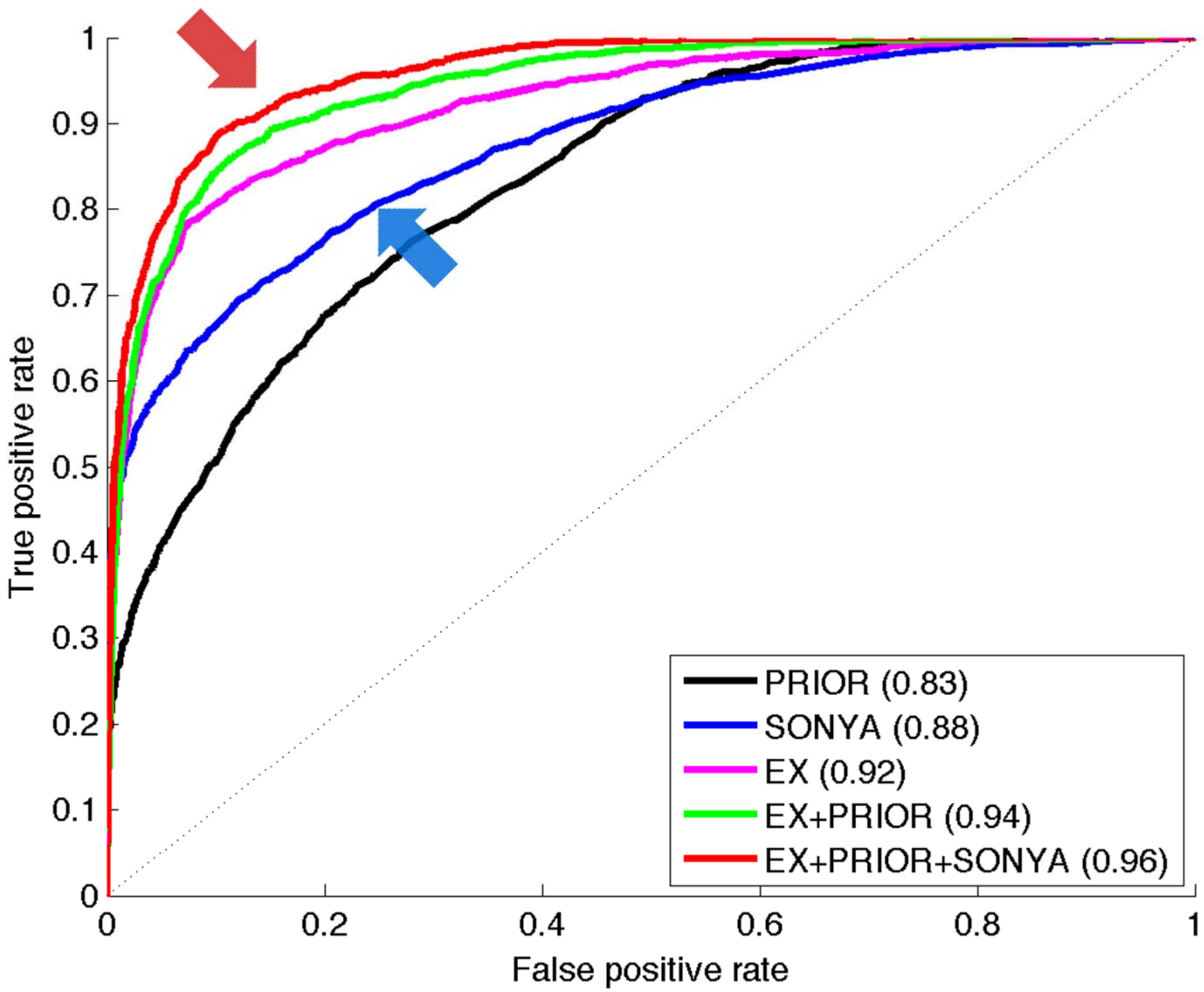


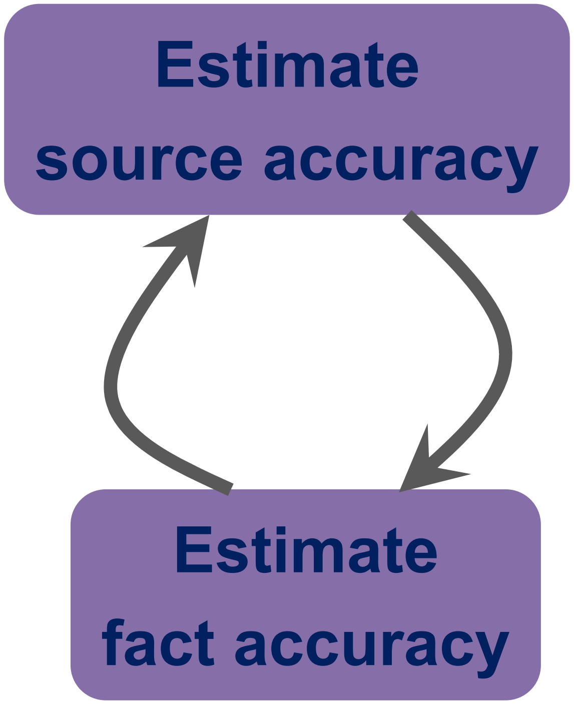
Should we trust all sources equally ?



Joint modeling of source and fact accuracy

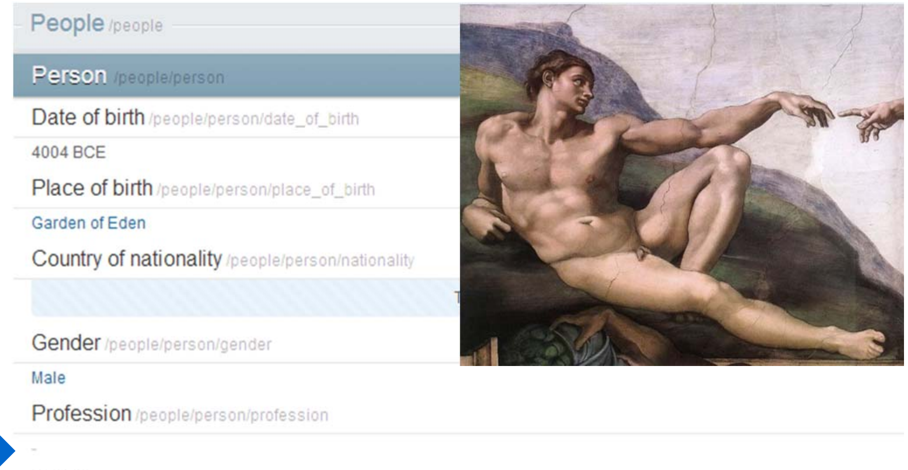
[Dong et al., VLDB '09]





Automatic knowledge base completion (focused extraction)

Relation	% unknown in Freebase
Profession	68%
Place of birth	71%
Nationality	75%
Education	91%
Spouse	92%
Parents	94%

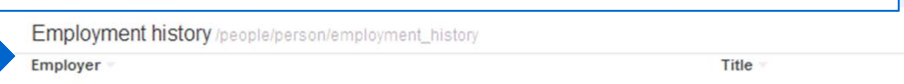


(Genesis 2)

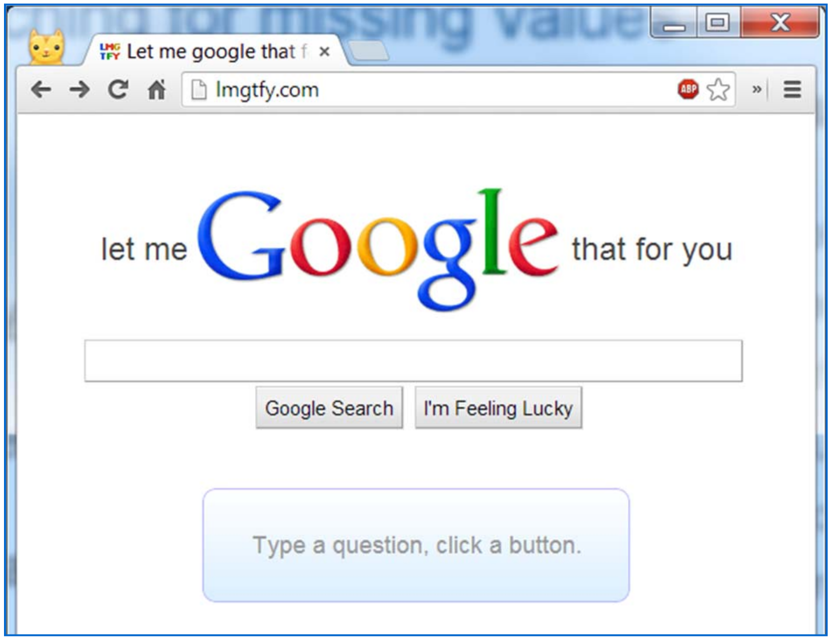
8 And the LORD God **planted a garden eastward in Eden; and there he put the man whom he had formed**.

15 Then the LORD God **took the man and put him in the garden of Eden to tend and keep it**.

19 And out of the ground the LORD God formed every beast of the field, and every fowl of the air; and **brought them unto Adam to see what he would call them: and whatsoever Adam called every living creature, that was the name thereof**.

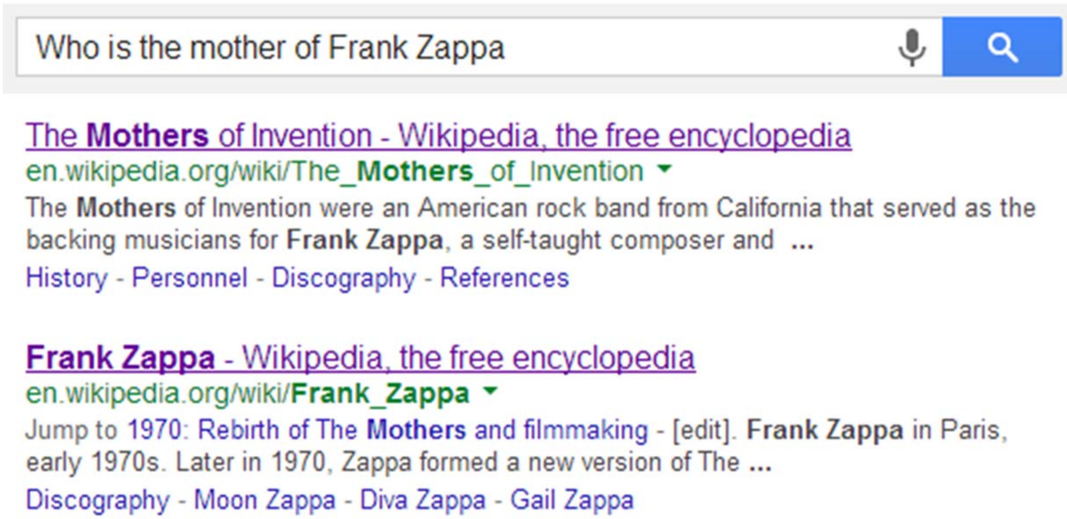


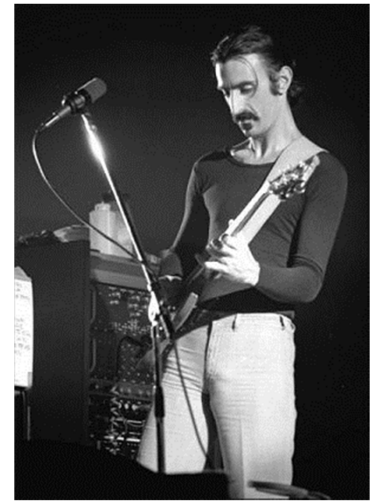
Proactively searching for missing values [West et al., WWW '14]

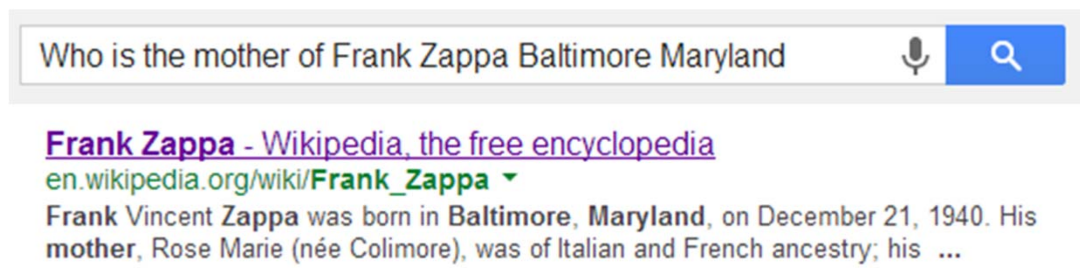


- Mine search logs for best query templates (per relation)
- Augment queries with disambiguating information
- Thou shalt ask in moderation
  - Asking too much may be harmful!

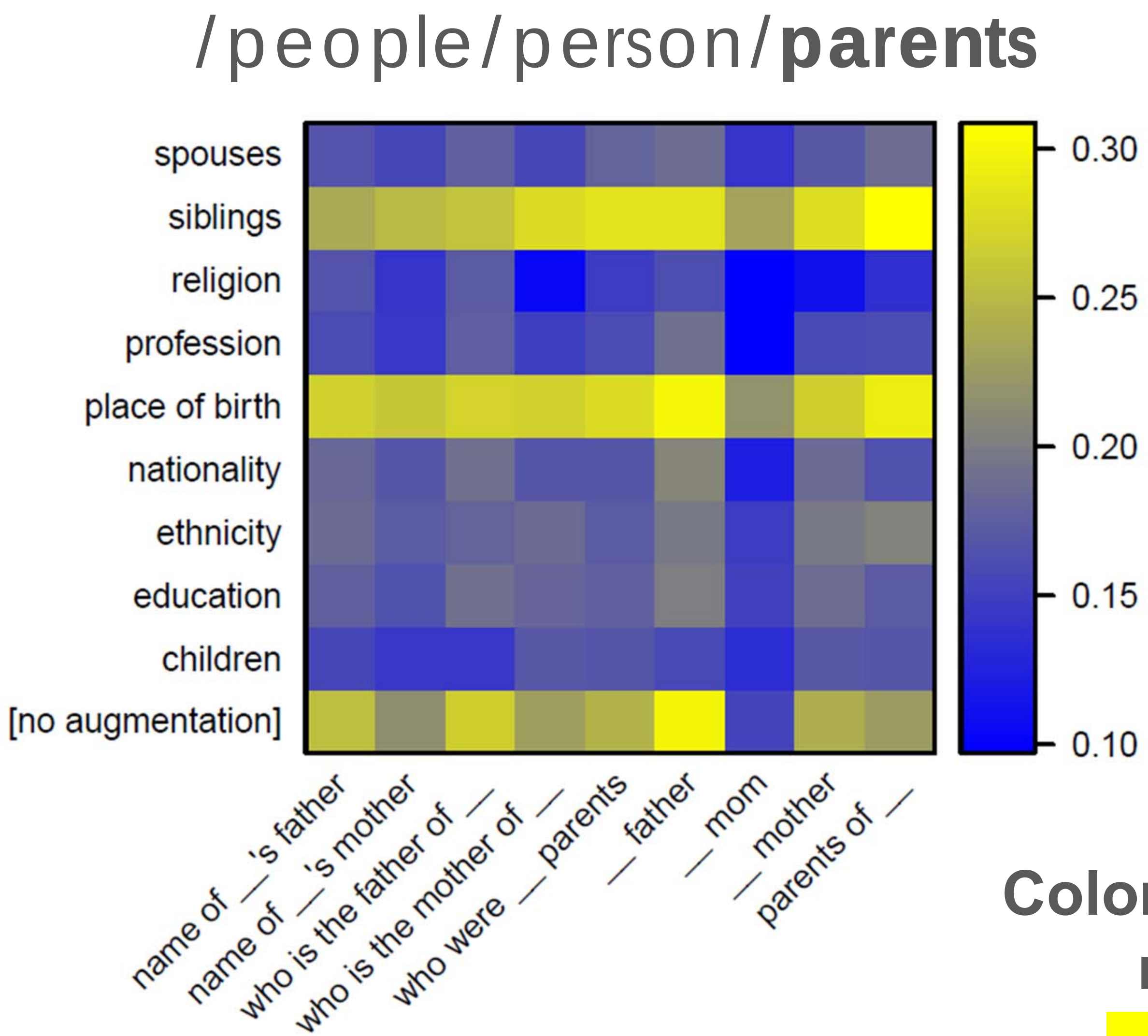
The importance of query augmentation







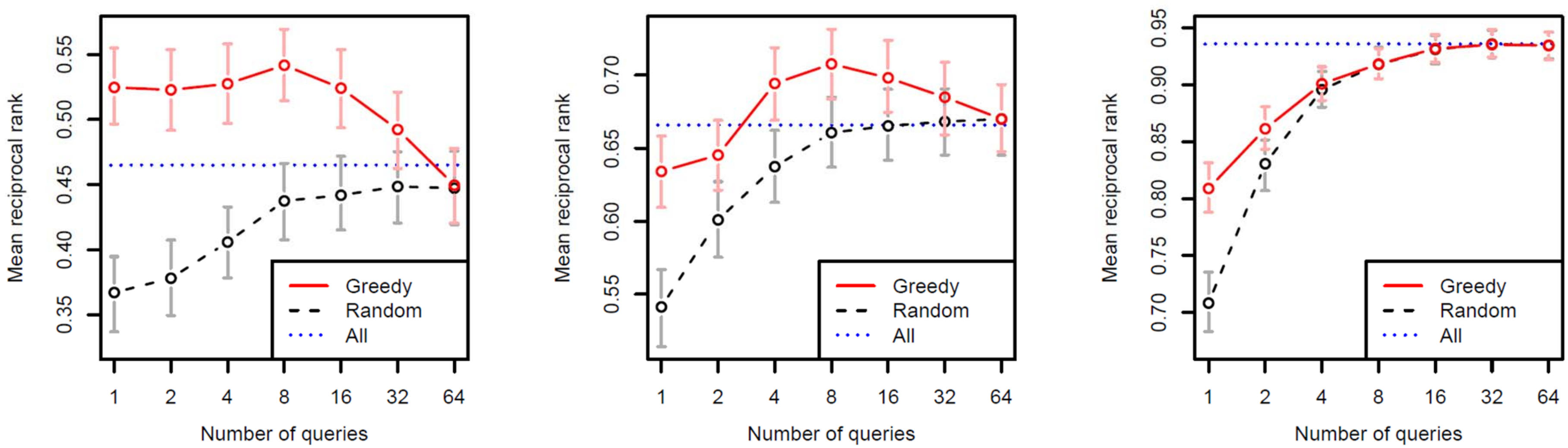
Learning to query



Color = mean reciprocal rank of true answer

GOOD **BAD**

Asking the right (number of) questions



(a) SPOUSES (b) PLACE OF BIRTH (c) NATIONALITY

PART 2: METHODS AND TECHNIQUES

Methods and techniques

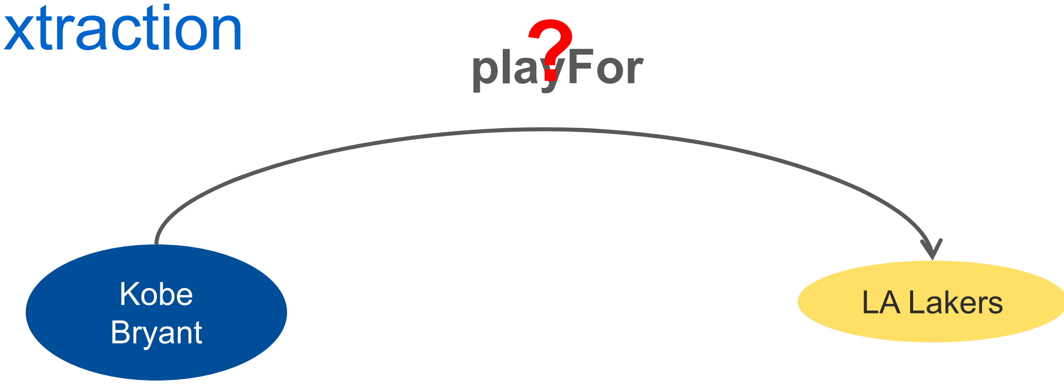
1. Relation extraction:
   - Supervised models
   - Semi-supervised models
   - Distant supervision
2. Entity resolution
   - Single entity methods
   - Relational methods
3. Link prediction
   - Rule-based methods
   - Probabilistic models
   - Factorization methods
   - Embedding models

Not in this tutorial:

- Entity classification
- Group/expert detection
- Ontology alignment
- Object ranking

RELATION EXTRACTION

Relation Extraction



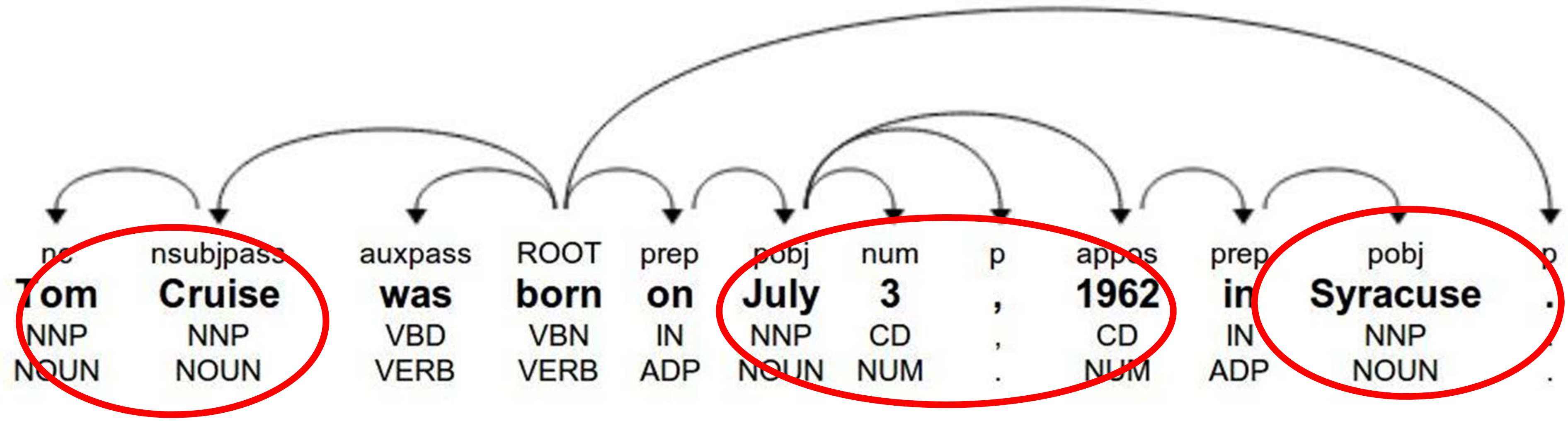
“Kobe Bryant,	the franchise player of	the Lakers”
“Kobe	once again saved	his team”
“Kobe Bryant	man of the match for	Los Angeles”

- Extracting semantic relations between sets of [grounded] entities
- Numerous variants:
  - Undefined vs pre-determined set of relations
  - Binary vs n-ary relations, facet discovery
  - Extracting temporal information
  - Supervision: {fully, un, semi, distant}-supervision
  - Cues used: only lexical vs full linguistic features

Supervised relation extraction

- Sentence-level labels of relation mentions
  - "**Apple** CEO **Steve Jobs** said.." => (SteveJobs, CEO, Apple)
  - "**Steve Jobs** said that **Apple** will.." => NIL
- Traditional relation extraction datasets
  - ACE 2004
  - MUC-7
  - Biomedical datasets (e.g BioNLP callenges)
- Learn classifiers from +/- examples
- Typical features: context words + POS, dependency path between entities, named entity tags, token/parse-path/entity distance

Examples of features



X was born on DDDD in Y

- DEP: X <*nsubjpass* / born *prep*> on *pobj*> DATE *prep*> in *pobj*> Y
- NER: X = PER, Y = LOC
- POS: X = NOUN, NNP; Y = NOUN, NNP
- Context: born, on, in , "born on"

Supervised relation extraction

- Used to be the “traditional” setting [Riloff et al., 06; Soderland et al., 99]

- **Pros**
  - High quality supervision
  - Explicit negative examples

- **Cons**
  - Very expensive to generate supervision
  - Not easy to add more relations
  - Cannot generalize to text from different domains

Semi-supervised relation extraction

- **Generic algorithm**
  1. Start with seed triples / golden seed patterns
  2. Extract patterns that match seed triples/patterns
  3. Take the top-k extracted patterns/triples
  4. Add to seed patterns/triples
  5. Go to 2

- Many published approaches in this category:
  - Dual Iterative Pattern Relation Extractor [Brin, 98]
  - Snowball [Agichtein & Gravano, 00]
  - TextRunner [Banko et al., 07] – almost unsupervised

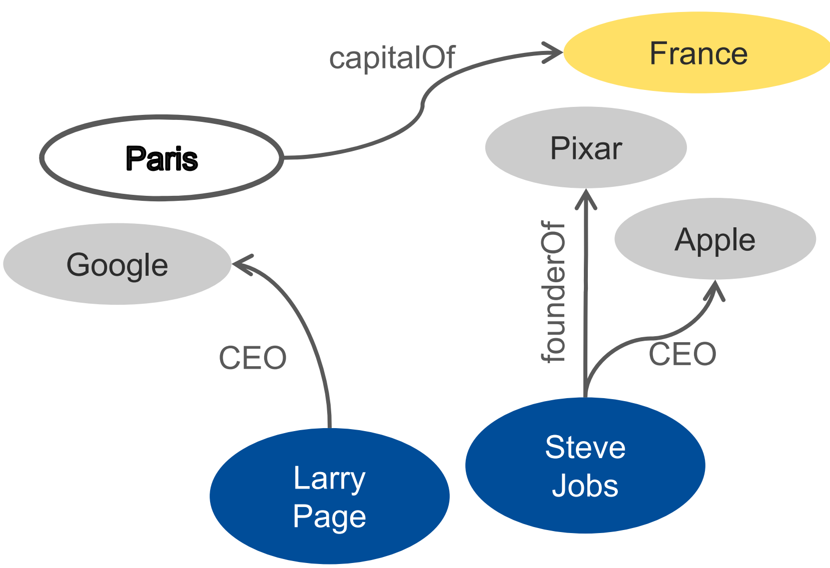
- Differ in pattern definition and selection

TextRunner [Banko et al., 07]

- Almost unsupervised
  - Relations not fixed: does not follow Knowledge Graph schema (~~growing~~)
  - No labeled data
  - Mostly unlabeled text
  - Uses heuristics to self-label a starting corpora (using a parser), such as
    - Path length < k
    - Path does not cross sentence-like boundaries like relative clauses
    - Neither entity is a pronoun
- Self-training
  - Generate +/- examples → learn classifier
  - Extract new relation mentions using this classifier
  - Generate triples from aggregated mentions, assign probabilistic score using [Downey et. al., 2005]
- Later improved in Reverb [Fader et al., 11]

Distantly-supervised relation extraction

- Existing knowledge base + unlabeled text → generate examples
  - Locate pairs of related entities in text
  - Hypothesizes that the relation is expressed



Google CEO Larry Page announced that...

Steve Jobs has been Apple for a while...

Pixar lost its co-founder Steve Jobs...

I went to Paris, France for the summer...

Distant supervision: modeling hypotheses

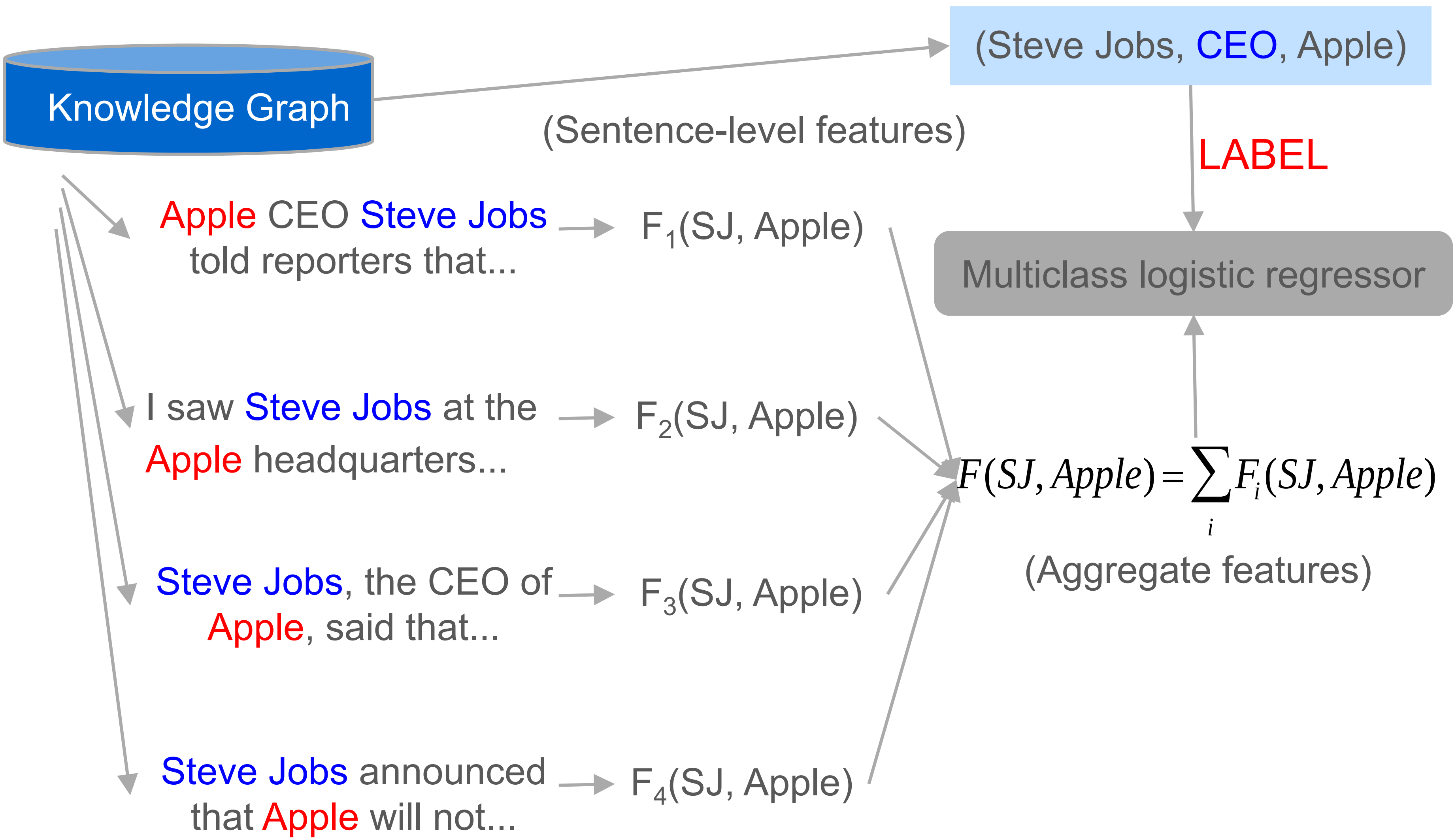
Typical architecture:

1. Collect many pairs of entities co-occurring in sentences from text corpus

2. If 2 entities participate in a relation, several hypotheses:
   1. **All** sentences mentioning them express it [Mintz et al., 09]

"**Barack Obama** is the 44th and current President of **the US**." → (BO, employedBy, USA)

[Mintz et al., 09]



[Mintz et al., 09]

- Classifier: multiclass logistic regressor

(Steve Jobs, Apple, *AggFeatures*)

CEO

PrimeMinister

Spokesman

CapitalOf

NotRelated

- Negative examples
  - Randomly sample unrelated entity pairs occurring in the same sentence
  - > 98% such pairs actually unrelated

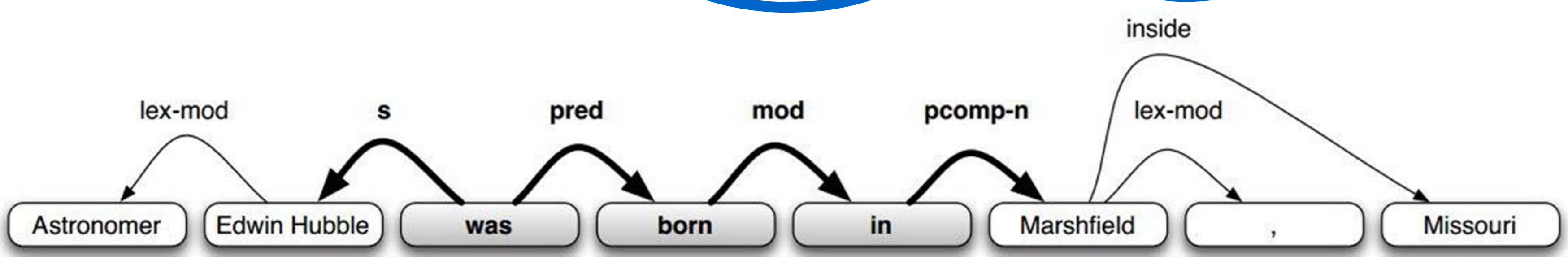
Sentence-level features

- Lexical: words in between and around mentions and their parts-of-speech tags (conjunctive form)
- Syntactic: dependency parse path between mentions along with side nodes
- Named Entity Tags: for the mentions
- Conjunctions of the above features
  - Distant supervision is used on to lots of data → sparsity of conjunctive forms not an issue

Sentence-level features

Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]

Table 3: Features for 'Astronomer Edwin Hubble was born in Marshfield, Missouri'.



Examples of top features

Relation	Feature type	Left window	NE1	Middle	NE2	Right window
/architecture/structure/architect	LEX$\curvearrowleft$		ORG	, the designer of the	PER	
	SYN	designed $\Uparrow_s$	ORG	$\Uparrow_s$ designed $\Downarrow_{by-subj}$ by $\Downarrow_{pcn}$	PER	$\Uparrow_s$ designed
/book/author/works_written	LEX		PER	s novel	ORG	
	SYN		PER	$\Uparrow_{pcn}$ by $\Uparrow_{mod}$ story $\Uparrow_{pred}$ is $\Downarrow_s$	ORG	
/book/book_edition/author_editor	LEX$\curvearrowleft$		ORG	s novel	PER	
	SYN		PER	$\Uparrow_{nn}$ series $\Downarrow_{gen}$	PER	
/business/company/founders	LEX		ORG	co - founder	PER	
	SYN		ORG	$\Uparrow_{nn}$ owner $\Downarrow_{person}$	PER	
/business/company/place_founded	LEX$\curvearrowleft$		ORG	- based	LOC	
	SYN		ORG	$\Uparrow_s$ founded $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC	
/film/film/country	LEX		PER	, released in	LOC	
	SYN	opened $\Uparrow_s$	ORG	$\Uparrow_s$ opened $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC	$\Uparrow_s$ opened
/geography/river/mouth	LEX		LOC	, which flows into the	LOC	
	SYN	the $\Downarrow_{det}$	LOC	$\Uparrow_s$ is $\Downarrow_{pred}$ tributary $\Downarrow_{mod}$ of $\Downarrow_{pcn}$	LOC	$\Downarrow_{det}$ the

Distant supervision: modeling hypotheses

Typical architecture:

1. Collect many pairs of entities co-occurring in sentences from text corpus

2. If 2 entities participate in a relation, several hypotheses:
   1. **All** sentences mentioning them express it [Mintz et al., 09]
   2. **At least one** sentence mentioning them express it [Riedel et al., 10]

"**Barack Obama** is the 44th and current President of **the US**." → (BO, employedBy, USA)

"**Obama** flew back to **the US** on Wednesday." → (BO, employedBy, USA)

[Riedel et al., 10]

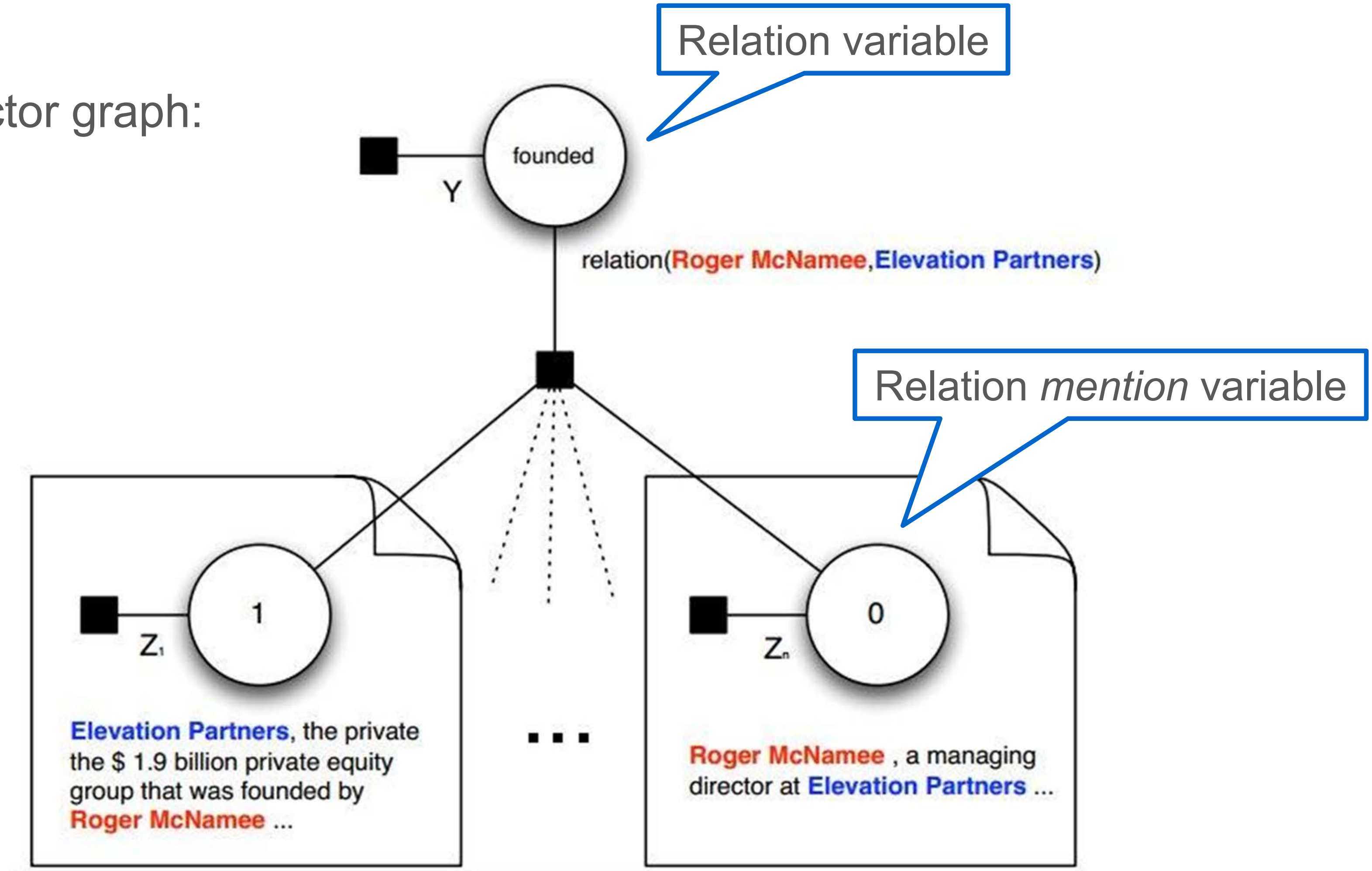
- Every mention of an entity-pair does not express a relation

Relation Type	New York Times	Wikipedia
nationality	38%	20%
place_of_birth	35%	20%
contains	20%	10%

- Violations more in news than encyclopediac articles
- Assert triple from only a few mentions, not all

[Riedel et al., 10]

- Factor graph:



- Multiple-instance setting

Distant supervision: modeling hypotheses

Typical architecture:

1. Collect many pairs of entities co-occurring in sentences from text corpus

2. If 2 entities participate in a relation, several hypotheses:
   1. **All** sentences mentioning them express it [Mintz et al., 09]
   2. **At least one** sentence mentioning them express it [Riedel et al., 10]
   3. **At least one** sentence mentioning them express it and 2 entities can express **multiple relations** [Hoffmann et al., 11] [Surdeanu et al., 12]

"**Barack Obama** is the 44th and current President of **the US**." → (BO, employedBy, USA)

"**Obama** flew back to **the US** Wednesday." → (BO, employedBy, USA)

"**Obama** was born in **the US** just as he always said." → (BO, bornIn, USA)

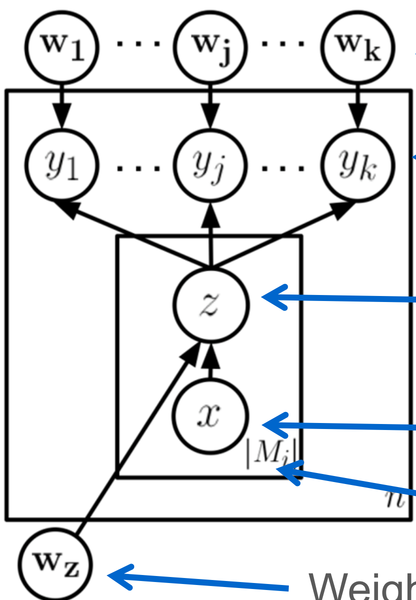
[Surdeanu et al., 12]

- Relation extraction is a multi-instance multi-label problem.

"**Barack Obama** is the 44th and current President of **the US**." → (BO, *employedBy*, USA)

"**Obama** was born in **the US** just as he always said." → (BO*, bornIn*, USA)

"**Obama** flew back to **the US** on Wednesday." → *NIL*



Weight vector for the binary top-level classifier for *j*th relation

Top-level classification decision for *j*th relation

Latent relation classification, among relations and NIL.

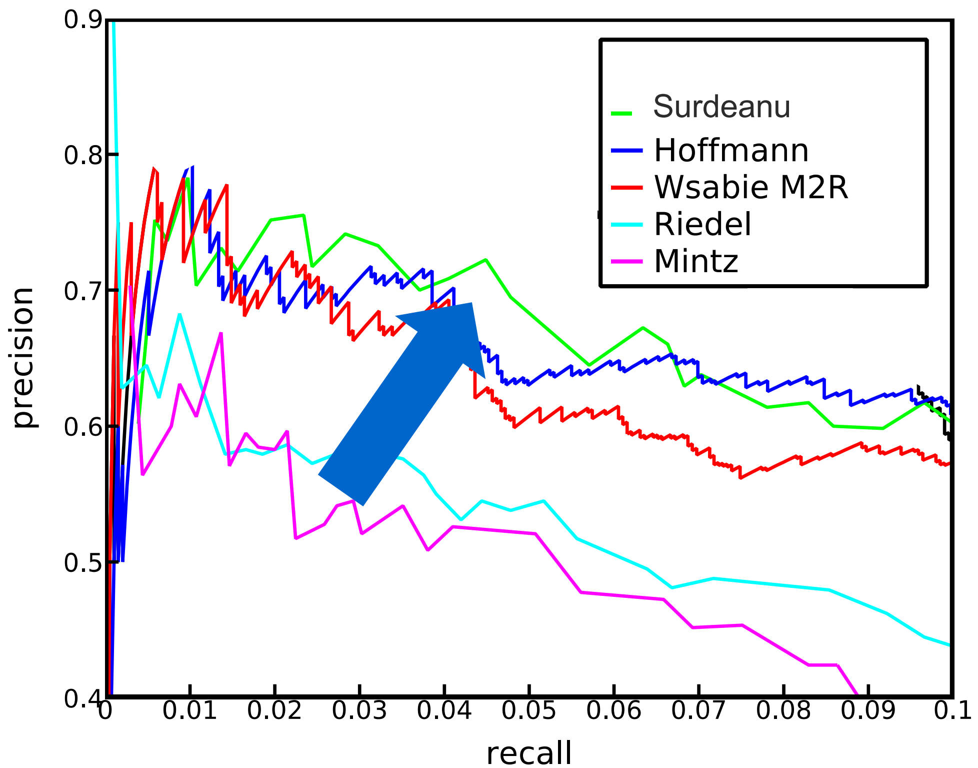
Sentence

Set of extraction for pair *i* of entities

Weight vector for multi-class mention classifier

- Training via EM with initialization with [Mintz et al., 09]

Relaxing hypotheses improves precision



Precision-recall curves on extracting from New York Times articles to Freebase [Weston et al., 13]

Distant supervision

- **Pros**
  - Can scale to the web, as no supervision required
  - Generalizes to text from different domains
  - Generates a lot more supervision in one iteration

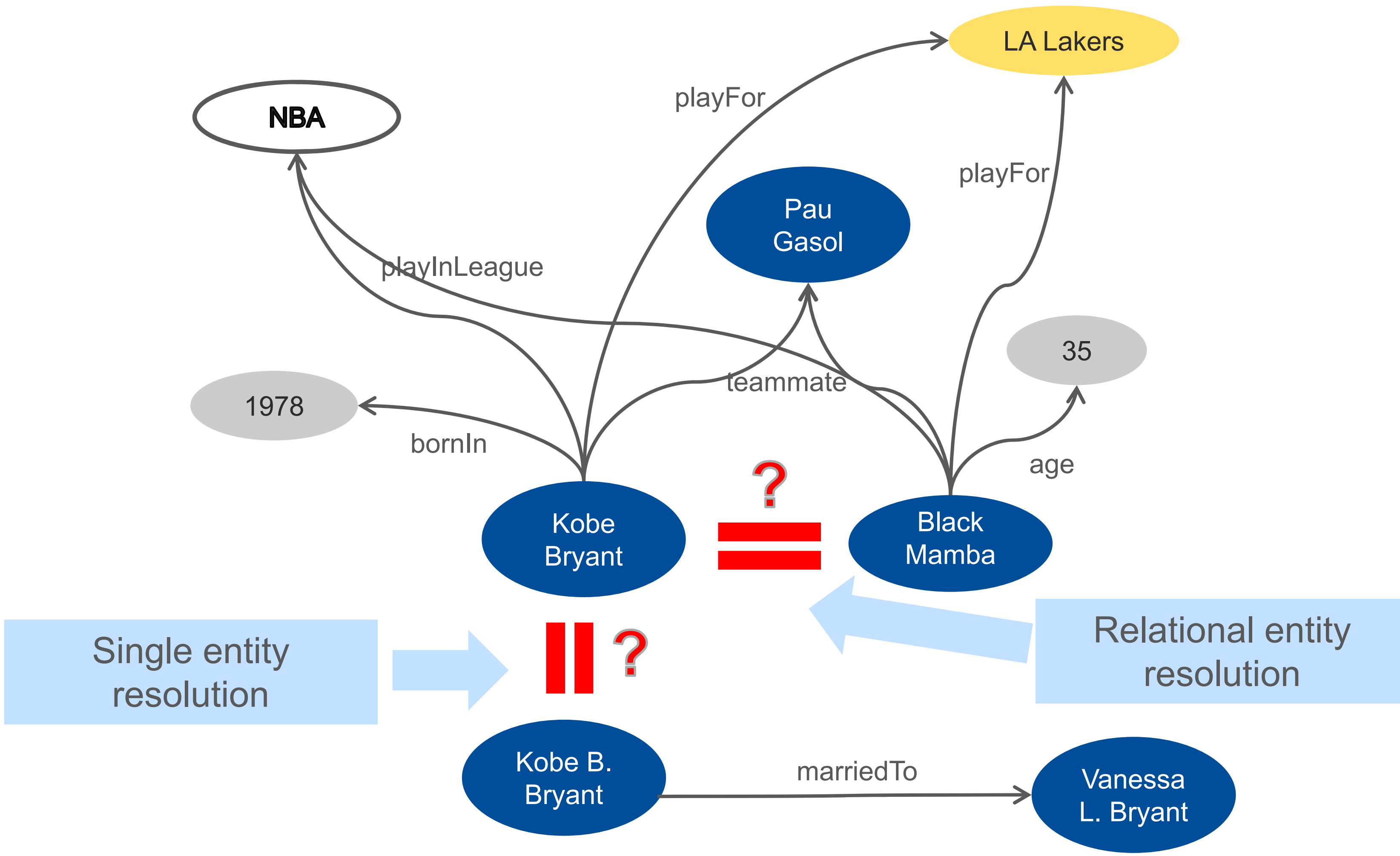
- **Cons**
  - Needs high quality entity-matching
  - Relation-expression hypothesis can be wrong
    - Can be compensated by the extraction model, redundancy, language model
  - Does not generate negative examples
    - Partially tackled by matching unrelated entities

Plenty of extensions

- Using language models [Downey et al., 07]
  - Do two entities seem to express a given relation, given the context?
- Joint relation extraction + other NLP tasks
  - Named Entity tagging [Yao et al., 10]
  - Possibly with entity resolution and/or coreference
- Jointly + repeatedly training multiple extractors [Carlson et. al., 10]
- Unsupervised extraction [Poon & Domingos, 10]
- **Jointly perform relation extraction and link prediction** [Bordes et al., 12; Weston et al., 13; Riedel et al., 13]

ENTITY RESOLUTION

Entity resolution

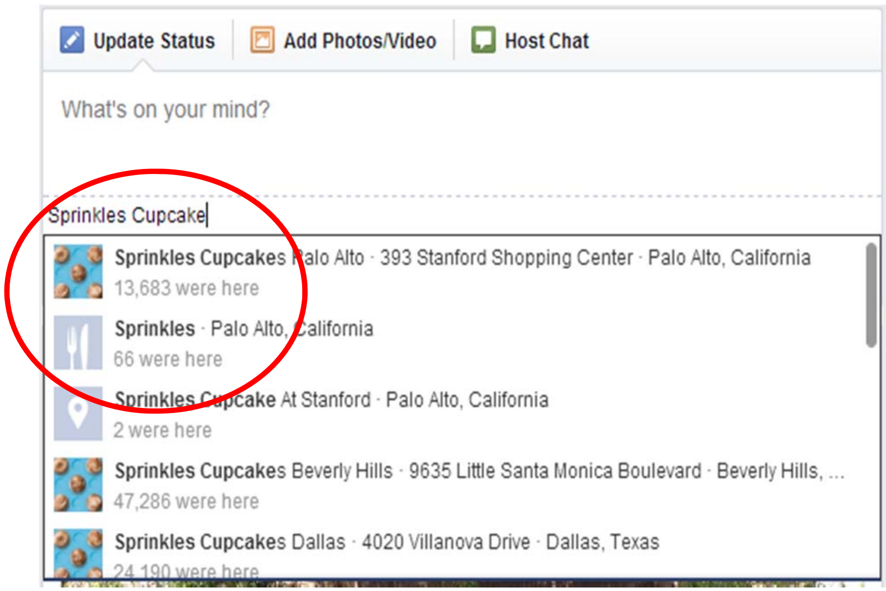


Single-entity entity resolution

- Entity resolution without using the relational context of entities
- Many distances/similarities for single-entity entity resolution:
  - Edit distance (Levenshtein, etc.)
  - Set similarity (TF-IDF, etc.)
  - Alignment-based
  - Numeric distance between values
  - Phonetic Similarity
  - Equality on a boolean predicate
  - Translation-based
  - Domain-specific

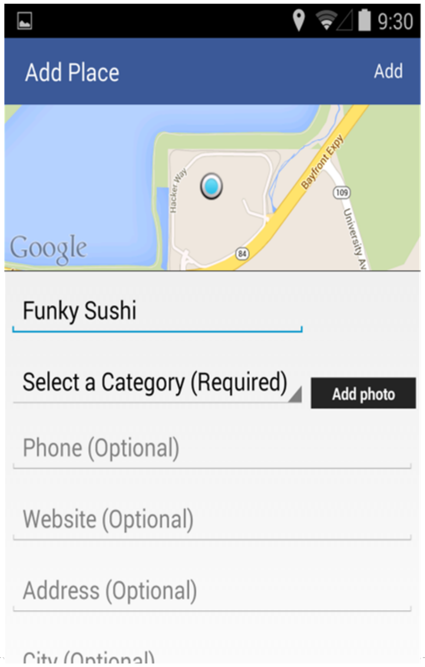
Case study: deduplicating places [Dalvi et al., 14]

- Multiple mentions of the same place is wrong and confusing.



Origin of duplicates

- Duplicates are often created during check-in:
  - Different spellings
  - GPS Errors
  - Wrong checkins
- Frequently, these duplicates have:
  - few attribute values
  - names were typed hurriedly



Effectively matching place names is hard

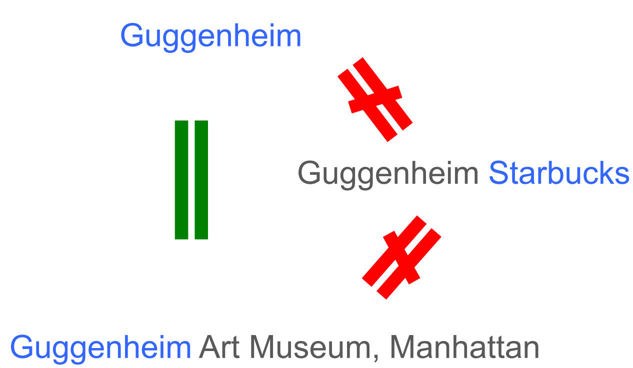
Good Matches (Help Recall)	
Guggenheim Art Museum Manhattan	Guggenheim
DishDash	Dish Dash Restaurant
Ippudo New York	Ipudo
Central Park Café (Sunnyvale)	Central Park Restaurant (Sunnyvale)

Bad Matches (Hurt Precision)	
Guggenheim	Guggenheim Starbucks
Central Park Café (NYC)	Central Park Restaurant (NYC)
Glen Park	Glen Canyon Park

- Easy to find cases where the "bad match" pair is more similar than the "good match" pair
- Existing similarity metrics (TF-IDF, Levenshtein, Learned-weight edit distance, etc.) generally fail to handle this level of variability

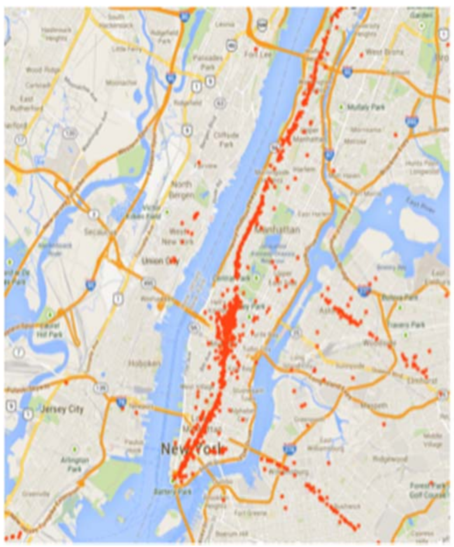
Idea 1: core words

- A core word = a word a human would use to refer to the place, if only a single word were allowed
- **Goal:** try to identify the core word, use it for comparisons



Idea 2: spatial context model

- Tokens vary in importance based on geographic context
  - Central Park is common/meaningless in NYC
- **Goal:** filter out context-specific tokens when matching names

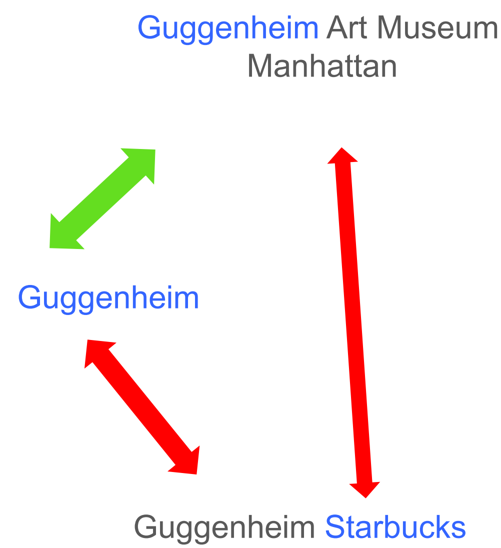


Broadway

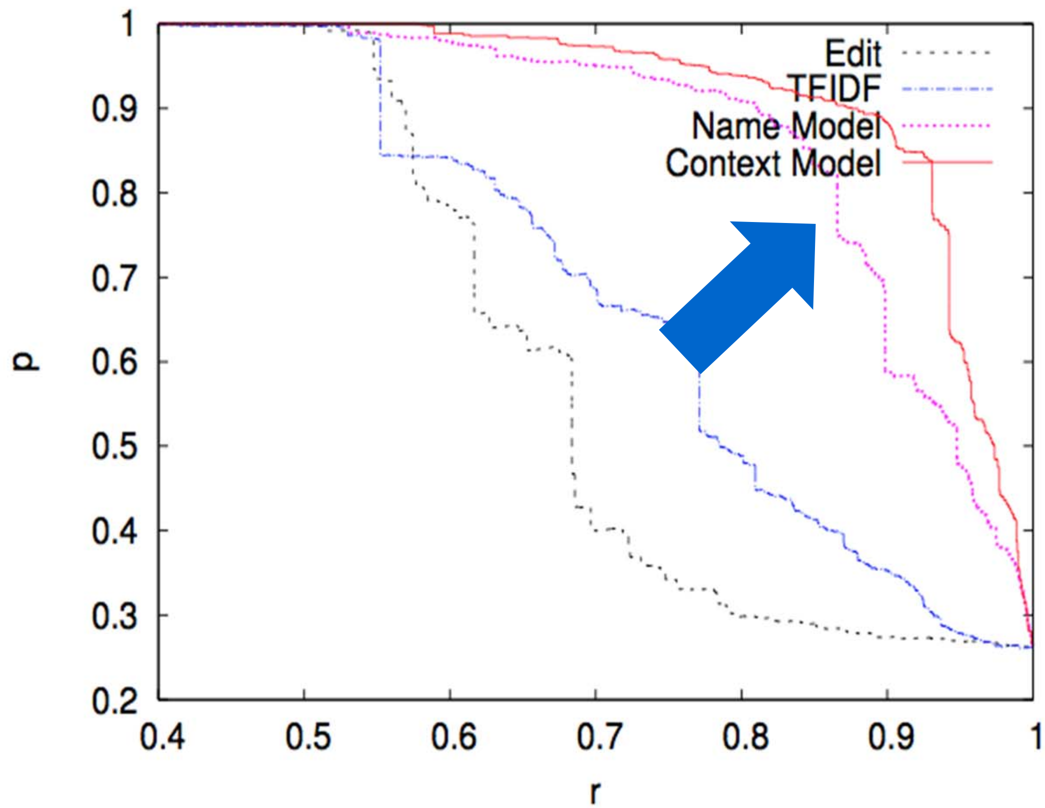
Times Square

Convert into an edit distance

- We match N_1 with N_2 given:
  - Core words model
  - Spatial contextual model
- Treat N_1, N_2 as bag of words, and require:
  - Core words match
  - Any words that match are either core or background in both N1 and N2
- Extend this to Levenshtein-like edit distance

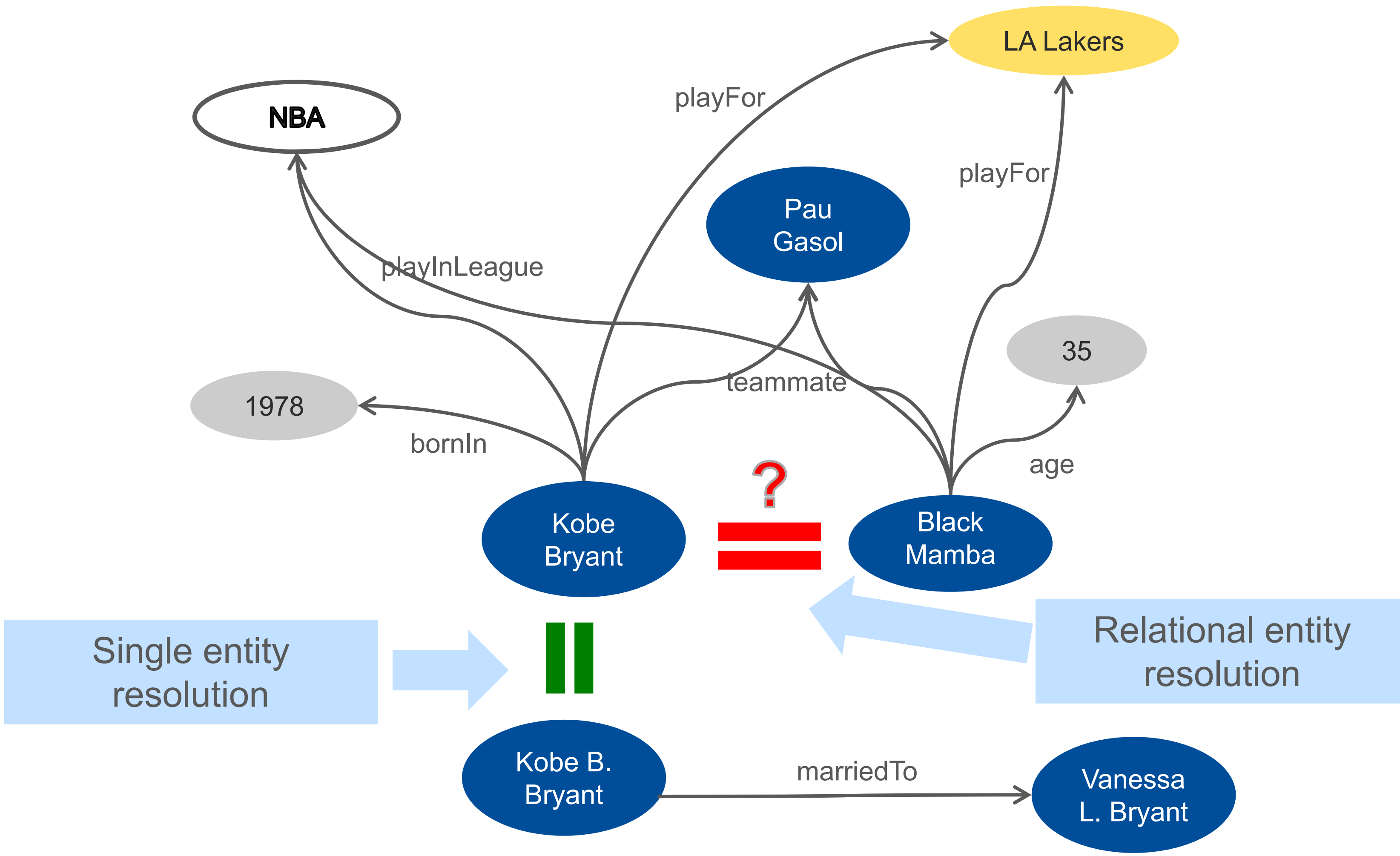


Deduplication results [Dalvi et al., 14]



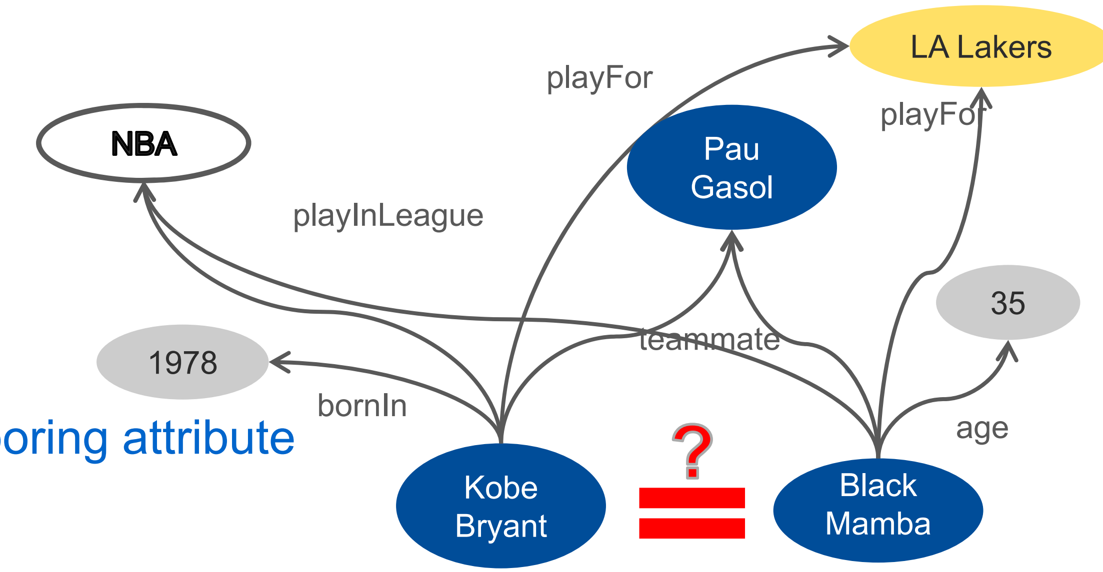
- Edit: Levenshtein distance between place names
- TF-IDF: cosine similarity of TF-IDF weighted vector of overlapping names

Entity resolution



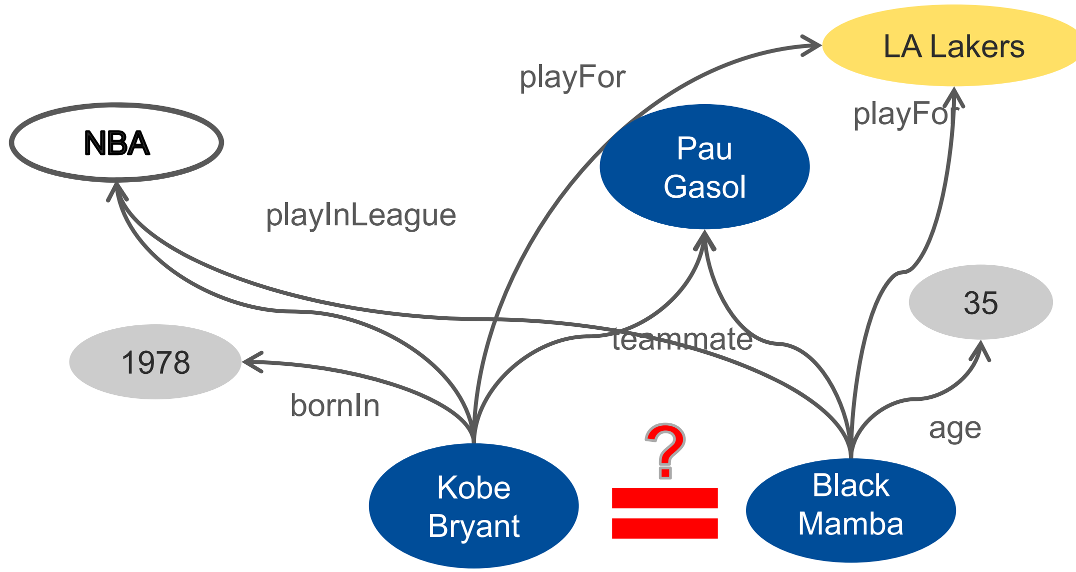
Relational entity resolution – Simple strategies

- Enrich model with relational features → richer context for matching



- Relational features:
  - Value of edge or neighboring attribute
  - Set similarity measures
    - *Overlap/Jaccard*
    - *Average similarity* between set members
    - *Adamic/Adar*: two entities are more similar if they share more items that are overall less frequent
    - *SimRank*: two entities are similar if they are related to similar objects
    - *Katz score:* two entities are similar if they are connected by shorter paths

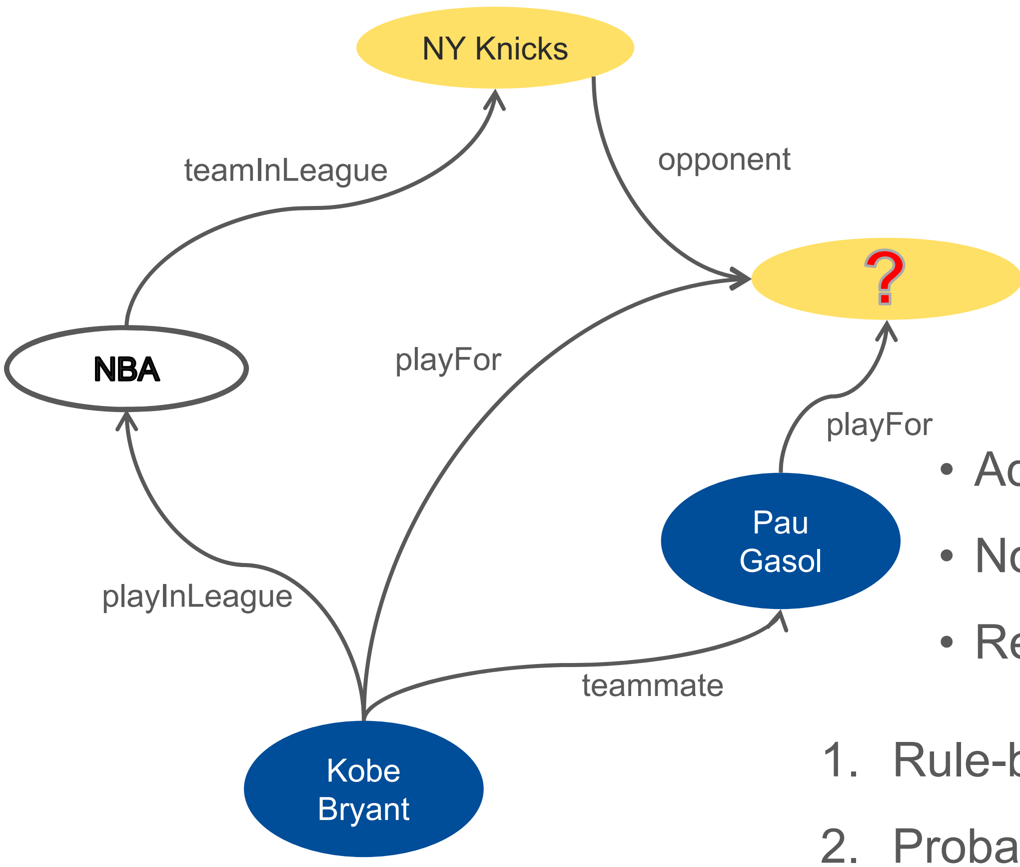
Relational entity resolution – Advanced strategies



- Dependency graph approaches [Dong et al., 05]
- Relational clustering [Bhattacharya & Getoor, 07]
- **Probabilistic Relational Models** [Pasula et al., 03]
- **Markov Logic Networks** [Singla & Domingos, 06]
- **Probabilistic Soft Logic** [Broecheler & Getoor, 10]

LINK PREDICTION

Link prediction

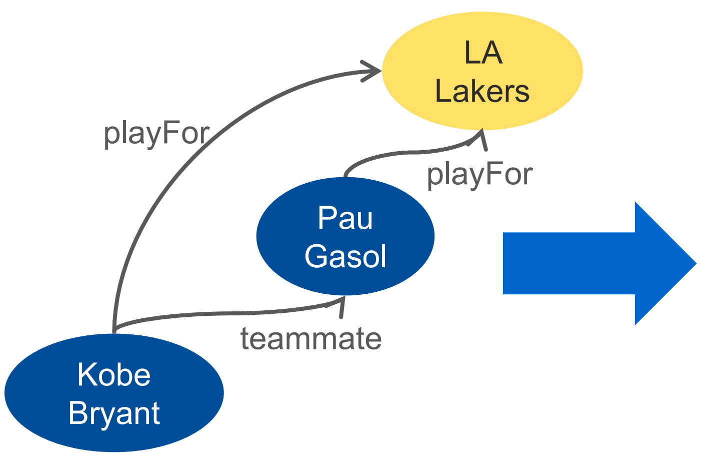


- Add knowledge from existing graph
- No external source
- Reasoning within the graph

1. Rule-based methods
2. Probabilistic models
3. Factorization models
4. Embedding models

First Order Inductive Learner

- FOIL learns function-free Horn clauses:
  - given positive negative examples of a concept
  - a set of background-knowledge predicates
  - FOIL inductively generates a logical rule for the concept that cover all + and no -

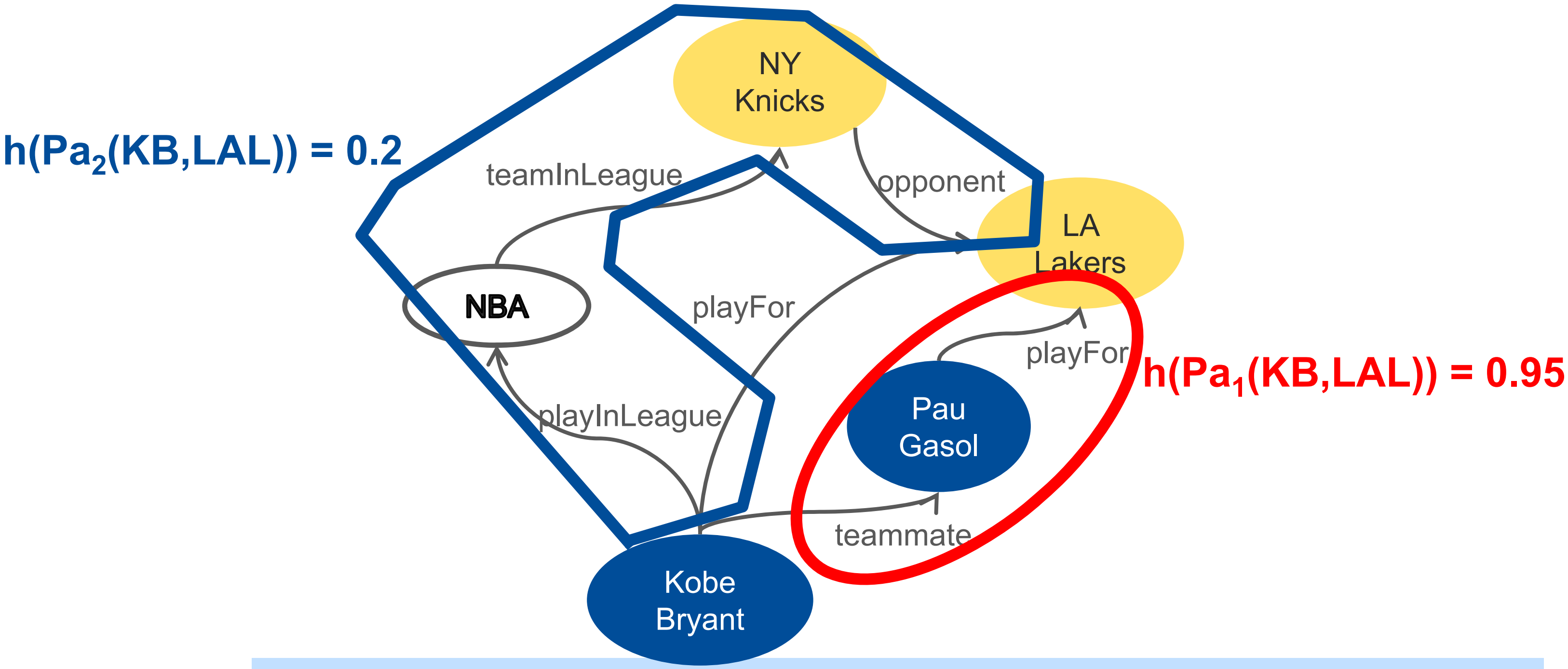


`teammate(x,y)∧ playFor(y,z) ⇒ playFor(x,z)`

- Computationally expensive: huge search space large, costly Horn clauses
- Must add constraints → high precision but low recall
- Inductive Logic Programming: deterministic and potentially problematic

Path Ranking Algorithm [Lao et al., 11]

- Random walks on the graph are used to sample paths
- Paths are weighted with probability of reaching target from source
- Paths are used as ranking experts in a scoring function



$$S(KB, playFor, LAL) = \sum_{i \in paths} \theta^{i}_{playFor} h(pa_i(KB, LAL))$$

Link prediction with scoring functions

- A scoring function alone does not grant a decision
- **Thresholding**: determine a threshold θ

  `(KB, playFor, LAL)` is *True* iff $S(KB, playFor, LAL) > \theta$

- **Ranking**:
  - The most likely relation between Kobe Bryant and LA Lakers is:

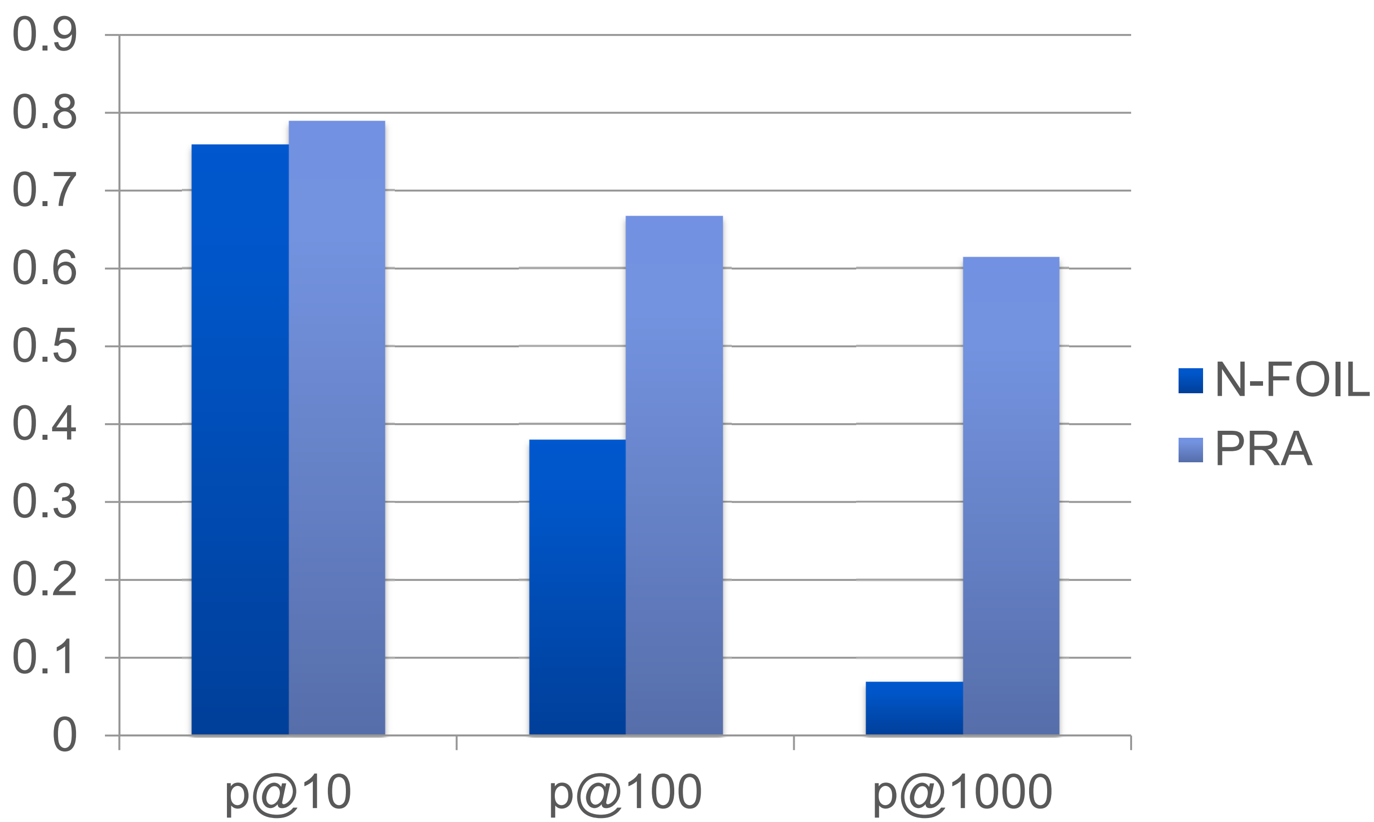
$$rel = \arg\max_{r' \in rels} S(KB, r', LAL)$$

  - The most likely team for Kobe Bryant is:

$$obj = \arg\max_{e' \in ents} S(KB, playFor, e')$$

- **As prior** for extraction models (cf. Knowledge Vault)
- No calibration of scores like probabilities

Random walks boost recall

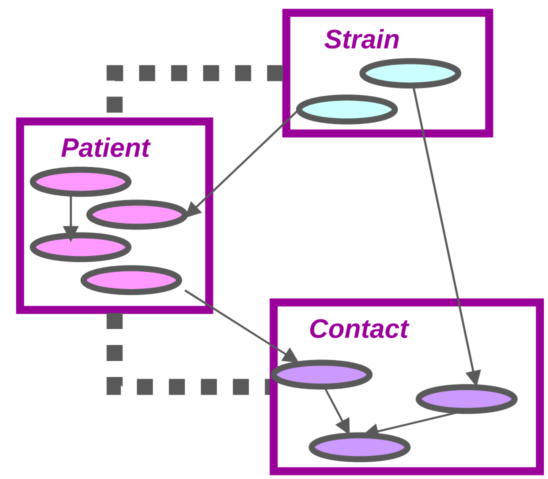


Precision of generalized facts for three levels of recall (Lao et al. 11)

Probabilistic Relational Models [Friedman et al., 99]

- **Probabilistic Relational Models** are directed graphical models that can handle link and feature uncertainty

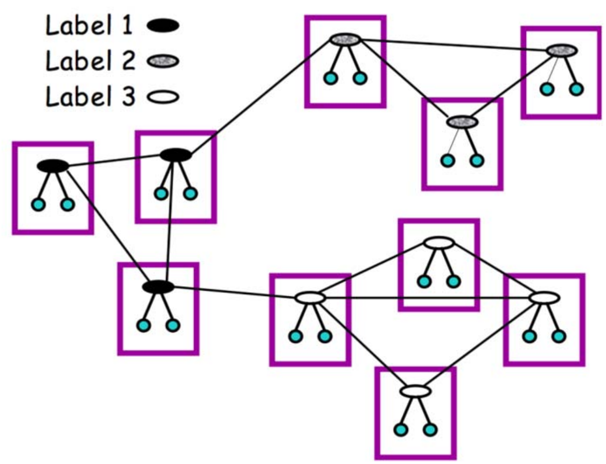
- Probabilistic inference to predict links but also duplicates, classes, clusters, etc. based on conditional probability distributions



- **Limitations:**
  - Careful construction: must avoid cycles
  - Generative process that models both observations and unknowns
  - Tractability issues

Relational Markov Networks [Taskar et al., 02]

- Discriminative model: performs inference over the unknowns only
- Discriminant function: $P(X = x) = \frac{1}{Z} \exp(\sum_i w_i f_i(x))$



- **Drawbacks:**
  - 1 feature for each state of each clique (large)
  - MAP estimation with belief propagation (slow)

Markov Logic Networks [Richardson & Domingos, 06]

- Knowledge graph = set of hard constraints on the set of possible worlds
  - Markov logic make them soft constraints
  - When a world violates a formula, it becomes less probable but not impossible
- Formulas
  - Constants: `KB, LAL, NBA`
  - Variables: `x, y` ranging over their domains (person, team, etc.).
  - Predicates: `teammate(x, y)`
  - Atom: `teammate(KB, x)`
  - Ground atom: `teammate(KB, PG)`

Number of true groundings of formula *i*

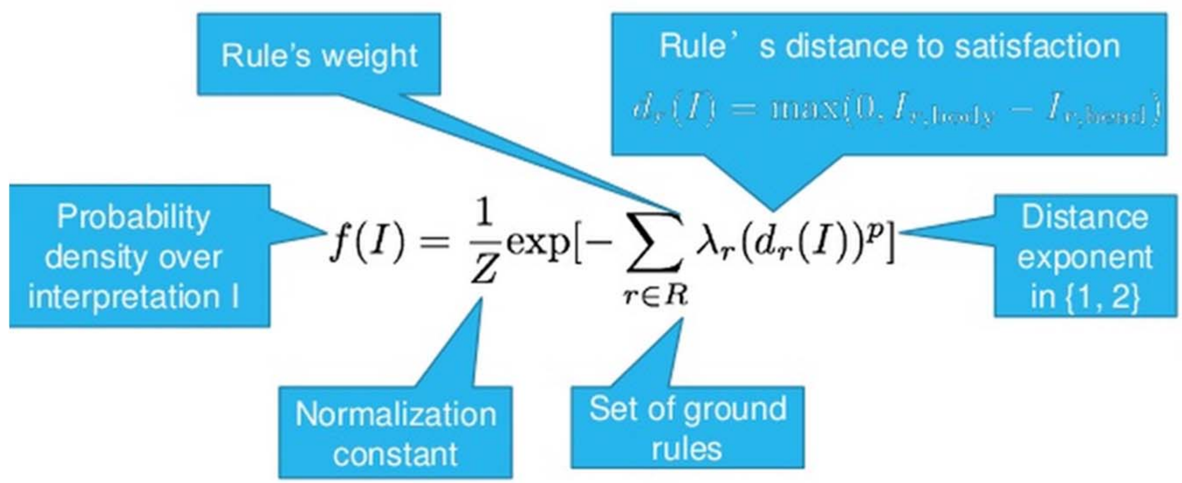
Weight of formula *i*

- A Markov Logic Network (*w, F*) is a set of weighted first-order formulas
  - Probability of a grounding *x:*
  - **Higher weight □ stronger constraint**

$$P(X = x) = \frac{1}{Z}\exp(\sum_{i \in F} w_i n_i(x))$$

Probabilistic Soft Logic [Bach et al., 13]

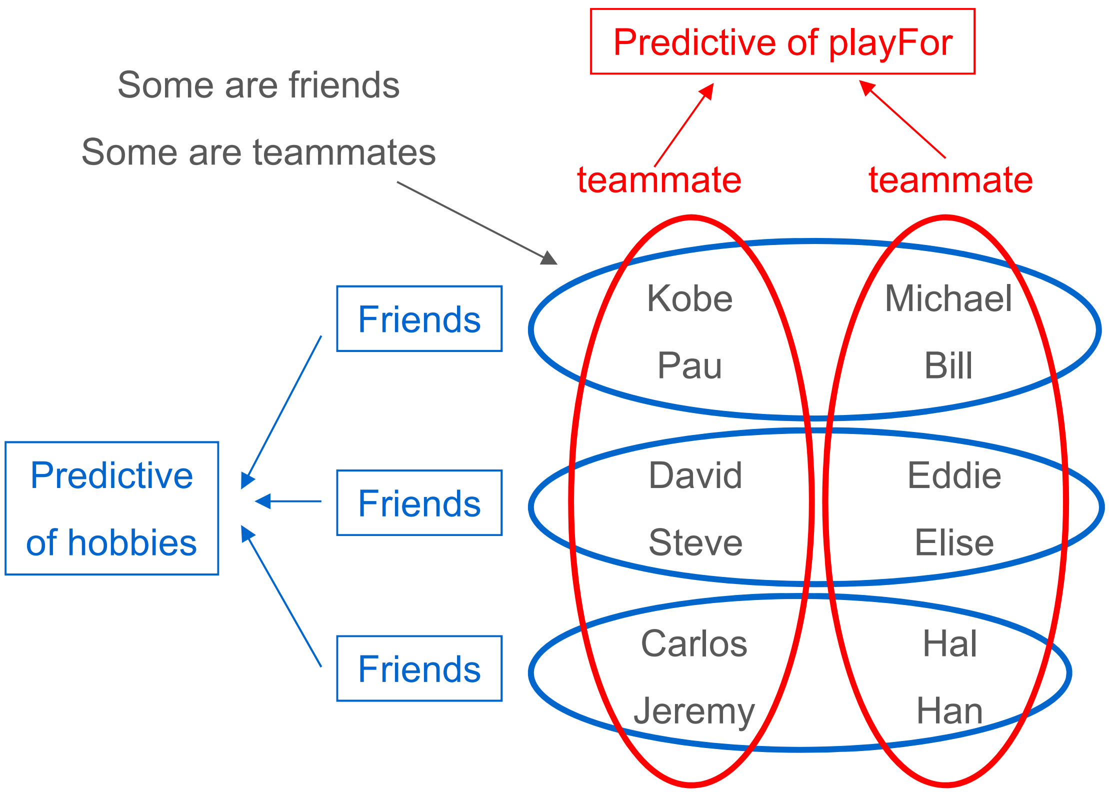
- Framework where rules have continuous truth values
- Atoms like `teammate(KB, x)` are continuous random variables
- Each predicate has a weight like in MLNs
- Probability of a grounding:



- Inference is very tractable: convex optimization problem.

Multiple Relational Clustering [Kok & Domingos, 07]

- **Hypothesis:** multiple clusterings are necessary to fully capture the interactions between entities



Multiple Relational Clustering [Kok & Domingos, 07]

- **Markov Logic framework:**
  - Create an unary predicate for each cluster e.g. `cluster22(x)`
  - Multiple partitions are learnt together
  - Use connections:
    - Cluster relations by entities they connect and vice versa
  - Use types:
    - Cluster objects of same type
    - Cluster relations with same arity and argument types
- Learning by greedy search and multiple restarts maximizing posterior
- Link prediction is determined by evaluating truth value of grounded atoms such as `playFor(KB, LAL)`

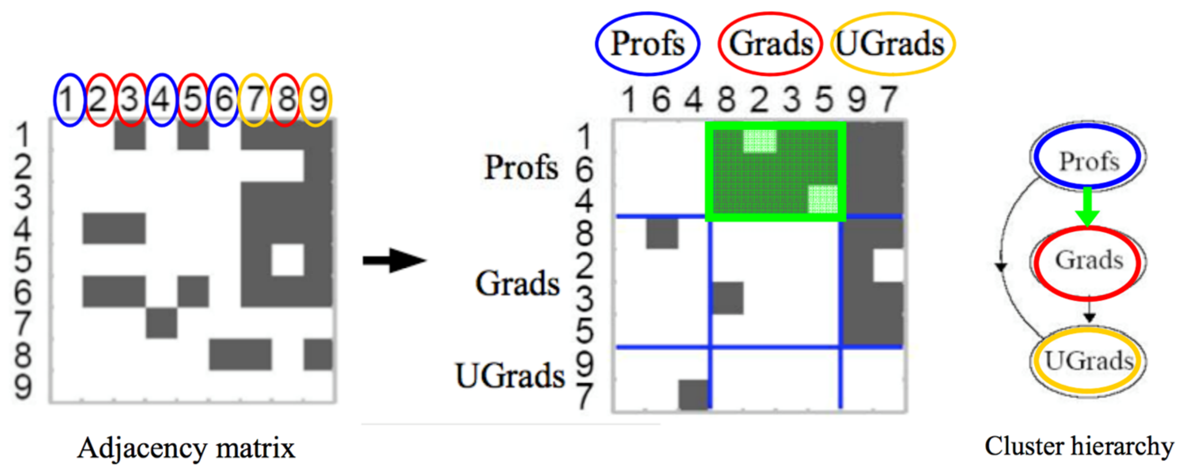
Stochastic Blockmodels [Wang & Wong, 87]

- **Blockmodels:** learn partitions of entities and of predicates
  - Partition entities/relations into subgroups based on equivalence measure.
  - For each pair of positions presence or absence of relation.
  - Structural equivalence: entities are structurally equivalent if they have identical relations to and from all the entities of the graph

- **Stochastic blockmodels:**
  - Underlying probabilistic model
  - Stochastic equivalence: two entities or predicates are stochastically equivalent if they are "exchangeable" w.r.t. the probability distribution

Infinite Relational Models [Kemp et al., 05]

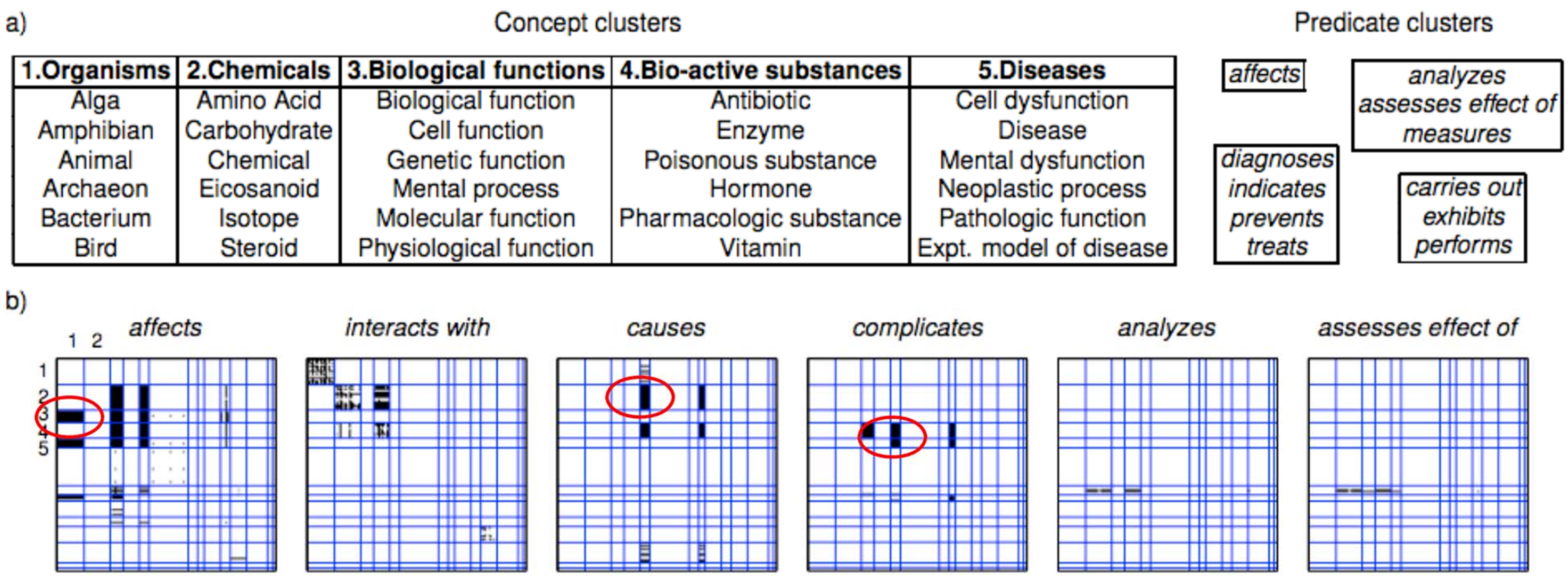
- **Infinite**: number of clusters increases as we observe more data
- **Relational**: it applies to relational data



- Prior assigns a probability to all possible partitions of the entities
- Allow number of clusters to adjust as we observe more data
- Chinese Restaurant Process: each new object is assigned to an existing cluster with probability proportional to the cluster size.

Example

- Semantic network with 135 concepts and 49 binary predicates.
- Finds 14 entities clusters and 21 predicate clusters



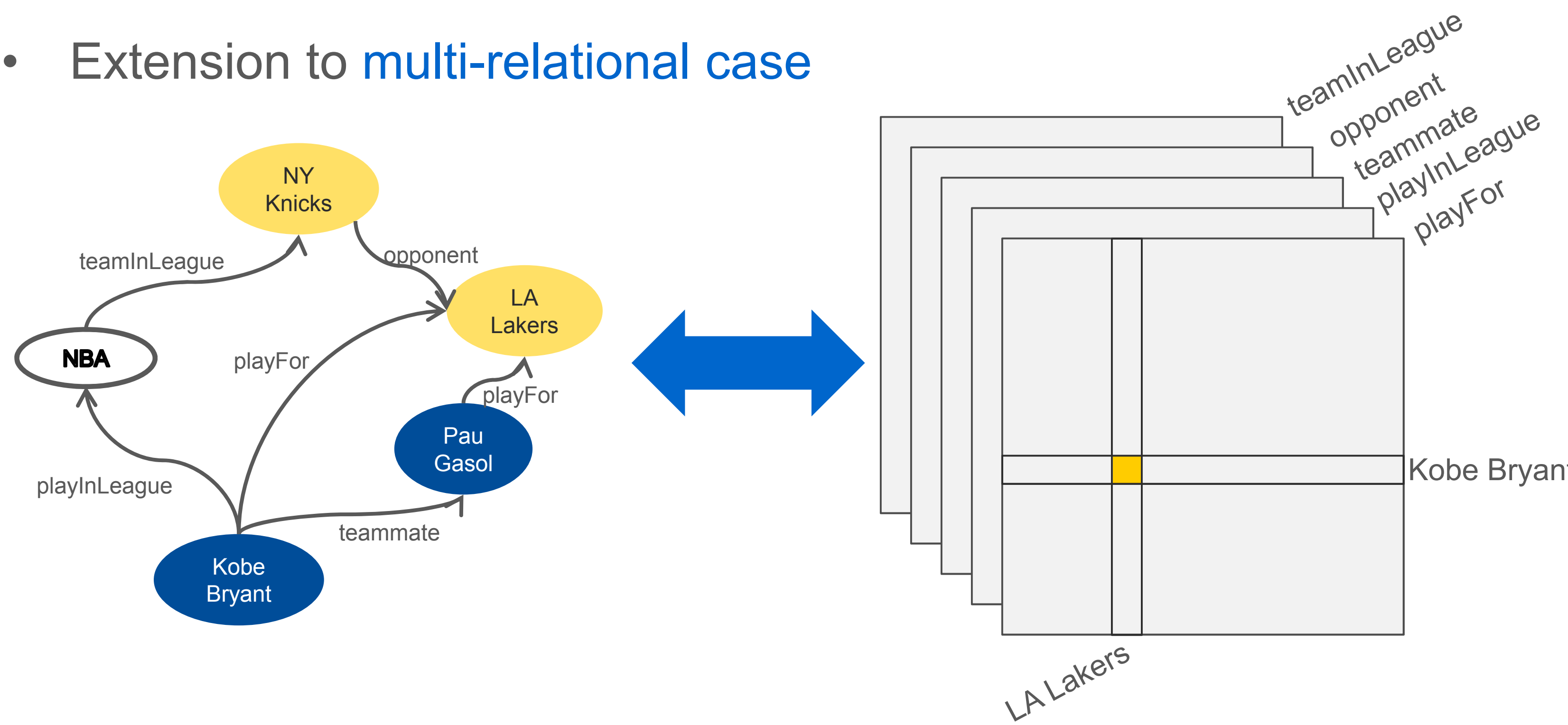
- Scalability issues with very large knowledge graphs

Variants of SBMs

- Mixed membership stochastic block models [Airoldi et al., 08]
- Nonparametric latent feature relational model [Miller et al., 09]
- Hybrid with tensor factorization [Sutskever et al., 09]

Factorization methods

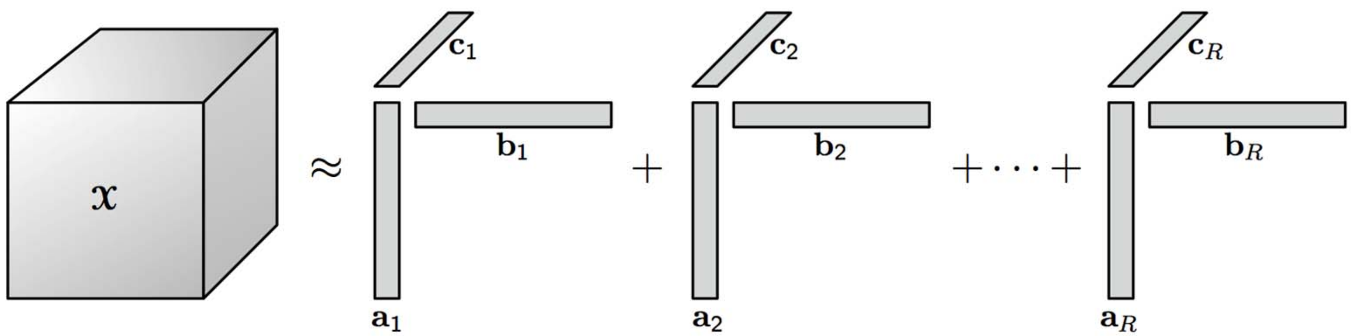
- Matrix factorization is successful: collaborative filtering, recommendation, etc.
- Extension to multi-relational case



- Collective matrix factorization or tensor factorization

Tensor factorization

- Many methods available: PARAFAC, Tucker, DEDICOM, etc.
- Example of PARAFAC [Harschman, 70]



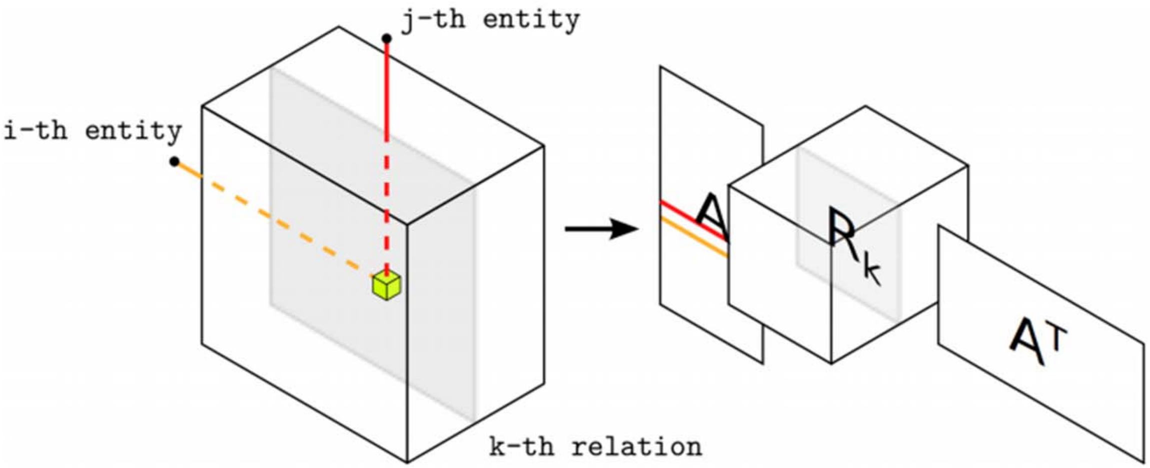
  - Decomposition as a sum of rank-one tensors

$$S(KB, playFor, LAL) = \sum_{i=1}^{R} a^i_{KB} b^i_{LAL} c^i_{playFor}$$

  - *A*, *B* and *C* are learned by alternating least squares
  - Does not take advantage of the symmetry of the tensor

RESCAL [Nickel et al., 11]

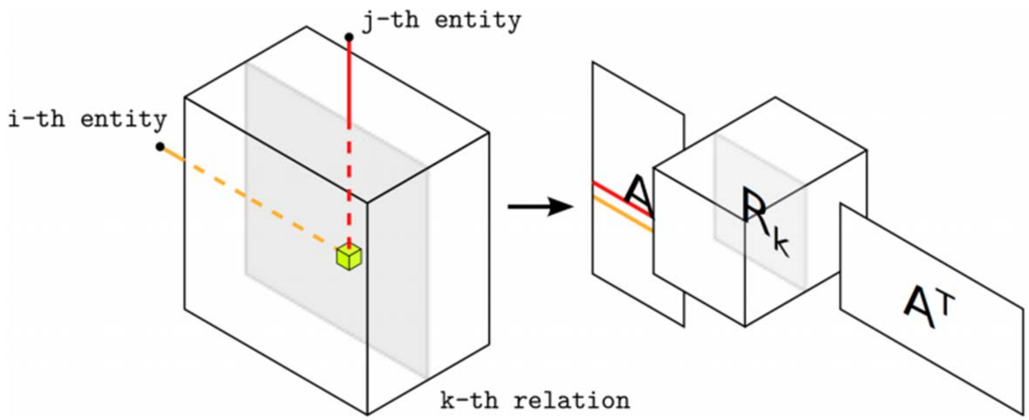
- Collective matrix factorization inspired by DEDICOM



  - A single matrix ***A*** stores latent representations of entities (vectors)
  - Matrices ***R_k*** store latent representations of relations

  - Score: $S(KB, playFor, LAL) = a_{KB} R_{playFor} a_{LAL}^T$

RESCAL [Nickel et al., 11]

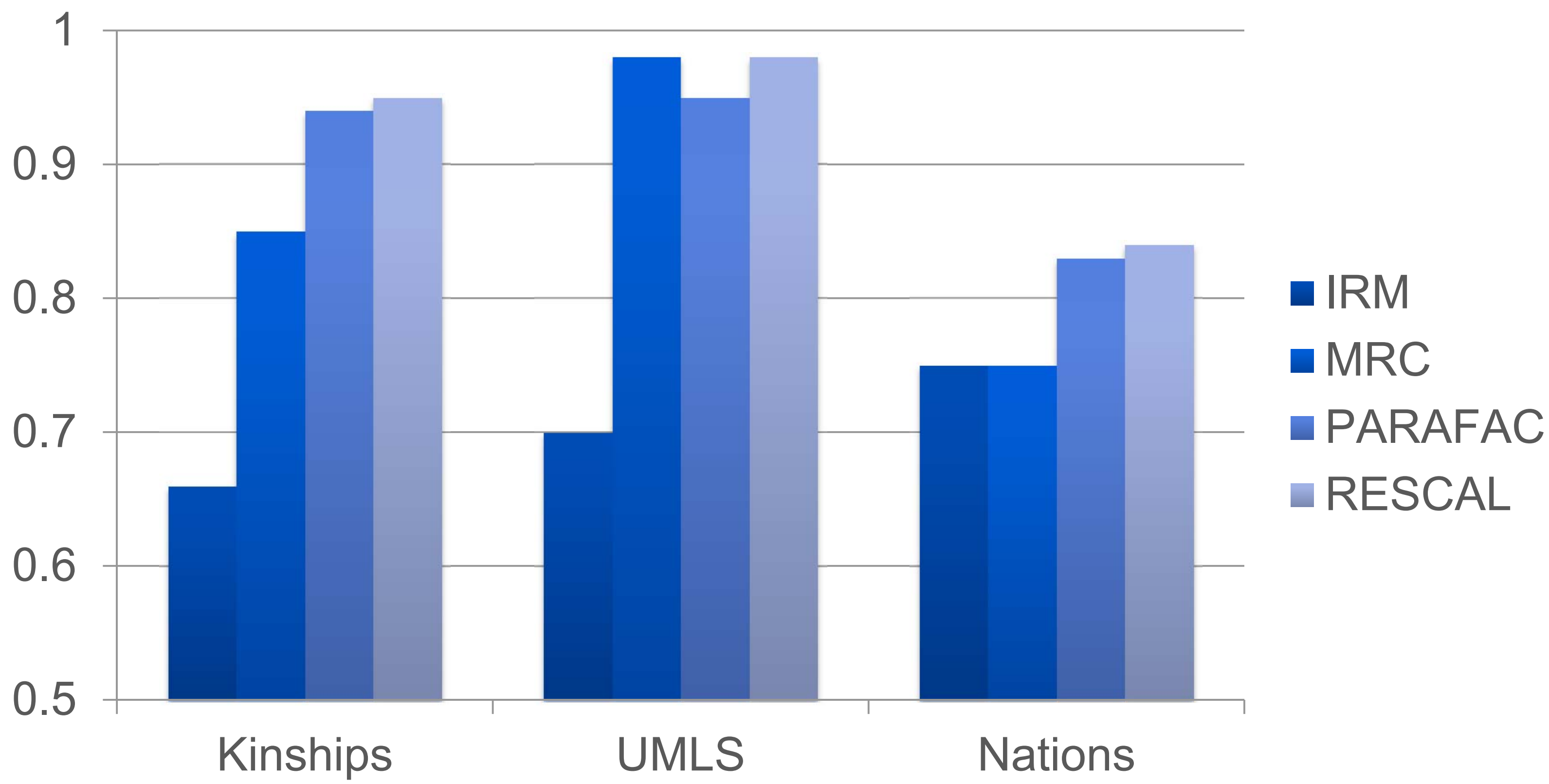


- Training with reconstruction objective:

$$\min_{A,R} \frac{1}{2}\left(\sum_k \| X_k - AR_kA^T \|_F^2\right) + \lambda_A \| A \|_F^2 + \lambda_R \sum_k \| R_k \|_F^2$$

- Optimization with alternating least squares on *A* and R_k
- Faster than PARAFAC.

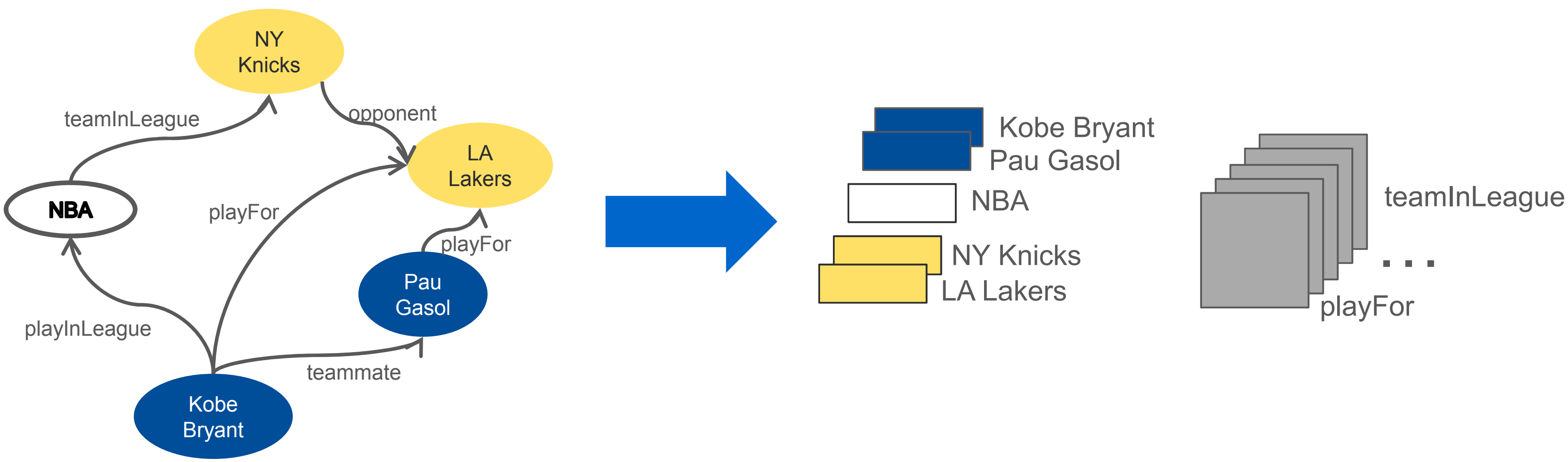
Factorization outperforms clustering



F1-score in link prediction on 3 benchmarks (Nickel et al. 11)

Embedding models

- Related to Deep Learning methods
- Entities are vectors (low-dimensional sparse)
- Relation types are operators on these vectors



- Embeddings trained to define a **similarity score** on triples such that:

$$S(KB, playFor, LAL) > S(KB, playFor, NYK)$$

Training embedding models

- Training by ranking triples from the KG vs negative (generated)
- For each triple from the training set such as `(KB, playFor, LAL)`:
  1. Unobserved facts (false?) are sub-sampled:
     - `(Kobe Bryant, opponent, LA Lakers)`
     - `(Kobe Bryant, playFor, NY Knicks)`
     - `(NBA, teammate, LA Lakers)`
     - Etc…
  2. It is checked that the similarity score of the true triple is lower:

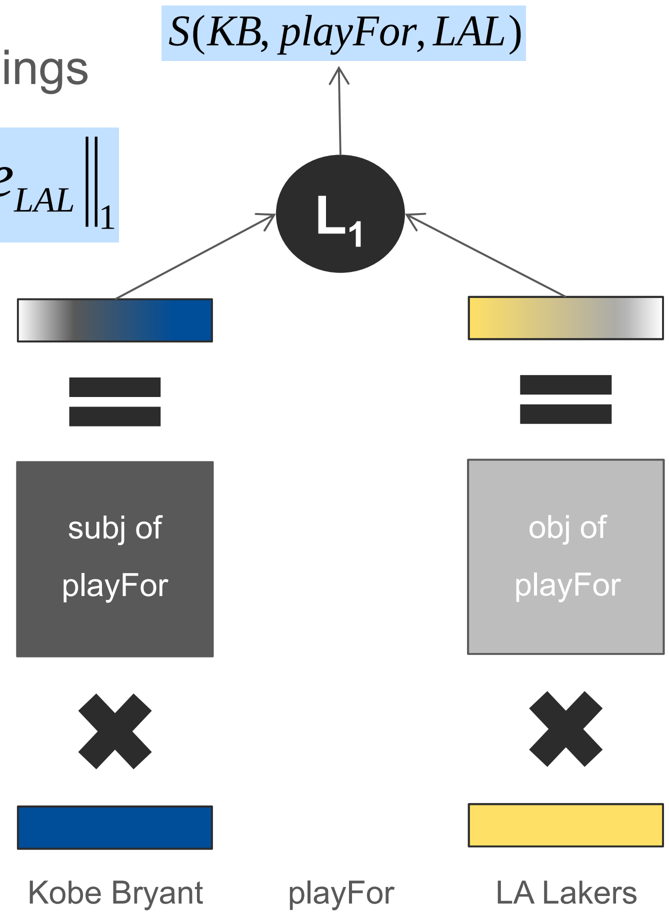
$$S(KB, playFor, LAL) > S(KB, playFor, NYK) + 1$$

  3. **If not**, parameters of the considered triples are updated.
- Optimization via Stochastic Gradient Descent

Structured Embeddings [Bordes et al., 11]

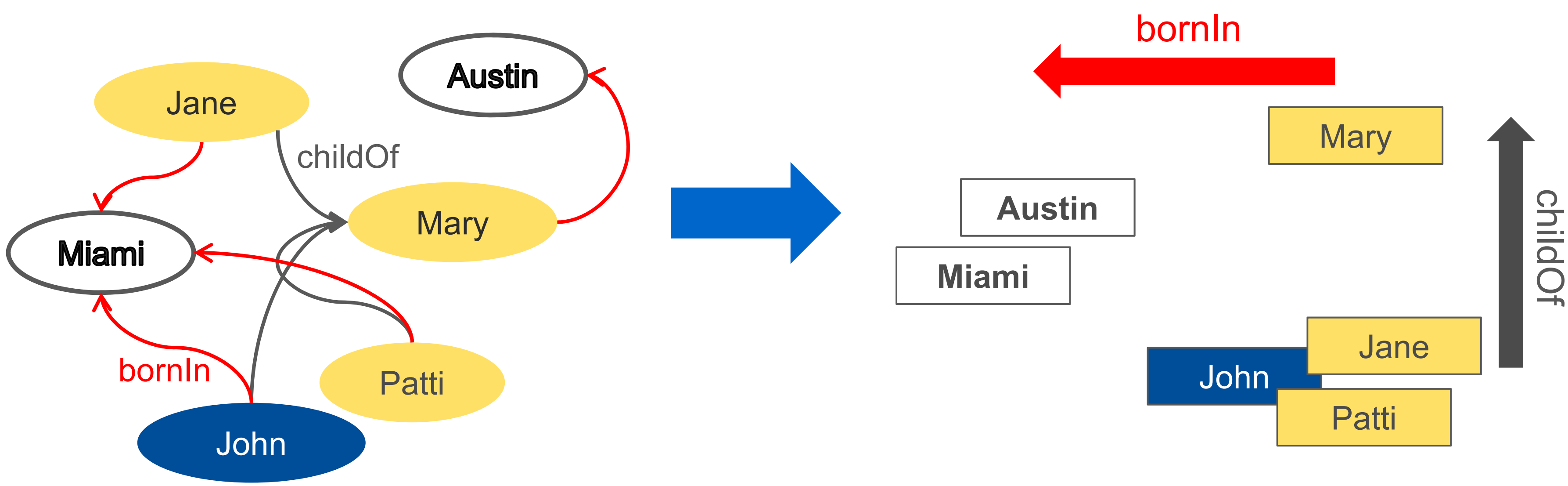
- Each entity = 1 vector
- Each relation = 2 matrices
- Score: L1 distance between projected embeddings

$$S(KB, playFor, LAL) = \left\| M_{playFor}^{sub} e_{KB} - M_{playFor}^{obj} e_{LAL} \right\|_1$$



Translating Embeddings [Bordes et al. 13]

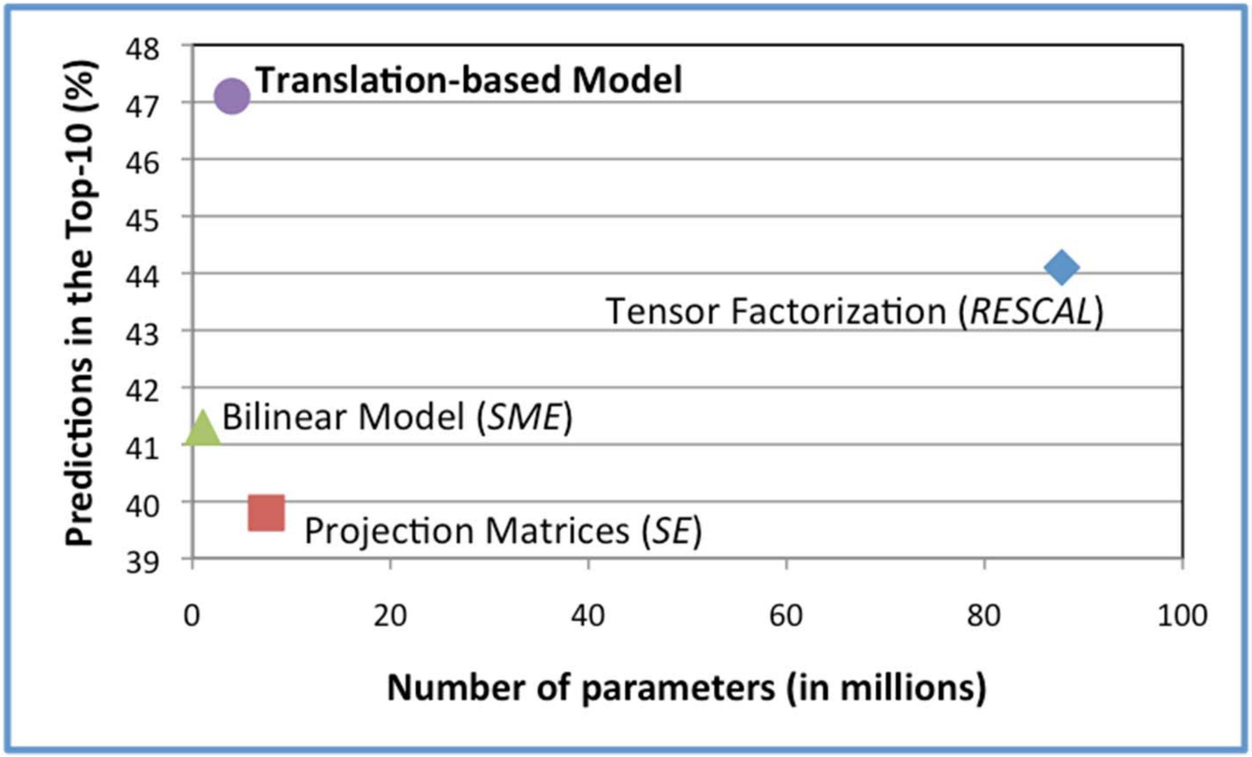
- *Simpler model:* relation types are translation vectors



$$S(john, bornIn, miami) = \left\| e_{john} + e_{bornIn} - e_{miami} \right\|_2$$

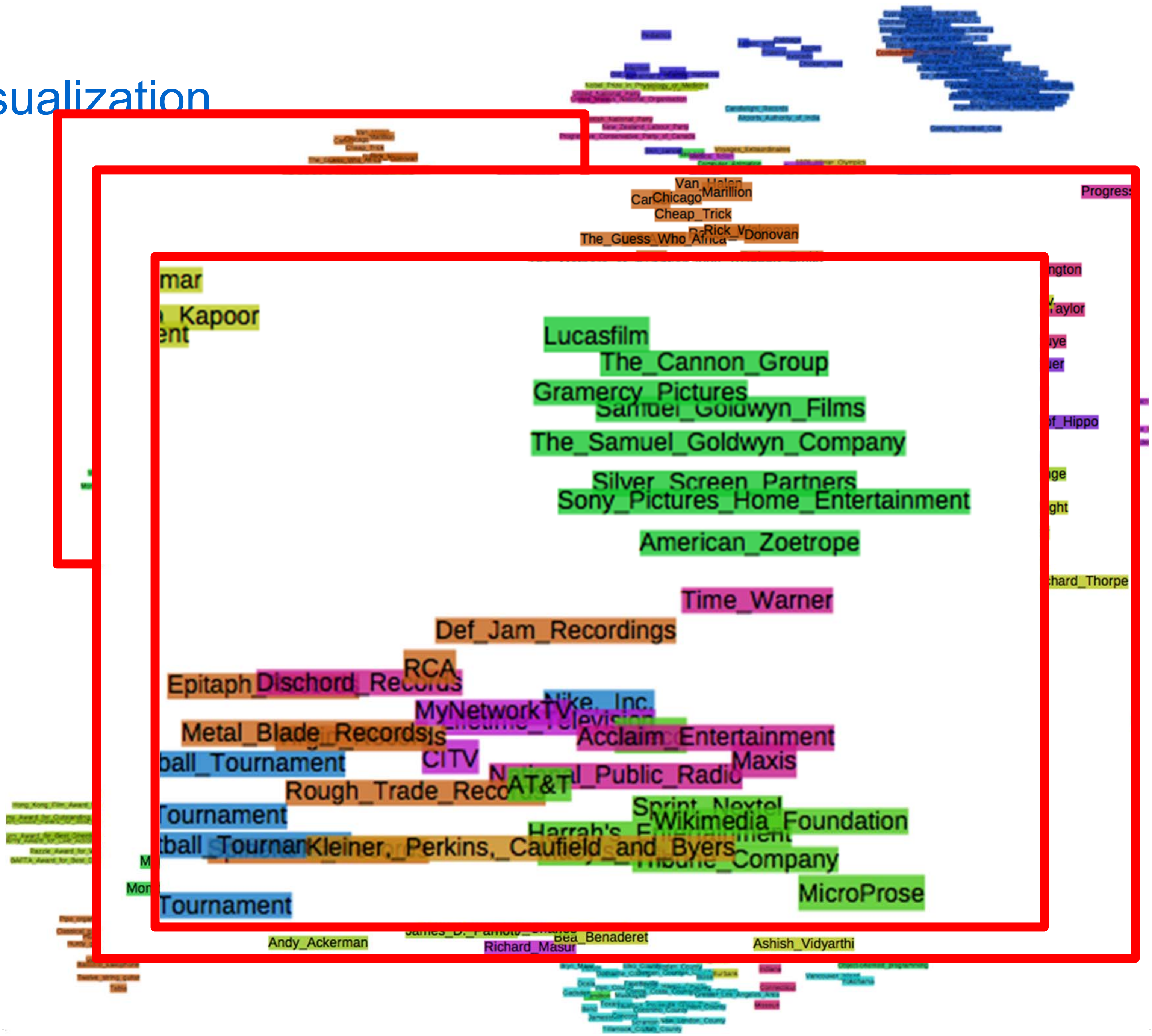
- Much fewer parameters (1 vector per relation).

The simpler, the better



Ranking object entities on a subset of Freebase [Bordes et al. 13]

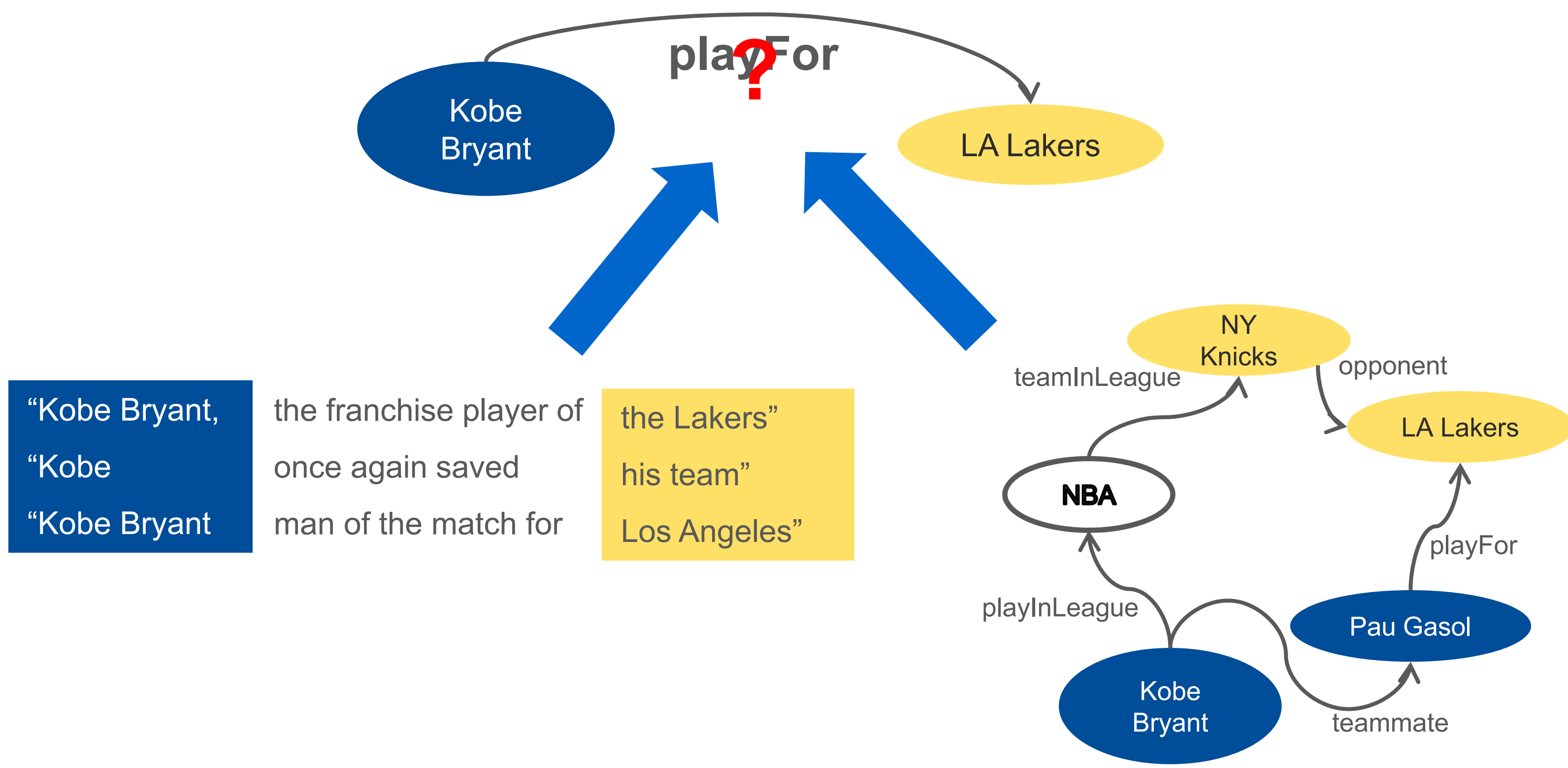
Visualization



Using knowledge graph and text together

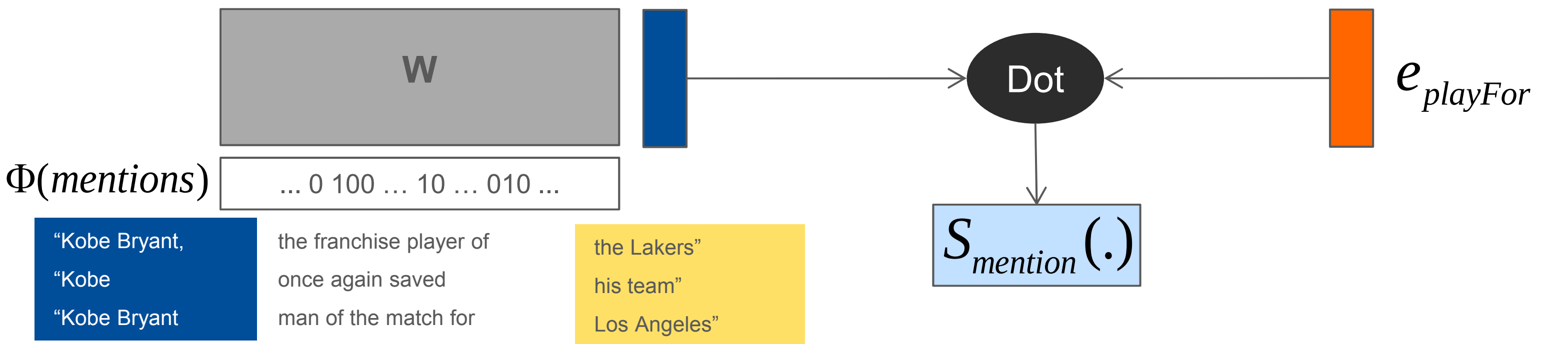
Why not merging relation extraction and link prediction in the same model?

Extracted facts should agree **both with the text and the graph!**

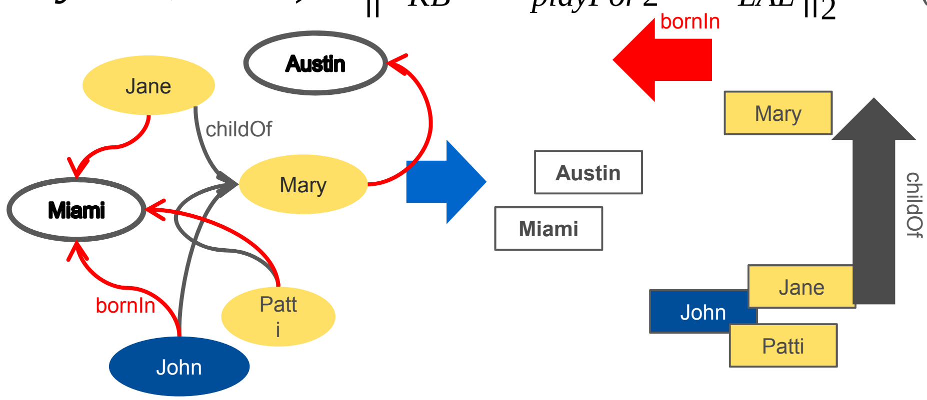


Joint embedding models [Bordes et al., 12; Weston et al., 13]

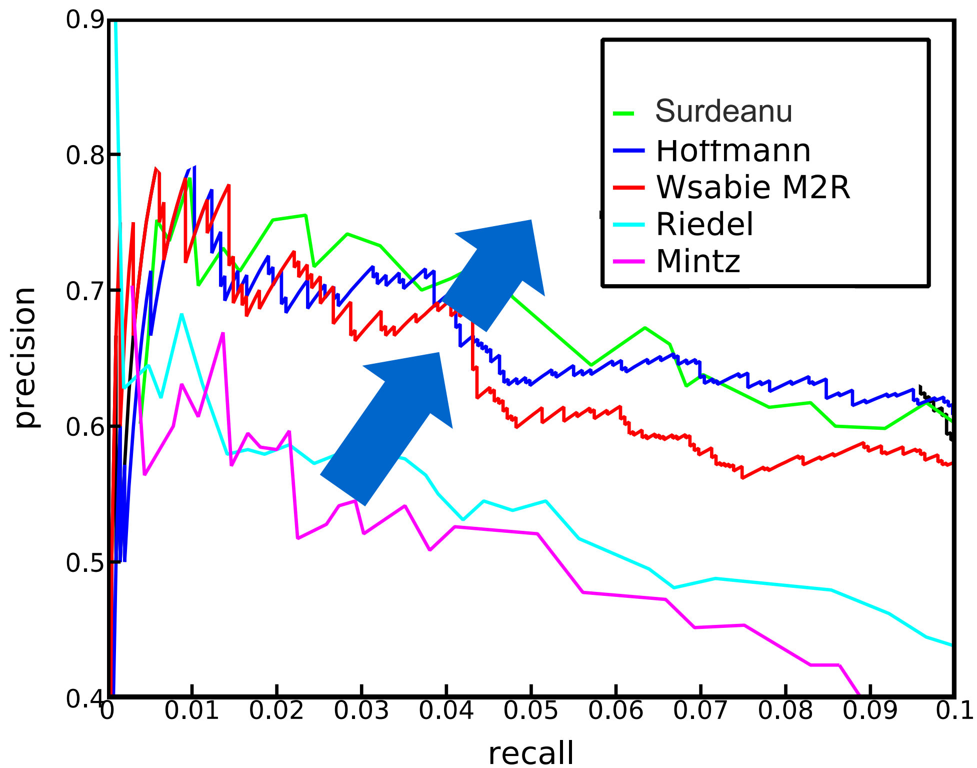
- Combination of two scores: $\boxed{S(.) = S_{text}(.) + S_{FB}(.)}$ (trained separately)
- $S_{text}(KB, playFor, LAL) = \left\langle W^T \Phi(m), e_{playFor1} \right\rangle$ inspired by WSABIE (Weston et al., 10)



- $S_{FB}(KB, playFor, LAL) = \left\| e_{KB} + e_{playFor2} - e_{LAL} \right\|_2$ (translating embeddings)



Using stored information improves precision even more

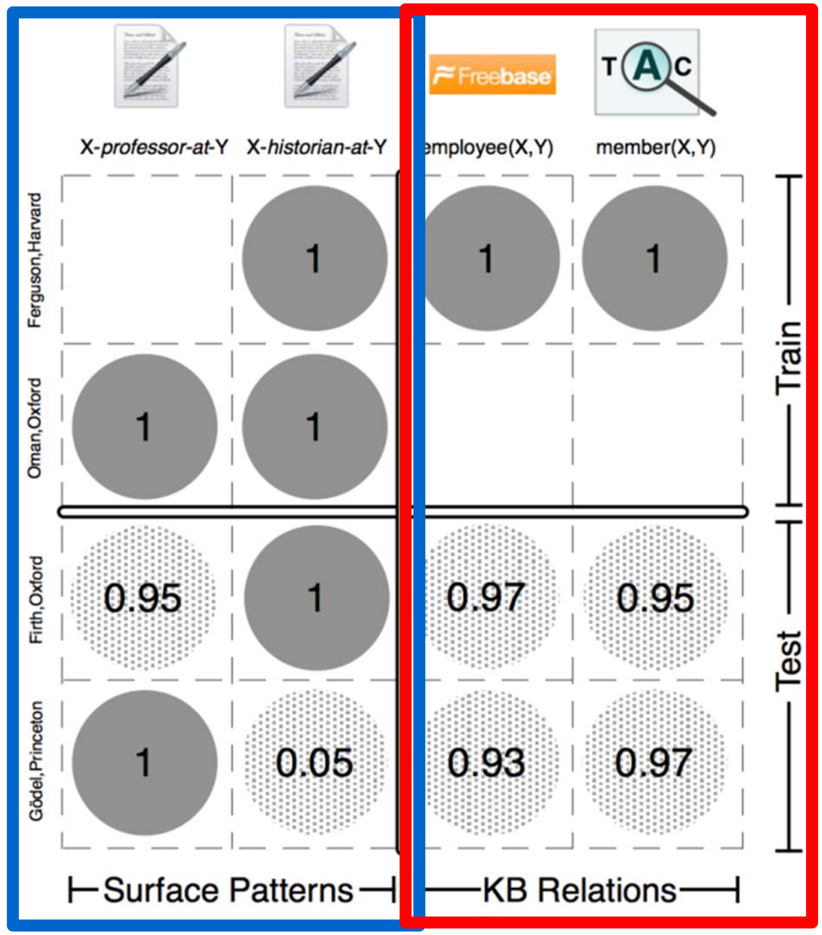


Precision-recall curves on extracting from New York Times articles to Freebase [Weston et al., 13]

Universal schemas [Riedel et al., 13]

- Join in a single learning problem link prediction and relation extraction

- The same model can score triples made of entities linked with:
  - extracted surface forms from text
  - predicates from a knowledge base



Universal schemas [Riedel et al., 13]

- Combination of three scores: $\boxed{S(.) = S_{mention}(.) + S_{FB}(.) + S_{neighbors}(.)}$

$$S_{mention}(KB, playFor, LAL) = \left\langle e_{mention}, e_{playFor1} \right\rangle$$

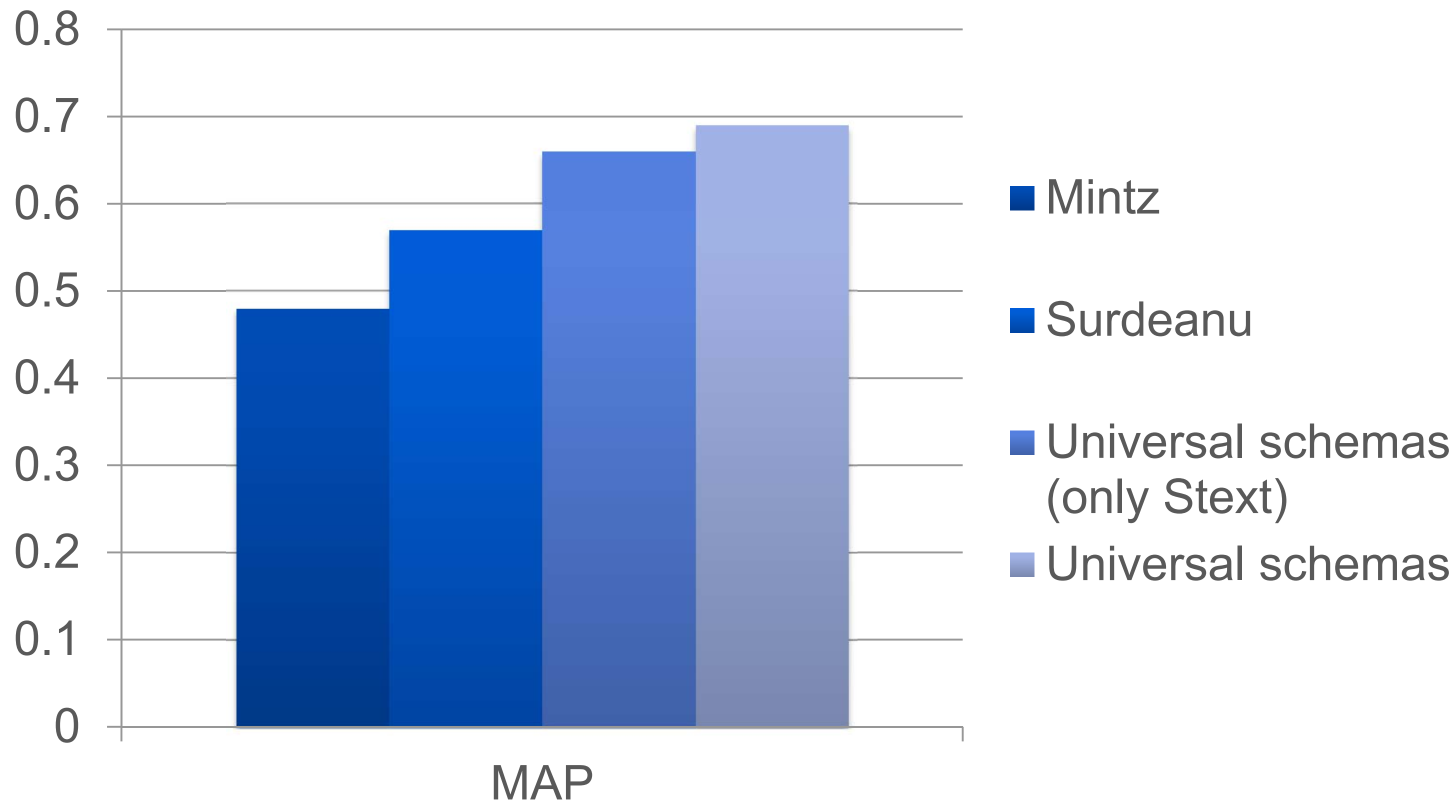
$$S_{FB}(KB, playFor, LAL) = \left\langle e^{sub}_{playFor2}, e_{KB} \right\rangle + \left\langle e^{obj}_{playFor2}, e_{LAL} \right\rangle$$

$$S_{neighbors}(KB, playFor, LAL) = \sum_{\substack{(KB, rel', LAL) \\ rel' \neq playFor}} w^{playFor}_{rel'}$$

- Embeddings for entities, relations and mentions.

- Training by ranking observed facts versus others and making updates using Stochastic Gradient Descent.

Using stored information (still) improves precision



Weighted Mean Averaged Precision on a subset of relations of Freebase [Riedel et al. 13]

RESOURCES

Related tutorial – here at KDD (later today) !

Bringing **Structure** to **Text**: Mining Phrases, Entity Concepts, Topics & Hierarchies

by Jiawei Han, Chi Wang and Ahmed El-Kishky

Today, 2:30pm

Relevant datasets

- Wikipedia
  - http://en.wikipedia.org/wiki/Wikipedia:Database_download
- Freebase
  - https://developers.google.com/freebase/data
- YAGO
  - http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/
- DBpedia
  - http://wiki.dbpedia.org/Datasets
- OpenIE/Reverb
  - http://reverb.cs.washington.edu/

Relevant competitions, evaluations, and workshops

- Knowledge Base Population (KBP) @ TAC

  http://www.nist.gov/tac/2014/KBP/

- Knowledge Base Acceleration (KBA) @ TREC

  http://trec-kba.org/

- Entity Recognition and Disambiguation (ERD) Challenge @ SIGIR 2014

  http://web-ngram.research.microsoft.com/erd2014/

- INEX Link the Wiki track

  http://link.springer.com/chapter/10.1007/978-3-642-23577-1_22

- CLEF eHealth Evaluation Lab

  http://link.springer.com/chapter/10.1007/978-3-642-40802-1_24

Relevant competitions, evaluations, and workshops (cont'd)

- Named Entity Extraction & Linking (NEEL) Challenge (#Microposts2014)

  http://www.scc.lancs.ac.uk/microposts2014/challenge/

- LD4IE 2014 Linked Data for Information Extraction

  http://trec-kba.org/

Tutorials

- Entity linking and retrieval tutorial (Meij, Balog and Odijk)
  - http://ejmeij.github.io/entity-linking-and-retrieval-tutorial/
- Entity resolution tutorials (Getoor and Machanavajjhala)
  - http://www.umiacs.umd.edu/~getoor/Tutorials/ER_VLDB2012.pdf
  - http://linqs.cs.umd.edu/projects/Tutorials/ER-AAAI12/Home.html
- Big data integration (Dong and Srivastava)
  - http://lunadong.com/talks/BDI_vldb.pptx
- Tensors and their applications in graphs (Nickel and Tresp)
  - http://www.cip.ifi.lmu.de/~nickel/iswc2012-learning-on-linked-data/
- Probabilistic soft logic (Bach et Getoor)
  - http://psl.umiacs.umd.edu/

Data releases from Google

1. Automatic annotation of ClueWeb09 and ClueWeb12 with Freebase entities (**800M documents, 11B entity mentions**)
2. Similar annotation of several TREC query sets (**40K queries**)
3. Human judgments of relations extracted from Wikipedia (**50K instances, 250K human judgments**)
4. Triples deleted from Freebase over time (**63M triples**)

Mailing list:

goo.gl/MJb3A

SUMMARY

Knowledge is crucial yet difficult to acquire

- Knowledge is crucial for many AI tasks
- Knowledge acquisition
  - From experts: slow and mostly reliable
  - From non-experts: faster and not always reliable
  - Automated: fastest and most scalable, yet noisiest
- Knowledge availability
  - A lot can be found online
  - A lot cannot be found
  - A lot cannot be extracted using today's methods

Where we are today

- We can extract a lot of knowledge from text and model its correctness
- Enforcing structure makes the extraction problem easier yet imposes limitations
- Leveraging existing knowledge repositories helps a lot

Next steps

- We need new extraction methods, from new sources
- Extracting from modalities other than text appears promising yet mostly unexplored

Plenty to be learned, problems are far from solved!

- Vibrant research area
- Numerous open research questions

This is a perfect time to work in this area!