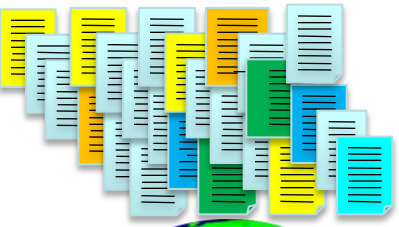
max planck institut
informatik

From Names and Phrases to Entities and Relations

Gerhard Weikum
Max Planck Institute for Informatics
& Saarland University
http://www.mpi-inf.mpg.de/~weikum/

From Language to Knowledge

and More and Deeper Knowledge



knowledge acquisition

Language

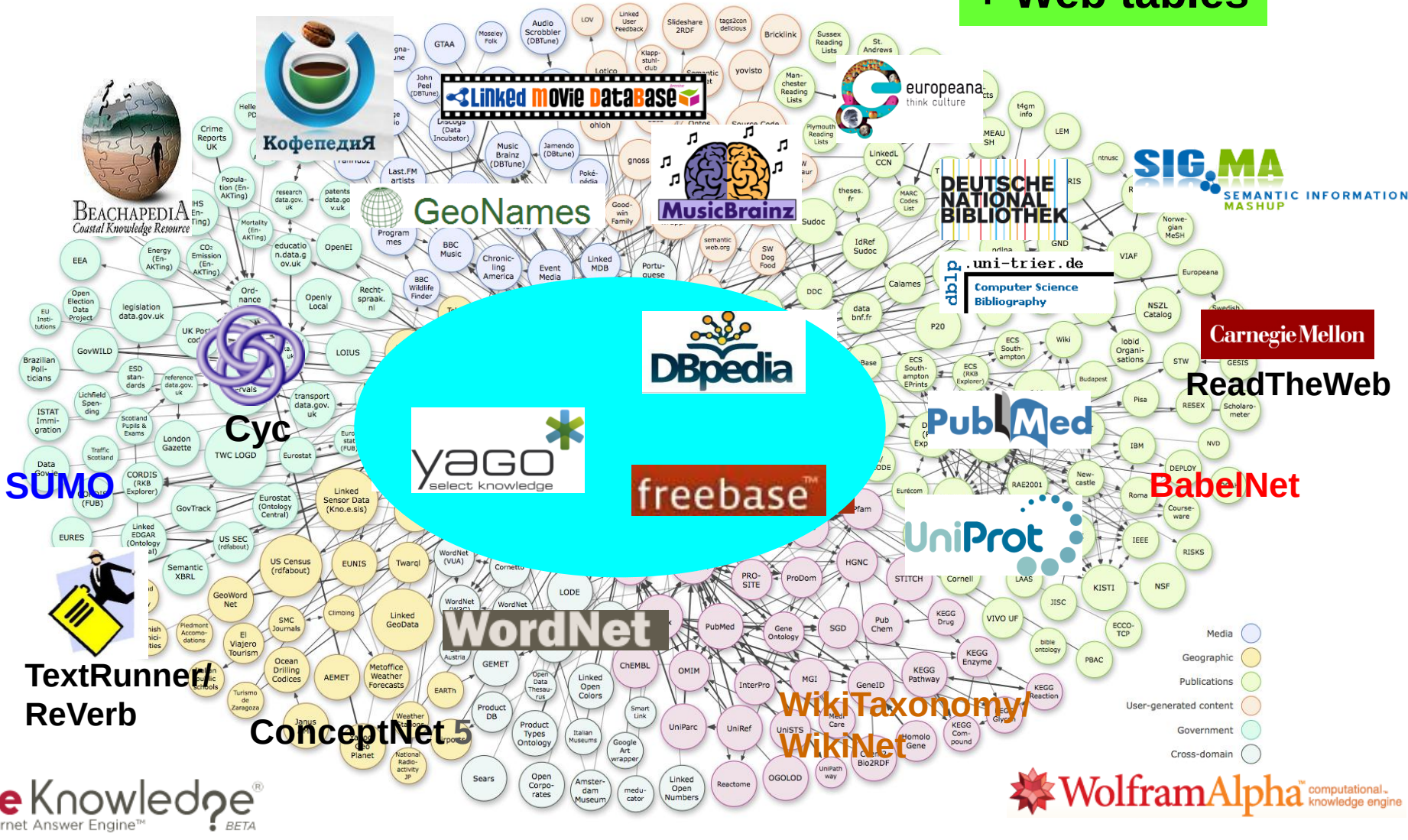
Knowledge

intelligent interpretation

Web of Data & Knowledge

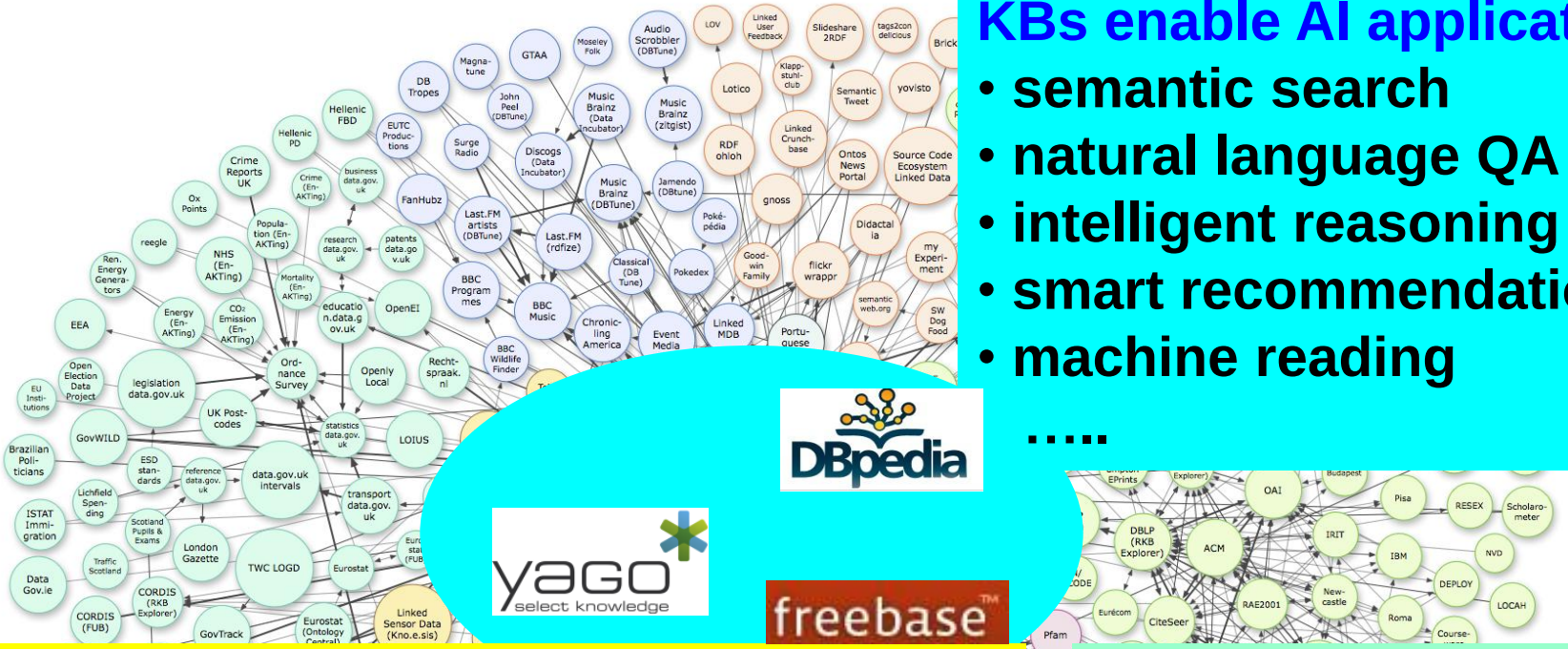
62 Bio. SPO triples (RDF) from 870 sources, and growing

+ Web tables



http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.png

Knowledge for Intelligence



KBs enable AI applications:

- semantic search
- natural language QA
- intelligent reasoning
- smart recommendations
- machine reading

.....

Yimou_Zhang type movie_director
Yimou_Zhang type olympic_games_participant
movie_director subclassOf artist
Yimou_Zhang directed Flowers_of_War
Christian_Bale actedIn Flowers_of_War
id11: Yimou_Zhang memberOf Beijing_film_academy
id11 validDuring [1978, 1982]
Yimou_Zhang „was classmate of“ Kaige_Chen
Yimou_Zhang „had love affair with“ Li_Gong
Li_Gong knownAs „China‘s most beautiful“

taxonomic knowledge

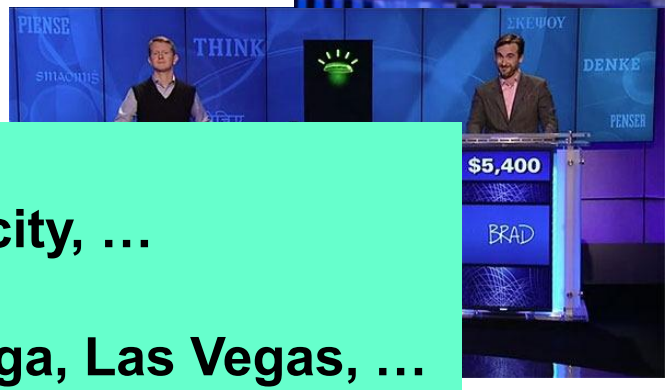
factual knowledge

temporal knowledge

open knowledge

terminological knowledge

Use Case: Question Answering



This town is known as "Sin City" & its downtown is "Glitter Gulch"

Q: Sin City ?
→ movie, graphical novel, nickname for city, …
A: Vegas ? Strip ?
→ Vega (star), Suzanne Vega, Vincent Vega, Las Vegas, …
→ comic strip, striptease, Las Vegas Strip, …

This American city has two airports named after a war hero and a WW II battle

question classification & decomposition

knowledge back-ends

D. Ferrucci et al.: Building Watson. AI Magazine, Fall 2010.
IBM Journal of R&D 56(3/4), 2012: This is Watson.

Use Case: Question Answering

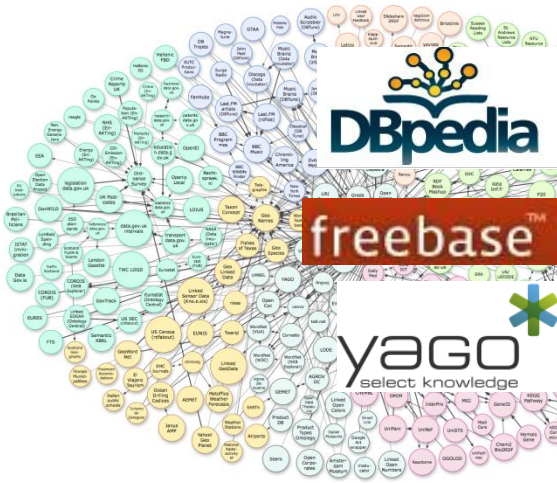
This town is known as "Sin City" & its downtown is "Glitter Gulch"

Q: Sin City ?
→ movie, graphical novel, nickname for city, …
A: Vegas ? Strip ?
→ Vega (star), Suzanne Vega, Vincent Vega, Las Vegas, …
→ comic strip, striptease, Las Vegas Strip, …

question

structured query

```
Select ?t Where {
  ?t type location .
  ?t hasLabel "Sin City" .
  ?t hasPart ?d .
  ?d hasLabel "Glitter Gulch" . }
```



Linked Data
Big Data
Web tables

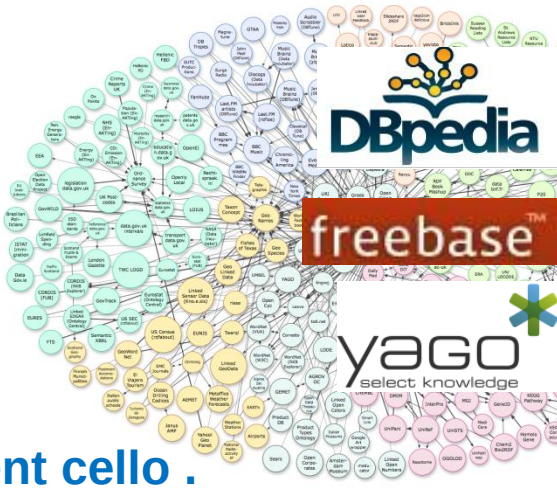
Use Case: Question Answering

Which classical cello player covered a composition from The Good, the Bad, the Ugly?

Q: Good, Bad, Ugly ?
covered ?

A: western movie ? Big Data – NSA - Snowden ?
played ? performed ?

question → ? → structured query



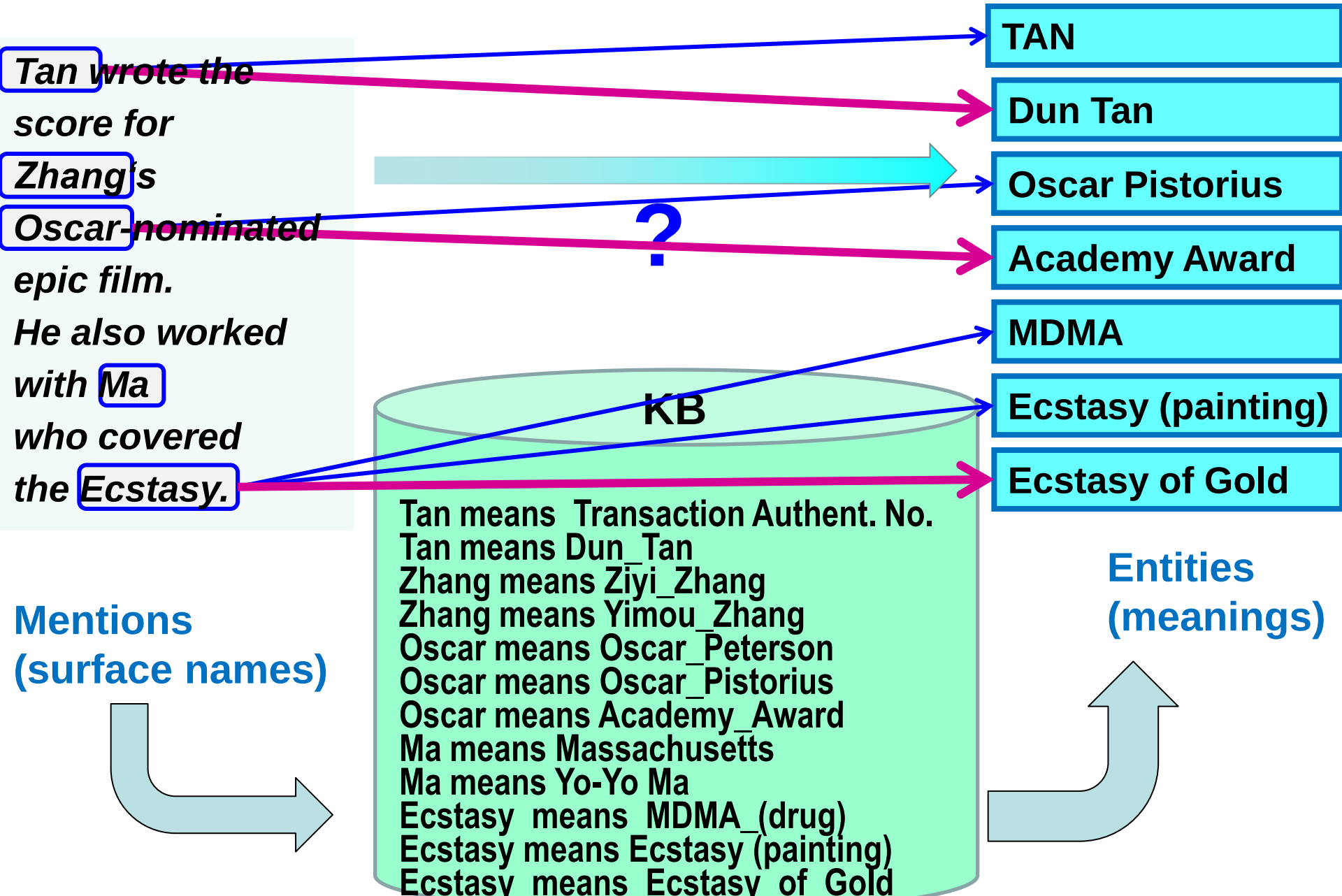

```
Select ?m Where {
  ?m type musician . ?m playsInstrument cello .
  ?m performed ?c .  ?c partOf ?f .
  ?f type movie .
  ? hasLabel "The Good, the Bad, the Ugly". }
```

Linked Data
Big Data
Web tables

Outline

- ✓ Motivation
- ★ Entity Name Disambiguation
- ★ Relational Paraphrases
- ★ Translating Questions into Queries
- ★ Wrap-Up

Named Entity Disambiguation



Mention-Entity Graph

weighted undirected graph with two types of nodes

Tan wrote the score for Zhang's Oscar-nominated epic film. He also worked with Ma who covered the Ecstasy.

bag-of-words or language model: words, bigrams, phrases

TAN

Dun Tan

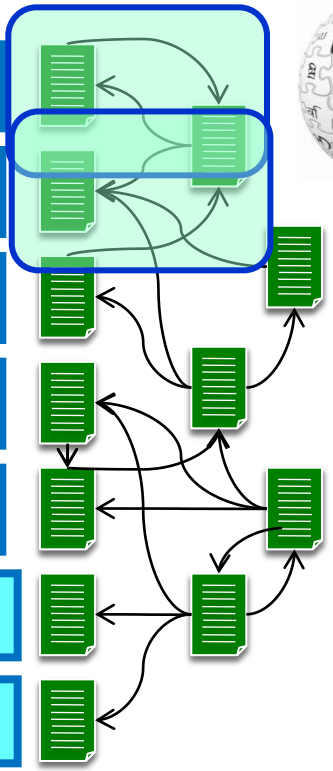
Oscar Pistorius

Academy Award

MDMA

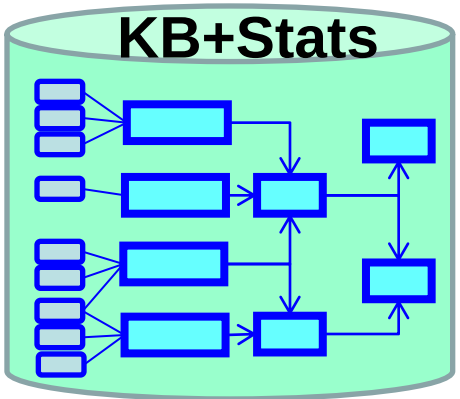
Ecstasy (painting)

Ecstasy of Gold



Similarity (m,e):

- **cos/Dice/KL (context(m), context(e))**



Mention-Entity Graph

weighted undirected graph with two types of nodes

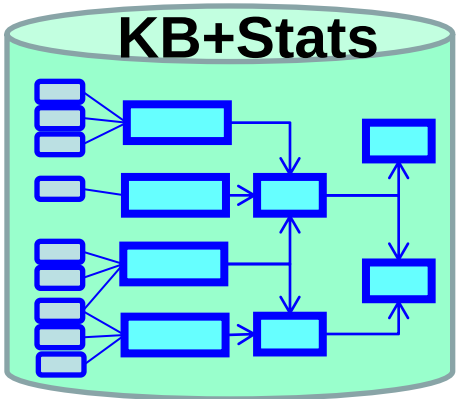
Tan wrote the score for Zhang's Oscar-nominated epic film. He also worked with Ma who covered the Ecstasy.

joint mapping

- TAN
- Dun Tan
- Oscar Pistorius
- Academy Award
- MDMA
- Ecstasy (painting)
- Ecstasy of Gold

Similarity (m,e):

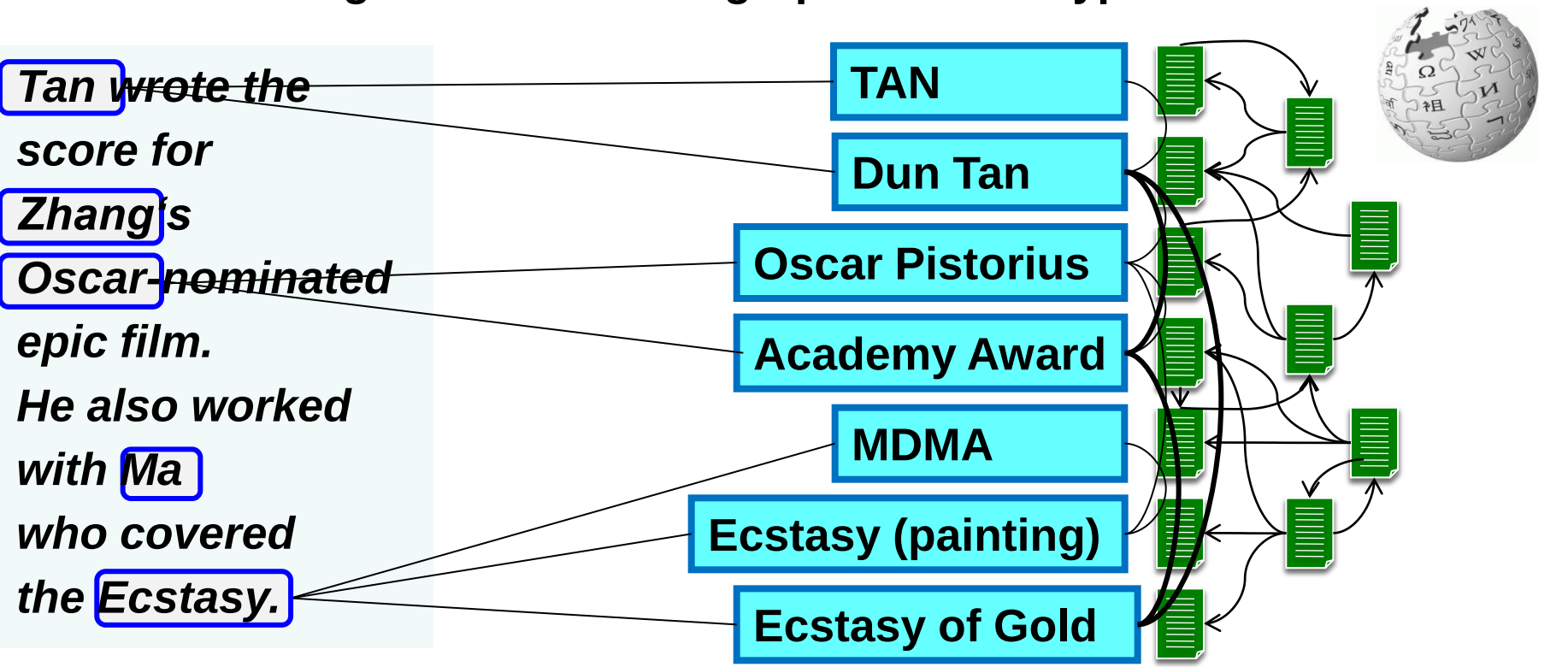
- cos/Dice/KL (context(m), context(e))



Mention-Entity Graph

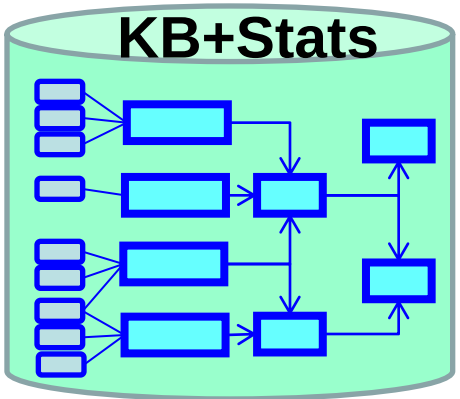
weighted undirected graph with two types of nodes

Tan wrote the score for Zhang's Oscar-nominated epic film. He also worked with Ma who covered the Ecstasy.



Similarity (m,e):

- cos/Dice/KL (context(m), context(e))



Coherence (e,e‘):

- dist(types)
- overlap(links)
- overlap (keyphrases)

Mention-Entity Graph

weighted undirected graph with two types of nodes

Tan wrote the score for Zhang's Oscar-nominated epic film. He also worked with Ma who covered the Ecstasy.

- TAN
- Dun Tan
- Oscar Pistorius
- Academy Award
- MDMA
- Ecstasy (painting)
- Ecstasy of Gold

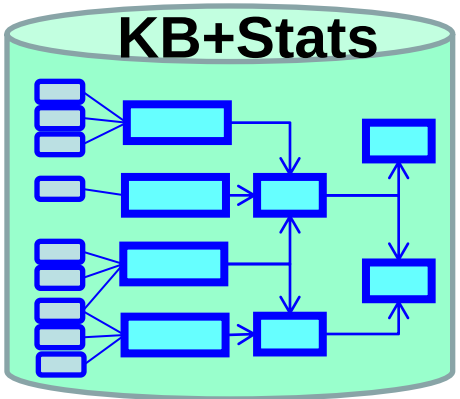
Chinese violinists
film music composers
artists
Academy Award winners

American film awards
annual awards
honors

Metallica songs
Ennio Morricone songs
artifacts
soundtrack music

Similarity (m,e):

- cos/Dice/KL (context(m), context(e))



Coherence (e,e‘):

- dist(types)
- overlap(links)
- overlap (keyphrases)

Mention-Entity Graph

weighted undirected graph with two types of nodes

Tan wrote the score for Zhang's Oscar-nominated epic film. He also worked with Ma who covered the Ecstasy.

- TAN
- Dun Tan
- Oscar Pistorius
- Academy Award
- MDMA
- Ecstasy (painting)
- Ecstasy of Gold

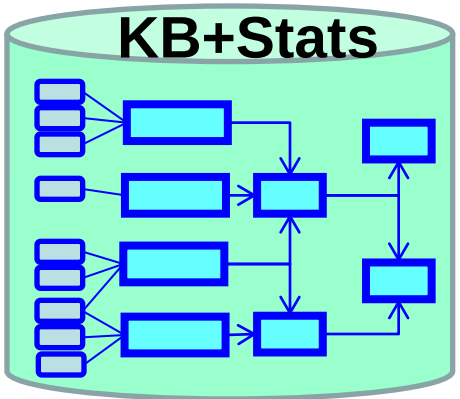
http://.../wiki/Hero_(movie)
http://.../wiki/Crouching_Tiger, Hidd
http://.../wiki/Yo-Yo_Ma
http://.../wiki/Academy_Award_winn

http://.../wiki/American_film_awards
http://.../wiki/Hollywood
http://.../wiki/Crouching_Tiger, Hidd
http://.../wiki/Ennio_Morricone

http://.../wiki/Ennio_Morricone
http://.../wiki/The_Good,_the_Bad,_
http://.../wiki/Yo-Yo_Ma
http://.../wiki/Metallica

Similarity (m,e):

- cos/Dice/KL (context(m), context(e))



Coherence (e,e‘):

- dist(types)
- overlap(links)
- overlap (keyphrases)

Mention-Entity Graph

weighted undirected graph with two types of nodes

Tan wrote the score for Zhang's Oscar-nominated epic film. He also worked with Ma who covered the Ecstasy.

TAN

Dun Tan

Oscar Pistorius

Academy Award

MDMA

Ecstasy (painting)

Ecstasy of Gold

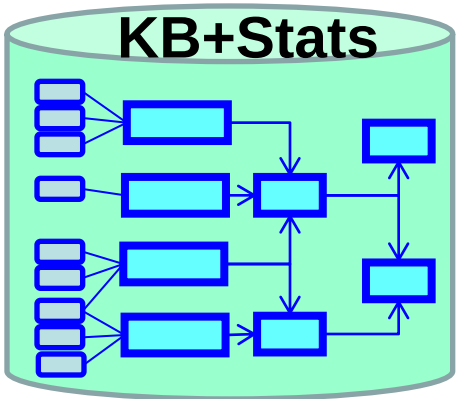
Crouching Tiger, Hidden Dragon
ceremony for Beijing olympics
Oscar for best original score
American cellist Yo-Yo Ma

Hollywood film industry
best foreign language film
best original score
Maestro Morricone life-time award

The Good, the Bad, and the Ugly
Man with No Name trilogy
soundtrack by Ennio Morricone
Yo-Yo Ma cover

Similarity (m,e):

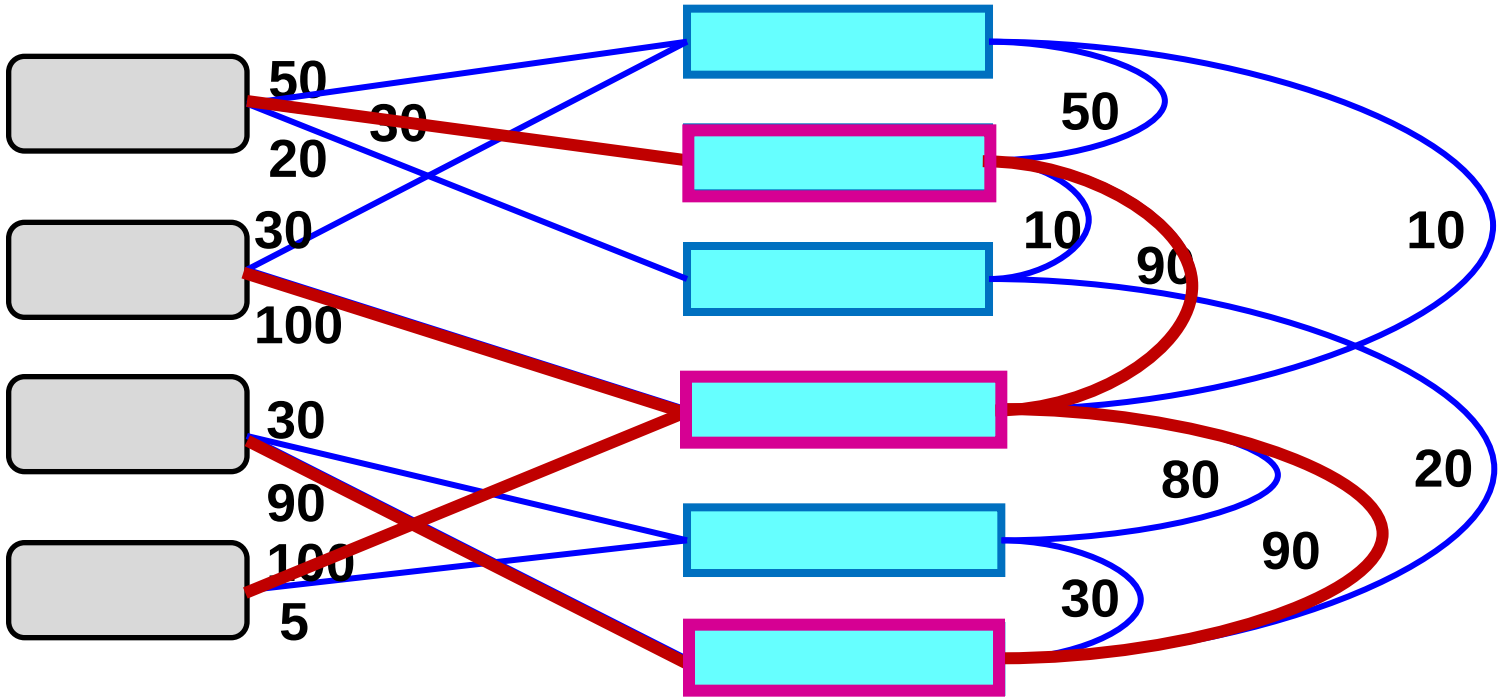
- cos/Dice/KL (context(m), context(e))



Coherence (e,e‘):

- dist(types)
- overlap(links)
- overlap (keyphrases)

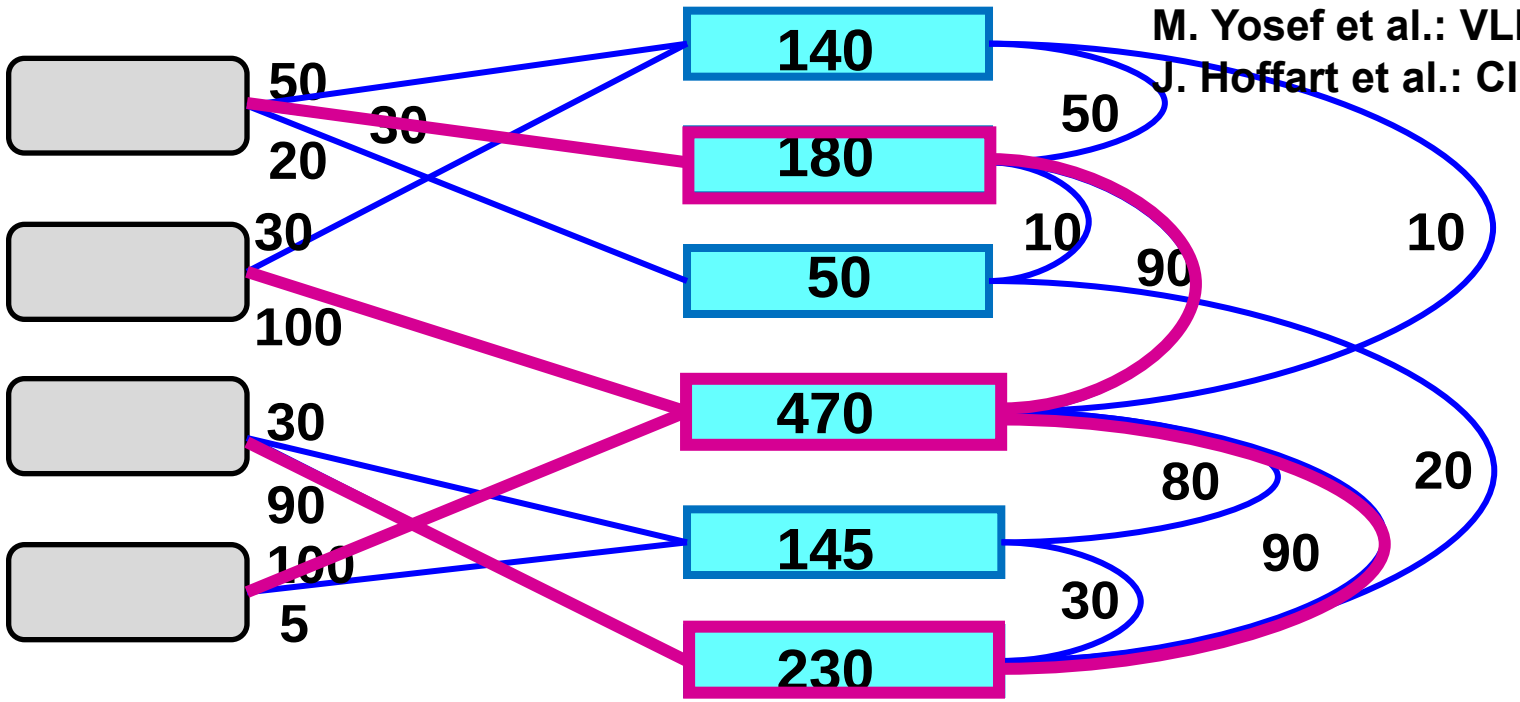
Joint Mapping



- Build **mention-entity graph** or **joint-inference factor graph** from knowledge and statistics in **YAGO** (or other KB)
- Compute **high-likelihood mapping** (ML or MAP) or **dense subgraph** such that:
  each m is **connected to exactly one** e (or **at most one** e)

Coherence Graph Algorithm

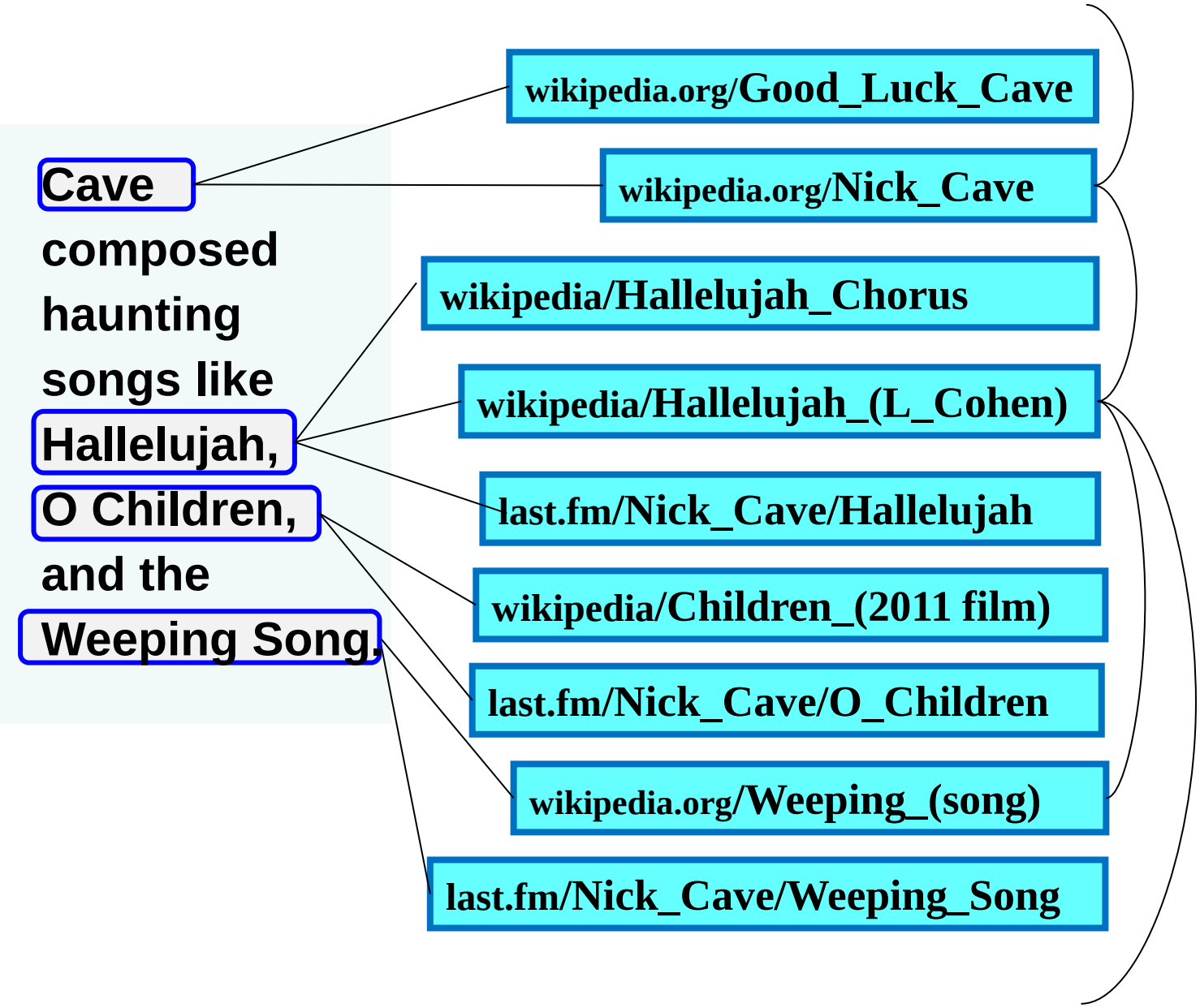
[J. Hoffart et al.: EMNLP'11
M. Yosef et al.: VLDB'11
J. Hoffart et al.: CIKM'12]



- Compute dense subgraph to
  maximize min weighted degree among entity nodes
  such that:
  each m is connected to exactly one e (or at most one e)
- Approx. algorithms (greedy, randomized, …), hash sketches, …
- 82% precision on CoNLL'03 benchmark
- Open-source software & online service AIDA

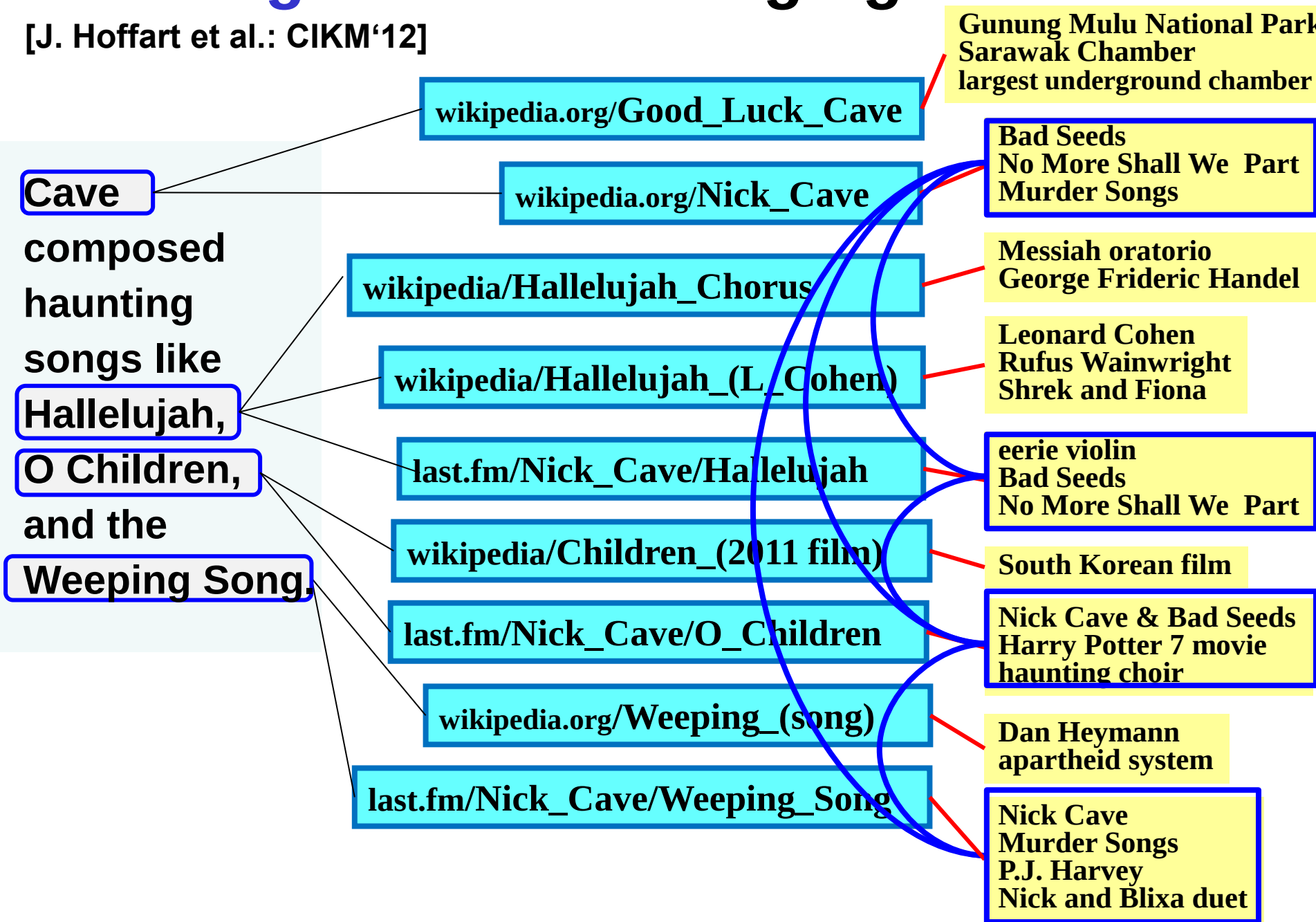
http://www.mpi-inf.mpg.de/yago-naga/aida/

Long-Tail and Emerging Entities

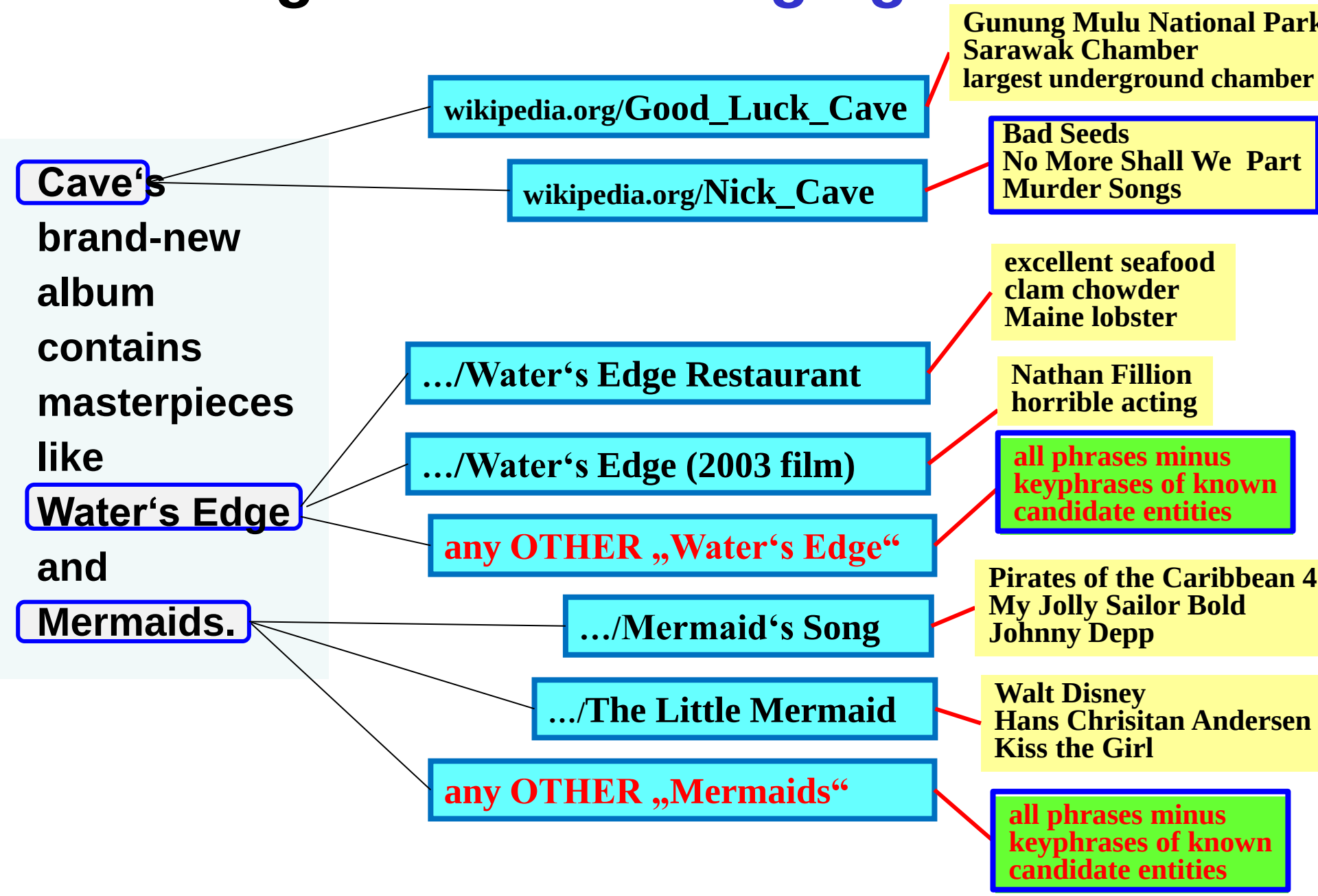


Long-Tail and Emerging Entities

[J. Hoffart et al.: CIKM'12]



Long-Tail and Emerging Entities



Open Challenges

High-throughput NERD (semantic indexing)
Low-latency NERD (speed-reading)

- **popular vs. long-tail entities, general vs.specific domain**

Deep-parsing vs. language-agnostic features

Dun Tan wrote the score for Zhang‘s Hero

谭盾曾为张艺谋的电影《英雄》创作配乐。

KB life-cycle for emerging entities

WSD for classes, relations, general concepts

- **for Web tables, questions, dialogs, summarization, …**

Outline

- ✓ Motivation
- ✓ Entity Name Disambiguation
- ★ Relational Paraphrases
- ★ Translating Questions into Queries
- ★ Wrap-Up

Diversity and Ambiguity of Relational Phrases

question: Who covered whom?

Amy Winehouse's concert included cover songs by the Shangri-Las

Amy's souly interpretation of Cupid, a classic piece of Sam Cooke

Nina Simone's singing of Don't Explain revived Holiday's old song

Cat Power's voice is sad in her version of Don't Explain

16 Horsepower played Sinnerman, a Nina Simone original

Cale performed Hallelujah written by L. Cohen

Cave sang Hallelujah, his own song unrelated to Cohen's

{cover songs, interpretation of, singing of, voice in, …} ⇔ SingerCoversSong

{classic piece of, 's old song, written by, composition of, …} ⇔ MusicianCreatesSong

Scalable Mining of SOL Patterns

[N. Nakashole et al.: EMNLP-CoNLL'12, VLDB'12]

***Syntactic-Lexical-Ontological (SOL)* patterns**

- **Syntactic-Lexical:** surface words, wildcards, POS tags
- **Ontological:** semantic classes as entity placeholders
  <singer>, <musician>, <song>, …
- **Type signature** of pattern: <singer> × <song>, <person> × <song>
- **Support set** of pattern: set of entity-pairs for placeholders
  → support and confidence of patterns

SOL pattern: <singer> 's ADJECTIVE voice * in <song>

Matching sentences:

Amy Winehouse's soul voice in her song 'Rehab'
Jim Morrison's haunting voice and charisma in 'The End'
Joan Baez's angel-like voice in 'Farewell Angelina'

Support set:

(Amy Winehouse, Rehab)
(Jim Morrison, The End)
(Joan Baez, Farewell Angelina)

Pattern Dictionary for Relations

[N. Nakashole et al.: EMNLP-CoNLL'12, VLDB'12]

WordNet-style dictionary/taxonomy for relational phrases based on SOL patterns (syntactic-lexical-ontological)

Relational phrases are typed

- *<person>* graduated from *<university>*
- *<singer>* covered *<song>*
- *<book>* covered *<event>*

Relational phrases can be synonymous

- "graduated from" ⇔ "obtained degree in * from"
- "and PRONOUN ADJECTIVE advisor" ⇔ "under the supervision of"

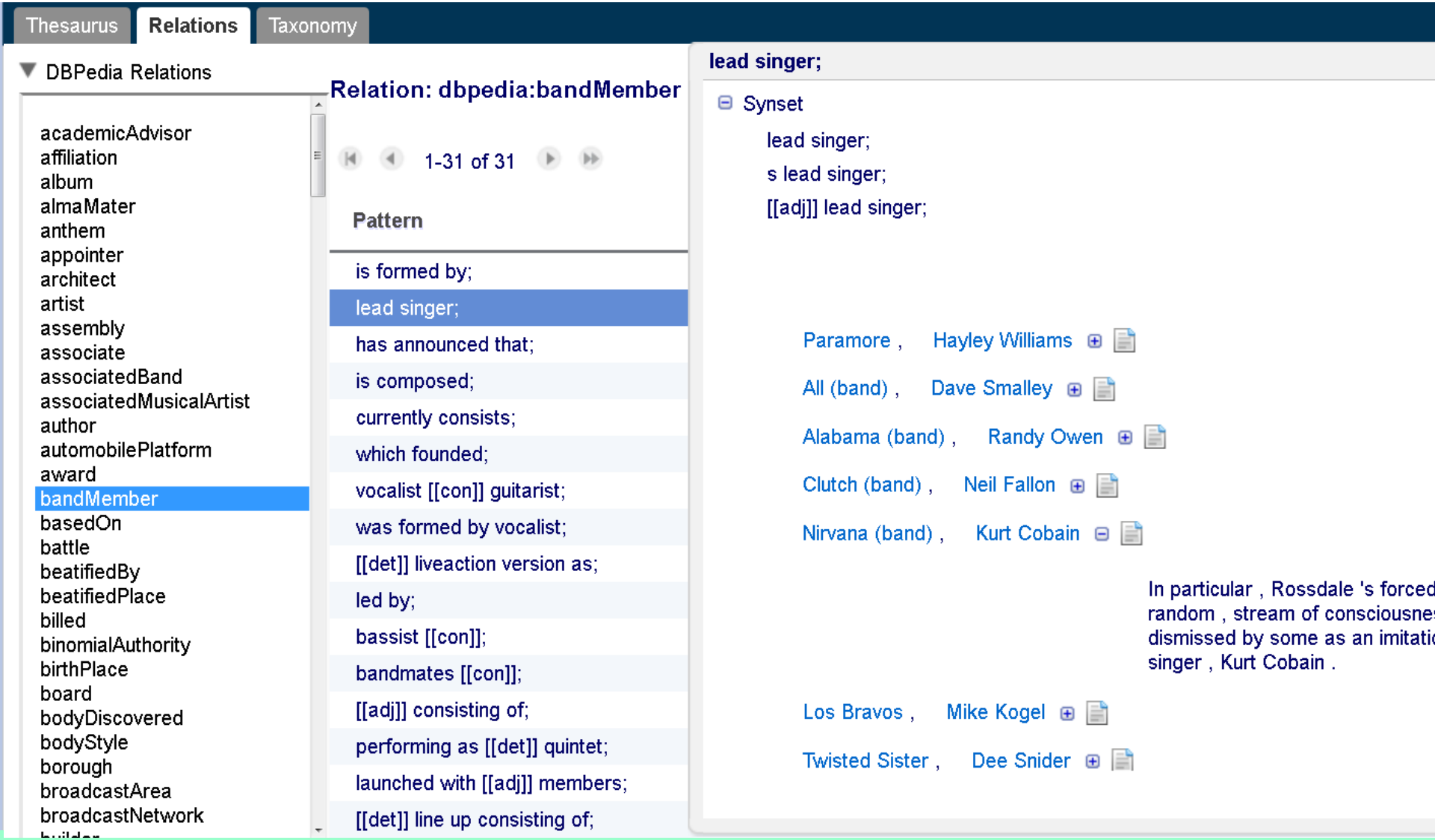
One relational phrase can subsume another

- "wife of" ⇒ " spouse of"

350 000 SOL patterns from Wikipedia, NYT archive, ClueWeb

http://www.mpi-inf.mpg.de/yago-naga/patty/

PATTY: Pattern Taxonomy for Relations

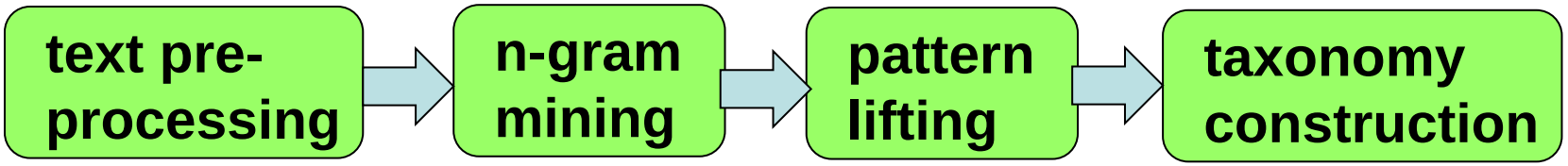


350 000 SOL patterns with 4 Mio. instances
accessible at: www.mpi-inf.mpg.de/yago-naga/patty

Big Data Algorithms at Work

Frequent sequence mining
with generalization hierarchy for tokens
Examples: famous → ADJECTIVE → *
her → PRONOUN → *
<singer> → <musician> → <artist> → <person>

Map-Reduce-parallelized on Hadoop:
- identify entity-phrase-entity occurrences in corpus
- compute frequent sequences
- repeat for generalizations



Semantic Typing of Emerging Entities

[N. Nakashole et al.: ACL 2013, T. Lin et al.: EMNLP 2012]

Problem: what to do with newly emerging entities

Idea: infer their semantic types using PATTY patterns

Sandy *threatens to hit* New York
Nive Nielsen *and her band performing* Good for You
Nive Nielsen*'s warm voice in* Good for You

Given triples (x, p, y) with new x
and all type triples (t1, p, t2) for known entities:

- score (x,t) ~ $\Sigma_{p:(x,p,y)}$ P [t | p,y]
- corr(t_1,t_2) ~ Pearson coefficient $\in$ [-1,+1]

For each new e and all candidate types t_i:

max $\alpha \Sigma_i$ score(e,t_i) X_i + $\beta \Sigma_{ij}$ corr(t_i,t_j) Y_{ij}

s.t. X_i, $Y_{ij} \in \{0,1\}$ and $Y_{ij} \leq X_i$ and $Y_{ij} \leq X_j$ and $X_i + X_j - 1 \leq Y_{ij}$

Semantic Typing of Emerging Entities

[N. Nakashole et al.: ACL 2013]

Entity	Inferred Type	Source Sentence (s)
Lochte	medalist	**Lochte** won America's lone gold on the first day of swimming competition.
Malick	director	Turn the clock back 15 months, and Brad Pitt, Sean Penn and Jessica Chastain all graced the red carpet in Cannes for **Malick**'s 2011 movie , " The Tree of Life'".
Bonamassa	musician	**Bonamassa** recorded Driving Towards the Daylight in Las Vegas with a mix of veteran studio musicians including drummer Anton Fig from the Late Show with David Letterman band and Nashville bass ace Michael Rhodes. At the age of 12, **Bonamassa** opened for B.B. King in Rochester , N.Y. "It was a thrill", he said and in 2009 he fulfilled a dream by performing at the Royal Albert Hall in London, where Eric Clapton made a guest appearance.
Analog Man	album	Analog Man is Joe Walsh's first solo album in 20 years.
Rep. Debbie Wasserman Schultz	person	Thomas Roberts speaks with **Rep. Debbie Wasserman Schultz**, chair of the Democratic National Committee, about a new Quinnipiac Poll that shows ...
LightSquared	organization	**LightSquared** paid Boeing some $1 billion for two satellites with the largest antenna receivers ever put into space, one of which was launched and is circling the Earth now.
Melinda Liu	journalist	"My fervent hope is that it would be possible for me and my family to leave for the U.S. on Hillary Clinton's plane," Chen said in a telephone interview with journalist **Melinda Liu** of the Daily Beast.
U.S. Border Patrol Agent Brian Terry	military officer	The inspector general determined that ATF agents and federal prosecutors had enough evidence to arrest and charge Jaime Avila, a Phoenix gun smuggler, months before **Border Patrol Agent Brian Terry** was killed near Tucson in December 2010.
RealtyTrac	publication	Earlier this month, **RealtyTrac** reported that for the first time since it began compiling foreclosure statistics in 2005, Illinois had the highest foreclosure rate among all the states in August.

Open Challenges

Higher coverage of relations
Cleaner subsumption hierarchy

Harness human computing (games?)
- **HIGGINS (Kondreddi et al.: ICDE'14)**

Combine Patty with grammar-based & learning methods
- **OLLIE (Mausam et al.: EMNLP-CoNLL'12)**
- **ClausIE (del Corro et al.: WWW'13)**

Beyond binary relations
- **verbs with two objects**
  Goldie Hawn presented the Oscar to Tan Dun for his music in …
- **spatial & temporal & other modifiers**
  Bayern Munich won this year's Champion's league in Wembley
- **provenance of statements**
  the city's spokesperson said that Munich's mayor is a fan of the Bayern team

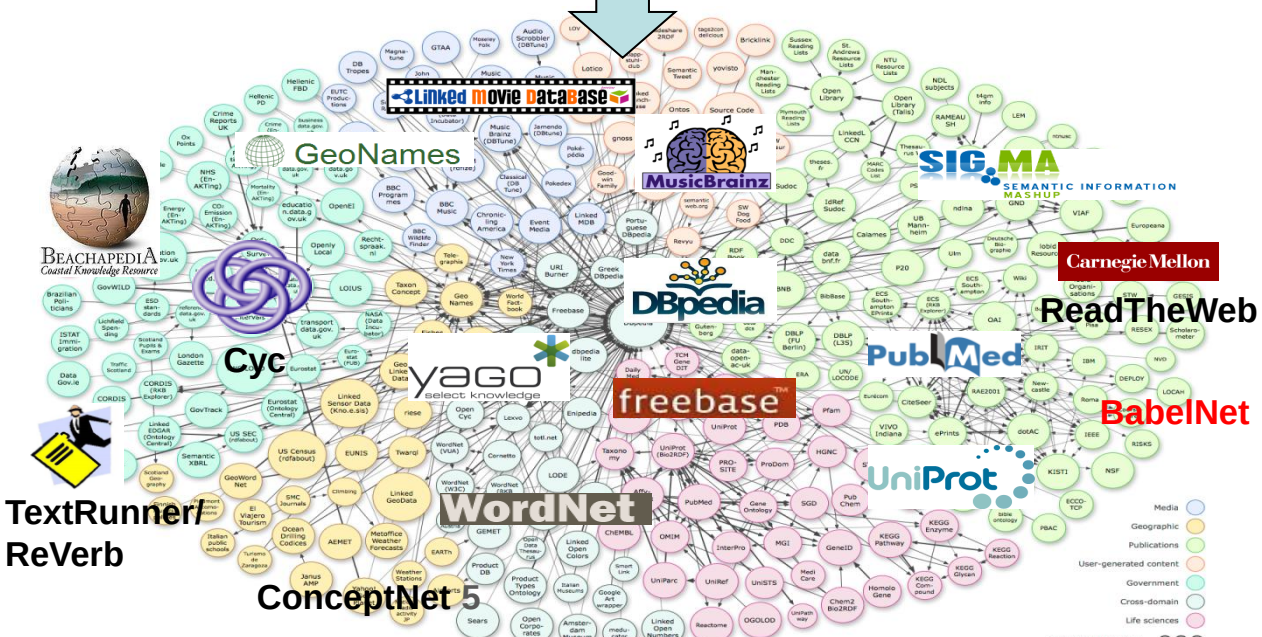
Outline

- ✓ **Motivation**
- ✓ **Entity Name Disambiguation**
- ✓ **Relational Paraphrases**
- ★ **Translating Questions into Queries**
- ★ **Wrap-Up**

Language Understanding: Questions into Queries

question: Who composed scores for westerns and is from Rome?

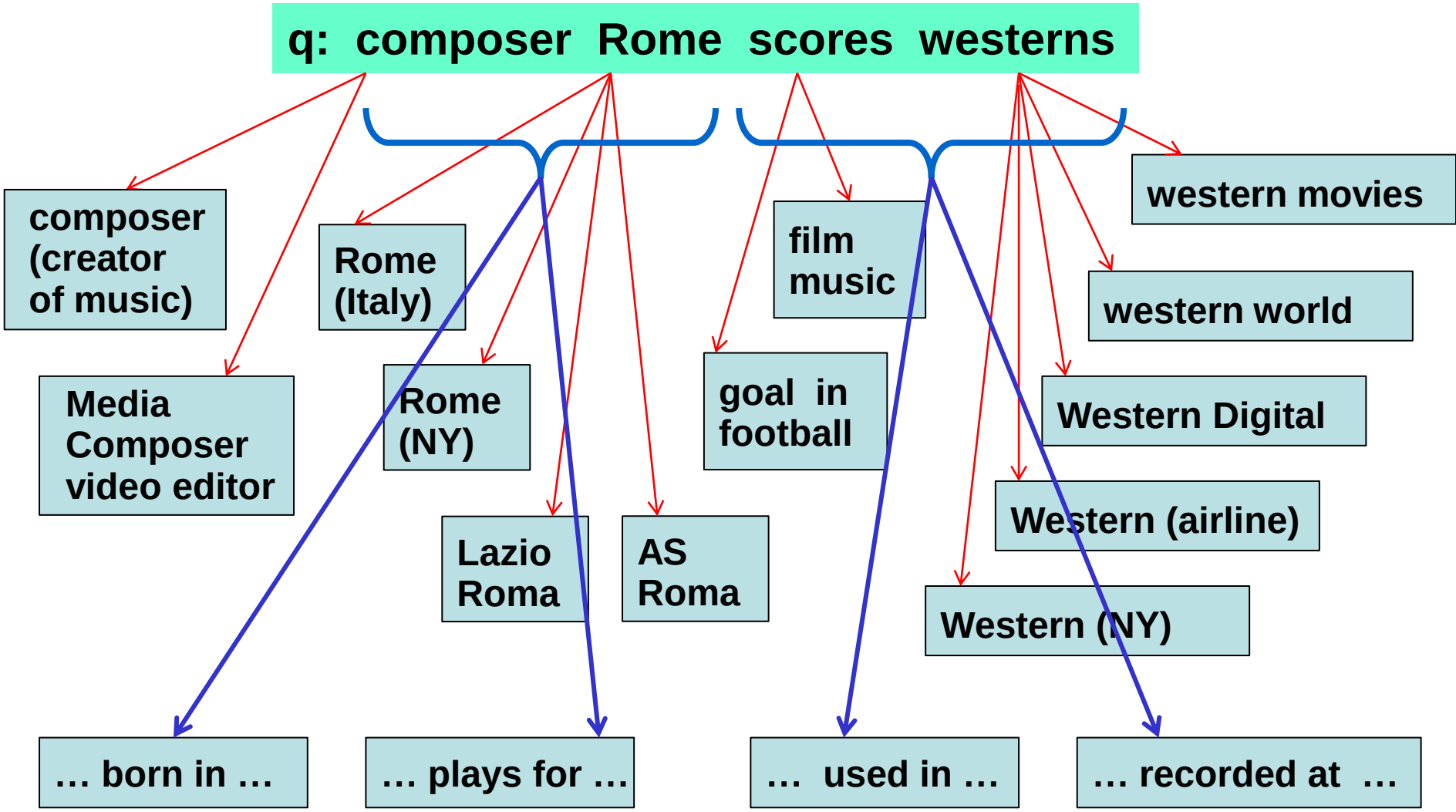
```
Select ?x Where {   ?x  created  ?s .
                    ?s  contributesTo  ?m .
                    ?m  type  westernMovie .
                    ?x  bornIn  Rome . }
```



Semantic Keyword Search

[Ilyas et al. Sigmod'10]

question: Who composed scores for westerns and is from Rome?

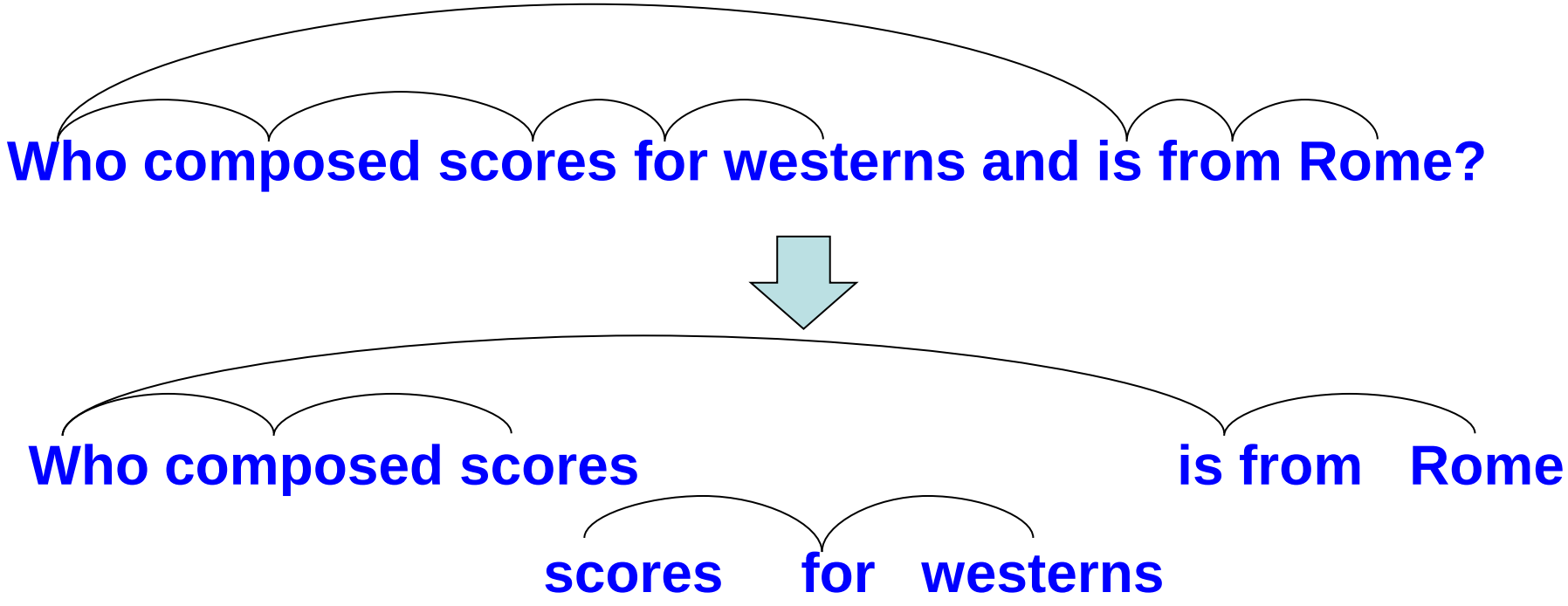


From Questions to Queries

translate question into Sparql query:

- **dependency parsing to decompose question**
- **mapping of question units onto entities, classes, relations**

question:

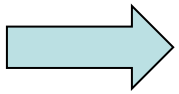


From Triploids to SPO Triple Patterns

[M. Yahya et al.: EMNLP'12, CIKM'13]

Who composed scores for westerns and is from Rome?

Who composed scores

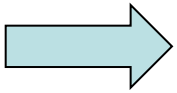


?x created ?s

?x type composer

?s type music

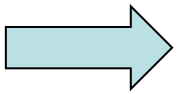
scores for westerns



?s contributesTo ?y

?y type westernMovie

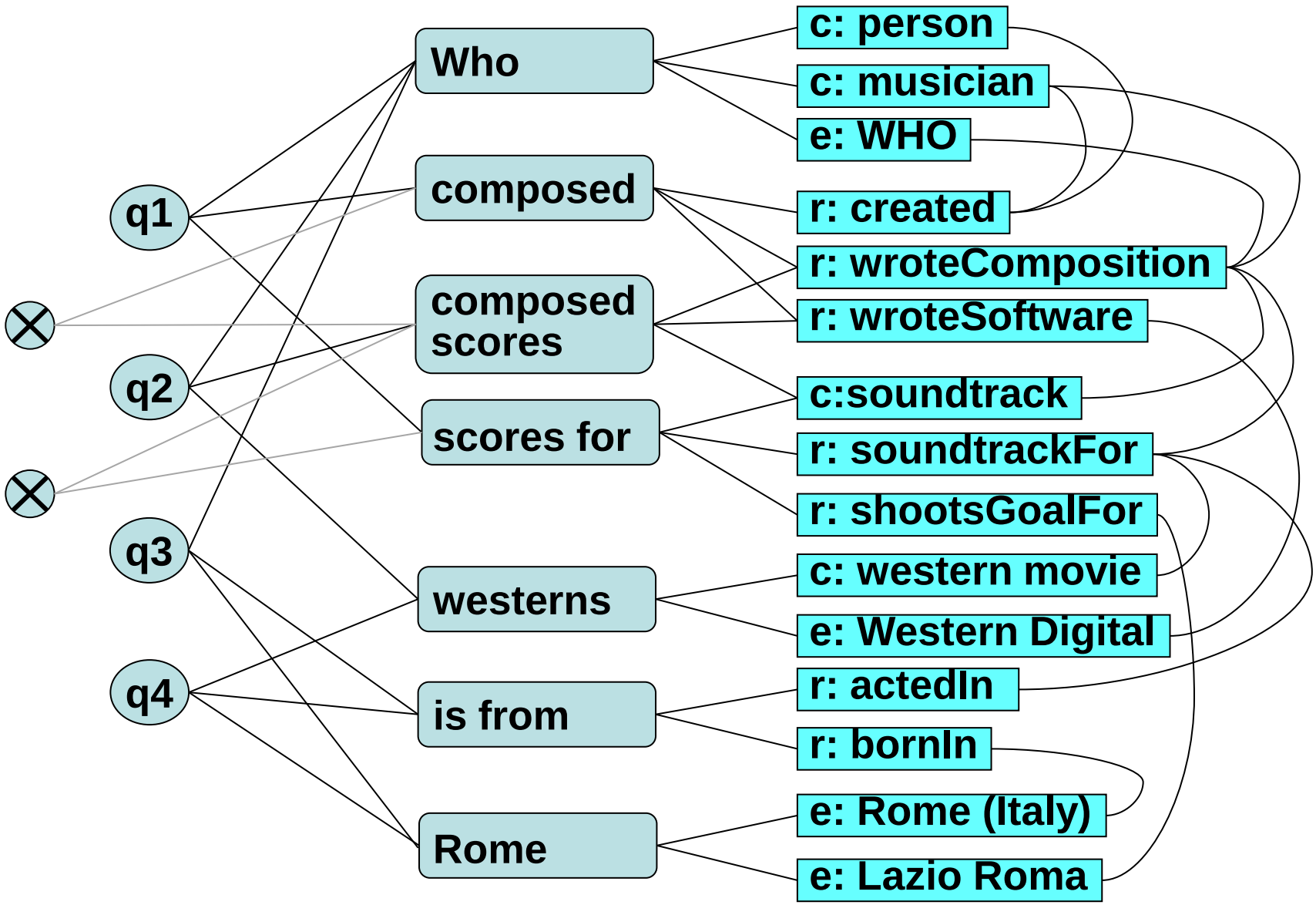
Who is from Rome



?x bornIn Rome

Disambiguation Mapping for Triploids

Who composed scores for westerns and is from Rome?

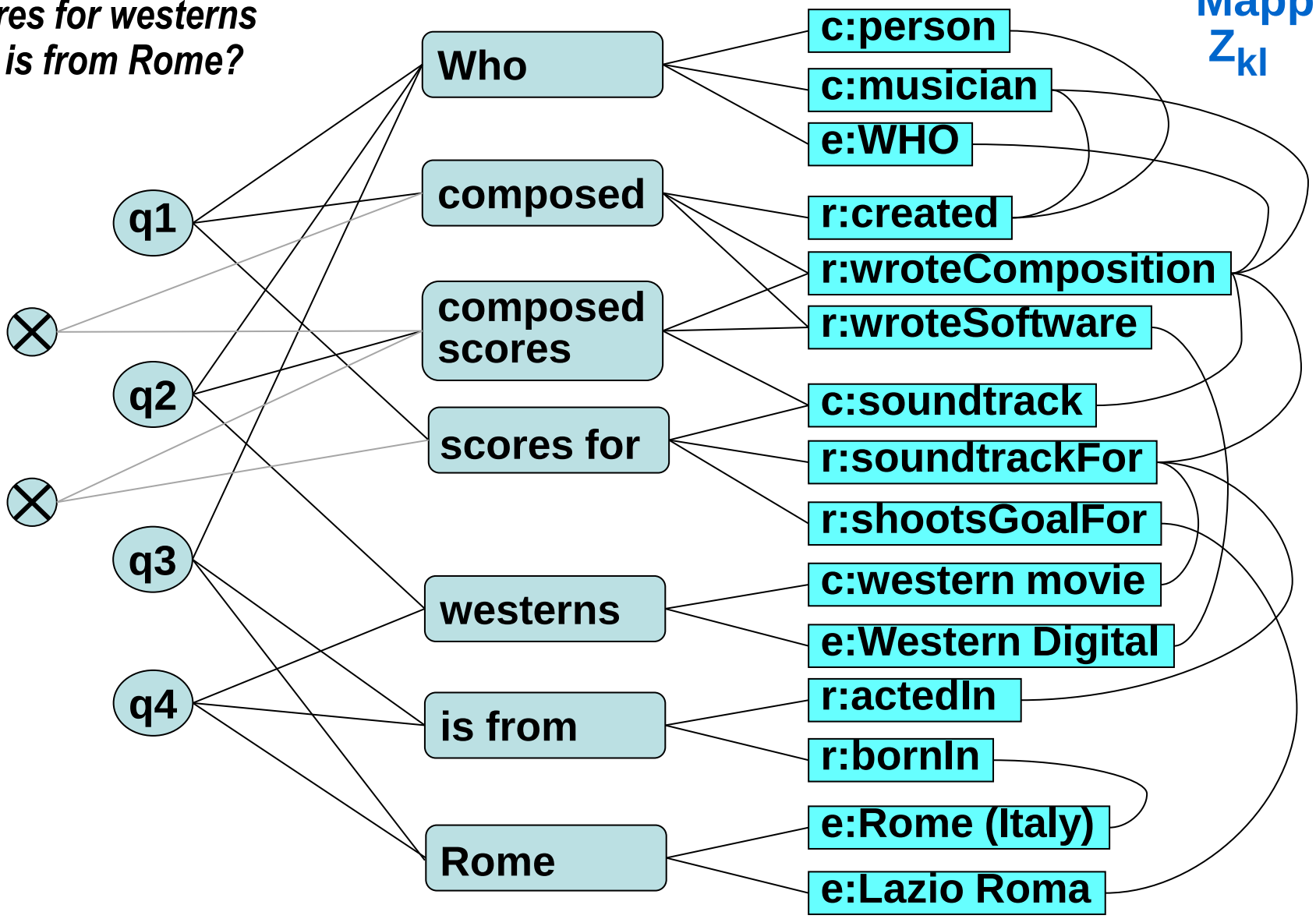


Combinatorial Optimization by ILP (with type constraints etc.)

Disambiguation Mapping

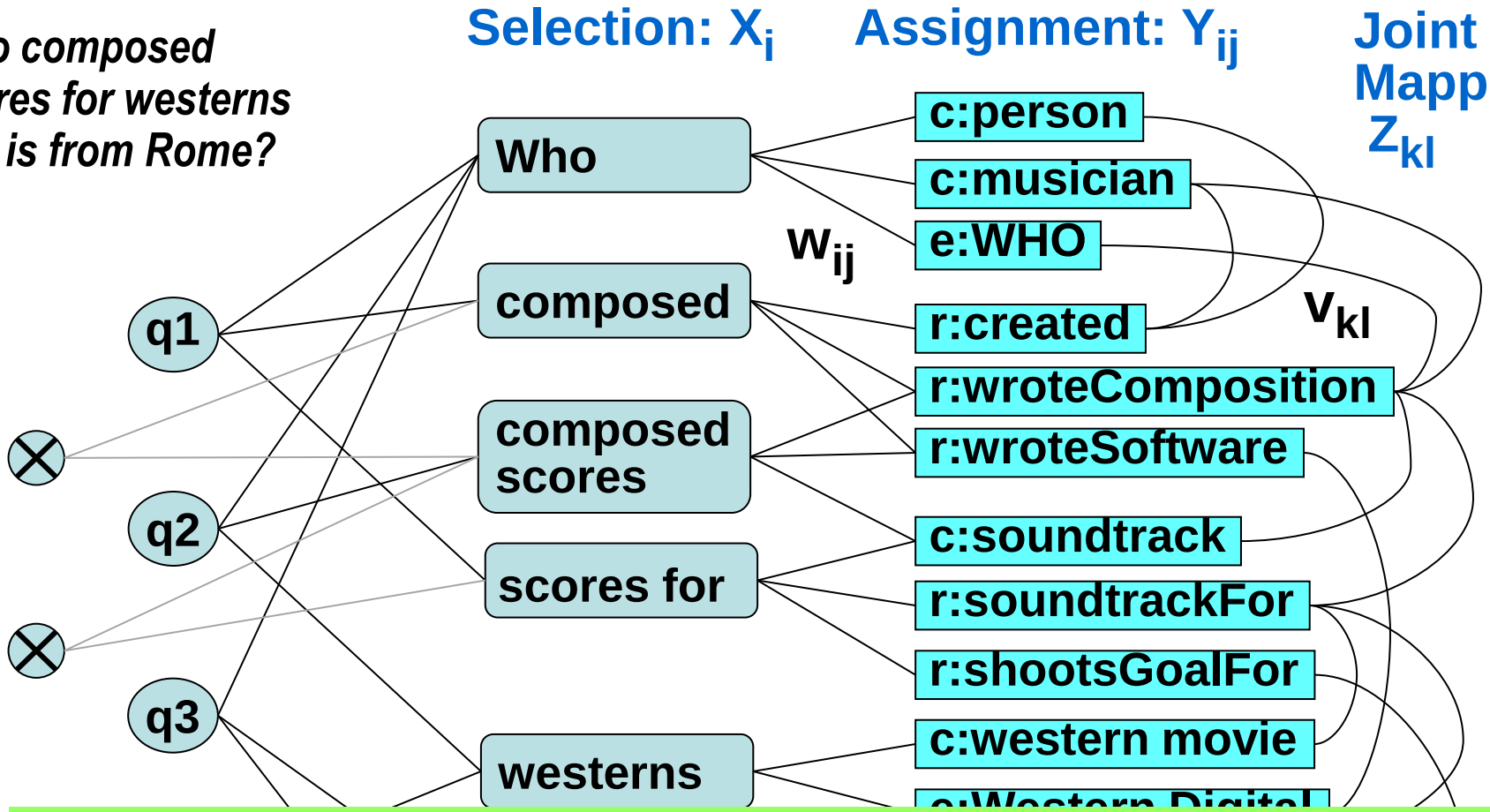
[M.Yahya et al.: EMNLP'12, CIKM'13]

Who composed scores for westerns and is from Rome?



Disambig. Mapping: Objective Function

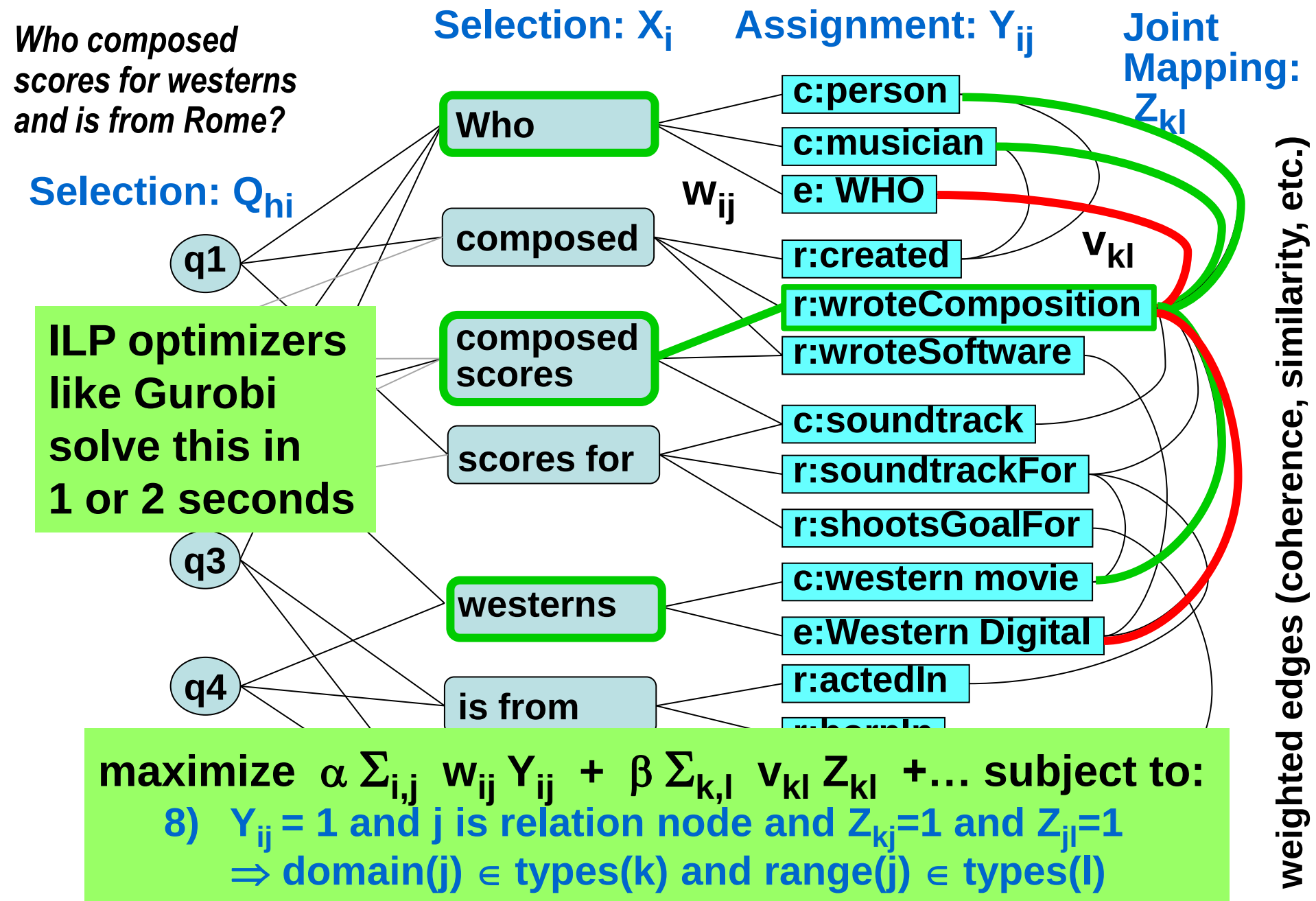
Who composed scores for westerns and is from Rome?



maximize $\alpha \Sigma_{i,j}\ w_{ij}\ Y_{ij} + \beta \Sigma_{k,l}\ v_{kl}\ Z_{kl} + \ldots$ **subject to:**

1) $Y_{ij} \leq X_i$ for all i,j
2) $\Sigma_j\ Y_{ij} \leq 1$ for all i
3) $Z_{kl} \leq \Sigma_{i,j}\ Y_{ik}$ and $Z_{kl} \leq \Sigma_j\ Y_{il}$ for all k,l
4) $X_i, Y_{ij}, Z_{kl} \in \{0,1\}$

Disambig. Mapping: Type Constraints



Open Challenges

Robustness

- training collections (QALD & others) ?

Questions in dialog

Someone covered Morricone‘s Ecstasy in a heavy metal style.
Who did this?

Multimodal dialog, with gestures …

Temporal, spatial & other modifiers

Which films did Roland Emmerich direct after 2012?
Which films did Yimou Zhang direct after 2012?
Which songs did Leonard Cohen write after Hallelujah?

Beyond English? Language-agnostic?

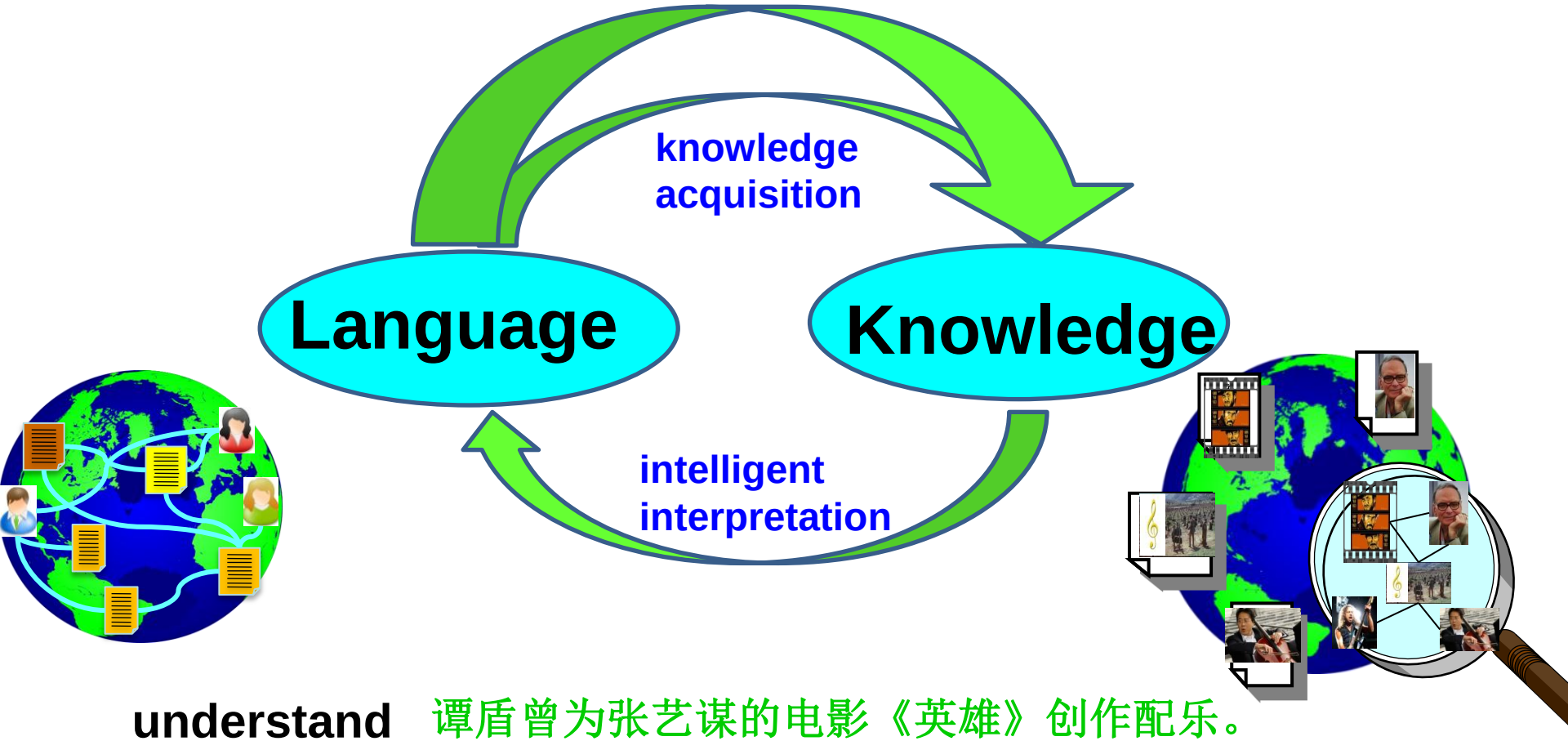
Outline

- ✓ Motivation
- ✓ Entity Name Disambiguation
- ✓ Relational Paraphrases
- ✓ Translating Questions into Queries
- ★ Wrap-Up

Summary

- **Web of Data & Knowledge & Text :**
  Entities, Classes & Relations
- **Diversity & Ambiguity of Names and Phrases**
  Calls for Disambiguation Mapping
- **Strong Story for Entity Name Disambiguation**
- **Ongoing Work on Relation Phrase Disambiguation**
- **Cornerstone of Question Answering and**
  Natural Language Dialog

Take-Home Message



understand 谭盾曾为张艺谋的电影《英雄》创作配乐。
& answer "Who composed the Ecstasy
and other pieces for westerns?"

⇒ can map ambiguous names and phrases
into entities and relations
for QA, search, machine reading, big text analytics, …

Acknowledgements

LAMA

SEVENTH FRAMEWORK PROGRAMME

M2CI CLUSTER OF EXCELLENCE

DFG Deutsche Forschungsgemeinschaft

Google

mpii max planck institut informatik