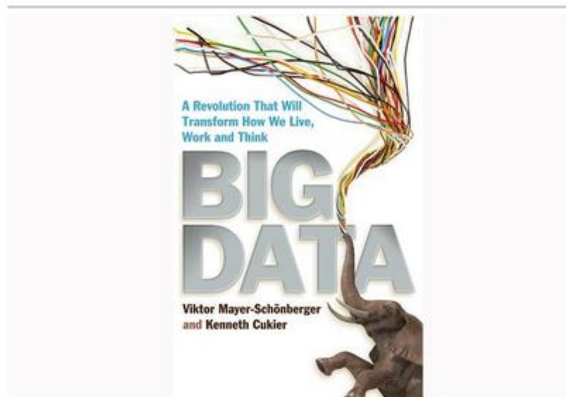
From Text to Entities and from Entities to Insight:
a Perspective on Unstructured Big Data

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

Why Do We Work on Big Data?

“Why do you want to climb it?“

“Because it‘s there!“
(George Mallory
1886-1924)



THE TIMES
Non-fiction
News | Opinion | Business | Money | Sport | Life | Arts | Puzzles | Papers

Welcome to your preview of The Times

Big Data: A revolution that will transform the way we live, work and think

The Promise of Big Data

Energy
Traffic
Health
Business
…..

Big Data Analytics

Platforms & Algorithms
Scale, Scale, Scale …

make the world a better place !

???

News
Books
Social Media
Scholarly Publ.
…..

Outline

- Interesting Data
- From Names to Entities
- From Phrases to Relations
- From Text Analytics to Insight
- Wrap-Up

Structured vs. Unstructured Data

precise & insightful

?

noisy & hopeless

Location	Month	Temp
Northern Territory	April	31.4159°C
12°27'S 130°50'E	April	34.5°C

In Darwin, the capital of NT, the average temperature in April was consistently above 30 degrees and reached a peak of 34.5 degrees on April 23, 2013.

Levothyroxine side effects:

- weight loss
- tremor
- headache
- nausea
- vomiting
- diarrhea
- stomach cramps
- nervousness
- irritability
- insomnia
- excessive sweating

....

Nervous system side effects of levothyroxine have rarely included seizures during initiation of therapy.
Dermatologic side effects including hair loss have been reported during the initial months of therapy.

I took levothyroxine for the past four days.
I got a spell that lasts for a couple of hours.
This spell consists of tremors (mostly of the hands).

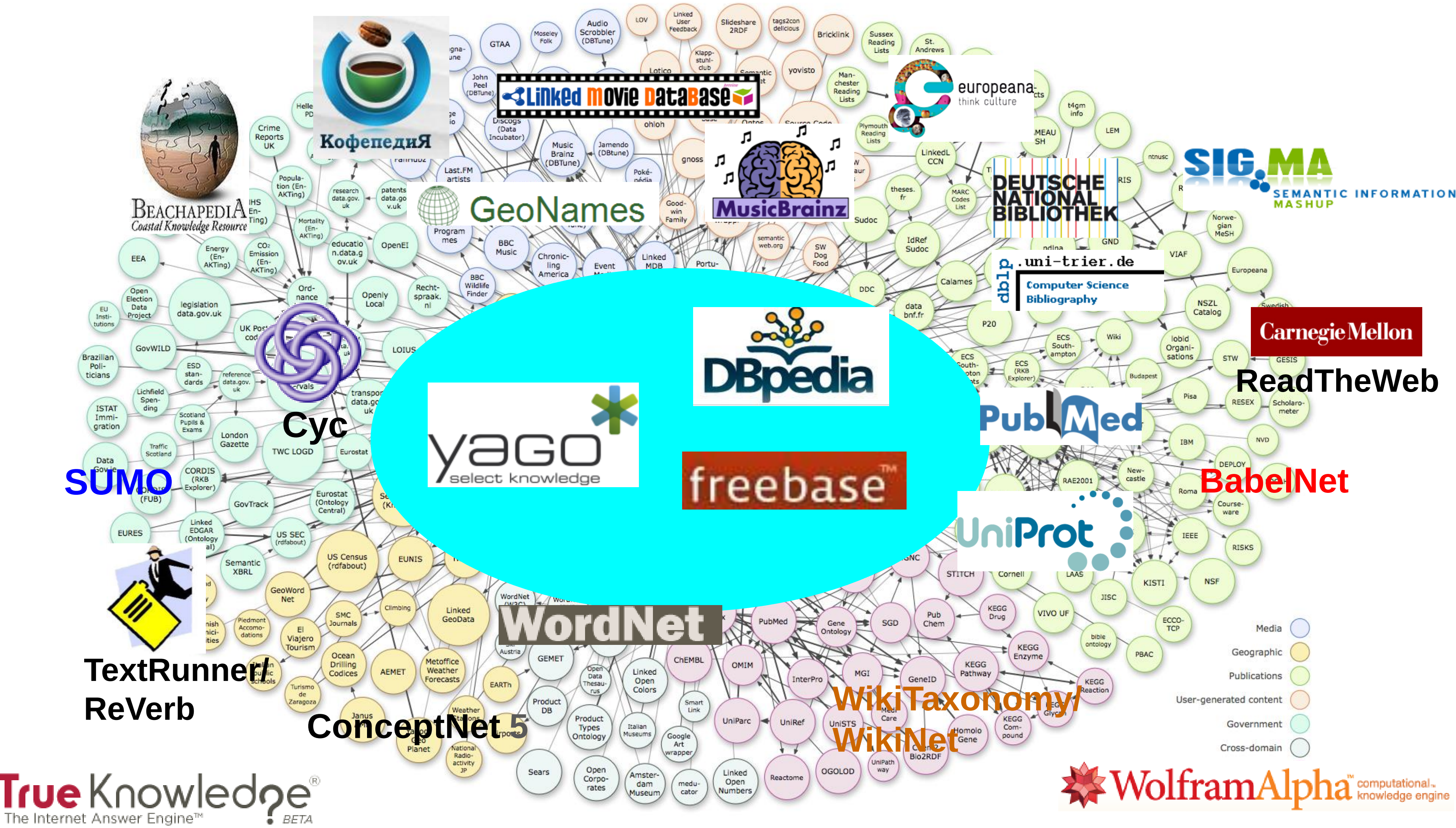
unclear to interpret

!

insightful to human

Interesting Data at Scale: LinkedOpenData

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



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

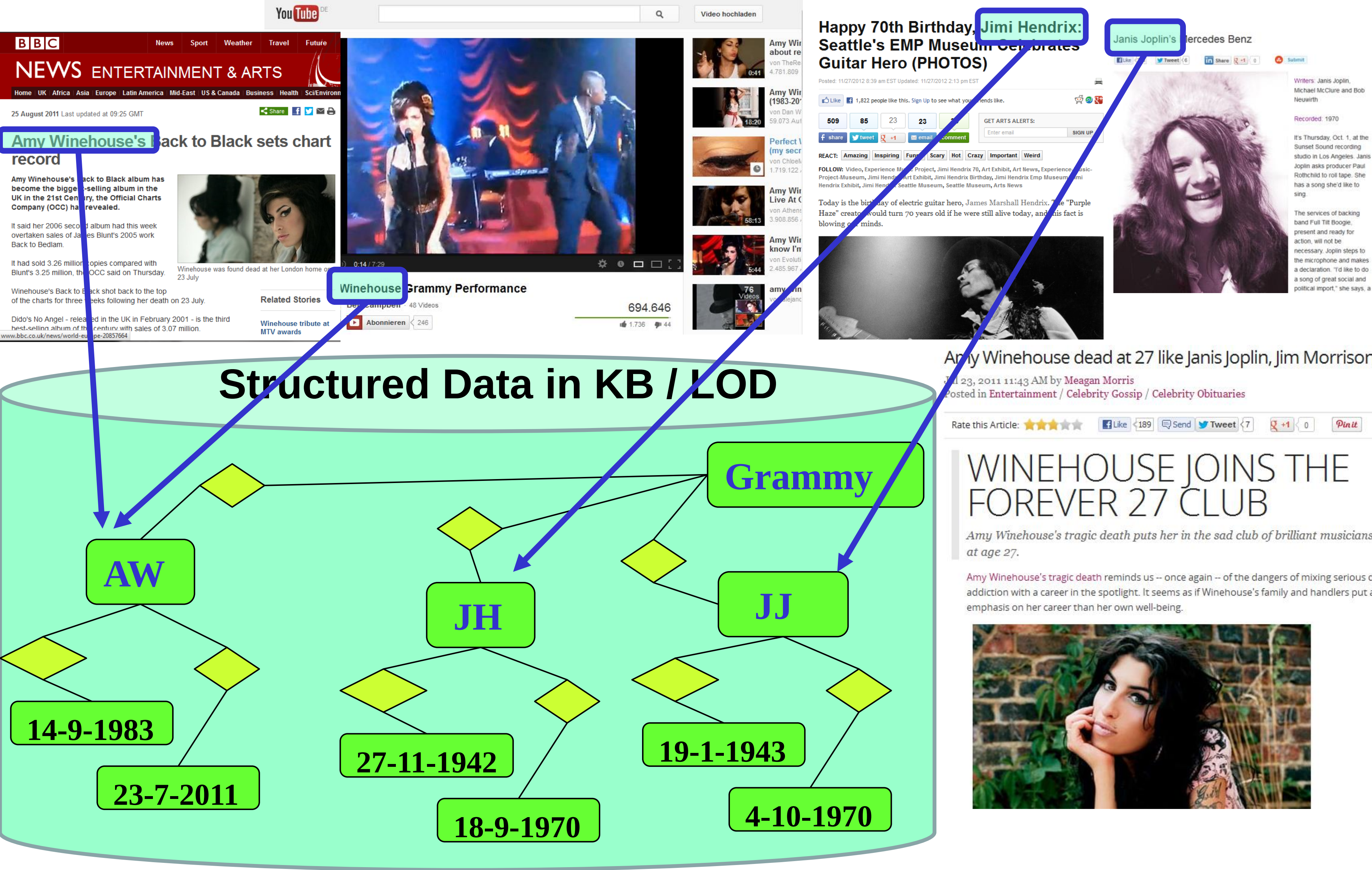
Linked Open Data

62 Bio. SPO triples (RDF) **+ linked text sources**

Amy_Winehouse **type** singer
Amy_Winehouse **type** GrammyAwardWinner
singer **subclassOf** musician
Amy_Winehouse **bornIn** Southgate
Southgate **locatedIn** London
Amy_Winehouse **diedOn** 23-July-2011
Amy_Winehouse **created** Back_to_Black
Amy_Winehouse **performed** Cupid

Web Entities

50 Bio. Web pages and Social-Media postings



Microdata & Tables in HTML

SECONDHANDSONGS

Search all | Find cover songs, artists and more

Explore

100 Mio's of structured tables

Second Hand Songs ▶ Database ▶ Artist ▶ Amy Winehouse

Artist: Amy Winehouse

Real name	Amy Jade Winehouse
Born	September 14, 1983
Died	July 23, 2011
Country	United Kingdom
Family	Mitch Winehouse *Father*

Covers

		Title	Performer	Release date	Originally by
1	YouTube	(There Is) No Greater Love	Amy Winehouse	October 2003	Isham Jones and His Orchestra
2	YouTube	A Song for You	Amy Winehouse	December 2, 2011	Leon Russell
3	YouTube	Body and Soul	Tony Bennett with Amy Winehouse	September 20, 2011	Ambrose and His Orchestra
4	YouTube	Cupid	Rhythms del Mundo featuring Amy Whinehouse	2009	Sam Cooke
5	YouTube	Cupid	Amy Winehouse	October 1, 2007	Sam Cooke
6	YouTube	Don't Go to Strangers			

http://www.secondhandsongs.com/artist/22392/

Microdata & Tables in HTML

SHS SECONDHANDSONGS

Artists | Ennio Morricone

Explore

100 Mio's of structured tables

Second Hand Songs ▶ Database ▶ Artist ▶ Ennio Morricone

Artist: Ennio Morricone

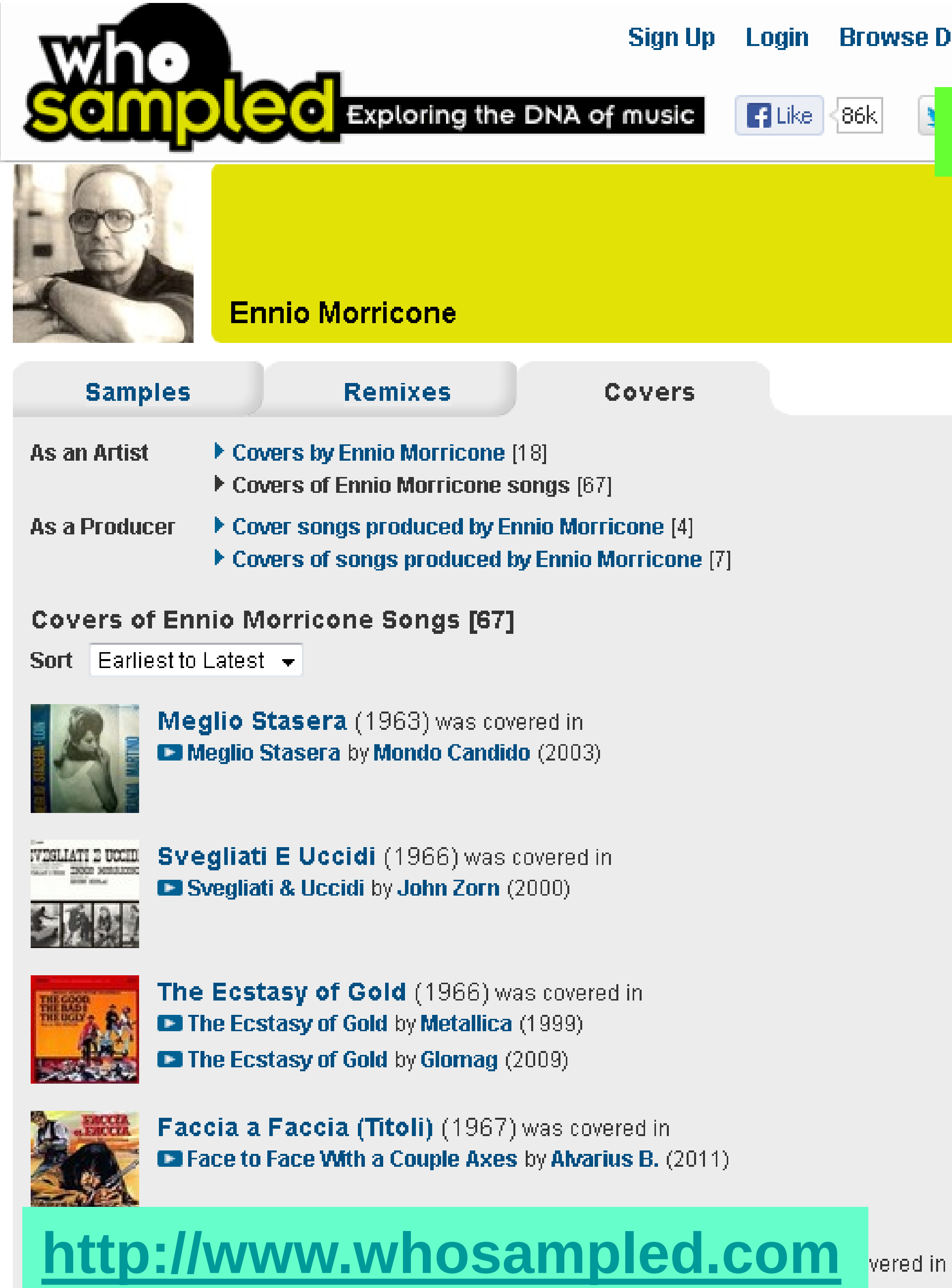
Aliases	Ennio Morricone e la sua orchestra
Born	November 10, 1928
Country	Italy
Comments	Composer.
Family	Andrea Morricone *son*

Covers

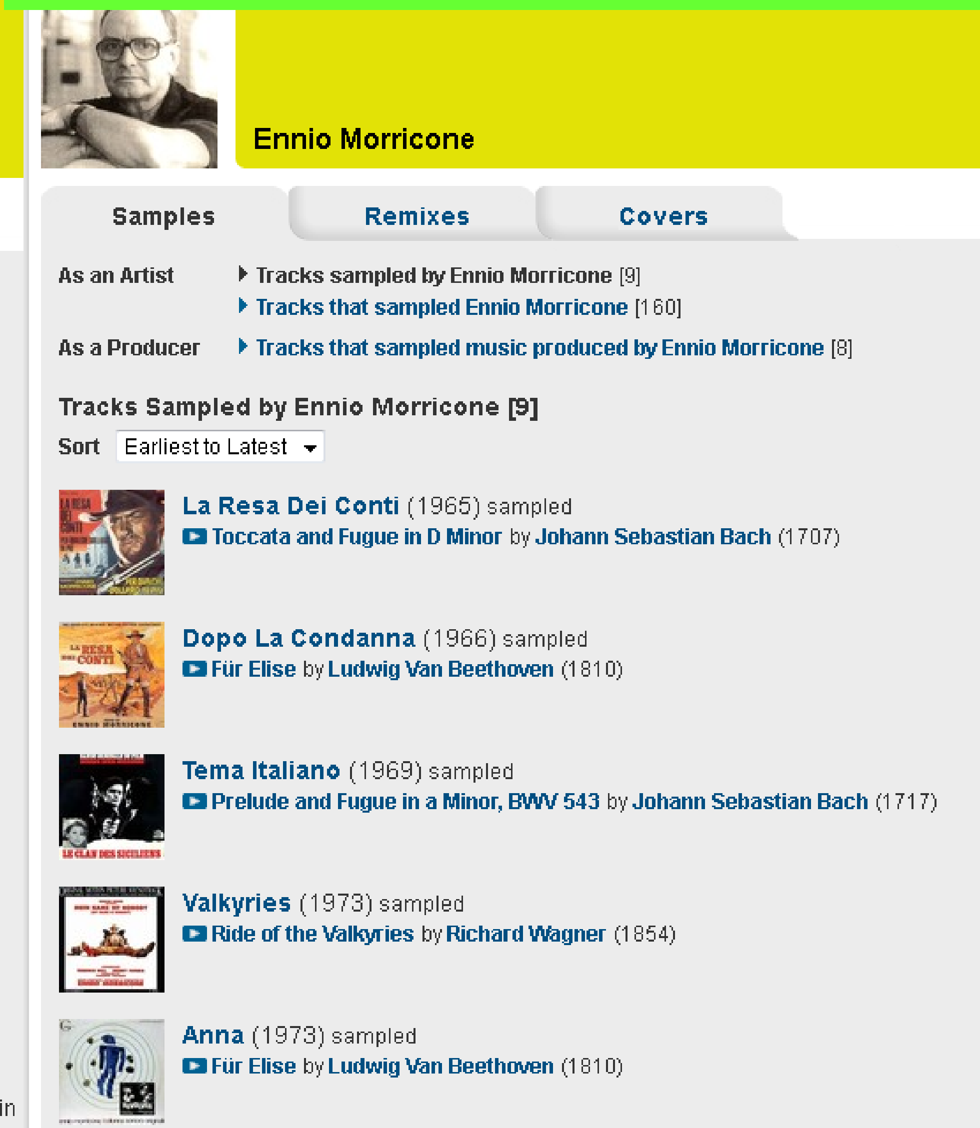
		Title	Performer	Release date	Originally by
1	You Tube	A Rose Among Thorns	Dulce Pontes & Ennio Morricone	2003	Ennio Morricone
2		Amapola	Ennio Morricone	1984	Miguel Fleta
3		Che cosa c'è	Ennio Morricone e la sua orchestra	1964	Ornella Vanoni
4		Chi mai	Milva & Ennio Morricone	1972	
5		Ciao ciao bambina (Piove)	Ennio Morricone e la sua orchestra	1964	Domenico Modugno
6		Deborah's Theme: I Knew I Loved You	Hayley Westenra & Ennio Morricone	April 18, 2011	Celine Dion - Ennio Morricone with Edda Dell'Orso
7		Here's to You	Hayley Westenra & Ennio Morricone	April 18, 2011	Ennio Morricone & Joan Baez
8		House of No Regrets	Dulce Pontes & Ennio Morricone	2003	Ennio Morricone
9		Io che amo solo te	Ennio Morricone e la sua orchestra	1964	Sergio Endrigo
10		La califfa			

http://www.secondhandsongs.com/artist/2257

Microdata & Tables in HTML

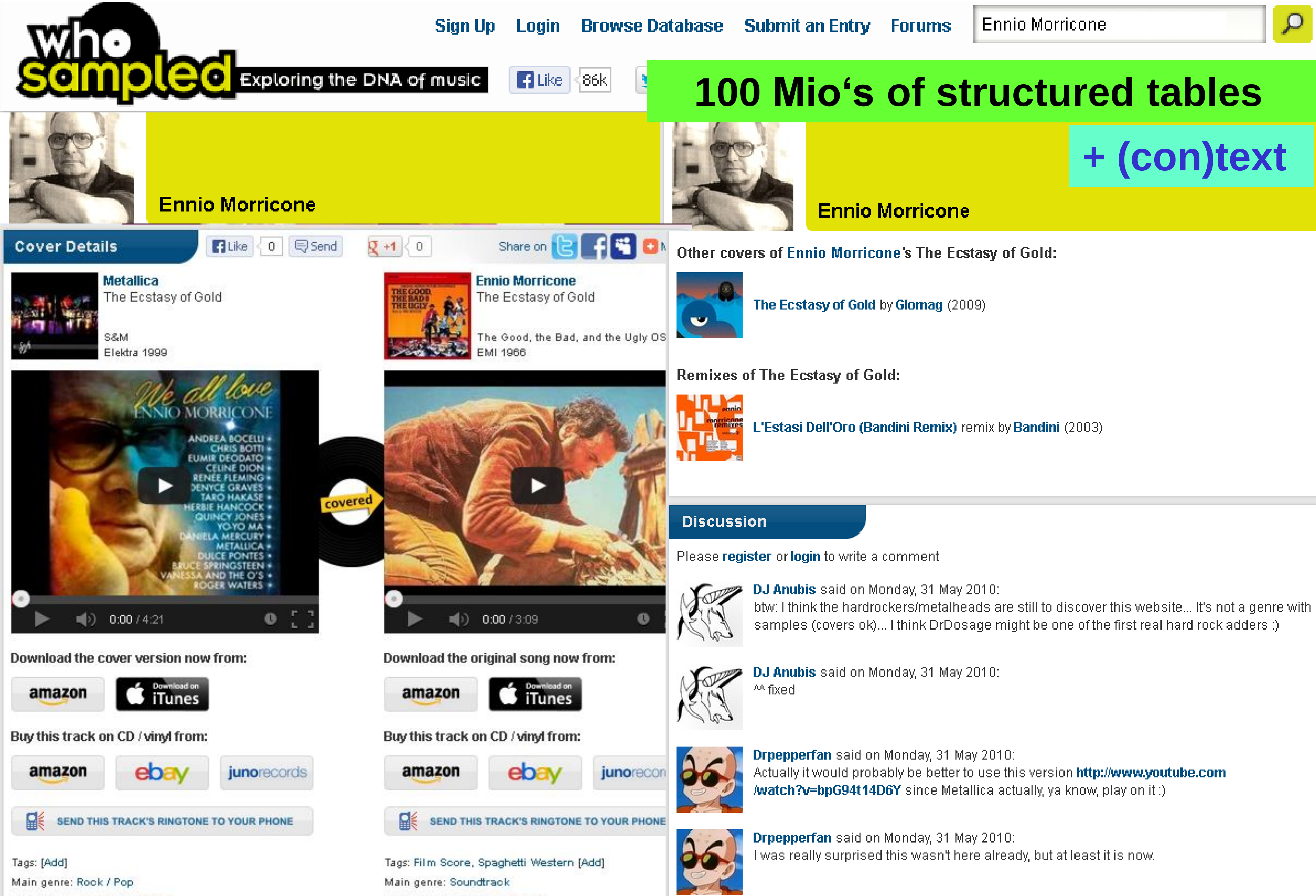


100 Mio's of structured tables



http://www.whosampled.com

Microdata & Tables in HTML



Big Data+Text Challenge

Entertainment Analytics – using only public data+text:

Who covered which other singer?
Which versions were most successful?
Who influenced which other musicians?

Health: **Which drug (combination)**
has which side effects under which conditions,
and how frequent are they observed?

Politics, Business, Energy, Traffic, Biodiversity, …

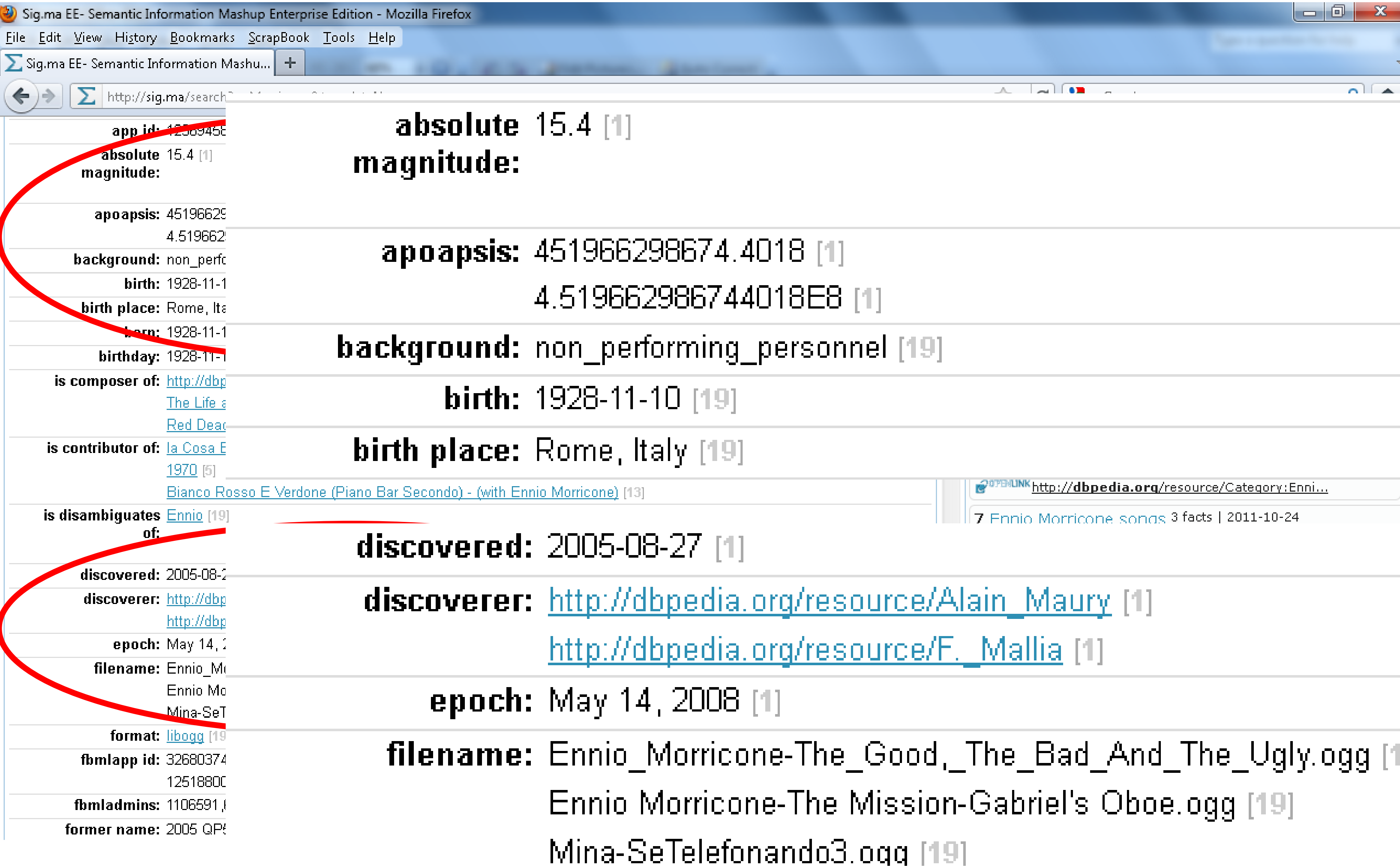
General Design Pattern:

- **Identify entities of interest & their relationships**
- **Position in time & space**
- **Group and aggregate**
- **Find insightful patterns & predict trends**

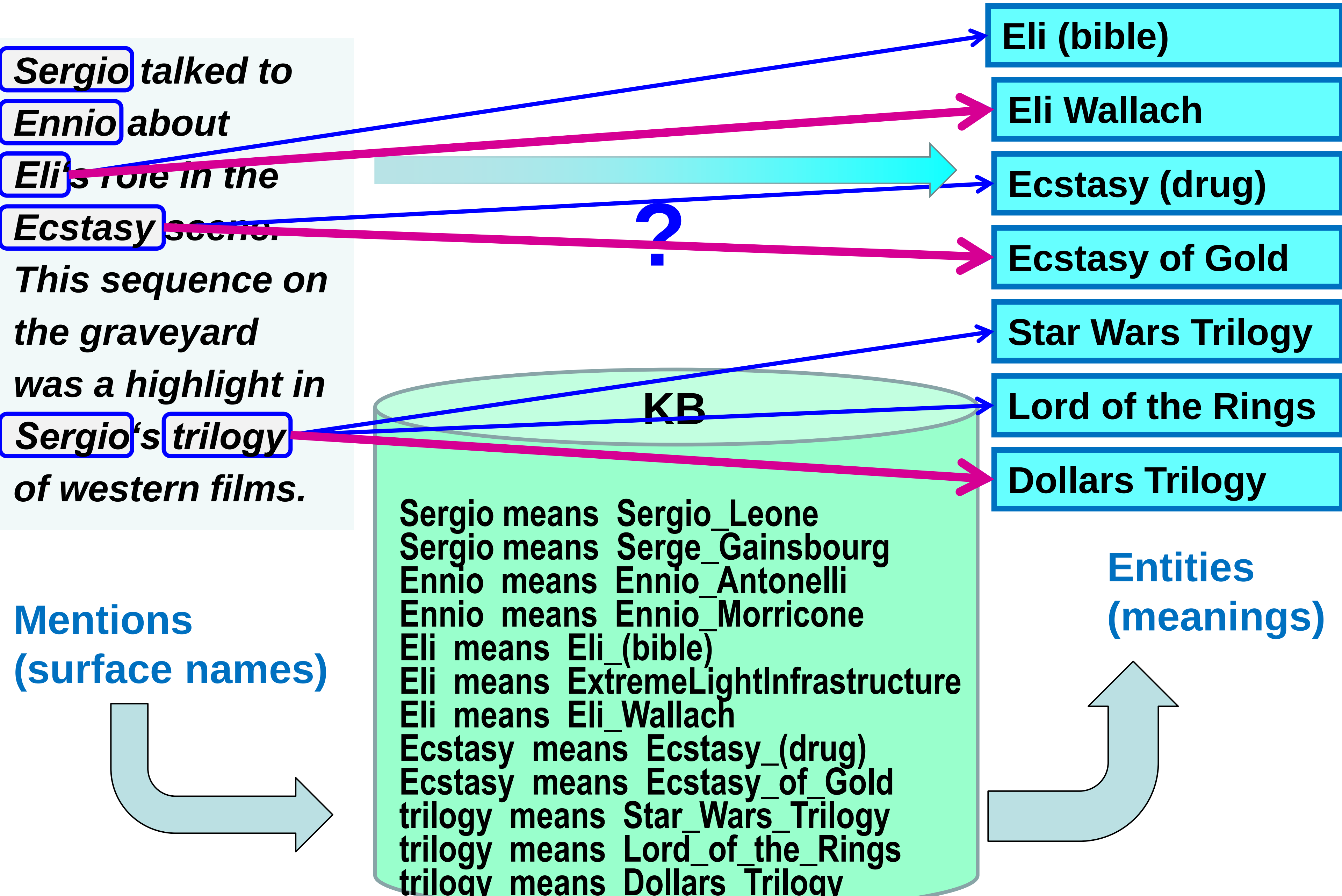
Outline

- ✓ Interesting Data
- ★ From Names to Entities
- ★ From Phrases to Relations
- ★ From Text Analytics to Insight
- ★ Wrap-Up

Names vs. Entities



Named Entity Disambiguation



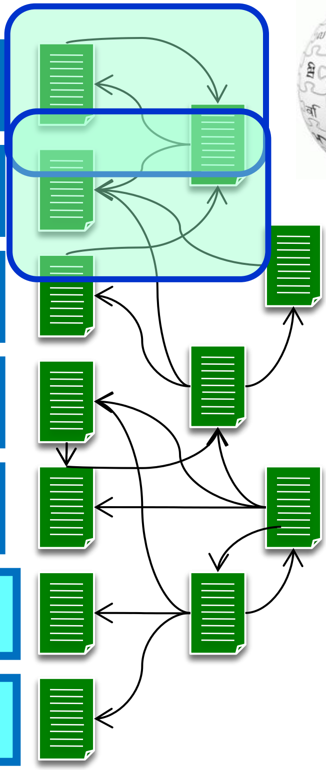
Mention-Entity Graph

weighted undirected graph with two types of nodes

Sergio talked to Ennio about Eli's role in the Ecstasy scene. This sequence on the graveyard was a highlight in Sergio's trilogy of western films.

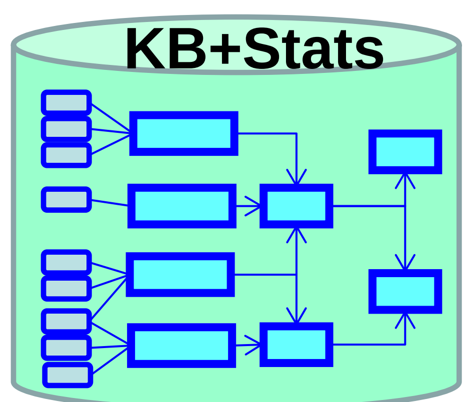
bag-of-words or language model: words, bigrams, phrases

Eli (bible)
Eli Wallach
Ecstasy (drug)
Ecstasy of Gold
Star Wars
Lord of the Rings
Dollars Trilogy



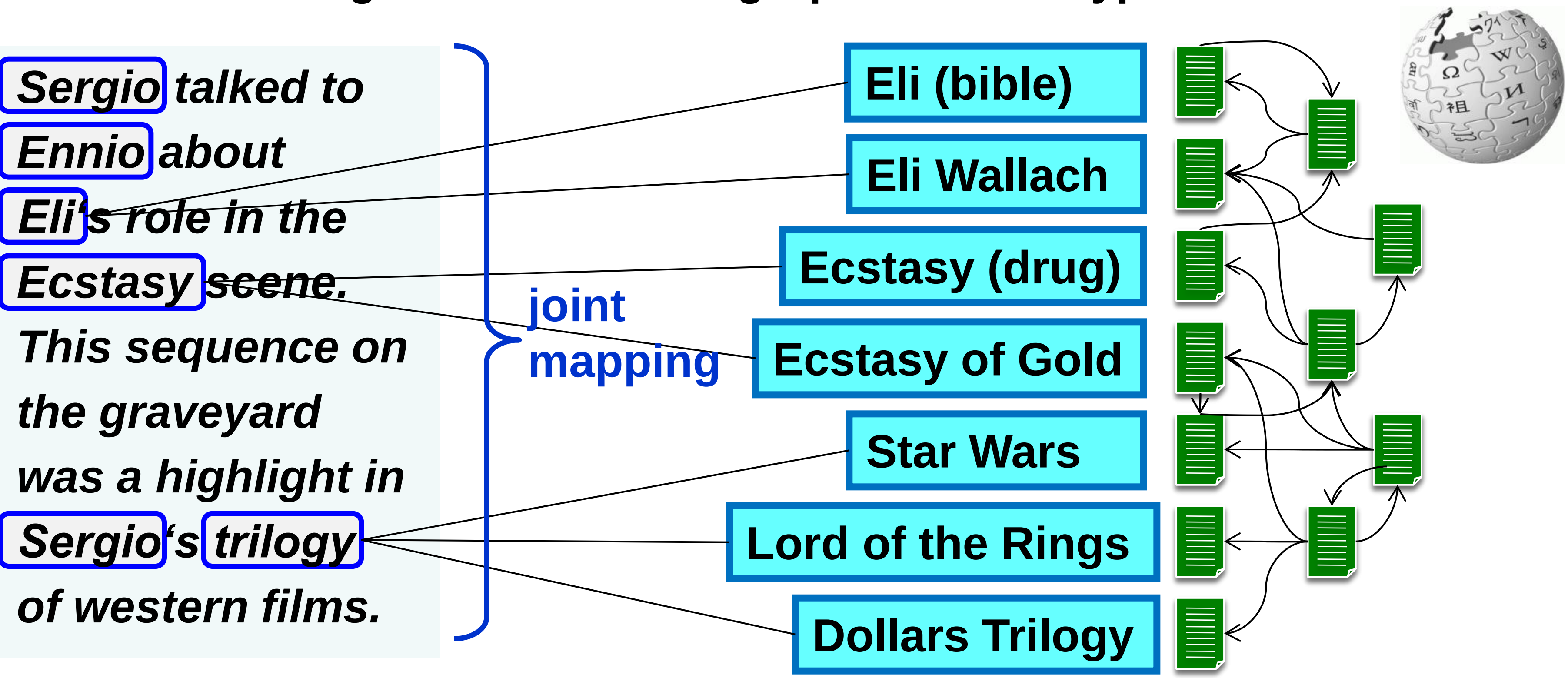
Popularity (m,e):
- freq(e|m)
- length(e)
- #links(e)

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



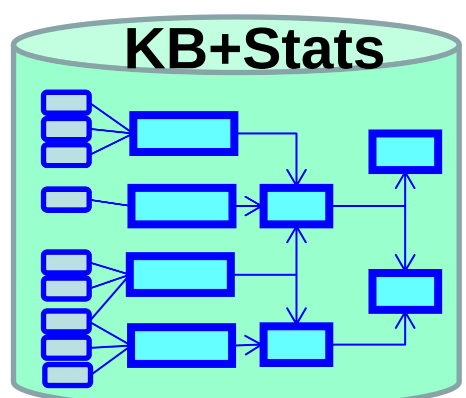
Mention-Entity Graph

weighted undirected graph with two types of nodes



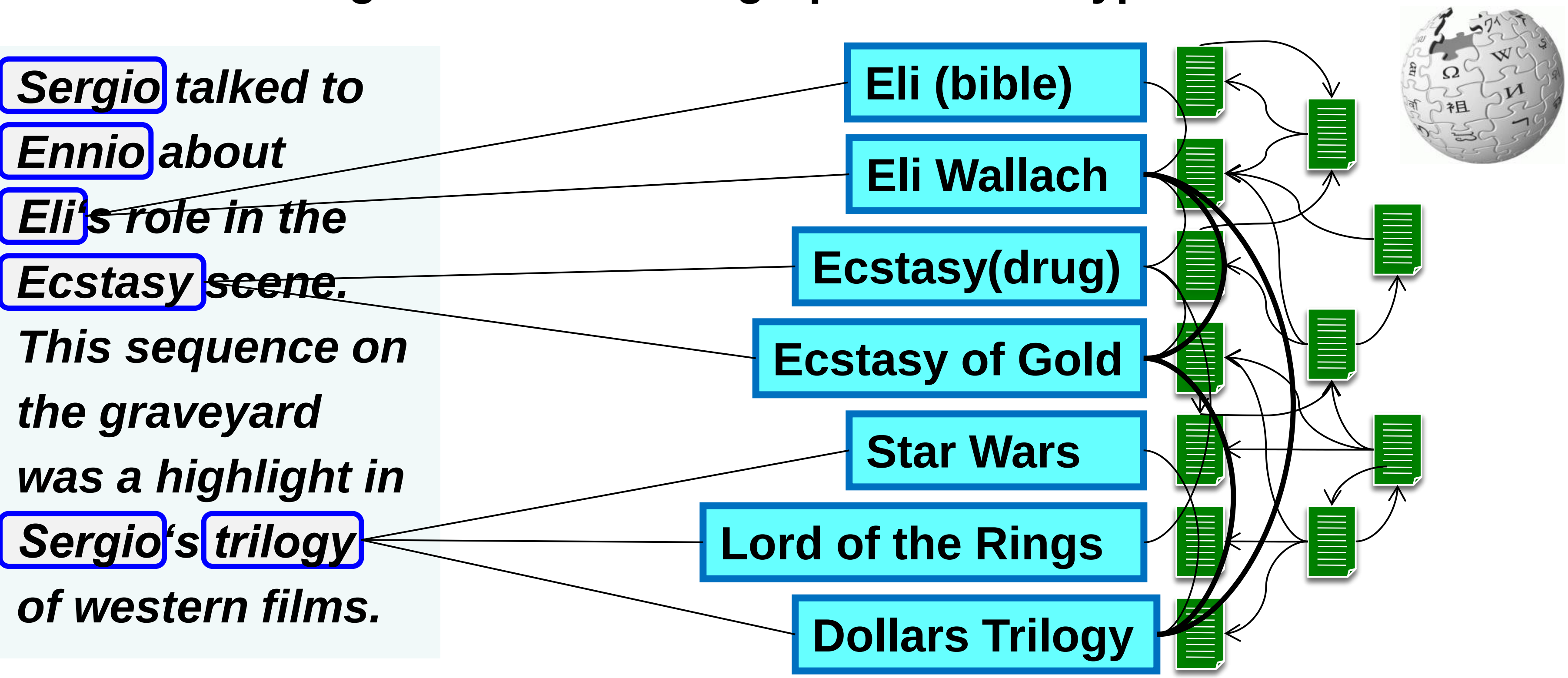
Popularity (m,e):
- freq(e|m)
- length(e)
- #links(e)

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



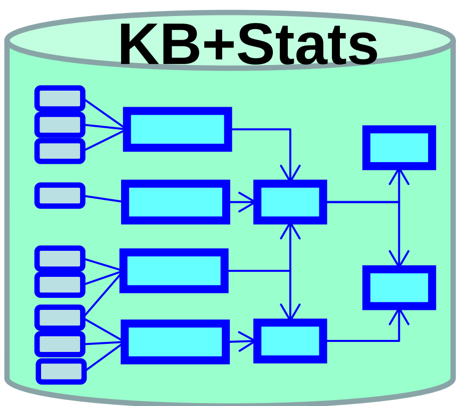
Mention-Entity Graph

weighted undirected graph with two types of nodes



Popularity (m,e):
- freq(m,e|m)
- length(e)
- #links(e)

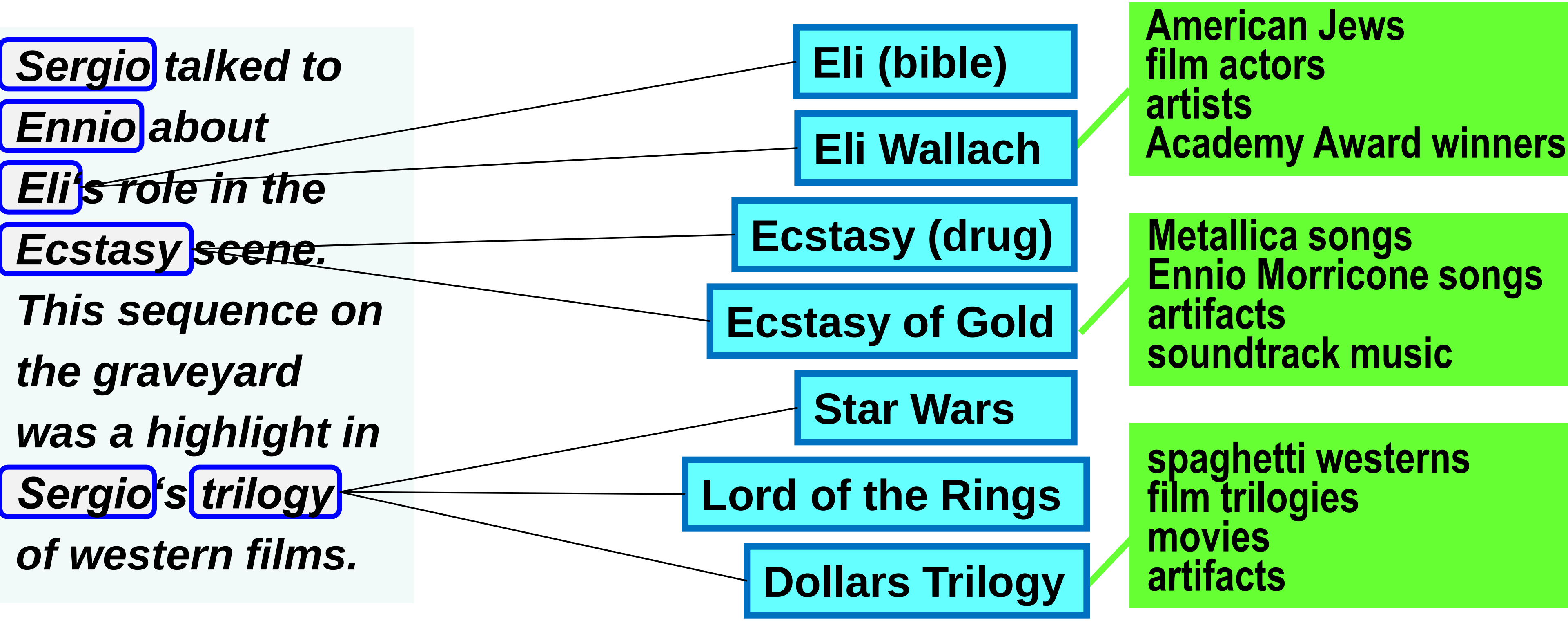
Similarity (m,e):
- cos/Dice/KL (context(m), context(e))



Coherence (e,e'):
- dist(types)
- overlap(links)
- overlap (keyphrases)

Mention-Entity Graph

weighted undirected graph with two types of nodes

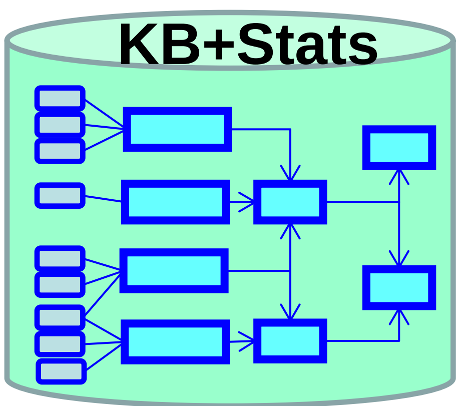


Popularity (m,e):

- freq(m,e|m)
- length(e)
- #links(e)

Similarity (m,e):

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



Coherence (e,e'):

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

Mention-Entity Graph

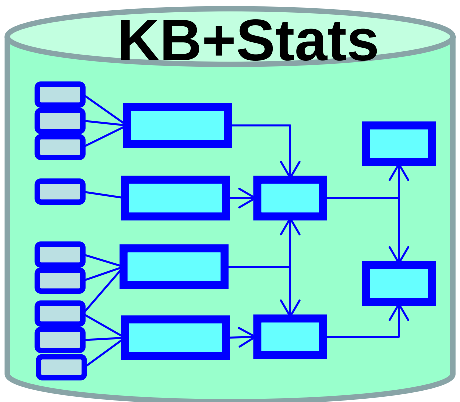
weighted undirected graph with two types of nodes

Sergio talked to Ennio about Eli's role in the Ecstasy scene. This sequence on the graveyard was a highlight in Sergio's trilogy of western films.

- Eli (bible)
- Eli Wallach
  - http://.../wiki/Dollars_Trilogy
  - http://.../wiki/The_Good,_the_Bad,_
  - http://.../wiki/Clint_Eastwood
  - http://.../wiki/Honorary_Academy_A
- Ecstasy (drug)
- Ecstasy of Gold
  - http://.../wiki/The_Good,_the_Bad,_t
  - http://.../wiki/Metallica
  - http://.../wiki/Bellagio_(casino)
  - http://.../wiki/Ennio_Morricone
- Star Wars
- Lord of the Rings
- Dollars Trilogy
  - http://.../wiki/Sergio_Leone
  - http://.../wiki/The_Good,_the_Bad,_
  - http://.../wiki/For_a_Few_Dollars_M
  - http://.../wiki/Ennio_Morricone

Popularity (m,e):
- freq(m,e|m)
- length(e)
- #links(e)

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

Sergio talked to Ennio about Eli's role in the Ecstasy scene. This sequence on the graveyard was a highlight in Sergio's trilogy of western films.

- Eli (bible)
- Eli Wallach
- Ecstasy (drug)
- Ecstasy of Gold
- Star Wars
- Lord of the Rings
- Dollars Trilogy

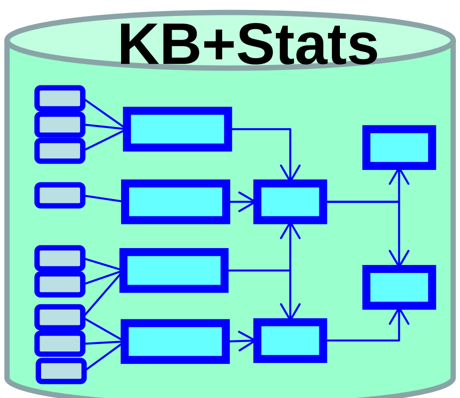
The Magnificent Seven
The Good, the Bad, and the Ugly
Clint Eastwood
University of Texas at Austin

Metallica on Morricone tribute
Bellagio water fountain show
Yo-Yo Ma
Ennio Morricone composition

For a Few Dollars More
The Good, the Bad, and the Ugly
Man with No Name trilogy
soundtrack by Ennio Morricone

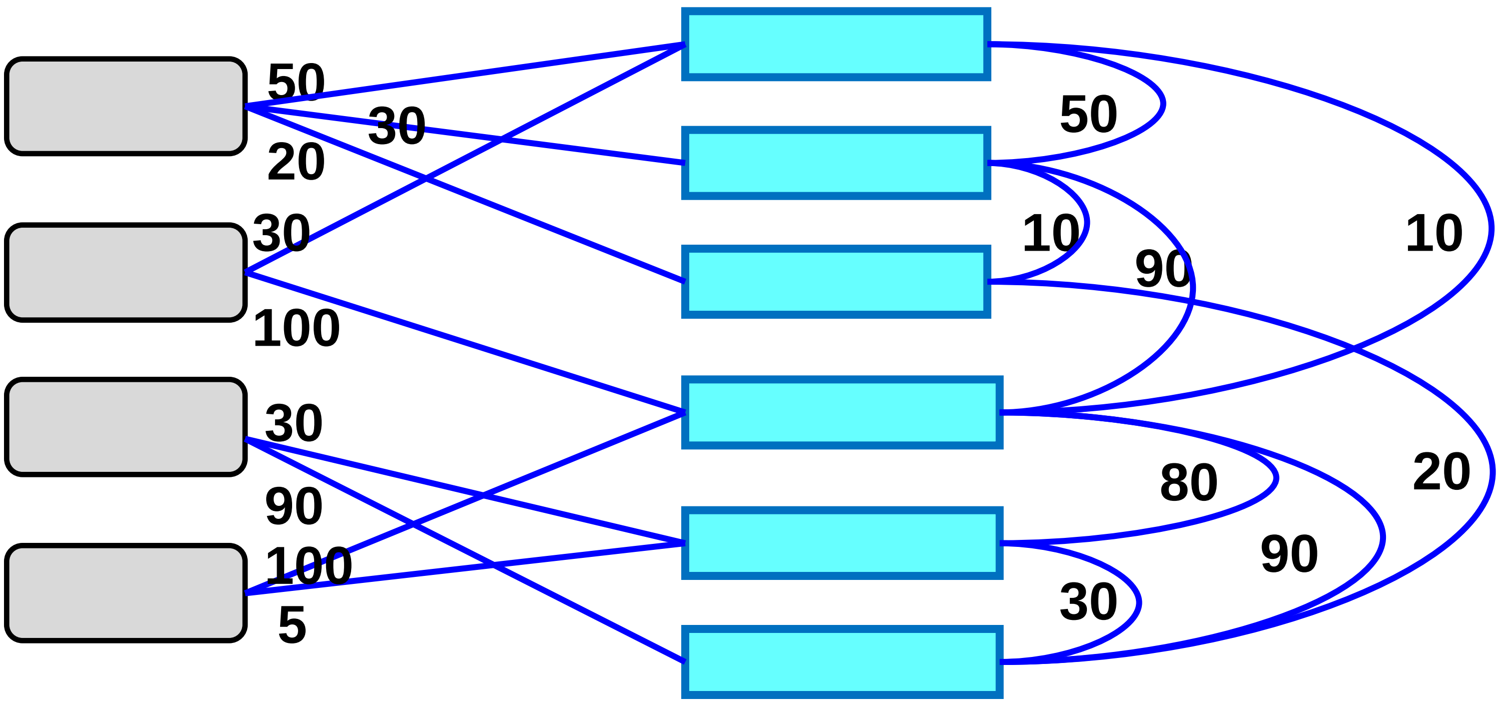
Popularity (m,e):
- freq(m,e|m)
- length(e)
- #links(e)

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



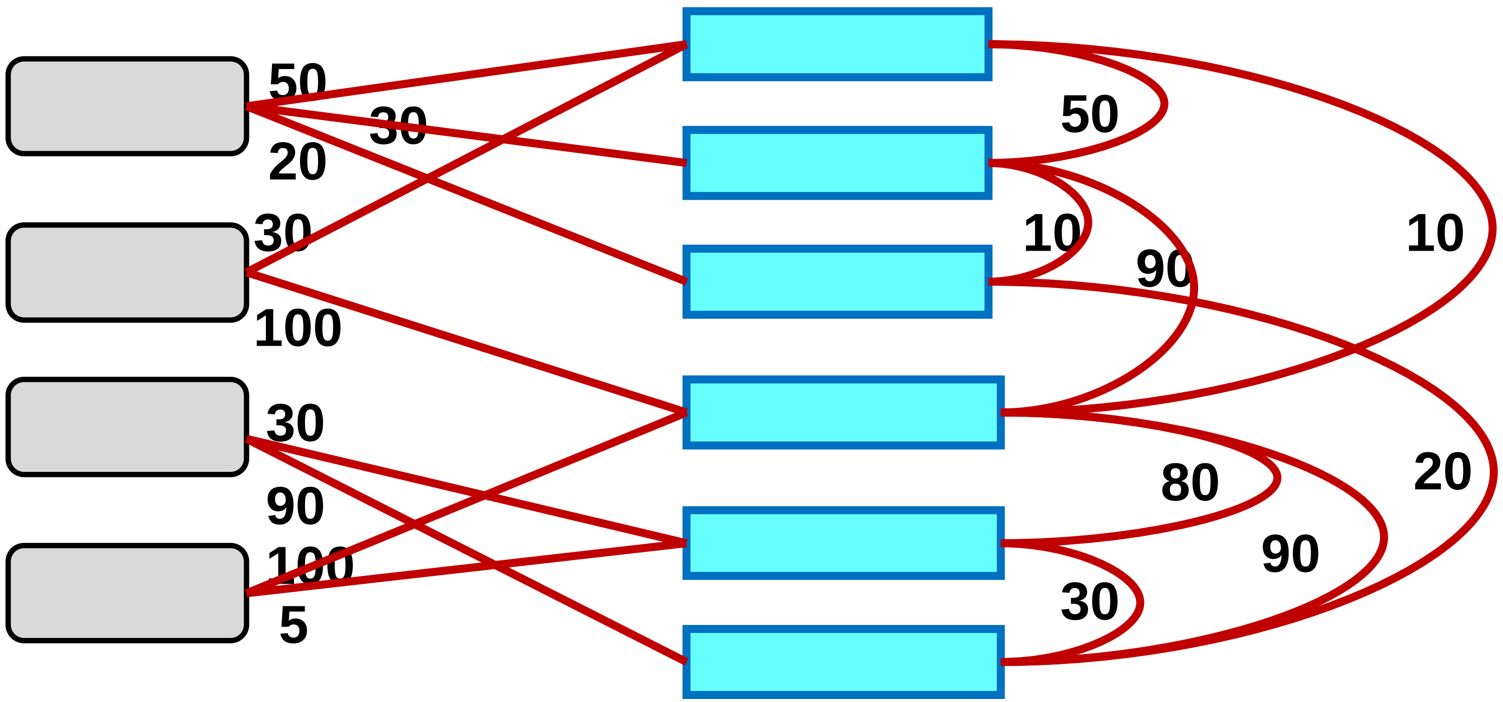
Coherence (e,e'):
- dist(types)
- overlap(links)
- overlap (keyphrases)

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)

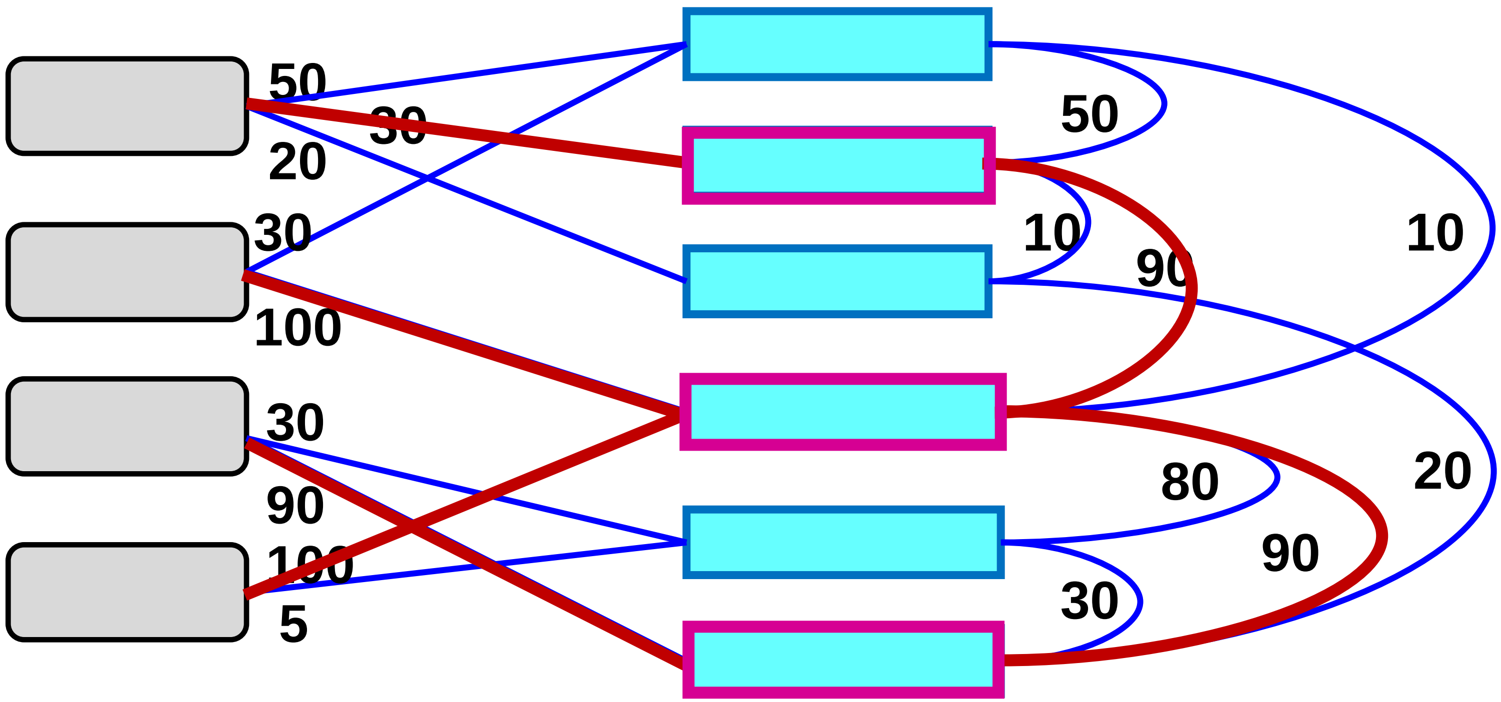
Joint Mapping: Prob. Factor Graph



Collective Learning with Probabilistic Factor Graphs
[Chakrabarti et al.: KDD'09]:

- **model P[m|e] by similarity and P[e1|e2] by coherence**
- **consider likelihood of P[m1 … mk | e1 … ek]**
- **factorize by all m-e pairs and e1-e2 pairs**
- **use MCMC, hill-climbing, LP etc. for solution**

Joint Mapping: Dense Subgraph



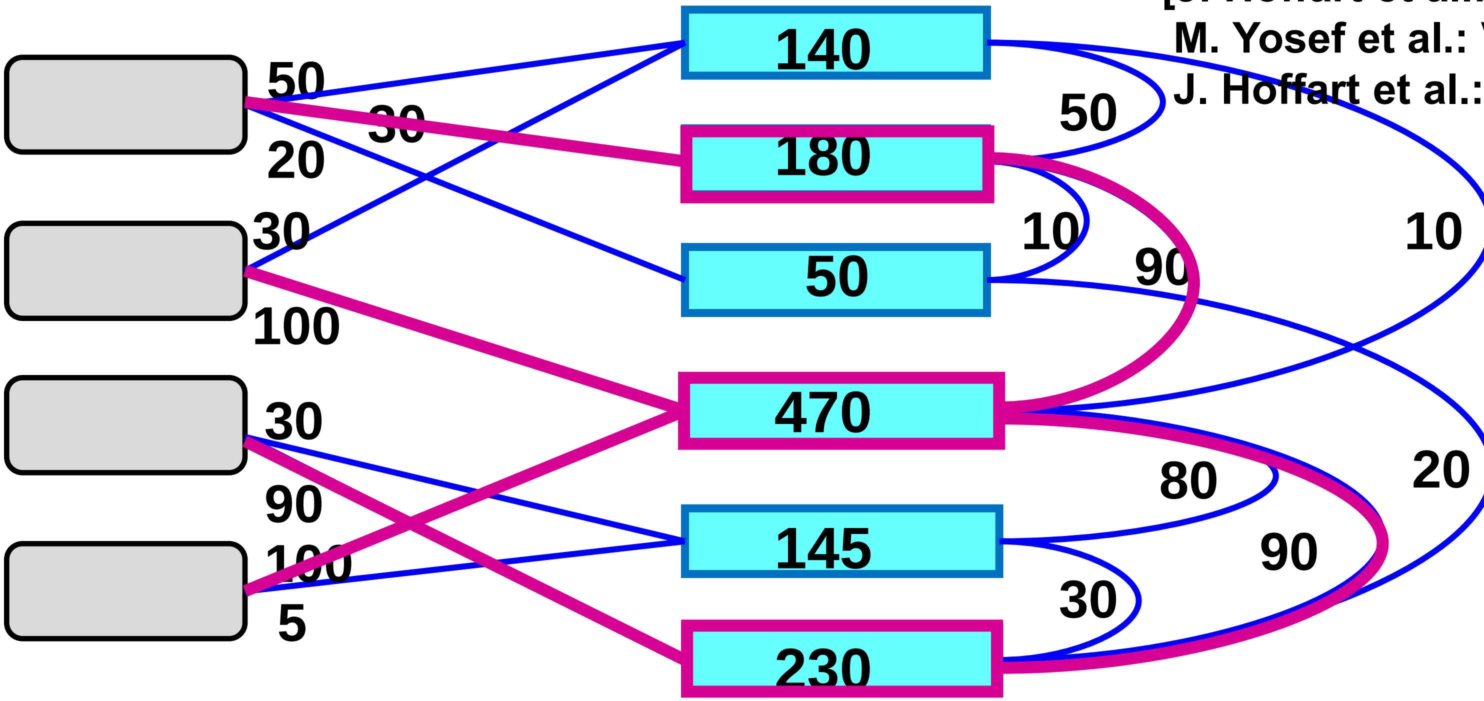
- Compute dense subgraph:

  Maximize total edge weight in subgraph such that each m is connected to exactly one e (or at most one e)
- NP-hard $\rightarrow$ approximation algorithms
- Alt.: feature engineering for similarity-only method

  [Bunescu/Pasca 2006, Cucerzan 2007, Milne/Witten 2008, …]

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/

Keyphrases for Mention-Entity Similarity

Precompute characteristic keyphrases q for each entity e: anchor texts or noun phrases in e page with high PMI:

$$weight(q,e) = \log \frac{freq(q,e)}{freq(q)\, freq(e)}$$

„Metallica tribute to Ennio Morricone"

Match keyphrase q of candidate e in context of mention m

$$score(q\,|\,e) \sim \frac{\#\, matching\ words}{length\ of\ cover(q)} \left(\frac{\sum_{w \in cover(q)} weight(w\,|\,e)}{\sum_{w \in q} weight(w\,|\,e)} \right)^{1+\gamma}$$

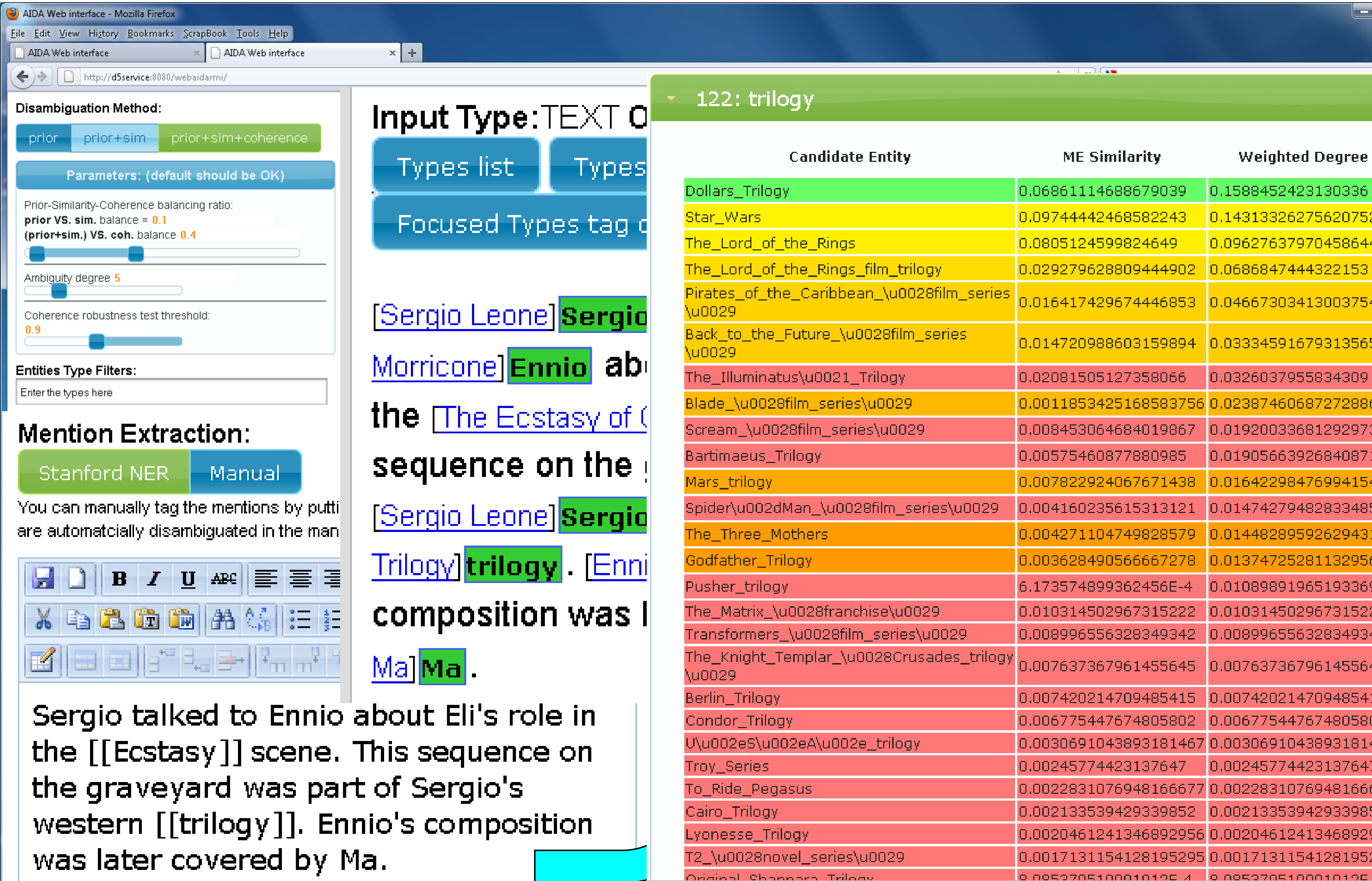
Extent of partial matches **Weight of matched words**

The Ecstasy piece was covered by $\boxed{\text{Metallica on the Morricone tribute}}$ album.

Compute overall similarity of context(m) and candidate e

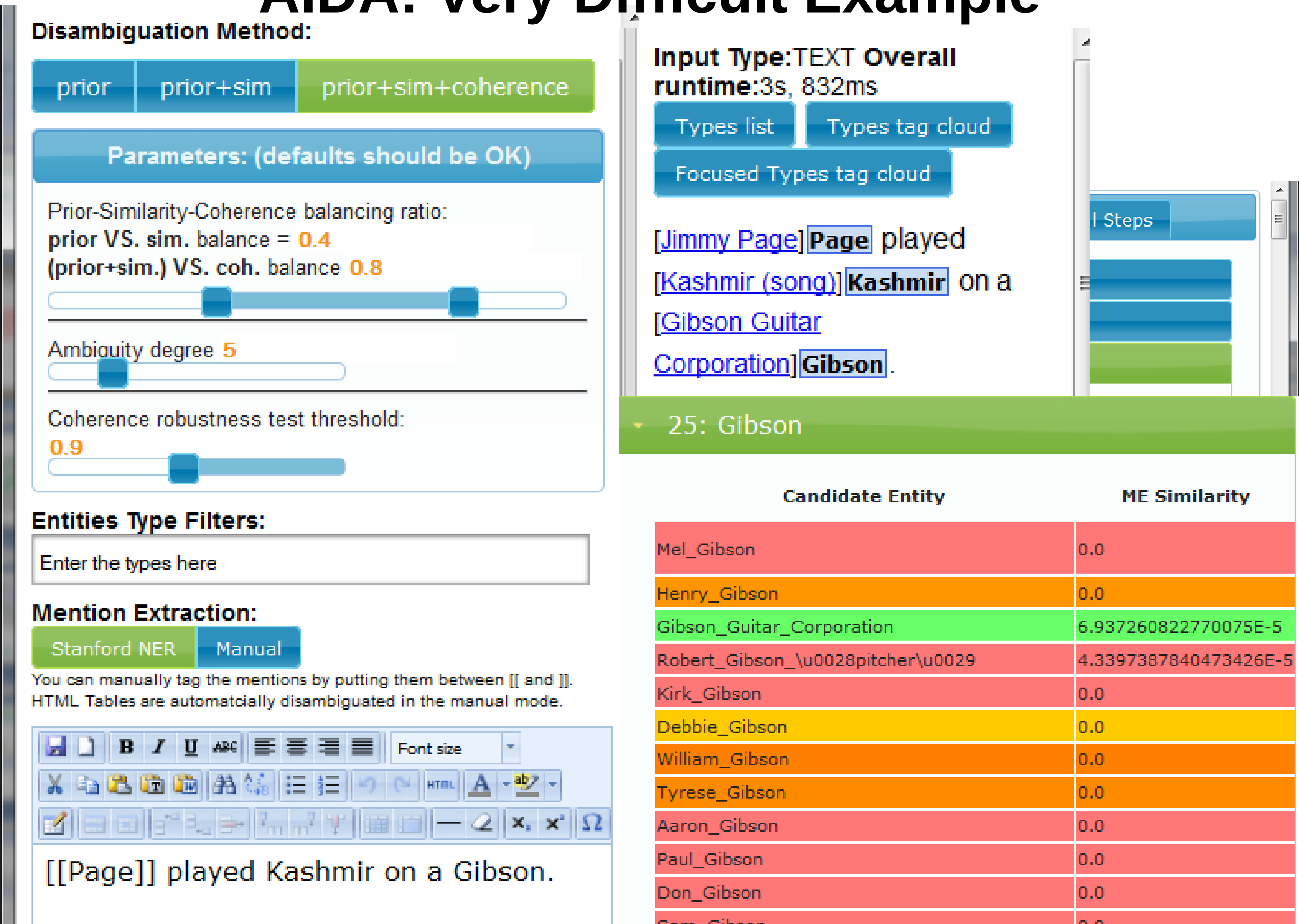
$$score(e\,|\,m) \sim \sum_{\substack{q \in keyphrases(e) \\ in\ context(m)}} score(q)\, dist(cover(q), m)^{-\alpha}$$

AIDA: Accurate Online Disambiguation

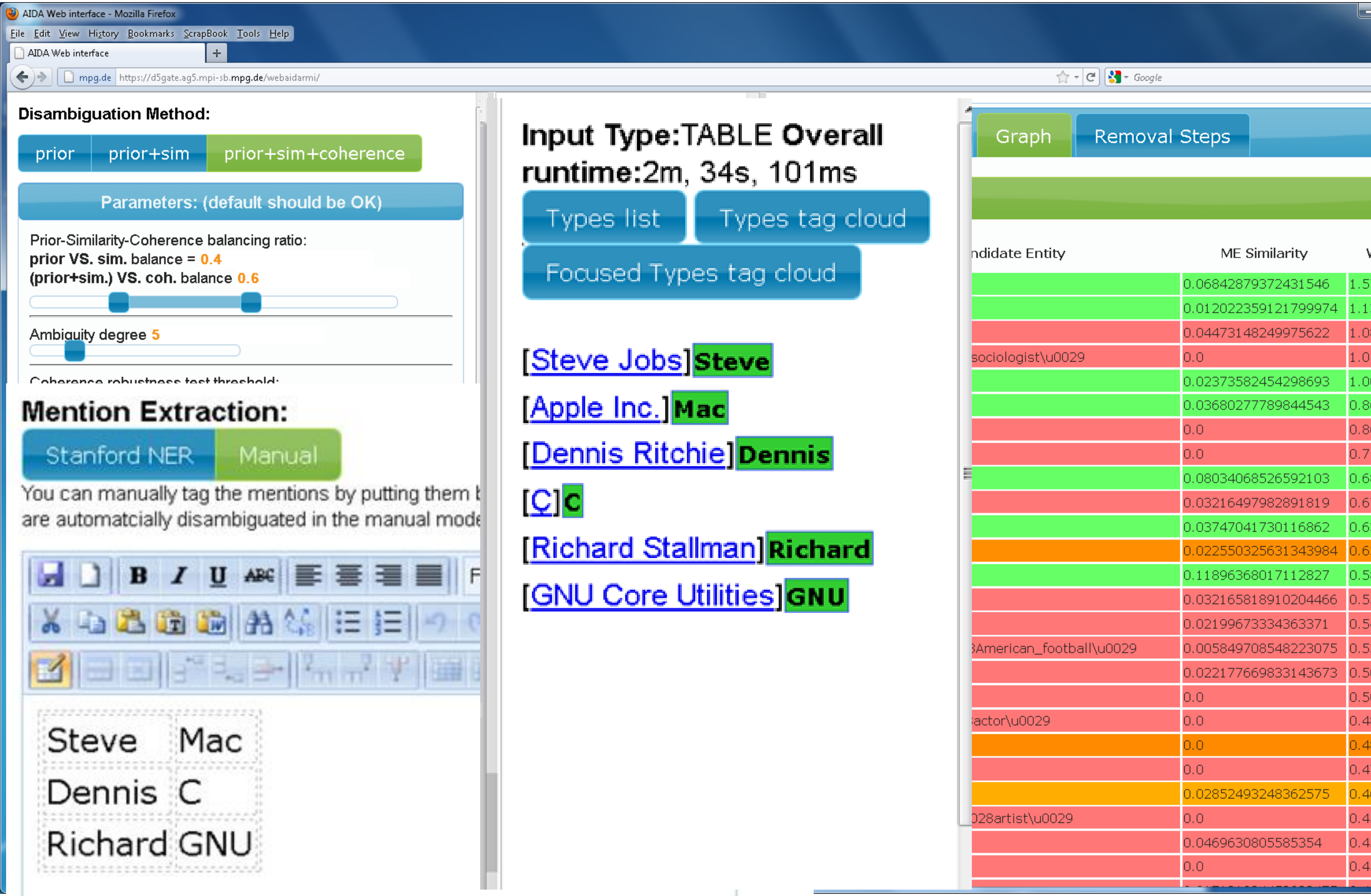


://www.mpi-inf.mpg.de/yago-naga/aida/

AIDA: Very Difficult Example



AIDA: Web Tables



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

NED: Experimental Evaluation

Benchmark:

- Extended CoNLL 2003 dataset: 1400 newswire articles
- originally annotated with mention markup (NER), now with NED mappings to Yago and Freebase
- difficult texts:

… Australia beats India … → Australian_Cricket_Team
… White House talks to Kreml … → President_of_the_USA
… EDS made a contract with … → HP_Enterprise_Services

Results:

Best: AIDA method with prior+sim+coh + robustness test
82% precision @100% recall, 87% mean average precision
Comparison to other methods, see [Hoffart et al.: EMNLP'11]

see also [P. Ferragina et al.: WWW'13] for NERD benchmarks

NERD Online Tools

J. Hoffart et al.: EMNLP 2011, VLDB 2011
https://d5gate.ag5.mpi-sb.mpg.de/webaida/

P. Ferragina, U. Scaella: CIKM 2010
http://tagme.di.unipi.it/

R. Isele, C. Bizer: VLDB 2012
http://spotlight.dbpedia.org/demo/index.html

Reuters Open Calais: http://viewer.opencalais.com/

Alchemy API: http://www.alchemyapi.com/api/demo.html

S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti: KDD 2009
http://www.cse.iitb.ac.in/soumen/doc/CSAW/

D. Milne, I. Witten: CIKM 2008
http://wikipedia-miner.cms.waikato.ac.nz/demos/annotate/

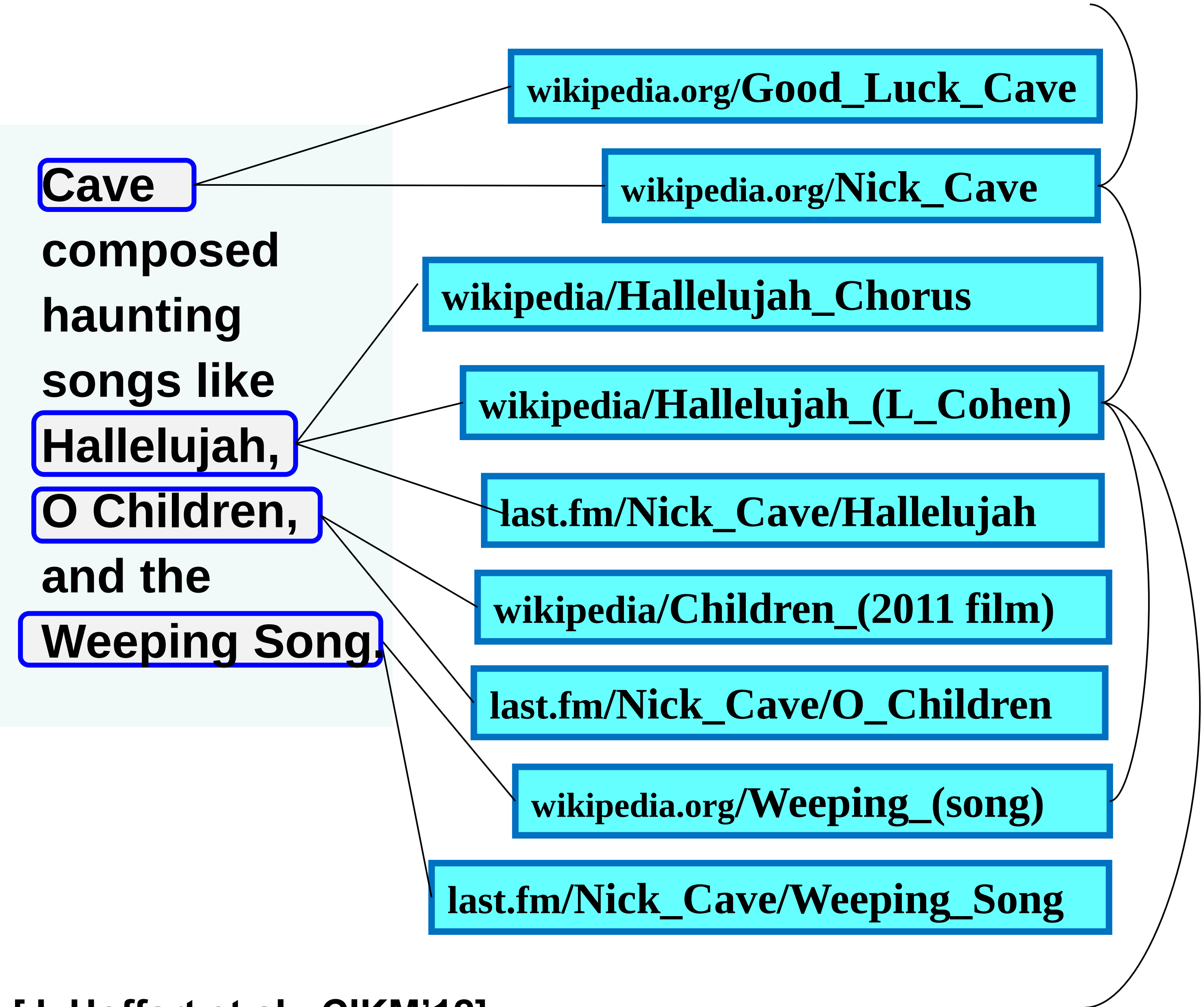
L. Ratinov, D. Roth, D. Downey, M. Anderson: ACL 2011
http://cogcomp.cs.illinois.edu/page/demo_view/Wikifier

some use Stanford NER tagger for detecting mentions
http://nlp.stanford.edu/software/CRF-NER.shtml

Ongoing Research & Remaining Challenges

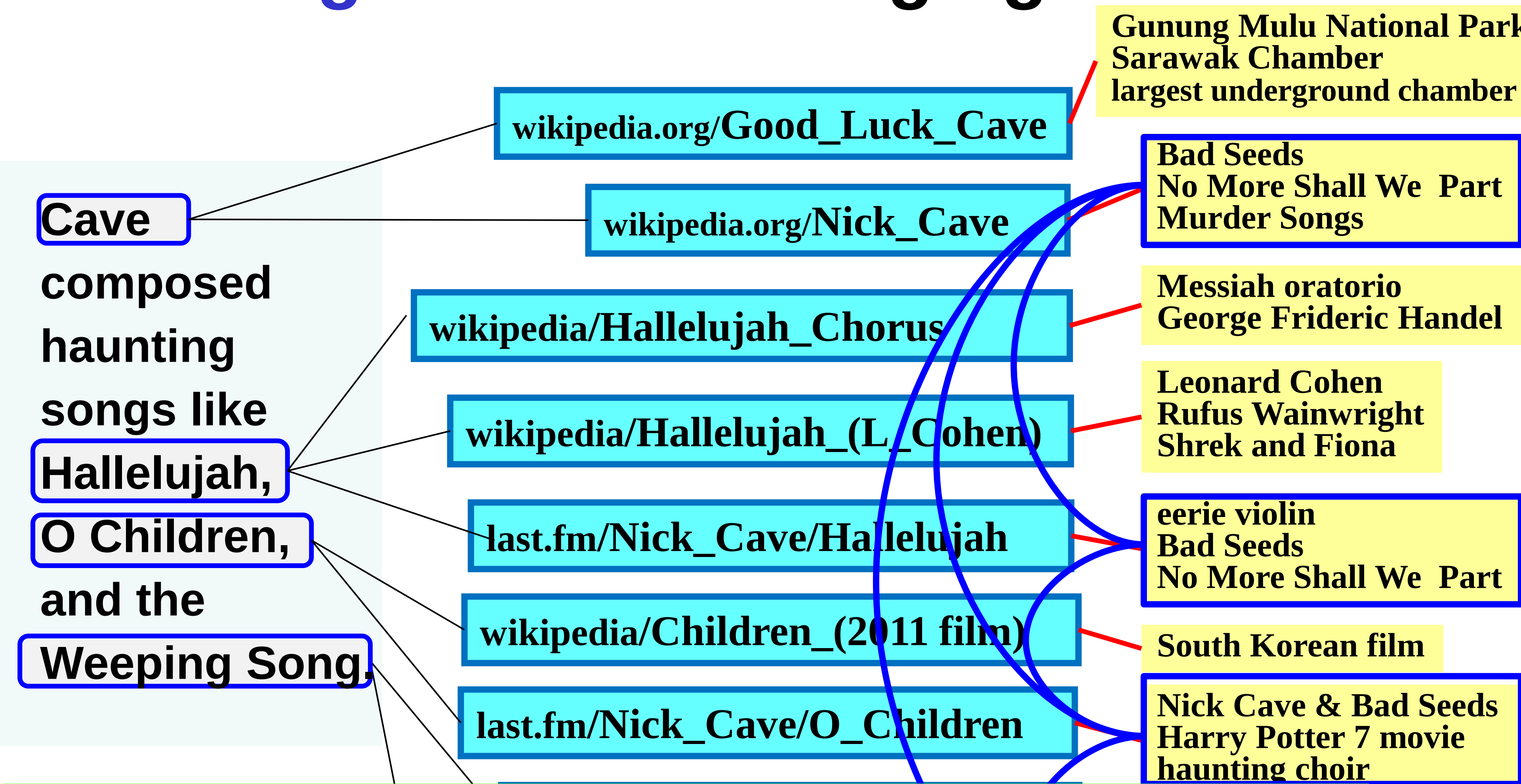
- More efficient graph algorithms (multicore, etc.)
- High-throughput NERD with batch or stream input
- Long-tail and newly emerging entities
- Short, very long, and difficult texts:
  - tweets, headlines, books, dissertations, etc.
  - fictional texts: novels, song lyrics, etc.
- Structured Web data: tables and lists
- Disambiguation beyond entity names:
  - coreferences: pronouns, paraphrases, etc.
  - common nouns, verbal phrases (general WSD)

Long-Tail and Emerging Entities



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

Long-Tail and Emerging Entities



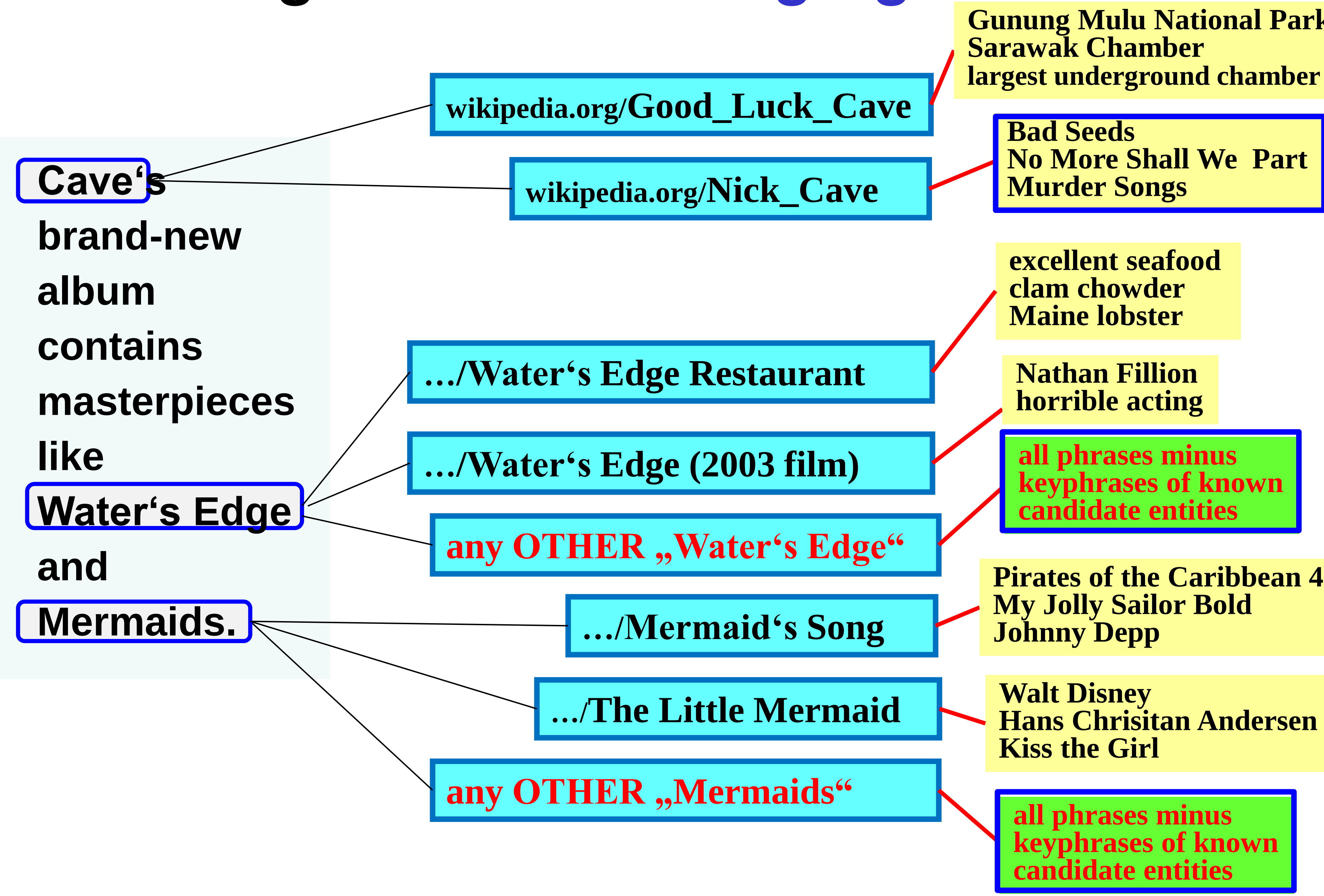
KO (p,q) = $\frac{\sum_t min(weight(t\ in\ p), weight(t\ in\ q))}{\sum_t max(weight(t\ in\ p), weight(t\ in\ q))}$

KORE (e,f) ~ $\sum_{p\in e, q\in f)} KO(p,q)^2 \times min(weight(p\ in\ e), weight(q\ in\ f))$

implementation uses min-hash and LSH

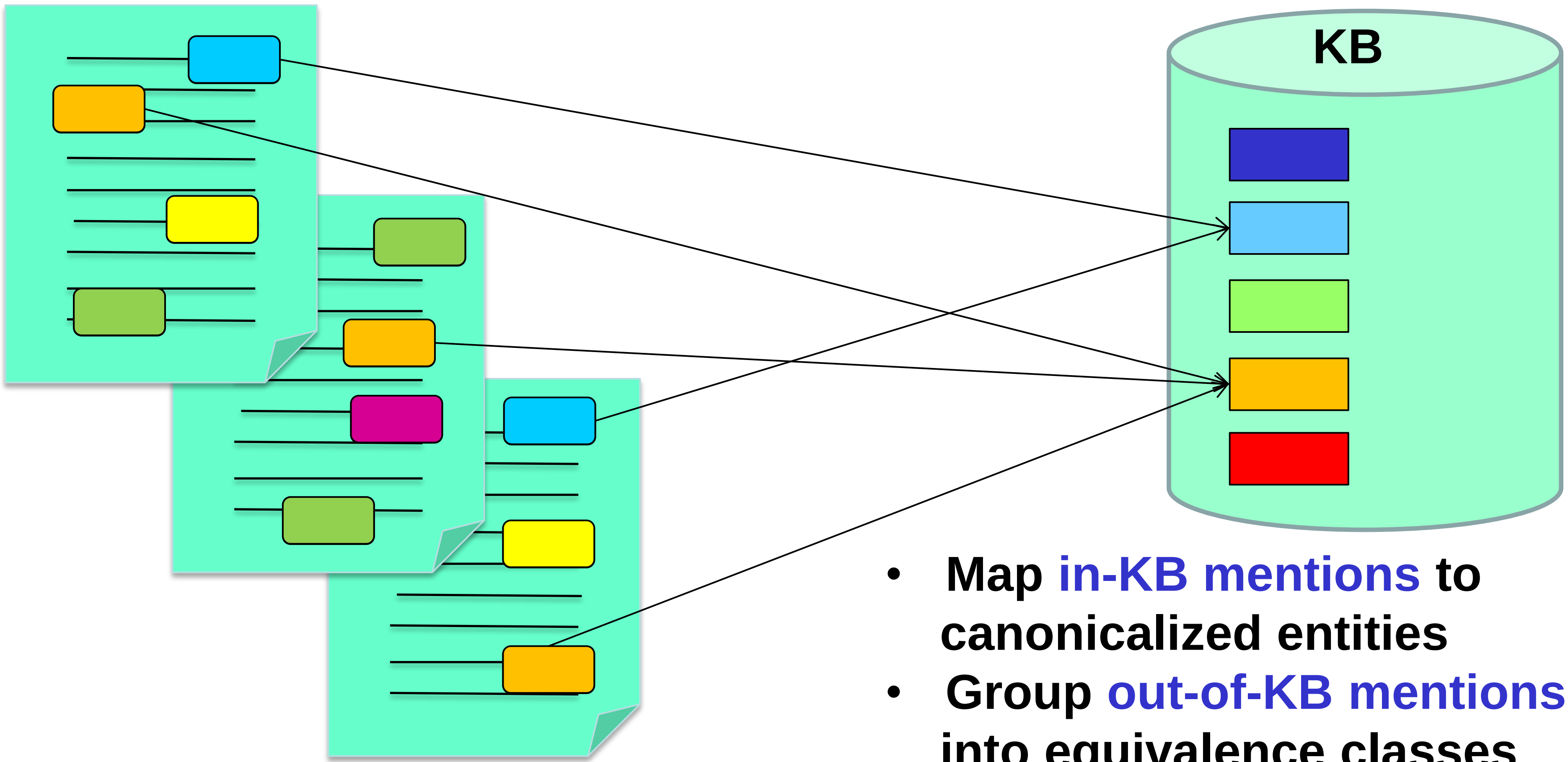
[J. Hoffart et al.: CIKM'12]

Long-Tail and Emerging Entities



Towards Integrated NERD and CCR

CCR = Cross-Document Coreference Resolution
(text counterpart of Entity Resolution for Struct. DB‘s)



- **Map in-KB mentions to canonicalized entities**
- **Group out-of-KB mentions into equivalence classes**

Opportunity & Challenge:
exploit mutual reinforcement of good NERD and good CCR

Big Data Algorithms at Work

Web-scale keyphrase mining

Web-scale entity-entity statistics

MAP on large prob. factor graph or dense subgraphs in large graph

data+text queries on huge KB or LOD

Applications to large-scale input batches:

- **discover all musicians in a week's social media postings**
- **identify all diseases & drugs in a month's publications**
- **track a (set of) politician(s) in a decade's news archive**

Outline

- ✓ What and Why
- ★ From Names to Entities
- ★ From Phrases to Relations
- ★ From Text Analytics to Insight
- ★ Wrap-Up

Diversity and Ambiguity of Relational Phrases

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**

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 soulful 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

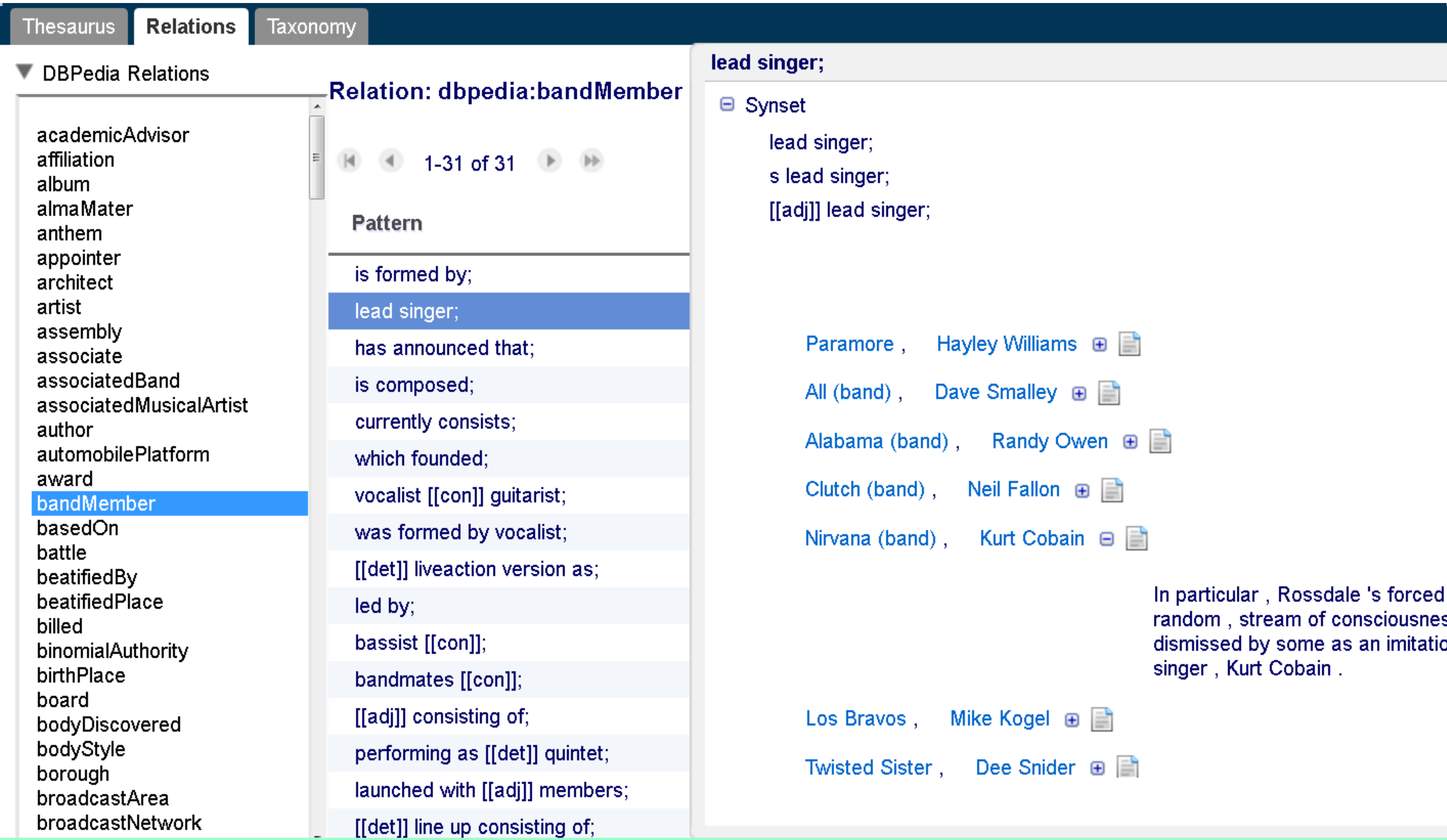
"wife of" ⇒ " spouse of"

350 000 SOL patterns from Wikipedia, NYT archive, ClueWeb

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

PATTY: Pattern Taxonomy for Relations

[N. Nakashole et al.: EMNLP 2012, demo at VLDB 2012]



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

text pre-processing → n-gram mining → pattern lifting → taxonomy construction

Ongoing Research & Remaining Challenges

- Countering sparseness to refine the pattern subsumption taxonomy
- Coping with (even) larger-scale input (social media, query-and-click logs, …)
- Cost-efficient crowdsourcing for higher coverage & accuracy
- Exploit pattern type signatures for discovering and organizing new entities [N.Nakashole et al.: ACL'13]
- Exploiting pattern synsets for translating questions to queries [M. Yahya et al.: EMNLP'12]

Semantic Typing of Emerging Entities

[N. Nakashole et al.: ACL 2013]

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,y
and all type triples (t1, p, t2) for known entities:

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

For each new e and all candidate types t_i:

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

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

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.

Outline

- ✓ What and Why
- ★ From Names to Entities
- ★ From Phrases to Relations
- ★ From Text Analytics to Insight
- ★ Wrap-Up

Big Data+Text Applications

Entertainment:

Who covered which other singer?
Who influenced which other musicians?

Health: Drugs (combinations) and their side effects

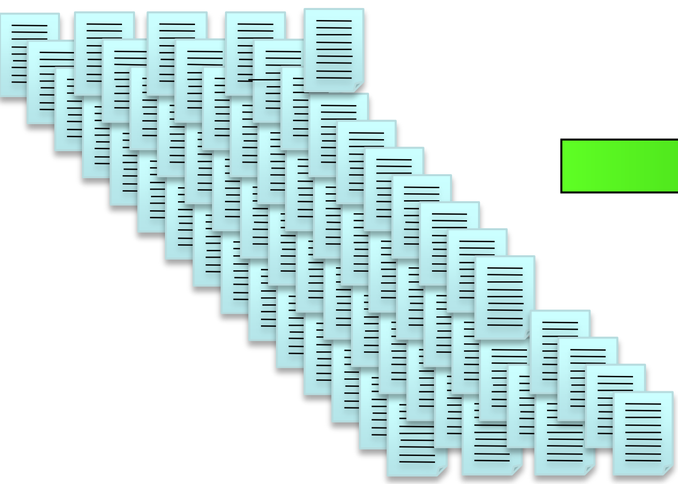
Politics: Politicians' positions on key topics and their involvement with industry

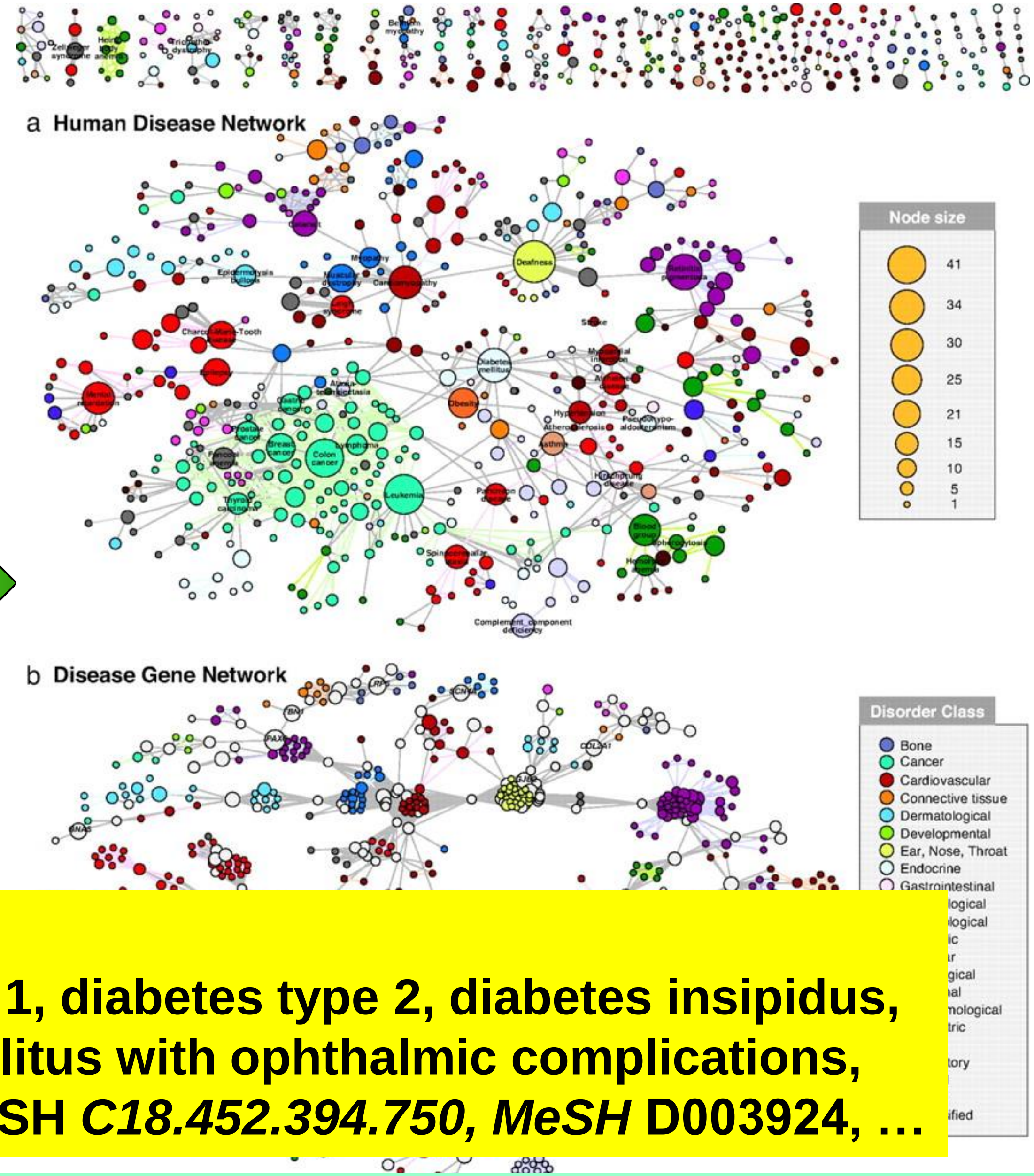
Business: Customer opinions on small-company products, gathered from social media

General Design Pattern:

- Identify entities of interest & their relationships
- Position in time & space
- Group and aggregate
- Find insightful patterns & predict trends

Big Data Analytics for Disease Networks





But try this with:

diabetes mellitus, diabetis type 1, diabetes type 2, diabetes insipidus, insulin-dependent diabetes mellitus with ophthalmic complications, ICD-10 E23.2, OMIM 304800, MeSH *C18.452.394.750, MeSH* D003924, …

K.Goh,M.Kusick,D.Valle,B.Childs,M.Vidal,A.Barabasi: The Human Disease Network, PNAS, May 2007

Big Analytics on Data + Text

Example task: Opinion Map on Controversial Topic

consider all news articles and social media postings related to firearms in private homes

- **Find all pro/con opinions, the opinion-holding entities, and their political parties**
- **Group and analyze by party/org, gender, geo-region over time, especially after major incidents**

Challenges at Web Scale:

- **Phrase mining (variable-length n-grams) for direct & indirect sentiments**
- **Entity recognition & disambiguation for people, organizations, locations, events**
- **Classification models for gender, pro/con, …**

Outline

- ✓ What and Why
- ★ From Names to Entities
- ★ From Phrases to Relations
- ★ From Text Analytics to Insight
- ★ Wrap-Up

Summary

- **Structured and Unstructured Data :**
  Entities & Relations are Key to Connect Both Worlds
- **Diversity & Ambiguity of Names and Phrases**
  Calls for Disambiguation Mapping
- **Good Story for Entity Name Disambiguation**
- **Ongoing Work on Relation Phrase Disambiguation**
- **Entities for Big Data Analytics:**
  Web contents, Data+Text, …, with KB, …
- **Key to Future Tapping into Speech, Video, …**

Take-Home Message

Solve analytic task: „Who covered whom?“

⇒ map ambiguous names and phrases into entities and relations for Big Data analytics over text, speech, …