



max planck institut
informatik

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?

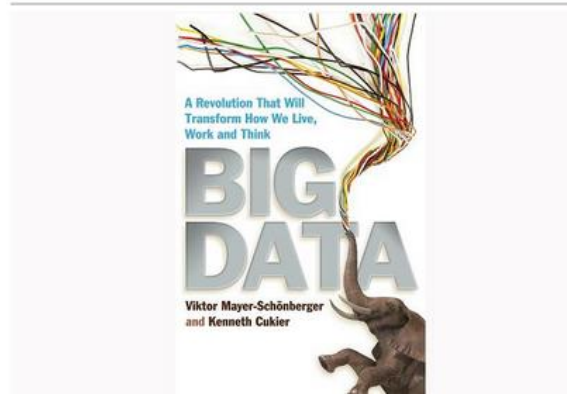


“Why do you want to climb it?”

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

Welcome to your preview of The Times

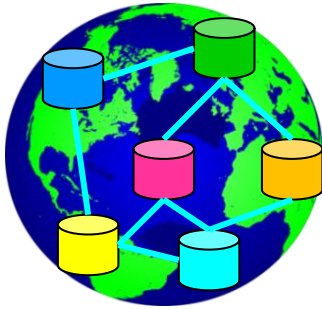
Big Data: A revolution that will transform the way we live, work and think



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

The Promise of Big Data

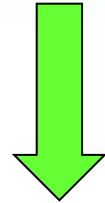
Energy
Traffic
Health
Business
.....



**Big Data
Analytics**



*Platforms & Algorithms
Scale, Scale, Scale ...*



News
Books
Social Media
Scholarly Publ.
.....



???

**make the world
a better place !**

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 12°27'S 130°50'E	April	31.4159°C
	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).*

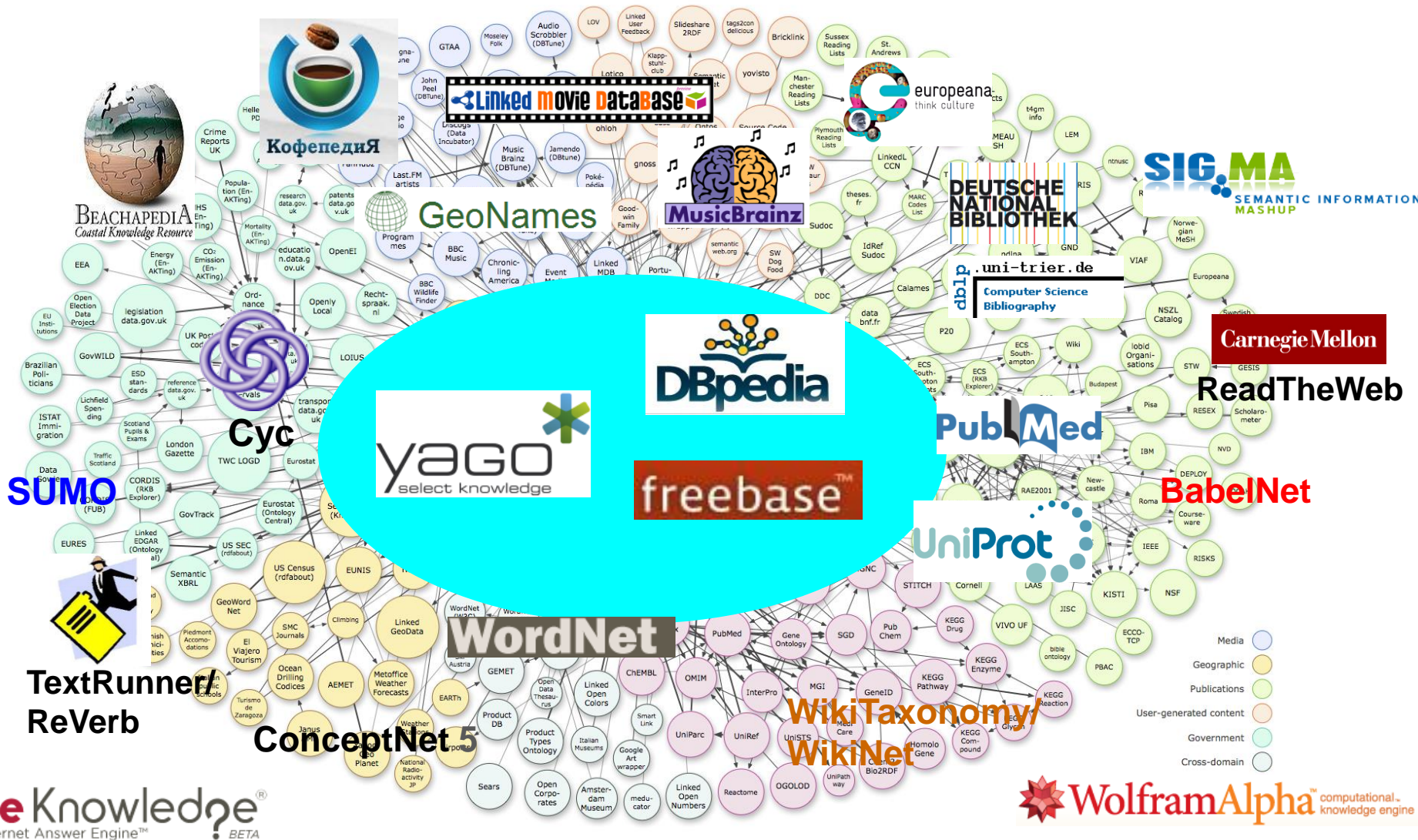
unclear to interpret



insightful to human

Interesting Data at Scale: LinkedOpenData

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



Linked Open Data

62 Bio. SPO triples (RDF)

+ linked text sources

Amy Winehouse's Back to Black sets chart record

Amy Winehouse's *Back to Black* album has become the biggest-selling album in the UK in the 21st Century, the Official Charts Company (OCC) has revealed.

It said her 2006 second album had this week overtaken sales of James Blunt's 2005 work *Back to Bedlam*.

It had sold 3.26 million copies compared with Blunt's 3.25 million, the OCC said on Thursday.

Winehouse's *Back to Black* shot back to the top of the charts for three weeks following her death on 23 July.

Dido's *No Angel* - released in the UK in February 2001 - is the third best-selling album of the century with sales of 3.07 million.

www.bbc.co.uk/news/world-europe-20857664



Winehouse was found dead at her London home on 23 July

Related Stories

Winehouse tribute at MTV awards

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

BBC NEWS ENTERTAINMENT & ARTS

Amy Winehouse's Back to Black sets chart record

25 August 2011 Last updated at 09:25 GMT

Amy Winehouse's Back to Black album has become the biggest-selling album in the UK in the 21st Century, the Official Charts Company (OCC) has revealed.

It said her 2006 second album had this week overtaken sales of James Blunt's 2005 work Back to Bedlam.

It had sold 3.26 million copies compared with Blunt's 3.25 million, the OCC said on Thursday.

Winehouse was found dead at her London home on 23 July.

Winehouse's Back to Black shot back to the top of the charts for three weeks following her death on 23 July.

Related Stories

Winehouse tribute at MTV awards

Dido's No Angel - released in the UK in February 2001 - is the third best-selling album of the century with sales of 3.07 million.

www.bbc.co.uk/news/world-20857664

YouTube

Winehouse Grammy Performance

6:14 / 7:29

694,646 views

Abonnieren 246

1,736 likes

44 comments

Happy 70th Birthday, Jimi Hendrix: Seattle's EMP Museum Celebrates Guitar Hero (PHOTOS)

Posted: 11/27/2012 9:39 am EST updated: 11/27/2012 2:13 pm EST

1,822 people like this. Sign up to see what your friends like.

609 likes, 85 tweets, 23 shares, 23 comments

GET ARTS ALERTS: Enter email SIGN UP

REACT: Amazing Inspiring Funny Scary Hot Crazy Important Weird

FOLLOW: Video, Experience Music Project, Jimi Hendrix 70, Art Exhibit, Art News, Experience Music Project Museum, Jimi Hendrix Art Exhibit, Jimi Hendrix Birthday, Jimi Hendrix EMP Museum, Jimi Hendrix Exhibit, Jimi Hendrix Seattle Museum, Seattle Museum, Arts News

Today is the birthday of electric guitar hero, James Marshall Hendrix. The "Purple Haze" creator would turn 70 years old if he were still alive today, and this fact is blowing my mind.

Happy 70th Birthday, Jimi Hendrix: Seattle's EMP Museum Celebrates Guitar Hero (PHOTOS)

Mercedes Benz

Janis Joplin's

Writers: Janis Joplin, Michael McClure and Bob Newirth

Recorded: 1970

It's Thursday, Oct. 1, at the Sunset Sound recording studio in Los Angeles. Janis Joplin asks producer Paul Rothchild to roll tape. She has a song she'd like to sing.

The services of backing band Full Tilt Boogie, present and ready for action, will not be necessary. Joplin steps to the microphone and makes a declaration. "I'd like to do a song of great social and political import," she says, a

Amy Winehouse dead at 27 like Janis Joplin, Jim Morrison

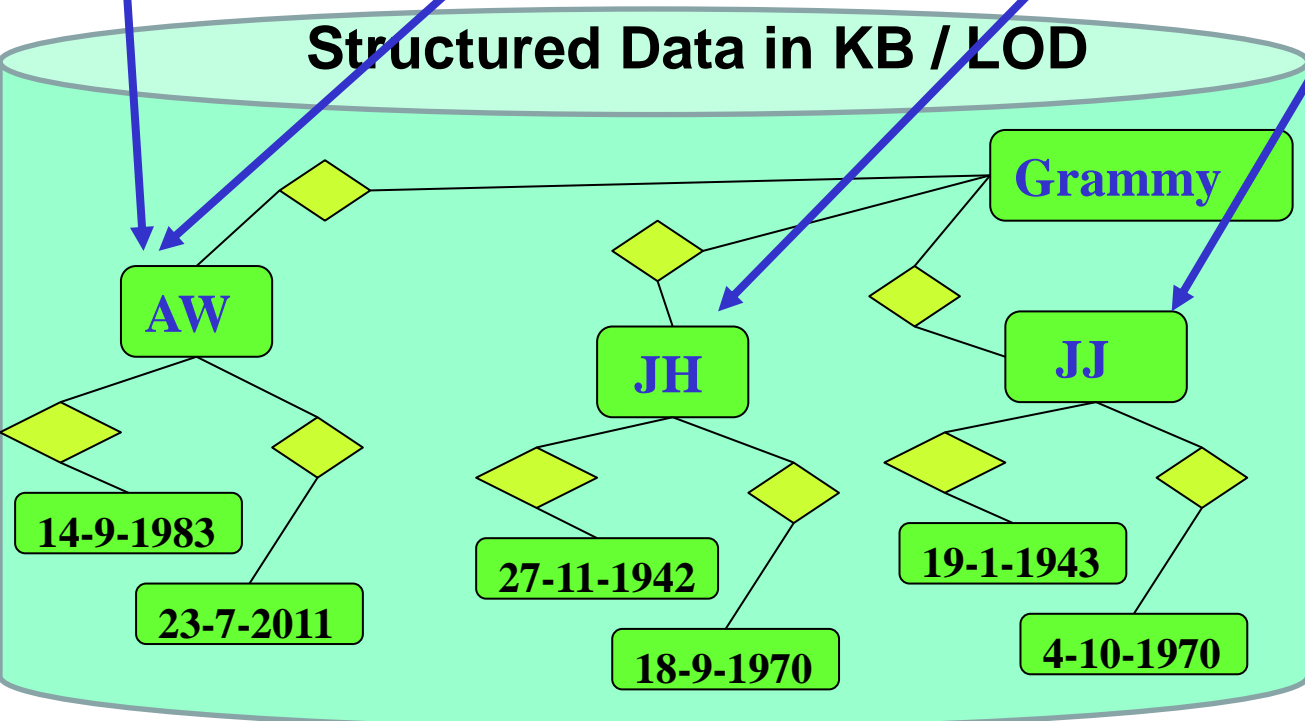
Jan 23, 2011 11:43 AM by Meagan Morris

Posted in Entertainment / Celebrity Gossip / Celebrity Obituaries

WINEHOUSE JOINS THE FOREVER 27 CLUB

Amy Winehouse's tragic death puts her in the sad club of brilliant musicians at age 27.

Amy Winehouse's tragic death reminds us -- once again -- of the dangers of mixing serious addiction with a career in the spotlight. It seems as if Winehouse's family and handlers put an emphasis on her career than her own well-being.



Microdata & Tables in HTML



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

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

Microdata & Tables in HTML



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	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



Ennio Morricone



Ennio Morricone

Samples

Remixes

Covers

- As an Artist
 - Covers by Ennio Morricone [18]
 - Covers of Ennio Morricone songs [67]
- As a Producer
 - Cover songs produced by Ennio Morricone [4]
 - Covers of songs produced by Ennio Morricone [7]

Covers of Ennio Morricone Songs [67]

Sort



Meglio Stasera (1963) was covered in
▶ **Meglio Stasera** by **Mondo Candido** (2003)



Svegliati e Uccidi (1966) was covered in
▶ **Svegliati & Uccidi** by **John Zorn** (2000)



The Ecstasy of Gold (1966) was covered in
▶ **The Ecstasy of Gold** by **Metallica** (1999)
▶ **The Ecstasy of Gold** by **Glomag** (2009)



Faccia a Faccia (Titoli) (1967) was covered in
▶ **Face to Face With a Couple Axes** by **Alvarius B.** (2011)

<http://www.whosampled.com>

Samples

Remixes

Covers

- As an Artist
 - Tracks sampled by Ennio Morricone [9]
 - Tracks that sampled Ennio Morricone [160]
- As a Producer
 - Tracks that sampled music produced by Ennio Morricone [8]

Tracks Sampled by Ennio Morricone [9]

Sort



La Resa Dei Conti (1965) sampled
▶ **Tocatta and Fugue in D Minor** by **Johann Sebastian Bach** (1707)



Dopo La Condanna (1966) sampled
▶ **Für Elise** by **Ludwig Van Beethoven** (1810)



Tema Italiano (1969) sampled
▶ **Prelude and Fugue in a Minor, BWV 543** by **Johann Sebastian Bach** (1717)



Valkyries (1973) sampled
▶ **Ride of the Valkyries** by **Richard Wagner** (1854)



Anna (1973) sampled
▶ **Für Elise** by **Ludwig Van Beethoven** (1810)

Microdata & Tables in HTML

100 Mio's of structured tables

+ (con)text

Ennio Morricone

Ennio Morricone

Cover Details

[Like](#) 0 [Send](#) [+1](#) 0 [Share on](#) [t](#) [f](#) [+](#)

Metallica
The Ecstasy of Gold
S&M
Elektra 1999

Ennio Morricone
The Ecstasy of Gold
The Good, the Bad, and the Ugly OST
EMI 1966

We all love
ENNIO MORRICONE

- ANDREA BOCELLI
- CHRIS BOTTI
- EUMIR DEODATO
- CELINE DION
- RENEE FLEMING
- DENYCE GRAVES
- TARO HAKASE
- HERBIE HANCOCK
- QUINCY JONES
- YO-YO MA
- DANIELA MERCURY
- METALLICA
- DULCE PONTES
- BRUCE SPRINGSTEEN
- VANESSA AND THE O'S
- ROGER WATERS

The Good, the Bad, and the Ugly

Other covers of Ennio Morricone's The Ecstasy of Gold:

The Ecstasy of Gold by **Glomag** (2009)

Remixes of The Ecstasy of Gold:

L'Estasi Dell'Oro (Bandini Remix) remix by **Bandini** (2003)

Discussion

Please [register](#) or [login](#) to write a comment

DJ Anubis said on Monday, 31 May 2010:
btw: I think the hardrockers/metalheads are still to discover this website... It's not a genre with samples (covers ok)... I think DrDosage might be one of the first real hard rock adders :)

DJ Anubis said on Monday, 31 May 2010:
^^ fixed

Drpepperfan said on Monday, 31 May 2010:
Actually it would probably be better to use this version <http://www.youtube.com/watch?v=bpG94t14D6Y> since Metallica actually, ya know, play on it :)

Drpepperfan said on Monday, 31 May 2010:
I was really surprised this wasn't here already, but at least it is now.

Download the cover version now from:

[amazon](#) [Download on iTunes](#)

Buy this track on CD / vinyl from:

[amazon](#) [ebay](#) [junorecords](#)

[SEND THIS TRACK'S RINGTONE TO YOUR PHONE](#)

Tags: [Add]
Main genre: Rock / Pop

Download the original song now from:

[amazon](#) [Download on iTunes](#)

Buy this track on CD / vinyl from:

[amazon](#) [ebay](#) [junorecords](#)

[SEND THIS TRACK'S RINGTONE TO YOUR PHONE](#)

Tags: Film Score, Spaghetti Western [Add]
Main genre: Soundtrack

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**

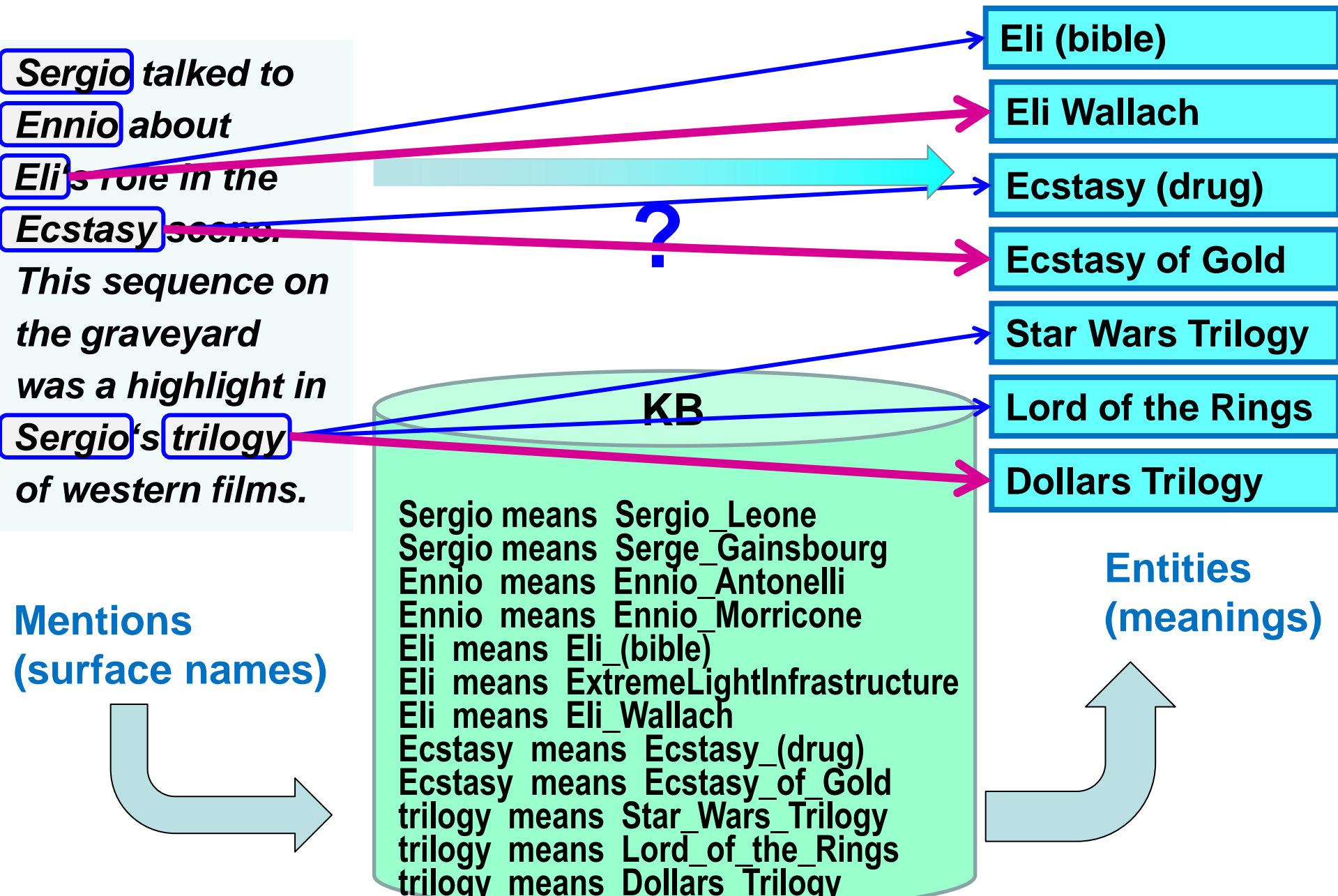
★ From **Text Analytics** to **Insight**

★ Wrap-Up

Names vs. Entities

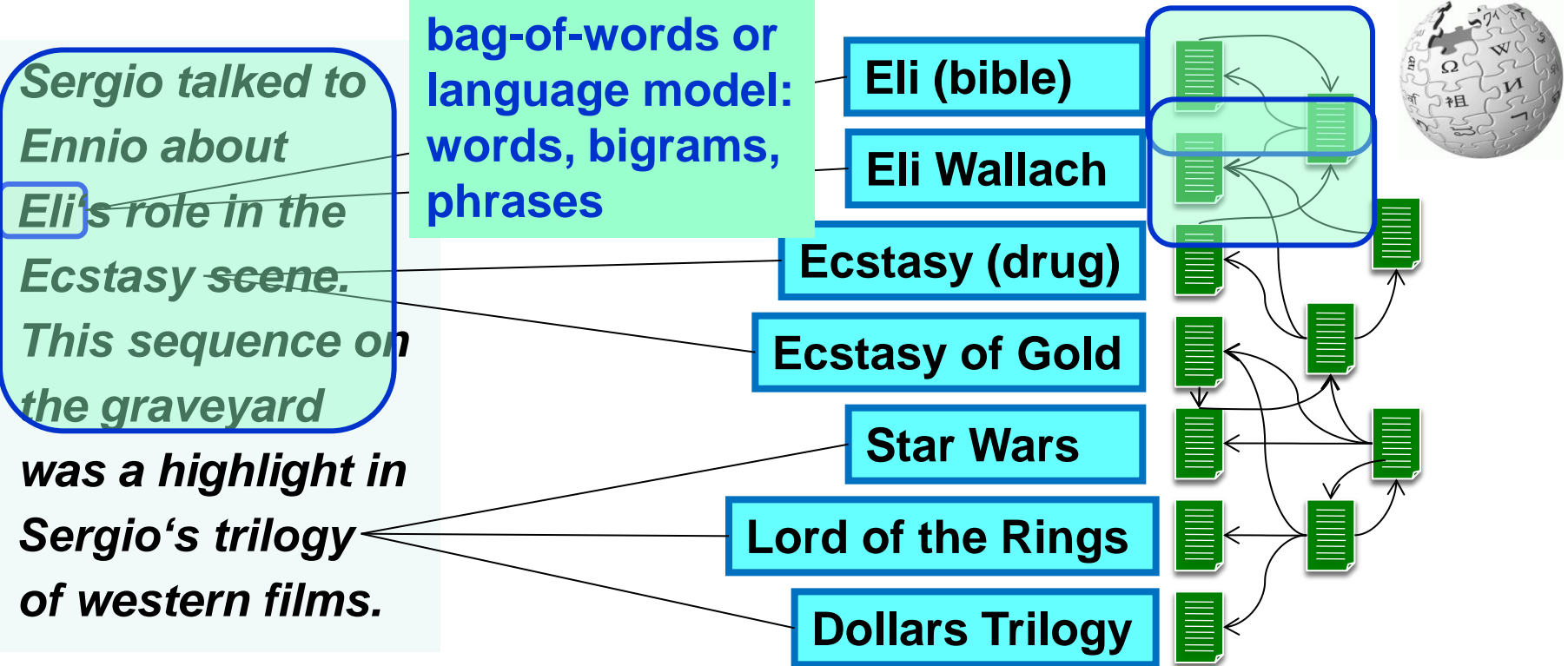
app id: 12369458	absolute magnitude: 15.4 [1]
apoapsis: 45196629 4.519662	apoapsis: 451966298674.4018 [1] 4.519662986744018E8 [1]
background: non_performing_personnel	background: non_performing_personnel [19]
birth: 1928-11-10	birth: 1928-11-10 [19]
birth place: Rome, Italy	birth place: Rome, Italy [19]
is composer of: http://dbpedia.org/resource/Ennio_Morricone The Life of Ennio Morricon Red Dear	discovered: 2005-08-27 [1]
is contributor of: la Cosa Bella 1970 [5] Bianco Rosso E Verdone (Piano Bar Secondo) - (with Ennio Morricon) [13]	discoverer: http://dbpedia.org/resource/Alain_Maury [1] http://dbpedia.org/resource/F._Mallia [1]
is disambiguates of: Ennio [19]	epoch: May 14, 2008 [1]
discovered: 2005-08-27	filename: Ennio_Morricon Ennio Morricon Mina-SeTelefonando3.ogg [19]
discoverer: http://dbpedia.org/resource/Alain_Maury http://dbpedia.org/resource/F._Mallia	former name: 2005 QP:
epoch: May 14, 2008	
filename: Ennio_Morricon Ennio Morricon Mina-SeTelefonando3.ogg [19]	

Named Entity Disambiguation

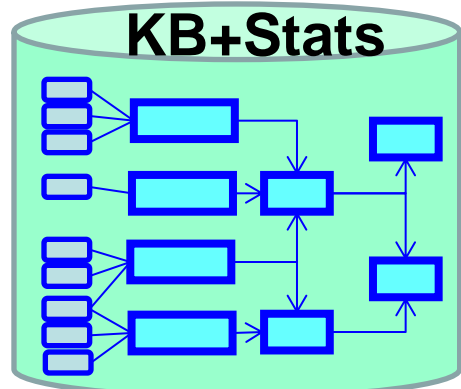


Mention-Entity Graph

weighted undirected graph with two types of nodes

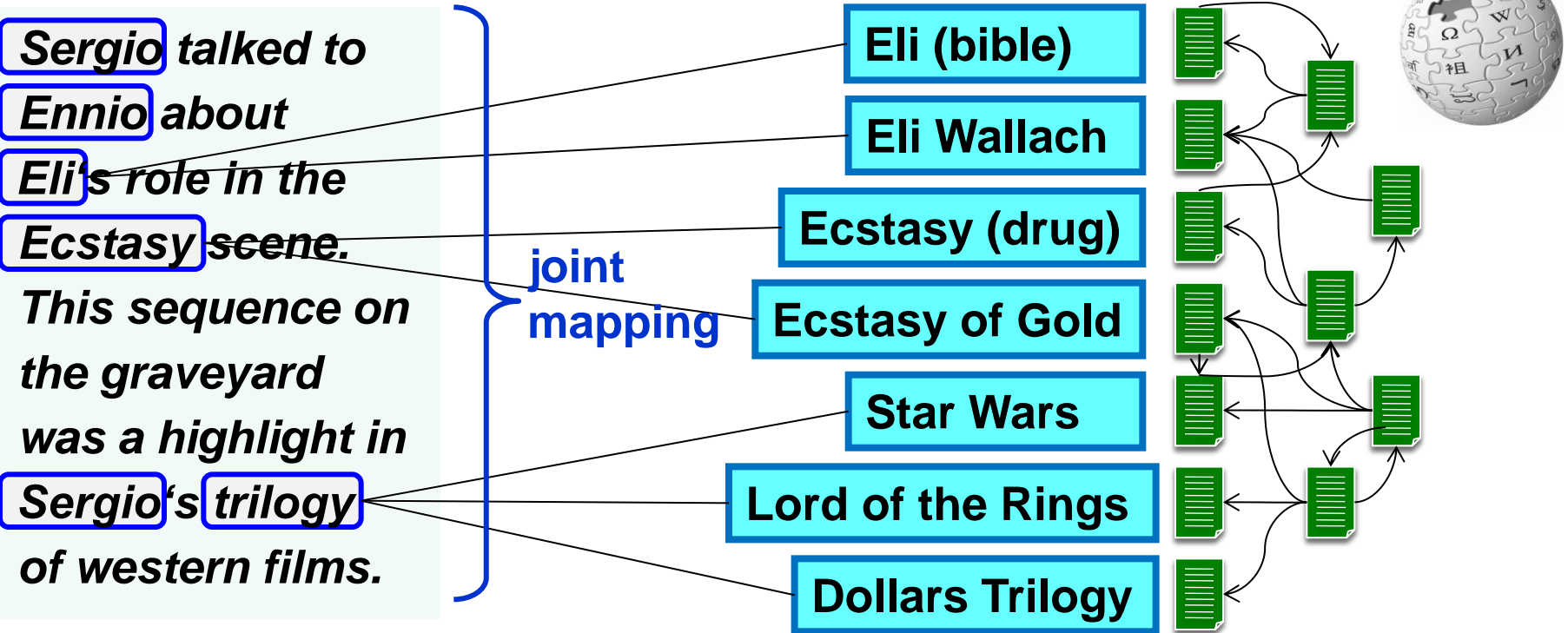


- Popularity (m,e):**
 - $\text{freq}(e|m)$
 - $\text{length}(e)$
 - $\#\text{links}(e)$
- Similarity (m,e):**
 - $\text{cos/Dice/KL}(\text{context}(m), \text{context}(e))$



Mention-Entity Graph

weighted undirected graph with two types of nodes

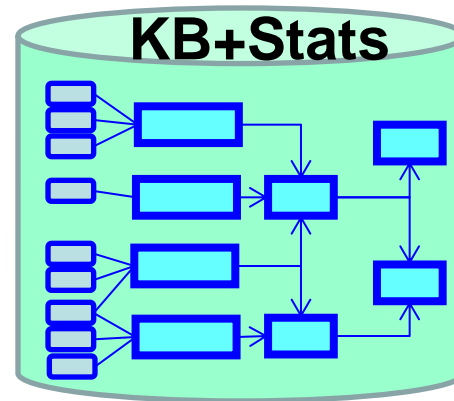


Popularity
(m,e):

- $\text{freq}(e|m)$
- $\text{length}(e)$
- $\#\text{links}(e)$

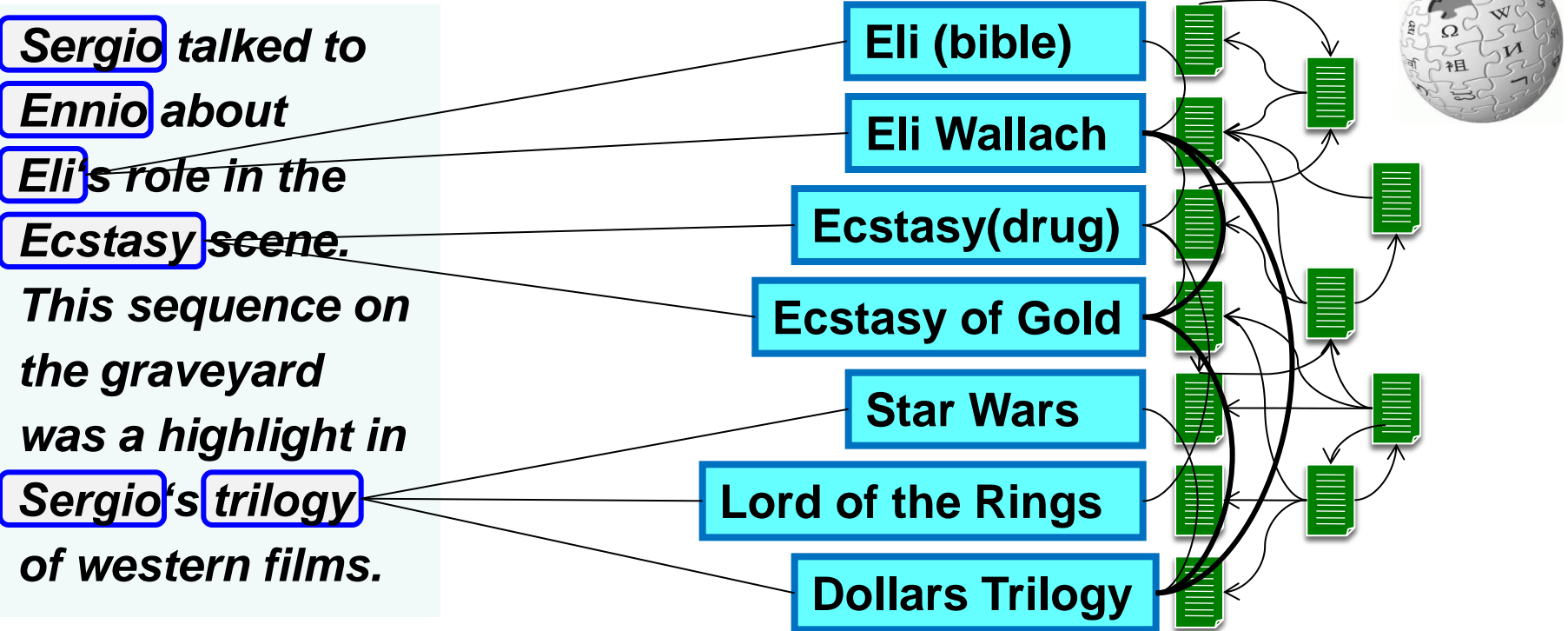
Similarity
(m,e):

- $\text{cos/Dice/KL}(\text{context}(m), \text{context}(e))$



Mention-Entity Graph

weighted undirected graph with two types of nodes

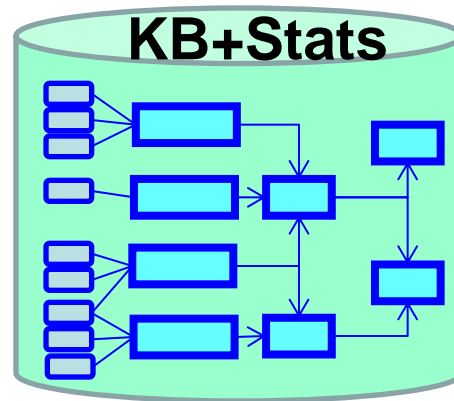


Popularity (m,e):

- $\text{freq}(m,e|m)$
- $\text{length}(e)$
- $\#\text{links}(e)$

Similarity (m,e):

- $\text{cos/Dice/KL}(\text{context}(m), \text{context}(e))$

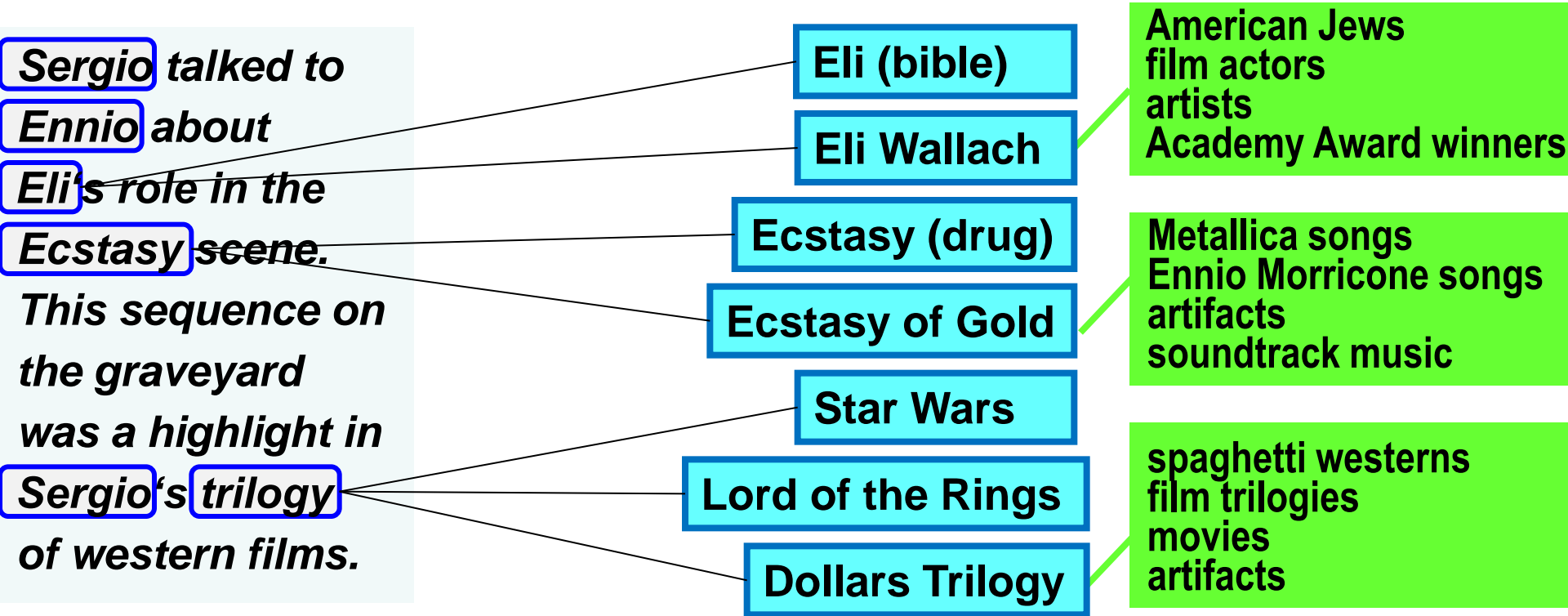


Coherence (e,e'):

- $\text{dist}(\text{types})$
- $\text{overlap}(\text{links})$
- $\text{overlap}(\text{keyphrases})$

Mention-Entity Graph

weighted undirected graph with two types of nodes

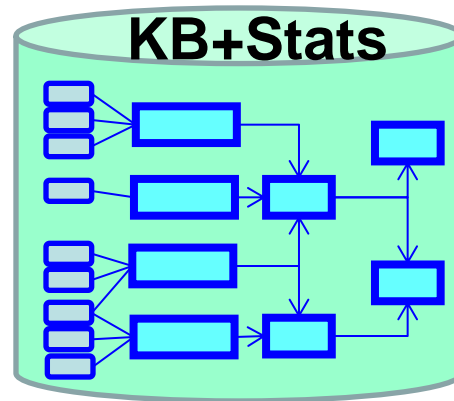


Popularity (m,e):

- $\text{freq}(m,e|m)$
- $\text{length}(e)$
- $\#\text{links}(e)$

Similarity (m,e):

- $\text{cos/Dice/KL}(\text{context}(m), \text{context}(e))$

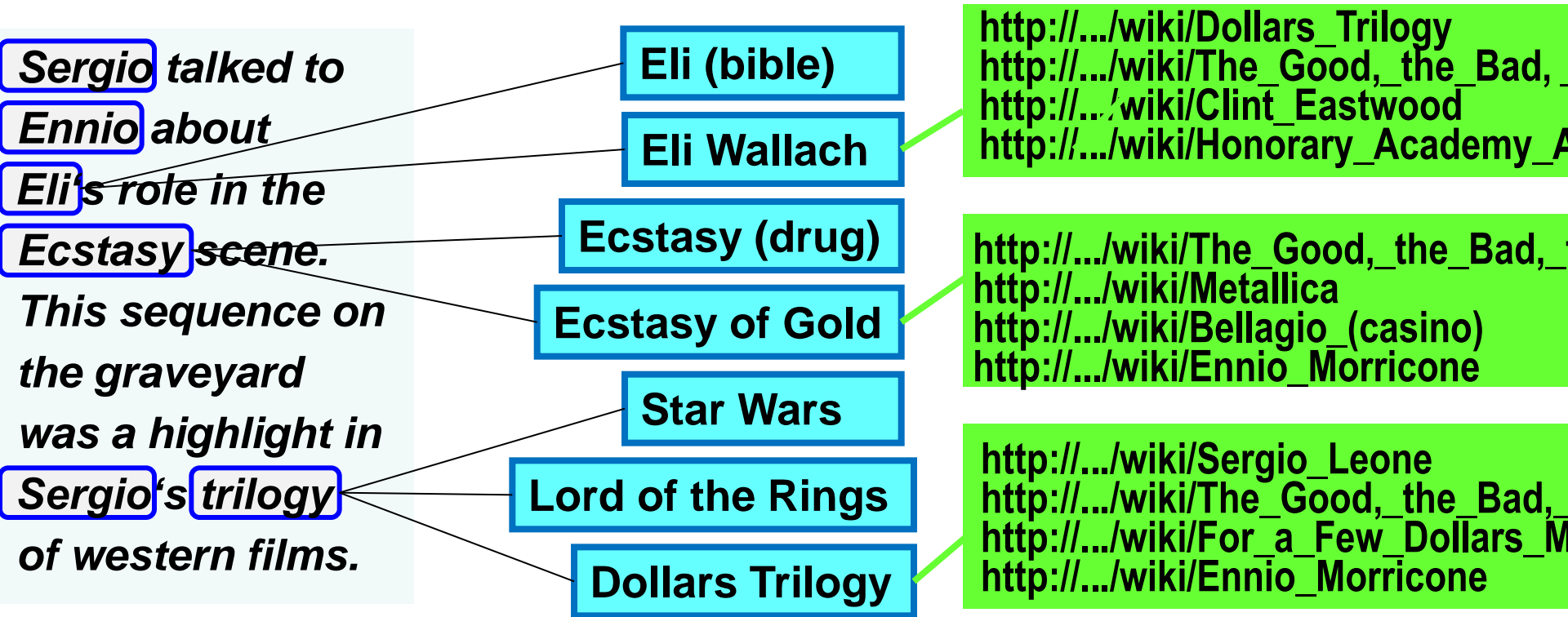


Coherence (e,e'):

- $\text{dist}(\text{types})$
- $\text{overlap}(\text{links})$
- $\text{overlap}(\text{keyphrases})$

Mention-Entity Graph

weighted undirected graph with two types of nodes

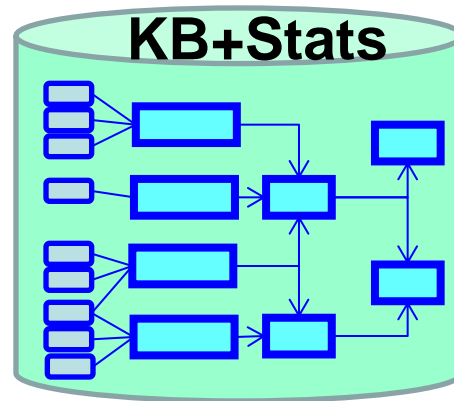


Popularity (m,e):

- $\text{freq}(m,e|m)$
- $\text{length}(e)$
- $\#\text{links}(e)$

Similarity (m,e):

- $\text{cos/Dice/KL}(\text{context}(m), \text{context}(e))$

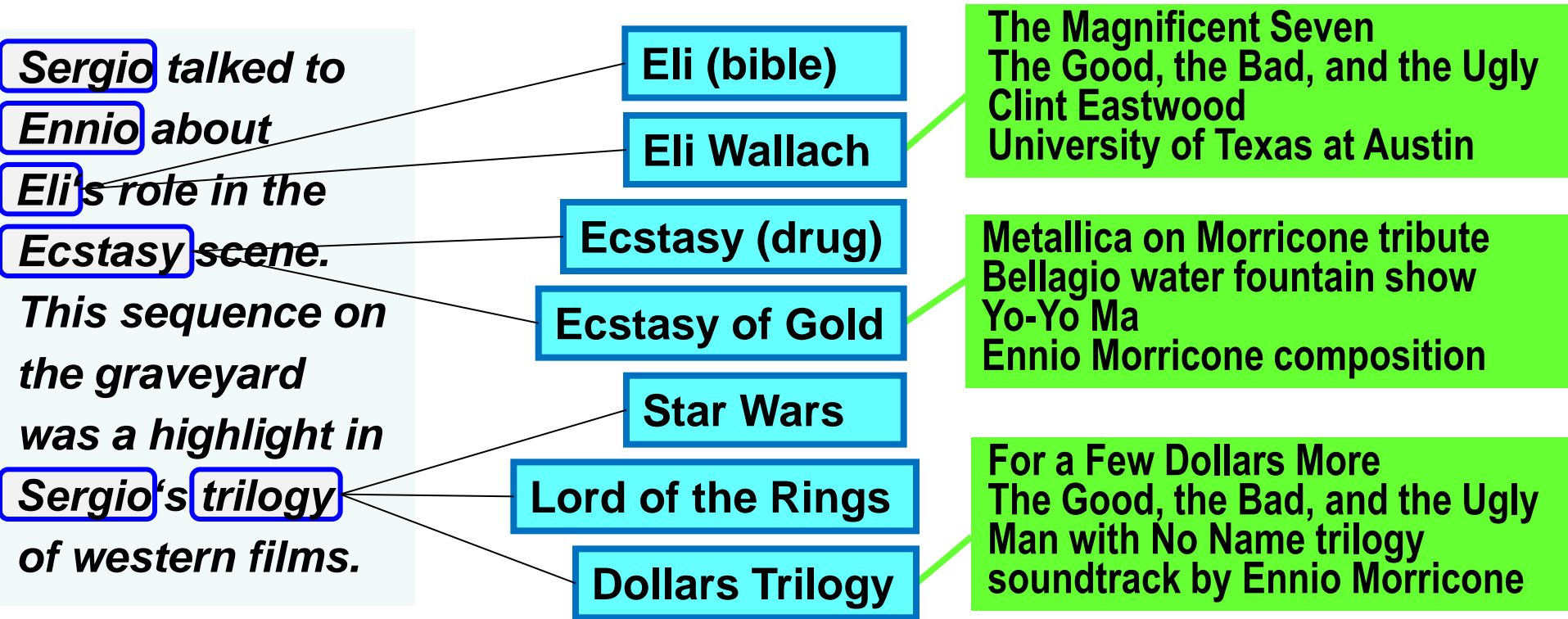


Coherence (e,e'):

- $\text{dist}(\text{types})$
- $\text{overlap}(\text{links})$
- $\text{overlap}(\text{keyphrases})$

Mention-Entity Graph

weighted undirected graph with two types of nodes

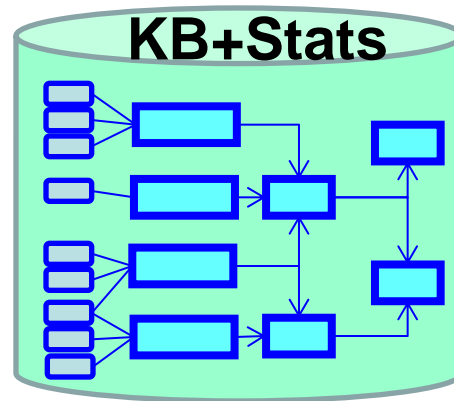


Popularity (m,e):

- $\text{freq}(m,e|m)$
- $\text{length}(e)$
- $\#\text{links}(e)$

Similarity (m,e):

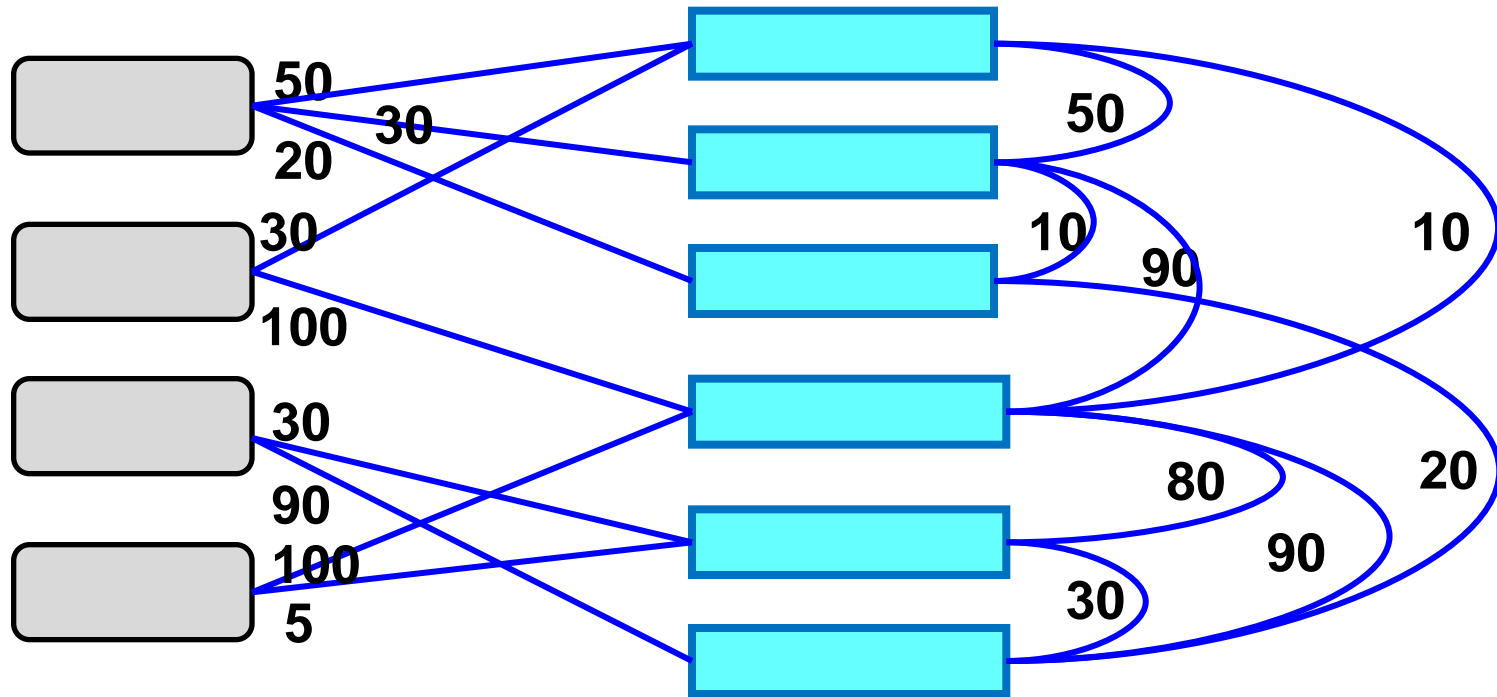
- $\text{cos/Dice/KL}(\text{context}(m), \text{context}(e))$



Coherence (e,e'):

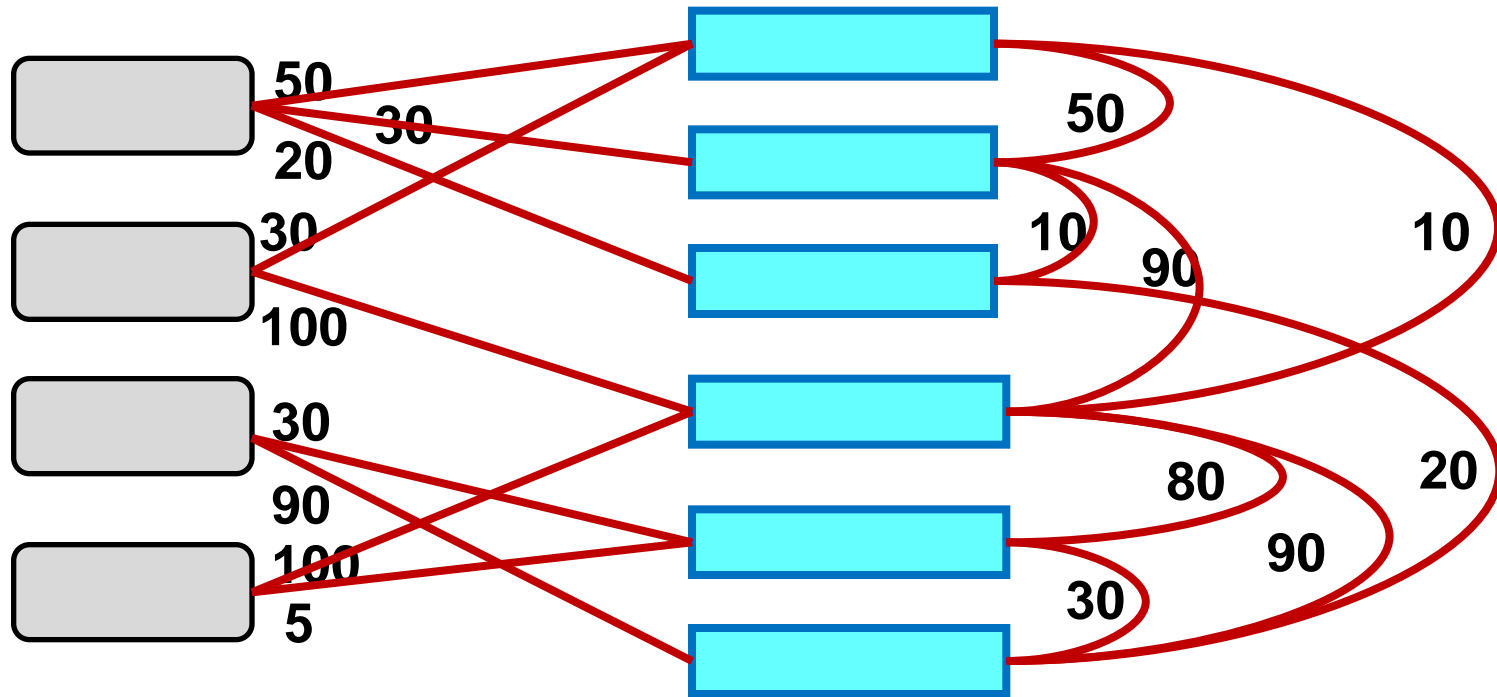
- $\text{dist}(\text{types})$
- $\text{overlap}(\text{links})$
- $\text{overlap}(\text{keyphrases})$

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

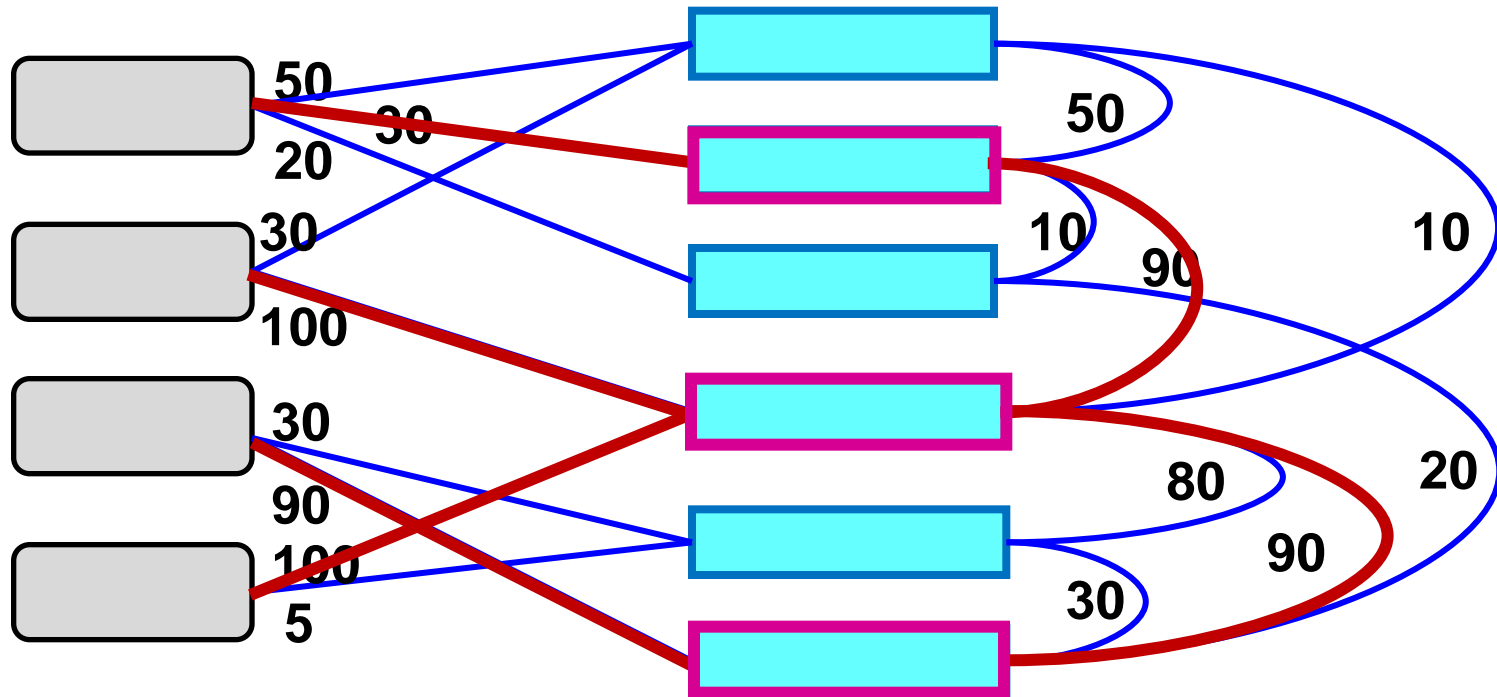


Collective Learning with Probabilistic Factor Graphs

[Chakrabarti et al.: KDD'09]:

- model $P[m|e]$ by similarity and $P[e_1|e_2]$ by coherence
- consider likelihood of $P[m_1 \dots m_k | e_1 \dots e_k]$
- factorize by all m - e pairs and e_1 - e_2 pairs
- use MCMC, hill-climbing, LP etc. for solution

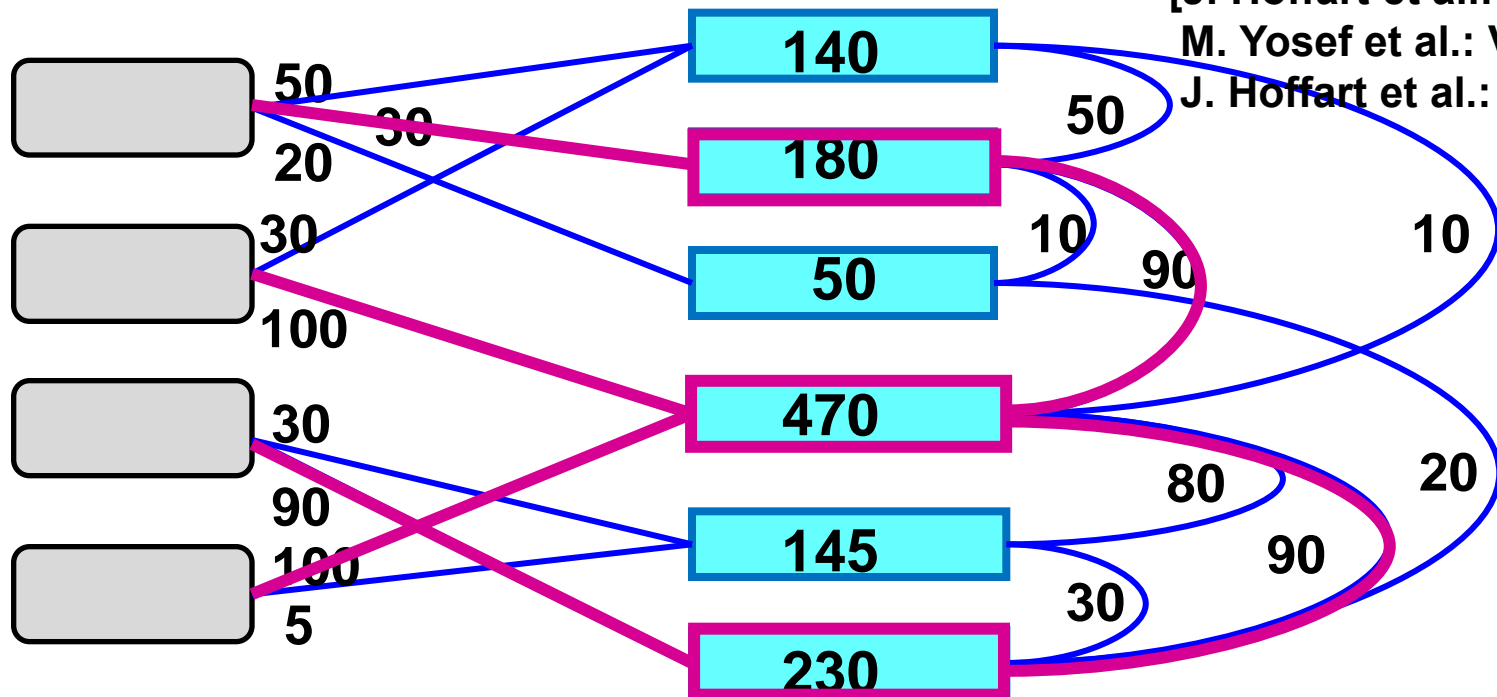
Joint Mapping: Dense Subgraph



- 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

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

Extent of partial matches

Weight of matched words

The **Ecstasy** piece was covered by **Metallica on the Morricone tribute** album.

Compute **overall similarity** of context(m) and candidate e

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

AIDA: Accurate Online Disambiguation

The screenshot displays the AIDA web interface in a Mozilla Firefox browser. The interface is divided into several sections:

- Disambiguation Method:** Includes radio buttons for 'prior', 'prior+sim', and 'prior+sim+coherence'. The 'prior+sim+coherence' method is selected. Below this are sliders for 'Parameters: (default should be OK)', 'Ambiguity degree 5', and 'Coherence robustness test threshold: 0.9'.
- Entities Type Filters:** A text input field labeled 'Enter the types here'.
- Mention Extraction:** Features buttons for 'Stanford NER' and 'Manual'. A text area contains the sentence: "Sergio talked to Ennio about Eli's role in the [[Ecstasy]] scene. This sequence on the graveyard was part of Sergio's western [[trilogy]]. Ennio's composition was later covered by Ma." The words 'Sergio Leone', 'Ennio', 'The Ecstasy of...', 'trilogy', and 'Ma' are highlighted with colored boxes (blue, green, and red).
- Input Type:** A dropdown menu is set to 'TEXT'. Below it are buttons for 'Types list' and 'Types', and a 'Focused Types tag' field.
- Candidate Entity List:** A table with columns 'Candidate Entity', 'ME Similarity', and 'Weighted Degree'. The table lists various entities such as 'Dollars_Triology', 'Star_Wars', 'The_Lord_of_the_Rings', etc., with their respective similarity and weighted degree values.

AIDA: Very Difficult Example

Disambiguation Method:

prior prior+sim **prior+sim+coherence**

Parameters: (defaults should be OK)

Prior-Similarity-Coherence balancing ratio:

prior VS. sim. balance = **0.4**

(prior+sim.) VS. coh. balance **0.8**



Ambiguity degree **5**



Coherence robustness test threshold:

0.9



Entities Type Filters:

Enter the types here

Mention Extraction:

Stanford NER **Manual**

You can manually tag the mentions by putting them between [[and]].
HTML Tables are automatically disambiguated in the manual mode.

[[Page]] played Kashmir on a Gibson.

Input Type:TEXT Overall runtime:3s, 832ms

Types list

Types tag cloud

Focused Types tag cloud

[Jimmy Page] **Page** played
[Kashmir (song)] **Kashmir** on a
[Gibson Guitar Corporation] **Gibson**.

25: Gibson

Candidate Entity

ME Similarity

Mel_Gibson	0.0
Henry_Gibson	0.0
Gibson_Guitar_Corporation	6.937260822770075E-5
Robert_Gibson_\u0028pitcher\u0029	4.3397387840473426E-5
Kirk_Gibson	0.0
Debbie_Gibson	0.0
William_Gibson	0.0
Tyrese_Gibson	0.0
Aaron_Gibson	0.0
Paul_Gibson	0.0
Don_Gibson	0.0
Case_Gibson	0.0

AIDA: Web Tables

Disambiguation Method:

prior prior+sim **prior+sim+coherence**

Parameters: (default should be OK)

Prior-Similarity-Coherence balancing ratio:
prior VS. sim. balance = 0.4
(prior+sim.) VS. coh. balance 0.6

Ambiguity degree 5

Coherence robustness test threshold:

Mention Extraction:

Stanford NER **Manual**

You can manually tag the mentions by putting them in boxes. They are automatically disambiguated in the manual mode.

Steve Mac
Dennis C
Richard GNU

Input Type:TABLE Overall runtime:2m, 34s, 101ms

Types list Types tag cloud
Focused Types tag cloud

[[Steve Jobs](#)] **Steve**
[[Apple Inc.](#)] **Mac**
[[Dennis Ritchie](#)] **Dennis**
[[C](#)] **C**
[[Richard Stallman](#)] **Richard**
[[GNU Core Utilities](#)] **GNU**

Graph **Removal Steps**

Candidate Entity	ME Similarity	W
	0.06842879372431546	1.5
	0.012022359121799974	1.1
	0.04473148249975622	1.0
sociologist\u0029	0.0	1.0
	0.02373582454298693	1.0
	0.03680277789844543	0.8
	0.0	0.8
	0.0	0.7
	0.08034068526592103	0.6
	0.03216497982891819	0.6
	0.03747041730116862	0.6
	0.022550325631343984	0.6
	0.11896368017112827	0.5
	0.032165818910204466	0.5
	0.02199673334363371	0.5
3American_football\u0029	0.005849708548223075	0.5
	0.022177669833143673	0.5
	0.0	0.5
actor\u0029	0.0	0.4
	0.0	0.4
	0.0	0.4
	0.02852493248362575	0.4
028artist\u0029	0.0	0.4
	0.0469630805585354	0.4
	0.0	0.4

<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

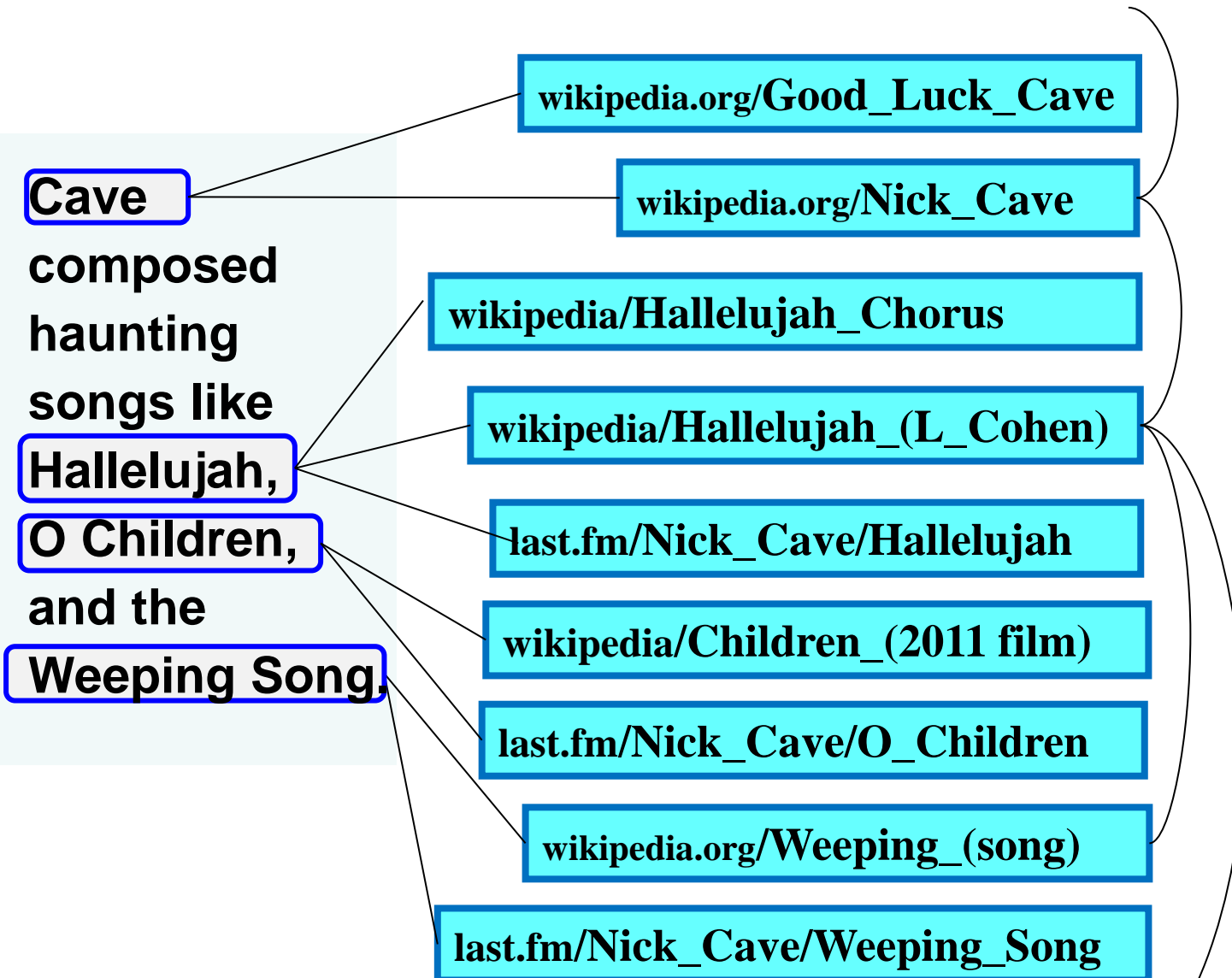
some use Stanford NER tagger for detecting mentions

<http://nlp.stanford.edu/software/CRF-NER.shtml>

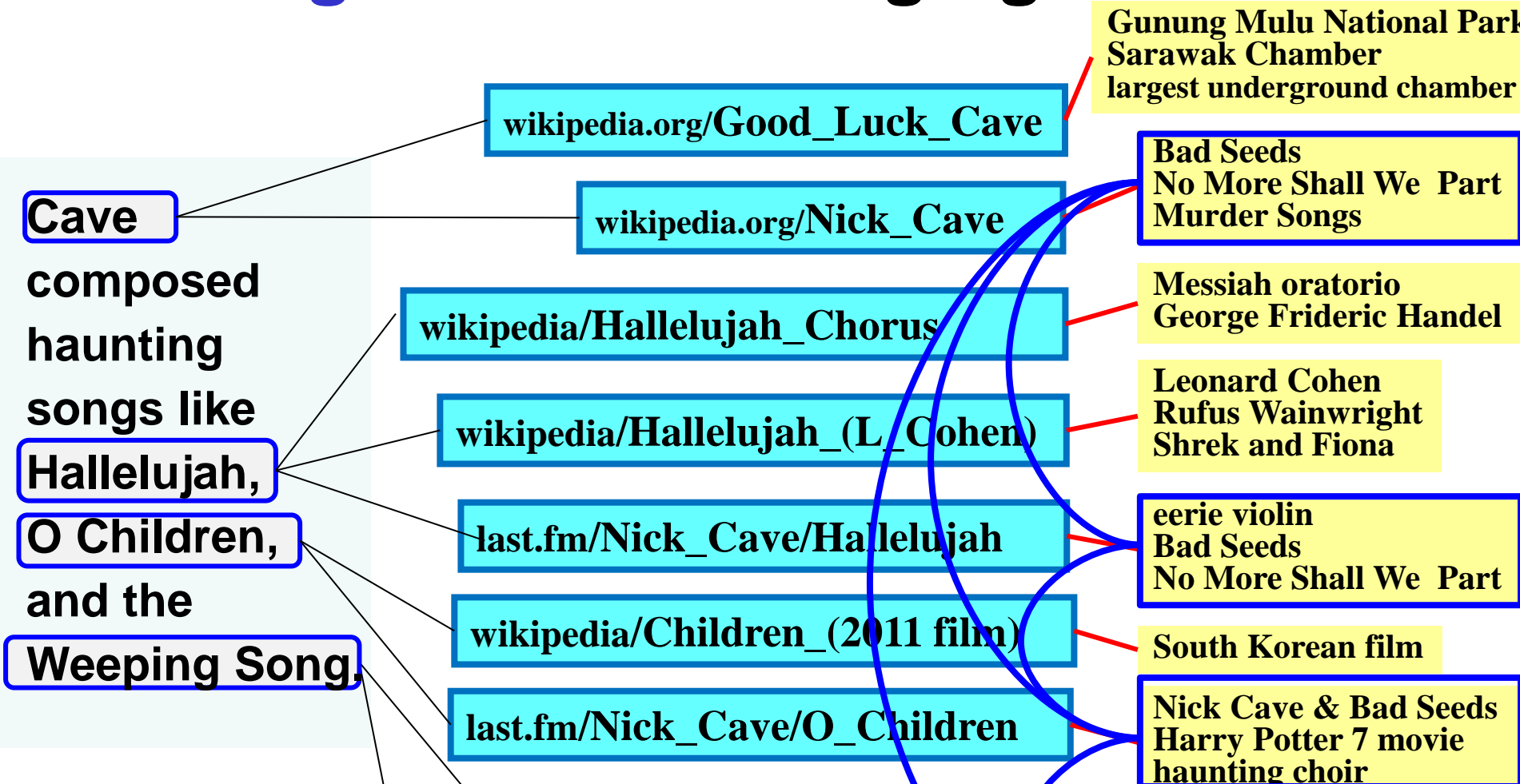
Ongoing Research & Remaining Challenges

- 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



Long-Tail and Emerging Entities



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

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

implementation uses min-hash and LSH

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

duet

Long-Tail and Emerging Entities

Cave's

brand-new
album
contains
masterpieces
like

Water's Edge

and

Mermaids.

wikipedia.org/Good_Luck_Cave

wikipedia.org/Nick_Cave

.../Water's Edge Restaurant

.../Water's Edge (2003 film)

any OTHER „Water's Edge“

.../Mermaid's Song

.../The Little Mermaid

any OTHER „Mermaids“

Gunung Mulu National Park
Sarawak Chamber
largest underground chamber

Bad Seeds
No More Shall We Part
Murder Songs

excellent seafood
clam chowder
Maine lobster

Nathan Fillion
horrible acting

all phrases minus
keyphrases of known
candidate entities

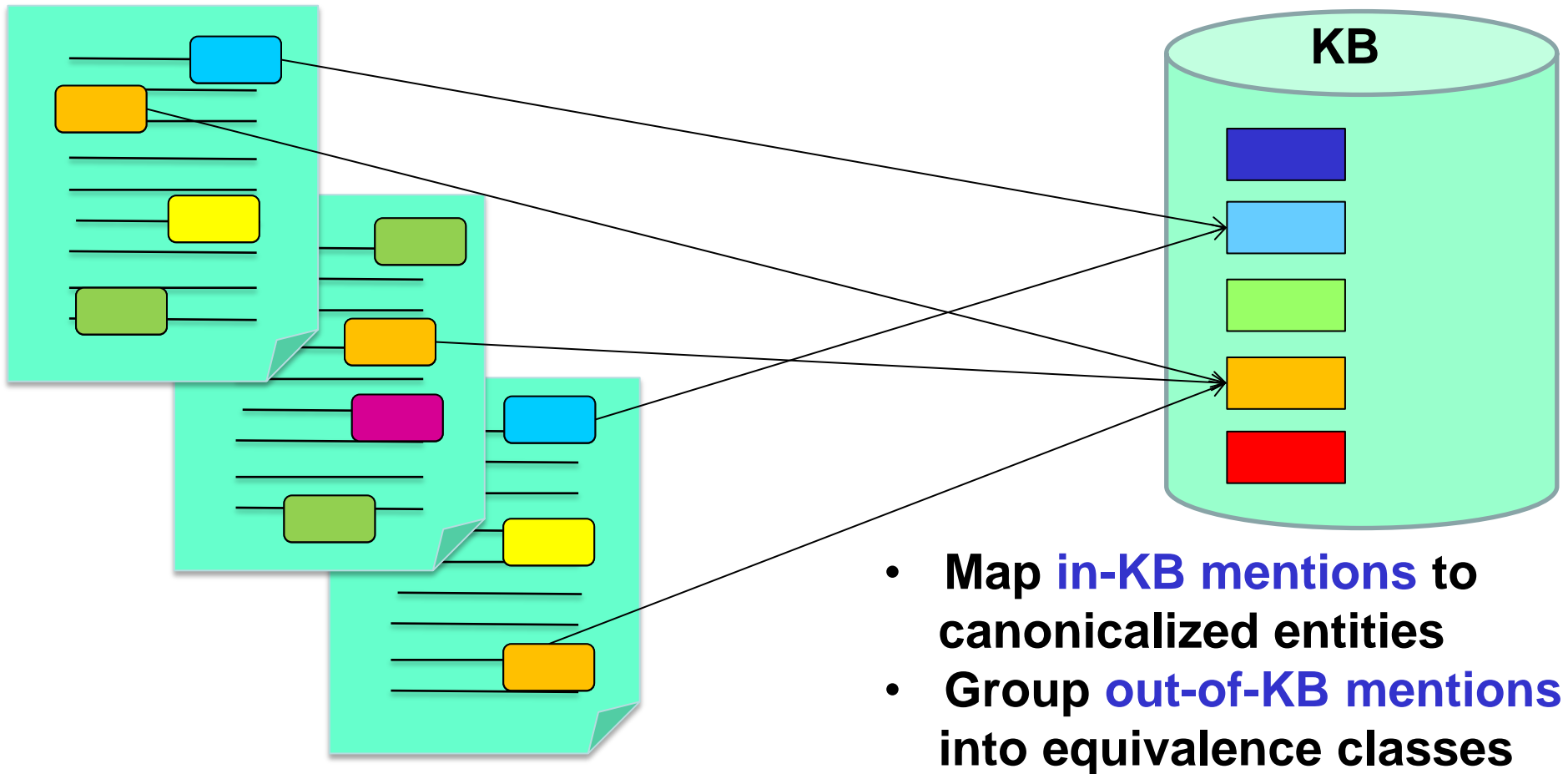
Pirates of the Caribbean 4
My Jolly Sailor Bold
Johnny Depp

Walt Disney
Hans Christian Andersen
Kiss the Girl

all phrases minus
keyphrases of known
candidate entities

Towards Integrated NERD and CCR

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



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 soulful 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” \Leftrightarrow *“obtained degree in * from”*

“and PRONOUN ADJECTIVE advisor” \Leftrightarrow *“under the supervision of”*

One relational phrase can **subsume** another

“wife of” \Rightarrow *“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]

Thesaurus Relations Taxonomy

▼ DBpedia Relations

academicAdvisor
affiliation
album
almaMater
anthem
appointer
architect
artist
assembly
associate
associatedBand
associatedMusicalArtist
author
automobilePlatform
award
bandMember
basedOn
battle
beatifiedBy
beatifiedPlace
billed
binomialAuthority
birthPlace
board
bodyDiscovered
bodyStyle
borough
broadcastArea
broadcastNetwork

Relation: dbpedia:bandMember

1-31 of 31

Pattern

is formed by;
lead singer;
has announced that;
is composed;
currently consists;
which founded;
vocalist [[con]] guitarist;
was formed by vocalist;
[[det]] liveaction version as;
led by;
bassist [[con]];
bandmates [[con]];
[[adj]] consisting of;
performing as [[det]] quintet;
launched with [[adj]] members;
[[det]] line up consisting of;

lead singer;

☐ Synset

lead singer;
s lead singer;
[[adj]] lead singer;

Paramore , Hayley Williams + 📄

All (band) , Dave Smalley + 📄

Alabama (band) , Randy Owen + 📄

Clutch (band) , Neil Fallon + 📄

Nirvana (band) , Kurt Cobain ☐ 📄

In particular , Rossdale 's forced random , stream of consciousness dismissed by some as an imitatio singer , Kurt Cobain .

Los Bravos , Mike Kogel + 📄

Twisted Sister , Dee Snider + 📄

350 000 SOL patterns with 4 Mio. instances

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

Big Data Algorithms at Work

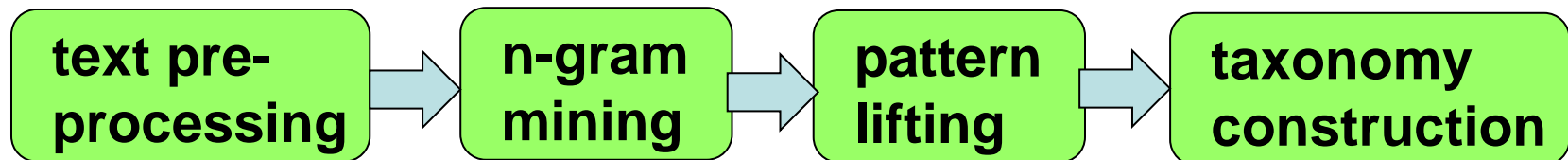
Frequent sequence mining

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



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 (t_1, p, t_2) for known entities:

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

For each new e and all candidate types t_i :

$$\max \alpha \sum_i \text{score}(e, t_i) X_i + \beta \sum_{ij} \text{corr}(t_i, t_j) Y_{ij}$$

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

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.

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

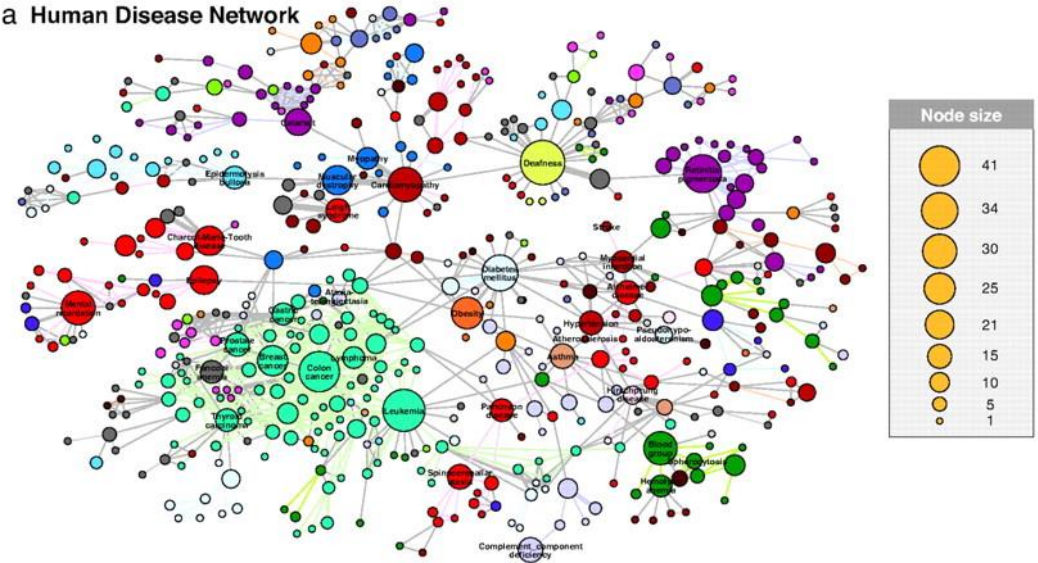
General Design Pattern:

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

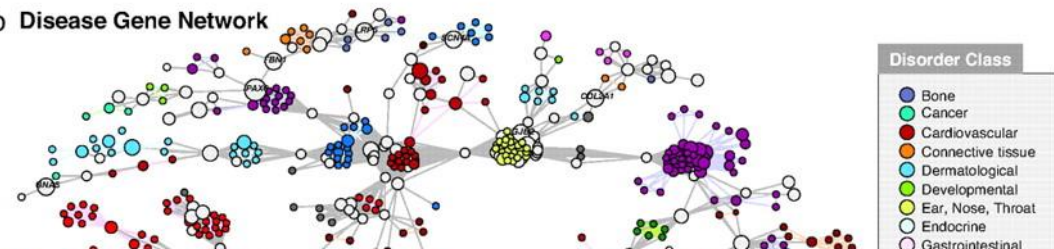
Big Data Analytics for Disease Networks



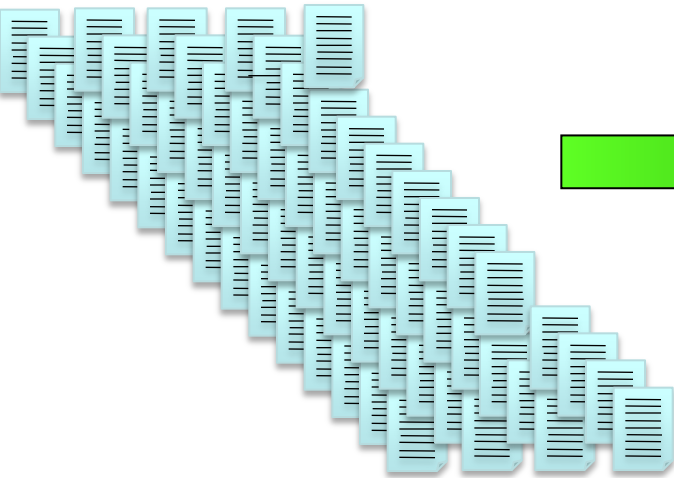
a Human Disease Network



b Disease Gene Network



PubMed



But try this with:
diabetes mellitus, diabetes 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, ...



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, ...

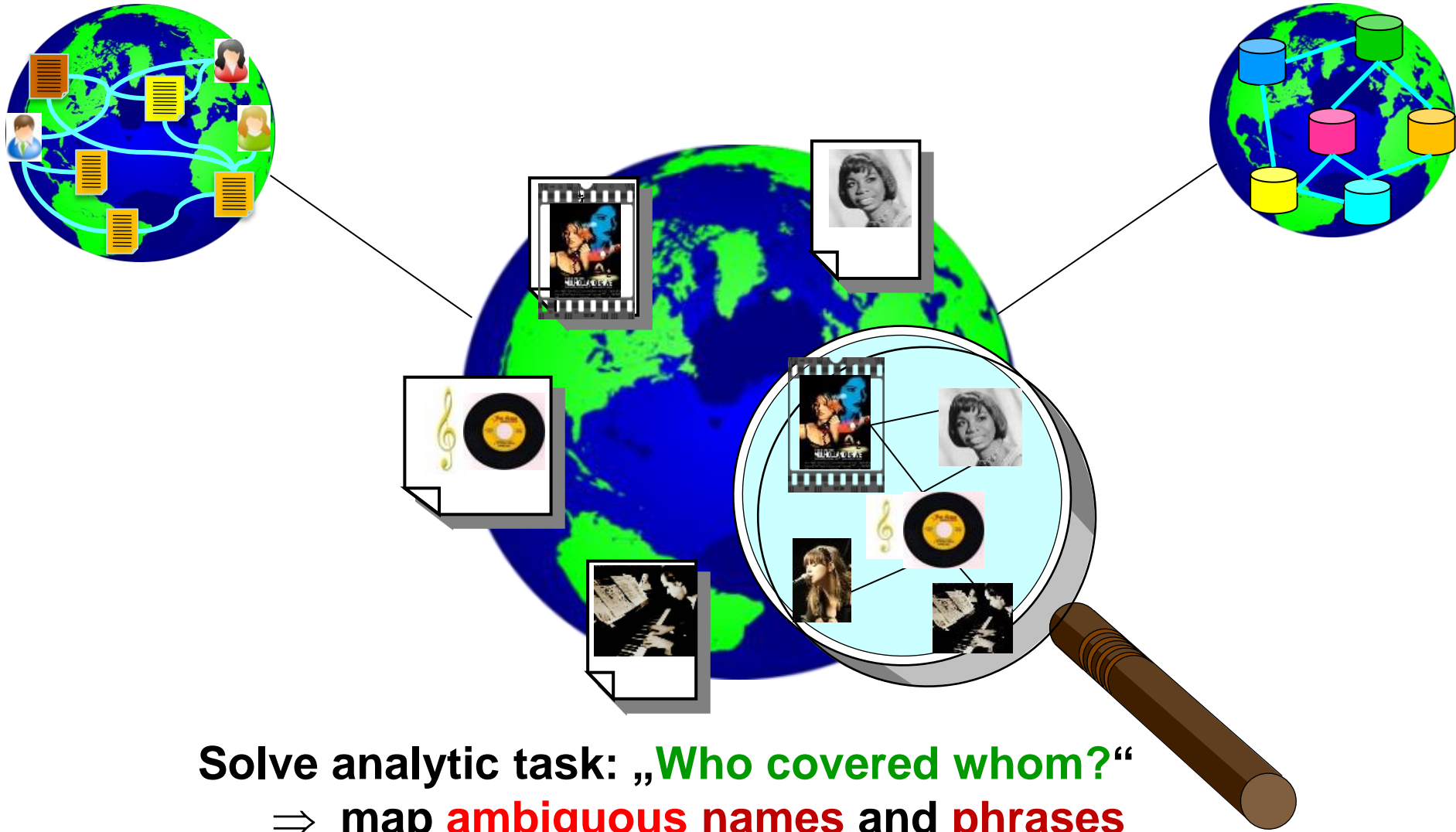
Outline

- ✓ **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, ...