

KDD Tutorial T39

# Building a Large-scale, Accurate and Fresh Knowledge Graph

**Yuqing Gao, Jisheng Liang, Benjamin Han, Mohamed Yakout, Ahmed Mohamed**  
**Satori Group, AI+R, Microsoft**



## Disclaimers

© 2018 Microsoft Corporation. All rights reserved. This document is provided "as-is." Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it.

Some examples are for illustration only and are fictitious. No real association is intended or inferred. This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes.

# Outline

- Logistics (5 min)
- Part I: Introduction (30 min)
- Part II: Acquiring Knowledge in the Wild (55 min)
- Break (2:30 – 3:00pm, 30 min)
- Part III: Building Knowledge Graph (70 min)
- Break (20 min)
- Part IV: Serving Knowledge to the World (30 min)

# Part I: Introduction

**Yuqing Gao**

Partner General Engineering Manager, Satori Group, Microsoft AI+R

[yuga@microsoft.com](mailto:yuga@microsoft.com)



# What is Knowledge

- Plato's definition: Justified true belief

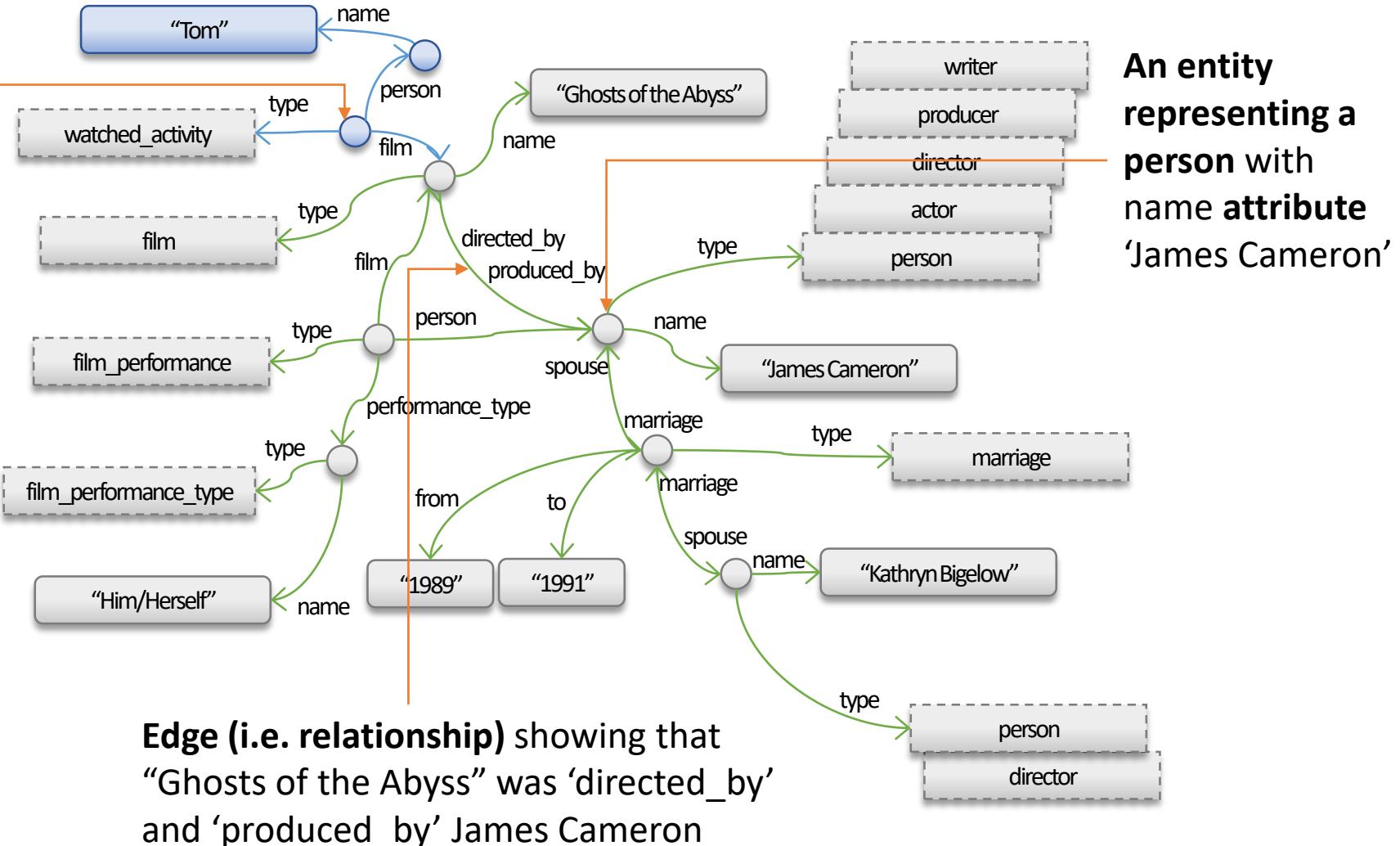
## Mission:

- Build the best Knowledge Graph in industry that will provide the highest quality of world's knowledge and personal knowledge measured by **correctness, coverage, freshness & usage**, to enable **Agile, intelligent knowledge experiences**

# What is a Knowledge Graph?

Knowledge represented as entities, edges and attributes

**Personal entity** showing  
that Tom watched Ghosts  
of the Abyss



Key concepts

<b>Entity</b>	Represent something in the real world
<b>Edge</b>	Represent relationship
<b>Attribute</b>	Represent something about an entity
<b>Ontology</b>	Definition of possible types of entities, relationships and attributes

# State of the art knowledge graphs

Minimum set of characteristics of knowledge graphs:

1. mainly describes real world entities and their interrelations, organized in a graph.
2. defines possible classes and relations of entities in a schema.
3. allows for potentially interrelating arbitrary entities with each other.
4. covers various topical domains.

State of Art KGs:

- Cyc and Open Cyc
- Freebase
- Wikidata
- DBpedia
- YAGO
- NELL
- Google Knowledge Vault
- **Google KG**
- **Microsoft Satori KG**

Large vertical KGs

- Facebook (social network)
- LinkedIn (people graph)
- Amazon (product graph)



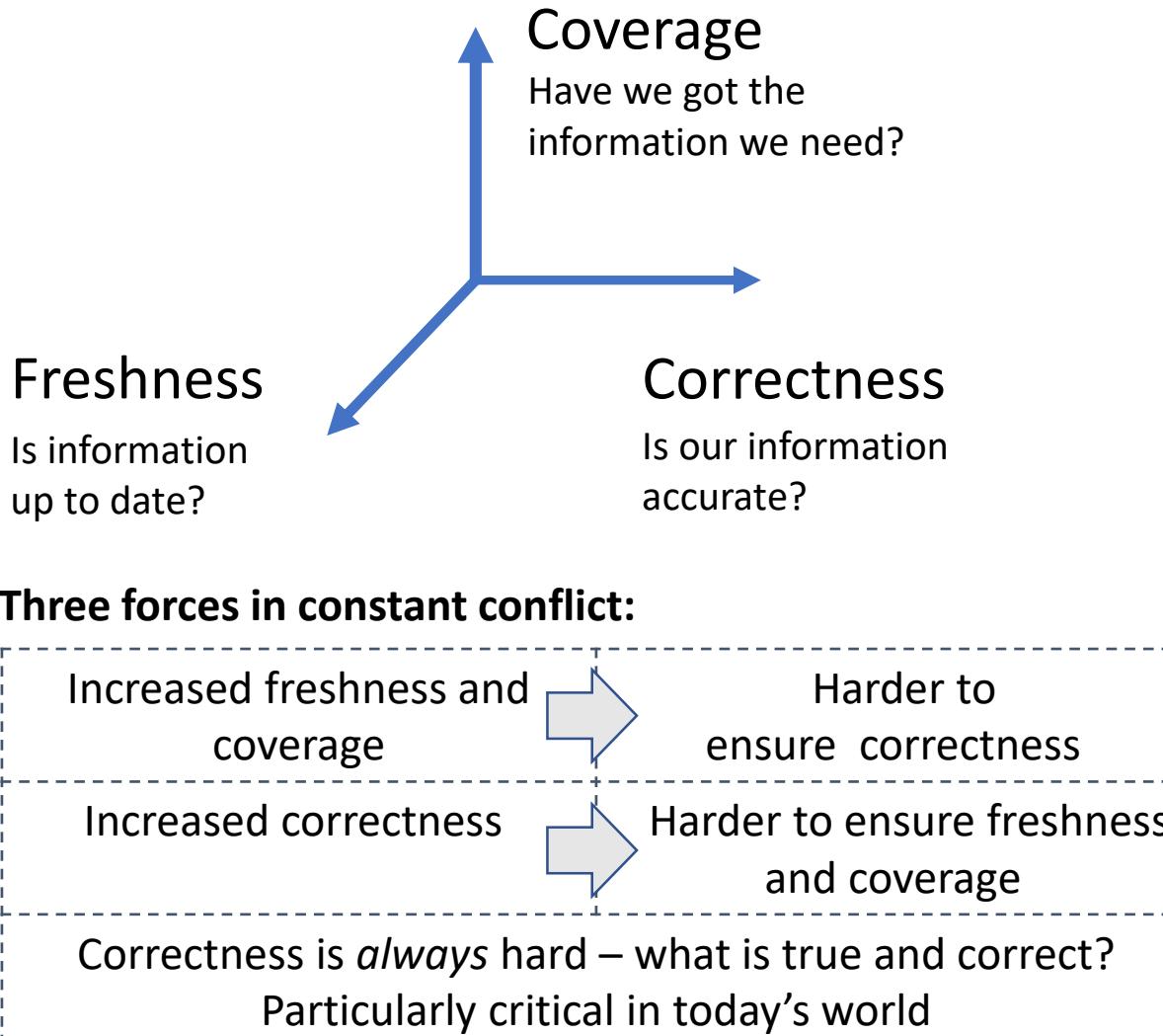
Large production KGs support Google and Bing Search

# Research fields

- Research related to knowledge graph refinement:
  - Ontology learning mainly deals with learning a concept level description of a domain, such as a hierarchy (e.g., Cities are Places)
- Approaches for Completion of Knowledge Graphs
  - Methods for Completing Type Assertions
  - Methods for Predicting Relations
- Approaches for Error Detection in Knowledge Graphs
  - Methods for Finding Erroneous Type Assertions
  - Methods for Finding Erroneous Relations
  - Methods for Finding Erroneous Literal Values
- Knowledge extraction
  - Entity linking and disambiguation
  - Fact extraction and verification

# Challenges of scaled KGs

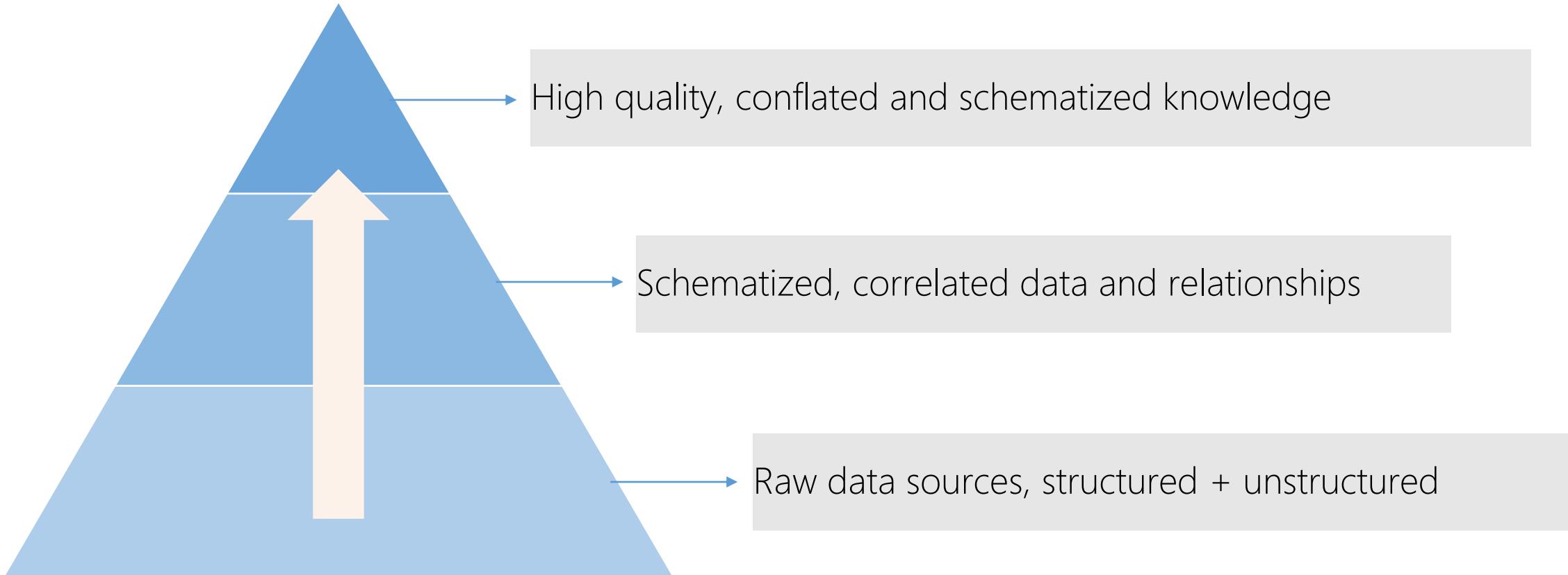
Building a small KG is easy - building a vast system like Satori is a huge challenge



**Will Smith:** Single entity, 108K facts assembled from 41 web sites.  
There are 200 Will Smiths on Wikipedia alone.

# Creating High Quality Web-scale Knowledge

AI: ML + NL + Conflation + Inference



# Knowledge flywheel in action: World graph

Search queries, views, click throughs, ...



Web pages, Web documents, Images, ...

## World graph

- People
- Places
- Things
- Actions
- ... ....

2B+ entities  
130B+ Web pages

# Knowledge flywheel in action: Domain-specific graph

Knowledge acquisition, search, recommendation ...



Authors, institutions, articles, conferences ...

## Domain-specific graph

- People
- Publications
- Fields of Study
- Venues

1B+ Scholarly articles

48K+ Journals

211M+ Authors

# Knowledge flywheel in action: Work graph

Messages read/sent, Document author/shared, ...



Emails, Messages, Documents, Meetings, ...

## Work graph

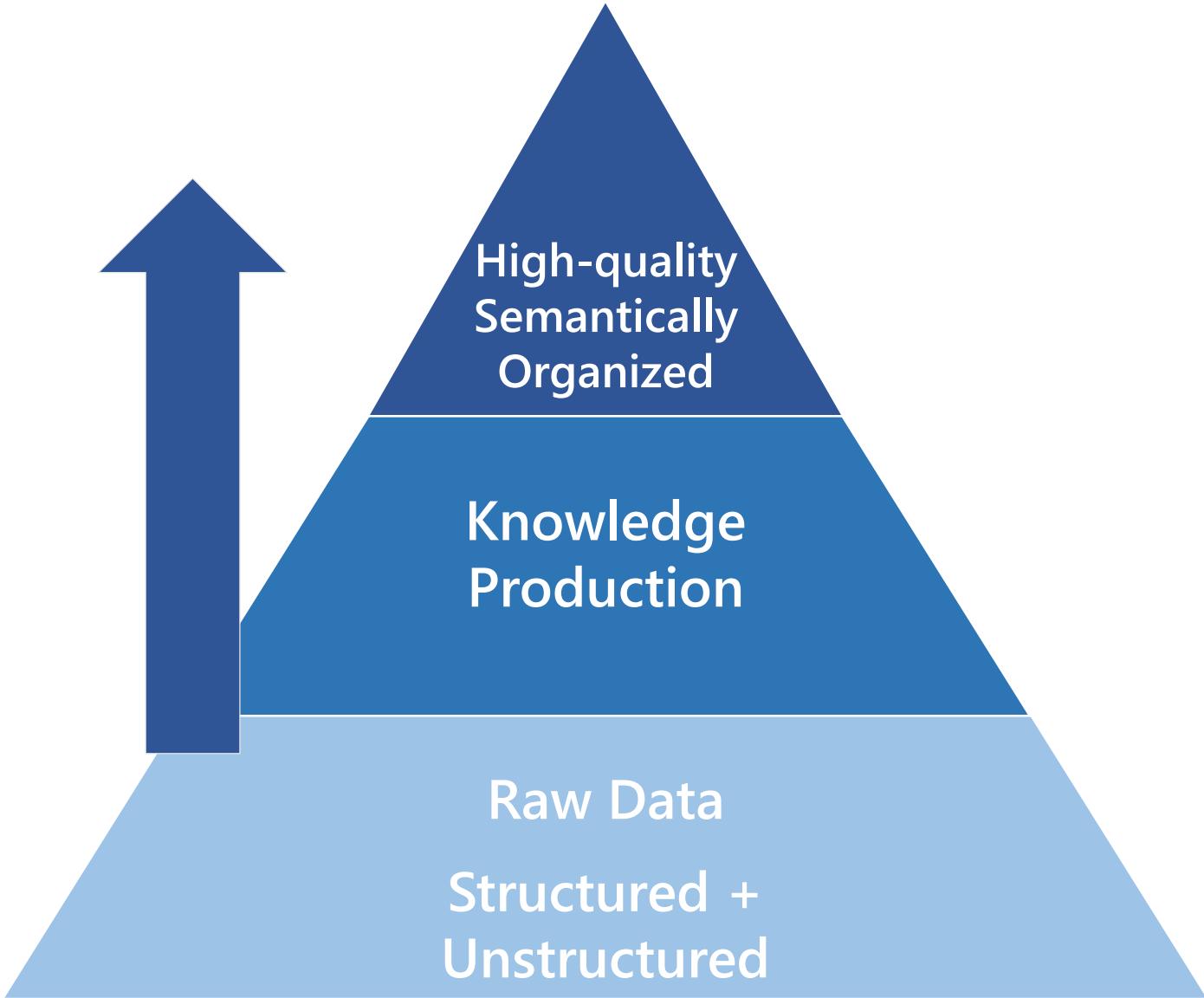
- People
- Groups
- Messages
- Activities

8T+ entities

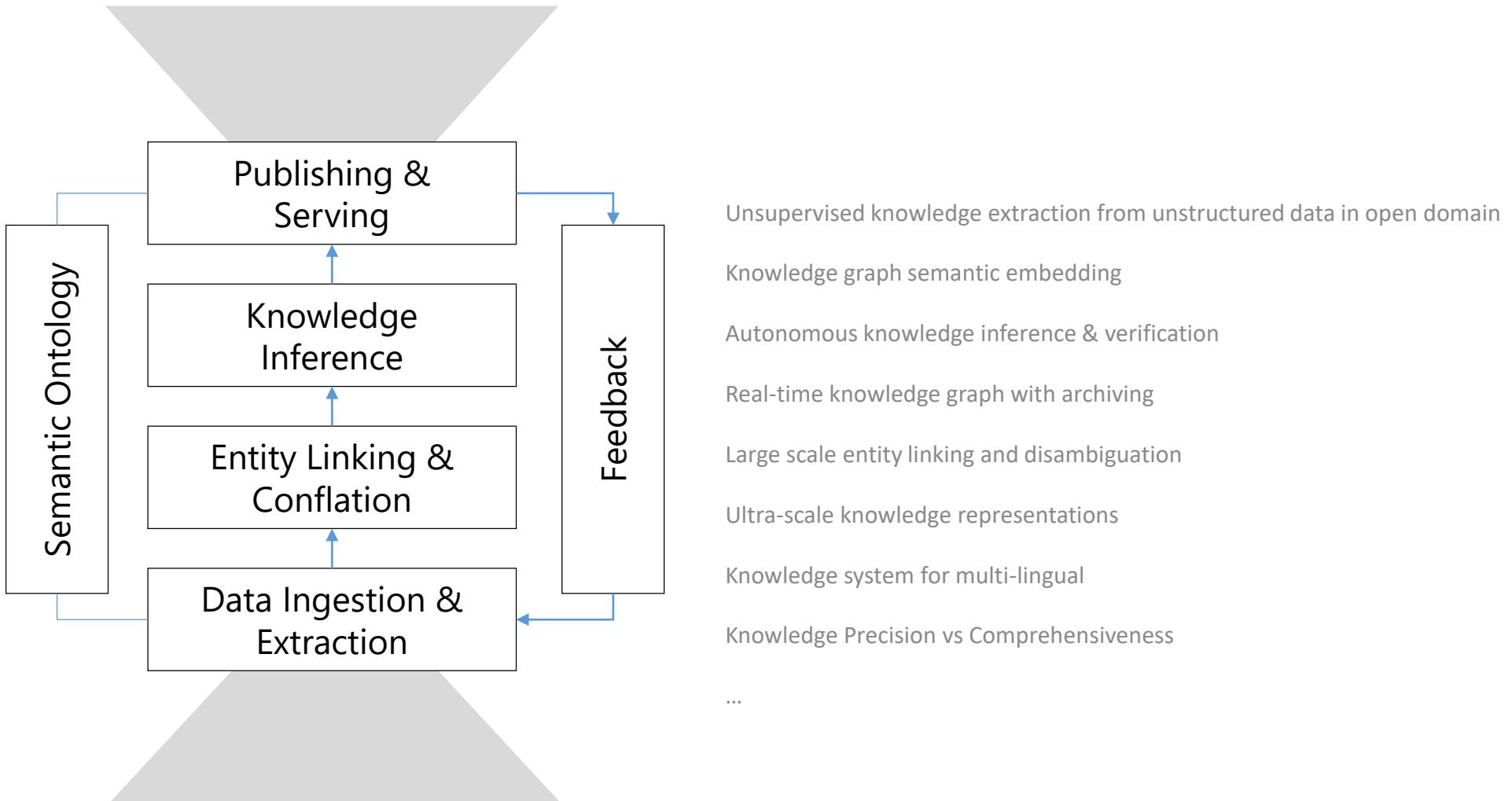
240+ markets

44+ languages

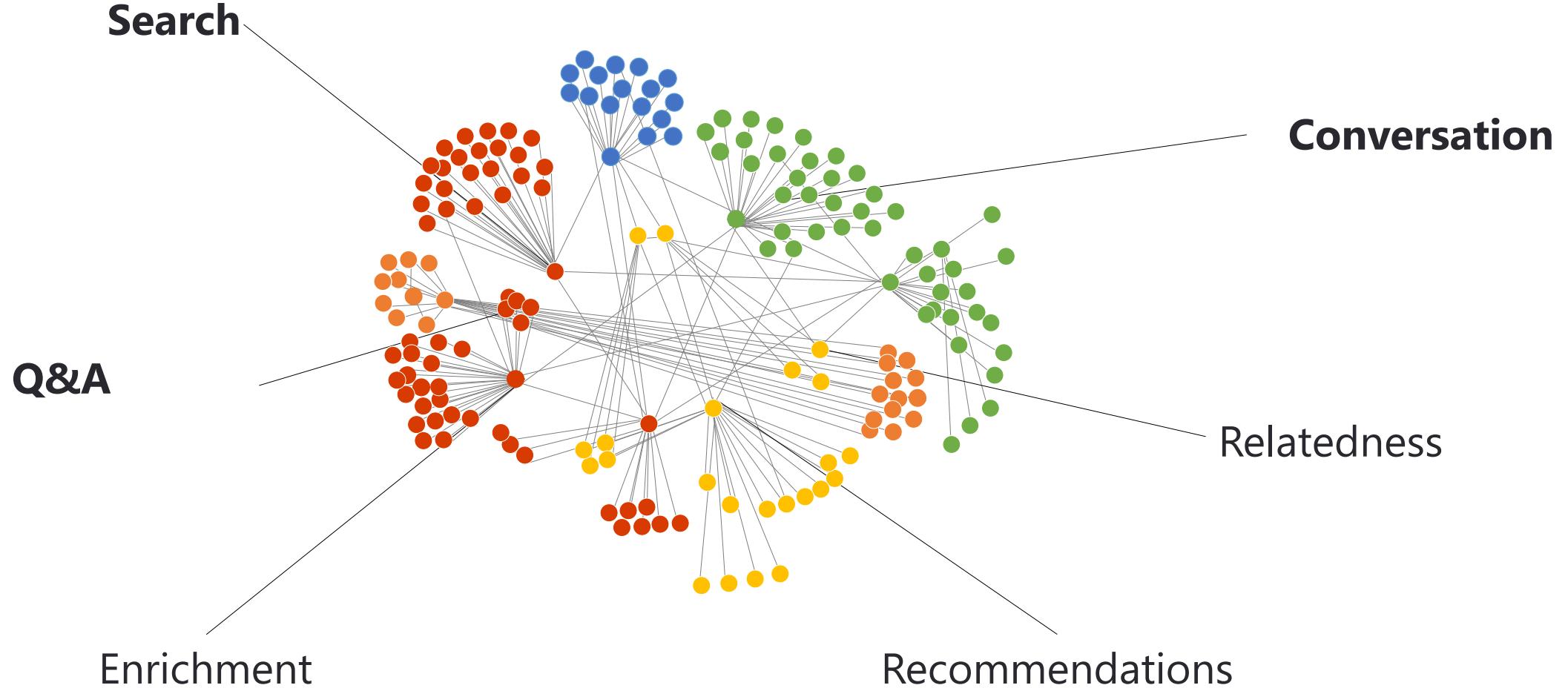
# How do we bring knowledge systems to life?



# Active research and product efforts in knowledge



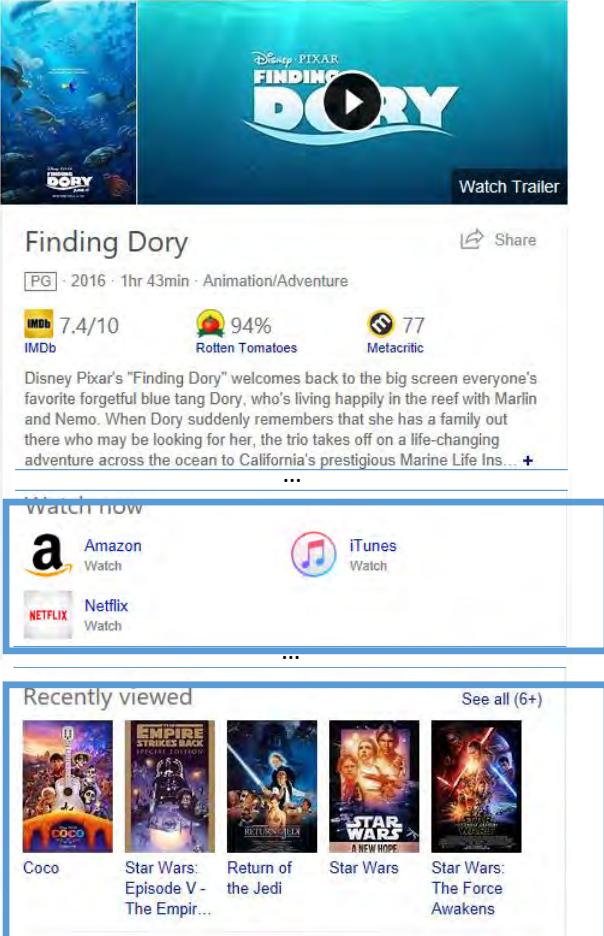
# Infusing knowledge: From search to conversation



# Satori powering Bing Search

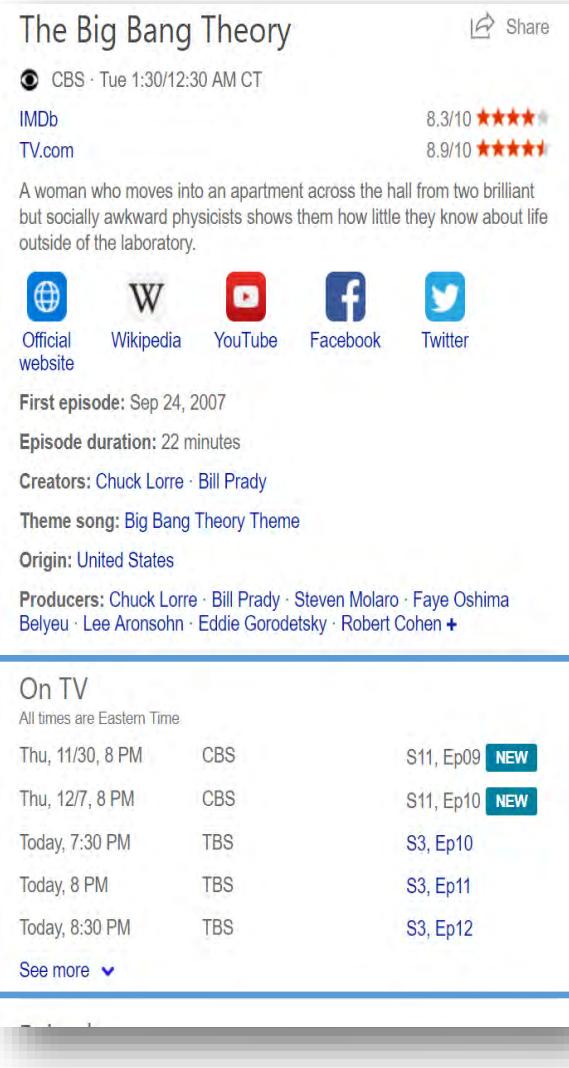
Watch now actions for movie entities

Recently viewed shows personal history



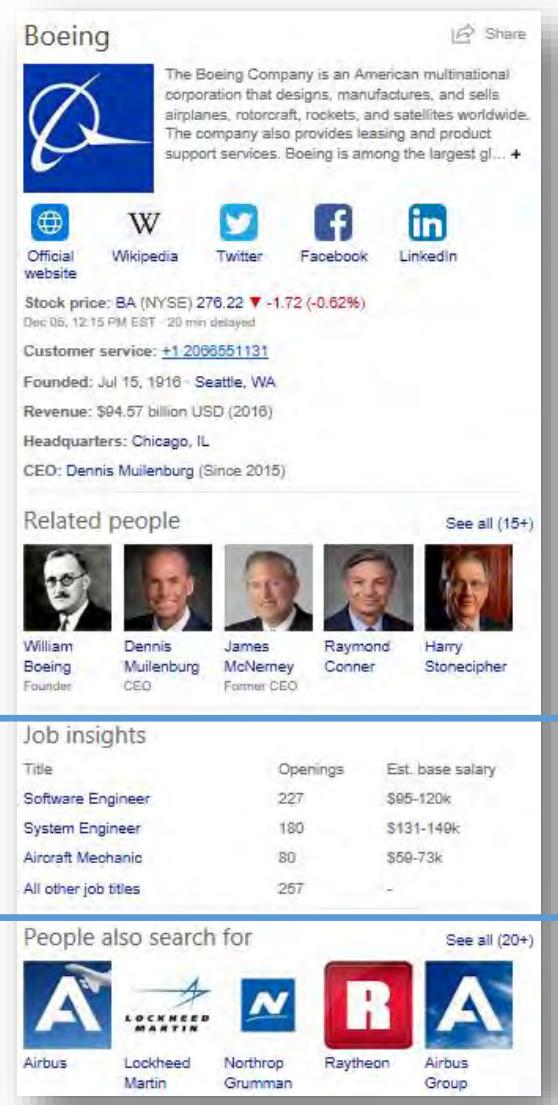
The screenshot shows search results for the movie "Finding Dory". At the top is a large thumbnail image of the movie poster. Below it is a summary card for "Finding Dory" with a play button, a "Watch Trailer" link, and a "Share" icon. The card includes the movie's title, rating (PG), release year (2016), duration (1hr 43min), genre (Animation/Adventure), and links to IMDb (7.4/10), Rotten Tomatoes (94%), and Metacritic (77). A detailed plot summary follows, mentioning Dory's return to the big screen and her adventure across California's Marine Life Ins... A "Watch now" section lists platforms like Amazon Watch, iTunes Watch, and Netflix Watch. A "Recently viewed" section shows thumbnails for "Coco", "Star Wars: Episode V - The Empire...", "Return of the Jedi", "Star Wars: A New Hope", and "Star Wars: The Force Awakens", along with a "See all (6+)" link.

TV listings for TV show entities



The screenshot shows search results for the TV show "The Big Bang Theory". It features a summary card with the show's title, network (CBS), broadcast time (Tue 1:30/12:30 AM CT), and ratings from IMDb (8.3/10) and TV.com (8.9/10). A plot summary describes the show as a woman moving into an apartment across the hall from two brilliant but socially awkward physicists. Below the card are social sharing icons for Official website, Wikipedia, YouTube, Facebook, and Twitter. A "First episode" section indicates the debut on Sep 24, 2007. Further details include episode duration (22 minutes), creators (Chuck Lorre, Bill Prady), theme song (Big Bang Theory Theme), origin (United States), and producers (Chuck Lorre, Bill Prady, Steven Molaro, Faye Oshima, Belyeu, Lee Aronsohn, Eddie Gorodetsky, Robert Cohen). A "On TV" section lists upcoming airings: Thu, 11/30, 8 PM on CBS (S11, Ep09 NEW); Thu, 12/7, 8 PM on CBS (S11, Ep10 NEW); Today, 7:30 PM on TBS (S3, Ep10); Today, 8 PM on TBS (S3, Ep11); and Today, 8:30 PM on TBS (S3, Ep12). A "See more" link is at the bottom.

Job insights for companies



The screenshot shows search results for the company "Boeing". It includes a summary card with the company's logo, name, and a brief description: "The Boeing Company is an American multinational corporation that designs, manufactures, and sells airplanes, rotorcraft, rockets, and satellites worldwide. The company also provides leasing and product support services. Boeing is among the largest gl...". Below the card are links to the company's official website, Wikipedia, Twitter, Facebook, and LinkedIn. Current stock information shows a price of \$276.22, down 1.72 (-0.62%) on Dec 06, 12:15 PM EST (20 min delayed). Other details include customer service number (+1 2068551131), founding date (Jul 15, 1916 - Seattle, WA), revenue (\$94.57 billion USD (2016)), headquarters (Chicago, IL), and CEO (Dennis Muilenburg (Since 2015)). A "Related people" section shows portraits and names for William Boeing (Founder), Dennis Muilenburg (CEO), James McNerney (Former CEO), Raymond Conner, and Harry Stonecipher. A "Job insights" section displays openings and estimated base salaries for various job titles:

Title	Openings	Est. base salary
Software Engineer	227	\$95-120k
System Engineer	180	\$131-149k
Aircraft Mechanic	80	\$59-73k
All other job titles	257	-

A "People also search for" section lists logos for Airbus, Lockheed Martin, Northrop Grumman, Raytheon, and Airbus Group, with a "See all (20+)" link.

# Richer Data for Entity Pane, Carousel, and Facts Across Segments



**Amy Klobuchar**  
United States Senator

Amy Jean Klobuchar is an American former prosecutor, author, and politician serving as the senior United States Senator from Minnesota. She is a member of the Minnesota Democratic-Farmer-Labor Party, an affiliate of the Democratic Party, and Minnesota's first elected female U.S. Senator.

[Official website](#) [Wikipedia](#) [Twitter](#) [Instagram](#) [YouTube](#)

Born: May 25, 1960 (age 57) • Plymouth, MN  
Mailing address: 302 Hart Senate Office Building Washington DC 20510  
Phone: (202) 224-3244  
Office: United States Senator MN (Since 2007)  
Party: Democratic Party  
Previous office: County Attorney of Hennepin County (1999 - 2007)

**Sponsored bills**

Introduced	Number	Title
Apr 26, 2018	S.2774	A bill to reauthorize the COPS ON THE BEAT grant program
Apr 23, 2018	S.2728	Social Media Privacy Protection and Consumer Rights Act of 2018
Apr 18, 2018	S.Res.476	A resolution designating April 2018 as "National 9-1-1 Education Month"

**Cosponsored bills**

**Timeline**

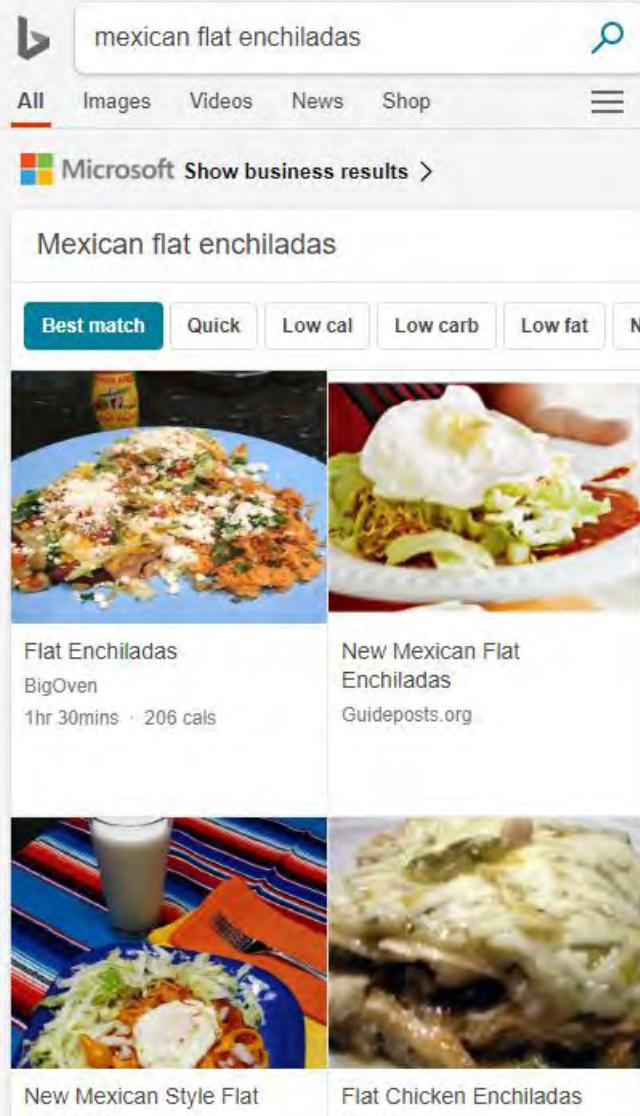
1986: In 1986 she published *Uncovering the Dome*, a case study of the 10-year political struggle behind the building of the Hubert H. Humphrey Metrodome.

1993: Klobuchar and Beissler were married in 1993.

2001: As Hennepin County Attorney, she was named by Minnesota Lawyer in 2001 as "Attorney of the Year" and received a leadership award from Mothers Against Drunk Driving for advocating for successful passage of Minnesota's first felony DWI law.

Show more

**SBS +10.85 weak**

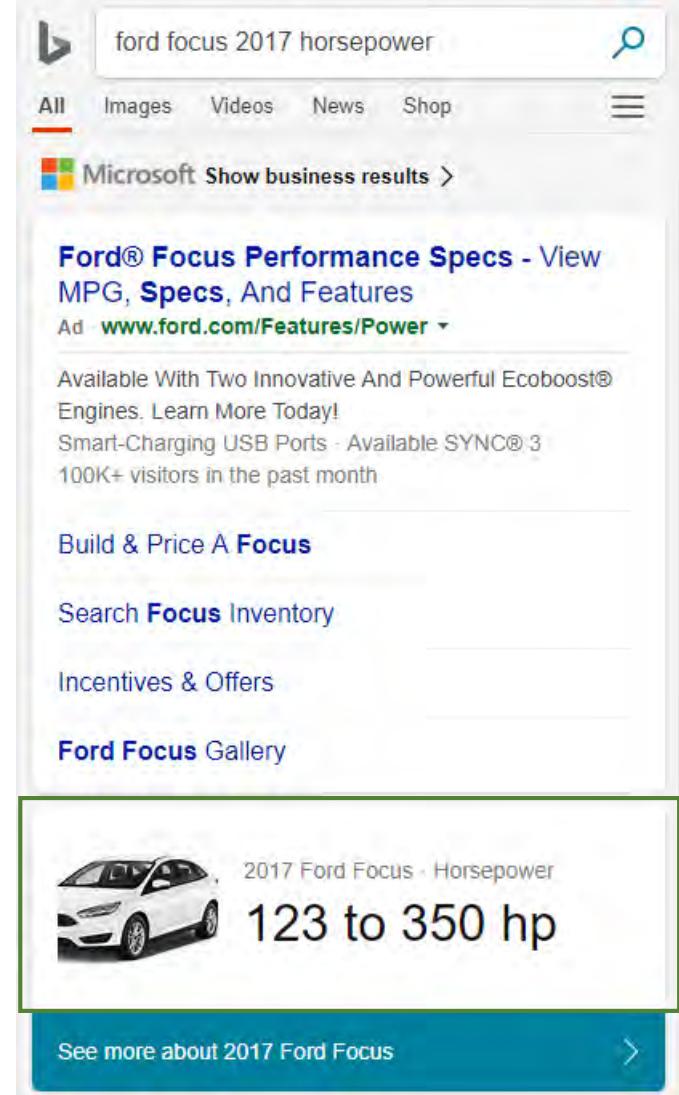
  
mexican flat enchiladas

All Images Videos News Shop

**Mexican flat enchiladas**

**Best match** Quick Low cal Low carb Low fat N

 Flat Enchiladas BigOven 1hr 30mins • 206 cals	 New Mexican Flat Enchiladas Guideposts.org
 New Mexican Style Flat Enchiladas	 Flat Chicken Enchiladas Recipe

  
ford focus 2017 horsepower

All Images Videos News Shop

**Ford® Focus Performance Specs - View MPG, Specs, And Features**  
Ad [www.ford.com/Features/Power](http://www.ford.com/Features/Power)

Available With Two Innovative And Powerful Ecoboost® Engines. Learn More Today!  
Smart-Charging USB Ports • Available SYNC® 3  
100K+ visitors in the past month

**Build & Price A Focus**

**Search Focus Inventory**

**Incentives & Offers**

**Ford Focus Gallery**

 2017 Ford Focus • Horsepower  
**123 to 350 hp**

See more about 2017 Ford Focus >

# Knowledge powered Q&A

## Text-based Q&A

Q Will I qualify for OSAP if I'm new in Canada?

### Selected Passages

"Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free."

Source: <http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/>

"To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD)."

Source: <http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/>

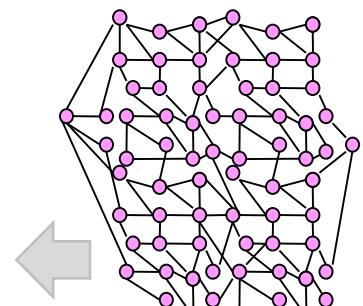
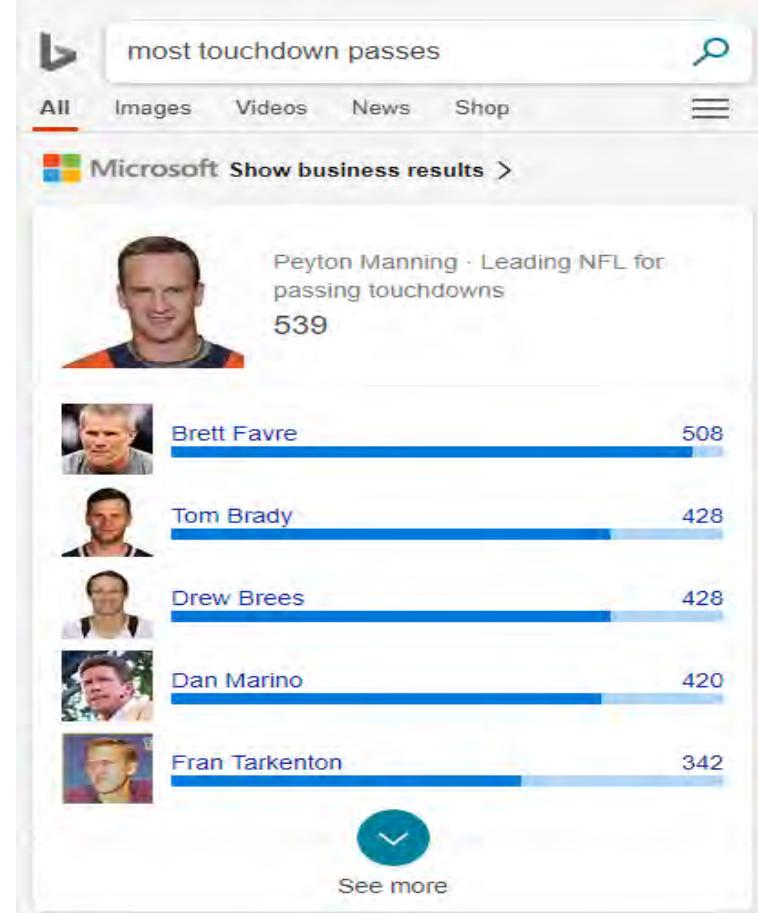
"You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans program. There are also grants, bursaries and scholarships available for both full-time and part-time students."

Source: <http://www.campusaccess.com/financial-aid/osap.html>

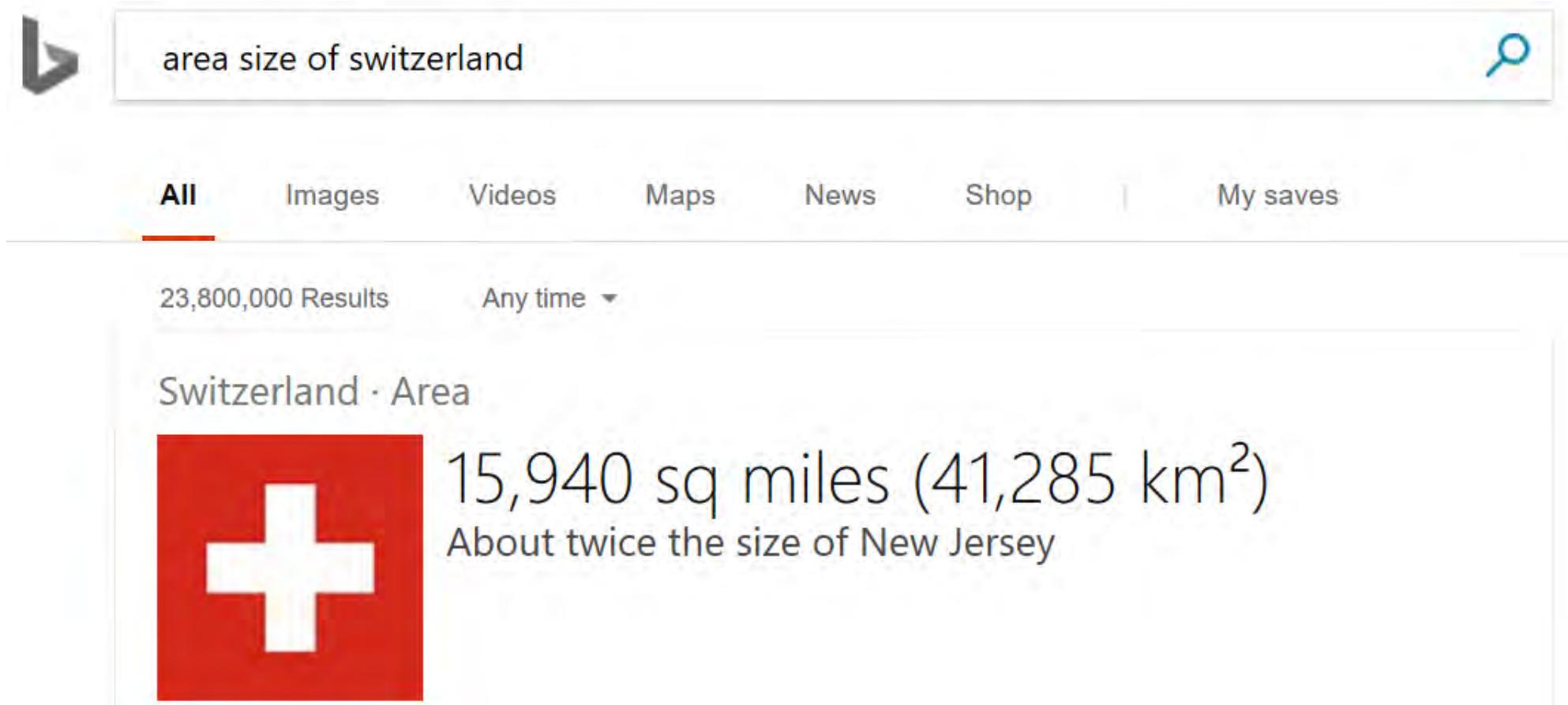
### Answer

No. You won't qualify.

## Knowledge-based Q&A



# Bing – knowledge in answers



The image shows a screenshot of a Bing search results page. At the top left is the Bing logo. The search bar contains the query "area size of switzerland". To the right of the search bar is a magnifying glass icon. Below the search bar is a navigation bar with tabs: "All" (which is underlined in red), "Images", "Videos", "Maps", "News", "Shop", and "My saves". Underneath the navigation bar, it says "23,800,000 Results" and "Any time". The main content area features a large image of the Swiss flag (red with a white cross) on the left. To its right, the text "Switzerland · Area" is displayed above the area information. The area information consists of the text "15,940 sq miles (41,285 km<sup>2</sup>)" followed by the comparison "About twice the size of New Jersey".

area size of switzerland

All Images Videos Maps News Shop My saves

23,800,000 Results Any time

Switzerland · Area



15,940 sq miles (41,285 km<sup>2</sup>)  
About twice the size of New Jersey

# Knowledge graph serves NL fact answers

2017 turing award winners

All Images Videos Maps News Shop | My saves

249,000 Results Any time ▾

Turing Award - Winner(s) in 2017

John L. Hennessy  David A. Patterson 

IBM revenue

All Images Videos Maps News Shop

14,700,000 Results Any time ▾

IBM revenue

US\$ 79.139 billion (2017)

Natural language Fact Answers

how many medals did michael phelps win in rio 

All Images Videos Maps News Shop | My saves

218,000 Results Any time ▾

6

In the 2016 Olympics, Michael Phelps won 5 gold medals for 200m Butterfly, 200m Individual Medley, 4×100m Freestyle Relay, 4×100m Medley Relay, and 4×200m Freestyle Relay and 1 silver medal for 100m Butterfly, for a total of 6 medals overall.

Year	Event	Medal
2016	Swimming 200m Butterfly	Gold
2016	Swimming 200m Individual Medley	Gold
2016	Swimming 4×100m Freestyle Relay	Gold
2016	Swimming 4×100m Medley Relay	Gold
2016	Swimming 4×200m Freestyle Relay	Gold
2016	Swimming 100m Butterfly	Silver

Learn more: [en.wikipedia.org/wiki/Michael\\_Phelps](https://en.wikipedia.org/wiki/Michael_Phelps)

Is this answer helpful?  

# Knowledge graph serves carousel of information

latest films by the director of titanic

All Images Videos Maps News Shop | My saves

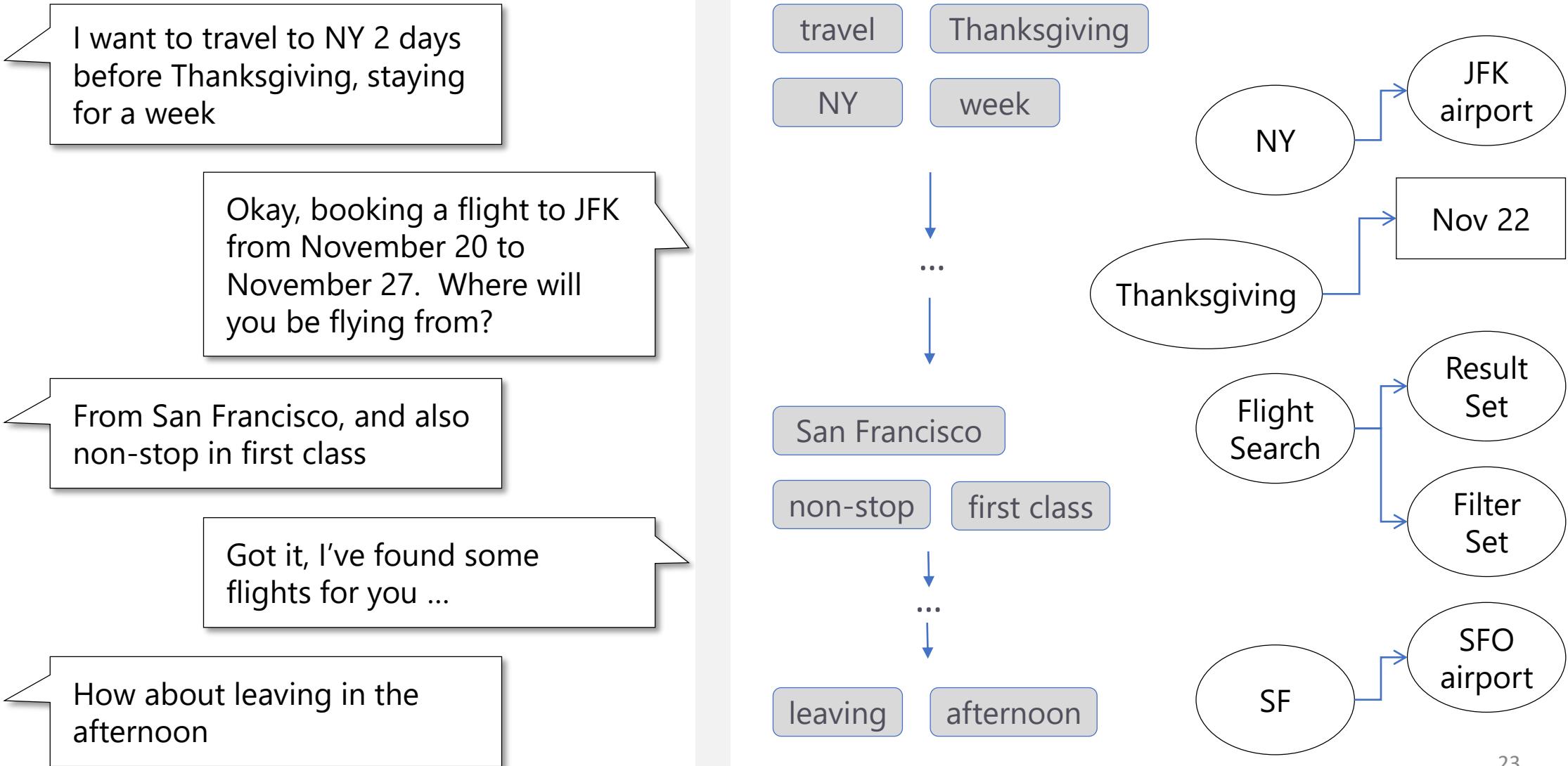
Homes for sale in Bellevue >\$4M Any beds Any baths Any year Compare

Address	Description	Price	Beds	Baths	Sq Ft
809 97th Ave SE, Bellevue, WA 98004	4 bed · 4.75 bath · 6,220 sq ft.	\$4,580,000	4	4.75	6,220
719 96th Ave SE, Bellevue, WA 98004	5 bed · 5.75 bath · 14,140 sq ft.	\$9,988,000	5	5.75	14,140
355 Shoreland Dr SE, Bellevue, WA 98004	5 bed · 4.75 bath · 6,500 sq ft.	\$4,988,000	5	4.75	6,500
12210 NE 33rd St, Bellevue, WA 98005	6 bed · 6.5 bath · 10,088 sq ft	\$6,888,000	6	6.5	10,088
24 Columbia Ky, Bellevue, WA 98006	5 bed · 4 bath · 5,090 sq ft	\$5,400,000	5	4	5,090
4648 NE 95th Ave, Bellevue, WA 98004	4 bed · 5.5 bath · 6,100 sq ft	\$9,400,000	4	5.5	6,100

Latest films by the director of Titanic

<a href="#">Avatar 4</a> Dec 20, 2024 (U)	<a href="#">Avatar 3</a> Dec 17, 2021 (U)	<a href="#">Avatar 2</a> Dec 18, 2020 (U)	<a href="#">Avatar</a> Dec 18, 2009 (PG-13)	<a href="#">Aliens of the Deep</a> Jan 28, 2005 (PG-13)	<a href="#">Ghosts of the Abyss</a> Mar 31, 2003 (PG-13)	<a href="#">Expedition: Bismarck</a> Dec 8, 2002 (U)	<a href="#">Titanic</a> Dec 19, 1997 (PG-13)

# Knowledge-powered Conversation



# Part II: Acquiring Knowledge in the Wild

**Benjamin Han**

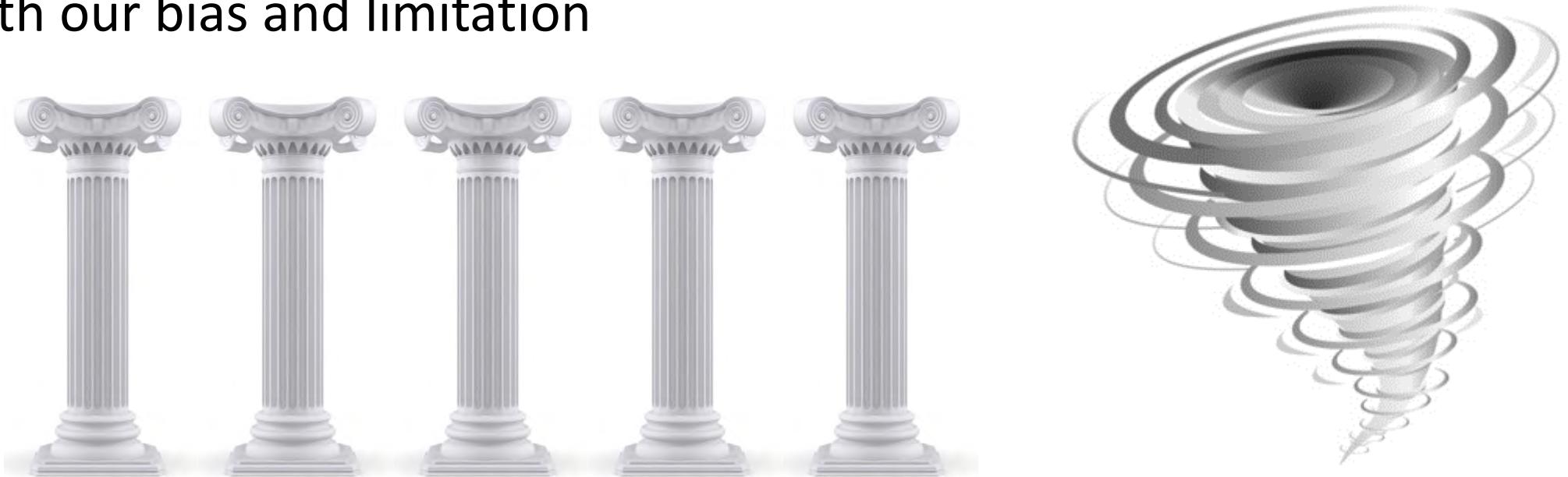
Principal Machine Learning & Data Scientist, Satori Group, Microsoft AI+R

[diha@microsoft.com](mailto:diha@microsoft.com)



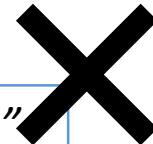
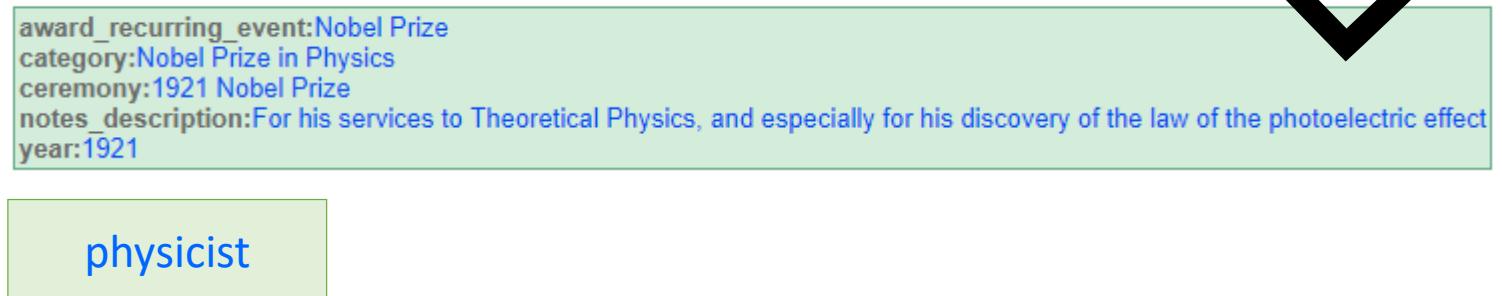
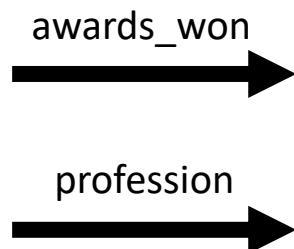
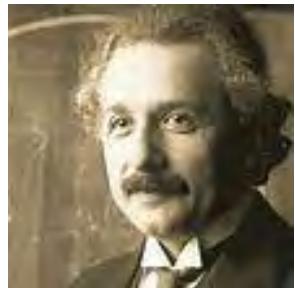
# Goals

- Identify the five pillars of a high-quality Knowledge Acquisition system
- Survey: a whirlwind tour of the proposed approaches
- With our bias and limitation



# Knowledge Graph (KG)

- What is a Knowledge Graph? [\[Paulheim 2016\]](#)
  - KG describes real-world *entities* and their *relations*, organized in a *graph*.
  - Possible classes and relations are defined by schemas.
  - Focus on *instance* aspect of knowledge (A-Box in [Description Logic](#)), not the schema aspect (T-Box in DL).



"If someone won a Nobel Prize in Physics, he must be a physicist."

# Knowledge Acquisition (KA) in the Wild

- Heterogeneous sources/formats/modalities.
- Different domains of knowledge.

*The Hobbit: An Unexpected Journey*

From Wikipedia, the free encyclopedia

**The Hobbit: An Unexpected Journey** is a 2012 epic fantasy adventure film directed by Peter Jackson. It is the first installment in a three-part film adaptation based on the 1937 novel *The Hobbit* by J. R. R. Tolkien. It is followed by *The Desolation of Smaug* (2013) and *The Battle of the Five Armies* (2014), and together they act as a prequel to Jackson's *The Lord of the Rings* film trilogy. The film's screenplay was written by Peter Jackson, his longtime collaborators Fran Walsh and Philippa Boyens, and Guillermo del Toro, who was originally chosen to direct the film before leaving the project in 2010.

The story is set in Middle-earth sixty years before the events of *The Lord of the Rings*, and portions of the film are adapted from the appendices to Tolkien's *The Return of the King*.<sup>10</sup> *An Unexpected Journey* tells the tale of Bilbo Baggins (Martin Freeman), who is convinced by the wizard Gandalf the Grey (Ian McKellen) to accompany thirteen Dwarves, led by Thorin Oakenshield (Richard Armitage), on a quest to reclaim the Lonely Mountain from the dragon Smaug. The ensemble cast also includes James Nesbitt, Ken Stott, Cate Blanchett, Ian Holm, Christopher Lee, Hugo Weaving, Elijah Wood and Andy Serkis, and features Sylvester McCoy, Barry Humphries and Manu Bennett.

*An Unexpected Journey* premiered on November 28, 2012 in New Zealand and was released internationally on December 12, 2012.<sup>11</sup> The film has grossed over \$1 billion at the box office, surpassing both *The Fellowship of the Ring* and *The Two Towers* (nominally, becoming the fourth highest-grossing film of 2012 and the 18th highest grossing film of all time). The film was nominated for three Academy Awards for Best Visual Effects, Best Production Design, and Best Makeup and Hairstyling.<sup>12</sup> It was also nominated for three BAFTA Awards.<sup>13</sup>

Contents [hide]

- 1 Plot
- 2 Cast
- 3 Production
  - 3.1 High frame rate
  - 3.2 Sound
- 4 Distribution
  - 4.1 Marketing
  - 4.1.1 Video game
  - 4.2 Theatrical release
  - 4.3 Home media
- 5 Reception
  - 5.1 Box office
  - 5.2 Critical response
  - 5.3 Accolades

The Hobbit: An Unexpected Journey

Top 5000

Your rating: ★★★★☆ (11)

Rating: 8.0/10 from 575,277 users Metascore: 58/100

Reviews: 1,361 user 638 critic 40 from Metacritic.com

Contact the filmmakers on IMDbPro >

Director: Peter Jackson

Writers: Fran Walsh (screenplay), Philippa Boyens (screenplay), 3 more credits >

Stars: Martin Freeman, Ian McKellen, Richard Armitage See full cast and crew >

Science Home News Journals Topics Careers

Log in | My account | Contact us

Once Again, Physicists Debunk Faster-Than-Light Neutrinos

By Adrian Cho | Jun. 8, 2012, 3:39 PM

Enough already. Five different teams of physicists have now independently verified that elusive subatomic particles called neutrinos do *not* travel faster than light. New results, announced today in Japan, contradict those announced last September by a 170-member crew working with the OPERA particle detector in Italy's subterranean Gran Sasso National Laboratory. The OPERA team made headlines after they suggested neutrinos traveled **0.002% faster** than the speed of light, as predicted by the theory of special relativity. The OPERA theory, however. So instead of the nail in the coffin new suite of results is more like the sod

Zuckerberg and wife Priscilla Chan welcome second daughter August

BTW

Mark Zuckerberg and Priscilla Chan have welcomed their second child, a daughter named August, the pair announced in (of course) a Facebook post Monday afternoon.

"Priscilla and I are so happy to welcome our daughter August!" Zuckerberg wrote in the caption of a life event status update.

CITES CoP17

September 24, 2016 @ 08:00 - October 5, 2016 @ 17:00

The 17th meeting of the Conference of the Parties to CITES (CoP17) will take place in Johannesburg, South Africa from 24 September to 5 October 2016 at the Sandton Convention Center. This will be the fourth meeting of the Conference of the Parties to CITES held on the African continent since CITES came into force on 1 July 1975, but it will be the first held on the continent since 2000. CITES is an international agreement between governments. Its aim is to ensure that international trade in specimens of wild animals and plants does not threaten their survival.

[+ GOOGLE CALENDAR](#) [+ ICAL EXPORT](#)

**Details**

**Start:** September 24, 2016 @ 08:00

**End:** October 5, 2016 @ 17:00

**Event Category:** Business

**Event Tags:** CITES, South Africa

**Websites:** <http://cites.org>

**Porter Hayden Company**

Founded: 1966

Years Operated: 1966-2005

Headquarters: Baltimore, Maryland

Business: Manufactured insulation

Asbestos Trust: Yes

Bankruptcy Status: Filed in 2005 and reorganized in 2007

The logo consists of a stylized industrial building or factory structure with smokestacks emitting smoke, enclosed within a blue square border.

# Knowledge Acquisition (KA) in the Wild

- Hard to ascertain veracity.
- Constantly changing.
- Training data is hard to come by.

**Adam Sandler actor and comedian, found dead at 49**



US actor and comedian Adam Sandler has been found dead, aged 49, in an apparent suicide.

Marin County Police in California said he was pronounced dead at his home shortly after officials responded to an emergency call around noon local time.

Sandler was famous for such films as Happy Gilmore, The Wedding Singer, 50 First Dates, Mr Deeds and more than 40 others.

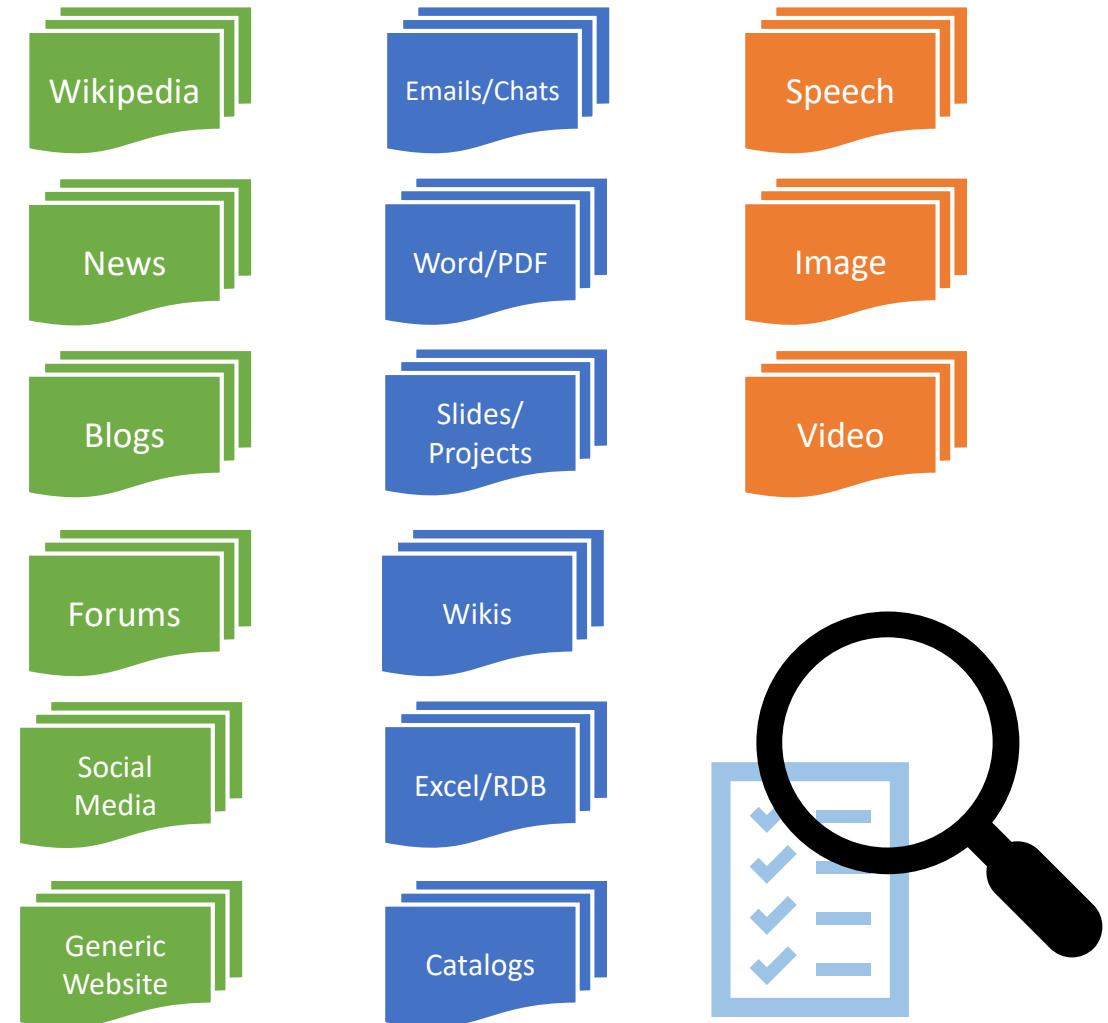


**WHO'S DATED WHO**

The screenshot shows a search bar with 'Search or browse celebrities' and a magnifying glass icon. A red speech bubble above the search bar contains the text 'WHO'S DATED WHO'. Below the search bar, there's a 'Trending Today' section featuring Jeffree Star (American Singer) and Taylor Schilling (American Actress). Both profiles include a green 'Dating' status indicator with a count (21+423 and 29+150 respectively). Below this, there's a grid of seven other celebrities: George Gray, Demi Lovato, Chris Pratt, Selena Gomez, Kourtney Kardashian, Henry Cavill, and Demi Moore, each with their names and a small image.

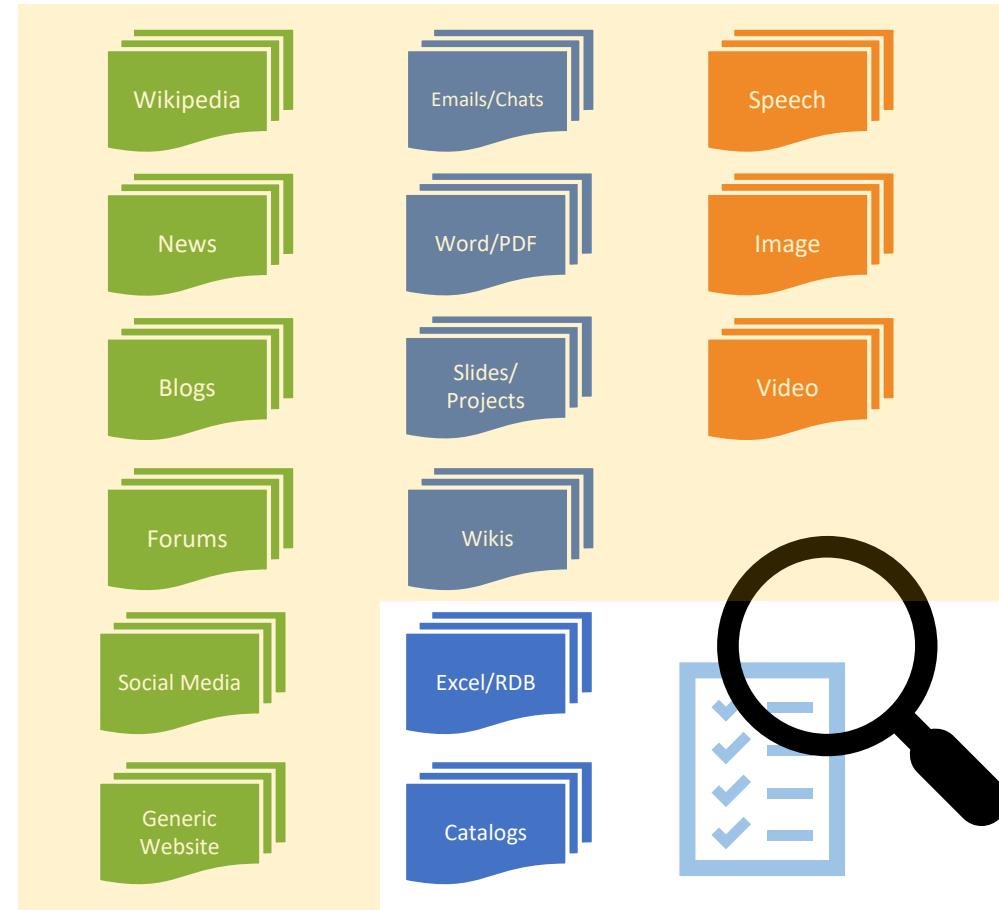
# Five Pillars of High-Quality KA for KG

- Wide Coverage
- High precision
- Verifiable knowledge
- More efficient human intervention
- High system maintainability



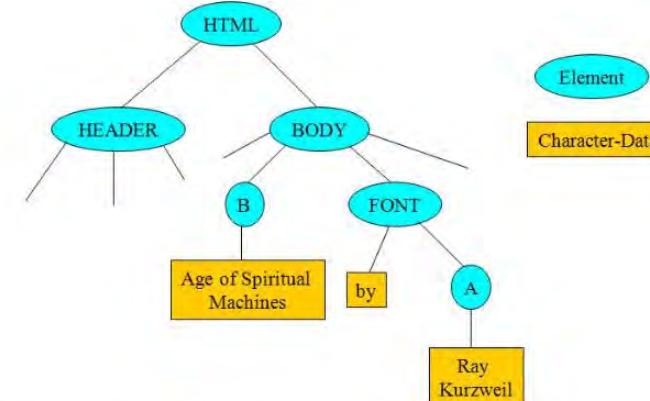
# Wide Coverage

- Knowledge can come from many sources and in many forms
  - Structured sources
    - Relational databases
    - Feeds
    - Catalogues, directories etc
  - Unstructured sources



# Unstructured Sources: Web Pages

- Web wrappers [\[Ferrara+ 2014\]](#)
  - Procedures for extracting user-designated data from web resources to structured form.
  - Major approaches
    - Rule-based: regular expressions, wrapper programming languages etc.
    - Tree-based: segment DOM into data regions, then extract with partial alignment.
    - Machine-learning-based.



Title: HTML → BODY → B → CharacterData

Author: HTML → BODY → FONT → A → CharacterData

[\(source\)](#)

# Degradation of Web Content Extractors

- Web content extractors degrade over time [\[Weninger+ 2015\]](#)
  - Algorithms reflected the state of web at the time.
  - Use of JavaScript and CSS made static HTML much less reliable to extract from.
  - Future: extraction should target *visual rendering*.

Algorithm	Year
Body Text Extractor (BTE)	[11] 2001
Largest Size Increase (LSI)	[16] 2001
Document Slope Curve (DSC)	[30] 2002
Link Quota Filter	[21] 2005
K-Feature Extractor (KFE)	[10] 2005
Advanced DSC (ADSC)	[13] 2007
Content Code Blurring (CCB)	[14] 2008
RoadRunner* (RR)	[7] 2008
Content Extraction via Tag Ratios (CETR)	[31] 2010
BoilerPipe	[17] 2010
Eatiht	[27] 2015

TABLE IV: Content extraction algorithms, with their citation and publication date. \*RoadRunner is a wrapper induction algorithm; all others are heuristic methods.

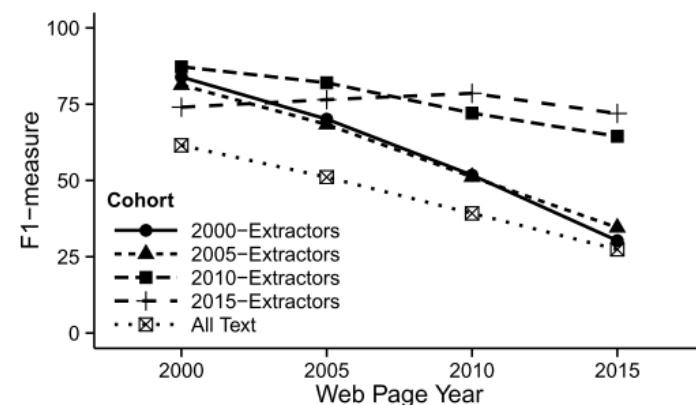


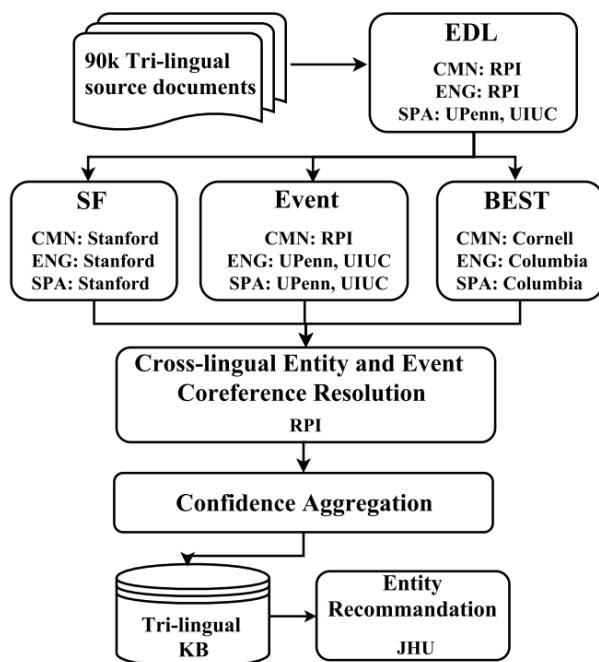
FIG. 3: Mean average  $F_1$  measure per cohort over each lustrum.

# Unstructured Sources: Texts – 1a

- News and forums
  - Continuing from the MUC and ACE, the most important evaluation is TAC KBP (Knowledge Base Population) organized by NIST. [\[Getman+ 2017\]](#)
    - In 2017 five *trilingual* tracks were offered: *Cold Start KB construction*, Entity Discovery & Linking, Slot Filling (relation extraction), Event, and Belief and Sentiment.
      - Cold Start KB: builds a knowledge base from scratch using a given document collection and a predefined KB schema.
      - KB schema: entities, Slot filler relations (finding values for pre-defined attributes), event nuggets and arguments, and sentiments.
    - Datasets include newswire and discussion forums, in English, Chinese and Spanish.

# Unstructured Sources: Texts – 1b

- TACKBP 2017 CSKB best system: Tinkerbell [\[AI-Badrashiny+ 2017\]](#)
  - First end-to-end trilingual system combining multiple building blocks from member institutions..

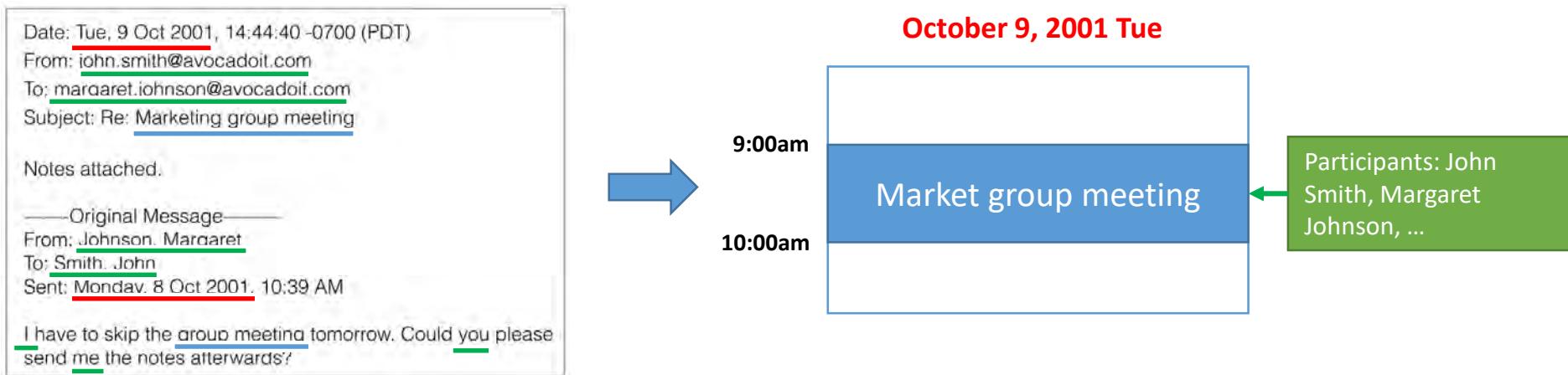


slot types	#justifications	TinkerBell	Human	% Human
all	3	7.56%	47.1%	16.1%
all	1	13.32%	59.77%	22.3%
SF	3	11.43%	40.97%	27.9%
SF	1	17.30%	41.53%	41.7%



# Unstructured Sources: Texts - 2

- Emails & calendars: What can we learn from them?
  - Personal/professional information about people: person entity linking in emails [\[Gao+ 2017\]](#)
  - Information about organization mentions [\[Gao+ 2016\]](#)
  - Linking meeting mentions from emails to calendars [\[Gao+ 2018\]](#)
  - Finding “topics” through clustering and expertise [\[Tang+ 2014\]](#)
  - Extracting problem solving traces in *professional* emails [\[Francois+ 2015\]](#)



# Unstructured Sources: Texts - 3

- Social media: what can we learn from them?
  - Twitter text normalization and named entity recognition [\[Baldwin+ 2015\]](#)
  - Two shared tasks held in 2015

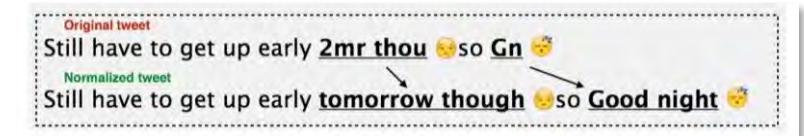
Team name	Precision	Recall	F1	Method highlights
NCSU_SAS_NING	0.9061	0.7865	0.8421	Random Forest
NCSU_SAS_WOOKHEE	0.9136	0.7398	0.8175	Lexicon + LSTM
NCSU_SAS_SAM	0.9012	0.7437	0.8149	ANN
IITP	0.9026	0.7191	0.8005	CRF + Rule
DCU-ADAPT	0.8190	0.5509	0.6587	Generalized Perceptron
LYSGROUP	0.4646	0.6281	0.5341	Spanish Normalization Adaption

Table 3: Results of the constrained systems for the lexical normalization shared task

Team name	Precision	Recall	F1	Method highlights
IHS_RD	0.8469	0.8083	0.8272	Lexicon + CRF + DidYouMean
USZEGED	0.8606	0.7564	0.8052	CRF + n-gram[ s]
BEKLI	0.7743	0.7416	0.7571	Lexicon + Rule + Ranker
GIGO	0.7593	0.6963	0.7264	N/A
LYSGROUP	0.4592	0.6296	0.5310	Spanish Normalization Adaption

Table 4: Results of the unconstrained systems for the lexical normalization shared task

Text normalization



	POS	Orthographic	Gazetteers	Brown clustering	Word embedding	ML
BASELINE	-	✓	✓	-	-	CRFsuite
Hallym	✓	-	-	✓	correlation analysis	CRFsuite
iitp	✓	✓	✓	-	-	CRF++
lattice	✓	✓	-	✓	-	CRF wapiti
multimedialab	-	-	-	-	word2vec	FFNN
NLANGP	-	✓	✓	✓	word2vec & GloVe	CRF++
nrc	-	-	✓	✓	word2vec	semi-Markov MIRA
ousia	✓	✓	✓	-	✓	entity linking
USFD	✓	✓	✓	✓	-	CRF L-BFGS

Table 7: Features and machine learning approach taken by each team.

	Precision	Recall	$F_{\beta=1}$		Precision	Recall	$F_{\beta=1}$
ousia	57.66	55.22	56.41	ousia	72.20	69.14	70.63
NLANGP	63.62	43.12	51.40	NLANGP	67.74	54.31	60.29
nrc	53.24	38.58	44.74	USFD	63.81	56.28	59.81
multimedialab	49.52	39.18	43.75	multimedialab	62.93	55.22	58.82
USFD	45.72	39.64	42.46	nrc	62.13	54.61	58.13
iitp	60.68	29.65	39.84	iitp	63.43	51.44	56.81
Hallym	39.59	35.10	37.21	Hallym	58.36	48.5	53.01
lattice	55.17	9.68	16.47	lattice	58.42	25.72	35.71
BASELINE	35.56	29.05	31.97	BASELINE	53.86	46.44	49.88

Table 8: Results segmenting and categorizing entities into 10 types.

Table 9: Results on segmentation only (no types).

NER

# Unstructured Sources: Texts - 4

- Extracting events and attributes [\[Wang+ 2015\]](#)
- Extracting user profiles [\[Jiwei+ 2015\]](#)
- Extracting computer security events [\[Ritter+ 2015\]](#)
- Extracting emerging entities using seeds [\[Brambilla+ 2017\]](#)
- Quantitative Information Extraction From Social Data [\[Alonso & Sellam 2018\]](#)

Victim	Date	Category	Sample Tweet
namecheap	Feb-20-2014	DDoS	My site was down due to a DDoS attack on NameCheap's DNS server. Those are lost page hits man...
bitcoin	Feb-12-2014	DDoS	Bitcoin value dramatically drops as massive #DDOS attack is waged on #Bitcoin <a href="http://t.co/YdoygOGmhv">http://t.co/YdoygOGmhv</a>
europe	Feb-20-2014	DDoS	Record-breaking DDoS attack in Europe hits 400Gbps.
barcelona	Feb-18-2014	Account Hijacking	Lmao, the official Barcelona account has been hacked.
adam	Feb-16-2014	Account Hijacking	@adamlambert You've been hacked Adam! Argh!
dubai	Feb-09-2014	Account Hijacking	Dubai police twitter account just got hacked!
maryland	Feb-20-2014	Data Breach	SSNs Compromised in University of Maryland Data Breach: <a href="https://t.co/j69VeJC4dw">https://t.co/j69VeJC4dw</a>
kickstarter	Feb-15-2014	Data Breach	I suspect my card was compromised because of the Kickstarter breach. It's a card I don't use often but have used for things like that.
tesco	Feb-14-2014	Data Breach	@directhex @Tesco thanks to the data breach yesterday it's clear no-one in Tesco does their sysadmin housekeeping!

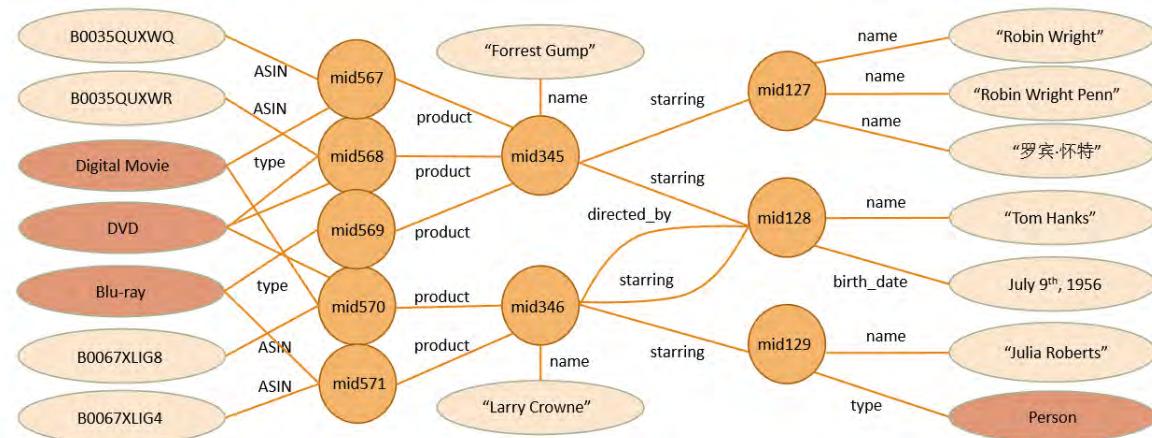
Table 6: Example high-confidence events extracted using our system.

Topic	P@5	sample of quant frags
Tax reform	0.8	corporate tax rate will be 21%; single mother a 70% tax cut
Tubbs fire	0.8	killed at least 11 people and destroyed 1,500 homes; 30–50 mph winds expected till midday
Harvey	0.8	category 4 hurricane with maximum sustained winds of 130 mph; once in 500 year flood
Irma	0.8	category 5 hurricane with 175 mph winds; three hurricanes simultaneously in the Atlantic
Superbowl 2018	0.8	NBC going dark for 30 seconds; Only two QBs have ever beaten Tom Brady
SOTU 2018	0.6	allocated \$700 billion for military; PolitiFact 498 times
Oscars 2017	0.8	At age 98, her story continues to inspire; Jackie Chan has been in films since the 1960's
The Last Jedi	0.6	highest rated at 96% with 83 reviews; \$200 million-plus opening weekend
Las Vegas	0.8	50+ dead, 200+ injured; 14 in critical condition one suspect down

Table 3: Precision evaluation for topics.

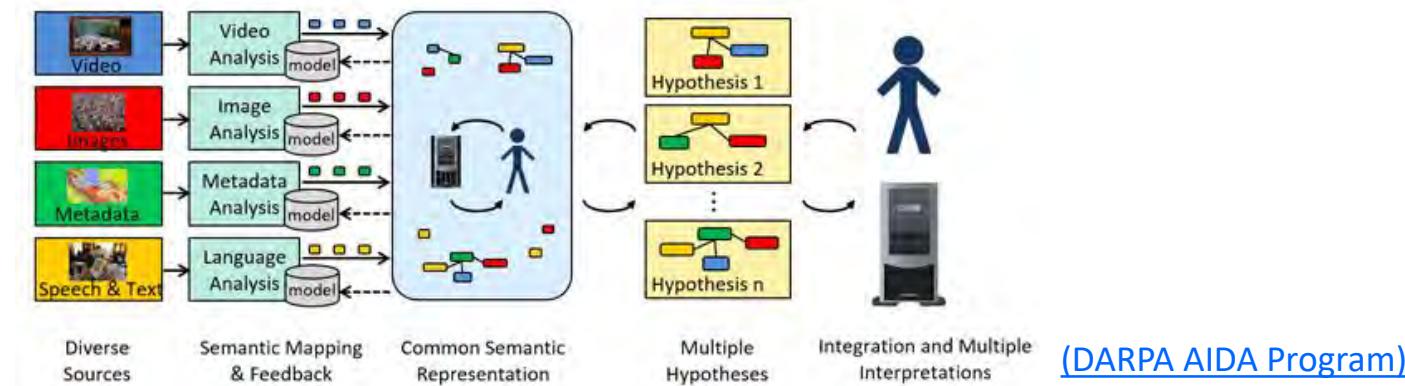
# Unstructured Sources: Texts - 5

- Catalog: Product Knowledge Graph [\[Dong 2017\]](#)
  - No major sources to curate product knowledge from
  - Wikipedia does not help too much
  - A lot of structured data buried in text descriptions in Catalog
  - Retailers gaming with the system so noisy data
  - Large # of products and categories, changing everyday
  - Many entities are not named



# Unstructured Sources: Other Modalities

- Speech, images, video
  - ImageCLEF competition [\[Ionescu+ 2017\]](#)
    - Lifelogging data retrieval and summarization; medical images to textual description/classification; discover unknown info from Earth observation images
  - TACKBP 2018 – Streaming Multimedia Knowledge Base Population [\[web\]](#)
    - Evaluate systems for extracting and aggregating knowledge from heterogeneous sources such as multilingual multimedia sources including text, speech, images, videos, and pdf files, and developing hypotheses interpreting the input.



# Coverage: Extracting Entities - 1

- *Joint* entity and relation extraction
  - Incremental joint extraction [\[Li & Ji 2014\]](#)
  - With a novel tagging scheme [\[Zheng+ 2017\]](#)
  - **With knowledge bases** [\[Ren+ 2016\]](#)

# Coverage: Extracting Entities - 2

- [Ren+ 2016] Framework CoType

- Produce candidate entity mentions using POS then candidate relation mentions; generate training set using the labels from KB
- Jointly embed relation and entity mentions, text features and labels
- Estimate type labels for test relation mentions and their argument mentions

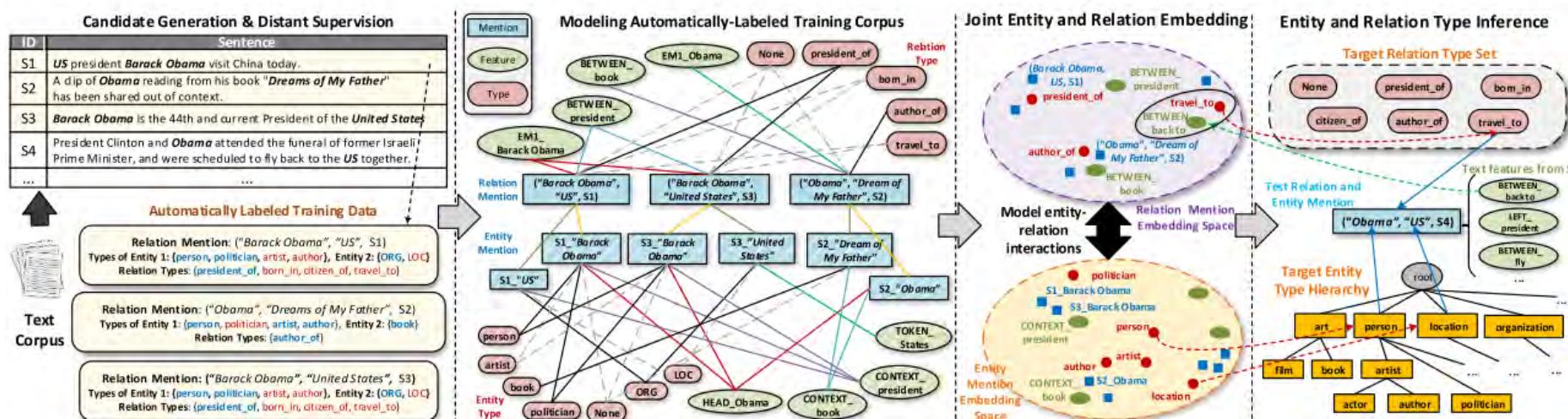


Figure 2: Framework Overview of CoType.

# Coverage/Precision: Entity Linking

- Disastrous result if linking failed, even with perfect extraction



# NEMO (Named Entities Made Obvious) - 1

[\[Cucerzan 2007\]](#)

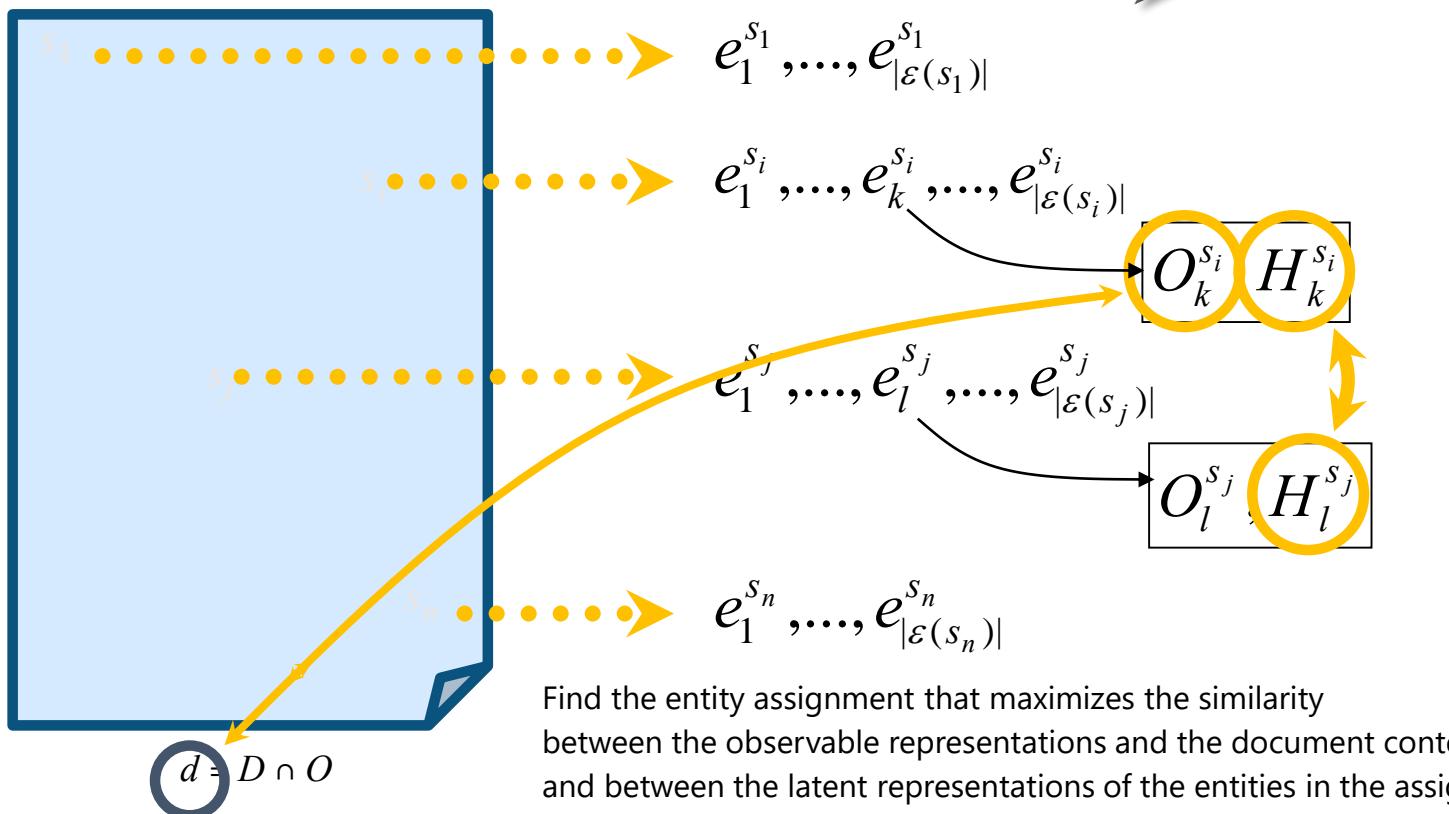
- The best evidence for entity disambiguation is provided by the set of co-occurring entities
  - Extract and disambiguate jointly all entities in a target document
  - Employ both observable attributes (known values, contexts) and latent attributes (e.g. entity relationships, topics)
- Syntax and local context are important - one-sense-per-discourse does not hold
  - Employ both whole-document and local context features

# NEMO - 2

Disambiguation - Intuition

Each entity has multiple vectorial representations

Text document  $D$



[Cucerzan 2007]

# NEMO - 3

## NIST/LDC Evaluations

Accuracy	NEMO system (2014)	best result in the TAC evaluation
TAC 2011 test set	89.3 %	86.8% (MSR/NEMO)
TAC 2012 test set	80.4 %	76.2% (MSR/NEMO)
TAC 2013 test set	85.2 %	83.2% (MSR/NEMO)
TAC 2014 test set	86.8 %	86.8% (MSR/NEMO)

Google-Microsoft-Yahoo ERD Challenge (best participating system) [\[Carmel+ 2014\]](#)

	Precision	Recall	F-measure
ERD 2014 train set	83.7%	72.6%	0.778
ERD 2014 test set	83.3%	69.9%	0.760

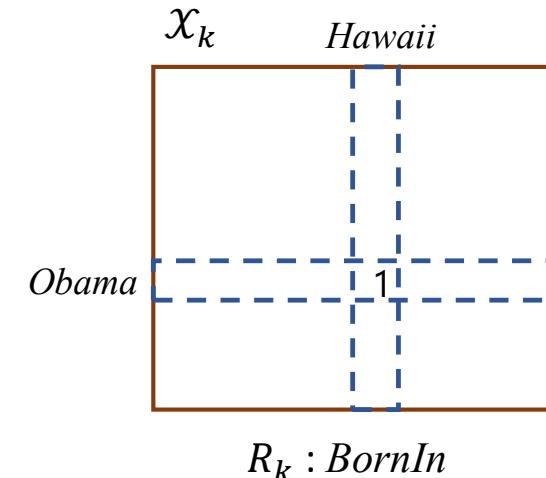
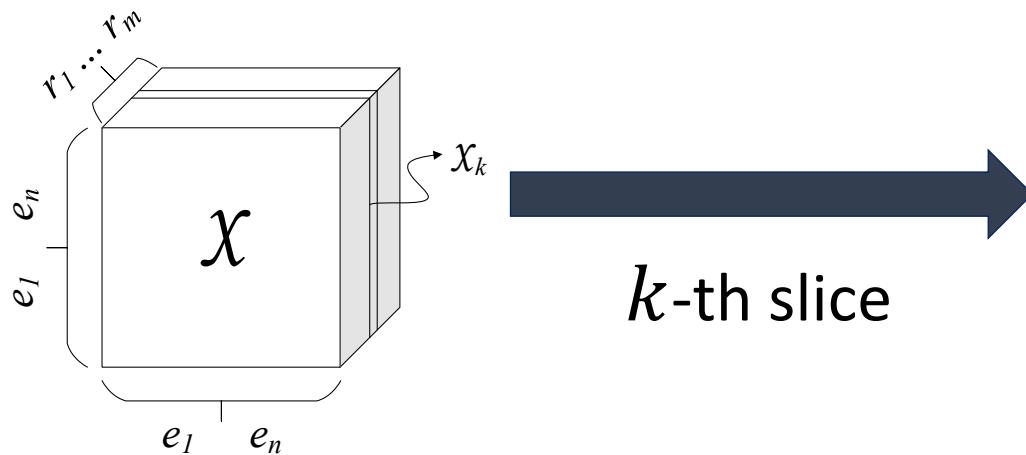
# Coverage: Extracting Relations

- Predicting relations based on existing ones using Tensor NN [\[Socher+ 2013\]](#)
- Universal Schemas [\[Riedel+ 2013\]](#)
- Type-constrained learning in KG [\[Krompaß+ 2015\]](#)
- Association rules mining [\[Kolthoff & Dutta 2015\]](#)
- Embedding-based methods [\[Zhao+ 2015\]](#) [\[Bishan+ 2015\]](#) [\[Toutanova 2015\]](#)  
[\[Goyal & Ferrara 2017\]](#) [\[Shen+ 2016\]](#)
- Reinforcement Learning [\[Feng+ 2018\]](#)
- Open IE [\[Cui+ 2018\]](#)
- Web search [\[West+ 2014\]](#)
- Survey of relational ML for Knowledge Graphs [\[Nickel+ 2015\]](#)

# Embedding Methods for KB Completion - 1

- Each entity in a KB is represented by an  $R^d$  vector
- Predict whether  $(e_1, r, e_2)$  is true by  $f_r(\mathbf{v}_{e_1}, \mathbf{v}_{e_2})$
- Work on KB embedding
  - Tensor decomposition
    - RESCAL [\[Nickel+ ICML-11\]](#), TRESCAL [\[Chang+ EMNLP-14\]](#)
  - Neural networks
    - SME [\[Bordes+ AISTATS-12\]](#), NTN [\[Socher+ NIPS-13\]](#), TransE [\[Bordes+ NIPS-13\]](#)

# Embedding Methods for KB Completion - 2



- Objective: 
$$\frac{1}{2} \left( \underbrace{\sum_k \|X_k - A\mathcal{R}_k A^T\|_F^2}_{\text{Reconstruction Error}} \right) + \frac{1}{2} \left( \underbrace{\|A\|_F^2 + \sum_k \|\mathcal{R}_k\|_F^2}_{\text{Regularization}} \right)$$

$$X_k \approx A \times \mathcal{R}_k \times A^T$$

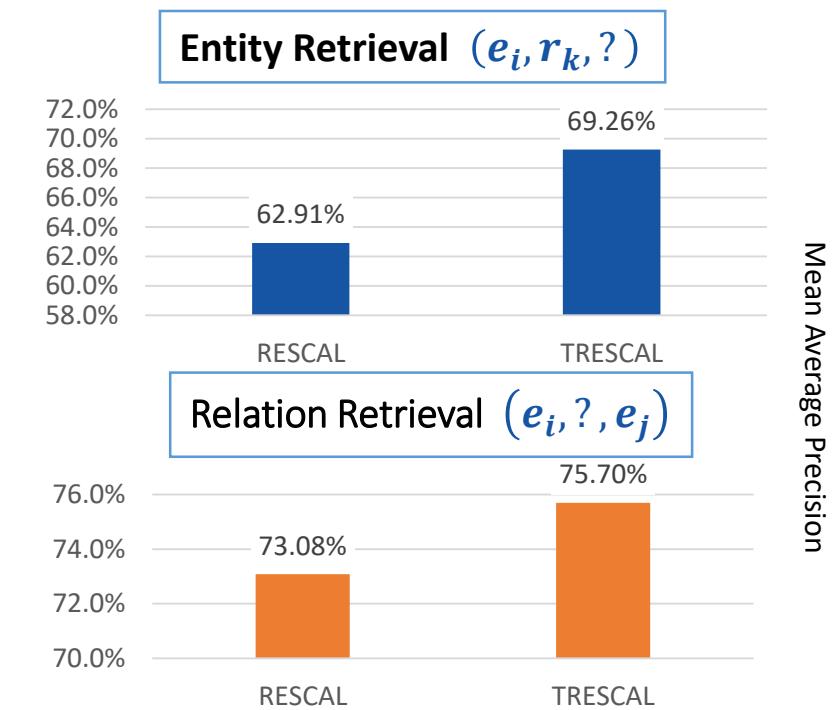
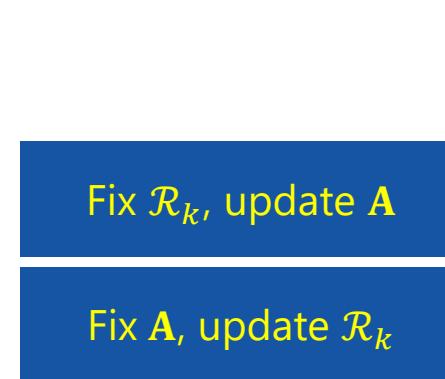
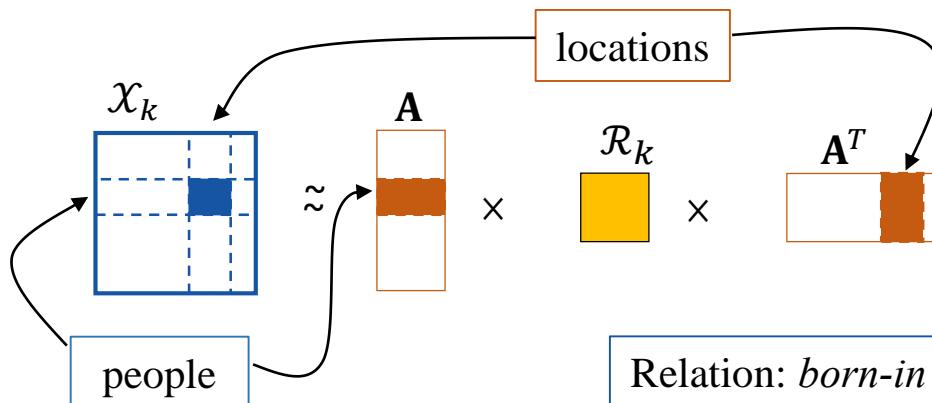
*k-th relation*

$$f_{BornIn}(Obama, Hawaii) = A_{Obama,:} \mathcal{R}_{BornIn} A_{Hawaii,:}^T$$

RESCAL [\[Nickel+ ICML-11\]](#)

# Embedding Methods for KB Completion - 3

- Typed tensor decomposition (TRESCAL) [\[Chang+ EMNLP-14\]](#)
  - Only legitimate entities are included in the loss
  - Faster model training time (4.6x speedup), highly scalable, higher accuracy
  - Reconstruction error:  $\frac{1}{2} \sum_k \|x_k - A\mathcal{R}_k A^T\|_F^2$
  - Training: Alternating Least-Squares (ALS)



# Relation Extraction from Semi-Structured Sources

- Wikipedia tables [Muñoz+ 2013](#).
- Wikipedia list pages [Paulheim & Ponzetto 2013](#)
- Web tables [Ritze+ 2015](#)
- [Microsoft Kable](#): Large scale unsupervised template learning

# Verifiable Knowledge - 1

- Not everything accurately extracted is *fact*
  - Knowledge-based Trust [\[Dong+ 2015\]](#)
- Many recent efforts on assessing truth and finding supports
  - Multilingual answer validation [\[Rodrigo+ 2009\]](#) [\[Kobayashi+ 2017\]](#)
  - FactChecker [\[Nakashole & Mitchell 2014\]](#)
  - PolitiFact [\[Vlachos & Riedel 2014\]](#), [\[Wang 2017\]](#)
  - Fake News challenge [\[Pomerleau & Rao 2017\]](#)
  - Fake news detection via crowd signals [\[Tschiatschek+ 2018\]](#)
  - **Fact Verification competition** [\[Thorne+ 2018\]](#)

# Verifiable Knowledge - 2

- Fact Verification competition (FEVER) [\[Thorne+ 2018\]](#)
  - Goal: given a claim
    - Label claim SUPPORTS, REFUTES, or NOT-ENOUGH-INFO
    - For the first two classes, select relevant sentences from Wikipedia intro sections.
  - Largest annotated fact sets
    - 185,445 annotated claims.
    - Claims generated by mutating Wikipedia sentences: paraphrasing, negation, substitution of entity/relation, generalize/specialize claims.

**Claim:** The Rodney King riots took place in the most populous county in the USA.

[\[wiki/Los Angeles Riots\]](#)

The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[\[wiki/Los Angeles County\]](#)

Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

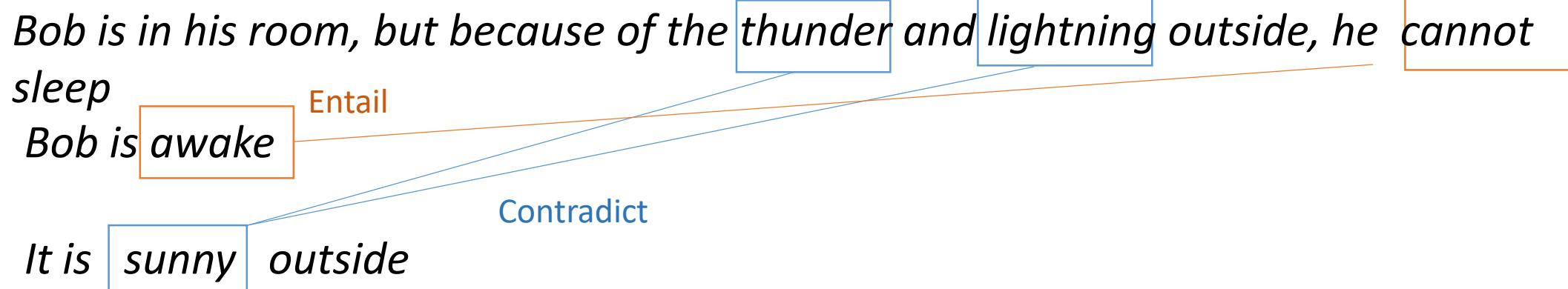
**Verdict:** Supported

# Verifiable Knowledge - 3

- FEVER baseline – sentence classification [\[Thorne+ 2018\]](#)

Basic idea: align parts of the text in sentences **a** and **b** and then aggregate info to predict the label

## Example



# Verifiable Knowledge - 4

- Decomposable Attention model (DA) [\[Thorne+ 2018\]](#)
  - Attend
    - Create soft alignment matrix to produce aligned subphrases between **a** and **b**
    - Alignments are learned using feedforward model **F**
  - Compare
    - Score aligned subphrases using a function **G**
    - **G** is a feedforward model which produces comparison vectors
  - Aggregate
    - Sum over comparison vectors and produce final score using feedforward model **H**

20	jamesthorne	FEVER Baseline	0.1826	0.4884	0.2745
----	-------------	----------------	--------	--------	--------

# Verifiable Knowledge - 4

#	User	Team Name	Evidence F1	Label Accuracy	FEVER Score
1	chaonan99	UNC-NLP	0.5296	0.6821	0.6421
2	tyoneda	UCL Machine Reading Group	0.3497	0.6762	0.6252
3	littsler	Athene UKP TU Darmstadt	0.3697	0.6546	0.6158
4	papelo		0.6485	0.6108	0.5736
5	chidey		0.2969	0.5972	0.4994
6	Tuhin	ColumbiaNLP	0.3533	0.5745	0.4906
7	nanjiang	The Ohio State University	0.5853	0.5012	0.4342
8	wotto	gesis cologne	0.1960	0.5415	0.4077
9	tomoki	Fujixerox	0.1649	0.4713	0.3881
10	nayeon7lee		0.4912	0.5125	0.3859

(web)

# Verifiable Knowledge - 5

- Techniques rooted in core NLP fields
  - Textual Entailment [\[Dagan+ 2006\]](#)
  - Natural language inference [\[Angeli & Manning 2014\]](#)

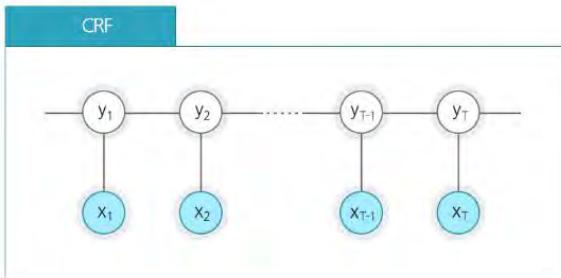
# More Efficient Human Intervention - 1

- Slot tagging using search click logs [\[Kim & Sarikaya 2015\]](#)
  - Slot tagging for queries: “when is the new bill murray movie release date?”
  - Weakly supervised: project labels from structured data found in *click logs*.



# More Efficient Human Intervention - 2

- **[Kim & Sarikaya 2015]** CRF variants to learn from partially labeled sequences

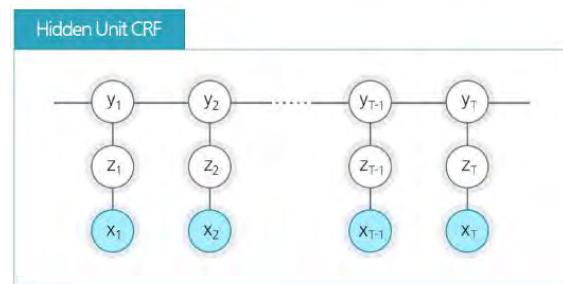


$$p_\theta(y|x) = \frac{\exp(\theta^\top \Phi(x, y))}{\sum_{y' \in \mathcal{Y}(x)} \exp(\theta^\top \Phi(x, y'))}$$

$$\theta^* = \arg \max_{\theta \in \mathbb{R}^d} \sum_{i=1}^N \log p_\theta(y^{(i)}|x^{(i)}) - \frac{\lambda}{2} \|\theta\|^2$$

Initialization:

- Cluster unlabeled data
- Train fully supervised HUCRF with cluster labels
- Keep learned  $\theta$  (between input  $x$  and hidden  $z$ ) and start task-specific training



$$p_{\theta, \gamma}(y, z|x) = \frac{\exp(\theta^\top \Phi(x, z) + \gamma^\top \Psi(z, y))}{\sum_{\substack{z' \in \{0,1\}^n \\ y' \in \mathcal{Y}(x, z')}} \exp(\theta^\top \Phi(x, z') + \gamma^\top \Psi(z', y'))}$$

$$p_{\theta, \gamma}(y|x) = \sum_{z \in \{0,1\}^n} p_{\theta, \gamma}(y, z|x)$$

**Partially observed CRF**

$$p_\theta(\mathcal{Y}(x, \tilde{y})|x) = \sum_{y \in \mathcal{Y}(x, \tilde{y})} p_\theta(y|x)$$

$$\mathcal{Y}(x_j, \tilde{y}_j) = \begin{cases} \{\tilde{y}_j\} & \text{if } \tilde{y}_j \text{ is given} \\ \mathcal{Y}(x_j) & \text{otherwise} \end{cases}$$

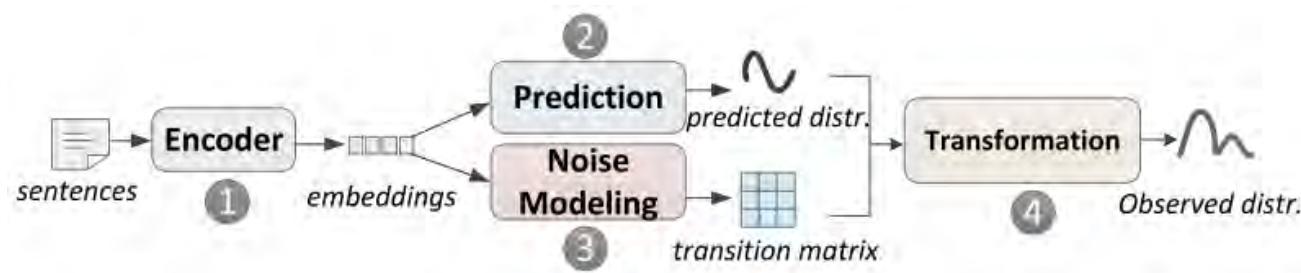
$$\theta^* = \arg \max_{\theta \in \mathbb{R}^d} \sum_{i=1}^N \log p_\theta(\mathcal{Y}(x^{(i)}, \tilde{y}^{(i)})|x^{(i)}) - \frac{\lambda}{2} \|\theta\|^2$$

Domains	games	music	movies	AVG.
CRF	74.21	37.13	68.58	59.97
POCRF	77.23	44.55	76.89	66.22
POHCRF	78.93	46.81	76.46	67.40
POHCRF+	<b>79.28</b>	<b>47.35</b>	<b>78.33</b>	<b>68.32</b>

F1 scores

# More Efficient Human Intervention - 3

- Distant supervision (DS) [Mintz+ 2009] [Gerber & Ngomo 2011] [Gerber+ 2013]
  - Enhance DS with dynamic transition matrix [\[Luo+ 2017\]](#)
    - Problem of DS: label noise
      - Triple <Donald Trump, born-in, New York> picked “Donald Trump worked in New York City” as positive example.
      - Solution: model noise via a transition matrix  $T_{ij}$  indicating the conditional probability for the input sentence to be labeled as relation  $j$  by DS, given  $i$  as the true relation.



Transition matrix:  $T_{ij} = \frac{\exp(\mathbf{w}_{ij}^T \mathbf{x}_n + b)}{\sum_{j=1}^{|\mathcal{C}|} \exp(\mathbf{w}_{ij}^T \mathbf{x}_n + b)}$  output:  $\mathbf{o} = \mathbf{T}^T \cdot \mathbf{p}$  ( $\mathbf{p}$  is prediction)

# More Efficient Human Intervention - 4

- [\[Luo+ 2017\]](#)
  - Training can be done on sentence level or bag level [\[Carboneau+ 2016\]](#)
  - How to train transition matrix w/o humans? Curriculum learning. [\[Bengio+ 2009\]](#)
    - $\text{trace}(T)$ : the larger (more similar to identity matrix) the lower the noise – regularize  $\text{trace}(T)$ .
    - Training: initially set  $\alpha, \beta = 1$  to learn  $p$  (prediction) only, then schedule to decrease  $\alpha, \beta$  to learn more about noise.

$$L = \sum_{i=1}^N -((1-\alpha)\log(o_i y_i) + \alpha\log(p_i y_i)) \\ - \beta \text{trace}(\mathbf{T}^i)$$

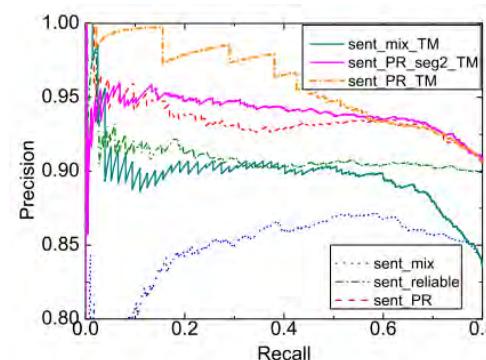


Figure 2: Sentence Level Results on TIMERE

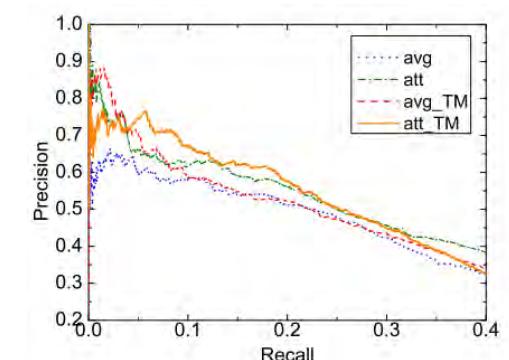


Figure 5: Results on ENTITYRE (bag level)  
63

# More Efficient Human Intervention - 5

- DS relation extraction from semi-structured web [Lockard+ 2018](#)
- Effective crowdsourcing [Chang+ 2017](#)
- More accessible ML tools [Yang+ 2018](#)

# High Maintainability - 1

- "High Interest Credit Card of Technical Debt" [\[Sculley+ 2014\]](#)
  - Complex Models Erode Boundaries
    - CACG (changing anything changes everything)
    - Hidden feedback loops
    - Undeclared customers
  - Data Dependencies Cost More than Code Dependencies
    - Unstable data dependencies
    - Underutilized data dependencies
    - Difficult to do static analysis of data dependencies
    - Danger in creating error-correction models
  - System-level spaghetti
    - Glue code
    - Pipeline Jungles
    - Dead experiment codepaths
    - Configuration debt
  - Dealing with changing world



# High Maintainability - 2

- Classifier error discovery through semantic data exploration [\[Chen+ 2018\]](#)

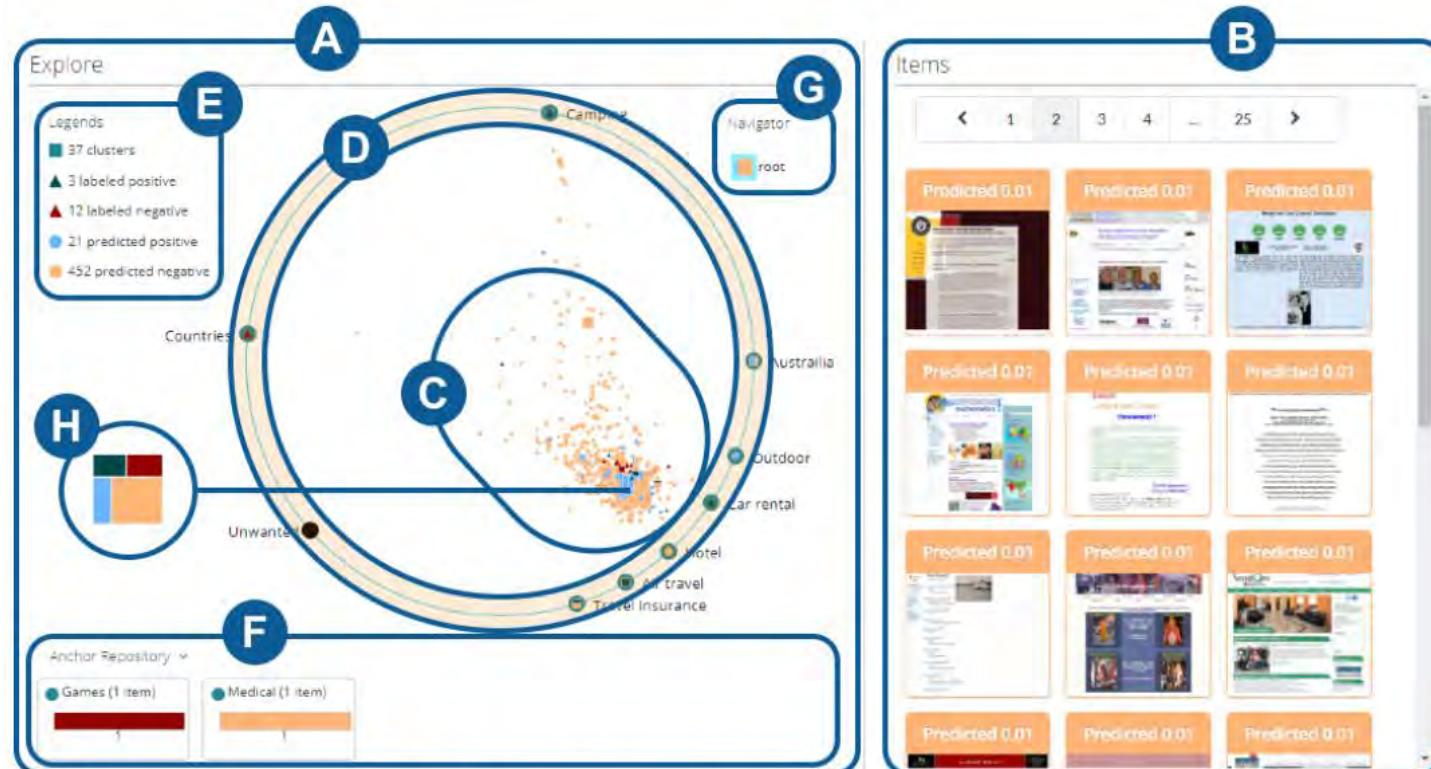


Figure 1. Overview of AnchorViz interface. The interface has Explore pane (A) that includes the visualization and Items pane (B) which shows a paginated grid of thumbnails of all currently visible items in the left pane. The visualization shows data points (C) within the outer circle (D) where anchors are positioned. The legend for data points (E) also acts as filters. Anchor repository (F) contains unused anchors. The navigator (G) shows which cluster the visualization is displaying in the current navigation stack. Clusters are represented as treemap-style squares (H).

# Summary

- Majority of the approaches still relies on textual data
- Providing constant stream of high-quality training data with minimal human intervention is still the key
- Knowledge verification and correction will become even more important
- Model and system maintainability requires a fresh take over the traditional ways of dealing with software engineering tasks

# Part III: Building Knowledge Graph

**Mohamed Yakout**

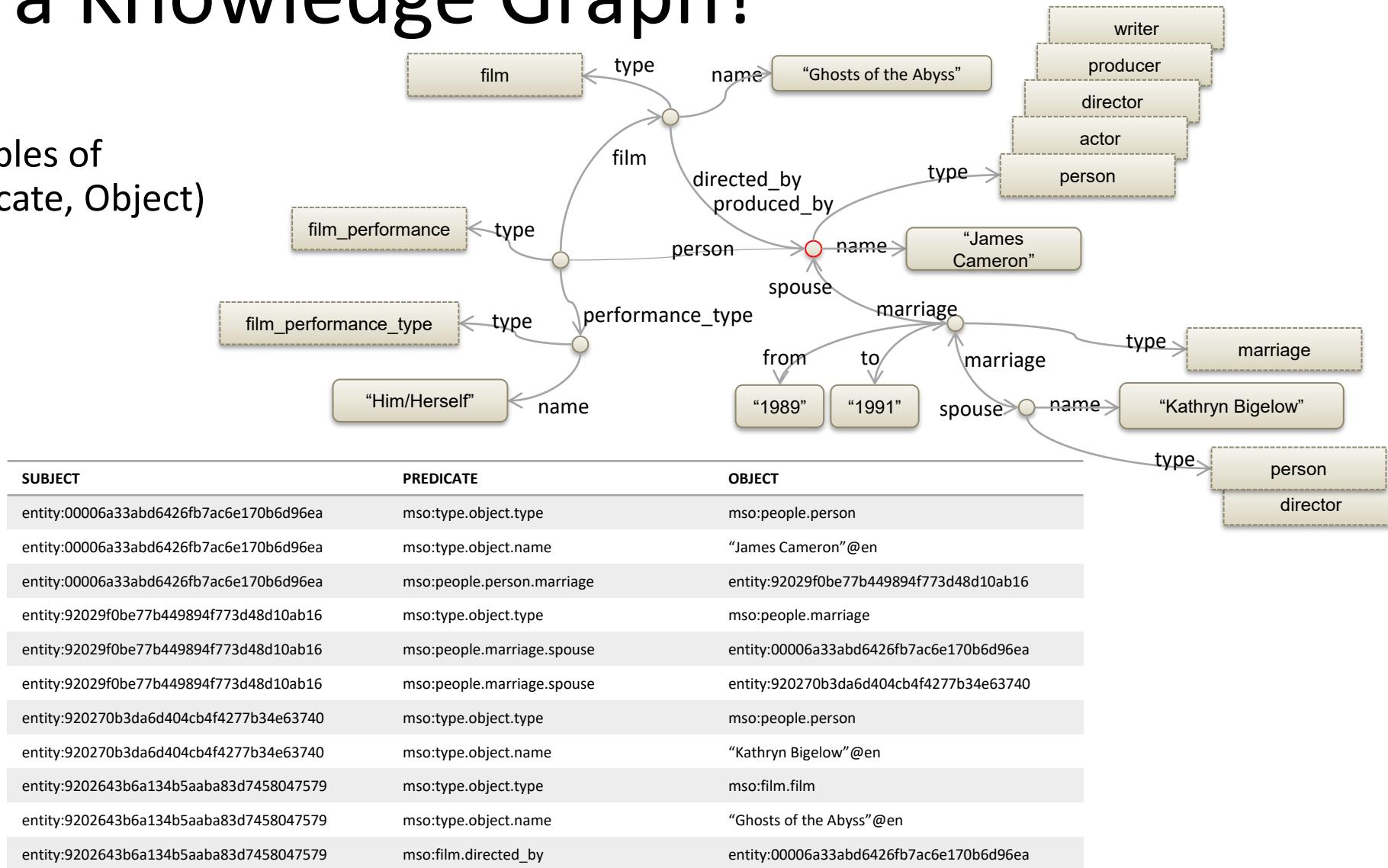
Principal Applied Science Manager, Satori Group, Microsoft AI+R

[myakout@microsoft.com](mailto:myakout@microsoft.com)



# What is a Knowledge Graph?

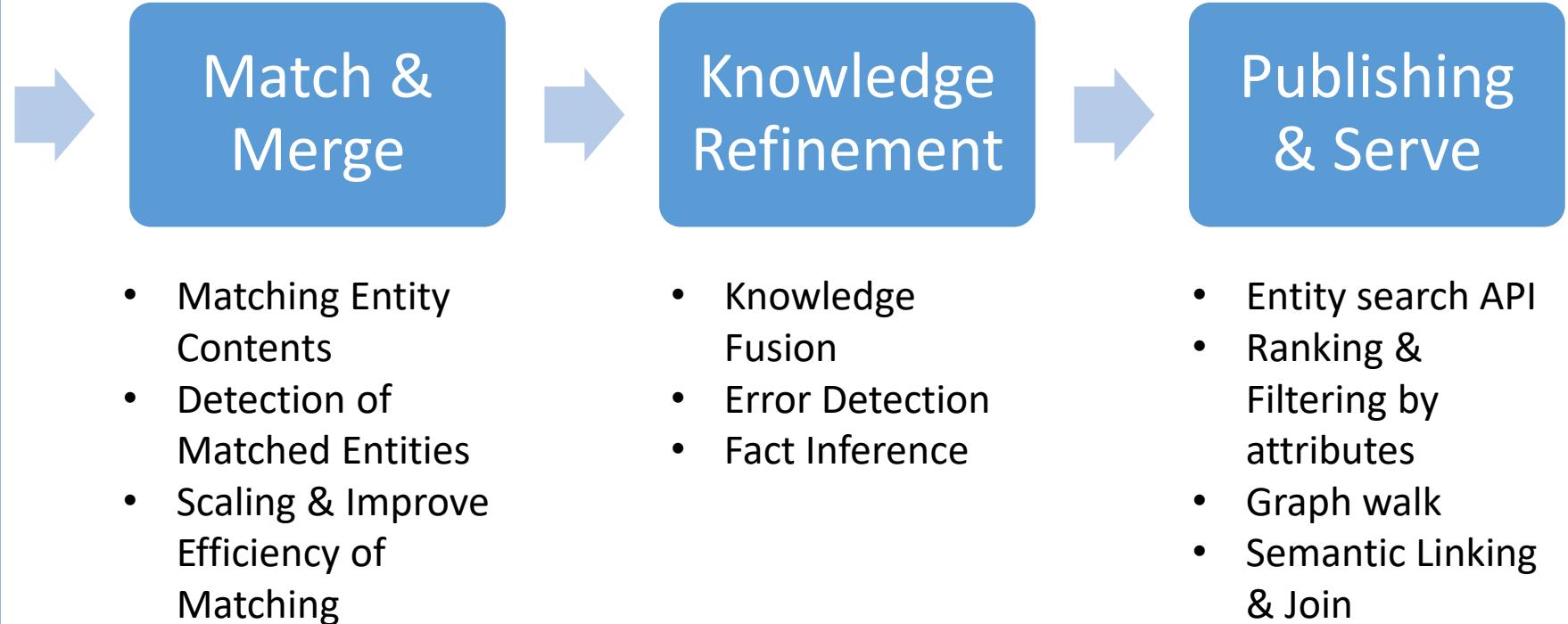
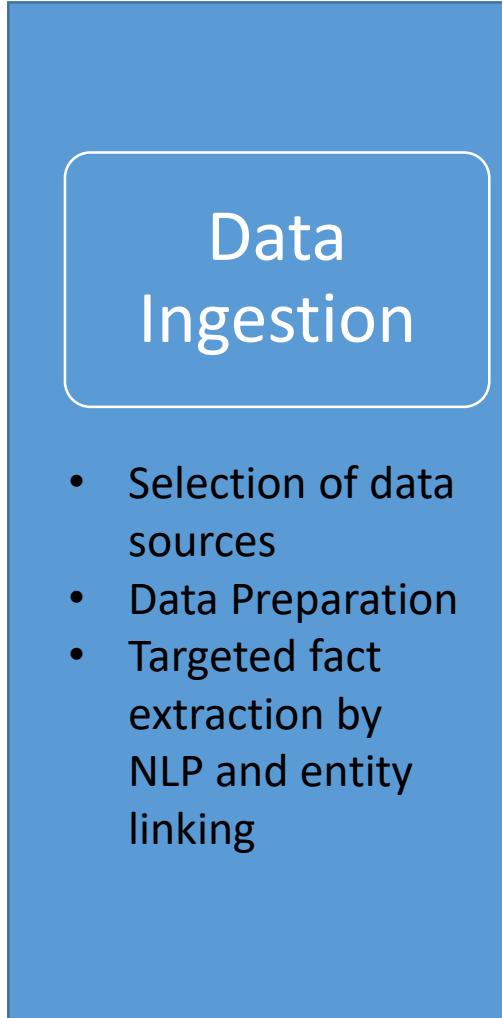
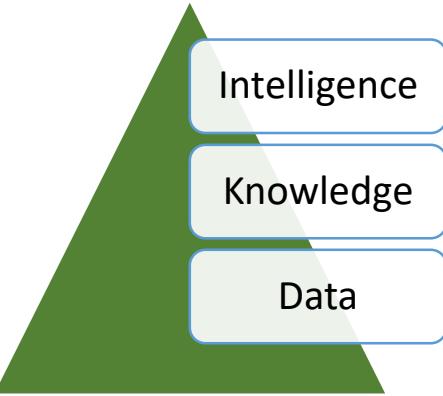
- Graph: RDF Triples of (Subject, Predicate, Object)



# Ontology Basics

- A complete, consistent, non-redundant, machine-readable representation of the world:
  - Allow data from various sources to be merged
  - Allow data to be shared across applications.
- Three elements: entities, properties, and types.
  - **Entities:** individuals, i.e. named objects in the world.
  - **Properties:** relationships between two entities or an entity and a **literal**, e.g. people.person.friends, people.person.employer, people.person.first\_name, time.event.start\_date, etc.
  - **Types:** sets or classes of entities:
    - **Primary entity types:** represent natural kinds or groupings, e.g. books, films, people, etc.
    - **Enumeration types:** Values that are standard but do not correspond to real objects in the world.
    - **Relationship types:** used to represent associations between more than two things, e.g. marriage (the people involved, when it started, where it began, etc.)

# Satori Graph Build

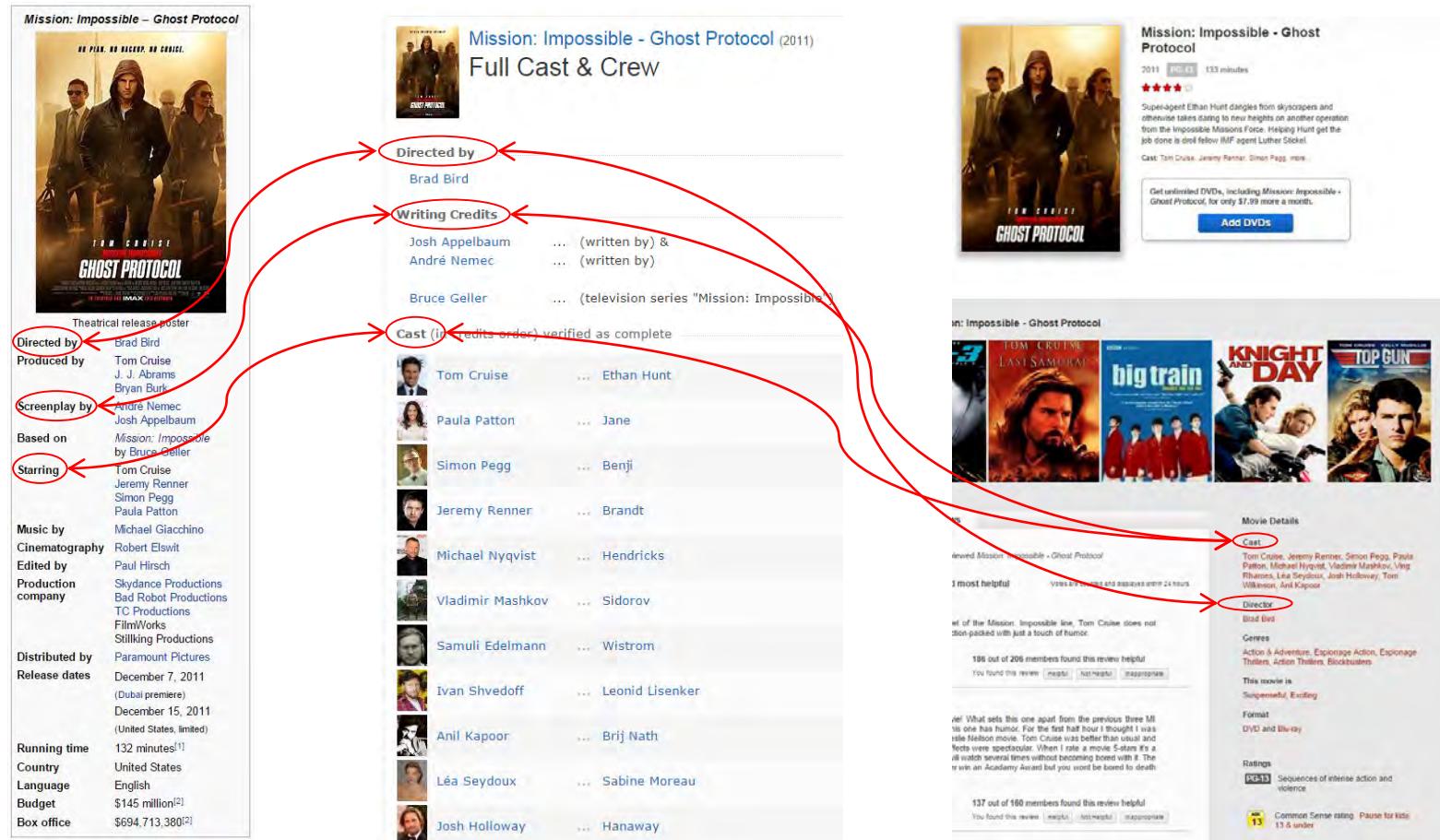
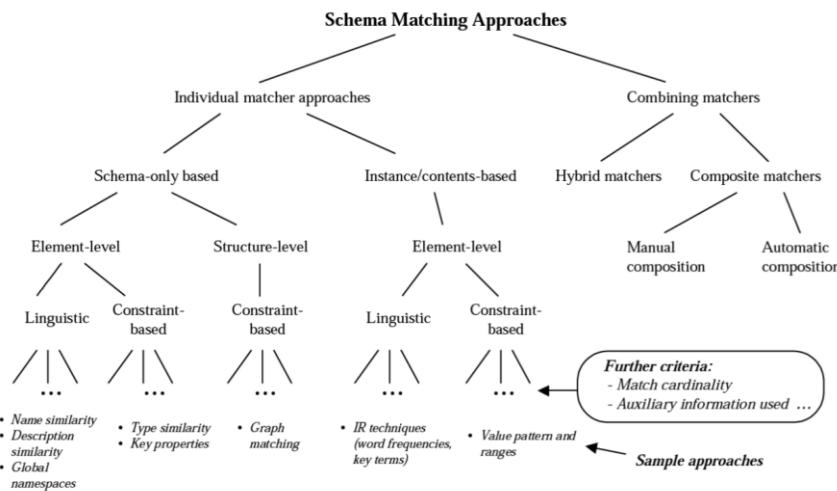


# Data Preparation

- Storing the data in a uniform manner.
  - **Parsing:** locate, identify and isolate data elements
  - **Data Transformation and Standardization:**
    - “44 West Fourth Street” or “44 West 4<sup>th</sup> St.”
    - 8 inches or 20 cm
    - July 28, 1999 or 07/28/1999 or 28/07/99
- Next, identify which fields to be compared.

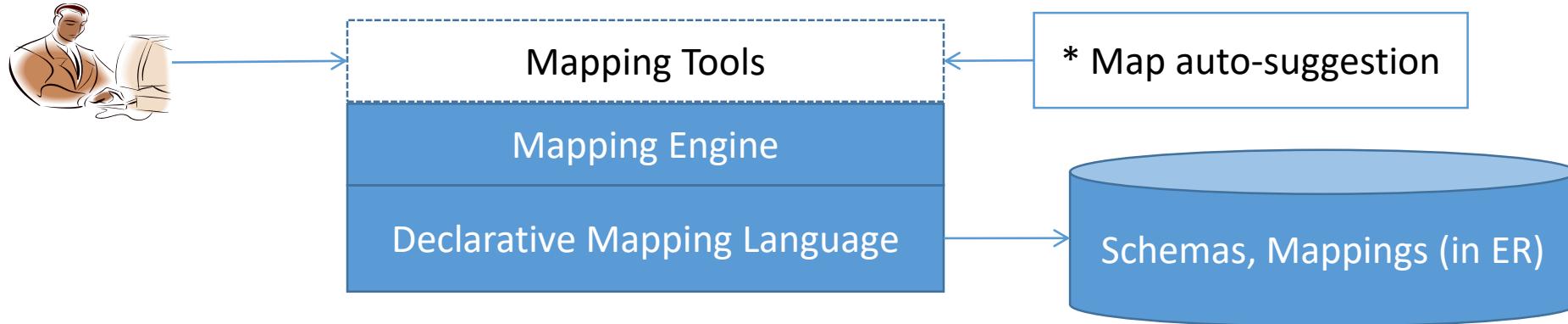
# Data Preparation

- Schema Matching
- Mapping to Microsoft Ontology



# Schema Mapping and Management

- Schema mapping: Declarative language, versioned and managed mappings, validation of mapping with schema change tracking

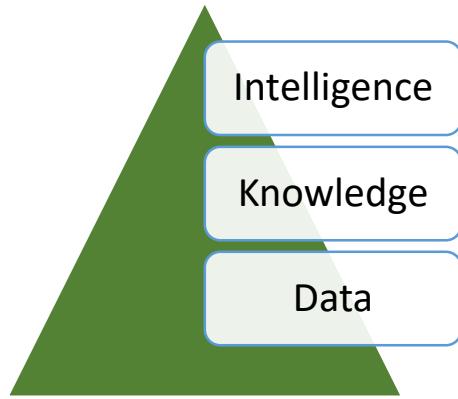
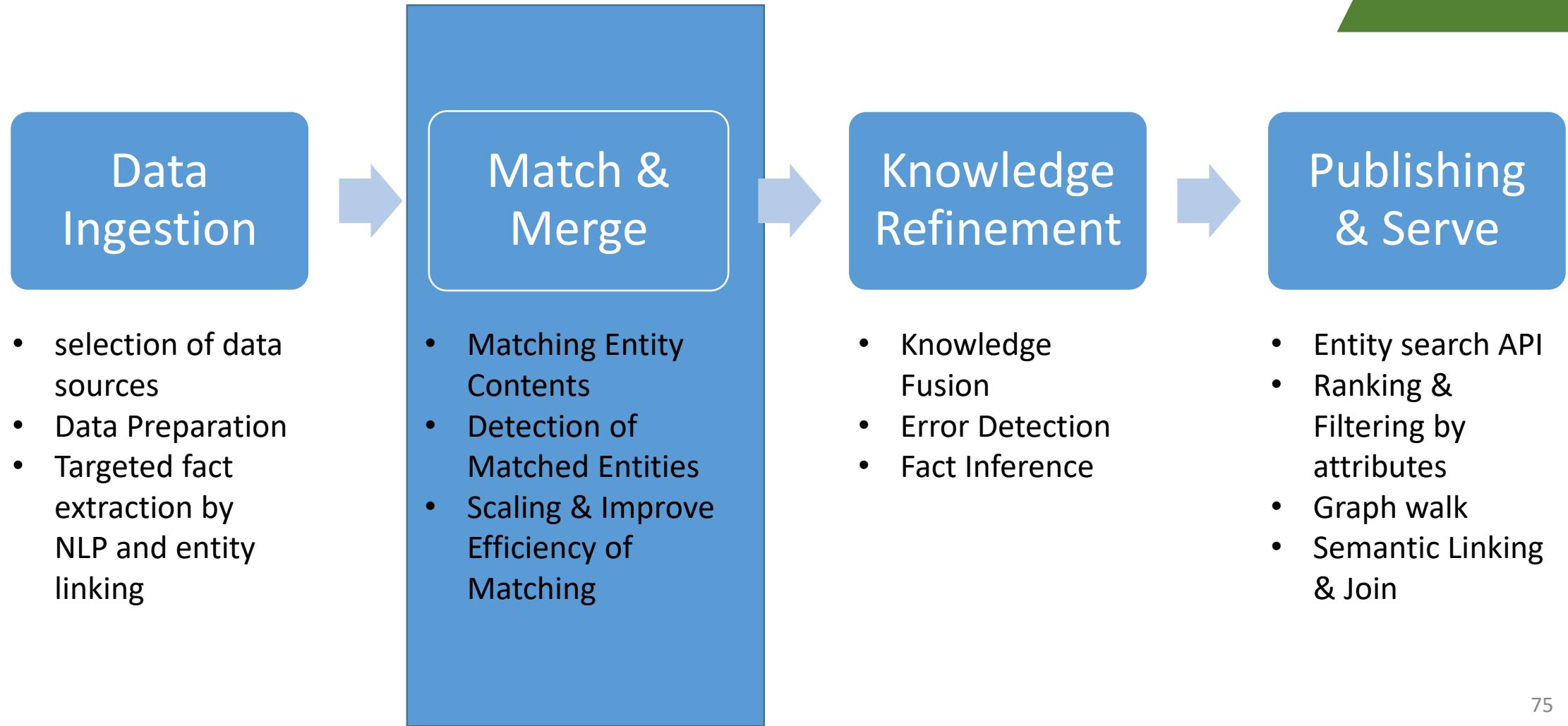


Example mapping for Music data to Satori ontology

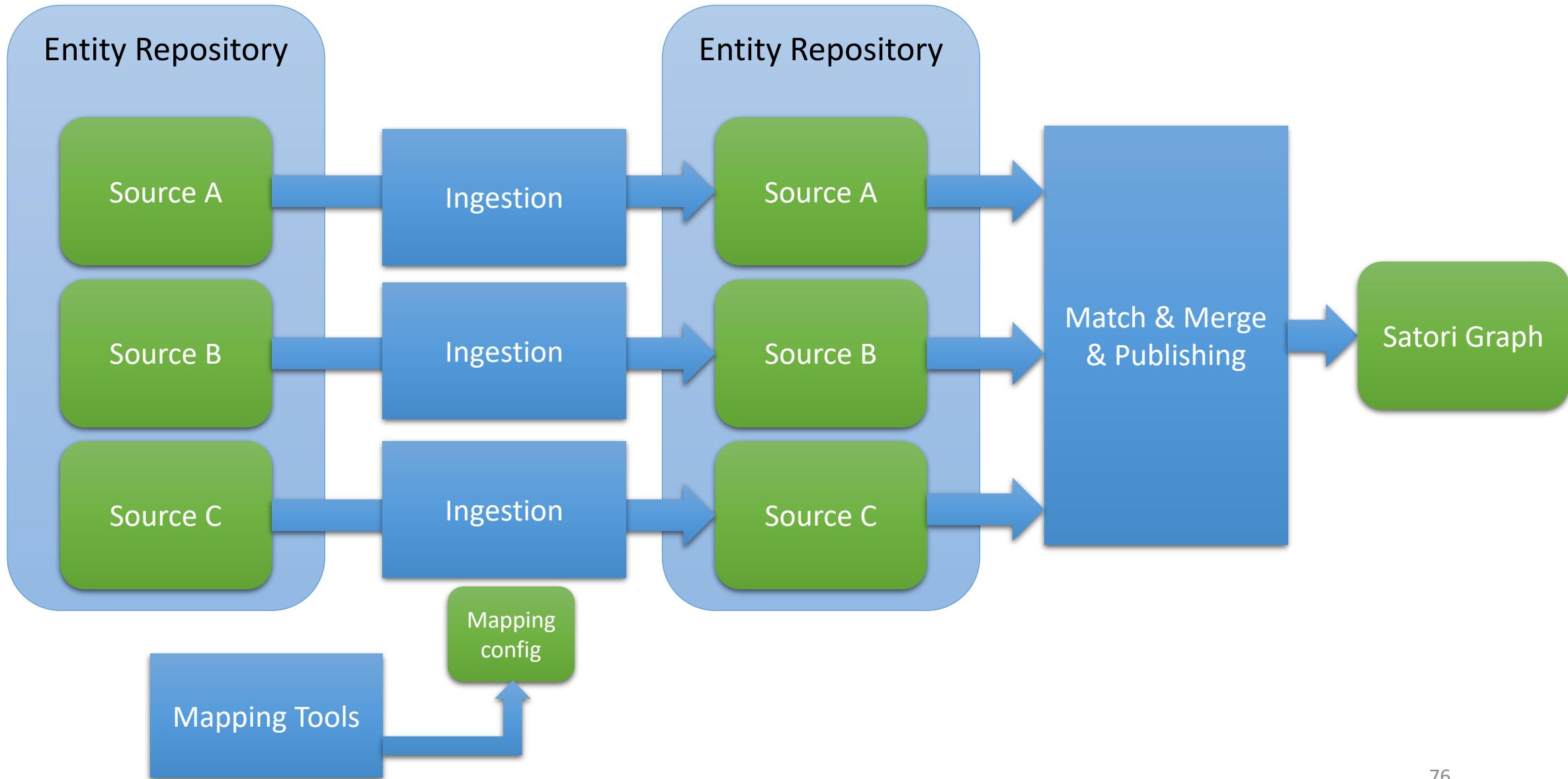
```
<ElementMap id='albumEntityPrimary.Album' elementName='Album' className='#Album@1.0'>
  <PropertyMaps>
    <ElementMap expression="'music.album'" elementName='type.object.type' />
    <ElementMap propertyPath='./Title' elementName='type.object.name' />
    <ElementMap propertyPath='./ID/ZuneMediaId' elementName='type.object.key' />
    <ElementMap propertyPath='./ReleaseDate' elementName='music.album.release_date' />
    <ElementMap propertyPath='./Label' elementName='music.album.record_label' />
    <ElementMap propertyPath='./Artists/Artist/Title' elementName='music.album.artist' />
    <ElementMap propertyPath='./Tracks/Track/Title' elementName='music.album.track' multiplicity='MultiValued' />
    <ElementMap expression="SUM(./Tracks/Track/DurationSeconds)" elementName='music.album.length' />
    <ElementMap propertyPath='./Genres/Genre/Genre' elementName='music.album.genre' multiplicity='MultiValued' />
  </PropertyMaps>
</ElementMap>
```

Closest approach in literature is Beaver: Jin et al, "Beaver: Towards a Declarative Schema Mapping" HILDA 2018

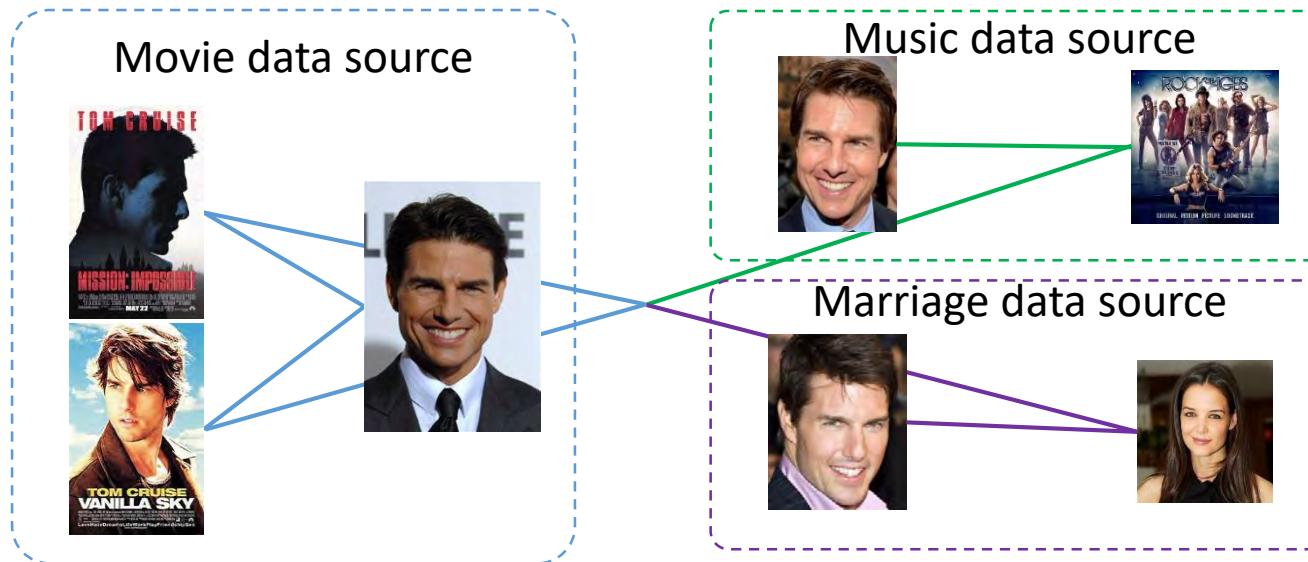
# Satori Graph Build



# Ingestion Flow



# Entity Matching



- **Well known problem:** Identify and discover instances referring to the same real-world entity.
- **Objective:**
  - Data Enrichment
  - Improve Data Quality by identifying and removing duplicates
  - Supporting fact correctness by merging duplicate facts from multiple sources
- **Synonyms:** Entity Linking, Entity Resolution, Reference Reconciliation, Deduplication, Match/Merge, Merge/Purge



## The Hobbit: An Unexpected Journey

12A · 2012 · 2 hr 49 min · Fantasy/Family

7.9/10  
IMDb

65%  
Rotten Tomatoes

[Share](#)

The adventure follows the journey of title character Bilbo Baggins, who is swept into an epic quest to reclaim the lost Dwarf Kingdom of Erebor from the fearsome dragon Smaug. Approached out of the blue by the wizard Gandalf the Grey, Bilbo finds himself joining a company of thirteen dwarves led by the legendary warrior, Thorin Oakenshield. Their jour... +

[IMDb](#)   [Wikipedia](#)   [Facebook](#)   [Official site](#)

Release date: 28 Nov 2012 (New Zealand)

Director: Peter Jackson

Gross revenue: \$1.021 billion USD

Films in series: The Hobbit: The Desolation of Smaug (Sequel) · The Hobbit: The Desolation of Smaug (Sequel) · The Hobbit: The Desolation of Smaug (Sequel)

Story by: J. R. R. Tolkien

Screenwriters: Peter Jackson · Fran Walsh · Philippa Boyens · Guillermo del Toro

### Cast



### Critic reviews

The Hobbit plays younger and lighter than Fellowship and its follow-ups, but does right by the faithful and has a strength in Martin Freeman's Bilbo that may yet see this trilogy measure up to the last one... [Full review](#)

[Empire](#) by Dan Jolin



## The Hobbit: An Unexpected Journey

From Wikipedia, the free encyclopedia

***The Hobbit: An Unexpected Journey*** is a 2012 epic fantasy adventure film directed by Peter Jackson. It is the first installment in a three-part film adaptation based on the 1937 novel *The Hobbit* by J. R. R. Tolkien. It is followed by *The Desolation of Smaug* (2013) and *The Battle of the Five Armies* (2014), and together they act as a prequel to Jackson's *The Lord of the Rings* film trilogy. The film's screenplay was written by Peter Jackson, his longtime collaborators Fran Walsh and Philippa Boyens, and Guillermo del Toro, who was originally chosen to direct the film before leaving the project in 2010.

The story is set in Middle-earth sixty years before the events of *The Lord of the Rings*, and portions of the film are adapted from the appendices to Tolkien's *The Return of the King*.<sup>[6]</sup> *An Unexpected Journey* tells the tale of Bilbo Baggins (Martin Freeman), who is convinced by the wizard Gandalf the Grey (Ian McKellen) to accompany thirteen Dwarves, led by Thorin Oakenshield (Richard Armitage), on a quest to reclaim the Lonely Mountain from the dragon Smaug. The ensemble cast also includes James Nesbitt, Ken Stott, Cate Blanchett, Ian Holm, Christopher Lee, Hugo Weaving, Elijah Wood and Andy Serkis, and features Sylvester McCoy, Barry Humphries and Manu Bennett.

*An Unexpected Journey* premiered on November 28, 2012 in New Zealand and was released internationally on December 12, 2012.<sup>[7]</sup> The film has grossed over \$1 billion at the box office, surpassing both *The Fellowship of the Ring* and *The Two Towers* nominally, becoming the fourth highest-grossing film of 2012 and the 18th highest grossing film of all time. The film was nominated for three Academy Awards for Best Visual Effects, Best Production Design, and Best Makeup and Hairstyling.<sup>[8]</sup> It was also nominated for three BAFTA Awards.<sup>[9]</sup>

### Contents [hide]

- 1 Plot
- 2 Cast
- 3 Production
  - 3.1 High frame rate
  - 3.2 Score
- 4 Distribution
  - 4.1 Marketing
    - 4.1.1 Video games
    - 4.2 Theatrical release
    - 4.3 Home media
- 5 Reception
  - 5.1 Box office
  - 5.2 Critical response
- 6 Accolades

### The Hobbit: An Unexpected Journey

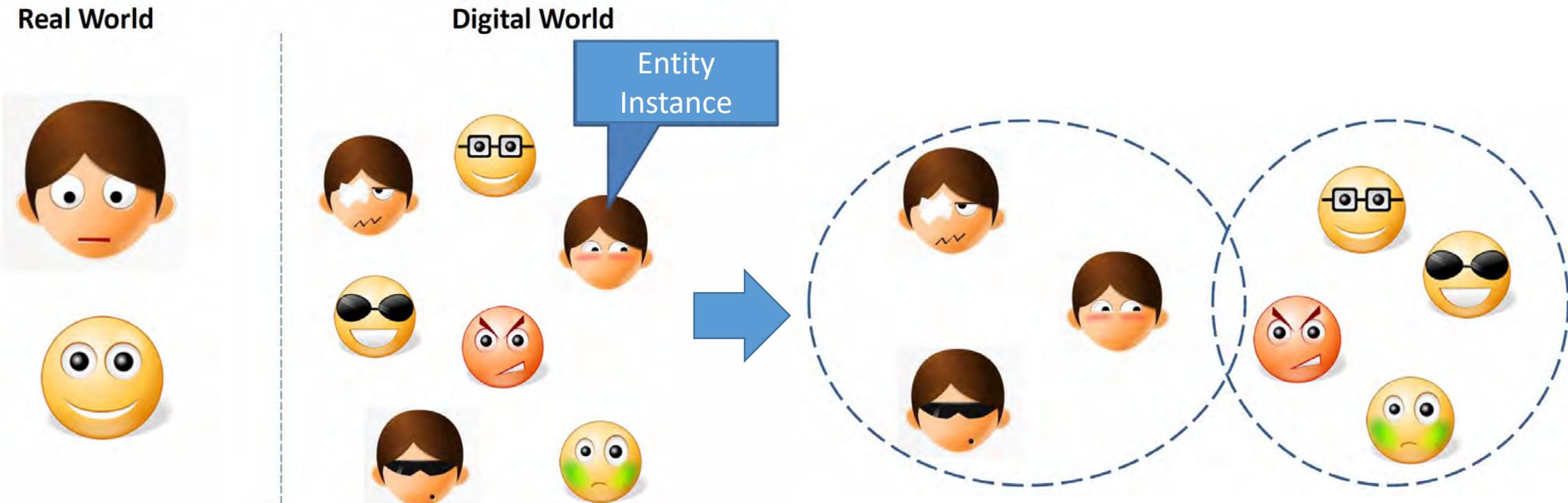


Directed by Peter Jackson  
Produced by Carolynne Cunningham  
Zane Weiner  
Fran Walsh  
Peter Jackson  
Screenplay by Fran Walsh  
Philippa Boyens  
Peter Jackson  
Guillermo del Toro  
Based on *The Hobbit* by J. R. R. Tolkien  
Starring Ian McKellen  
Martin Freeman  
Richard Armitage

# Entity Matching References

- Book / Survey Articles
  - Data Quality and Record Linkage Techniques [T. Herzog, F. Scheuren Scheuren, W. Winkler Winkler, Springer Springer, '07]
  - Duplicate Record Detection [A. Elmagard, P. Ipeirotis, V. Verykios, TKDE '07]
  - An Introduction to Duplicate Detection [F. Naumann, M. Herschel, M&P synthesis lectures 2010]
  - Evaluation of Entity Resolution Approached on Real-world Match Problems [H. Kopke, A. Thor, E. Rahm, PVLDB 2010]
  - A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication [P. Christen TKDE '11]
  - Data Matching [P. Christen, Springer 2012]
- Tutorials
  - Record Linkage: Similarity measures and Algorithms [N. Koudas, S. Sarawagi, D. Srivatsava SIGMOD '06]
  - Data fusion--Resolving data conflicts for integration [X. Dong, F. Naumann VLDB '09]
  - Entity Resolution: Resolution: Theory, Practice Practice and Open Challenges Challenges [L. Getoor, A. Machanavajjhala VLDB '12]
  - Entity Resolution in the Web of Data: Tutorial [Kostas Stefanidis CIKM 2013]
- Systems
  - SecondString, <http://secondstring.sourceforge.net/>
  - Simmetrics: <http://sourceforge.net/projects/simmetrics/>
  - LingPipe, <http://alias-i.com/lingpipe/index.html>

# Data Quality Challenge

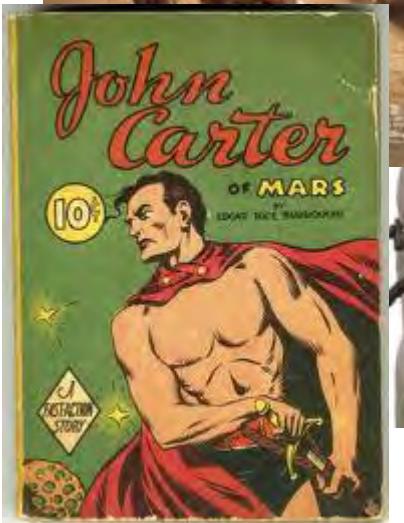


Missing Data  
Data error due to IE tech or human errors  
Abbreviations and truncation

# Open Domain Entity Matching (Disambiguation Challenge)



**John Carter** 3rd  
**Quantitative Headhunter**  
Greater New York City Area | Fi



**John Carter** 3rd  
**Head of Own Brand and Prod**  
**Cash and Carry**  
Munich Area, Germany | Consu



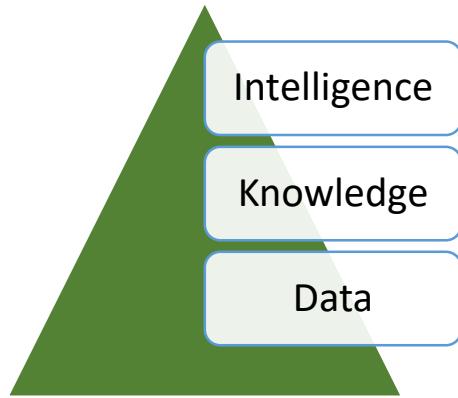
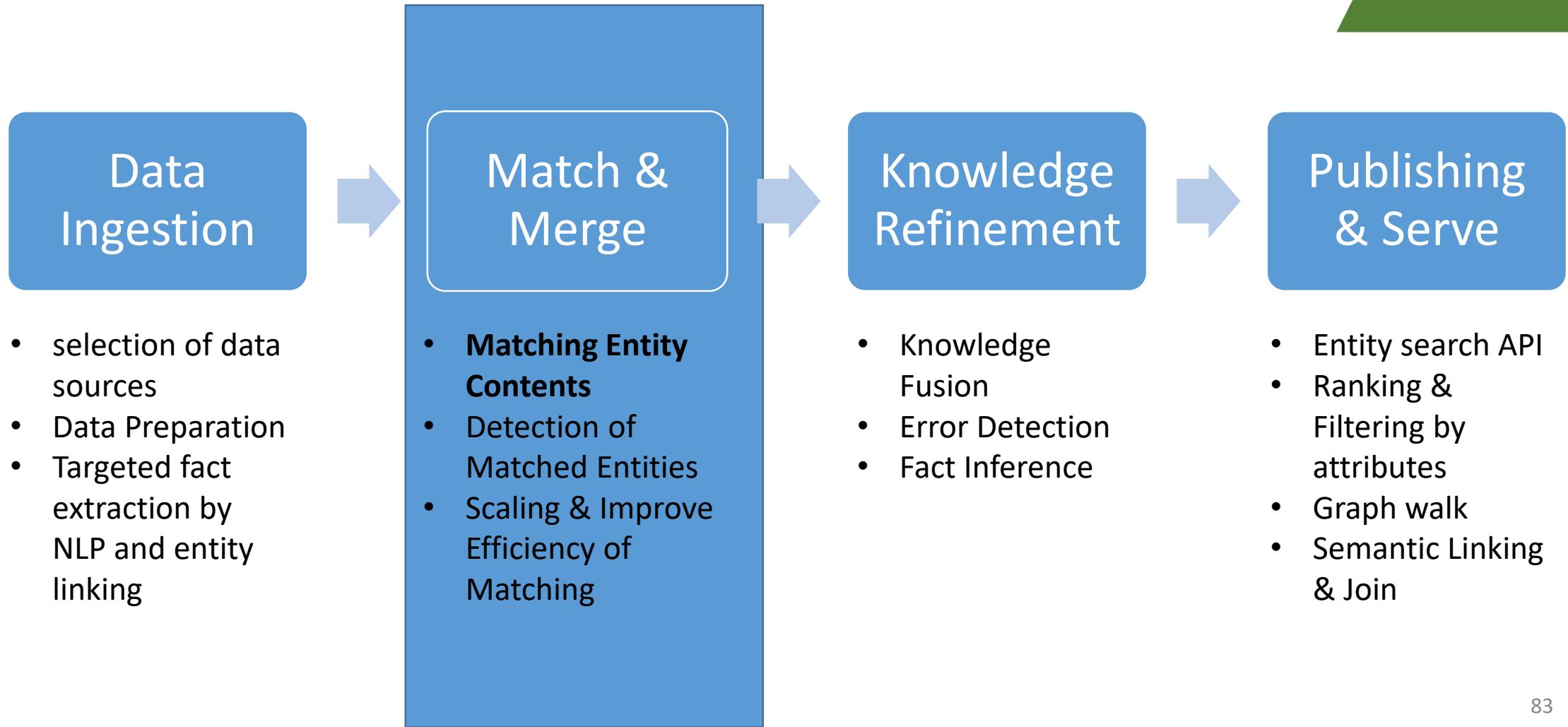
**John Carter** 3rd  
**Recruiter at STERIS Corporat**  
Cleveland/Akron, Ohio Area | M

**JOHN H. CARTER CO., INC.**  
"An Emerson Process Management Representative"

# EM Big Data Challenge

- **Larger Datasets:** Need Faster, Efficient, Parallel techniques.
- **Multi-Domain:** Need different matching methods and a technique to manage executions within and across domains
- **Linked, Connected and Relational data:** Need techniques to leverage the diversity of connections and representation.

# Satori Graph Build



# Matching Entity Contents

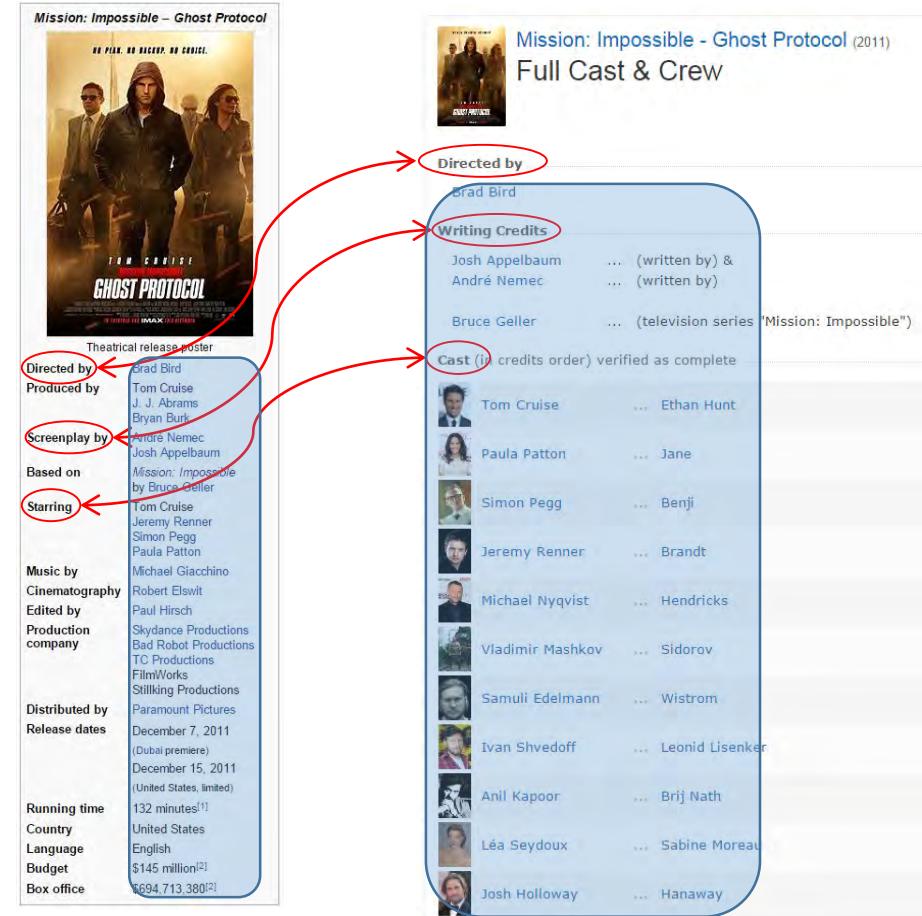
- Matching Functions

- Generic Functions

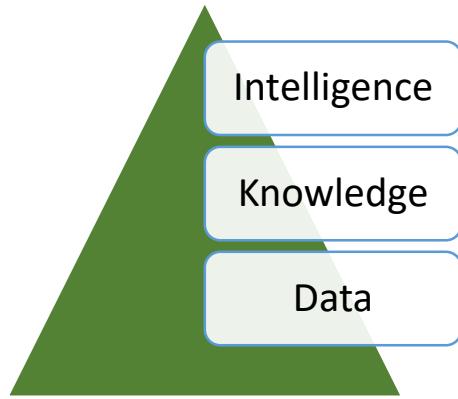
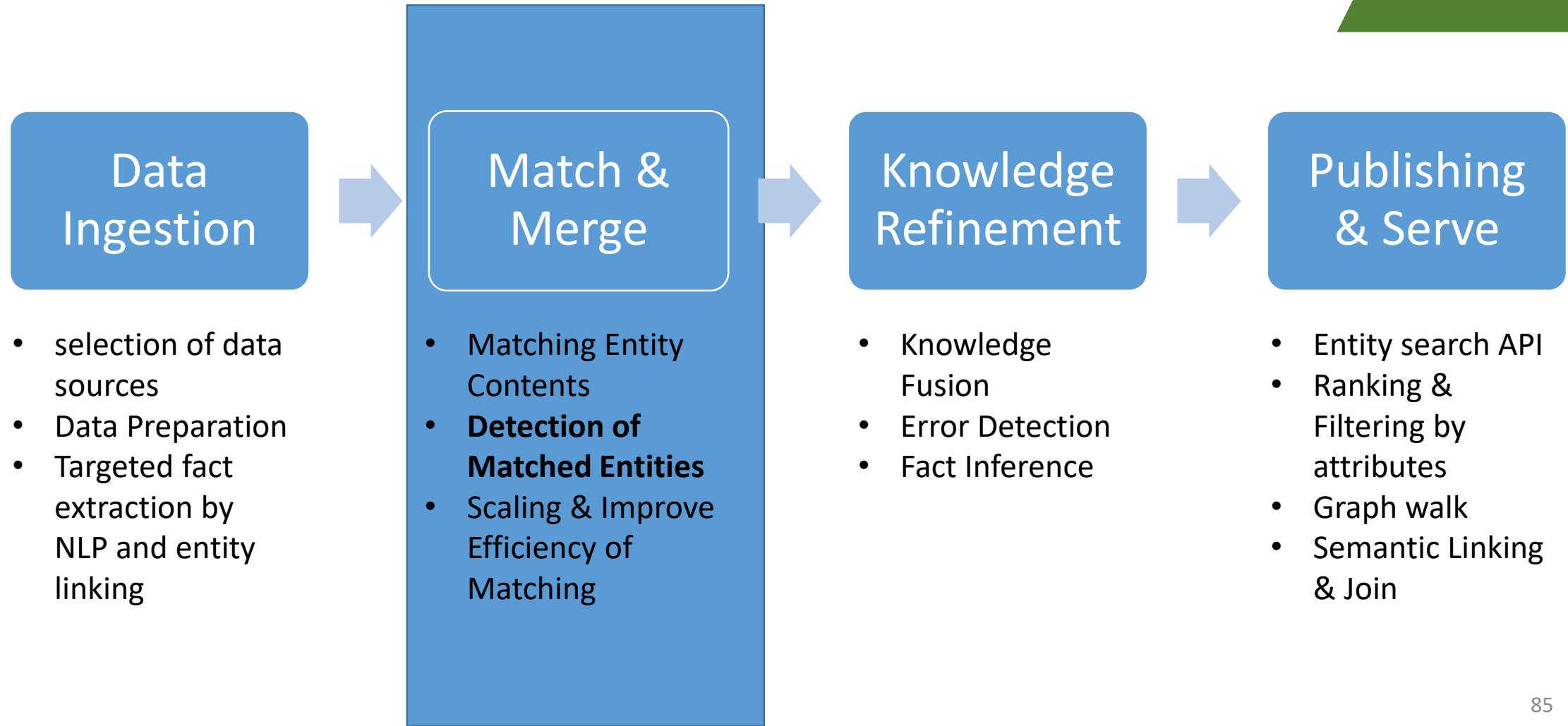
- Character Based Functions
  - Token Based Functions
  - Phonetic Based Functions
  - Transformation Rule Based Matching Functions
  - Value-Set Matching Functions

- Specific Functions

- Numeric Matching Functions (Numbers, Dates, ... etc)
  - Special Matching Functions (Zip codes, Phone Numbers, Address ... etc)



# Satori Graph Build



# Detection of Matched Entities

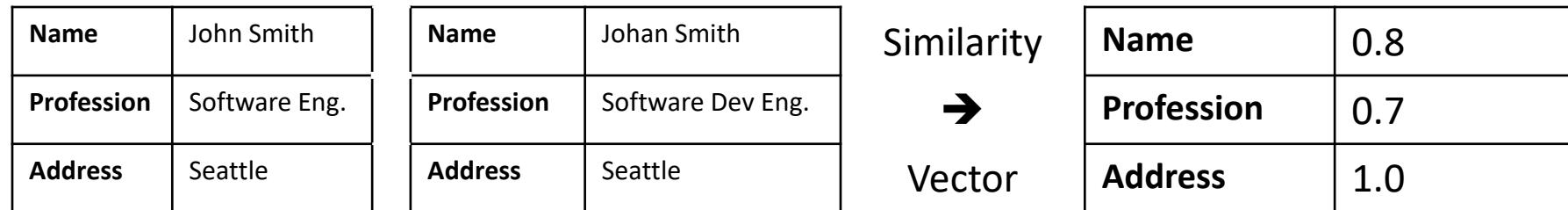
- Probabilistic Matching Models
  - Supervised and Semi-supervised Learning
  - Unsupervised learning
  - Active Learning Based
- Distance Based
  - Threshold
  - Neighborhood exploration
- Declarative Matching Rules and Constraints
  - Disjunction of conjunction
  - Constraint base clustering
- Collective Resolution in Linked Data
  - Similarity signals propagation
  - Entity similarity based on connections

# Detection of Matched Entities

- **Probabilistic Matching Models**
  - Supervised and Semi-supervised Learning
  - Unsupervised learning
  - Active Learning Based
- Distance Based
  - Threshold
  - Neighborhood exploration
- Declarative Matching Rules and Constraints
  - Disjunction of conjunction
  - Constraint base clustering
- Collective Resolution in Linked Data
  - Similarity signals propagation
  - Entity similarity based on connections

# Detection of Matched Entities

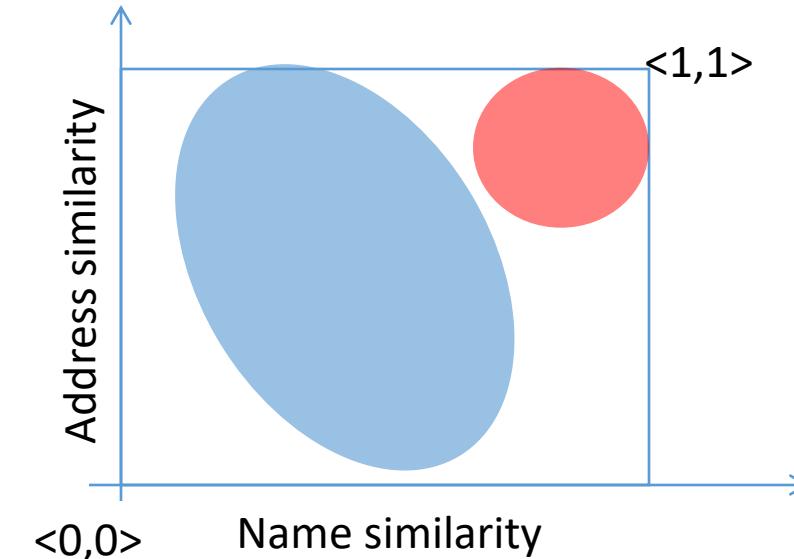
- Compute similarity vector
- Classify the vectors as Match and UnMatch.



The similarity vector  $<0.8, 0.7, 1.0>$

# Detection of Matched Entities: Probabilistic Matching Models

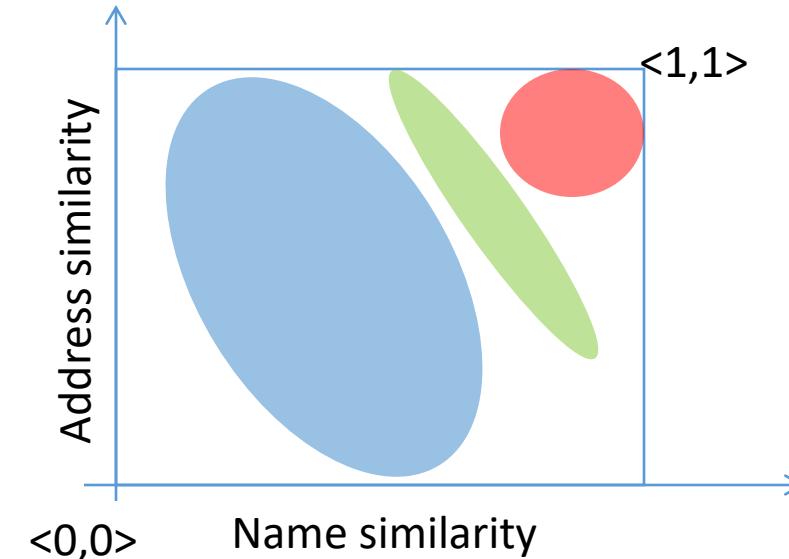
- Supervised and Semi-supervised Learning
  - Map the similarity vector to two classes (M, U)



• .

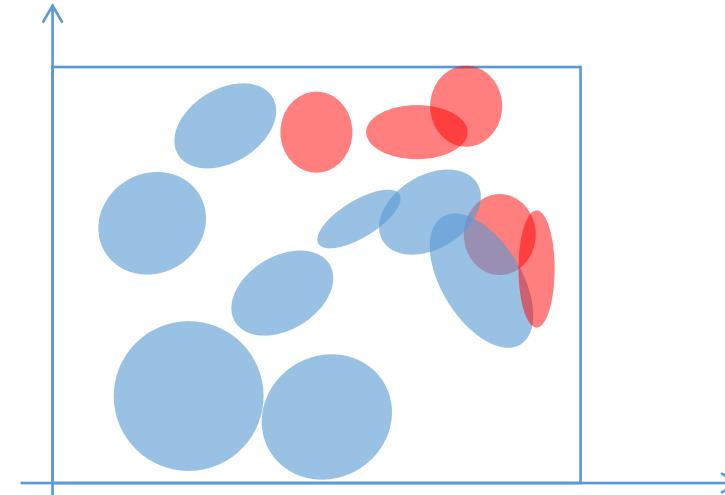
# Detection of Matched Entities: Probabilistic Matching Models

- Supervised and Semi-supervised Learning
  - Map the similarity vector to two classes (M, U)
  - Later on a rejection or uncertain rejoin is considered (M, R, U)
- Rely on the existence of training data, pair of records pre-labeled match or not. Do we have that??!



# Detection of Matched Entities: Probabilistic Matching Models

- Pairs Sampling for training
  - Random Sample
    - Most of space contains non-matched pairs
  - Sample from blocks
    - Apply blocking
    - Random Sample a set of blocks
    - Get pairs from the randomly sampled blocks
  - Stratified Sample
    - Cluster the similarity vectors
    - Sample from clusters



# Detection of Matched Entities: Probabilistic Matching Models

- Active Learning
  - Train an initial ML model by an initial small sample
  - While (user is not happy with predictions)
    - Foreach Pair  $p$  in all pairs
      - Apply the model on  $p$
      - Get the prediction probability and compute uncertainty
    - Sort all pairs based on uncertainty
    - Display pairs with the highest uncertainty first to user for labeling
    - Re-train the model.

# Detection of Matched Entities: Probabilistic Matching Models

- Active Learning

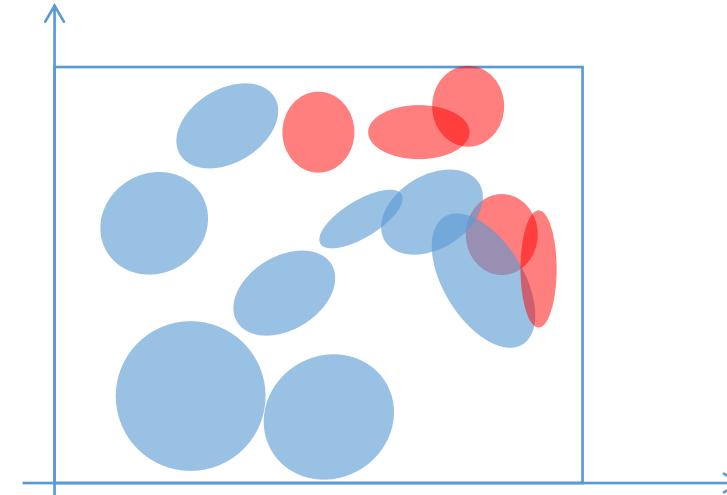
- Train an initial ML model by an **initial small sample**
- While (user is not happy with predictions)
  - **Foreach Pair  $p$  in all pairs**
    - Apply the model on  $p$
    - Get the prediction probability and compute uncertainty
  - Sort all pairs based on uncertainty
  - Display pairs with the highest uncertainty first to user for labeling
  - Re-train the model.

How to get a good initial sample? The initial model will be biased and we may not see a lot of cases during interaction because of the initial model

We cannot afford doing that with millions of pairs in an online interactive system

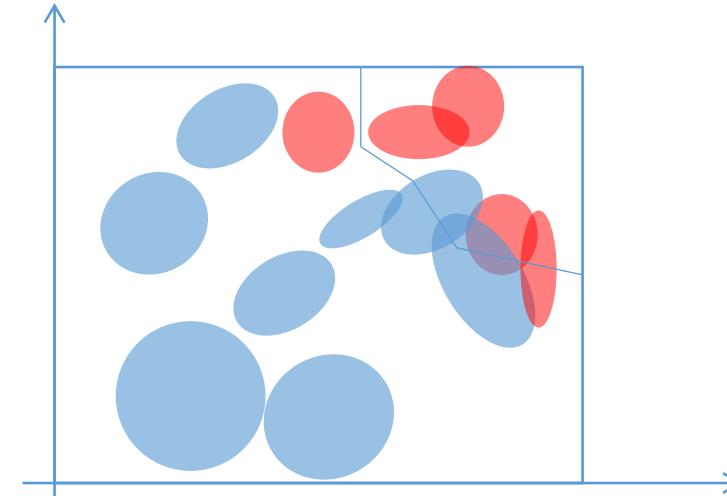
# Detection of Matched Entities: Probabilistic Matching Models

- **Effective Active Learning for Entity Matching**
  - Better control on the space of similarities.
  - Clustering for all vectors
  - Offline sample from clusters host locally
  - **Active Learning Guided by the Clusters** through:
    - Focus on clusters with high uncertainty
    - Cover clusters with less training samples
    - From a cluster, sampling positive uncertain cases improves precision
    - From a cluster, sampling negative uncertain cases improves recall.
  - Uncertainty can be computed from the entropy of model's probability.



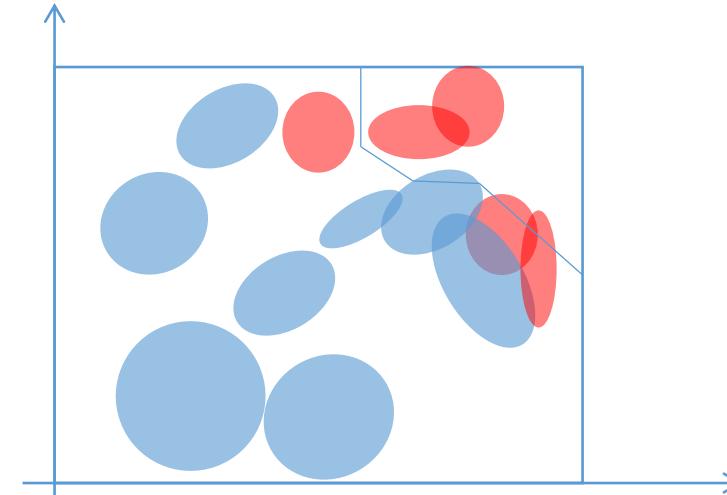
# Detection of Matched Entities: Probabilistic Matching Models

- **Effective Active Learning for Entity Matching**
  - Better control on the space of similarities.
  - Clustering for all vectors
  - Offline sample from clusters host locally
  - **Active Learning Guided by the Clusters** through:
    - Focus on clusters with high uncertainty
    - Cover clusters with less training samples
    - From a cluster, sampling positive uncertain cases improves precision
    - From a cluster, sampling negative uncertain cases improves recall.
  - Uncertainty can be computed from the entropy of model's probability.



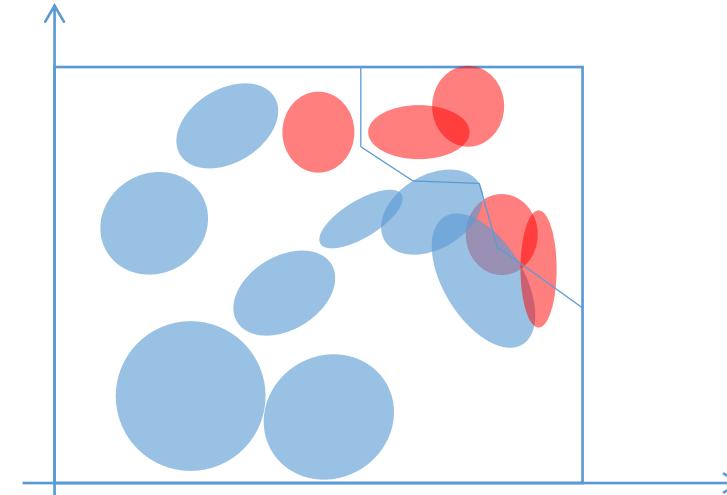
# Detection of Matched Entities: Probabilistic Matching Models

- **Effective Active Learning for Entity Matching**
  - Better control on the space of similarities.
  - Clustering for all vectors
  - Offline sample from clusters host locally
  - **Active Learning Guided by the Clusters** through:
    - Focus on clusters with high uncertainty
    - Cover clusters with less training samples
    - From a cluster, sampling positive uncertain cases improves precision
    - From a cluster, sampling negative uncertain cases improves recall.
  - Uncertainty can be computed from the entropy of model's probability.



# Detection of Matched Entities: Probabilistic Matching Models

- **Effective Active Learning for Entity Matching**
  - Better control on the space of similarities.
  - Clustering for all vectors
  - Offline sample from clusters host locally
  - **Active Learning Guided by the Clusters** through:
    - Focus on clusters with high uncertainty
    - Cover clusters with less training samples
    - From a cluster, sampling positive uncertain cases improves precision
    - From a cluster, sampling negative uncertain cases improves recall.
  - Uncertainty can be computed from the entropy of model's probability.



# Detection of Matched Entities

- Probabilistic Matching Models
  - Supervised and Semi-supervised Learning
  - Unsupervised learning
  - Active Learning Based
- **Distance Based**
  - Threshold
  - Neighborhood exploration
- Declarative Matching Rules and Constraints
  - Disjunction of conjunction
  - Constraint base clustering
- Collective Resolution in Linked Data
  - Similarity signals propagation
  - Entity similarity based on connections

# Detection of Matched Entities: Distance Based

- Threshold
  - If  $w_1 sim(name_1, name_2) + w_2 sim(address_1, address_2) > t$  .  
Then it a match.



- Neighborhood exploration
  - Matches are “closer” to each other than to others
    - A “**Compact Set**” criteria
  - The local neighborhood of matched entities is sparse
    - A “**Sparse Neighborhood**” criteria
  - Requires an overall matching or distance function for two entities



# Detection of Matched Entities

- Probabilistic Matching Models
  - Supervised and Semi-supervised Learning
  - Unsupervised learning
  - Active Learning Based
- Distance Based
  - Threshold
  - Neighborhood exploration
- **Declarative Matching Rules and Constraints**
  - Disjunction of conjunction
  - Constraint base clustering
- Collective Resolution in Linked Data
  - Similarity signals propagation
  - Entity similarity based on connections

# Detection of Matched Entities: Matching Rules and Constraints

- Disjunction of Conjunction (Simple)
  - **Match(movie\_name) AND Match(release\_date)**  
**OR Match(movie\_name) AND Match(director) → Match**
- Constraints based clustering and matching (e.g., Dedupalog)
  - Encoding of rules and constraints and then cluster entities to **satisfy hard constraints and minimize soft rules violations**. Example:
    - No researcher has published more than five AAAI papers in a year
    - If two citations match, then their authors will be matched in order
    - Papers with similar titles should likely be clustered together”
  - The framework is domain independent. **But how realistic is this to compile these rules?**

# Detection of Matched Entities

- Probabilistic Matching Models
  - Supervised and Semi-supervised Learning
  - Unsupervised learning
  - Active Learning Based
- Distance Based
  - Threshold
  - Neighborhood exploration
- Declarative Matching Rules and Constraints
  - Disjunction of conjunction
  - Constraint base clustering
- **Collective Resolution in Linked Data**
  - Similarity signals propagation
  - Entity similarity based on connections

# Detection of Matched Entities: Collective Resolution in Linked Data

Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist

Watch Interstellar, Vikings and more on Prime Video Start your 30-day free trial

FULL CAST AND CREW TRIVIA USER REVIEWS IMDbPro MORE SHARE

+ Mission: Impossible III (2006) ★ 6.9 271,578 Rate This

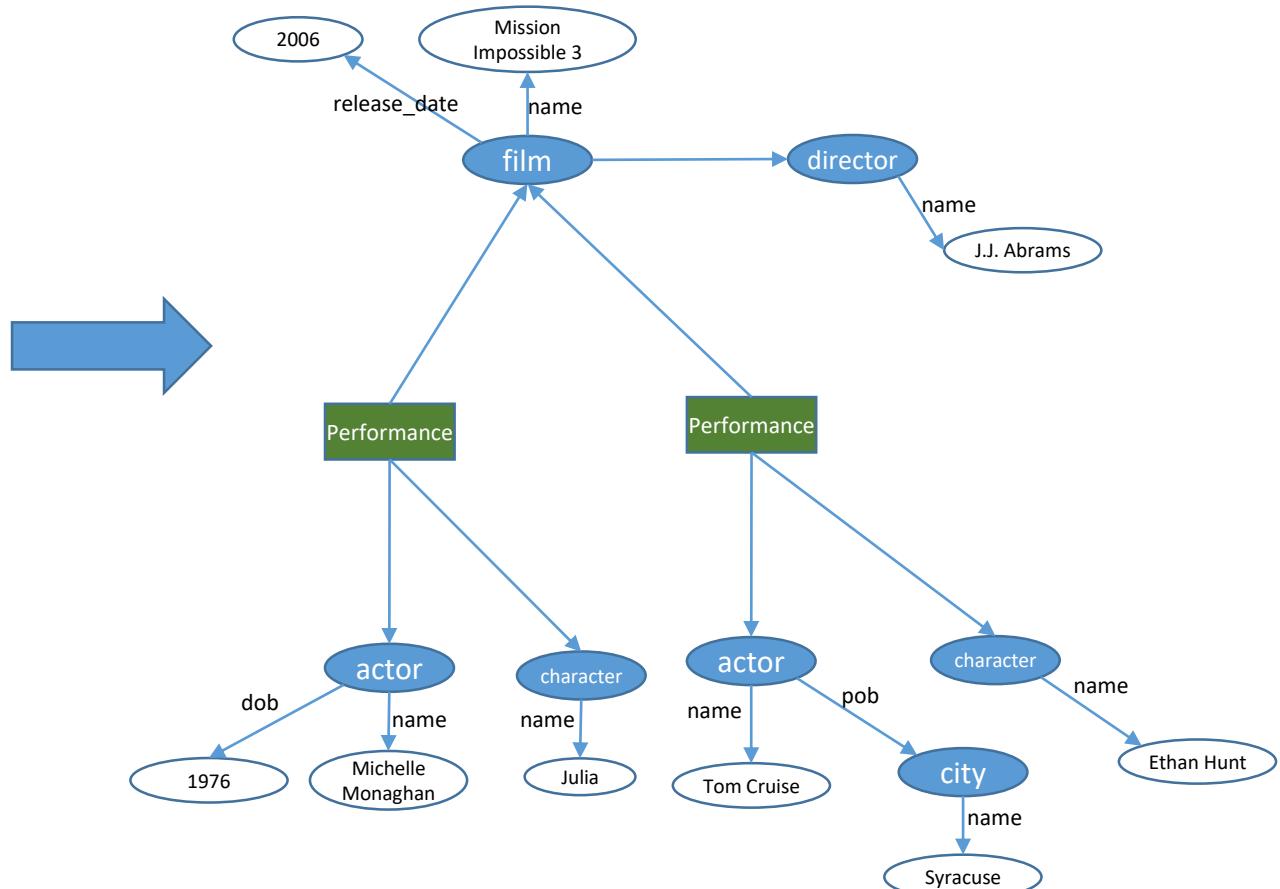
PG-13 | 2h 6min | Action, Adventure, Thriller | 5 May 2006 (USA)

Agent Ethan Hunt comes into conflict with a dangerous and sadistic arms dealer who threatens his life and his fianceé in response .

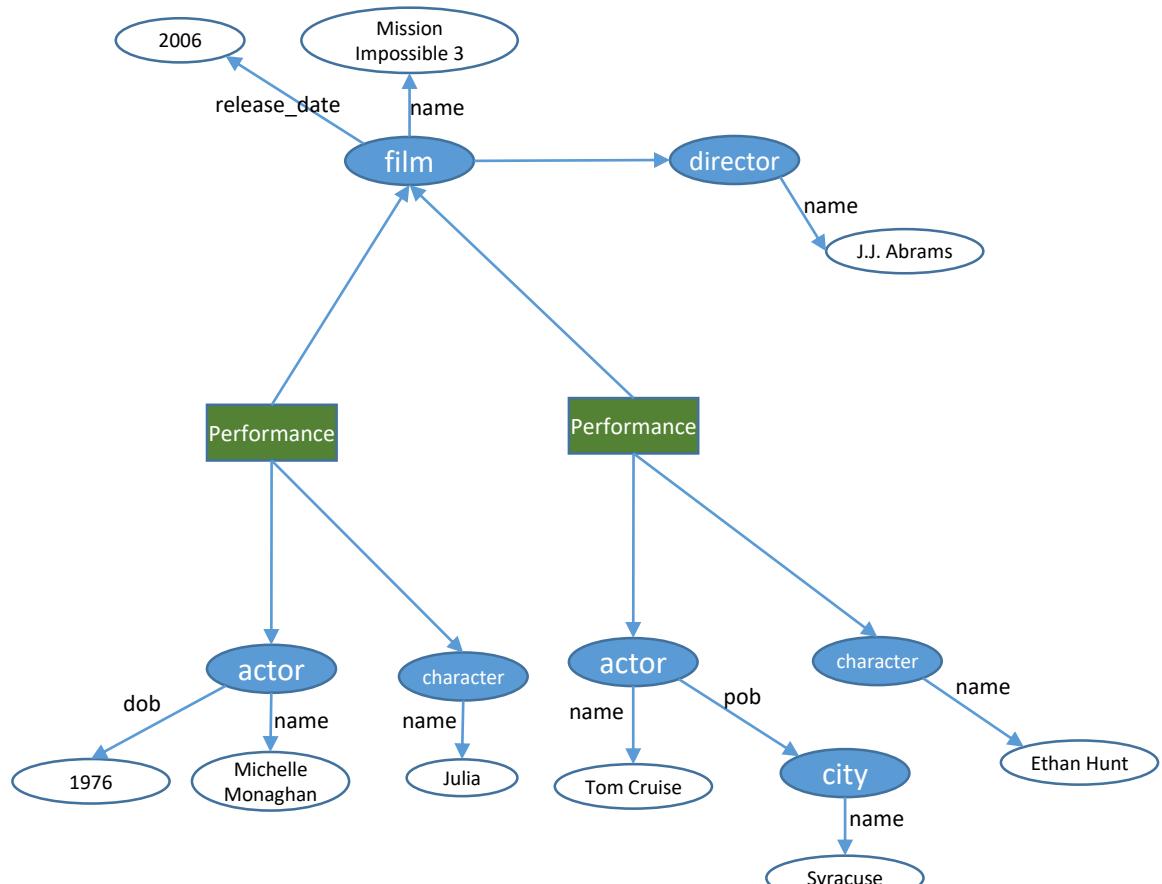
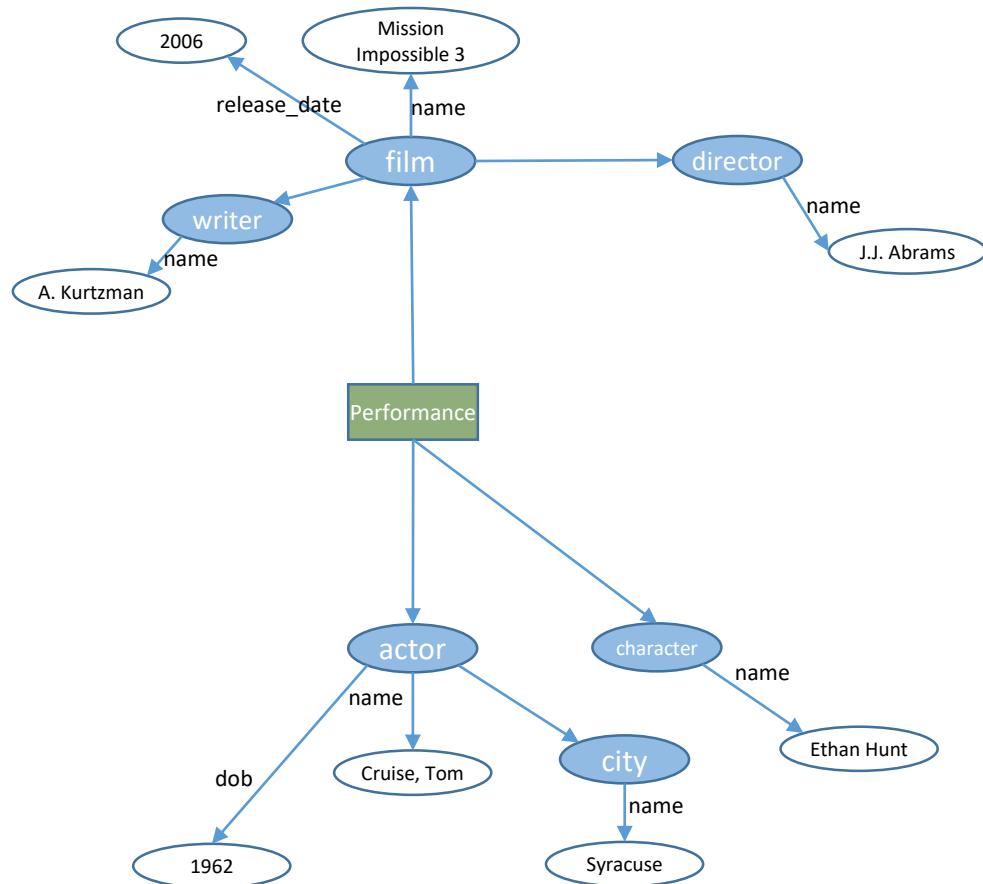
Director: J.J. Abrams  
Writers: Alex Kurtzman, Roberto Orci | 2 more credits »  
Stars: Tom Cruise, Michelle Monaghan, Ving Rhames | See full cast & crew »

TOM CRUISE MISSION IMPOSSIBLE III THE MISSION BEGINS MAY 5

66 Metascore From metacritic.com Reviews 875 user | 308 critic Popularity 725 (★ 534)

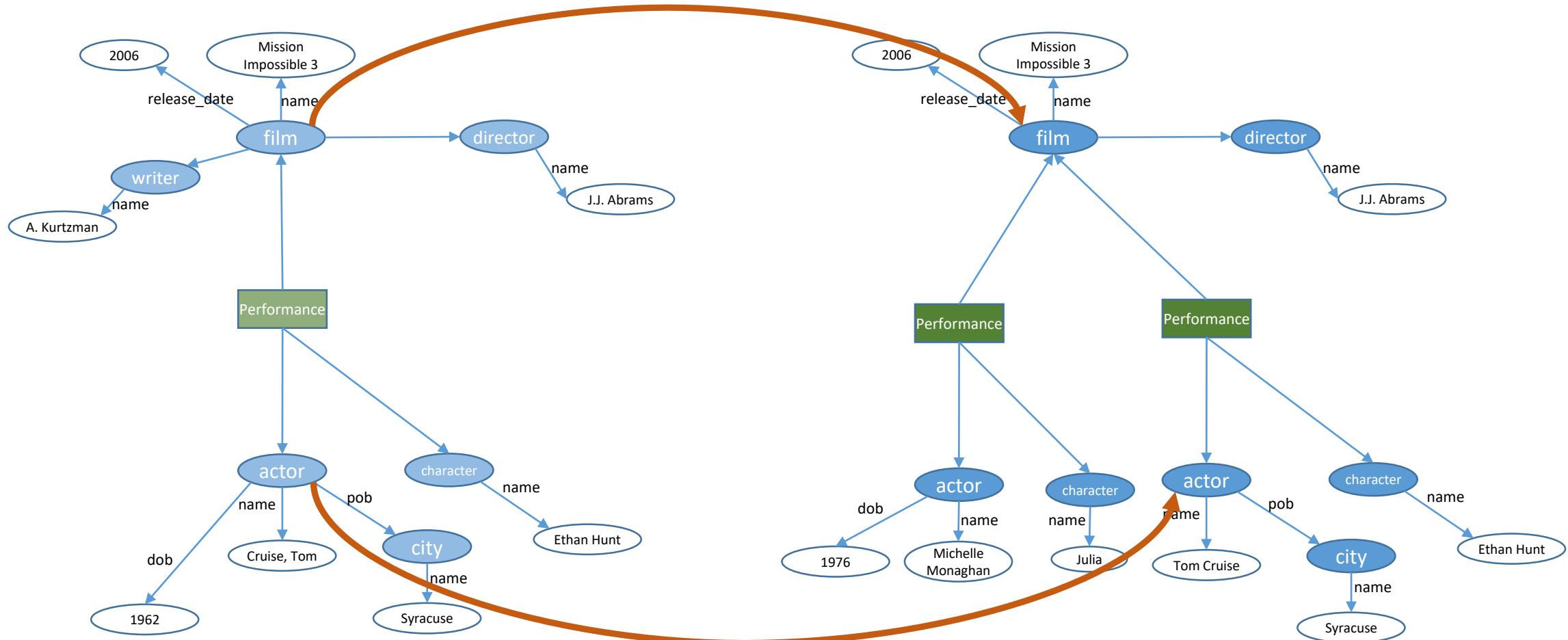


# Graph Data Model and Conflation

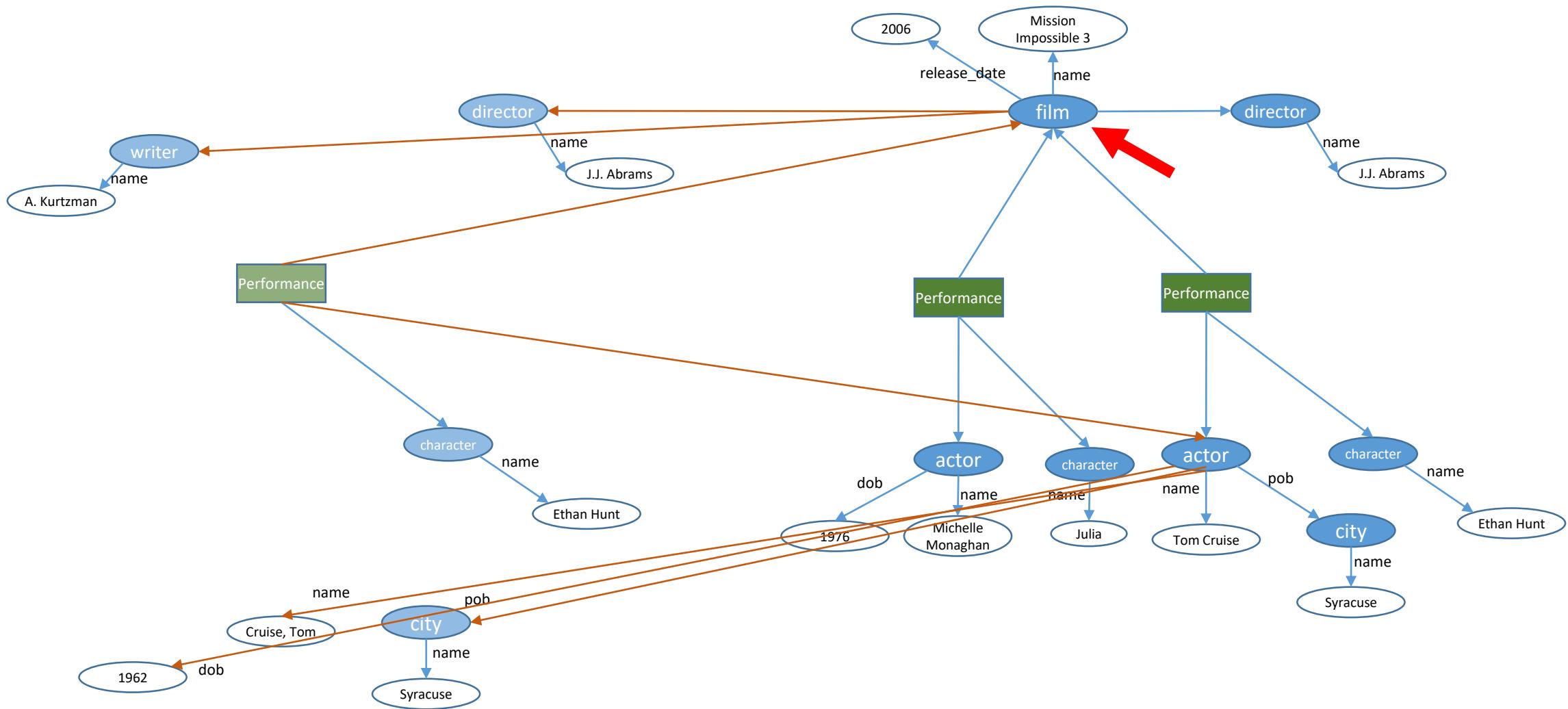


Similarity Signals Propagation

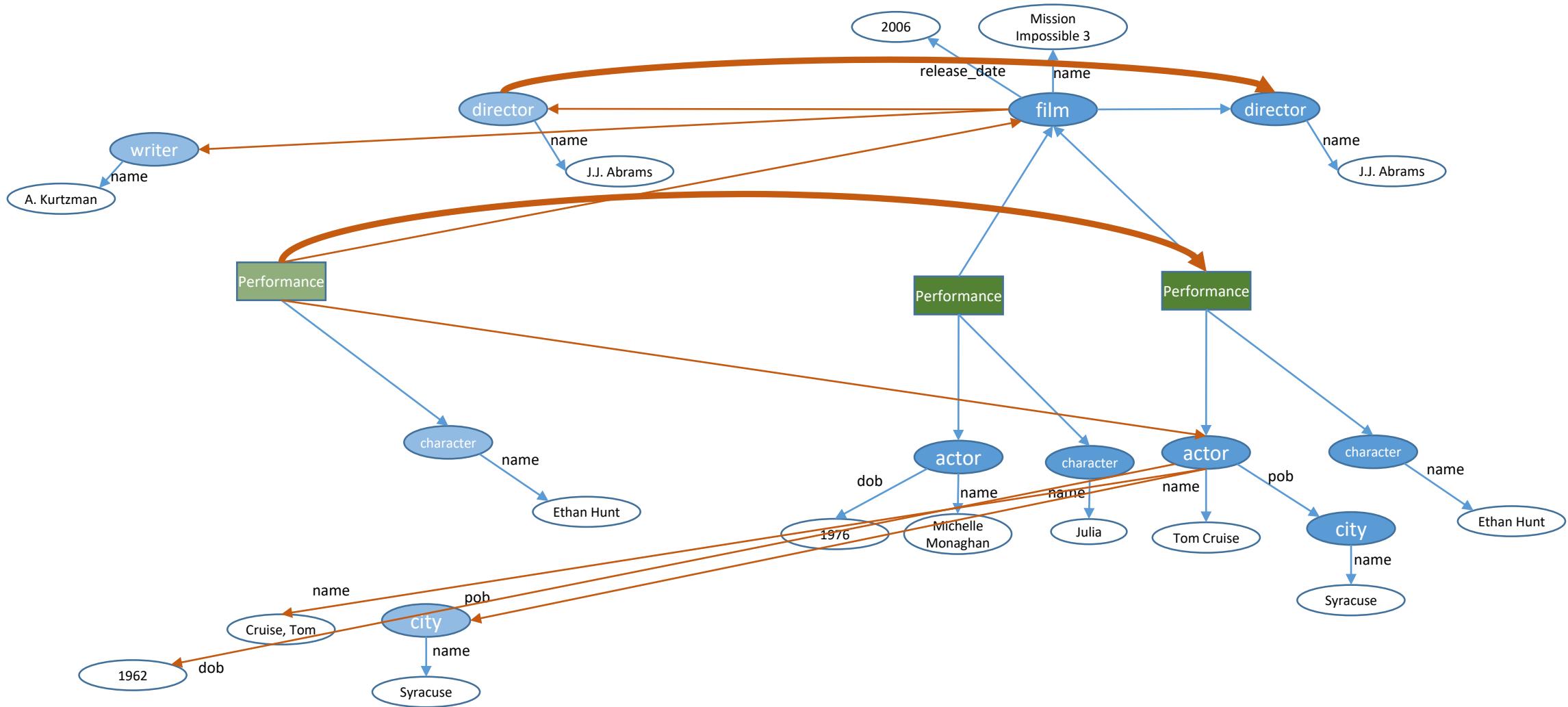
# Graph Data Model and Conflation



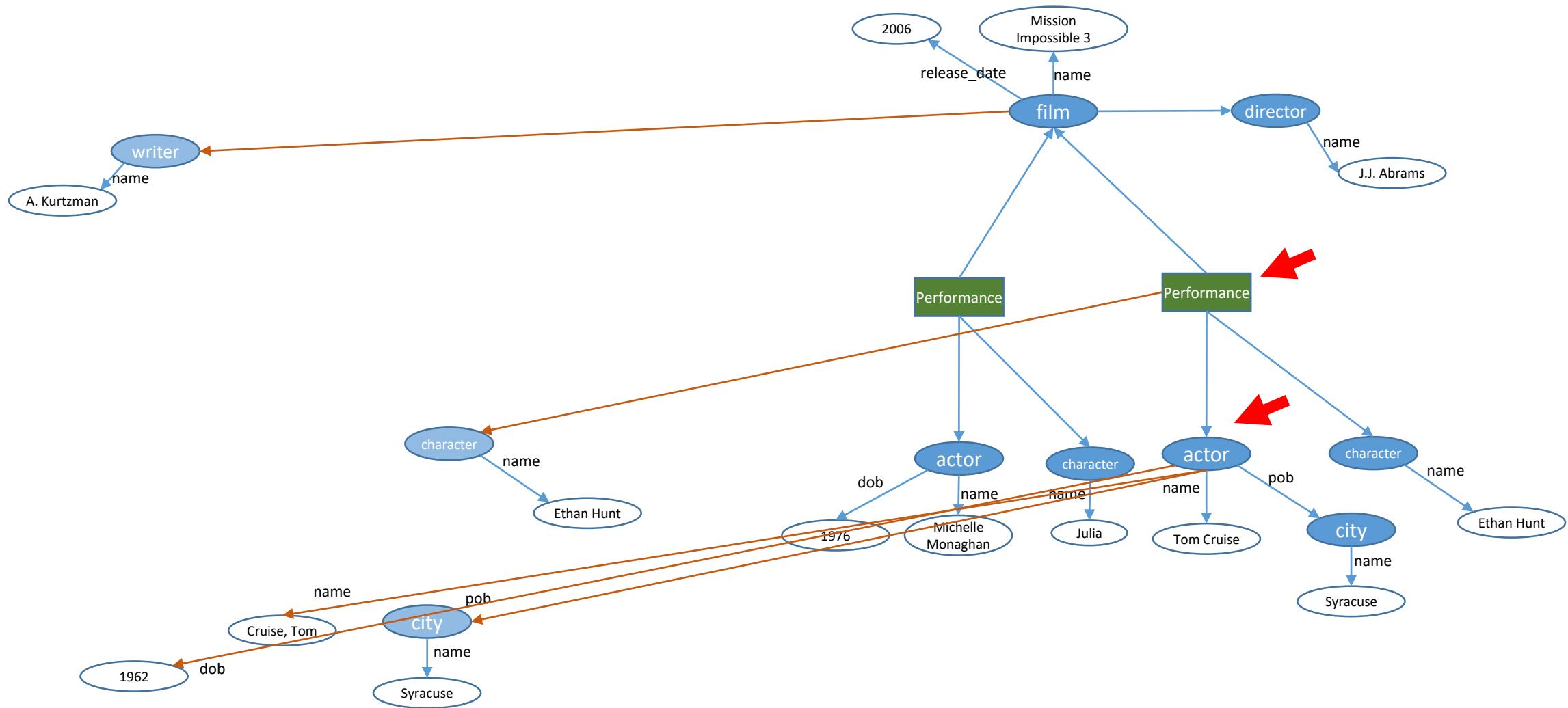
# Graph Data Model and Conflation



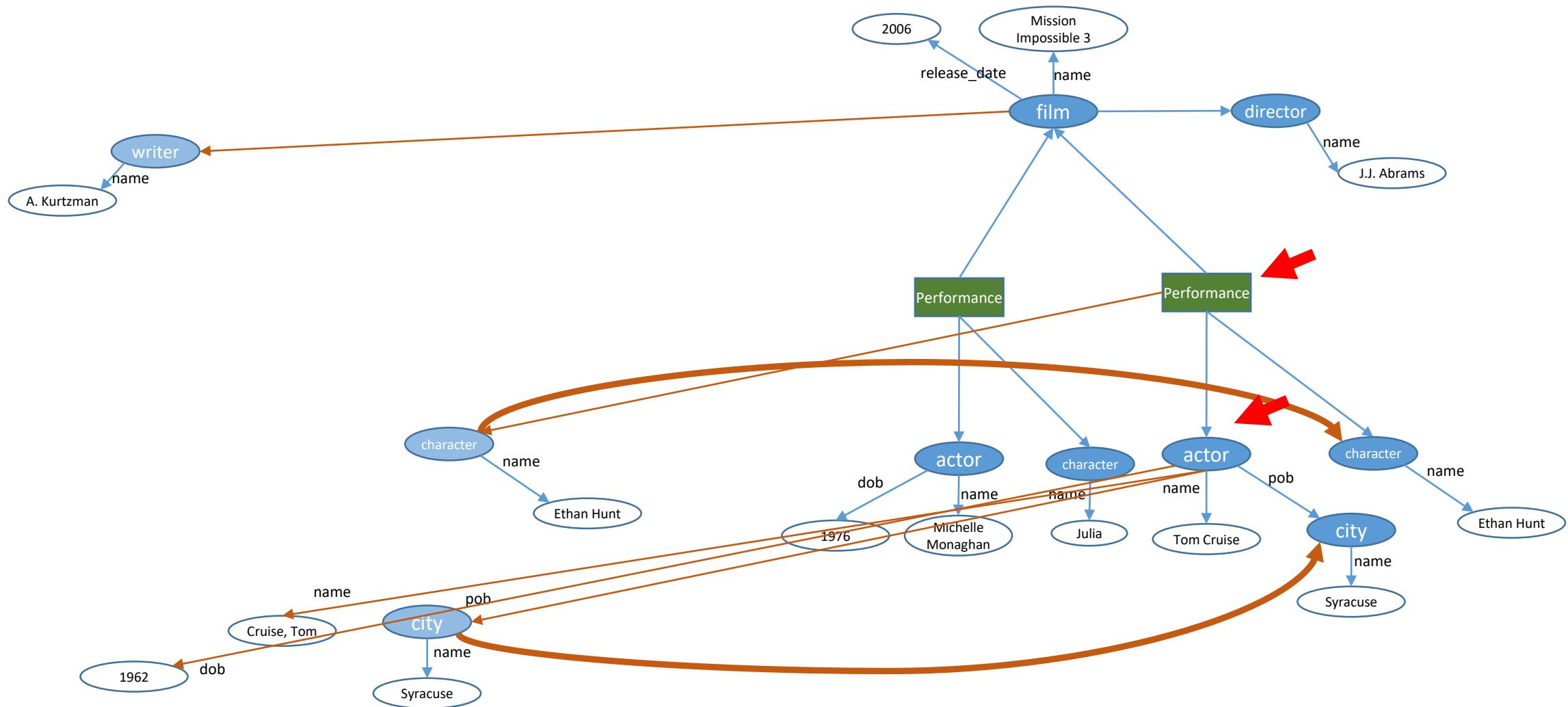
# Graph Data Model and Conflation



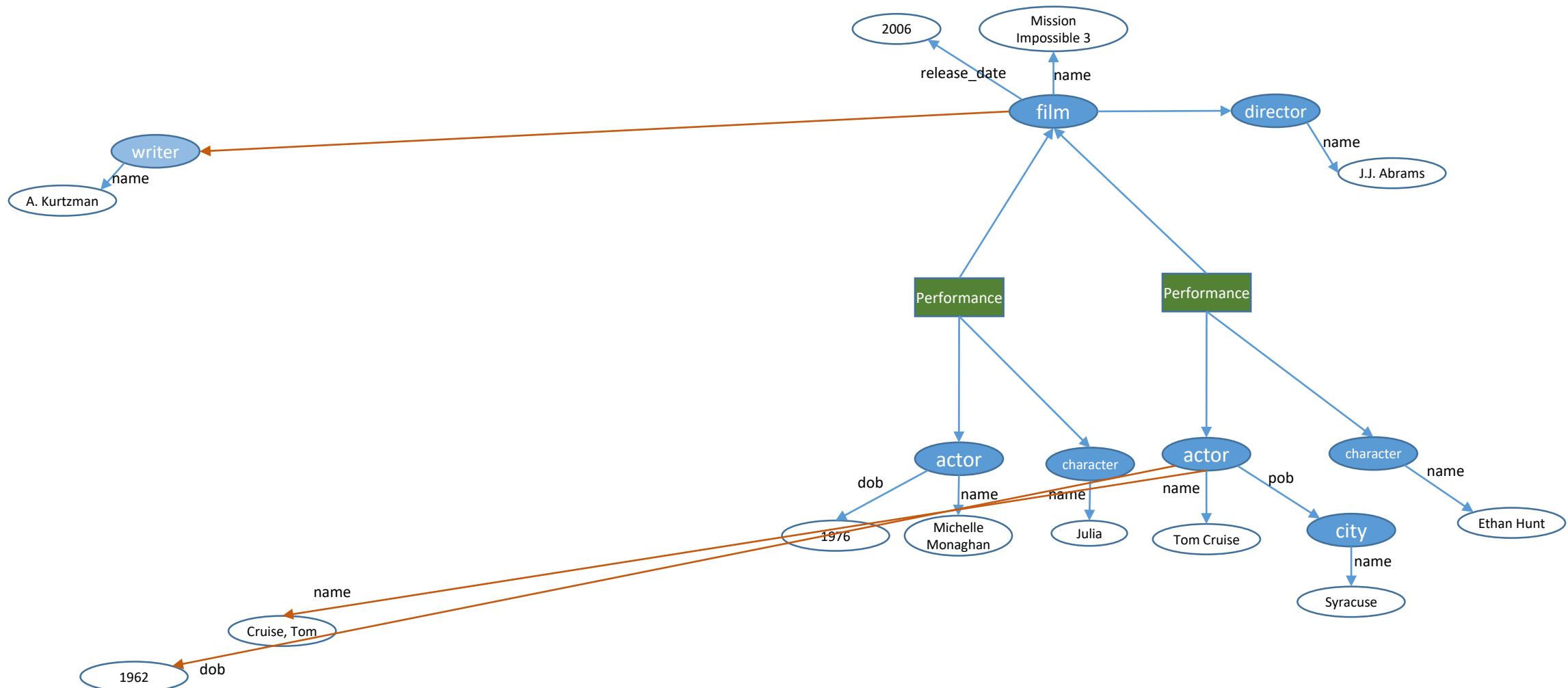
# Graph Data Model and Conflation



# Graph Data Model and Conflation



# Graph Data Model and Conflation



# Detection of Matched Entities: Collective Resolution in Linked Data

- Entity similarity based on connections
- Measures
  - **Adamic/Adar Measure:** Two nodes are more similar if they share more items that are overall less frequent

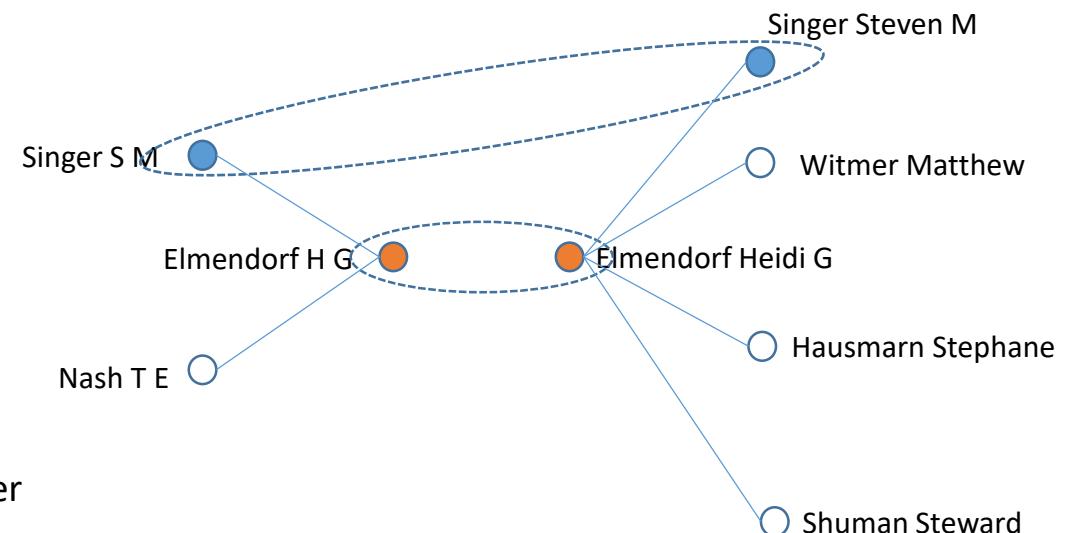
$$sim(a, b) = \sum_{i \in shared} \frac{1}{\log(freq(i))}$$

- **SimRank:** Two objects are similar if they are related to similar objects

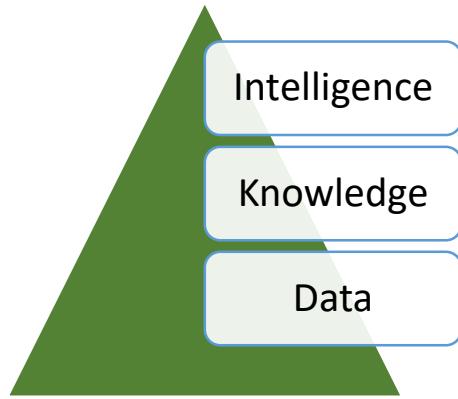
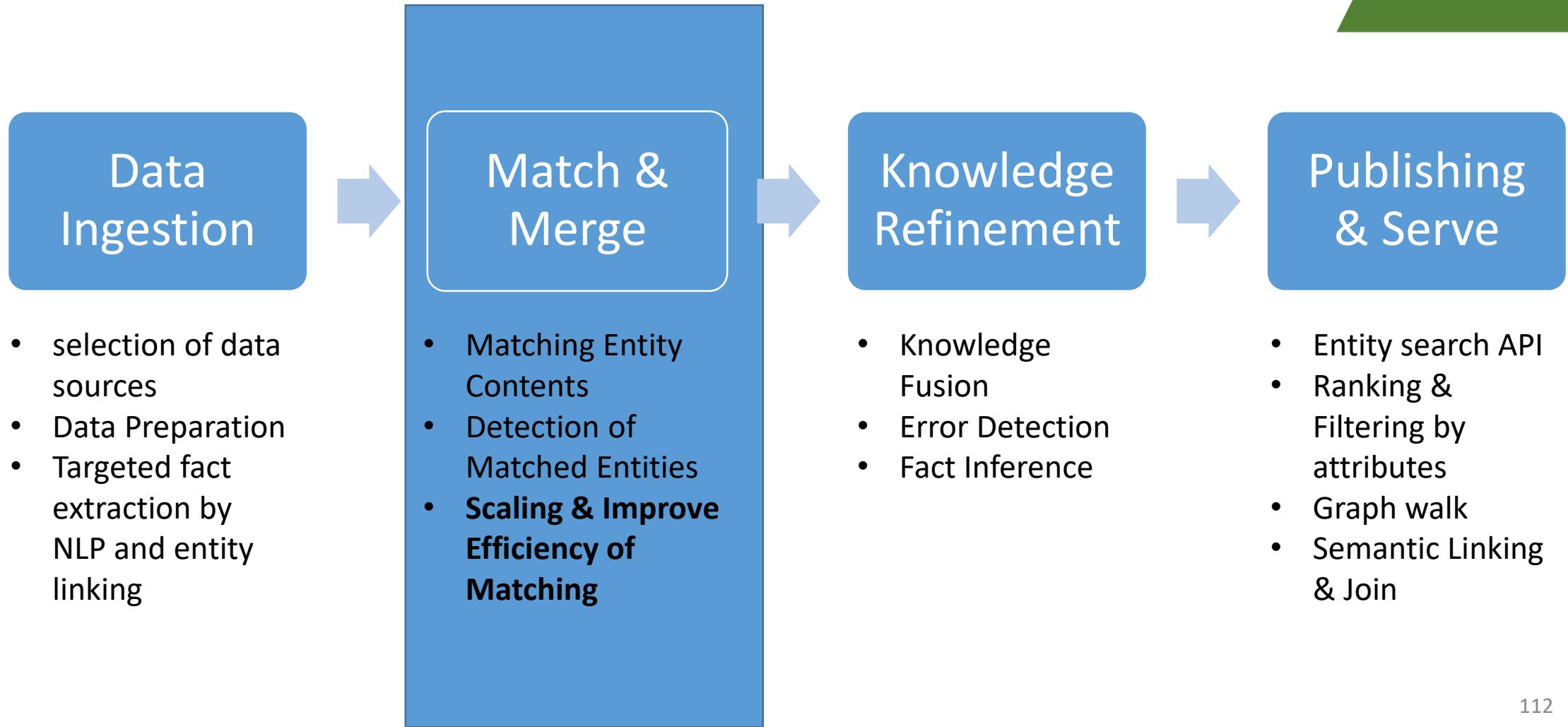
$$sim(a, b) = \frac{c}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} sim(I_i(a), I_j(b))$$

- **Katz Score:** Two objects are similar if they are connected by shorter paths

$$sim(a, b) = \sum_{l=1}^{\infty} \beta^l \cdot |paths^{(l)}(a, b)|$$



# Satori Graph Build



# Improve Efficiency of Matching

- Matching two data sources each with 1 M entities
- $1M \times 1M$  with an entity pair comparison time of 5  $\mu$ s
- 160 years
- 300K machines to finish in 5 hrs

- **Solution:** Blocking or Indexing

- Efficiency or Reduction Ratio = 
$$\frac{|\text{compared pairs}|}{m \times n}$$

- Recall or pairs completeness = 
$$\frac{|\text{True Matches compared}|}{|\text{All existing true matches}|}$$

Entity	Hashes
E1	h1, h2
E2	h1, h3, h4
E3	h3
E4	h4



Inverted Index

Key	Post List
h1	E1, E2
h2	E1
h3	E1, E2, E3
h4	E2, E4

# Improve Efficiency of Matching

- Reduce the number of entities comparisons  
(Indexing or Blocking)
  1. Identify blocking attributes
  2. Hashing Functions
  3. Retrieval of pairs

Entity	Hashes
E1	h1, h2
E2	h1, h3, h4
E3	h3
E4	h4



Inverted Index

Key	Post List
h1	E1, E2
h2	E1
h3	E1, E2, E3
h4	E2, E4

# Improve Efficiency of Matching

- Reduce the number of entities comparisons  
(Indexing or Blocking)

## 1. Identify Blocking Attributes

- **Quality of values** in the attributes may directly cause recall loss
- **Frequency and distribution of values** directly impact performance and recall.
- Best practice:
  - Use several attributes with combinations
  - Estimate and/or learn **Identity Attributes**
    - Movie name and release date –or– movie name, producer and director
    - Person name, date of birth and place of birth –or person name, affiliation and age

Entity	Hashes
E1	h1, h2
E2	h1, h3, h4
E3	h3
E4	h4



Inverted Index

Key	Post List
h1	E1, E2
h2	E1
h3	E1, E2, E3
h4	E2, E4

# Improve Efficiency of Matching

- Reduce the number of entities comparisons (Indexing or Blocking)

## 2. Hashing Functions

- PassThrough:  $H(\text{Tom Cruse}) = \{\text{Tom Cruse}\}$
- TokenSequence:  $H(\text{Tom Cruse}) = \{\text{crusetom}\}$
- Metaphone:  $H(\text{Robert})=H(\text{Rupert})$
- Q-Gram (a lot of hashes per value)
  - $H(\text{Smith})=\{\text{smmiith, miith, smith, smmith, smmiit}\}$
  - (1) Compute grams (2) concat except one
  - 2-Gram( $\text{smith}\text{)}=\{\text{sm, mi, it, th}\}$
  - $H(\text{Smith}) \cap H(\text{Smithy}) \cap H(\text{Smithe}) = \{\text{smmiith}\}$
- Suffix Array (a lot of duplicate post list)
  - $H(\text{Catherine})=\{\text{catherine, atherine, therine, herine}\}$
- Minhash

# Improve Efficiency of Matching: Hashing Functions

- **MinHash: min-wise independent permutations**
  - Convert the **string** to a **set of elements**
  - **Random function** to give a **random order** for all the elements in the universe
  - For two sets of elements  $S_1, S_2$
  - $J(S_1, S_2) = P_r(\text{minhash}(S_1) = \text{minhash}(S_2))$
- Example:
  - $s_1 = \{c, d, e, f, g\}$      $s_2 = \{c, d, x, f, g\}$
  - Order1: a,b,c,d,e,f,g, ...x
    - $s_1 = \{g, f, e, d, c\}$      $s_2 = \{x, g, f, d, c\}$
    - $\text{minhash}(s_1)=c$      $\text{minhash}(s_2)=c$
  - Order2:a,g,d,x,e,b,f,c...
    - $s_1 = \{c, f, e, d, g\}$      $s_2 = \{c, f, x, d, g\}$
    - $\text{minhash}(s_1)=g$      $\text{minhash}(s_2)=g$
- If  $\text{sim}(s_1, s_2) = 0.6$ , then by generating **two minhashes**, they will **overlap with probability**  
$$1 - [(1 - 0.6)(1 - 0.6)] = 1 - (0.4 \times 0.4) = 1 - 0.16 = 0.84$$

# Improve Efficiency of Matching

- Reduce the number of entities comparisons (Indexing or Blocking)

## 3. Retrieval

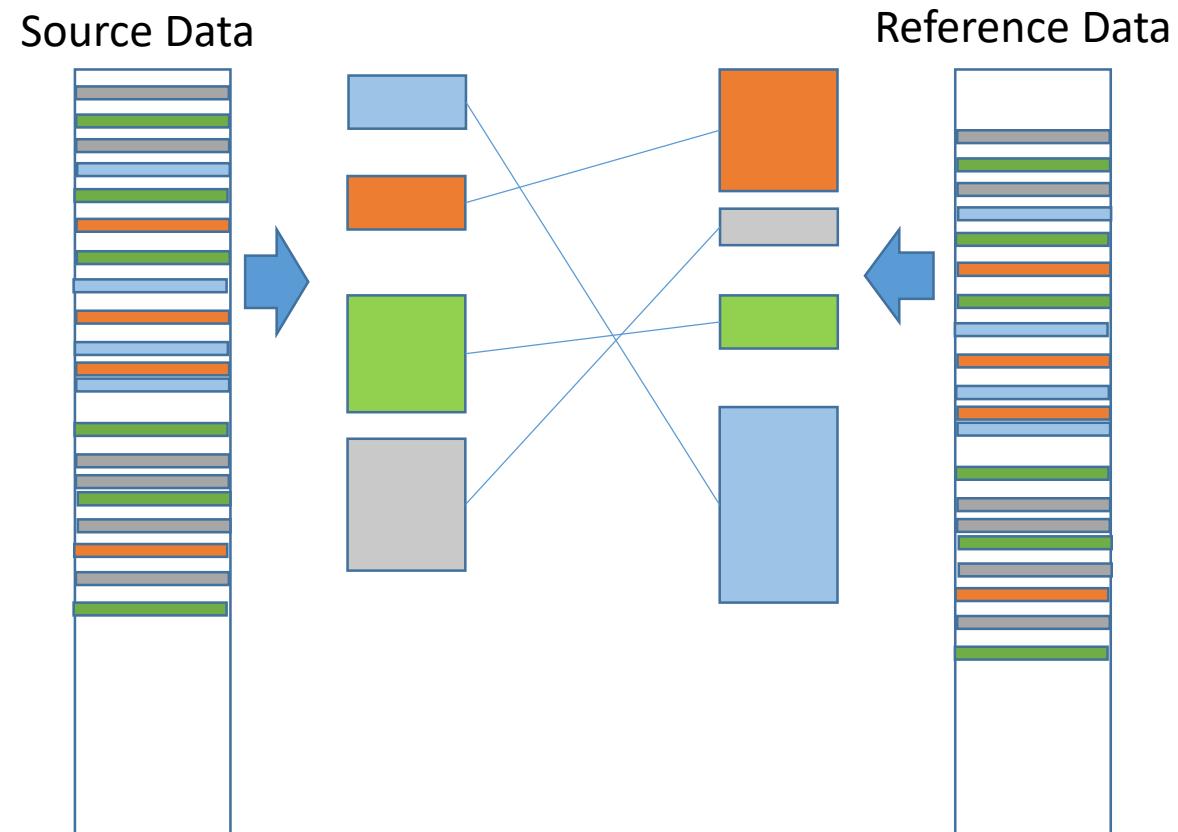
- Within blocks comparison
- Sorted Neighborhood
- Canopy Clustering (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
- Entity Index Join

# Improve Efficiency of Matching

- Reduce the number of entities comparisons (Indexing or Blocking)

## 3. Retrieval

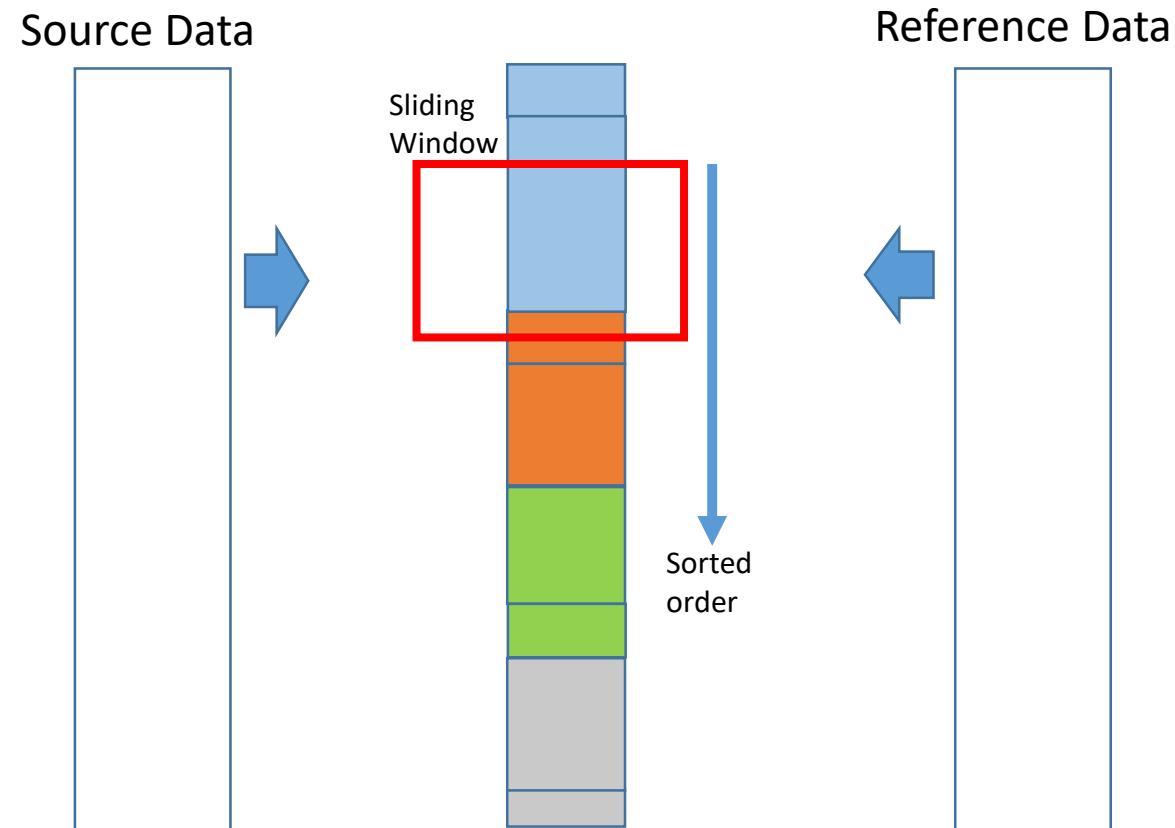
- **Within blocks comparison**
- Sorted Neighborhood
- Canopy Clustering (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
- Entity Index Join



# Improve Efficiency of Matching

- Reduce the number of entities comparisons (Indexing or Blocking)

3. Retrieval
  - Within blocks comparison
  - **Sorted Neighborhood**
  - Canopy Clustering (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
  - Entity Index Join

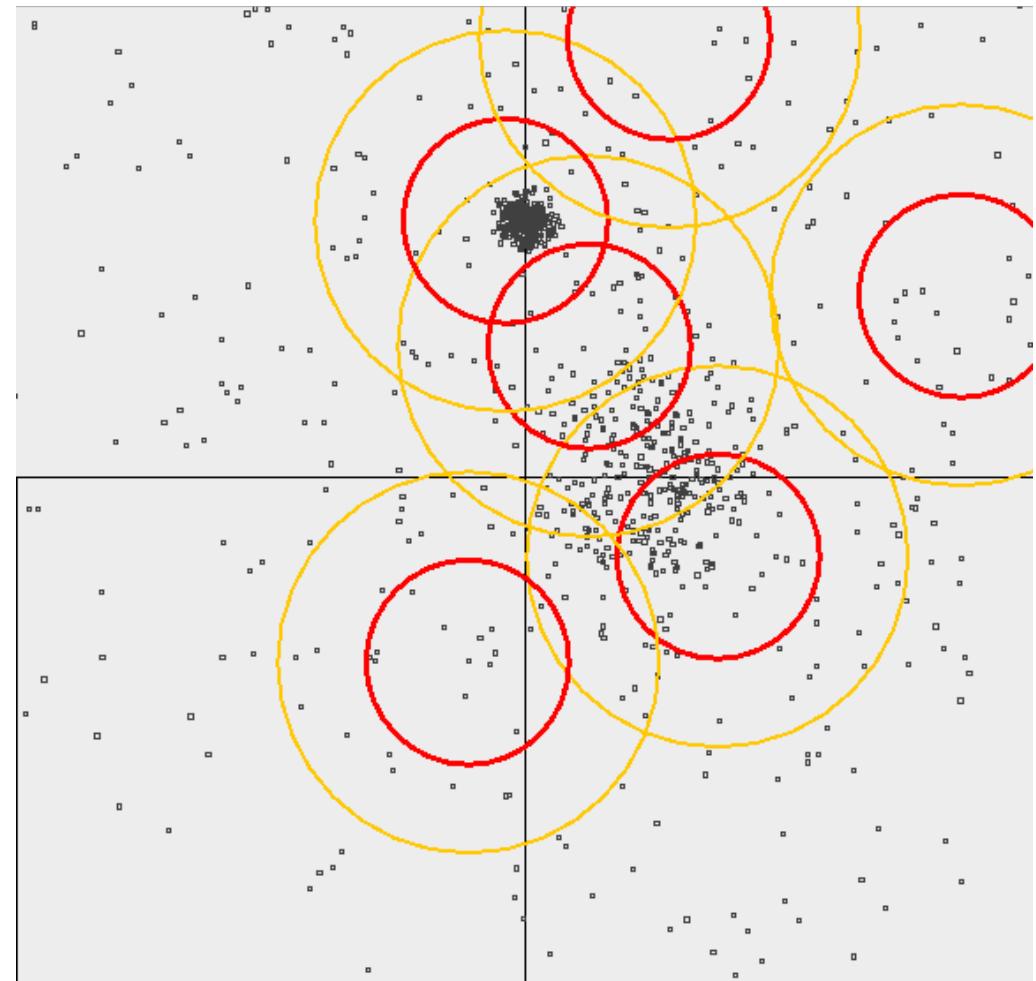


# Improve Efficiency of Matching

- Reduce the number of entities comparisons (Indexing or Blocking)

## 3. Retrieval

- Within blocks comparison
- Sorted Neighborhood
- **Canopy Clustering** (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
- Entity Index Join



# Improve Efficiency of Matching

- Reduce the number of entities comparisons (Indexing or Blocking)

## 3. Retrieval

- Within blocks comparison
- Sorted Neighborhood
- Canopy Clustering (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
- **Entity Index Join**

$$- E_1 = \{(h, idf_1(h)) \dots\}$$

$$- E_2 = \{(h, idf_2(h)) \dots\}$$

$$L_1(E_1, E_2) = \sum_{\forall h} idf_1(h) \times idf_2(h)$$

Entity	Hashes
E1	h1, h2
E2	h1, h3, h4
E3	h3
E4	h4

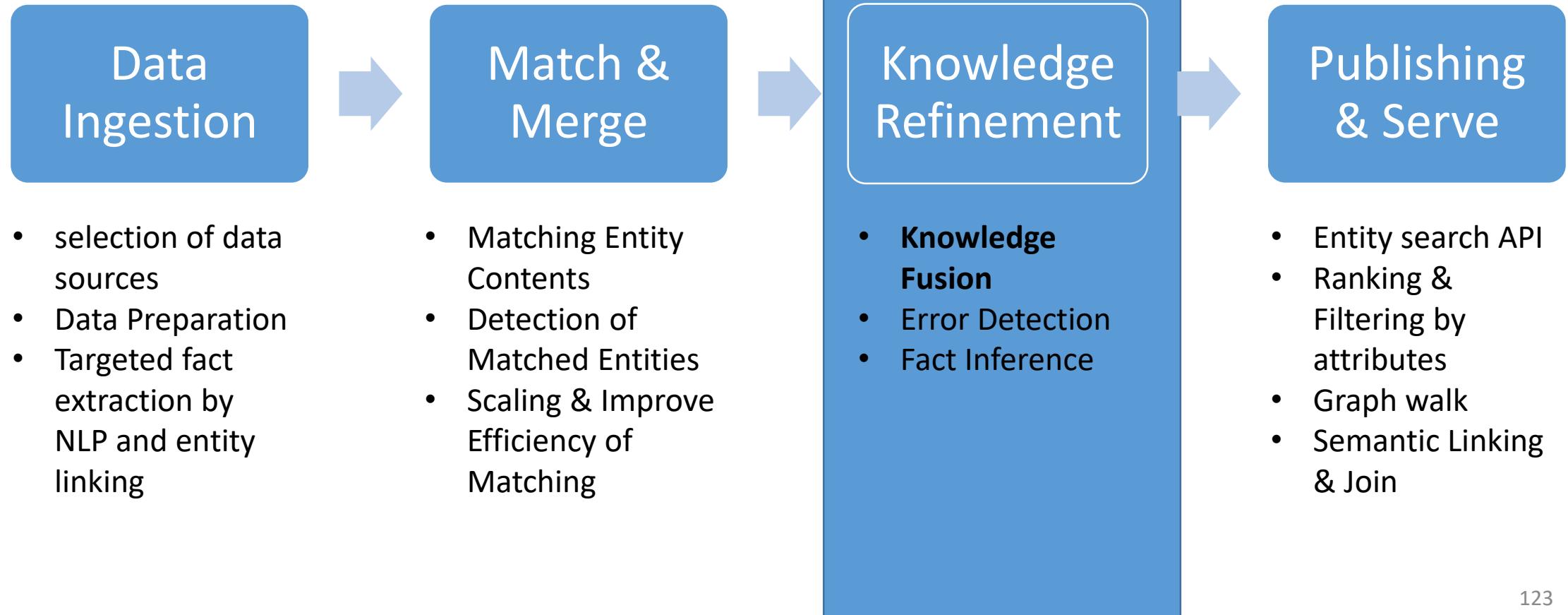
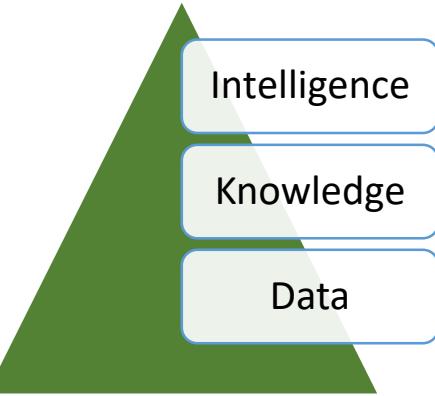


Inverted Index

Key	IDF	Post List
h1	IDF(h1)	E1, E2
h2	IDF(h2)	E1
h3	IDF(h3)	E1, E2, E3
h4	IDF(h4)	E2, E4

Return Top K entities for each other entity

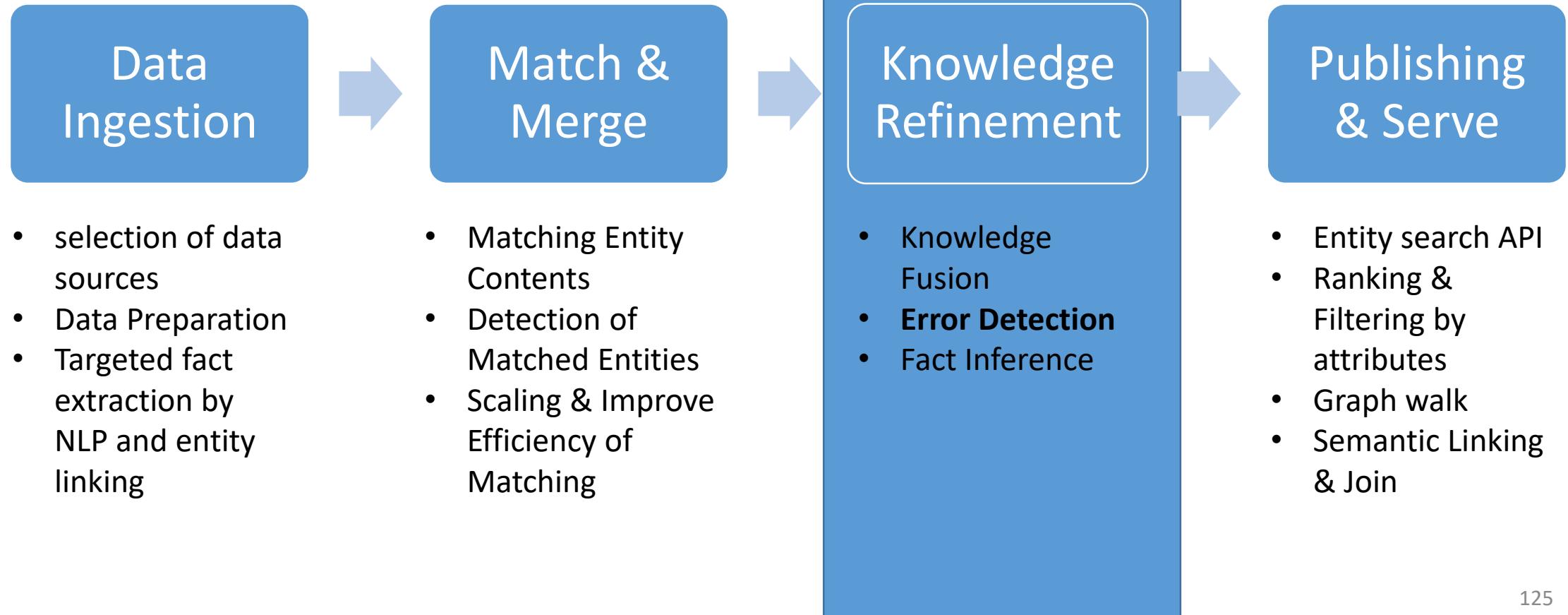
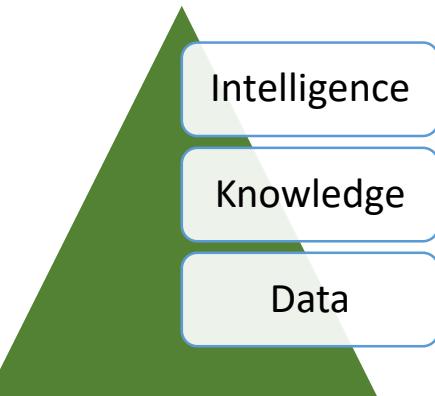
# Satori Graph Build



# Knowledge Fusion (Merging Entities)

- After merging entity nodes in the graph, we end up with conflicting facts and connections
- Resolving facts (and finding truth)
  - Majority Voting
  - Identify Authoritative Sources
  - Fact Checker
    - Gather evidence from different sources
    - Evaluate evidences
    - Model joint interactions
    - Aggregate evidence and predict

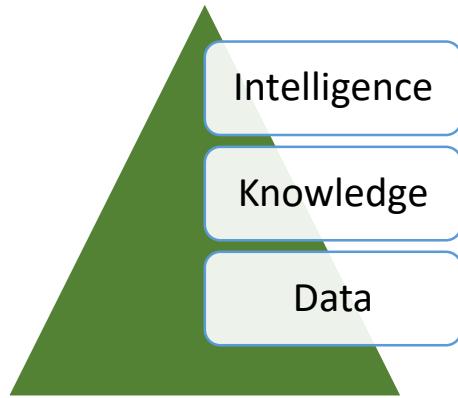
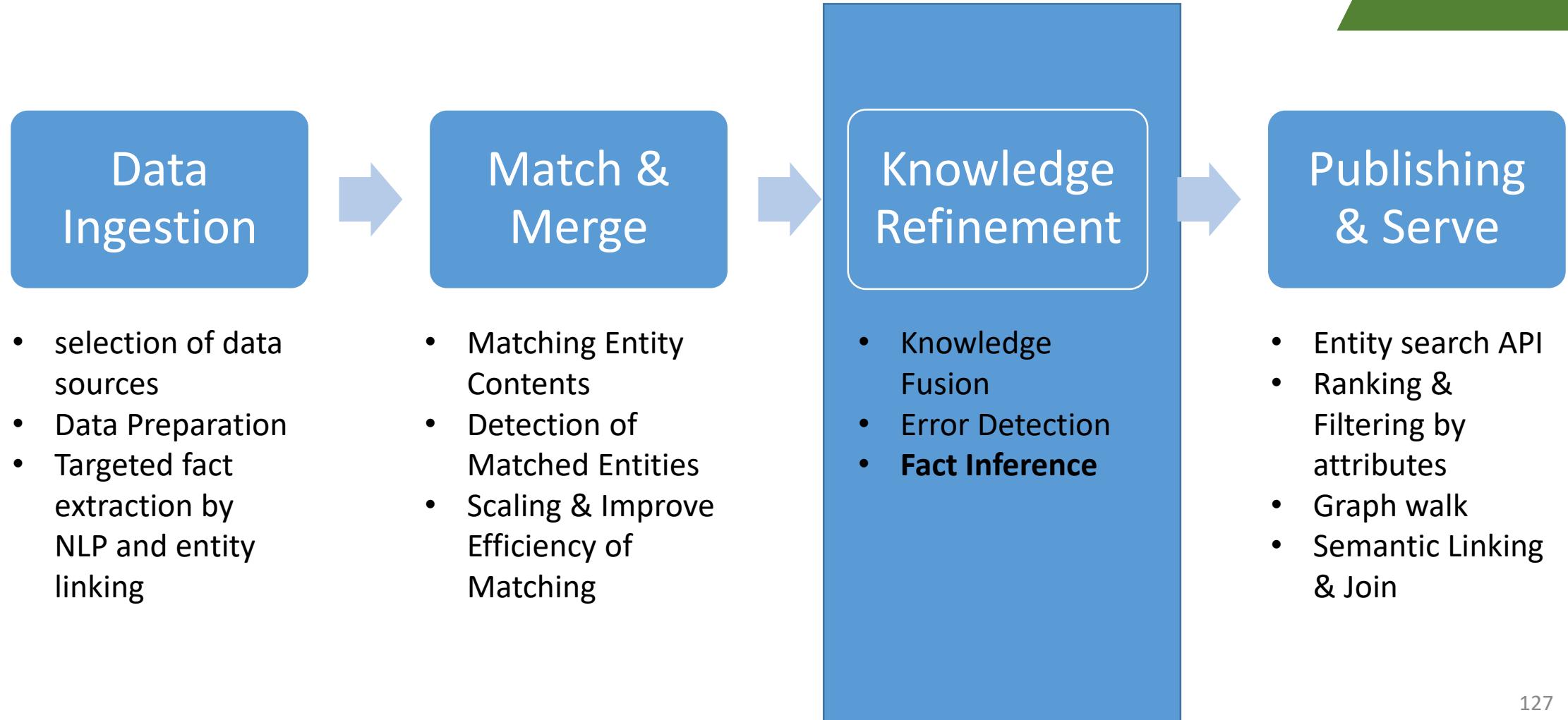
# Satori Graph Build



# Error Detection

- Error Detection
  - Data Quality Rules
    - Functional Dependency and its conditional variation  
e.g.; Zip → City
    - Inconsistency  
`Entity cannot be a movie and book`  
`Date_of_birth < date_of_death`
    - Outliers detection
  - External signals for relationship validation (e.g.; co-clicks)
  - NLP features (e.g.; deadlive)

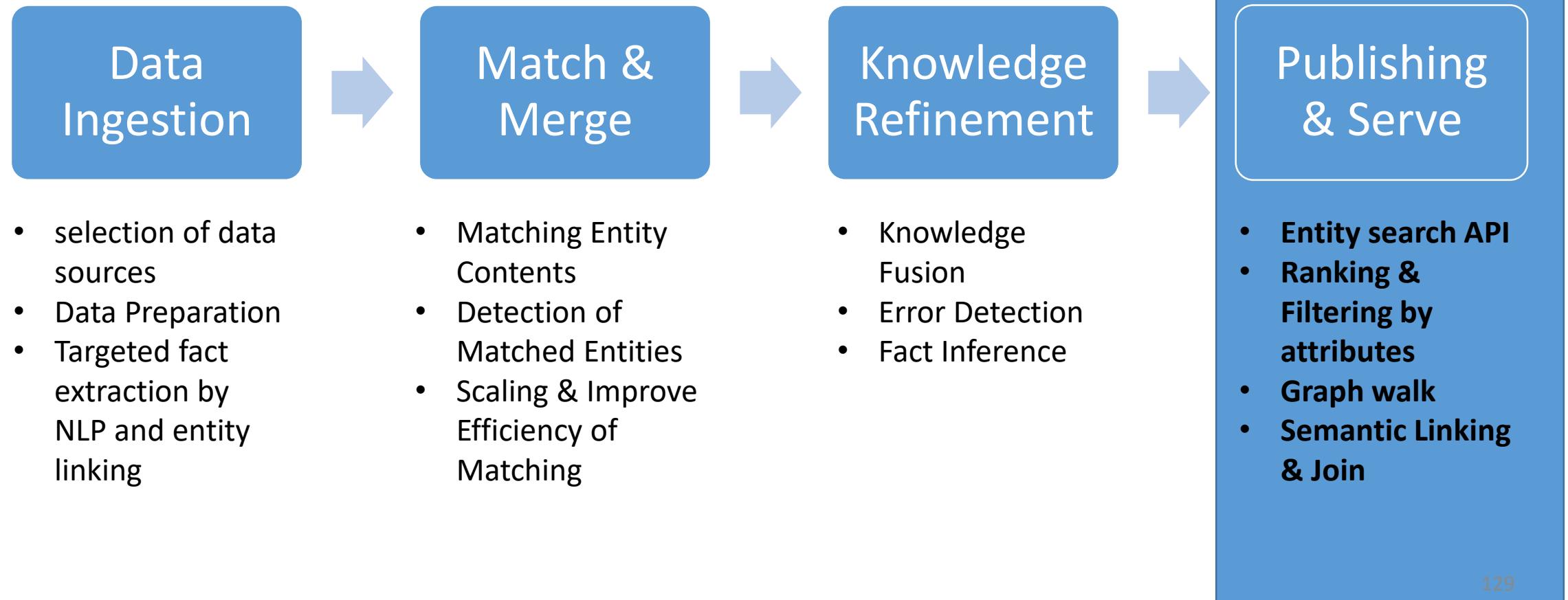
# Satori Graph Build



# Fact Inference

- Further Enrichment/Data completion
  - Internal: Dominant type and Label
  - External: search engine method for enriching social links

# Satori Graph Build



# Part IV: Serving Knowledge to the World

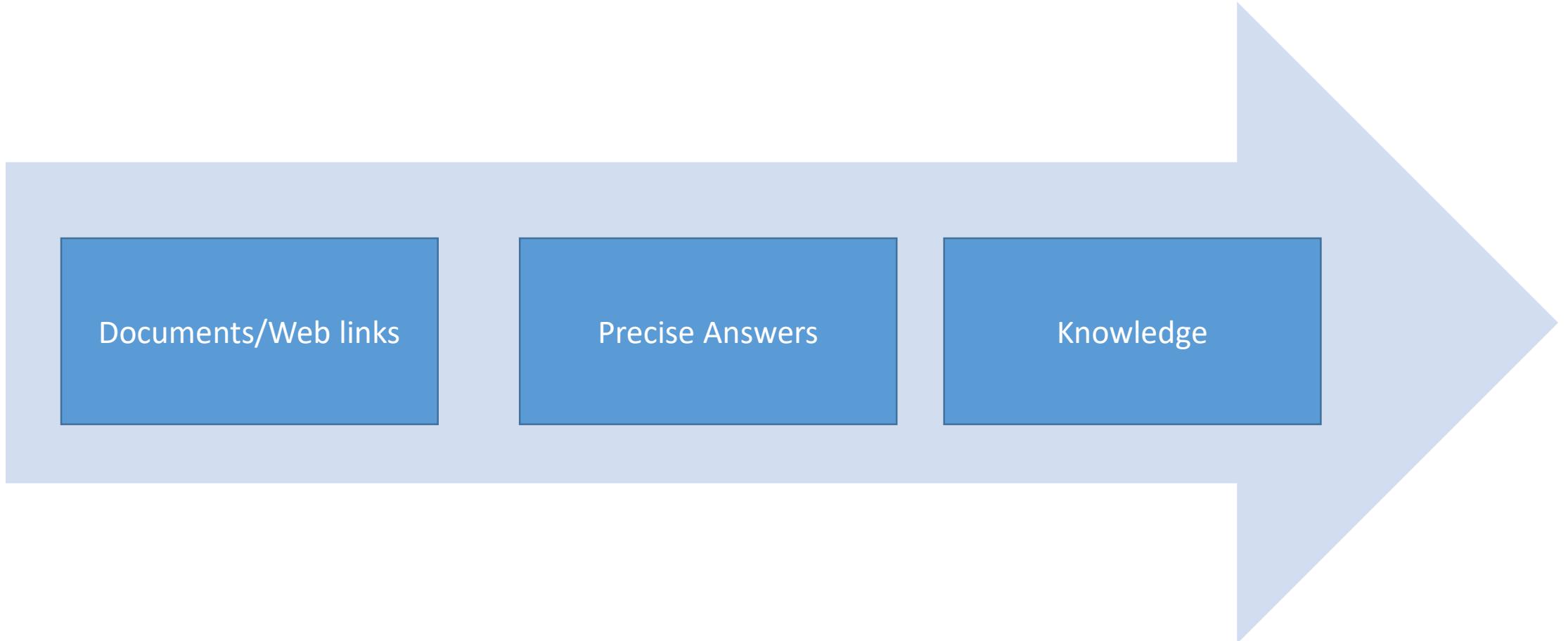
**Ahmed K. A. Mohamed**

Senior Applied Data Scientist, Satori Group, Microsoft AI+R

[ahmedat@microsoft.com](mailto:ahmedat@microsoft.com)



# Why Knowledge Graph Serve



# Satori Knowledge Graph Application Areas: Bing&Cortana

Satori data and serve APIs has a tremendous impact on all Bing impressions for e.g.



Finding Dory

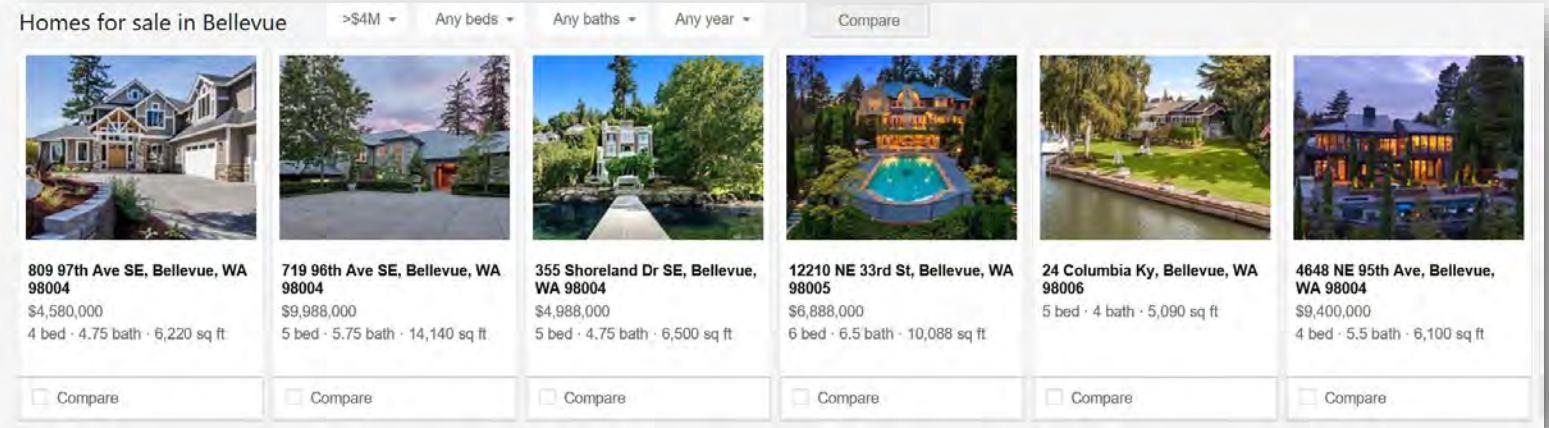
PG · 2016 · 1hr 43min · Animation/Adventure

IMDb 7.4/10 Rotten Tomatoes 94% Metacritic 77

Disney Pixar's "Finding Dory" welcomes back to the big screen everyone's favorite forgetful blue tang Dory, who's living happily in the reef with Marlin and Nemo. When Dory suddenly remembers that she has a family out there who may be looking for her, the trio takes off on a life-changing adventure across the ocean to California's prestigious Marine Life Ins... +

Watch now actions for movie entities

Recently viewed shows personal history



Homes for sale in Bellevue

>\$4M · Any beds · Any baths · Any year · Compare

Address	Price	Beds	Baths	Size
809 97th Ave SE, Bellevue, WA 98004	\$4,580,000	4 bed	4.75 bath	6,220 sq ft
719 96th Ave SE, Bellevue, WA 98004	\$9,988,000	5 bed	5.75 bath	14,140 sq ft
355 Shoreland Dr SE, Bellevue, WA 98004	\$4,988,000	5 bed	4.75 bath	6,500 sq ft
12210 NE 33rd St, Bellevue, WA 98005	\$6,888,000	6 bed	6.5 bath	10,088 sq ft
24 Columbia Ky, Bellevue, WA 98006	\$5,688,000	5 bed	4.5 bath	5,090 sq ft
4648 NE 95th Ave, Bellevue, WA 98004	\$9,400,000	4 bed	5.5 bath	6,100 sq ft

Carousel of information from Satori

Watch now

Amazon Watch

iTunes Watch

Netflix Watch

...

Recently viewed

Coco

Star Wars: Episode V - The Empire Strikes Back

Return of the Jedi

Star Wars: A New Hope

Star Wars: The Force Awakens

See all (6+)



Carousel

bing MS Beta

what are the tallest mountains in washington

Web Images Videos Maps News More

21,900,000 RESULTS Any time

Washington - Highest mountains

Mountain	Height (feet)
Mount Rainier	14,411 feet
Mount Adams	12,281 feet
Little Tahoma Peak	11,138 feet
Mount Baker	10,781 feet

bing MS Beta

what is the periodic symbol for silicon

Web Images Videos Maps News More

932,000 RESULTS Any time

Silicon symbol

Si

Data from wikipedia

bing MS Beta

who was grace hopper married to

Web Images Videos Maps News More

1,830,000 RESULTS Any time

Grace Hopper spouse

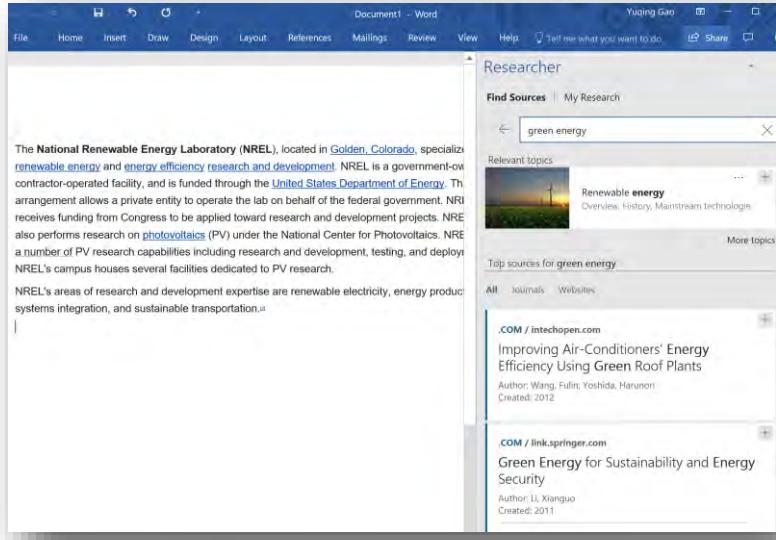
Vincent Foster Hopper

(m. 1930-1945)

Data from whosdatedwho

# Satori Knowledge Graph Application Areas: Office

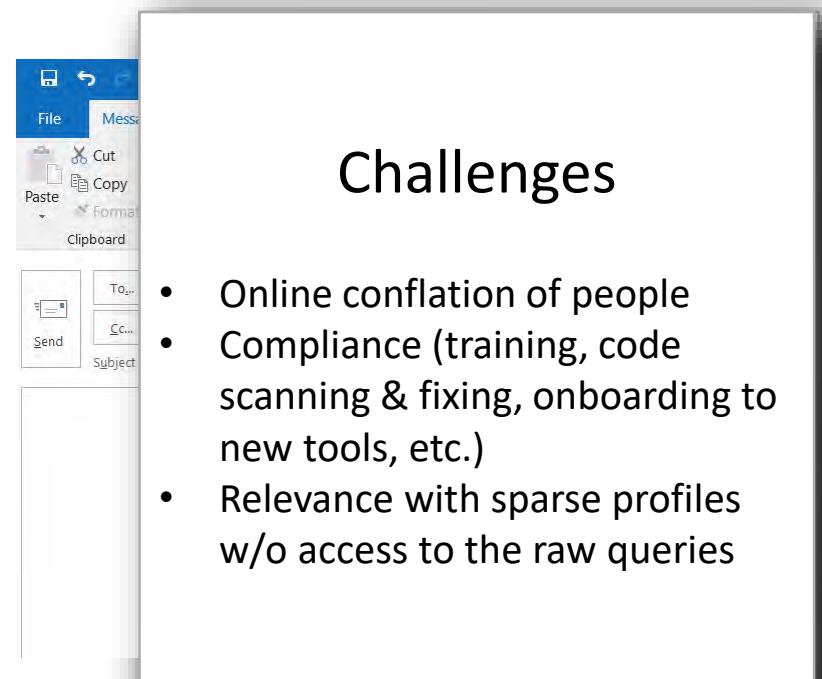
## Enriching the Office experience with Satori data



**Researcher in Word & OneNote**  
Get topic information straight  
into your documents

A	B	C	D	E
2	Stock	52 week high	52 week low	Exchange
3	Microsoft Corp	96.07	63.62	NASDAQ
4	Alphabet Inc	1,186.89	803.37	NASDAQ
5	Apple Inc			
6	Amazon.com Inc			
7	Walt Disney Co			
8	Starbucks Corp			
9	Nike Inc			
10	AT&T Inc			
11	HP Inc			
12	International Busin			
13				
14				
15				
16				
17				
18				
19				

**Project Yellow (Excel)**  
Finance and demographic  
information available based  
on cell contents



## Challenges

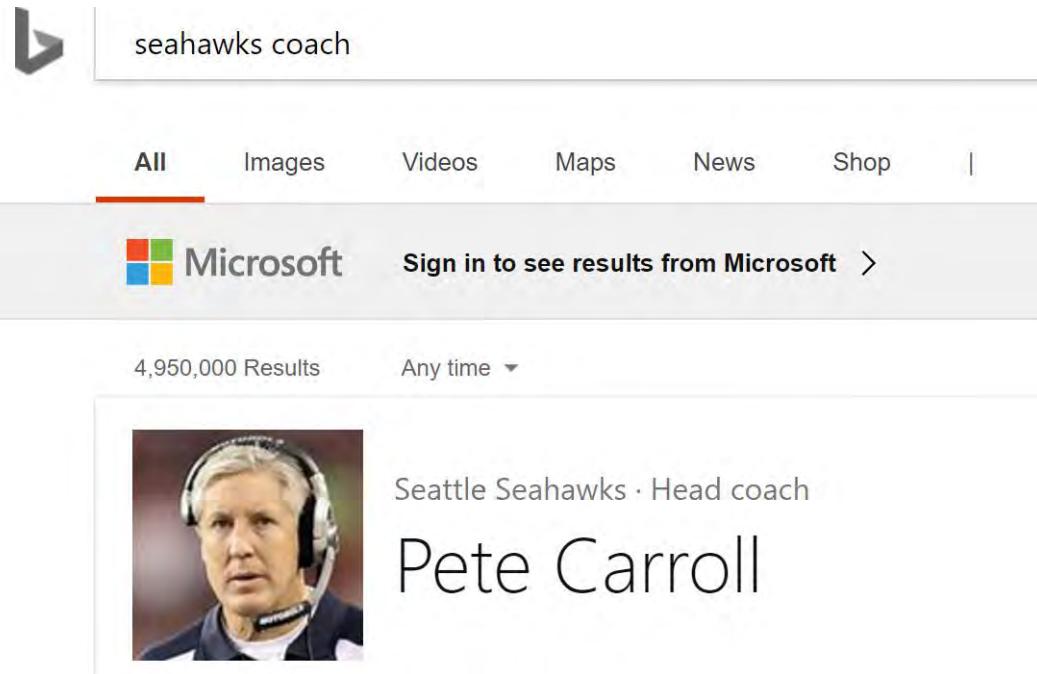
- Online conflation of people
- Compliance (training, code scanning & fixing, onboarding to new tools, etc.)
- Relevance with sparse profiles w/o access to the raw queries

LinkedIn profile information  
visible in O365 People Card  
through Satori

# Serving Knowledge by Answering Questions

- Given:
  - Knowledge graph ingested from unstructured, structured, and semi-structured data sources
- Input:
  - Natural language query
- Output:
  - Answer in the form of knowledge

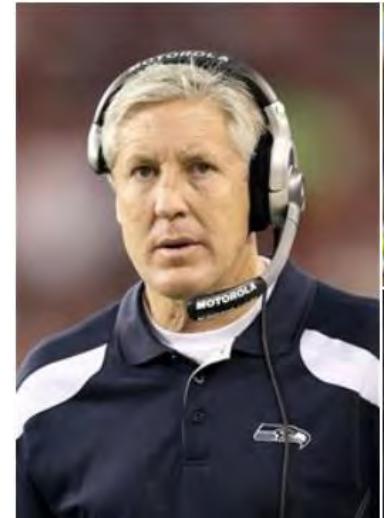
# Serving Knowledge by Answering Questions



A screenshot of a Bing search results page. The search query "seahawks coach" is entered in the search bar. Below the search bar, there are tabs for All, Images, Videos, Maps, News, and Shop. The "All" tab is selected. A Microsoft sign-in prompt "Sign in to see results from Microsoft" is visible. The search results show 4,950,000 results and are filtered by "Any time". The top result is a large image of Pete Carroll, identified as "Seattle Seahawks · Head coach".



Seattle Seahawks /American\_football\_team\_current\_head\_coach



# Serving Knowledge by Answering Questions

- Challenges:
  - Matching language
    - There are many ways to ask the same query e.g. {who directed titanic}, {what is the name of the person who directed titanic}, {in the movie titanic, who was the director}, ...etc
    - Scalable entity linking
    - Word sense disambiguation
    - Semantic roles and relationships extraction
  - Large search space
    - Every entity can have hundreds of edges and every entity instance can have hundreds of millions of edges/facts
  - Compositionality
    - {Movies starring the first wife of tom hanks}, {movies directed by the director of titanic}

# Serving Knowledge by Answering Questions

- Approaches:
  1. Semantic parsing approaches (serving graph as output):
    - 1.1 Generic semantic parsing followed by ontology grounding
    - 1.2 Knowledge base specific semantic parsing
    - 1.3 Knowledge embedding
  2. Information extraction approaches (serving passage outputs):
    - 2.1 Information retrieval methods with semantic enrichment

# 1. Semantic Parsing

2017 movies starring the actor that played batman in batman 

All Images Videos Maps News Shop | My saves

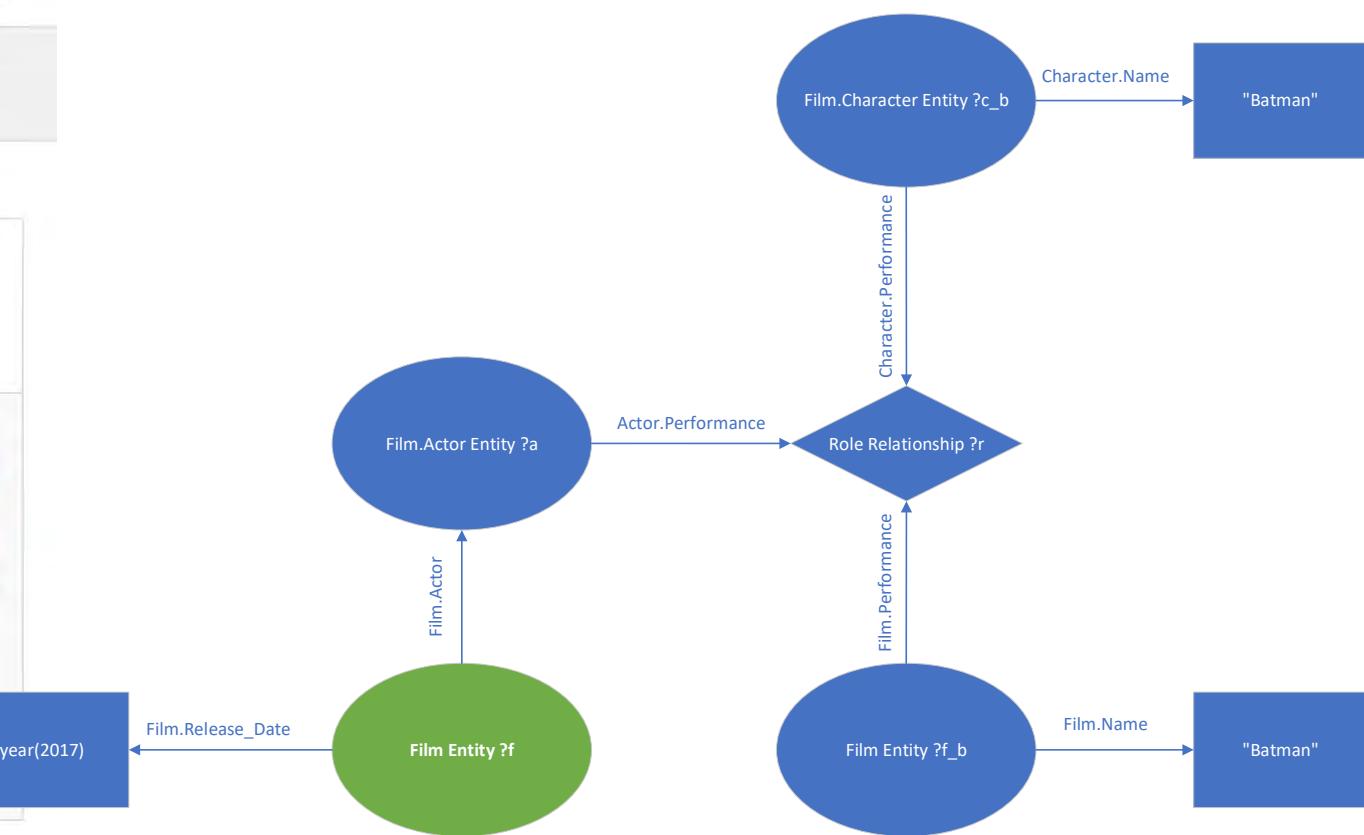
Microsoft Sign in to see results from Microsoft >

12,900,000 Results Any time ▾

2017 Movies starring Michael Keaton who played  ...  
Characters named Batman who acted in Batman (1989)

**Spider-Man: Homecoming**   
**American Assassin** 2017 · Suspense   
**The Founder** 2017 · Biography 

year(2017) 



# 1.1 Generic Semantic Parsing

- In this approach as in the example provided by [Kwiatkoski 13], we:
  1. Perform a generic semantic parsing of the utterances
  2. Perform ontology matching on relationships
- For e.g. {who is Donald Trump's Daughter}
  1.  $\lambda x. daughter\_of(Donald\ Trump, x)$
  2.  $\lambda x. child\_of(Donald\ Trump, x) \wedge gender(x, female)$
- This semantic expression can be then compiled into a knowledge graph database query e.g. SPARQL and executed to return the results

# Dependency parsers: Arc-standard [Nivre 2004]

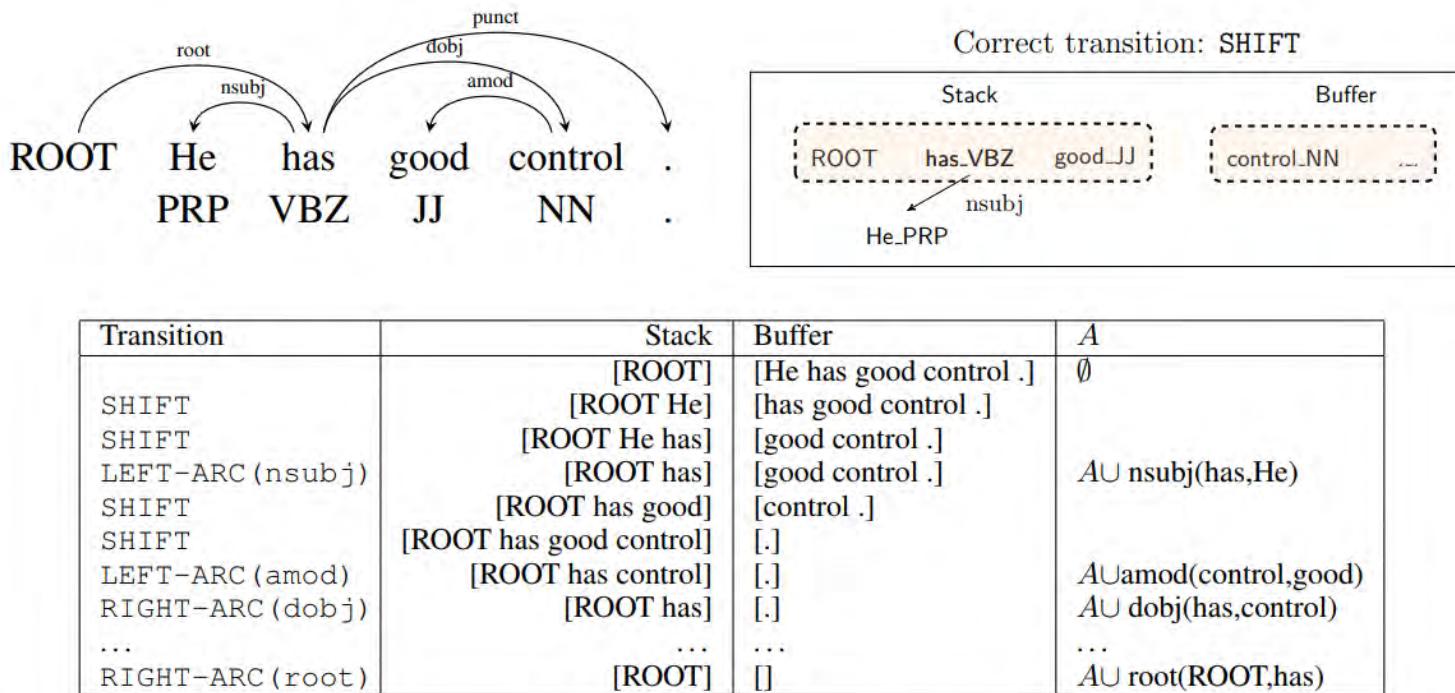


Figure 1: An example of transition-based dependency parsing. Above left: a desired dependency tree, above right: an intermediate configuration, bottom: a transition sequence of the arc-standard system.

Arc-standard actions are then learned using for e.g. stack LSTM [Dyer 2015]

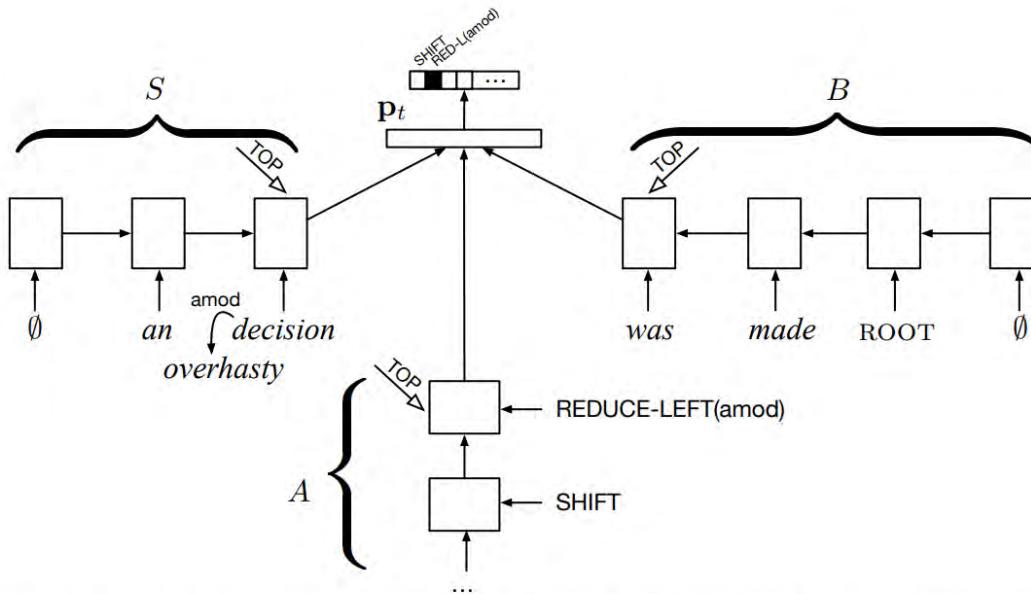
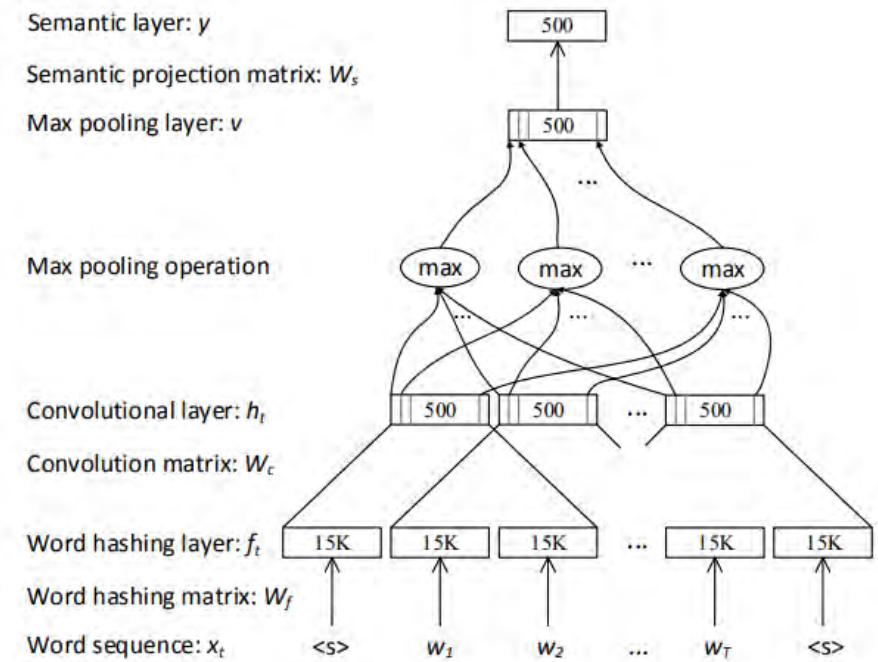
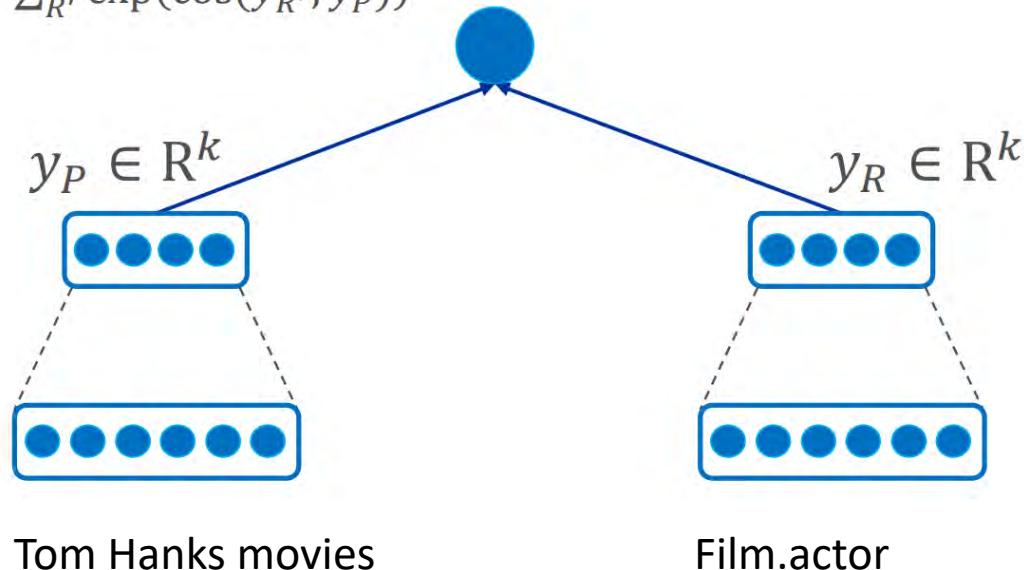


Figure 2: Parser state computation encountered while parsing the sentence "*an overhasty decision was made.*" Here  $S$  designates the stack of partially constructed dependency subtrees and its LSTM encoding;  $B$  is the buffer of words remaining to be processed and its LSTM encoding; and  $A$  is the stack representing the history of actions taken by the parser. These are linearly transformed, passed through a ReLU nonlinearity to produce the parser state embedding  $p_t$ . An affine transformation of this embedding is passed to a softmax layer to give a distribution over parsing decisions that can be taken.

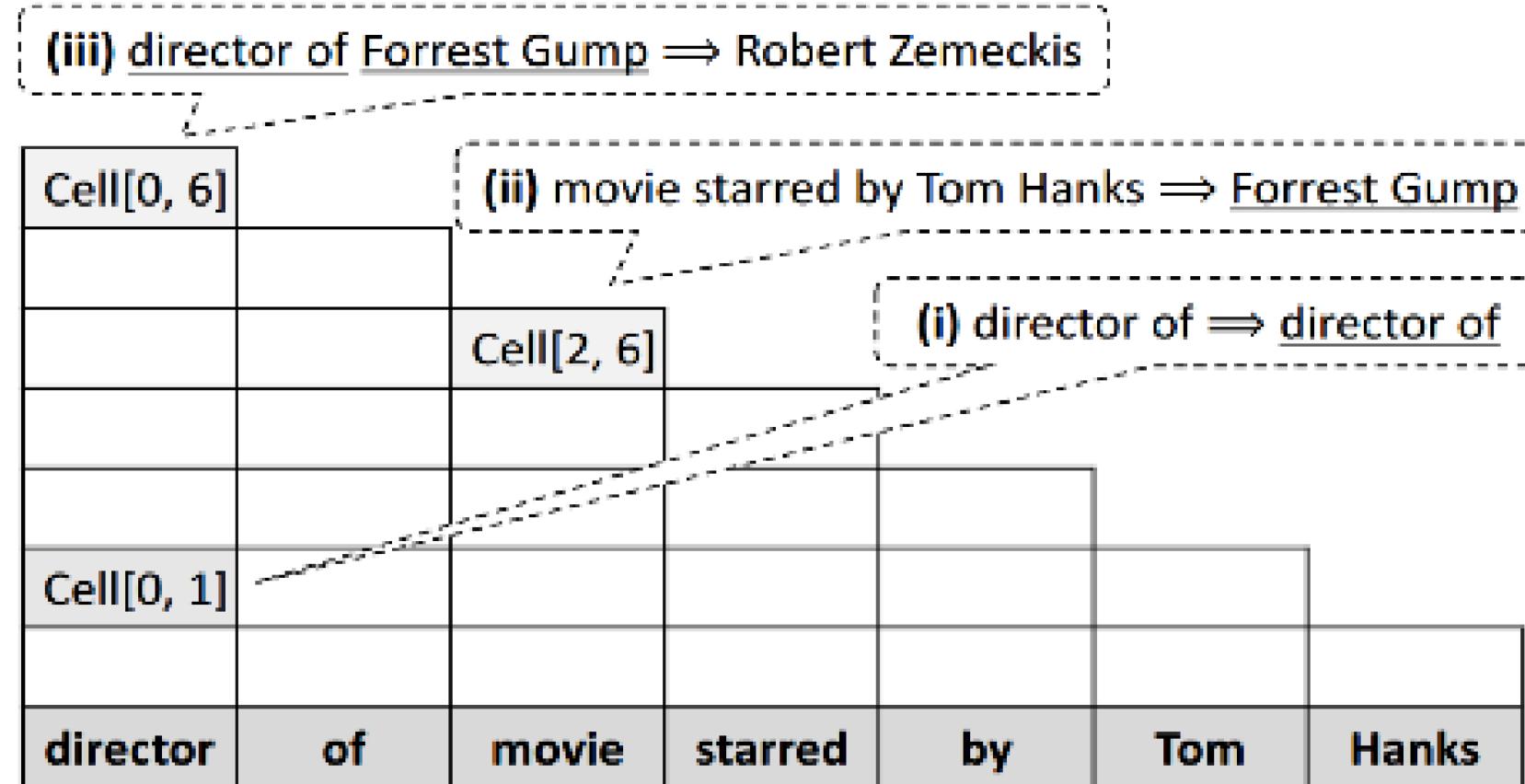
# Ontology Matching on Relationships using DSSM [Shen+ 14]

- Input is mapped into two  $k$  dimensional vectors
- Probability is determined by softmax of their cosine similarity

$$P(R|P) = \frac{\exp(\cos(y_R, y_P))}{\sum_{R'} \exp(\cos(y_{R'}, y_P))}$$



# 1.2 Knowledge base specific semantic parsing



Constituency parsers:  
PCFG Chart Parsing

Grammar is learned  
independently from an  
annotated dataset

Fig.1 of [Bao et al., 2014]

# 1.3 Knowledge Embedding for e.g. [Bordes 2014]

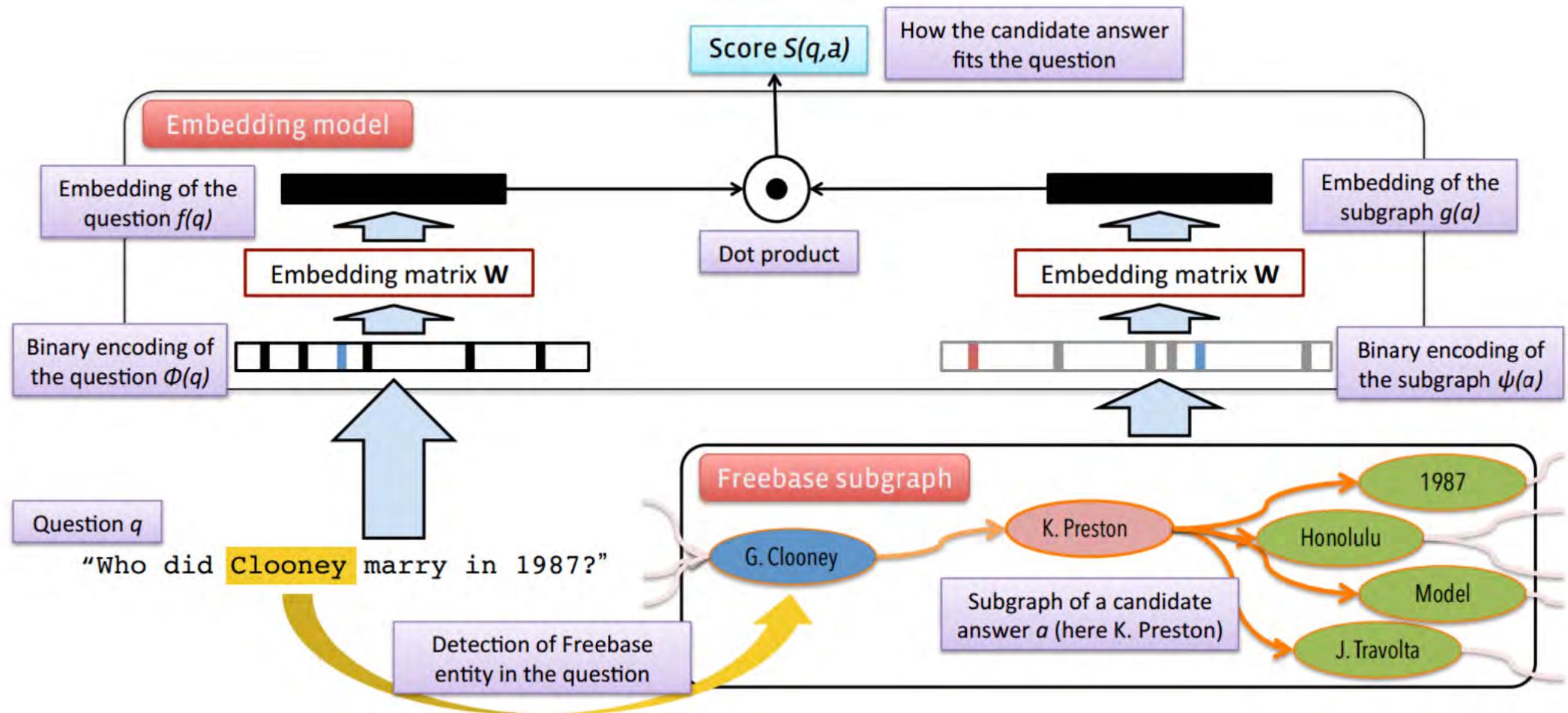
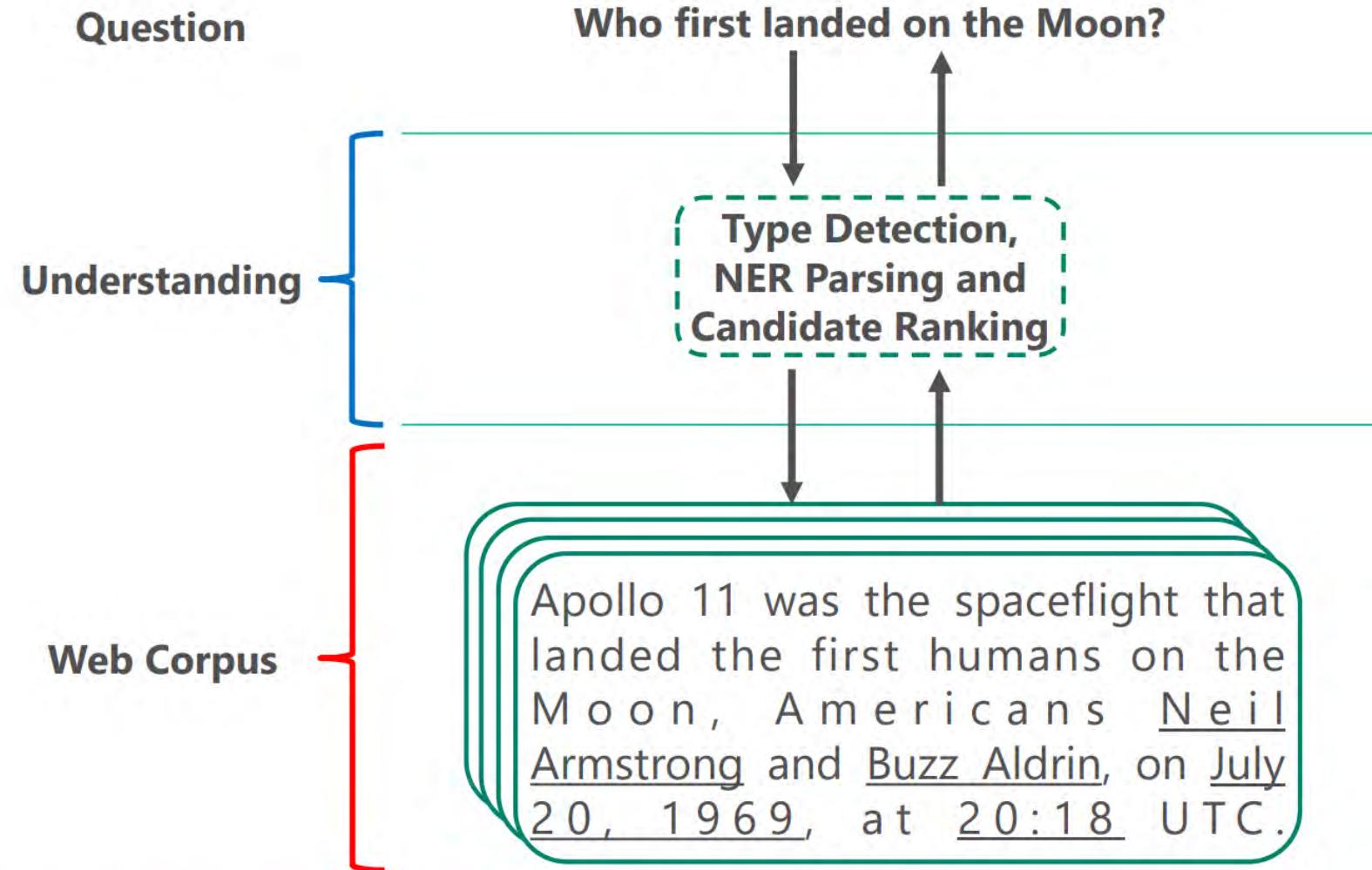


Fig. 1 of [Bordes et al., 2014] 144

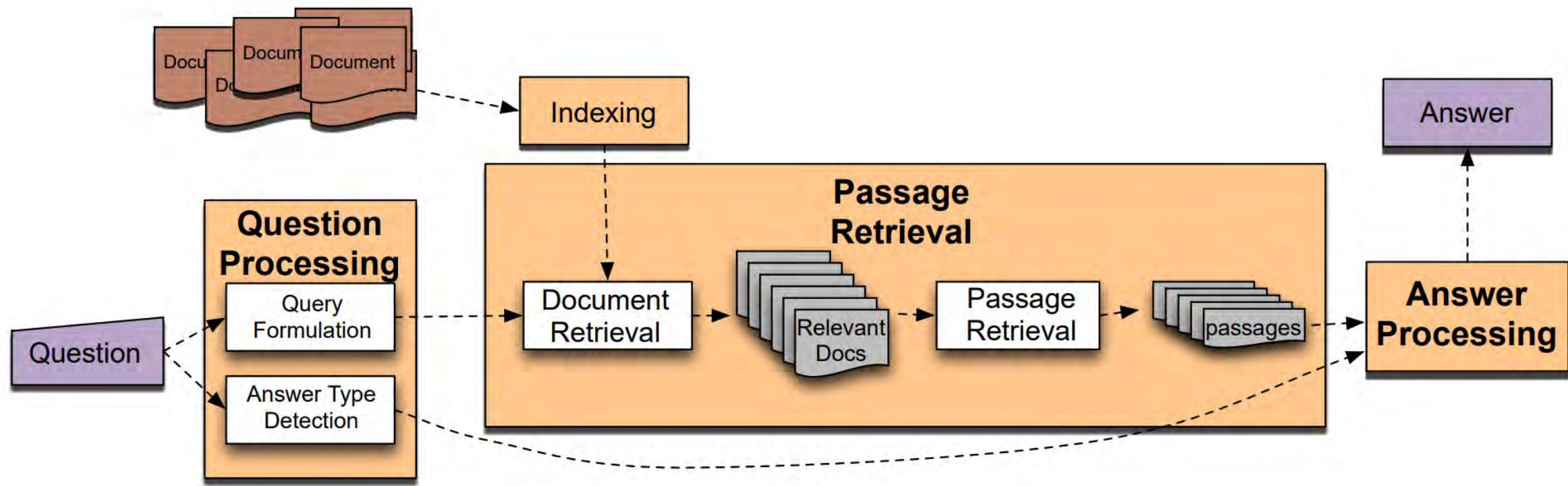
## 2. Information extraction approaches

- Extracting answers on the fly.
- These approaches provide ways to leverage the knowledge graph in cases where the question cannot be covered by the ontology or the data or both.

# Information extraction approaches



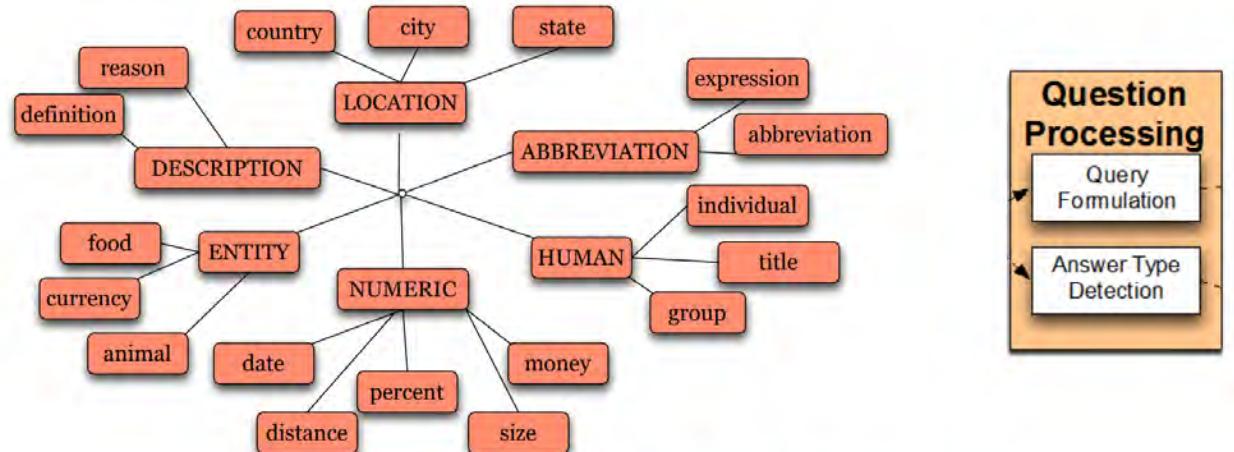
# Information Extraction Approach



Question Answering [Dan Jurafsky, Stanford]

# Answer Type Detection

- Who first landed on the moon => Person
- Where is the headquarters of Microsoft => Location
- What is the largest country in population => Country
- Highest flying bird => Animal/Bird



Learning Question Classifiers [Xin Li & Dan Roth, COLING 2002]

Question Answering [Dan Jurafsky, Stanford]

# Answer Type Detection

- Rules:
  - Grammar for e.g. who be/... => Person
  - Head word for e.g. which **city** is the largest
- Learned type classifier e.g. SVM utilizing features like question words, phrases, POS tags, headwords, mentioned entities, ...etc [Dan Jurafsky]

# Passage Retrieval

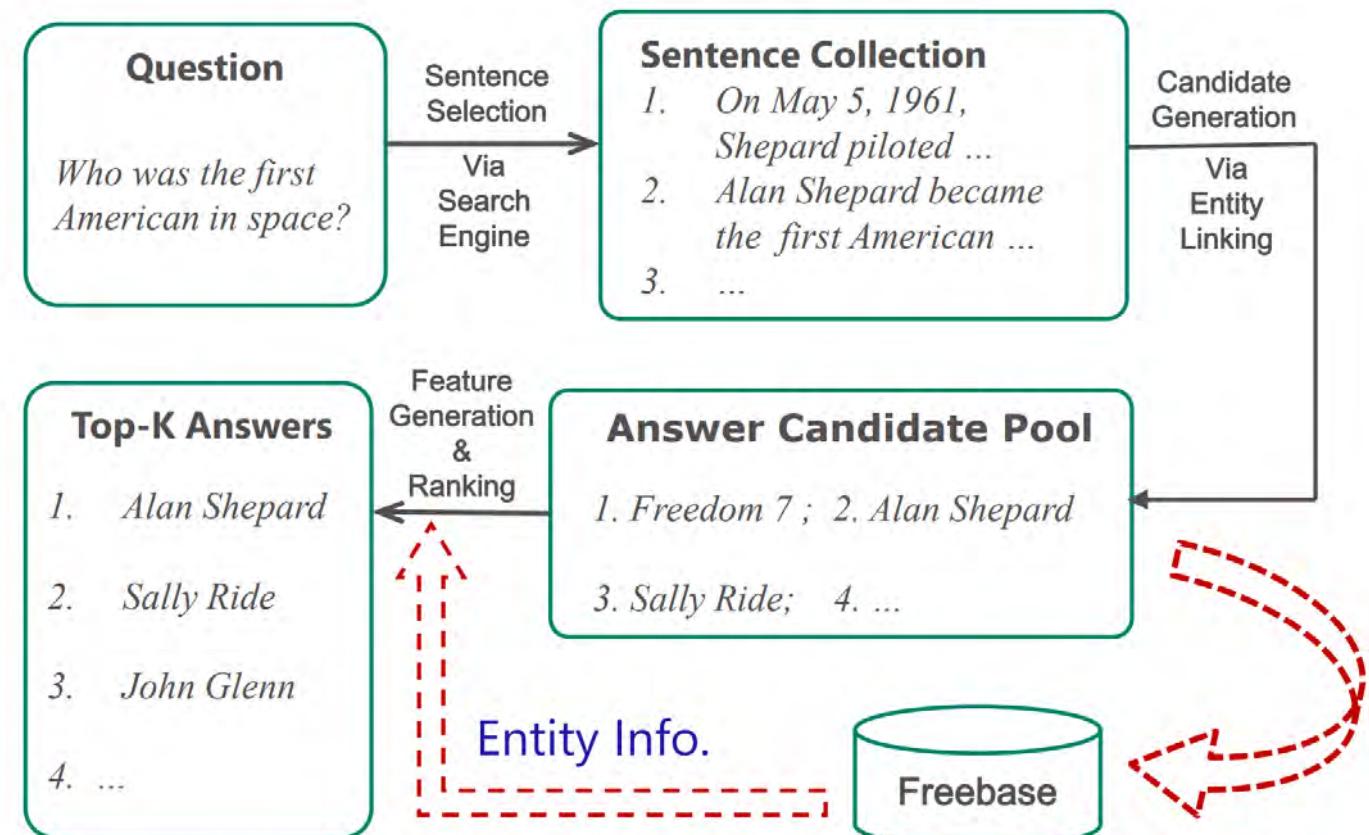
- Retrieve documents using expanded query terms + search engine
- Segment the documents into smaller units e.g. passages/paragraphs
- Rank passages using learned model utilizing features like:
  - Number of named entities of the right type in the passage
  - Number of query words in the passage
  - Number of question n-grams in the passage
  - Proximity of query words in the passage
  - Longest sequence of question words
  - Rank of document containing passage,...etc

# Process Answer

- Detect answer entity by running NER on the passage
- Mark the answer entity in the passage
- How many bones in an adult human body? (**Number**)
  - The human skeleton is the internal framework of the body. It is composed of 270 bones at birth – this total decreases to **206 bones** by adulthood after some bones have fused together.

# Answer Semantic Enrichment using KB [Huan Sun, et al., WWW 2015]

- 5-20% MRR improvement



Open Domain Question and Answering via Semantic Enrichment [Huan Sun, et al., WWW 2015]

# Serving Knowledge Through Dialogs

- Approaches:
  - E2E Seq2seq (Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015)
  - Knowledge based ontological slot filling (Dai+ 2017)
  - Knowledge grounded neural approaches (Ghazvininejad+ 2018)

# E2E Dialog Systems (e.g. Sordoni et al. 2015)

- Suitable for chitchat kind of bots.
- Predicted target sequences are usually free from facts

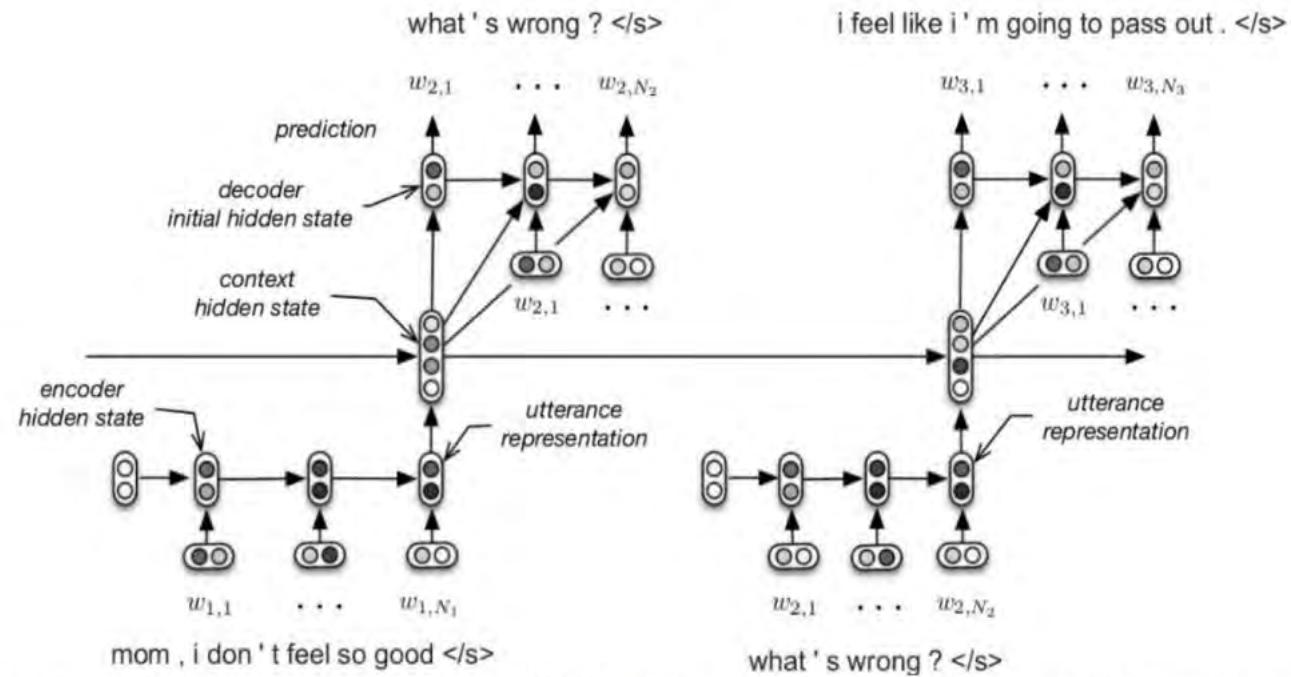
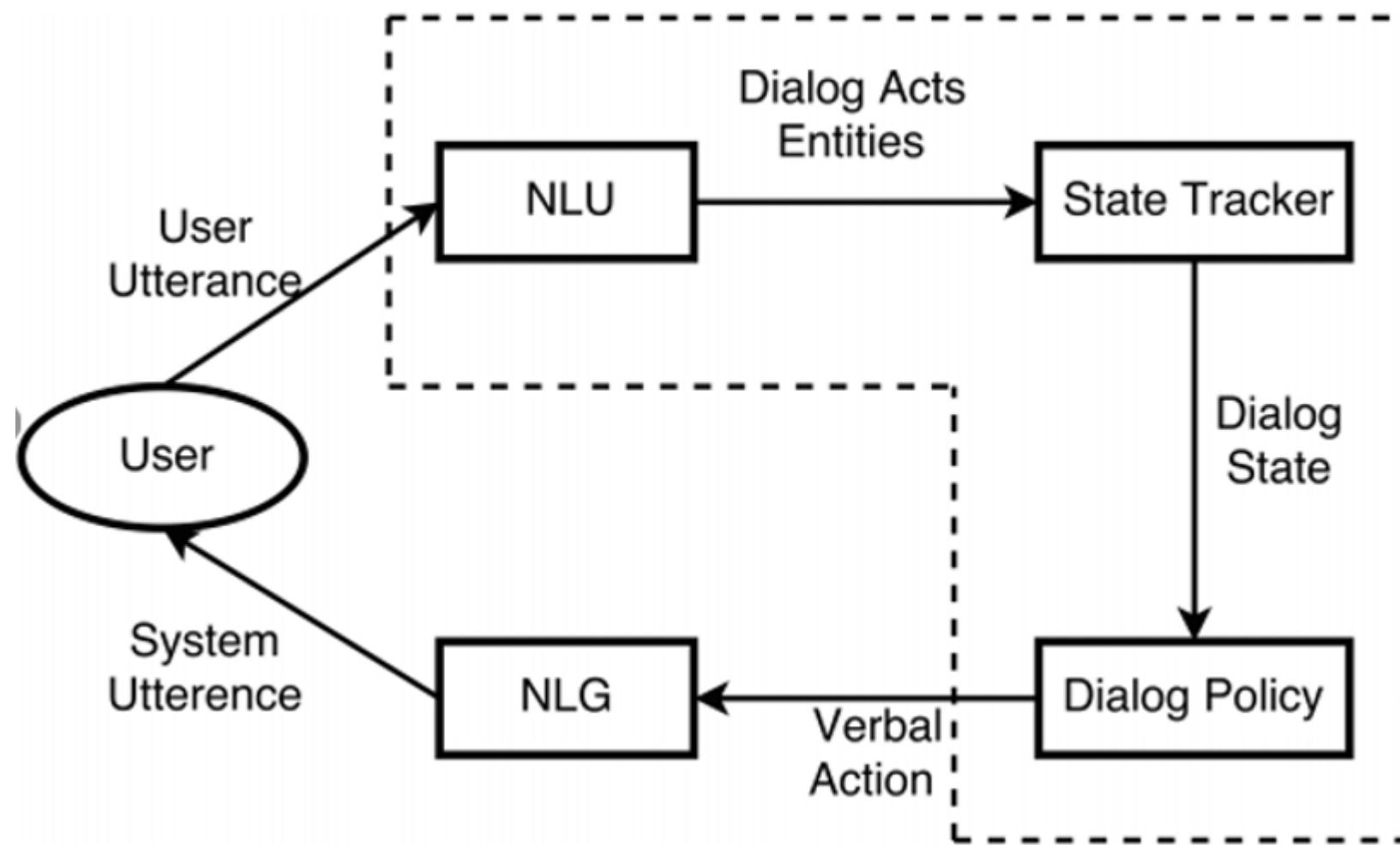


Figure 8: A computational graph representing the HRED architecture for dialogue over a span of three turns. The major addition to the architecture is a higher-level *context-RNN* keeping track of past utterances by progressively processing over time each utterance vector and conditioning the decoding on the last hidden state of the context vector (middle).

# Knowledge Based Ontological Slot Filling



# Knowledge Grounded Neural Approaches e.g. [Ghazvininejad+ 2018]

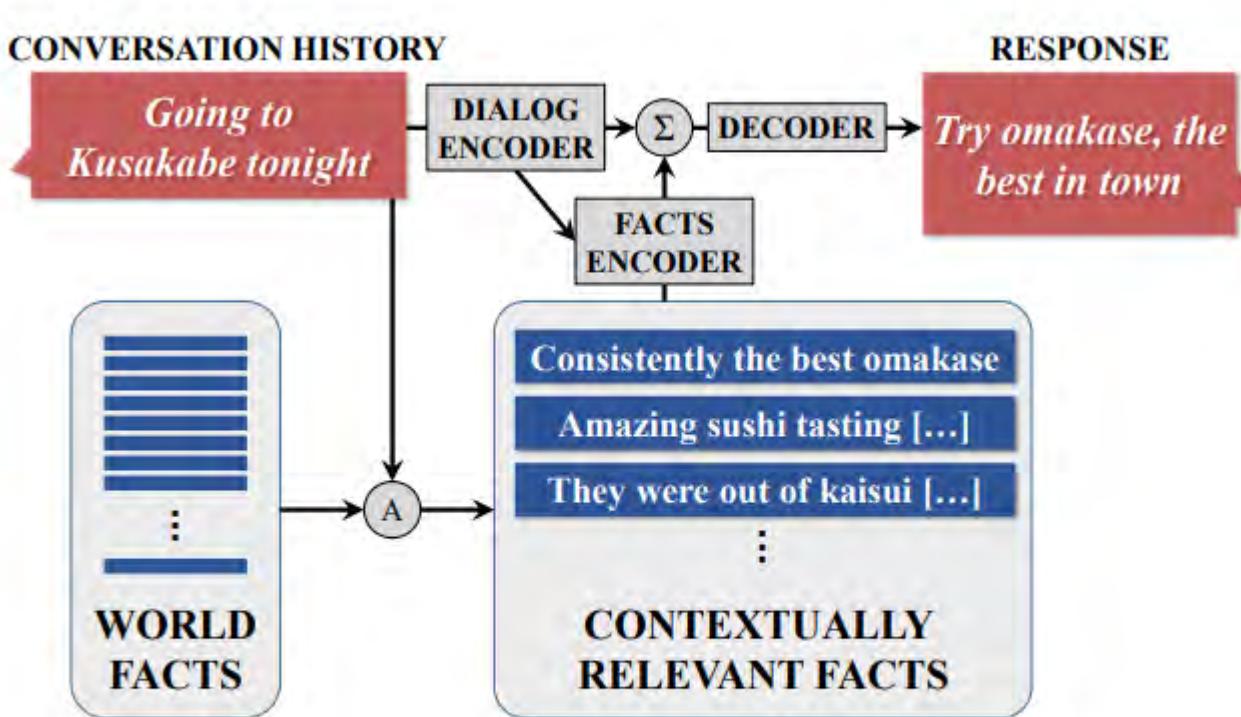


Figure 3: Knowledge-grounded model architecture.

# Enterprise Scenarios

- All the challenges mentioned previously plus the following:
- Compliance
- Different data formats: databases, emails, chat logs, discussion forums, web blogs, pdfs, PowerPoint/Word/Excel documents etc.
- Different schemas: schema mapping and merging, and new schema discovery.
- Consumption via dialog systems, search interface, mobile devices or other modalities, API.
- Highly domain-specific models required, bootstrapped by pre-trained models. Need on-prem domain-adaptation.

# Questions

# Closing