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

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

Microsoft

What is Knowledge

- Plato’s definition: Justified true belief

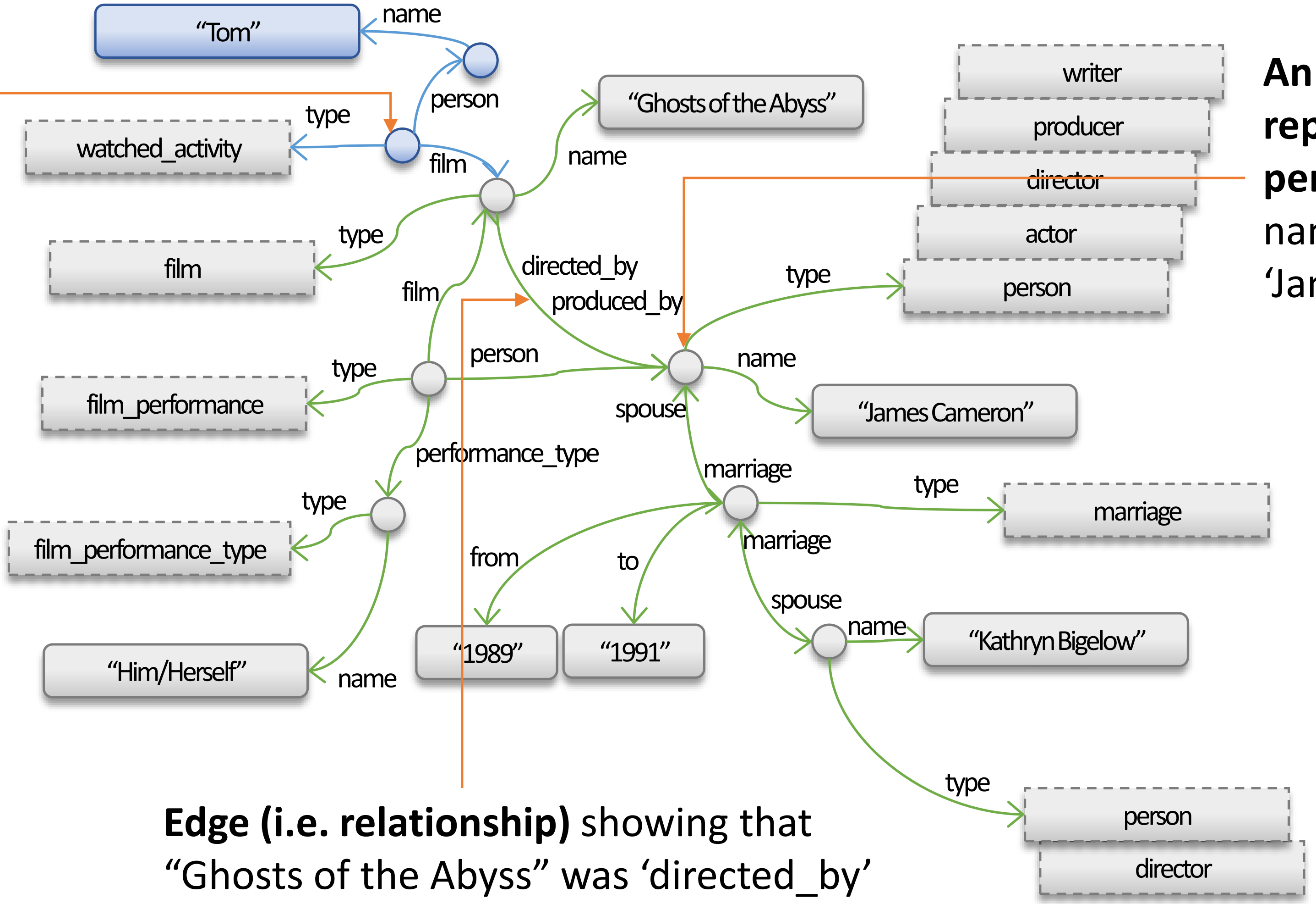
Mission:

- 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



An entity representing a person with name **attribute** 'James Cameron'

Edge (i.e. relationship) showing that "Ghosts of the Abyss" was 'directed_by' and 'produced_by' James Cameron

Key concepts

Entity	Represent something in the real world
Edge	Represent relationship
Attribute	Represent something about an entity
Ontology	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 production KGs support Google and Bing Search

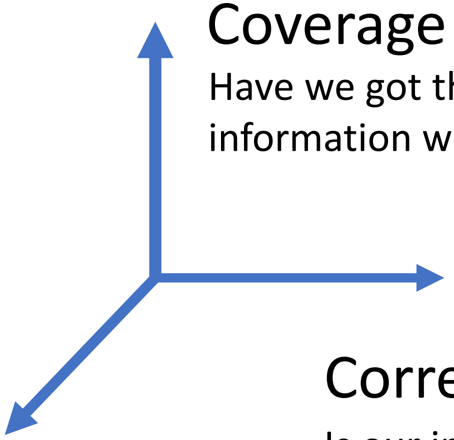
Large vertical KGs
- Facebook (social network)
- LinkedIn (people graph)
- Amazon (product graph)

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 disambiguition
  - Fact extraction and verification

Challenges of scaled KGs

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



Coverage
Have we got the information we need?

Freshness
Is information up to date?

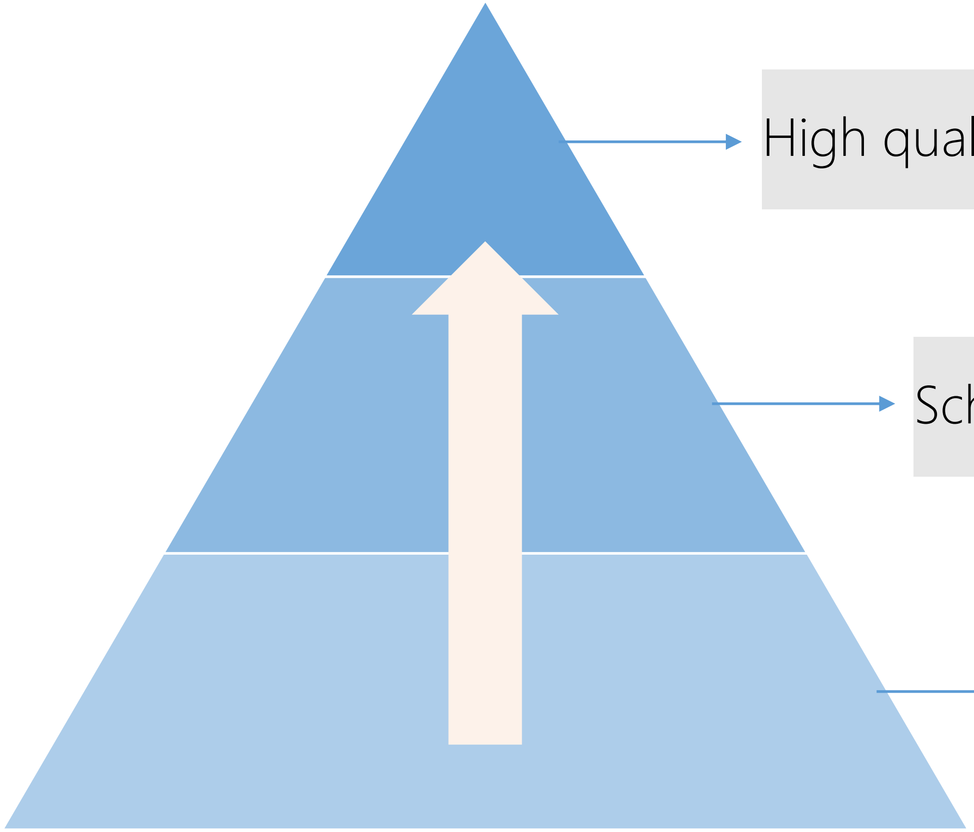
Correctness
Is our information accurate?

Three forces in constant conflict:

Increased freshness and coverage	Harder to ensure correctness
Increased correctness	Harder to ensure freshness and coverage
Correctness is *always* hard – what is true and correct? Particularly critical in today's world	

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

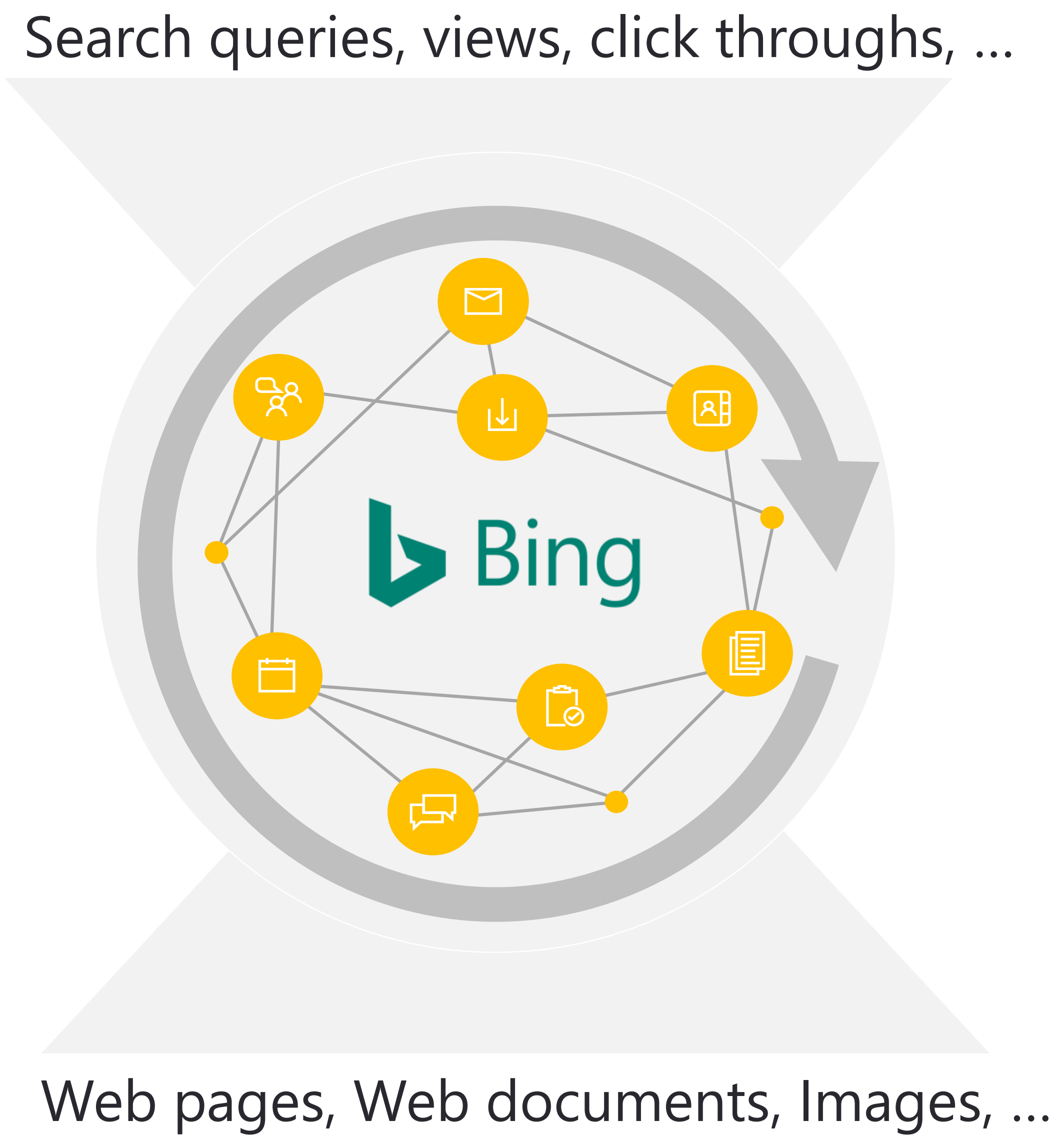


High quality, conflated and schematized knowledge

Schematized, correlated data and relationships

Raw data sources, structured + unstructured

Knowledge flywheel in action: World graph



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, …

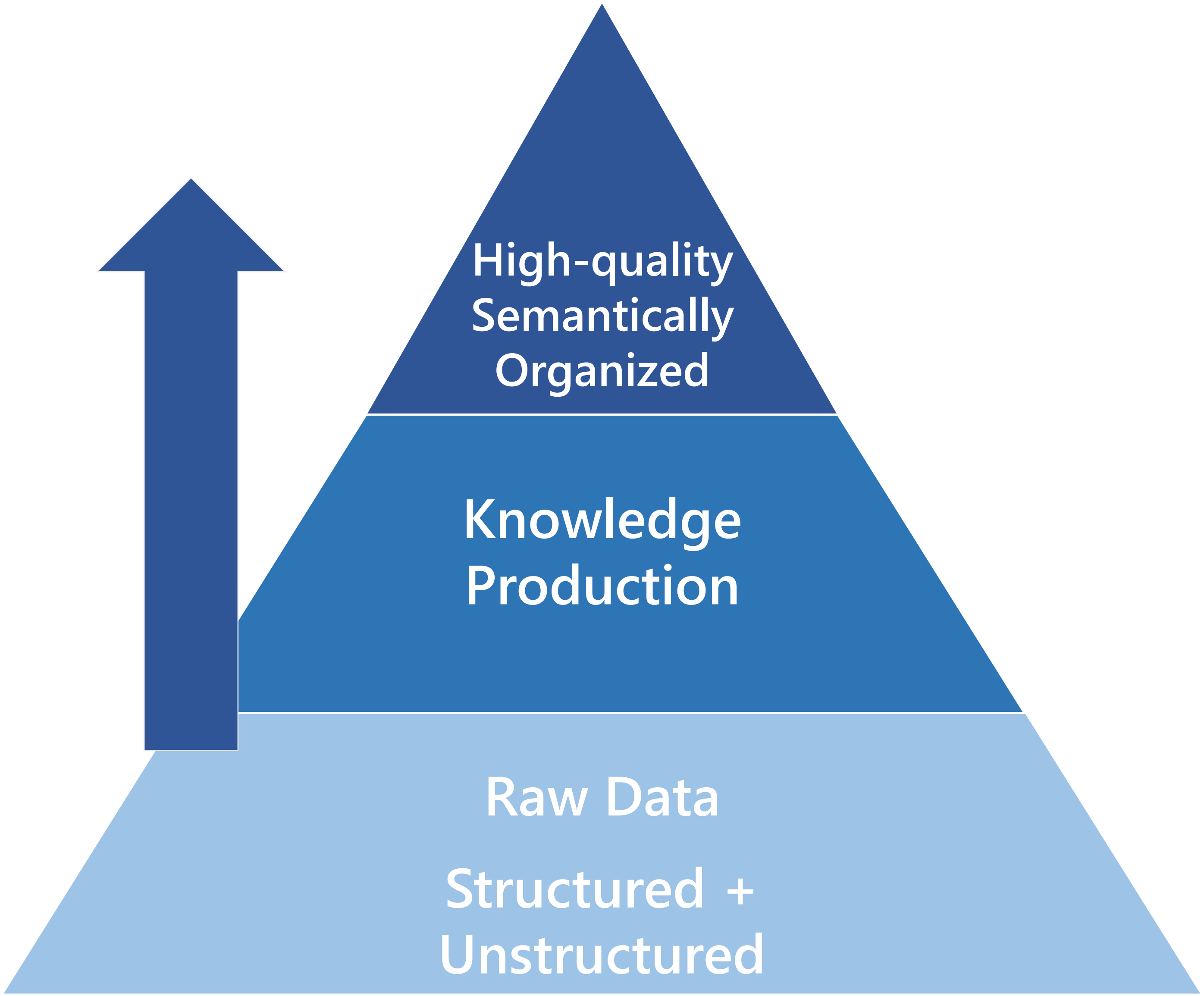
Emails, Messages, Documents, Meetings, …

Work graph

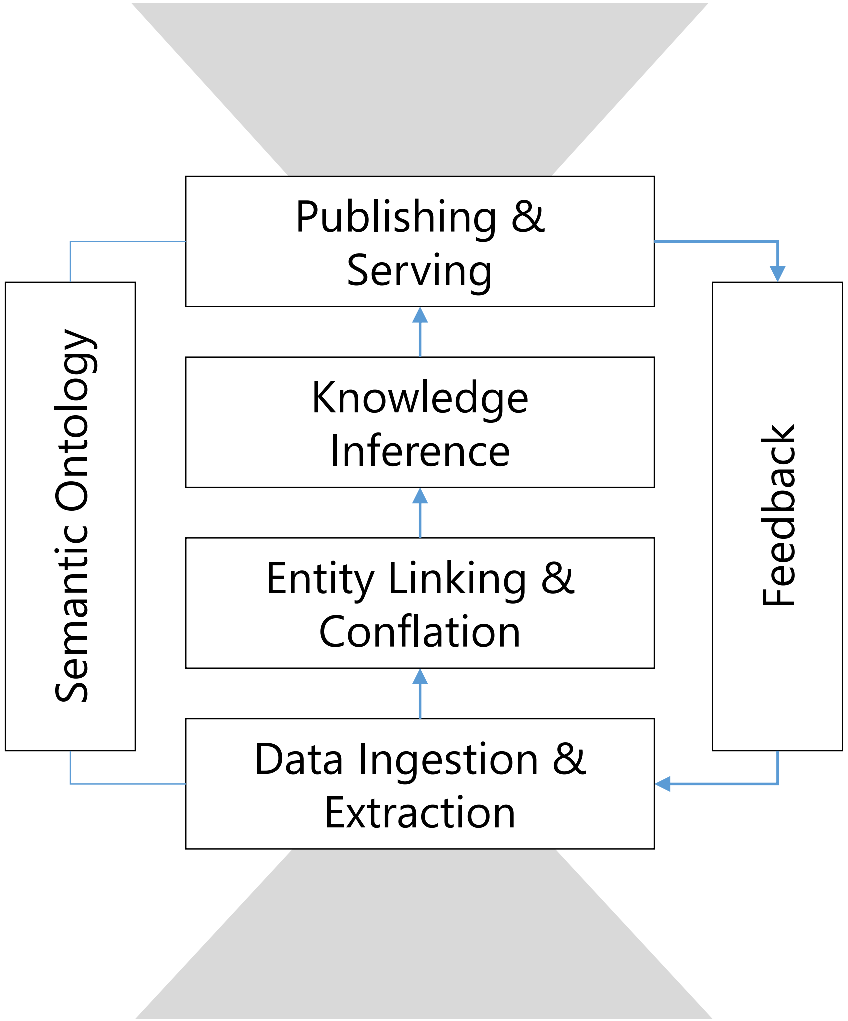
- People
- Groups
- Messages
- Activities

8T+ entities
240+ markets
44+ languages

How do we bring knowledge systems to life?



Active research and product efforts in knowledge



Unsupervised knowledge extraction from unstructured data in open domain

Knowledge graph semantic embedding

Autonomous knowledge inference & verification

Real-time knowledge graph with archiving

Large scale entity linking and disambiguation

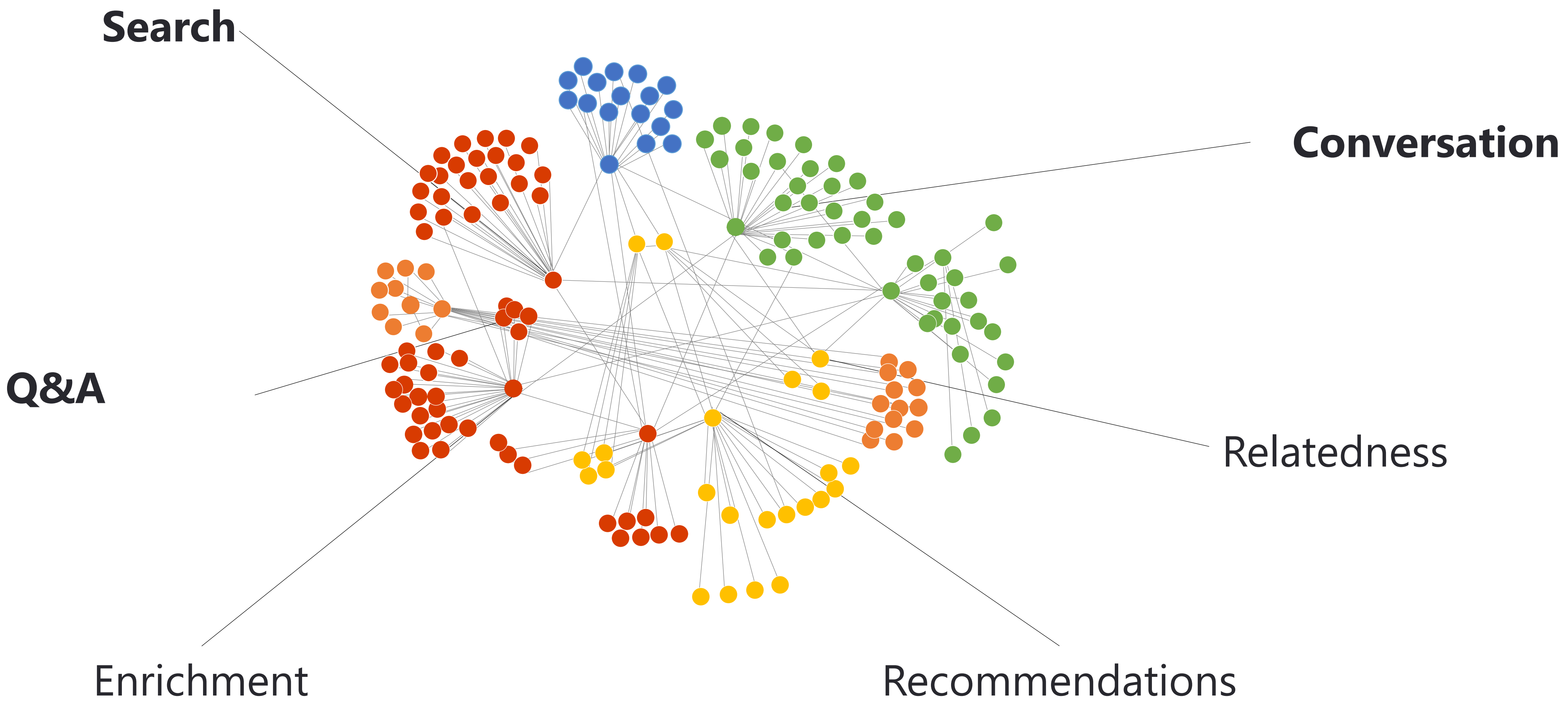
Ultra-scale knowledge representations

Knowledge system for multi-lingual

Knowledge Precision vs Comprehensiveness

…

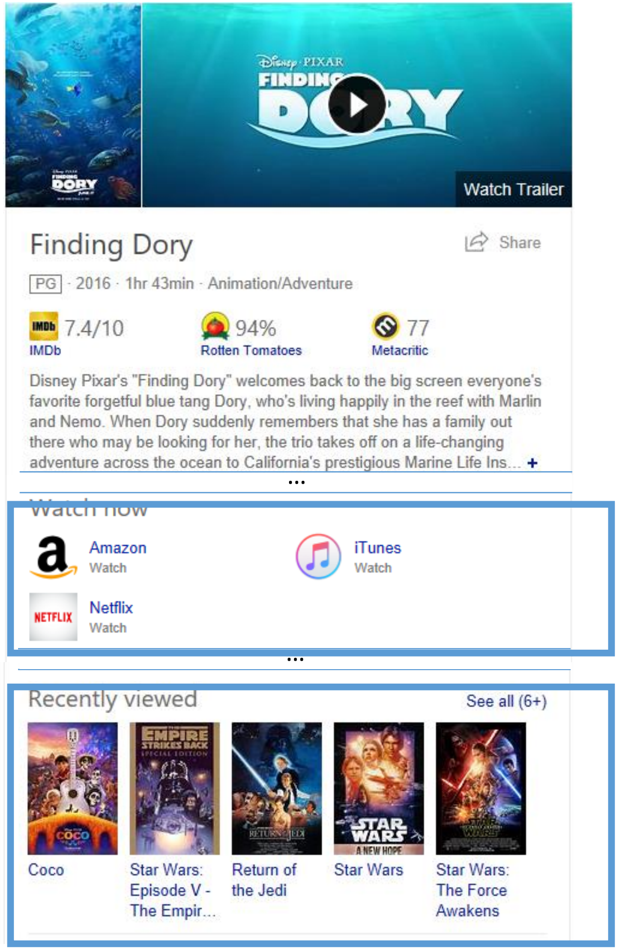
Infusing knowledge: From search to conversation



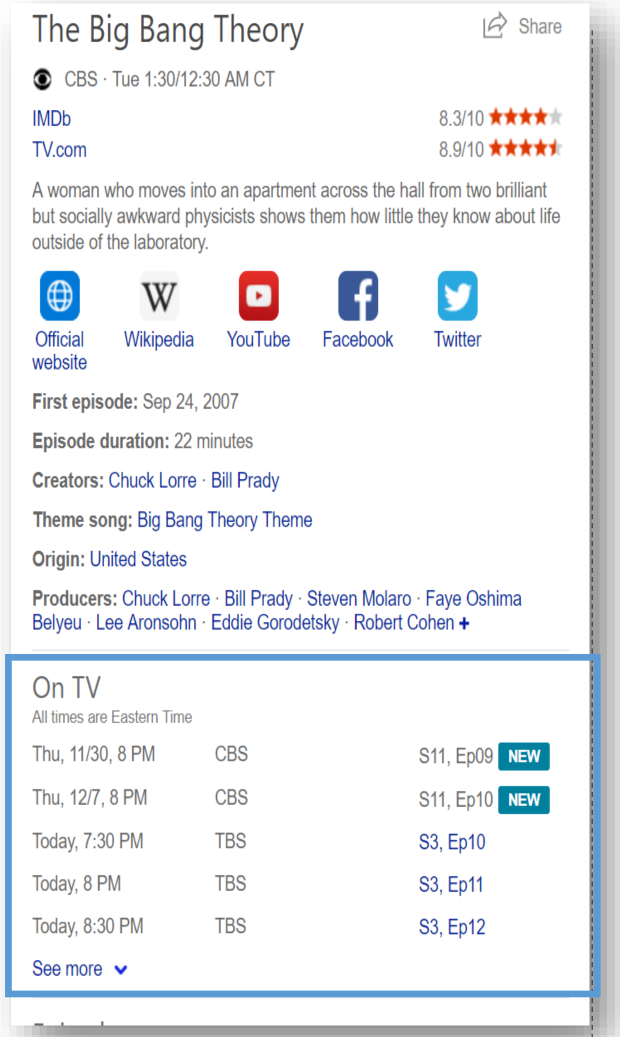
Satori powering Bing Search

Watch now actions for movie entities

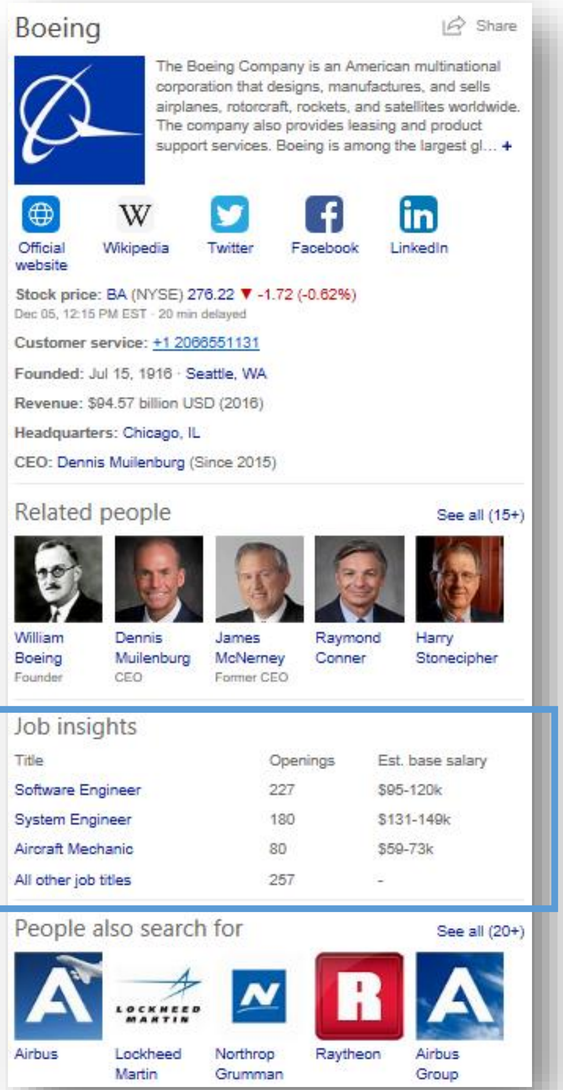
Recently viewed shows personal history



TV listings for TV show entities

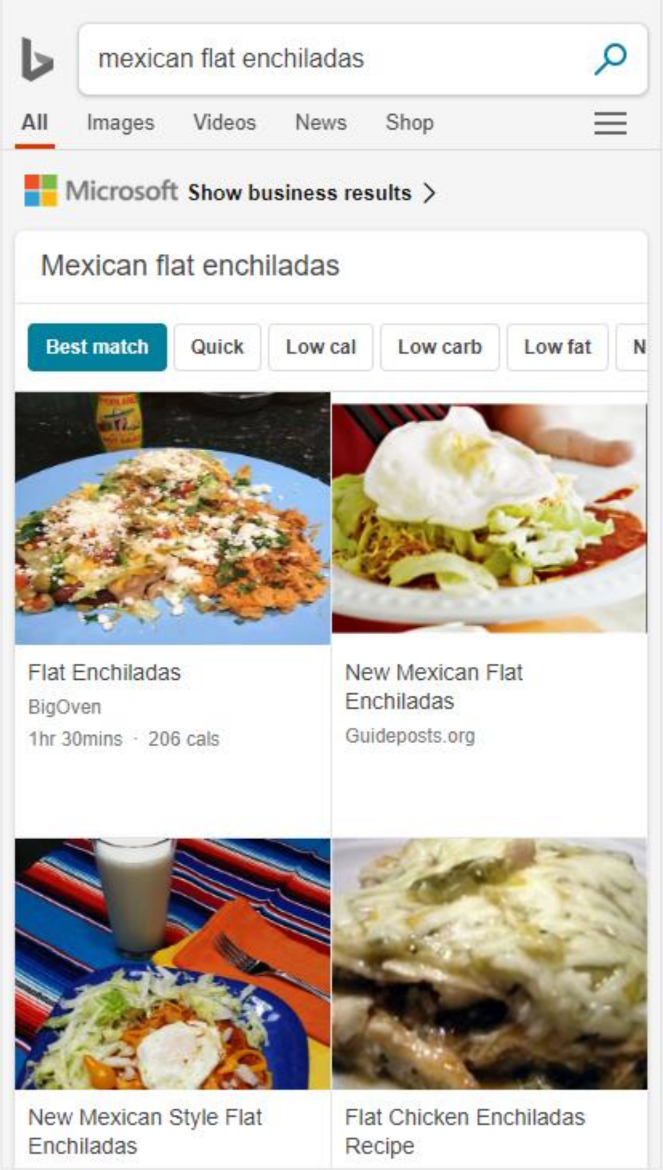


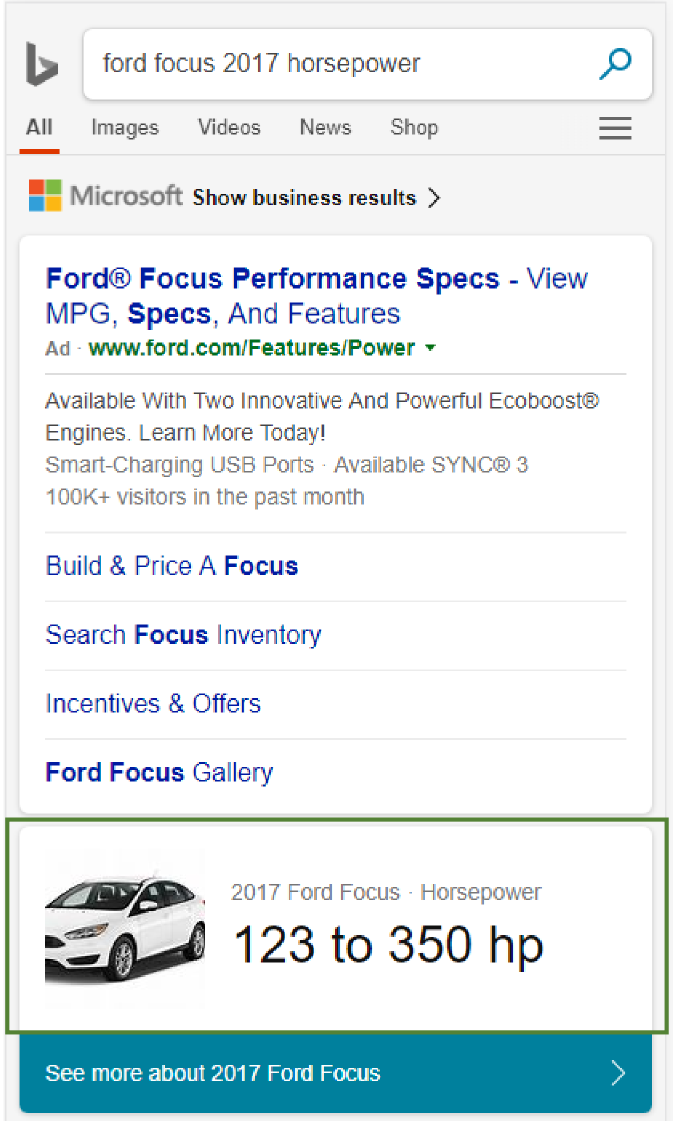
Job insights for companies



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

SBS +10.85 weak





Knowledge powered Q&A

Text-based Q&A

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/

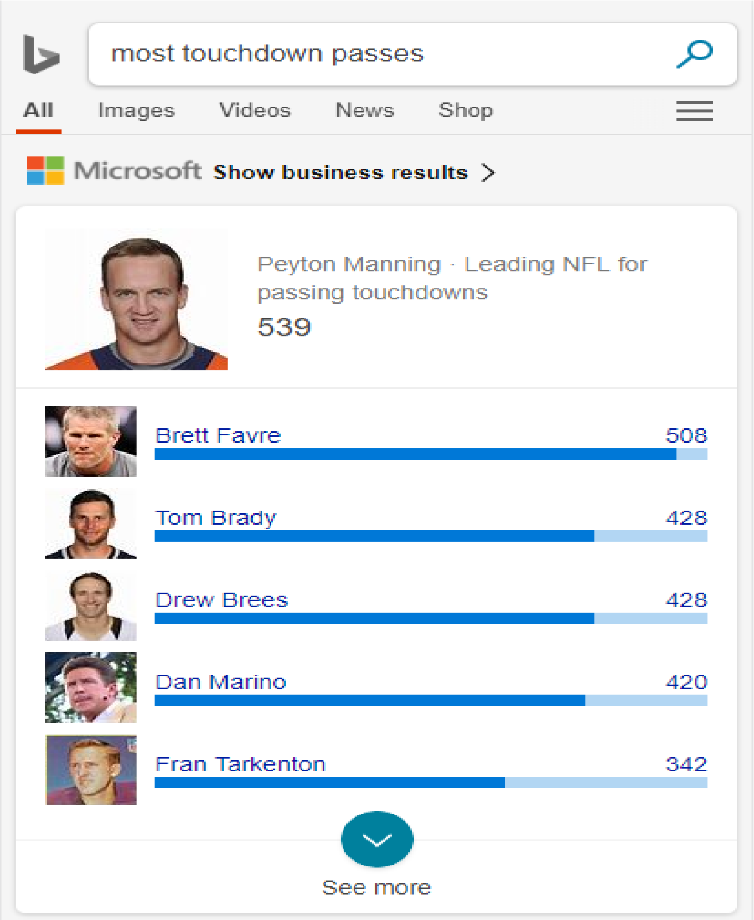
"You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans program. There are also grants, bursaries and scholarships available for both full-time and part-time students."

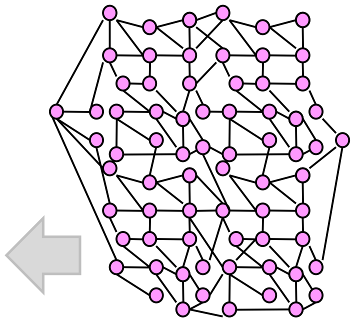
Source: http://www.campusaccess.com/financial-aid/osap.html

Answer

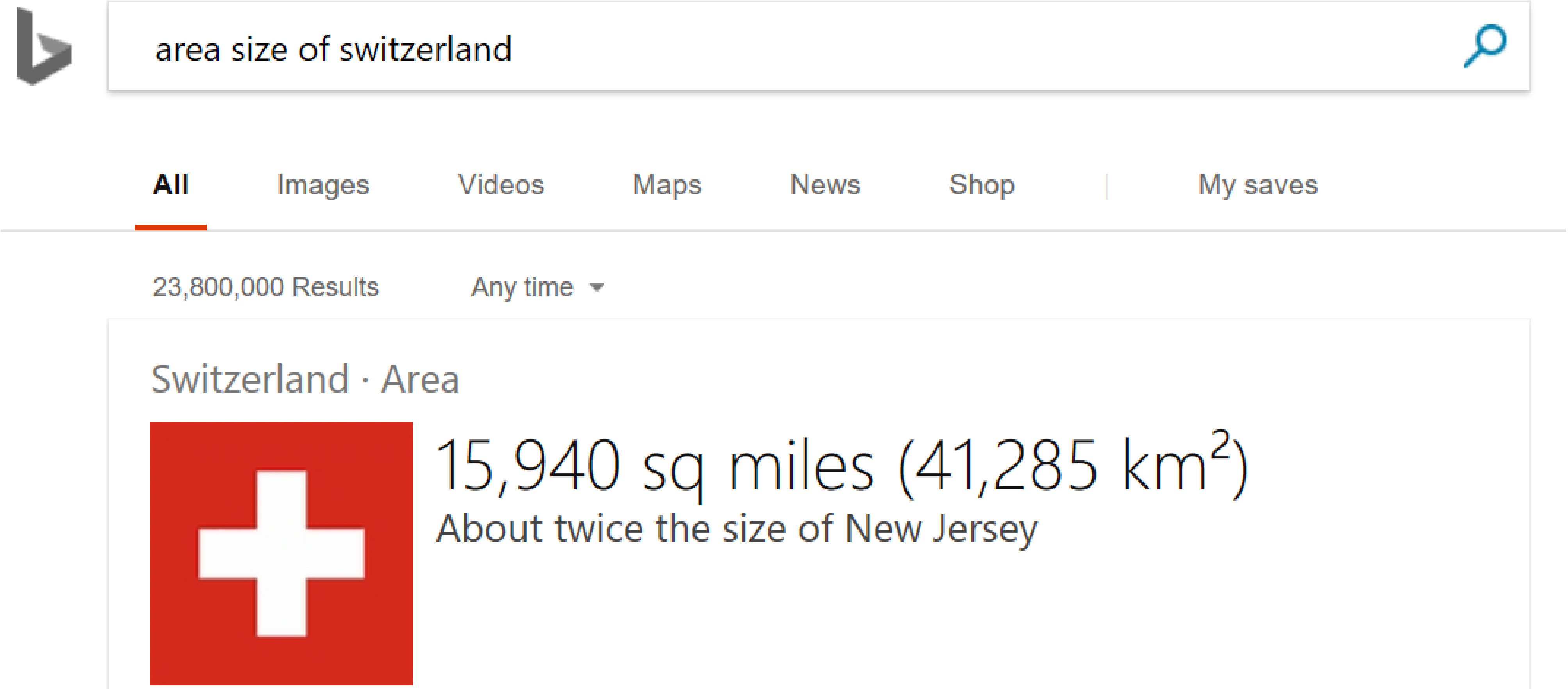
No. You won't qualify.

Knowledge-based Q&A





Bing – knowledge in answers



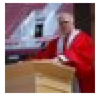
Knowledge graph serves NL fact answers

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 ▾

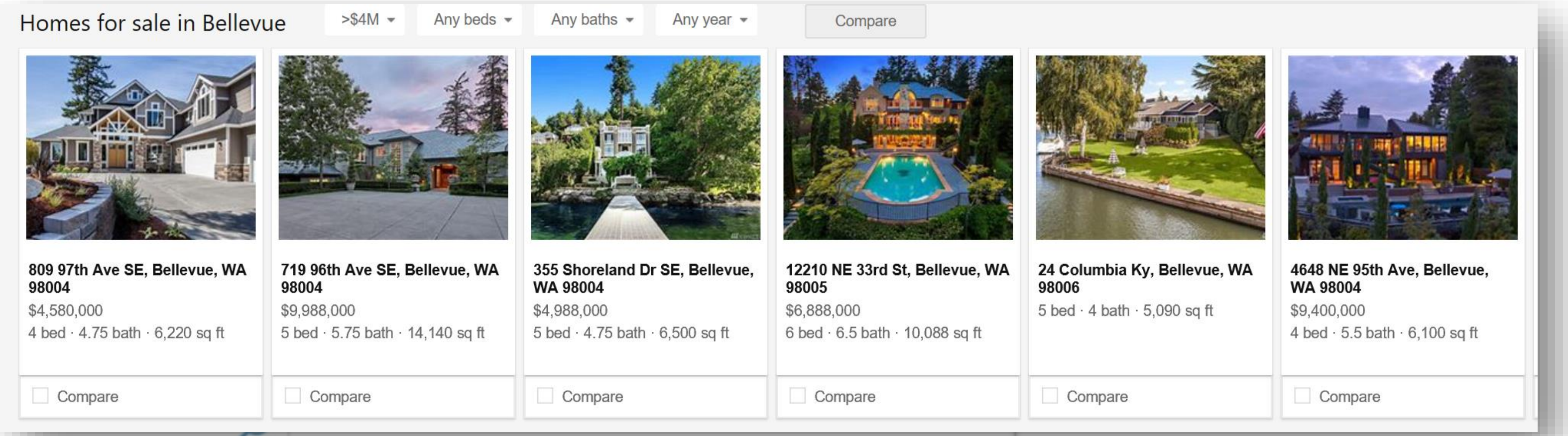
6

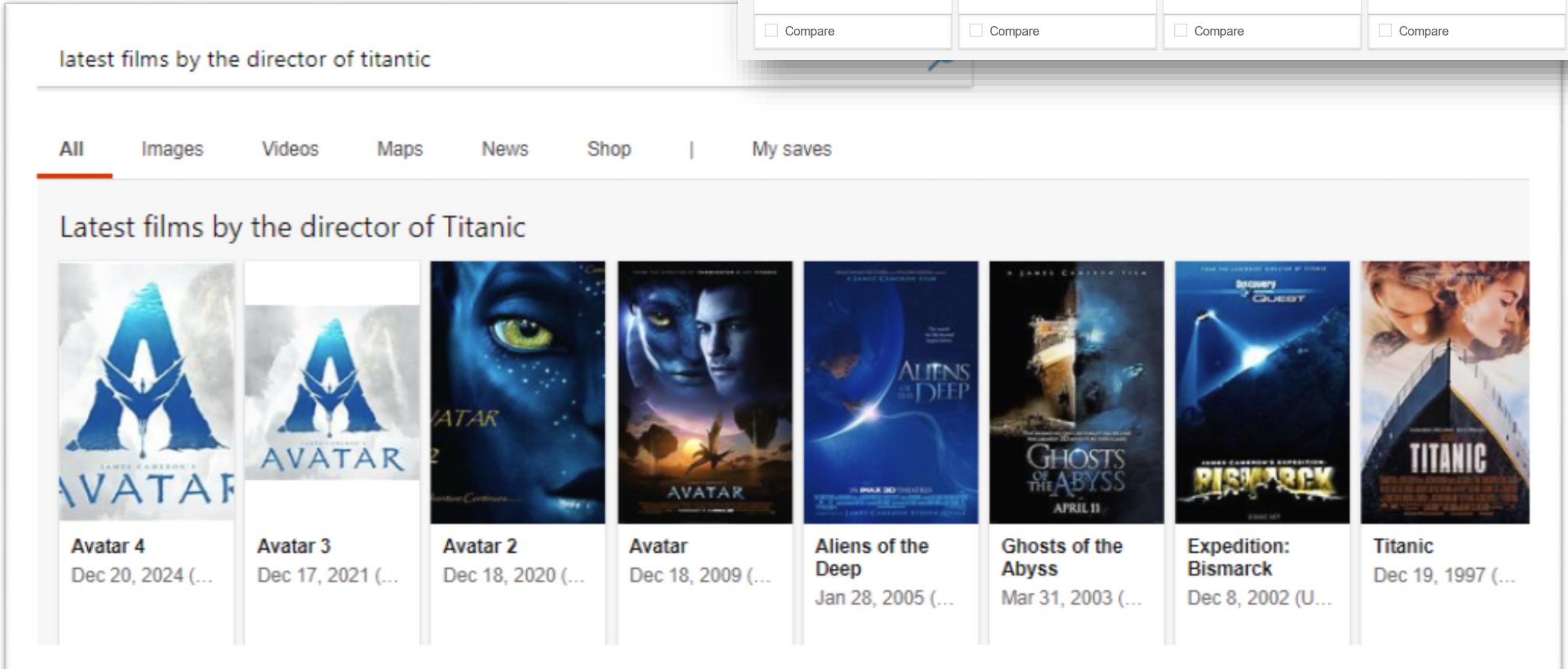
In the 2016 Olympics, Michael Phelps won 5 gold medals for 200m Butterfly, 200m Individual Medley, 4×100m Freestyle Relay, 4×100m Medley Relay, and 4×200m Freestyle Relay and 1 silver medal for 100m Butterfly, for a total of 6 medals overall.

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

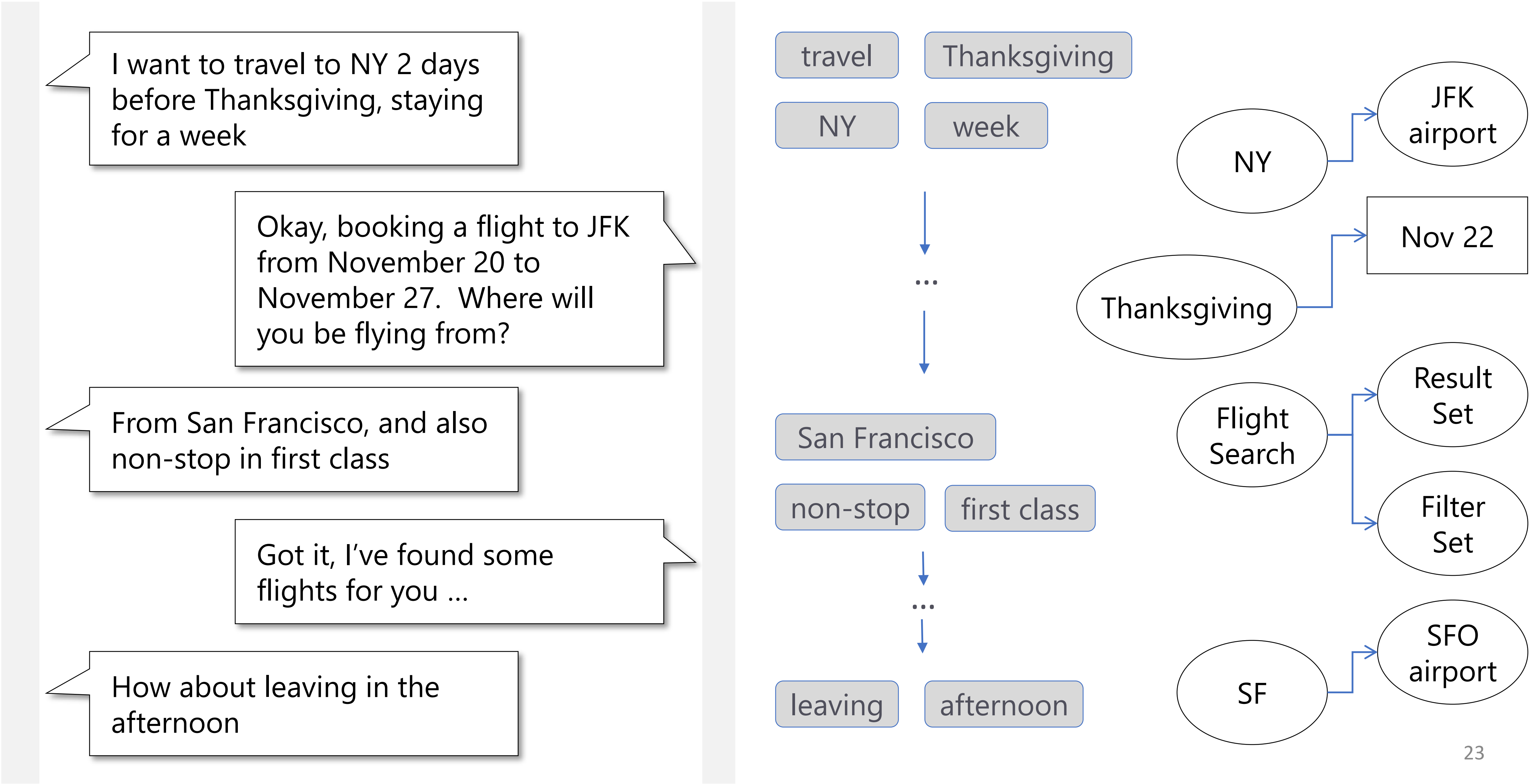
Learn more: en.wikipedia.org/wiki/Michael_Phelps   Is this answer helpful?

Knowledge graph serves carousel of information





Knowledge-powered Conversation



Part II: Acquiring Knowledge in the Wild

Benjamin Han

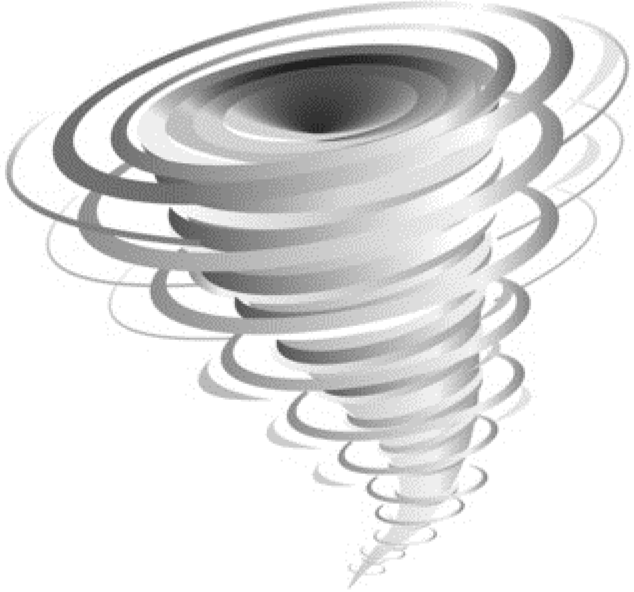
Principal Machine Learning & Data Scientist, Satori Group, Microsoft AI+R

diha@microsoft.com

Microsoft

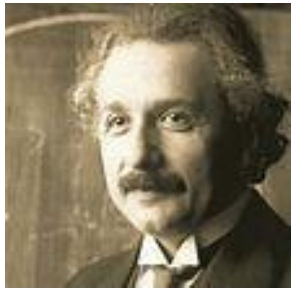
Goals

- 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).



awards_won →

award_recurring_event:Nobel Prize
category:Nobel Prize in Physics
ceremony:1921 Nobel Prize
notes_description:For his services to Theoretical Physics, and especially for his discovery of the law of the photoelectric effect
year:1921

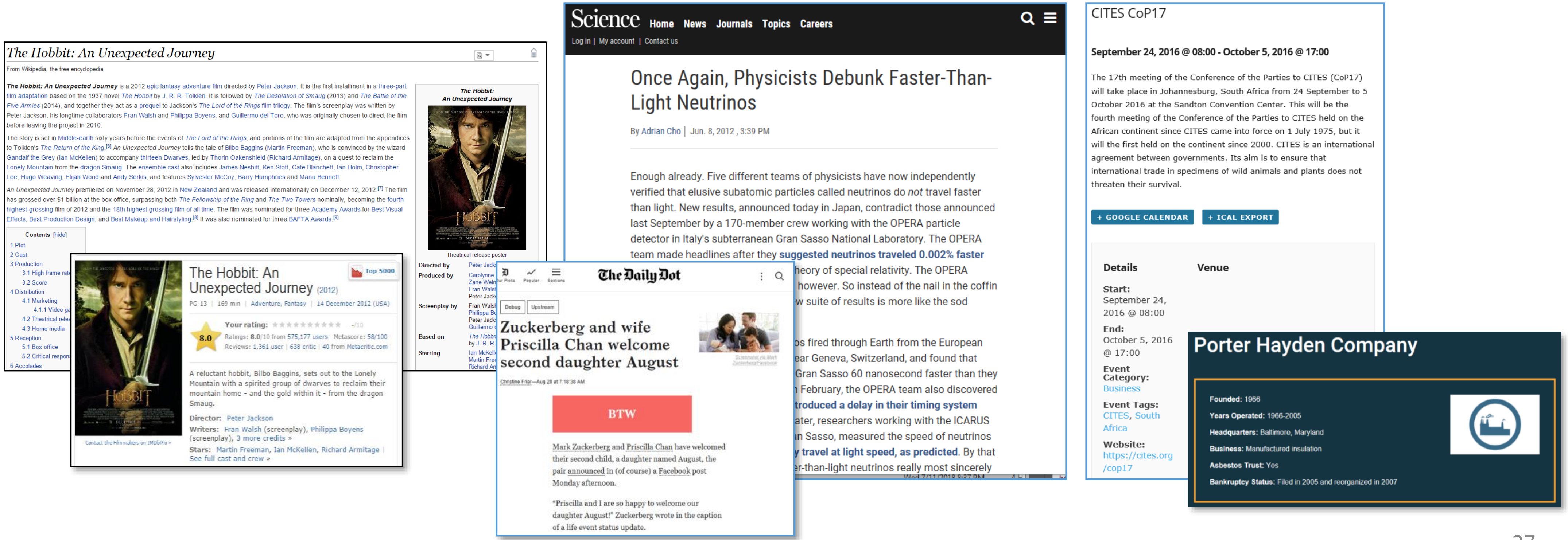
profession →

physicist

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

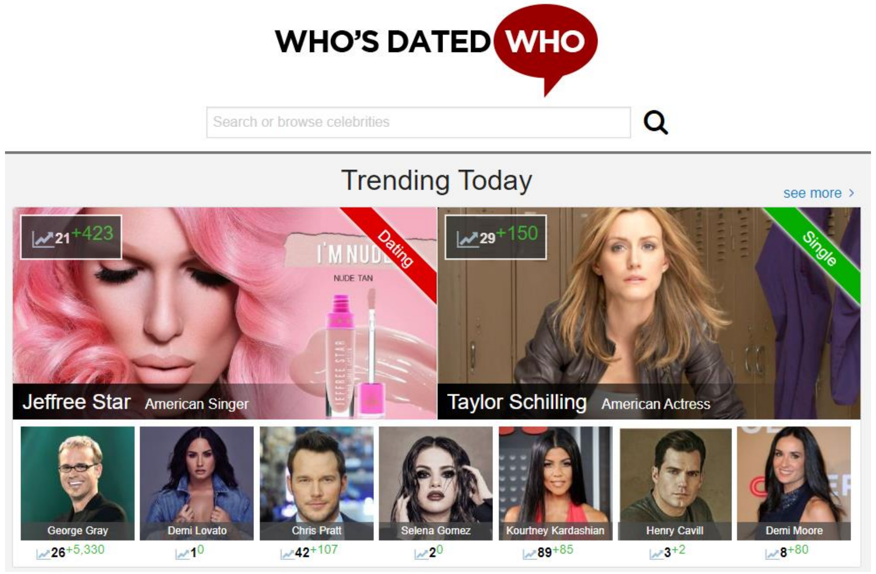
Knowledge Acquisition (KA) in the Wild

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



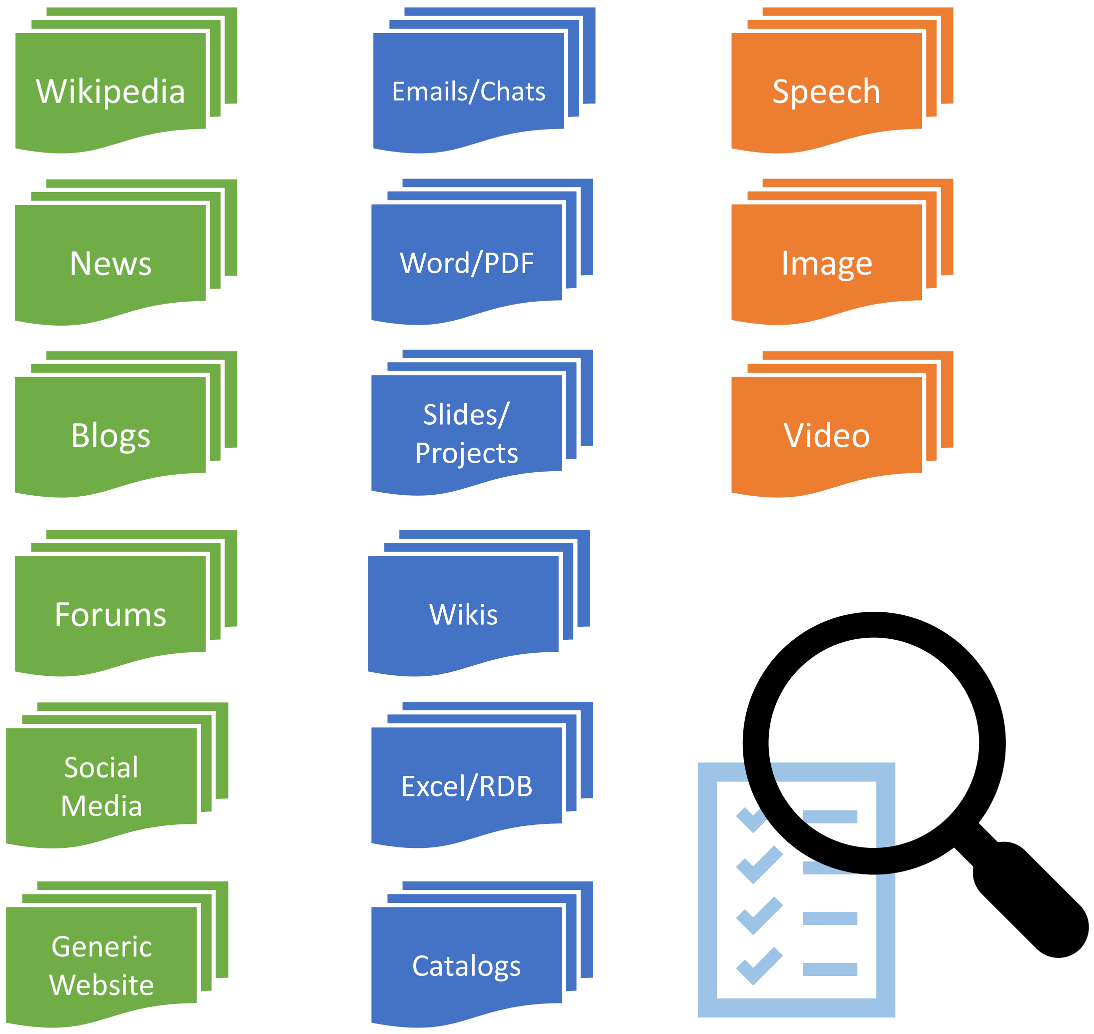
Knowledge Acquisition (KA) in the Wild

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



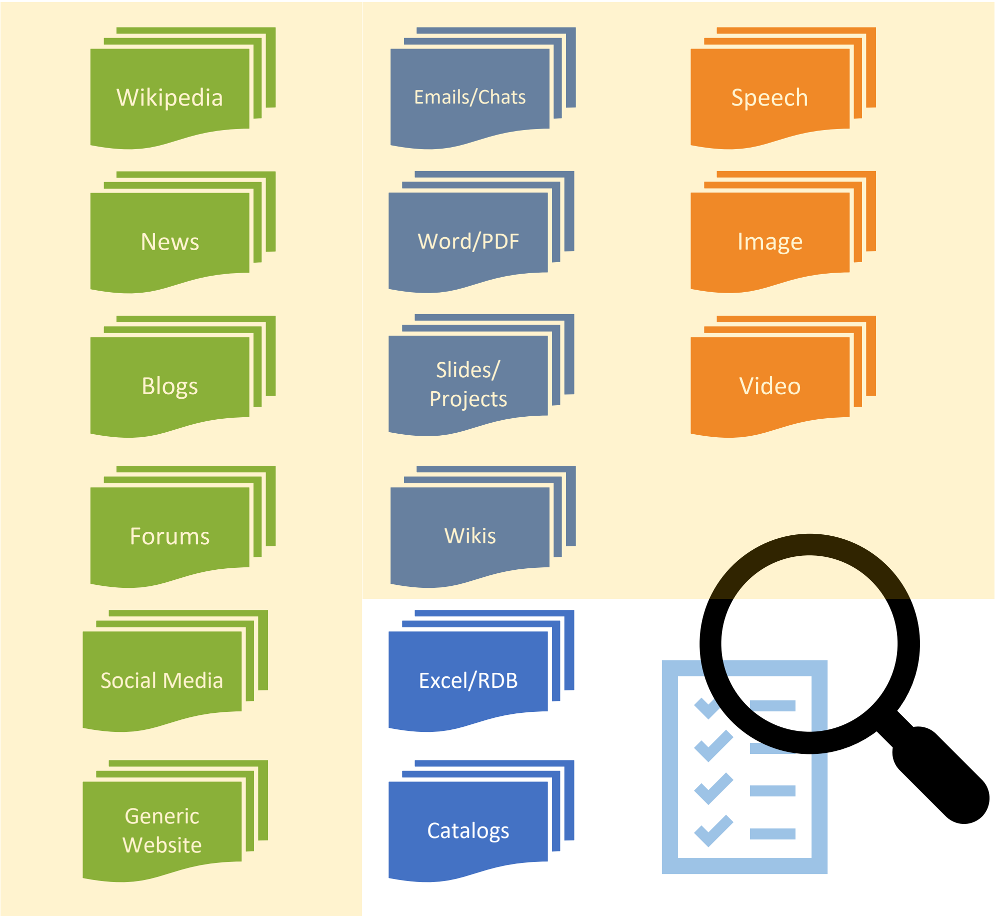
Five Pillars of High-Quality KA for KG

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



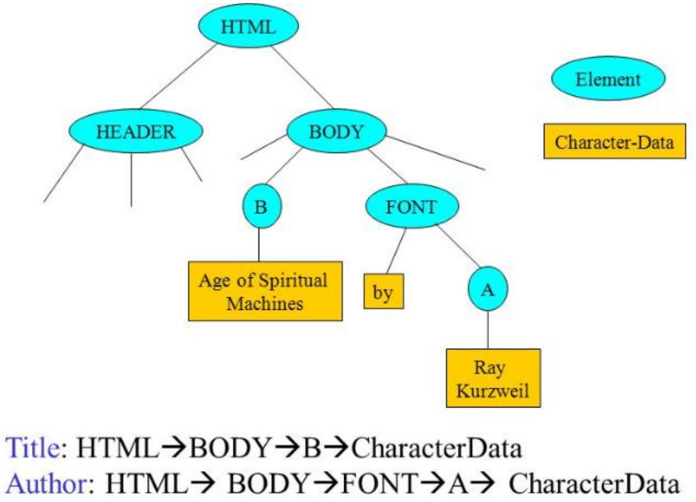
Wide Coverage

- 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 in to data regions, then extract with partial alignment.
    - Machine-learning-based.



(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.

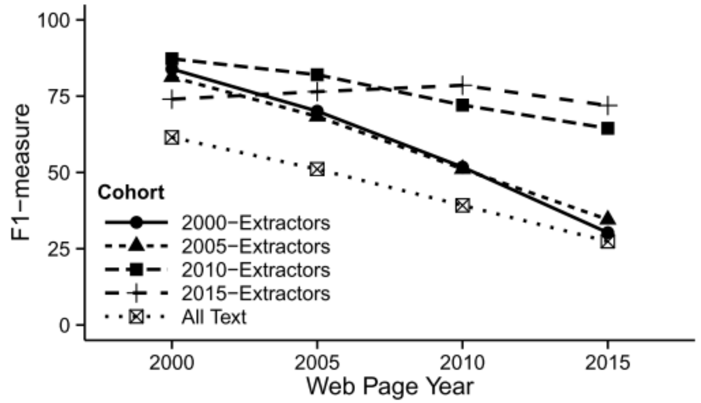


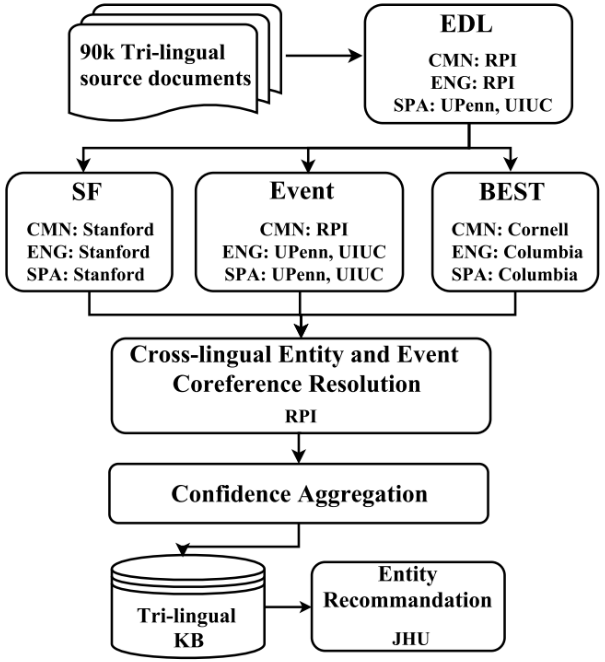
FIG. 3: Mean average F_1 measure per cohort over each lustrum.

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

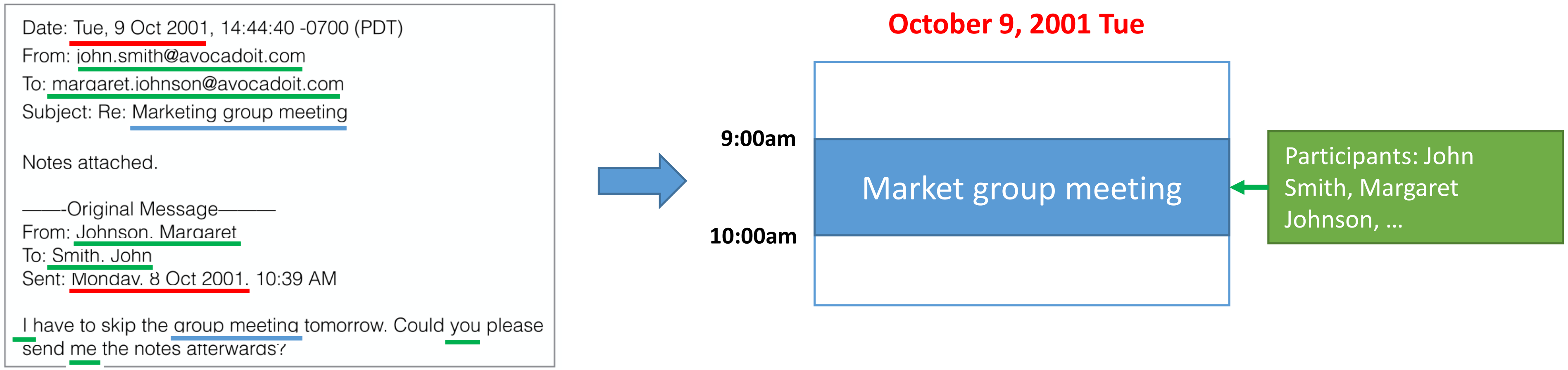
- TACKBP 2017 CSKB best system: Tinkerbell [Al-Badrashiny+ 2017]
  - First end-to-end trilingual system combining multiple building blocks from member institutions..



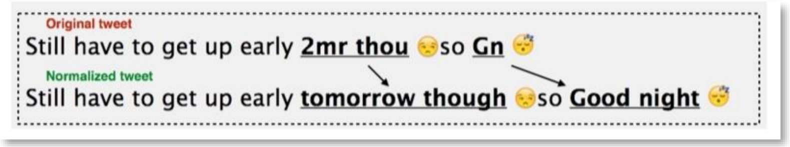
slot types	#justifications	TinkerBell	Human	% Human
all	3	7.56%	47.1%	16.1%
all	1	13.32%	59.77%	22.3%
SF	3	11.43%	40.97%	27.9%
SF	1	17.30%	41.53%	41.7%

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	$F_{\beta=1}$
ousia	57.66	55.22	56.41
NLANGP	63.62	43.12	51.40
nrc	53.24	38.58	44.74
multimedialab	49.52	39.18	43.75
USFD	45.72	39.64	42.46
iitp	60.68	29.65	39.84
Hallym	39.59	35.10	37.21
lattice	55.17	9.68	16.47
BASELINE	35.56	29.05	31.97

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

	Precision	Recall	$F_{\beta=1}$
ousia	72.20	69.14	70.63
NLANGP	67.74	54.31	60.29
USFD	63.81	56.28	59.81
multimedialab	62.93	55.22	58.82
nrc	62.13	54.61	58.13
iitp	63.43	51.44	56.81
Hallym	58.36	48.5	53.01
lattice	58.42	25.72	35.71
BASELINE	53.86	46.44	49.88

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 http://t.co/YdoygOGmhv
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: https://t.co/j69VeJC4dw
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!

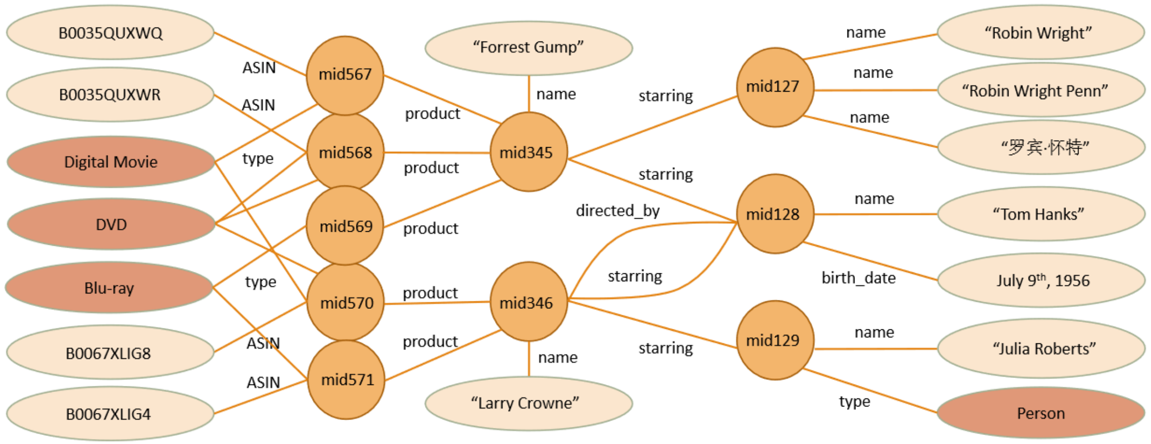
Table 6: Example high-confidence events extracted using our system.

Topic	P@5	sample of quantfrags
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.

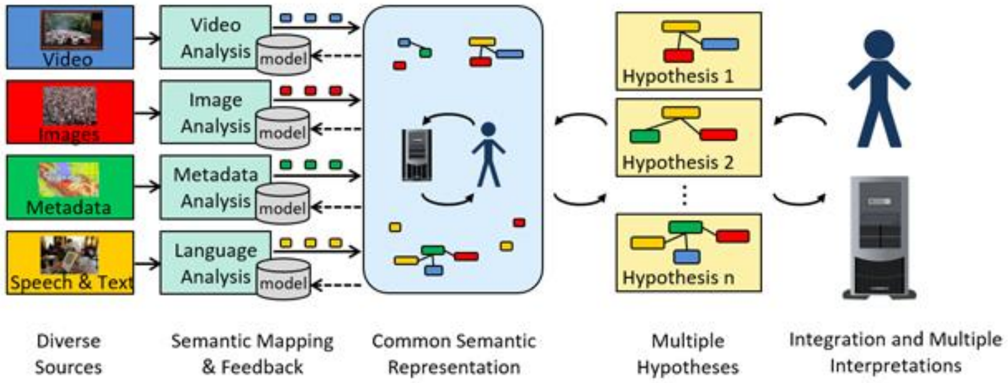
Unstructured Sources: Texts - 5

- 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.



(DARPA AIDA Program)

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

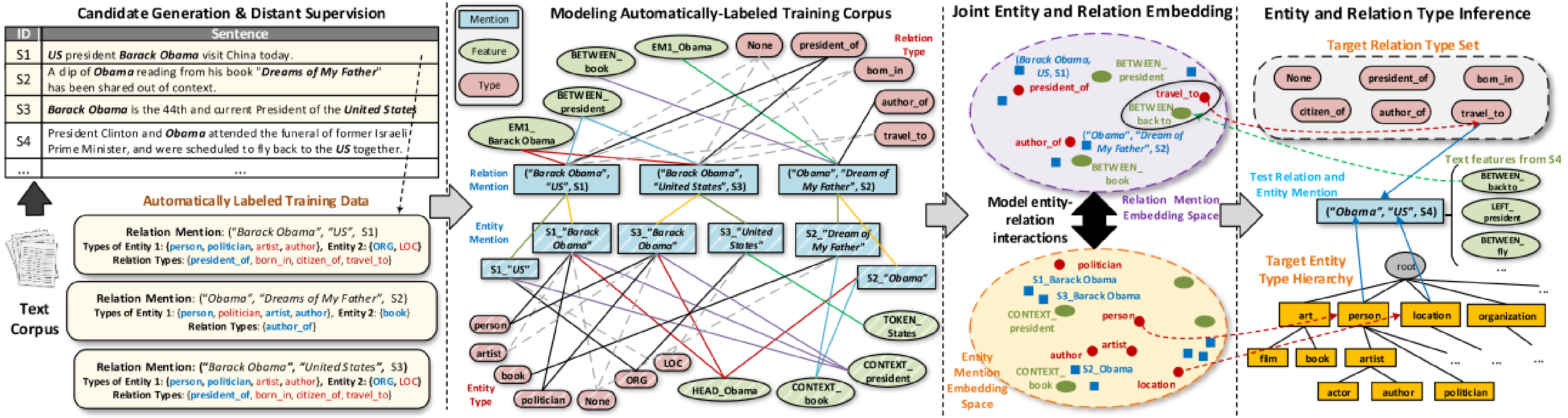
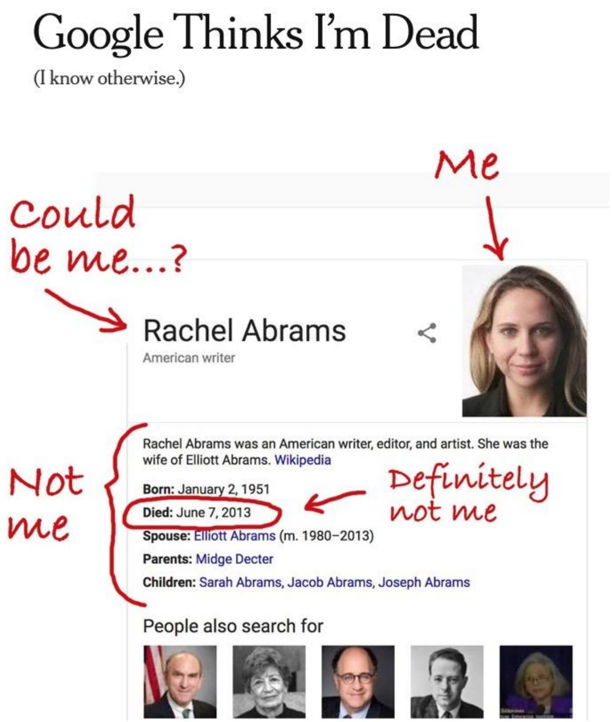


Figure 2: Framework Overview of CoType.

Coverage/Precision: Entity Linking

- Disastrous result if linking failed, even with perfect extraction



(source)

NEMO (Named Entities Made Obvious) - 1

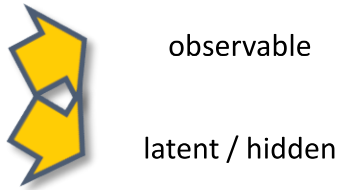
[Cucerzan 2007]

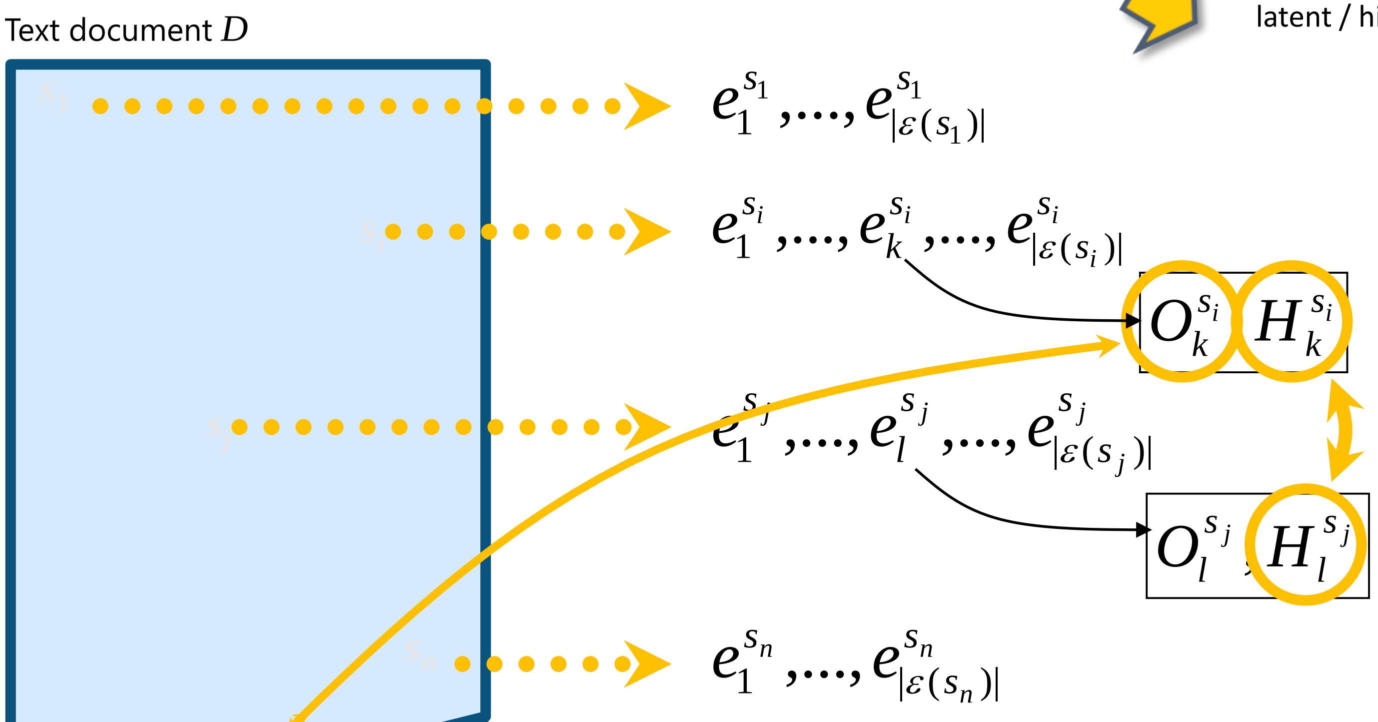
- The best evidence for entity disambiguation is provided by the set of co-occurring entities
  - Extract and disambiguate jointly all entities in a target document
  - Employ both observable attributes (known values, contexts) and latent attributes (e.g. entity relationships, topics)
- Syntax and local 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





$d = D \cap O$

Find the entity assignment that maximizes the similarity between the observable representations and the document context d and between the latent representations of the entities in the assignment.

[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

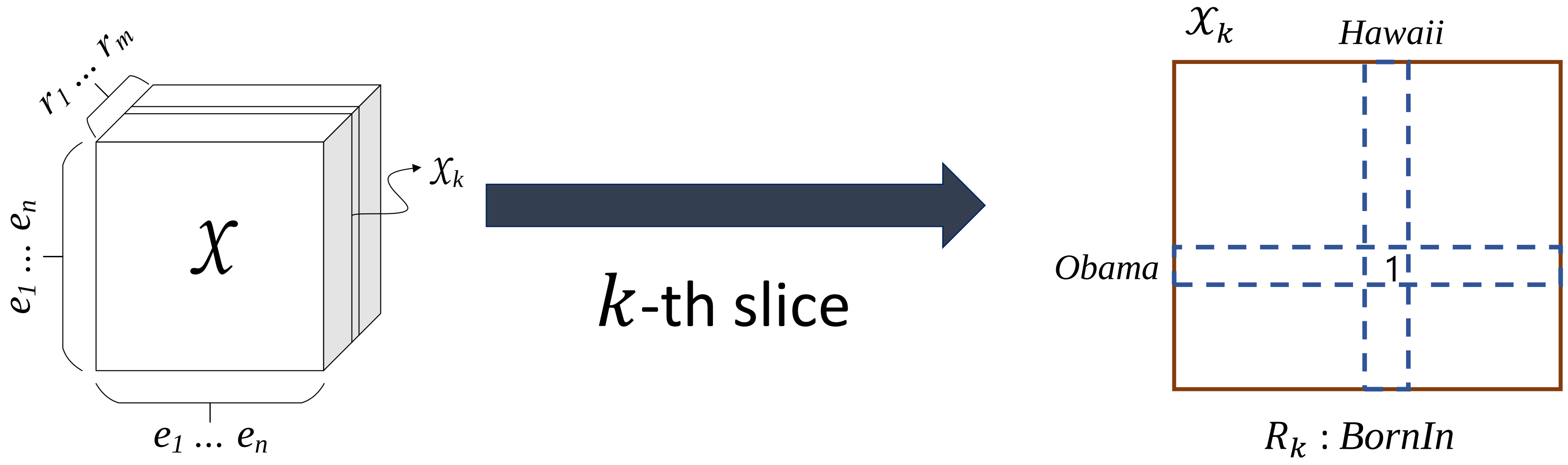
Coverage: Extracting Relations

- 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]

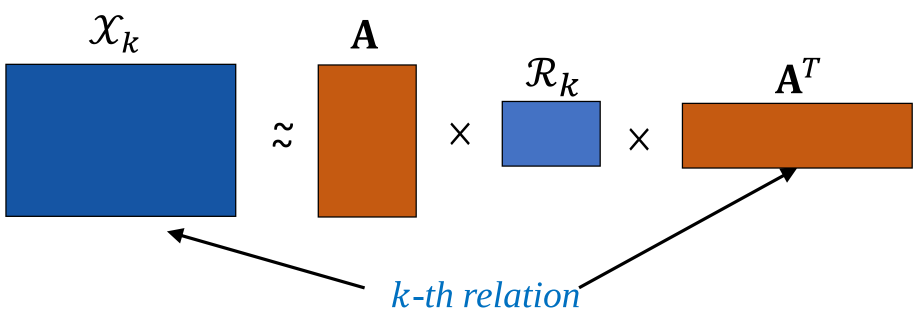
Embedding Methods for KB Completion - 1

- Each entity in a KB is represented by an R^d vector
- Predict whether (e_1, r, e_2) is true by $f_r(\boldsymbol{v}_{e_1}, \boldsymbol{v}_{e_2})$
- 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: $\underbrace{\frac{1}{2}\left(\sum_k \|\mathcal{X}_k - \mathbf{A}\mathcal{R}_k\mathbf{A}^T\|_F^2\right)}_{\textit{Reconstruction Error}} + \underbrace{\frac{1}{2}\left(\|A\|_F^2 + \sum_k \|\mathcal{R}_k\|_F^2\right)}_{\textit{Regularization}}$

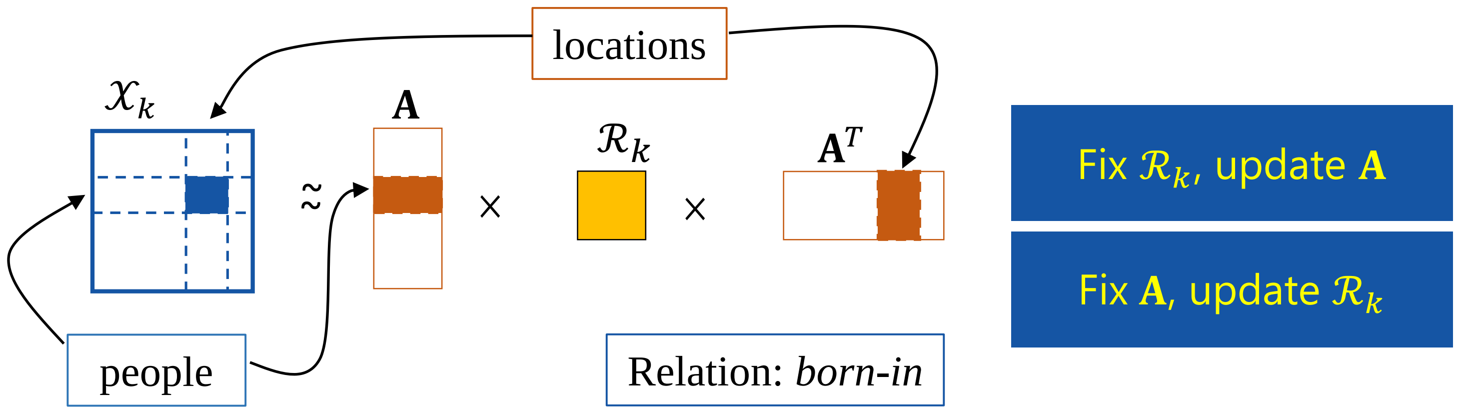


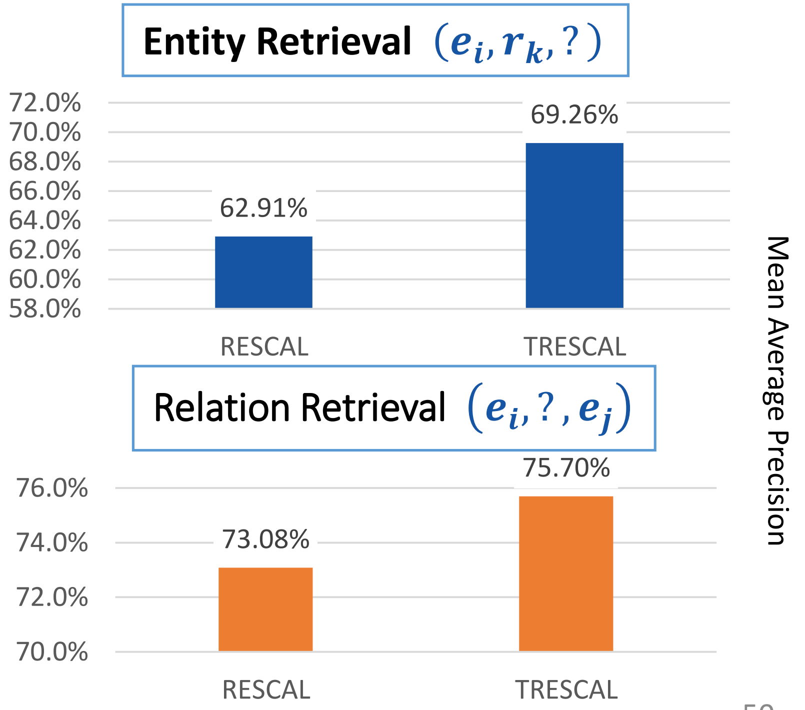
$$f_{\mathrm{BornIn}}(\mathrm{Obama}, \mathrm{Hawaii})$$
$$= \mathbf{A}_{\mathrm{Obama},:}\, \mathcal{R}_{\mathrm{BornIn}}\, \mathbf{A}^{\mathrm{T}}_{\mathrm{Hawaii},:}$$

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: $\frac{1}{2}\sum_k \|\mathcal{X}_k - \mathbf{A}\mathcal{R}_k\mathbf{A}^T\|_F^2$
  - 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

Bob is in his room, but because of the thunder and lightning outside, he cannot sleep

Entail

Bob is awake

Contradict

It is sunny outside

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]

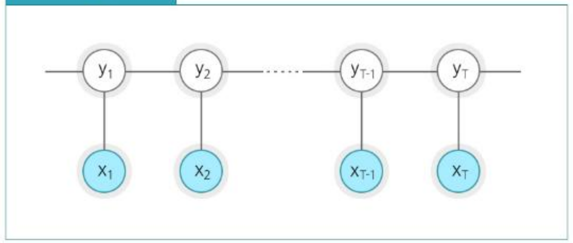
More Efficient Human Intervention - 1

- 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

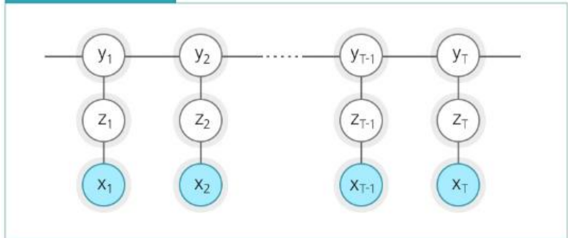
CRF



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

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

Hidden Unit CRF



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

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

Partially observed CRF

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

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

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

Initialization:

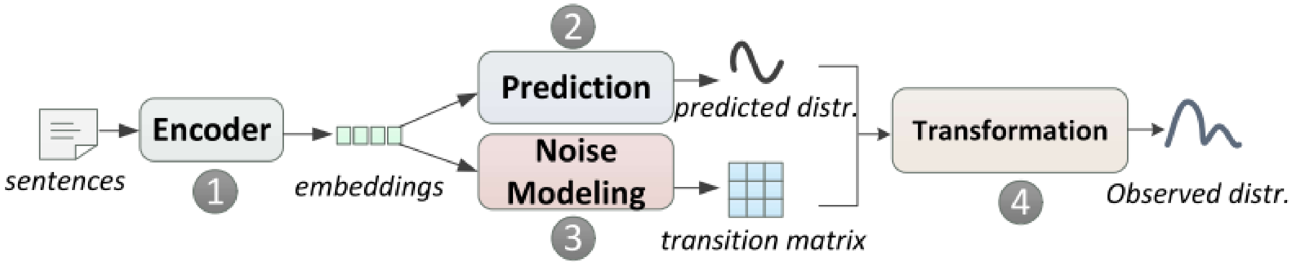
- Cluster unlabeled data
- Train fully supervised HUCRF with cluster labels
- Keep learned θ (between input *x* and hidden *z*) and start task-specific training

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+	**79.28**	**47.35**	**78.33**	**68.32**

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

More Efficient Human Intervention - 4

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

$$L = \sum_{i=1}^{N} -((1-\alpha)log(o_{iy_i}) + \alpha log(p_{iy_i})) - \beta trace(\mathbf{T}^i)$$

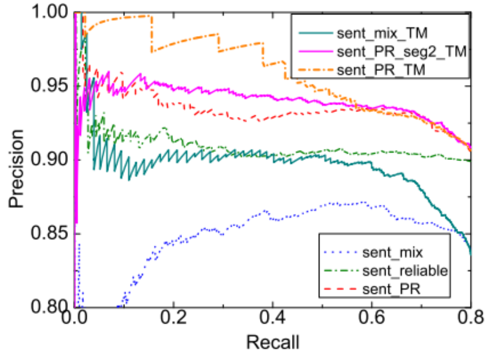


Figure 2: Sentence Level Results on TIMERE

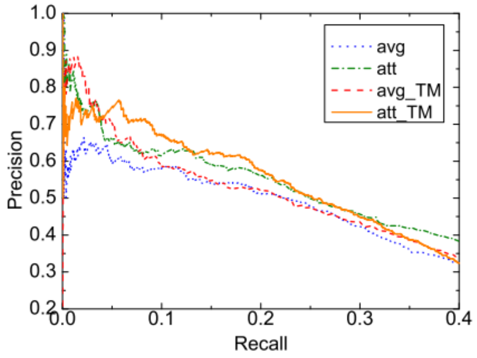


Figure 5: Results on ENTITYRE (bag level)

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]

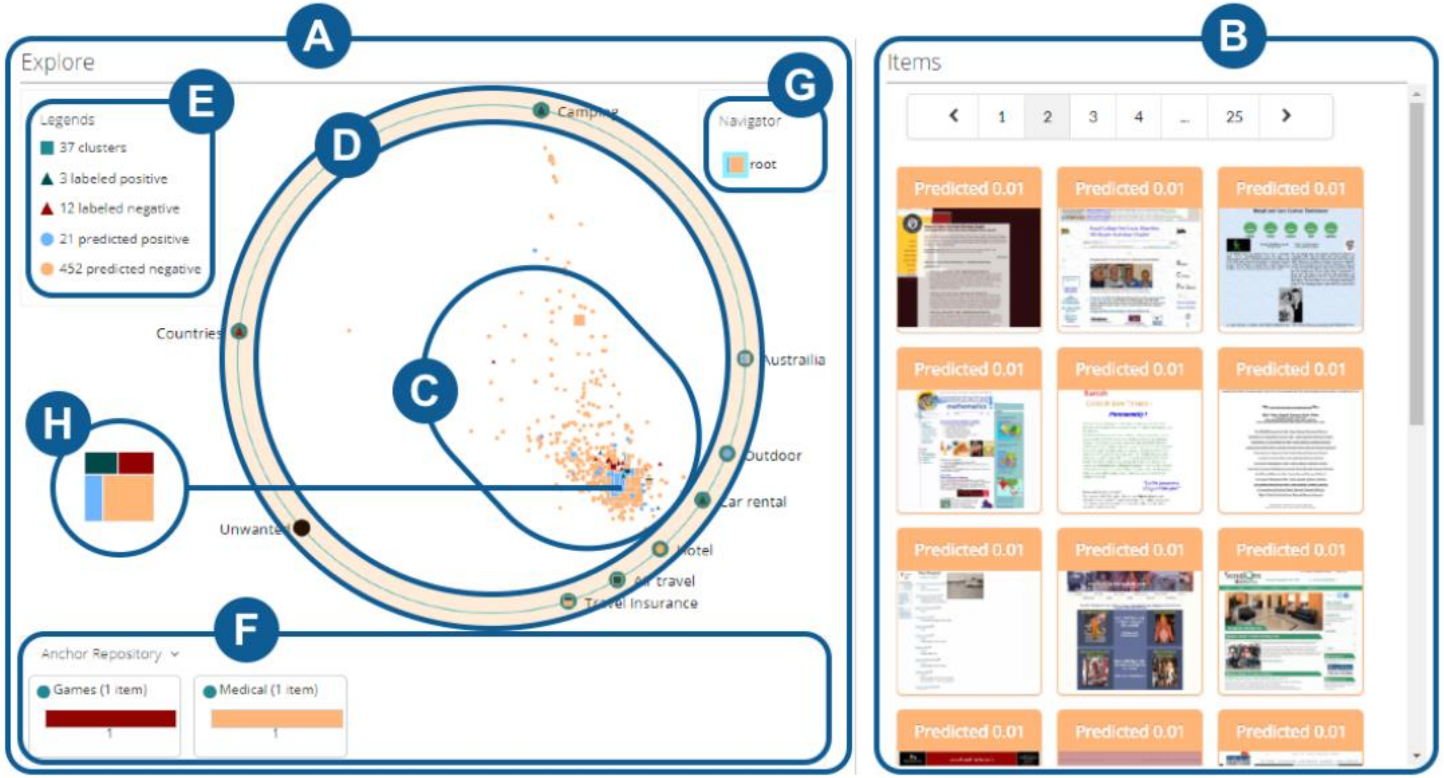


Figure 1. Overview of AnchorViz interface. The interface has Explore pane (A) that includes the visualization and Items pane (B) which shows a paginated grid of thumbnails of all currently visible items in the left pane. The visualization shows data points (C) within the outer circle (D) where anchors 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

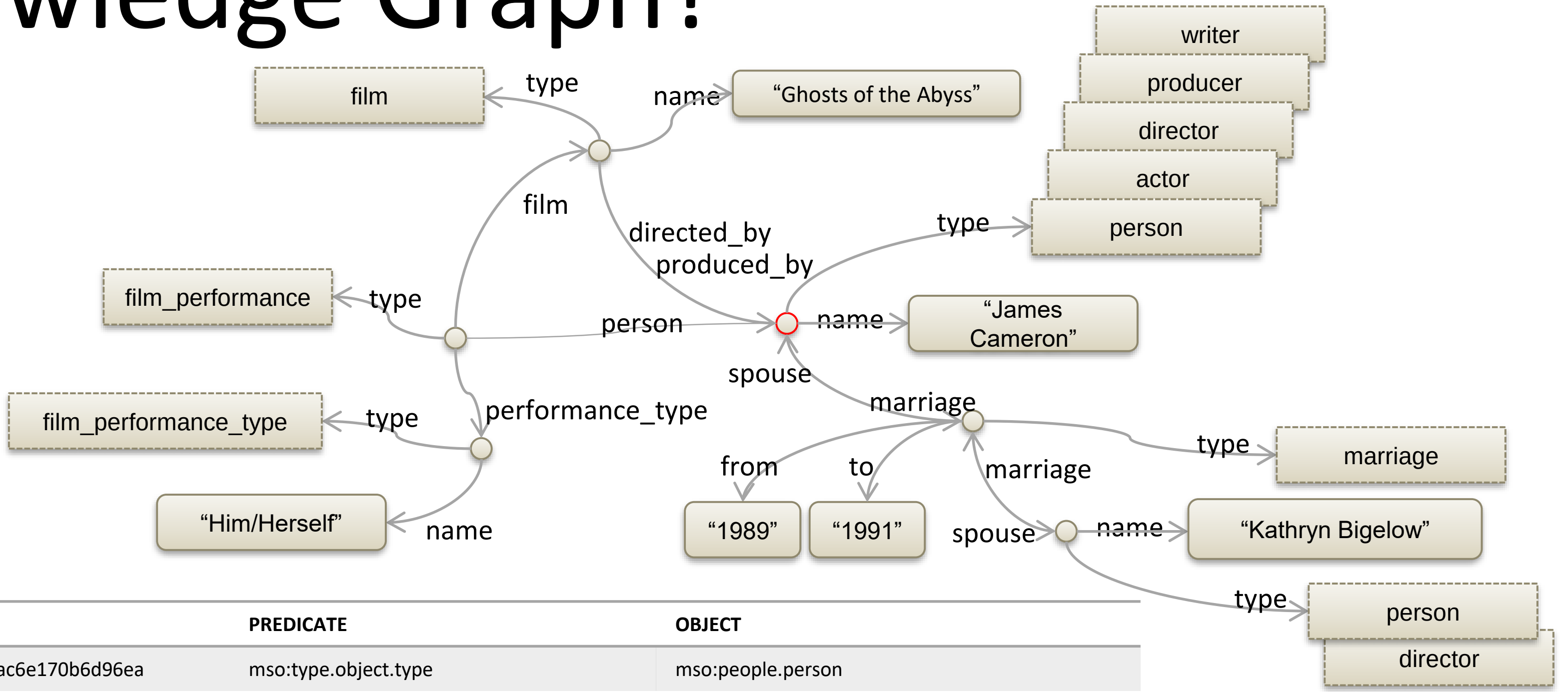
Mohamed Yakout

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

myakout@microsoft.com

What is a Knowledge Graph?

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



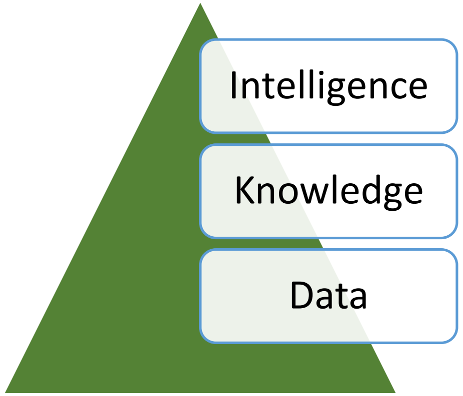
SUBJECT	PREDICATE	OBJECT
entity:00006a33abd6426fb7ac6e170b6d96ea	mso:type.object.type	mso:people.person
entity:00006a33abd6426fb7ac6e170b6d96ea	mso:type.object.name	"James Cameron"@en
entity:00006a33abd6426fb7ac6e170b6d96ea	mso:people.person.marriage	entity:92029f0be77b449894f773d48d10ab16
entity:92029f0be77b449894f773d48d10ab16	mso:type.object.type	mso:people.marriage
entity:92029f0be77b449894f773d48d10ab16	mso:people.marriage.spouse	entity:00006a33abd6426fb7ac6e170b6d96ea
entity:92029f0be77b449894f773d48d10ab16	mso:people.marriage.spouse	entity:920270b3da6d404cb4f4277b34e63740
entity:920270b3da6d404cb4f4277b34e63740	mso:type.object.type	mso:people.person
entity:920270b3da6d404cb4f4277b34e63740	mso:type.object.name	"Kathryn Bigelow"@en
entity:9202643b6a134b5aaba83d7458047579	mso:type.object.type	mso:film.film
entity:9202643b6a134b5aaba83d7458047579	mso:type.object.name	"Ghosts of the Abyss"@en
entity:9202643b6a134b5aaba83d7458047579	mso:film.directed_by	entity:00006a33abd6426fb7ac6e170b6d96ea

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 Ingestion → **Match & Merge** → **Knowledge Refinement** → **Publishing & Serve**

Data Ingestion

- Selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge

- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement

- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve

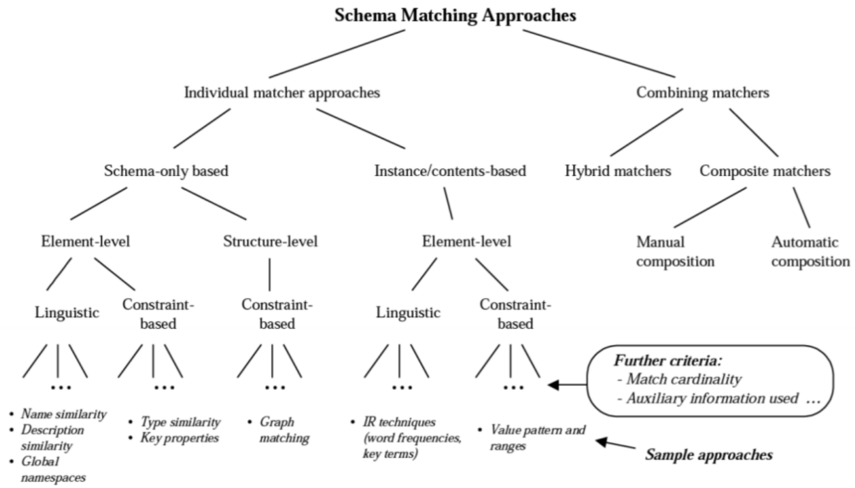
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join

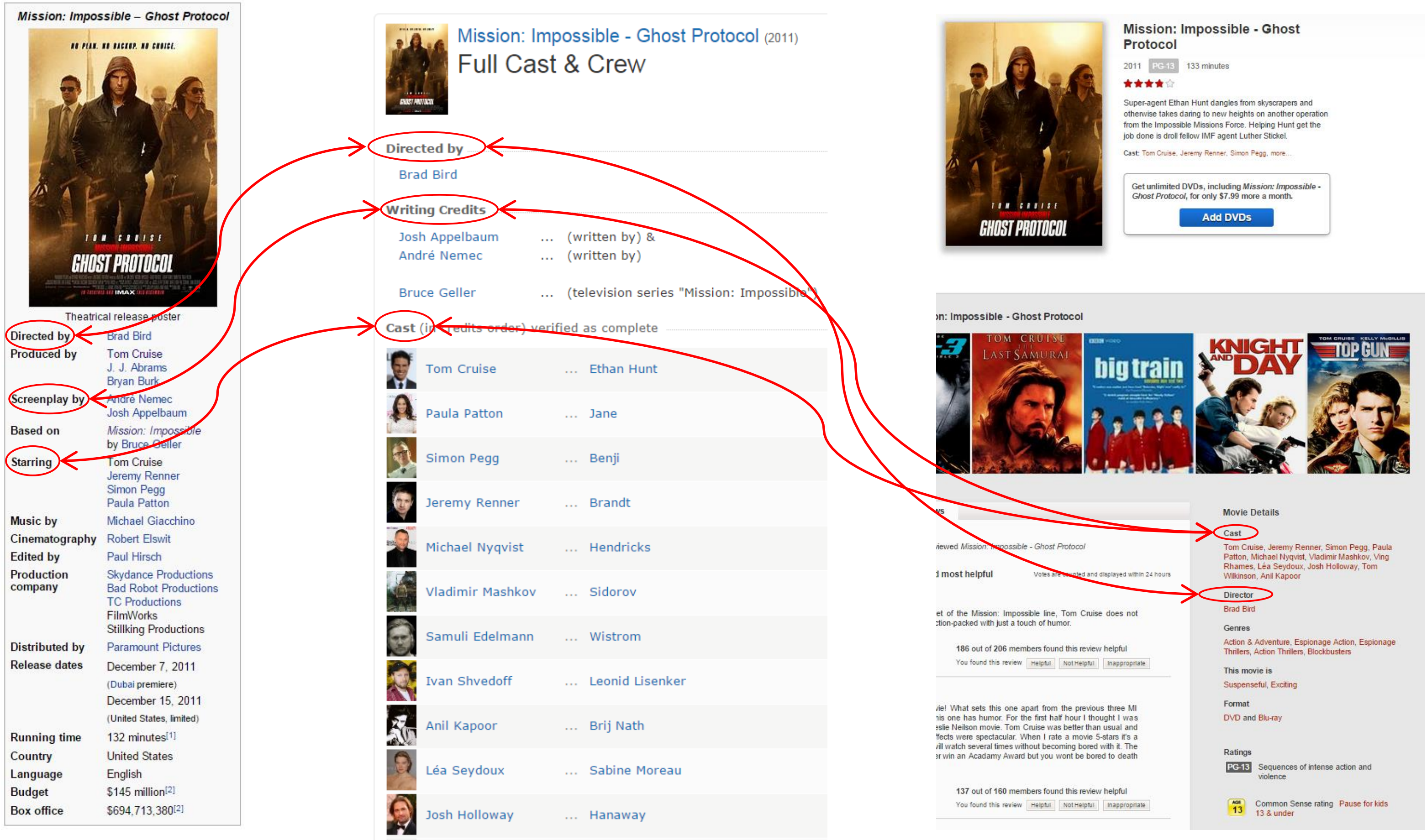
Data Preparation

- 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^{th} 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

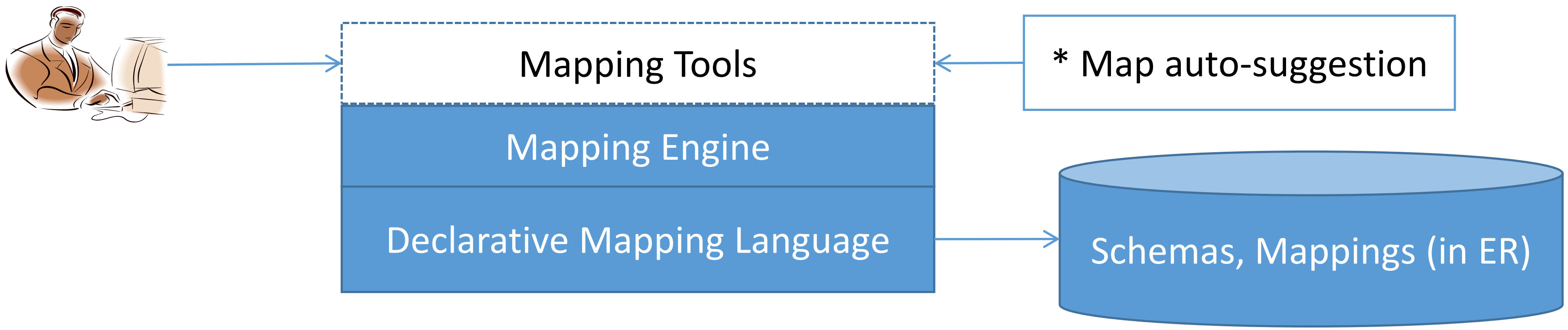




Erhard et al, "A survey of approaches to automatic schema matching", VLDB 2001

Schema Mapping and Management

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

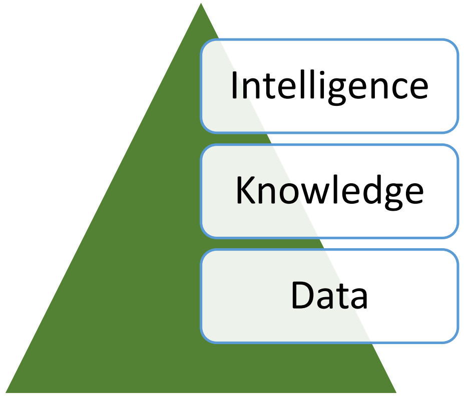


Example mapping for Music data to Satori ontology

```
<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



Data Ingestion

- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge

- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

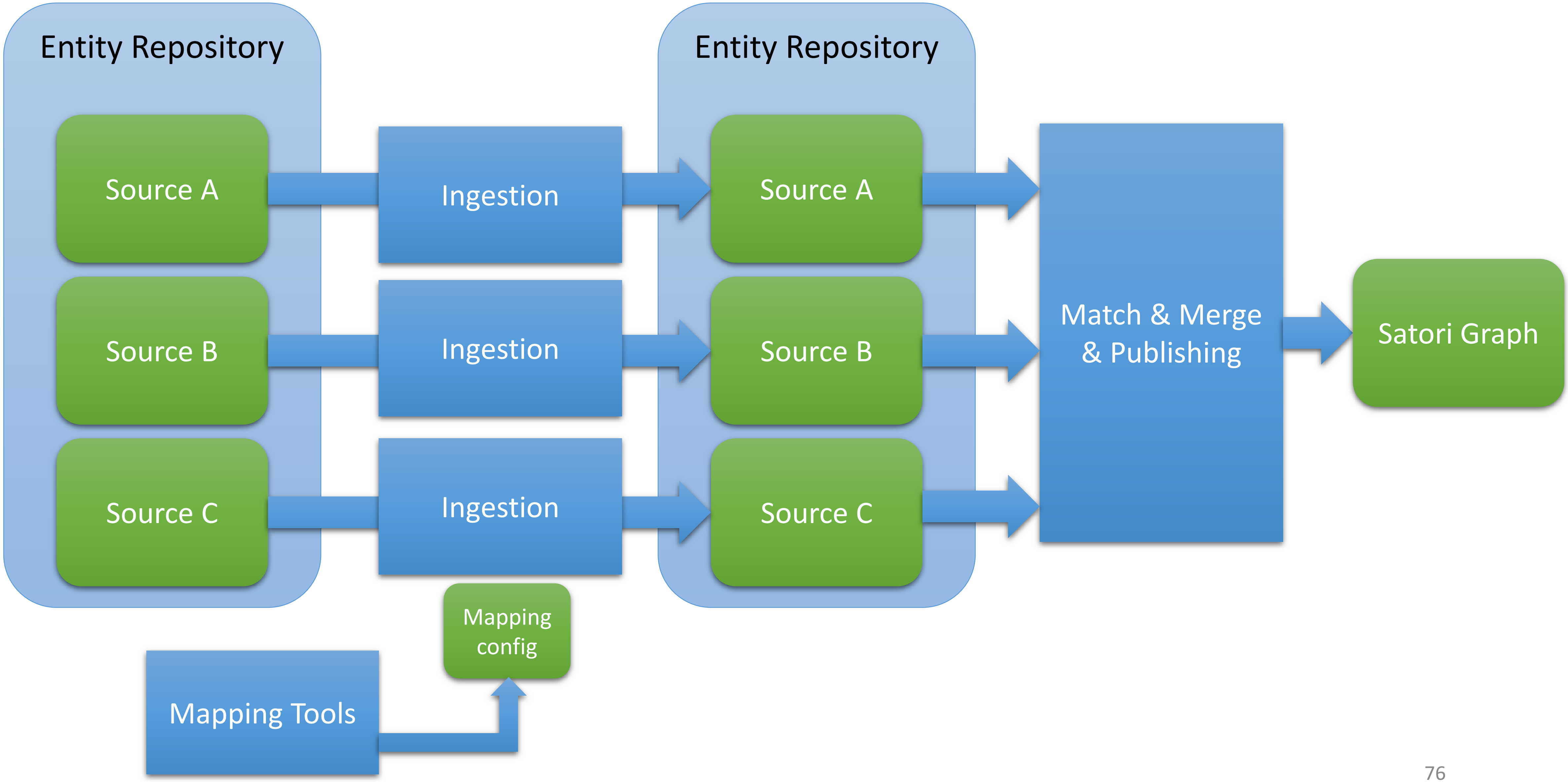
Knowledge Refinement

- Knowledge Fusion
- Error Detection
- Fact Inference

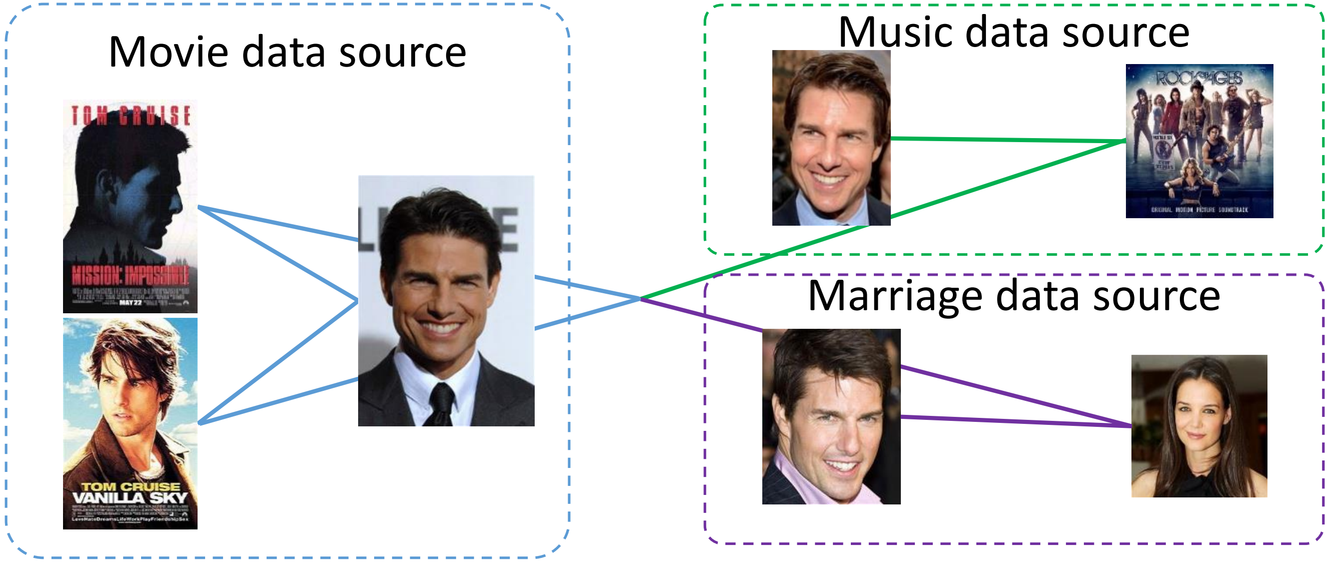
Publishing & Serve

- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join

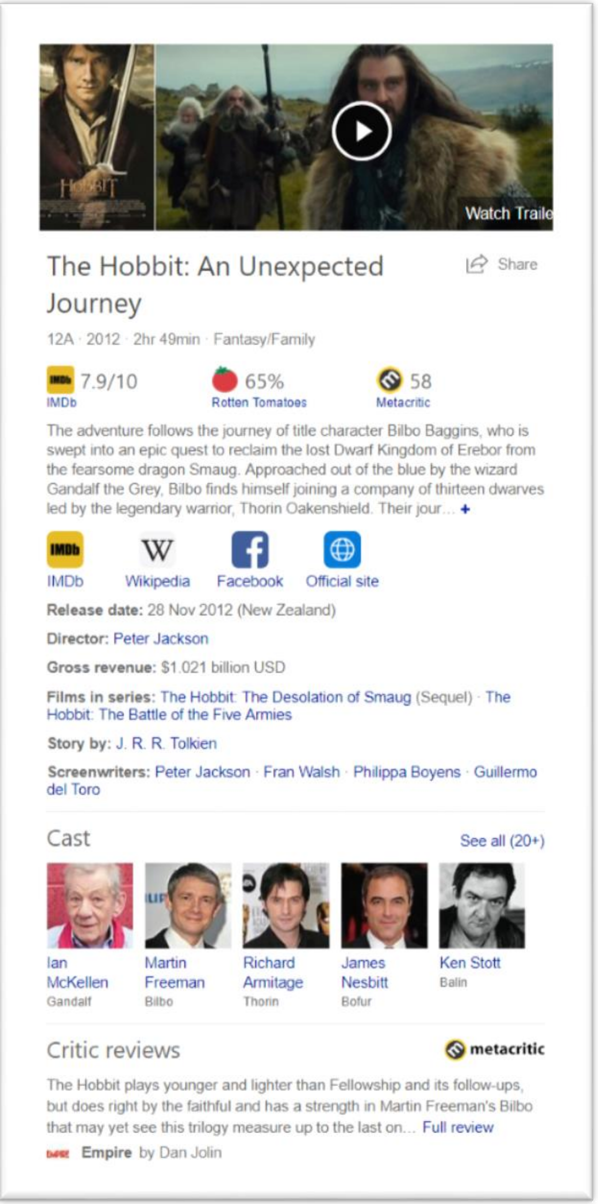
Ingestion Flow

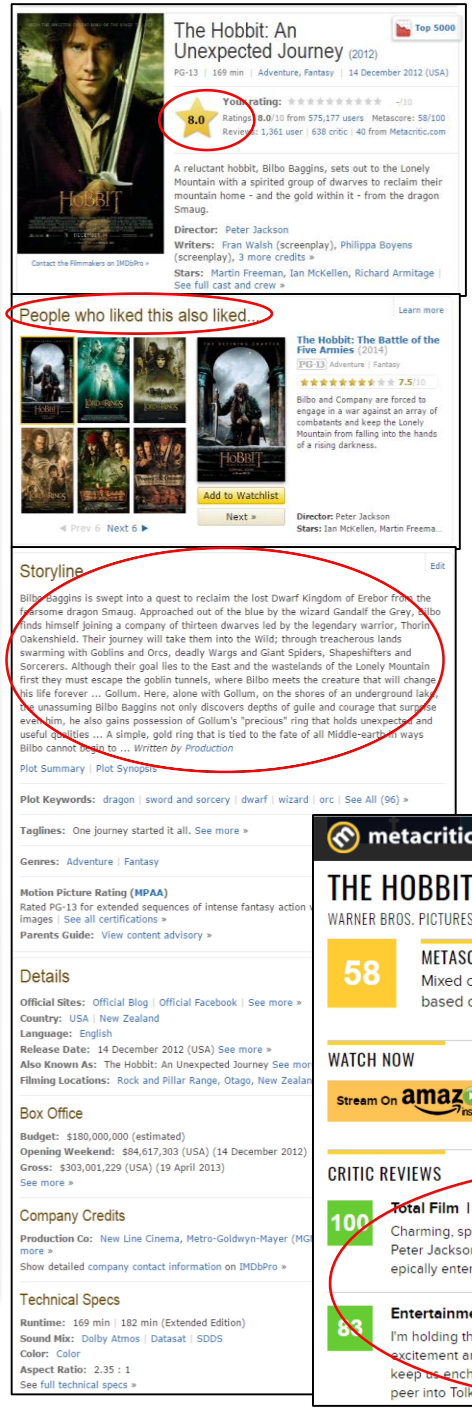


Entity Matching



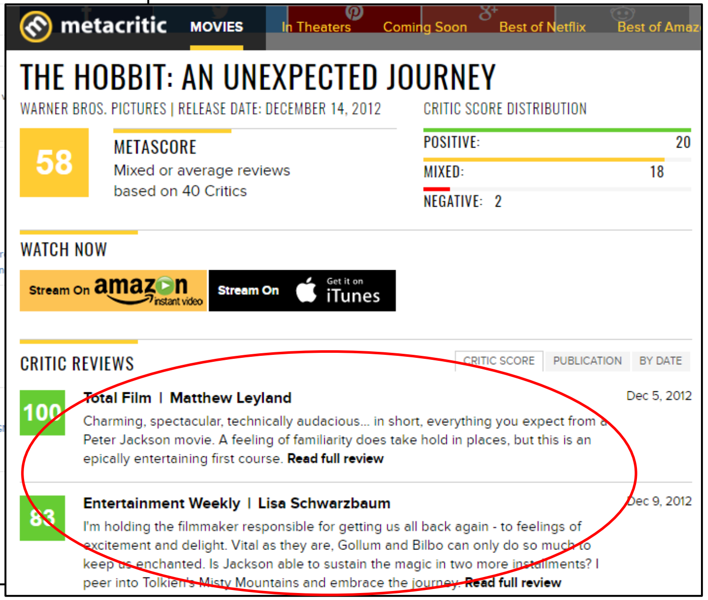
- **Well known problem**: Identify and discover instances referring to the same real-world entity.
- **Objective**:
  - Data Enrichment
  - Improve Data Quality by identifying and removing duplicates
  - Supporting fact correctness by merging duplicate facts from multiple sources
- **Synonyms**: Entity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





Integration means more information and enrichment

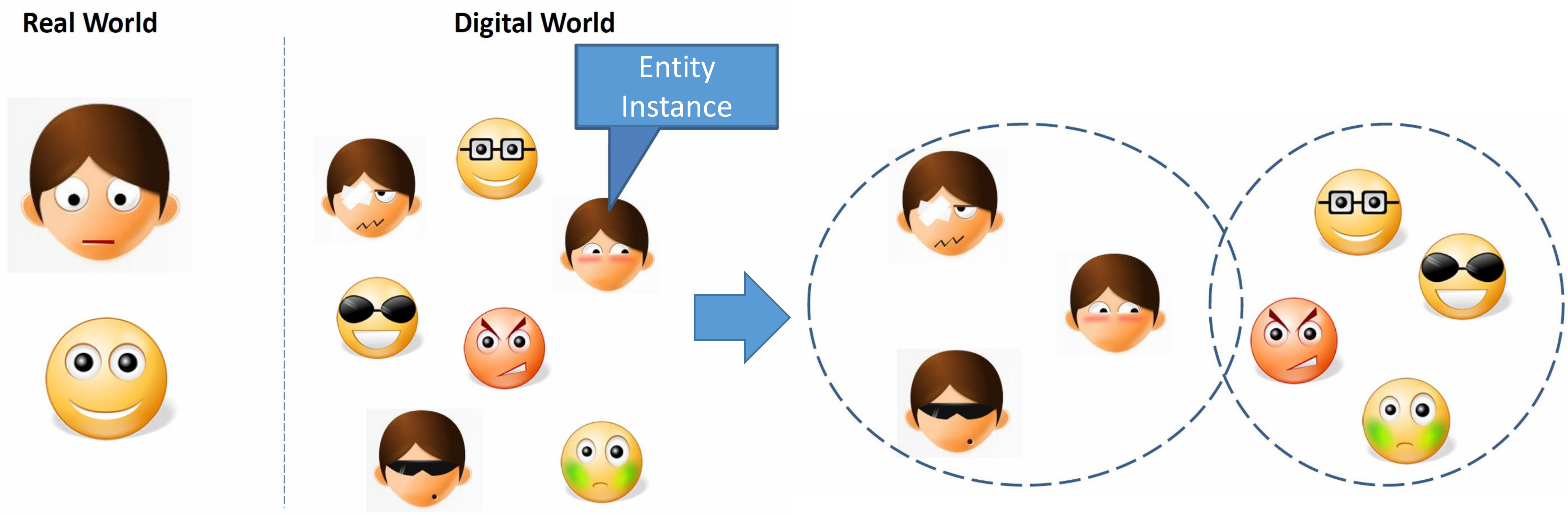
Knowledge Evolution



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. Elmagarid, 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

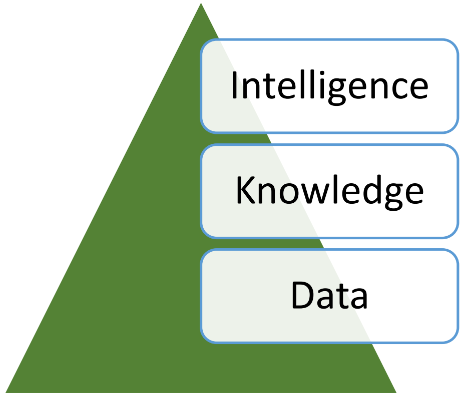
Getoor et al, “Entity Resolution: Theory, Practice & Open Challenges”, VLDB 2012

Open Domain Entity Matching (Disambiguation Challenge)

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



Data Ingestion

- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge

- **Matching Entity Contents**
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement

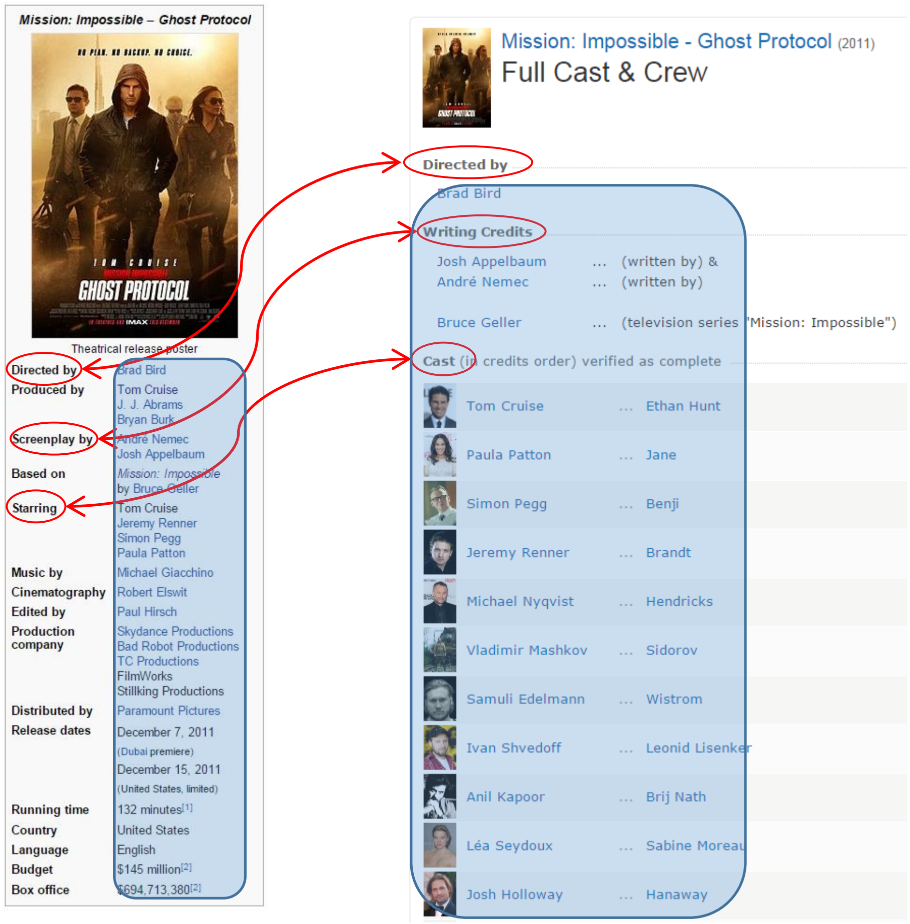
- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve

- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join

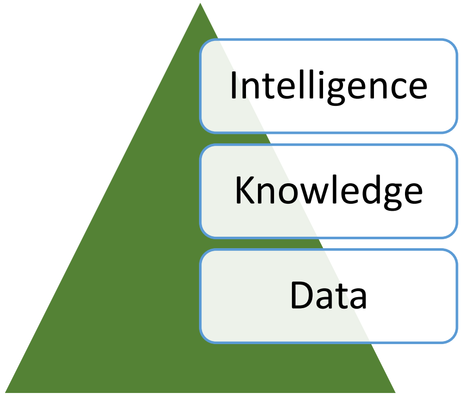
Matching Entity Contents

- 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)



Elmagarmid et al, “Duplicate Record Detection” TKDE 2007]

Satori Graph Build



Data Ingestion → **Match & Merge** → **Knowledge Refinement** → **Publishing & Serve**

Data Ingestion

- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge

- Matching Entity Contents
- **Detection of Matched Entities**
- Scaling & Improve Efficiency of Matching

Knowledge Refinement

- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve

- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join

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.

Name	John Smith
Profession	Software Eng.
Address	Seattle

Name	Johan Smith
Profession	Software Dev Eng.
Address	Seattle

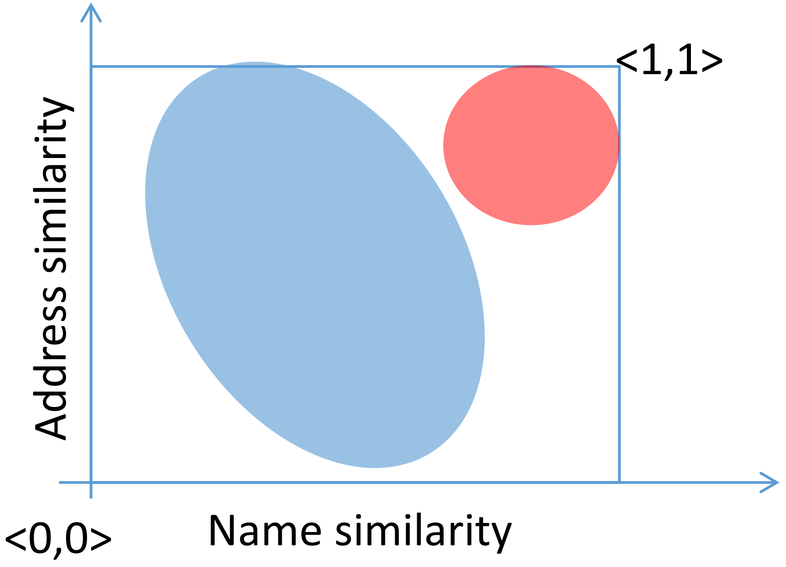
Similarity ➔ Vector

Name	0.8
Profession	0.7
Address	1.0

The similarity vector <0.8, 0.7, 1.0>

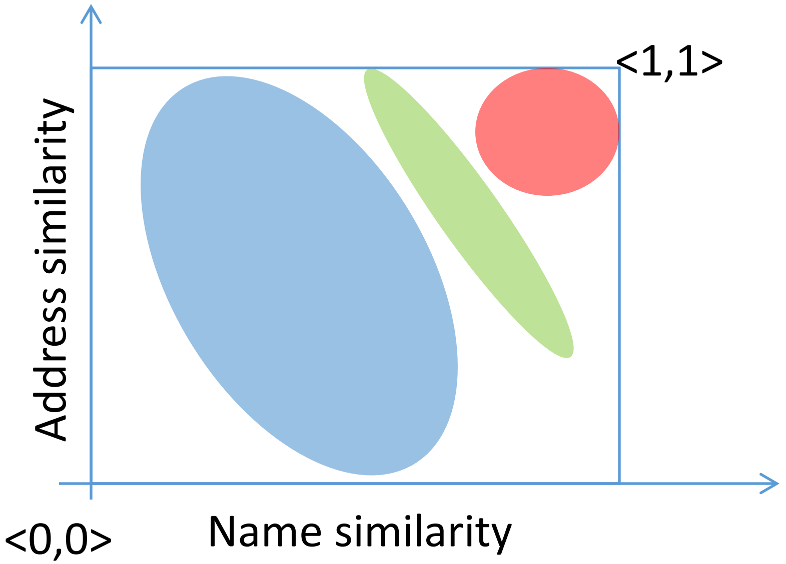
Detection of Matched Entities: Probabilistic Matching Models

- 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)

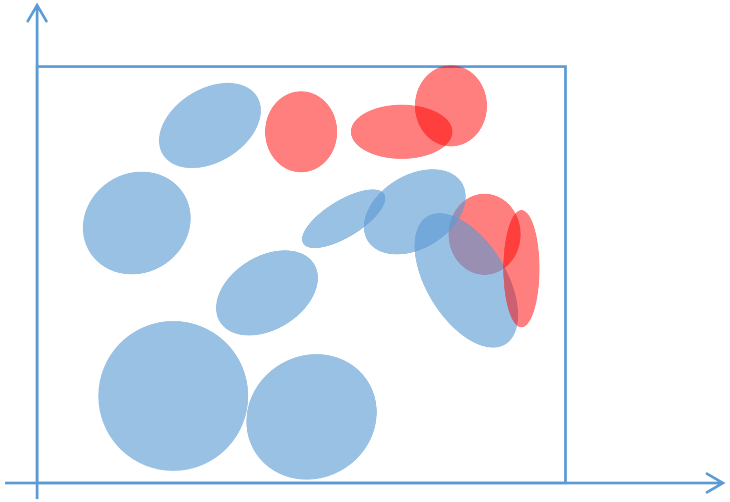


  - Rely on the existence of training data, pair of records pre-labeled match or not. Do we have that?!!

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

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

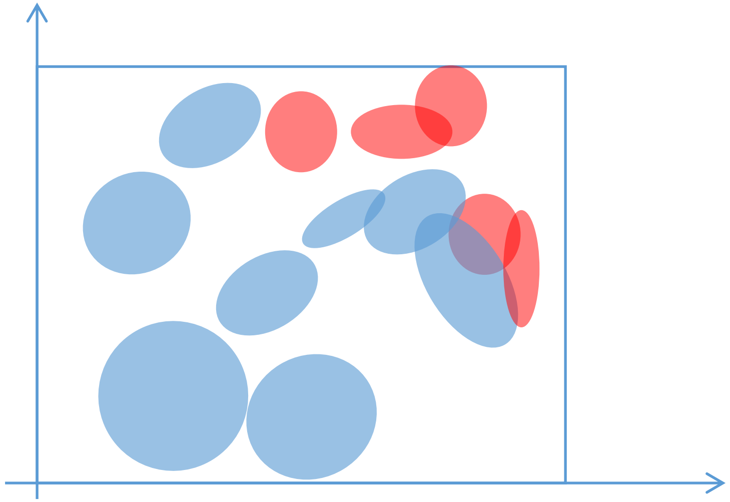
  - Train an initial ML model by an initial small sample
  - While (user is not happy with predictions)
    - Foreach Pair p in all pairs
      - Apply the model on p
      - Get the prediction probability and compute uncertainty
    - Sort all pairs based on uncertainty
    - Display pairs with the highest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

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**

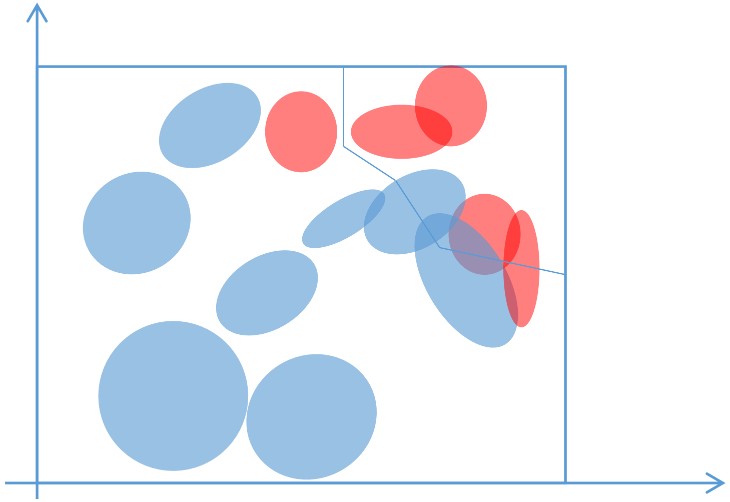
- Better control on the space of similarities.
- Clustering for all vectors
- Offline sample from clusters host locally
- **Active Learning Guided by the Clusters** through:
  - Focus on clusters with high uncertainty
  - Cover clusters with less training samples
  - From a cluster, sampling positive 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**

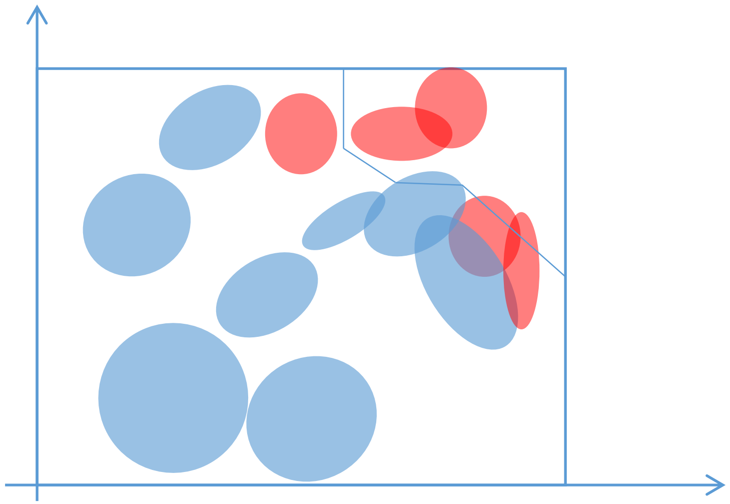
- Better control on the space of similarities.
- Clustering for all vectors
- Offline sample from clusters host locally
- **Active Learning Guided by the Clusters** through:
  - Focus on clusters with high uncertainty
  - Cover clusters with less training samples
  - From a cluster, sampling positive 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**

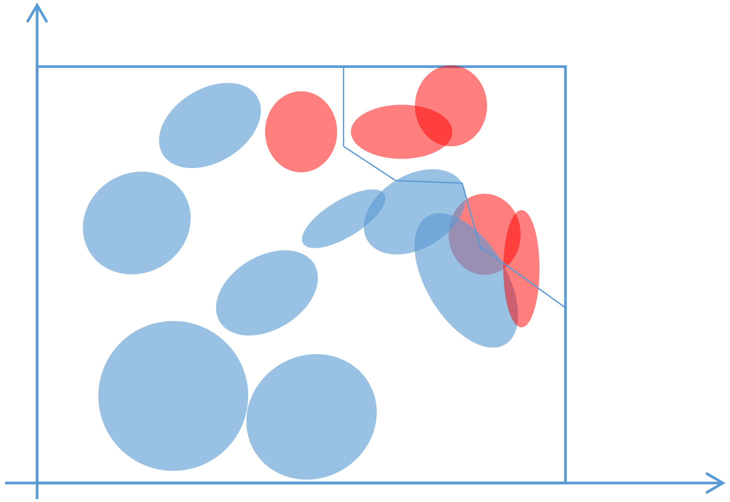
- Better control on the space of similarities.
- Clustering for all vectors
- Offline sample from clusters host locally
- **Active Learning Guided by the Clusters** through:
  - Focus on clusters with high uncertainty
  - Cover clusters with less training samples
  - From a cluster, sampling positive 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.

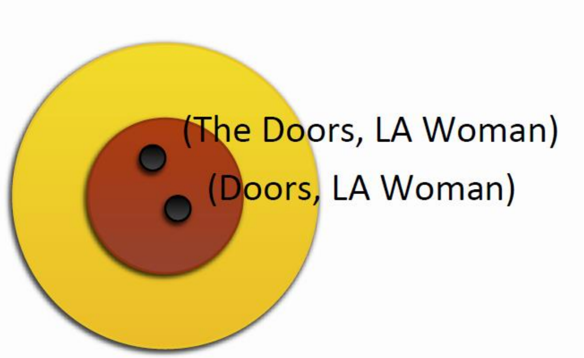


Detection of Matched Entities

- 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

Chaudhuri et all, "Robust Identification of Fuzzy Duplicates", ICDE 2005

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 constrains 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?

Arasu et al, "Large-Scale Deduplication with Constraints Using Dedupalog", ICDE 2009

Detection of Matched Entities

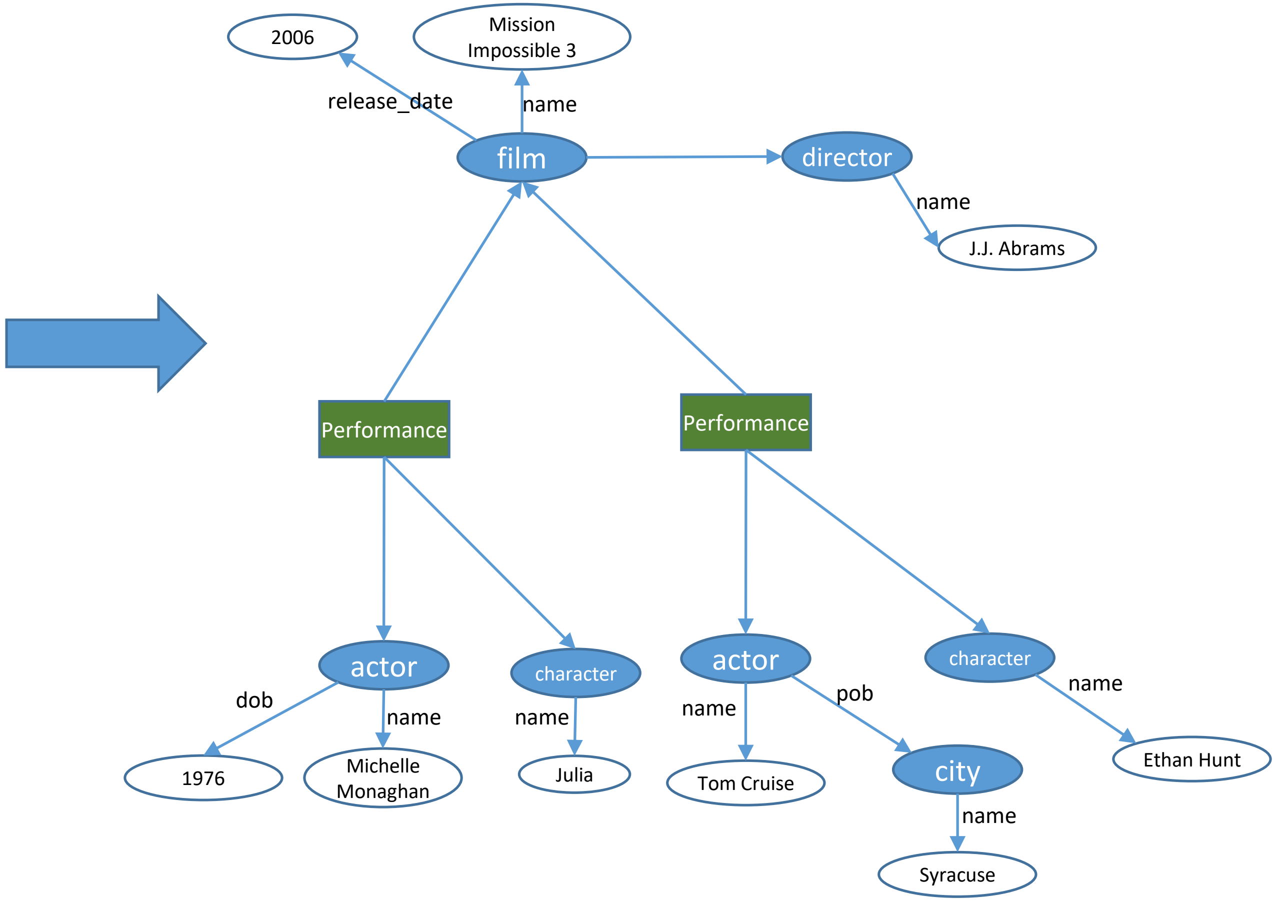
- Probabilistic Matching Models
  - Supervised and Semi-supervised Learning
  - Unsupervised learning
  - Active Learning Based

- Distance Based
  - Threshold
  - Neighborhood exploration

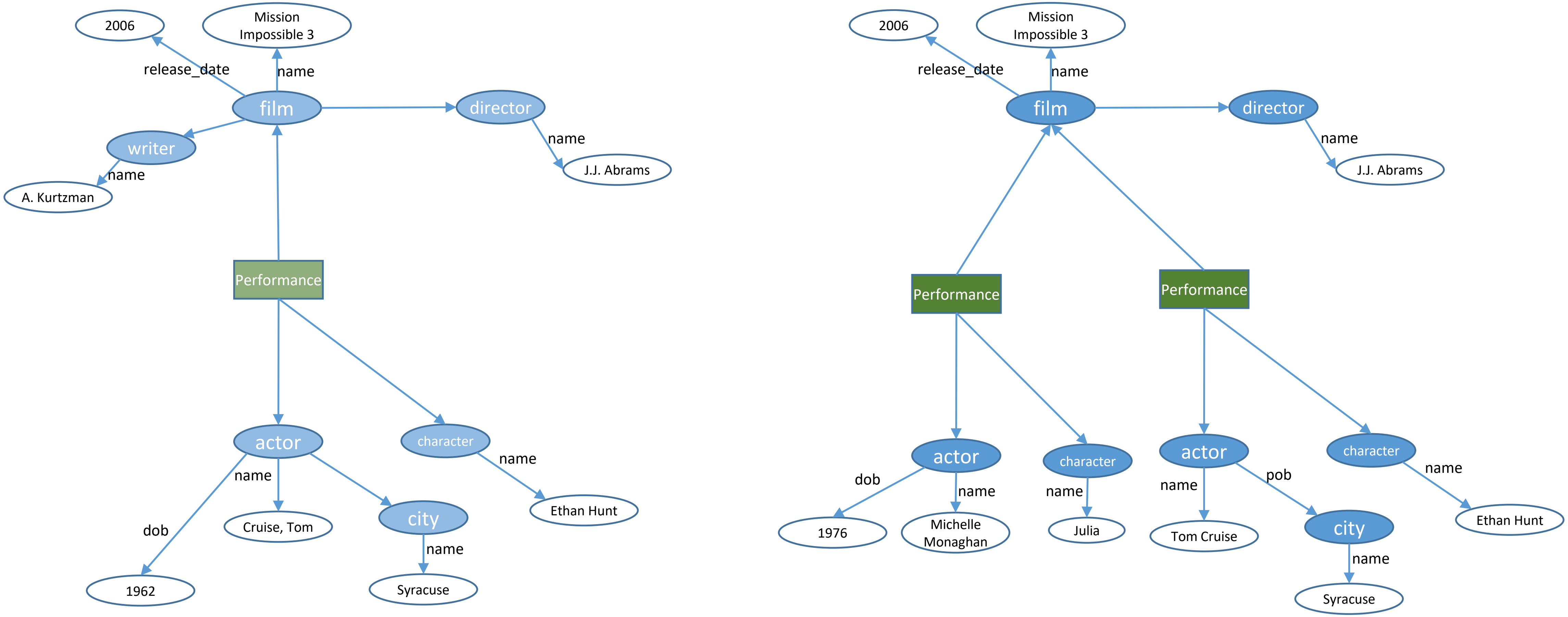
- Declarative Matching Rules and Constraints
  - Disjunction of conjunction
  - Constraint base clustering

- **Collective Resolution in Linked Data**
  - Similarity signals propagation
  - Entity similarity based on connections

Detection of Matched Entities: Collective Resolution in Linked Data

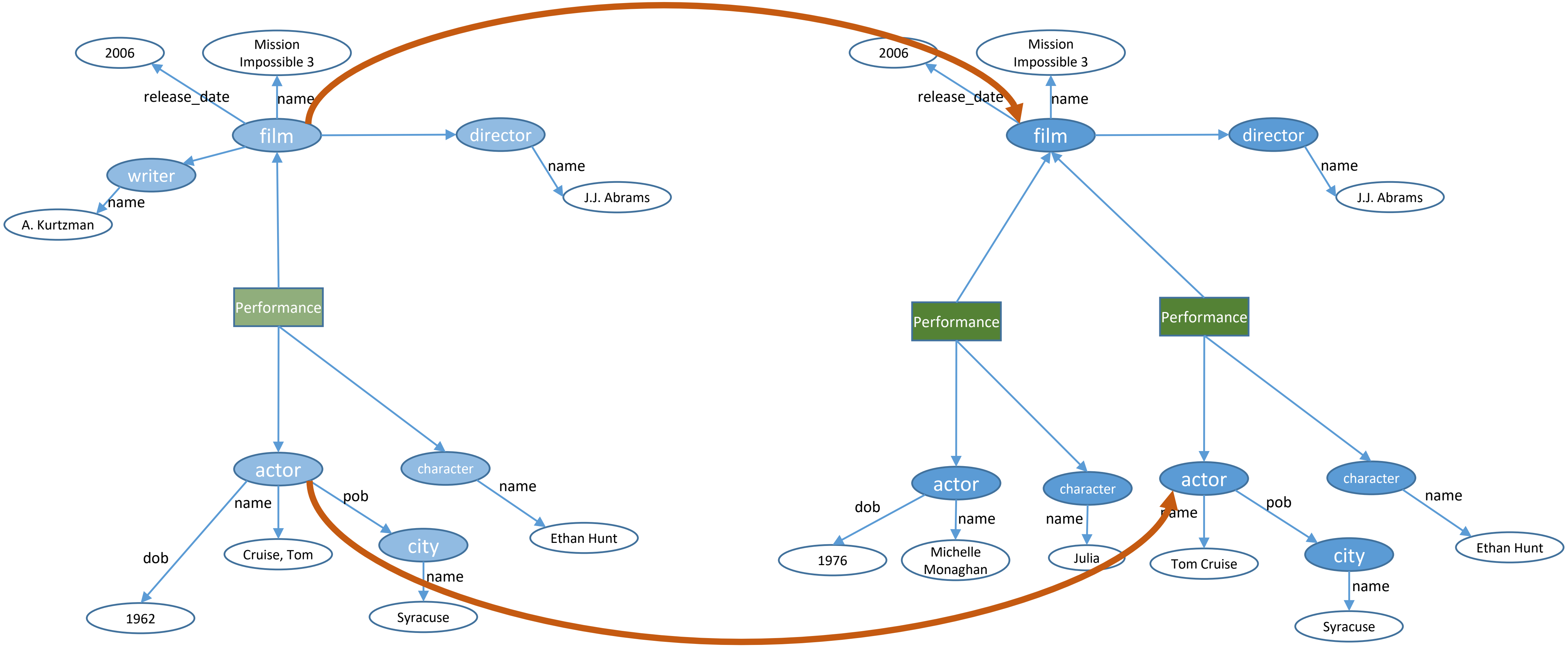


Graph Data Model and Conflation

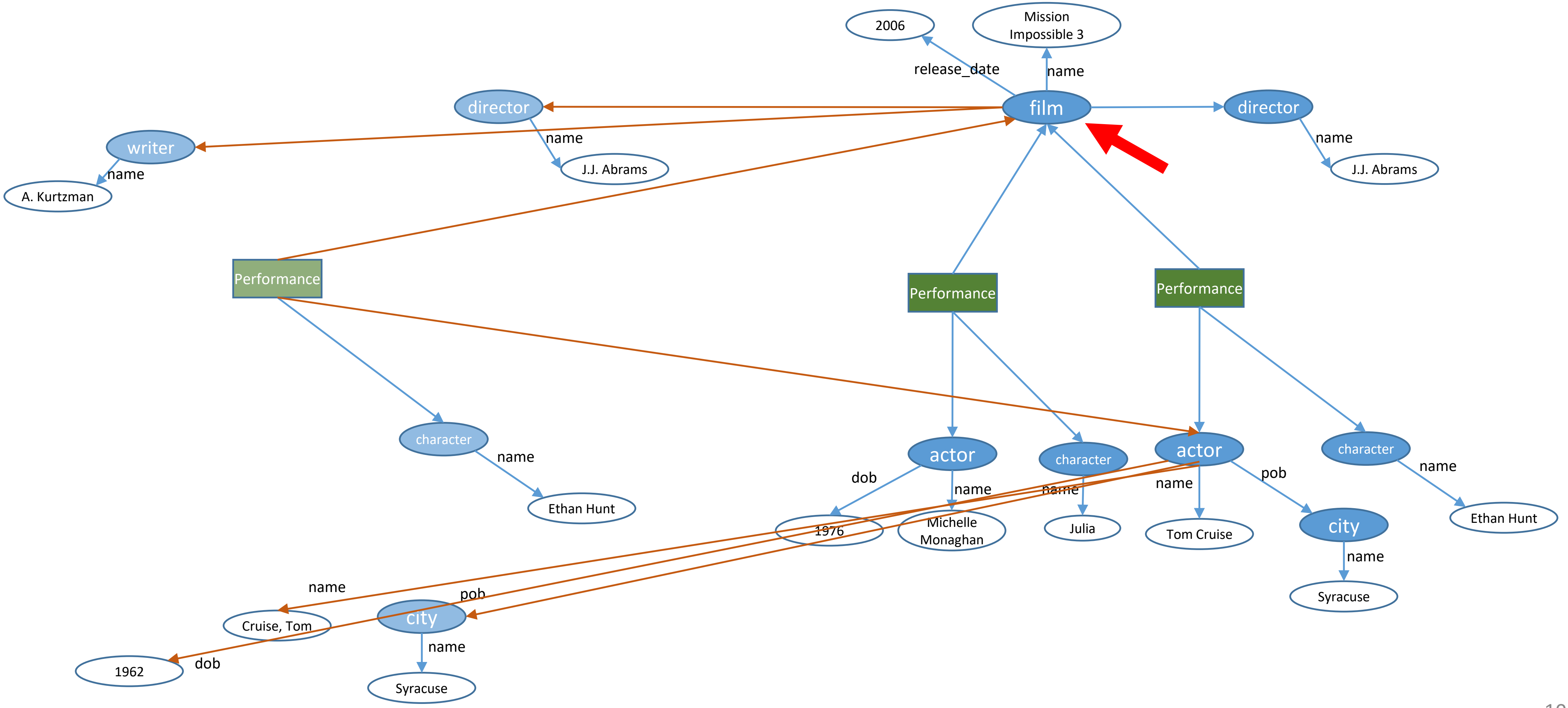


Similarity Signals Propagation

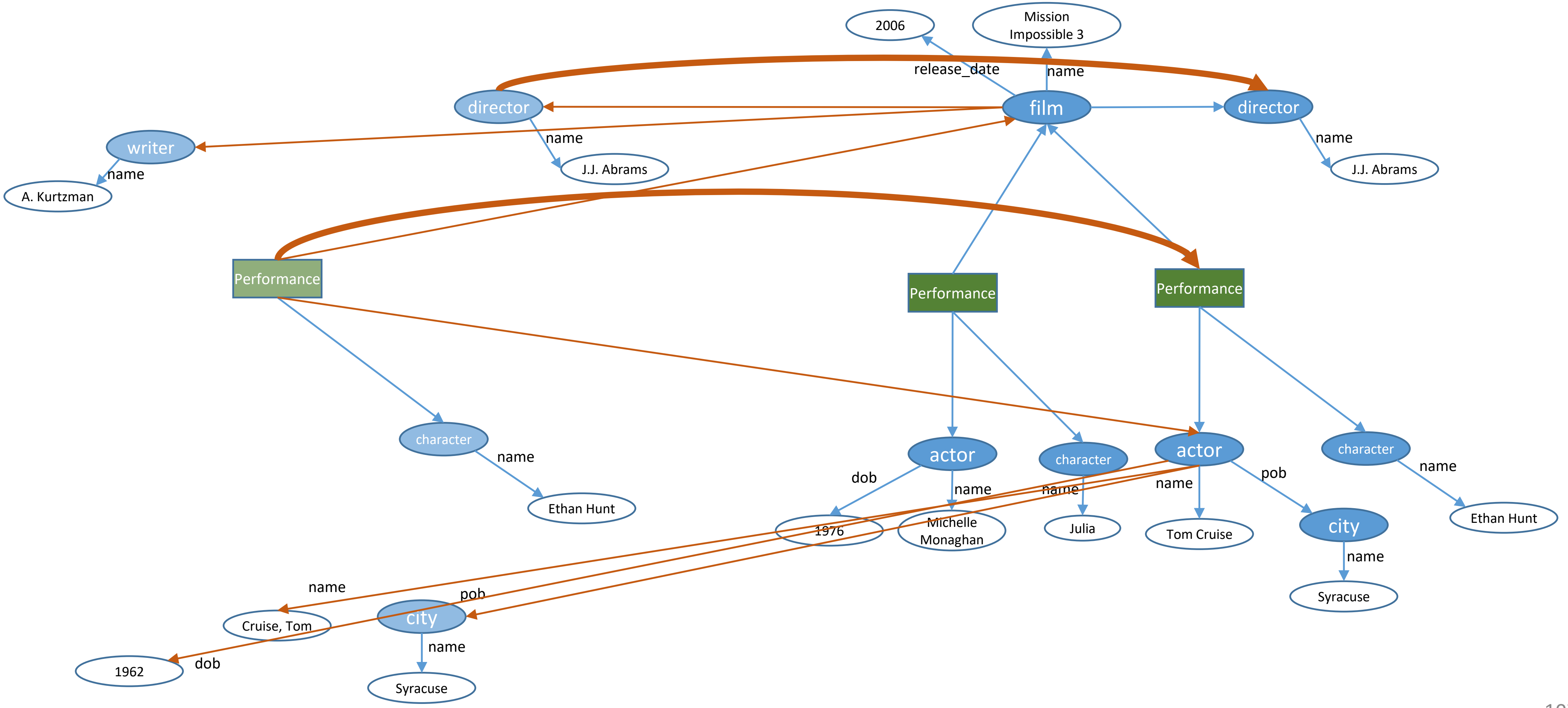
Graph Data Model and Conflation



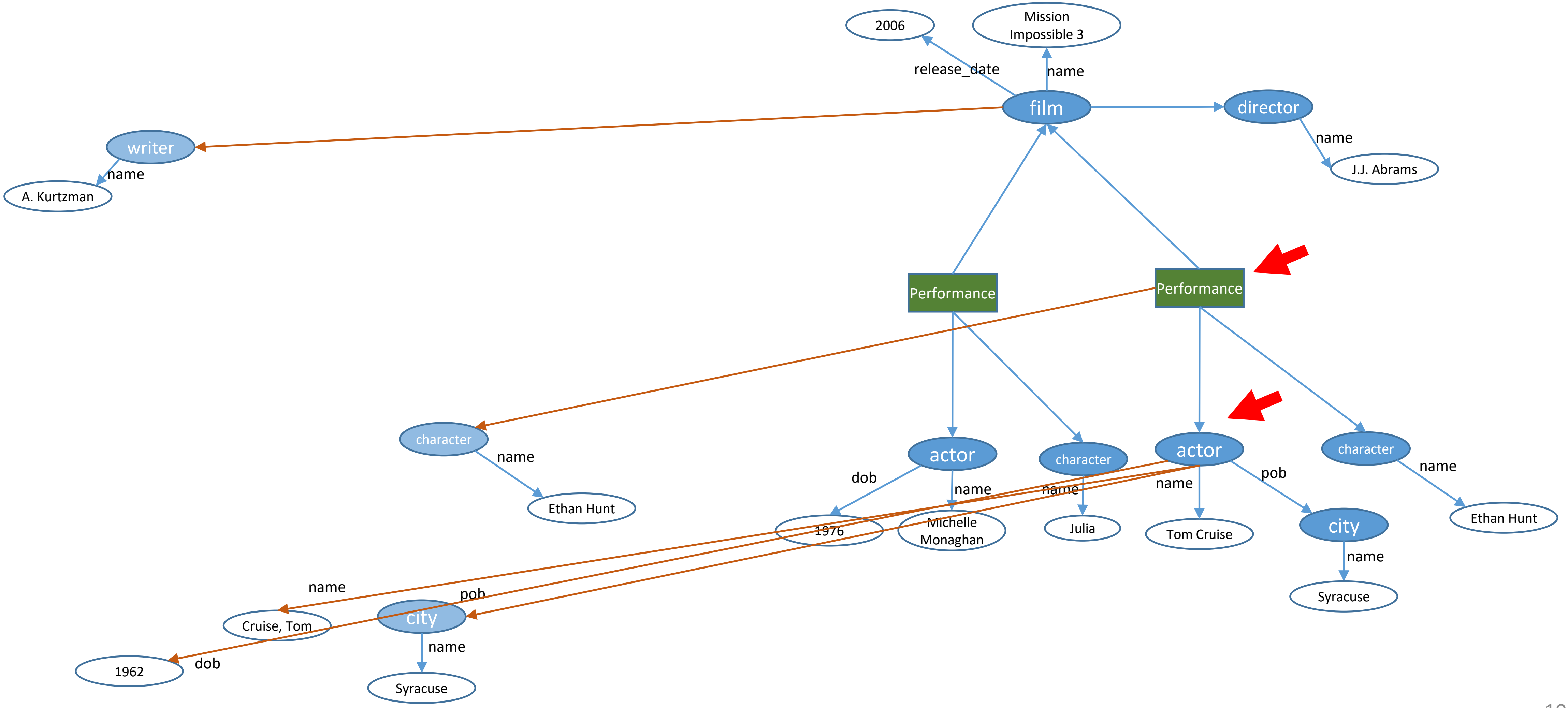
Graph Data Model and Conflation



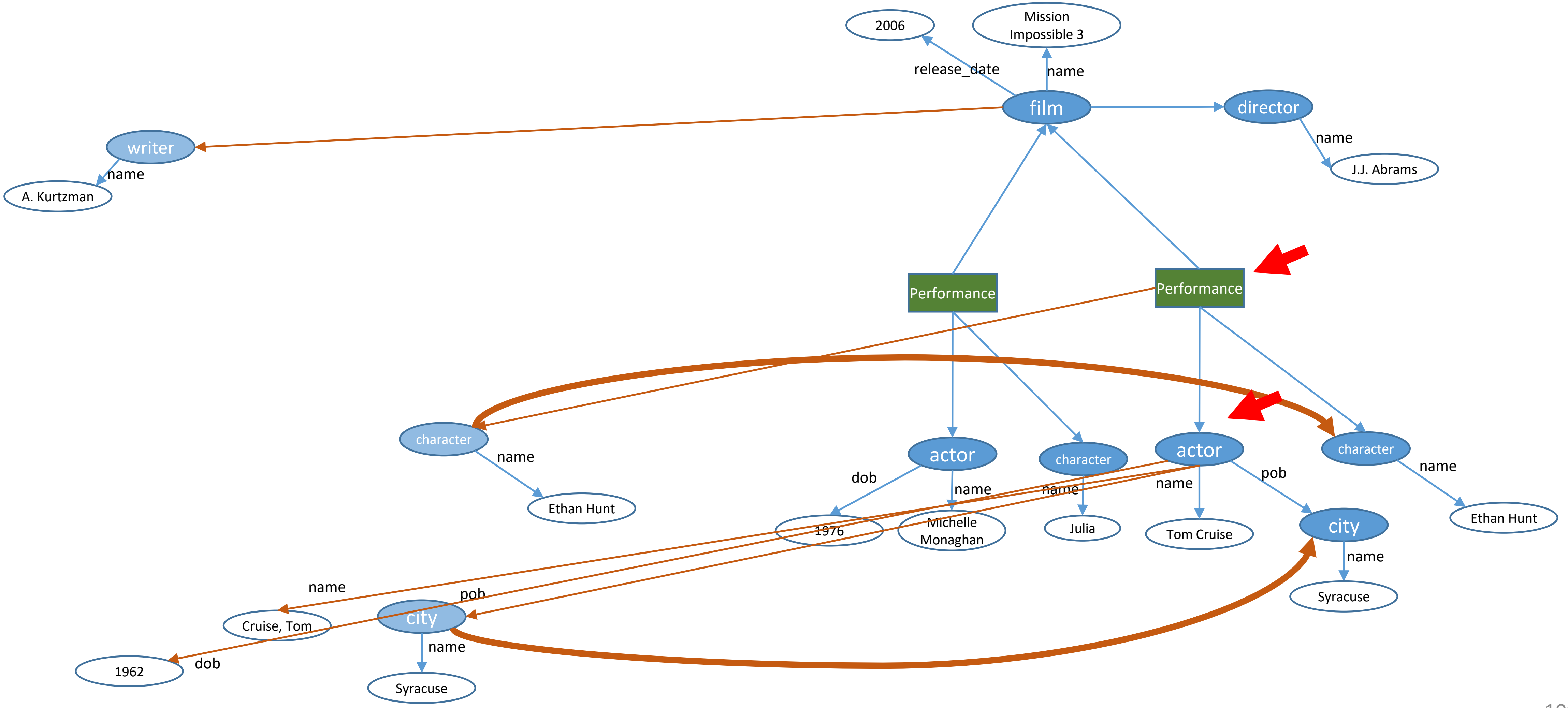
Graph Data Model and Conflation



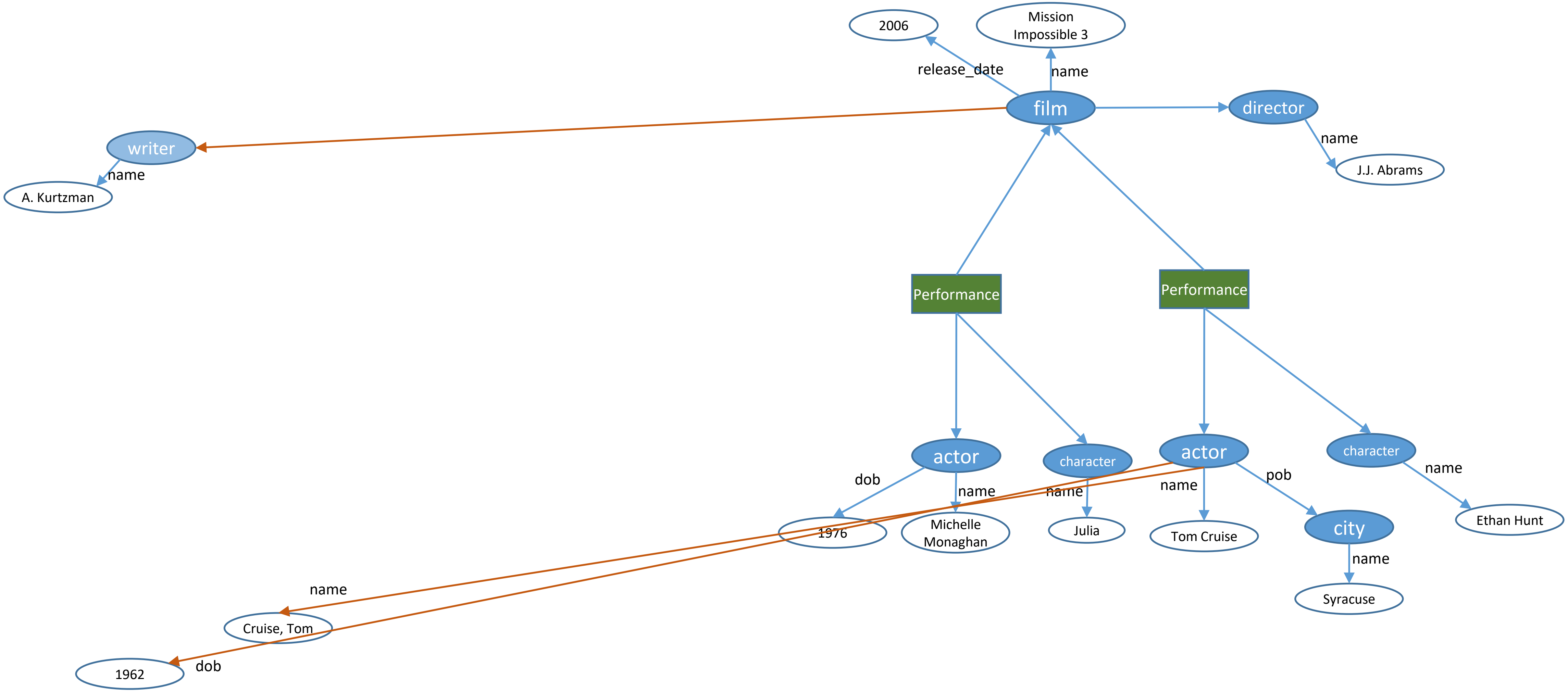
Graph Data Model and Conflation



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

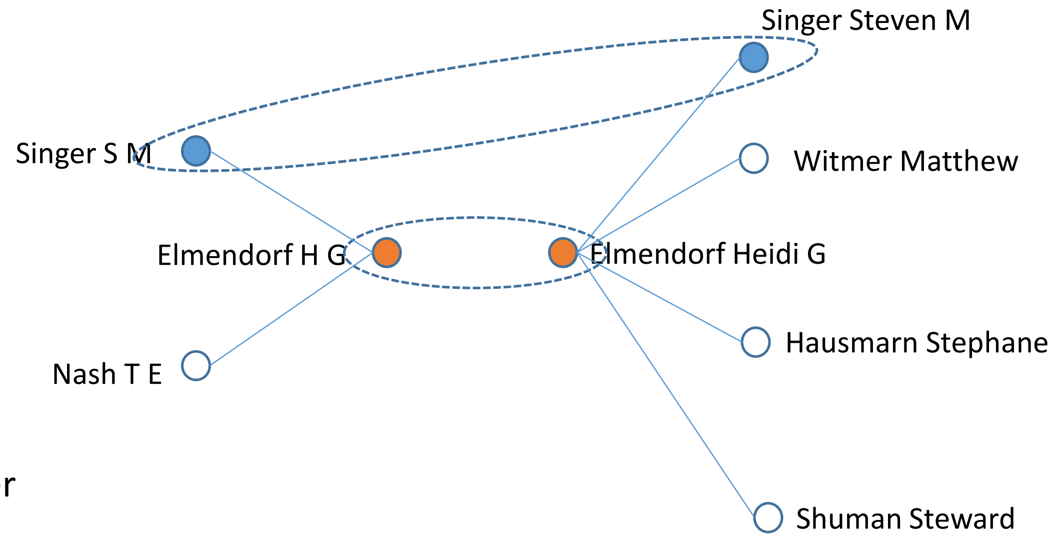
$$sim(a,b) = \sum_{i \in shared} \frac{1}{\log(freq(i))}$$

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

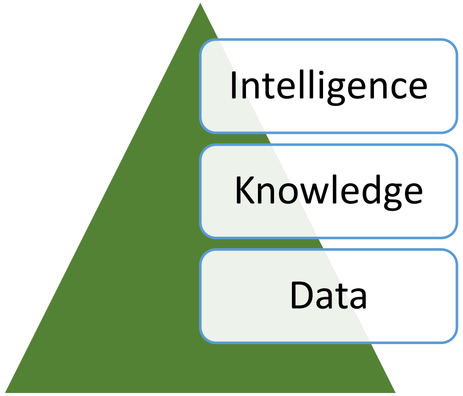
$$sim(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} sim(I_i(a), I_j(b))$$

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

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



Satori Graph Build



Data Ingestion

- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge

- Matching Entity Contents
- Detection of Matched Entities
- **Scaling & Improve Efficiency of Matching**

Knowledge Refinement

- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve

- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join

Improve Efficiency of Matching

- Matching two data sources each with 1 M entities
- 1M x 1M with an entity pair comparison time of 5 µs
- 160 years
- 300K machines to finish in 5 hrs

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

- **Solution:** Blocking or Indexing
  - Efficiency or Reduction Ratio = $\frac{|compared\ pairs|}{m \times n}$
  - Recall or pairs completeness = $\frac{|True\ Matches\ compared|}{|All\ existing\ true\ matches|}$

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(Tom Cruse)= {Tom Cruse}
- TokenSequence: H(Tom Cruse) = {crusetom}
- Metaphone: H(Robert)=H(Rupert)

- Q-Gram (a lot of hashes per value)
  - H(Smith)={smmiitth, miitth, smitth, smmith, smmiit}
  - (1) Compute grams (2) concat except one
  - 2-Gram(smith)={sm, mi, it, th}
  - H(Smith) ∩ H(Smithy) ∩ H(Smithe) ={smmiith}

- Suffix Array (a lot of duplicate post list)
  - H(Catherine)={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
- $J(S_1, S_2) = P_r(minhash(S_1) = minhash(S_2))$

- 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}
  - minhash(s_1)=c   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}
  - minhash(s_1)=g   minhash(s_2)=g
- If 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 x 0.4) = 1 – 0.16 = 0.84

Improve Efficiency of Matching

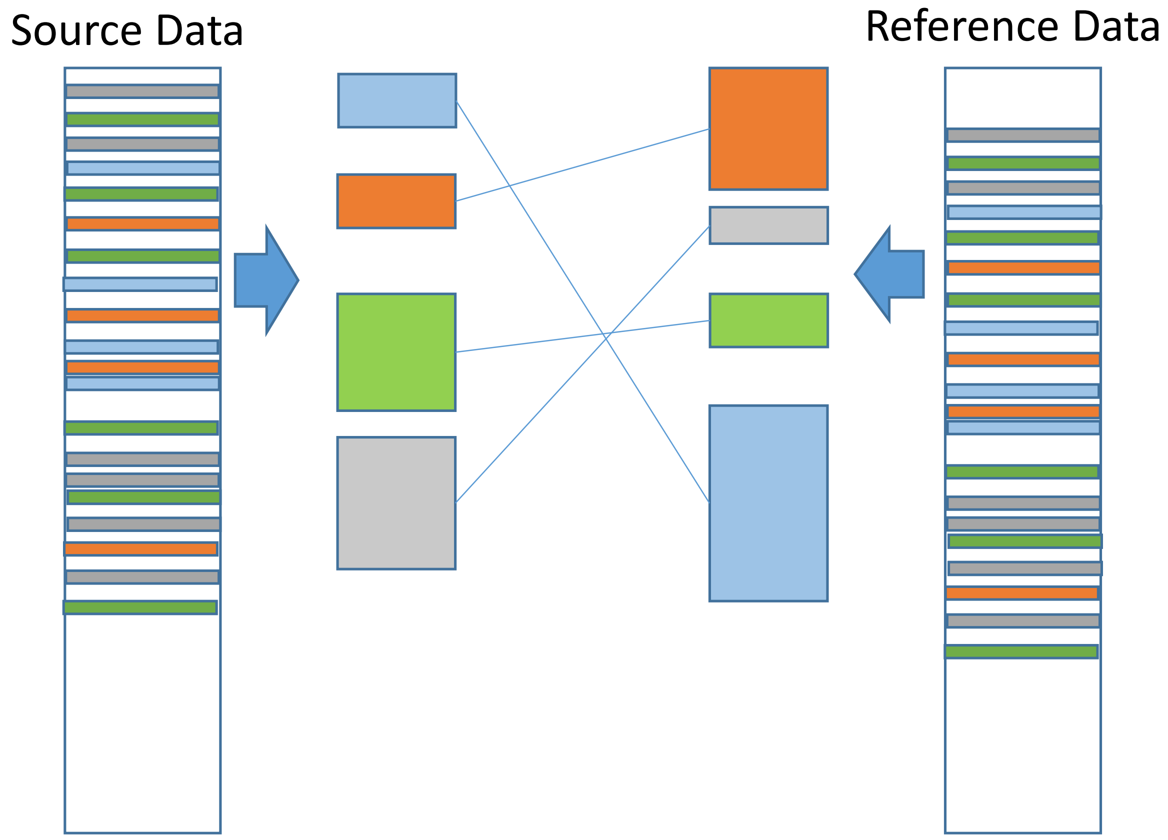
- Reduce the number of entities comparisons (Indexing or Blocking)

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)

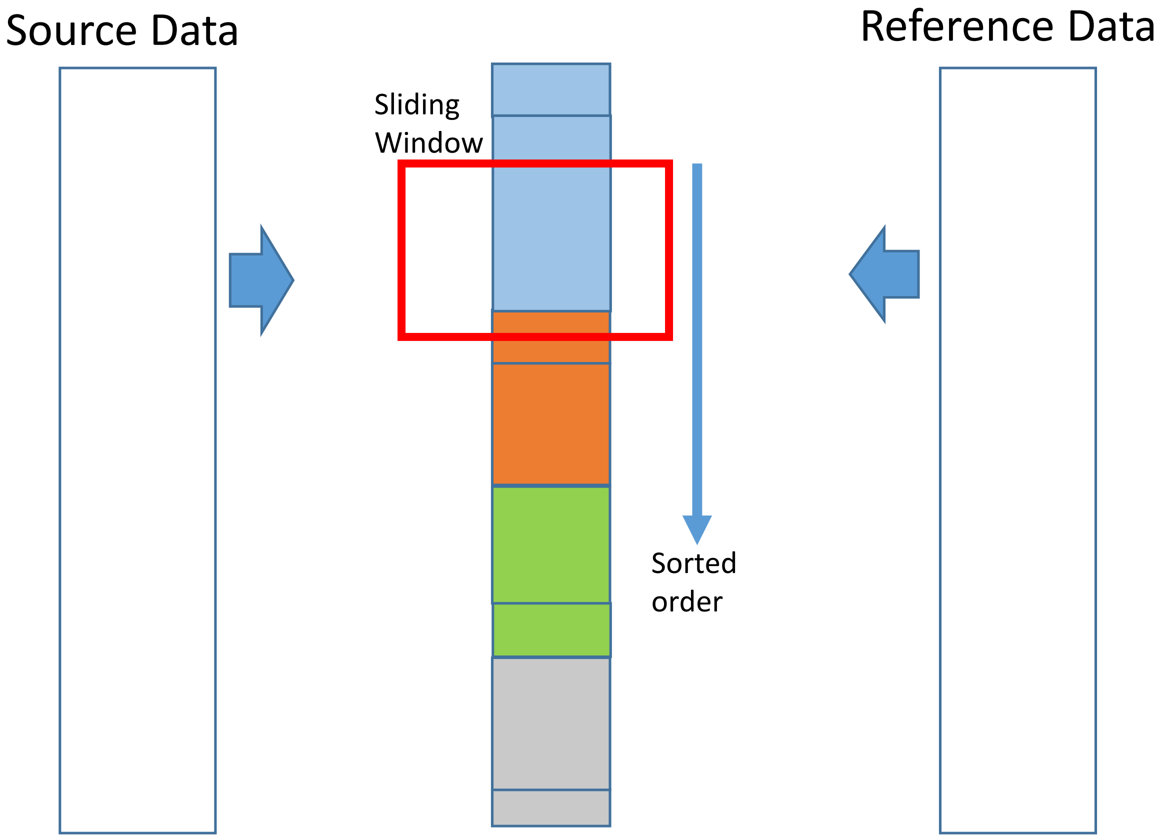
3. Retrieval
    - **Within blocks comparison**
    - Sorted Neighborhood
    - Canopy Clustering (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
    - Entity Index Join



Improve Efficiency of Matching

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

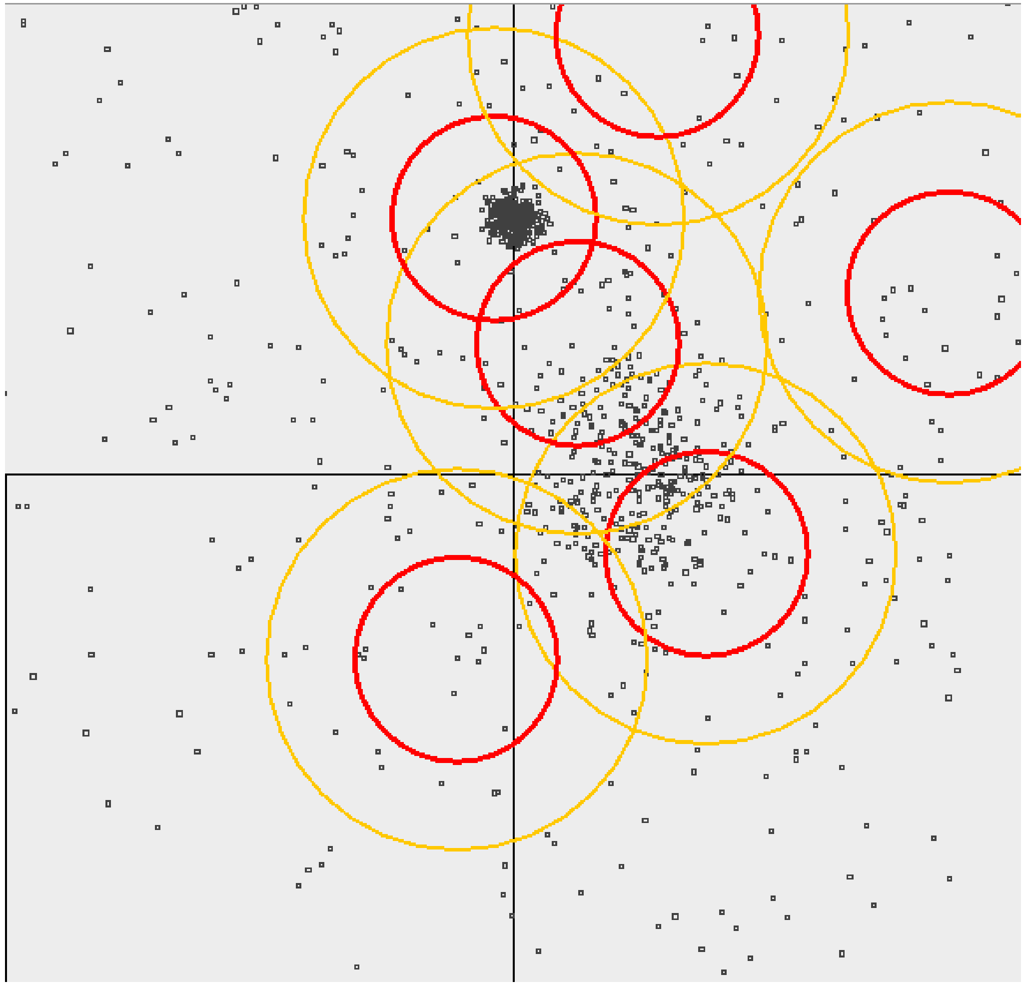
3. Retrieval
   - Within blocks comparison
   - **Sorted Neighborhood**
   - Canopy Clustering (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
   - Entity Index Join



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**

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

Inverted Index

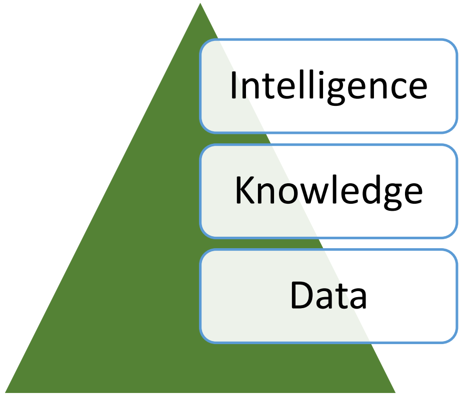
Key	IDF	Post List
h1	IDF(h1)	E1, E2
h2	IDF(h2)	E1
h3	IDF(h3)	E1, E2, E3
h4	IDF(h4)	E2, E4

- $E_1 = \{(h, idf_1(h)) \dots\}$
- $E_2 = \{(h, idf(h)) \dots\}$

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

Return Top K entities for each other entity

Satori Graph Build



Data Ingestion

- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge

- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement

- **Knowledge Fusion**
- Error Detection
- Fact Inference

Publishing & Serve

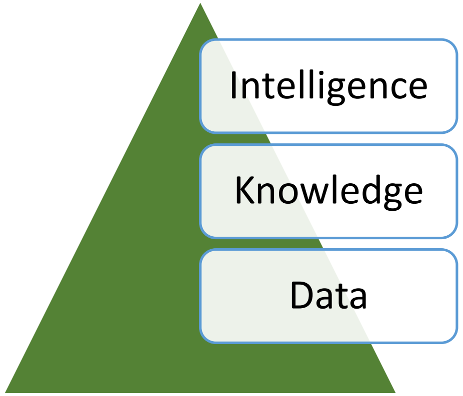
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join

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

Dong el al, "From data fusion to knowledge fusion" VLDB 2014
Dong et al, "Fact checking: theory and practice." KDD 2018

Satori Graph Build



Data Ingestion

- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge

- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement

- Knowledge Fusion
- **Error Detection**
- Fact Inference

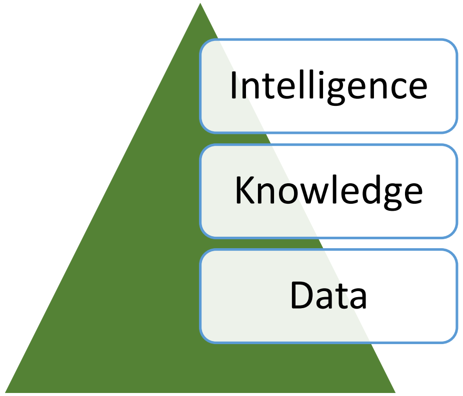
Publishing & Serve

- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join

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



Data Ingestion → **Match & Merge** → **Knowledge Refinement** → **Publishing & Serve**

Data Ingestion

- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge

- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement

- Knowledge Fusion
- Error Detection
- **Fact Inference**

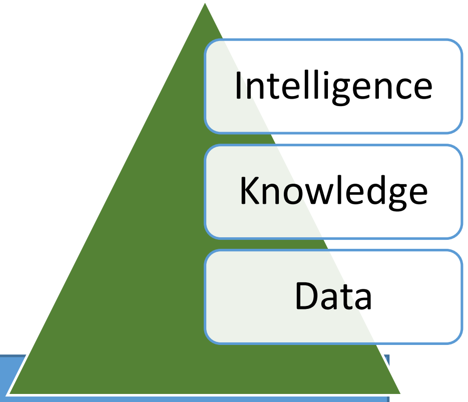
Publishing & Serve

- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join

Fact Inference

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

Satori Graph Build



Data Ingestion → **Match & Merge** → **Knowledge Refinement** → **Publishing & Serve**

Data Ingestion

- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge

- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement

- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve

- **Entity search API**
- **Ranking & Filtering by attributes**
- **Graph walk**
- **Semantic Linking & Join**

Part IV: Serving Knowledge to the World

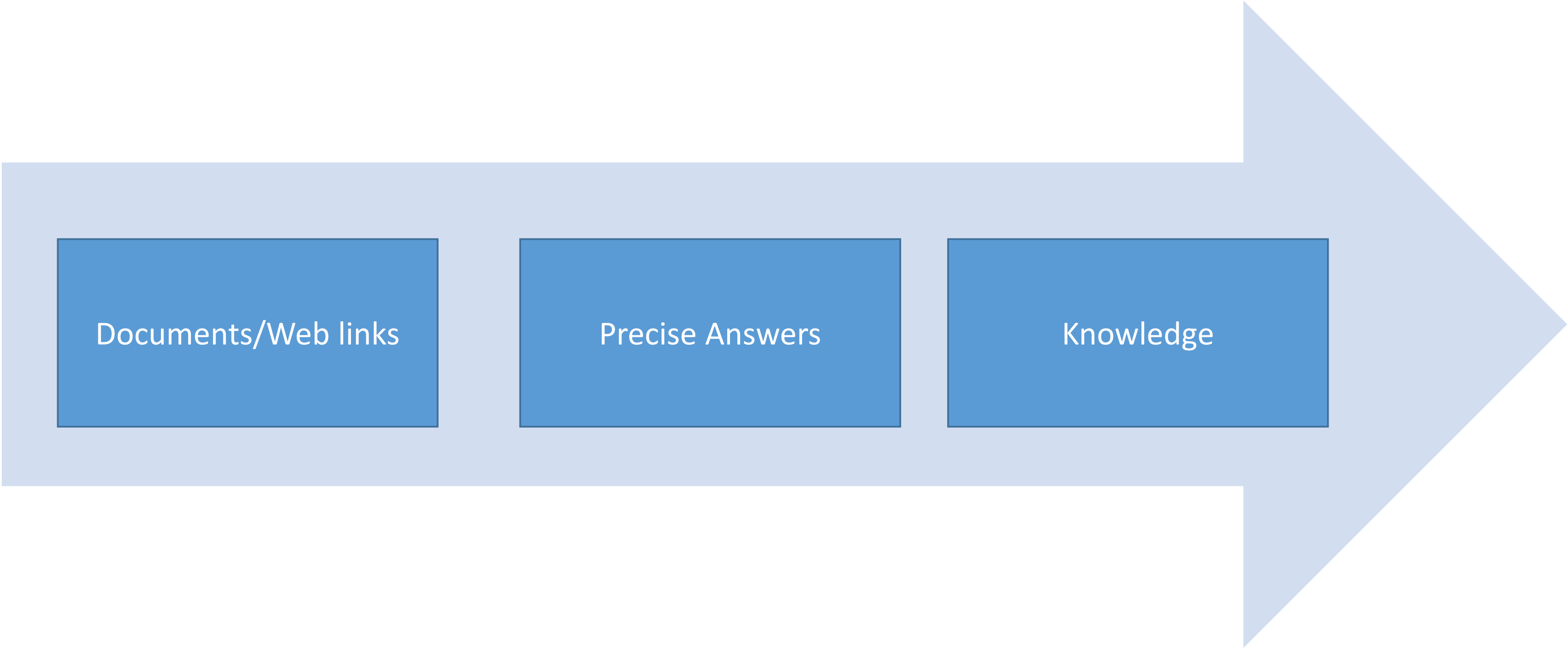
Ahmed K. A. Mohamed

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

ahmedat@microsoft.com

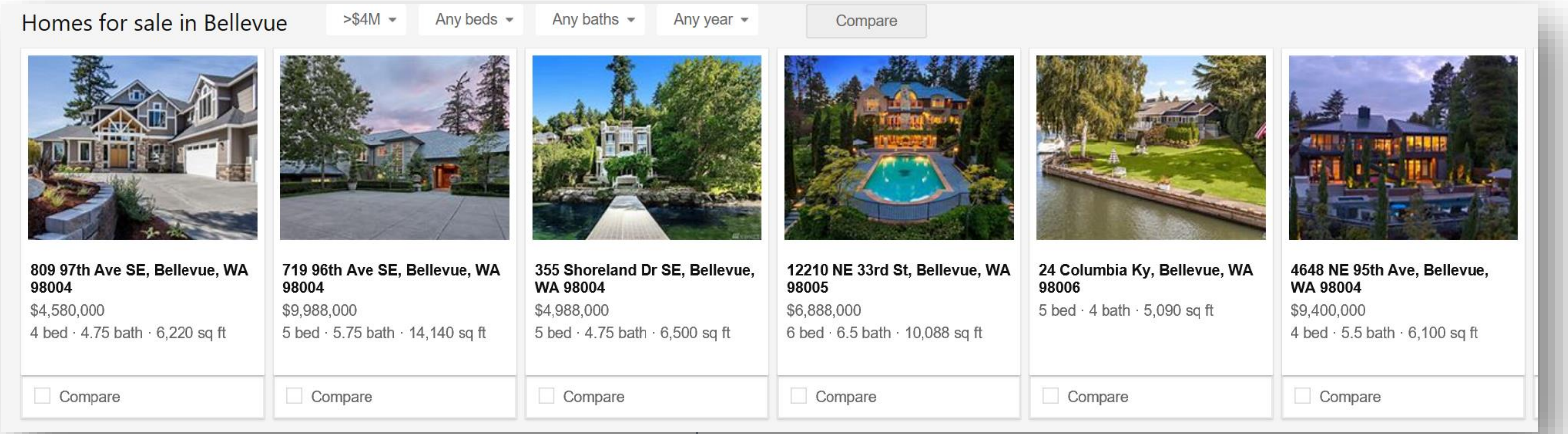
Microsoft

Why Knowledge Graph Serve

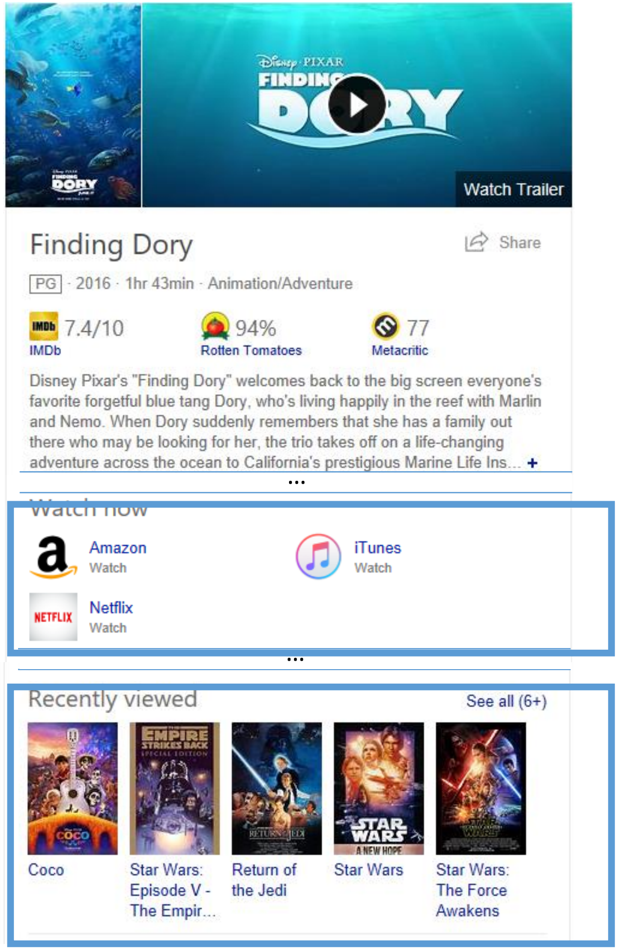


Satori Knowledge Graph Application Areas: Bing&Cortana

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



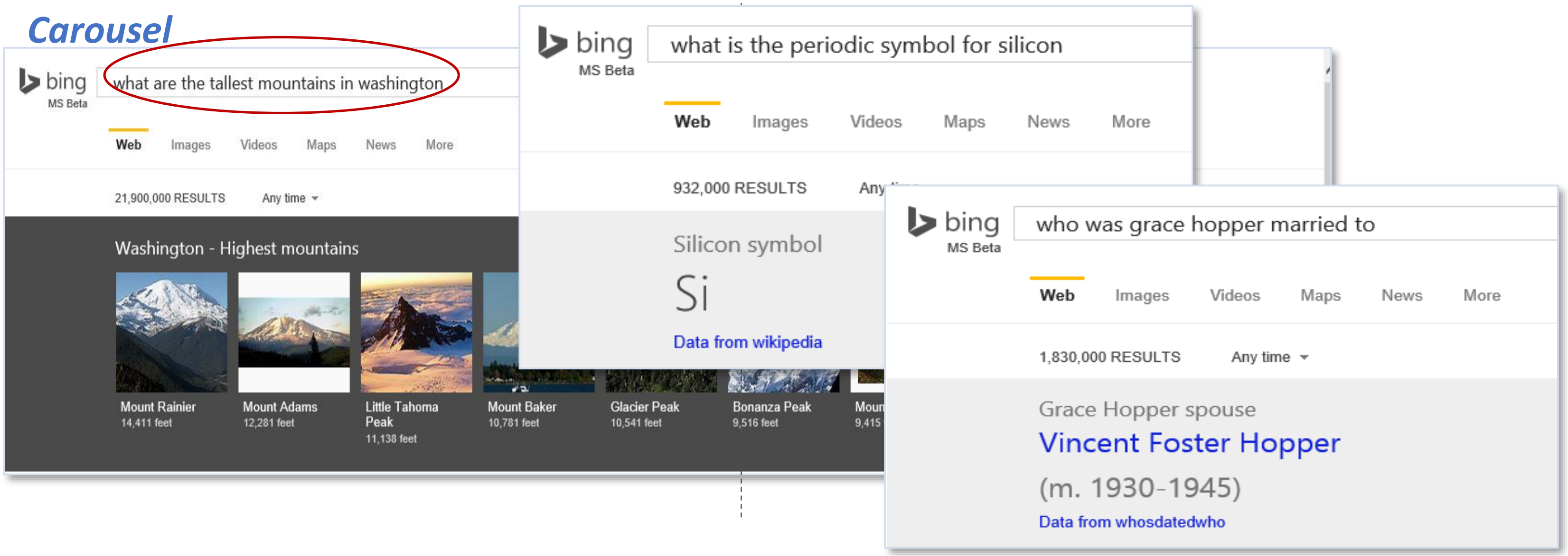
Carousel of information from Satori



Watch now actions for movie entities

Recently viewed shows personal history

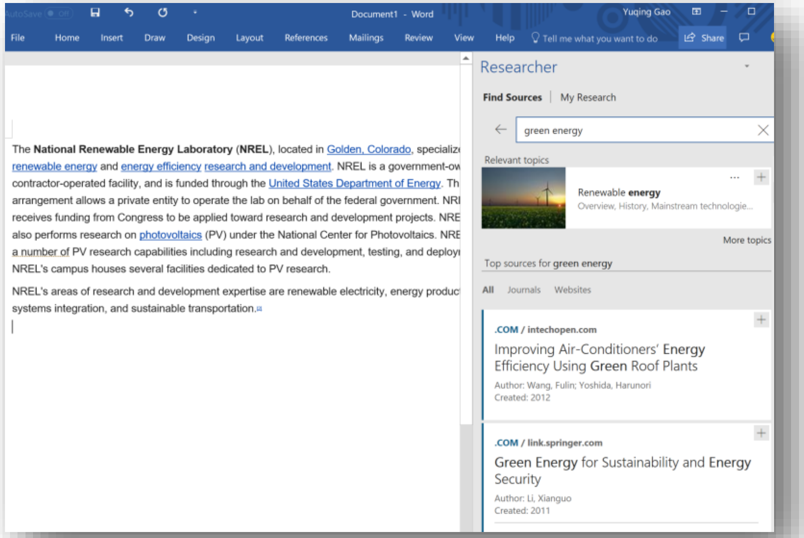
Carousel



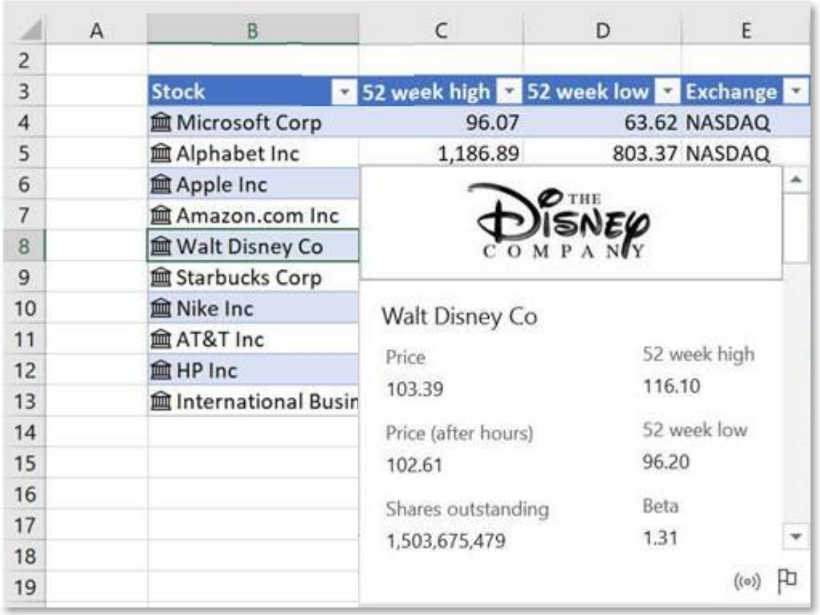
Natural language **Fact Answers**

Satori Knowledge Graph Application Areas: Office

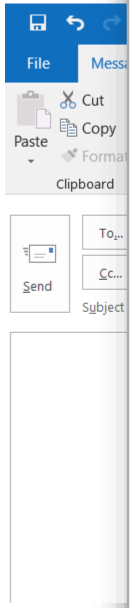
Enriching the Office experience with Satori data



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



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



Challenges

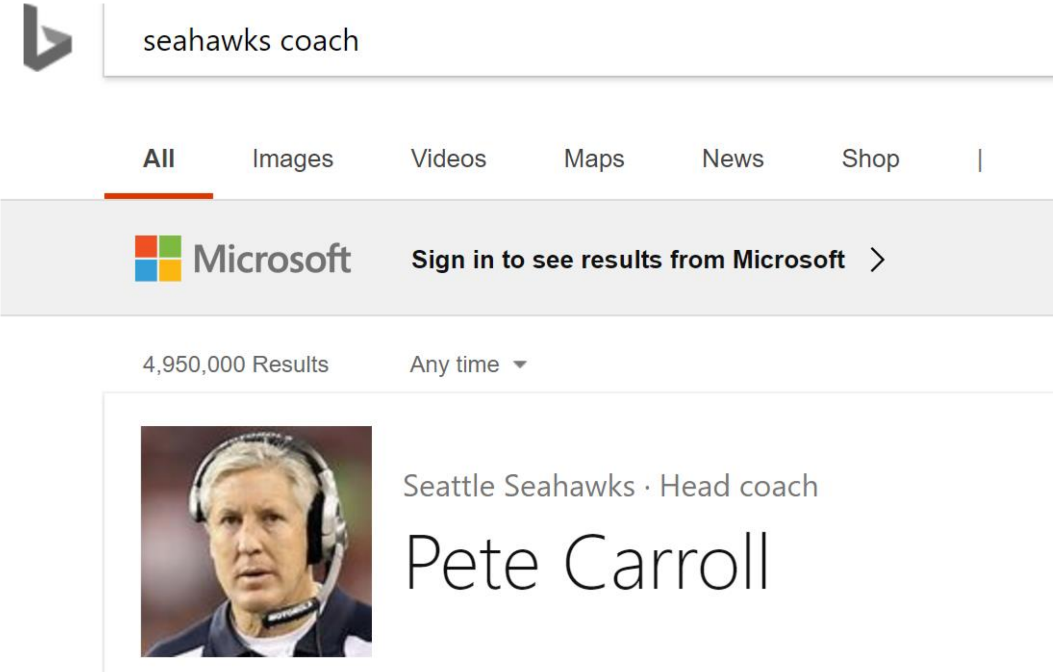
- Online conflation of people
- Compliance (training, code scanning & fixing, onboarding to new tools, etc.)
- Relevance with sparse profiles w/o access to the raw queries

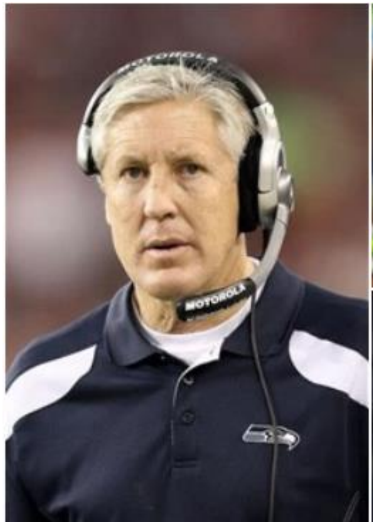
LinkedIn profile information visible in O365 People Card through Satori

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





Seattle Seahawks /American_football_team_current_head_coach

Serving Knowledge by Answering Questions

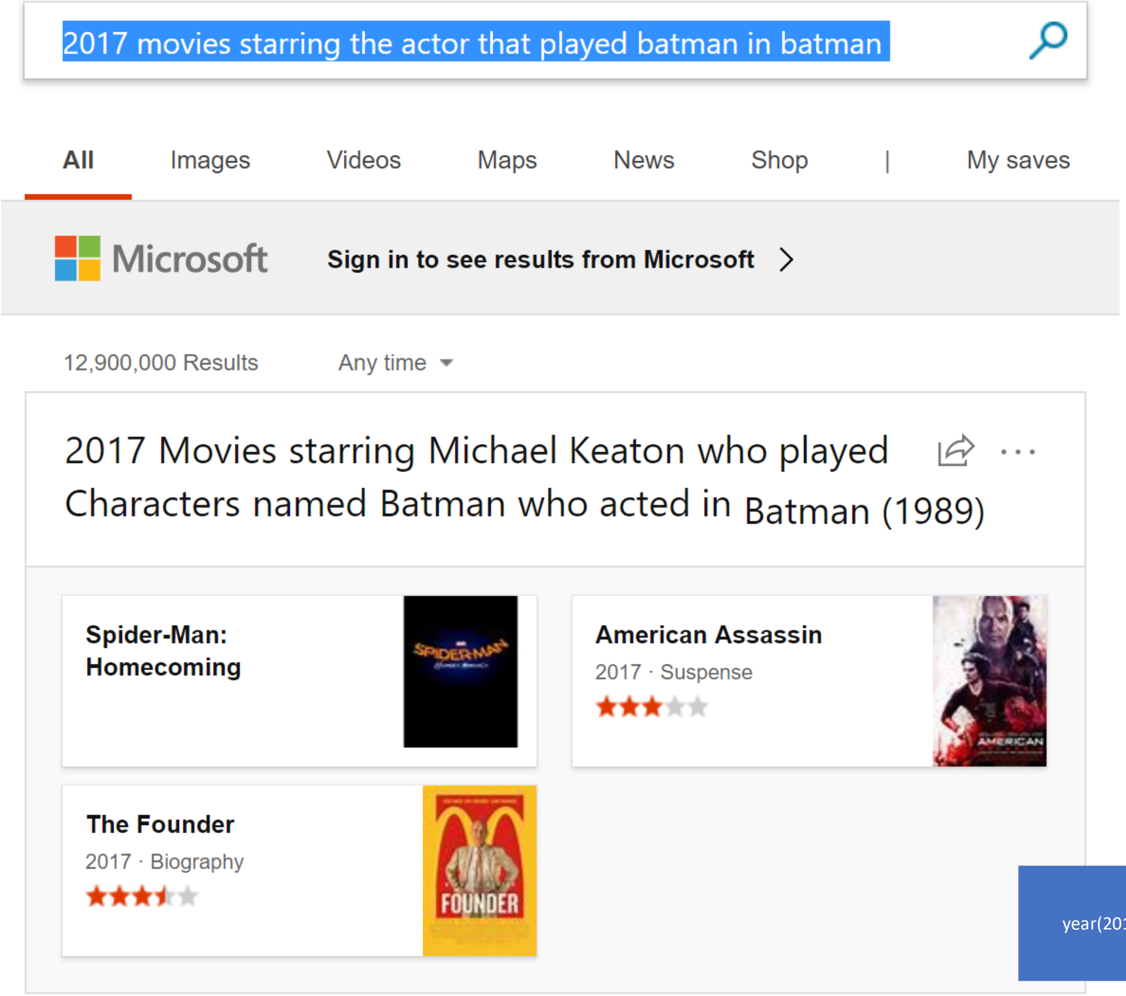
- Challenges:
  - Matching language
    - There are many ways to ask the same query e.g. {who directed titanic}, {what is the name of the person who directed titanic}, {in the movie titanic, who was the director}, …etc
    - Scalable entity linking
    - Word sense disambiguation
    - Semantic roles 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}

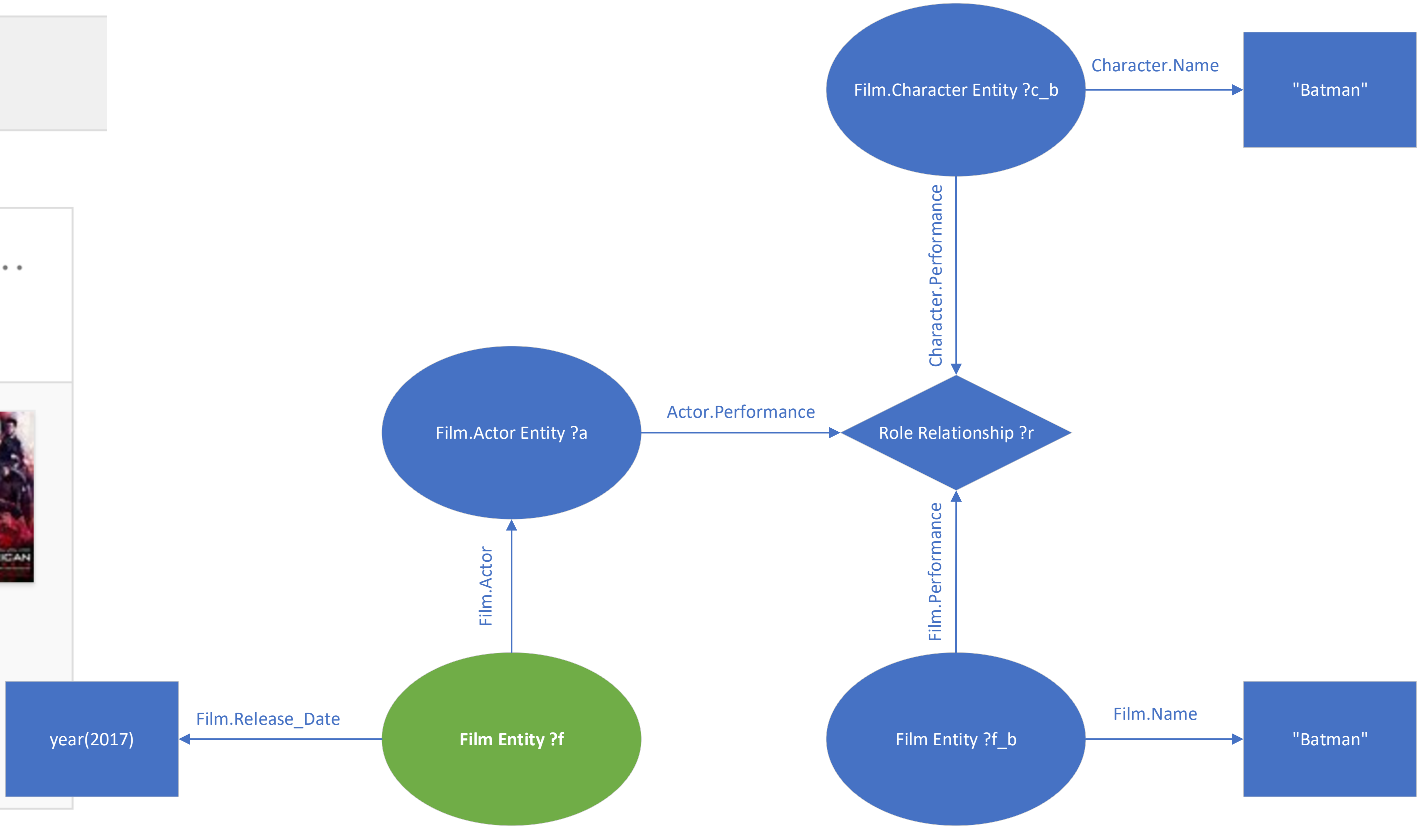
Serving Knowledge by Answering Questions

- 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





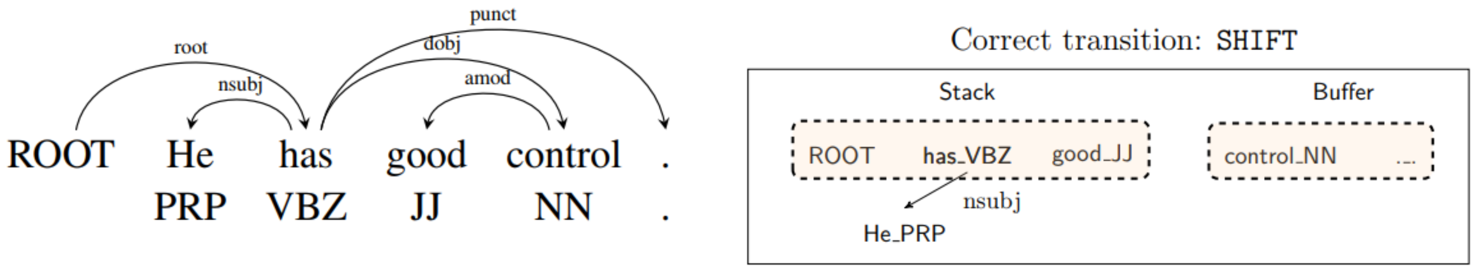
1.1 Generic Semantic Parsing

- 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. $\lambda x. child\_of(Donald\ Trump, x)$ ^ $gender\ (x, female)$

- 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]



Transition	Stack	Buffer	A
	[ROOT]	[He has good control .]	∅
SHIFT	[ROOT He]	[has good control .]	
SHIFT	[ROOT He has]	[good control .]	
LEFT-ARC(nsubj)	[ROOT has]	[good control .]	$A\cup$ nsubj(has,He)
SHIFT	[ROOT has good]	[control .]	
SHIFT	[ROOT has good control]	[.]	
LEFT-ARC(amod)	[ROOT has control]	[.]	$A\cup$amod(control,good)
RIGHT-ARC(dobj)	[ROOT has]	[.]	$A\cup$ dobj(has,control)
...	...	...	...
RIGHT-ARC(root)	[ROOT]	[]	$A\cup$ root(ROOT,has)

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]

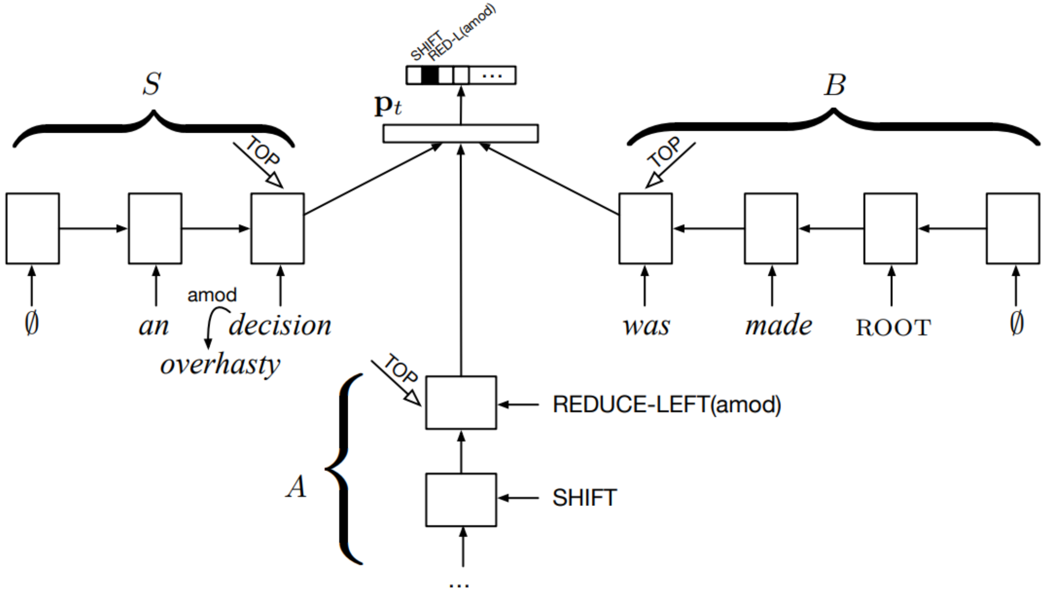
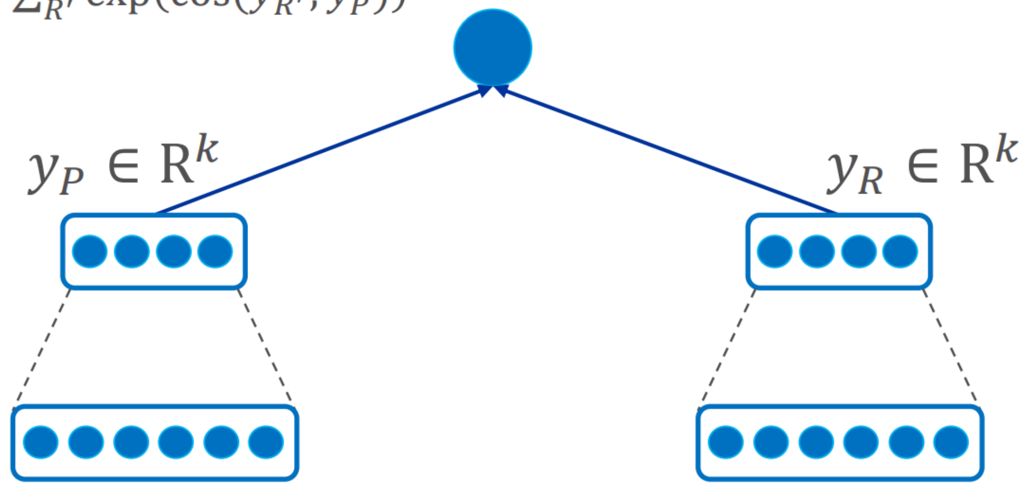


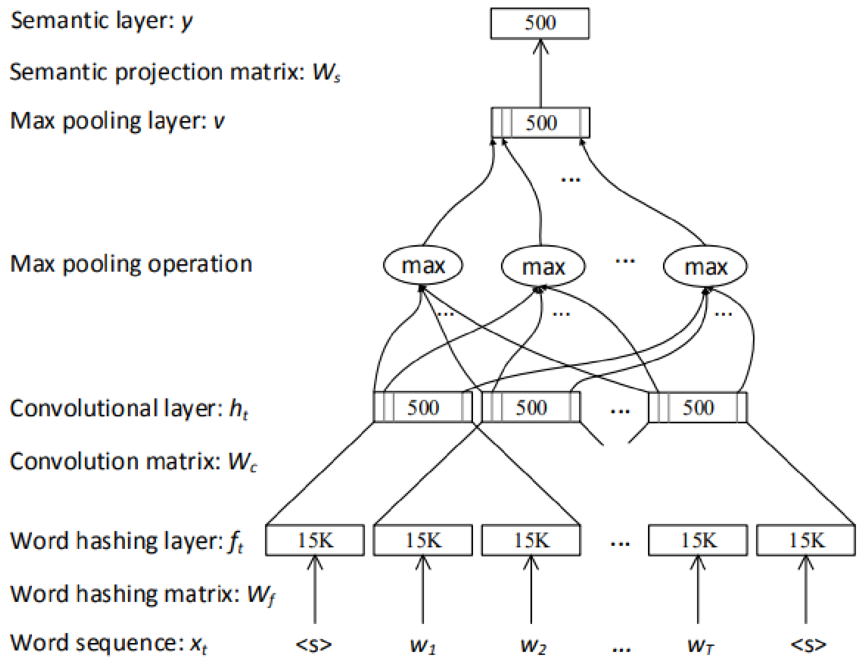
Figure 2: Parser state computation encountered while parsing the sentence "*an overhasty decision was made*." Here S designates the stack of partially constructed dependency subtrees and its LSTM encoding; B is the buffer of words remaining to be processed and its LSTM encoding; and A is the stack representing the history of actions taken by the parser. These are linearly transformed, passed through a ReLU nonlinearity to produce the parser state embedding $\mathbf{p}_t$. 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

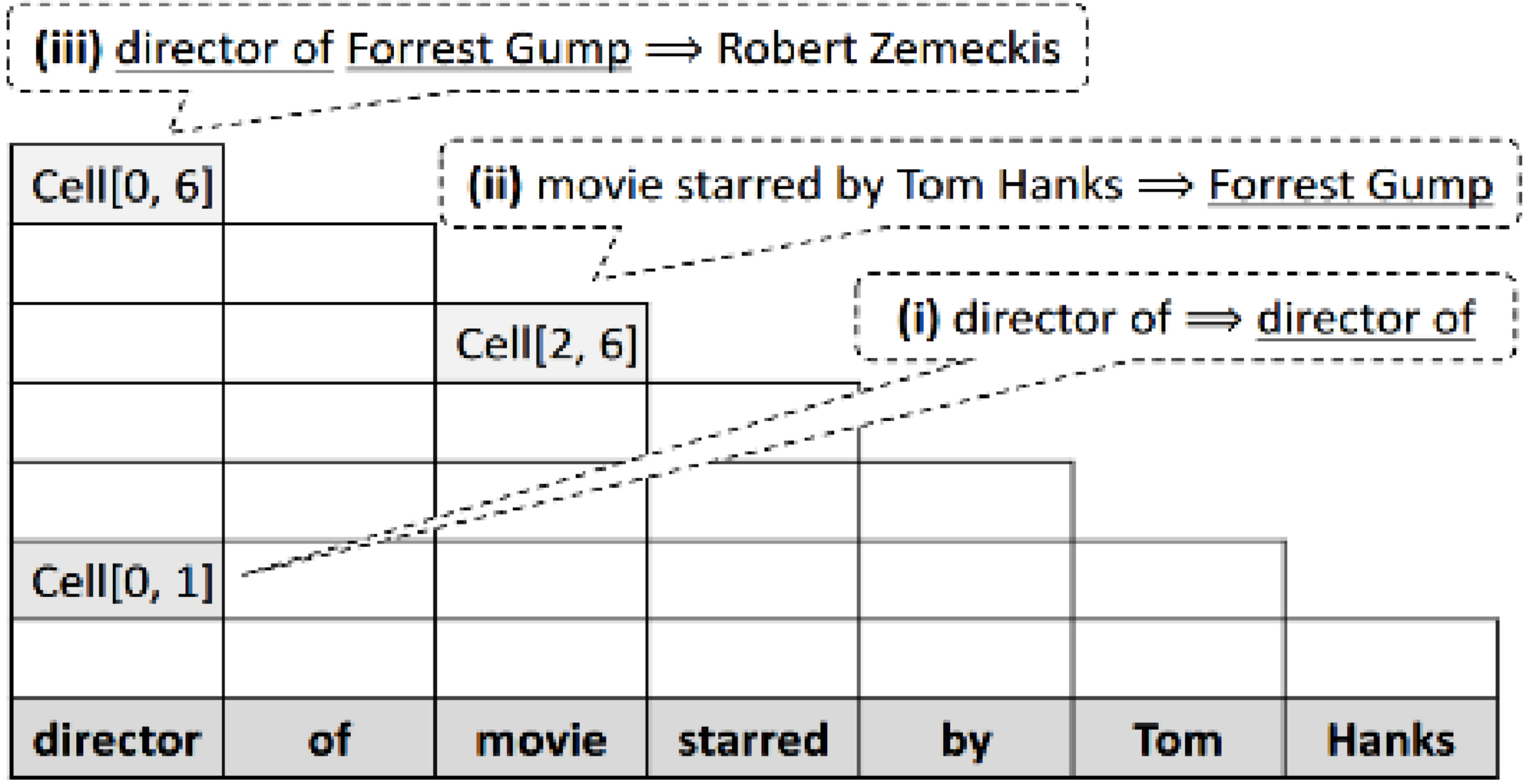


Fig.1 of [Bao et al., 2014]

Constituency parsers: PCFG Chart Parsing

Grammar is learned independently from an annotated dataset

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

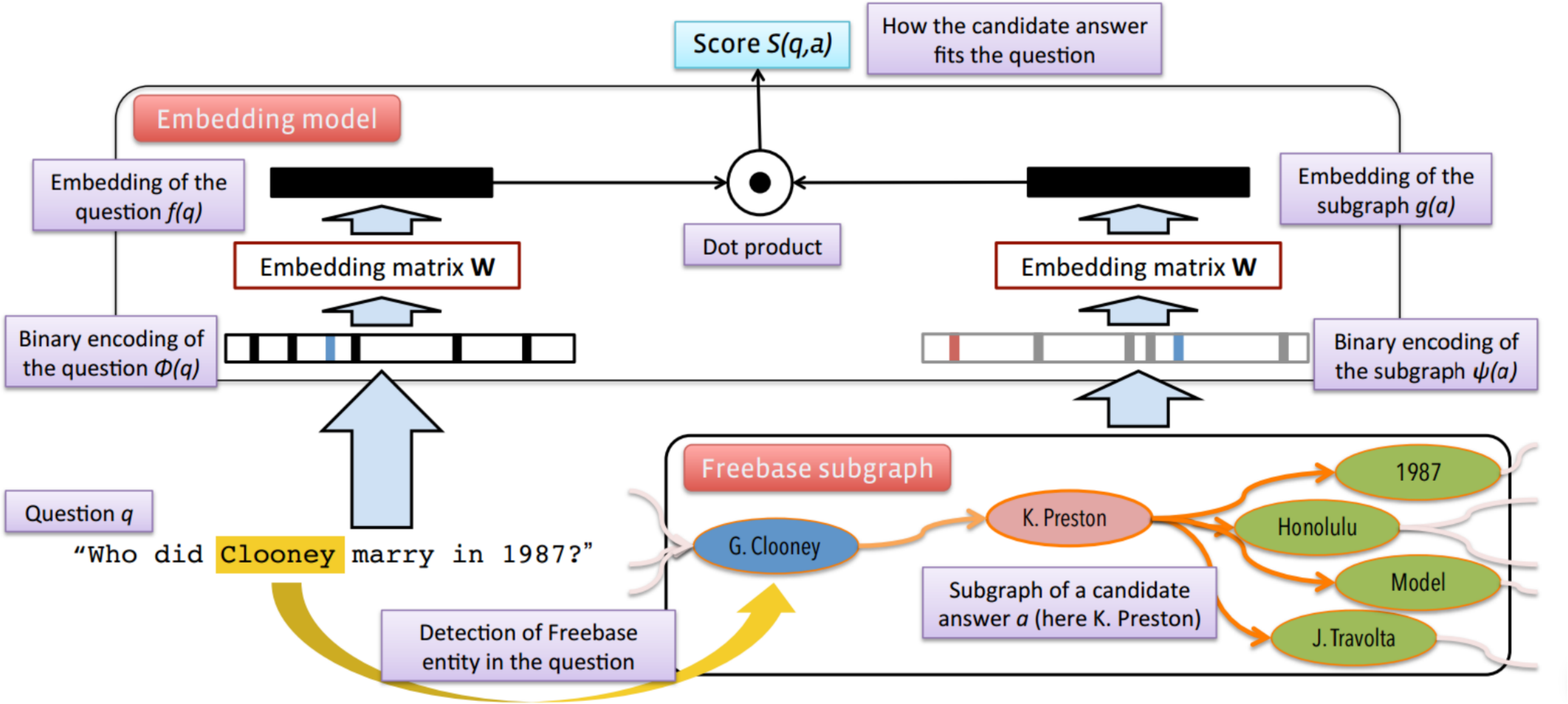
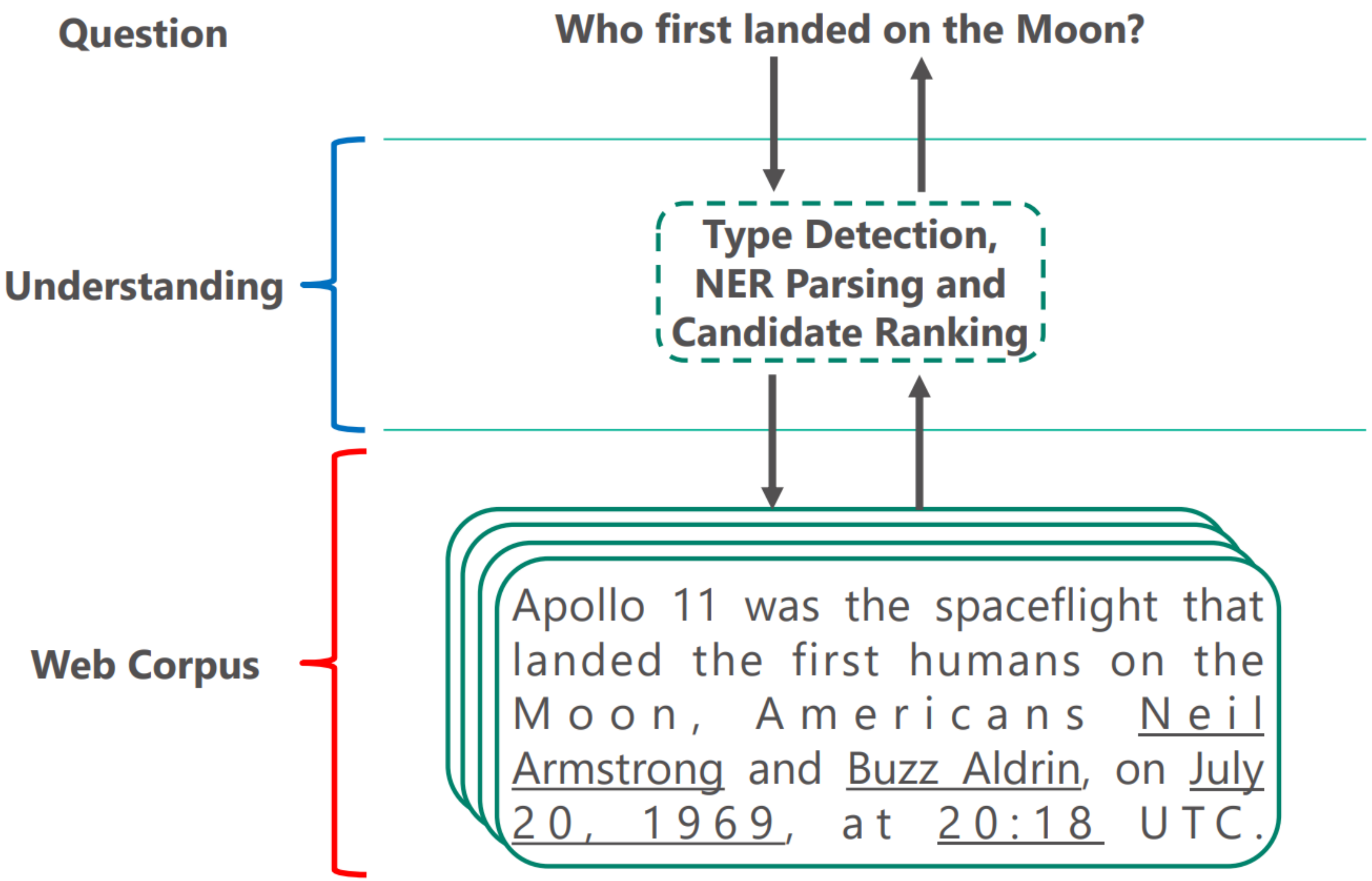


Fig.1 of [Bordes et al., 2014]

2. Information extraction approaches

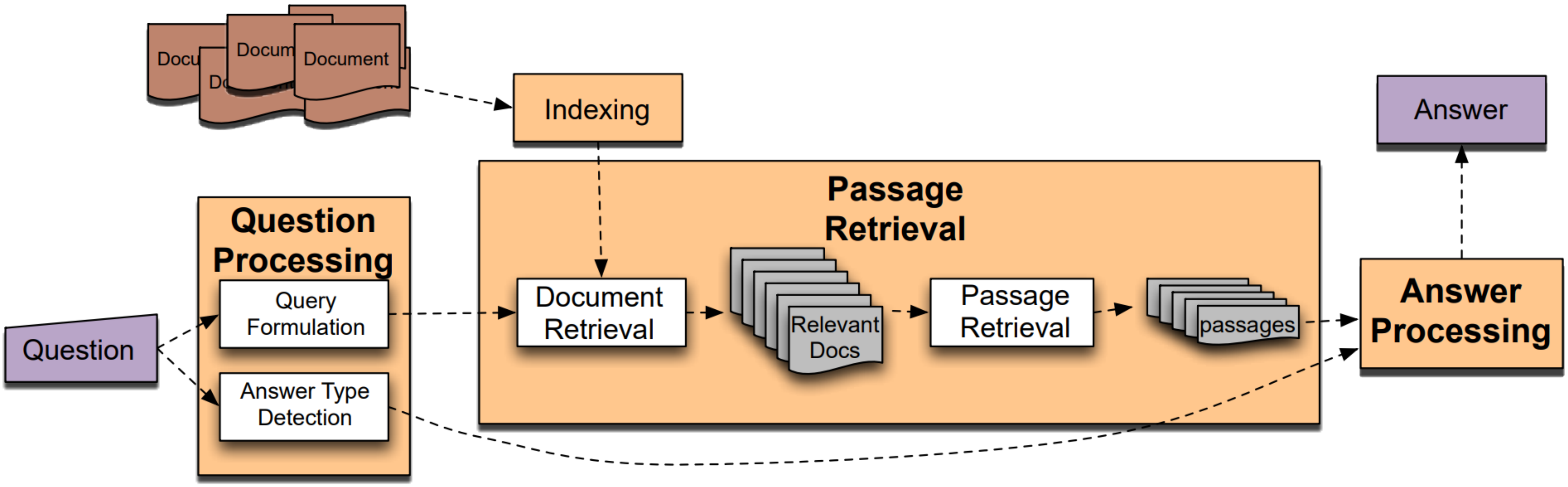
- Extracting and answers on the fly.
- These approaches provide ways to leverage the knowledge graph in cases where the question cannot be covered by the ontology or the data or both.

Information extraction approaches



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

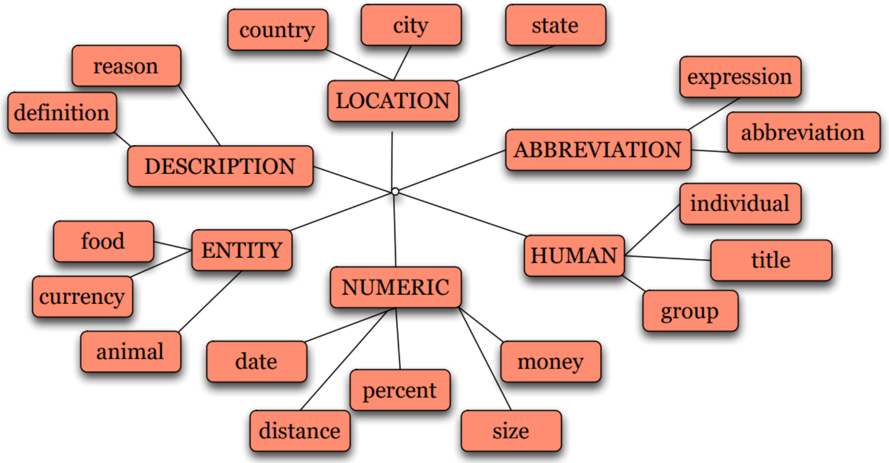
Information Extraction Approach

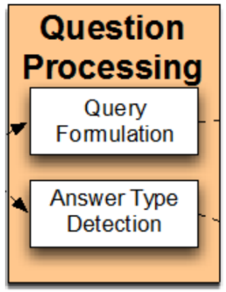


Question Answering [Dan Jurafsky, Stanford]

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 Retreival

- 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

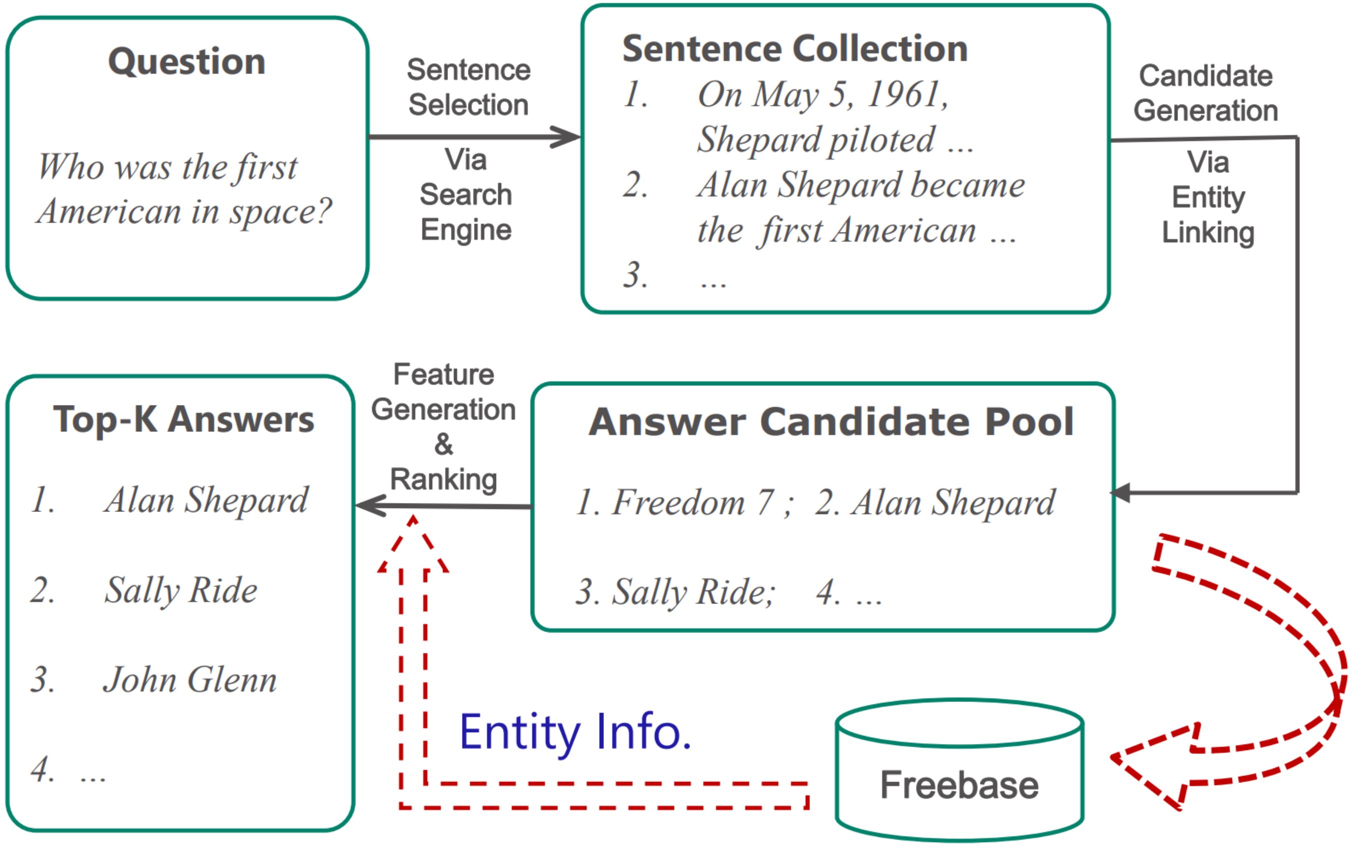
- Detect answer entity by running NER on the passage
- Mark the answer entity in the passage

- 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.

Question Answering [Dan Jurafsky, Stanford]

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

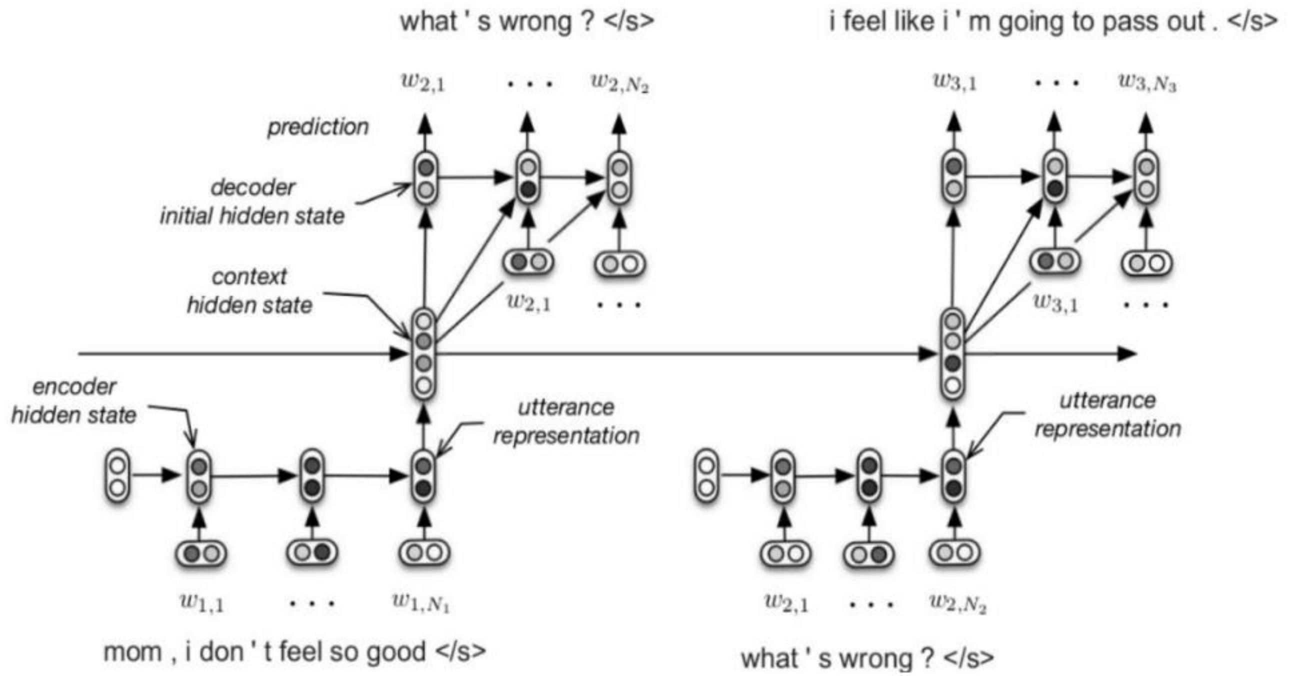
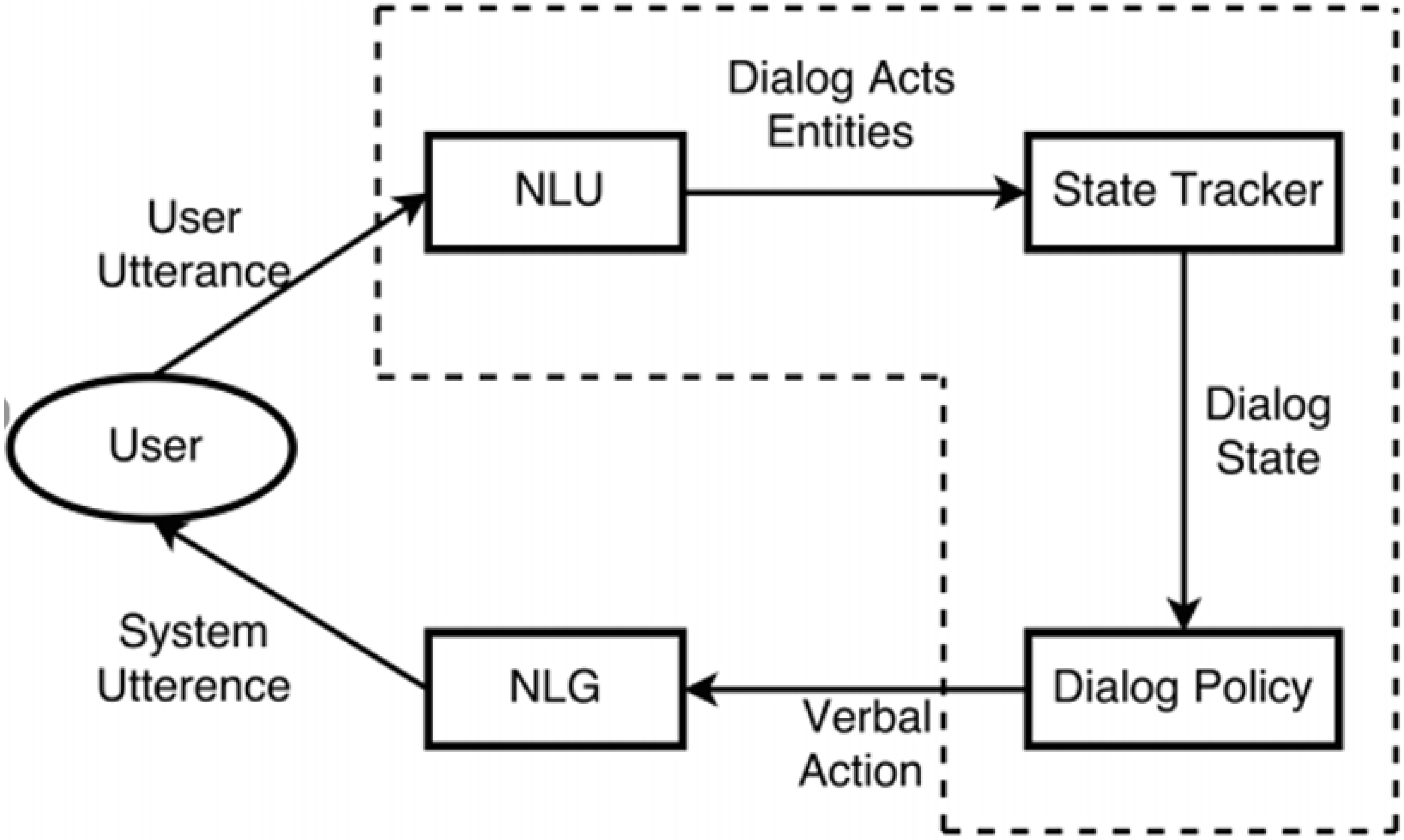


Figure 8: A computational graph representing the HRED architecture for dialogue over a span of three turns. The major addition to the architecture is a higher-level *context*-RNN keeping track of past utterances by progressively processing over time each utterance vector and conditioning the decoding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

Knowledge Based Ontological Slot Filling



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

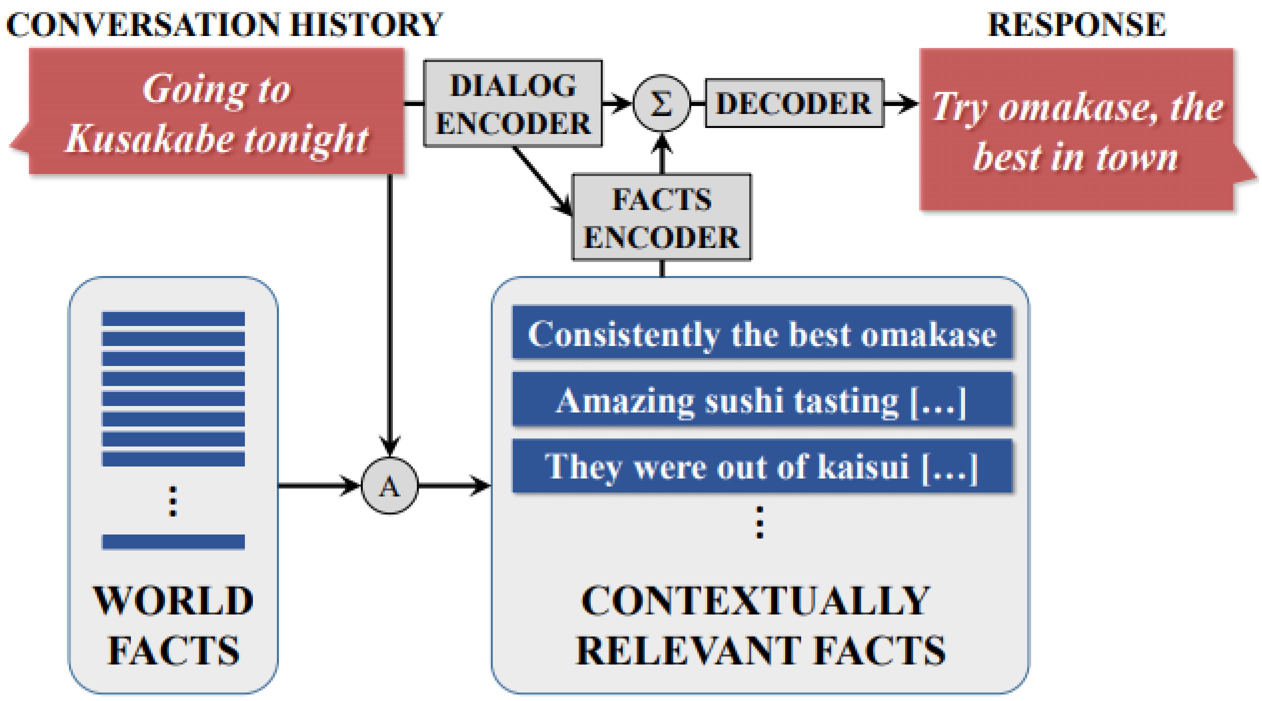


Figure 3: Knowledge-grounded model architecture.

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