Knowledge graph (KG)

COMPCSI 752: Big data management

University of Auckland

Slides are collected and edited from https://web.stanford.edu/class/cs520/ (Galkin) and https://kgtutorial.github.io/ (WSDM 2018 Tutorial) and https://kdd2018tutorialt39.azurewebsites.net/ (KDD 2018 Tutorial)

Credit to Ninh Pham for these slides

Outline

- Introduction of knowledge graph
- How to construct knowledge graph
- Challenges and success stories

Outline

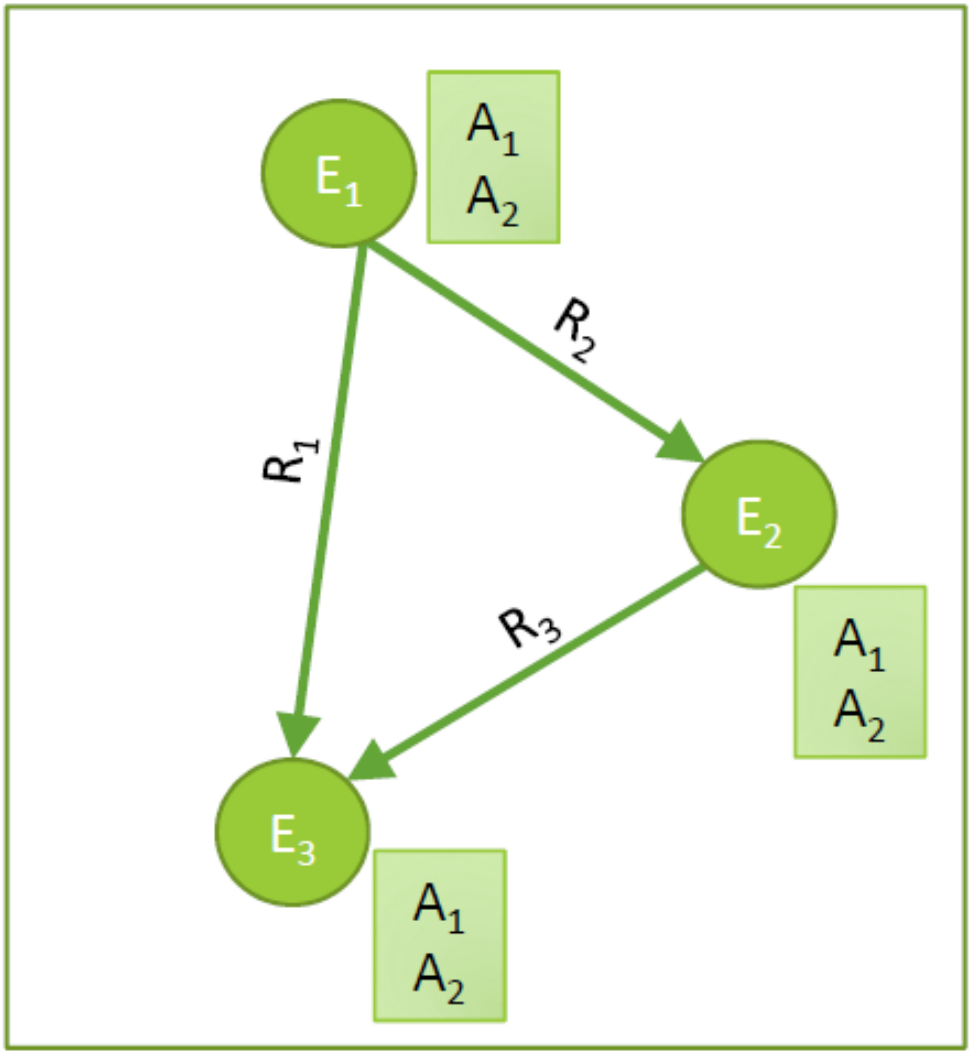
- Introduction of knowledge graph
  - Introduction and applications
  - Viewpoints: Logic, Databases, CV, NLP
- How to construct knowledge graph
- Challenges and success stories

KG: Google's knowledge base

https://www.youtube.com/watch?v=mmQl6VGvX-c&t=75s

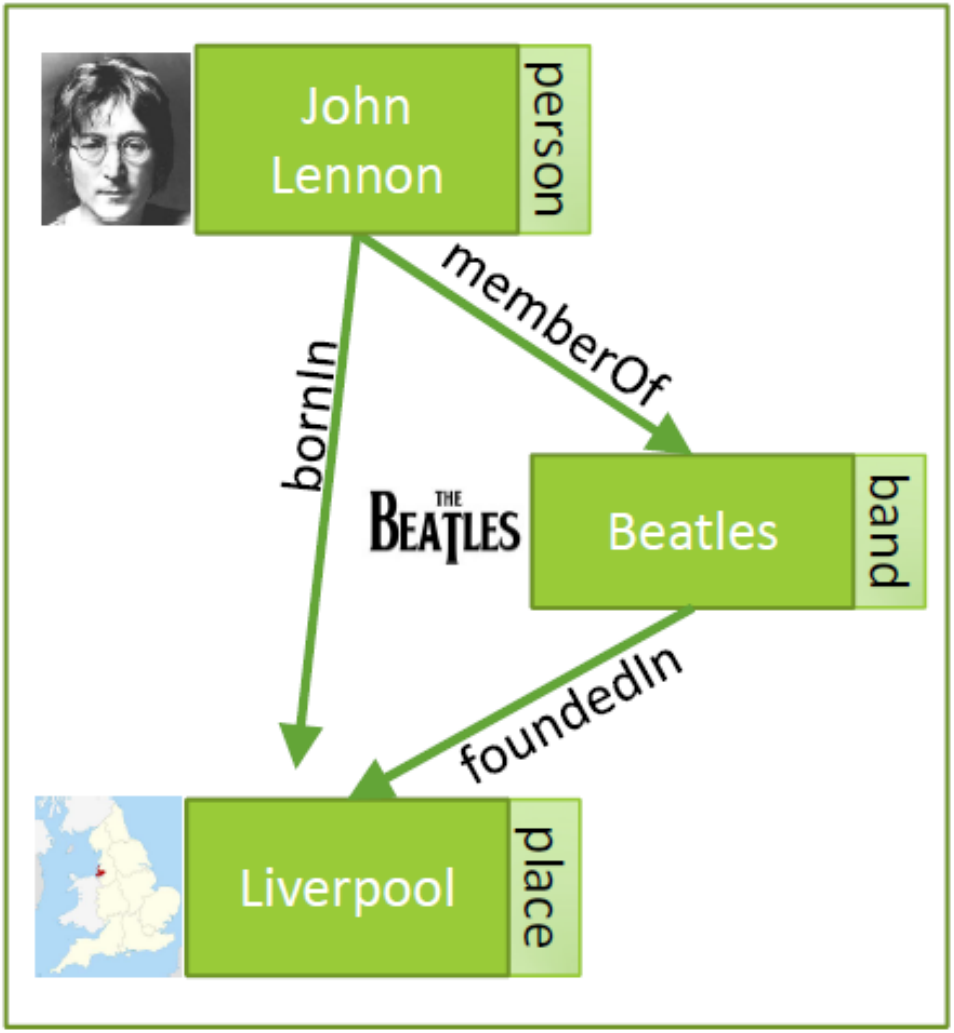
What is KG?

- Knowledge in graph form!
- Capture entities, attributes and relationships
- Nodes are entities
- Nodes are labeled with attributes (e.g. types)
- Edges between 2 nodes capture a relationship between entities



KG example

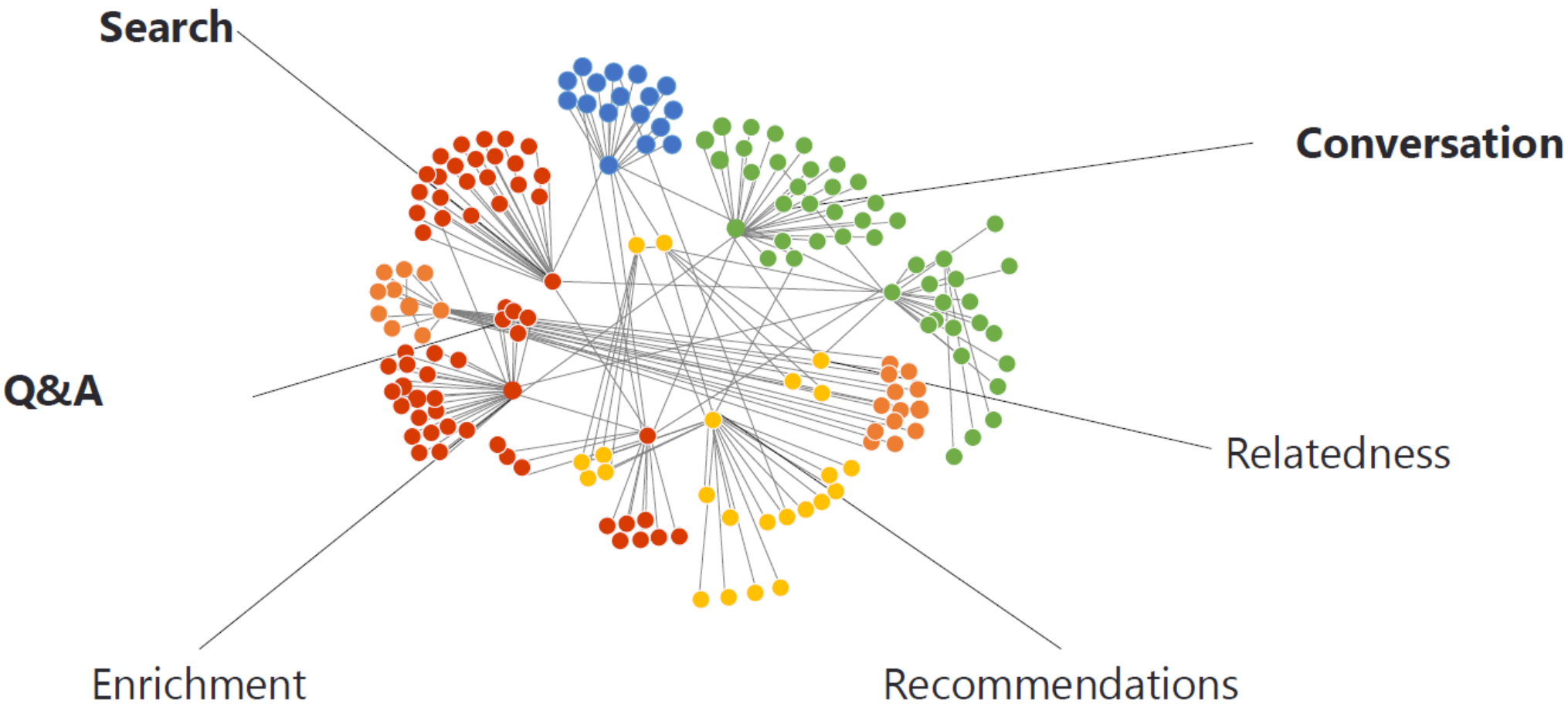
- Knowledge in graph form!
- Capture entities, attributes and relationships
- Nodes are entities
- Nodes are labeled with attributes (e.g. types)
- Edges between 2 nodes capture a relationship between entities



Why KGs?

- Human:
  - Combat information (from big data) overload
  - Explore via intuitive structure
  - Tool for supporting knowledge-driven tasks
- AIs:
  - Key ingredient for many AI tasks
  - Bridge from data to human semantics
  - Use decades of work on graph analysis

Applications



Powering Google search

big bang theory

All | Images | News | Videos | Maps | More | Settings | Tools

About 161,000,000 results (0.85 seconds)

en.wikipedia.org › wiki › The_Big_Bang_Theory

The Big Bang Theory - Wikipedia

The **Big Bang Theory** is an American television sitcom created by Chuck Lorre and Bill Prady, both of whom served as executive producers on the series, along with Steven Molaro. All three also served as head writers.

Related shows: Young Sheldon **Producer(s)**: Faye Oshima Belyeu
No. of seasons: 12 **Original release**: September 24, 2007 –; May 1...

List of The Big Bang Theory ... · Sheldon Cooper · Category:The Big Bang Theory

www.imdb.com › title

The Big Bang Theory (TV Series 2007–2019) - IMDb

★★★★☆ Rating: 8.1/10 - 684,017 votes

Created by Chuck Lorre, Bill Prady. With Johnny Galecki, Jim Parsons, Kaley Cuoco, Simon Helberg. A woman who moves into an apartment across the hall ...

People also ask

- What is Big Bang theory of universe?
- Why did Big Bang Theory end?
- Is Big Bang Theory on a break?
- Is Big Bang Theory on Netflix?

Feedback

The Big Bang Theory

2007 · Sitcom · 12 seasons

8.1/10	82%	8.9/10
IMDb	Rotten Tomatoes	TV.com

94% liked this TV show

Google users

The lives of four socially awkward friends, Leonard, Sheldon, Howard and Raj, take a wild turn when they meet the beautiful and free-spirited Penny.

First episode date: 24 September 2007

Final episode date: 16 May 2019

No. of episodes: 279 (list of episodes)

Knowledge-powered Q&A

who is vice chancellor of the university of auckland

All Images News Videos Shopping Web Forums More Tools

AI Overview

Learn more

The current Vice-Chancellor of the University of Auckland is Professor Dawn Freshwater.

Role:
She is the university's chief academic and administrative officer, and the employer of all staff.

Responsibilities:
She is responsible for academic leadership, effective management, strategic planning, and resource allocation.

Office Location:

Show more

Management team - The University of Auckland
The University of Auckland

Vice-Chancellor - The University of Auckland
Professor Dawn Freshwater, BA(Hons) Manc., PhD Nott.
The University of Auckland

Law school stoush: Auckland Uni vice-chancellor faces backlash

Professor Dawn Freshwater

Professor Dawn Freshwater, BA(Hons) Manc., PhD Nott.

The University of Auckland
https://www.auckland.ac.nz › about-us › the-university

Vice-Chancellor - The University of Auckland

About featured snippets • Feedback

not static.

change overtime.

Knowledge-powered conversation

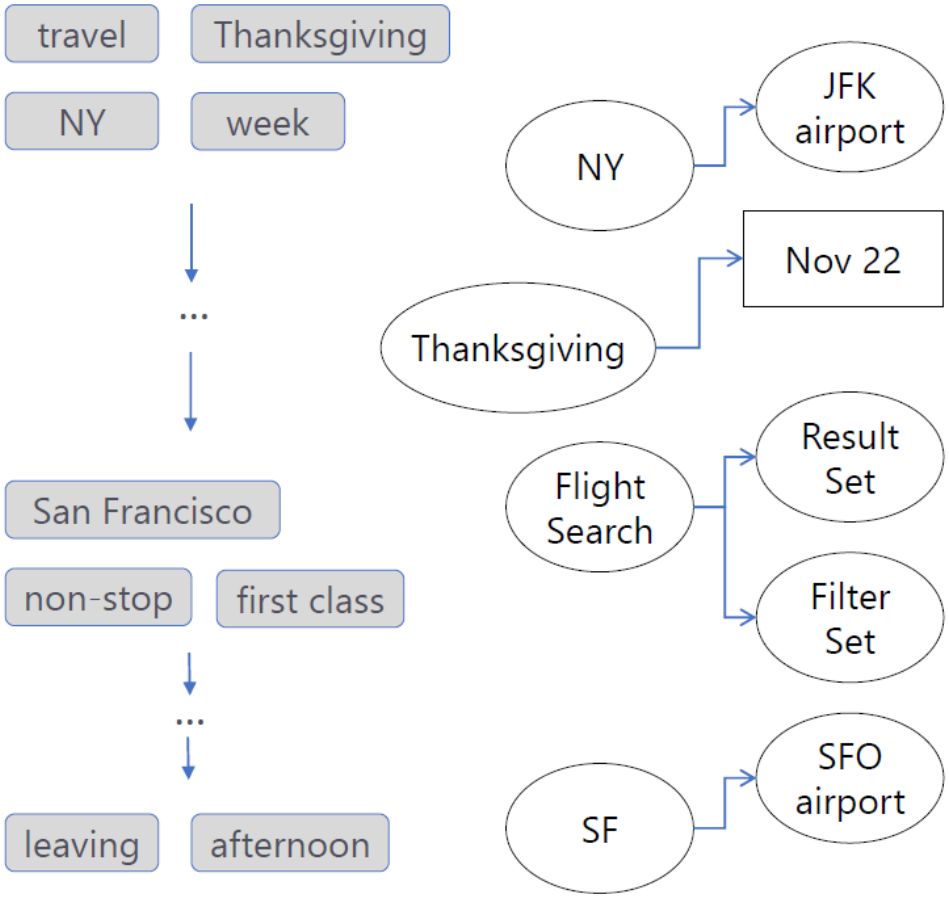
I want to travel to NY 2 days before Thanksgiving, staying for a week

Okay, booking a flight to JFK from November 20 to November 27. Where will you be flying from?

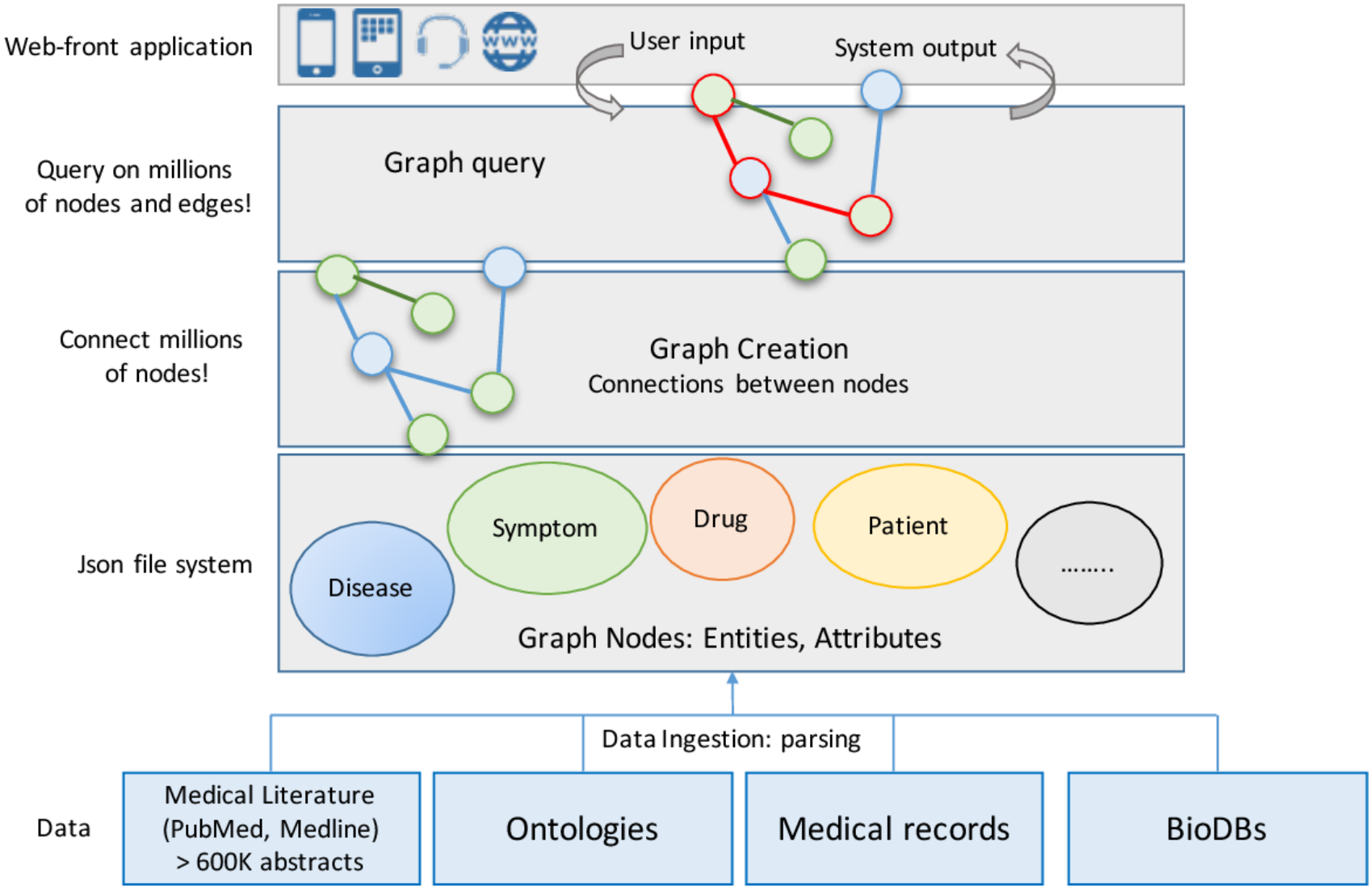
From San Francisco, and also non-stop in first class

Got it, I've found some flights for you ...

How about leaving in the afternoon



Decision support

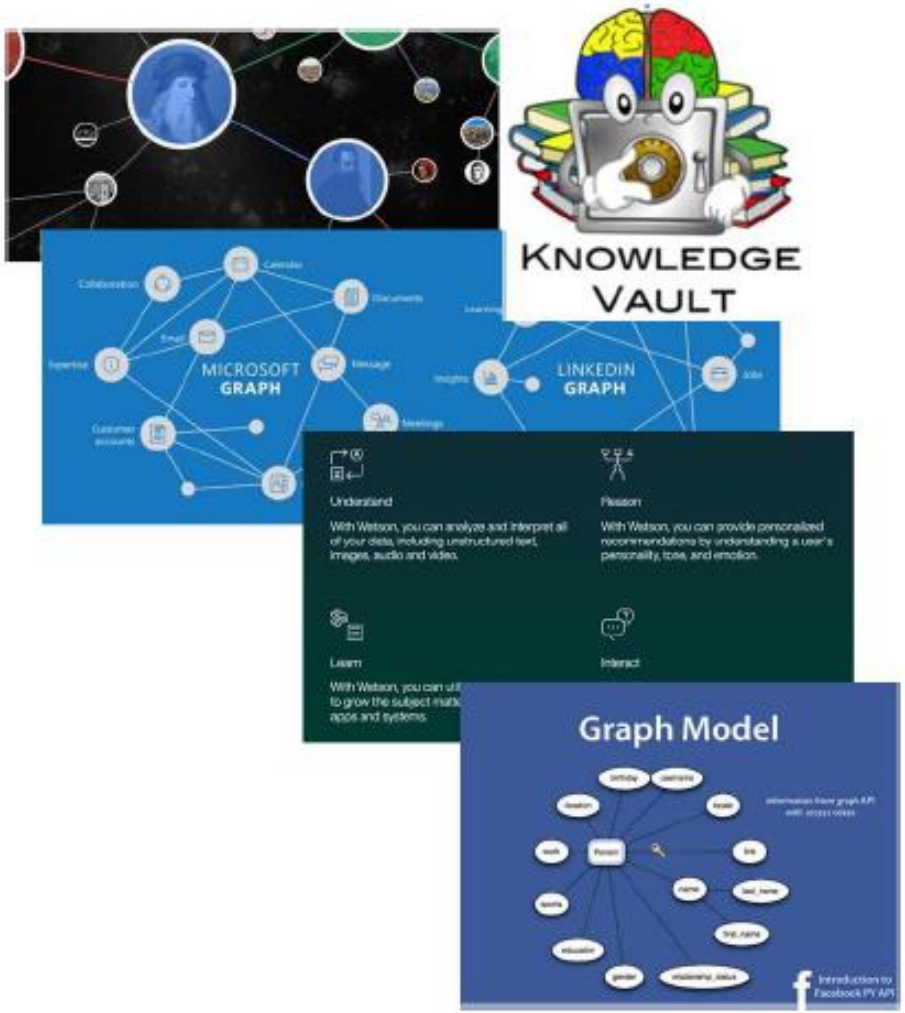


most work

Source: Cognitive DDx Assistant in Rare Diseases (Reumann et al. 2018)

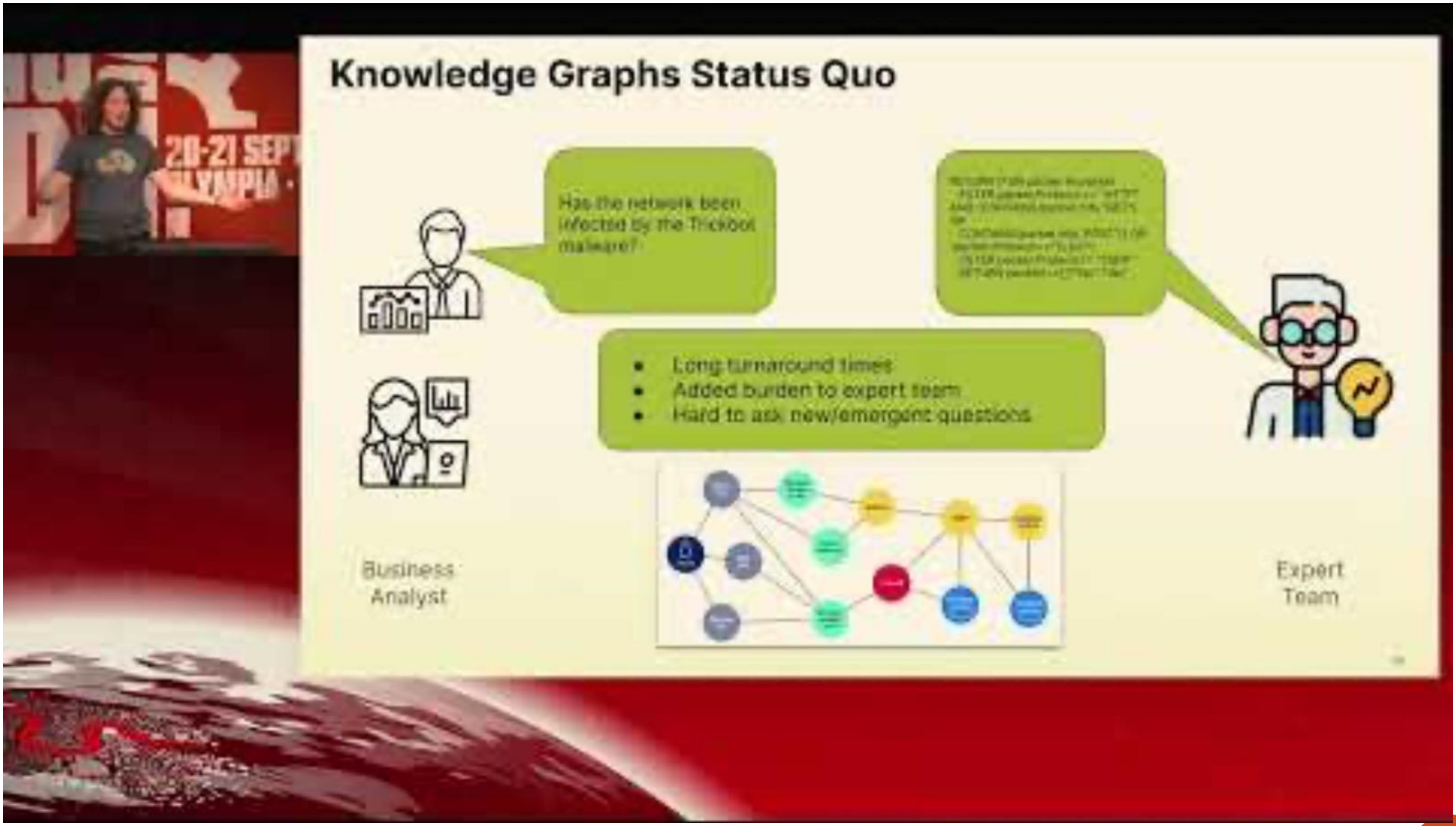
KG and industry

- Google KG
- Amazon Product Graph
- Facebook Graph API
- IBM Watson
- Microsoft Satori
- LinkedIn KG
- Yandex Object Answer
- …



大语言模型

KGs meet LLMs



https://www.youtube.com/watch?v=qCT1a2jdpk4&ab_channel=BigDataLDN

Where does knowledge come from?

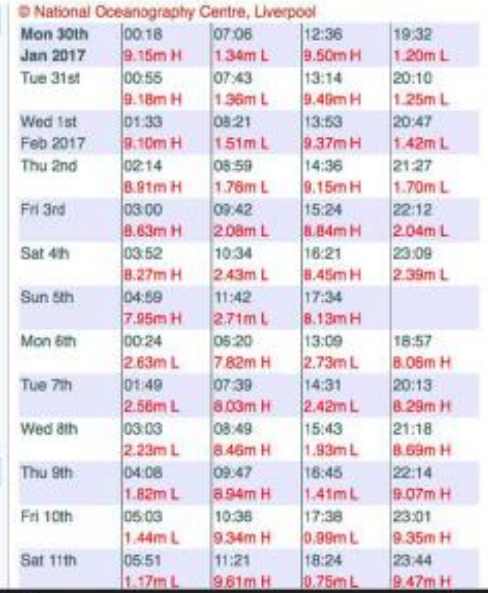
- Structured text:
  - Wikipedia, tables, databases

The Beatles

The "Fab Four" Beatles lineup in 1964. Clockwise from top left: John Lennon, Paul McCartney, Ringo Starr and George Harrison

Background information

Origin	Liverpool, England, United Kingdom
Genres	Rock · pop
Years active	1960–1970
Labels	EMI · Polyc Swan · Vee United Artis
Associated acts	The Quarry Preston · P
Website	thebeatles.
Past members	John Lenn Paul McCa George Ha Ringo Star



© National Oceanography Centre, Liverpool

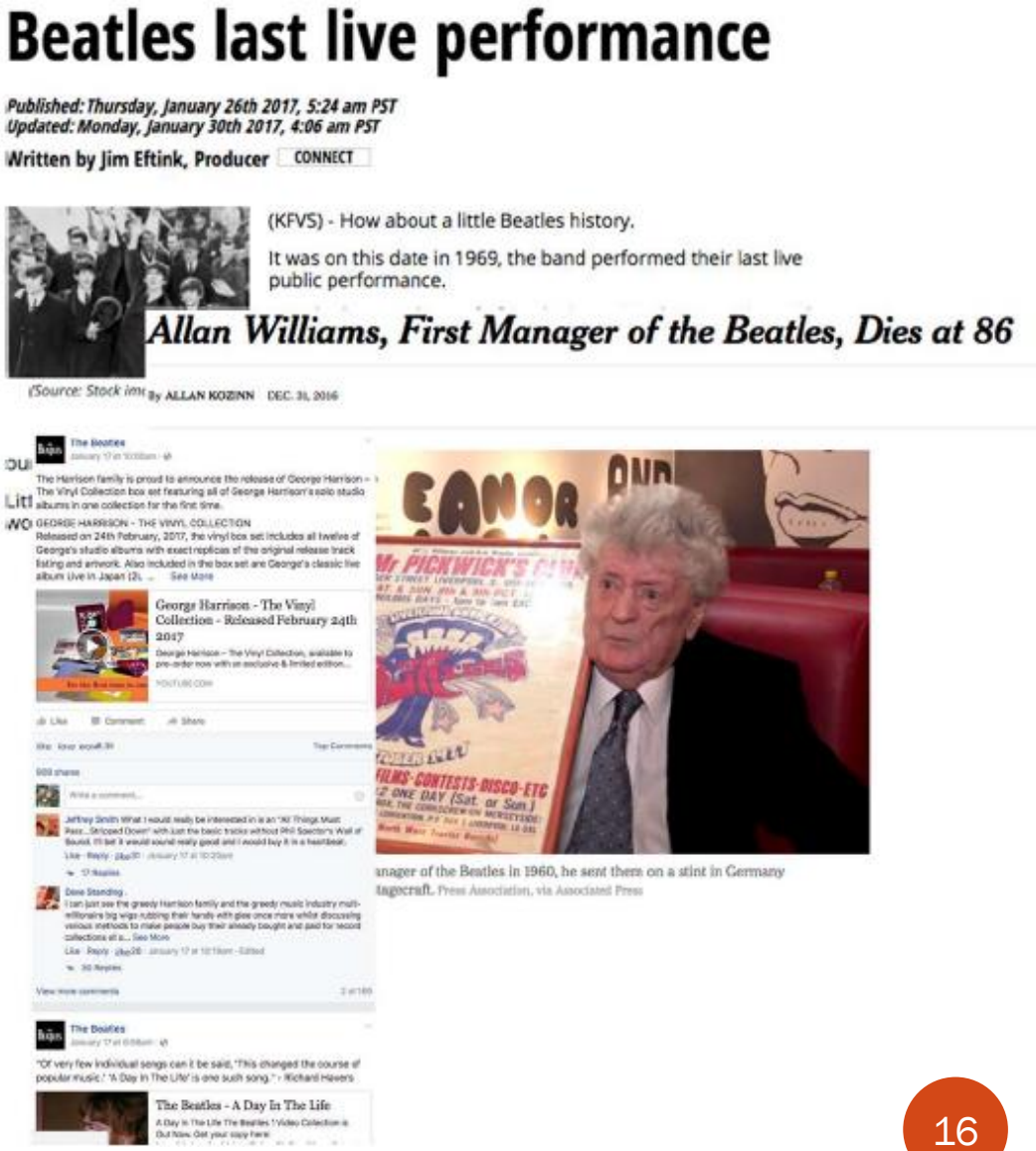
Mon 30th Jan 2017	00:18 9.15m H	07:06 1.34m L	12:36 9.50m H	19:32 1.20m L
Tue 31st	00:55 9.18m H	07:43 1.36m L	13:14 9.49m H	20:10 1.25m L
Wed 1st Feb 2017	01:33 9.10m H	08:21 1.51m L	13:53 9.37m H	20:47 1.42m L
Thu 2nd	02:14 8.91m H	08:59 1.76m L	14:36 9.15m H	21:27 1.70m L
Fri 3rd	03:00 8.63m H	09:42 2.08m L	15:24 8.84m H	22:12 2.04m L
Sat 4th	03:52 8.27m H	10:34 2.43m L	16:21 8.45m H	23:09 2.39m L
Sun 5th	04:59 7.95m H	11:42 2.71m L	17:34 8.13m H	
Mon 6th	00:24 2.63m L	06:20 7.82m H	13:09 2.73m L	18:57 8.06m H
Tue 7th	01:49 2.56m L	07:39 8.03m H	14:31 2.42m L	20:13 8.29m H
Wed 8th	03:03 2.23m L	08:49 8.46m H	15:43 1.93m L	21:18 8.69m H
Thu 9th	04:08 1.82m L	09:47 8.94m H	16:45 1.41m L	22:14 9.07m H
Fri 10th	05:03 1.44m L	10:38 9.34m H	17:38 0.99m L	23:01 9.35m H
Sat 11th	05:51 1.17m L	11:21 9.61m H	18:24 0.75m L	23:44 9.47m H

The Beatles Total Album Sales Statistics	Data
Total number of Beatles albums sold	2,303,100,000
Total Albums Sold on iTunes	765,000
Total Singles Sold on iTunes	3,600,000
Sales By Available Markets	
United States	209.1 Million
Canada	13.6 Million
United Kingdom	7.6 Million
Germany	7.3 Million
France	3.1 Million
Australia	2.8 Million
Japan	1.9 Million
Argentina	1.6 Million
Brazil	606,000
Sweden	564,000
Austria	570,000
Switzerland	450,000
Beatles Billboard Chart Statistics	
Total weeks on chart	1,278 weeks
Total number ones	15
Total weeks at number one	175 weeks
Album with longest time spent at number one ("Please Please Me")	30 weeks

PREV DATA SET NEXT DATA SET

Where does knowledge come from?

- Structured text:
  - Wikipedia, tables, databases
- Unstructured text:
  - WWW, news, social media



Beatles last live performance

Published: Thursday, January 26th 2017, 5:24 am PST
Updated: Monday, January 30th 2017, 4:06 am PST

Written by Jim Eftink, Producer CONNECT

(KFVS) - How about a little Beatles history.

It was on this date in 1969, the band performed their last live public performance.

Allan Williams, First Manager of the Beatles, Dies at 86

(Source: Stock im By ALLAN KOZINN DEC. 31, 2016

The Beatles

The Harrison family is proud to announce the release of George Harrison - The Vinyl Collection box set featuring all of George Harrison's solo studio albums in one collection for the first time.

GEORGE HARRISON - THE VINYL COLLECTION
Released on 24th February, 2017, the vinyl box set includes all twelve of George's studio albums with exact replicas of the original release track listing and artwork. Also included in the box set are George's classic live album Live in Japan (2). ... See More

George Harrison - The Vinyl Collection - Released February 24th 2017

YOUTUBE.COM

Like Comment Share

Top Comments

Write a comment...

Jeffrey Smith What I would really be interested in is an "All Things Must Pass...Stripped Down" with just the basic tracks without Phil Spector's Wall of Sound. I'll bet it would sound really good and I would buy it in a heartbeat.
Like · Reply · January 17 at 10:20am

Dave Standing
I can just see the greedy Harrison family and the greedy music industry multi-millionaire big wigs rubbing their hands with glee once more whilst discussing various methods to make people buy their already bought and paid for record collections at a... See More
Like · Reply · January 17 at 10:15am · Edited

View more comments

The Beatles

"Of very few individual songs can it be said, 'This changed the course of popular music.' 'A Day In The Life' is one such song." - Richard Havers

The Beatles - A Day In The Life

A Day In The Life The Beatles 1 Video Collection is Out Now. Get your copy here!

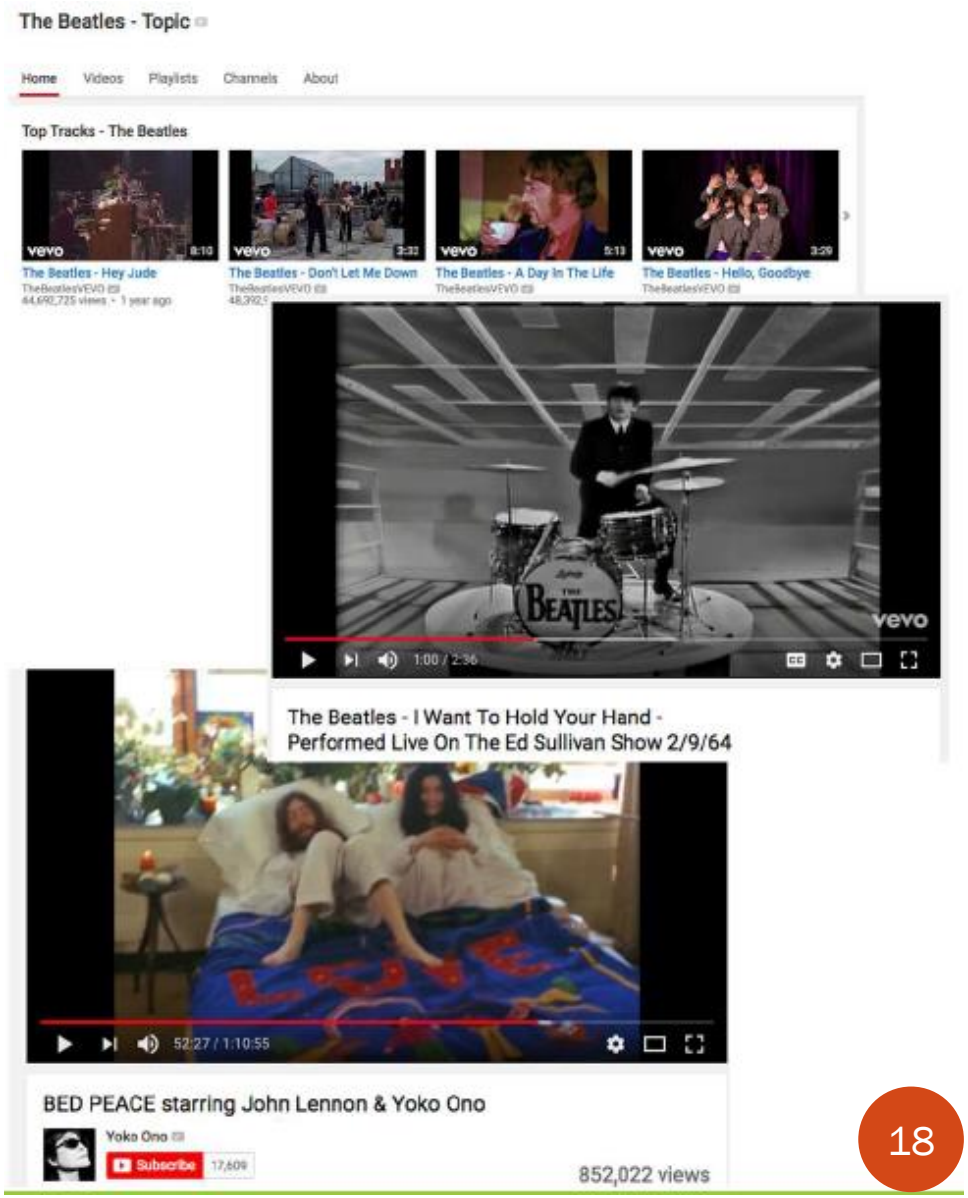
anager of the Beatles in 1960, he sent them on a stint in Germany
tagecraft. Press Association, via Associated Press

Where does knowledge come from?

- Structured text:
  - Wikipedia, tables, databases
- Unstructured text:
  - WWW, news, social media
- Images

Where does knowledge come from?

- Structured text:
  - Wikipedia, tables, databases
- Unstructured text:
  - WWW, news, social media
- Images
- Video:
  - Youtube, video feeds



Outline

- Introduction of knowledge graph
  - Introduction and applications
  - Viewpoints: Logic, Databases, CV, NLP

computer vision.

- How to construct knowledge graph

- Application of knowledge graph

Knowledge representation

Given entities E, relations R, KG is a directed multi-relational graph G that comprises triples (s, p, o)

$$\mathcal{G} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$$

$$(s, p, o) \in \mathcal{G}$$

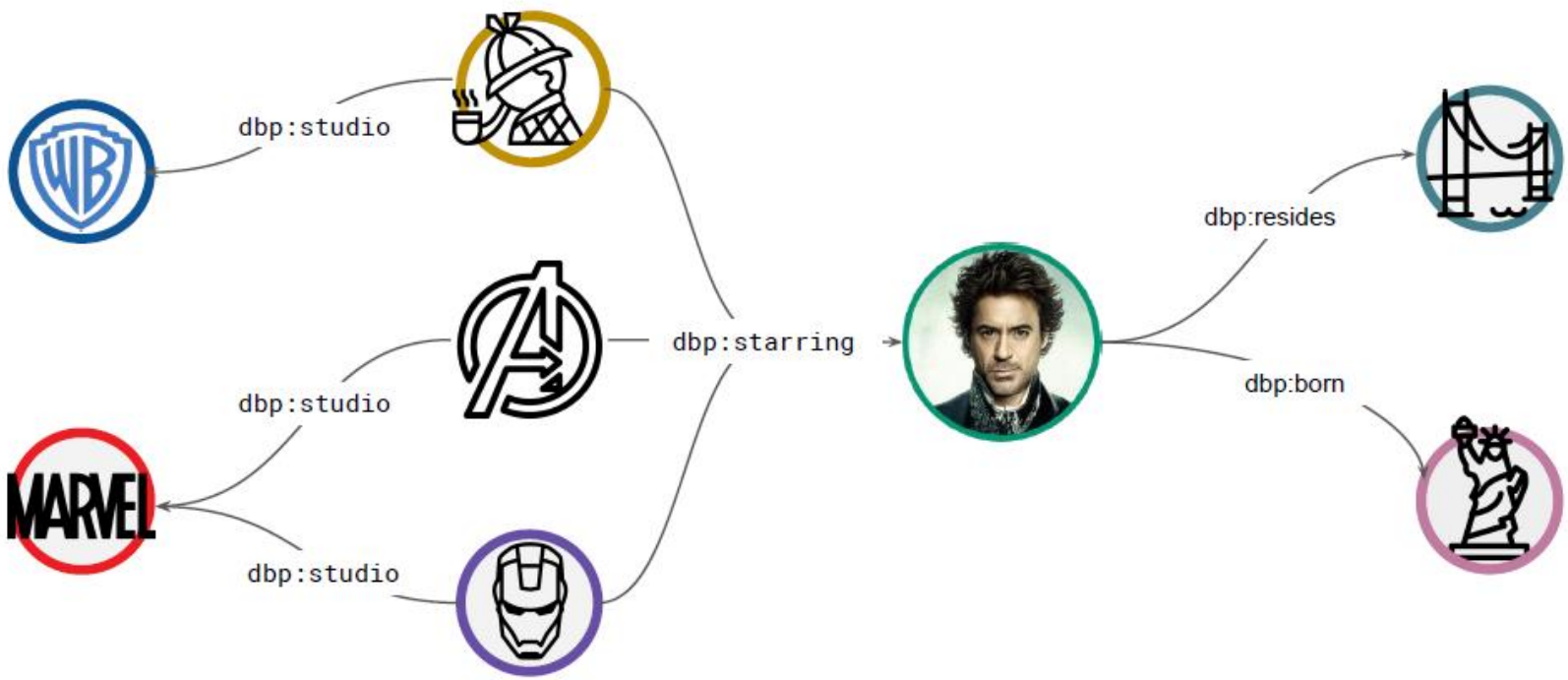
"Schema (TBox) with instances (ABox)"

* describes entities and relations
* defines a schema
* interrelating arbitrary entities
* various topical domains

"Every RDF / LPG / RDF* graph is a knowledge graph"

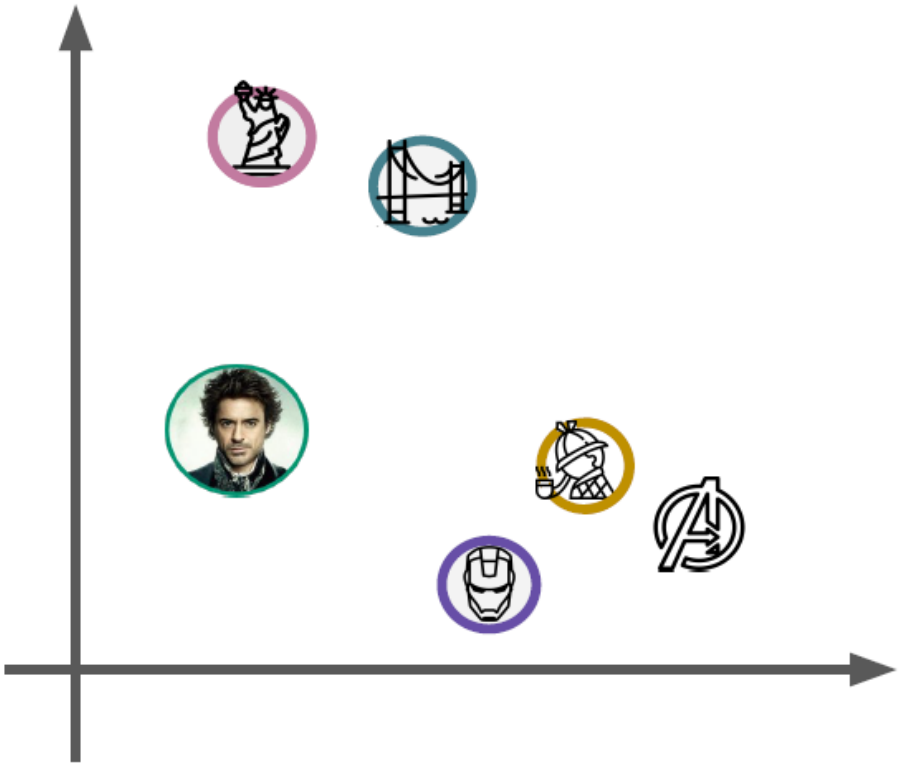
Graph-structured world model

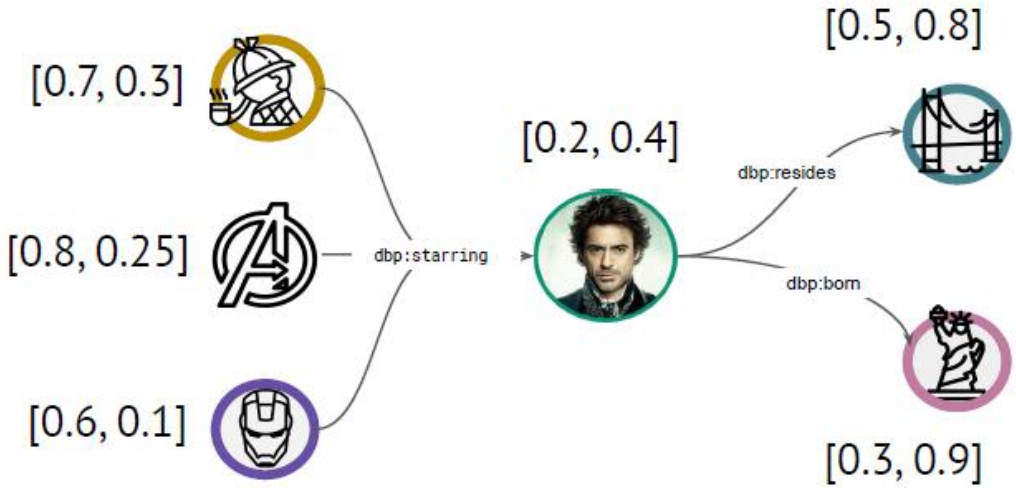
Symbolic: Triples



RDJ	dbp:resides	SF .	Avengers	dbp:studio	Marvel .
RDJ	dbp:born	NY .	Avengers	dbp:starring	RDJ .
Sherlock_Holmes	dbp:studio	WB .	Iron_Man	dbp:studio	Marvel .
Sherlock_Holmes	dbp:starring	RDJ.	Iron_Man	dbp:starring	RDJ .

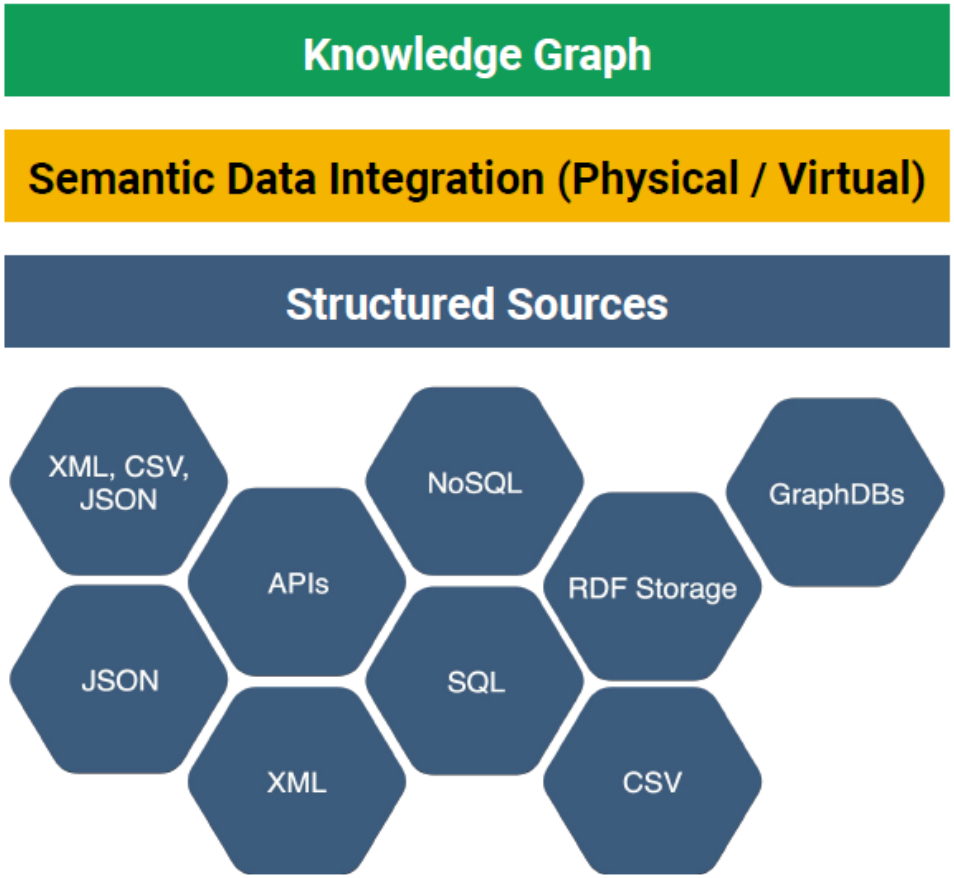
Vector: Embeddings





http://rdf2vec.org/

View: Database – Data integration



Entities

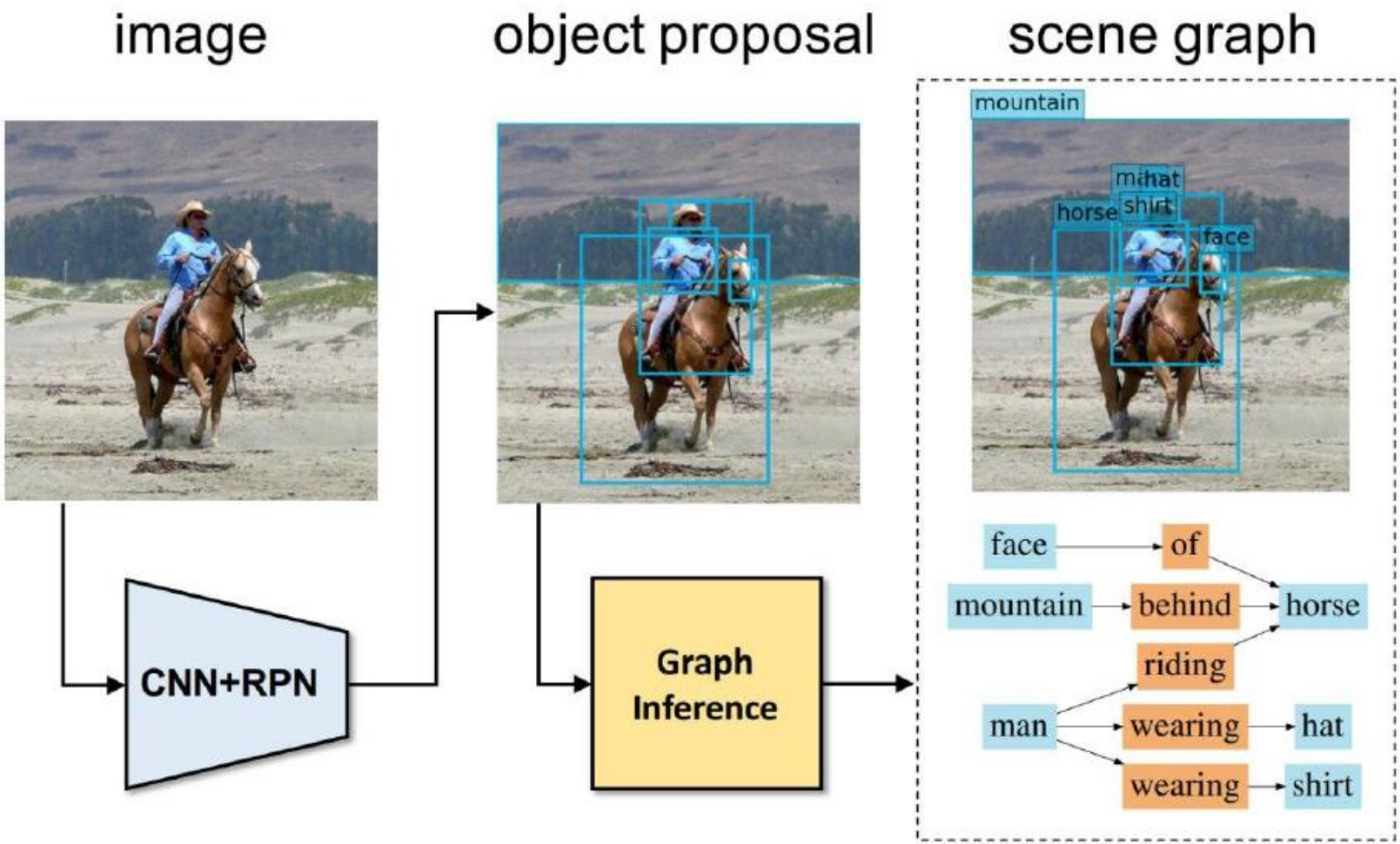
Cells - classes, instances

Relations

Column headers

Author	Notable work	Date of birth
Stan Lee	Iron Man	12/28/1922
Bob Kane	Batman	10/24/1915

View: Computer vision



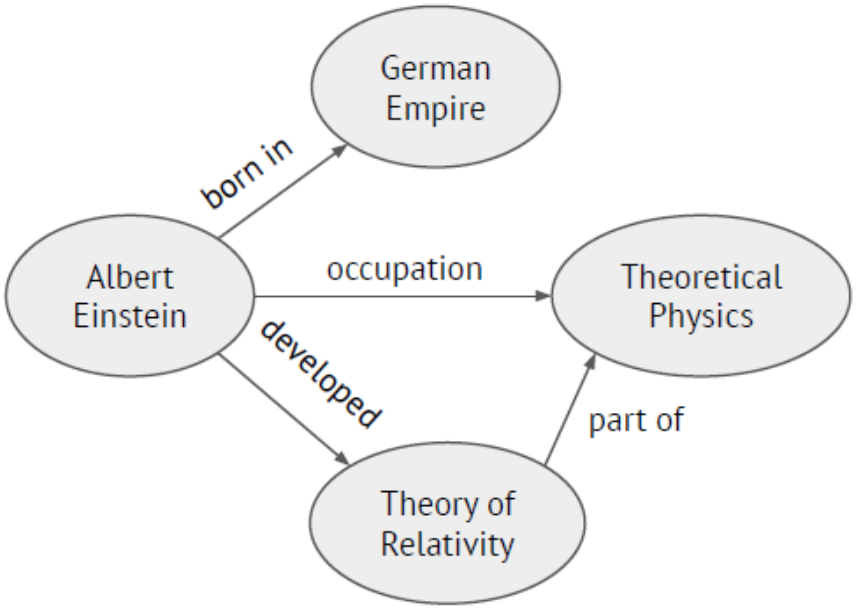
Entities: Identified objects

Relations: Inferred as to objects arrangement

Xu et al. Scene Graph Generation by Iterative Message Passing. CVPR 2017

View: NLP – Build KGs from texts

Albert Einstein was a German-born theoretical physicist who developed the theory of relativity.



Knowledge Graph

Information Retrieval

Unstructured Sources

View: NLP – Named entity recognition

Entities in a Knowledge Graph

apple (Q89)

fruit of the apple tree

apples

Apple (Q1754545)

1990 album by Mother Love bone

Apple (Q213710)

UK international record label; imprint of Apple Corps Ltd.

LC 01074 | LC 1074 | Apple Records

Apple Inc. (Q312)

American producer of hardware, software, and services, based in Cupertino, California

Apple Computer, Inc. | Apple Computer | Apple Computer Inc | Apple | Apple Incorporated | Apple Computer Incorporated |

Who is the CEO of **Apple**?

Apple belongs to which genus?

movie character

Downey played **Iron Man** in which year?

Who is the alter ego of **Iron man**?

comic character

View: NLP – Relation linking

ask some thing

Relations in a Knowledge Graph

- List of known relations
- Surface forms (synonyms), easily multi-lingual
- Relations constraints
- Relations hierarchy
- Most used types of subjects and objects

wdt:P161

Name all the movies in which Robert Downey Jr **acted**?

Find me all the films **casting** Robert Downey Jr ?

List all the movies **starring** Robert Downey Junior?

RDJ **has acted** in which movies?

cast member (P161)

actor in the subject production |

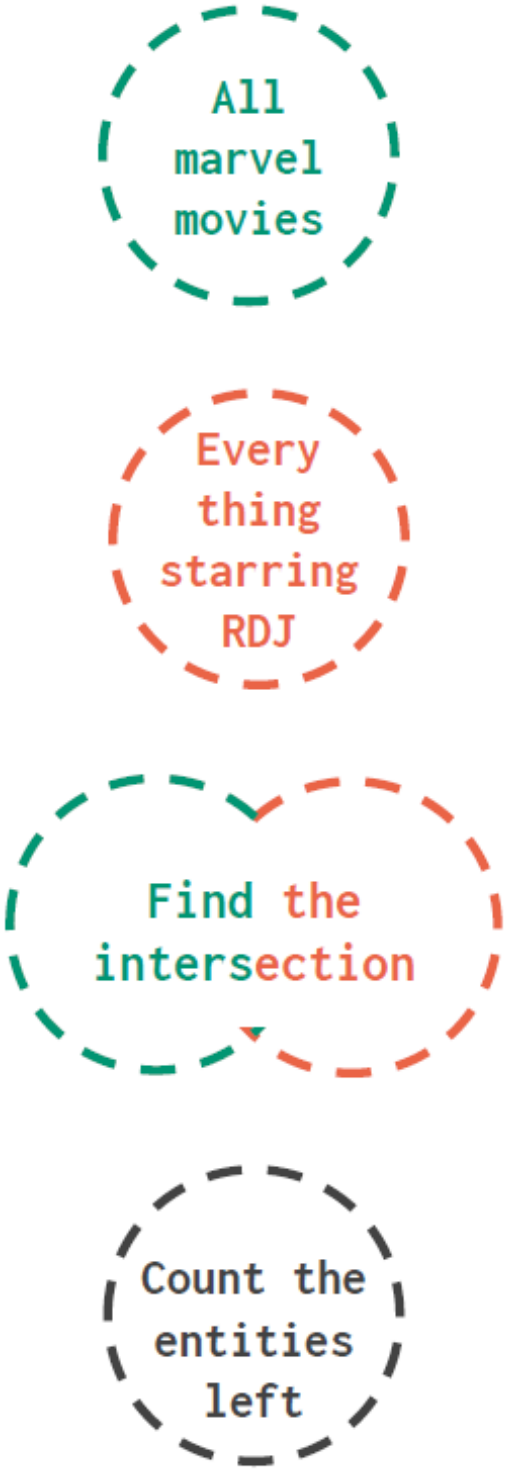
starring | film starring | actor | actress | contestant or a play

performer (P175)

actor, musician, band or other performer associated with this role or musical work

artist | musician | played by | portrayed by | recorded by | recording by | dancer | actor | musical artist

View: NLP – Q&A

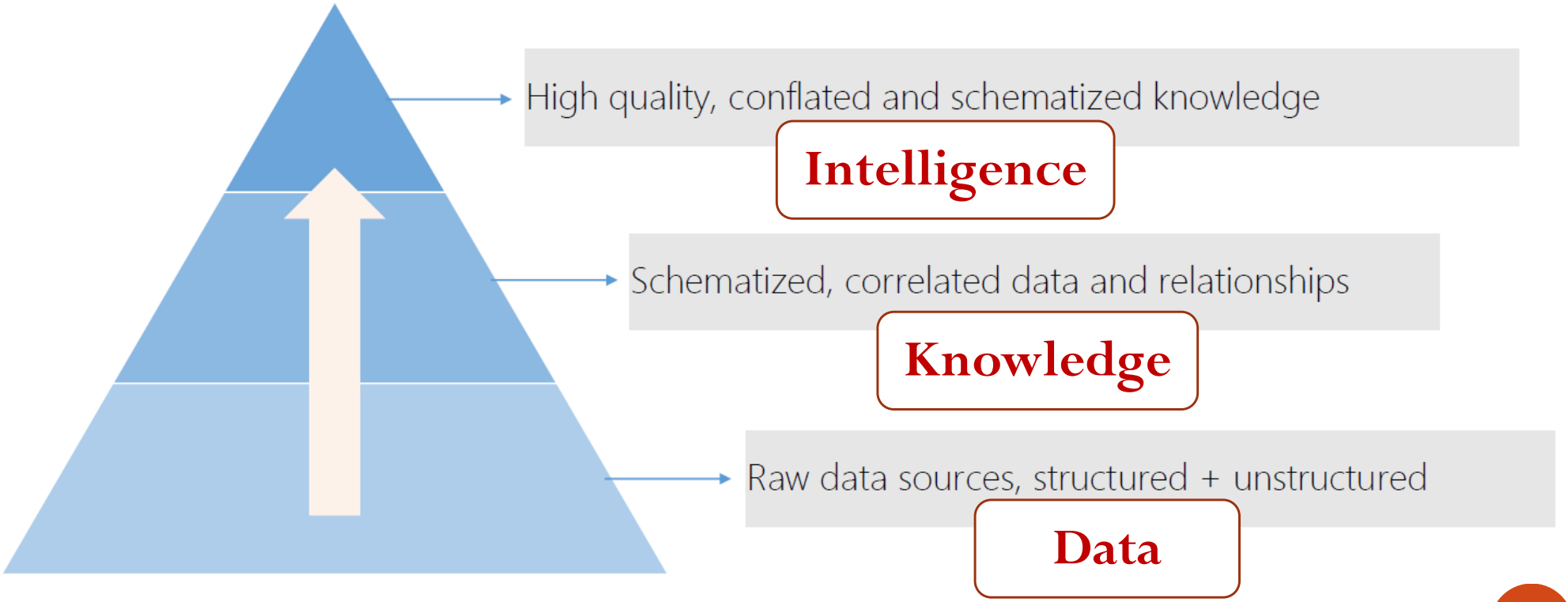


How many Marvel movies was Robert Downey Jr. casted in?

```
SELECT COUNT(?uri) WHERE {
    ?uri dbp:studio dbr:Marvel_Studios.
    ?uri dbo:starring dbr:Robert_Downey_Jr
}
```

Create high-quality KGs

- Require machine learning, natural language processing, conflation and inference

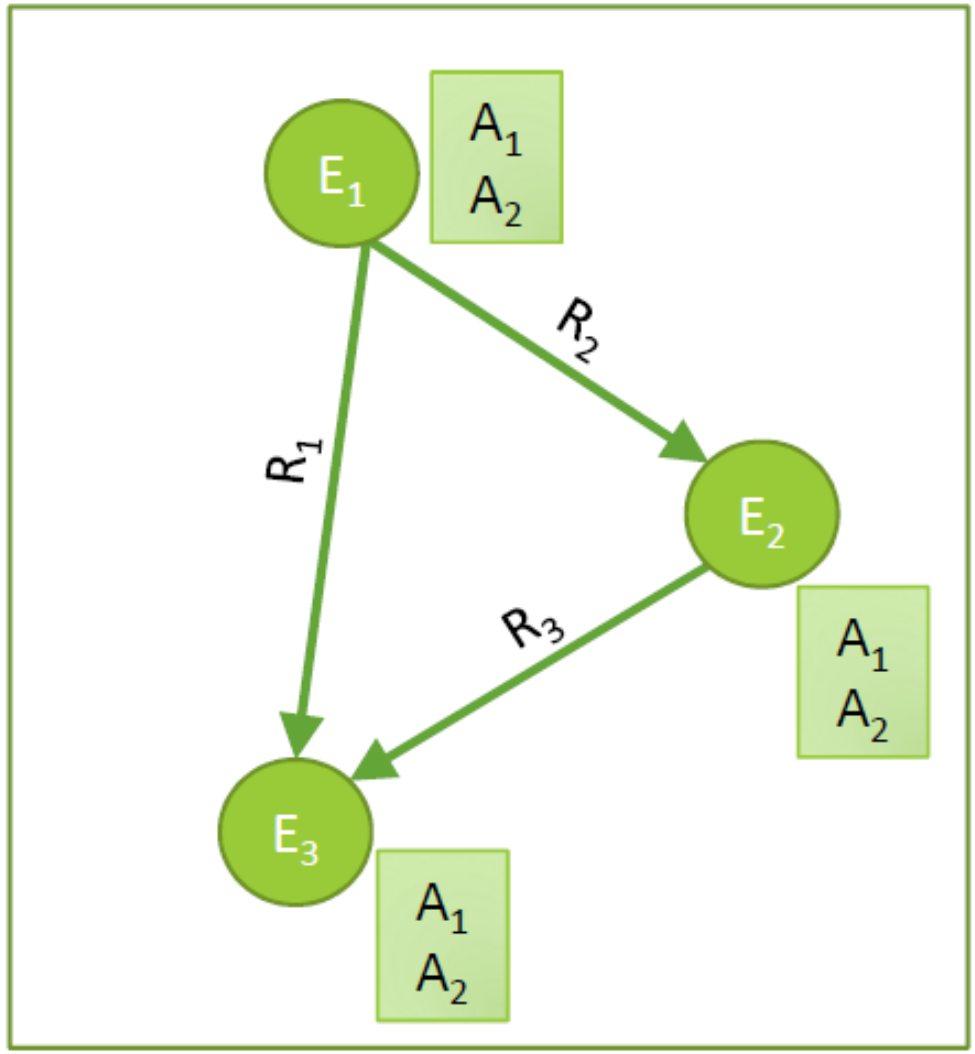


Outline

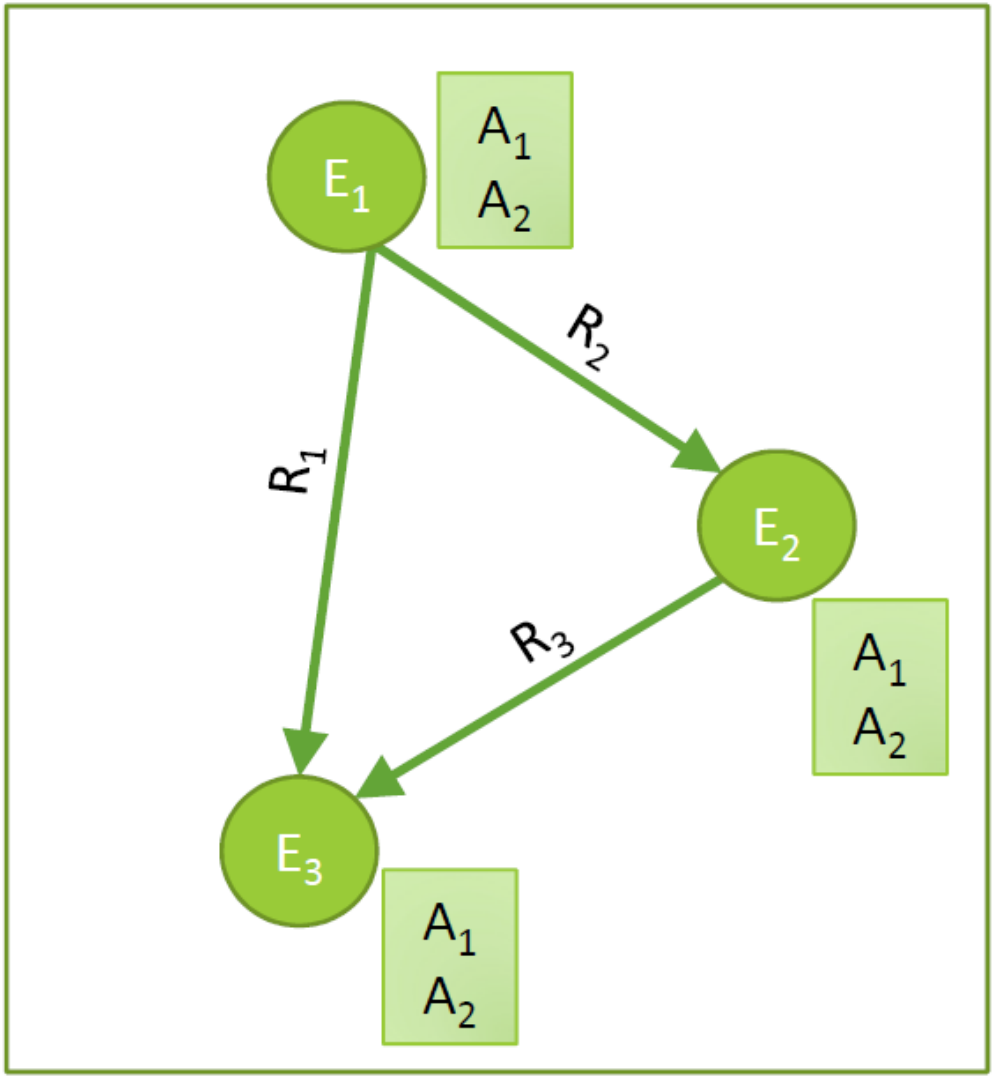
- Introduction of knowledge graph
- How to construct knowledge graph
  - Extract information (from text)
  - Match/merge entities
  - Knowledge refinement
- Challenges and success stories

What is KG?

- Knowledge in graph form!
- Capture entities, attributes and relationships
- Nodes are entities
- Nodes are labeled with attributes (e.g. types)
- Edges between 2 nodes capture a relationship between entities

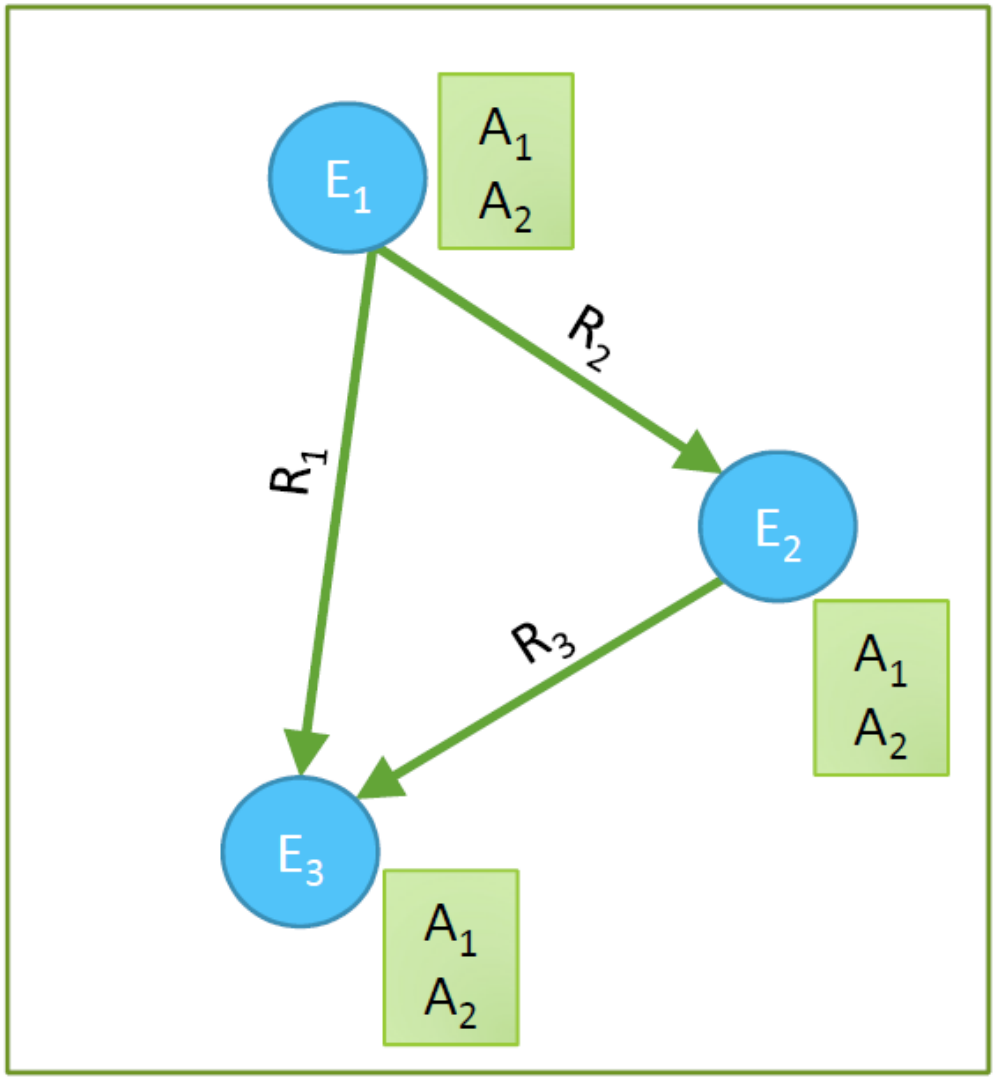


Basic problems



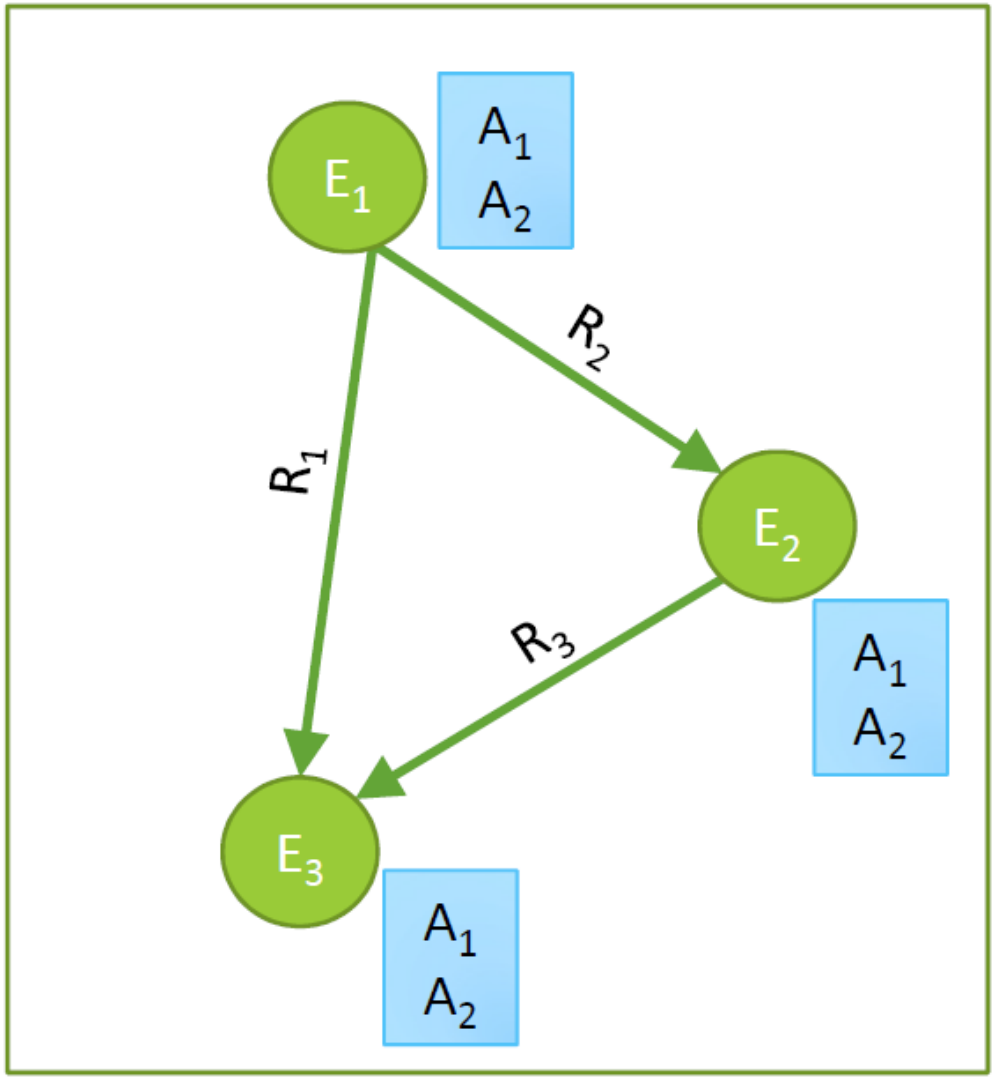
Basic problems

- Who are the entities (nodes) in graph?



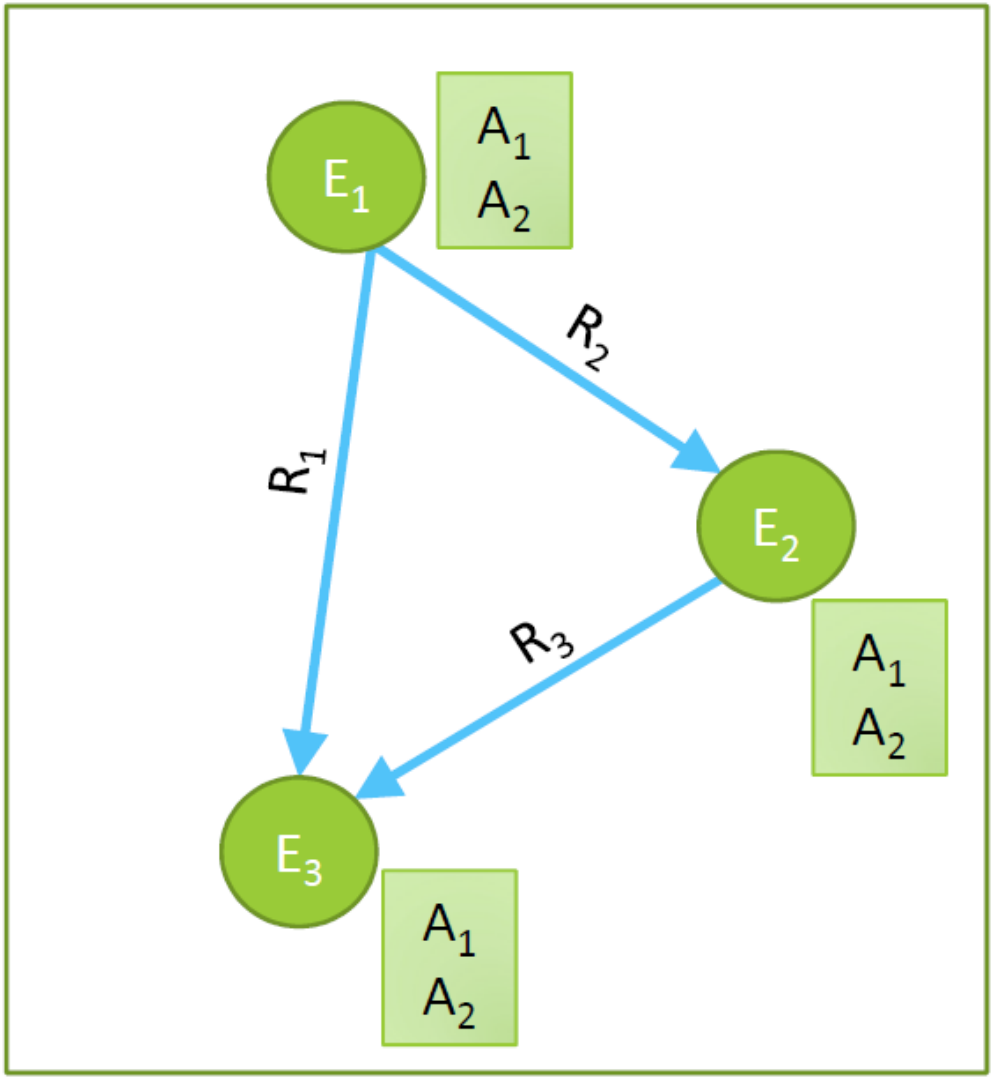
Basic problems

- Who are the entities (nodes) in graph?
- What are their attributes and types (labels)?



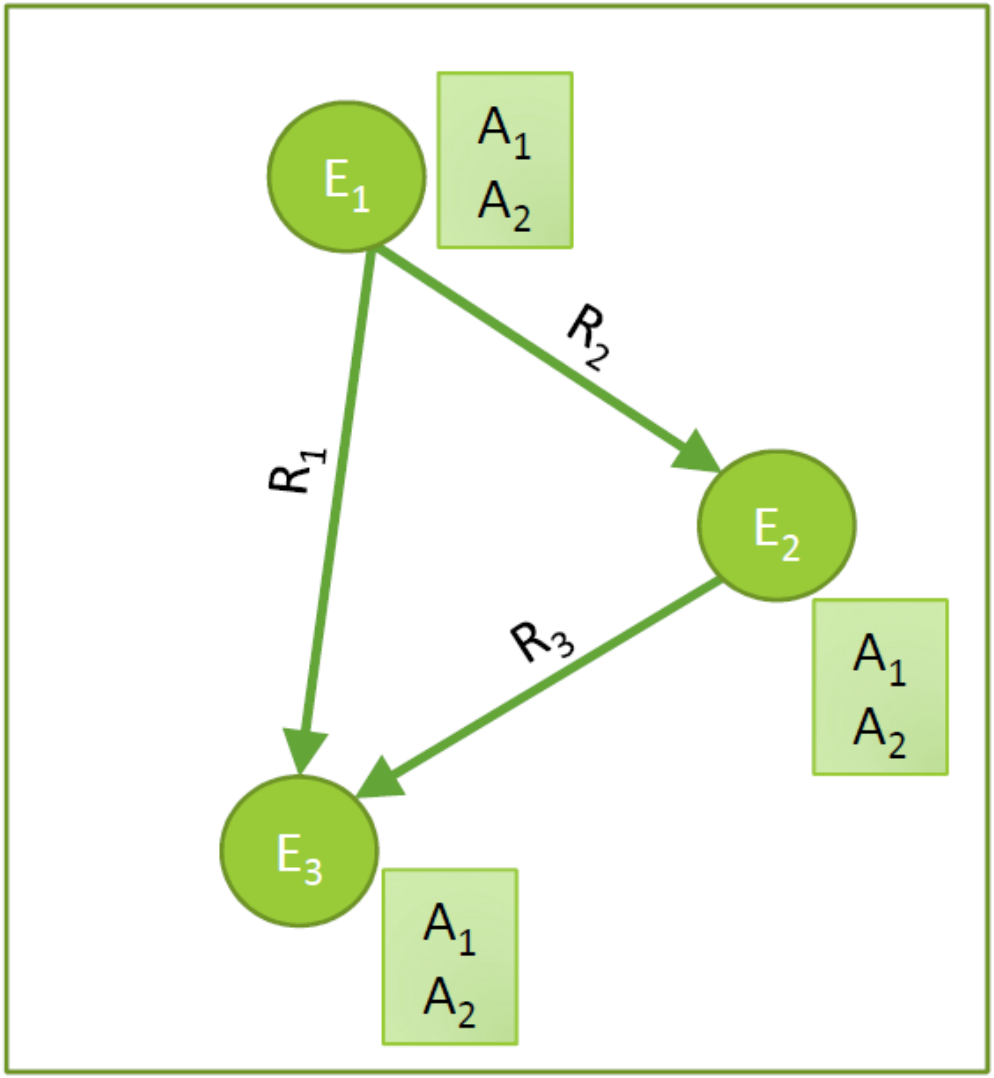
Basic problems

- Who are the entities (nodes) in graph?
- What are their attributes and types (labels)?
- How are they related (edges)?



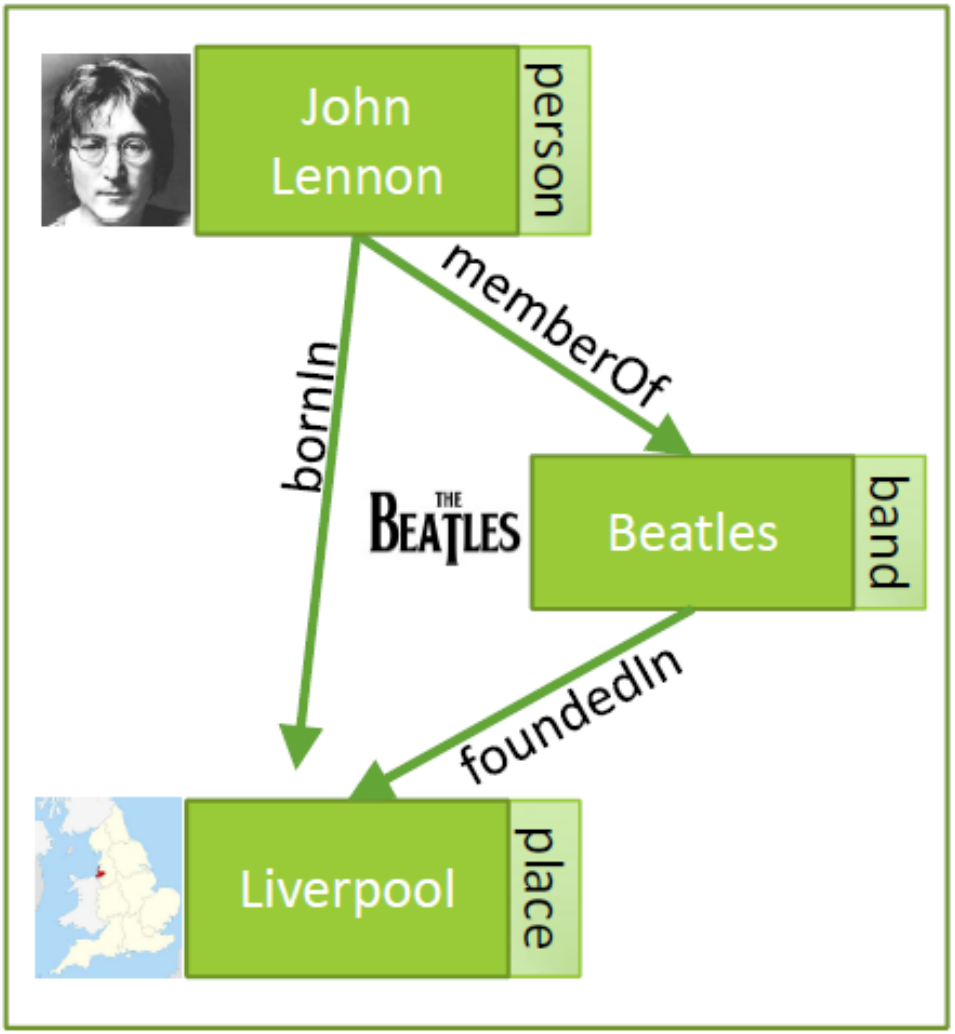
Basic problems

- Who are the entities (nodes) in graph?
- What are their attributes and types (labels)?
- How are they related (edges)?



How difficult it is

- Easiest: Only Who
- Easy: Who + How
- Difficult: Who + How + What
- Very difficult: Who + How + What + Constraints



Construct KG

most work

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

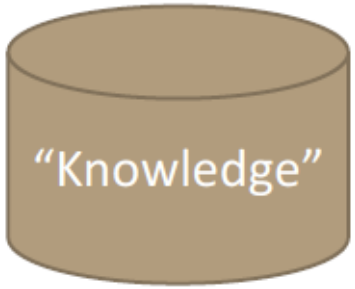
- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve

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

Knowledge extraction

Knowledge extraction from text



Unstructured
Ambiguous
Lots and lots of it!

Structured
Precise, Actionable
Specific to the task

Humans can read them, but
… very slowly
… can't remember all
… can't answer questions

Can be used for downstream applications, such as creating Knowledge Graphs!

Knowledge extraction using NLP

John was born in Liverpool, to Julia and Alfred Lennon.

Text

NLP

Lennon..
John Lennon...

the Pool

Mrs. Lennon..
.. his mother ..

his father
he
Alfred

Person | Location | Person | Person

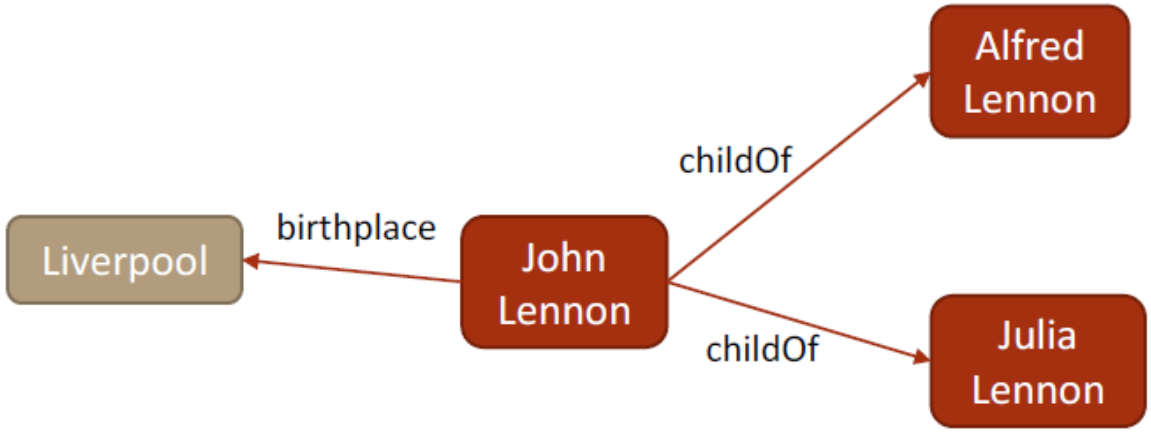
John was born in Liverpool, to Julia and Alfred Lennon.

NNP | VBD | VBD | IN | NNP | TO | NNP | CC | NNP | NNP

Annotated text

Extraction graph

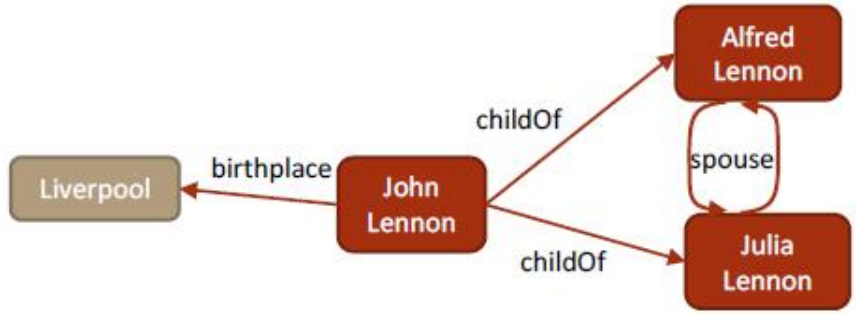
Information Extraction



Breaking it down

Information Extraction

Entity resolution,
Entity linking,
Relation extraction…



Document

Coreference Resolution…

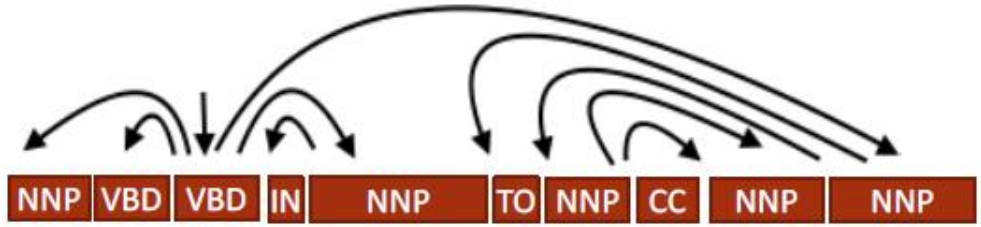
Lennon..
John Lennon…
the Pool
Mrs. Lennon..
.. his mother ..
his father
he
Alfred

Person Location Person Person

John was born in Liverpool, to Julia and Alfred Lennon.

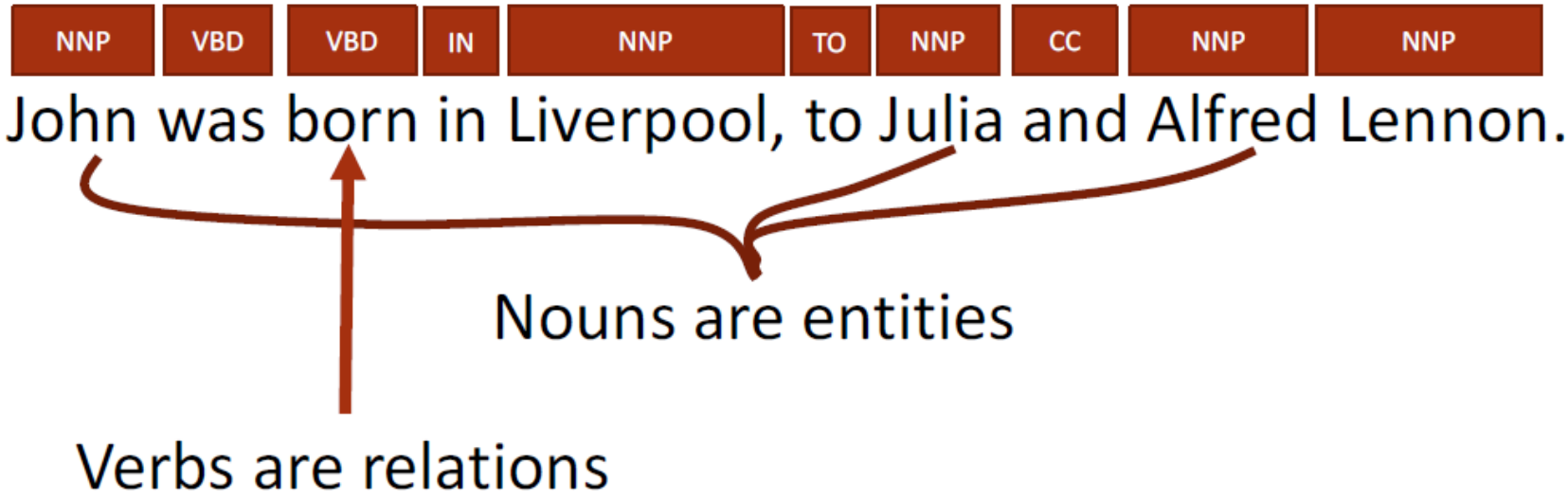
Sentence

Dependency Parsing,
Part of speech tagging,
Named entity recognition…



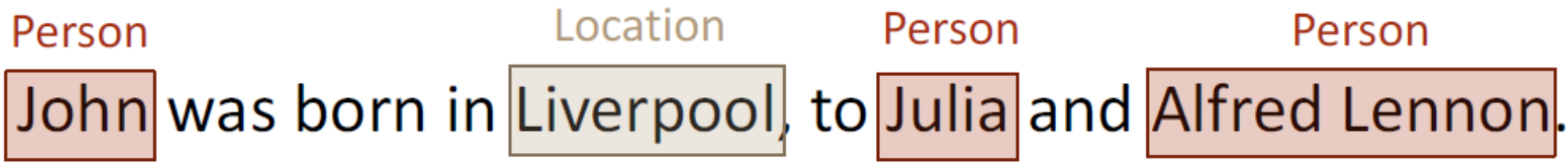
John was born in Liverpool, to Julia and Alfred Lennon.

Tagging the parts of speech (POS)



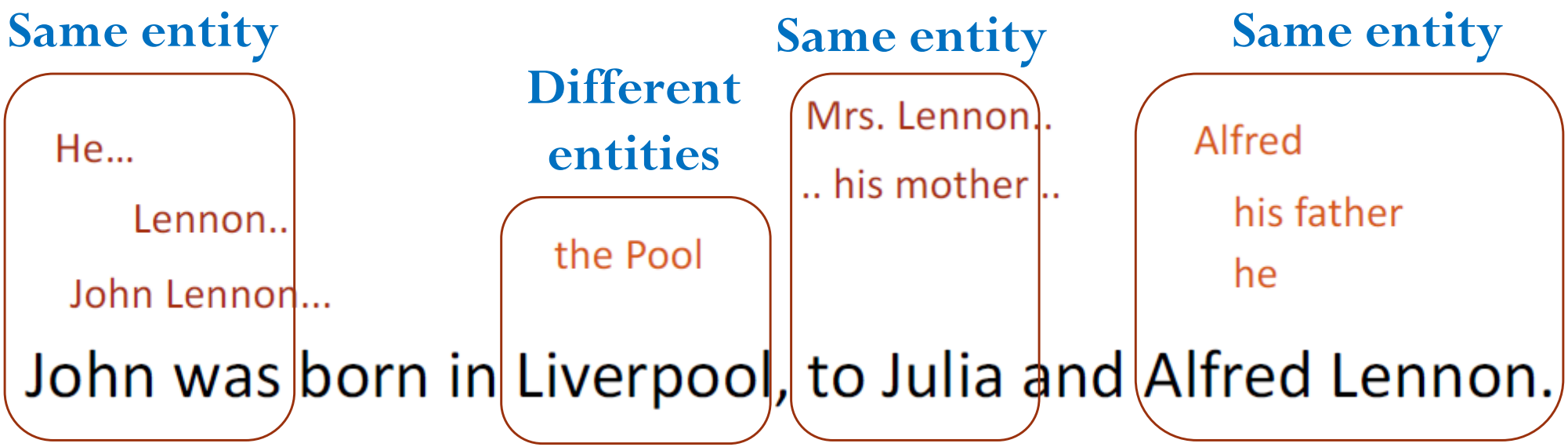
- POS resources:
  - https://parts-of-speech.info/
  - http://nlp.stanford.edu:8080/parser/
  - Python packages: NLTK (https://www.nltk.org/), pyStatParser, TextBlob, Pattern,…

Named entities recognition (NER)



- Machine learning for NLP:
  - Structured classification to add category label (e.g. Person, Location, Organization,…) for entity mentions and entity types
- NER resources:
  - https://spacy.io/ (Python package: SpaCy)
  - http://textanalysisonline.com/spacy-named-entity-recognition-ner

Within-document co-reference



- Machine learning for NLP: Pairwise model for each noun/pronounce to consolidate information
  - Resources: https://stanfordnlp.github.io/CoreNLP/

Named entity resolution

...during the late 60′s and early 70′s, **Kevin Smith** worked with several local...

...the term hip-hop is attributed to **Lovebug Starski**. What does it actually mean...

Like Back in 2008, the Lions drafted **Kevin Smith**, even though Smith was badly...

... backfield in the wake of **Kevin Smith**′s knee injury, and the addition of Haynesworth...

The filmmaker **Kevin Smith** returns to the role of Silent Bob...

Nothing could be more irrelevant to **Kevin Smith**′s audacious ′′Dogma′′ than ticking off...

... The Physiological Basis of Politics," by **Kevin Smith**, Douglas Oxley, Matthew Hibbing...

Entity name: Two main problems

Entities with Same Name

Same type of entities share names

Kevin Smith, John Smith, Springfield, …

Things named after each other

Clinton, Washington, Paris, Amazon, Princeton, Kingston, …

Partial Reference

First names of people, Location instead of team name, Nick names

Different Names for Entities

Nick Names

Bam Bam, Drumpf, …

Typos/Misspellings

Baarak, Barak, Barrack, …

Inconsistent References

MSFT, APPL, GOOG…

Entity resolution approach

Washington drops 10 points after game with UCLA Bruins.

Candidate Generation

Washington DC, George Washington, Washington state, Lake Washington, Washington Huskies, Denzel Washington, University of Washington, Washington High School, …

Entity Types LOC/ORG

Washington DC, ~~George Washington~~, Washington state, Lake Washington, Washington Huskies, ~~Denzel Washington~~, University of Washington, Washington High School, …

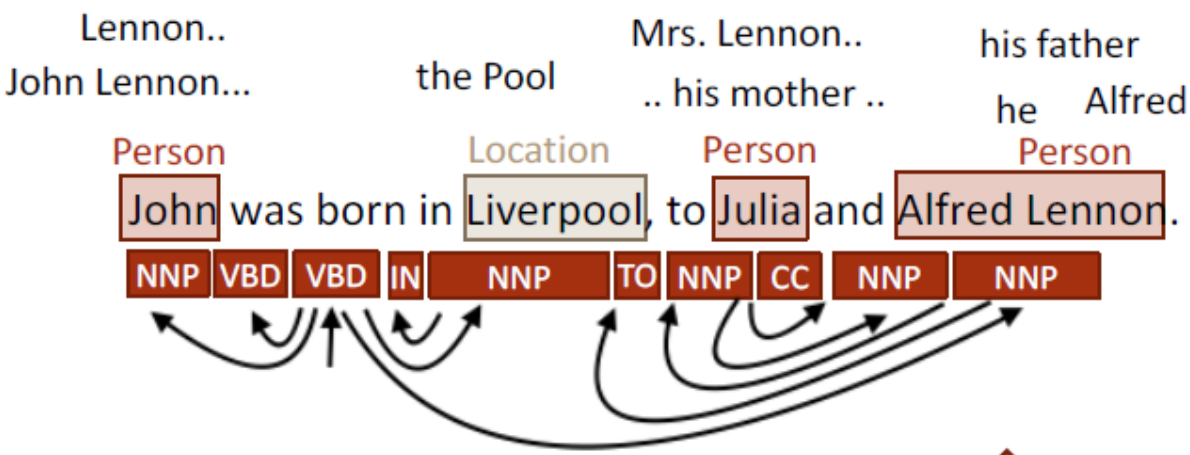
Coreference UWashington, Huskies

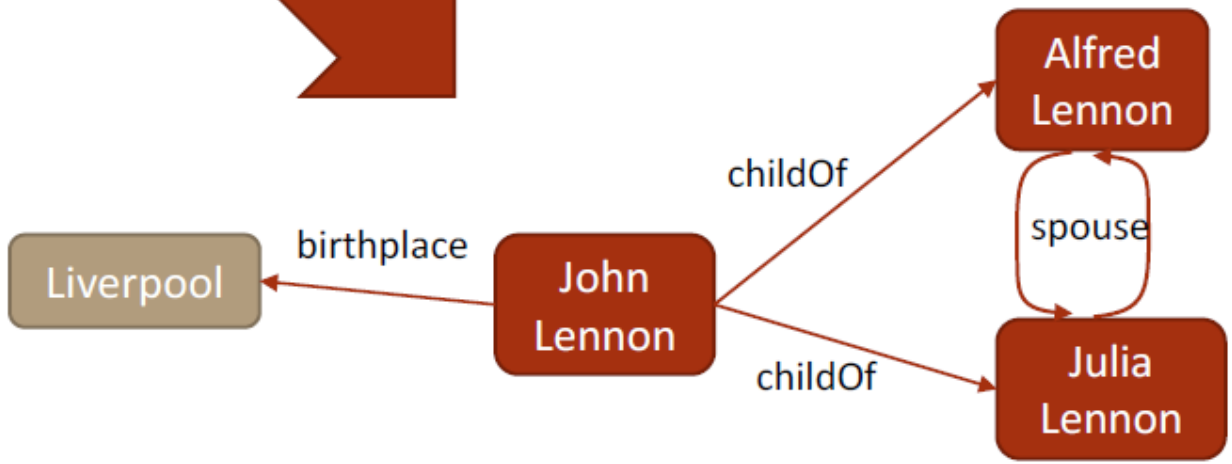
~~Washington DC~~, ~~George Washington~~, ~~Washington state~~, ~~Lake Washington~~, Washington Huskies, ~~Denzel Washington~~, University of Washington, ~~Washington High School~~, …

Coherence UCLA Bruins, USC Trojans

~~Washington DC~~, ~~George Washington~~, ~~Washington state~~, ~~Lake Washington~~, Washington Huskies, ~~Denzel Washington~~, ~~University of Washington~~, ~~Washington High School~~, …

Knowledge extraction





General framework and challenge

timeliness

ID	Sentence
S1	***Obama*** was born in Honolulu, Hawaii, ***USA*** as he has always said.
S2	President Clinton and ***Obama*** attended the funeral of former Israeli Prime Minister, and were scheduled to fly back to the ***US*** together.
S3	***Barack Obama*** is the 44th and current President of the ***United States***
S4	A clip of ***Barack Obama*** reading from his book "***Dreams of My Father***" has been shared out of context.
...	...

Text Corpus

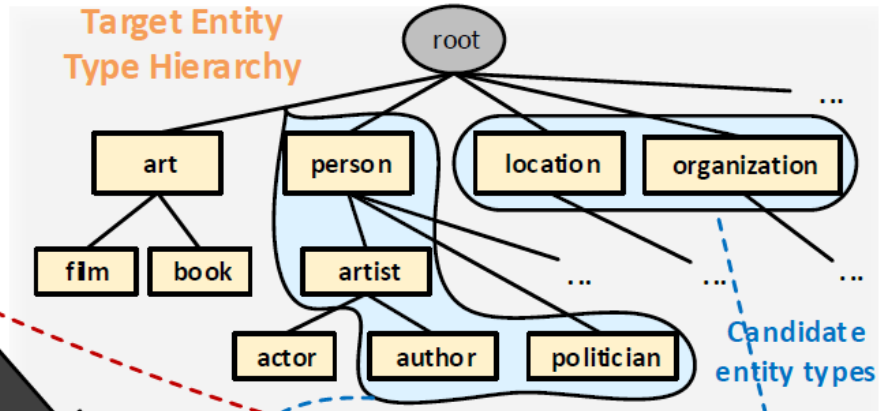
Automatically Labeled Training Data

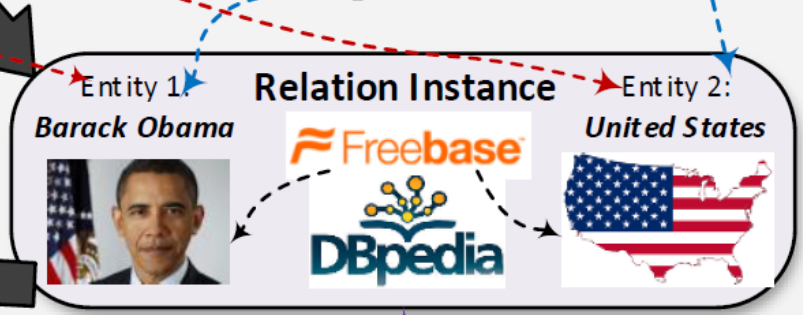
Relation Mention: (*"Obama", "USA"*, S1)
Types of Entity 1: {person, politician, artist, author}, **Entity 2:** {ORG, LOC}
Relation Types: {president_of, born_in, citizen_of, travel_to}

Relation Mention: (*"Obama", "US"*, S2)
Types of Entity 1: {person, politician, artist, author}, **Entity 2:** {ORG, LOC}
Relation Types: {president_of, born_in, citizen_of, travel_to}

Relation Mention: (*"Barack Obama", "United States"*, S3)
Types of Entity 1: {person, politician, artist, author}, **Entity 2:** {ORG, LOC}
Relation Types: {president_of, born_in, citizen_of, travel_to}

Relation Mention: (*"Barack Obama", "Dreams of My Father"*, S4)
Types of Entity 1: {person, politician, artist, author}, **Entity 2:** {book}
Relation Types: {author_of}





Candidate relation types

Relation Type	Entity 1	Entity 2
president_of	*Barack Obama*	*United States*
born_in	*Barack Obama*	*United States*
citizen_of	*Barack Obama*	*United States*
visit	*Barack Obama*	*United States*

KB Relations of Target Types

Source: CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases (Ren at al. 2017)

Information extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring the facts

×

3 LEVELS OF SUPERVISION

Supervised

Semi-supervised

Unsupervised

Effect of supervision on extraction

Precision,
Human efforts

Recall,
Speed

Information extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring the facts

×

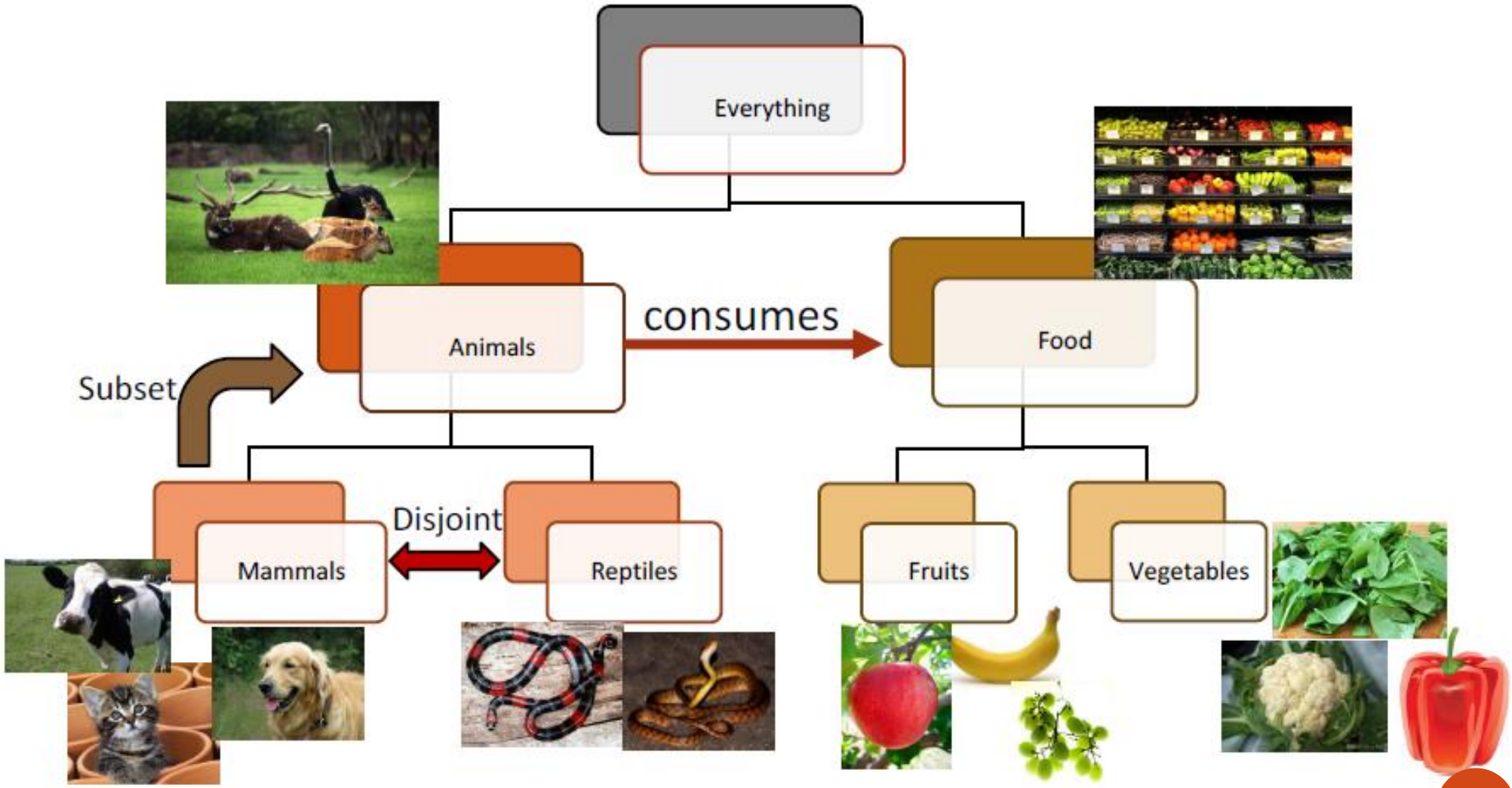
3 LEVELS OF SUPERVISION

Supervised

Semi-supervised

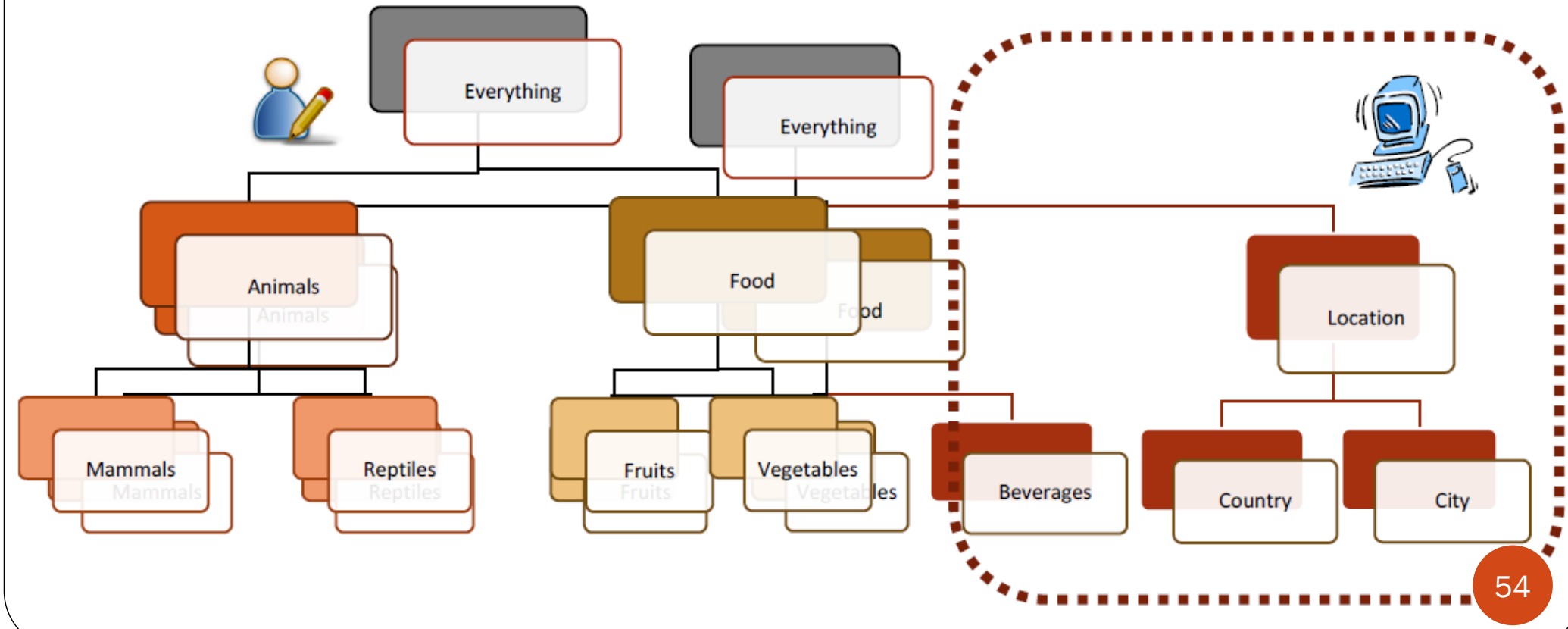
Unsupervised

Defining domain: Manual



Defining domain: Semi-automatic

- Subset of types are manually defined
- SSL methods discover new types from unlabeled data



Define domain: Automatic

- Any noun phrase is a candidate entity
  - Dog, cat, cow, reptile, mammal, apple, greens, mixed greens, lettuce,…
- Any verb phrase is a candidate relation
  - Eats, feats on, grazes, consumes,…

Information extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring candidate facts

×

3 LEVELS OF SUPERVISION

Supervised

Semi-supervised

Unsupervised

Learning extractors

- Supervised: high precision patterns
  - <PERSON> plays in <BAND>

- Semi-supervised: Bootstrapping to learn patterns
  - Create examples (John Lennon, Beatles), find patterns
  - Manually correct incorrect patterns

- Unsupervised: cluster phrases with constraints
  - Identify candidate verb phrases, find candidate arguments, cluster by NER types

Information extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring candidate facts

×

3 LEVELS OF SUPERVISION

Supervised

Semi-supervised

Unsupervised

Scoring the candidate facts

- Human defined scoring function or

  Scoring function learnt using supervised ML with large amount of training data
  {expensive, high precision}

- Small amount of training data is available

  scoring refined over multiple iterations using both labeled and unlabeled data

- Completely automatic (Self-training)

  Confidence(extraction pattern) $\propto$ (#unique instances it could extract)
  Score(candidate fact) $\propto$ (#distinct extraction patterns that support it)
  {cheap, leads to semantic drift}

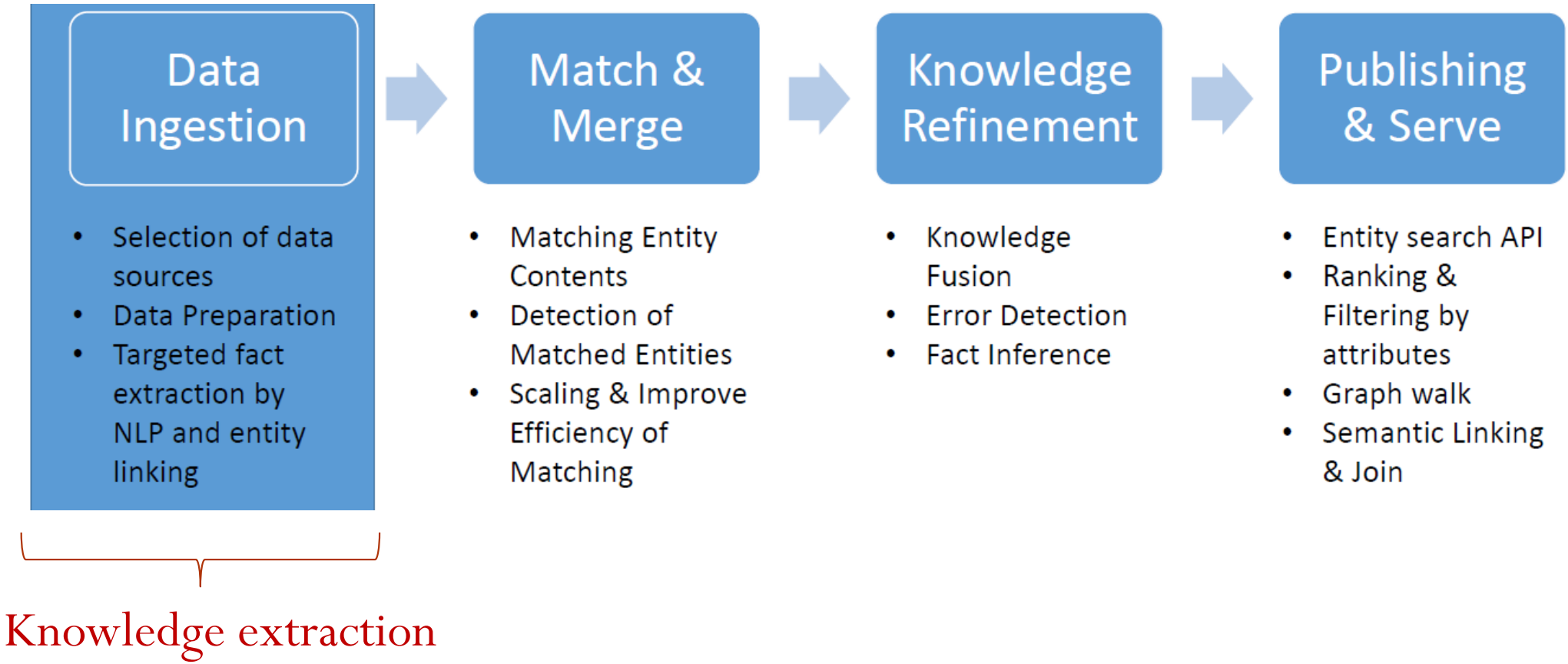
Information extraction in practice

	Defining domain	Learning extractors	Scoring candidate facts	Fusing extractors
ConceptNet				
NELL				Heuristic rules
Knowledge Vault				Classifier
OpenIE				

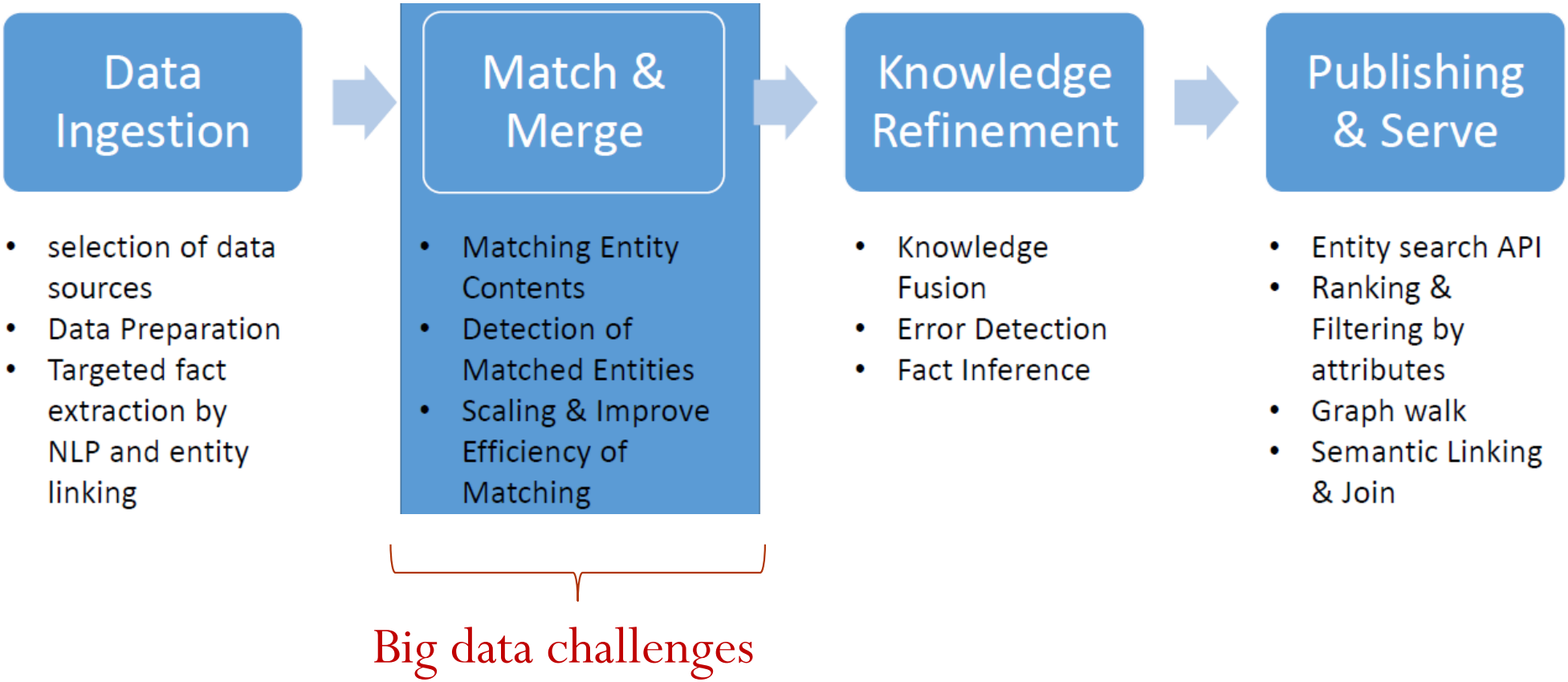
Knowledge extraction: Key points

- Built on the foundation of NLP techniques
  - POS tagging, dependency parsing, named entity recognition, co-reference resolution…
  - Challenging research problems with very useful outputs
- Information extraction techniques using NLP and ML to define domains, extract entities and relations, score candidates
  - Trade-off between manual & automatic methods

Construct KG

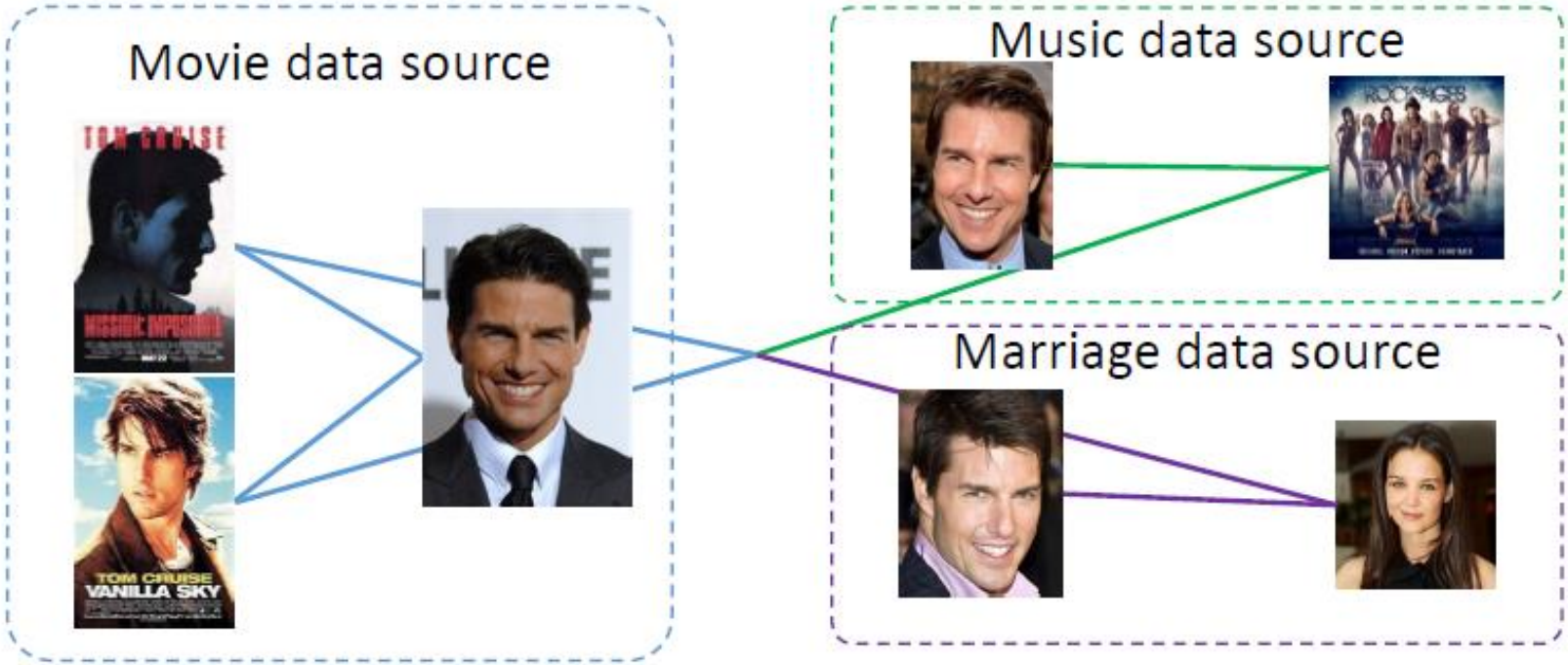


Construct KG



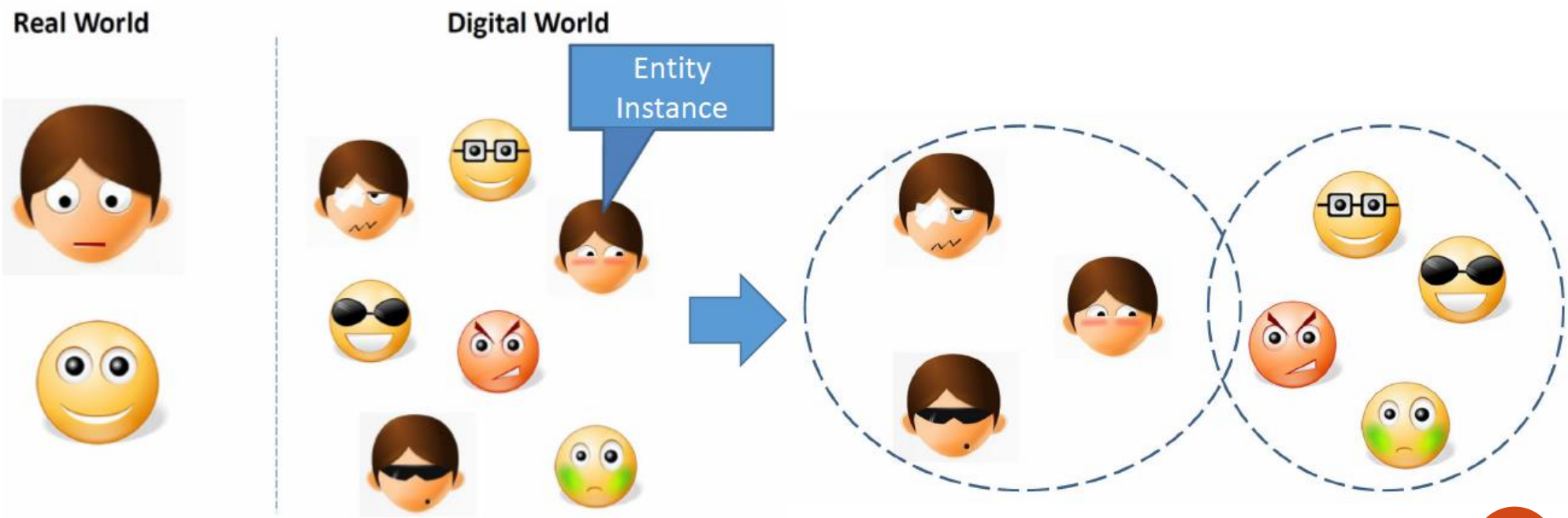
Entity matching (EM)

- Identify and discover instances referring to the same real-world entity to increase data quality
- Synonyms: Entity linking, entity resolution, reference reconciliation, deduplication,…

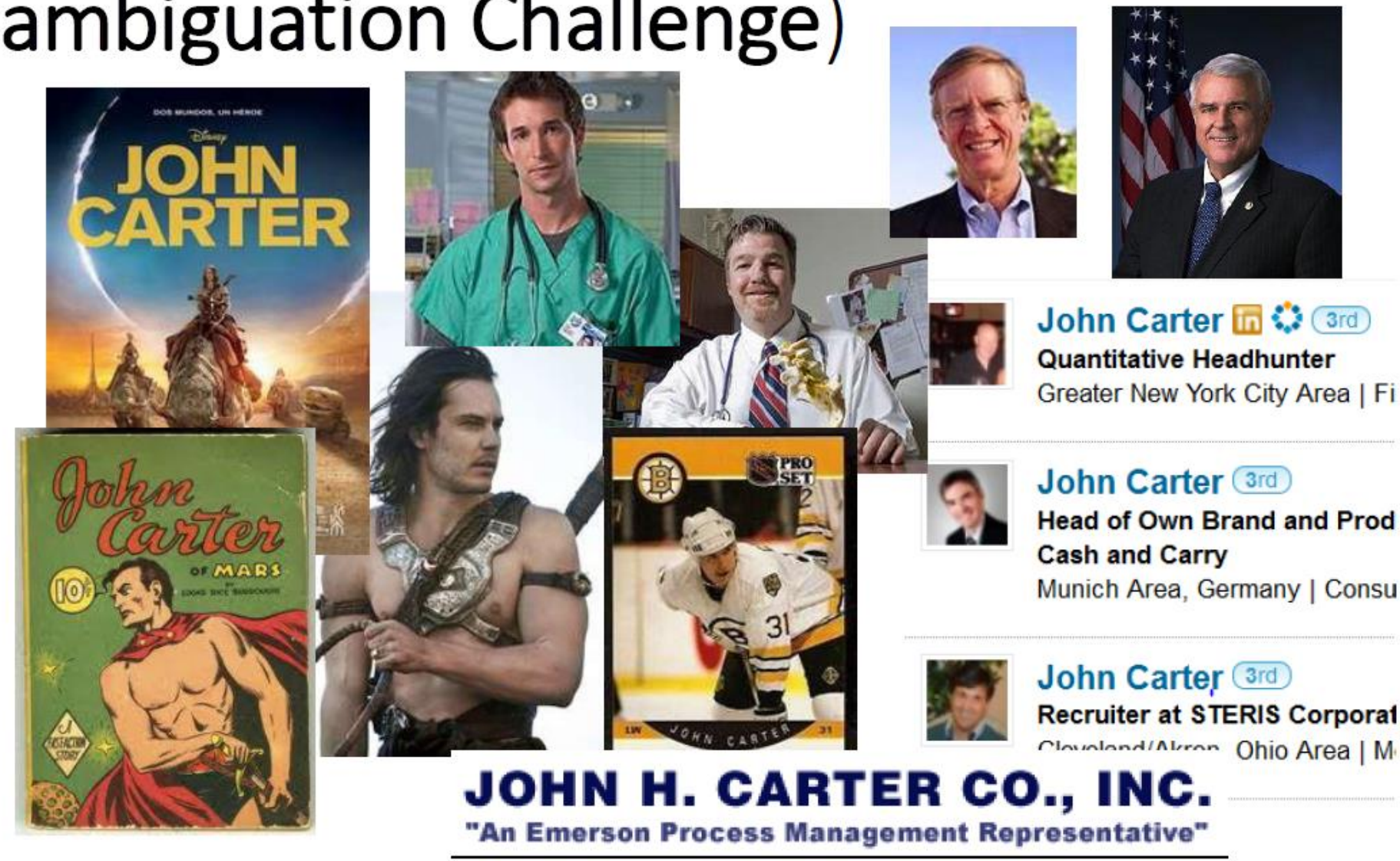


Data quality challenge

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



Open Domain Entity Matching (Disambiguation Challenge)



EM challenges in big data

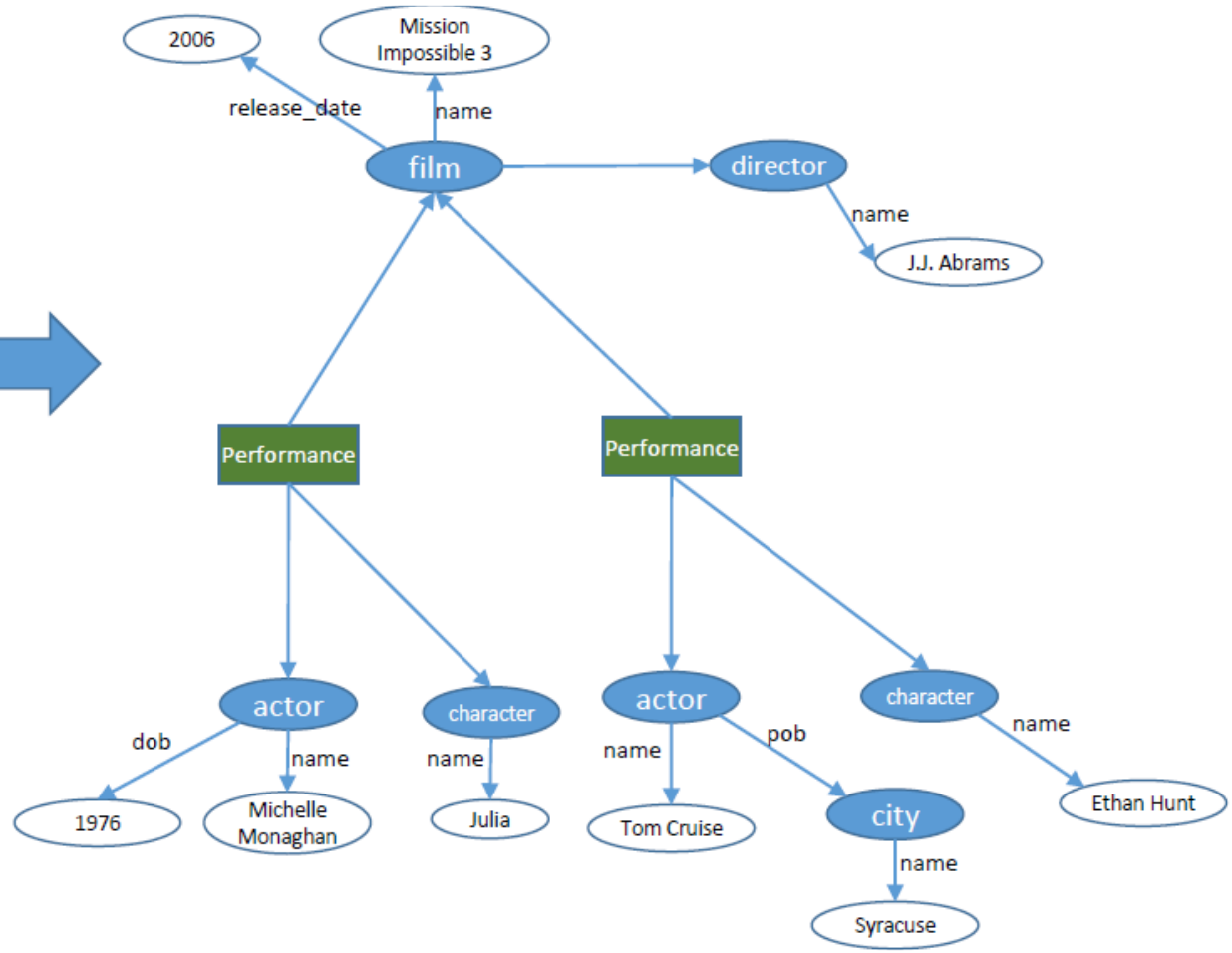
可扩展性

- Big data set: Scalable, efficient, parallel techniques
- Multi-domain: Efficient techniques to manage executions within and across domains
- Linked, connected and relational data: Efficient techniques to leverage the diversity of connections and representation

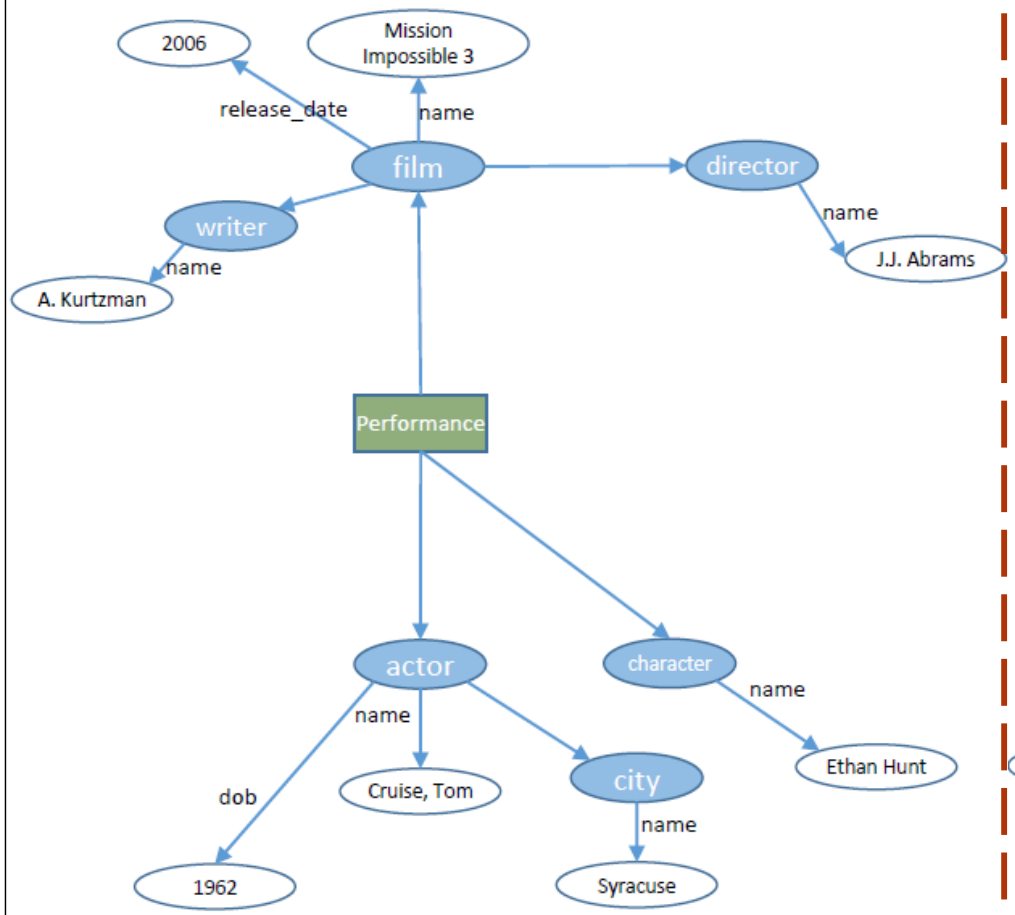
Techniques for EM

- Collective resolution in linked data
- Distance based models
- Probabilistic matching models (not covered)
  - Supervised and semi-supervised learning
  - Active learning
- Declarative matching rules and constraints (not covered)
  - Disjunction of conjunction rules
  - Constraint-based clustering

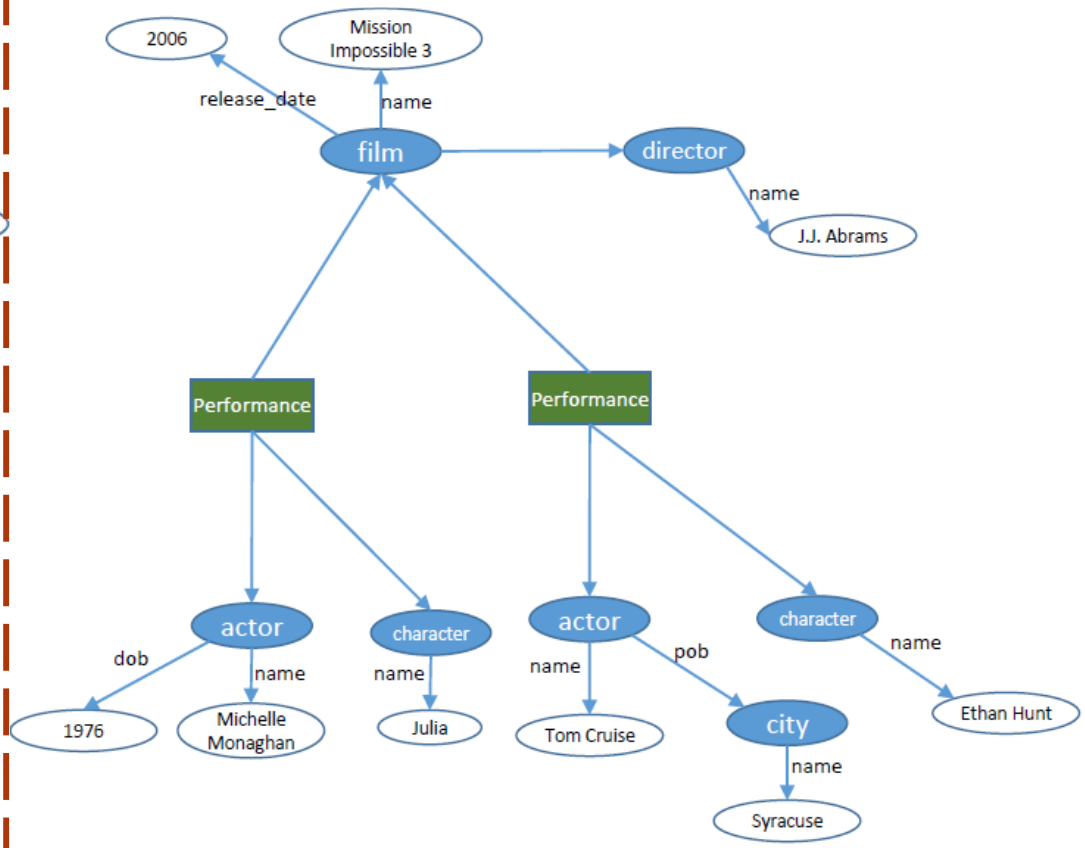
Collective resolution in linked data



Graph model and conflation

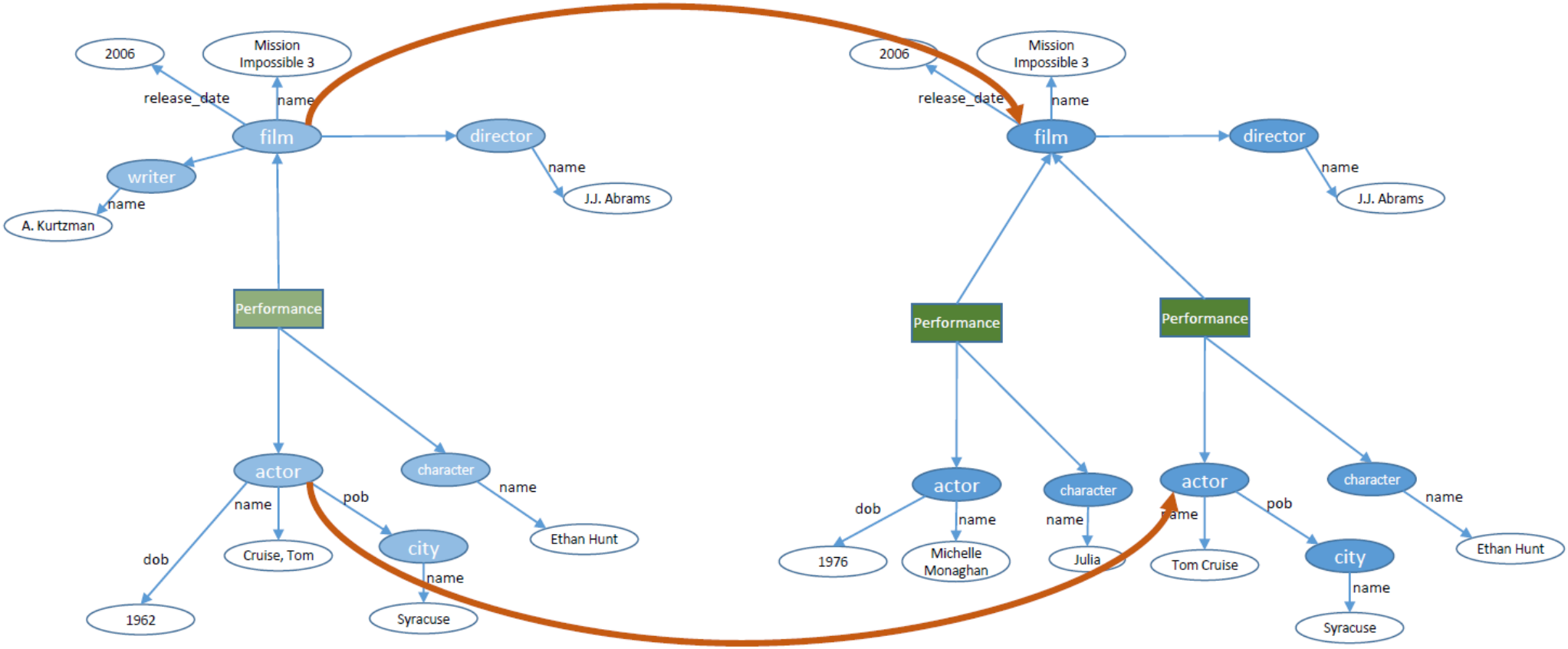


Database 1

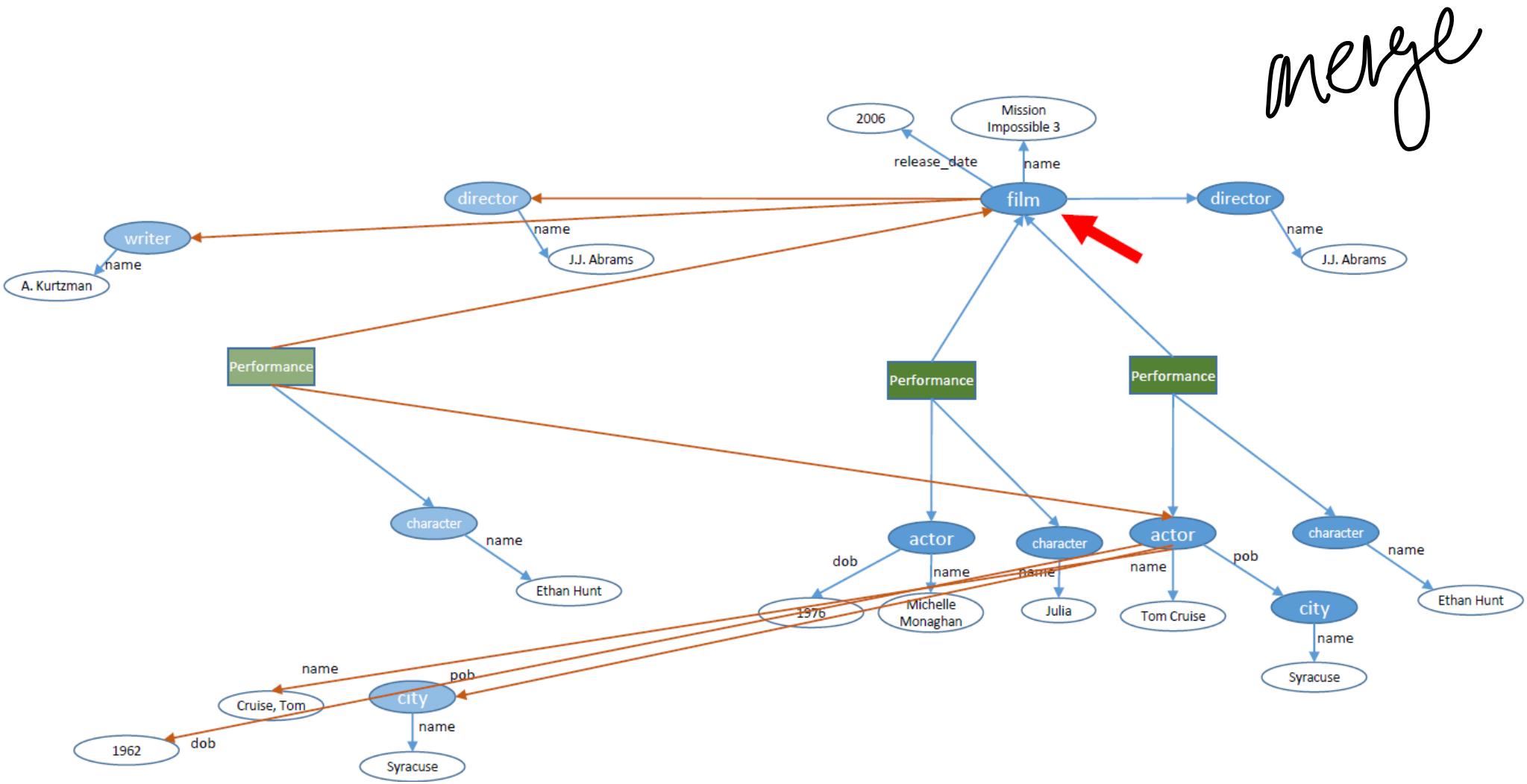


Database 2

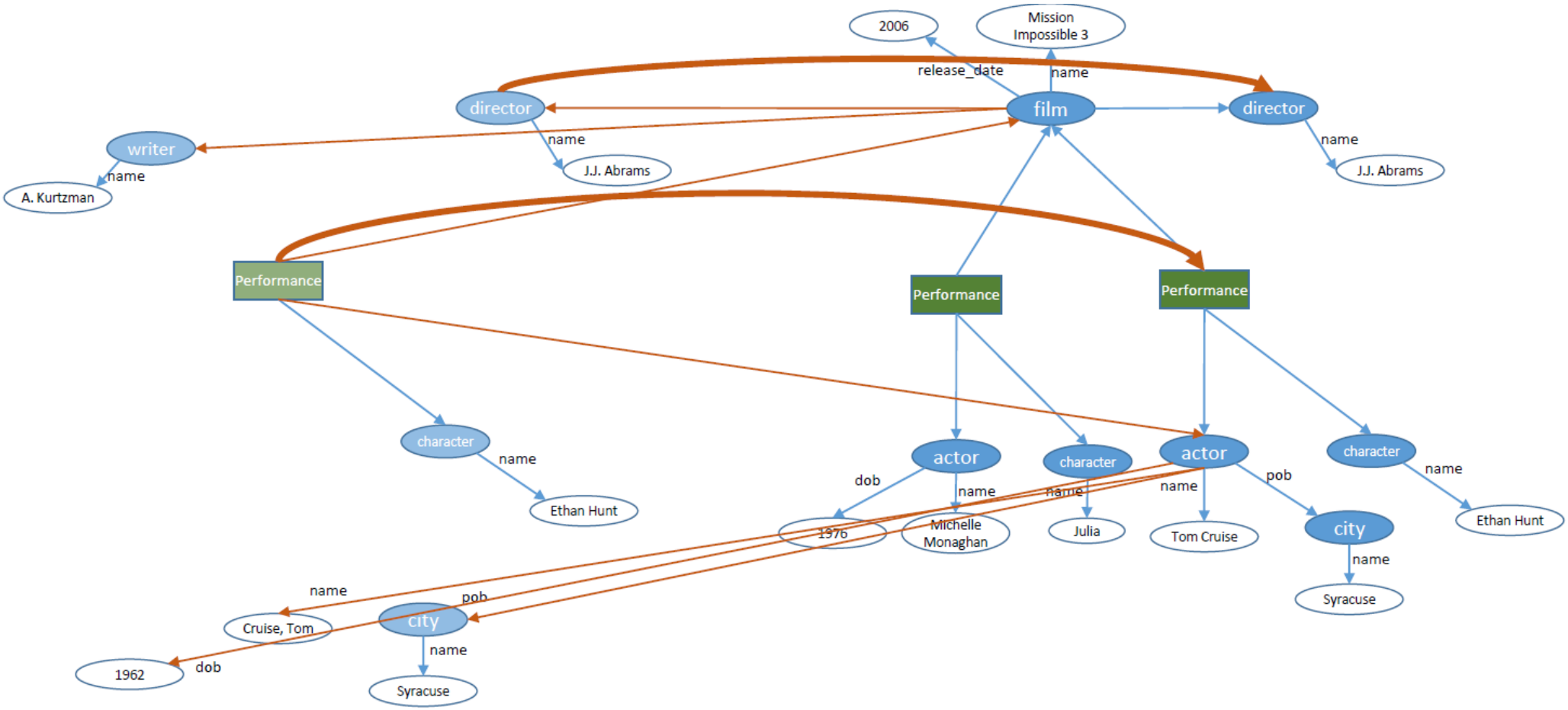
Graph model and conflation



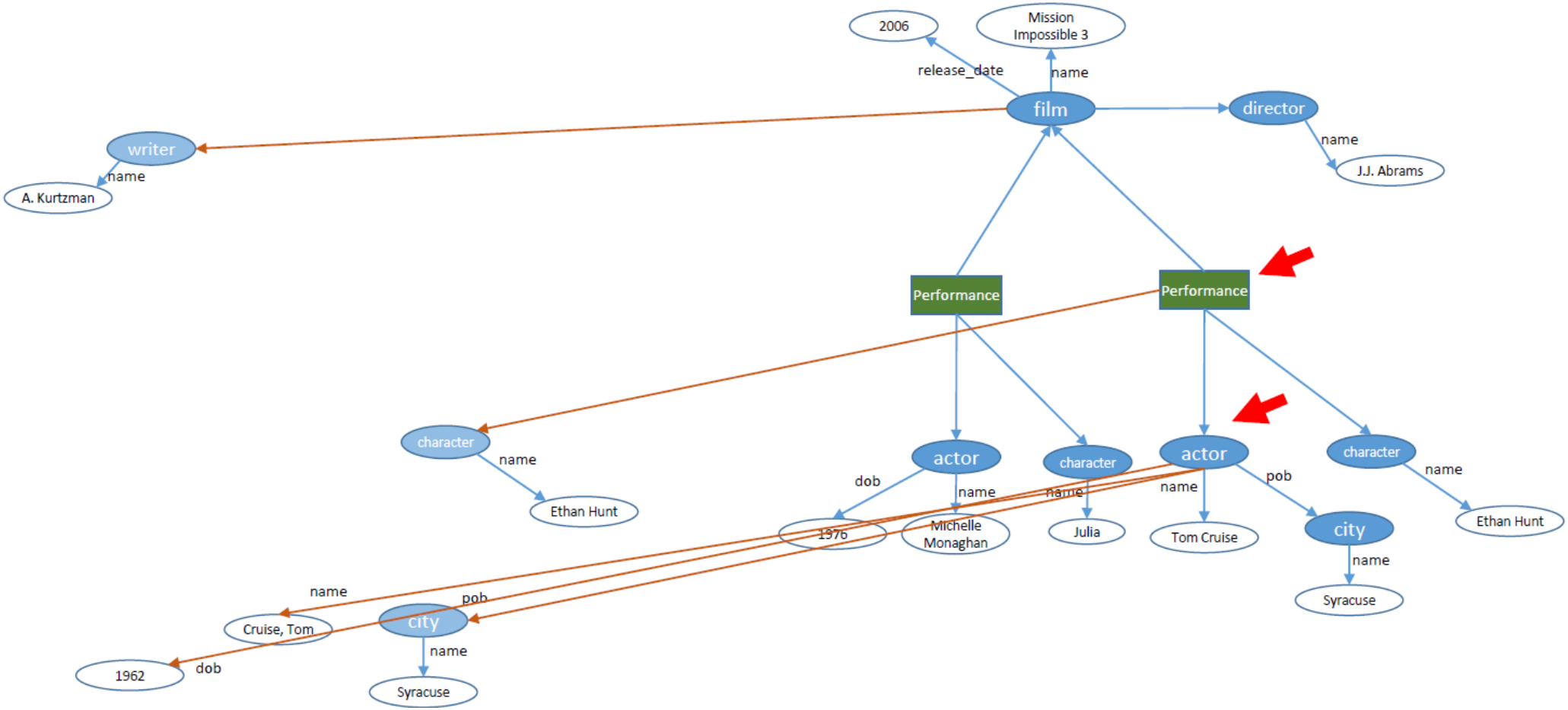
Graph model and conflation



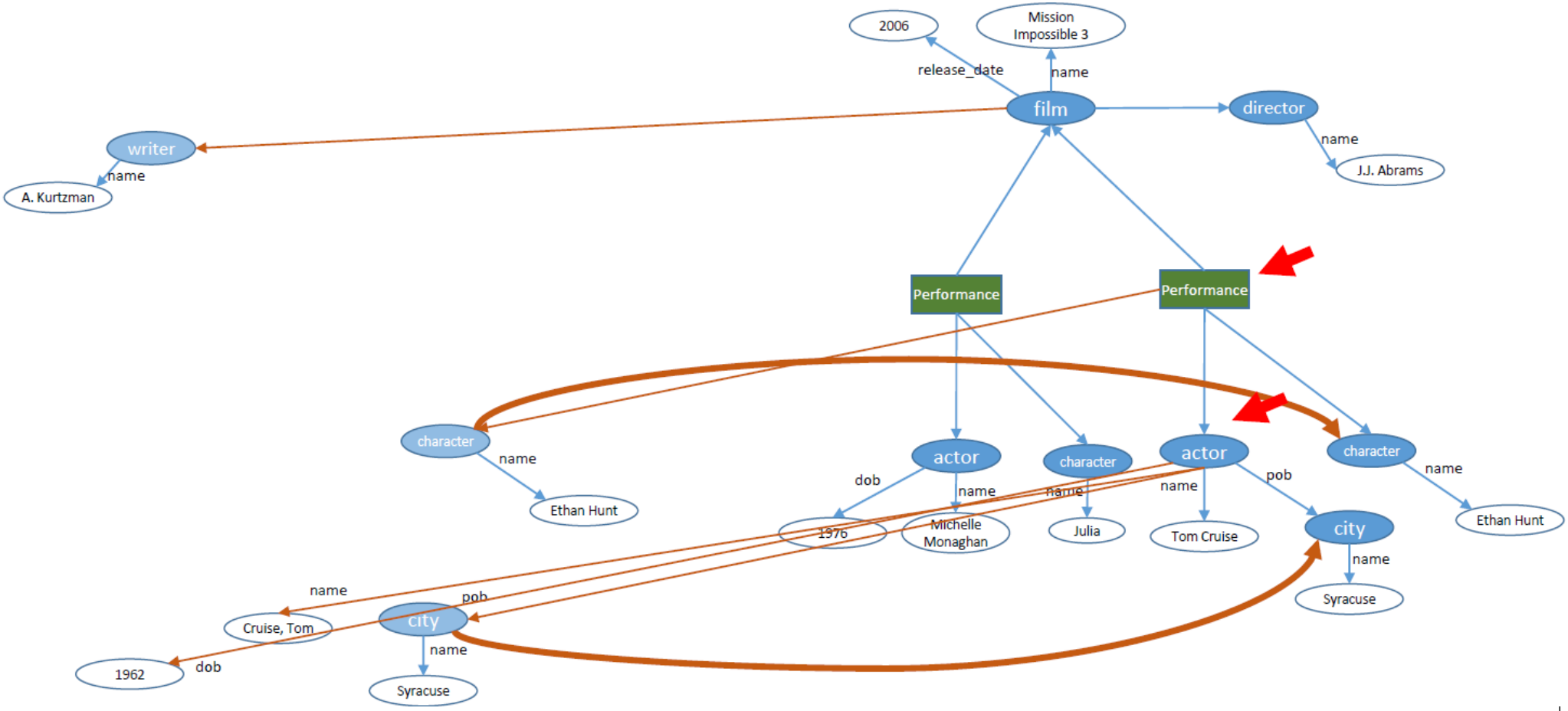
Graph model and conflation



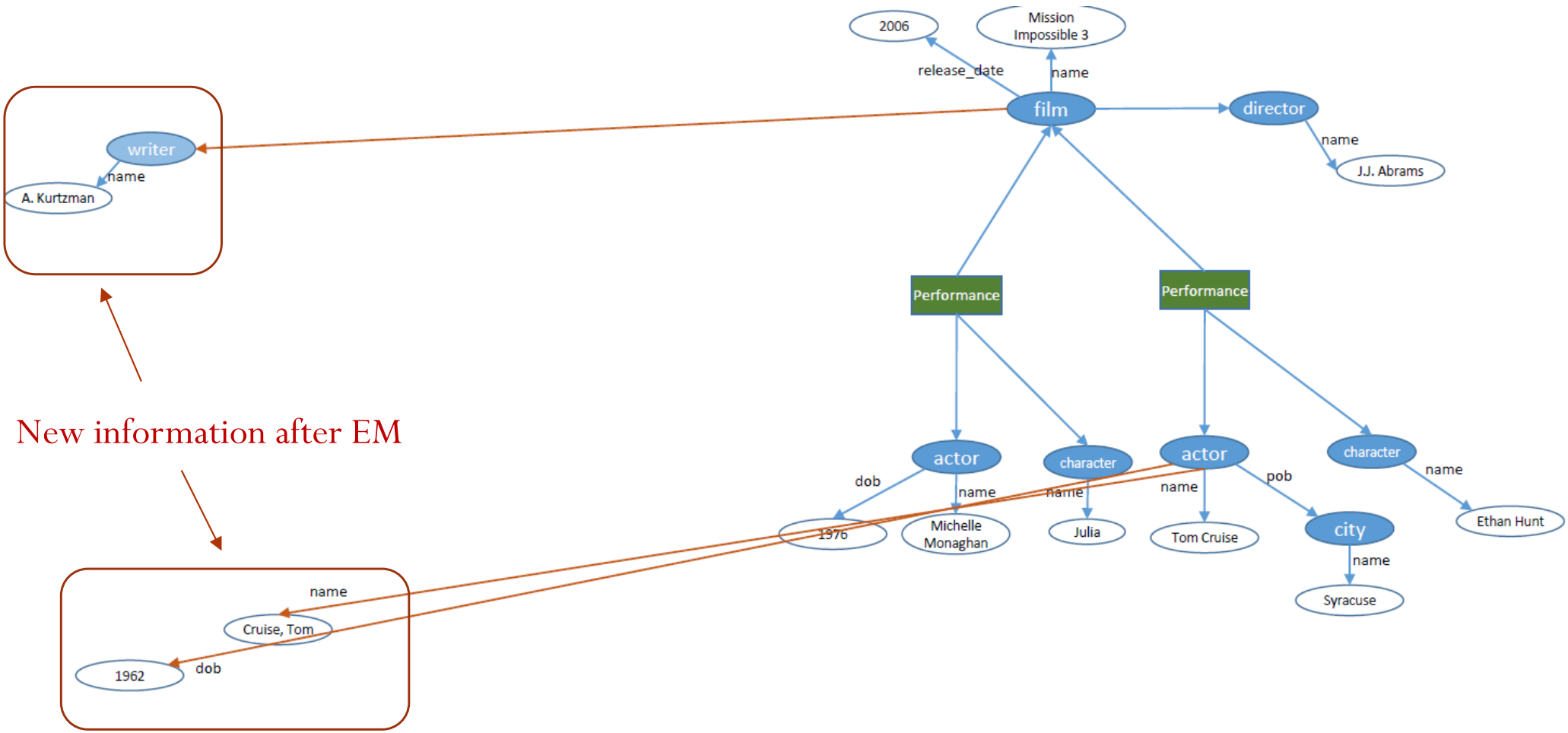
Graph model and conflation



Graph model and conflation



Graph model and conflation



New information after EM

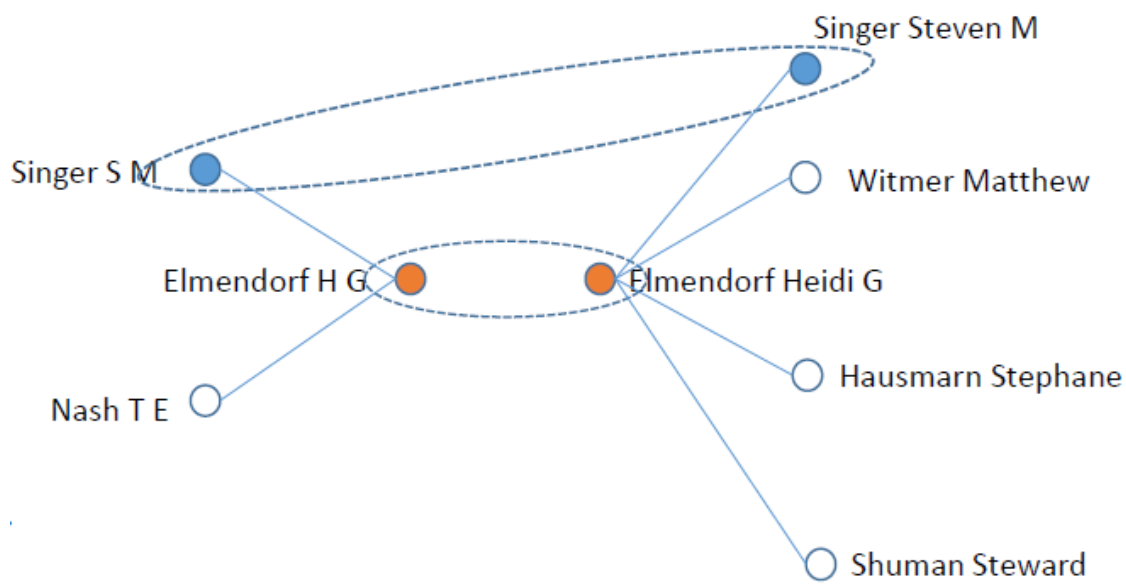
Collective resolution in linked data

- Entity similarity based on connection 相似度得分
- Measures:
  - SimRank: Two nodes are similar if they are referenced by similar nodes

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

where **I(a)** is the neighbors of the node **a**

  - Adar measure
  - Katz score
  - …



Distance-based EM models

- Threshold method:
  - Compute similarity vector using some similarity measure
  - Classify the vectors as Match if the aggregate similarity is larger than a threshold **t**; otherwise, UnMatch
- Example:
  - Similarity vector **<0.8, 0.7, 1.0>**
  - Verify: $\mathbf{sim(Name_1, Name_2) + sim(Prof_1, Prof_2) + sim(Add_1, Add_2) > t}$

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

Distance-based EM models

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

- Solution:
  - Indexing
  - Locality-sensitive hashing (CS 753)
  - MapReduce-based solution (MapReduce lecture)

subtask

Distance-based EM models

- Inverted index:

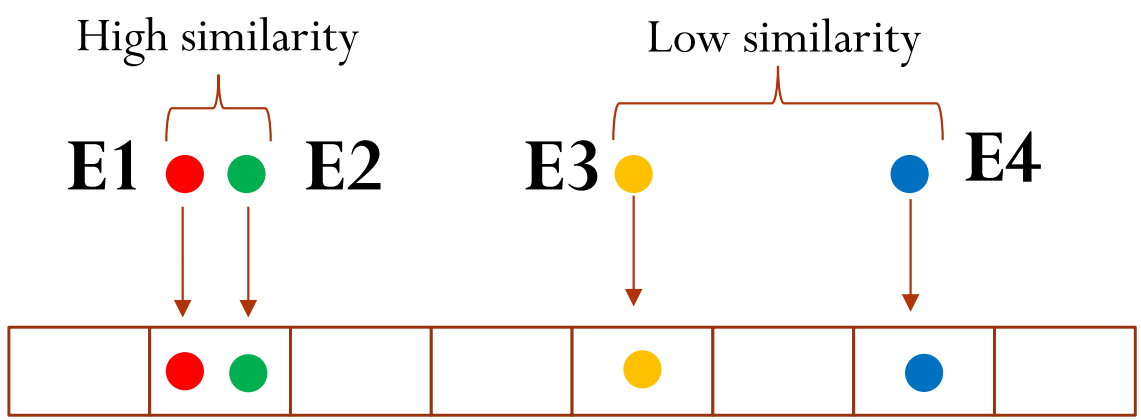
Entity	Features
E1	F1, F2
E2	F1, F3, F4
E3	F3
E4	F4

Key	Post list
F1	E1, E2
F2	E1
F3	E2, E3
F4	E2, E4

- Why efficient?
  - Candidates **(E1, E2), (E2, E3), (E2, E4)** compared to all possible **6** pairs **(E1, E2), (E1, E3), …, (E3, E4)**
  - Can extend to idf weight for text/document

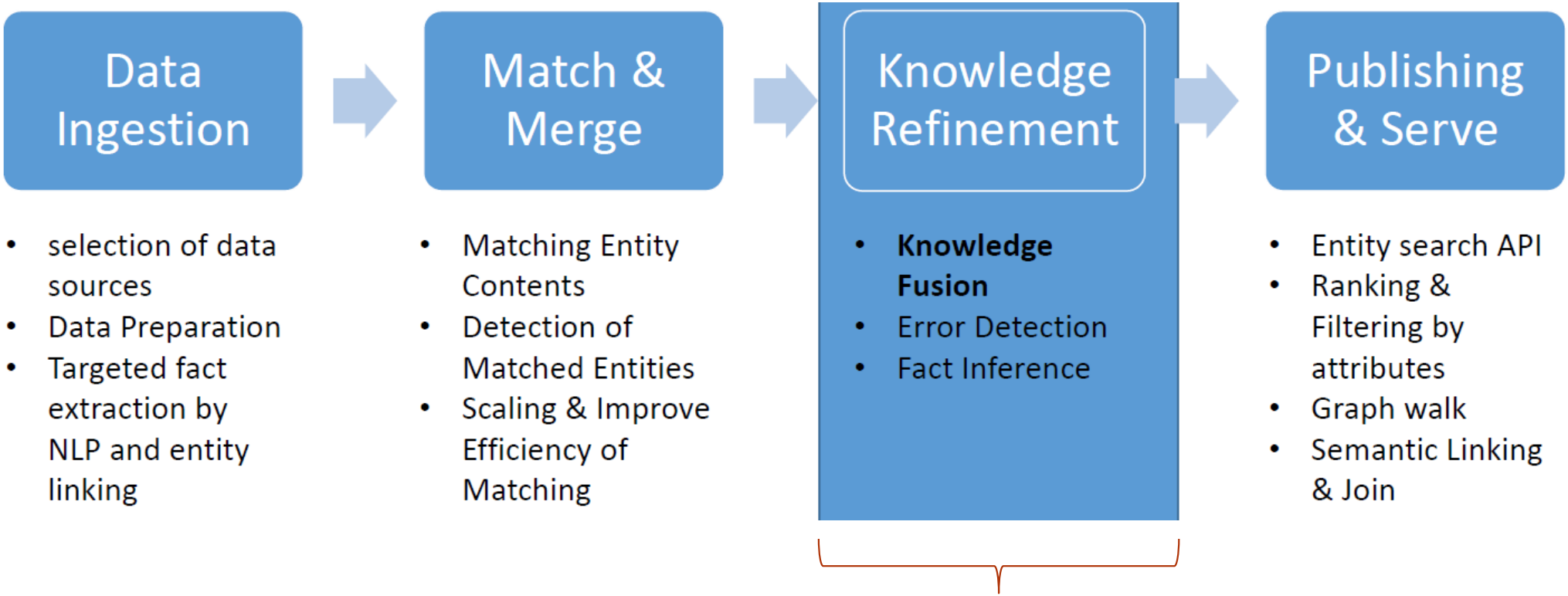
Distance-based EM models

- Locality-sensitive hashing:



- Why efficient?
  - Only check candidate pairs **(E1, E2)** in the bucket compared to all possible **6** pairs **(E1, E2), (E1, E3), …, (E3, E4)**
  - Can extend to entities (images, text) with metric distance

Construct KG



Post-processing after merging entities

Knowledge fusion 融合

- After merging entity nodes, there are conflicting facts and connections
- Solutions: Majority voting 多数

	Name	Birth date	Birthplace	Age	Children
S1	**Barack Obama**	**August 4, 1961**	**Hawaii, US**	**56**	**Malia, Sashia**
S2	**Barack Hussein Obama**	**08/04/1961**	**Honolulu**	**56**	**Malia**
S3	**Barack H. Obama**	07/04/1961	Kenya	**56**	Ivanka

Majority voting and correctly identify similar values. Accuracy < 1.0

Knowledge fusion

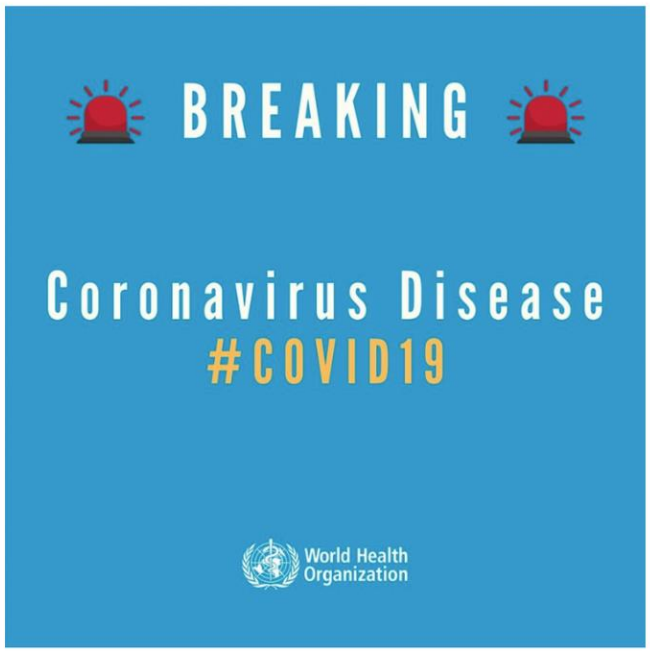
- After merging entity nodes, there are conflicting facts and connections
- Solutions: Identify authoritative source

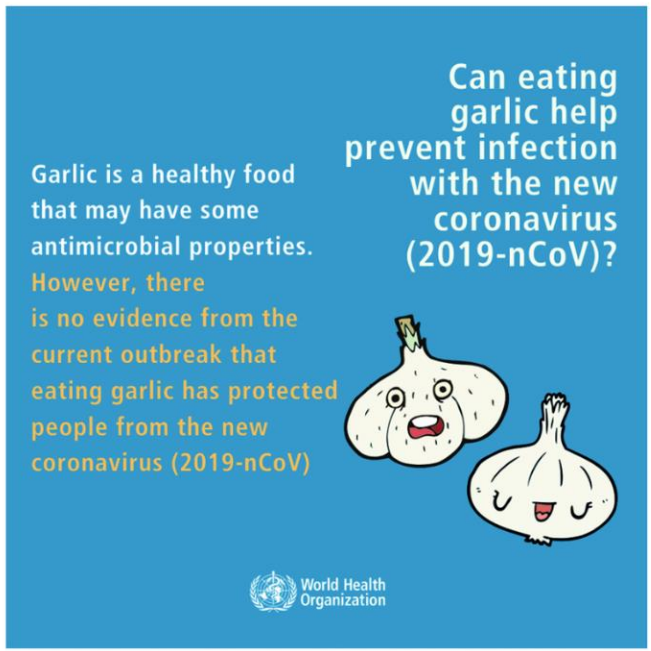
	Name	Birth date	Birthplace	Age	Children
S1	**Barack Obama**	**August 4, 1961**	**Hawaii, US**	**56**	**Malia, Sashia**
S2	**Barack Hussein Obama**	**08/04/1961**	**Honolulu**	**56**	**Malia**
S3	**Barack H. Obama**	07/04/1961	Kenya	**56**	Ivanka

Accurate source, Accuracy = 1.0

Knowledge fusion

- After merging entity nodes, there are conflicting facts and connections
- Solutions:
  - Fact checking tutorial (https://shiralkarprashant.github.io/fact-checking-tutorial-KDD2018/): not only for KGs but also for fake news



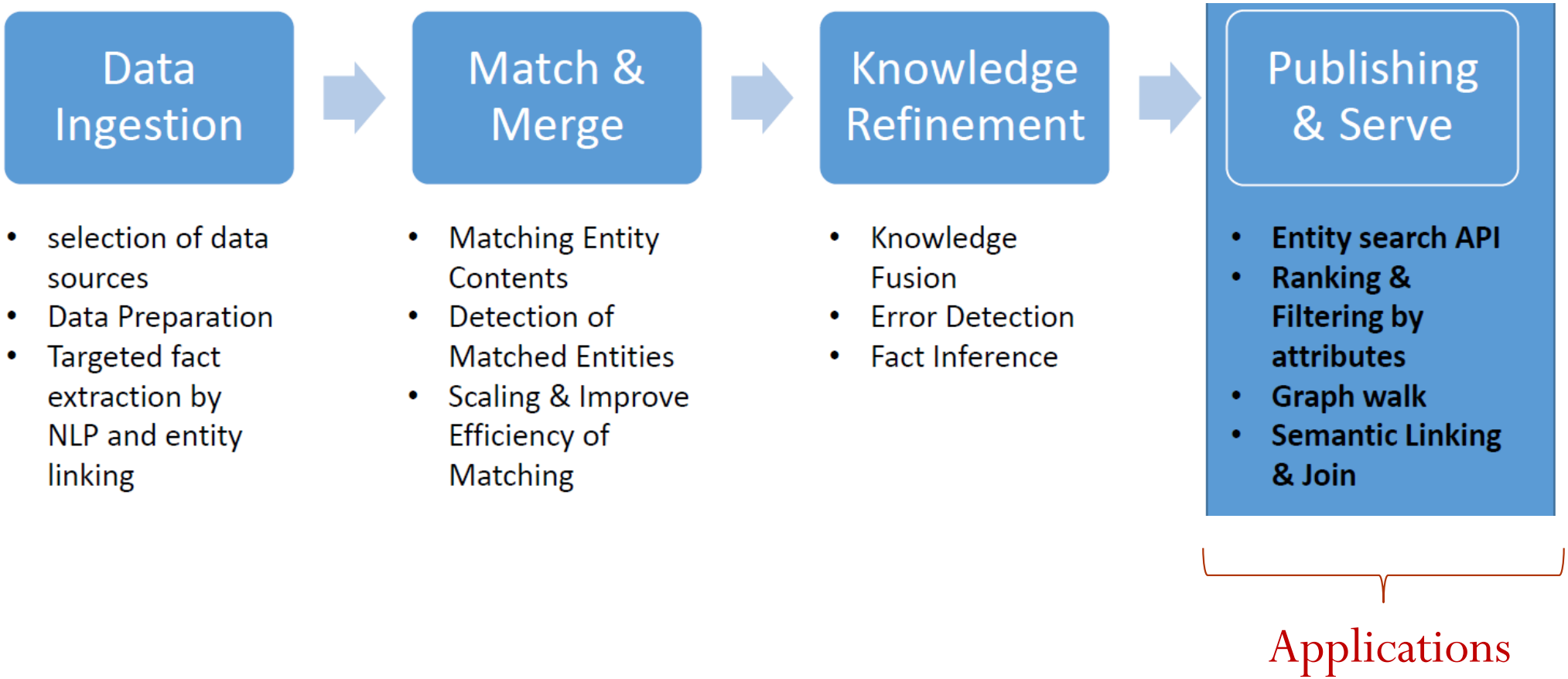


Error detection and fact inference

- Error detection using data quality rules:
  - Functional dependency (e.g. Zip → City)
  - Inconsistency (e.g. data_of_birth > date_of_death)
  - Outlier detection from graph

- Fact inference:
  - Further enrichment or data completion by internal information (e.g. dominant type and label) or external information (e.g. feedback)

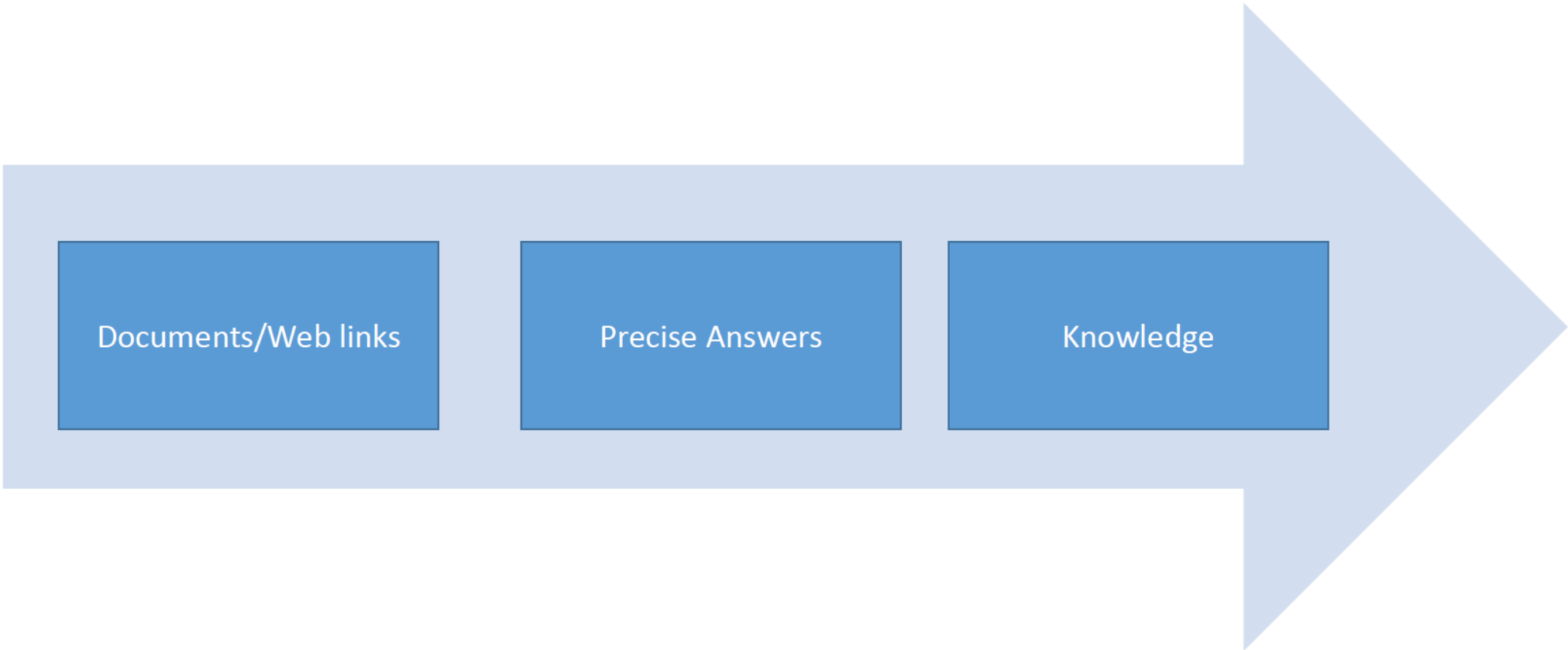
Construct KG



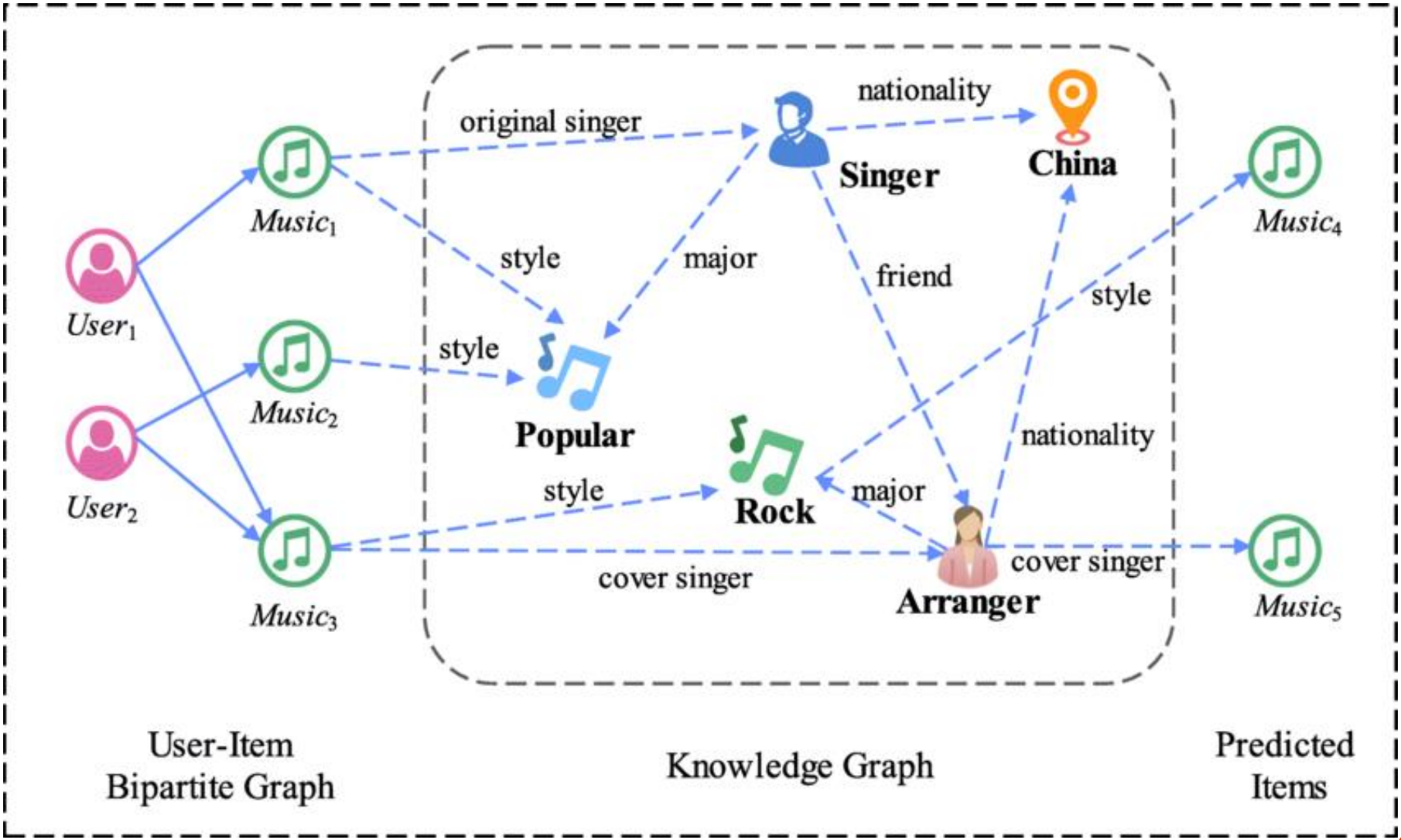
Outline

- Introduction of knowledge graph
- How to construct knowledge graph
- Challenges and success stories

Why KGs serve?

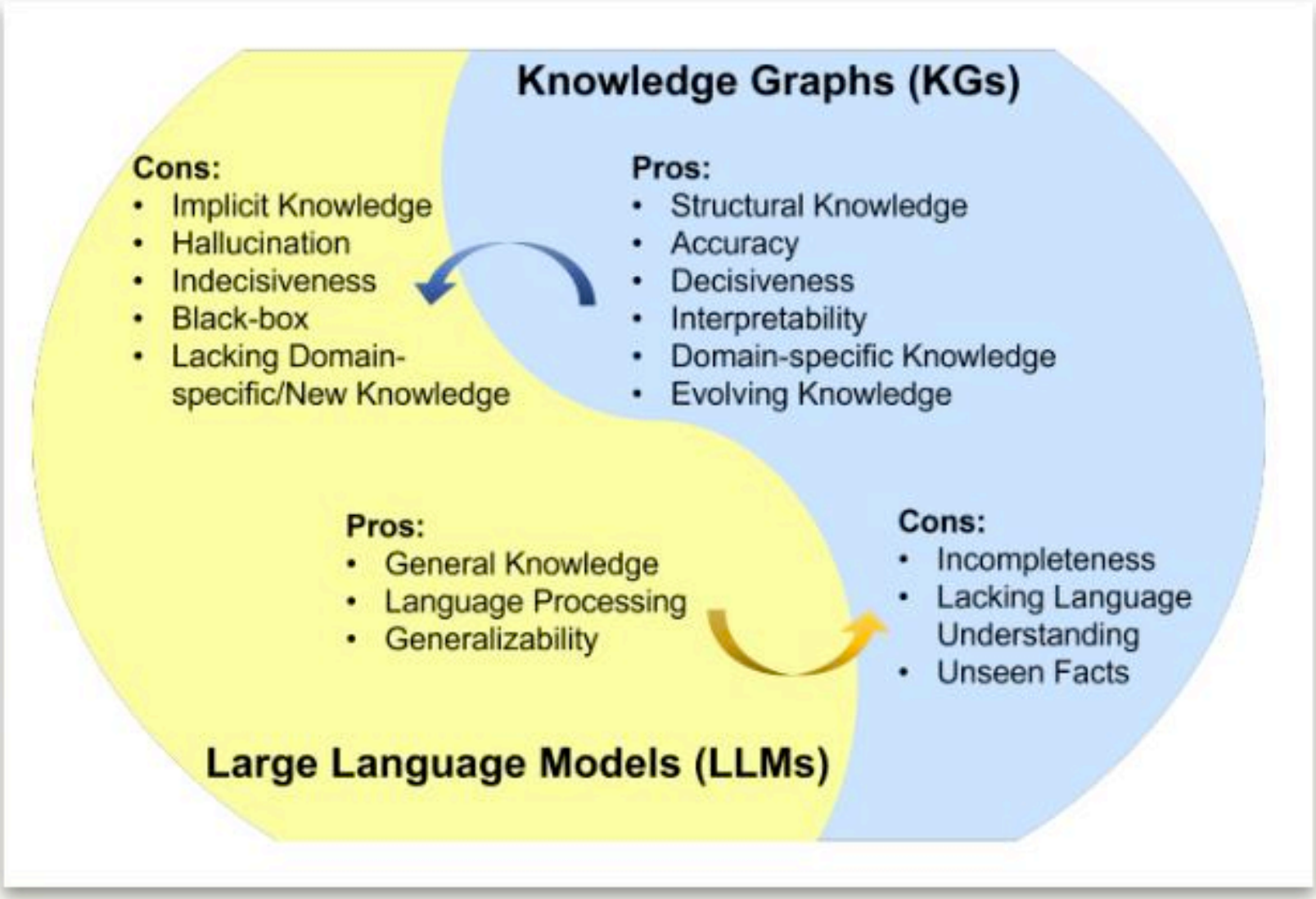


KGs used in recommender systems



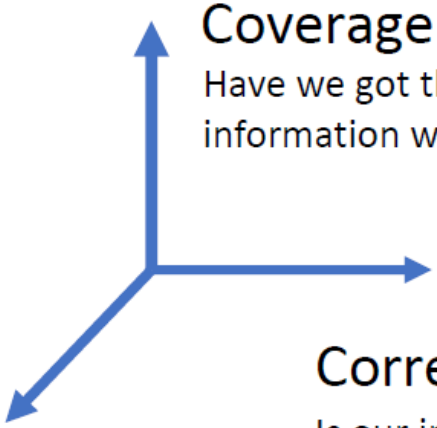
Source: FIRE: knowledge-enhanced recommendation with feature interaction and intent-aware attention networks (2022)

Fusion of KGs and LLMs



https://arangodb.com/2023/08/bridging-knowledge-and-language-arangodb-empowers-large-language-models-for-real-world-applications/

Challenges of scaled KGs



Coverage
Have we got the information we need?

Freshness
Is information up to date?

Correctness
Is our information accurate?

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

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

Success stories

openie.allenai.org

Open Information Extraction

yago
select knowledge

ConceptNet
An open, multilingual knowledge graph

NELL Knowledge Base Browser
CMU Read the Web Project

Open Information Extraction

Hosted by AI2 Created at

Argument 1: Beatles **Relation:** **Argument 2:** All Search

302 answers from **2640 sentences** (cached)

all | location (27) | film location (25) | statistical region (22) | name source (22) | governmental jurisdiction (19) | misc. | more types

- citytown (19)
- tv actor (19)
- employer (18)
- person (18)
- hall of fame inductee (16)
- organization scope (15)
- deceased person (15)
- fictional setting (14)
- artist (14)
- actor (13)
- award winner (12)
- person or entity appearing in film (12)
- place of interment (11)

- were bigger than **Jesus** (104)
- came to America (95)
- appeared on **The Ed Sullivan Show** (81)
- Here Comes the Sun (64)
- broke up in 1970 (56)
- are a great band (50)
- came to America (45)
- perform on **The Ed Sullivan Show** (39)
- were **Musical ensemble** (36)
- are now on **ITunes** (29)
- released their first album (28)
- came on the scene (27)

ConceptNet

An open, multilingual knowledge graph

Documentation FAQ Chat

en beatles

An English term in ConceptNet 5.7

Sources: Open Mind Common Sense contributors, DBPedia 2015, JMDict 1.07, OpenCyc 2012, English Wiktionary, French Wiktionary, and Open Multilingual WordNet
View this term in the API

Synonyms

- fi Beatles (n, group)
- fr Beatles (n, group)
- ja ビートルズ (n)
- ja ビートルズ (n, group)
- th วงเดอะ บีทเทิลส์ (n, group)
- ar بيتلز
- ar بيتيلز
- az bitlz
- be бітлз
- bg бийтълс
- bn দ্য বিটল্স
- bn বিটল্স
- el μπητλσ

Related terms

- en sixties (n, time)
- en beatle (n)
- en beatlehead (n)
- en beatlemania (n)
- en british
- en operate
- en quartet
- en rock music
- fr beatles
- en ex beatle (n)
- en lennon (n)
- en mccartney (n)
- es beatles (n)

Links to other sites

- dbpedia.org The_Beatles
- umbel.org Beatle
- wikidata.dbpedia.org Q1299
- sw.opencyc.org Beatle
- wordnet-rdf.princeton.edu 400520405-N
- wordnet-rdf.princeton.edu 108386847-n
- dbpedia.org The_Beatles_(No._1)
- wikidata.dbpedia.org Q738260
- dbpedia.org The_Beatles_(The_Original_Studio_Recordings)
- wikidata.dbpedia.org Q603122
- en.wiktionary.org Beatles
- fr.wiktionary.org Beatles

Derived terms

- en beatle
- en beatledom
- en beatlemania
- en beatlesque
- en ex beatle
- en fourth beatle

YAGO: A High-Quality Knowledge Base

YAGO is a large knowledge base with general knowledge about people, cities, countries, movies, and organizations.

NELL Knowledge Base Browser

CMU Read the Web Project

log in | preferences | help/instr

- religion
- mlalgorithm
- date
  - year
  - month
  - dateliteral
  - dayofweek
- medicalprocedure
- sociopolitical
- language
- profession
  - jobposition
- mlarea
- color
- creativework
  - book
  - lyrics
  - poem
  - musicalbum
  - movie
  - televisionshow
  - visualartform
  - musicsong
- politicaloffice
- sportsteamposition
- charactertrait
- time
- celltype
- physicalaction

(category) *Any algorithm or class of algorithms that is used at any st*

View list | map | metadata

3,514 instances, 1 page

instance	iteration	date learned
clustering_as_optimization	888	04-dec-2014
conditional_random_fields_and_rich_feature_sets	888	04-dec-2014
cyclic_graphical_models	888	04-dec-2014
decision_trees_and_boosting_methods	888	04-dec-2014
discriminative_random_fields	888	04-dec-2014
em_algorithm_for_maximum_likelihood_calculation	888	04-dec-2014
forward_backward_algorithm	0	28-oct-2010
generalized_linear_models	1103	18-mar-2018
generalized_regression	888	04-dec-2014
hidden_markov_models	120	19-jun-2010
hidden_markov_models_and_viterbi	888	04-dec-2014
hidden_markov_model_hybrids	888	04-dec-2014
hmms	120	19-jun-2010
in_dynamic_bayesian_networks	888	04-dec-2014
kruskal_wallis_one_way_analysis	1095	18-jan-2018
latent_semantic_analysis_and_plsa	888	04-dec-2014
likelihood_ratio	888	04-dec-2014
likelihood_ratio_test	829	10-apr-2014
linear_prediction	883	02-nov-2014
linear_regression_models	104	23-may-2010

startUS INSIGHTS

277

STARTUPS ANALYZED

5 Top Emerging Knowledge Representation & Reasoning Startups

This Global Startup Heat Map illustrates geographical distribution of 277 analyzed as well as 5 selected startups. Data from October 2020.