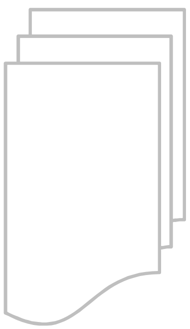
Tutorial Overview

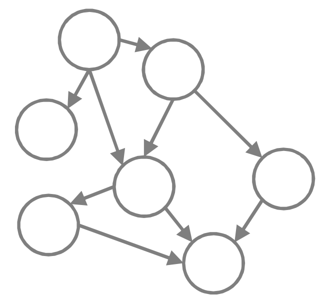
https://kgtutorial.github.io

Part 1: Knowledge Graphs



Part 2: Knowledge Extraction

Part 3: Graph Construction



Part 4: Critical Analysis

Tutorial Outline

1. **Knowledge Graph Primer** **[Jay]** 
2. Knowledge Extraction Primer [Jay] 

Coffee Break 

3. Knowledge Graph Construction
   a. Probabilistic Models [Jay] 
   b. Embedding Techniques [Sameer] 
4. Critical Overview and Conclusion [Sameer] 

Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING RESEARCH DIRECTIONS

Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING RESEARCH DIRECTIONS

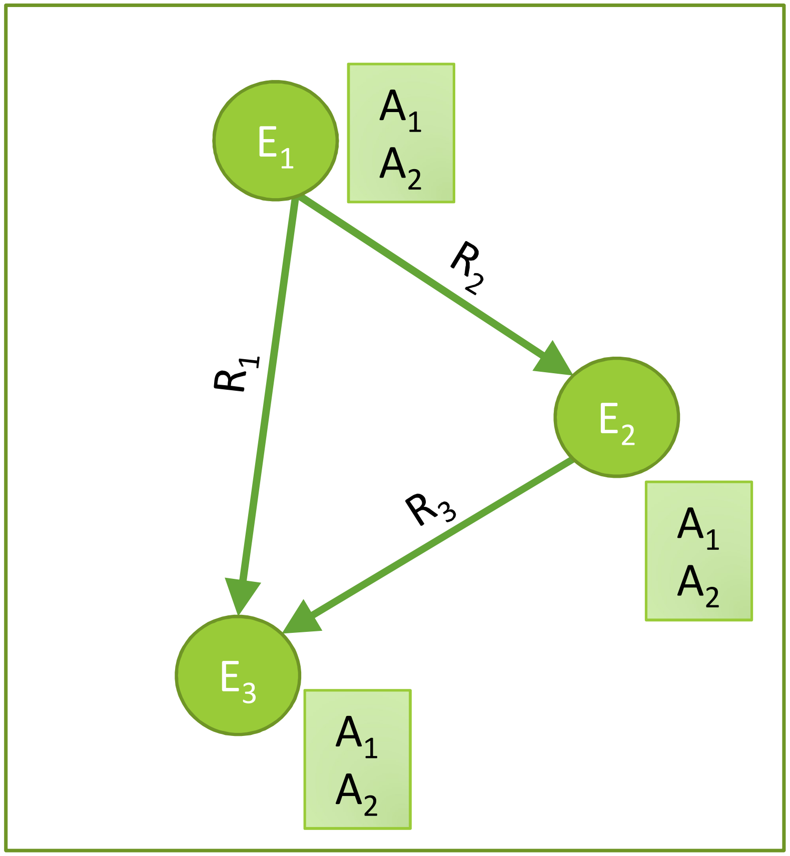
Why do we need Knowledge graphs?

- Humans can explore large database in intuitive ways

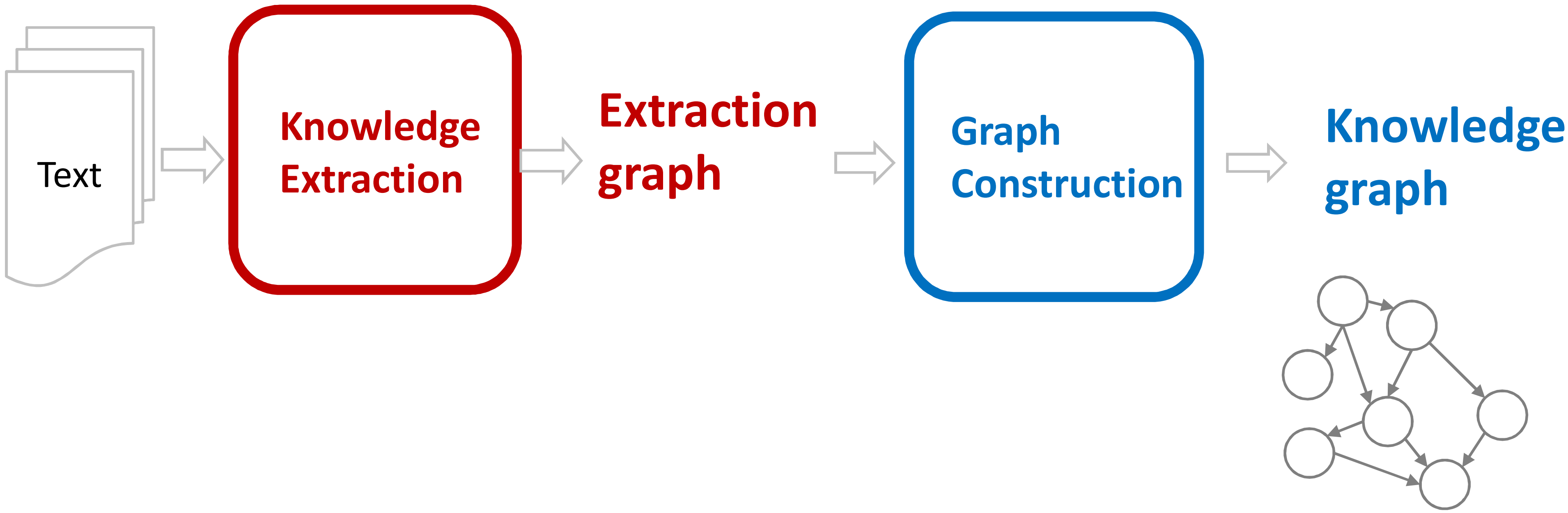
- AI agents get access to human common sense knowledge

Knowledge graph construction

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



Knowledge Graph Construction

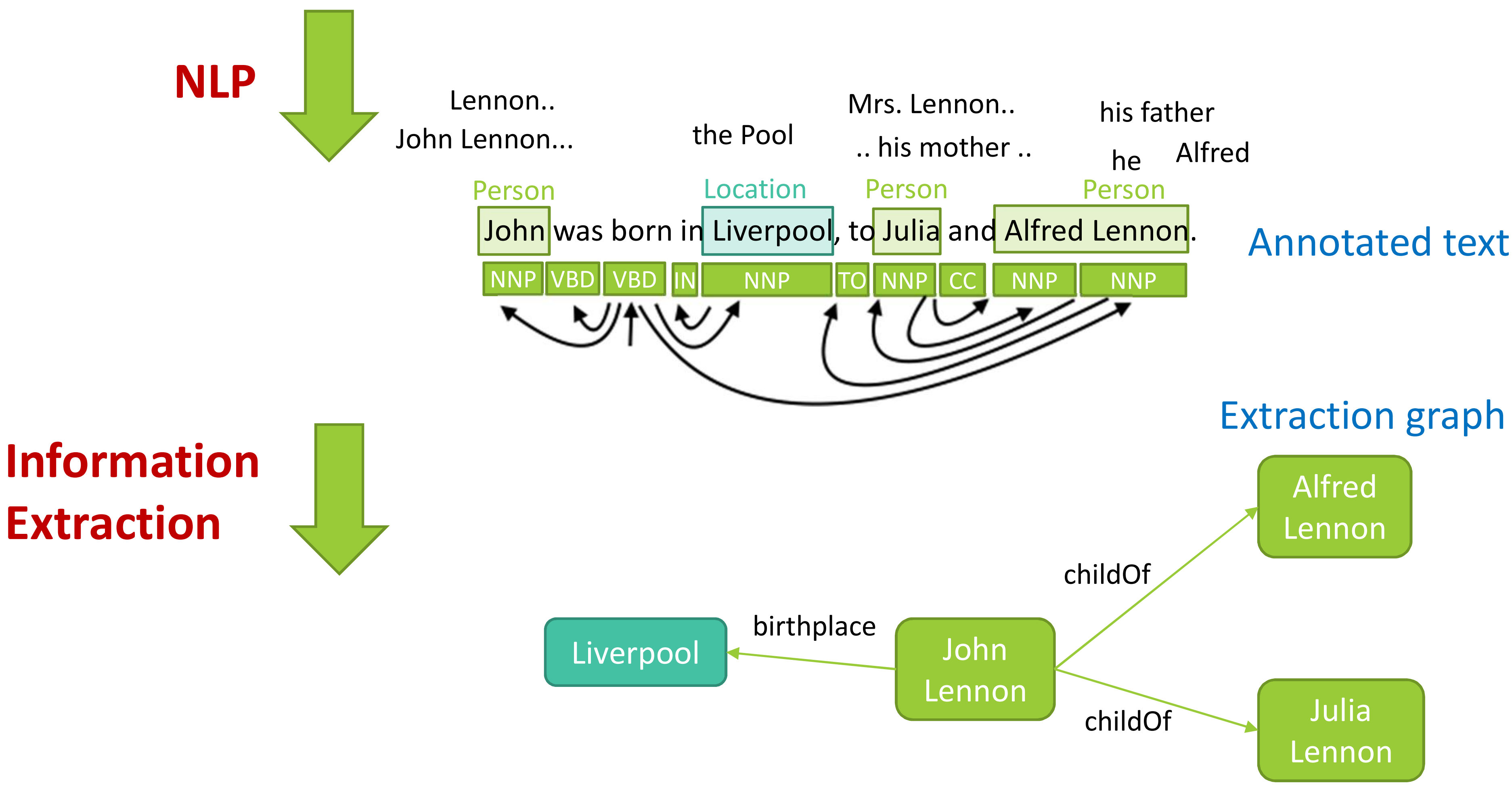


Two perspectives

	Extraction graph	Knowledge graph
Who are the entities? (nodes)	• Named Entity Recognition • Entity Coreference	• Entity Linking • Entity Resolution
What are their attributes? (labels)	• Entity Typing	• Collective classification
How are they related? (edges)	• Semantic role labeling • Relation Extraction	• Link prediction

Knowledge Extraction

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



Information Extraction

Single extractor

Defining domain

Learning extractors

Scoring candidate facts

×

Supervised 

Semi-supervised 

Unsupervised 

Fusing multiple extractors

Knowledge Graph Construction

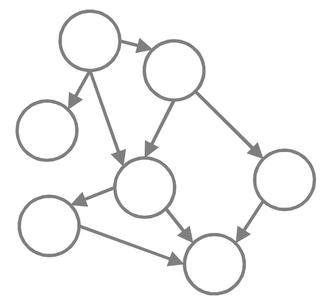
Text

Part 2: Knowledge Extraction

Extraction graph

Part 3: Graph Construction

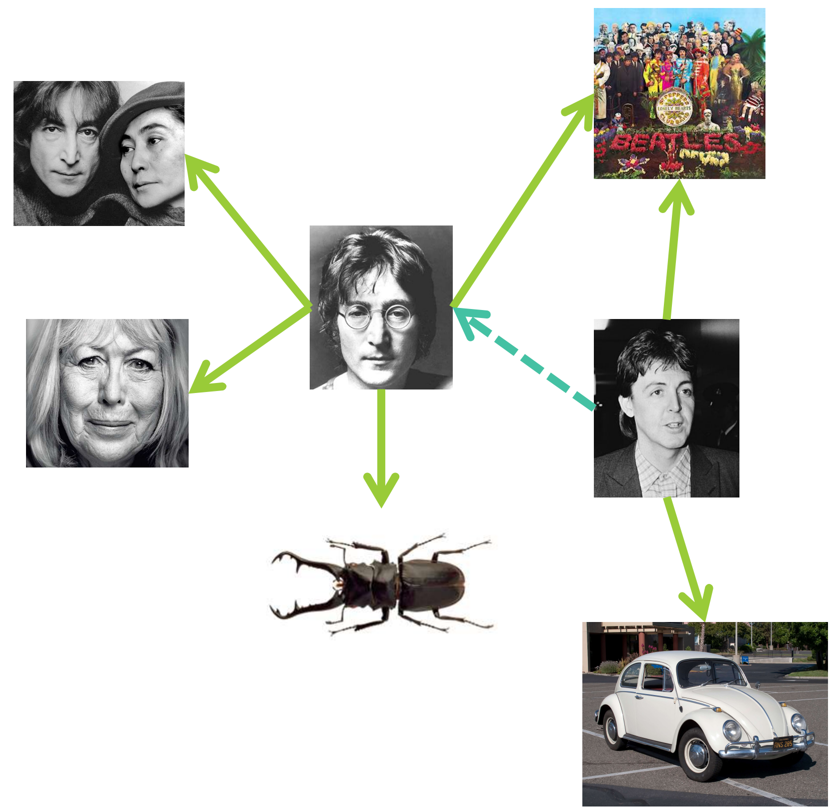
Knowledge graph



Issues with Extraction Graph

Extracted knowledge could be:

- ambiguous
- incomplete
- inconsistent



Two approaches for KG construction

PROBABILISTIC MODELS

EMBEDDING BASED MODELS

Two approaches for KG construction

PROBABILISTIC MODELS

EMBEDDING BASED MODELS

Two classes of Probabilistic Models

GRAPHICAL MODEL BASED

- Possible facts in KG are variables
- Logical rules relate facts
- Probability $\propto$ satisfied rules
- Universal-quantification

RANDOM WALK BASED

- Possible facts posed as queries
- Random walks of the KG constitute "proofs"
- Probability $\propto$ path lengths/transitions
- Local grounding

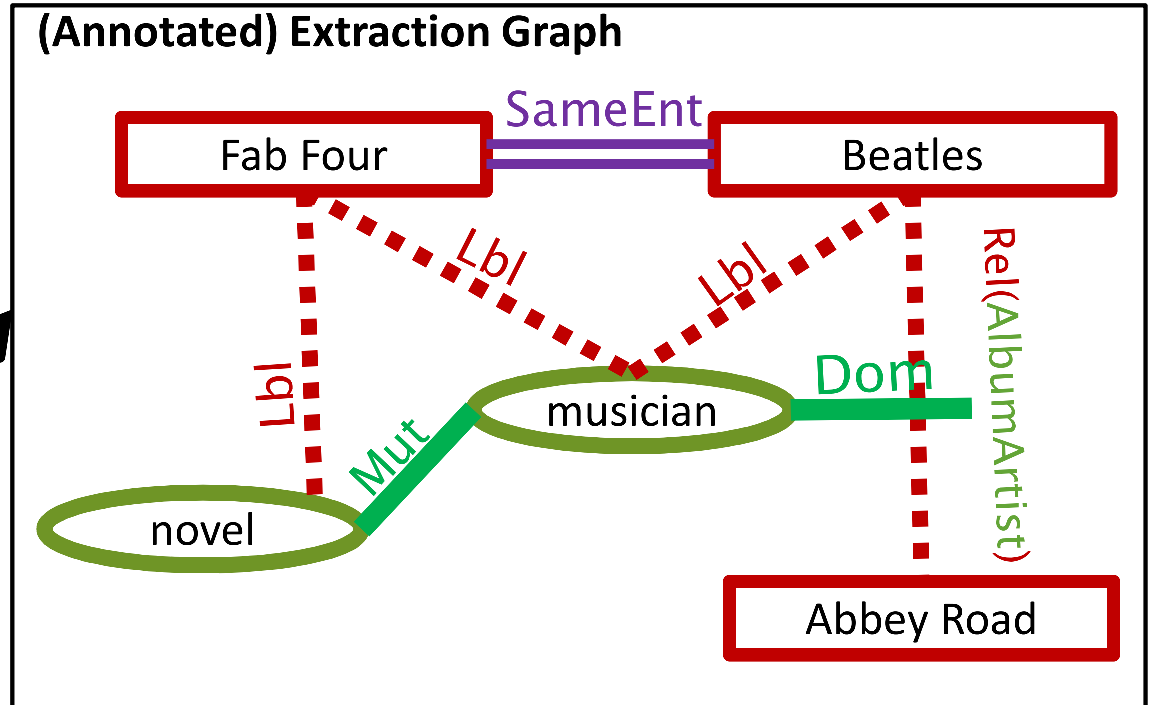
Illustration of KG Identification

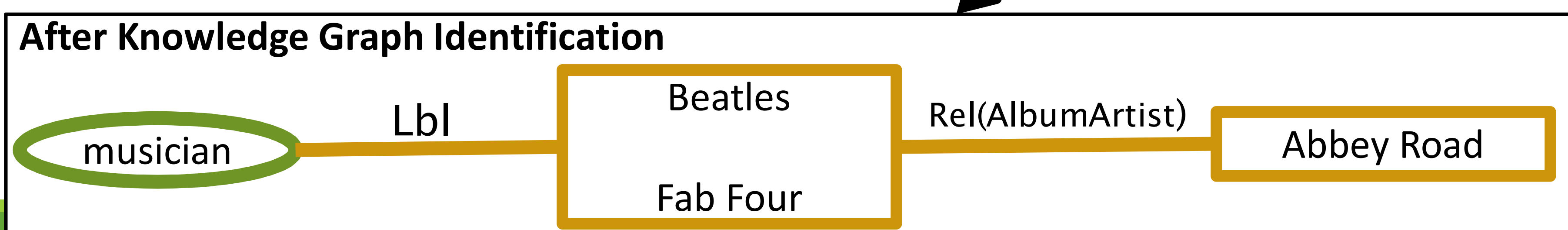
Uncertain Extractions:
.5: Lbl(Fab Four, novel)
.7: Lbl(Fab Four, musician)
.9: Lbl(Beatles, musician)

.8: Rel(Beatles,AlbumArtist,
Abbey Road)

Ontology:
Dom(albumArtist, musician)
Mut(novel, musician)

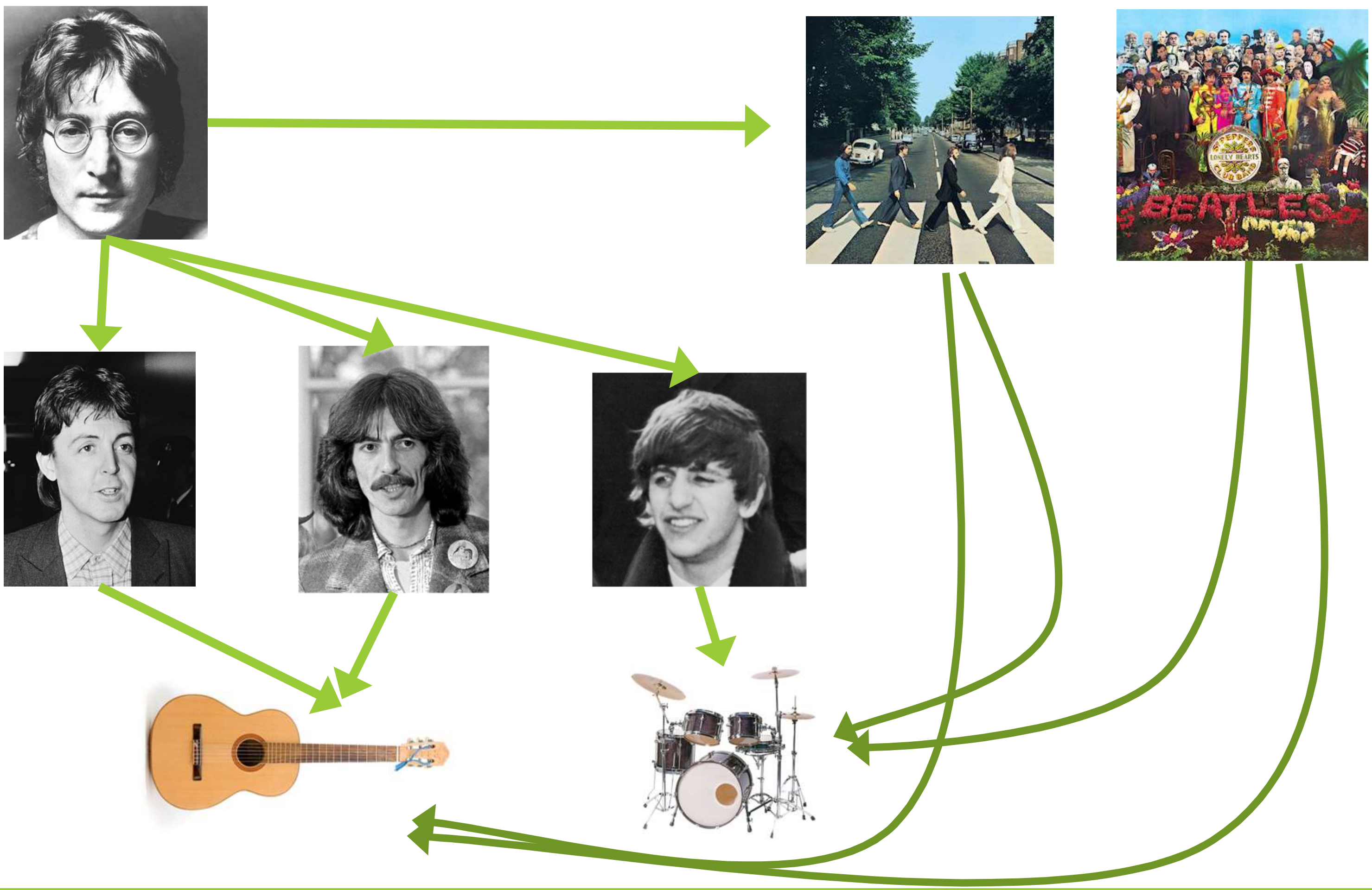
Entity Resolution:
SameEnt(Fab Four, Beatles)





Random Walk Illustration

Query: R(Lennon, PlaysInstrument, ?)



Two approaches for KG construction

PROBABILISTIC MODELS

EMBEDDING BASED MODELS

Why embeddings?

Limitations of probabilistic models

Limitation to Logical Relations

- Representation restricted by manual design
  - Clustering? Asymmetric implications?
  - Information flows through these relations
- Difficult to generalize to unseen entities/relations

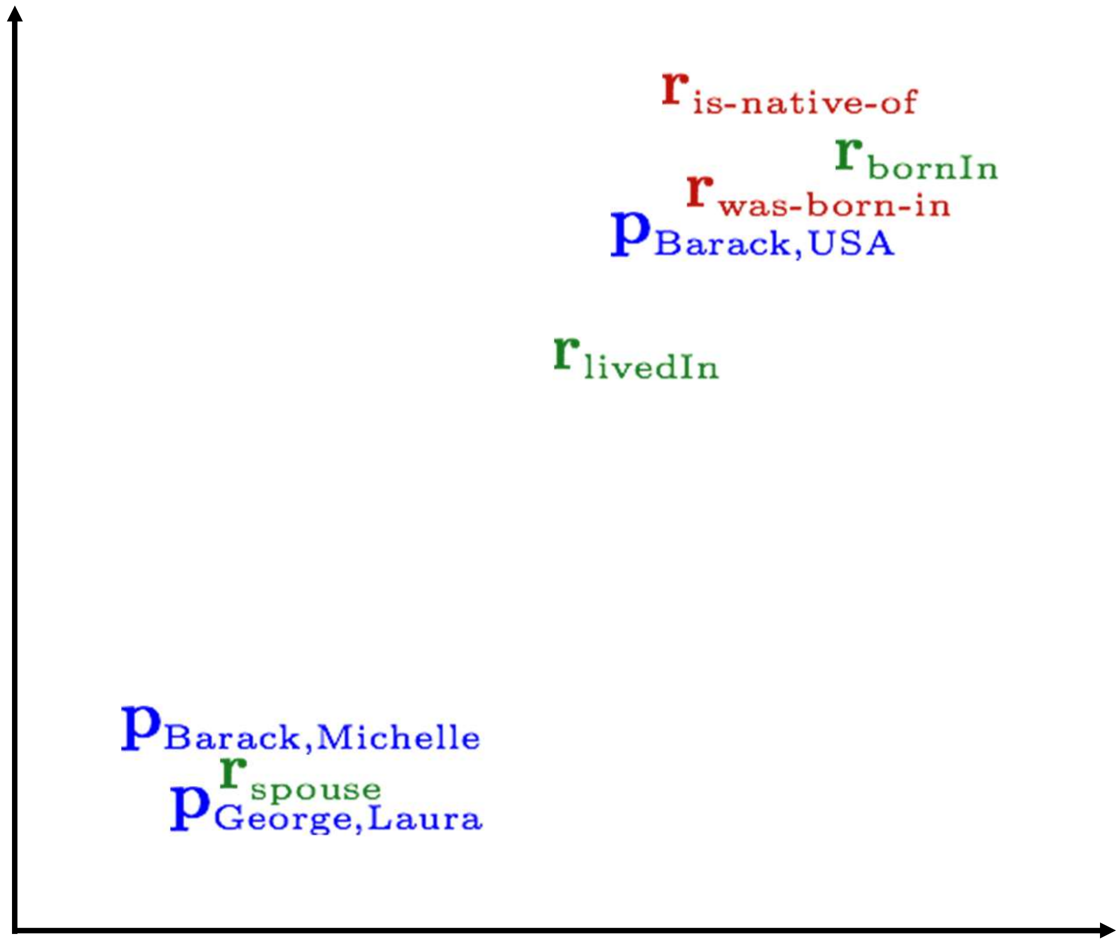
Computational Complexity of Algorithms

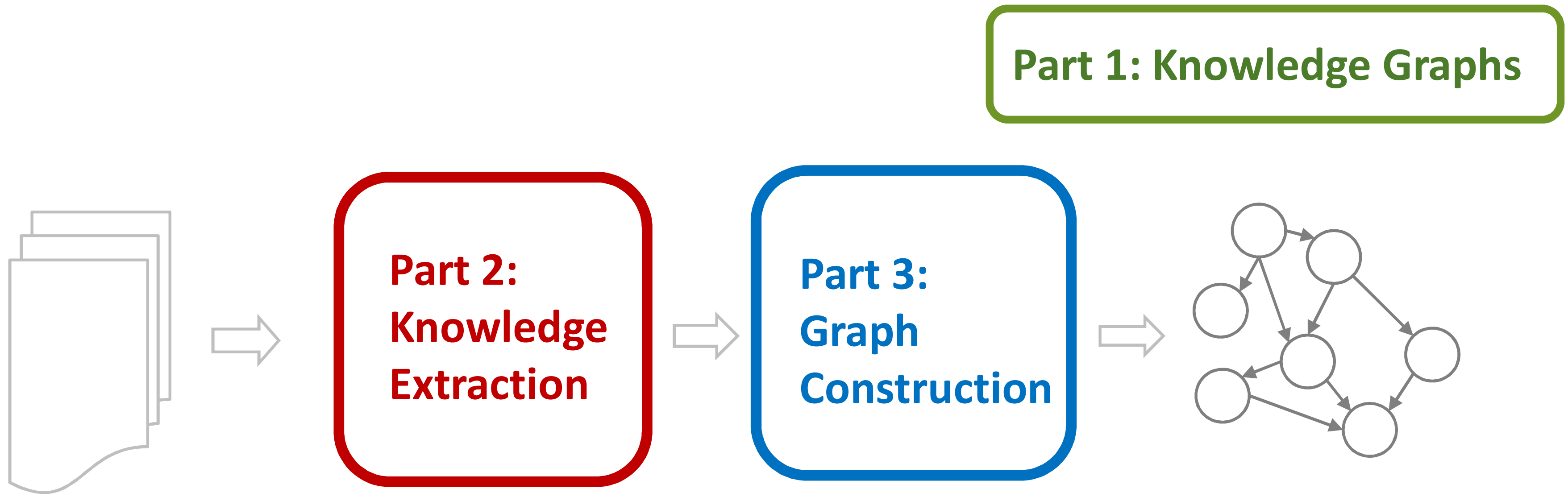
- Learning is NP-Hard, difficult to approximate
- Query-time inference is also NP-Hard
- Not easy to parallelize, or use GPUs
- Scalability is badly affected by representation

Embedding based models

- Can generalize to unseen entities and relations
- Efficient inference at large scale

Relation Embeddings




Part 1: Knowledge Graphs
Part 2: Knowledge Extraction
Part 3: Graph Construction

Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING RESEARCH DIRECTIONS

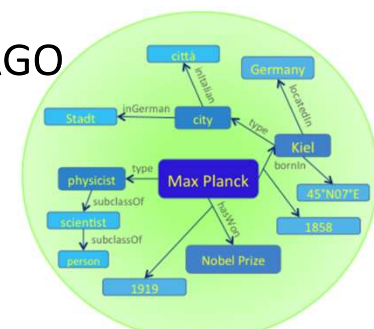
Success stories

Open Information Extraction

NELL Knowledge Base Browser
CMU Read the Web Project

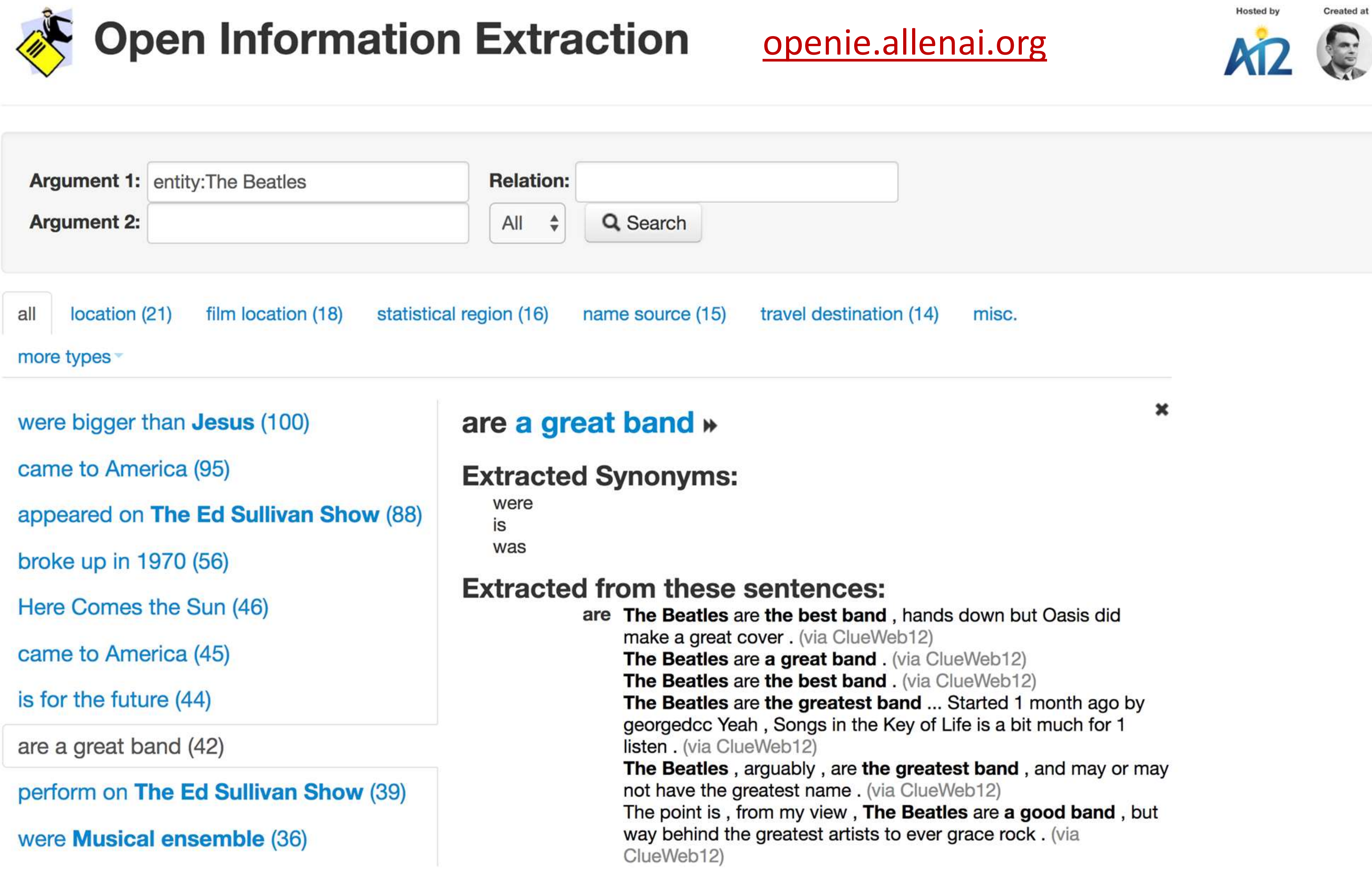
ConceptNet
An open, multilingual knowledge graph

YAGO

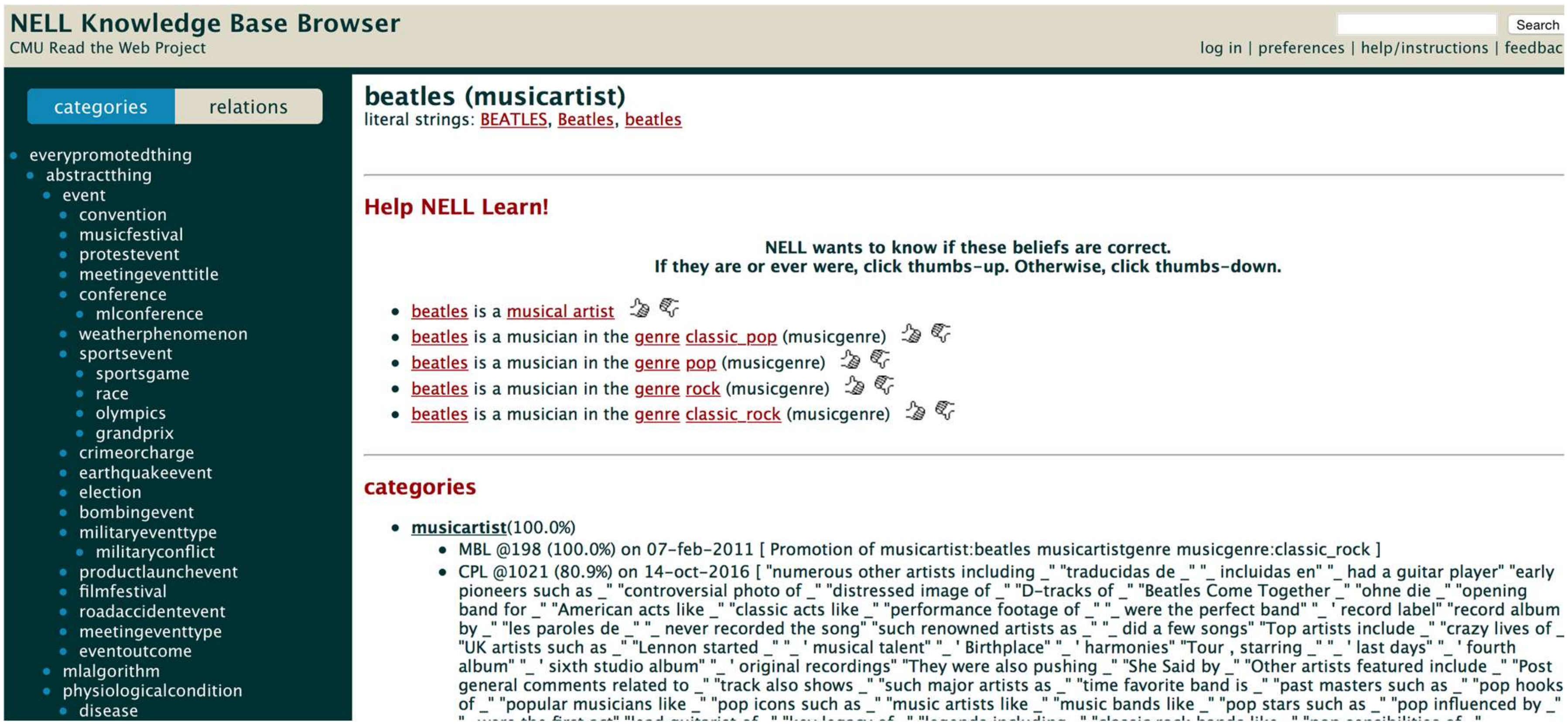


DeepDive v0.8.0
Think about **features**, not **algorithms**.

Success story: OpenIE (ReVerb)

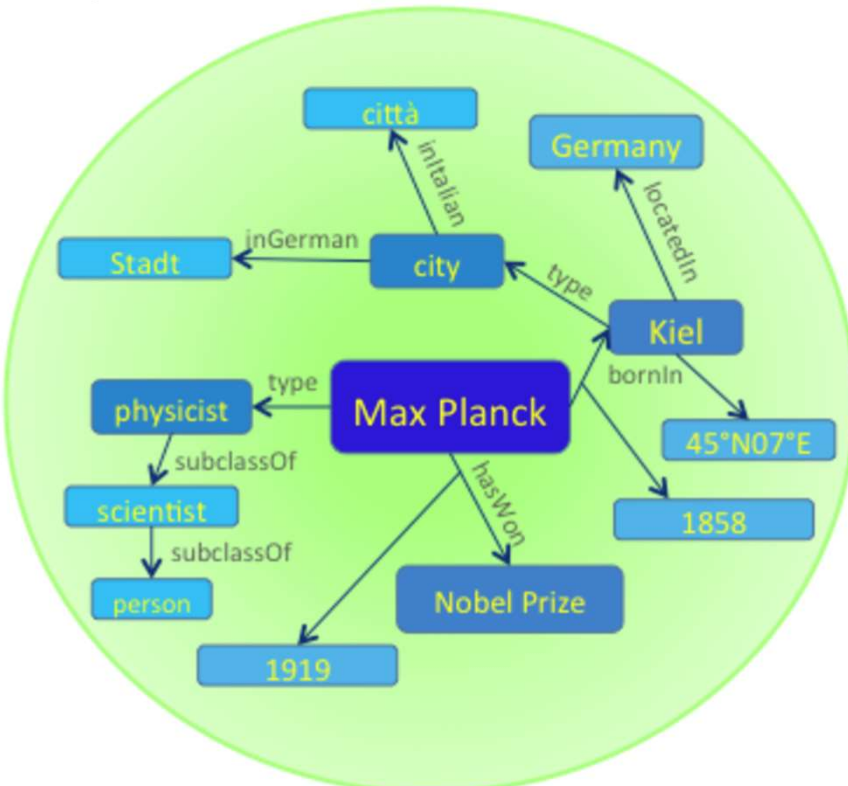


Success story: NELL



Success story: YAGO

- **Input:** Wikipedia infoboxes, WordNet and GeoNames
- **Output:** KG with 350K entity types, 10M entities, 120M facts
- Temporal and spatial information



ConceptNet

An open, multilingual knowledge graph

[Link]

en beatles

An English term in ConceptNet 5.5

Derived terms

- en beatle →
- en beatledom →
- en beatlemania →
- en beatlesque →
- en fourth beatle →

beatles is a type of…

- en a British band →
- en man (n) →
- en band (n) →
- en musician (n) →
- en album (n) →

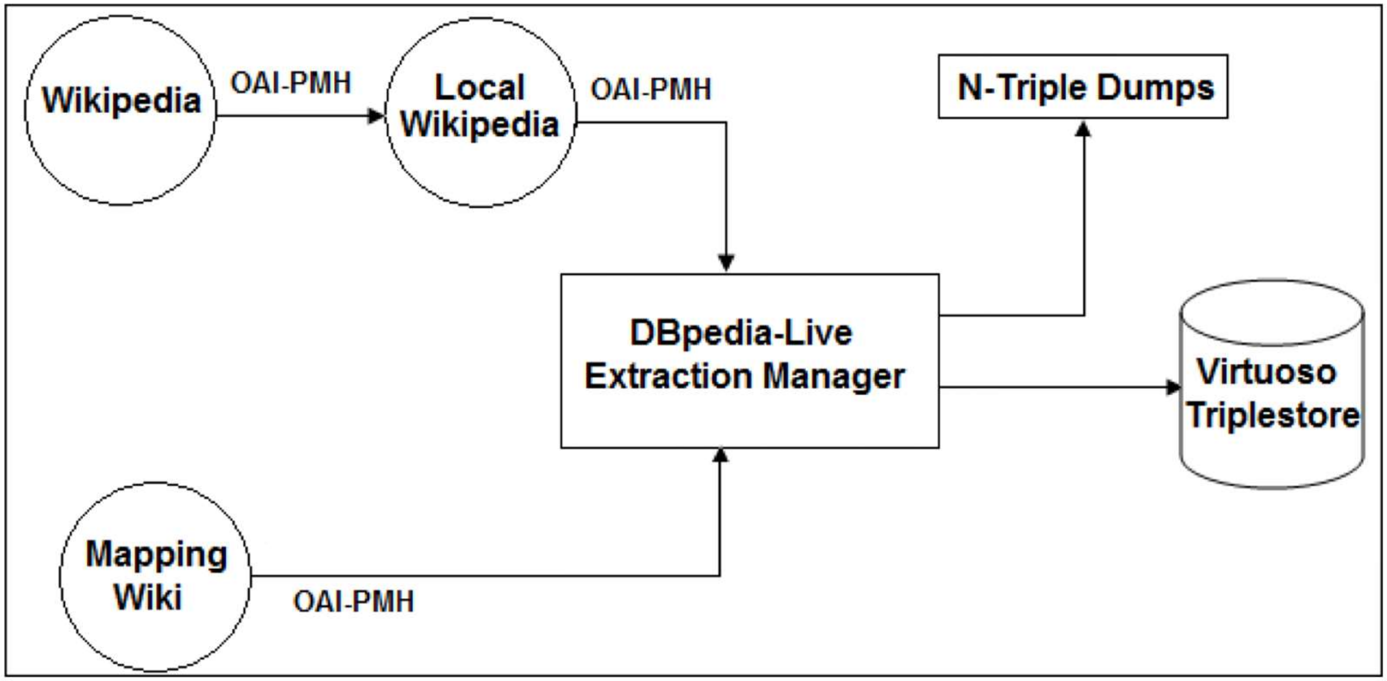
Links to other sites

- dbpedia.org The Beatles →
- sw.opencyc.org Beatle →
- umbel.org Beatle →
- wordnet-rdf.princeton.edu 400520405-N →
- wordnet-rdf.princeton.edu 108386847-n →
- wikidata.dbpedia.org Q1299 →
- en.wiktionary.org Beatles →
- dbpedia.org The Beatles (No. 1) →
- wikidata.dbpedia.org Q738260 →
- fr.wiktionary.org Beatles →
- dbpedia.org The Beatles (The Original Studio Recordings) →
- wikidata.dbpedia.org Q603122 →

Success story

Towards a Public Data Infrastructure for a Large, Multilingual, Semantic Knowledge Graph

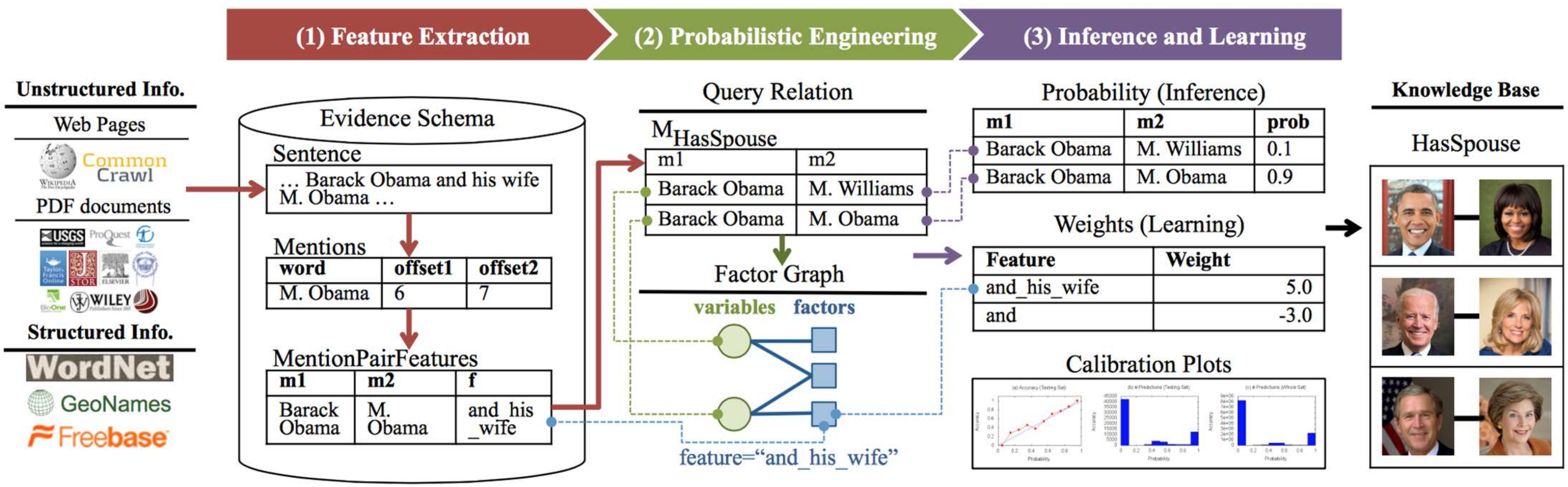
- DBPedia is automatically extracted structured data from Wikipedia
  - 17M canonical entities
  - 88M type statements
  - 72M infobox statements



DeepDive

DeepDive v0.8.0

Think about **features**, not **algorithms**.



- Best Precision/recall/F1 in KBP-slot filling task 2014 evaluations (31 teams participated)

IE systems in practice

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

Critical Overview

SUMMARY

SUCCESS STORIES

<u>DATASETS, TASKS, SOFTWARES</u>

EXCITING RESEARCH DIRECTIONS

Datasets

- KG as datasets
  - FB15K-237 Knowledge base completion dataset based on Freebase[1]
  - DBPedia Structured data extracted from Wikipedia
  - NELL Read the web datasets
  - AristoKB Tuple knowledge base for Science domain

- Text datasets
  - Clueweb09: 1 billion webpages (sample of Web)
  - FACC1: Freebase Annotations of the Clueweb09 Corpora
  - Gigaword: automatically-generated syntactic and discourse structure
  - NYTimes: The New York Times Annotated Corpus

- Datasets related to Semi-supervised learning for information extraction
  Link: entity typing, concept discovery, aligning glosses to KB, multi-view learning

[1]see Dettmers et al, 2017 for details (https://arxiv.org/pdf/1707.01476.pdf)

Shared tasks

- Text Analysis Conference on Knowledge Base Population (TAC KBP)
  - **Slot filling task**
  - **Cold Start KBP Track**
  - **Tri-Lingual Entity Discovery and Linking Track (EDL)**
  - **Event Track**
  - **Validation/Ensembling Track**

Software: NLP

- Stanford CoreNLP: a suite of core NLP tools [link] (Java code)
- FIGER: fine-grained entity recognizer assigns over 100 semantic types
  link (Java code)
- FACTORIE: out-of-the-box tools for NLP and information integration
  link (Scala code)
- EasySRL: Semantic role labeling
  link (Java code)

Software: Extracting and Reasoning

- **Open IE**
  (University of Washington)
  Open IE 4.2 link (Scala code)
  Stanford Open IE link (Java code)

- **Interactive Knowledge Extraction** (IKE)
  (Allen Institute for Artificial Intelligence)
  link (Scala code)

- **PSL:** Probabilistic soft logic
  link (Java code)

- **ProPPR:** Programming with Personalized PageRank
  link (Java code)

Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING RESEARCH DIRECTIONS

Exciting Active Research

- INTERESTING APPLICATIONS OF KG
- MULTI-MODAL INFORMATION EXTRACTION
- KNOWLEDGE AS SUPERVISION
- COMMON KNOWLEDGE

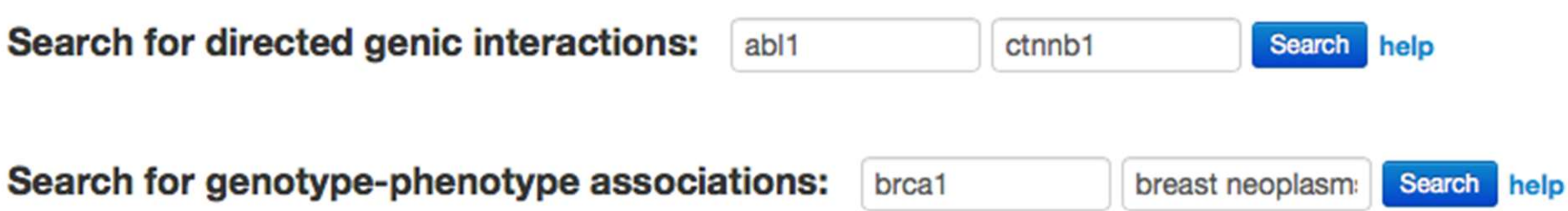
Exciting Active Research

- INTERESTING APPLICATIONS OF KG
- MULTI-MODAL INFORMATION EXTRACTION
- KNOWLEDGE AS SUPERVISION
- COMMON KNOWLEDGE

Interesting application of Knowledge Graphs

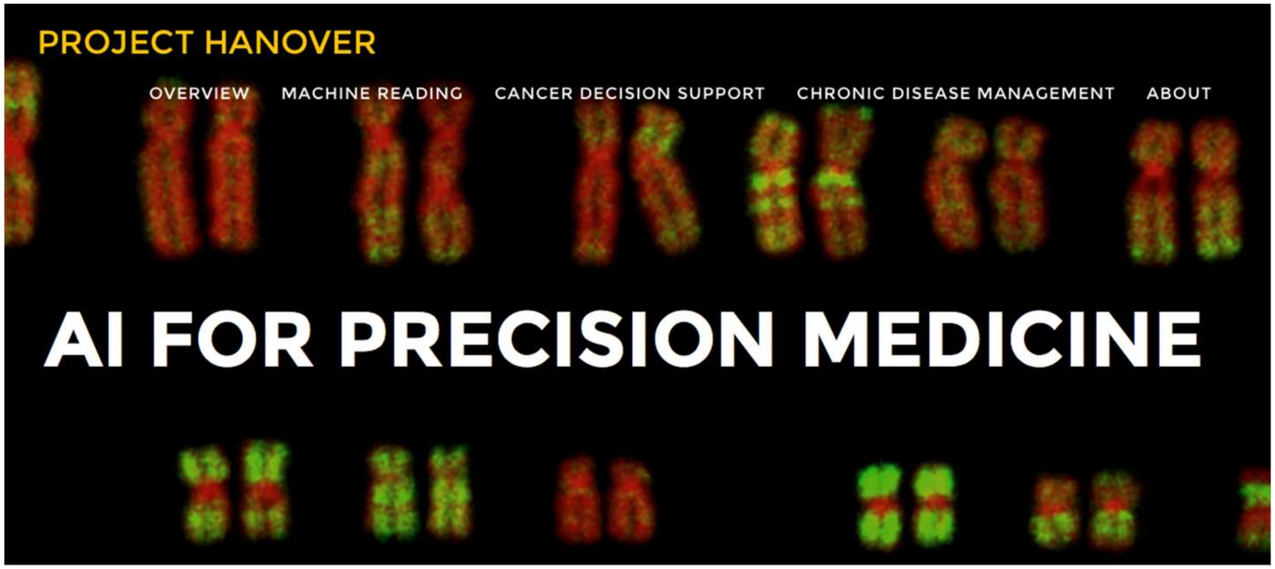
The Literome Project [link]

- Automatic system for extracting genomic knowledge from PubMed articles
- Web-accessible knowledge base



***Literome: PubMed-Scale Genomic Knowledge Base in the Cloud**, Hoifung Poon et al., Bioinformatics 2014*

Interesting application of Knowledge Graphs



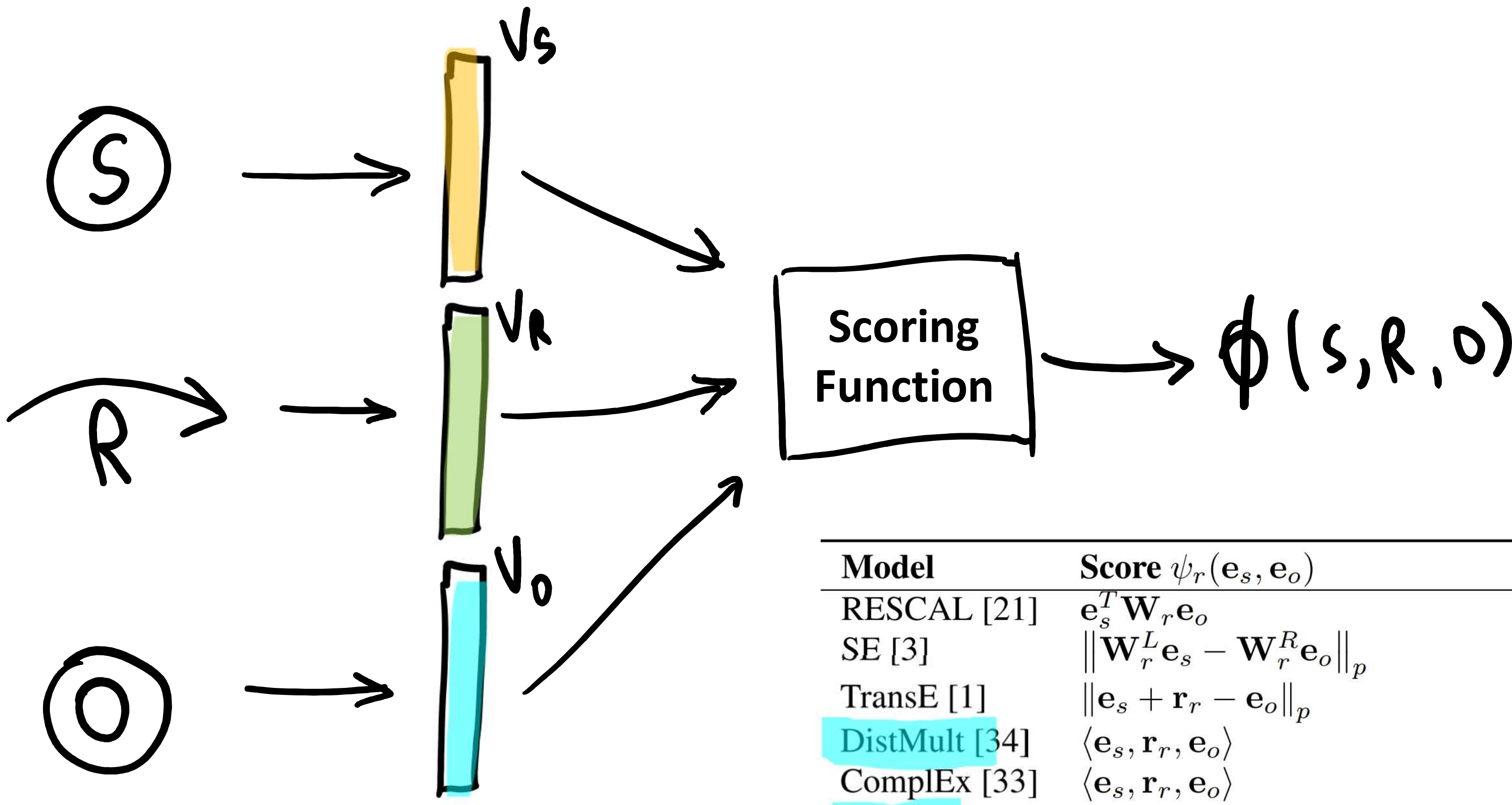
Microsoft Research

Chronic disease management:
develop AI technology for predictive and preventive personalized medicine to reduce the national healthcare expenditure on chronic diseases (90% of total cost)

Exciting Active Research

- INTERESTING APPLICATIONS OF KG
- MULTI-MODAL INFORMATION EXTRACTION
- KNOWLEDGE AS SUPERVISION
- COMMON KNOWLEDGE

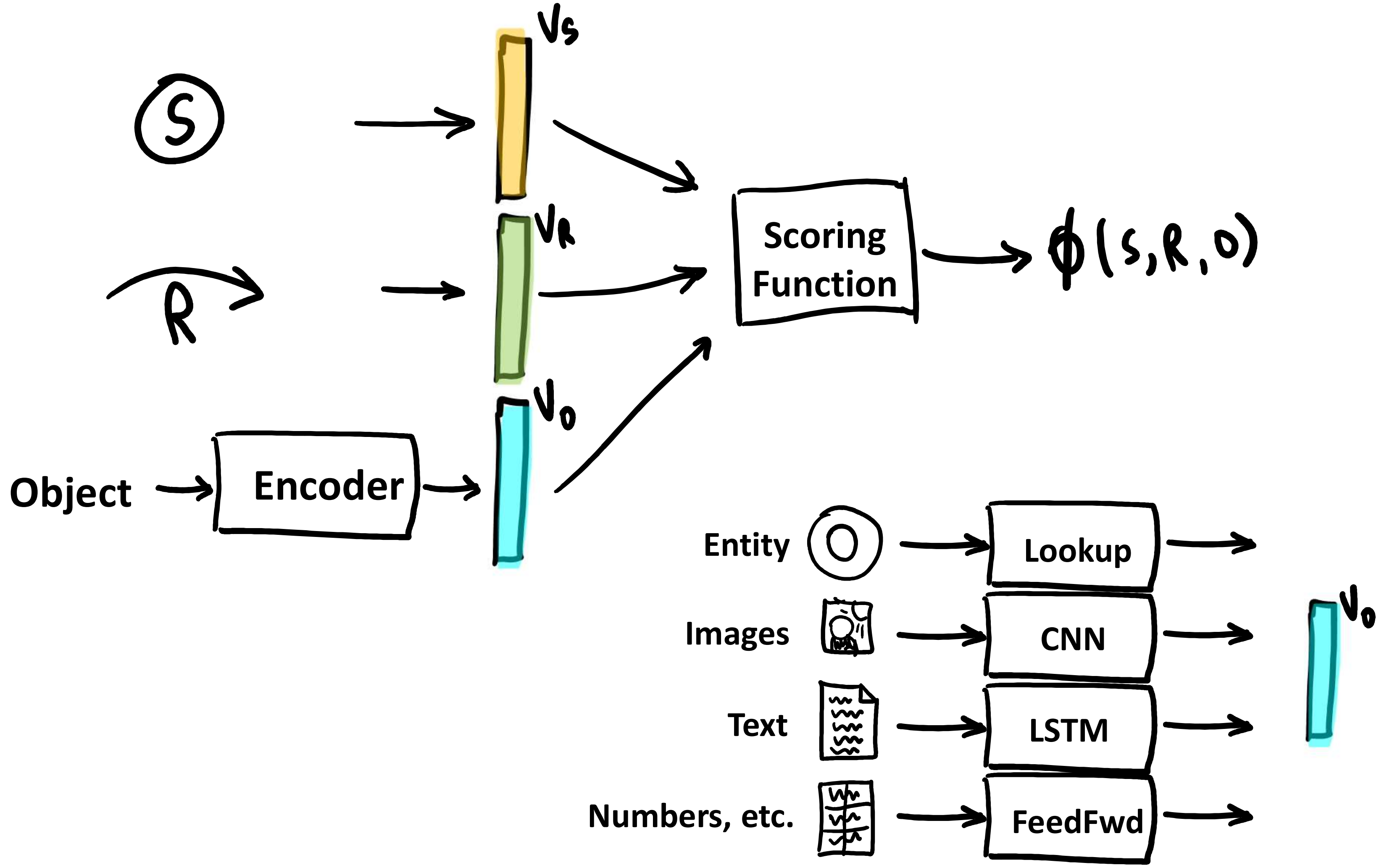
Knowledge Base Completion



Model	Score $\psi_r(\mathbf{e}_s, \mathbf{e}_o)$
RESCAL [21]	$\mathbf{e}_s^T \mathbf{W}_r \mathbf{e}_o$
SE [3]	$\left\|\mathbf{W}_r^L \mathbf{e}_s - \mathbf{W}_r^R \mathbf{e}_o\right\|_p$
TransE [1]	$\|\mathbf{e}_s + \mathbf{r}_r - \mathbf{e}_o\|_p$
DistMult [34]	$\langle \mathbf{e}_s, \mathbf{r}_r, \mathbf{e}_o \rangle$
ComplEx [33]	$\langle \mathbf{e}_s, \mathbf{r}_r, \mathbf{e}_o \rangle$
ConvE	$f\left(\mathrm{vec}\left(f\left([\overline{\mathbf{e}_s}; \overline{\mathbf{r}_r}] * \omega\right)\right) \mathbf{W}\right) \mathbf{e}_o$

Table from Dettmers, et al. (2017)

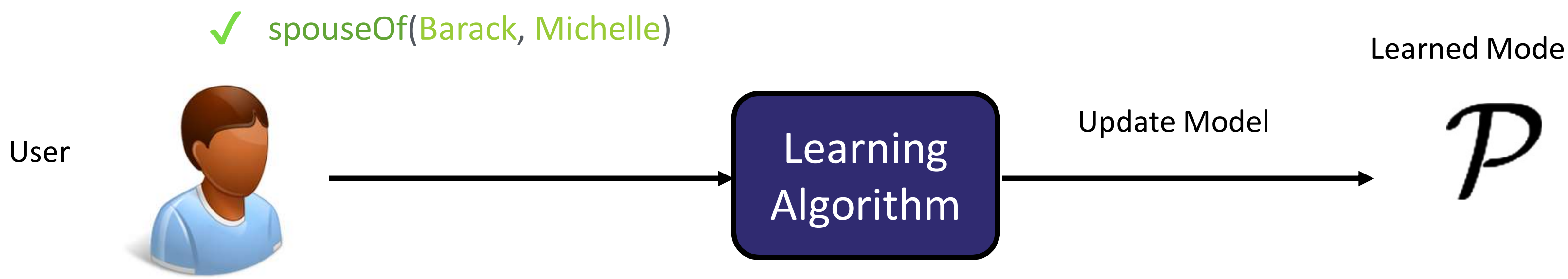
Multimodal KB Embeddings



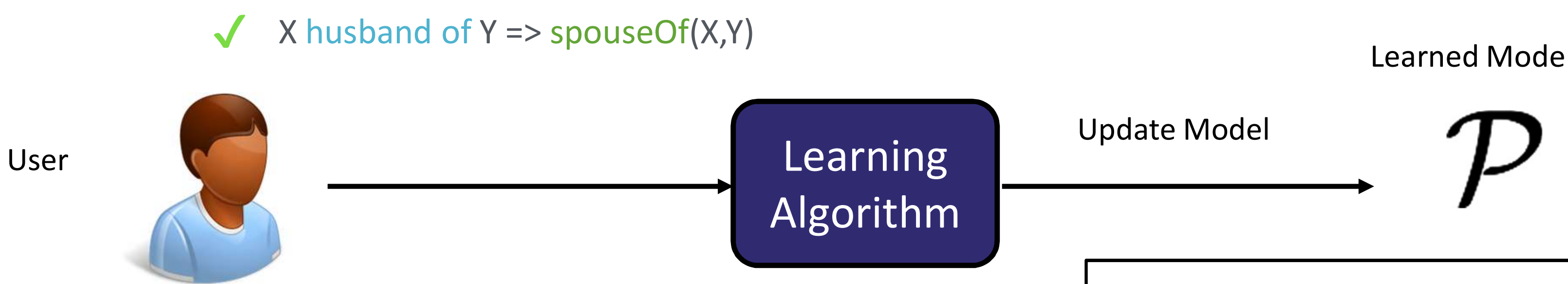
Exciting Active Research

- INTERESTING APPLICATIONS OF KG
- MULTI-MODAL INFORMATION EXTRACTION
- KNOWLEDGE AS SUPERVISION
- COMMON KNOWLEDGE

Knowledge as Supervision



Problem 1: Each annotation takes time
Problem 2: Each annotation is a drop in the ocean



Many different options

- Generalized Expectation
- Posterior Regularization
- Labeling functions in SNORKEL

Exciting Active Research

- INTERESTING APPLICATIONS OF KG
- MULTI-MODAL INFORMATION EXTRACTION
- KNOWLEDGE AS SUPERVISION
- <u>COMMON KNOWLEDGE</u>

Aristo Science QA challenge

- Science questions dataset
  ~5K 4-way multiple choice questions

  Frogs lay eggs that develop into tadpoles and then into adult frogs. This sequence of changes is an example of how living things ______

  (A) go through a life cycle

  (B) form a food web

  (C) act as a source of food

  (D) affect other parts of the ecosystem

Science knowledge

frog's life cycle, metamorphosis

+

Common sense knowledge

frog is an animal, animals have life cycle

Future……

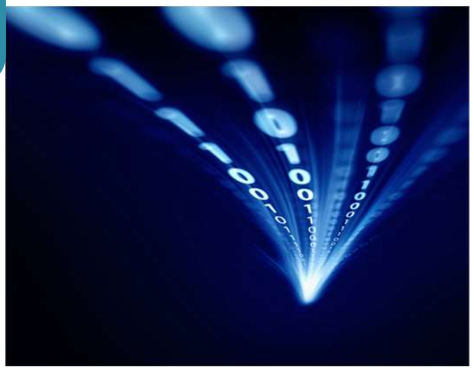
Future KG construction system

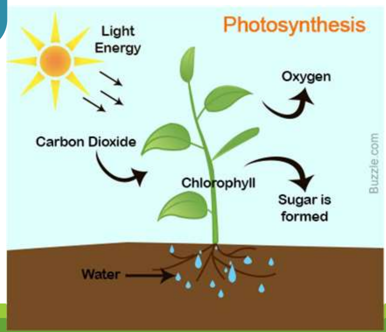
Consume online streams of data

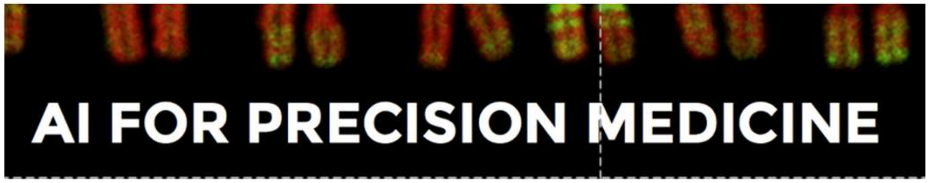
Represent context beyond facts

Supports humanity

Corrects its own mistakes







Thank You

Jay Pujara
jaypujara.org
jay@cs.umd.edu
@jay_mlr

Sameer Singh
sameersingh.org
sameer@uci.edu
@sameer_

Two perspectives

	Extraction graph	Knowledge graph
Who are the entities? (nodes)		
What are their attributes? (labels)		
How are they related? (edges)		

Natural Language Processing

Document

Within-doc Coreference...

Lennon..
John Lennon...

the Pool

Mrs. Lennon..
.. his mother ..

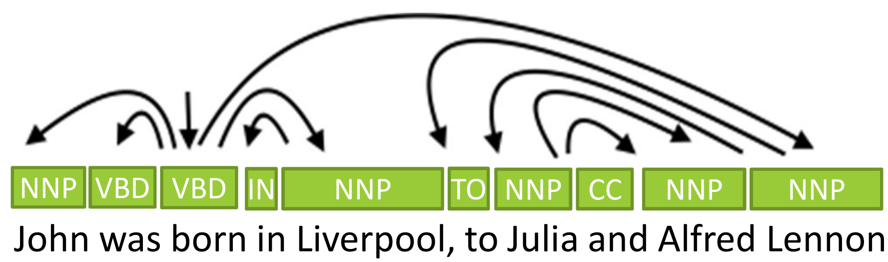
his father
he Alfred

Person | Location | Person | Person

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

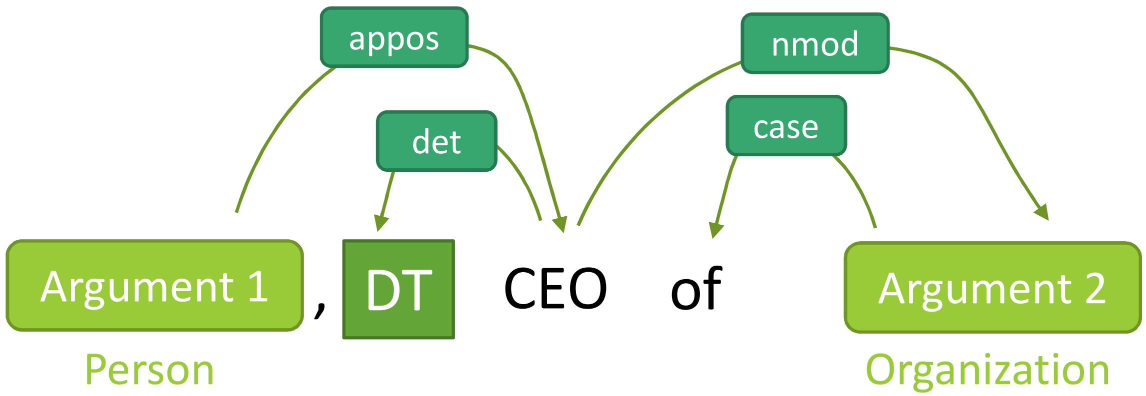
Sentence

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



NLP annotations → features for IE

Combine tokens, dependency paths, and entity types to define rules.



Bill Gates, the CEO of Microsoft, said …
Mr. Jobs, the brilliant and charming CEO of Apple Inc., said …
… announced by Steve Jobs, the CEO of Apple.
… announced by Bill Gates, the director and CEO of Microsoft.
… mused Bill, a former CEO of Microsoft.
and many other possible instantiations…

Success story: OpenIE

- **Key contributions:**
  - No need for human defined relation schemas
  - First ever successful open-source open domain IE system

- **ReVerb**
  - Input = Clueweb09 corpus (1B web pages)
  - Output = 15M high-precision extractions

Open IE Systems

2007	2010	2012	2014	2016
OpenIE v 1.0 TextRunner	v 2.0 ReVerb	v 3.0 OLLIE	OpenIE 4.0	OpenIE 5.0
CRF Self-training	POS-tag based relation extraction	Dependency parse based extraction	SRL-based extraction; temporal, spatial extractions	Supports compound noun phrases; numbers; lists

Increase in precision, recall, expressiveness

Success story: NELL

- **Key technical contributions:**
  - “Never ending learning” paradigm
  - “Coupled bootstrap learning” to reduce semantic drift

- Input: Clueweb09 corpus (1B web pages)
- Ontology: ~2K predicates
  O(100K)constraints between predicates
- Output: 50 million candidate facts
  3 million high-confidence facts

Success story: YAGO

- **Key contributions:**
  - **Rich Ontology:** Linking Wikipedia categories to WordNet
  - **High Quality:** High precision extractions (~95%)

Success story: ConceptNet

- Commonsense knowledge base
- **Key contributions:**
  - **Freely available resource:** covers wide range of common sense concepts and relations organized in a easy-to-use semantic network
  - **NLP toolkit based on this resource:** supports analogy, text summarization, context dependent inferences
- ConceptNet4 was manually built using inputs from thousands of people
  - 28 million facts expressed in natural language
  - spanning 304 different languages

DeepDive

DeepDive v0.8.0
Think about **features**, not **algorithms**.

- Machine learning based extraction system
- Key contributions:
  - **scalable, high-performance inference and learning engine**
  - **Developers contribute features (rules) not algorithms**
  - **Combines data from variety of sources (webpages, pdf, figures, tables)**

Future……

Aristo ScienceKB

- AI2's TupleKB dataset: link
- **Open problems**
  - Best KR for Science domain
  - Domain targeted KB completion
  - Measuring recall w.r.t. end task

(1) Future research directions:
Going beyond facts

- Most of the existing KGs are designed to represent and extract binary relations → good enough for search engines

- Applications like QA demand in depth knowledge about higher level structures like activities, events, processes

(2) Future research directions:
Online KG Construction

- One shot KG construction → Online KG construction
  - Consume online stream of data
  - Temporal scoping of facts
  - Discovering new concepts automatically
  - Self-correcting systems

(2) Future research directions:
Online KG Construction

- **Continuously learning and self-correcting systems**
  - *[Selecting Actions for Resource-bounded Information Extraction using Reinforcement Learning, Kanani and McCallum, WSDM 2012]*
    - Presented a reinforcement learning framework for budget constrained information extraction
  - *[Never-Ending Learning, Mitchell et al. AAAI 2015]*
    - Tom Mitchell says "Self reflection and an explicit agenda of learning subgoals" is an important direction of future research for continuously learning systems.

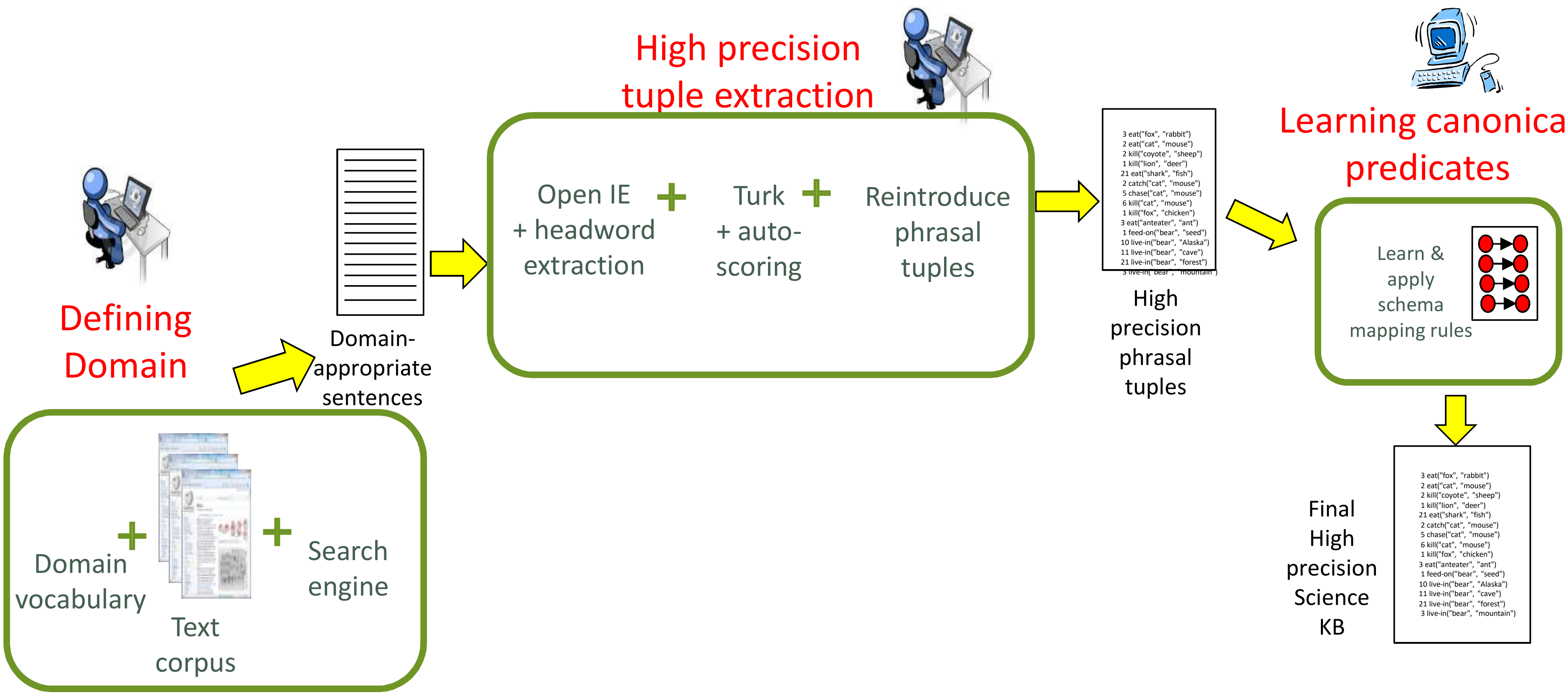
AI2's ScienceKB

Existing knowledge graphs

- Too named entity centric (no domain relevance)
- Too noisy (not directly usable by inference systems)

**Upcoming article on ``High Precision Knowledge Extraction for Science domain''

AI2's ScienceKB



**Upcoming article on ``High Precision Knowledge Extraction for Science domain''

AI2's ScienceKB

Hybrid Approach: Adding structure to Open domain IE

	Defining domain	Learning extractors	Scoring candidate facts
Open domain IE			

+

Distant supervision to add structure

AI2's TupleKB dataset: link
> 300K common-sense and science facts
> 80% precision

**Upcoming article on ``High Precision Knowledge Extraction for Science domain''

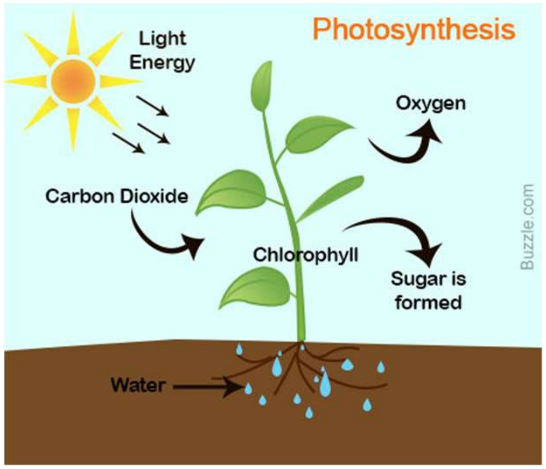
Future research directions:
Going beyond facts

- **Fact:** Individual knowledge tuples
  (plant, take in, CO_2)

- **Event frame:**
  more context how, when, where?

subject	plant
predicate	Take in
object	CO2
time	daytime

- **Processes:**
  representing larger structures, sequence of events
  e.g. Photosynthesis



[*Modeling Biological Processes for Reading Comprehension, Berant et al., EMNLP 2014*]

(3) Exciting active research:
Ambitious Project

ALLEN INSTITUTE
for ARTIFICIAL INTELLIGENCE

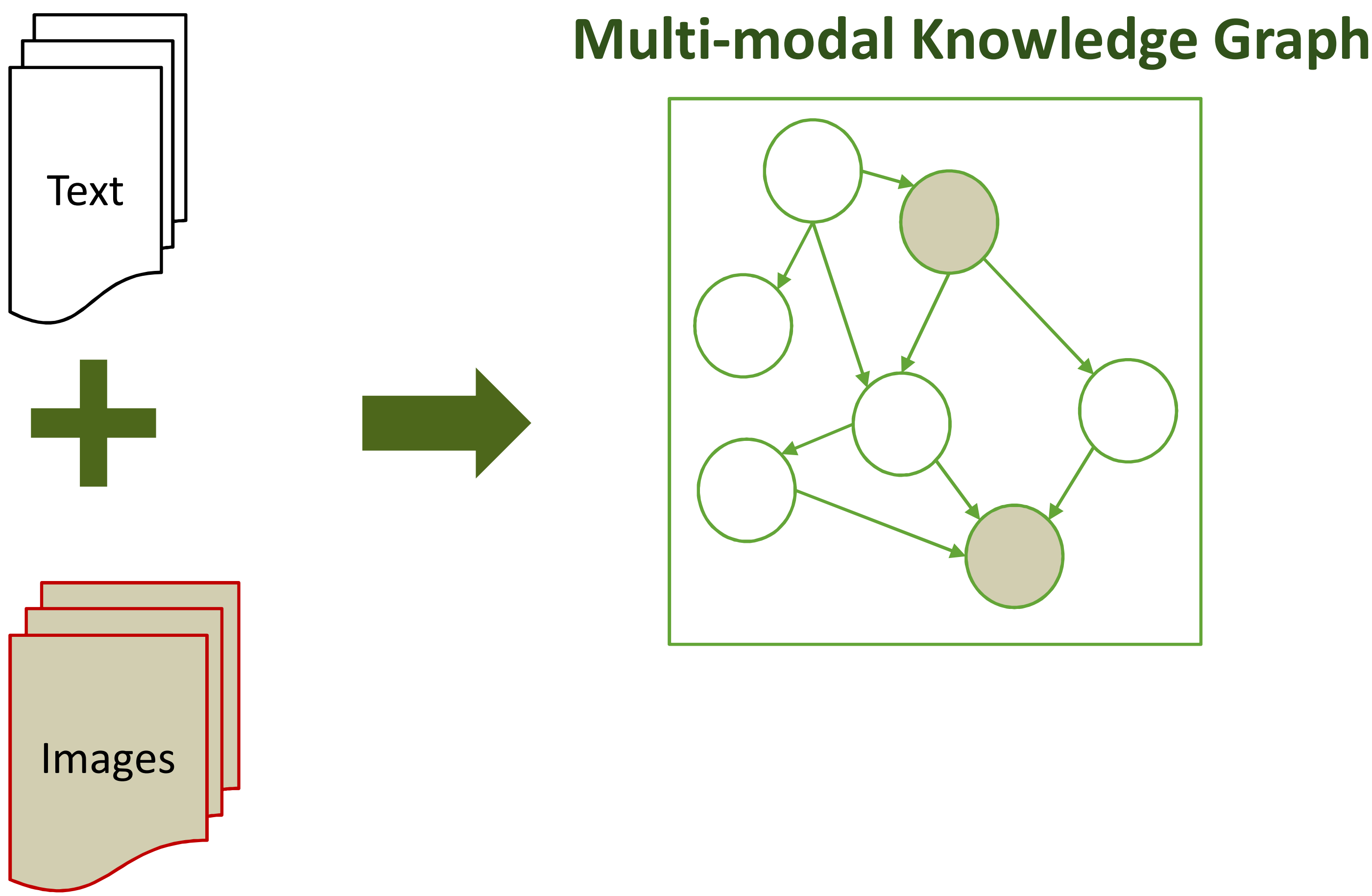
The Allen AI Science Challenge

Is your model smarter than an 8th grader?

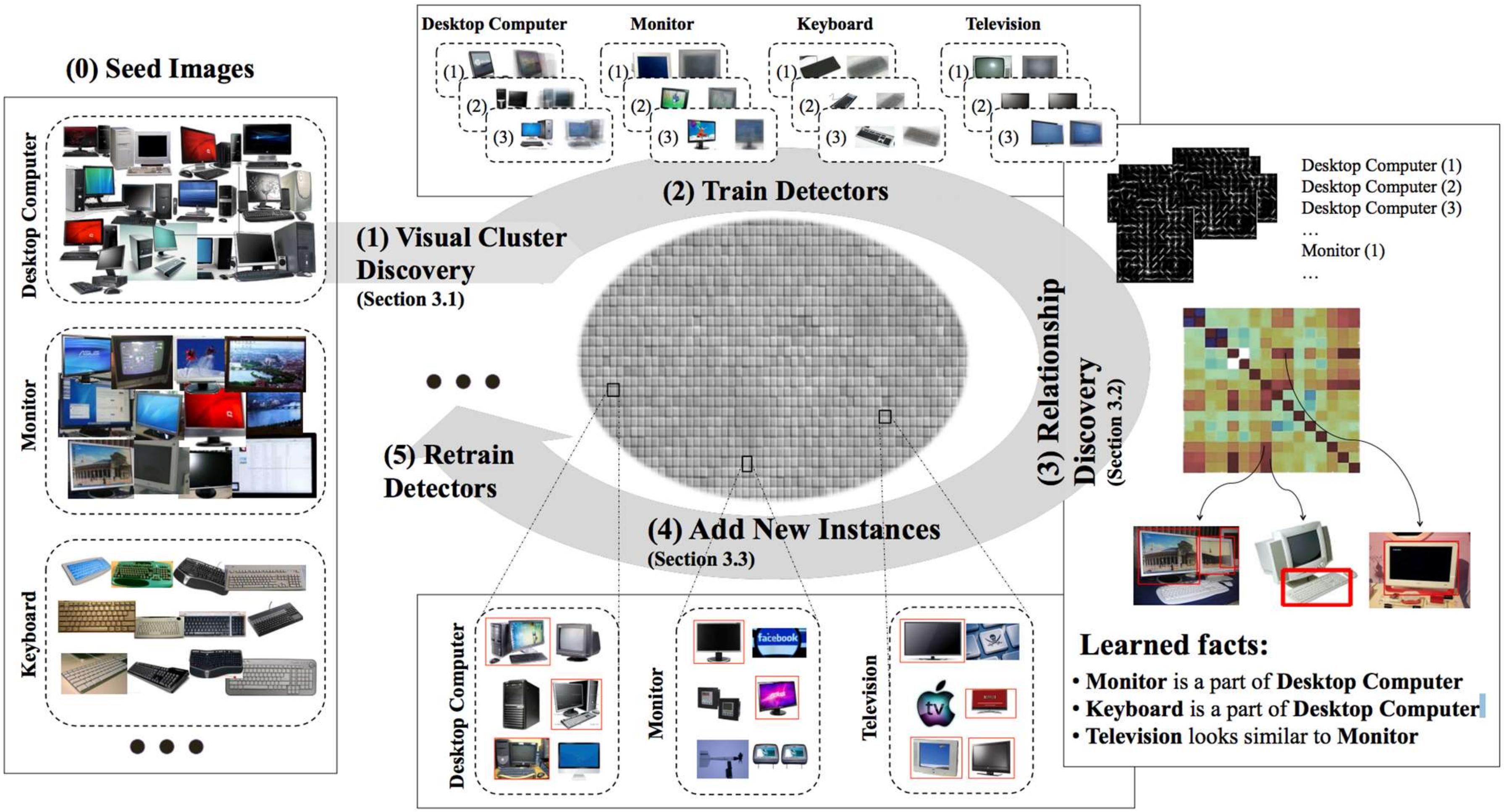
$80,000 · a year ago

#	△1w	Team Name ✱ in the money	Kernel	Team Members	Score	Entries	Last
1	—	✱ Alejandro Mosquera			0.59375	2	1y
2	new	✱ Cardal			0.59000	2	1y
3	new	✱ poweredByTalkwalker		+4	0.59000	4	1y

(2) Exciting active research:
Multi-modal information extraction

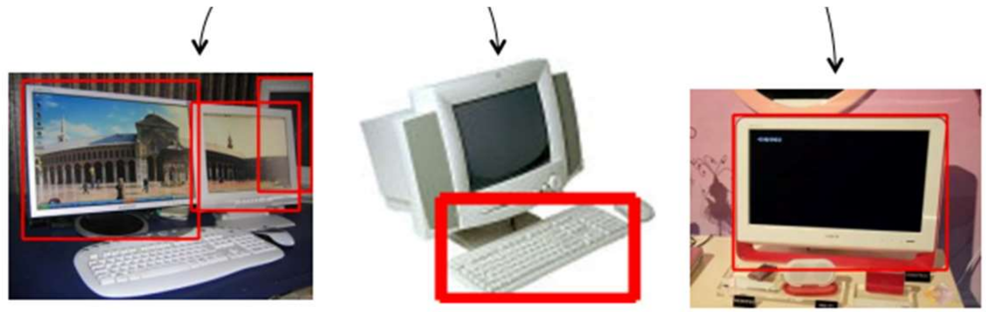


NEIL: Extracting Visual Knowledge from Web Data



[Chen et al., "NEIL: Extracting Visual Knowledge from Web Data," *ICCV* 2013]

NEIL: Extracting Visual Knowledge from Web Data

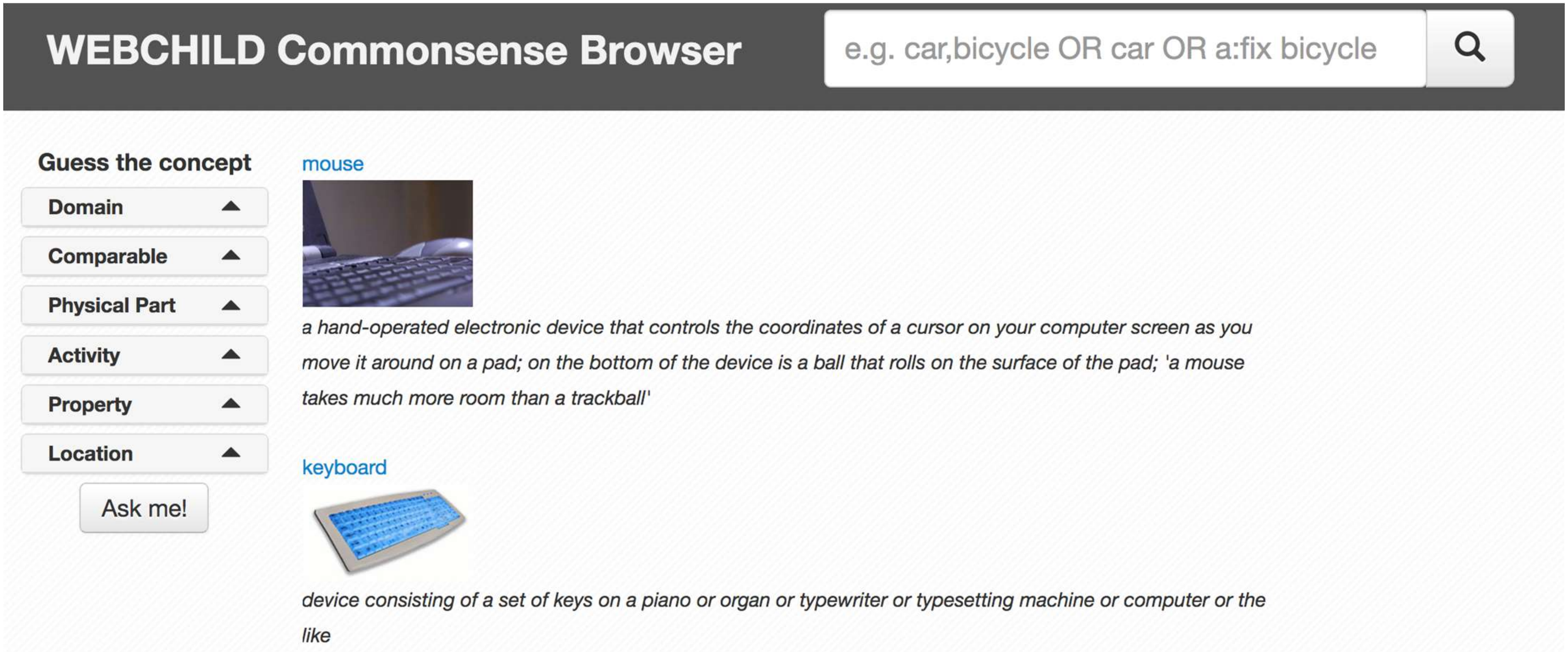


Learned facts:

- **Monitor** is a part of **Desktop Computer**
- **Keyboard** is a part of **Desktop Computer**
- **Television** looks similar to **Monitor**

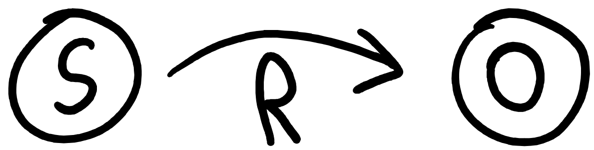
[Chen et al., "NEIL: Extracting Visual Knowledge from Web Data," *ICCV* 2013]

WebChild: Text + Images

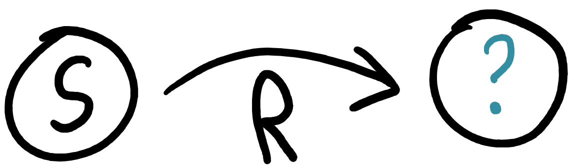


[Tandon et al. "Commonsense in Parts: Mining Part-Whole Relations from the Web and Image Tags." AAAI '16]

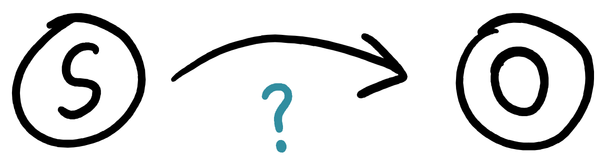
Knowledge Base Completion



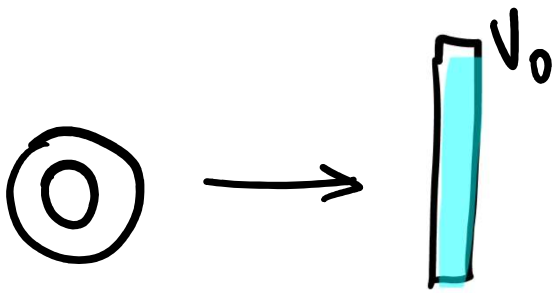
Entity Prediction



Link Prediction



Restrictions in the Model



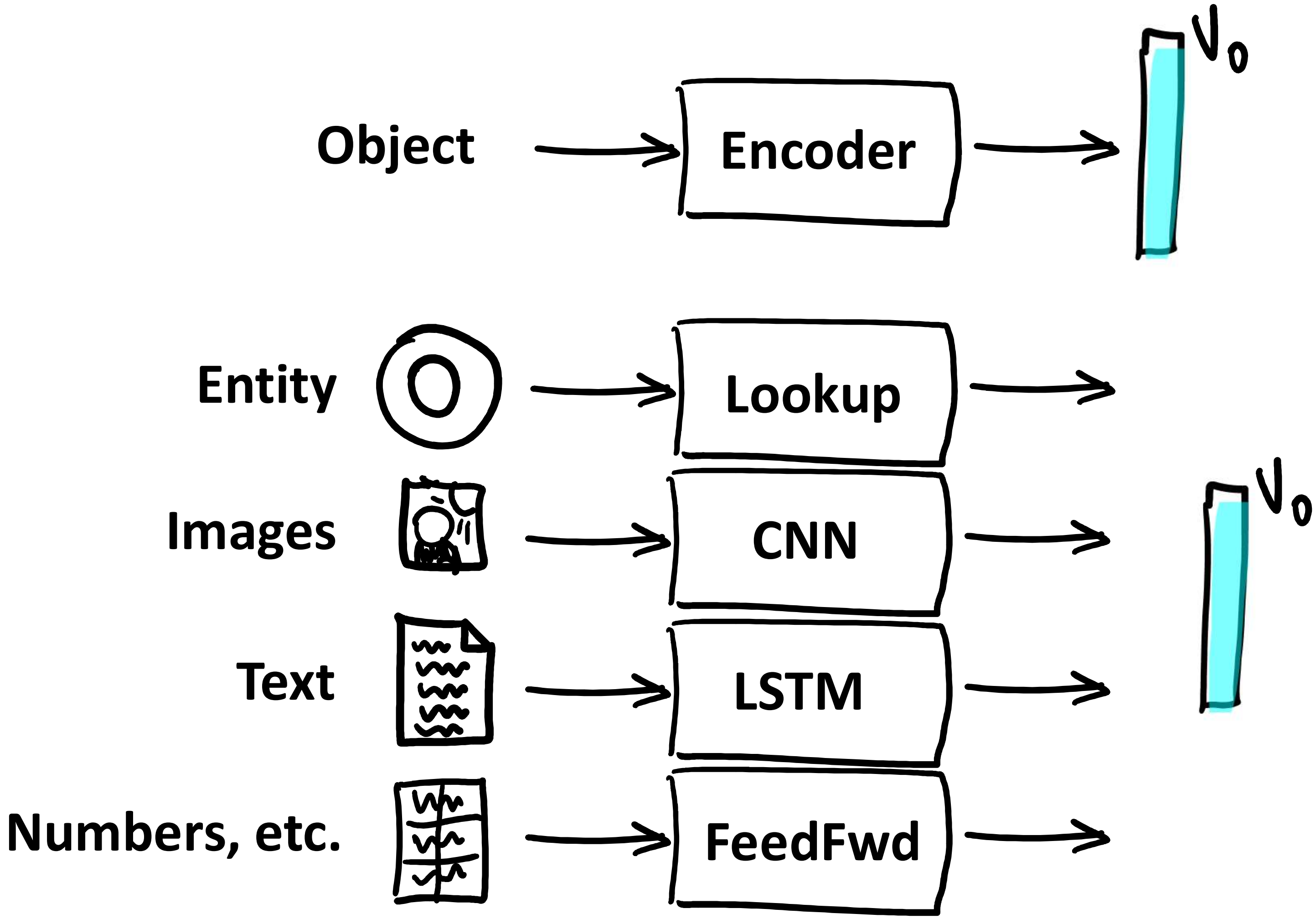
Each object has a vector representation:

- Limits number of objects
- Large number of parameters
- Is not compositional (doesn't generalize)

What about other kinds of objects?

- Dates and Numbers: should generalize
- Text: Names and Descriptions
- Images: Portraits, Posters, etc.

Multimodal KB Embeddings



Augmenting Existing Datasets

MovieLens-100k-plus	
Relations	13
Users	943
Movies	1682
Posters	1651
Ratings	100,000

YAGO3-10-plus	
Relations	37 → 45
Entities	123,182
Structure Triples	1,079,040
Numbers (Years)	1651
Descriptions	107,326
Images	61,246