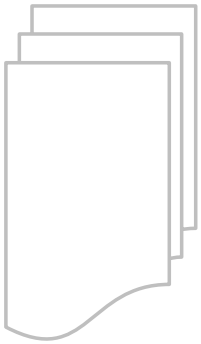
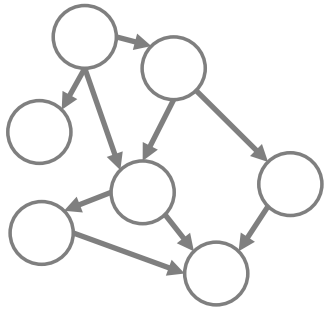
Part 1: Knowledge Graphs



Part 2: Knowledge Extraction

Part 3: Graph Construction



Part 4: Critical Analysis

# Tutorial Outline

1. Knowledge Graph Primer [Jay] 
2. Knowledge Extraction from Text
    a. NLP Fundamentals [Sameer] 
    b. Information Extraction [Bhavana] 

Coffee Break 

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

# Critical Overview

# Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING ACTIVE RESEARCH

FUTURE RESEARCH DIRECTIONS

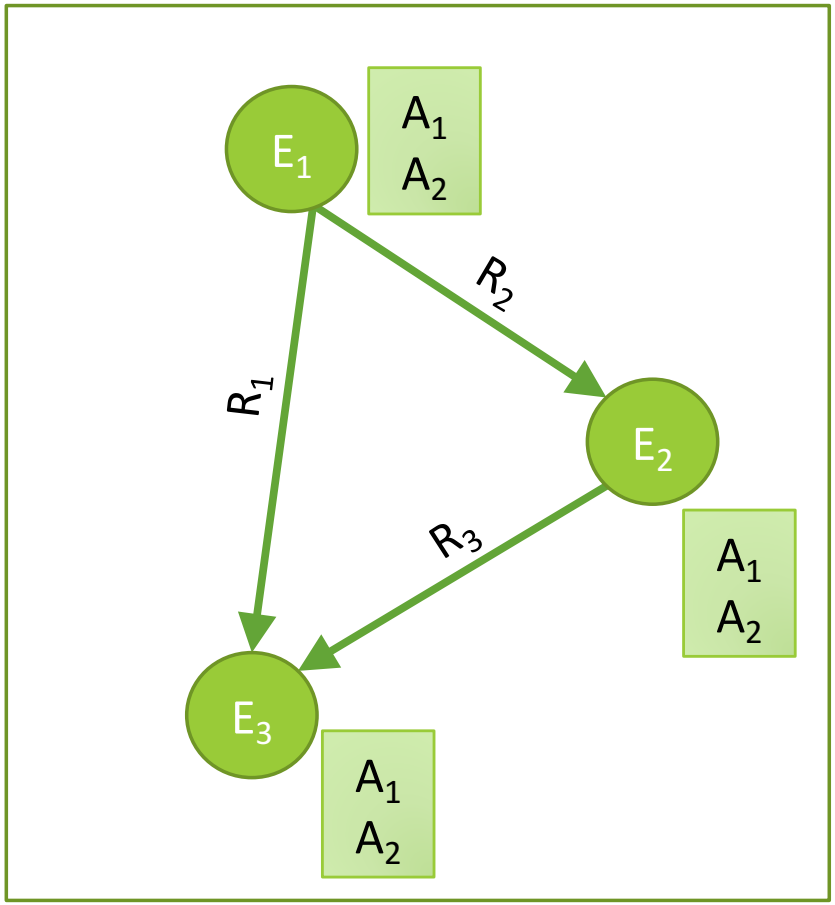
# Why do we need Knowledge graphs?

- Humans:
  - Combat information overload
  - Explore via intuitive structure
  - Tool for supporting knowledge-driven tasks

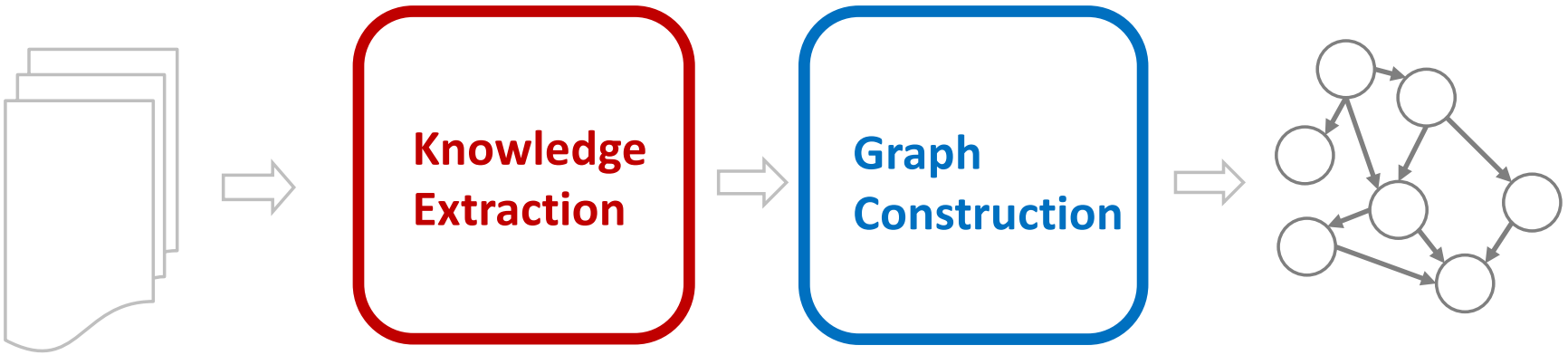
- AIs:
  - Key ingredient for many AI tasks
  - Bridge from data to human semantics
  - Use decades of work on graph analysis

# 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)** | | |
| **What are their attributes? (labels)** | | |
| **How are they related? (edges)** | | |

# Two perspectives

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

# Knowledge Extraction

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

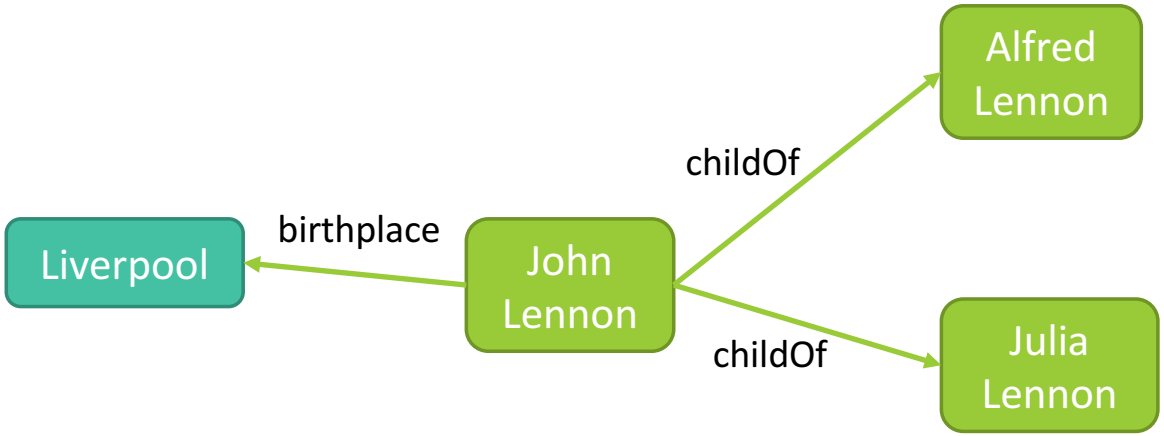
John was born in Liverpool, to Julia and Alfred Lennon.

NNP VBD VBD IN NNP TO NNP CC NNP NNP

***Annotated text***

**Information Extraction**

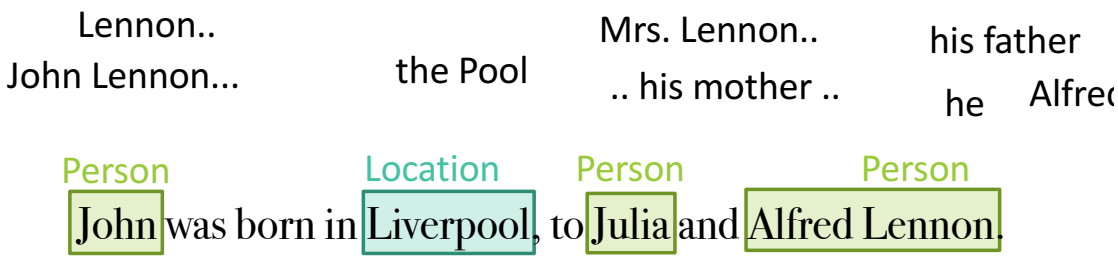
***Extraction graph***



# NLP

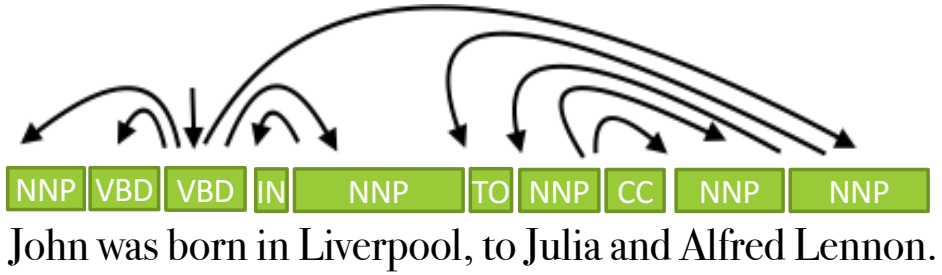
Document

Within-doc Coreference...

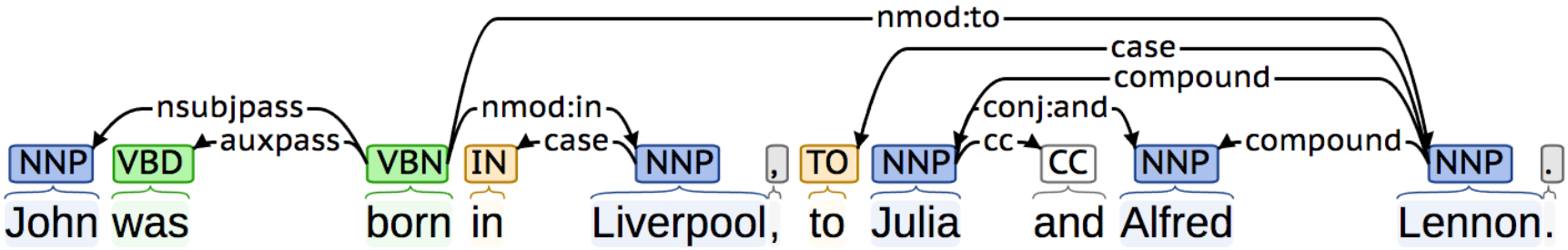



Sentence

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



# Dependency Parsing




## Uses in KG Construction:

- Incredibly useful for relations!
  - What verb is attached?
  - Relation to which mention?
- Incredibly useful for attributes!
  - Appositives: "X, the CEO, …"
- Paths are used as surface relations

# Information Extraction

***Single extractor***

Defining domain

Learning extractors

Scoring the extractions

×

Manual

Semi-automatic

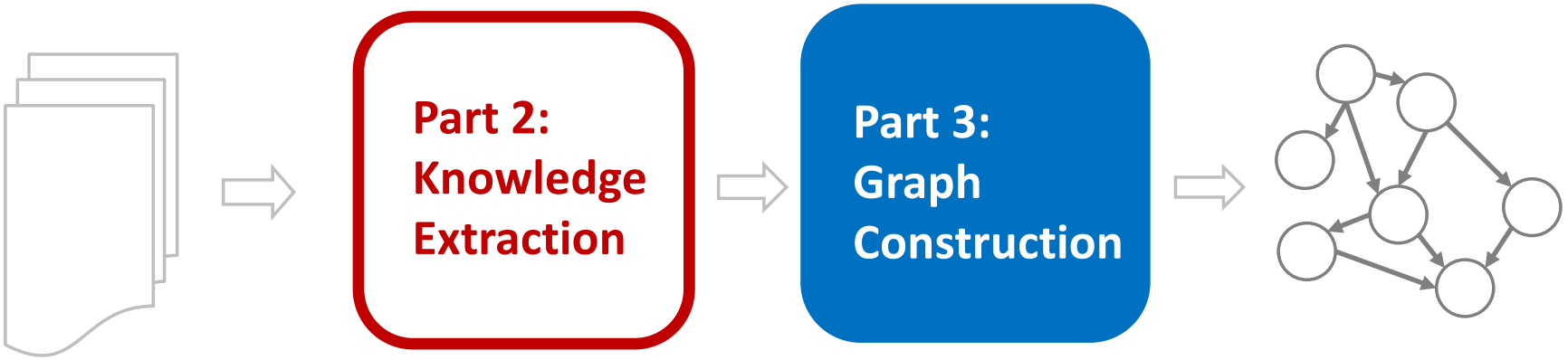
Automatic

Fusing multiple extractors

# IE systems in practice

| | Defining domain | Learning extractors | Scoring extractions | Fusing extractors |
|---|---|---|---|---|
| ConceptNet |  |  |  | |
| NELL |  |  |  | Heuristic rules |
| Knowledge Vault |  |  |  | Classifier |
| OpenIE |  |  |  | |

# Graph Construction



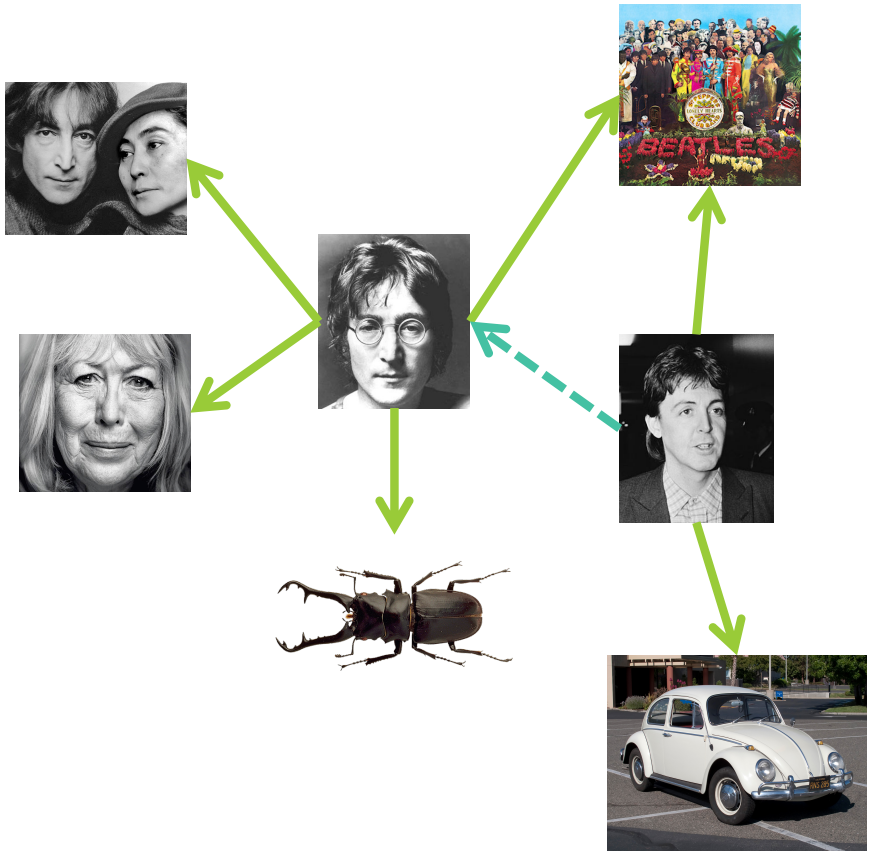

**Extraction graph**

**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
- Universally-quantified

## RANDOM WALK BASED

- Possible facts posed as queries
- Random walks of the KG constitute “proofs”
- Probability $\propto$ path lengths/transitions
- Locally grounded

# Two approaches for KG construction

PROBABILISTIC MODELS

EMBEDDING BASED MODELS

# Why embeddings?

Embeddings

## Limitation to Logical Relations

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

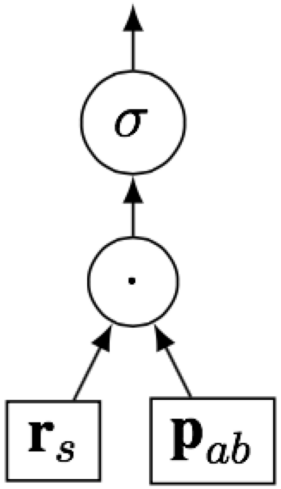
- Everything as dense vectors
- Captures many relations
- Learned from data

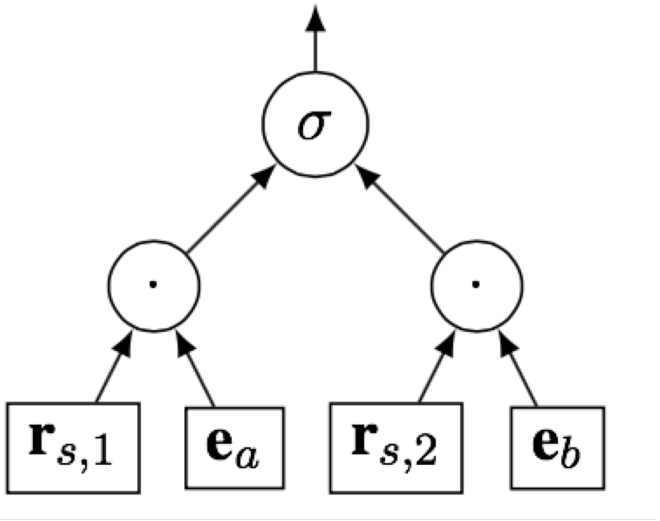
## Computational Complexity of Algorithms

- 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

- Learning using stochastic gradient, back-propagation
- Querying is often cheap
- GPU-parallelism friendly

# Matrix vs Tensor Factorization

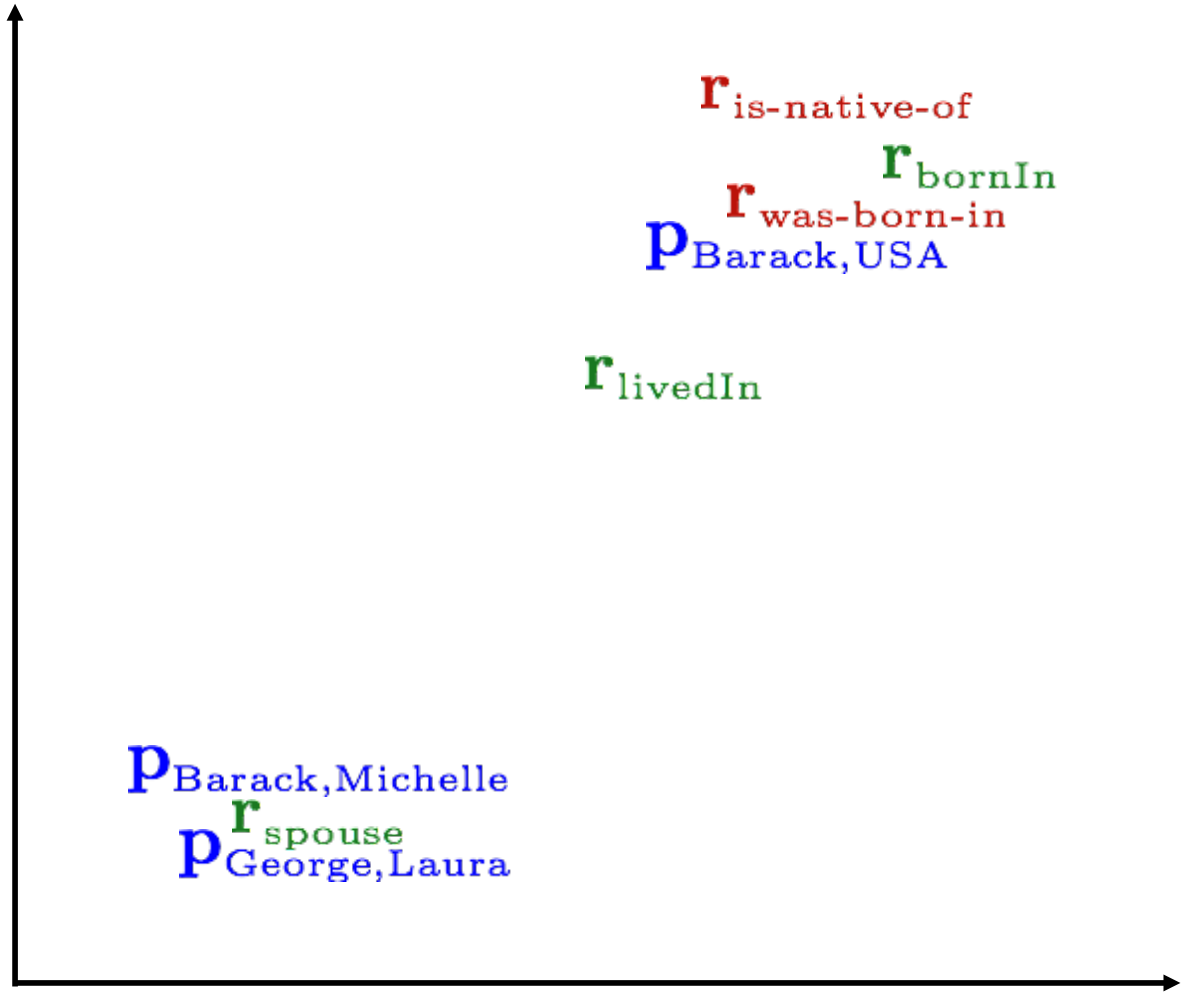


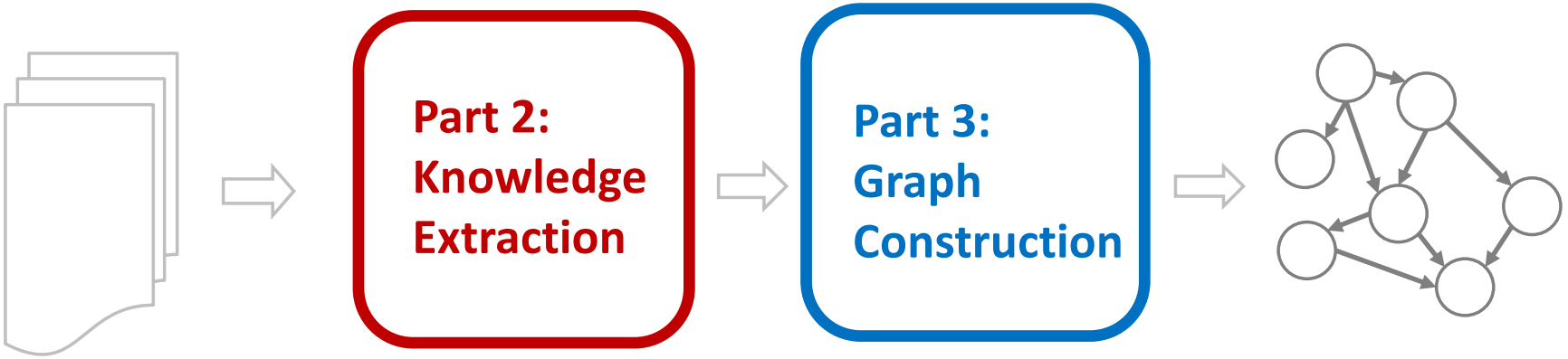




- No encoding of type information
- Can only predict for entity pairs that appear in text together
- Sufficient evidence has to be seen for each entity pair

- Assume low-rank for pairs
- But many relations are not!
- Spouse: you can have only ~1
- Cannot learn pair specific information

# Relation Embeddings




Part 2:
Knowledge
Extraction
Part 3:
Graph
Construction

# Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING ACTIVE RESEARCH

FUTURE RESEARCH DIRECTIONS

# 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

# Success story: OpenIE (ReVerb)

**Open Information Extraction** openie.allenai.org

Hosted by AI2 Created at

**Argument 1:** entity:The Beatles **Relation:**

**Argument 2:** All Search

all | location (21) | film location (18) | statistical region (16) | name source (15) | travel destination (14) | misc.

more types

- were bigger than **Jesus** (100)
- came to America (95)
- appeared on **The Ed Sullivan Show** (88)
- broke up in 1970 (56)
- Here Comes the Sun (46)
- came to America (45)
- is for the future (44)
- are a great band (42)
- perform on **The Ed Sullivan Show** (39)
- were **Musical ensemble** (36)

**are a great band**

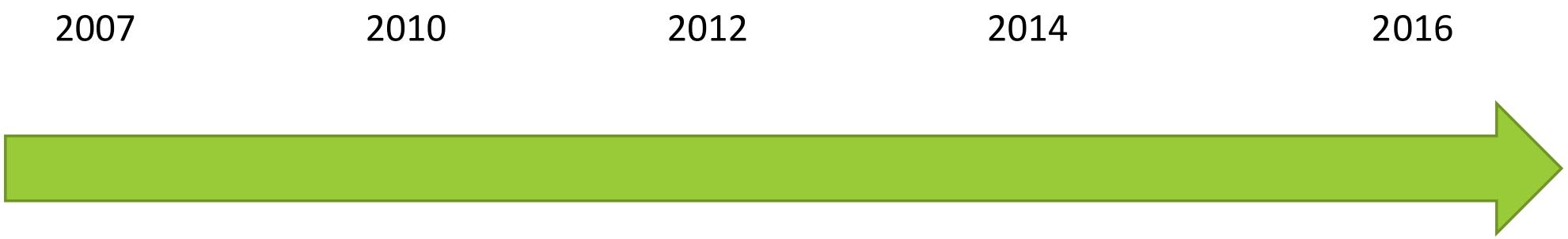
**Extracted Synonyms:**

were
is
was

**Extracted from these sentences:**

**are** **The Beatles** are **the best band** , hands down but Oasis did make a great cover . (via ClueWeb12)
**The Beatles** are **a great band** . (via ClueWeb12)
**The Beatles** are **the best band** . (via ClueWeb12)
**The Beatles** are **the greatest band** ... Started 1 month ago by georgedcc Yeah , Songs in the Key of Life is a bit much for 1 listen . (via ClueWeb12)
**The Beatles** , arguably , are **the greatest band** , and may or may not have the greatest name . (via ClueWeb12)
The point is , from my view , **The Beatles** are **a good band** , but way behind the greatest artists to ever grace rock . (via ClueWeb12)

# Open IE Systems

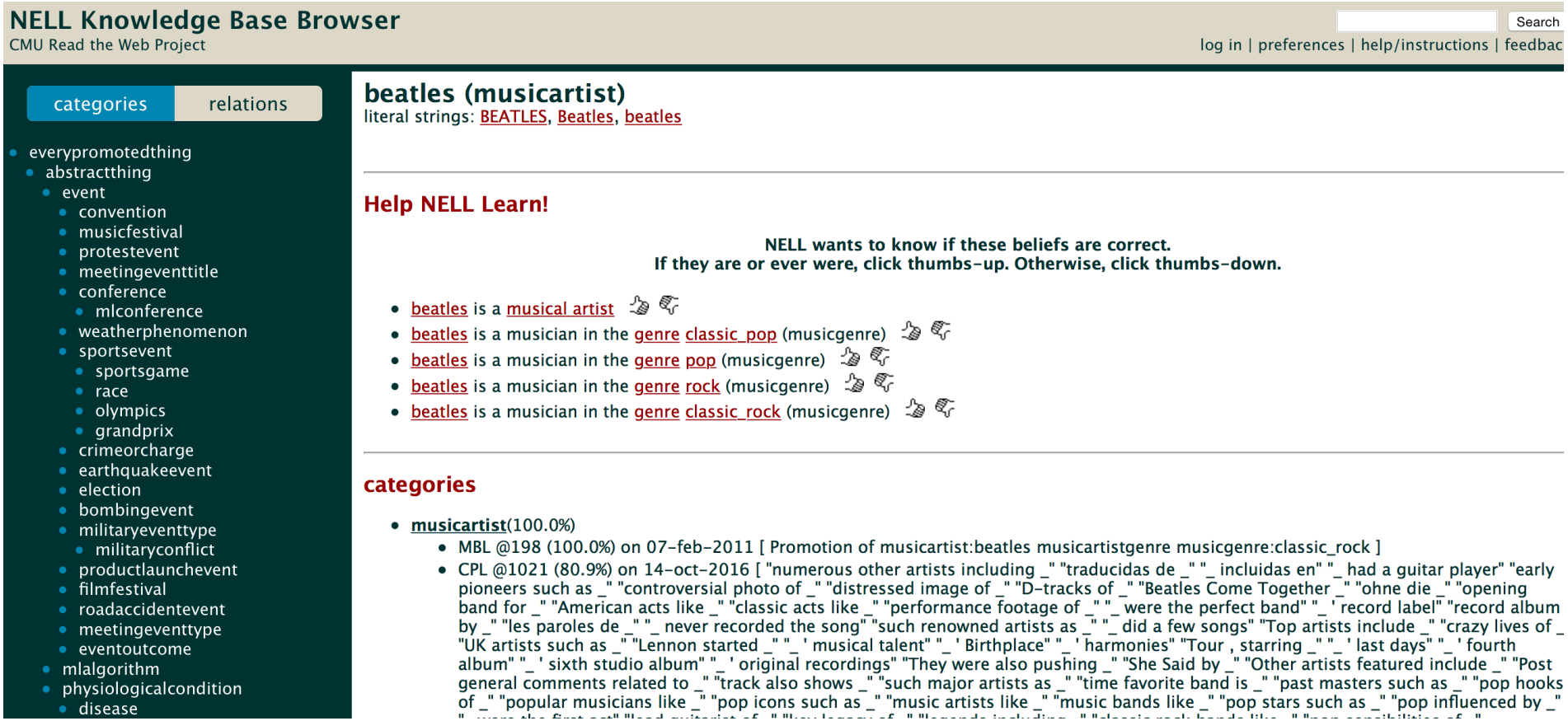


# Success story: NELL

- **Key technical contributions:**
  - Never ending learning,
  - Coupled bootstrapping

- Input: Clueweb09 corpus (1B web pages)
- Ontology: ~2K predicates
  ~700K constraints between predicates
- Output: 50 million candidate facts
  3 million high-confidence facts

# Success story: NELL

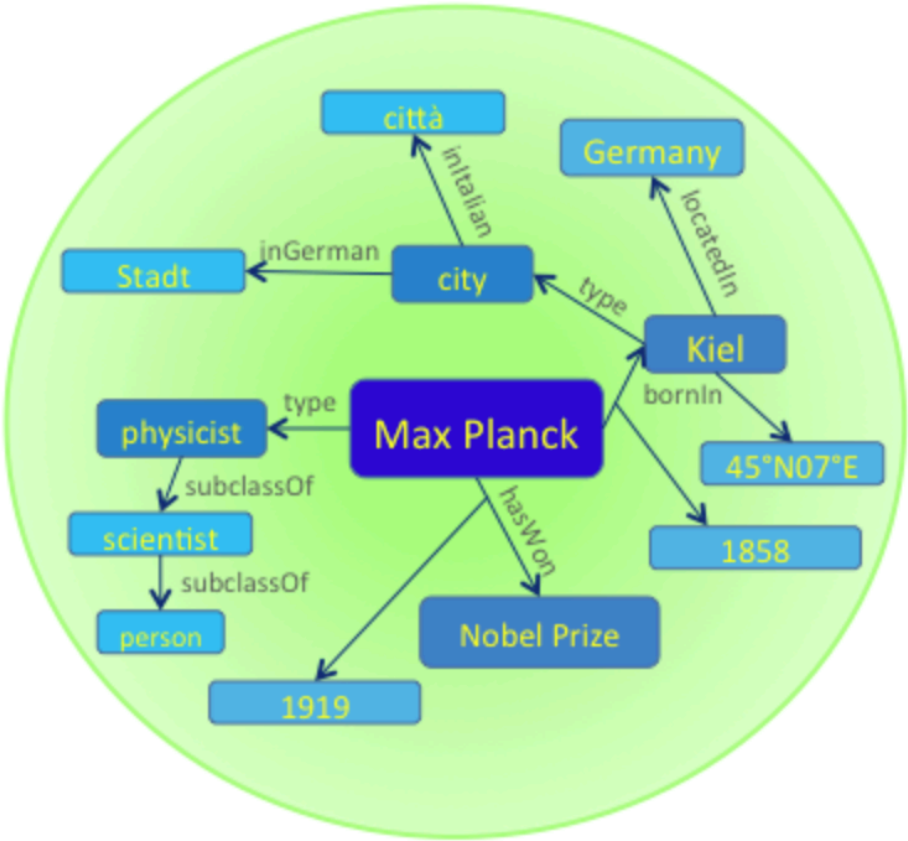


# Success story: YAGO

- **Key contributions:**
  - **Rich Ontology:** Linking Wikipedia categories to WordNet, providing a rich taxonomy
  - **High Quality:** High precision extractions (accuracy ~95%)

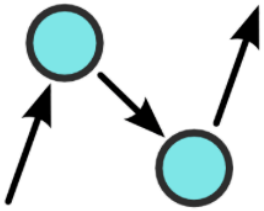
# Success story: YAGO

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



# 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



# ConceptNet

An open, multilingual knowledge graph

## 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) →

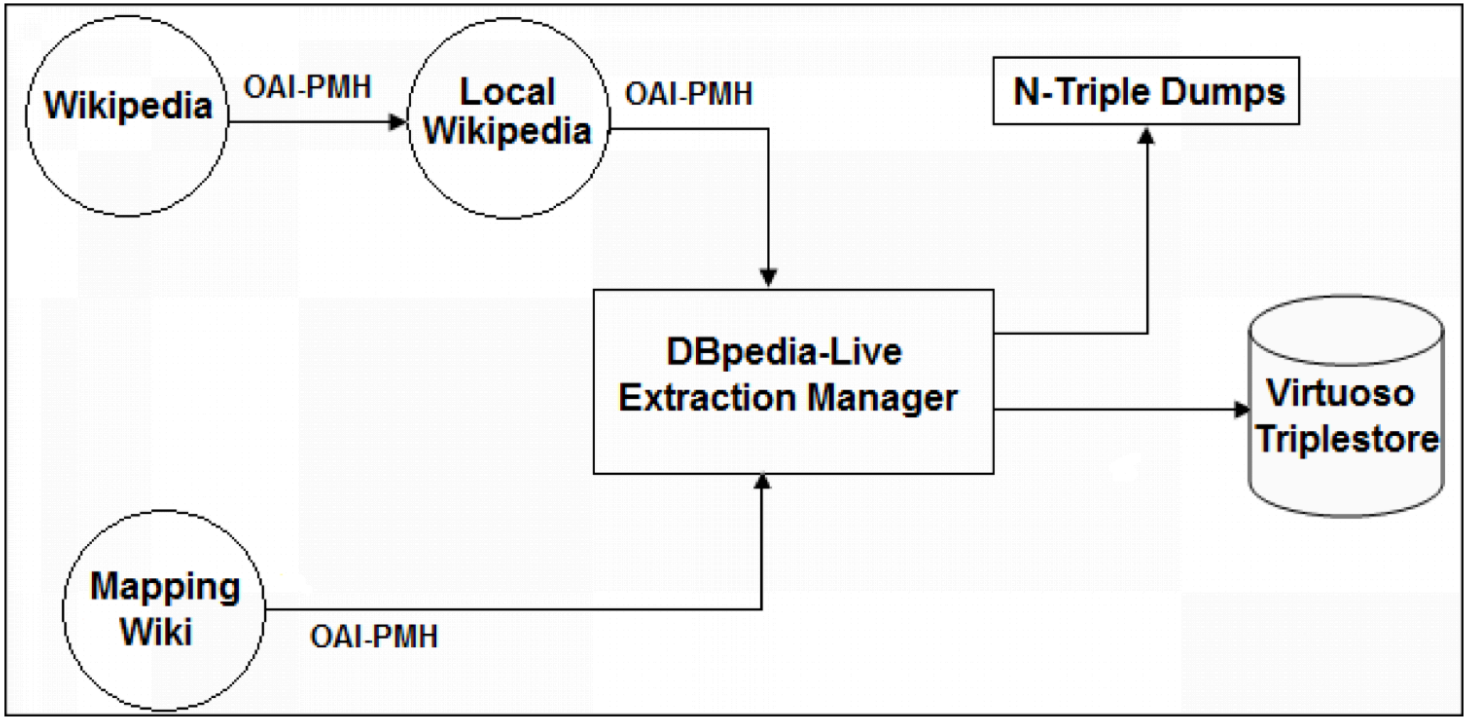
### Links to other sites

- 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

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

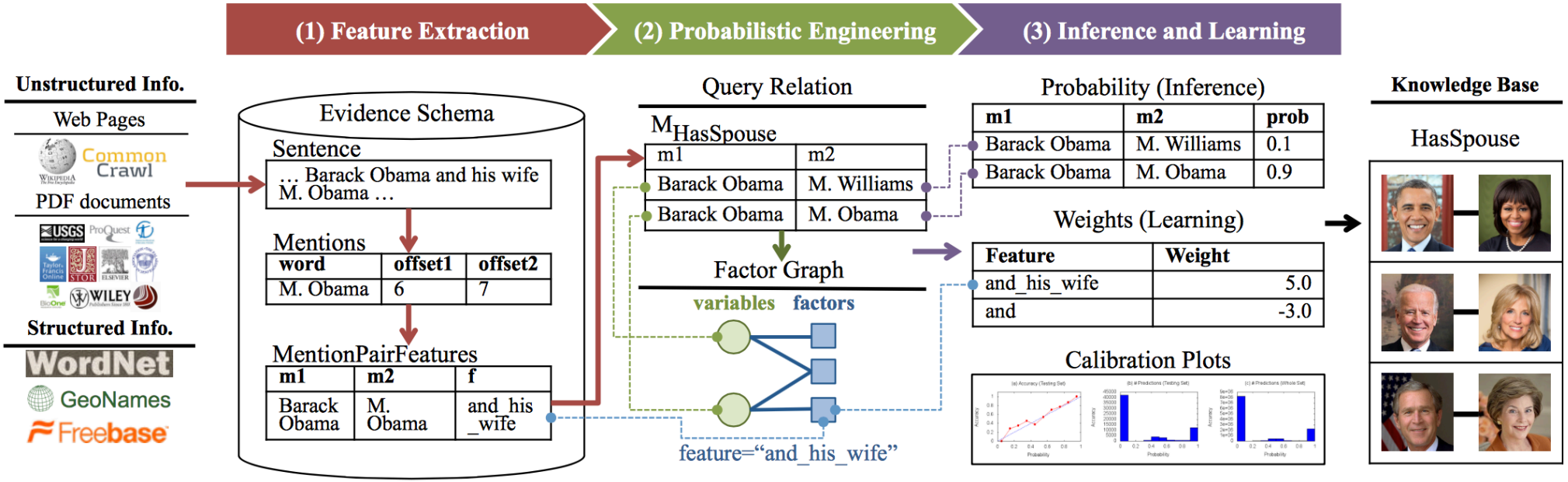
- Wikipedia is manually built encyclopedia project
- DBPedia is automatically extracted structured data from Wikipedia
  - 17M canonical instances
  - 88M type statements
  - 72M infobox statements



# DeepDive


- 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)**

# DeepDive




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

# Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING ACTIVE RESEARCH

FUTURE RESEARCH DIRECTIONS

# Datasets

- KG as datasets
  - FB15K-237 Knowledge base completion dataset based on Freebase
  - 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

# Shared tasks

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

# Softwares: 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)

UNIVERSITY of WASHINGTON

- FACTORIE: out-of-the-box tools for NLP and information integration
  link (Scala code)


- EasySRL: Semantic role labeling
  link (Java code)

UNIVERSITY of WASHINGTON

# Softwares: Information Extraction

- 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)

- Universal schema
  link (Java + Scala code)

UMASS
AMHERST

# Softwares: Knowledge Graphs

- **AMIE:** Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases
  link (Java code)


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


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

Carnegie Mellon University

- **Junto:** graph based semi-supervised learning methods
  link (Scala code)

# Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING ACTIVE RESEARCH

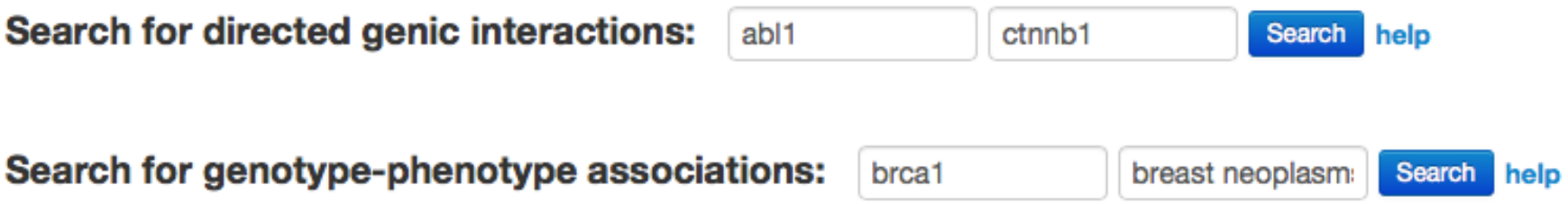
FUTURE RESEARCH DIRECTIONS

# Exciting Active Research

- INTERESTING APPLICATIONS OF KG
- AMBITIOUS PROJECTS
- MULTI-MODAL INFORMATION EXTRACTION

# (1) Exciting active research:
## Interesting application of Knowledge Graphs

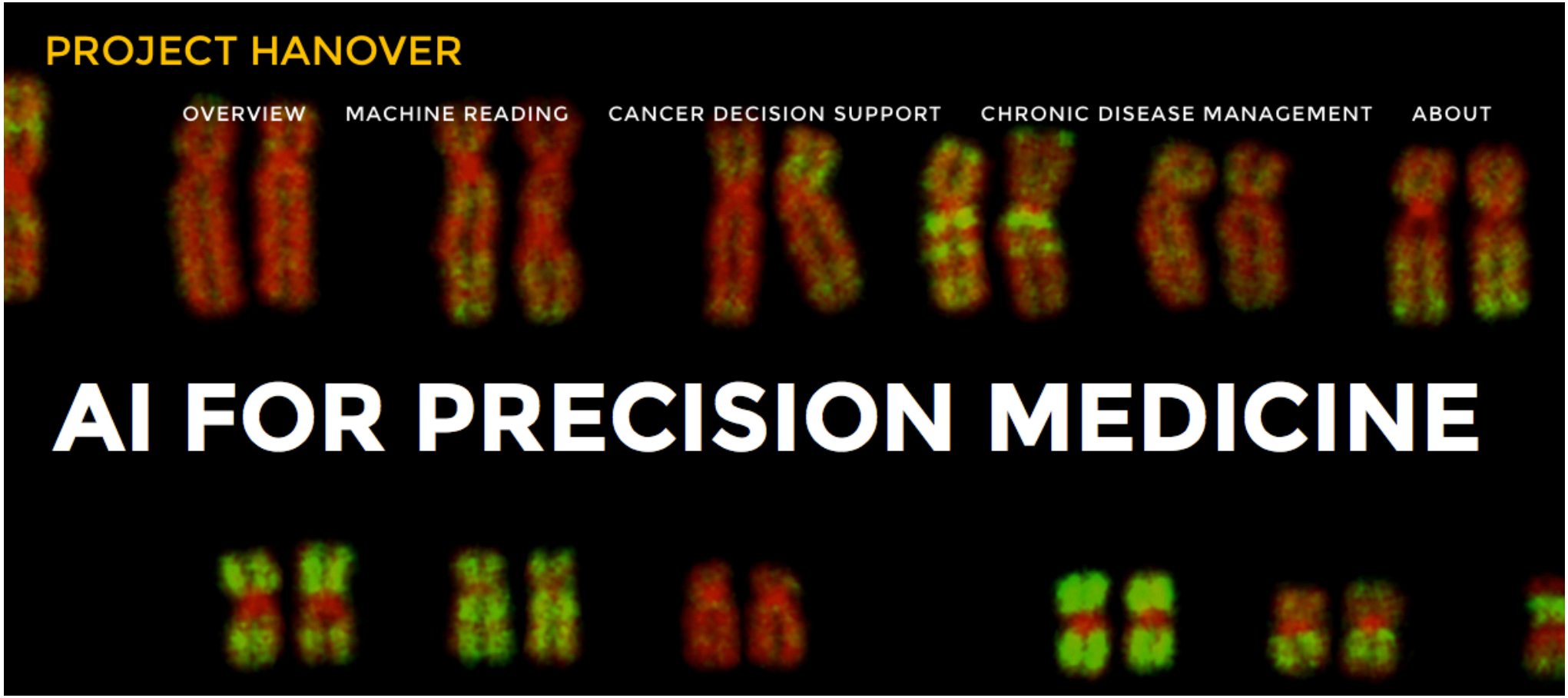
- **The Literome Project [link]**




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

# (1) Exciting active research:
## Interesting application of Knowledge Graphs

Microsoft Research



# (2) Exciting active research:
## Ambitious Projects


**ARISTO**

*Answering Science Questions using Machine Reading*

What is Aristo?

**Question:** During which season of the year would a rabbit's fur be thickest?

Hide

**Answer:** (D) winter

**Because:** A bear's fur would be thickest during winter.

**Confidence:** 60.14%

**Source:** Barrons 4th Grade Study Guide

# 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

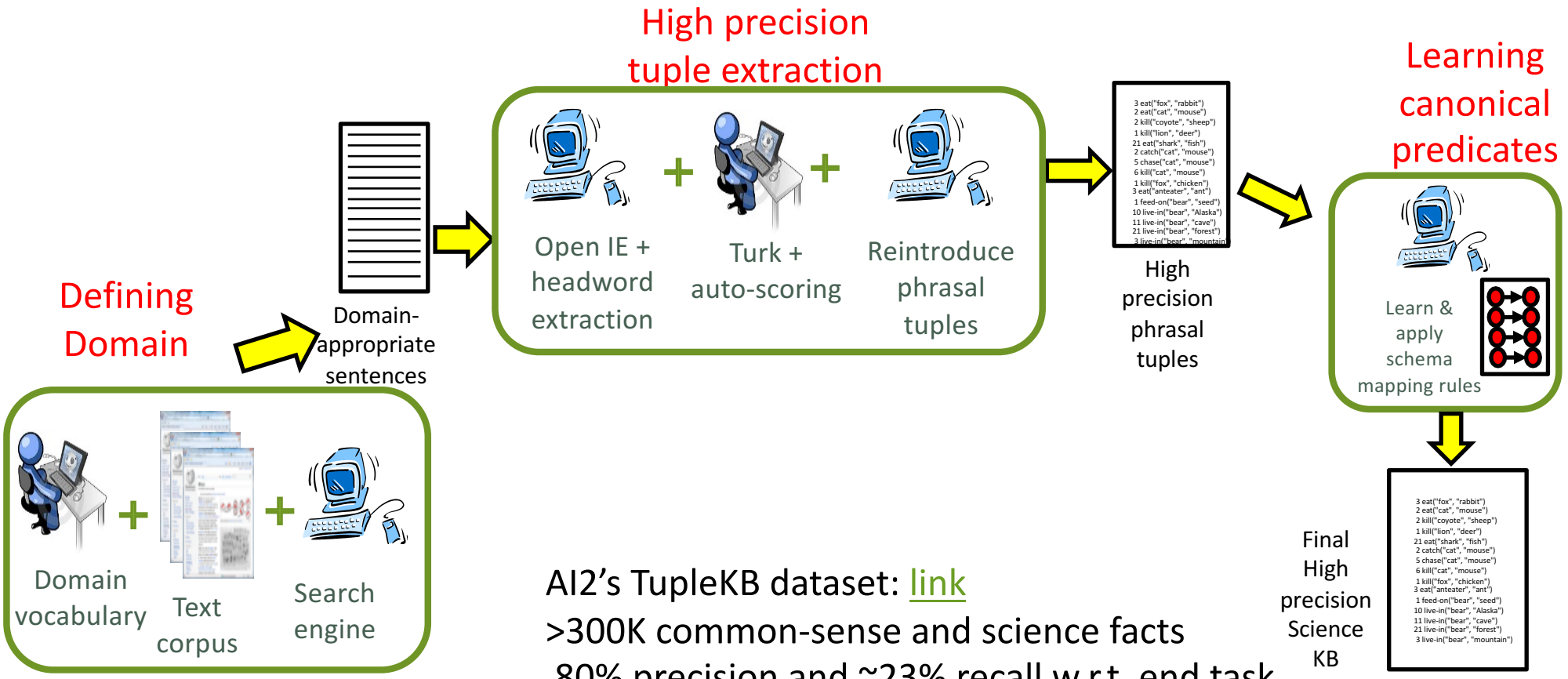
frog's life cycle, metamorphosis

Common sense knowledge

frog is an animal, animals have life cycle

# AI2's ScienceKB

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE




AI2's TupleKB dataset: link

>300K common-sense and science facts

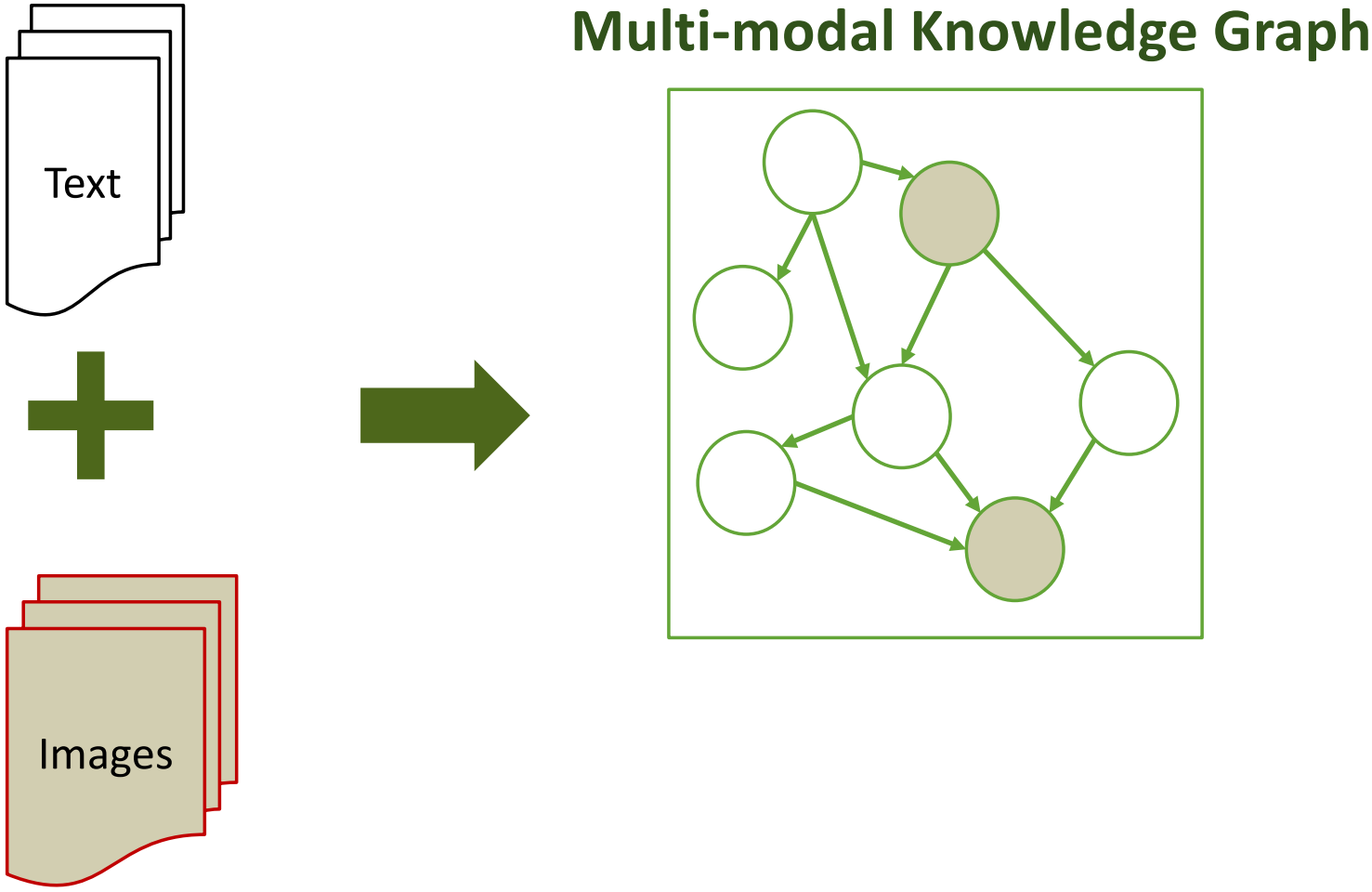
80% precision and ~23% recall w.r.t. end task.

**Upcoming article in 2017 ``High Precision Knowledge Extraction for Science domain''

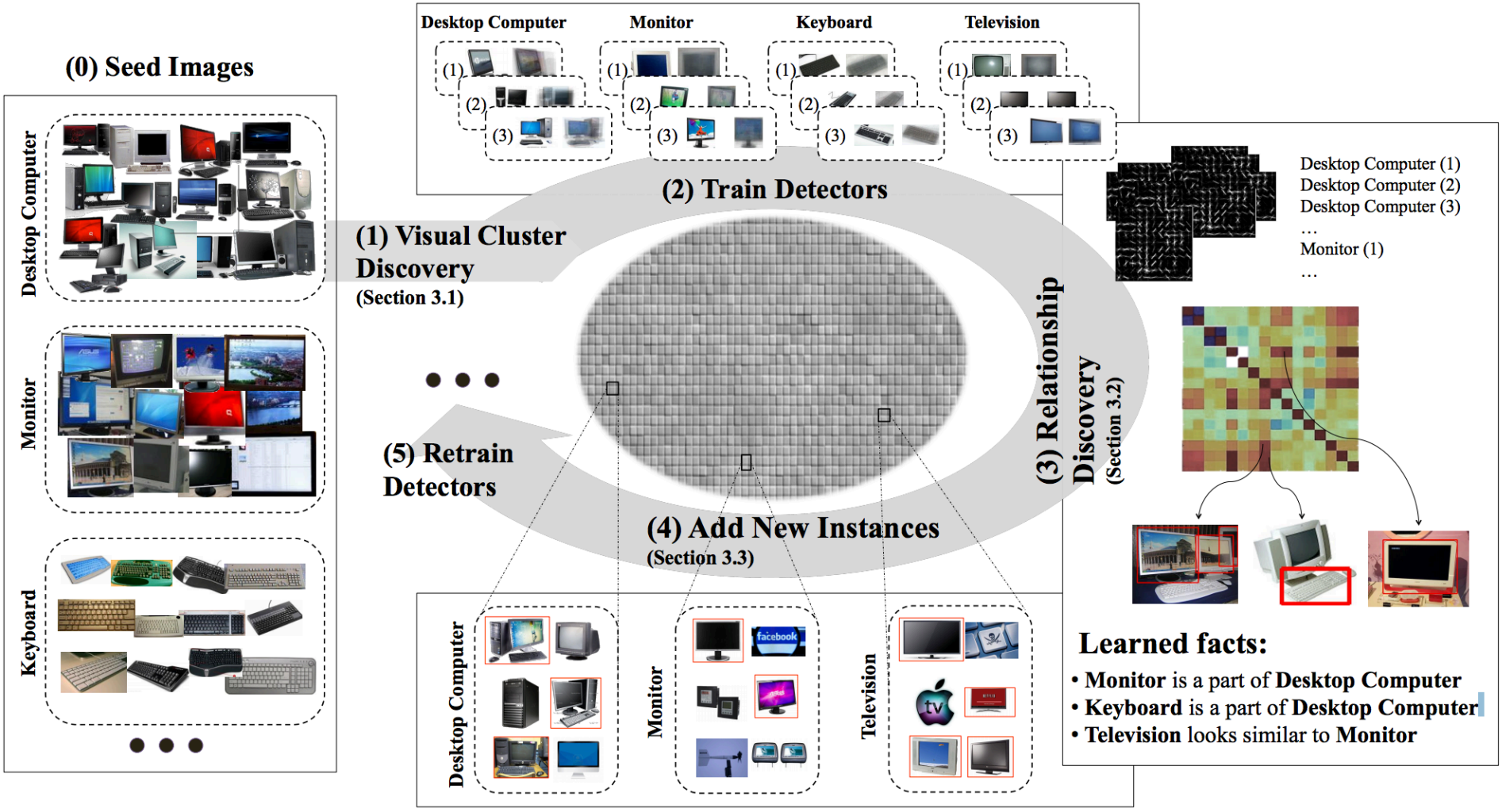
# Aristo ScienceKB

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

# (3) 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

**WEBCHILD Commonsense Browser**

e.g. car,bicycle OR car OR a:fix bicycle

**Guess the concept**

- Domain
- Comparable
- Physical Part
- Activity
- Property
- Location

Ask me!

mouse

*a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad; 'a mouse takes much more room than a trackball'*

keyboard

*device consisting of a set of keys on a piano or organ or typewriter or typesetting machine or computer or the like*

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

# Critical Overview

SUMMARY

SUCCESS STORIES

DATASETS, TASKS, SOFTWARES

EXCITING ACTIVE RESEARCH

FUTURE RESEARCH DIRECTIONS

# Future Research Directions

GOING BEYOND FACTS

CONTINUOUSLY LEARNING AND SELF-CORRECTING SYSTEMS

# (1) Future research directions:
## Going beyond facts

- Representing and learning complex models
- Represent higher level structures like activities, events, processes

# Future research directions:
## Going beyond facts

- **Fact:** Individual knowledge tuples
  (plant, take in, CO2)

- **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 ]*

# (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.

# Future……

- Knowledge graphs construction systems

Can consume online stream of data at Web scale,

Represent context beyond facts,

Support domains like medicine, science to help humanity

Can correct its own mistakes over time

# Thank You