

**Part 1: Knowledge Graphs**

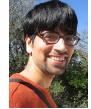
**Part 2:  
Knowledge  
Extraction**

**Part 3:  
Graph  
Construction**

**Part 4: Critical Analysis**

# Tutorial Outline

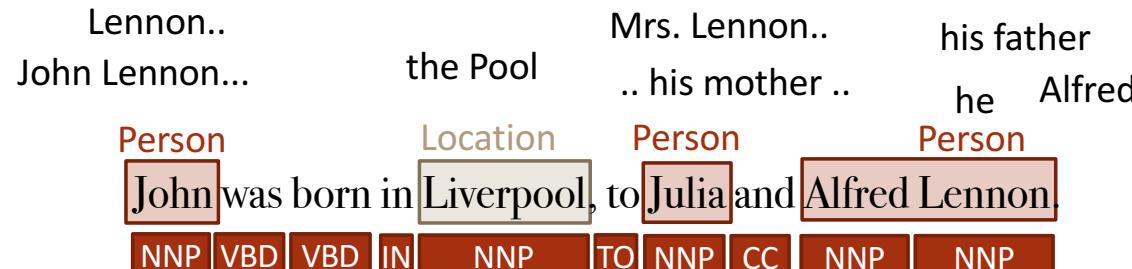
---

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] 

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

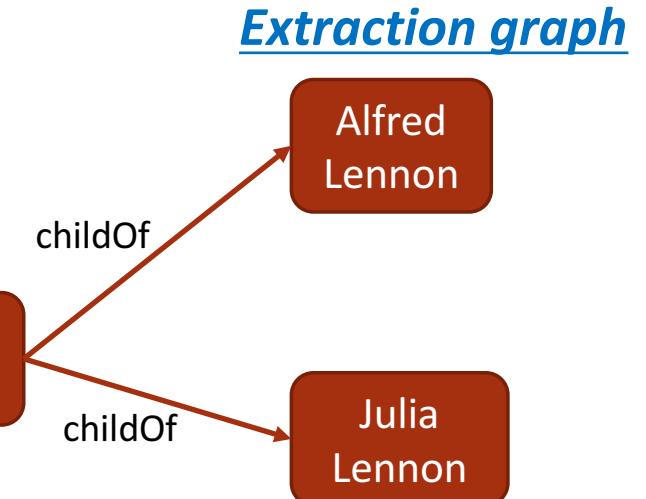
Text

NLP



Annotated text

Information Extraction



# Information Extraction

---

## 3 IMPORTANT SUB-PROBLEMS

CATEGORIES OF IE TECHNIQUES

KNOWLEDGE FUSION

IE SYSTEMS IN PRACTICE

# Information Extraction

## 3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring the extractions

## 3 LEVELS OF SUPERVISION

Supervised



Semi-supervised



Unsupervised



# Information Extraction

## 3 CONCRETE SUB-PROBLEMS

### Defining domain

Learning extractors

Scoring the extractions



## 3 LEVELS OF SUPERVISION

Supervised



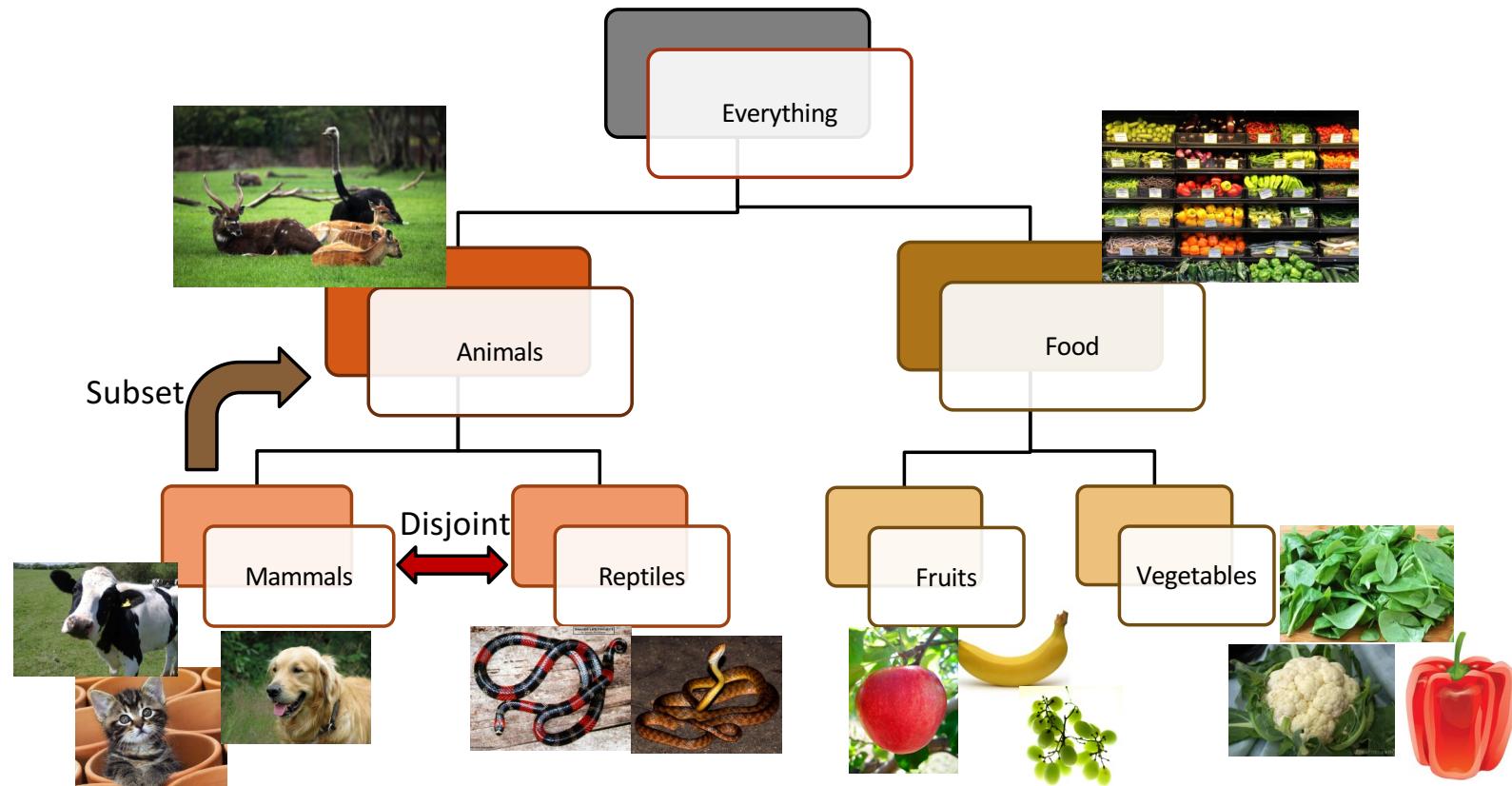
Semi-supervised



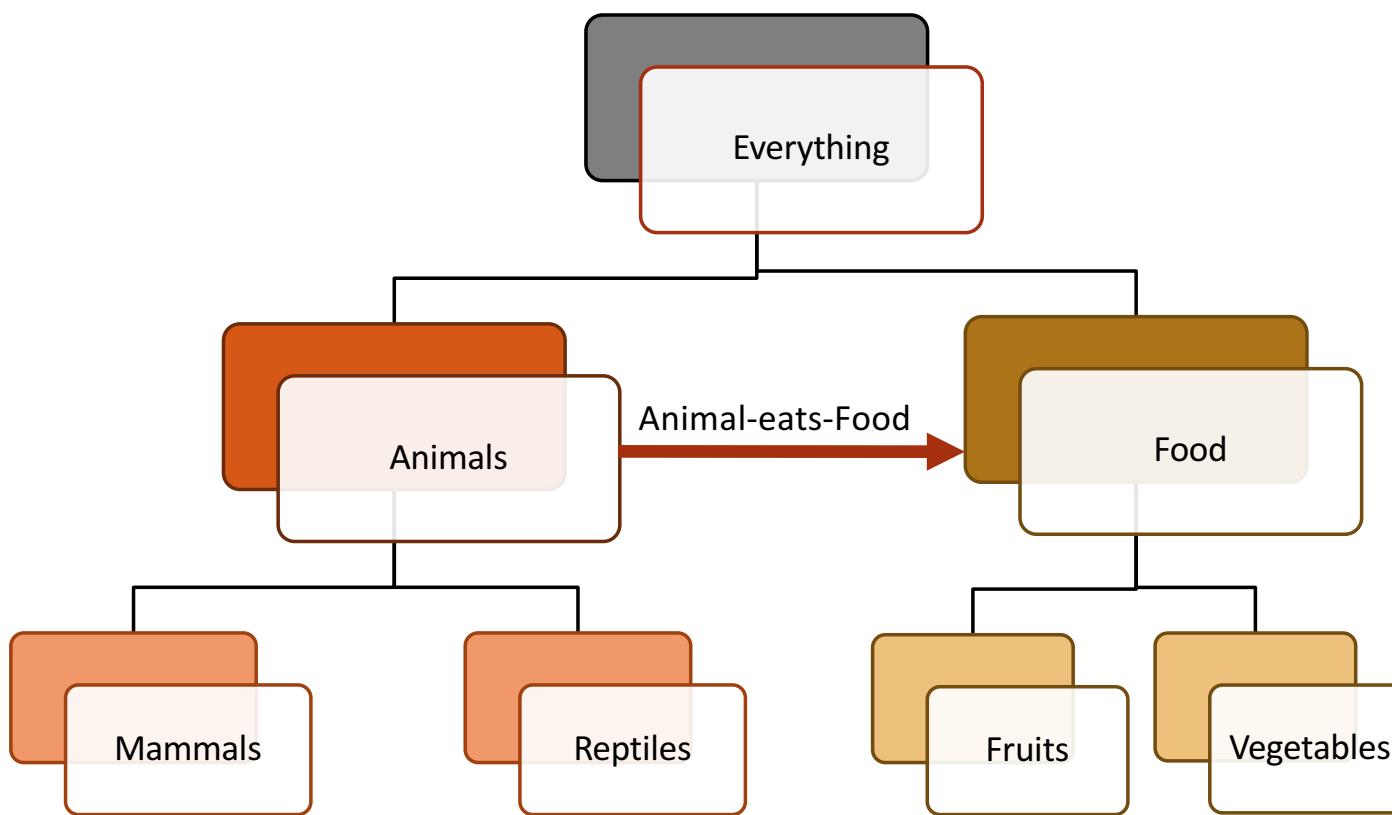
Unsupervised



# Defining Domain: Manual



# Defining Domain: Manual

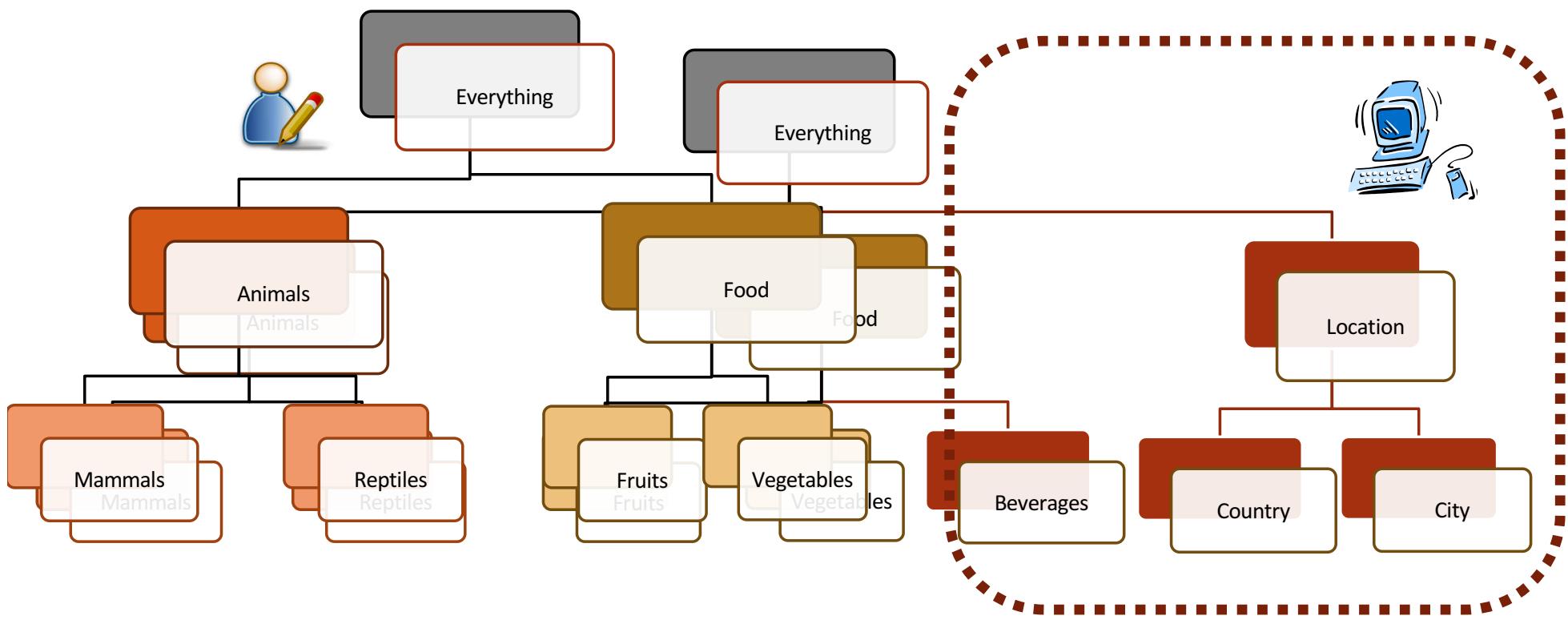


- Highly semantic ontology
- Leads to high precision extractions
- Expensive to create
- Requires domain experts

# Defining Domain: Semi-automatic



- Subset of types are manually defined
- More types are discovered from data



# Defining Domain: Semi-automatic



- Types and type hierarchy is manually defined  
E.g. River, City, Food, Chemical, Disease, Bacteria
- Relations are automatically discovered using clustering methods

- Easier to derive types using existing resources
- Relations are discovered from the corpus
- Leads to moderate precision extractions
- Partially semantic ontology

Discovered relation	Patterns	Seed instances
River -in heart of- City	“in heart of” “in the center of” “which flows through”	“Seine, Paris”, “Nile, Cairo” “Tiber river, Rome” “River arno, Florence”
Food -to produce- Chemical	“to produce” “to make” “to form”	“Salt, Chlorine” “Sugar, Carbon dioxide” “Protein , Serotonin”
Disease -caused by- Bacteria	“caused by” “is the causative agent of” “is the cause of”	“pneumonia, legionella” “mastitis, staphylococcus aureus” “gonorrhea, neisseria gonorrhoeae”

# Defining Domain: Automatic

---



- Any noun phrase is a candidate entity
- Any verb phrase is a candidate relation

- **Cheapest way to induce types/relations from corpus**
- **Little expert annotations needed**
- **Limited semantics**
- **Leads to noisy extractions**

# Information Extraction

## 3 CONCRETE SUB-PROBLEMS

Defining domain

## Learning extractors

Scoring the extractions



## 3 LEVELS OF SUPERVISION

Supervised



Semi-supervised



Unsupervised



# Learning Extractors: Manual



- Human defined high-precision extraction patterns for each relation

Person-member of-Band



<PERSON> works for <BAND>  
<PERSON> is part of <BAND>

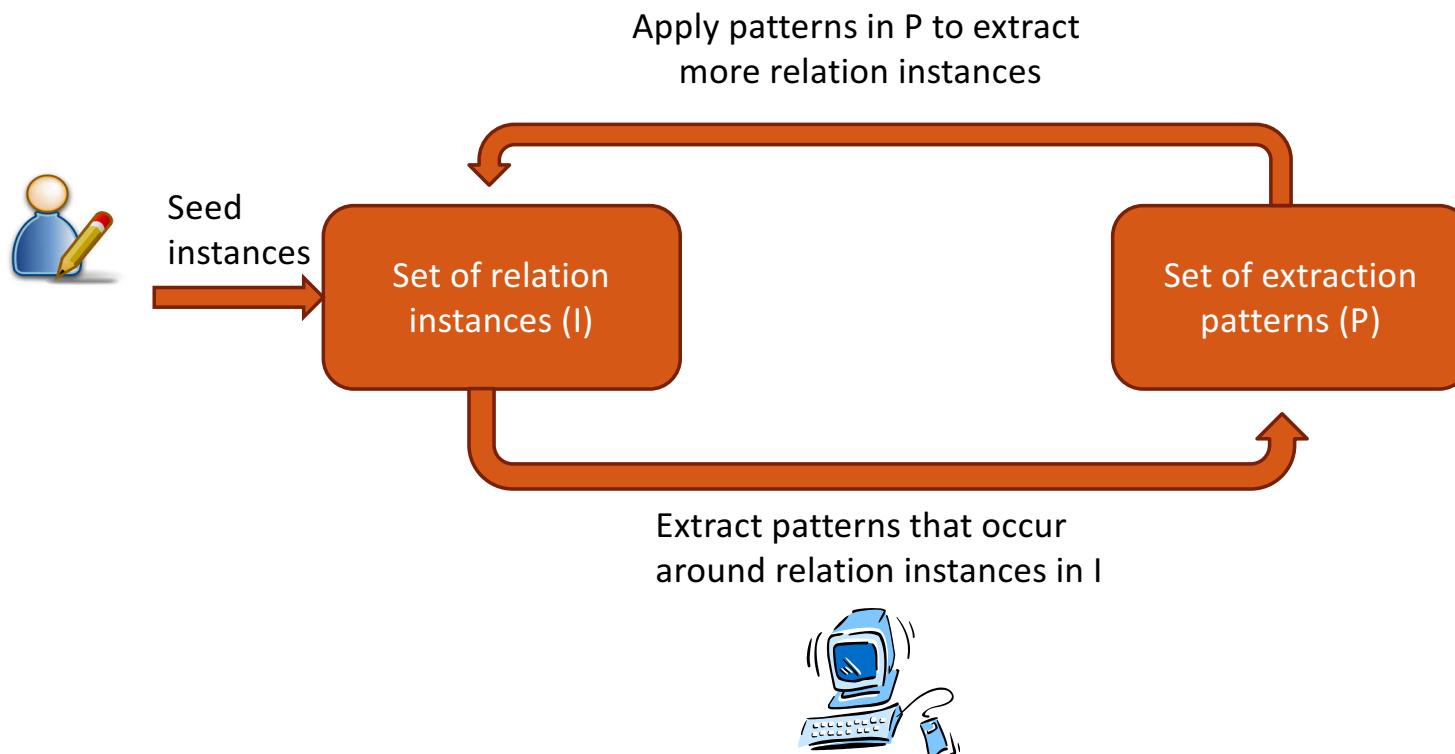


Extract relation instances  
(John Lennon, The Beatles)  
(Brian Jones, The Rolling Stones)

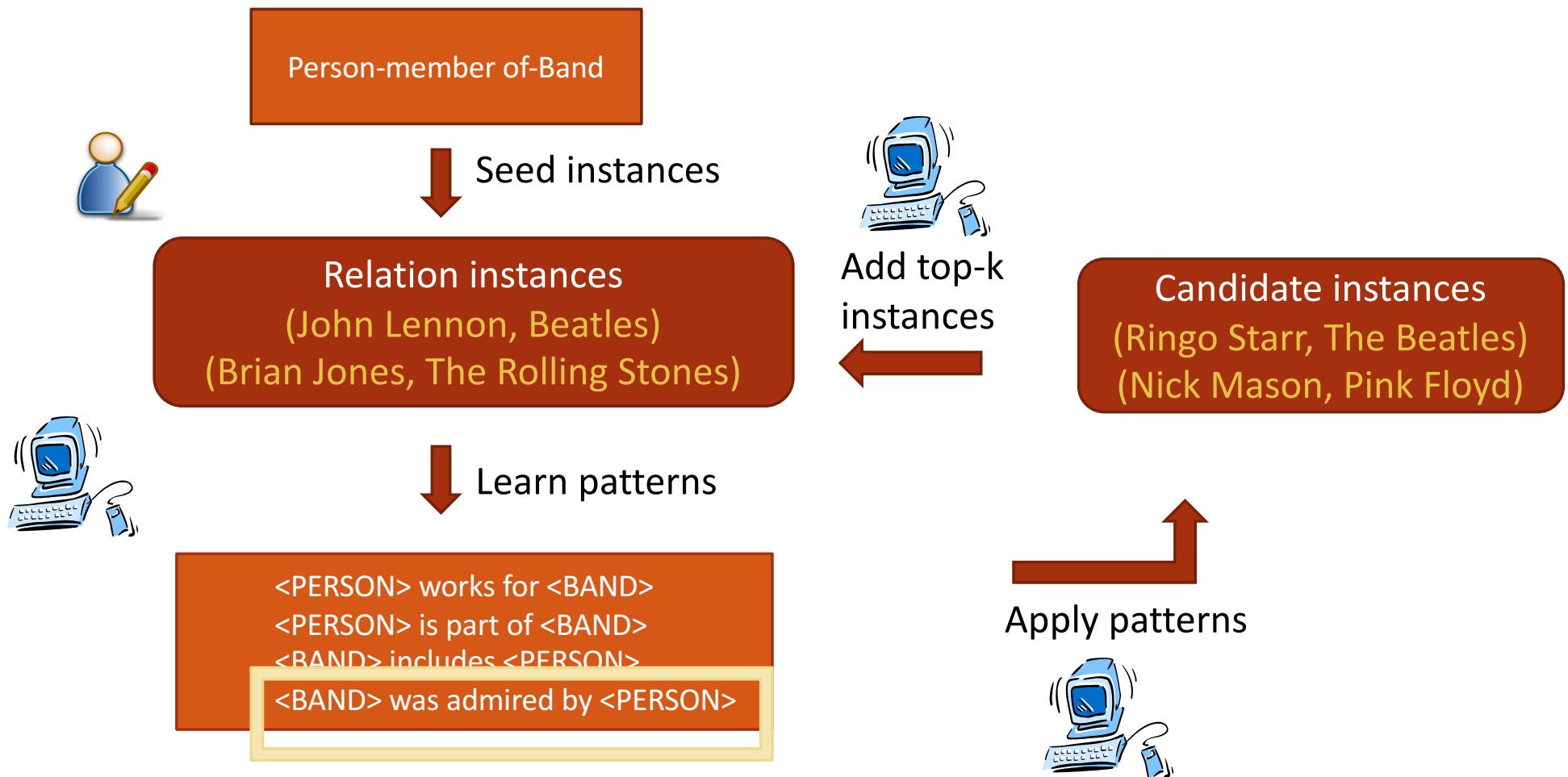
# Learning Extractors: Semi-supervised



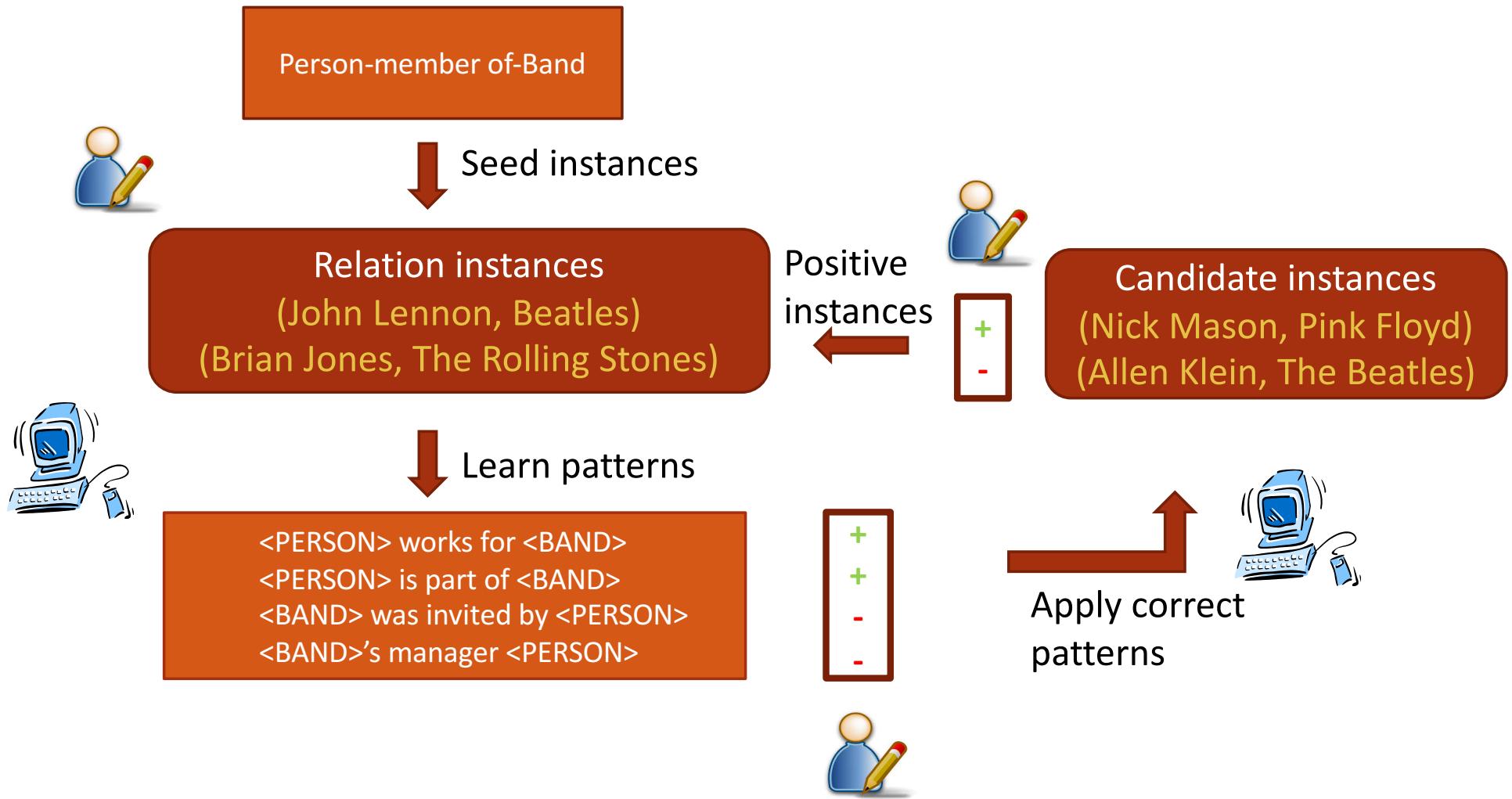
## Bootstrapping



# Learning Extractors: Semi-supervised



# Learning Extractors : Interactive



# Learning Extractors : Unsupervised



- Identify candidate relations:  
for each verb find the longest sequence of words  
s.t. syntactic and lexical constraints are satisfied
- Identify arguments for each relation:  
For each identified relation phrase  $r$ ,  
find the closest noun-phrases on the left and right of  $r$   
satisfying certain syntactic constraints

Syntactic constraint

Regular expressions of POS tags

Lexical constraint

| distinct arguments |  
a relation phrase takes

# Learning Extractors : Unsupervised



Hudson was born in Hampstead, which is a suburb of London.

- e1: (Hudson, was born in, Hampstead)
- e2: (Hampstead, is a suburb of, London)

# Information Extraction

## 3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

## Scoring the extractions



## 3 LEVELS OF SUPERVISION

Supervised



Semi-supervised



Unsupervised



# Scoring the candidate extractions

---

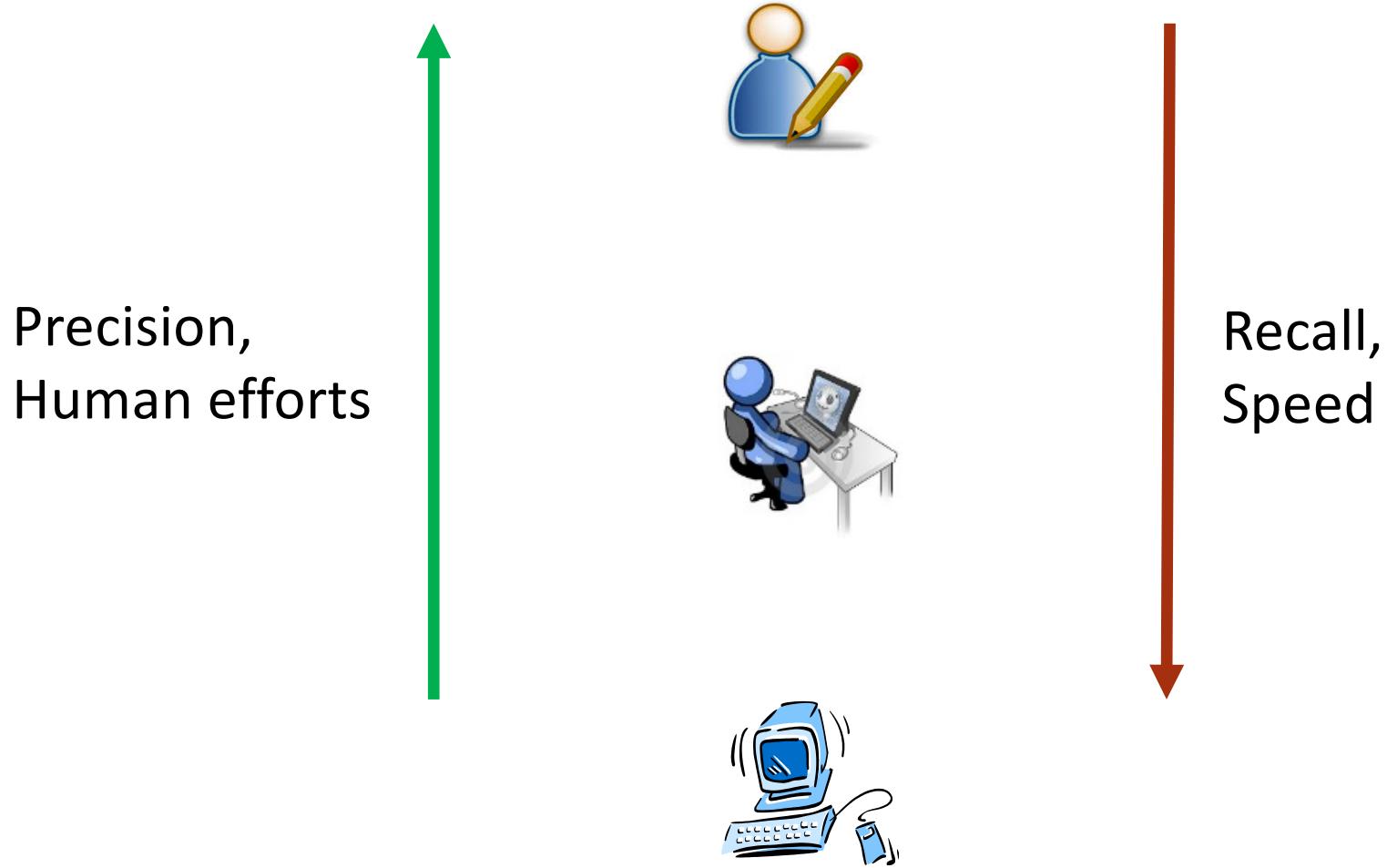


- Human defined scoring function  
(expensive, high precision, low recall)
- Expert comes up with features  
Crowdsourced true/false evaluation of training data  
Scoring function is learnt using standard ML

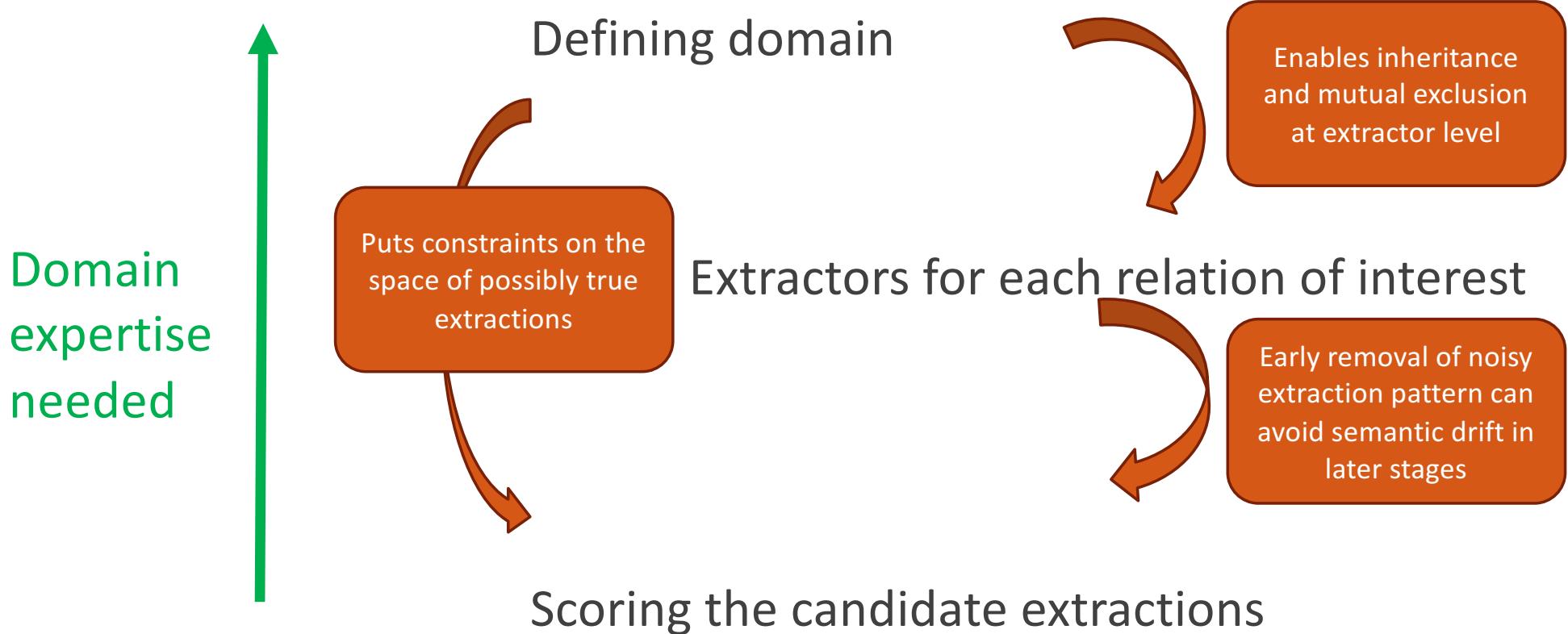


- Completely automatic (Self-training)  
Updated set of instances → weights of extraction patterns → more instances → ....  
(cheap, leads to semantic drift)

# Effect of supervision on extractions



# Impact of early supervision



# Information Extraction

---

3 IMPORTANT SUB-PROBLEMS

CATEGORIES OF IE TECHNIQUES

KNOWLEDGE FUSION

IE SYSTEMS IN PRACTICE

# Categories of IE Techniques

## 3 concrete sub-problems

Defining domain

Learning extractors

Scoring the extractions



## 3 levels of supervision

Supervised



Semi-supervised



Unsupervised



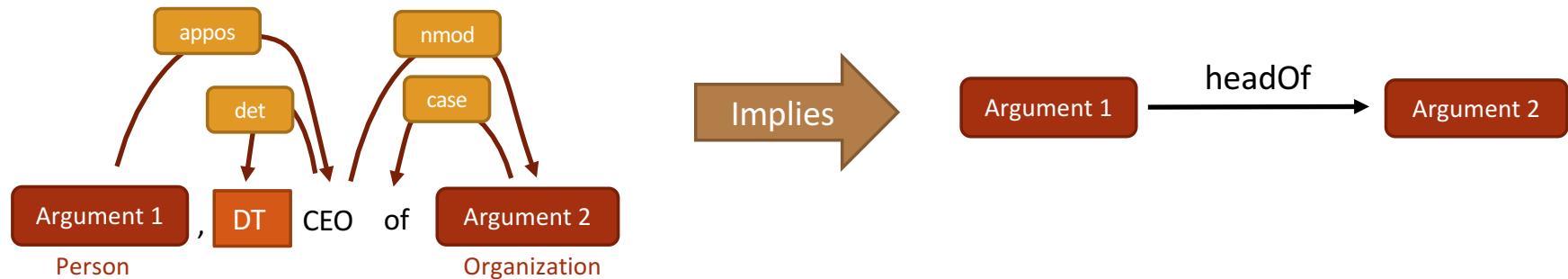
# Categories of IE Techniques

---

1. Narrow domain patterns
2. Ontology based extraction
3. Interactive extraction
4. Open domain IE
5. Hybrid approach (Adding structure to OpenIE KB)

# (1) Narrow domain patterns

Use a collection of rules as the system itself



**High precision:** when it fires, it's correct  
Easy to explain predictions  
Easy to fix mistakes

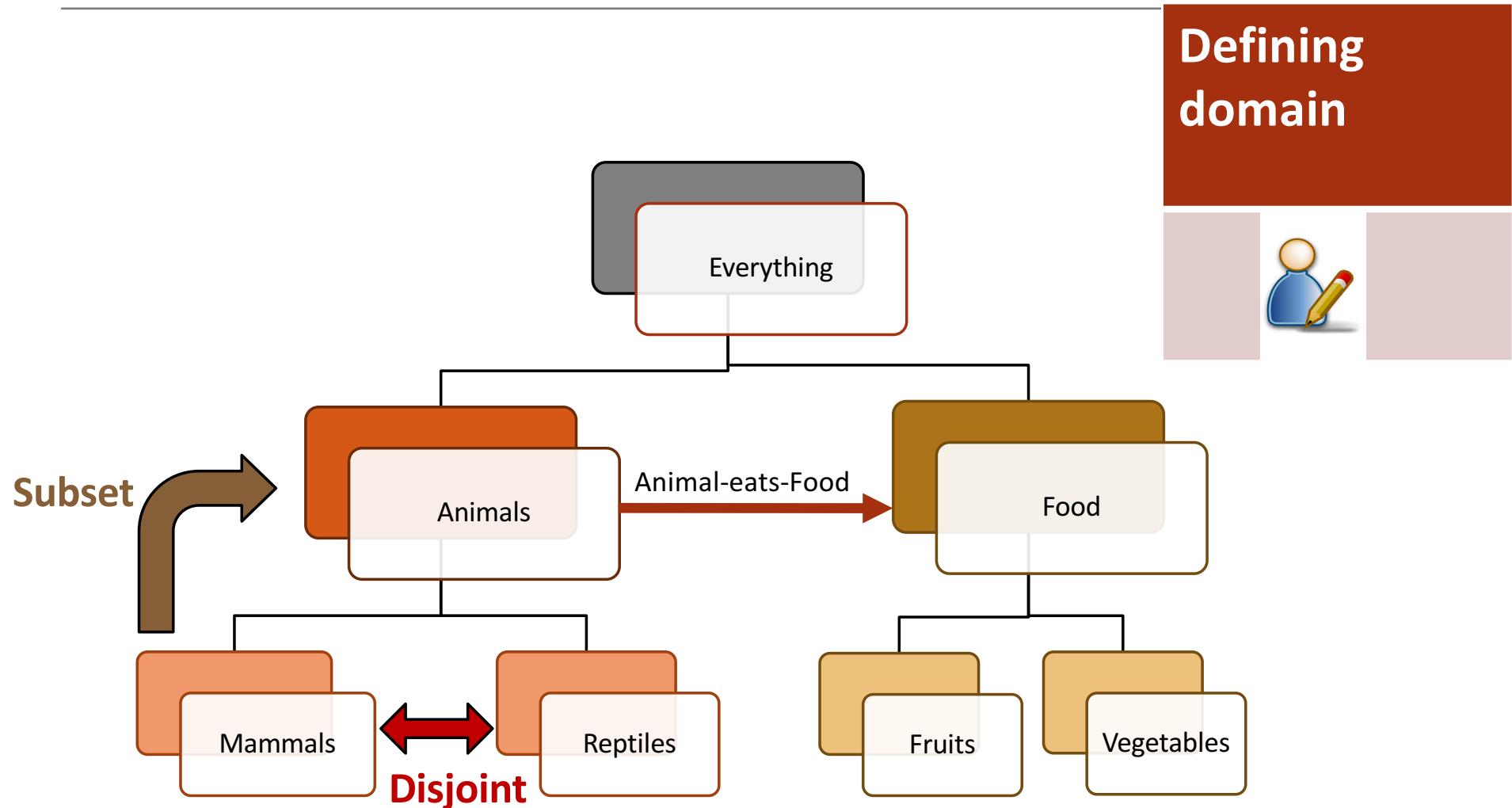
However...  
Only work when the rules fire  
**Poor recall:** Do not generalize!

# (1) Narrow domain patterns

---

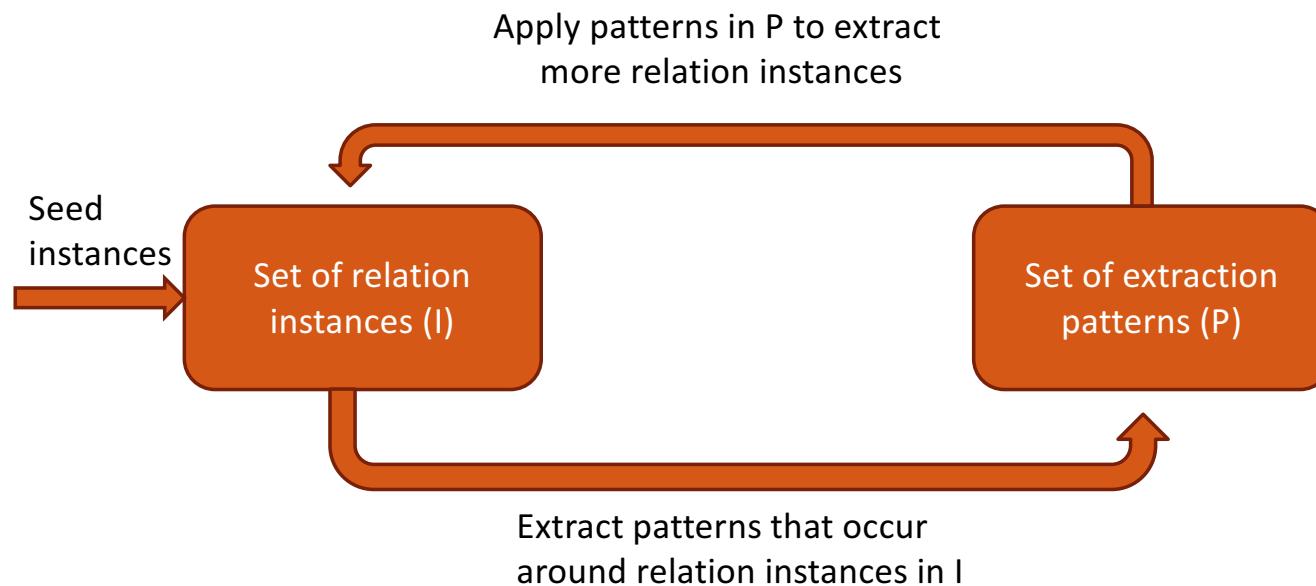
Defining domain	Learning extractors	Scoring extractions
		

# (2) Ontology based extraction

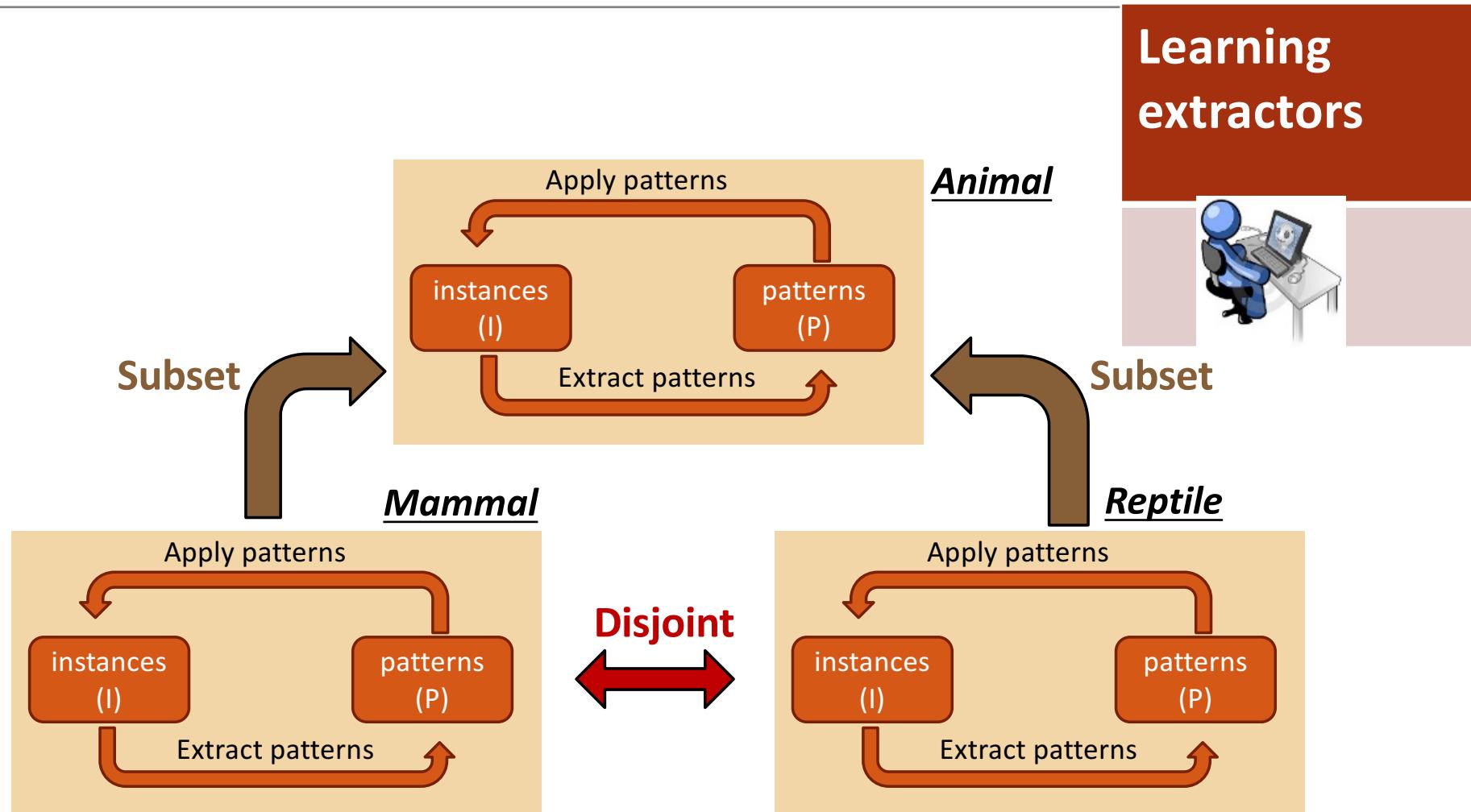


# Semi-supervised learning (bootstrapping)

---

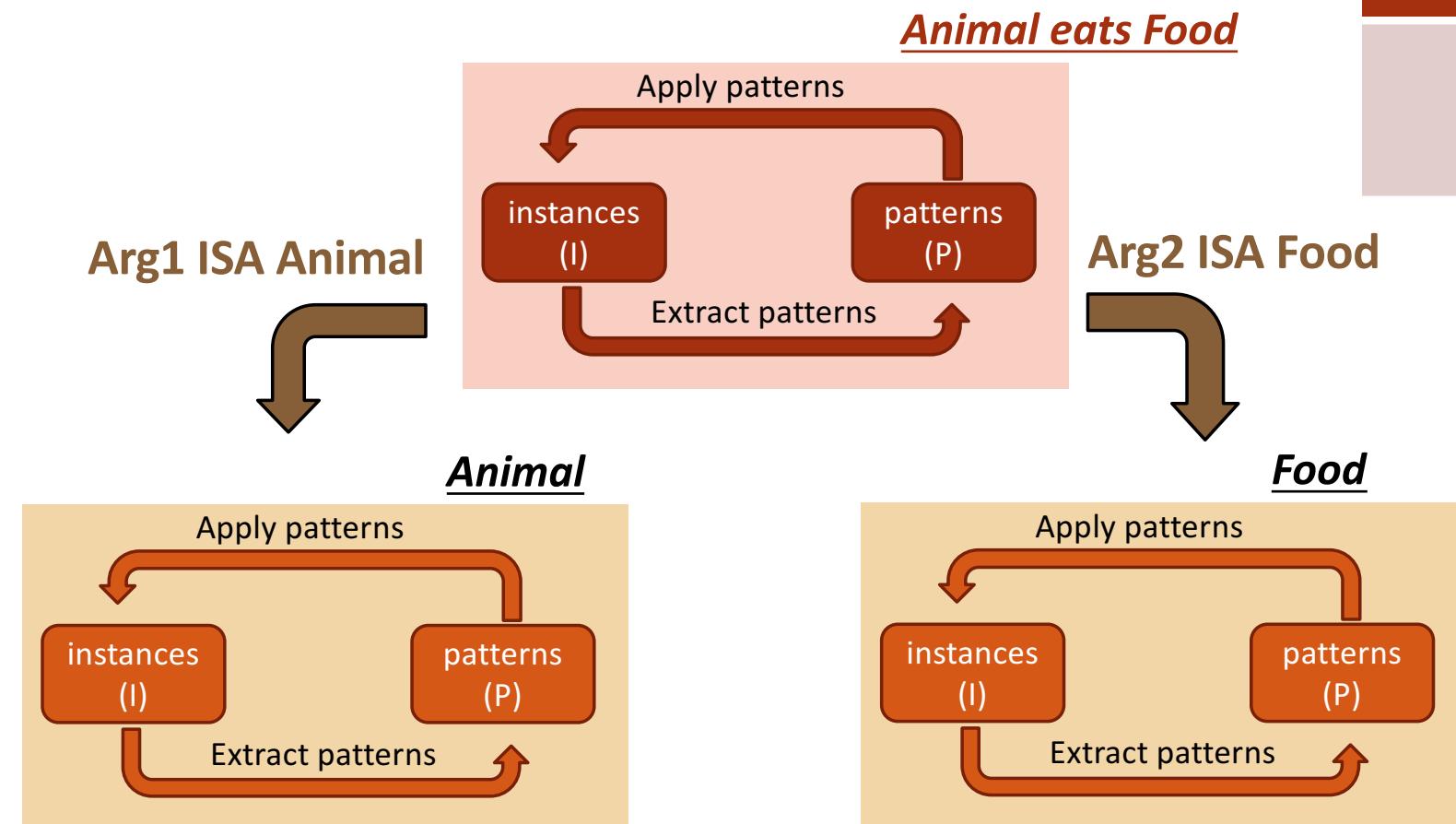


# Coupled bootstrap learning



# Coupled bootstrap learning

Scoring  
extractions



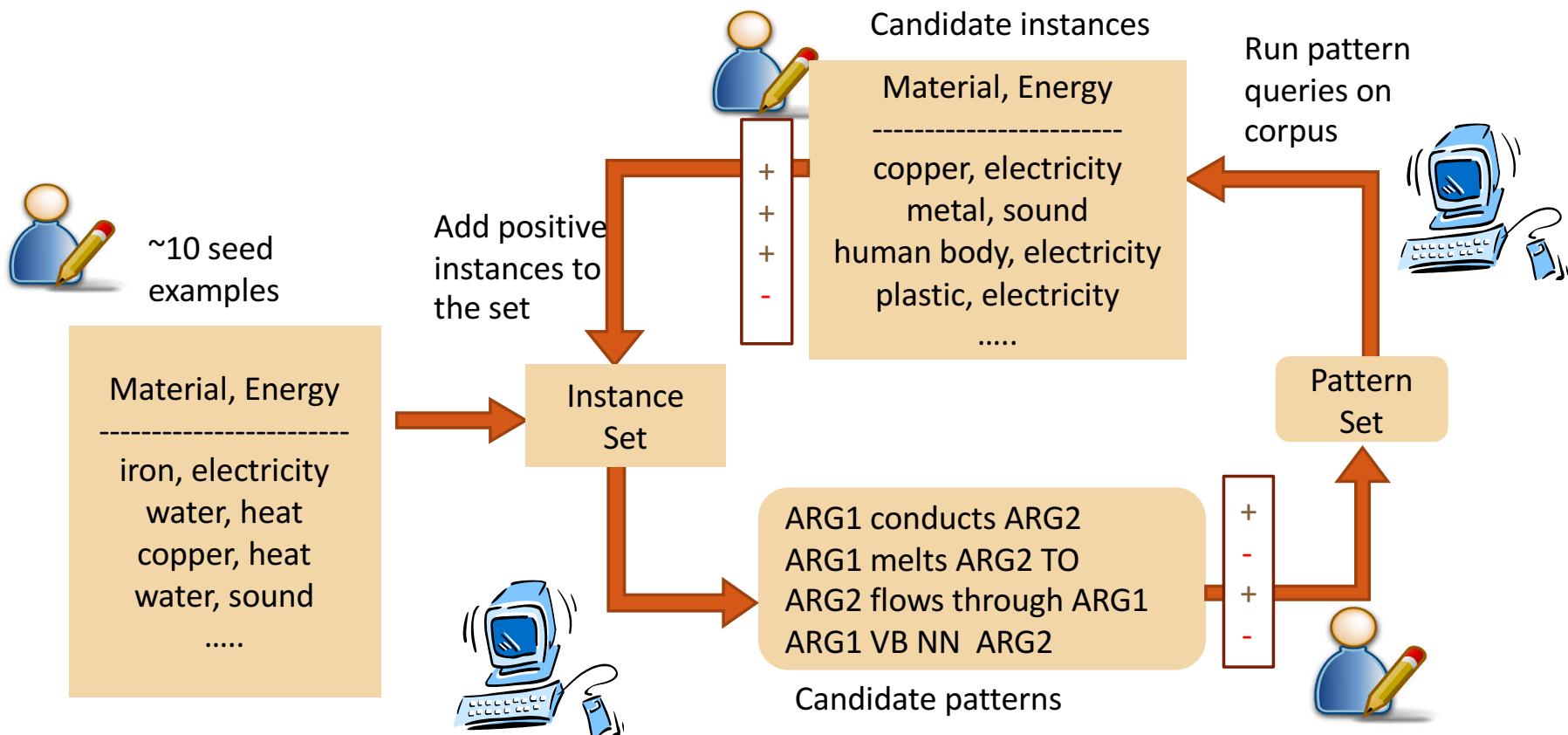
# (2) Ontology based extraction

---

Defining domain	Learning extractors	Scoring extractions
		

# (3) Interactive Extraction

## Interactive Bootstrapping



# (3) Interactive Extraction

---

Defining domain	Learning extractors	Scoring extractions
		

# (4) Open domain IE

Open domain  
any NP is a candidate entity  
Any VP is a candidate relation



Hudson was born in Hampstead, which is a suburb of London.

Scoring based on classifier  
(features: POS tags,  
dependency parse ...)

(Hudson, was born in, Hampstead) : 0.88  
(Hampstead, is a suburb of, London) : 0.9

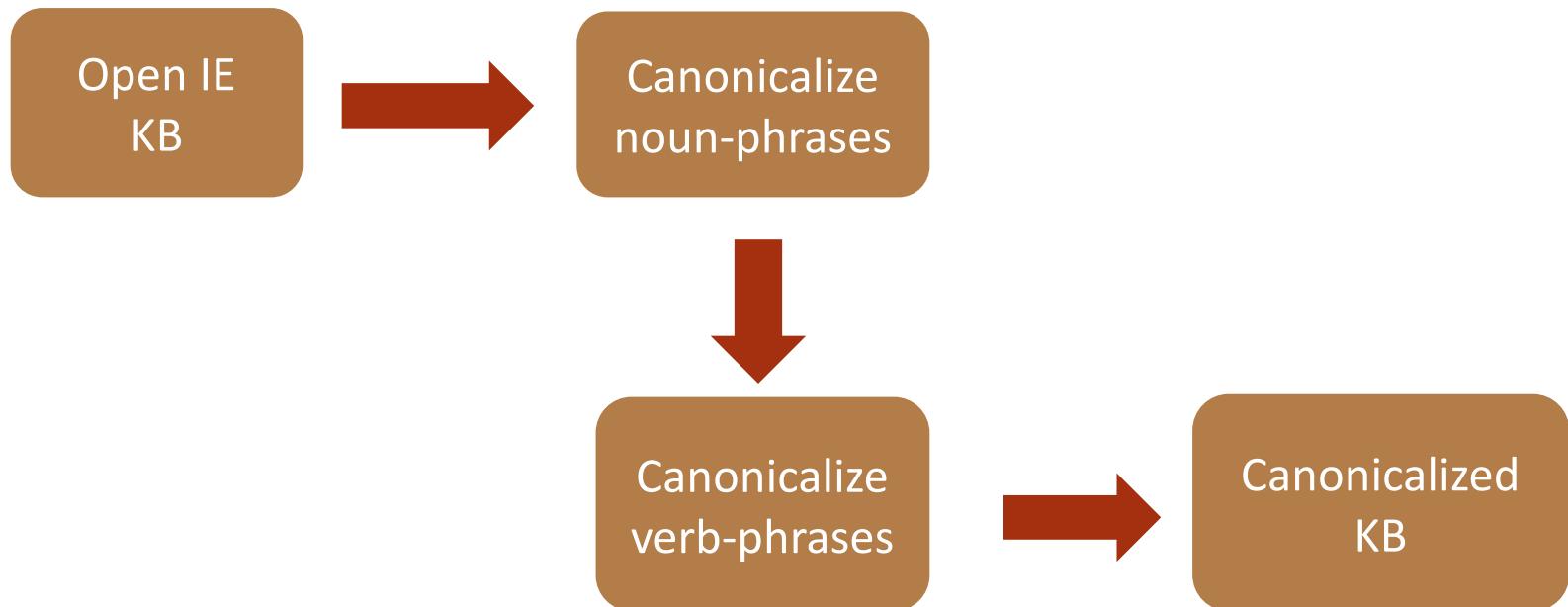
# (4) Open domain IE

---

Defining domain	Learning extractors	Scoring extractions
		

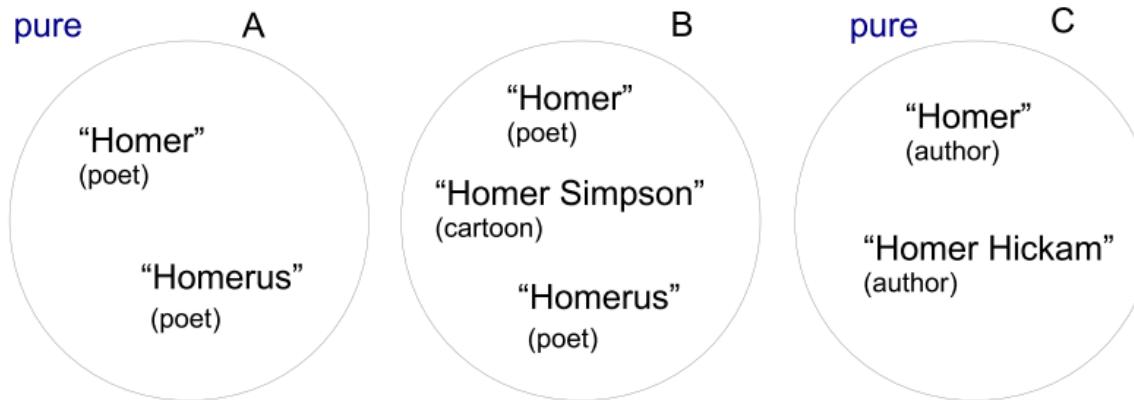
# (5) Hybrid approach (adding structure to Open IE KB)

---



# (5) Hybrid approach

- *Canonicalizing noun phrases*

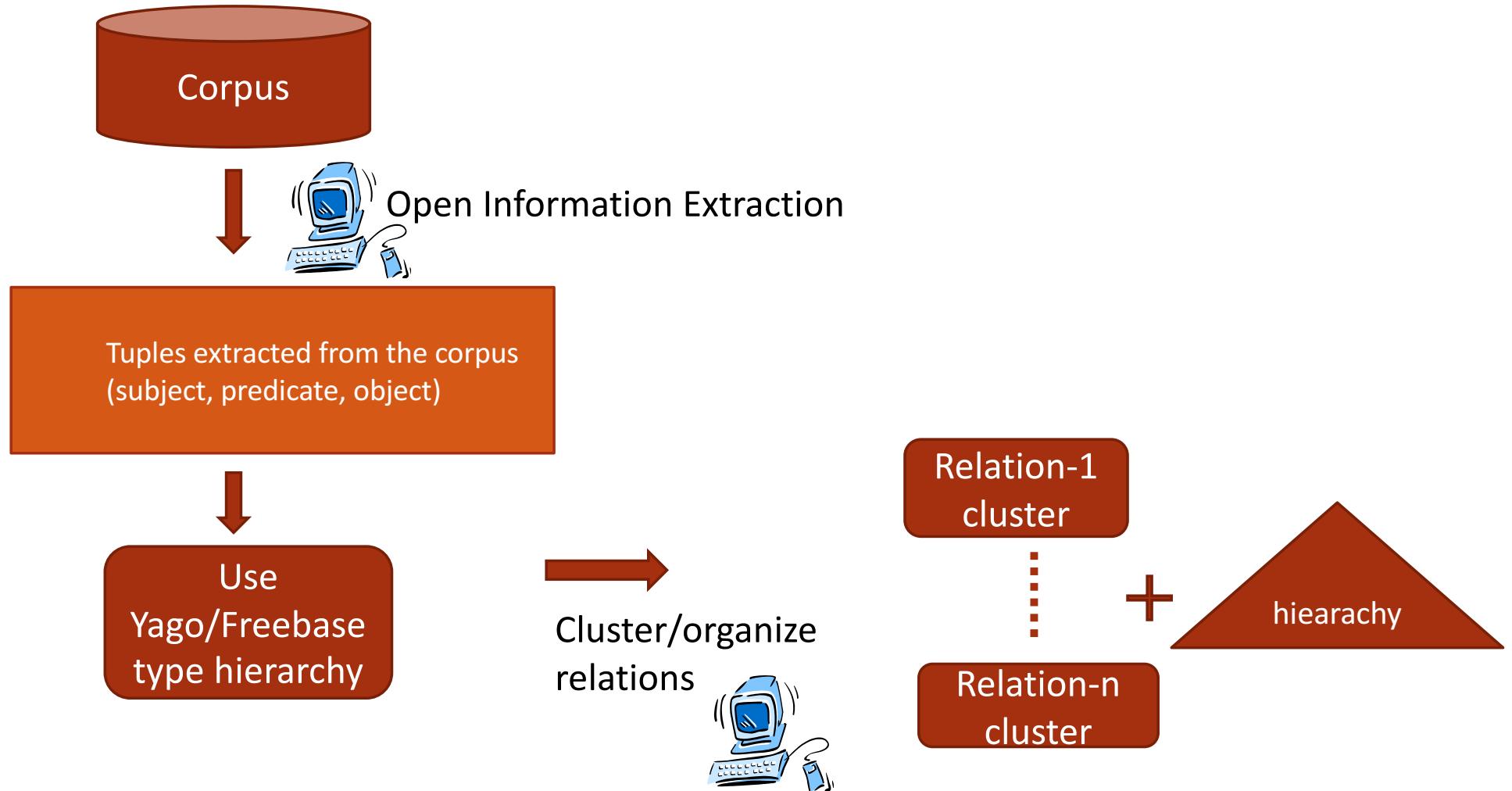


- *Canonicalizing verb phrases*

Verb phrases	Freebase relation
be an abbreviation-for, be known as, stand for, be an acronym for be spoken in, be the official language of, be the national language of be bought, acquire	- location.country.official_language organization.organization.acquired_by

Clustering based on  
statistical features

# (5) Hybrid approach



# (5) Hybrid approach

---

Defining domain	Learning extractors	Scoring extractions
		

# Information Extraction

---

3 IMPORTANT SUB-PROBLEMS

CATEGORIES OF IE TECHNIQUES

**KNOWLEDGE FUSION**

IE SYSTEMS IN PRACTICE

# Knowledge fusion

## Single extractor

Defining domain

Learning extractors

Scoring the extractions

Manual



Semi-automatic



Automatic



## Fusing multiple extractors

# Knowledge fusion with multiple extractors

---

- Voting (and vs or of extractors)
- Co-training (multiple extraction methods)
- Multi-view learning (multiple data sources)
- Classifier

# Multiple extractors

---

- **Extractor 1:** text patterns to extract ISA relations  
e.g. coupled pattern learner
- **Extractor 2:** learning wrappers for HTML pages to extract ISA relations from structured text

# (1) Voting Schemes

---

- ***AND of two extractors:***

- For a candidate extraction to be promoted to a fact in KB, both the extractors should support the fact
- $\text{score}(\text{fact}) = \text{Min}(\text{score\_extractor1}(\text{fact}), \text{score\_extractor2}(\text{fact}))$

- ***OR of two extractors***

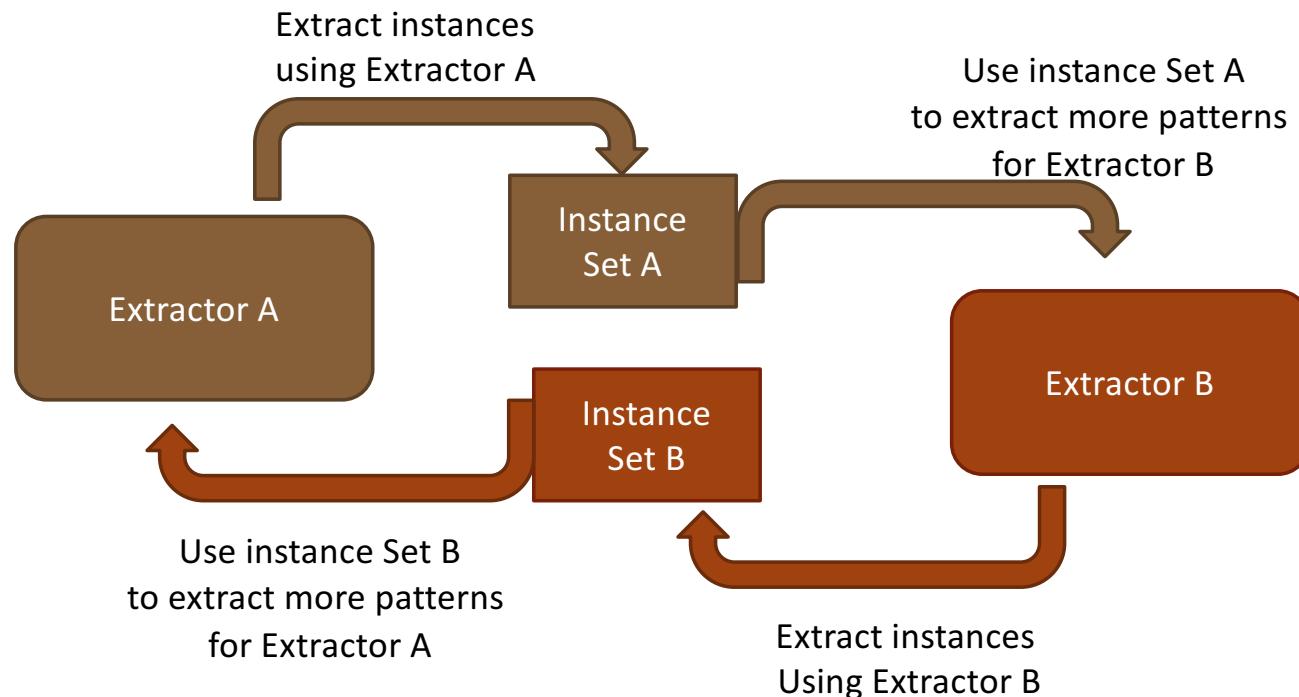
- For a candidate extraction to be promoted to a fact in KB, both the extractors should support the fact
- $\text{score}(\text{fact}) = \text{Max}(\text{score\_extractor1}(\text{fact}), \text{score\_extractor2}(\text{fact}))$

- **Hand-coded heuristic rules**

- E.g. (at least one extractor has confidence > 0.9) or  
(two extractors support the fact with confidence > 0.6)

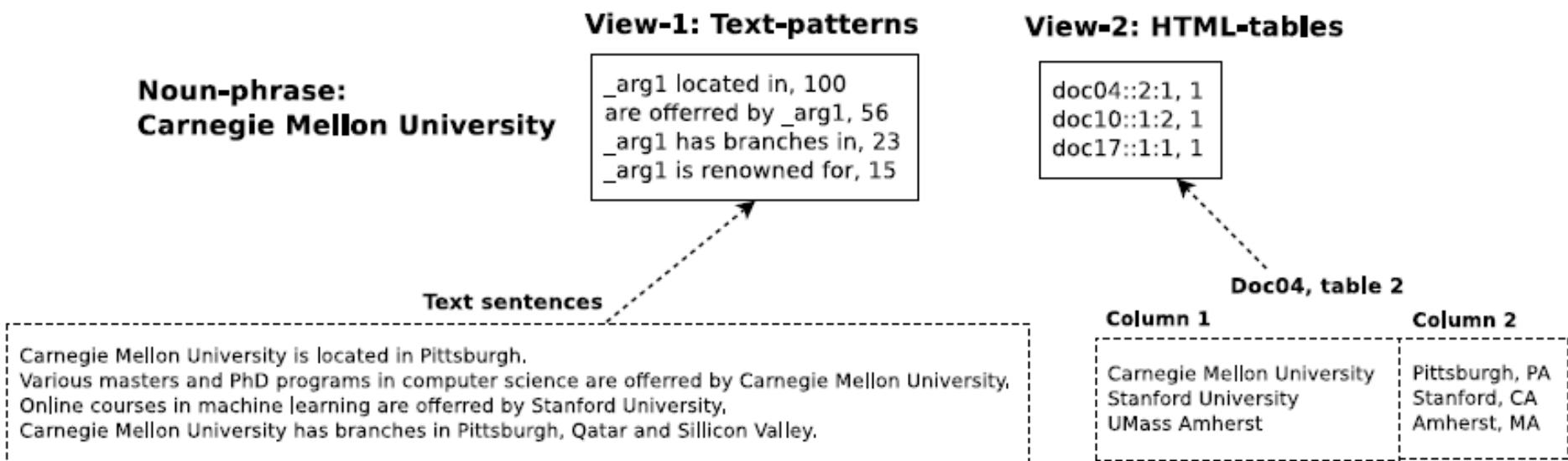
.....

# (2) Co-training

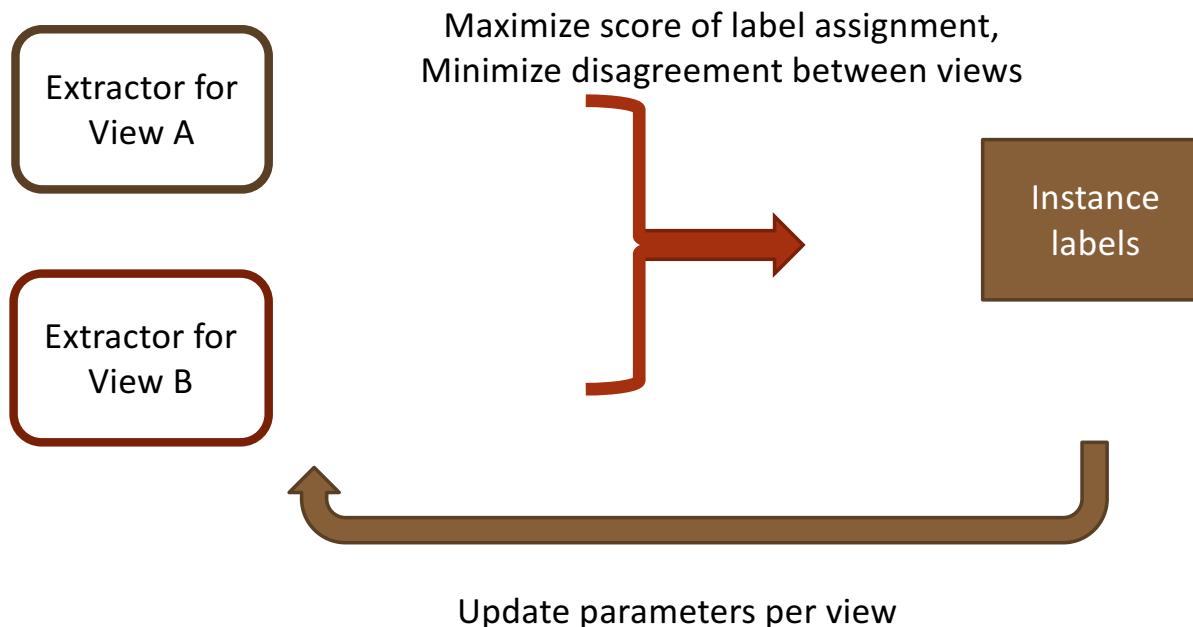


# (3) Multi-view learning

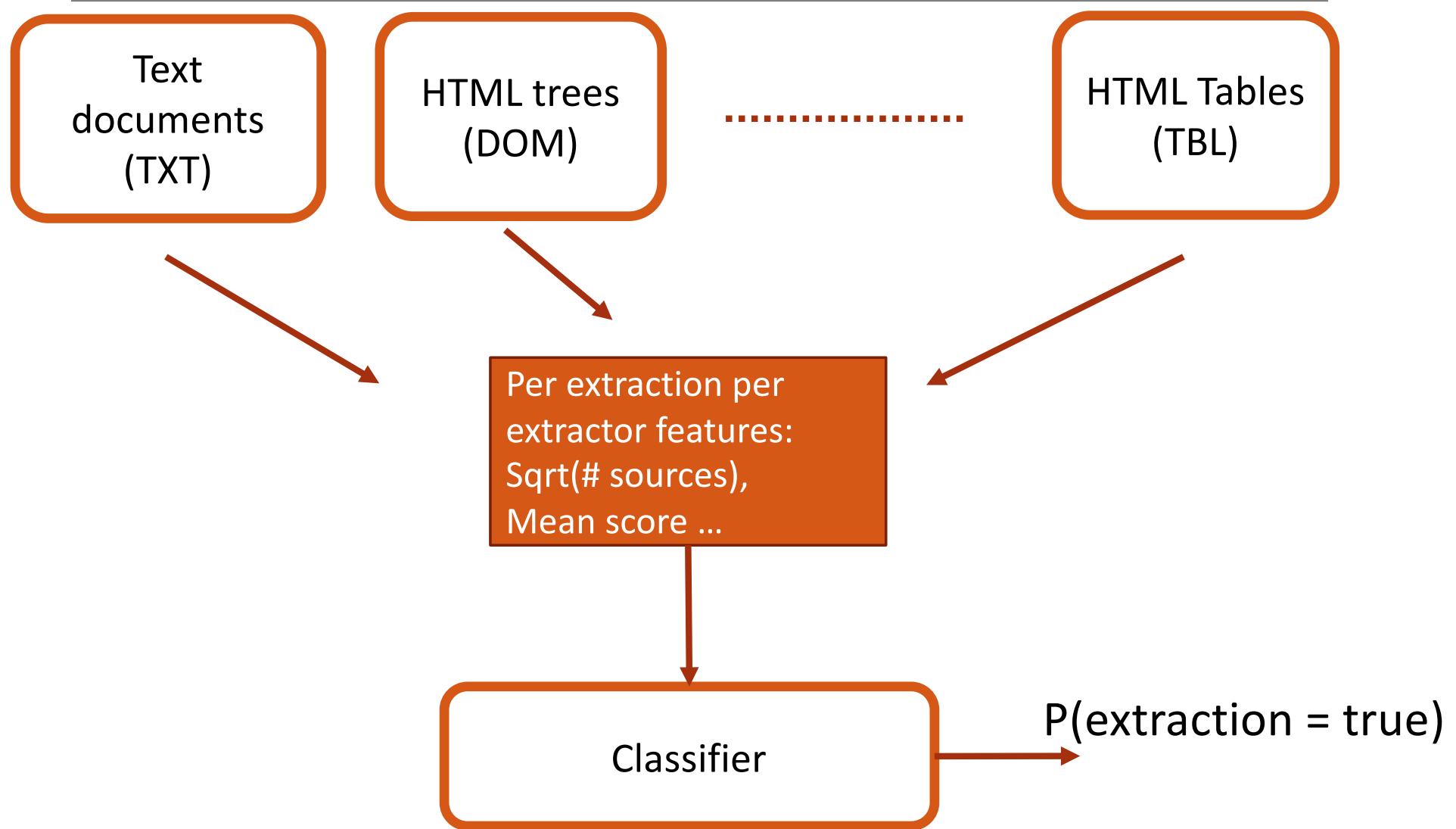
- NP “Carnegie Mellon University” can be represented in two different ways based on its occurrence in text documents and HTML tables.



# (3) Multi-view learning



# (4) Classifiers



# Information Extraction

---

3 IMPORTANT SUB-PROBLEMS

CATEGORIES OF IE TECHNIQUES

KNOWLEDGE FUSION

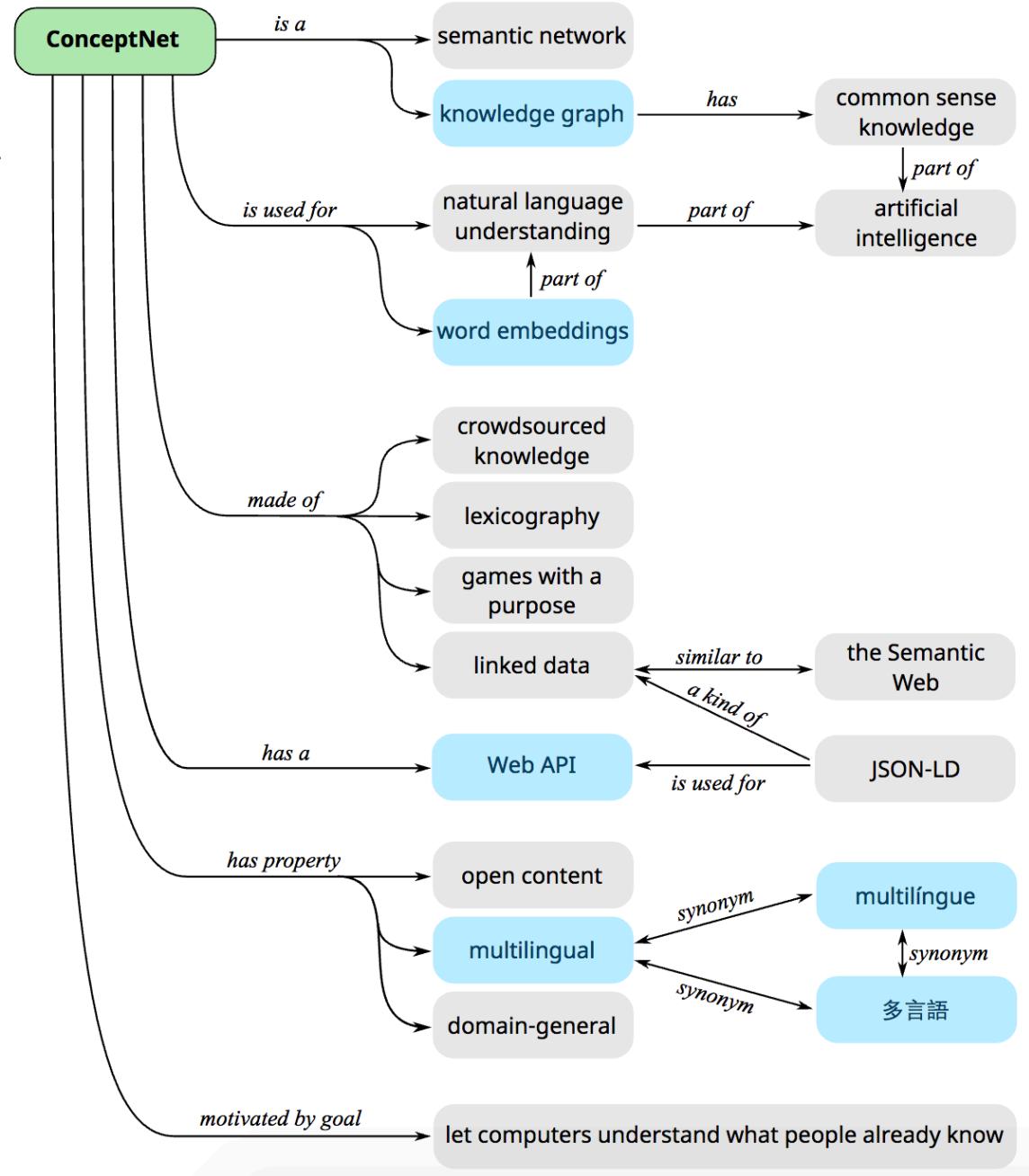
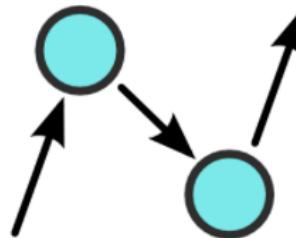
**IE SYSTEMS IN PRACTICE**

# IE systems in practice

---

- Conceptnet
- NELL
- Knowledge vault
- Open IE

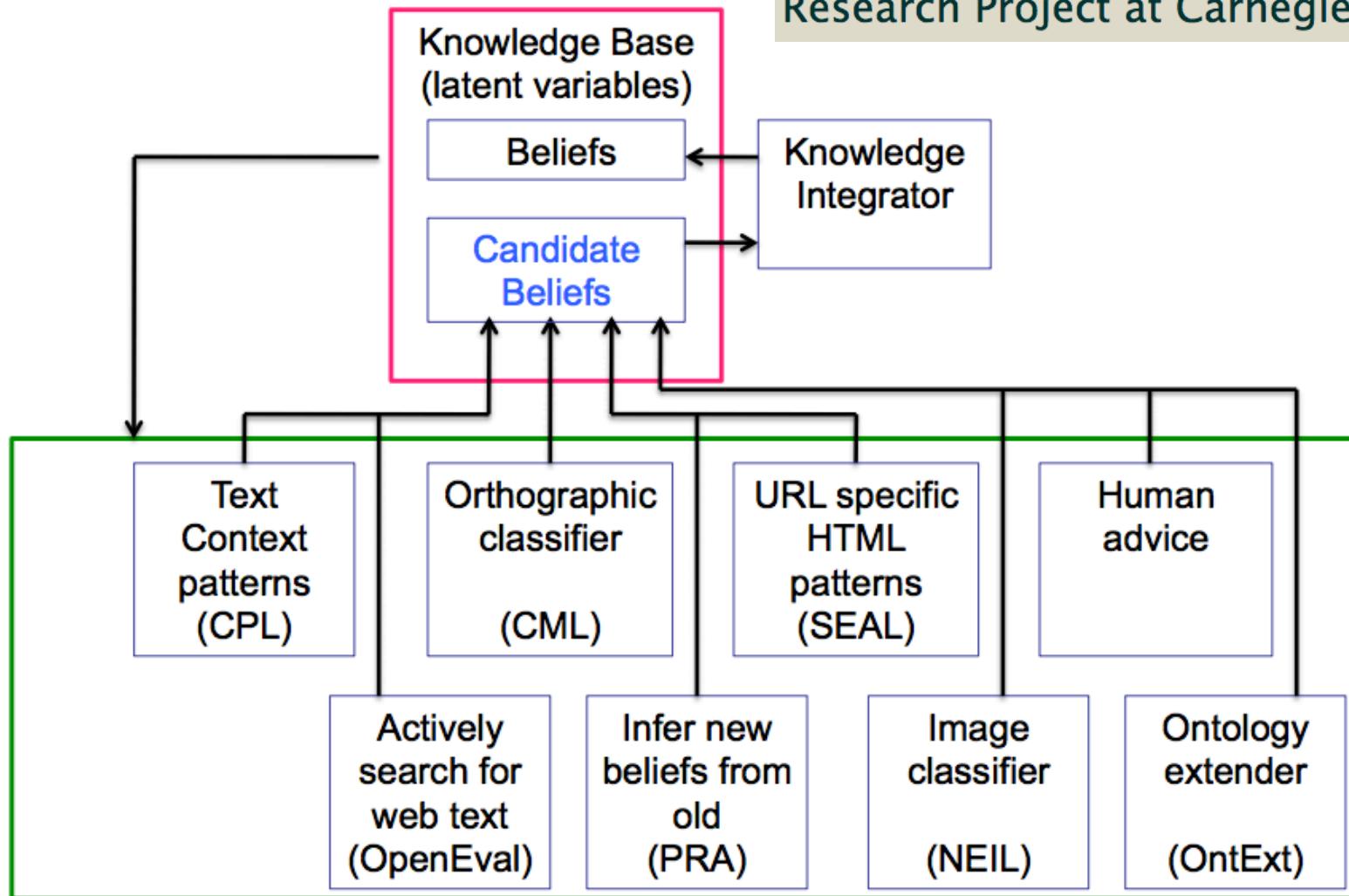
# ConceptNet

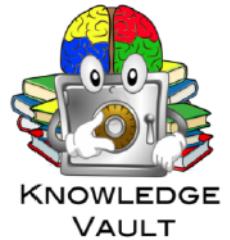


# Never Ending Language Learning (NELL)

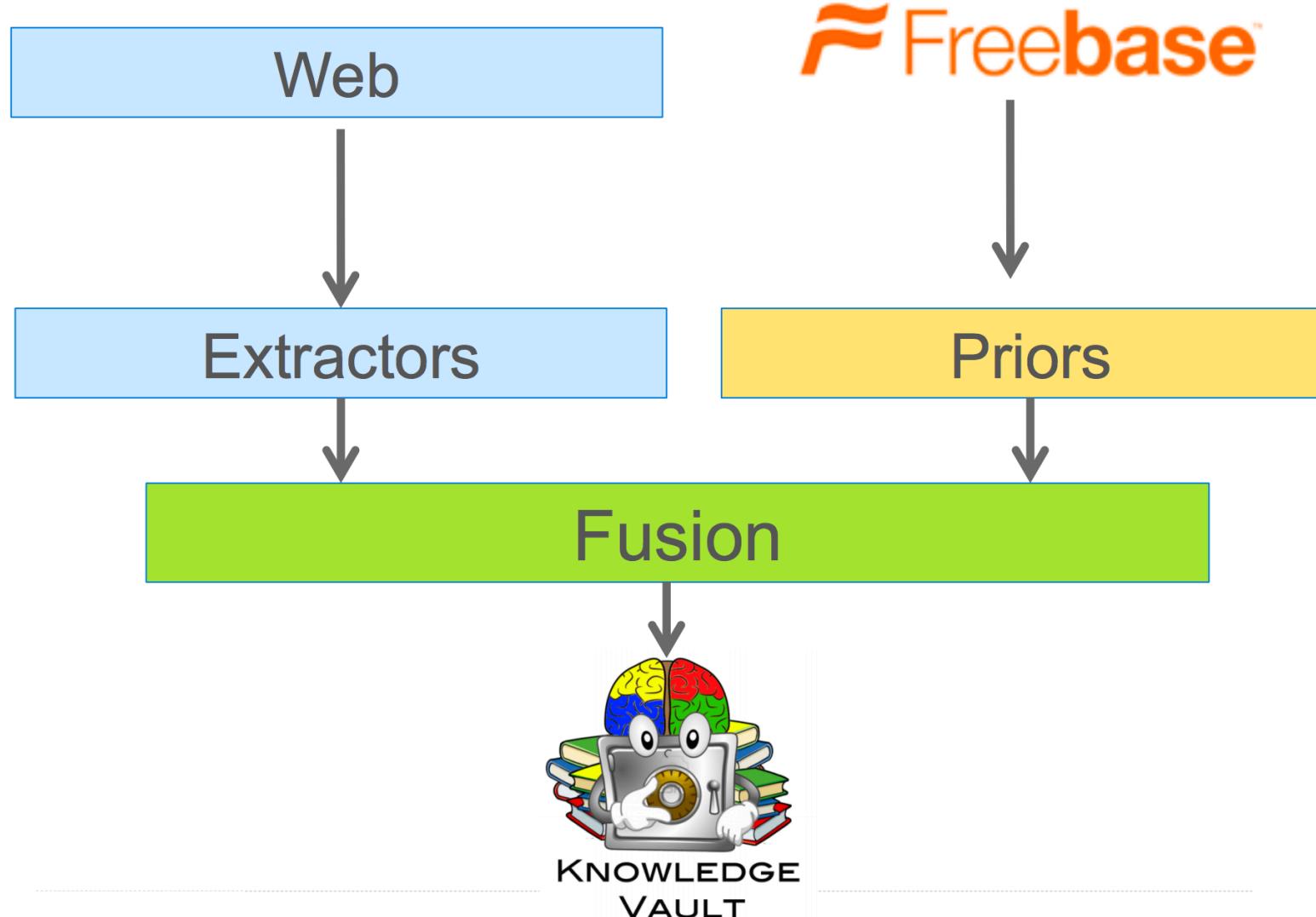
## Read the Web

Research Project at Carnegie Mellon University





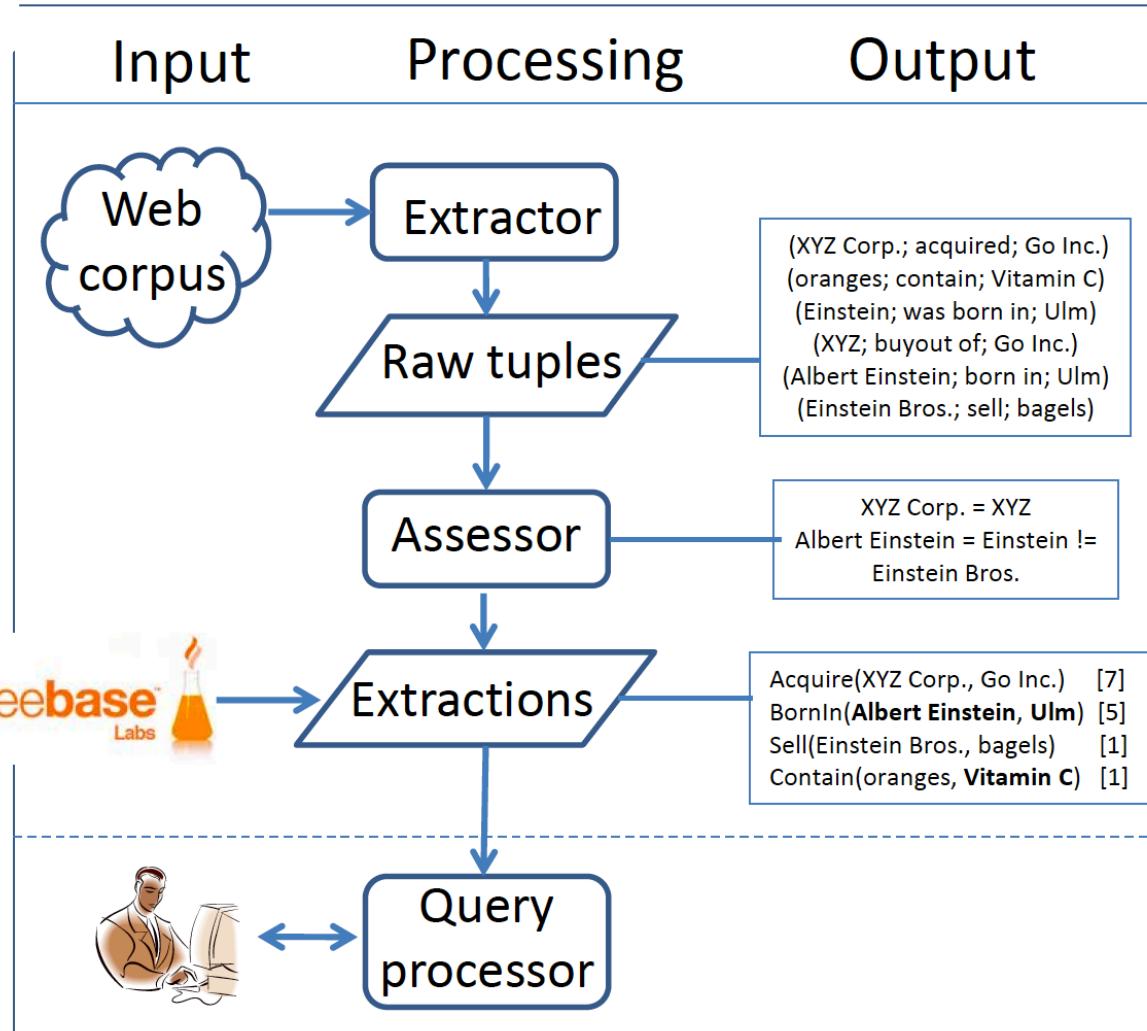
# Knowledge Vault



# Open IE (KnowItAll)



## Open Information Extraction



Relation-independent extraction

Synonyms, Confidence

Index in Lucene;  
Link entities

# IE systems in practice

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

# 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] 

# Thank You

---



SEE YOU AFTER THE COFFEE BREAK!

