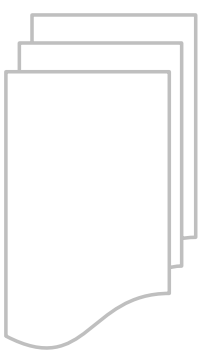
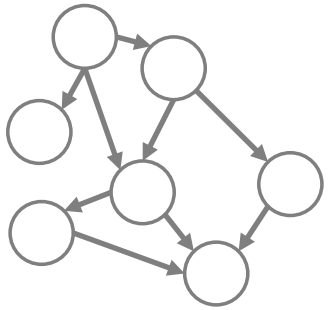
Part 1: Knowledge Graphs



Part 2: Knowledge Extraction

Part 3: Graph Construction



Part 4: Critical Analysis

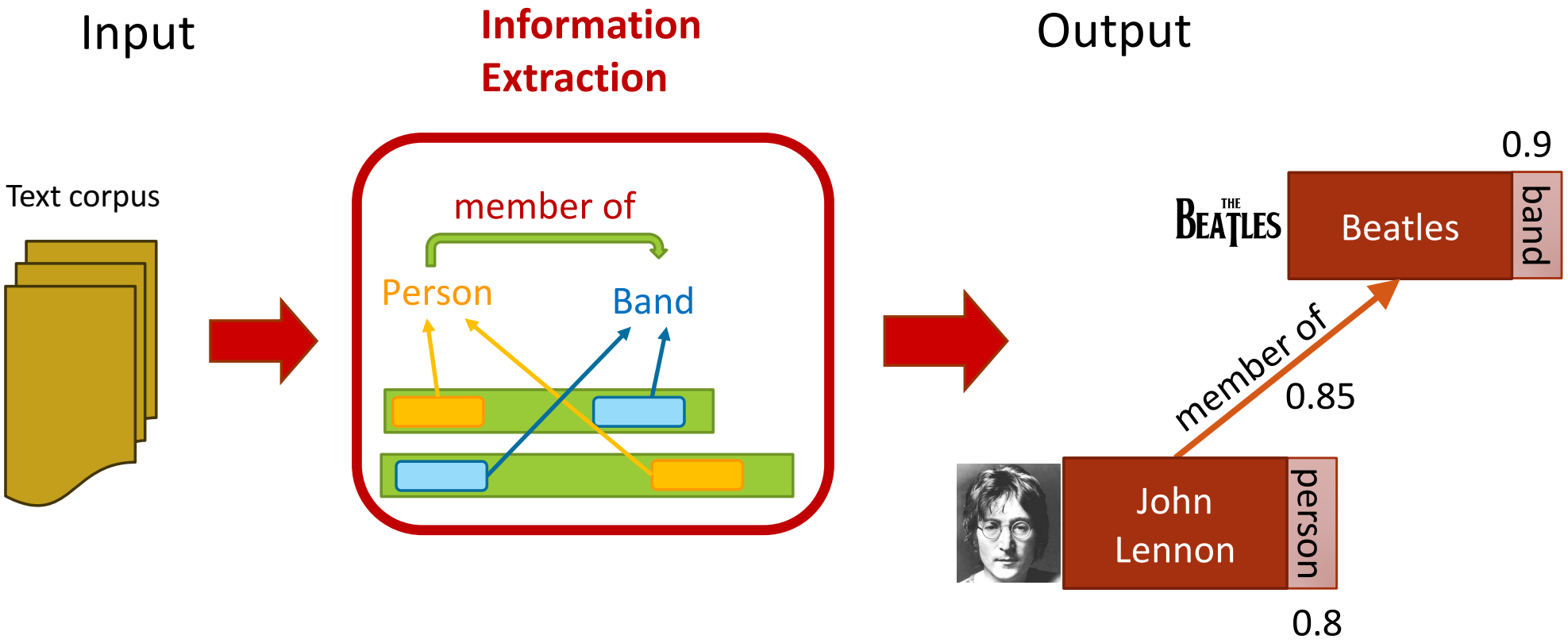
Tutorial Outline

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] 

Information Extraction



Information Extraction

3 IMPORTANT SUB-PROBLEMS

(DEFINE DOMAIN, LEARN EXTRACTORS, SCORE EXTRACTIONS)

3 LEVELS OF SUPERVISION

(MANUAL, SEMI-SUPERVISED, UNSUPERVISED)

KNOWLEDGE FUSION WITH MULTIPLE EXTRACTORS

(CO-TRAINING, MULTI-VIEW LEARNING)

EXAMPLE IE SYSTEMS

Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring the extractions

×

3 LEVELS OF SUPERVISION

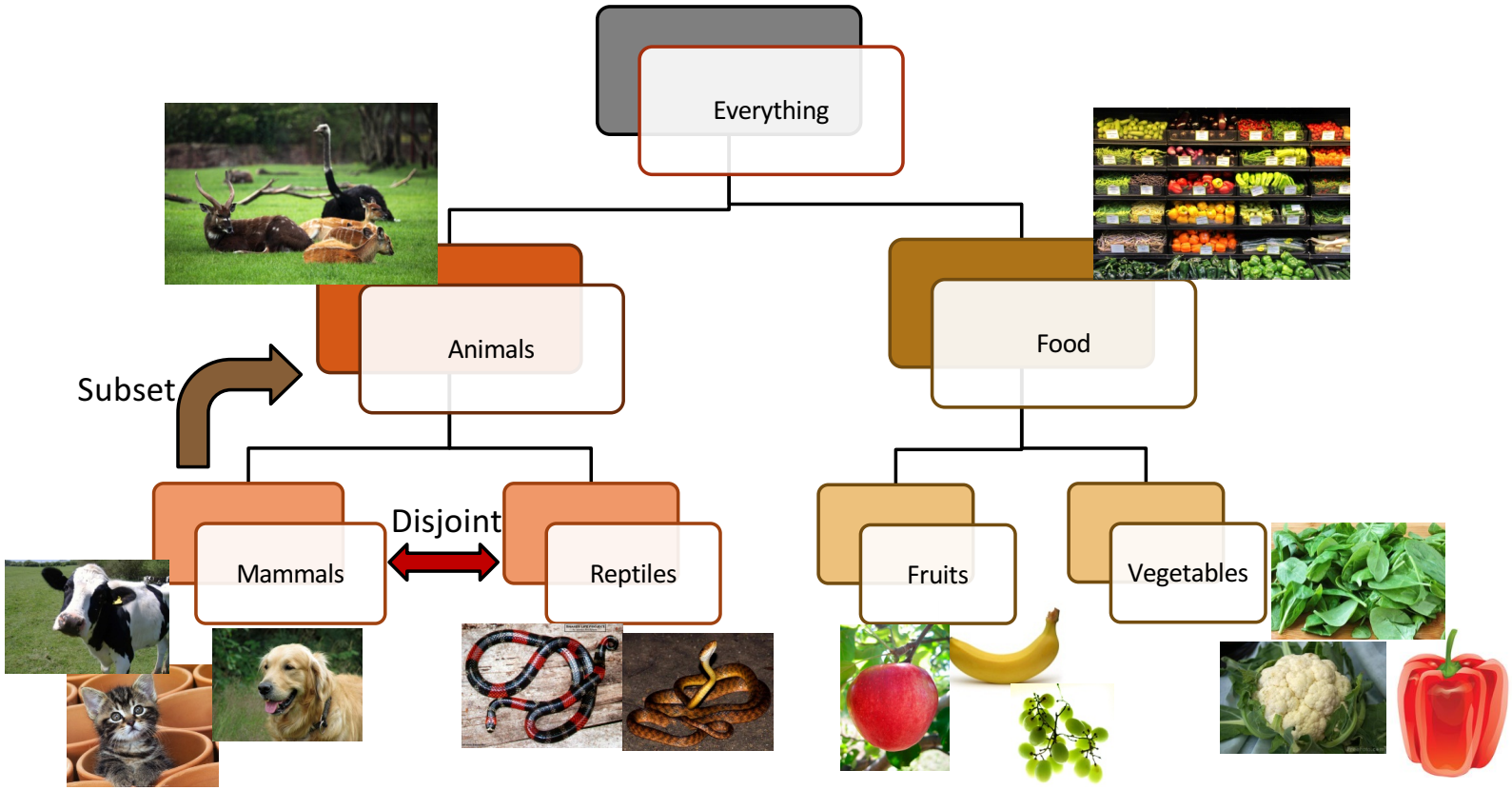
Manual

Semi-automatic

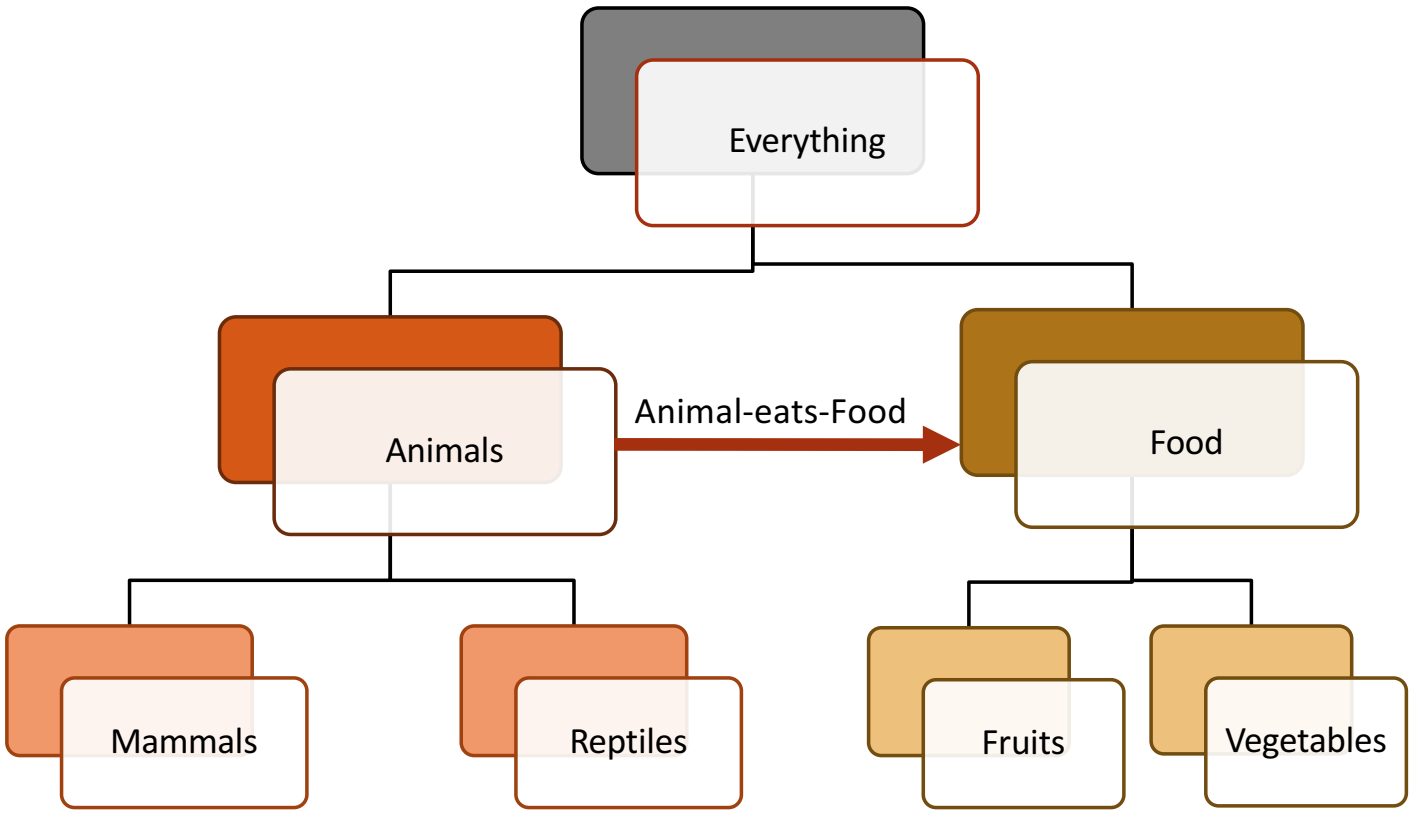
Automatic

Defining domain: types/relations of interest

Defining Domain: Manual



Defining Domain: Manual

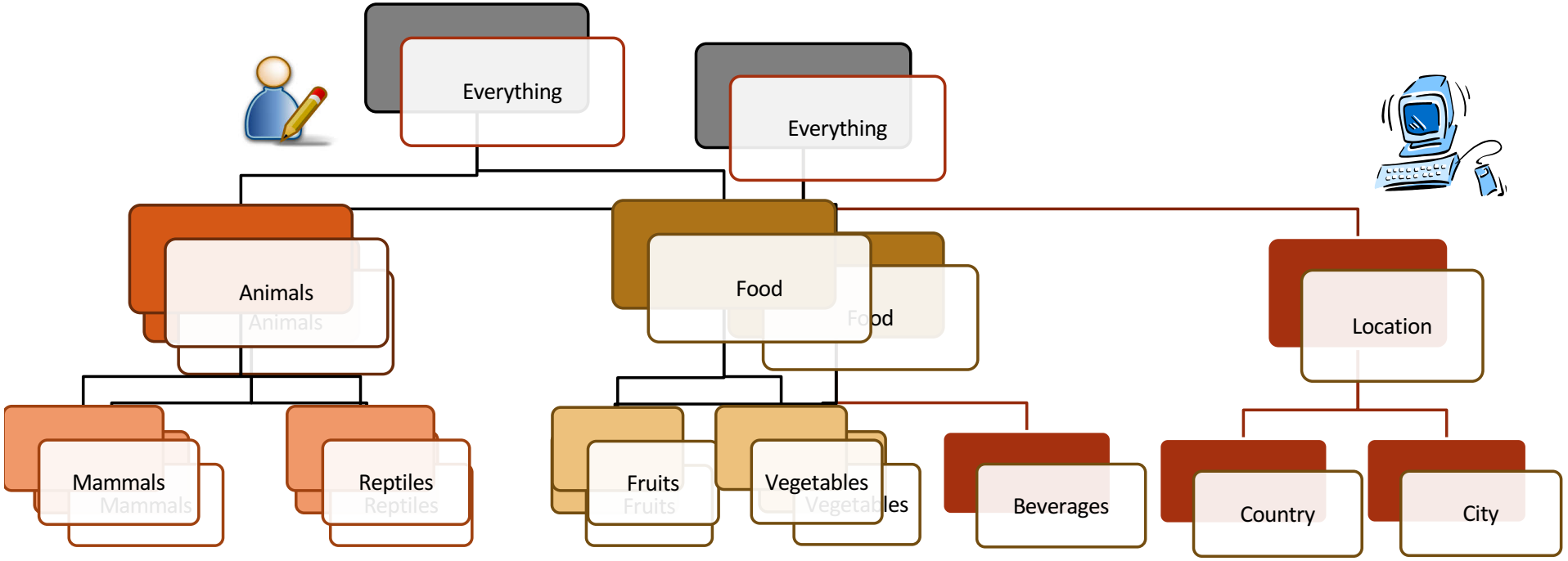


- **Highly semantic ontology**
- **Leads to high precision extractions**
- **Expensive to create**
- **Requires domain experts**

[Toward an Architecture for Never-Ending Language Learning, Carlson et al. AAAI 2010]

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

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"

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

[Discovering Relations between Noun Categories, Mohamed et al., EMNLP 2011]

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/no expert annotations needed**
- **Limited semantics**
- **Leads to noisy extractions**

*[**Open Information Extraction from the Web**, Banko et al., IJCAI 2007]*

Extractors for each relation of interest

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

Person-member of-Band

Seed instances

Add top-k instances

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

Candidate instances
(Ringo Starr, The Beatles)
(Nick Mason, Pink Floyd)

Learn patterns

<PERSON> works for <BAND>
<PERSON> is part of <BAND>
<BAND> includes <PERSON>
<BAND>'s manager <PERSON>

Apply patterns

Learning Extractors : Interactive

Person-member of-Band

Seed instances

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

Positive instances

+
-

Candidate instances
(Nick Mason, Pink Floyd)
(Allen Klein, The Beatles)

Learn patterns

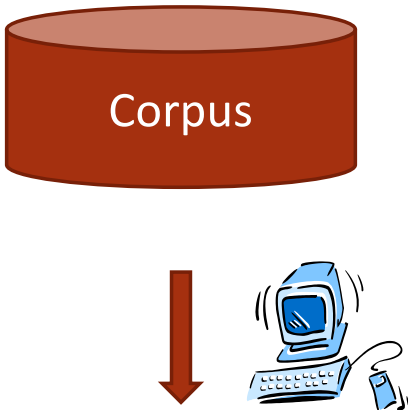
<PERSON> works for <BAND>
<PERSON> is part of <BAND>
<BAND> was invited by <PERSON>
<BAND>'s manager <PERSON>

+
+
-
-

Apply correct patterns

*[**Open information extraction to KBP relations in 3 hours**, Soderland et al., TAC KBP 2013]*

Learning Extractors : Unsupervised



Scoring the candidate extractions

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

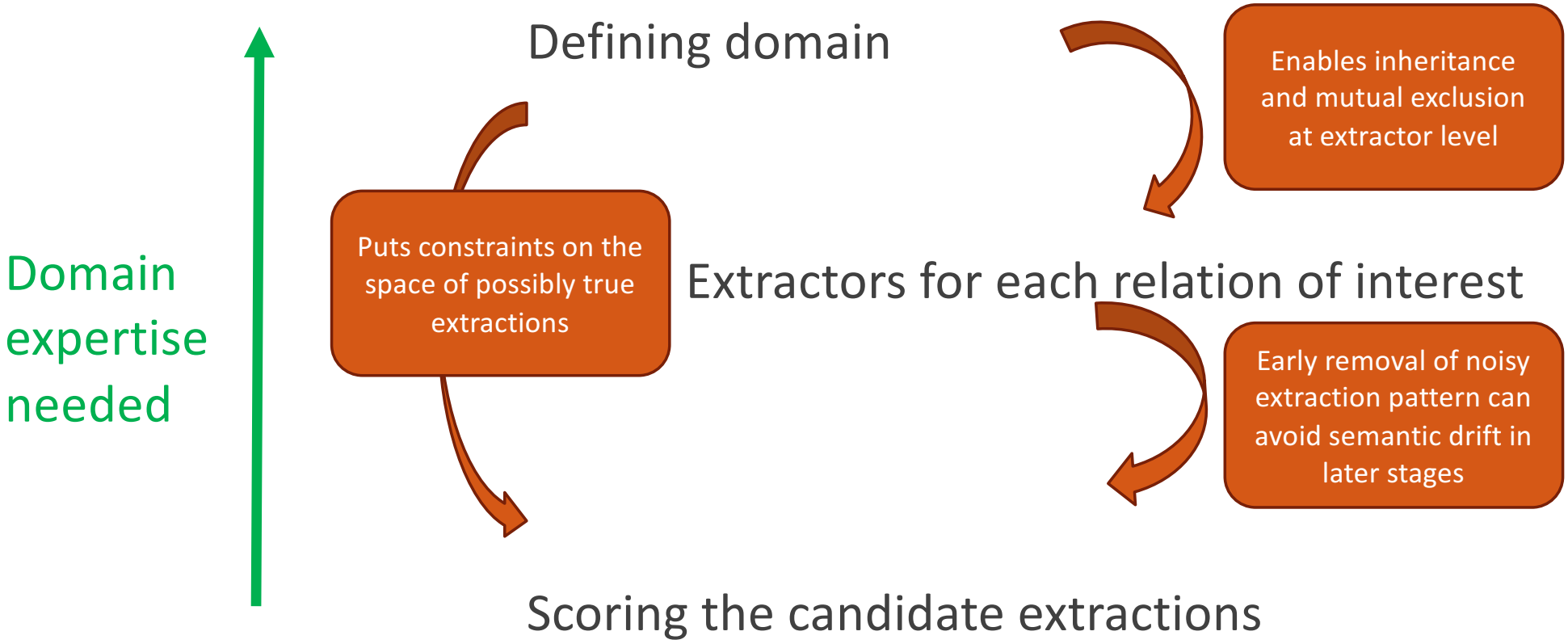
- Completely automatic (Self-training)
  Updated set of instances → weights of extraction patterns → more instances → ….
  (cheap, leads to semantic drift)

Effect of supervision on extractions

Precision,
Human efforts

Recall,
Speed

Impact of early supervision



Example Information Extraction Techniques

3 concrete sub-problems

Defining domain

Learning extractors

Scoring the extractions

×

3 levels of supervision

Manual

Semi-automatic

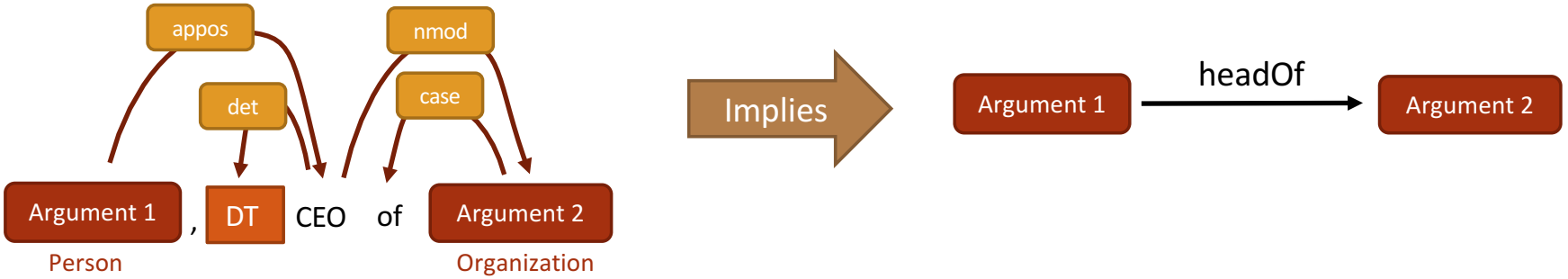
Automatic

(1) Narrow domain patterns

Defining domain	Learning extractors	Scoring extractions
		

(1) Narrow domain patterns

Use a collection of rules as the system itself



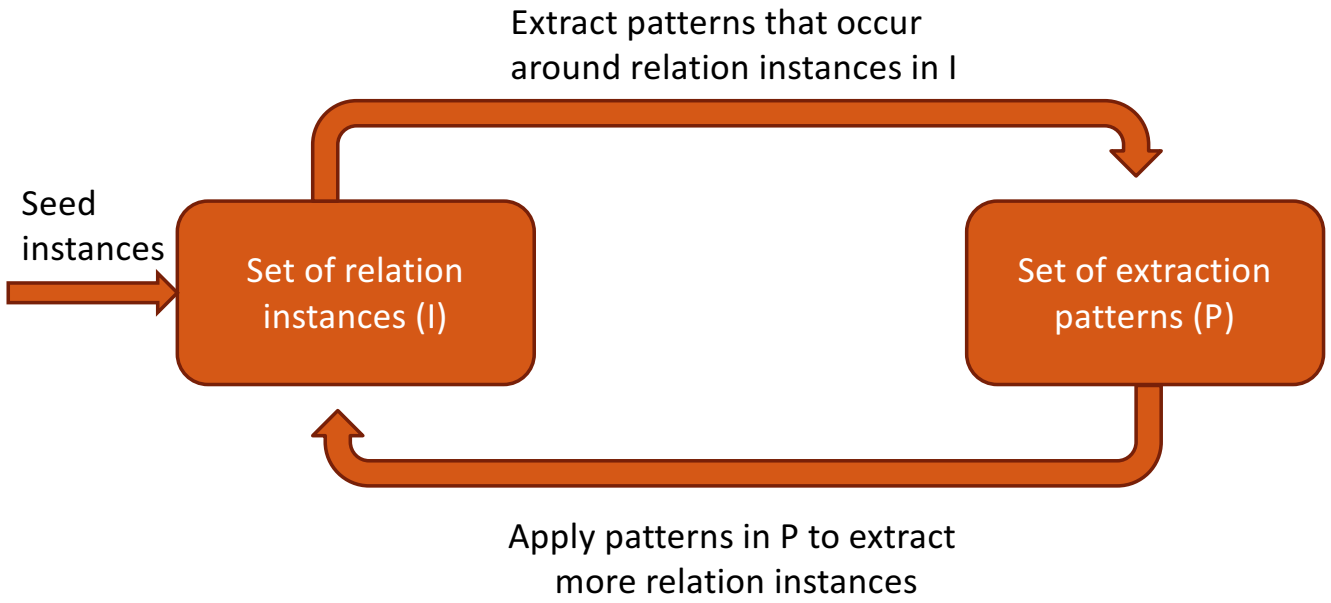
High precision: when it fires, it's correct
Easy to explain predictions
Easy to fix mistakes

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

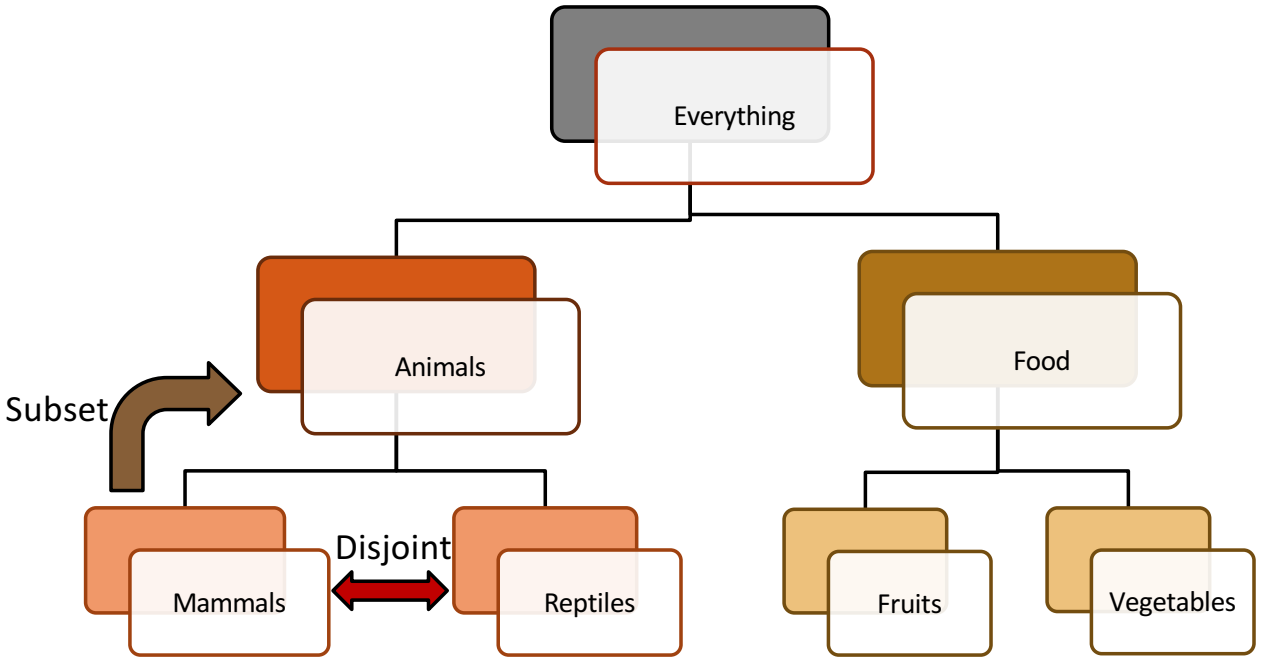
(2) Ontology based extraction

Defining domain	Learning extractors	Scoring extractions
		

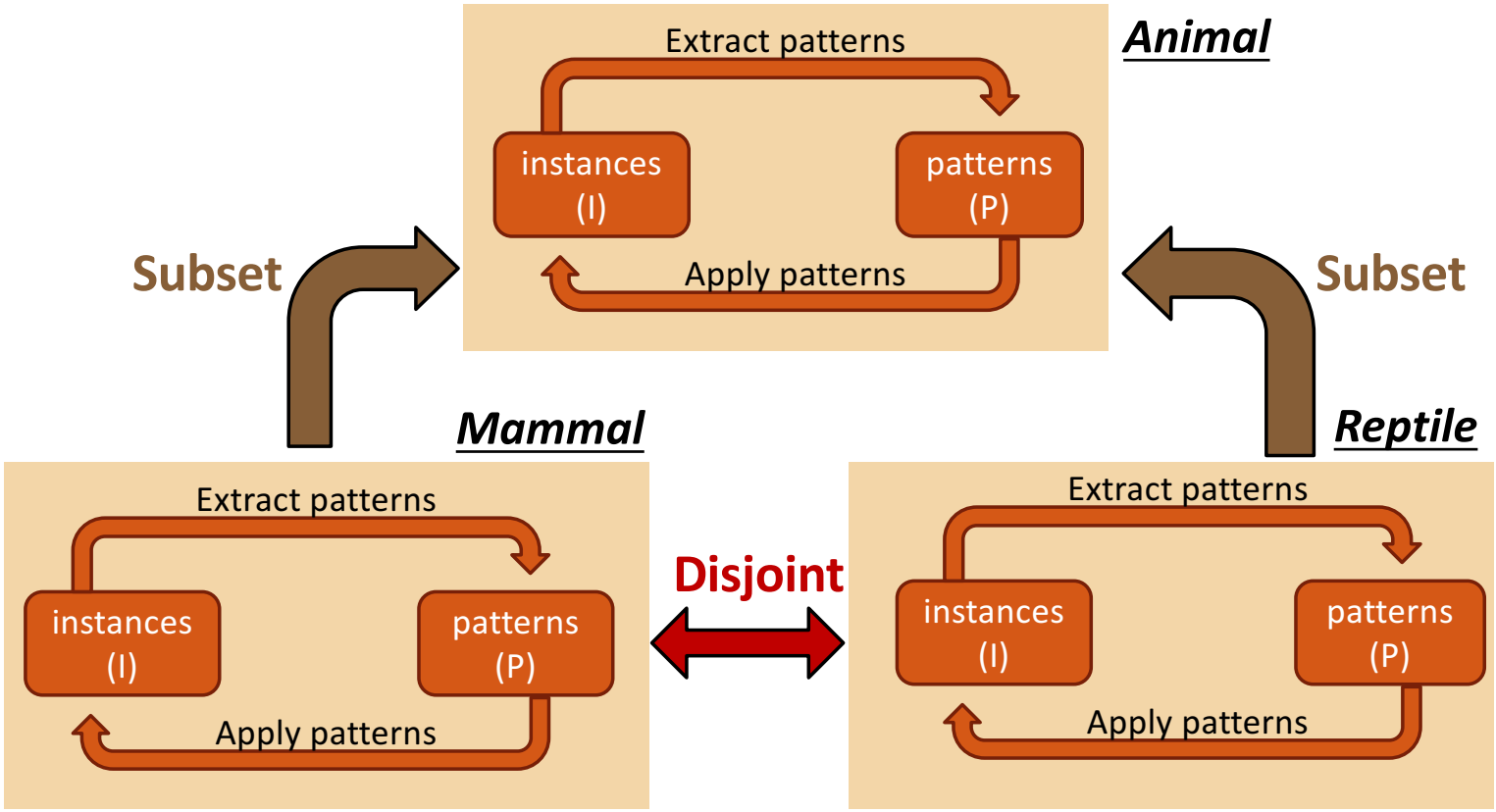
Semi-supervised learning (bootstrapping)



Coupling Constraints (Ontology)



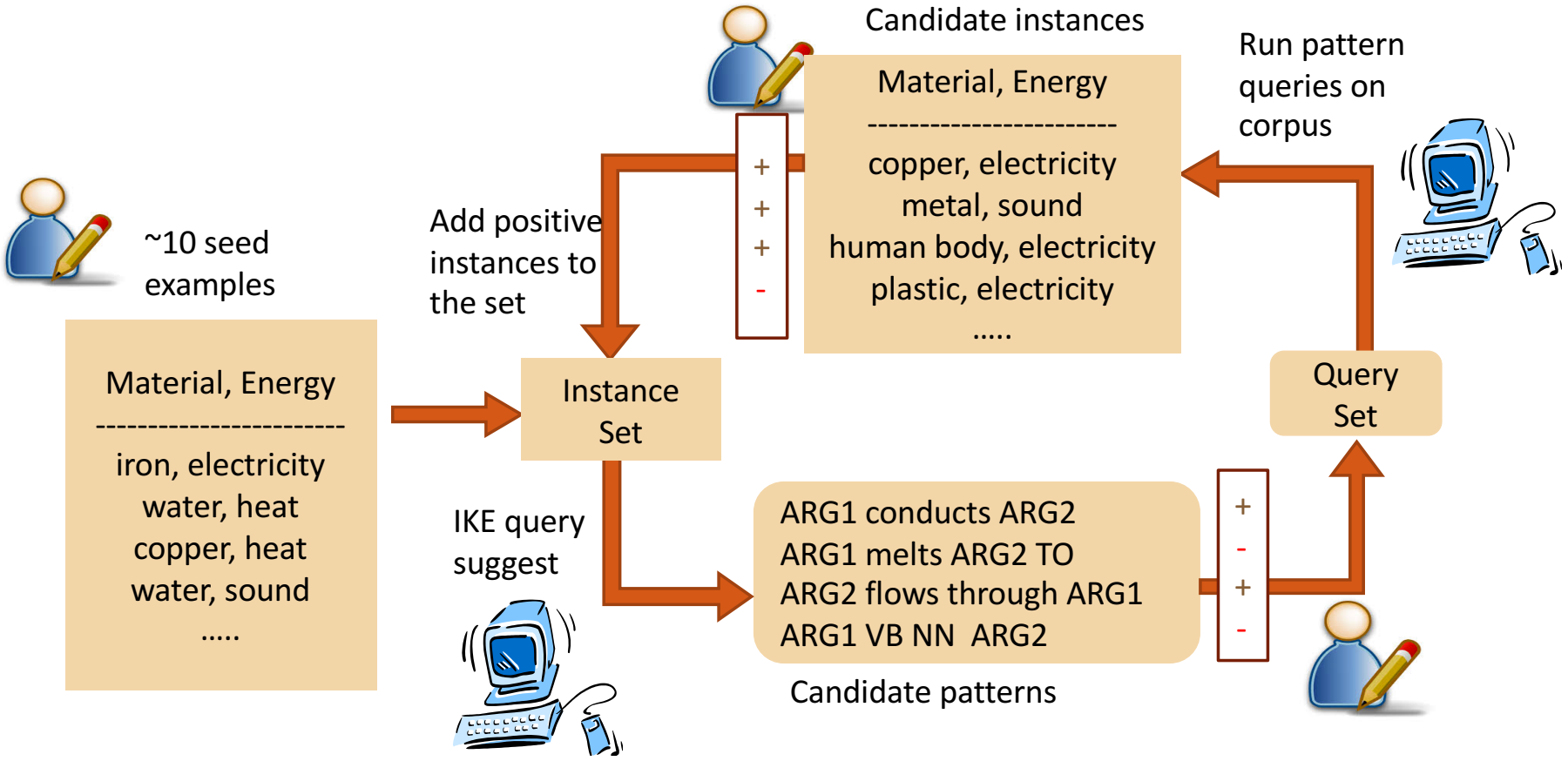
Coupled bootstrap learning



(3) Interactive Bootstrapping (IKE)

Defining domain	Learning extractors	Scoring extractions
		

(3) Interactive Bootstrapping (IKE)



(4) Open Domain IE

Defining domain	Learning extractors	Scoring extractions
		

(4) Open domain IE

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

John Lennon was an English music artist who gained worldwide fame as one of the members of the Beatles.

Open IE

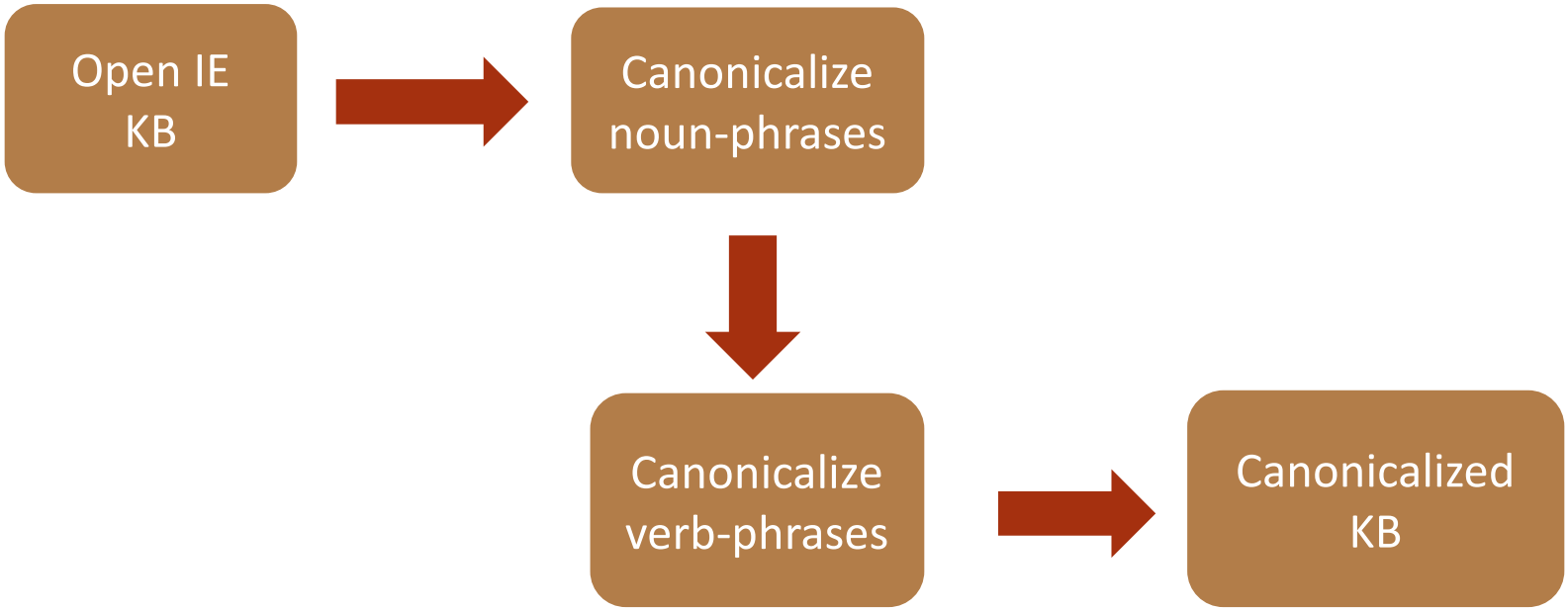
- 0.95 (John Lennon; **was**; an English music artist)
- 0.94 (an English music artist; **gained**; L:worldwide; fame; as one of the members of the Beatles)

(5) Hybrid approach

Adding structure to Open KB

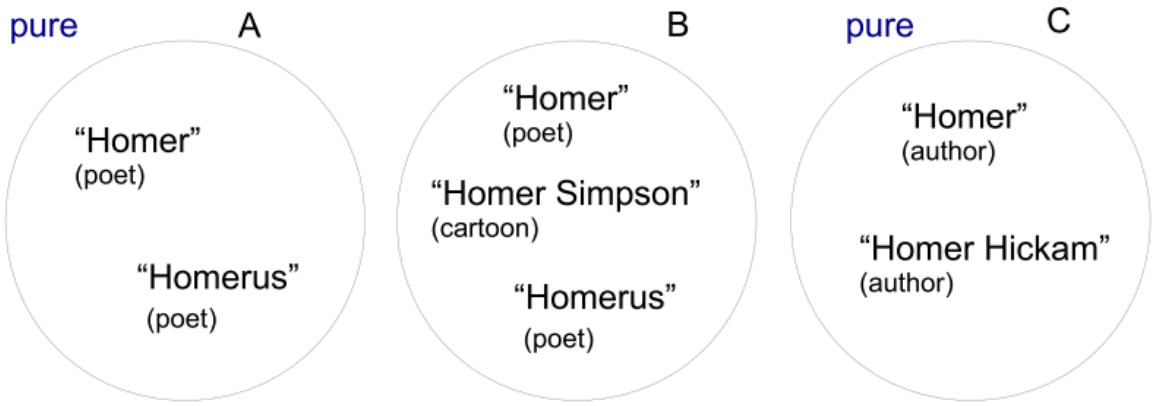
Defining domain	Learning extractors	Scoring extractions
		

(5) Hybrid approach



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

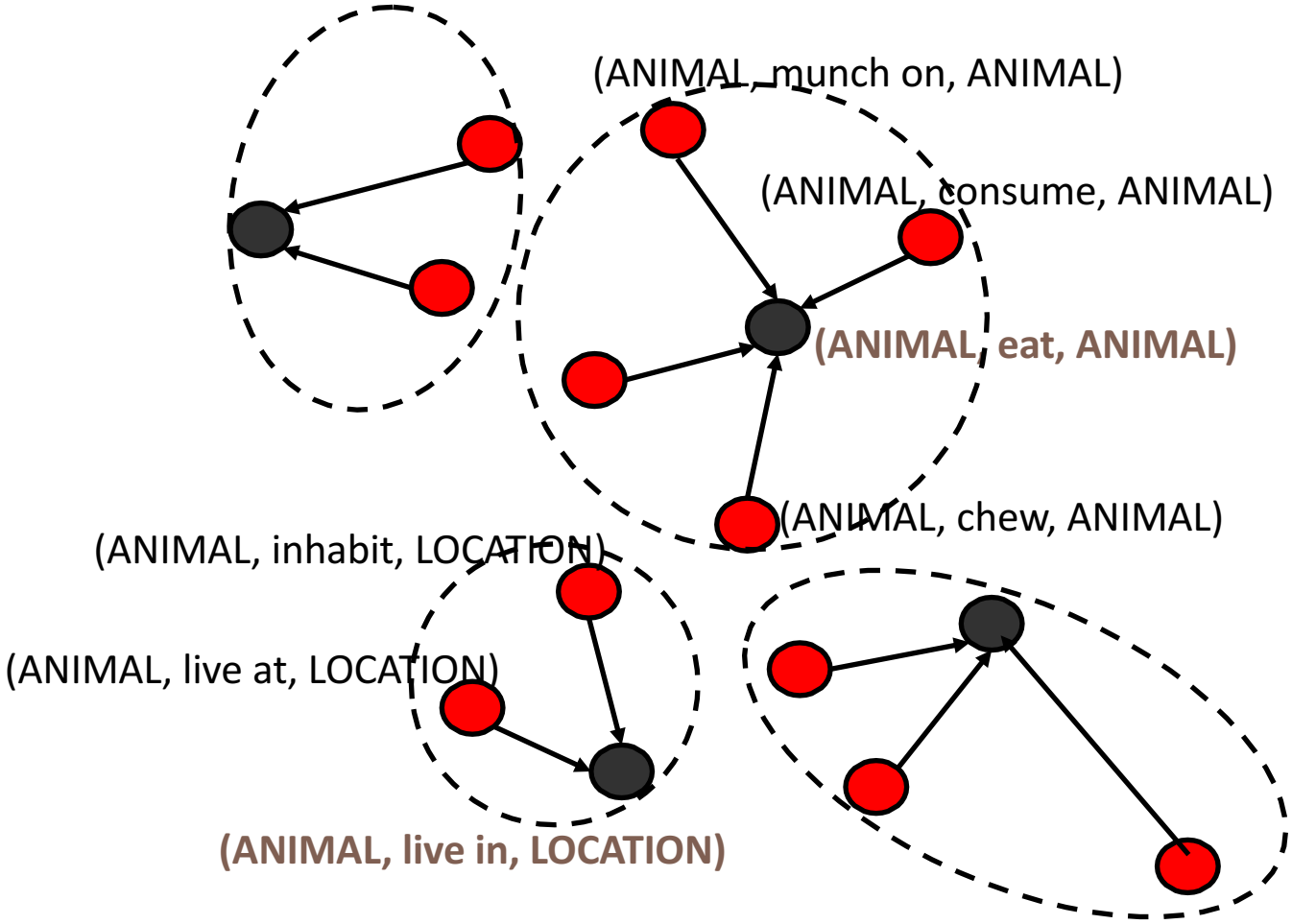
- ***Canonicalizing noun phrases***



- ***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	`location.country.official_language`
be bought, acquire	`organization.organization.acquired_by`

Canonical schema induction



[Upcoming article in TACL 2017 ``Domain-Targeted, High Precision Knowledge Extraction'']

Knowledge fusion with multiple extractors

VOTING (AND VS OR OF EXTRACTORS)

CO-TRAINING (MULTIPLE EXTRACTION METHODS)

MULTI-VIEW LEARNING (MULTIPLE DATA SOURCES)

MACHINE LEARNING FOR KNOWLEDGE FUSION

Information Extraction

Single extractor

Defining domain

Learning extractors

Scoring the extractions

×

Manual

Semi-automatic

Automatic

Fusing multiple extractors

Multiple weak extractors

- ***Extractor 1:*** text patterns to extract ISA relations
  e.g. coupled pattern learner in NELL

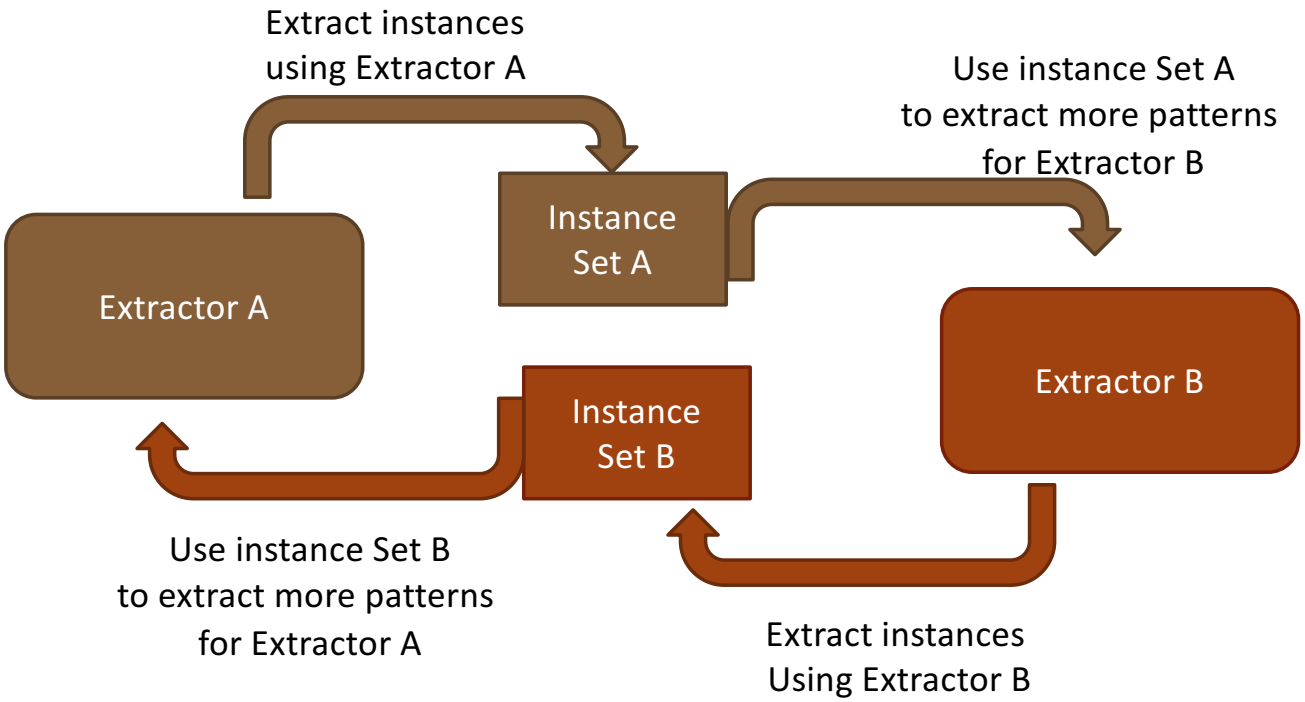
- ***Extractor 2:*** learning wrappers for HTML pages to extract ISA relations from structured text

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
  - score(fact) = score_extractor1(fact) * scrore_extractor2(fact)

- ***OR of two extractors***
  - For a candidate extraction to be promoted to a fact in KB, both the extractors should support the fact
  - score(fact) = score_extractor1(fact) * scrore_extractor2(fact)

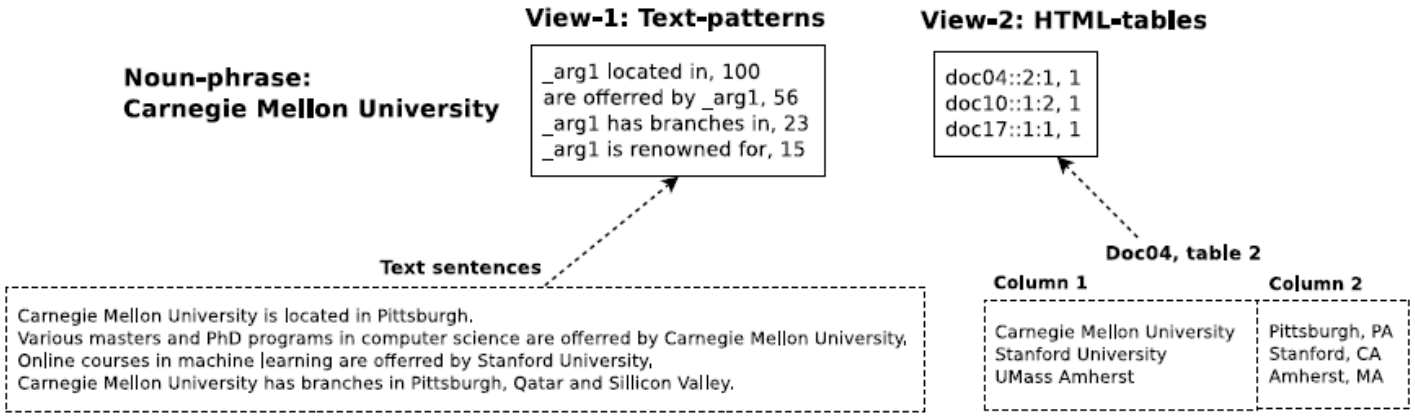
Co-training



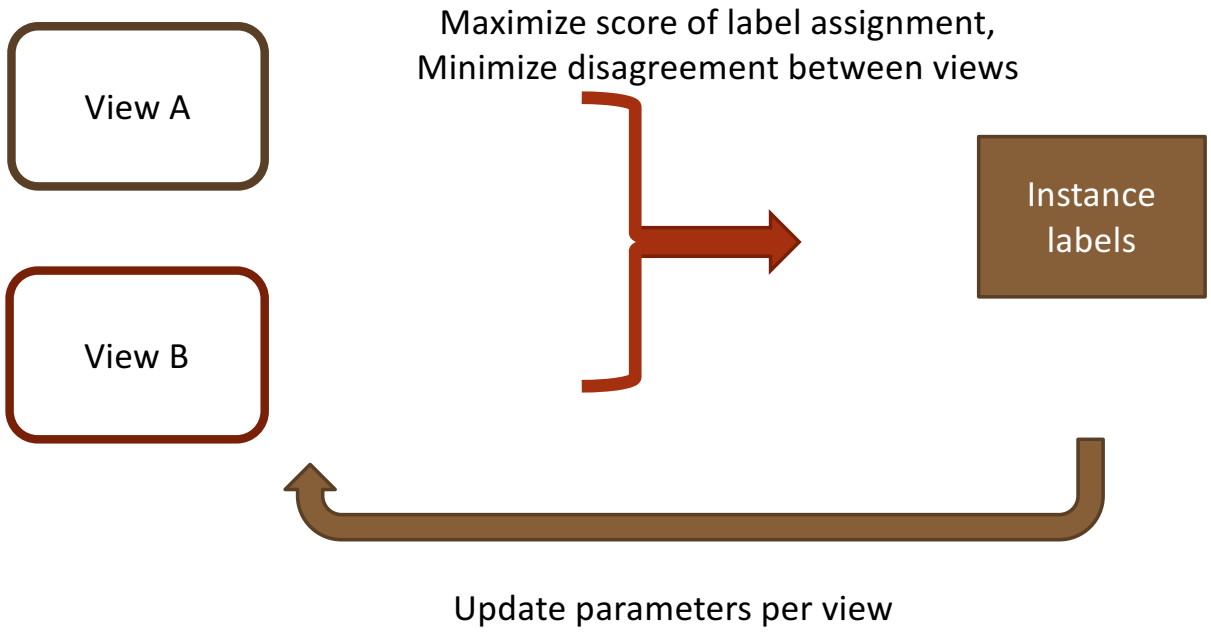
[**Combining Labeled and Unlabeled Data with Co-Training**, *Blum and Mitchell, CoLT 1998]*

Multiple data-views

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

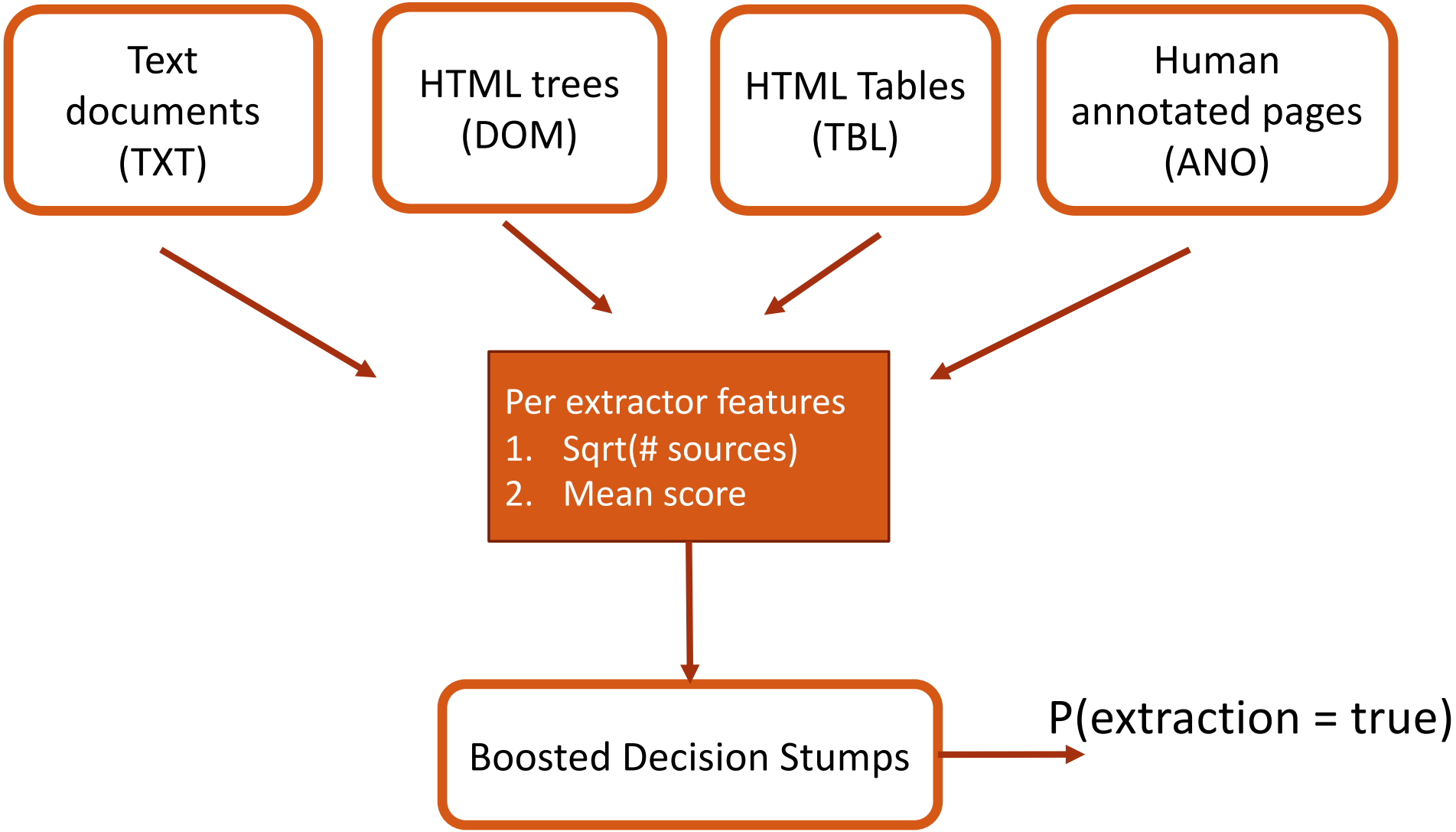


Multi-view learning



[**Combining Labeled and Unlabeled Data with Co-Training,** *Blum and Mitchell, CoLT 1998]*

Knowledge vault: fusing the extractors



Example IE Systems

OPEN IE

NELL

KNOWLEDGE VAULT

Open IE (KnowItAll)

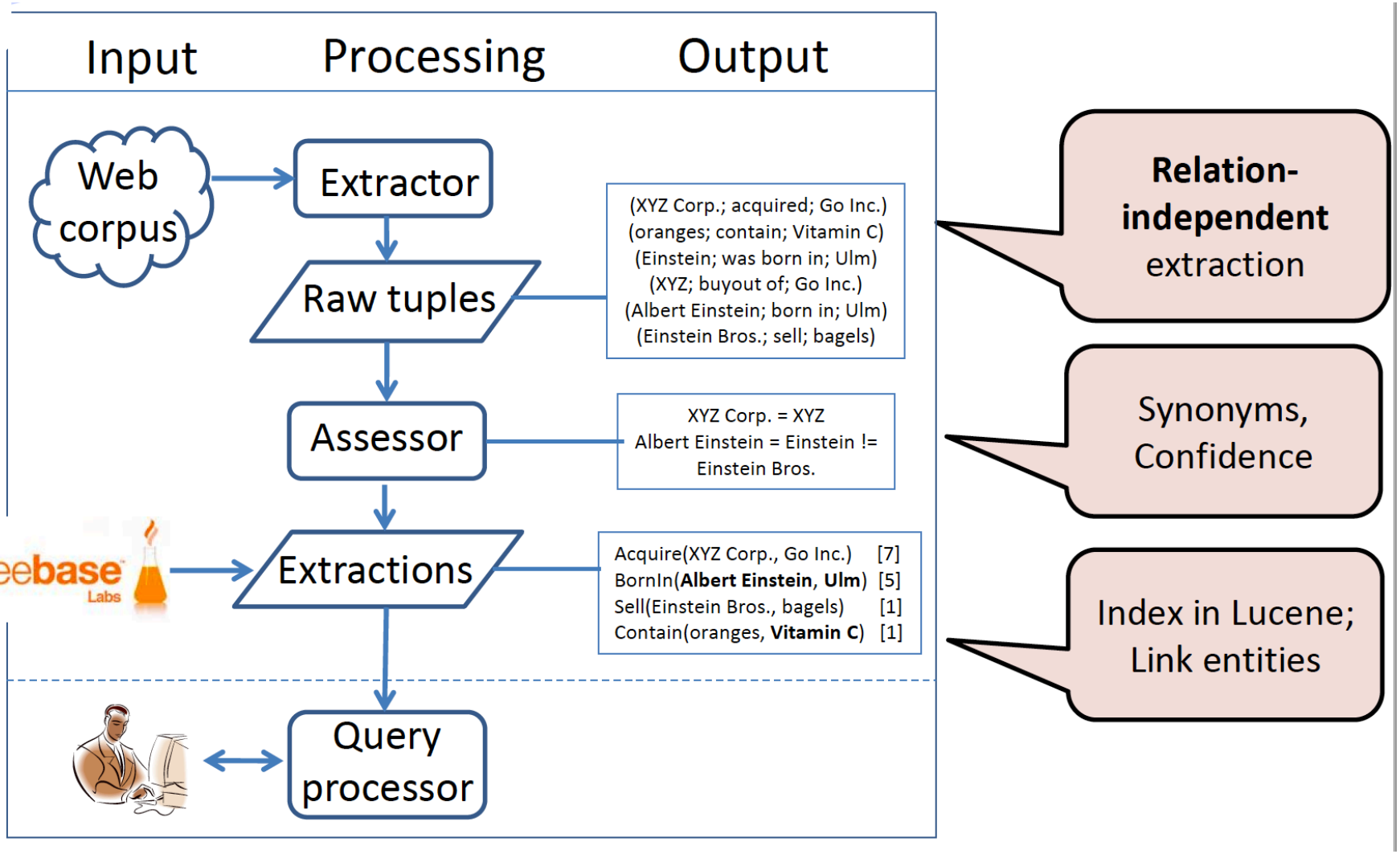
Open Information Extraction

Defining domain	Learning extractors	Scoring extractions
		

[Architecture diagram taken from Oren Etzioni's slides]

Open IE (KnowItAll)

Open Information Extraction



[Architecture diagram taken from Oren Etzioni's slides]

Never Ending Language Learning (NELL)

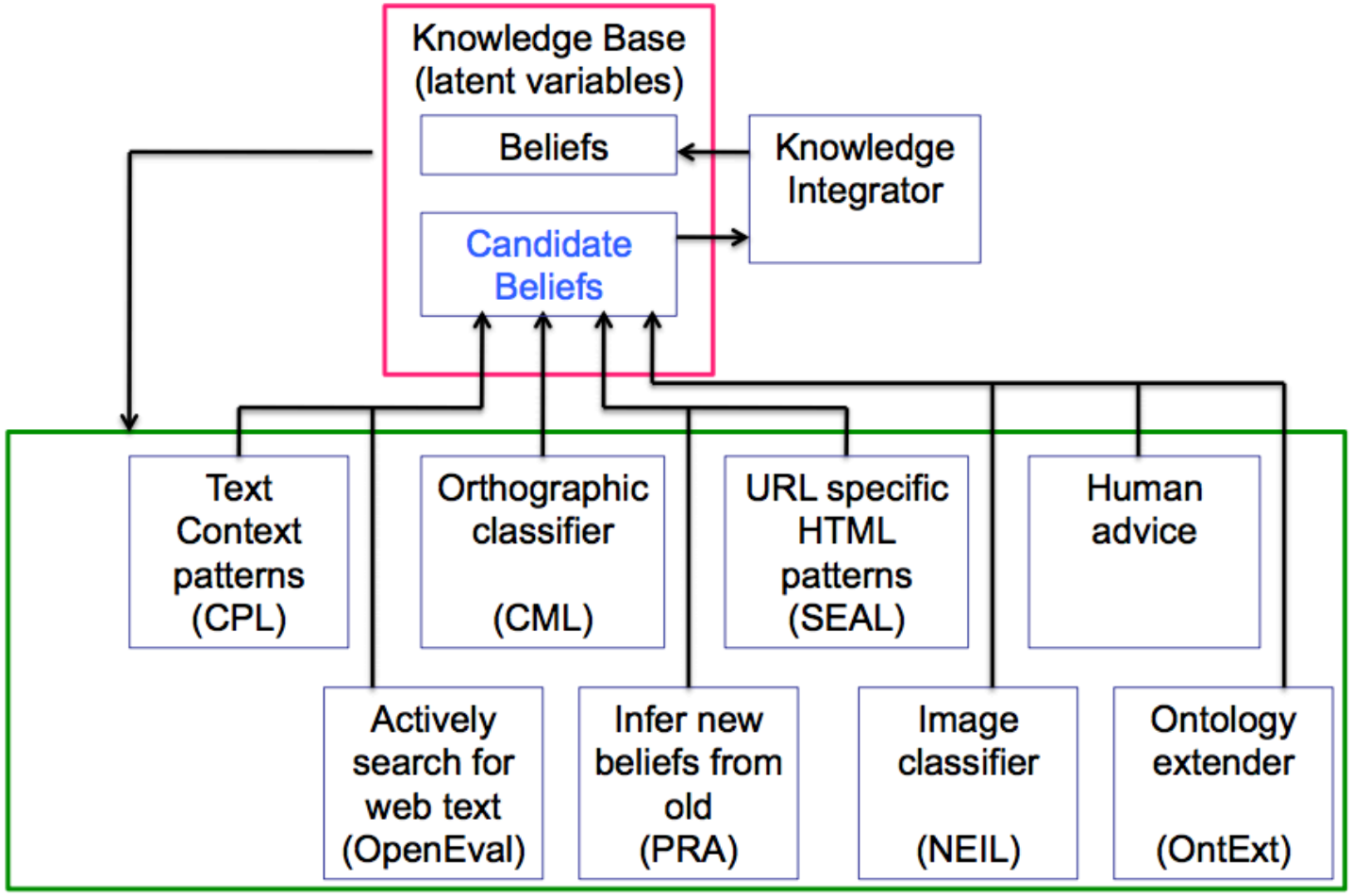
Ontology based extraction

NELL Knowledge Base Browser
CMU Read the Web Project

categories | relations

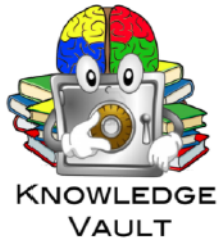
Defining domain	Learning extractors	Scoring extractions	Fusing extractors
			Voting + local constraint satisfaction

Never Ending Language Learning (NELL)



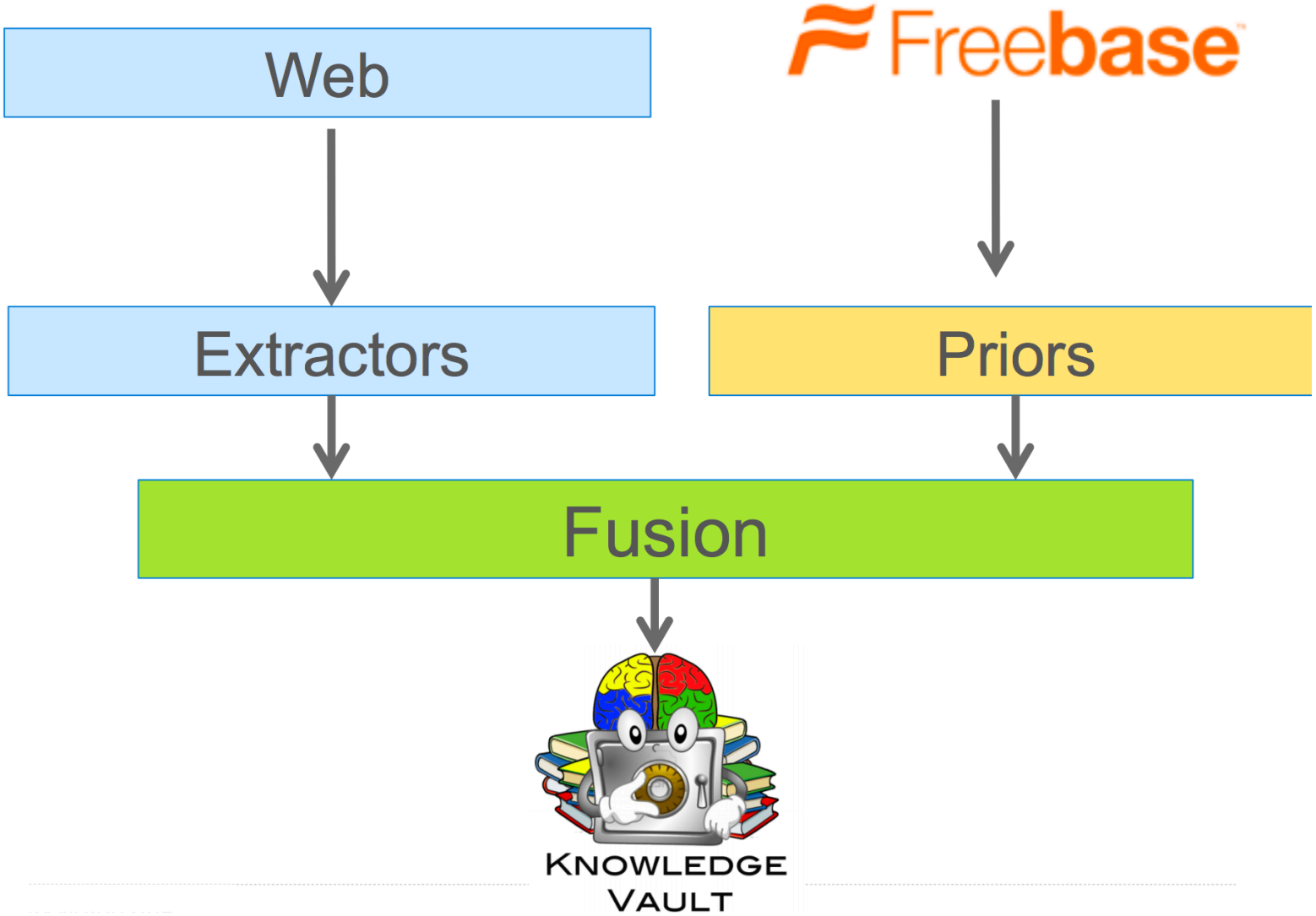
*[**Never-Ending Learning**, Mitchell et al., AAAI 2015]*

Knowledge Vault



Defining domain	Learning extractors	Scoring extractions	Fusing extractors
			Classifier (Boosted decision stumps)

Knowledge Vault



[Architecture diagram taken from Kevin Murphy's slides]

Summary: Information Extraction

3 IMPORTANT SUB-PROBLEMS

(DEFINE DOMAIN, LEARN EXTRACTORS, SCORE EXTRACTIONS)

3 LEVELS OF SUPERVISION

(MANUAL, SEMI-SUPERVISED, UNSUPERVISED)

KNOWLEDGE FUSION WITH MULTIPLE EXTRACTORS

(CO-TRAINING, MULTI-VIEW LEARNING)

EXAMPLE IE SYSTEMS

Thank You

SEE YOU AFTER THE COFFEE BREAK!