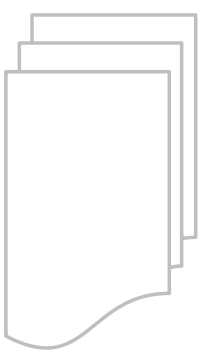
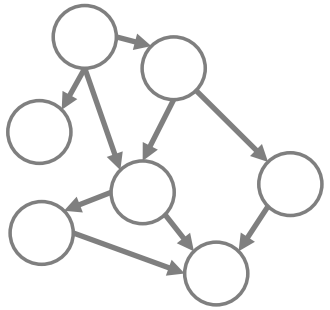
Part 1: Knowledge Graphs



Part 2: Knowledge Extraction

Part 3: Graph Construction



Part 4: Critical Analysis

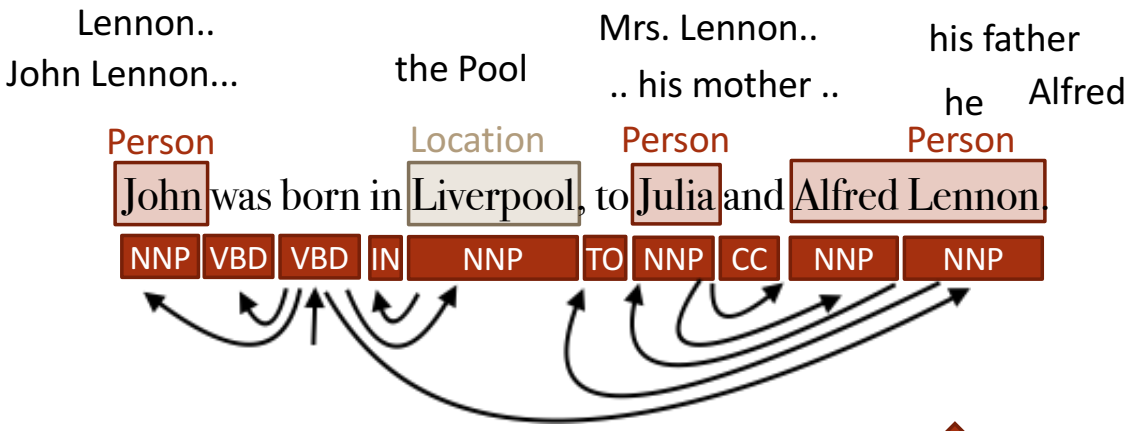
Tutorial Outline

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

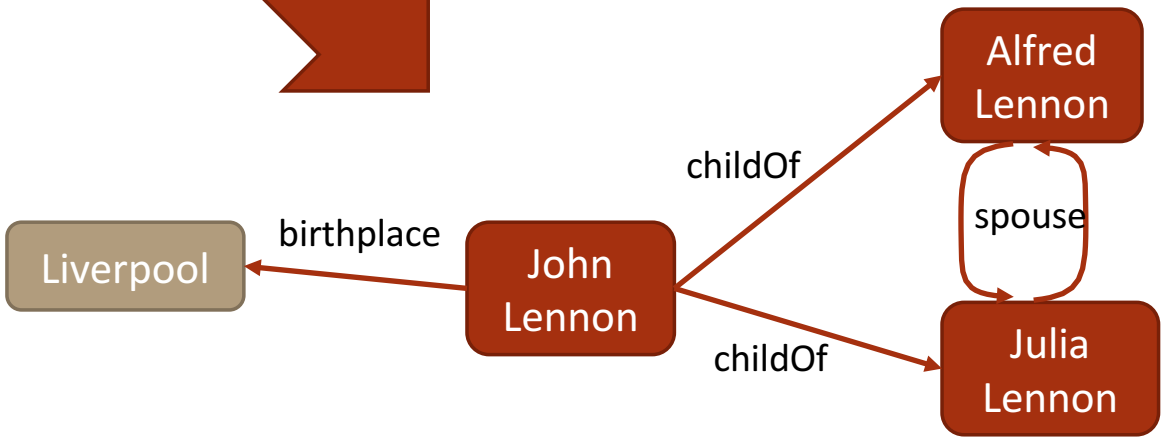
Coffee Break 

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

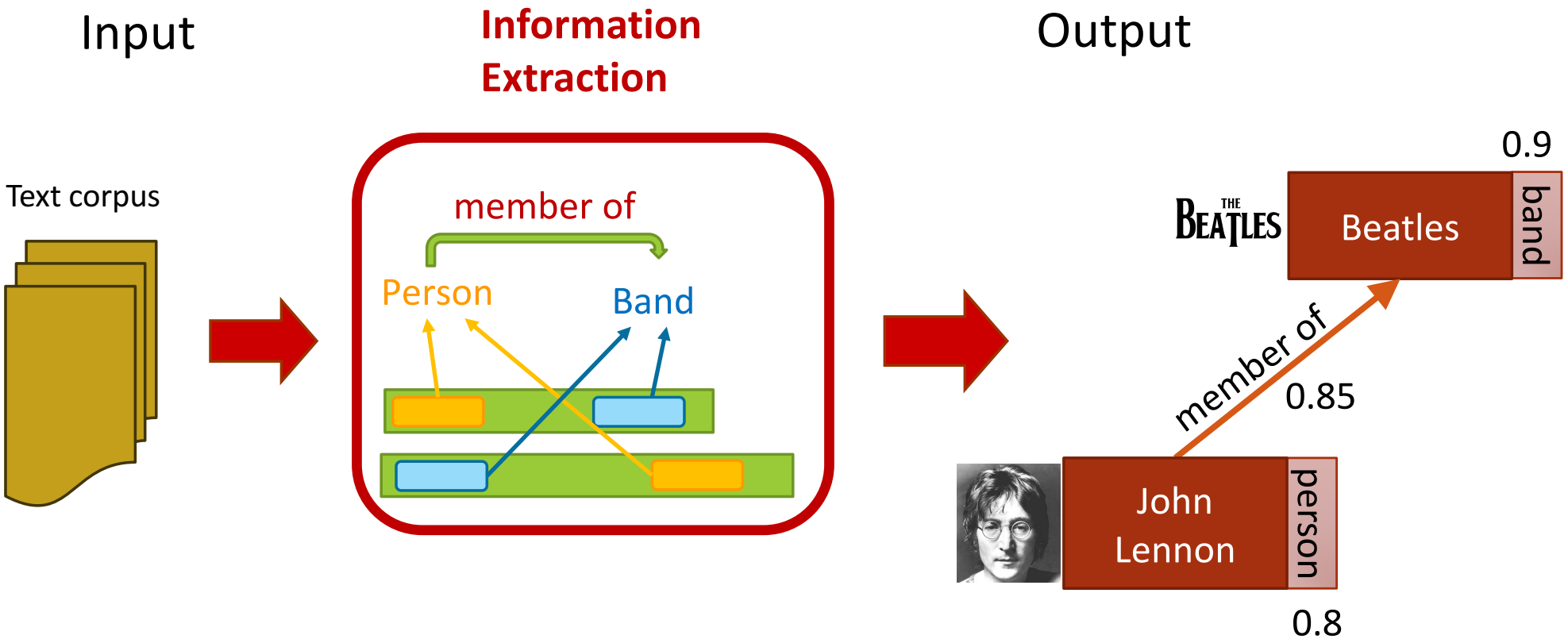
What is Information Extraction?



Information Extraction



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

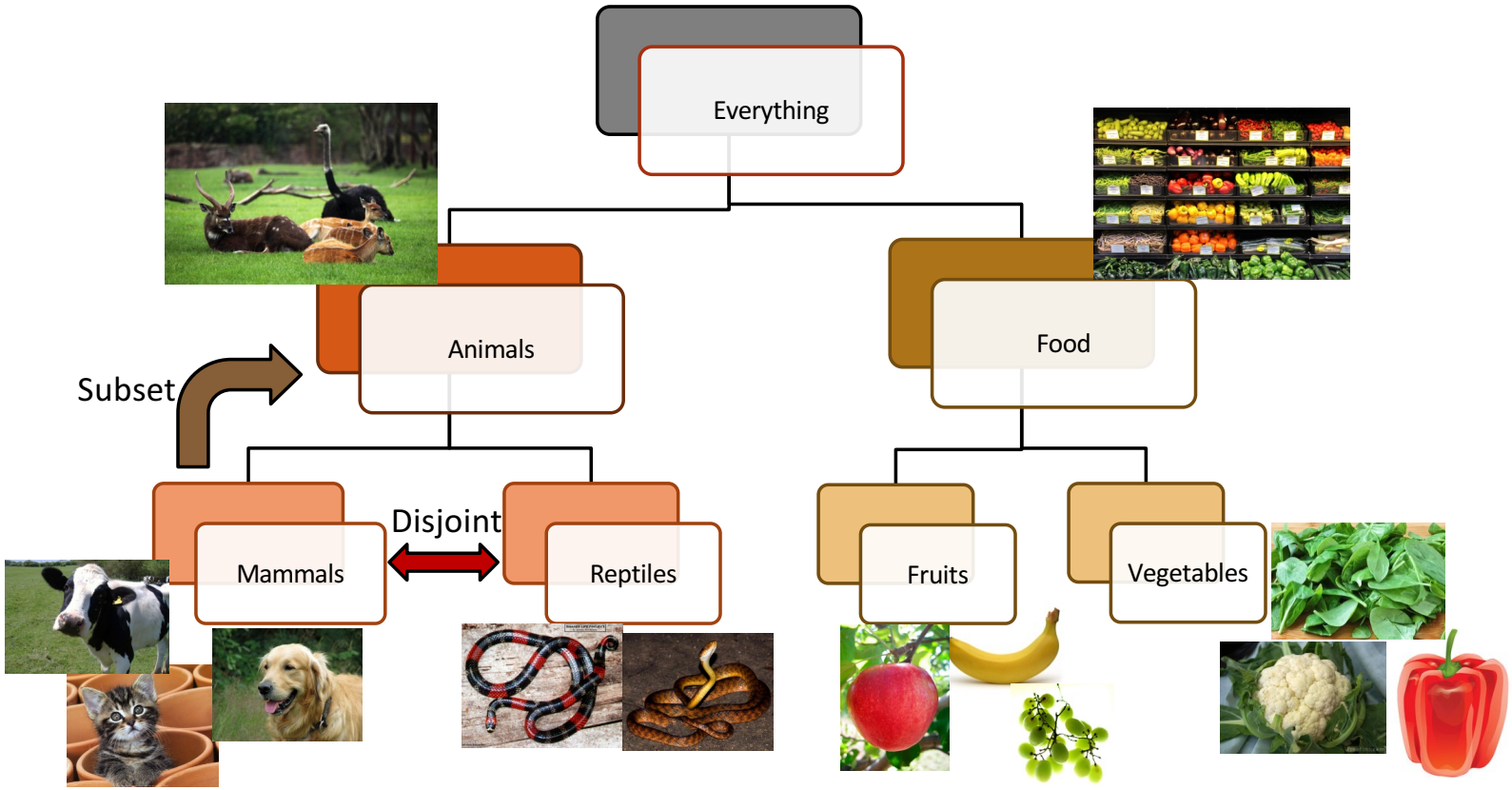
Supervised

Semi-supervised

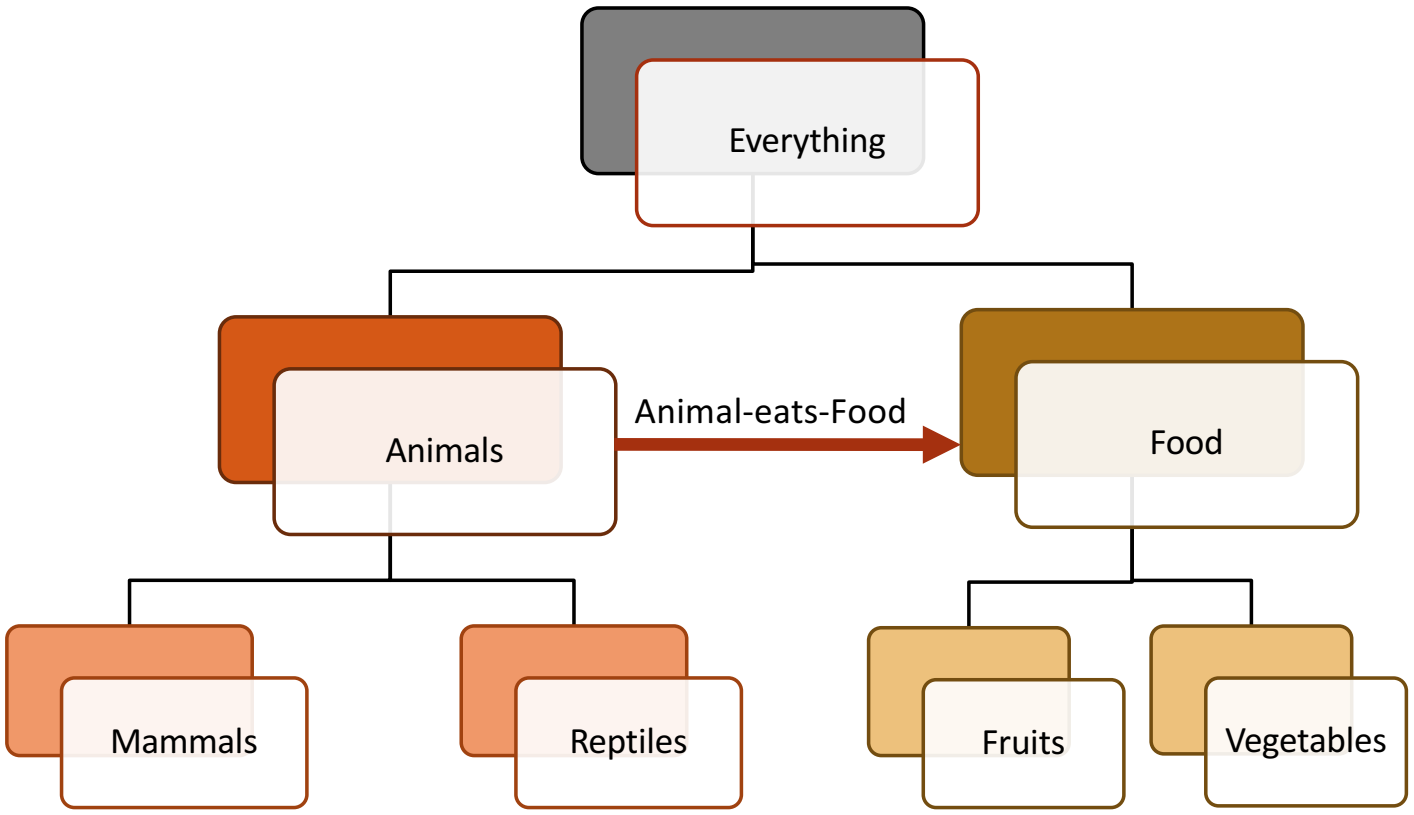
Unsupervised

Defining domain: types/relations of interest

Defining Domain: Manual



Defining Domain: Manual

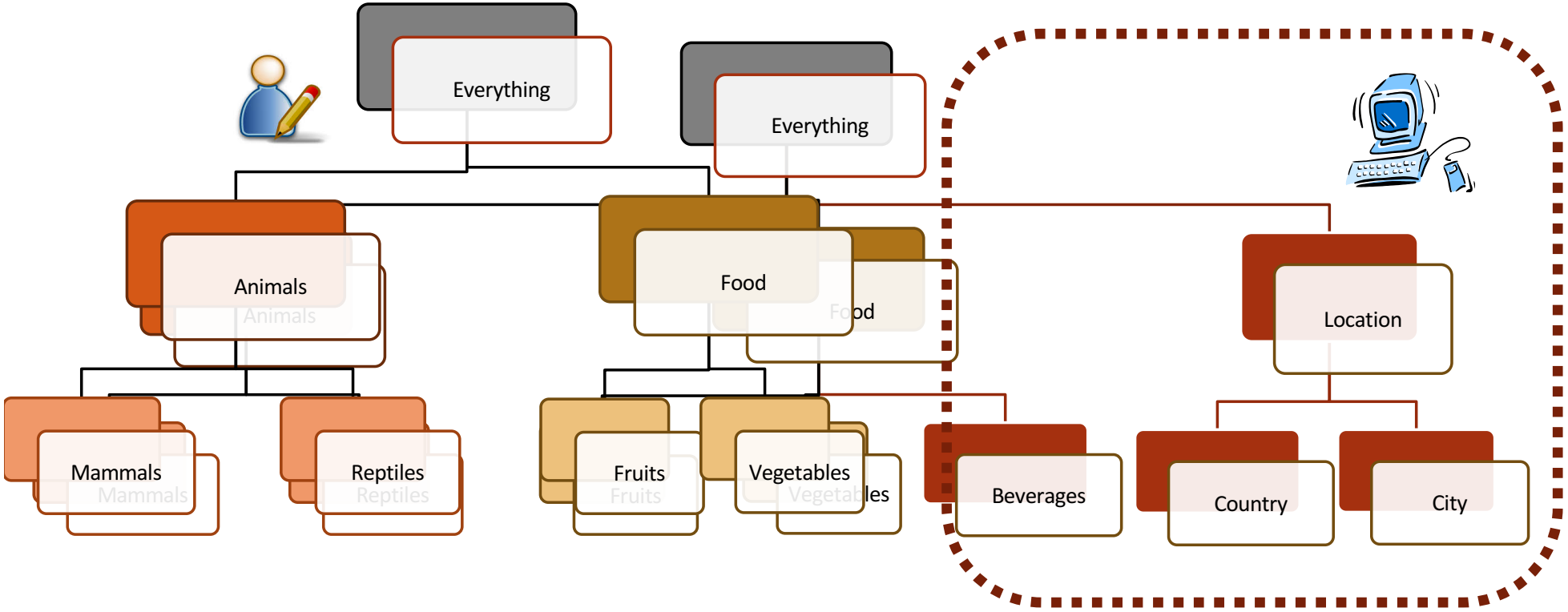


- **Highly semantic ontology**
- **Leads to high precision extractions**
- **Expensive to create**
- **Requires domain experts**

[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

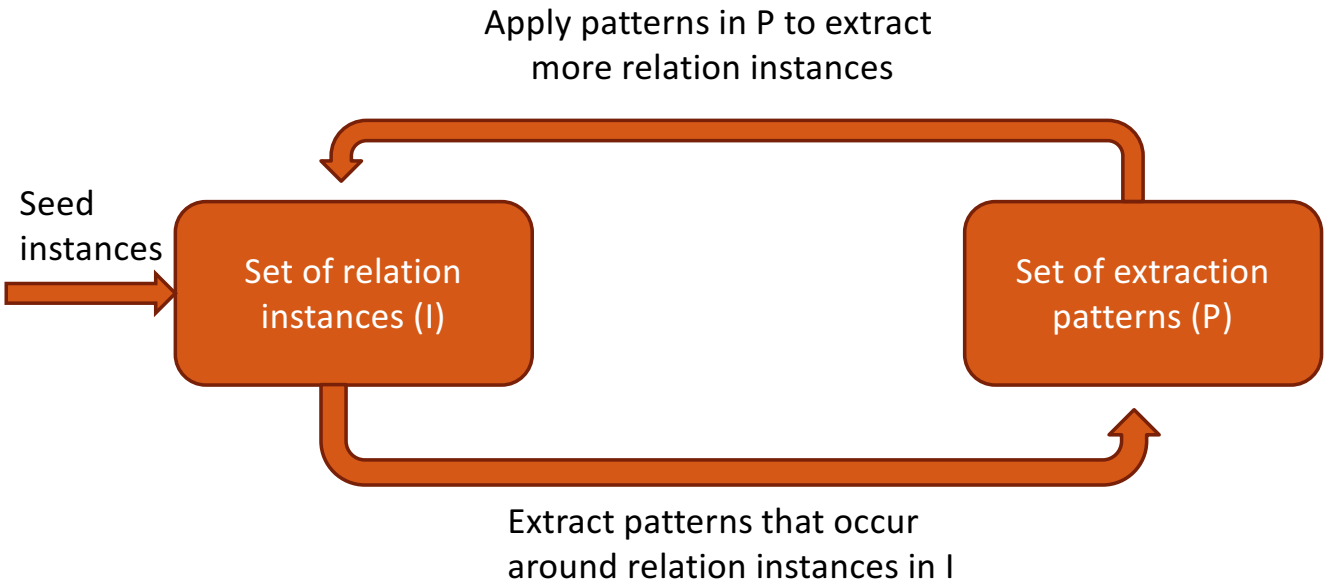
Person-member of-Band

<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

Bootstrapping



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

- 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:

Syntactic constraint

Regular expressions of POS tags

Lexical constraint

|distinct arguments|
a relation phrase takes

/**Identifying Relations for Open Information Extraction, Fader et al.**, EMNLP 2011/

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)

*[**Identifying Relations for Open Information Extraction, Fader et al.**, EMNLP 2011]*

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

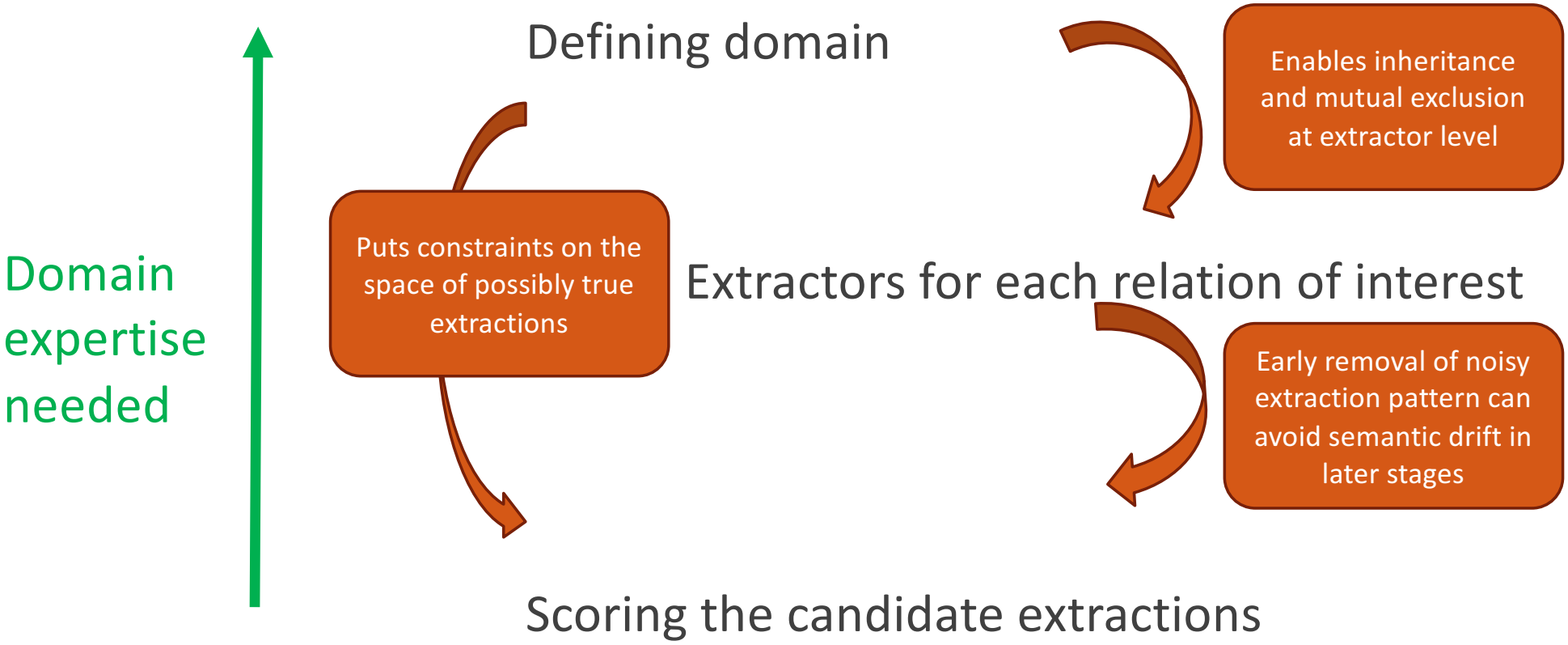
- Completely automatic (Self-training)
  Updated set of instances → weights of extraction patterns → more instances → ….
  (cheap, leads to semantic drift)

Effect of supervision on extractions

Precision,
Human efforts

Recall,
Speed

Impact of early supervision



Categories of IE Techniques

3 concrete sub-problems

Defining domain

Learning extractors

Scoring the extractions

×

3 levels of supervision

Manual

Semi-automatic

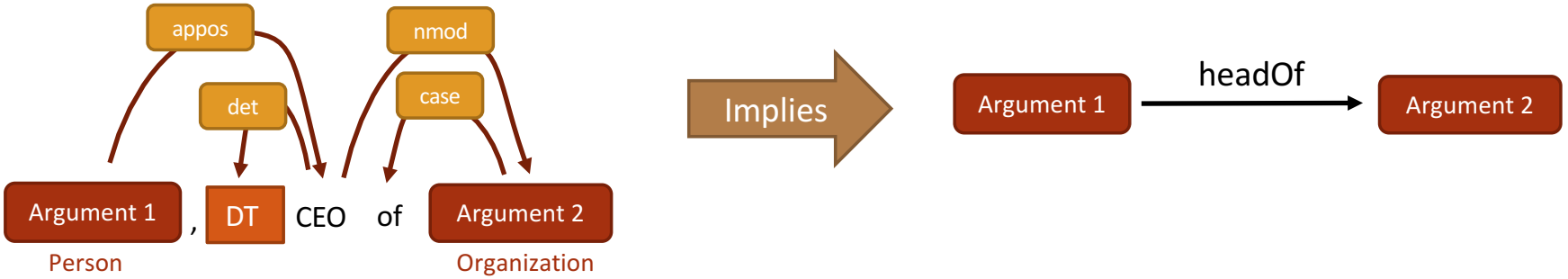
Automatic

Categories of IE Techniques

- Narrow domain patterns
- Ontology based extraction
- Interactive extraction
- Open domain IE
- Hybrid approach (Adding structure to OpenIE KB)

(1) Narrow domain patterns

Use a collection of rules as the system itself



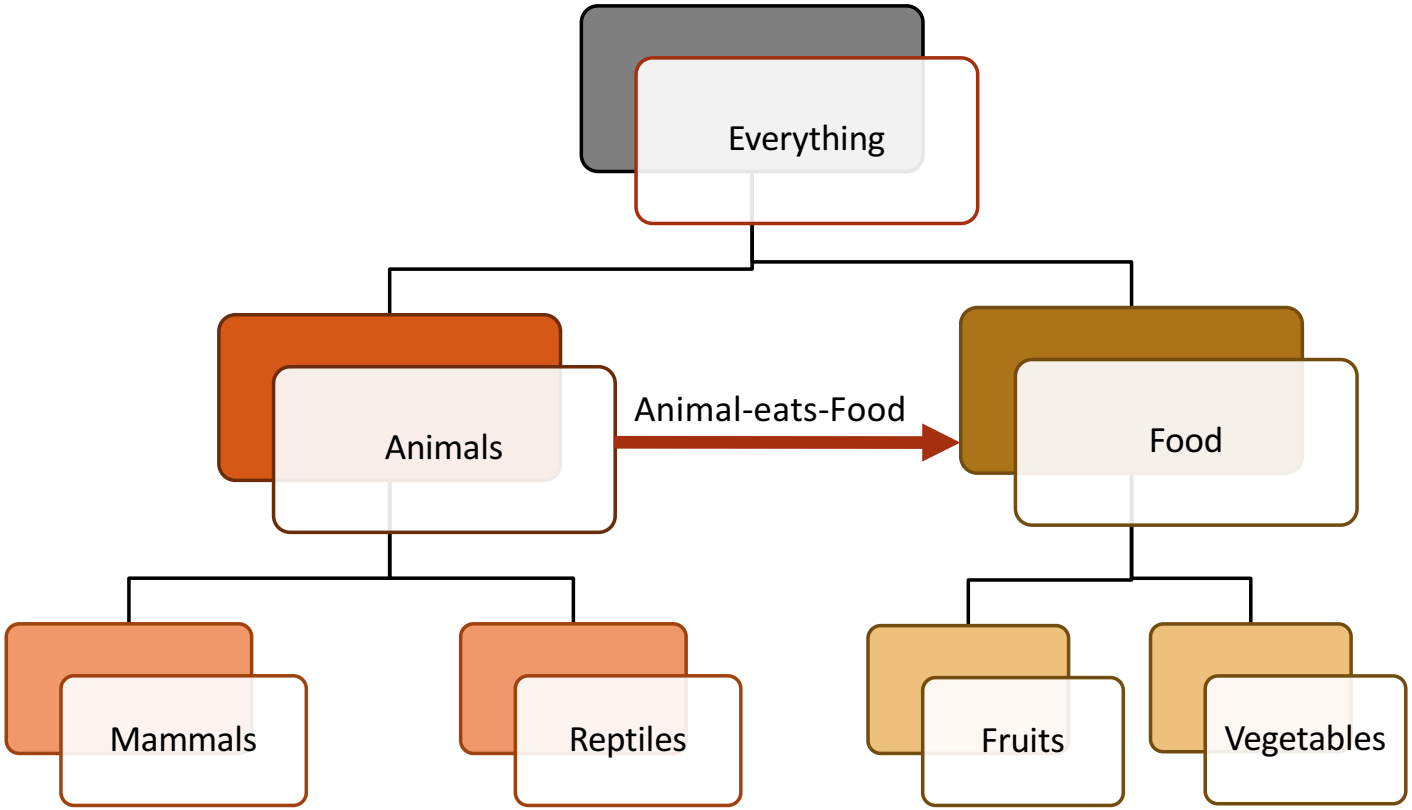
High precision: when it fires, it's correct
Easy to explain predictions
Easy to fix mistakes

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

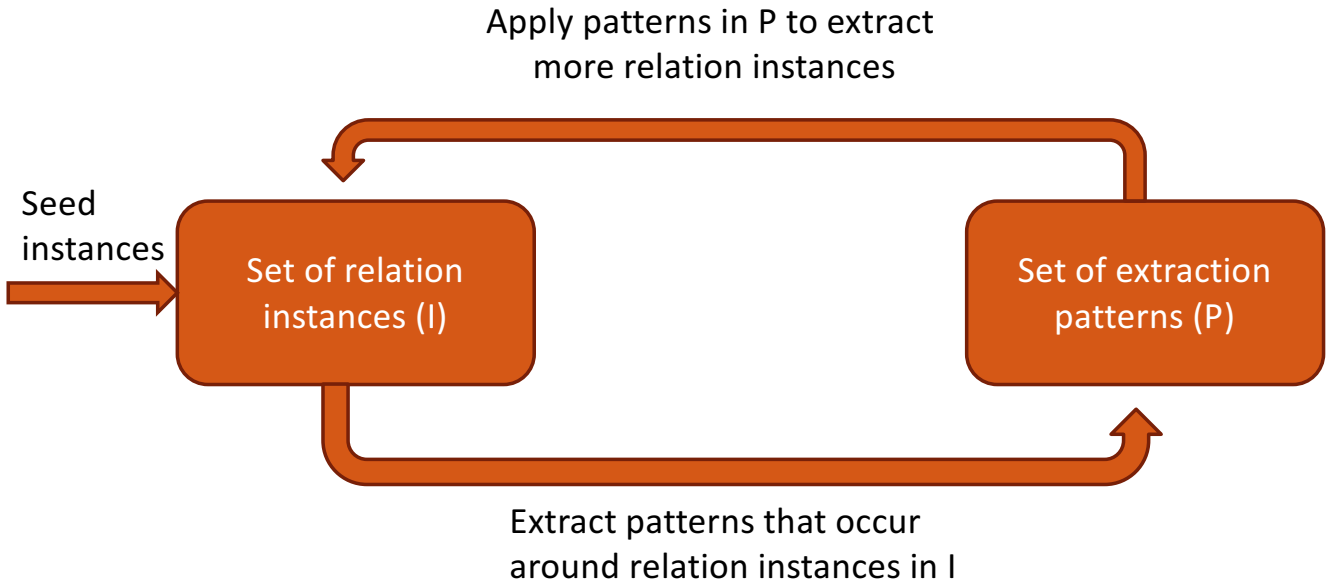
(1) Narrow domain patterns

Defining domain	Learning extractors	Scoring extractions
		

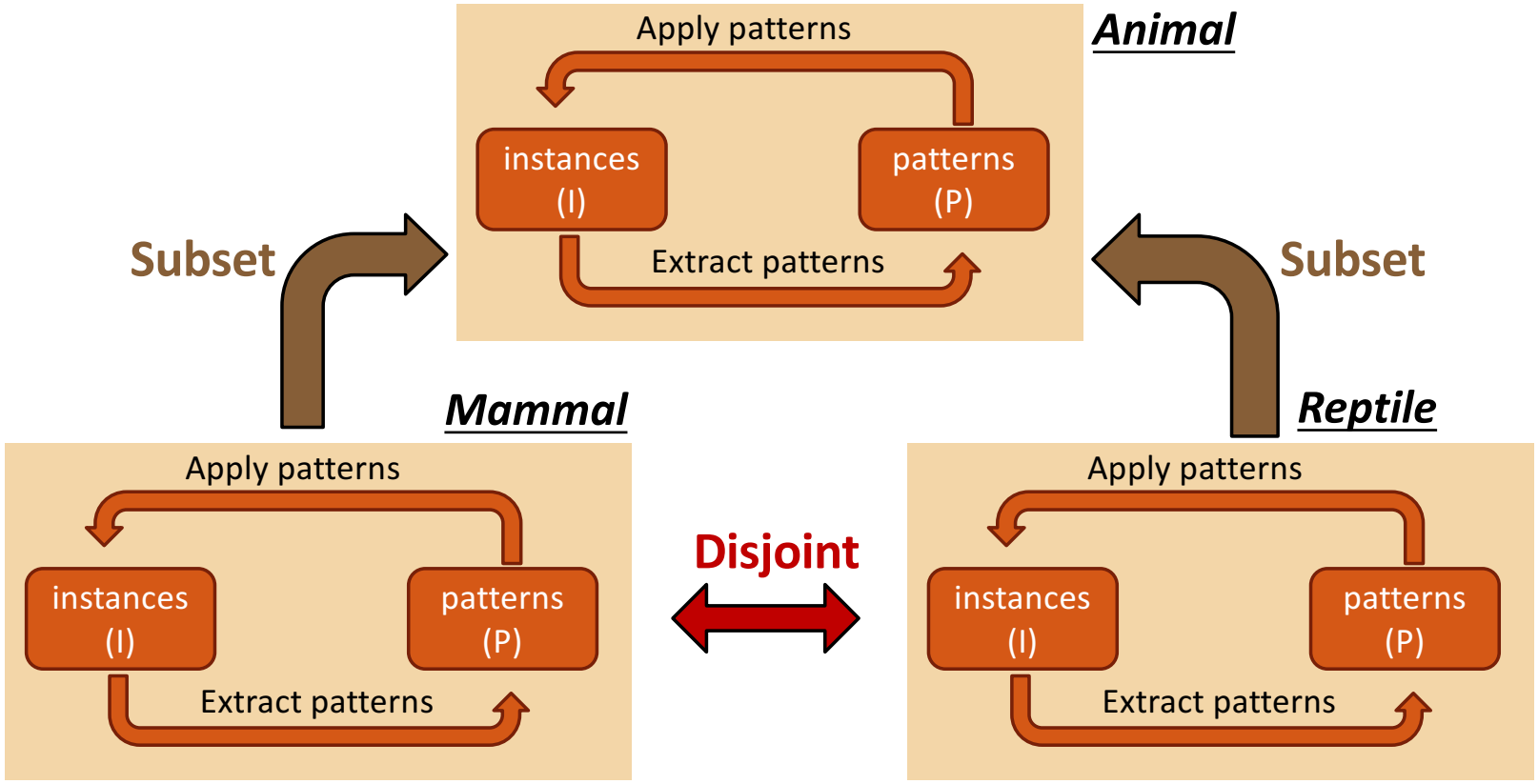
(2) Ontology based extraction



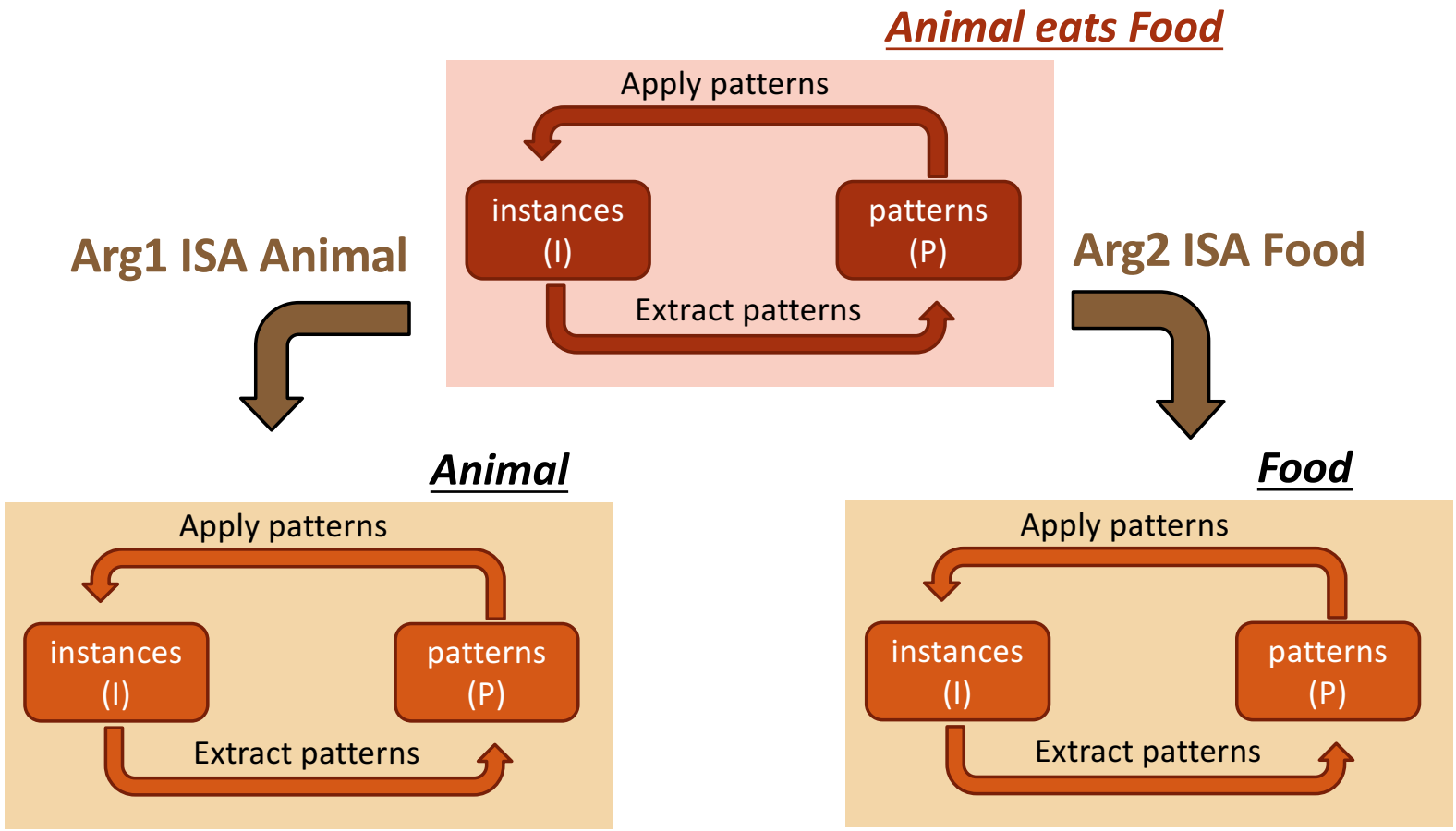
Semi-supervised learning (bootstrapping)



Coupled bootstrap learning



Coupled bootstrap learning

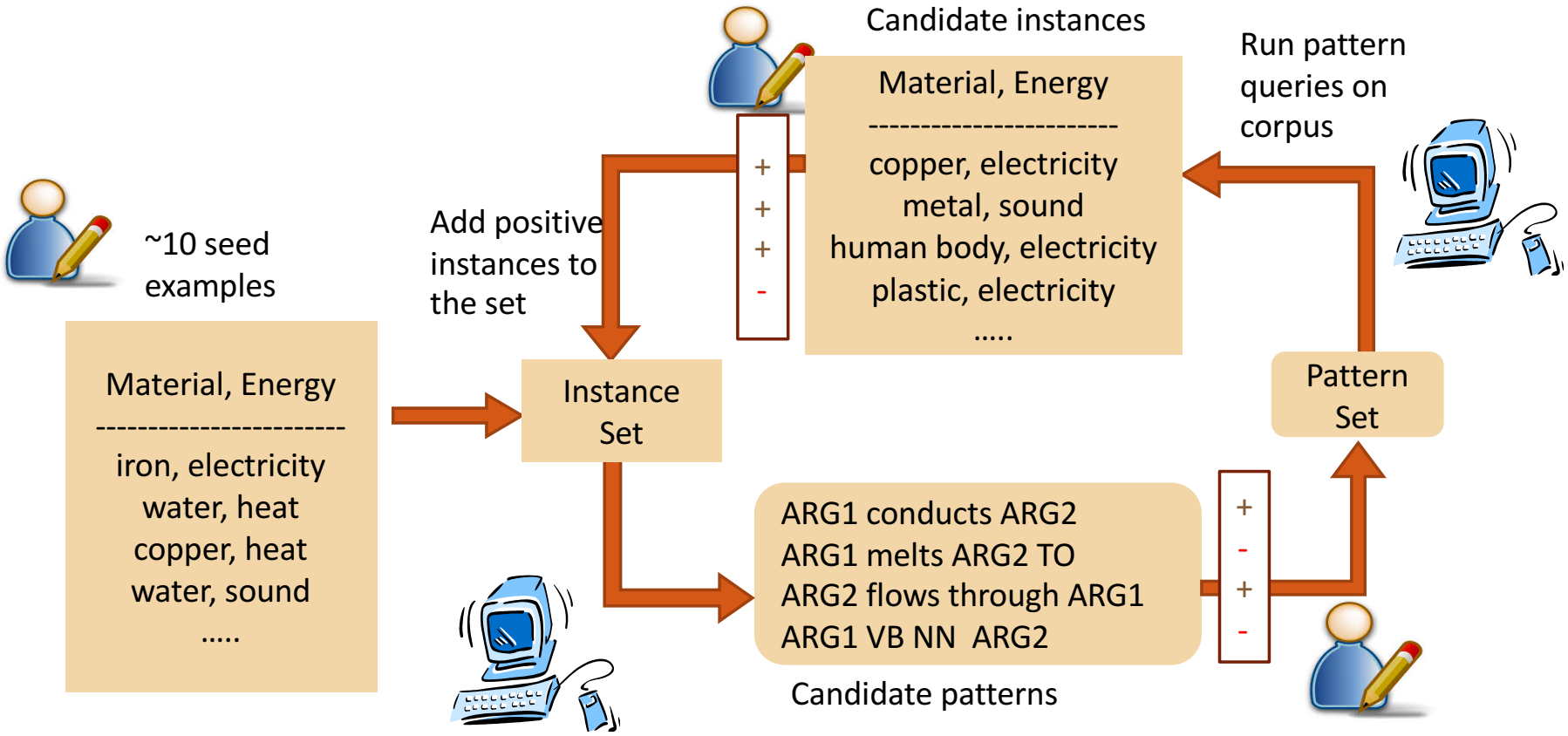


(2) Ontology based extraction

Defining domain	Learning extractors	Scoring extractions
		

(3) Interactive Extraction

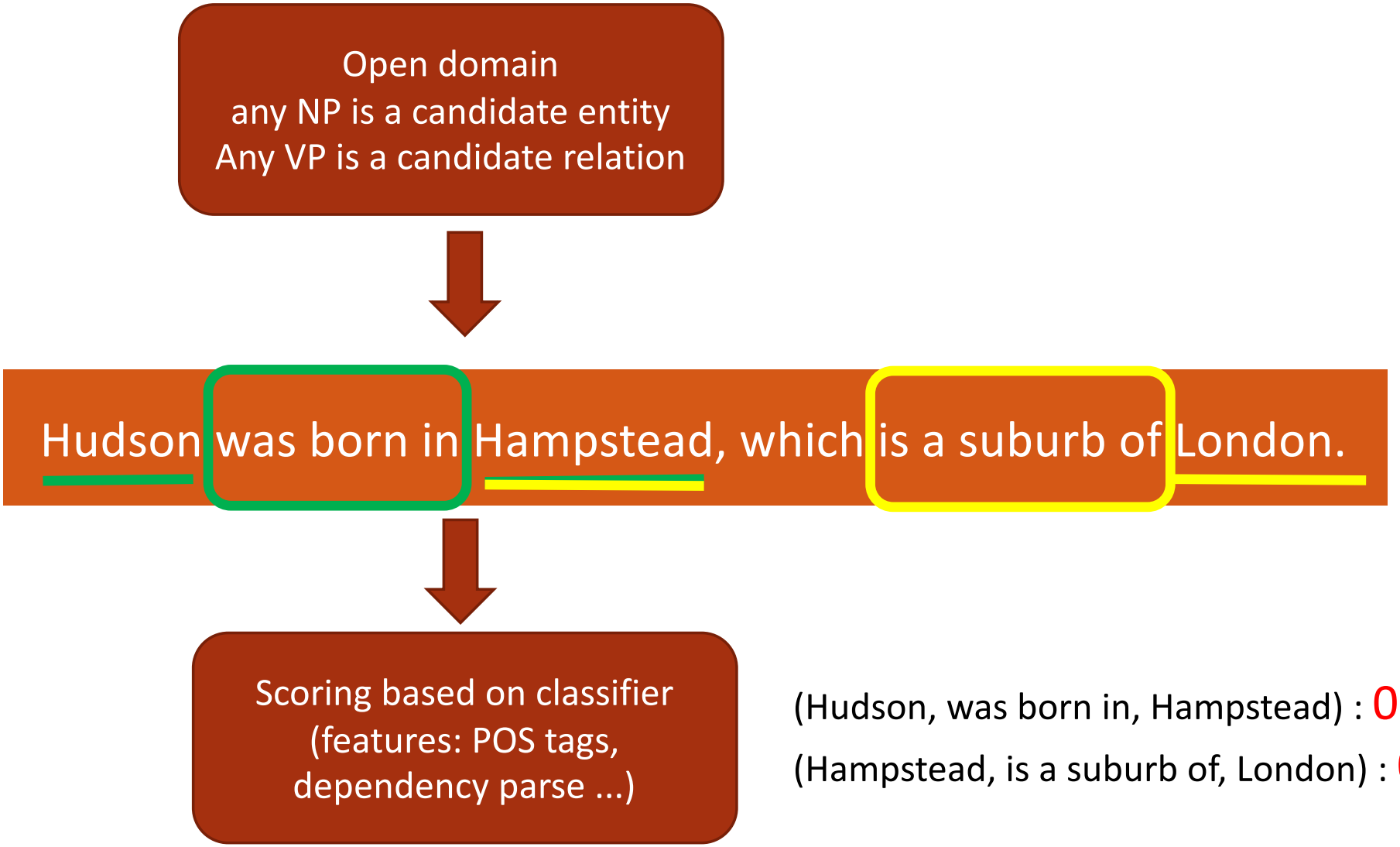
Interactive Bootstrapping



(3) Interactive Extraction

Defining domain	Learning extractors	Scoring extractions
		

(4) Open domain IE



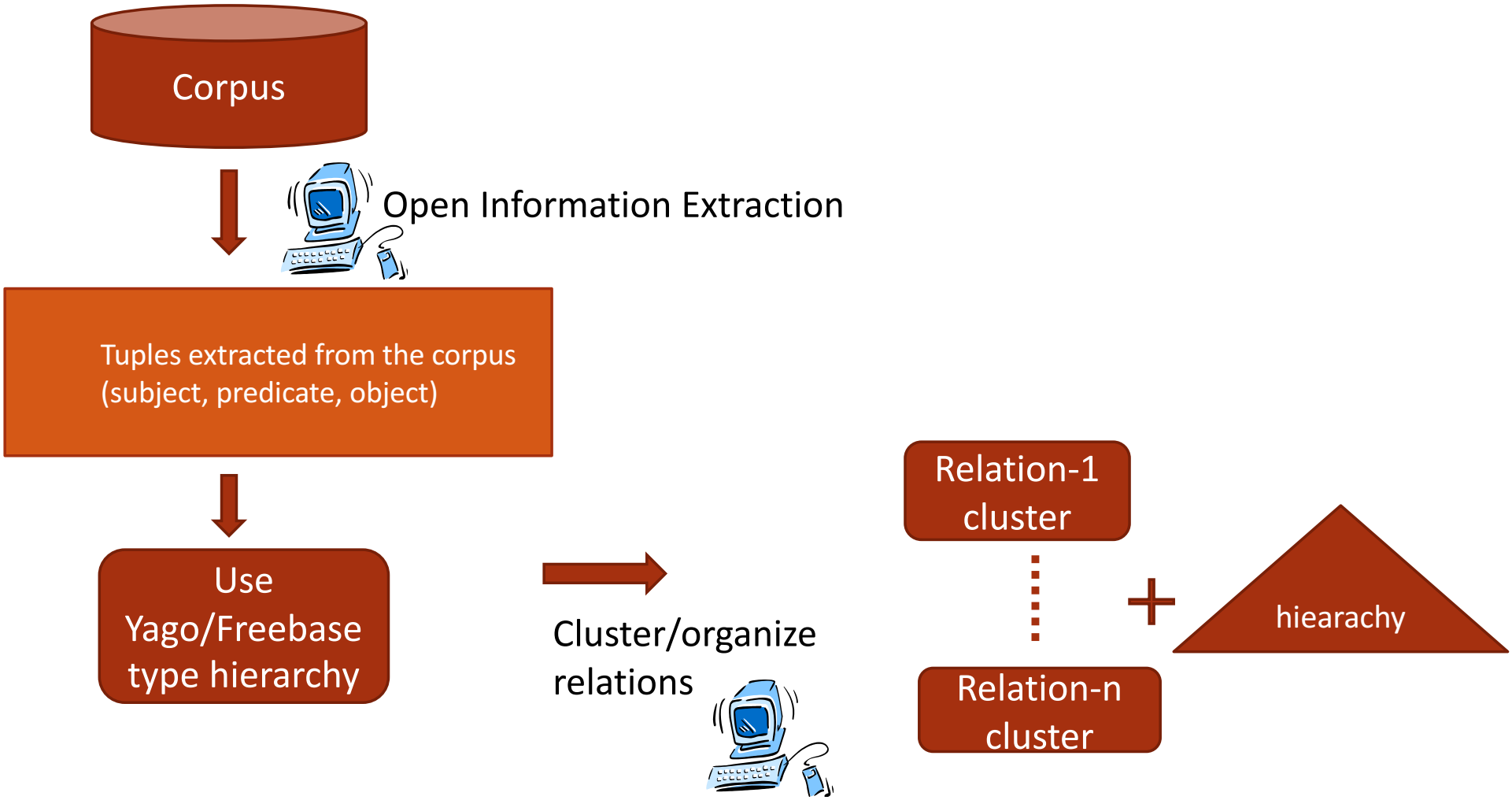
[Identifying Relations for Open Information Extraction, Fader et al, EMNLP 2011]

(4) Open domain IE

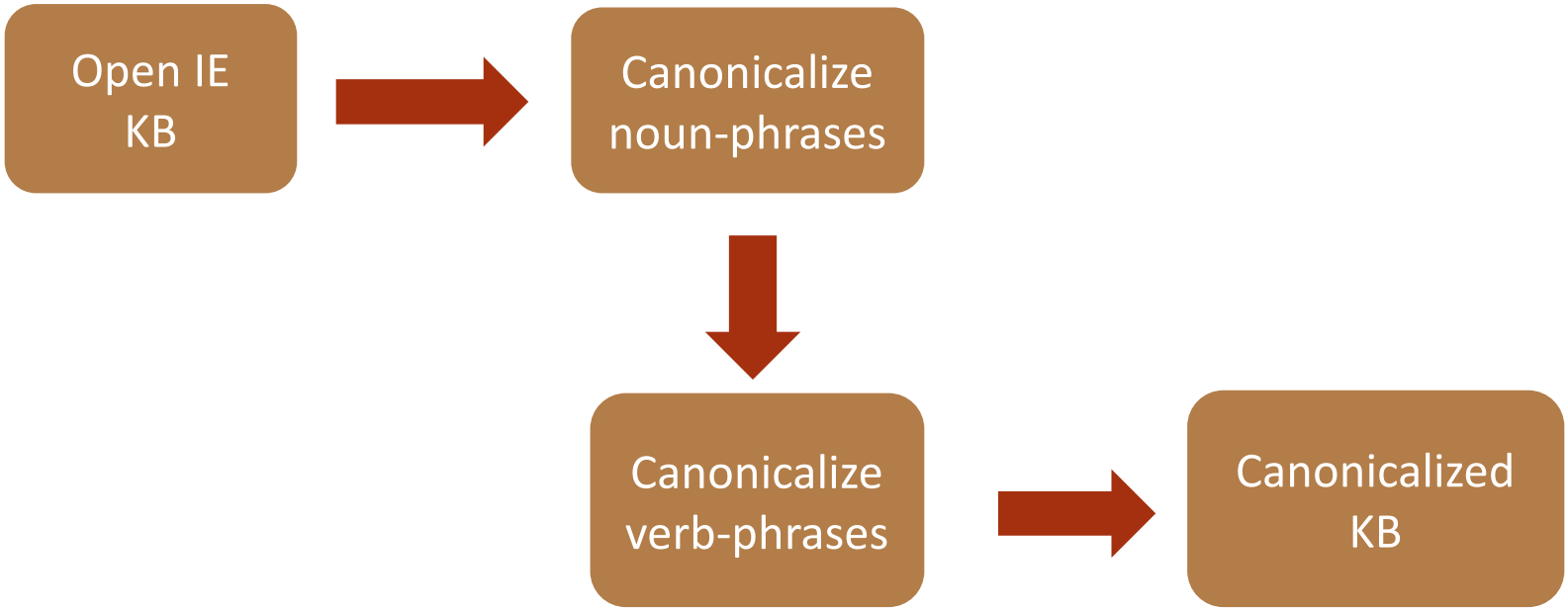
Defining domain	Learning extractors	Scoring extractions
		

(5) Hybrid approach

(adding structure to Open IE KB)

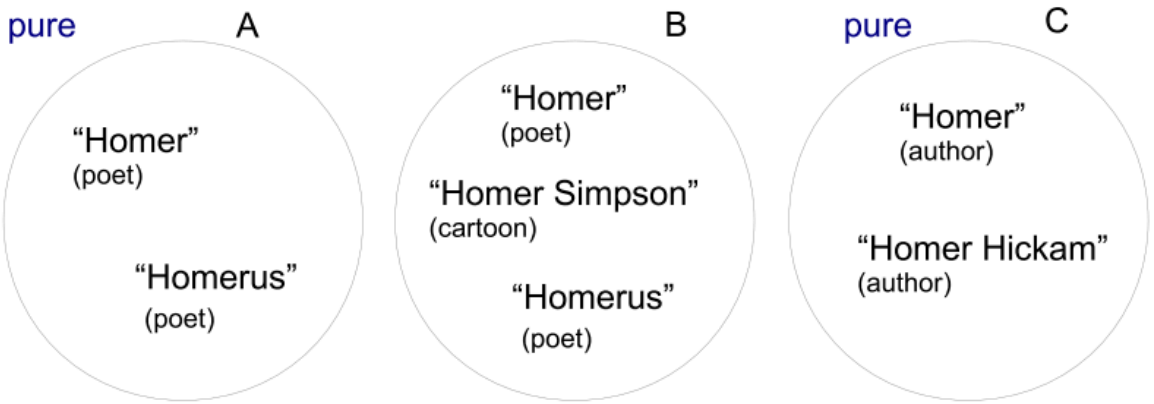


(5) Hybrid approach



(5) Hybrid approach

- ***Canonicalizing noun phrases***

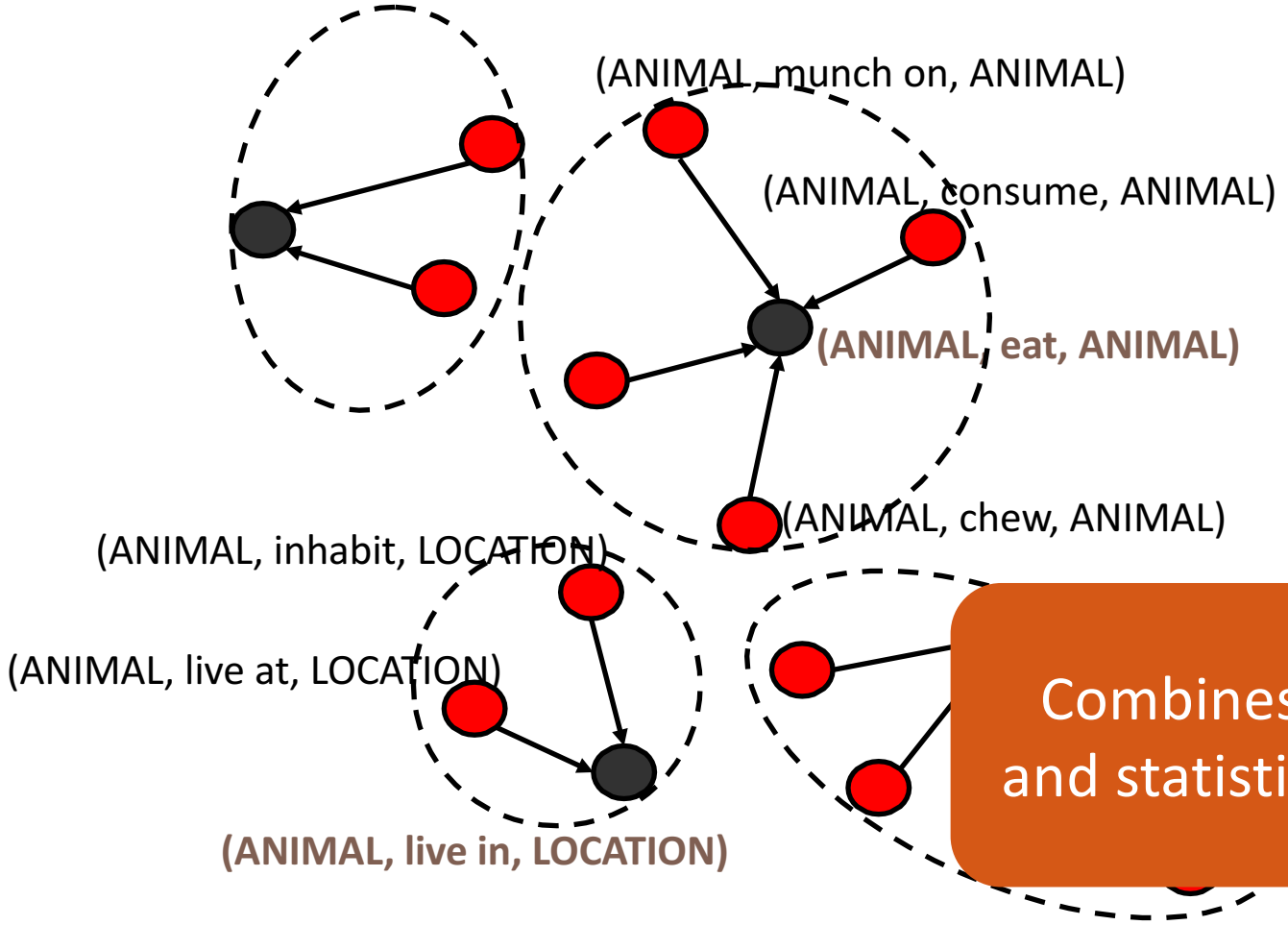


- ***Canonicalizing verb phrases***

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

Clustering based on statistical features

(5) Hybrid approach
Canonical schema induction (CASI)



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

(5) Hybrid approach

Defining domain	Learning extractors	Scoring extractions
		

Knowledge fusion with multiple extractors

VOTING (AND VS OR OF EXTRACTORS)

CO-TRAINING (MULTIPLE EXTRACTION METHODS)

MULTI-VIEW LEARNING (MULTIPLE DATA SOURCES)

CLASSIFIER

Information Extraction

Single extractor

Defining domain

Learning extractors

Scoring the extractions

×

Manual

Semi-automatic

Automatic

Fusing multiple extractors

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

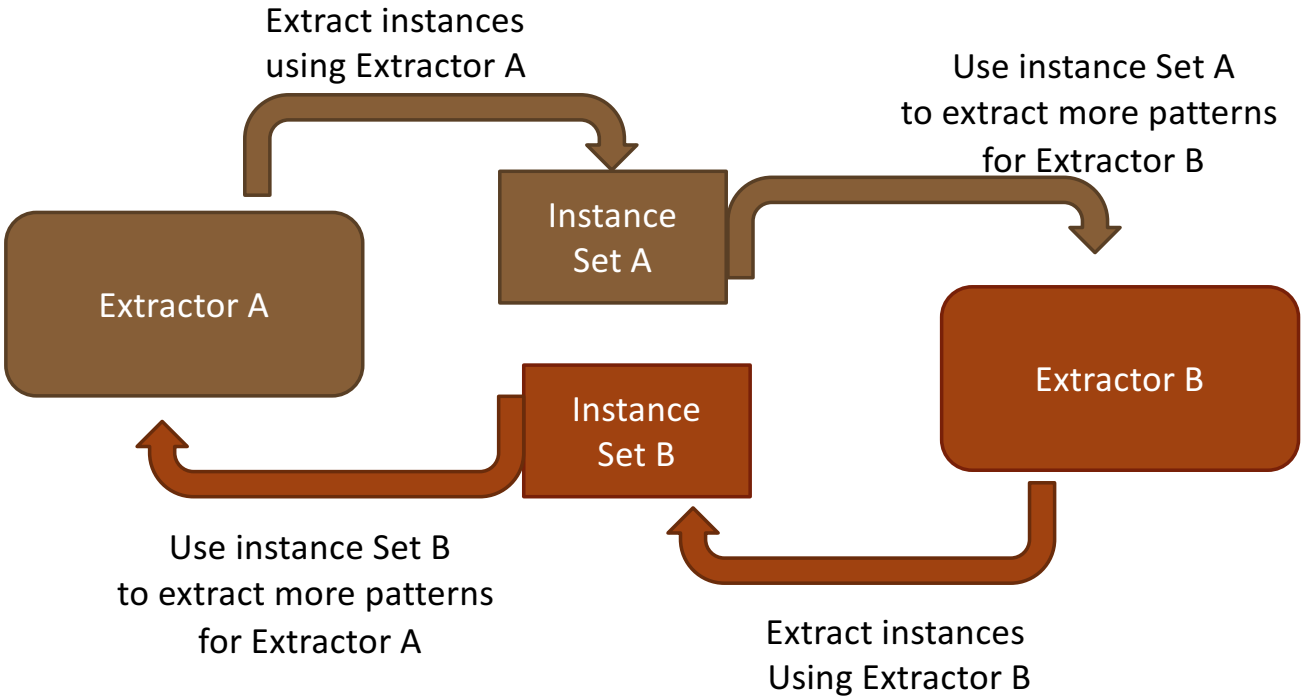
(1) Voting Schemes

- ***AND of two extractors:***
  - For a candidate extraction to be promoted to a fact in KB, both the extractors should support the fact
  - score(fact) = Min(score_extractor1(fact), score_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) = Max(score_extractor1(fact) , score_extractor2(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)
    …..

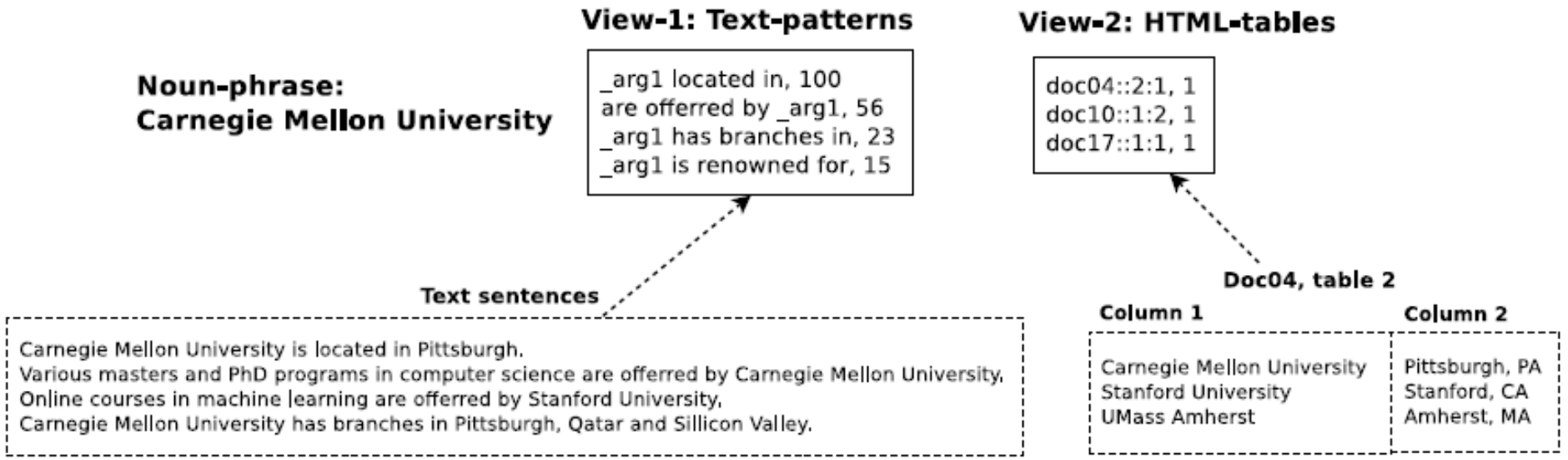
(2) Co-training



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

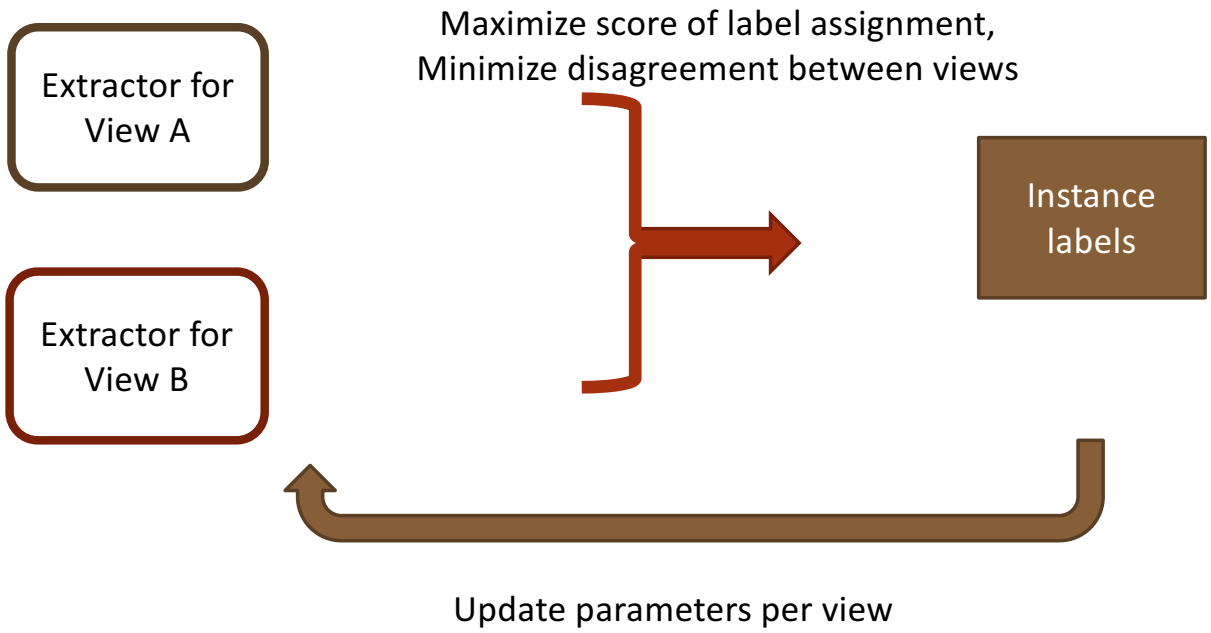
(3) Multi-view learning

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



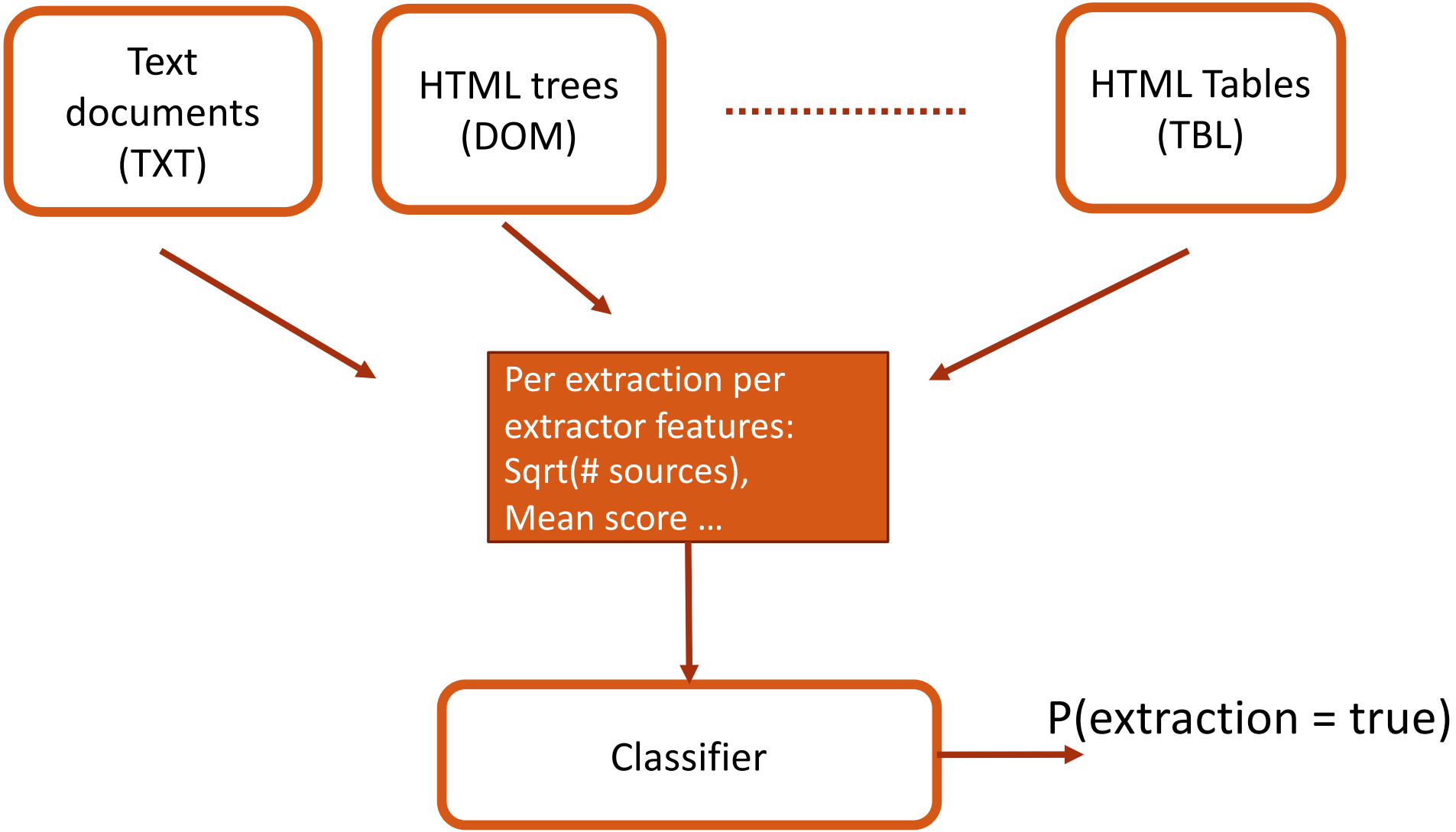
*[*Multi-View Hierarchical Semi-supervised Learning by Optimal Assignment of Sets of Labels to Instances, Dalvi et al. in preparation, link]

(3) Multi-view learning



*[*Multi-View Hierarchical Semi-supervised Learning by Optimal Assignment of Sets of Labels to Instances, Dalvi et al. in preparation, PDF]

(4) Classifiers



*[*Dong, Xin et al. **"Knowledge vault: a web-scale approach to probabilistic knowledge fusion."** *KDD* (2014)*]*

IE systems in practice

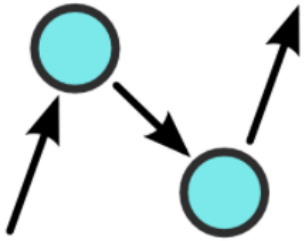
CONCEPTNET

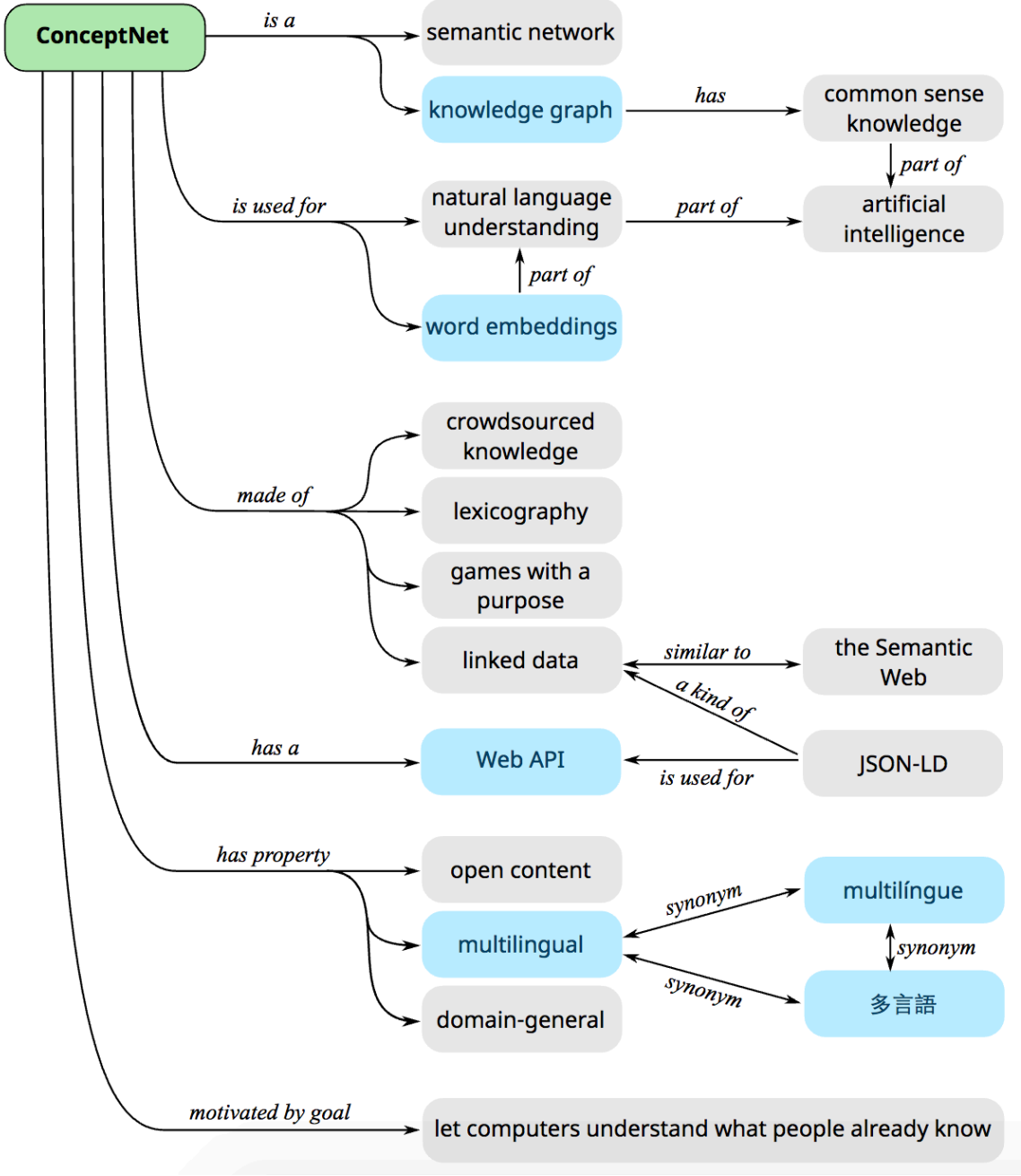
NELL

KNOWLEDGE VAULT

OPEN IE

ConceptNet

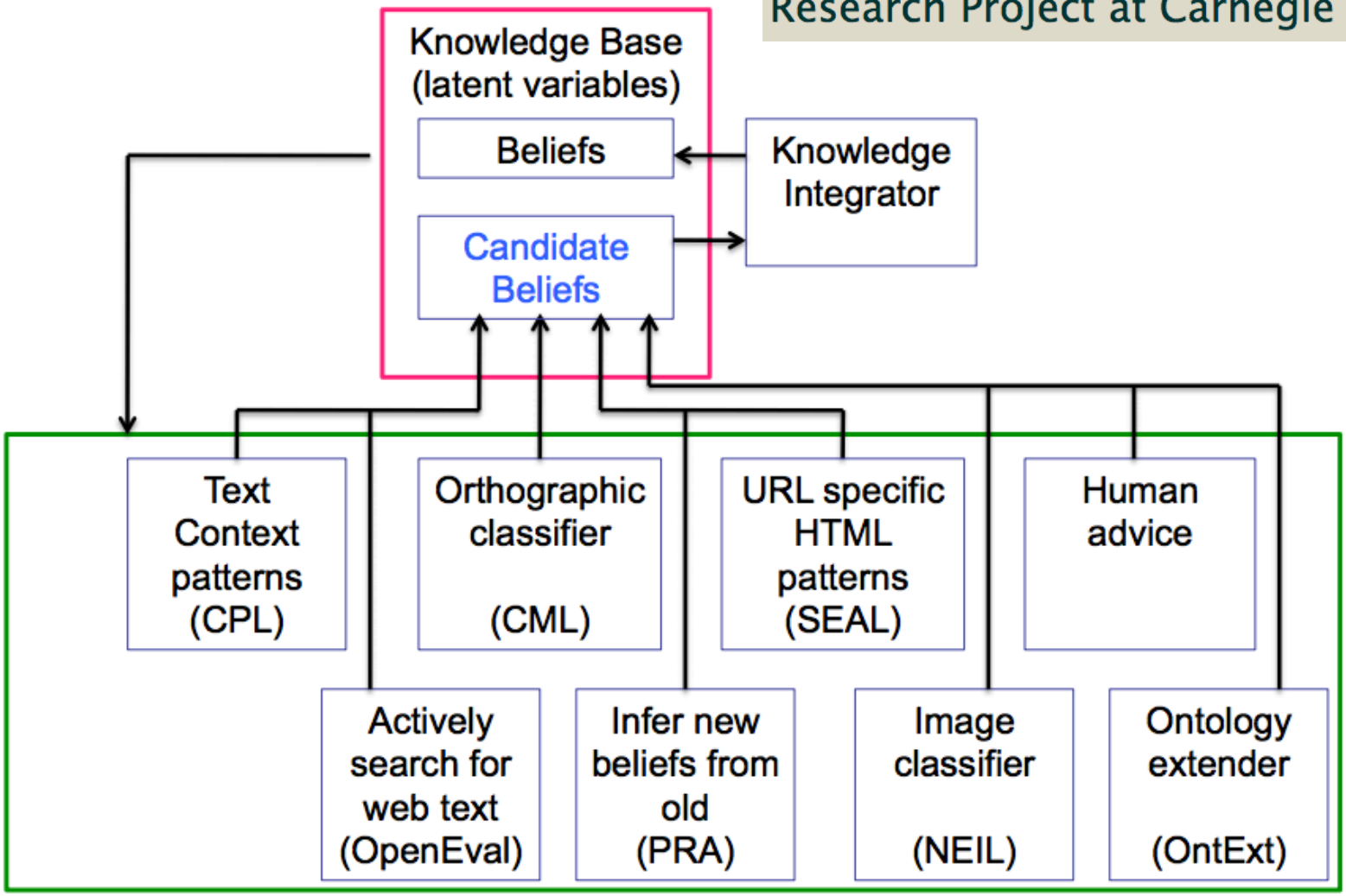



ConceptNet
is a
semantic network
knowledge graph
has
common sense knowledge
part of
artificial intelligence
is used for
natural language understanding
part of
part of
word embeddings
crowdsourced knowledge
made of
lexicography
games with a purpose
linked data
similar to
the Semantic Web
a kind of
has a
Web API
is used for
JSON-LD
has property
open content
multilingual
domain-general
synonym
multilíngue
synonym
synonym
多言語
motivated by goal
let computers understand what people already know

Never Ending Language Learning (NELL)

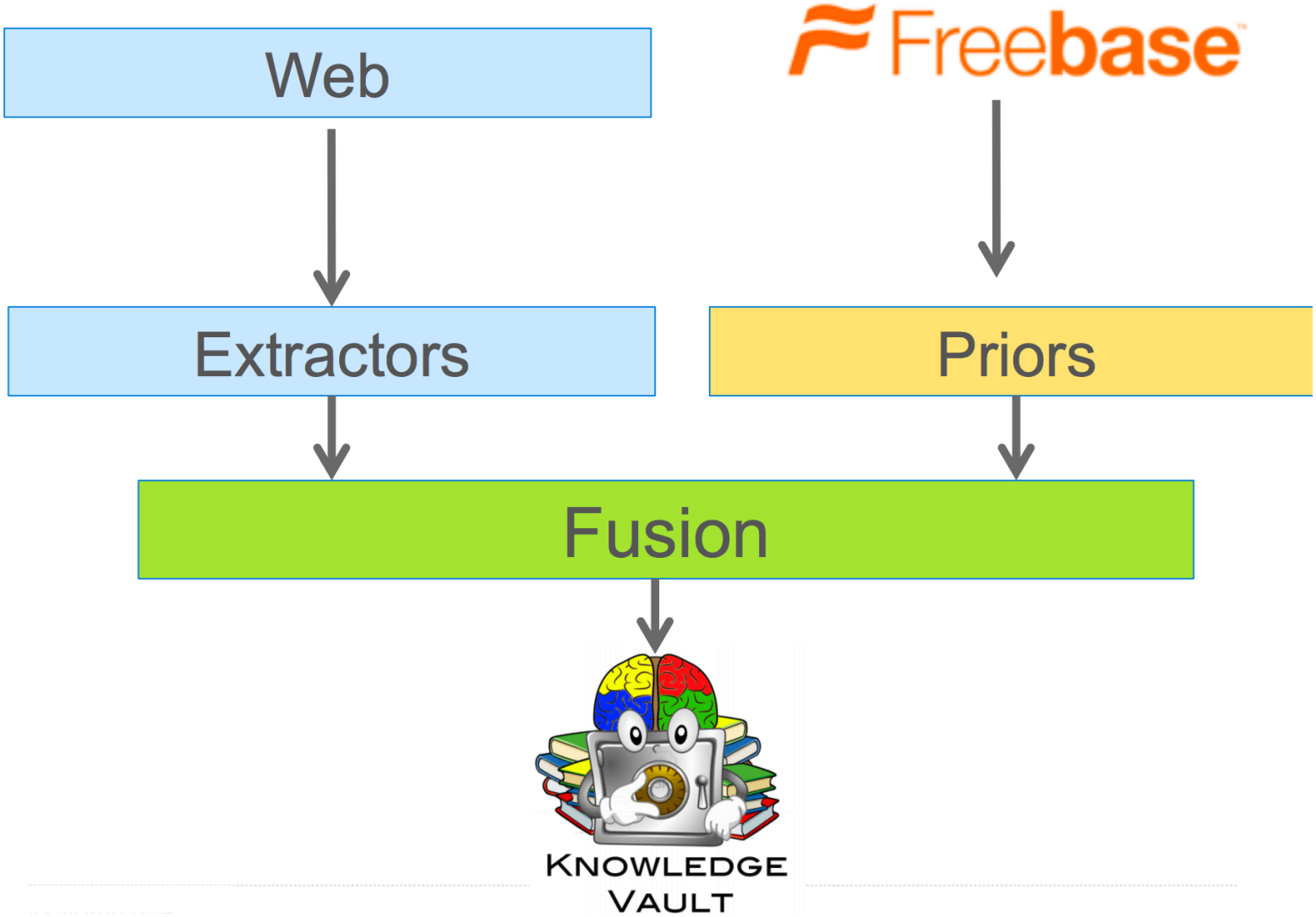
Read the Web

Research Project at Carnegie Mellon University



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

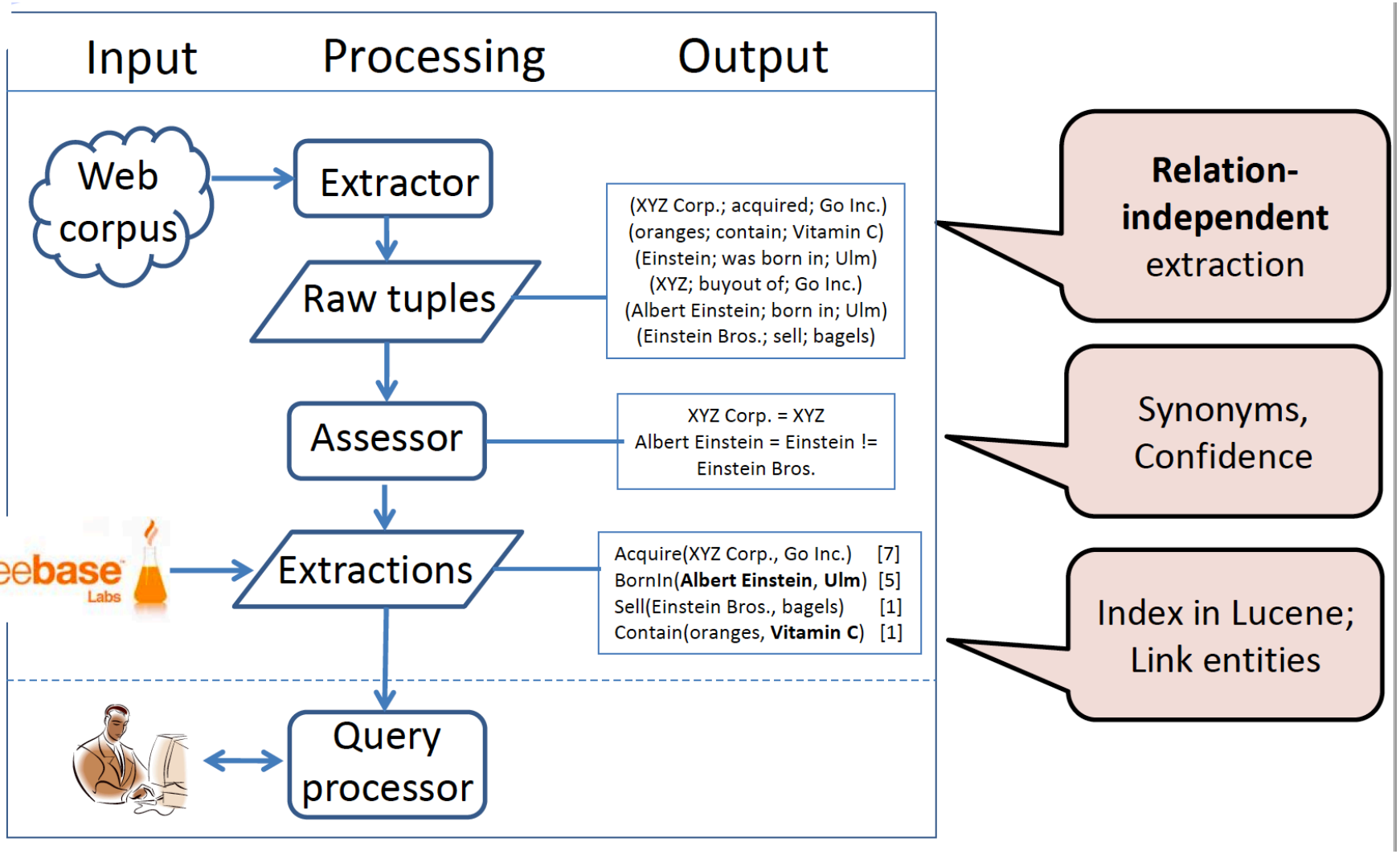
Knowledge Vault



[Architecture diagram taken from Kevin Murphy's slides]

Open IE (KnowItAll)

Open Information Extraction



[Architecture diagram taken from Oren Etzioni's slides]

IE systems in practice

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

Summary: Information Extraction

3 IMPORTANT SUB-PROBLEMS
(DEFINE DOMAIN, LEARN EXTRACTORS, SCORE EXTRACTIONS)

3 LEVELS OF SUPERVISION
(MANUAL, SEMI-SUPERVISED, UNSUPERVISED)

CATEGORIES OF IE TECHNIQUES

KNOWLEDGE FUSION WITH MULTIPLE EXTRACTORS
(CO-TRAINING, MULTI-VIEW LEARNING)

IE SYSTEMS IN PRACTICE

Thank You

SEE YOU AFTER THE COFFEE BREAK!