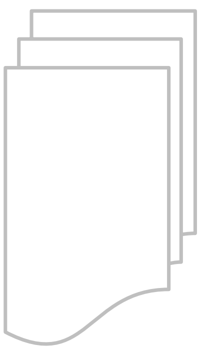
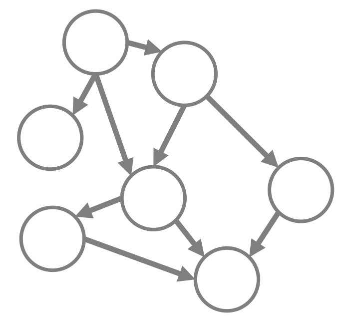
Part 1: Knowledge Graphs



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] 

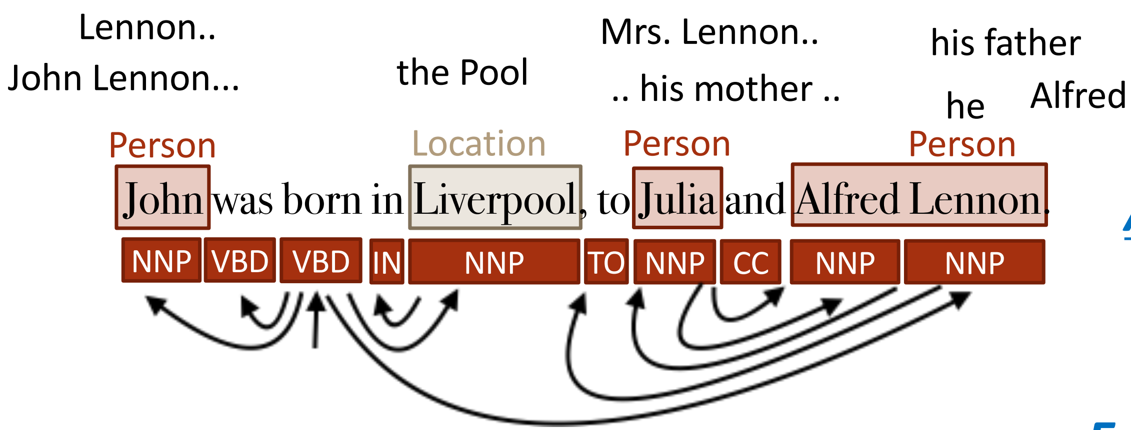
Coffee Break 

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

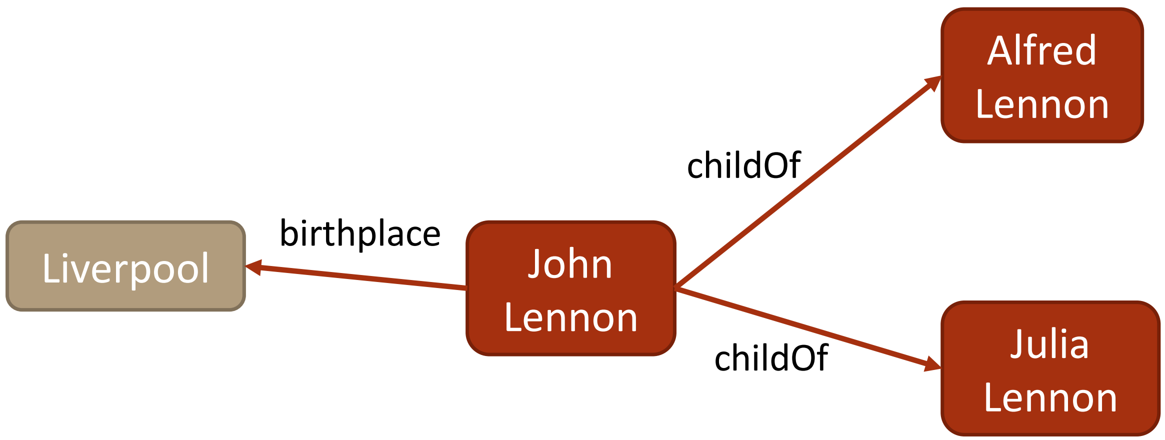
NLP



Annotated text

Information Extraction

Extraction graph



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 facts

×

3 LEVELS OF SUPERVISION

Supervised

Semi-supervised

Unsupervised

Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring the facts

×

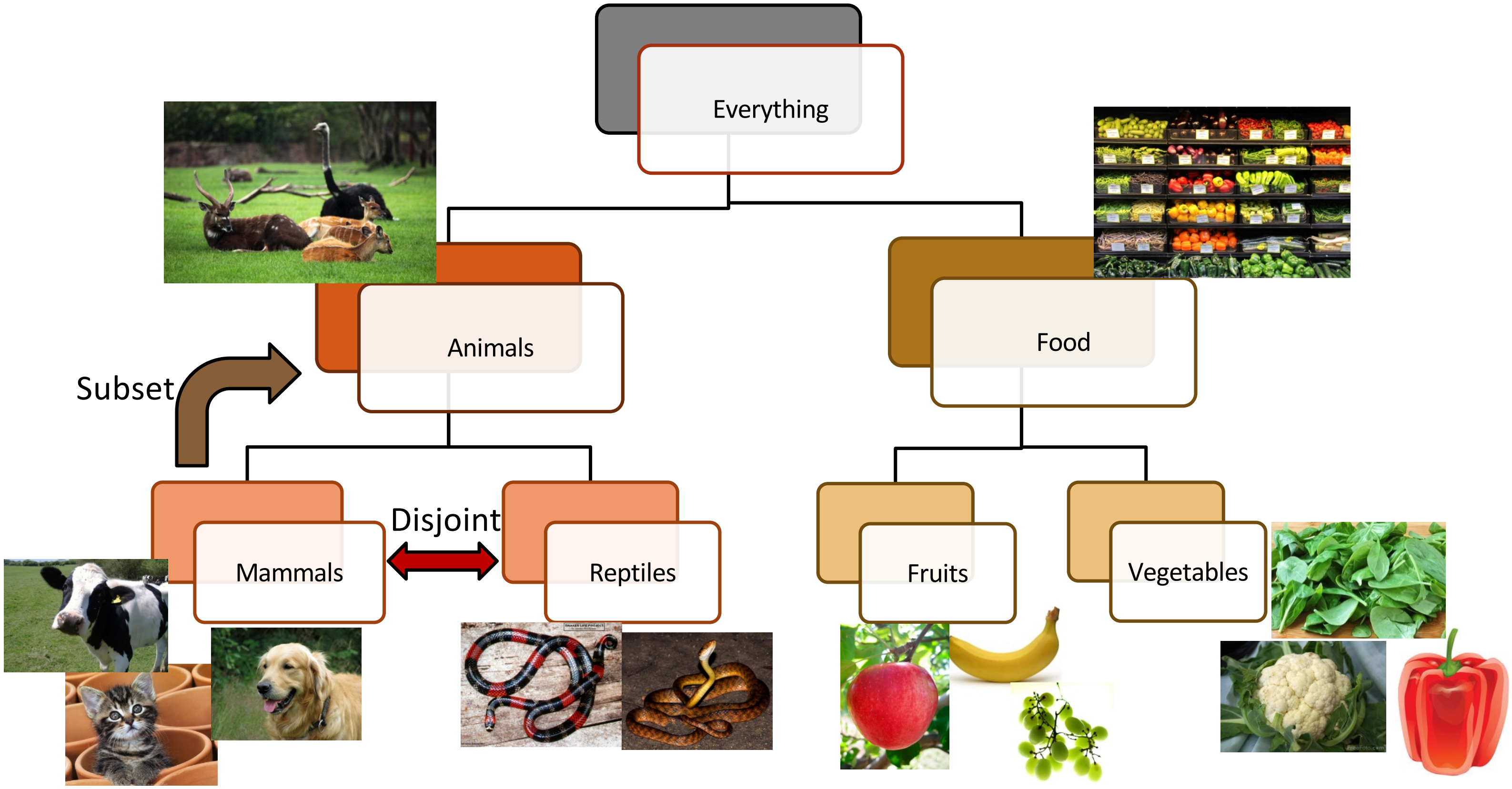
3 LEVELS OF SUPERVISION

Supervised 

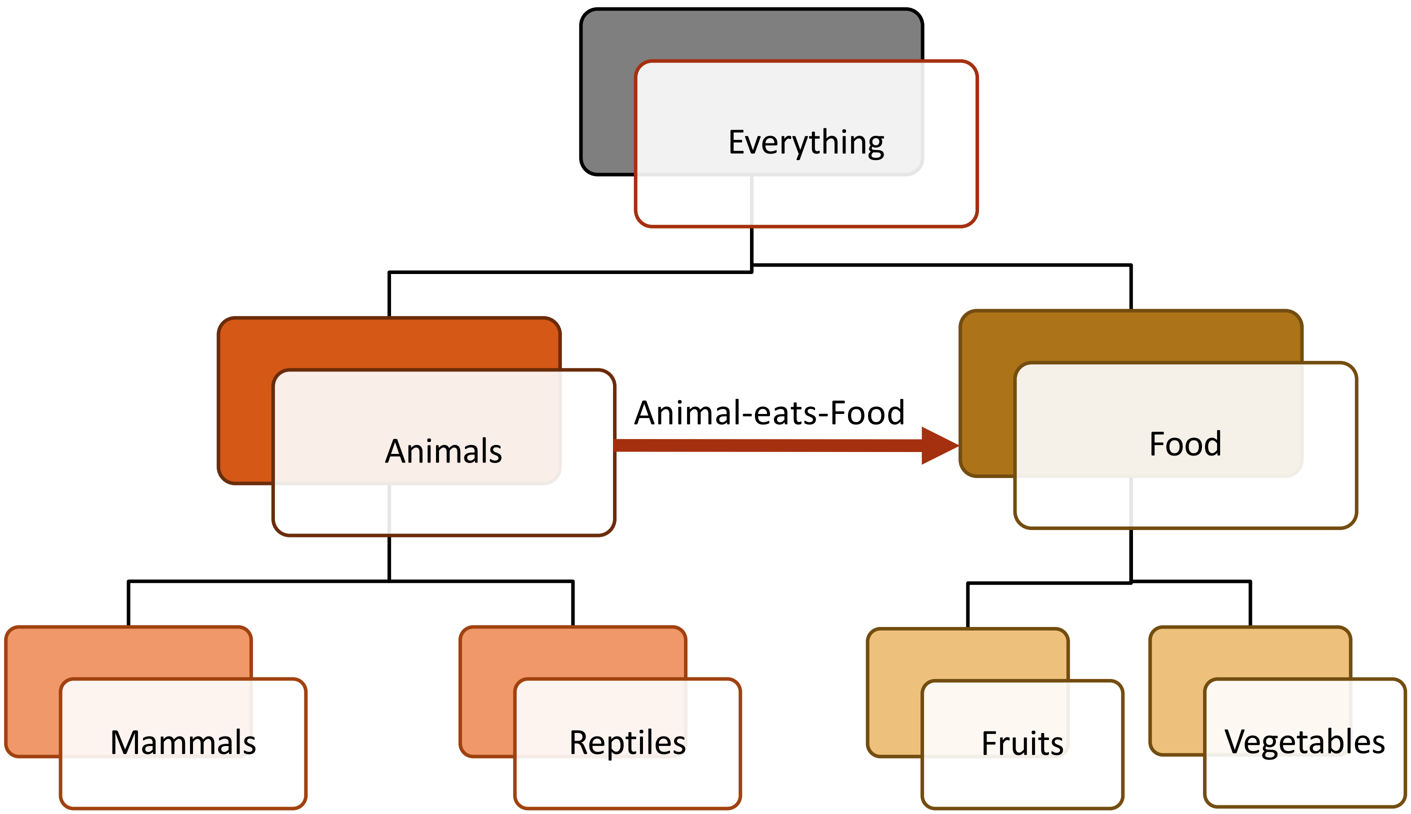
Semi-supervised 

Unsupervised 

Defining Domain: Manual



Defining Domain: Manual

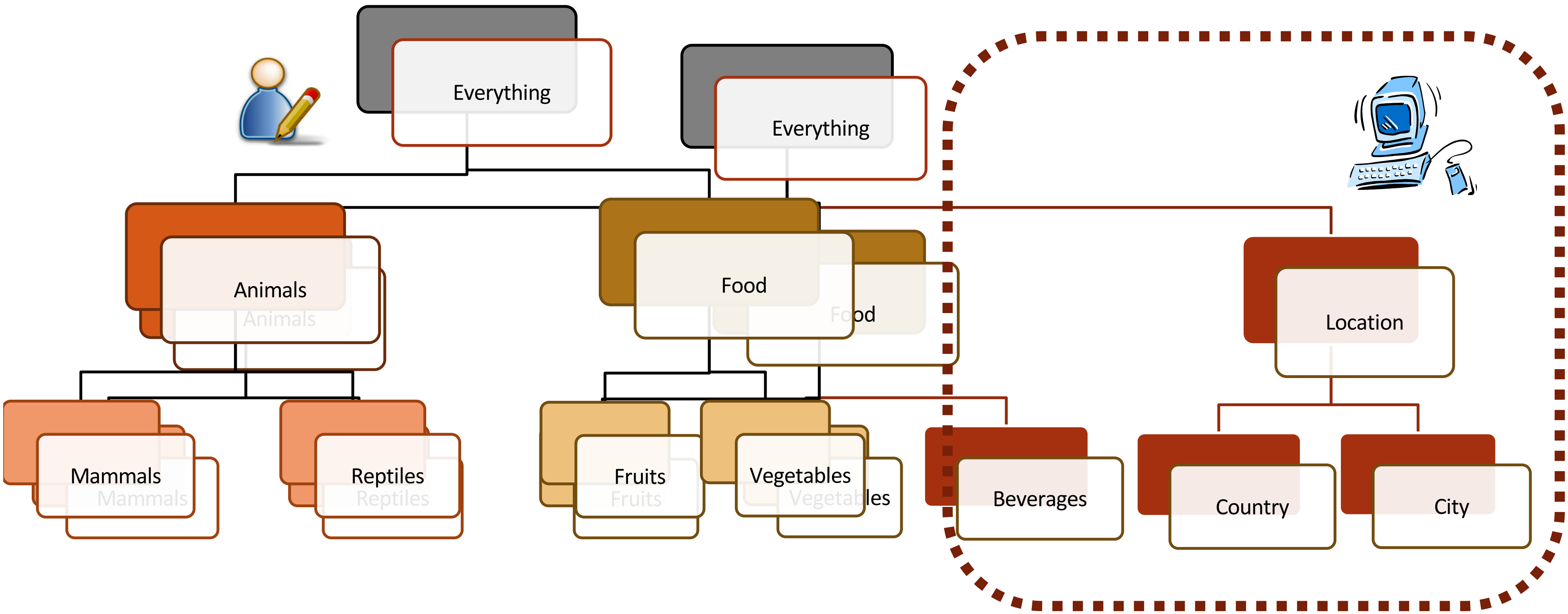


- Highly semantic ontology
- Leads to high precision extractions
- Expensive to create
- Requires domain experts

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

Defining Domain: Semi-automatic

- Subset of types are manually defined
- SSL methods discover new types from unlabeled data



Defining Domain: Semi-automatic

- Assume: 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 expert annotations needed**
- **Limited semantics**
- **Leads to noisy extractions**

Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring candidate facts

×

3 LEVELS OF SUPERVISION

Supervised

Semi-supervised

Unsupervised

Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring candidate facts

×

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)

Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring candidate facts

×

3 LEVELS OF SUPERVISION

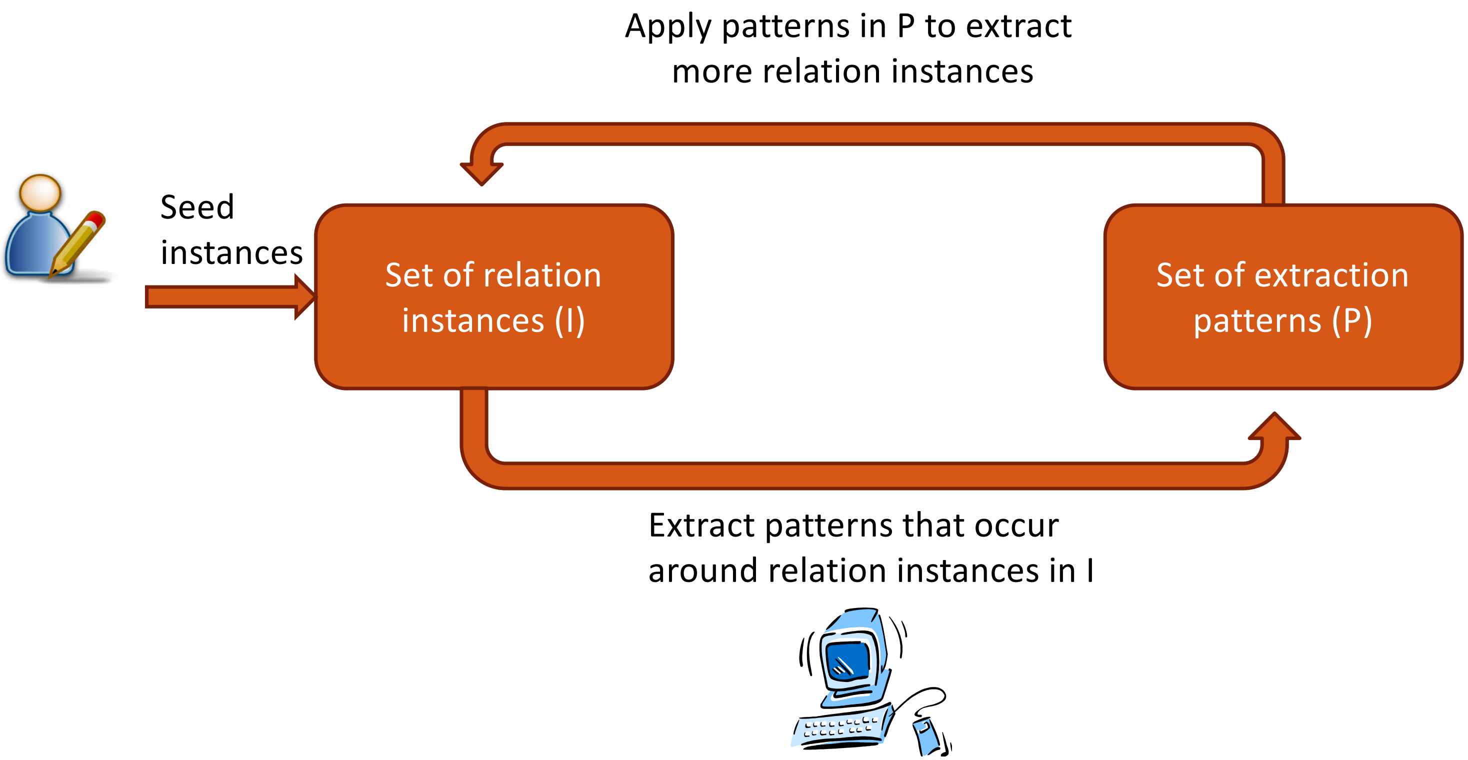
Supervised

Semi-supervised

Unsupervised

Learning Extractors: Semi-supervised

Bootstrapping



Learning Extractors: Semi-supervised

Person-member of-Band

Seed instances

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

Add top-k instances

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

Learn patterns

<PERSON> works for <BAND>
<PERSON> is part of <BAND>
<BAND> includes <PERSON>
<BAND> was admired by <PERSON>

Apply patterns

Semantic Drift!

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

Learning Extractors : Interactive

Person-member of-Band

Seed instances

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

Positive instances

+
-

Candidate facts
(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

Helps reduce semantic drift!

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

Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring candidate facts

×

3 LEVELS OF SUPERVISION

Supervised 

Semi-supervised 

Unsupervised 

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

*[**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)

Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring candidate facts

3 LEVELS OF SUPERVISION

Supervised 

Semi-supervised 

Unsupervised 

Scoring the candidate facts

- Human defined scoring function or
  Scoring function learnt using supervised ML with large amount of training data
  {expensive, high precision}

- Small amount of training data is available
  scoring refined over multiple iterations
  using both labeled and unlabeled data

- Completely automatic (Self-training)
  Confidence(extraction pattern) $\propto$ (#unique instances it could extract)
  Score(candidate fact) $\propto$ (#distinct extraction patterns that support it)
  {cheap, leads to semantic drift}

Impact of early supervision

Domain expertise needed

Defining domain

Enables inheritance and mutual exclusion at extractor level

Puts constraints on the space of possibly true extractions

Extractors for each relation of interest

Early removal of noisy extraction pattern can avoid semantic drift in later stages

Scoring the candidate facts

Effect of supervision on extractions

Precision,
Human efforts

Recall,
Speed

Information Extraction

3 IMPORTANT SUB-PROBLEMS

CATEGORIES OF IE TECHNIQUES

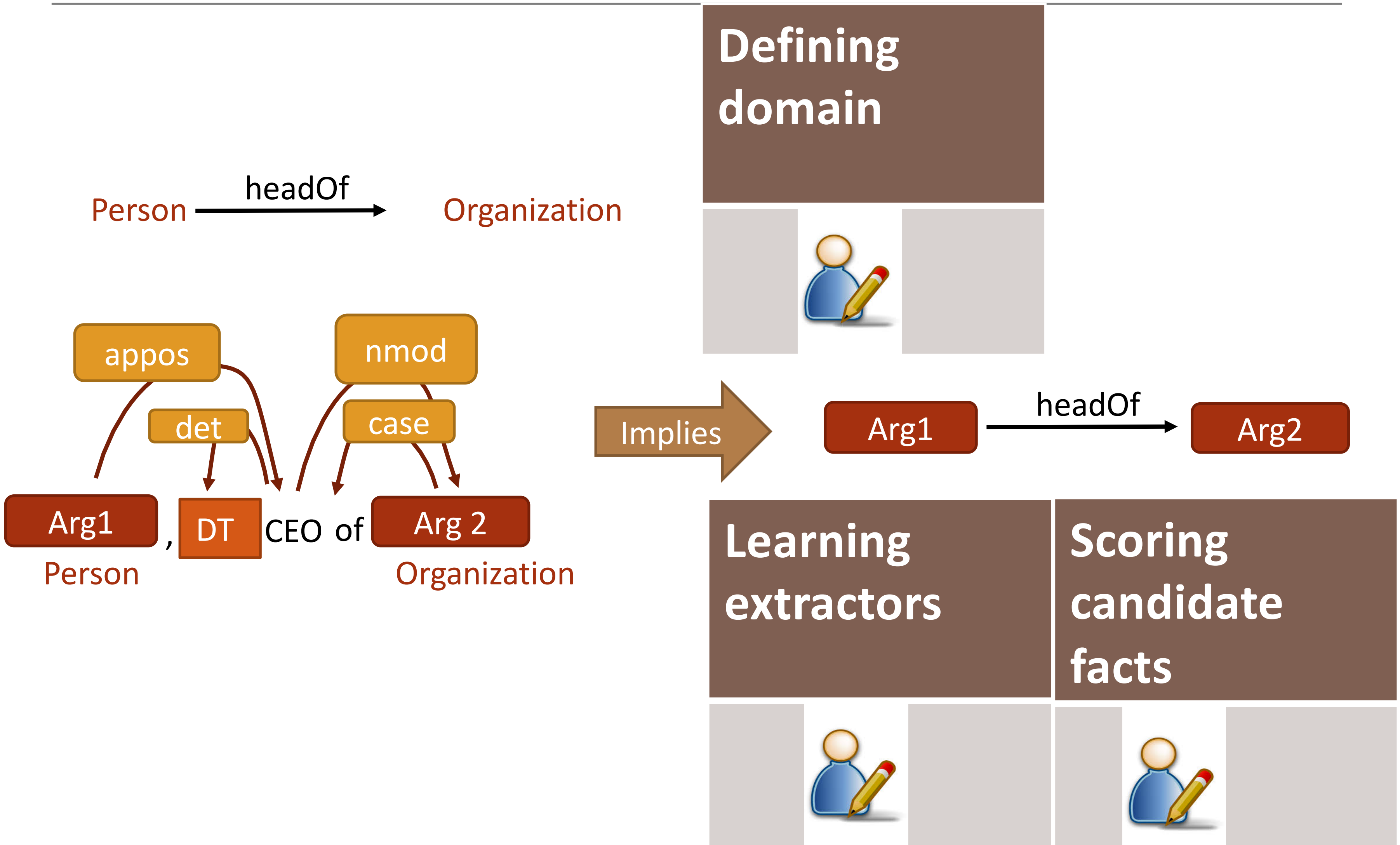
KNOWLEDGE FUSION

IE SYSTEMS IN PRACTICE

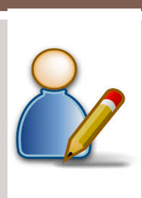
Categories of IE Techniques

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

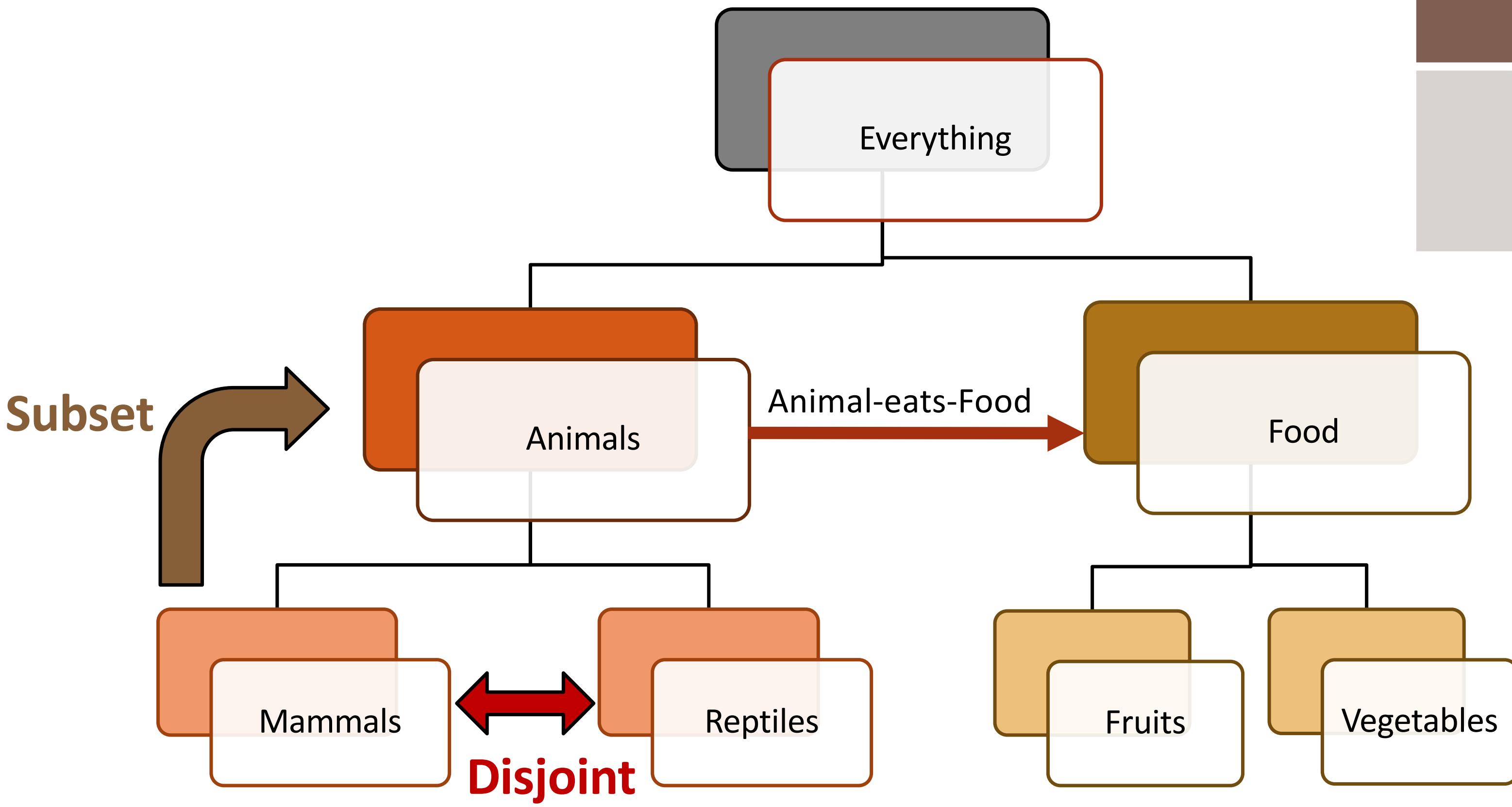


(1) Narrow domain patterns

Defining domain	Learning extractors	Scoring candidate facts
		

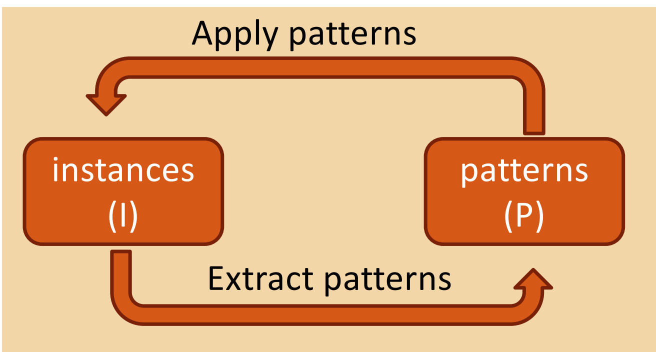
(2) Ontology based extraction

Defining domain

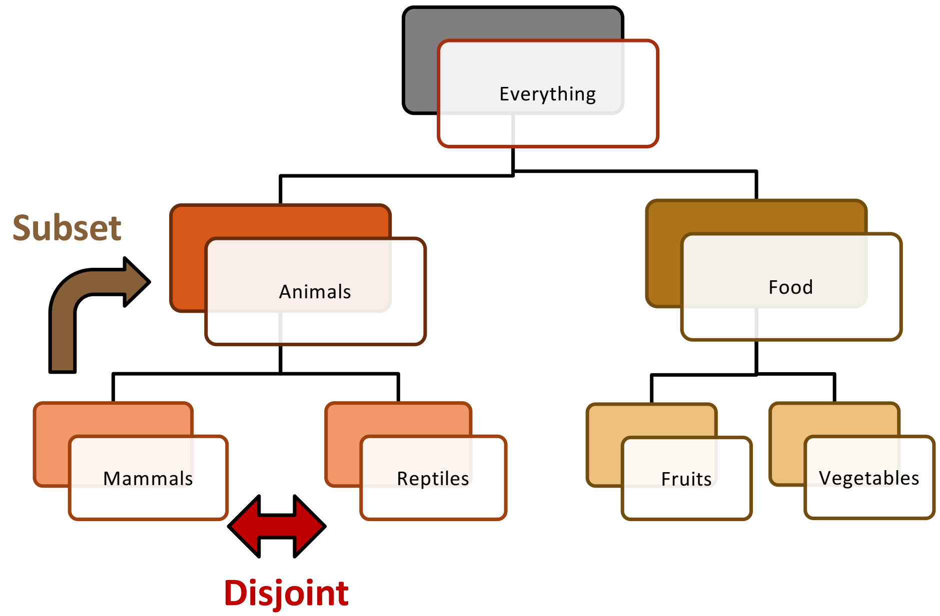


(2) Ontology based extraction

Bootstrapping



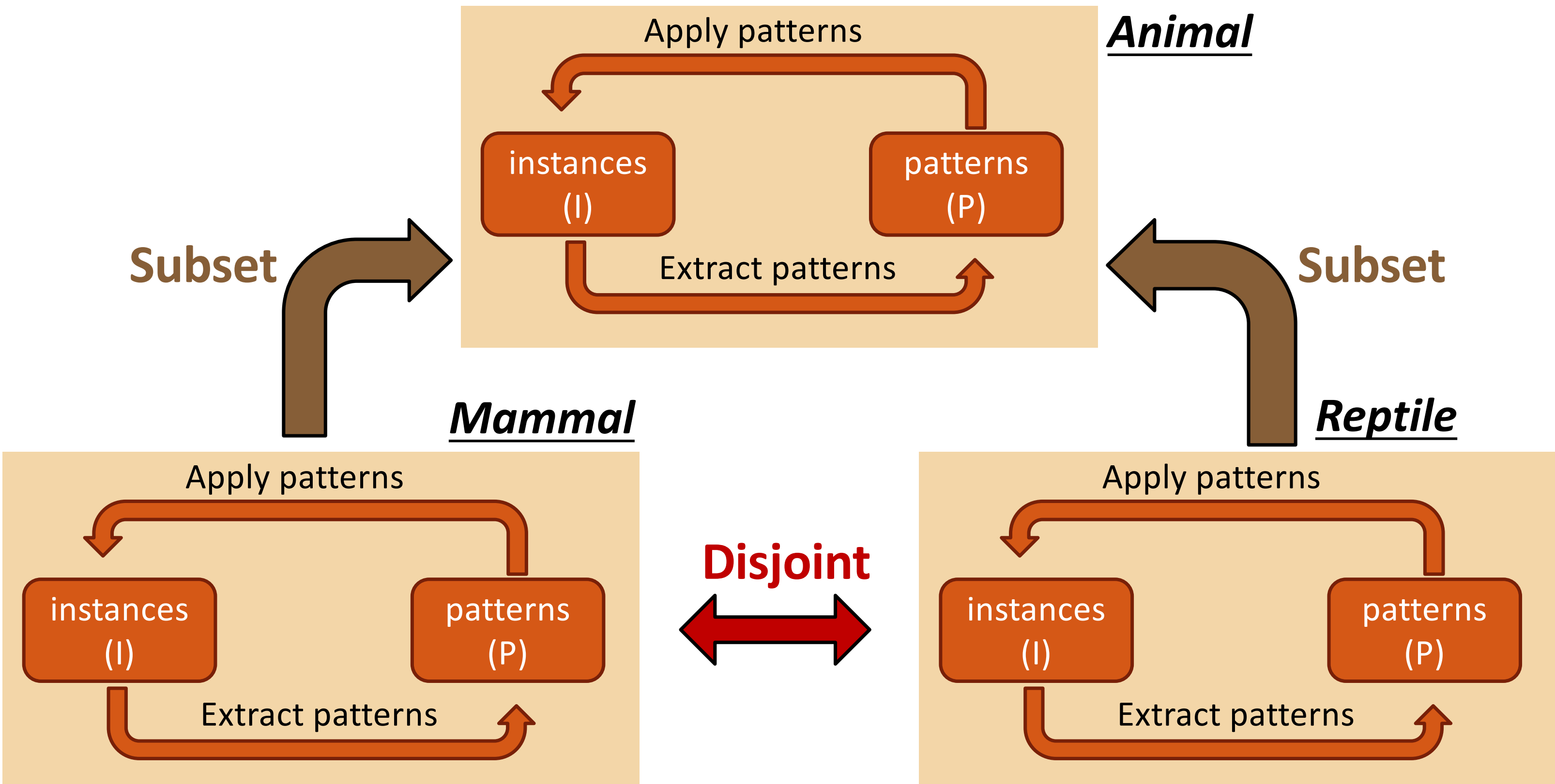
Ontological constraints



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

(2) Ontology based extraction

Coupled Bootstrap learning

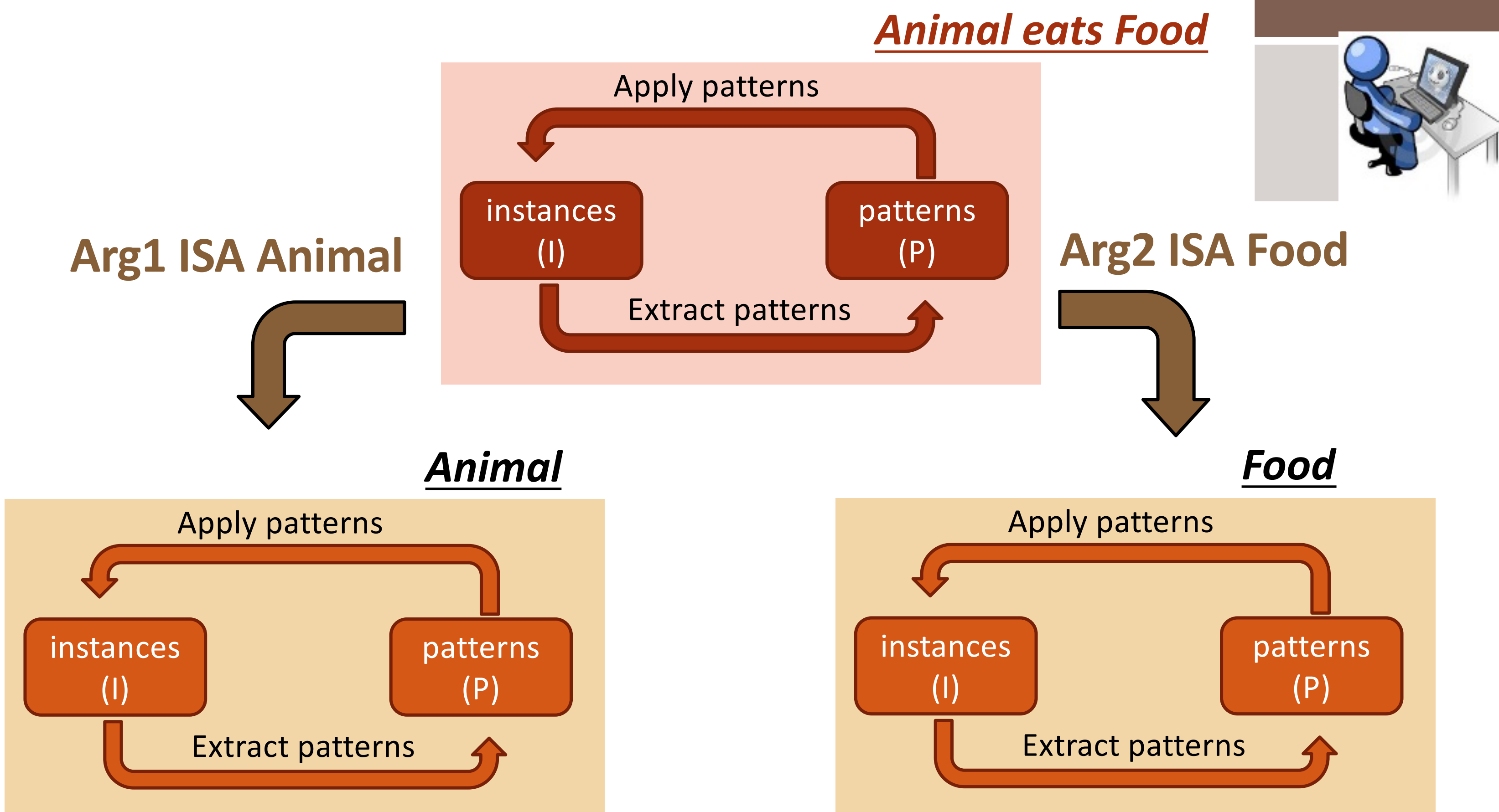


*[**Toward an Architecture for Never-Ending Language Learning**, Carlson et al. AAAI 2010]*

(2) Ontology based extraction

Learning extractors

Coupled Bootstrap learning



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

(2) Ontology based extraction

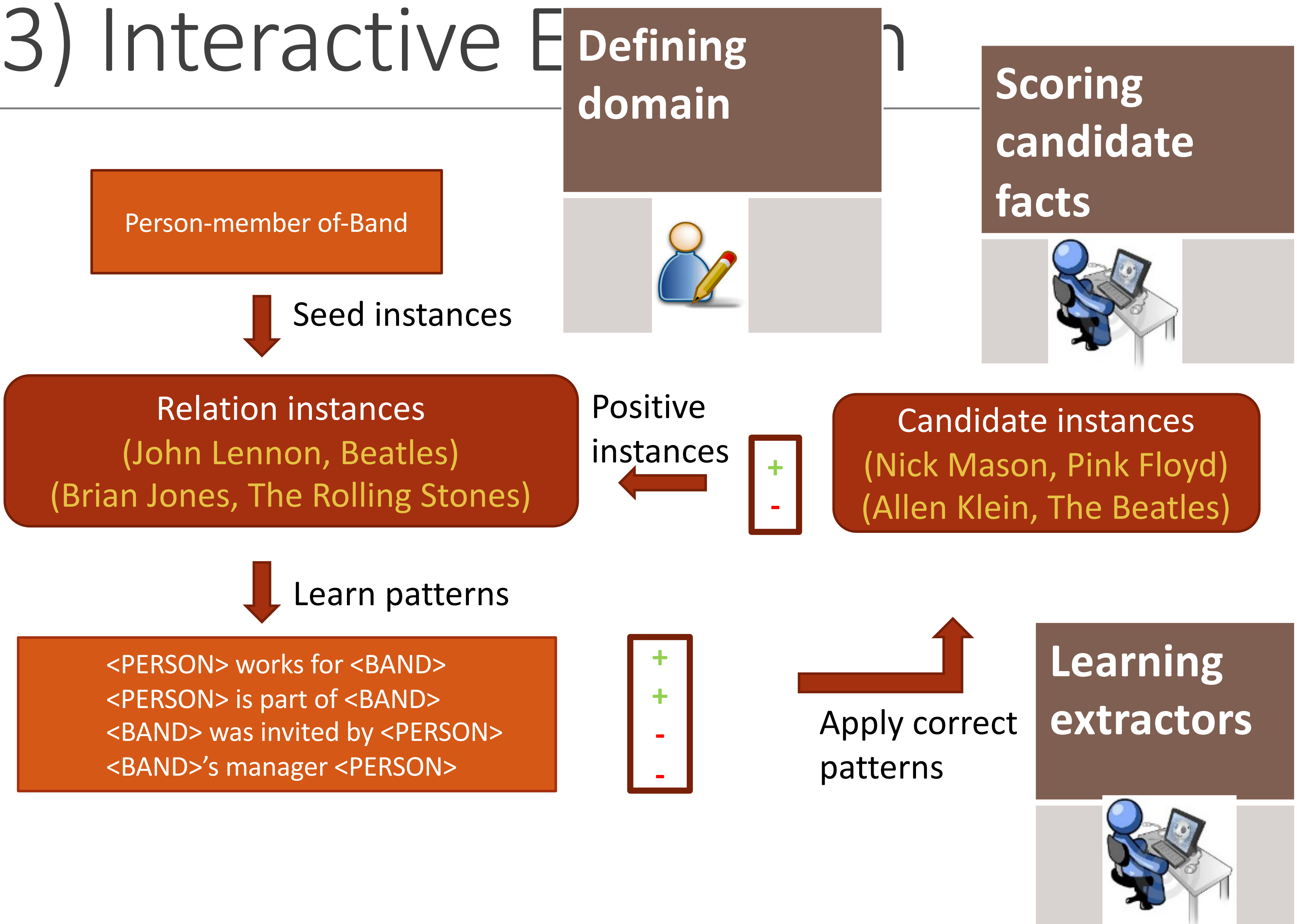
- Self-training for scoring candidate facts
  - Confidence(extraction pattern) ∝ (#unique instances it could extract)
  - Score(candidate fact) ∝ (#distinct extraction patterns that support it)

Scoring candidate facts

(2) Ontology based extraction

Defining domain	Learning extractors	Scoring candidate facts
		

(3) Interactive E n



(3) Interactive Extraction

Defining domain	Learning extractors	Scoring candidate facts
		

Can we do Web-scale IE?

1. Narrow domain patterns
2. Ontology based extraction
3. Interactive extraction

Assume expert input
Biased towards high precision
High costs

4. Open domain IE
5. Hybrid approach
   (Adding structure to OpenIE KB)

(4) Open domain IE

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

Defining domain

Learning extractors

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

Scoring candidate facts

n in, Hampstead) : 0.88

uburb of, London) : 0.9

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

(4) Open domain IE

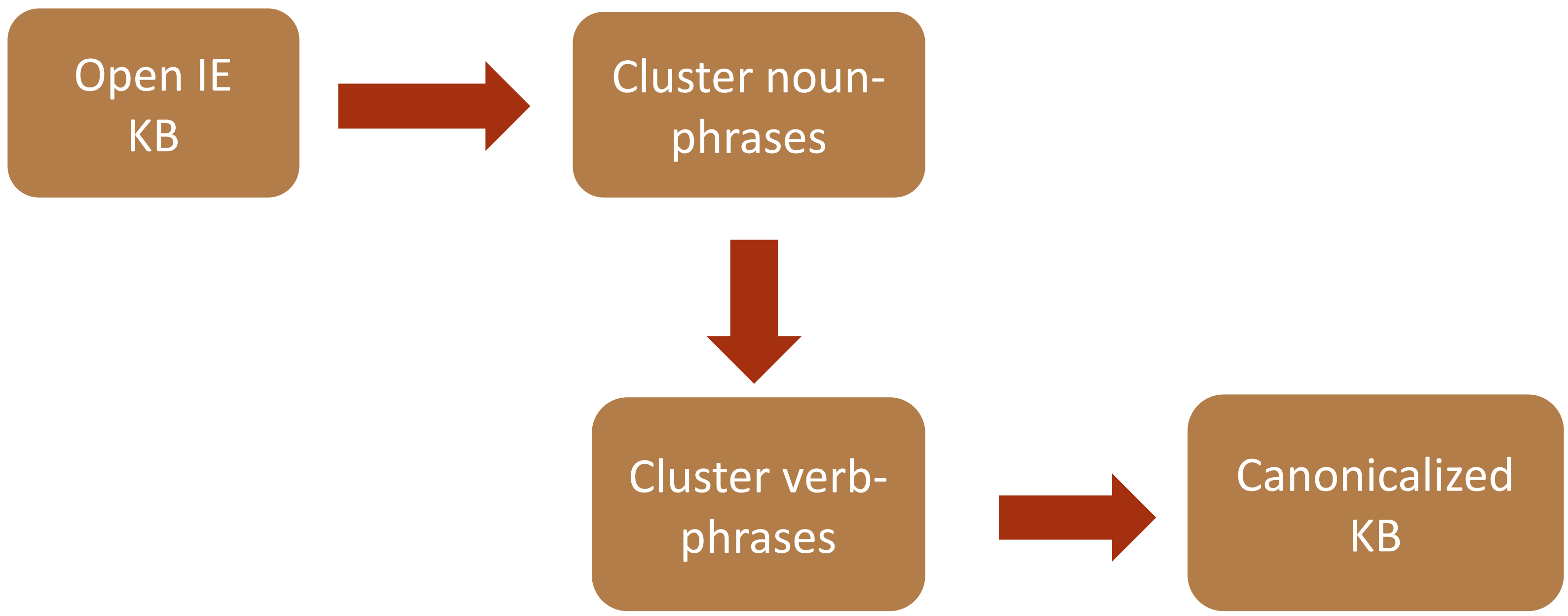
Defining domain	Learning extractors	Scoring candidate facts
		

Pros and Cons of Open domain IE

- Open domain IE paradigm can be easily applied
  - on a large scale corpus
  - in a new domain (no training data)

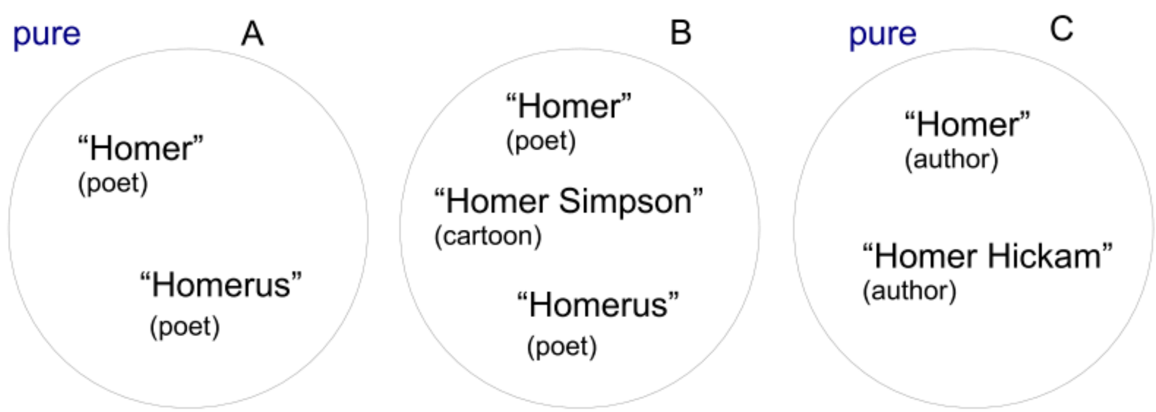
- **Main disadvantages**
  - Poor aggregation
    Doesn't detect different surface forms for same entity or relation
  - Lack of semantics
    OpenIE merely tells us how many times the lexical fact occurred in a corpus

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



(5) Hybrid approach

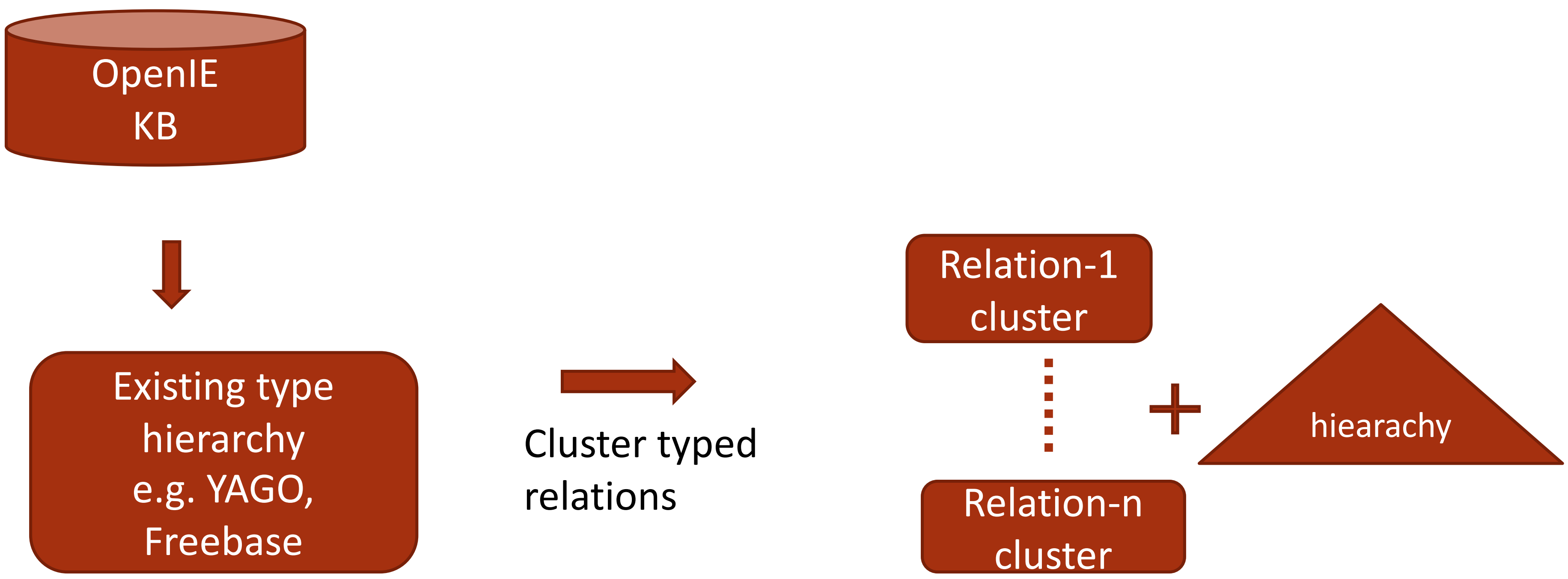
- ***Clustering entities***



- ***Clustering relations***

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`

(5) Hybrid approach



(5) Hybrid approach

	Defining domain	Learning extractors	Scoring candidate facts
Open domain IE			

+

Distant supervision to add structure

Categories of IE Techniques

1. Narrow domain patterns
2. Ontology based extraction
3. Interactive extraction

Assume expert input
Biased towards high precision
High cost

4. Open domain IE
5. Hybrid approach
   (Adding structure to OpenIE KB)

No expert annotations
Biased towards high recall
Low cost

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 candidate facts

×

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

Knowledge fusion schemes

- Voting (AND vs OR of extractors)
- Co-training (multiple extraction methods)
- Multi-view learning (multiple data sources)
- Classification

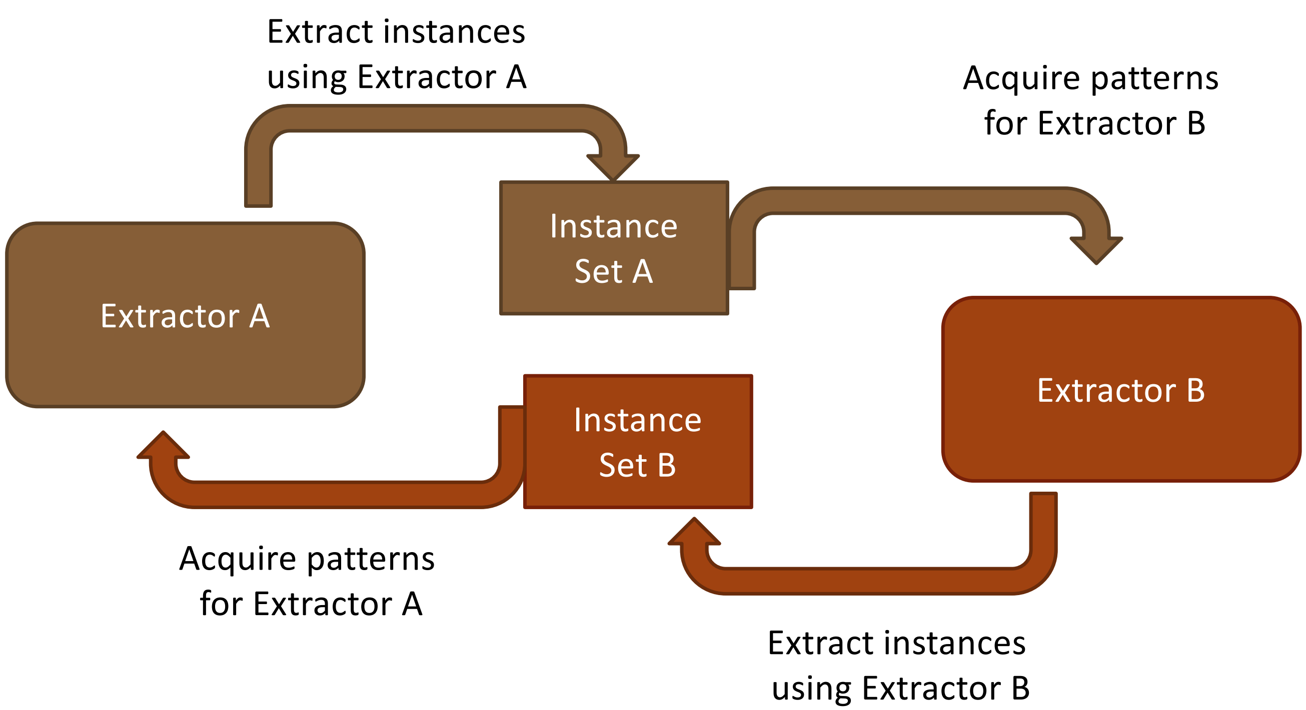
(1) Voting Schemes

- ***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))

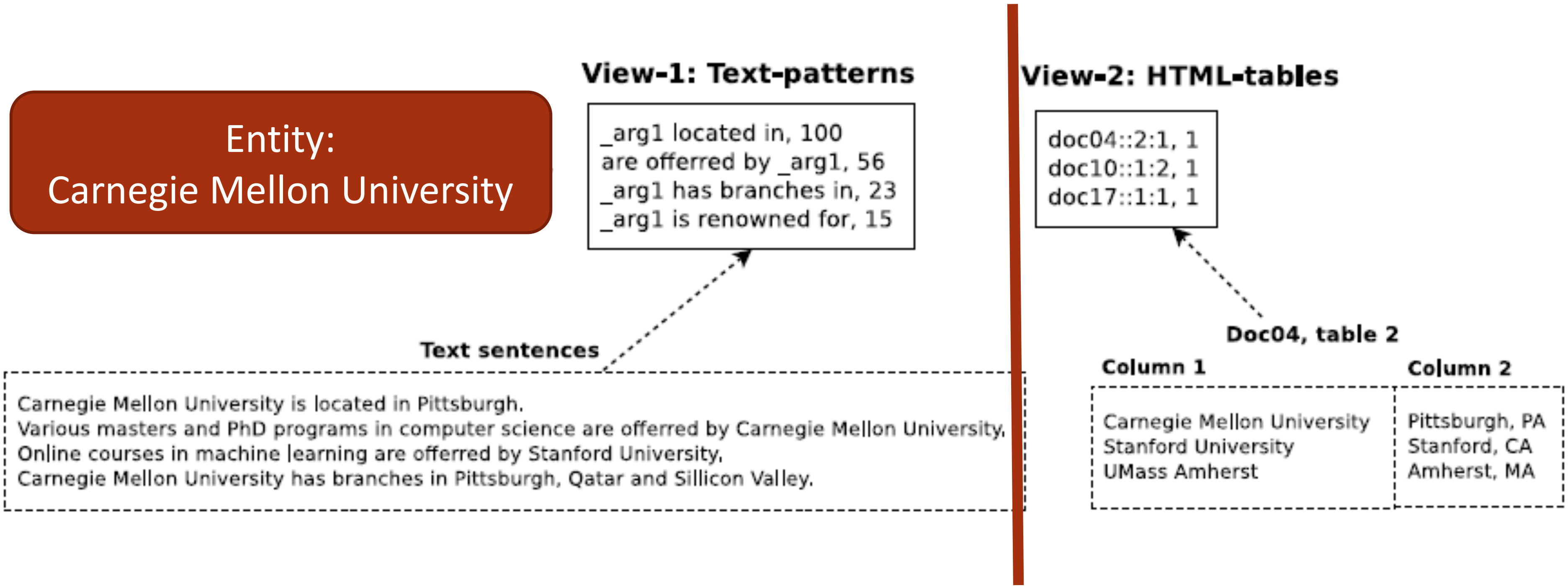
- **Hand-coded heuristic rules**
  - E.g. (at least one extractor has confidence > 0.9) or
    (two extractors support the fact with confidence > 0.6)
    …..

(2) Co-training



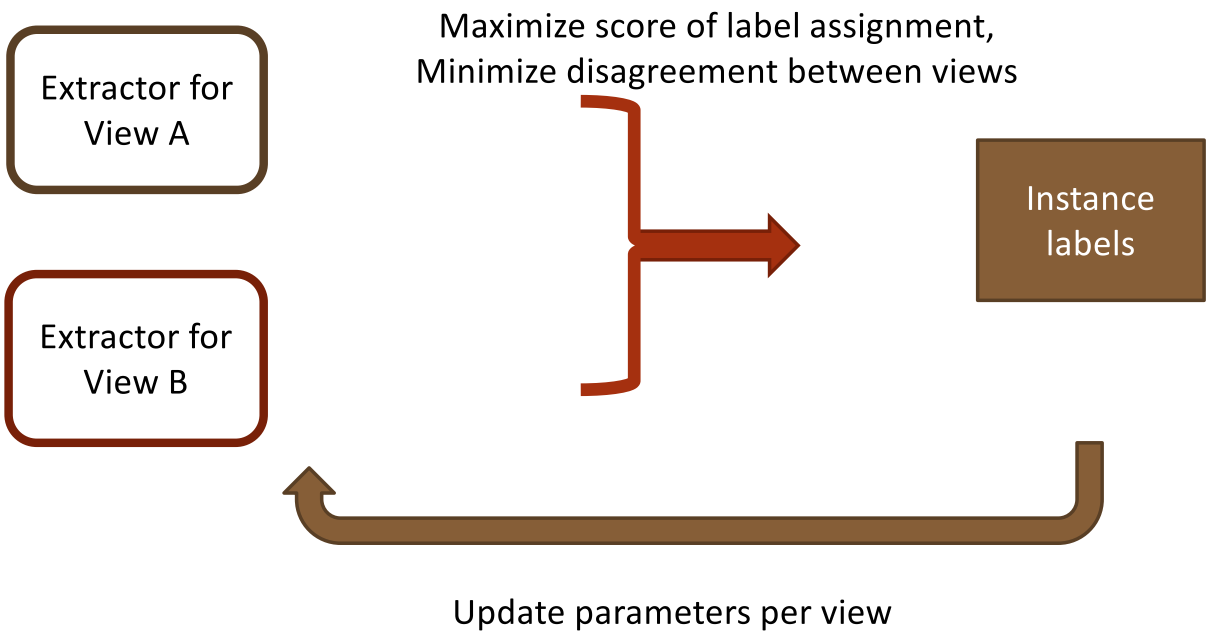
(3) Multi-view learning

- Task: Entity typing
- Each entity can be represented using two independent data views



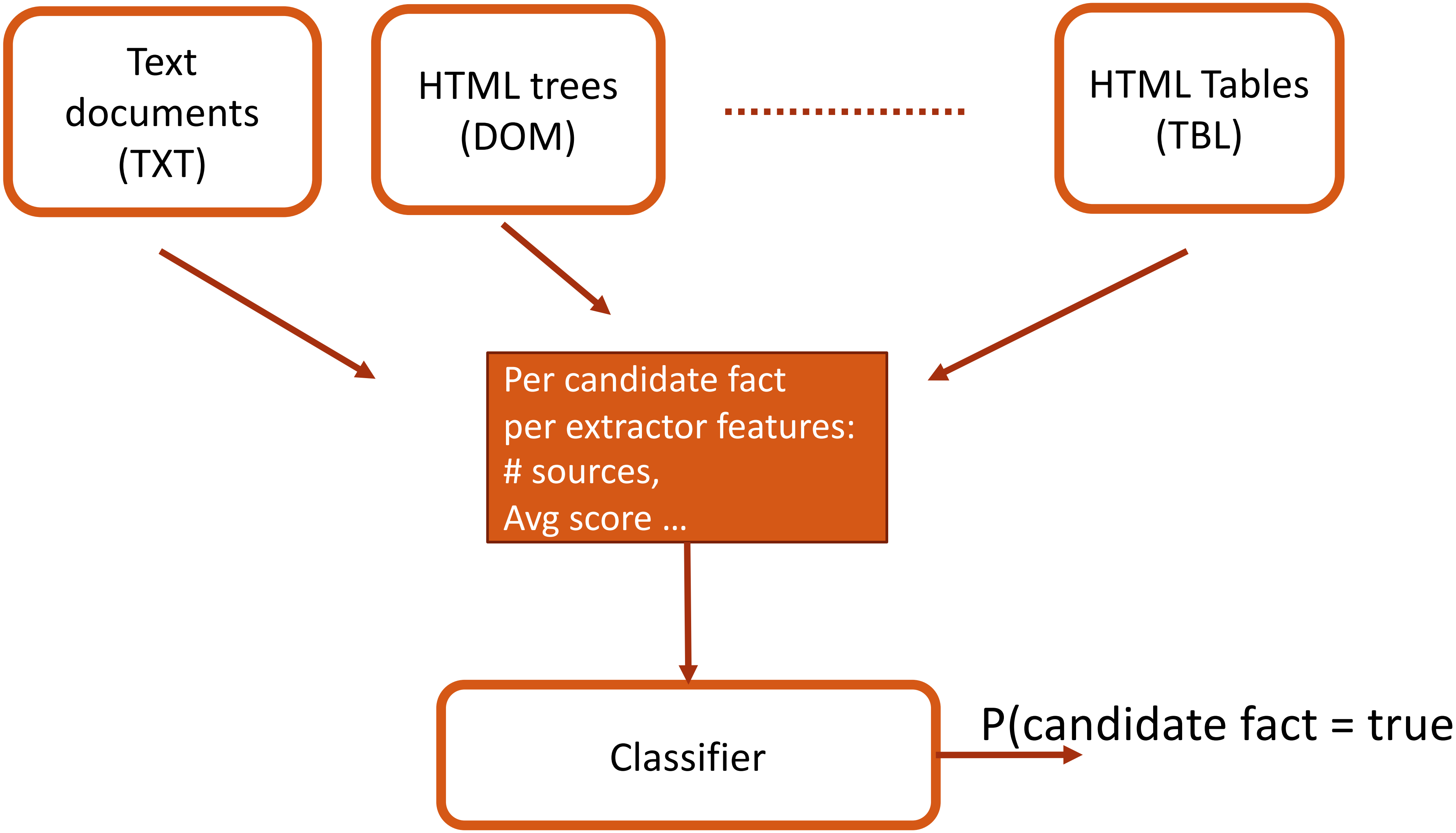
[Multi-View Hierarchical Semi-supervised Learning by Optimal Assignment of Sets of Labels to Instances, Dalvi et al. in preparation, link]

(3) Multi-view learning



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

(4) Classification



Knowledge fusion schemes

- Voting (AND vs OR of extractors)
- Co-training (multiple extraction methods)
- Multi-view learning (multiple data sources)
- Classification

Information Extraction

3 IMPORTANT SUB-PROBLEMS

CATEGORIES OF IE TECHNIQUES

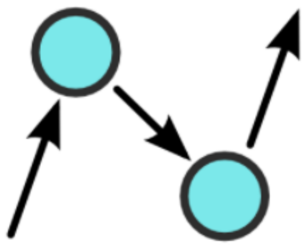
KNOWLEDGE FUSION

IE SYSTEMS IN PRACTICE

IE systems in practice

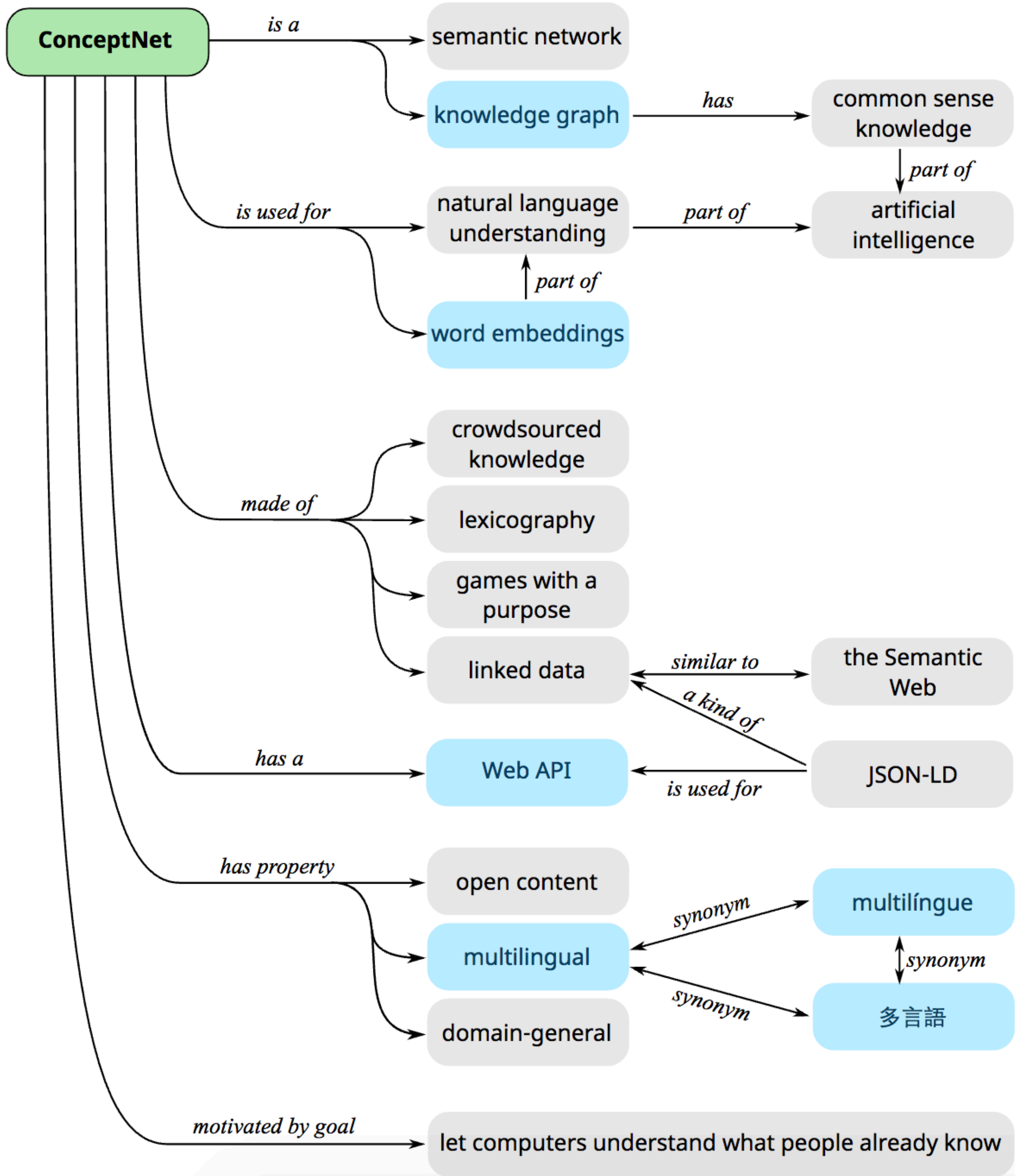
- Conceptnet
- NELL
- Knowledge vault
- Open IE

ConceptNet

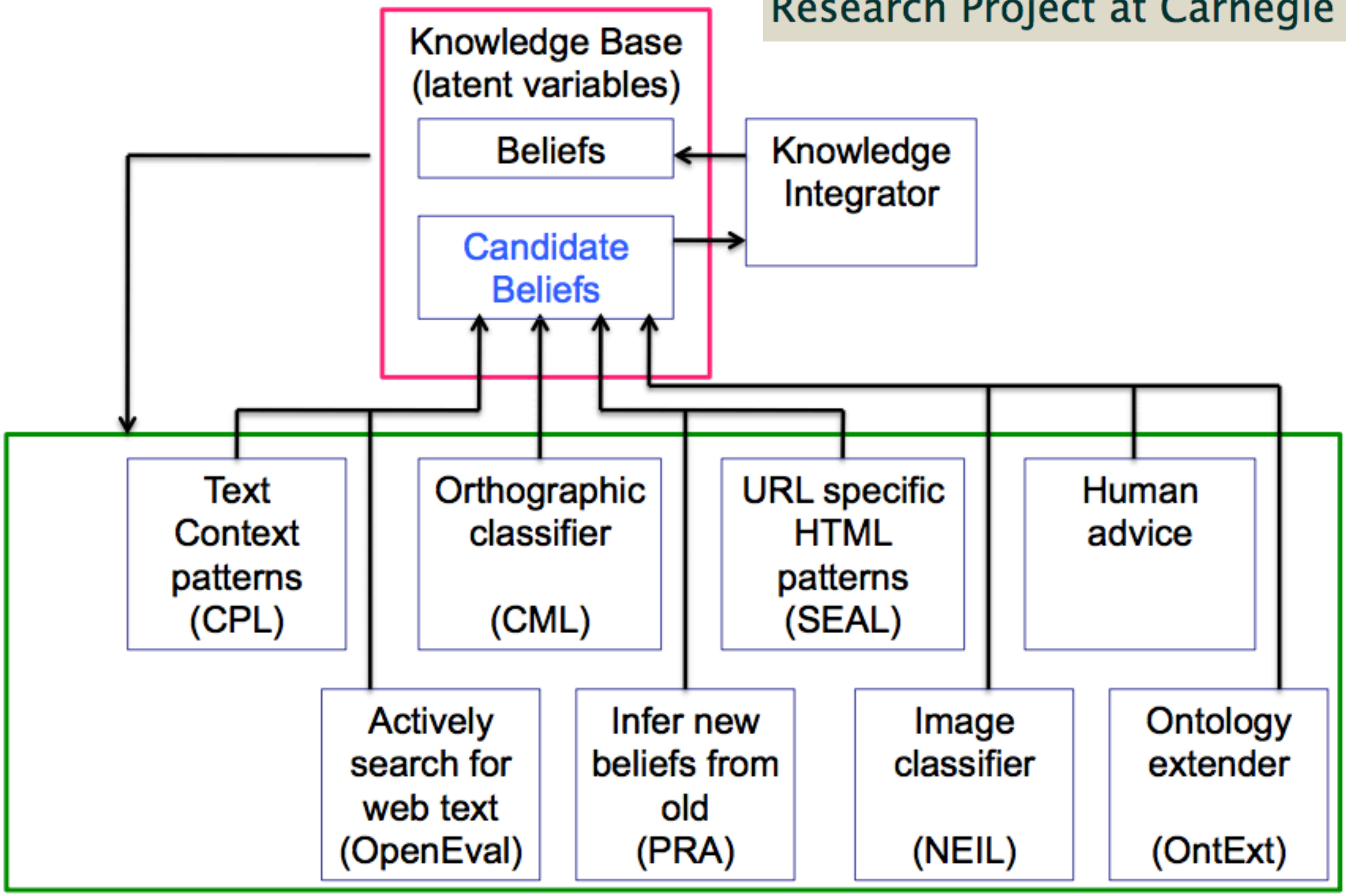


ConceptNet is a freely-available semantic network, designed to help computers understand the meanings of words that people use.

This knowledge was derived from thousands of human contributors.

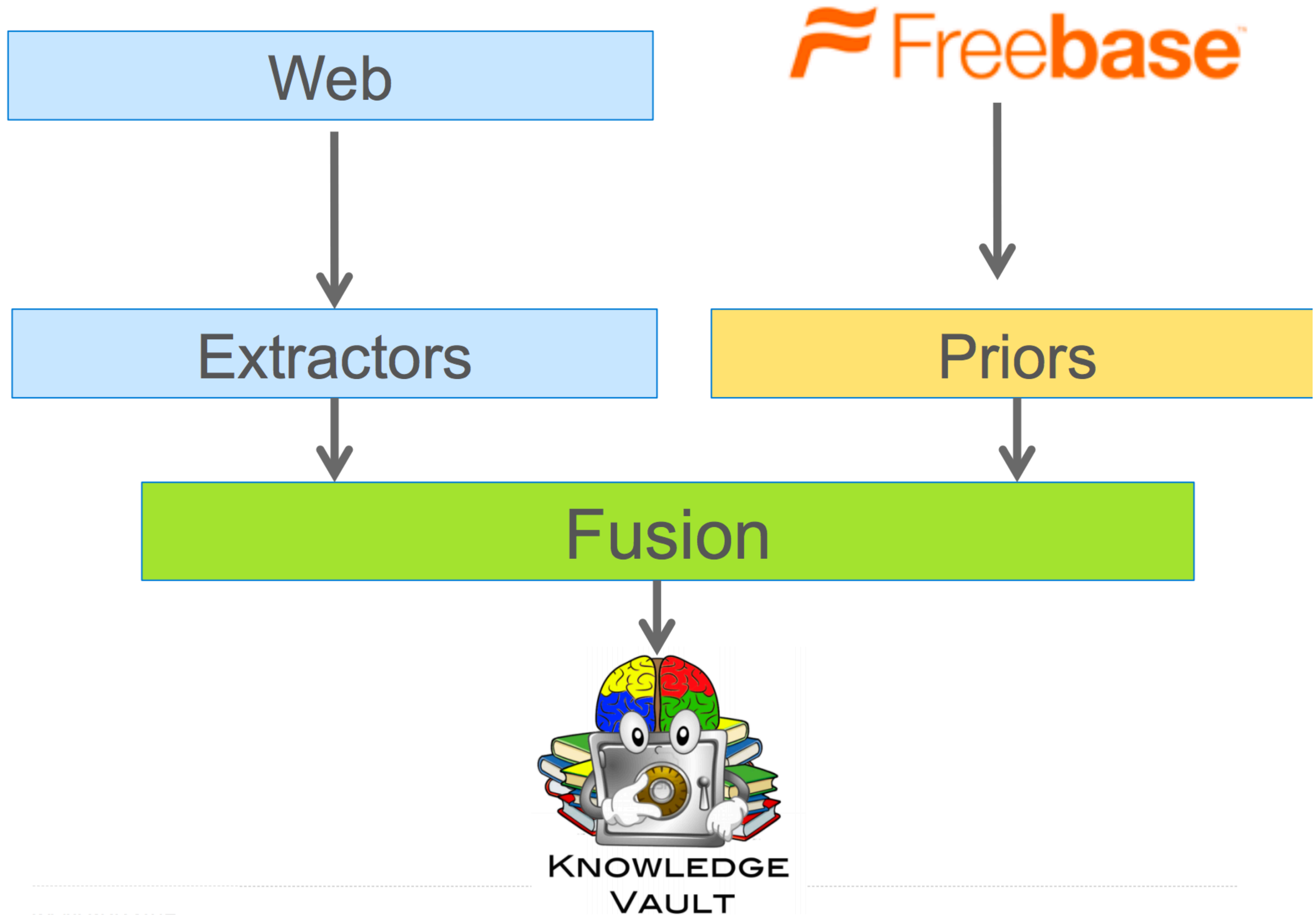


Never Ending Language Learning (NELL)



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

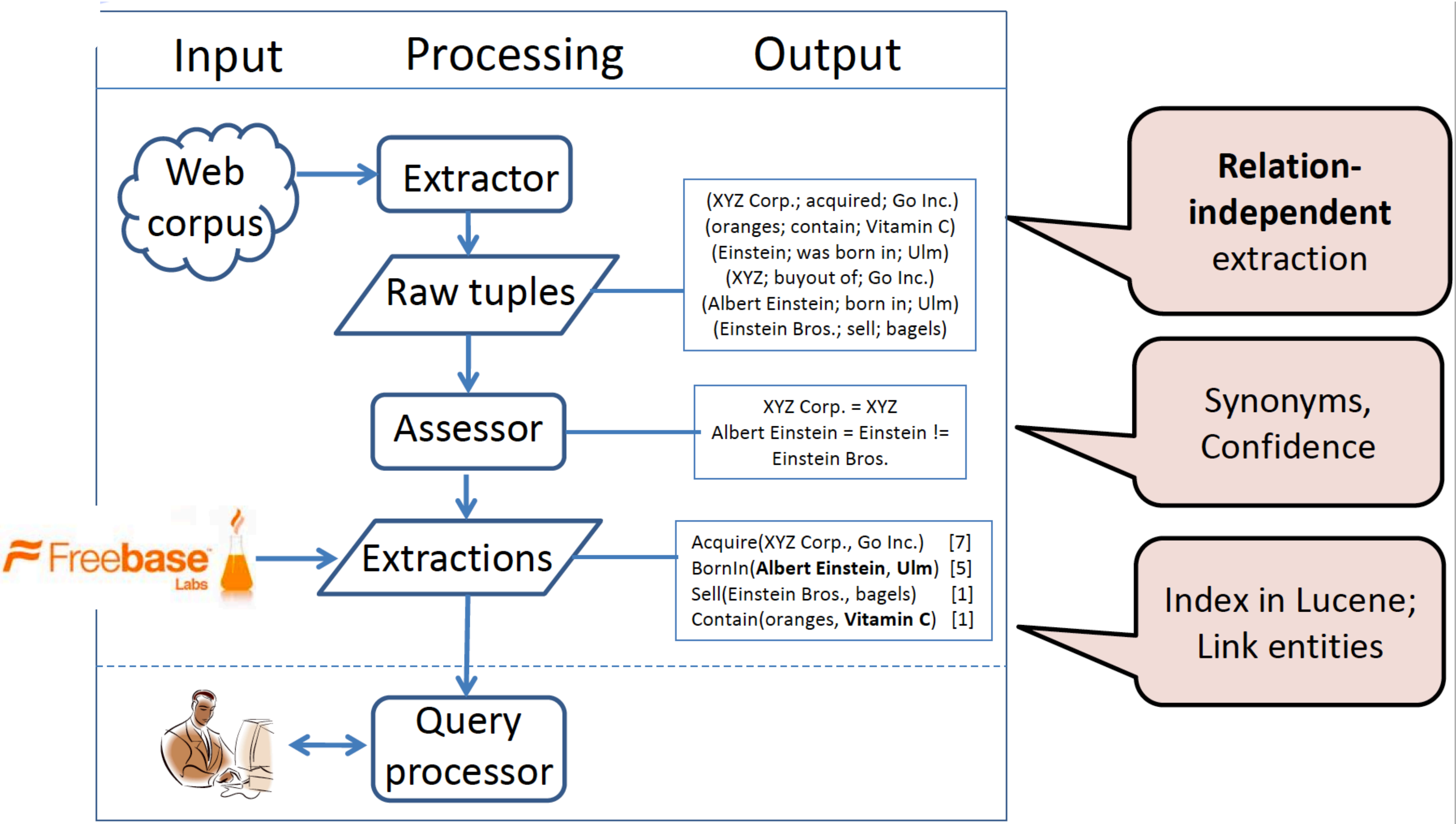
Knowledge Vault



[Architecture diagram taken from Kevin Murphy's slides]

Open IE (KnowItAll)

Open Information Extraction



[Architecture diagram taken from Oren Etzioni's slides]

IE systems at a glance

	Defining domain	Learning extractors	Scoring candidate facts	Fusing extractors

IE systems at a glance

	Defining domain	Learning extractors	Scoring candidate facts	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!