Quick Start: Information Extraction

DSCI-558, SPRING 2021

JAY PUJARA

NLP Fundamentals

EXTRACTING STRUCTURES FROM LANGUAGE

What is NLP?

Unstructured
Ambiguous
Lots and lots of it!

Structured
Precise, Actionable
Specific to the task

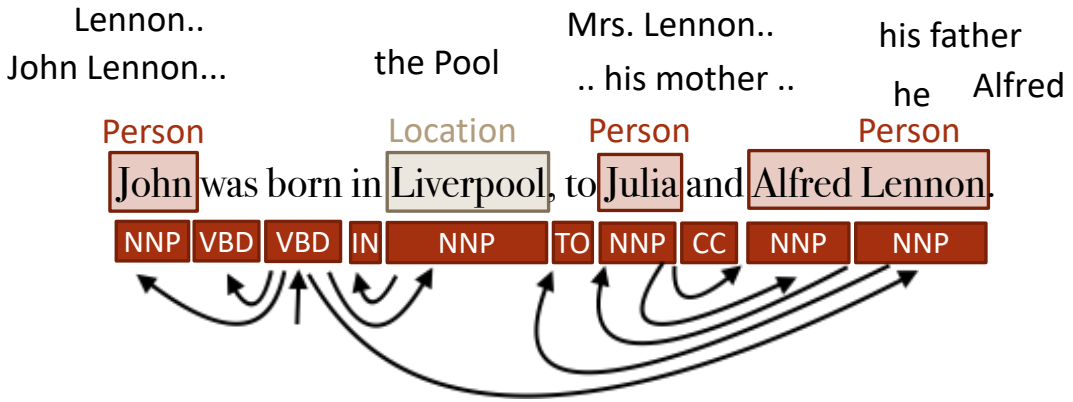
Humans can read them, but
… very slowly
… can’t remember all
… can’t answer questions

Can be used for downstream applications, such as creating Knowledge Graphs!

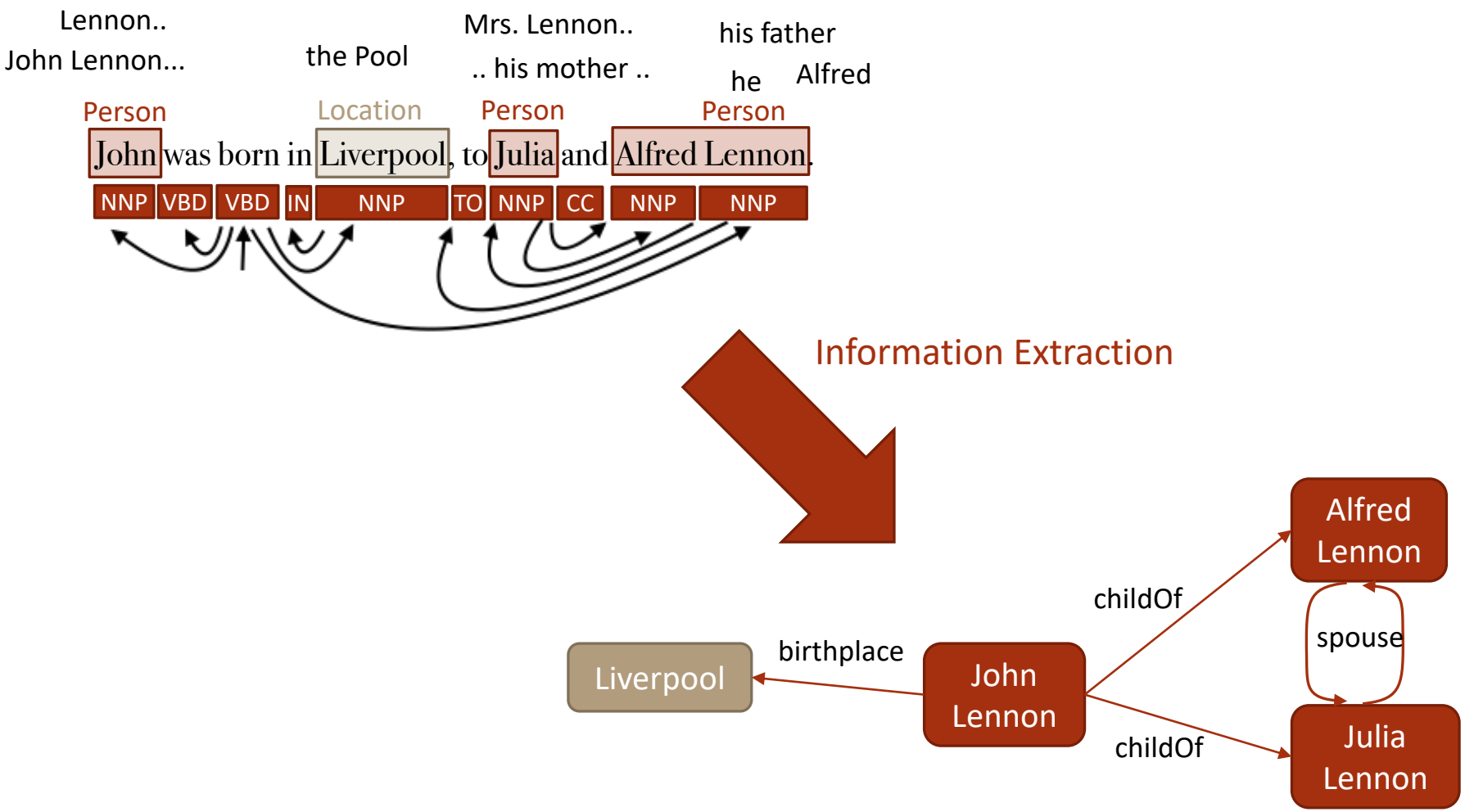
What is NLP?

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

Natural Language Processing



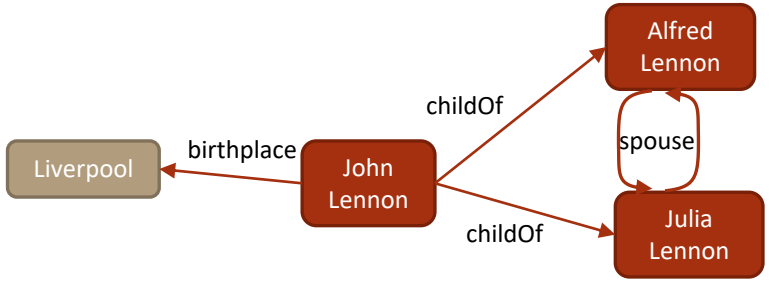
What is Information Extraction?



Breaking it Down

Information Extraction

Entity resolution,
Entity linking,
Relation extraction…



Document

Coreference Resolution…

Lennon..
John Lennon…

the Pool

Mrs. Lennon..
.. his mother ..

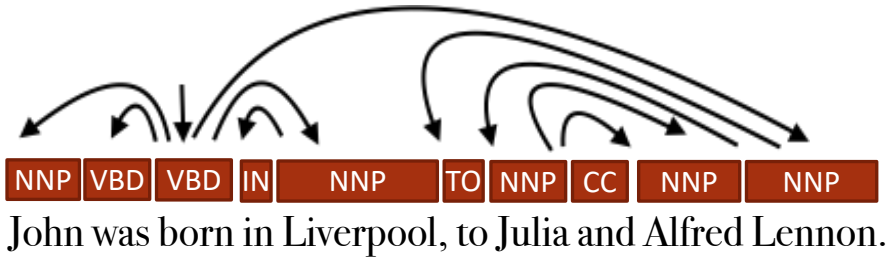
his father
he
Alfred

Person Location Person Person

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

Sentence

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



Tokenization & Sentence Splitting

"Mr. Bob Dobolina is thinkin' of a master plan. Why doesn't he quit?"

[Mr.] [Bob] [Dobolina] [is] [thinkin'] [of] [a] [master] [plan] [.]
[Why] [doesn't] [he] [quit] [?]

How it is done:

- Regular expressions, but not trivial
  - Mr., Yahoo!, lower-case
- For non-English, incredibly difficult!
  - Chinese: no "space" character
- Non-trivial for some domains…
  - What is a "token" in BioNLP?

Uses in KG Construction:

- Strictly constrains other NLP tasks
  - Parts of Speech
  - Dependency Parsing
- Directly effects KG nodes/edges
  - Mention boundaries
  - Relations within sentences

Tagging the Parts of Speech

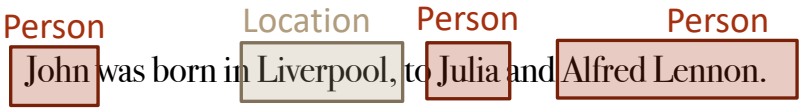
How it is done:

- Context is important!
  - run, table, bar, …
- Label whole sentence together
  - "Structured prediction"
- Conditional Random Fields, ..
- Now: CNNs, LSTMs, …

Uses in KG Construction:

- Entities appear as nouns
- Verbs are very useful
  - For identifying relations
  - For identifying entity types
- Important for downstream NLP
  - NER, Dependency Parsing, …

Detecting Named Entities



How it is done:

- Context is important!
  - Georgia, Washington, …
  - John Deere, Thomas Cook, …
  - Princeton, Amazon, …
- Label whole sentence together
  - Structured prediction again

Uses in KG Construction:

- Mentions describes the nodes
- Types are incredibly important!
  - Often restrict relations
- Fine-grained types are informative!
  - Brooklyn: city
  - Sanders: politician, senator

NER: Entity Types

Stanford CoreNLP

3 class:	Location, Person, Organization
4 class:	Location, Person, Organization, Misc
7 class:	Location, Person, Organization, Money, Percent, Date, Time

spaCy.io

PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FACILITY	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LANGUAGE	Any named language.

From Stanford CoreNLP (http://nlp.stanford.edu/software/CRF-NER.shtml)

NER: Entity Types

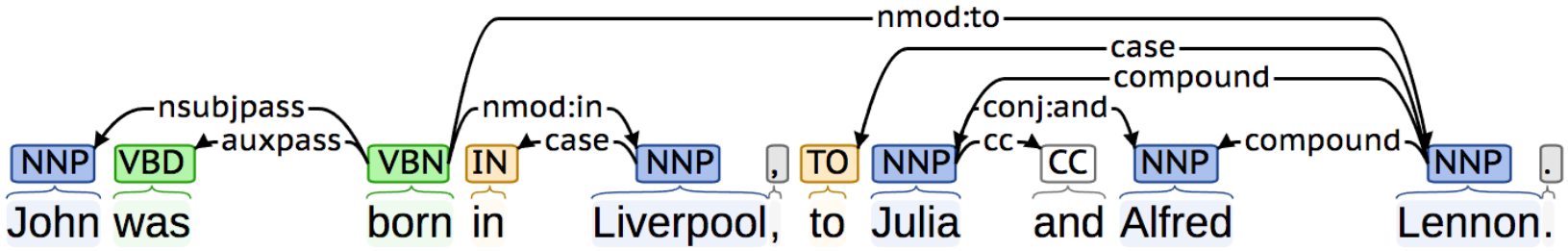
Fine-grained Types

person		doctor	**organization**		terrorist_organization		
actor		engineer	airline		government_agency		
architect		monarch	company		government		
artist		musician	educational_institution		political_party		
athlete		politician	fraternity_sorority		educational_department		
author		religious_leader	sports_league		military		
coach		soldier	sports_team		news_agency		
director		terrorist					
location		body_of_water	**product**		camera	**art**	written_work
city		island	engine		mobile_phone	film	newspaper
country		mountain	airplane		computer	play	music
county		glacier	car		software		
province		astral_body	ship		game	**event**	military_conflict
railway		cemetery	spacecraft		instrument	attack	natural_disaster
road		park	train		weapon	election	sports_event
bridge						protest	terrorist_attack
building	time		chemical_thing		website		
airport	color		biological_thing		broadcast_network		
dam	award		medical_treatment		broadcast_program		
hospital	educational_degree		disease		tv_channel		
hotel	title		symptom		currency		
library	law		drug		stock_exchange		
power_station	ethnicity		body_part		algorithm		
restaurant	language		living_thing		programming_language		
sports_facility	religion		animal		transit_system		
theater	god		food		transit_line		

- More on this later…

From Ling & Weld. AAAI 2012 (http://aiweb.cs.washington.edu/ai/pubs/ling-aaai12.pdf)

Dependency Parsing



How it is done:

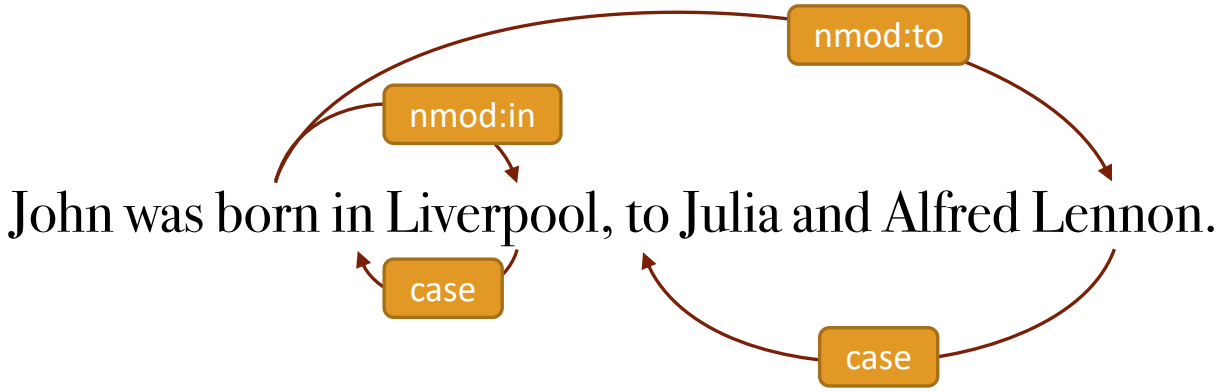
- Model: score trees using features
  - Lexical: words, POS, …
  - Structure: distance, …
- Prediction: Search over trees
  - greedy, spanning tree, belief propagation, dynamic prog, …

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

Using http://nlp.stanford.edu:8080/corenlp/process

Dependency Paths



	Text Patterns	Dependency Paths
John, Liverpool	“was born in”	“was born in”
John, Julia	“was born in Liverpool, to”	“was born to”
John, Alfred Lennon	“was born in Liverpool, to Julia and”	“was born to”

Within-document Coreference

He…

Lennon..

John Lennon...

the Pool

Mrs. Lennon..

.. his mother ..

Alfred

his father

he

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

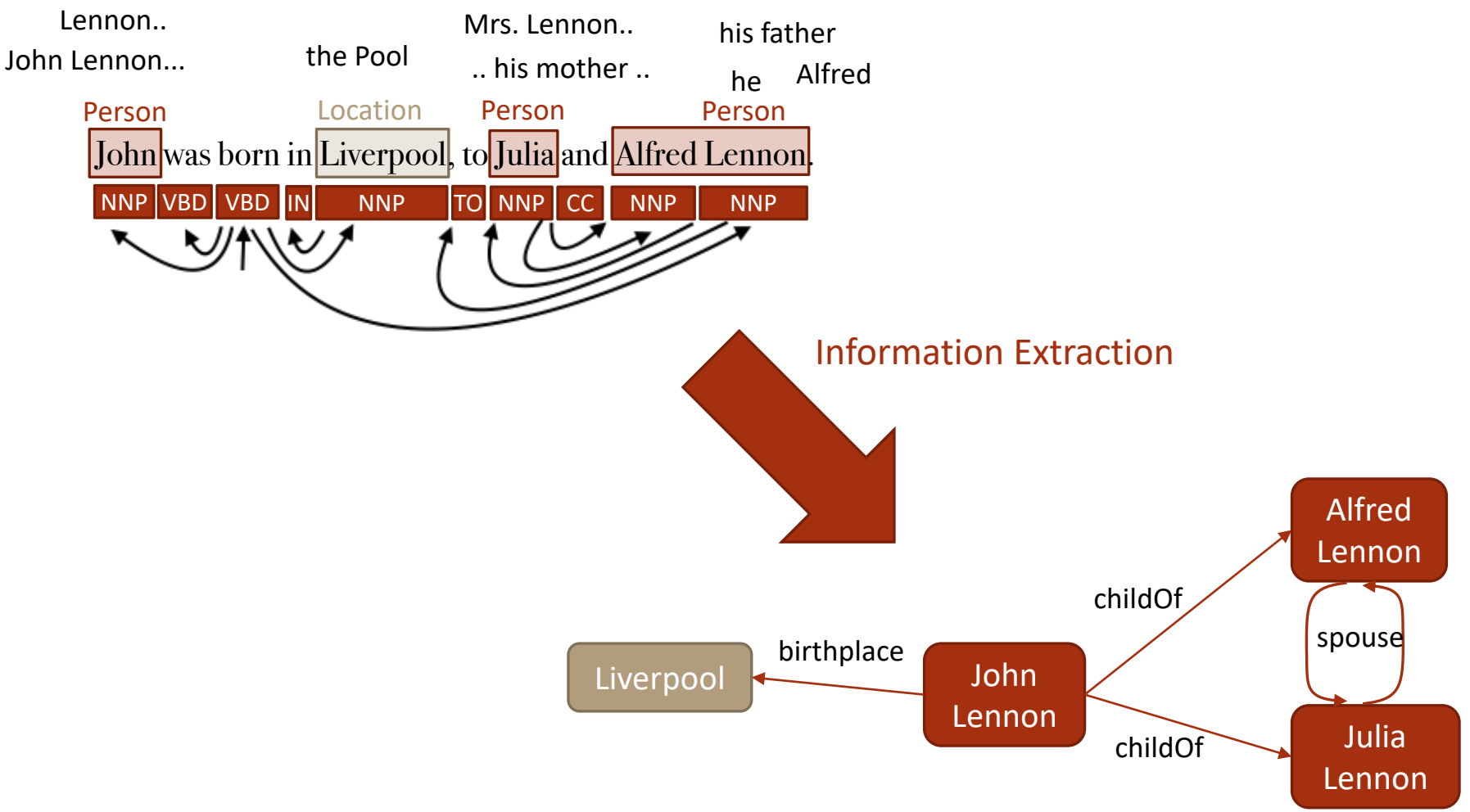
How it is done:

- Model: score pairwise links
  - dep path, similarity, types, …
  - "representative mention"
- Prediction: Search over clusterings
  - greedy (left to right), ILP, belief propagation, MCMC, …

Uses in KG Construction:

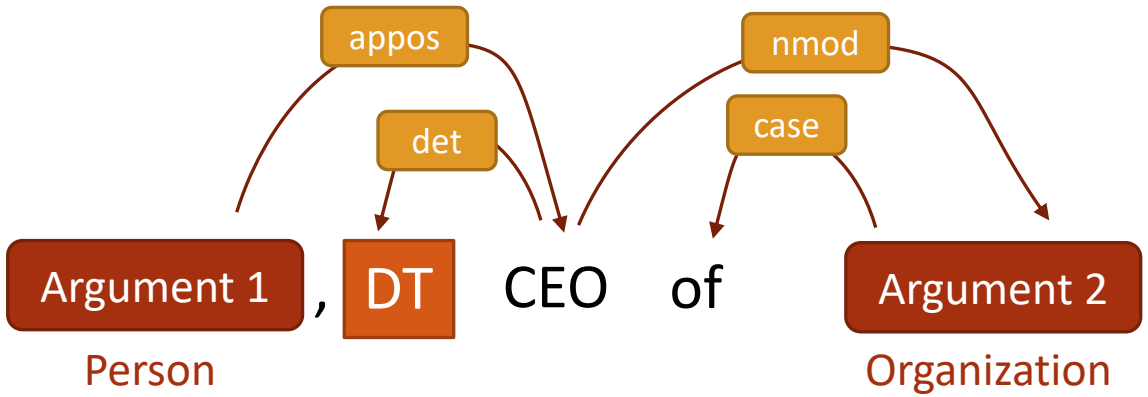
- More context for each entity!
- Many relations occur on pronouns
  - "He is married to her"
- Coref can be used for types
  - Nominals: The president, …
- Difficult, so often ignored

Information Extraction



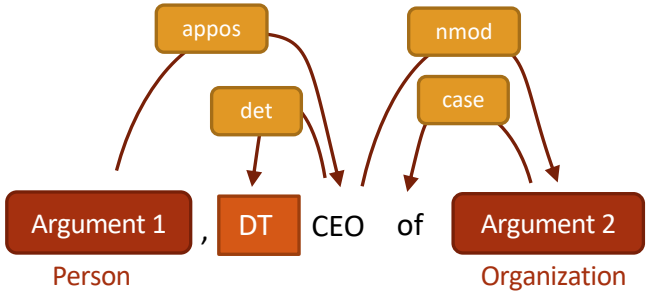
Surface Patterns

Combine tokens, dependency paths, and entity types to define rules.



Bill Gates, the CEO of Microsoft, said …
Mr. Jobs, the brilliant and charming CEO of Apple Inc., said …
… announced by Steve Jobs, the CEO of Apple.
… announced by Bill Gates, the director and CEO of Microsoft.
… mused Bill, a former CEO of Microsoft.
and many other possible instantiations…

Rule-Based Extraction



Use a collection of rules as the system itself

Variations

Source:
- Manually specified
- Learned from Data

Multiple Rules:
- Attach priorities/precedence
- Attach probabilities (more later)

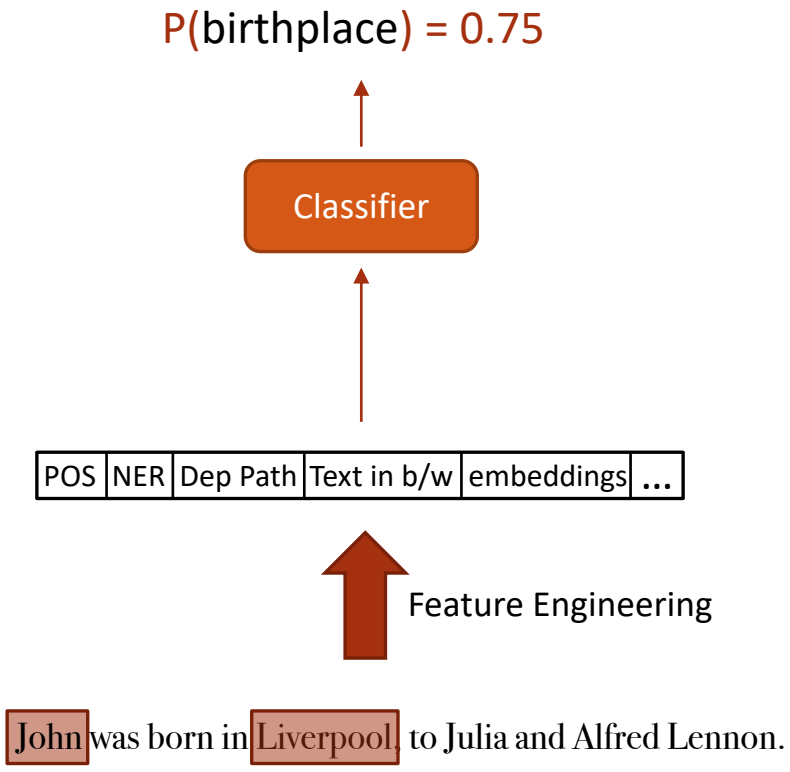
High precision: when it fires, it's correct
Easy to explain predictions
Easy to fix mistakes

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

Supervised Extraction

Machine Learning: hopefully, generalizes the labels in the *right way*

Use all of NLP as features: words, POS, NER, dependencies, embeddings



However

Usually, a lot of labeled data is needed, which is expensive & time consuming. Requires a lot of feature engineering!

Entity Resolution & Linking

…during the late 60′s and early 70′s, **Kevin Smith** worked with several local…

…the term hip-hop is attributed to **Lovebug Starski**. What does it actually mean…

Like Back in 2008, the Lions drafted **Kevin Smith**, even though Smith was badly…

… backfield in the wake of **Kevin Smith**′s knee injury, and the addition of Haynesworth…

The filmmaker **Kevin Smith** returns to the role of Silent Bob…

Nothing could be more irrelevant to **Kevin Smith**′s audacious ′′Dogma′′ than ticking off…

… The Physiological Basis of Politics,” by **Kevin Smith**, Douglas Oxley, Matthew Hibbing…

Entity Names: Two Main Problems

Entities with Same Name

Same type of entities share names

Kevin Smith, John Smith, Springfield, …

Things named after each other

Clinton, Washington, Paris, Amazon, Princeton, Kingston, …

Partial Reference

First names of people, Location instead of team name, Nick names

Different Names for Entities

Nick Names

Bam Bam, Drumpf, …

Typos/Misspellings

Baarak, Barak, Barrack, …

Inconsistent References

MSFT, APPL, GOOG…

Entity Linking Approach

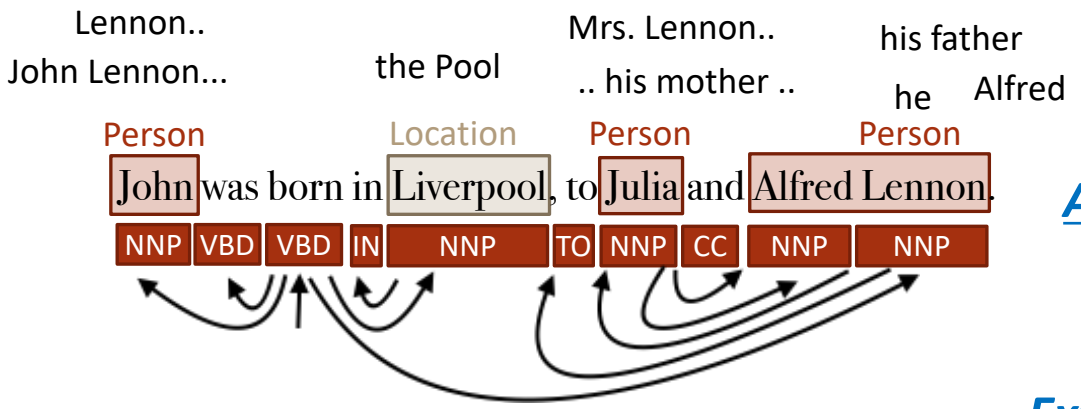
Washington drops 10 points after game with UCLA Bruins.

Candidate Generation		Washington DC, George Washington, Washington state, Lake Washington, Washington Huskies, Denzel Washington, University of Washington, Washington High School, …
Entity Types	LOC/ORG	Washington DC, ~~George Washington~~, Washington state, Lake Washington, Washington Huskies, ~~Denzel Washington~~, University of Washington, Washington High School, …
Coreference	UWashington, Huskies	~~Washington DC~~, ~~George Washington~~, ~~Washington state~~, ~~Lake Washington~~, Washington Huskies, ~~Denzel Washington~~, University of Washington, ~~Washington High School~~, …
Coherence	UCLA Bruins, USC Trojans	~~Washington DC~~, ~~George Washington~~, ~~Washington state~~, ~~Lake Washington~~, Washington Huskies, ~~Denzel Washington~~, ~~University of Washington~~, ~~Washington High School~~, …

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

Text

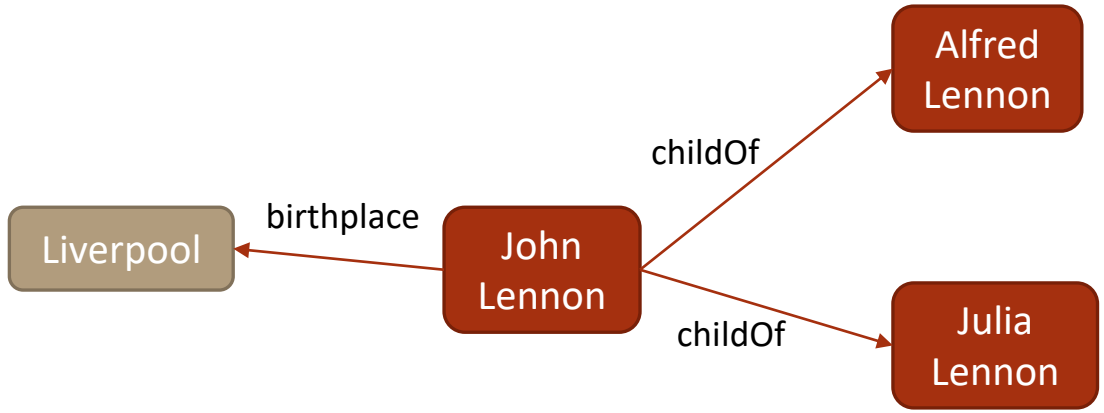
NLP



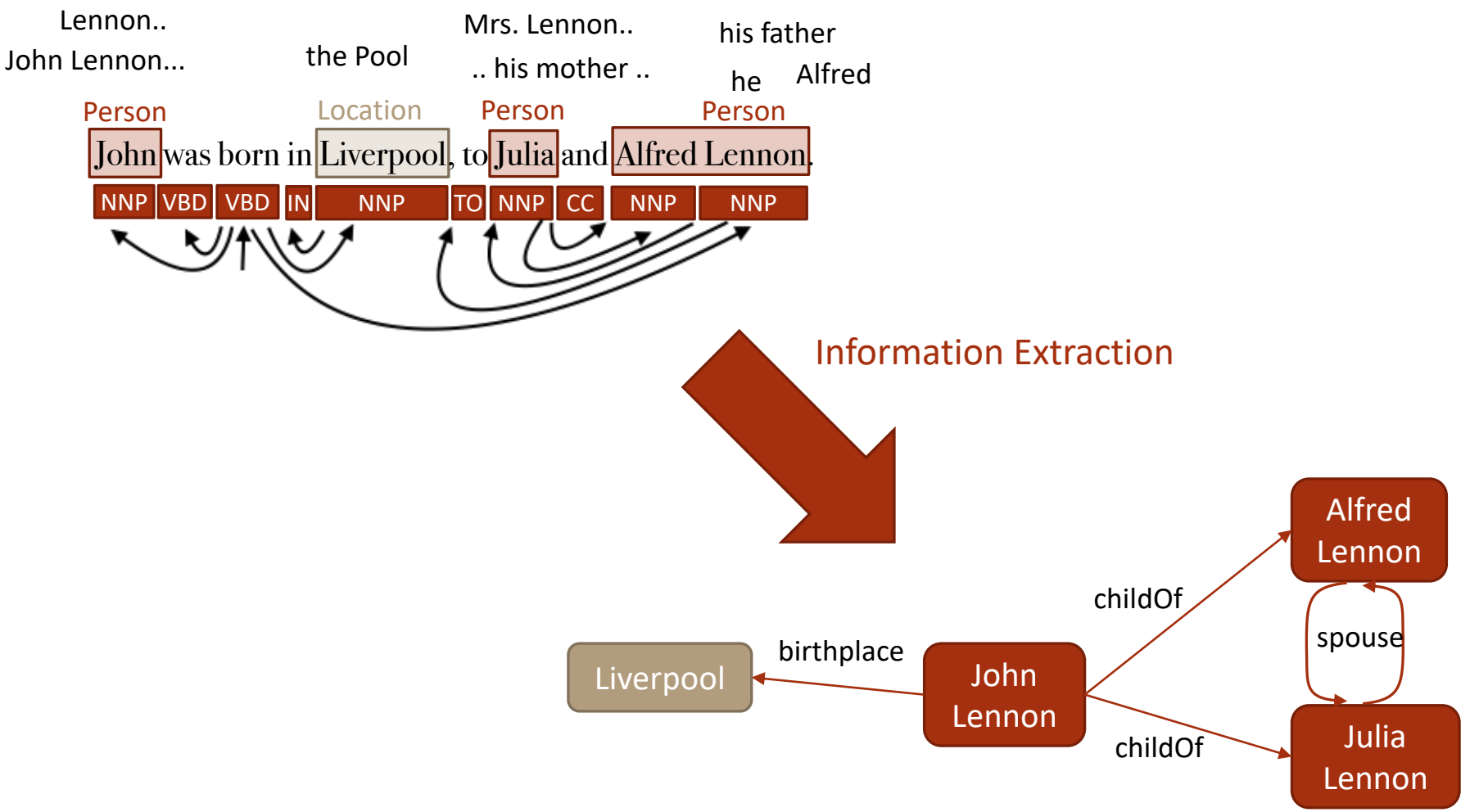
Annotated text

Information Extraction

Extraction graph



Information Extraction



Demo of spaCy

In the next homework, you will be using spaCy to process documents you have crawled and perform information extraction.

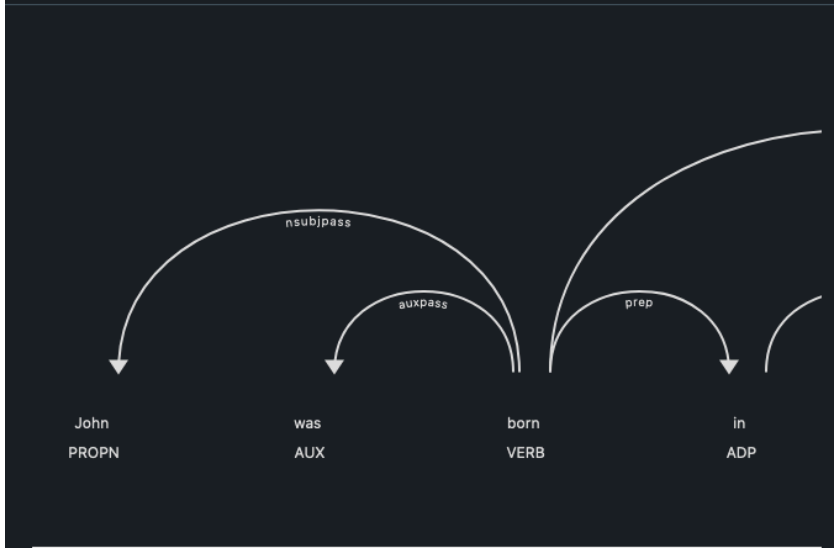
If you're looking for a good time, try https://spacy.io/usage/spacy-101

```
import spacy
from spacy import displacy

# Load English tokenizer, tagger, parser, NER and word vectors
nlp = spacy.load("en_core_web_sm")

# Process whole documents
text = ("John was born in Liverpool, to Julia and Alfred Lennon.")
doc = nlp(text)
displacy.render(doc, style="dep")
# Analyze syntax
print("Noun phrases:", [chunk.text for chunk in doc.noun_chunks])
print("Verbs:", [token.lemma_ for token in doc if token.pos_ == "VERB"])
# Find named entities, phrases and concepts
for entity in doc.ents:
    print(entity.text, entity.label_)
```

RUN




```
Noun phrases: ['John', 'Liverpool', 'Julia', 'Alfred Lennon']
Verbs: ['bear']
John PERSON
Liverpool GPE
Julia GPE
Alfred Lennon PERSON
```

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

×

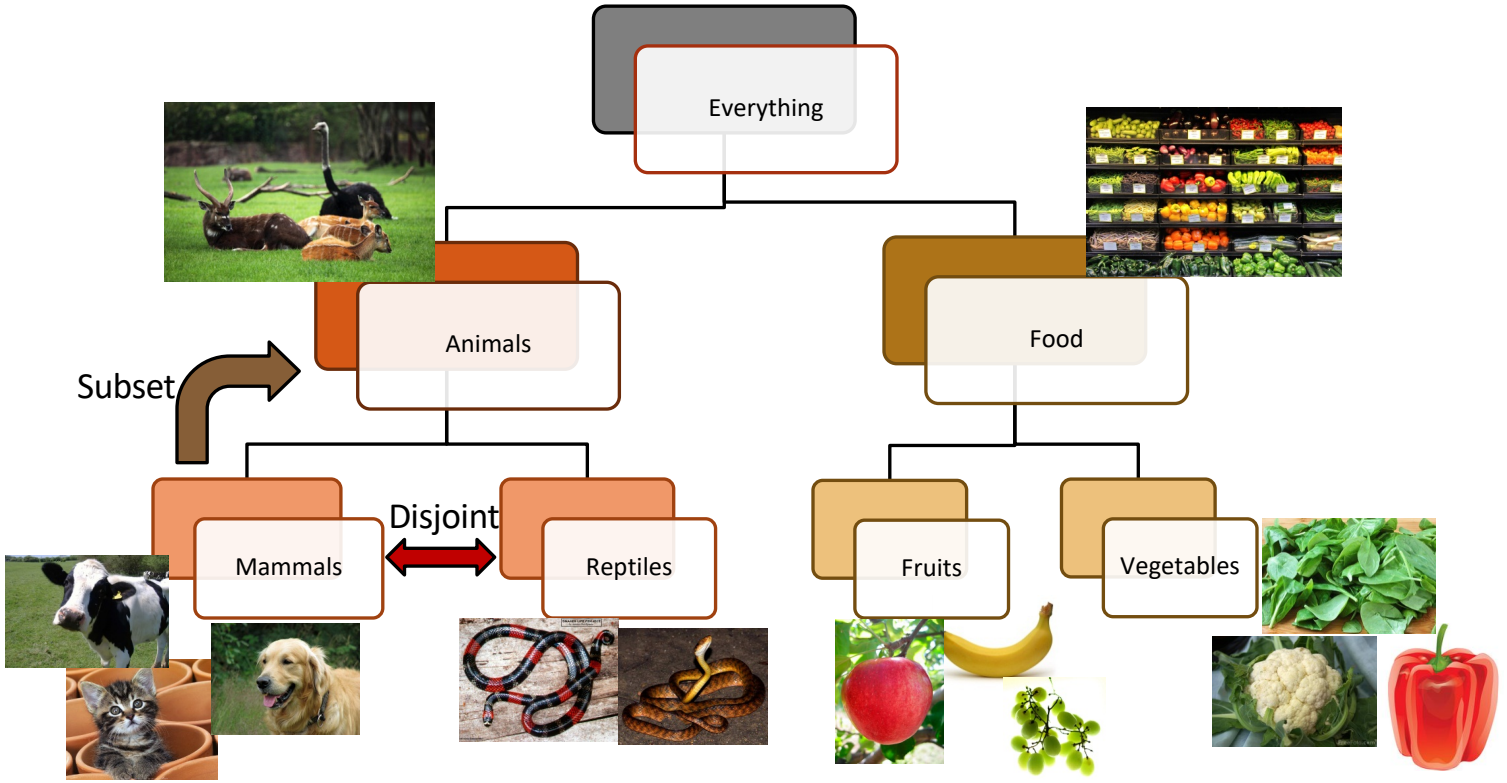
3 LEVELS OF SUPERVISION

Supervised

Semi-supervised

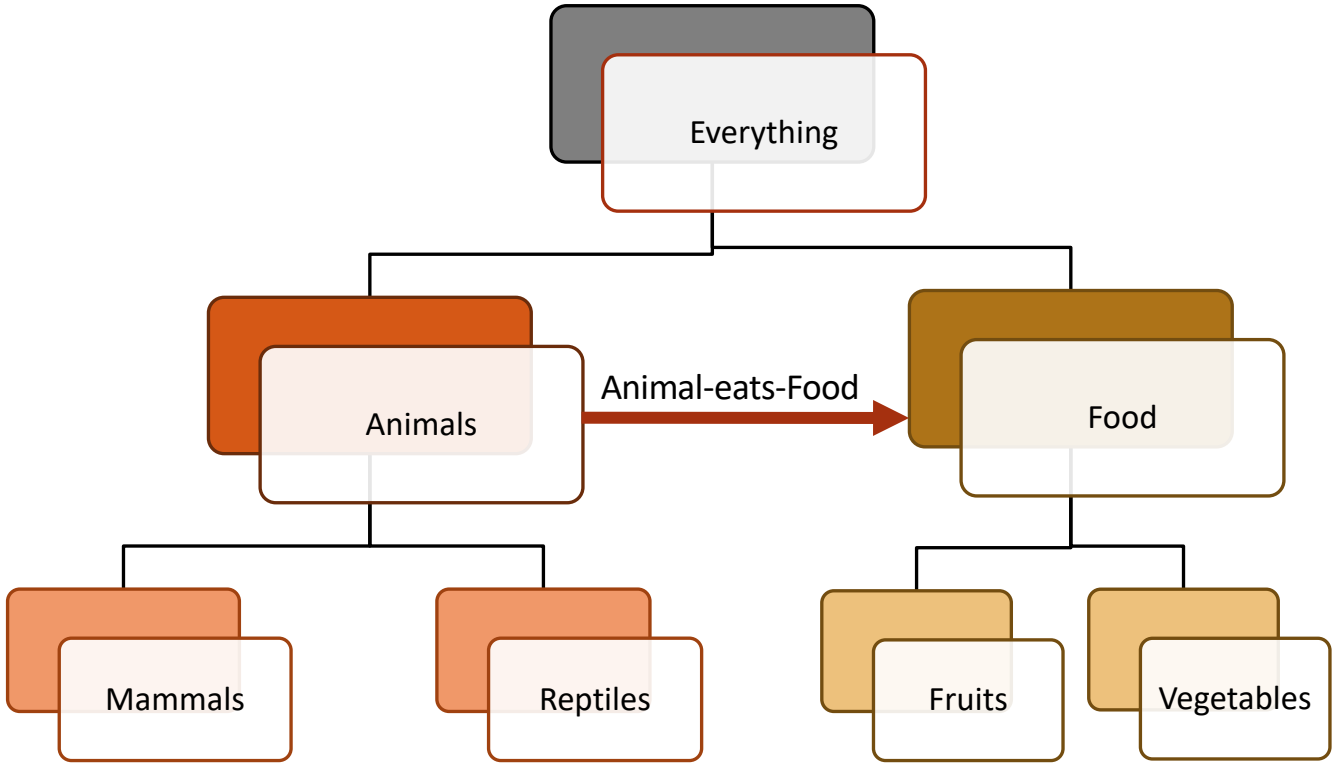
Unsupervised

Defining Domain: Manual



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

Defining Domain: Manual

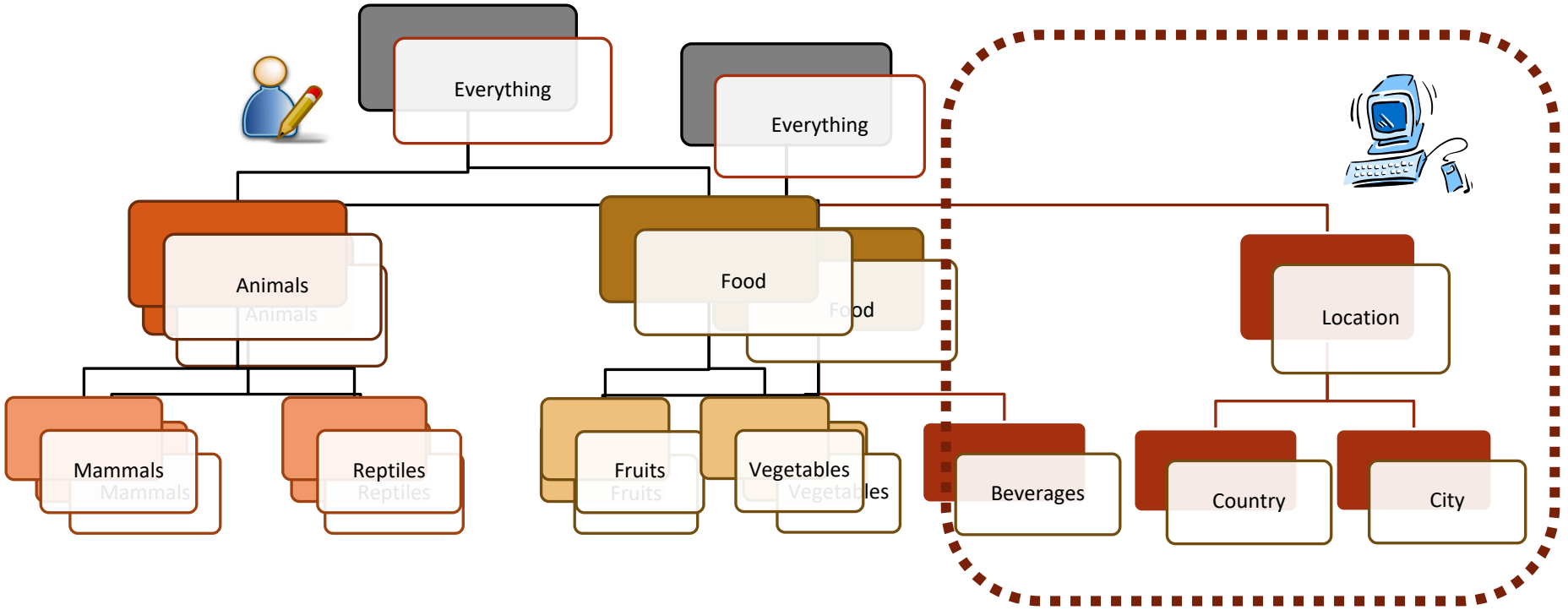


- **Highly semantic ontology**
- **Leads to high precision extractions**
- **Expensive to create**
- **Requires domain experts**

[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



*[**Exploratory Learning**, Dalvi et al., ECML 2013]*
*[**Hierarchical Semi-supervised Classification with Incomplete Class Hierarchies, Dalvi et al., WSDM 2016**]*

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**

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

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

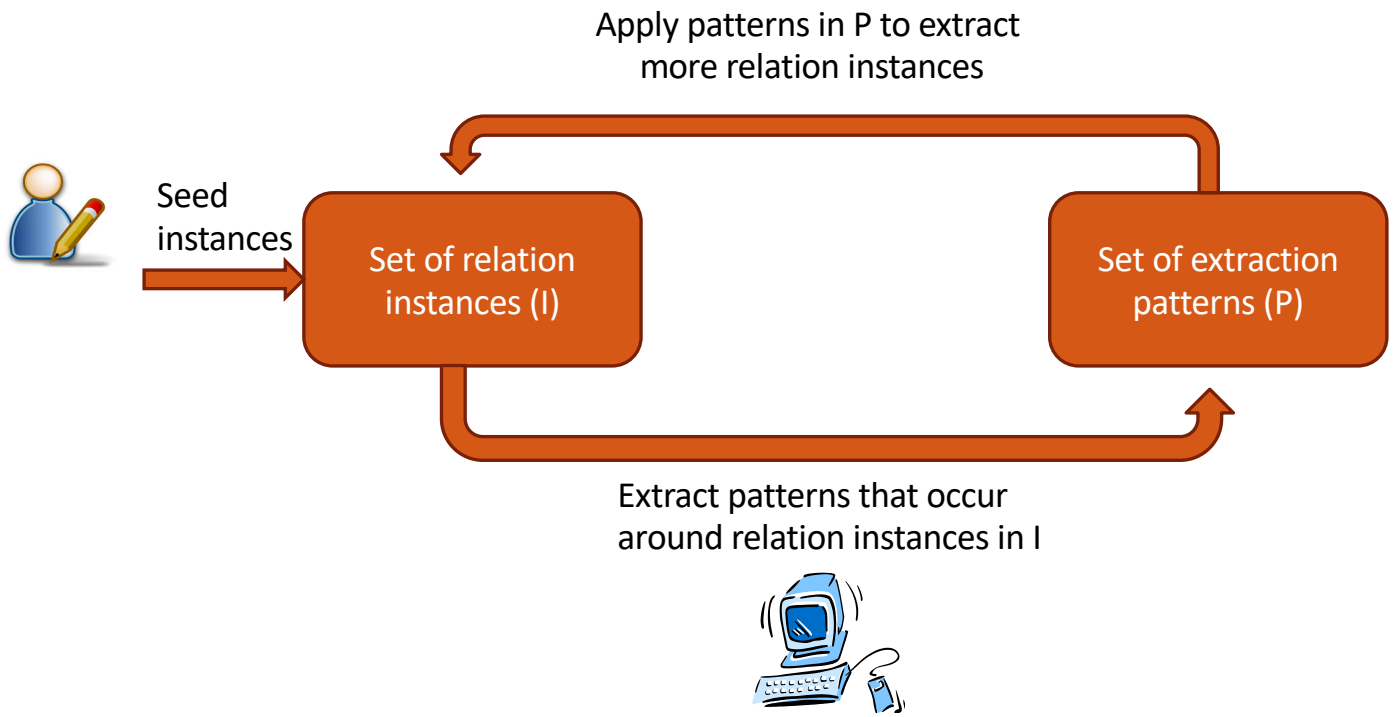
Supervised

Semi-supervised

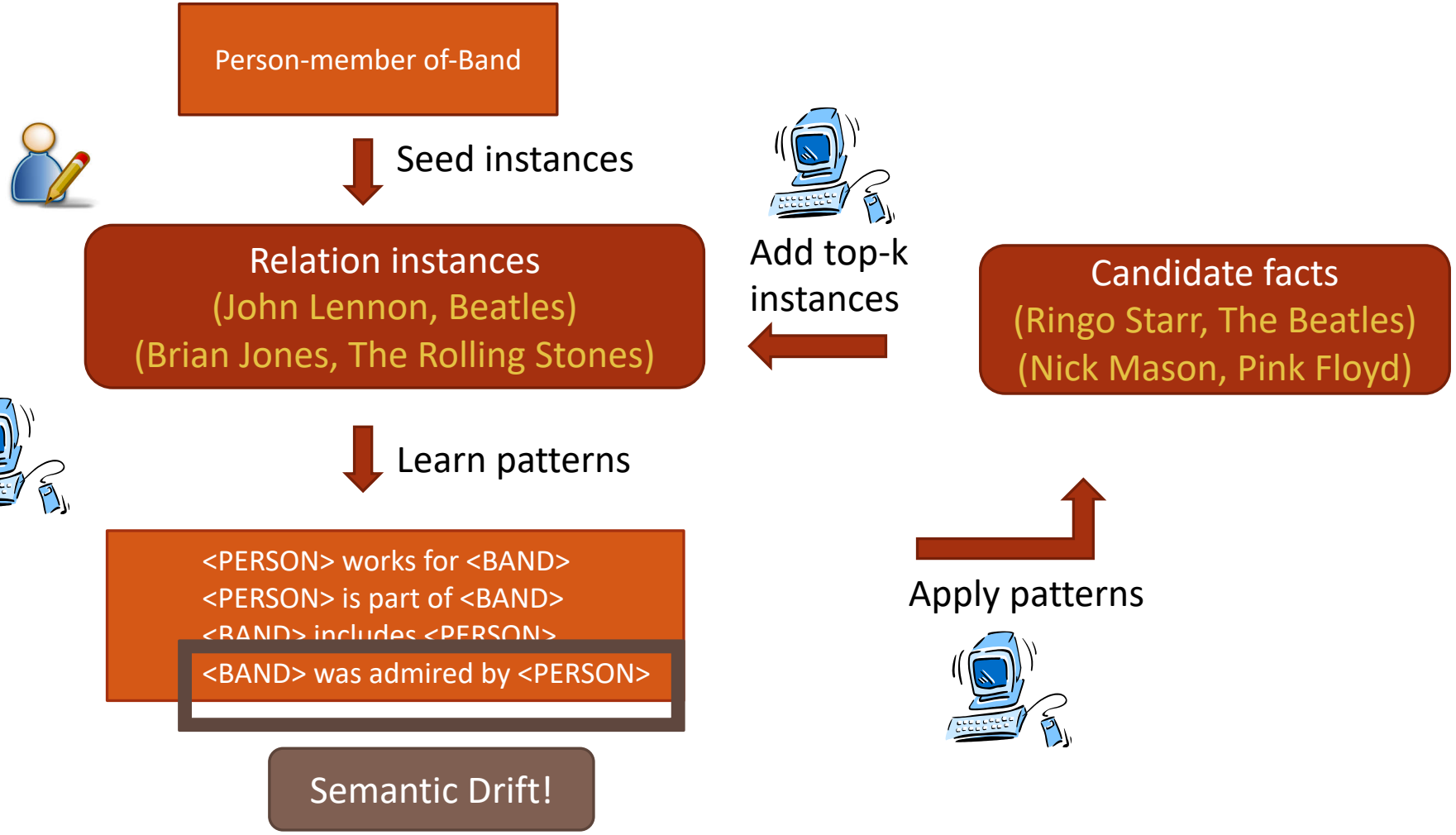
Unsupervised

Learning Extractors: Semi-supervised

Bootstrapping

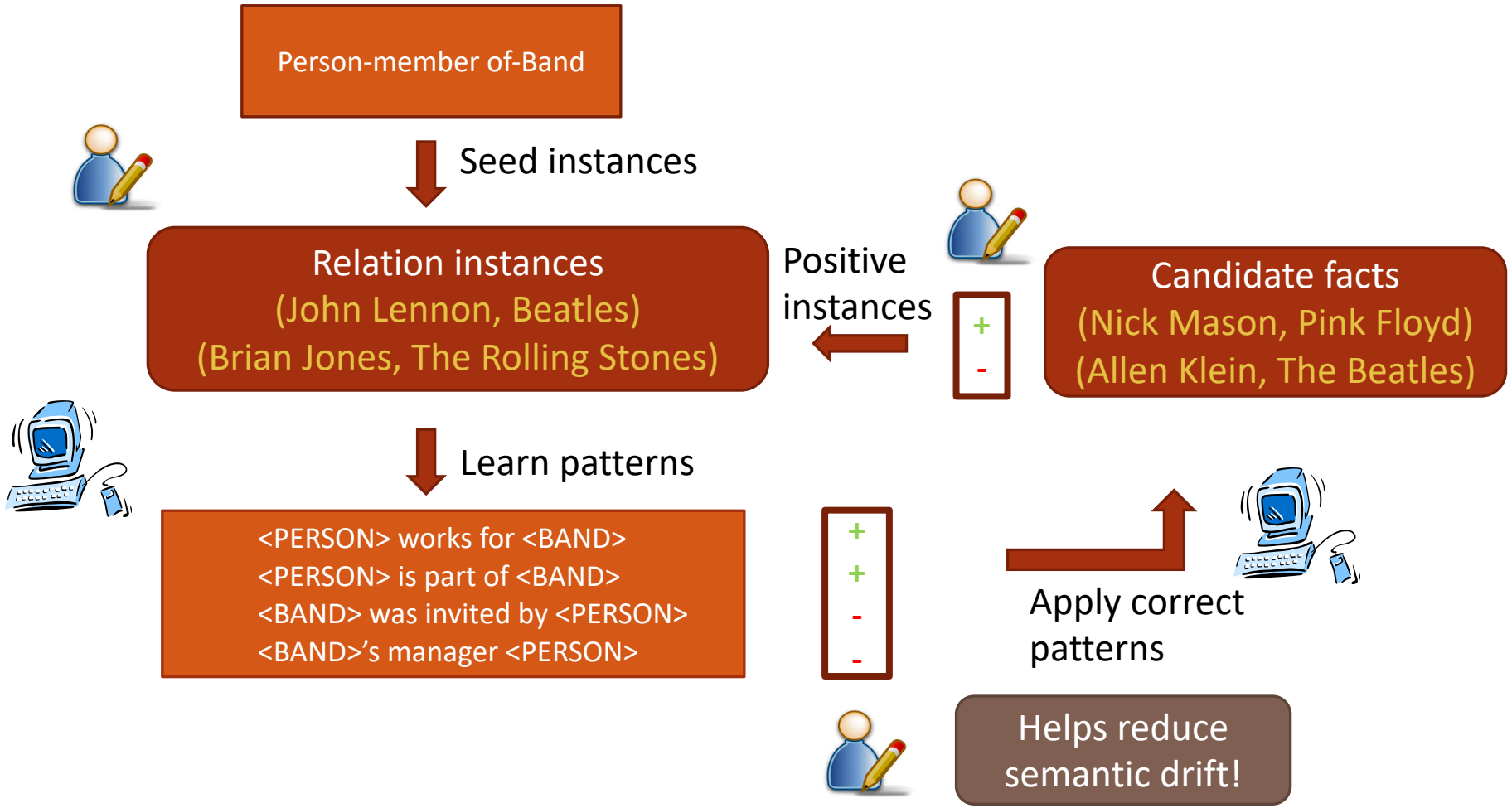


Learning Extractors: Semi-supervised



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

Learning Extractors : Interactive



*[**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)

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

Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring candidate facts

×

3 LEVELS OF SUPERVISION

Supervised

Semi-supervised

Unsupervised

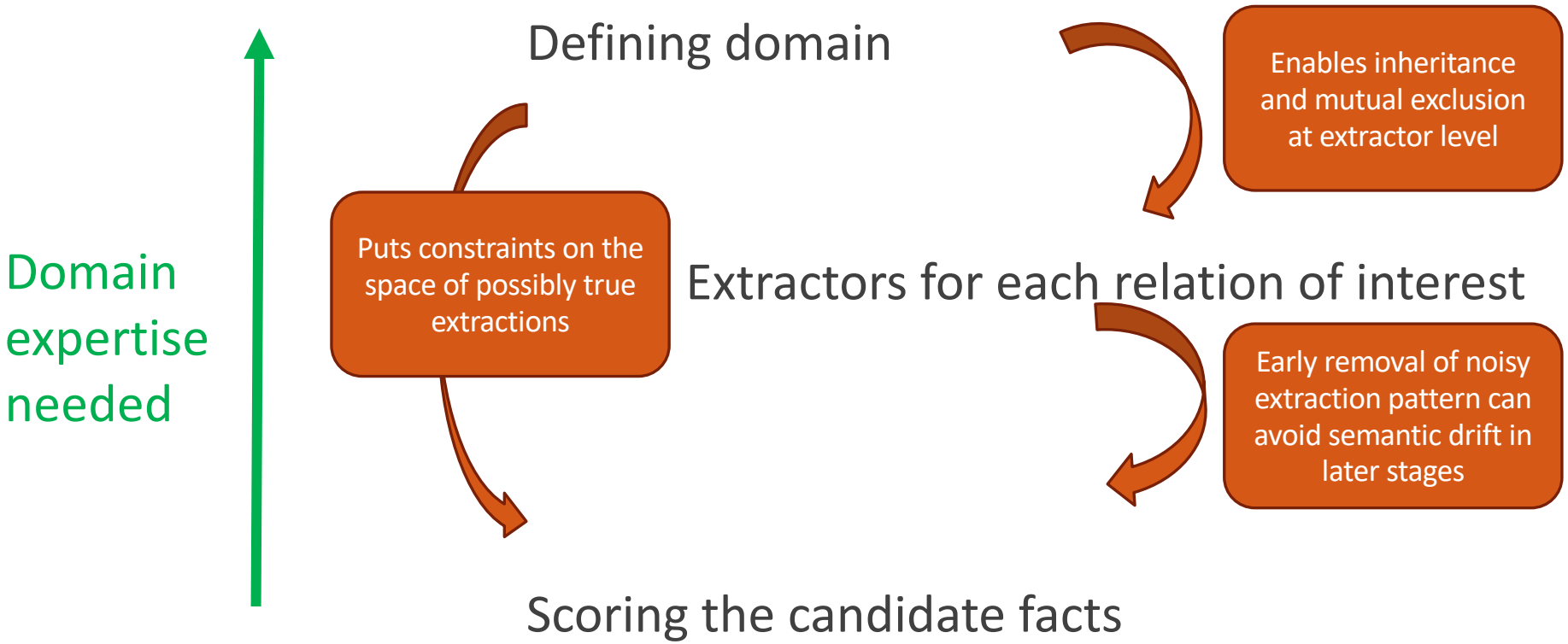
Scoring the candidate facts

- 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) ∝ (#unique instances it could extract)
  Score(candidate fact) ∝ (#distinct extraction patterns that support it)
  {cheap, leads to semantic drift}

Impact of early supervision



Effect of supervision on extractions

Precision,
Human efforts

Recall,
Speed

Information Extraction

3 IMPORTANT SUB-PROBLEMS

CATEGORIES OF IE TECHNIQUES

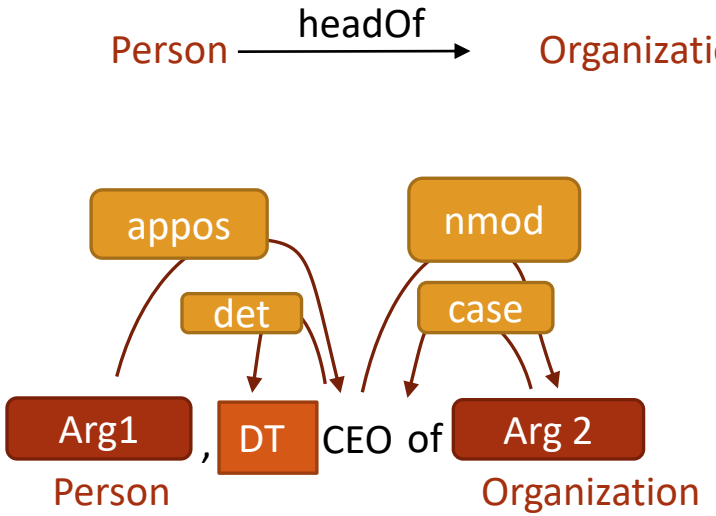
KNOWLEDGE FUSION

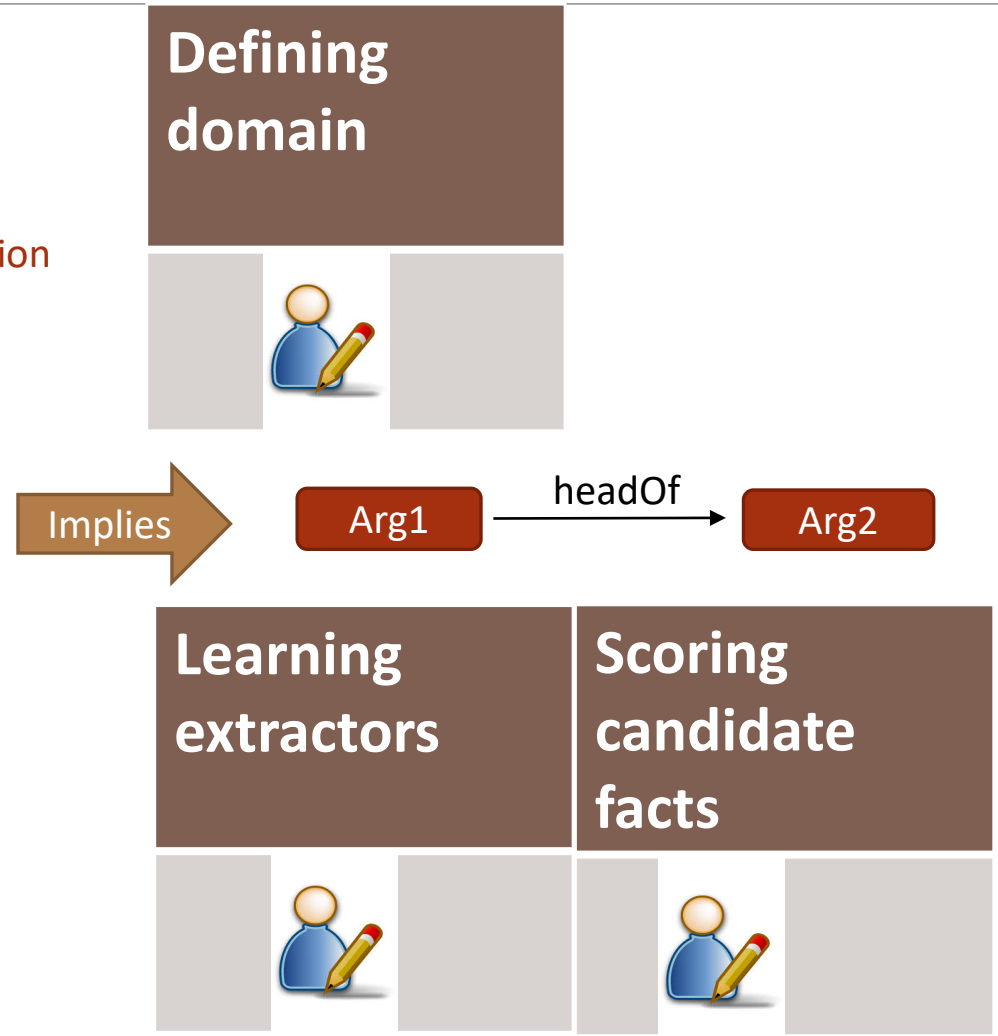
IE SYSTEMS IN PRACTICE

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

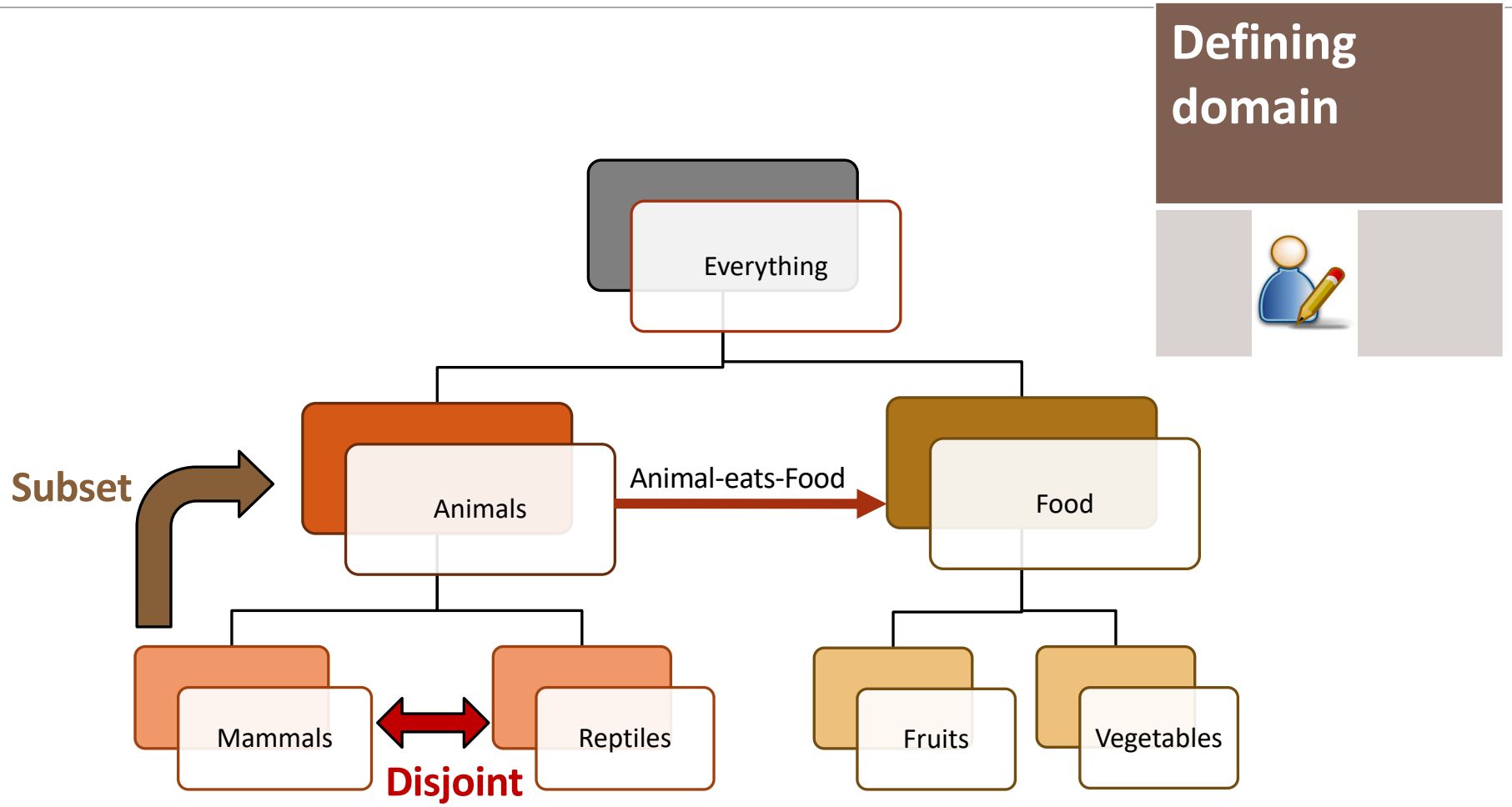




(1) Narrow domain patterns

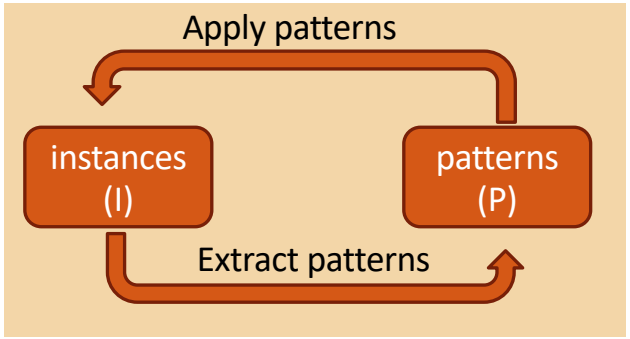
Defining domain	Learning extractors	Scoring candidate facts
		

(2) Ontology based extraction

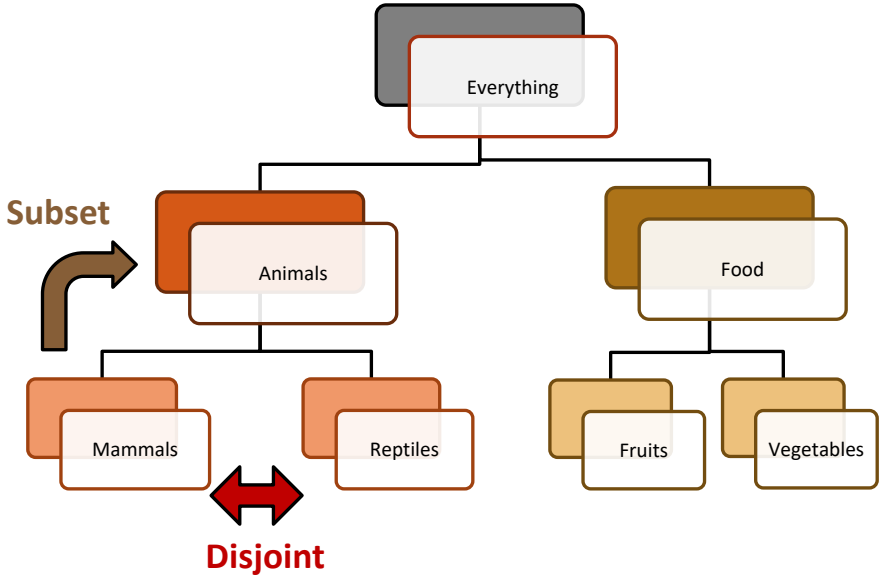


(2) Ontology based extraction

Bootstrapping



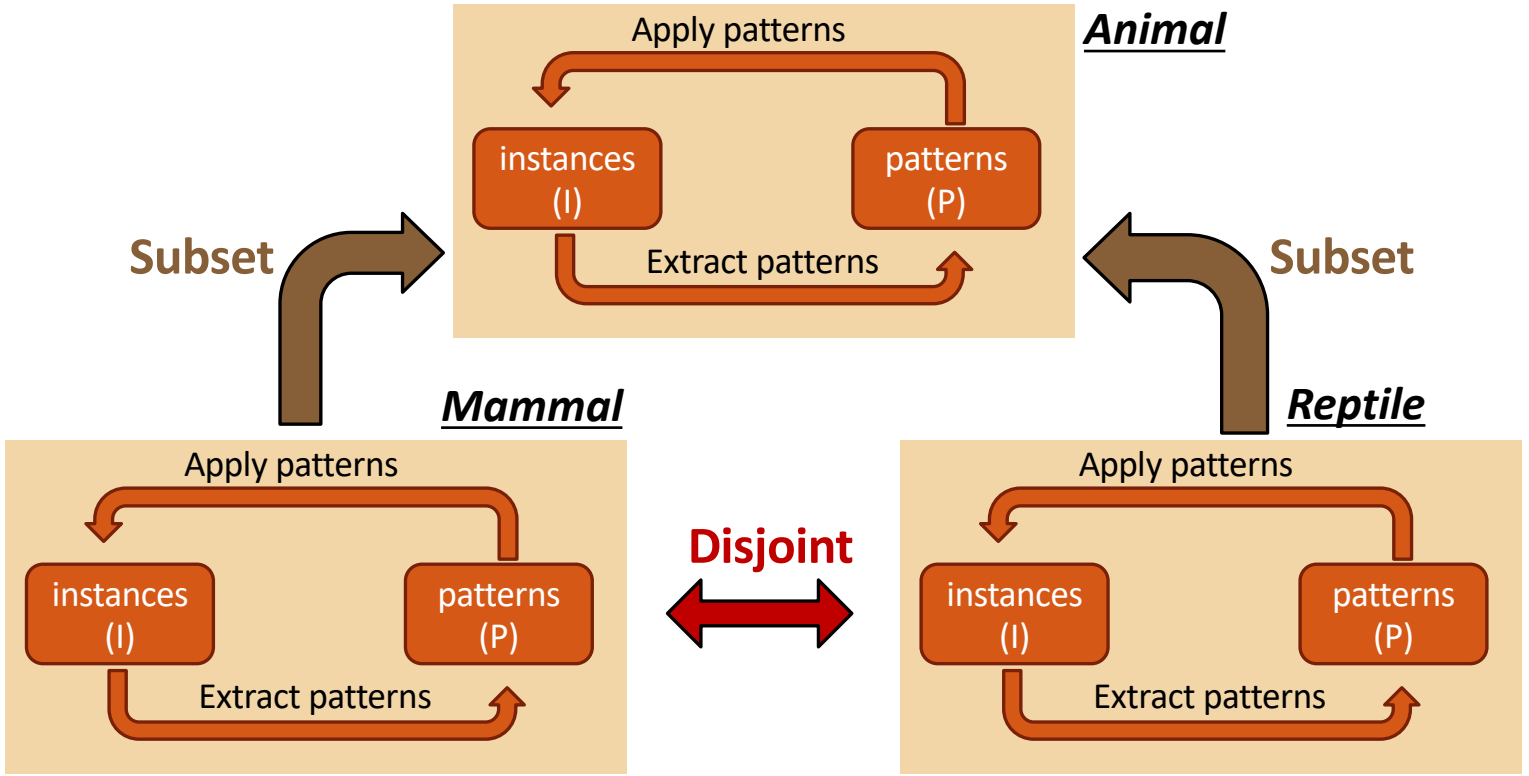
Ontological constraints



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

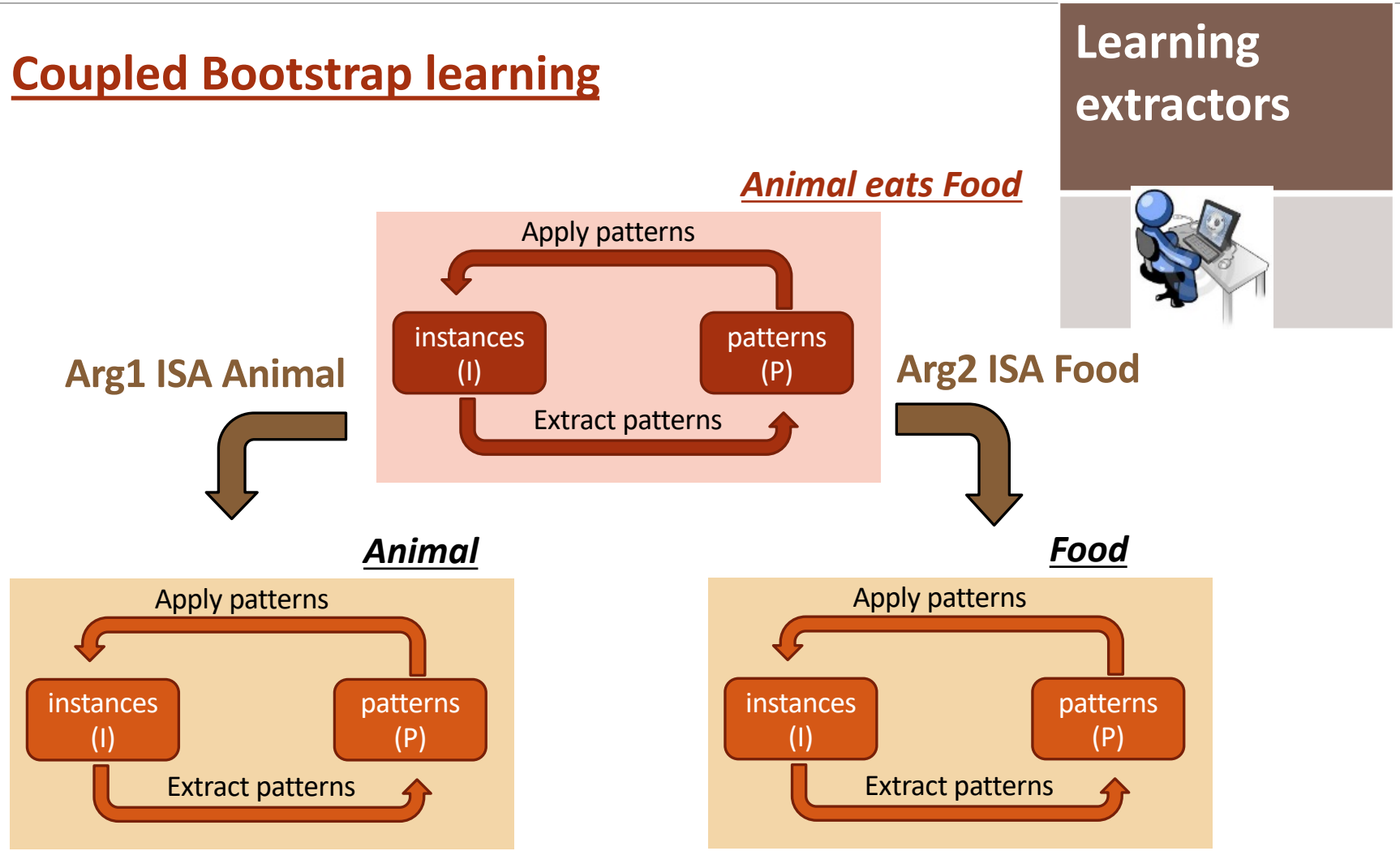
(2) Ontology based extraction

Coupled Bootstrap learning



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

(2) Ontology based extraction



[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)

(2) Ontology based extraction

Defining domain	Learning extractors	Scoring candidate facts
		

(3) Interactive Extrac

Defining domain

Scoring candidate facts

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

Learning extractors

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

Scoring candidate facts

n in, Hampstead) : 0.88

suburb of, London) : 0.9

(4) Open domain IE

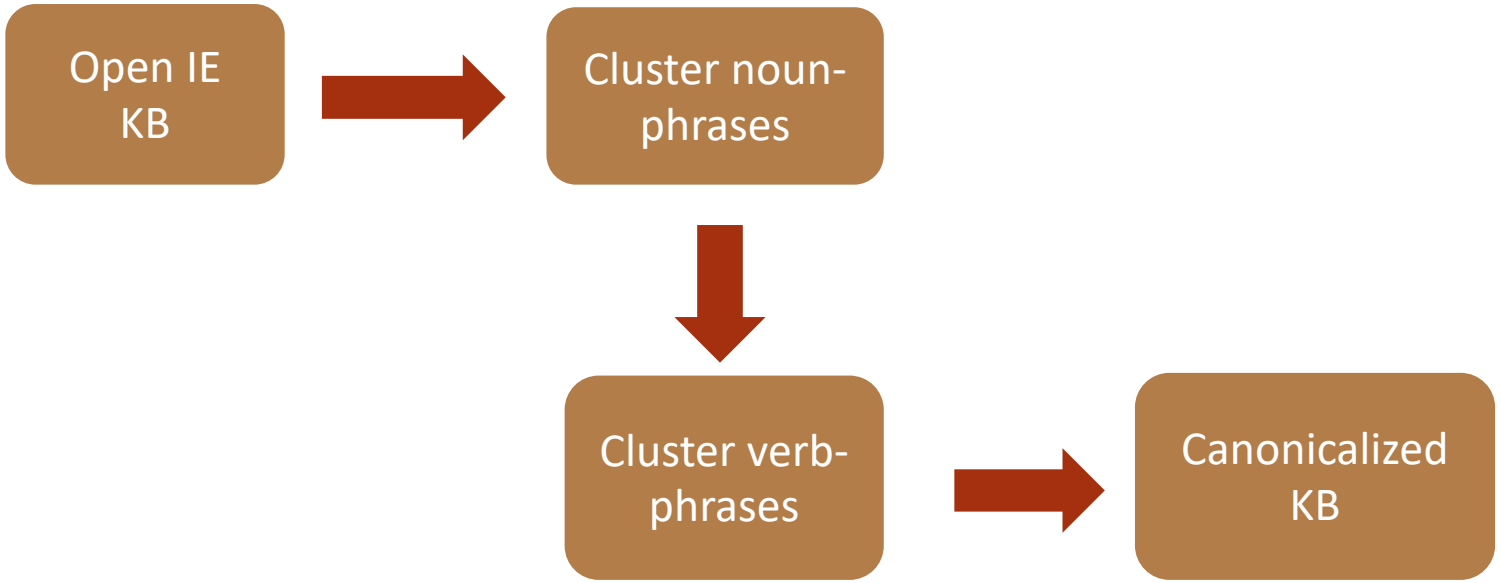
Defining domain	Learning extractors	Scoring candidate facts
		

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

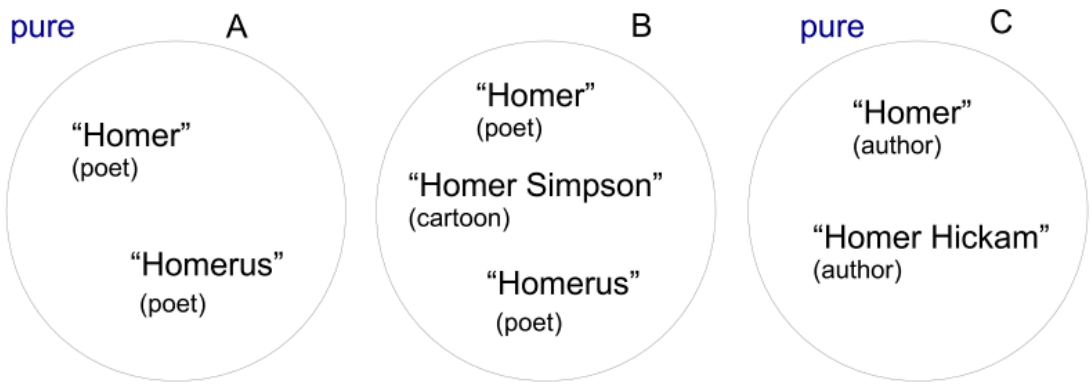
(5) Hybrid approach (adding structure to Open IE KB)



*[**Canonicalizing Open Knowledge Bases, Galárraga at al.,** CIKM 2014]*

(5) Hybrid approach

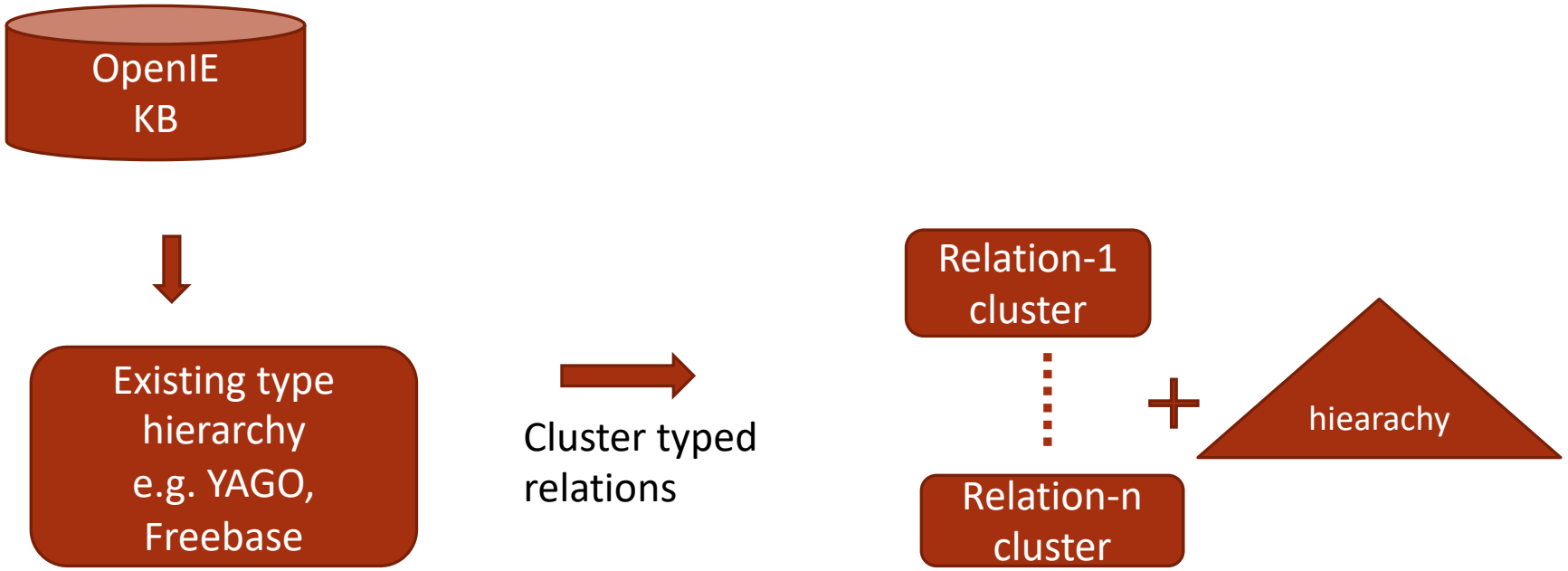
- ***Clustering entities***



- ***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

Open domain IE

Defining domain	Learning extractors	Scoring candidate facts
		

+

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

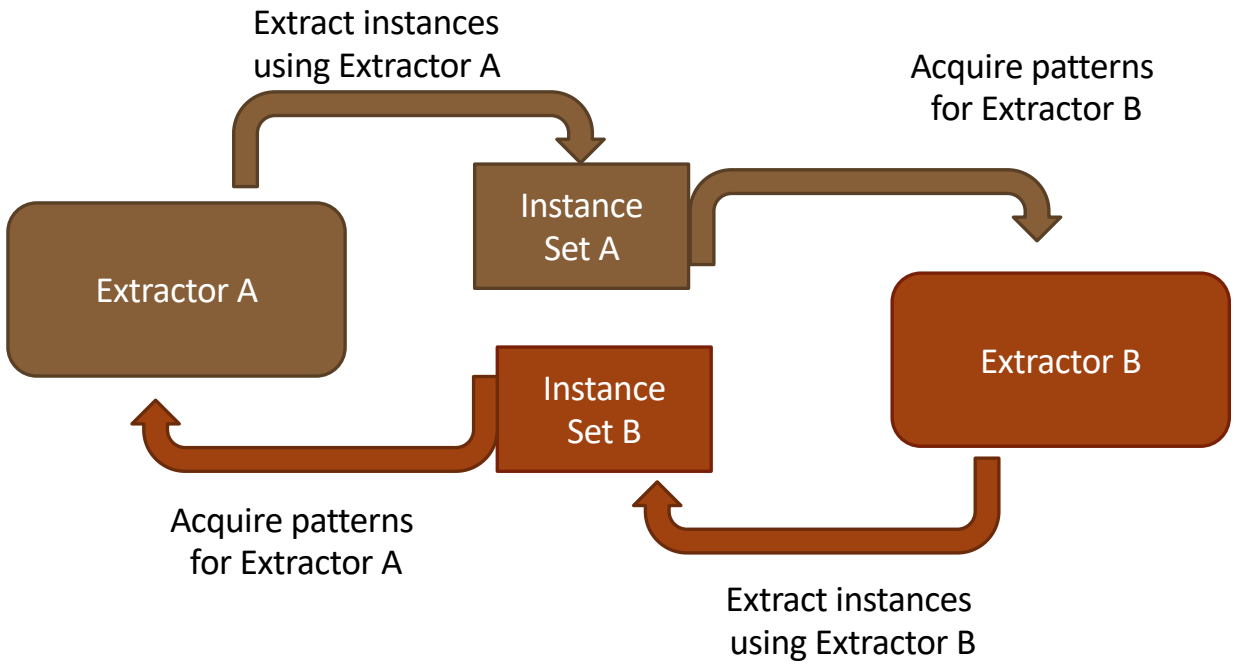
(1) Voting Schemes

- ***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)
    …..

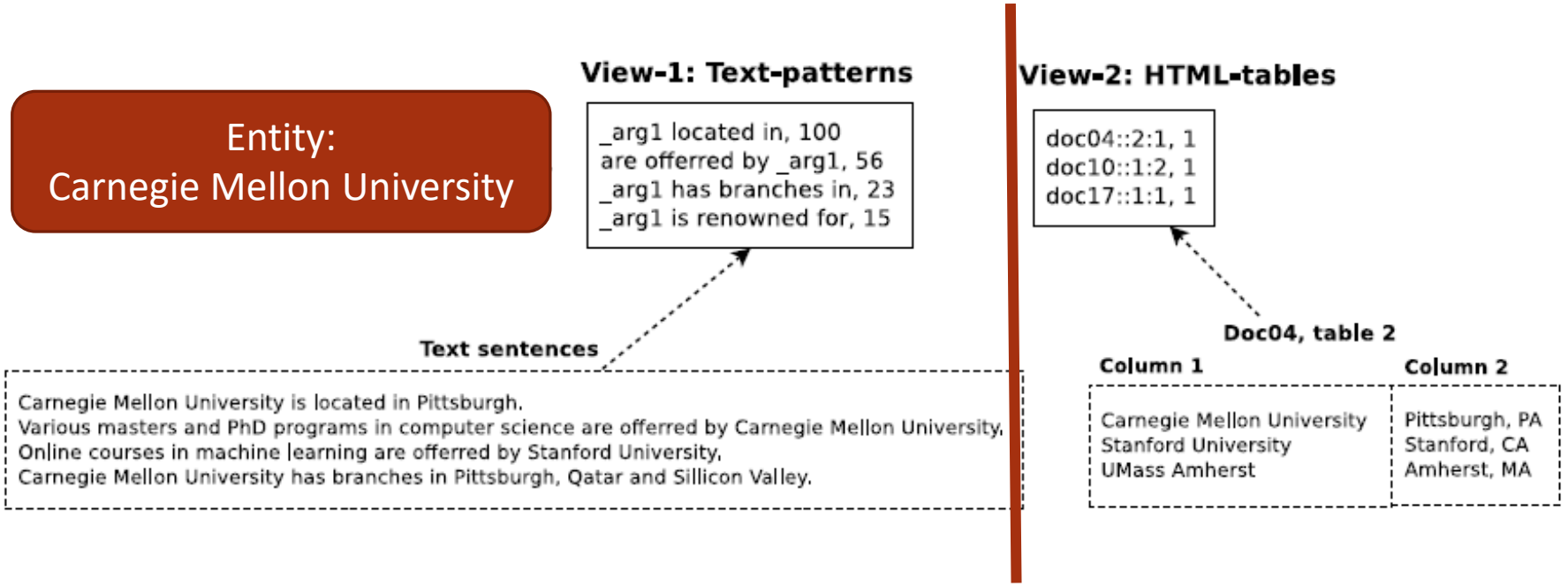
(2) Co-training



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

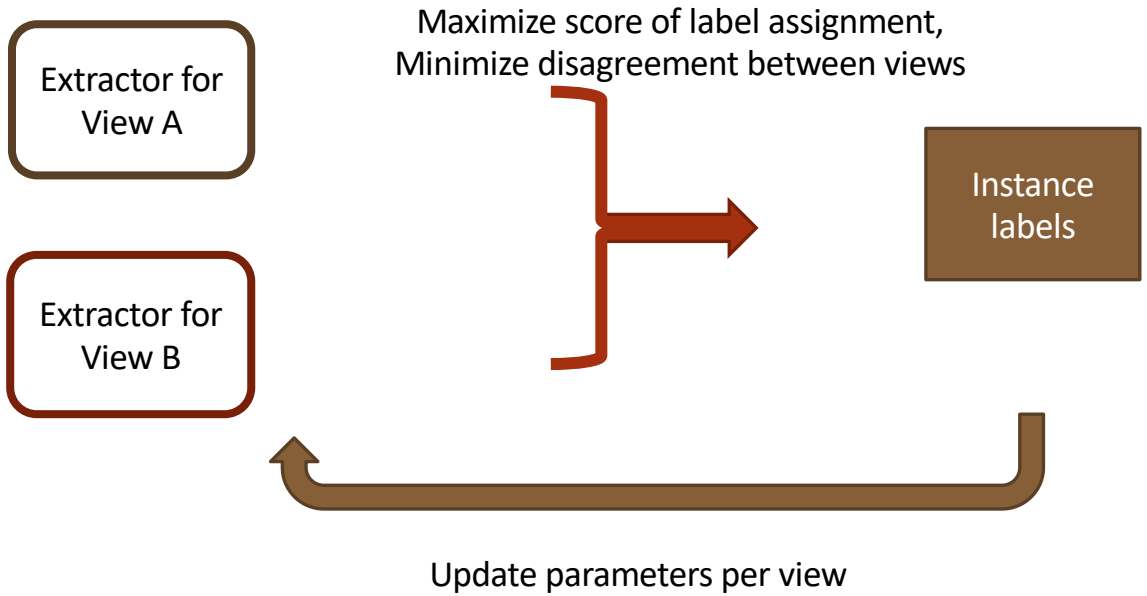
(3) Multi-view learning

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



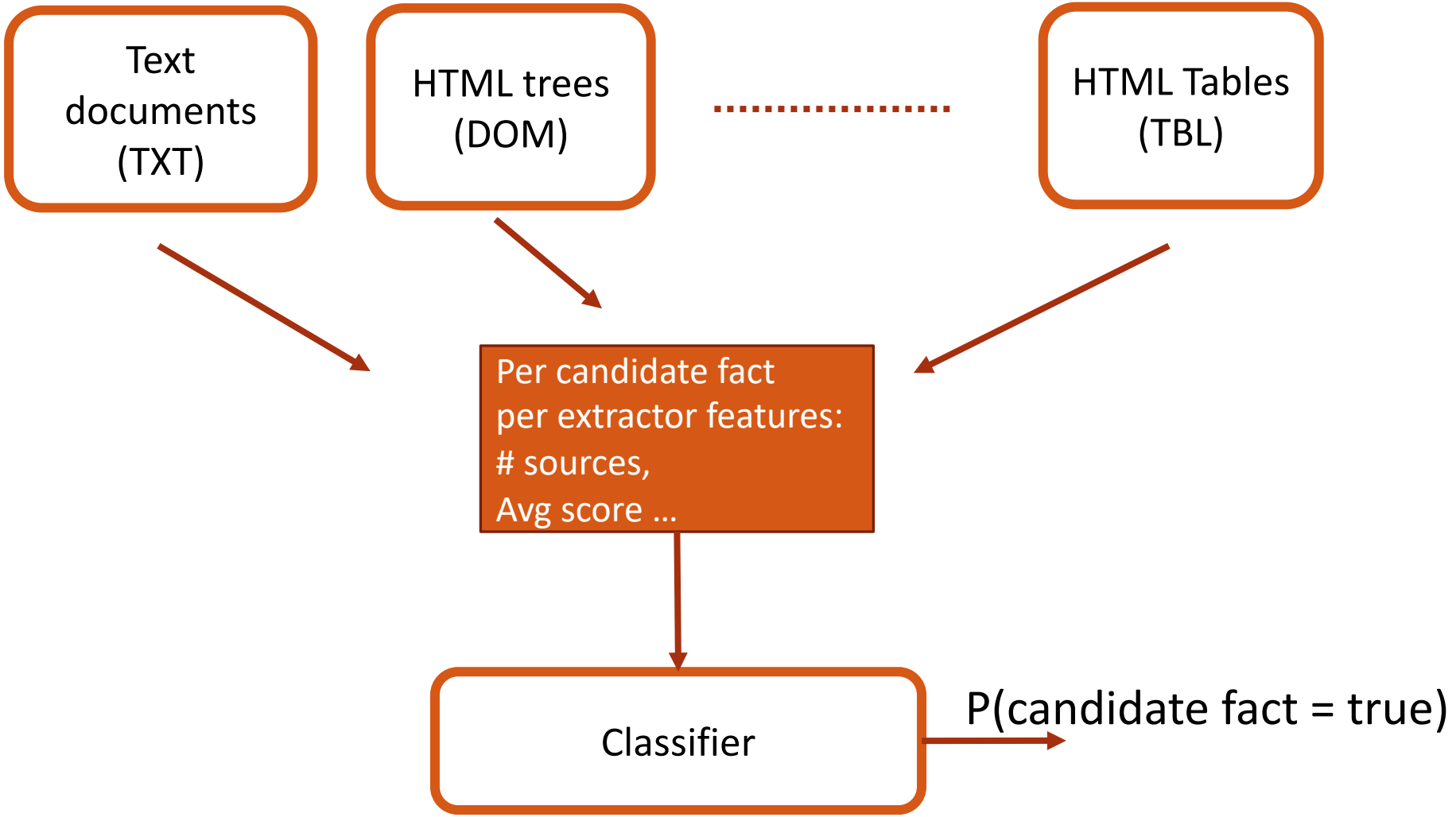
*[*Multi-View Hierarchical Semi-supervised Learning by Optimal Assignment of Sets of Labels to Instances, Dalvi et al. in preparation, link]

(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



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

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

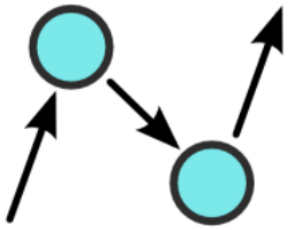
KNOWLEDGE FUSION

IE SYSTEMS IN PRACTICE

IE systems in practice

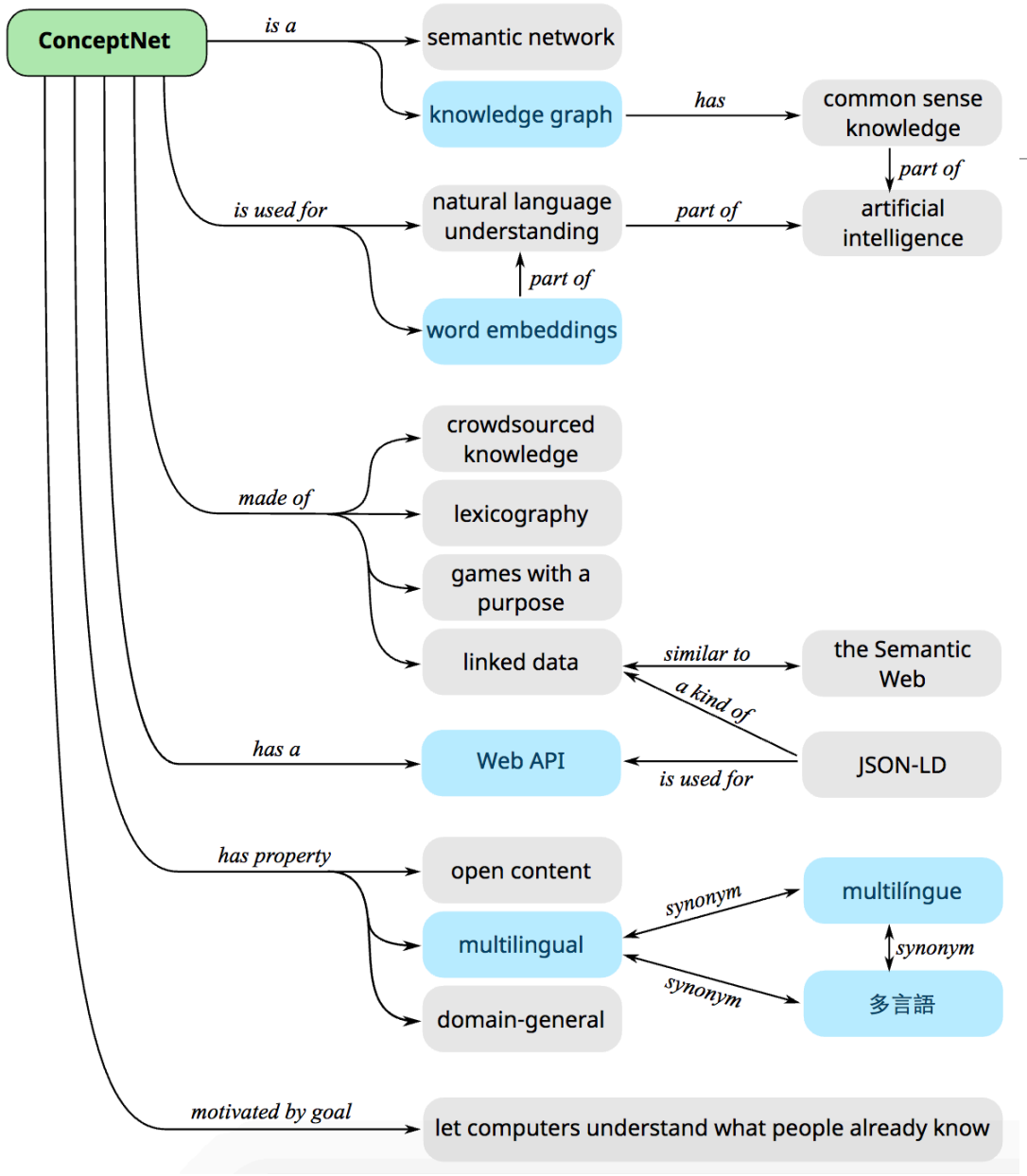
- Conceptnet
- NELL
- Knowledge vault
- Open IE

ConceptNet



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

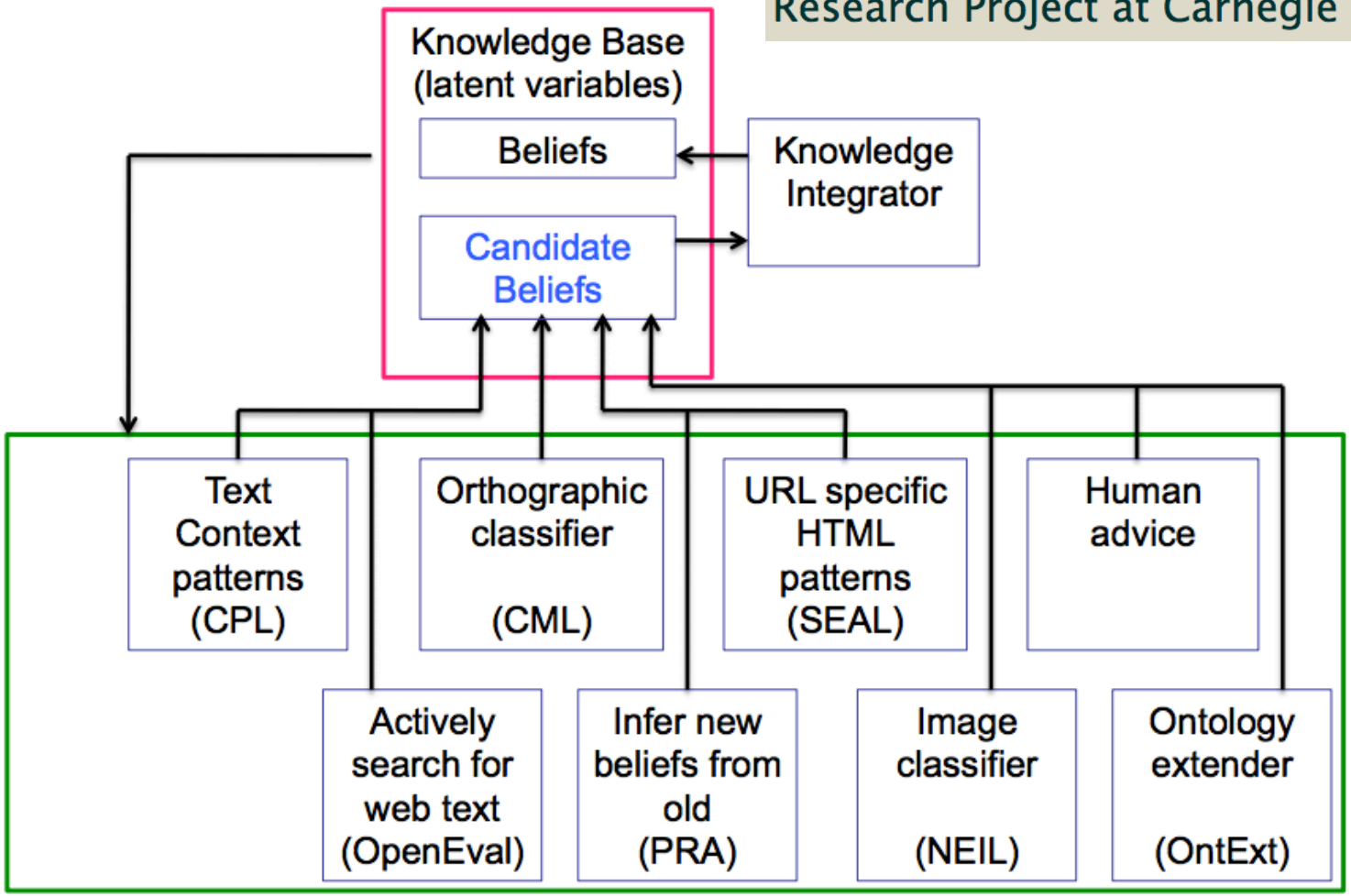
This knowledge was derived from thousands of human contributors.



Never Ending Language Learning (NELL)

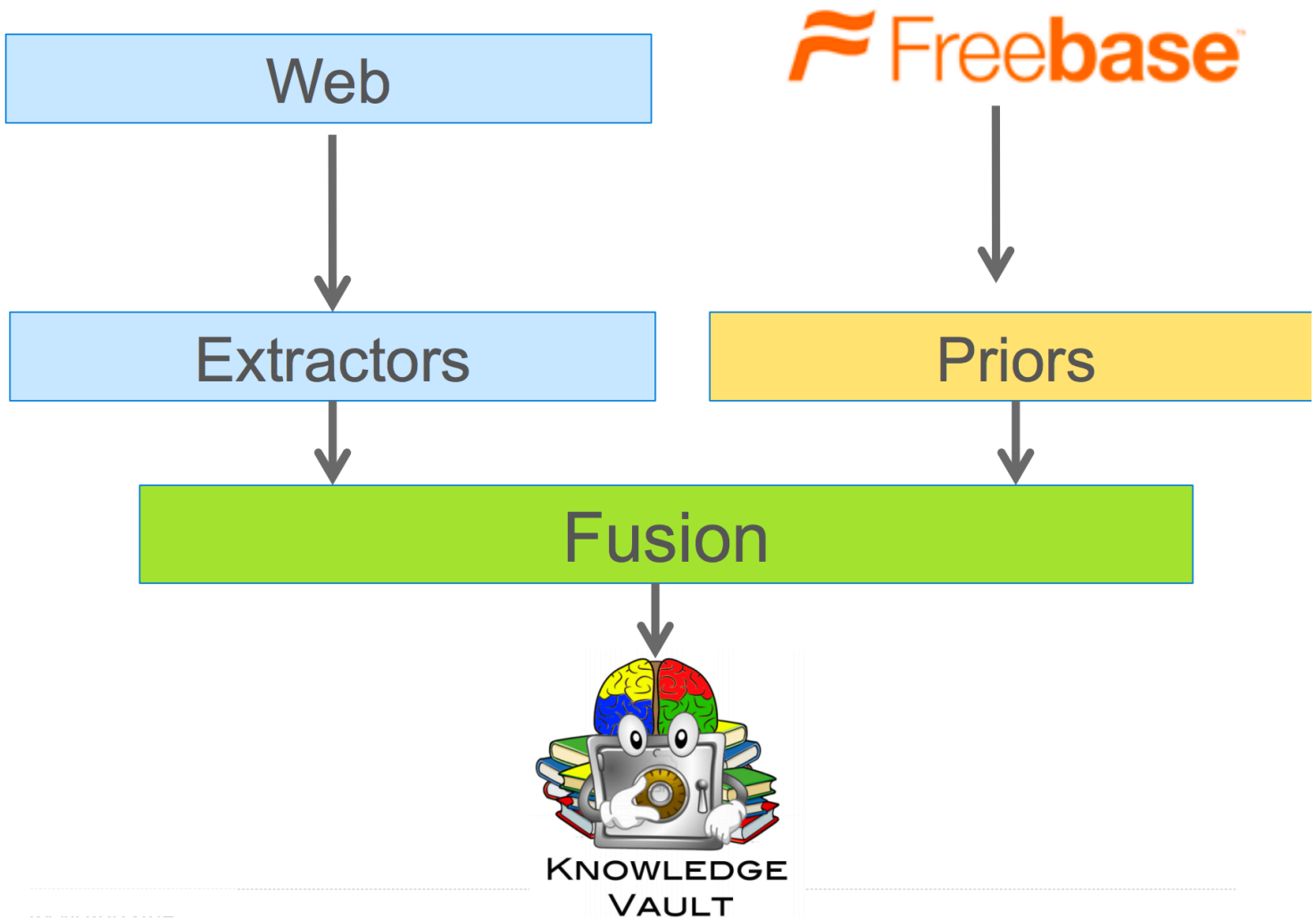
Read the Web

Research Project at Carnegie Mellon University



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

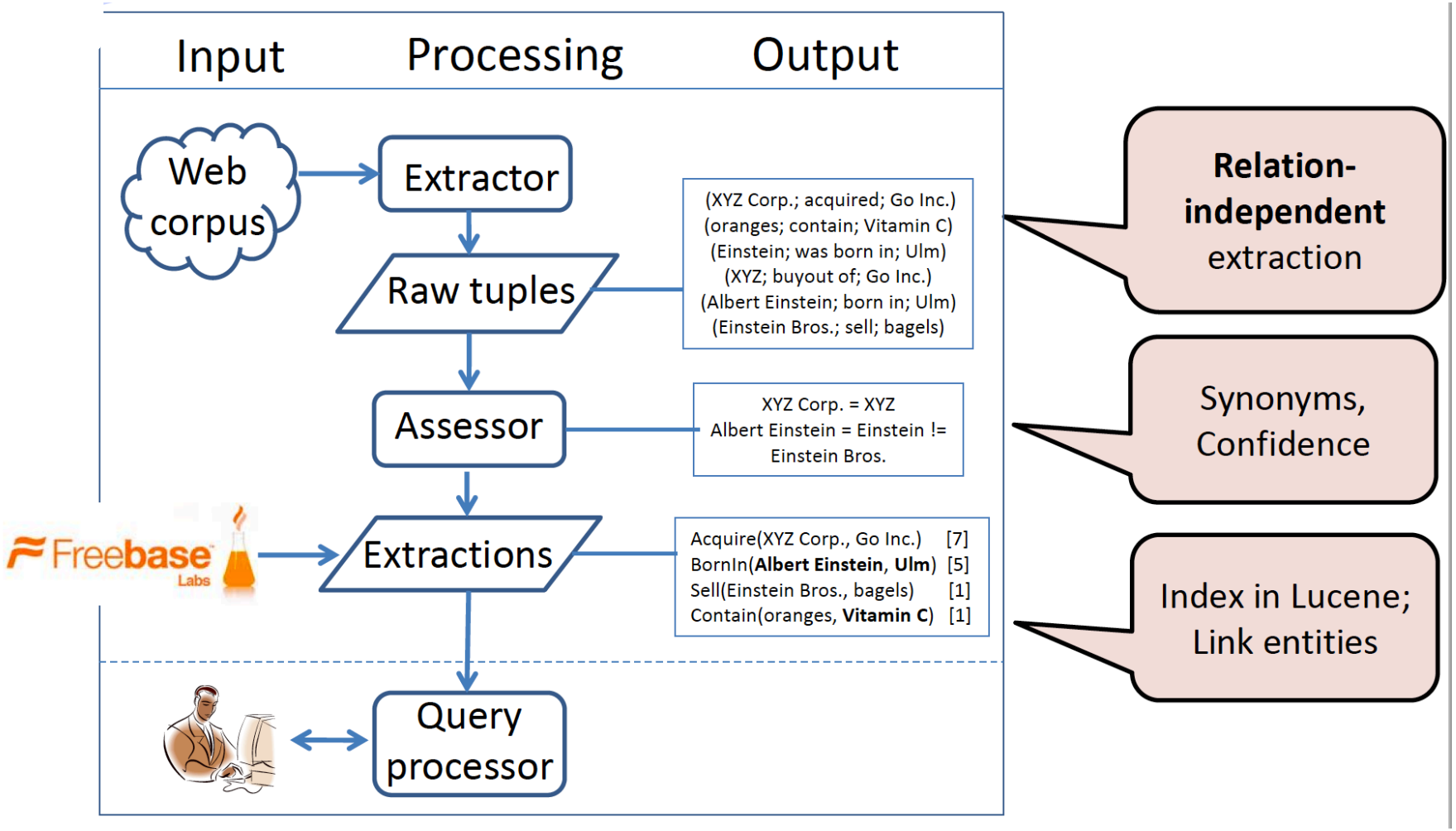
Knowledge Vault



[Architecture diagram taken from Kevin Murphy's slides]

Open IE (KnowItAll)

Open Information Extraction



[Architecture diagram taken from Oren Etzioni's slides]

IE systems 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