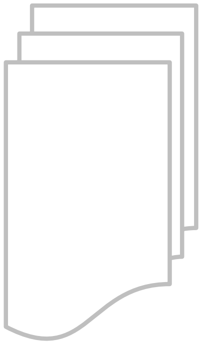
Mining Knowledge Graphs from Text

WSDM 2018

JAY PUJARA, SAMEER SINGH

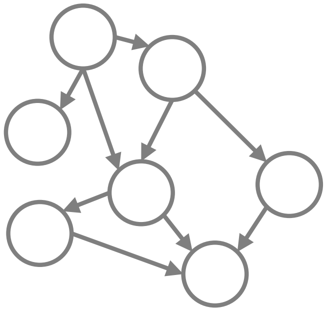
Tutorial Overview

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 Primer [Jay]** 
3. Knowledge Graph Construction
   a. Probabilistic Models [Jay] 

Coffee Break 

   b. Embedding Techniques [Sameer] 

4. Critical Overview and Conclusion [Sameer] 

What is NLP?

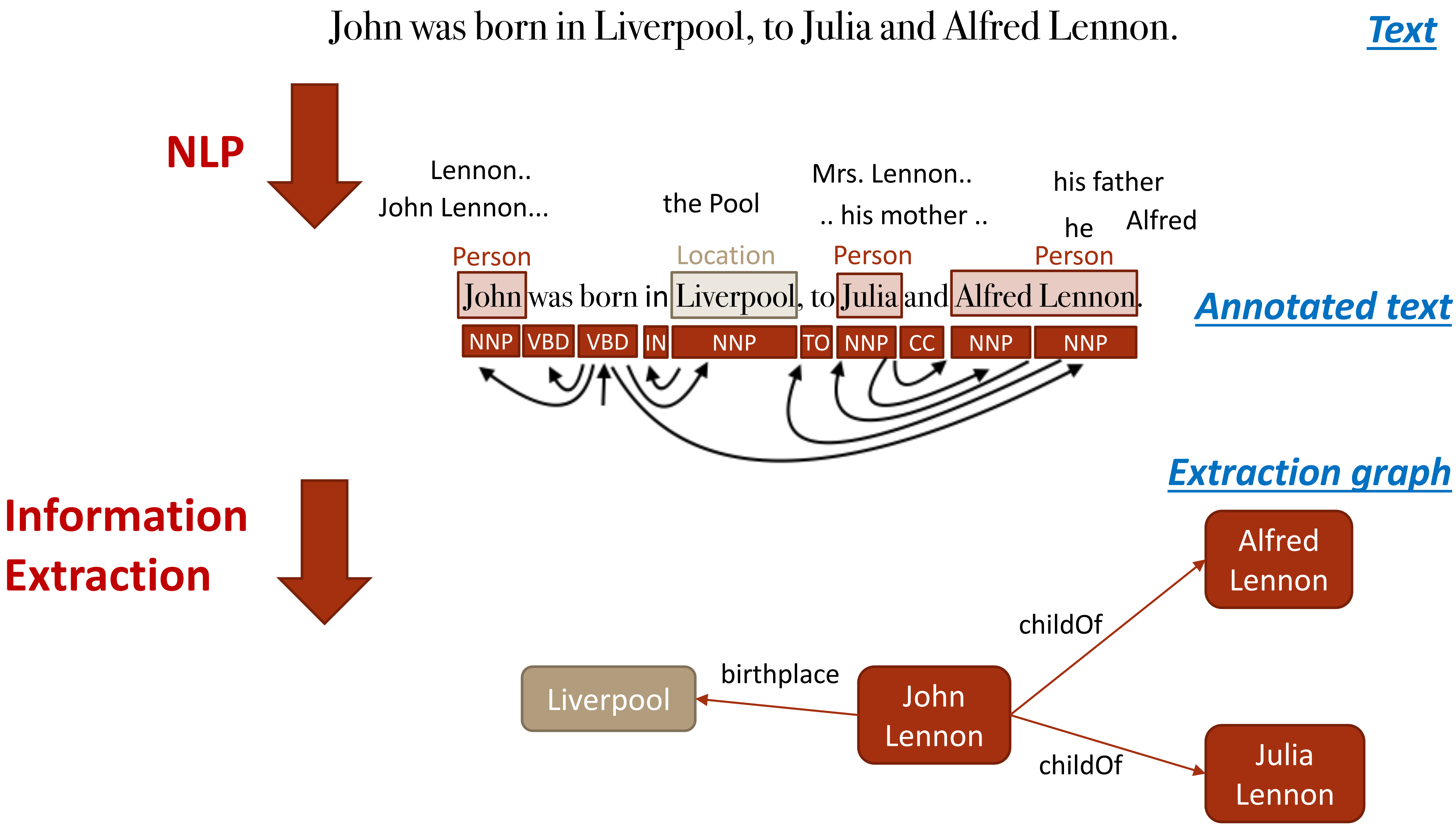
Unstructured
Ambiguous
Lots and lots of it!

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!

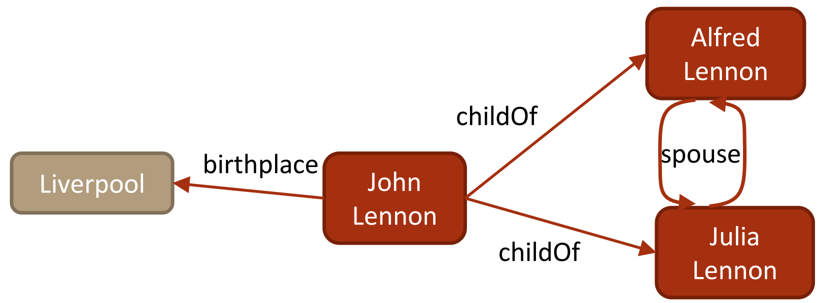
Knowledge Extraction



Breaking it Down

Information Extraction

Entity resolution,
Entity linking,
Relation extraction…



Document

Coreference Resolution…

Lennon..
John Lennon…

the Pool

Mrs. Lennon..
.. his mother ..

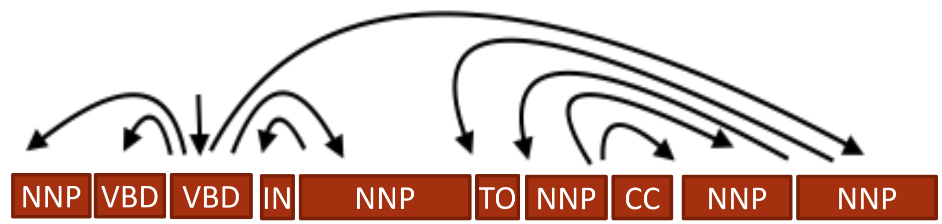
his father
he
Alfred

Person | Location | Person | Person

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

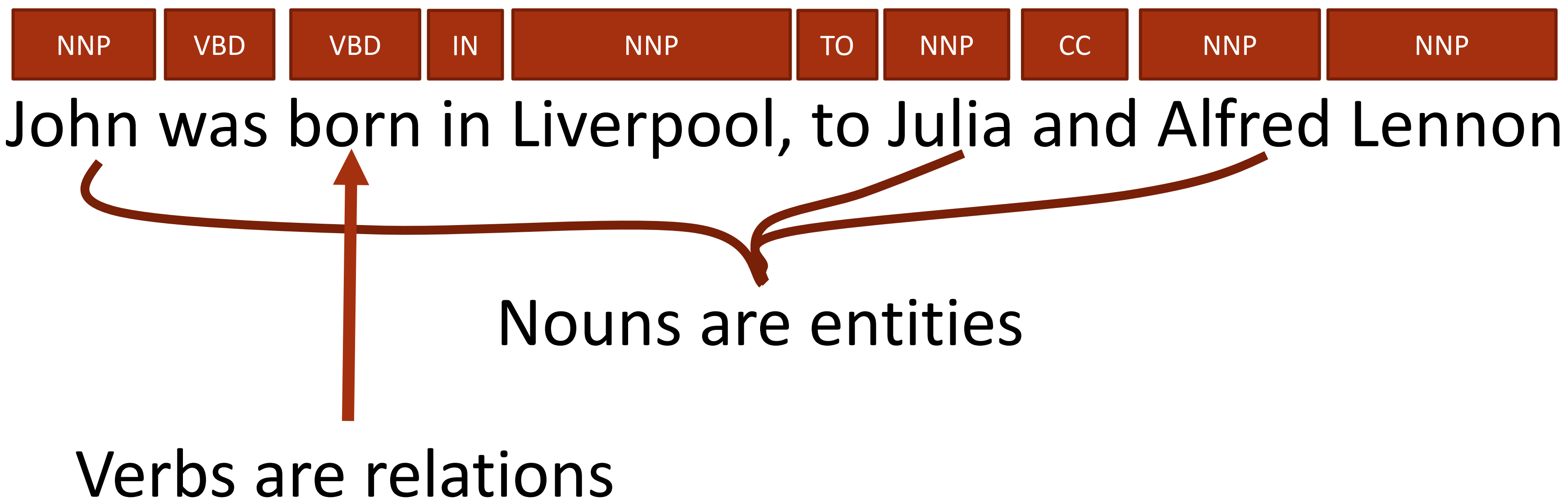
Sentence

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



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

Tagging the Parts of Speech



- Common approaches include CRFs, CNNs, LSTMs

Detecting Named Entities

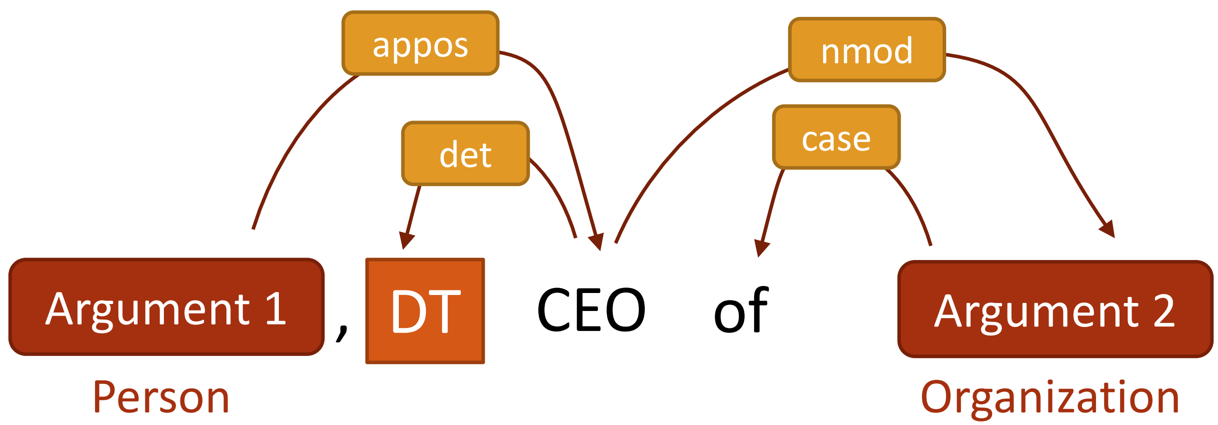
Person Location Person Person

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

- Structured prediction approaches
- Capture entity mentions and entity types

NLP annotations → features for IE

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…

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.

- Pairwise model for each noun/pronoun
- Can consolidate information, provide context

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, …

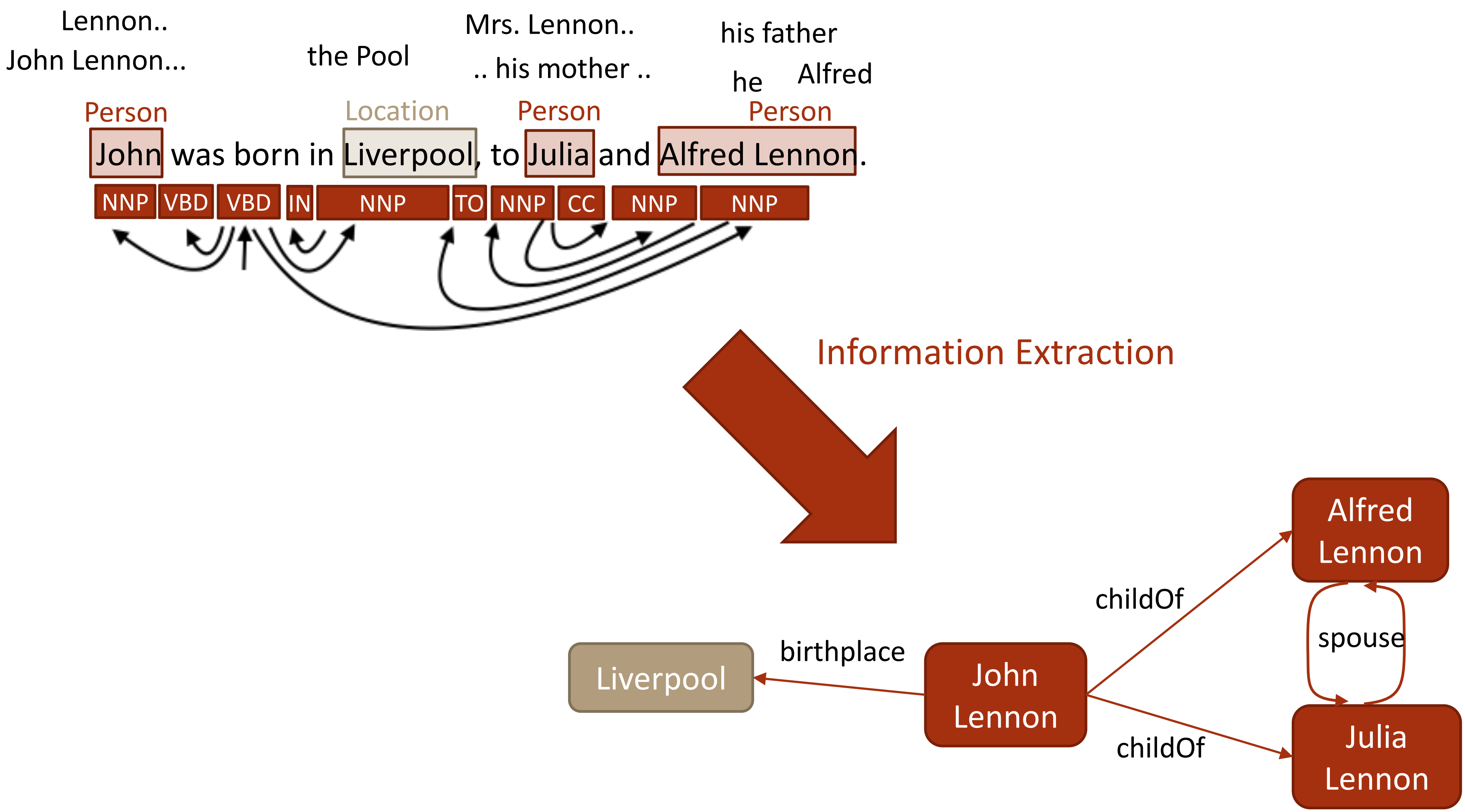
Coreference UWashington, Huskies

~~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~~, …

Information Extraction



Information Extraction

<u>3 CONCRETE SUB-PROBLEMS</u>

Defining domain

Learning extractors

Scoring the facts

×

<u>3 LEVELS OF SUPERVISION</u>

Supervised

Semi-supervised

Unsupervised

Effect of supervision on extractions

Precision,
Human efforts

Recall,
Speed

Information Extraction

3 CONCRETE SUB-PROBLEMS

Defining domain

Learning extractors

Scoring the facts

×

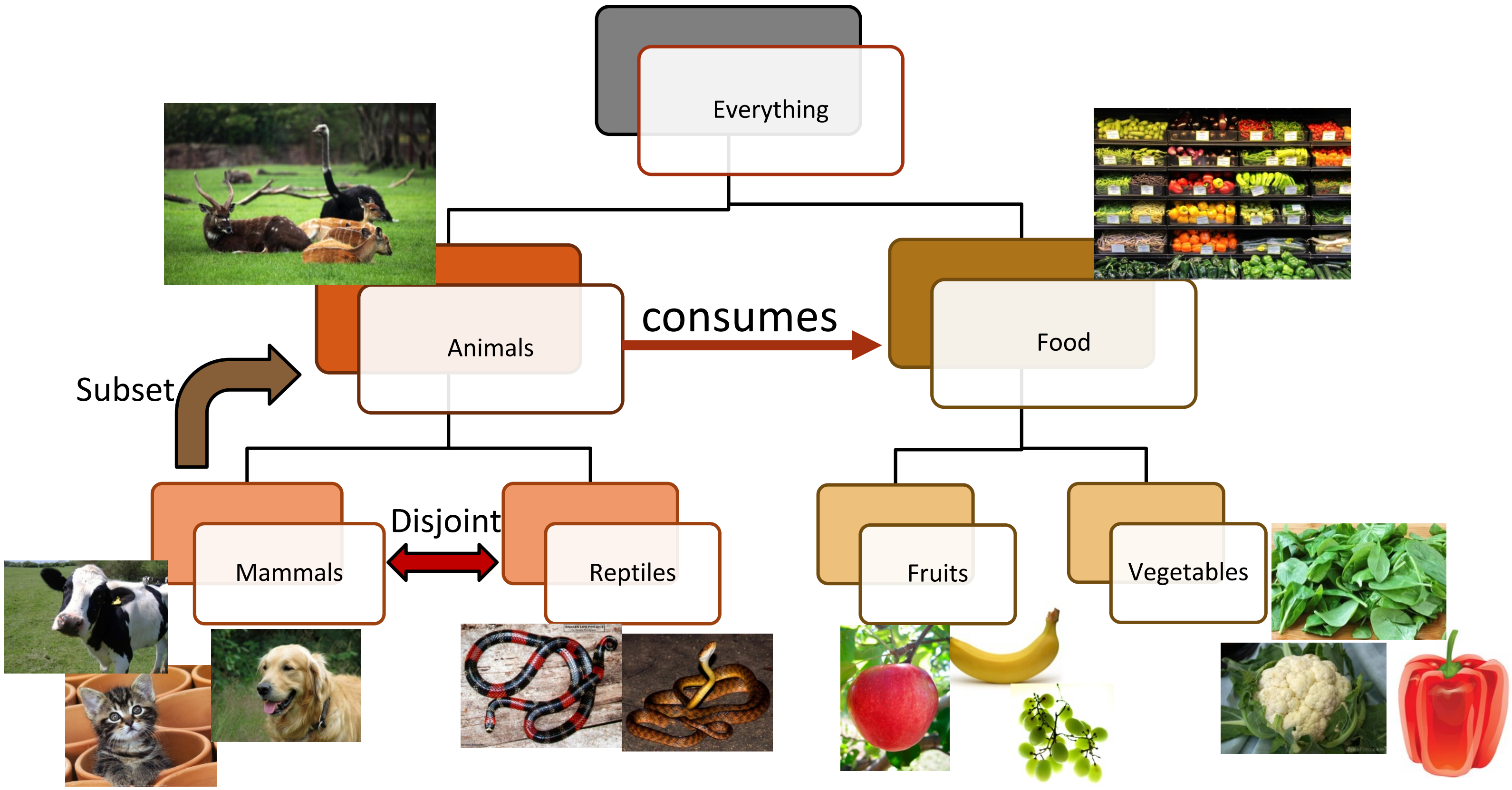
3 LEVELS OF SUPERVISION

Supervised 

Semi-supervised 

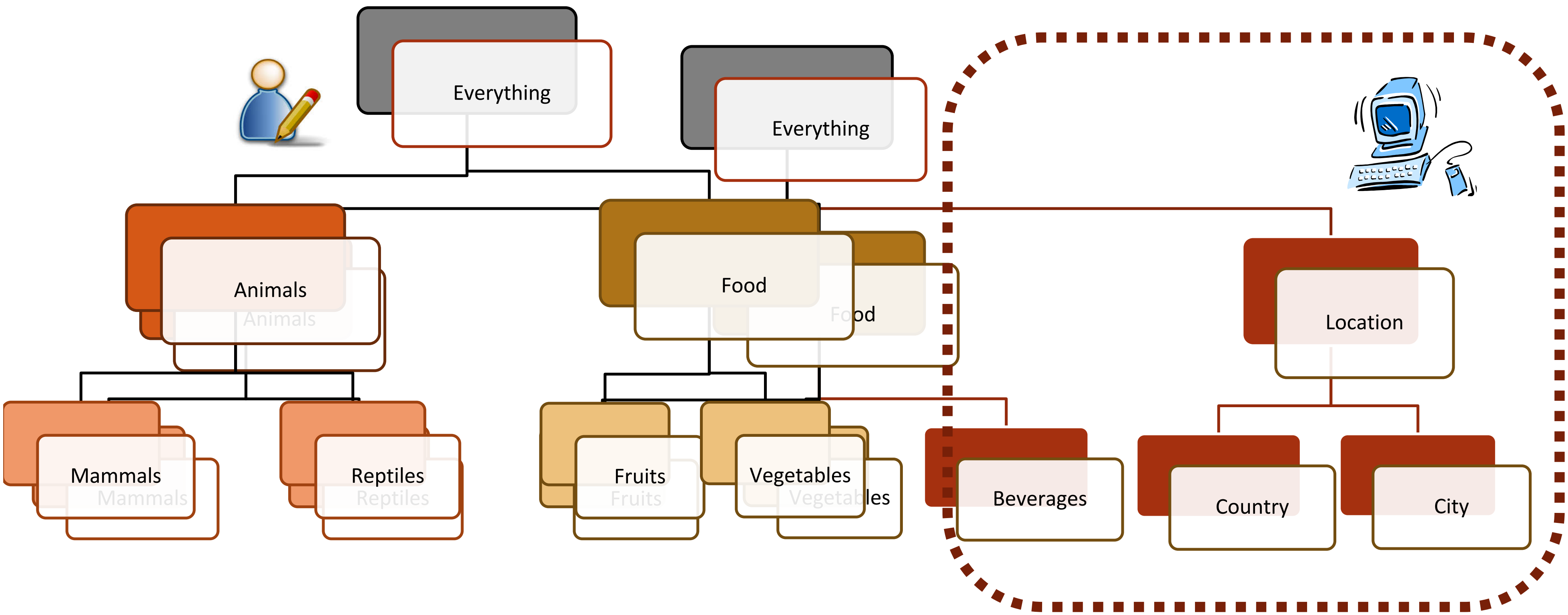
Unsupervised 

Defining Domain: Manual



Defining Domain: Semi-automatic

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



Defining Domain: Automatic

- Any noun phrase is a candidate entity
  - Dog, cat, cow, reptile, mammal, apple, greens, mixed greens, lettuce, red leaf lettuce, romaine lettuce, iceberg lettuce…
- Any verb phrase is a candidate relation
  - Eats, feasts on, grazes, consumes,

[**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 

Learning Extractors

- Supervised: high precision patterns
  - <PERSON> plays in <BAND>

- Semi-supervised: Bootstrapping to learn patterns
  - Create examples (John Lennon, Beatles), find patterns
  - Manually correct incorrect patterns

- Unsupervised: cluster phrases with constraints
  - Identify candidate verb phrases, find candidate arguments, cluster by NER types

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

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

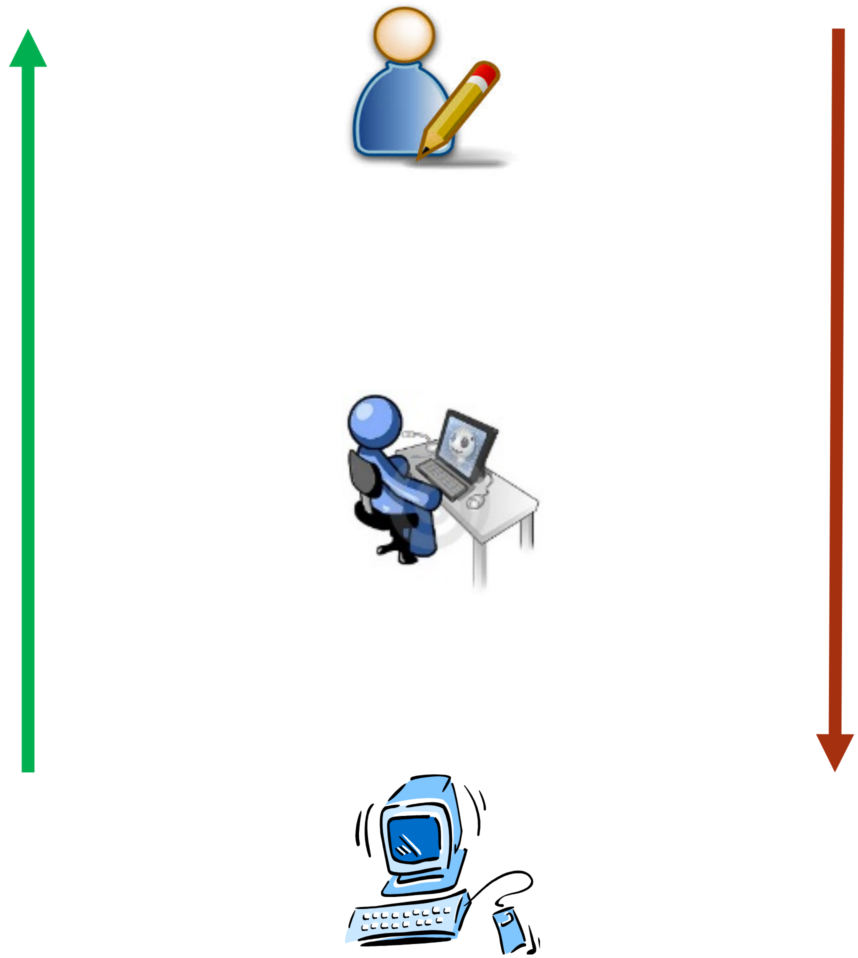
Early removal of noisy extraction pattern can avoid semantic drift in later stages

Scoring the candidate facts

Domain expertise needed

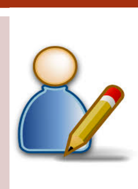
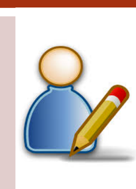
Effect of supervision on extractions

Precision,
Human efforts



Recall,
Speed

IE systems in practice

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

Knowledge Extraction: Key Points

- Built on the foundation of NLP techniques
  - Part-of-speech tagging, dependency parsing, named entity recognition, coreference resolution…
  - Challenging problems with very useful outputs
- Information extraction techniques use NLP to:
  - define the domain
  - extract entities and relations
  - score candidate outputs
- Trade-off between manual & automatic methods