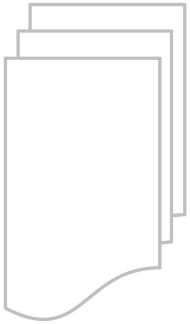
NLP Fundamentals

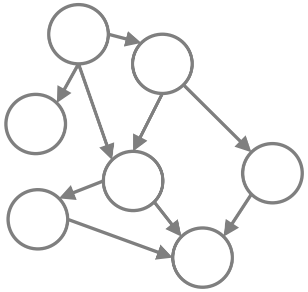
EXTRACTING STRUCTURES FROM LANGUAGE

Part 1: Knowledge Graphs

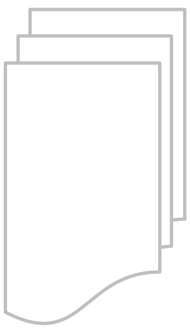


Part 2: Knowledge Extraction

Part 3: Graph Construction



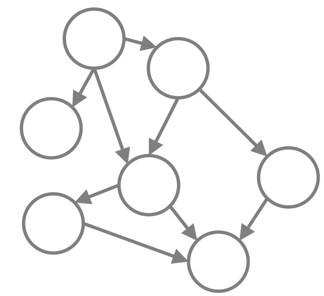
Part 4: Critical Analysis



Part 1: Knowledge Graphs

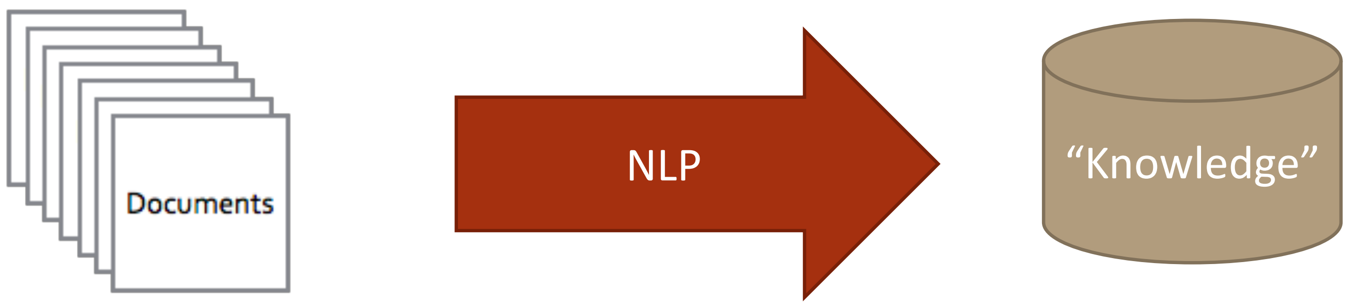
Part 2: Knowledge Extraction

Part 3: Graph Construction



Part 4: Critical Analysis

What is NLP?



Unstructured
Ambiguous
Lots and lots of it!

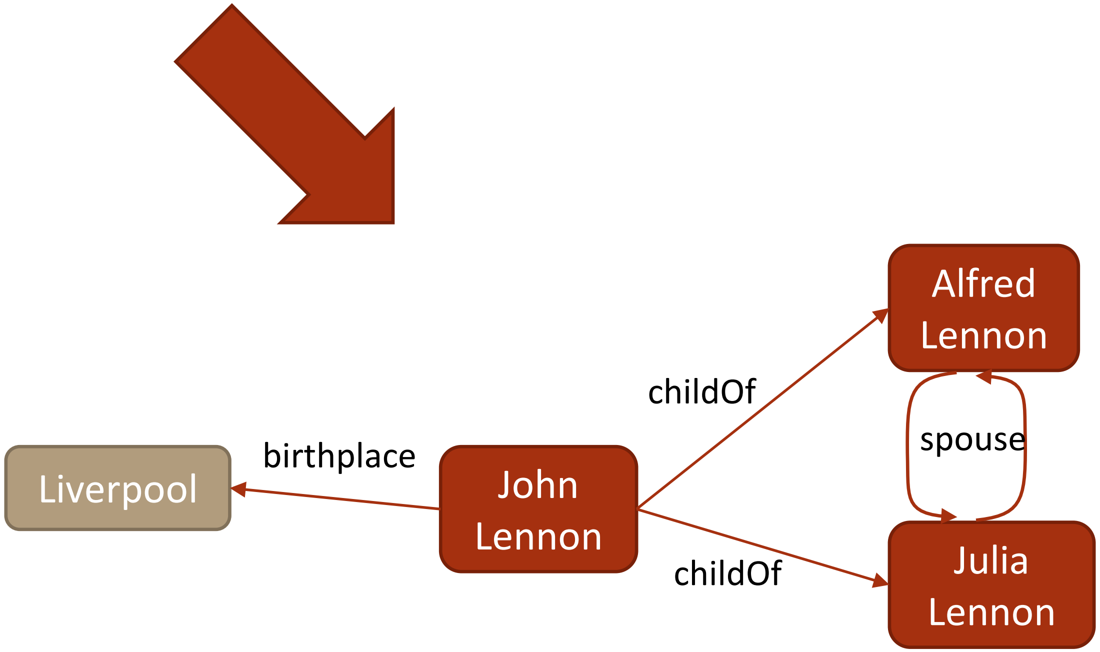
Humans can read them, but
… very slowly
… can’t remember all
… can’t answer questions

Structured
Precise, Actionable
Specific to the task

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

Why do we need NLP?

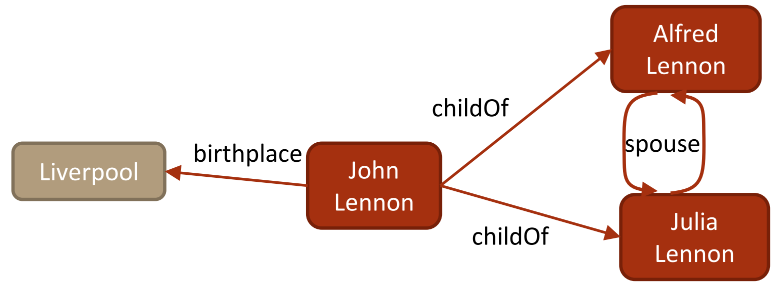
John was born in Liverpool, to Julia and Alfred Lennon.



Breaking it Down

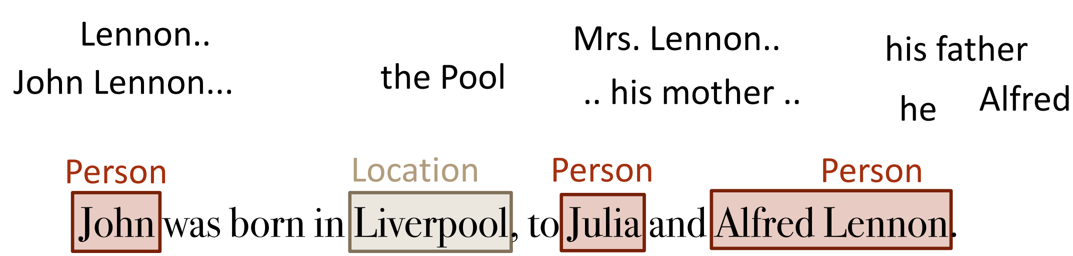
Corpus

Entity resolution,
Entity linking,
Relation extraction…



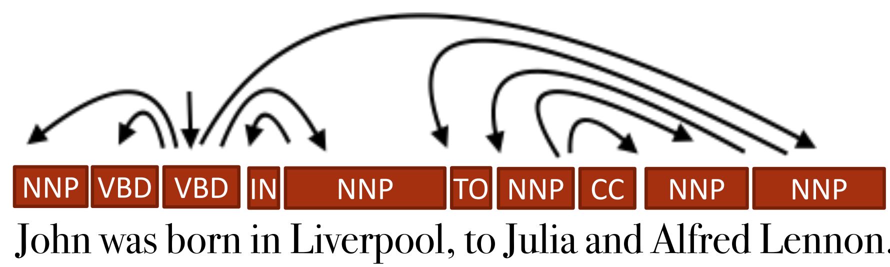
Document

Within-doc Coreference…



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

NNP	VBD	VBD	IN	NNP	TO	NNP	CC	NNP	NNP
John	was	born	in	Liverpool,	to	Julia	and	Alfred	Lennon.

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

Person: John — Location: Liverpool — Person: Julia — Person: Alfred Lennon

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

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		

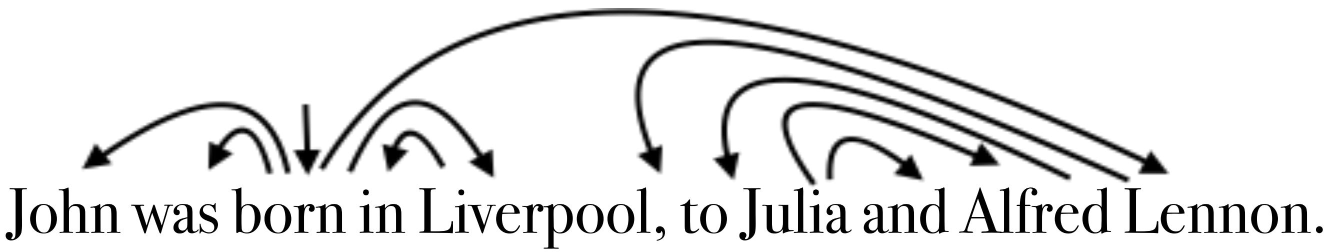
location	body_of_water	**product**	camera	**art**	written_work
city	island	engine	mobile_phone	film	newspaper
country	mountain	airplane	computer	play	music
county	glacier	car	software	**event**	military_conflict
province	astral_body	ship	game	attack	natural_disaster
railway	cemetery	spacecraft	instrument	election	sports_event
road	park	train	weapon	protest	terrorist_attack
bridge					

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)

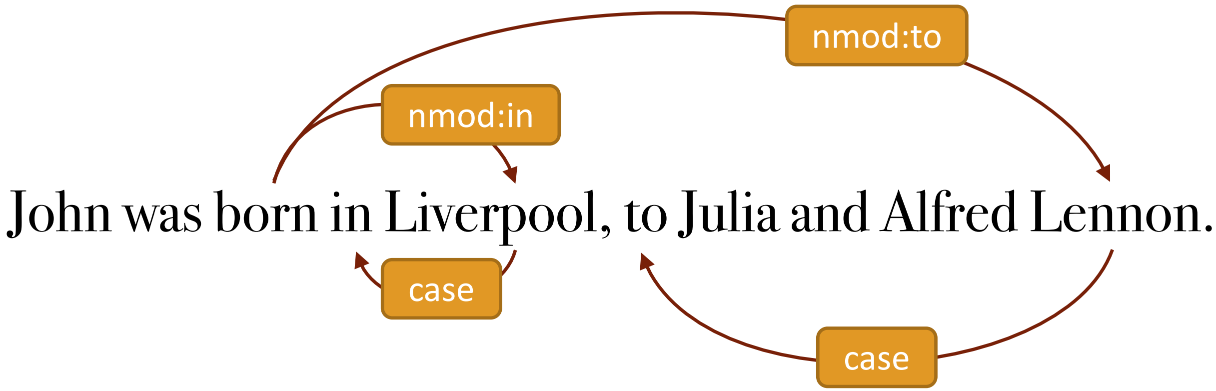
Dependency Parsing



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

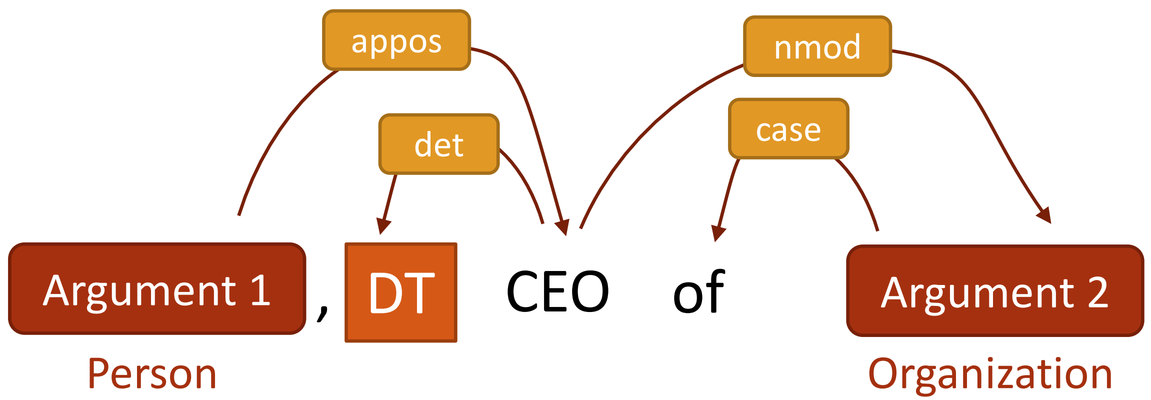
Dependency Paths



	Text Patterns	Dependency Paths
Sanders, Brooklyn	"was born in"	"was born in"
Sanders, Dorothy	"was born in Brooklyn, to"	"was born to"
Sanders, Eli Sanders	"was born in Brooklyn, to Dorothy and"	"was born to"

Surface Patterns

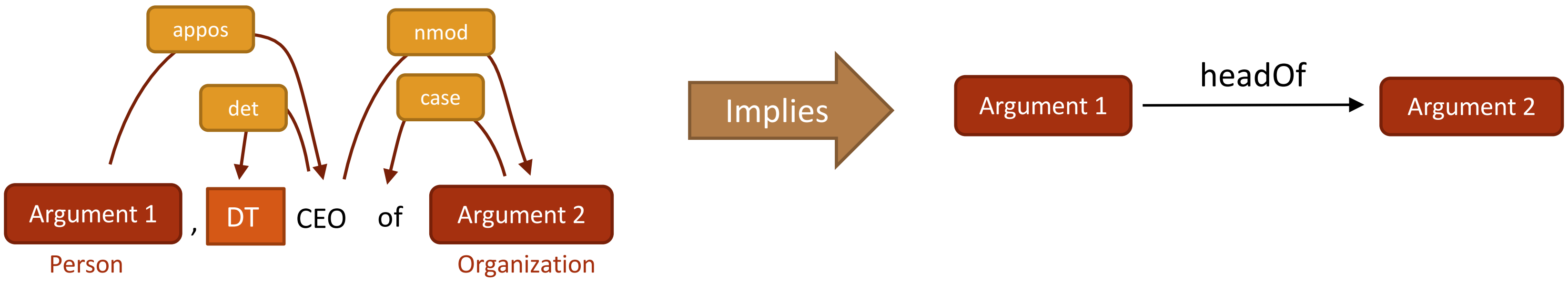
Combine tokens, dependency paths, and entity types to define rules.



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

Rule-Based Relation 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!

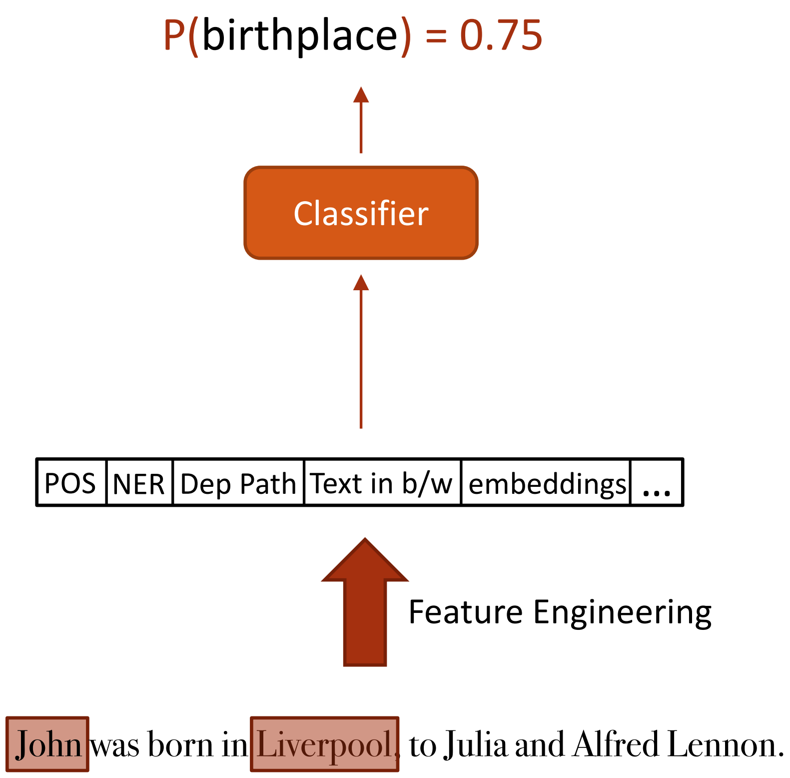
Supervised Relation Extraction

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!



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

Entity Disambiguation & 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, …

Typos/Misspellings

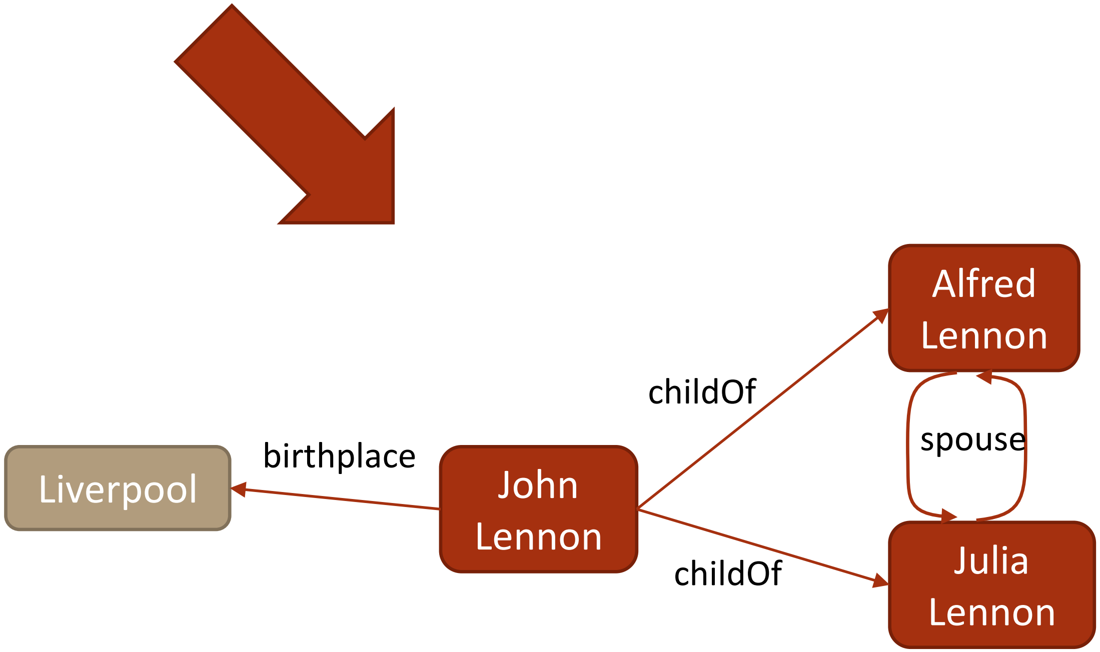
Baarak, Barak, Barrack, …

Inconsistent References

Bam Bam, …

Review: What NLP gives us

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



Information Extraction