

Information Extraction

Prof. Sameer Singh

CS 295: STATISTICAL NLP

WINTER 2017

February 21, 2017

Outline

What is Information Extraction

Named Entity Recognition

Homework 3

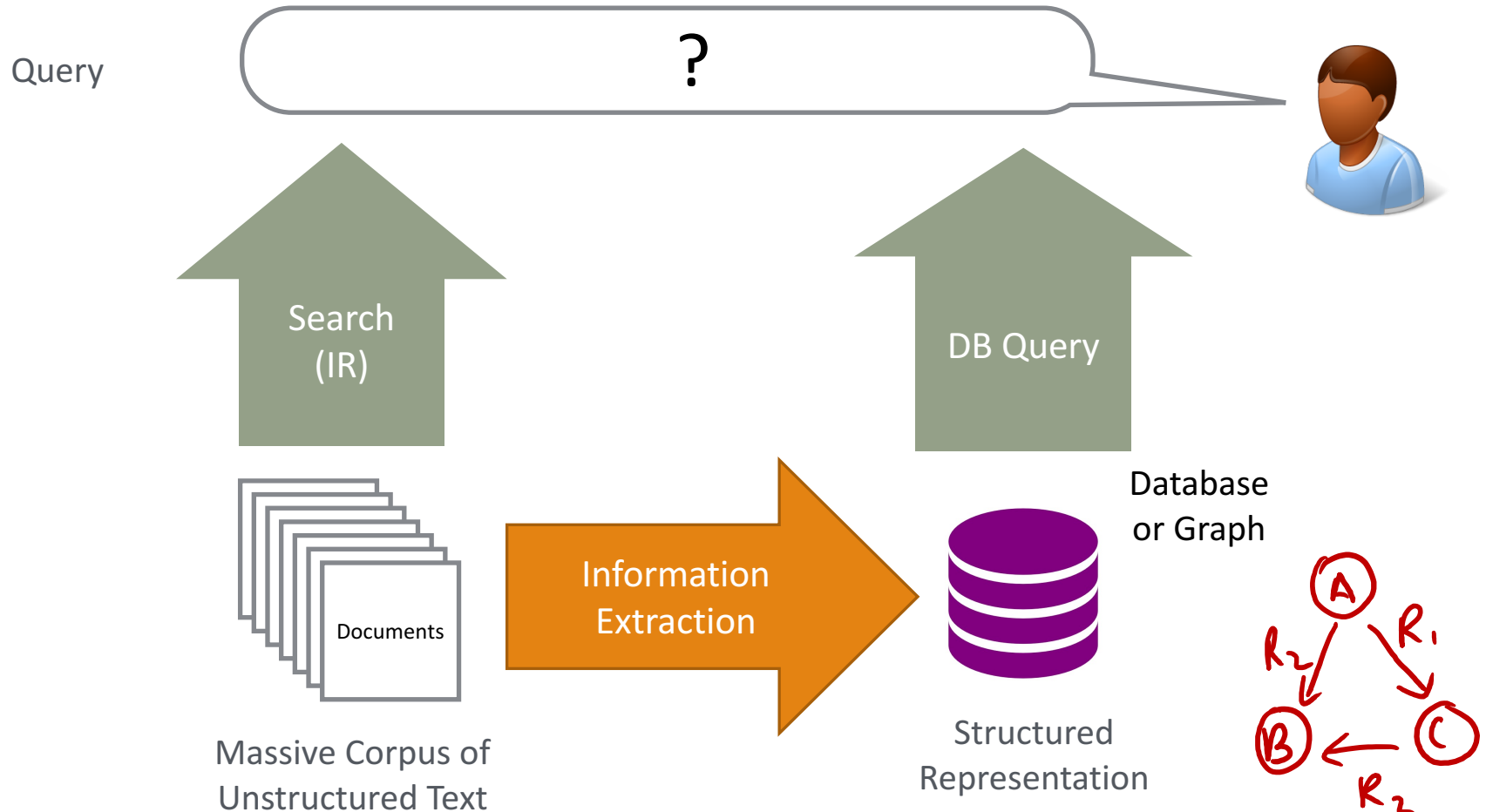
Outline

What is Information Extraction

Named Entity Recognition

Homework 3

Making Sense of Text



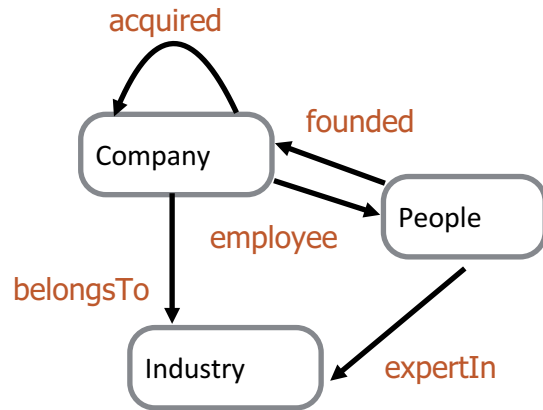
News Articles

Query

Which AI startups have been acquired by Tech companies?



Structured
Representation



Information
Extraction

Massive Corpus of
News Articles



The New York Times

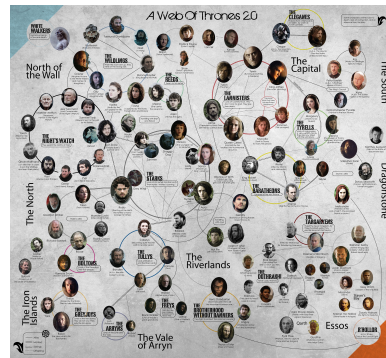


Fiction

Query

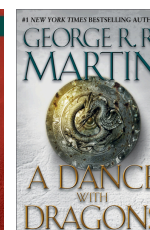
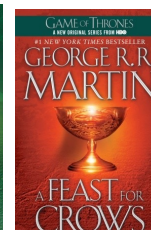
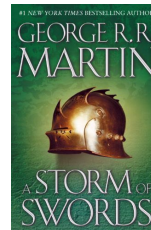
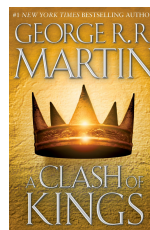
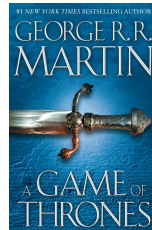
Which two characters are not related by blood?

Structured
Representation



Information
Extraction

Collection of
Books

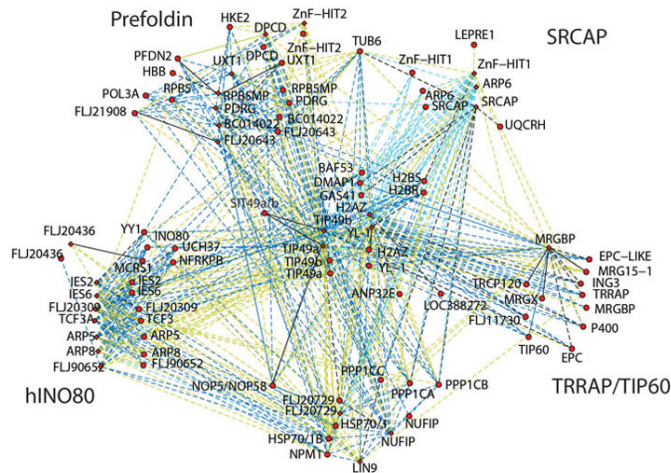


Academic Research

Query

What is the interaction pathway between YY1 and TIP60?

Structured
Representation



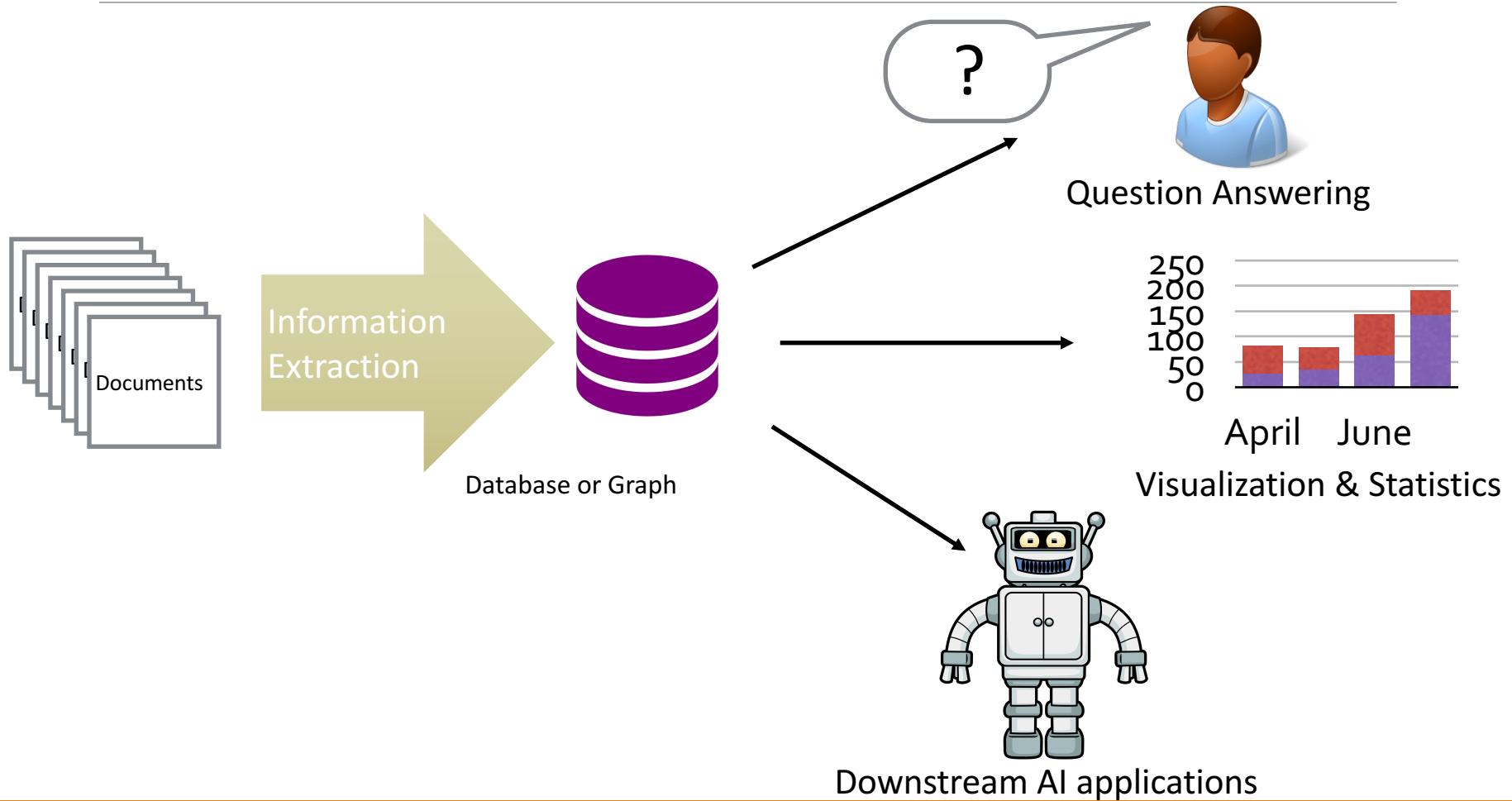
Massive Corpus of
Scientific Papers



Information
Extraction



Applications





Low-level Info. Extraction

The Los Altos Robotics Board of Directors is having a potluck dinner Friday January 6, 2012 and the upcoming [Botball](#) and FRC ([MVHS Eagle Strike Robotics](#)) seasons. You are back and it was a of these dinners three years

Create New iCal Event...
Show This Date in iCal...

Copy

Slightly better...


 


[All](#) [News](#) [Shopping](#) [Videos](#) [Maps](#) [More](#) [Settings](#) [Tools](#)

About 15,300,000 results (0.66 seconds)


Super Bowl LI

Super Bowl
Sunday, February 5, 6:30 PM on FOX
NRG Stadium, Houston, Texas

 **New England Patriots**

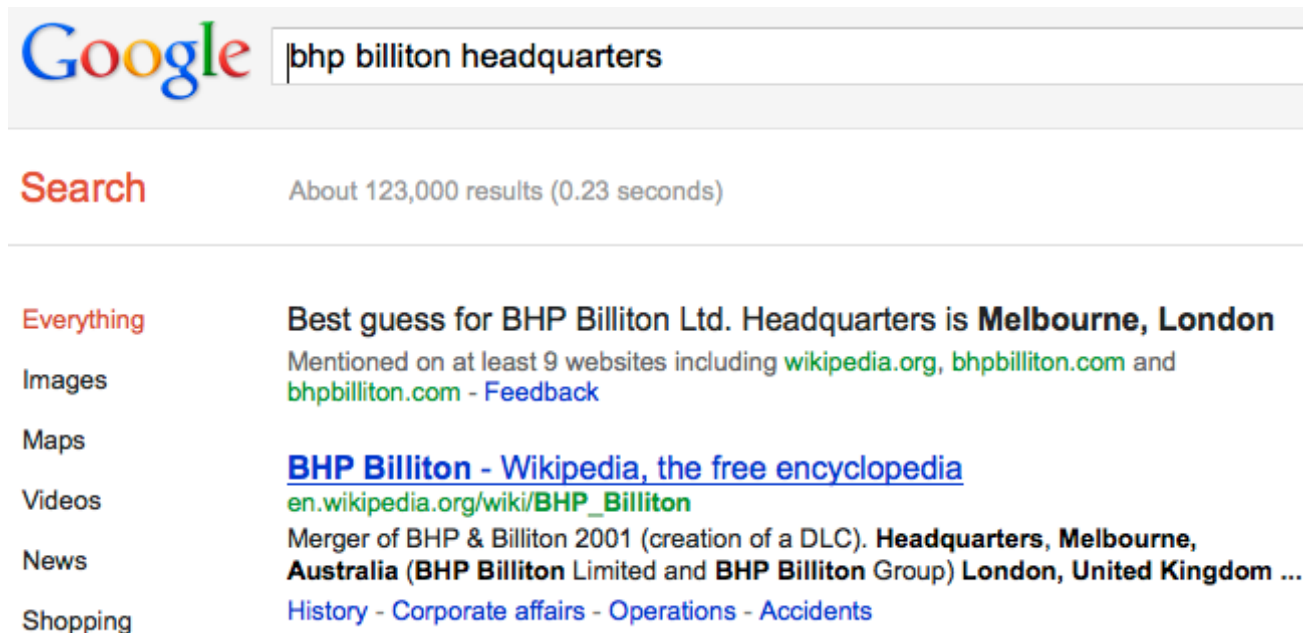
 **Atlanta Falcons**

[Tickets - Preview](#)

 3:43

All times are in Eastern Time

Slightly better?



The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.

headquarters("BHP Biliton Limited", "Melbourne, Australia")

In the industry...

Google Knowledge Graph

- Google Knowledge Vault

Amazon Product Graph

Facebook Graph API

IBM Watson

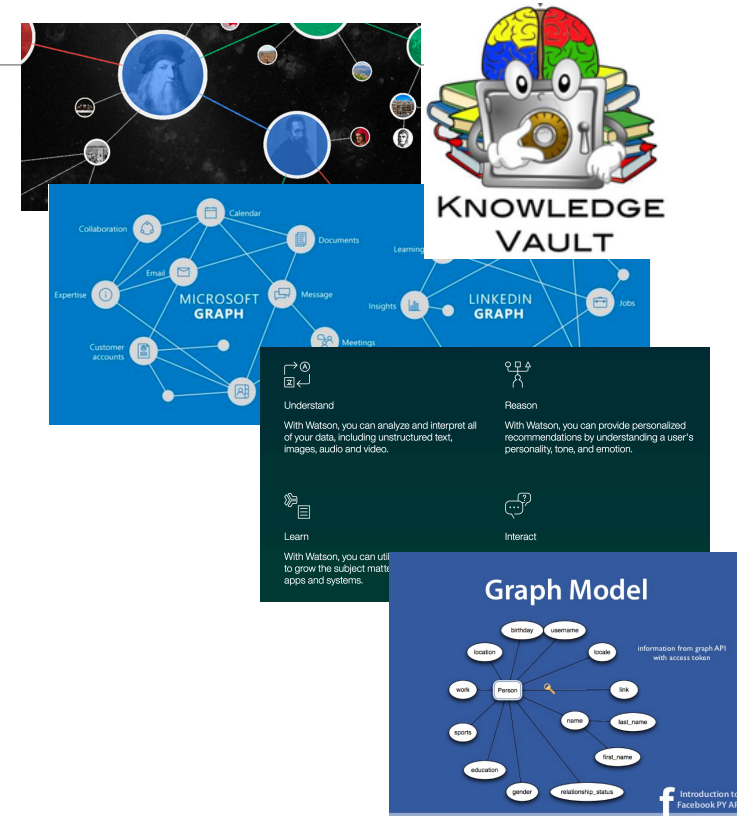
Microsoft Satori

- Project Hanover/Literome

LinkedIn Knowledge Graph

Yandex Object Answer

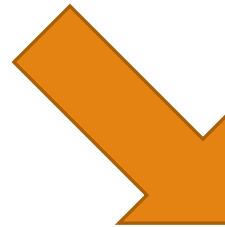
Diffbot, GraphIQ, Maana, ParseHub, Reactor Labs, SpazioDati



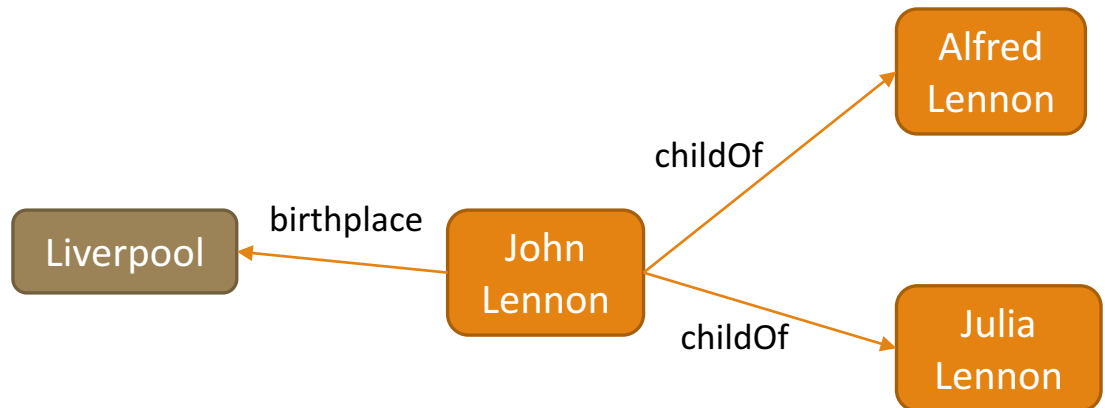
Knowledge Extraction

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

Text



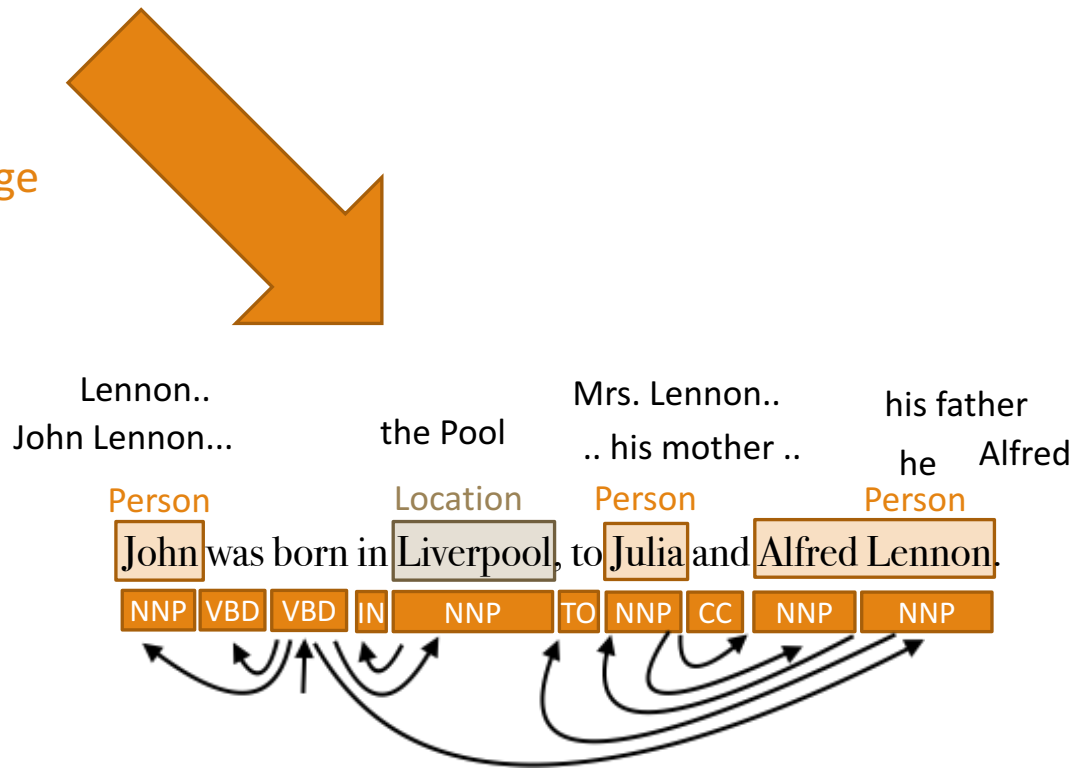
Literal Facts



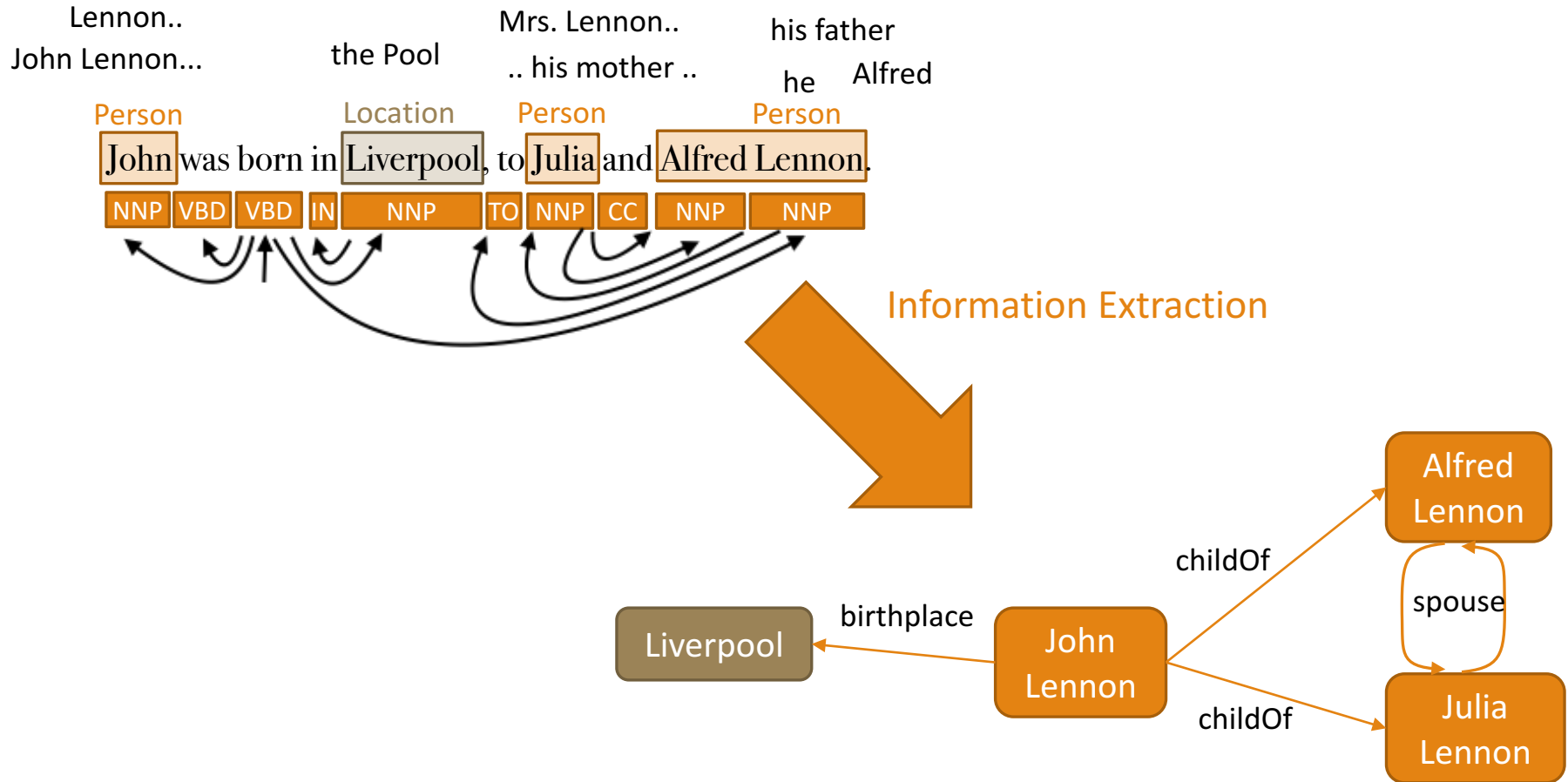
Role of NLP?

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

Natural Language
Processing



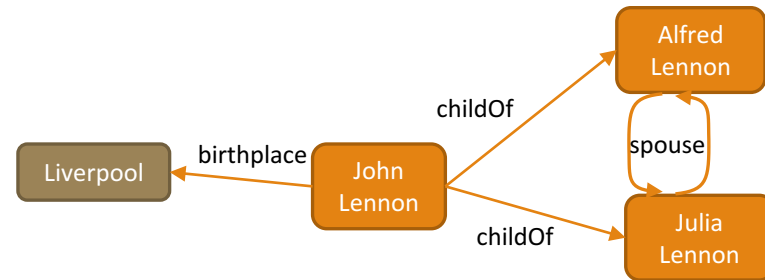
Information Extraction



Breaking it Down

Information
Extraction

Entity resolution,
Entity linking,
Relation extraction...



Document

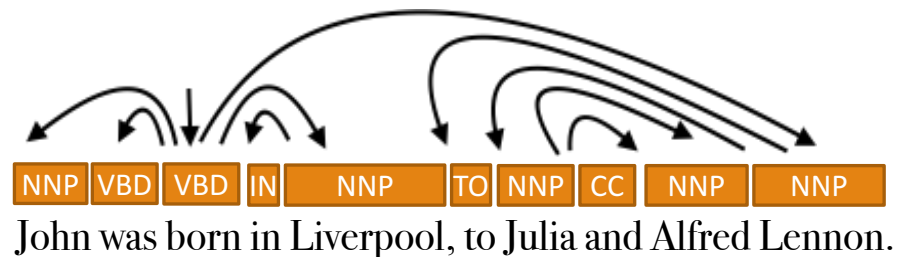
Coreference Resolution...

Lennon.. the Pool Mrs. Lennon.. his father
John Lennon... .. his mother .. 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...



Outline

What is Information Extraction

Named Entity Recognition

Homework 3

Relation Extraction

Named Entity Recognition

An important sub-task: find and classify names in text, for example:

- The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition

An important sub-task: **find** and classify names in text, for example:

- The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie**, **Rob Oakeshott**, **Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Named Entity Recognition

An important sub-task: **find** and **classify** names in text, for example:

- The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie**, **Rob Oakeshott**, **Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Person
Date
Location
**Organi-
zation**

Detecting Named Entities

Person Location Person Person
[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 Knowledge Extraction:

- 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.

NER: Entity Types

Fine-grained Types

person		organization	
actor	doctor	airline	terrorist_organization
architect	engineer	company	government_agency
artist	monarch	educational_institution	government
athlete	musician	fraternity_sorority	political_party
author	politician	sports_league	educational_department
coach	religious_leader	sports_team	military
director	soldier		news_agency
	terrorist		
location	body_of_water	product	art
city	island	camera	written_work
country	mountain	engine	film
county	glacier	airplane	newspaper
province	astral_body	car	play
railway	cemetery	ship	event
road	park	spacecraft	military_conflict
bridge		train	attack
			natural_disaster
			election
			sports_event
			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

Animals with Misleading Names

Electric Eel



Not an eel.

Mountain Goat



Not a goat.

Maned Wolf



Not a wolf.

King Cobra



Not a cobra. Also, snakes are typically self-governing.

Peacock Mantis Shrimp



Not a peacock.
Not a mantis.
Also, not a shrimp.

Horny Toad



Not a toad.
Only thinks of you as a friend.

Mayfly



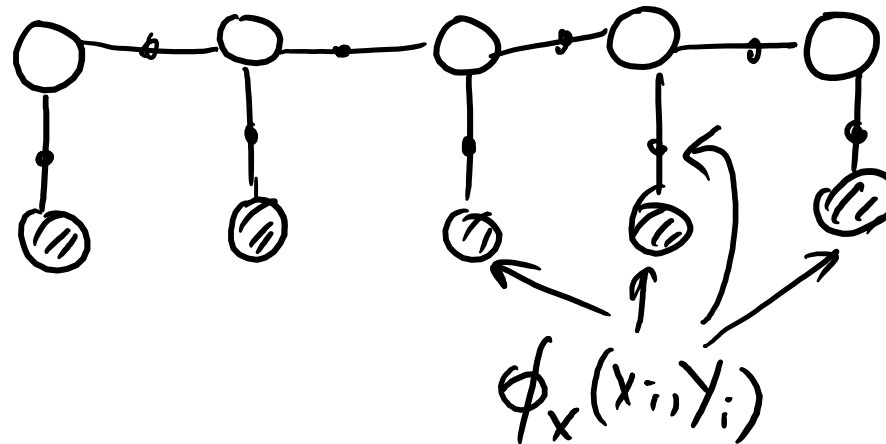
Active through the spring and summer.

Eastern Kingbird



Found in the West.
Many birds do not recognise its authority.

Sequence Labeling for NER

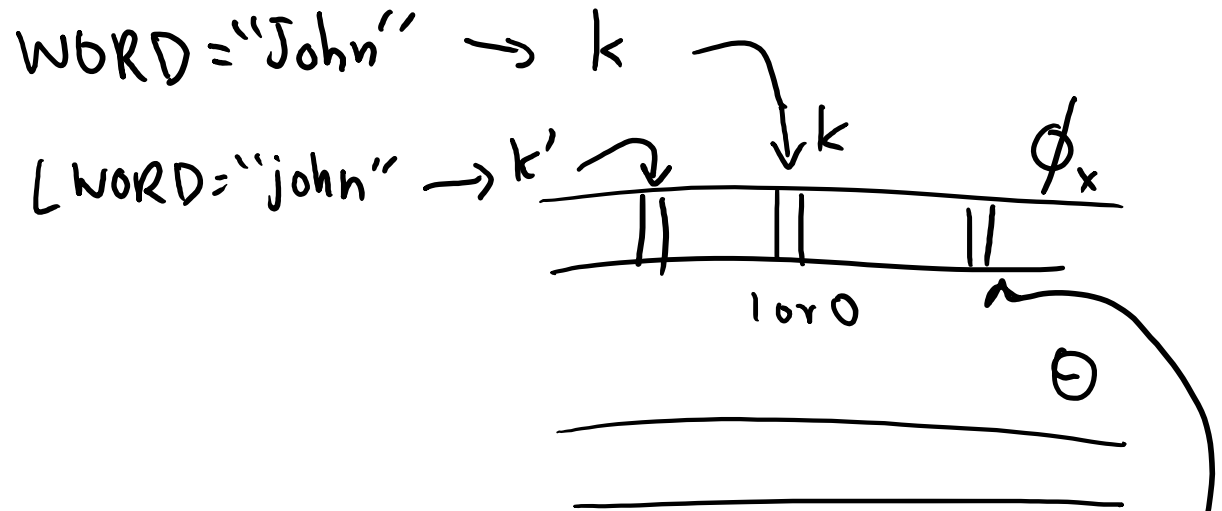


$$\underset{y}{\operatorname{argmax}} \quad \theta \cdot \vec{\phi}(\vec{x}, \vec{y})$$

$$\hookrightarrow \sum_{i=1}^n \phi(x_i, y_i, y_{i-1})$$

Features: Words and Lexicons

Words



Lexicons

People names \leftarrow US Census
"Appears in FName.100"
 $\hookrightarrow k$

movies, places, ...

Features: Prefixes/Suffixes

John

prefix 1 = "J"

prefix 2 = "Jo"

⋮
prefix 4 = "John"

suffix 1 = "h"

⋮

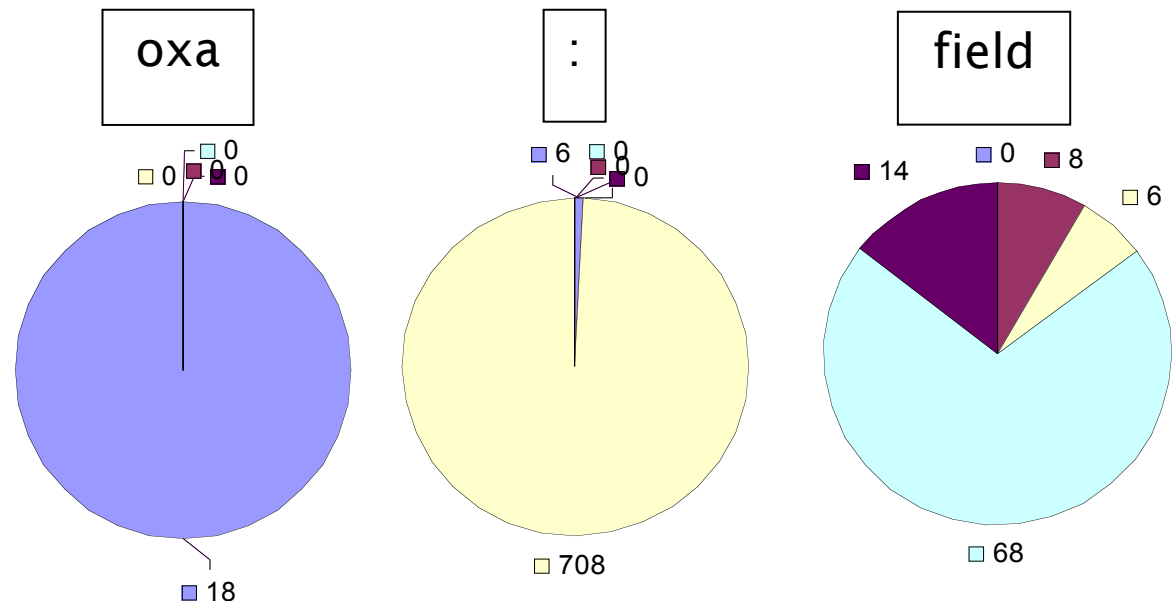
Features: Substrings of Words



Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion

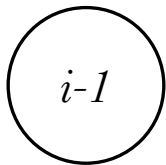


Features: Word Shapes

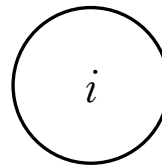
$Shape(c) =$	$\left\{ \begin{array}{ll} \text{if A-Z} & X \\ \text{if a-z} & x \\ \text{if 0-9} & d \\ \text{o.w.} & c \end{array} \right.$			
		John	DC-100	CamelCase
	Word shapes	Xxxx	XX-ddd	XxxxxXxxx
	Short shapes	Xx	X-d	XxXx

Features: Surrounding Context

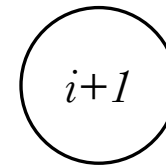
John



Deere



announced



NEXT_BIAS
NEXT_WORD=Deere
NEXT_LWORD=deere
NEXT_FIRSTCAP=True
NEXT_SSHAPE=Xx
NEXT_LEXICON=company
...

BIAS
WORD=Deere
LWORD=deere
FIRSTCAP=True
SSHAPE=Xx
LEXICON=company
...

PREV_BIAS
PREV_WORD=Deere
PREV_LWORD=deere
PREV_FIRSTCAP=True
PREV_SSHAPE=Xx
PREV_LEXICON=company
...

Outline

What is Information Extraction

Named Entity Recognition

Homework 3

Relation Extraction

Sequence Tagging on Twitter

Parts of Speech

What a productive day . Not .

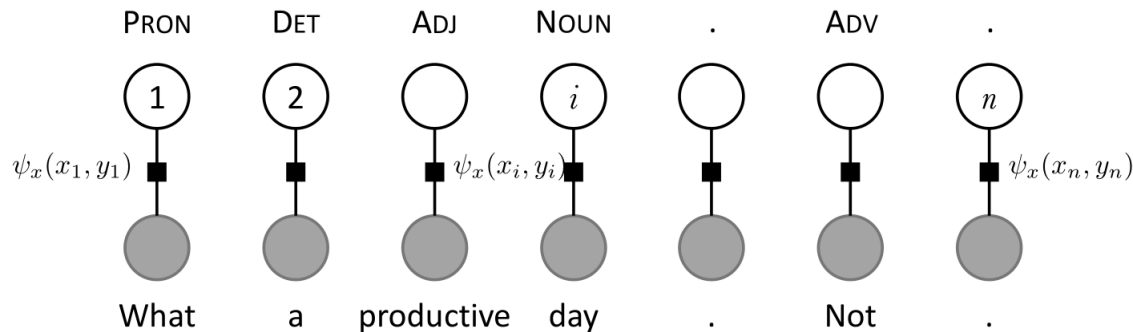
PRON DET ADJ NOUN . ADV .

Named Entity Recognition

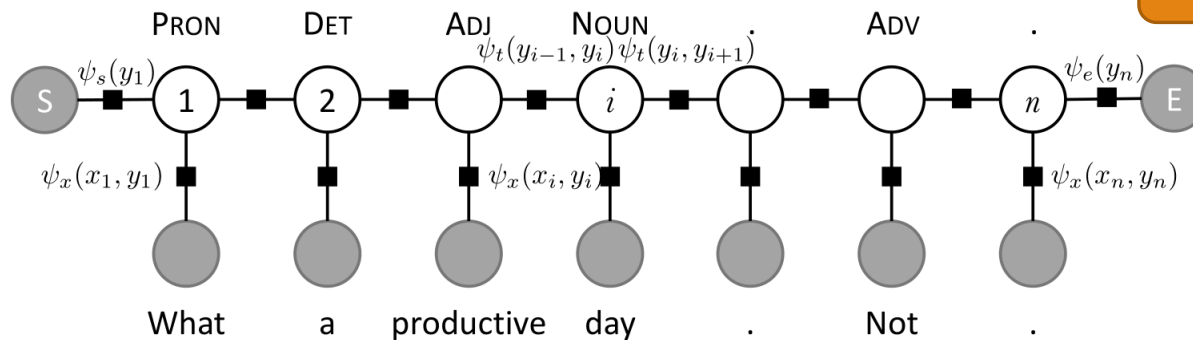
‘ Breaking Dawn ’ Returns to Vancouver on January 11th

O B-MOVIE I-MOVIE O O O B-GEO-LOC O O O

Sequence Tagging Models



Logistic Regression

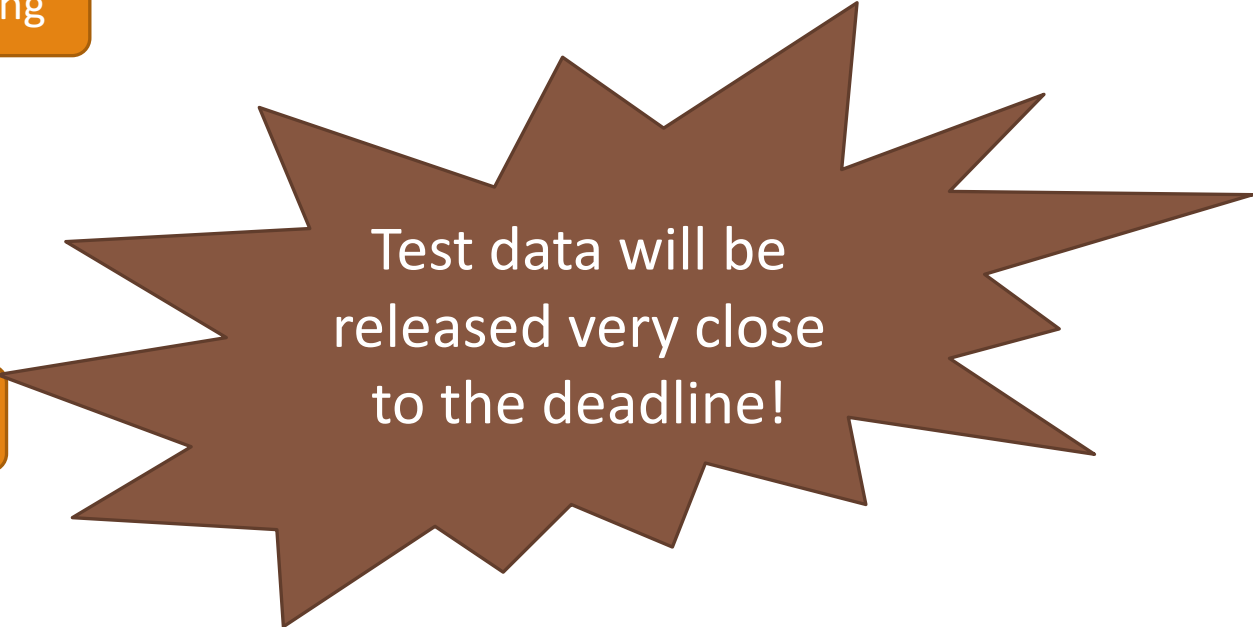


Conditional Random Fields

What do you have to do?

Feature Engineering

Viterbi Algorithm



Test data will be
released very close
to the deadline!

Upcoming...

Homework

- Homework 3 is due on **February 27**
- Write-up and data has been released.

Project

- Status report due in 1.5 weeks: **March 2, 2017**
- Instructions coming soon
- Only **5 pages**

Summaries

- Paper summaries: **February 28**, March 14
- Only **1 page** each