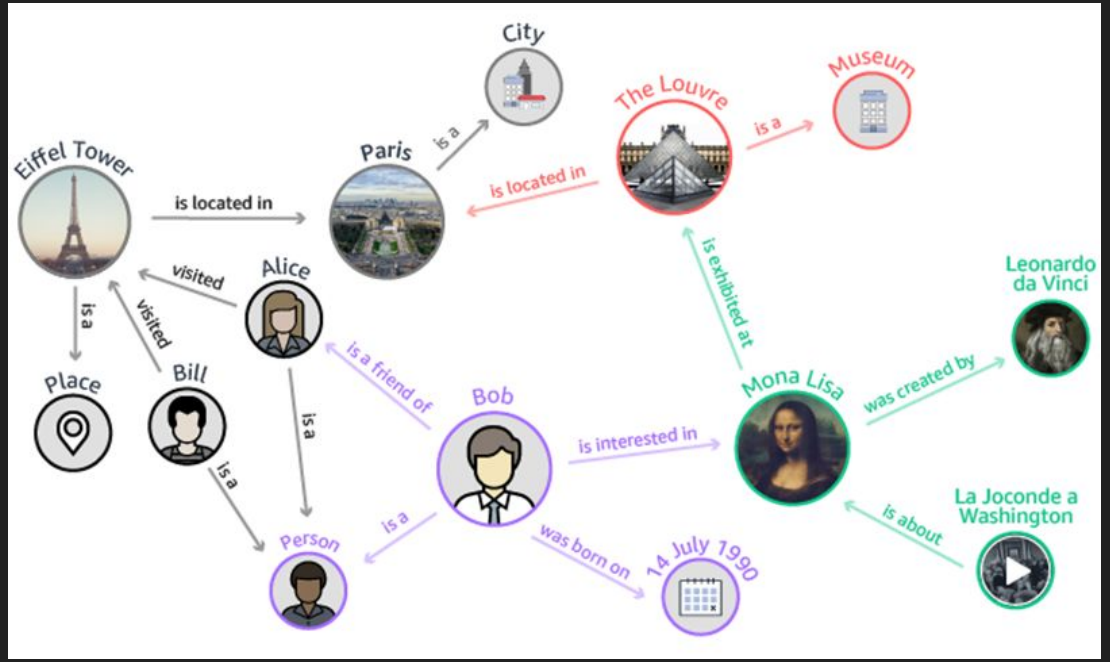
From Unstructured Text to Knowledge Graphs

Rajat Dange (rdange)
Rutvik Kolhe (rkolhe)
Varun Shah (vshah)

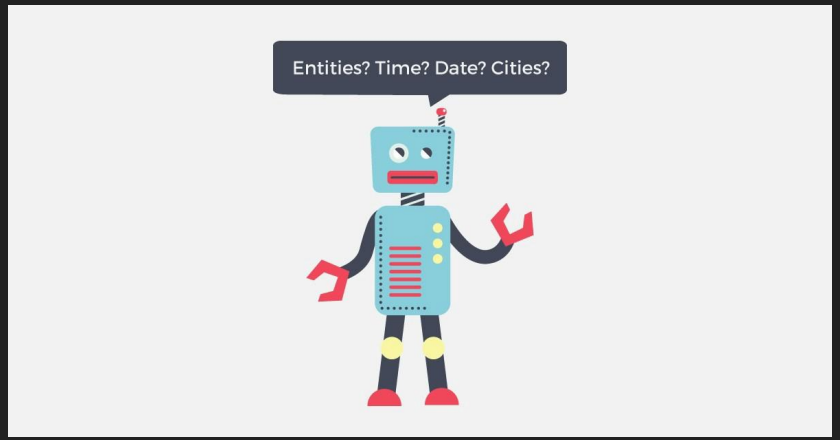
What are Knowledge Graphs?

- Models entities and relations between them
- A graph database
- Entities are nodes, relations are edges

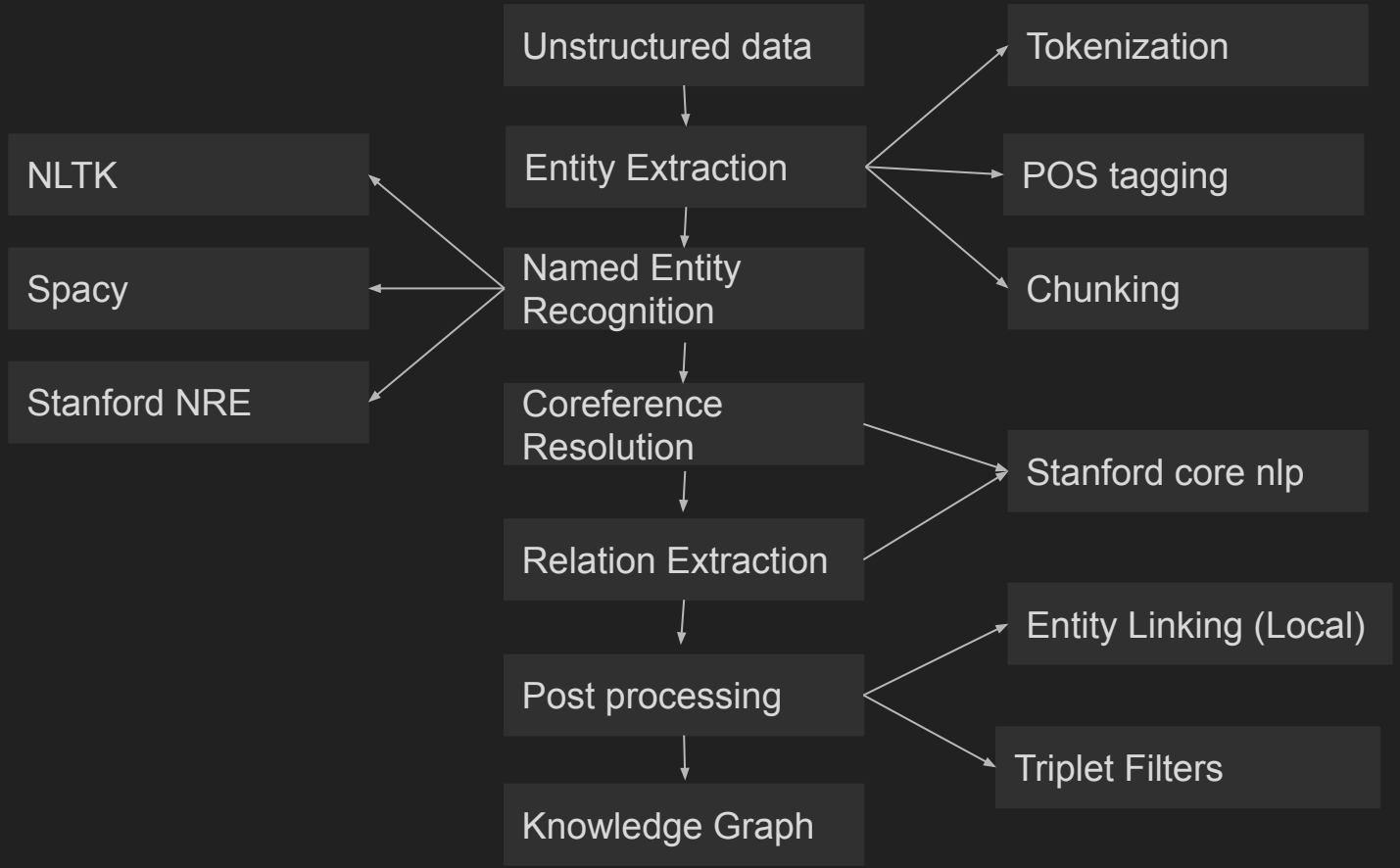


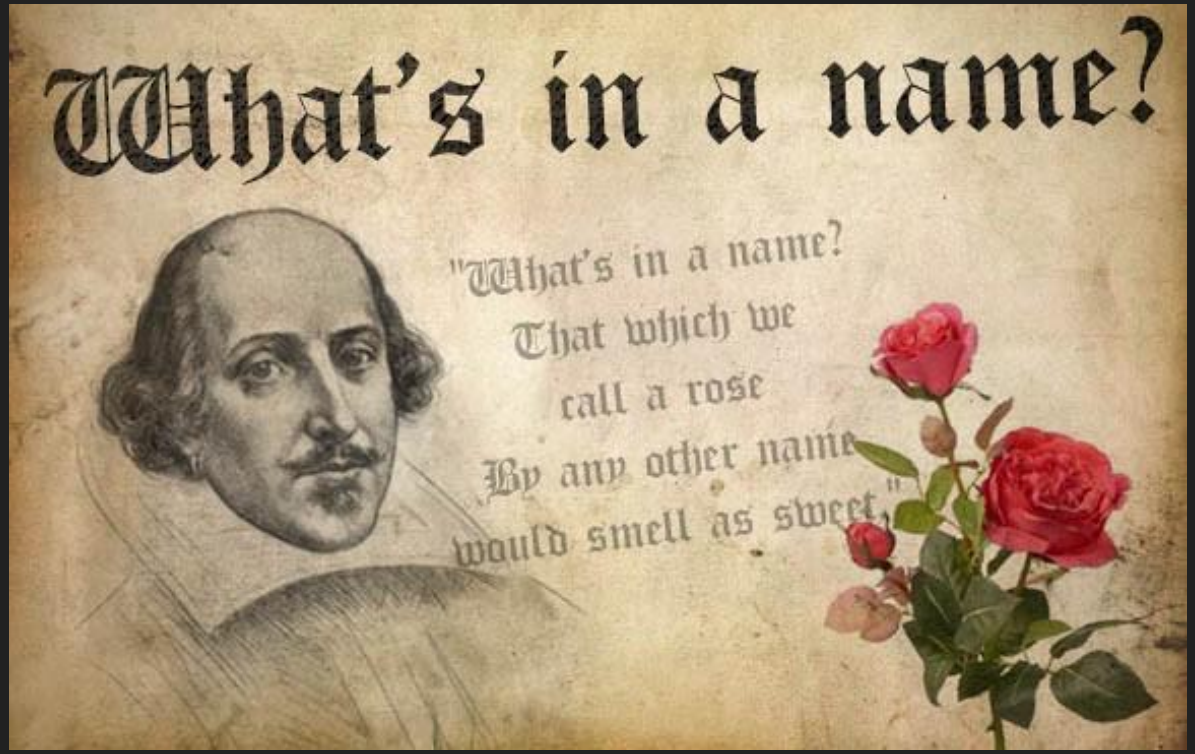
Objective

- To develop an end-to-end solution for building a knowledge graph from unstructured text using NLP tools
- **Input**: Unstructured Text from the internet
- **Output**: Visualization of structured text containing named entities and the relationship between them using graph

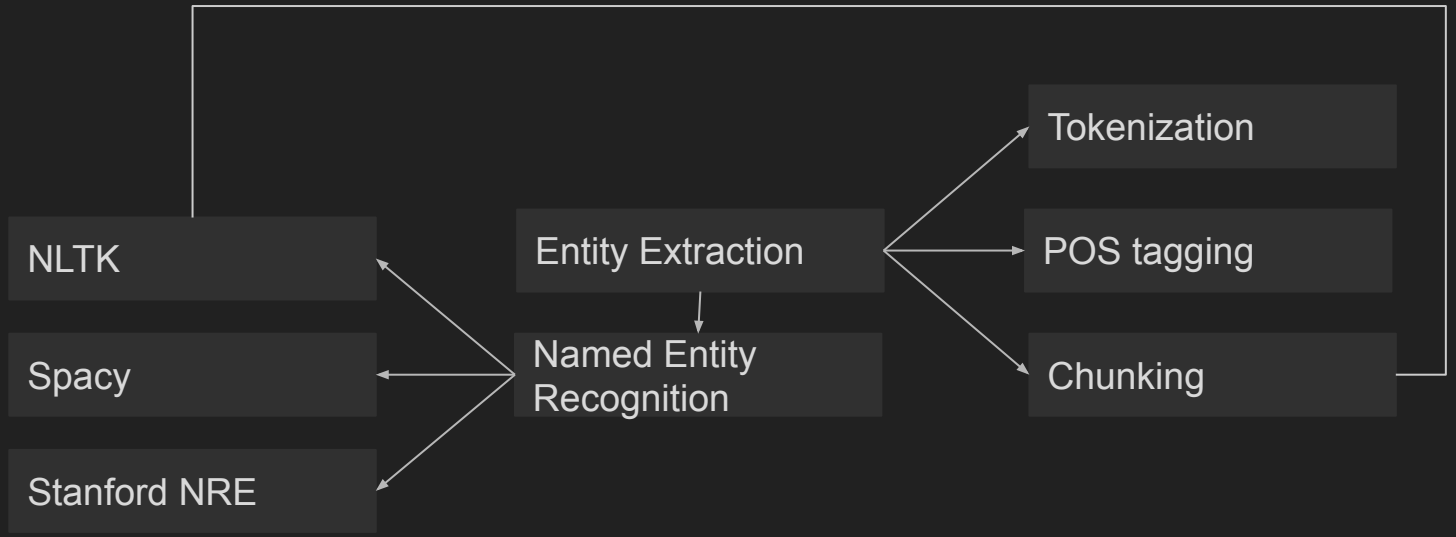


Process - Overview




What's in a name?
"What's in a name?
That which we
call a rose
By any other name
would smell as sweet."

Named Entity Recognition



Firing Mr. Strzok PERSON , however, removes a favorite target of Mr. Trump PERSON from the ranks of the F.B.I. GPE and gives Mr. Bowdich PERSON and the F.B.I. GPE director, Christopher A. Wray PERSON , a chance to move beyond the president's ire.

NER using NLTK

```
('film', 'NN'), ('directed', 'VBN'), ('by', 'IN'), Tree('PERSON', [('Francis', 'NNP'), ('Ford', 'NNP'), ('Coppola', 'NNP')]), ('and', 'CC'), ('pro
duced', 'VBN'), ('by', 'IN'), Tree('PERSON', [('Albert', 'NNP'), ('S.', 'NNP'), ('Ruddy', 'NNP')]), (',', ','), ('based', 'VBN'), ('on', 'IN'), Tr
```

NER using Stanford

```
... Running stanford for NER; this may take some time ...
[[('The', 'O'), ('Godfather', 'O'), ('is', 'O'), ('a', 'O'), ('1972', 'O'), ('American', 'O'), ('crime', 'O'), ('film', 'O'), ('directed', 'O'), (
'by', 'O'), ('Francis', 'PERSON'), ('Ford', 'PERSON'), ('Coppola', 'PERSON'), ('and', 'O'), ('produced', 'O'), ('by', 'O'), ('Albert', 'PERSON'),
('S.', 'PERSON'), ('Ruddy', 'PERSON'), (',', 'O'), ('based', 'O'), ('on', 'O'), ('Mario', 'PERSON'), ('Puzo', 'PERSON'), ("'s", 'O'), ('best-selli
```

NER using Spacy

```
using Spacy for NER

[('Godfather', 'ORG'), ('1972', 'DATE'), ('American', 'NORP'), ('Francis Ford Coppola', 'PERSON'), ('Albert S. Ruddy', 'PERSON'), ("Mario Puzo's",
 'PERSON'), ('Marlon Brando', 'PERSON'), ('Al Pacino', 'PERSON'), ('New York', 'GPE'), ('1945', 'DATE'), ('1955', 'DATE'), ('Vito Corleone (', 'PR
ODUCT'), ('Michael Corleone', 'PERSON'), ('Pacino', 'NORP'), ('1972[5', 'CARDINAL')]
```

Coreference Resolution

What?

- “D. Trump is president of United States. He was elected to office in 2016”
- Resolving: He ≡ Trump

Why?

- Most text reference main entities only once
- Different nodes in relation extraction will render that aspect unusable

How?

- Stanford Core NLP gives related word phrases
- Identify the coreference to be replaced for each list
    - Uses recognized Named Entities
- Determine which entity needs to be replaced in the sentence
- Replace if replacement is valid.

Coreference Resolution - output

```
Original:  The story, spanning 1945 to 1955, chronicles the family under the patriarch Vito Corleone (Brando), focusing on the transformation of M
ichael Corleone (Pacino) from reluctant family outsider to ruthless mafia boss.
To replace: the patriarch Vito Corleone -LRB- Brando -RRB-  | at: 13 20 With:  Michael Corleone
Result:  The story , spanning 1945 to 1955 , chronicles the family under Michael Corleone , focusing on the transformation of Michael Corleone ( P
acino ) from reluctant family outsider to ruthless mafia boss .


Original:  The story, spanning 1945 to 1955, chronicles the family under the patriarch Vito Corleone (Brando), focusing on the transformation of M
ichael Corleone (Pacino) from reluctant family outsider to ruthless mafia boss.
To replace: Brando  | at: 18 19 With:  Marlon Brando
Result:  The story , spanning 1945 to 1955 , chronicles the family under the patriarch Vito Corleone ( Marlon Brando ) , focusing on the transform
ation of Michael Corleone ( Pacino ) from reluctant family outsider to ruthless mafia boss .


Original:  The film was the highest-grossing film of 1972[5] and was for a time the highest-grossing film ever made.
To replace: the highest-grossing film  | at: 17 20 With:  The film
Result:  The film was the highest-grossing film of 1972 [ 5 ] and was for a time The film ever made .
```

Relation Extraction

- Relation Extraction lies in the center of our pipeline
- Here we form raw triplets of the form --> “Entity-1, Relation, Entity-2”.

- To achieve this we are using python wrapper for Stanford OpenIE and saving the output as a csv file
- As you can see below, the output is redundant and raw, so we do some post processing

Godfather	is	1972 American crime film directed by Francis Ford Coppola
Godfather	is	1972 American crime film
Godfather	is	1972 American crime film directed
Godfather	is	1972 crime film
Godfather	is	1972 crime film directed
Mario Puzo	on	best-selling novel of same name
Godfather	is	1972 crime film directed by Francis Ford Coppola
Godfather	is	American
It	stars	Marlon Brando
It	stars Marlon Brando as	leaders of fictional New York crime family
It	stars Marlon Brando as	leaders
story	focusing on	transformation from reluctant family outsider to ruthless mafia boss
story	chronicles family under	patriarch Vito Corleone
story	focusing on	transformation of Michael Corleone from reluctant family outsider to ruthless mafia boss
story	focusing on	transformation of Michael Corleone from reluctant family outsider

Post Processing

- The format of the expected structured output (shown below) for graph visualization is different from what is obtained from Relation Extraction

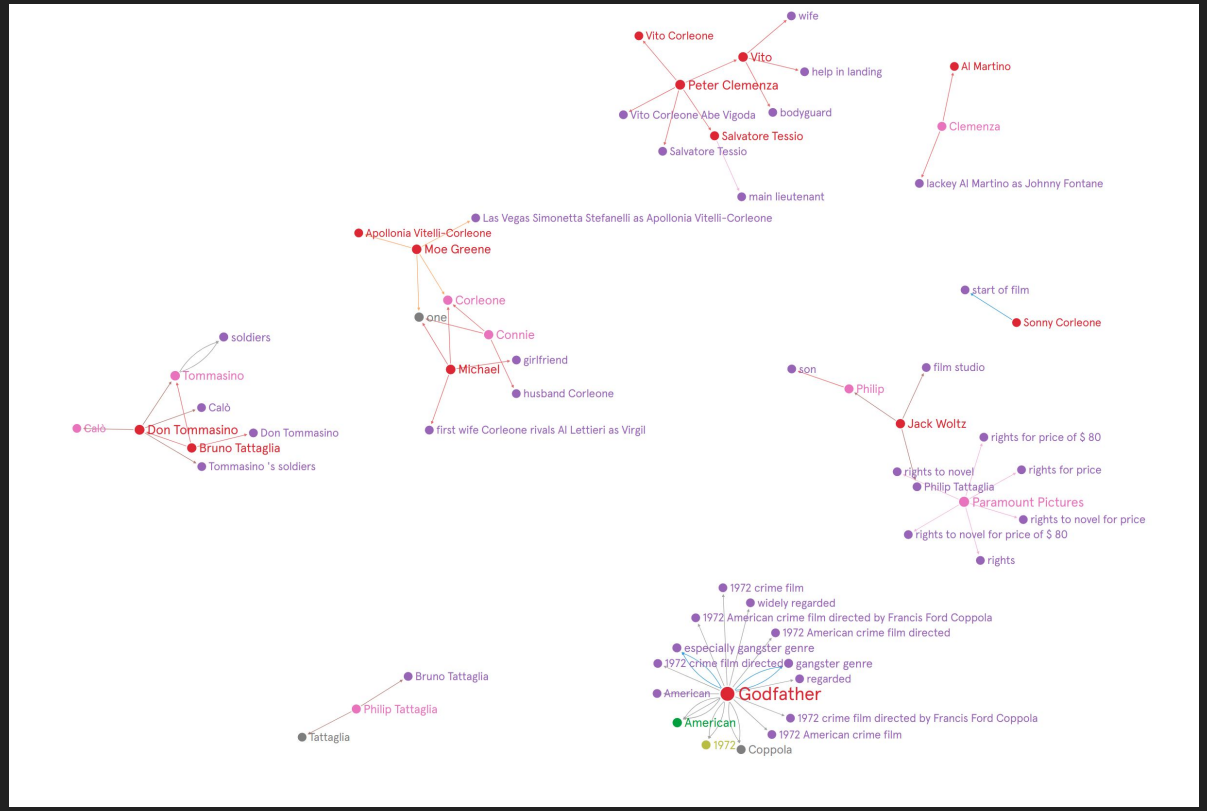
```
FromType, FromName, Edge, ToType, ToName
Person, John Doe, FRIENDS, Person, Emily
Person, John Doe, MEMBER OF, Organization, Acme
```

- FromName and ToName represents *entity1* and *entity2* respectively
- Edge represents *relationship* between the two named entities

Post processing output

PERSON	Godfather	is	GPE	1972 crime film directed by Francis Ford Coppola
ORG	Paramount Pictures	obtained	O	rights for price of $ 80
PERSON	Godfather	is	NORP	1972 American crime film directed by Francis Ford Coppola
PERSON	Godfather	is	DATE	1972 crime film directed by Francis Ford Coppola
PERSON	Jack Woltz	Rendina as	ORG	Philip Tattaglia
PERSON	Godfather	is	DATE	1972 American crime film directed
PERSON	Godfather	is	GPE	1972 American crime film directed by Francis Ford Coppola
ORG	Connie	has	ORG	husband Corleone

Results https://graphcommons.com/graphs/df34b339-b7ef-4a25-a3f6-63e431d4644c



References

- Entity Relation Extraction (http://ceur-ws.org/Vol-1733/paper-10.pdf)
- Current approaches for Relation Extraction (http://nlpprogress.com/english/relationship_extraction.html)
- Mining knowledge graphs from text (https://kgtutorial.github.io/)
- NER recognition tutorial (https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da)
- Relation extraction using NLTK (https://stackoverflow.com/questions/7851937/extract-relationships-using-nltk)
- NER using NLTK (https://www.nltk.org/book/ch07.html)
- Stanford NER (https://nlp.stanford.edu/software/CRF-NER.shtml)
- NLTK - Stanford wrapper (http://www.nltk.org/api/nltk.tag.html#module-nltk.tag.stanford)
- Coreference resolution (https://stackoverflow.com/questions/39410282/coreference-resolution-in-python-nltk-using-stanford-corenlp)
- Python Wrapper for Coreference Resolution (https://github.com/dasmith/stanford-corenlp-python)
- Spacy (https://spacy.io/usage/spacy-101)
- Graph commons (https://graphcommons.com/)