Named Entity Disambiguation Using Graphical Approaches

•••

Brian Lin
Cory Williams
Matteo Zhang
Shane Ong

Who is Kensho?

Technology company under S&P Global

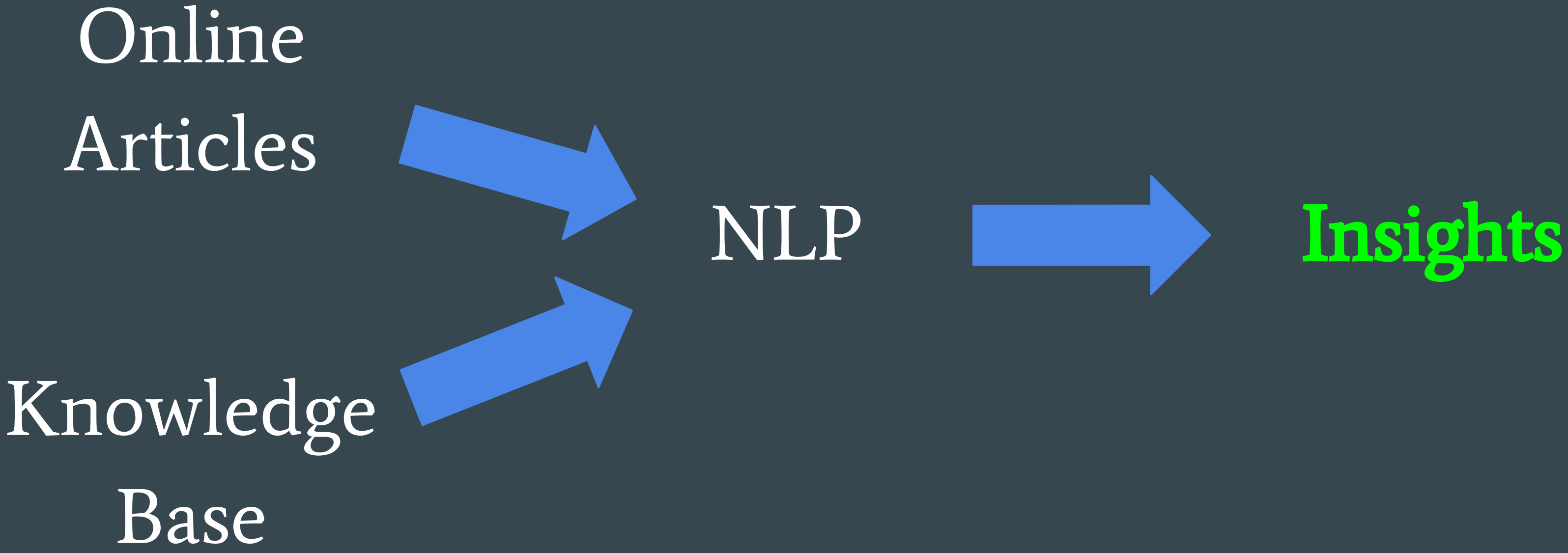
Products

- Data analytics tools/platforms
- Machine Learning algorithms (AI Lab)
- **Insights**

Insights

- Financial market
- Healthcare
- Societal trends
- And much more!

Motivation for our project



Online Articles

Knowledge Base

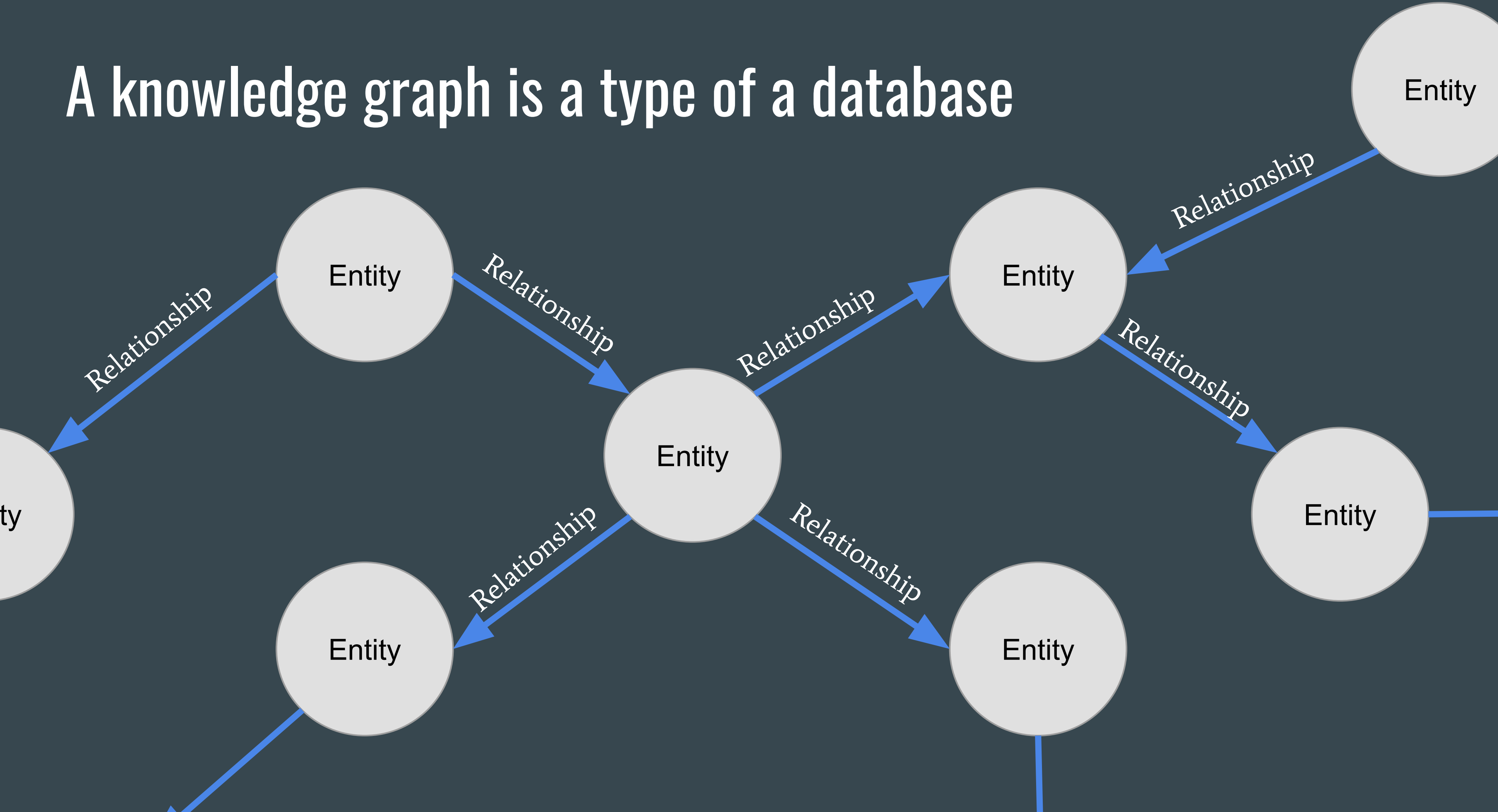
NLP

Insights

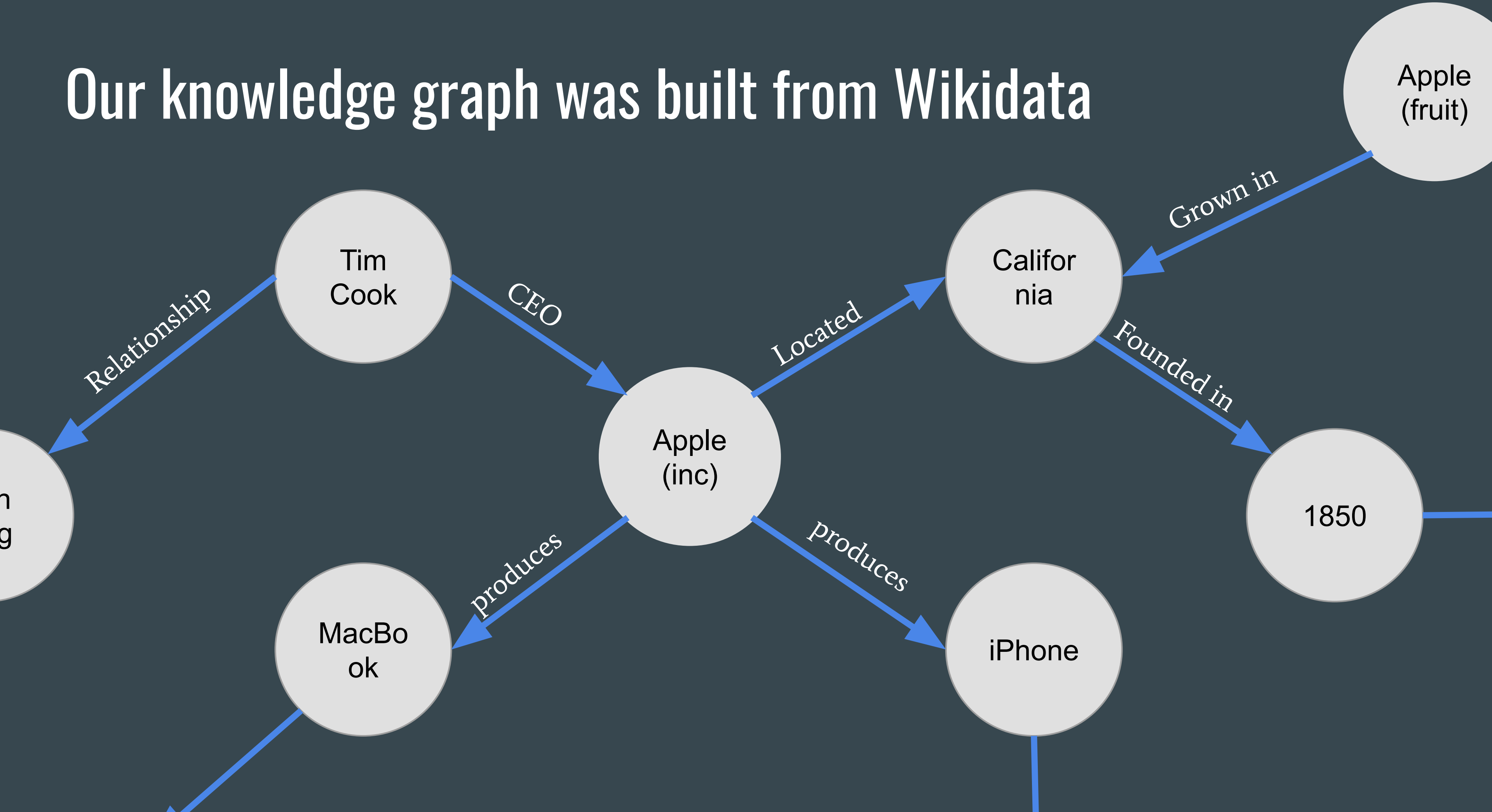
Named Entity Disambiguation (NED) is an NLP task

“Apple shares suffer worst week of 2019 as investors fear China trade turmoil threatens iPhone growth”[1]

A knowledge graph is a type of a database

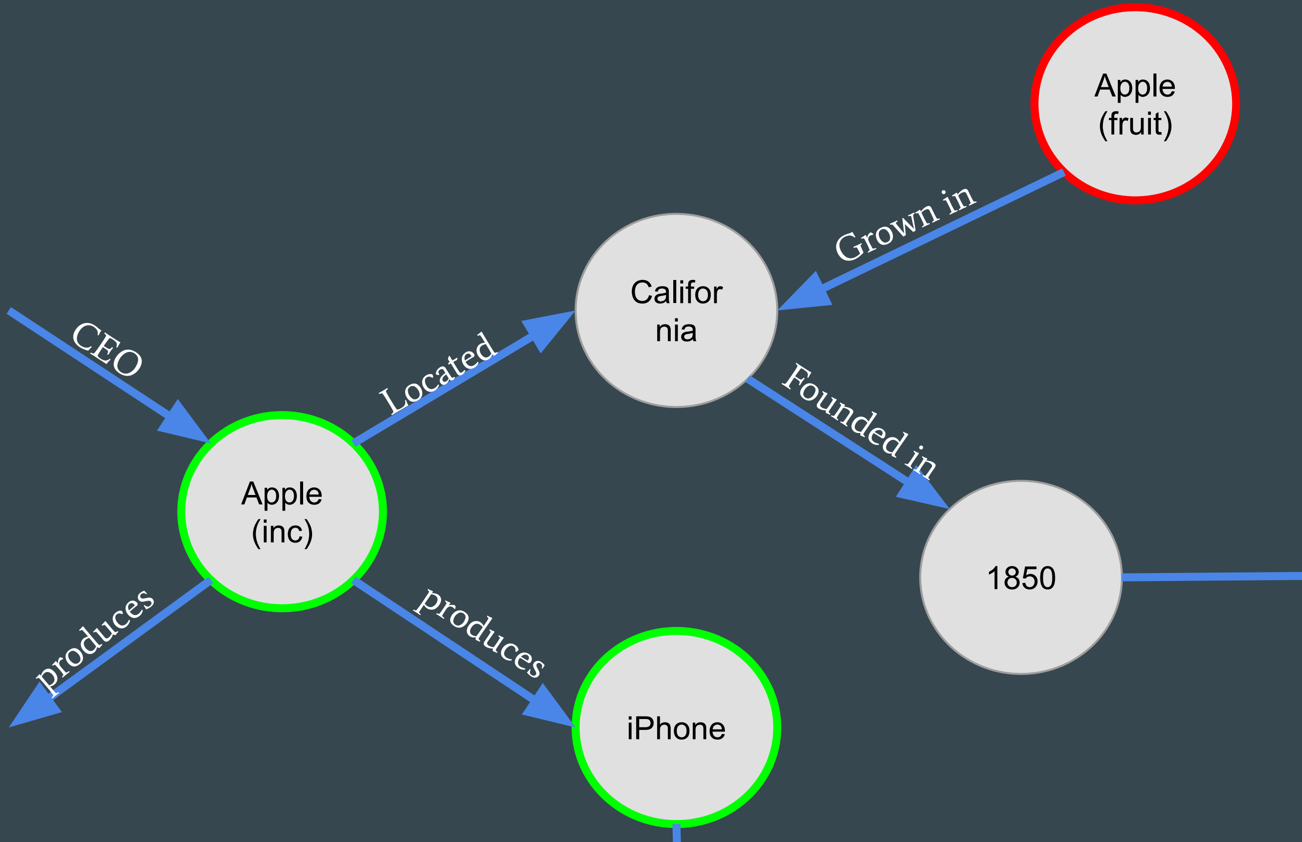


Our knowledge graph was built from Wikidata



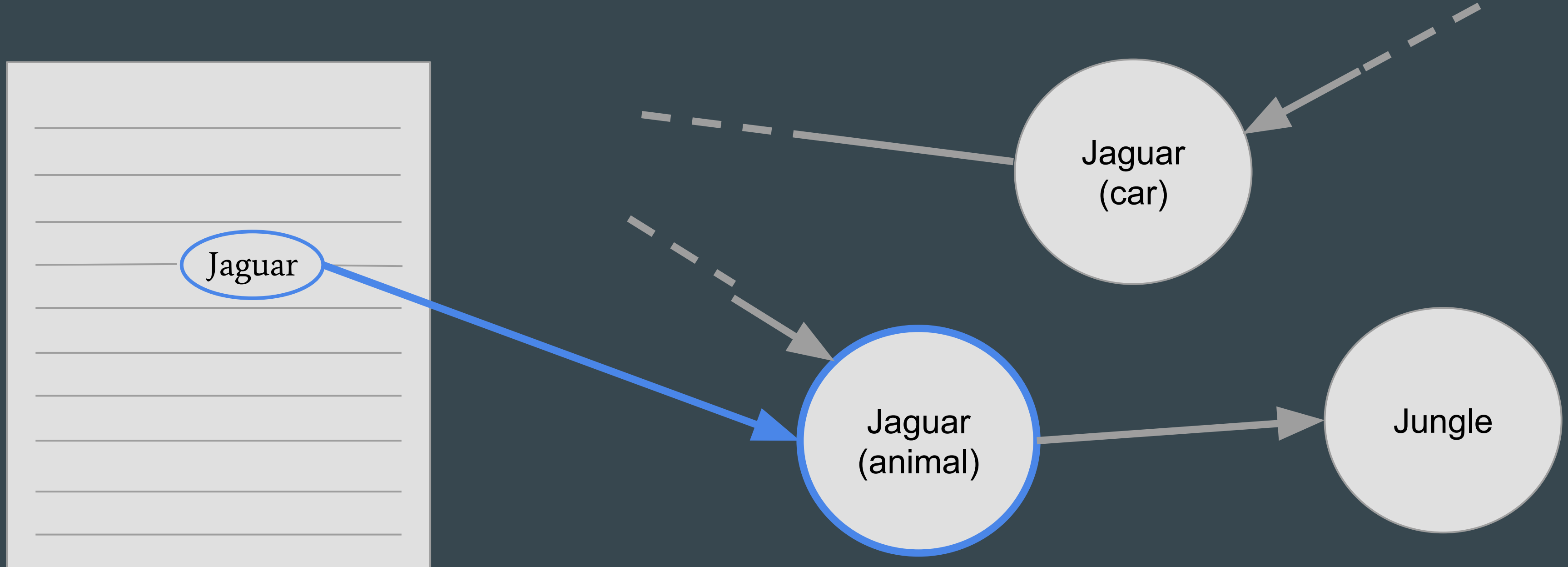
We can use the knowledge graph for NED

“Apple shares suffer worst week of 2019 as investors fear China trade turmoil threatens iPhone growth”[1]



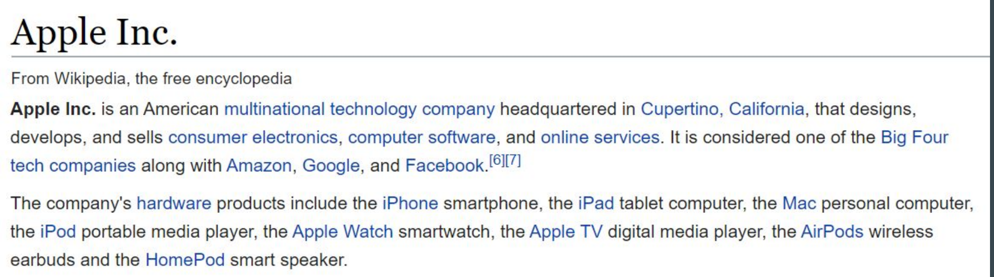
Problem Statement

Develop an algorithm to perform named entity disambiguation by using graphical methods, as well as merging graphs with deep learning methods.



Data

- **Wikipedia** is the encyclopedia, **Wikidata** is the knowledge base

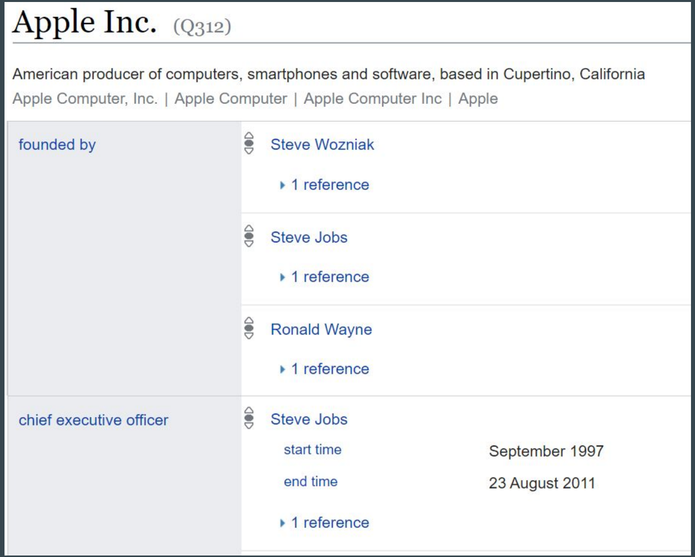


Apple Inc.

From Wikipedia, the free encyclopedia

Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, and online services. It is considered one of the Big Four tech companies along with Amazon, Google, and Facebook.[6][7]

The company's hardware products include the iPhone smartphone, the iPad tablet computer, the Mac personal computer, the iPod portable media player, the Apple Watch smartwatch, the Apple TV digital media player, the AirPods wireless earbuds and the HomePod smart speaker.



Apple Inc. (Q312)

American producer of computers, smartphones and software, based in Cupertino, California

Apple Computer, Inc. | Apple Computer | Apple Computer Inc | Apple

founded by

Steve Wozniak

▸ 1 reference

Steve Jobs

▸ 1 reference

Ronald Wayne

▸ 1 reference

chief executive officer

Steve Jobs

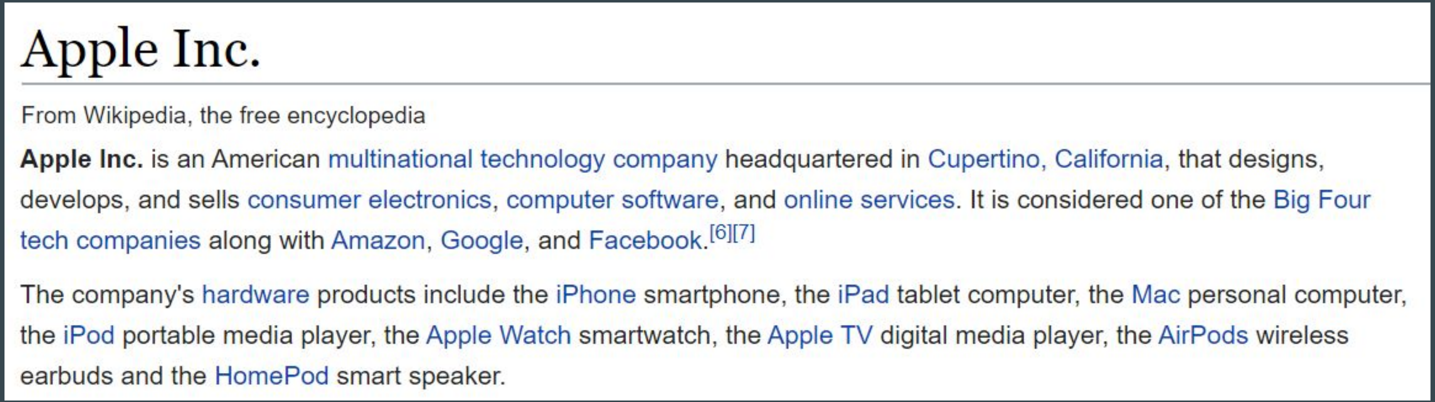
start time September 1997

end time 23 August 2011

▸ 1 reference

Data

- **Wikipedia data**: text, link anchors, and redirects
  - 5.9 million unique pages
  - 135 million links to other Wikipedia pages
  - 7.8 million unique anchor text and link pairs



Apple Inc.

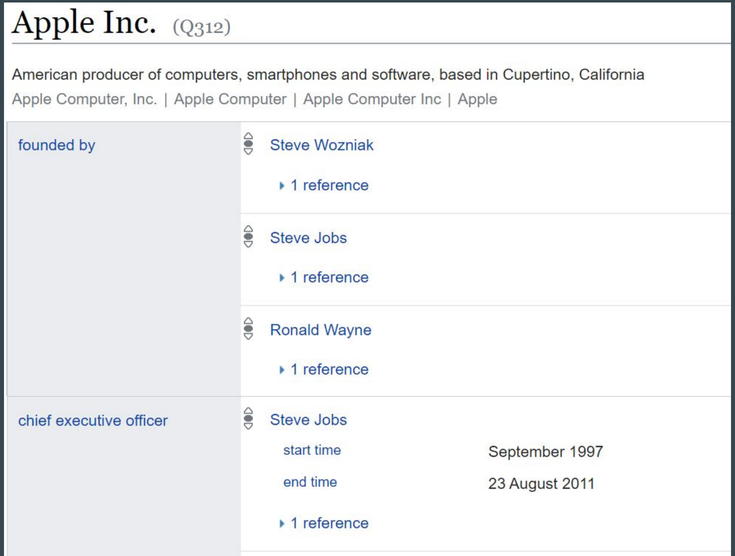
From Wikipedia, the free encyclopedia

Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, and online services. It is considered one of the Big Four tech companies along with Amazon, Google, and Facebook.[6][7]

The company's hardware products include the iPhone smartphone, the iPad tablet computer, the Mac personal computer, the iPod portable media player, the Apple Watch smartwatch, the Apple TV digital media player, the AirPods wireless earbuds and the HomePod smart speaker.

Data

- **Wikidata data:** triplets (nodes, edges, targets) ⟷ (entity 1, relation, entity 2)
  - 383.2 million triplets
  - 58.7 million entities
  - 6490 types of relations



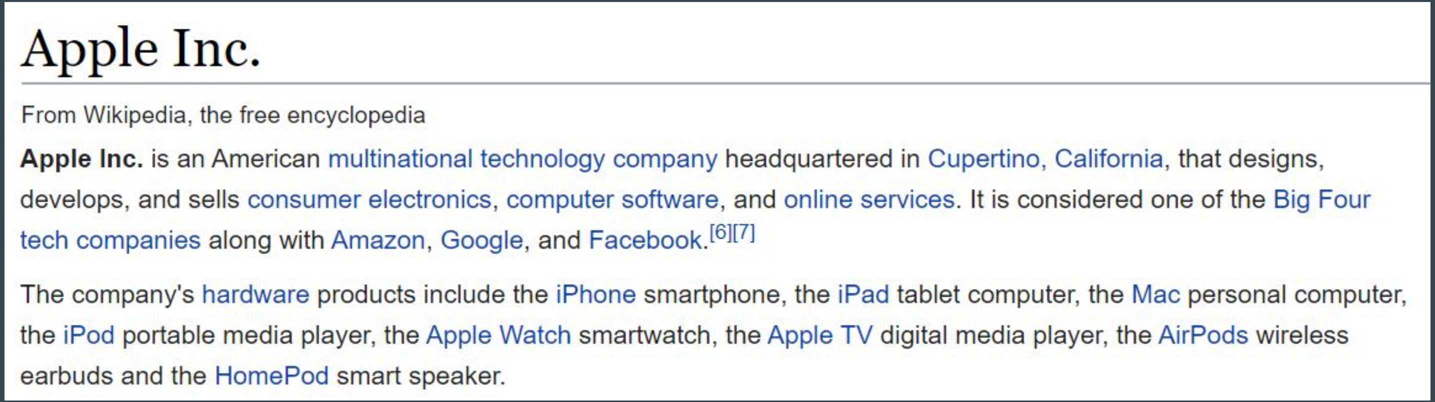
Apple Inc. (Q312)

American producer of computers, smartphones and software, based in Cupertino, California

Apple Computer, Inc. | Apple Computer | Apple Computer Inc | Apple

founded by	Steve Wozniak ▸ 1 reference Steve Jobs ▸ 1 reference Ronald Wayne ▸ 1 reference
chief executive officer	Steve Jobs start time September 1997 end time 23 August 2011 ▸ 1 reference

Data

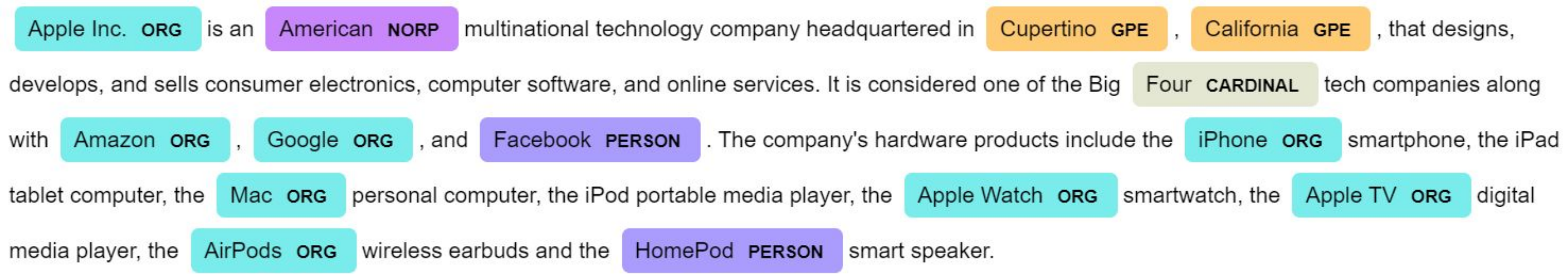


Apple Inc.

From Wikipedia, the free encyclopedia

Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, and online services. It is considered one of the Big Four tech companies along with Amazon, Google, and Facebook.[6][7]

The company's hardware products include the iPhone smartphone, the iPad tablet computer, the Mac personal computer, the iPod portable media player, the Apple Watch smartwatch, the Apple TV digital media player, the AirPods wireless earbuds and the HomePod smart speaker.



Apple Inc. **ORG** is an American **NORP** multinational technology company headquartered in Cupertino **GPE** , California **GPE** , that designs, develops, and sells consumer electronics, computer software, and online services. It is considered one of the Big Four **CARDINAL** tech companies along with Amazon **ORG** , Google **ORG** , and Facebook **PERSON** . The company's hardware products include the iPhone **ORG** smartphone, the iPad tablet computer, the Mac **ORG** personal computer, the iPod portable media player, the Apple Watch **ORG** smartwatch, the Apple TV **ORG** digital media player, the AirPods **ORG** wireless earbuds and the HomePod **PERSON** smart speaker.

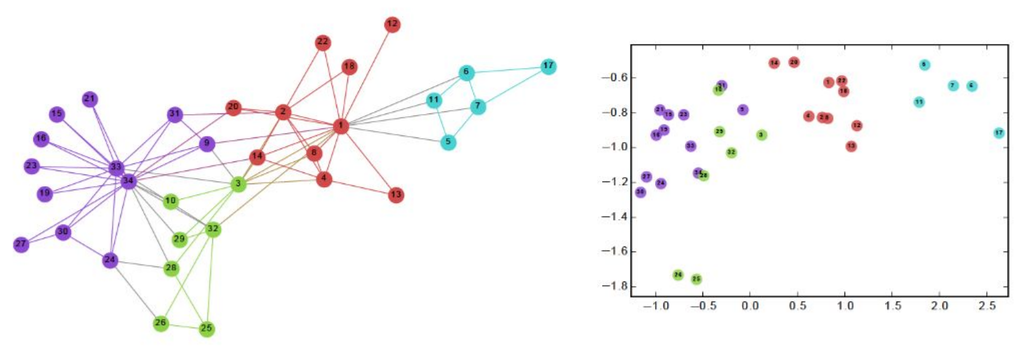
Data Preprocessing

- Word Embeddings
  - Vector representations of words that capture semantic meaning within a document
  - word2vec, doc2vec
- Graph Embeddings
  - Vector representations of nodes/graphs that capture the structure and topology of a graph
  - node2vec, graph2vec

Conceptually, a sentence is essentially just a more restricted graph structure.

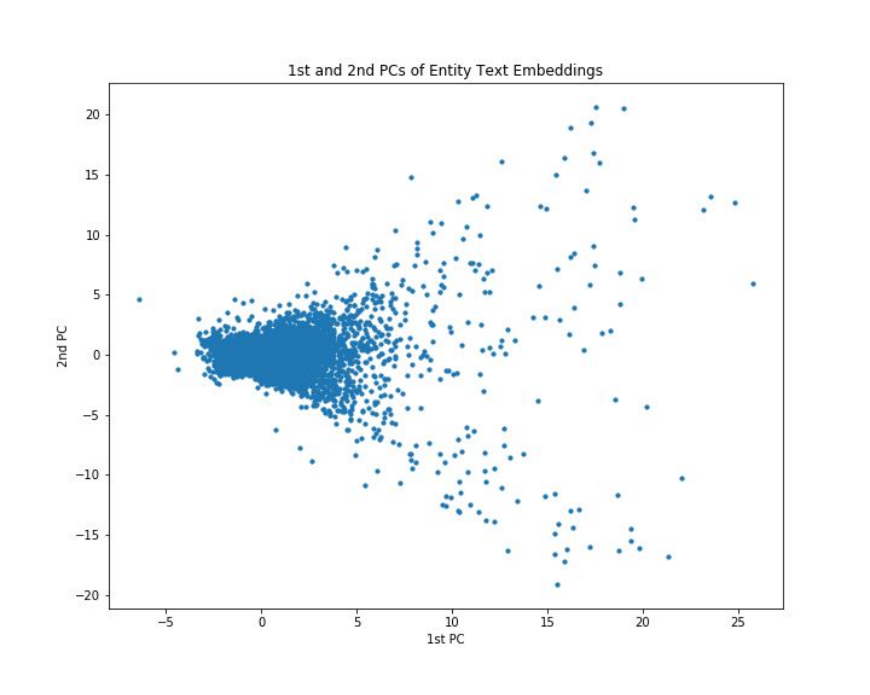
Data Preprocessing

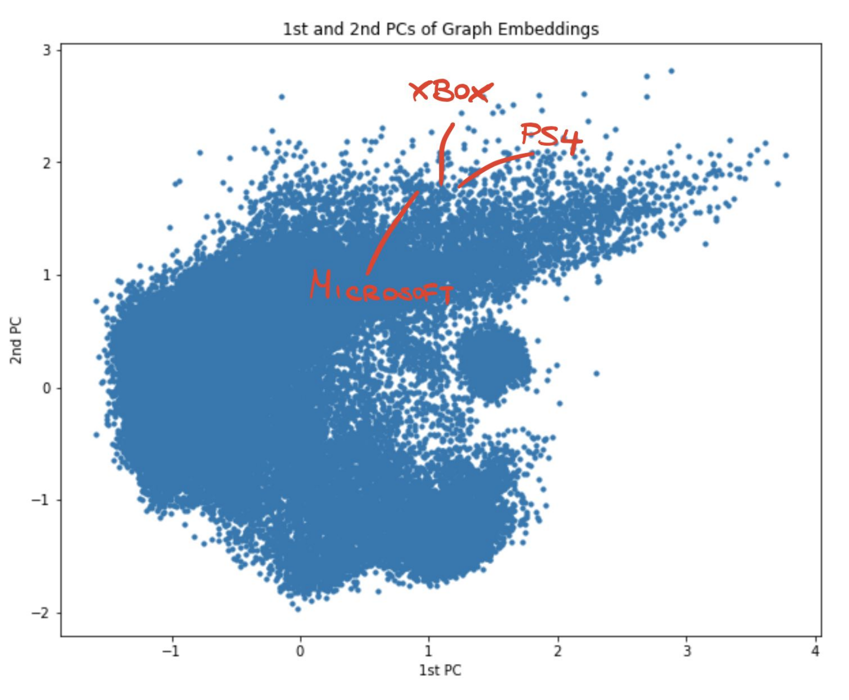
- Node2vec
  - Uses random walk to traverse a graph to learn embeddings
  - Each node is represented by a vector
  - Allows balance between trade-off of breadth-first-search (local) AND depth-first-search (global)



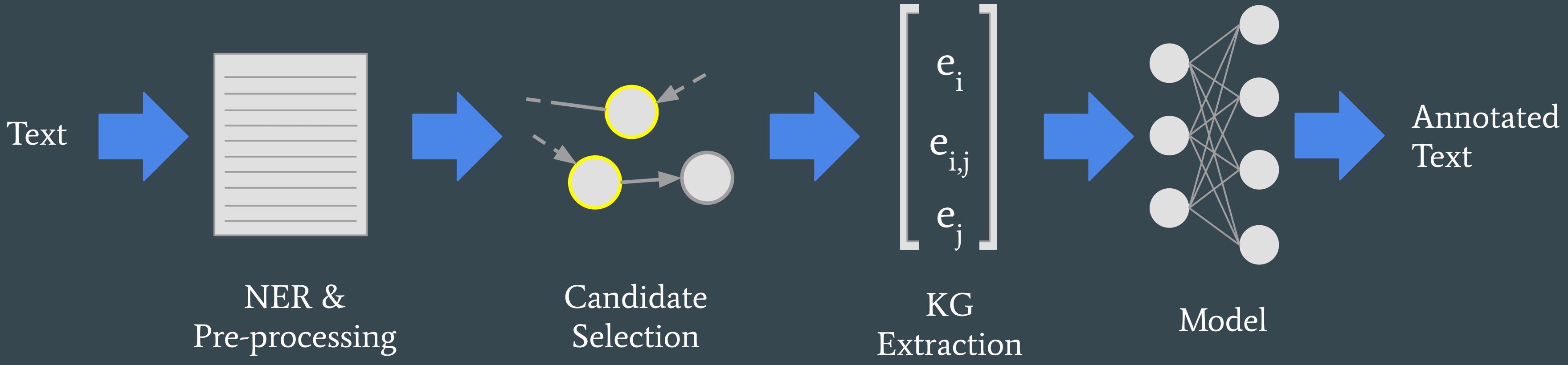
The original graph and it's embedding (Courtesy of the DeepWalk research team)

Data Exploration





Overall Pipeline (Deliverable)



Text

NER & Pre-processing

Candidate Selection

$$\begin{bmatrix} e_i \\ e_{i,j} \\ e_j \end{bmatrix}$$

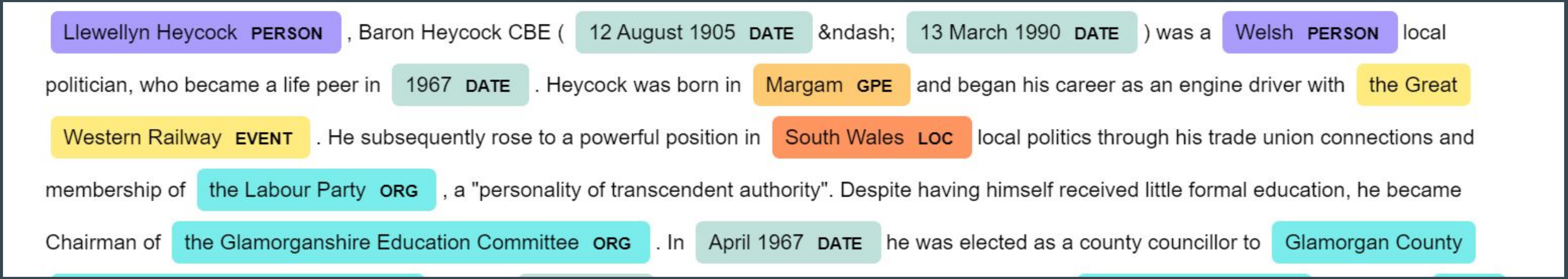
KG Extraction

Model

Annotated Text

NER and Preprocessing

- For text, extract all NER identified entities (SpaCy), locations within text
- Sample: 5000 Wikipedia article introductions
- True labels of entities are existing Wikipedia link anchors

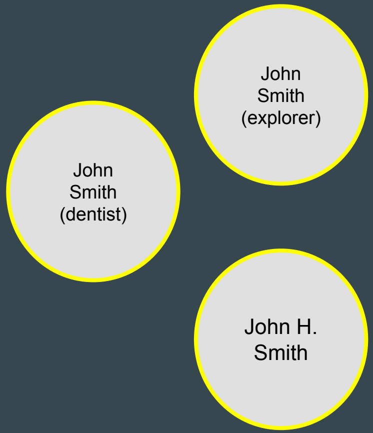


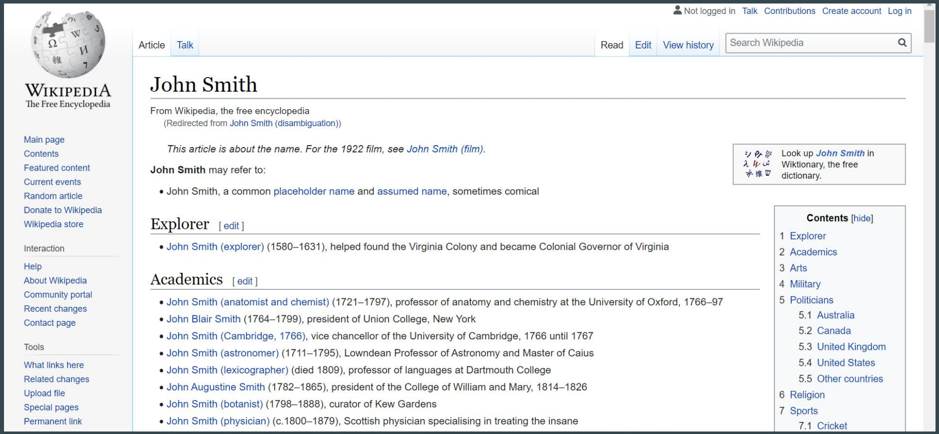
Llewellyn Heycock PERSON , Baron Heycock CBE (12 August 1905 DATE &ndash; 13 March 1990 DATE) was a Welsh PERSON local politician, who became a life peer in 1967 DATE . Heycock was born in Margam GPE and began his career as an engine driver with the Great Western Railway EVENT . He subsequently rose to a powerful position in South Wales LOC local politics through his trade union connections and membership of the Labour Party ORG , a "personality of transcendent authority". Despite having himself received little formal education, he became Chairman of the Glamorganshire Education Committee ORG . In April 1967 DATE he was elected as a county councillor to Glamorgan County

Candidate Selection

- For each identified entity, run a text similarity matching with Wikidata items
- Select a candidate group of possible entities for disambiguation

… the renowned John Smith Jr. …





John Smith

From Wikipedia, the free encyclopedia
(Redirected from John Smith (disambiguation))

This article is about the name. For the 1922 film, see John Smith (film).

Look up ***John Smith*** in Wiktionary, the free dictionary.

John Smith may refer to:

- John Smith, a common placeholder name and assumed name, sometimes comical

Explorer [edit]

- John Smith (explorer) (1580–1631), helped found the Virginia Colony and became Colonial Governor of Virginia

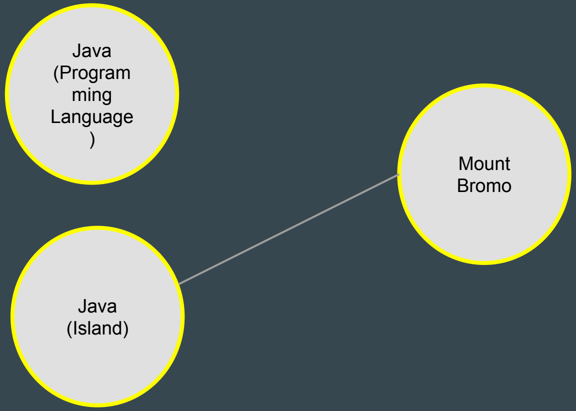
Academics [edit]

- John Smith (anatomist and chemist) (1721–1797), professor of anatomy and chemistry at the University of Oxford, 1766–97
- John Blair Smith (1764–1799), president of Union College, New York
- John Smith (Cambridge, 1766), vice chancellor of the University of Cambridge, 1766 until 1767
- John Smith (astronomer) (1711–1795), Lowndean Professor of Astronomy and Master of Caius
- John Smith (lexicographer) (died 1809), professor of languages at Dartmouth College
- John Augustine Smith (1782–1865), president of the College of William and Mary, 1814–1826
- John Smith (botanist) (1798–1888), curator of Kew Gardens
- John Smith (physician) (c.1800–1879), Scottish physician specialising in treating the insane

Knowledge Graph Extraction & Base Model

- Find all triplets that involve candidate Wikidata items
- Use centrality measures to determine best candidate for each entity
    - Choose candidate with highest centrality for each entity
    - For degree centrality, only look at direct relations between candidate Wikidata items

"Mount Bromo is one of Java's most popular tourist attractions."



Base Model Results

Sentence: This apple tastes great!

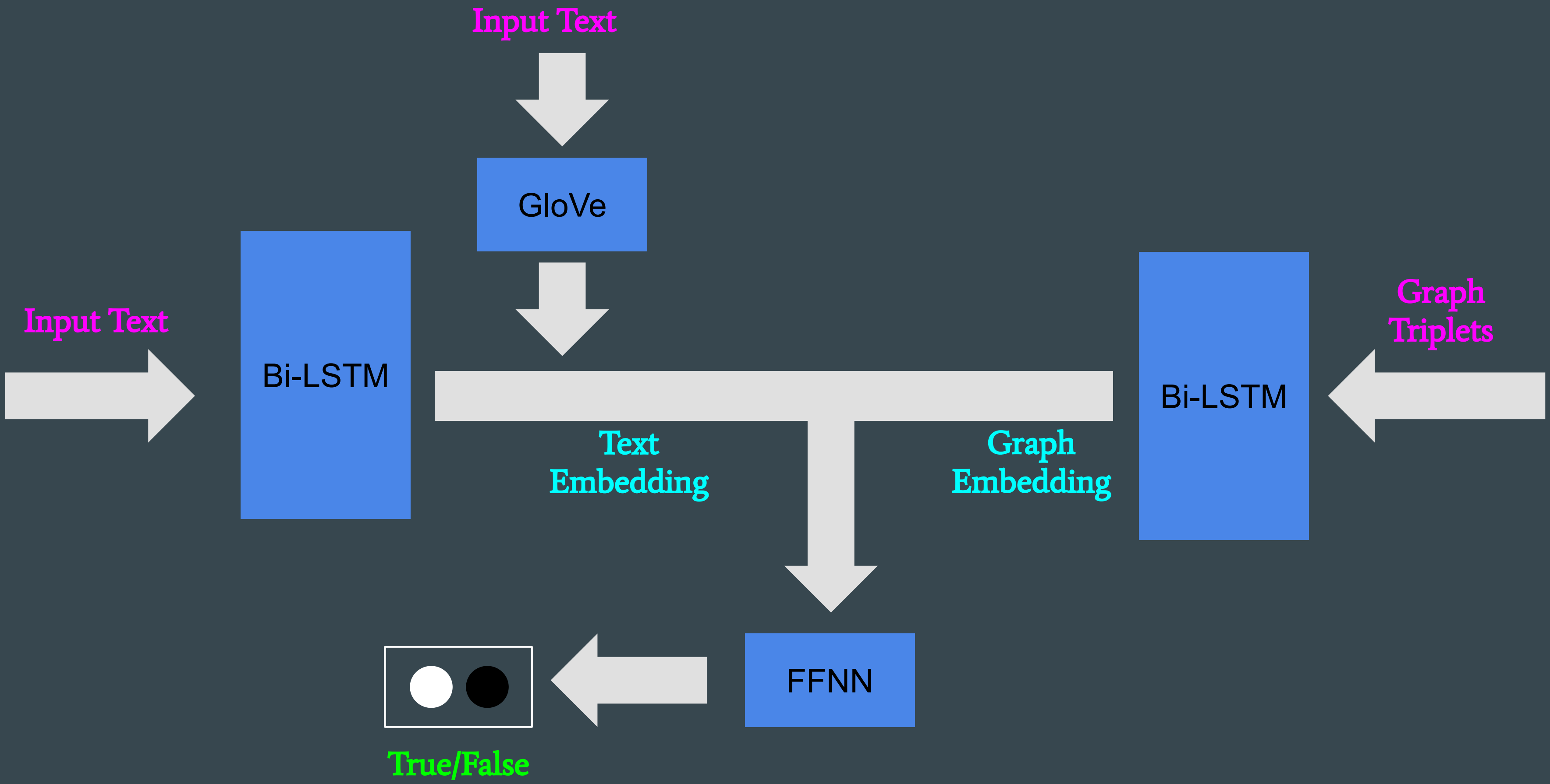
Top 5 Results from NEL: [Apple (company), apple (fruit), ...]

Top 1 Recall: 0%
Top 5 Recall: 100%

Top K	Recall (overall pipeline)	Recall (individual)
1	26%	48%
5	40%	74%
10	47%	86%

Top 1 Recall (overall pipeline) using Neural EL by Kolitsas et al. 2018: 47%

Deep Learning - Text and Graph Embeddings



Input Text

GloVe

Input Text

Bi-LSTM

Text Embedding

Graph Embedding

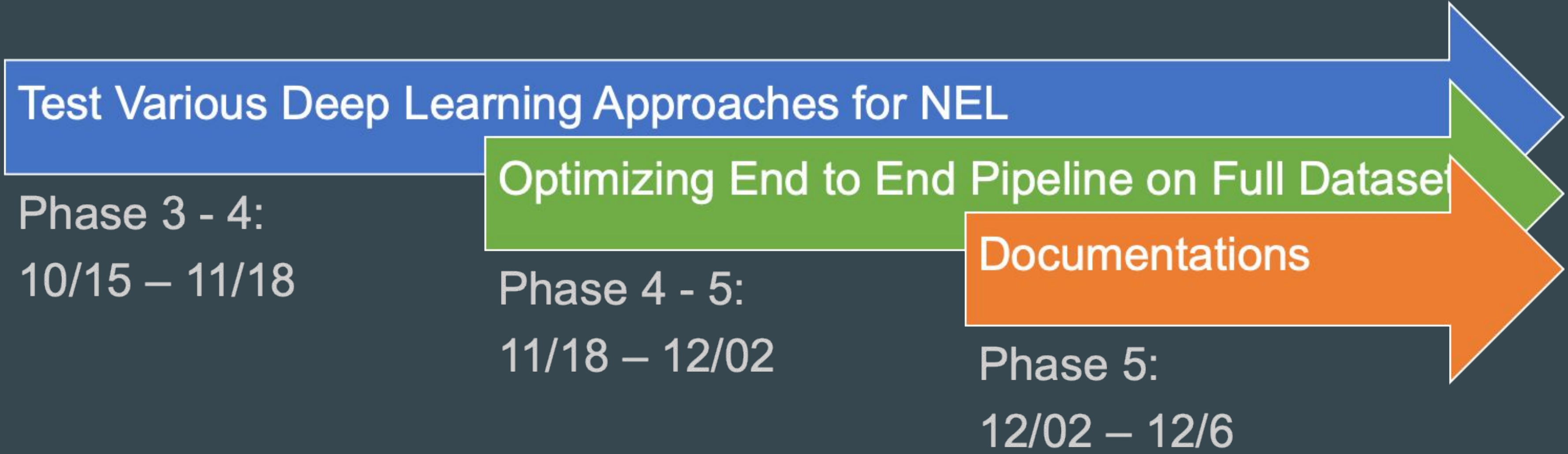
Bi-LSTM

Graph Triplets

FFNN

True/False

Timeline



Test Various Deep Learning Approaches for NEL

Phase 3 - 4:

10/15 – 11/18

Optimizing End to End Pipeline on Full Dataset

Phase 4 - 5:

11/18 – 12/02

Documentations

Phase 5:

12/02 – 12/6