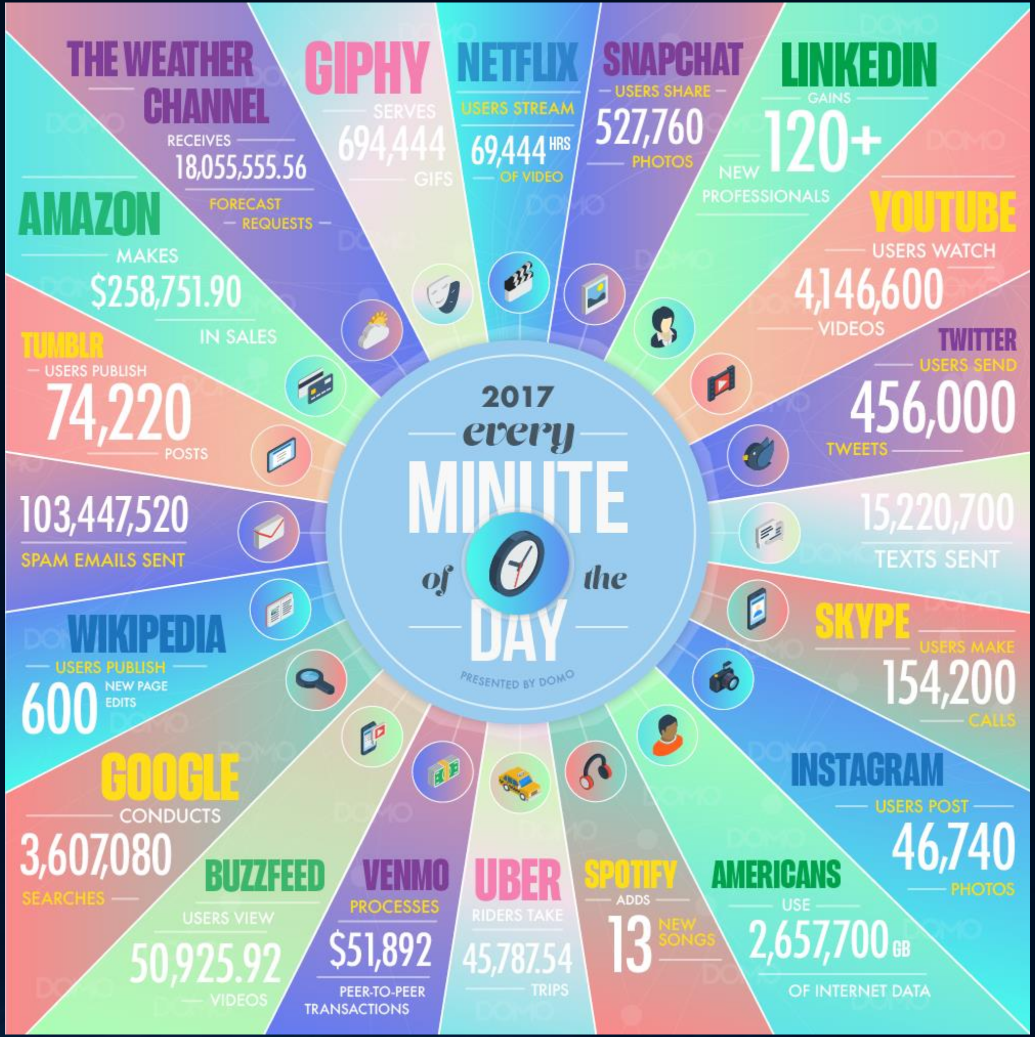
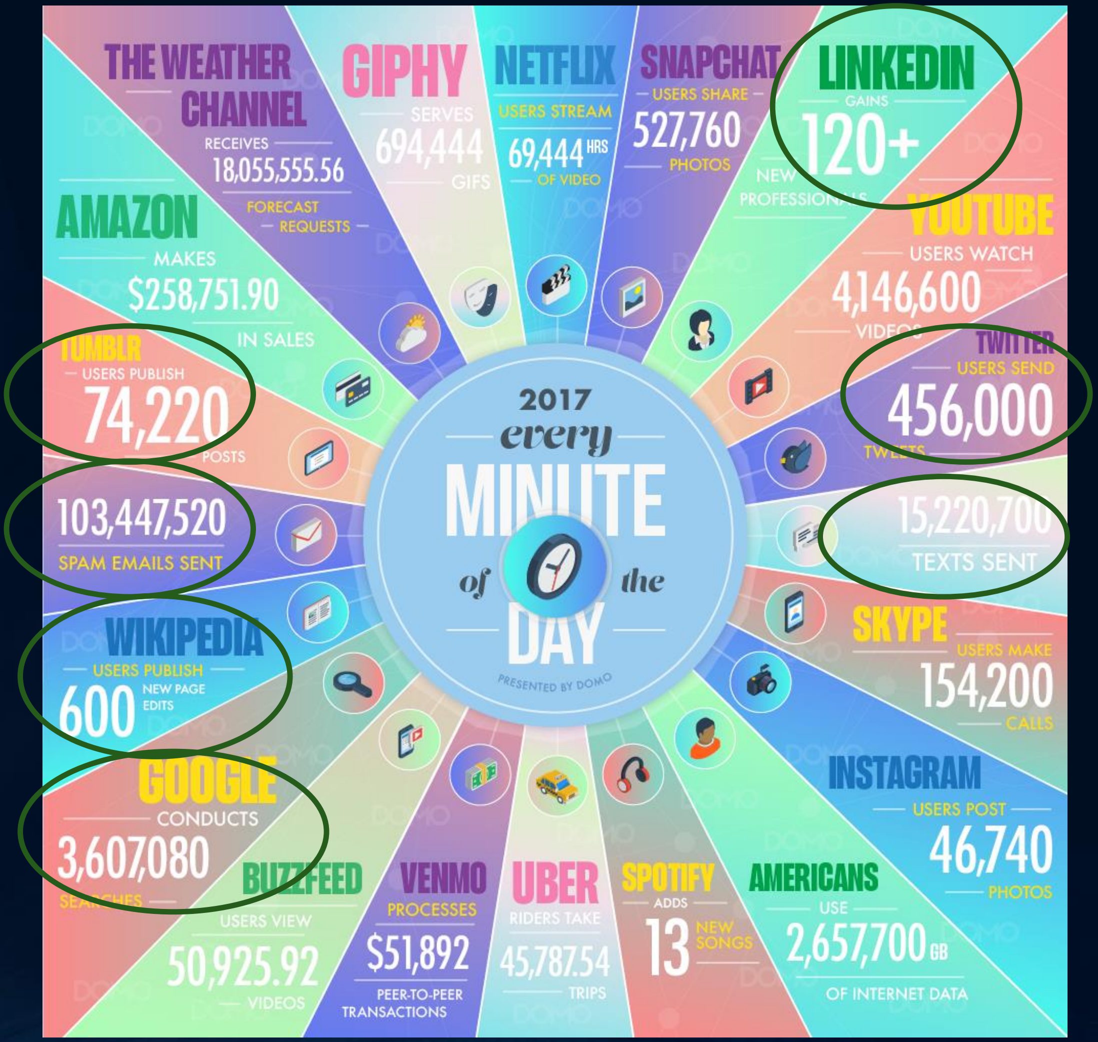
Introduction to Natural Language Processing

CHRISTOPHE SERVAN, PHD

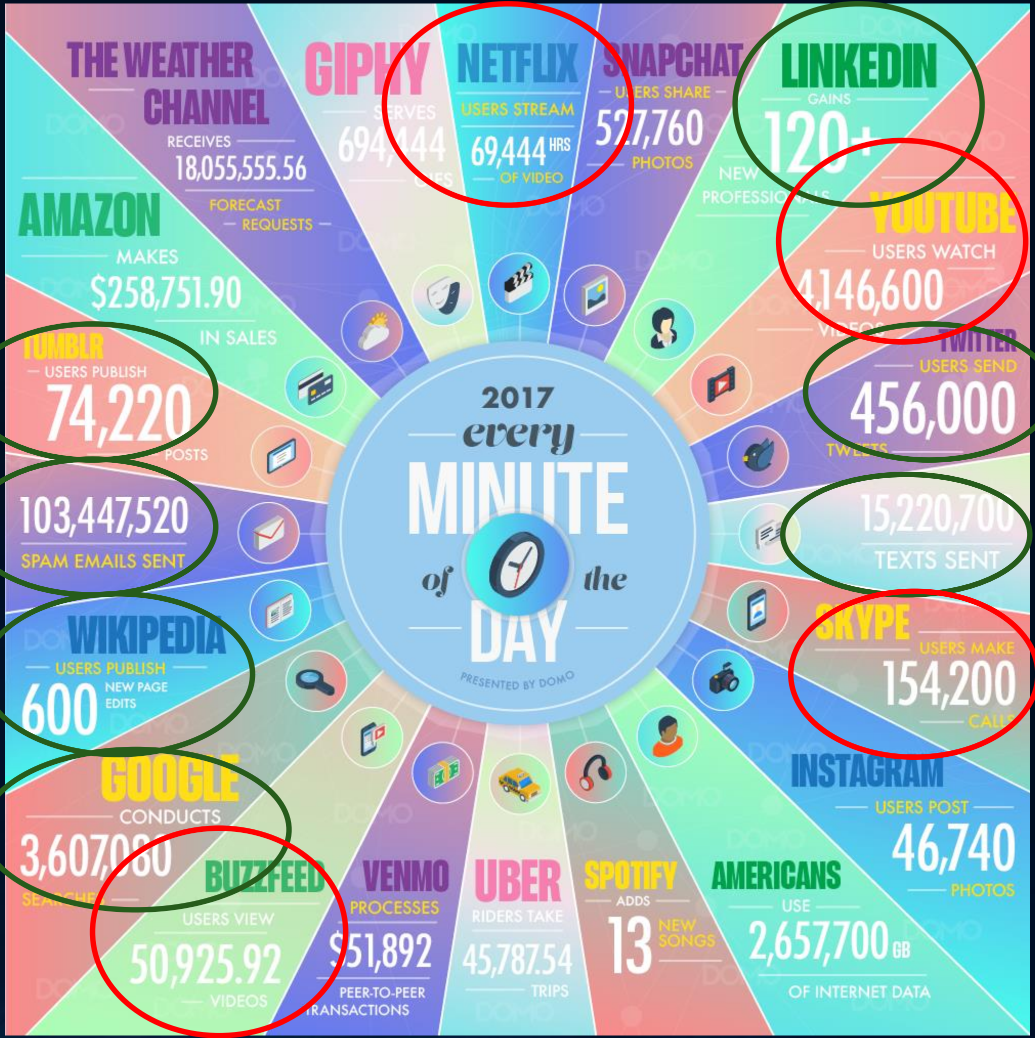
Data generated every minutes (2017)



Data generated every minutes (2017)



Data generated every minutes (2017)



Natural Language Processing (NLP)

- NLP is a research field which studies how computers can analyse, understand, generate and derive meaning from the Human language.
- How language looks to Human:

Nikola Tesla (10 July 1856 – 7 January 1943) was a Serbian-American inventor, electrical engineer, mechanical engineer, and futurist who is best known for his contributions to the design of the modern alternating current (AC) electricity supply system.

- How language looks to computers:

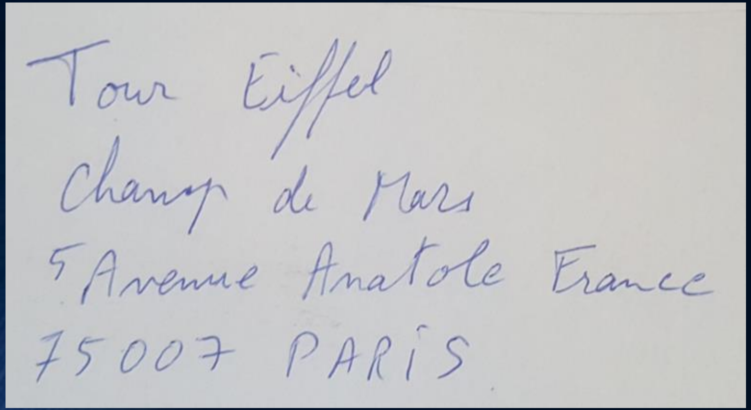
Никола Те́сла (10 июля 1856 — 7 января 1943) — изобретатель в области электротехники и радиотехники сербского происхождения, учёный, инженер, физик.

Documents

- Written
- Speech
- Web
- Text Structure
- Layout analysis

Written documents

- Optical Character Recognition (OCR):
  converting image which contain text (typed, handwritten, or printed) into machine-encoded text



Tour Eiffel
Champ de Mars
5 Avenue Anatole France
75007 PARIS

Tour Eiffel
Champ de Mars
5 Avenue Anatole France
75007 PARIS

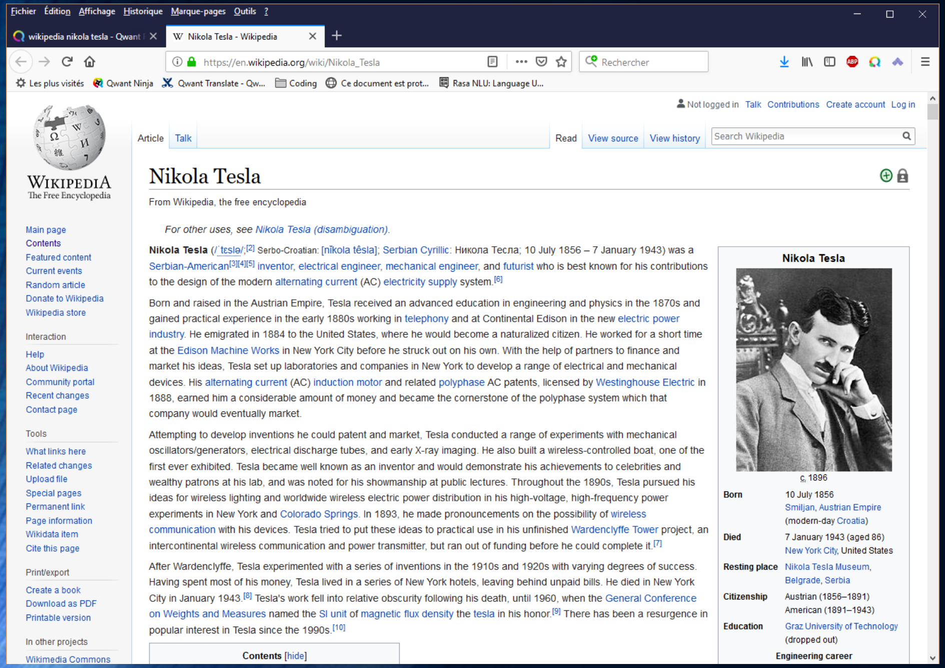
Spoken documents

- Automatic Speech Recognition (ASR): transcribe a spoken language into text

Tour Eiffel
Champ de Mars
5 Avenue Anatole France
75007 PARIS

Web documents

- Crawling:
  exploring the Internet and sytematically downloading and analysing web pages



Nikola Tesla

From Wikipedia, the free encyclopedia

For other uses, see Nikola Tesla (disambiguation).

Nikola Tesla (/ˈtɛslə/;[2] Serbo-Croatian: [nǐkola têsla]; Serbian Cyrillic: Никола Тесла; 10 July 1856 – 7 January 1943) was a Serbian-American[3][4][5] inventor, electrical engineer, mechanical engineer, and futurist who is best known for his contributions to the design of the modern alternating current (AC) electricity supply system.[6]

Born and raised in the Austrian Empire, Tesla received an advanced education in engineering and physics in the 1870s and gained practical experience in the early 1880s working in telephony and at Continental Edison in the new electric power industry. He emigrated in 1884 to the United States, where he would become a naturalized citizen. He worked for a short time at the Edison Machine Works in New York City before he struck out on his own. With the help of partners to finance and market his ideas, Tesla set up laboratories and companies in New York to develop a range of electrical and mechanical devices. His alternating current (AC) induction motor and related polyphase AC patents, licensed by Westinghouse Electric in 1888, earned him a considerable amount of money and became the cornerstone of the polyphase system which that company would eventually market.

Attempting to develop inventions he could patent and market, Tesla conducted a range of experiments with mechanical oscillators/generators, electrical discharge tubes, and early X-ray imaging. He also built a wireless-controlled boat, one of the first ever exhibited. Tesla became well known as an inventor and would demonstrate his achievements to celebrities and wealthy patrons at his lab, and was noted for his showmanship at public lectures. Throughout the 1890s, Tesla pursued his ideas for wireless lighting and worldwide wireless electric power distribution in his high-voltage, high-frequency power experiments in New York and Colorado Springs. In 1893, he made pronouncements on the possibility of wireless communication with his devices. Tesla tried to put these ideas to practical use in his unfinished Wardenclyffe Tower project, an intercontinental wireless communication and power transmitter, but ran out of funding before he could complete it.[7]

After Wardenclyffe, Tesla experimented with a series of inventions in the 1910s and 1920s with varying degrees of success. Having spent most of his money, Tesla lived in a series of New York hotels, leaving behind unpaid bills. He died in New York City in January 1943.[8] Tesla's work fell into relative obscurity following his death, until 1960, when the General Conference on Weights and Measures named the SI unit of magnetic flux density the tesla in his honor.[9] There has been a resurgence in popular interest in Tesla since the 1990s.[10]

Contents [hide]

Nikola Tesla

c. 1896

Born	10 July 1856 Smiljan, Austrian Empire (modern-day Croatia)
Died	7 January 1943 (aged 86) New York City, United States
Resting place	Nikola Tesla Museum, Belgrade, Serbia
Citizenship	Austrian (1856–1891) American (1891–1943)
Education	Graz University of Technology (dropped out)

Engineering career

Nikola Tesla (10 July 1856 – 7 January 1943) was a Serbian-American inventor, electrical engineer, mechanical engineer, and futurist who is best known for his contributions to the design of the modern alternating current (AC) electricity supply system.

Documents

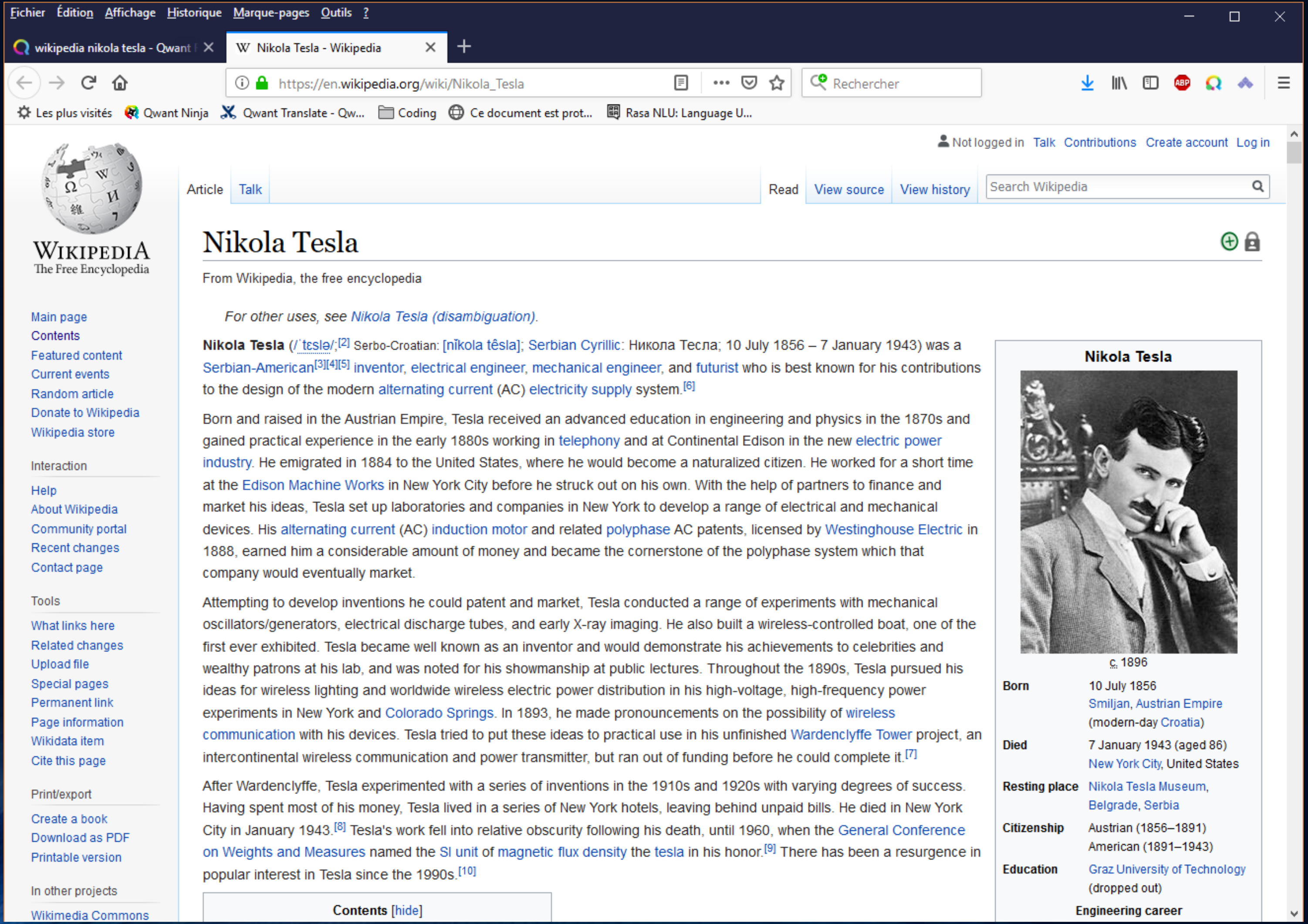
Unstructured or Structured information ?

- **Unstructured Information**: no meaning, no meta data, no data structured, no type associated to text

- **Structured Information**: layout, word position in the sentence, punctuation, sentence position in the document

Documents

- Document Layout Analysis



Nikola Tesla

From Wikipedia, the free encyclopedia

For other uses, see Nikola Tesla (disambiguation).

Nikola Tesla (/ˈtɛslə/;[2] Serbo-Croatian: [nǐkola têsla]; Serbian Cyrillic: Никола Тесла; 10 July 1856 – 7 January 1943) was a Serbian-American[3][4][5] inventor, electrical engineer, mechanical engineer, and futurist who is best known for his contributions to the design of the modern alternating current (AC) electricity supply system.[6]

Born and raised in the Austrian Empire, Tesla received an advanced education in engineering and physics in the 1870s and gained practical experience in the early 1880s working in telephony and at Continental Edison in the new electric power industry. He emigrated in 1884 to the United States, where he would become a naturalized citizen. He worked for a short time at the Edison Machine Works in New York City before he struck out on his own. With the help of partners to finance and market his ideas, Tesla set up laboratories and companies in New York to develop a range of electrical and mechanical devices. His alternating current (AC) induction motor and related polyphase AC patents, licensed by Westinghouse Electric in 1888, earned him a considerable amount of money and became the cornerstone of the polyphase system which that company would eventually market.

Attempting to develop inventions he could patent and market, Tesla conducted a range of experiments with mechanical oscillators/generators, electrical discharge tubes, and early X-ray imaging. He also built a wireless-controlled boat, one of the first ever exhibited. Tesla became well known as an inventor and would demonstrate his achievements to celebrities and wealthy patrons at his lab, and was noted for his showmanship at public lectures. Throughout the 1890s, Tesla pursued his ideas for wireless lighting and worldwide wireless electric power distribution in his high-voltage, high-frequency power experiments in New York and Colorado Springs. In 1893, he made pronouncements on the possibility of wireless communication with his devices. Tesla tried to put these ideas to practical use in his unfinished Wardenclyffe Tower project, an intercontinental wireless communication and power transmitter, but ran out of funding before he could complete it.[7]

After Wardenclyffe, Tesla experimented with a series of inventions in the 1910s and 1920s with varying degrees of success. Having spent most of his money, Tesla lived in a series of New York hotels, leaving behind unpaid bills. He died in New York City in January 1943.[8] Tesla's work fell into relative obscurity following his death, until 1960, when the General Conference on Weights and Measures named the SI unit of magnetic flux density the tesla in his honor.[9] There has been a resurgence in popular interest in Tesla since the 1990s.[10]

Contents [hide]

Nikola Tesla

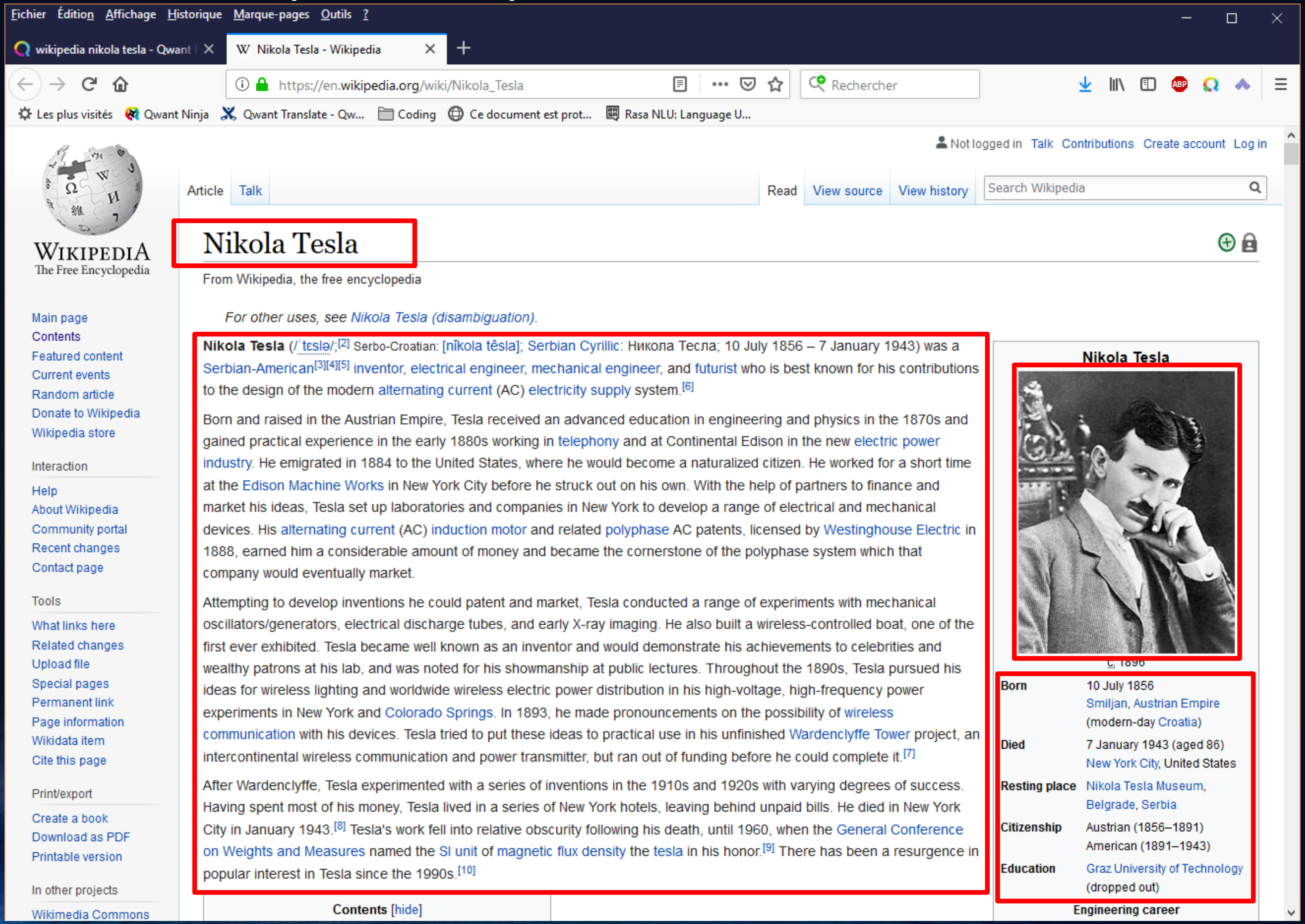
c. 1896

Born	10 July 1856 Smiljan, Austrian Empire (modern-day Croatia)
Died	7 January 1943 (aged 86) New York City, United States
Resting place	Nikola Tesla Museum, Belgrade, Serbia
Citizenship	Austrian (1856–1891) American (1891–1943)
Education	Graz University of Technology (dropped out)

Engineering career

Documents

- Document Layout Analysis



Nikola Tesla

From Wikipedia, the free encyclopedia

For other uses, see Nikola Tesla (disambiguation).

Nikola Tesla (/ˈtɛslə/;[2] Serbo-Croatian: [nǐkola têsla]; Serbian Cyrillic: Никола Тесла; 10 July 1856 – 7 January 1943) was a Serbian-American[3][4][5] inventor, electrical engineer, mechanical engineer, and futurist who is best known for his contributions to the design of the modern alternating current (AC) electricity supply system.[6]

Born and raised in the Austrian Empire, Tesla received an advanced education in engineering and physics in the 1870s and gained practical experience in the early 1880s working in telephony and at Continental Edison in the new electric power industry. He emigrated in 1884 to the United States, where he would become a naturalized citizen. He worked for a short time at the Edison Machine Works in New York City before he struck out on his own. With the help of partners to finance and market his ideas, Tesla set up laboratories and companies in New York to develop a range of electrical and mechanical devices. His alternating current (AC) induction motor and related polyphase AC patents, licensed by Westinghouse Electric in 1888, earned him a considerable amount of money and became the cornerstone of the polyphase system which that company would eventually market.

Attempting to develop inventions he could patent and market, Tesla conducted a range of experiments with mechanical oscillators/generators, electrical discharge tubes, and early X-ray imaging. He also built a wireless-controlled boat, one of the first ever exhibited. Tesla became well known as an inventor and would demonstrate his achievements to celebrities and wealthy patrons at his lab, and was noted for his showmanship at public lectures. Throughout the 1890s, Tesla pursued his ideas for wireless lighting and worldwide wireless electric power distribution in his high-voltage, high-frequency power experiments in New York and Colorado Springs. In 1893, he made pronouncements on the possibility of wireless communication with his devices. Tesla tried to put these ideas to practical use in his unfinished Wardenclyffe Tower project, an intercontinental wireless communication and power transmitter, but ran out of funding before he could complete it.[7]

After Wardenclyffe, Tesla experimented with a series of inventions in the 1910s and 1920s with varying degrees of success. Having spent most of his money, Tesla lived in a series of New York hotels, leaving behind unpaid bills. He died in New York City in January 1943.[8] Tesla's work fell into relative obscurity following his death, until 1960, when the General Conference on Weights and Measures named the SI unit of magnetic flux density the tesla in his honor.[9] There has been a resurgence in popular interest in Tesla since the 1990s.[10]

Nikola Tesla

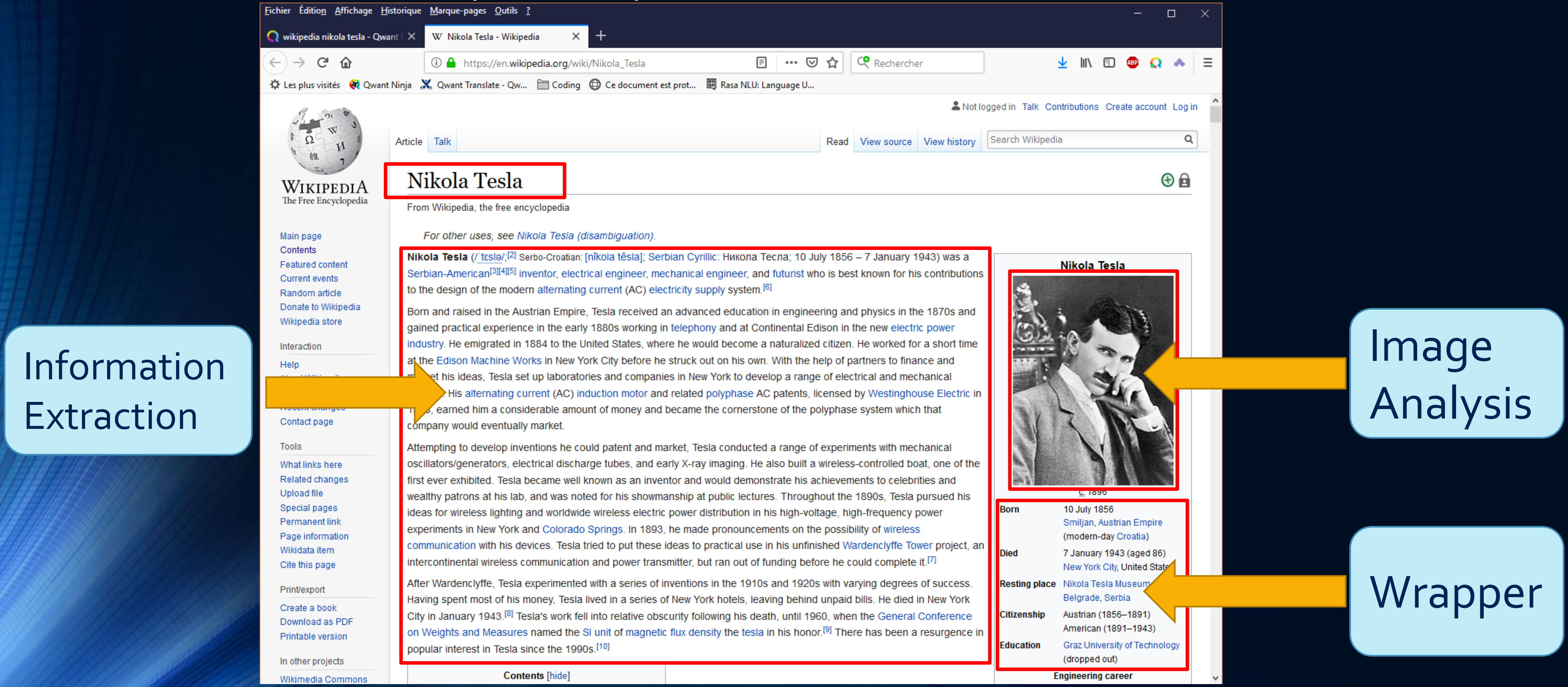
c. 1896

Born	10 July 1856 Smiljan, Austrian Empire (modern-day Croatia)
Died	7 January 1943 (aged 86) New York City, United States
Resting place	Nikola Tesla Museum, Belgrade, Serbia
Citizenship	Austrian (1856–1891) American (1891–1943)
Education	Graz University of Technology (dropped out)

Engineering career

Documents

- Document Layout Analysis



Nikola Tesla

From Wikipedia, the free encyclopedia

For other uses, see Nikola Tesla (disambiguation).

Nikola Tesla (/ˈtɛslə/;[2] Serbo-Croatian: [nǐkola têsla]; Serbian Cyrillic: Никола Тесла; 10 July 1856 – 7 January 1943) was a Serbian-American[3][4][5] inventor, electrical engineer, mechanical engineer, and futurist who is best known for his contributions to the design of the modern alternating current (AC) electricity supply system.[6]

Born and raised in the Austrian Empire, Tesla received an advanced education in engineering and physics in the 1870s and gained practical experience in the early 1880s working in telephony and at Continental Edison in the new electric power industry. He emigrated in 1884 to the United States, where he would become a naturalized citizen. He worked for a short time at the Edison Machine Works in New York City before he struck out on his own. With the help of partners to finance and m[illegible]et his ideas, Tesla set up laboratories and companies in New York to develop a range of electrical and mechanical [illegible] His alternating current (AC) induction motor and related polyphase AC patents, licensed by Westinghouse Electric in 1[illegible]s, earned him a considerable amount of money and became the cornerstone of the polyphase system which that company would eventually market.

Attempting to develop inventions he could patent and market, Tesla conducted a range of experiments with mechanical oscillators/generators, electrical discharge tubes, and early X-ray imaging. He also built a wireless-controlled boat, one of the first ever exhibited. Tesla became well known as an inventor and would demonstrate his achievements to celebrities and wealthy patrons at his lab, and was noted for his showmanship at public lectures. Throughout the 1890s, Tesla pursued his ideas for wireless lighting and worldwide wireless electric power distribution in his high-voltage, high-frequency power experiments in New York and Colorado Springs. In 1893, he made pronouncements on the possibility of wireless communication with his devices. Tesla tried to put these ideas to practical use in his unfinished Wardenclyffe Tower project, an intercontinental wireless communication and power transmitter, but ran out of funding before he could complete it.[7]

After Wardenclyffe, Tesla experimented with a series of inventions in the 1910s and 1920s with varying degrees of success. Having spent most of his money, Tesla lived in a series of New York hotels, leaving behind unpaid bills. He died in New York City in January 1943.[8] Tesla's work fell into relative obscurity following his death, until 1960, when the General Conference on Weights and Measures named the SI unit of magnetic flux density the tesla in his honor.[9] There has been a resurgence in popular interest in Tesla since the 1990s.[10]

Nikola Tesla

c. 1896

Born	10 July 1856 Smiljan, Austrian Empire (modern-day Croatia)
Died	7 January 1943 (aged 86) New York City, United State
Resting place	Nikola Tesla Museum Belgrade, Serbia
Citizenship	Austrian (1856–1891) American (1891–1943)
Education	Graz University of Technology (dropped out)

Engineering career

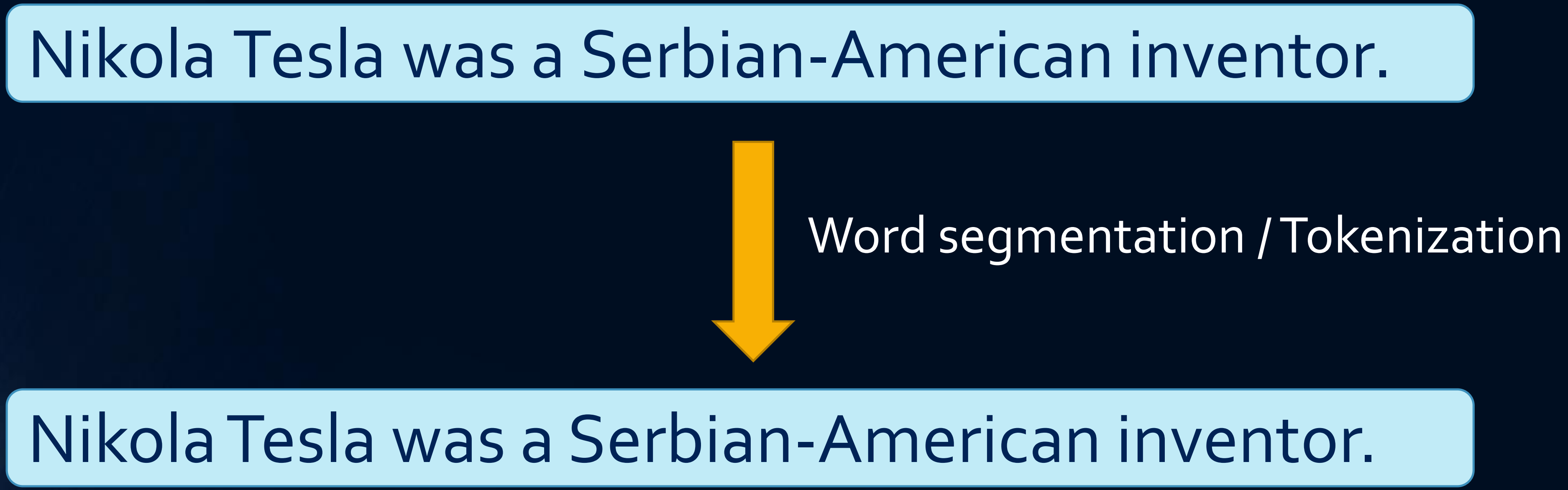
Information Extraction

Image Analysis

Wrapper

Character-level analysis

Tokenization: task of splitting text into words or tokens.



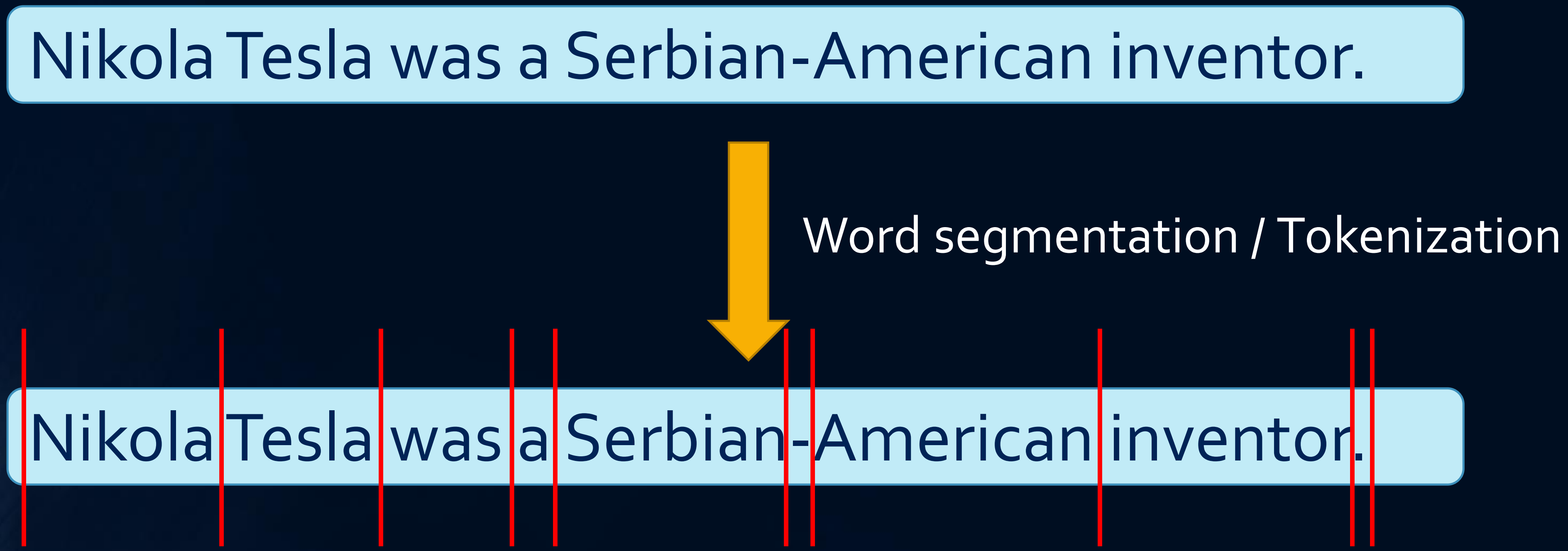
Nikola Tesla was a Serbian-American inventor.

Word segmentation / Tokenization

Nikola Tesla was a Serbian-American inventor.

Character-level analysis

Tokenization: task of splitting text into words or tokens.



Nikola Tesla was a Serbian-American inventor.

Word segmentation / Tokenization

Nikola Tesla was a Serbian-American inventor.

Word-level analysis

- Morphological analysis
- Lemmatization
- Stemming
- Word Sense Desambiguation
- Part-of-Speech Tagging
- Name Entity Recognition
- Entity Linking

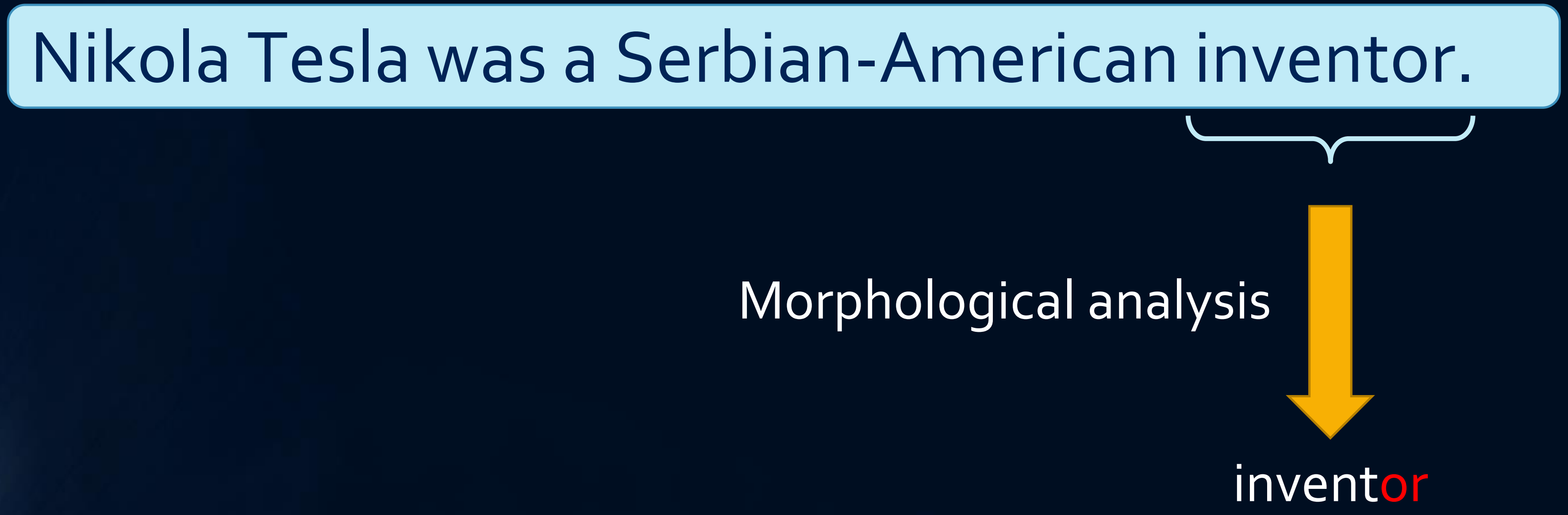
Morphological analysis

Splitting words into text into compoments (morphemes).

Nikola Tesla was a Serbian-American inventor.

Morphological analysis

Splitting words into text into compoments (morphemes).



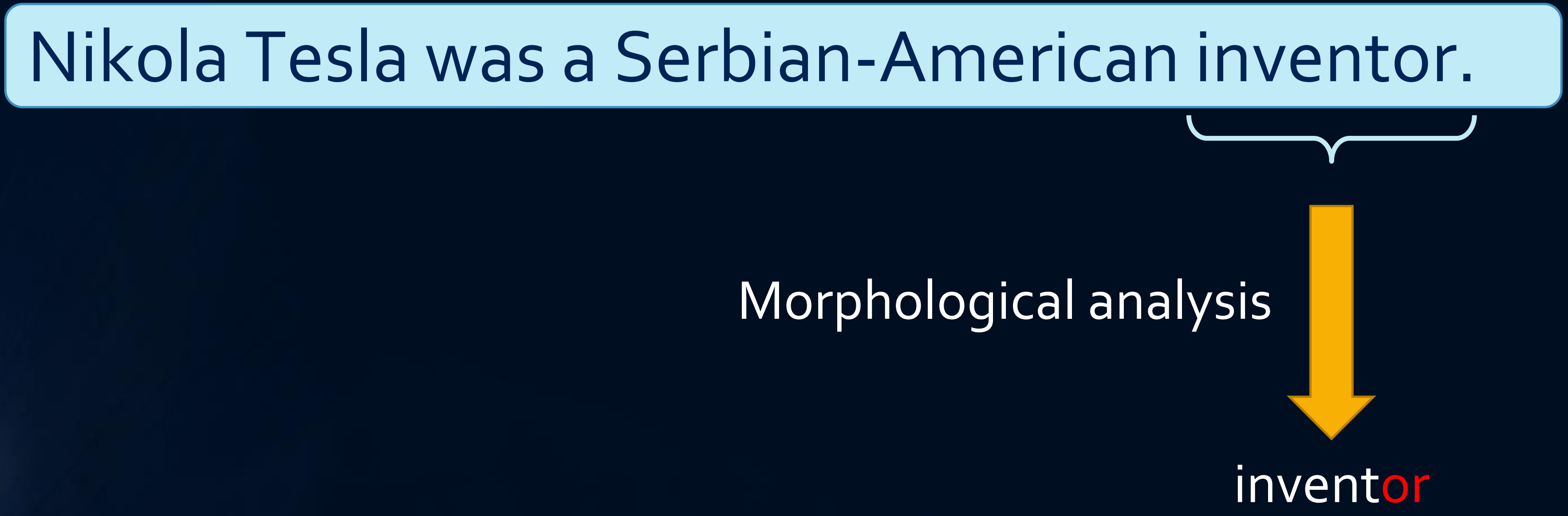
Nikola Tesla was a Serbian-American inventor.

Morphological analysis

inventor

Morphological analysis

Splitting words into text into compoments (morphemes).



Nikola Tesla was a Serbian-American inventor.

Morphological analysis

inventor

- Impossible
- Parisiennes

→

Im (prefix) + possible (verb)
Paris (Noun) + ien (suffix) + ne (female suffix) + s (plural suffix)

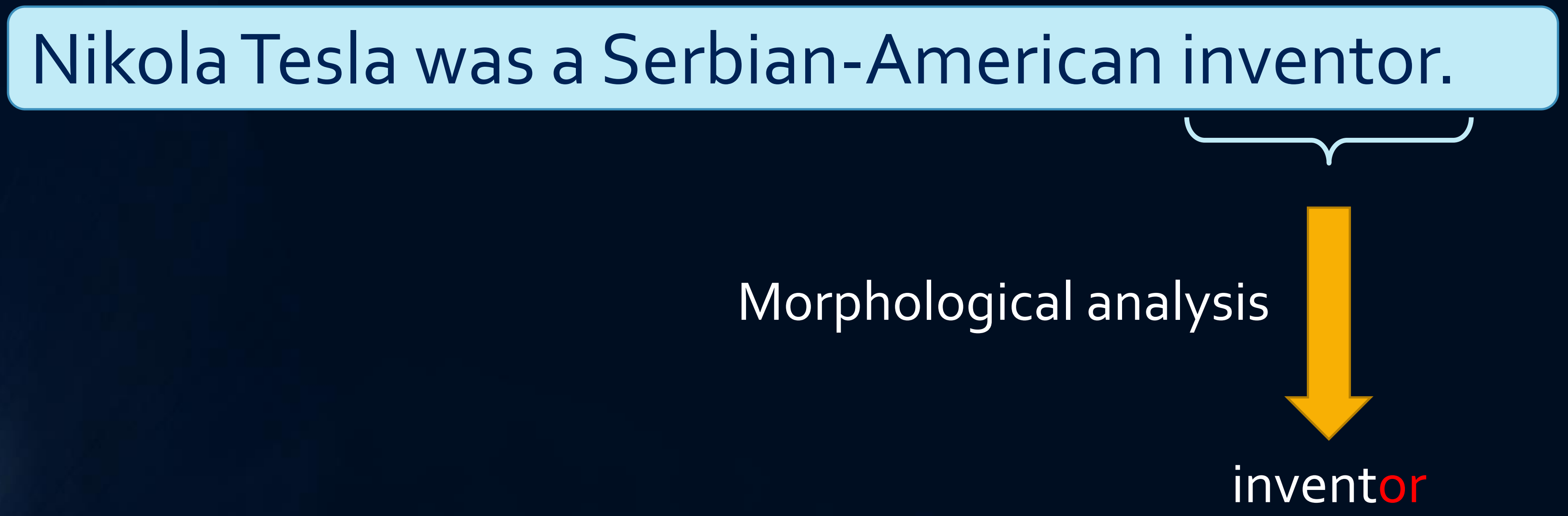
Lemmatization

map words to lemmas (word roots).

Nikola Tesla was a Serbian-American inventor.

Lemmatization

map words to lemmas (word roots).



Nikola Tesla was a Serbian-American inventor.

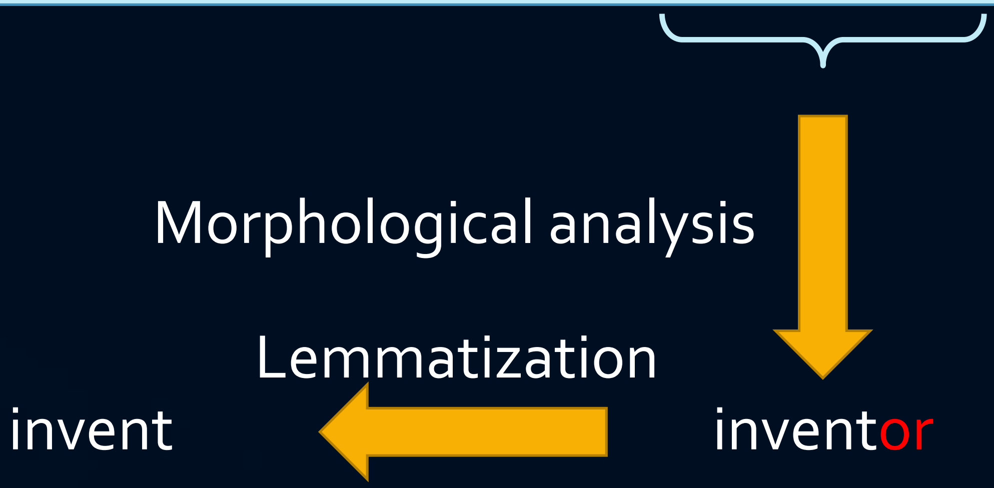
Morphological analysis

inventor

Lemmatization

map words to lemmas (word roots).

Nikola Tesla was a Serbian-American inventor.



Morphological analysis

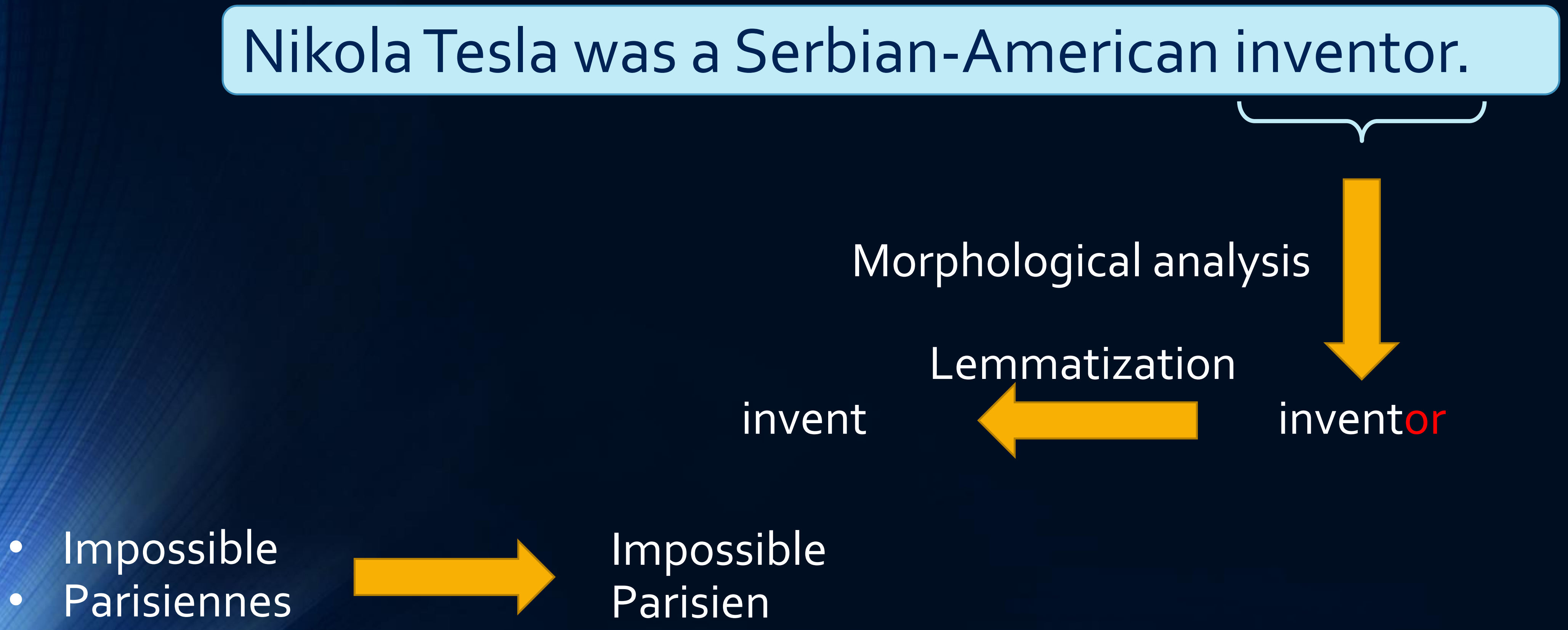
Lemmatization

invent

inventor

Lemmatization

map words to lemmas (word roots).



Nikola Tesla was a Serbian-American inventor.

Morphological analysis

Lemmatization

invent

inventor

- Impossible
- Parisiennes

Impossible
Parisien

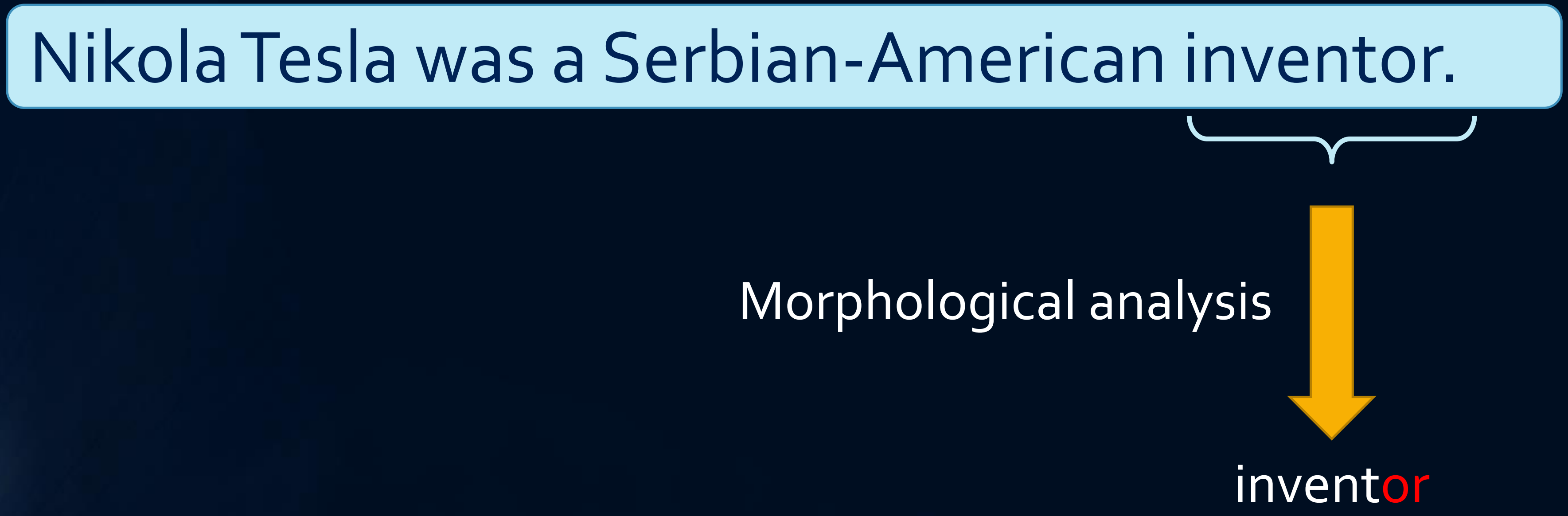
Stemming

map words to stems (word radicals). The easiest word simplification

Nikola Tesla was a Serbian-American inventor.

Stemming

map words to stems (word radicals). The easiest word simplification



Nikola Tesla was a Serbian-American inventor.

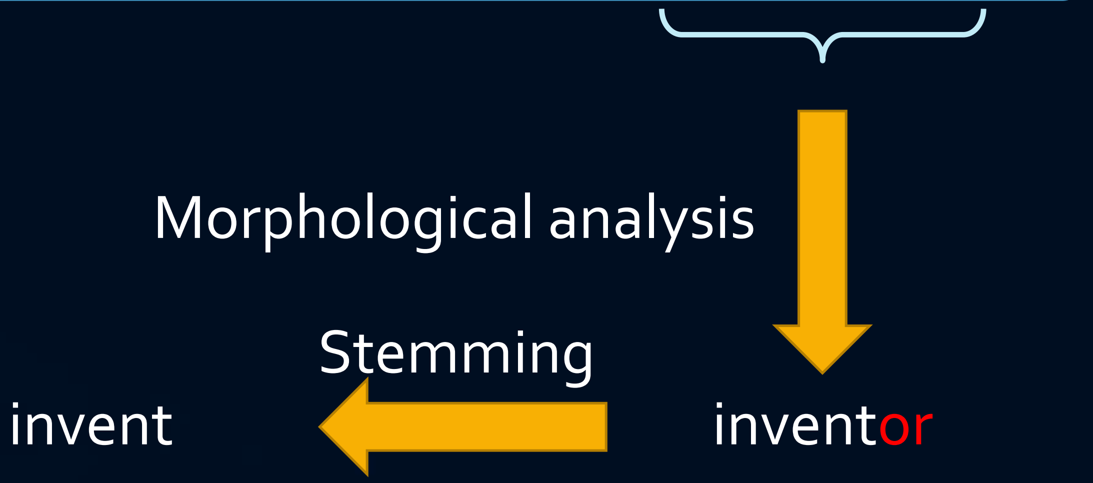
Morphological analysis

inventor

Stemming

map words to stems (word radicals). The easiest word simplification

Nikola Tesla was a Serbian-American inventor.



Morphological analysis

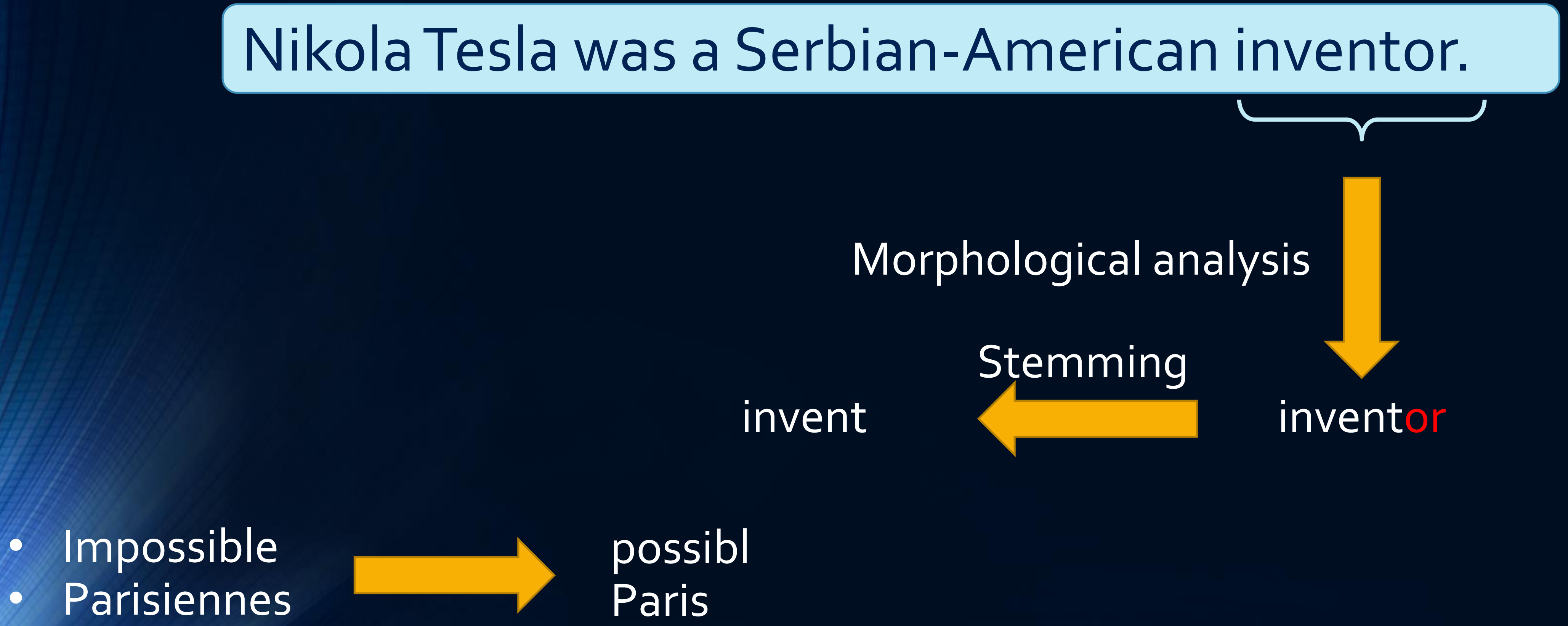
Stemming

invent

inventor

Stemming

map words to stems (word radicals). The easiest word simplification



Nikola Tesla was a Serbian-American inventor.

Morphological analysis

inventor

Stemming

invent

- Impossible
- Parisiennes

possibl
Paris

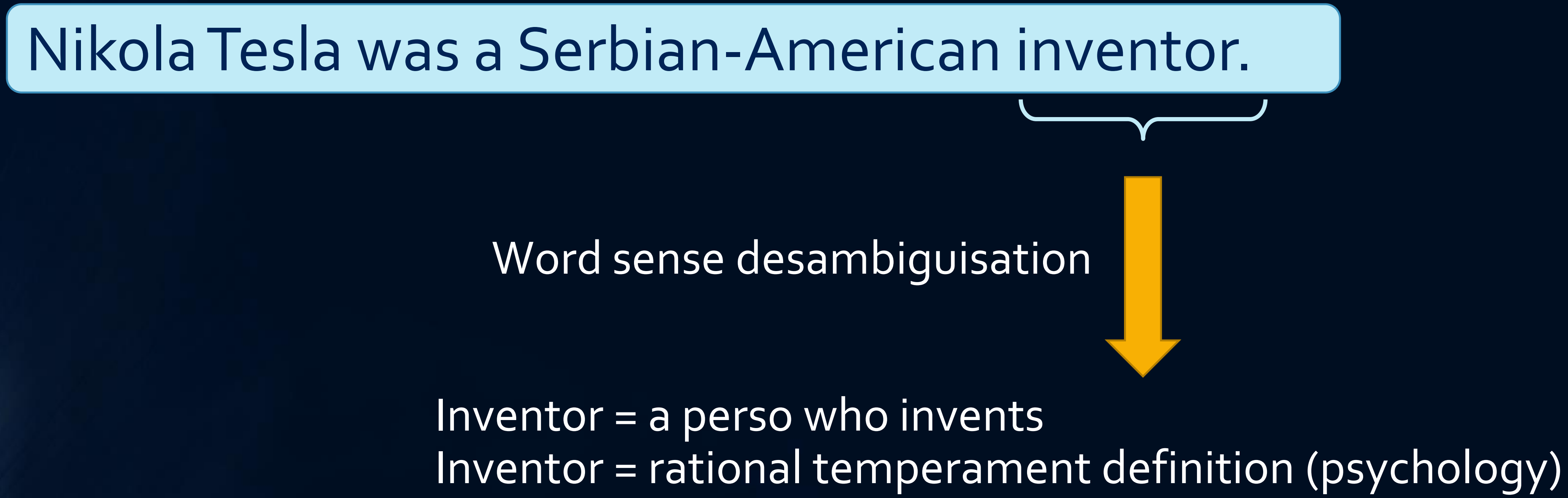
Word sense desambiguisation (WSD)

Identify the meaning of a word.

Nikola Tesla was a Serbian-American inventor.

Word sense desambiguisation (WSD)

Identify the meaning of a word.



Nikola Tesla was a Serbian-American inventor.

Word sense desambiguisation

Inventor = a perso who invents
Inventor = rational temperament definition (psychology)

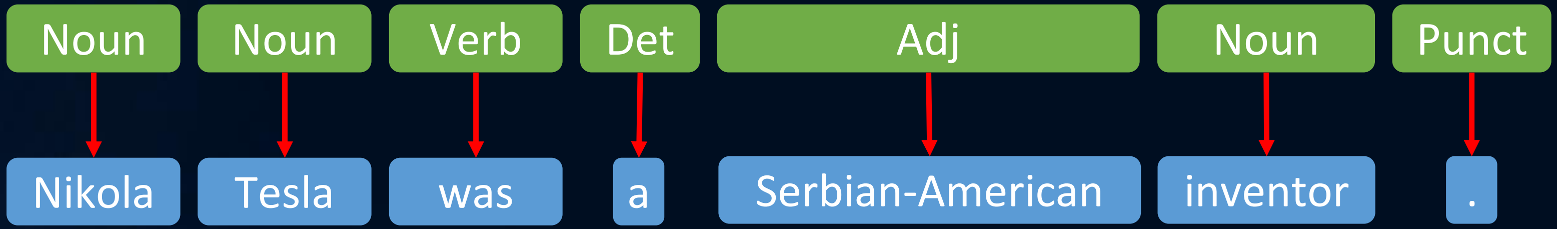
Part-of-Speech tagging (PoS)

Label a word (often a lexical category).

Nikola Tesla was a Serbian-American inventor .

Part-of-Speech tagging (PoS)

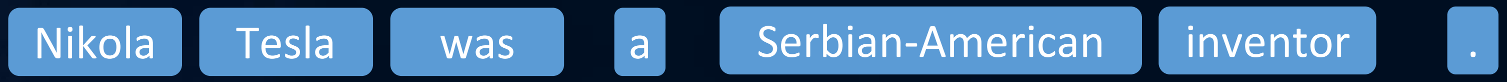
Label a word (often a lexical category).



Noun	Noun	Verb	Det	Adj	Noun	Punct
Nikola	Tesla	was	a	Serbian-American	inventor	.

Name Entity Recognition (NER)

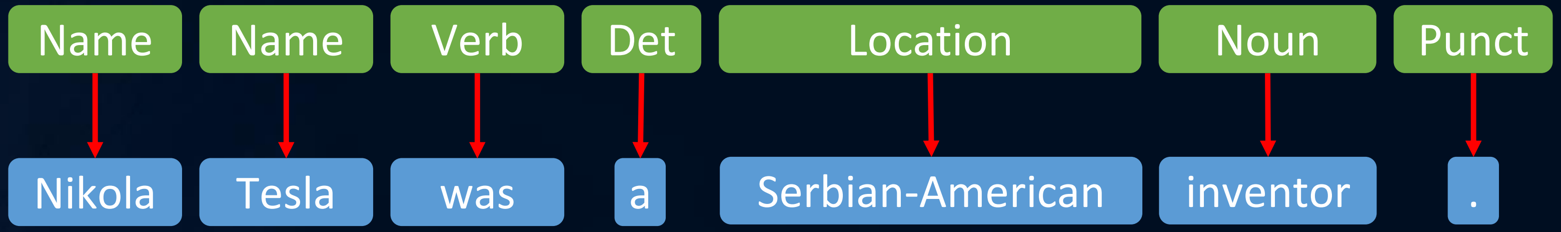
Extract entities (names, numbers, etc.)



Nikola Tesla was a Serbian-American inventor .

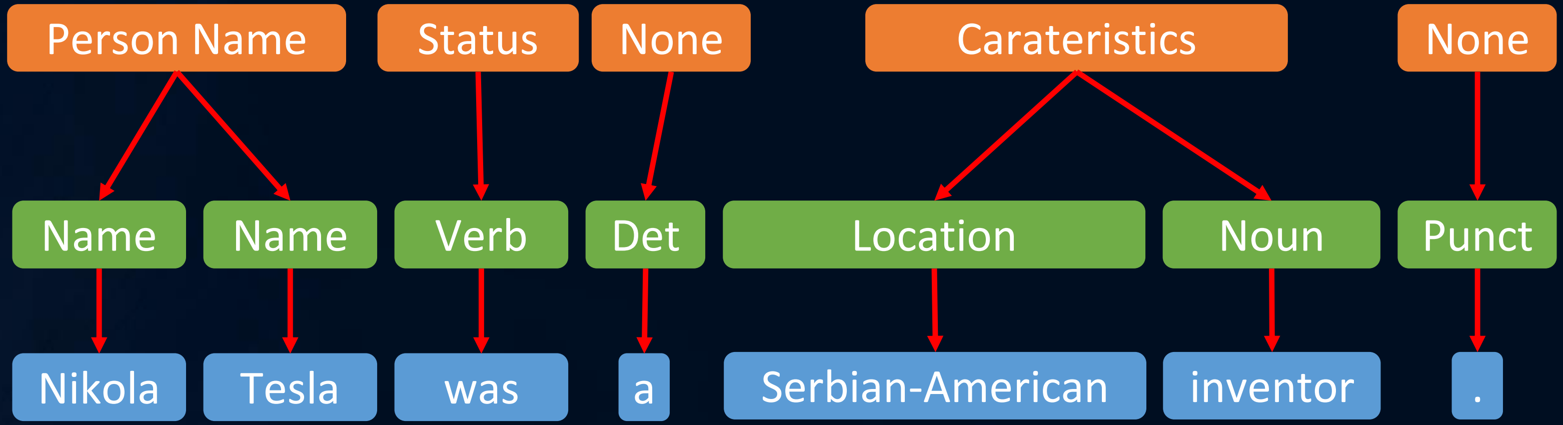
Name Entity Recognition (NER)

Extract entities (names, numbers, etc.)



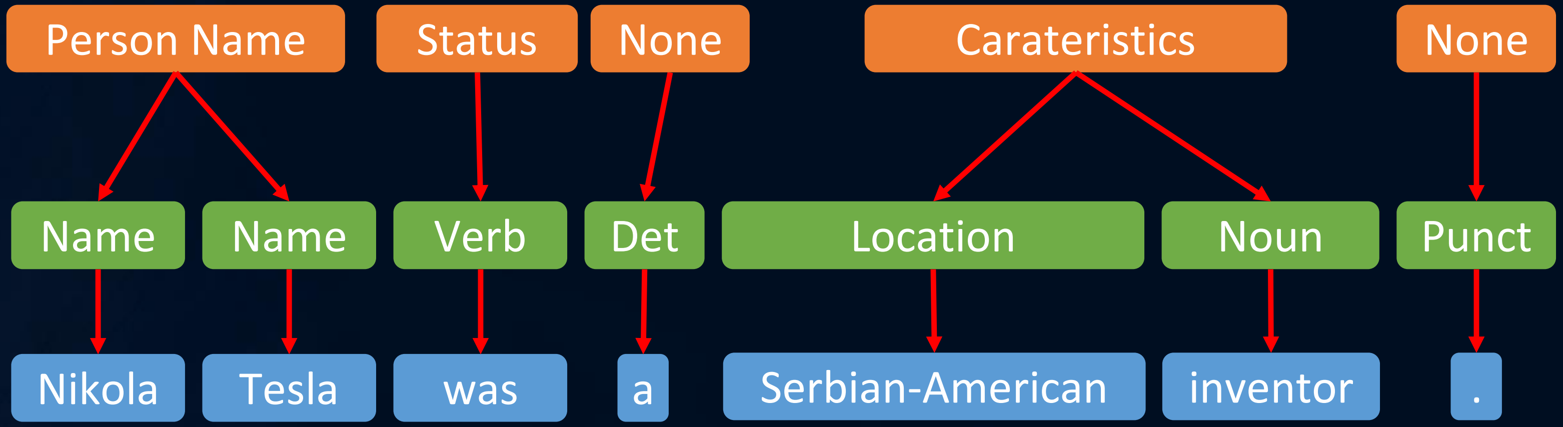
Name Entity Recognition (NER)

Extract entities (names, numbers, etc.)



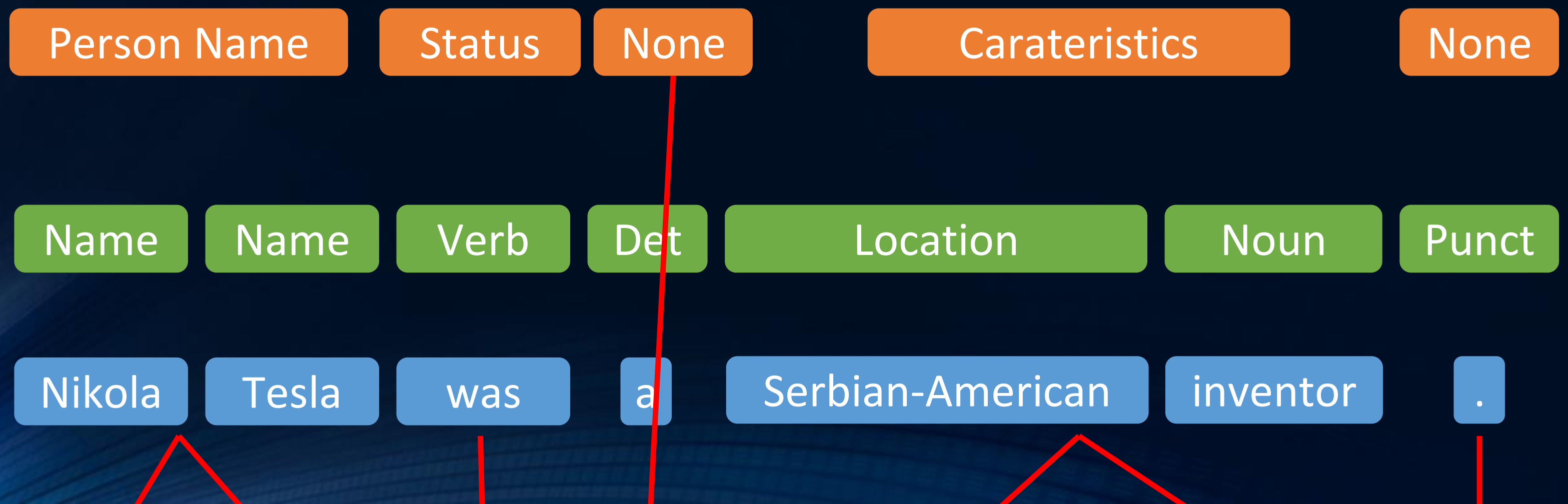
Entity Linking

Do the correspondance with entities in a database



Entity Linking

Do the correspondance with entities in a database



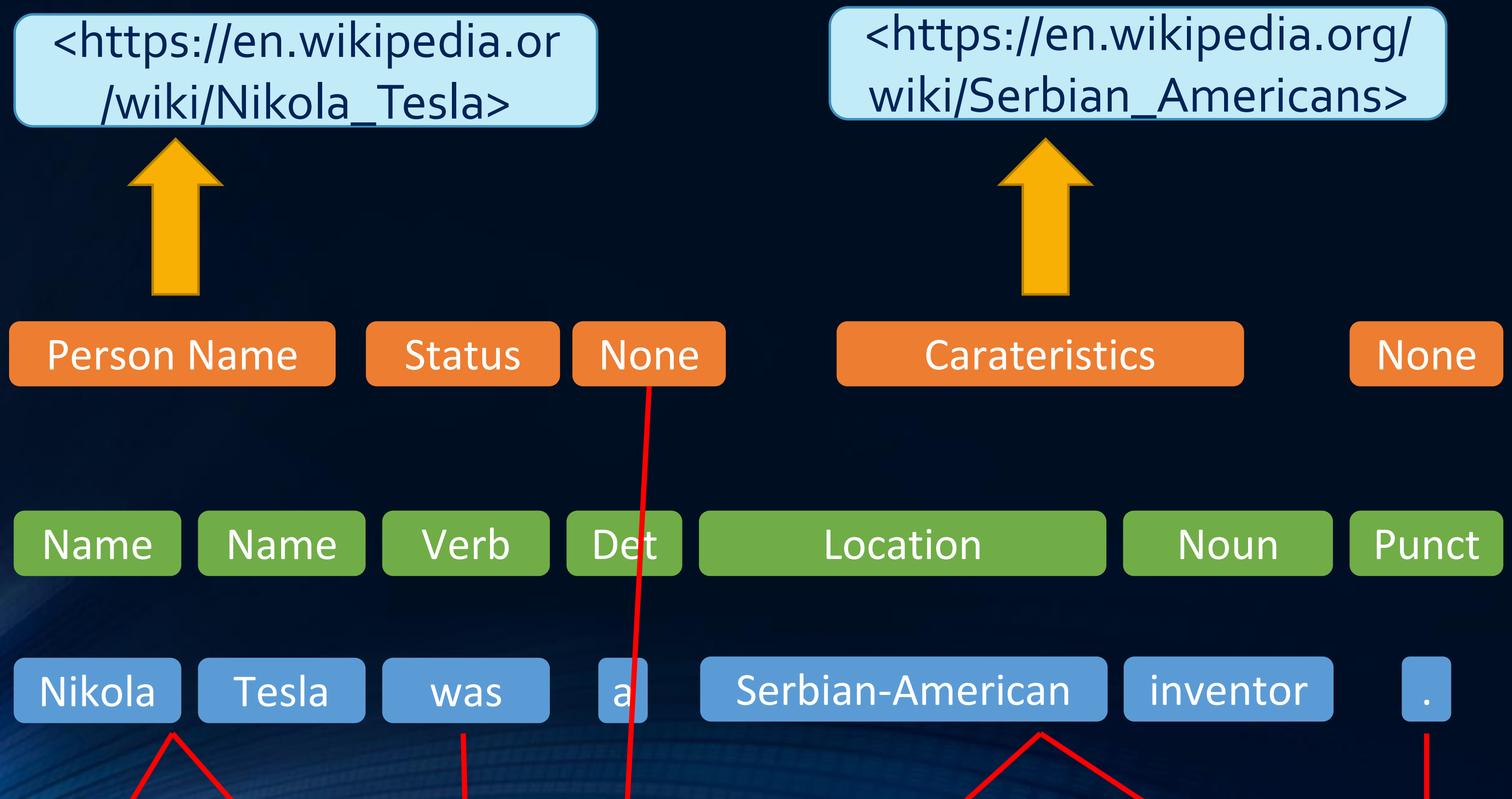
Person Name Status None Carateristics None

Name Name Verb Det Location Noun Punct

Nikola Tesla was a Serbian-American inventor .

Entity Linking

Do the correspondance with entities in a database



<https://en.wikipedia.org/wiki/Nikola_Tesla>

<https://en.wikipedia.org/wiki/Serbian_Americans>

Person Name Status None Carateristics None

Name Name Verb Det Location Noun Punct

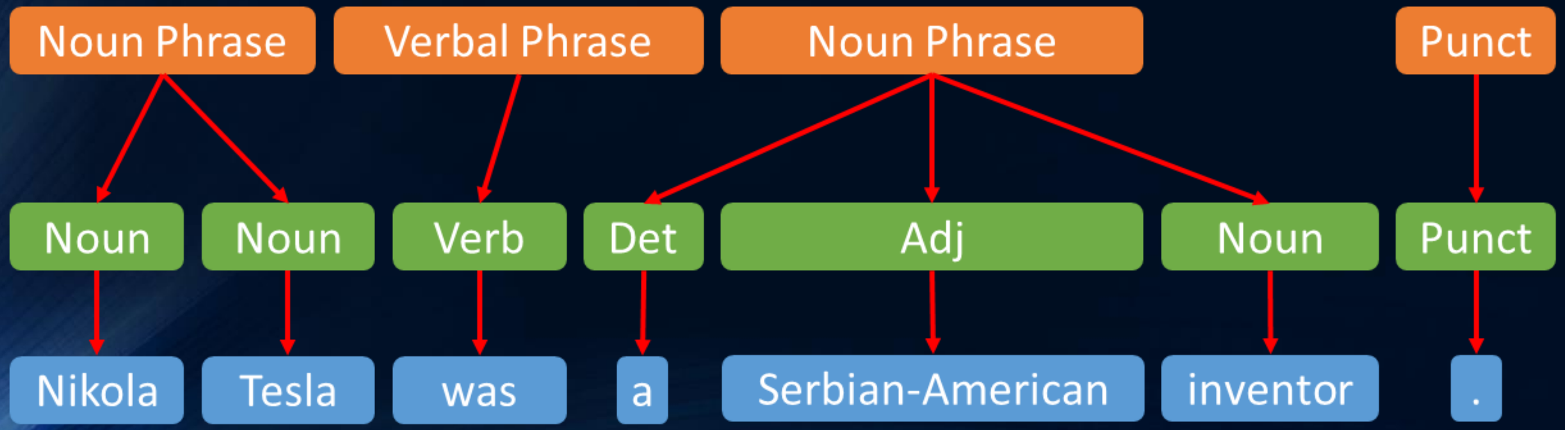
Nikola Tesla was a Serbian-American inventor .

Sentence-level analysis

- Syntactic analysis
- Dependency Analysis
- Semantic analysis
- Coreference resolution
- Information extraction
- Applications

Syntactic Analysis

Analyse the sentence structure.



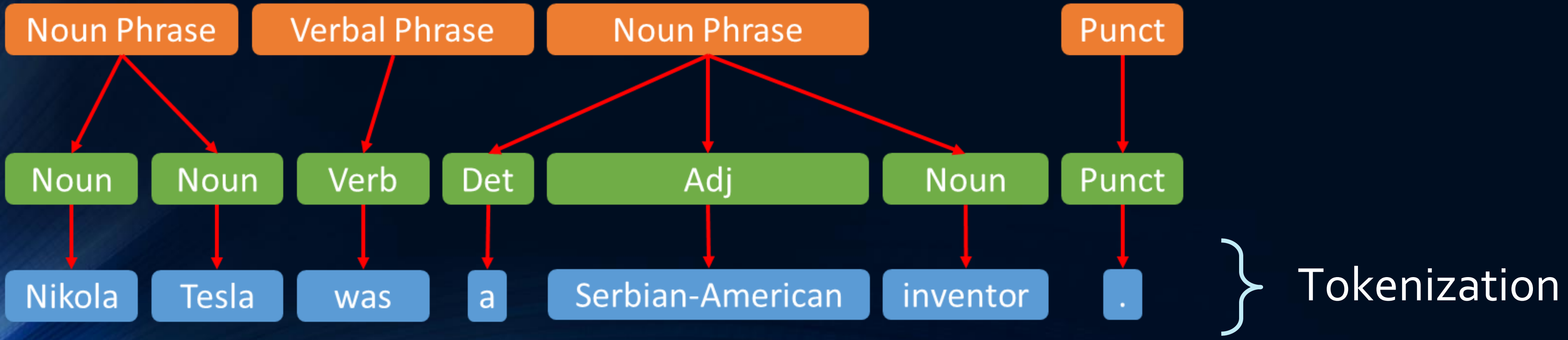
Noun Phrase Verbal Phrase Noun Phrase Punct

Noun Noun Verb Det Adj Noun Punct

Nikola Tesla was a Serbian-American inventor .

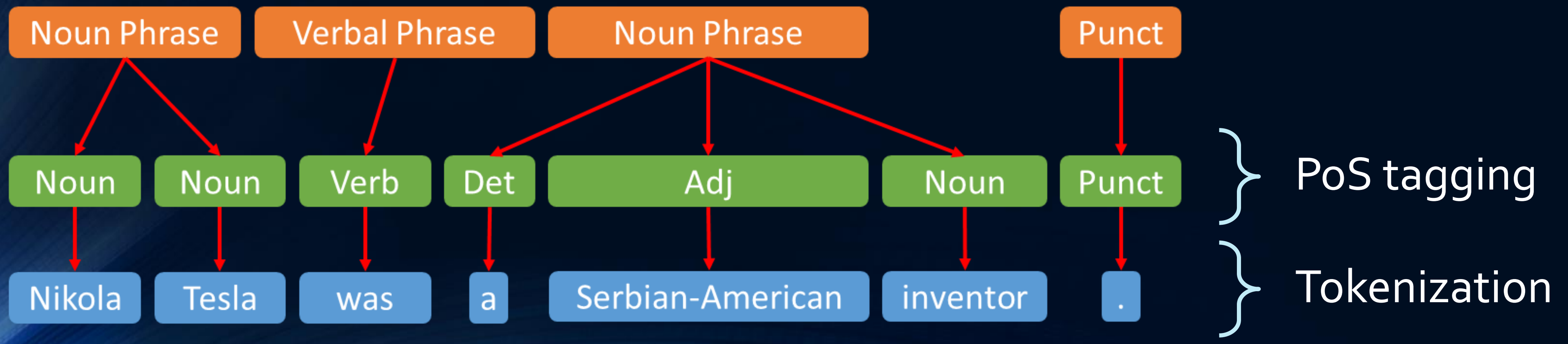
Syntactic Analysis

Analyse the sentence structure.



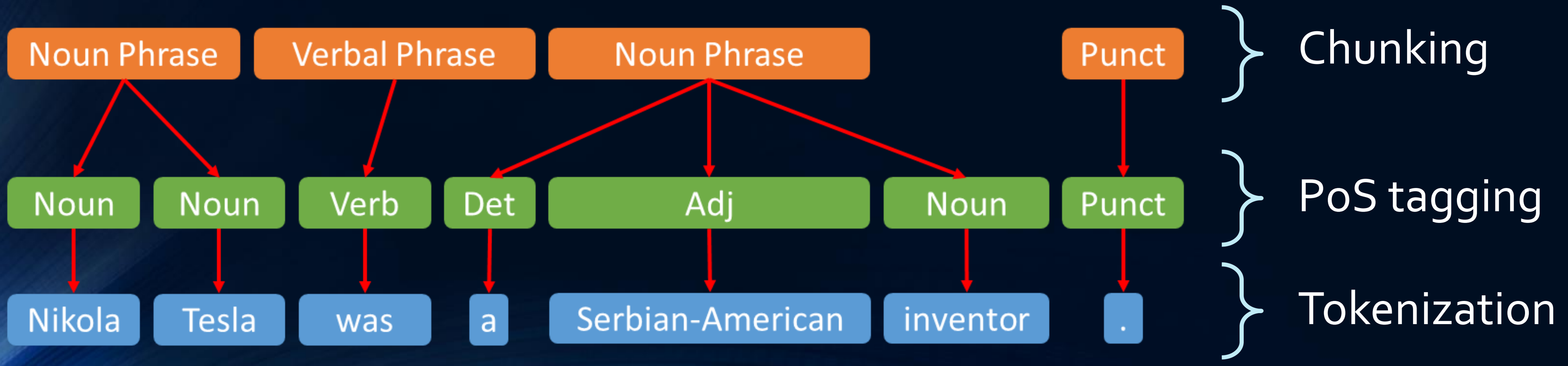
Syntactic Analysis

Analyse the sentence structure.



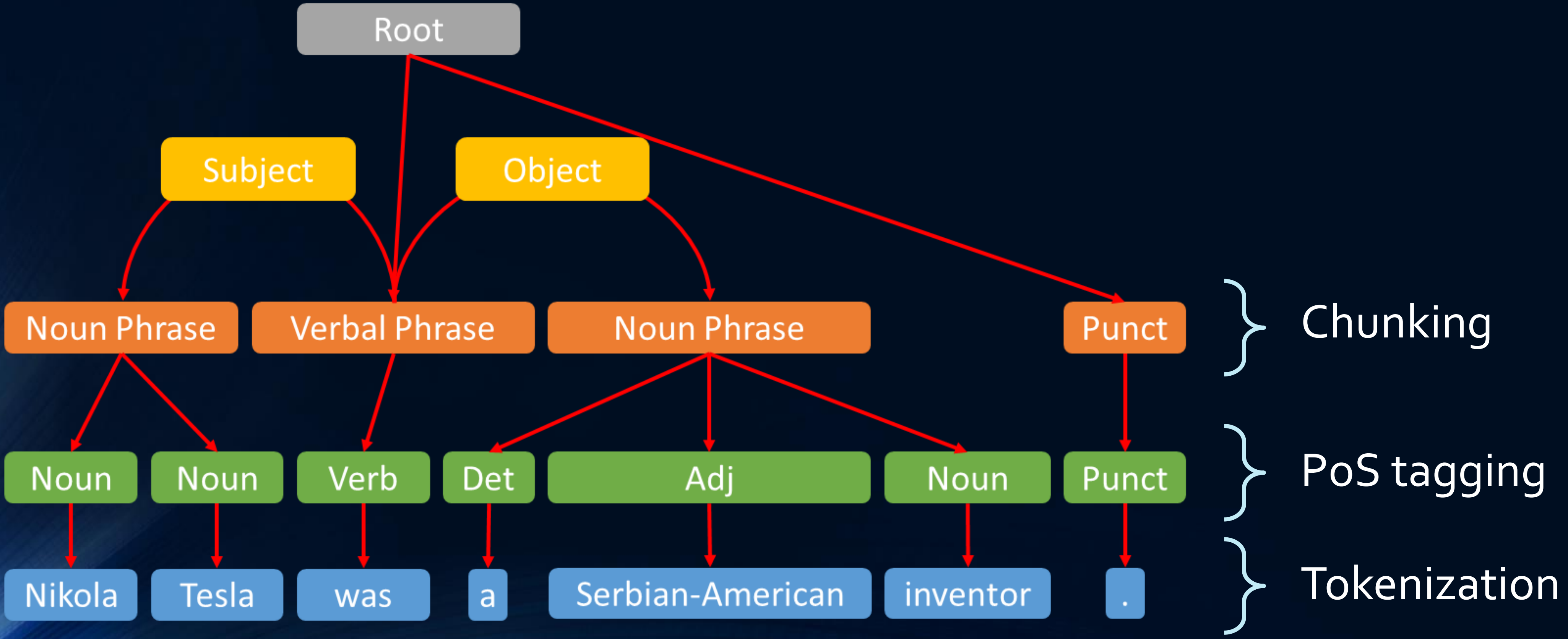
Syntactic Analysis

Analyse the sentence structure.



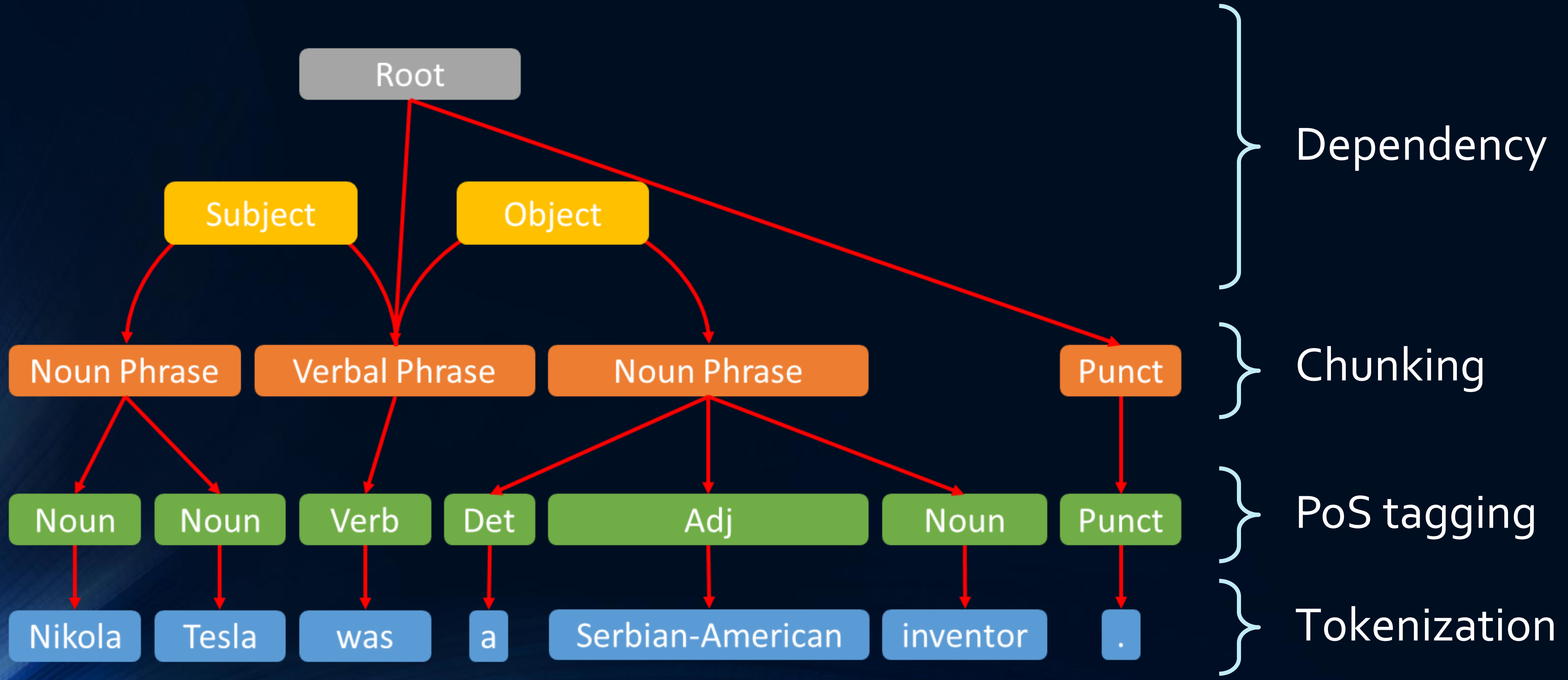
Syntactic Dependency Analysis

Link the sentence structure.



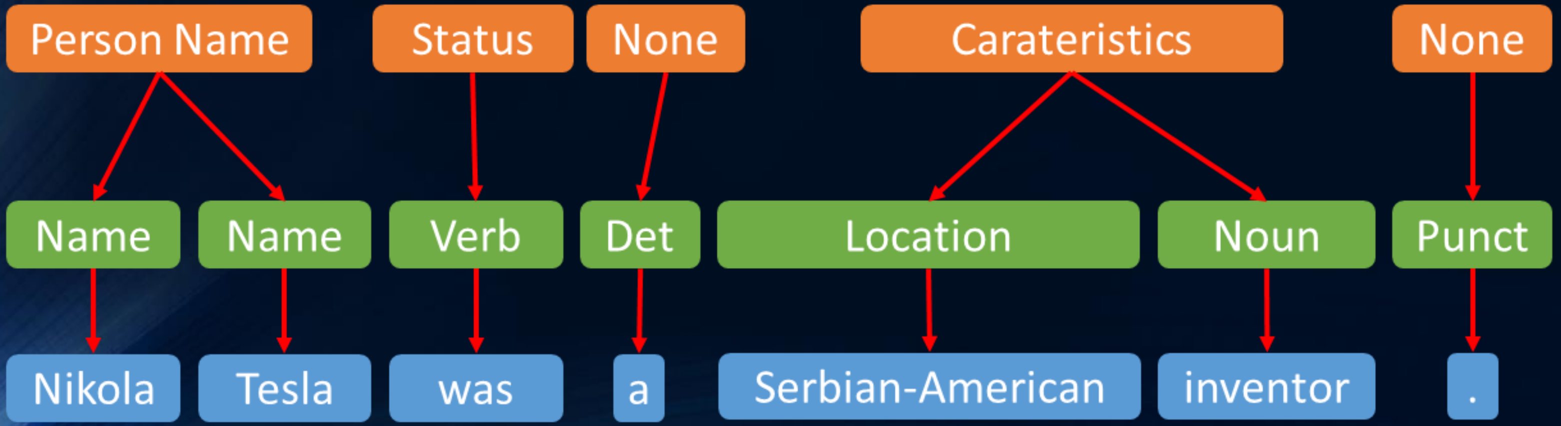
Syntactic Dependency Analysis

Link the sentence structure.



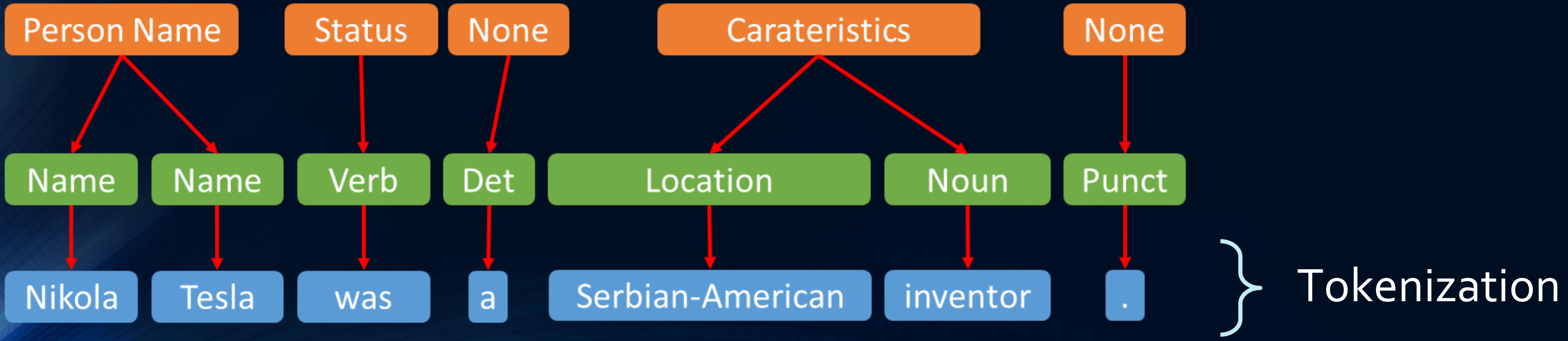
Semantic Analysis

Add concepts to the structure.



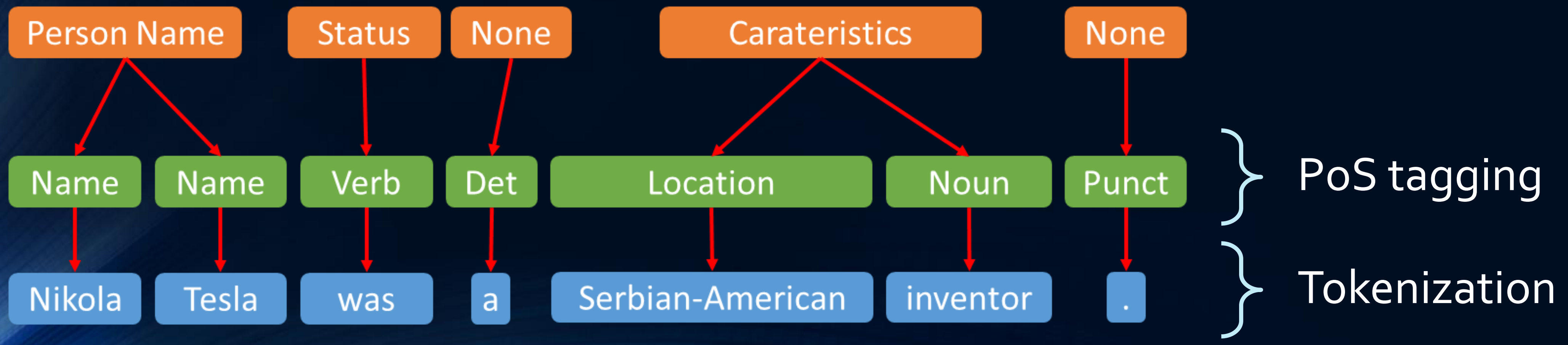
Semantic Analysis

Add concepts to the structure.



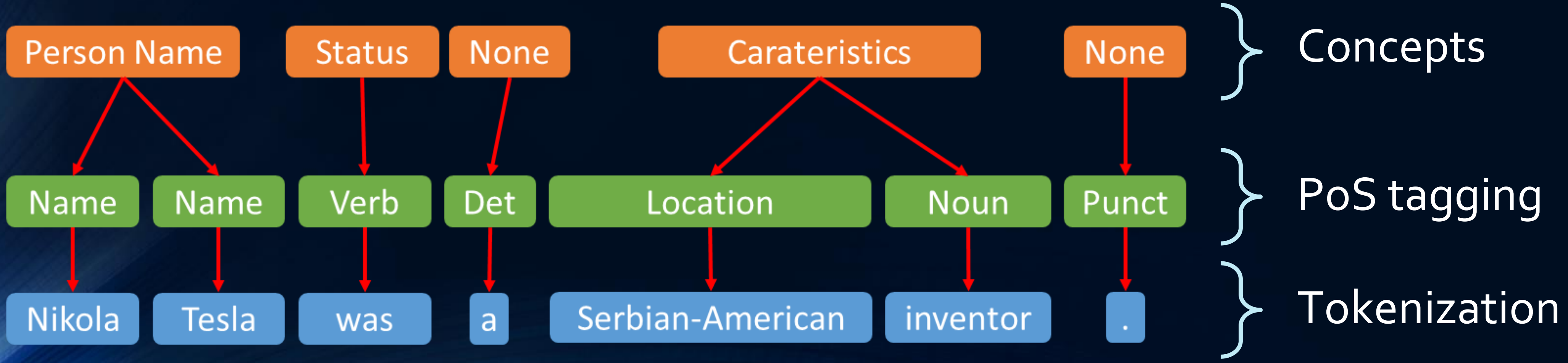
Semantic Analysis

Add concepts to the structure.



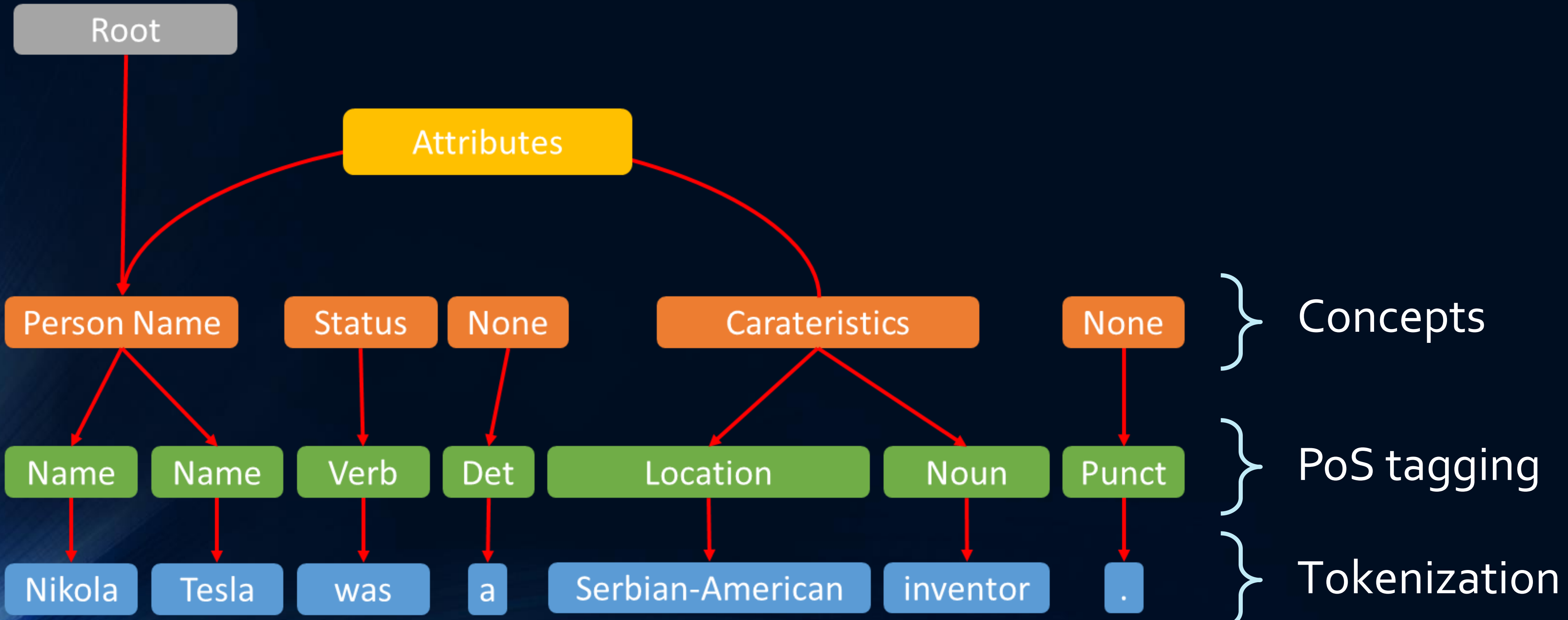
Semantic Analysis

Add concepts to the structure.



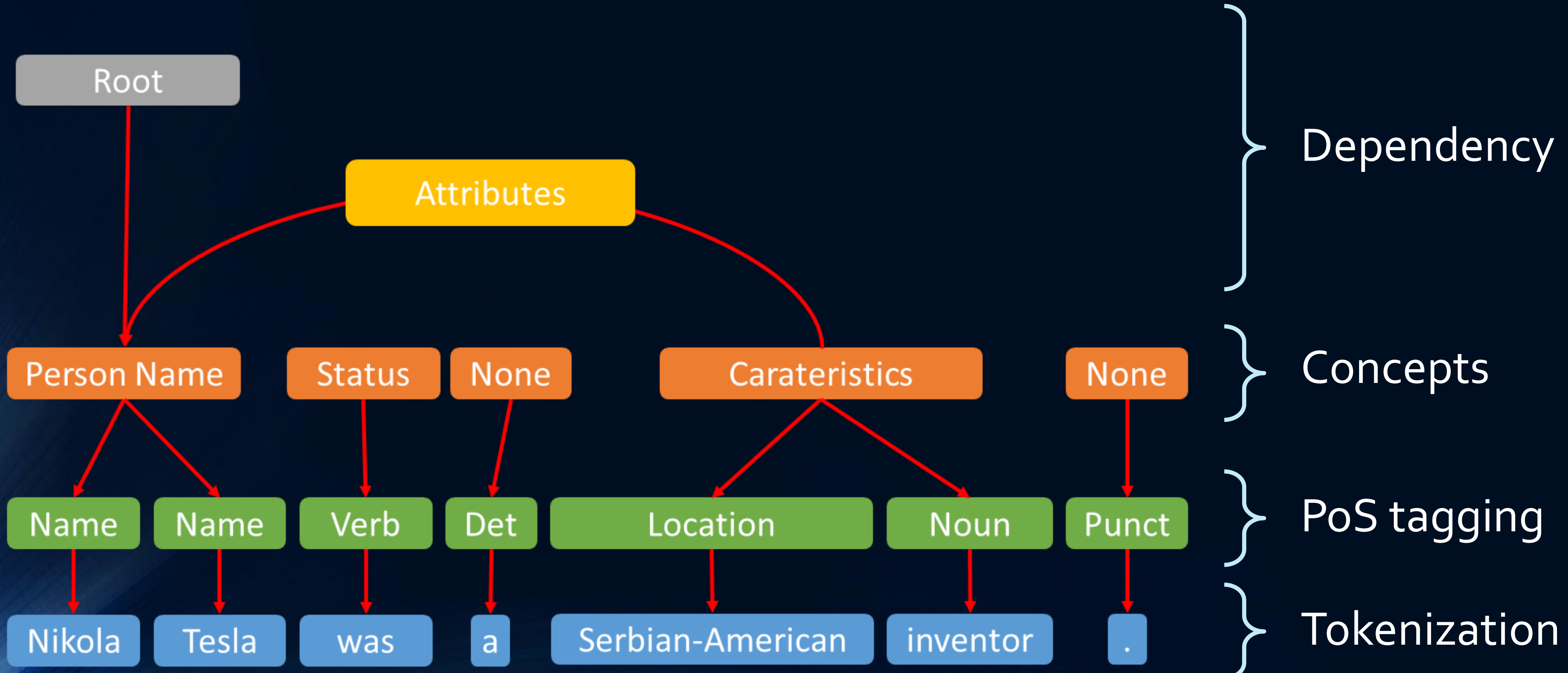
Semantic Dependency Analysis

Link the conptual structure.



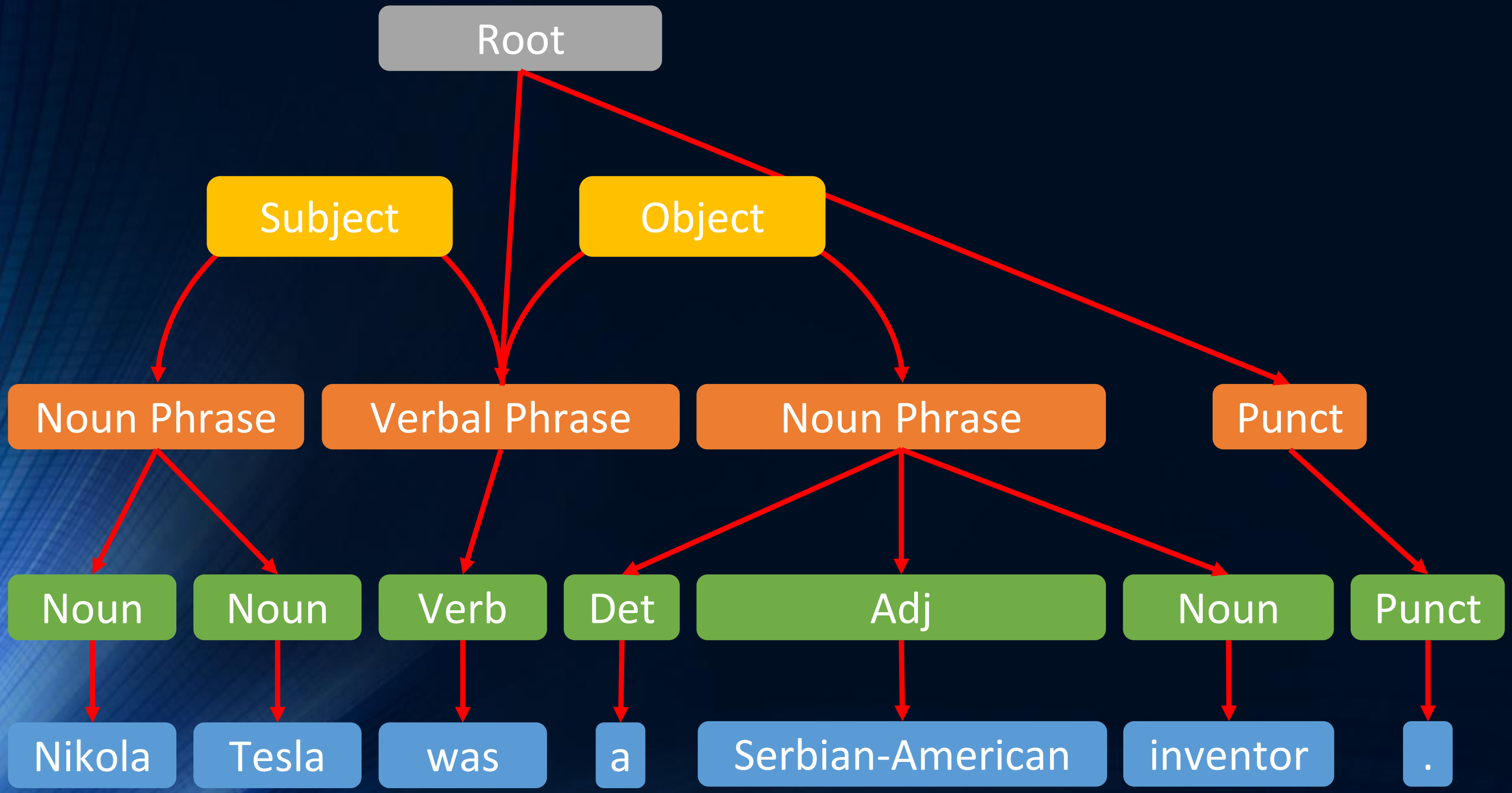
Semantic Dependency Analysis

Link the conptual structure.



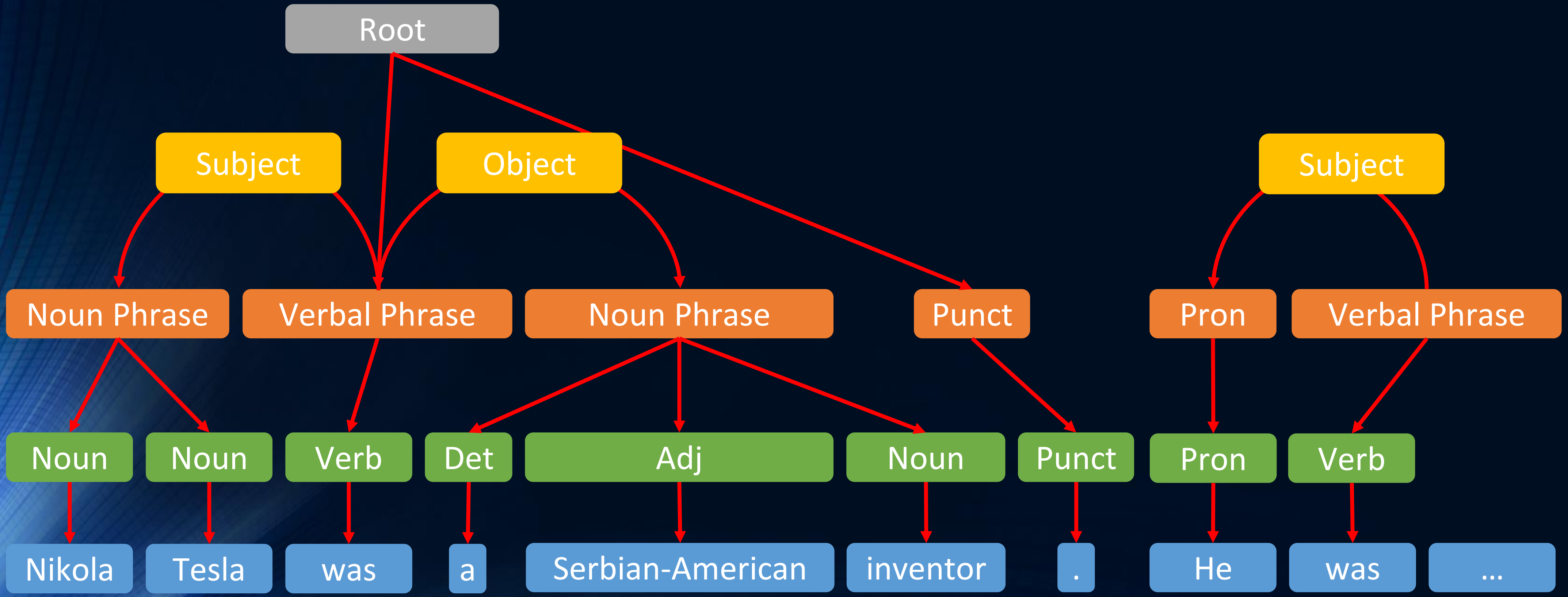
Coreference resolution

Find the right word it refers.



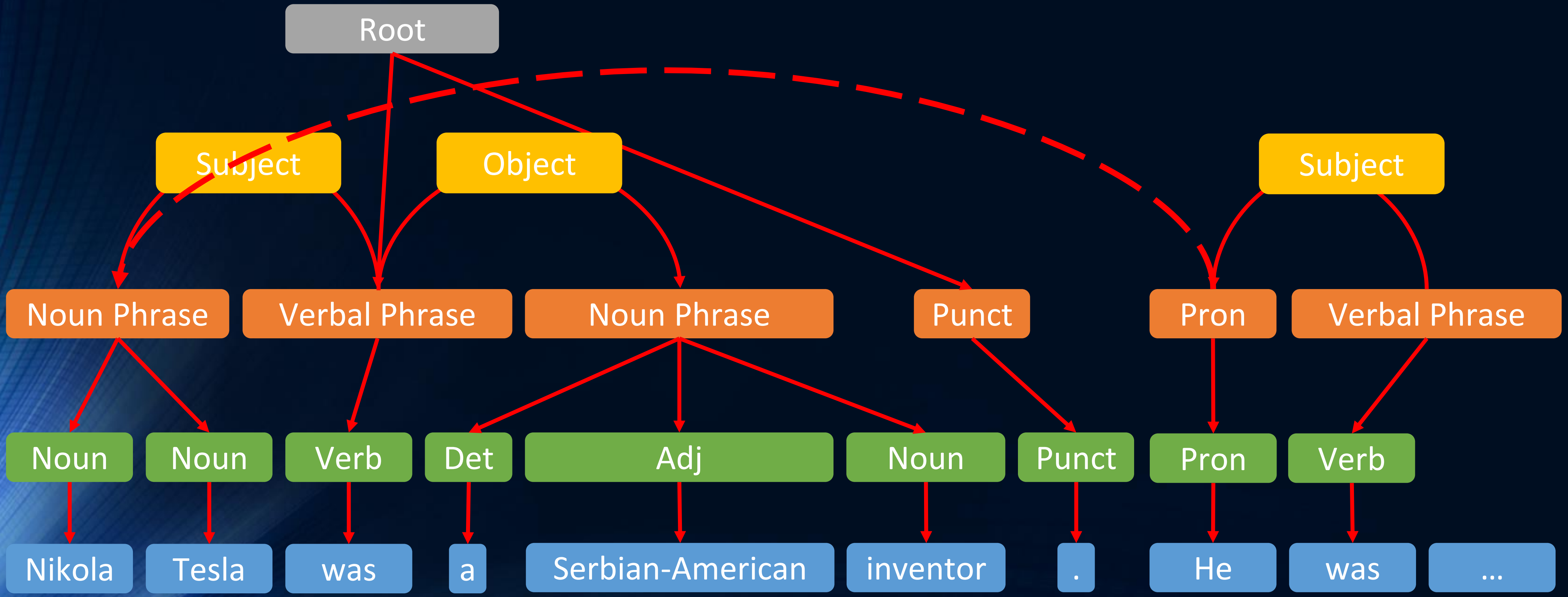
Coreference resolution

Find the right word it refers.



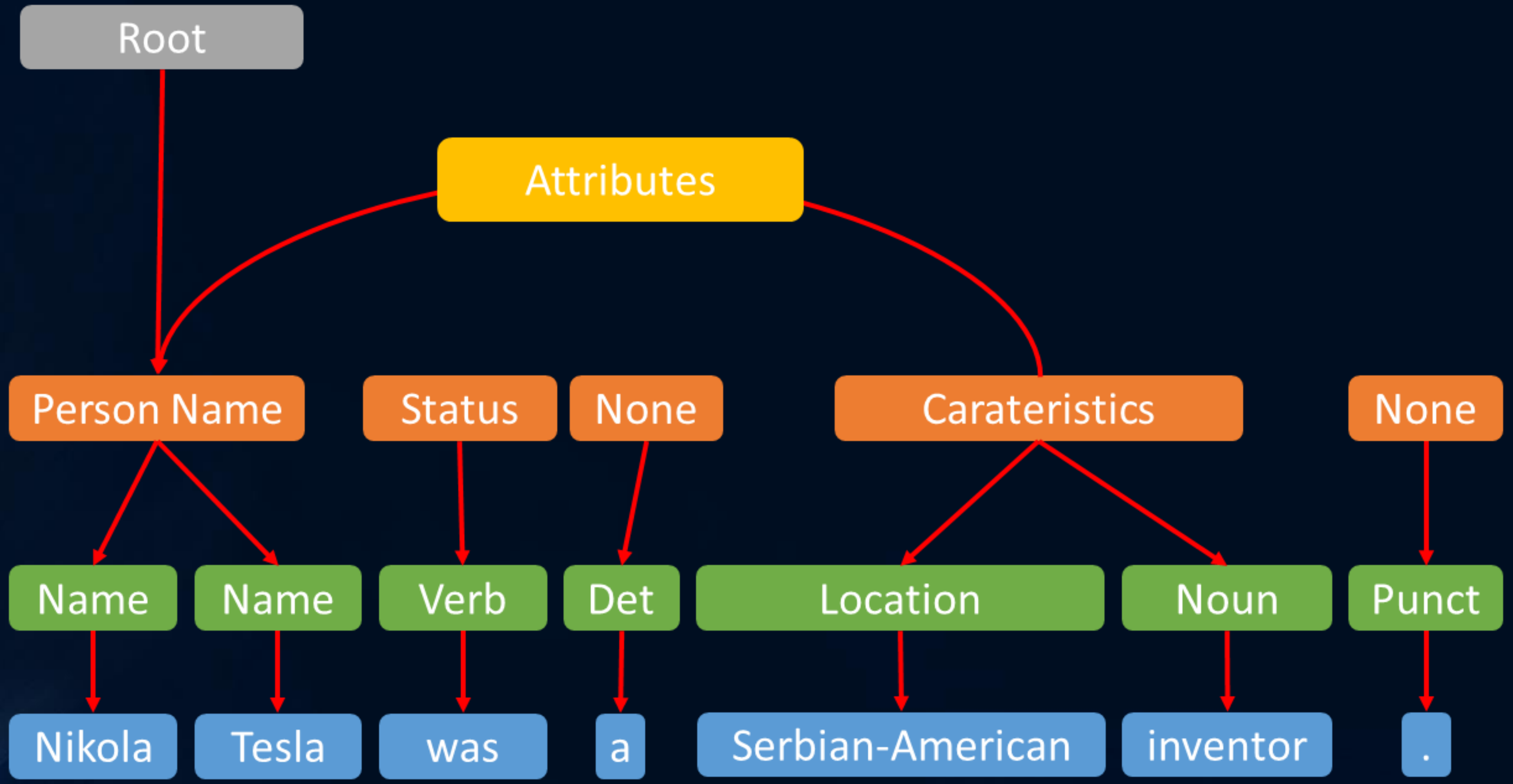
Coreference resolution

Find the right word it refers.



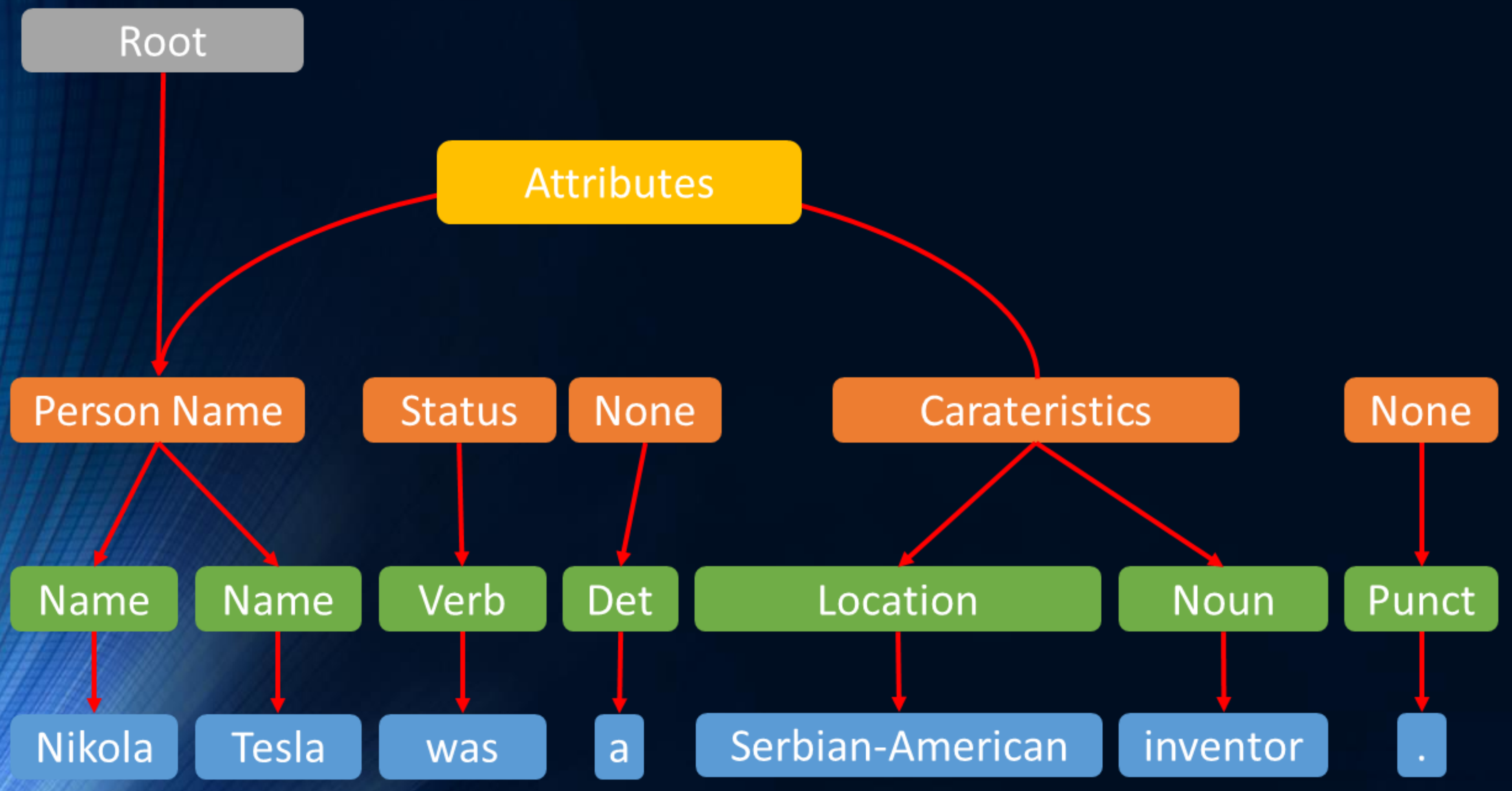
Information extraction

Extract logical relations or representations.



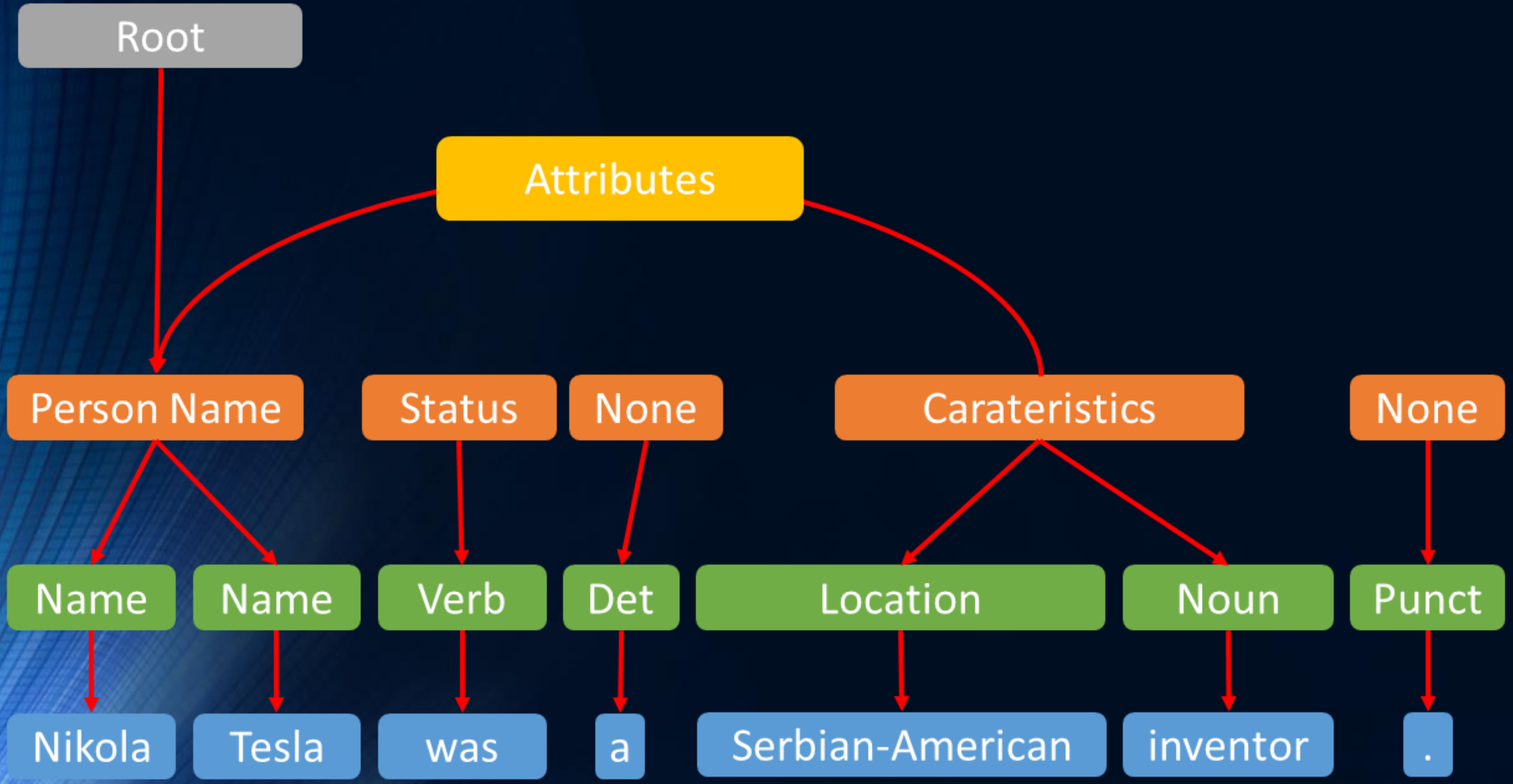
Information extraction

Extract logical relations or representations.



Information extraction

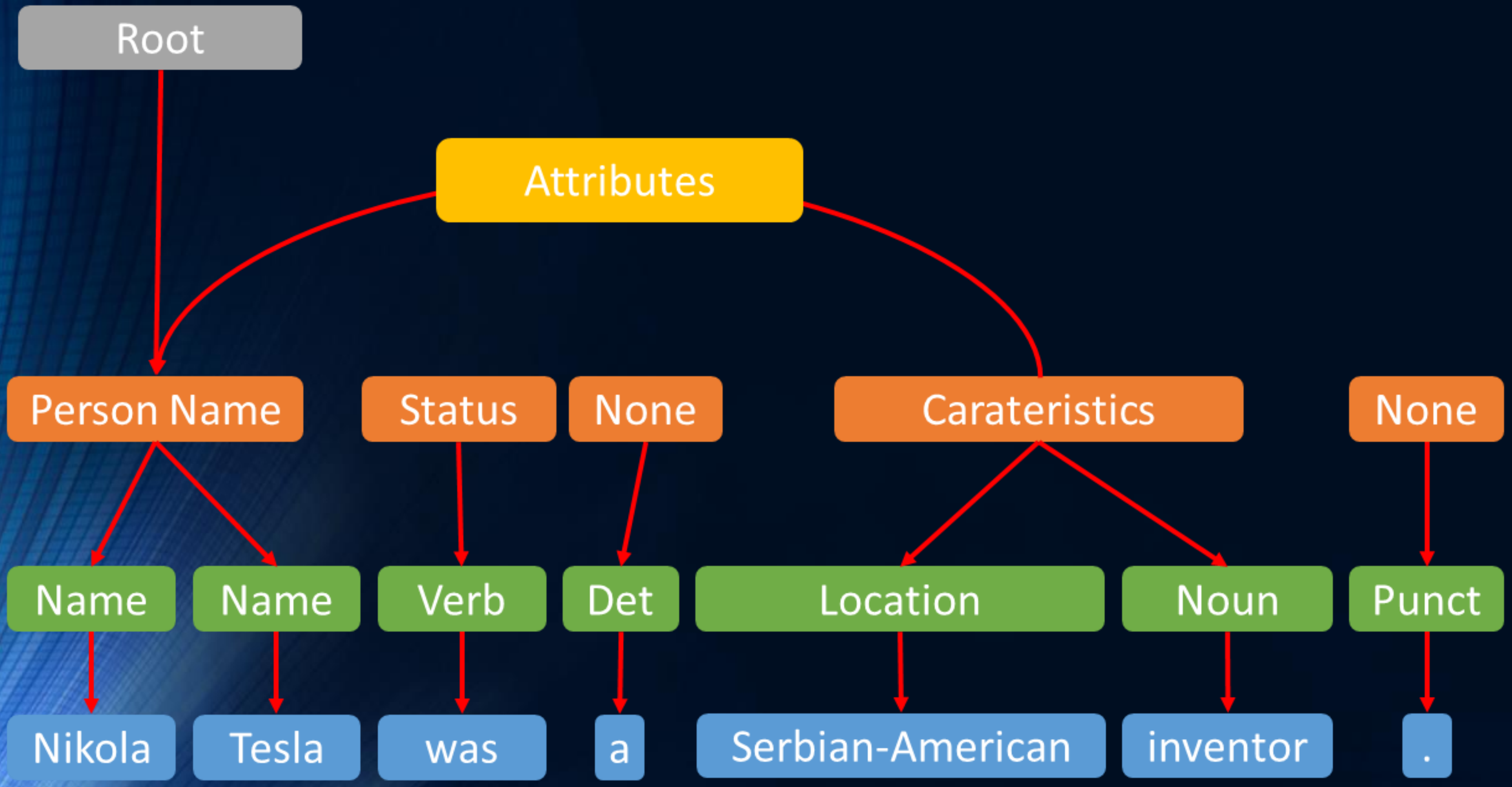
Extract logical relations or representations.



Nationality:
{"Nikolas Tesla",
"Serbian-American"}

Information extraction

Extract logical relations or representations.



Nationality:
{"Nikolas Tesla",
"Serbian-American"}

Job:
{"Nikolas Tesla",
"Inventor"}

Applications

- Question Answering
- Textual Entailment
- Reasoning
- Knowledge Base
- Semantic Web
- Natural Language Generation
- Speech Synthesis
- Dialogue Systems

Question Answering (QA)

Jeopardy!

- Nationality:
  {“Nikolas Tesla”, “Serbian-American”}
- Job:
  {“Nikolas Tesla”, “Inventor”}

Question Answering (QA)

Jeopardy!

- Nationality:
  {“Nikolas Tesla”, “Serbian-American”}
- Job:
  {“Nikolas Tesla”, “Inventor”}

→ What was Nikolas Tesla? **Inventor**

Textual Entailment (TE)

Verify assumptions

- Nationality:
  {“Nikolas Tesla”, “Serbian-American”}
- Job:
  {“Nikolas Tesla”, “Inventor”}

Textual Entailment (TE)

Verify assumptions

- Nationality:
  {“Nikolas Tesla”, “Serbian-American”}
- Job:
  {“Nikolas Tesla”, “Inventor”}

➡ Did Nikolas Tesla lived in the US? **Yes**

Reasoning

Induce knowledge from what we already knows.

- Nationality:
  {“Nikolas Tesla”, “Serbian-American”}
- Job:
  {“Nikolas Tesla”, “Inventor”}

Reasoning

Induce knowledge from what we already knows.

- Nationality:
  {“Nikolas Tesla”, “Serbian-American”}
- Job:
  {“Nikolas Tesla”, “Inventor”}

→ Nikolas Tesla did not lived in the Classical Ages

Knowledge bases construction

It aims to create a fact collection using semantic.

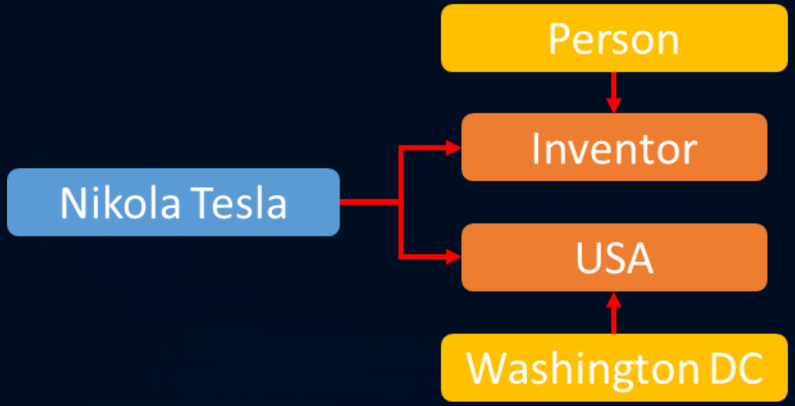
- Nationality:
  {"Nikolas Tesla", "Serbian-American"}
- Job:
  {"Nikolas Tesla", "Inventor"}

Knowledge bases construction

It aims to create a fact collection using semantic.

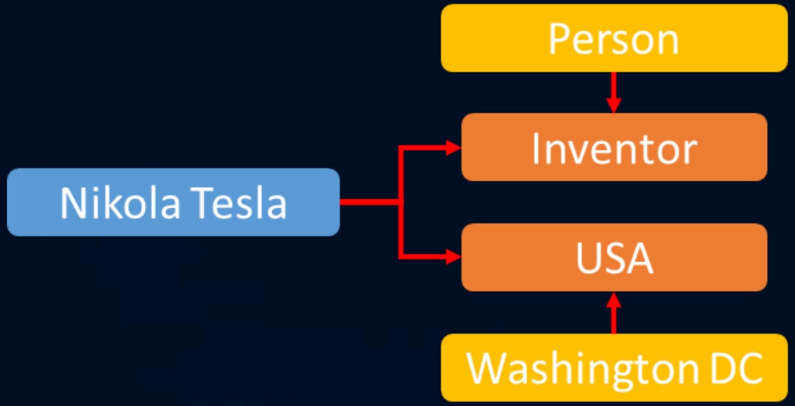
- Nationality:
  {“Nikolas Tesla”, “Serbian-American”}
- Job:
  {“Nikolas Tesla”, “Inventor”}

Nikolas Tesla



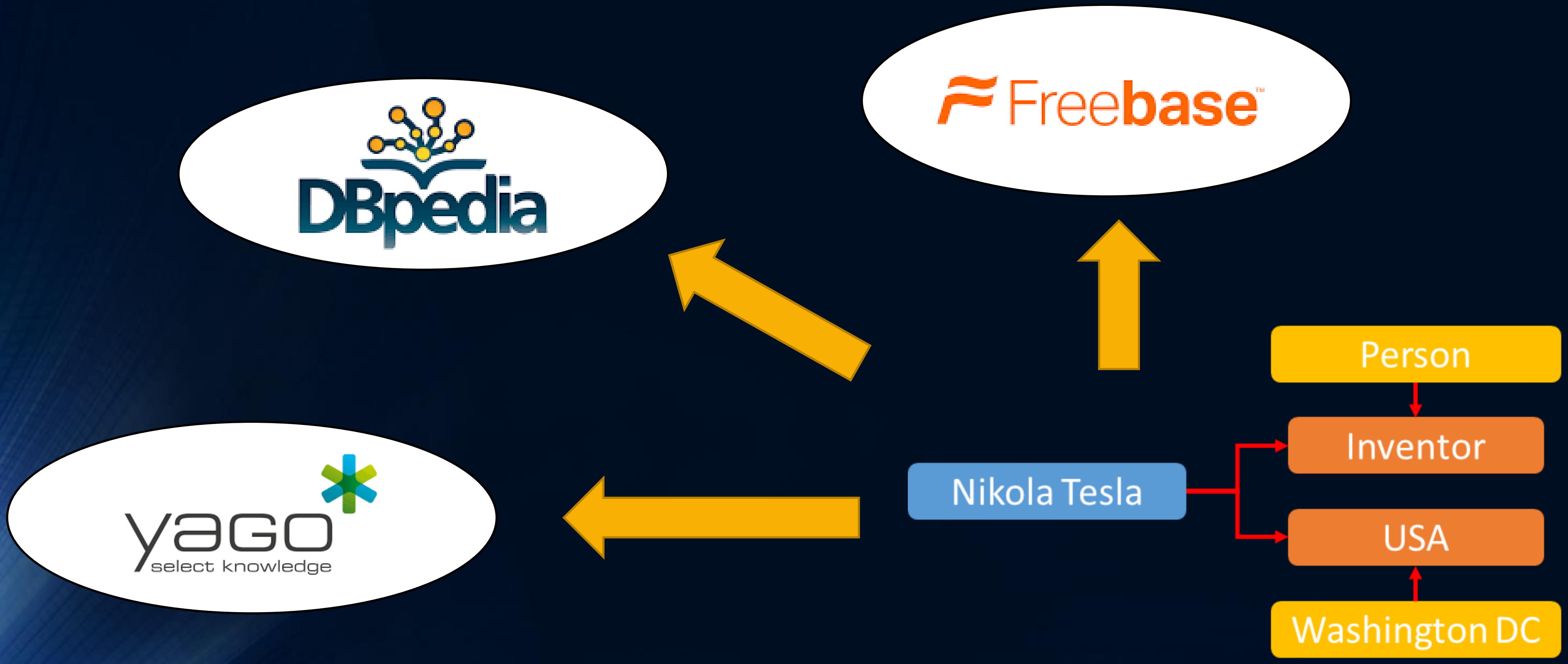
Semantic web

It is a set of computer-readable KB (RDF, SparQL, etc.) from Internet.



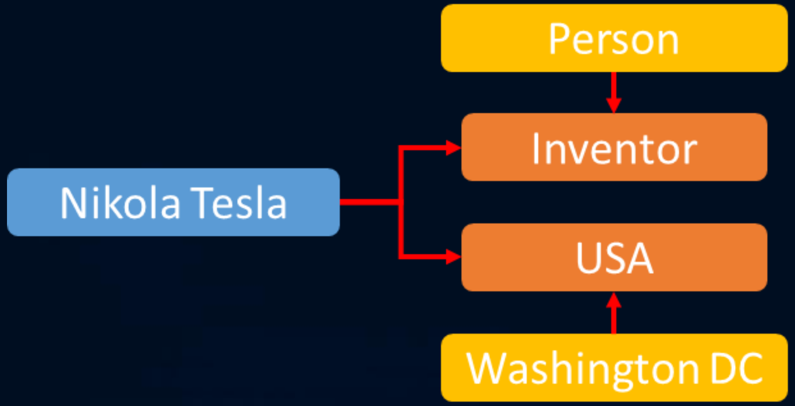
Semantic web

It is a set of computer-readable KB (RDF, SparQL, etc.) from Internet.



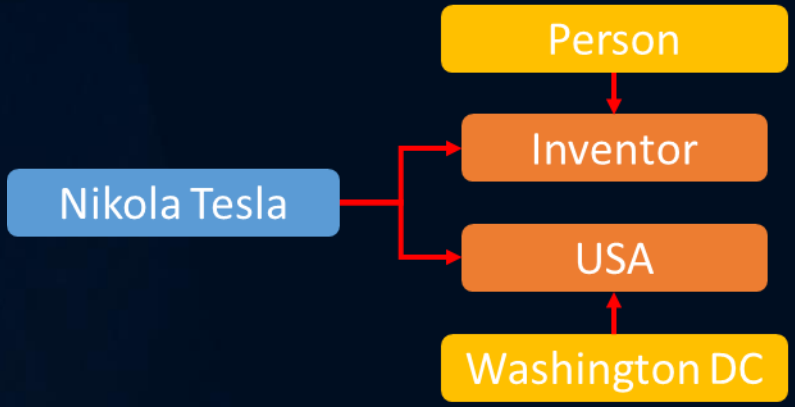
Natural Language Generation

From KB generate natural language.



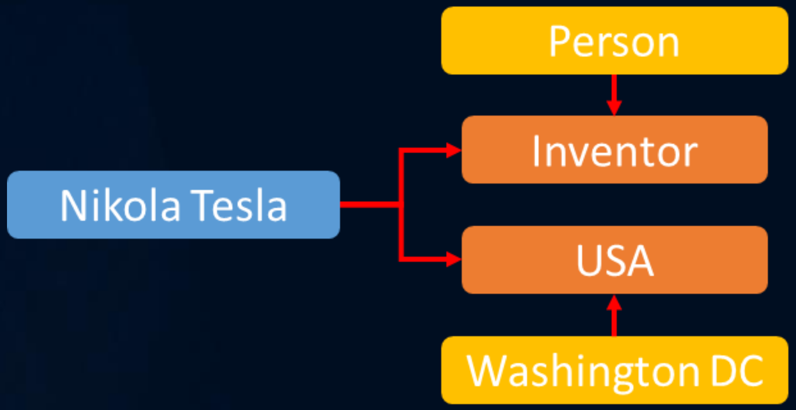
Natural Language Generation

From KB generate natural language.



Natural Language Generation

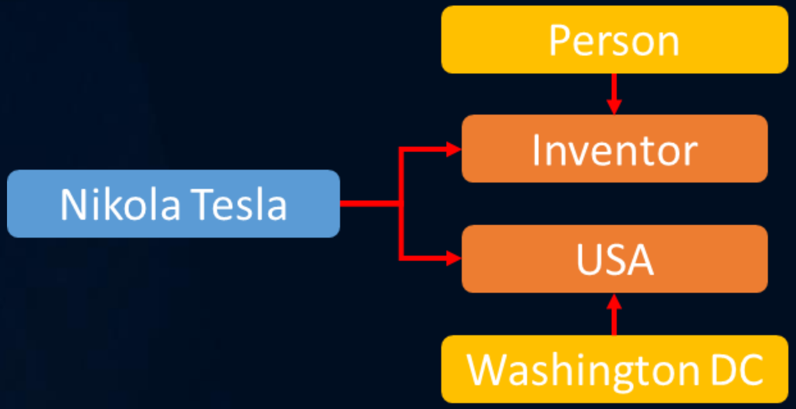
From KB generate natural language.



Nikolas Tesla is an American inventor

Speech Synthesis (Text-to-Speech)

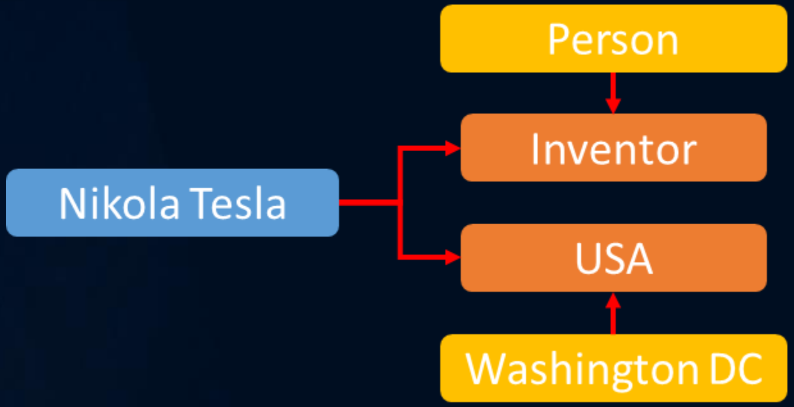
Generate an audio from text



Nikolas Tesla is an American inventor

Speech Synthesis (Text-to-Speech)

Generate an audio from text



Person

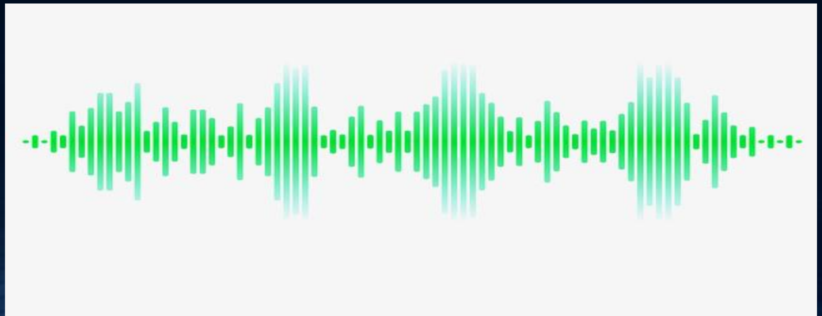
Inventor

Nikola Tesla

USA

Washington DC

Nikolas Tesla is an American inventor

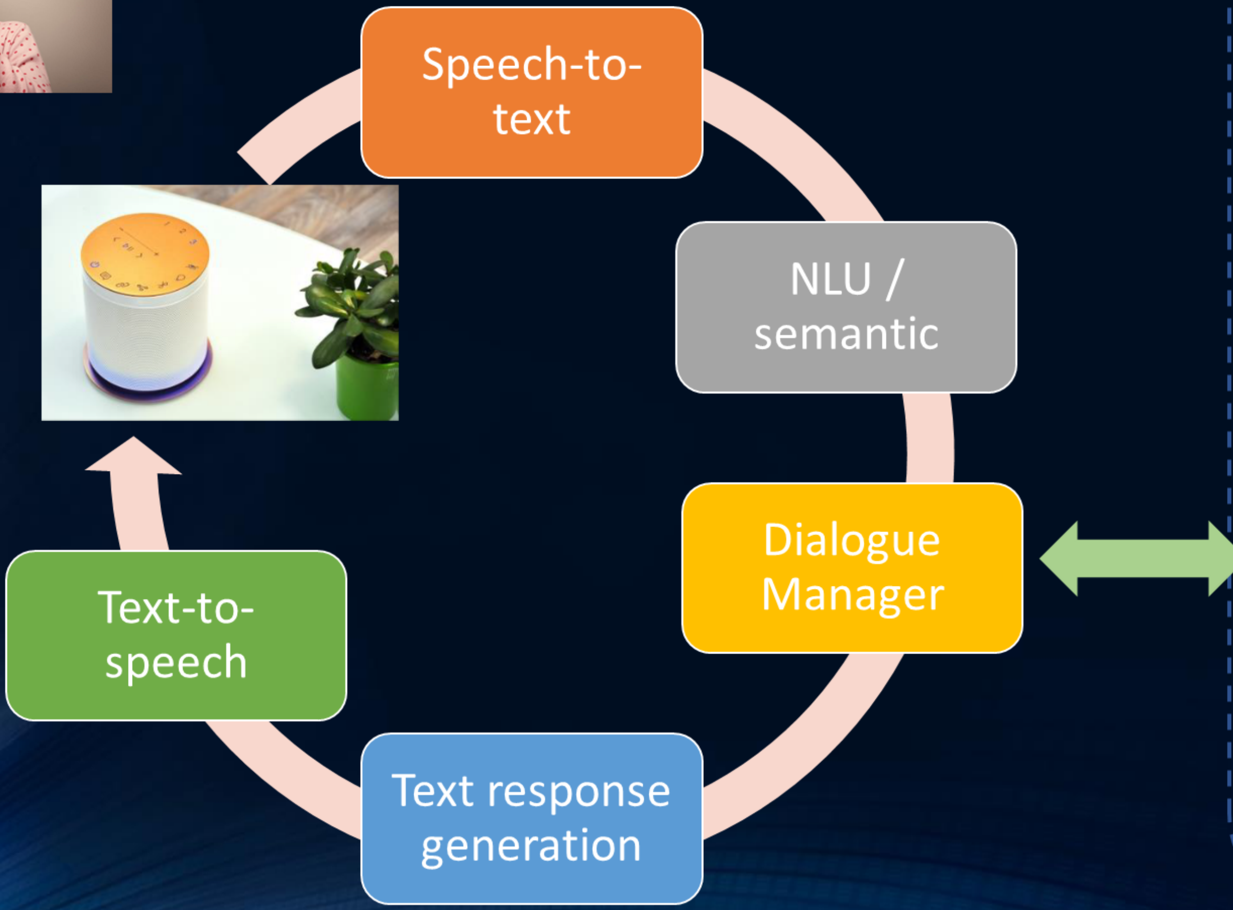


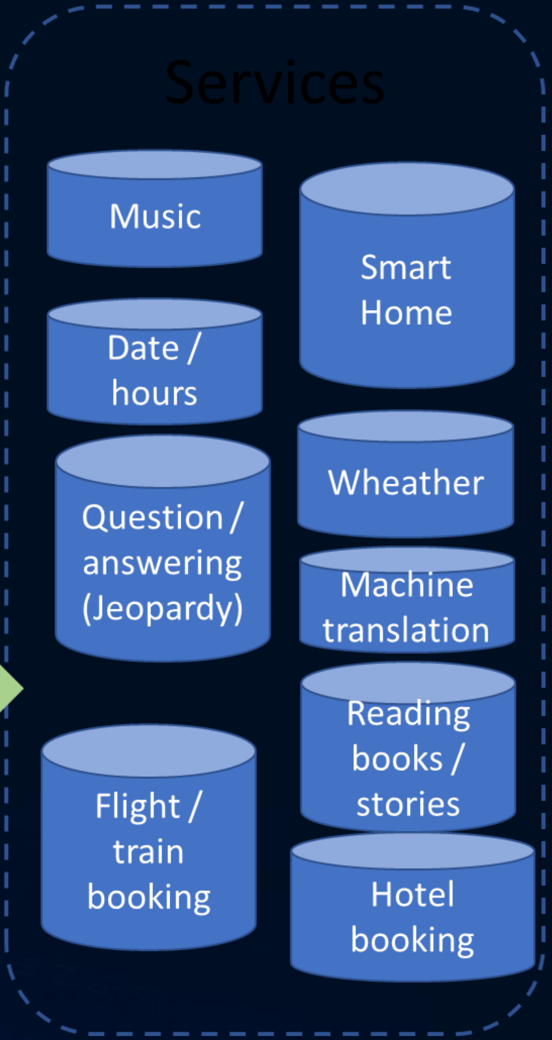
Dialogue Systems

Full process: Analysis and Applications

Dialogue Systems

Full process: Analysis and Applications





Approaches

- Rules
- Statistical / Machine Learning
- Deep Learning

Example NER

- Aim: identify “Name” and “Date” from this example

Nikola Tesla was born in 1856

Rule-Based Approach

- Design rules manually done by a Human expert.

Nikola Tesla was born in 1856

Example of rules:
"Name": two Words with capital letters before the sequence of words "was born"
"Date": 4 digits numbers after the sequence of words "born in"

Pros:

- Easy to implements
- Easy to debug
- Easy to explain (tractable)

Cons:

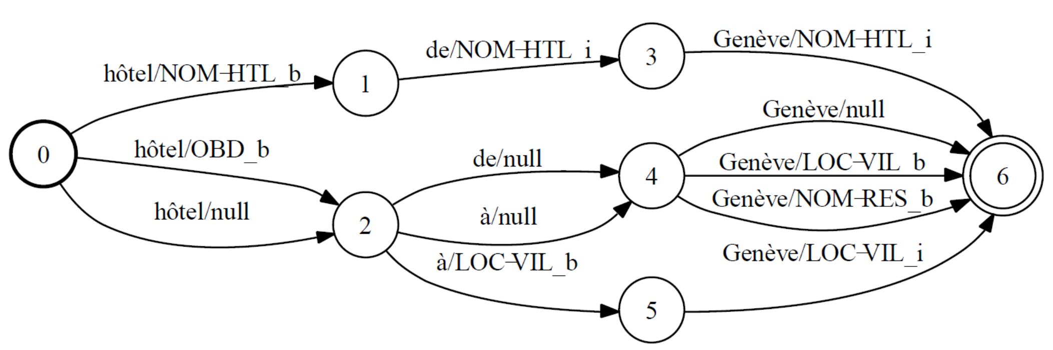
- Can't deal with unseen cases (here the middle name)
- Manual rules need expertise and time
- Ad Hoc application (e.g. not generalizable)

Statistical / Machine learning

- Need training data annotated by expert

Nikola Tesla was born in 1856

Generative (or graphical) models:
Finite State Machines



Pros:
- State of the Art in 2000's
- Tractable
- Can be mixed with manual rules

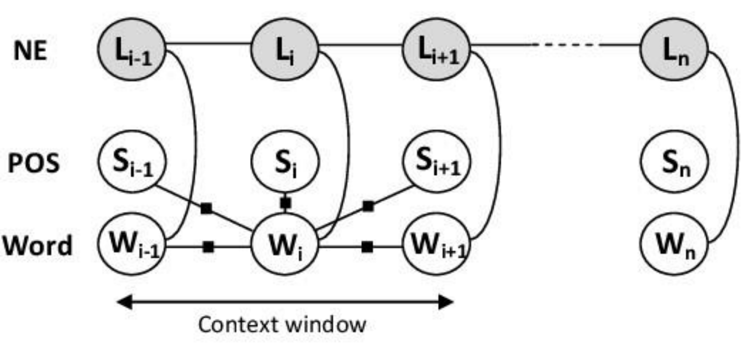
Cons:
- Can't deal with unseen cases
- Ad Hoc application (e.g. not generalizable)

Statistical / Machine learning

- Need training data annotated by expert

Nikola Tesla was born in 1856

Discriminative models:
Conditional Random Fields



Pros:
- State of the Art early 2010's
- Can deal with unseen cases
- More generalizable

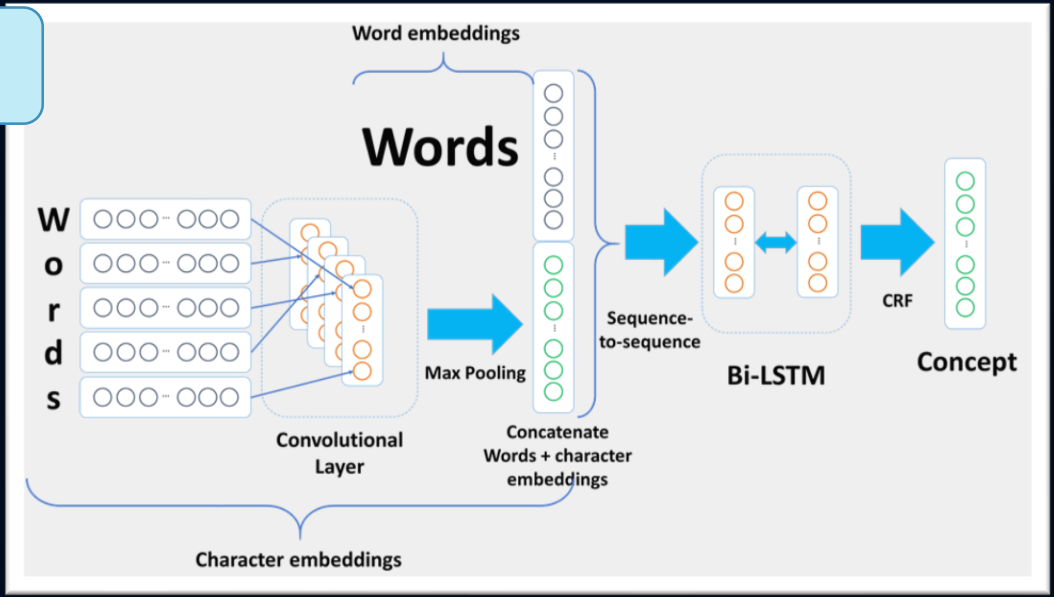
Cons:
- Not tractable
- Can't be mixed with manual rules
- Need of clean training data

Deep learning

- Need a lot of training data annotated

Nikola Tesla was born in 1856

Discriminative models



Pros:

- State of the Art end of 2010's
- Can deal with unseen cases
- Generalizable at will

Cons:

- Definitely NOT tractable
- Can't be mixed with manual rules
- Need of huge amount of training data

Conclusion

- Introduction to NLP
- Preprocessing
- Processing data through several analysis
- Applications examples
- Methodology
  - Rule-based, ML, DL