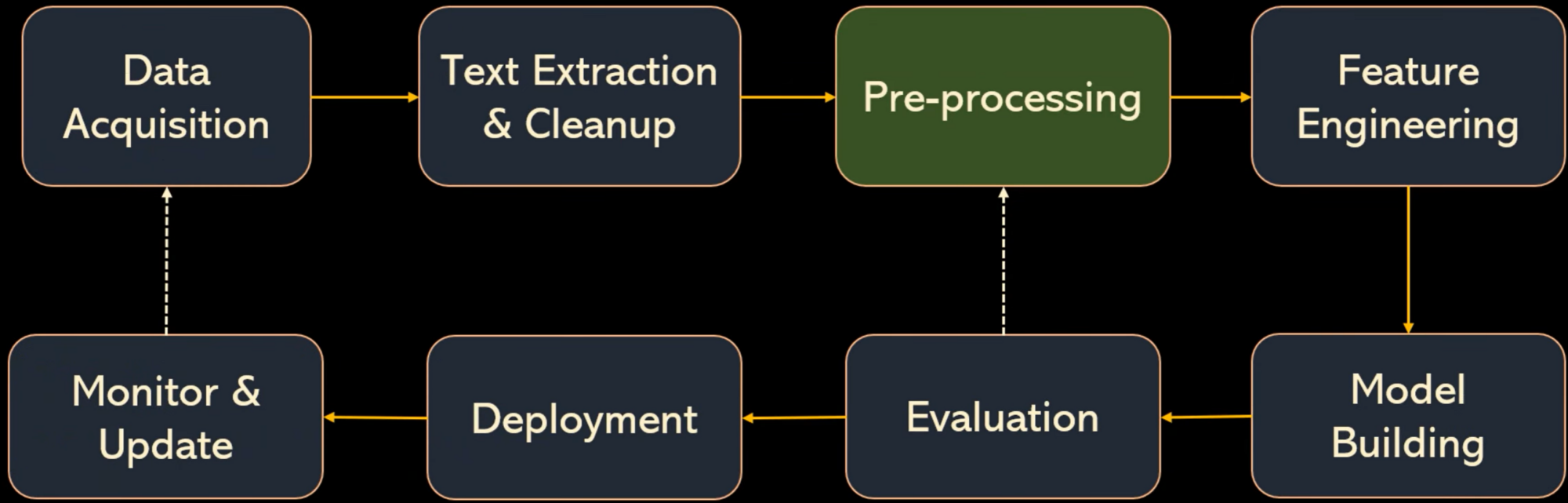
Tokenization in Spacy


Data Acquisition
Text Extraction & Cleanup
Pre-processing
Feature Engineering
Model Building
Evaluation
Deployment
Monitor & Update

Credits: Pra

Pre-Processing

"camtasia closing and crashing during ba
another post complaining about somethi
times before: During video production (tl
production as that's what I usually use),
The crash is basically.."

Sentence Tokenization

"camtasia closing and crashing during

“So here’s another post complaining

“The crash is basically.."

Pre-Processing

"camtasia closing and crashing during ba
another post complaining about somethi
times before: During video production (t
production as that's what I usually use),
The crash is basically.."

Sentence Tokenization

"camtasia closing and crashing during

"So here's another post complaining

"The crash is basically.."

Word Tokenization

camtasia

closing

crashing

Pre-Processing

"camtasia closing and crashing during ba
another post complaining about somethi
times before: During video production (t
production as that's what I usually use),
The crash is basically.."

Sentence Tokenization

"camtasia closing and crashing during

"So here's another post complaining

"The crash is basically.."

Word Tokenization

camtasia | closing | crashing

Stemming, Lemmatization

camtasia | clo

Tokenization is a process of splitting text into meaningful segments


WHAT'S THE BIG DEAL IN
TOKENIZATI
imgflip.com

Dr. strange ordered samosas, ravioli etc. for his lunch.

Dr. strange ordered samosas, ravioli etc. for his lunch.

Sentence # 1	Dr
Sentence # 2	strange ordered samosas, ravioli etc
Sentence # 3	for his lunch

Dr. strange ordered samosas, ravioli etc. for his lunch.

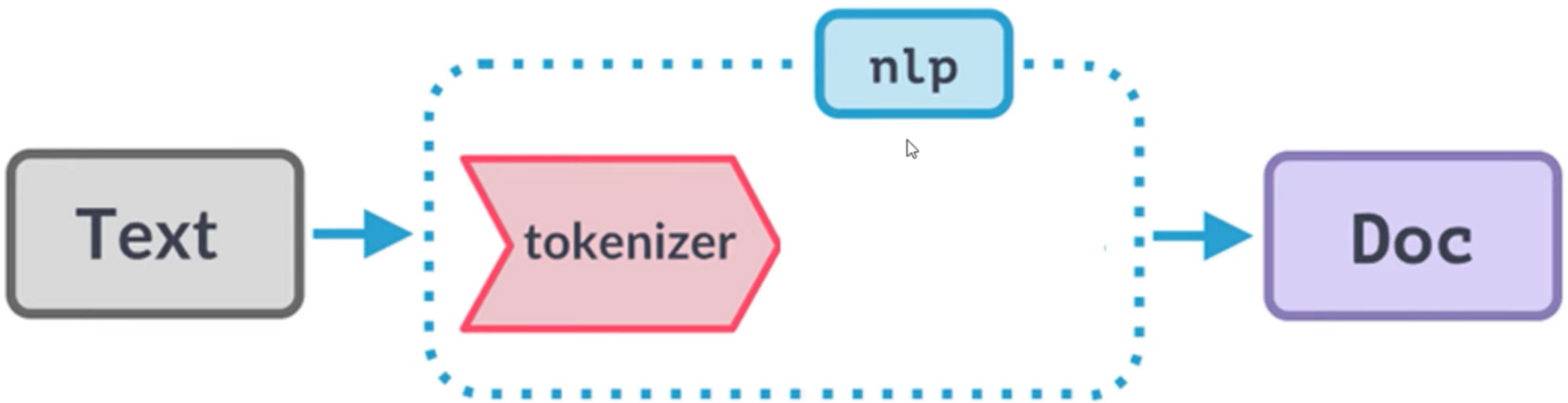
Sentence # 1	Dr
Sentence # 2	strange ordered samosas, ravioli etc
Sentence # 3	for his lunch

You see, it is not just about splitting a sentence by .

Or

Splitting words in a sentence by spaces

nlp = spacy.blank("en")



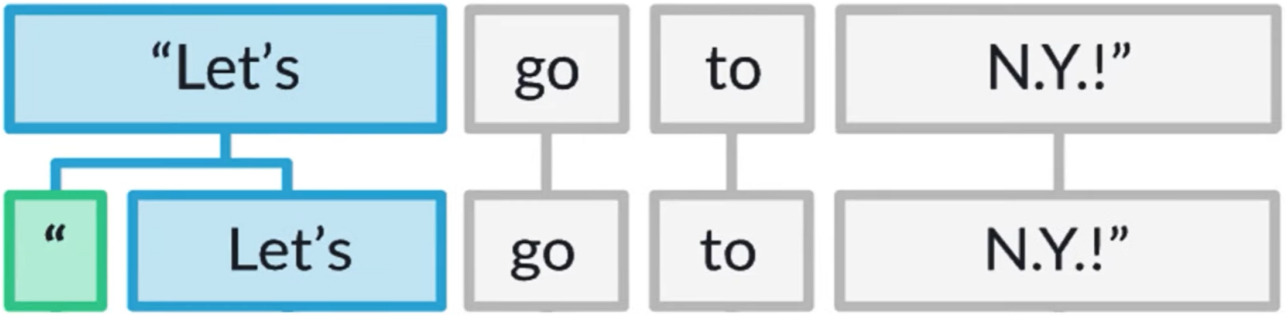
Credits: https://spacy.io/

“Let’s go to N.Y!”

“Let’s go to N.Y!”

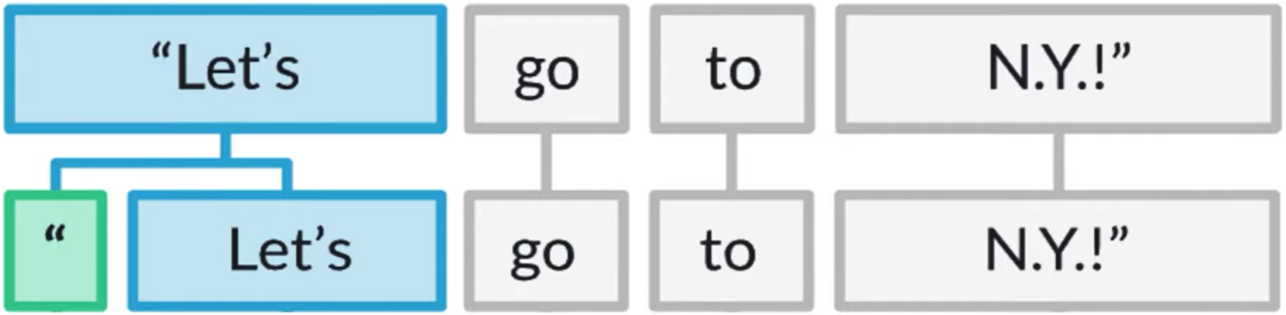
Split the words by spaces. i.e. text.split(‘ ‘)

“Let’s	go	to	N.Y.!”



PREFIX

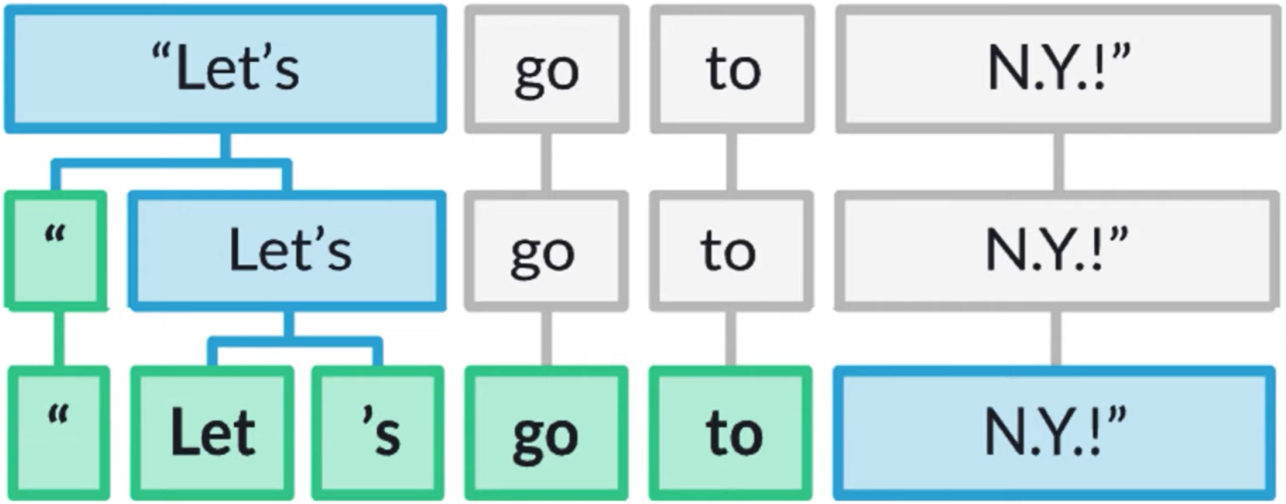
Credits: https://spacy.io/



PREFIX

$ (“

Credits: https://spacy.io/

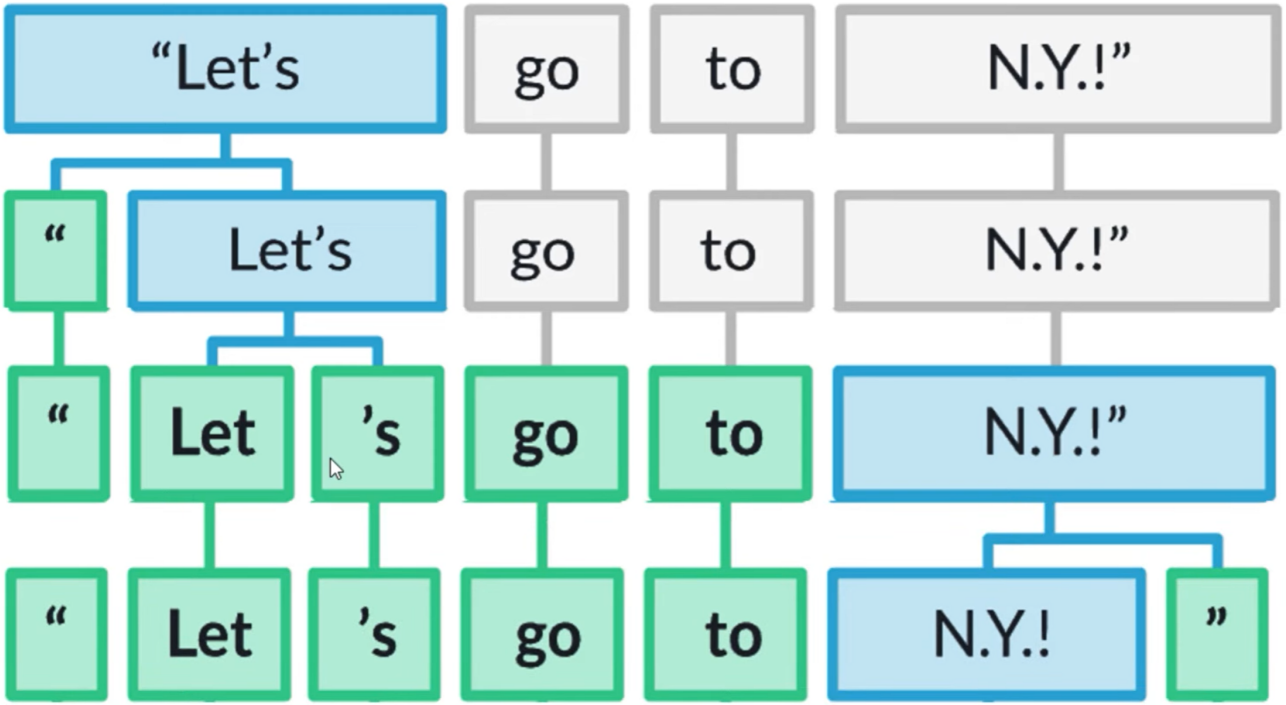


PREFIX

$ ("

EXCEPTION

Credits: https://spacy.io/



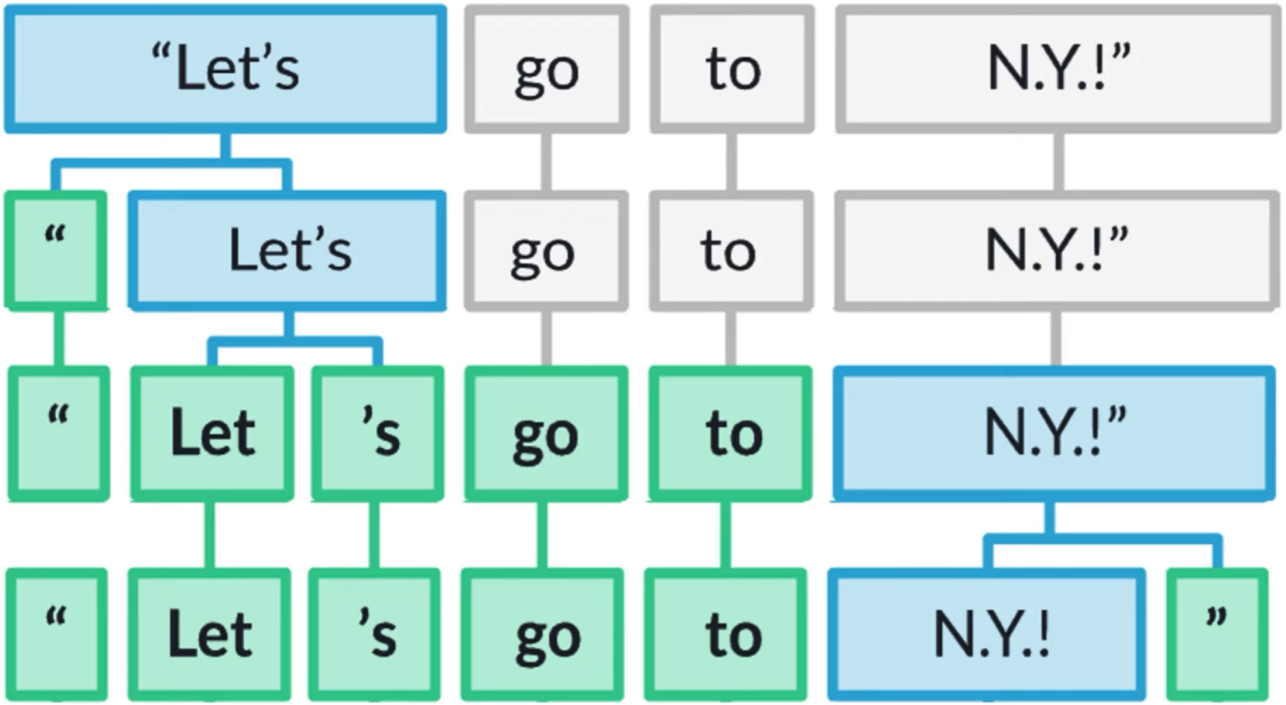
PREFIX

EXCEPTION

SUFFIX

$ (“

Credits: https://spacy.io/



PREFIX

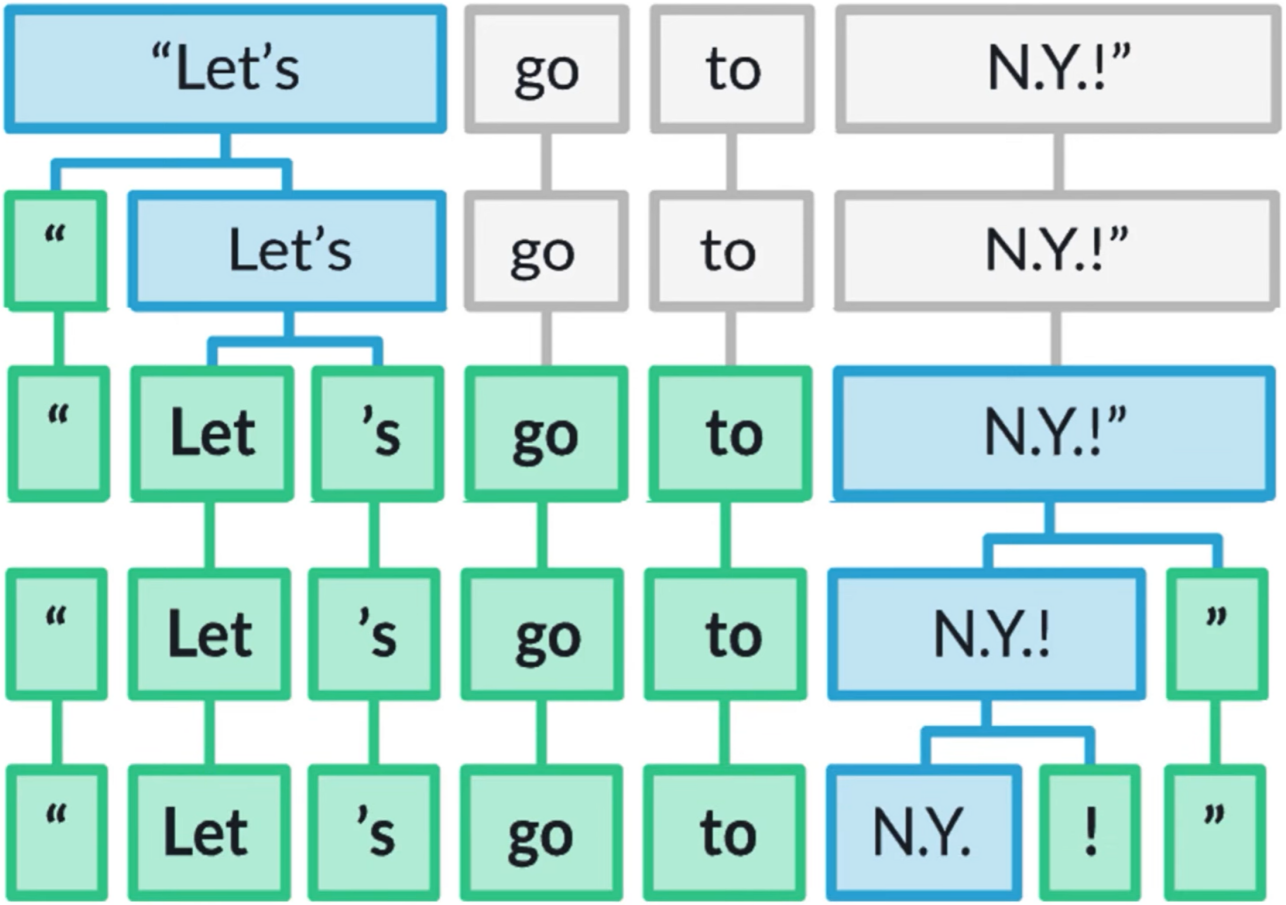
$ ("

EXCEPTION

SUFFIX

km) ! "

Credits: https://spacy.io/



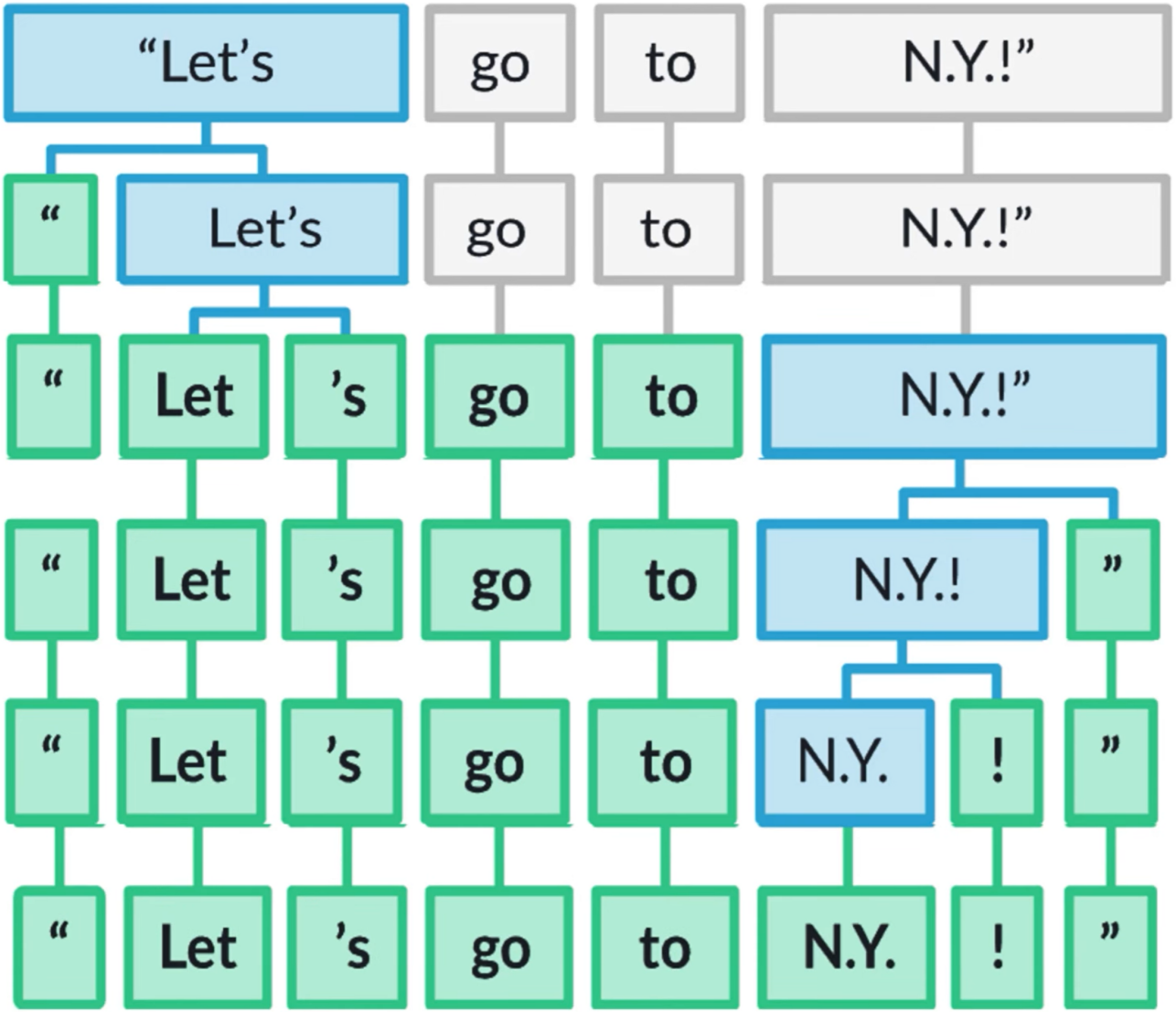
PREFIX $ ("

EXCEPTION

SUFFIX km) ! "

SUFFIX

Credits: https://spacy.io/



PREFIX $ ("

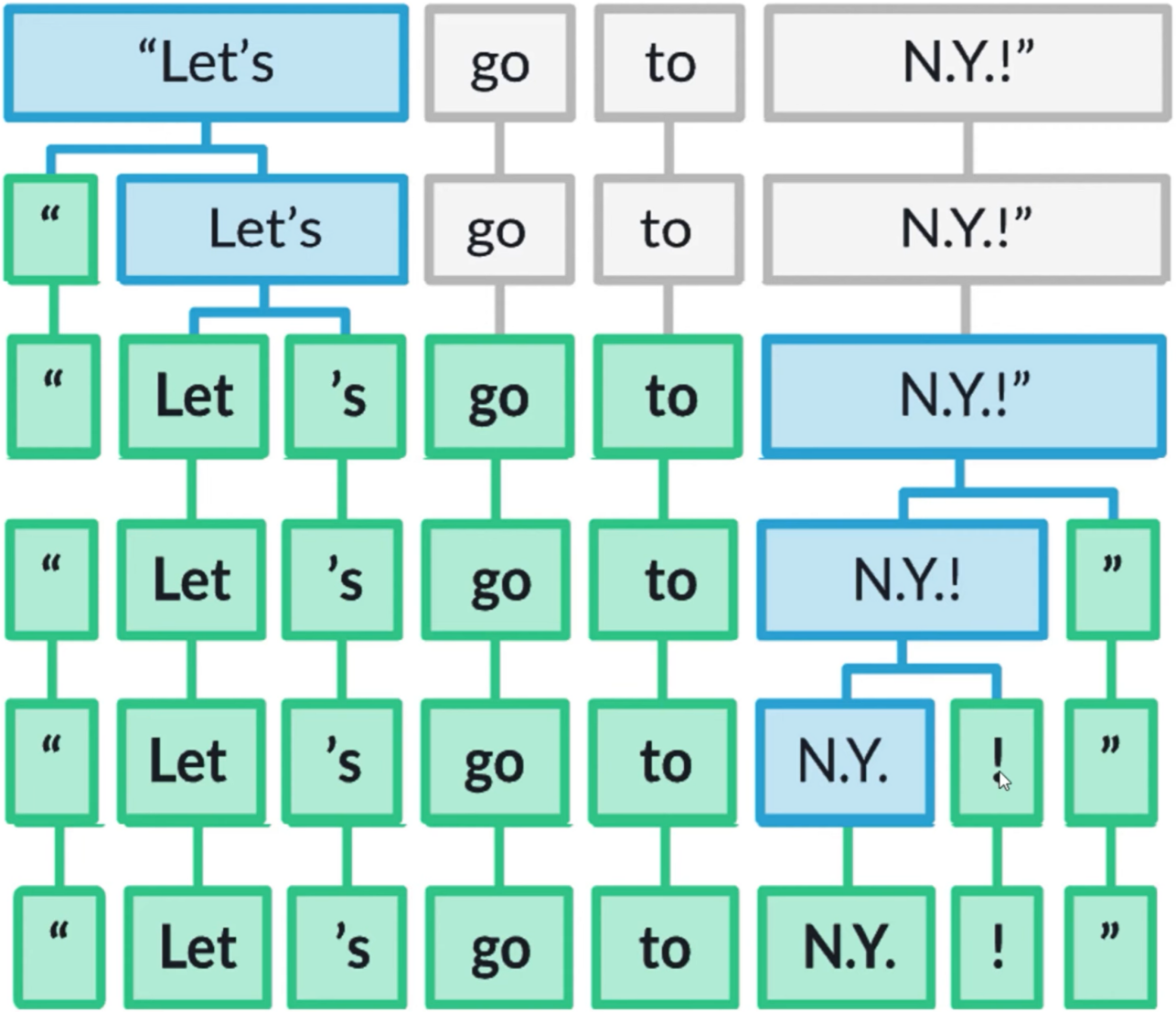
EXCEPTION

SUFFIX km) ! "

SUFFIX

EXCEPTION

Credits: https://spacy.io/



PREFIX

EXCEPTION

SUFFIX

SUFFIX

EXCEPTION

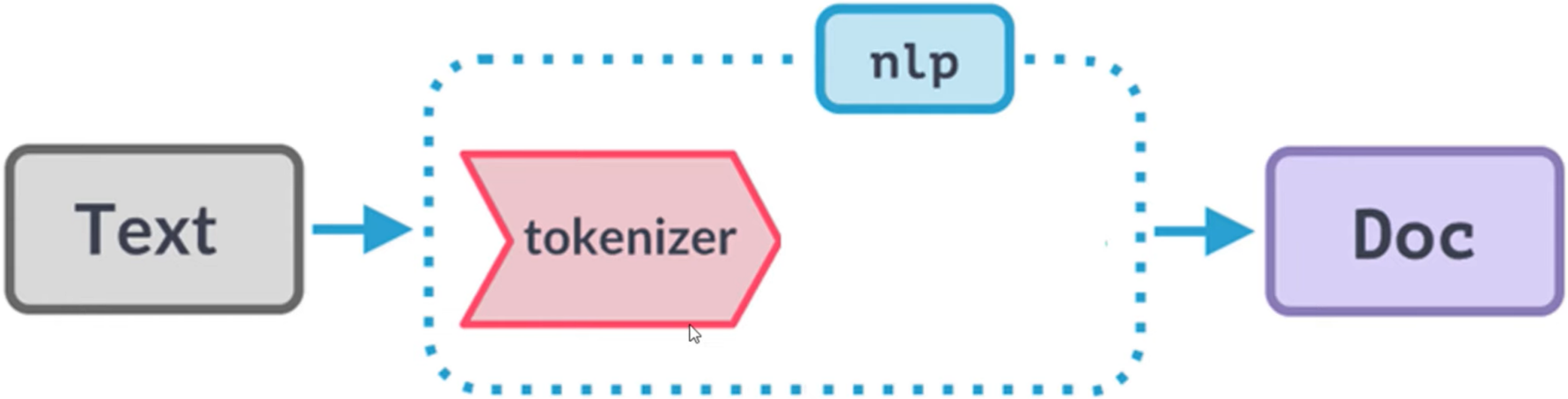
DONE

$ ("

km) ! "

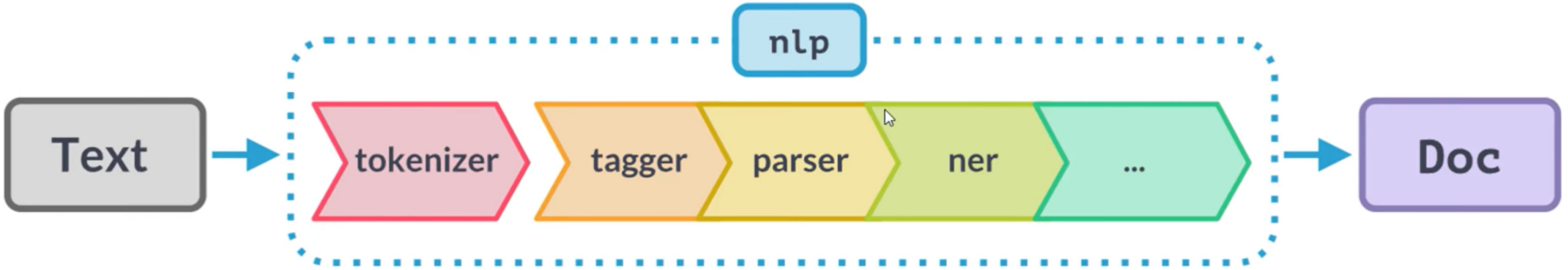
Credits: https://spacy.io/

nlp = spacy.blank("en")



Credits: https://spacy.io/

nlp = spacy.load("en_core_web_sm")



Credits: https://spacy.io/