SENSE CORP

Houston Data Analytics: Intro to Natural Language Processing (NLP)

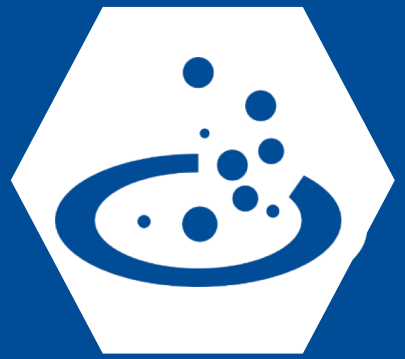
Sense Corp Presentation

- September 2018

SENSE CORP

Powering Insight-Driven Organizations

We turn data into actionable insights and transform organizations for the digital era.



Data

Digital

Transformation

SENSE CORP

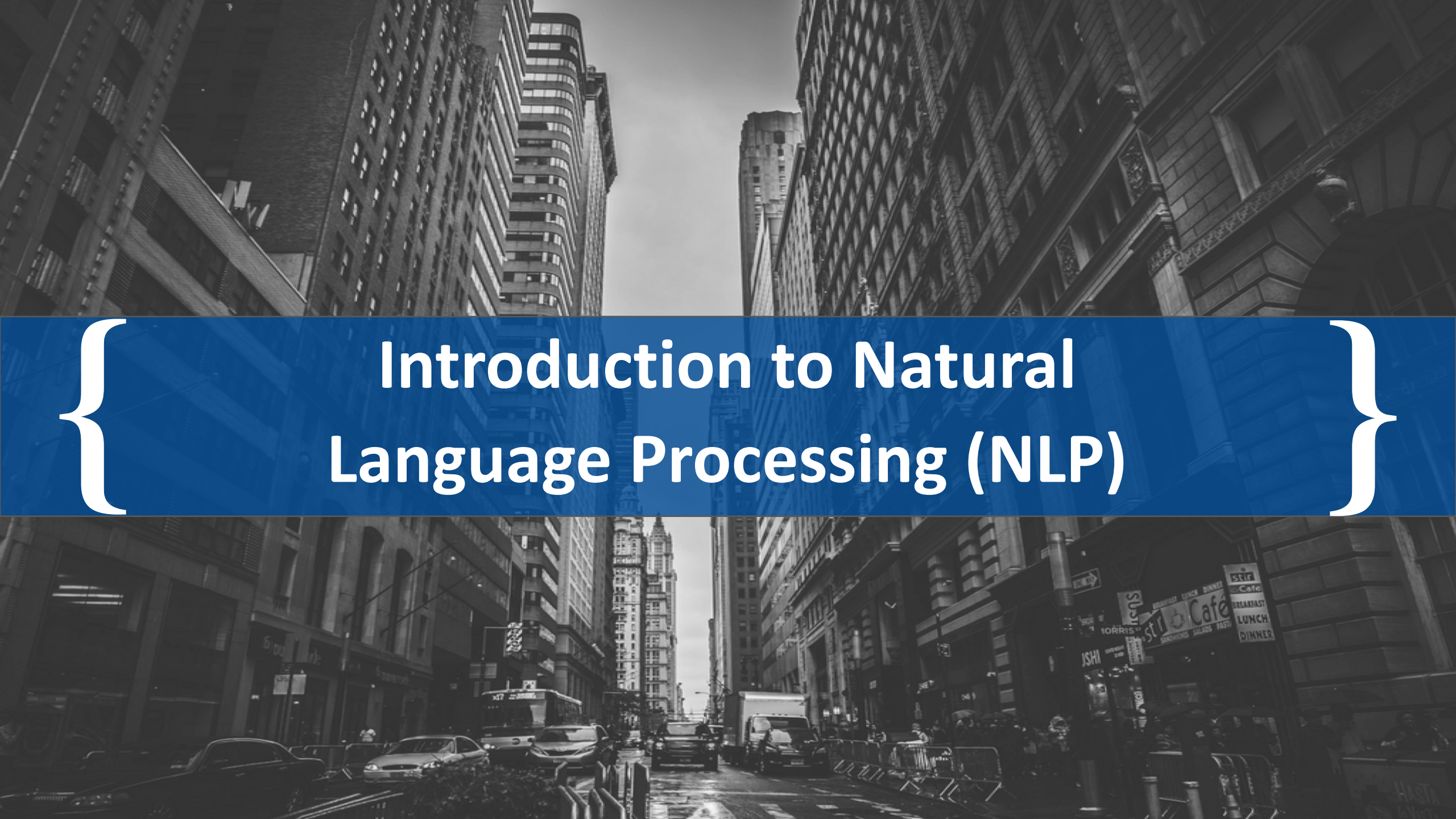
Powering Insight-Driven Organizations

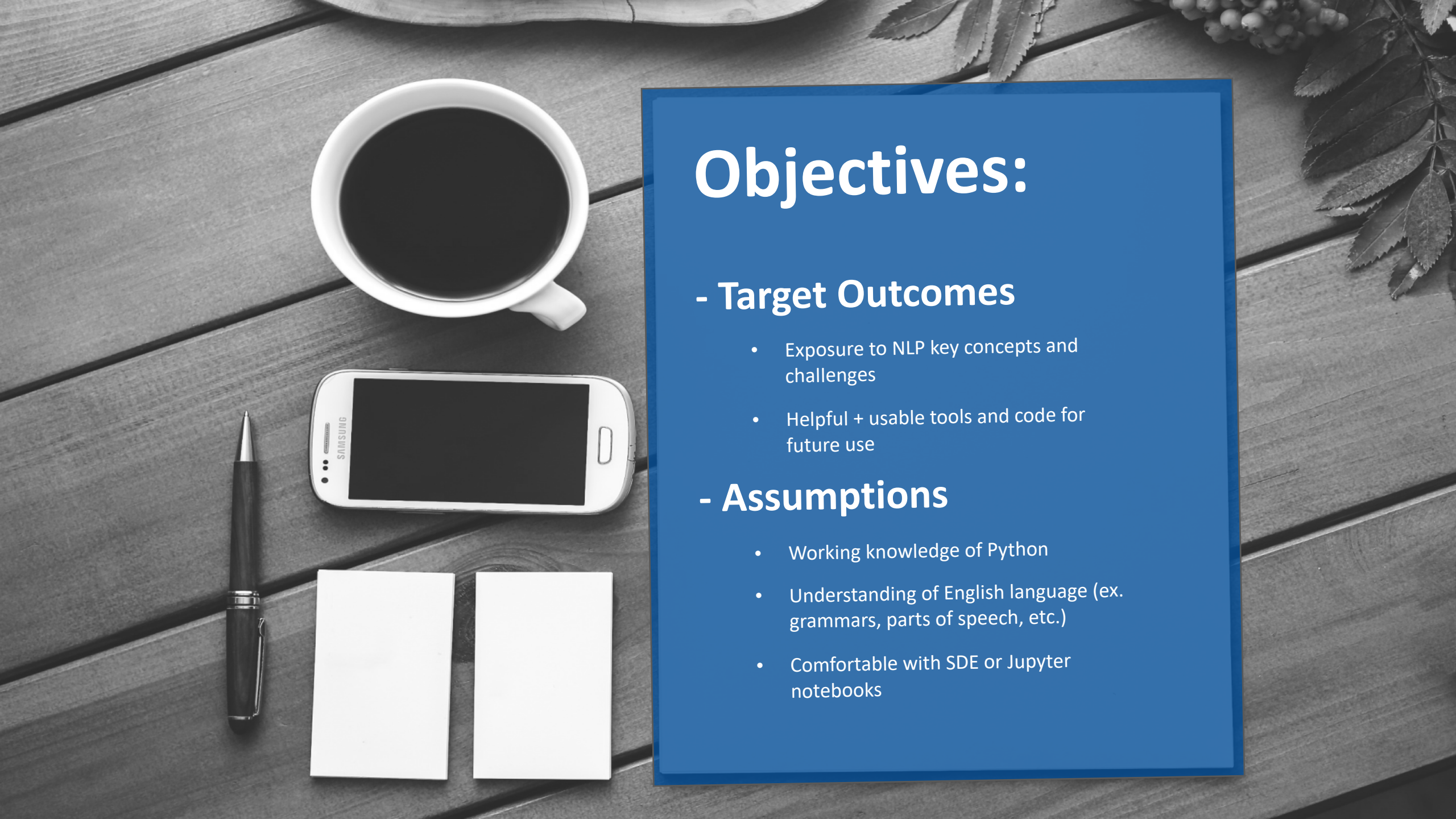
- Enterprise Data Strategy
- Data Visualization
- Data Governance
- Advanced Analytics & Data Science
- MDM
- Big Data
- Data Engineering
- Data Security

... we turn data into actionable insights and transform organizations for the digital era.

Sense Corp has delivered exceptional results to hundreds of the world's largest global organizations and complex government agencies.

PUBLIC SECTOR	TELECOM/MEDIA	HEALTHCARE	FINANCIAL SERVICES	ENERGY	OTHER
TEXAS Health and Human Services	Time Warner Cable	ASCENSION	WELLS FARGO	PLAINS ALL AMERICAN PIPELINE, L.P.	LENNOX
Texas Department of Transportation	COMCAST	BlueCross BlueShield	NEW YORK LIFE	bhpbilliton	enterprise FLEET MANAGEMENT
Office of the Comptroller Texas	AT&T	UnitedHealthcare	MasterCard	OCEANEERING	TOYOTA
ercot	Spectrum	MTM	FARM CREDIT BANK OF TEXAS	Bristow	DELL
TXGLO	CenturyLink	OPTUM	USAA	HALLIBURTON	Builders FirstSource
HARRIS COUNTY FLOOD CONTROL DISTRICT	suddenlink	GEISINGER	Bank of America	Shell	Panera BREAD
DIR		WELLMED	Dimensional	devon	Sysco
		CENTENE Corporation		CAMERON	


Introduction to Natural
Language Processing (NLP)



Objectives:

- Target Outcomes

- Exposure to NLP key concepts and challenges
- Helpful + usable tools and code for future use

- Assumptions

- Working knowledge of Python
- Understanding of English language (ex. grammars, parts of speech, etc.)
- Comfortable with SDE or Jupyter notebooks

What is NLP?

- **Wiki: Natural language processing** (**NLP**) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages.

- Identify the structure and meaning of words, sentences, texts and conversations
- Deep understanding of broad language
- NLP is all around us

What are some applications?

- *Machine translation*
- *Dialog systems*
- *Sentiment analysis*
- *Text classification*
- *Question answering*
- *Digital assistants*
- *Information retrieval*

Example: Language Comprehension

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh.** As **a boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

Q: who wrote Winnie the Pooh?

Q: where is Chris lived?

Why is NLP hard?

News Headlines

- Ban on Nude Dancing on Governor's Desk
- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Local High School Dropouts Cut in Half
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, but More Lies Ahead
- Hospitals Are Sued by 7 Foot Doctors
- Kids Make Nutritious Snacks

Semantic Disambiguation

Sentence	Relation
I ate spaghetti with meatballs.	(ingredient of spaghetti)
I ate spaghetti with salad.	(side dish of spaghetti)
I ate spaghetti with abandon.	(manner of eating)
I ate spaghetti with a fork.	(instrument of eating)
I ate spaghetti with a friend.	(accompanier of eating)

NLP: Where do we start?

Like any other data analytics/science project: **Exploration!**

Code workshop: NLTK and Genesis (Bible)

- Lexical Dispersion Plot
- Basic statistics
  - Concordances, similarities, dispersion plots
  - Corpus Size (Vocabulary and lexical diversity)
  - Frequency Distributions
  - Collocations
  - Overall statistics (number of words, average word length, words per sentence)
  - Sentiment analysis

(Larger) Example: The News

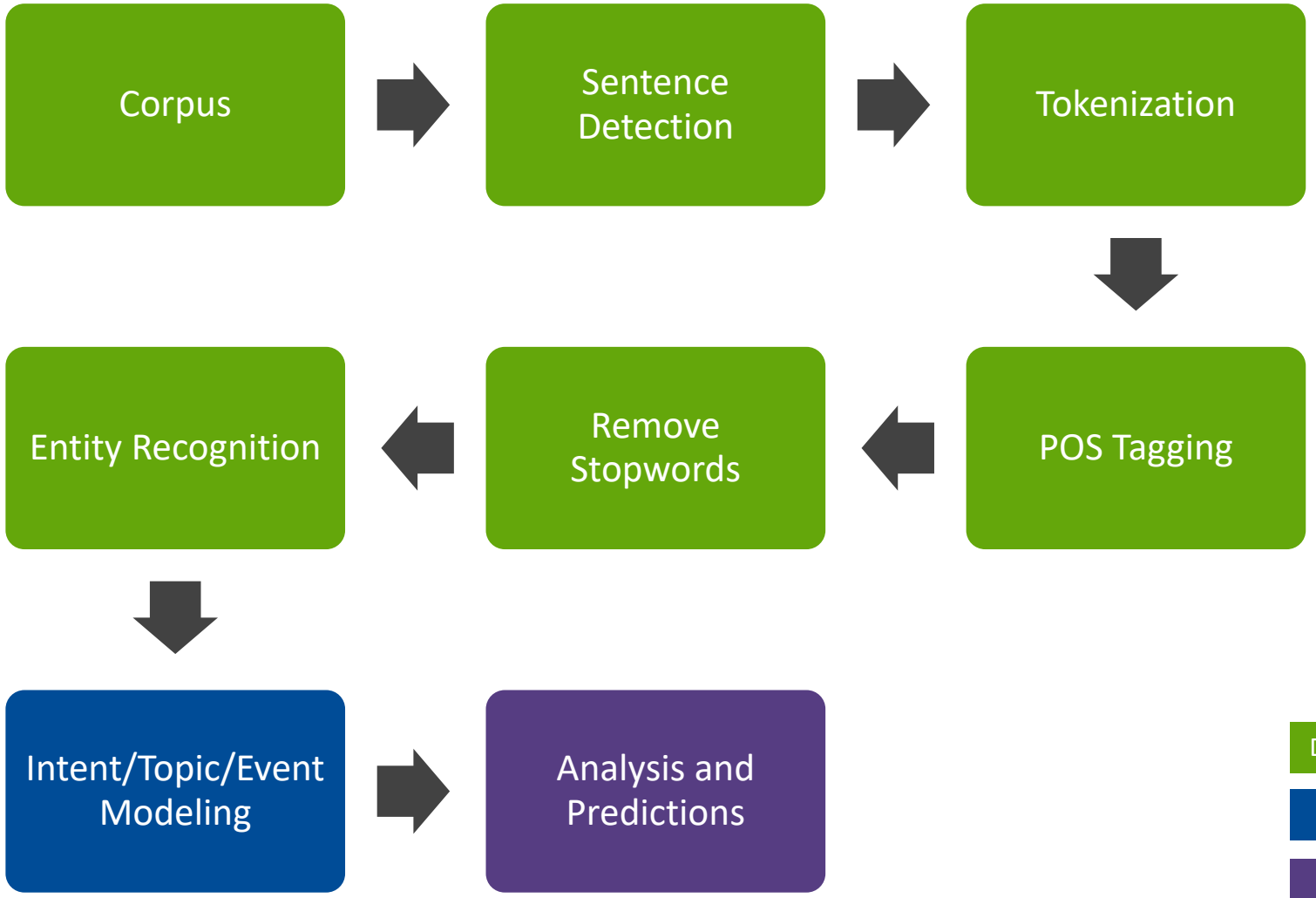
There are over 2M news articles published every day…

What is news?

News is whatever you want it to be

Let’s classify it (with NLP)

Walkthrough: NLP Pipeline



- General Framework (analogous to data processing pipeline)
- Data processing vs. modeling
- Extracting Signal vs. Removing Noise

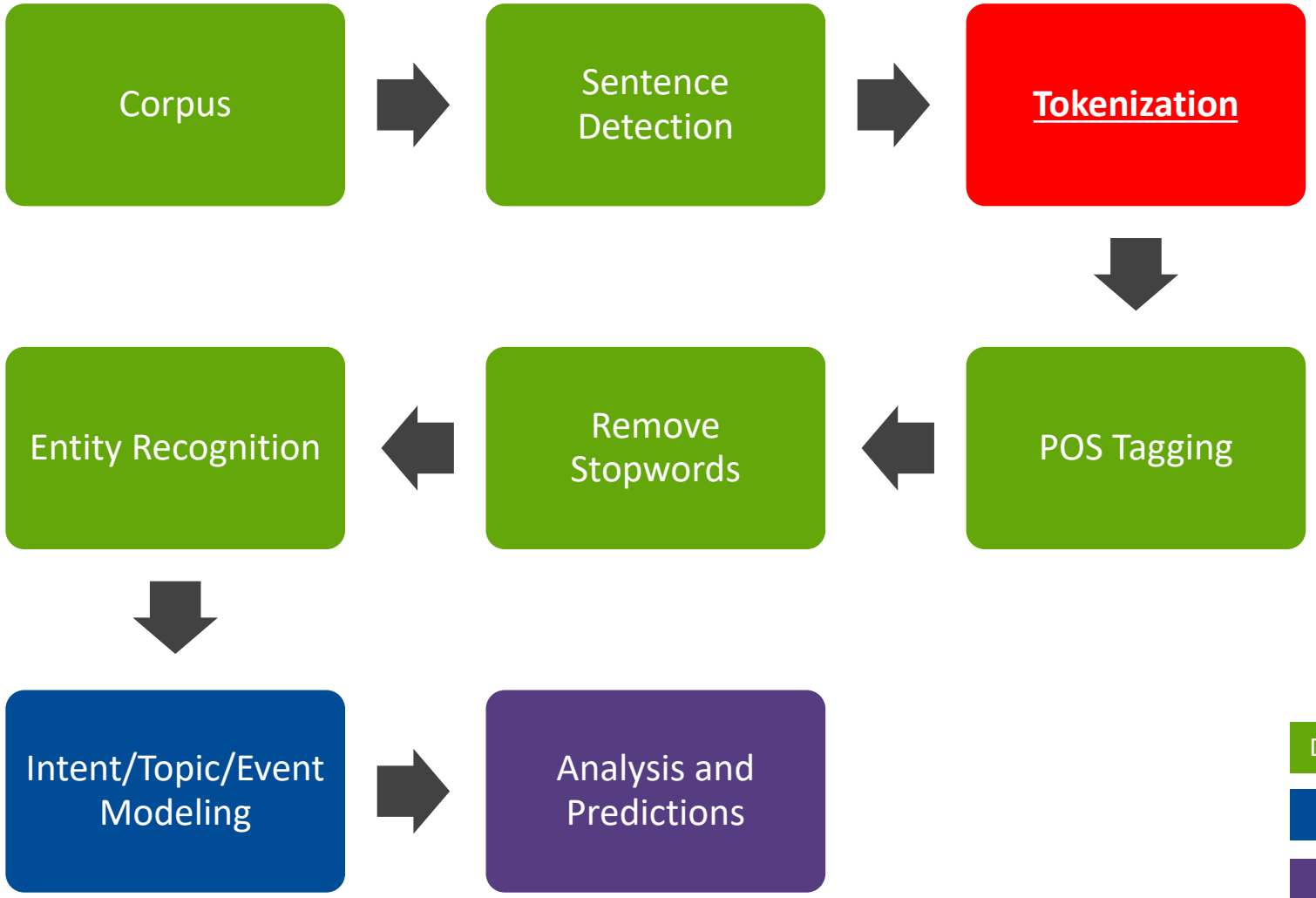
--LEGEND--

Data Prep/Processing/Wrangling

Actual Modeling

Outputs

Walkthrough: NLP Pipeline



- Tokenization is ‘chopping up’ a text into pieces
- Tokens are also called ‘grams’ and can be designated to be any length

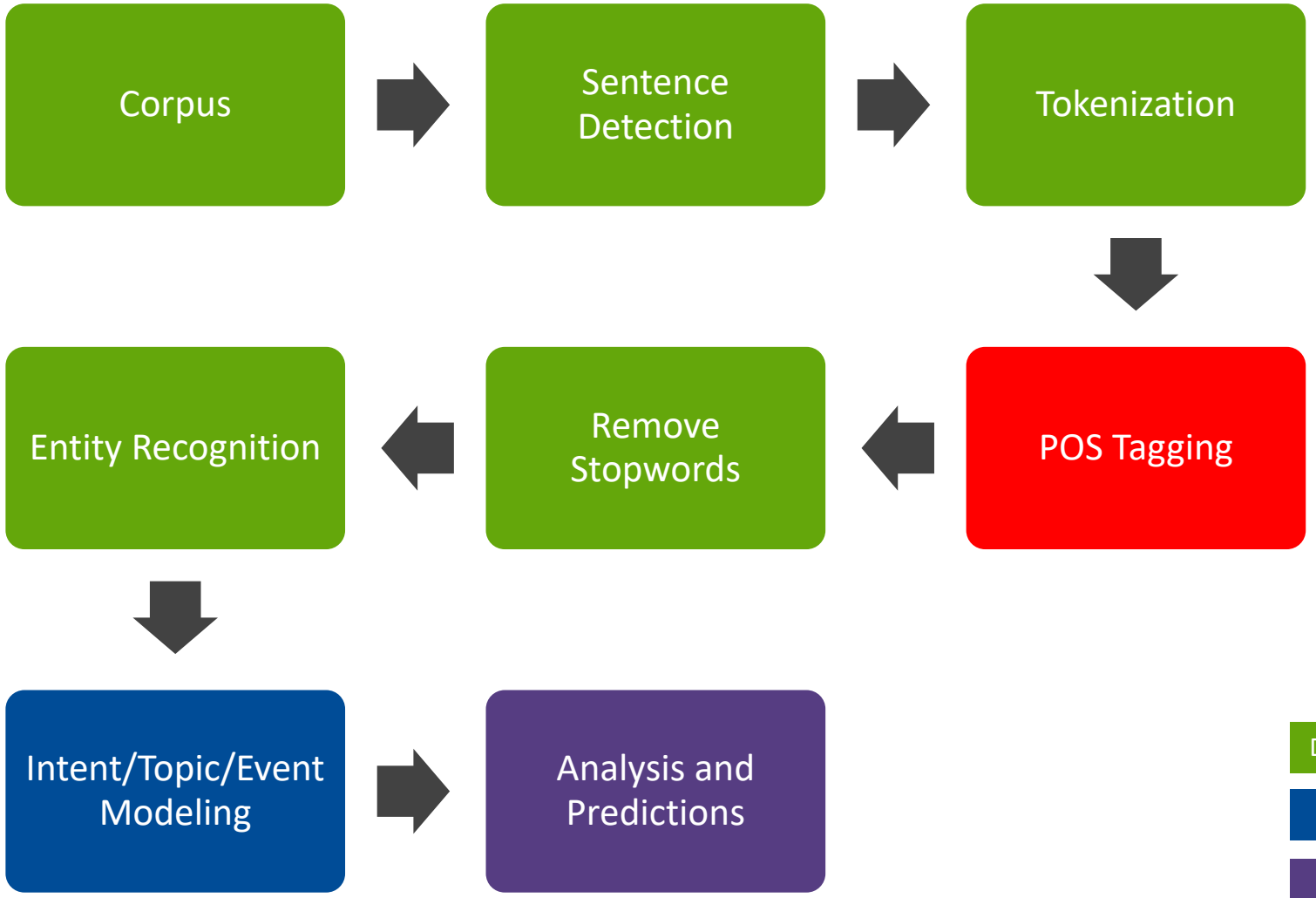
--LEGEND--

Data Prep/Processing/Wrangling

Actual Modeling

Outputs

Walkthrough: NLP Pipeline



- POS Tagging, or "Part-of-speech tagging," is identifying the part of speech of each token such as noun, adjective, etc.
- POS Tagging can be complex! The POS Tag is dependent on the definition of the word AND the usage.

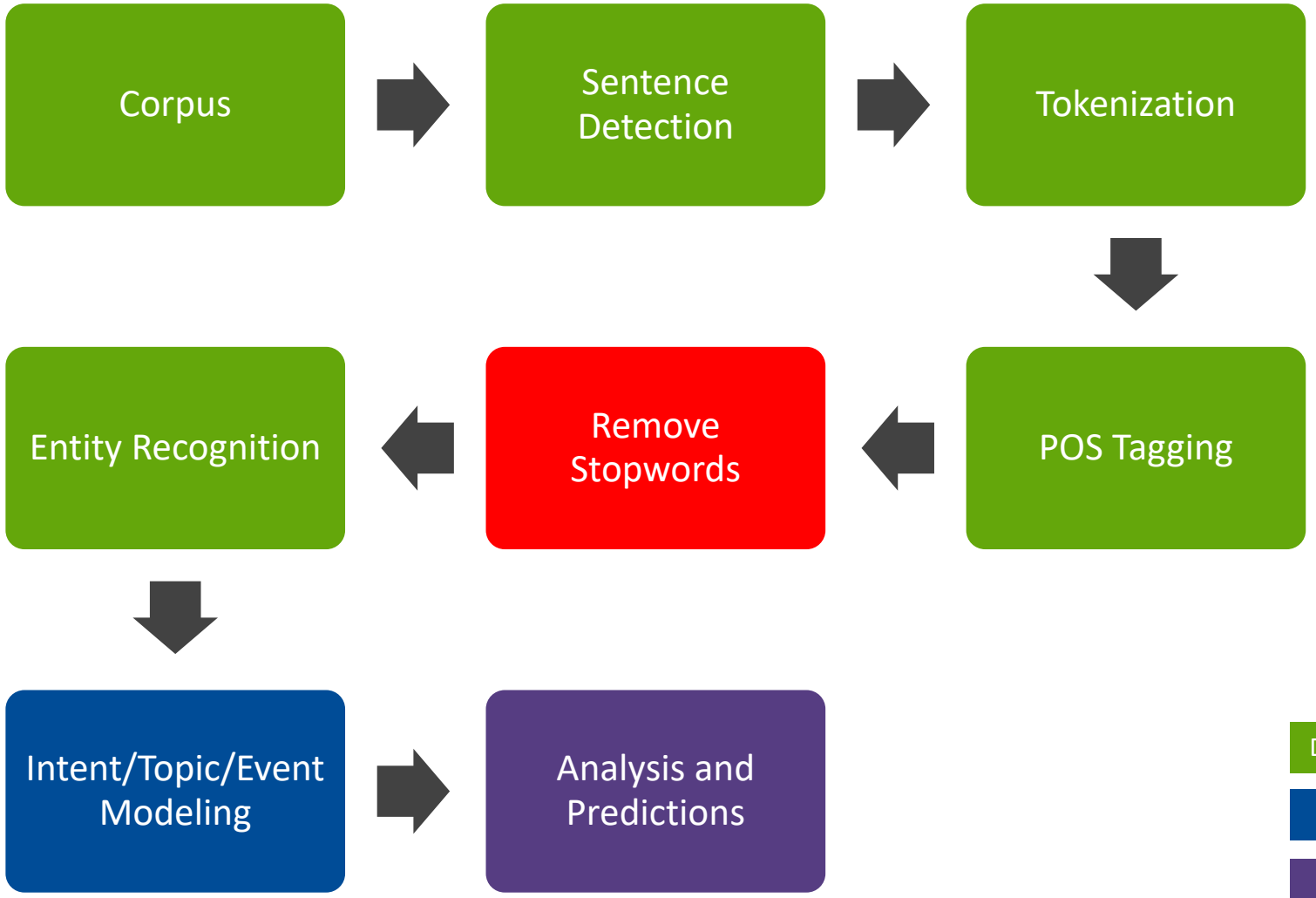
--LEGEND--

Data Prep/Processing/Wrangling

Actual Modeling

Outputs

Walkthrough: NLP Pipeline

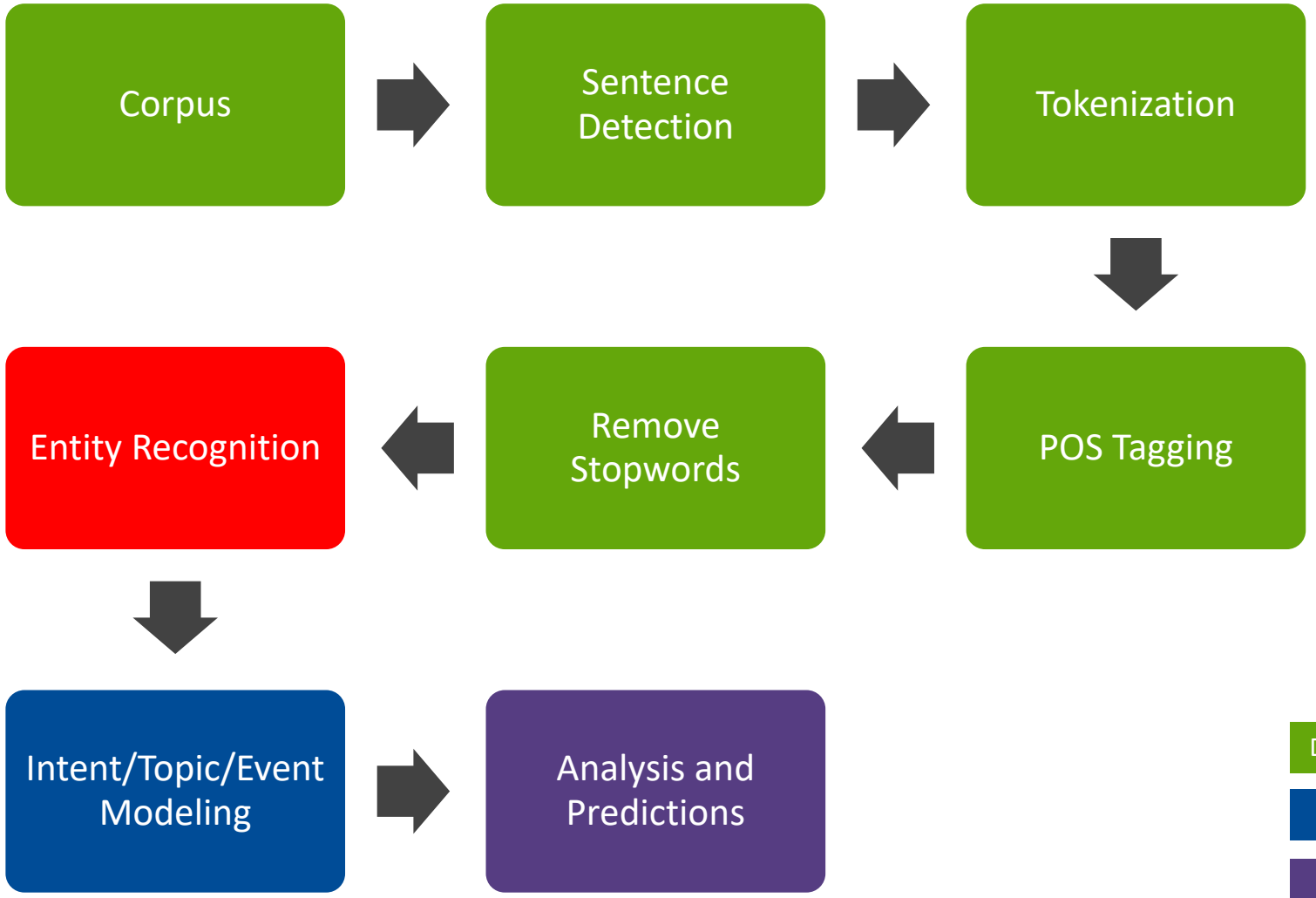


- Stopwords are extremely common words that have very little value, like ‘the’, ‘is’, etc.
- Because they are so common, they add significant ‘noise’ to our analysis

--LEGEND--

- Data Prep/Processing/Wrangling
- Actual Modeling
- Outputs

Walkthrough: NLP Pipeline



- Entity recognition is the task of identifying key nouns such as persons, places, etc.
- Entities can be identified by pre-defined dictionaries and/or grammars (POS tags).

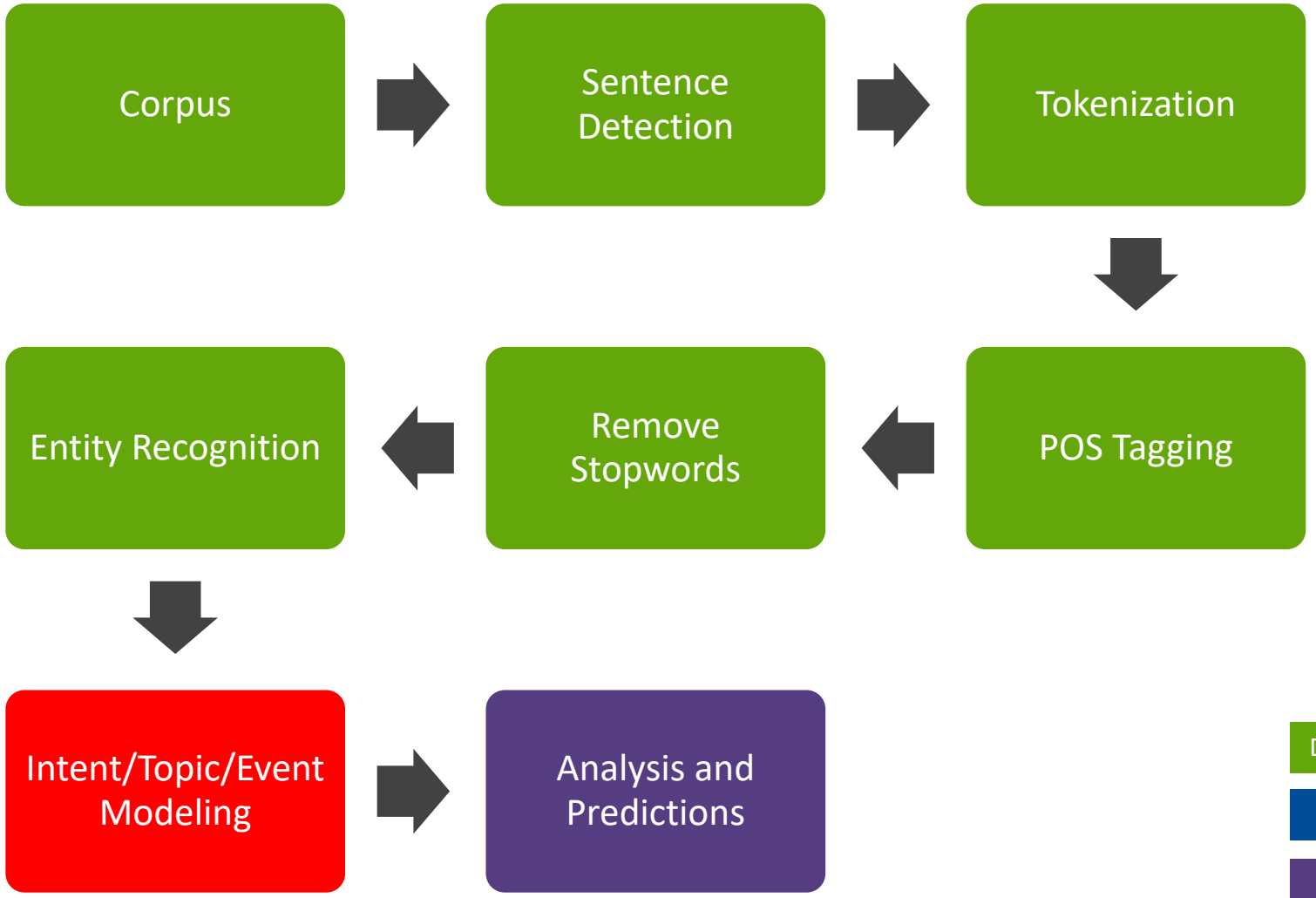
--LEGEND--

Data Prep/Processing/Wrangling

Actual Modeling

Outputs

Walkthrough: NLP Pipeline



- For this example, we will use latent semantic analysis (LSA).
- LSA is a matrix decomposition technique that aims to map terms to a hidden topic layer. It is based on single value decomposition (SVD).

--LEGEND--

- Data Prep/Processing/Wrangling
- Actual Modeling
- Outputs

Useful Tools

- Helpful Python packages
  - NLTK
    - Wordnet
  - GenSim
  - Textblob
  - SpaCy
  - Polyglot
  - Scikit-Learn
- 3rd Party Tools
  - Google Cloud NLP (Ex. Knowledge Graph)
  - Stanford Core NLP

Takeaways

NLP have practical applications, but none do a great job in an open-ended domain

Sentences are understood through grammar, parsing and lexicons

Choosing a good interpretation of a sentence requires evidence from many sources

Most interesting NLP comes in connected discourse rather than in isolated sentences

THANKS!

Justin J. Nguyen
jnguyen@sensecorp.com

Pratish Kanani
pkanani@sensecorp.com

SENSE CORP