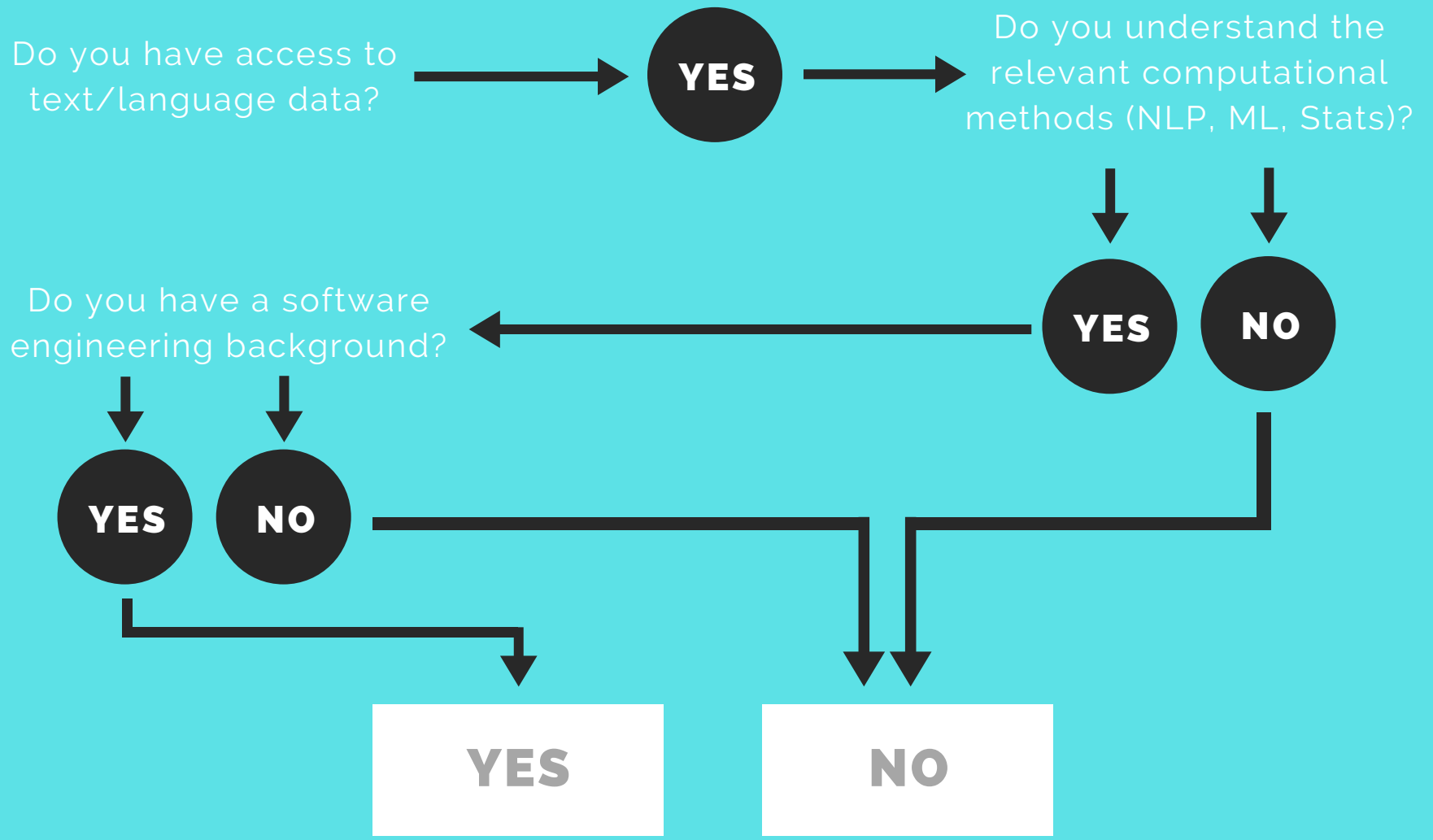
UNIVERSITY OF WASHINGTON
DEPARTMENT OF LINGUISTICS

HACKABLE UNIFIED TEXT PROCESSING PLATFORM

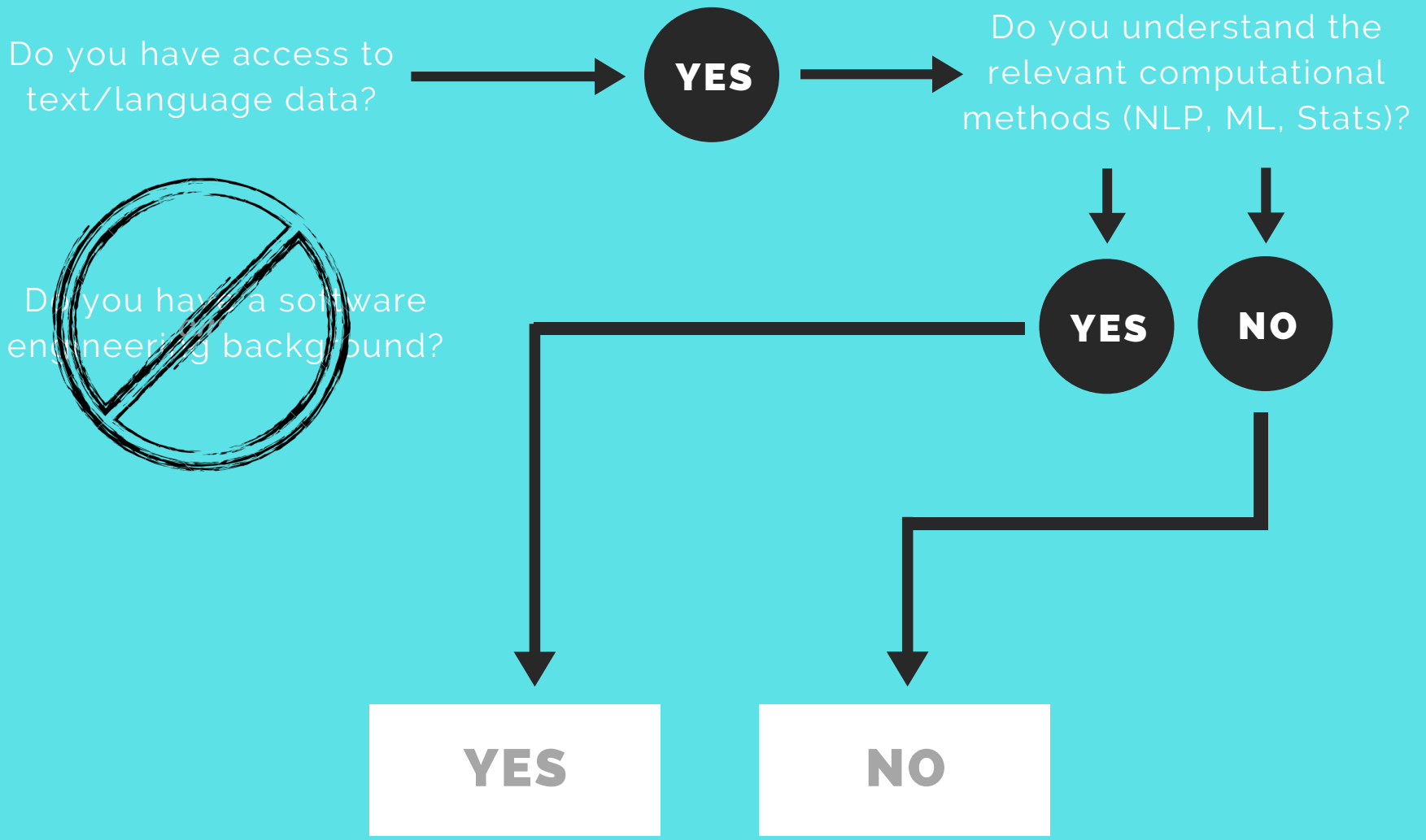
TEV'N POWERS

COMPUTATIONAL LINGUISTIC MASTER'S THESIS

WHAT MAKES AN NLP PRACTITIONER?



WHO CAN BE AN NLP PRACTITIONER?



WHO ARE THE BUILDERS AND USERS OF NLP TOOLS?

ACADEMIA

Students and researchers usually write code for their peers in the academic community, with a goal to make advancements in the fields of linguistics, natural language processing, and/or machine learning.

INDUSTRY

Software engineers and/or data scientists in industry usually build language features specific to the needs of a product to be marketed or sold to customers.

OPEN SOURCE

Open source project contributors and maintainers may share some of the same interests as those in academia or industry, as well as building systems or tools for other developers.

?

WHO COULD BE BUILDERS AND USERS OF NLP TOOLS?

ACADEMIA

Students and researchers usually write code for their peers in the academic community, with a goal to make advancements in the fields of linguistics, natural language processing, and/or machine learning.

INDUSTRY

Software engineers and/or data scientists in industry usually build language features specific to the needs of a product to be marketed or sold to customers.

OPEN SOURCE

Open source project contributors and maintainers may share some of the same interests as those in academia or industry, as well as building systems or tools for other developers.

EVERYONE ELSE

Anyone who works with or has access to text data should be able to leverage NLP techniques to gather valuable insights from their data.

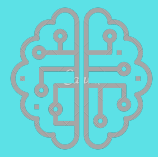
DATA PROCESSING PIPELINE

Load Data

- File System
- Database
- Local machine (on computer)
- Cloud server
- Various file formats

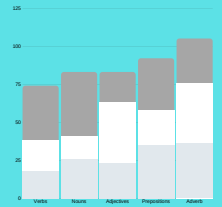
Text Featurization

- Tokenization
- POS-Tagging
- Dependency Parsing
- Stemming
- Lemmatization



Modeling

- Clustering
- Nearest Neighbors
- Classification
- Sentiment Analysis
- Text Generation



Insights

- Predictions
- Visualizations

HACKABLE UNIFIED TEXT PROCESSING PLATFORM

TEXT PROCESSING PLATFORM

A desktop application that provides the framework for executing NLP, ML, and statistical techniques on text data provided by a user.

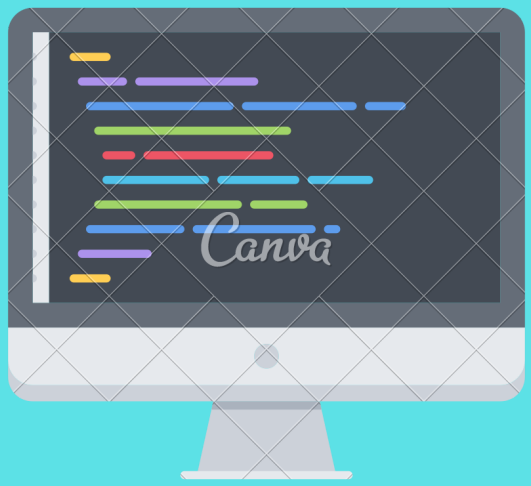
HACKABLE

Every user has access to create "plugins" which support their own text processing needs or the needs of others in the community.

UNIFIED

A common data schema across all plugins will enable users to import and export data to/from arbitrary formats. Users can sequence text processing modules together to form text processing pipelines.

HACKABLE UNIFIED TEXT PROCESSING PLATFORM



Data

Users can import data from files or a database on their computer or cloud storage provider.

Data Labeling

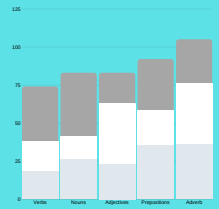
Text can be annotated via a simple data set interface in the application.

Text Featurization

Text features are created via text transformation modules that augment input data.

Modeling

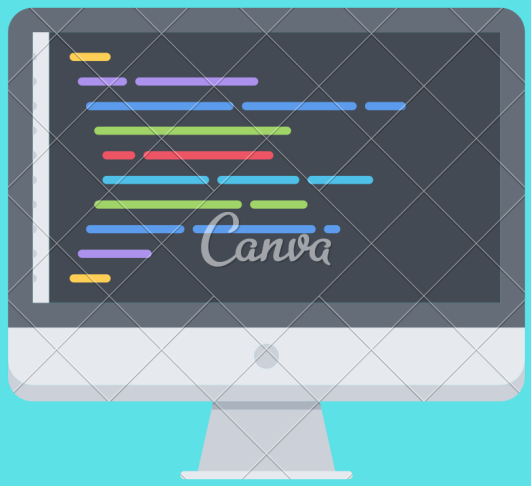
Modeling plugins are fit to learn patterns in data and make predictions on unseen data.



Insights

System output can take the form of annotated data, salient insights, or visualizations.

HACKABLE UNIFIED TEXT PROCESSING PLATFORM



Data

Users can import data from files or a database on their computer or cloud storage provider.

Data Labeling

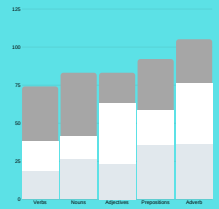
Text can be annotated via a simple data set interface in the application.

Text Featurization

Text features are created via text transformation modules that augment input data.

Modeling

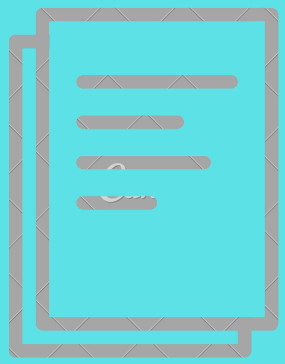
Modeling plugins are fit to learn patterns in data and make predictions on unseen data.



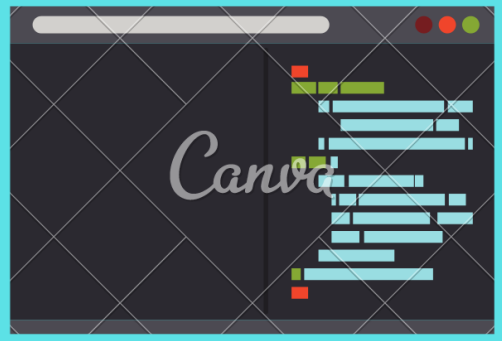
Insights

System output can take the form of annotated data, salient insights, or visualizations.

HACKABLE **UNIFIED** TEXT PROCESSING PLATFORM



Standardized Data Format



Standardized Data Module API

A FANFICTION EXAMPLE

SAM

- Fanfiction writer
- Interested in understanding differences in writers and fandoms in the fanfiction dataset

ALEX

- Beginner Software Engineer
- Wants to become more proficient with NLP techniques

RYAN

- Computational Linguist
- Works as an NLP engineer and contributes to various OSS projects

A FANFICTION EXAMPLE

ALEX

- Writes a plugin that tokenizes and pos-tags each document in a dataset (e.g. a fanfict story)
- Generates basic exploratory data analysis (EDA) statistics for each document

RYAN

- Writes a plugin that extracts the highest tf-idf features from each document in a dataset & visualizes via word cloud
- Her plugin can treat each document separately, or group by an arbitrary key (e.g. author or fandom)

SAM

- Sees graphs and plots about the statistics of each story in the Fanfiction data set
- Produces a word cloud of the most meaningfully different tokens in each story, author, or fandom.

SOURCE CODE

https://github.com/tevnpowers/thesis