



Introduction



Global Lead Data Science Instructor, General Assembly

Managing Partner, BetaVector

Marketing + Comms Director, Statistics Without Borders

Previously:

Data Science @ Optimus Consulting

Enterprise Analytics @ Smucker's

M.S. Statistics @ The Ohio State University

Recommended Reads:

Data-Driven Thinking: "Factfulness"

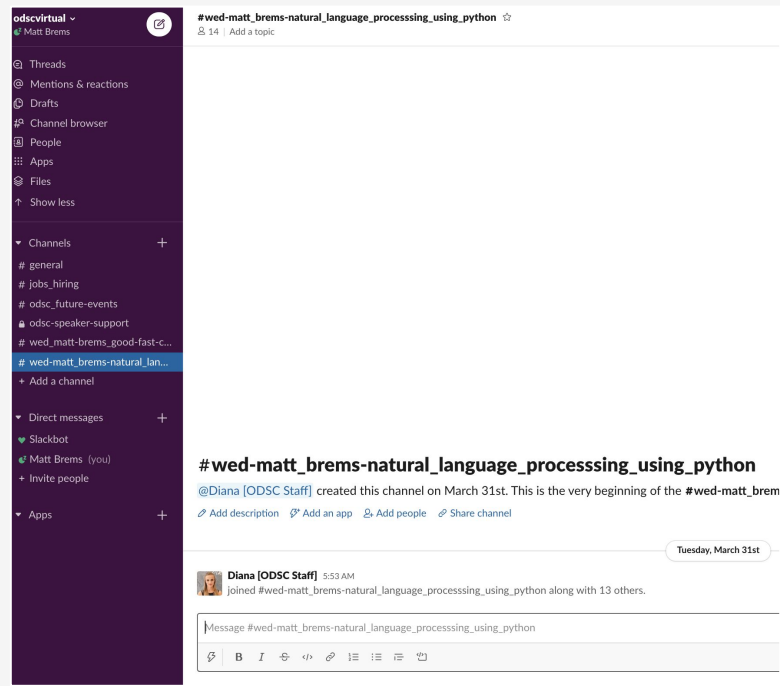
Data Visualization: "Storytelling with Data"

Data Science: "An Introduction to Statistical Learning with Applications in R"

Engagement Today: Slack!

We will be using Slack in three main ways today:

1. For you to ask questions of me!



Engagement Today: Slack!

We will be using Slack in three main ways today:

1. For you to ask questions of me!
2. For me to ask open-ended questions of you!

Reply to thread



Matt Brems (he/him) 2:31 PM

What might be some advantages to the bag-of-words approach? (THREAD)

Engagement Today: Slack!

We will be using Slack in three main ways today:

1. For you to ask questions of me!
2. For me to ask open-ended questions of you!
3. For me to ask closed-ended questions of you!



Polly APP 2:29 PM

Are you familiar with the term bag-of-words?

Yes

Maybe

No

What is Natural Language Processing?



What is Natural Language Processing?

```
model.doesnt_match(['angioplasty', 'appendectomy', 'cabg', 'bronchoscopy'])  
  
'appendectomy'
```

What is Natural Language Processing?

English - detected ▾



↔



German ▾

Can you please help me find the bathroom?

×

Können Sie mir bitte helfen, das Badezimmer zu finden?

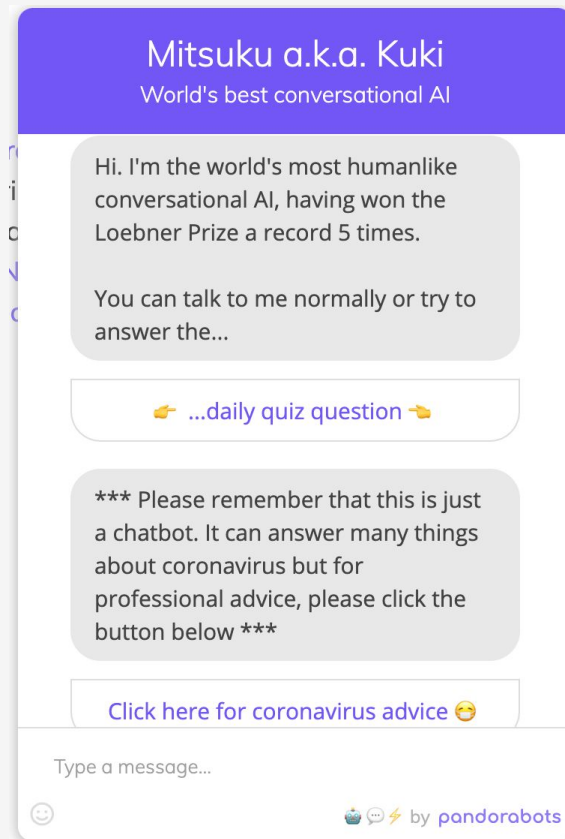
Open in Google Translate

Feedback

translate.google.com ▾

Google Translate

What is Natural Language Processing?



Our goals with Natural Language Processing

1. **Our broad goal** with natural language processing is to get computers to understand language more like how humans understand language.
2. **Our more specific goal** with natural language processing in traditional machine learning is to convert our semi-structured text data into a dataframe of real numbers.
 - X is our input data.
 - Y is our output data.

WARNING

Analysis with NLP can only be as good as the data you provide it.

- If your data are biased, then your results will be biased.
- If your data are not biased... you're probably wrong.

Jeopardy Primer

- Round 1: Jeopardy!
 - Five clues in six categories.
 - Dollar amounts range from \$200 to \$1,000.
- Round 2: Double Jeopardy!
 - Five clues in six categories.
 - Dollar amounts range from \$400 to \$2,000.
- Round 3: Final Jeopardy!
 - One category, one question, dollar amount is a wager.



To the notebook!

THANK YOU



LinkedIn: Matthew Brems

Twitter: @matthewbrems

Github: matthewbrems

Email: matt@betavector.com



hello@betavector.com