

# Course Introduction

CS 485, Fall 2024

Applications of Natural Language Processing

Brendan O'Connor

College of Information and Computer Sciences  
University of Massachusetts Amherst

# What

- Learn fundamental principles in natural language processing, and how to practically use them
- Hands-on experience: data collection and implementation
- Appreciation of basic linguistic issues
- Know when NLP works and when it doesn't
- Course topics may include: text processing, corpus data, probabilistic language models, syntax models, word embeddings, text classification, sentiment analysis, question answering, deep learning methods, large language models...

# Requirements

- (10%) Quizzes / exercises (approx. weekly)
- (30%) Problem sets
  - Written: math and concepts
  - Programs: in Python
- (30%) In-class Midterm
- (30%) Final projects (groups of 2-3)  
*[Choose a topic, or select a suggested topic]*
  - Project Proposal
  - Progress Report
  - In-class presentations
  - Final Report

# Resources

- Webpage (and Canvas' link to it) will have lectures/readings and links to all other sites
  - Canvas? for communication
  - Gradescope for turning in assignments
  - Canvas/Echo360 for video recordings

# Upcoming

- Exercise #1: Class survey, on Gradescope.  
Releasing this week due next week
- First homework coming next week

# Ambiguity in language

*Why NLP is hard...*

- Red Tape Holds Up Bridges
- March Planned for Next August
- At the drop of a hat

## **A Huge Threat to the U.S. Budget Has Receded. No One Is Sure Why.**

Instead of growing and growing, as it always had, spending per Medicare beneficiary has nearly leveled off over more than a decade.

## Levels of Linguistic Structure

This is a simple sentence.



## Levels of Linguistic Structure: Characters

T	h	i	s		i	s		a		s	i	m	p	l	e		s	e	n	t	e	n	c	e	.
---	---	---	---	--	---	---	--	---	--	---	---	---	---	---	---	--	---	---	---	---	---	---	---	---	---

## Levels of Linguistic Structure: Morphology

This	is	a	simple	sentence	.
------	----	---	--------	----------	---

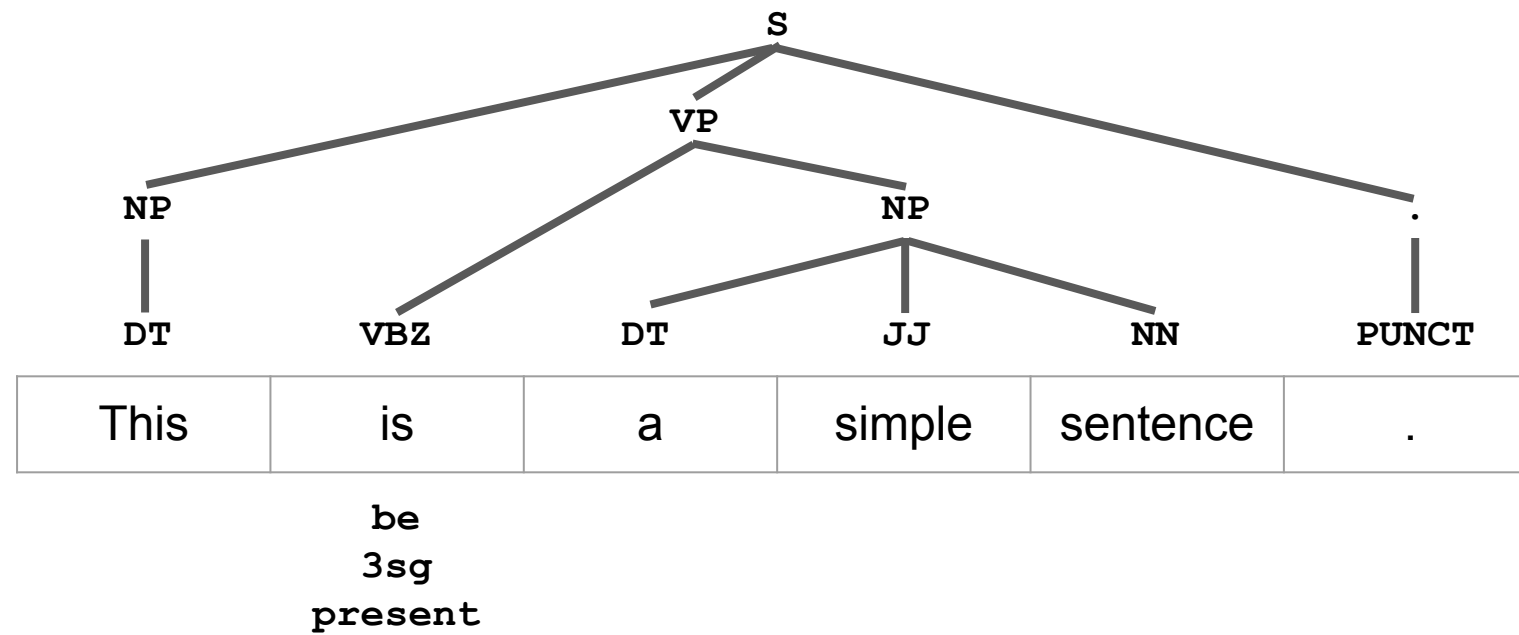
**be**  
**3sg**  
**present**

## Levels of Linguistic Structure: Parts of Speech

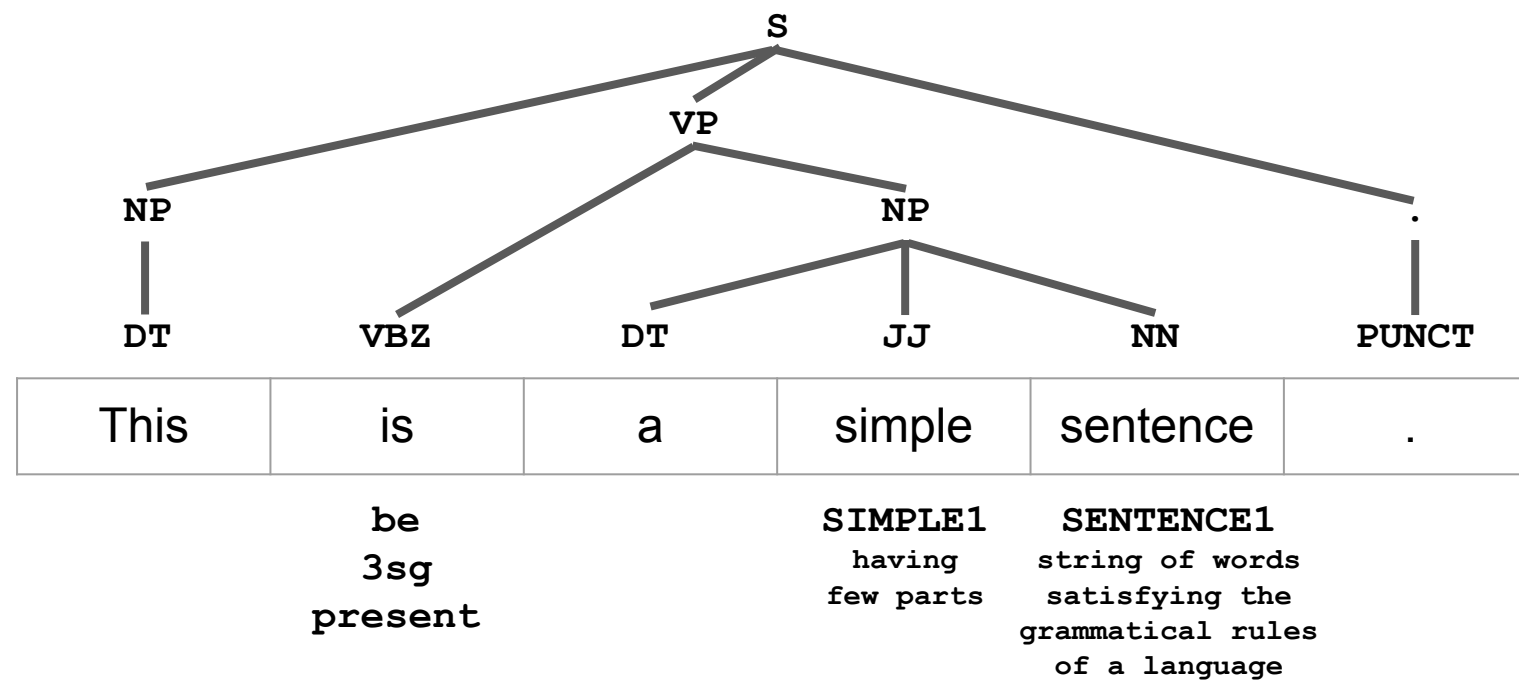
DT	VBZ	DT	JJ	NN	PUNCT
This	is	a	simple	sentence	.

be  
3sg  
present

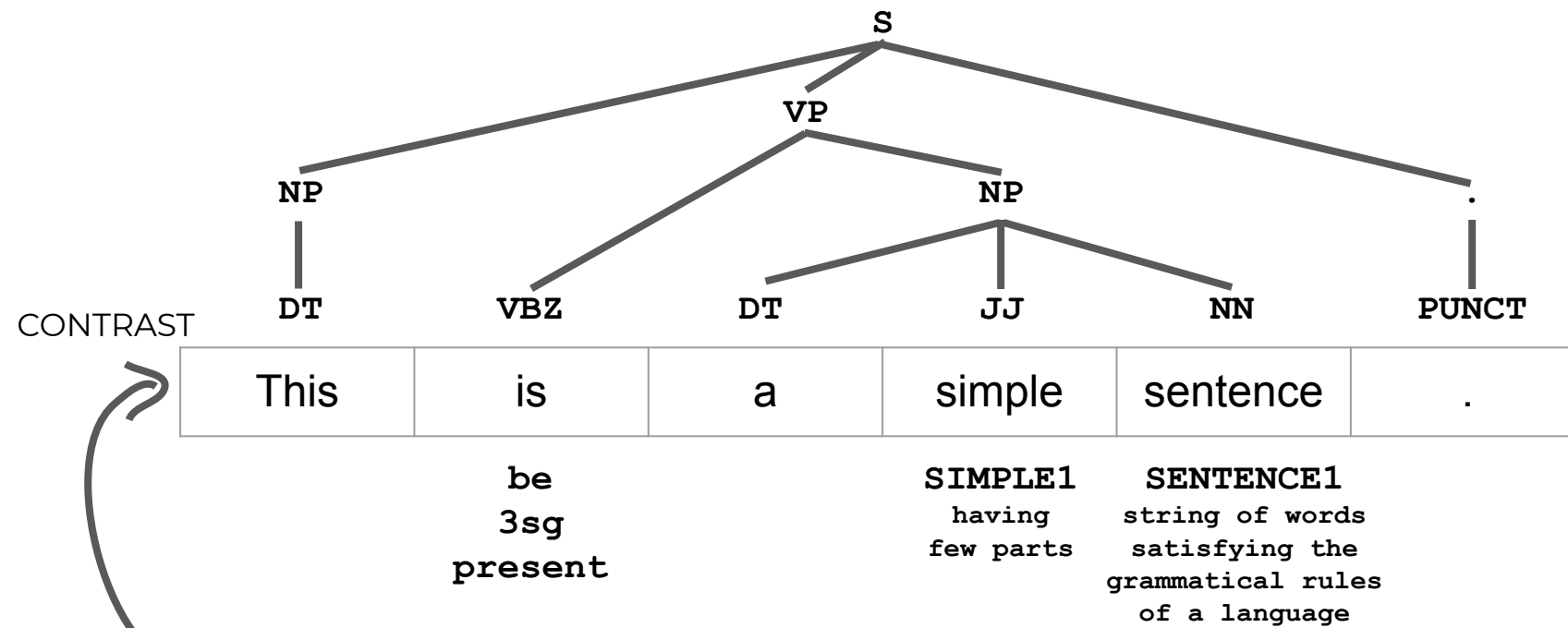
## Levels of Linguistic Structure: Syntax



## Levels of Linguistic Structure: Semantics



# Levels of Linguistic Structure: Discourse



CONTRAST



But it is instructive.

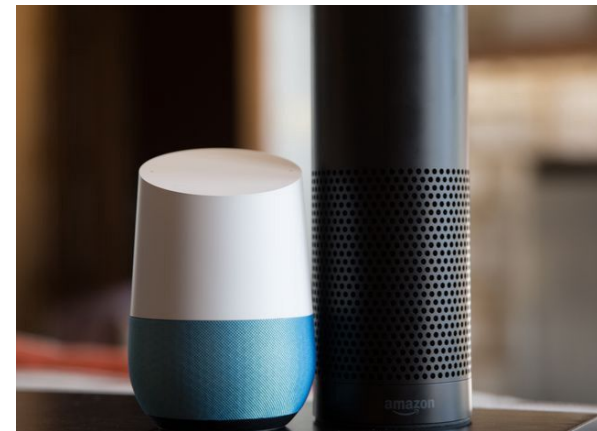
## NLP Today: Speech Interfaces & Voice Assistants



## NLP Today: Speech Interfaces & Voice Assistants



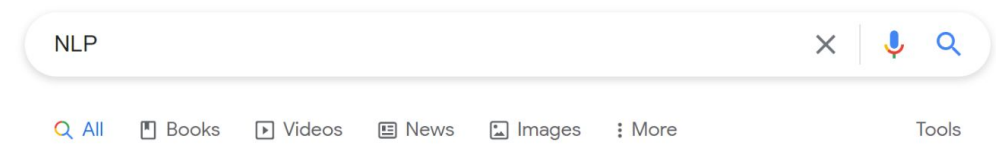
Ophelia, call the  
police



Sure. Playing F\*\*\* tha  
Police by N.W.A.



# NLP Today: Search & Summarization



About 65,300,000 results (0.74 seconds)



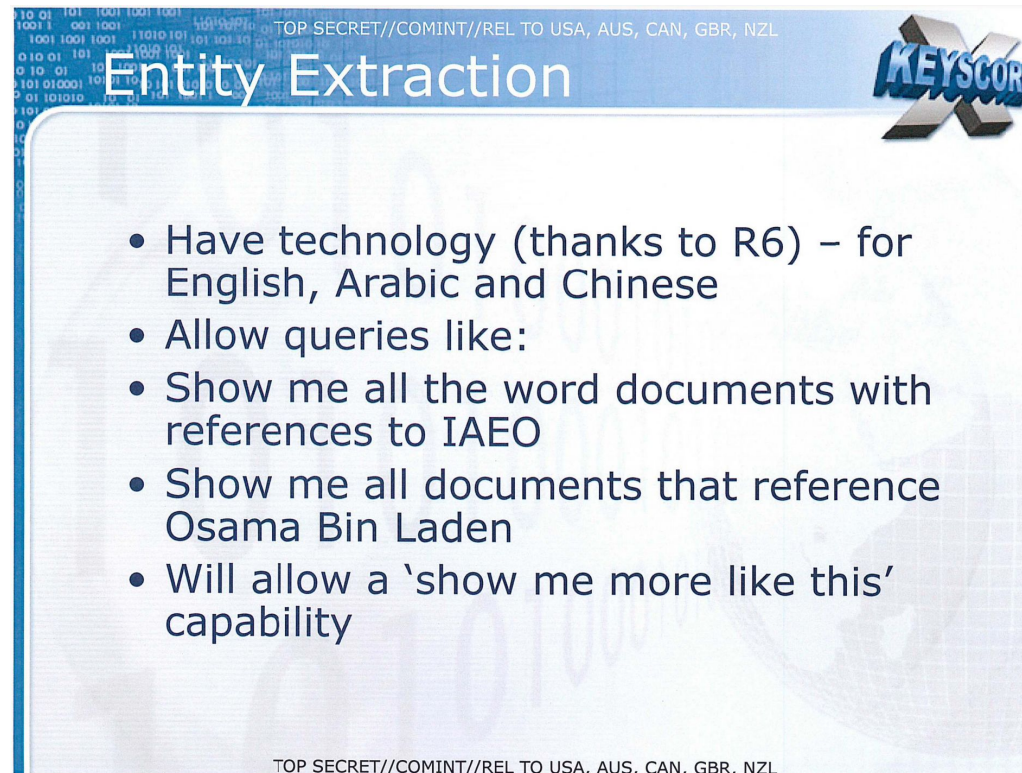
**Natural language processing (NLP)** refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

Jul 2, 2020

<https://www.ibm.com> > Cloud > Cloud Learn

[natural language processing \(NLP\) - IBM](#)

## NLP Today: Search & Summarization



Entity Extraction

KEYSCORE

- Have technology (thanks to R6) – for English, Arabic and Chinese
- Allow queries like:
- Show me all the word documents with references to IAEO
- Show me all documents that reference Osama Bin Laden
- Will allow a 'show me more like this' capability

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

# NLP Today: Machine Translation

## 三体 (小说) [编辑]

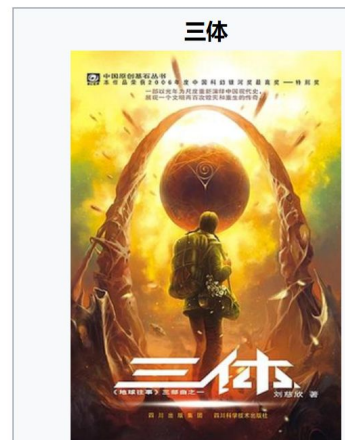
维基百科，自由的百科全书

本条目介绍的是刘慈欣所著之科幻小说。关于天体力学中的基本力学模型，请见「[三体问题](#)」。

《**三体**》是中國大陸作家**刘慈欣**于2006年5月至12月在《科幻世界》杂志上连载的一部长篇科幻小说，出版后成为中国大陆最畅销的科幻长篇小说之一<sup>[1]</sup>。2008年，该书的单行本由重庆出版社出版。本书是**三体系列**（系列原名为：地球往事三部曲）的第一部，该系列的第二部《**三体II：黑暗森林**》已经于2008年5月出版。2010年11月，第三部《**三体III：死神永生**》出版发行。2011年，“地球往事三部曲”在台湾陆续出版。小说的英文版获得美国科幻奇幻作家協會2014年度「**星云奖**」提名<sup>[2]</sup>，并荣获2015年**雨果奖**最佳小说奖。<sup>[3]</sup>

### 目录 [隐藏]

- 主体脉络
- 世界观设定
- 内容概述
- 出场角色
- 用语
- 改动



《三体》封面

作者 [刘慈欣](#)  
类型 [科幻小说](#)  
系列 [中国科幻基石丛书](#)

## Three Body (Novel) [Edit]

Wikipedia, the free encyclopedia

This article introduces science fiction r celestial mechanics, please see "[Three-Body Problem](#)".

"**Three Body**" is a full-length science fiction novel serialized by the Chinese writer [Liu Cixin](#) in the "[World of Science Fiction](#)" magazine from May to December 2006. After publication, it became one of the best-selling science fiction novels in [Mainland China](#)<sup>[1]</sup>. In 2008, the pamphlet of the book was published by Chongqing Publishing House. This book is the first part of the [three-body series](#) (the series was originally called the [Trilogy on Earth](#)). The second part of the series "[Three-body II: Dark Forest](#)" was published in May 2008. In November 2010, the third "[Three-Body III: Death Eternal Life](#)" was published. In 2011, the "Earth Past Trilogy" was successively published in Taiwan. The English version of the novel was nominated for the 2014 "[Nebula Award](#)" by the American Association of Science Fiction and Fantasy Writers<sup>[2]</sup> and won the 2015 [Hugo Award](#) for Best Novel.<sup>[3]</sup>



"Three Body" Cover

author [Liu Cixin](#)  
type [science fiction](#)  
series [China Science Fiction Cornerstone Series](#)

Chinese (Simplified) **English** ⋮ ×

Always translate Chinese (Simplified)

Google Translate

## NLP Today: Machine Translation

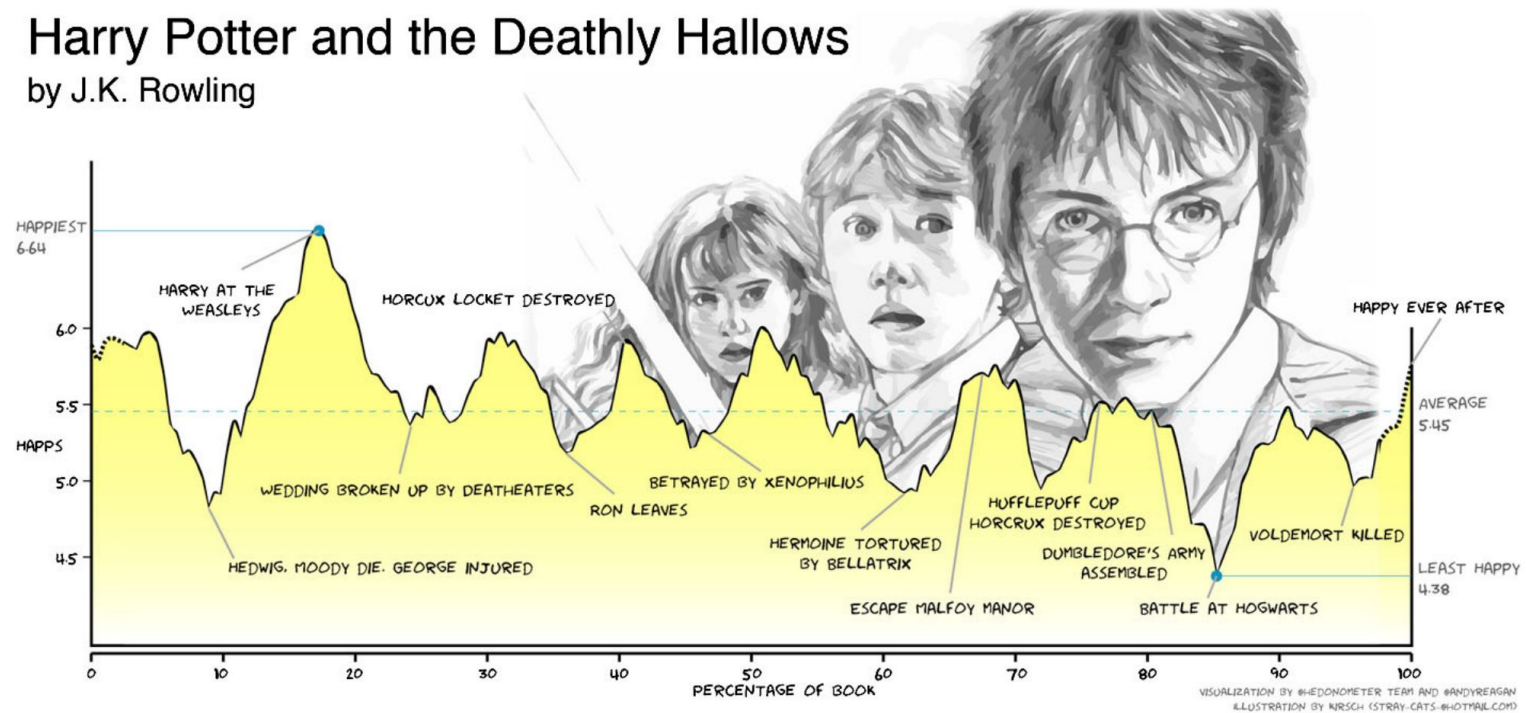
LATIN - DETECTED   MACEDONIAN   BELARUSIAN   ▾   ↔   ENGLISH   LATIN   ESTONIAN   ▾

Ense petit placidam sub libertate quietem | ×   With his sword he seeks a peaceful repose under freedom

*By the sword we seek peace, but peace only under liberty*

# NLP Today: Sentiment Analysis

Harry Potter and the Deathly Hallows  
by J.K. Rowling



Reagan et al. 2016



# ChatGPT



Rewrite the following paragraph in the style of a breezy, readable airport book:

A spectre is haunting Europe — the spectre of communism. All the powers of old Europe have entered into a holy alliance to exorcise this spectre: Pope and Tsar, Metternich and Guizot, French Radicals and German police-spies.

## From other samples



Picture t  
shadows  
that send  
forming

- [...] with some seriously big names. We're talking the Pope and the Tsar, Metternich and Guizot, and even those fiery French Radicals and crafty German police-spies.
- Picture this: the Pope and the Tsar, Metternich and Guizot,

everyone on board: the Pope and the Tsar, Metternich and Guizot, even the French Radicals and those sneaky German police-spies. It's a real clash of the titans, and they're ready to wage a battle against this spooky ideology. Hold on tight, because things are about to get interesting!

# Approximate topics list

- Introduction
- Words and regular expressions
- Text Classification and Naive Bayes
- Logistic Regression Classification
- Logistic Regression, continued
- Classification Evaluation and Generalization
- Annotations
- Text Data and Final Project Overview
- Tagging
- Context-Free Grammars
- Parsing (CKY algorithm)
- Dependencies and project work session
- N-Gram Language Models
- Word Embeddings (I)
- Word Embeddings (II)
- Midterm review
- In-class midterm
- Neural networks
- Transformers and BERT
- Generative LLMs (incl. GPT)
- Challenges and Issues for LLMs
- Final presentations