



IIT RTC Conference  
Illinois Tech @ Chicago  
Oct 8-10, 2024



IEEE

# Building Multiple Natural Language Processing Models to Work In Concert Together

David vonThenen  
Software Engineer/Developer Advocate  
  @davidvonthenen

# David vonThenen

- Are you Human or an AI?
- I want 5 Kubernetes
- Virtual Machines are Real
- Replacing Myself with Bots...
- Cloudy, cloudy, cloudy...
- There is storage for that!

  [@davidvonthenen](https://www.linkedin.com/in/davidvonthenen)



IIT RTC Conference  
Illinois Tech @ Chicago  
Oct 8-10, 2024



# Agenda

- **"Hello World": Question vs Sentence**
- **Named Entity Recognition (NER)**
  - **Obtaining/Finding the Data**
  - **Grooming and Formatting the Data**
  - **Processing Data and Building the Model**
- **Demo: Multiple NLP Models**
- **Q&A**



# **Our First NLP Model**

**Machine Learning Terms, Basics, Etc**

# Level Set with ML Models

- Data(set)
  - Domain of Problem, "Examples"
  - Search/Pattern Amongst
- Tokenzier
  - BERT uncased, DeBERTa, etc
- ML Framework
  - PyTorch, Tensorflow, fastai
- Tensor – A Measurement (Multi-Dimensional Matrix of Measured Data)
- Supporting Libraries
  - pandas, NumPy, etc





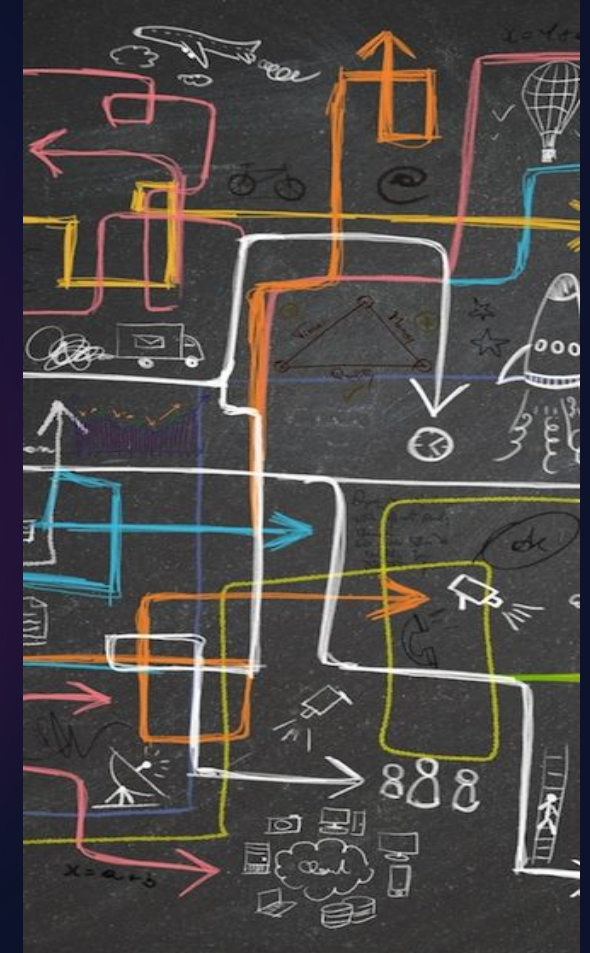
# Building Your First NLP Model

- Classification Models Easier to Understand
- Starter Model: Sentence or Question?
- Off-the-Shelf/Curated Datasets
  - Data... Lots of Data
  - Stanford Question Answering Dataset (SQuAD)
- Classify the Data:
  - Yes or No, 1 or 0



# More Complex That You Think...

- While This Seems Straightforward
  - Couldn't You Just Look For "?"
- Consider These Examples:
  - Is this an example sentence?
  - My name is John Doe.
  - How are you doing my friend
  - Tell me about the history of the United States.



# Demo

## Question vs Non-Question



IIT RTC Conference  
Illinois Tech @ Chicago  
Oct 8-10, 2024





# Qs vs non-Qs Recap

- People Don't Conform to Language Rules
- Things to Consider. Not All...
  - Questions End With a Question Mark
  - Sentences End With a Period
- More Complex Than We Think
  - Not All Question Start With:
    - Who, What, When, Where, Why, How
  - Some "Questions" End With a Period

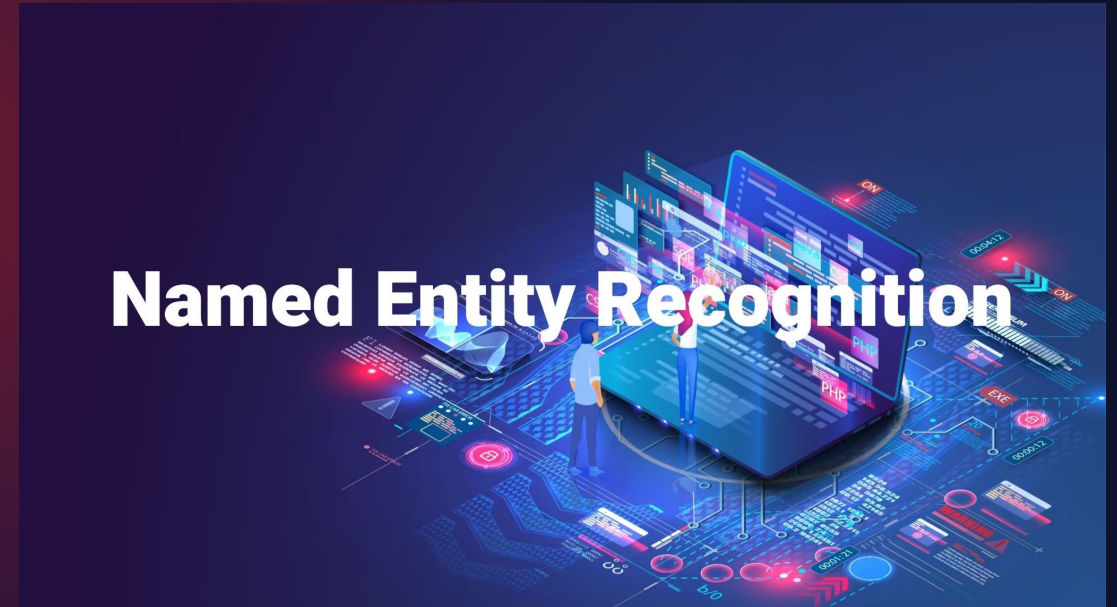


# **Building NLP Models**

## **Named Entity Recognition**

# What Are Named Entities?

- Extracting and Classifying "Things" Mentioned in Unstructured Text into Predefined Categories
- Typically Means:
  - Personally Identifiable Info
    - Name, Age, SSN, IP Address
  - Protected Health Info
    - Blood Type, Drug, Injury
  - Payment Card Industry
    - Credit Card #, CVV
- More Basic, It's Just a Label



# Obtaining/Finding the Data

- Most Difficult Part is Getting the Data
- Look Everywhere...
  - GitHub – [Entity Recognition Repo](#)<sup>1</sup>
  - [Huggingface](#)
  - [Kaggle](#) – Projects w/ Datasets
  - [Academic Torrents](#)
- and Get Creative...
  - Any [CoNLL](#)<sup>2</sup> Formatted Dataset
  - Ask Researchers! Some Will Share!
  - Synthetic Data – Be careful With This!



1: [GitHub](#) for [Juan Diego Rodriguez](#)

2: [Conference on Computational Natural Language Learning \(CoNLL\)](#)





# Grooming and Formatting

- CoNLL Format Desirable Due to Availability
  - "Standard" Widely Available Format
- The Simplistic View...
  - Capture Words in Sentences
  - Each Word is Labelled
  - Labels Apply to Multiple Words
    - United States of America
- Label = Classification!
  - PII, PHI, PCI SSC



IIT RTC Conference  
Illinois Tech @ Chicago  
Oct 8-10, 2024





# CoNLL Format – Good

## 4 Columns (Space Delimited)

<b>Word</b>	<b>Part of Speech</b>	<b>Syntactic Chunk</b>	<b>Entity Tag</b>
United	NNP	I-NP	B-ORG
Nations	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	B-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	B-LOC
.	.	O	O

# CoNLL Format – Bad

Word	Part of Speech	Syntactic Chunk	Entity Tag
United	NNP	I-NP	O
Nations	NNP	I-NP	O
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-LOC
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-PER
.	.	O	O



# Processing and Building

- After Data Is Formatted, We Need Structure!
- Word, "Tag Map" Or...

Word	O (No Entity)	B-ORG	I-ORG	I-TIME	...
United	0	1	0	0	...
Nations	0	0	1	0	...
is	0	0	0	0	...

- Tokenizer = bert-base-uncased
- Each Sentence Composed of a Tensors for:
  - Tokens, Entity Labels, Attention Mask (Padding)



# Demo

## NER + Mult-Model

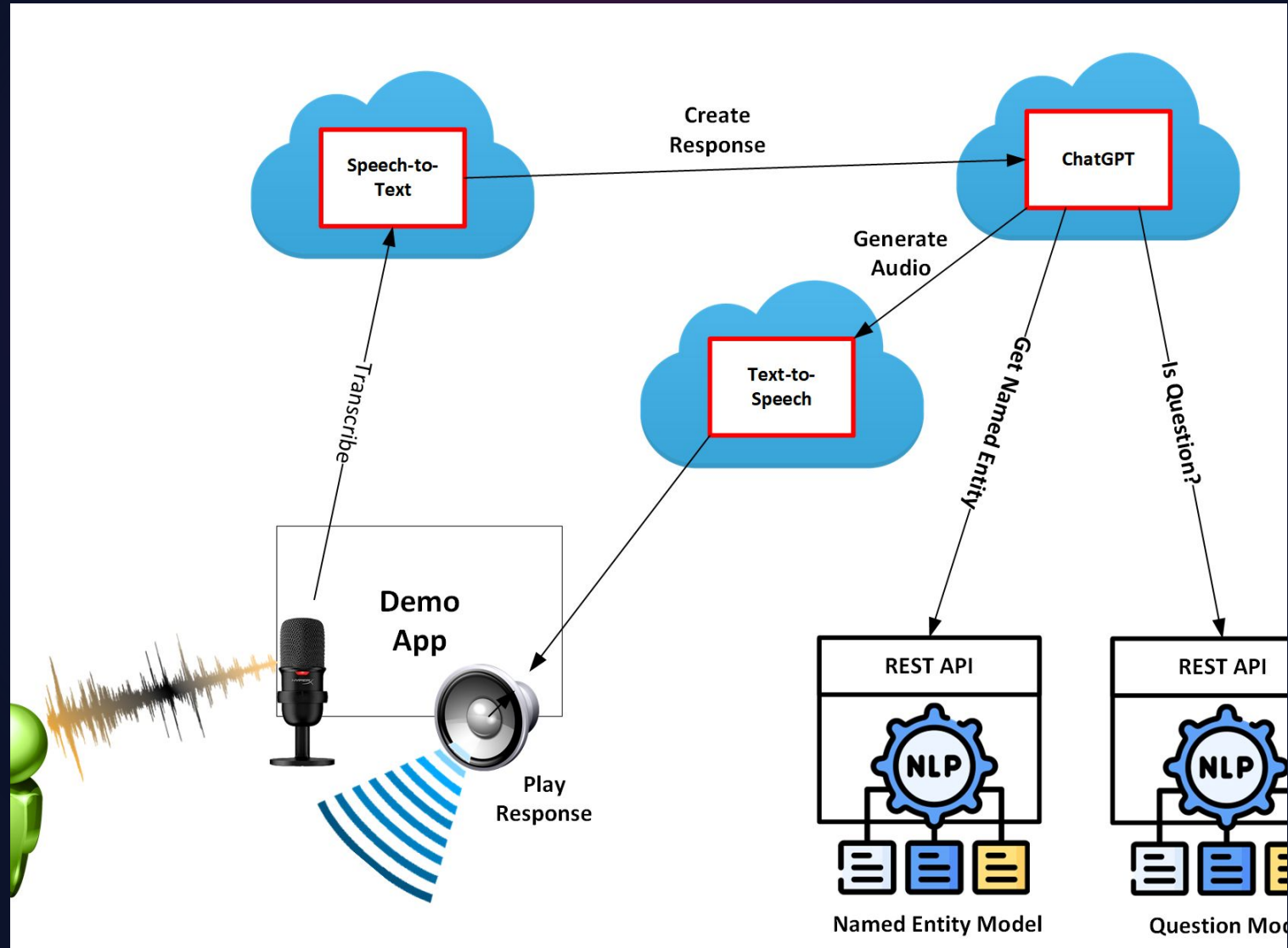
<https://youtu.be/EVUXxm8gSzl>



IIT RTC Conference  
Illinois Tech @ Chicago  
Oct 8-10, 2024



# Demo Components





# Named Entity Recognition Recap

- Find Datasets and Start With Low-Hanging Fruit
  - Custom Data(sets)?
- Most Difficult: Grooming the Data
  - Does Data Accurately Reflect the Problem
  - Fix the Data! Correct the Errors
- Structure the Data for ML Training
- Generate the Model, Does It Work?
- Rinse and Repeat, Always Outliers
- Iterative Improvements, Refinement



# Presentation Resources

# Resources

[\[CLICK HERE\] for All Material Contained in this Session \[CLICK HERE\]](#)

Code with Instructions for:

- Part 1: [Building a Question Classification Model](#)
- Part 2: [Building a Named Entity Recognition Model](#)
- Part 3: [Final Demo Used in this Presentation](#)

Other Resources:

- [Deepgram Speech-to-Text: API and Docs](#)
- [Deepgram Text-to-Speech: API and Docs](#)
- [Juan Diego Rodriguez](#) – [Named Entity Repo](#)



IIT RTC Conference  
Illinois Tech @ Chicago  
Oct 8-10, 2024





# Questions?



**IIT RTC Conference**  
**Illinois Tech @ Chicago**  
**Oct 8-10, 2024**



# Thank You!

**David vonThenen**

  [@dvonthenen](https://www.linkedin.com/in/dvonthenen)

<https://linktr.ee/davidvonthenen>



**IIT RTC Conference**  
**Illinois Tech @ Chicago**  
**Oct 8-10, 2024**

