Harvard John A. Paulson
School of Engineering
and Applied Sciences

Spotify

Disagreement Detection

Final Presentation

Connor Capitolo *Max Li* *Morris Reeves* *Jiahui Tang*

Overview

- Problem statement & motivation
- Data
- Literature
- Modeling approach
- Evaluation & interpretation
- Conclusion

What is disagreement?

Disagreement comes in different forms...

Playful

Pie in America

Serious

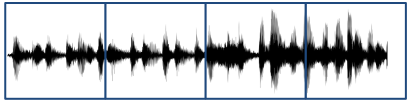
Australian Teens Discussing Abortion

What are we trying to achieve?

Generate insights about disagreement based on the Spotify English-Language 100k Podcast Dataset

Inputs

Audio and text



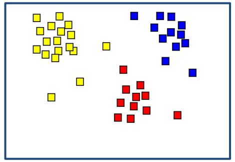
No, that's absurd..	Ok, moving on...	This is a good podcast	No, it's been quite bad...

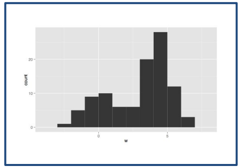
Outputs

Disagreement/Non-disagreement

D	N	N	D

Insights





Harvard John A. Paulson School of Engineering and Applied Sciences

Why is disagreement detection important?

Better User Experience
Tailored recommendations

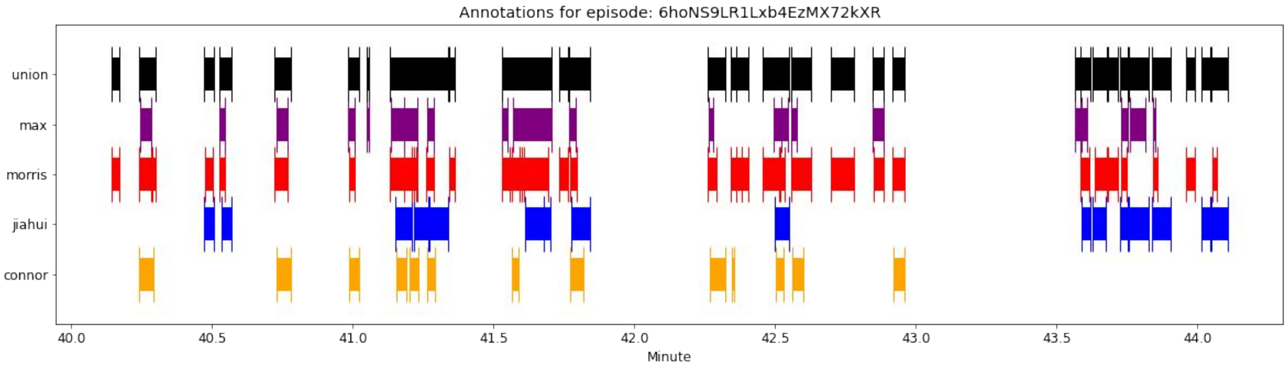
Drive audience engagement
Improve style or moderation

Better Content Moderation
Happier Users + Happier Creators = Happier Spotify

Definition of Disagreement

- Generate disagreement labels that can be used for modeling
- Definition: a speaker is **directly contradicting or rejecting** another person's idea where it is **immediately perceptible** by the listener
    - Rule of thumb: If you handed this podcast to a stranger, they would know it's disagreement

Dog clip: a couple discussing what it means to have a dog when single



Harvard John A. Paulson School of Engineering and Applied Sciences

Data

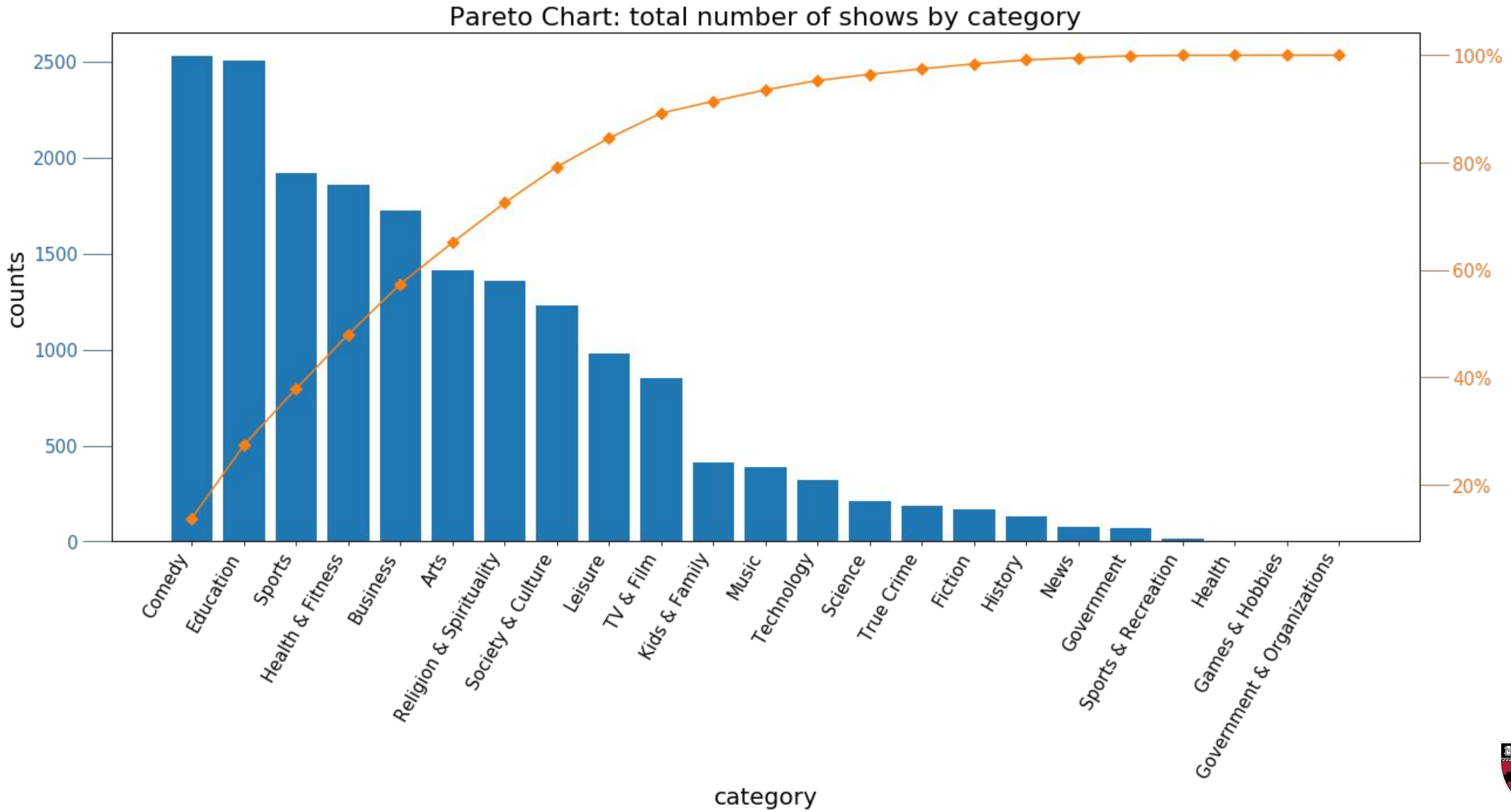
- **105,360 episodes**
- **8,360 shows**
- **23 genres**

audio
~ 50k hours

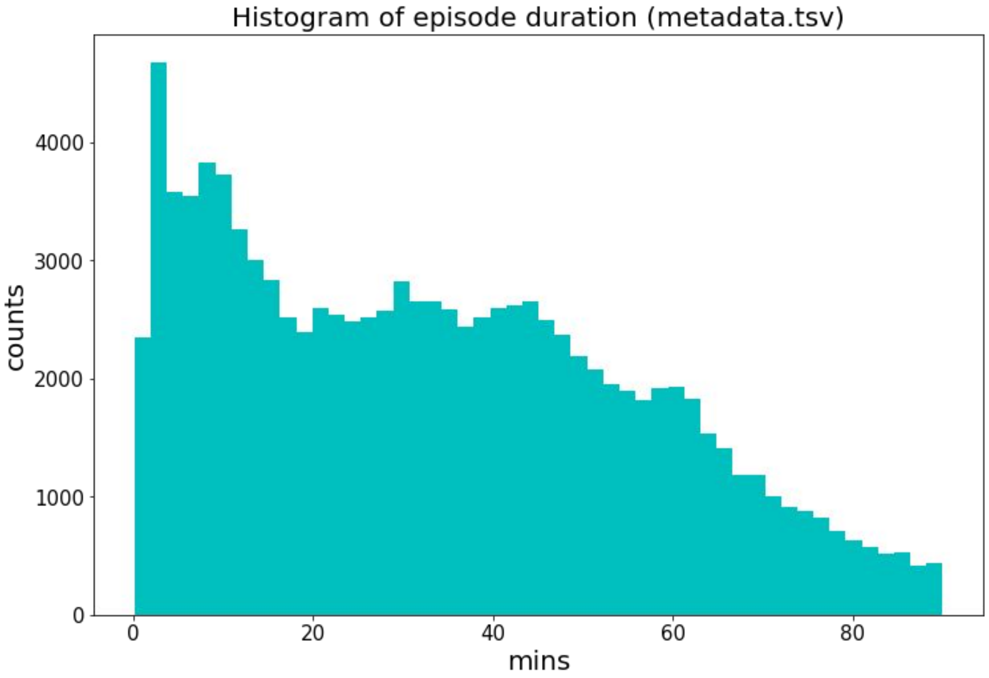
text
> 600M words

	min	average	max
minutes	<1	31.6	305.0
words	11	5,728	43,504

Data - podcast genre



Data - episode duration



Data

- **105,360 episodes**
- **8,360 shows**
- **23 genres**

audio
~ 50k hours

text
> 600M words

	min	average	max
minutes	<1	31.6	305.0
words	11	5,728	43,504

How do we handle all this data?
Where are disagreements?

Data

	getting target labels	finding how to annotate	finding episodes to annotate	combining annotations	discretizing audio and text data
Problem	definition of disagreement	consistent formatting & ease of use	random podcast selection not efficient	multiple annotators	segment & merge
Solution	immediately perceptible contradiction	prodigy	functions for searching metadata & transcripts	functions for compiling & unioning annotation	sliding window with adjustable parameters

text
audio (and text)

Literature overview

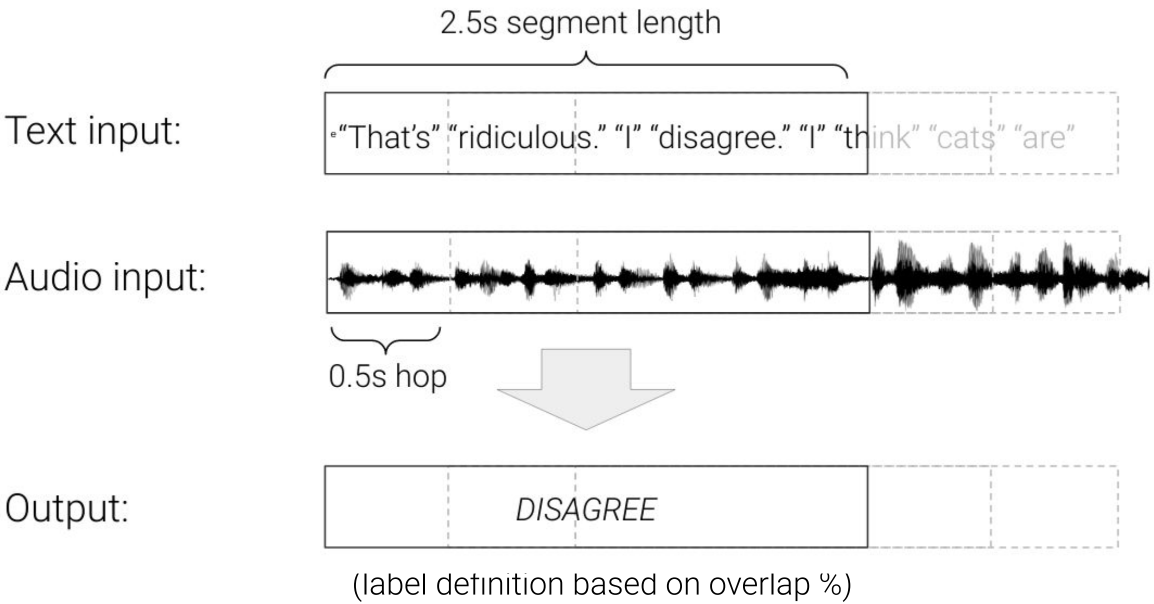
	Xu et al. (2019)	Gokcen and de Marneffe (2015)	Wang and Cardie (2014)	Wang et al. (2011)	Hillard et al. (2003)
Model class	RNN	Logistic regression	Conditional random field	Conditional random field	Decision tree
Features	GloVe word vectors	n-grams speech acts dependencies	n-grams dependencies TFIDF	n-grams speech rate pitch	words pause fundamental frequency
Data	Tweets	Online forums (IAC)	Wiki Discussions	English Broadcast Conversations	Meetings (ICSI)

Harvard John A. Paulson School of Engineering and Applied Sciences

Literature: our approach in context

- Modeling on **podcast data** is relatively unexplored
  - Diversity of podcast episode genres, topics
  - Disagreement in podcasts may be more subdued
- Limited exploration of **multivariate audio features** (DWT, spectrogram)
  - Existing approaches: primarily count-based or univariate features

Modeling: inputs and outputs



Harvard John A. Paulson School of Engineering and Applied Sciences

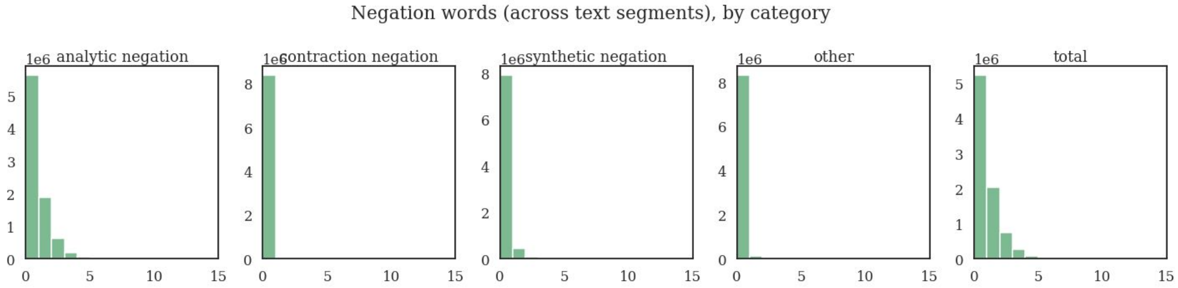
Model overview

Category	***Input Features***	***Model Class***
Baseline	Random coin flip	*NA*
Text	Negation word counts	Logistic Regression
	Avg. word2vec embedding	Logistic Regression
Audio	Mel spectrogram	CNN
	Discrete Wavelet Transform (DWT)	Logistic Regression
	DWT	Random Forest
	DWT	XGBoost
	DWT	LSTM

Text: count-based model

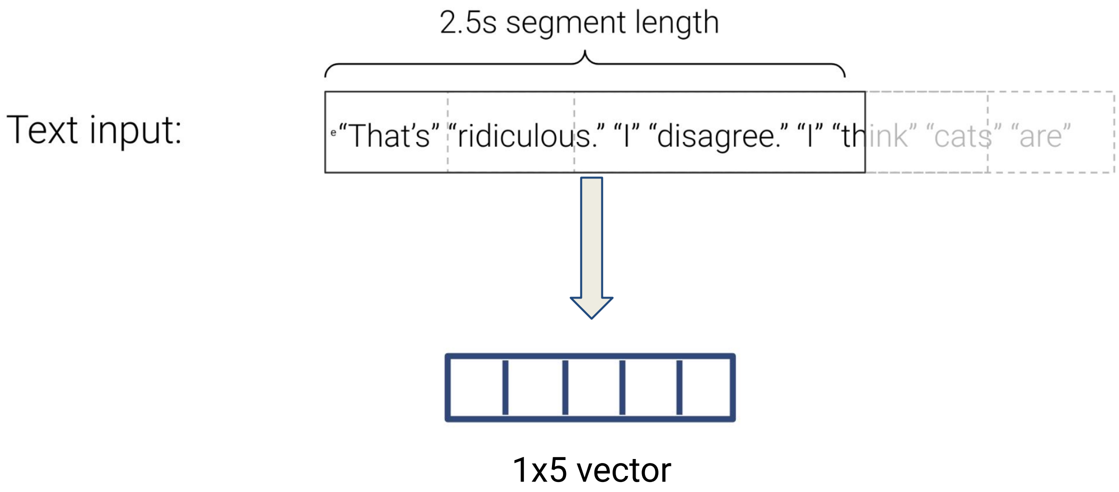
Hypothesis: p(disagreement) = f(count of “negation words”)

Category	Words
Analytic negation	no, not
Contraction negation	ain’t, aren’t, didn’t, shouldn’t, etc.
Synthetic negation	neither, never, nor, none, nobody, noone, no-one
Other	disagree, incorrect, wrong, ridiculous, absurd



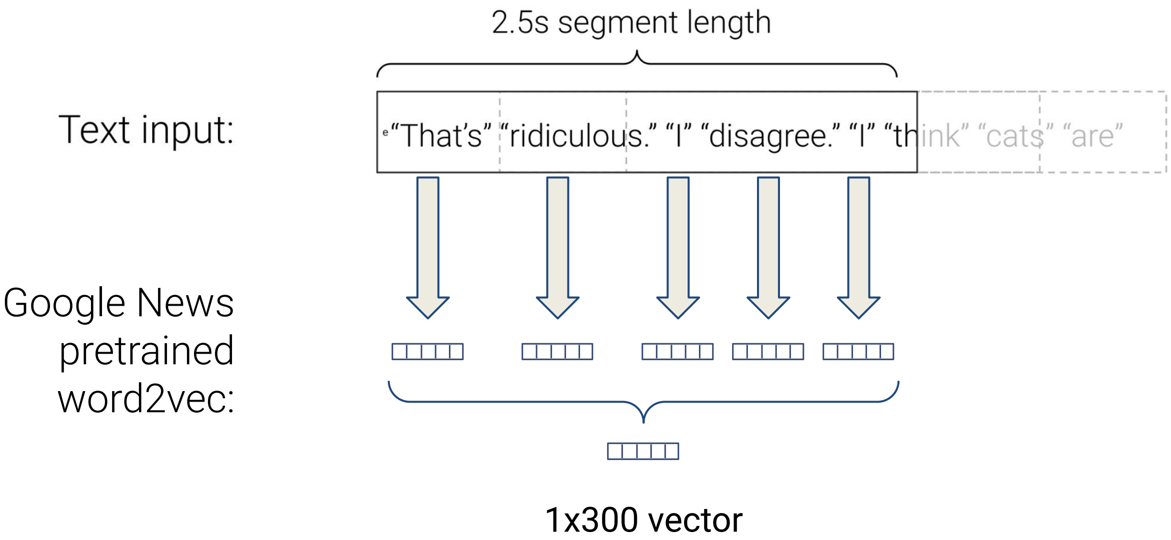
Text: count-based model

Hypothesis: p(disagreement) = f(count of "negation words")



Text: embedding-based model

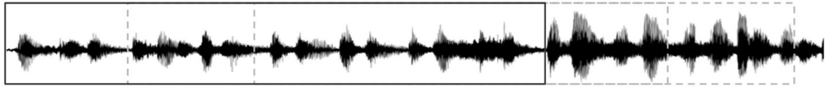
Hypothesis: p(disagreement) = f(word embeddings)



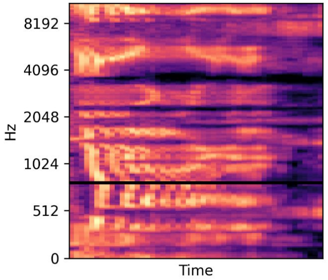
Audio: Spectrogram CNN

Hypothesis: p(disagreement) = f(changes in frequency, amplitude over time)

Audio input:



Mel spectrogram:
FFT: time → frequency
STFT: FFT on windows
Mel: nonlinear freq. scale



CNN

Discrete Wavelet Transform

Audio input:

N observations of 2.5 second sliding window audio chunks

features: *discrete wavelet coefficients*

0.003	0.219	…			

Models using DWT

- **Logistic Regression**
- **Random Forest**
- Boosting
  - **XGBoost**
  - AdaBoost
  - GradientBoost
- **LSTM**
  - time series don't have i.i.d assumptions
  - LSTM could better capture patterns in sequential data

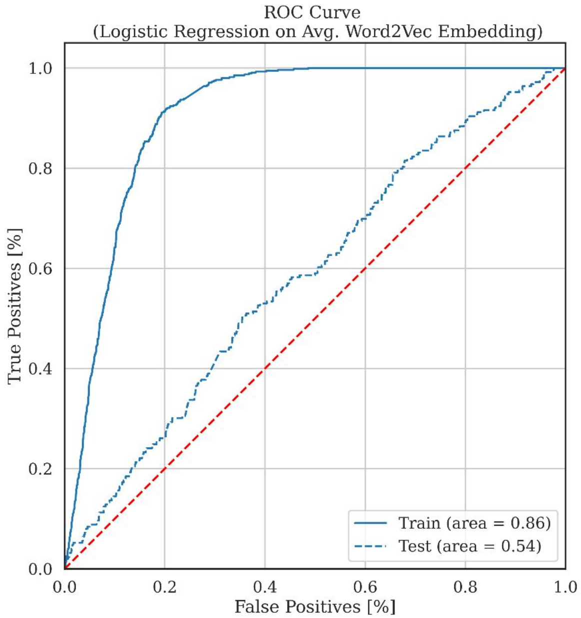
Evaluation/Results

Category	***Name***	***Precision: Test Set 1*** *(Australian teens)*	***Precision: Test Set 2*** *(2 "Hot Take" eps.)*
Baseline	Naive	0.04	0.04
Text	Word count	0.09	0.07
	Word2vec	0.06	0.03
Audio	Spectrogram	0.04	0.06
	Wavelet - Logreg	0.04	0.02
	Wavelet - RF	0.00	0.00
	Wavelet - XGBoost	0.00	0.00
	LSTM	0.03	0.07

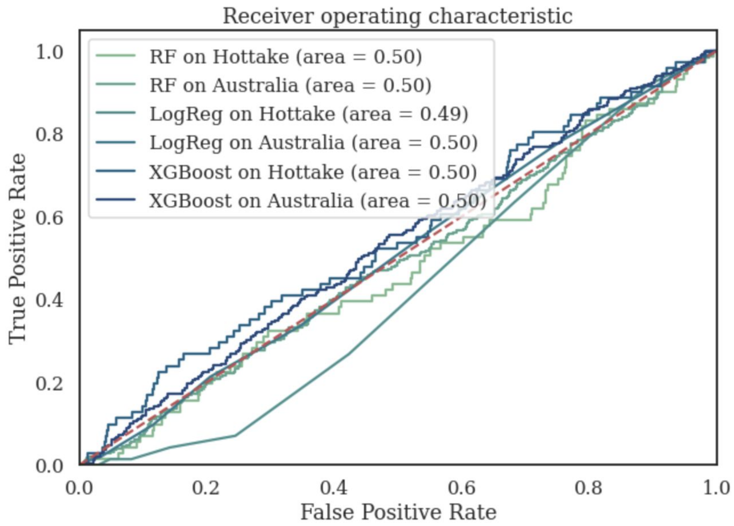
doesn't generalize to different test episodes/domains

Evaluation/Results

Word2vec model



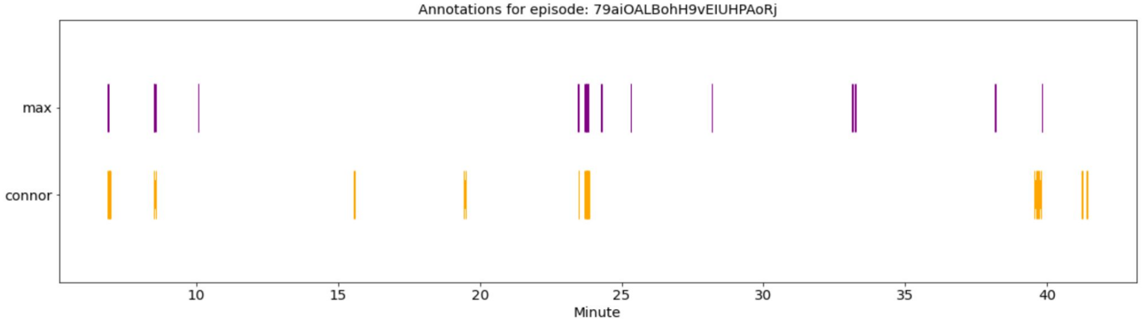
Audio models



Interpretation

Sparsity of Disagreement (Class Imbalance)

Dating + Food episode: 4 people discussing dating and restaurants in NYC



- Severe class imbalance makes classification difficult (only 4% of either test set is disagreement), even with class weights and augmentation

Harvard John A. Paulson School of Engineering and Applied Sciences

Interpretation

Challenges in Generalization

- Episodes span different genres + styles, making it difficult to generalize
- Train-test splits within the same podcast episode yield much higher precision (0.4 - 0.5) than train-test splits across episodes
  - ***One potential solution:*** pre-train a large model with general dataset and fine-tune it over first few mins of testing episode

IID Assumption

- Models treat segments as independent, constraining performance

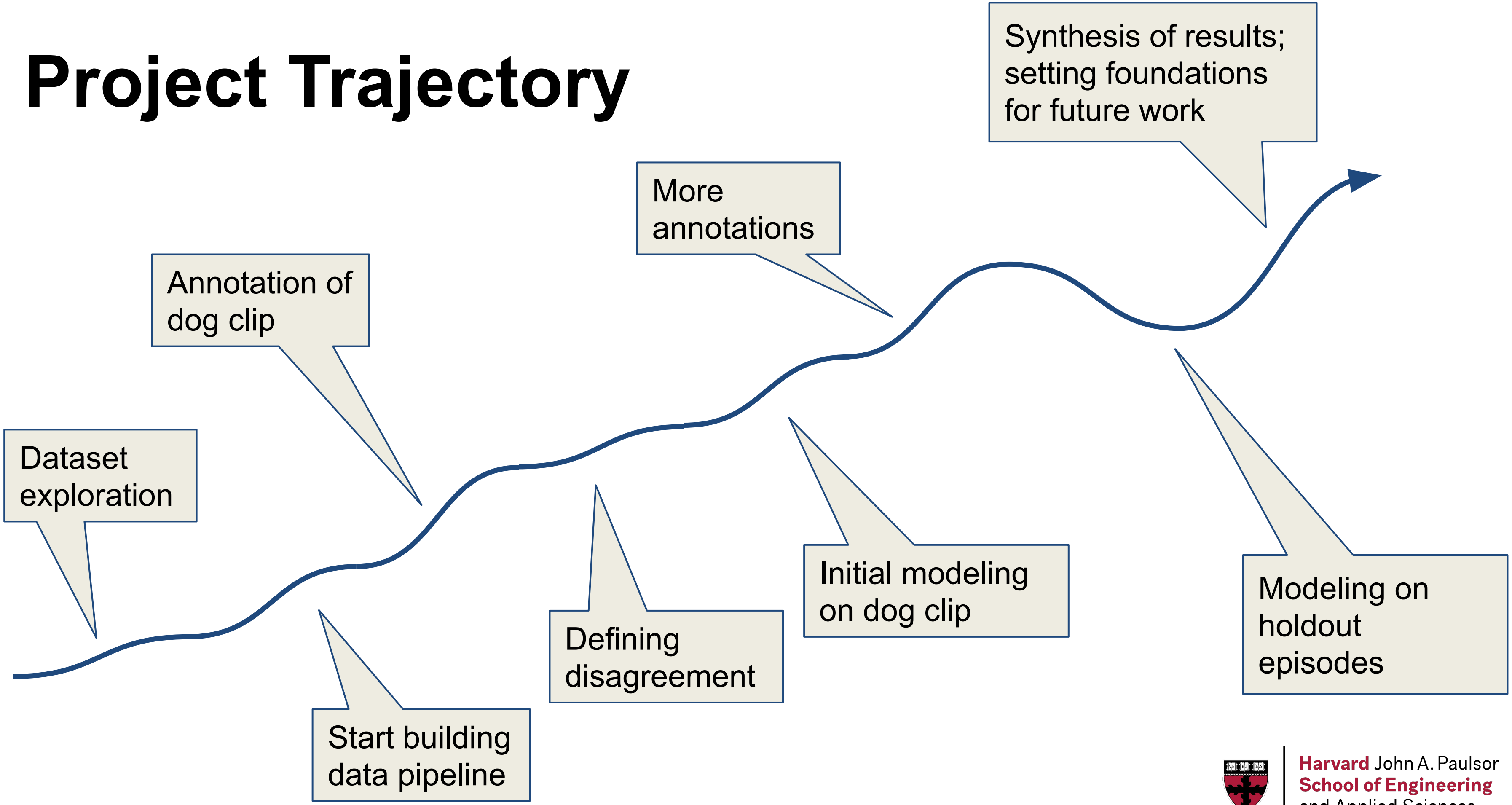
Future suggestions

Use **speaker changes** as feature

Perform **data augmentation** for invariance to speaker characteristics

Consider longer **contexts** and autoregressive approaches

Project Trajectory



Key Contributions

- Creation of first-in-kind annotated disagreement **dataset**
- Data **pipeline** for discretizing audio and text data
- Identification of **categories** of disagreement
- Baseline **modeling** approach
- Identification of key **challenges** that a successful model need to address: sparse target class, data augmentation, generalization

Special Thanks

Our partners: **Rosie Jones** and **Jussi Karlgren**

Our TFs: **Eagon Meng** and **Nick Stern**

Our Professor: **Chris Tanner**

Q&A