Project STATIC

Scanning
Transmissions
And
Transcribing
Information
Cartographically

May 2020

Alex Golden
Jungmoon Ham
Luke Podsiadlo
Zach Tretter

Project 5 (Client Project)
Data Science Immersion Cohort 11
General Assembly, Boston

Table of Contents

- Problem Statement
- Bottom Line Up Front
- Project Methodology
- Data Collection
- Audio Processing
- Natural Language Processing
- Geocoding & Mapping
- Conclusions & Next Steps

Prompt & Problem Statement

<u>Using live police radio reports for real time identification of people needing assistance.</u>

"Currently, FEMA identifies areas that require immediate attention (for search and rescue efforts) either by responding to reports and requests put directly by the public or, recently, using social media posts. This tool will utilize live police radio reports to identify hot spots representing locations of people who need immediate attention. The tool will flag neighborhoods or specific streets where the police and first-respondents were called to provide assistance related to the event."

Questions we sought to answer

- What geographic and situational information is provided by (live) police radio?
- How can this information be displayed on a map?

Principal Data Science Elements - Webscraping, APIs, NLP

Bottom Line Up Front Results

- Police radio provides some geographic information
- Translating radio to discrete information inhibited by
  - Poor audio quality of raw feed
  - Inherent limitations of speech-to-text tools
- Mapping of geographic information in radio transcript obfuscated by
  - Unintelligibility of police chatter to layman
  - Ambiguity of street names in an urban area
  - Uncertainty of precise location even with a clear address

Limitations in Raw Data Impact Utility of an Automated Speech-to-Text-to-Location Tool

Approach to Problem

- Reviewed and leveraged work by prior DSI cohorts
- Split project into modules by data type and workflow step

Project Step	Input	Output
Audio Processing	Raw Audio	Structured Audio
Speech Recognition	Structured Audio	Raw Text
Natural Language Processing	Raw Text	Structured Text
Geocoding	Structured Text	Map

- Built small scale model then scaled to larger data set
- Managed and tracked tasks via GIT project board

Review of Work by Prior Cohorts

Project Title	"Camp Fire Radio-to-Location"	"Red Siren"	"Mapping Emergency Dispatch Transmission"	"San Francisco Dispatch Audio Mapping"
Authors	DSI-NYC-7 (Spring 2019) Link to GitHub Mitchell BohmanNour Zahlan, Masiur Abik	DSI-CHI-7 (Spring 2019) Link to GitHub Lance Carroll, Neal Manahan Rodolfo Flores Méndez Sadorkhon Tursunov, Blake Wallace	DSI-ATL-8 (Summer 2019) Link to GitHub Anthony Chapman, Carol Chiu, Kwamae Delva, Joseph Hopkins	DSI-SF-9 (Fall 2019) Link to GitHub Grant Wilson, J. Hall, Gabriel Perez Prieto
Project Focus	The Camp Fire (Butte County California) in November 2018	"...various emergencies and non-emergencies..." across multiple years	Atlanta Police Zone 5 and Fire Dispatch from late July 2019	San Francisco City Police Dispatch from early November 2019
Data Sources	Broadcastify	Broadcastify Youtube	Broadcastify Broadcastify Archive Toolkit	Broadcastify Broadcastify Archive Toolkit
Suggested Focus Areas for Future Work	Managing the large quantities of large audio files	Google Cloud's speech-to-text LiBrosa Features	Encourage future cohorts to leverage or iterate the (broadcastify-archtk)	Dolby API

Project Workflow

- Downloaded archived police audio from Broadcastify.com
- Webscraped Greater Boston Area Street Names
- Webscraped Police Codes

- pydub to splice audio
- Dolby API to:
  - Background Noise/ Static
  - Isolate Speech
- Google Transcribe API to for Speech to Text

- Dataframe of raw and structured text built with spacy and ntlk
- Tokenized raw text to match with list of streets
- Generated list of potential street addresses

- Import dataframe with potential addresses
- Find latitude & longitude via google geocoding
- Plot geocodes on map and visualize with folium

Principal Data Sources

Audio Files

Archived radio feeds from Broadcastify using the **B**roadcastify **A**rchive **T**ool

Boston Metro Area Police and Fire West (Feed ID 25818)

Street Names

Compiled a complete list of all greater Boston area street names by scraping 'geographic.org' using BeautifulSoup

Police Codes

Gathered a data dictionary .csv file of all Police codes from 'bearcat1.com'

*Not used

Case Study

Transcript from April 2013 manhunt of marathon bombers

Yielded 1512 row Dataframe

32 hours of raw audio spliced into 1512 chunks

1512 audio chunks transcribed

NLP marked 417 as having address data

1418 potential addresses to geocode

Audio Processing - Overview

Clean

Using the DOLBY API we cleaned audio to isolate speech from background noise and static.

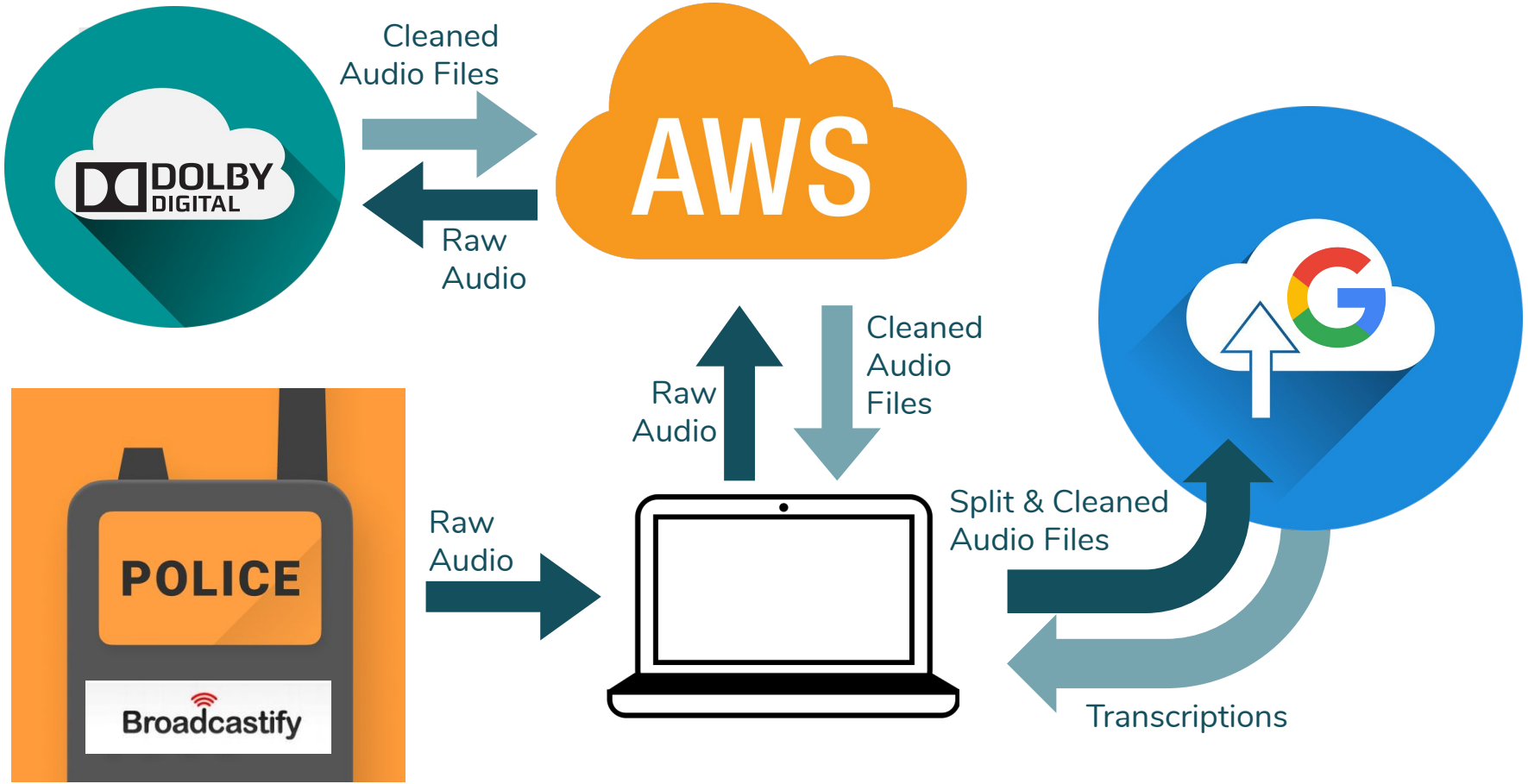
Split

Split the longer audio files into individual segments.

Transcribe

Using the GOOGLE Transcribe API we transcribed each segment.

Audio Processing - Workflow



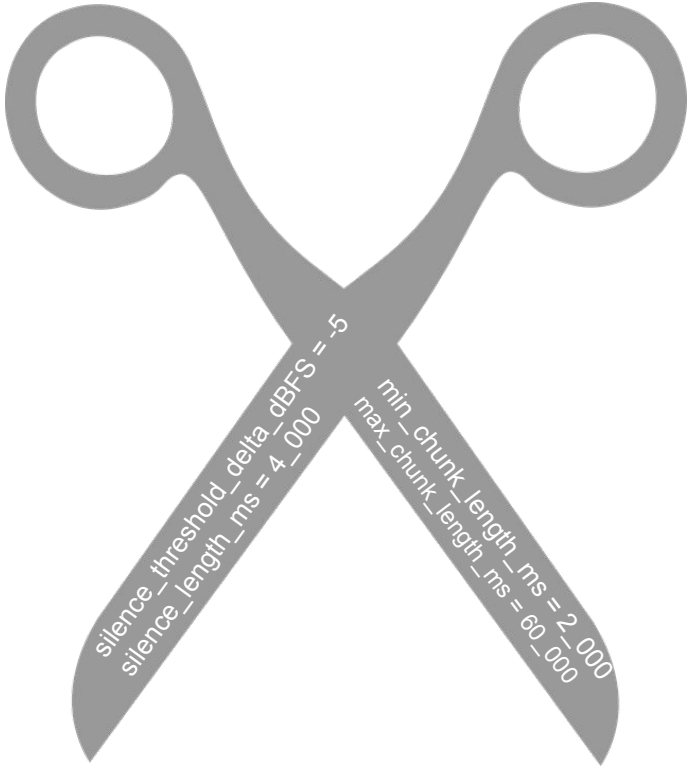
Audio Processing - Cleaning

01	Loudness	• Applying loudness correction scales the loudness of the audio segments. • This increases loudness of very quiet segments.
02	Dynamics	• The dynamics algorithm provides basic correction of effects from bad microphones or mixing. • This setting was turned up to ‘max’.
03	Noise	• This feature reduces background noise such as HVAC, or machine hums. It targets sound that is consistent throughout a recording.
04	Speech	• Enhances the quality of recordings through speech isolation and sibilance smoothing.

Audio Processing - Splitting

- In order to identify individual transmissions, the large audio files are split into small segments.

- Splitting is done with the **PYDUB** audio package.
  - The method splits based on silence like a string method would split on commas



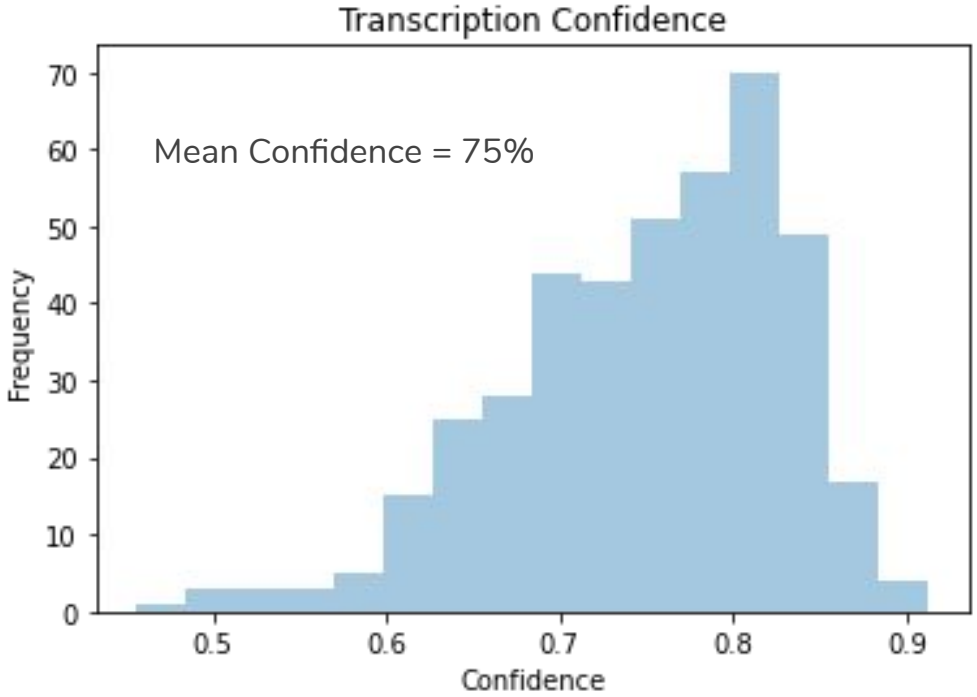
Audio Processing - Transcribing

Transcription was done through Google Cloud Services which uses its own python wrapper: **speech_v1p1beta1**

Providing expected phases such as street names decreased the quality of the transcriptions considerably .

Confidence in this model is defined by Google as “likelihood that the individual words were recognized correctly”.

Google builds multiple alternative transcripts and the ‘best’ is not solely based on confidence, for example sentence context also factors in.



Natural Language Processing

- Libraries and Primary Functions Used
- Process
- Results
- Synopsis

NLP - Libraries and Primary Functions Used

spacy - Advanced Natural Language Processing Library

- *Matcher* - matches sequences of tokens together

NLTK - Classic Natural Language Processing Library

- *RegexpTokenizer* - Breaks blocks of text into a list of words

usaddress - Python library designed specifically for parsing unstructured address data

NLP - Identify and Match Street Names

Spacy Matcher takes tokenized transcript data and extracts tokens that match with the reference list of greater Boston area street names

01	**Transcript**	• **Thomas 155 Lexington Street South Korea parking lot for your nearest box for four or five like this year Santa Fe 12:05**
02	**Tokenized Transcript (Input)**	• **[Thomas, Lexington, Street, South, Korea, parking, lot, for, your, nearest, box, for, four, or, five, like, this, year, Santa, Fe]**
03	**Possible Street Names (Output)**	• **Thomas** • **Lexington** • **South**

NLP - Find Possible Street Numbers and Append to Address

usadress extracts street address numbers from a string

01	Transcript	• Thomas 155 Lexington Street South Korea parking lot for your nearest box for four or five like this year Santa Fe 12:05
02	Street Address Number (Output) - usaddress	• 155
03	Possible Street Names (Output) - spacy matcher	• Thomas • Lexington • South
04	Potential Full Addresses (Final Output)	• 155 Thomas • 155 Lexington • 155 South

Excerpt of NLP Built Dataframe

	file_name	confidence	transcript	tokens	streets	numbers	full_streets
0	sample116-25818-20200501-0941.wav	0.764874	returning from the second alarm to go off in fact about	[returning, from, the, second, alarm, to, go, off, in, fact, about]	The , Off	[]	[]
1	sample92-25818-20200501-0941.wav	0.692422	having a t show jokes to 19219 South Street	[having, a, t, show, jokes, to, South, Street]	South	[19219]	[19219 South]
2	sample635-25818-20200501-1439.wav	0.643082	but I'm trying to we went through a window we found a party took social housing waiting for the window	[but, I, trying, to, we, went, through, a, window, we, found, a, party, took, social, housing, waiting, for, the, window]	The	[]	[]
3	sample396-25818-20200501-1140.wav	0.787601	CMS said that they were told by somebody up here that they were up and they walked off and went downstairs that needs to be fine at least jumped on	[CMS, said, that, they, were, told, by, somebody, up, here, that, they, were, up, and, they, walked, off, and, went, downstairs, that, needs, to, be, fine, at, least, jumped, on]	Off	[]	[]
4	sample217-25818-20200501-0612.wav	0.722337	for calling Broadway	[for, calling, Broadway]	Broadway	[]	[]
...	...	...	...	...	...	...	...
1508	sample-9-25818-20200502-2159-19588.wav	0.832902	off of 104 1st Ave on the Callback	[off, of, Ave, on, the, Callback]	Off , 1st , The	[104]	[104 1st, 104 Off, 104 The]
1509	sample-9-25818-20200502-2229-19283.wav	0.721903	chat again you have like a beeping noise in the background	[chat, again, you, have, like, a, beeping, noise, in, the, background]	The	[]	[]

Correctly identified street address

No address content

NLP identifies multiple address options

NLP - Synopsis

Pros

- Relatively simple process once setup
- Quick to process thousands of lines of transcript

Cons

- Often extracts multiple potential addresses from same line of transcript
- Confuses words that are not streets with street names (The, off, etc.)
- By necessity strips identifying words like street, ave, blvd
- Difficult to scale beyond local use

Mapping Overview

Mapping Process

1. Build list of potential addresses from NLP dataframe
2. Geocode (find latitude and longitude) via google geocoding API
3. Plot geocodes on map using folium package

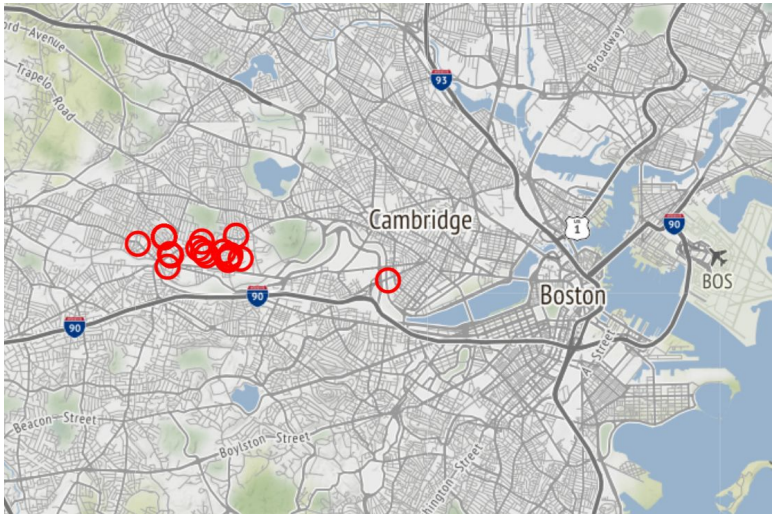
Limitations

- Can’t verify accuracy of address
- Street name by itself does not provide precise location

Mapping Example

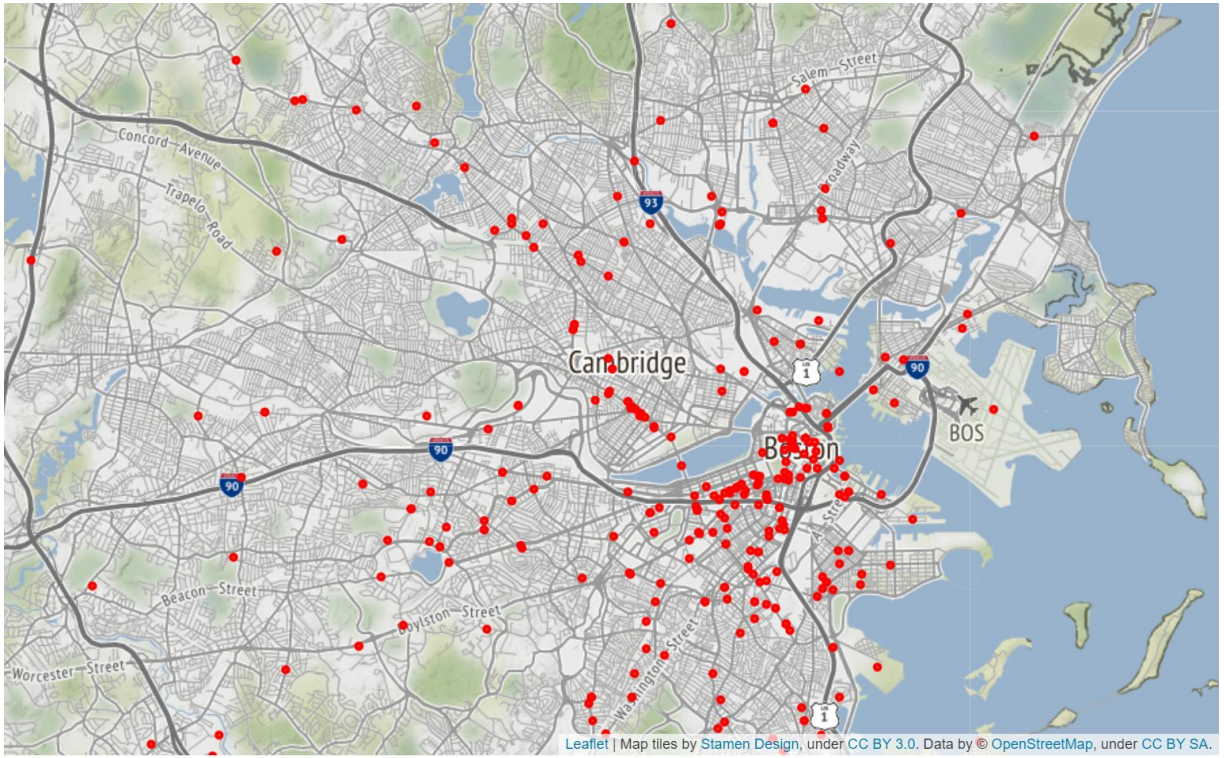
Watertown Manhunt/Shootout of Marathon Bombers (18 Apr 2013)
Note - Audio was manually transcribed from a recording

full_streets	transcript
[816 Memorial, 816 Black, 816 Drive, 816 Eastern, 816 The, 816 Cambridge, 816 Middle, 816 Station]	One 5'7", the second with darker skin, both suspects armed with firearms, driving a Black Mercedes SUV. Carjacked at 816 memorial drive at the gas station in Cambridge, Suspects are two middle eastern males, on is 5'7", second one with
[609 The]	Is there an officer driving the 609 right now?
[94 Long, 94 Spruce]	Shots Fired! Shots Fired! Officers pinned down 94 Spruce Street. I need backup. 94 Spruce Street I need long guns. I need long guns!
[111 Dexter, 111 Hazel]	111 there is an officer down at Hazel and Dexter.
[94 Long, 94 Spruce]	94 Spruce Street, I need long guns. 94 Spruce.
[982, Watertown]	982, Patch it up with probably Watertown, yea Watertown for now -there is a report of second officer down, there is definitely hand grenades and automatic gunfire.
[1181 Lincoln, 1181 Spruce]	1181 I am at Spruce and Lincoln
[13, Dexter, 13, Laurel, 13, The]	13, CP I am on foot in the backyards between Dexter and Laurel.
[2 Spruce]	Okay once we complete those 2 blocks - expand it two more blocks from those areas. okay - we have plenty of police officers lets start using them - from 98 Spruce, first 2 blocks then 4 blocks.
[543 The, 543 Auburn]	inaudible.....we have a package on the ground at Mt Auburn - 543 Mt. Auburn.
[144 Dexter, 144 Laurel]	inaudible.....at 144 Dexter Ave and Laurel Street
[144 Dexter]	Squad 2 144 Dexter Ave - Officer down.
[19, The]	19, I am looking at the two right now - I am looking at the one a
[526 Spruce, 526 Auburn, 526 The, 526 Upland]	526 Mt Auburn and the one at Upland and Mt. Auburn(officer at Spruce) correct and there definitely the suspects?
[526 Auburn]	I don't know who this second person is at 526 Mt Auburn
[19, Auburn, 19, Dexter, 19, Laurel, 19, Upland, 19, The]	19, I don't know - I gotten word that there's one at Upland and Mt. Auburn in custody, I've gotten word there's another one at Dexter and Laurel - fallen down waiting for EMS - the second party is not out of his ....inaudible



Mapping Example

Boston Police Feed 01 May 2020 to 02 May 2020



Successes

Geographic information can be extracted from police radio.

Establishing a clean workflow to download, clean, parse, and transcribe audio files.

- Successfully integrated a number of APIs
- Extracted geolocations and plotted them on a map

Limitations

Quality of the raw audio from police/dispatch radios.

Limitations of transcription software

Ambiguous nature of police chatter

- Garbage in → Garbage Out
- No speech-to-text tool can create more information than exists in the raw data
- A cop says they're reporting to a disturbance on Mass Ave...well Mass Ave is 16 miles long...

Future Work

If our group had more time we would...

Explore different audio processing tools/techniques, research standards/rules for police chatter on police radio to tailor model to recognize jargon, integrate input from live broadcasts, expand NLP for multi word text strings

Potential unique explorations for other groups

- Non-urban radio source
- Dedicated 911/fire vs police
- Focus on audio processing
- Focus on NLP for street names