48 United States,

4 years, and

3,000,000 Traffic Accidents

Lynn Pepin

sorry hawaii and alaska

Overview

Plan overview

- 5 min presentation
- 2min short Q&A

Do the following:

1. Goal(s) and overview of your project – Invent a project name or your dashboard name.
2. List examples of problems and examples of analyses you plan to support through the data visualization and through your dashboard
3. Identify and list data sources

Dashboard visualization

Include some visualization

1. Include mockups
2. **Short demo of prototype dashboard** developed in HW 6/7 (!!!)

Not MusicNet

- 23GB of data, hard to work with
- Couldn't find anything interesting for data visualization :(
- Swapped to Traffic Accidents ~2 weeks ago

Data: **US Accidents**

US Accidents on Kaggle (from Sobhan Moosavi)
https://www.kaggle.com/sob hanmoosavi/us-accidents

- Data from Feb 2016 to Dec 2020 across 48 states.
- About **3.0 million total.**
- Contains (among other things) **time**, traffic delay **severity**, and **location.**
- Location (latitude/longitude) can be paired with map data.
- Weather, visibility, and road conditions, plus other details!

Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Description	...	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop	Sunrise_Sunset	Civil_Twilight	Nautical_Twilight	Astronomical_Twilight
3	2016-02-08 00:37:08	2016-02-08 06:37:08	40.10891	-83.09286	40.11206	-83.03187	3.230	Between Sawmill Rd/Exit 20 and OH-315/Olentang...	...	False	False	False	False	False	False	Night	Night	Night	Night
2	2016-02-08 05:56:20	2016-02-08 11:56:20	39.86542	-84.06280	39.86501	-84.04873	0.747	At OH-4/OH-235/Exit 41 - Accident.	...	False	False	False	False	False	False	Night	Night	Night	Night
2	2016-02-08 06:15:39	2016-02-08 12:15:39	39.10266	-84.52468	39.10209	-84.52396	0.055	At I-71/US-50/Exit 1 - Accident.	...	False	False	False	False	False	False	Night	Night	Night	Day
2	2016-02-08	2016-02-08	39.10148	-84.52341	39.09841	-84.52241	0.219	At I-71/US-50/Exit	...	False	False	False	False	False	False	Night	Night	Night	Day

Motivation: Traffic accidents are a problem!

- We can't control **the risk** of driving.
- People feel comfortable with driving.[citation needed]
- How can we **help people understand these stats?**

The answer: **Through an interactive data dashboard!**

Objective: **Create a tool** that lets people explore **where** and **when** traffic accidents happen in the continental United States.

Hypotheses

1. "There more crashes during dawn, dusk, night"
2. "There are more crashes in rain"
3. "Crashes occur more often at intersections"

Motivation: Traffic accidents are a problem!

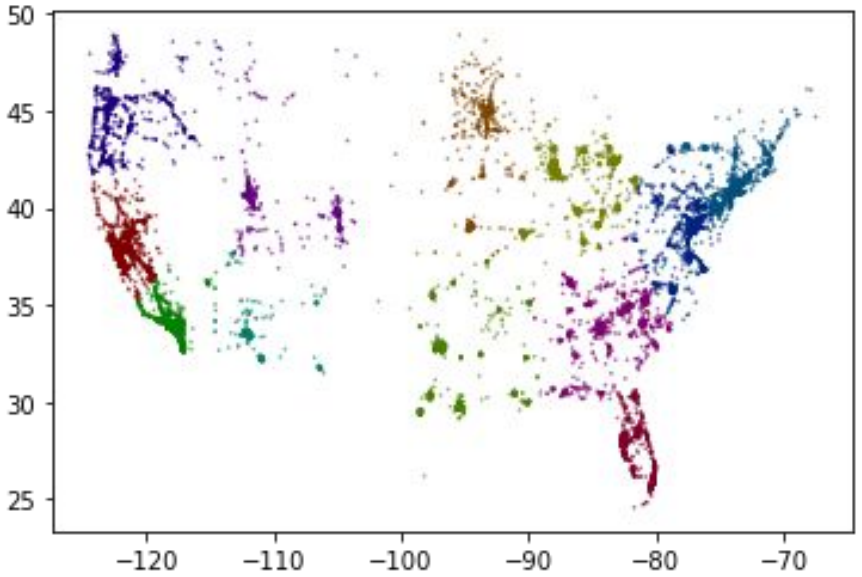
- We can't control **the risk** of driving.
- People feel comfortable with driving.[citation needed]
- *(Why can't we have public transport?)*
- How can we **help people understand stats?**

The answer: **Through an interactive data dashboard!**

Objective: **Create a tool** that lets people explore **where** and **when** traffic accidents happen in the continental United States.

Goals of the design:

- Sortable by **state**
- Sortable by **time**
- Users must be able to **compare** different accidents.
- **Natural-language** simple script language



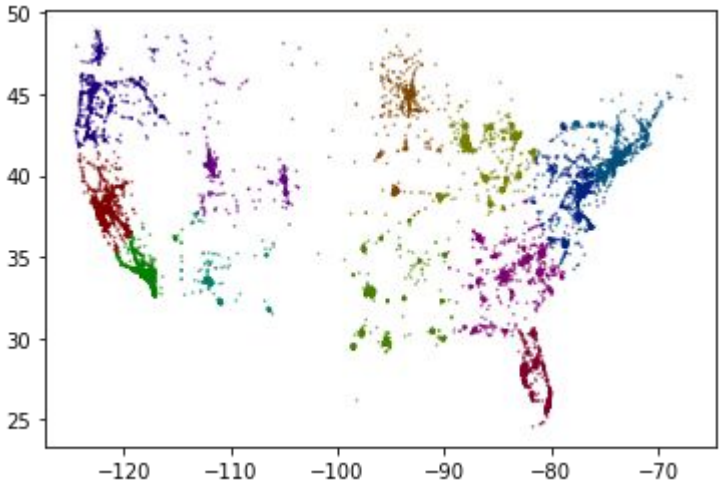
Goals of the design:

- Sortable by **state**
- Sortable by **time**
- Users must be able to **compare** different accidents.

Questions a user might ask:

1. How do December 25th accident rates compare to the average?
2. How does Friday December 25th compare to the average Friday?
3. Where did the most severe accidents in Connecticut happen during 2017?

example from K-means clustering the accident data



latitude/longitude on the axes

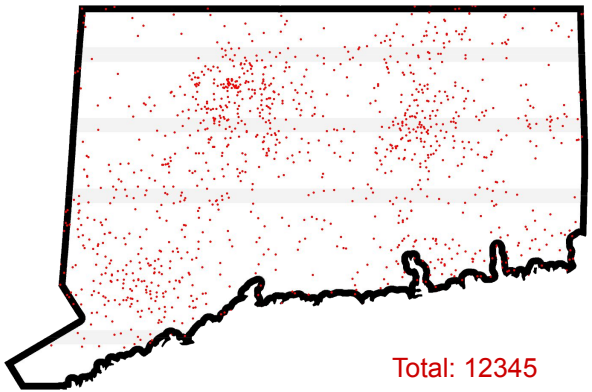
Ugly Mockup 1: Analyzing a specific state

State

⬇️ Connecticut

Severity

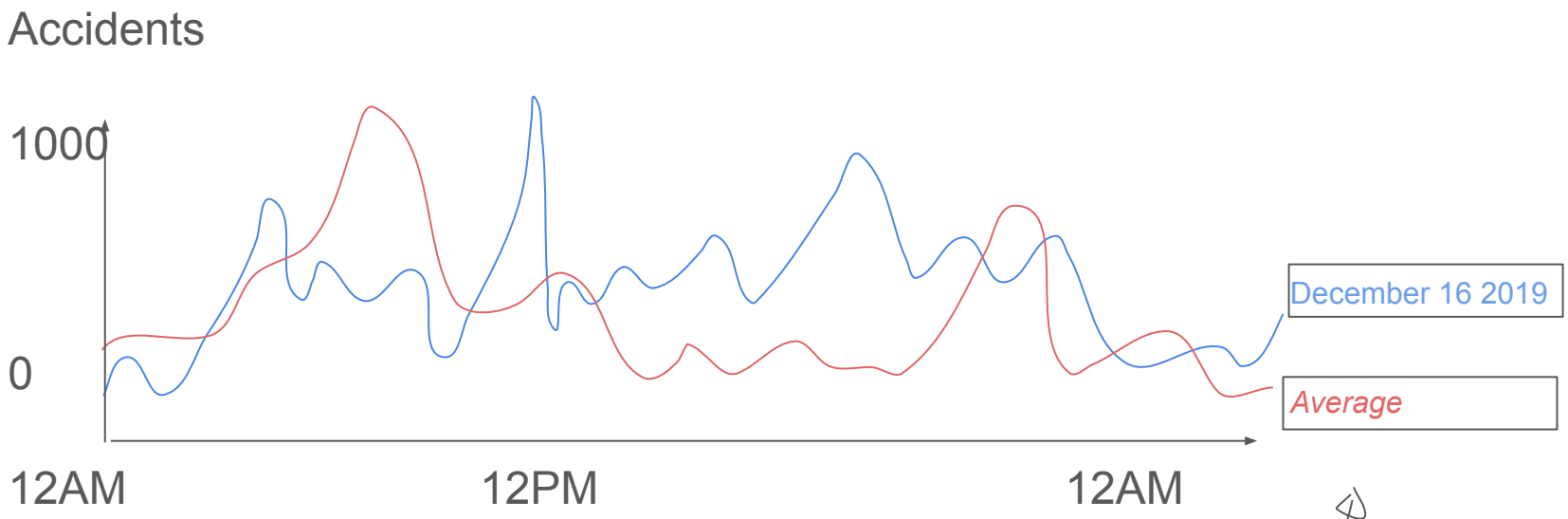
- ☑ 1
- ☑ 2
- ☐ 3
- ☐ 4



Time

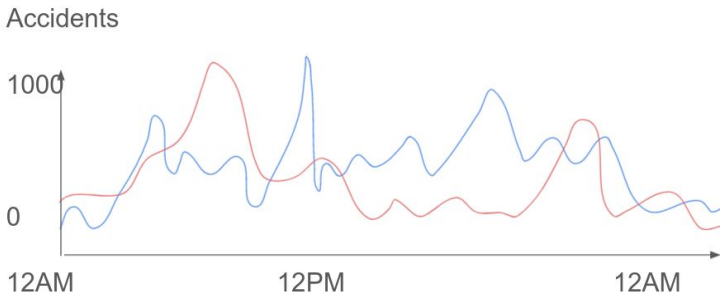
mar 2016 — nov 2017

Ugly Mockup 2: Analyzing across time



A simple script language will be used to specify timelines.

Examples of "natural language"



Goal:

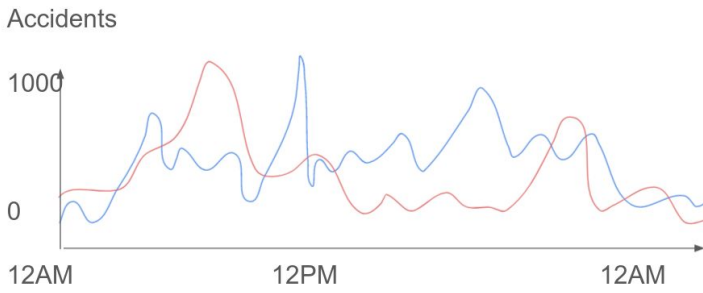
Users should be able to **use natural language to describe subsets of data.**

- "2016 December"
  - Compiles to "All states; From 2016-12-01 00:00 to 2016-12-31 11:59"
- "Tuesdays"
  - Compiles to "All states; From 2016-02-02 00:00 to 2016-02-02 11:59, and From 2016-02-09 00:00 to 2016-02-09 11:59, and …"
- "Saturdays and Sunday in California in December"

  etc...

"Oh no this sounds super hard!" Worry not, `set()` does most of the work

Natural language vs interface module?



One-box "natural language" process

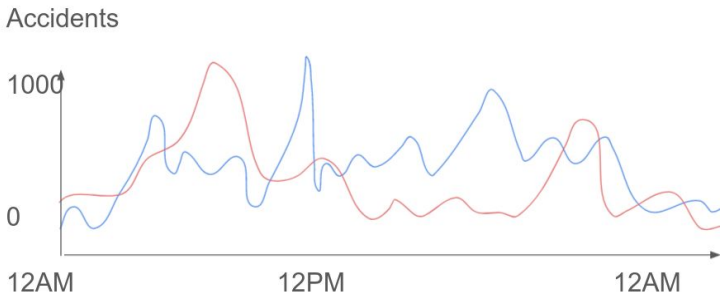
Fridays in Connecticut and Massachusetts from 2017 to 2018

versus

Many elements

Weekday	Year	Month	Day	States
Fri	2017, 2018	*all*	*all*	CT, MA

Example interpreter

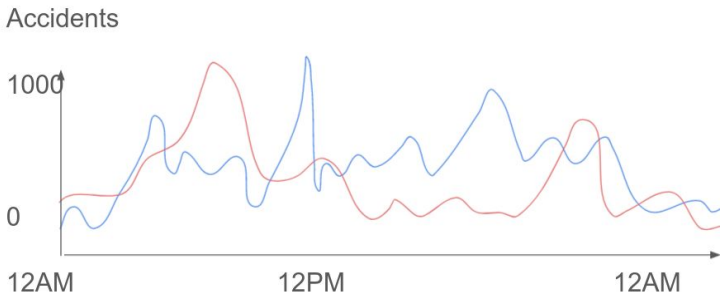


1. Raw string
   `` `Fridays in Connecticut and Massachusetts from 2017 to 2018` ``
2. Tokenize by space
   `['Fridays', 'in', 'Connecticut', 'and', 'Massachusetts', 'from', '2017', 'to', '2018']`
3. Ignore non-keywords
   (Keywords include numbers, full-names for months, weekdays, states, and abbreviations, as well as 'to' and 'and'.)
   `['Fridays', 'Connecticut', 'Massachusetts', '2017', 'to', '2018']`
4. Interpret joiner 'to' and set lowercase
   `['fridays', 'connecticut', 'massachusetts', '2017 to 2018']`

   ('to' can join weekdays, or up to three neighboring specifiers. Redundant weekdays are ignored when date is fully specified, e.g. as in Wednesday November 3rd 2021.)
5. Map to keys for index matching
   `['fri', 'ct, 'ma', '2017 to 2018']`
6. Use `set()` over strings and use that to select indices.

Limitations: No boolean operators, no fuzzy matching, etc.

Text processing for selecting a time



Users can:

- **Year, Month, Day, Hour.**
  Users can specify no year, one year, or a range of unit.
  This can be start, or end.
- **Weekday:**
  Users can specify none, or up to 7 weekdays.
- **State:**
  Users can specify none, or up to 50 states.

Dashboards!

Temporal data , **spatial** data

Questions?


My roommate's cat,
Catsidy