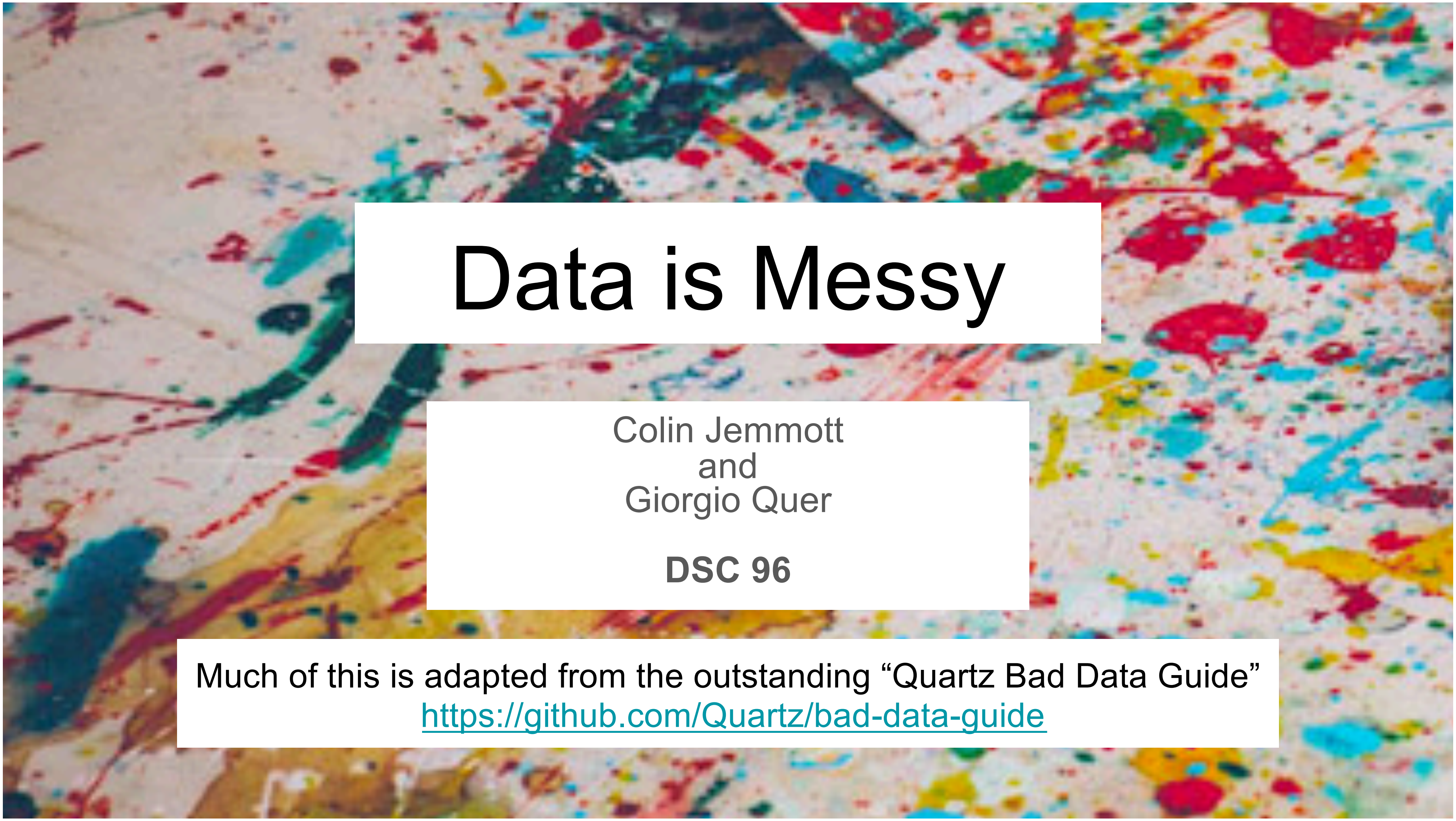
Announcements

- Office Hours:
  - Tuesdays, 5:00 PM - 6:00 PM, in HDSI Office E203 (SDSC).

- Readings (Week 2 due Friday 1/17)
  - Email: gquer@ucsd.edu
  - Subject line:
    - [DSC96 W20]: Week 02, Sec A/B, YourFirstName YourLastName
  - Email content:
    - Your comments/ questions/ observations on the proposed lectures

- Assignment 1:
  - SDPD workbook with dashboard
  - due: 1/21



Data is Messy

Colin Jemmott
and
Giorgio Quer

DSC 96

Much of this is adapted from the outstanding “Quartz Bad Data Guide”
https://github.com/Quartz/bad-data-guide

Data collection problems

- We have a great dataset:
  - Physical activity for 1 year from 10M people in US with an activity tracker!
  - We want to describe the physical activity of US citizens !
  - **Can we?**

Data collection problems

- We have a great dataset:
  - Physical activity for 1 year from 10M people in US who bought an activity tracker!
  - We want to describe the physical activity of US citizens !
  - Can we?
- Ok, let’s collect the data properly:
  - 1000 people randomly selected (any age or physical status or income) in San Diego county
  - 3 months of data (May, June, July)
  - Are we ok now?

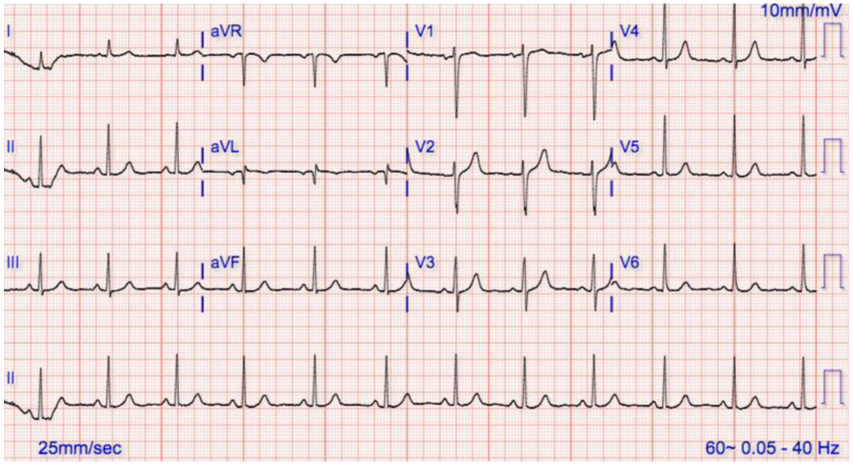
Data collection problems

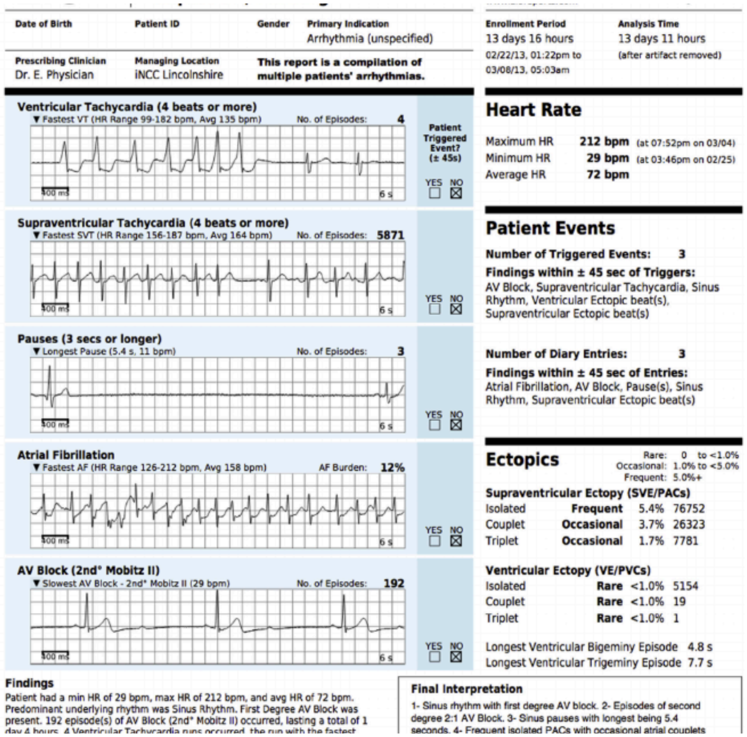
- Sample is not random
  - You have the number of steps, but the population is composed of very active people
- Seasonal variation
  - You have number of steps from a good population, but only in summer time
- Results are p-hacked
  - The data collection stopped once a significant result was observed

Other data types

Data doesn’t always come in in nicely formatted packages

- CSV, escaping, and the lack of standards
- Data are in a PDF - what now?
- Images and sound recordings as data





Date of Birth	**Patient ID**	**Gender**	**Primary Indication** Arrhythmia (unspecified)
Prescribing Clinician Dr. E. Physician	**Managing Location** iNCC Lincolnshire	**This report is a compilation of multiple patients' arrhythmias.**	

Enrollment Period
13 days 16 hours
02/22/13, 01:22pm to
03/08/13, 05:03am

Analysis Time
13 days 11 hours
(after artifact removed)

Ventricular Tachycardia (4 beats or more)
▼ Fastest VT (HR Range 99-182 bpm, Avg 135 bpm) No. of Episodes: **4**
400 ms 6 s
Patient Triggered Event? (± 45s)
YES □ NO ☒

Supraventricular Tachycardia (4 beats or more)
▼ Fastest SVT (HR Range 156-187 bpm, Avg 164 bpm) No. of Episodes: **5871**
400 ms 6 s
YES □ NO ☒

Pauses (3 secs or longer)
▼ Longest Pause (5.4 s, 11 bpm) No. of Episodes: **3**
400 ms 6 s
YES □ NO ☒

Atrial Fibrillation
▼ Fastest AF (HR Range 126-212 bpm, Avg 158 bpm) AF Burden: **12%**
400 ms 6 s
YES □ NO ☒

AV Block (2nd° Mobitz II)
▼ Slowest AV Block - 2nd° Mobitz II (29 bpm) No. of Episodes: **192**
400 ms 6 s
YES □ NO ☒

Heart Rate

Maximum HR	**212 bpm**	(at 07:52pm on 03/04)
Minimum HR	**29 bpm**	(at 03:46pm on 02/25)
Average HR	**72 bpm**	

Patient Events

Number of Triggered Events: 3

Findings within ± 45 sec of Triggers:
AV Block, Supraventricular Tachycardia, Sinus Rhythm, Ventricular Ectopic beat(s), Supraventricular Ectopic beat(s)

Number of Diary Entries: 3

Findings within ± 45 sec of Entries:
Atrial Fibrillation, AV Block, Pause(s), Sinus Rhythm, Supraventricular Ectopic beat(s)

Ectopics

Rare:	0 to <1.0%
Occasional:	1.0% to <5.0%
Frequent:	5.0%+

Supraventricular Ectopy (SVE/PACs)

Isolated	**Frequent**	5.4%	76752
Couplet	**Occasional**	3.7%	26323
Triplet	**Occasional**	1.7%	7781

Ventricular Ectopy (VE/PVCs)

Isolated	**Rare**	<1.0%	5154
Couplet	**Rare**	<1.0%	19
Triplet	**Rare**	<1.0%	1

Longest Ventricular Bigeminy Episode	4.8 s
Longest Ventricular Trigeminy Episode	7.7 s

Findings
Patient had a min HR of 29 bpm, max HR of 212 bpm, and avg HR of 72 bpm. Predominant underlying rhythm was Sinus Rhythm. First Degree AV Block was present. 192 episode(s) of AV Block (2nd° Mobitz II) occurred, lasting a total of 1 day 4 hours. 4 Ventricular Tachycardia runs occurred, the run with the fastest

Final Interpretation
1- Sinus rhythm with first degree AV block. 2- Episodes of second degree 2:1 AV Block. 3- Sinus pauses with longest being 5.4 seconds. 4- Frequent isolated PACs with occasional atrial couplets

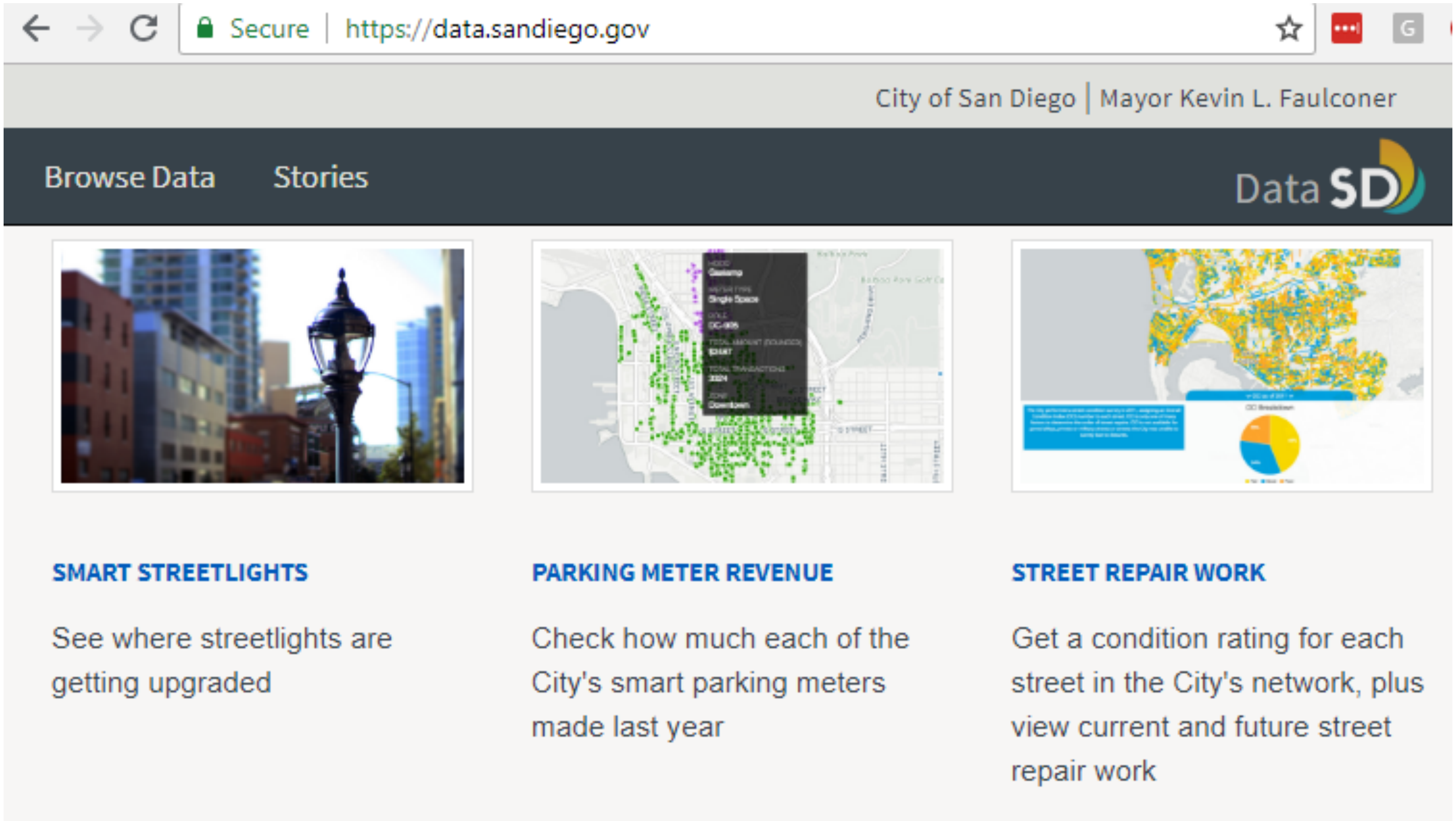
from: Barrett et al, "Comparison of 24-hour Holter Monitoring with 14-dayNovel Adhesive Patch Electrocardiographic Monitoring"

Vehicle Stop Data

DSC 96

Data Source

https://data.sandiego.gov/datasets/police-vehicle-stops/


Secure
https://data.sandiego.gov
City of San Diego | Mayor Kevin L. Faulconer
Browse Data
Stories
Data SD

SMART STREETLIGHTS

See where streetlights are getting upgraded

PARKING METER REVENUE

Check how much each of the City's smart parking meters made last year

STREET REPAIR WORK

Get a condition rating for each street in the City's network, plus view current and future street repair work

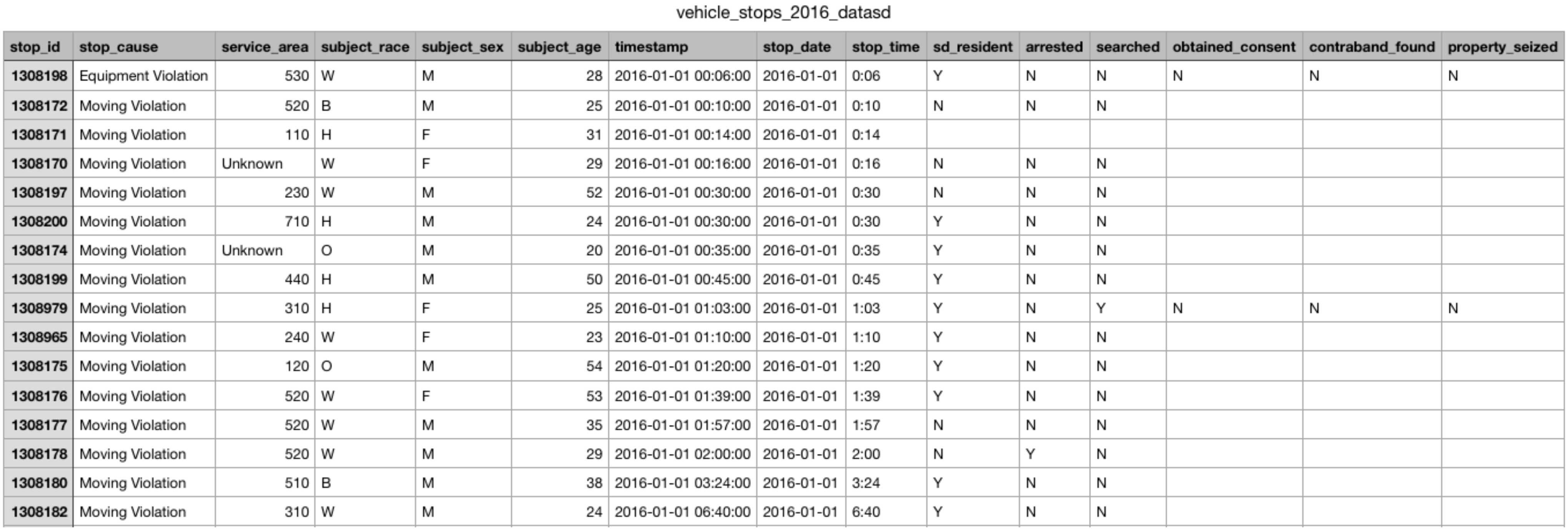
Why Police Data?

- Where is it?
  - https://github.com/gquer/dsc96_W20/blob/master/Projects/02%20SDPD/vehicle_stops_2016_datasd.csv

- Where do we start?
  - https://github.com/gquer/dsc96_W20/blob/master/Projects/02%20SDPD/vehicle_stops_2016_datasd_example.png
  - https://github.com/gquer/dsc96_W20/blob/master/Projects/02%20SDPD/README.md

Police Vehicle Stops

Vehicle stops made by the San Diego Police Department. Vehicle Stops files contain all vehicle stops for a given year.



vehicle_stops_2016_datasd

stop_id	stop_cause	service_area	subject_race	subject_sex	subject_age	timestamp	stop_date	stop_time	sd_resident	arrested	searched	obtained_consent	contraband_found	property_seized
1308198	Equipment Violation	530	W	M	28	2016-01-01 00:06:00	2016-01-01	0:06	Y	N	N	N	N	N
1308172	Moving Violation	520	B	M	25	2016-01-01 00:10:00	2016-01-01	0:10	N	N	N			
1308171	Moving Violation	110	H	F	31	2016-01-01 00:14:00	2016-01-01	0:14						
1308170	Moving Violation	Unknown	W	F	29	2016-01-01 00:16:00	2016-01-01	0:16	N	N	N			
1308197	Moving Violation	230	W	M	52	2016-01-01 00:30:00	2016-01-01	0:30	N	N	N			
1308200	Moving Violation	710	H	M	24	2016-01-01 00:30:00	2016-01-01	0:30	Y	N	N			
1308174	Moving Violation	Unknown	O	M	20	2016-01-01 00:35:00	2016-01-01	0:35	Y	N	N			
1308199	Moving Violation	440	H	M	50	2016-01-01 00:45:00	2016-01-01	0:45	Y	N	N			
1308979	Moving Violation	310	H	F	25	2016-01-01 01:03:00	2016-01-01	1:03	Y	N	Y	N	N	N
1308965	Moving Violation	240	W	F	23	2016-01-01 01:10:00	2016-01-01	1:10	Y	N	N			
1308175	Moving Violation	120	O	M	54	2016-01-01 01:20:00	2016-01-01	1:20	Y	N	N			
1308176	Moving Violation	520	W	F	53	2016-01-01 01:39:00	2016-01-01	1:39	Y	N	N			
1308177	Moving Violation	520	W	M	35	2016-01-01 01:57:00	2016-01-01	1:57	N	N	N			
1308178	Moving Violation	520	W	M	29	2016-01-01 02:00:00	2016-01-01	2:00	N	Y	N			
1308180	Moving Violation	510	B	M	38	2016-01-01 03:24:00	2016-01-01	3:24	Y	N	N			
1308182	Moving Violation	310	W	M	24	2016-01-01 06:40:00	2016-01-01	6:40	Y	N	N			

SDPD Vehicle Stop Data

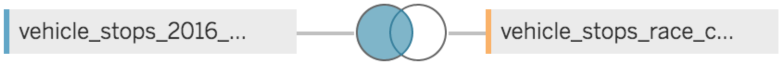
- Create SDPD workbook
- Assignment 1:

  Your dashboard should be completed by 1/21

- Number of stops per age group?
  - Attention: what is Measures and what is Dimensions here?
  - Notice any issues? What should we do (use Groups?)
  - How are they distributed?
  - Divide by sex and age
- Time series plots!
  - E.g., stops by hour of day, day of week, month, might be interesting.
  - What happens if we plot stops for any minute of the day? are there any abnormality low/high to discuss?
- Explore the “stop cause” variable. Notice any issues? What should we do?
- Race bias: can you plot number of stop vs race?
- What is the fraction of searched

Join two datasets

- Ethnicity:
    - which races do you see? Can you rename them? -> join two datasets
    - which are more represented? should we group them?
- Click on: Data –> vehicle_stops_2016_datasd
- Drop vehicle_stops_race_codes
- Different ways to join:
    - Inner Join. The records where the IDs match in both data sets.
    - Left Join. You get all the records from the data on the left side of your equation and any time the IDs match, you also get the records from the right side of the equation.
    - Right Join. You get all the records from the data on the right side of your equation and any time the IDs match, you also get the records from the right side of the equation.
    - Outer Join. You add all the records from each data set together, even when there is no join. (all information)
- Attention: you should specify which fields to join



Calculated field

- Searched: data is Y N n Null
- We would like to have a binary outcome (1 if searched, 0 otherwise)
- Change format:
  - Create -> Calculated Field:

    ```
    IF [Searched]= 'Y' THEN 1 ELSE 0 END
    ```
- Move to Measures
  - Now we can calculate the average !!!
- Question:

  What is the probability of being searched, when stopped, for:
  - Male vs female
  - Race: White vs Hispanic vs African American vs Asian vs Other

    (how do we define races in this way?)