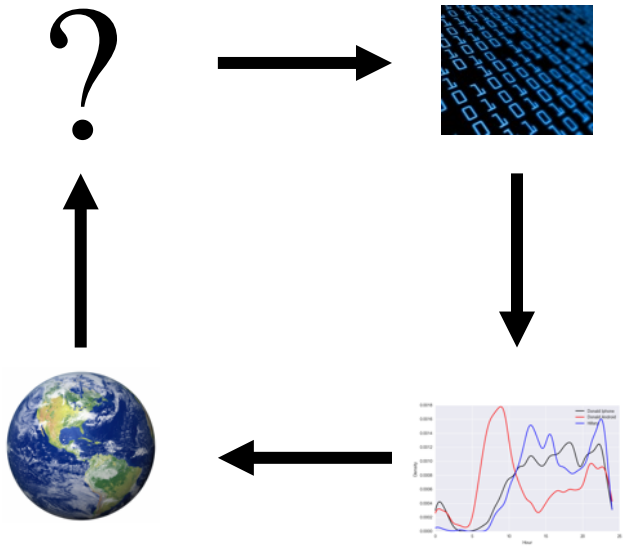
# Data Science 100

## *Lecture 5: Exploratory Data Analysis*

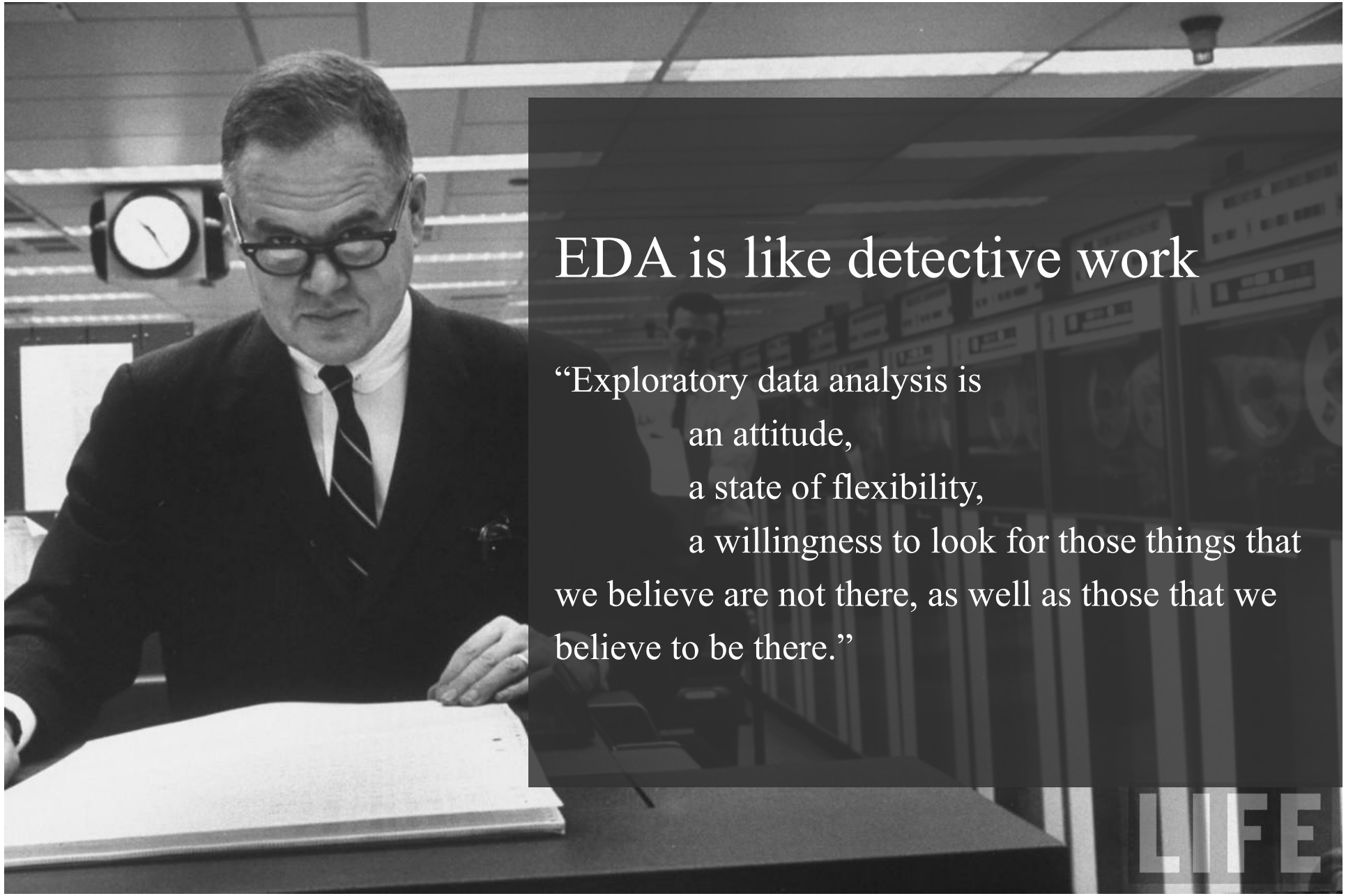
Slides by:

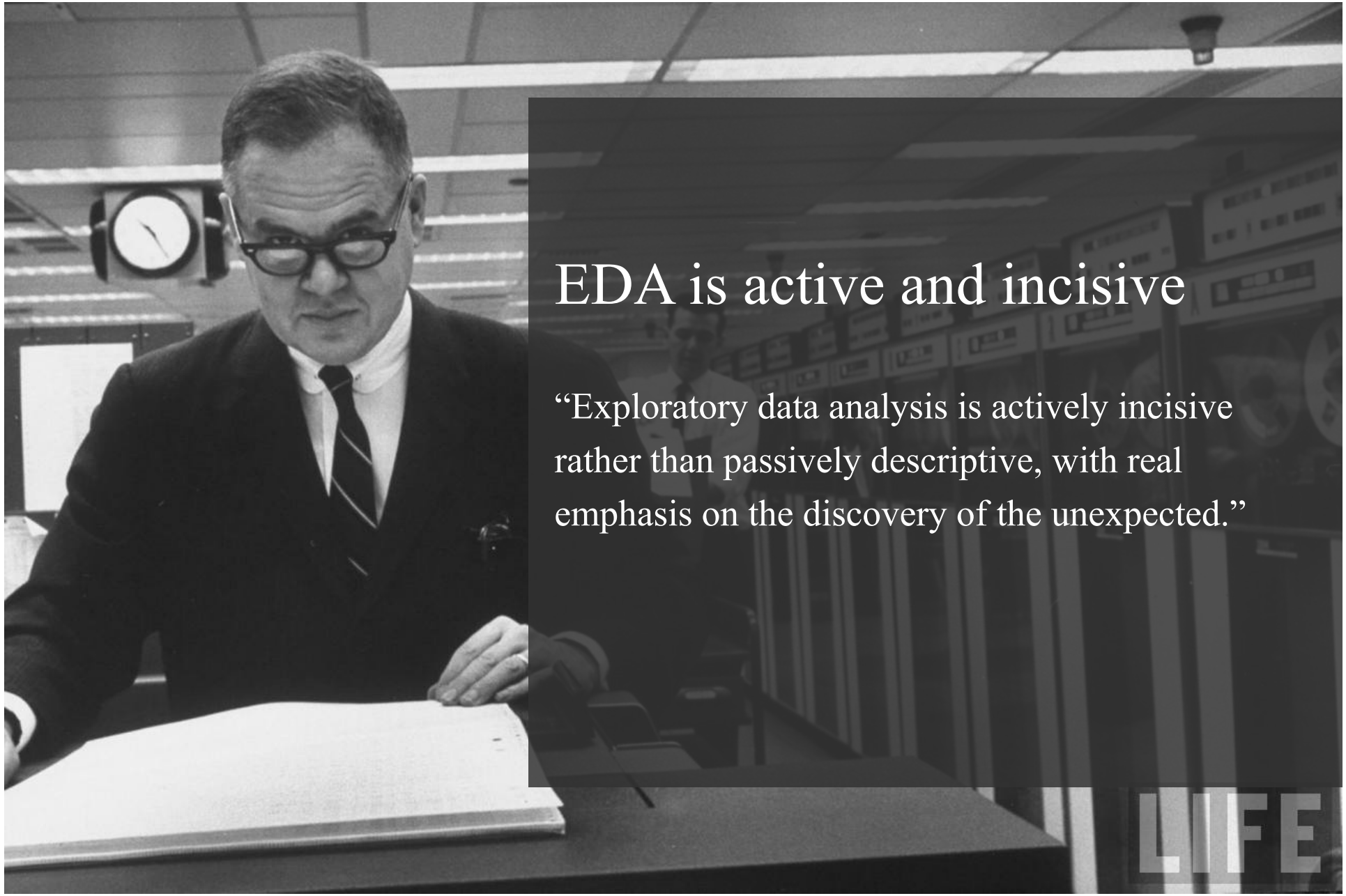
**Deb Nolan**

deborah_nolan@berkeley.edu




Data Analysis & Statistics, Tukey 1965


EDA is like detective work
“Exploratory data analysis is
an attitude,
a state of flexibility,
a willingness to look for those things that we believe are not there, as well as those that we believe to be there.”
LIFE


EDA is active and incisive
“Exploratory data analysis is actively incisive rather than passively descriptive, with real emphasis on the discovery of the unexpected.”
LIFE

# Philosophy of EDA

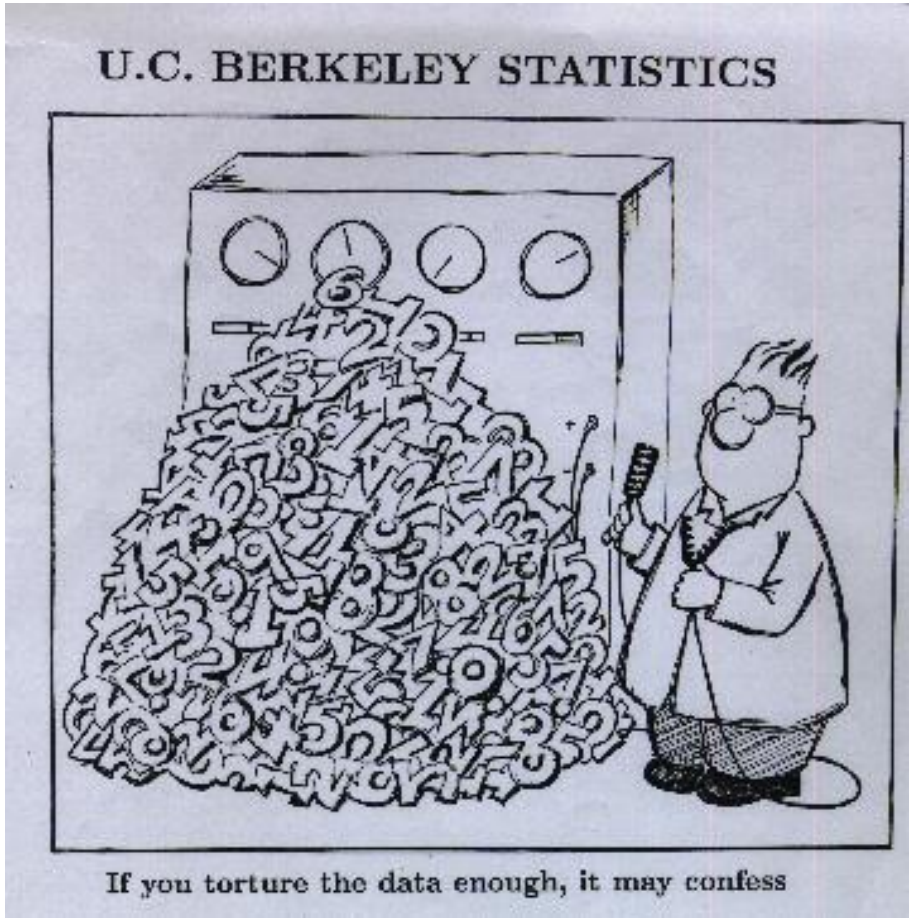
- Confirm understanding of the data
- Keep an open mind and be willing to find something surprising
- Iterate
  - Uncover new aspects of our data
  - Re-examine our understanding of the data
  - Continue exploration

# Where does EDA help?

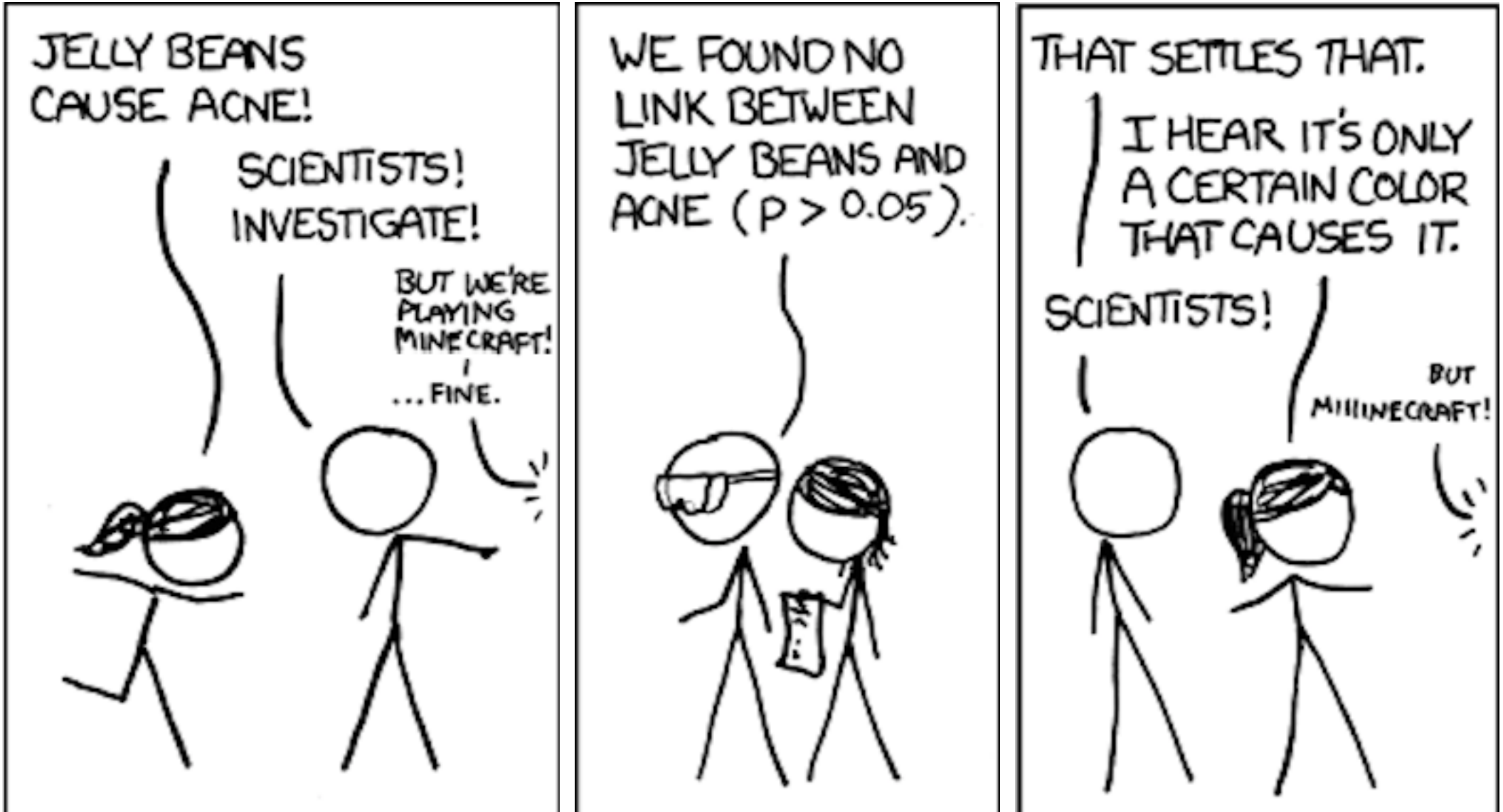
- Clean data
- Transform variables and derive new variables – put data in format suitable for analysis
- Better understand data/situation (no formal analysis pursued)
- Inform formal analysis – uncover important features that impact the analysis
- Examine formal analysis – explore output from an analysis, e.g., predictions, parameter estimates, model fit

# Caution about EDA

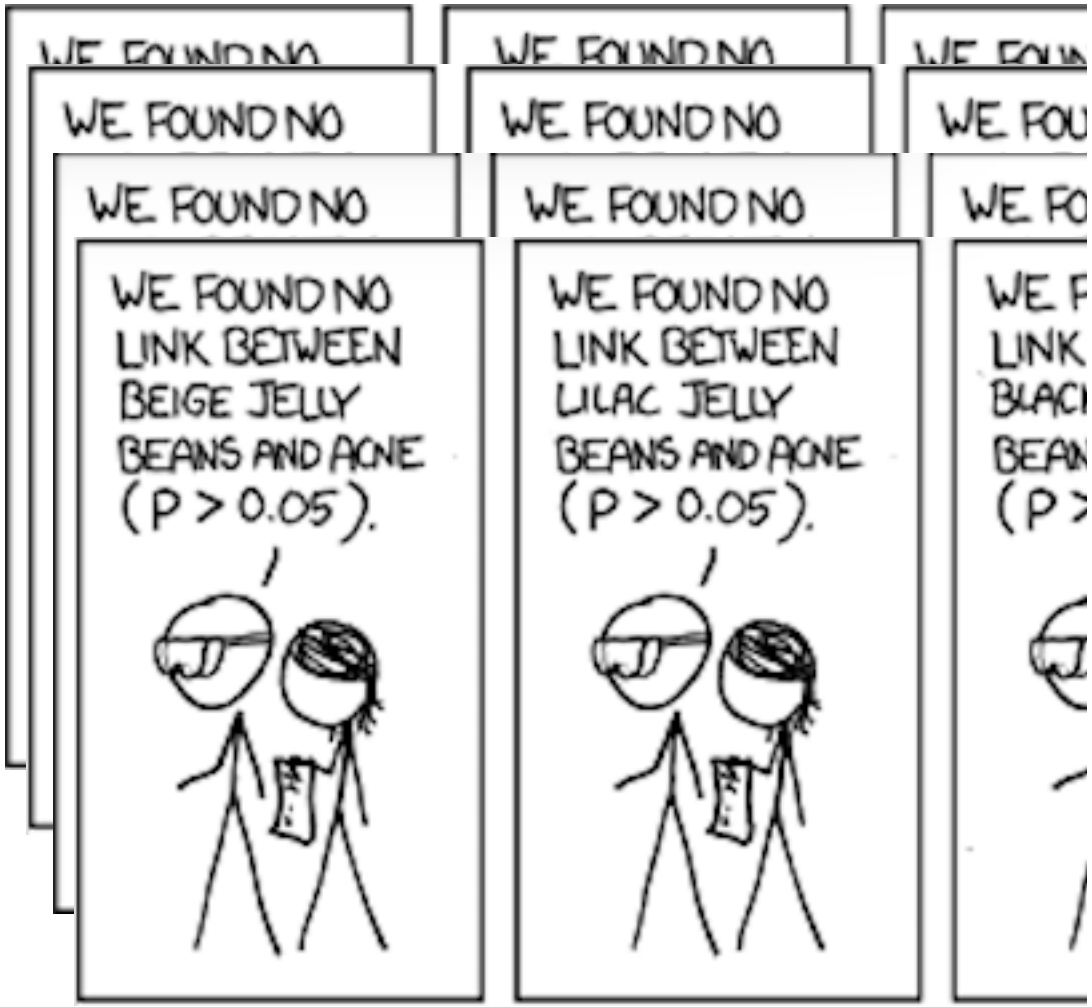
- Considered data snooping
- With enough data, if you look hard enough you will find something “interesting”
- Related to p-hacking

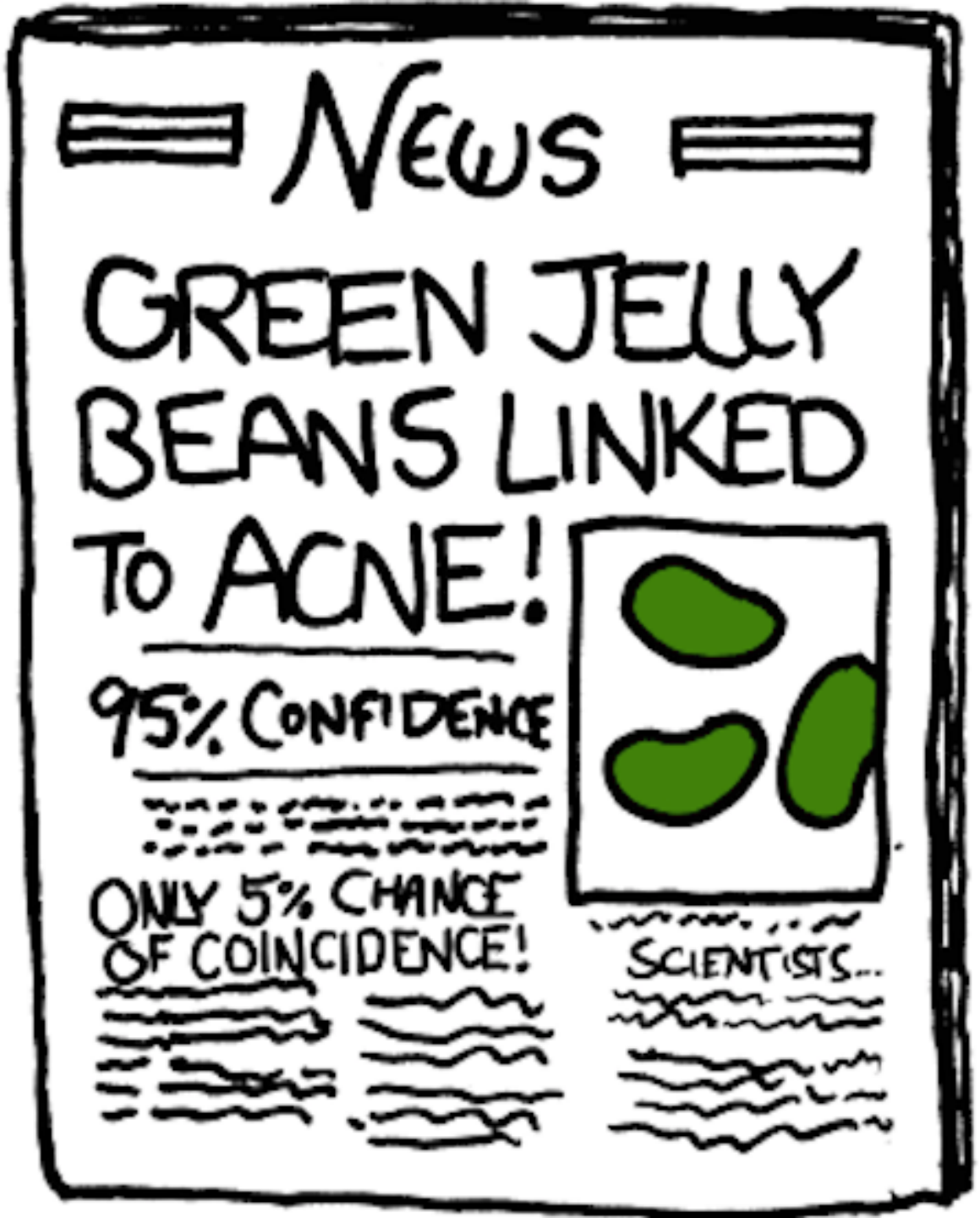


# Caution about EDA – ala xkcd



# Caution about EDA -



# On The Other Hand

EDA can provide valuable insights about a model and its assumptions

# What to do?

- Additional Data
  - Pilot study
  - Perform a second experiment
  - Collect more data
  - Short of that – divide data in two at random and explore half
- Reporting
  - Describe EDA that was performed
  - Understand and describe the limitations of your analysi
- Differentiate between descriptive analysis and formal inference

# How to Carry out EDA?

- Plot data in multiple ways to get different insights
- Transform variables to symmetrize distributions
- Transform to straighten relationships
- Derive new variables
- Consider effect of other variables on distributions & relationships

**Connect** what you find to the question and context

# A Case Study

# Question


Taxi driver claim:

When traffic on a freeway begins to break down, the fasts lane breaks down first so you should move immediately to the right.

Can we confirm this phenomena?

Can we turn this claim into a statistics question that is answerable with data?

# Question

When traffic on a freeway begins to break down, the fast lane breaks down first so they move immediately to the right.

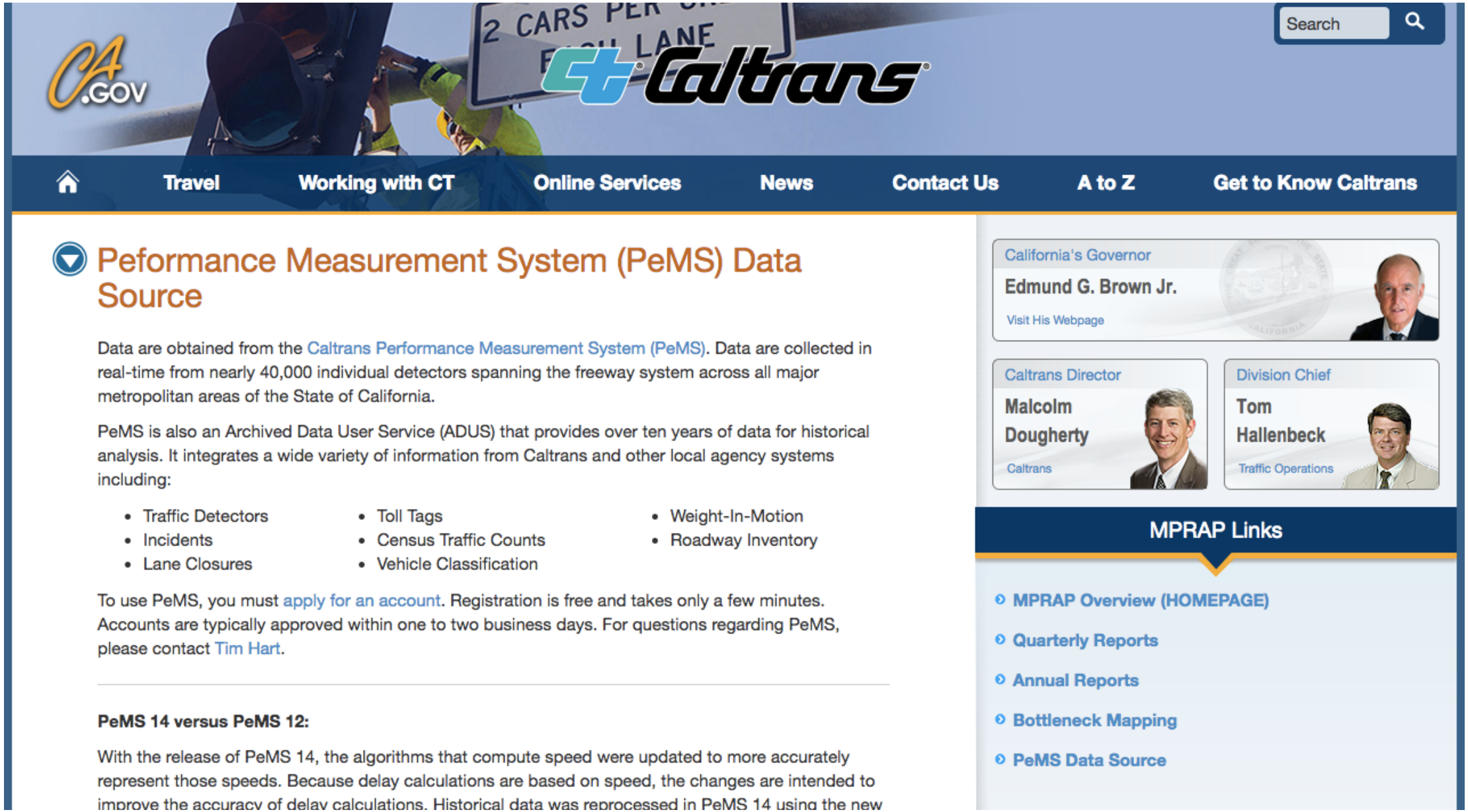
What information do we need to study this question? –

- Observe traffic on a freeway
- Include time to see when traffic is heavy and break downs
- Record traffic in lanes to differentiate traffic levels between in lanes

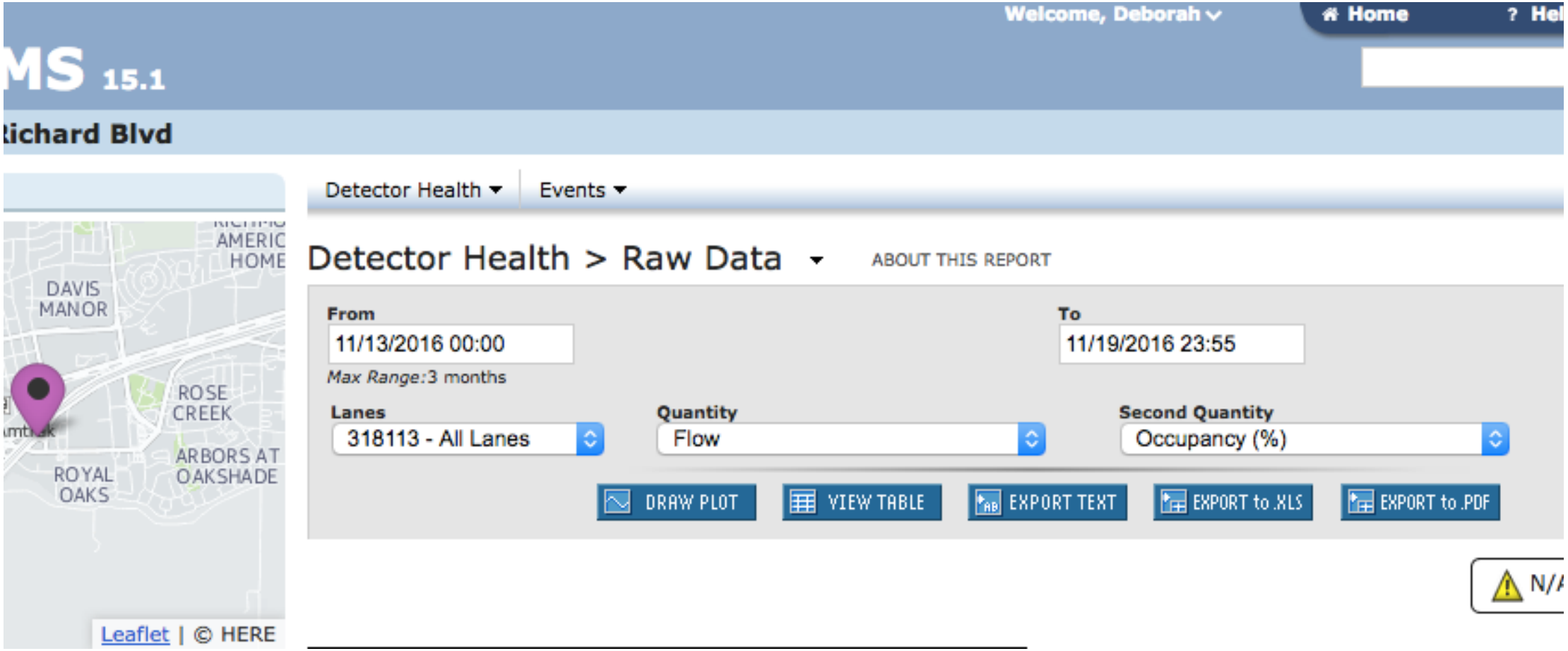
# Simpler Related Questions

- Do the 3 lanes serve the same amount of traffic?
- When 1 lane is congested, are the other lanes also congested?
- What's the relationship between amount of traffic served and congestion?
- How are answers to these questions impacted by time of day and day of week?

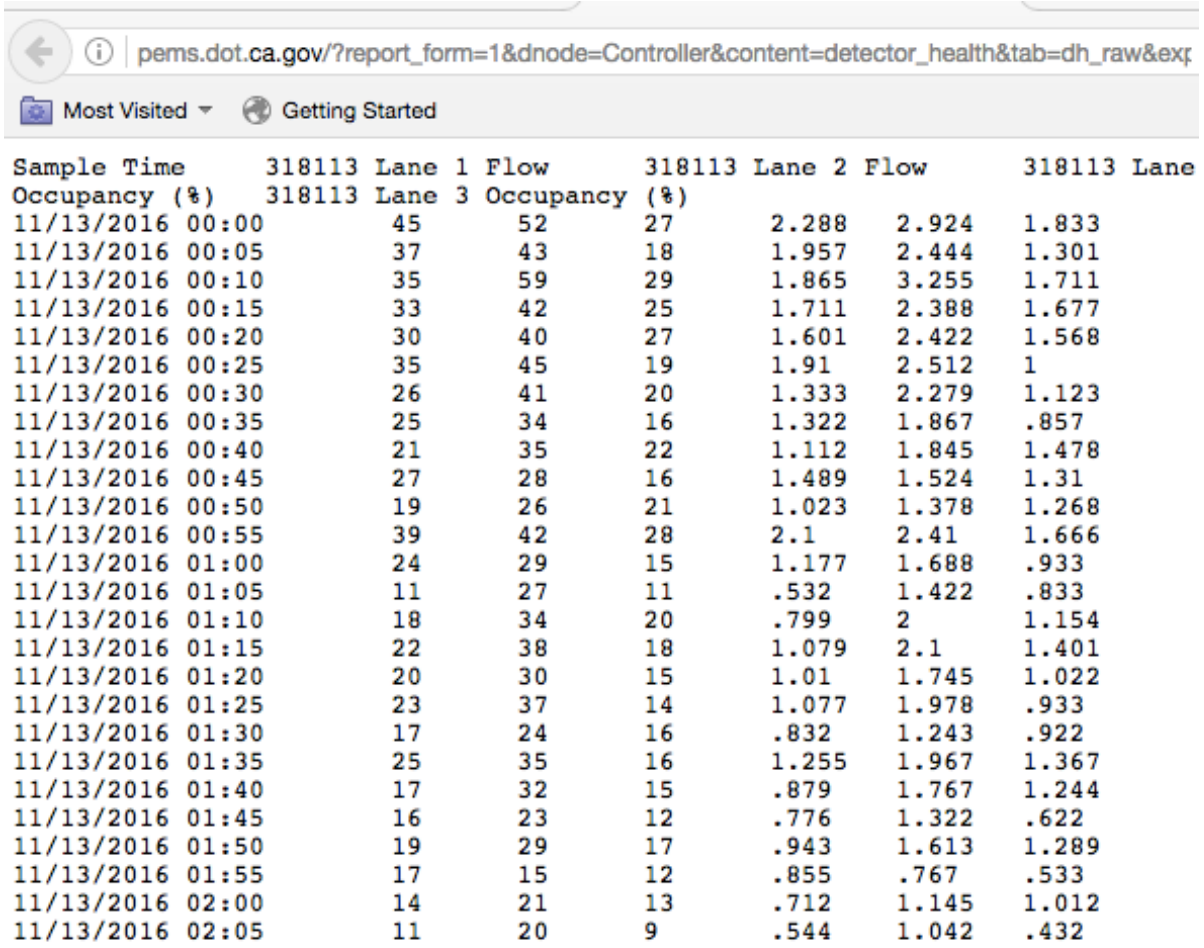
# Let’s Get Some Data


CA.GOV
Caltrans
Search
Travel
Working with CT
Online Services
News
Contact Us
A to Z
Get to Know Caltrans
Peformance Measurement System (PeMS) Data Source
Data are obtained from the Caltrans Performance Measurement System (PeMS). Data are collected in real-time from nearly 40,000 individual detectors spanning the freeway system across all major metropolitan areas of the State of California.
PeMS is also an Archived Data User Service (ADUS) that provides over ten years of data for historical analysis. It integrates a wide variety of information from Caltrans and other local agency systems including:
Traffic Detectors
Incidents
Lane Closures
Toll Tags
Census Traffic Counts
Vehicle Classification
Weight-In-Motion
Roadway Inventory
To use PeMS, you must apply for an account. Registration is free and takes only a few minutes. Accounts are typically approved within one to two business days. For questions regarding PeMS, please contact Tim Hart.
PeMS 14 versus PeMS 12:
With the release of PeMS 14, the algorithms that compute speed were updated to more accurately represent those speeds. Because delay calculations are based on speed, the changes are intended to
California's Governor
Edmund G. Brown Jr.
Visit His Webpage
Caltrans Director
Malcolm Dougherty
Caltrans
Division Chief
Tom Hallenbeck
Traffic Operations
MPRAP Links
MPRAP Overview (HOMEPAGE)
Quarterly Reports
Annual Reports
Bottleneck Mapping
PeMS Data Source

# PEMS Data – Available in online form



# PEMS Data – Available in online form




| Sample Time | 318113 Lane 1 Flow | | 318113 Lane 2 Flow | | 318113 Lane |
|---|---|---|---|---|---|
| Occupancy (%) | 318113 Lane 3 Occupancy (%) | | | | |
| 11/13/2016 00:00 | 45 | 52 | 27 | 2.288 | 2.924 | 1.833 |
| 11/13/2016 00:05 | 37 | 43 | 18 | 1.957 | 2.444 | 1.301 |
| 11/13/2016 00:10 | 35 | 59 | 29 | 1.865 | 3.255 | 1.711 |
| 11/13/2016 00:15 | 33 | 42 | 25 | 1.711 | 2.388 | 1.677 |
| 11/13/2016 00:20 | 30 | 40 | 27 | 1.601 | 2.422 | 1.568 |
| 11/13/2016 00:25 | 35 | 45 | 19 | 1.91 | 2.512 | 1 |
| 11/13/2016 00:30 | 26 | 41 | 20 | 1.333 | 2.279 | 1.123 |
| 11/13/2016 00:35 | 25 | 34 | 16 | 1.322 | 1.867 | .857 |
| 11/13/2016 00:40 | 21 | 35 | 22 | 1.112 | 1.845 | 1.478 |
| 11/13/2016 00:45 | 27 | 28 | 16 | 1.489 | 1.524 | 1.31 |
| 11/13/2016 00:50 | 19 | 26 | 21 | 1.023 | 1.378 | 1.268 |
| 11/13/2016 00:55 | 39 | 42 | 28 | 2.1 | 2.41 | 1.666 |
| 11/13/2016 01:00 | 24 | 29 | 15 | 1.177 | 1.688 | .933 |
| 11/13/2016 01:05 | 11 | 27 | 11 | .532 | 1.422 | .833 |
| 11/13/2016 01:10 | 18 | 34 | 20 | .799 | 2 | 1.154 |
| 11/13/2016 01:15 | 22 | 38 | 18 | 1.079 | 2.1 | 1.401 |
| 11/13/2016 01:20 | 20 | 30 | 15 | 1.01 | 1.745 | 1.022 |
| 11/13/2016 01:25 | 23 | 37 | 14 | 1.077 | 1.978 | .933 |
| 11/13/2016 01:30 | 17 | 24 | 16 | .832 | 1.243 | .922 |
| 11/13/2016 01:35 | 25 | 35 | 16 | 1.255 | 1.967 | 1.367 |
| 11/13/2016 01:40 | 17 | 32 | 15 | .879 | 1.767 | 1.244 |
| 11/13/2016 01:45 | 16 | 23 | 12 | .776 | 1.322 | .622 |
| 11/13/2016 01:50 | 19 | 29 | 17 | .943 | 1.613 | 1.289 |
| 11/13/2016 01:55 | 17 | 15 | 12 | .855 | .767 | .533 |
| 11/13/2016 02:00 | 14 | 21 | 13 | .712 | 1.145 | 1.012 |
| 11/13/2016 02:05 | 11 | 20 | 9 | .544 | 1.042 | .432 |

Loop Detector Data:

Number of vehicles that pass over a loop detector in a 5-minute period

Percentage of time in the 5-minute period that a vehicle was over the loop detector

# Statistical Issues:
# What is the context Question?
# What is the Population?
# Data Represent Population?

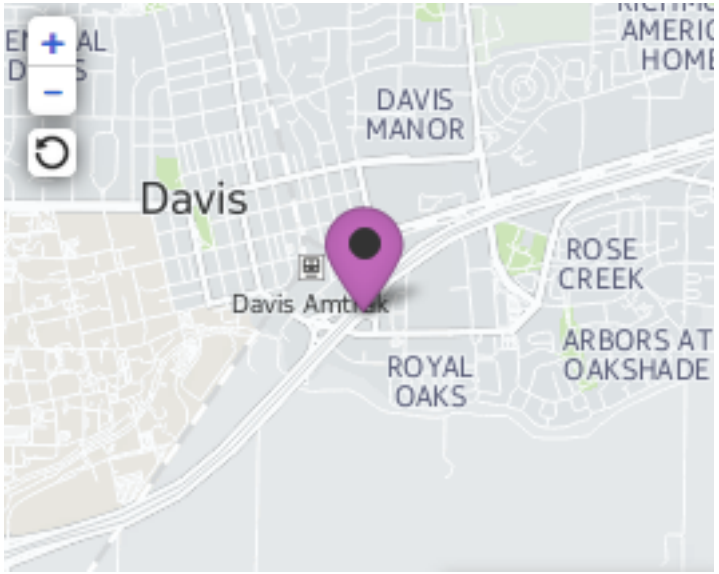
# Population

5-minute intervals at any time and place on a freeway

Data:

Time: Nov 13 to Nov 19, 2016 (one week)

Place: Loop 318113 on I 80 Eastbound



# Representative

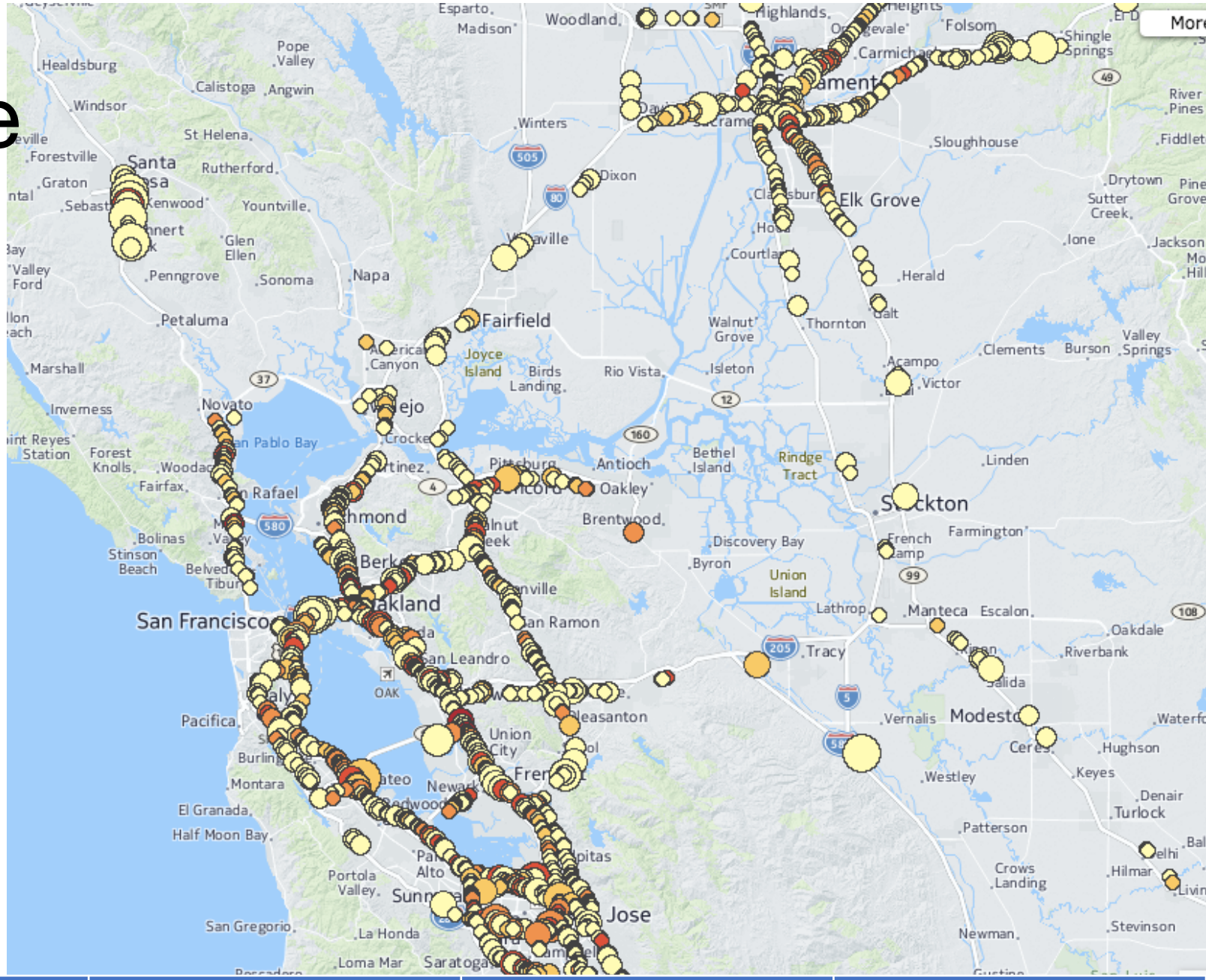
Traffic may behave differently at:

Other locations

Other times of year

Such as?

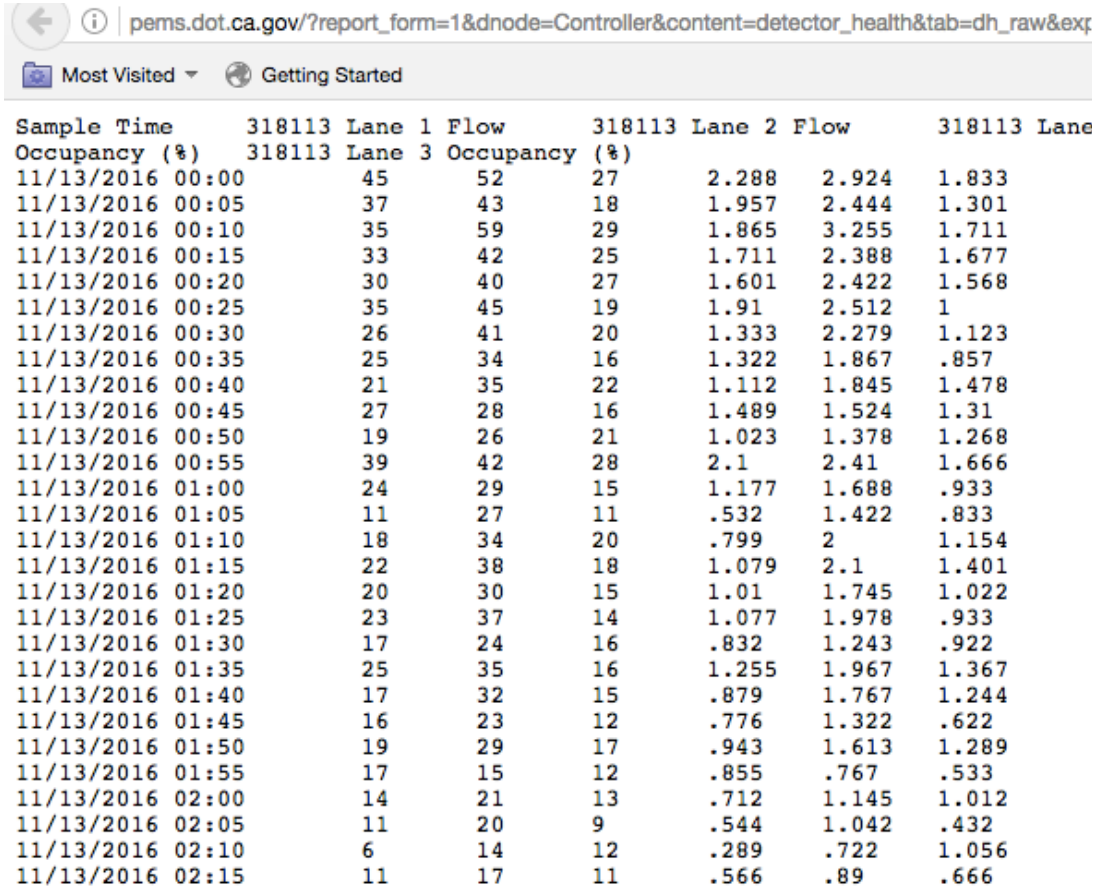
Keep this in mind as we analyze the data



# Data Wrangling Issues:
Structure
Granularity
Faithfulness
Time
Scope

# Structure

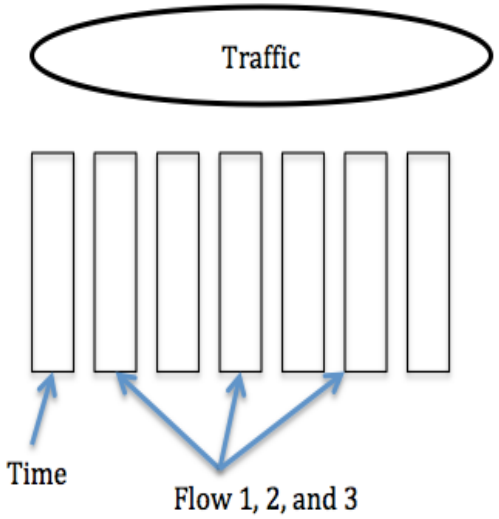
- Rectangular?
- Record delimiter?
- Variable value delimiter?
- Primitive Data Types?

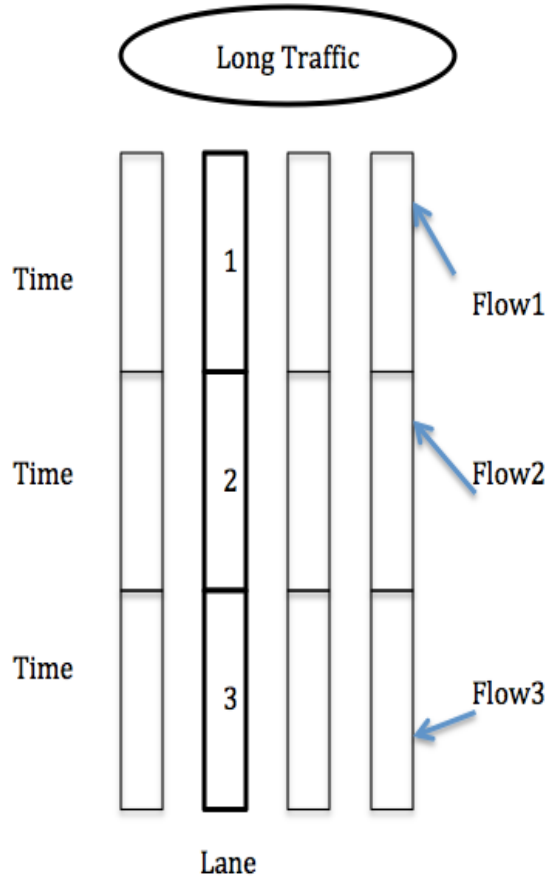



| Sample Time | 318113 Lane 1 Flow | | 318113 Lane 2 Flow | | 318113 Lane |
|---|---|---|---|---|---|
| Occupancy (%) | 318113 Lane 3 Occupancy (%) | | | | |
| 11/13/2016 00:00 | 45 | 52 | 27 | 2.288 | 2.924 | 1.833 |
| 11/13/2016 00:05 | 37 | 43 | 18 | 1.957 | 2.444 | 1.301 |
| 11/13/2016 00:10 | 35 | 59 | 29 | 1.865 | 3.255 | 1.711 |
| 11/13/2016 00:15 | 33 | 42 | 25 | 1.711 | 2.388 | 1.677 |
| 11/13/2016 00:20 | 30 | 40 | 27 | 1.601 | 2.422 | 1.568 |
| 11/13/2016 00:25 | 35 | 45 | 19 | 1.91 | 2.512 | 1 |
| 11/13/2016 00:30 | 26 | 41 | 20 | 1.333 | 2.279 | 1.123 |
| 11/13/2016 00:35 | 25 | 34 | 16 | 1.322 | 1.867 | .857 |
| 11/13/2016 00:40 | 21 | 35 | 22 | 1.112 | 1.845 | 1.478 |
| 11/13/2016 00:45 | 27 | 28 | 16 | 1.489 | 1.524 | 1.31 |
| 11/13/2016 00:50 | 19 | 26 | 21 | 1.023 | 1.378 | 1.268 |
| 11/13/2016 00:55 | 39 | 42 | 28 | 2.1 | 2.41 | 1.666 |
| 11/13/2016 01:00 | 24 | 29 | 15 | 1.177 | 1.688 | .933 |
| 11/13/2016 01:05 | 11 | 27 | 11 | .532 | 1.422 | .833 |
| 11/13/2016 01:10 | 18 | 34 | 20 | .799 | 2 | 1.154 |
| 11/13/2016 01:15 | 22 | 38 | 18 | 1.079 | 2.1 | 1.401 |
| 11/13/2016 01:20 | 20 | 30 | 15 | 1.01 | 1.745 | 1.022 |
| 11/13/2016 01:25 | 23 | 37 | 14 | 1.077 | 1.978 | .933 |
| 11/13/2016 01:30 | 17 | 24 | 16 | .832 | 1.243 | .922 |
| 11/13/2016 01:35 | 25 | 35 | 16 | 1.255 | 1.967 | 1.367 |
| 11/13/2016 01:40 | 17 | 32 | 15 | .879 | 1.767 | 1.244 |
| 11/13/2016 01:45 | 16 | 23 | 12 | .776 | 1.322 | .622 |
| 11/13/2016 01:50 | 19 | 29 | 17 | .943 | 1.613 | 1.289 |
| 11/13/2016 01:55 | 17 | 15 | 12 | .855 | .767 | .533 |
| 11/13/2016 02:00 | 14 | 21 | 13 | .712 | 1.145 | 1.012 |
| 11/13/2016 02:05 | 11 | 20 | 9 | .544 | 1.042 | .432 |
| 11/13/2016 02:10 | 6 | 14 | 12 | .289 | .722 | 1.056 |
| 11/13/2016 02:15 | 11 | 17 | 11 | .566 | .89 | .666 |

# Granularity

- Key – date/time (5 minute interval
- Alternative Key – more useful in our analysis:
  - Date + Lane
  - Need to stack our data table
  - Need to create a new variable



# Faithfulness

- Census of all 5-minute intervals in the time period for all 3 lanes
- Check detector health to see if all working in that time period, missing and faulty values are imputed
- Dependency –
  - percentages between 0 & 100
  - Number of records 12 (intervals/hr) * 24 (hr/day) * 7 day/wk = 2016

# Temporality & Scope

- Temporality - One week in November 2016
- Scope – Recordings for some (which?) 5-minute intervals in the time period may be imputed from historical data

# Ready for 2nd EDA

Do the 3 lanes serve the same amount of traffic?

When 1 lane is congested, are the other lanes also congested?

What's the relationship between amount of traffic served and congestion?

How are answers to these questions impacted by time of day and day of week?

# Distribution of Variable Values

Always a good place to start

Get your head in the game

# Distribution of Values

- Rug plots show exact location of values

- This tends not to be useful when we have many values
- Typically, we want a summary of this distribution

# Histograms & Density Curves

- Histogram bar:

  Height * Width = Area = Proportion (or count)

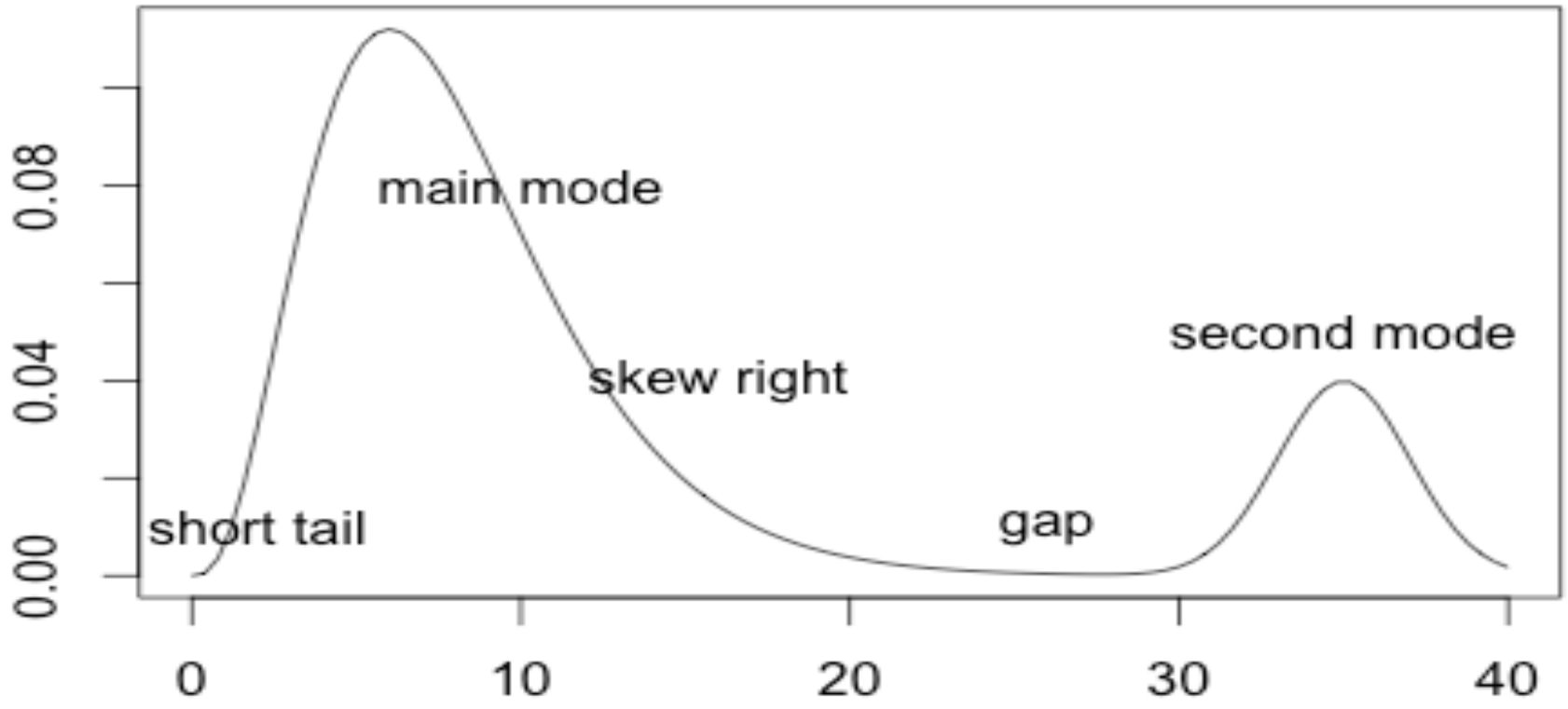
- Similar property for area beneath a density curve

- No longer see individual values
- Smooth data values over a bin/region (histogram/density curve)
- Focus on the main features of the distribution
- Caution: Smoothing can reveal or disguise features

# Features Seen in a Visual Summary

- Mode(s) - values concentrate around particular points
- Symmetry – skew left, symmetric, skew right distribution of values about center
- Tails  - long, short, normal (what expect for normal distribution)
- Gaps - regions where no values observed
- Outliers - unusually large/small values

# Distribution of Values



# What to Expect for Traffic Flow Distribution?

- Skew or Symmetric? Why?
- Long or short tails?
- Number of modes?

# Flow Density

?

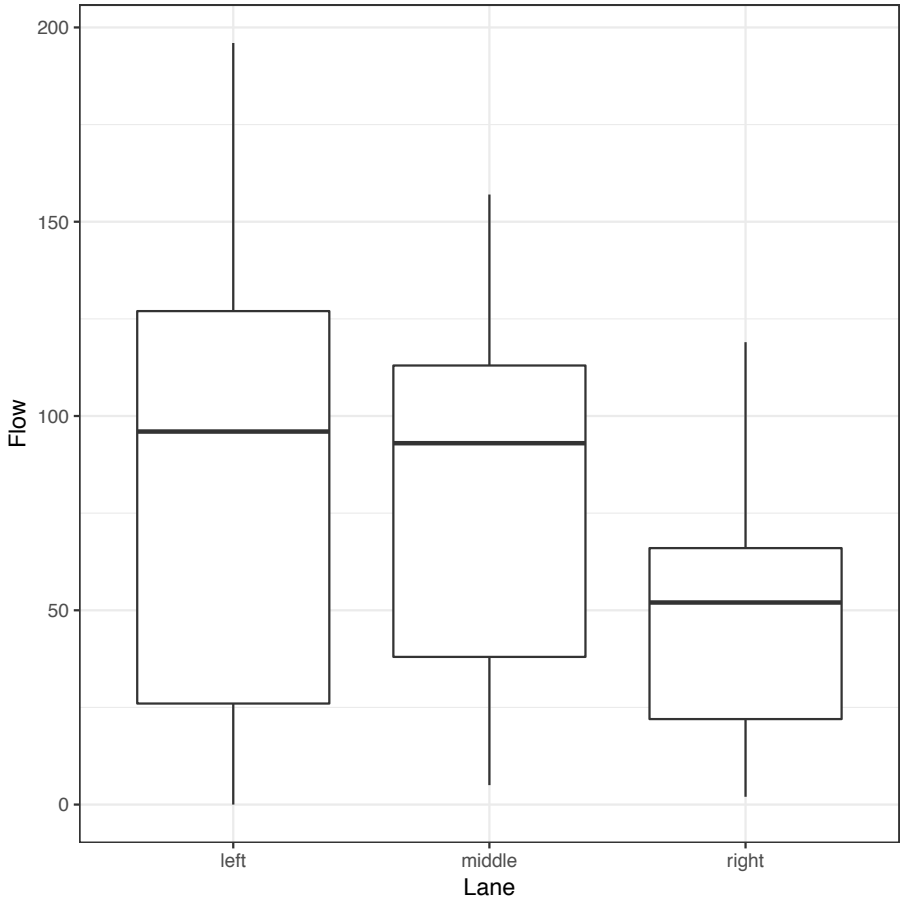
# Do the 3 lanes serve the same amount of traffic?

Flow = count of vehicles in 5-minute intervals

Sum flow - left: 11k, center: 13k, right: 8k lanes

# Compare Flow Across 3 Lanes

# Summary Stats for 3 Flow Distributions




- Quartiles
- Skewness
- Longer right tail

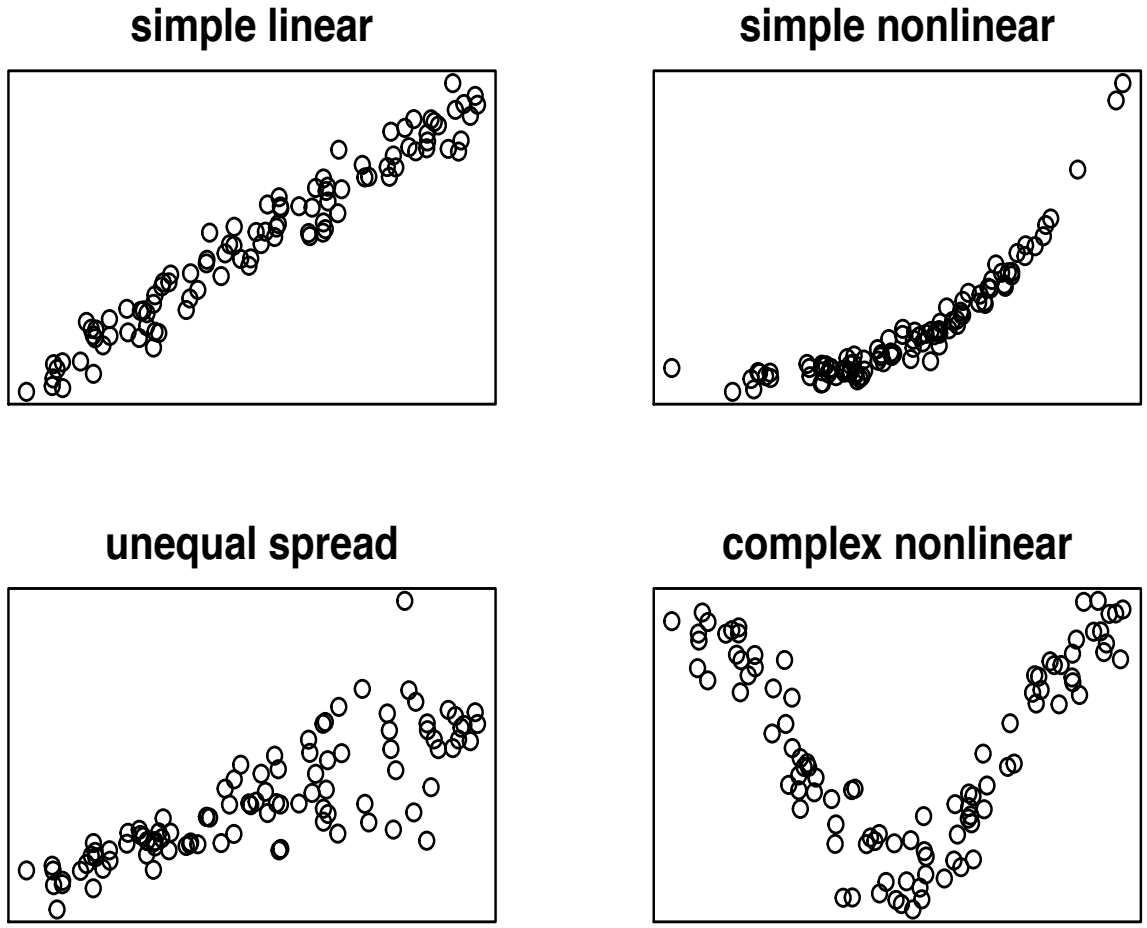
- What don’t we see?

# How Do Lane Flows Vary Together in Time?

# Plotting Pairs of Variables

- Scatter plot uncovers form of relationship between 2 variables
- Linear relationships are particularly simple to interpret
- Simple and elegant statistical theory for linear relationships
- Models are typically approximations, choose a simpler model over a complex one

# Plotting Pairs of Variables



# Flow in: Middle & Right Lanes and Left and Right Lanes

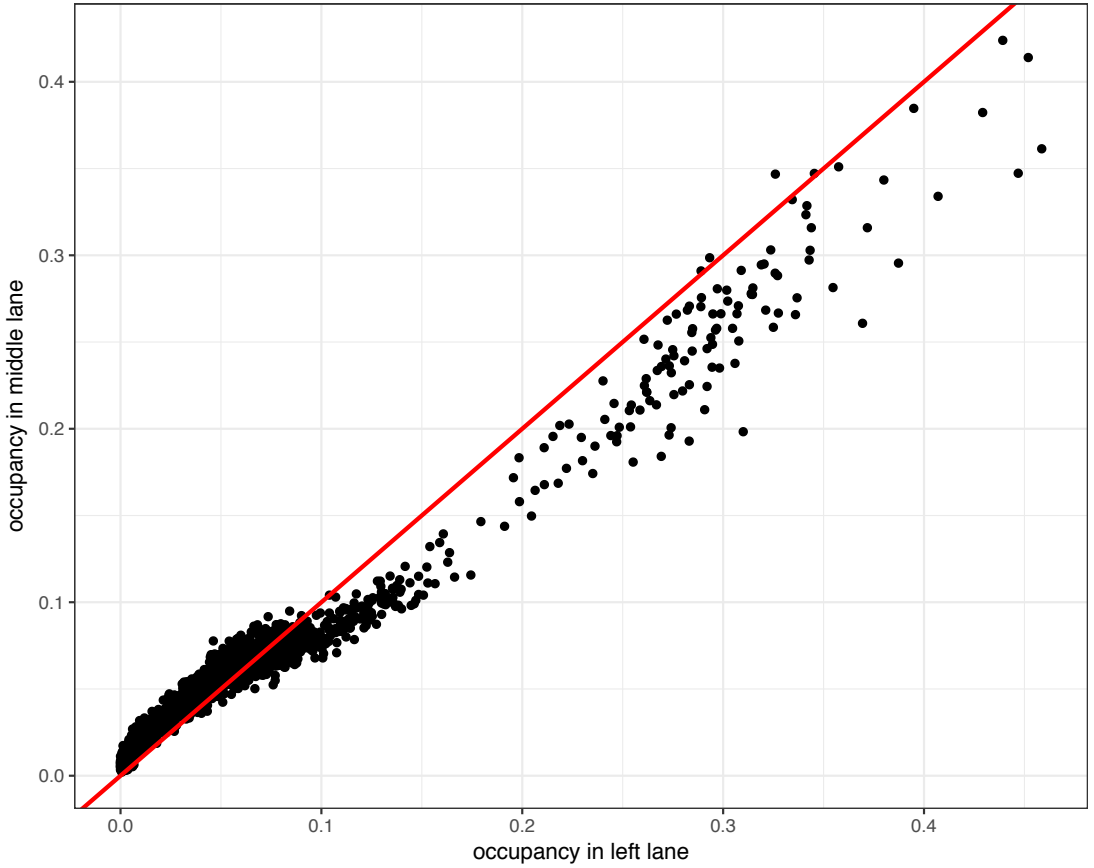
Do we expect:

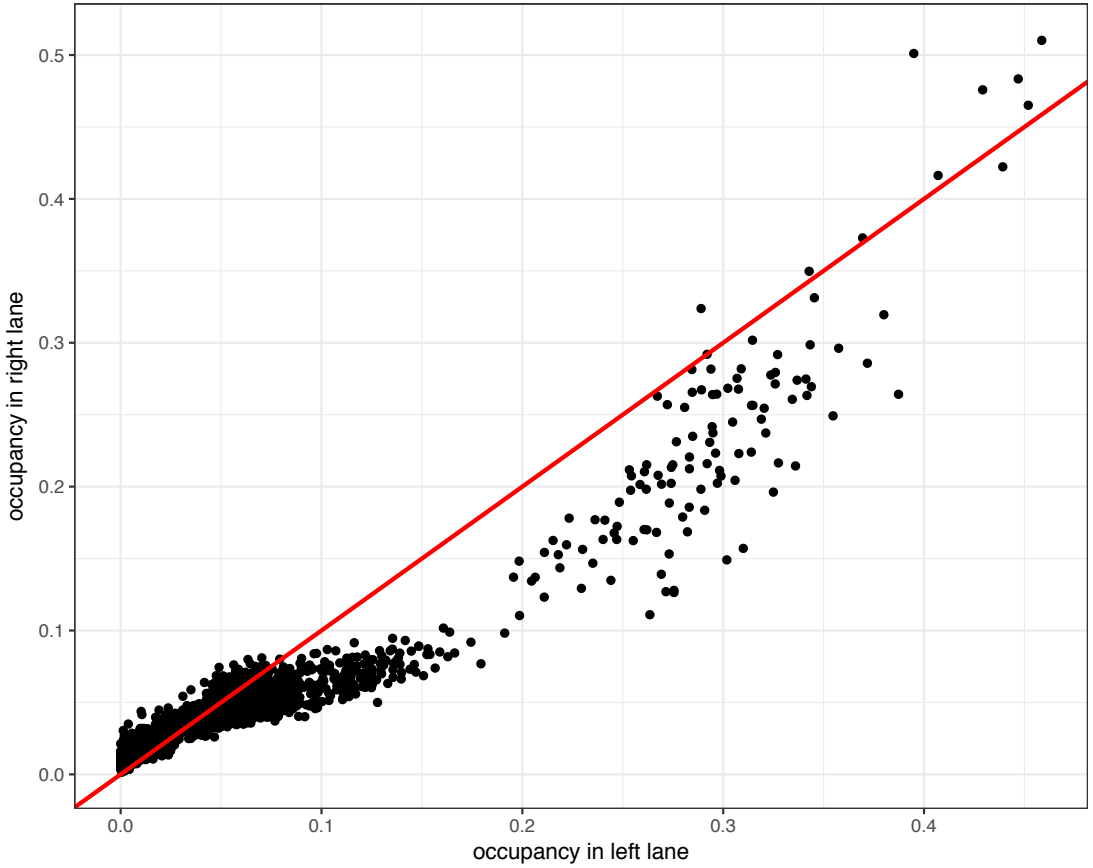
- flow in two lanes increase together?
- flow in two lanes to be linearly associated?
- Slope to be roughly 1?

# Flow in Lanes for the Same 5 Minutes

# When a lane is congested, are the other lanes congested too?

# Occupancy in Lanes for 5 Min. Intervals



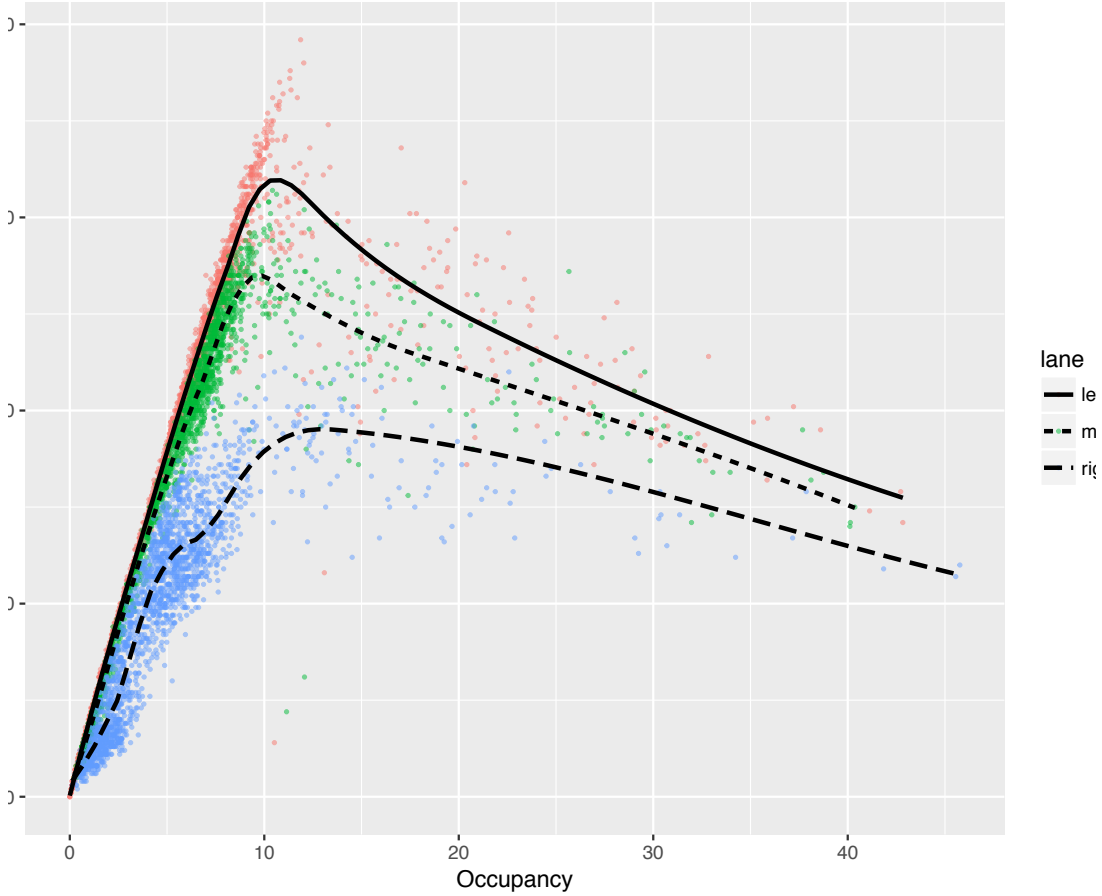
# What's the relationship between amount of traffic and congestion?

# Flow and Occupancy Pairs

Do we expect:

- flow and occupancy to increase together?
- flow and occupancy to be linearly associated?
- 3 lanes to have the same relationship?

# Flow & Occupancy

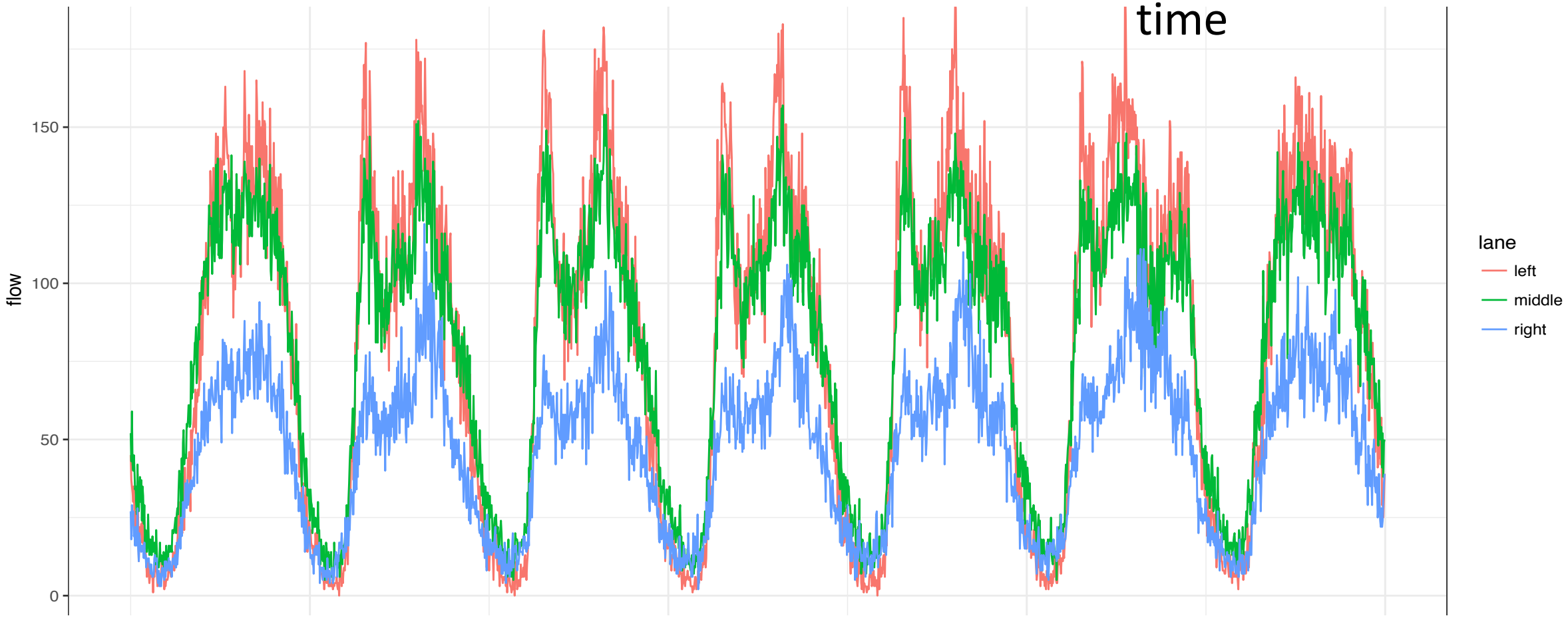


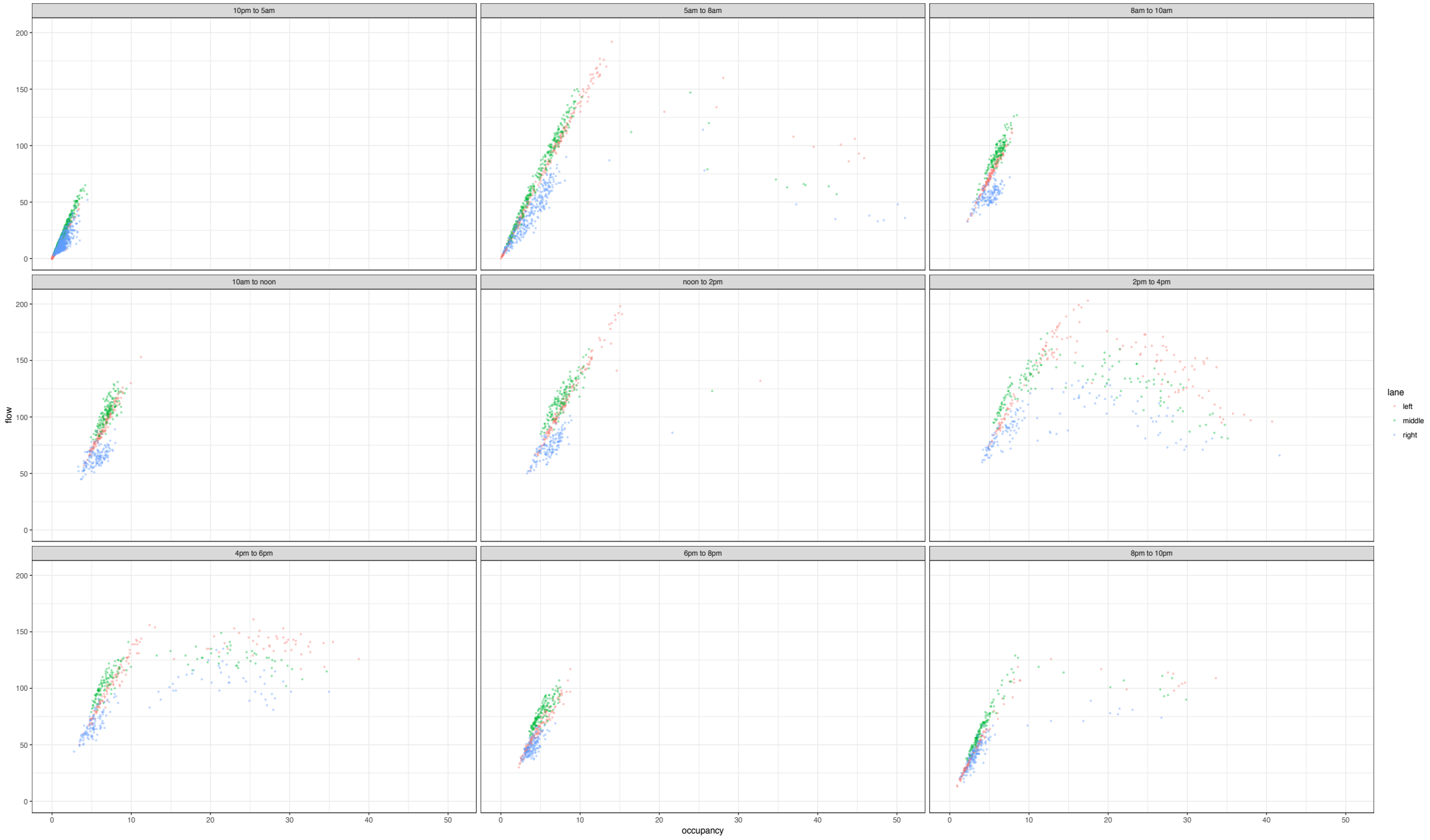

- Flow and occupancy are linearly related for occupancy below 10%
- As occupancy (congestion) increases traffic breaks down and flow decreases
- Breakdown appears to occur at different occupancy levels for the lanes

# How are these findings affected by time of day and day of week?

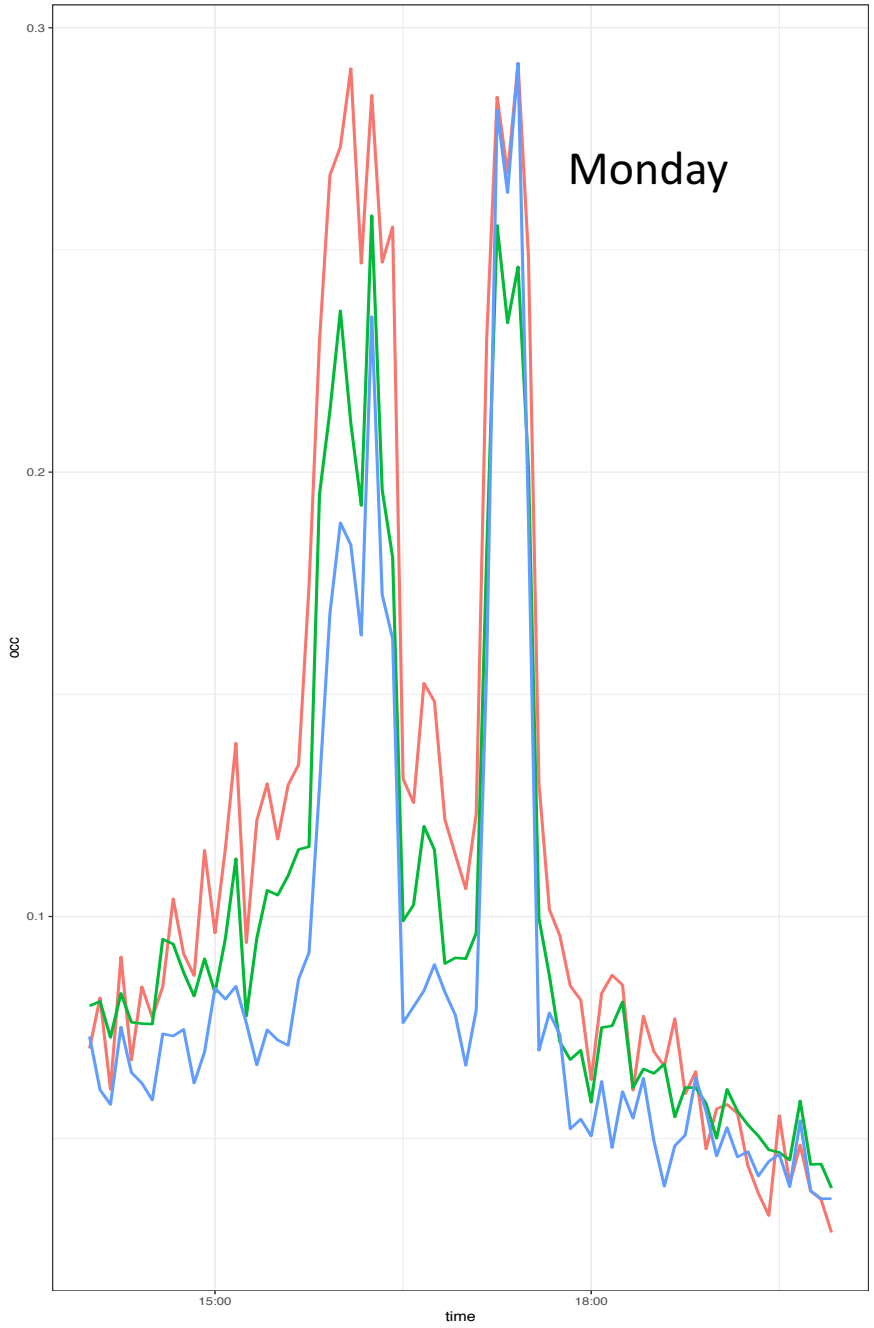
# Examine Flow in Time

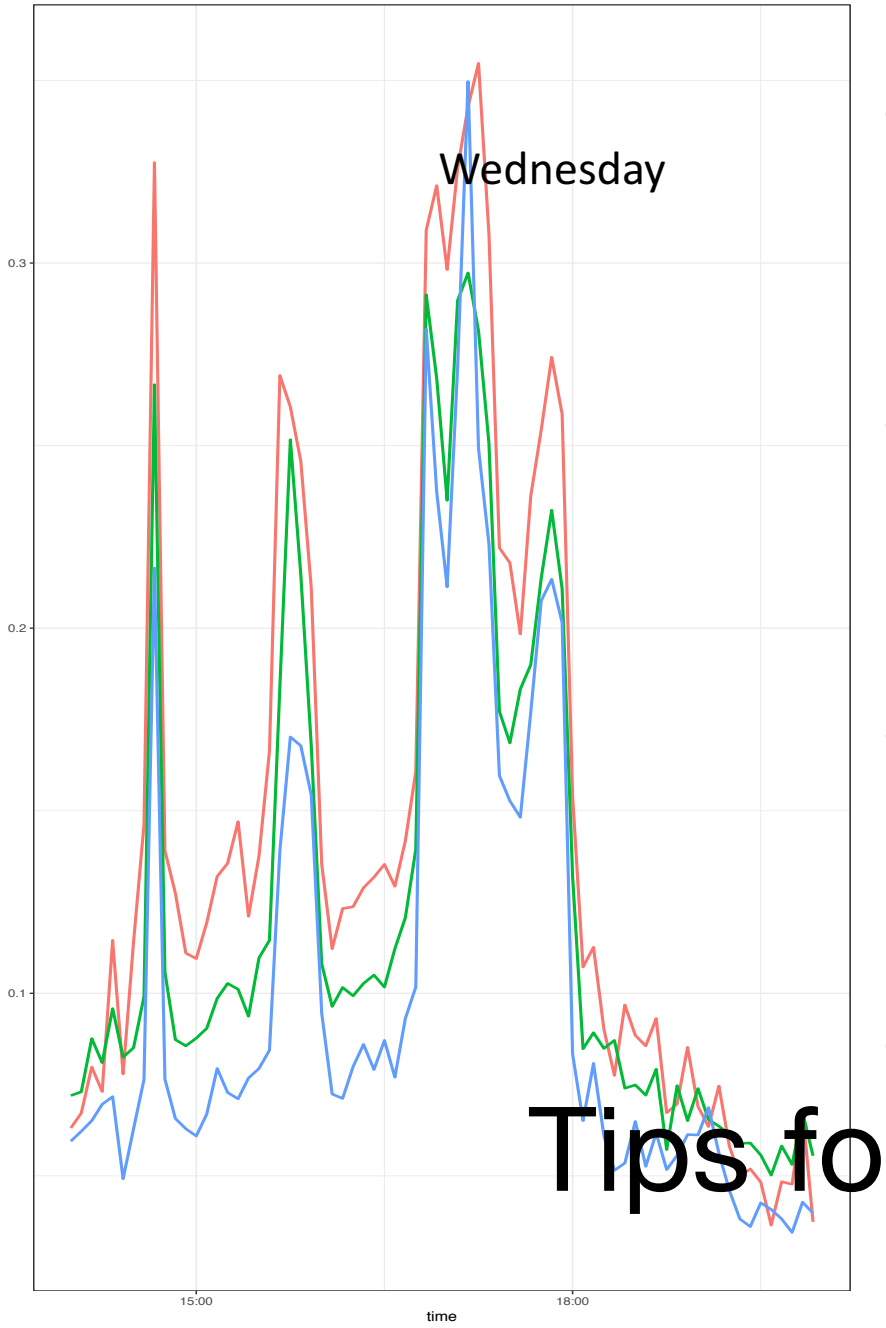
We want to see flow AND occupancy in time

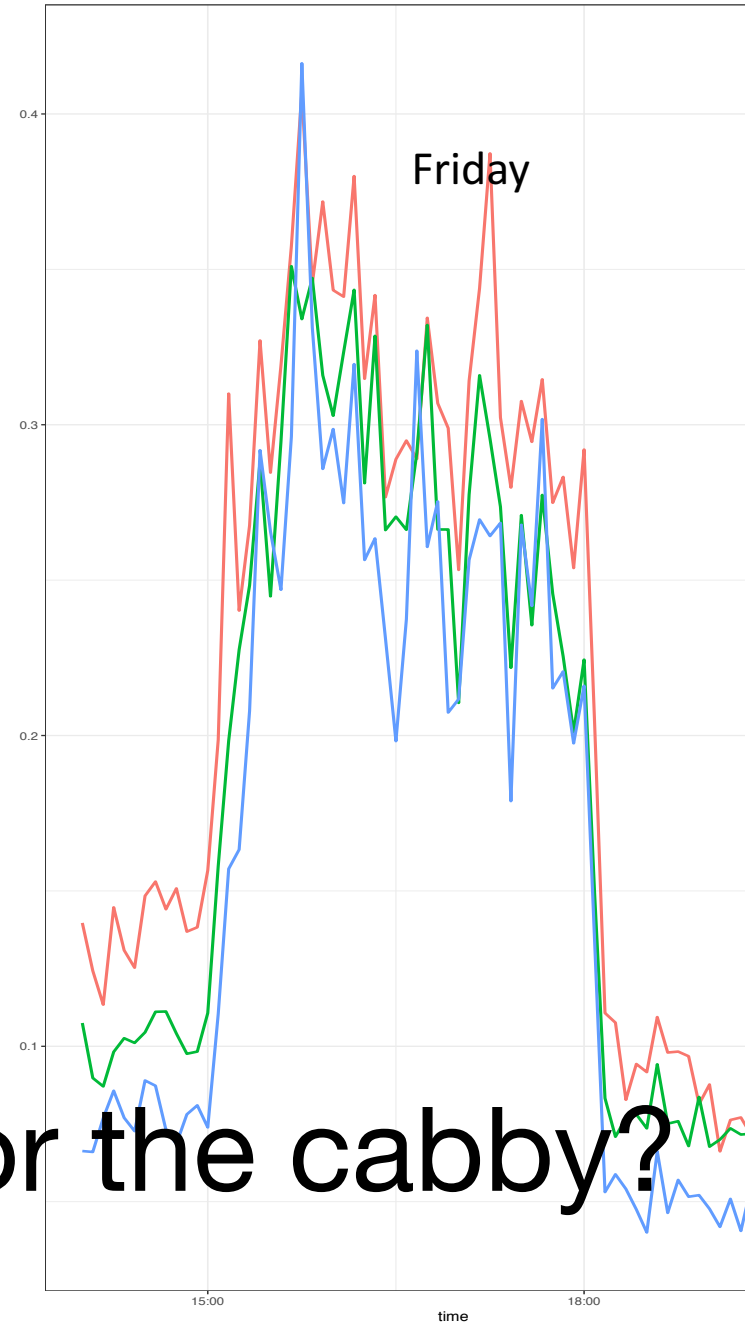



10pm to 5am
5am to 8am
8am to 10am
10am to noon
noon to 2pm
2pm to 4pm
4pm to 6pm
6pm to 8pm
8pm to 10pm
flow
occupancy
lane
left
middle
right

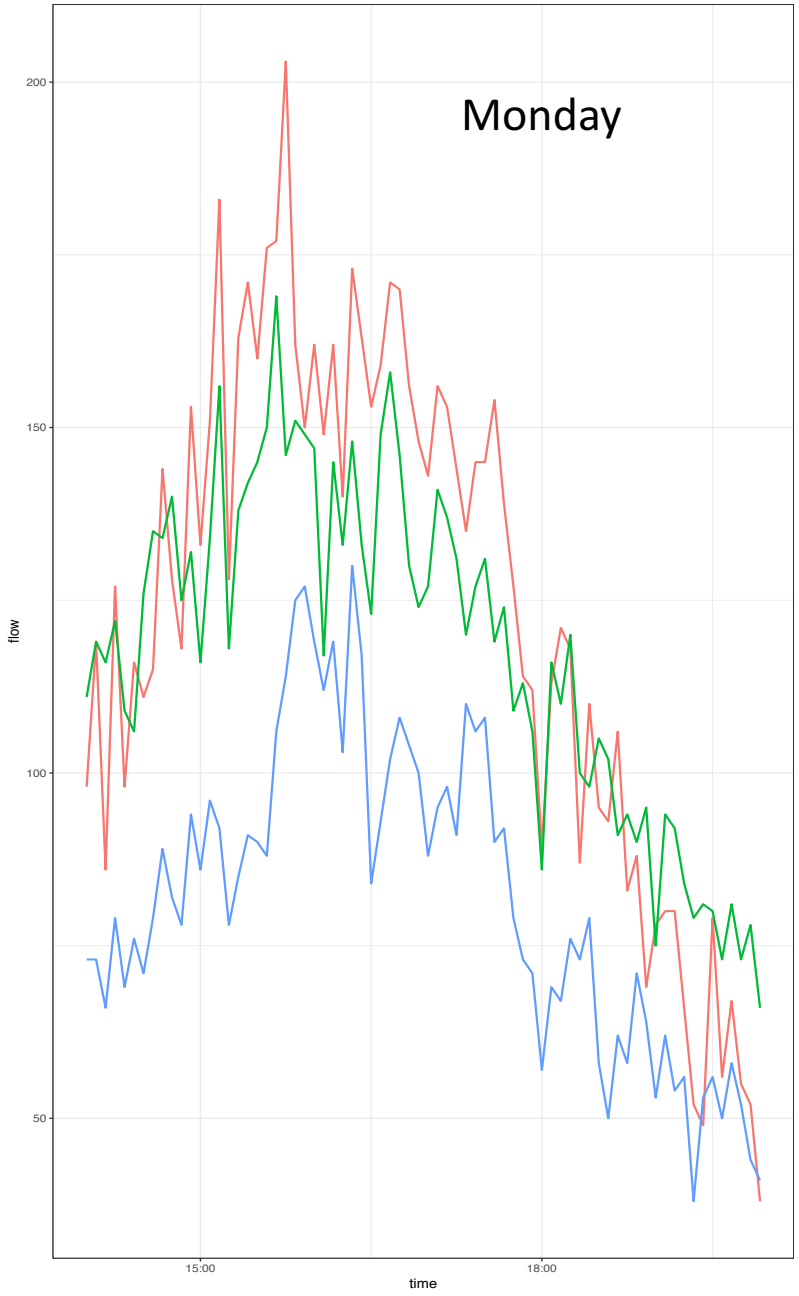
# Focus on Certain Times of Day

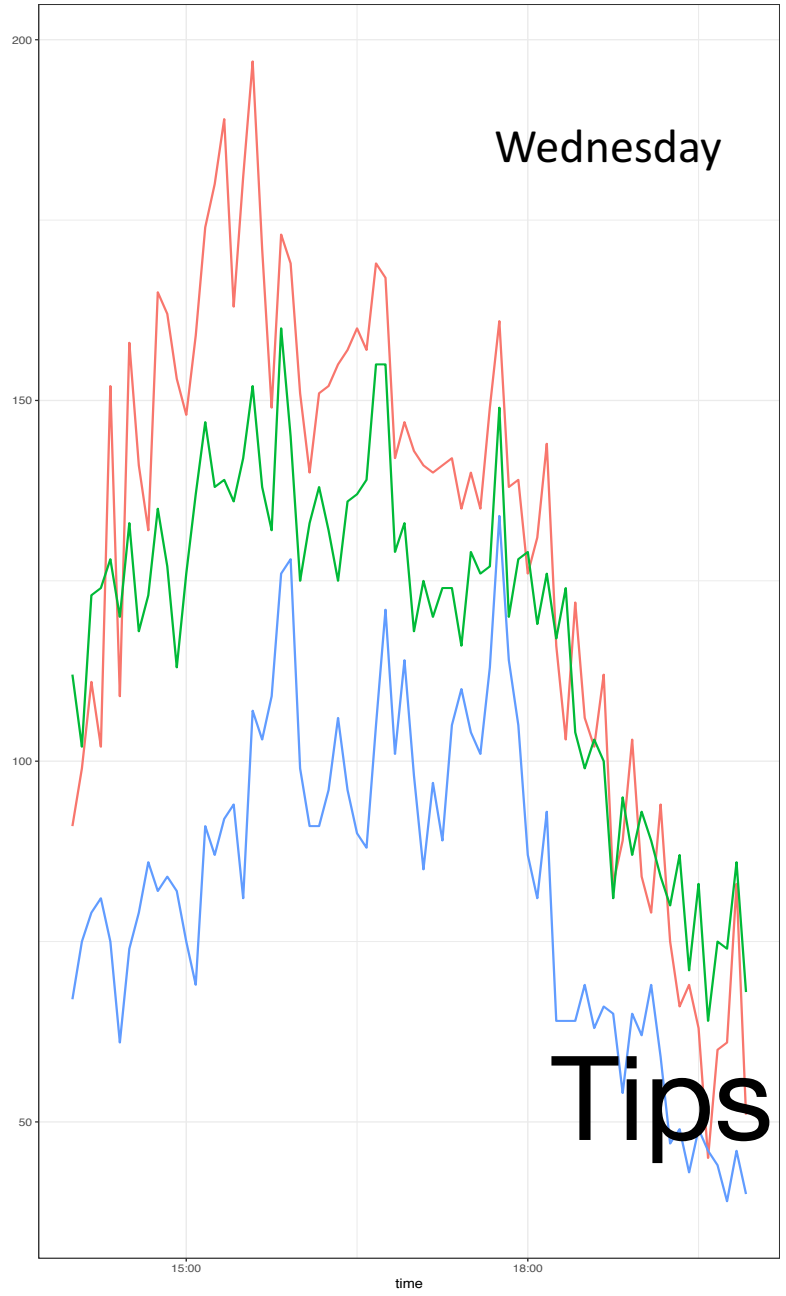


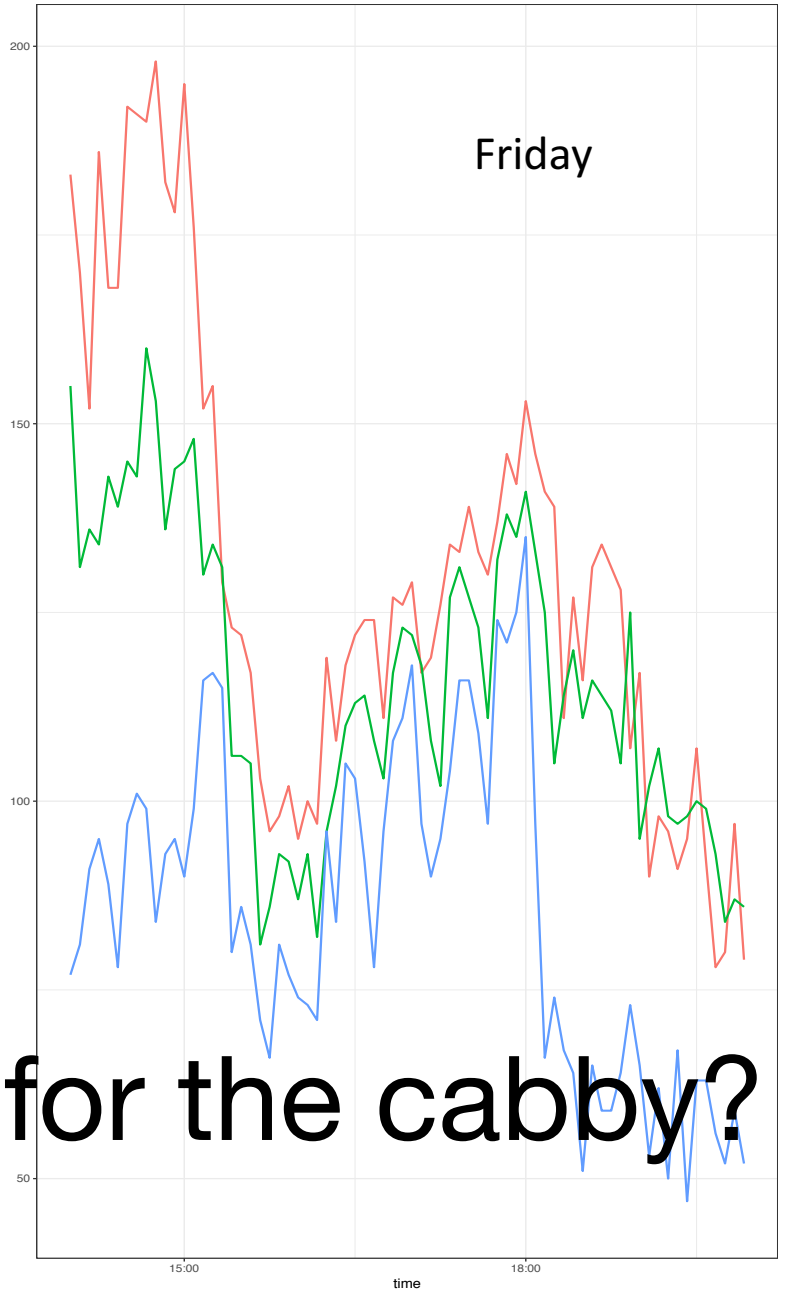




# Tips for the cabby?




# Tips for the cabby?

# Implications for Formal Analysis

- Lane matters – distributions are similar in shape, but locations of peaks and size of spread differ
- Relationship between flow and occupancy is linear until traffic breaks down
- Traffic jams do not have the same relationship, spread increases, negative association between flow and occupancy
- Distinct patterns within a day and for day of week