“CitySense”: Introduction to Probabilistic Modeling

Julia Kempe & David S. Rosenberg

CDS, NYU

March 5, 2019

Contents

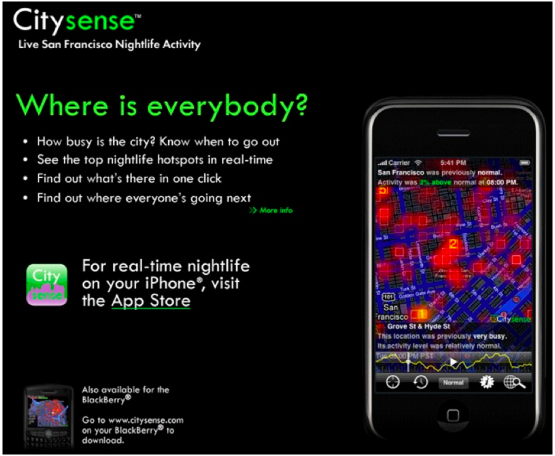
The CitySense Problem

Sense Networks

- Startup company incorporated around 2006.
- Objective: Develop and leverage expertise in **location data** analytics.
- First product was called CitySense[1] (2008).
  - A real-time, data-driven guide to nightlife in San Francisco.

[1]See "CitySense: Multiscale space time clustering of GPS points and trajectories" by Markus Loecher and Tony Jebara (2009). http://www.cs.columbia.edu/~jebara/papers/CitySense.JSM2009.pdf

CitySense (2008)



(Sadly, no longer in the App Store.)

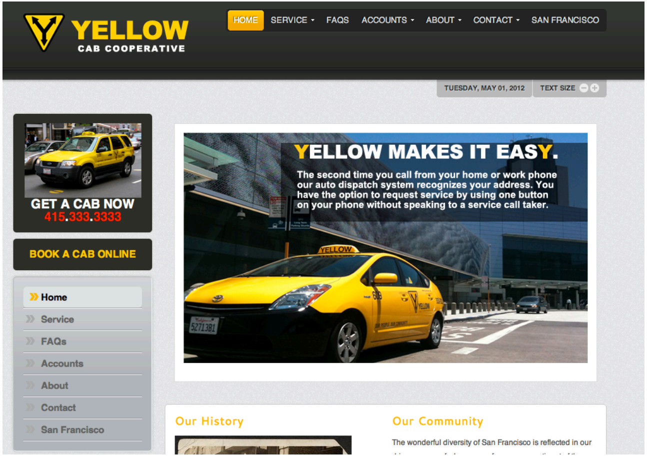
CitySense: Use Cases

Two use cases:

1. I'm new to the city – where does everybody hang out at night?
2. I know the city, but is there anything **special** going on tonight?

CitySense: Data Source

- Taxi GPS data for sale in San Francisco



CitySense

- Main Idea: Taxi destinations are a proxy for where people are going.
- Can use taxi data to bootstrap
  - Once we had users, we could use the locations from their phones.
- Taxi feed is **real-time**, so can use it to find those big secret parties.

CitySense

Data Science Strategy

1. Model "typical" behavior of each area of the city.
2. Rank areas with activity levels that are "most unusual".

We'll discuss modeling strategies shortly.

Plan for this lecture

- Examine the CitySense "anomaly detection" problem.
- But use the NYC taxi pickup data – more local and more recent.
- Our dataset is from 2009.
- Currently (2017/11/09) you can download 2013 data from https://github.com/andresmh/nyctaxitrips
- You can also request data directly from the NYC Taxi and Limousine Commission via the Freedom of Information Law. http://www.nyc.gov/html/tlc/html/passenger/records.shtml

The Case for Probability Models

Predicting Probability Distributions

So far we've discuss two problem classes:

- **Classification**
  - Outcome space $\mathcal{Y} = \{-1, 1\}$
  - Action space $\mathcal{A} = \mathbf{R}$ (threshold to get hard classifications)
- **Regression**
  - Outcome space $\mathcal{Y} = \mathbf{R}$
  - Action space $\mathcal{A} = \mathbf{R}$.
- Today we consider a third type of **action space**:

$$\mathcal{A} = \{\text{Probability distributions on outcome space } \mathcal{Y}\}$$

- Why?

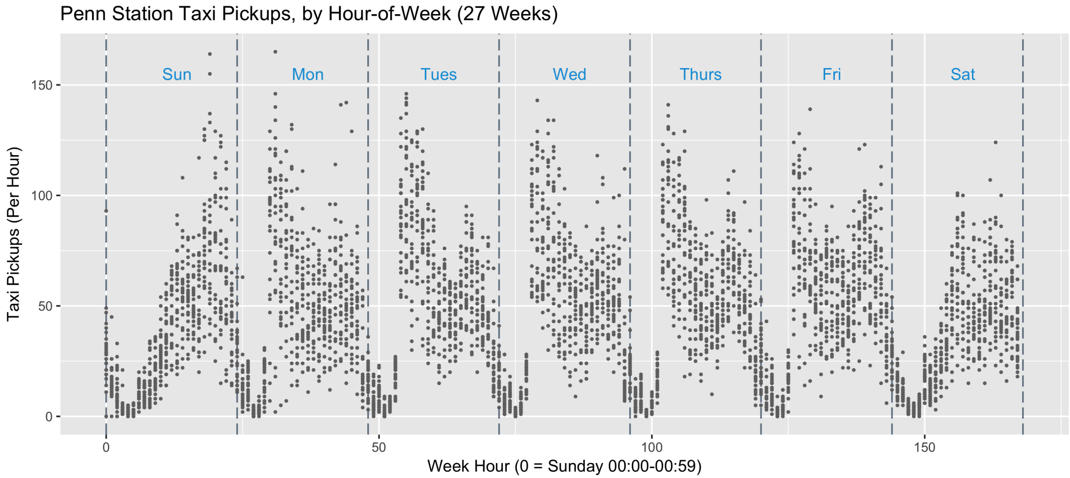
The Joy of Probability Distributions

- Outcome space $\mathcal{Y} = \mathbf{R}$ (some regression problem)
- For input x, suppose we produce a **probability density** on $\mathcal{Y}$:

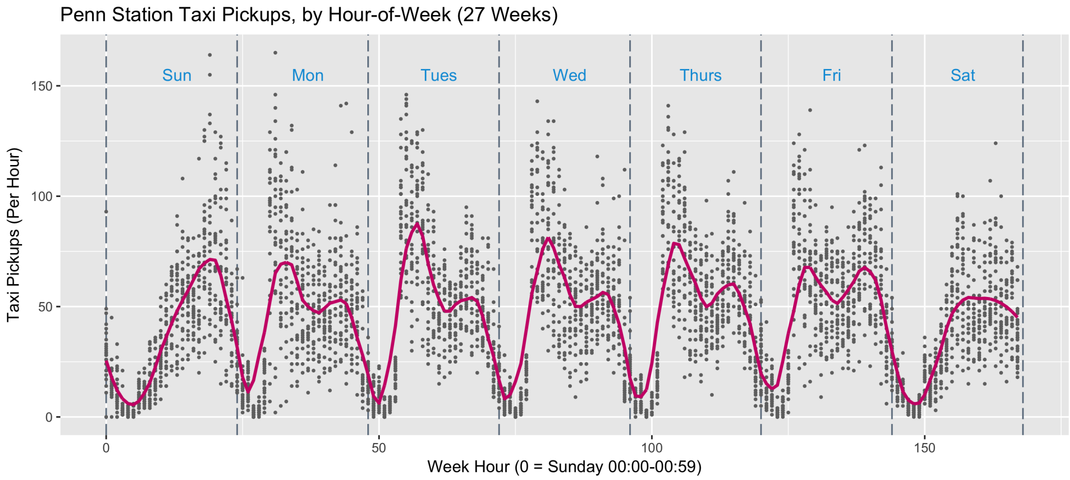
$$x \mapsto p(y)$$

- We can interpret this setting as modeling the **conditional probability density** $p(y \mid x)$.
- If we know $p(y \mid x)$, we can find a $\hat{y}$ that minimizes any loss function for a given x:
  - For square loss, give the mean of $p(y \mid x)$. [From homework]
  - For ℓ_1 loss, give the median of $p(y \mid x)$. [From homework]
  - Can produce a **prediction interval** that $p(y \mid x)$ assigns a 95% probability
-

Penn Station Taxi Pickup Counts - 27 Weeks

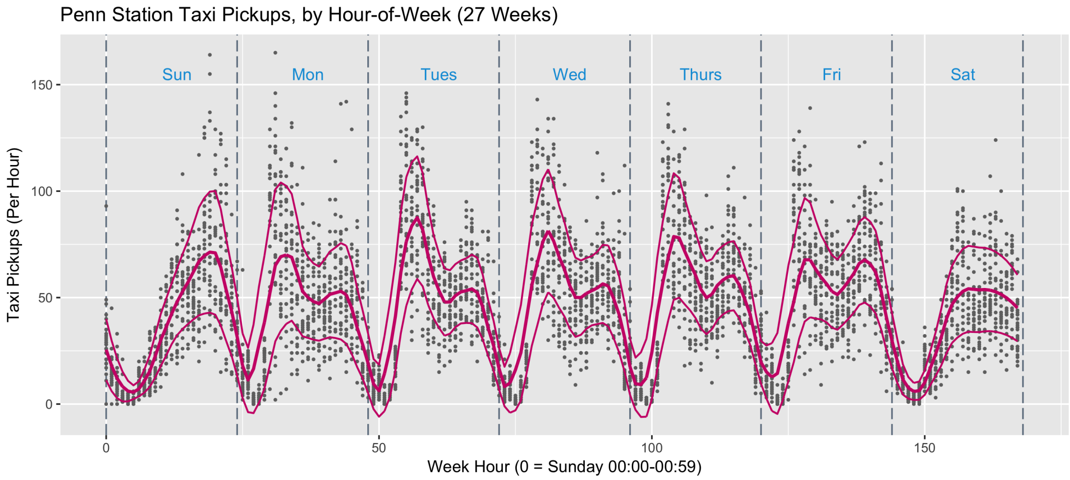


Penn Station Taxi Pickup Counts - Regression



Regression line predicts **mean pickups**. But what's the typical range?

Penn Station Taxi Pickup Counts - Prediction Intervals



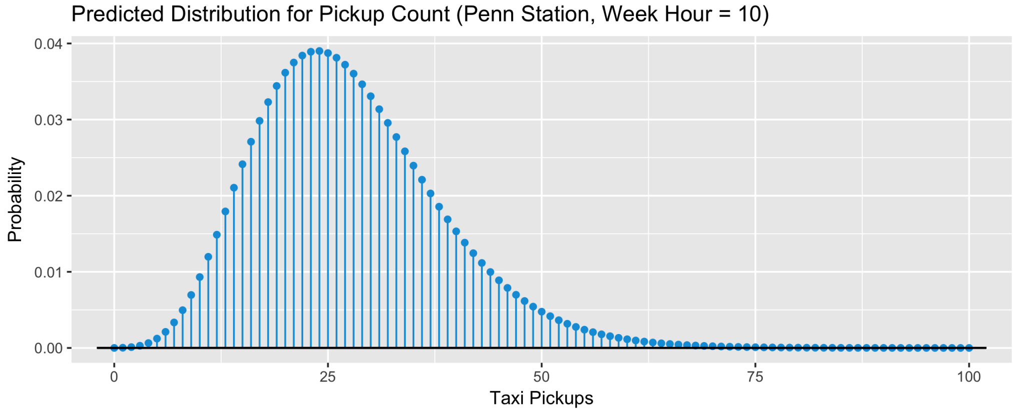
Here plotting estimated ± 1 standard deviation.

Penn Station Taxi Pickup Counts - Predictive Distribution

- Consider predictions for a particular weekhour $x \in \{0, \ldots, 167\}$, say $x = 10$.
- **Regression** gives a single number: $\mathbb{E}[y \mid x = 10] \approx \mathbf{30.1}$ taxi pickups
- A **prediction interval** gives two numbers: $\mathbb{P}(y \in [\mathbf{17.8}, \mathbf{42.3}] \mid x = 10) \approx 68\%$.
- We can also produce an estimate of the full **conditional probability distribution** for $p(y \mid x = 10)$...

Penn Station Taxi Pickup Counts - Predictive Distribution

- For weekhour 10 (i.e. $x = 10$), we predict the following distribution for $p(y \mid x = 10)$:



- According to this predictive distribution, how likely are we to get 90 taxi pickups?

Predictive Distributions for Anomaly Characterization

- At week-hour 10,
  - the expected number of taxi pickups 30.1.
  - the 68% prediction interval was $[17.8, 42.3]$.
- Suppose we observe 90 taxi pickups.
- How can we characterize how unusual this event is?
- We can directly calculate the probability of 90 or more taxi pickups:

$$\mathbb{P}(y \geqslant 90 \mid x = 10) = \sum_{c=90}^{\infty} p(y = c \mid x = 10)$$

  measures how unusual this event is.

Prediction Intervals from Probability Distributions

- Given a conditional probability distribution $p(y \mid x)$,
  - it's usually straightforward to compute a **prediction interval**.
- A 95% prediction interval is an interval $[a, b]$ such that

$$\mathbb{P}\left(y \in [a, b] \mid x\right) \approx .95$$

- We can get $[a, b]$ by finding the 2.5% and 97.5% quantiles of the distribution $p(y \mid x)$.
- [Alternatively, can do this with **quantile regression**.]

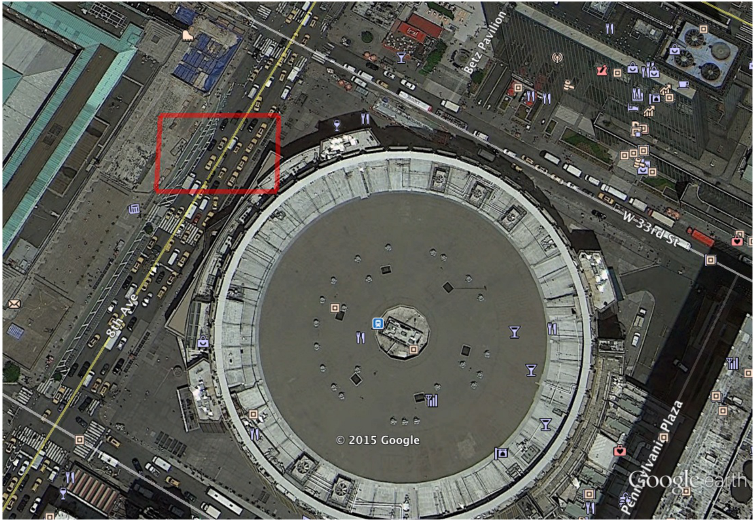
The Grid Cells

The Basic Approach

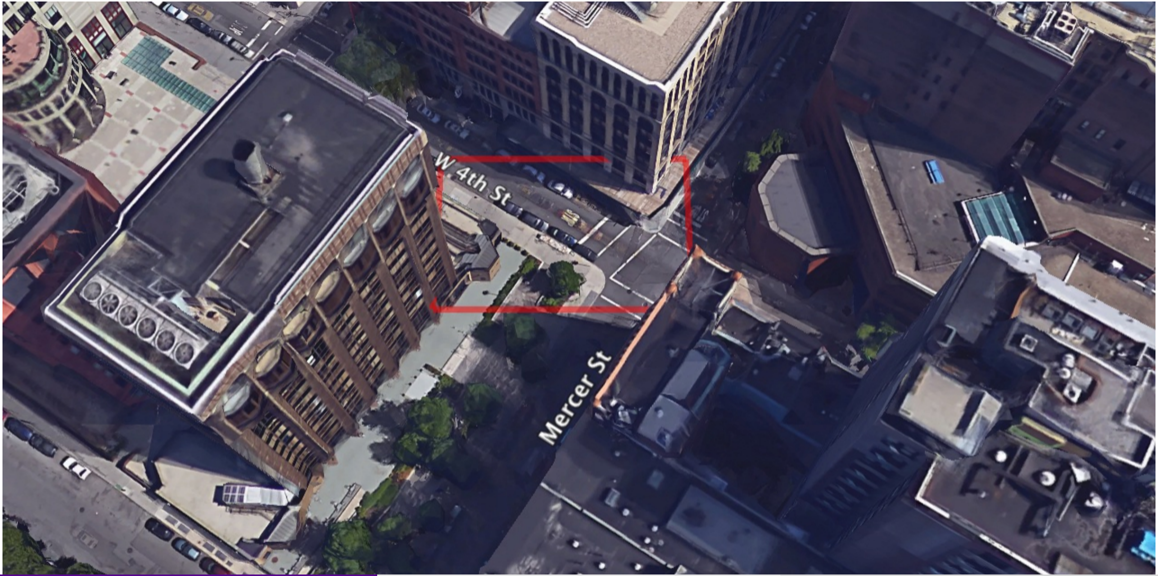
- Raw input is [roughly] continuous in
  - space (lat/lon) and
  - time (seconds since 1970-01-01).
- To make it easier to handle, we partition space and time into buckets.
- Spatial partitioning
  - Divide earth into regularly spaced grid cells.
  - About 400,000 grid cells to cover NYC
- Time partitioning
  - Only consider times at the hour level.
- Aggregate taxi pickup counts at the Grid Cell / Hour level.

Initial data analysis, including aggregation by grid cell and hour, was done by Blake Shaw.

Most Active Grid Cell: Penn Station (Grid ID 7750)

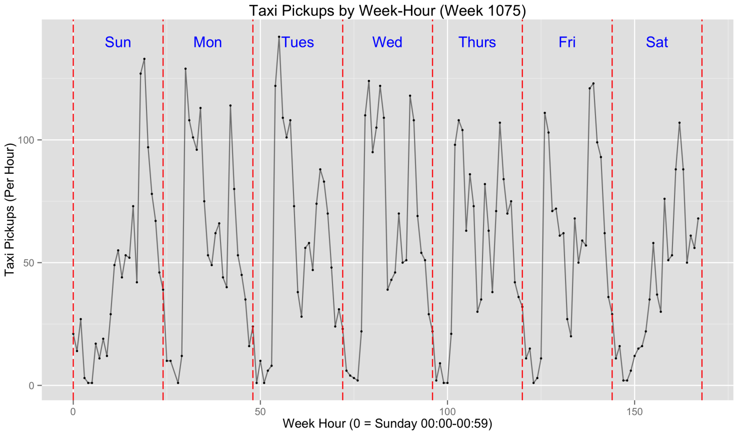


Courant Institute (Grid ID 21272)



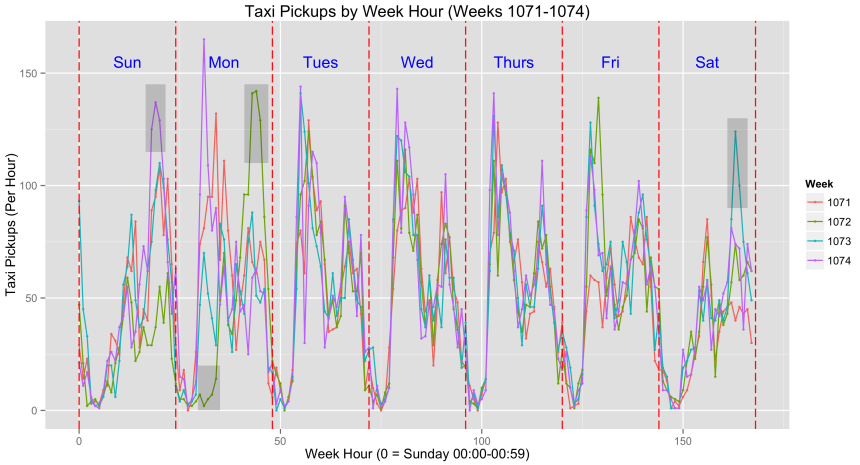
Data Visualization

Penn Station (Cell 7750): 1300 Taxi Pickups Per Day

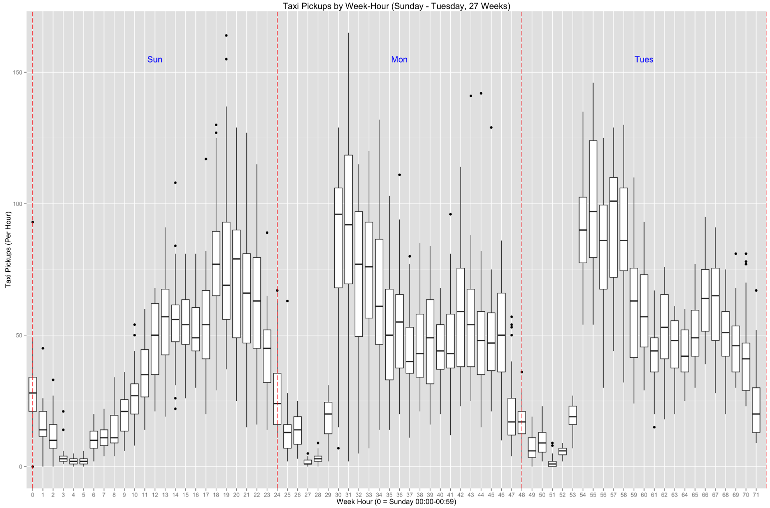


Note difference between weekend and weekday patterns.

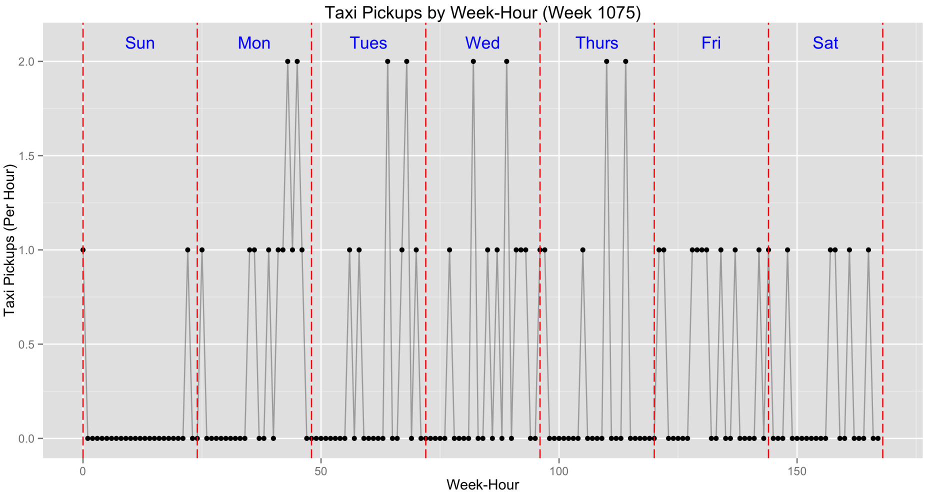
Penn Station (Cell 7750): Four Weeks, Some Outliers



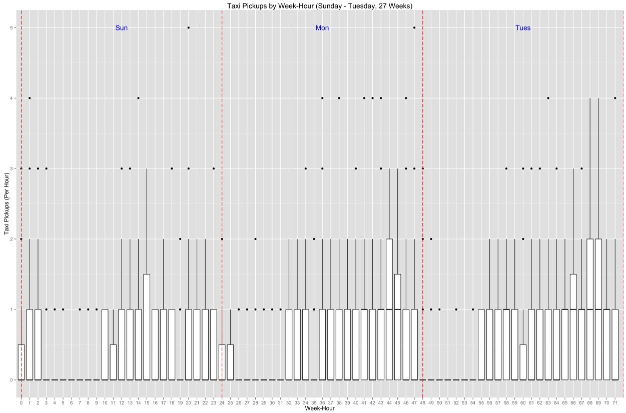
Penn Station: Sunday-Tuesday, 27 Weeks



Courant (Week 1075): 12 Taxi Pickups Per Day



Courant Institute: Sunday-Tuesday, 27 Weeks



Note: At least 25%, sometimes 75%+ of counts are zero.
Box plot clearly shows extreme values (ranging up to 5).

The Prediction Problem

The Prediction Problem

Somebody queries a **grid cell** and a **week-hour**, we tell them what to expect.

- Input space: $\mathcal{X} = \{(g,h) \mid g \in \{1,\ldots,398245\} \text{ and } h \in \{0,\ldots,167\}\}$, where
  - g is the grid Cell ID and
  - h is the week-hour
  - Possible future inputs: Holiday? Raining? Special event?
- Action space: $\mathcal{A} = \{\text{Probability distributions on number of pickups}\}$
- Outcome space: $\mathcal{Y} = \{0,1,2,3,\ldots\}$
  - Actual number of taxi pickups.
- Evaluation? Loss function? We'll come back to these questions...

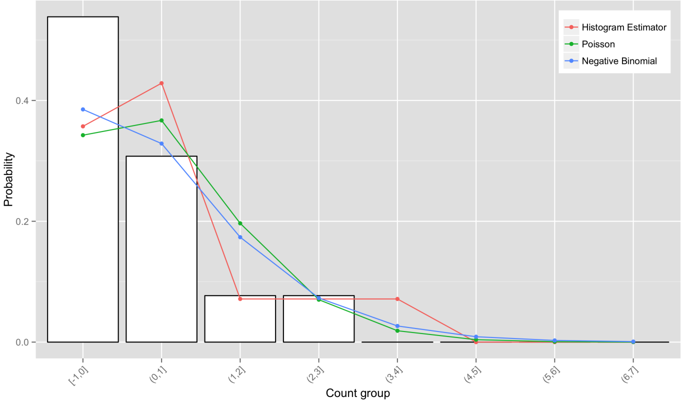
Setting up the Learning Problem

- Labeled data look like:
  - (Grid Cell = 10321, Week Hour = 120) $\mapsto$ Count = 3
  - (Grid Cell = 192001, Week Hour = 6) $\mapsto$ Count = 12
  - (Grid Cell = 1271, Week Hour = 154) $\mapsto$ Count = 0
- How to split the data into a training set and a test set?
- Our approach:
  - First 14 weeks are **training set**.
  - Last 13 weeks are **test set**.

Stratification Approaches

Approach 1: Full Stratification (Courant, Tuesdays 7-8pm)

- Estimate distribution for each grid cell / week hour pair.
- Colored lines are from training. White bars are from test.



Terminology: Stratification and Bucketing

Definition

We say we are **stratifying** if we partition our input space into groups, and treat each group separately. For example, in modeling we would build a separate model for each group, without information sharing across groups.

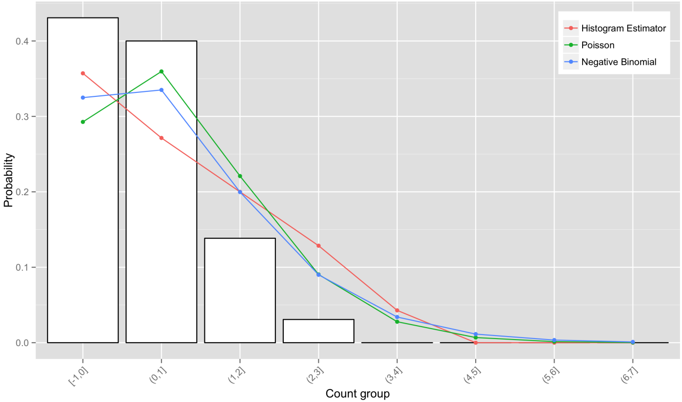
On the other hand,

Definition

We say we are **bucketing** (or **binning**) if we are combining natural groups in the data into a single group, rather than building a separate model for each group. For example, combining all weekdays together would be "bucketing".

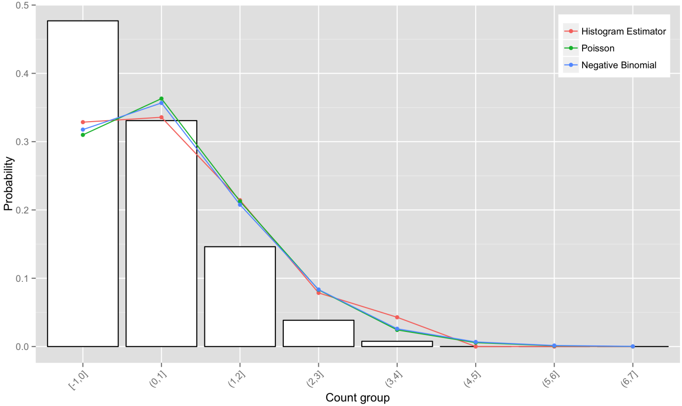
Approach 2: Weekday Bucketing (Courant, M-F 7-8pm)

- Data inspection suggests that day patterns are similar Mon-Fri.



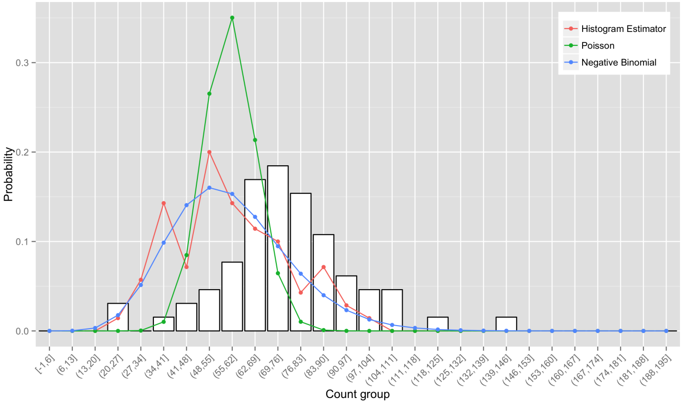
Approach 3: (Courant, M-F 6-8pm)

- Also, 6-7pm looks similar to 7-8pm, so join together



Penn Station, M-F 7-8pm

- Negative binomial fits empirical much better than Poisson. (overdispersion)
- Massive shift between train and test!



The Estimation/Approximation Tradeoff of Stratification

- With a separate probability distribution for every grid cell / week-hour pair, model is highly specific!
- Could capture idiosyncrasy of Friday @5pm that we would miss if combining all weekdays.
  - That is, we're decreasing approximation error.
- With relatively little data in a particular stratum, estimates may have high estimation error.
- By "bucketing", or combining strata:
  - We can reduce estimation error.
  - It may cost us in approximation error.
  - By bucketing in a smart way, you can minimize increase in approximation error.
- Note: This is often referred to as a **bias / variance tradeoff**:
  - bias $\approx$ approximation error; variance $\approx$ estimation error

Is there a more convenient way?

- We can tradeoff between estimation and approximation error by varying the stratification and the bucketing.
- It's a great way to start your data analysis.
    - You get a feel for the data and gain some intuition.
- Our classification and regression techniques also trade off between approximation and estimation:
    - We had to choose our features.
    - We had to tune our regularization parameter.
- Can we do something similar for predicting distributions?
- Yes – to be discussed in our module on conditional probability modeling.