ORCA 2017
One Regional Card for All

Project Lead & Data Scientists:
Mark Hallenbeck, Anat Caspi, Bryna Hazelton,
Michael Wolf, Jake VanderPlas

DSSG Fellows

Mayuree Binjolkar          Daniel Dylewsky          Andrew Ju          Wenonah Zhang
ORCA Transactions Data

Information helps to improve transit performance
First Step: Geolocate the ORCA boardings

ORCA Transactions Data

Automatic Vehicle Location Data
ORCA Boardings - Only Half of the Story
Estimate Destination of the Trip
Transfer Analysis

Objectives

Transfer Data

Real vs. Financial Transfer?

Cleaner Origin/Destination Data

- Headway
- Walking Time
- Transfer Duration
- # bus missed
Model Selection Stage

- No ground truth to conduct supervised learning
- Gaussian Mixture Model did not perform as well as expected
- K means Unsupervised learning oversimplified the clusters
- The amount of labeled data based on human intuition is not sufficient for supervised learning

Why Semi-supervised Learning with Label Spreading Algorithm?

- Performs well with a small amount of labeled data
- Considerable improvement in learning accuracy when use unlabeled data in conjunction with labeled data
Label Spreading Result

- Real Transfer: 5%
- Financial Transfer: 7%
- Unlabeled: 88%

- Real Transfer: 83%
- Financial Transfer: 17%
ORCA Data is Biased and Variable

APC-ORCA factors by TAZ region

APC - ORCA Factors
Zero Inflated Negative Binomial Regression

• Regression approach is highly interpretable to understand bias
• Count data is well suited for Poisson distribution
• Overdispersion (mean > variance) suggests a Negative Binomial variant of the Poisson distribution
• High number of 0s and noisy data encourages a zero inflated / mixed model approach
Distribution for APC Count (response variable)

Poisson Distribution vs. Data for All Routes/TAZ
Continuing Work...

- Validating semi supervised learning models for transfer analysis
- Zero Inflated Negative Binomial Model on Entire Network
- Neural Nets (scalability, diversity)