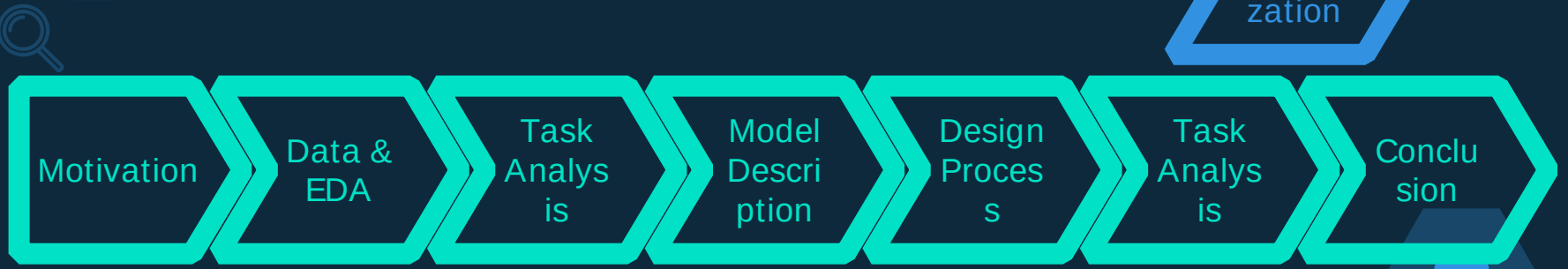
City Recommendation in the USA using Yahoo Flickr Creative Commons 100M Dataset

by Anant Jain, Ahmet Salih Gündoğdu

DS 5500 Fall 2018 --- Prof. Cody Dunne, Northeastern University

Layout


Motivation
Data & EDA
Task Analysis
Model Description
Design Process
Task Analysis
Conclusion
Final Visualization

1 Motivation

Let’s start with the first set of slides


facebook
The Batcave
ROBIN
The location of our Bat Cave is meant to be secret, so STOP CHECKING IN!

A picture is worth a thousand words

A complex idea can be conveyed with just a single still image, namely making it possible to absorb large amounts of data quickly.

2

Data & EDA

Yahoo Flickr Creative Commons 100M

In short, YFCC100M

- One of the largest assemblages of multimedia check-ins ever created
- Publicly hosted on AWS
- Released under the Yahoo Web-Scope program
- Hundred million media objects dating between 2004 and 2014

Pruning

ELIMINATED UNWANTED COLUMNS

- Workable with limited RAM
- Omitting records that weren't geo-tagged (i.e. more than 50%)
- Omitting records that came with a wrong date format (0.01%)

FILTER TO USA

- YFCC100M Places - Expansion Dataset
- Reverse geocode information of all records.

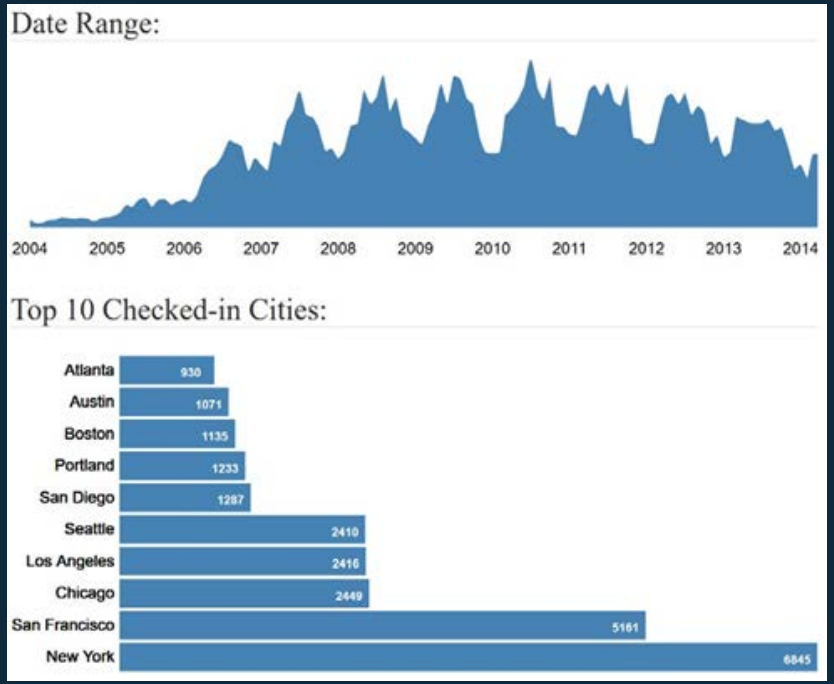
YFCC100M + Pruning + Merging + Cleaning = YFCC_USA16M

Columns

pid	Unique media identifier
user_nickname	User identifier
date	Date the media object was created
longitude	Longitude of the location the media object was checked at
latitude	Latitude of the location the media object was checked at
town	Town the media object was checked in
state	State the media object was checked in

EDA

EDA (contd.)


Date Range:
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Top 10 Checked-in Cities:
Atlanta
930
Austin
1071
Boston
1135
Portland
1233
San Diego
1287
Seattle
2410
Los Angeles
2416
Chicago
2449
San Francisco
5161
New York
6845

Objective

Utilize the travel check-in data and use data-based visualizations to explore, assess and evaluate multiple SVD algorithms for the purposes of identifying anomalies, generating trust and providing the best recommendation for cities to visit in the USA

3 Task Analysis

Tasks

Priority	Domain Task	Analytic Task	Search Task	Analyze Task
3	Examining and evaluating the model performance of the recommended places against the given user's travel history	Compare	Locate	Present
2	Generate a ranked list of recommendations	Sort	Explore	Present
1	Visualize different models and hyperparameters for assessment of the best set of modeling parameters to use.	Compare	Explore	Discover
4	Exploratory Data Analysis	Compare	Explore	Discover

Intended Users

Experts

Researchers and machine learning engineers who are interested in recommendation systems.

Travelers

Anybody who wants to get travel recommendations in the USA

4 Model Description

Backend

Assorted selection of Hyperparameters and Models

PREPROCESSING

- Numeric: #
- Binary: 1 or 0

MODELS

- SVD_explicit
- SVD_implicit: Alternating Least Squares

LATENT DIMENSIONS

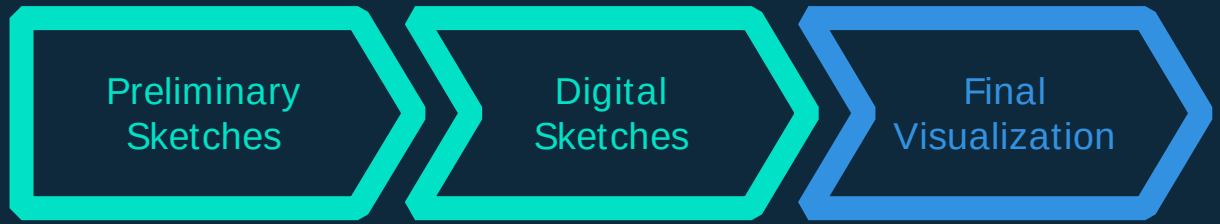
- Number of dimensions/features to extract for each user and location

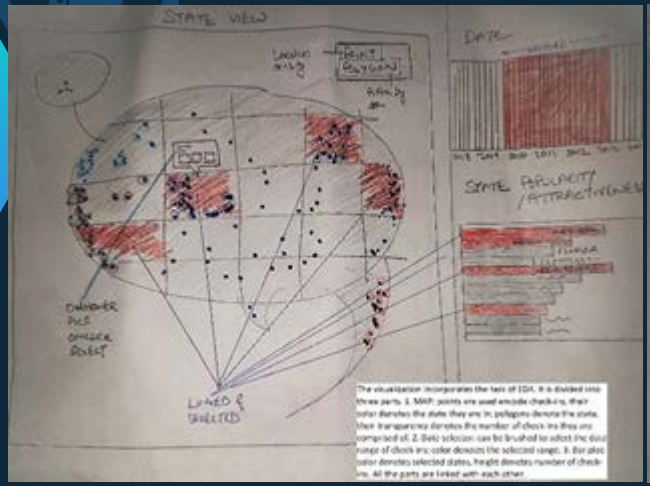
METRIC

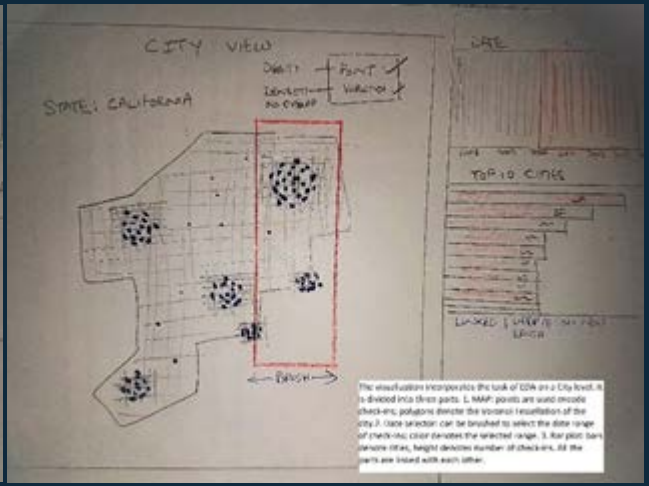
- Precision-Train Set
- Recall-Train Set
- Precision-Validation Set
- Recall-Validation Set

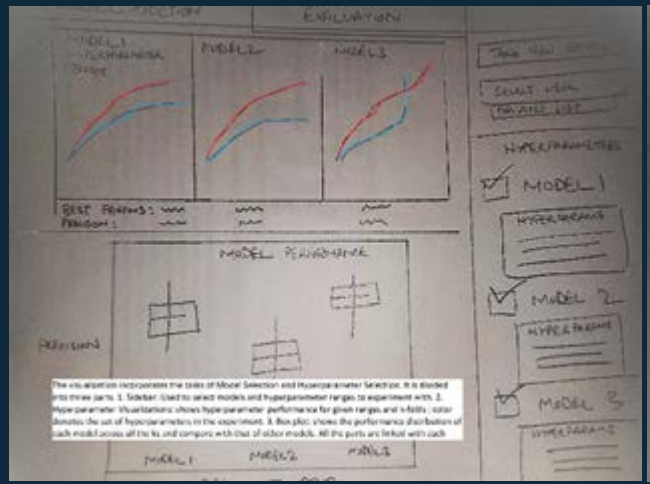
5 Design Process

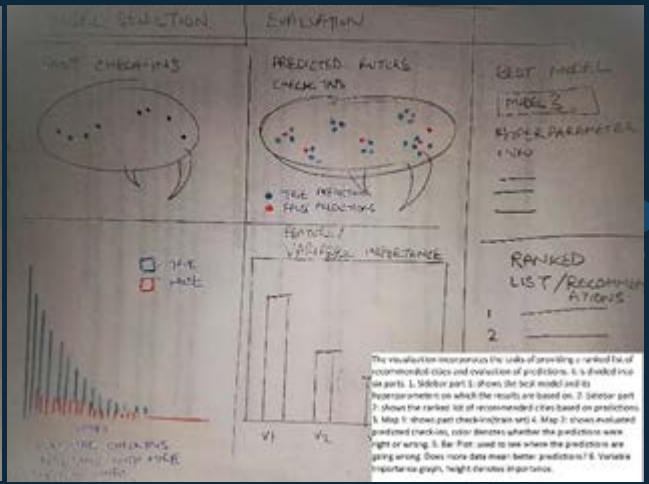
Design Process

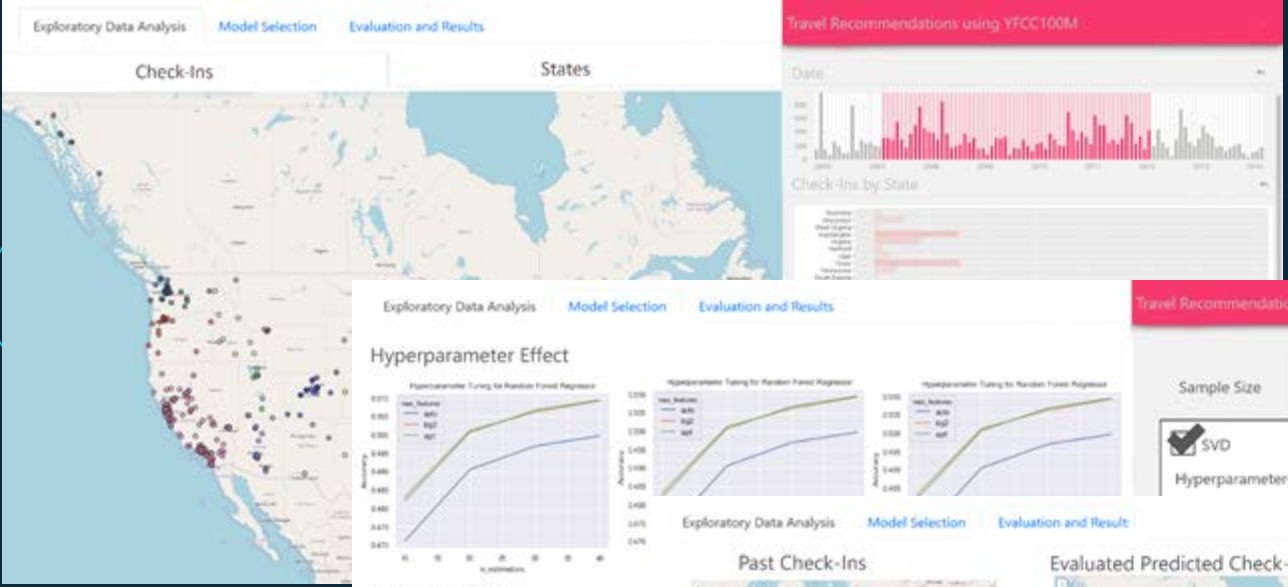



STATE VIEW
DATE
STATE POPULARITY /ATTRACTIVENESS
LINKED & SELECTED


CITY VIEW
STATE: CALIFORNIA
DATE
TOP 10 CITIES
BRUSH

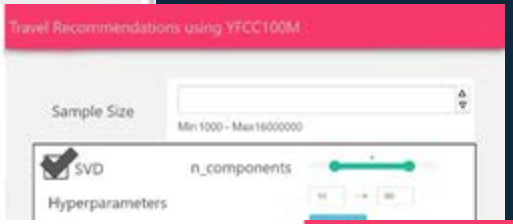

EVALUATION
MODEL 1
MODEL 2
MODEL 3
BEST PARAMS:
MODEL PERFORMANCE
PREDICTIONS
HYPERPARAMETERS
MODEL 1
MODEL 2
MODEL 3

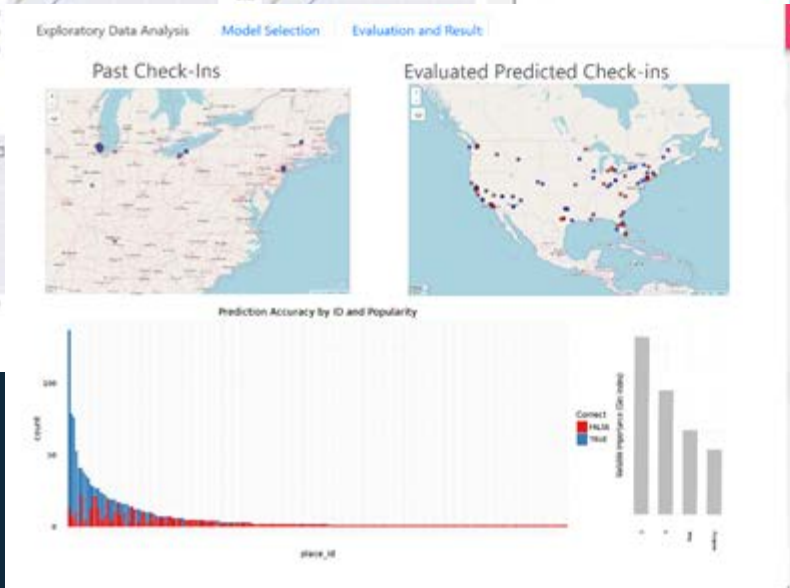

EVALUATION
PREDICTED RATERS
TRUE PREDICTIONS
FALSE PREDICTIONS
BEST MODEL
MODEL 2
HYPERPARAMETER INFO
RANKED LIST/RECOMMENDATIONS


Exploratory Data Analysis
Model Selection
Evaluation and Results
Check-Ins
States

Travel Recommendations using YFCC100M
Date
Check-Ins by State

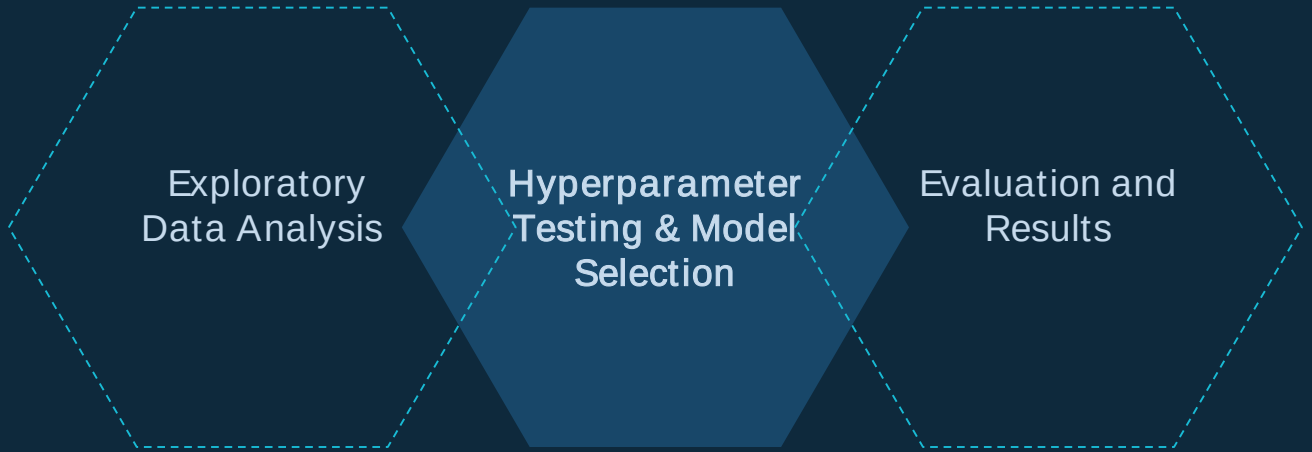
Exploratory Data Analysis
Model Selection
Evaluation and Results
Hyperparameter Effect
Model Selection


Travel Recommendations using YFCC100M
Sample Size
SVD
n_components
Hyperparameters

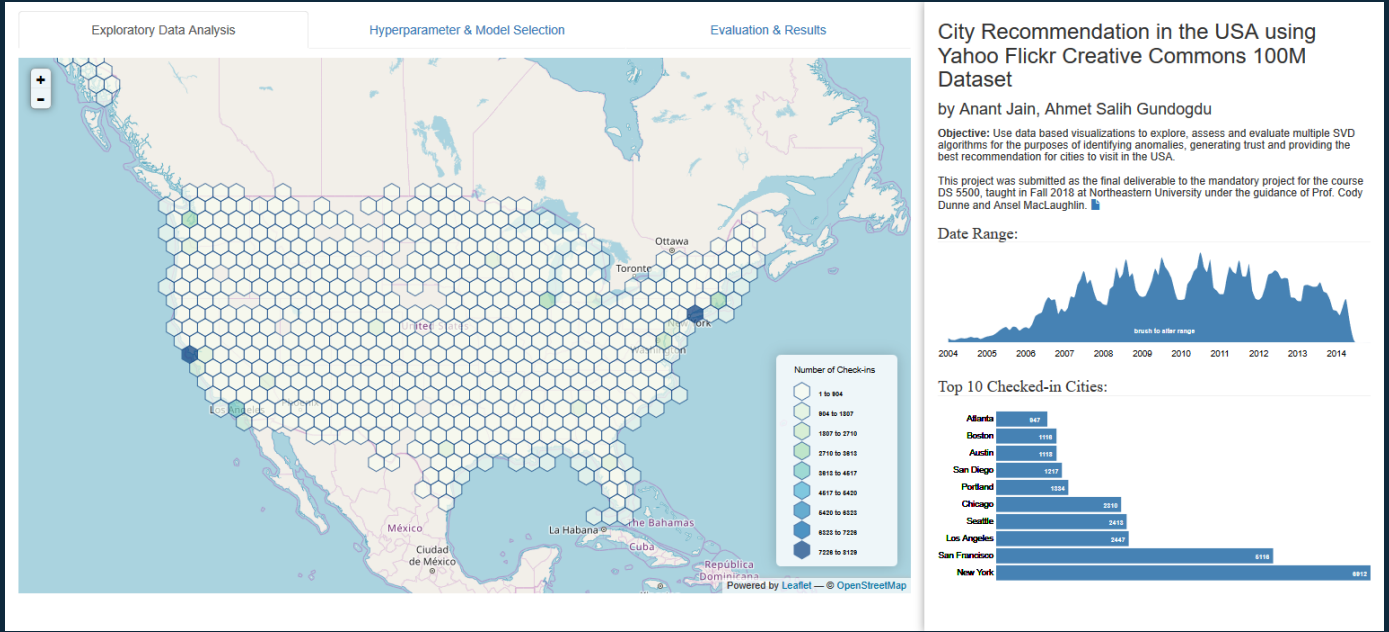

Exploratory Data Analysis
Model Selection
Evaluation and Result
Past Check-Ins
Evaluated Predicted Check-ins
Prediction Accuracy by ID and Popularity


Travel Recommendations using YFCC100M
Select User:
Train Size: 10000
Test Size: 3000
Best Model: SVD++
Hyperparametrs: n_estimators = 10
Ranked List of City Recommendations:
1 San Francisco, CA
2 Maui, Hawaii
3 San Diego, CA

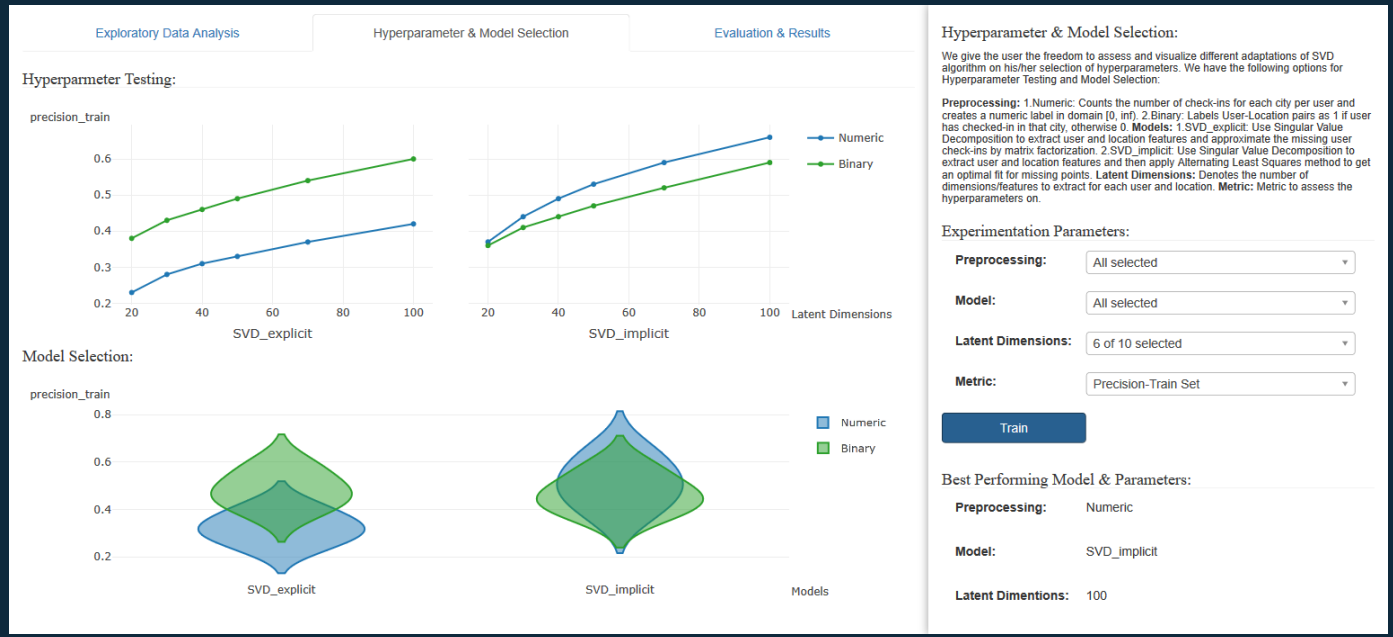
Final Visualization



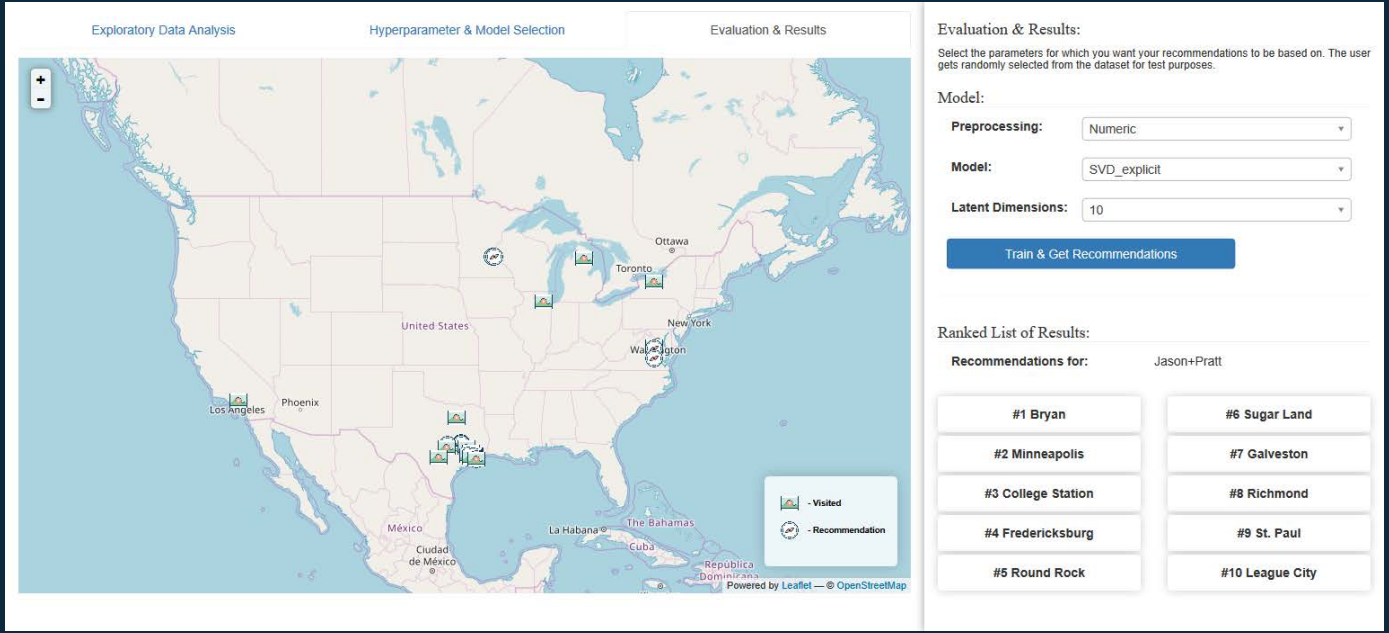
Exploratory Data Analysis



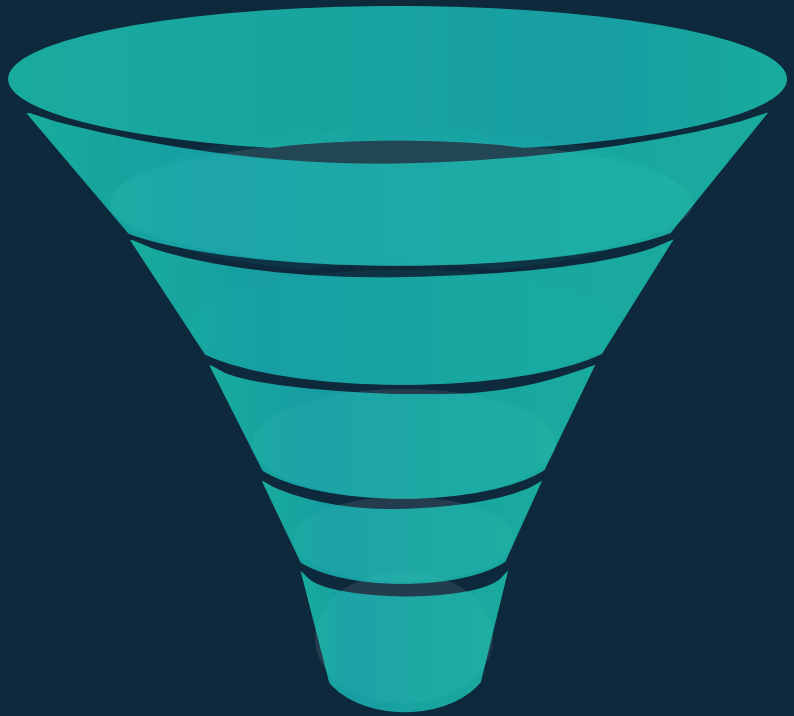
Hyperparameter Testing & Model Selection



Evaluation and Results



Conclusion



Bind ML with Visualizations

Proper Visual Encodings

Include User in the ML tasks

Build Trust in Results

Enjoyment :)

Future Work

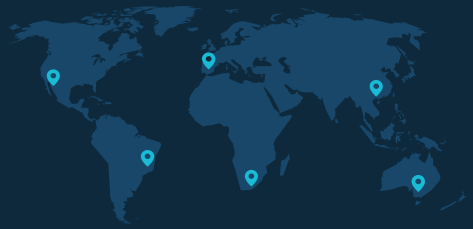
Integration of more complex models

E.x. Autoencoders

Better evaluation techniques

Distance-based, etc.

Scale to cover the whole world instead of just the US



Thanks!

Any questions?

You can find us at:

- https://github.com/antujn
- https://github.com/asgundogdu

GitHub

Github URLs are attached to the icons.

Credits

Special thanks to all the people who made and released these awesome resources for free:

- d3
- leaflet
- colorbrewer
- tipsy
- plotly
- flask
- bootstrap
- multi-select
- pylab
- implicit