

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

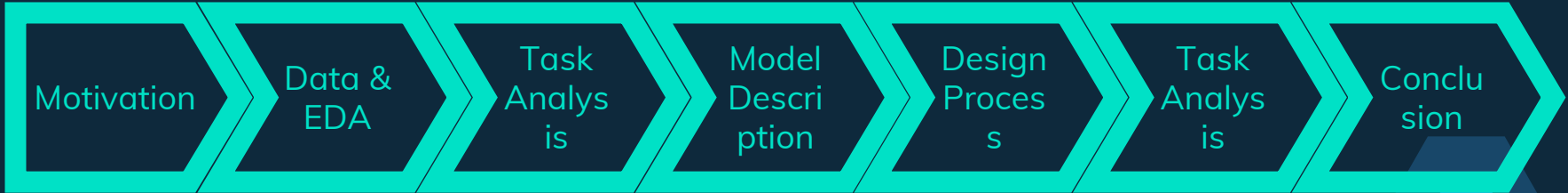




Layout



Final
Visuali
zation



Motivation

Data &
EDA


Task
Analysis

Model
Description

Design
Processes

Task
Analysis

Conclu
sion



A decorative pattern of hexagons in various shades of blue and cyan on the left side of the slide. Some hexagons contain icons: a lightbulb, a thumbs up, a smartphone, a magnifying glass, and a gear. A network diagram with a central node and five peripheral nodes is also visible.

1

Motivation

Let's start with the first set of slides





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.



A decorative pattern of hexagons in various shades of blue and cyan. Some hexagons contain icons: a lightbulb, a thumbs up, a network node, a smartphone, a magnifying glass, a gear, and a speech bubble. The pattern is arranged in a cluster on the left side of the slide.

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

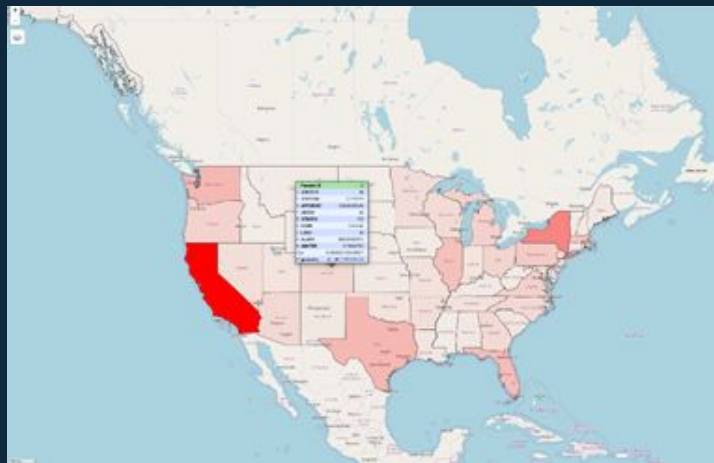
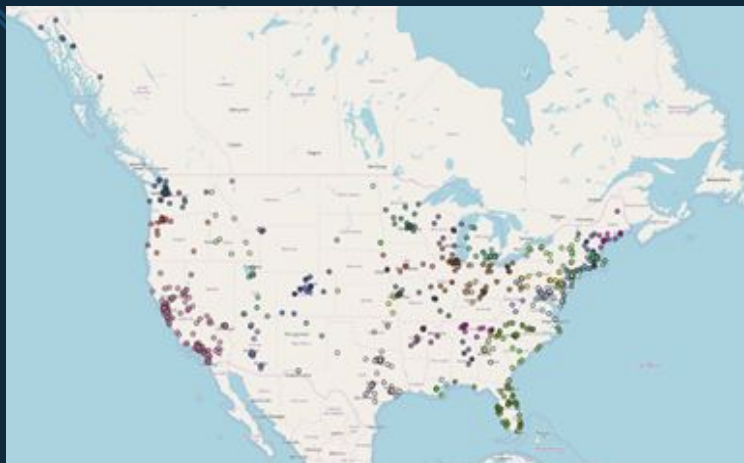


Columns

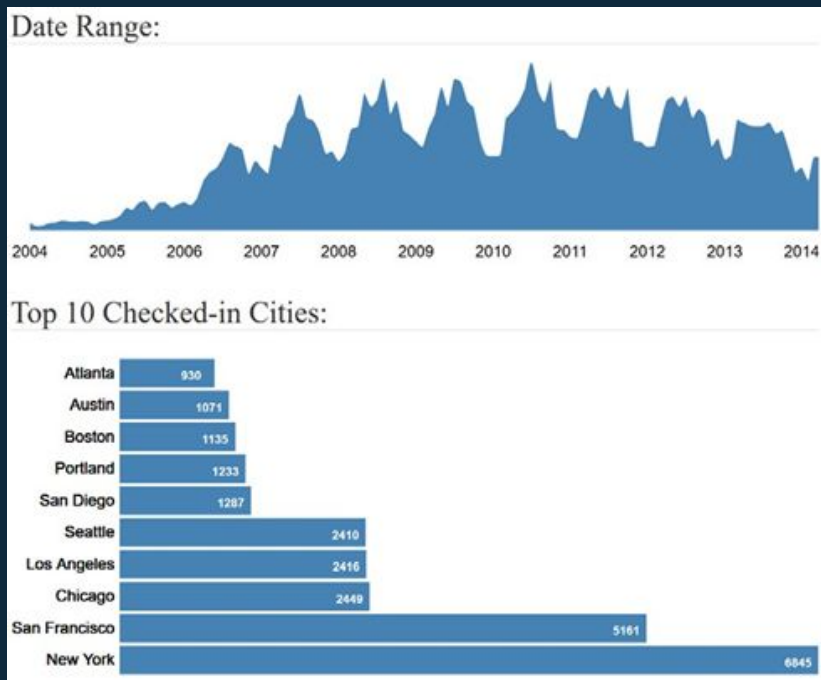
pid	Unique media identifier
user_nickname	User identifier
date	Date the media object was created
longitude	Longitude of the location the media object was checked at
latitude	Latitude of the location the media object was checked at
town	Town the media object was checked in
state	State the media object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



EDA



EDA (contd.)





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



A decorative graphic on the left side of the slide. It features a large, light blue hexagon in the center. Surrounding it are several smaller hexagons in various shades of blue and teal. Some of these smaller hexagons contain white icons: a lightbulb, a thumbs-up, a smartphone, a magnifying glass, and a gear. There is also a network-like icon with a central node and several smaller nodes connected by lines.


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



A decorative graphic on the left side of the slide. It features a large, light blue hexagon in the center. Surrounding it are several smaller hexagons in various shades of blue and teal. Some of these smaller hexagons contain white icons: a lightbulb, a thumbs-up, a smartphone, a magnifying glass, and a gear. There is also a network-like icon with a central node and several smaller nodes connected by lines.

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


A decorative graphic on the left side of the slide. It features a large cyan hexagon with the number '5' inside. Surrounding this central hexagon are several smaller hexagons and icons in various shades of blue and cyan. The icons include a lightbulb, a thumbs-up, a network of nodes, a smartphone, a magnifying glass, a gear, and a speech bubble.

5

Design Process



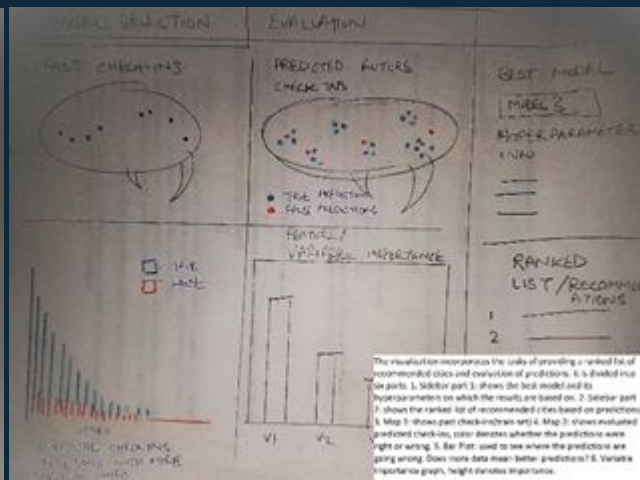
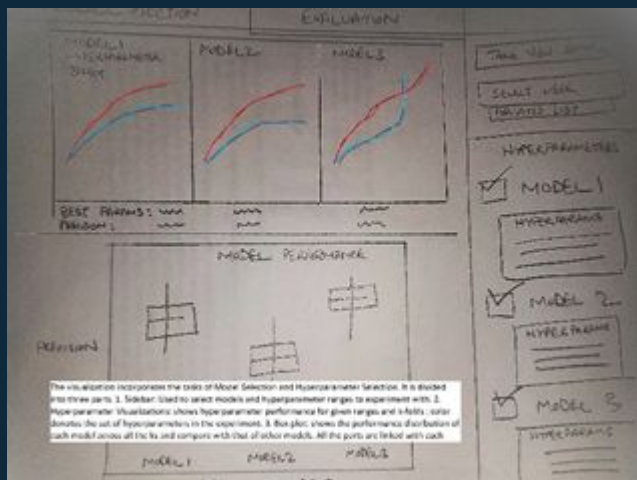
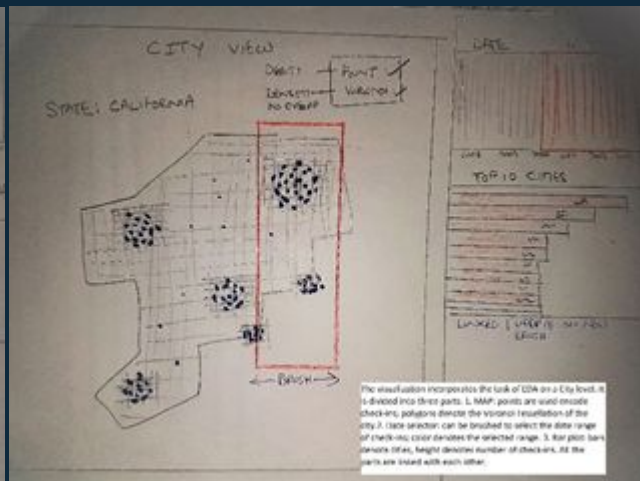
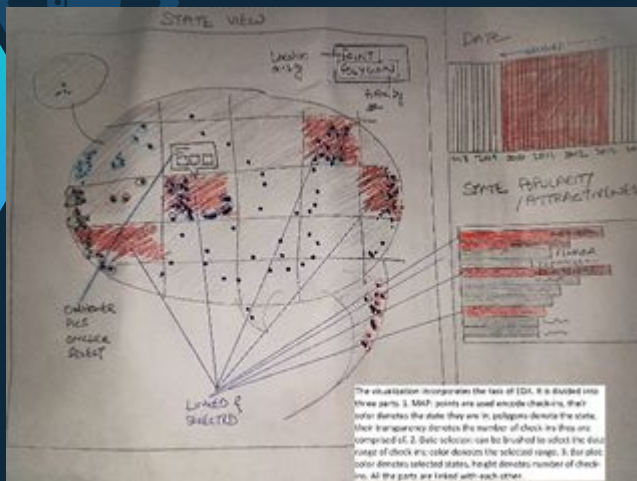
Design Process



Preliminary
Sketches

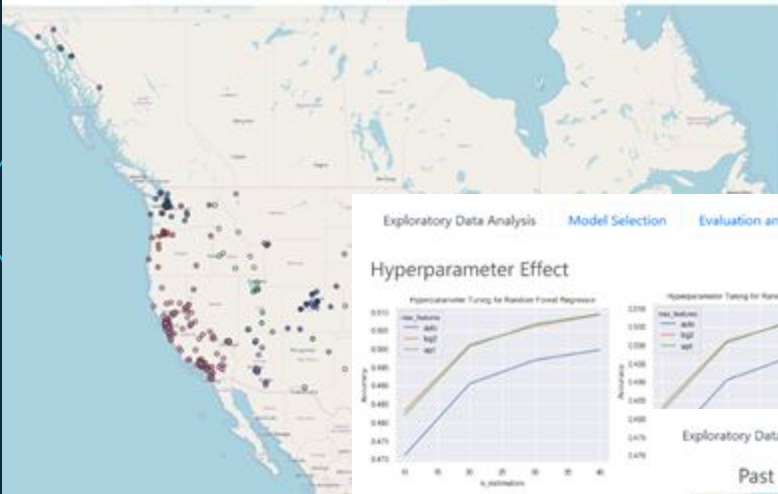
Digital
Sketches

Final
Visualization



Check-Ins

States



Travel Recommendations using YFCC100M



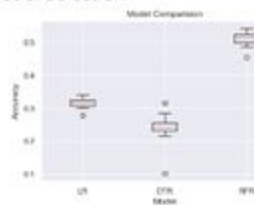
Exploratory Data Analysis Model Selection Evaluation and Results

Hyperparameter Effect



Exploratory Data Analysis Model Selection Evaluation and Results

Model Selection



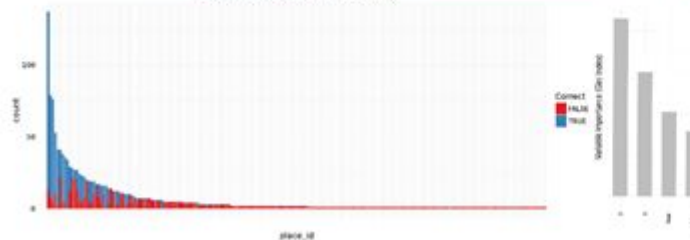
Past Check-Ins



Evaluated Predicted Check-ins



Prediction Accuracy by ID and Popularity



Travel Recommendations using YFCC100M

Sample Size

Min 1000 - Max 1600000



Hyperparameters

n_components



Travel Recommendations using YFCC100M

Select User:

Current System: California

- California ☒
- Florida ☒
- Illinois ☒
- Michigan ☒
- Ohio ☒
- Texas ☒
- Washington ☒

Train Size: 10000

Test Size: 3000

Best Model: SVD++

Hyperparameters: n_estimators = 10

Ranked List of City Recommendations:

- 1 San Francisco, CA
- 2 Maui, Hawaii
- 3 San Diego, CA



Final Visualization

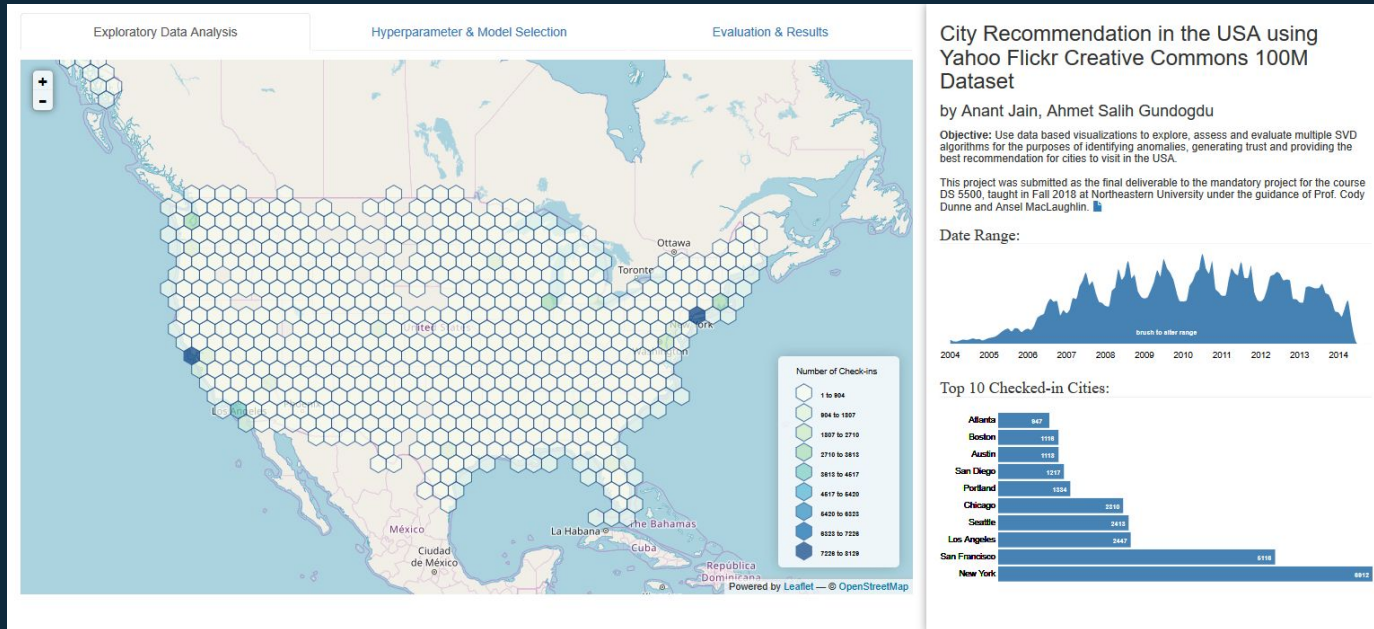
Video Walkthrough

Exploratory
Data Analysis

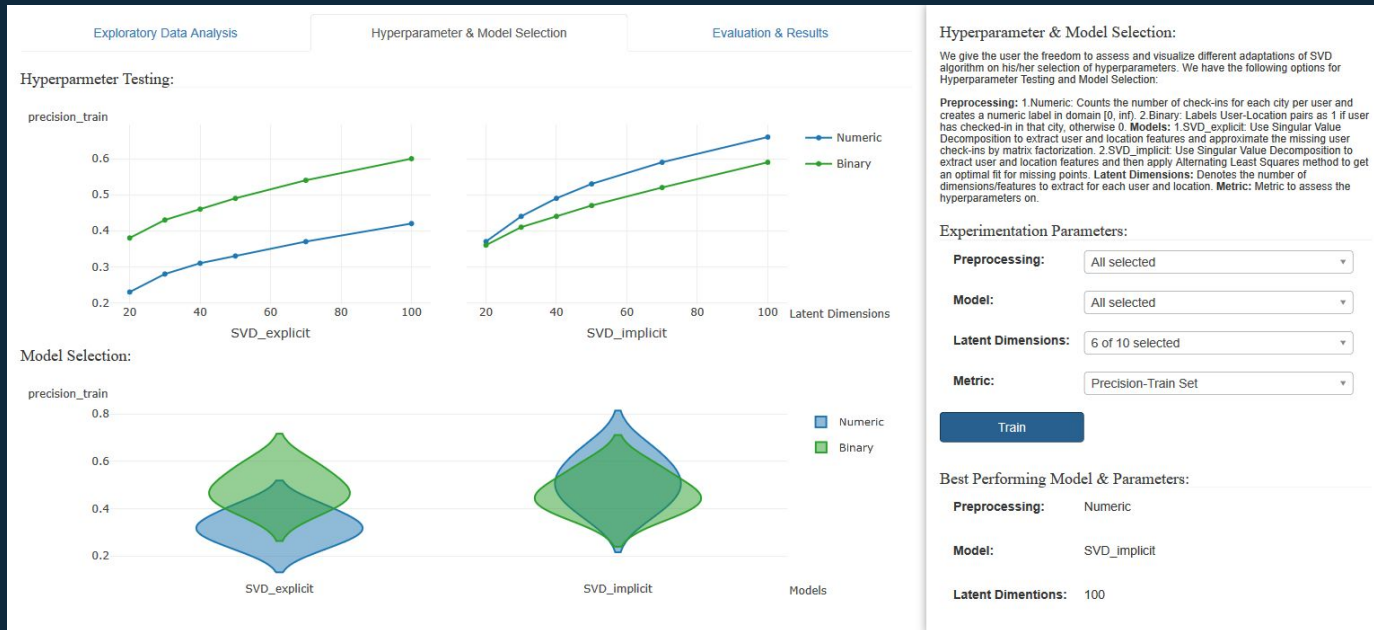
**Hyperparameter
Testing & Model
Selection**

Evaluation and
Results

Exploratory Data Analysis



Hyperparameter Testing & Model Selection



Evaluation and Results

Exploratory Data Analysis

Hyperparameter & Model Selection

Evaluation & Results

Evaluation & Results:

Select the parameters for which you want your recommendations to be based on. The user gets randomly selected from the dataset for test purposes.

Model:

Preprocessing:

Model:

Latent Dimensions:

Train & Get Recommendations

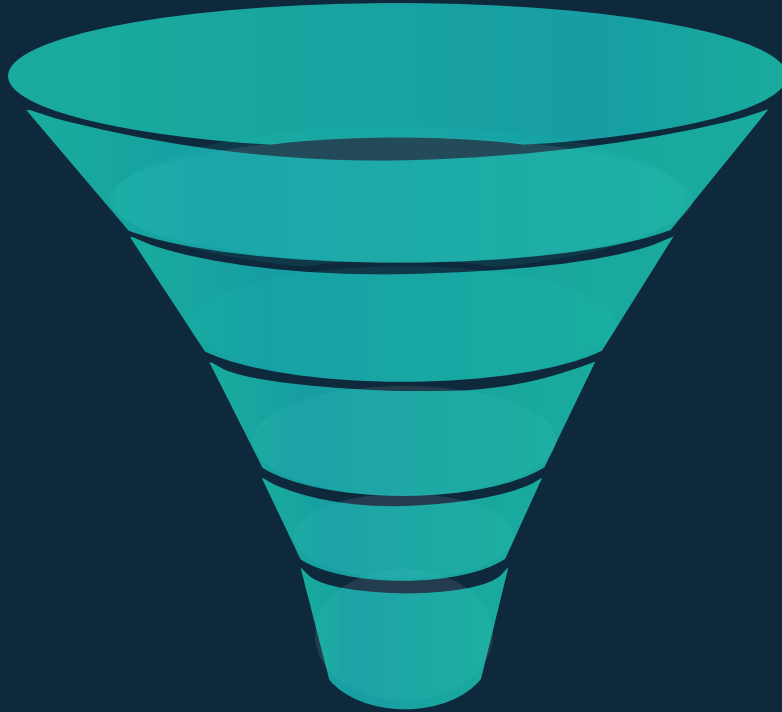
Ranked List of Results:

Recommendations for: Jason+Pratt

#1 Bryan	#6 Sugar Land
#2 Minneapolis	#7 Galveston
#3 College Station	#8 Richmond
#4 Fredericksburg	#9 St. Paul
#5 Round Rock	#10 League City



Conclusion



Bind ML with Visualizations

Proper Visual Encodings

Include User in the ML tasks

Build Trust in Results

Enjoyment :)





Future Work

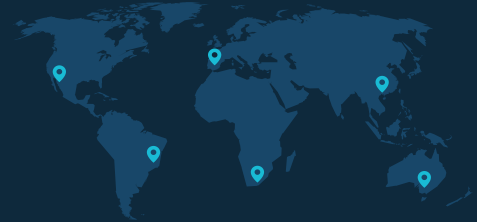
Integration of more complex models

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 URLs are attached to the icons.

Credits

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

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

