

Mining Hackathon Projects data

- Siddharth Patki, Kanishk Tripathi
December 11, 2015

What is a Hackathon?

- Hack + Marathon = Hackathon



Who conducts?



Idea



- Analyze what separates winning projects from others.
- Fit a classifier to predict for future entries.
- Find distribution of participants across the country

Submission Hub



Project Details

Open Channels

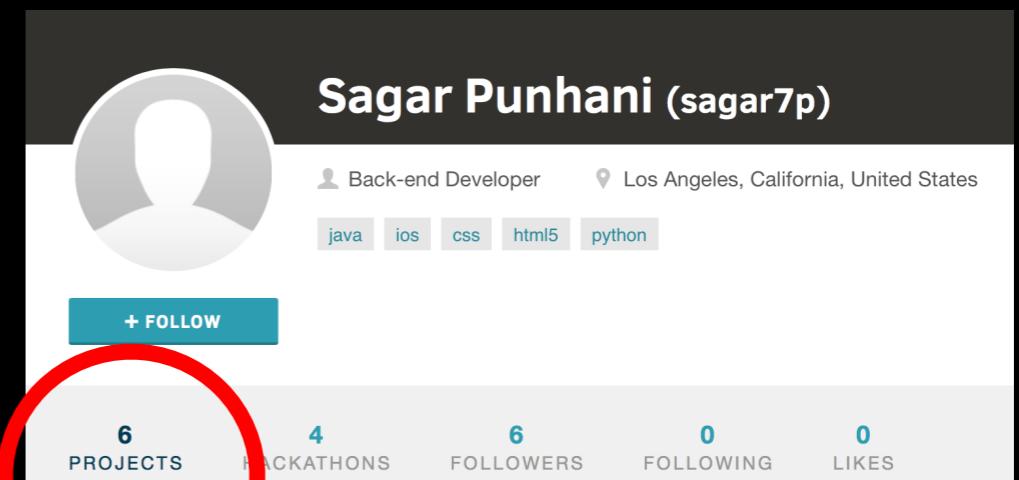
A revolution in visualization

Built With

c-sharp amazon-web-services unity myo android-studio vuforia augmented-reality social-action

data-visualization security game hardware

Participant Profiles



Sagar Punhani ([sagar7p](#))

Back-end Developer | Los Angeles, California, United States

java ios css html5 python

+ FOLLOW

6 PROJECTS 4 HACKATHONS 6 FOLLOWERS 0 FOLLOWING 0 LIKES

A participant profile card for Sagar Punhani. It includes a placeholder profile picture, the name "Sagar Punhani" with the handle "(sagar7p)", a title "Back-end Developer", a location "Los Angeles, California, United States", and skill tags "java", "ios", "css", "html5", and "python". A blue "+ FOLLOW" button is present. Below the bio, there are statistics: "6 PROJECTS" (circled in red), "4 HACKATHONS", "6 FOLLOWERS", "0 FOLLOWING", and "0 LIKES".

Basic Preprocessing

- Remove the data noise.
1% tags => 95% projects
Filter out noisy tags

Approximately
30,000 Projects

10,200 Projects

1,100 Winners

taylor-swift

coffee-oh-god-so-much-coffee
the-understanding-that-life-is-ultimately-pointless

Data Preprocessing

- Remove the entries not containing tags.
- Calculate the histogram of tags and remove the ones with frequency less than 50

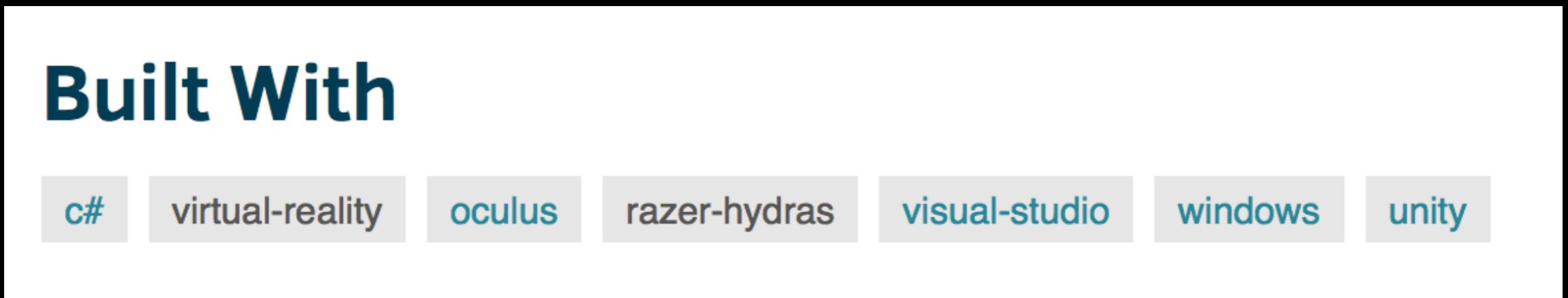


- Bin the numeric attributes (member count and experience)



Feature Vector

1. Tags



2. Number of team members



3. Team Experience



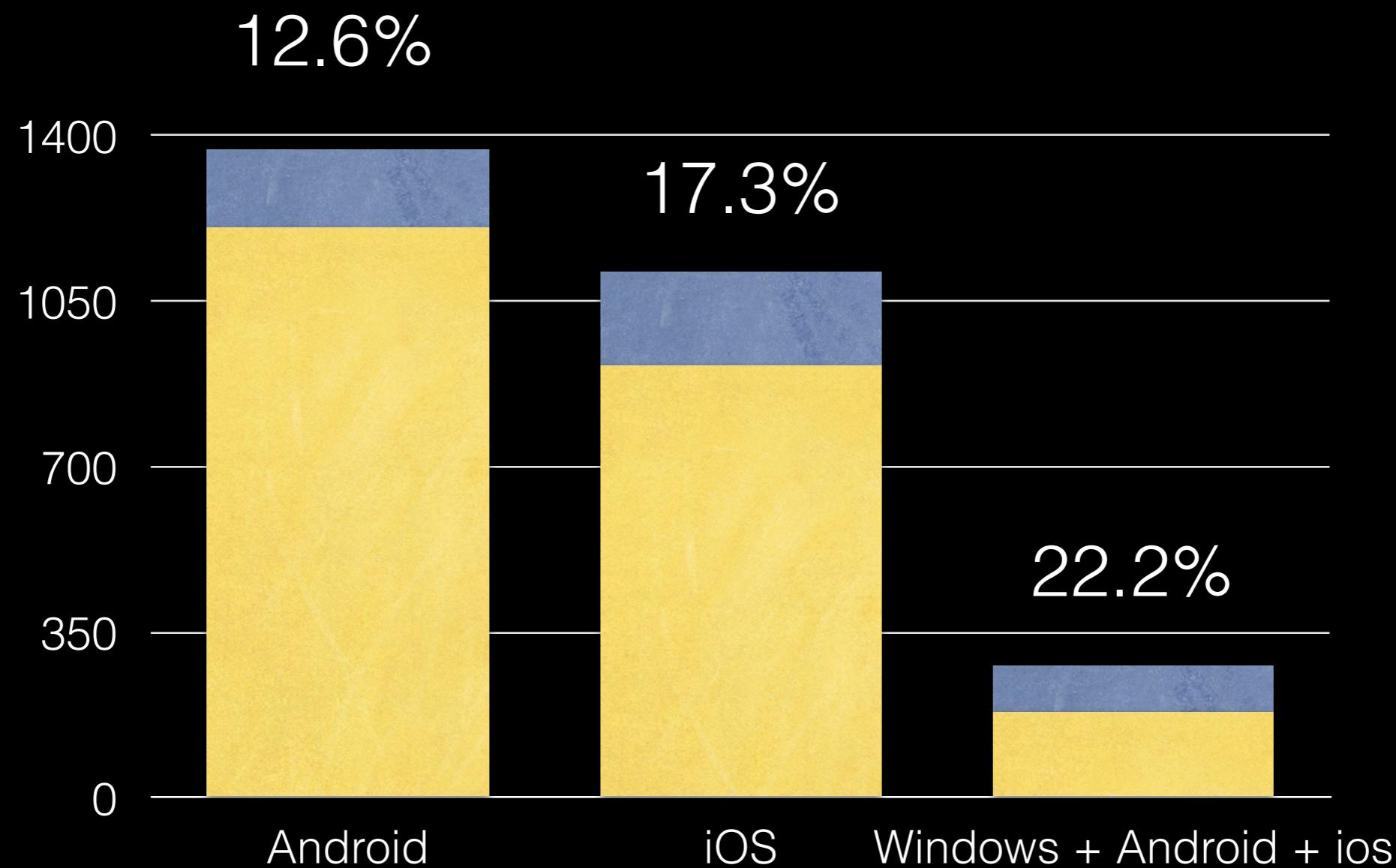


Overall Project tags

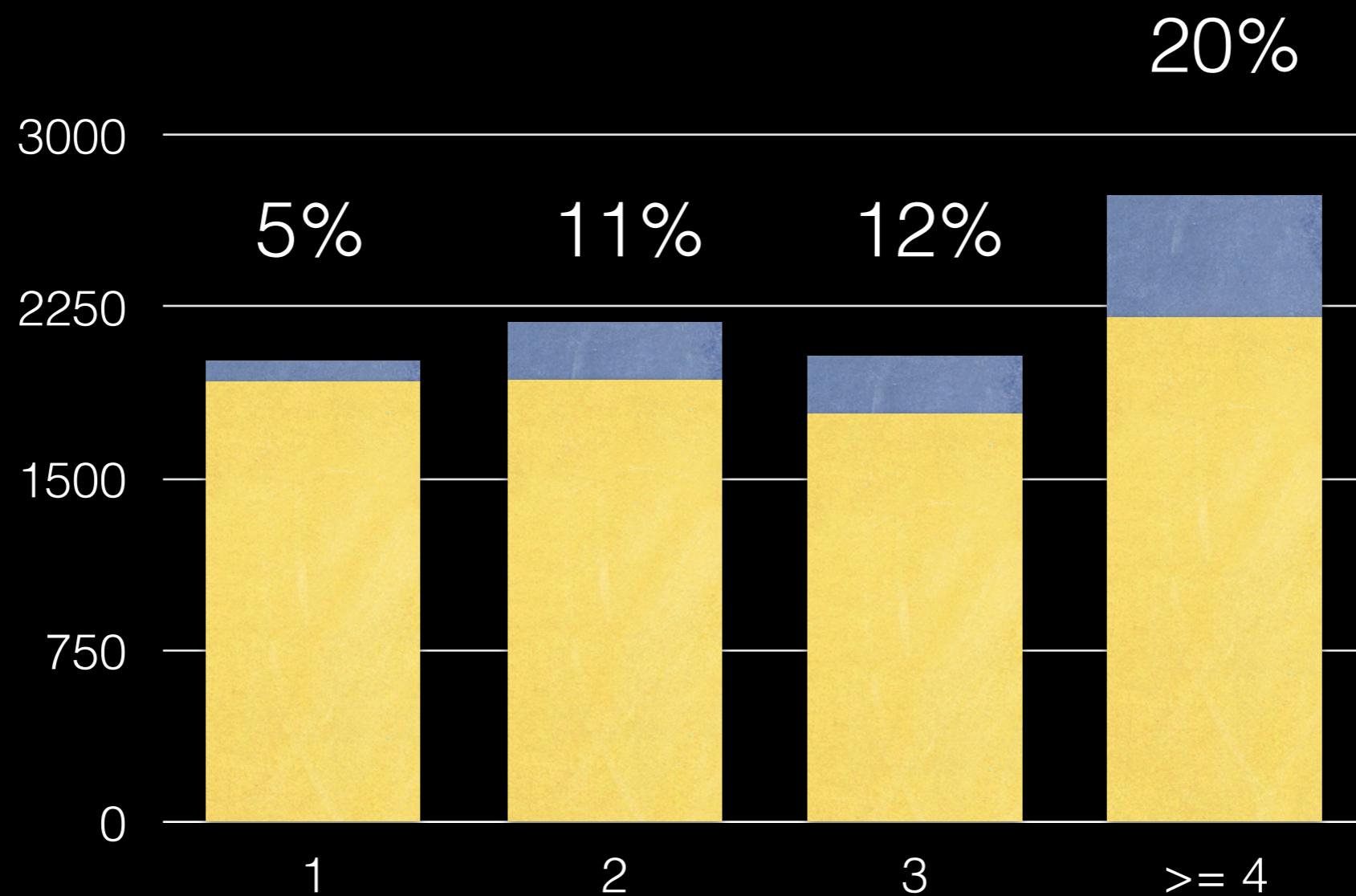


Winning Projects tags

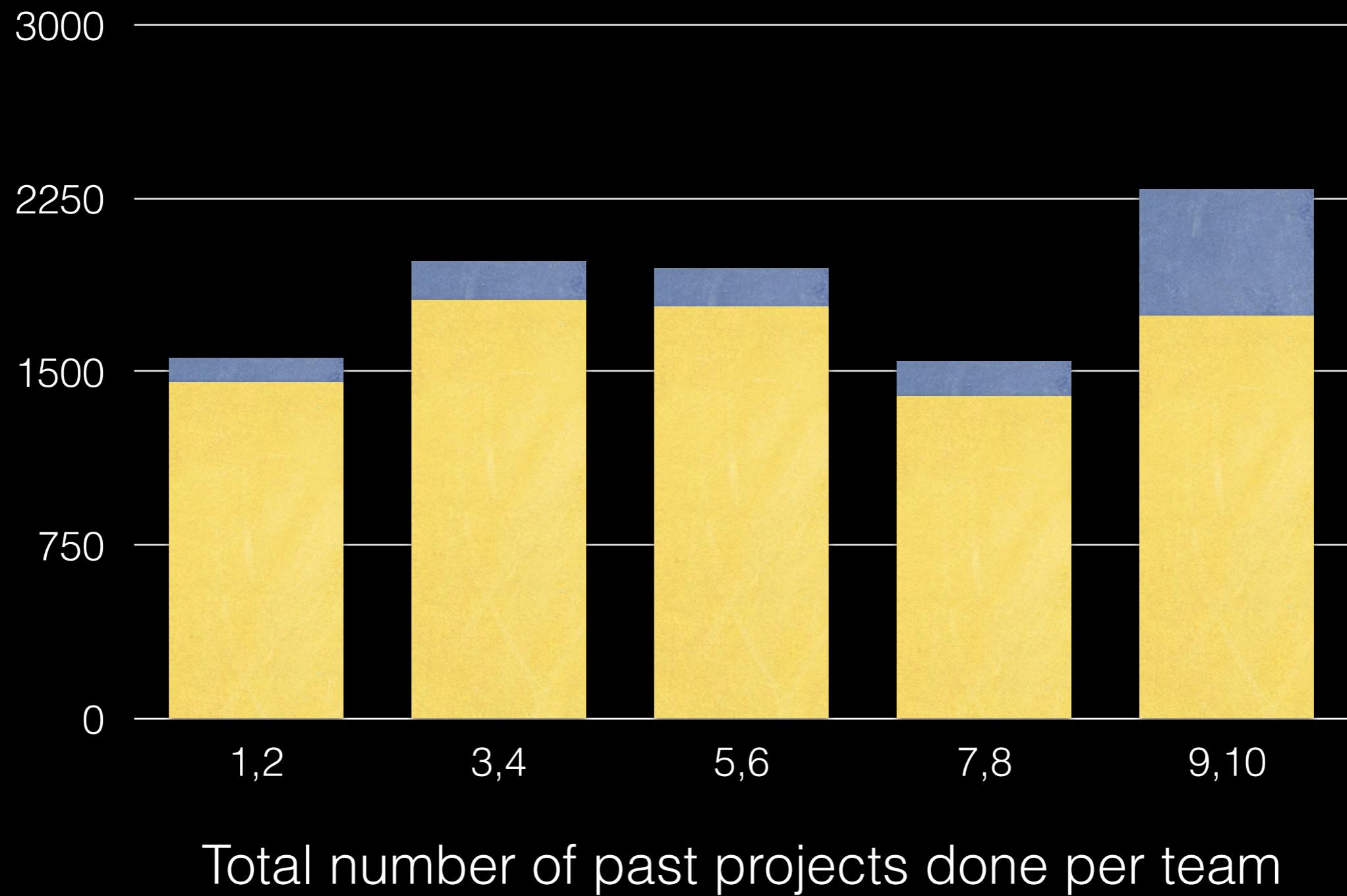
Trends - mobile app development



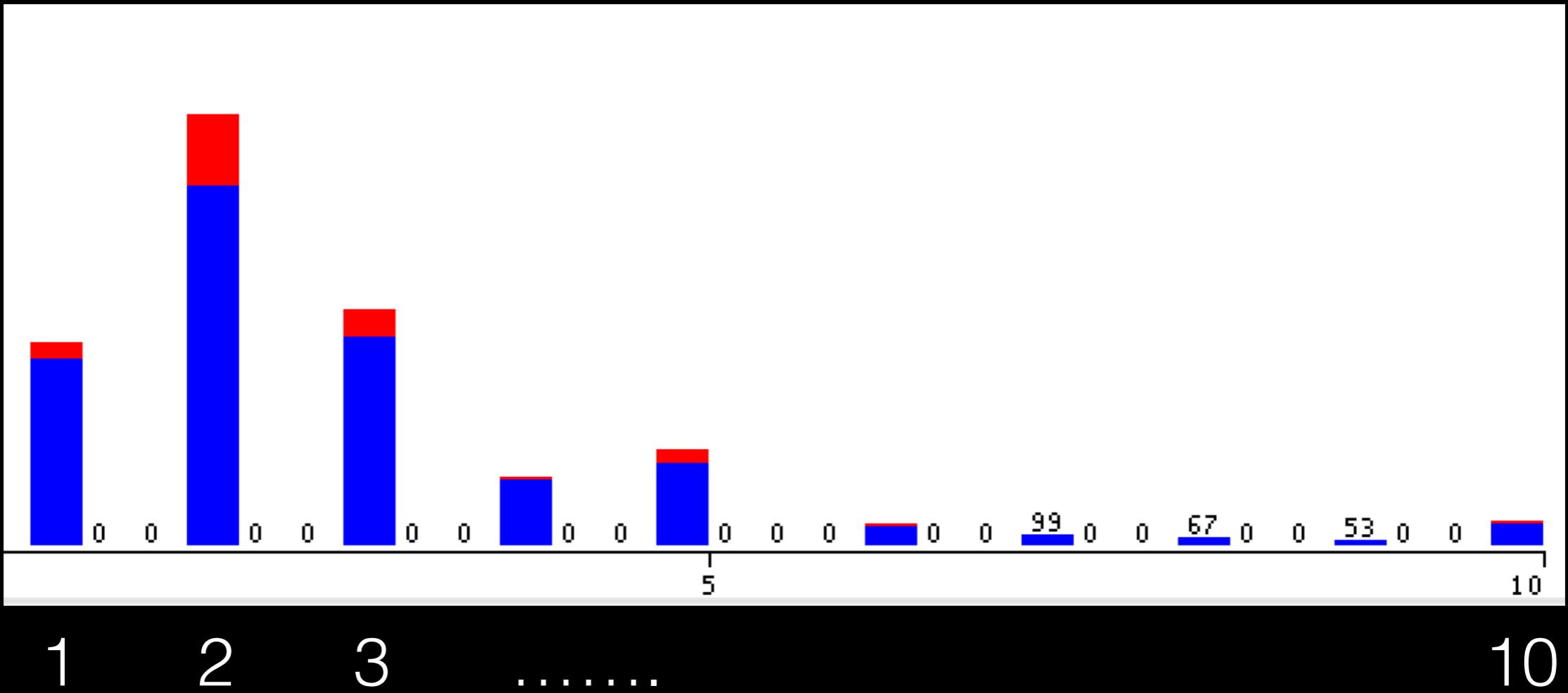
Trends - team members



Trends - team experience



Trends - avg. experience per member

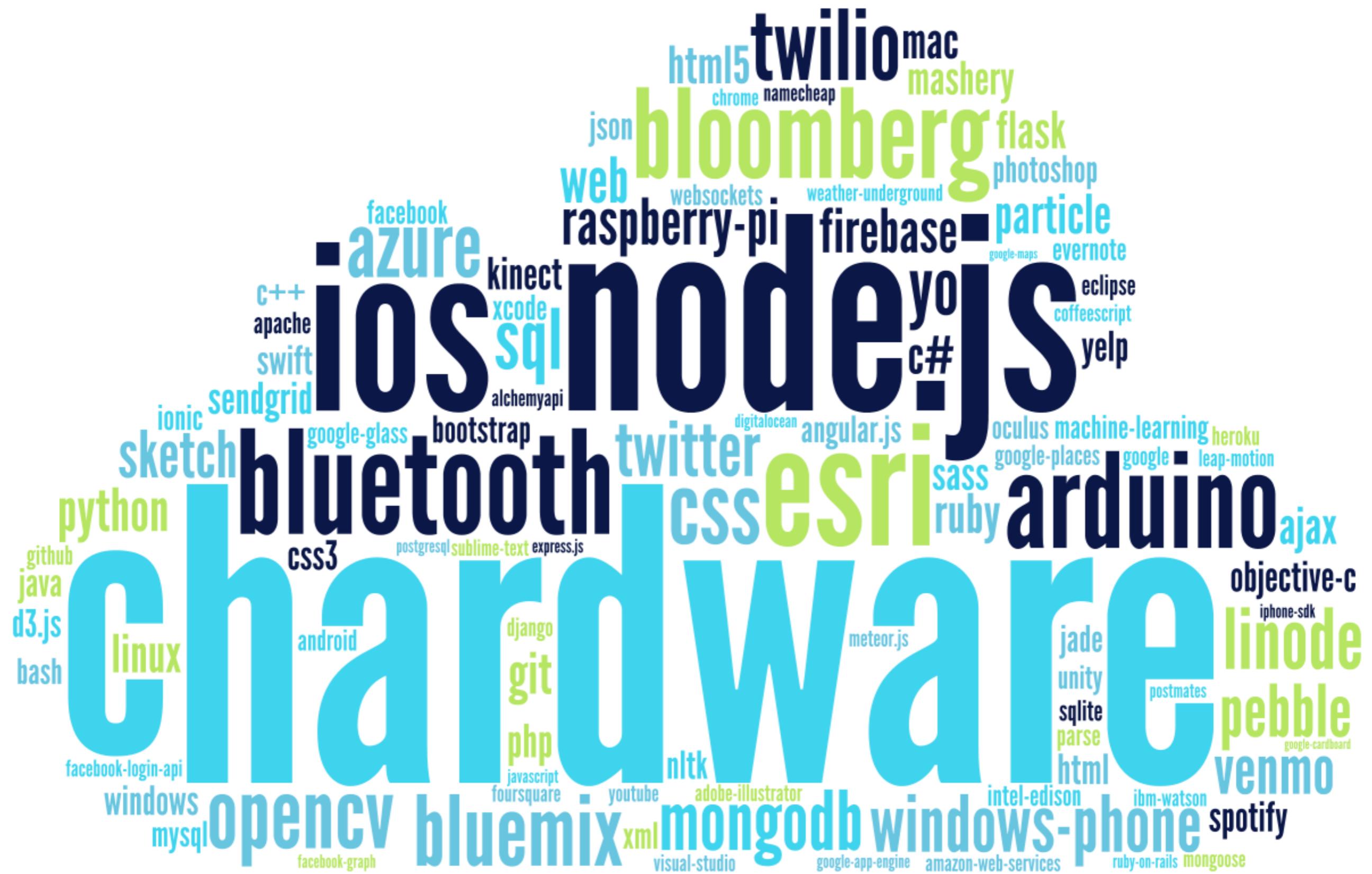


Frequent does not mean discriminative

Information Gain

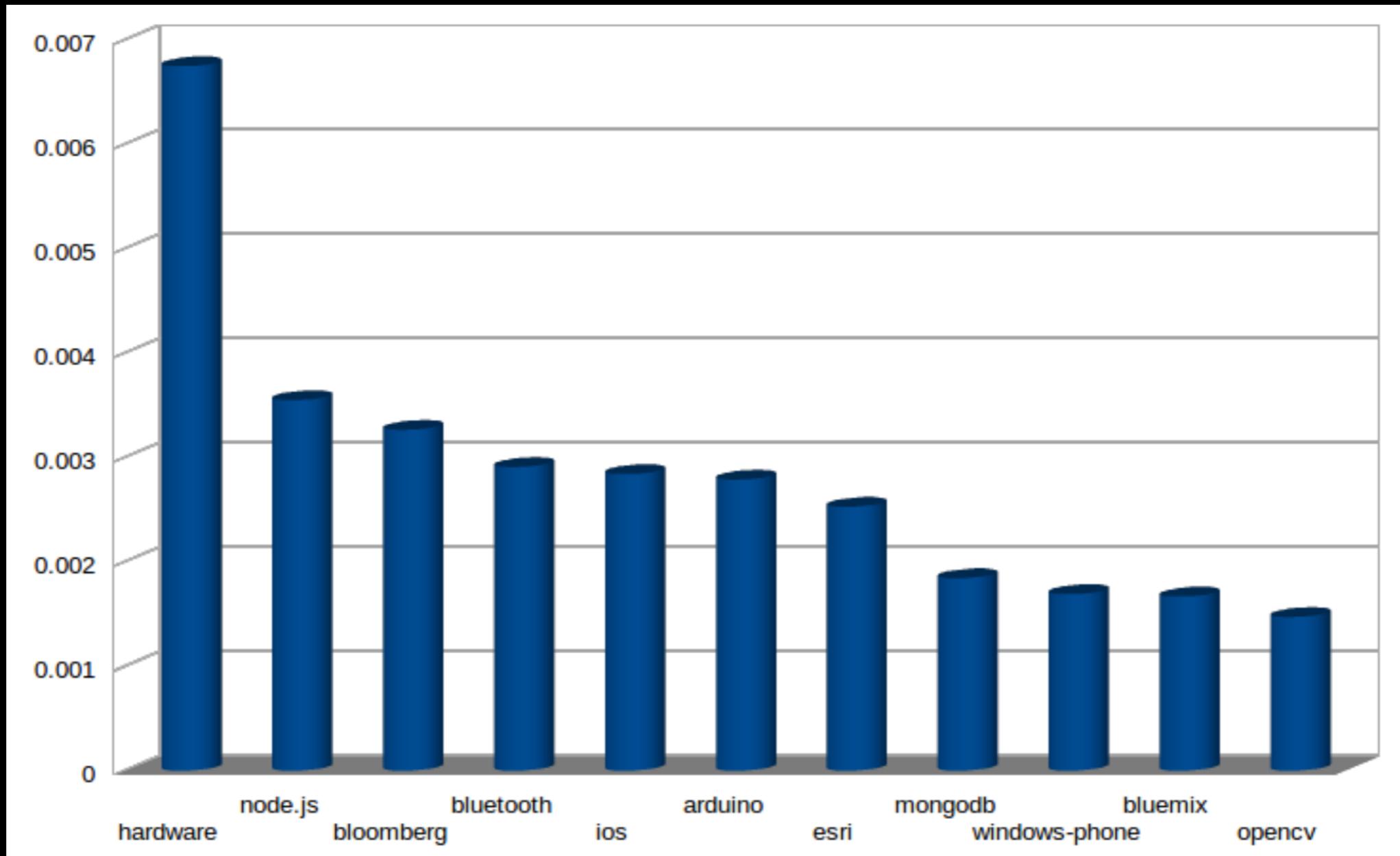
Rank tags according to their **information gain**

$$IG(T, a) = H(T) - H(T|a)$$



Information Gain

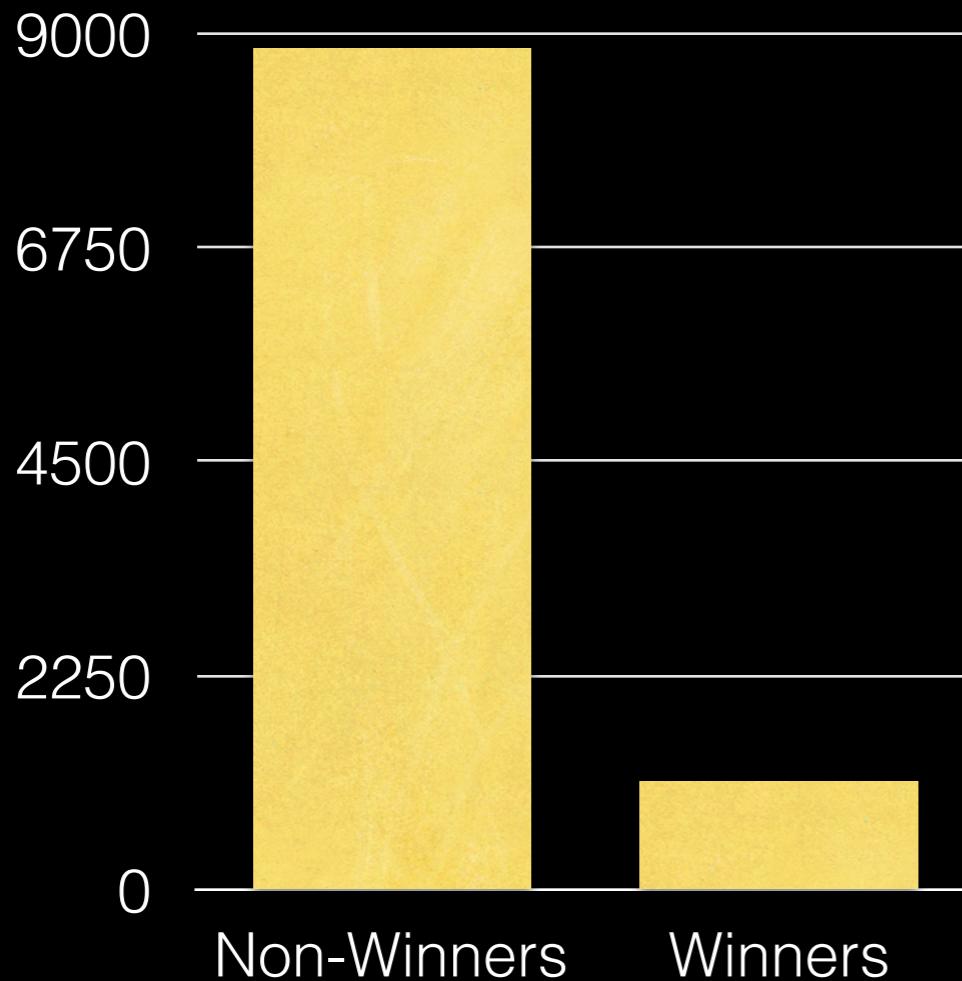
Discriminative Tags - continued



Separates two classes better

Classification

- Class imbalance Problem:
 - SMOTE
 - Synthetic minority oversampling Technique
 - New instances developed with reference to k-nearest neighbors



10 - fold cross validation

Naive Bayes

a	b	<-- classified as
7459	387	a = False
903	235	b = True

38% Precision, TPR = 20.6%

C 4.5

a	b	<-- classified as
7671	175	a = False
535	603	b = True

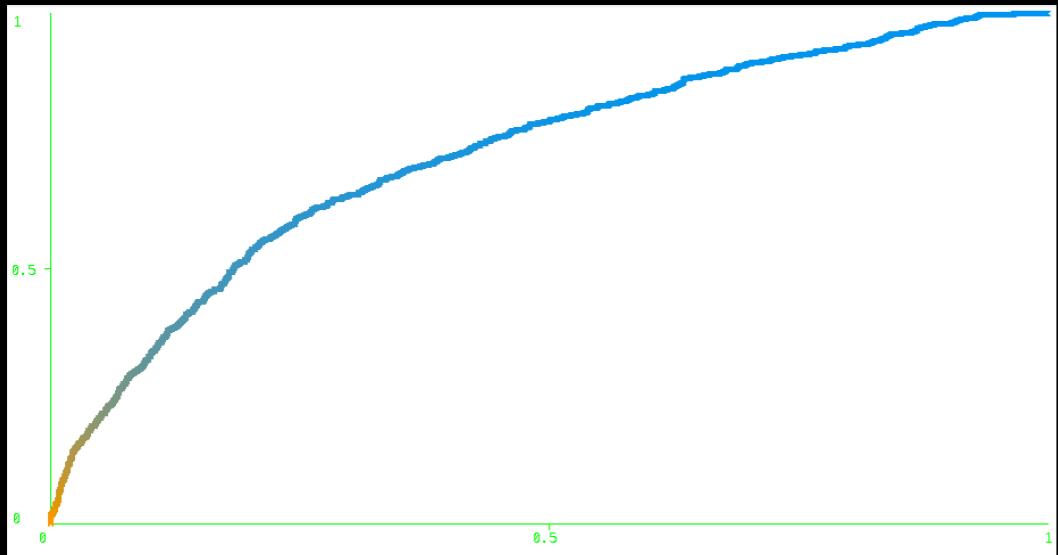
77.5% Precision, TPR = 53%

Random Forest

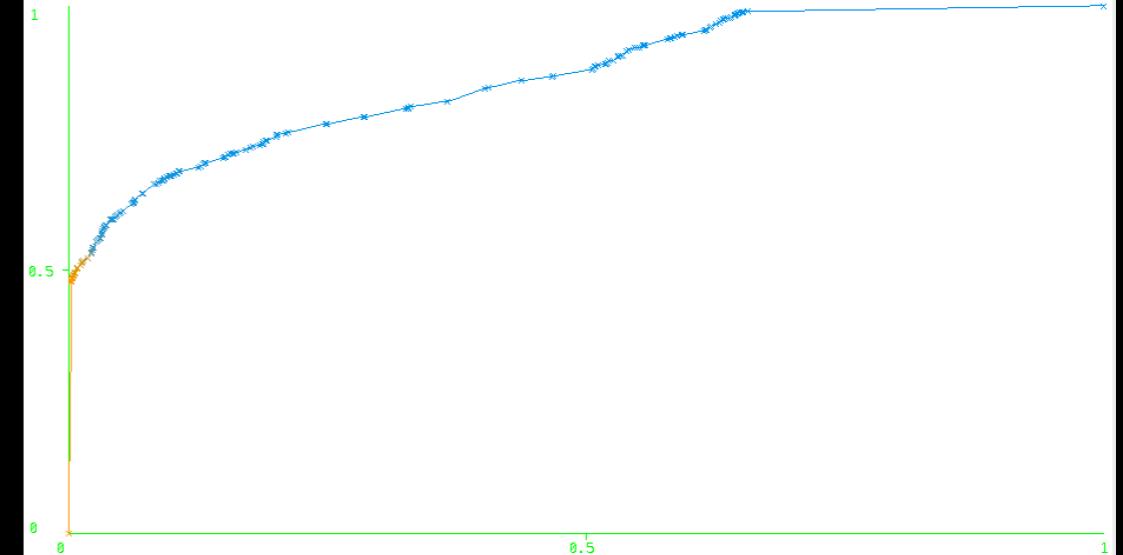
a	b	<-- classified as
7684	162	a = False
67	1071	b = True

86.8% Precision, TPR = 94%

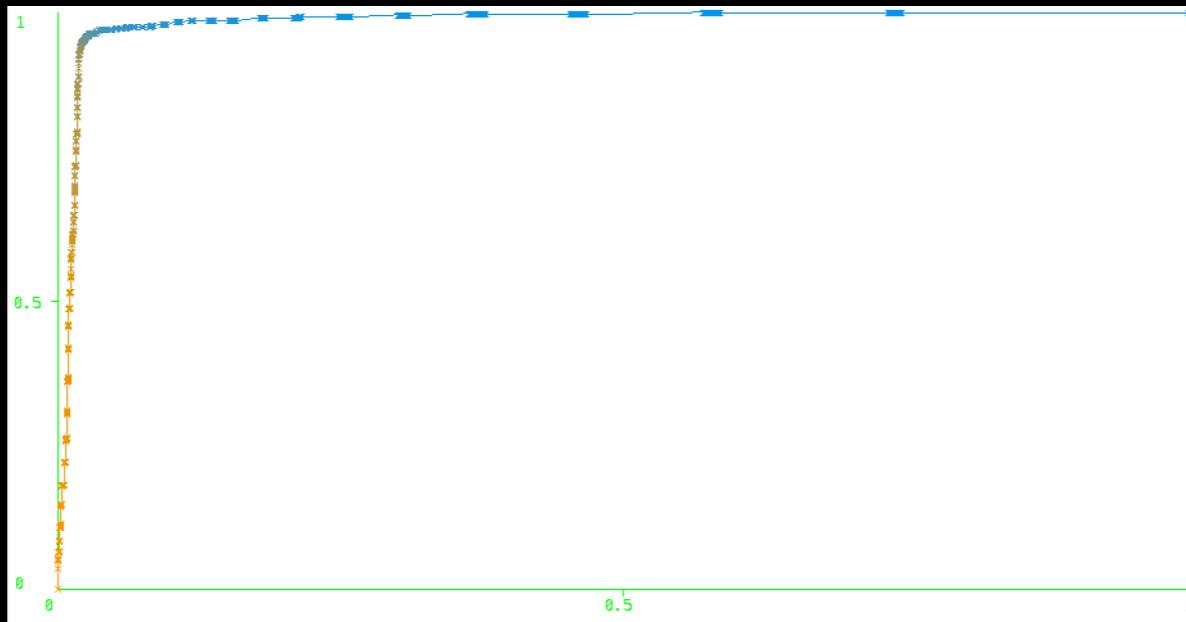
ROC



Naive Bayes

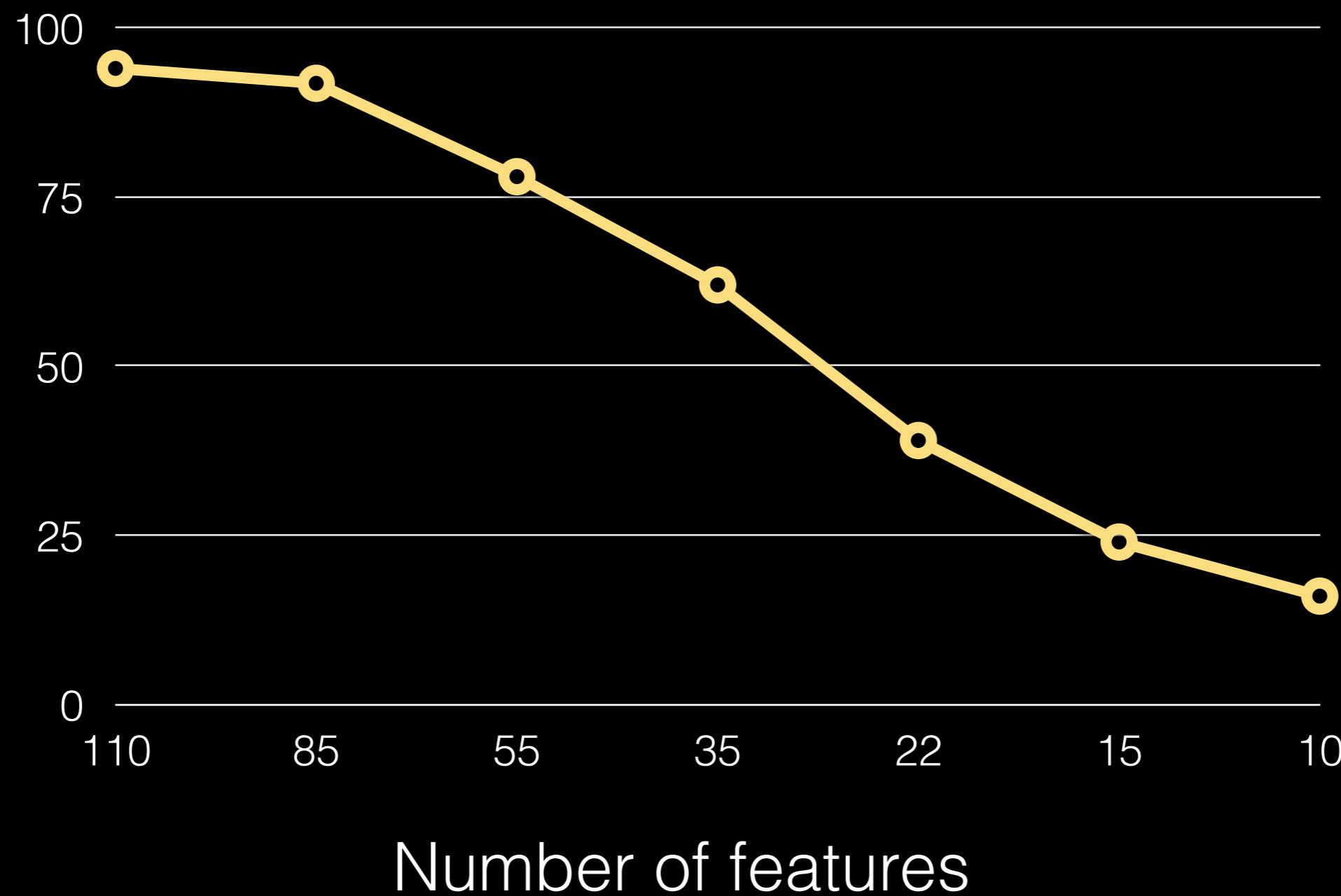


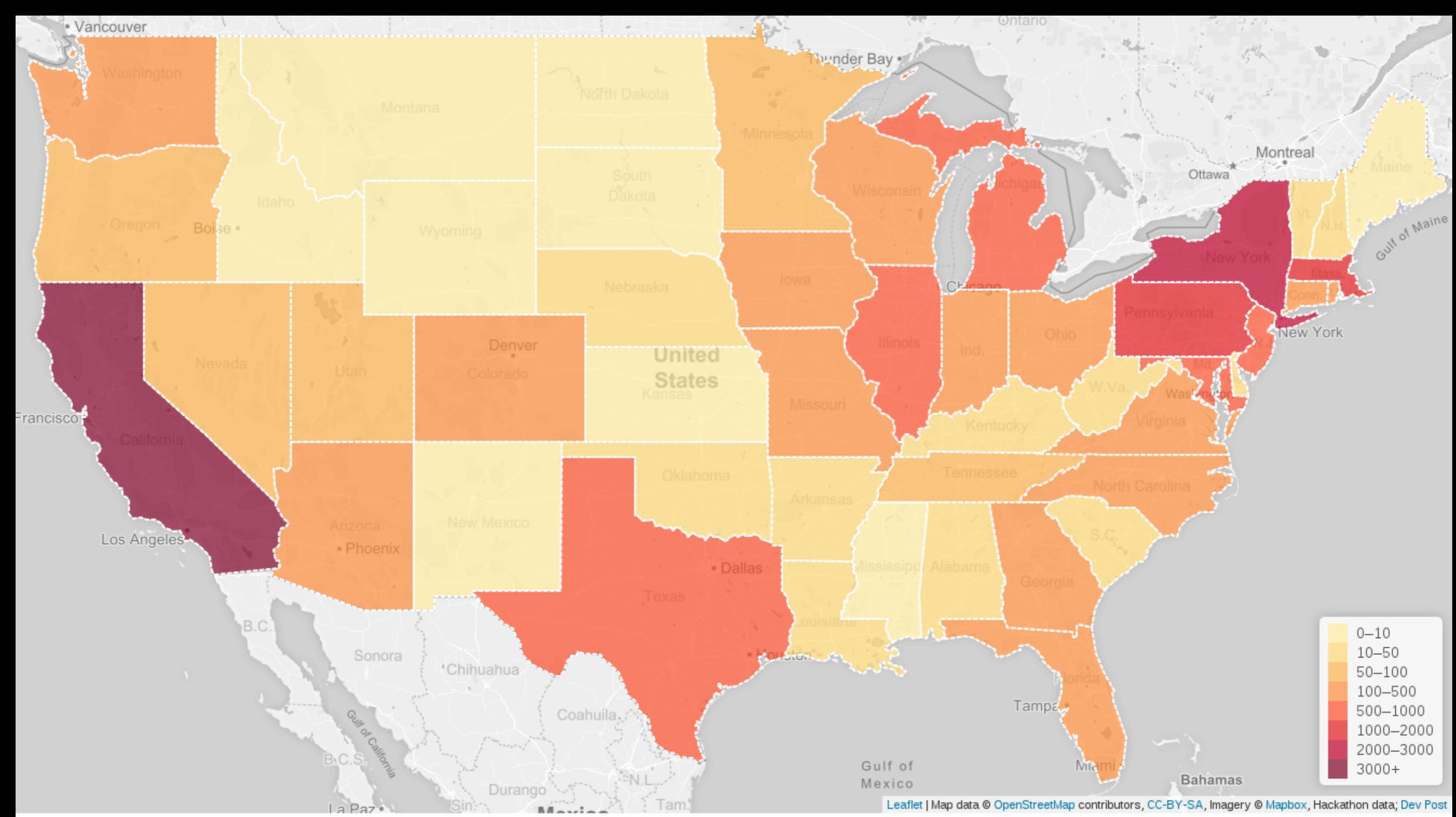
C 4.5



Random Forest

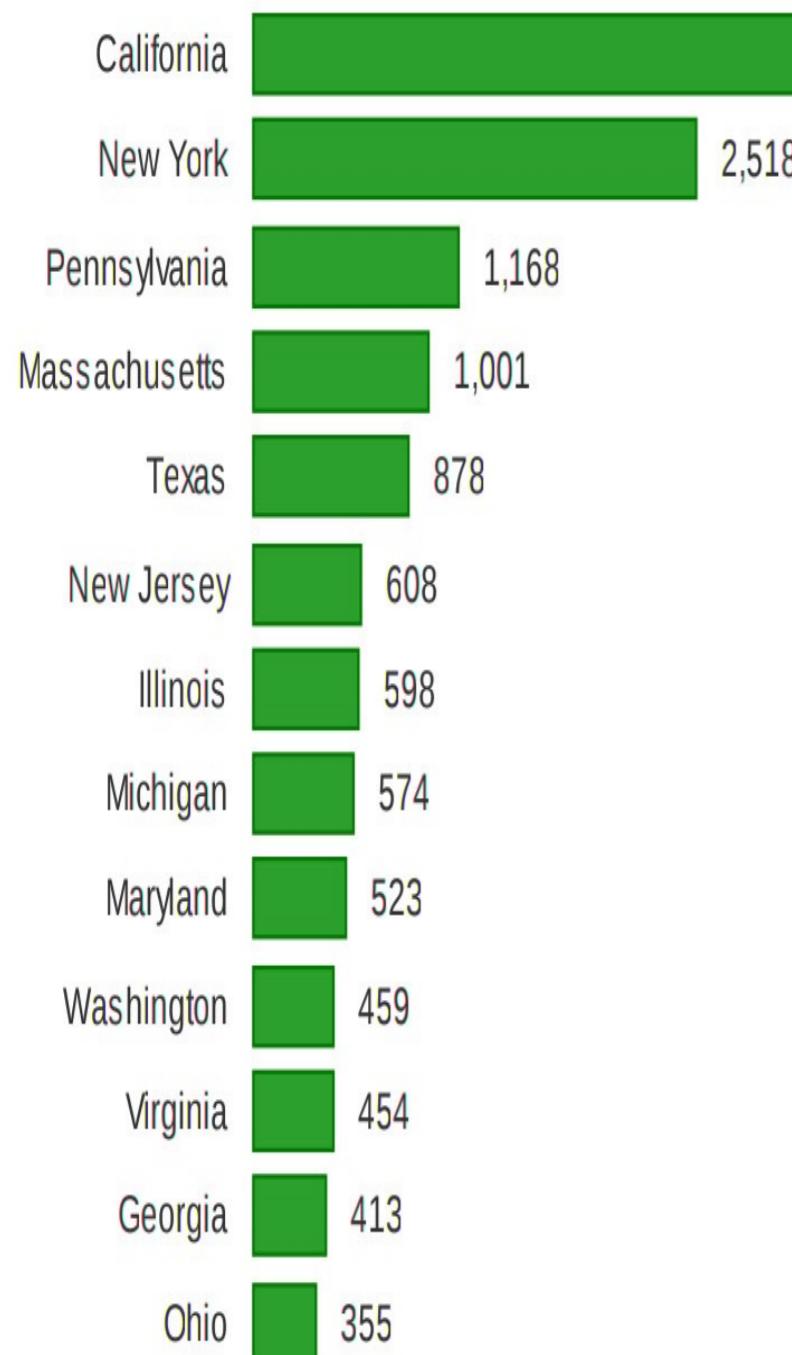
Accuracy Vs. Number of Features



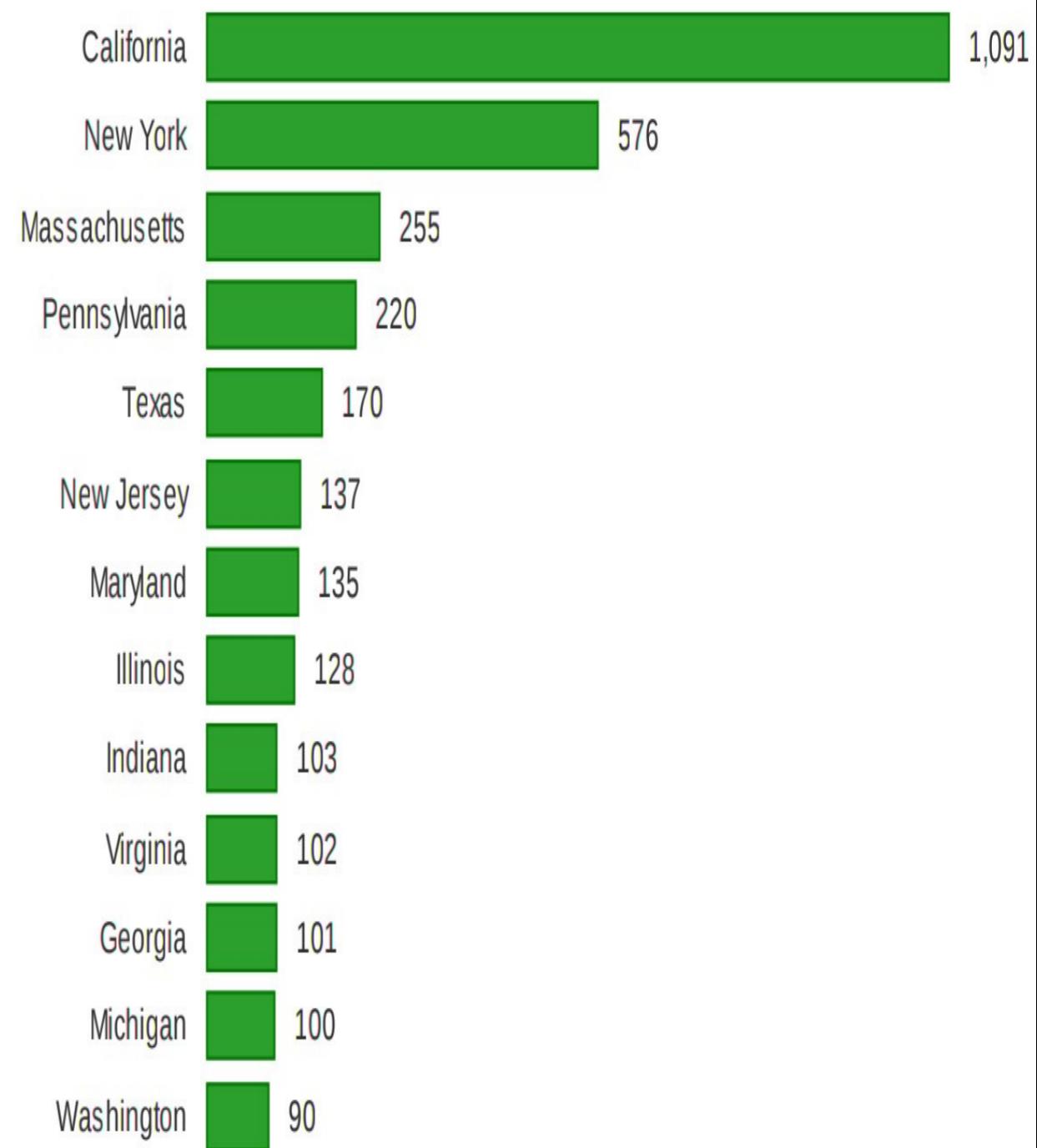


Location trends - Participation across states

participant_count

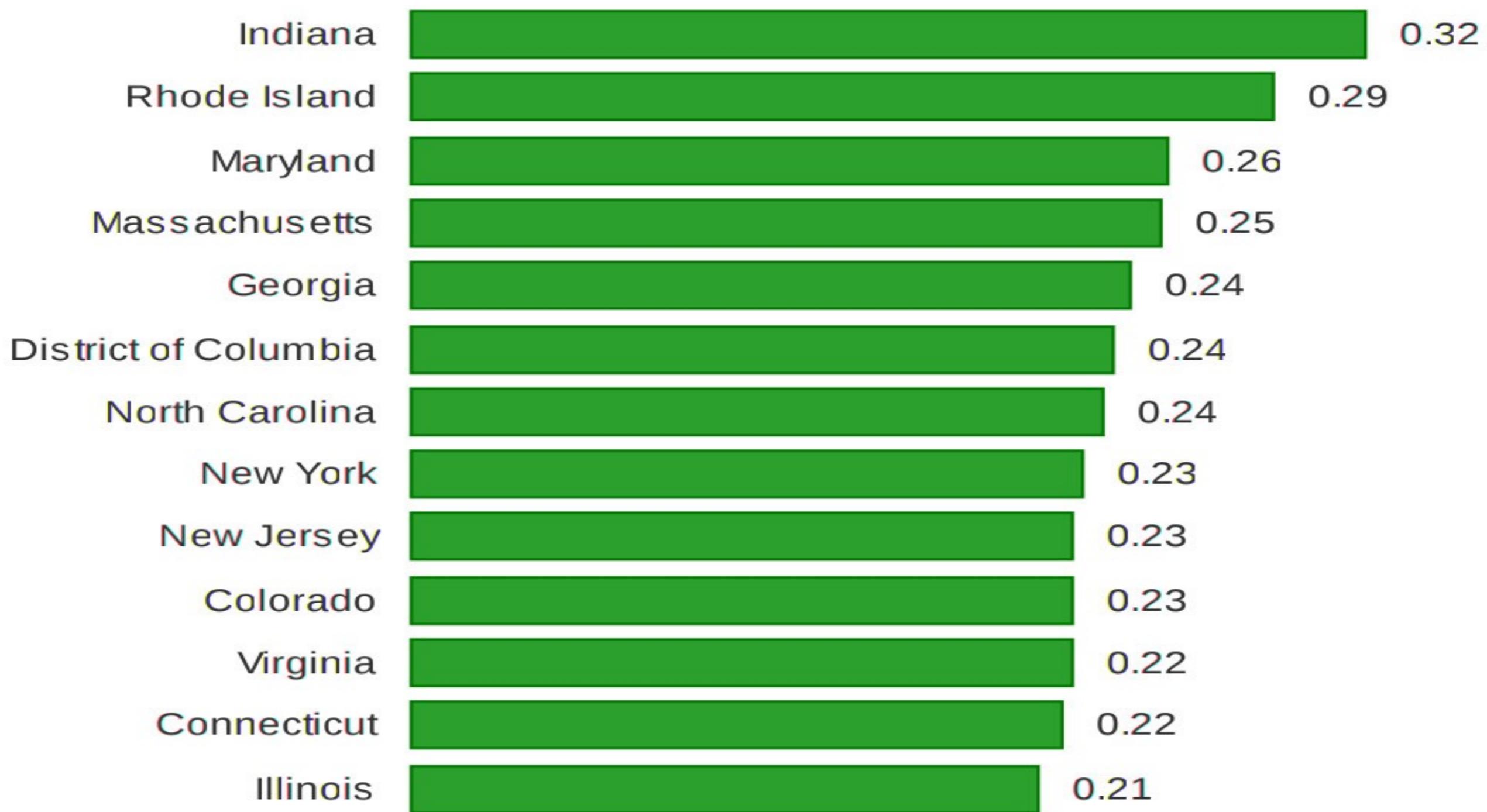


win_count

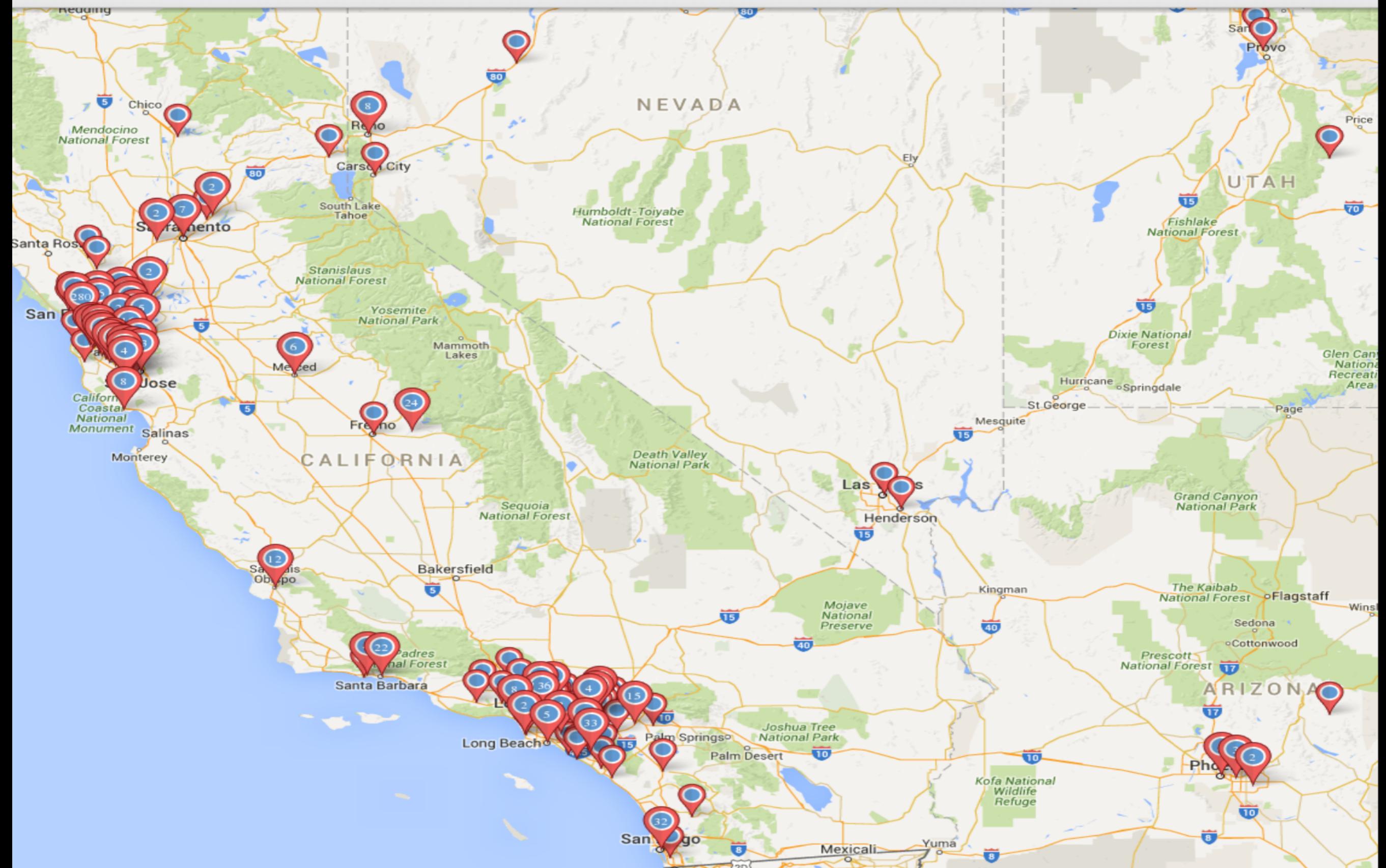


Participation across states

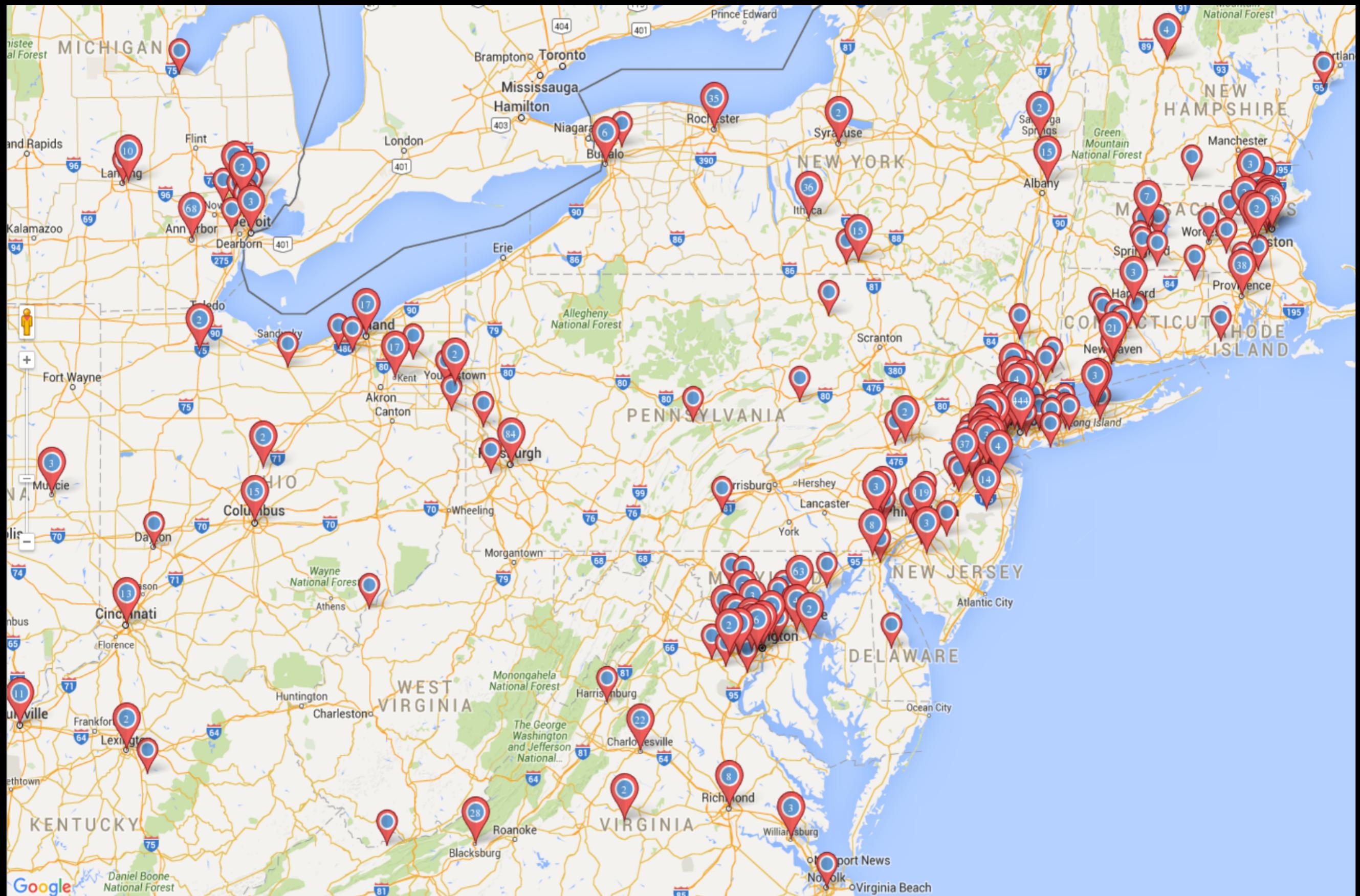
win_ratio



Participation to Win ratio across states



Winners : West coast



Winners : East coast

Future Work

- Recommendation system based on User collaboration, skill-set, location using clustering.

Thank You