Mining Hackathon Projects data

- Siddharth Patki, Kanishk Tripathi
December 11, 2015

What is a Hackathon?

- Hack + Marathon = Hackathon

Who conducts?

PENN APPS

HACKSC

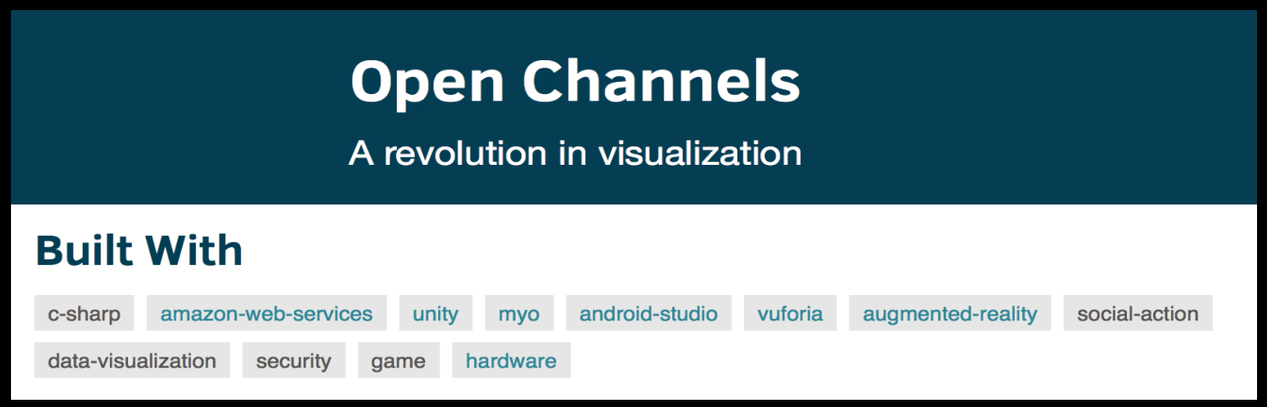
Idea

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

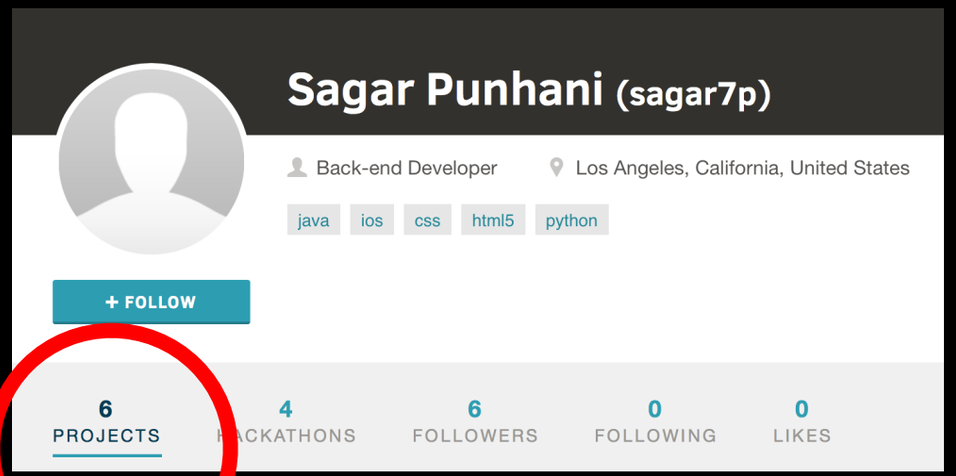
Submission Hub

DEVPOST

Project Details



Participent Profiles



Basic Preprocessing

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

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

Built With

c# | virtual-reality | oculus | razer-hydras | visual-studio | windows | unity

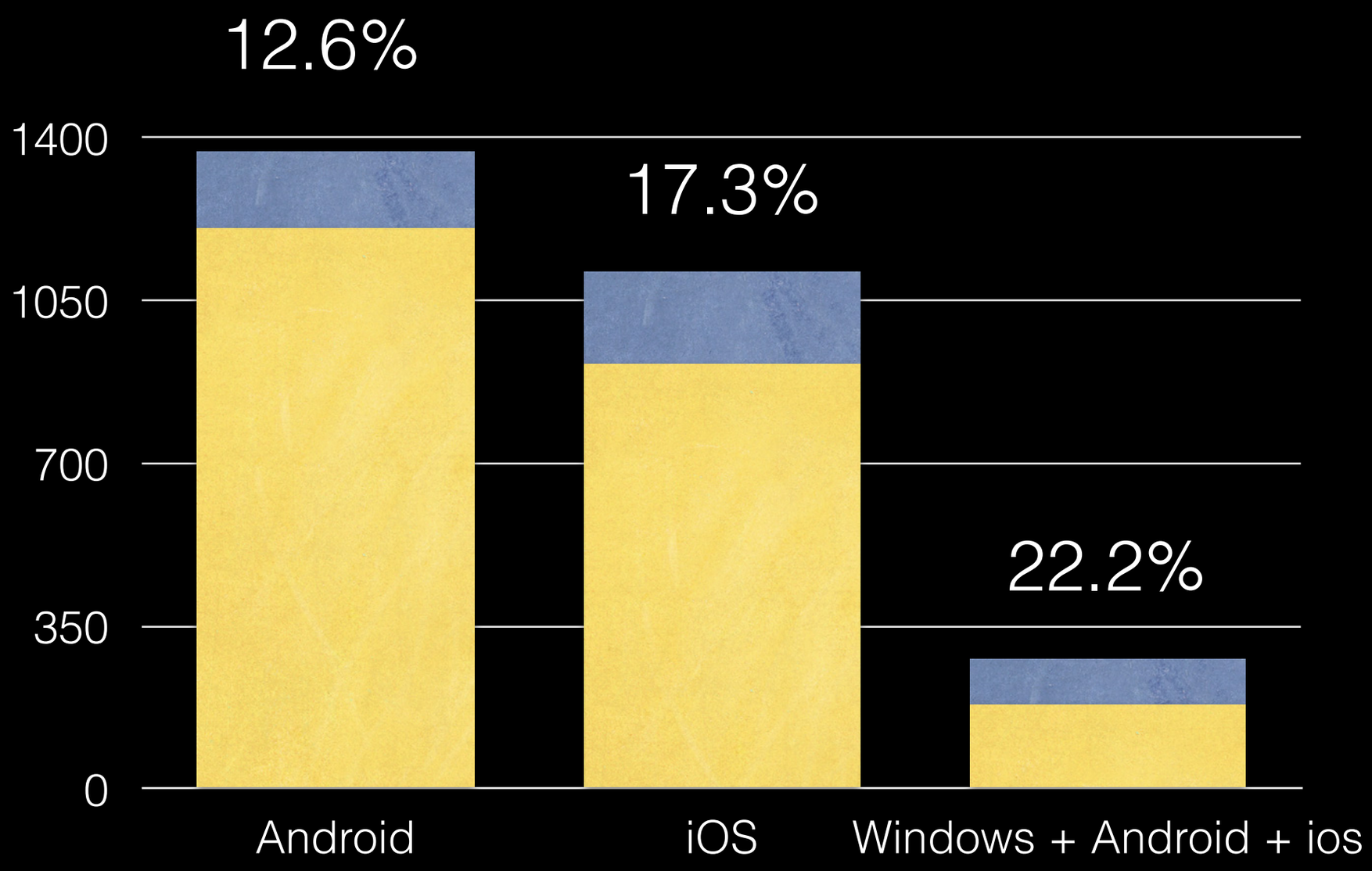
2. Number of team members

3. Team Experience

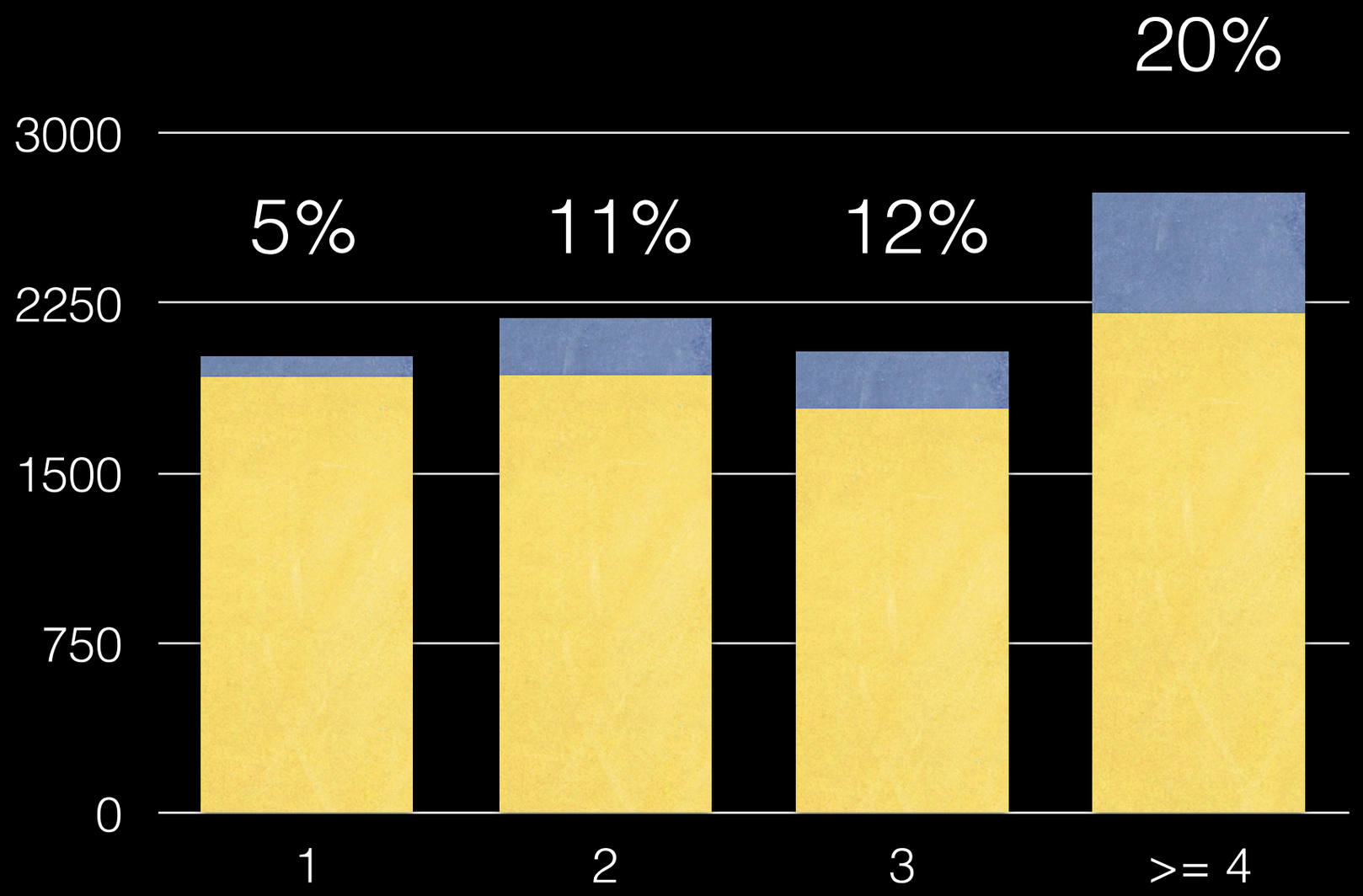
Overall Project tags

Winning Projects tags

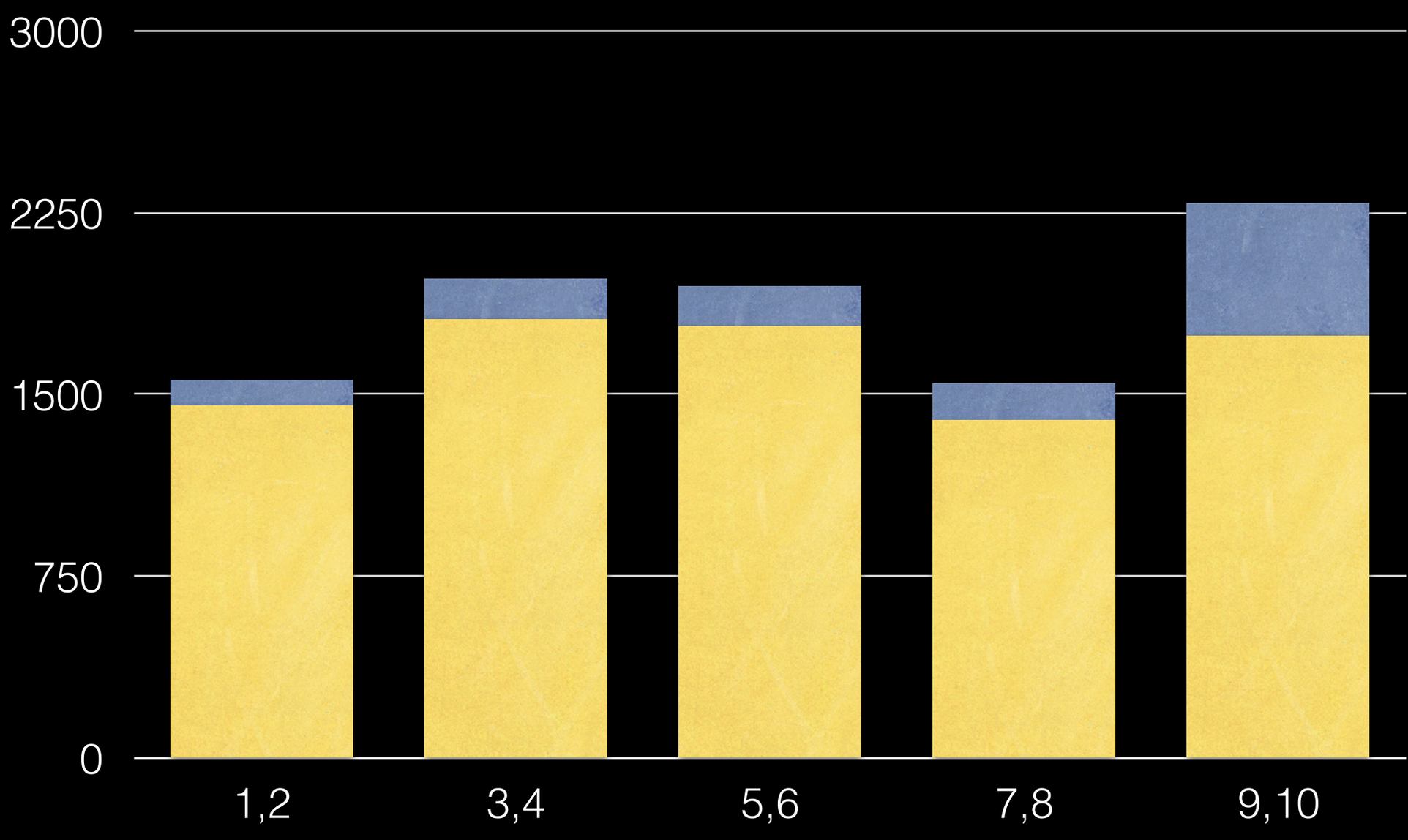
Trends - mobile app development



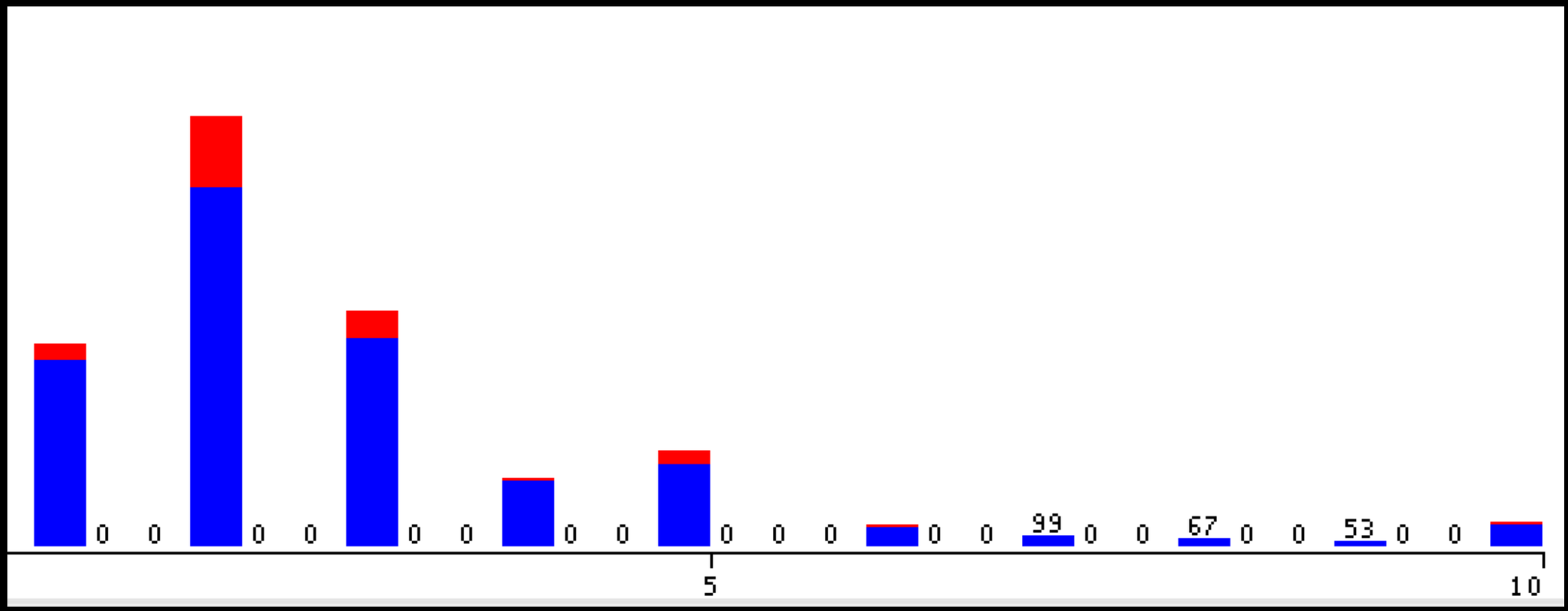
Trends - team members



Trends - team experience



Trends - avg. experience per member



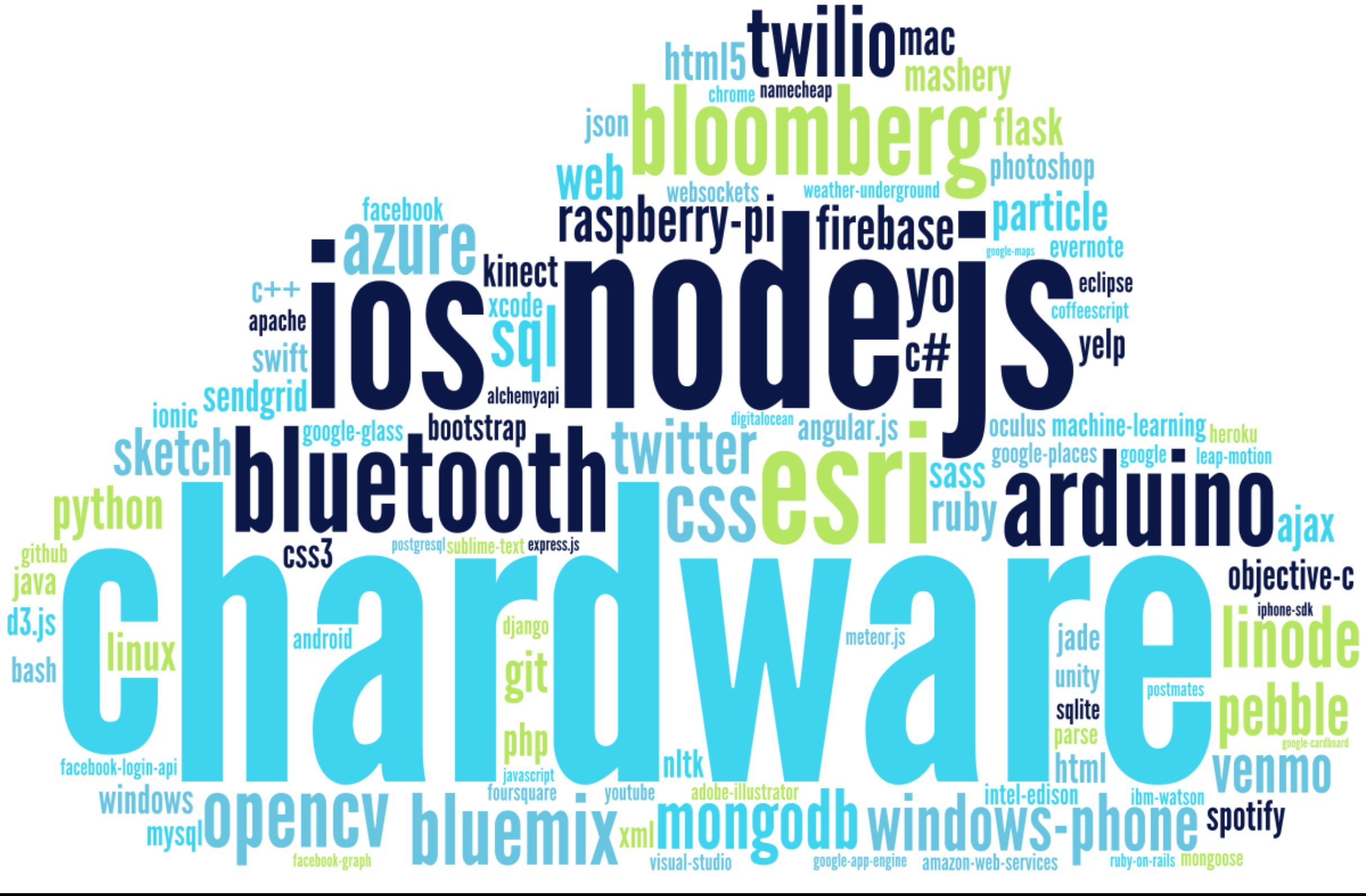
1 2 3 10

Frequent does not mean discriminative

Information Gain

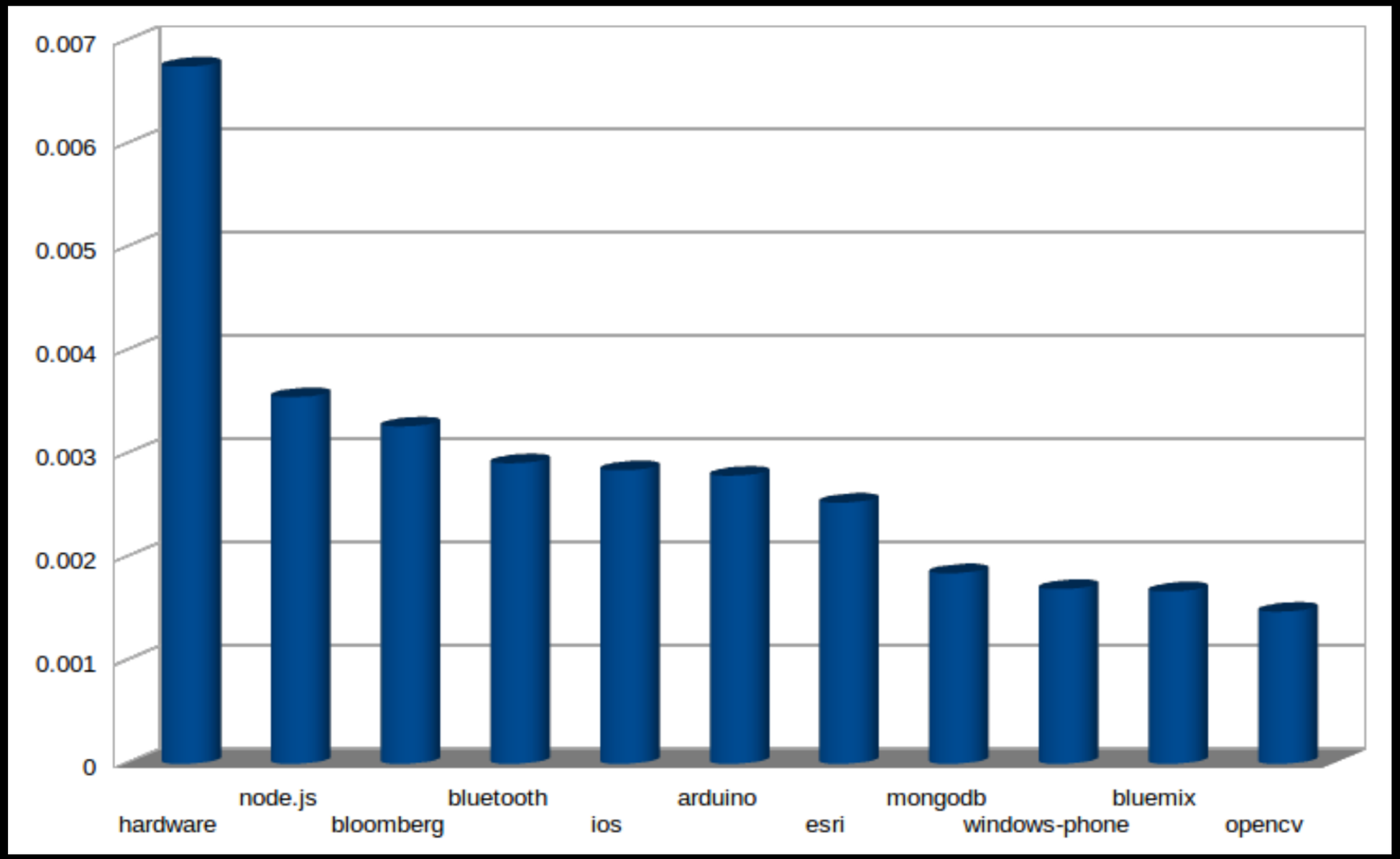
Rank tags according to their **information gain**

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



Information Gain

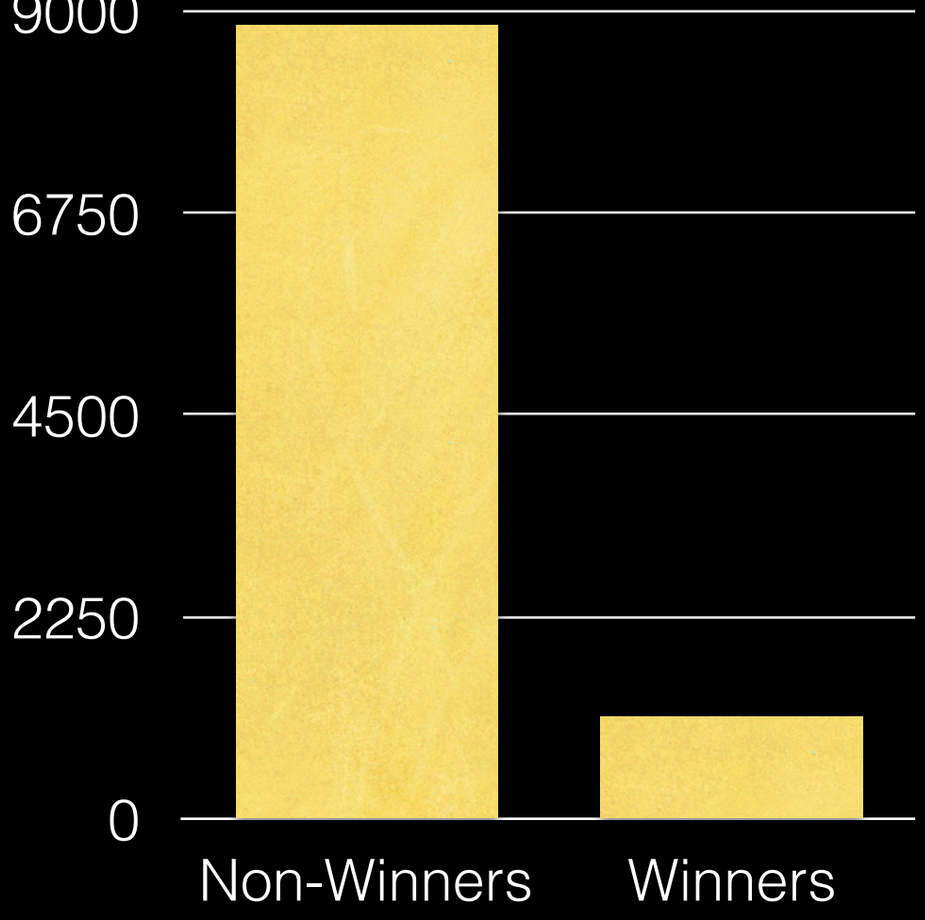
Discriminative Tags - continued



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
```

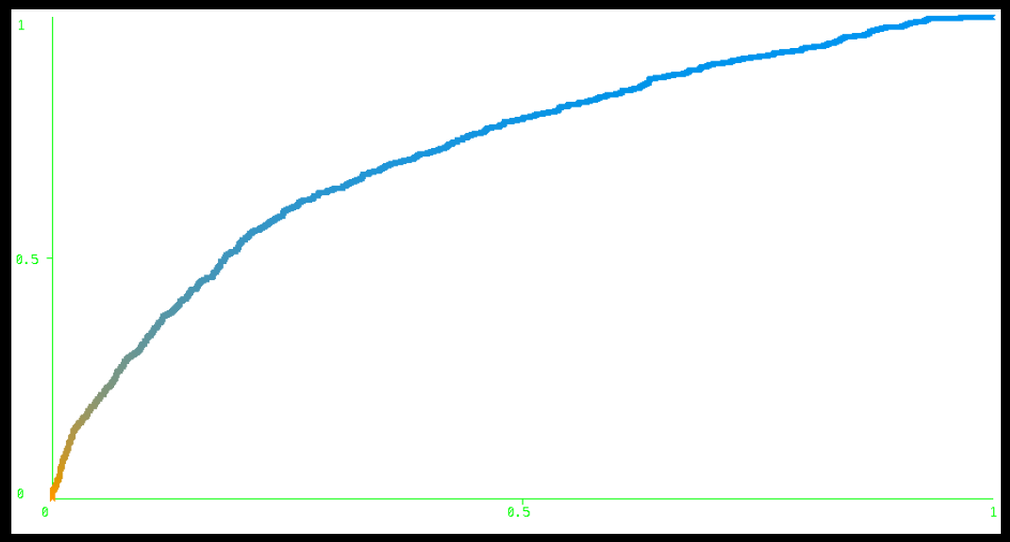
77.5% Precision, TPR = 53%

Random Forest

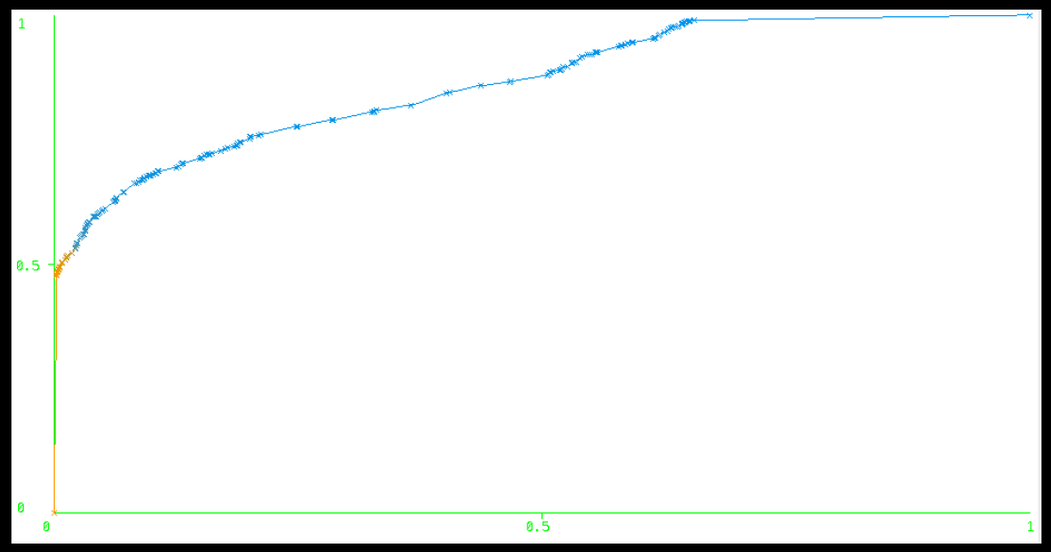
```
   a    b   <-- classified as
7684  162 |     a = False
  67 1071 |     b = True
```

86.8% Precision, TPR = 94%

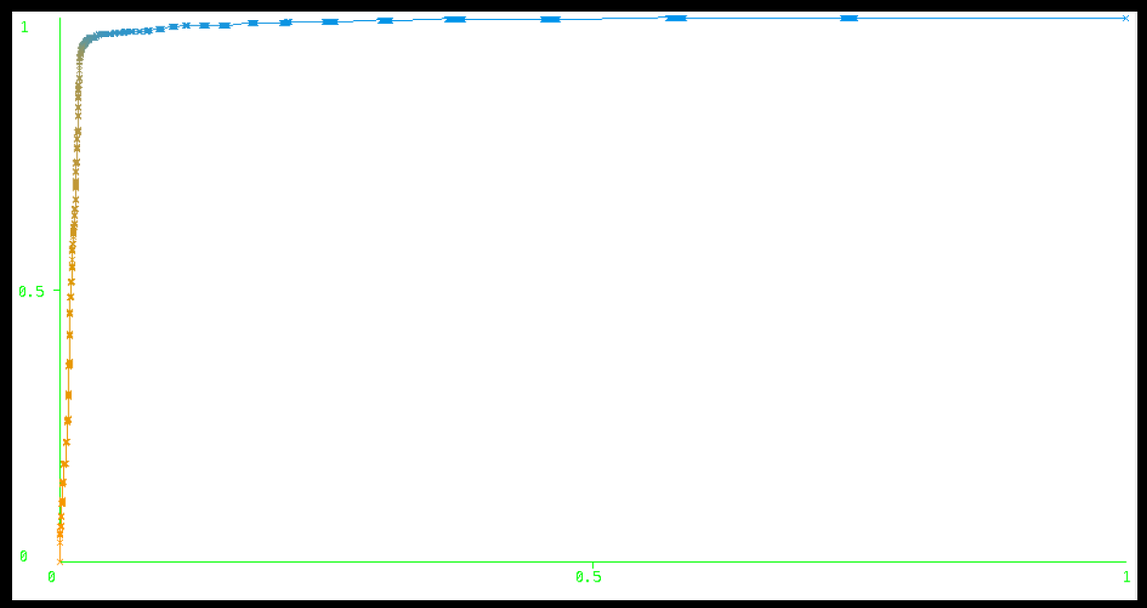
ROC



Naive Bayes

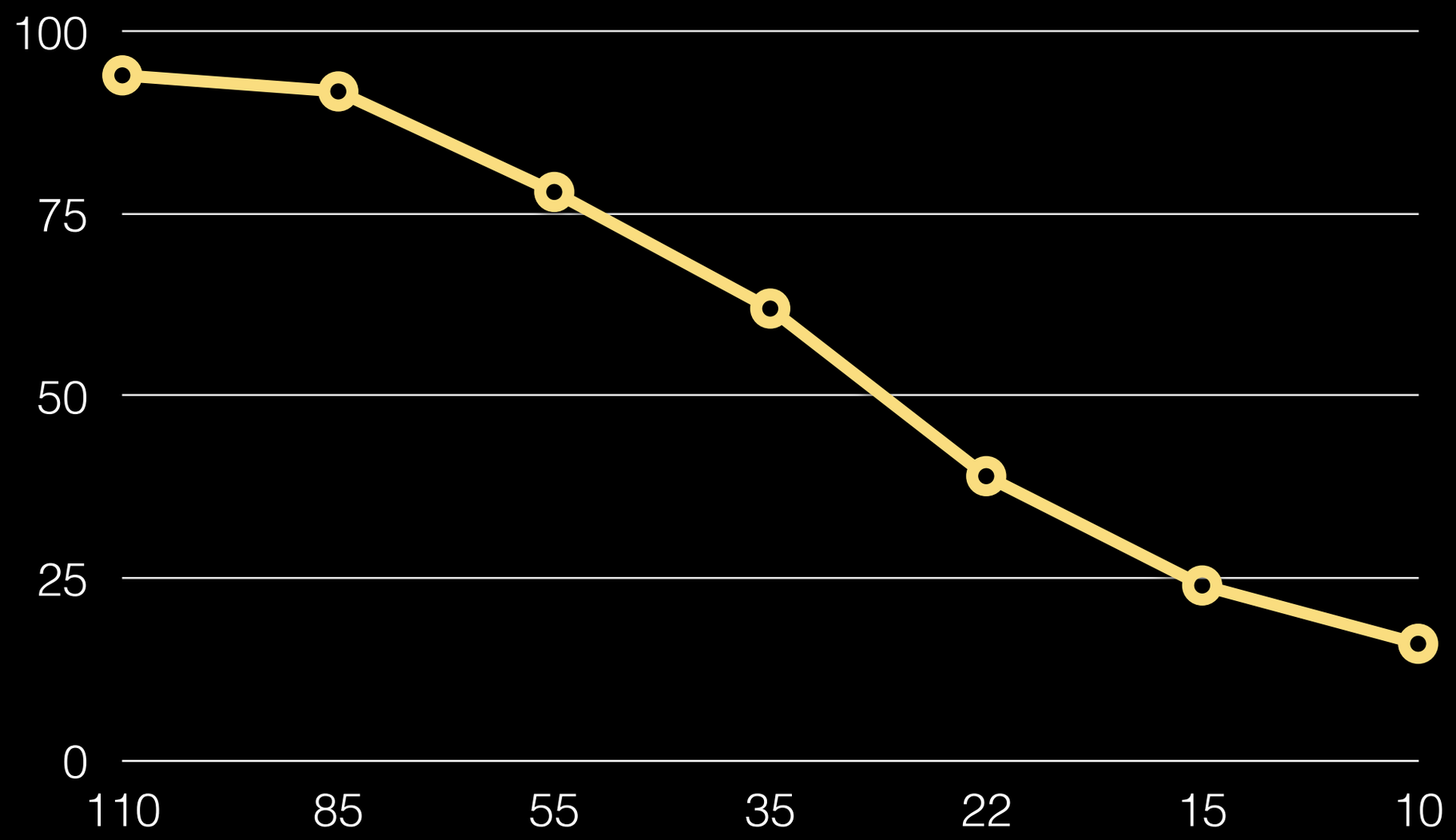


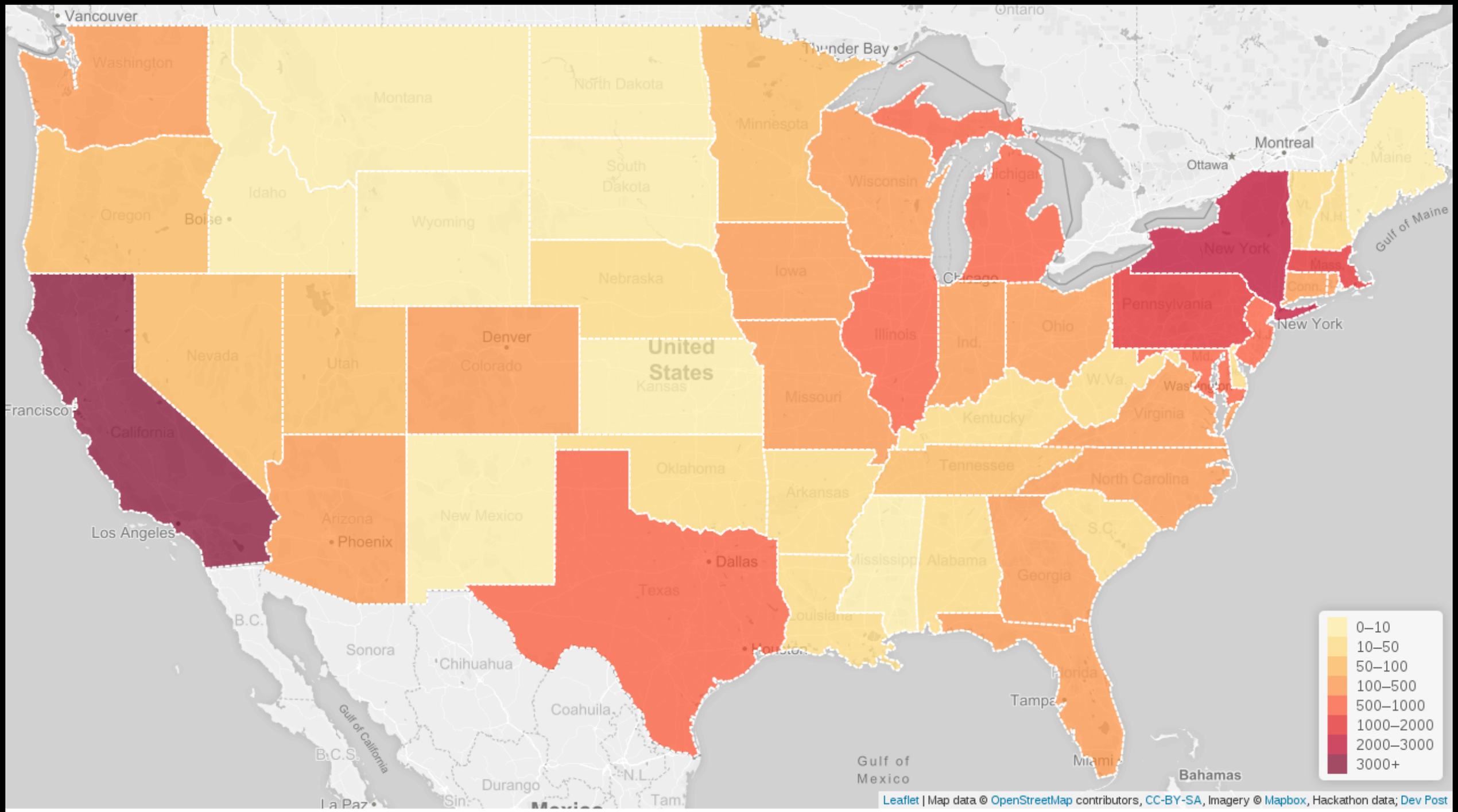
C 4.5



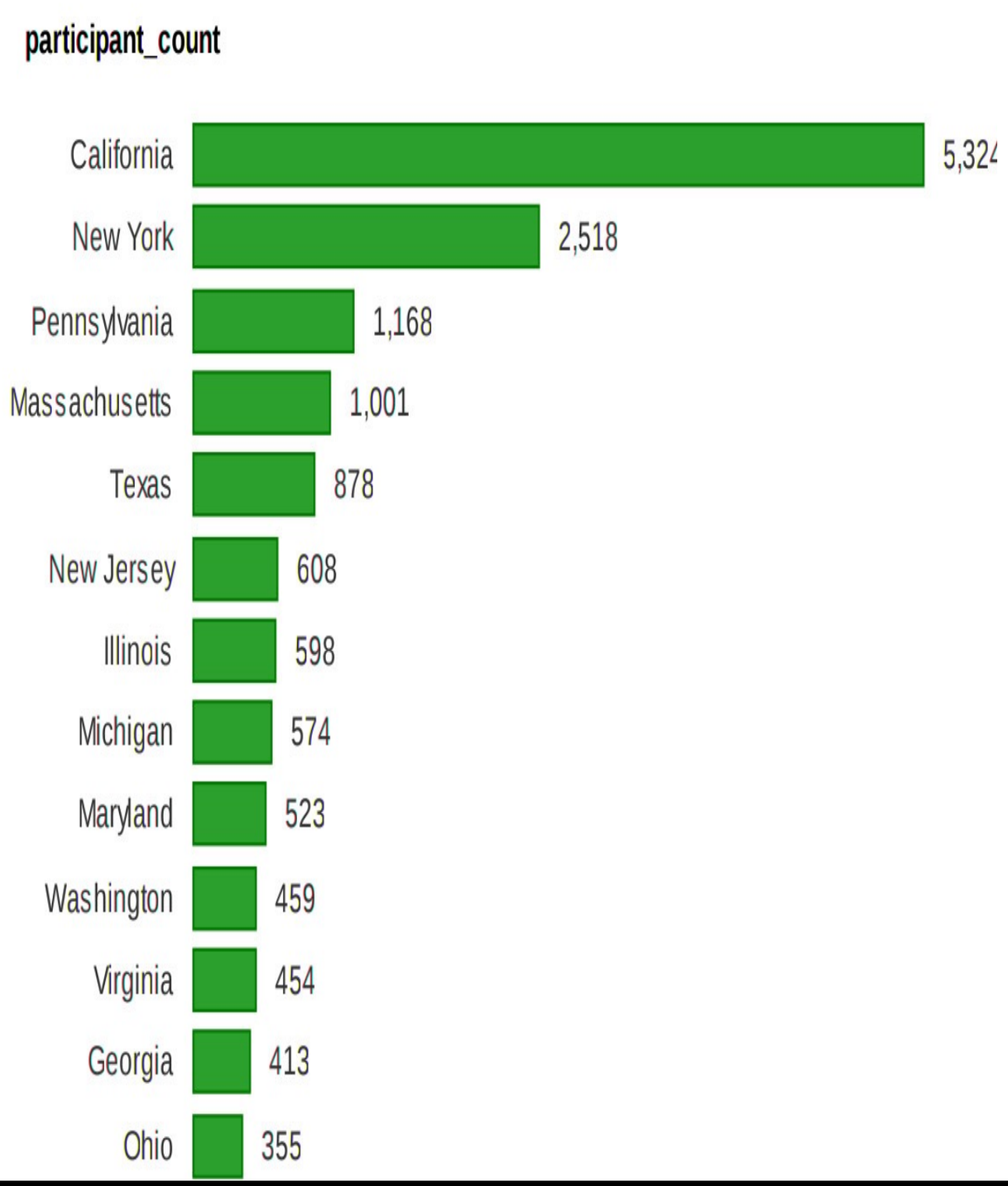
Random Forest

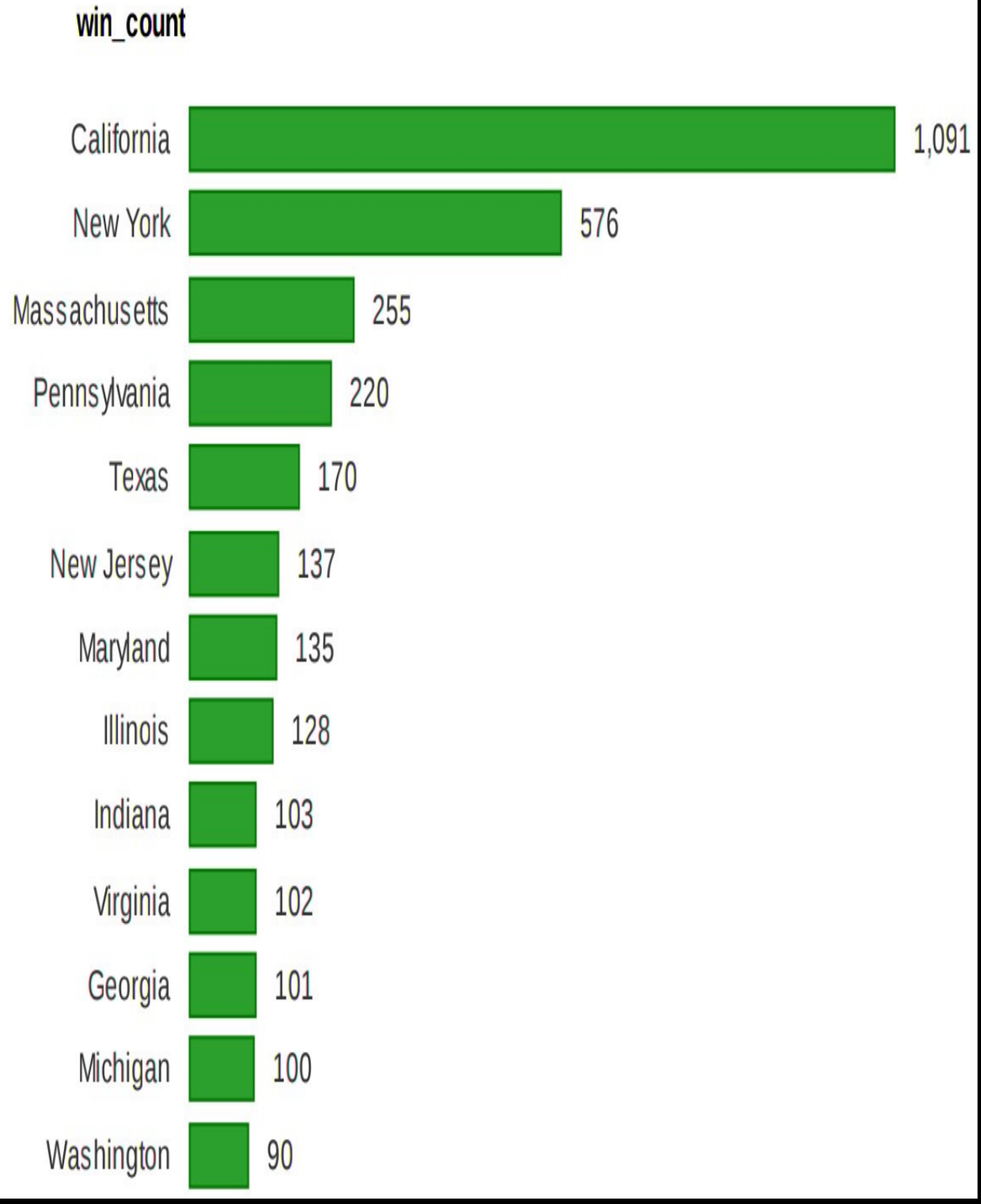
Accuracy Vs. Number of Features



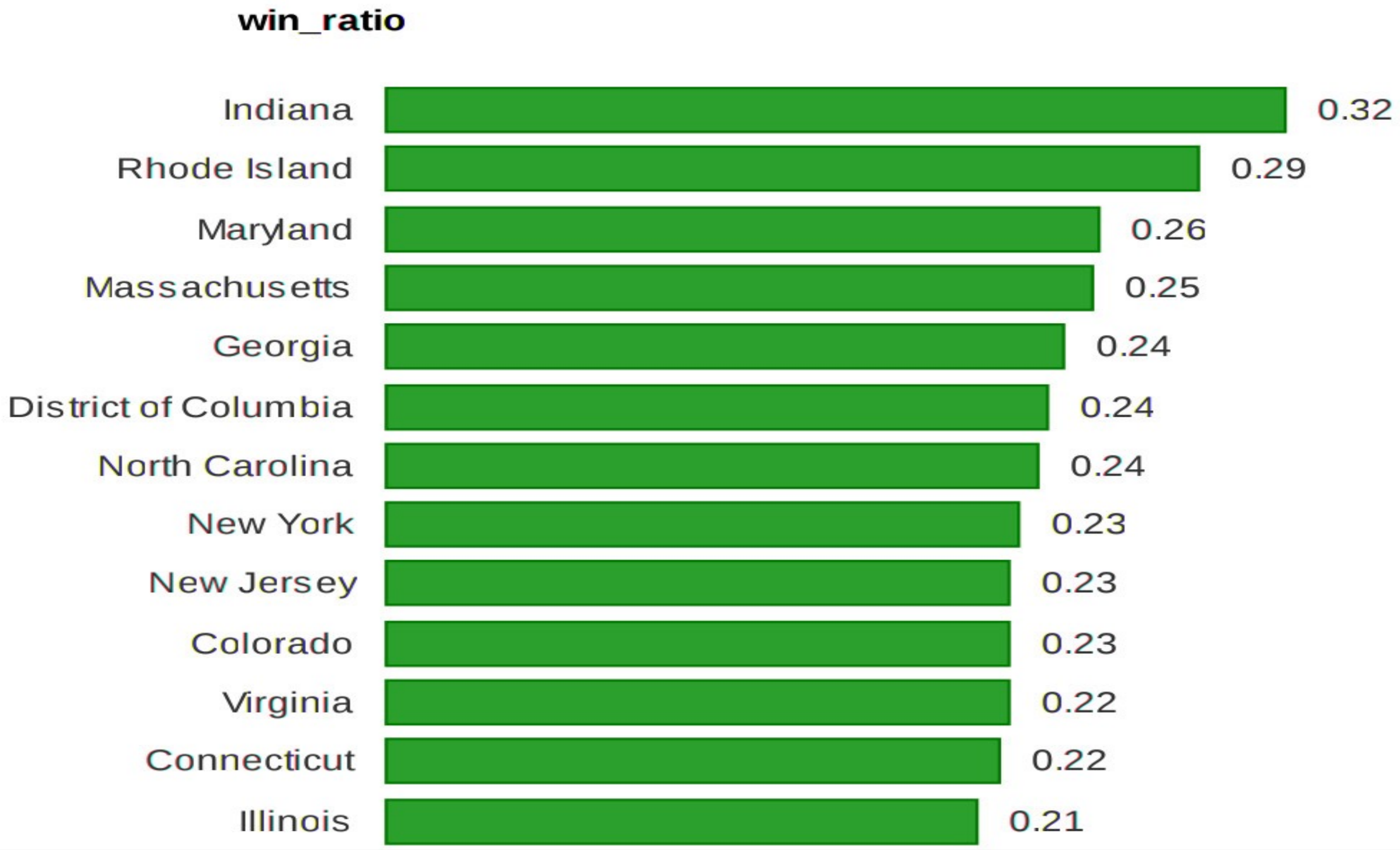


Location trends - Participation across states

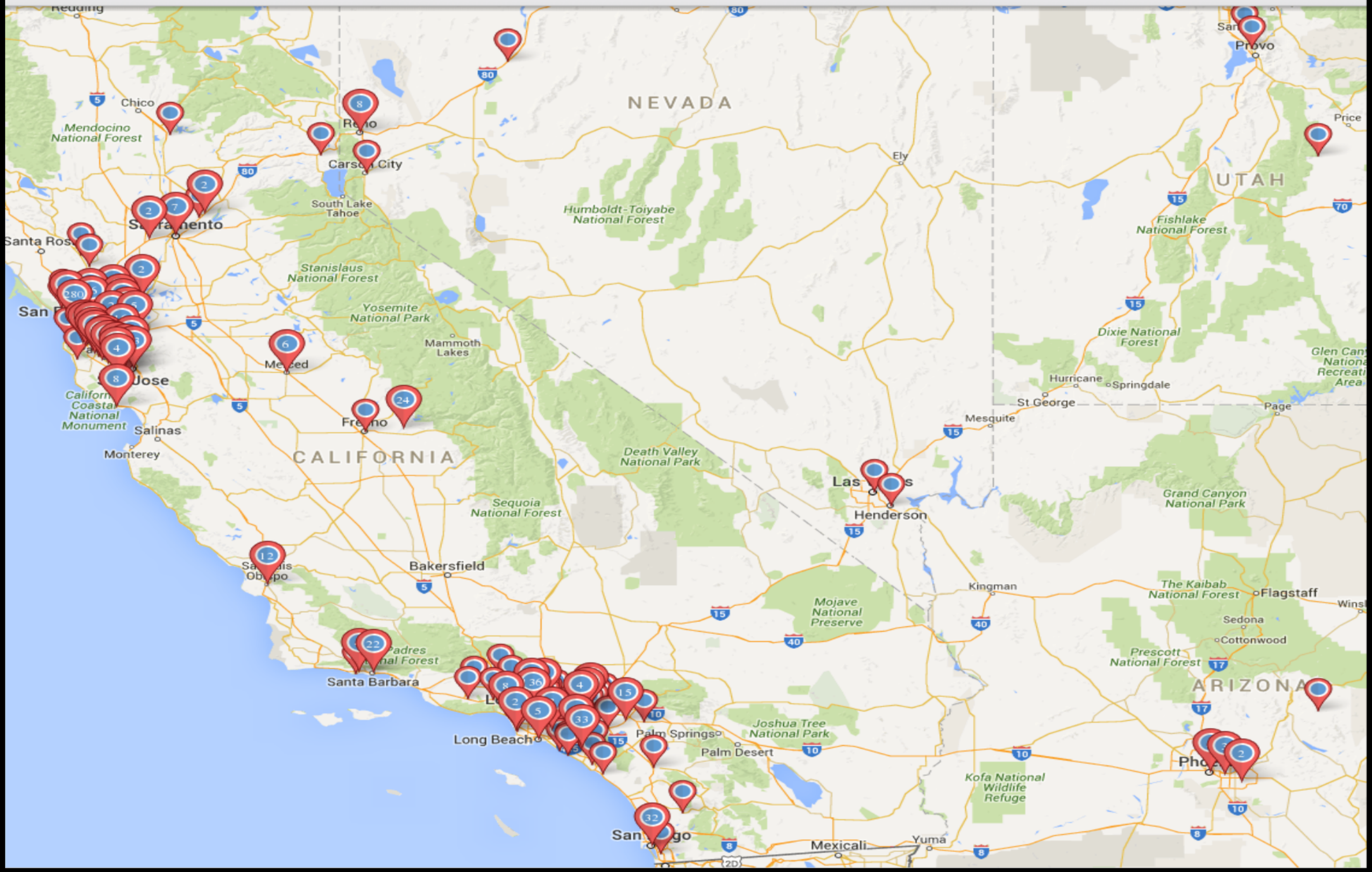




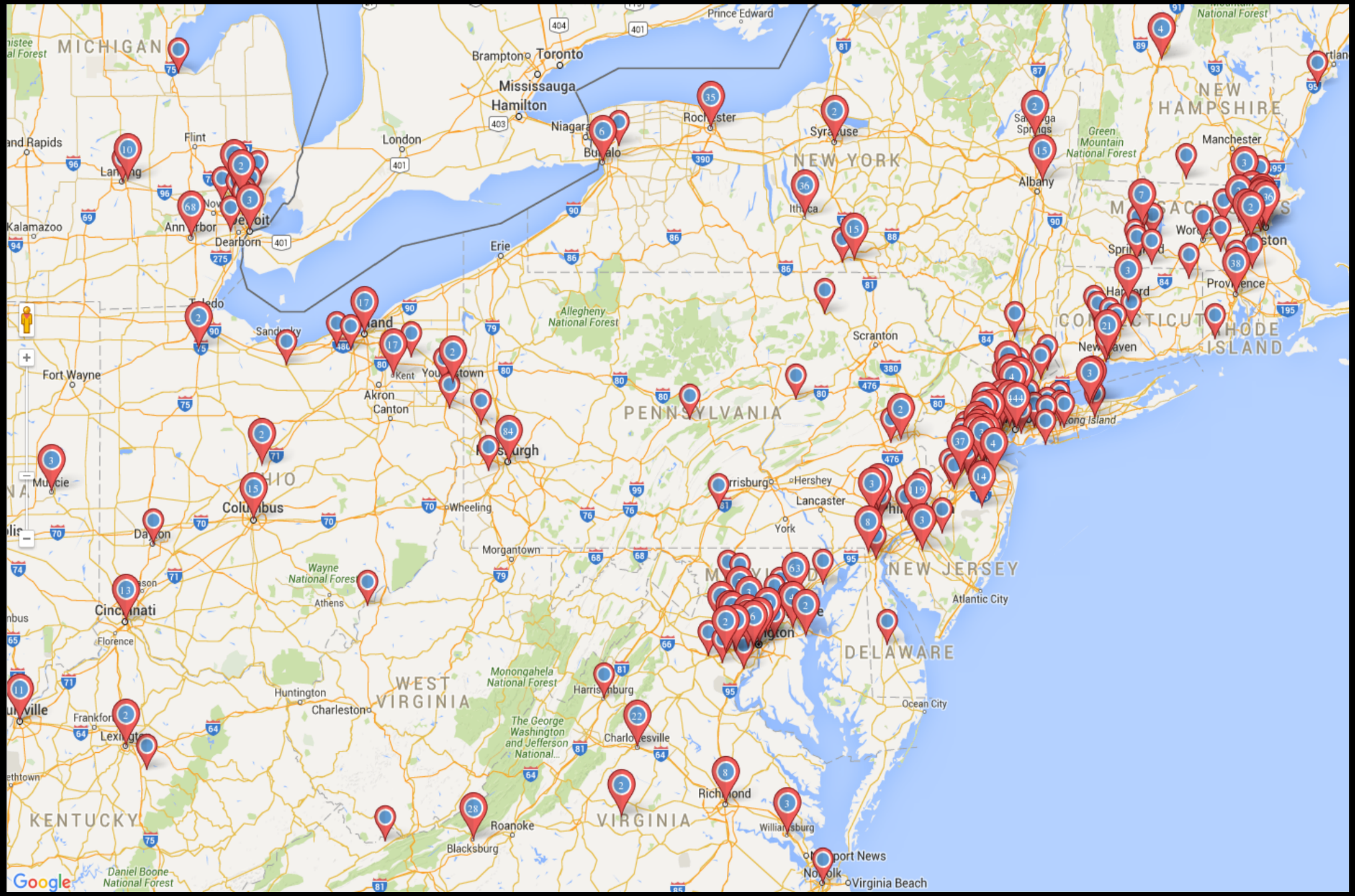
Participation across states



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