



Data Mining: Reddit Analysis



Hello!

Presentation by Nicholas Kovacs



1

Locational data:

Why and How?



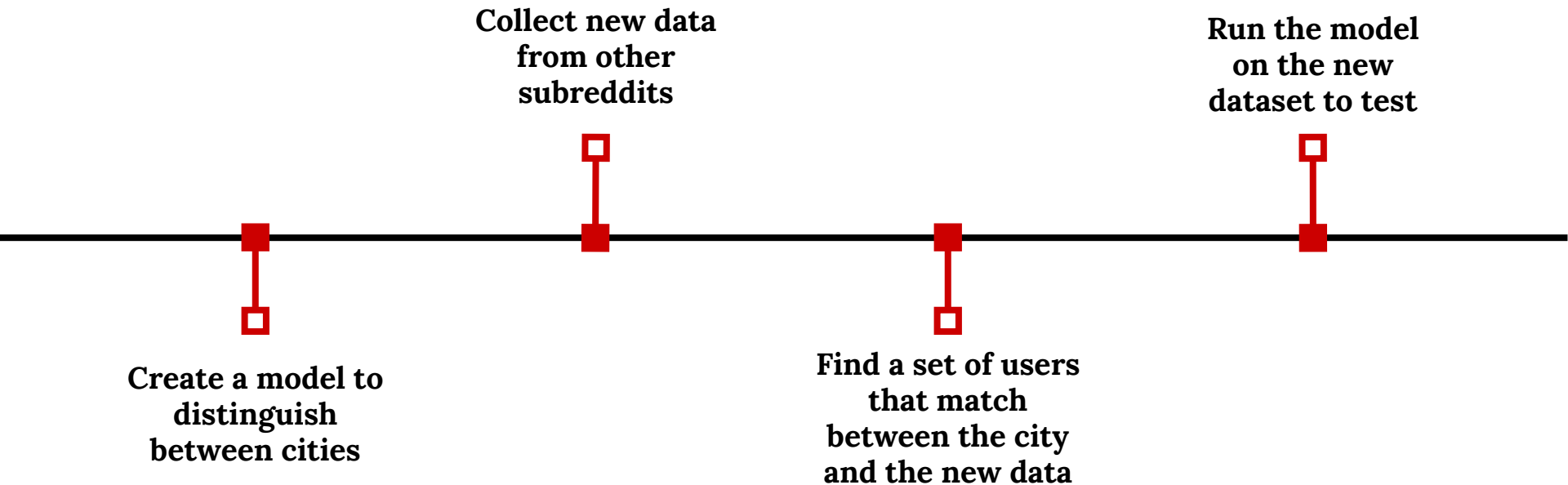
Why?

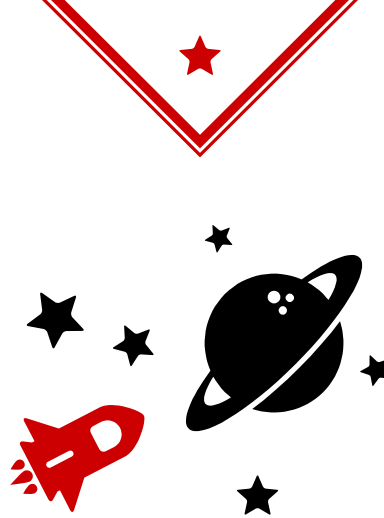
- ◆ Location-specific goods and services
- ◆ Targeted advertising
- ◆ Value of data






How?





Initial Models

Basic testing of several models to see if the
premise (distinguishing between cities) is even
possible



Models Testing

Models Tested

- ❖ Multinomial Naive Bayes
- ❖ K-Nearest Neighbors
- ❖ Random Forest
- ❖ SVM

Naive Bayes:

Accuracy: 67%

Baseline: 51%





Advancing the City Discriminator



Three cities, equally distributed: 51% accuracy rating
(still ~17% above baseline)



137 final posts

Not too many - but this was compensated for

34% accuracy

(it didn't beat the baseline)

1000 iterations

(it didn't just get unlucky)



Recommendations



Thanks!

Any questions?

Feel free to ask for clarification
on any point from the presentation.