Lyrics +
Predictive Analytics

Brendan Herger, Data Scientist @ Capital One
Hergertarian.com

Run the World (Girls)

6 writers
4 producers

Girls, we run this mother (yeah!) x4
Who run the world? Girls (girls)
We run this motha? Girls (girls)
Who run the world? Girls (girls)
Who run the world? Girls (girls)
Girls, we run this mother? Girls
Girls, we run this mother? Girls
Girls, we run this mother? Girls
Girls, we run this mother? Girls
Who run the world? Girls (girls)
We run this motha? Girls (girls)
Who run the world? Girls (girls)
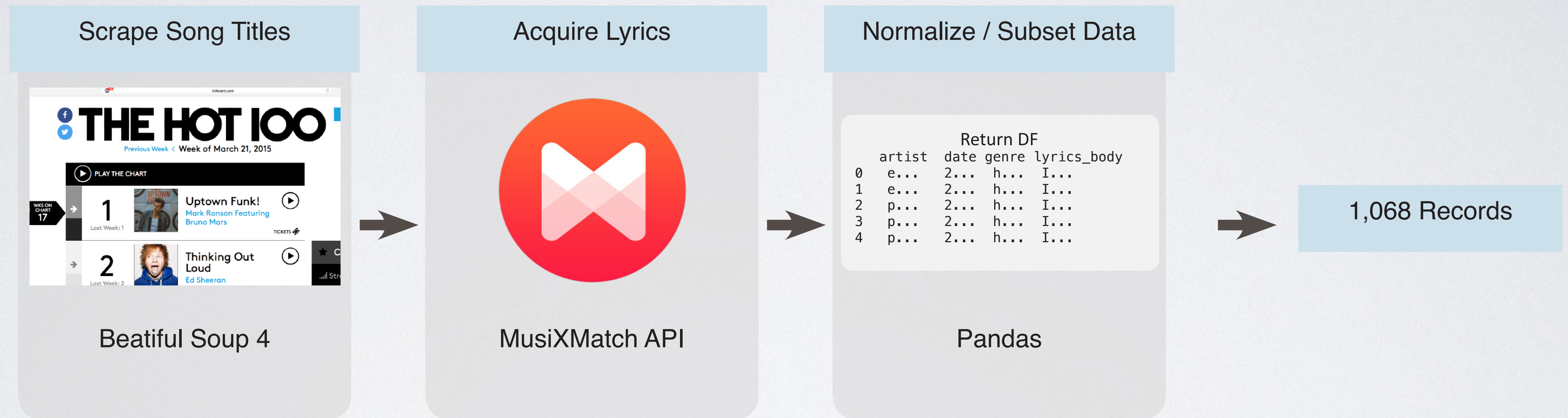Who run the world? Girls (girls)

Bohemian Rhapsody

1 writer
1 producer

Is this the real life?
Is this just fantasy?
Caught in a landslide,
No escape from reality
Open your eyes,
Look up to the skies and see,
I'm just a poor boy, I need no sympathy,
Because I'm easy come, easy go,
Little high, little low,
Any way the wind blows doesn't really matter to
me, to me
Mama,
I just killed a man,
Put a gun against his head, pulled my trigger
now he's dead
Mama... life had just begun,
But now I've gone and thrown it all away
Mamaaaaa oooh,
Didn't mean to make you cry,
If I'm not back again this time tomorrow,
Carry on, carry on as if nothing really matters

Can we identify genre from lyrics?

Data Acquisition
Feature Engineering
Statistical Modeling

Data Acquisition

Scrape Song Titles

Beatiful Soup 4

Acquire Lyrics

MusiXMatch API

Normalize / Subset Data

```
              Return DF
      artist  date genre lyrics_body
0     e...    2...  h...  I...
1     e...    2...  h...  I...
2     p...    2...  h...  I...
3     p...    2...  h...  I...
4     p...    2...  h...  I...
```

Pandas

1,068 Records

Feature Extraction

Words per line

Number of words

Unique word density

Noun density

Curse word density

Verb density

Stop word density

Curse Word Density

Models

Algorithm	
Logistic Regression	
Support Vector Classifier	
Random Forest	
AdaBoost	
KNN	
Stochastic Gradient Boost	

Accuracies

Algorithm	Accuracy
Logistic Regression	70.4%
Support Vector Classifier	74.1%
Random Forest	75.9%
AdaBoost	75.9%
KNN	79.6%
Stochastic Gradient Boost	81.5%

Accuracies

Algorithm	Accuracy
Logistic Regression	70.4%
Support Vector Classifier	74.1%
Random Forest	75.9%
AdaBoost	75.9%
KNN	79.6%
Stochastic Gradient Boost	81.5%
Naive Bayes	81.2%

Future Work

- Generalize past 2 genre

- Model probability of commercial success

- Author identification

Thanks!

GitHub: goo.gl/YVtW2u

Brendan Herger | Hergertarian.com

| 13herger@gmail.com