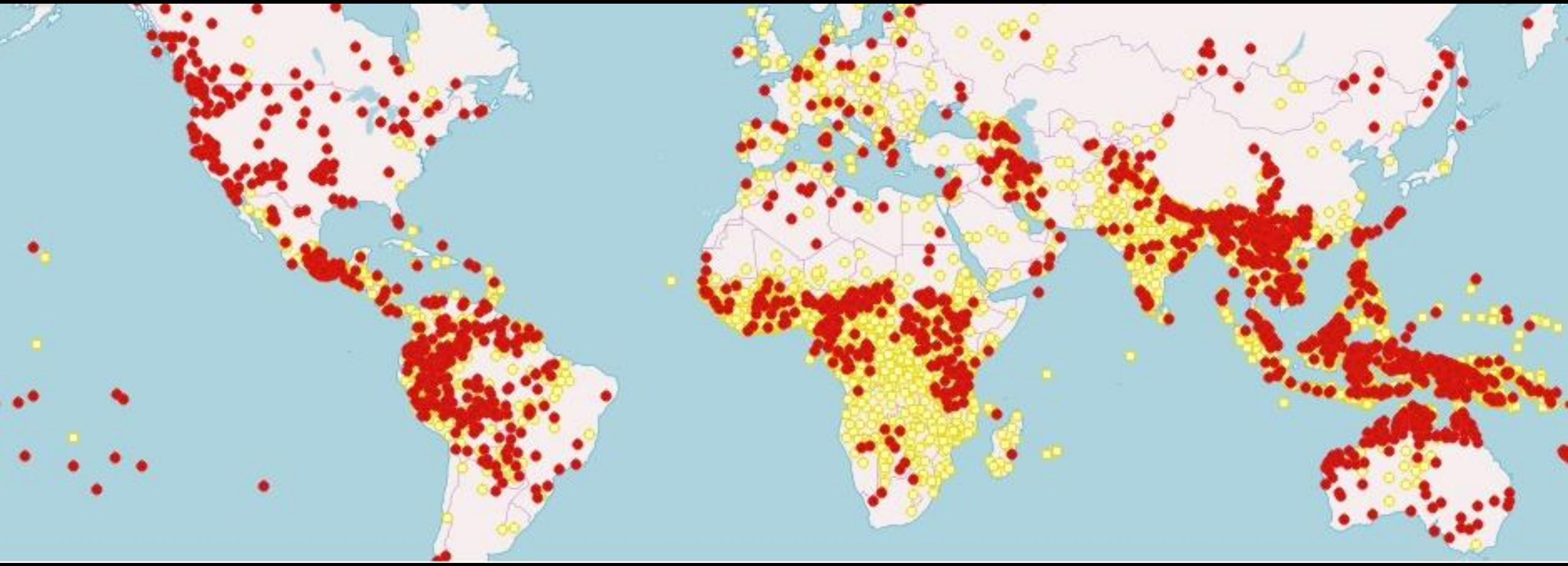


Saving Endangered Languages using Reddit Data

SEAN Y. LI

Background

60% of languages are endangered

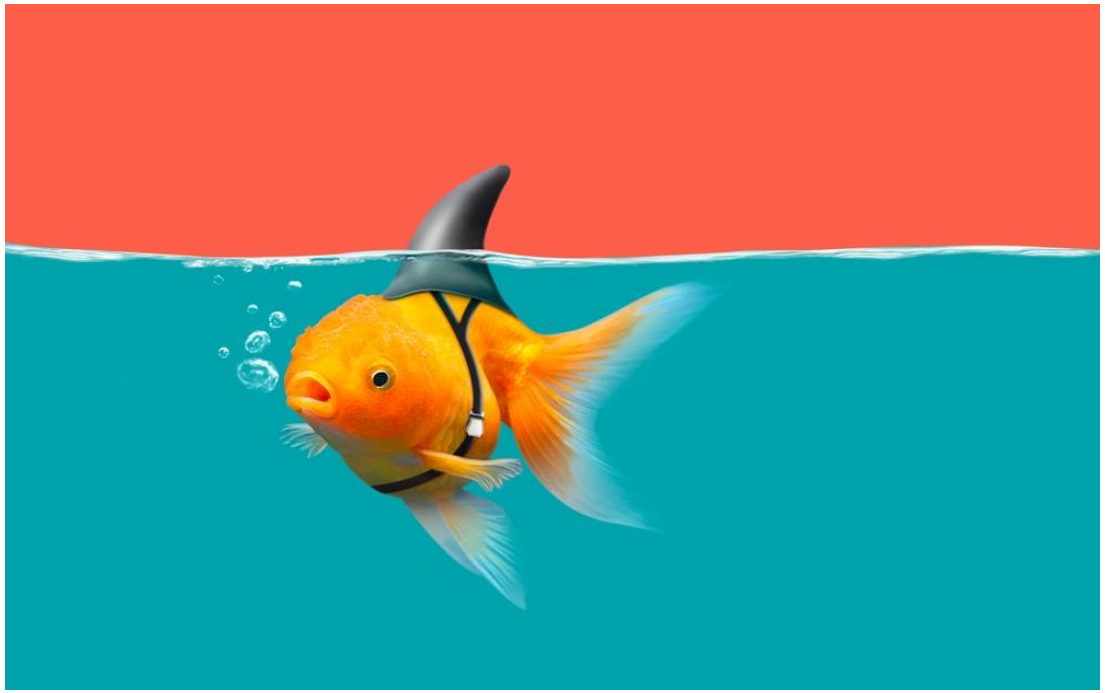
40% are moribund

Crowdsourcing University Labor

- Why?
  - Basic documentation simple
  - Keeping audio, video, and text data
- Further linguistic analysis requires more training

Roadblock

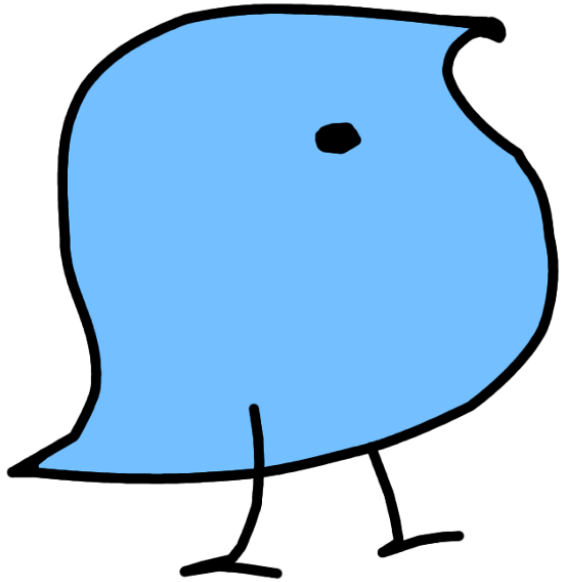
People are submitting fake data!



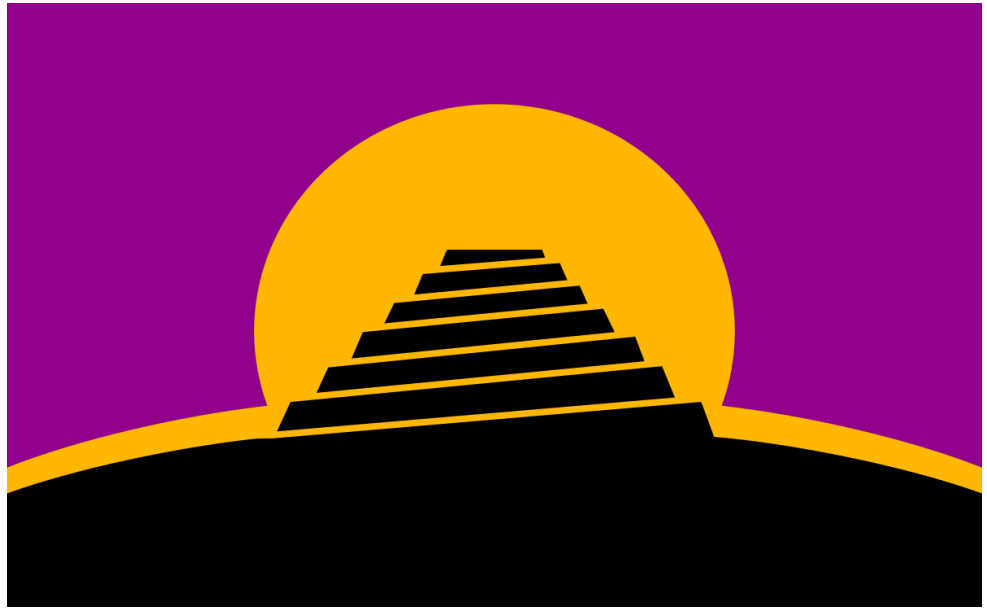
Solution

reddit

r/linguistics



r/conlangs



Problem Statement

- Can we build a NLP model based on reddit data to separate real linguistic data and conlang data?

Data Scraping

- Pushshift API
- 1000 posts from each subreddit

EDA (preliminary modeling)

- Dropped posts without description text
- GridSearch pipeline:
  - Vectorizer
    - TF-IDF Vectorizer
  - Classifier
    - Logistic regression
    - Multinomial Naive Bayes
    - *k*-NN classifier
    - Random Forest Classifier

Preliminary Model Results

1311 samples

50.1% r/linguistics

49.9% r/conlangs

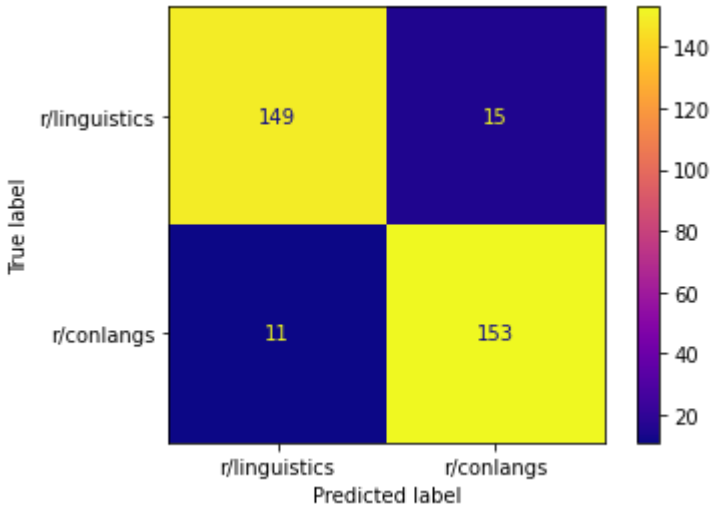
Preliminary Model Results

1311 samples

50.1% r/linguistics

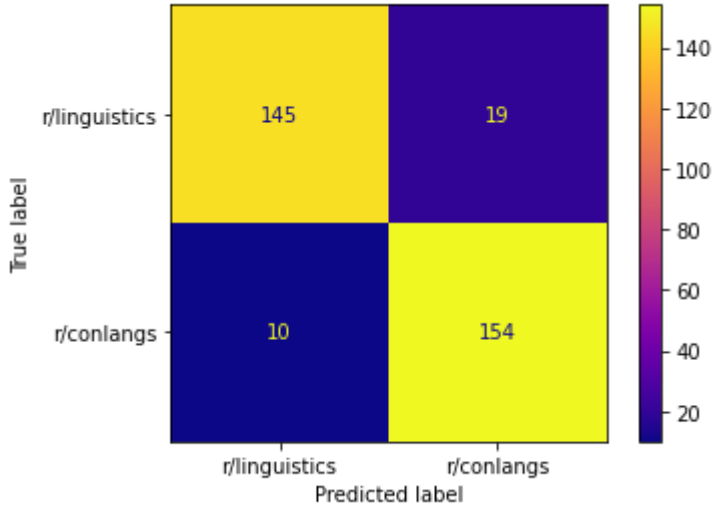
49.9% r/conlangs

Best Performing Models



Logistic Regression

Train: 97.56% Test: 92.07%



Random Forest Classifier

Train: 99.99% Test: 91.16%

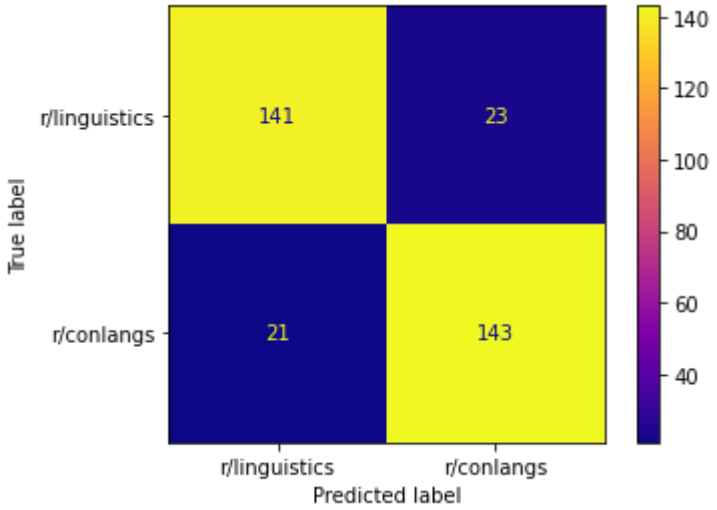
Preliminary Model Results

1311 samples

50.1% r/linguistics

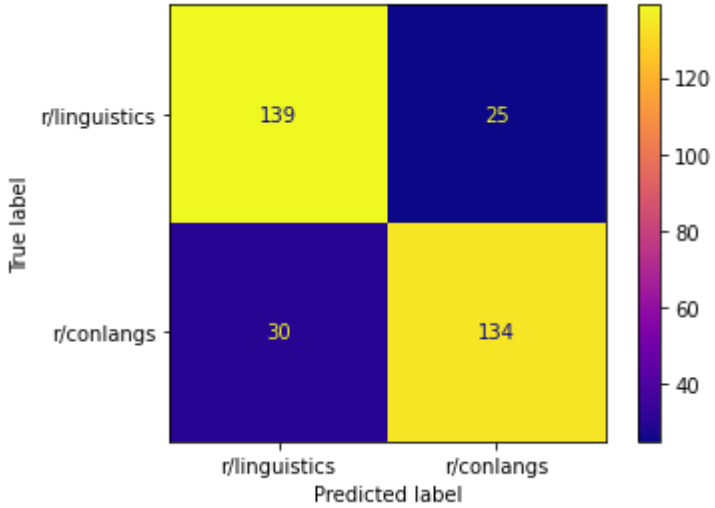
49.9% r/conlangs

Worst Performing Models



Multinomial Naive Bayes

Train: 95.73% Test: 86.59%



k-NN classifier

Train: 99.99% Test: 83.23%

Issues with models

- “conlang” not in stop words
- No lemmatization
- Many short and/or low-quality posts that could easily be manually checked
- Overfitting

Back to Data Cleaning

- Removed stopwords and “conlang” and “conlangs”
- Stemmed words with Porter Stemmer
- Dropped posts that had less than 300 characters long (about 40-80 words)

New Model Results

761 samples

60.7% r/linguistics

39.3% r/conlangs

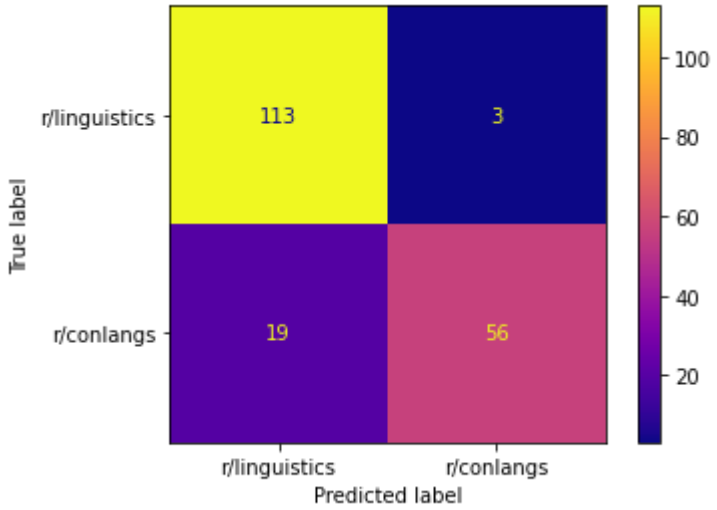
New Model Results

761 samples

60.7% r/linguistics

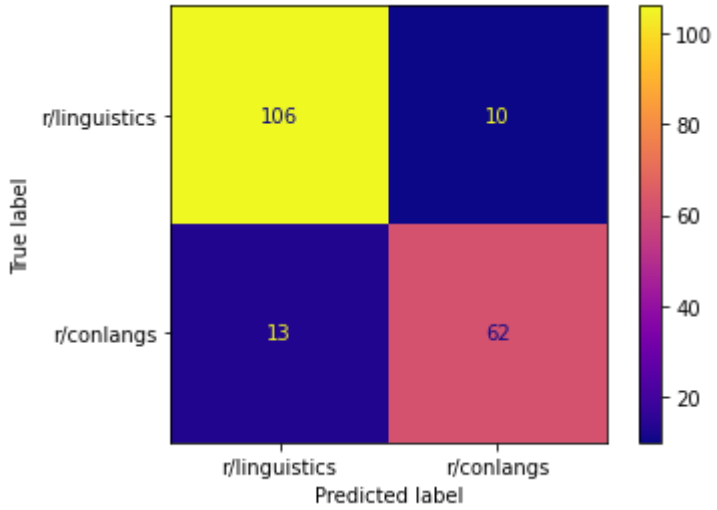
39.3% r/conlangs

Worst Performing Models



Logistic Regression

Train: 97.02% Test: 88.48%



Random Forest Classifier

Train: 100.00% Test: 87.96%

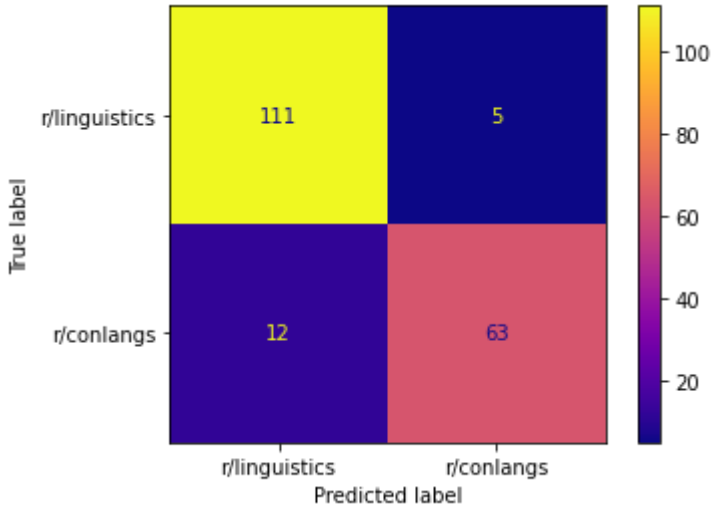
New Model Results

761 samples

60.7% r/linguistics

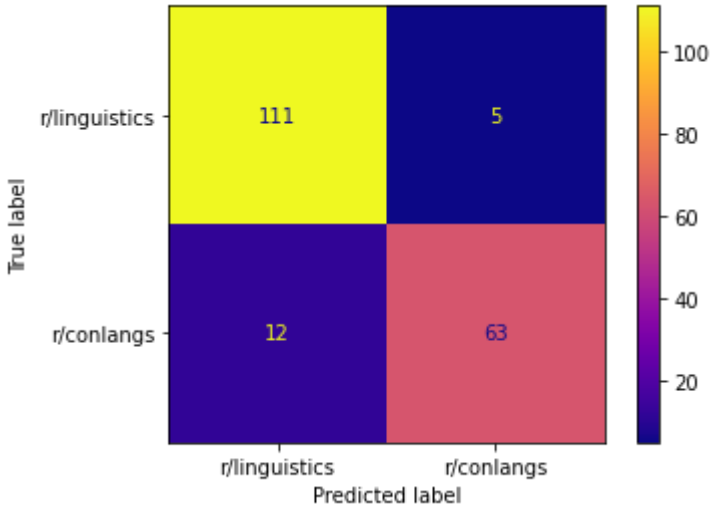
39.3% r/conlangs

Best Performing Models



Multinomial Naive Bayes

Train: 94.74% Test: 91.10%



k-NN classifier

Train: 99.82% Test: 91.10%

Findings and Areas for Interest

- Working model is possible
  - Apply model to real crowdsourced data
- Areas of improvement
  - Utilize more data (comments)
  - Maximize sensitivity (minimize false negatives)

Questions?