Classifying Refugee News Reports

Data Warehousing and Computing Lab

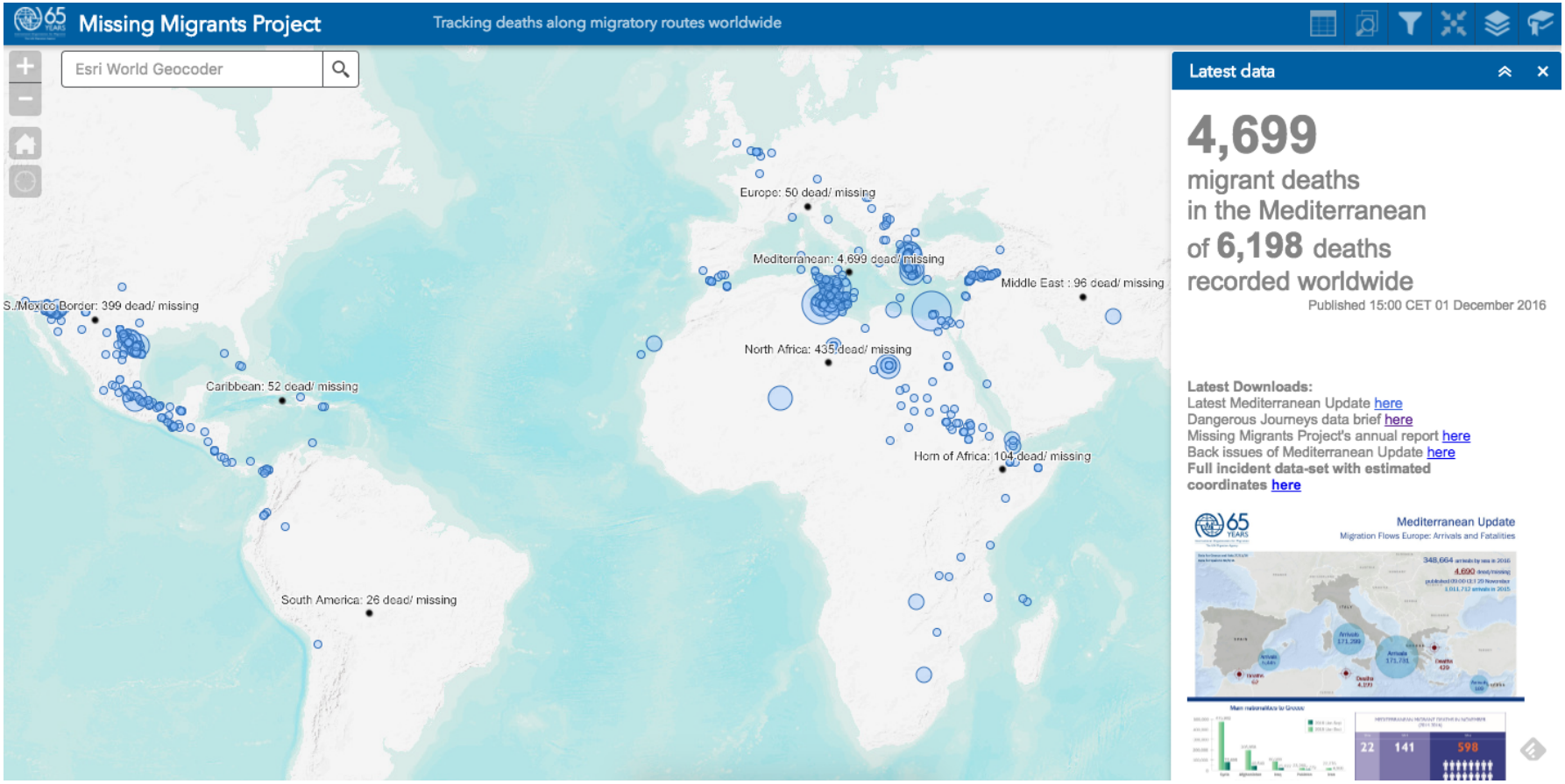
Euan Dowers, Robert Lange, Nandan Rao
Supervisors: G. Bartolozzi and C. Brownlees
Barcelona Graduate School of Economics

December 1, 2016

Outline

1. **Aim and Motivation**
   Working with IOM and Refugee News Flood
2. **Database Management**
   UI, Architecture, Persistance
3. **Text Processing**
   String Cleaning, Vectorization and tf-idf Representations
4. **Modelling the Data**
   Clustering and Cross-Validated Classifiers

Motivation - Missing Migrant Project (MMP)



▸ Link

Data Types, Challenges and a Solution

- Data Sources:
  - → Google Alert News Feeds
  - → Twitter Feeds
  - → Missing Migrant Project (MMP) data
- Data Management Challenges:
  - → One schema is not enough (different data types)
  - → Online data integration
  - → Pulling from different news feeds
- Data Management Solution: MongoDB

Architecture and Persistance

- MongoDB has several advantages:

Cleaning the strings

- Original text string - Label: Rejected

```
0  Italy Becomes A Leading Destination For Migrants, Matching Greece &amp;quot;Nobody died,&amp;quot;
 he says. With close to 160,000 arrivals this year, Italy could surpass Greece as Europe;s ma
jor migrant and refugee point of entry.
```

- Splitting text into tokens

```
0    [Italy, Becomes, A, Leading, Destination, For, Migrants, Matching, Greece, quot, Nobody,
 died, quot, he, says, With, close, to, 160,000, arrivals, this, year, Italy, could, surpass,
 Greece, as, Europe, s, major, migrant, and, refugee, point, of, entry]
```

- Removing stopwords and lemma transformation

```
0    [italy, becomes, leading, destination, migrant, matching, greece, quot, nobody, died, qu
ot, say, close, 160,000, arrival, year, italy, surpass, greece, europe, s, major, migrant, re
fugee, point, entry]
```

Constructing a Vectorized Representation

- Count Vectorization
- tf-idf: re-weighting by proportion of times a word appears in the document vs. corpus
- Deciding on dimensionality: bi-grams, tri-grams, etc.
  → Which representations do really matter?

Building a First Classification Model

- Many potential classifiers available: Logistic Regression, Naive Bayes, SVM, Decision Tree, Random Forest and NNs
- Idea: Start with MVP (minimal viable product) to grasp the problem $\rightarrow$ Generative Model: Naive Bayes

$$p(t_j|x_i) = p(x_i|t_j)p(t_j)p(x_i)$$

  - Assumption: Treat observations as iid $\rightarrow$ likelihood factorizes

$$p(x|t) = \prod_{k=1}^{d} p(x_k|t_j)$$

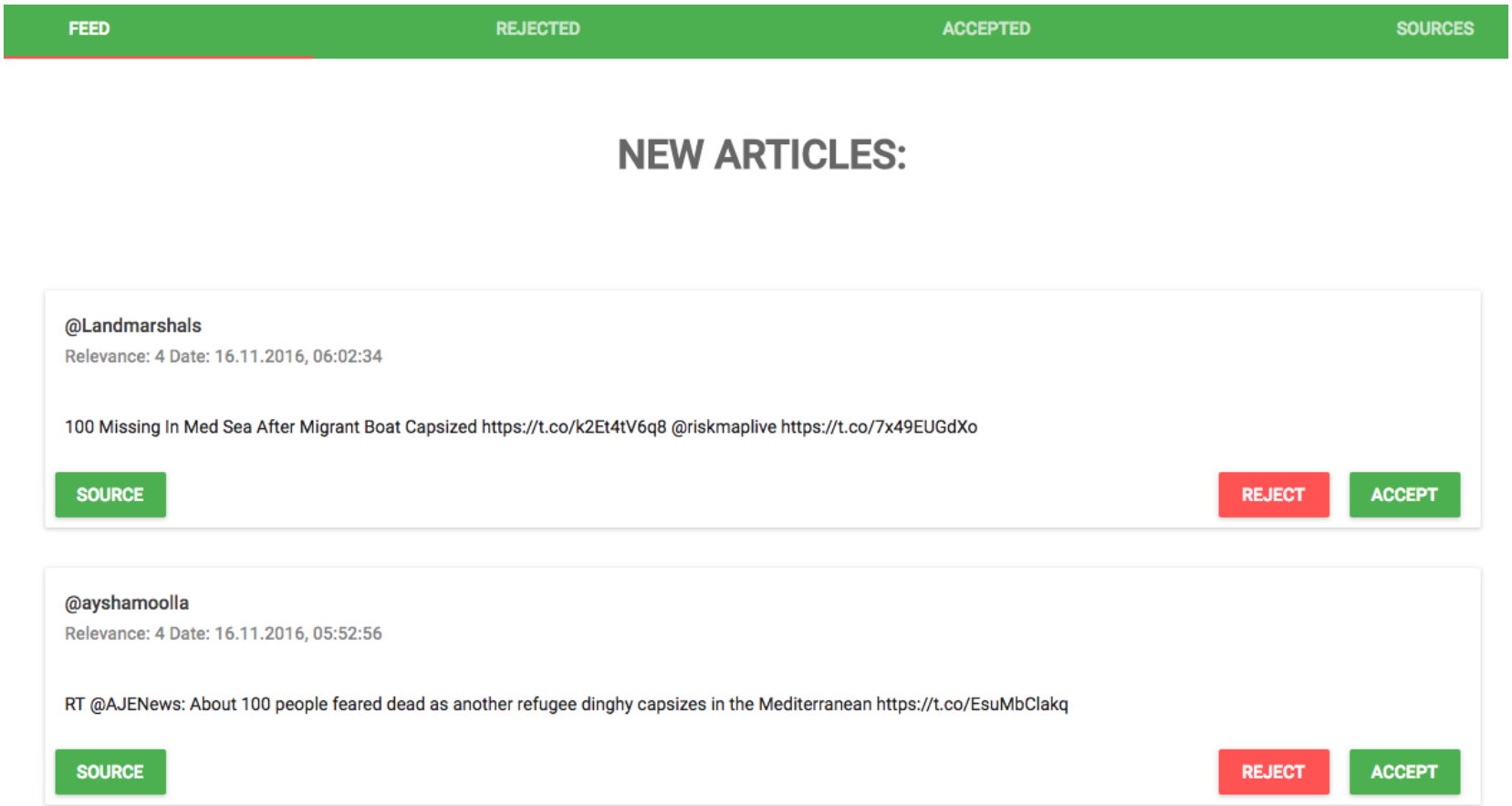
Model Evaluation

- Prior choice
- overfitting
- error evaluation depends on problem

Adding Complexity: Random Forests

- Theory

Automated Labelling Process and Lean UX



▸ Link

Clustering using DBSCAN - Density-based spatial clustering of applications with noise

- Density based clustering algorithm that takes as parameters ϵ and n minimum points in a cluster.
- DBSCAN allows points to be marked as noise - not in any cluster
- Do not need to specify number of clusters before running DBSCAN
- Intention is to cluster similar stories/tweets using vectorized presentation.

Problems and Improving Classification

- Sensitivity to specific "small" words: e.g. not
- Hyperparameter choices: 5-fold cross-validation and parameter grid search
- Adding non-parametric complexity: Random Forest

Conclusion

- Open research/work:
  1. Better understanding of the decision boundary problem
- *Any Questions?*
- *Thank you for your attention!*