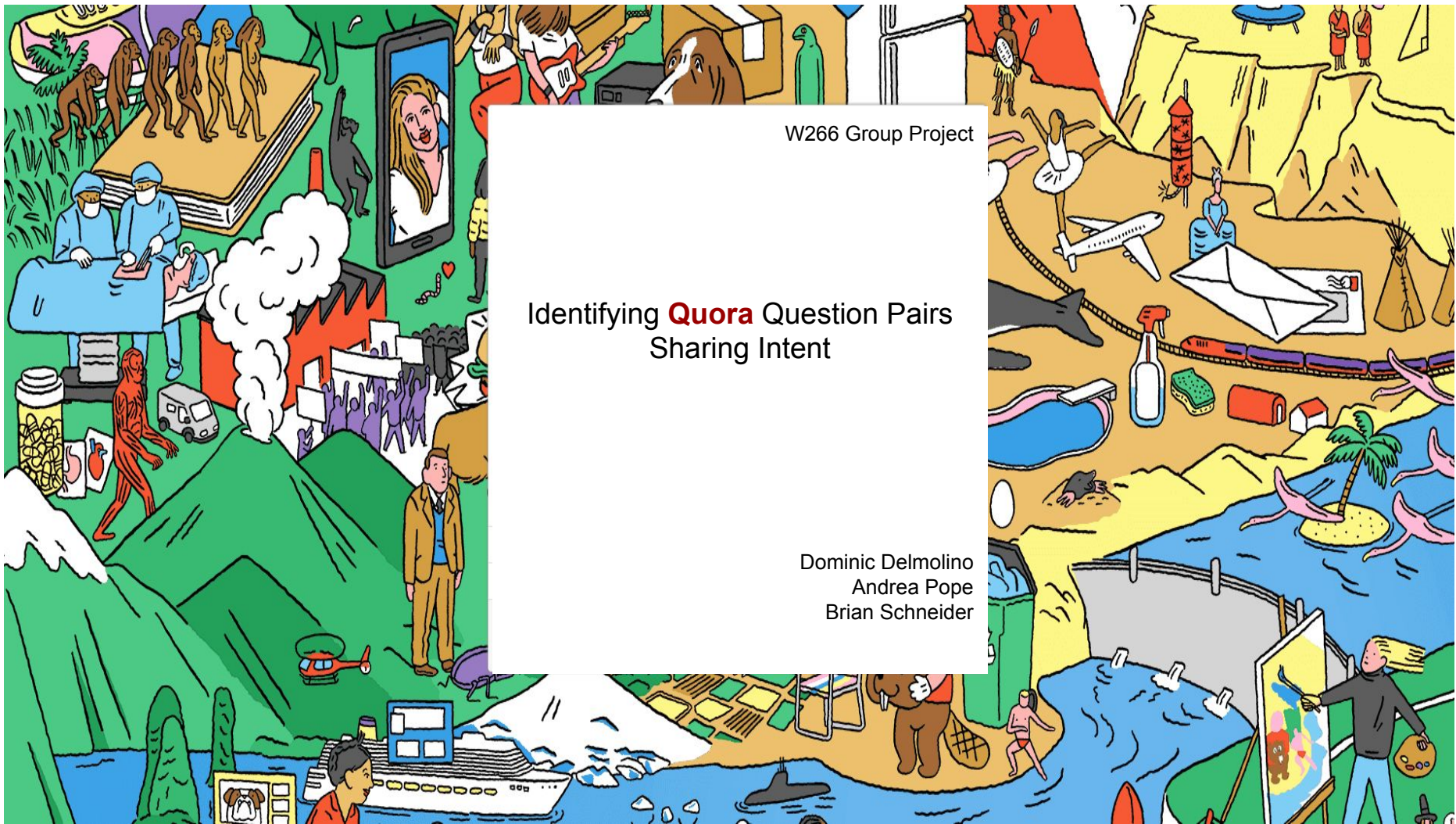


W266 Group Project

Identifying **Quora** Question Pairs Sharing Intent

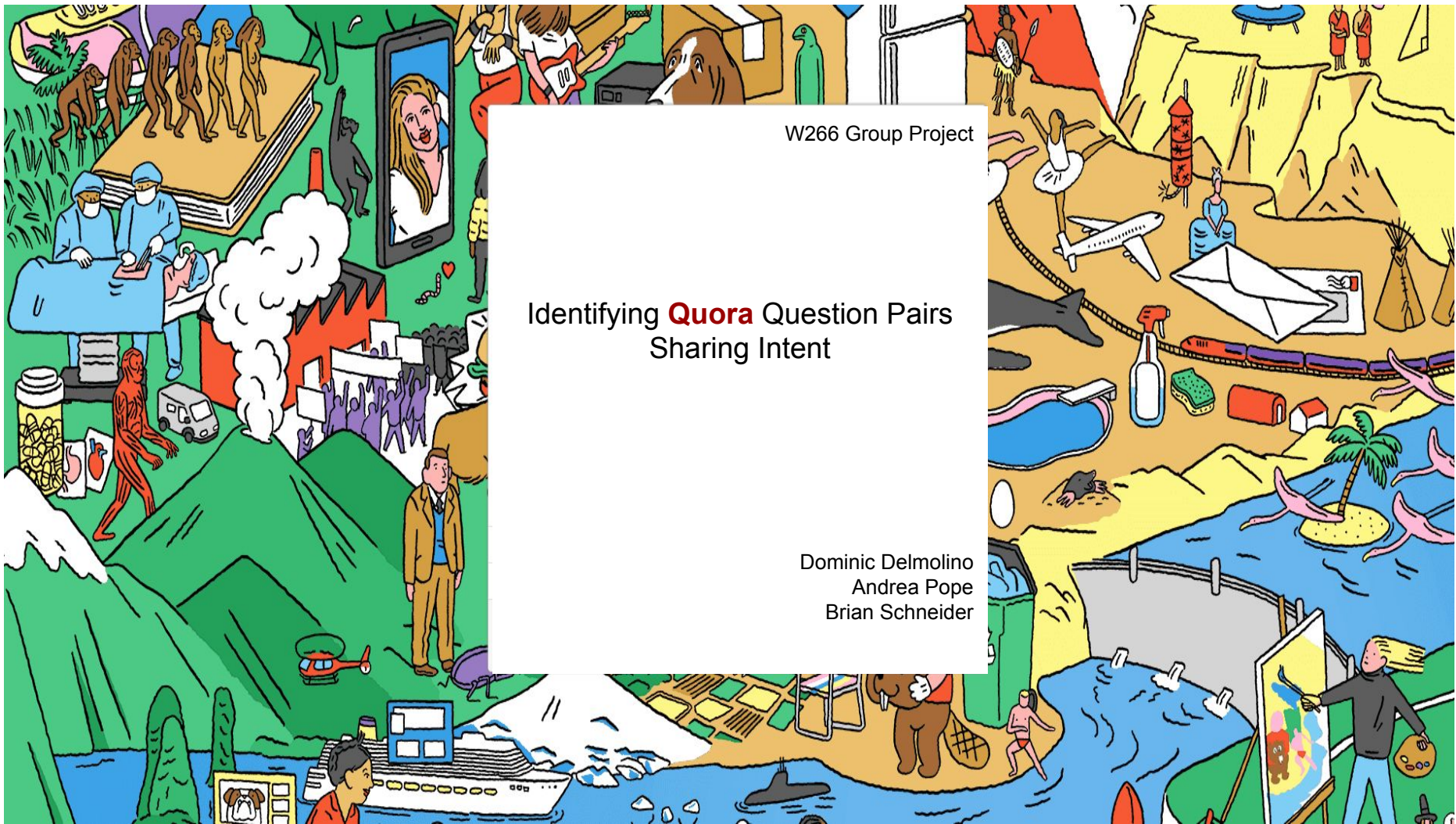
Dominic Delmolino
Andrea Pope
Brian Schneider



W266 Group Project

Identifying **Quora** Question Pairs Sharing Intent

Dominic Delmolino
Andrea Pope
Brian Schneider



W266 Group Project

Identifying **Quora** Question Pairs Sharing Intent

Dominic Delmolino
Andrea Pope
Brian Schneider

The Challenge: Kaggle - Quora Question Pairs

Question 1: How can I get a list of my Gmail accounts?

Question 2: How do I find my list of GMail addresses?

Are these questions effectively the same?

Can we use NLP techniques to help us pair like questions, to make searching for Q&A more effective on Quora?

The Adventure: Research to Execution, Simple to Not

Non-NN Models



RNN-LSTM



CNN*-RNN

Considerations:

Metric Kaggle uses log loss, we focused on accuracy

Data Enhance with POS tags?

Embeddings? What is the best approach?

Pre-trained vectors

(Co-occurrence Matrix w PPMI and SVD,
GloVe, Google)

Vectorizers

(Word2Vec, Sent2Vec, Doc2Vec)

Aids in the Adventure:

Keras: Faster model building

Layers: Facilitate dense and convolutional layers

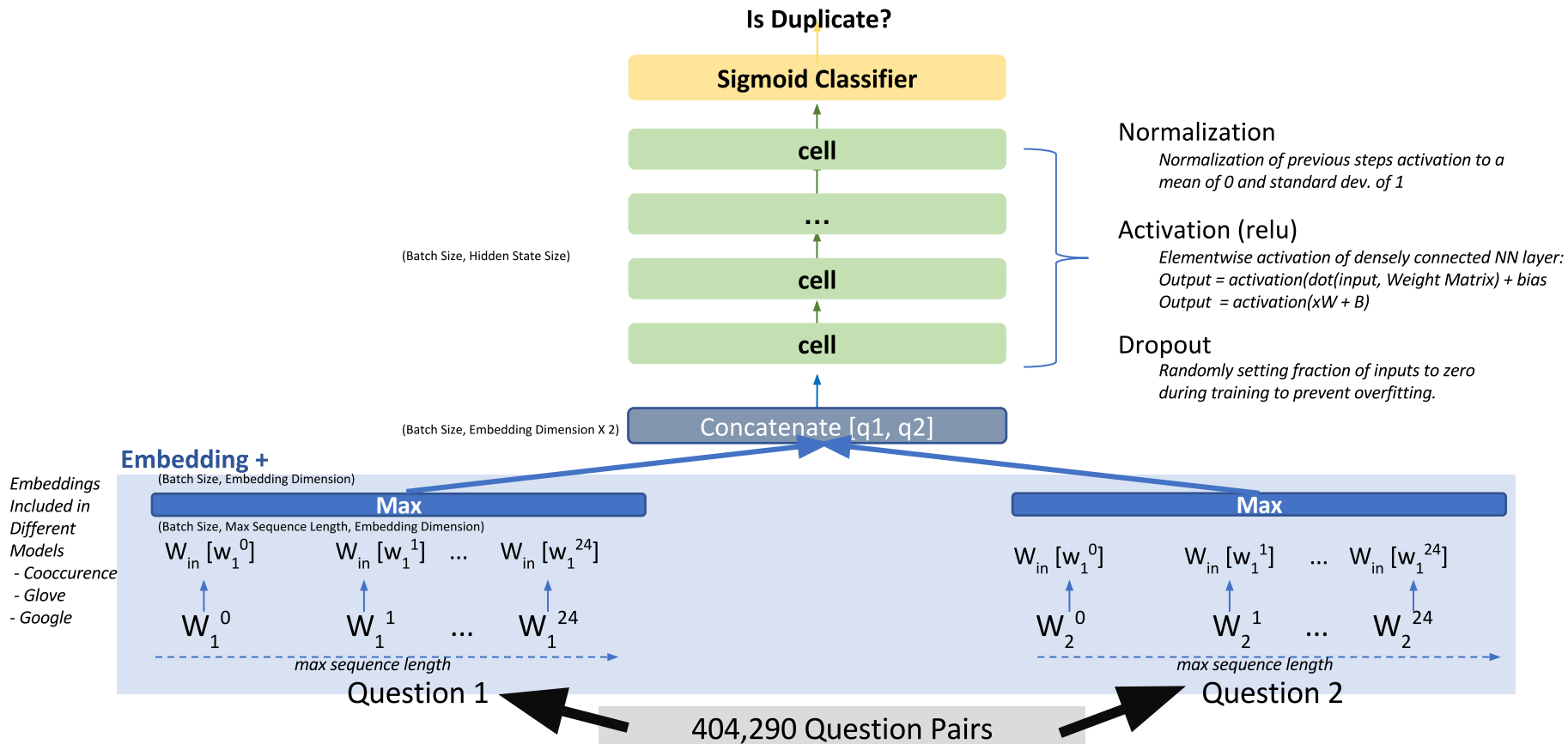
Attention: Focus on sentence sequencing

Dropout: Prevent overfitting

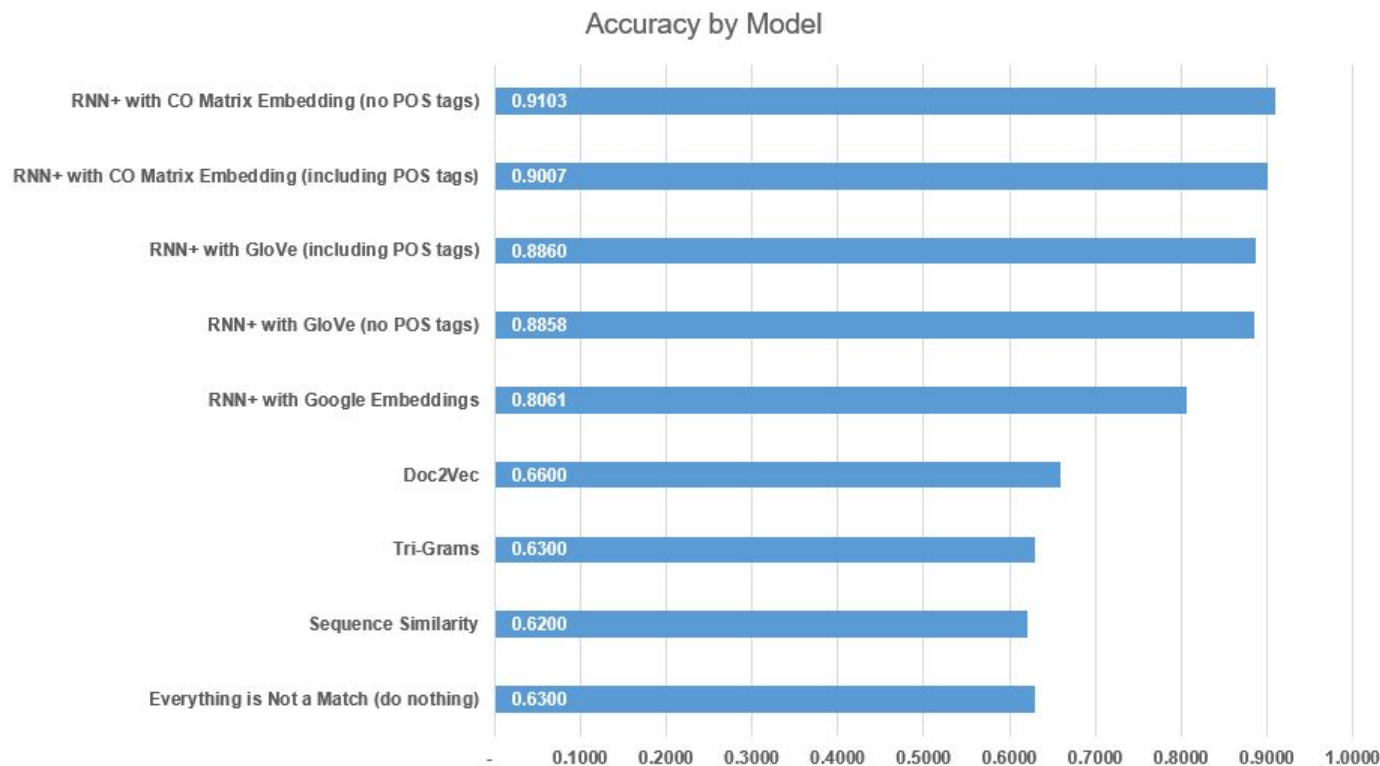


“The 10,000 Year
Clock”

The Most Successful Model



The Path: 44% improvement in training accuracy



APPENDIX

Potential Further Improvements

- Additional Feature Engineering
 - SVM & XGBoost
 - Magic features
 - A.k.a., “Last word in question1” = “Last word in question2”
- Extreme depth in layers (contest winner’s model took a week to run, and had 100s of layers at each step)
- Latent Semantic Indexing & Latent Semantic Analysis