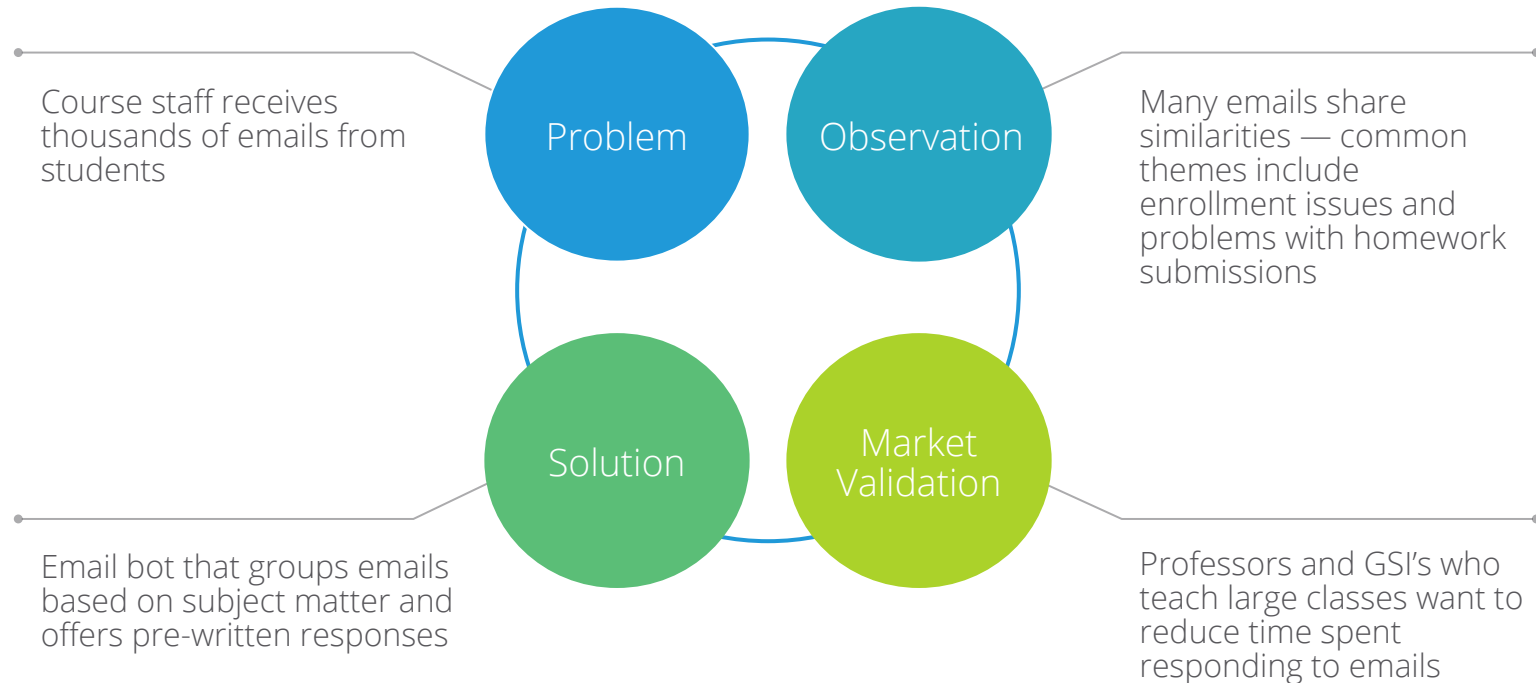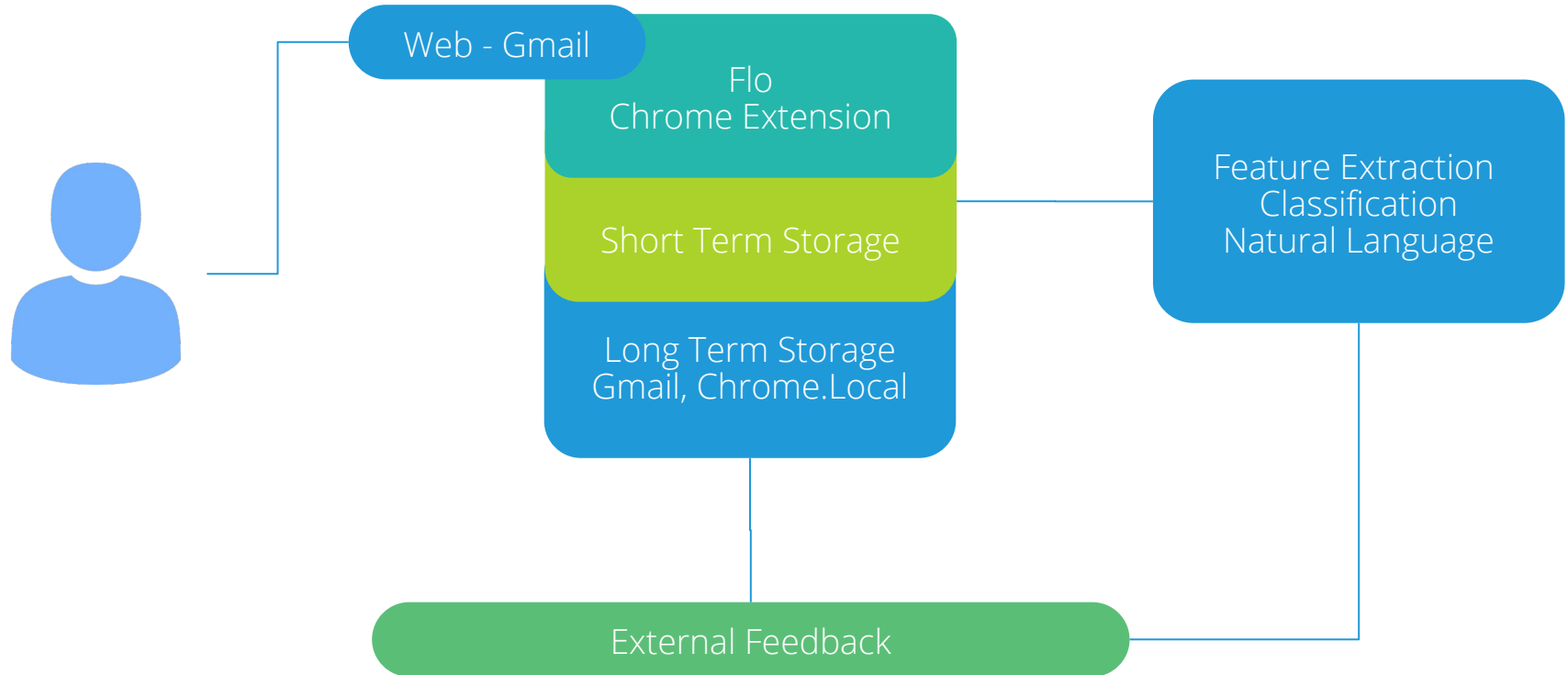# Flo

An Email Management System

Ndeye Fatou Diop, Keiko Kamei, Rohan Lageweg,
Ting Chih Lin, Joyce Siu Ying Lo, Kristian Rolland
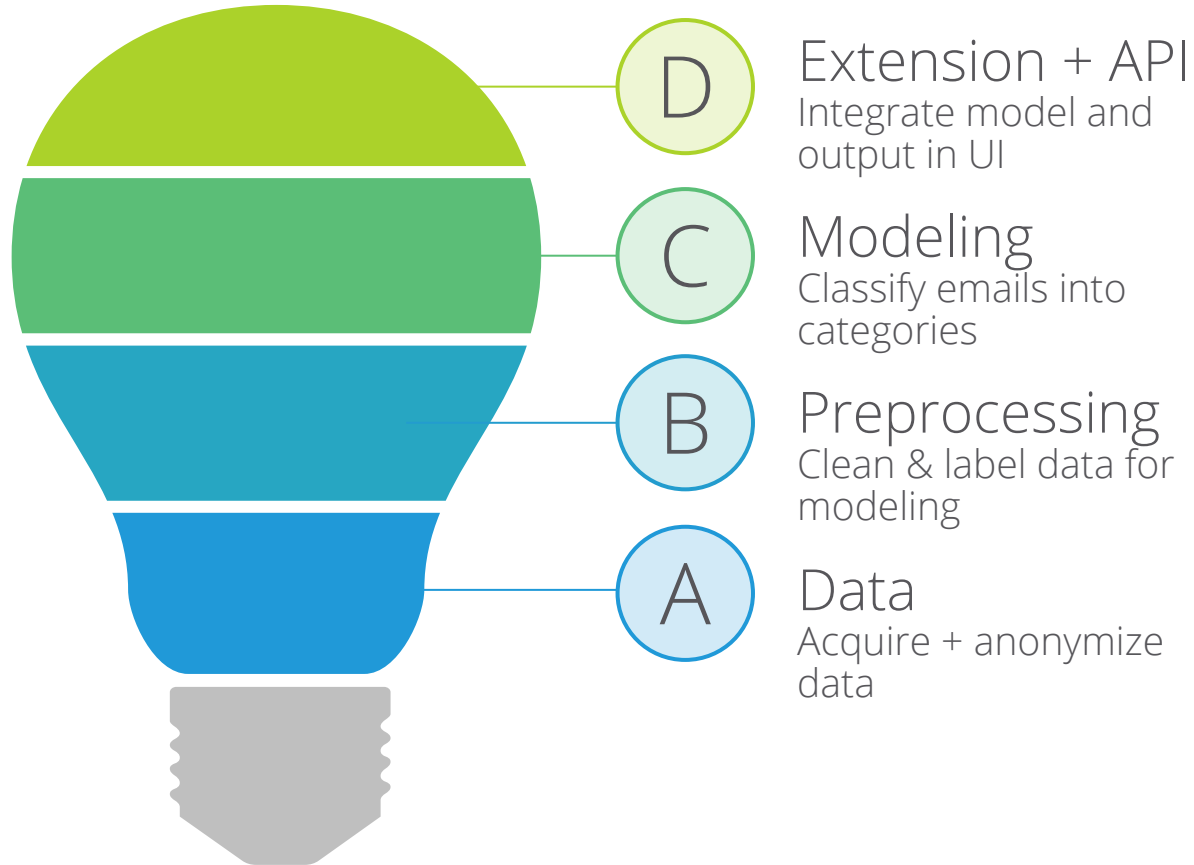
Flo: An Email Management System

Problem

Course staff receives thousands of emails from students

Observation

Many emails share similarities — common themes include enrollment issues and problems with homework submissions

Solution

Email bot that groups emails based on subject matter and offers pre-written responses

Market Validation

Professors and GSI's who teach large classes want to reduce time spent responding to emails

Architecture of Solution

Project Outline

D — Extension + API
Integrate model and output in UI

C — Modeling
Classify emails into categories

B — Preprocessing
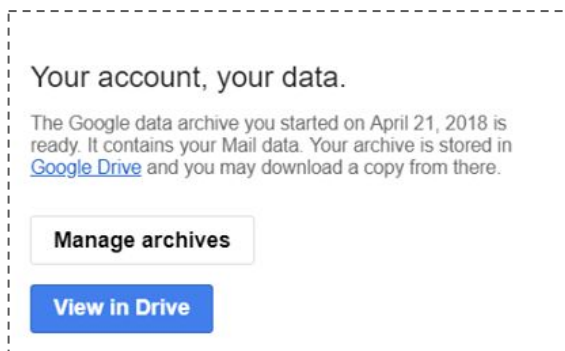Clean & label data for modeling

A — Data
Acquire + anonymize data

Data Acquisition

- With permission from course staff, downloaded MBOX file containing roughly 3 months worth of course emails

Your account, your data.

The Google data archive you started on April 21, 2018 is ready. It contains your Mail data. Your archive is stored in Google Drive and you may download a copy from there.

Manage archives

View in Drive

- Dataset:
  - Approximately 350 unique conversations
  - Equates to 1520 total course emails (including replies and forwards)

- Reached out to instructors & teaching staff of other courses but could not obtain data due to student privacy concerns
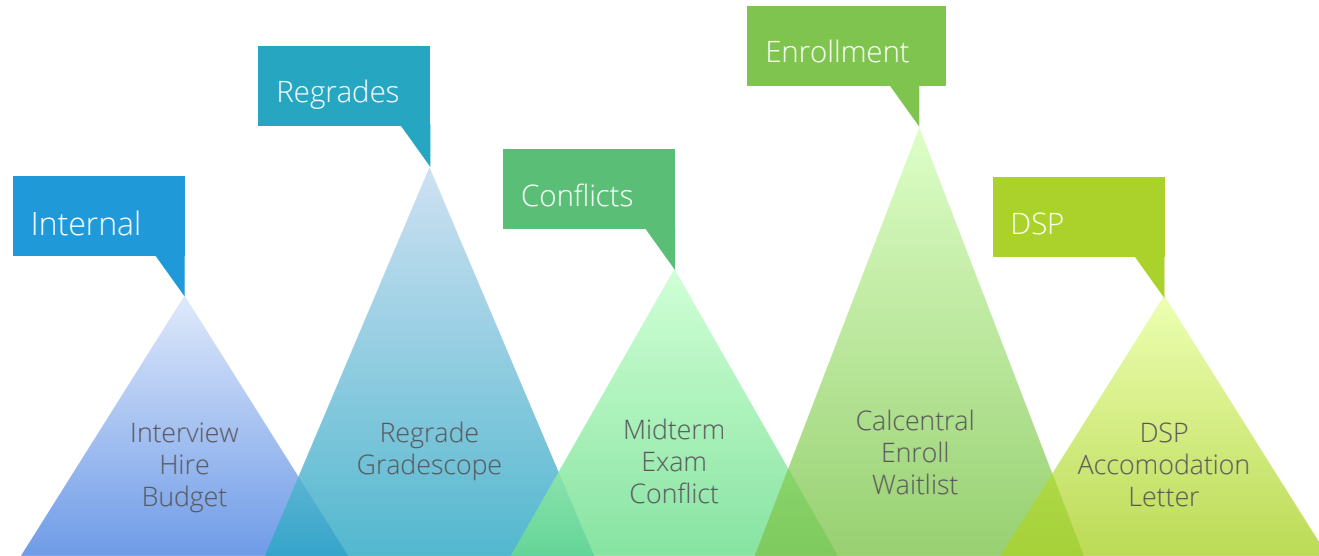
Cleaning & Anonymizing

- Removed any instance of course-identifying words/email ID's
- Filtered out indicators of forwarded messages and unwanted thread attachments via RegEx
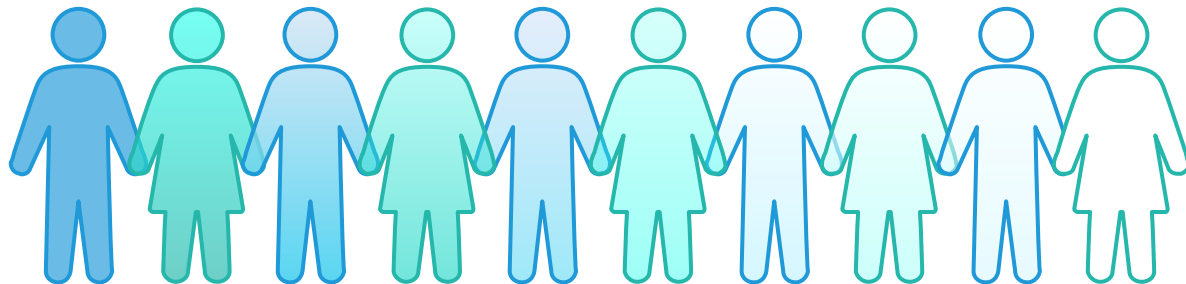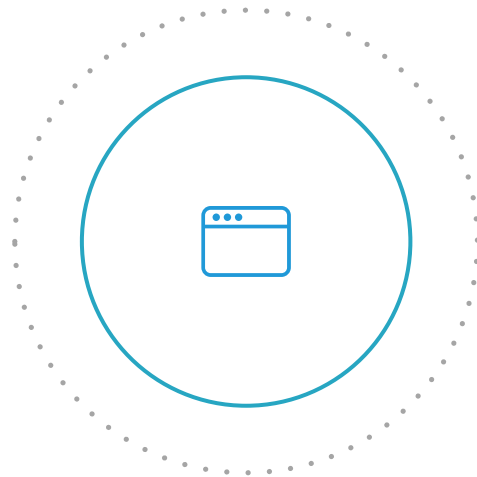
Labeling

1. Miscl.
2. Conflicts
3. Attendance
4. Assignments
5. Enrollments
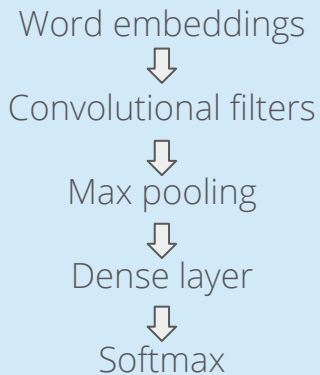6. Internal
7. DSP
8. Regrades

Our Models

Cos Classifier with LDA topics

LDA topics
⇩
Cosine Similarity

- Validation accuracy: 38.6%
- Pros: No training needed
- Cons: Depends heavily on topic vectors and too much variability

Random Forest

LDA topics
⇩
Random Forest Classifier

- Validation accuracy: 61.4%
- Pros: Easy to train
- Cons: Depends heavily on topic vectors

Cos Classifier with Word2vec

Wor2Vec vectors
⇩
Cosine Similarity

- Validation accuracy: 59.8%
- Pros: No training needed
- Cons: Depends heavily on topic vectors

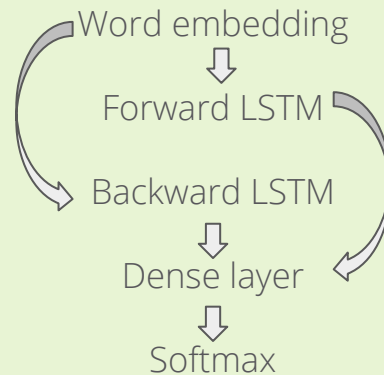Baseline accuracy : 37.9%

Our Models

Convolutional Neural Network

Word embeddings
⬇
Convolutional filters
⬇
Max pooling
⬇
Dense layer
⬇
Softmax

- Validation accuracy: 48.8%
- Pros: Easy to train
- Cons: Can't capture long-term dependencies

C-LSTM

Word embeddings
⬇
Convolutional filters
⬇
Windowed max pooling
⬇
LSTM
⬇
Dense layer
⬇
Softmax
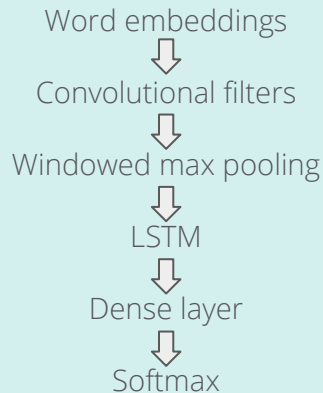
- Validation accuracy: 51.9%
- Pros: CNN can learn short-term and LSTM can capture long-term dependencies
- Cons: Operates in a single direction

Bidirectional LSTM

Word embedding
⬇
Forward LSTM

Backward LSTM
⬇
Dense layer
⬇
Softmax

- Validation accuracy: 74.1%
- Pros: Can effectively use past and future information
- Cons: Harder to train

Baseline accuracy : 37.9%

Chrome Extension

Server-side Implementation

Demo
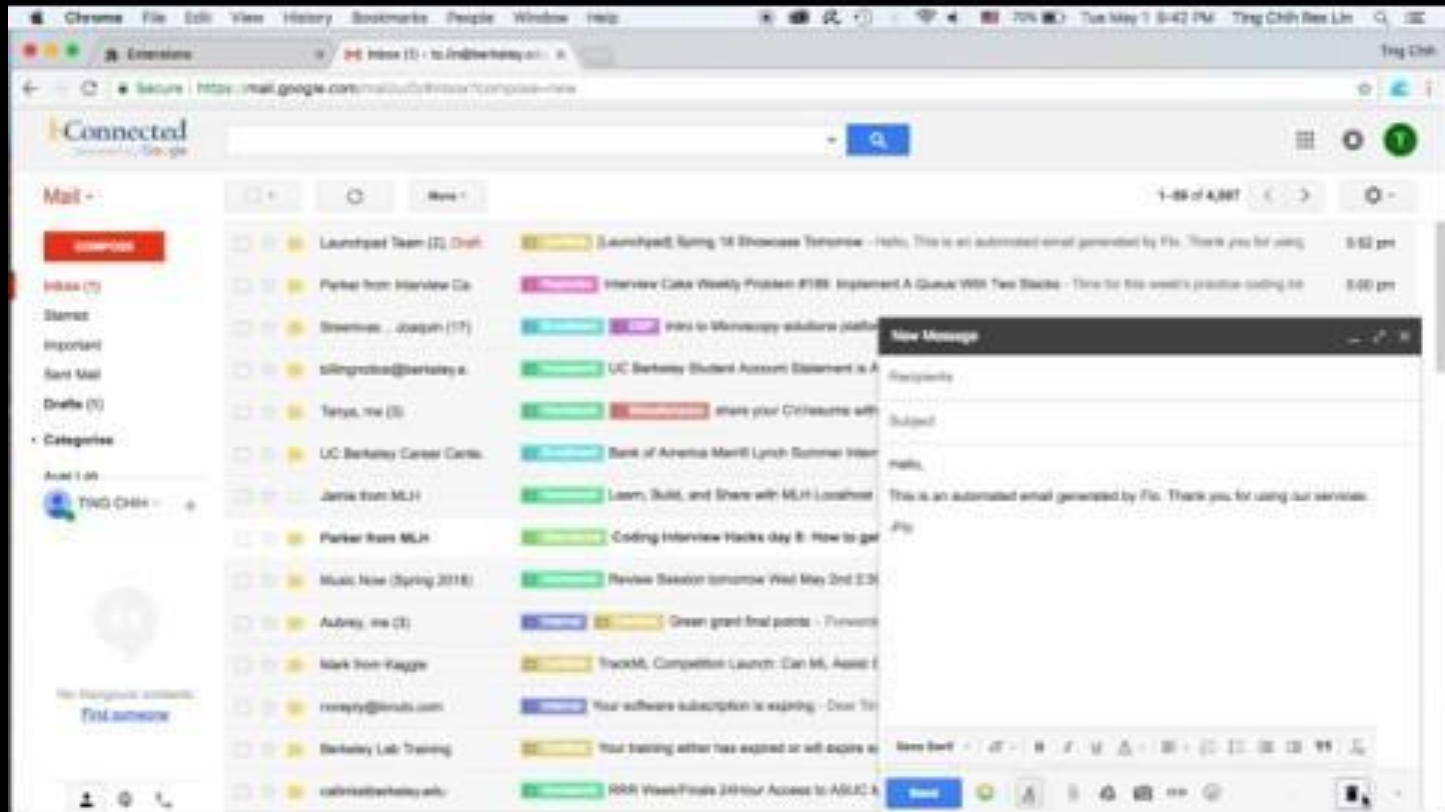
Learning Path Summary
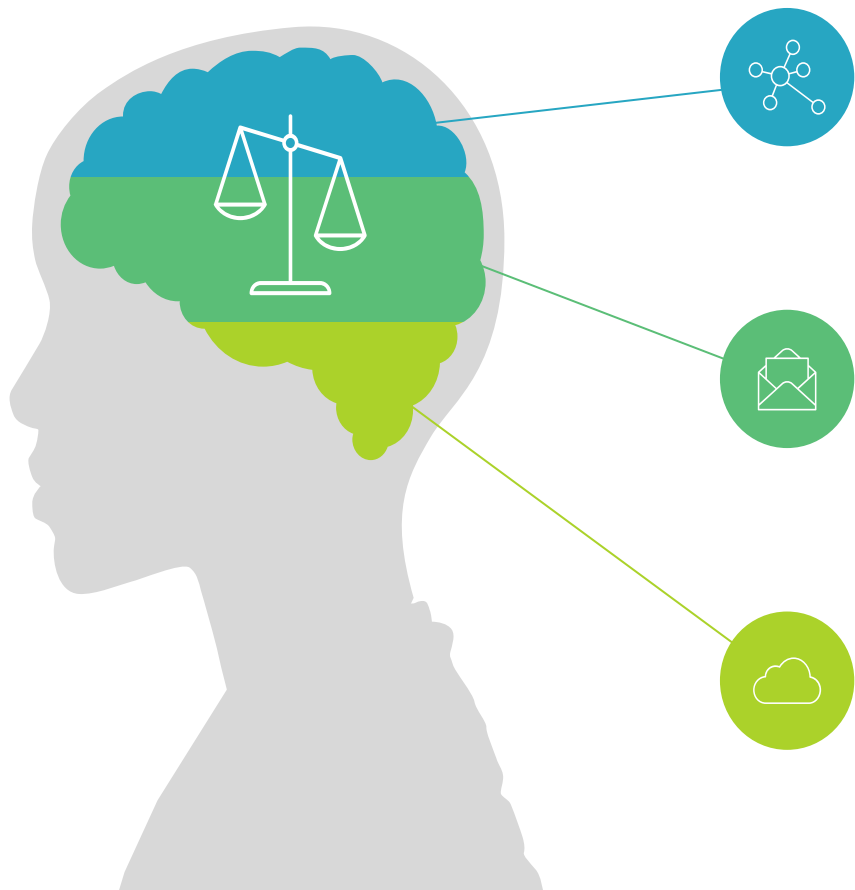
Model
- Corpus
  - Google News Vectors
- Word2Vec / Gensim
- Features of Model
  - Time correlation
  - Pre-assign weights

Email Data
- Understand & work with MBOX file type
- Extract relevant data
- Clean and anonymize
- Labeling

Chrome Extension

Looking Forward: Project Release For Summer 2018

1.   Data

- More data
- Further exploration for feature extraction
- Labeling variations

2. Models

- Improve categorization model
- Explore different models

3. Extension

- Launch model
- Feedback feature
- Advance Chrome Extension capabilities

4. Response

- Draft responses based on category

Team Flo

Ting-Chih (Rex) Lin
Fourth year MCB.

Keiko Kamei
Hi! I'm a third year studying Applied Math & Data Science and I will be a Team Lead for the Data Modules Program next semester -- Go Bears!

Rohan Lageweg
Hi! I'm a third year studying EECS/MSE and I'm fairly interested in data science. I'm currently a TA for an undisclosed course and Go Bears!

Joyce Lo
Hi, I'm a senior studying statistics and I'm passionate about data science. I'm currently a GSI for DS100 and I'll be working at Facebook after graduation!

Ndeye Fatou Diop
I am a Master of Engineering in IEOR and did my undergrad in France in Applied Math & Computer Science. I plan to work as a Software Engineer after graduation.

Kristian Rolland
Hi, I'm a junior studying Economics at Cal. My other interests include data science, coding, and music producing!

Notable References

- GloVe: Global Vectors for Word Representation
  Stanford Word Vector Paper Published In 2014
  Authors: Jeffrey Pennington, Richard Socher, Christopher D. Manning

- Smart Reply: Automated Response Suggestion For Email
  A Google Paper Published in June 2016
  Authors: Anjuli KAnnan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs Marina Gaenea

- Multi-Task Sequence To Sequence Learning
  A Google Brain Paper Published From March 2016

  Authors: Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, Lukasz Kaiser