# Regulations/Controls Mapping Automation with AI

Data Science Institute
COLUMBIA UNIVERSITY

**Patrick Kwon, Jong Hyuk Lee, Pranav Alva, Yinan Ling**
**Roshan Rao, Junghoon Woo, Akshat Palntikar**
**Sining Chen**

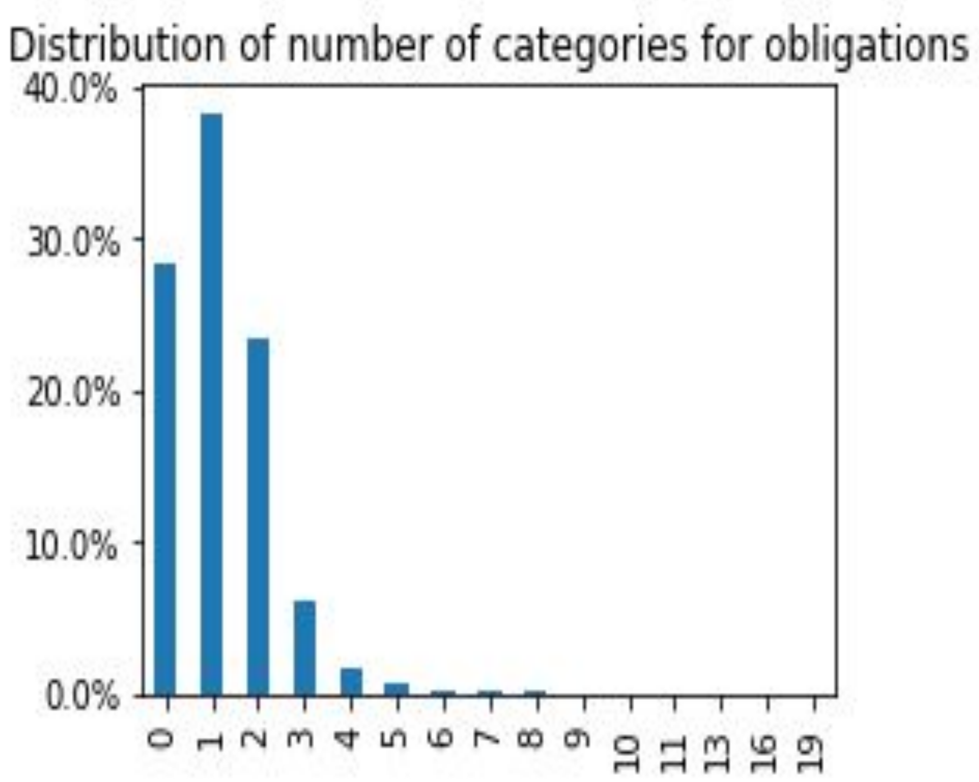
**Data Science Capstone Project with KPMG Lighthouse**

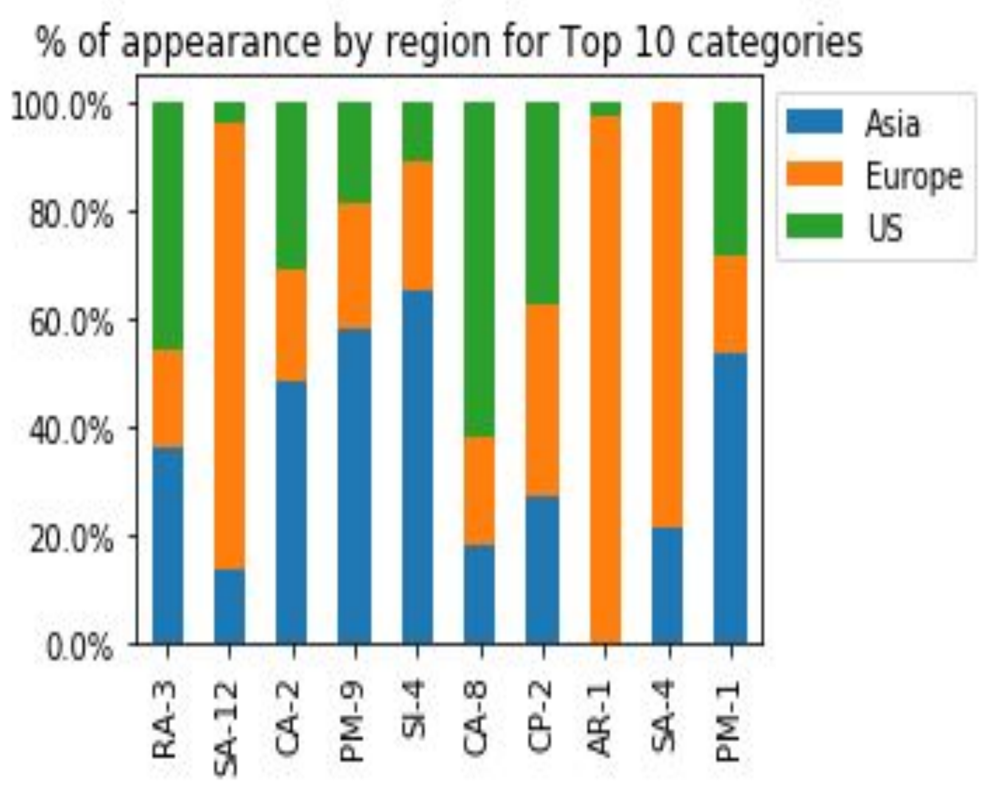
## Project Description

**Deliver machine learning capabilities that automatically maps requirements from a number of cyber privacy and information security regulations to the security controls and associated assessment procedures defined in National Institute of Standards and Technology (NIST) Special Publication 800-53**

## Data Analysis

**Most mapped obligations have only 1 or 2 regulation categories. Some categories tend to be dominated by a certain region, while others are more distributed. When analyzed by text, words like "information" and "data" are widely used by many obligations, and "maturity level" are rather dominated by a single region.**




| Word (unigram) | count | Word (bigram) | count |
|---|---|---|---|
| information | 488 | maturity level | 367 |
| data | 466 | personal data | 170 |
| level | 421 | level baseline | 166 |
| shall | 403 | service provider | 147 |
| maturity | 368 | level intermediate | 104 |

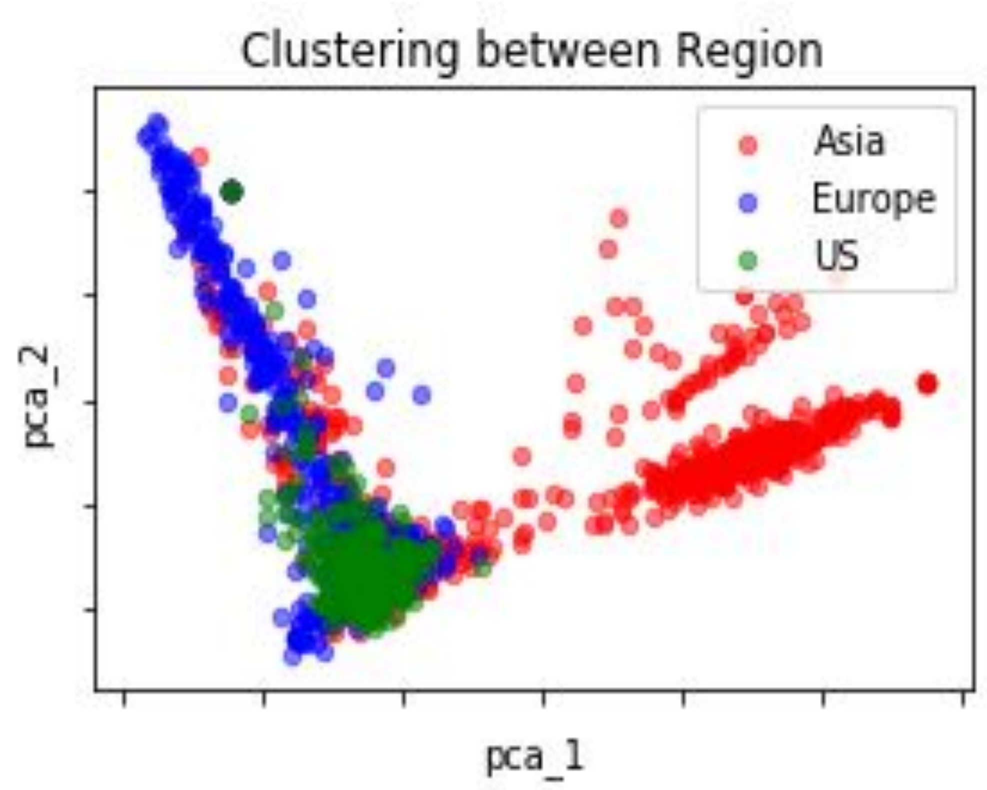
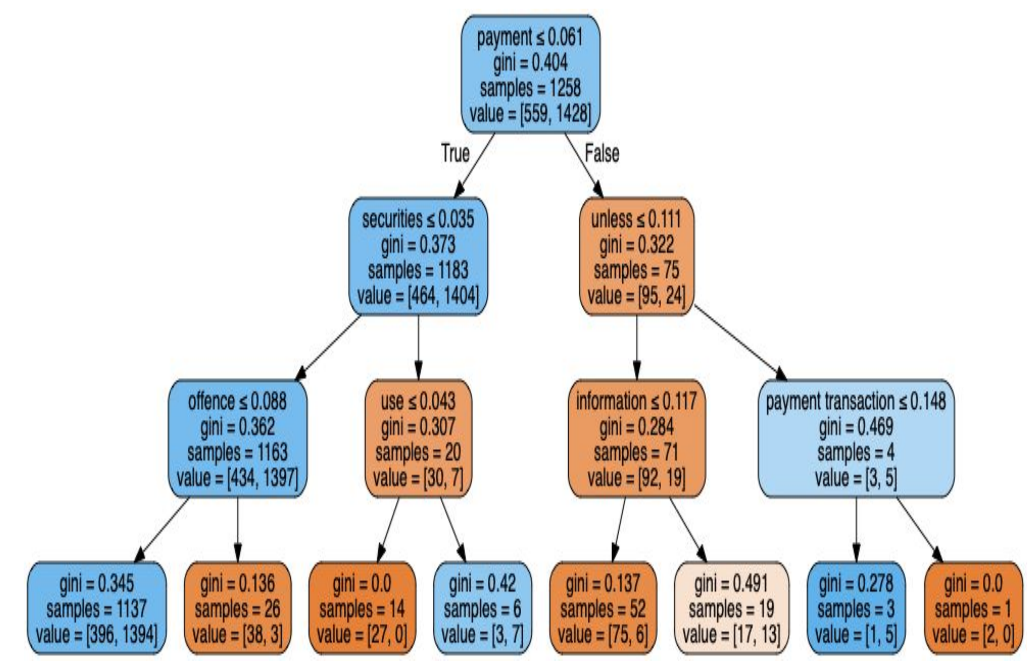



**Figure 2-1 : Distribution of number of categories for obligation**
**Figure 2-2: % of appearance by region for Top 10 categories**
**Figure 2-3: Top 5 Unigram/Bigrams on the obligations**
**Figure 2-4: PCA Clustering graph for obligations, based on region**

## Data Classification

**The obligations are encoded, and fed into machine-learning based classification methods. For the Map/No Map classification, glove embedding vectors/TFIDF vectors and random forest classification methods are used. The resulting recall was 97.4%**

**For the Category classification, spacy embedding vectors and a bi-direcitonal LSTM model (based on tensorflow-keras) was used. The resulting accuracy was 97.9% (1,356 predictions exactly matched the actual mapped categories, among 1,385 rows)**



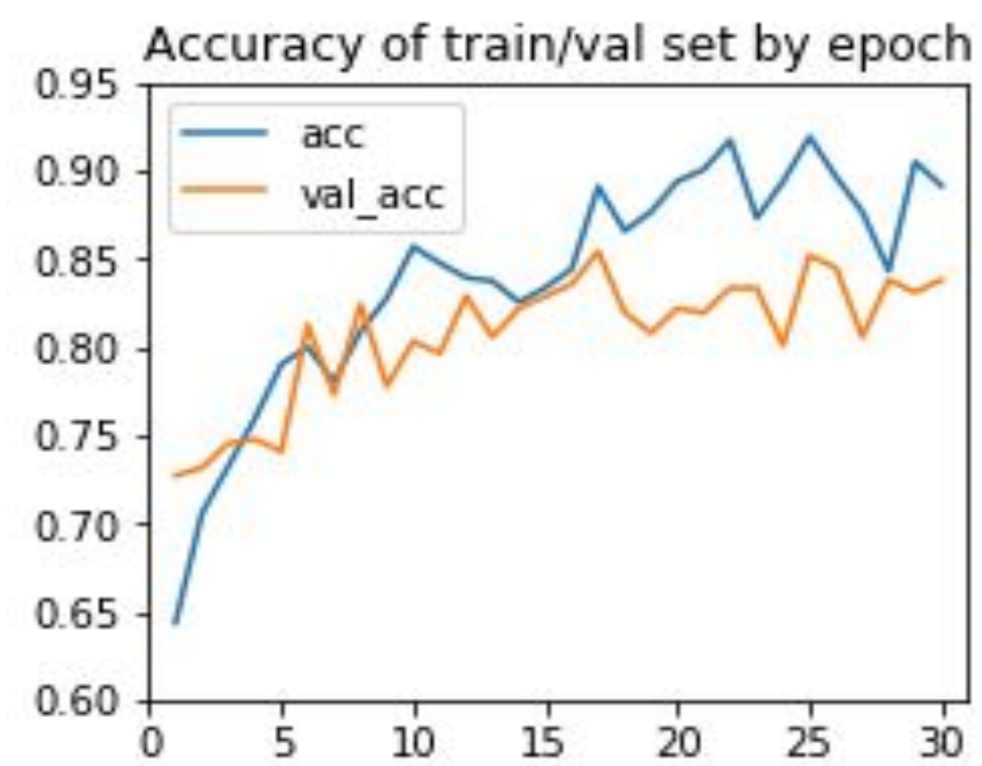



**Figure 3-1: Visualization of tree for Map/No Map classification with TFIDF vector**
**Figure 3-2: Accuracy graph for train/validation set along 30 epocs.**

## Conclusion

**The best performing obligation and category classification models both reach accuracy of over 97%. Alternative embedding/modeling methods such as BERT Embedding vectors, released by Google in 2018, and Attention Models with RNN might help improving the accuracy/recall better.**

## Acknowledgments (Calibri, 36 points, bold)

**We would like to express our great appreciation to the KPMG Lighthouse team and Sining for their valuable feedback on the project.**

## References (Calibri, 36 points, bold)

**National Institute of Standards and Technology Special Publication 800-53, Revision 4**