JOHN TAMANAS

WHAT YOUR GGI COFFEE MUG SAYS ABOUT YOU

DATA OVERVIEW

- 50 Participants (45 students, 5 professors)
- 23 Offices
- 21 Home Countries (across 6 continents)
- 14 Institution Countries
- 5 Mug Colors
  - Black, Yellow, Orange, Green, Red

DEALING WITH CATEGORICAL DATA

- There is not a well-defined way of determining distances between points
- Map the category, labels to a high dimensional vector space

STANDARD EMBEDDINGS

- One-hot-encoding is uninformative
- Word2Vec is trained on semantics
- Difficult to implement

CATEGORICAL ➔ NUMERICAL DATA

- Solution: Create new embedding using various national metrics
  - Capital's Distance to Florence
  - Population
  - Medal Points for the 2000 Summer Olympics in Sydney
  - Average years in school after age 15
  - Average number of hours on internet per year per person

CATEGORICAL ➔ NUMERICAL DATA (CONT.)

- Number of radios per person
- Number of newspapers
- Percentage of people between 15-64 years old
- Ratio of birth rates of males to females
- Length of country's full name
- Length of country's primary language's name

ANALYSES

K-MEANS CLUSTERING

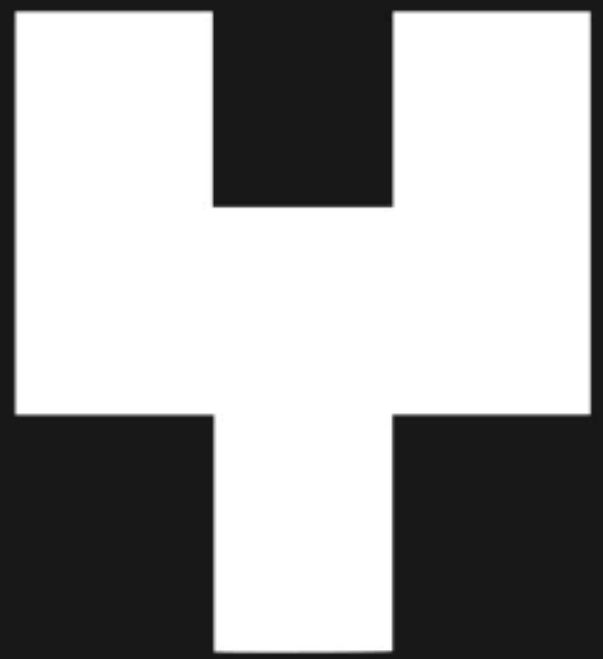
- Unsupervised
- Cluster data in embedded vector space into *k* groups
- Isotropic
  - Must normalize data
  - Try to limit degeneracy between data points

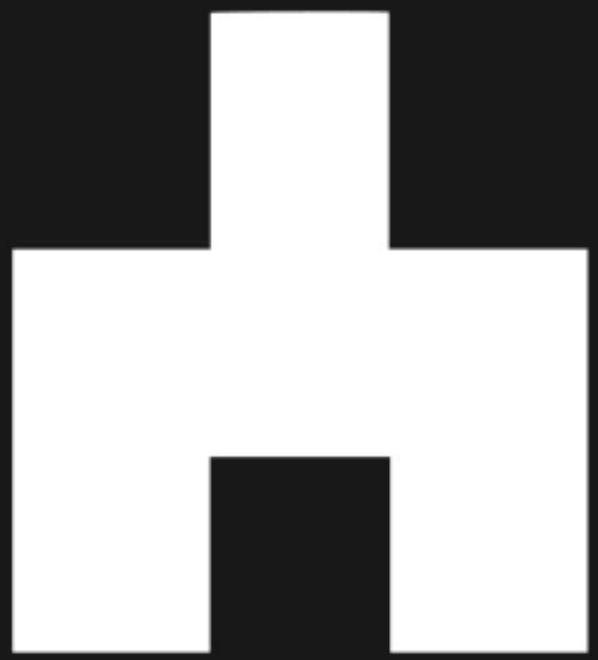
DEEP NEURAL NETWORK

- Olfactory Neural Network
- 2 hidden layers
- ~10^6 parameters

https://www.quantamagazine.org/new-ai-strategy-mimics-how-brains-learn-to-smell-20180918/

DECISION TREES



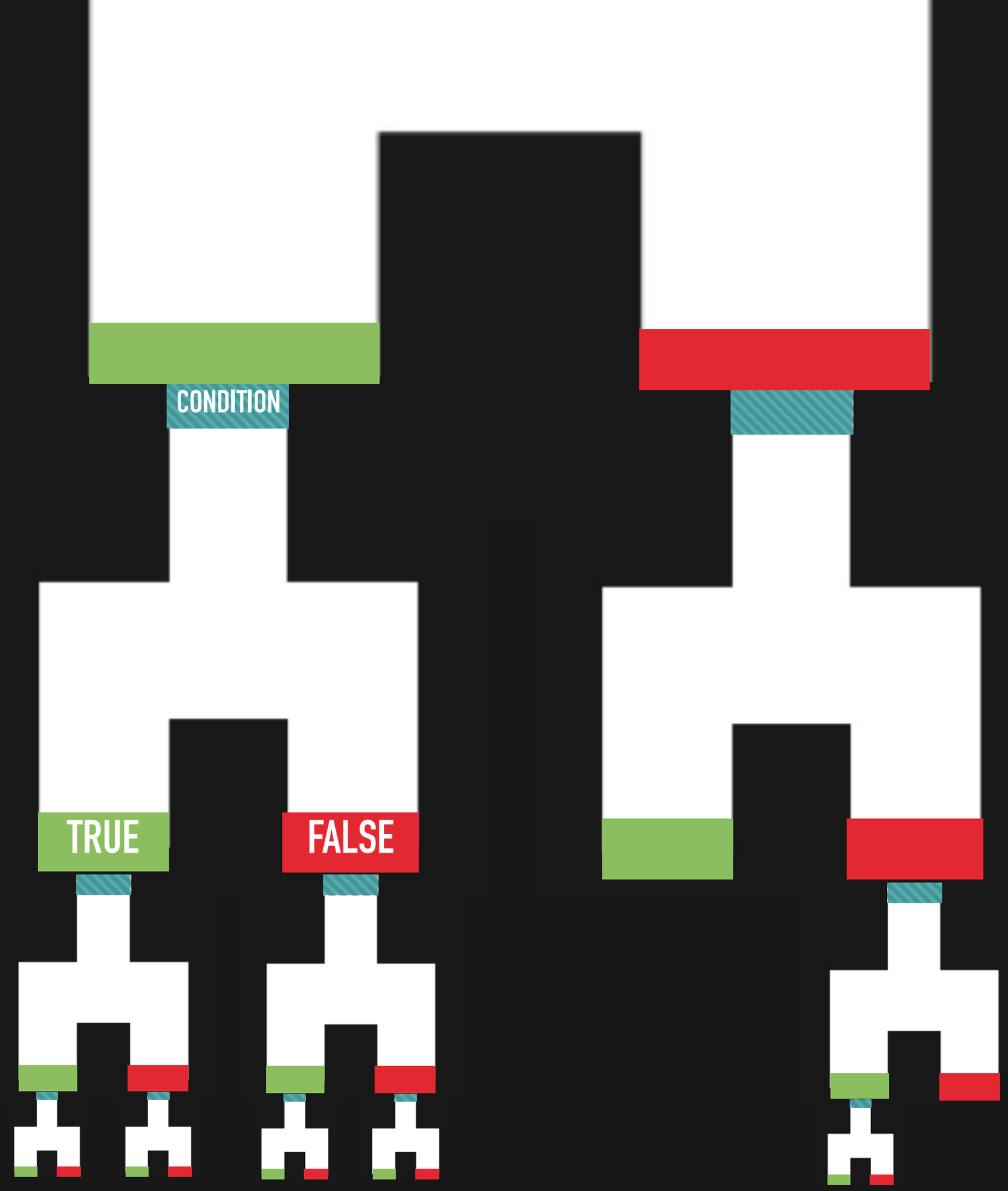


Continue?

YES

H*CK YEAH!!!

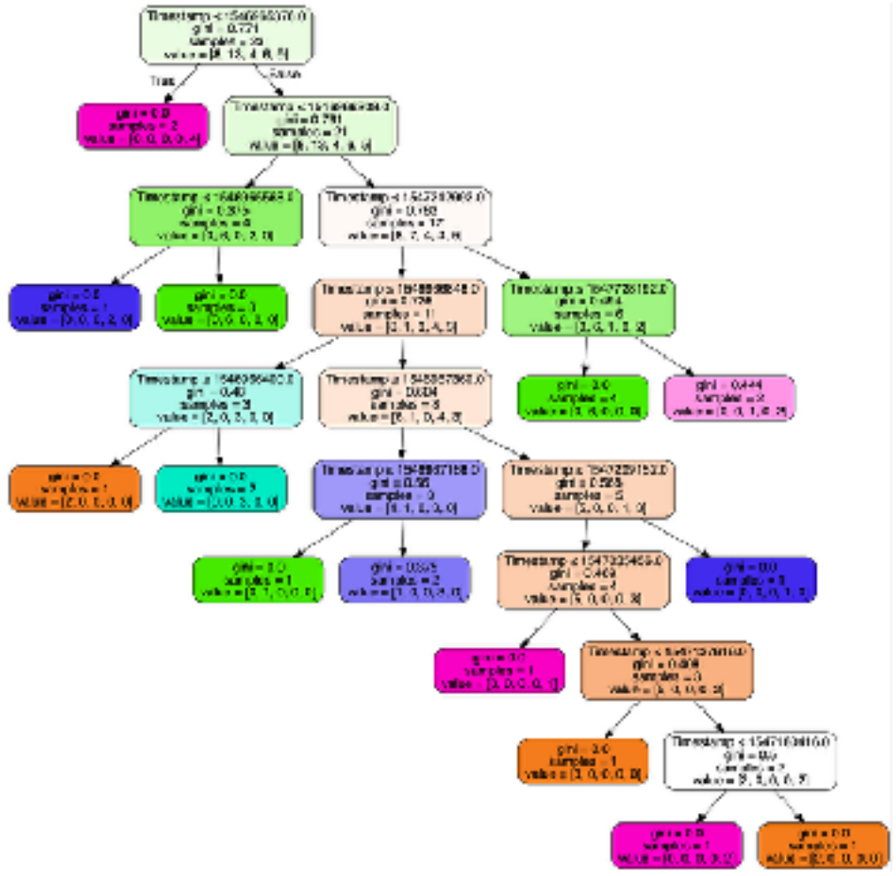
DECISION TREE



RESULTS

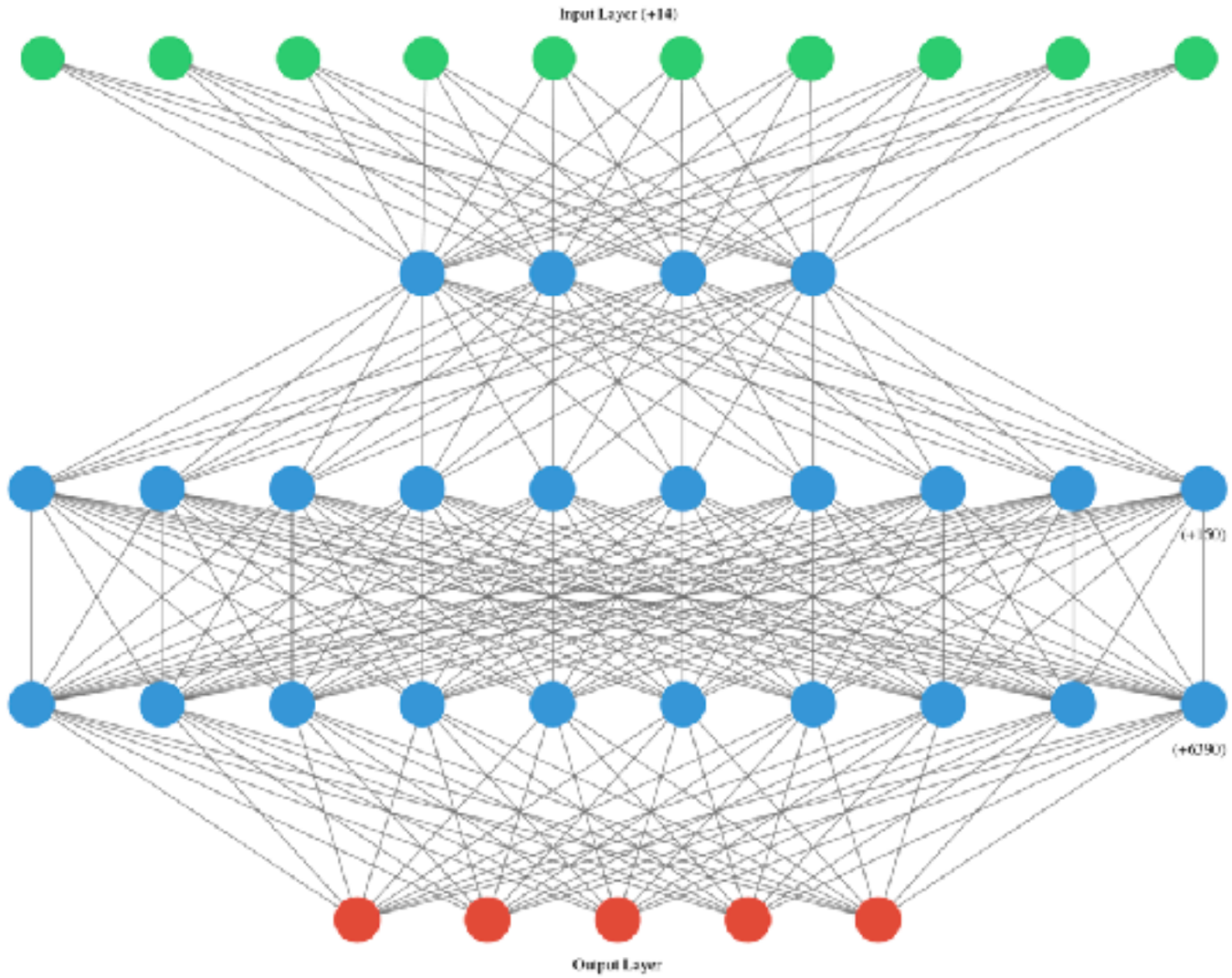
RANDOM FORESTS

- Compared 100 different random forests on real data vs shuffled data
- Found one significant result:
  - Institution's country's male-to-female birth ratio
    - ~14% accurate



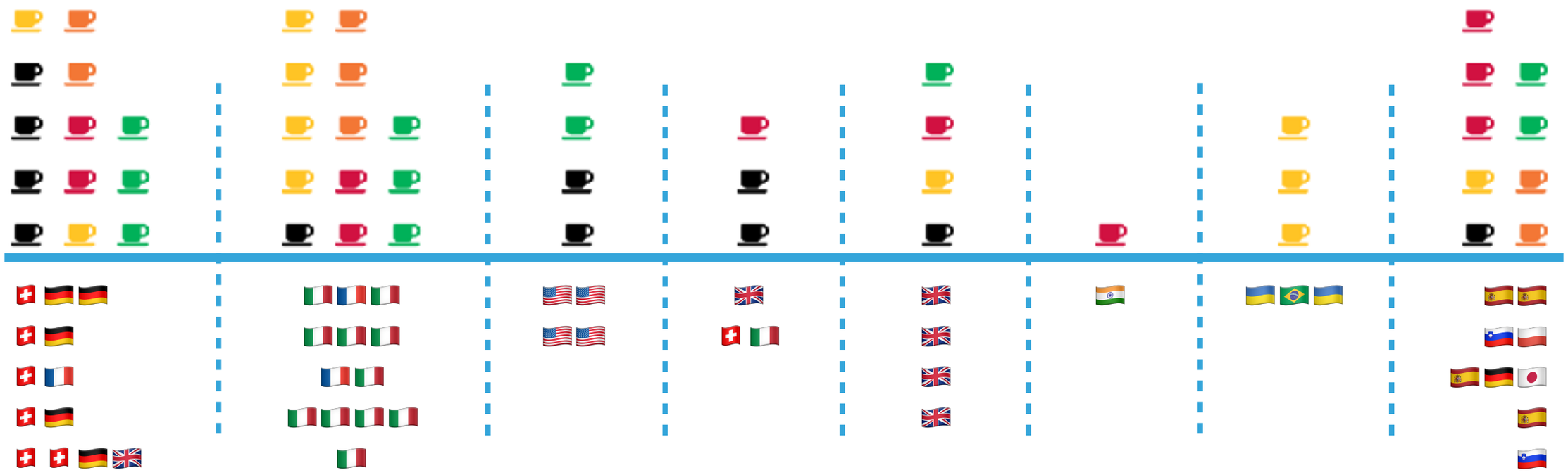
DEEP NEURAL NETWORK

- KFold Cross Validation
  - Randomize training sample
  - Evaluate on test samples
- (20 ± 14) % accuracy



K-MEANS CLUSTERING

- Divided into 8 different clusters
  - Primarily based on place of study (makes sense)



CONCLUSIONS AND FUTURE WORK

THANK YOU