Machine Learning applied to Planetary Sciences

PTYS 595B/495B

Leon Palafox

Final Project

- Final Project will be 60% of the Final Evaluation.
  - The final project will consist in the correct use and validation of one Machine Learning technique in a dataset related to the student field of study.
  - The students can form teams of up to **three persons**.
  - Graduate students and teams need to write an 8-page report on the data, methodology and conclusions. In addition, make a 10-minute presentation of their work.

Books

- Bishop, Christopher M. *Pattern recognition and machine learning.* Springer, 2006.
- Rogers, Simon, and Mark Girolami. *A first course in machine learning*. CRC Press, 2011. (http://www.dcs.gla.ac.uk/~srogers/firstcourseml/ , Available Online at the UA Library webpage)
- James, Gareth, et al. *An introduction to statistical learning.* New York: Springer, 2013. (http://www-bcf.usc.edu/~gareth/ISL/)
- Petersen, Kaare Brandt, and Michael Syskind Pedersen. *The matrix cookbook*. Technical University of Denmark 7 (2008): 15. (https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf)

Questions?

Group Activity

1st Message

- Experience
- Secretary
- State
- Foreign
- Policy
- Email scandal

2nd Message

- Rapists
- Muslims
- Great
- Winning
- Huge

How did you know?

- Frequent words are associated with each candidate
- Your brain does the correlation between words and candidates
- Your brain calculates the joint probability that a certain candidate is associated with all the words.

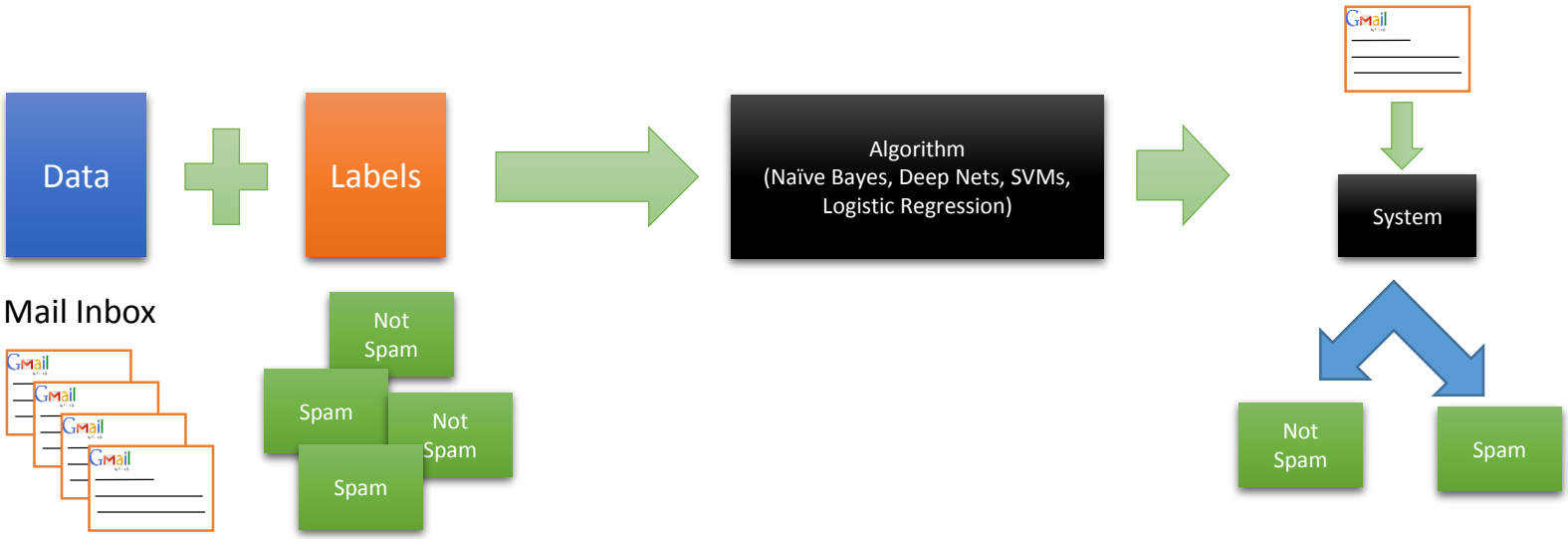
Rules of the land

- Data:
  - Documents - > Text
  - Images - > Pixels
  - Songs -> notes, tones
- Features:
  - Text -> Strings: hi, ho, friend, help
  - Images -> RGB colors, DN, grayscale, floats
  - Tones -> float values that represents tone.

Supervised training

- Set of labeled data:
  - Set of emails with spam/not spam tags.
  - Amazon reviews (Stars)
  - Facebook status like/not like.
  - Stock Market Prices - > Volume
- Algorithm
  - Linear Regression
  - Logistic Regression
  - Support Vector Machines
  - Deep Learning (Neural Networks and Convolutional NN)
- Classification tool

Supervised Learning



Each category (spam, not spam) will have features that will characterize them.

Spam: Offer, Viagra, medicine, Free, Conference in China

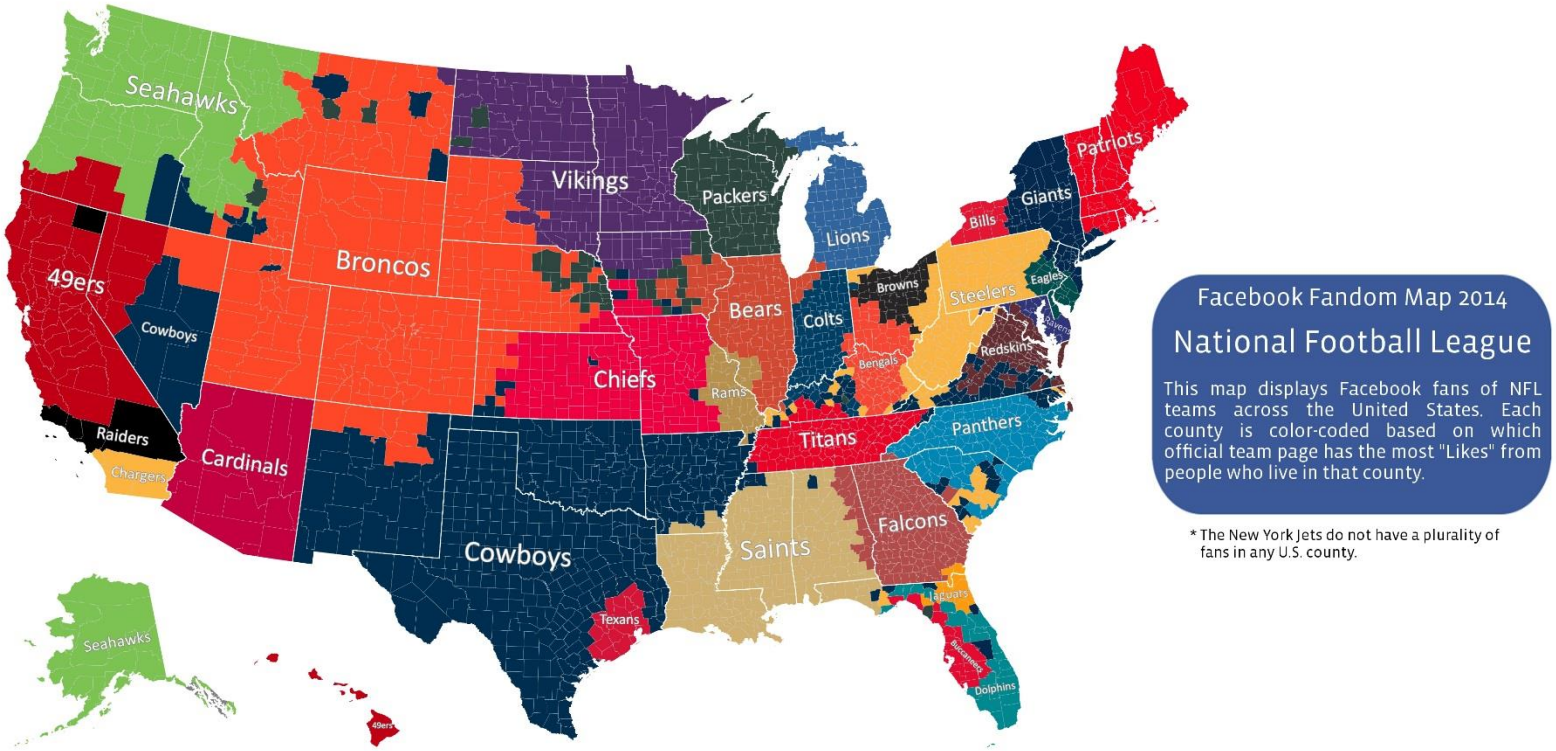
Not Spam: Hamilton, LPL, DTM, Mom, Dad

Validation Methods

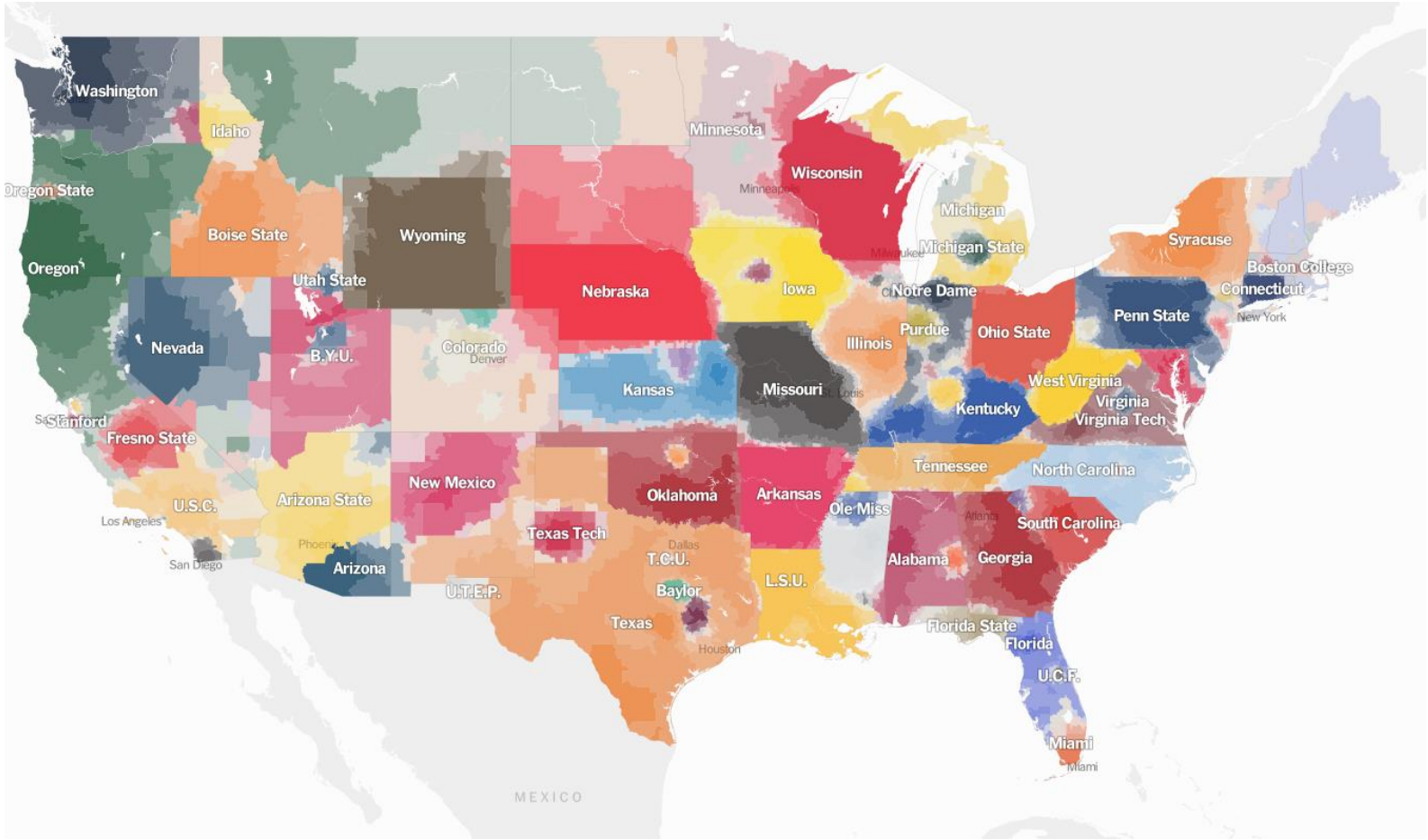
- Cross Validation
  - Prevent overfitting.
  - Find the best set of parameters.

- Bias-Variance Analysis
  - *“The needs of the Many outweigh the needs of the few”*
    - *Spock*
  - You don’t want to tell someone they have cancer, but you **really** don’t want to tell them they don’t if they do.

Group Activity



Group Activity



Knowledge discovery

- We don’t need labels to discover patterns.
- Data itself organizes (closeness)
- Most algorithms just figure out that organization.

Unsupervised Learning

Mail Inbox

The set of elements that describe a single datum are called features, in this case, the features are the words in the e-mails.

Each topic (clusters) will have features that will characterize them.

Research: Mars, Proposal, DTM, HiRISE, Machine Learning, Deep Nets, Bayesian

Family: Mom, House, Mexico

Promotions: Computer, PS4, Cheap, Amazon, Deal

Classes: Grades, Homework, Questions, Office Time

Questions?