GENERAL ASSEMBLY

INTRO TO UNSUPERVISED LEARNING

Matt Brems

DSI+

LEARNING OBJECTIVES

- By the end of this lesson, students should be able to:
  - Define supervised learning and unsupervised learning.
  - Identify strategies for unsupervised learning.
  - Identify problems where unsupervised learning can be applied.

DATA SCIENCE PROCESS

- Step 1: Define your problem.
- Step 2: Obtain the data.
- Step 3: Explore the data.
- Step 4: Model the data.
- Step 5: Evaluate the model.
- Step 6: Answer your problem.

RECAP: SUPERVISED LEARNING

- Let’s list the modeling techniques we’ve learned about.

WHAT IS OUR Y VARIABLE IN THESE CASES?

- I want to predict who is likely to vote in the 2020 election.
- I want to group stores by the demographic profiles of their consumers.
- I want to organize tweets by their topic.

UNSUPERVISED LEARNING

- **Unsupervised learning** is where we have, as part of our training data, no observed Y values.

- **Supervised learning** is where we have observed Y values as part of our training data.

STRATEGIES IN UNSUPERVISED LEARNING

1. Pick a proxy Y, then do supervised learning.

2. Use unsupervised learning as a stepping stone to get to supervised learning.

3. Try to organize observations by features.

STRATEGIES IN UNSUPERVISED LEARNING

1. Pick a proxy Y, then do supervised learning.

STRATEGIES IN UNSUPERVISED LEARNING

2. Use unsupervised learning as a stepping stone to get to supervised learning.

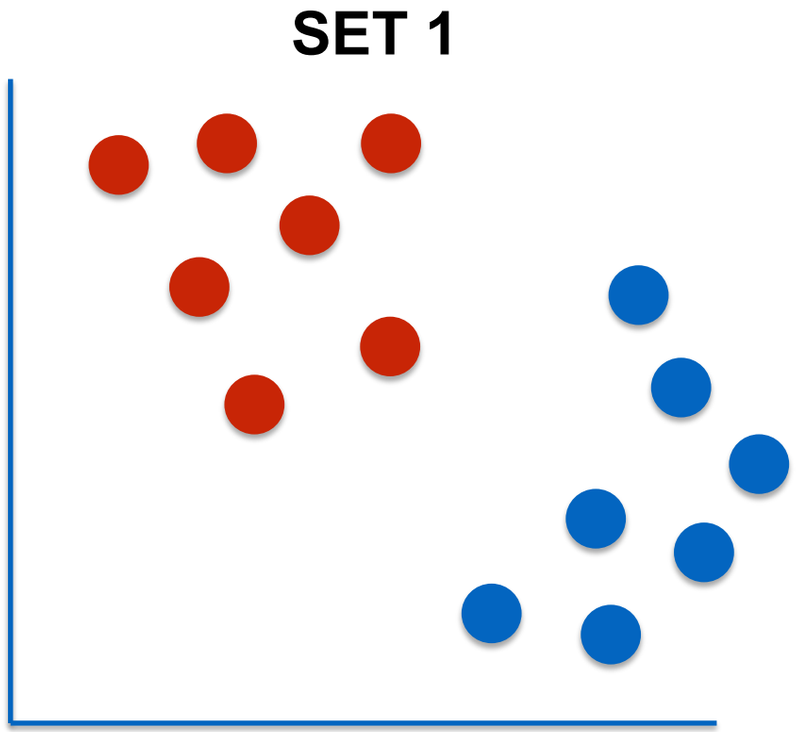
STRATEGIES IN UNSUPERVISED LEARNING

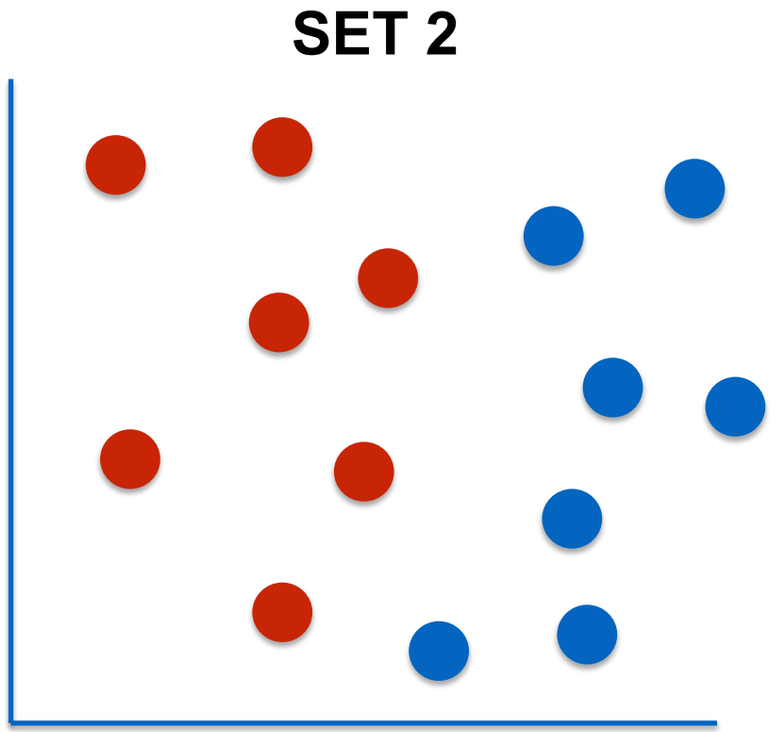
3. Try to organize observations by features.

EVALUATING UNSUPERVISED LEARNING MODELS

- In regression, we often use MSE and/or R^2 to evaluate models.
- In classification, we often use accuracy, sensitivity, specificity, precision, and/or AUC ROC to evaluate models.
- All of these rely on knowing our actual Y and predicted Y.
  - …so how do we evaluate models for unsupervised learning?

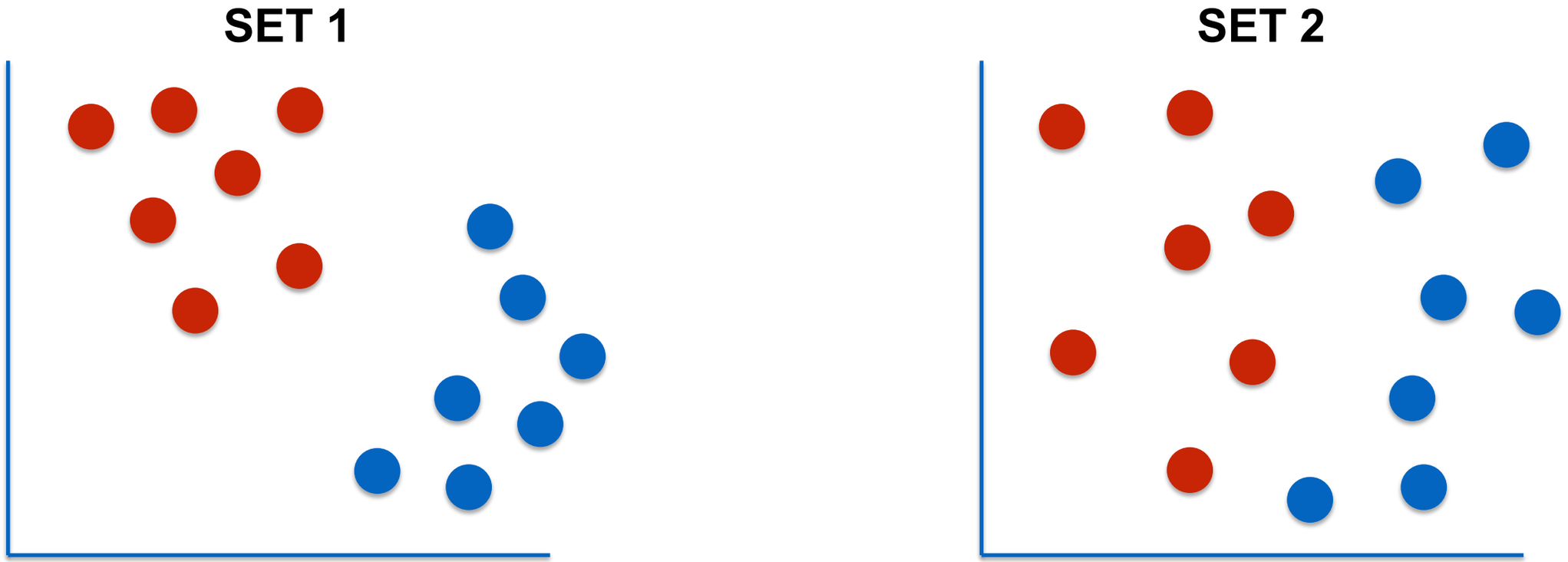
EVALUATING UNSUPERVISED LEARNING MODELS





- Would we say that one of these sets of clusters is “better?” Why or why not?

EVALUATING UNSUPERVISED LEARNING MODELS



- How might we quantify the “goodness” of a cluster?

EVALUATING UNSUPERVISED LEARNING MODELS

- Inertia

- Silhouette Score

- Three common measurements that require knowledge of “ground truth” a.k.a. Y labels!
  - Homogeneity, Completeness, V-Measure

UNSUPERVISED LEARNING TAKEAWAYS

- Unsupervised learning will almost always underperform supervised learning methods.

- It is quite difficult to learn without access to our *Y* variable… but we want to learn as much as we can from the data that we have!
  - Network Analysis
  - Topic Modeling
  - Clustering
  - Principal Component Analysis

NETWORK ANALYSIS

- As we discussed, **network analysis** has a number of common problems where we may use unsupervised learning methods.

- Let's say I have access to credit card transaction data. I want to know which transactions are likely to be fraudulent.

- Let's say I have access to social network data. Each node corresponds to a person and each edge connecting a person measures how connected two individuals are through social media, phone calls, emails, and text messages. I want to know if there exists a crime ring in the data.

TOPIC MODELING

- One example of an unsupervised learning problem is **topic modeling**.
- Suppose I want to look at how topics in magazines or academic articles have changed over time. How can I model this?
    - A topic model is where we use a statistical model to attempt to identify topics or categories in a set of documents.
    - Interested in more? Check out Latent Dirichlet Allocation (LDA) for a relatively easy algorithm to implement.

Interpretation of Results

Articles from *The Journal of Cell Biology*

23,896 articles from 1962 through present (2017)

2.2 million words

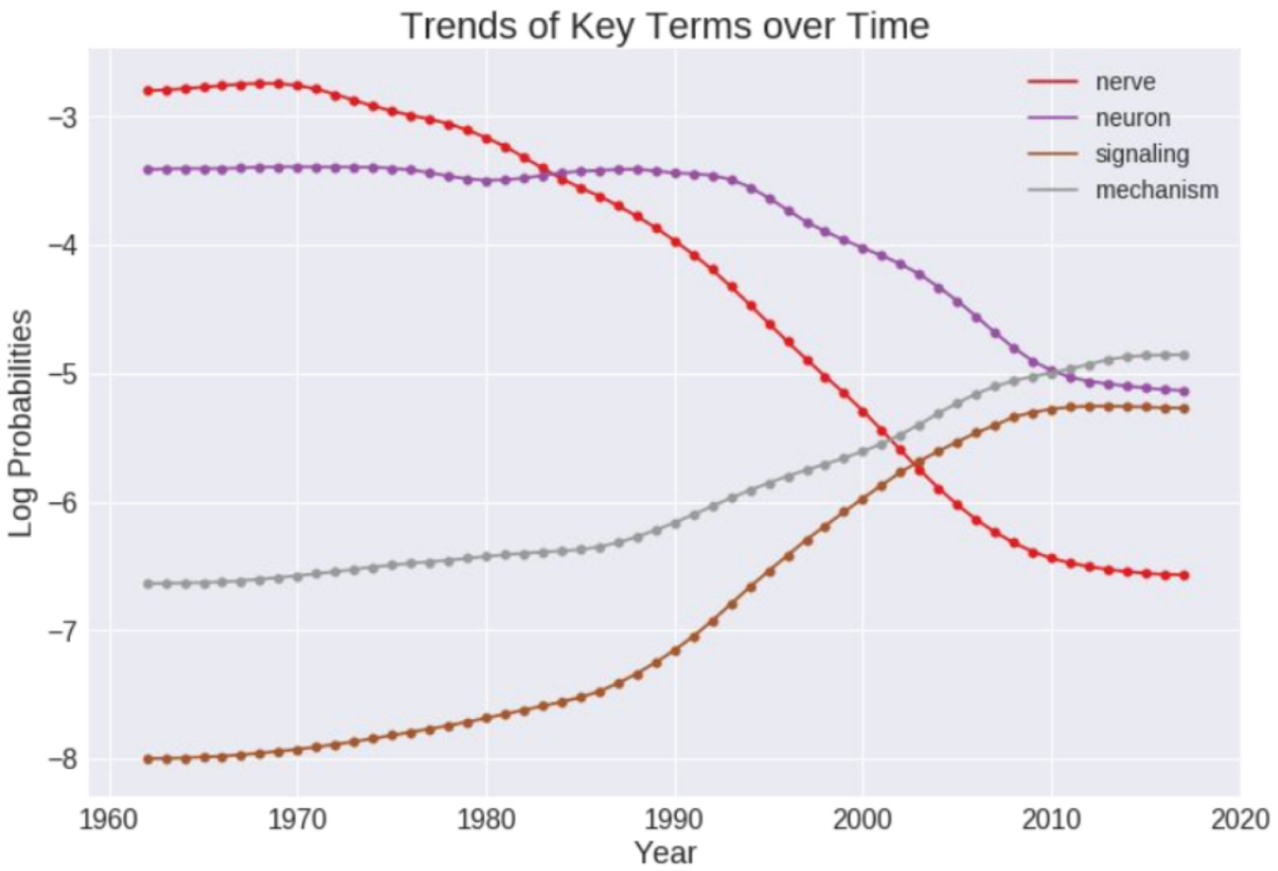
6,619 words in vocabulary after pruning

Estimated 10-component dynamic topic model:

- cytoskeletal systems
- inter-cell communications
- nucleus, cell replications and cycles
- inter- and intra-cell transport
- neuroscience
- cell signaling
- imaging techniques (esp. microscopy)
- gene transcription and translation
- cell/tissue cultures, cancer research
- mitochondria

Credit: Jingfei Cai, DSI-DC-5

“neuroscience”



1965: 'nerve', 'axon', 'neuron', 'myelin', 'sheath'

1975: 'nerve', 'axon', 'neuron', 'synaptic', 'terminal'

1985: 'cam', 'neuron', 'nerve', 'cell', 'axon', 'ngf'

1995: 'neuron', 'cell', 'axon', 'growth', 'apoptosis'

2005: 'cell', 'apoptosis', 'protein', 'neuron', 'death'

2015: 'cell', 'protein', 'function', 'cellular', 'mechanism'

CLUSTERING

- One of the most common approaches to tackling unsupervised learning problems is to cluster.

- Let's say I'm attempting to organize 1,400 Target stores by demographic profiles so I can better market to them. How might I do this?

- Let's say that I want to look at positions of basketball players on the court over time and see how many "real" positions there are. How might I do that?

PRINCIPAL COMPONENT ANALYSIS

- **Principal component analysis** is a commonly used technique for a few different reasons.

    1. In cases where we want to do dimensionality reduction, we can apply PCA so that we retain all of the “good” information from all of our features, but discard the redundant information from our features!

    2. In cases where we want to ensure our independent variables are independent of one another, PCA will make sure this happens.

ONE FINAL NOTE

- The methods you learn in unsupervised learning **can** be applied to supervised learning problems!
  - For example, you can apply PCA to reduce the number of variables you're passing into a regression model. (Check out Principal Component Regression in ISLR for more on this.)

- In these cases, you can use metrics like homogeneity, completeness, or even metrics like accuracy and sensitivity to evaluate your model.

- However, you cannot easily apply supervised learning techniques to situations without labeled Y data.

ADDITIONAL RESOURCES

- **Wired article on basketball positions:** https://www.wired.com/2012/04/analytics-basketball/

- **Jingfei Cai's capstone presentation**: https://github.com/Ailuropoda1864/PMC-text-mining/blob/master/presentation.pdf

- **Mini-book on Fraud Detection through unsupervised learning**: http://www.diva-portal.org/smash/get/diva2:897808/FULLTEXT01.pdf