

REVIEW

Week 8

DSI-NYC Triassics, General Assembly

WEEK 8 REVIEW

LEARNING OBJECTIVES

- Explain the concepts, models and tools we've covered in recent weeks
- Contextualize their place in your personal toolkits

REVIEW SESSION

LESSON PLAN

5 min	Opening	
20 min	Key concepts	Curse of dimensionality, train-validation-test, cross-validation, Bayes
10 min	New tech tools	Web architecture, Flask, R, regex
20 min	Unsupervised methods	Clustering, PCA
10 min	Supervised methods	Naive Bayes
20 min	Choosing models	(Part 1 of this discussion)
5 min	Conclusion	

WEEK 8 REVIEW

RECENT MATERIAL

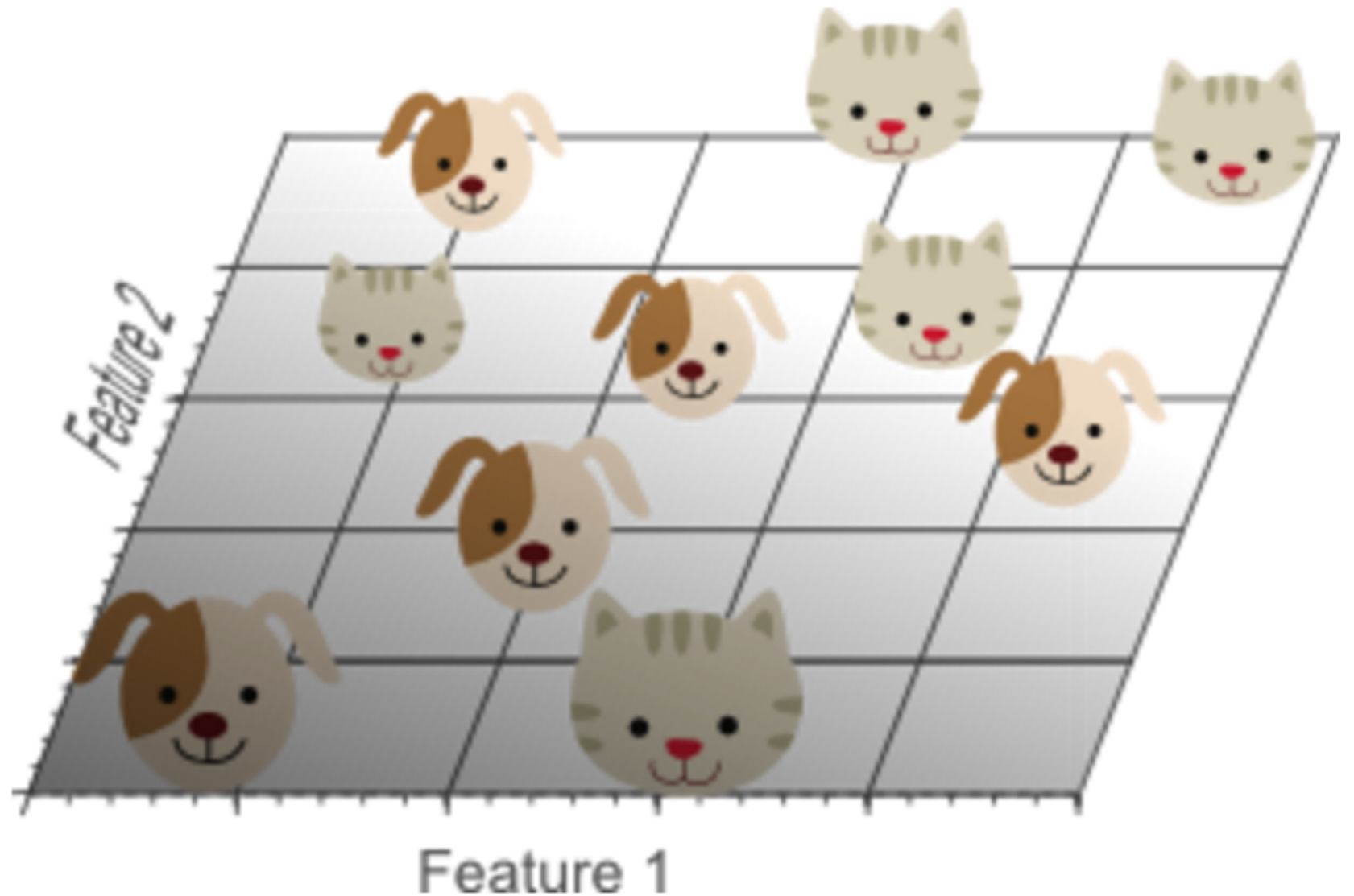
- Key concepts
 - Curse of dimensionality
 - Test-train-validation
 - Cross validation
 - Bayesian statistics
- Unsupervised methods
 - Clustering
 - K-means
 - Hierarchical
 - DBSCAN
 - PCA
- Supervised
 - Naive Bayes
- Tech
 - Web architecture
 - Flask
 - R
 - Regex

REVIEW

KEY CONCEPTS

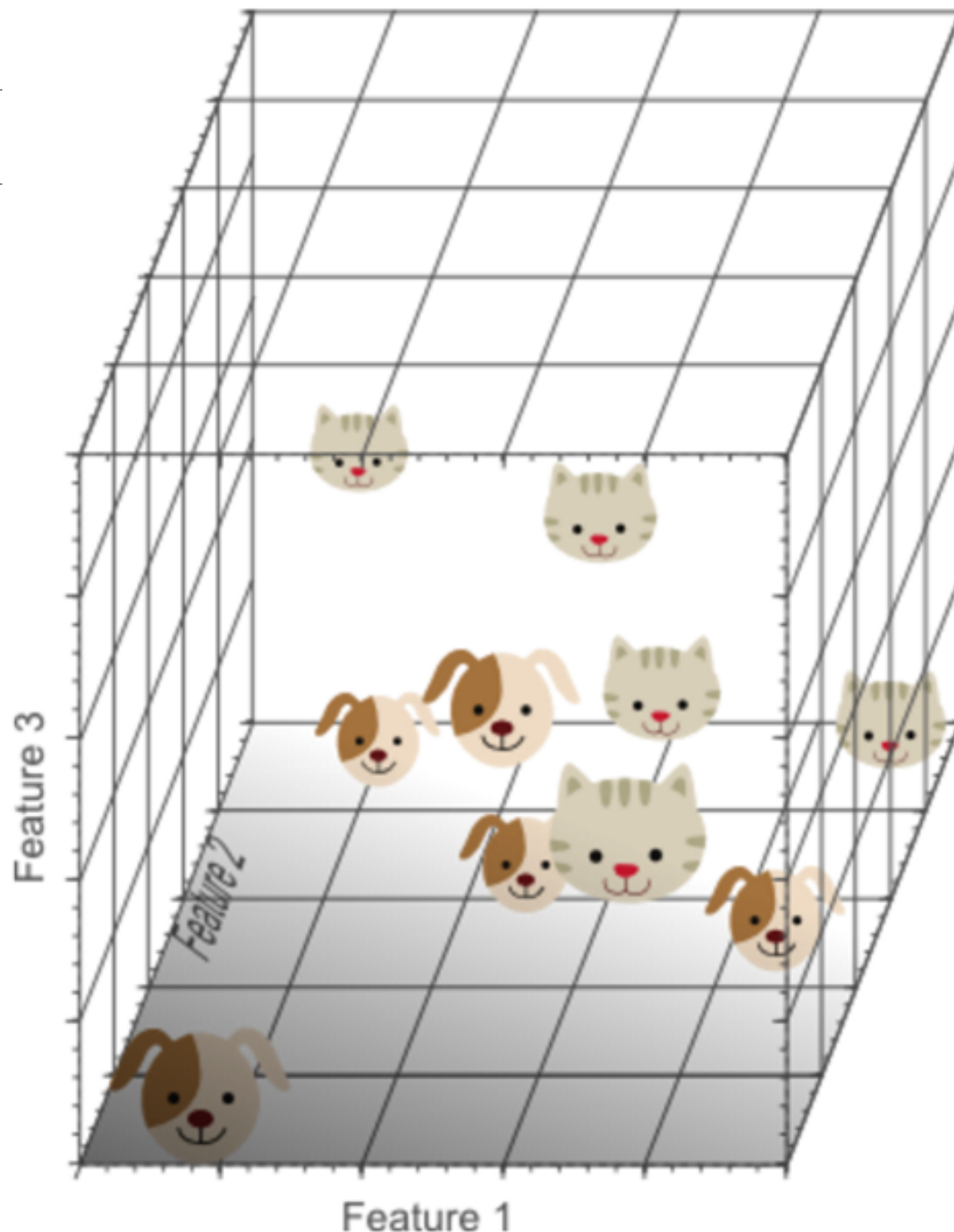
KEY CONCEPTS

CURSE OF DIMENSIONALITY



KEY CONCEPTS

CURSE OF DIMENSIONALITY



KEY CONCEPTS

TRAIN-TEST-VALIDATION

Fit model
on...

Train
data

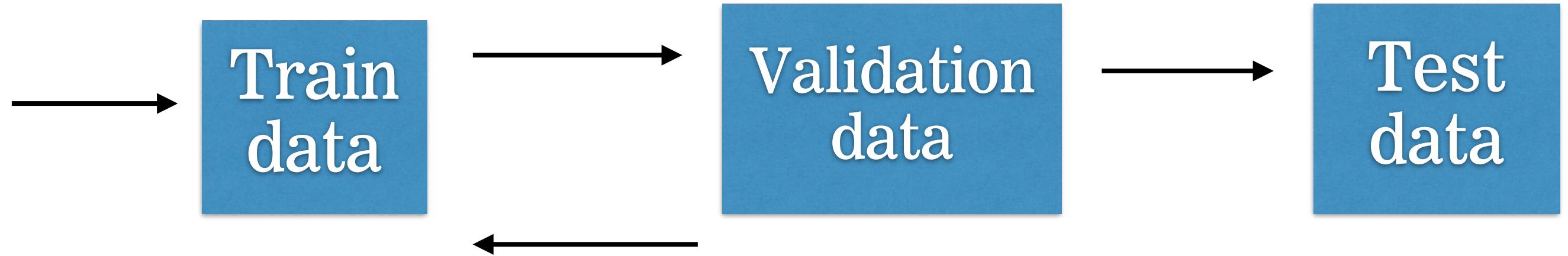
Score model
on....

Validation
data

Test 'best' model for
external validity on...

Test
data

Iterate
hyperparameter(s) and
fit revised model on....



KEY CONCEPTS

TRAIN-TEST-VALIDATION

- “The training set is used to fit the models; the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final chosen model. Ideally, the test set should be kept in a ‘vault,’ and be brought out only at the end of the data analysis.”
 - Elements of Statistical Learning
- Why? To avoid information leakage and overfitting
- Avoiding this 100% is tough. At least try to minimize the variance of your test error, e.g. when using cross-validation

KEY CONCEPTS

CROSS VALIDATION

- Use more (or all of) your dataset for training by iteratively splitting it into train and validation sets
- With the k-folds strategy, split into k folds, train on k - 1 folds, and validate (score) on remaining fold
- ‘cross_val_score’ IS NOT A SCORE. It is a function that does cross validation and returns a score for each iteration
 - The metric for this score is the estimator’s default method; you can also set the scoring parameter yourself

KEY CONCEPTS

BAYESIAN STATISTICS

- Probability is a representation of our uncertainty given what we know and believe to be true.
- “Data informs us about the distribution, and as we receive more data our view of the distribution can be updated, further confirming or denying our previous beliefs (but never in certainty).”

KEY CONCEPTS

BAYESIAN STATISTICS

Likelihood

How probable is the evidence
given that our hypothesis is true?

Prior

How probable was our hypothesis
before observing the evidence?

$$P(H | e) = \frac{P(e | H) P(H)}{P(e)}$$

Posterior

How probable is our hypothesis
given the observed evidence?
(Not directly computable)

Marginal

How probable is the new evidence
under all possible hypotheses?
 $P(e) = \sum P(e | H_i) P(H_i)$

Graphic from <http://www.psychologyinaction.org/2012/10/22/bayes-rule-and-bomb-threats/>

REVIEW

TECH

OTHER TECHNOLOGICAL TOOLS

WEB ARCHITECTURE

- The “web” is based on specific protocols for sharing information over the Internet
- If you want to do Internet things, learn about these protocols

OTHER TECHNOLOGICAL TOOLS

FLASK

- A very light-weight framework for creating applications that employ web protocols
- Great for internal prototypes or delicately used tools
- Not ideal for ‘real’ web apps*

* Alex and Dan disagree with this.

OTHER TECHNOLOGICAL TOOLS

R

- A programming language / computational environment
- Basic concepts are similar to Python
- Well suited to ad hoc analyses and EDA, less suited to production code

OTHER TECHNOLOGICAL TOOLS

REGULAR EXPRESSIONS

- A standardized language for creating text-based search patterns
- ‘Regex’ implementations are available from many languages, including Python

REVIEW

UNSUPERVISED METHODS

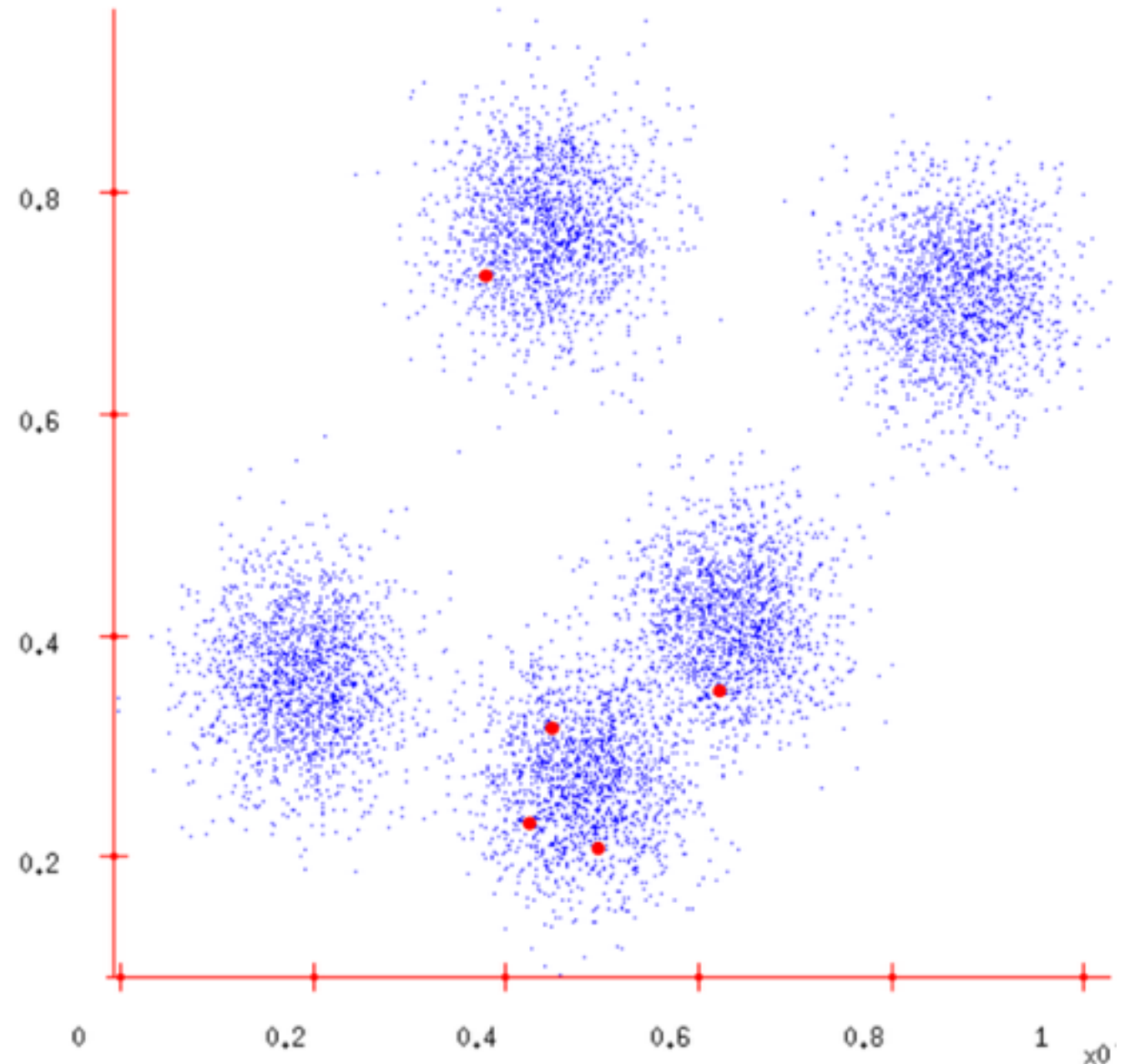
UNSUPERVISED METHODS

- Finding groupings in your data. Why?
 - Market segmentation
 - Internal structure in physical systems (e.g. genetics)
 - Dimensionality reduction for computational reasons
 - ...other suggestions?

UNSUPERVISED METHODS

K-MEANS CLUSTERING

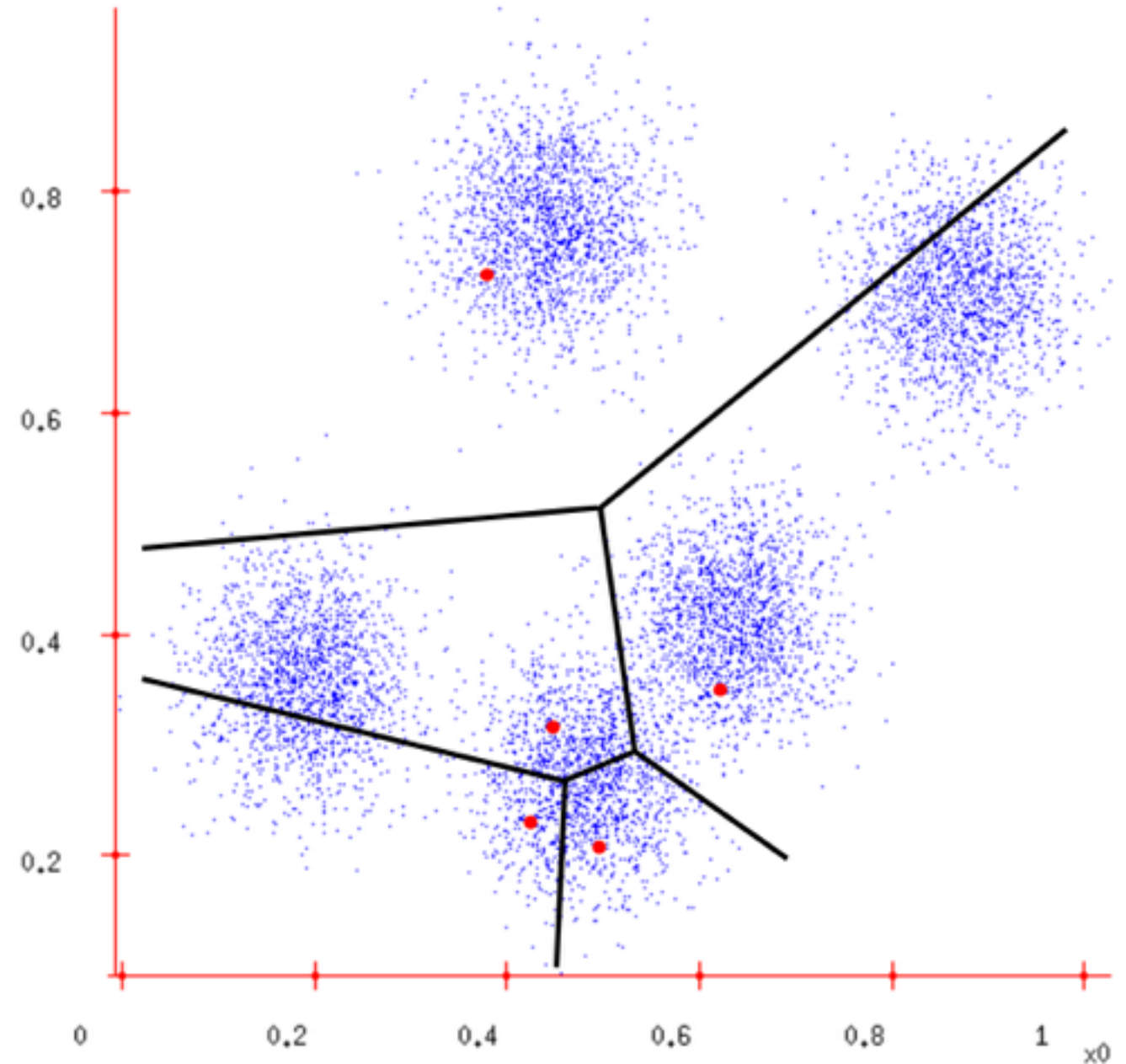
- Place an arbitrary number k centroids randomly in your feature space



UNSUPERVISED METHODS

K-MEANS CLUSTERING

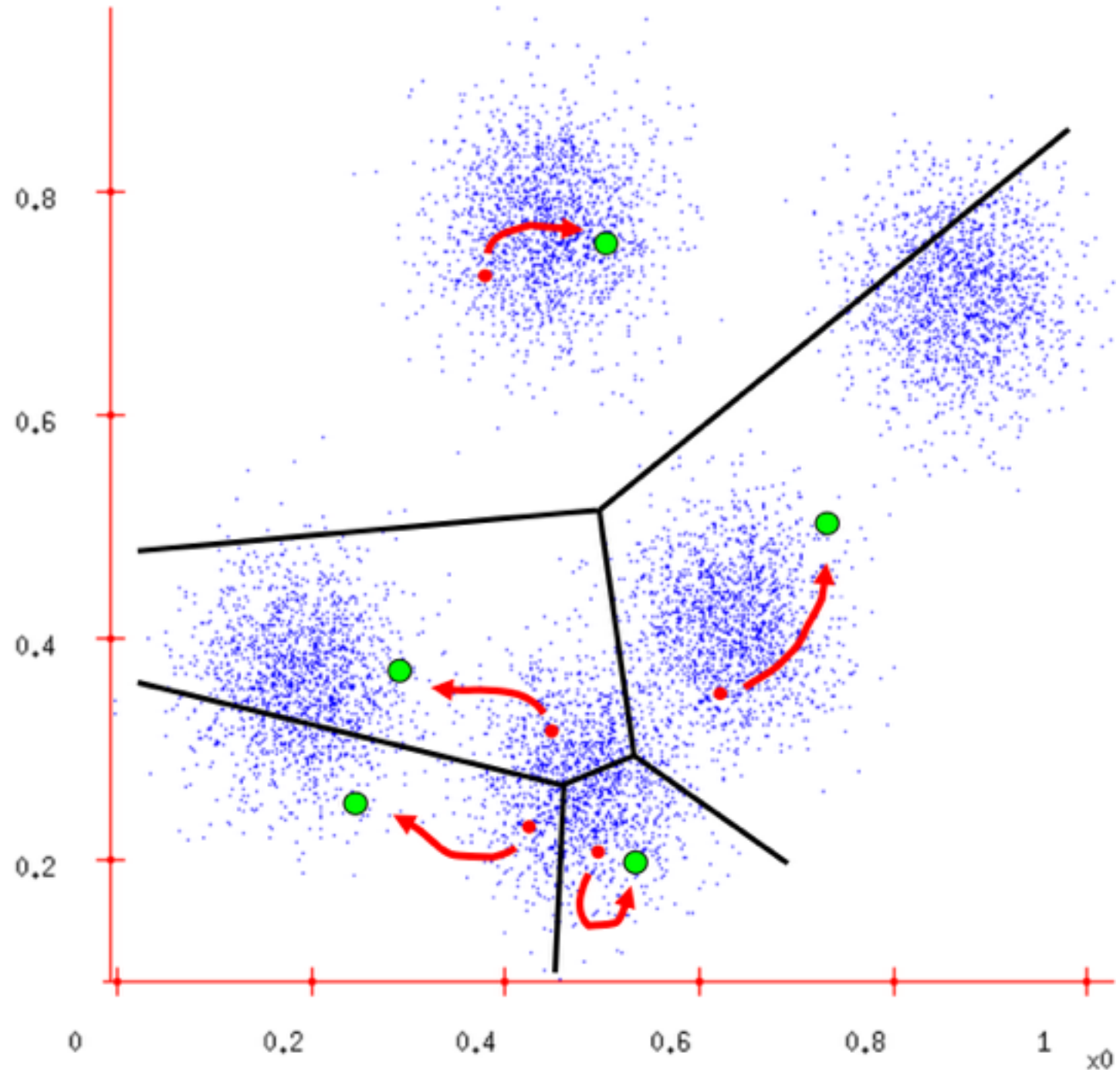
- Assign points to nearest centroid



UNSUPERVISED METHODS

K-MEANS CLUSTERING

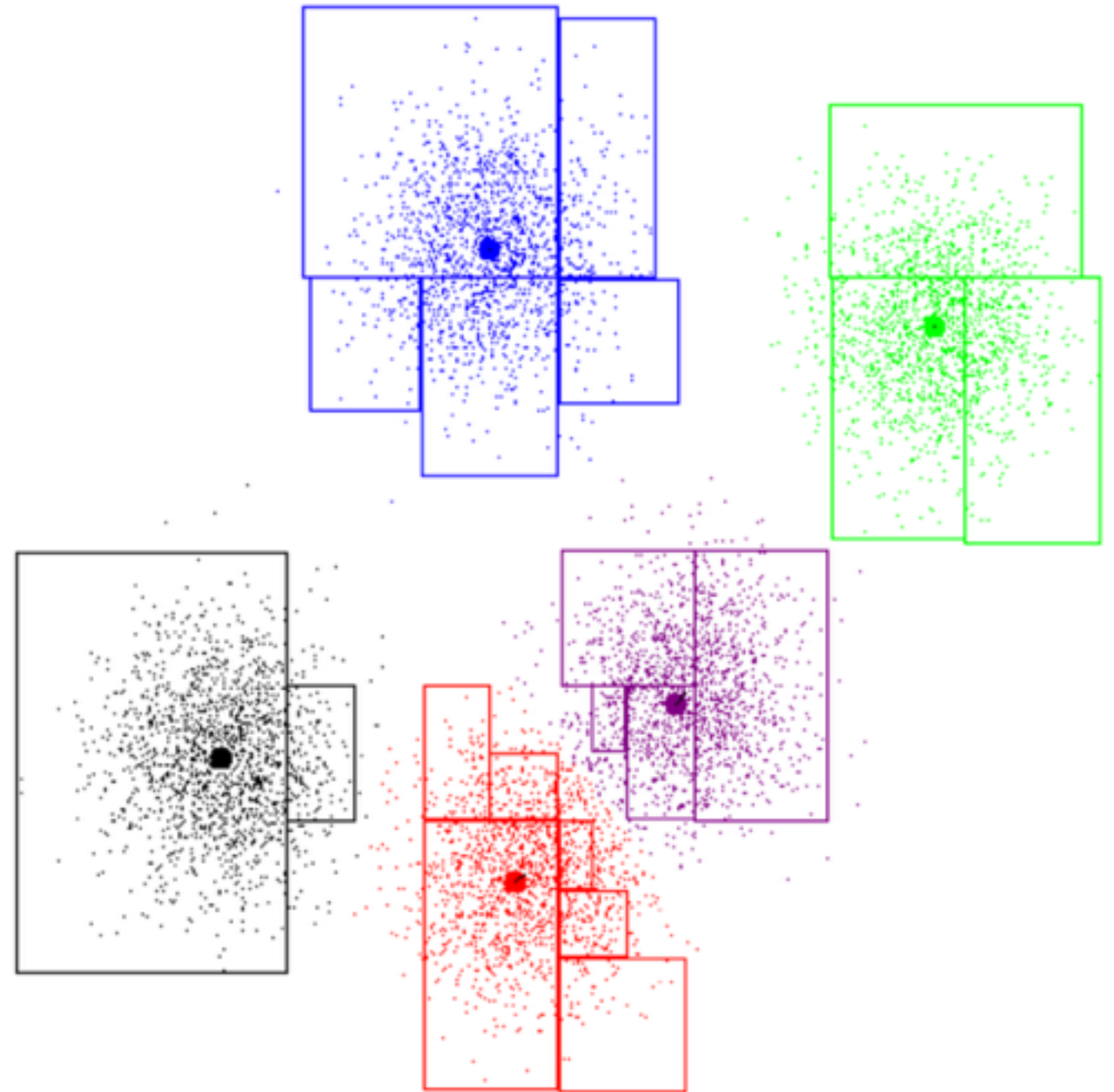
- Move centroids to Euclidean center of each cluster of point



UNSUPERVISED METHODS

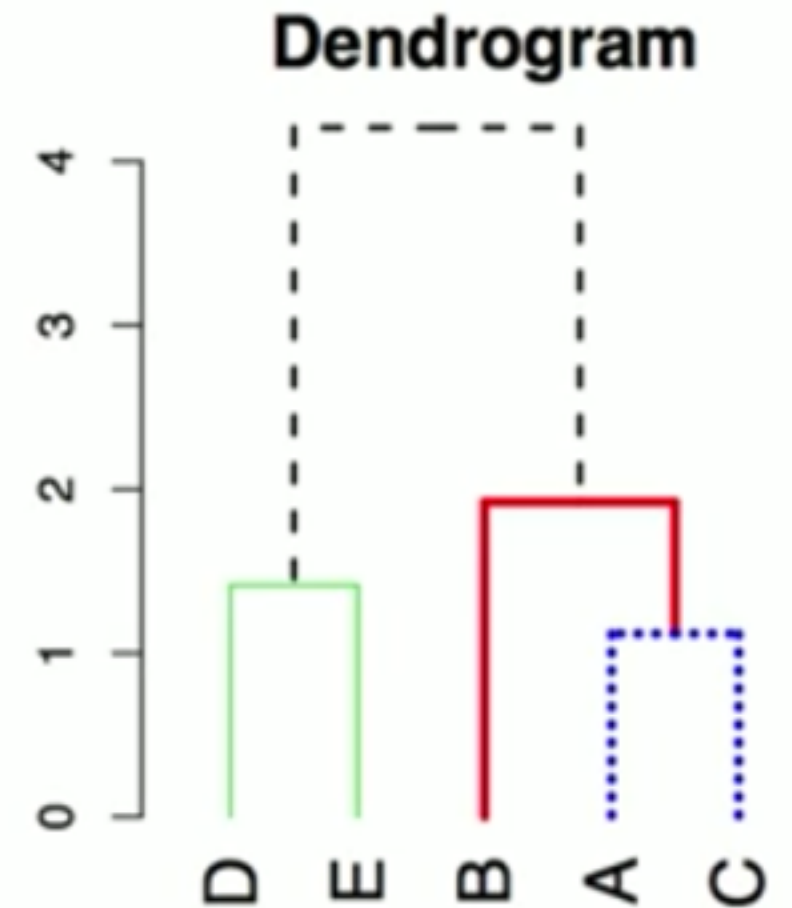
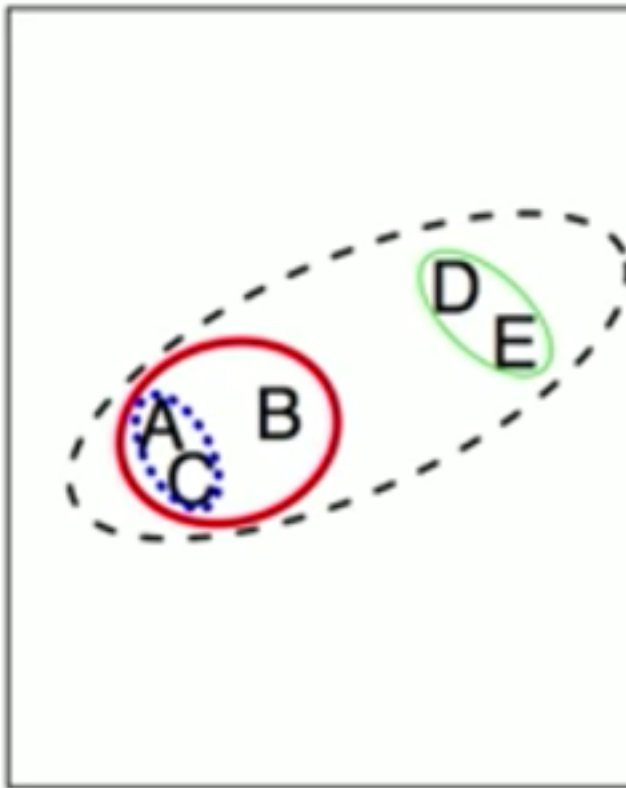
K-MEANS CLUSTERING

- Repeat until the centroids are \sim stationary



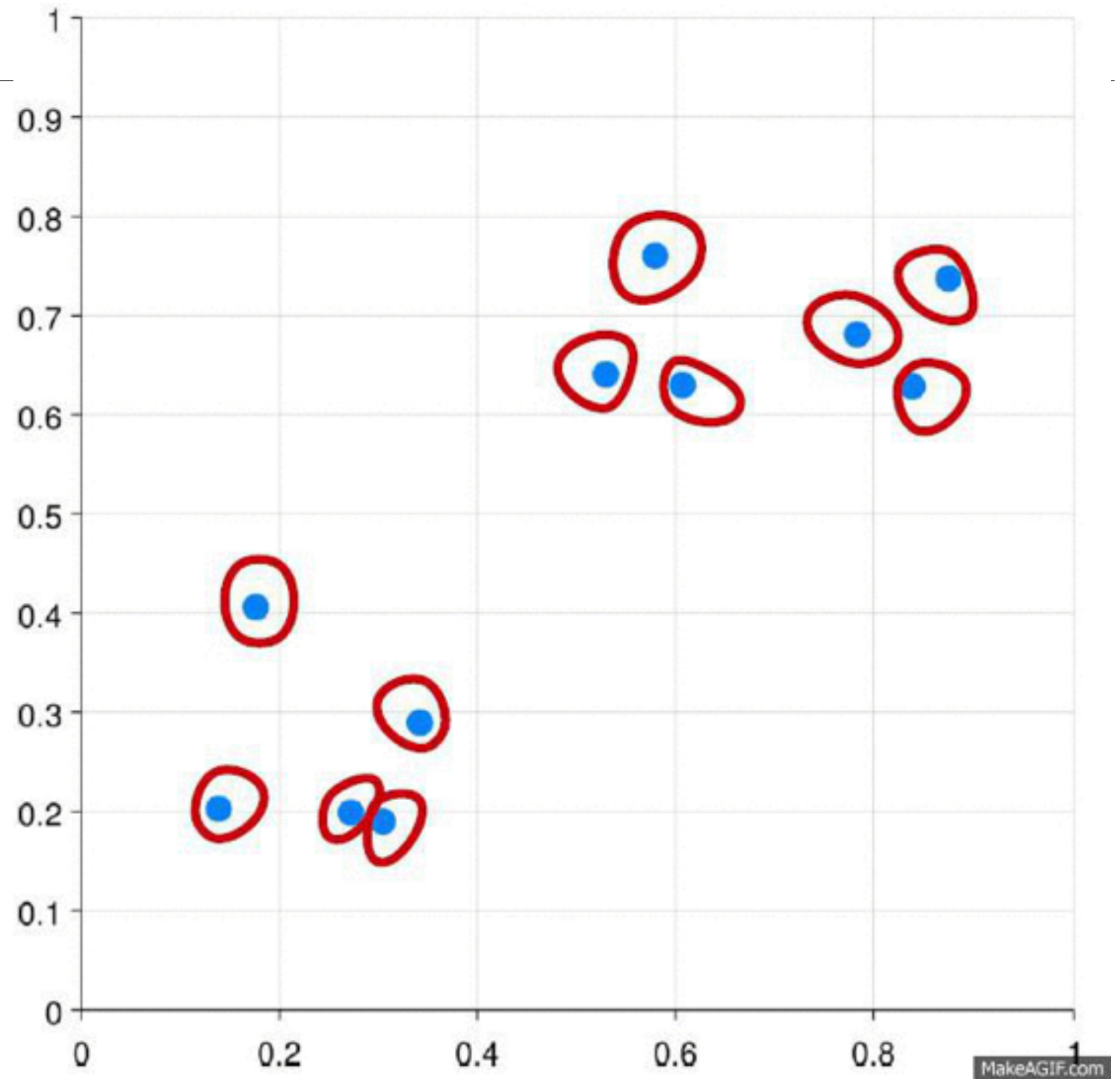
UNSUPERVISED METHODS

HIERARCHICAL CLUSTERING



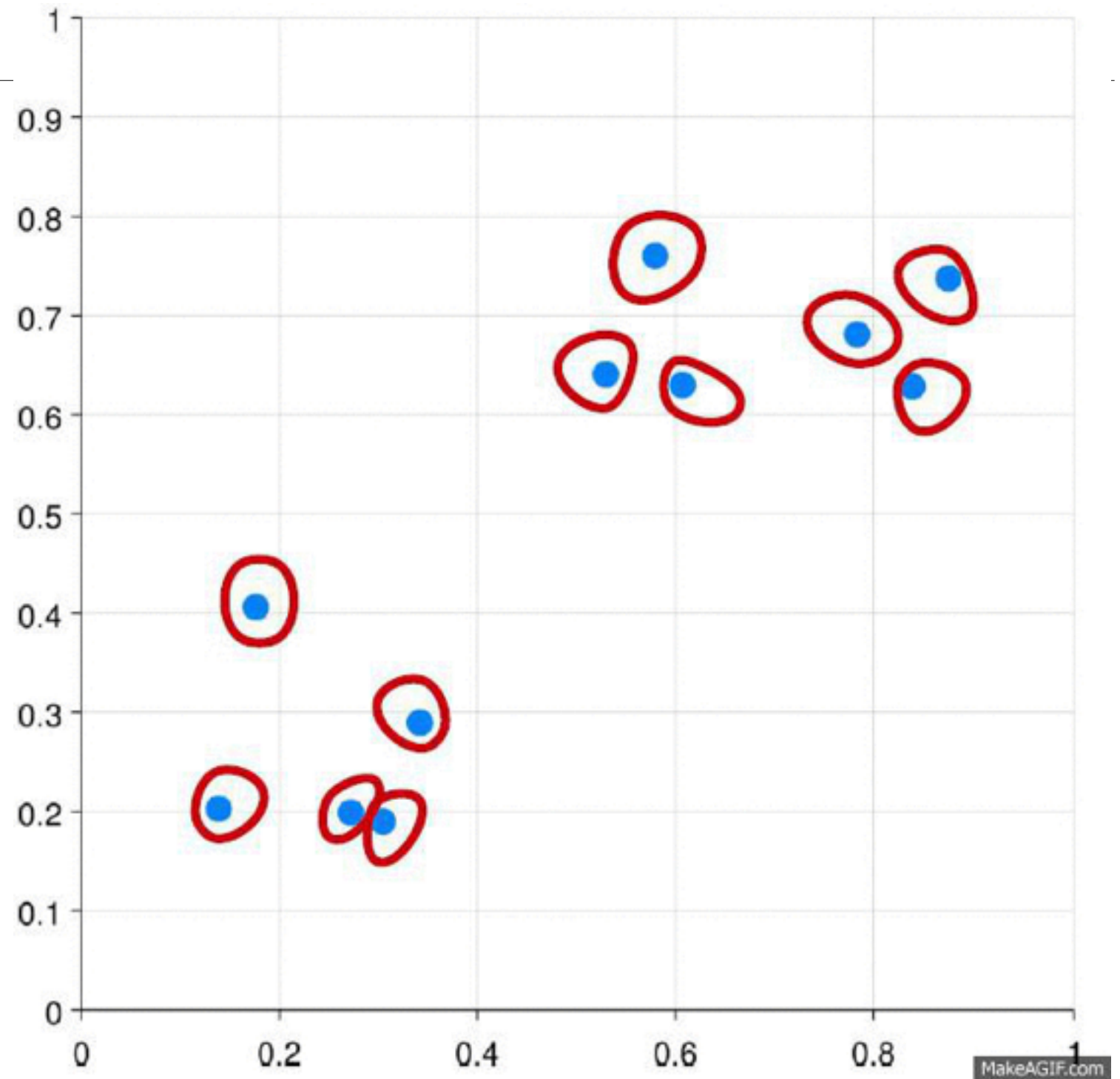
UNSUPERVISED METHODS

HIERARCHICAL CLUSTERING



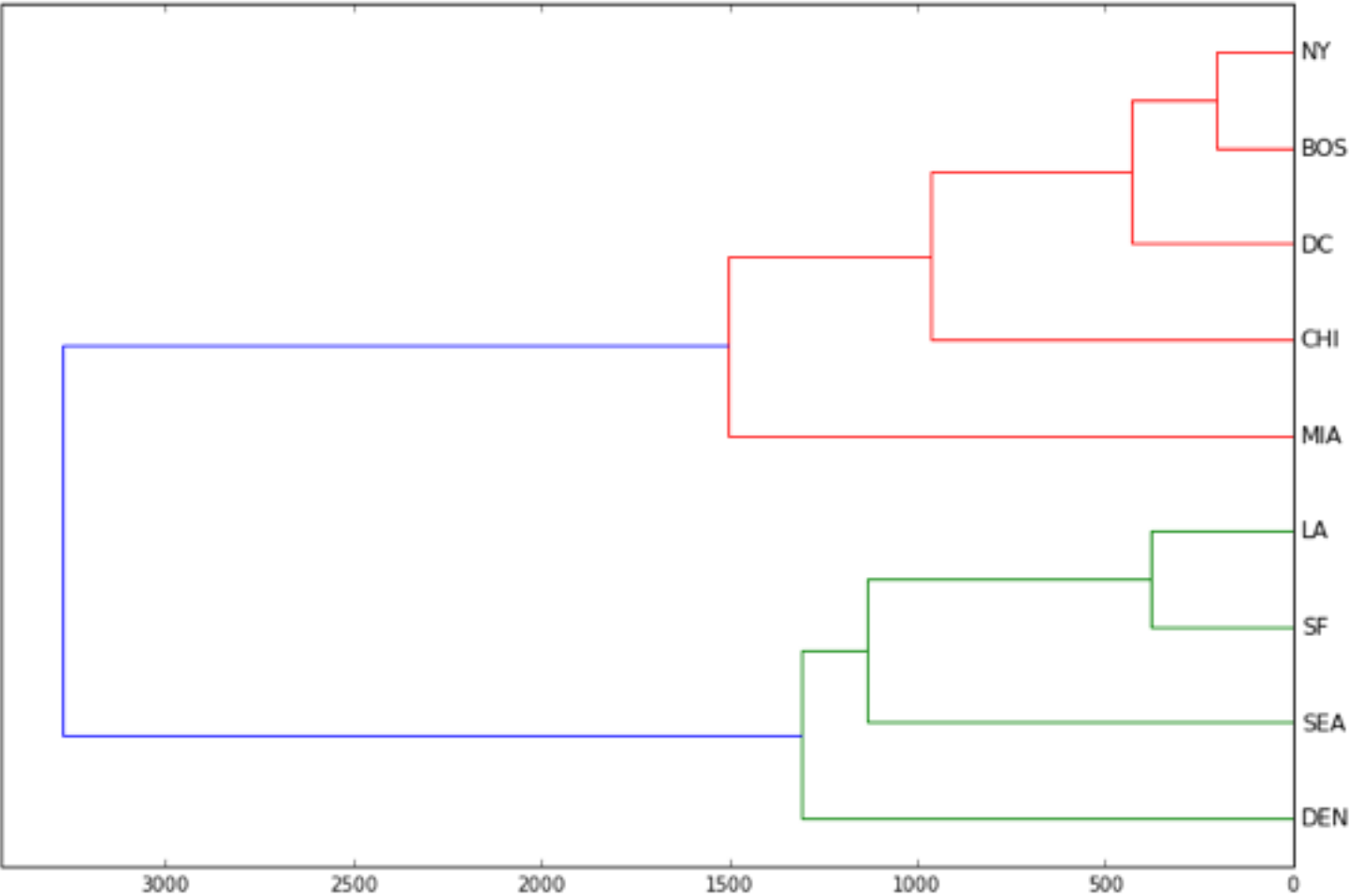
UNSUPERVISED METHODS

HIERARCHICAL CLUSTERING



UNSUPERVISED METHODS

HIERARCHICAL CLUSTERING



UNSUPERVISED METHODS

DBSCAN

- Density-Based Spatial Clustering for Applications with Noise
- <http://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>

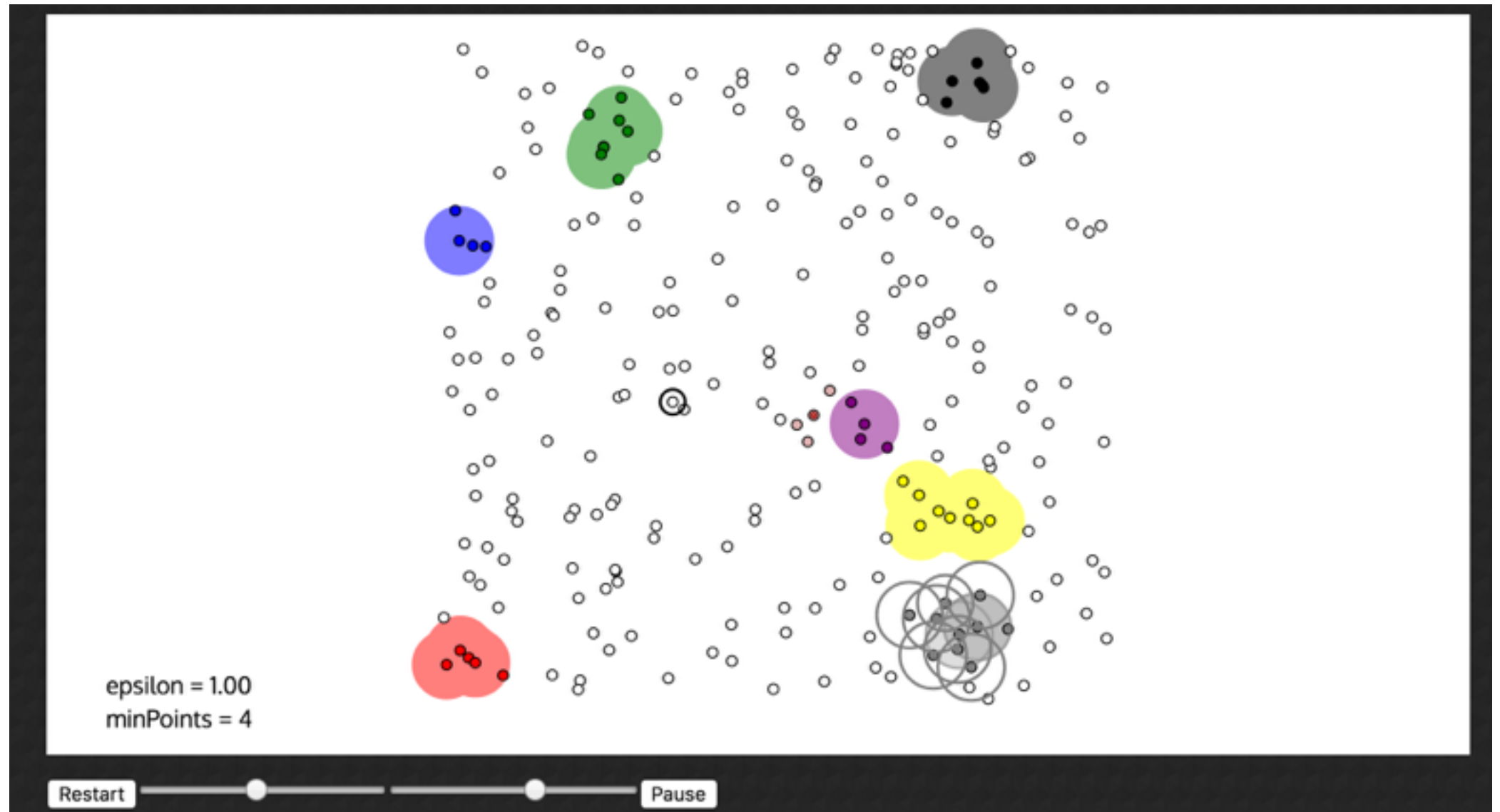
UNSUPERVISED METHODS

DBSCAN

- One arbitrary starting point.
- If enough points in its ϵ -neighborhood, start a cluster. Otherwise call it noise.
- If noise, select another random point, and begin again.
- At each cluster, add points the same way.
- Close the cluster when you can't add any more points and start again.

UNSUPERVISED METHODS

DBSCAN

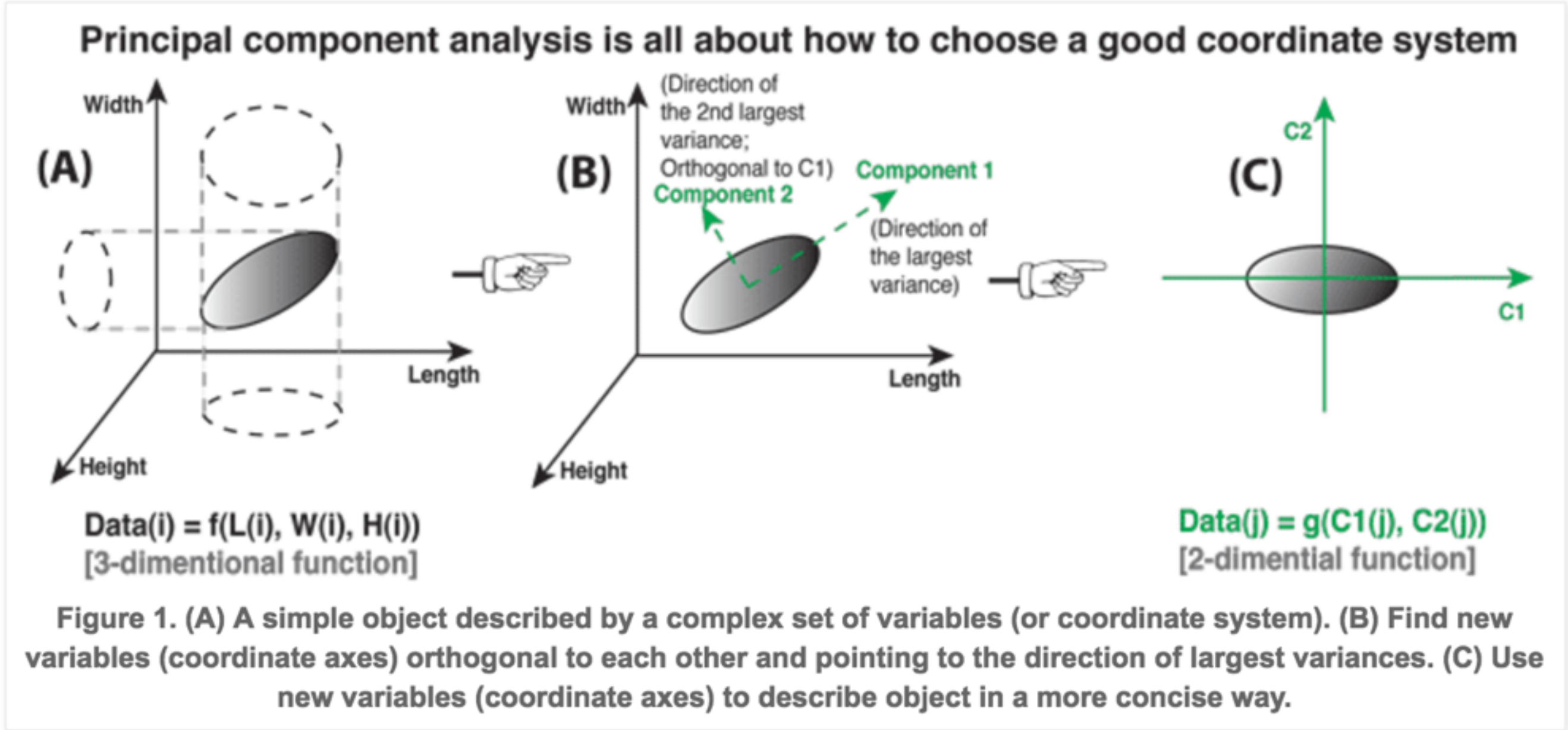






UNSUPERVISED METHODS

PCA

- Find the directions of greatest variance in your data
- These are the eigenvectors of your covariance matrix
- These are linear combinations of your current dimensions, and are orthogonal to each other
- Their eigenvalues indicate amount of variance explained
- If a subset of your principal components together explain most of the variance, use them to create a new coordinate space

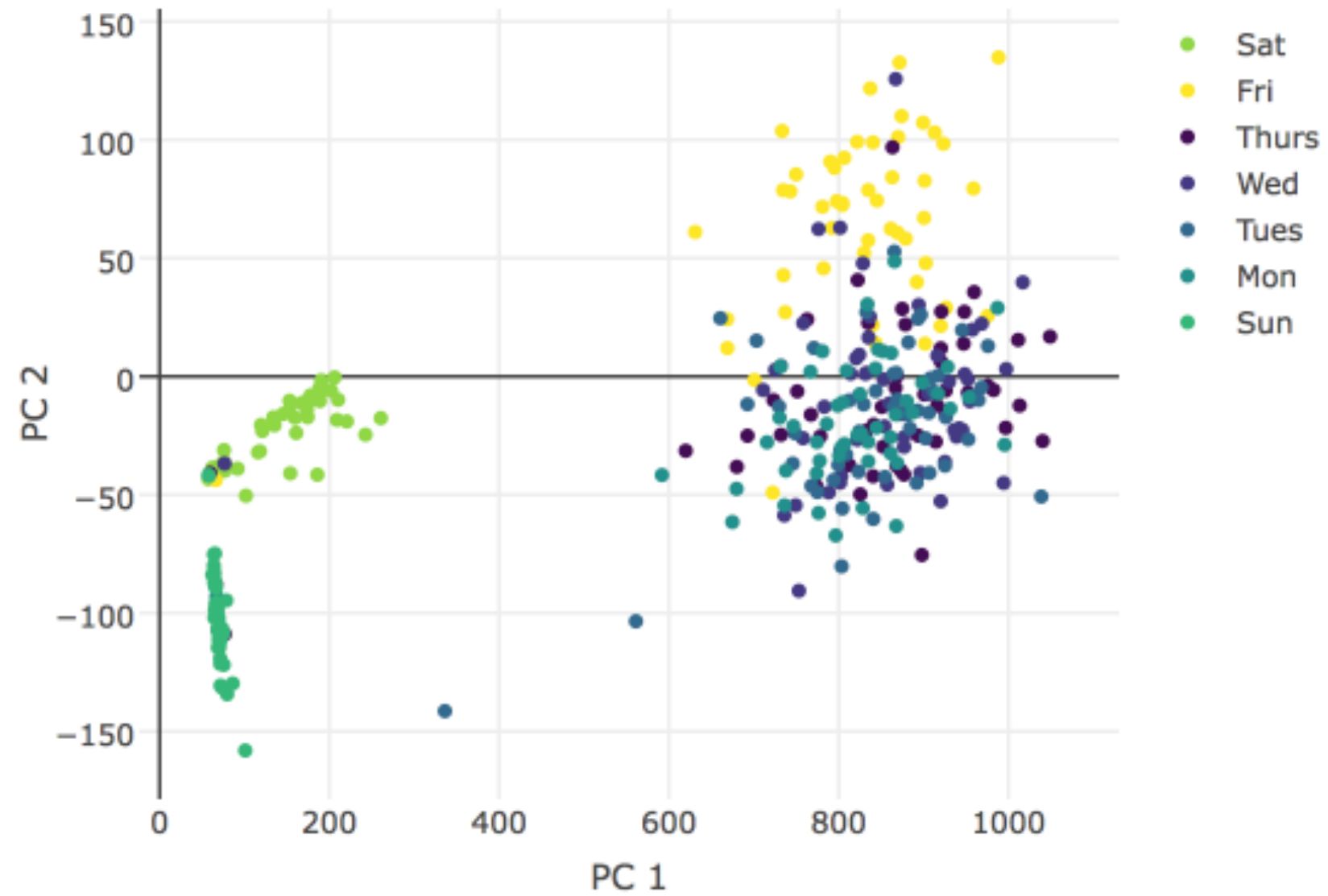
PCA is nothing but coordinate system transformation: A simple example



	DEMAND 	TIME 	HOUR 	DATE 
1	44.66143	2014-07-18 00:00:00	0	2014-07-18
2	45.81310	2014-07-18 01:00:00	1	2014-07-18
3	64.94065	2014-07-18 02:00:00	2	2014-07-18
4	136.03094	2014-07-18 03:00:00	3	2014-07-18
5	288.16307	2014-07-18 04:00:00	4	2014-07-18
6	320.10478	2014-07-18 05:00:00	5	2014-07-18
7	329.21464	2014-07-18 06:00:00	6	2014-07-18
8	324.05256	2014-07-18 07:00:00	7	2014-07-18
9	315.26830	2014-07-18 08:00:00	8	2014-07-18
10	321.05021	2014-07-18 09:00:00	9	2014-07-18
11	320.20072	2014-07-18 10:00:00	10	2014-07-18
12	253.04267	2014-07-18 11:00:00	11	2014-07-18
13	115.55841	2014-07-18 12:00:00	12	2014-07-18

UNSUPERVISED METHODS

PCA



UNSUPERVISED METHODS

PCA – USE CASES

- Visual EDA
- Dimensionality reduction as preprocessing (while reducing information loss)
- Data compression
- Analysis - may reveal ‘truthier’ dimensionality in data

REVIEW

SUPERVISED ML

SUPERVISED METHOD

NAIVE BAYES

- Leverages Bayes' Theorem
- Assumes independence of feature values, which allows:

$$P(y \mid x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i \mid y)$$

\Downarrow

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i \mid y),$$

REVIEW

WHEN TO DO WHAT AND WHY

PART I

WEEK 8 REVIEW

A CAVEAT

- There are few easy answers.
Machine learning in general is an area of intense academic research
- Evolving community understanding of what works when, and why
- Domain-specific practical experience == \$\$\$

WEEK 8 REVIEW

THAT SAID...

WEEK 8 REVIEW

COMMON CONSIDERATIONS AND HEURISTICS

- Is this supervised or unsupervised? Neither, exactly?
- Is this classification or regression?
- Is the target function linear?
- How much data do you have? Can you get more?
- Will you train incrementally or in full batches?

WEEK 8 REVIEW

COMMON CONSIDERATIONS AND HEURISTICS

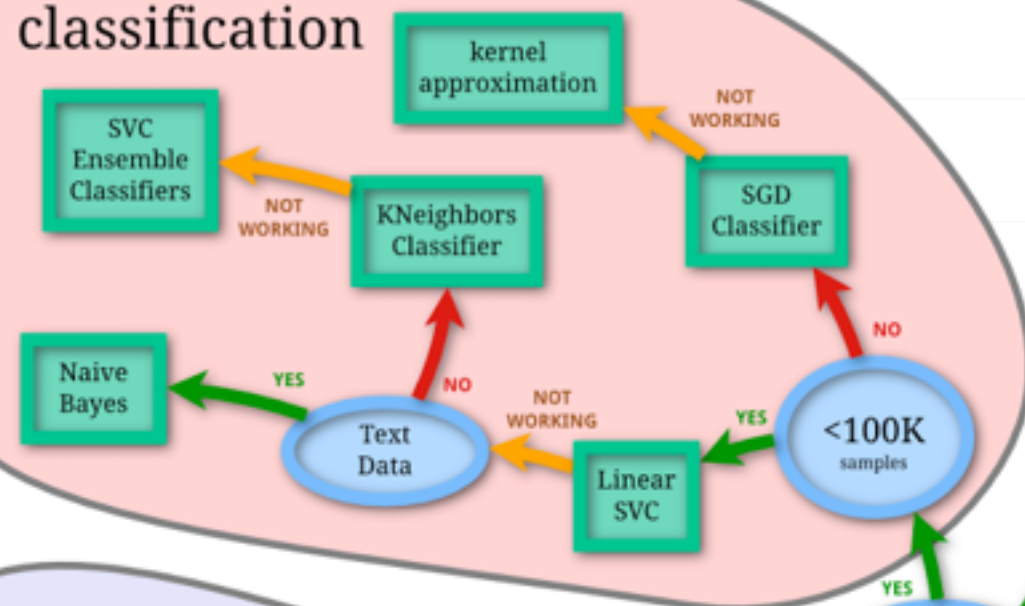
- What kinds of values do your features take?
- Do you believe your features are good ones? Can you engineer better ones?
- Does training speed matter?
- Does prediction speed matter?
- Does interpretability matter?

COMMON CONSIDERATIONS AND HEURISTICS

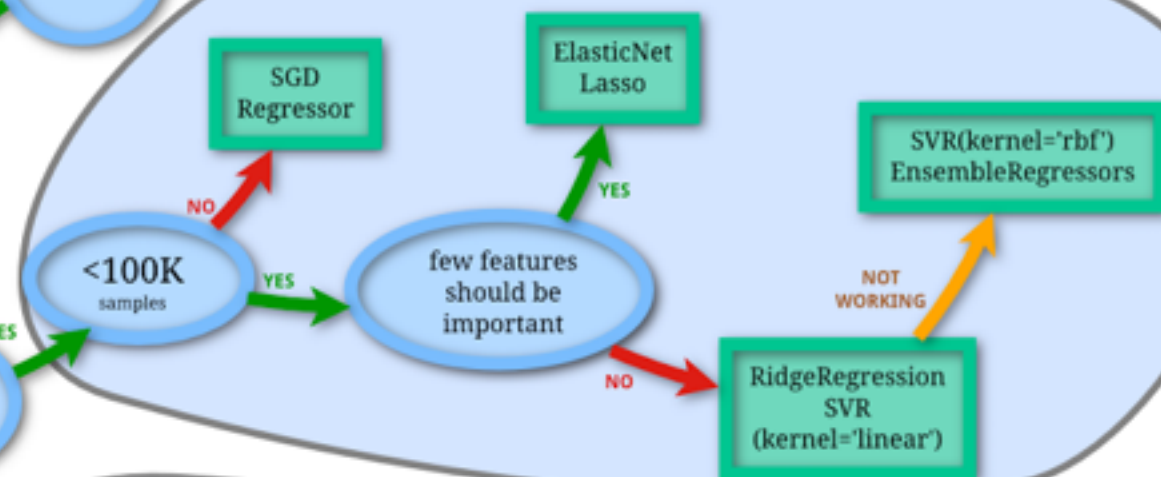
- Is it Kaggle?
 - XGBoost and neural networks are so hot right now.
 - Random forests are so 2015
 - Ensembles still work

scikit-learn algorithm cheat-sheet

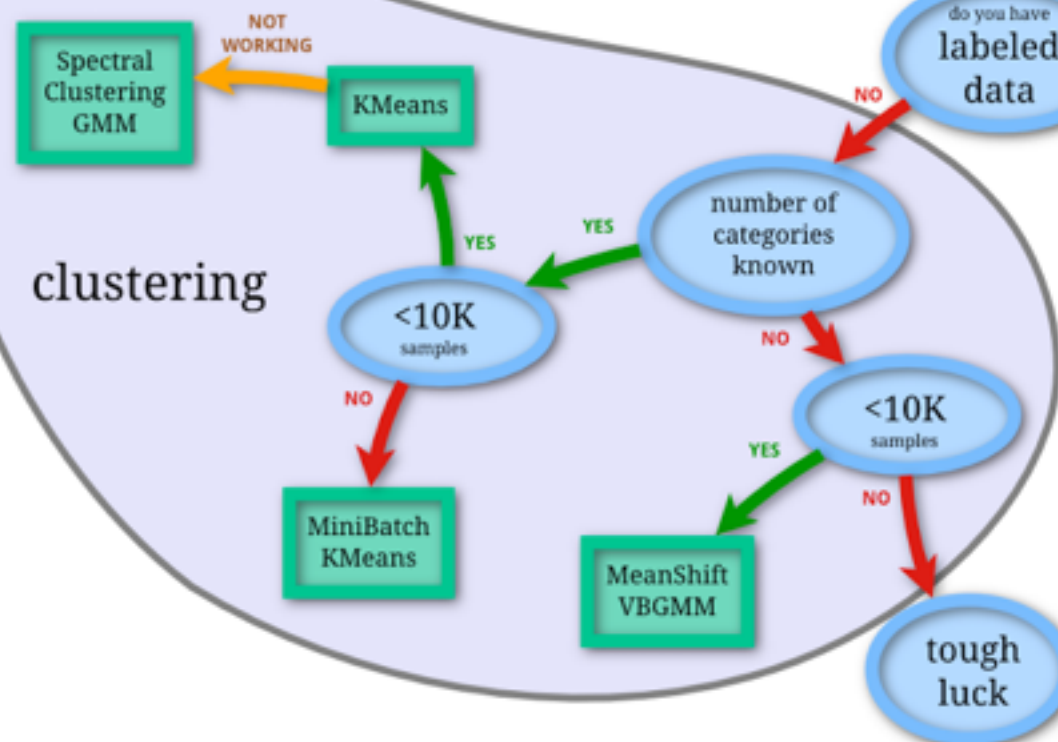
classification



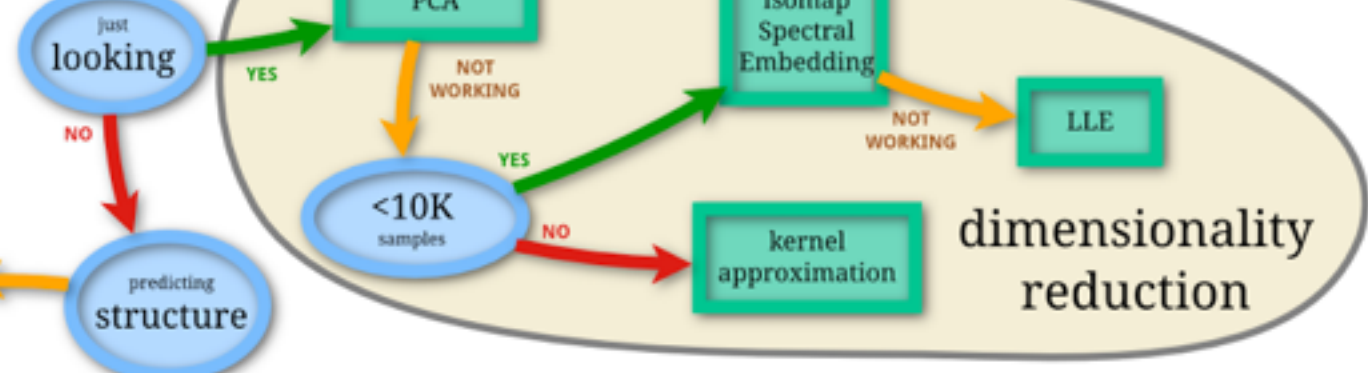
regression



clustering



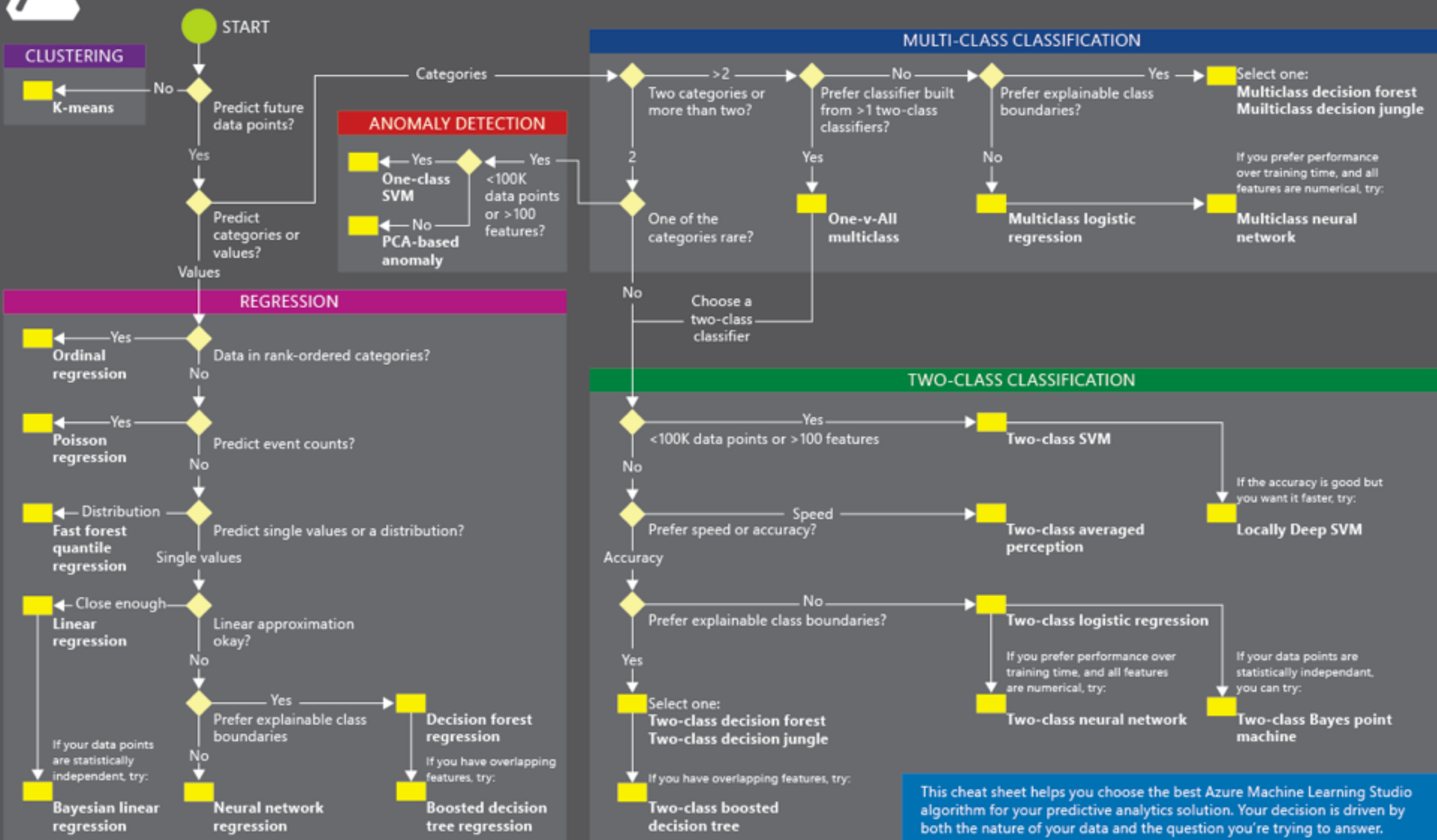
just looking



dimensionality reduction



Microsoft Azure Machine Learning: Algorithm Cheat Sheet



REVIEW

CONCLUSION