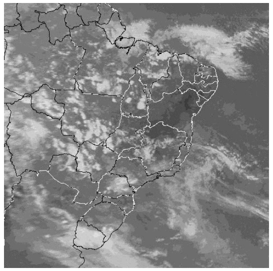


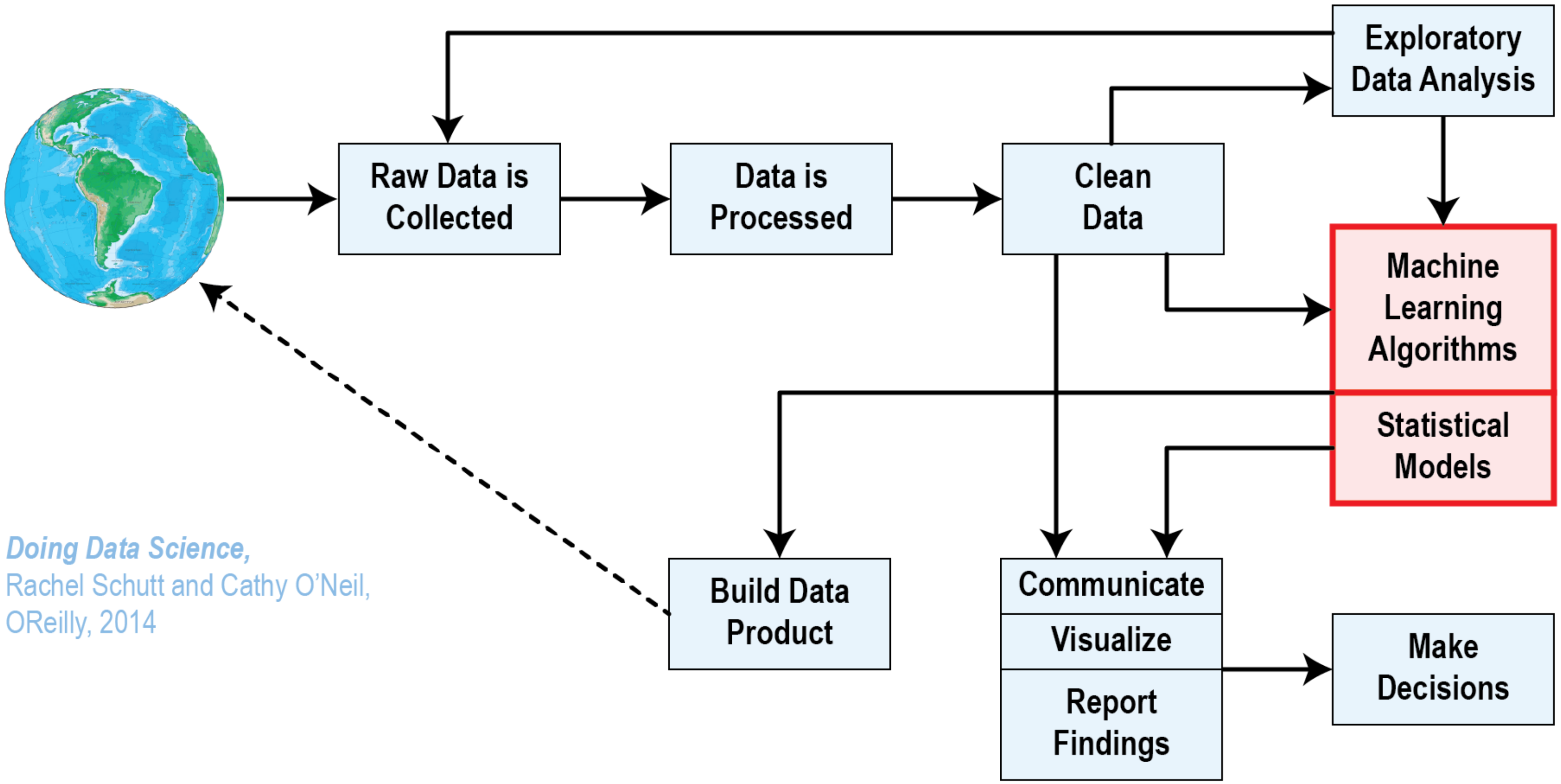
ELAC 2018 INTRODUCTION TO DATA SCIENCE

Day 4

Rafael Santos – rafael.santos@inpe.br
www.lac.inpe.br/~rafael.santos/talks.html

About this Lecture

Where are we?



Doing Data Science,
Rachel Schutt and Cathy O'Neil,
OReilly, 2014

Machine Learning

Classification

Classification

- Prediction of a category or discrete label.
- Model or Classifier creation:
  - Input: instances with known classes.
  - Output: model based on the data and algorithm.
- Classification:
  - Input: unlabeled data.
  - Output: labels for the unlabeled data based on the model.
- Post-processing: model evaluation.

Classification

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125	No
2	No	Married	100	No
3	No	Single	70	No
4	Yes	Married	120	No
5	No	Divorced	95	Yes
6	No	Married	60	No
7	Yes	Divorced	220	No
8	No	Single	85	Yes
9	No	Married	75	No
10	No	Single	90	Yes

We want to predict who will cheat on taxes based on other attributes.

Sort

Tid	Refund	Marital Status	Taxable Income	Cheat
7	Yes	Divorced	220	No
2	No	Married	100	No
4	Yes	Married	120	No
6	No	Married	60	No
9	No	Married	75	No
1	Yes	Single	125	No
3	No	Single	70	No
5	No	Divorced	95	Yes
8	No	Single	85	Yes
10	No	Single	90	Yes

Classification

Tid	Refund	Marital Status	Taxable Income	Cheat
2	No	Married	100	No
6	No	Married	60	No
9	No	Married	75	No
3	No	Single	70	No
7	Yes	Divorced	220	No
4	Yes	Married	120	No
1	Yes	Single	125	No
5	No	Divorced	95	Yes
8	No	Single	85	Yes
10	No	Single	90	Yes

Sort 2

Sort 1

Sort 1

Sort 2

Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No

Classification

Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No

Sort 1 Sort 2

Sort 1 Sort 2

Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No

Classification

- Bad rule: nobody cheats: 3/10 errors.
- Bad rule: those who don't get refunds, cheat: 4/10 errors.
- Better rule: if 85 ≤ income ≤ 100 then cheat: 1/10 errors.
- Even better rule: if 85 ≤ income ≤ 95 then cheat: 0/10 errors.

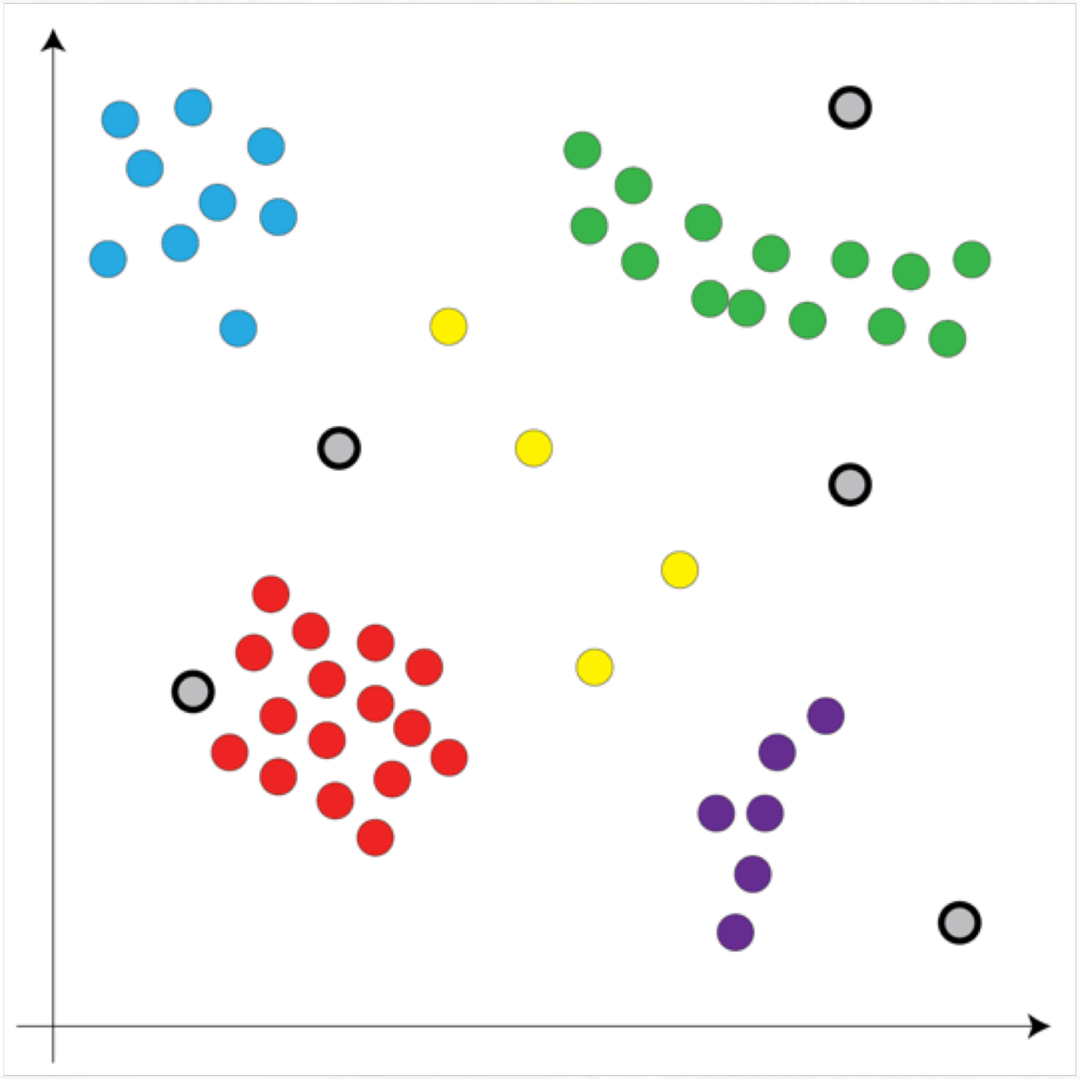
Tid	Refund	Marital Status	Taxable Income	Cheat
6	No	Married	60	No
3	No	Single	70	No
9	No	Married	75	No
8	No	Single	85	Yes
10	No	Single	90	Yes
5	No	Divorced	95	Yes
2	No	Married	100	No
4	Yes	Married	120	No
1	Yes	Single	125	No
7	Yes	Divorced	220	No

- Another perfect rule: if 75 ≤ income ≤ 95 *and* marital status is {single or divorced} then cheat: 0/10 errors.
- Another perfect rule: if 75 ≤ income ≤ 95 *and* marital status is {single or divorced} *and* refund is no then cheat: 0/10 errors.

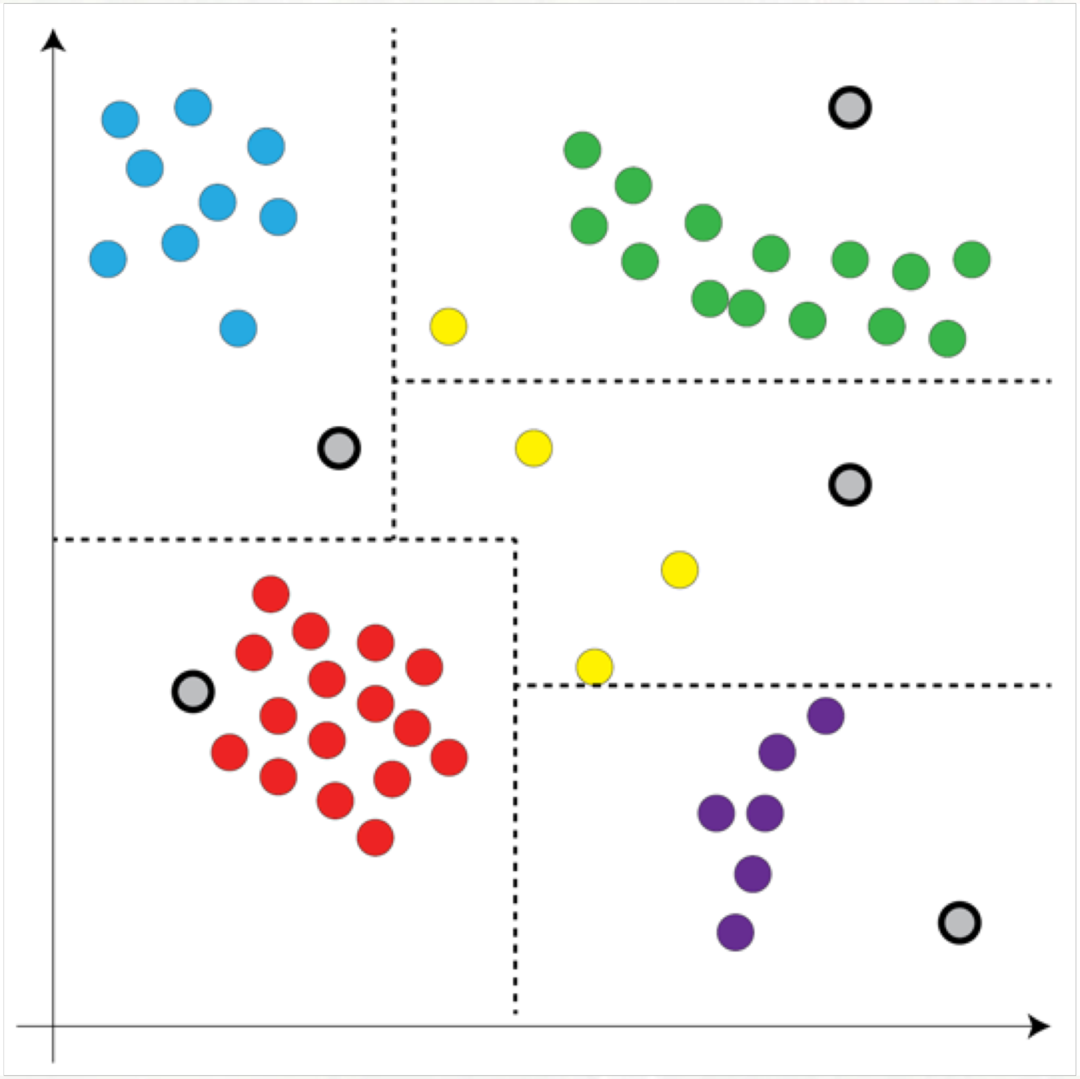
What do we want from a classifier?

- Classify unknown data.
  - Model must be robust enough to deal with previously unknown data – generalization!
- Explain our data, e.g. using statistics and rules.
  - Eventually there is no need to explain all data in intricate details: generalization again!

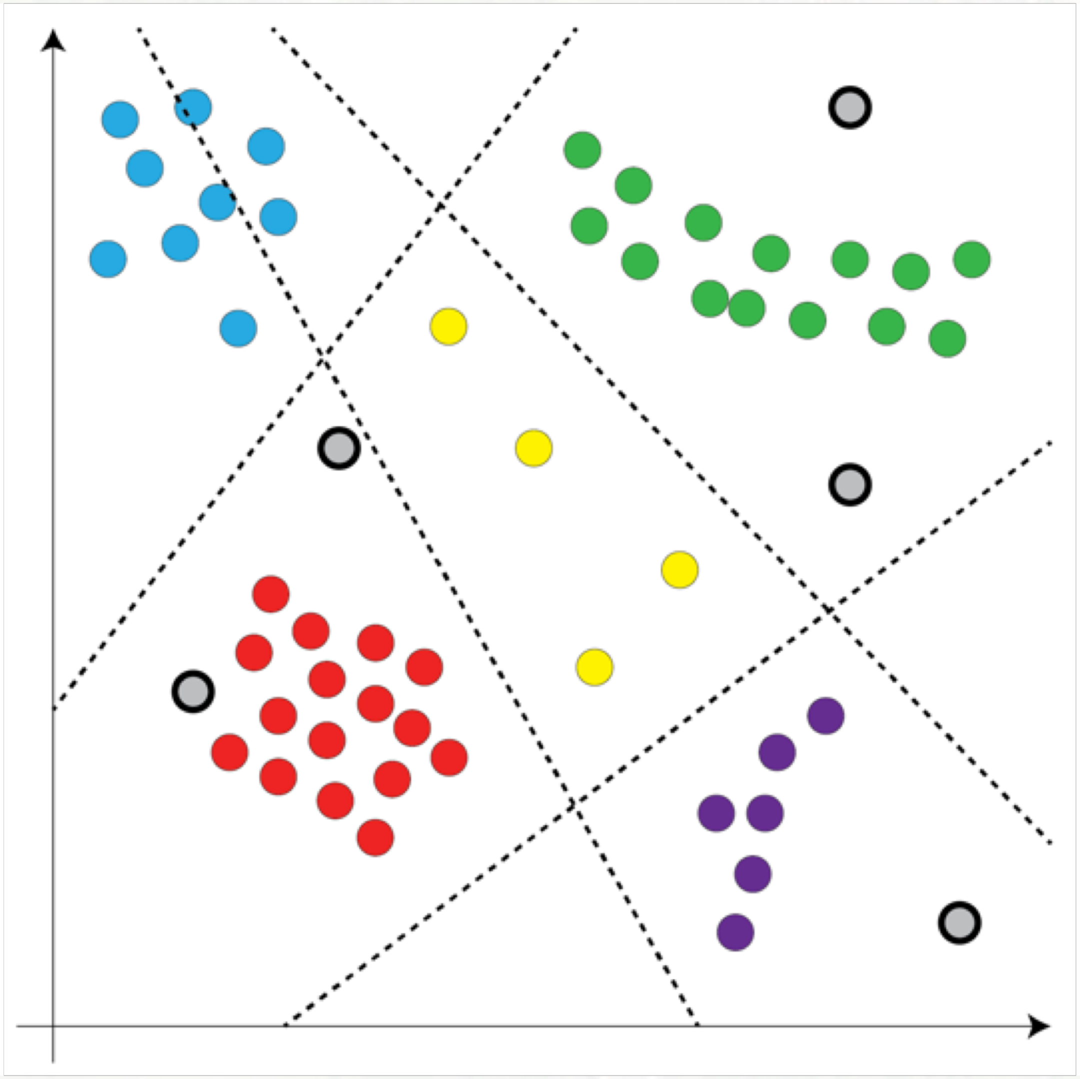
Classification



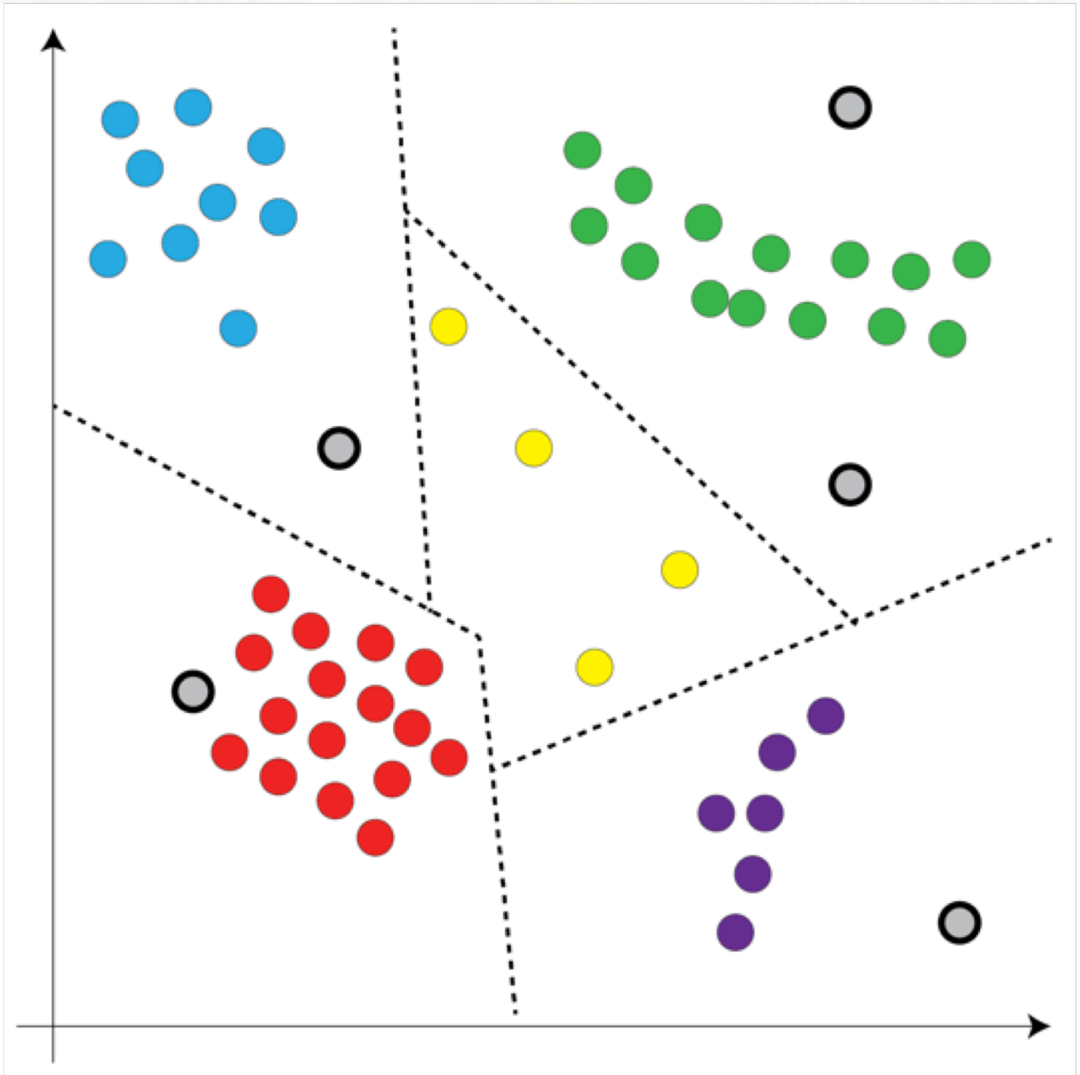
Classification



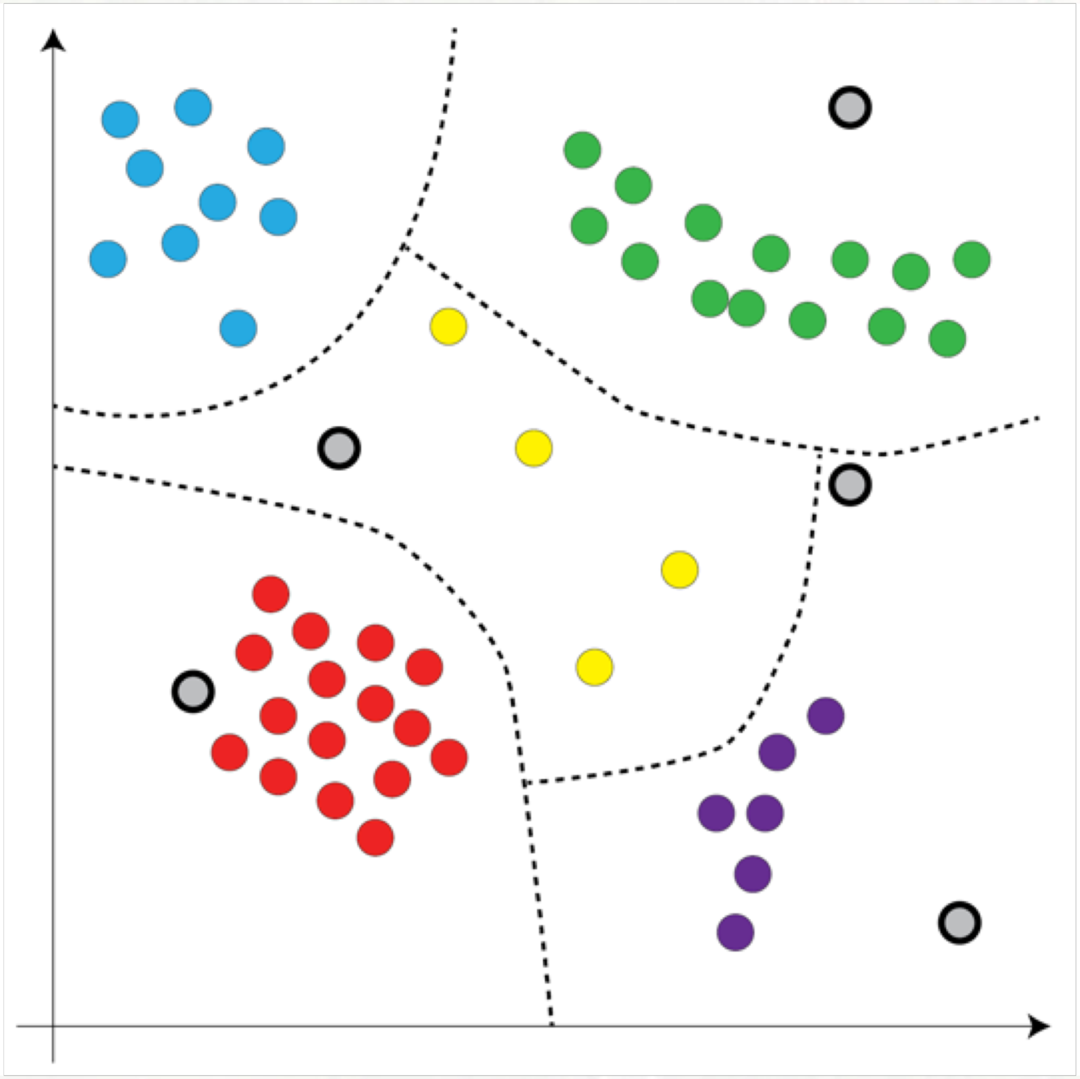
Classification



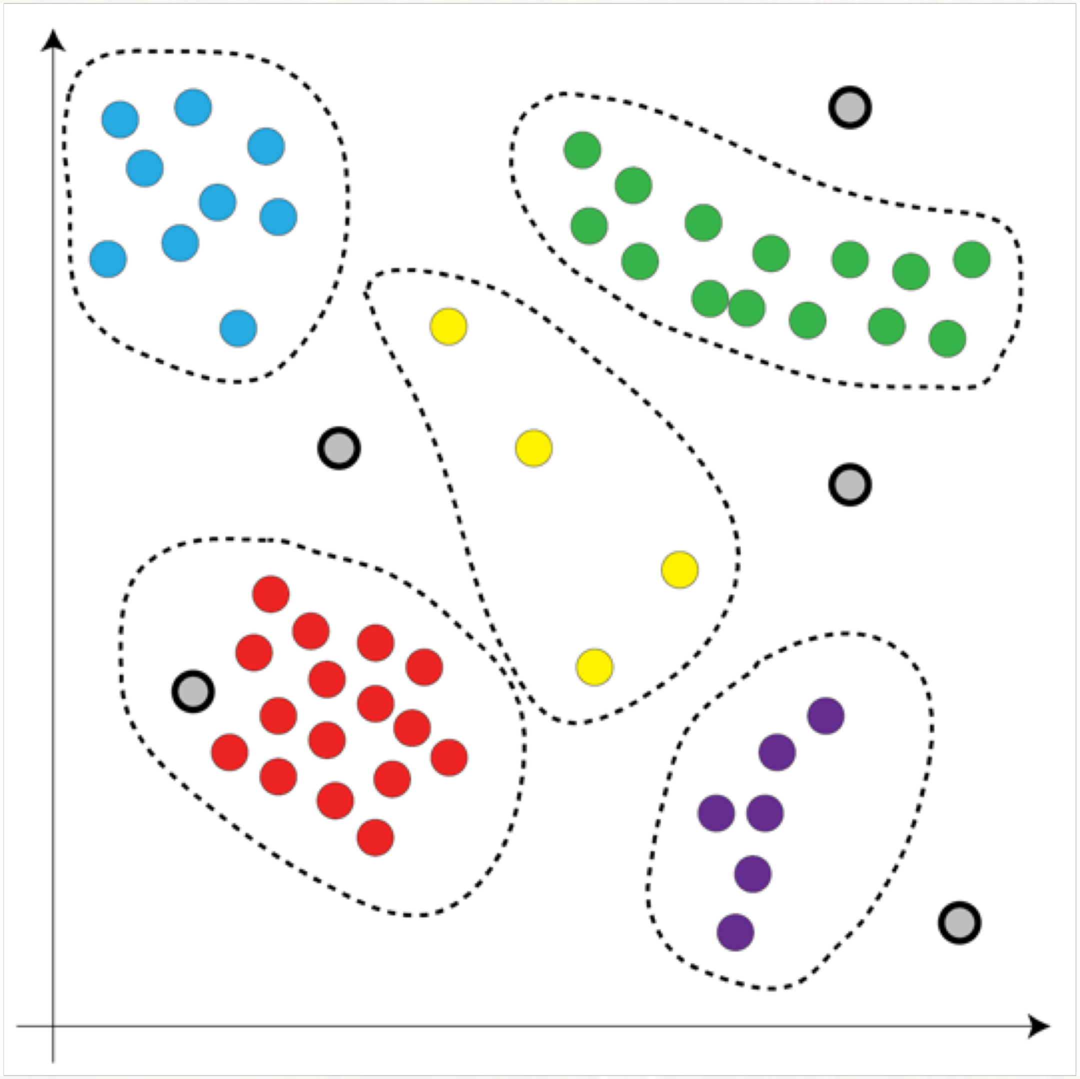
Classification



Classification



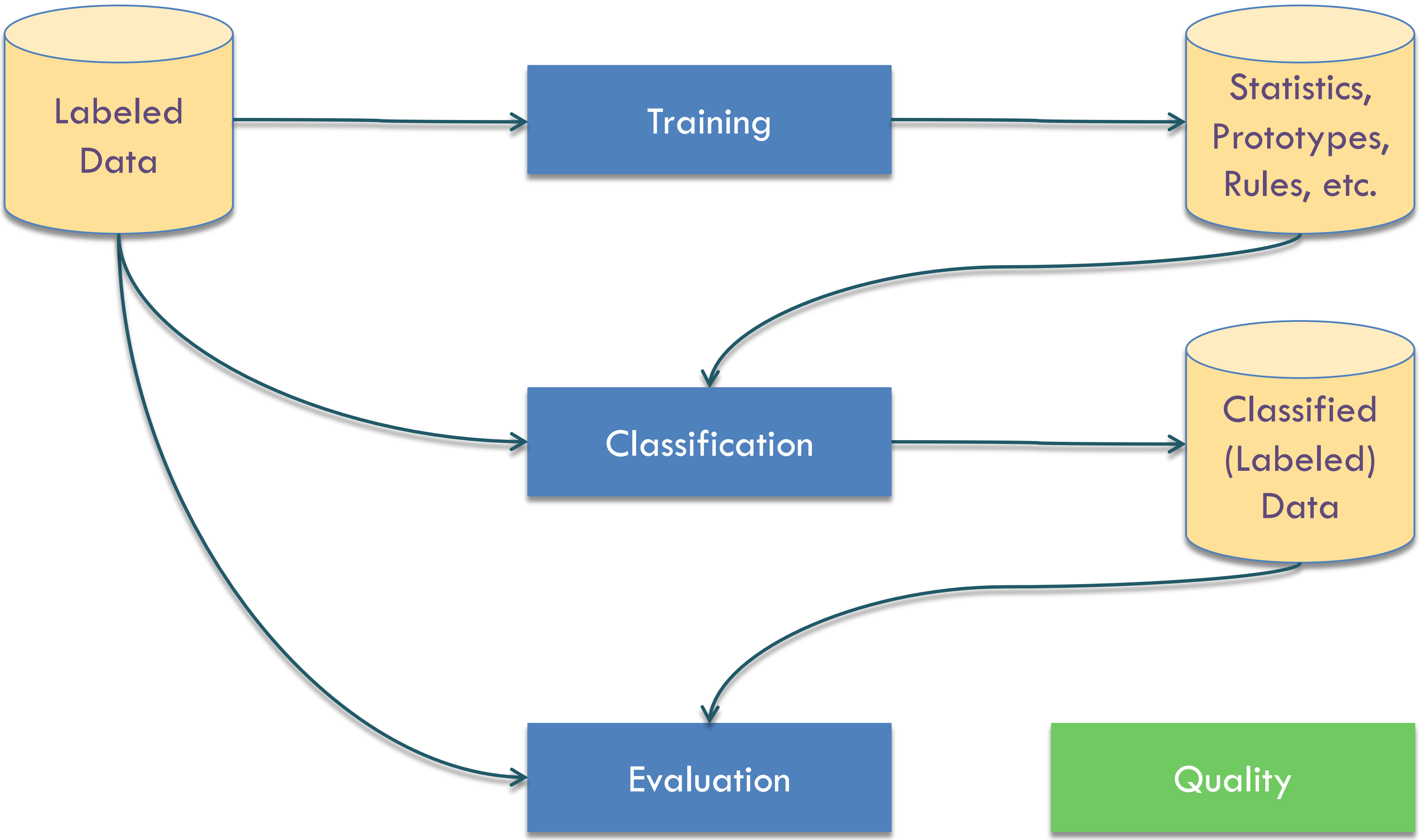
Classification



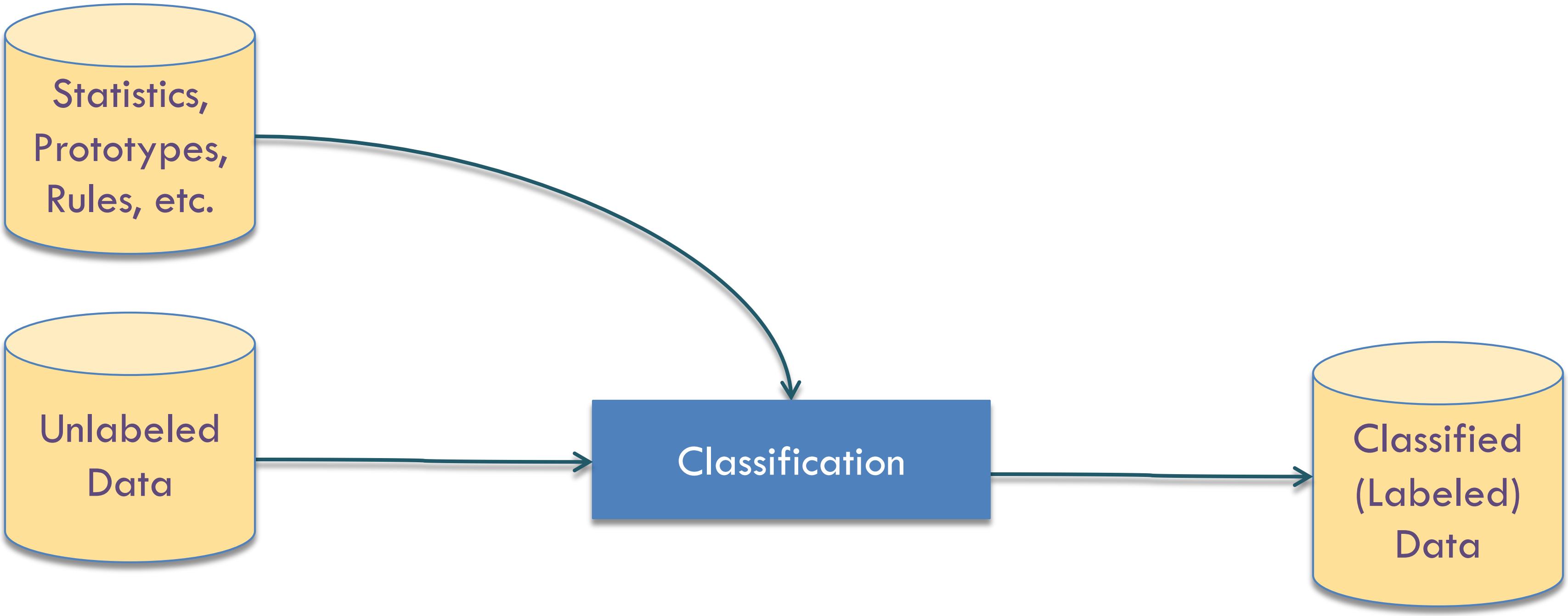
How to create a model?

- Different algorithms creates different models.
- Some models are inherently more precise, some are easier to understand.
- Some models are compact, some are extensive.
- Which is better?

Classification



Classification



Classification: Evaluation

- A simple evaluation technique: confusion matrix.
  - Classify labeled or known data.
  - Usually data used for training or a subset of it

Was classified as

Original Label is

	A	B	C
A	50	0	0
B	20	155	25
C	10	0	190

- Accuracy: Correct Classifications/All Classifications
  - 395/450 = 87.78%

More Metrics from Confusion Matrix

- Recall for a class (sensitivity or true positive rate)
  - Of the classified as ***X*** how many are really ***X*** (i.e. not other classes in X's boundary)?
  - TP/(TP+FN): 0.6250 for A; 1.0000 for B; 0.8837 for C
- Precision for a class (positive predictive value)
  - Of all the ***X*** how many were classified as ***X*** (i.e. not misclassified)?
  - TP/(TP+FP): 1.0000 for A; 0.7750 for B; 0.9500 for C

- True Positives
- True Negatives
- False Positives
- False Negatives

- For ***B***:

Classified as

Original

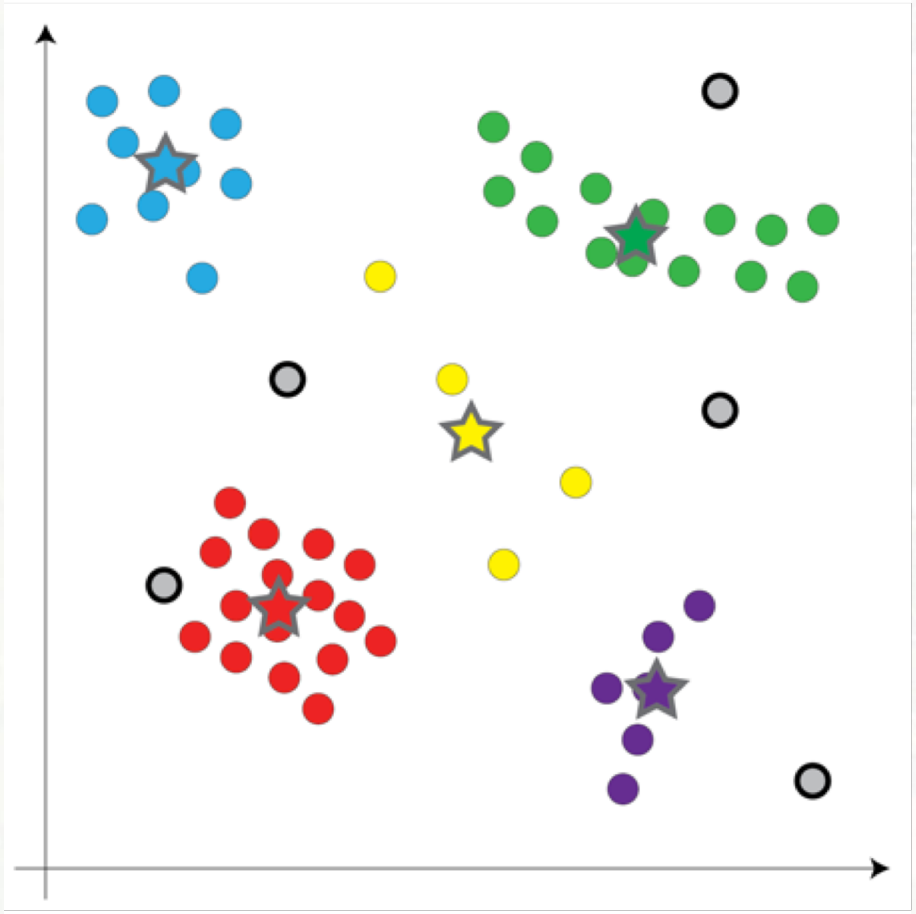
	A	**B**	C
A	50	0	0
B	20	155	25
C	10	0	190

Ideas from the evaluation process

- Labeled data: does it *really* corresponds to samples for a class?
- Are there mixed classes in our labels for class X?
- Are there really N classes (instead of N±n)?

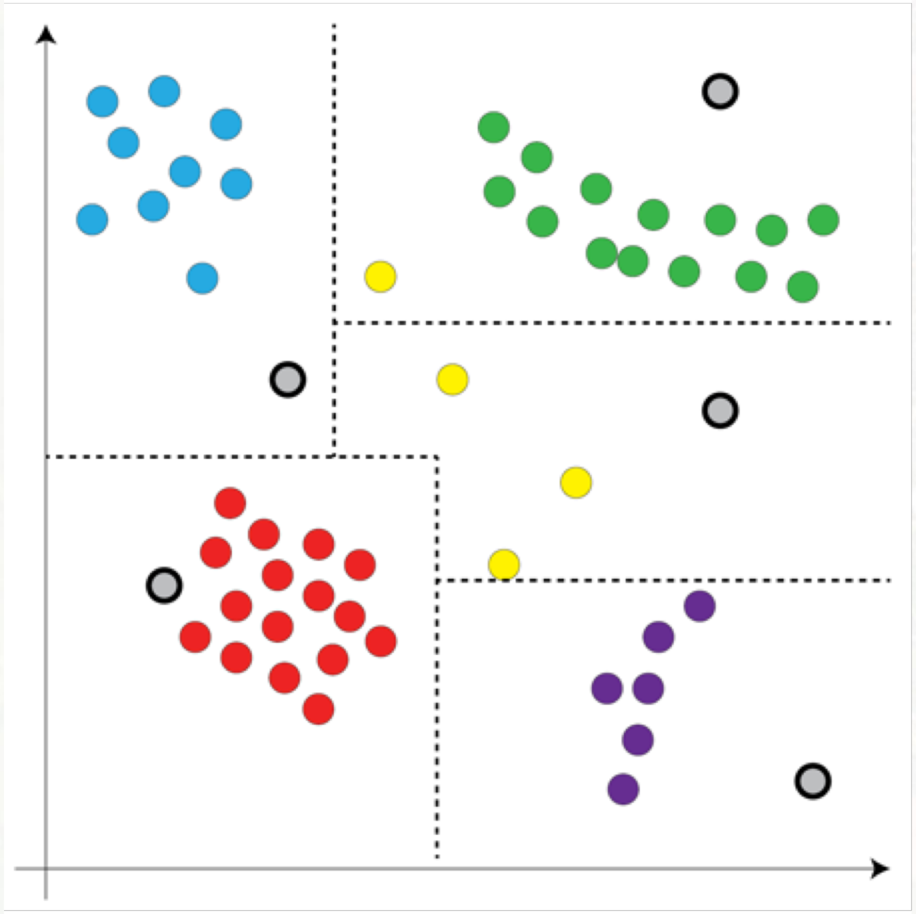
Classification: Minimum Distance

- Model is the average of the data points (geometric center).
- Class is determined from the minimum distance to center.
- Other metrics may be used.

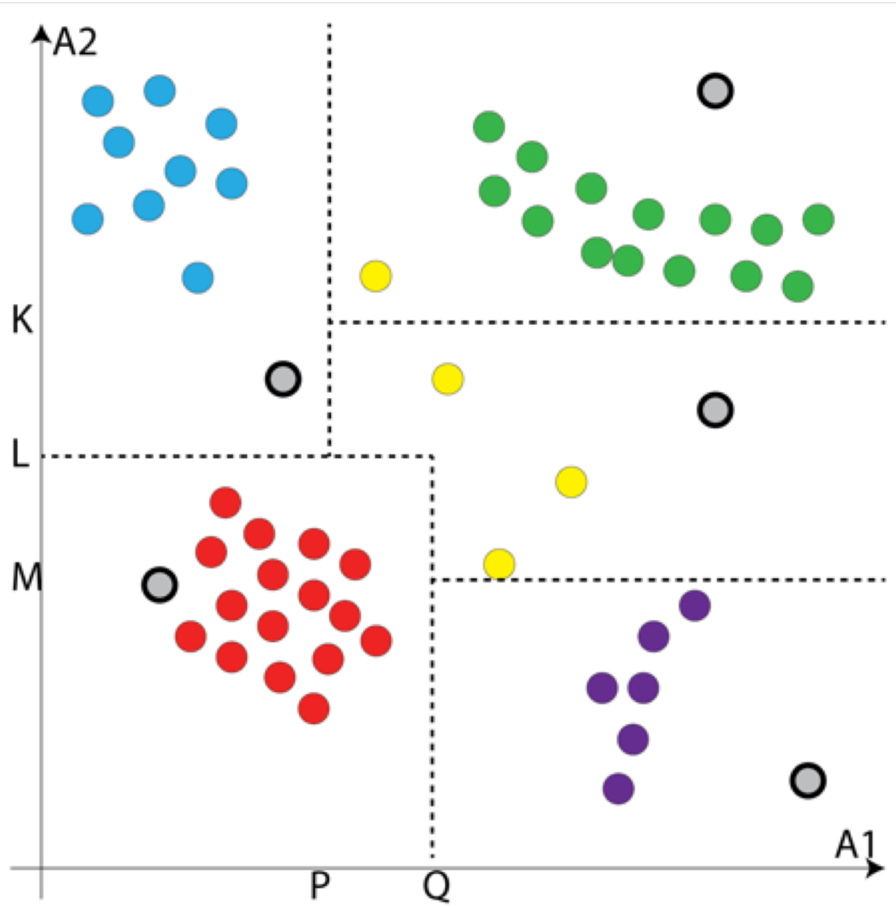


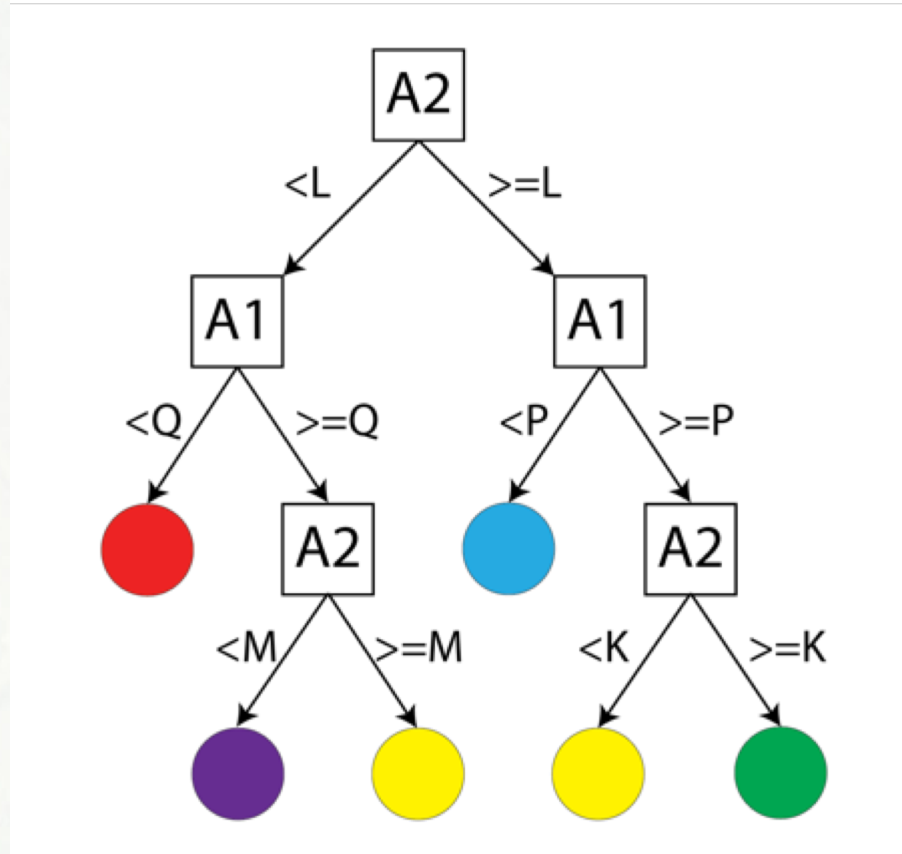
Classification: Decision Tree

- Model is the set of decision rules that best separates the classes.
- Class is determined from evaluation of the rules.



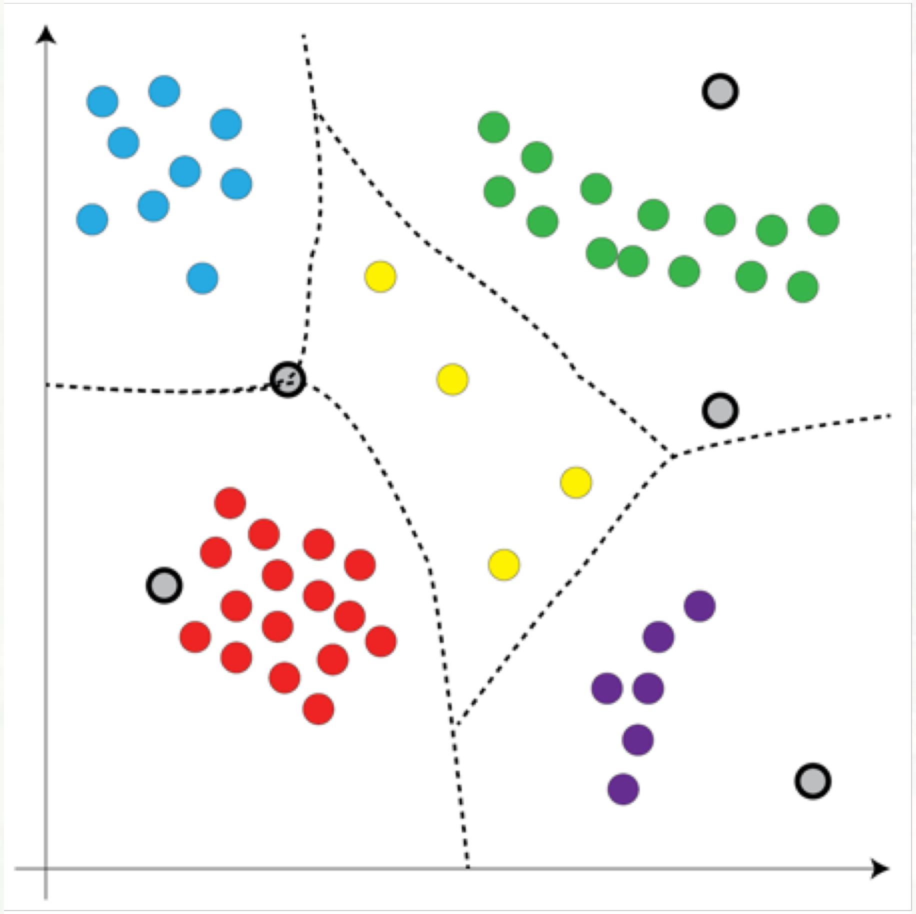
Classification: Decision Tree





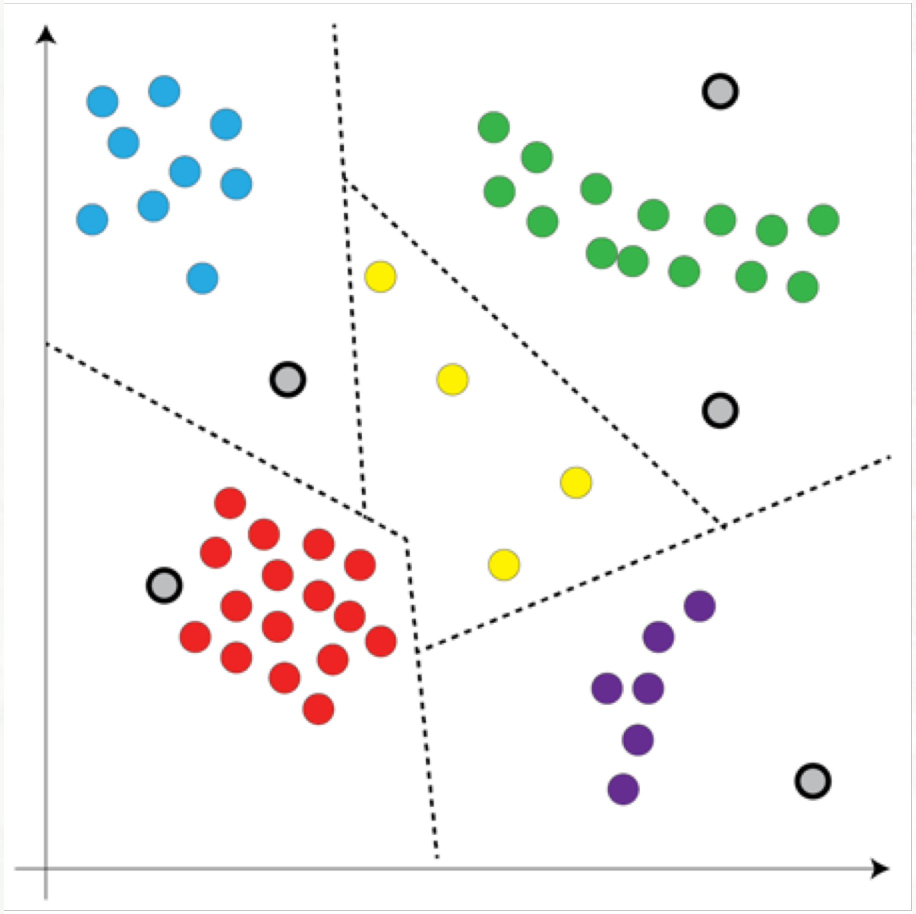
Classification: Nearest Neighbors

- No Model: uses labeled data points themselves.
  - Computationally intensive.
- Class is determined from majority of labeled nearest neighbors.



Classification: Neural Networks (MLPs)

- Model: parameters of a neural network, trained to separate classes.
  - Model is hard to understand.
  - Underfitting/Overfitting problem.



Clustering

Clustering

- Methods that find natural groups in data.
  - Data in the same cluster or group are somehow similar.
  - Data in different groups are somehow different.
  - We don’t need labels to train a classifier!
- Problems:
  - How to define similarity?
  - How many groups do we have?

Clustering

- Input:
  - Data
  - Similarity metrics, attributes
  - Number of clusters
- Output:
  - Assignment for each data point to each cluster (hard or soft)
  - Clustering quality metrics

K-Means

- Iterative algorithm:
  1. Start with K groups
  2. Calculate centers of groups based on membership
  3. Assign data to groups
  4. Repeat 2-4 until stopping condition

$$v_i = \frac{1}{n_i} \sum_{x_k \in C_i} x_k$$

$$J = \sum_{k=1}^{n} \sum_{x_k \in C_i} |x_k - v_i|^2$$

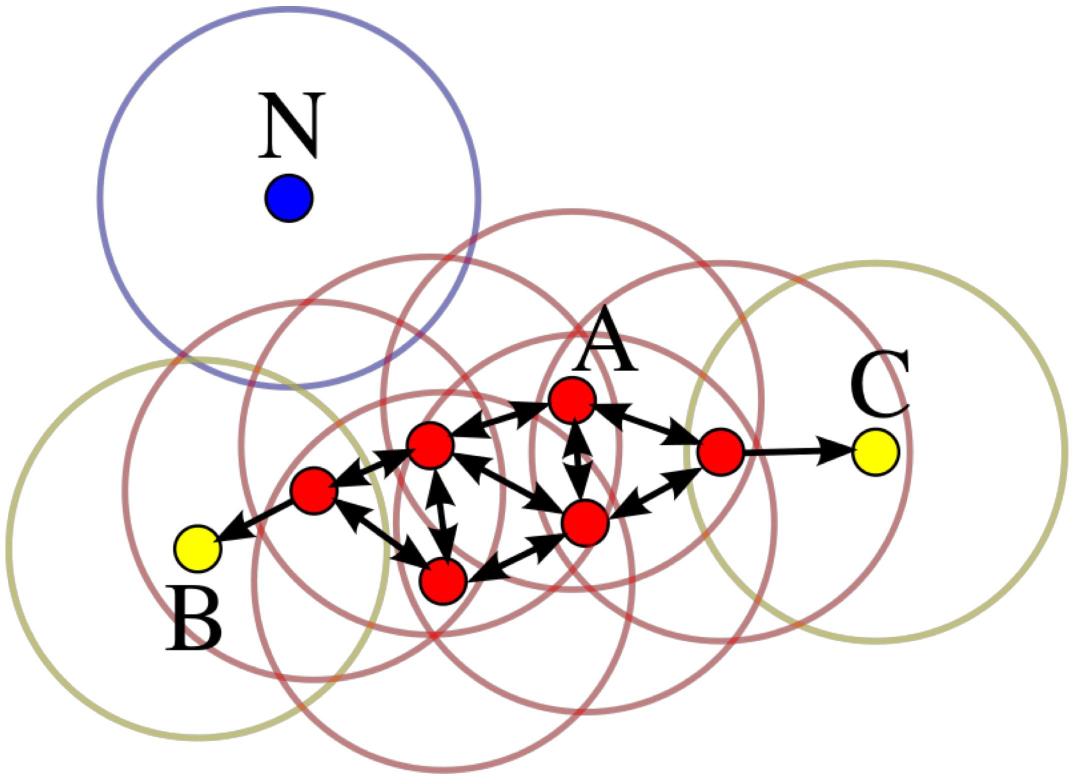
Fuzzy C-Means

- Similar to K-Means, but it uses a membership table.
- For cluster analysis, usually last step is defuzzification, but..
  - Allows "Plan B" clustering.
  - We can cut assignment to clusters if max(membership) is too low.

Instância	Classe A	Classe B	Classe C	Classe D
1	0.31	0.19	0.50	0.00
2	0.08	0.01	0.74	0.17
3	0.25	0.24	0.26	0.25
4	0.99	0.00	0.00	0.01
5	0.50	0.50	0.00	0.00

DBScan

- Identification of density-based clusters:
  - Find core points
  - Test reachability
  - Identify noise



Wikipedia

Hierarchical Clustering

- Methods that create several partitions in the data:
- Top-down: starts with all data in a single cluster, partitions every cluster until each data is in a single cluster.
- Bottom-up: starts with each data in a single cluster, merges data+clusters until all data is in a single cluster.

Hierarchical Clustering

	X	Y
1	0,25	0,27
2	0,32	0,91
3	0,33	0,80
4	0,18	0,33

	1	2	3	4
1	0,000	**0,644**	**0,536**	**0,092**
2	0,644	0,000	**0,110**	**0,597**
3	0,536	0,110	0,000	**0,493**
4	0,092	0,597	0,493	0,000

	X	Y
1+4	0,22	0,30
2	0,32	0,91
3	0,33	0,80

	1+4	2	3
1+4	0,000	**0,619**	**0,513**
2	0,619	0,000	**0,110**
3	0,513	0,110	0,000

	X	Y
1+4	0,22	0,30
2+3	0,33	0,86

	1+4	2+3
1+4	0,000	**0,566**
2+3	0,566	0,000

	X	Y
1+2+3+4	0,27	0,58