Machine Learning:
Where did this Random Forest come from?

Biostats Club

Dan Bunis

February 21, 2019

Machine Learning- da fuq?

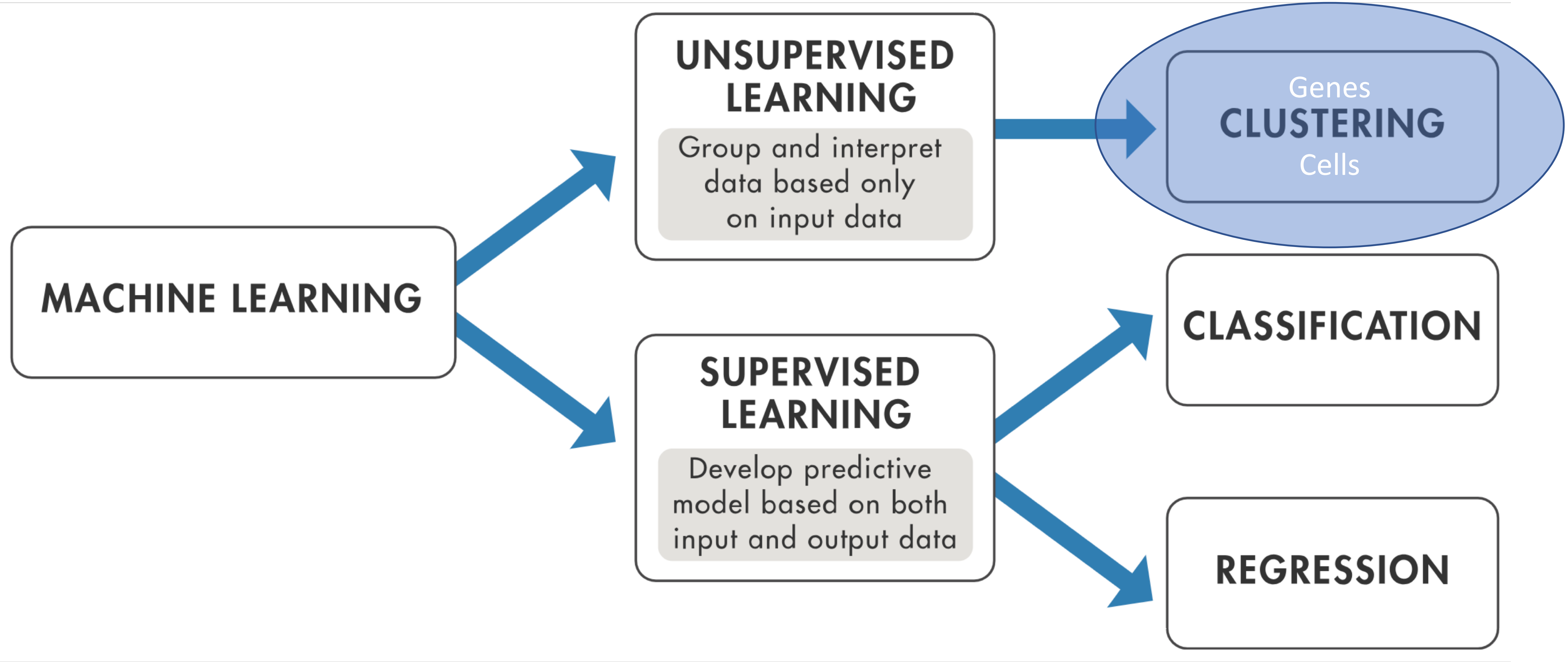
- Machine learning is any computational technique that involves “learning from experience”.
- Machine learning algorithms “learn” information directly from data, (often) without relying on a predetermined equation as a model.

- If iteration 2 involves information that came from iteration 1, that’s machine learning!

https://www.mathworks.com/discovery/machine-learning.html

We actually use machine learning all the time in analysis of big data without even thinking about it!

What can/do I do with machine learning?



MACHINE LEARNING

UNSUPERVISED LEARNING

Group and interpret data based only on input data

SUPERVISED LEARNING

Develop predictive model based on both input and output data

Genes

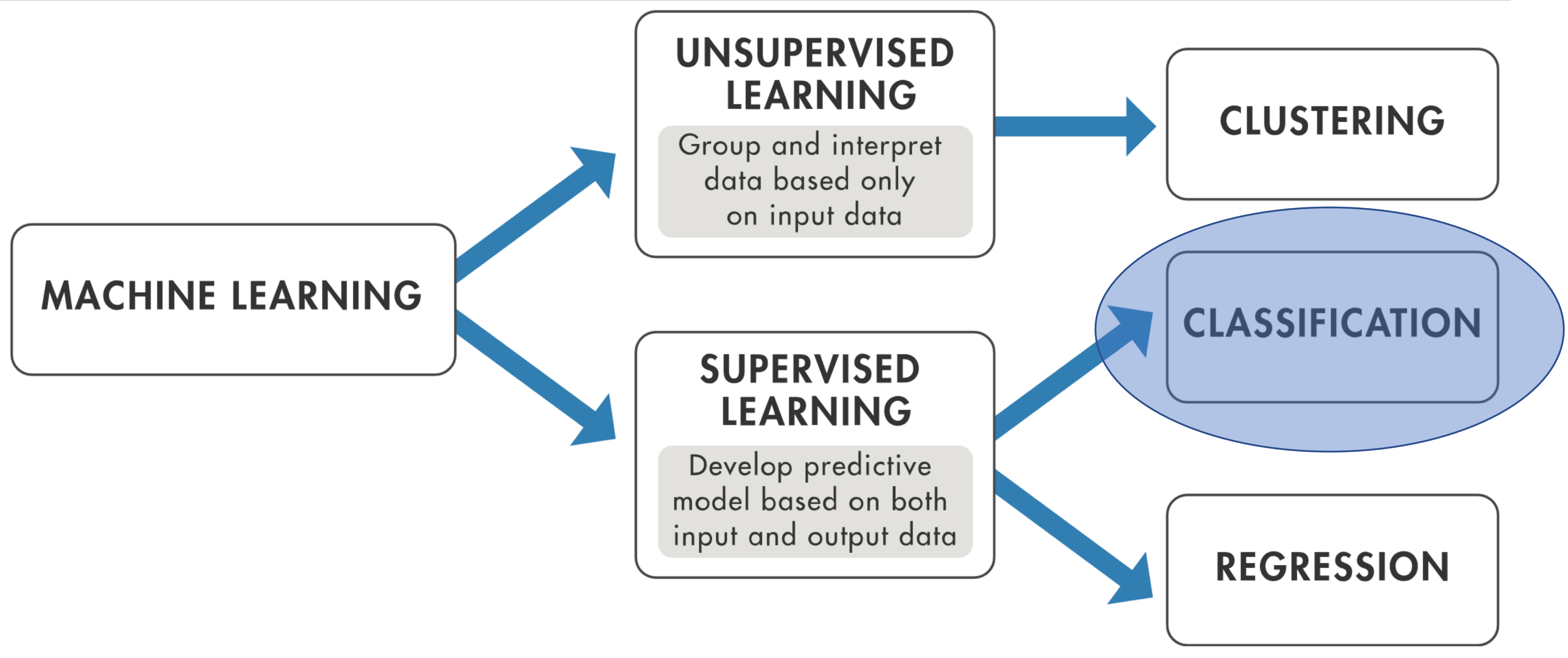
CLUSTERING

Cells

CLASSIFICATION

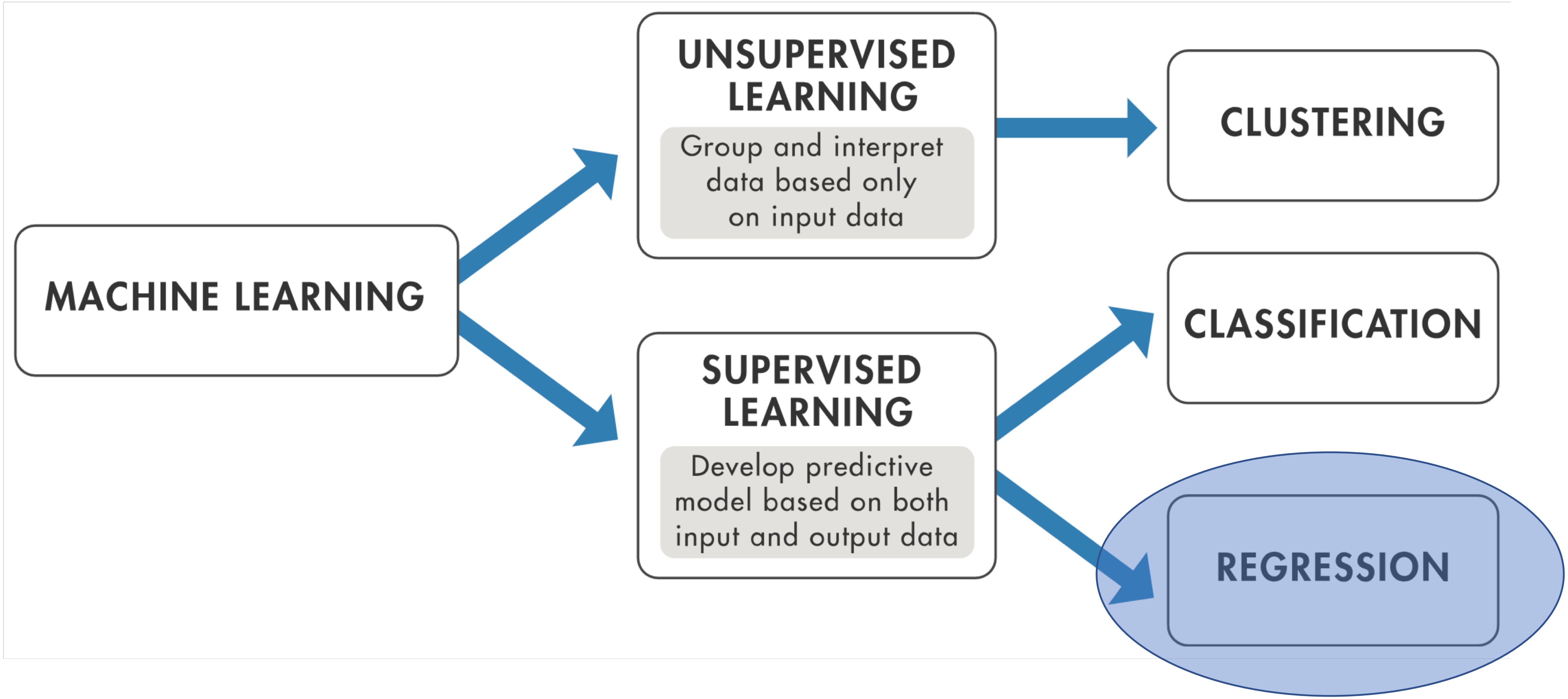
REGRESSION

What can/do I do with machine learning?



MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

What can/do I do with machine learning?



MACHINE LEARNING

UNSUPERVISED LEARNING

Group and interpret data based only on input data

SUPERVISED LEARNING

Develop predictive model based on both input and output data

CLUSTERING

CLASSIFICATION

REGRESSION

Unsupervised Learning: Group and interpret data based only on input data

- Why? To infer hidden associations or structure in the dataset

- Common Uses:
  - Hierarchical Clustering of cells / genes for a heatmap
  - Clustering of cells in single cell RNAseq data
  - Picking peaks that are above background level in ChIPseq data

Supervised Learning: Develop a predictive model based on input & output data

- Classification OR Regression
- Why? Build a model for scoring or classifying new datapoints!

- Common Uses:
  - Novel Gene Prediction (based on previous genomes, known genes/organisms)
  - Disease risk prediction (based on healthy vs diseased training sets)
  - Clustering of cells in single cell data (?)
  - Cell type classification (scRNAseq, based on pure reference datasets [Singler])

Not all highly computational analyses are machine learning

- t-SNE: YES
- PCA: NO
- Statistical significance testing: NO (most of the time)

10 tips for biological machine learning

1. **Adjust your data's structure:**
    - Start with enough data:
        - ideal = 10 samples for every feature
        - These both seem super low! (???)
    - Shuffle the order of your data
        - Removes any possibility of trends related to order
    - Clean the data
        - Remove corrupt, inaccurate, inconsistent, or outlier values
        - Normalize (= scale function) if needed (not needed for random forest)

Cicco D. "Ten quick tips for machine learning in computational biology," Bio Data Mining. 2017

10 tips for biological machine learning

2. Split into Training, Validation, and Test sets

- Common breakdown:
  - 50% Training
  - 30% Validation <- not always needed
  - 20% Test
- Training & validation: for training / adjusting “hyperparameters”.
- Test set: ONLY touch once you are all done.
- “the lock box approach should be employed by every machine learning project in every field”

Cicco D. “Ten quick tips for machine learning in computational biology,” Bio Data Mining. 2017

10 tips for biological machine learning

3. Understand the types of machine learning

- Unsupervised / supervised
  - Classification / Regression

- Unsupervised:
  - K-means clustering, Truncated singular value decomposition, Probabilistic latent semantic analysis, and more!
- Supervised:
  - Support Vector Machines, K- Nearest Neighbors, Regression modeling, Random Forest, and more!

Cicco D. “Ten quick tips for machine learning in computational biology,” Bio Data Mining. 2017

10 tips for biological machine learning

4. Pick the “right” algorithm: Start with the simplest

- Suggested that you use multiple, but 🤷‍♂️
- If you start simple, you can learn & you can probably manage the debugging

Cicco D. “Ten quick tips for machine learning in computational biology,” Bio Data Mining. 2017

10 tips for biological machine learning

3. Understand the types of machine learning

- Unsupervised / supervised
  - Classification / Regression

- Unsupervised:
  - K-means clustering, Truncated singular value decomposition, Probabilistic latent semantic analysis, and more!
- Supervised:
  - Support Vector Machines, K- Nearest Neighbors, **Random Forest**, and more!

Cicco D. "Ten quick tips for machine learning in computational biology," Bio Data Mining. 2017

10 tips for biological machine learning

5. **Take care of the imbalanced data problem**

- If 90% yes, 10% no: the algorithm may not learn to pick No’s well
- “Balance” your data
  - Ensure a 70%/30% split
  - (90%+50%)/2 & (10%+50%)/2

Cicco D. “Ten quick tips for machine learning in computational biology,” Bio Data Mining. 2017

10 tips for biological machine learning

6. Optimize your hyper-parameters

- Hyper-parameter = a parameter that the algorithm uses, but does not train
    - Ex: # groups in k-means, # of “neighbors” in k-NN clustering
- Method 1: Use a grid of values, then score with validation set
- Method 2: Use an automatic selection method, like X-means

Cicco D. “Ten quick tips for machine learning in computational biology,” Bio Data Mining. 2017

10 tips for biological machine learning

7. **Minimize overfitting**

- = algorithm memorizes the training set instead of learning the concepts required for the test
- = Great classification of training set, not great performance on validation/test set

1. Cross validation

2. Be aware of the possibility, and adjust for it when you see it.

Cicco D. “Ten quick tips for machine learning in computational biology,” Bio Data Mining. 2017

10 tips for biological machine learning

8. **Score the model with the Matthews correlation coefficient (MCC) or the Precision-Recall curve**

- MCC: has penalization when accuracy is low for certain groups
  - Important because we often have many negatives and few positives in biology!
  - Simple accuracy: If 95% of data is truly 1, and 5% is 2. A model that always goes with 1 would be 95% accurate.

- **My advice: Google the proper test, and consider step 10.**

Cicco D. "Ten quick tips for machine learning in computational biology," Bio Data Mining. 2017

10 tips

9. Use open source over proprietary algorithms

- Access to the actual code
- Enables collaboration regardless of license access

Cicco D. “Ten quick tips for machine learning in computational biology,” Bio Data Mining. 2017

10 tips

10. Get feedback from experts

- Statisticians think we don’t know what we are doing lol
  - It’s often true!

Cicco D. “Ten quick tips for machine learning in computational biology,” Bio Data Mining. 2017

10 tips: This become 5 separate steps

1. **Adjust your data's structure:**
2. **Split into Training, (Validation,) and Test sets**
3. **Understand the types of machine learning**
4. **Pick the "right" algorithm: Start with the simplest**
5. **Take care of the imbalanced data problem**
6. **Optimize your hyper-parameters**
7. **Minimize overfitting**
8. **Score the model's performance with a valid test**
9. **Use open source over proprietary algorithm**
10. **Get feedback from experts**

10 tips: This become 5 separate steps

1. **Adjust your data's structure:**
2. **Split into Training, Validation, and Test sets**
3. **Understand the types of machine learning**
4. **Pick the "right" algorithm: Start with the simplest**
5. **Take care of the imbalanced data problem**
6. **Optimize your hyper-parameters**
7. **Minimize overfitting**
8. **Score the model's performance with a valid test**
9. **Use open source over proprietary algorithm**
10. **Get feedback from experts**

Step 1

Step 2

Step 3

Step 4

Step 5

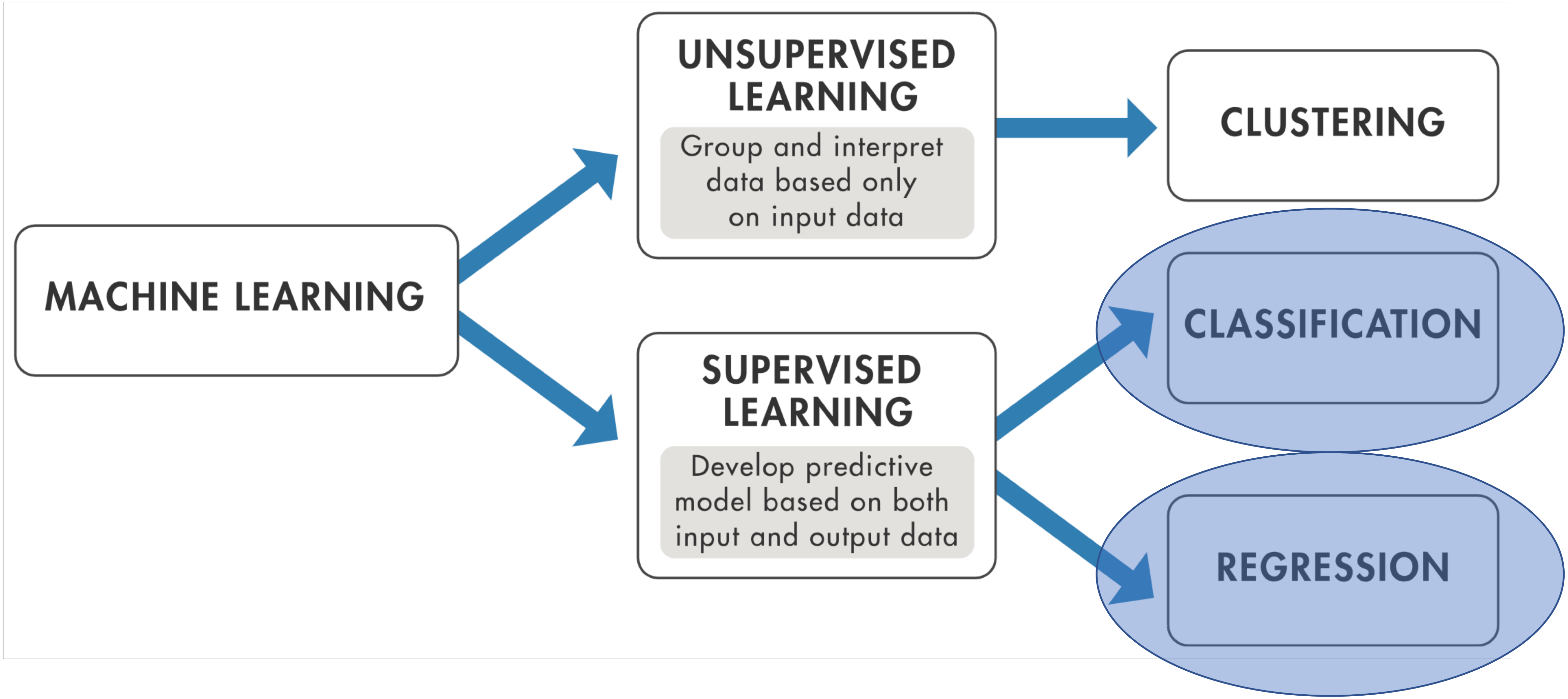
Synthesized tips

1. Use the simplest model you can!
2. Mind your model:
   - train/test/validate sets
   - Randomize data order
   - Remember assumptions & try not to over-extend your conclusions
     - Ex: if you use k-means with k=4, don’t say “the fact that we got 4 clusters is meaningful.”
3. Over-fitting: look for it, and adjust accordingly
4. Run your method by an expert
   - Example: selecting markers based on a PCA of the entire dataset = bad practice

Why Random Forest?

- Cuz trees fight climate change!
- Incredibly Powerful yet also Flexible
    - Regression or Classification

Why Random Forest?



MACHINE LEARNING

UNSUPERVISED LEARNING

Group and interpret data based only on input data

CLUSTERING

SUPERVISED LEARNING

Develop predictive model based on both input and output data

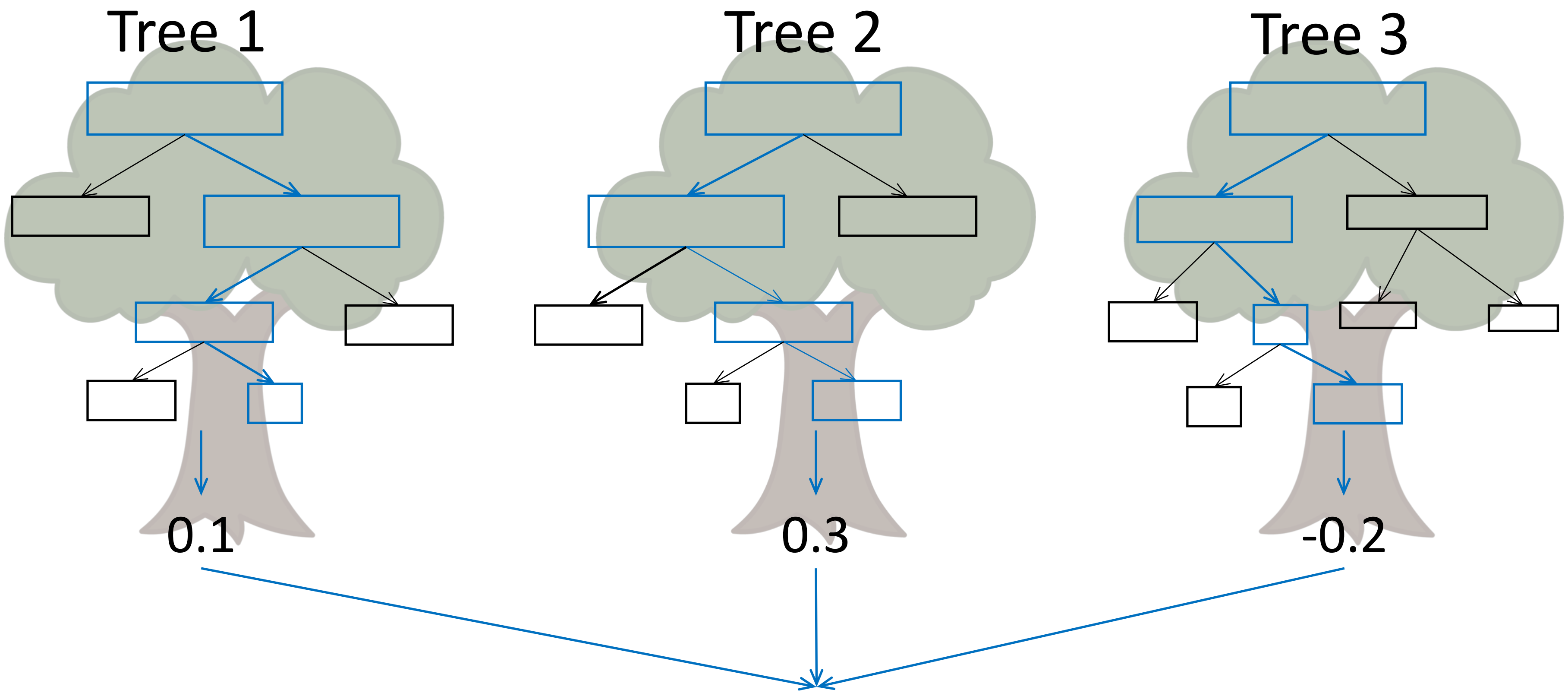
CLASSIFICATION

REGRESSION

Why Random Forest?

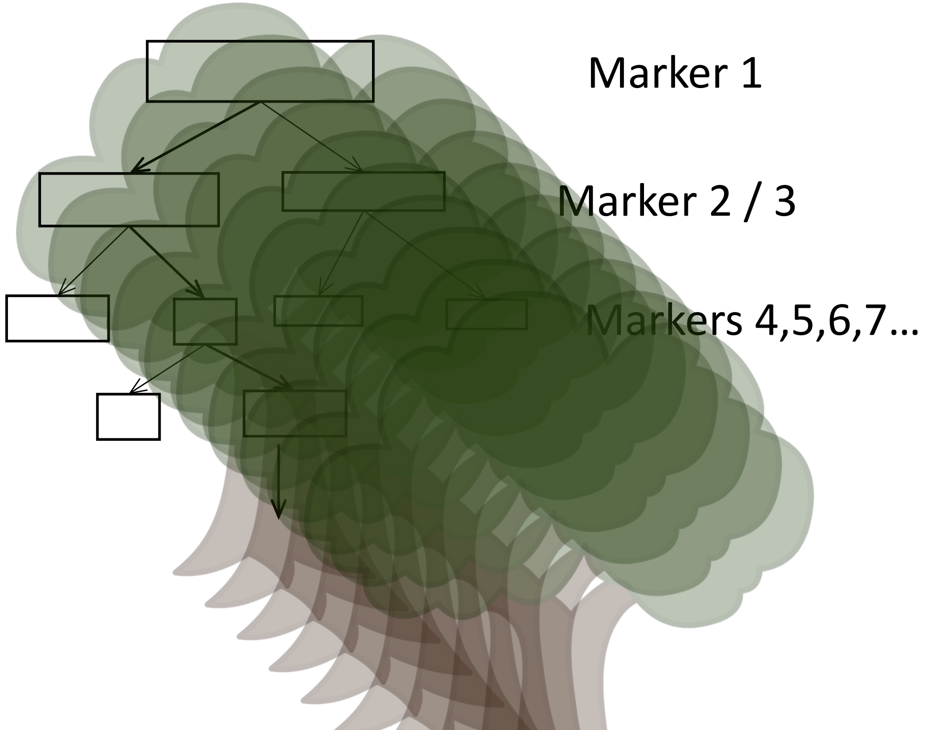
- Cuz trees fight climate change!
- Incredibly Powerful yet also Flexible
  - Regression or Classification
- Easy to use: few hyperparameters
- Training is quick
  - seconds or minutes
- Can often reverse engineer metrics about the markers used afterwards!

Why Random “Forest”?



What is “Random” Forest?

- “Random” = bagging = a random set of samples is picked
  - Code: sample(data, size = nrow(data), replace = T) → ~2/3 of data
- Each tree is iteratively built on a different random set
- 1st marker and break point are calculated
  - Selects feature & break point associated with maximal entropy reduction in the outcomes of the data groups after the split!
- Repeat for 2nd and 3rd markers/breaks…
- Done when N breaks made (=hyperparameter)
- Repeat bagged tree generation many times
  - How many trees is also a hyperparameter.


Marker 1
Marker 2 / 3
Markers 4,5,6,7…