Machine learning - what is it, and what can it offer biologists?

https://www.techemergence.com/what-is-machine-learning/

Learning objectives

Define machine learning

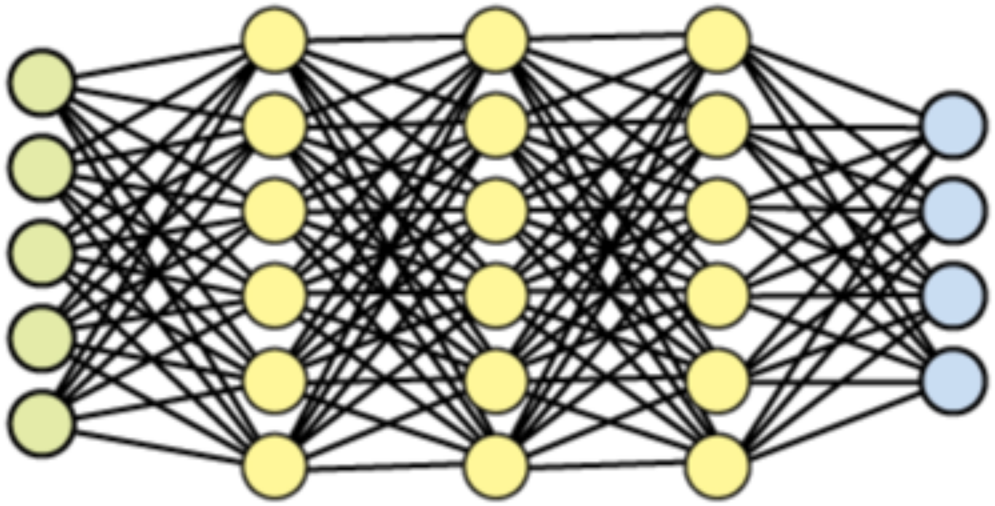
Understand the distinction between supervised and unsupervised machine learning

Identify the limits of prediction and classification schemes

Defining machine learning

“Machine learning enables computers to do tasks **without being explicitly programmed** for those tasks”

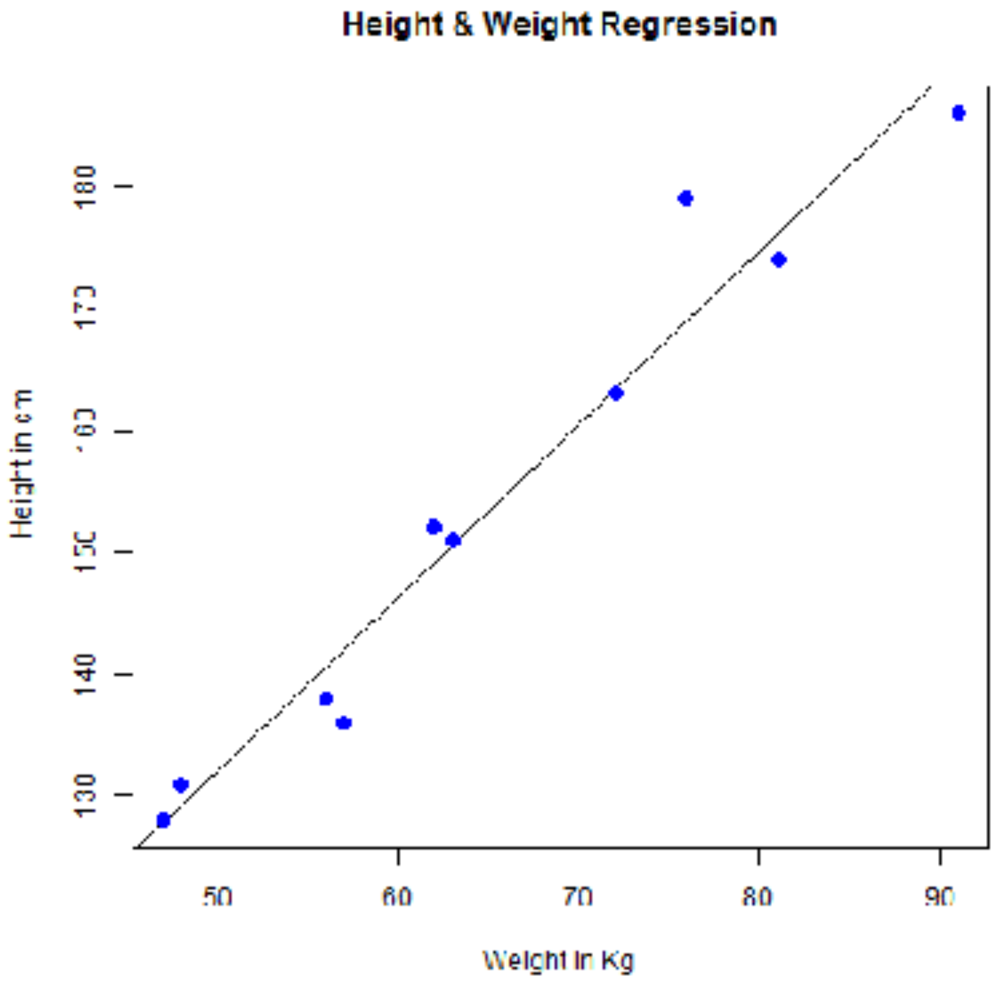
Not just this!!



Neural networks are a TYPE of machine learning, but there are many other types

You’re probably already doing machine learning!

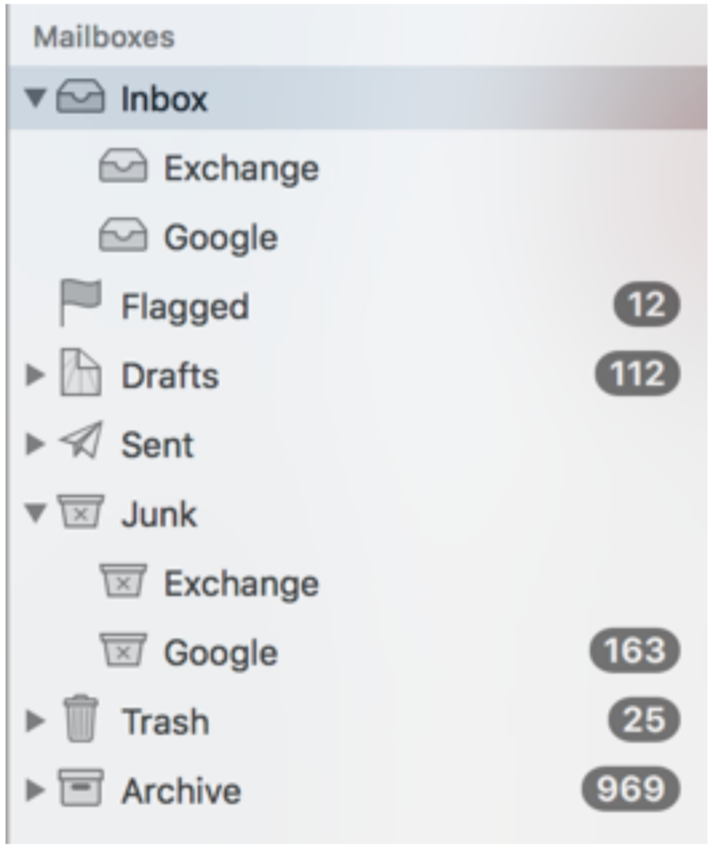
- Linear Regression
- Logistic Regression
- Decision Tree
- Support Vector Machines
- Naive Bayes
- k Nearest Neighbor
- K-Means
- Random Forest
- Dimensionality reduction
-



Machine learning algorithms are ubiquitous for commercial applications

NETFLIX

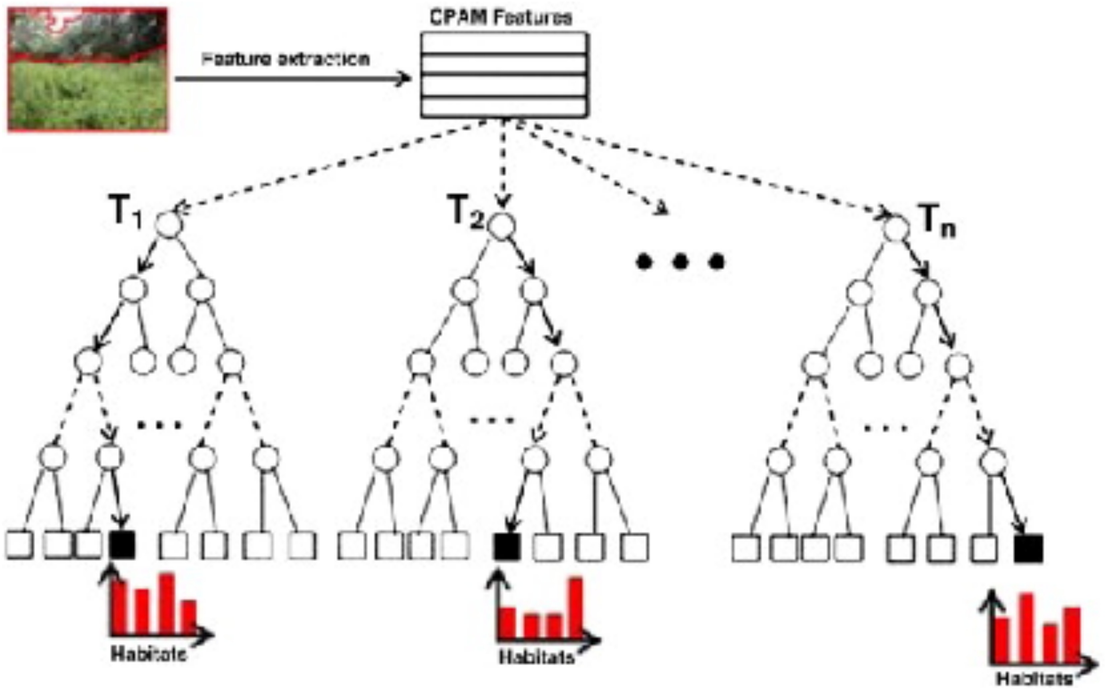
facebook

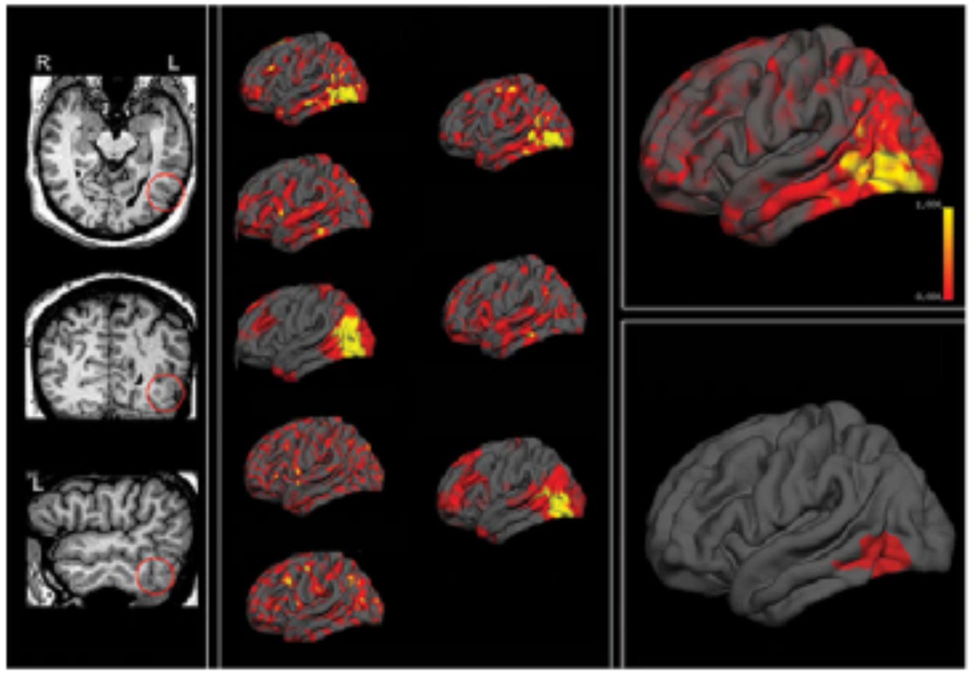


Mailboxes

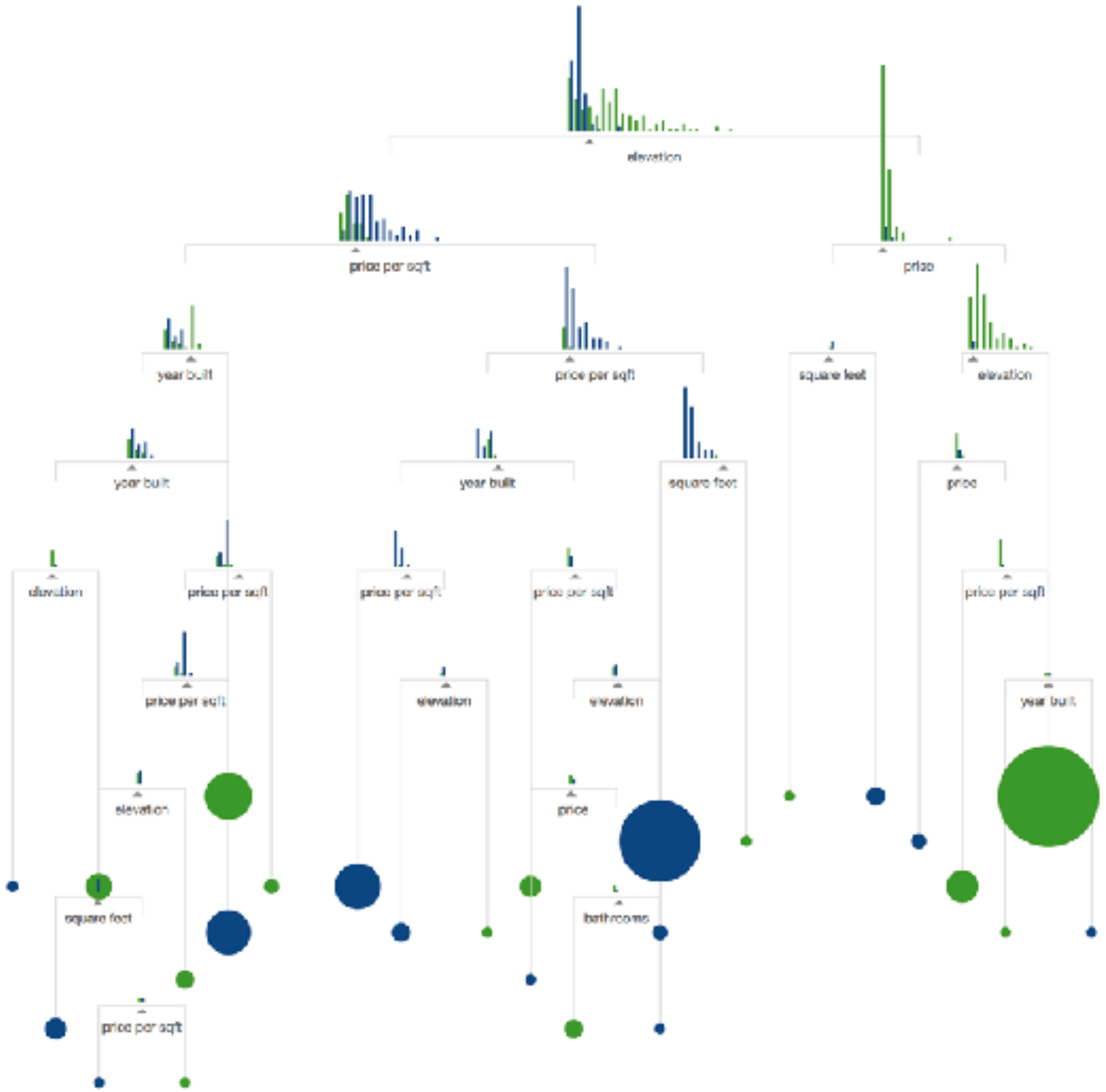
▼ Inbox
Exchange
Google
Flagged 12
▶ Drafts 112
▶ Sent
▼ Junk
Exchange
Google 163
▶ Trash 25
▶ Archive 969

Machine learning has lots of potential in biology





Classification and prediction are common machine learning problems



http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Supervised vs. unsupervised

Supervised machine learning requires labeled training data

Unsupervised algorithms look for patterns within the data without labeling categories

ACTIVITY: Classify the “organisms” provided based on the instructions you receive

Left side of the room: Procedure A

Classify candy into preferred and non-preferred

Use a subset for “training”

Test the model with another subset of the candy

Right side of the room: Procedure B

Identify the number of types of candy, based on features

Split into 2-3 categories

What are the key splits?

Discuss

1. Which approach is supervised learning, Procedure A or Procedure B? Which is unsupervised? Why?

2. How might sample size affect outcomes of your algorithm? What about the range of variation in your dataset?

3. What did you learn about underlying logic for assigning individual pieces of candy to one category or another? Is it easier to detect these mechanistic relationships with supervised or unsupervised learning?

Connections to biological problems

Would phylogeny be a supervised or unsupervised algorithm?

What about using DNA barcoding to identify thousands of individuals to species?

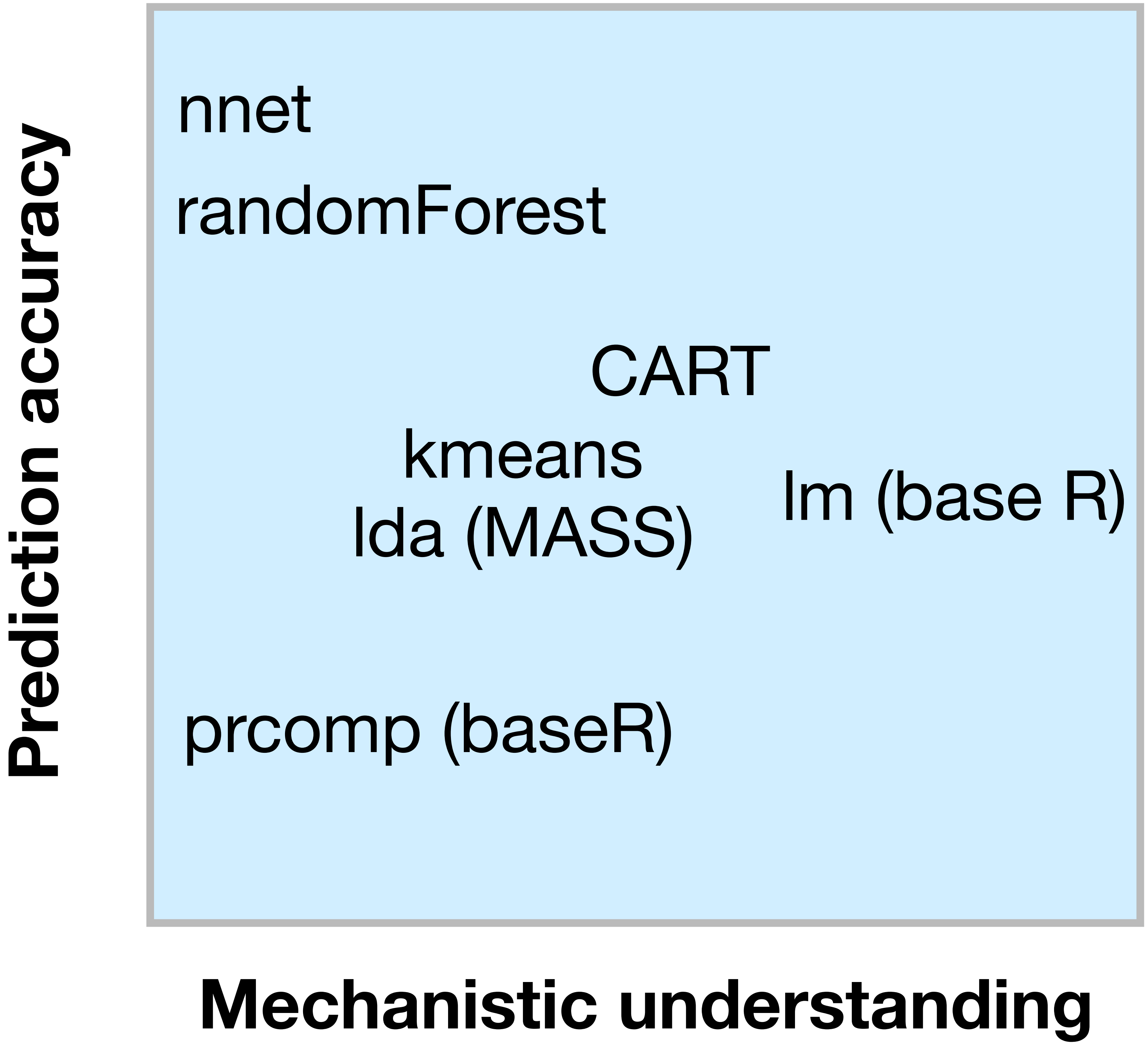
Applying machine learning to biology

Defining your question well is essential

Potential trade-off between prediction and mechanistic understanding

What are your goals, and what are your data like?

No shortage of easy-to-use ML tools in R



Prediction accuracy

nnet

randomForest

CART

kmeans

lda (MASS)

lm (base R)

prcomp (baseR)

Mechanistic understanding