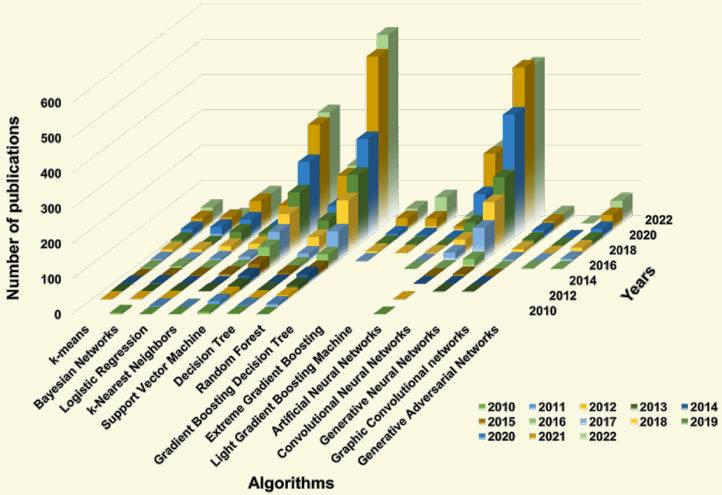
Machine Learning For Ecological Research

NAU

April 9th, 2024

Jeremy Forsythe

Machine Learning Publications in Ecology

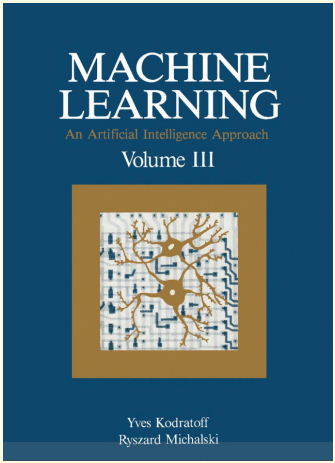


↑ Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.

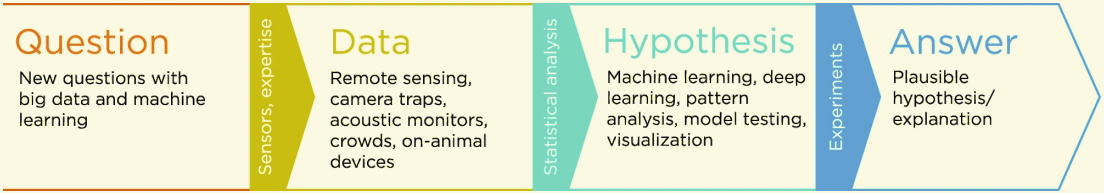
Machine Learning Definition

Widely Accepted Definition:

A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.



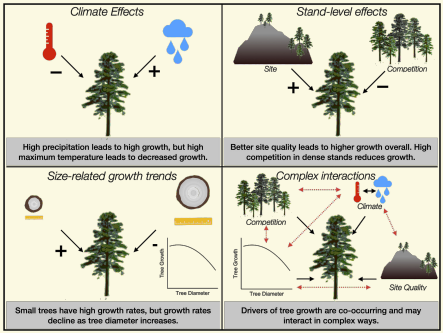
Scientific Method



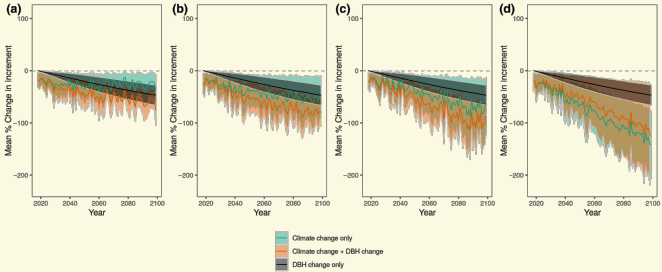
↑ Tuia et. al. (2022). Perspectives in machine learning for wildlife conservation. Nat Commun 13, 792.

Big Picture Goals?

Inference: Explanatory Power



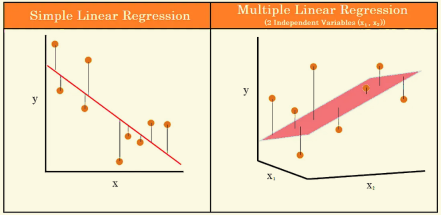
Prediction: Predictive Power



←↑ Heilman et. al. (2022). Ecological forecasting of tree growth: Regional fusion of tree-ring and forest inventory data to quantify drivers and characterize uncertainty. Global Change Biology, 28, 2442–2460.

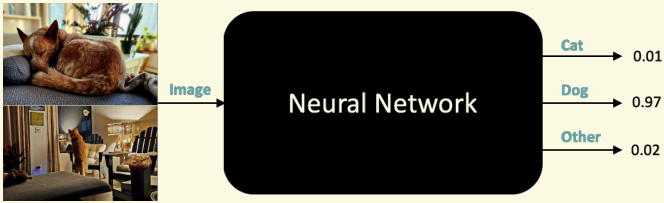
OLS vs. Machine Learning

OLS Regression



$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \varepsilon_i$$

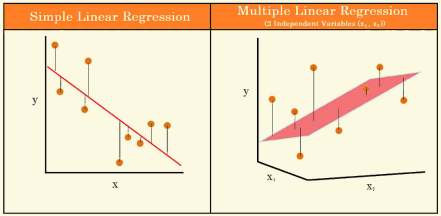
Machine Learning



$$y_i = ?x_{i\ldots}$$

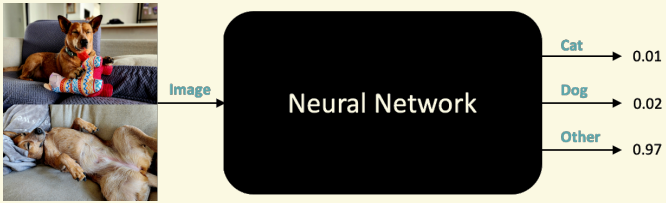
OLS vs. Machine Learning

OLS Regression



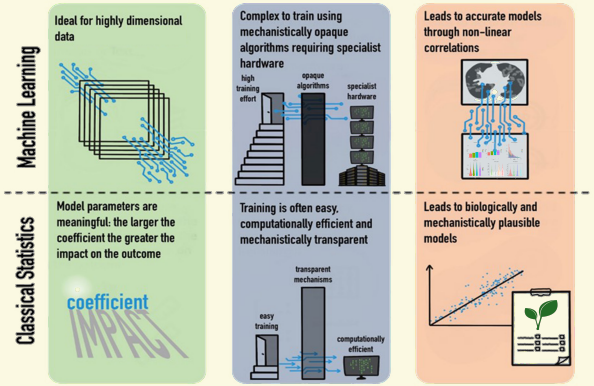
$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \varepsilon_i$$

Machine Learning

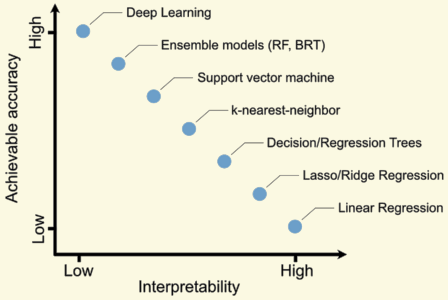


$$y_i = ?x_{i\ldots}$$

Tradeoffs?

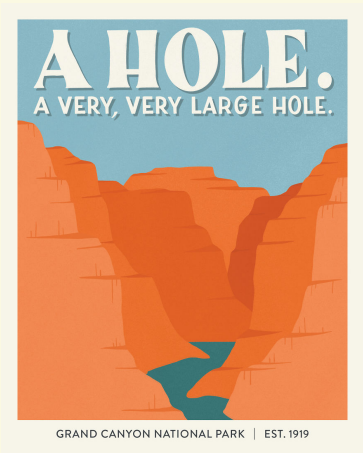


← Al-Hindawi et. al. 2021. A Pro-con Debate for Machine Learning vs. Traditional Statistics.

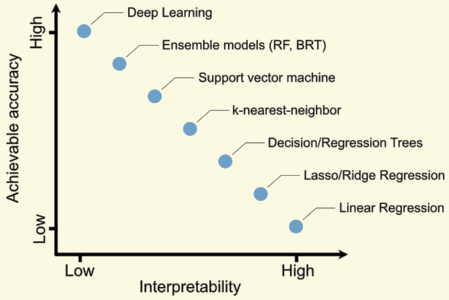


→ Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

Opinions?



← Poster made by Amber Share at Bored Panda from ACTUAL Yelp review of the Grand Canyon.



→ Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

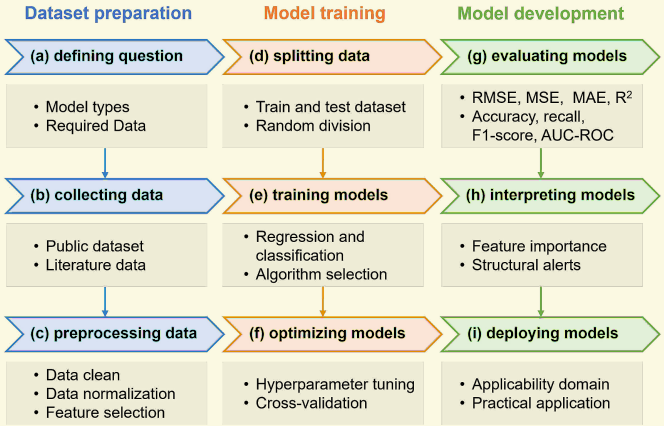
Back To Barney & Moe

“All models are wrong, but some are useful’’. - George Box

Widely Accepted Definition:

A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.

→ Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.



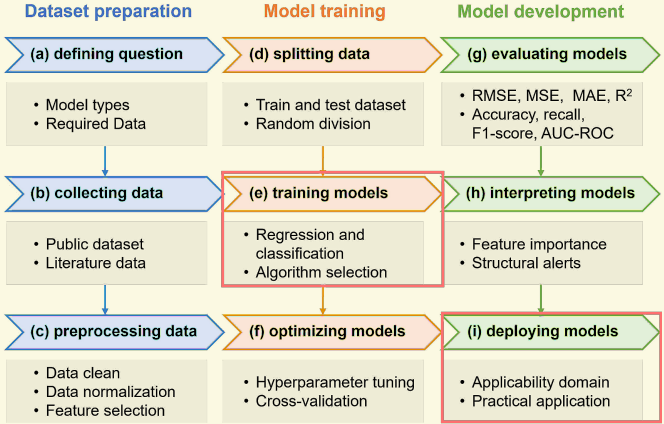
Task (T):

Classification: Which category does x_i belong to?

Regression: Given predictor(s), estimate a corresponding numerical target variable.

Anomaly Detection: Given a set of observations, flag the unusual ones.

→ Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.

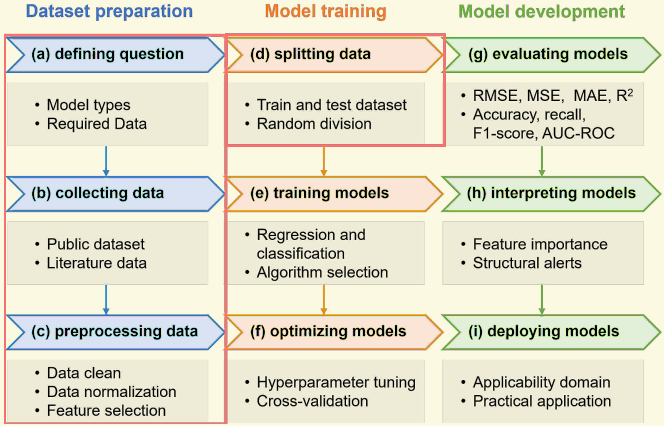


Experience (E):

Data: A Collection of examples, data points, observations.

Supervised or Unsupervised Learning

→ Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.



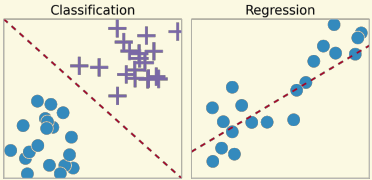
Supervised vs Unsupervised Learning

Overview

	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

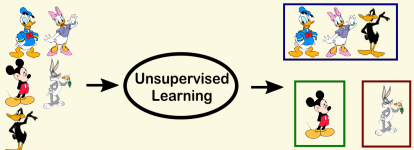
Supervised Learning

Learn a function that, given a sample of data and desired outputs, best approximates the relationship between input and output observable in the data.



Unsupervised Learning

Goal Is To Infer The Natural Structure Present Within A Set Of Data Points With Prior Expectations.

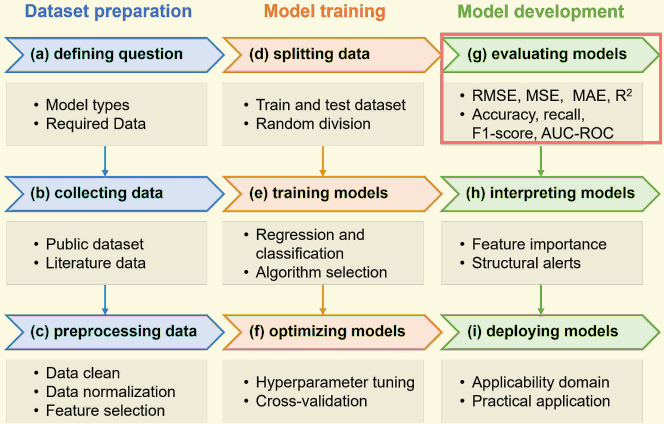


Performance Measure (P):

Evaluates the abilities of the machine learning system to perform the task (T).

For example, in regression you could use: **RMSE, R^2, RE**

→ Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.



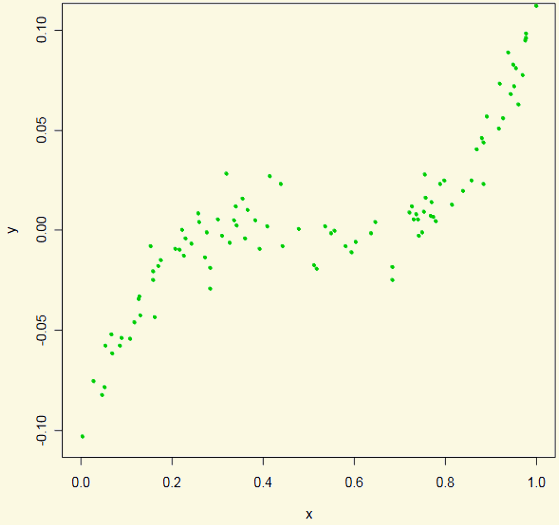
Given a set of *n* observations (instances or samples), **estimate** the relationship between *i* independent variables (predictors) and a dependent variable (target, response).

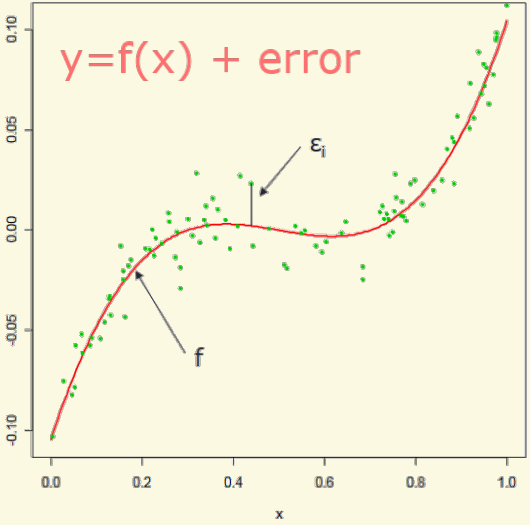
Given $x_i = (x_{i1}, x_{i2}, \ldots, x_in)$

and $y = y_1, y_2, \ldots, y_n$:

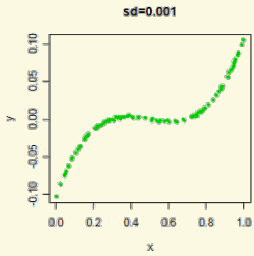
What is $y = f(x) + \varepsilon$?

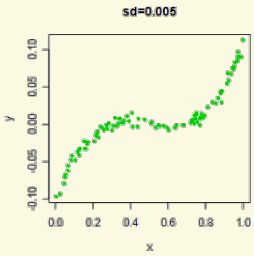
Machine Learning Basics : Example

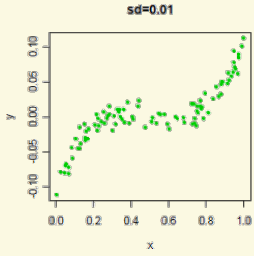


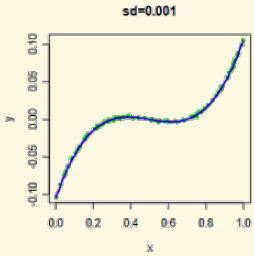


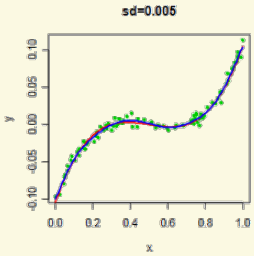
Machine Learning Basics : Noisy Signals

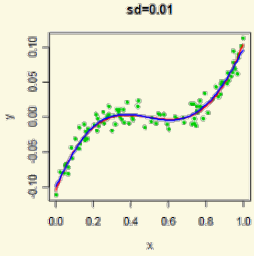


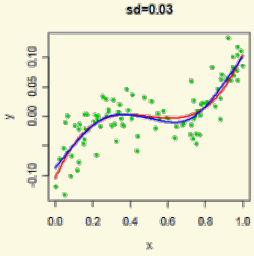












Machine Learning Basics : Recipe

1. Split Data Into **Training** and **Test** Datasets. (Cross Validation)
2. **Fit** Candidate Models on Training Dataset
3. **Assess** Performance of Candidate Models on Testing Dataset Using Same Set of Metrics
4. **Choose** Final Model Form
5. **Fit** Final Model Form
6. **Interpret** Model Output / **Predict** Future Responses
7. Walk Away Feeling **Empowered** $\rightarrow$

Tips & Tricks : Model Form

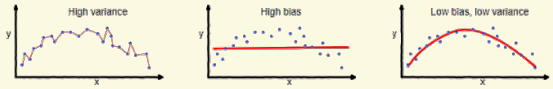
Parametric Models:

- Summarizes data with a fixed set of model parameters in a fixed model form.
- **Benefits:** Simple, Fast, Require Less Training Data
- **Limitations:** Constrained By Model Form, Model May Not Represent True Relationship
- **Examples:** Logistic Regression, Linear Discriminant Analysis

Nonparametric Models:

- Seek the best fit to the training data without a specified model form.
- **Benefits:** Flexible, Potentially Higher Performance
- **Limitations:** Requires More Data, Slower, Overfitting
- **Examples:** KNN, Support Vector Machines, Random Forests

Tips & Tricks : Bias & Overfitting



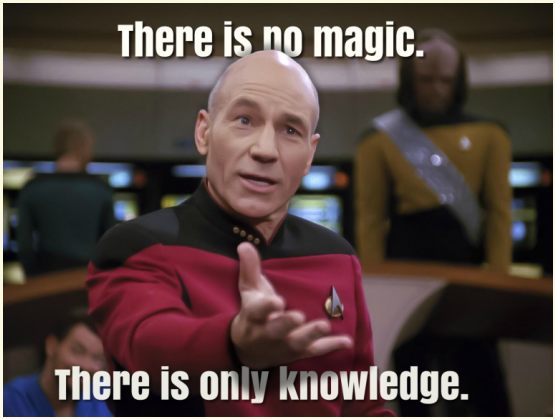
Bias-Variance Tradeoff

- Model is too simple, it will have very few parameters then it may have high bias and low variance.
- On the other hand if the model has large number of parameters then it's going to have high variance and low bias.
- Our job is to strike the correct balance.

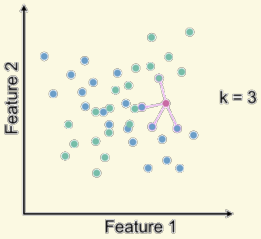
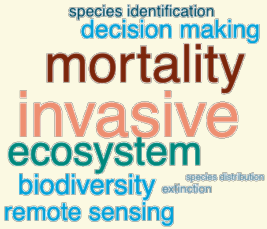
There Is No Magic...

There is no magic here, no one model is always better than another.

Fit the models, make an informed choice, enjoy your life!

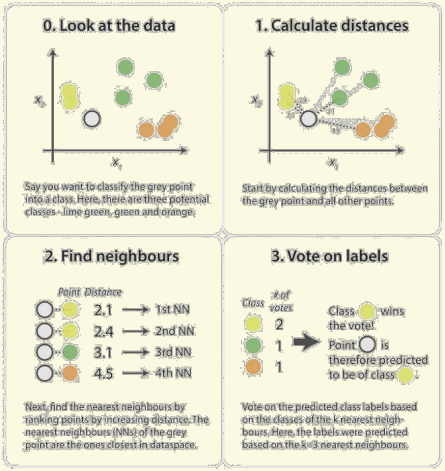


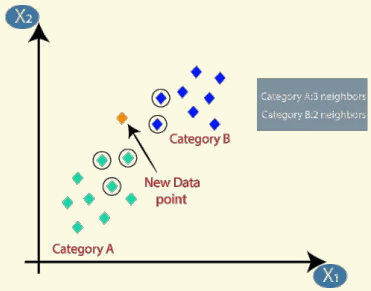
Machine Learning Model Overviews

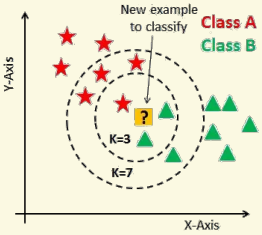
Machine learning model	Description	Data Type	Application areas
k-nearest-neighbor: 	K nearest neighbors in feature space decide response (e.g., by majority voting) + simple + no training - scales poorly - high dimensionality	Tabular Data: - Classification - Regression	

↑ Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

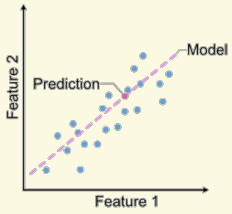
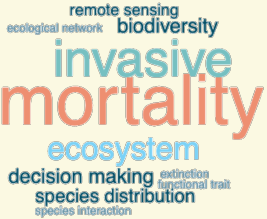
K Nearest Neighbor (KNN)







Machine Learning Model Overviews

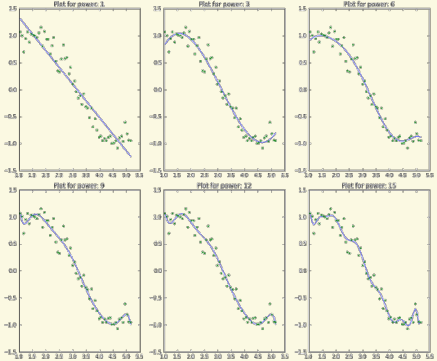
Machine learning model	Description	Data Type	Application areas
Lasso, Ridge Regression:  	Regression models with regularized coefficients + highly interpretable + few observations - Not very flexible	Tabular data: - Classification - Regression	

↑ Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

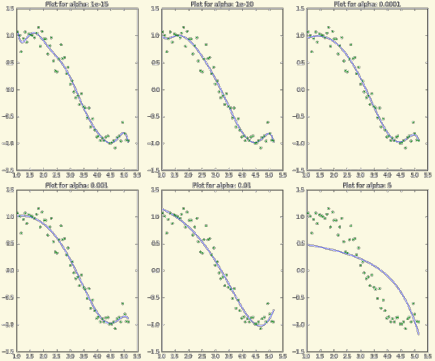
Lasso & Ridge Regression

Polynomial Regression (No Penalty)

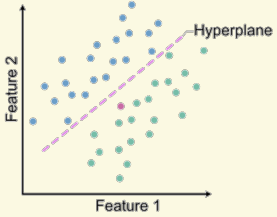
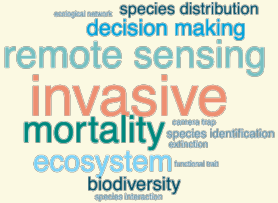
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \varepsilon$$



Ridge Regression (Overfitting Penalty)

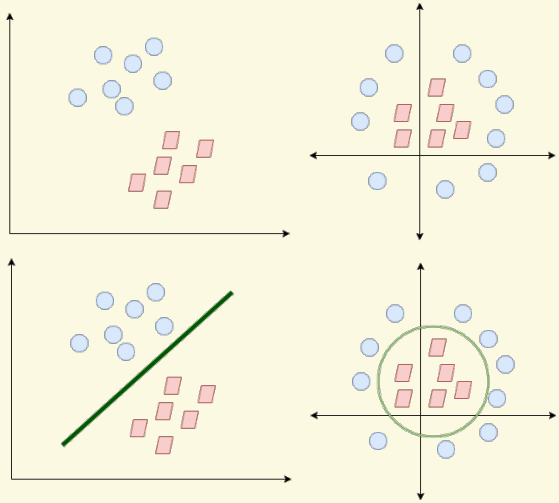


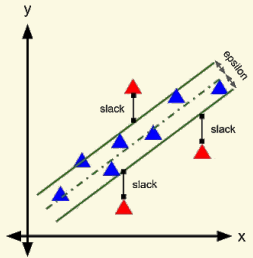
Machine Learning Model Overviews

Machine learning model	Description	Data Type	Application areas
Support vector machines: 	Hyperplane is optimized to separate response classes + fast and memory efficient + high dimensional data - kernel dependent - no probabilities	Tabular data: - Classification - Regression	

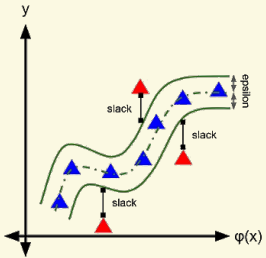
↑ Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

Support Vector Machines



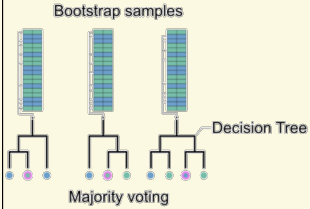
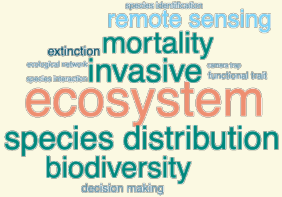


Linear



Non-linear

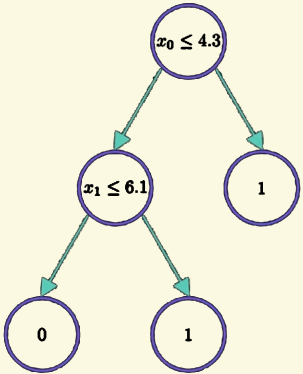
Machine Learning Model Overviews

Machine learning model	Description	Data Type	Application areas
Random Forest: 	N decision (regression) trees are fitted on bootstrap samples. Split variable is selected from random subset of variables + flexible + robust (e.g. outliers) + few hyper-parameters (+) variable importance - scales poorly	Tabular Data: - Classification - Regression	

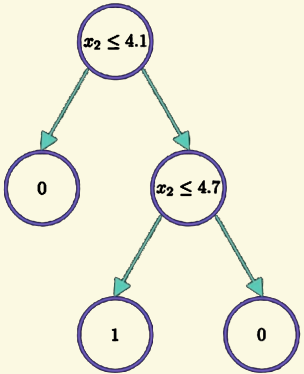
↑ Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

Decision Trees

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1



id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	6.5	4.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1



Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

id
2
1
3
1
4
4

id
4
1
3
0
0
2

id
3
3
2
5
1
2

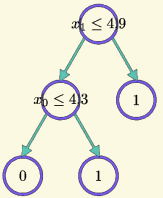
New Data:

2.8	6.2	4.3	5.3	5.5

Bootstrap + Aggregating

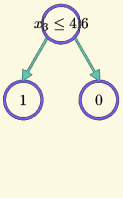
(Bagging)

x_0, x_1



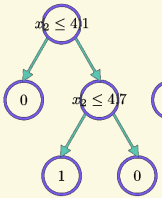
1

x_2, x_3



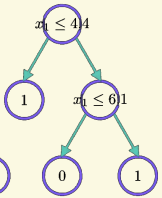
0

x_2, x_4



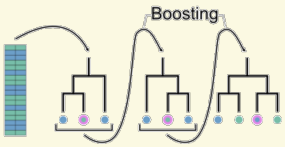
1

x_1, x_3



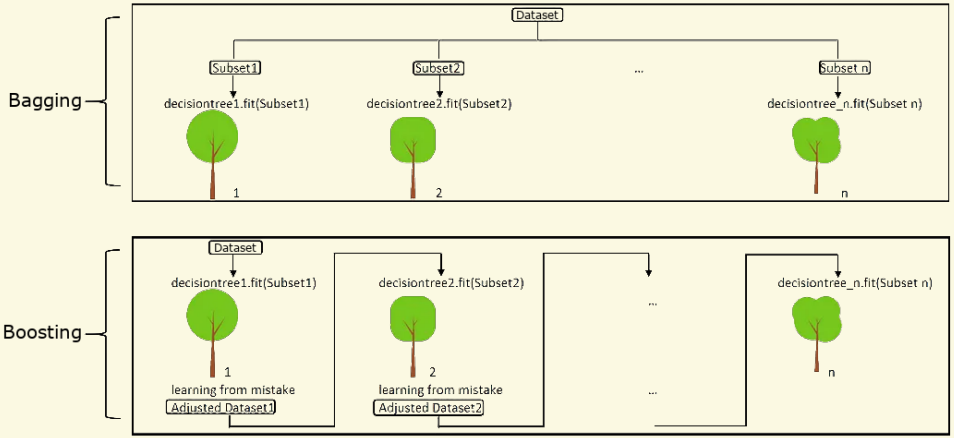
1

Machine Learning Model Overviews

Machine learning model	Description	Data Type	Application areas
Boosted Regression Trees: 	N trees are fitted sequentially to minimize an overall loss function + flexible (+) variable importance - many hyper-parameters - high complexity	Tabular Data: - Classification - Regression	

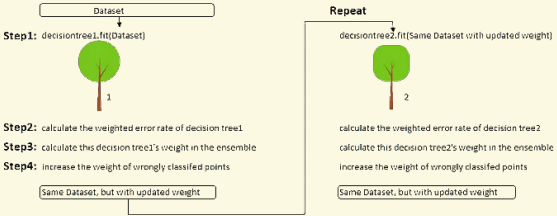
↑ Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

Boosting

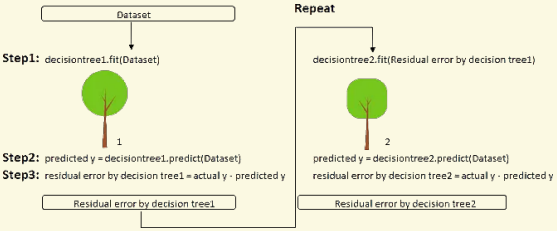

Bagging
Dataset
Subset1
Subset2
...
Subset n
decisiontree1.fit(Subset1)
decisiontree2.fit(Subset2)
decisiontree_n.fit(Subset n)
1
2
n
Boosting
Dataset
decisiontree1.fit(Subset1)
decisiontree2.fit(Subset2)
...
decisiontree_n.fit(Subset n)
1
2
n
learning from mistake
Adjusted Dataset1
learning from mistake
Adjusted Dataset2
...

Adaptive vs. Gradient Boosting

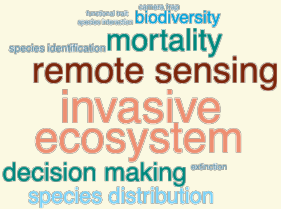
Adaptive Boosting:



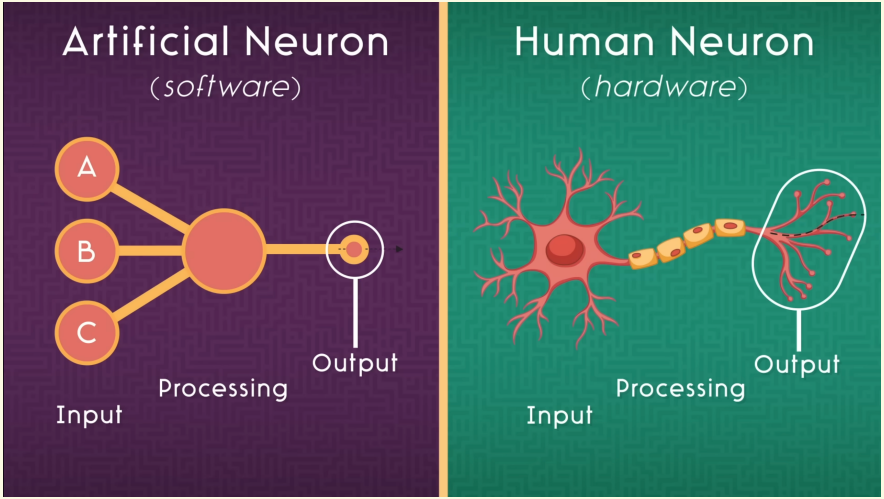
Gradient Boosting:



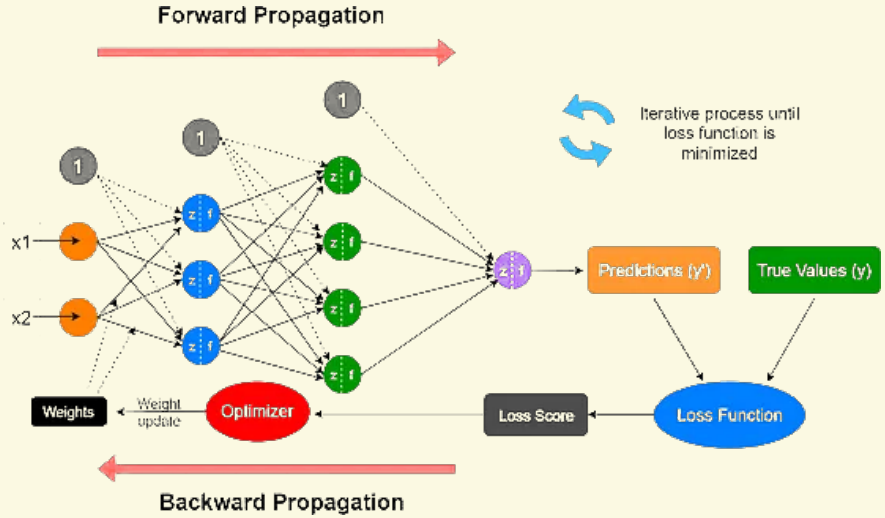
Machine Learning Model Overviews

Machine learning model	Description	Data Type	Application areas
Deep Neural Networks:  	Input (features) are passed through many hidden layers. Last layer maps into response space + flexible + adaptive to different tasks - many hyper-parameters - computationally expensive	Tabular Data: - Classification - Regression	

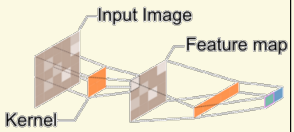
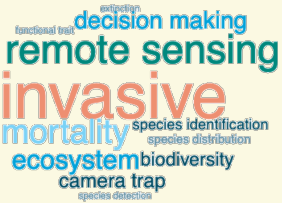
↑ Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.


Artificial Neuron
(software)
A
B
C
Input
Processing
Output
Human Neuron
(hardware)
Input
Processing
Output

Neural Networks

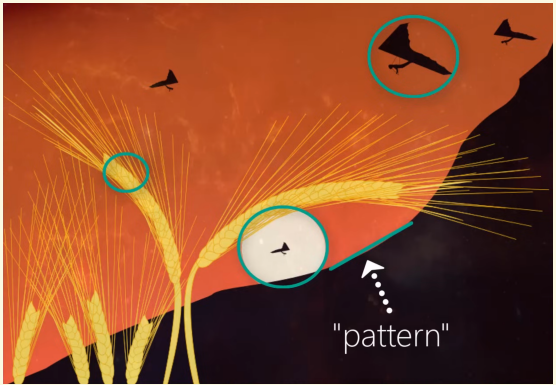


Machine Learning Model Overviews

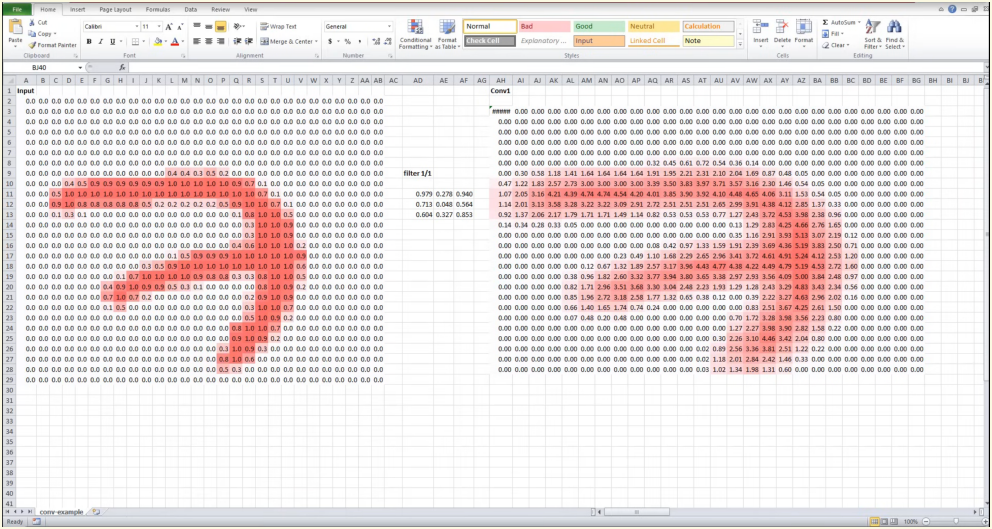
Machine learning model	Description	Data Type	Application areas
Convolutional Neural  Networks:	Small kernels (filters) processes images before passing it to fully connected layers + flexible + detecting shapes and edges - many hyper-parameters - computationally expensive	Images: - Classification - Object detection	

↑ Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

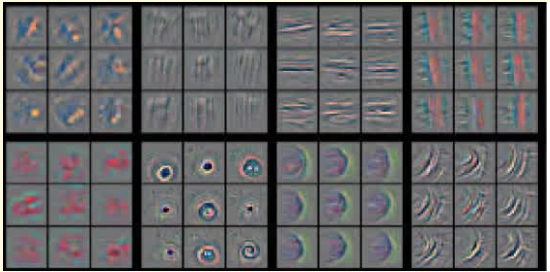
Convolutional Neural Networks

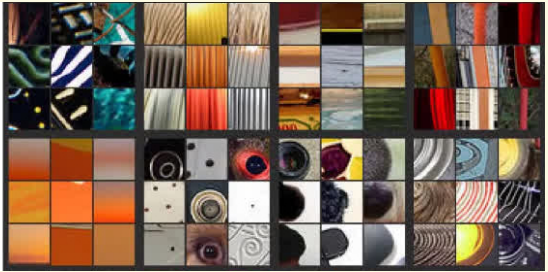


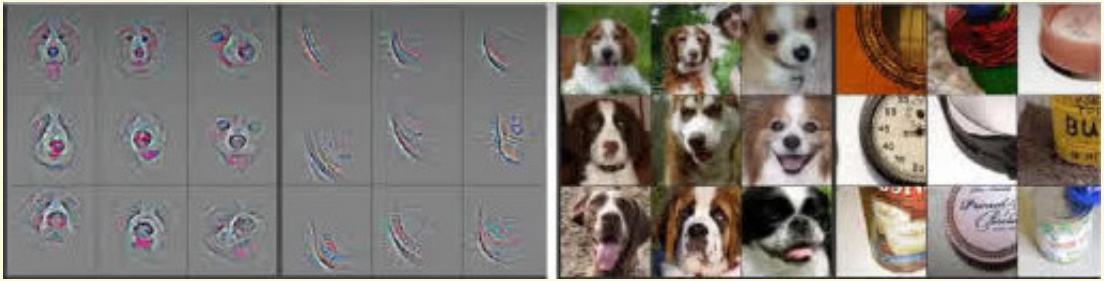
Convolutional Neural Networks




filter 1
-1 -1 -1
1 1 1
0 0 0
filter 2
-1 1 0
-1 1 0
-1 1 0
filter 3
0 0 0
1 1 1
-1 -1 -1
filter 4
0 1 -1
0 1 -1
0 1 -1
"bright" corresponds to white in this example







Classification
And
Regression
Training