

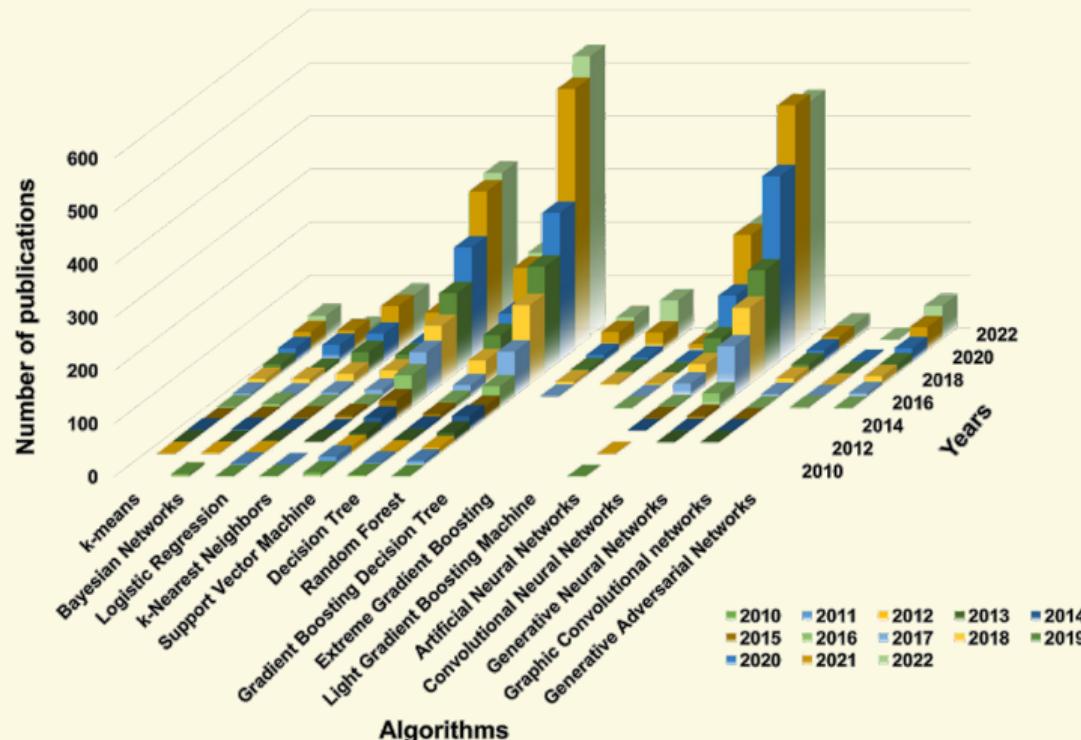
ECOML WORKSHOP:

MACHINE LEARNING FOR ECOLOGICAL RESEARCH

JEREMY FORSYTHE,
JAZLYNN HALL,
JATAN BUCH



MACHINE LEARNING PUBLICATIONS IN ECOLOGY

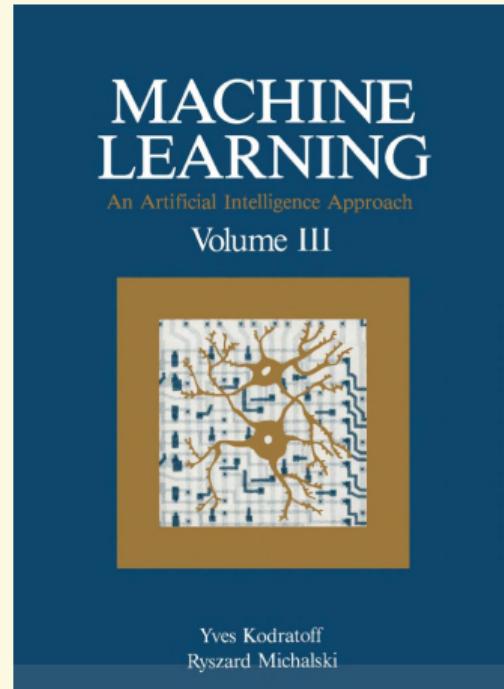


Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.

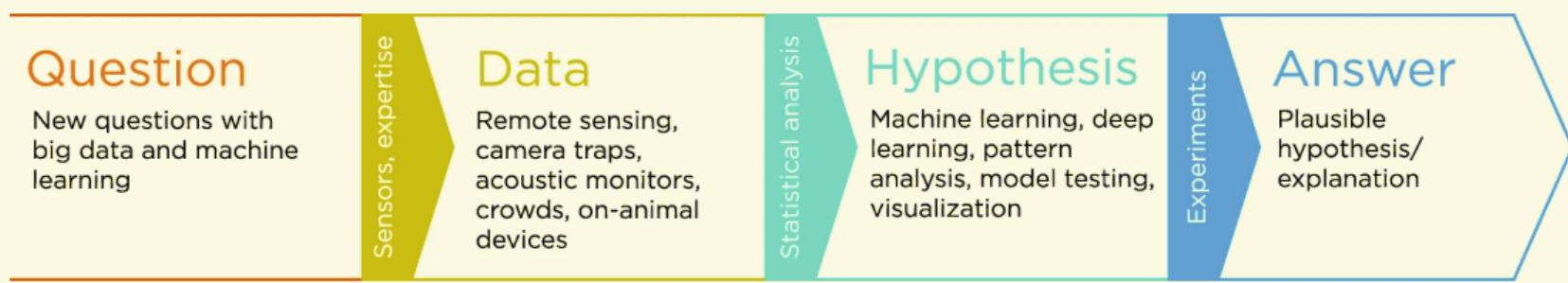
MACHINE LEARNING DEFINITION

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 .



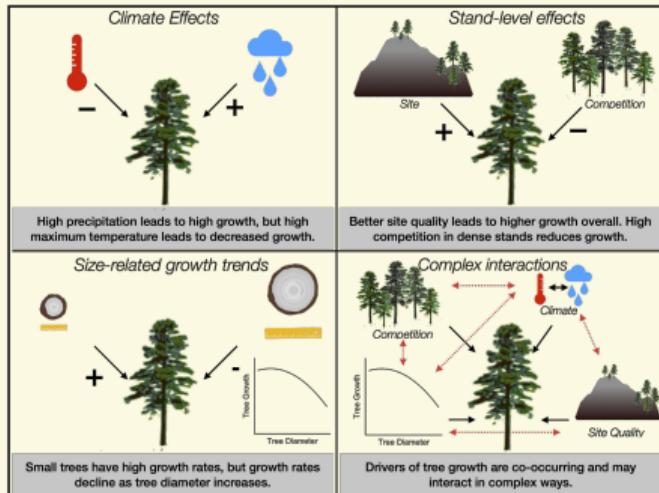
SCIENTIFIC METHOD



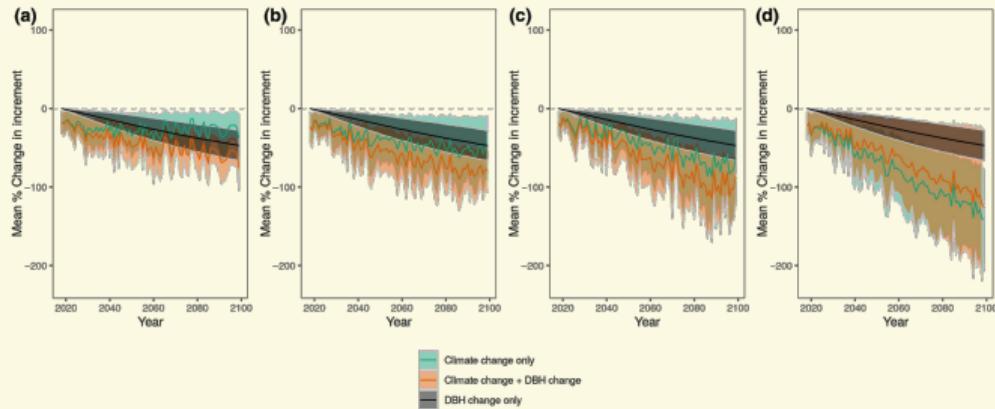
Tuia et. al. (2022). Perspectives in machine learning for wildlife conservation. Nat Commun 13, 792.

BIG PICTURE GOALS?

Inference: Explanatory Power



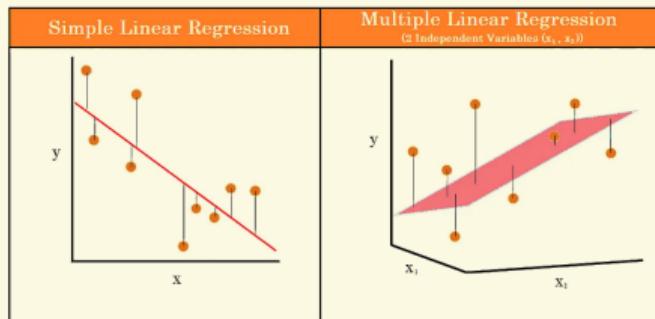
Prediction: Predictive Power



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, 24422460.

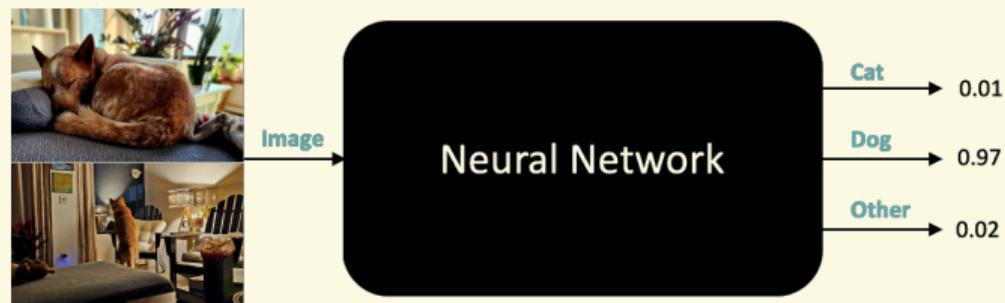
OLS vs. MACHINE LEARNING

OLS Regression



$$y_i = _1x_{i1} + _2x_{i2} + \dots + _ix_{ii}$$

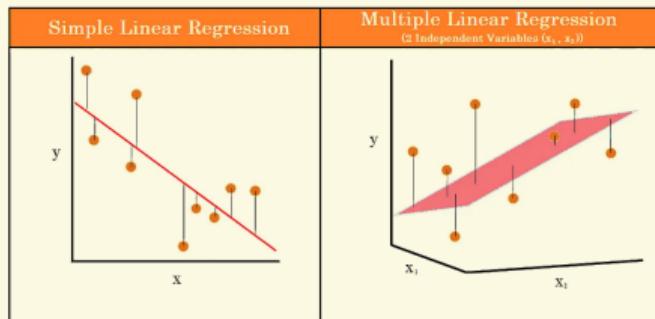
Machine Learning



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

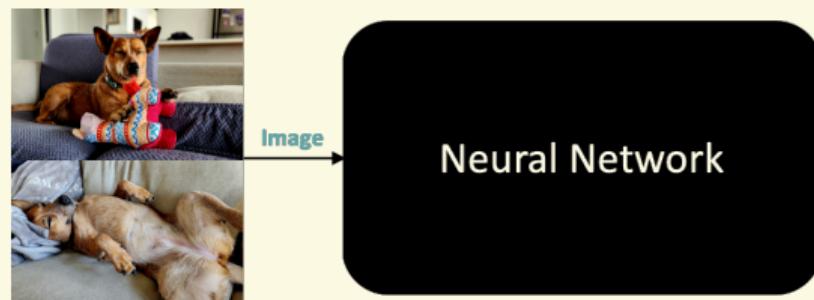
OLS vs. MACHINE LEARNING

OLS Regression



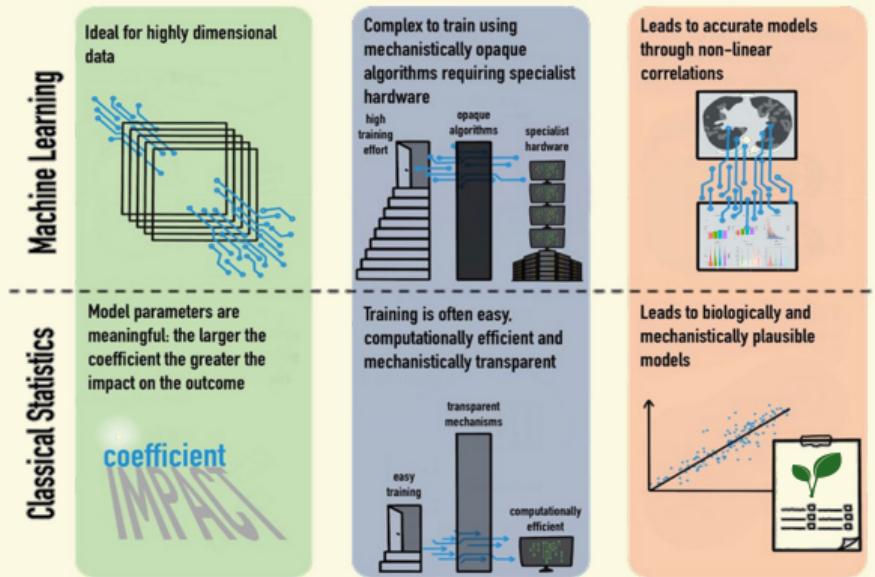
$$y_i = _1x_{i1} + _2x_{i2} + \dots + _i$$

Machine Learning

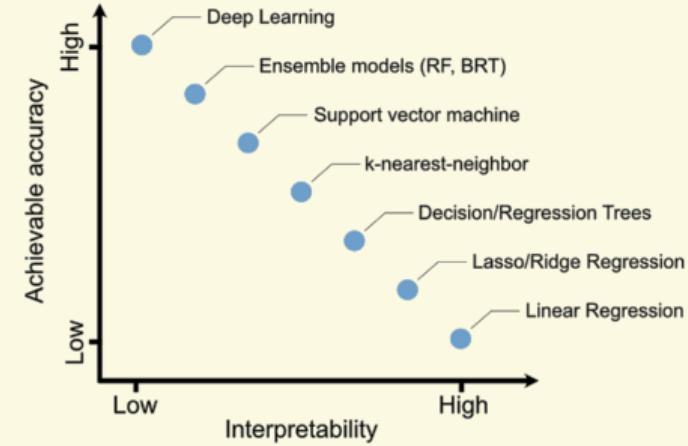


$$y_i = ?x_{i\dots}$$

TRADEOFFS?

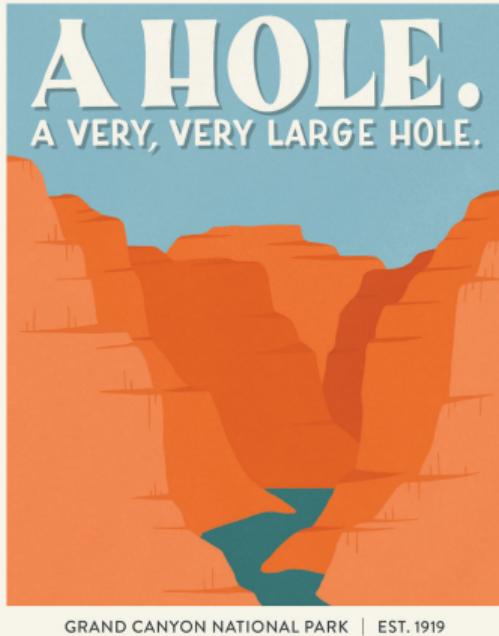


Al-Hindawi et. al. 2021. A Pro-con Debate for Machine Learning vs. Traditional Statistics.

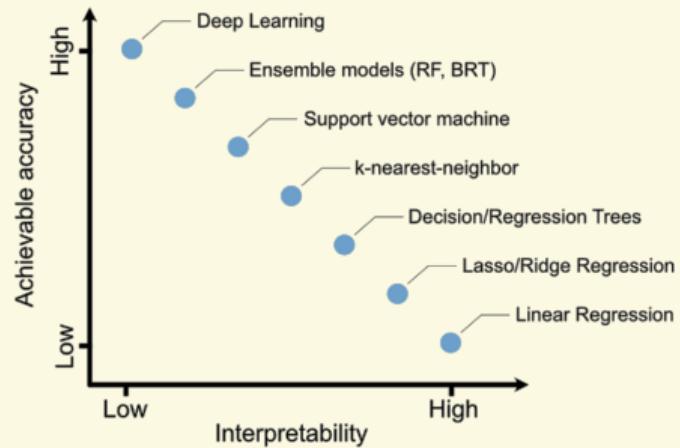


Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

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



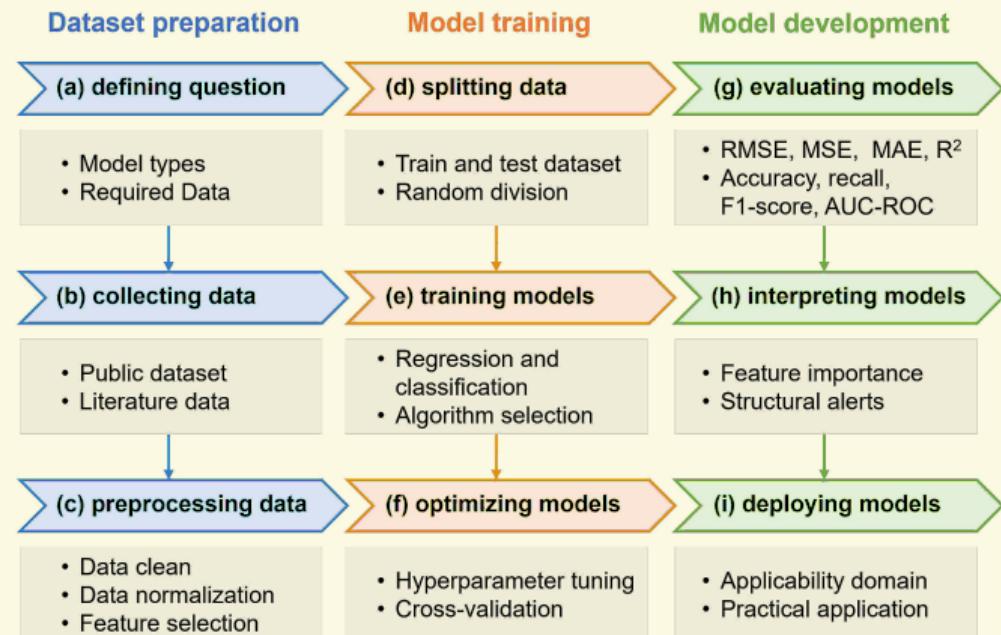
“All models are wrong, but some are useful”. - George Box

MACHINE LEARNING BASICS

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.



MACHINE LEARNING BASICS : TASK

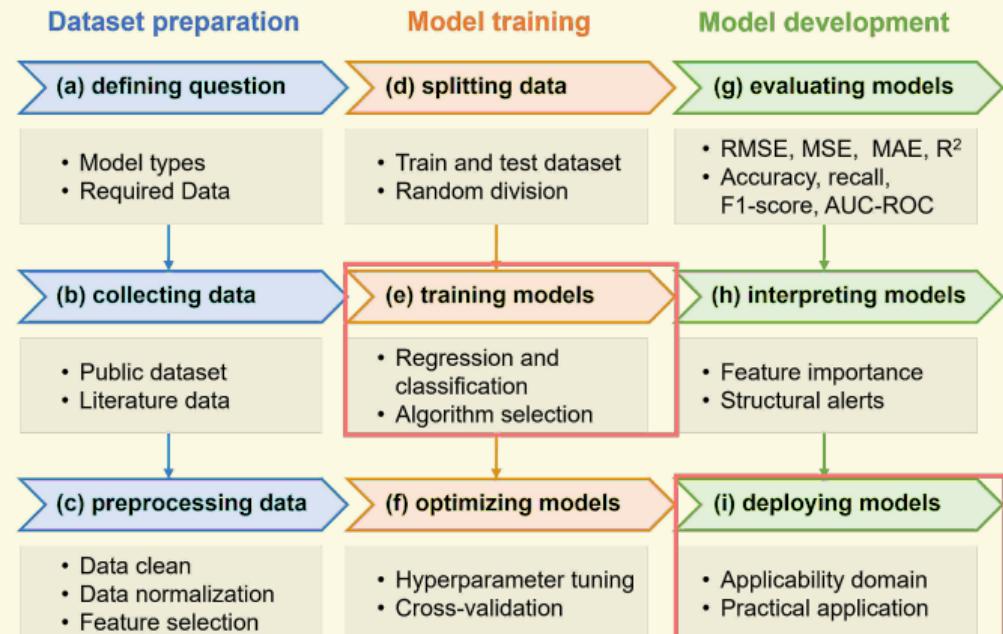
Task (T):

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.



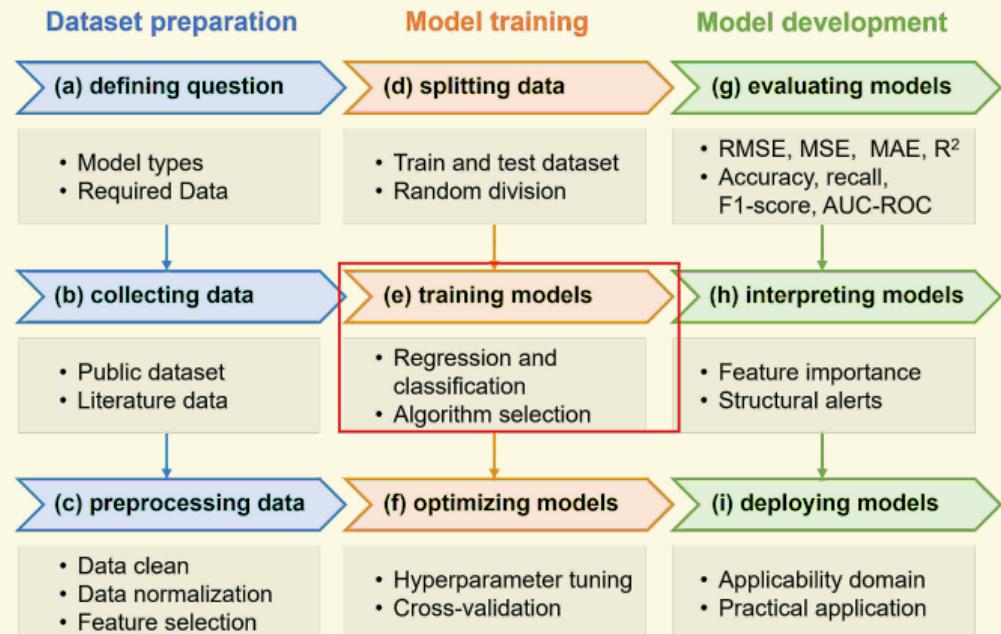
MACHINE LEARNING BASICS : EXPERIENCE

Experience (E):

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

Supervised or Unsupervised Learning

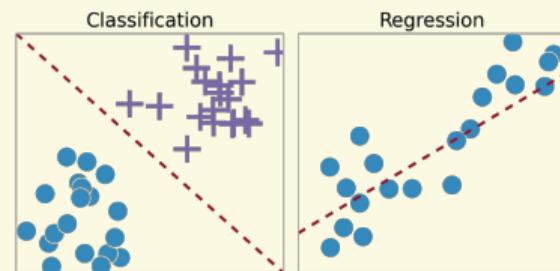
Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.



SUPERVISED VS UNSUPERVISED LEARNING

Overview

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

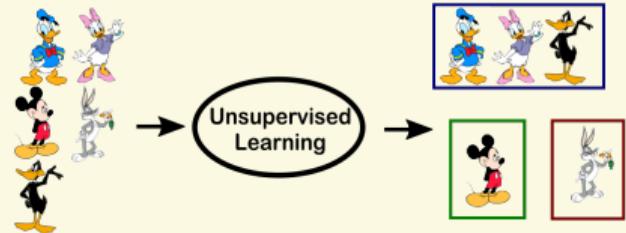


Supervised Learning

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.



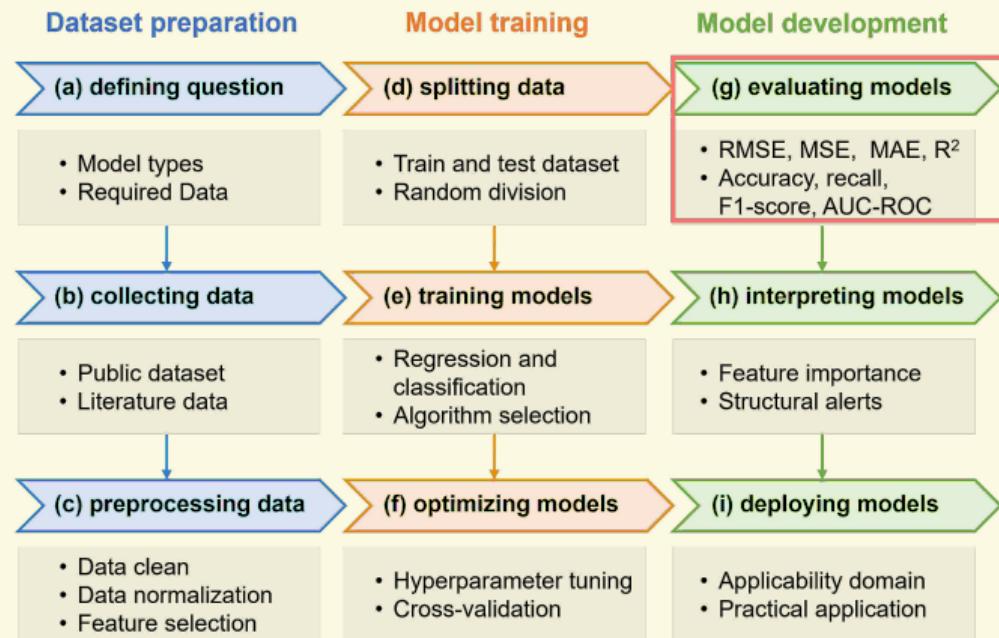
MACHINE LEARNING BASICS : PERFORMANCE MEASUREMENTS

Performance Measure (P):

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

For example, in regression you could use: **RMSE, R², RE**

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



MACHINE LEARNING BASICS : TERMS

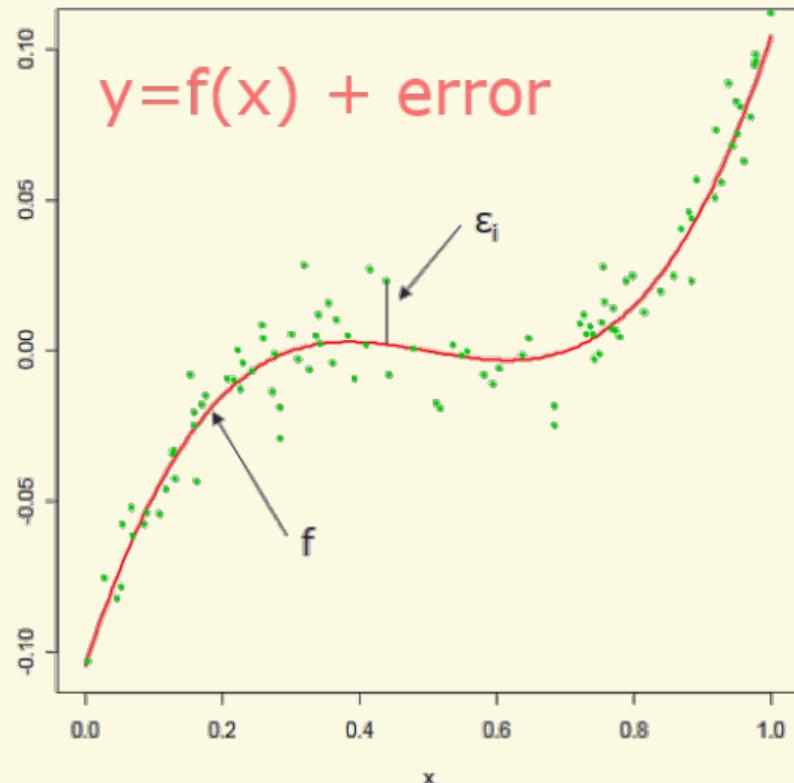
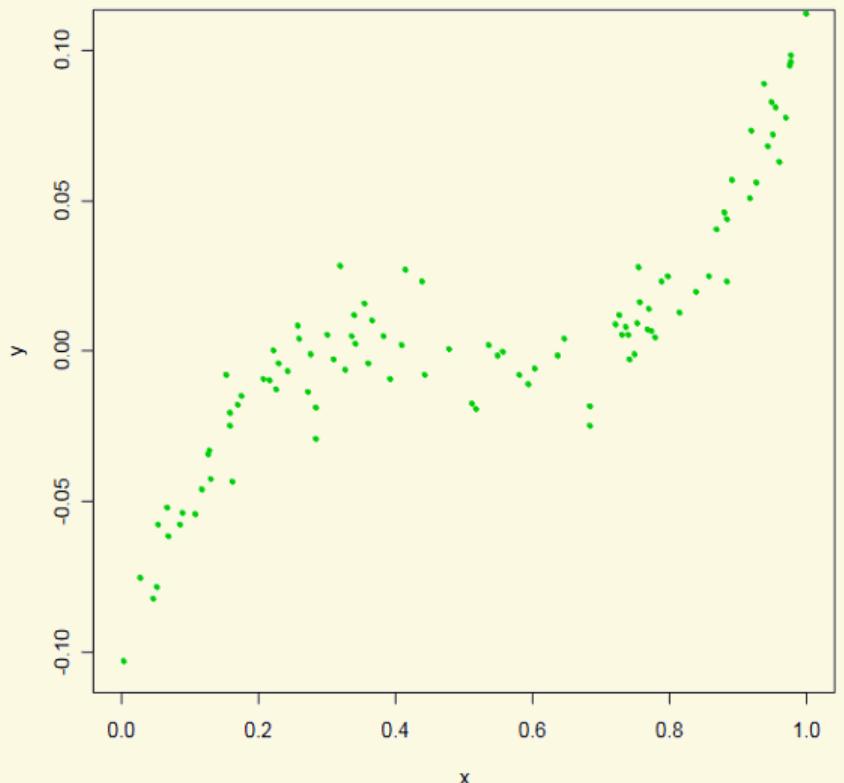
Given a set of n observations (instances or samples), **estimate** the relationship between i independent variables (predictors) and a dependent variable (target, response).

Given $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$

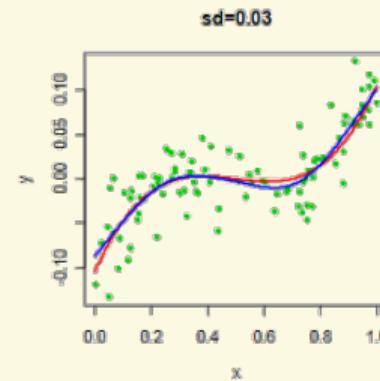
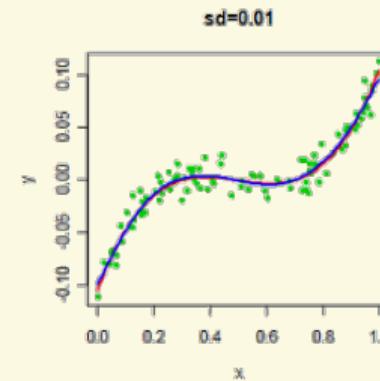
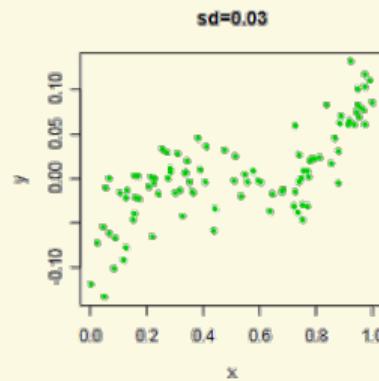
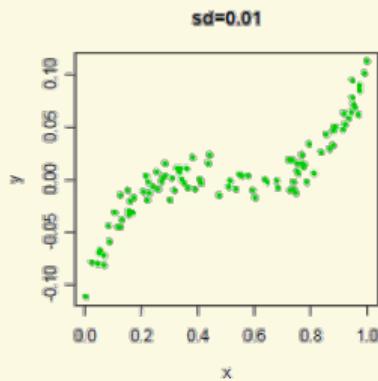
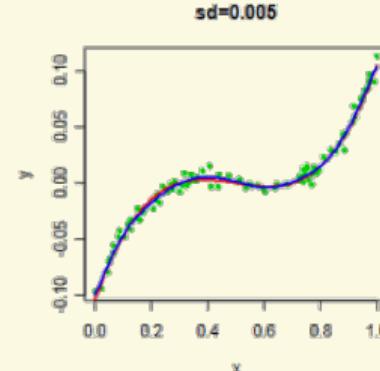
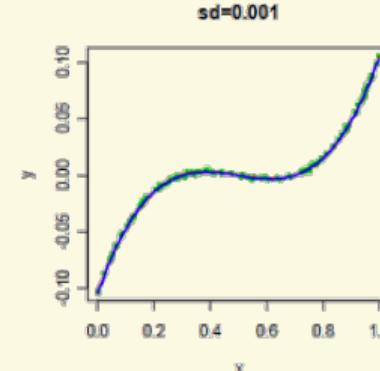
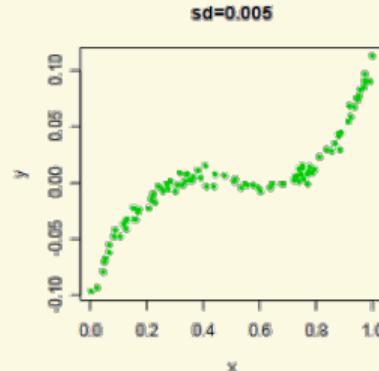
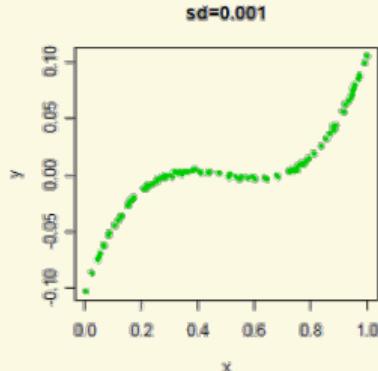
and $y = y_1, y_2, \dots, y_n$:

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

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**



TIPS & TRICKS : MODEL FORM

Parametric Models:

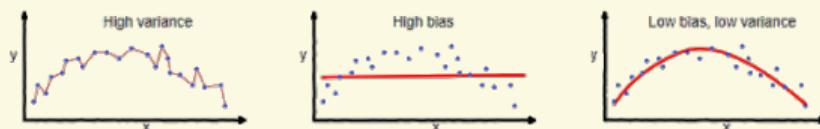
- Summarizes data with a fixed set of model parameters in a fixed model form.
- **Benefits:** Simple, Fast, Require Less Training Data
- **Limitations:** Constrained By Model Form, Model May Not Represent True Relationship
- **Examples:** Logistic Regression, Linear Discriminant Analysis

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

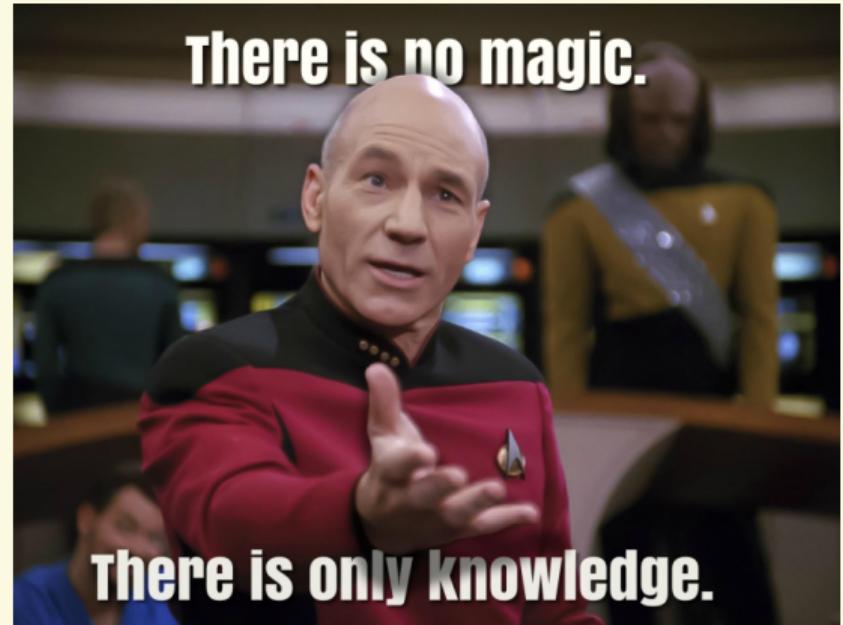


- Model is too simple, it will have very few parameters then it may have high bias and low variance.
- On the other hand if the model has large number of parameters then its going to have high variance and low bias.
- Our job is to strike the correct balance.

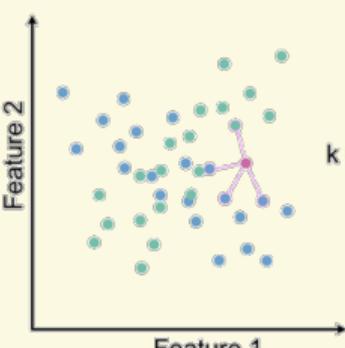
THERE IS NO MAGIC...

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

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

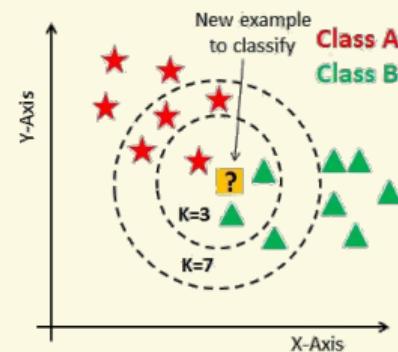
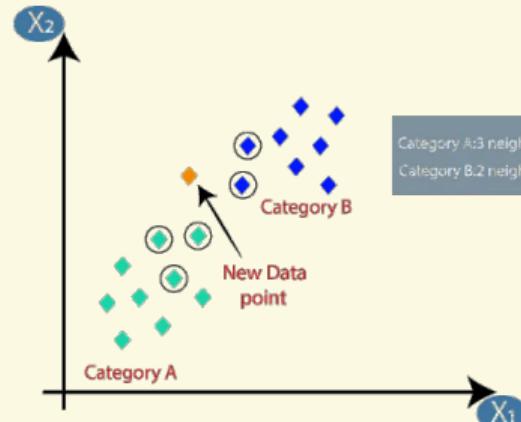
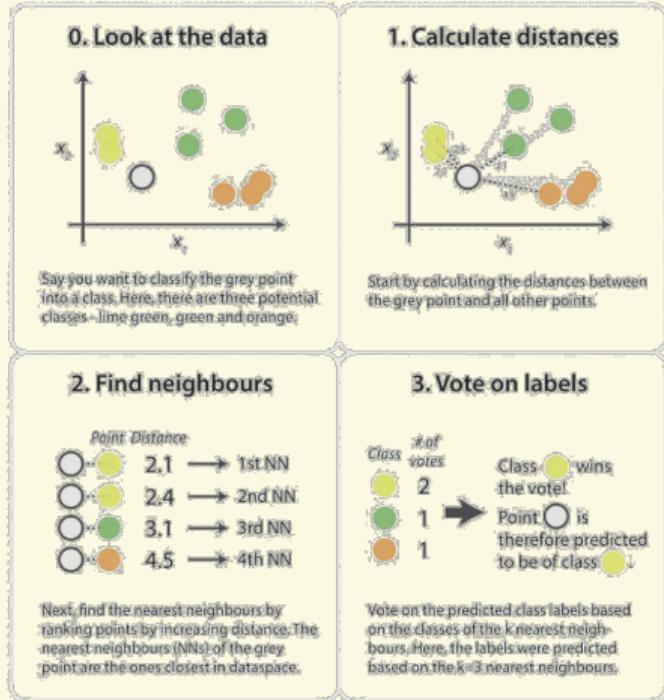


MACHINE LEARNING MODEL OVERVIEWS

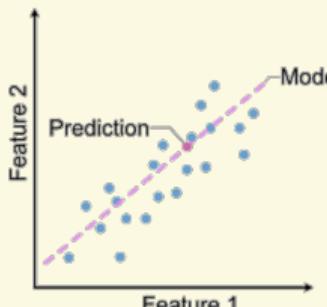
Machine learning model	Description	Data Type	Application areas
<p>k-nearest-neighbor:</p> 	<p>K nearest neighbors in feature space decide response (e.g., by majority voting)</p> <ul style="list-style-type: none">+ simple+ no training- scales poorly- high dimensionality	<p>Tabular Data:</p> <ul style="list-style-type: none">- Classification- Regression	<p>species identification decision making mortality invasive ecosystem biodiversity remote sensing species distribution extinction</p>

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

K NEAREST NEIGHBOR (KNN)



MACHINE LEARNING MODEL OVERVIEWS

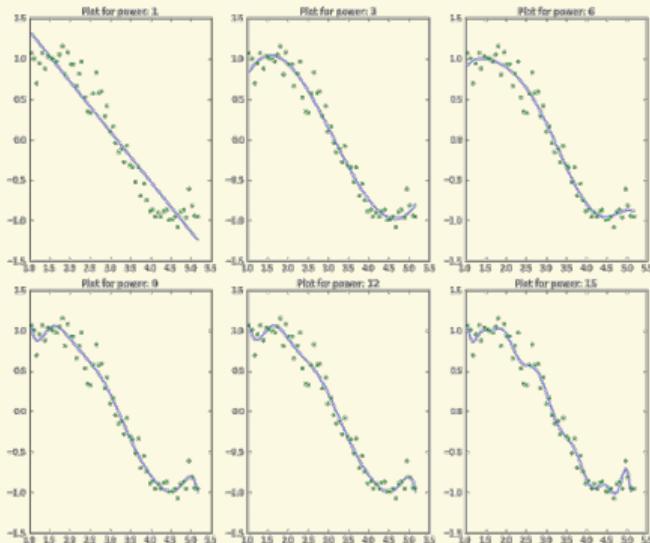
Machine learning model	Description	Data Type	Application areas
Lasso, Ridge Regression: 	<p>Regression models with regularized coefficients</p> <ul style="list-style-type: none">+ highly interpretable+ few observations- Not very flexible	<p>Tabular data:</p> <ul style="list-style-type: none">- Classification- Regression	<p>remote sensing ecological network biodiversity</p> <p>invasive mortality ecosystem</p> <p>decision making extinction species distribution functional trait species interaction</p>

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

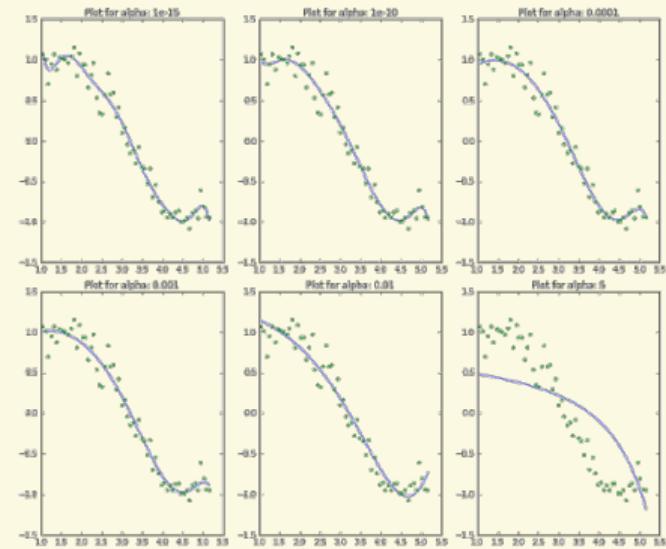
LASSO & RIDGE REGRESSION

Polynomial Regression (No Penalty)

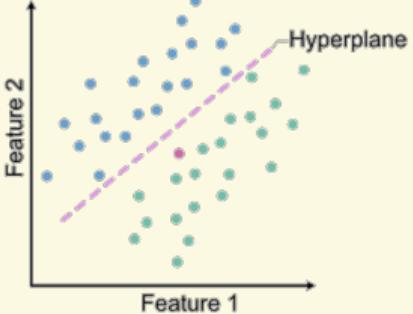
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_n x^n + \epsilon$$



Ridge Regression (Overfitting Penalty)

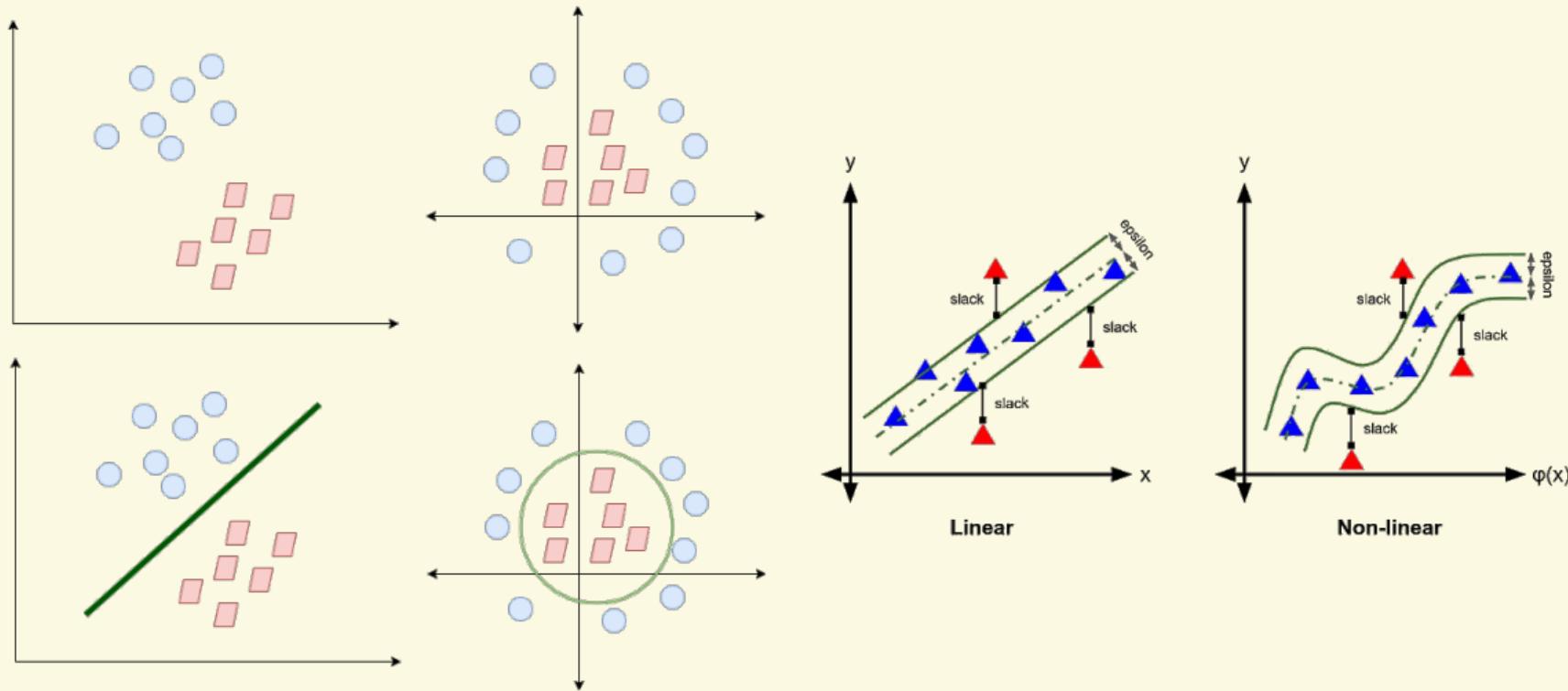


MACHINE LEARNING MODEL OVERVIEWS

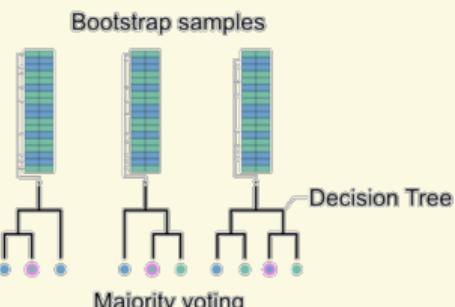
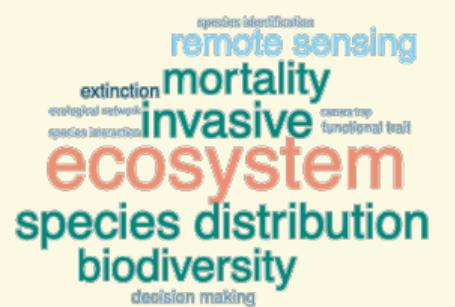
Machine learning model	Description	Data Type	Application areas
Support vector machines: 	<p>Hyperplane is optimized to separate response classes</p> <ul style="list-style-type: none">+ fast and memory efficient+ high dimensional data- kernel dependent- no probabilities	<p>Tabular data:</p> <ul style="list-style-type: none">- Classification- Regression	<p>ecological network species distribution decision making remote sensing invasive mortality ecosystem biodiversity</p> <p>camera trap species identification estimation functional trait species interactions</p>

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

SUPPORT VECTOR MACHINES

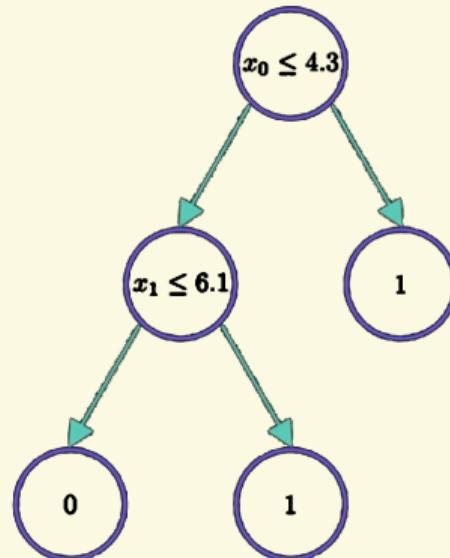


MACHINE LEARNING MODEL OVERVIEWS

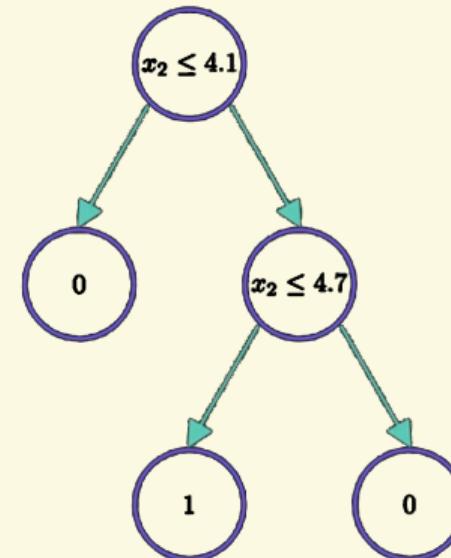
Machine learning model	Description	Data Type	Application areas
<p>Random Forest:</p>  <p>Bootstrap samples</p> <p>Decision Tree</p> <p>Majority voting</p>	<p>N decision (regression) trees are fitted on bootstrap samples. Split variable is selected from random subset of variables</p> <ul style="list-style-type: none">+ flexible+ robust (e.g. outliers)+ few hyper-parameters(+) variable importance- scales poorly	<p>Tabular Data:</p> <ul style="list-style-type: none">- Classification- Regression	

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

<i>id</i>	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

<i>id</i>
2
0
2
4
5
5

<i>id</i>
2
1
3
1
4
4

<i>id</i>
4
1
3
0
0
2

<i>id</i>
3
3
2
5
1
2

New Data:

2.8	6.2	4.3	5.3	5.5
-----	-----	-----	-----	-----

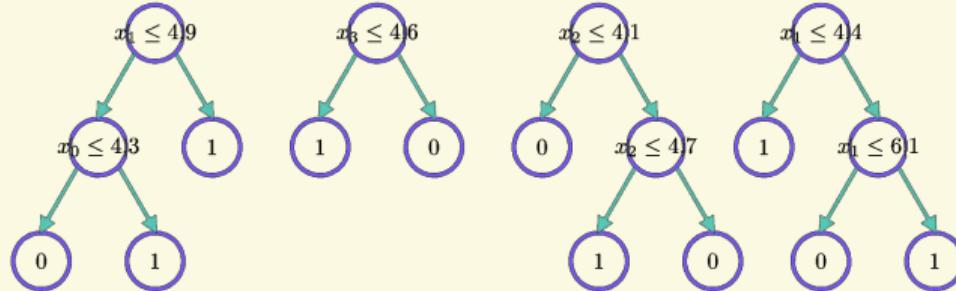
Bootstrap + Aggregating
(Bagging)

x_0, x_1

x_2, x_3

x_2, x_4

x_1, x_3



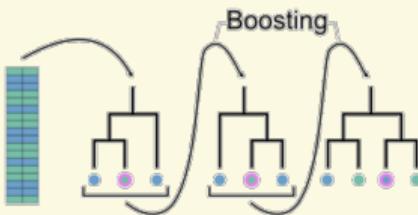
1

0

1

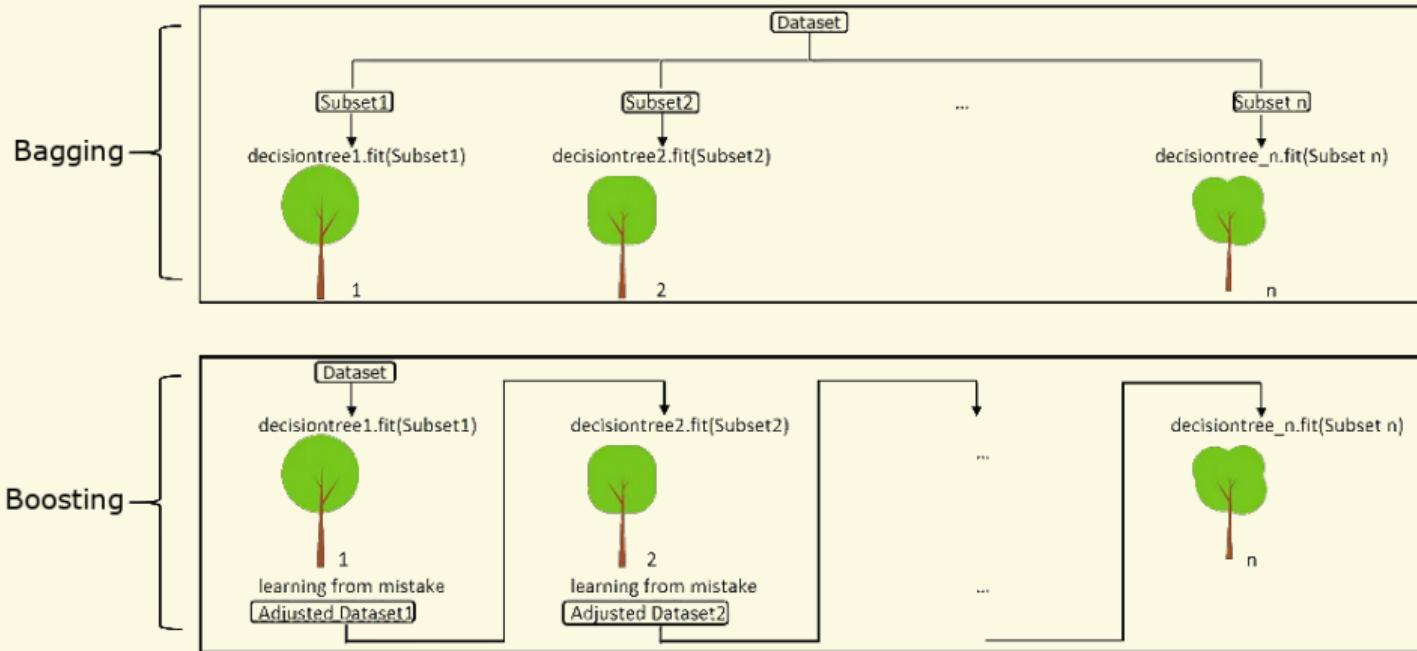
1

MACHINE LEARNING MODEL OVERVIEWS

Machine learning model	Description	Data Type	Application areas
Boosted Regression Trees: 	<p>N trees are fitted sequentially to minimize an overall loss function</p> <ul style="list-style-type: none">+ flexible(+) variable importance- many hyper-parameters- high complexity	<p>Tabular Data:</p> <ul style="list-style-type: none">- Classification- Regression	<p>functional traits decision making</p> <p>ecological contexts invasive ecosystem species distribution mortality biodiversity remote sensing</p> <p>extinction species distribution species interaction</p>

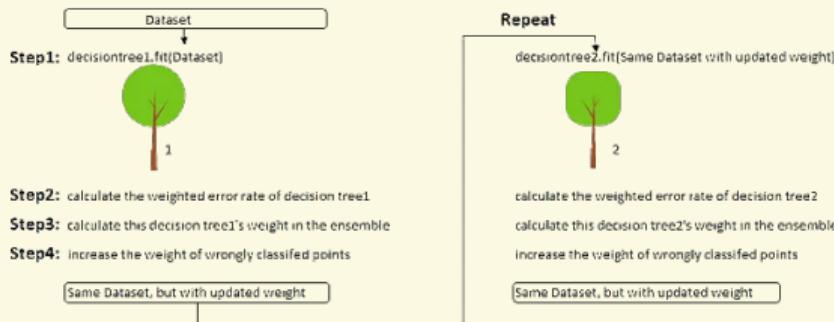
Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

BOOSTING

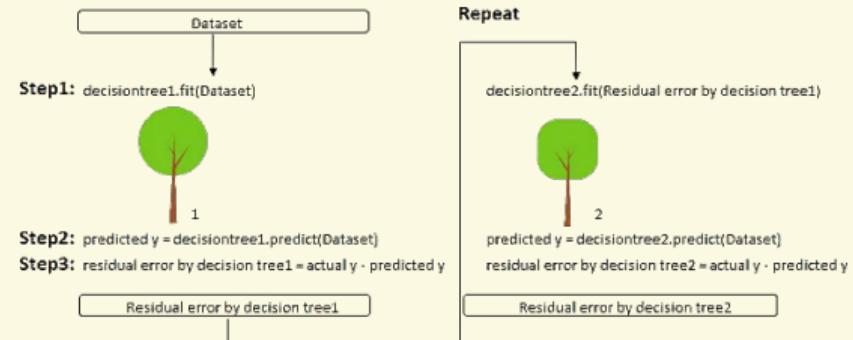


ADAPTIVE VS. GRADIENT BOOSTING

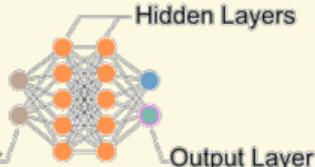
Adaptive Boosting:



Gradient Boosting:



MACHINE LEARNING MODEL OVERVIEWS

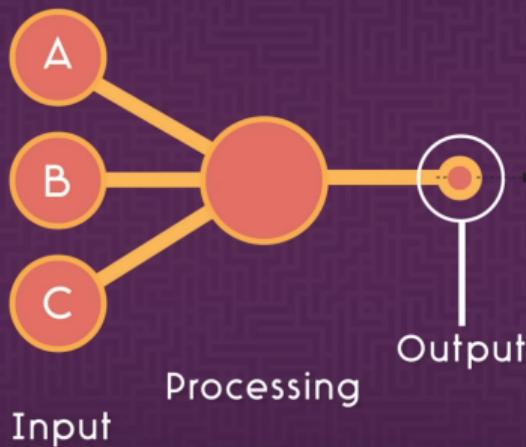
Machine learning model	Description	Data Type	Application areas
Deep Neural Networks: 	<p>Input (features) are passed through many hidden layers.</p> <p>Last layer maps into response space</p> <ul style="list-style-type: none">+ flexible+ adaptive to different tasks- many hyper-parameters- computationally expensive	<p>Tabular Data:</p> <ul style="list-style-type: none">- Classification- Regression	<p>species identification</p> <p>population trend</p> <p>species distribution</p> <p>camera trap</p> <p>biodiversity</p> <p>mortality</p> <p>remote sensing</p> <p>invasive</p> <p>ecosystem</p> <p>decision making</p> <p>extinction</p> <p>species distribution</p>

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

NEURAL NETWORKS

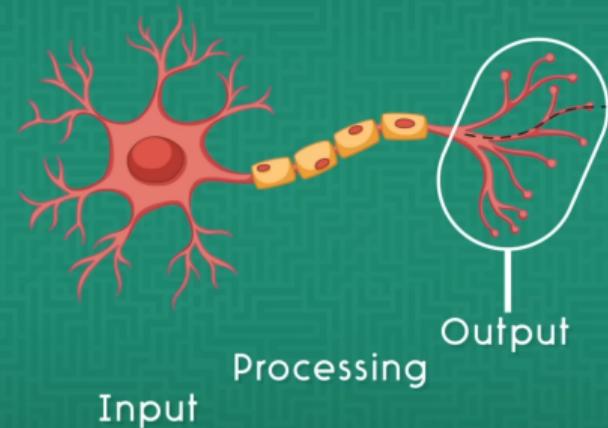
Artificial Neuron

(software)

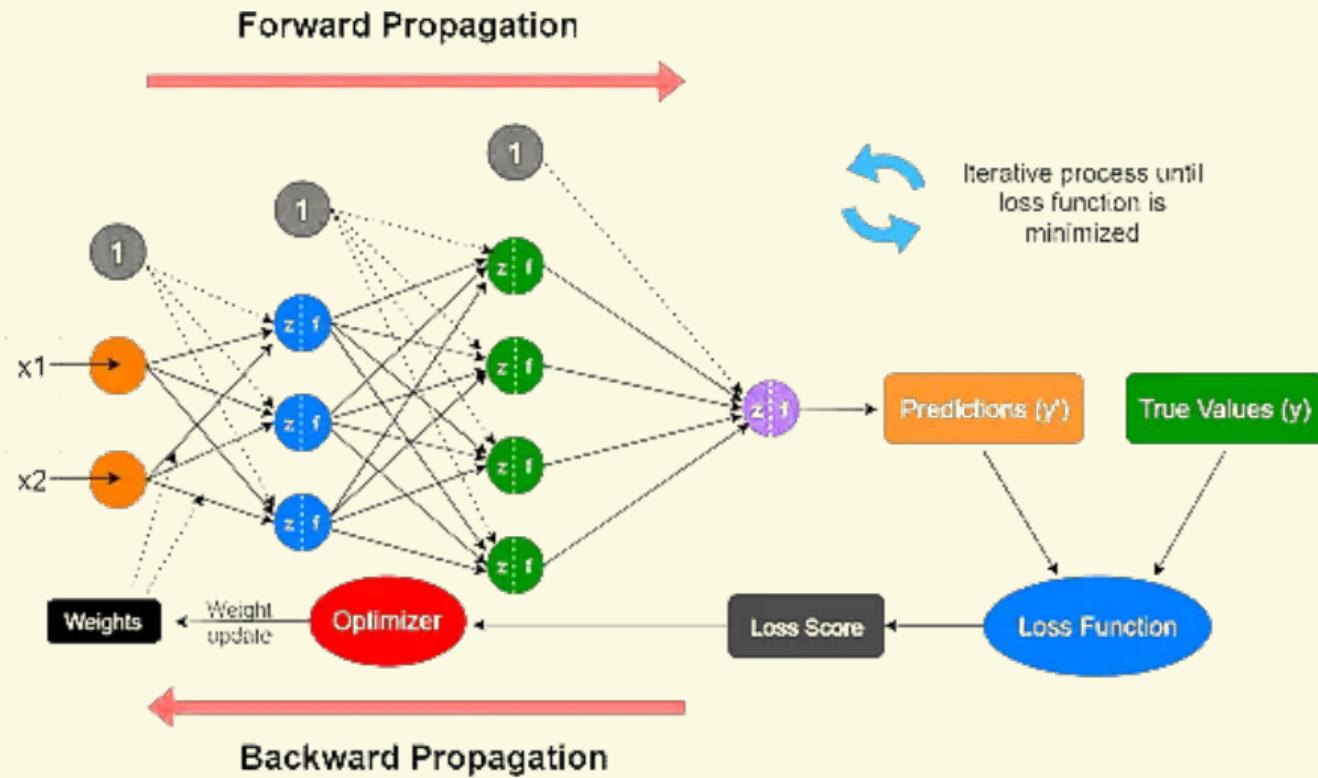


Human Neuron

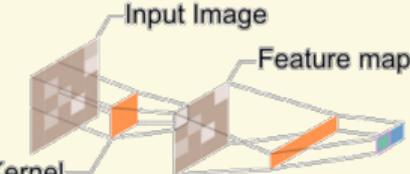
(hardware)



NEURAL NETWORKS

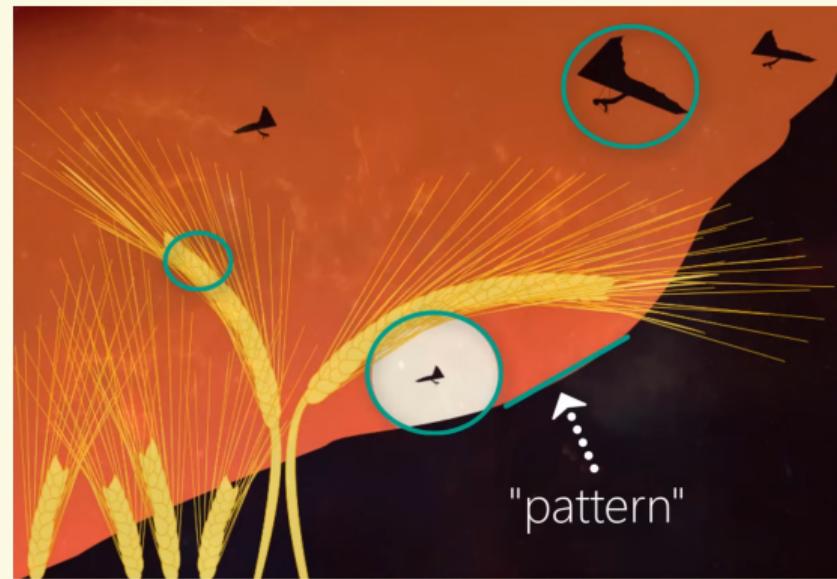


MACHINE LEARNING MODEL OVERVIEWS

Machine learning model	Description	Data Type	Application areas
<p>Convolutional Neural Networks:</p>  <p>Input Image</p> <p>Kernel</p> <p>Feature map</p>	<p>Small kernels (filters) processes images before passing it to fully connected layers</p> <ul style="list-style-type: none">+ flexible+ detecting shapes and edges- many hyper-parameters- computationally expensive	<p>Images:</p> <ul style="list-style-type: none">- Classification- Object detection	<p>functional trait</p> <p>extinction</p> <p>decision making</p> <p>remote sensing</p> <p>invasive</p> <p>mortality</p> <p>species identification</p> <p>species distribution</p> <p>ecosystem</p> <p>biodiversity</p> <p>camera trap</p> <p>species detection</p>

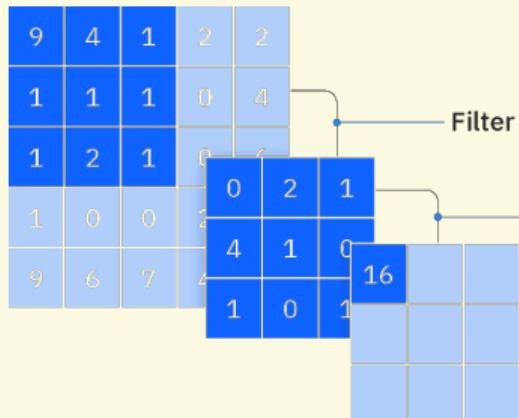
Pichler & Hartig (2022). Machine Learning and Deep Learning – A review for Ecologists. Preprint.

CONVOLUTIONAL NEURAL NETWORKS



CONVOLUTIONAL NEURAL NETWORKS

Input image

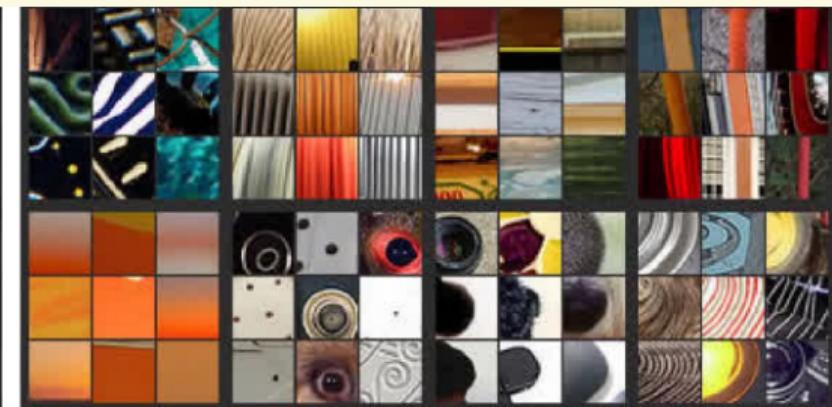
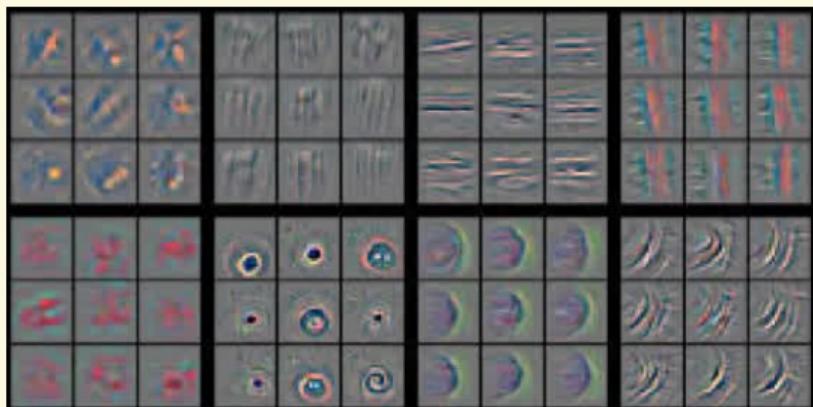


CONVOLUTIONAL NEURAL NETWORKS

CONVOLUTIONAL NEURAL NETWORKS



CONVOLUTIONAL NEURAL NETWORKS





Classification And Regression Training

