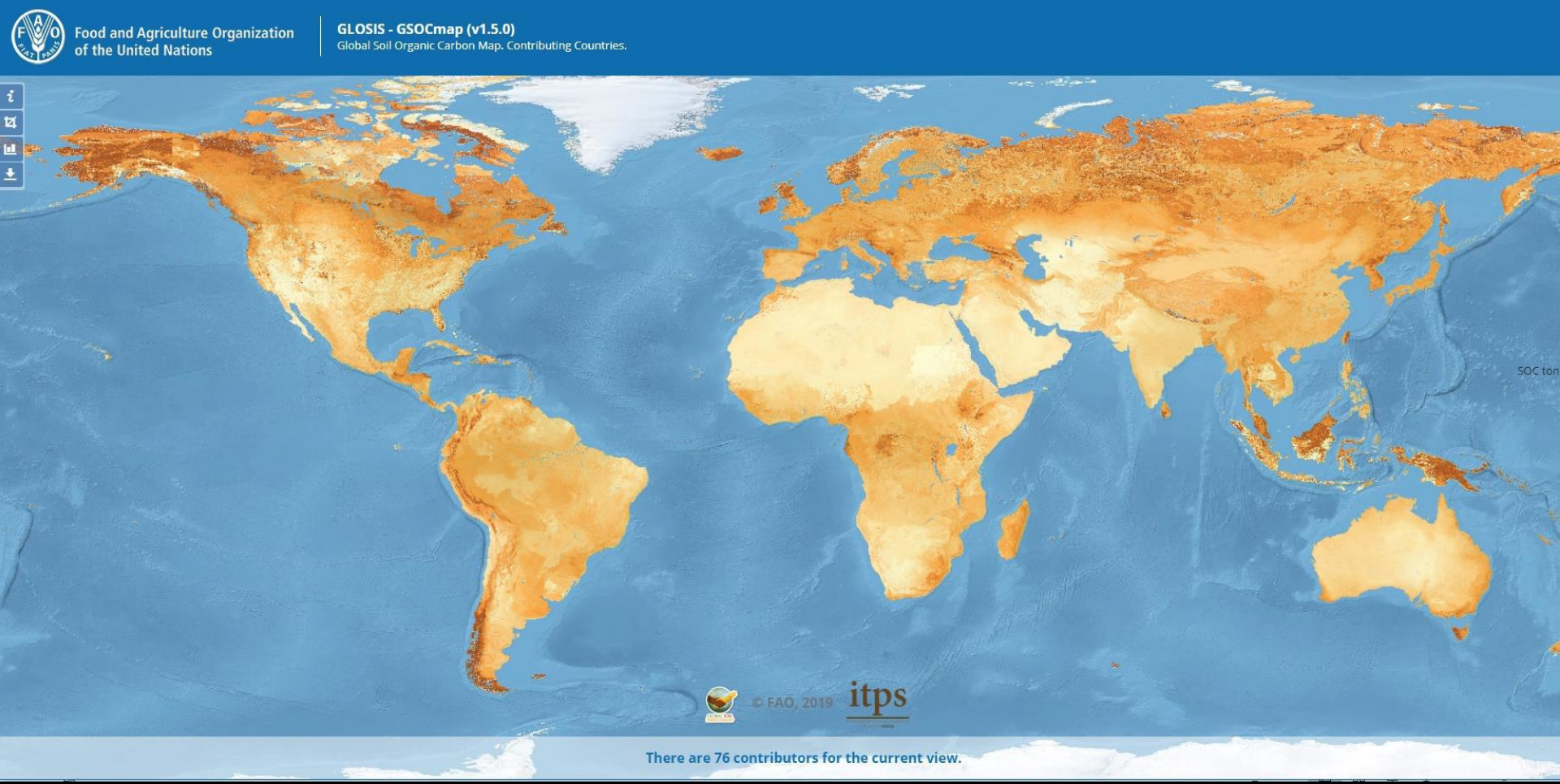


Machine Learning for Digital Soil Mapping-Part 2

Ruhollah Taghizadeh
Soil Science and Geomorphology

E-mail: *ruhollah.taghizadeh-mehrjardi@mnf.uni-tuebingen.de*

Content

- **Part 1**
  - **Digital Soil Mapping**
  - **Machine Learning**
- **Part 2**
  - **Lets practice!**

SCORPAN model

$$S = f\ (s, c, o, r, p, a, n) + \varepsilon$$

S : Soil, at a specific point in space and time: soil classes, Sc or soil attributes, Sa

From Jenny's Equation

c : climate, climate properties of the environment;
o : organisms, vegetation;
r : topography, landscape attributes;
p : parent material, lithology;
a : age or time factor;

Additions:

s : soil, prior knowledge of the soil at a point;
n : space, relative spatial position;
ε : auto-correlated random spatial variation.

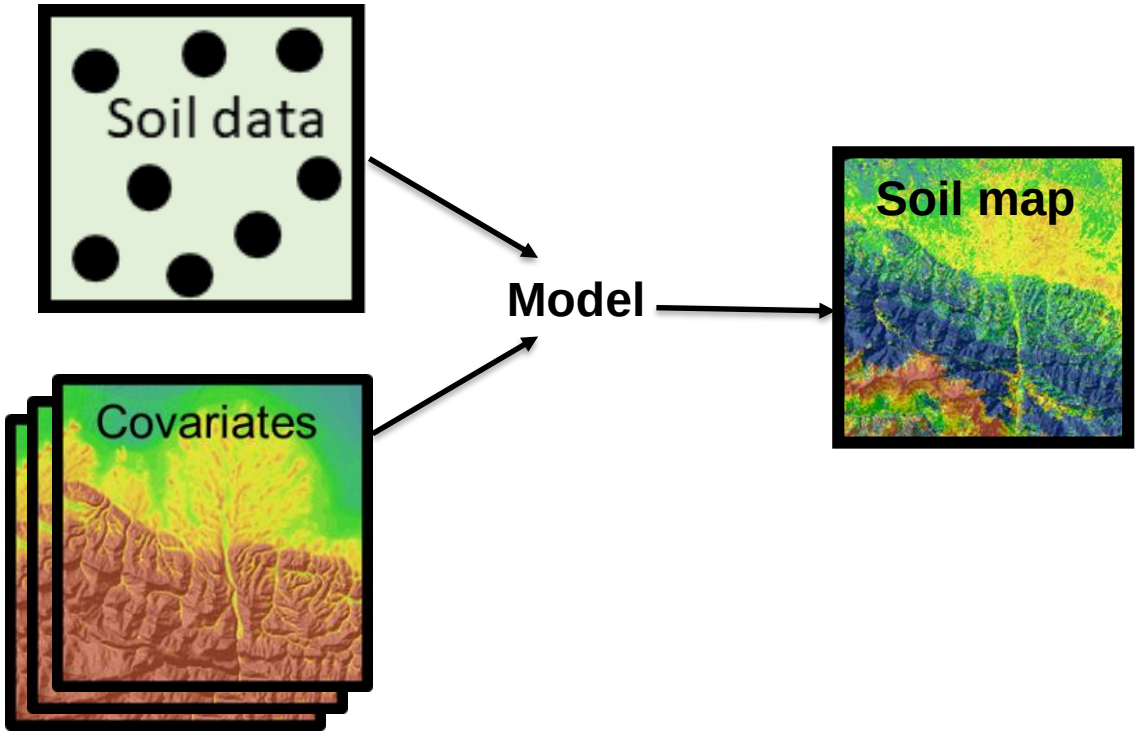
f() : Quantitative function *f* linking S to ***scorpan*** factors

McBratney et al., 2003

SCORPAN model

$$S = f(s, c, o, r, p, a, n) + \varepsilon$$

- S: Soil data
- f: Model
- (s, c, o, r, p, a, n): Covariates
- ε: spatially dependent residuals



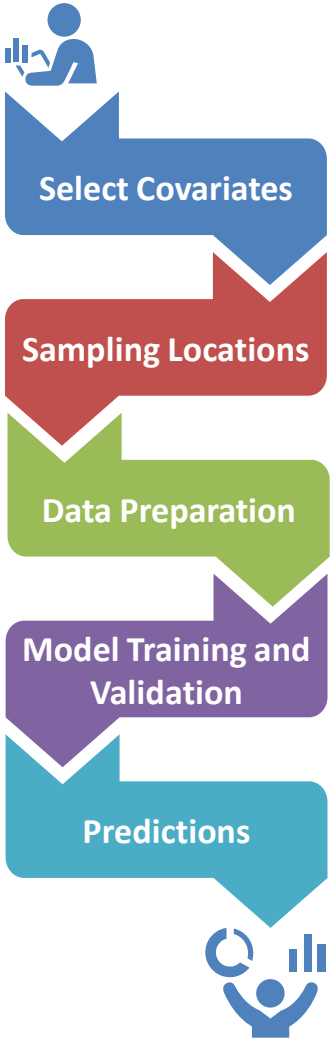
SCORPAN Model

$$S = f(s,c,o,r,p,a,n) + \varepsilon$$

Machine learning

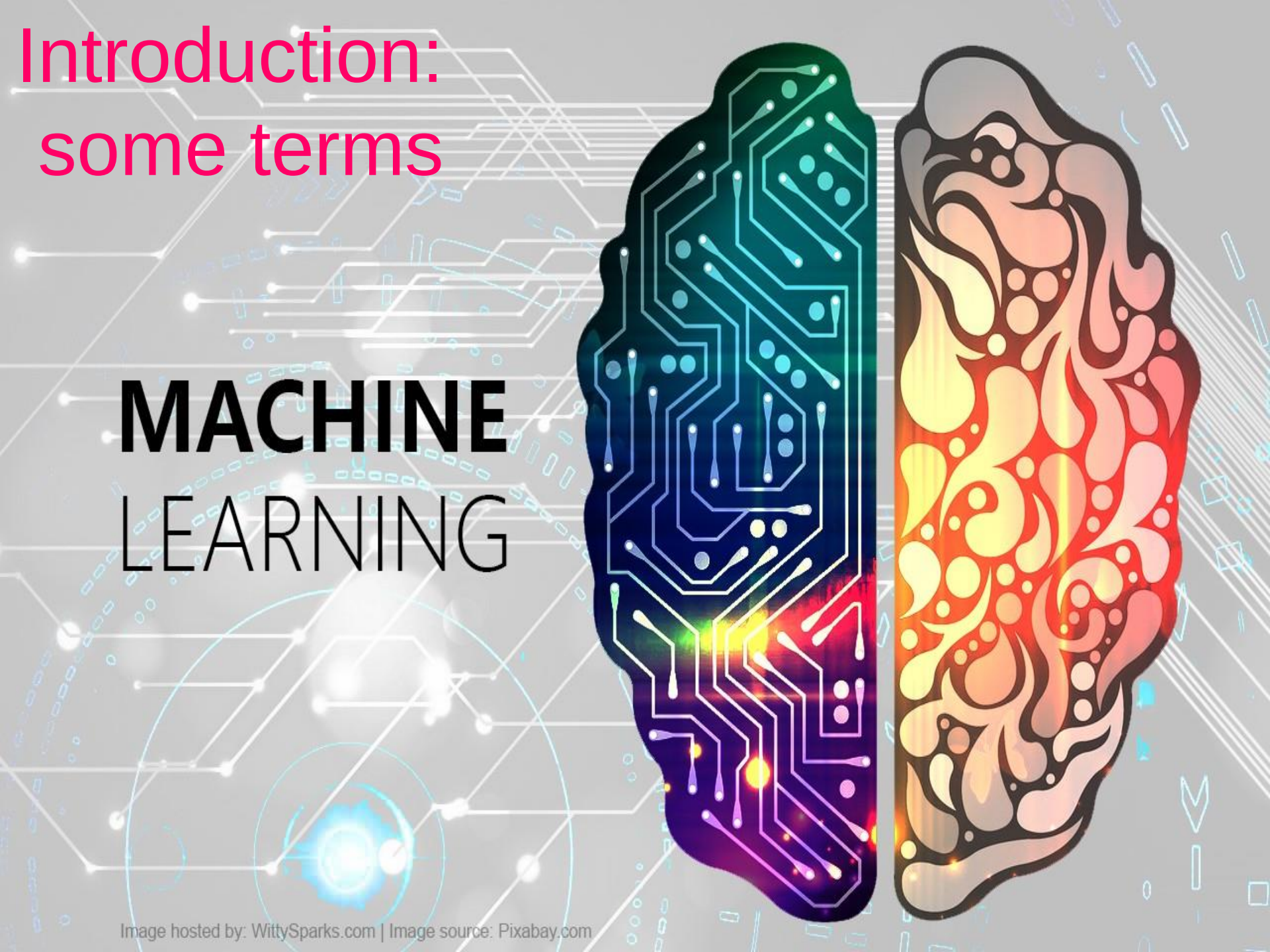
f() : Quantitative function *f* linking S to ***scorpan*** factors

Sequence of DSM Steps

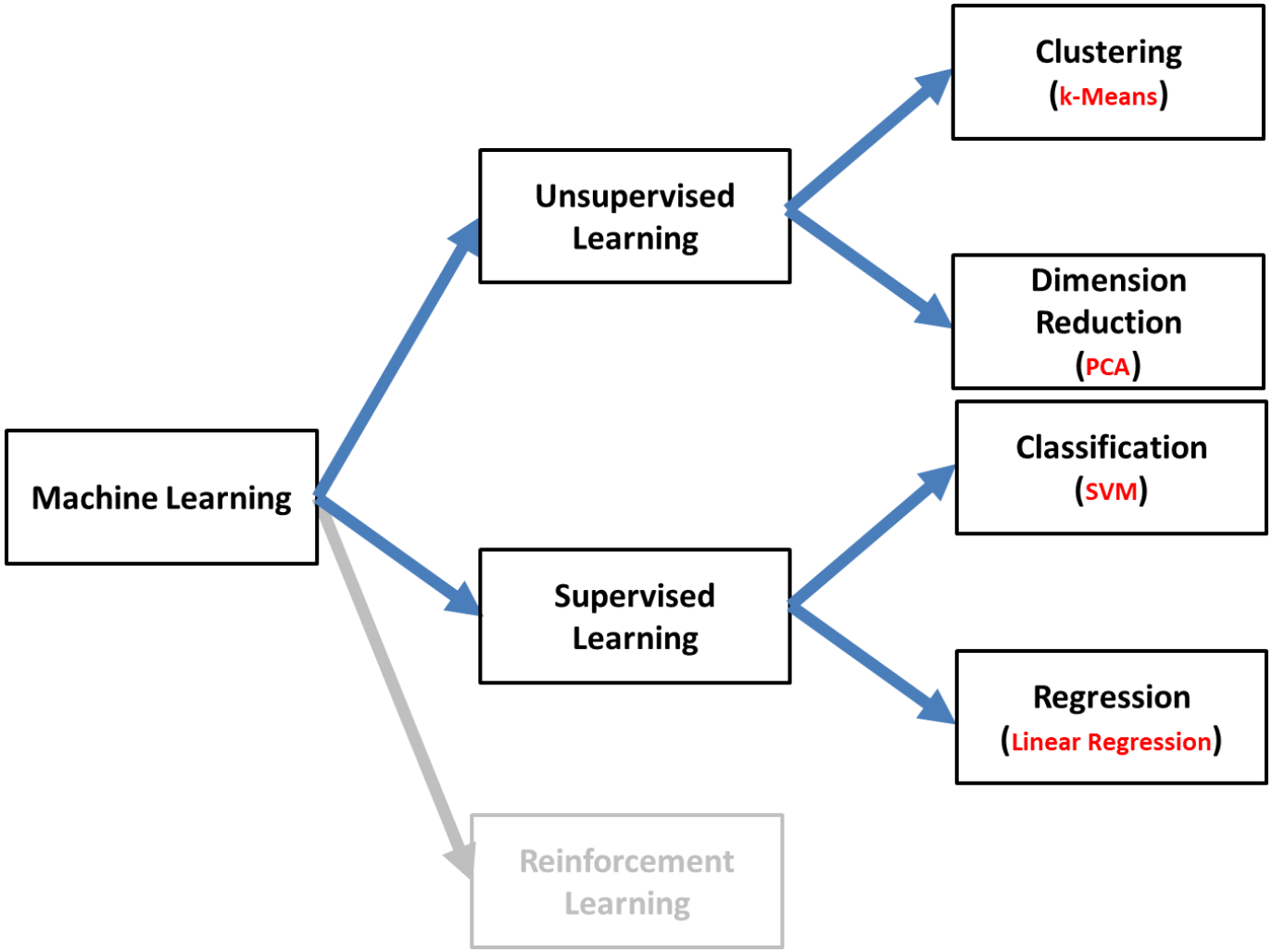


1. Environmental covariates, relevant as predictors of soil property/class, are derived from remote sensing, digital elevation, climatic datasets, …
2. Soil samples are collected at the specified locations (e.g., Latin hypercube sampling) and soil property is measured in the laboratory.
3. Intersecting the covariates with the soil point observations.
4. Machine learning models (e.g., random forest) are trained using training data, and accuracy assessment is carried out using the test data set.
5. The ML models are applied to the entire study area in order to produce a soil property/class map.

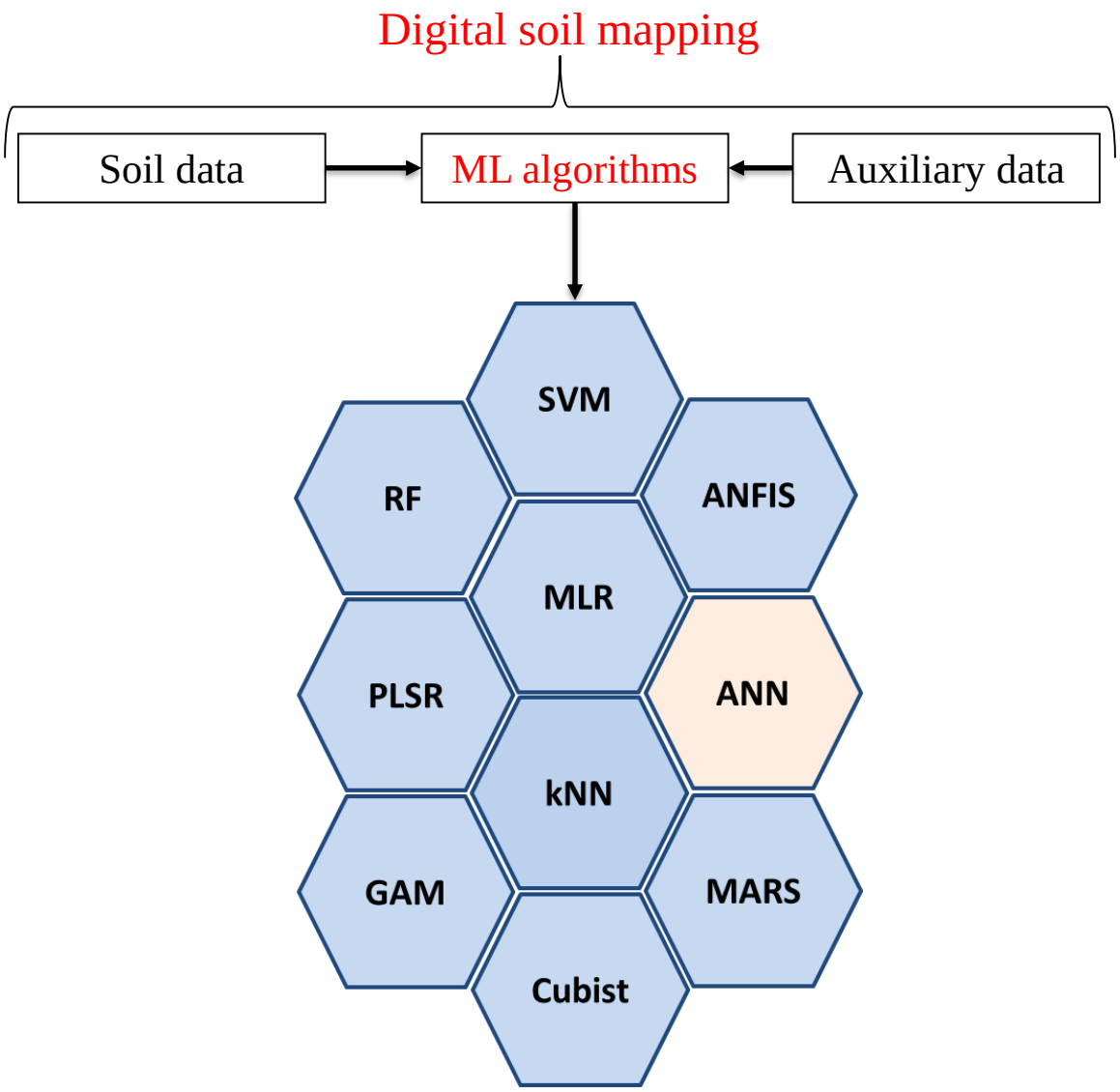
Introduction: some terms



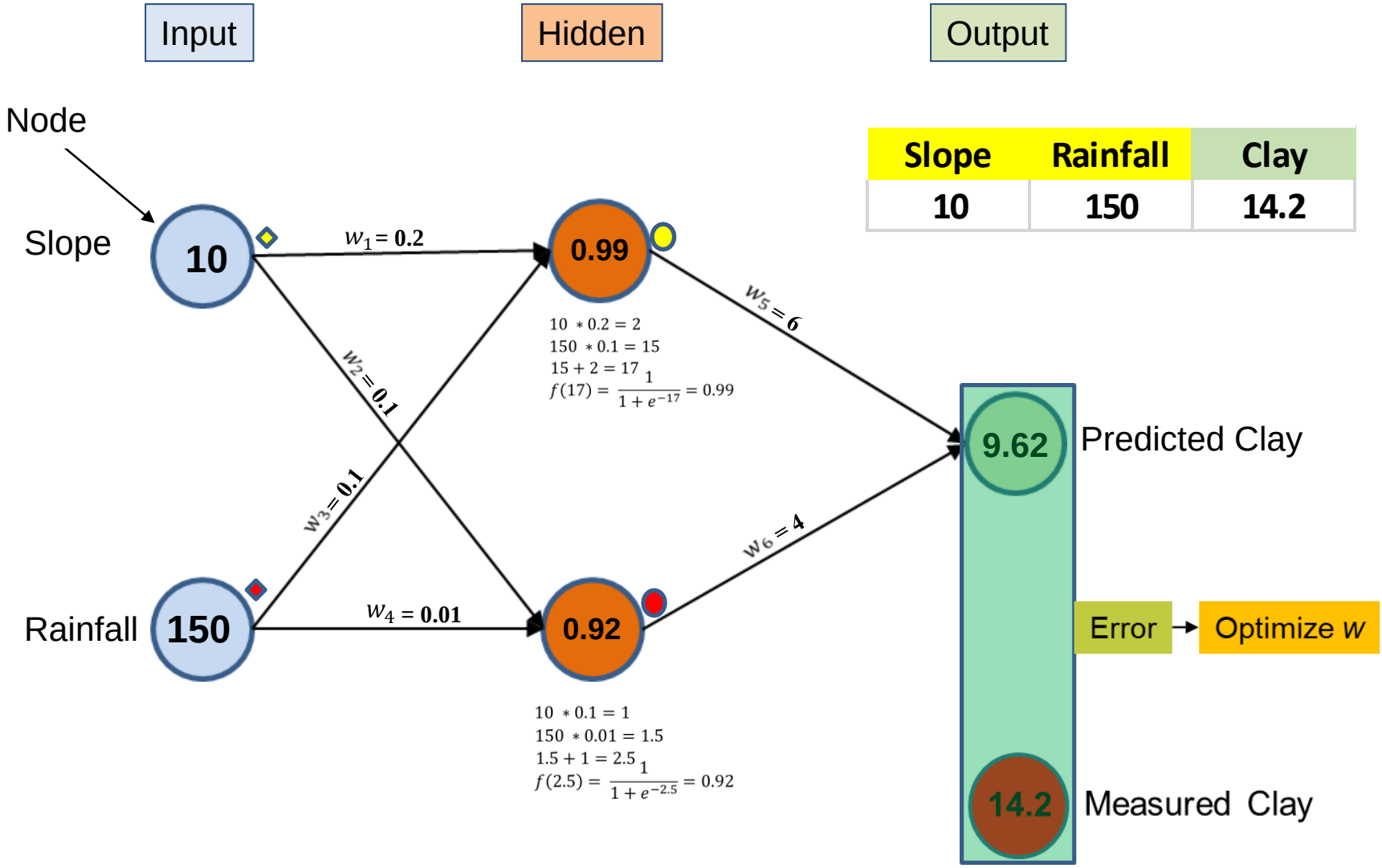
Types of Machine Learning



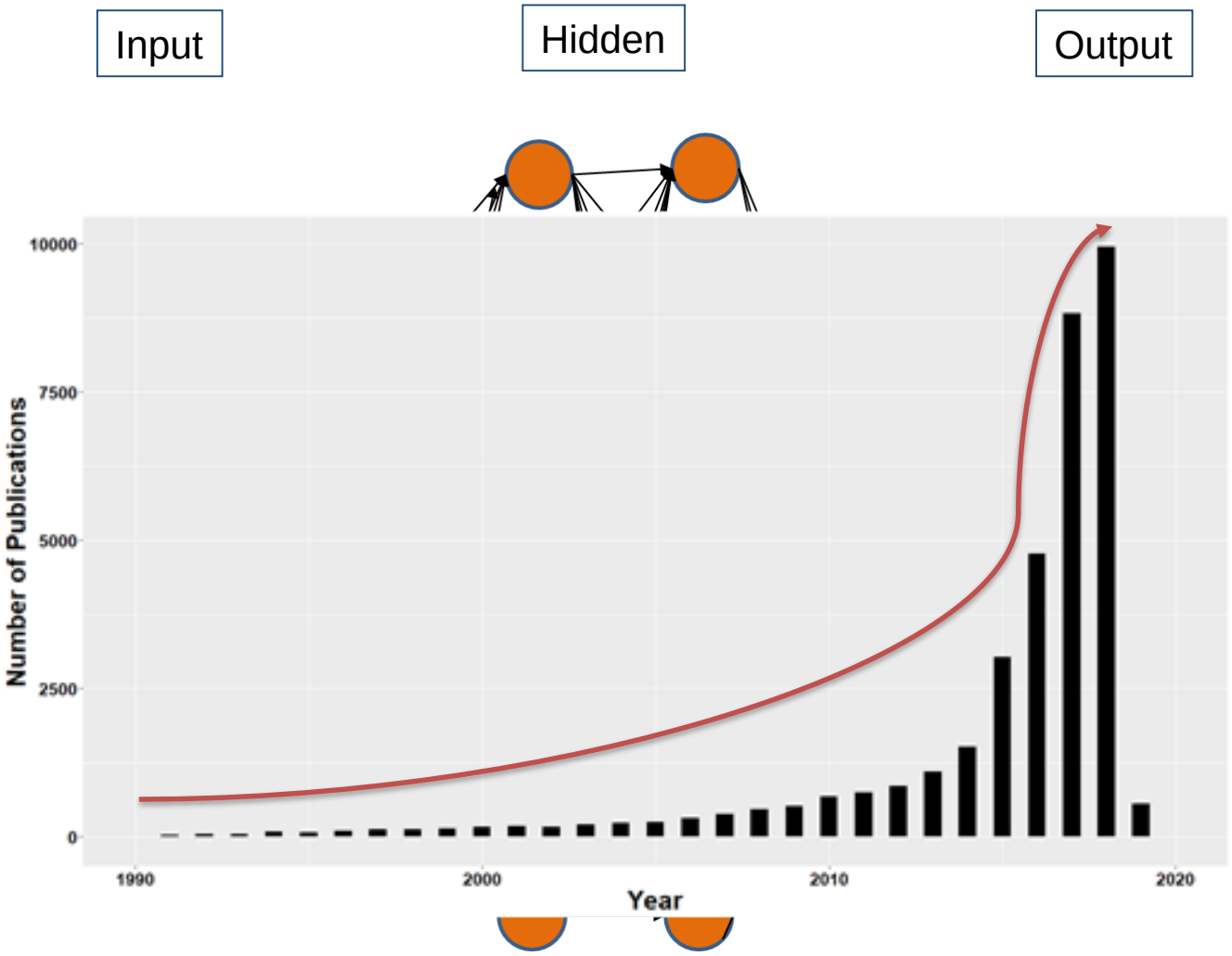
❑ Different ML algorithms in DSM



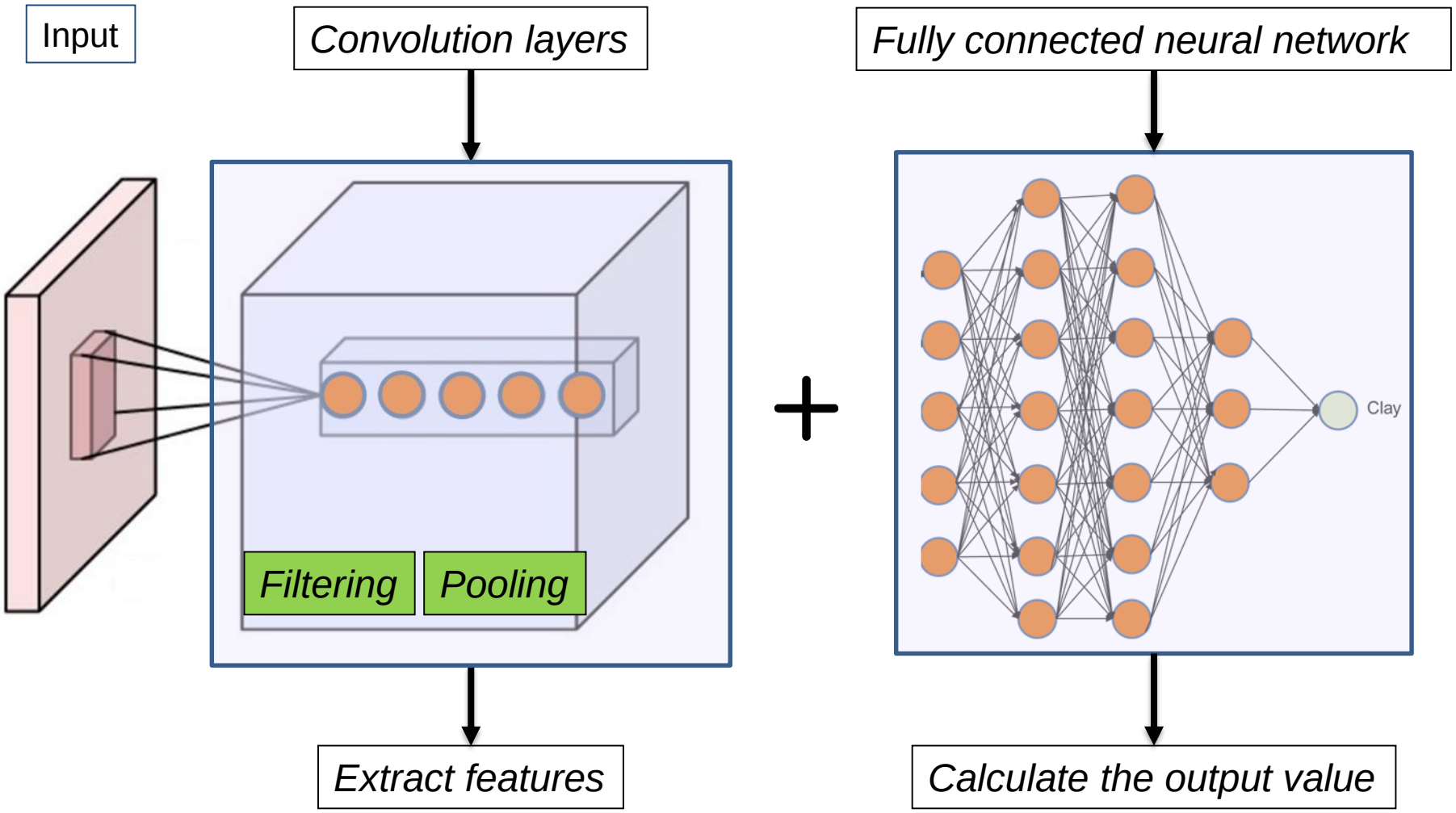
❑ Artificial neural networks (ANN)



❑ Deep artificial neural networks (DL)

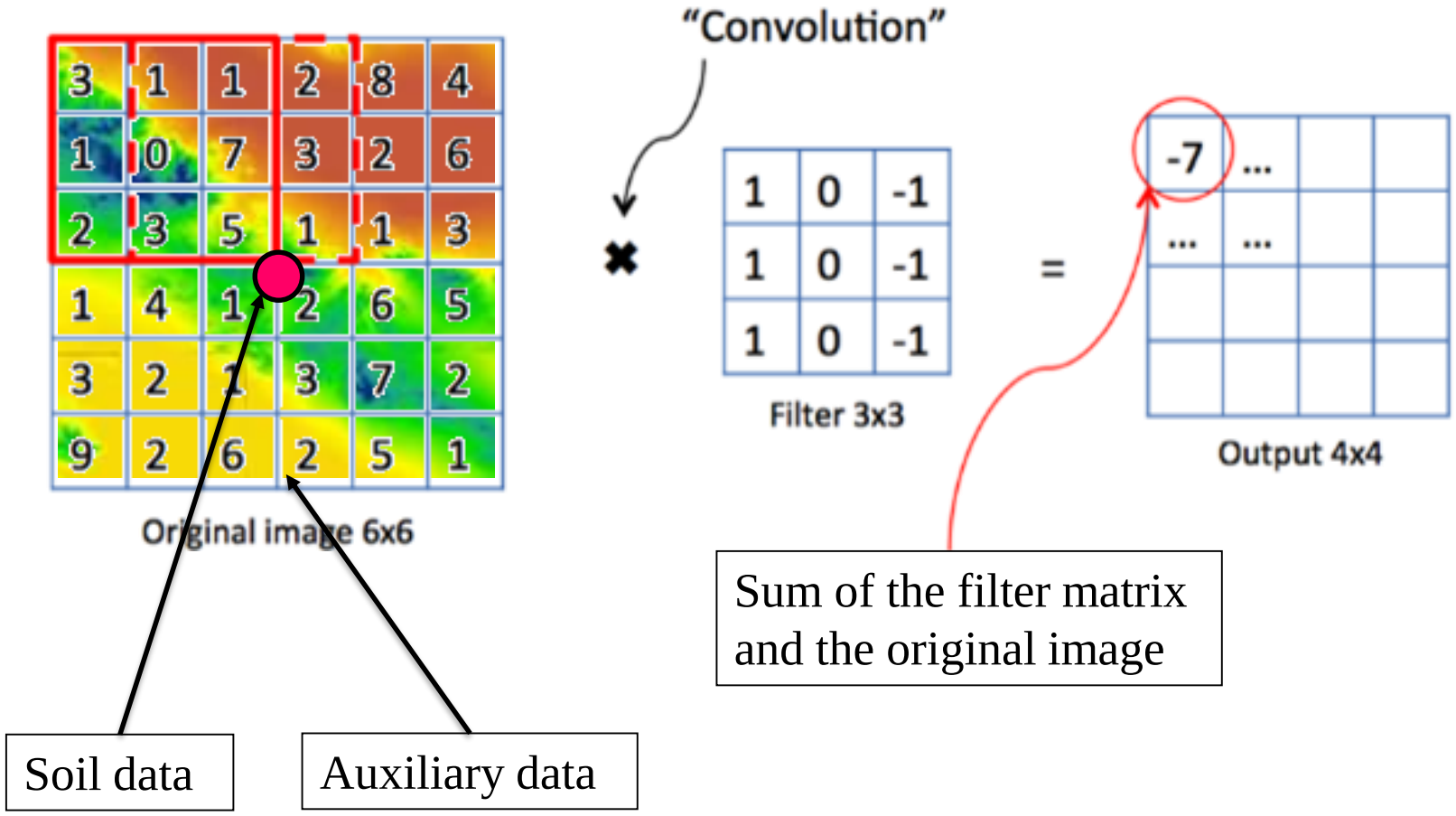


❑ Convolutional neural networks (CNN)



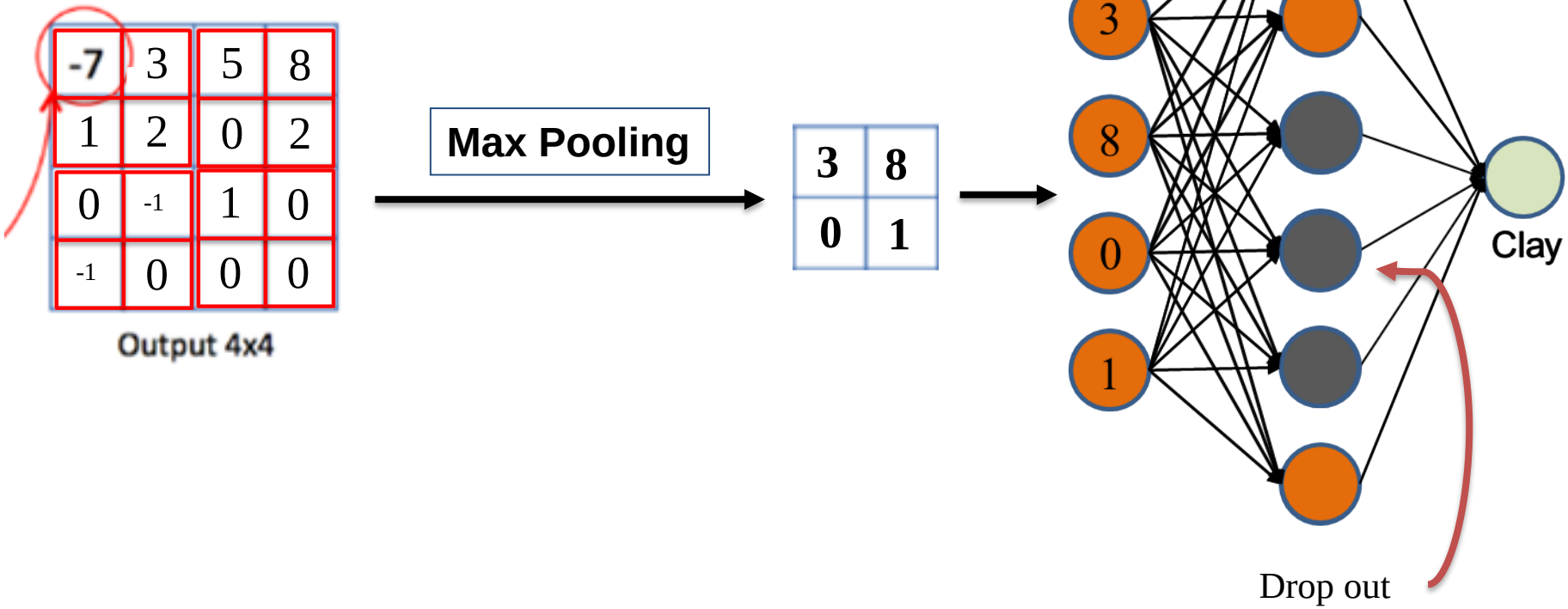
❑ Convolutional neural networks (CNN)

- Convolutional layers: Filters

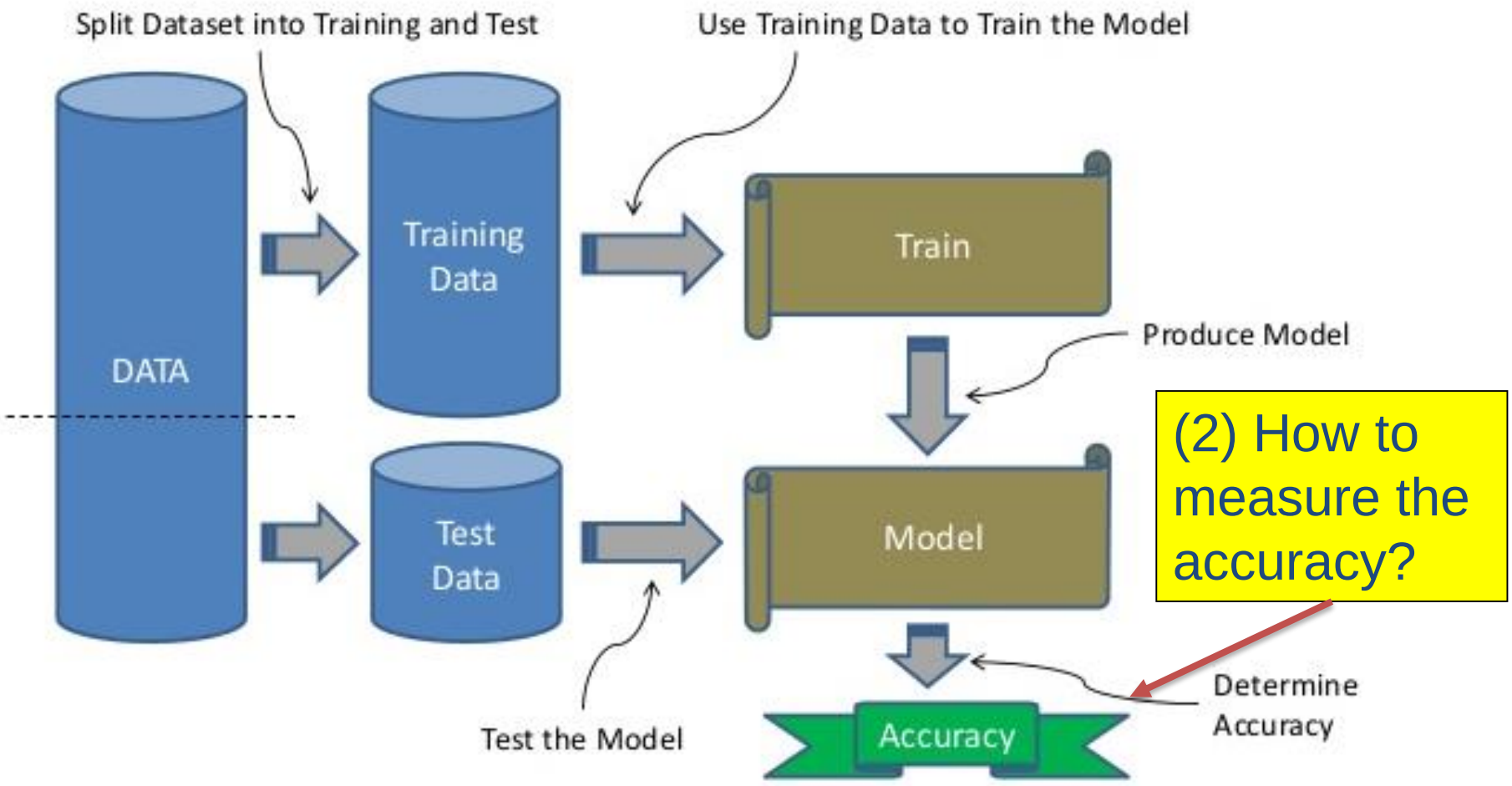


❑ Convolutional neural networks (CNN)

- Convolutional layers: Pooling
- Fully connected neural network



Accuracy Assessment of Models



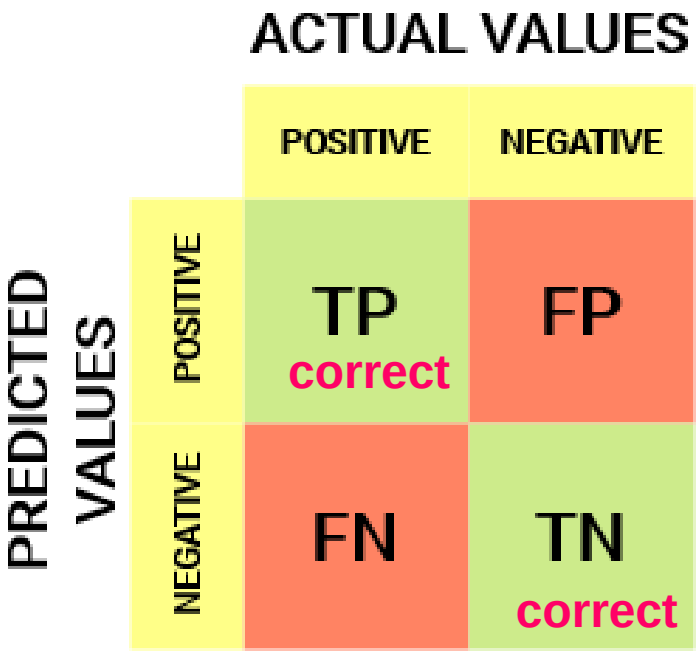
How to measure the accuracy?

- Performance Metrics:

Regression	Classification
• Mean Absolute Error (MAE) • Root Mean Squared Error (RMSE) • R-Squared and Adjusted R-Squared	• Recall • Precision • F1-Score • Accuracy • Area Under the Curve (AUC)

Performance Metrics for Classification

- Confusion Matrix: is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model.



where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively

Performance Metrics for Classification

- Confusion Matrix:

Class A: Aridisols

Class B: Entisols

	Class A: Aridisoils	Class B: Entisols
Class A: Aridisoils		
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

Class A: Aridisols

Class B: Entisols

Actual Test data

Predicted Test data

	Class A: Aridisoils	**Class B: Entisols**
Class A: Aridisoils		
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

Class A: Aridisols

Class B: Entisols

Actual Test data

1 2 3 4 5

Predicted Test data

1 2 3 4 5

	Class A: Aridisols	**Class B: Entisols**
Class A: Aridisols		
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

Class A: Aridisols

Class B: Entisols

Actual Test data

1 2 3 4 5

Predicted Test data

1 2 3 4 5

	Class A: Aridisols	**Class B: Entisols**
Class A: Aridisols	| |	
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

Class A: Aridisols

Class B: Entisols

Actual Test data

1 2 3 4 5

Predicted Test data

1 2 3 4 5

	Class A: Aridisols	**Class B: Entisols**
Class A: Aridisols	| |	
Class B: Entisols	|	

Performance Metrics for Classification

- Confusion Matrix:

Class A: Aridisols

Class B: Entisols

Actual Test data

1 2 3 4 5

Predicted Test data

1 2 3 4 5

	Class A: Aridisols	**Class B: Entisols**
Class A: Aridisols	| |	
Class B: Entisols	|	|

Performance Metrics for Classification

- Confusion Matrix:

Class A: Aridisols

Class B: Entisols

Actual Test data

1 2 3 4 5

Predicted Test data

1 2 3 4 5

	Class A: Aridisols	**Class B: Entisols**
Class A: Aridisols	| |	|
Class B: Entisols	|	|

Performance Metrics for Classification

- Confusion Matrix:

Class A: Aridisols

Class B: Entisols

Actual Test data

1 2 3 4 5

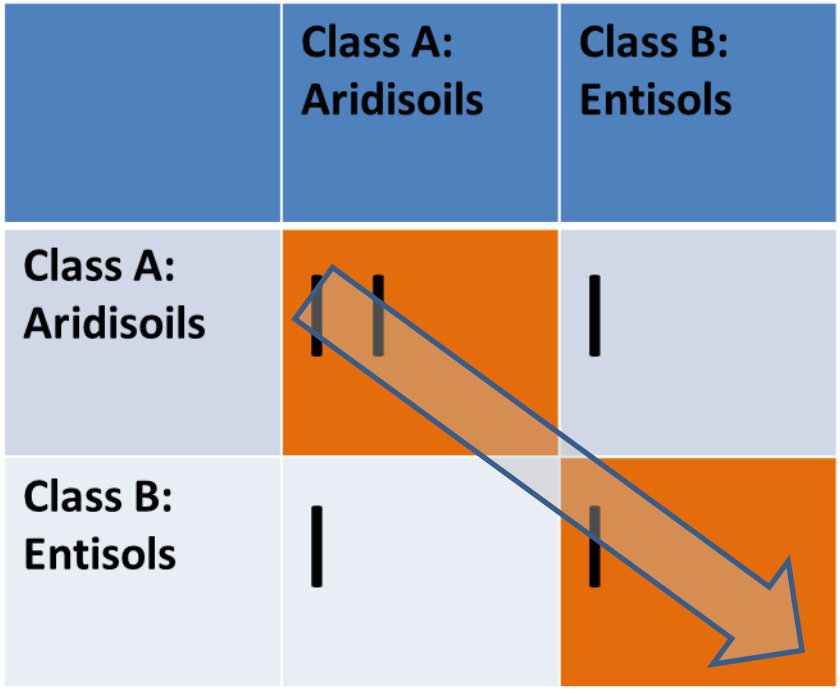
Predicted Test data

1 2 3 4 5

	Class A: Aridisols	Class B: Entisols
Class A: Aridisols	2	1
Class B: Entisols	1	1

Performance Metrics for Classification

- Overall Accuracy: is a metric calculating the classifier overall accuracy

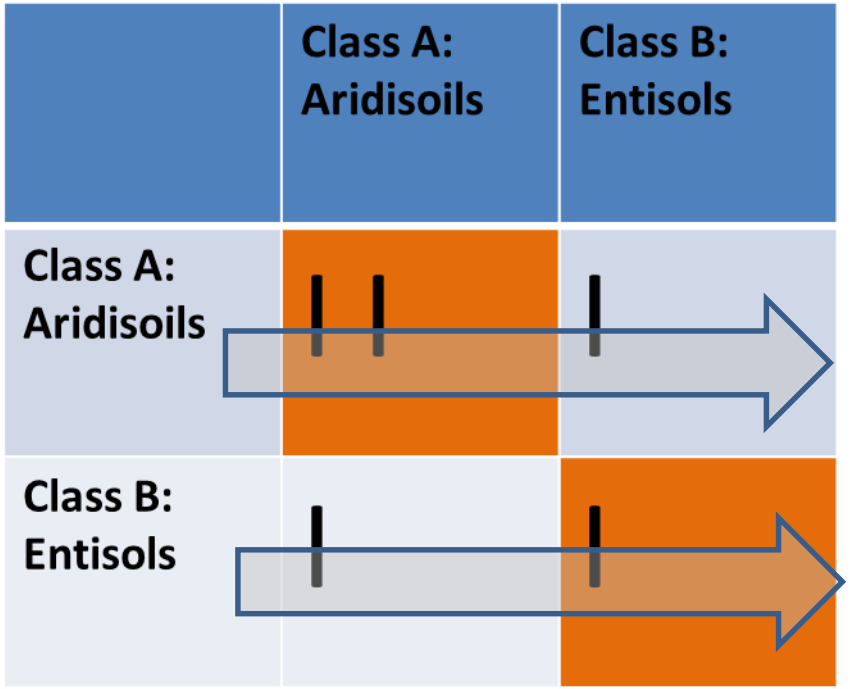


$$OA = \left(\frac{\text{Correctly classified}}{\text{Total number of test data}}\right) * 100$$

$$OA = \left(\frac{2+1}{5}\right) * 100 = 60\%$$

Performance Metrics for Classification

- Precision: is the proportion of those predicted instances that are correctly classified



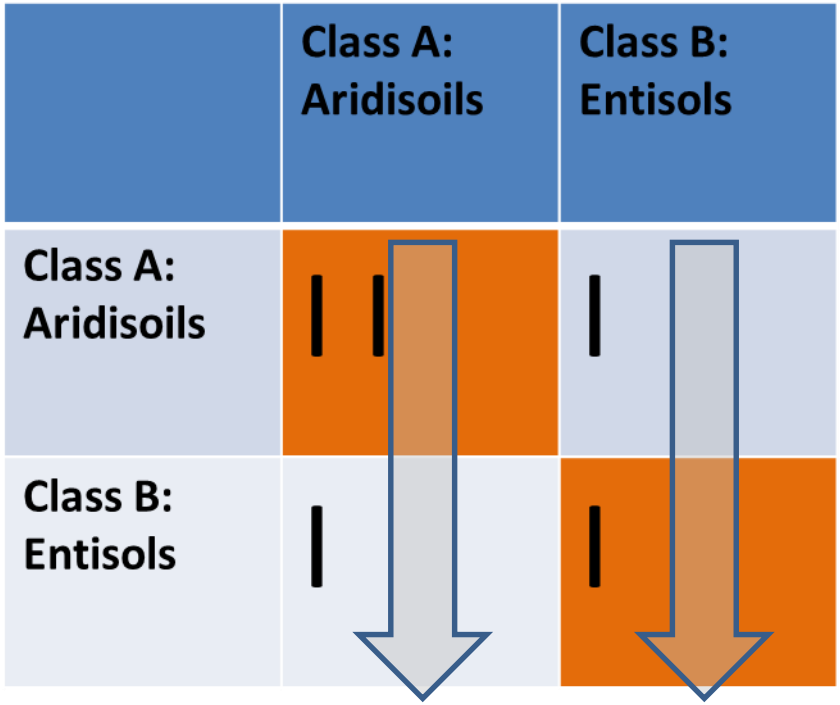
$$Pr = \left(\frac{\text{Correctly classified for each class}}{\text{Total number of row data}}\right) * 100$$

$$Pr(class\ A) = \left(\frac{2}{2+1}\right) * 100 = 66\%$$

$$Pr(class\ B) = \left(\frac{1}{1+1}\right) * 100 = 50\%$$

Performance Metrics for Classification

- Recall: is the proportion of those instances that are correctly classified



$$Re = \left(\frac{\text{Correctly classified for each class}}{\text{Total number of coloum data}}\right) * 100$$

$$Re(class\ A) = \left(\frac{2}{2+1}\right) * 100 = 66\%$$

$$Re(class\ B) = \left(\frac{1}{1+1}\right) * 100 = 50\%$$

Performance Metrics for Classification

- F-score: the F-score is the harmonic mean of precision and recall

$$F - score = \left(\frac{2 \times Preciosion \times Recall}{Preciosion + Recall} \right) * 100$$

Metric	Formula
Accuracy	$ACC = \frac{TP+TN}{TP+TN+FP+FN}$
Error rate	$ERR = \frac{FP+FN}{TP+TN+FP+FN}$
Precision	$PRC = \frac{TP}{TP+FP}$
Sensitivity Specificity	$SNS = \frac{TP}{TP+FN}$ $SPC = \frac{TN}{TN+FP}$
ROC	$ROC = \frac{\sqrt{SNS^2+SPC^2}}{\sqrt{2}}$
F_1 score	$F_1 = 2\frac{PRC \cdot SNS}{PRC+SNS}$
Geometric Mean	$GM = \sqrt{SNS \cdot SPC}$

Practice

Import covariates

Import point data

Overlay point data on covariates

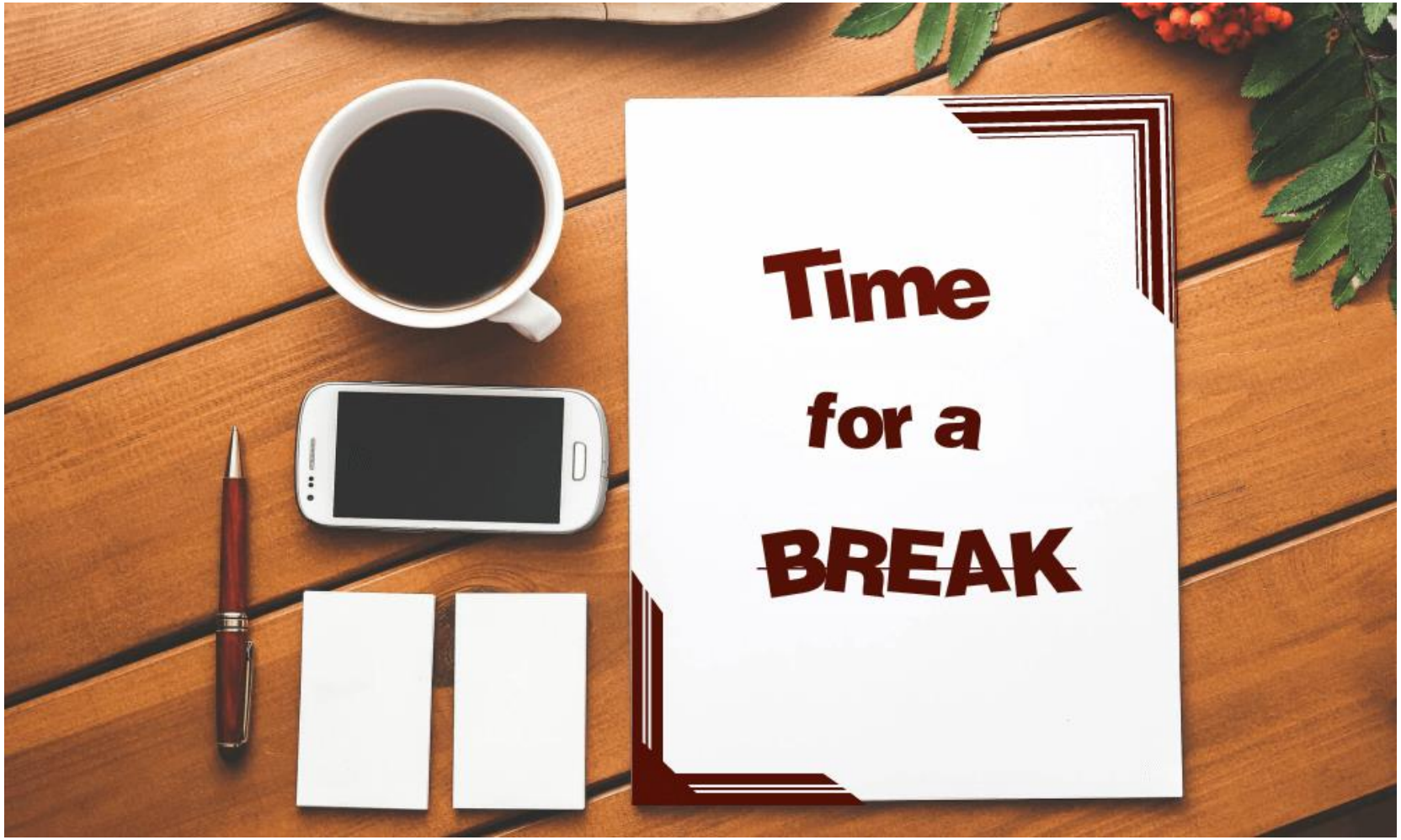
Make a geodatabase

Split the geodatabase to training and testing sets

Train machine learning models

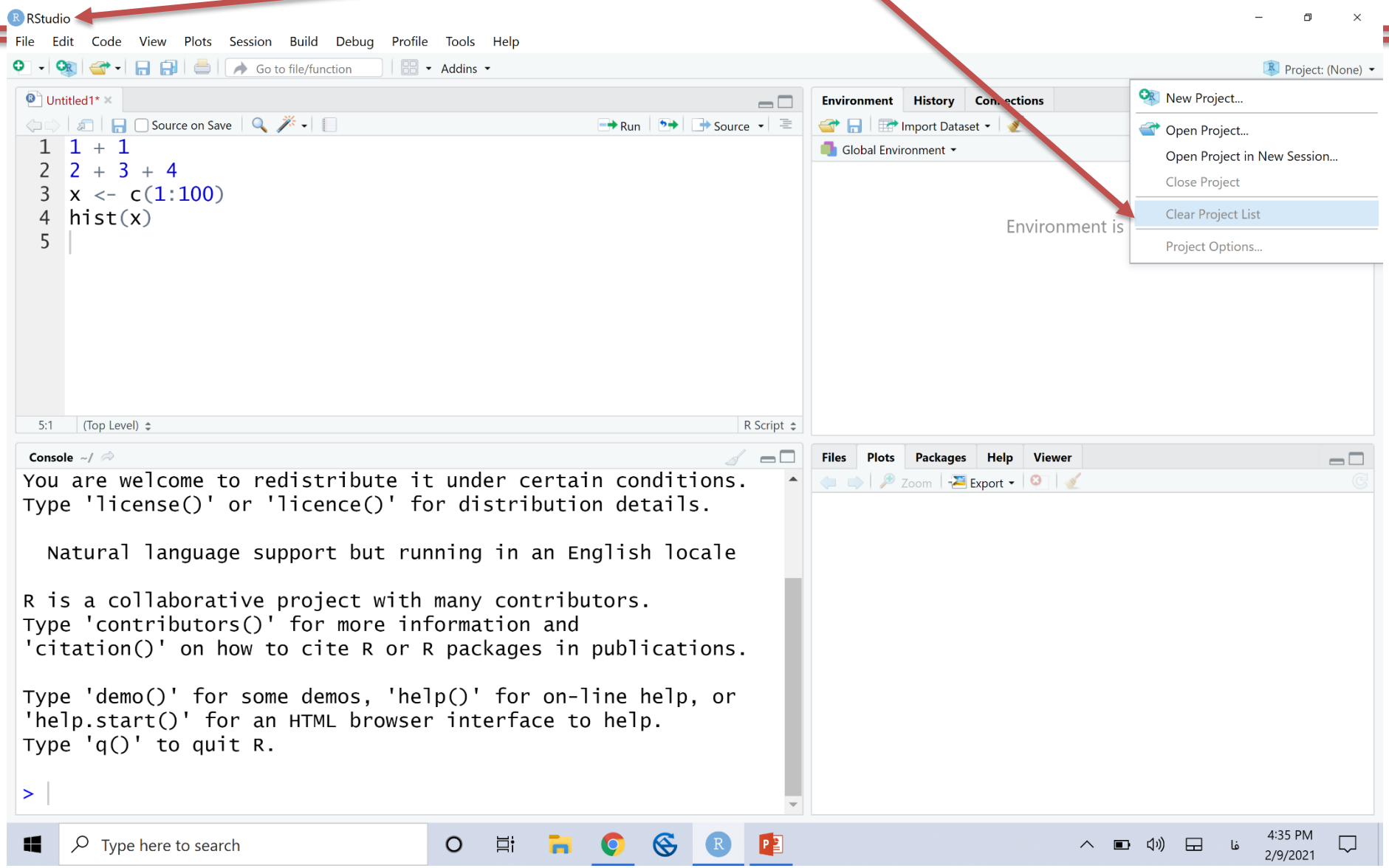
Test machine learning models

Predict soil maps

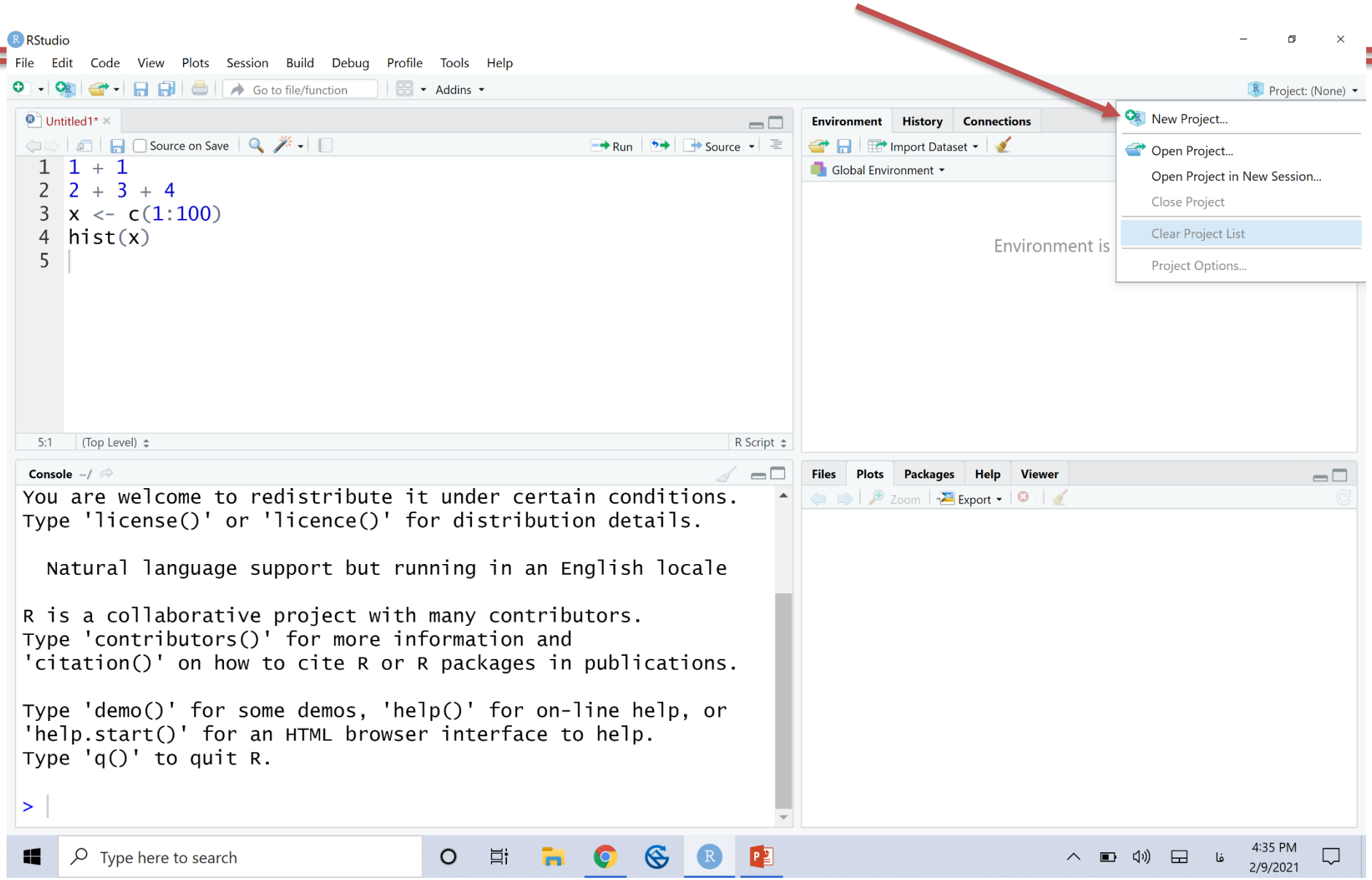

Time
for a
BREAK

1. Open Rstudio

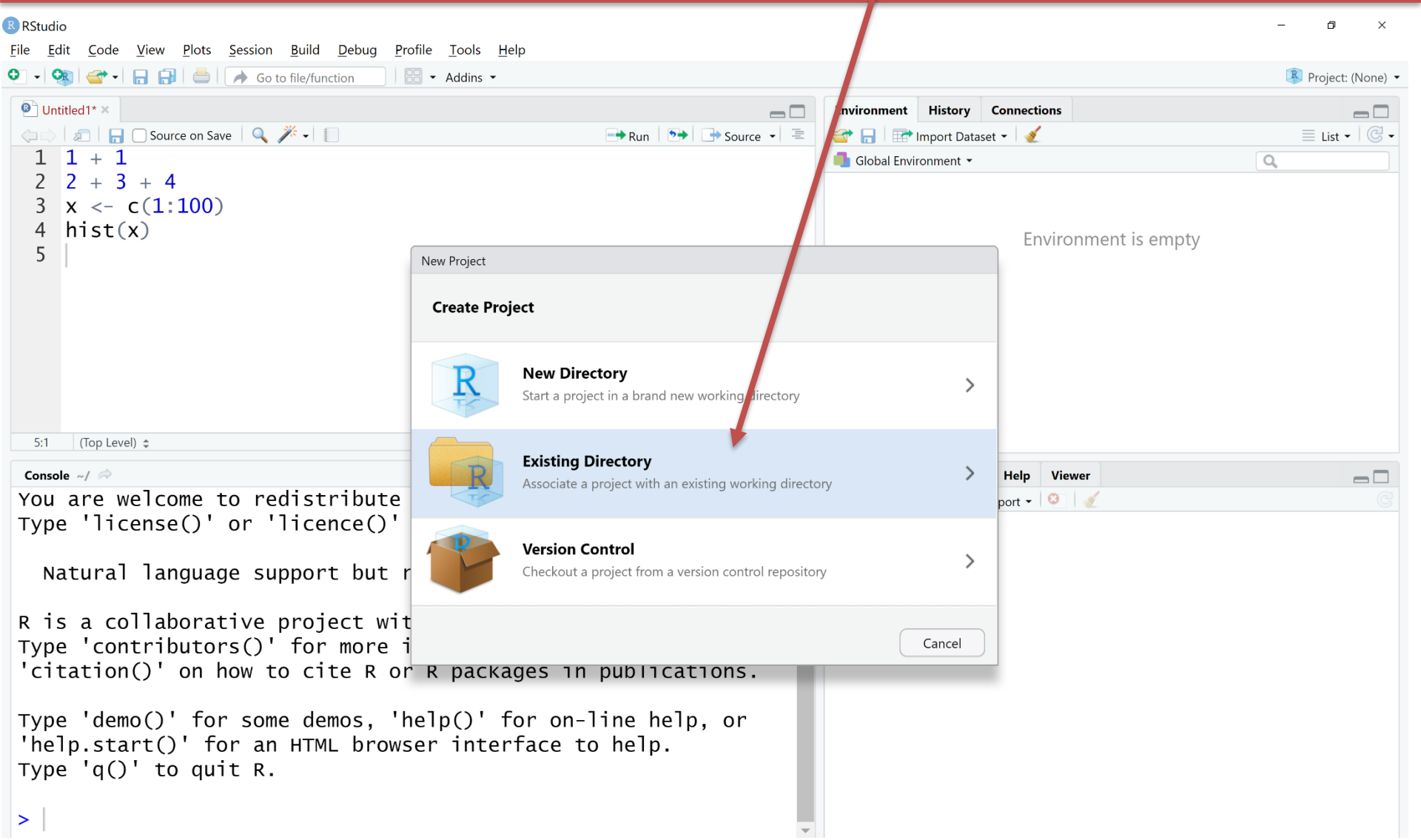
2. Clear Project List



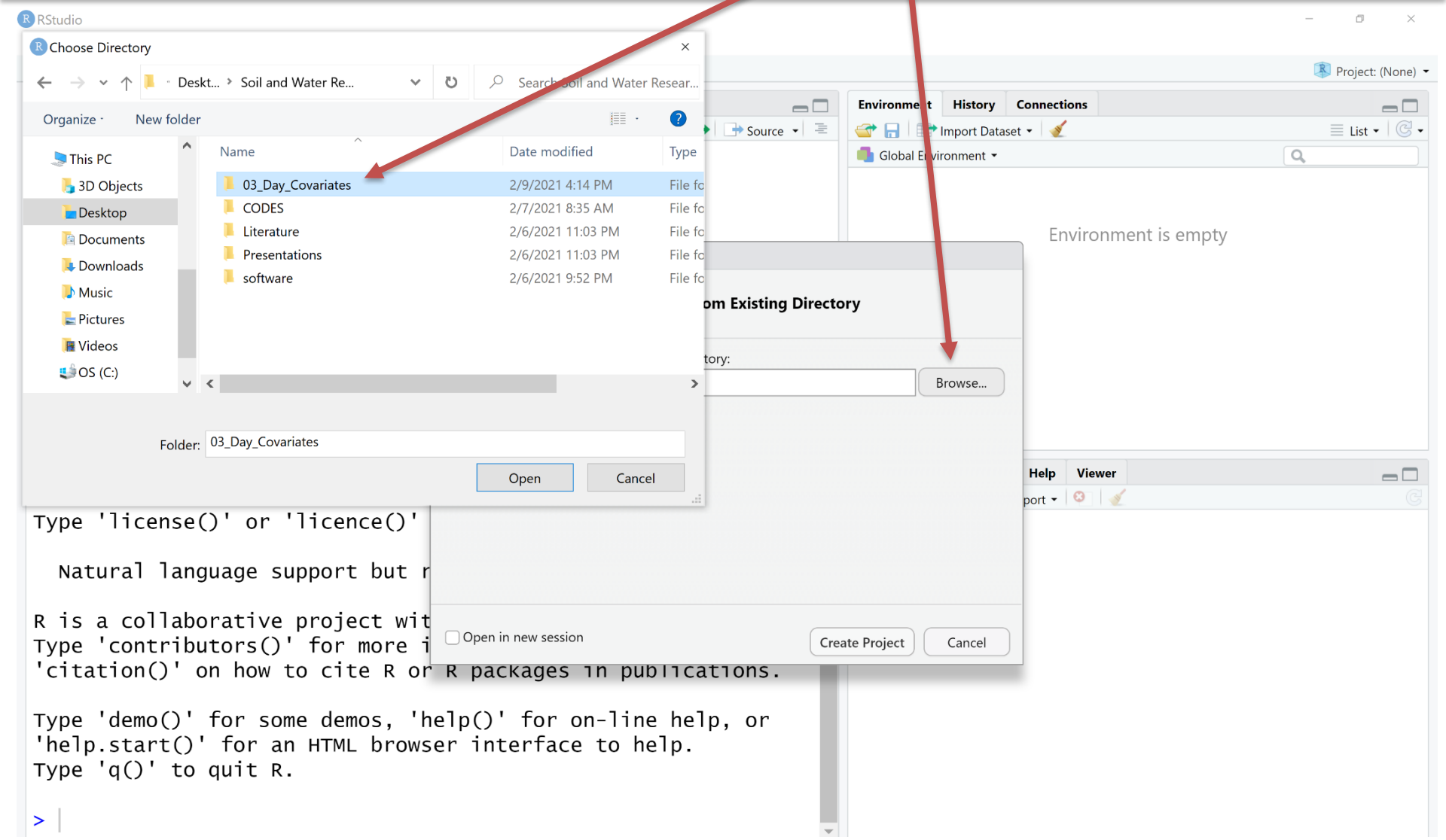
3. New Project



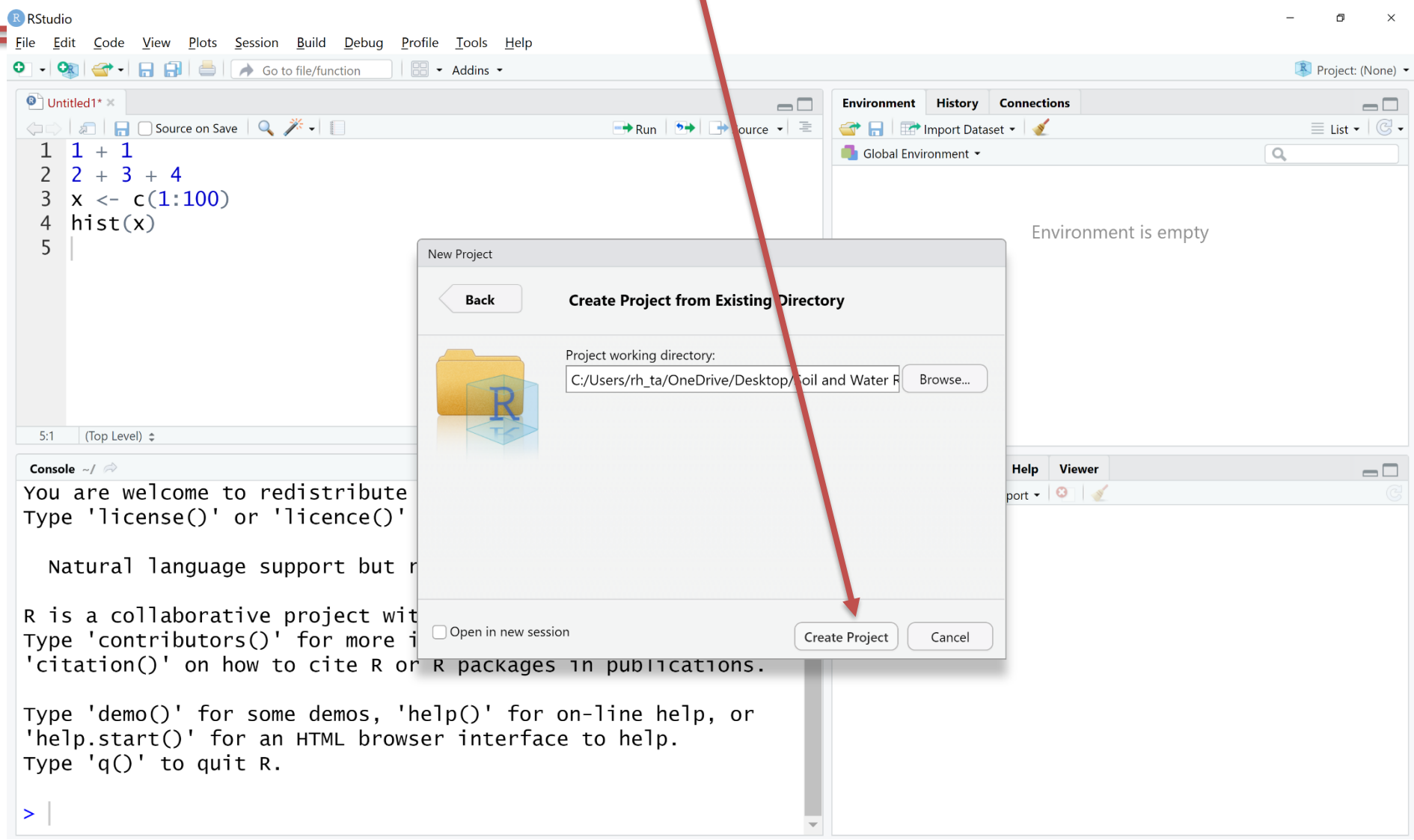
4. Existing Directory



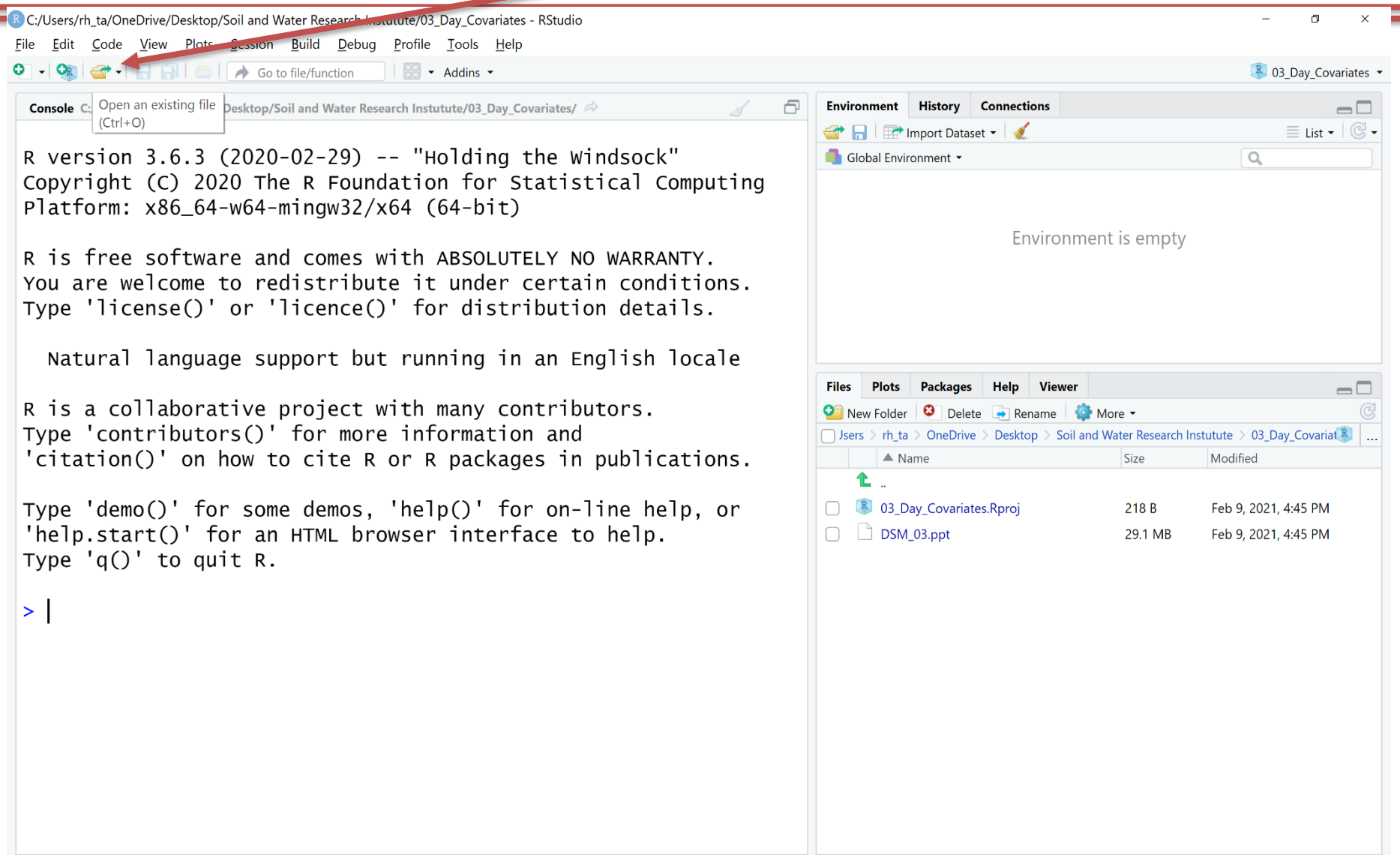
5. Find the folder and find **Day_05_Machine Learning-P2** and open



6. Create project



7. Open an existing file and find **05_R_Cov.R**



8. Open **05_R_Cov.R**