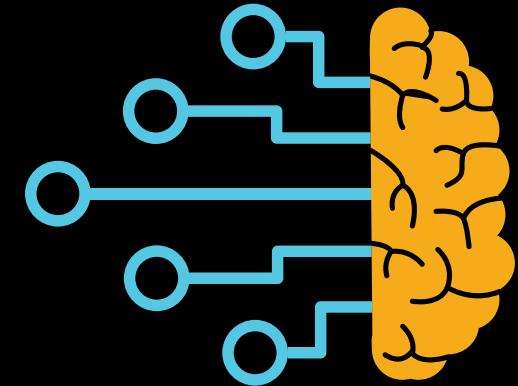


DCAO, UBA  
08-12 AGOSTO 2022

# APRENDIZAJE AUTOMÁTICO.

Fundamentos y Aplicaciones en  
Meteorología del Espacio



---

Dra María Graciela Molina

FACET-UNT / CONICET  
Tucumán Space Weather Center - TSWC

<https://spaceweather.facet.unt.edu.ar/>  
IG -> @spaceweatherargentina

gmolina@herrera.unt.edu.ar

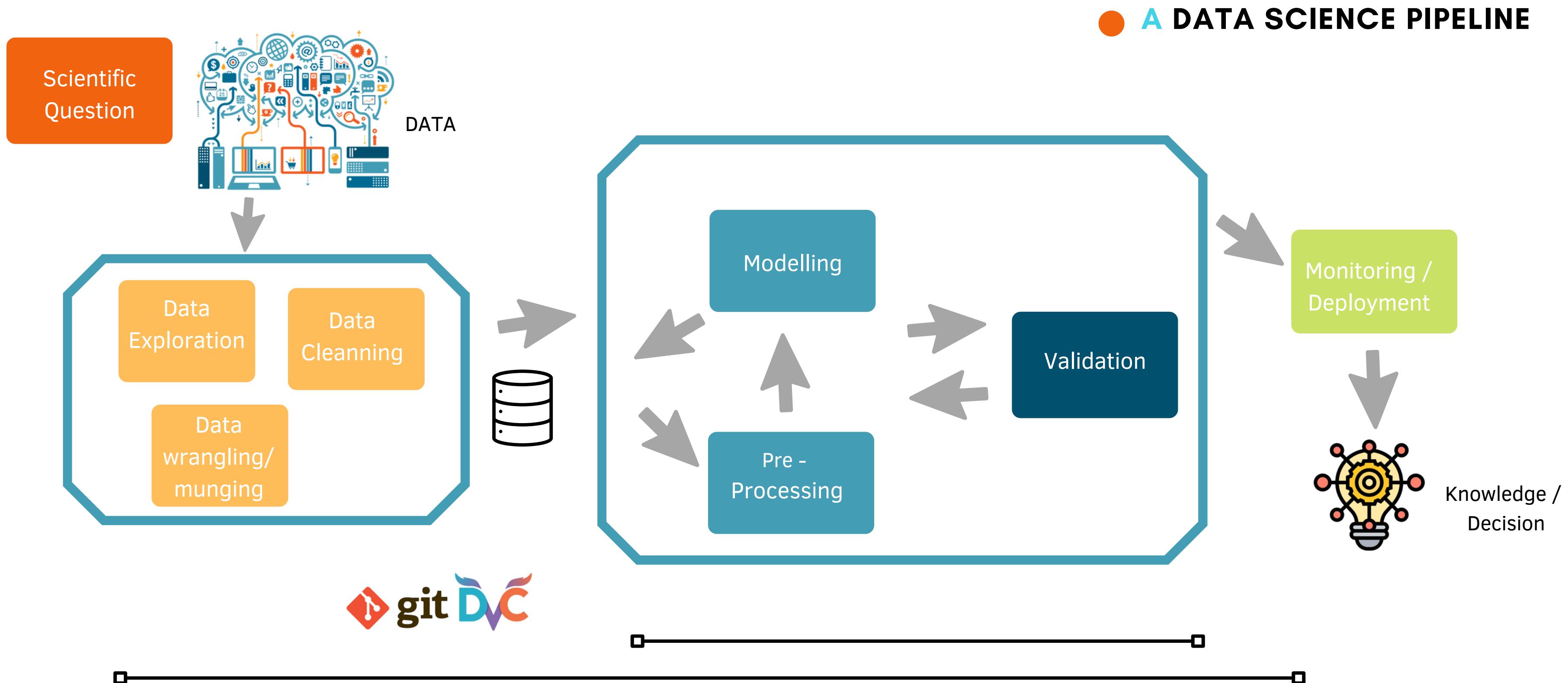
# Machine Learning

TO PROVE YOU'RE A HUMAN,  
CLICK ON ALL THE PHOTOS  
THAT SHOW PLACES YOU  
WOULD RUN FOR SHELTER  
DURING A ROBOT UPRISING.

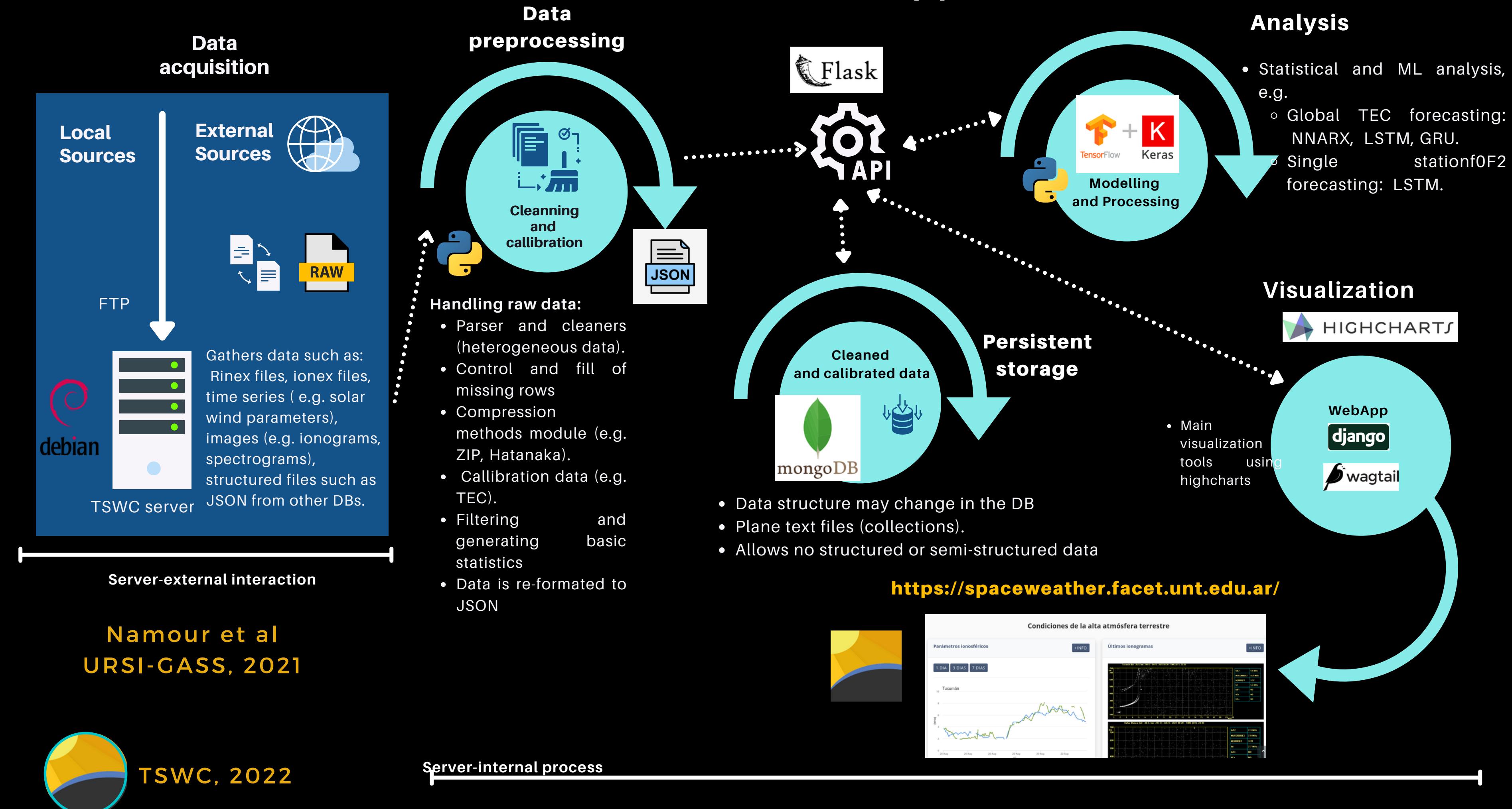


TSWC, 2022

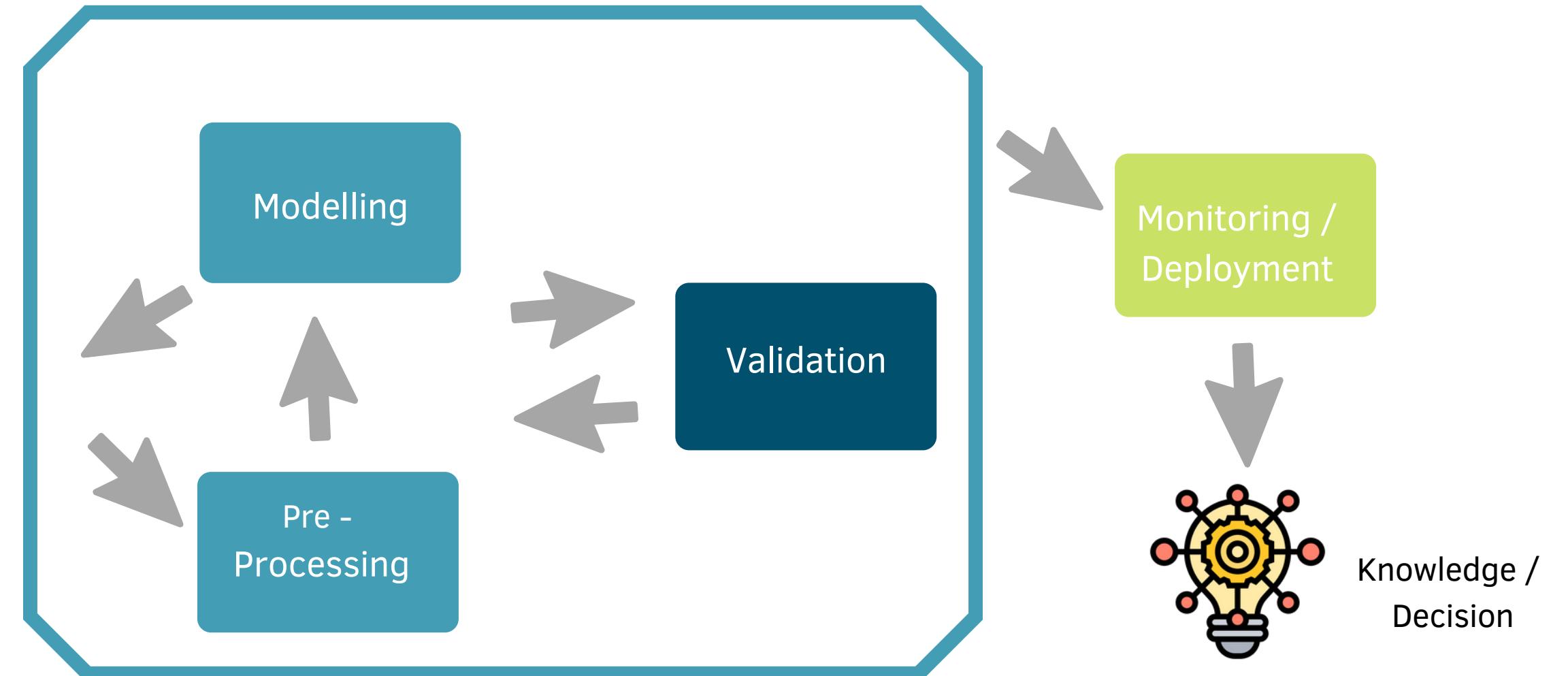
# Data-driven model



# General TSWC pipeline



# ML pipeline



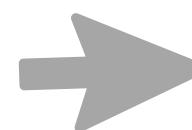
What / how to ask?

How is the answer?

DATA (input)



ML



MODEL  
(program)

Output



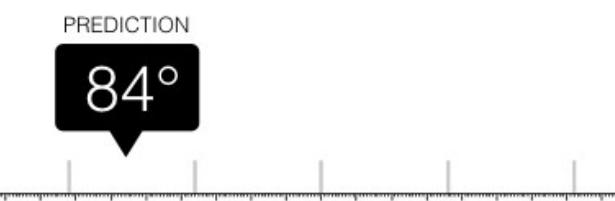
TSWC, 2022

## PROBLEMS (TYPES)



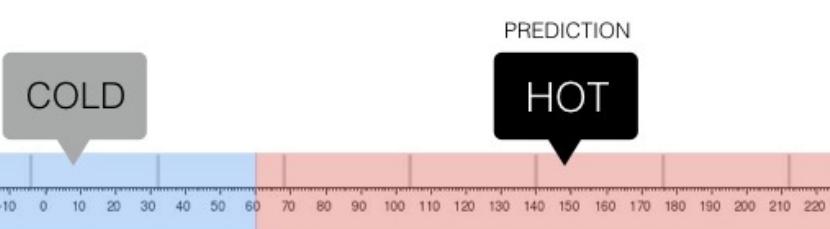
### Regression

What is the temperature going to be tomorrow?



### Classification

Will it be Cold or Hot tomorrow?

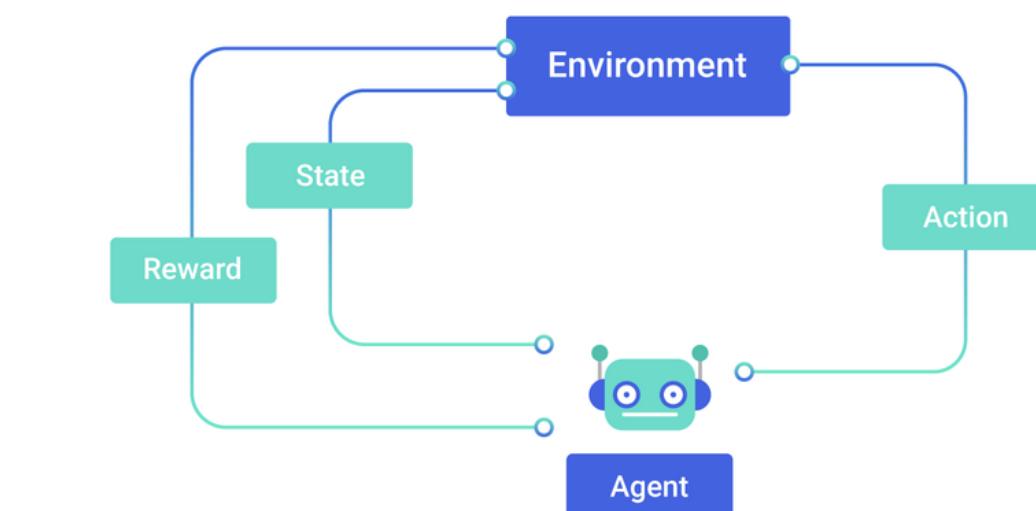
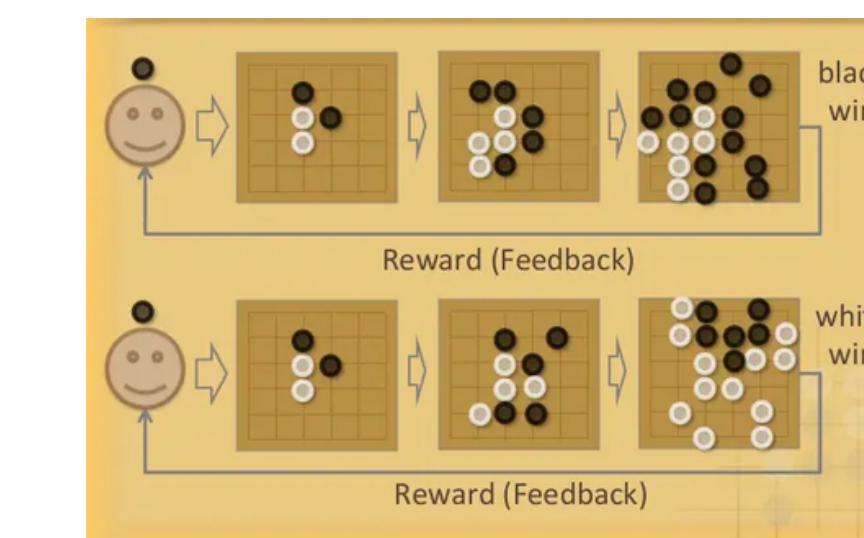
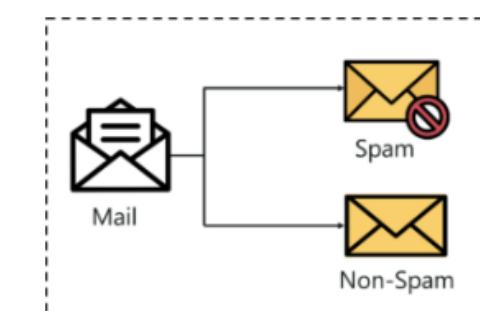
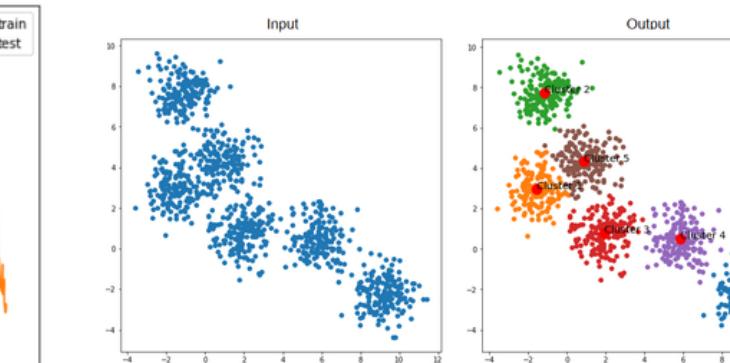
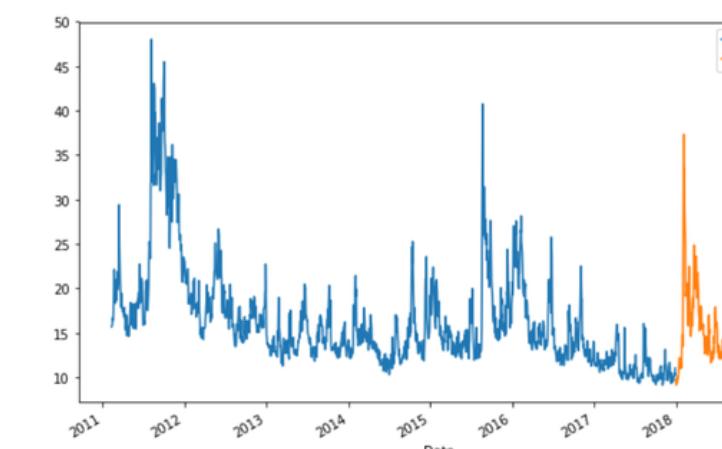


Regression: deals with predicting a continuous value

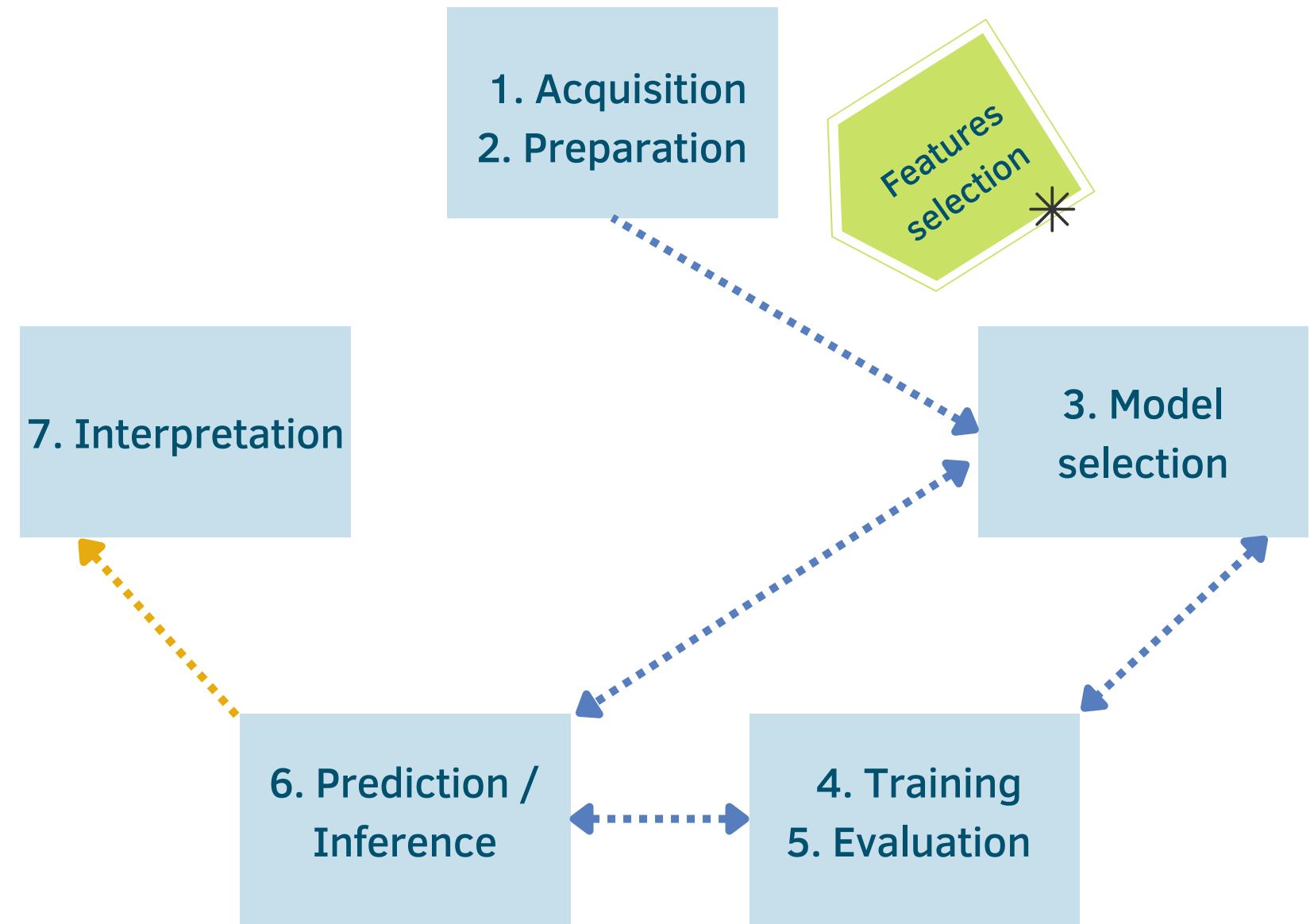
Classification: to predict output from a set of finite categorical values

## LEARNING (TYPES)

- **Supervised:** The target function is known. We have a labelled dataset
- **Non Supervised:** The dataset is not labelled
- **Semi-supervised:** The dataset is partially labelled.
- **Reinforcement learning:** ML system learns from the environment and it corrects itself by penalty or reward.



# ML-based modeling



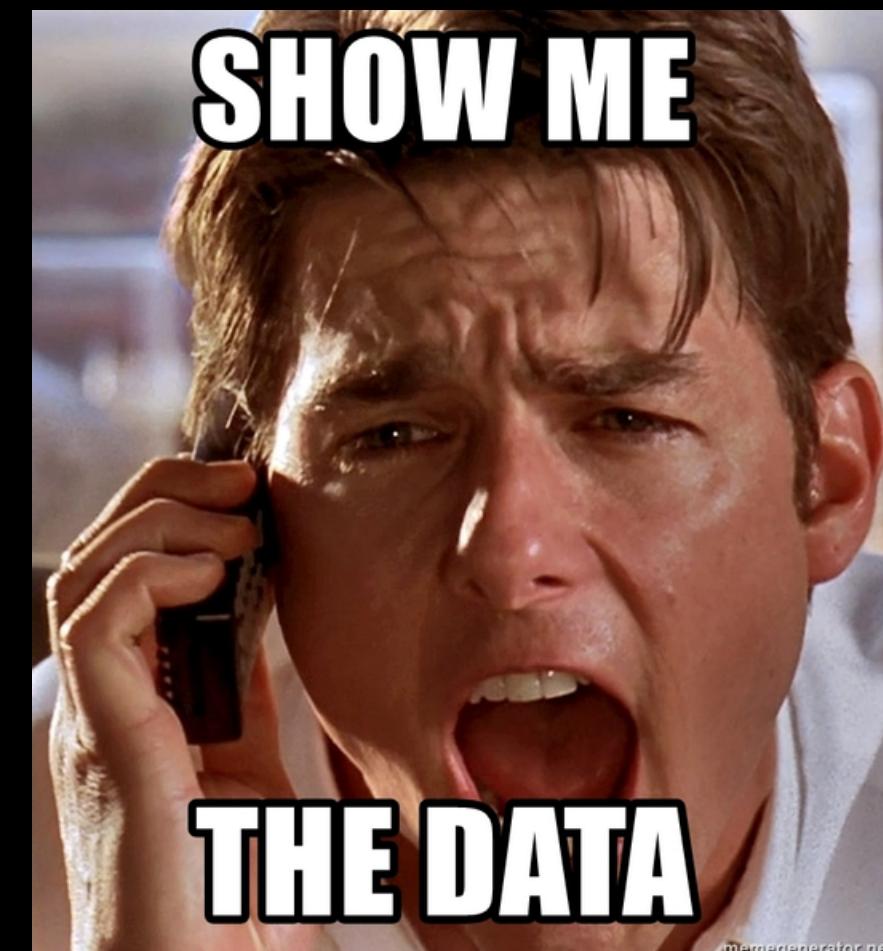
- Acquisition + preparation

- From ML public datasets:

<https://www.kaggle.com/datasets>

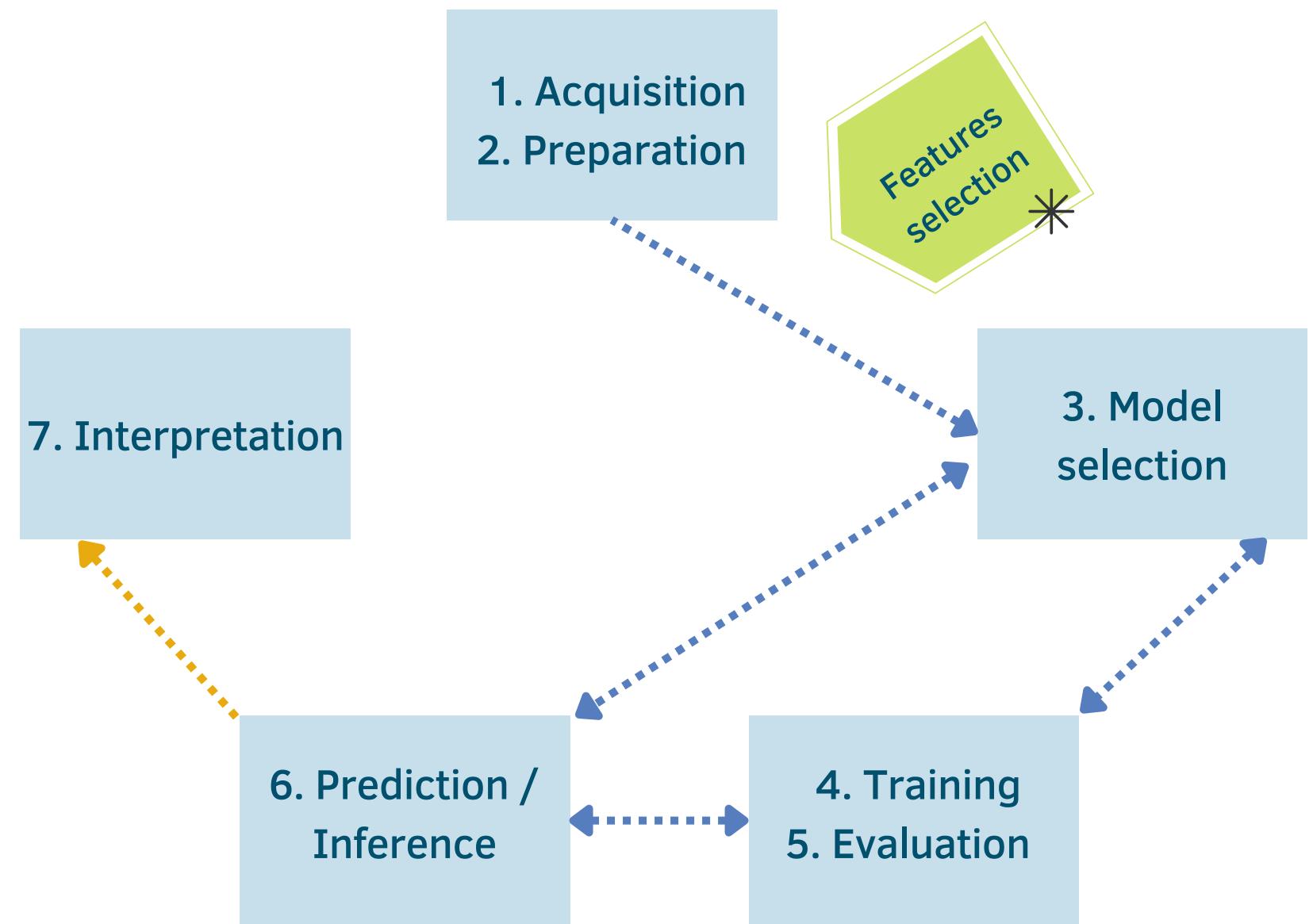
<https://archive.ics.uci.edu/ml/index.php>

<https://github.com/awesomedata/awesome-public-datasets>



- \* • By hand = ML - automatically = DL  
• Take care of dimensionality

# ML-based modeling



- Acquisition + preparation
  - Make our ML datasets:
    - Acquisition from experiments or instruments or - simulations, etc
    - Synthetic data: create datasets

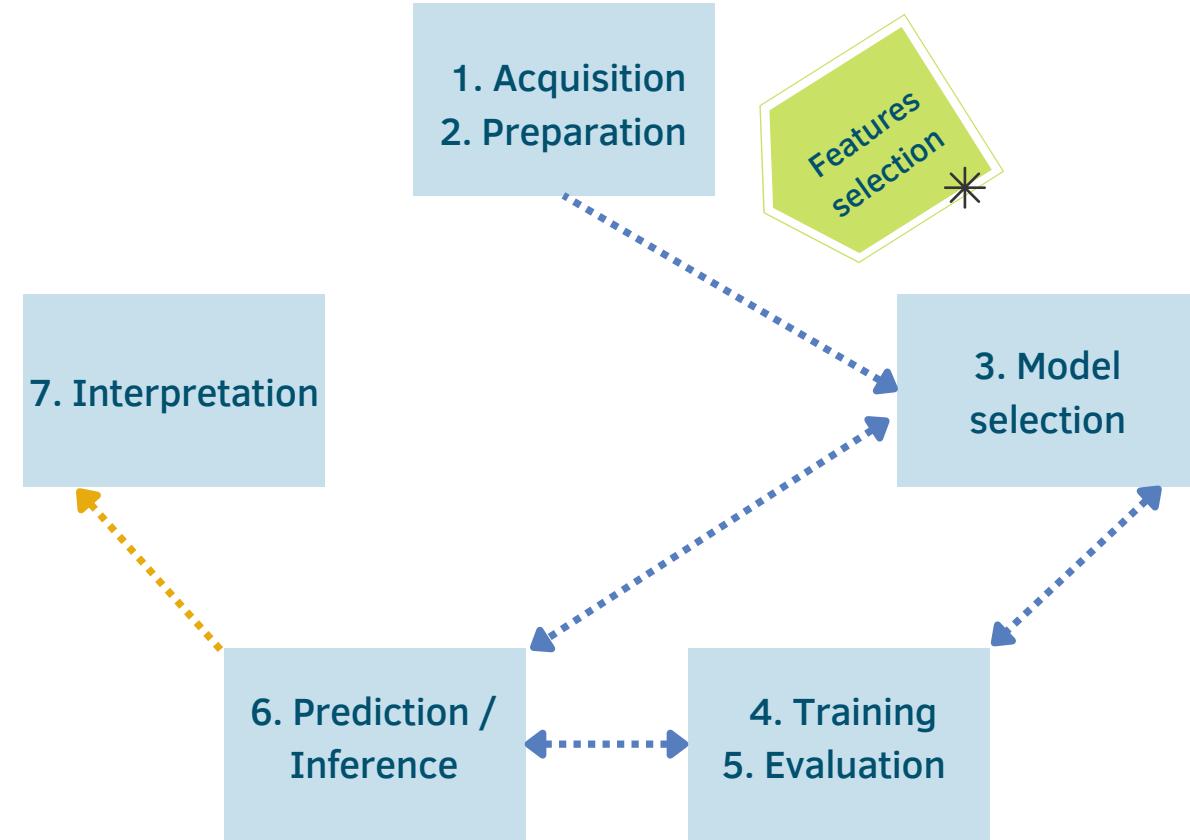
Prepare the dataset: sampling, formatting, metadata, resolution, balance, dataset size, curation, standarization/normalization, integration, storing, etc

## HOW TO CONFUSE MACHINE LEARNING



- \*
  - By hand = ML - automatically = DL
  - Take care of dimensionality

# ML-based modeling



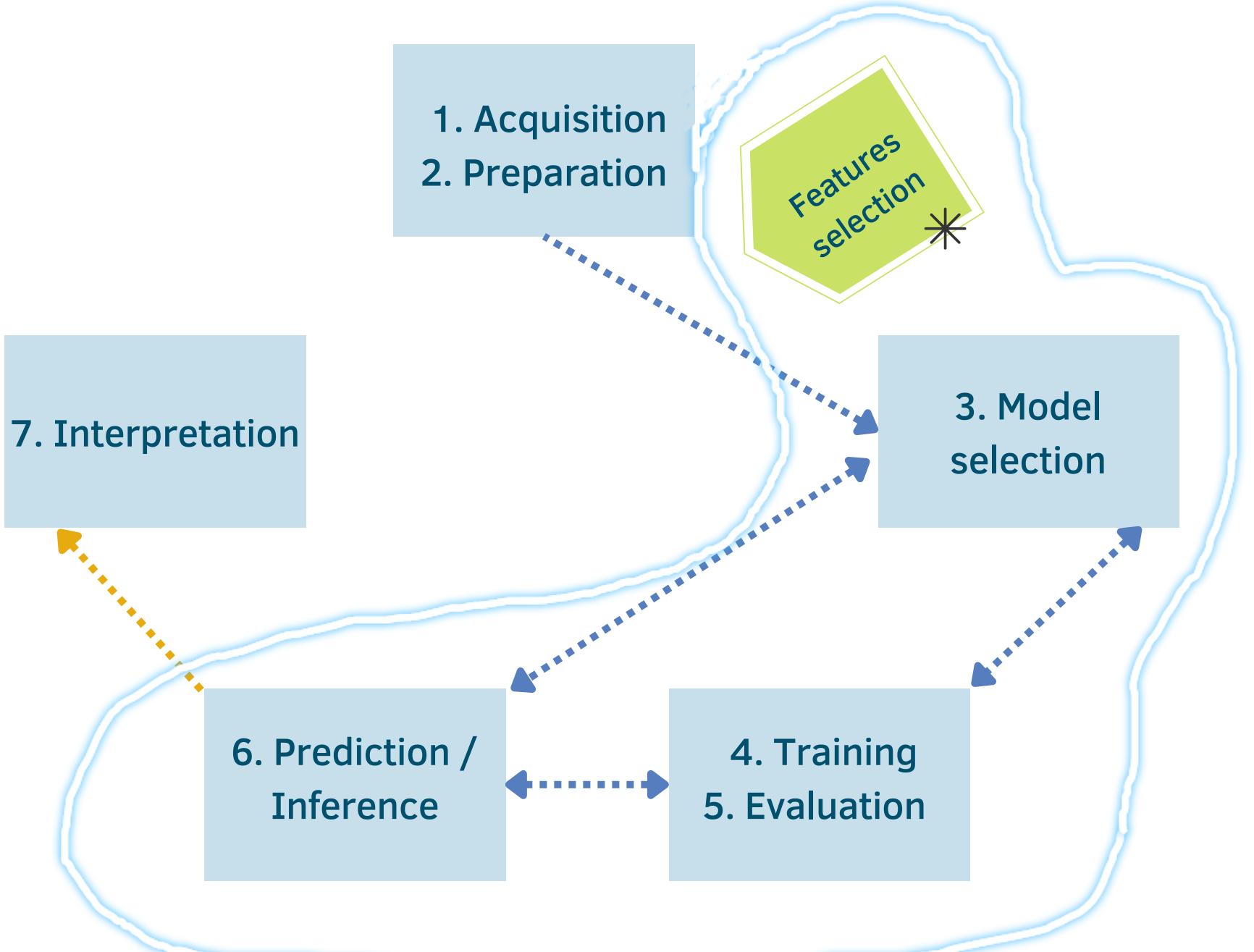
- Modeling

## DATA PREPARATION FOR MODELING

- Basic exploration: basic statistics about the dataset: # samples, # nulls, pdfs, plot, etc
- Balance / unbalanced datasets (the problems, how to solve them)
- Missing values: to input or not to input data!
- Outliers!
- Data transformation = binning, log scales; normalization/standarization.
- Dimensionality reduction if necessary
- Data traceability!

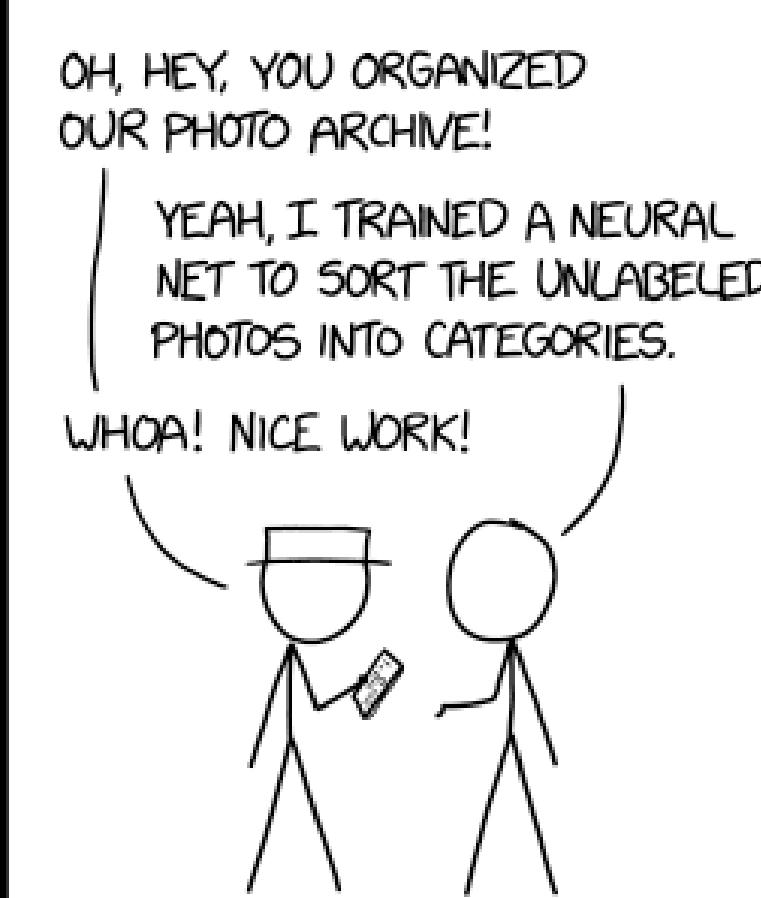


# ML-based modeling



- Modeling

## FEATURE ENGINEERING



ENGINEERING TIP:  
WHEN YOU DO A TASK BY HAND,  
YOU CAN TECHNICALLY SAY YOU  
TRAINED A NEURAL NET TO DO IT.

- Extract/choose relevant features from the dataset
- May include the creation of new features
- Requires experience and domain knowledge
- Feature engineering by hand: hard, slow, not robust, not scalable
- Explicit or/and implicit. (lagged features, statistics, data transformation )
- DL = automatic feature engineering



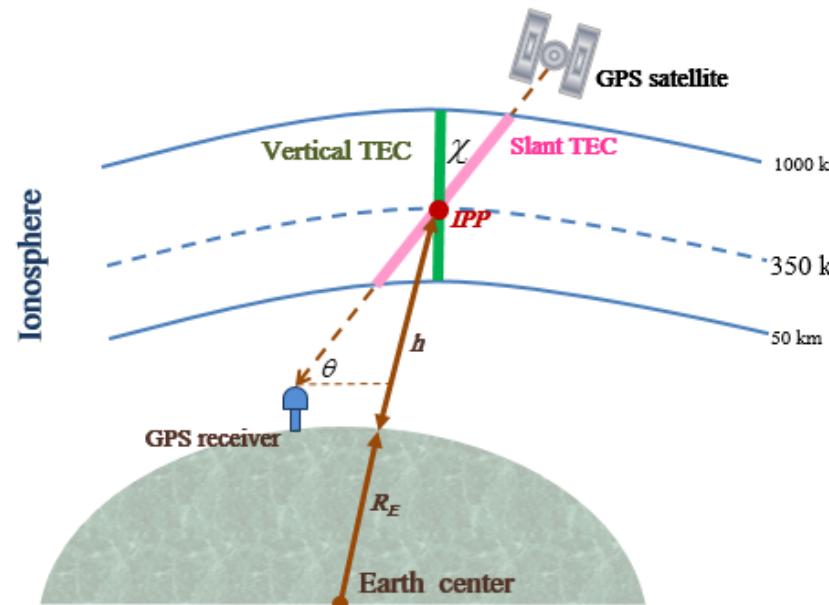
# ML-based modeling

## Example

- Let's forecast the ionospheric conditions! how hard could it be? Let's select the features!

Assumptions:

- single station modeling
- TEC derived from GNSS (using "some" calibration technique and "some" constellation)
- It is a time series

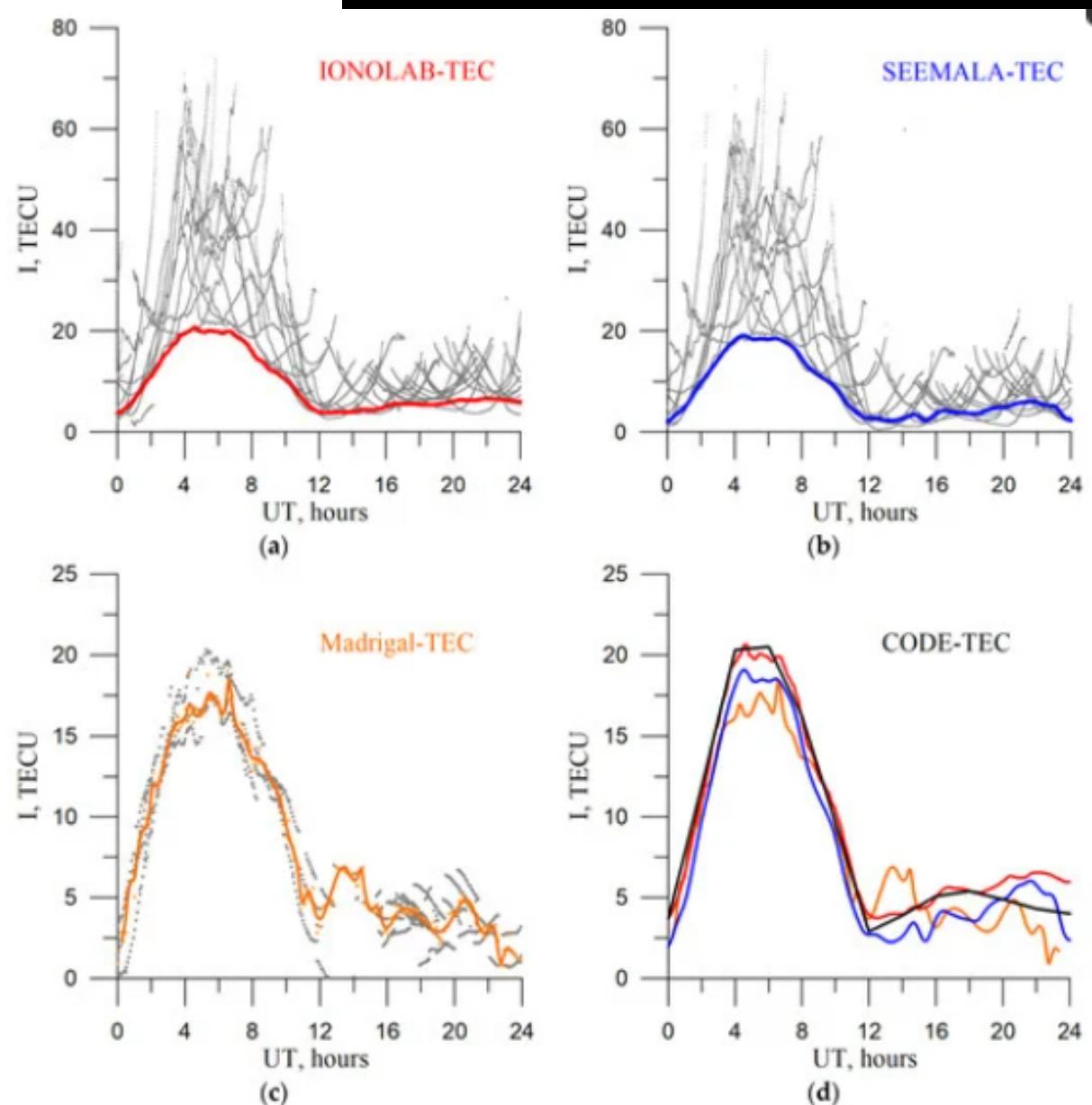


- Which is/are the feature/s?

- Modeling

## FEATURE ENGINEERING

- Extract/choose relevant features from the dataset
- May include the creation of new features
- Requires experience and domain knowledge
- Feature engineering by hand: hard, slow, not robust, not scalable
- Explicit or/and implicit. (lagged features, statistics, data transformation )
- DL = automatic feature engineering



# Parentesis (el medio bajo estudio)

Ionosfera: porción de la alta atmósfera terrestre donde tanto iones como electrones se encuentran presentes en cantidades suficientes como para afectar las ondas de radio (IEEE Std 211-1997).

Se caracteriza por la alta densidad de electrones e iones libres producidos principalmente por foto-ionización de rayos UV y X que arrivan desde el sol (y minoritariamente en altas latitudes por ionización corpuscular)

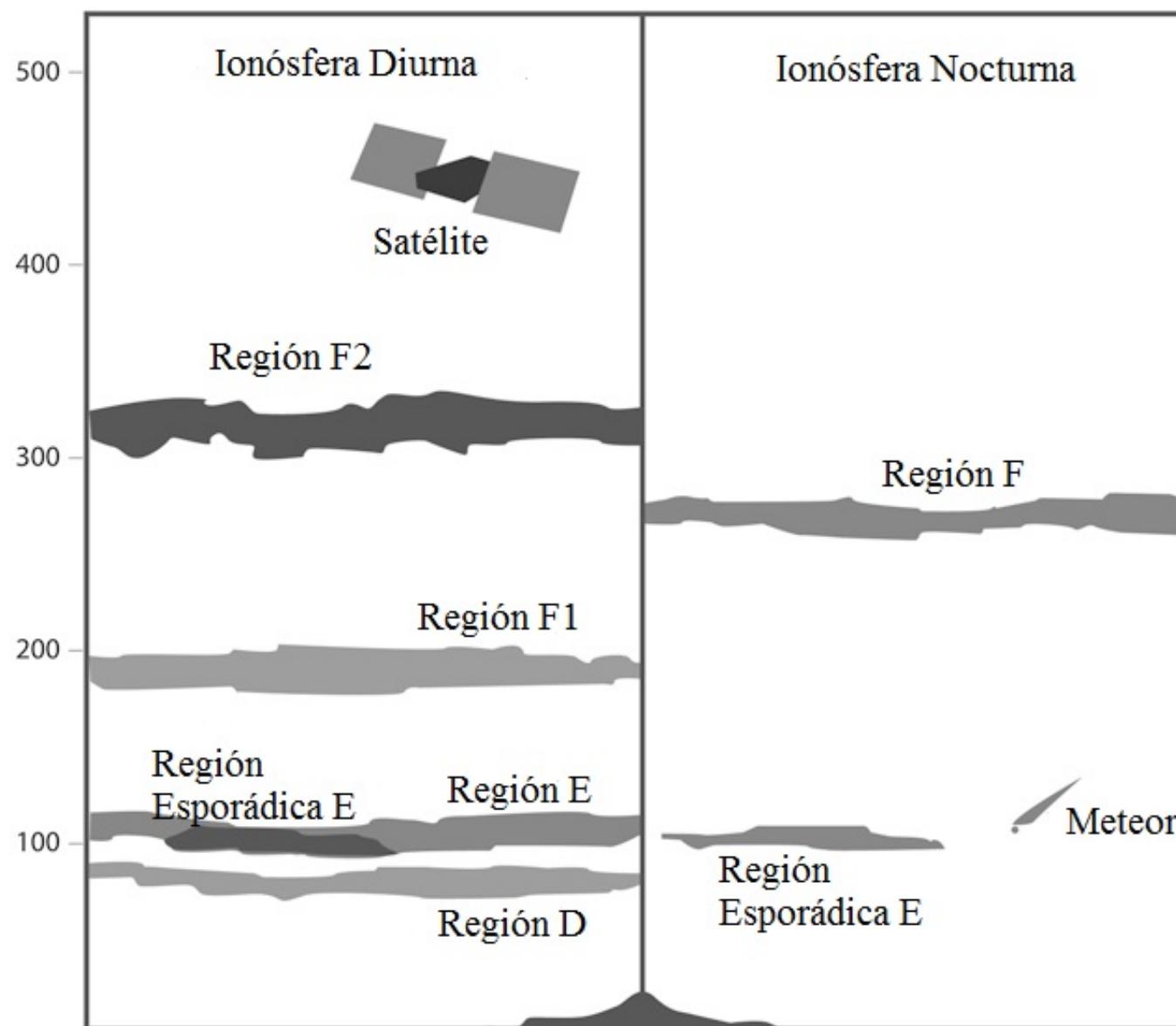
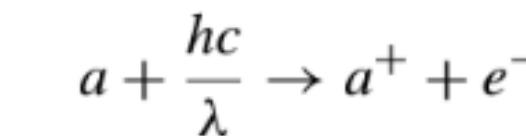
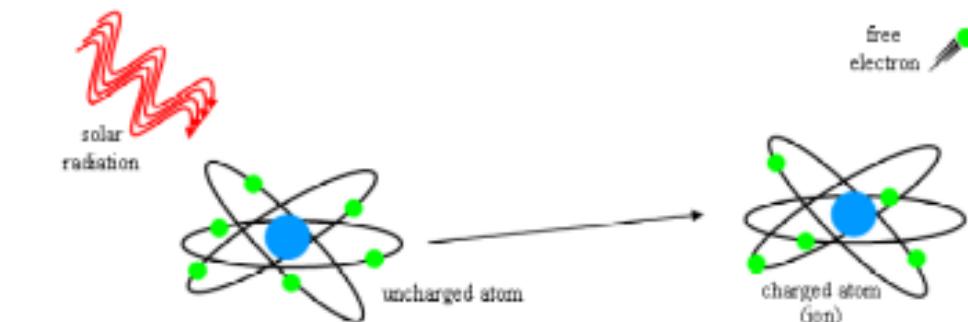


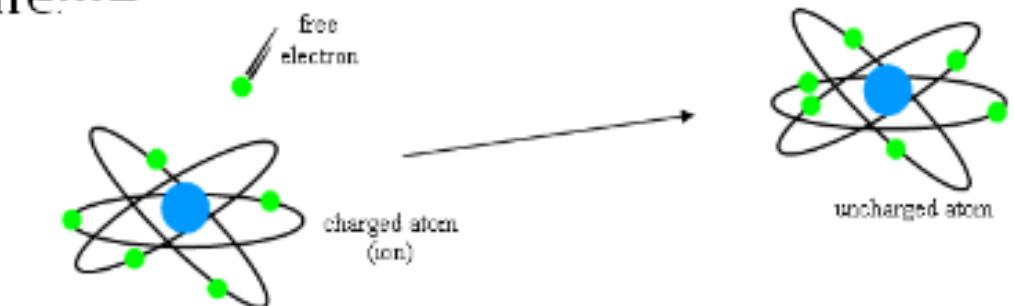
Foto-ionización (producción de iones y electrones libres)



$h =$ Plank const  $6.62 \times 10^{-34}$  Js,  $c =$  vel. Luz,  $\lambda =$  long de onda incidente



- Recombinación: fenómeno inverso. Los electrones libres se combinan con iones positivos para producir átomos neutros nuevamente



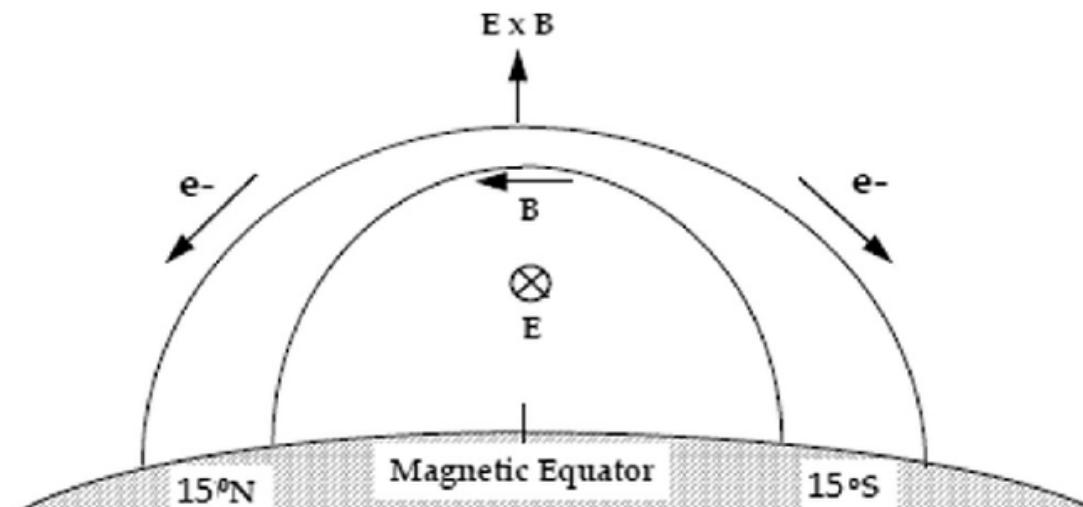
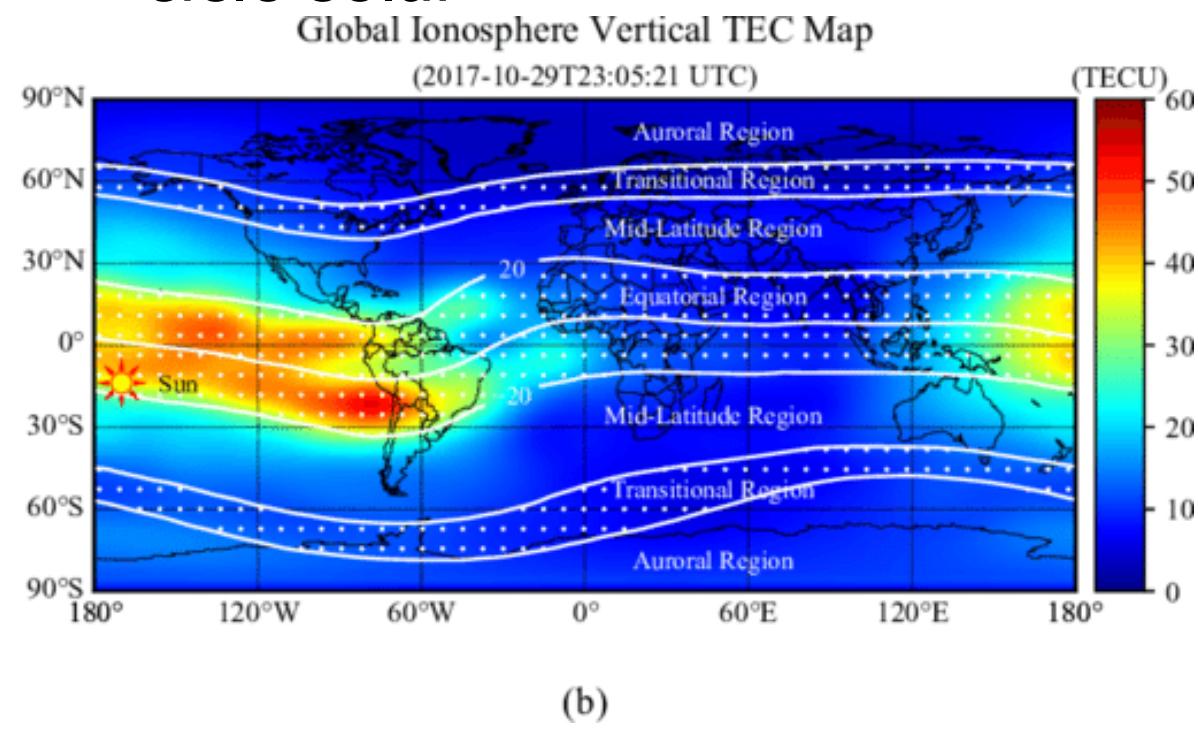
$dN/dt =$  varición densidad electronica en fn del tpo (unidad de vol.),  $q =$  producción de electrones,  $l =$  pérdida de electrones,  $d =$  factor (pérdida por difusión, vientos neutros, drift electromag. vertical)

$$\frac{dN}{dt} = q - l + d$$

# Parentesis (el medio bajo estudio)

## Variabilidad regular

- diaria
- estacional
- geográfica/geomagnética
- ciclo solar



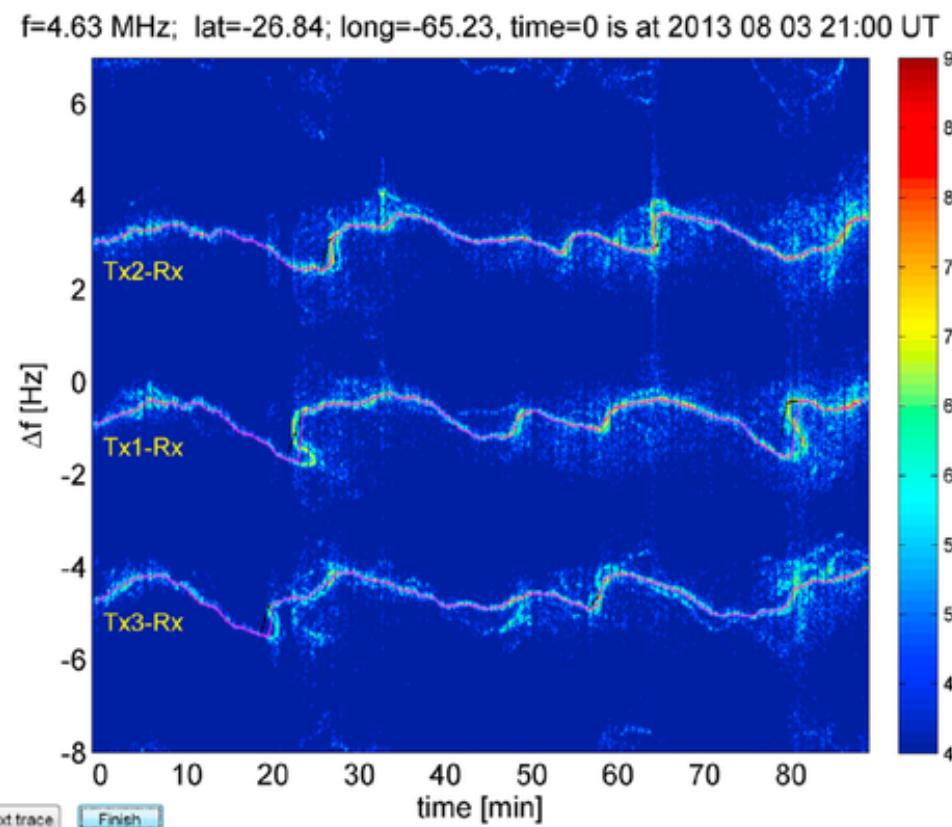
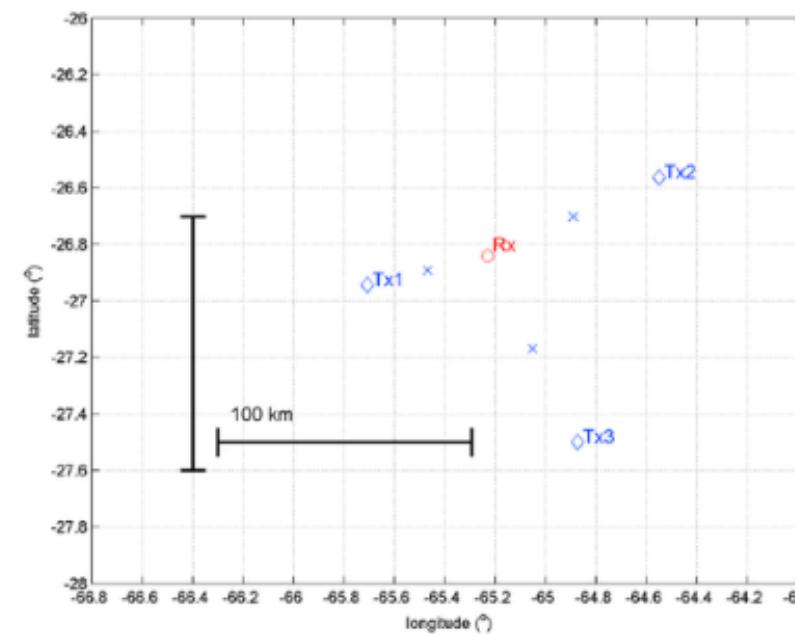
## Irregularidades

### Travelling ionospheric disturbances (TIDs)

Irregularidades de la región F expresadas como oscilaciones similares a ondas del contorno de la densidad electrónica que va descendiendo lentamente con el tiempo

Se clasifican en : **large-scale TIDs (LS-TIDs)** y **medium-scale TIDs (MS-TIDs)**.

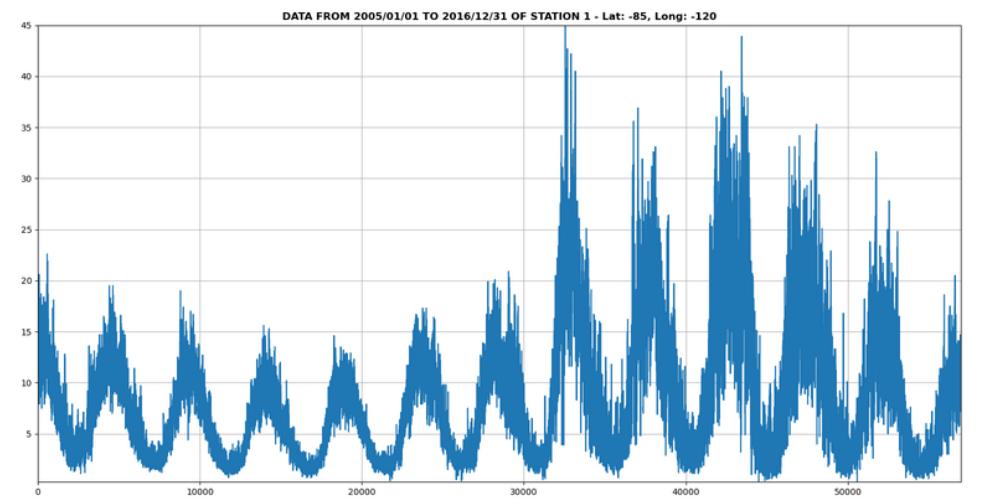
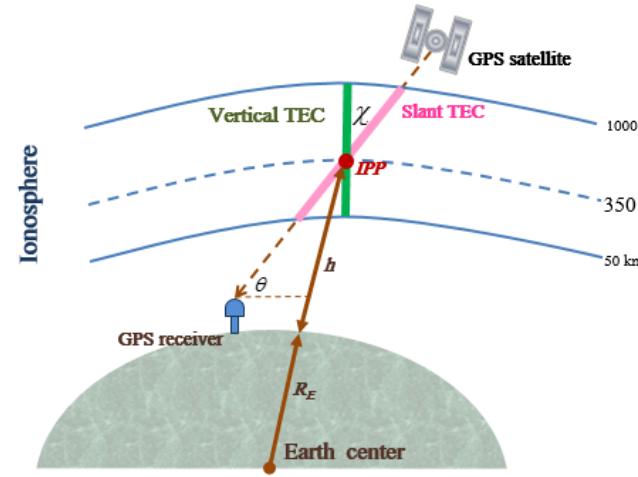
- LS-TIDs relacionadas con AGWs causadas por acoplamiento VS-M-I en zonas polares que se transportan a otras regiones (Space Weather)
- MS-TIDs: periodos de tiempo más cortos, se mueven más lento y generalmente están relacionadas a fenómenos como vientos neutros y al terminador solar que generan AGWs a alturas ionosféricas



● TSWC-CAS

# ML-based modeling

- Which is/are the feature/s?



- TEC -> time (sequence -order)
- what about shifted TEC series? e.g. 24 his shift? seasonal shift? other lags?
- Latitudinal characteristics (how to add?) - how important is this? How about irregularities? (scales!!!)
- TEC is a model! how much impact will have if we approximate TEC with different calibration techniques or the number of PRN/constellations? Which we should choose? is it relevant to add different time series (for different techniques or constellations?)

$$STEC = \int_s N_e ds \text{ (TECU)}$$

- and this is for a single station!

- Modeling

## FEATURE ENGINEERING

- Extract/choose relevant features from the dataset
- May include the creation of new features
- Requires experience and domain knowledge
- Feature engineering by hand: hard, slow, not robust, not scalable
- Explicit or/and implicit. (lagged features, statistics, data transformation )
- DL = automatic feature engineering

How many is too many?



TSWC, 2022

# ML-based modeling



## FEATURES (Selection)

- What if one of the features is strongly related to another feature? how to know they are related?
- How many features (dimensions) do the ML needs to get a satisfactory model?

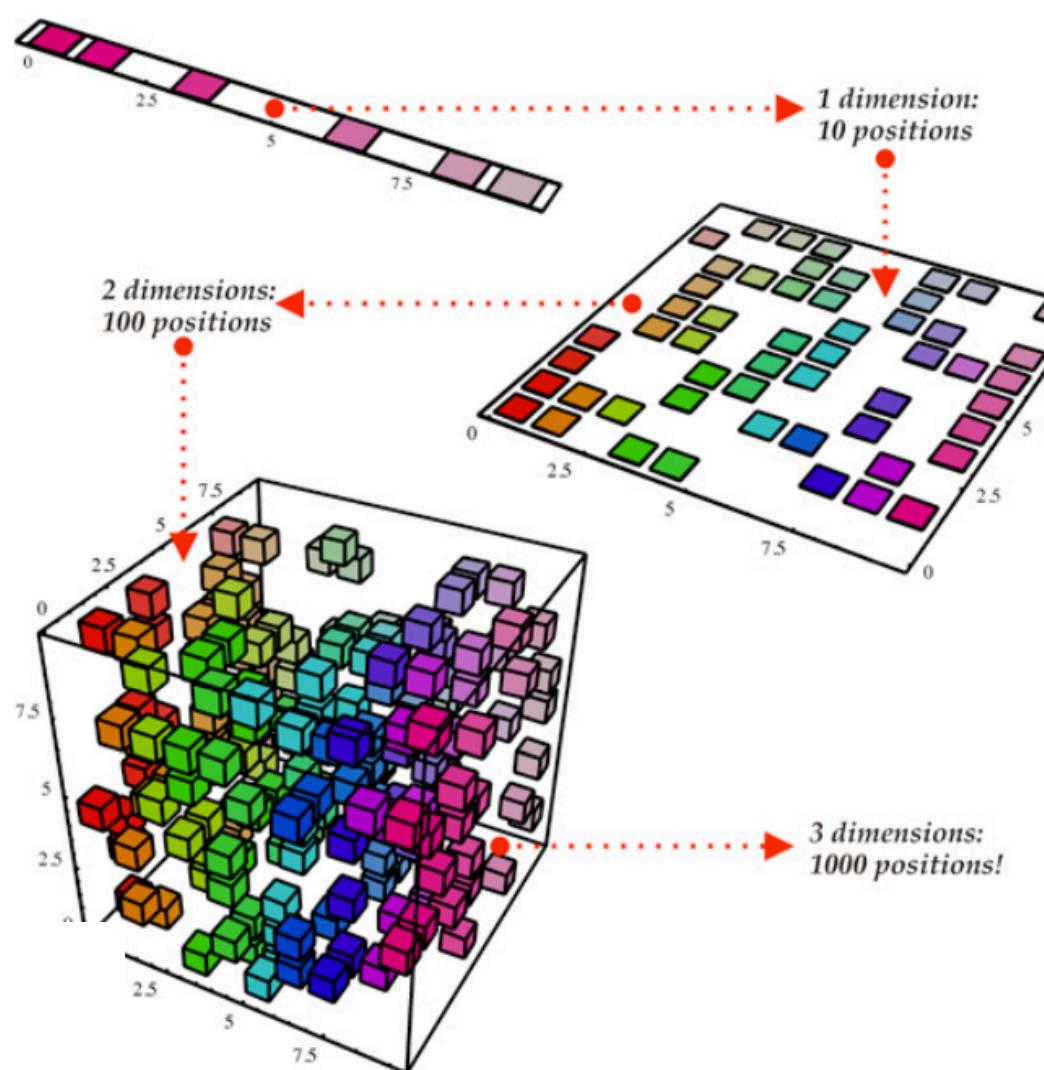
[nature](#) > [nature methods](#) > [this month](#) > [article](#)

This Month | [Published: 31 May 2018](#)

POINTS OF SIGNIFICANCE

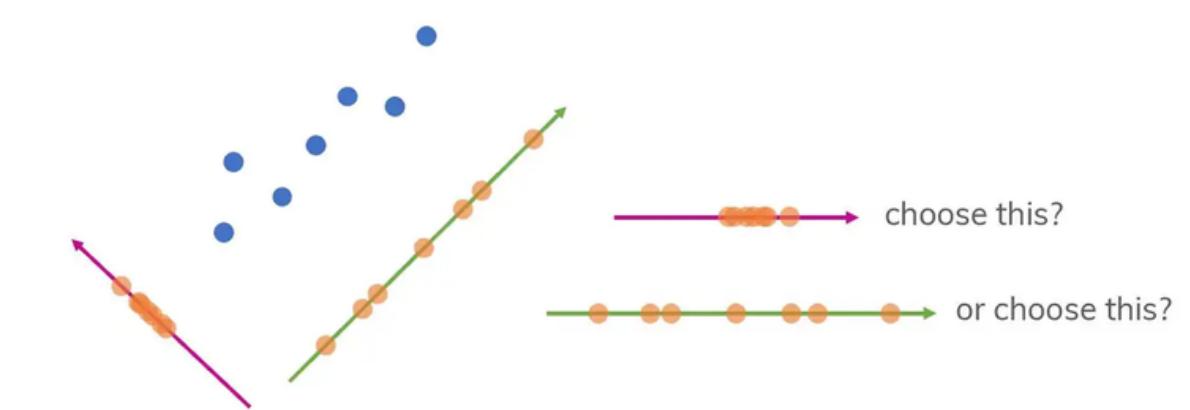
## The curse(s) of dimensionality

[Naomi Altman](#) & [Martin Krzywinski](#)

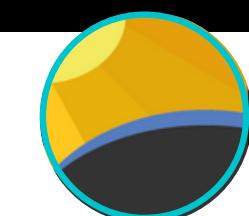


+ features + data + computing effort + sparse data

Distance concentration: many ML algorithms are based on the calculation of distances. The distance between 2 points tends to grow as the dimension grows.

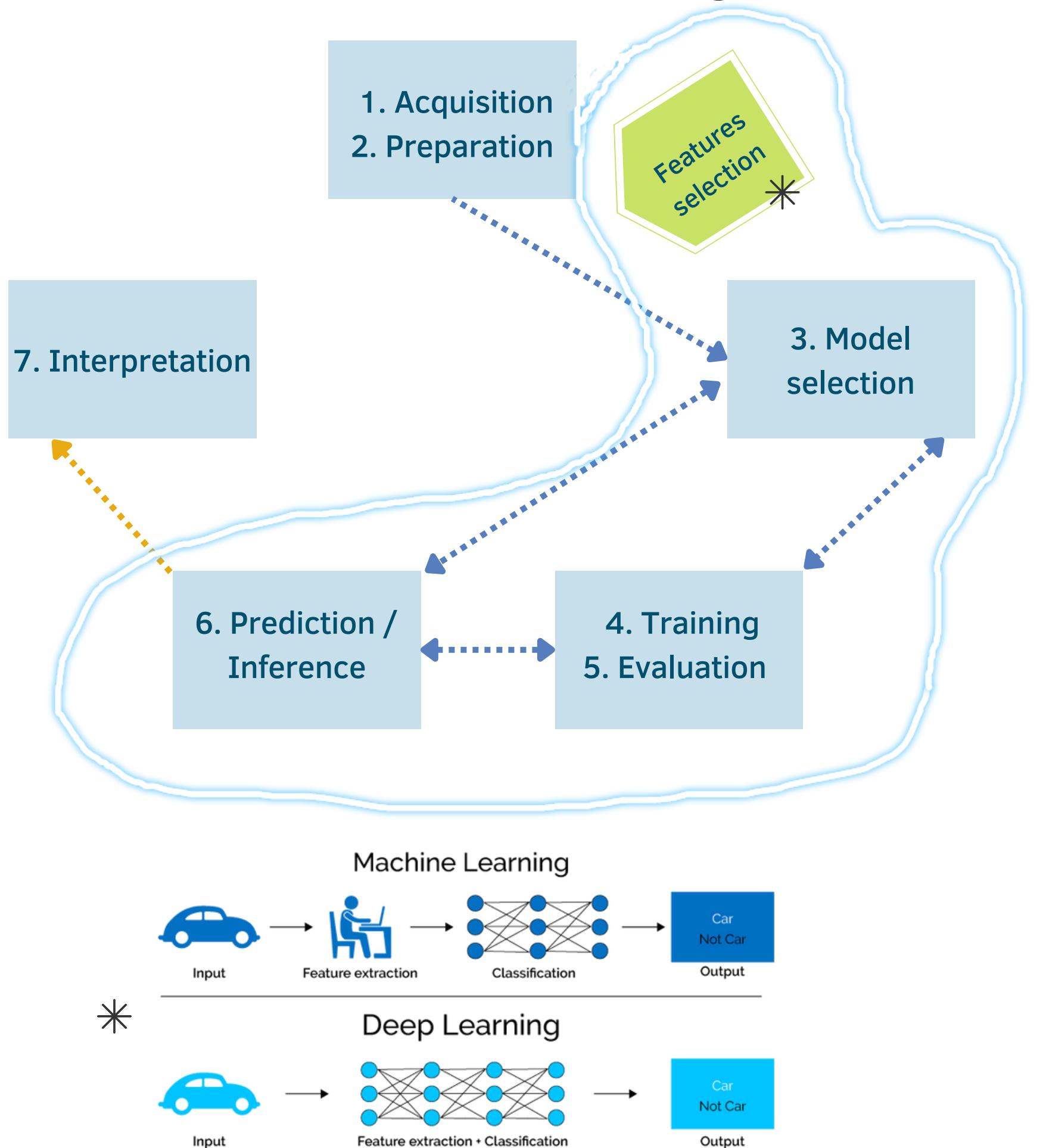


<https://www.nature.com/articles/s41592-018-0019-x>



TSWC, 2022

# ML-based modeling

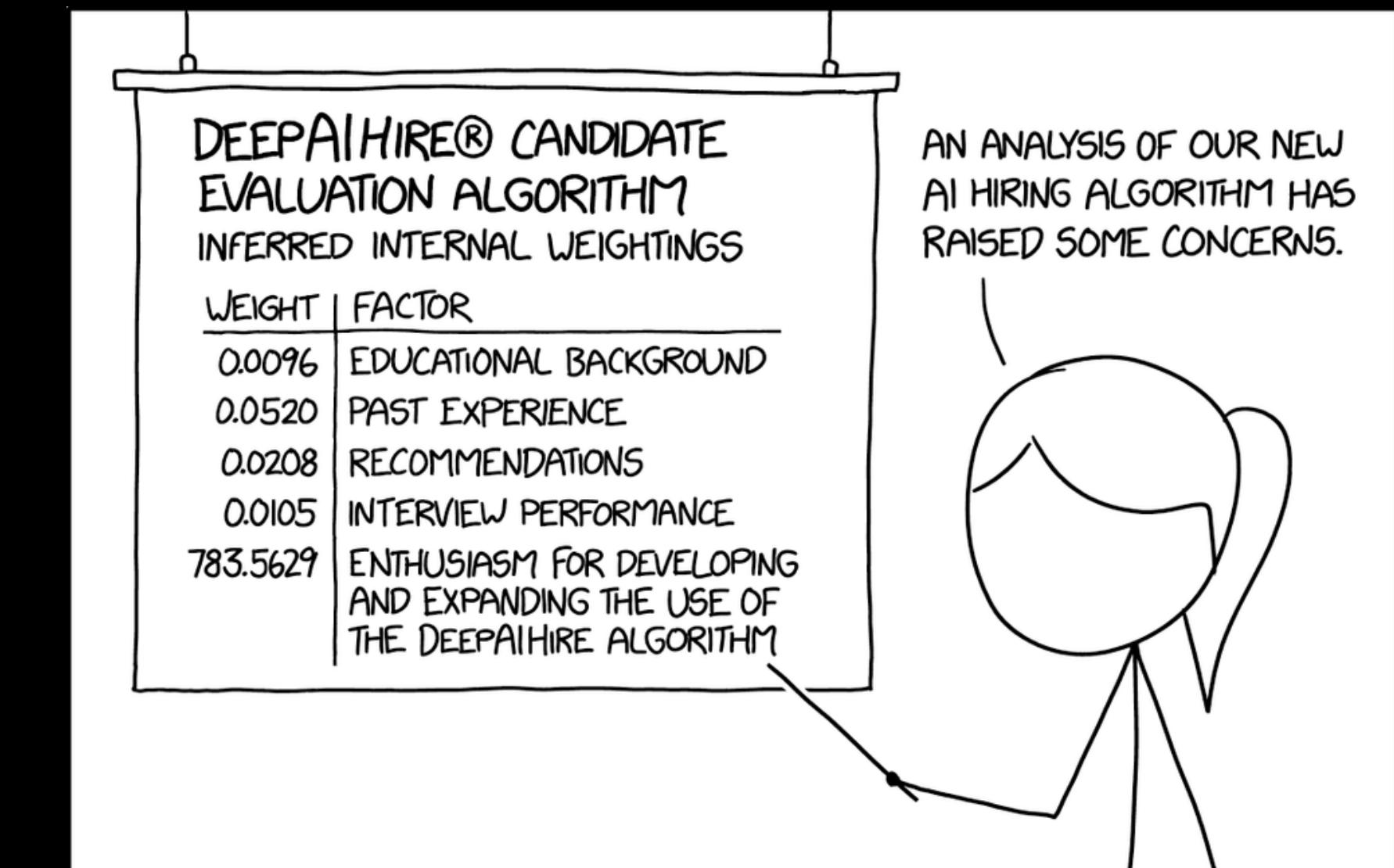


- Modeling

in detail in this course!

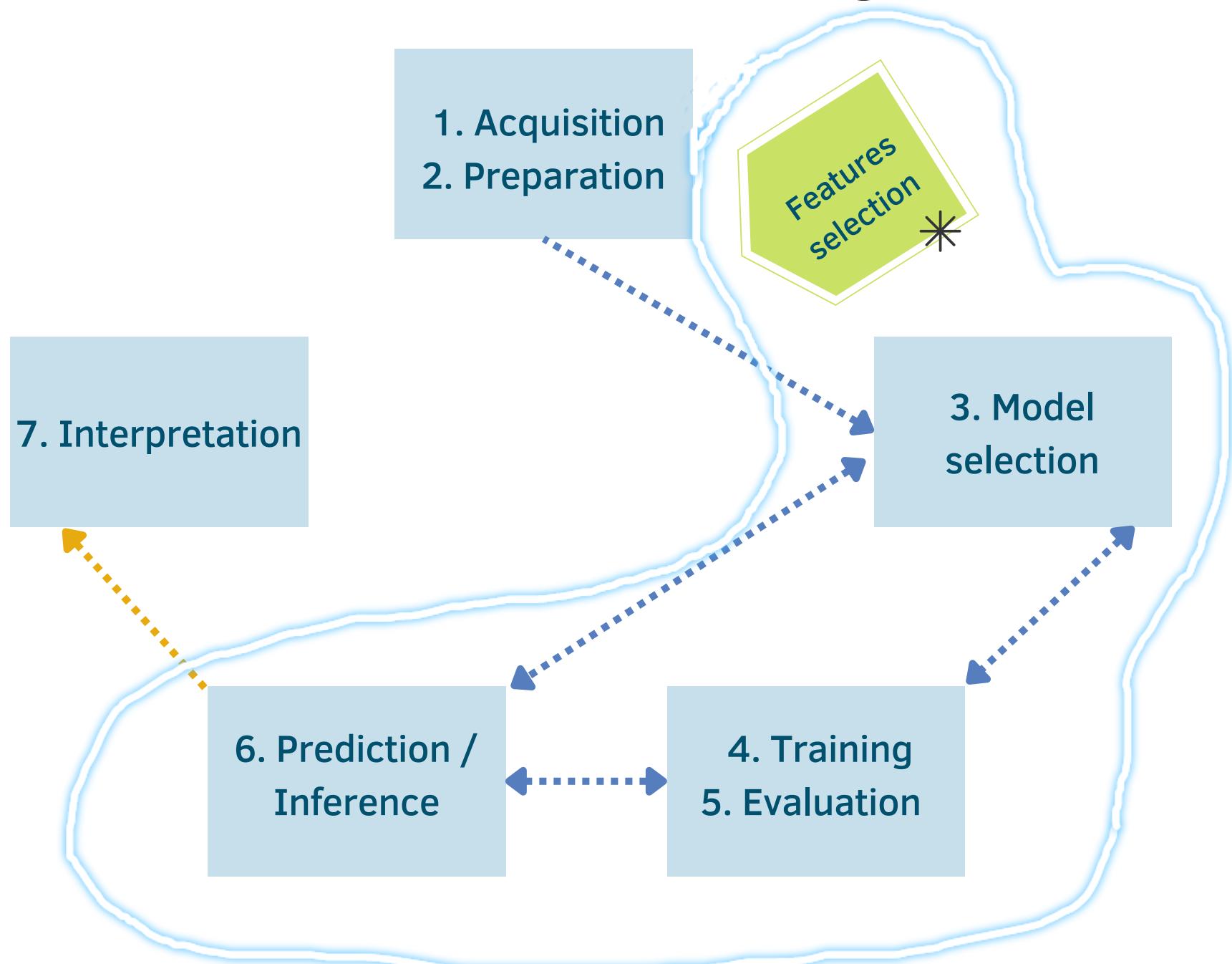
Iterative steps

- MODEL SELECTION
- TRAINING
- EVALUATION
- PREDICTION



TSWC, 2022

# ML-based modeling



- Modeling

## INTERPRETATION

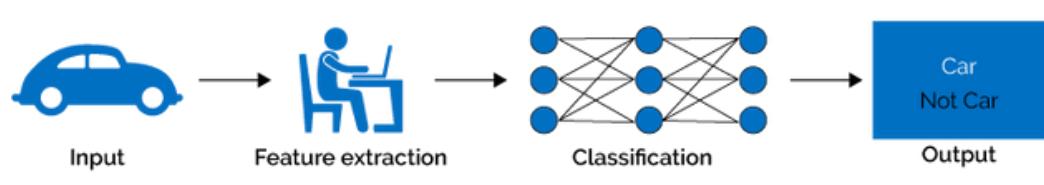
| XAI Goal          |
|-------------------|
| Trustworthiness   |
| Causality         |
| Transferability   |
| Informativeness   |
| Confidence        |
| Fairness          |
| Accessibility     |
| Interactivity     |
| Privacy awareness |

- **interpretability:** transparency.
- **explainability:** action or procedure taken by a model with the intent of clarifying or detailing its internal functions.

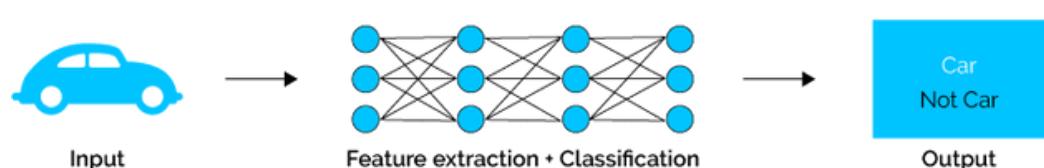
## eXplainable AI (XAI): ML techniques

- Explainable models while maintaining a high level of learning performance (e.g., prediction accuracy)
- Understand, trust, manage AI

### Machine Learning



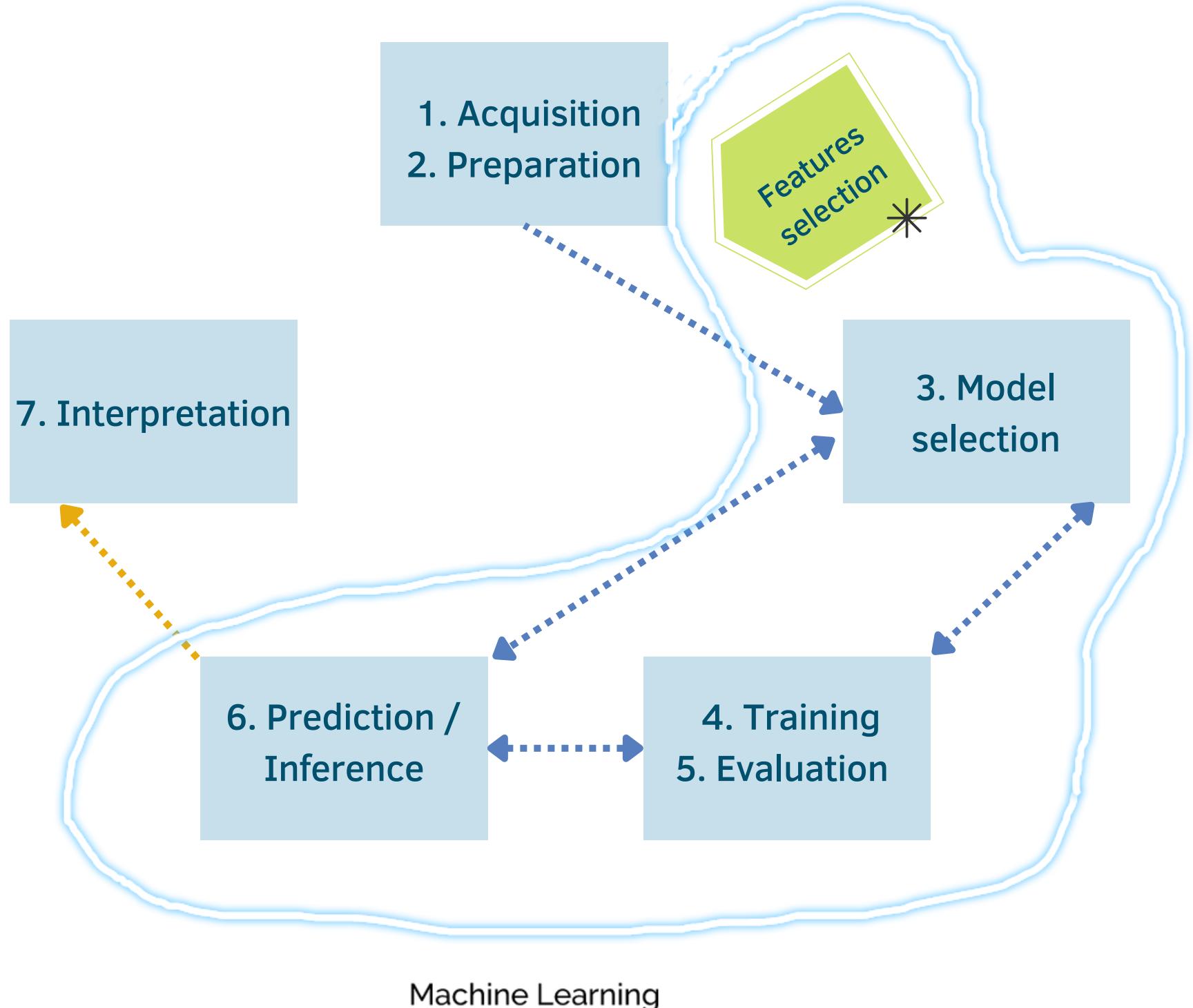
### Deep Learning



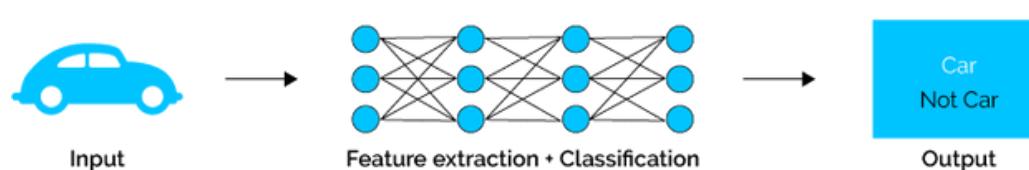
Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI  
Barredo Arrieta et.al. (2019)



# ML-based modeling

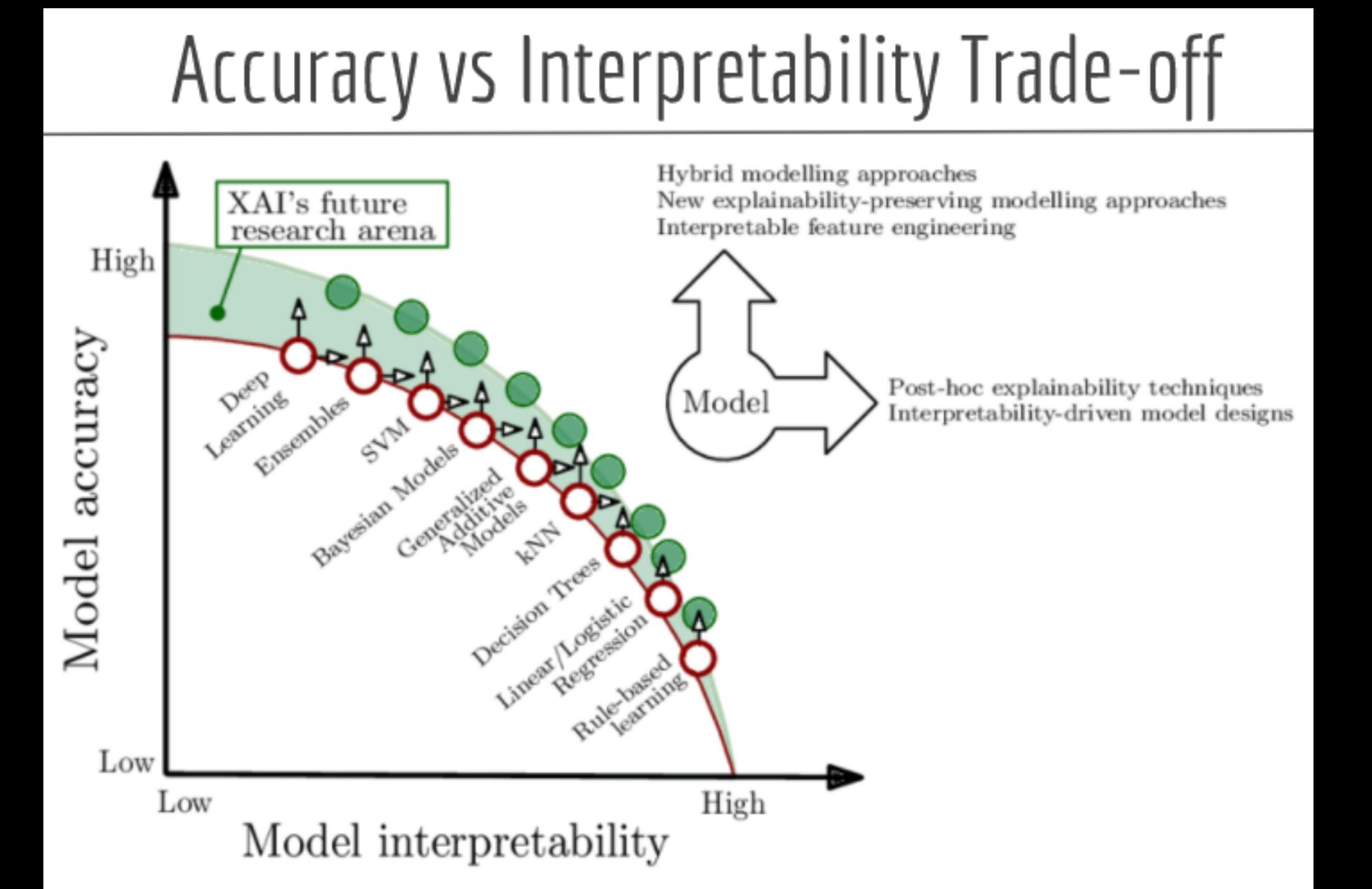


## Deep Learning



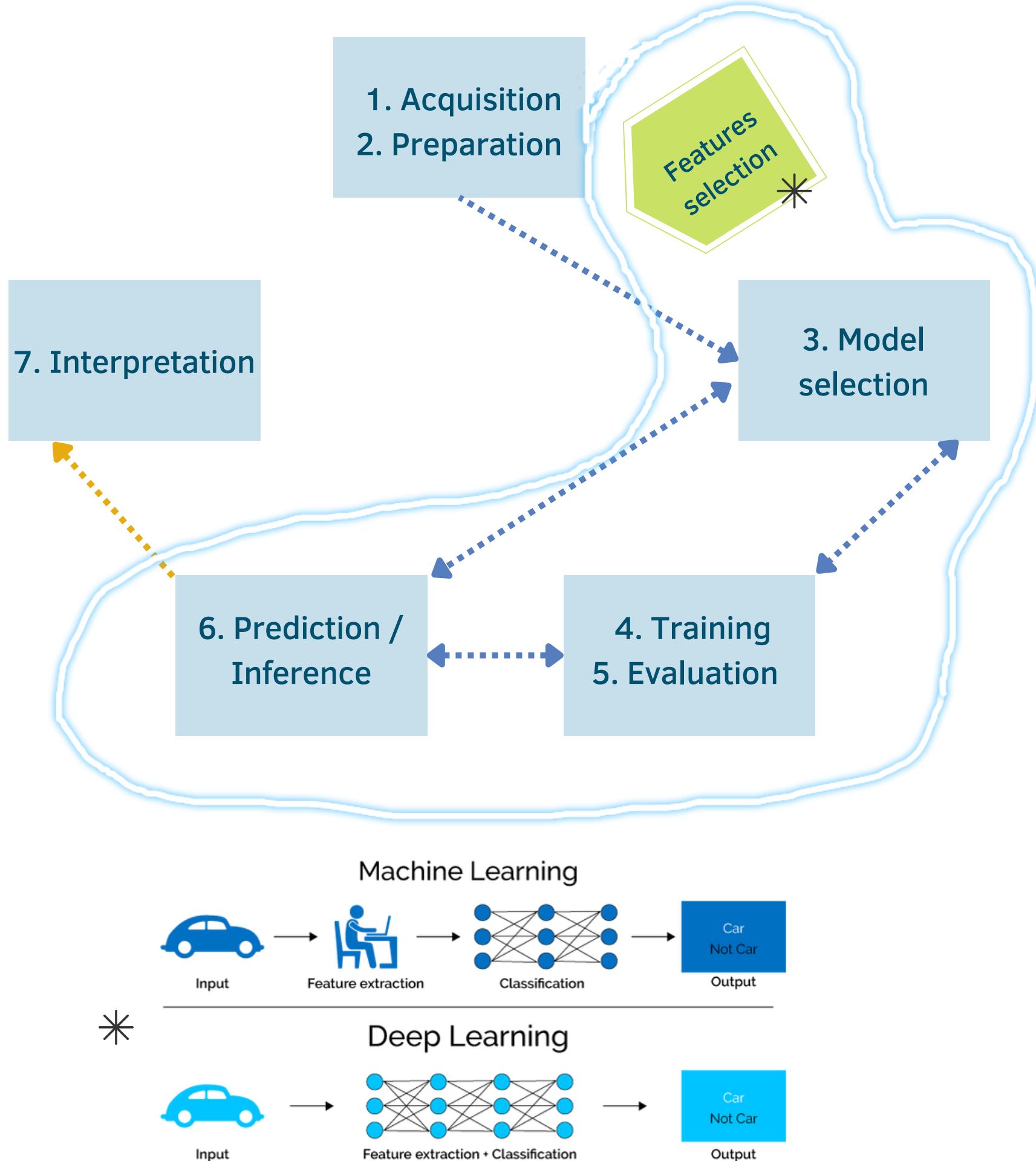
- Modeling

## INTERPRETATION



TSWC, 2022

# ML-based modeling

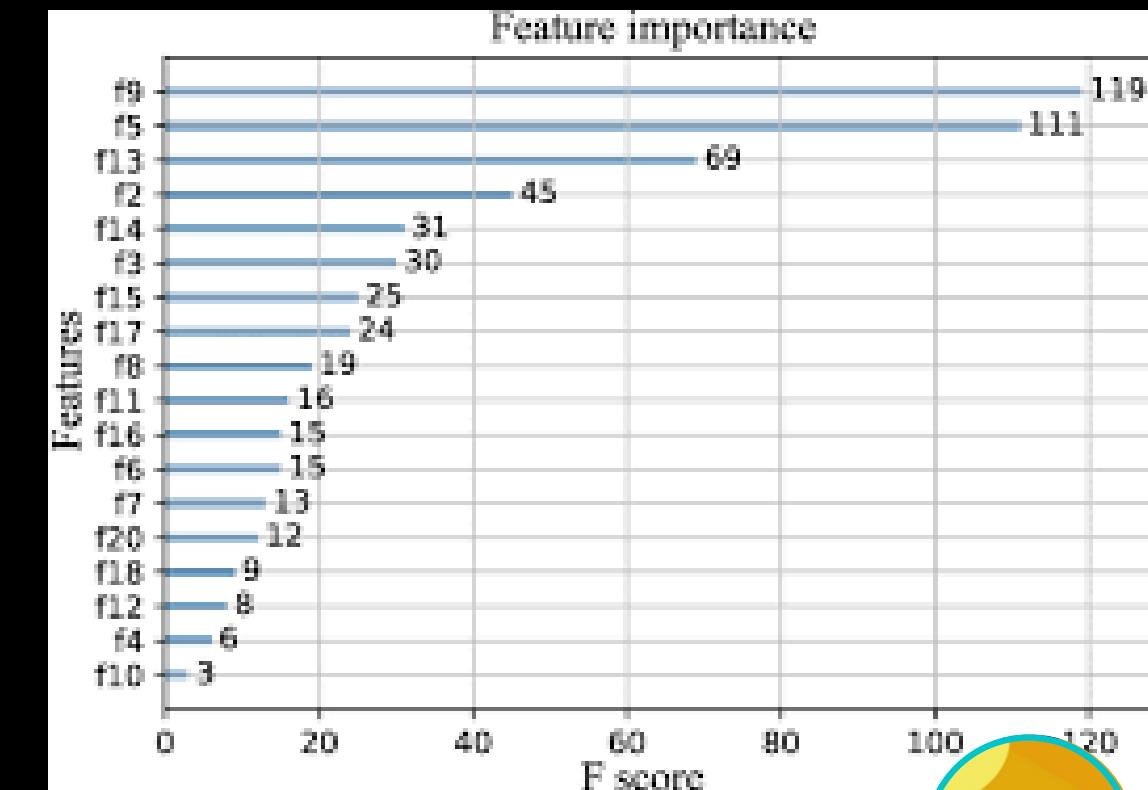


- Modeling

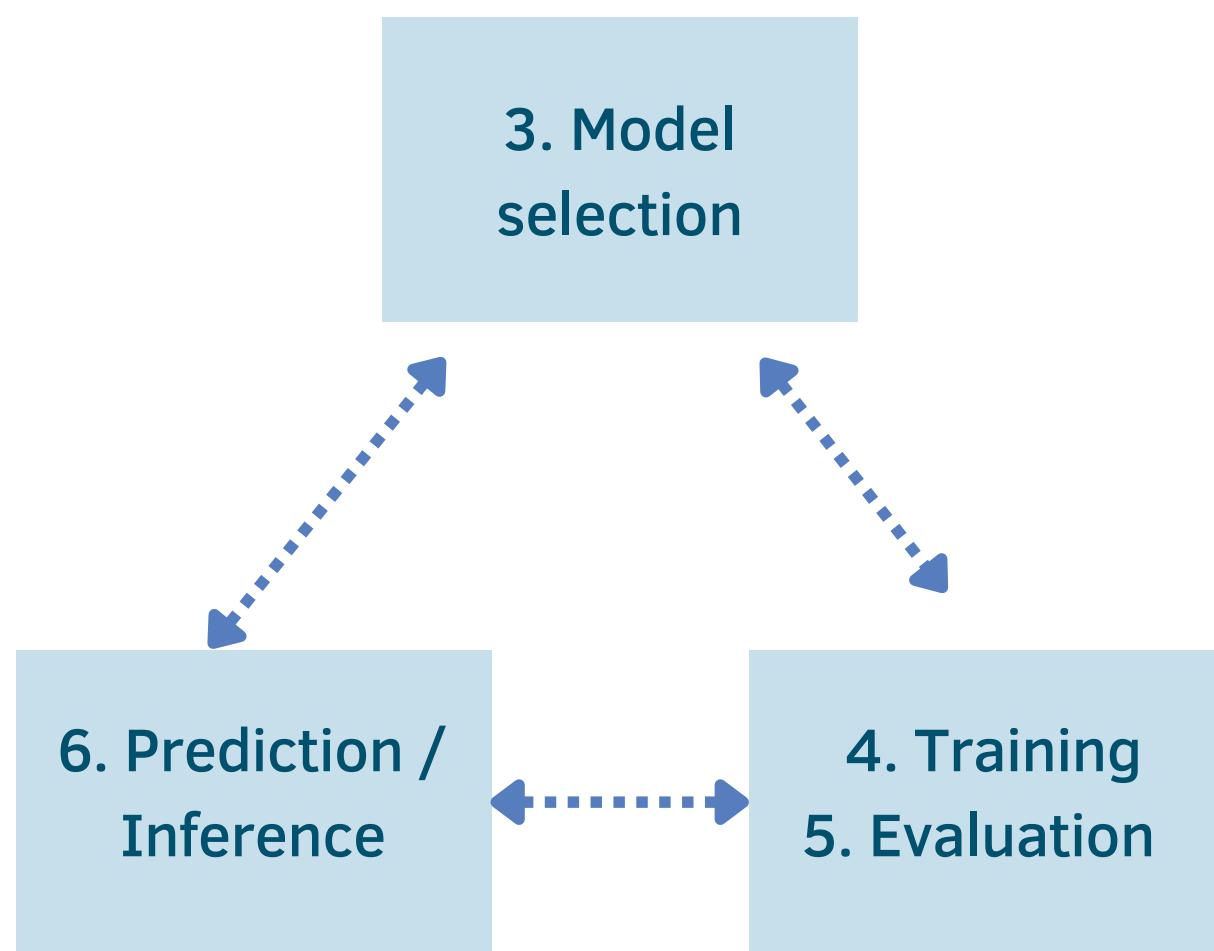
## INTERPRETATION

Post-hoc explainability techniques (explaining the black box problem)

- **Text explanations** (semantic mapping from model to symbols)
- **Visualizations** (of model's behaviour, e.g. dimensionality reduction),
- **Local explanations**,
- **Explanations by example** (centred in extracting representative examples that grasp the inner relationships and correlations found by the model being analyzed),
- Explanations by **simplification**
- **Feature relevance** (scoring of features).



# ML-based modeling



- Modeling

## MODEL SELECTION

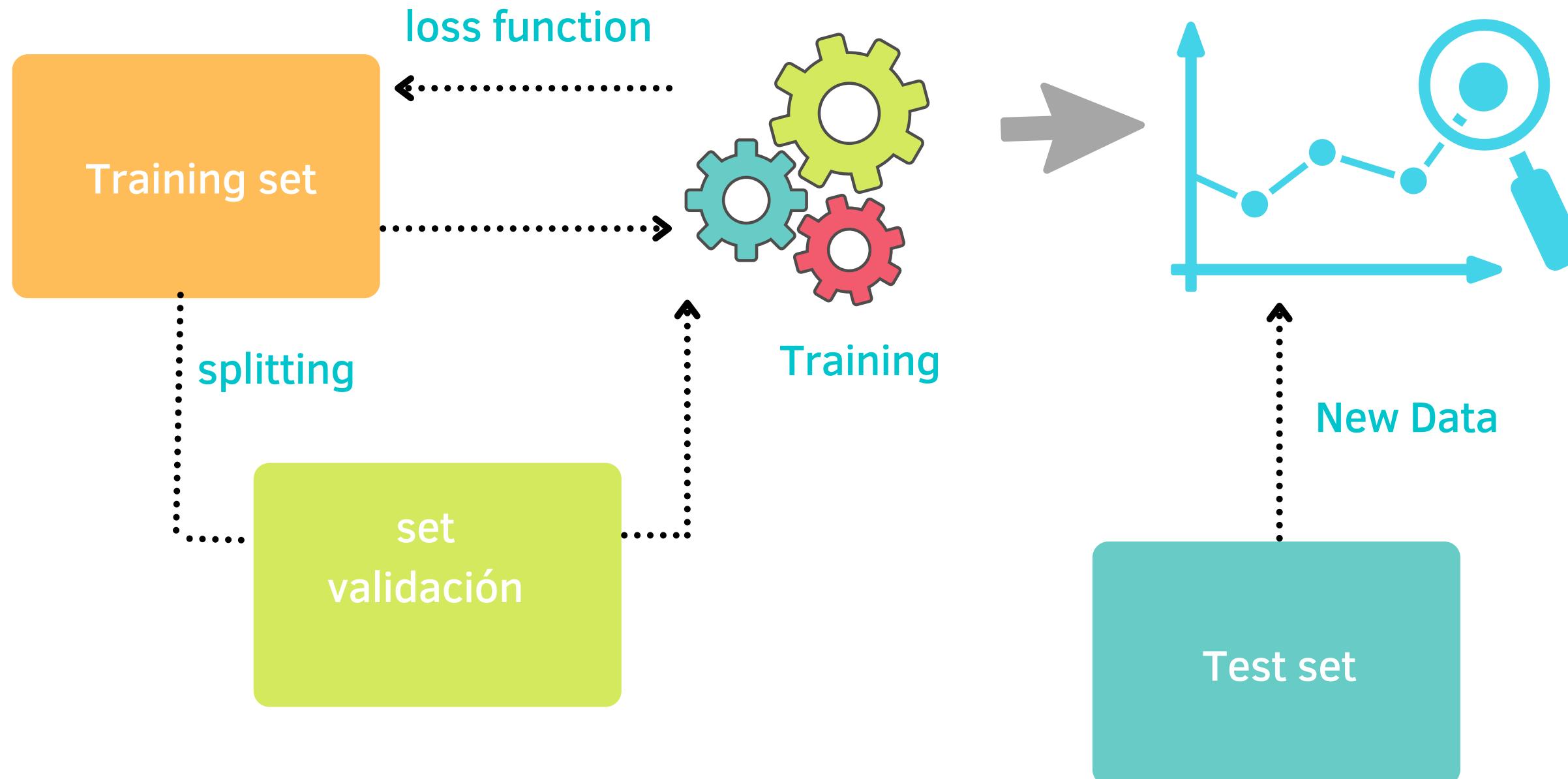
- The amount of data is important
- Do we have numeric or categorical data? Is it labeled? Or can it be labeled?
- How "transparent" do we need the model to be?
- Ask the right questions? (that an ML model can answer)
- Explore the bibliography related to your topic
- Compare more than one technique

MODEL IS CHOSEN!



TSWC, 2022

# ML-based modeling



## TRAINING

- Learning from the data

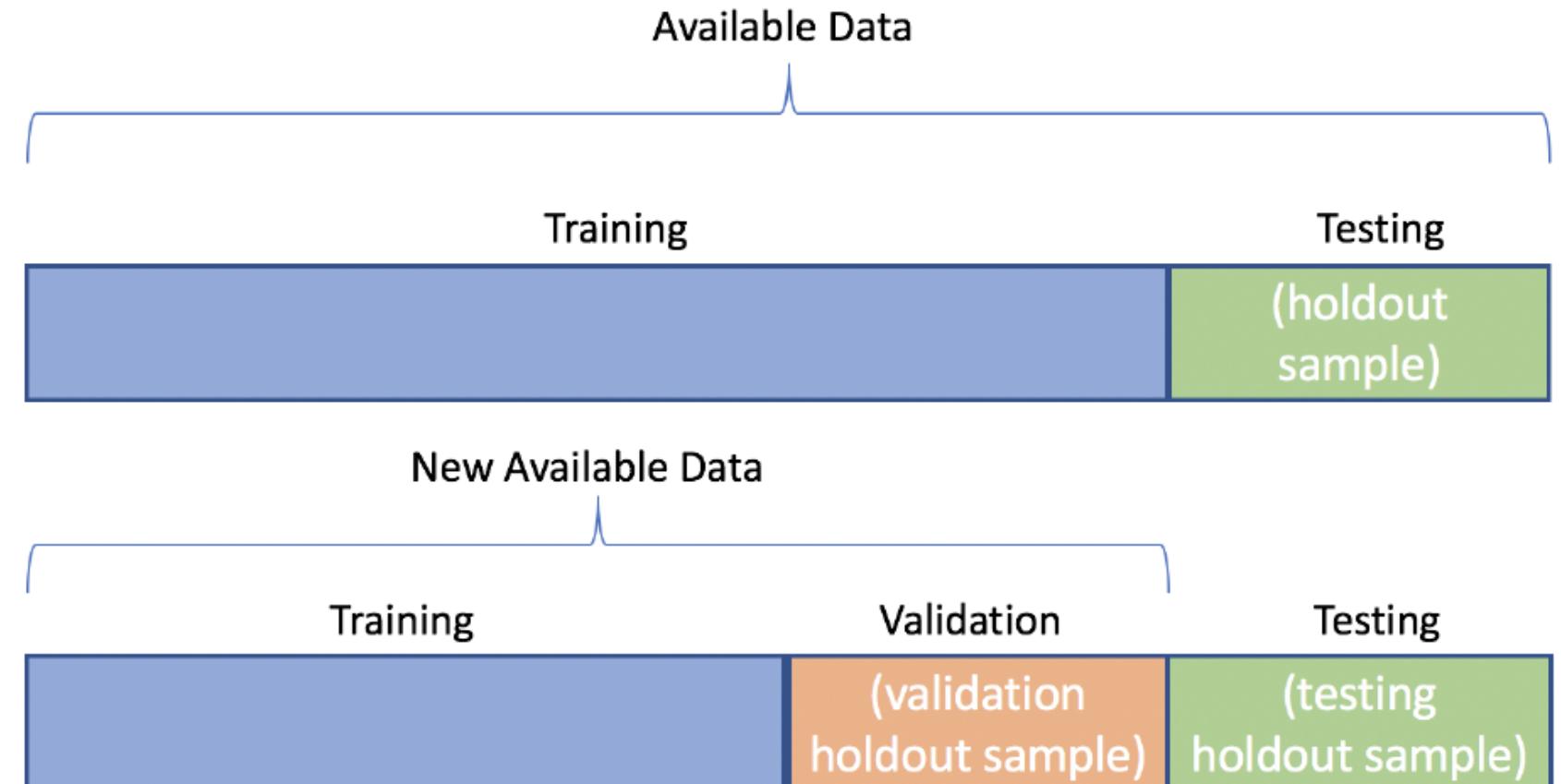
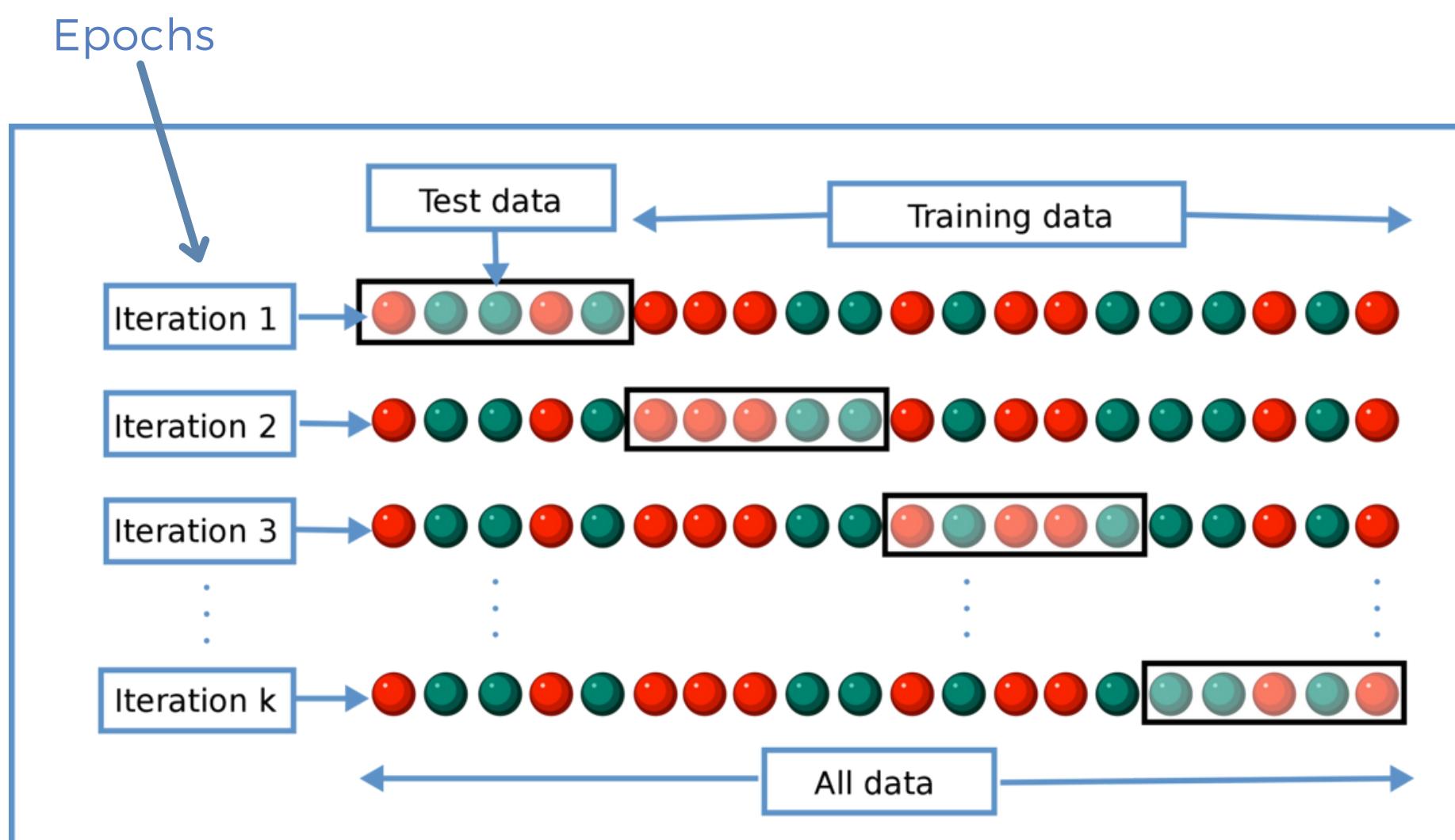


TSWC, 2022

# SPLITTING

split randomly (\*):

- Training set
- Validation set (to choose hyperparameters)
- Test set



## Hyperparameters

- Set before training
- E.g. #capas, dropout, learning rate, #neurons, loss function, etc

## Hyperparameter tuning (choose optimal parameters)

- Methods (e.g. grid search, evolutionary optimization )
- Previous domain knowledge (e.g. constraints)



# SPLITTING

- **data sampling strategies!**

Validation strategies for target prediction methods

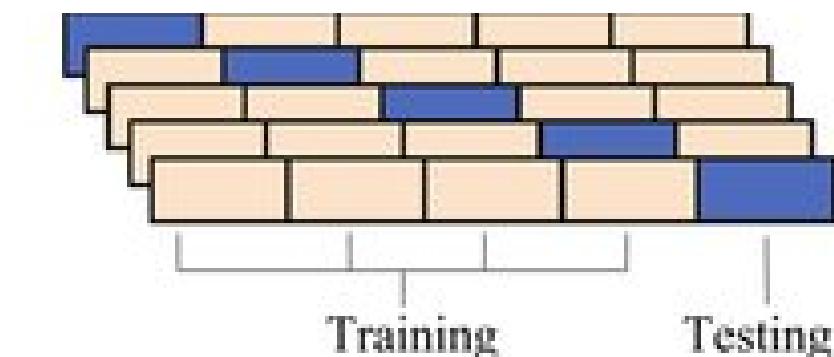
(Mathai et al, 2019)



Single train-test split



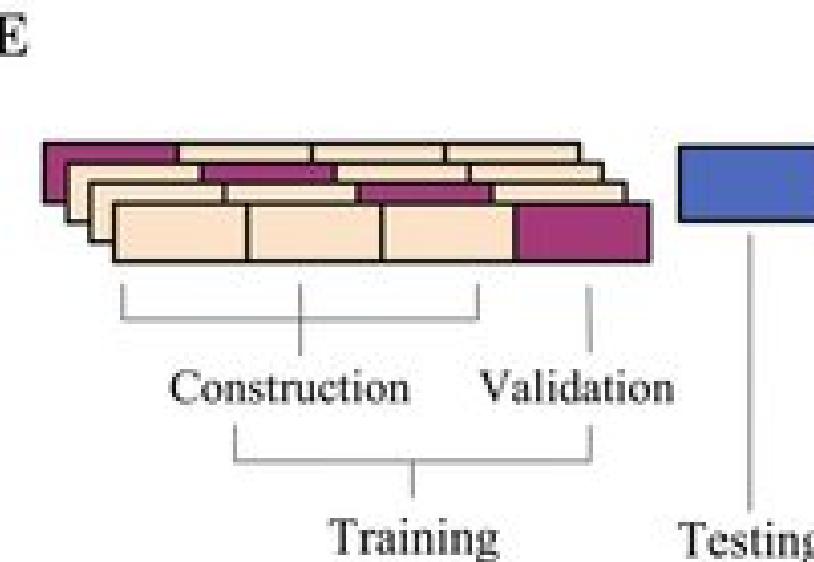
Single train-test time split



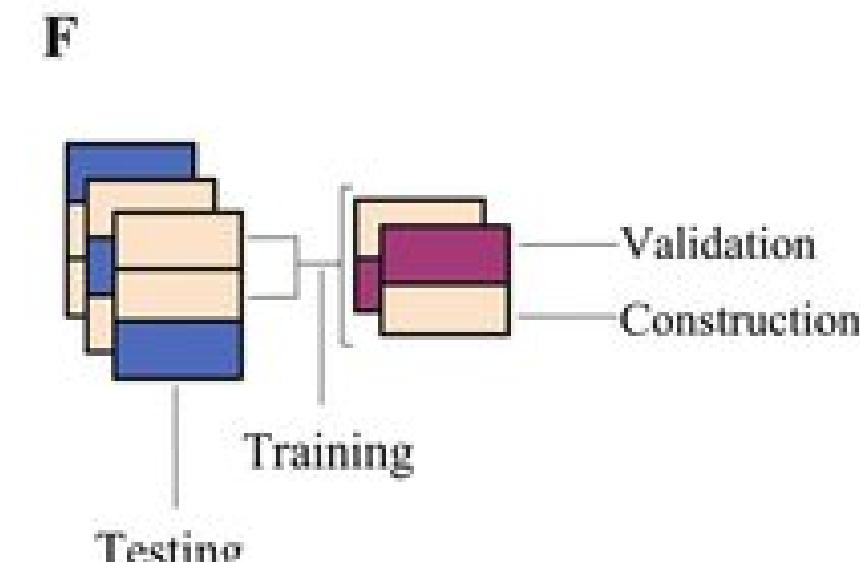
Cross-validation (5-fold)



Single train-test split with an external testing set



Cross-validation (4-fold) used for internal validation and an external testing set used for



Nested cross-validation with a 2-fold internal validation loop and a 3-fold external validation

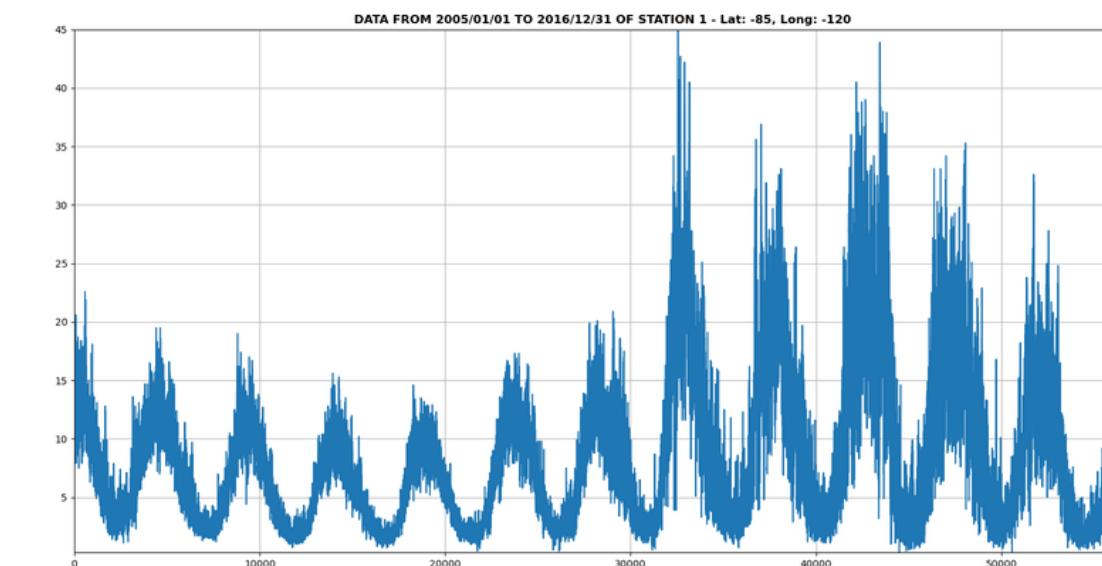
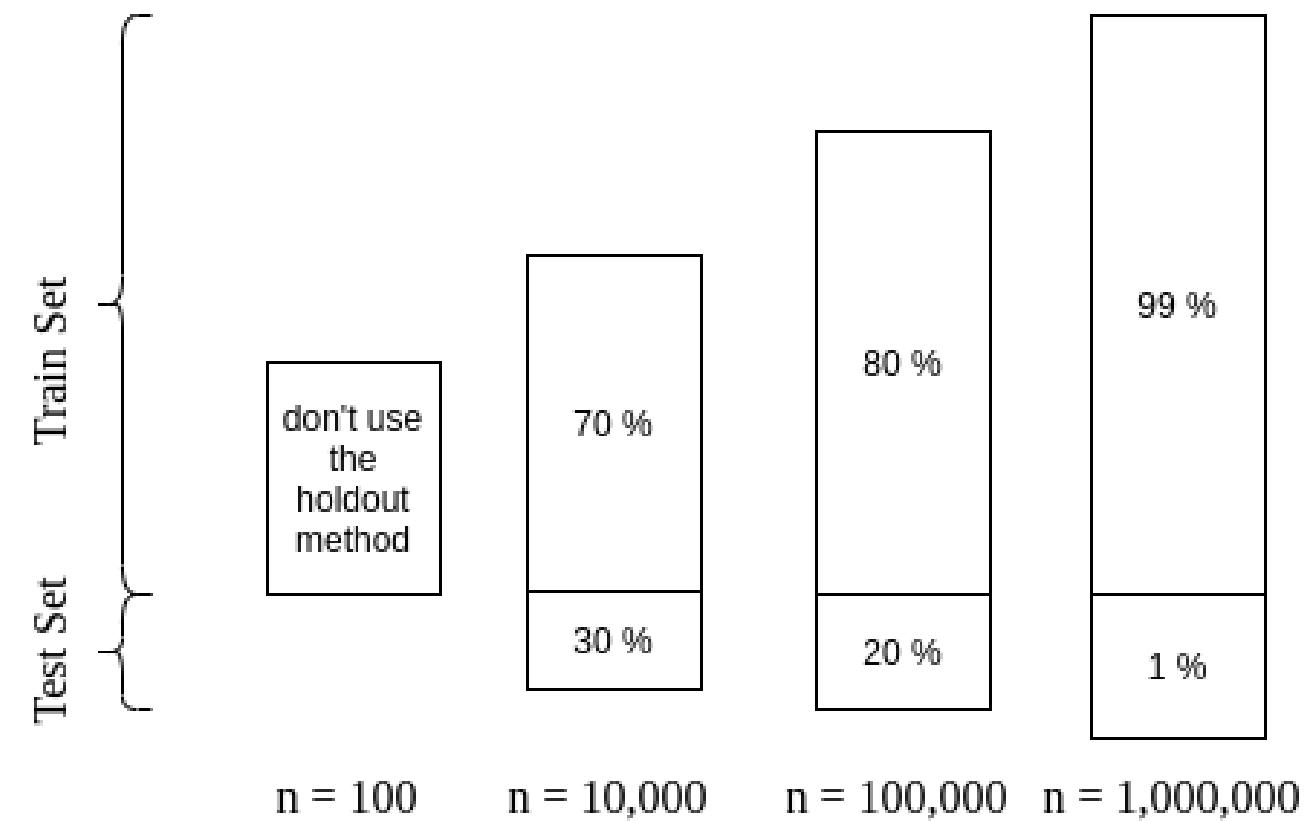


TSWC, 2022

# SPLITTING

- **data sampling strategies!**

Validation strategies for target prediction methods  
(Mathai et al, 2019)



- **Balanced/imbalanced datasets: another story!**



# LOSS FUNCTION

- or "cost function" or "target function"
- Measure the cost of inaccurate predictions
- measure how far an estimated value is from its true value
- is a method of evaluating how well specific algorithm models the given data.
- **Goal: Minimize the loss function during the training (optimization problem)**



WHILE NOT CONVERGE  
TRAIN  
MIN (loss function)



TSWC, 2022



# LOSS FUNCTION

- or "cost function" or "target function"
- Measure the cost of inaccurate predictions
- measure how far an estimated value is from its true value
- is a method of evaluating how well specific algorithm models the given data.
- Goal: Minimize the loss function during the training (optimization problem)

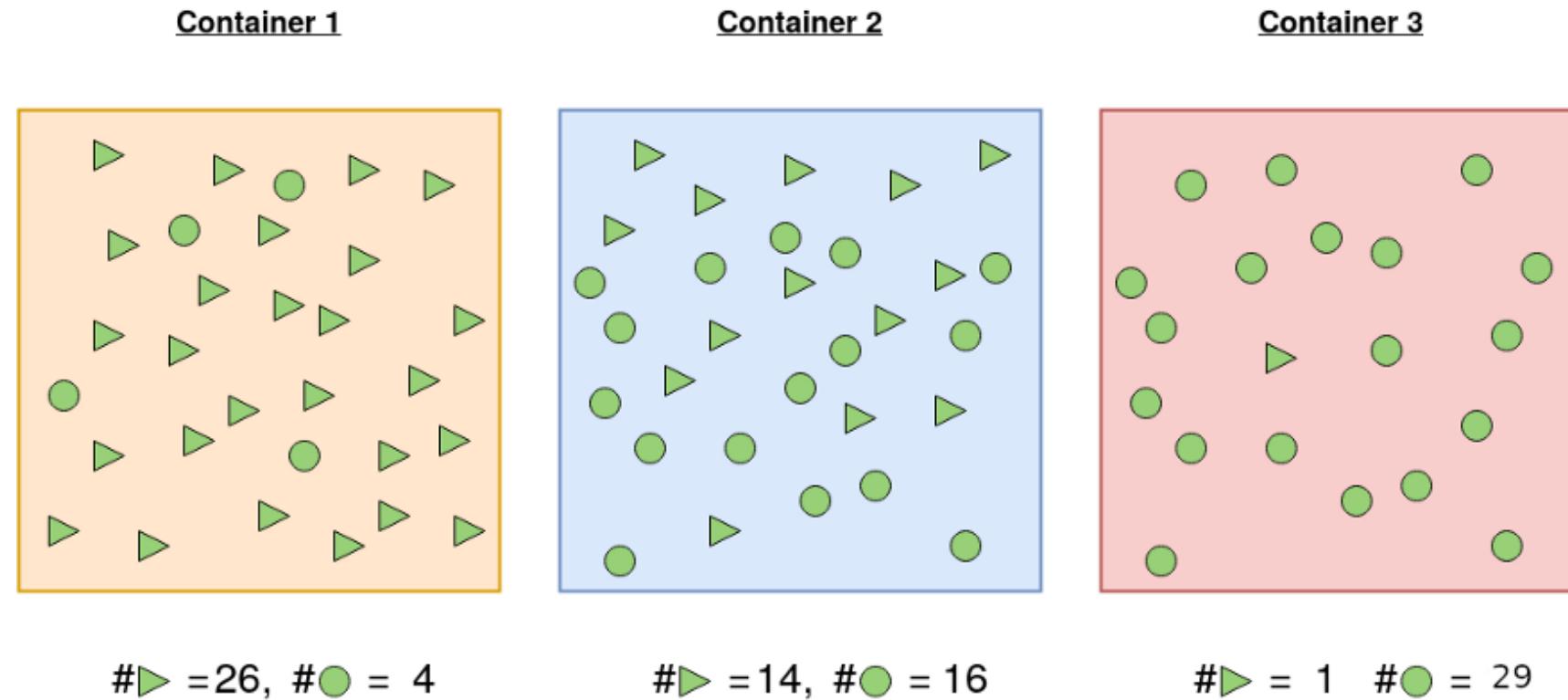
| <b>Tipo de problema</b> | <b>Loss function</b>   |             |
|-------------------------|--|-------------|
| Regresión               | Error cuadrático medio<br>(Mean Square Error)  | Cercano a 0 |
| Clasificación           | Cross entropy<br>Binary cross entropy<br>(mide la distancia entre las probabilidades de salida y la de los valores verdaderos) | Cercano a 0 |



# LOSS FUNCTION CLASSIFICATION

- Entropy of a random variable X is the level of uncertainty inherent in the variable's possible outcome.
- If the entropy is higher, that means we need more information to represent an event (info theory)

$$\text{Entropy}, H(X) = - \sum_{i=1}^n P(x_i) \log_2(P(x_i))$$



$$\begin{aligned} H(X) &= - \sum_x p(x) \log(p(x)) \\ &= -[p(x_1) \log_2(p(x_1)) + p(x_2) \log_2(p(x_2))] \\ &= -\left[\frac{26}{30} \log_2\left(\frac{26}{30}\right) + \frac{4}{30} \log_2\left(\frac{4}{30}\right)\right] \\ &= 0.5665 \end{aligned}$$

$$\begin{aligned} H(X) &= - \sum_x p(x) \log(p(x)) \\ &= -[p(x_1) \log_2(p(x_1)) + p(x_2) \log_2(p(x_2))] \\ &= -\left[\frac{14}{30} \log_2\left(\frac{14}{30}\right) + \frac{16}{30} \log_2\left(\frac{16}{30}\right)\right] \\ &= 0.9968 \end{aligned}$$

The entropy for the first and third container is smaller than the second one. This is because probability of picking a given shape is more certain in container 1 and 3 than in 2.

$$\begin{aligned} H(X) &= - \sum_x p(x) \log(p(x)) \\ &= -[p(x_1) \log_2(p(x_1)) + p(x_2) \log_2(p(x_2))] \\ &= -\left[\frac{1}{30} \log_2\left(\frac{1}{30}\right) + \frac{29}{30} \log_2\left(\frac{29}{30}\right)\right] \\ &= 0.2108 \end{aligned}$$



# LOSS FUNCTION CLASSIFICATION

- Cross-entropy (information theory) - the difference between two probability distributions
- Cross-entropy loss increases as the predicted probability diverge from the actual label.

## • Cross Entropy

$$H(p, q) = - \sum_{i=1}^n p(x_i) \log_2(q(x_i))$$

## • Binary Cross Entropy

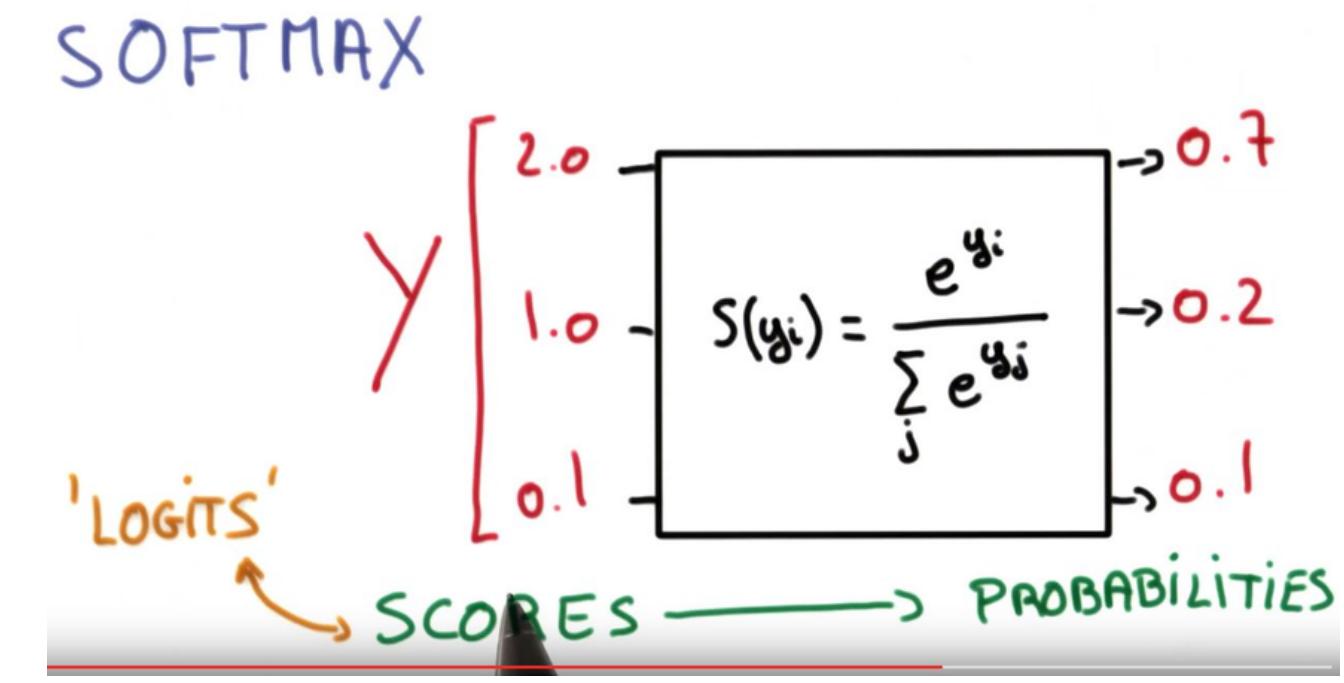
$$\begin{aligned} L &= - \sum_{i=1}^2 t_i \log(p_i) \\ &= - [t \log(p) + (1-t) \log(1-p)] \end{aligned}$$

where  $t_i$  is the truth value taking a value 0 or 1 and  $p_i$  is the Softmax probability for the  $i^{th}$  class.

- is often calculated as the average cross-entropy across all data examples

for  $N$  data points where  $t_i$  is the truth value taking a value 0 or 1 and  $p_i$  is the Softmax probability for the  $i^{th}$  data point.

$$L = -\frac{1}{N} \left[ \sum_{j=1}^N [t_j \log(p_j) + (1-t_j) \log(1-p_j)] \right]$$

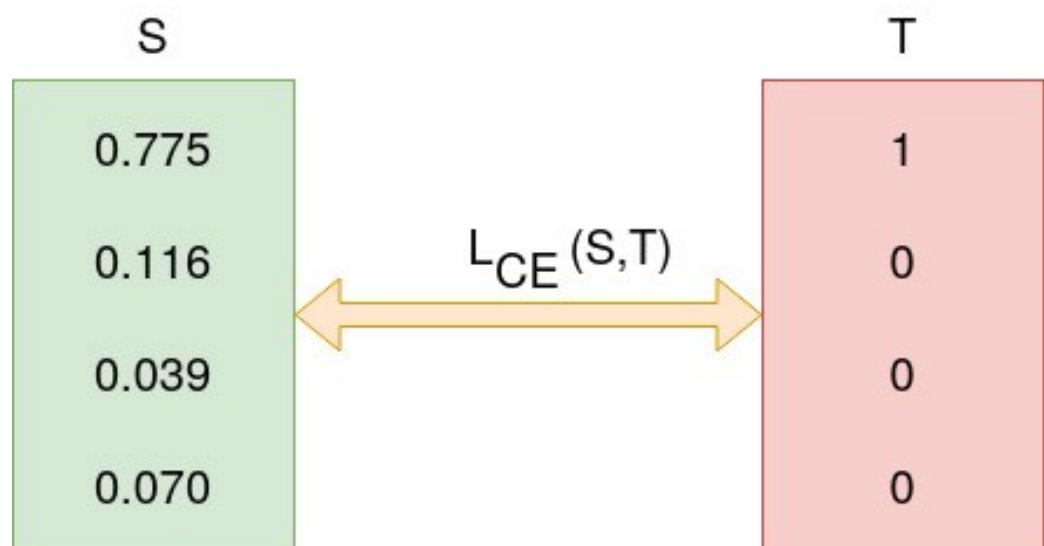


logits = unnormalised (or not-yet normalised)  
predictions (or outputs) of a model

# LOSS FUNCTION CLASSIFICATION

$$L = -\frac{1}{N} \left[ \sum_{j=1}^N [t_j \log(p_j) + (1 - t_j) \log(1 - p_j)] \right]$$

for  $N$  data points where  $t_i$  is the truth value taking a value 0 or 1 and  $p_i$  is the Softmax probability for the  $i^{th}$  data point.



$$\begin{aligned}
 L_{CE} &= - \sum_{i=1} T_i \log(S_i) \\
 &= - [1 \log_2(0.775) + 0 \log_2(0.126) + 0 \log_2(0.039) + 0 \log_2(0.070)] \\
 &= - \log_2(0.775) \\
 &= 0.3677
 \end{aligned}$$

- Notice that when actual label is 1 ( $t_i = 1$ ), second half of function =0 whereas in case actual label is 0 ( $t_i = 0$ ) first half is dropped off. In short, we are just multiplying the log of the actual predicted probability for the ground truth class.

- Each predicted class probability is compared to the actual class desired output 0 or 1 and a score/loss is calculated that penalizes the probability based on how far it is from the actual expected value. The penalty is logarithmic in nature yielding a large score for large differences close to 1 and small score for small differences tending to 0. During the training (different iterations using the dataset) the Loss function is optimized
- In binary classification it means reducing the cross-entropy

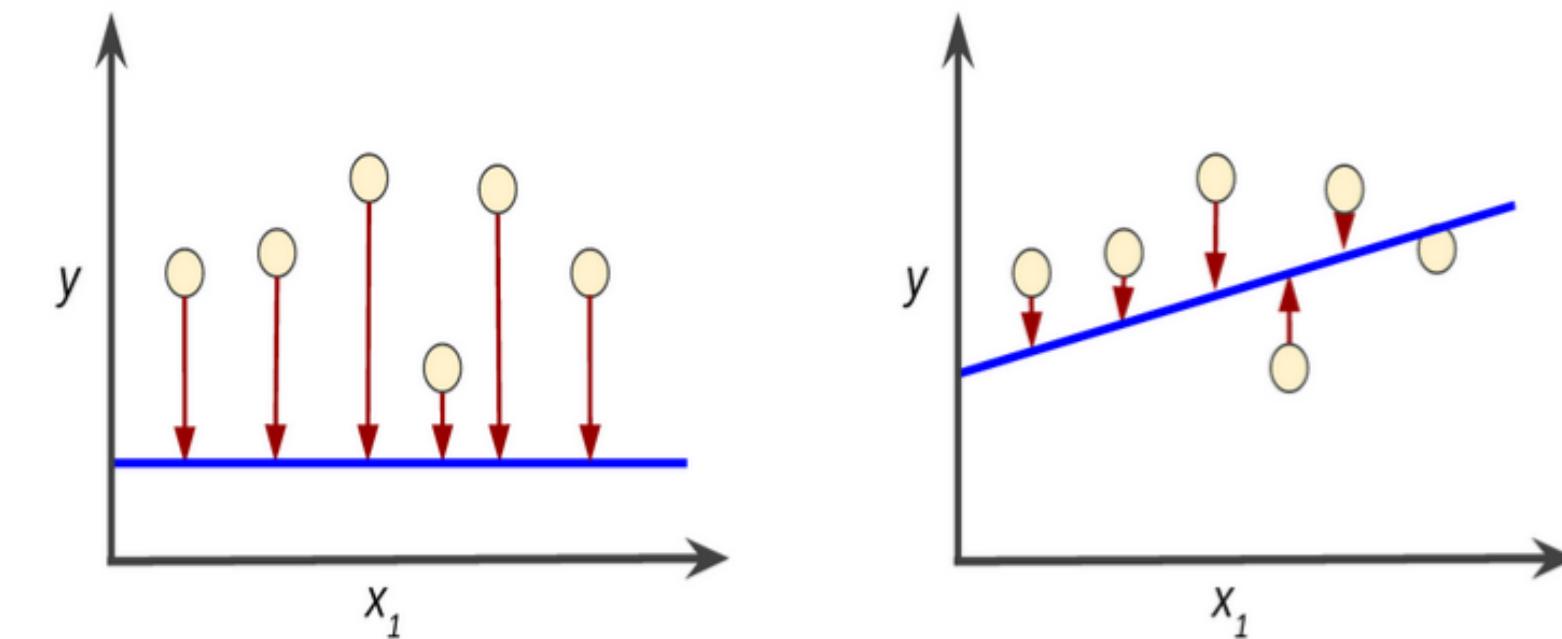




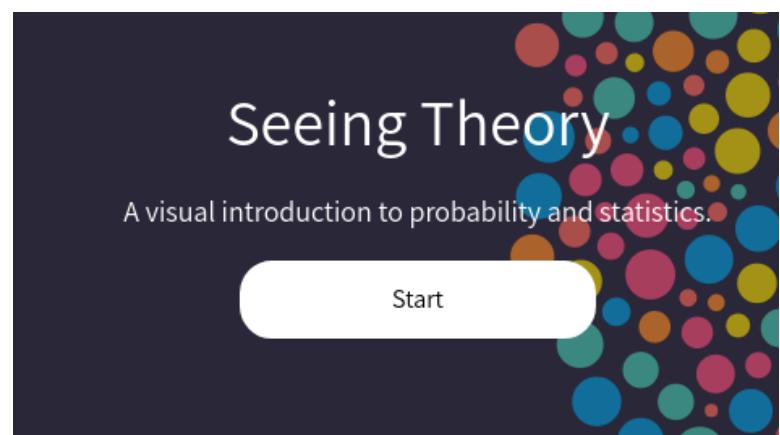
## LOSS FUNCTION

## REGRESSION

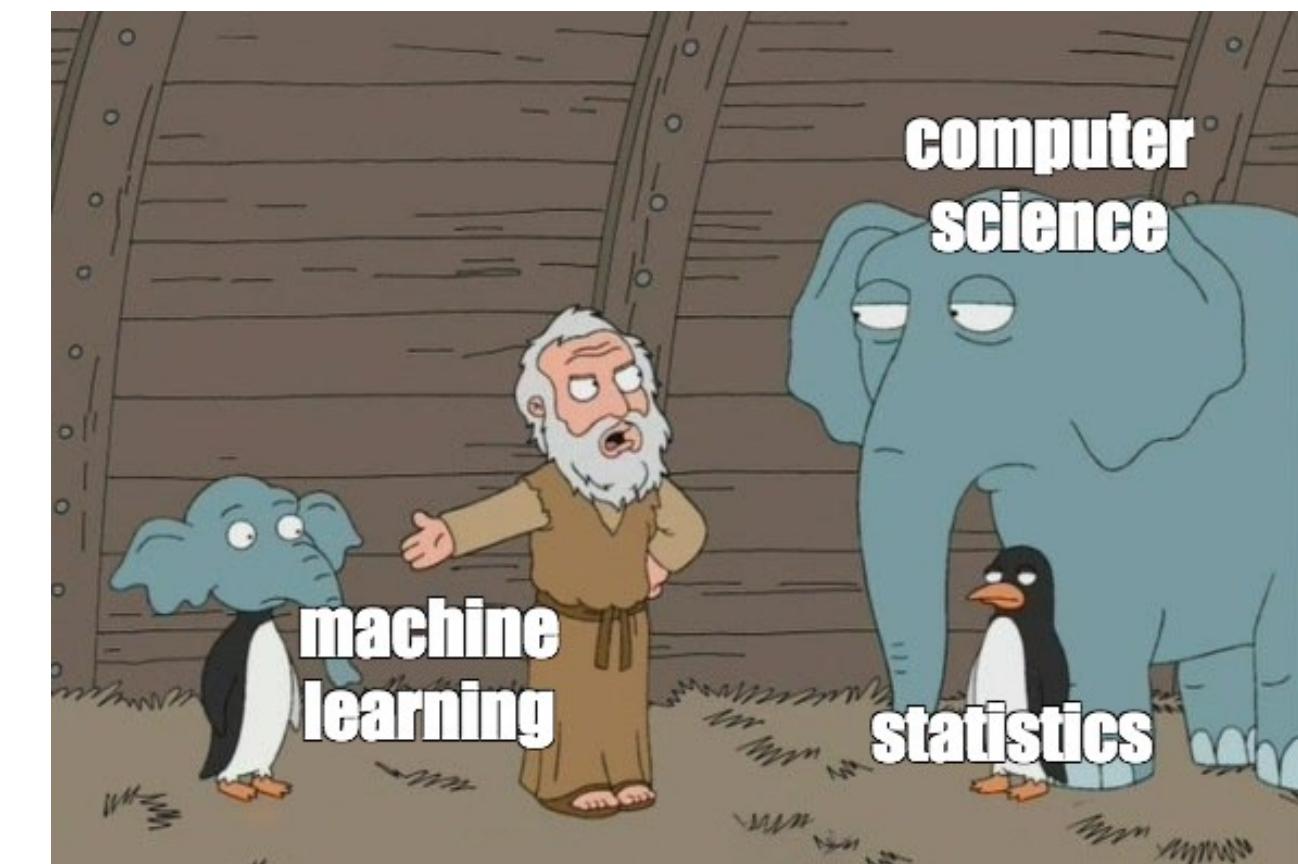
- Mean square error is measured as the average of the squared difference between predictions and actual observations. It's only concerned with the average magnitude of error irrespective of their direction. However, due to squaring, predictions that are far away from actual values are penalized heavily in comparison to less deviated predictions.

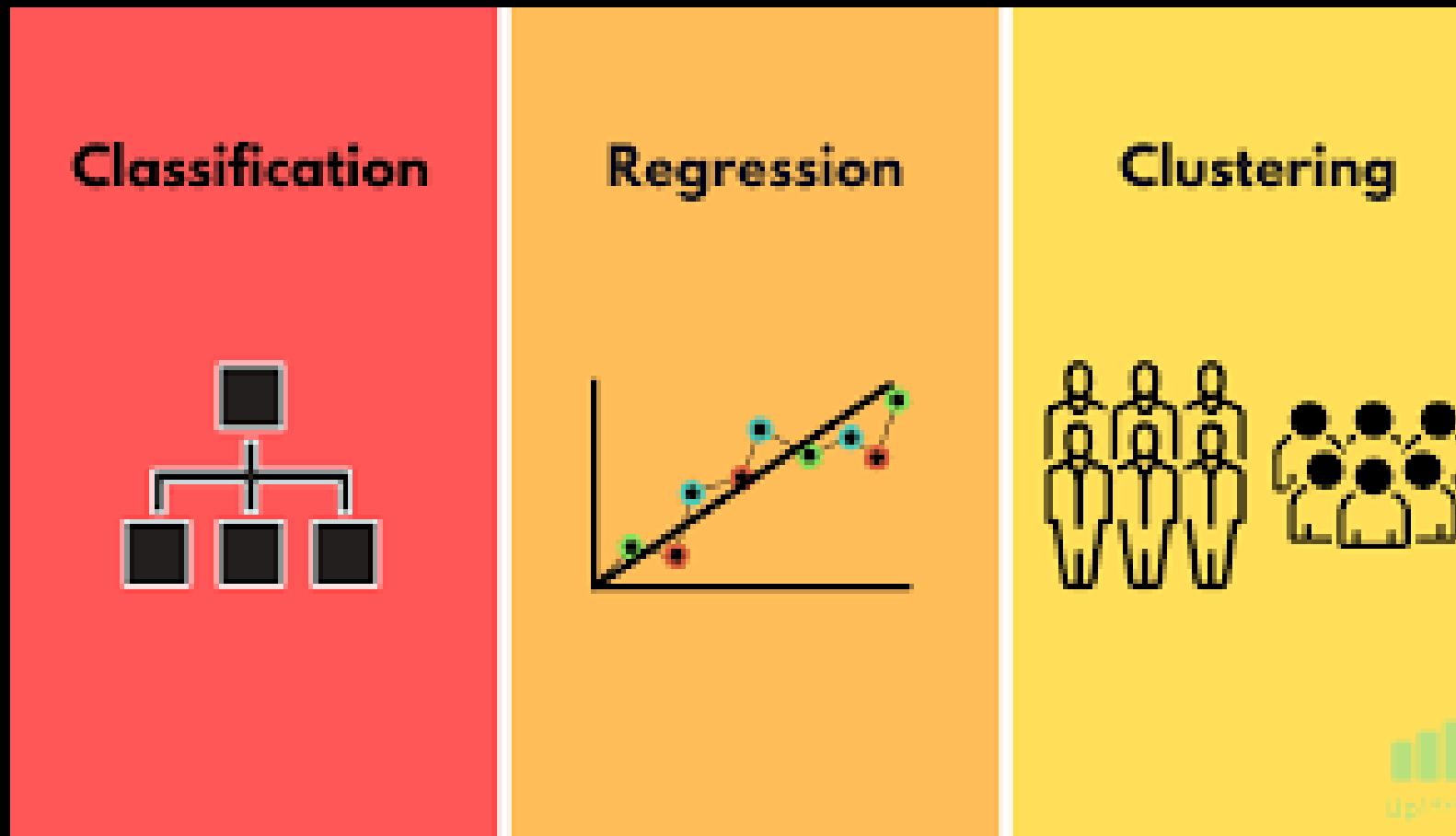


$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$



<https://seeing-theory.brown.edu>





- Classification: confusion matrix (binary), AUC & ROC (multiclass)
- Regression: MAE,MSE,RMSE (on the test set)
- Clustering: ? ? ? ?

## PERFORMANCE

- Sometimes also called model validation (!)
  - Uses the **testset** to compare against the prediction with ML
  - Metrics
- 



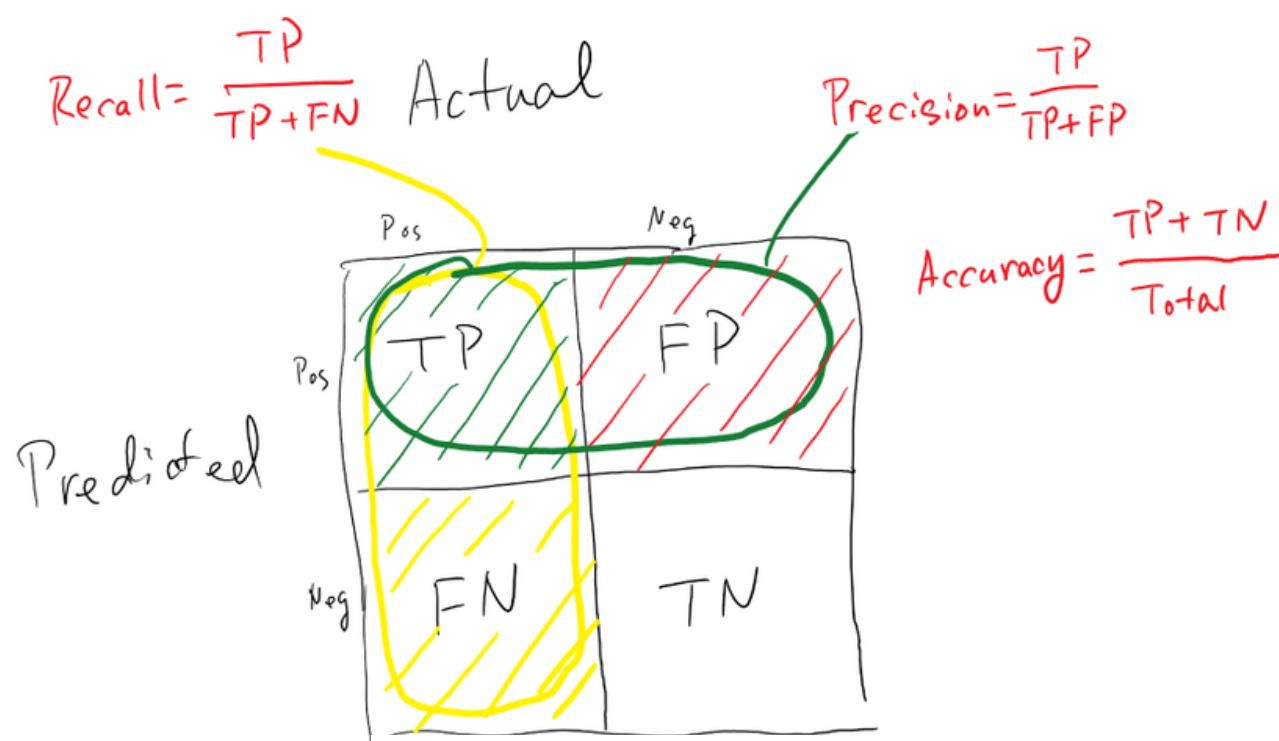
TSWC, 2022

# CLASSIFICATION



TSWC, 2022

|                 |          | True Class |          |
|-----------------|----------|------------|----------|
|                 |          | Positive   | Negative |
| Predicted Class | Positive | TP         | FP       |
|                 | Negative | FN         | TN       |



$$F\text{-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

- Accuracy: It is defined as the closeness of the predicted value to the actual value.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

- Precision: Precision is defined based on true positive values only out of all positive values.

$$\text{Precision} = TP / (TP + FP)$$

- Recall: It is also known as sensitivity or hit rate or true positive rate. It tells how good our estimator or model is to predict the positive values.

$$\text{Recall} = TP / (TP + FN)$$

- <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

# CLASSIFICATION



TSWC, 2022

|           | Elephant | Monkey | Fish | Lion |
|-----------|----------|--------|------|------|
| Actual    | 25       | 3      | 0    | 2    |
| Elephant  | 3        | 53     | 2    | 3    |
| Monkey    | 2        | 1      | 24   | 2    |
| Fish      | 1        | 0      | 2    | 71   |
| Lion      |          |        |      |      |
| Predicted | Elephant | Monkey | Fish | Lion |

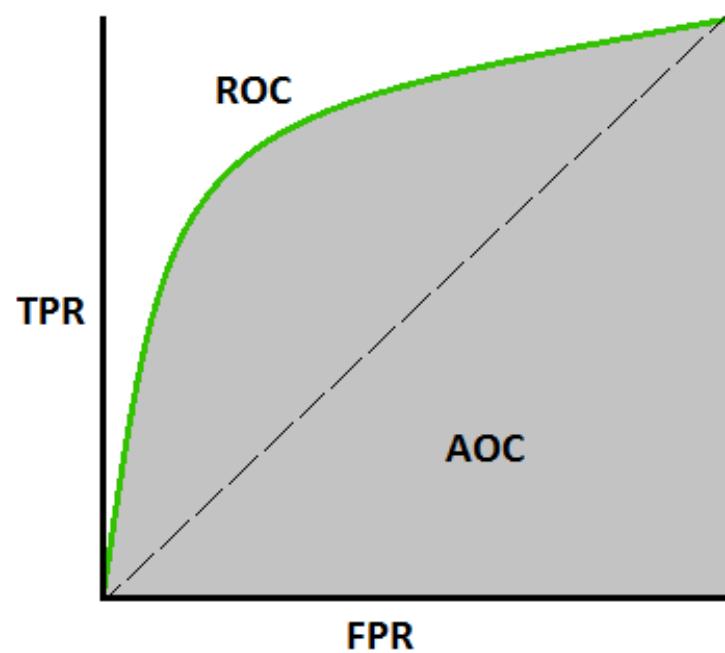
## Multiclass

- Confusion matrix
- AUC =Area Under The Curve
- ROC = Receiver Operating Characteristics curve.

better model, >> AUC

AUC = tells how much the model is capable of distinguishing between classes

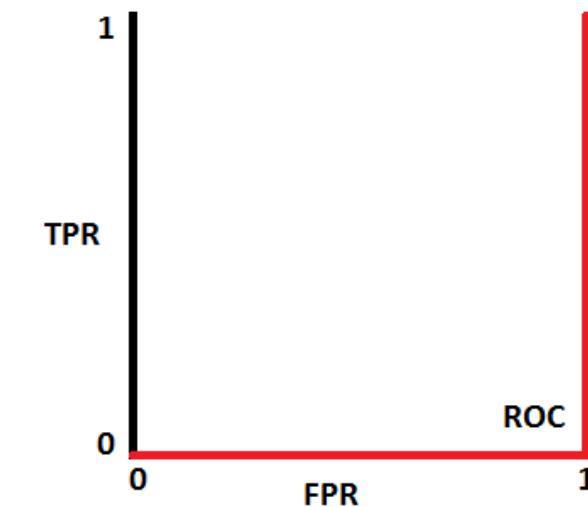
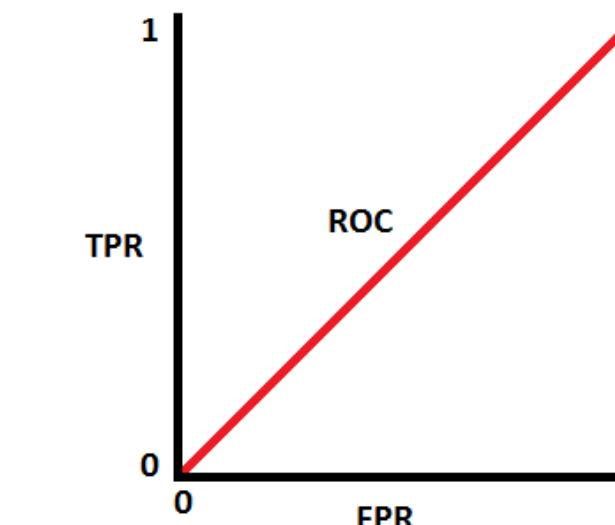
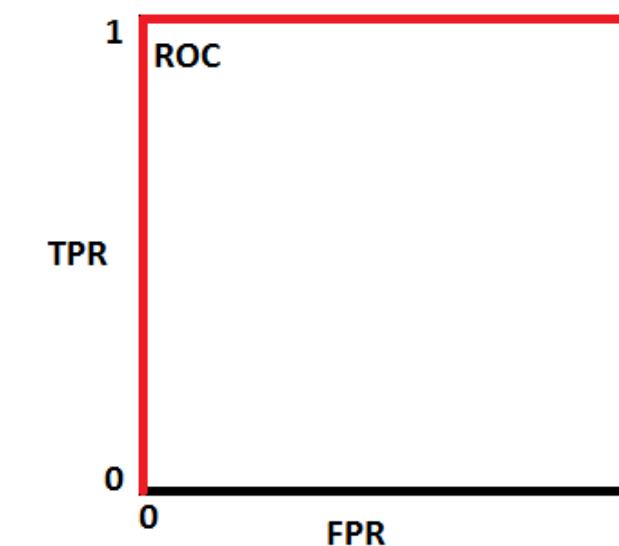
ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes



$$\text{TPR / Recall / Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

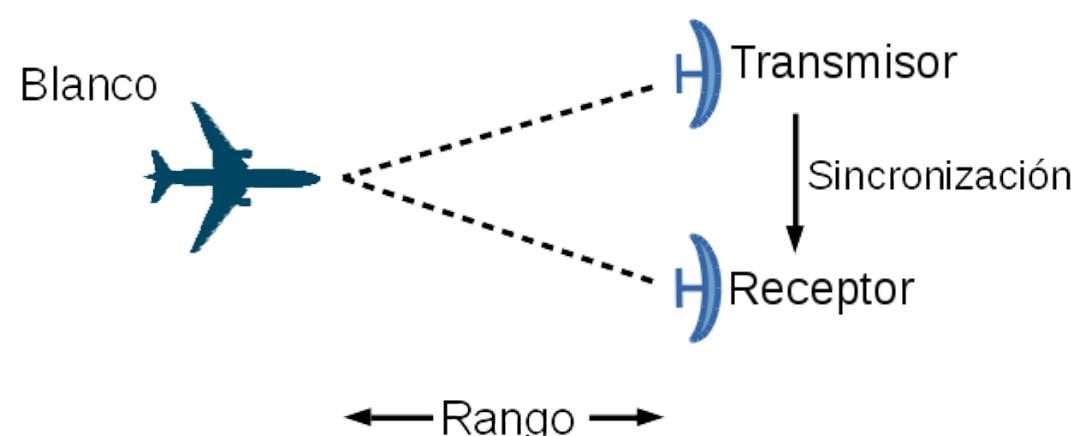
$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\begin{aligned} \text{FPR} &= 1 - \text{Specificity} \\ &= \frac{\text{FP}}{\text{TN} + \text{FP}} \end{aligned}$$



- <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

# CLASSIFICATION (Application: Echo detection)

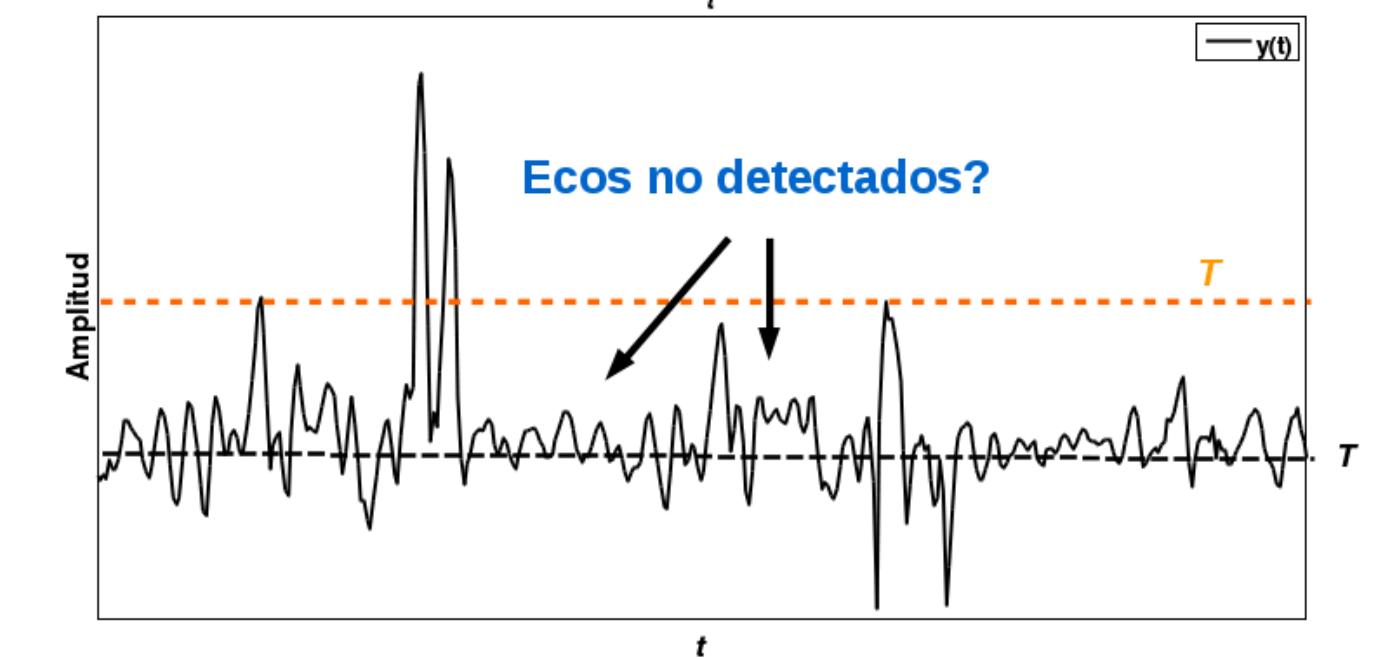
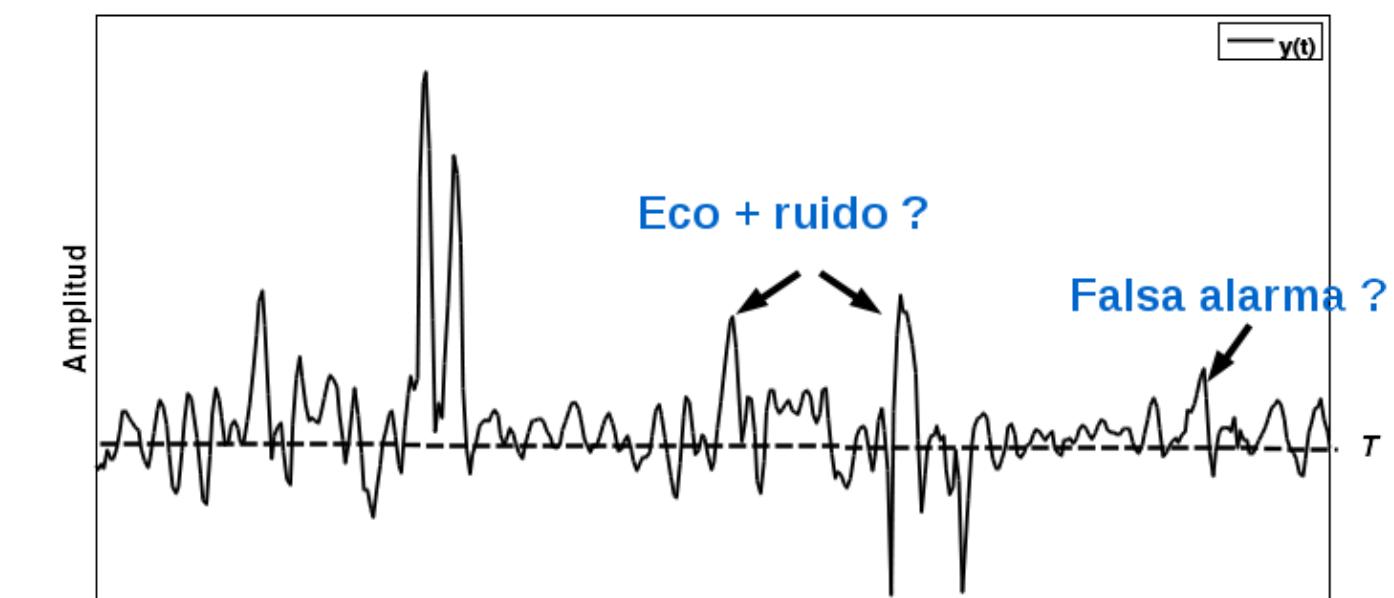


$H_0$ :  
noise/interference/jamming  
only

$H_1$ :  
noise/interference/jamming  
+ echo from a target

|       | $H_1$     | $H_0$                |
|-------|-----------|----------------------|
| $H_0$ | Target    | Positive Detection ✓ |
| $H_1$ | No Target | False Alarm ✗        |
|       |           | Miss ✗               |
|       |           | No Detection ✓       |

Cada  $y_i > T$  es una detección



Si  $P_d \uparrow$ , entonces  $P_{fa}$  también  $\uparrow$

Maximizar  $P_d$  sujeta a  $P_{fa} \leq \beta$

- $P_d$  ( $H_1$ )
- $P_{fa}$  ( $H_1$ )
- $P_m$  ( $H_0$ )

$$T = \alpha \hat{y}$$

- How to think the problem from ML point of view?



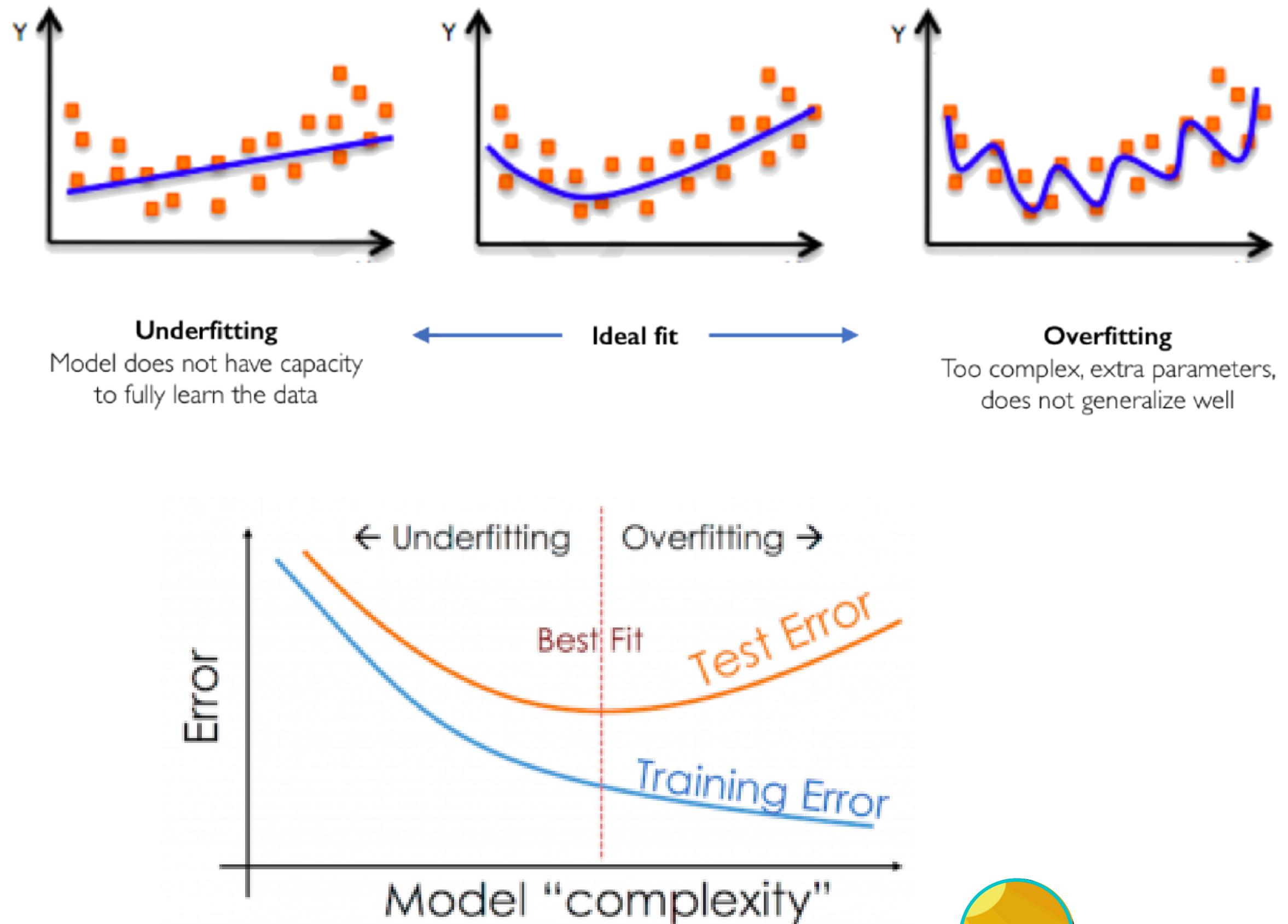
Automatic ionospheric layers detection: Algorithms analysis

Maria G. Molina <sup>a,b,\*</sup>, Enrico Zuccheretti <sup>c</sup>, Miguel A. Cabrera <sup>b</sup>, Cesidio Bianchi <sup>c</sup>, Umberto Sciacca <sup>c</sup>, James Baskaradas <sup>d</sup>



TSWC, 2022

# PERFORMANCE



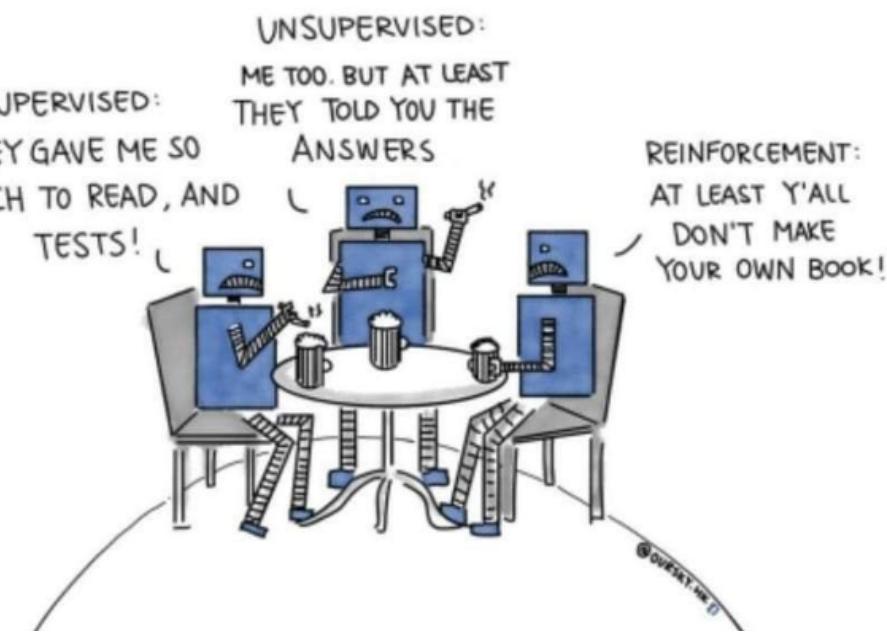
- There are techniques to overcome overfitting



TSWC, 2022

# Algorithms

# MACHINE LEARNING



## Supervised



### Regression

- Linear
- Polynomial

## Unsupervised



### Clustering

- K-means
- PCA
- SVD

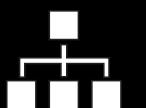
Continuous  
data



### Decision Tree



### Random Forest



### Classification

- KNN
- Logistic Regression
- Naive-Bayes
- SVM

Association Rule  
Learning

Categorical  
data

Hidden Markov  
Model

## Artificial Neural Networks

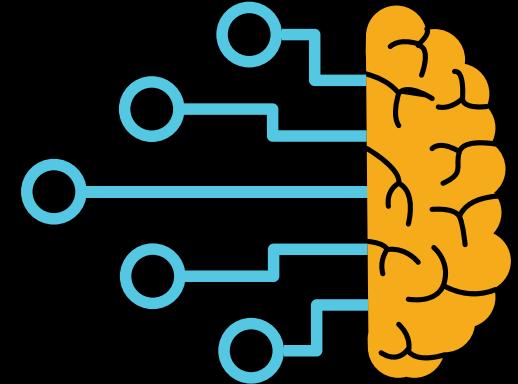


TSWC, 2022

DCAO, UBA  
08-12 AGOSTO 2022

# APRENDIZAJE AUTOMÁTICO.

Fundamentos y Aplicaciones en  
Meteorología del Espacio



---

Dra María Graciela Molina

FACET-UNT / CONICET  
Tucumán Space Weather Center - TSWC

<https://spaceweather.facet.unt.edu.ar/>  
IG -> @spaceweatherargentina

gmolina@herrera.unt.edu.ar

