Introduction to Machine Learning

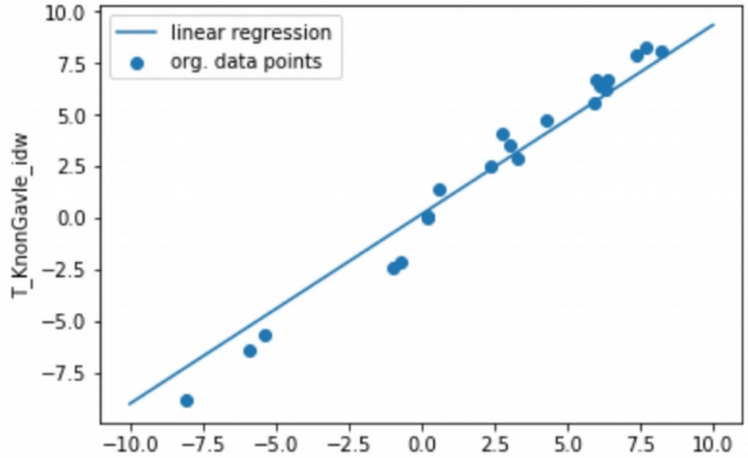
How to navigate through the universe of learning algorithms?

Sven Decker

GEO4300 Geophysical Data Science
Module 11: Machine Learning
University of Oslo
08. November 2018

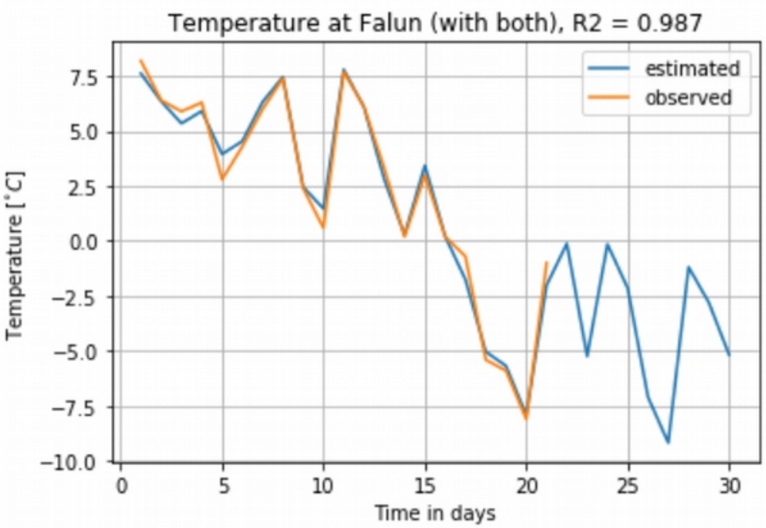
Is this machine learning?

Out[16]: <matplotlib.legend.Legend at 0x7f65b8118550>

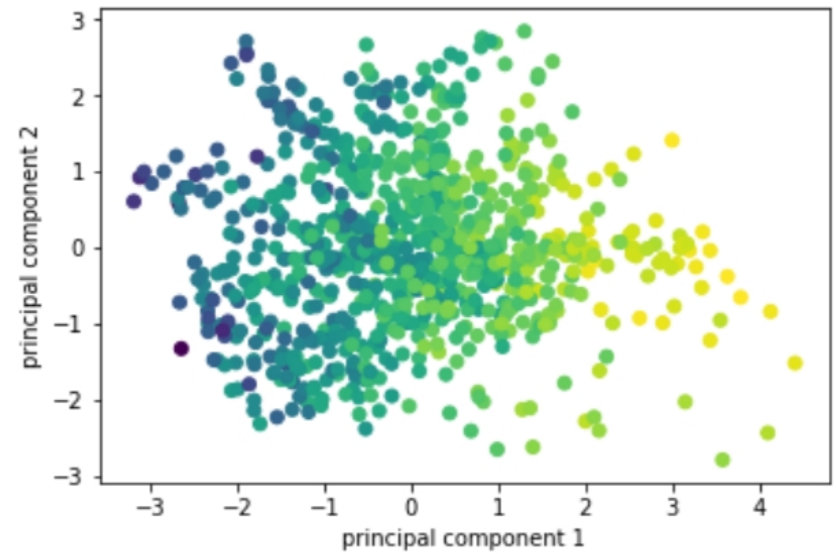


Out[4]:

	Enet	wind_speed	RH_air	T_air	hour
principal component 1	0.549267	-0.019078	-0.482102	0.557344	0.393558
principal component 2	0.188627	0.855545	0.160218	-0.274193	0.362783



Out[5]: Text(0,0.5,'principal component 2')

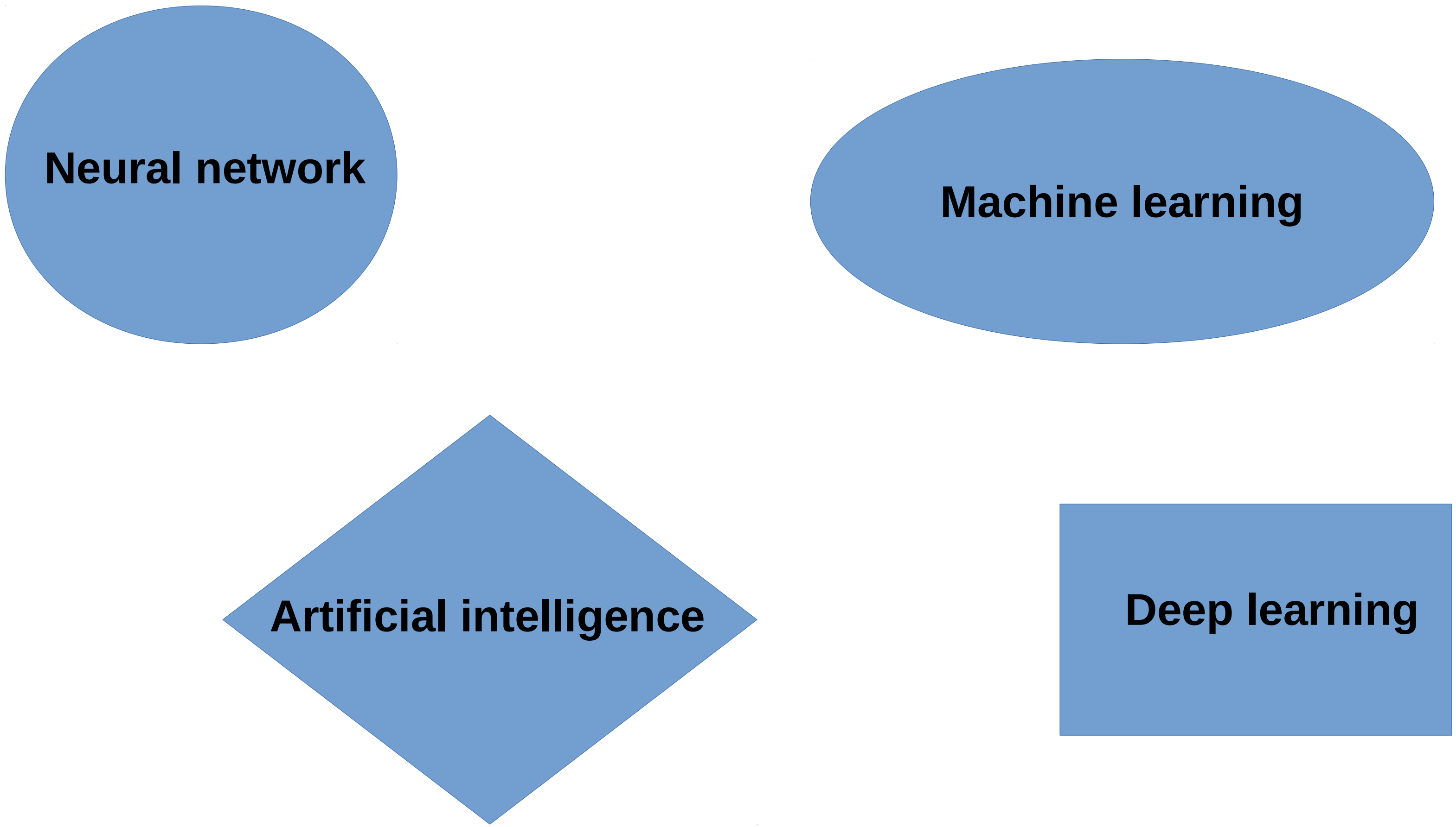


- Definition of machine learning
- How to structure machine learning algorithms

- The Machine Learning Design Cycle

- Examples

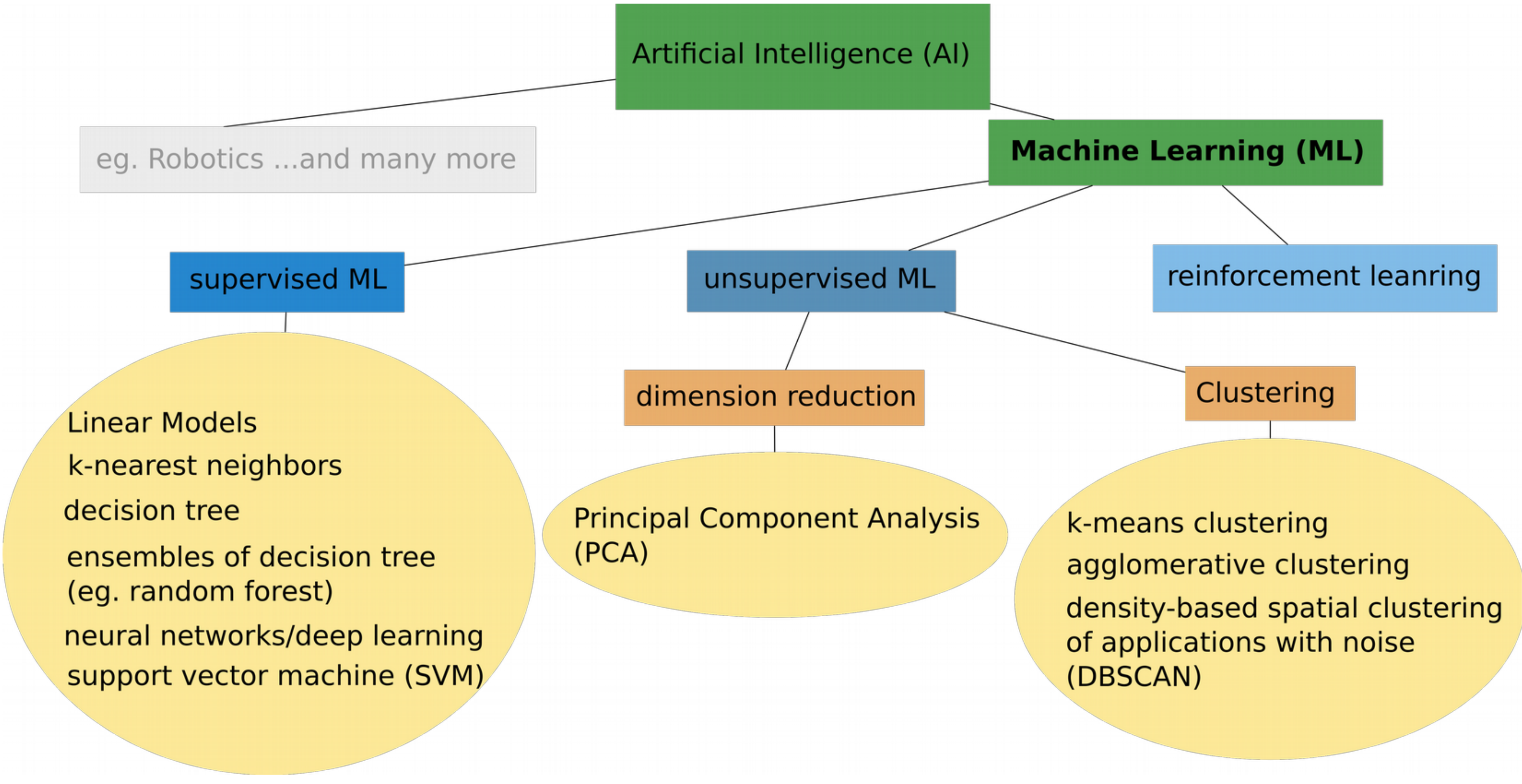
How to handle all these buzzwords?



Definition Machine Learning

- Definition by wikipedia:
  - ***Machine learning (ML)*** *is a field of* ***artificial intelligence (AI)*** *that uses statistical techniques to give computer systems the ability to "learn" from data, without being explicitly programmed.*
- **Machine Learning** is the basis for **deep learning** and **neural networks**.
- **Deep learning** and **neural networks** are synonymous

Map of machine learning



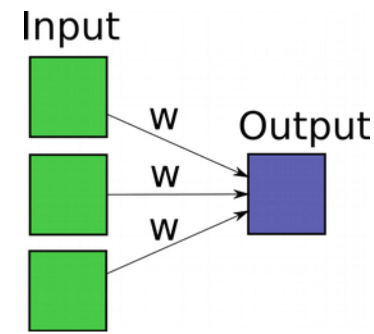
Unsupervised ML

- Dimension reduction
  - Principle component analysis (PCA)
- Clustering
  - K-means clustering
    - Tries to find cluster center representing certain areas of the dataset
  - Agglomerative clustering
    - Start with a cluster for each point and pools nearest clusters until stopping criteria
  - Density-based spatial clustering of applications with noise (DBSCAN)
    - No a-priori setting of number of clusters needed
    - Can capture more complex shapes

Reinforcement learning

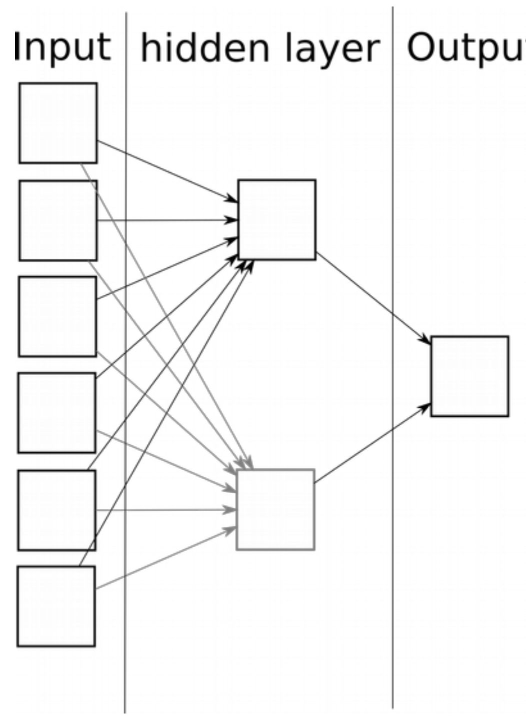
- The algorithm does a self-evaluation and decides how to change the model for further improvement

Supervised ML



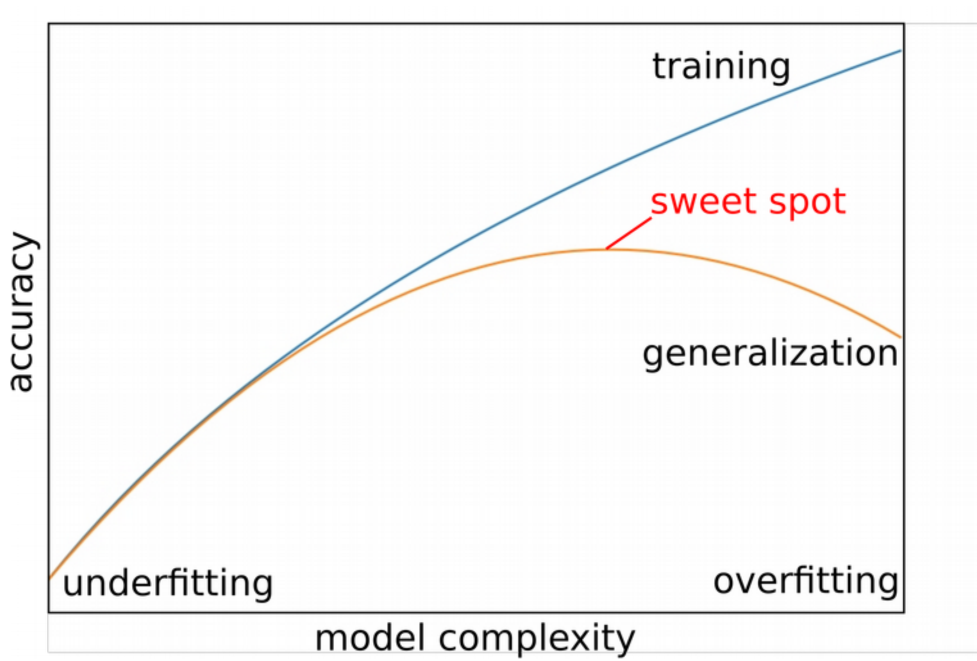
- Two major types:
  - Classification: predict a certain class, eg. vegetation class
  - Regression: predict a continuous number, eg. evapotranspiration
- Many methods can be used for both cases

- Linear Models
  - Use a linear function based on input features to predict
- K-nearest neighbors
  - The model stores the training dataset and calculates the distance to the k nearest neighbors for any prediction dataset.
- Decision tree and (random) forests
  - Decision trees learn a hierachy of if/else questions leading to a decision
- Neural network/deep learning
  - **Neural networks** are several layers of linear regressions with an activation functions
  - ‘Deep’ in **deep learning** refers to multi-layer (= deep) regressions.
  - Already two layers are considered as ‘deep’
- Support vector machine



Generalization and Overfitting

- Goal: train a model to predict accurate on unseen data
  - The model should be able to generalize from training data to test data.
- Simple models generalize better to new data

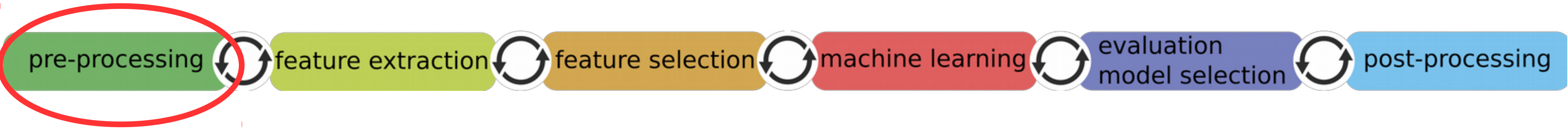


- Overfitting: too complex model, with bad generalization
- Underfitting: too simple model, with bad generalization
- The larger the variety of your dataset is, the more complex your model can be without overfitting.
- Regularization can avoid overfitting (and improve generalization)
  (each feature should have as little effect as possible on the outcome)

The Machine Learning Design Cycle

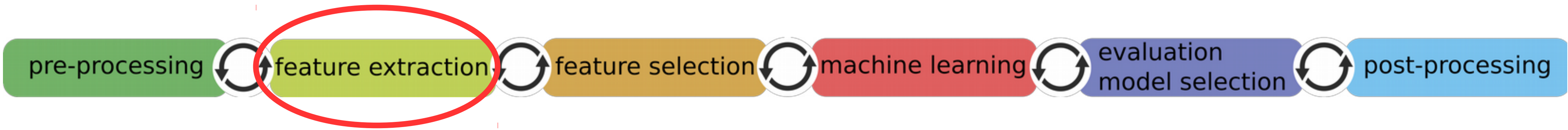
- Every single step is as important as the actual machine learning step
- Often the features used for the algorithm are more important than the algorithm itself.
- While working on your problem, you will go back and forth through this cycle many times – it is always important to keep in mind where in the cycle you are…
- Often many algorithms are tested and first towards the end the best or final one is selected.

Design Cycle – Pre-processing



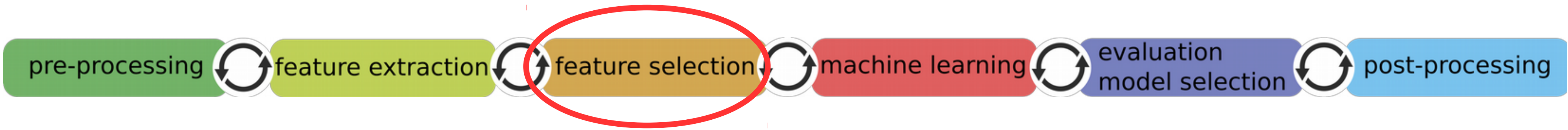
- Frame your problem as an ML problem
- Get the data, permissions etc.
- Clean the data
  - Where are missing values and how to treat them?
  - Are outliers meaningful or eg. sensor failing?
- Get the data into the right form (X, y)
- Split data into training, validation and test dataset (analogous to calibration and validation datasets in hydrologic modelling)

Design Cycle – Feature Extraction



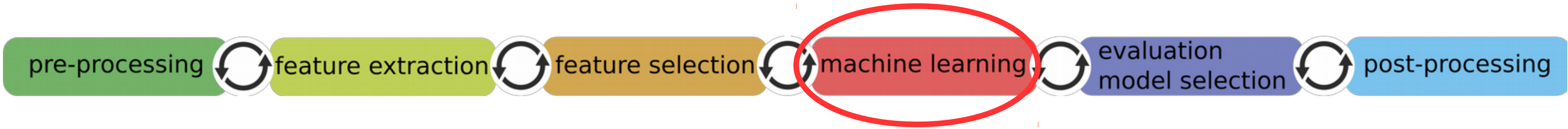
- Combined features are often more expressive than their components
  - EXAMMPLE
- Encode features if necessary
  - For categorical features
    - eg. snow crystal types, vegetation type, precipitation type
    - The algorithm cannot handle strings
    - One way is to use one-hot-encoding:
      - Rain = [1,0,0]
        Snow = [0,1,0]
        Sleet = [0,0,1]

Design Cycle – Feature Selection



- Sometimes too many features are available
  - Feature creation (eg. by combining and encoding) can blow up the number of features
- Too many features will break some algorithms or result in a bad performance
- One way to reduce the number of features:
  - Principle component analysis (PCA)

Design Cycle – machine learning



- ‘core’ of the workflow, but not more important than the other steps
- Different algorithms could be used to solve your problem
  - Computational efficiency
  - Statistical efficiency
- Can you phrase your problem in different ways or can you combine various machine learning paradigms?
  - eg. AlphaGo:
    - Classification to predict expert moves
    - Regression to evaluate board positions
    - Reinforcement learning by self-play
  - eg. learning in kids:
    - ‘Unsupervised’ perception of environment
    - ‘supervised’ signal from y: ‘Look, there is a cat!’ (only very few data points to learn)

Design Cycle – evaluation and model selection / post-processing

- Generalize
  - Perform well on unseen data
    - Precision measures
    - Calculation speed (for both training and testing)
- Which algorithm(s) do we choose?
  - Computational efficiency, accuracy
  - Generalization performance
    - Split dataset in training, validation and test set
    - Test set is only for final test
- Which hyperparameter?
  - eg. number of layer in netwokrs, constants on training the algorithm
  - In contrast to internal model parameters
- Transform model output to meaningful, understandable output

Take home messages

- Definition of machine learning
  - Part of AI
  - Three major sub-fields:
    - Supervised ML: known target, powerful for prediction
    - Unsupervised: unknown target, powerful for exploring
    - Reinforcement learning
  - Typical problems are classification or regression problems
- Design Cycle
  - Each step is equally important
  - While working on the problem, you will go back and forth
  - **Split your dataset in training and testing data and NEVER mix up**
  - Be careful about overfitting and generalization