Machine Learning Foundations

Jesper Dramsch

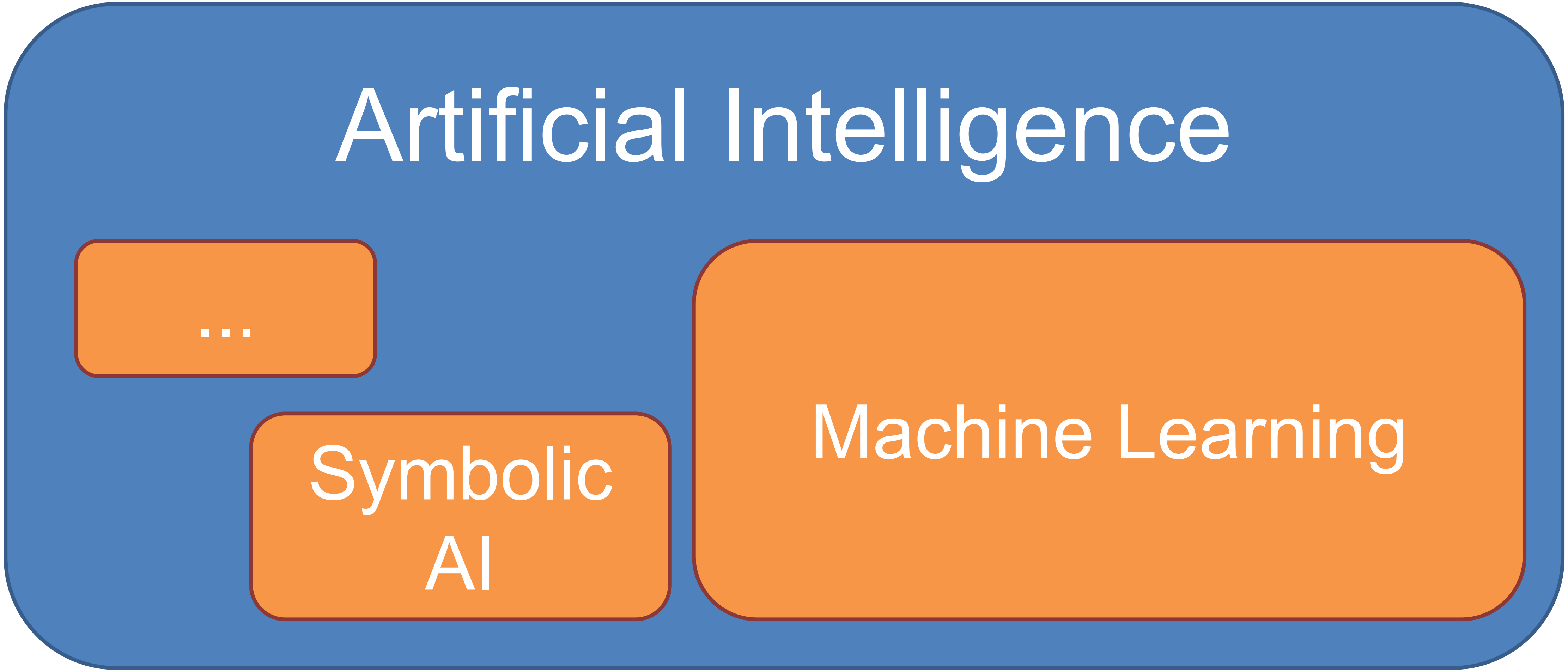
Scientist for Machine Learning

Jesper.Dramsch@ecmwf.int

Outline

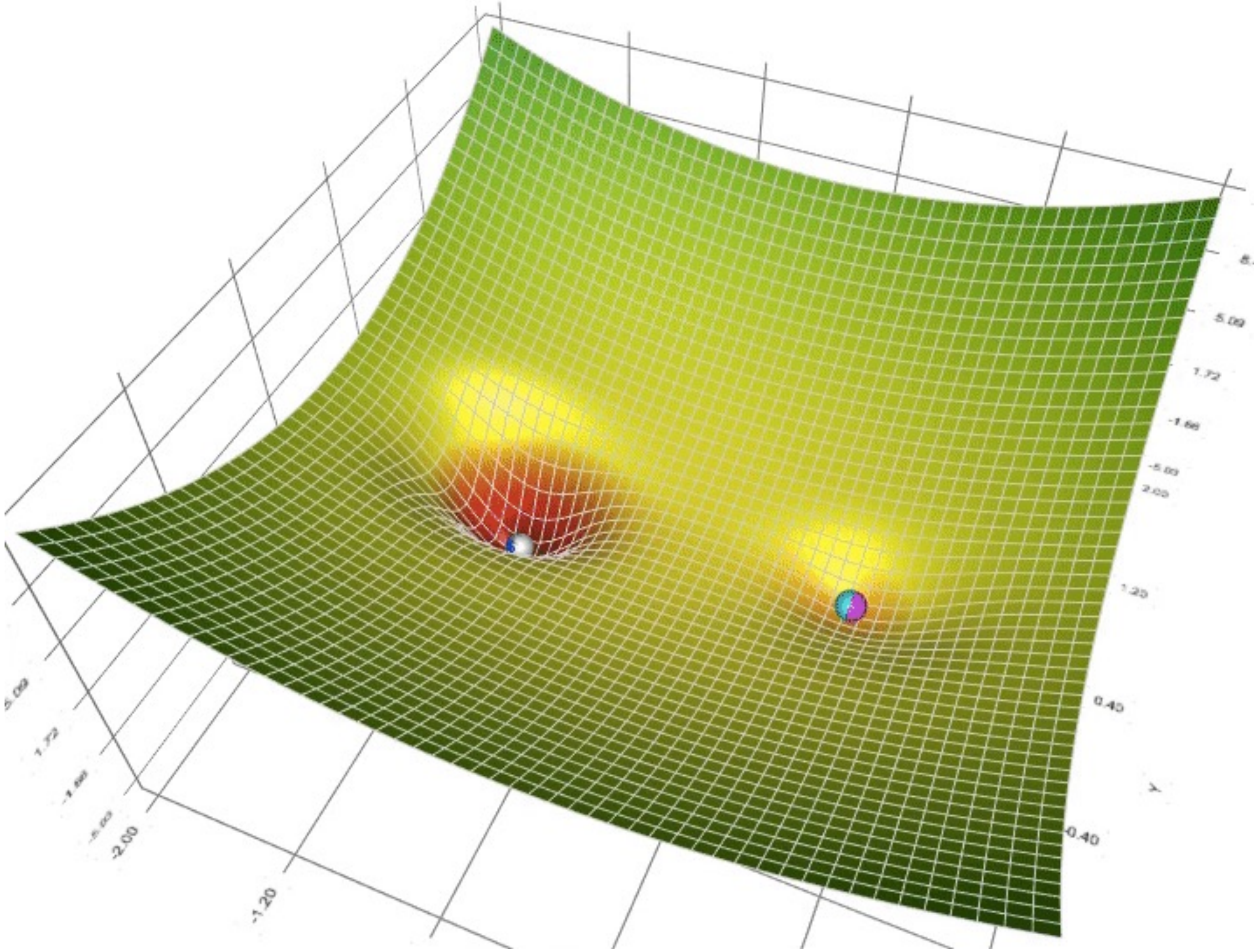
- Understanding AI & Machine Learning
- Types of Machine Learning
- Key Concepts in Machine Learning
- Dealing with Data
- Finding the Optimal Model

Understanding AI & ML


Artificial Intelligence
...
Symbolic AI
Machine Learning

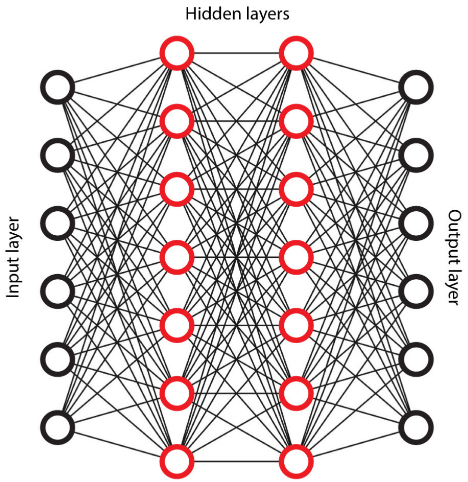
"Machine Learning is a set of **algorithms** that improve their **performance** on a set **task** through **experience**."

Statistical Methods & Numerical Optimisation

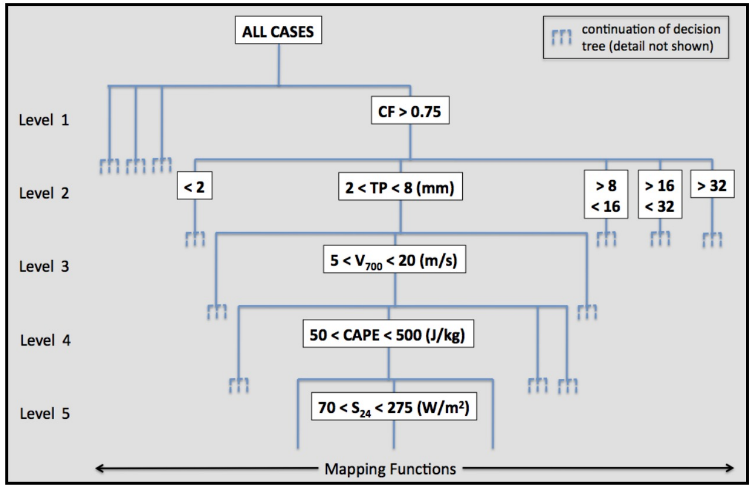


Many Models – Many Methods

Neural Networks

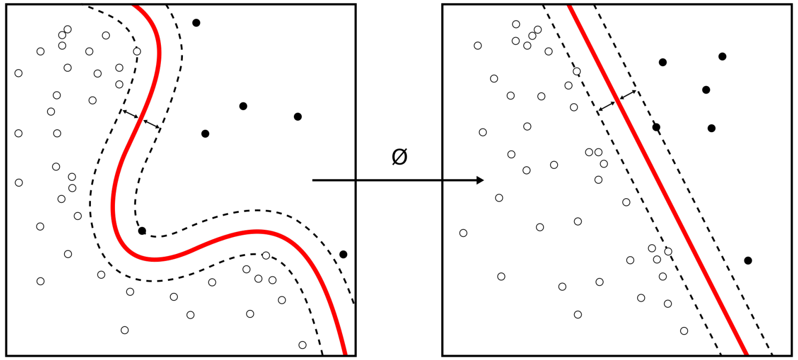


Decision Trees



Hewson and Pillosu 2020

Support-Vector Machine



Why I'm keeping the details on models short

- Modern software packages make machine learning easy!

Choose a model

```
>>> from sklearn import svm
>>> clf = svm.SVC(gamma=0.001, C=100.)
```

Fit the model to training data

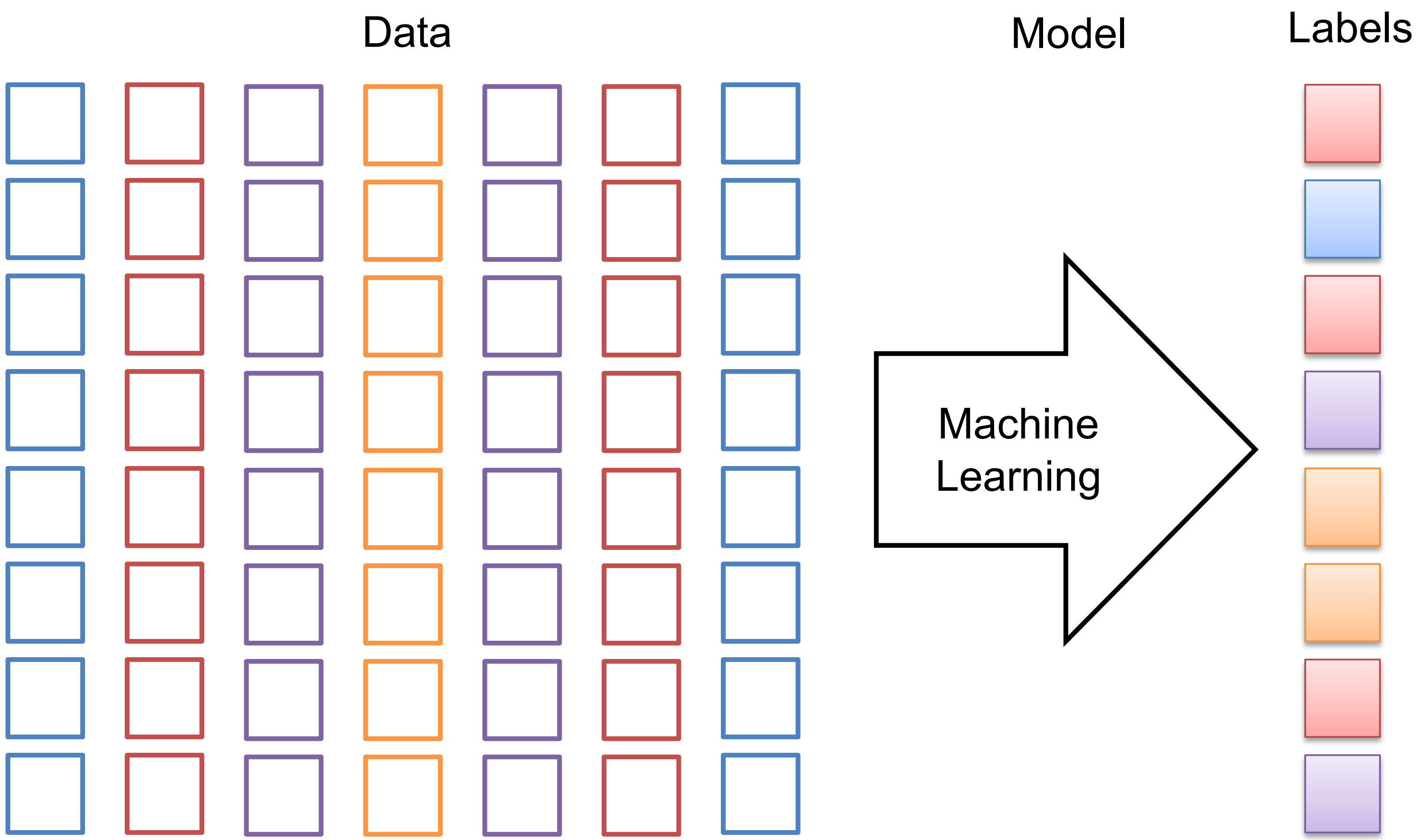
```
>>> clf.fit(digits.data[:-1], digits.target[:-1])
SVC(C=100.0, gamma=0.001)
```

Use model to predict

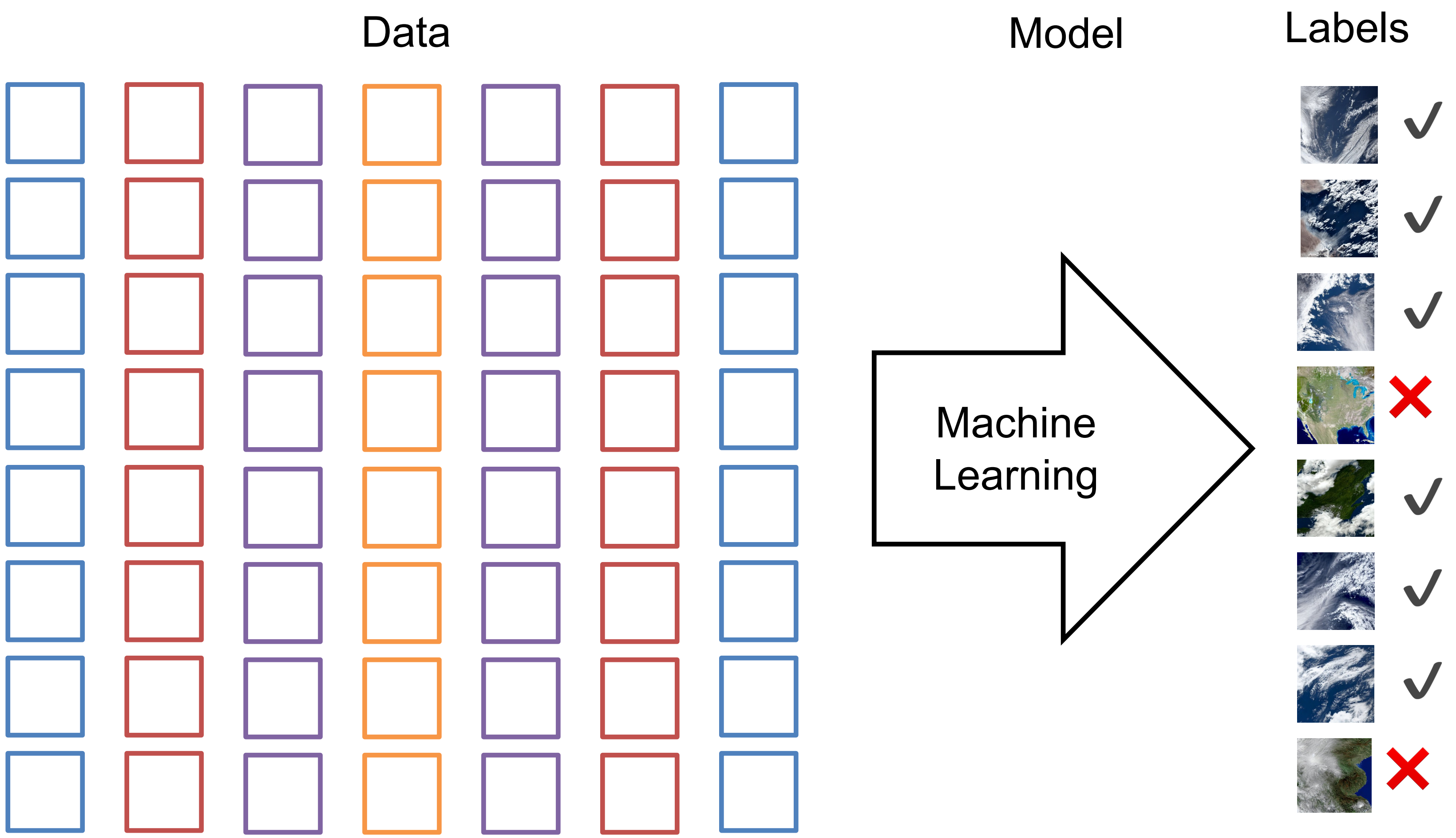
```
>>> clf.predict(digits.data[-1:])
array([8])
```

Types of Machine Learning

Supervised Learning

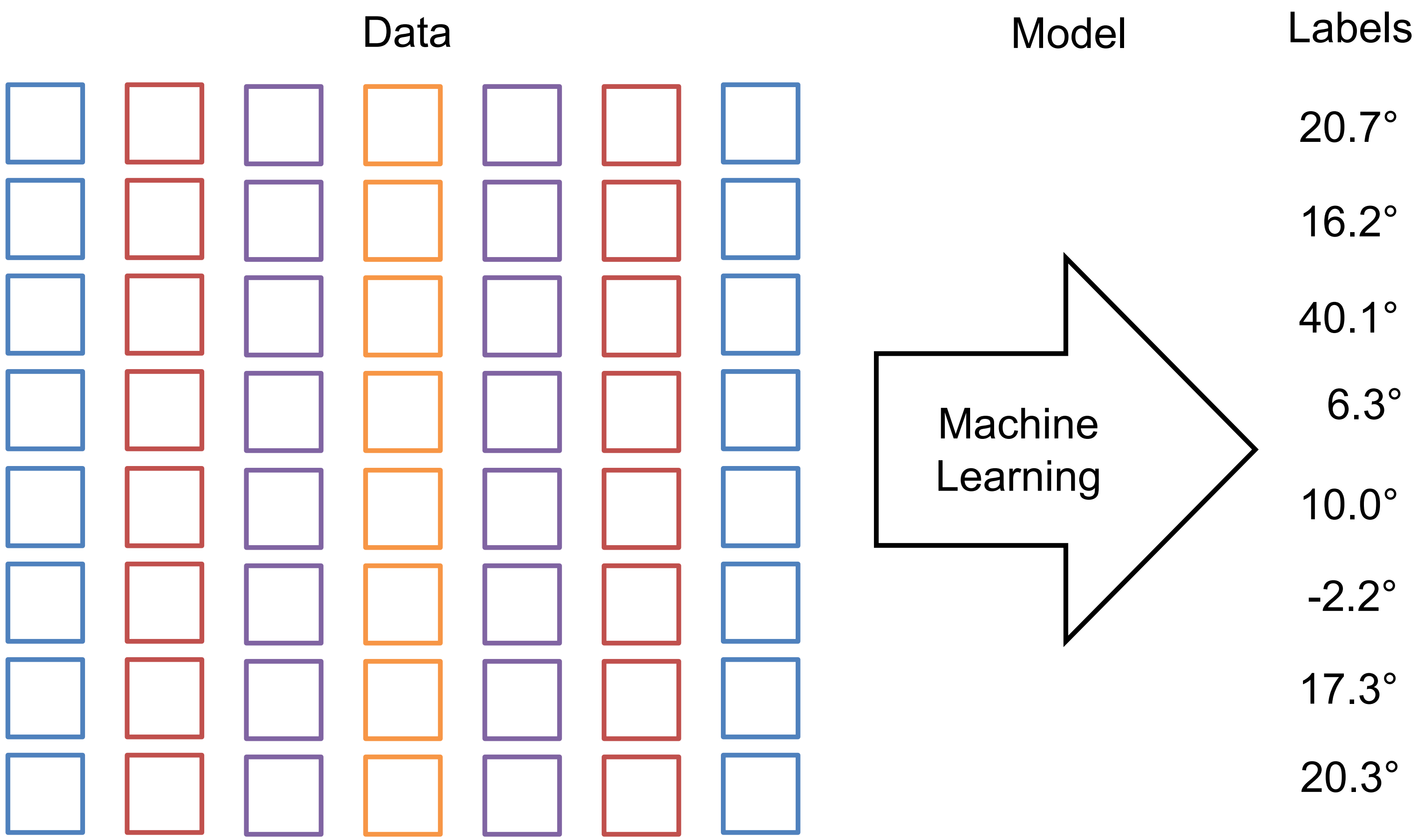


Supervised Learning – Classification

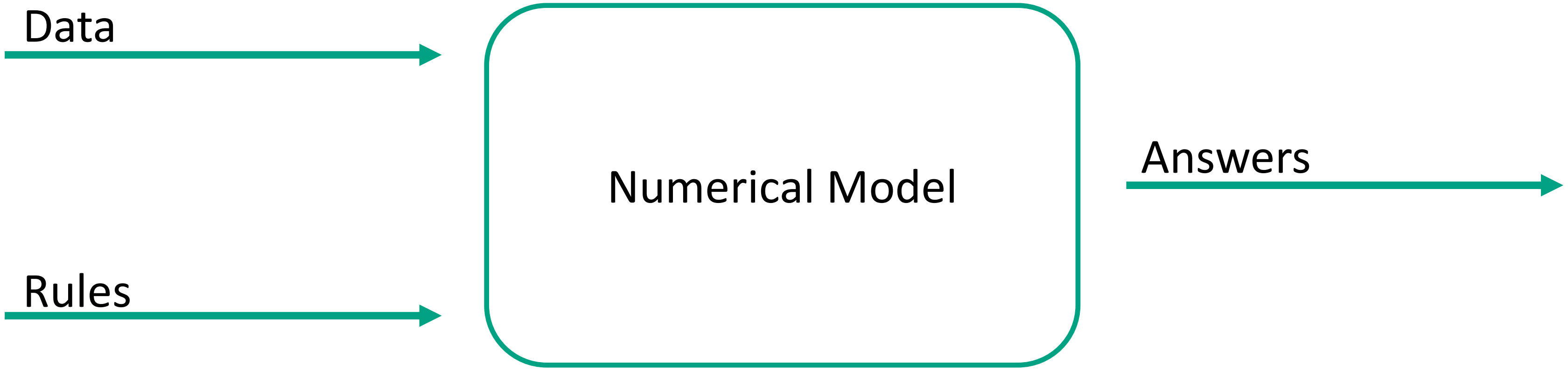


ECMWF

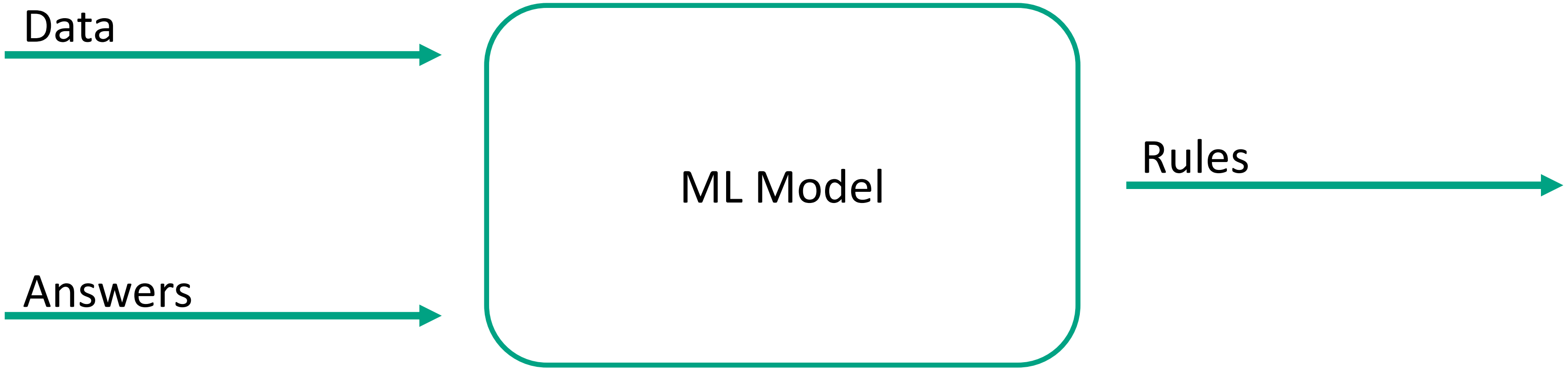
Supervised Learning – Regression



Classical Modelling

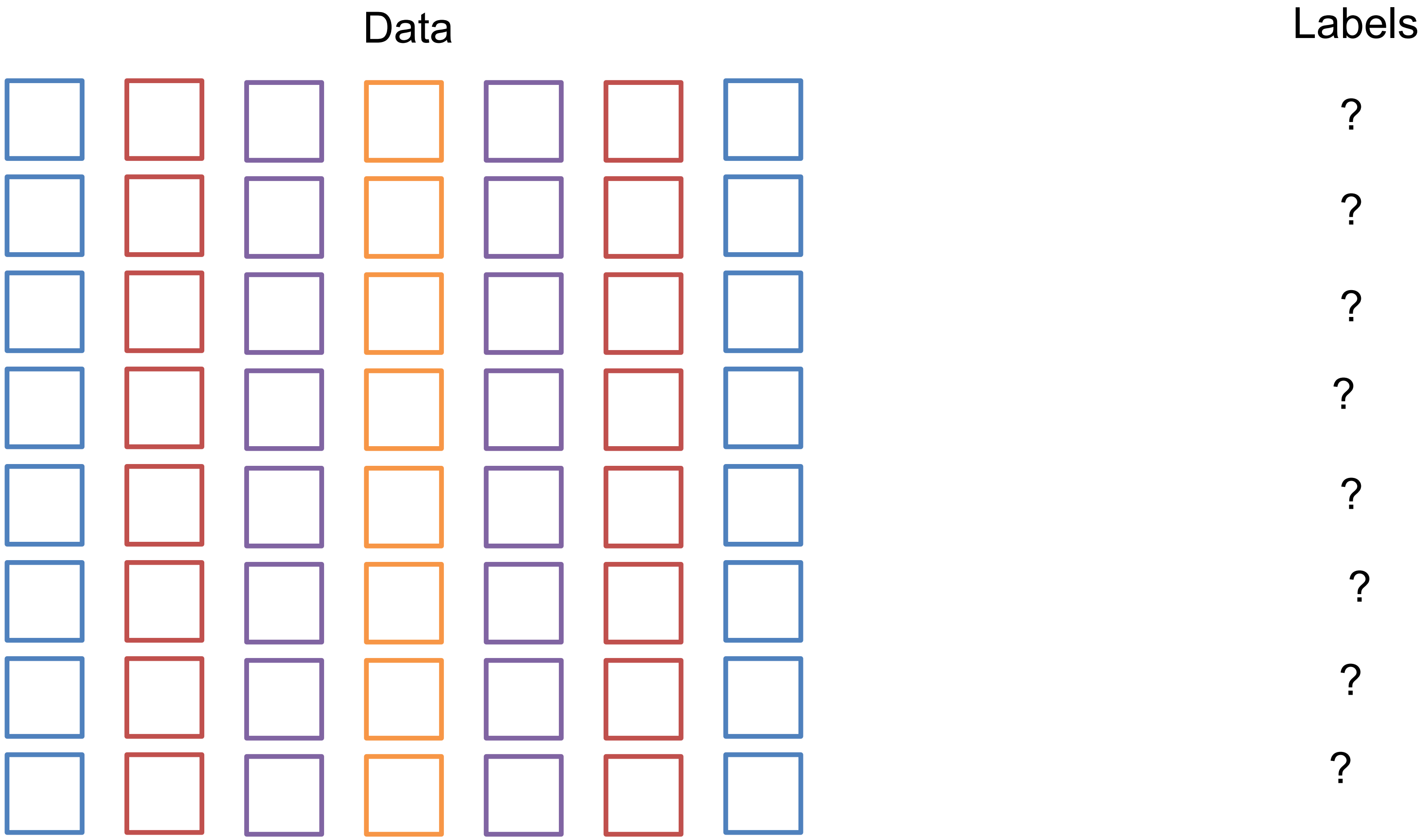


Supervised Machine Learning Modelling



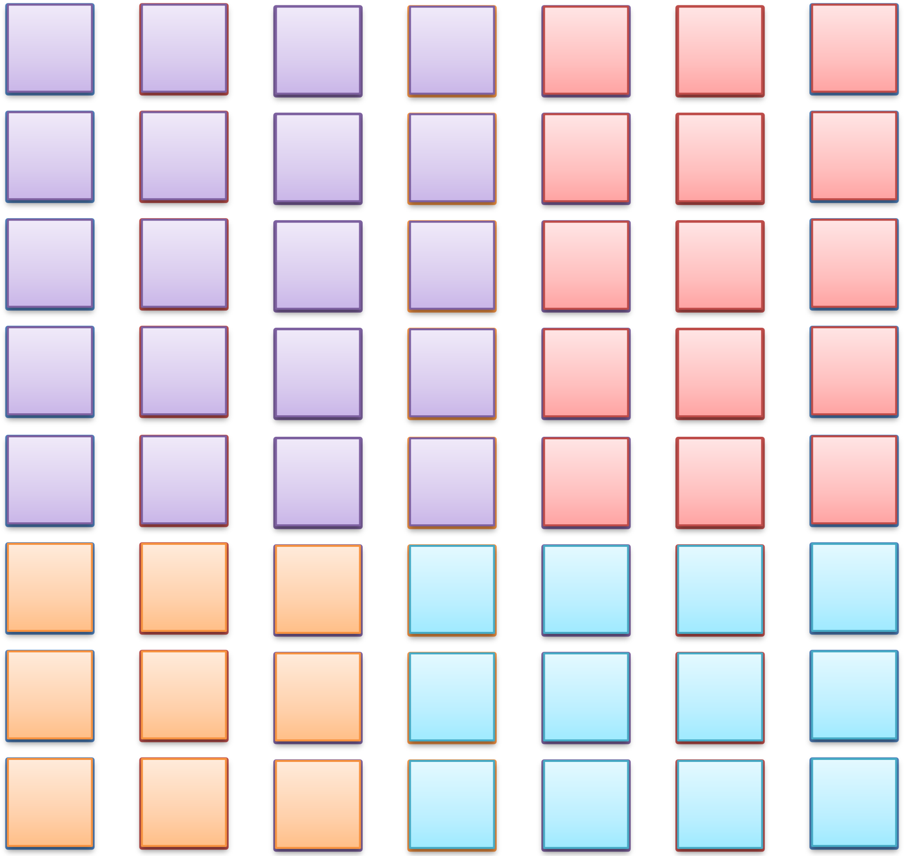
ECMWF

Unsupervised Learning



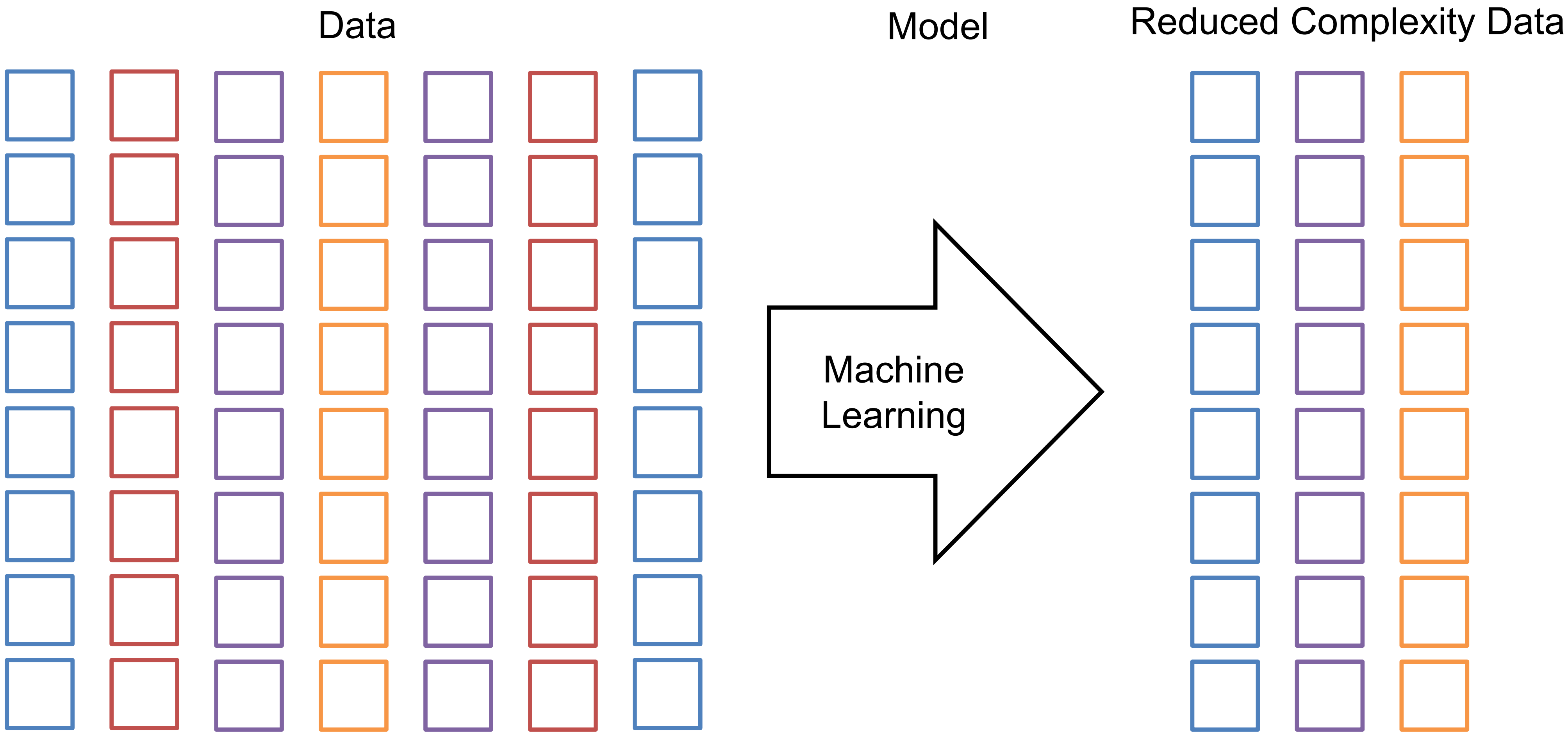
Unsupervised Learning – Clustering

Data



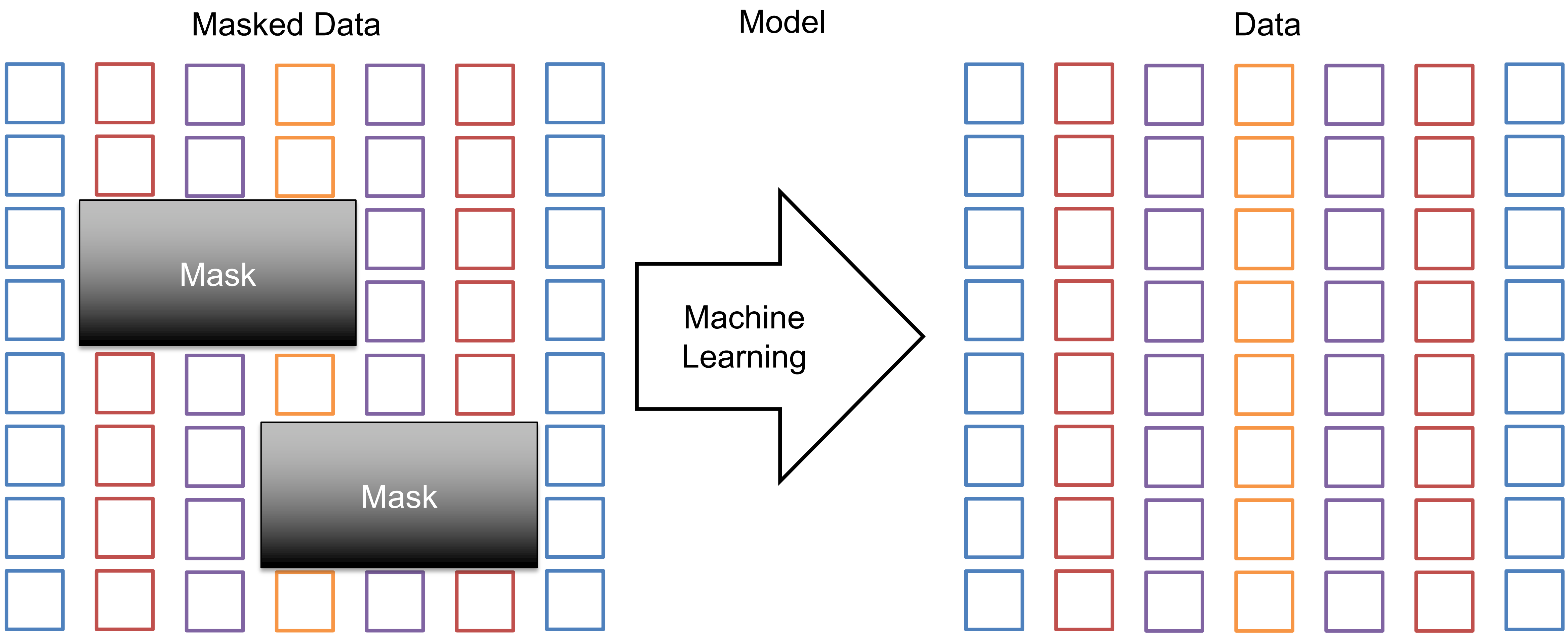
ECMWF

Unsupervised Learning – Dimensionality Reduction



ECMWF

Unsupervised Learning – Self-supervision

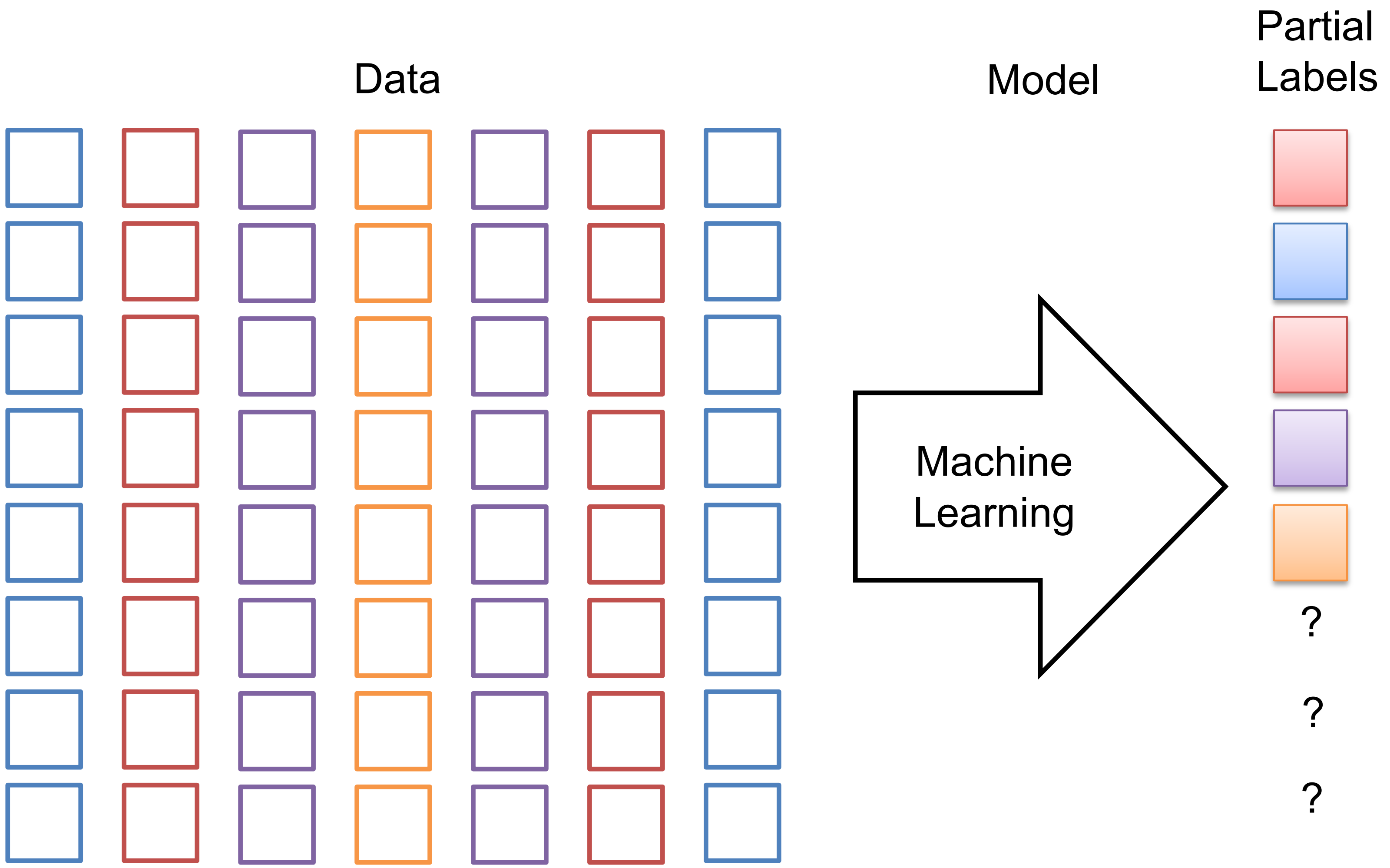


ECMWF

Unsupervised Machine Learning

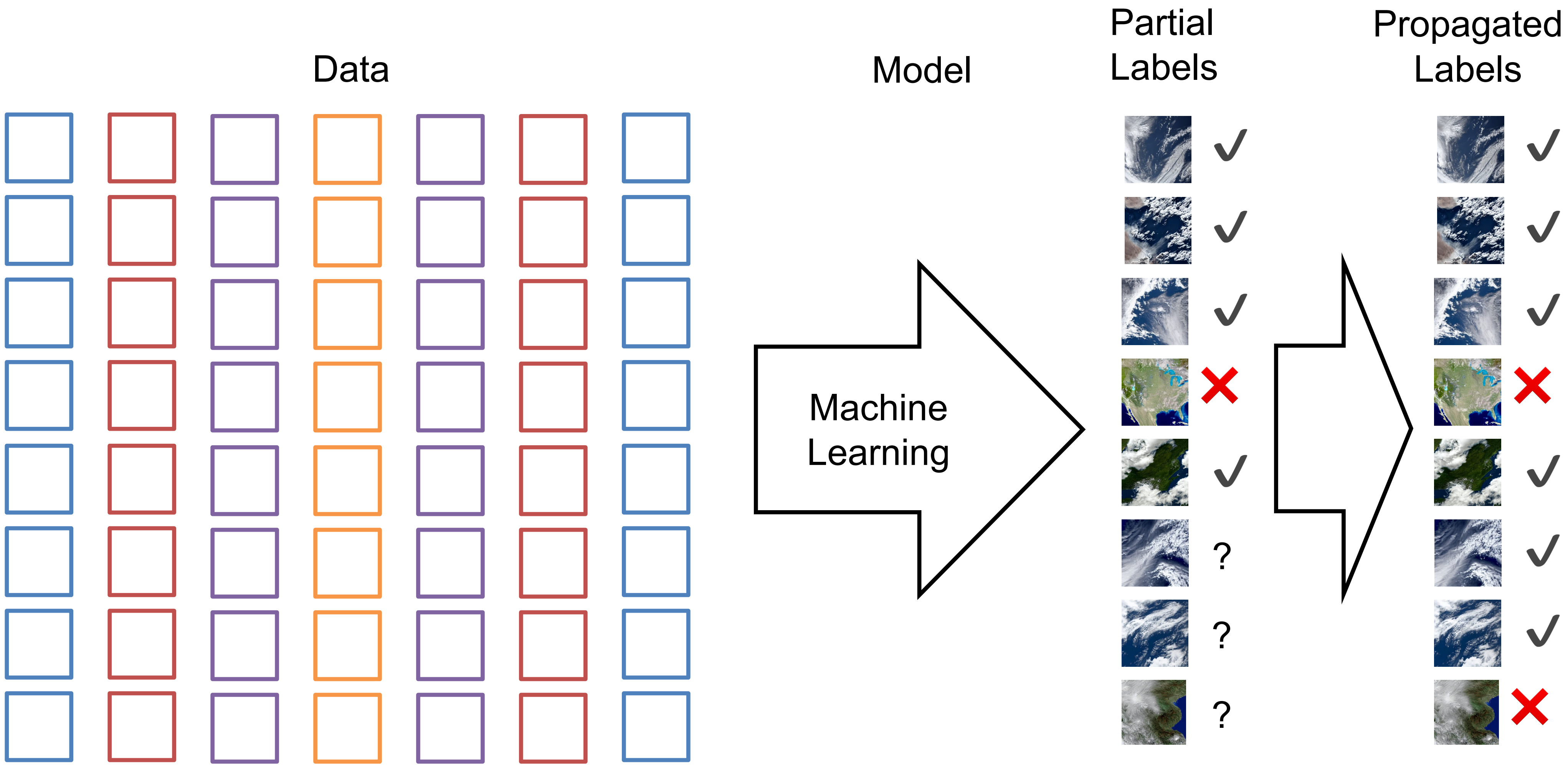
- Unlabeled data
  - Labeling needs expertise and is expensive
  - Labeling can introduce bias
- Exploits the internal structure of data
- Can accomplish different tasks
  - Assign Labels
  - Reduce complexity of data
  - Fill missing parts of data

Semi-Supervised Learning

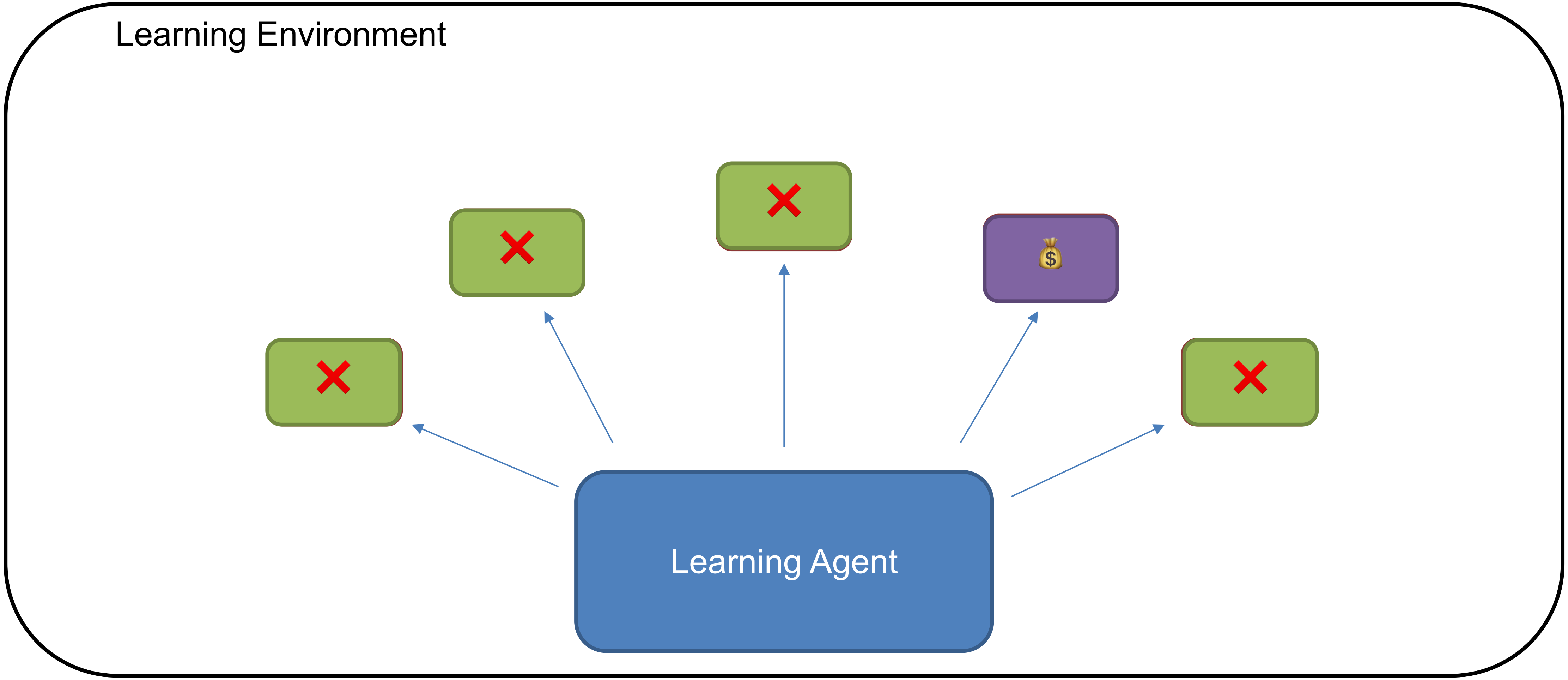


ECMWF

Semi-Supervised Learning – Cloud Classification



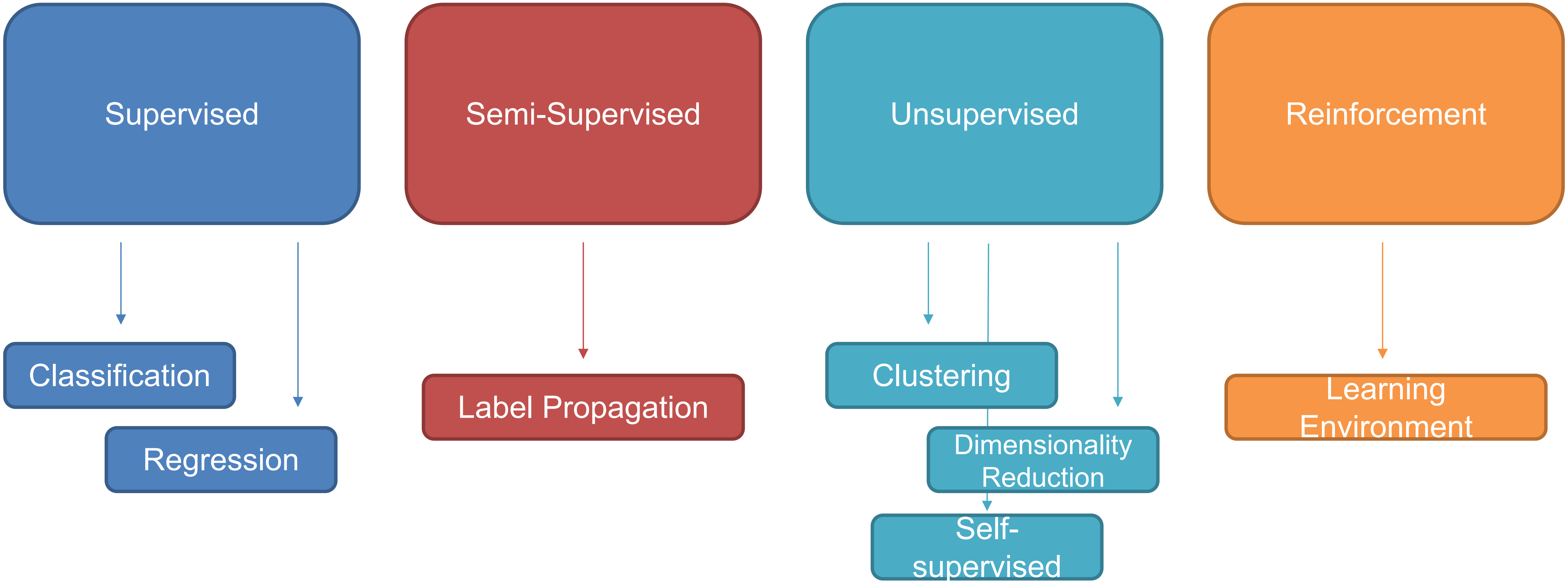
Reinforcement Learning



Reinforcement Learning – Games

Reinforcement Learning – Real World

The Types of Machine Learning



Other “Learning” which is not a “Type”

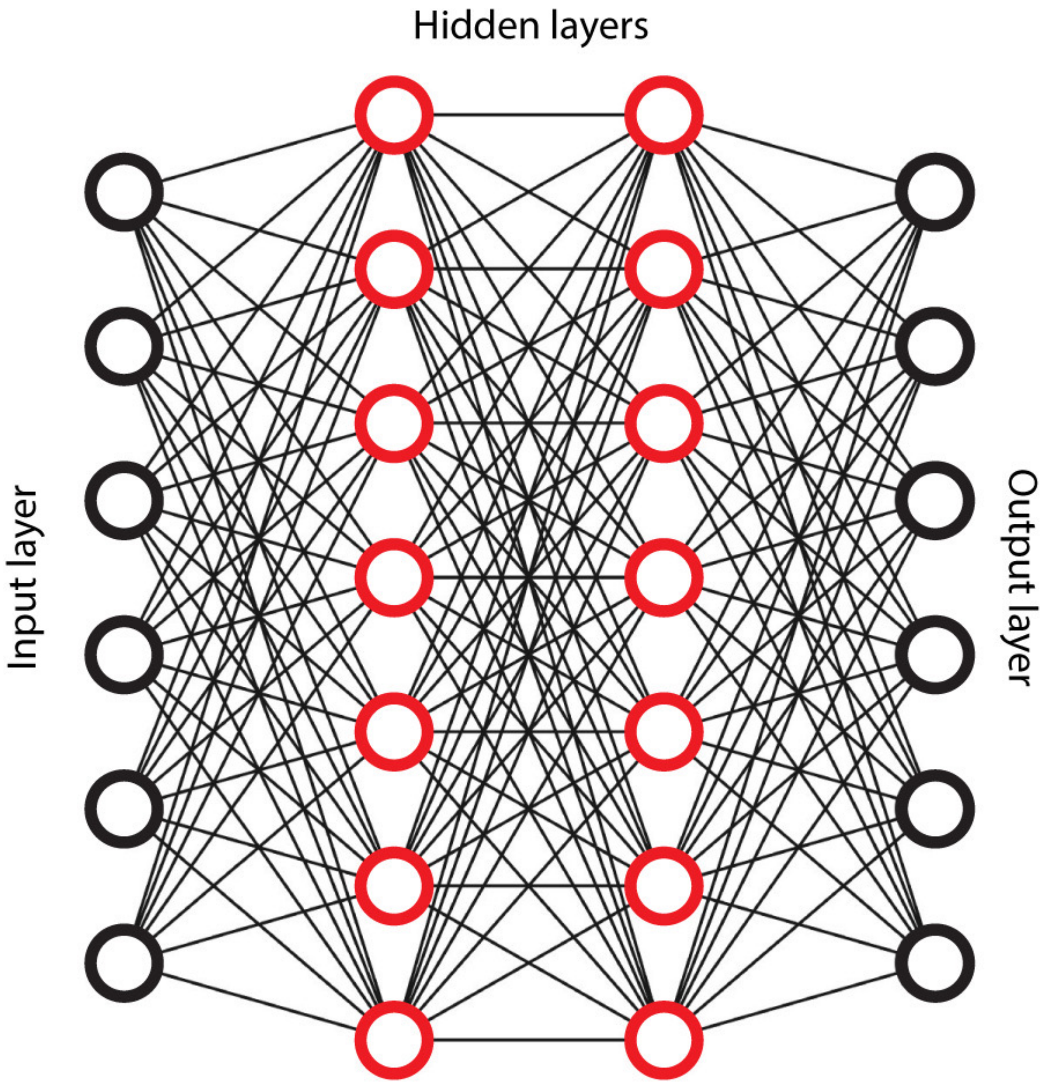
Deep learning and artificial neural networks as one example of machine learning

The concept:
Take input and output samples from a large data set
Learn to predict outputs from inputs
Predict the output for unseen inputs

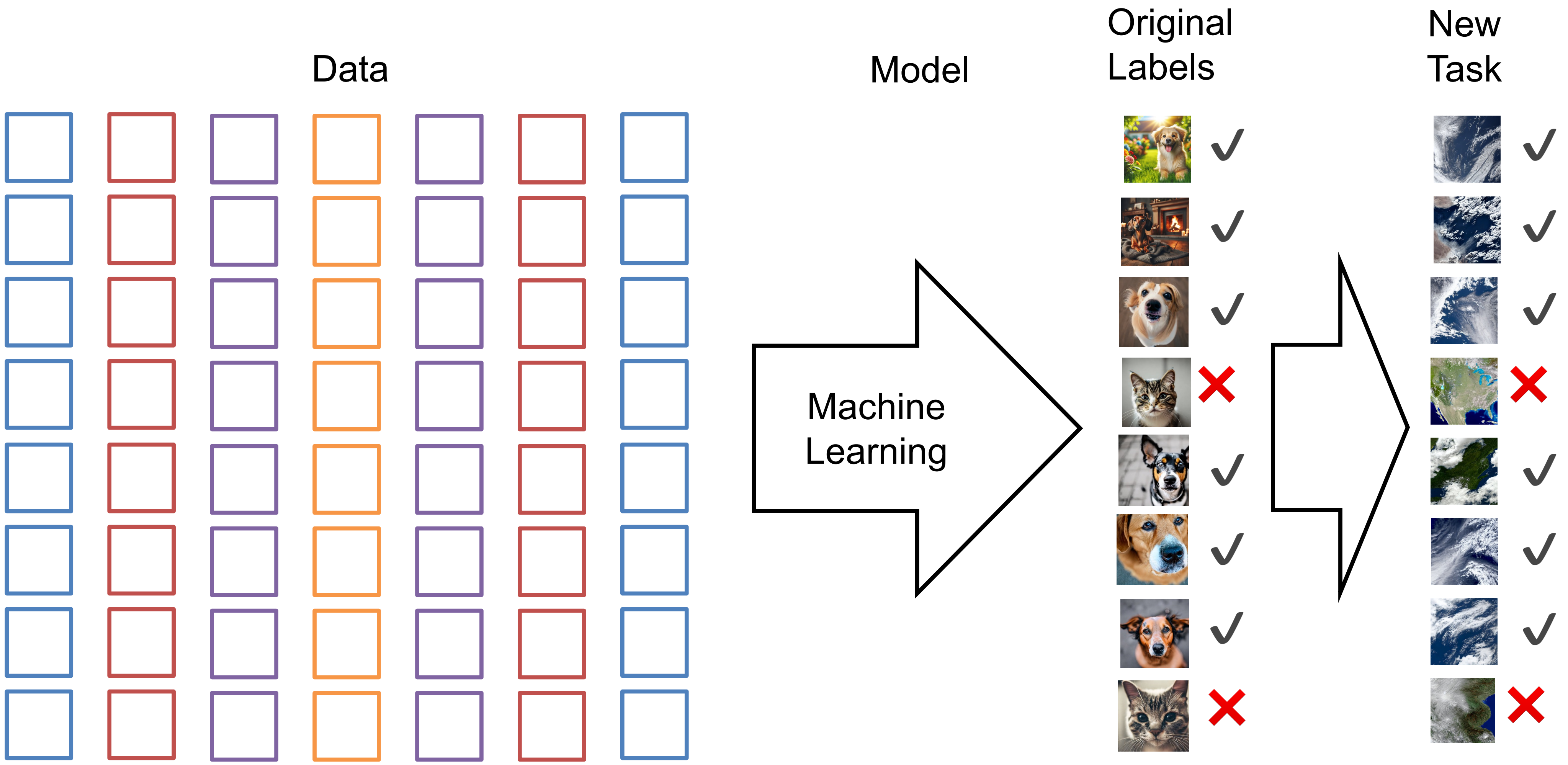
The key:
Neural networks can learn a complex task as a “black box”
No previous knowledge about the system is required
More data will allow for better networks

The number of applications is increasing by day:
Image recognition
Speech recognition
Healthcare
Gaming
Finance
Music composition and art
…

And weather/climate!



Transfer Learning and Domain Adaptation



ECMWF

Key Concepts in Machine Learning

Machine Learning Evaluation - Classification

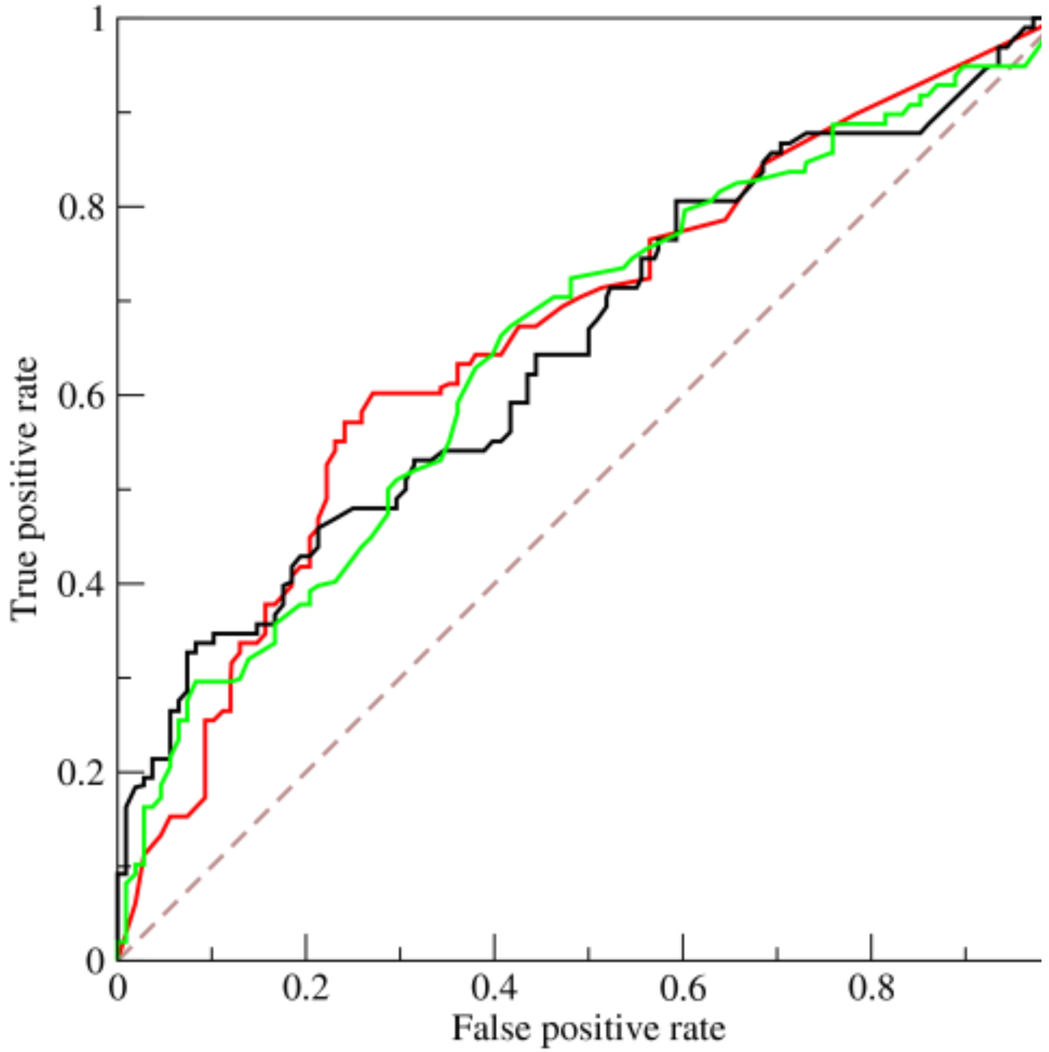
- Measure performance of model against ”answers we know”
- Confusion Matrix
  - True Positive
  - True Negative
  - False Positive
  - False Negative
  - Works with Multi-class
- Class Imbalance skews results

Predicted

Actual	Dog	Not Dog	Third Label
Dog	50	5	1
Not Dog	5	40	1
Third Label	1	10	1

Machine Learning Evaluation - Classification

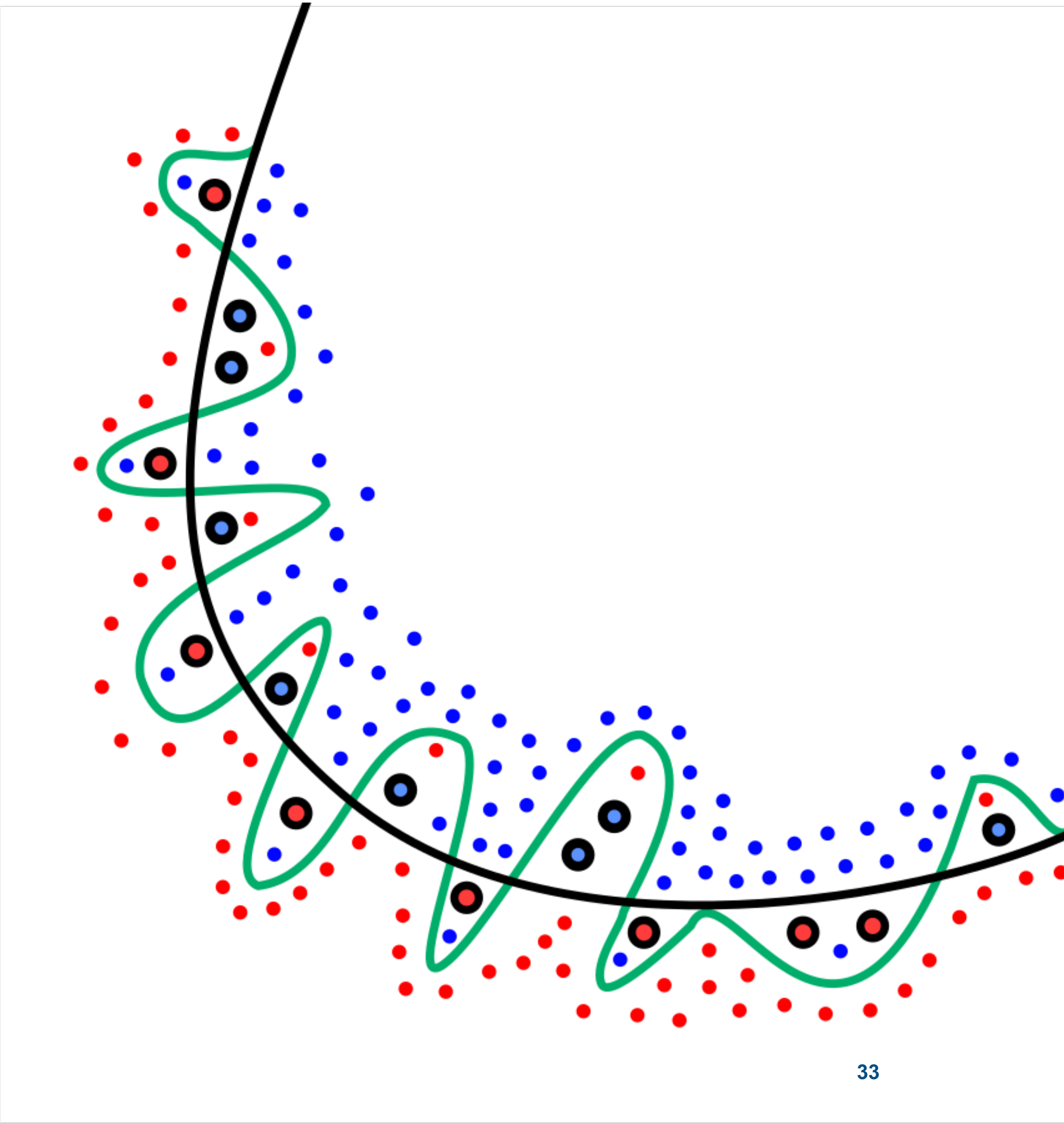
- Receiver Operator Characteristic (ROC)
  - Balances acceptable false positive rate with desired true positive rate
- Used to define class thresholds
- Works on balanced data
  - For imbalanced use Precision-Recall curves
- ML models rise and fall by their metrics



“Generalization is a ML **model’s ability** to generate accurate and reliable predictions on previously **unseen data**.”

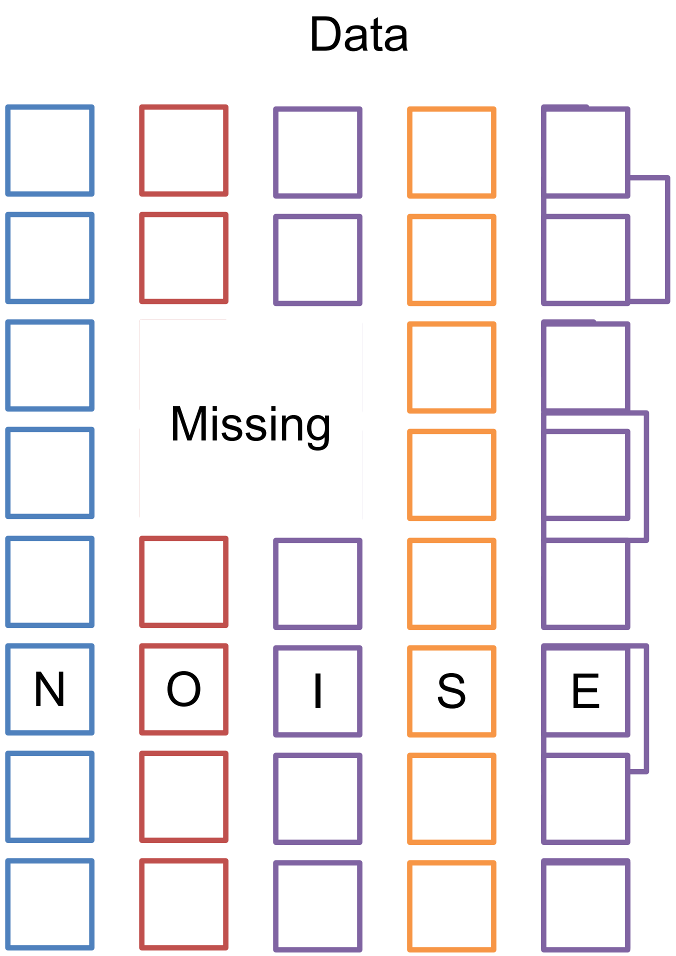
Generalization and Overfitting

- ML model learns from historic data
- Generalization for performance on unseen data
- Underfitting
  - Model can't fit the complex data
- Overfitting
  - Model exactly fits the training data
  - Does not generalize to unseen data
- Overfitting can be avoided by
  - Reducing model complexity
  - Regularization
  - Pruning
  - Etc.



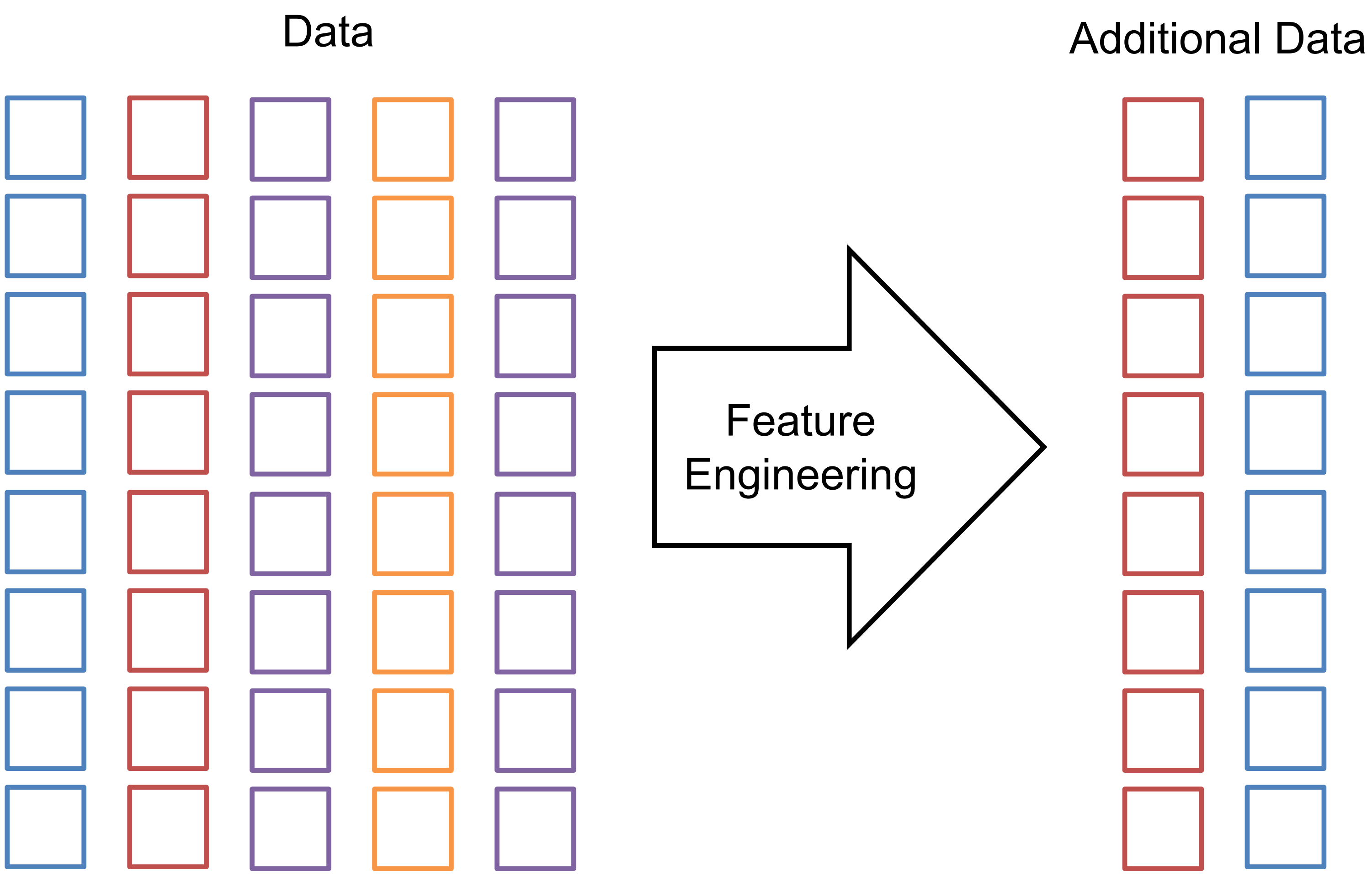
Dealing with Data

Data Preprocessing



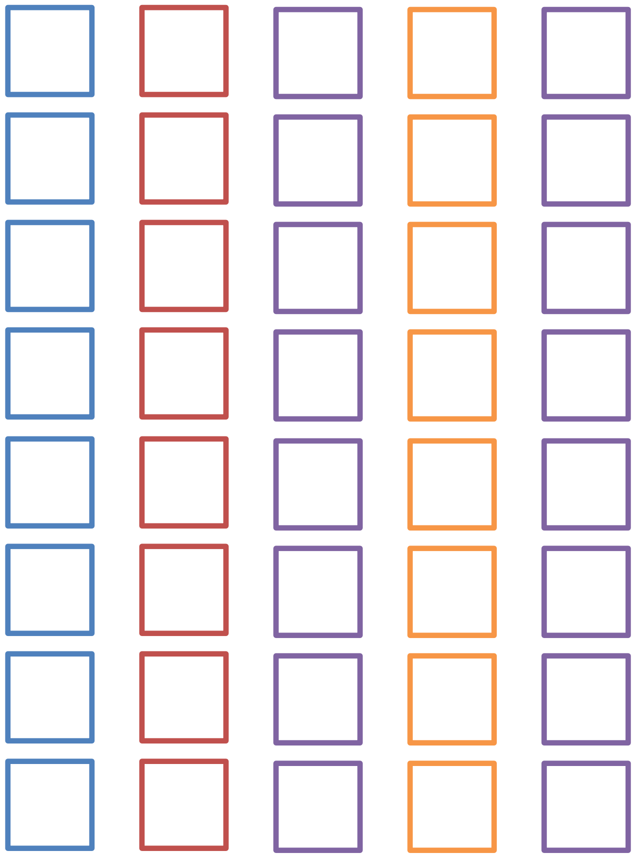
- Machine learning models struggle with irregular data
- Imputation
  - Filling in missing values
  - Often with Mean or Median
- Data Cleaning
  - Removing noise from data
  - Careful! Easy to ”over-clean”
  - Needs to be faithful to real-world data
- Normalisation
  - Standardization
  - Min-Max Scaling
- Transformations
  - Log-Scaling

Feature Engineering



Incremental Learning / Batch Processing

Big Data



Batch 1

Batch 2

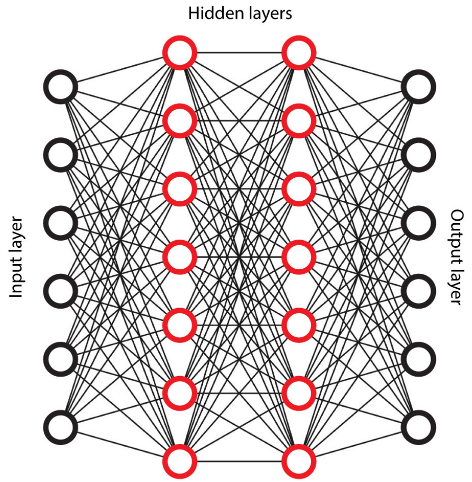
Hyperparameters and finding the optimal model

“There is no **Free Lunch**.”

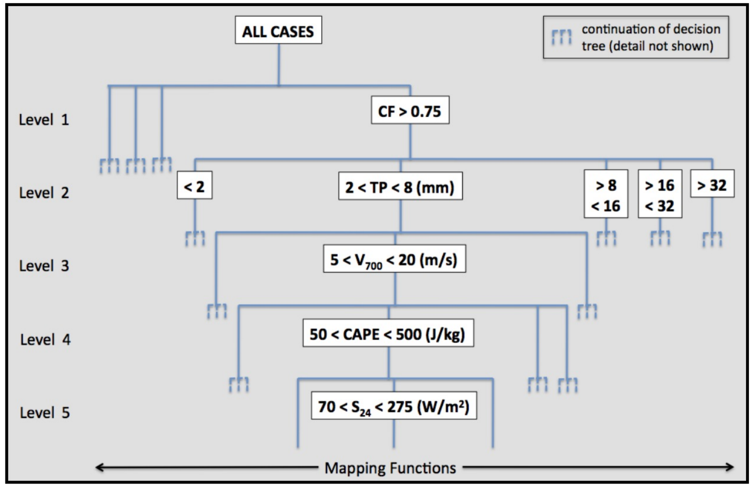
Hyperparameters and Tuning

- Parameters:
  - Statistical Term
  - E.g. Parametric / non-parametric model
- Hyperparameters:
  - ”Settings of Model”
- Examples of Hyperparameters:
  - Number of nodes
  - Number of layers
  - Number of Trees
  - Learning Rate of optimization process
  - Batch size, of incremental training

Neural Networks

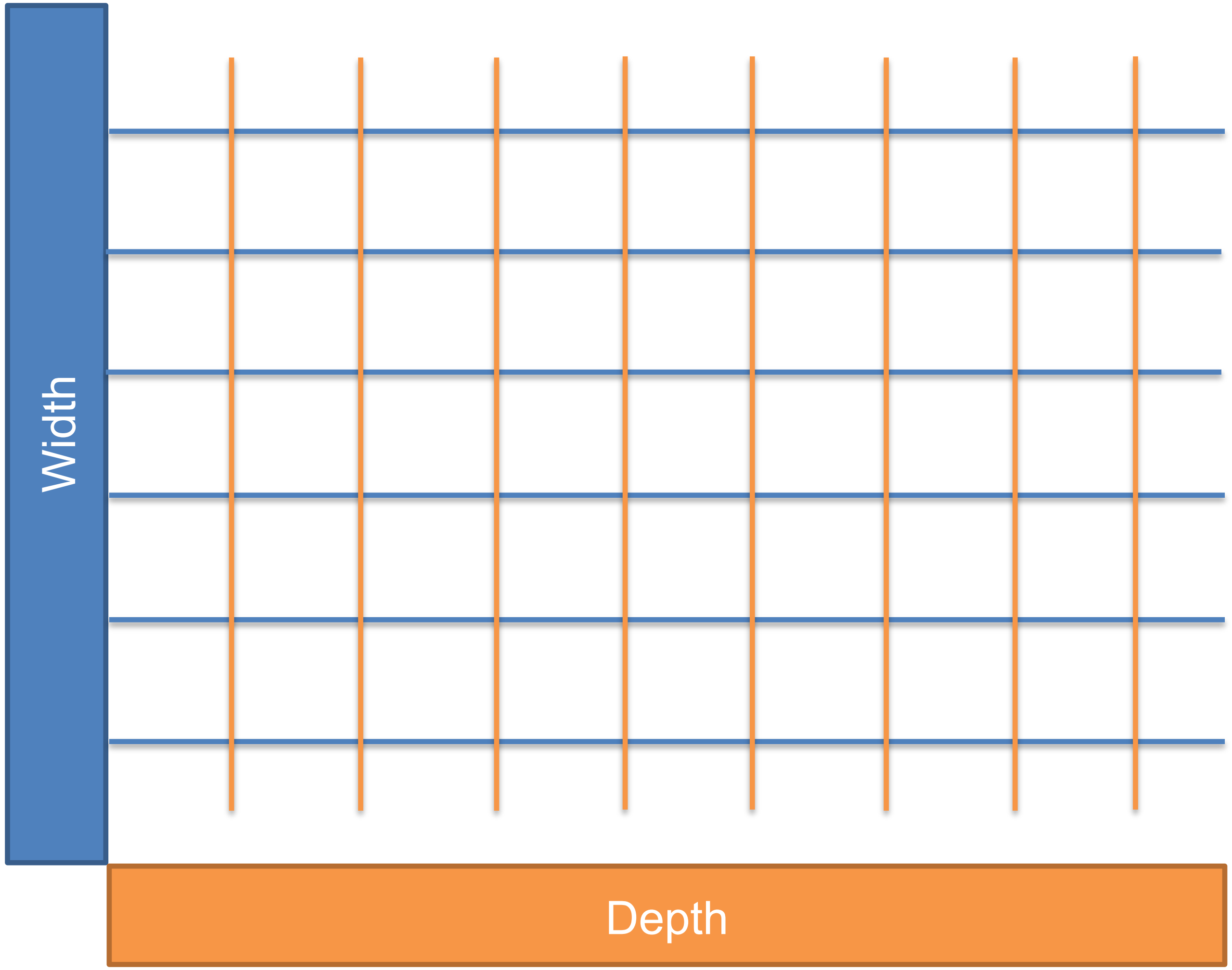


Decision Trees



Hewson and Pillosu 2020

Grid search



- Exhaustive Search
- Every Combination is Evaluated
- Combinatoric Explosion of Evals
- Inefficient searching beyond minimum
- Possible to miss optimal parameters

because explicit values are provided

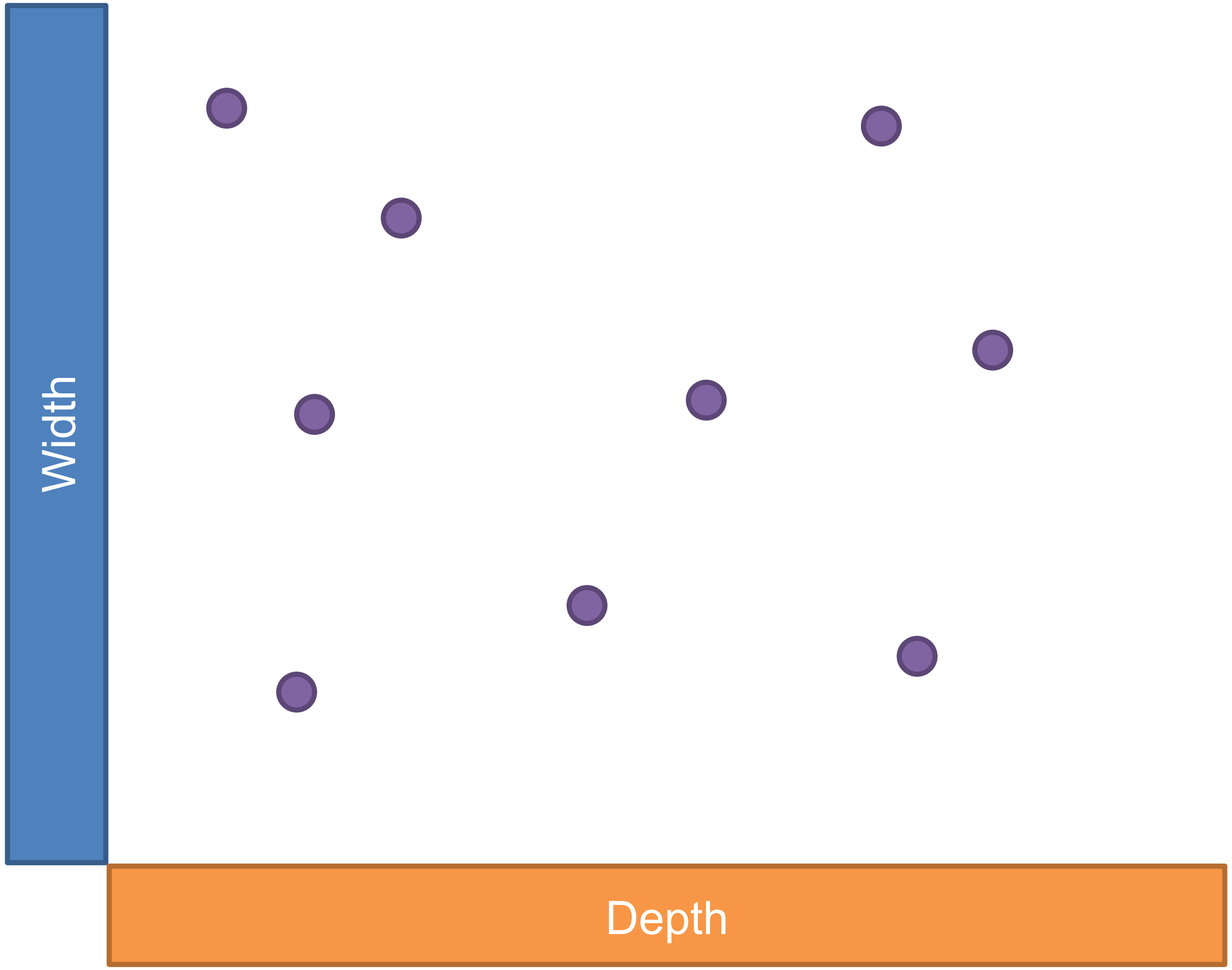
```python
from sklearn.model_selection import GridSearchCV

parameters = {'width':[5, 10, 15, 20],
              'depth':[1, 2, 3, 4, 5, 6],
              'activation':['tanh', 'relu']}

gridcv = GridSearchCV(neural_network, parameters)

gridcv.fit(X_train, y_train)
```

Randomized search



- Exhaustive Search
- Budget independent of No. parameters
- Adding Parameters not Inefficient
- Inefficient searching beyond minimum
- Possible to miss optimal parameters

because explicit values are provided

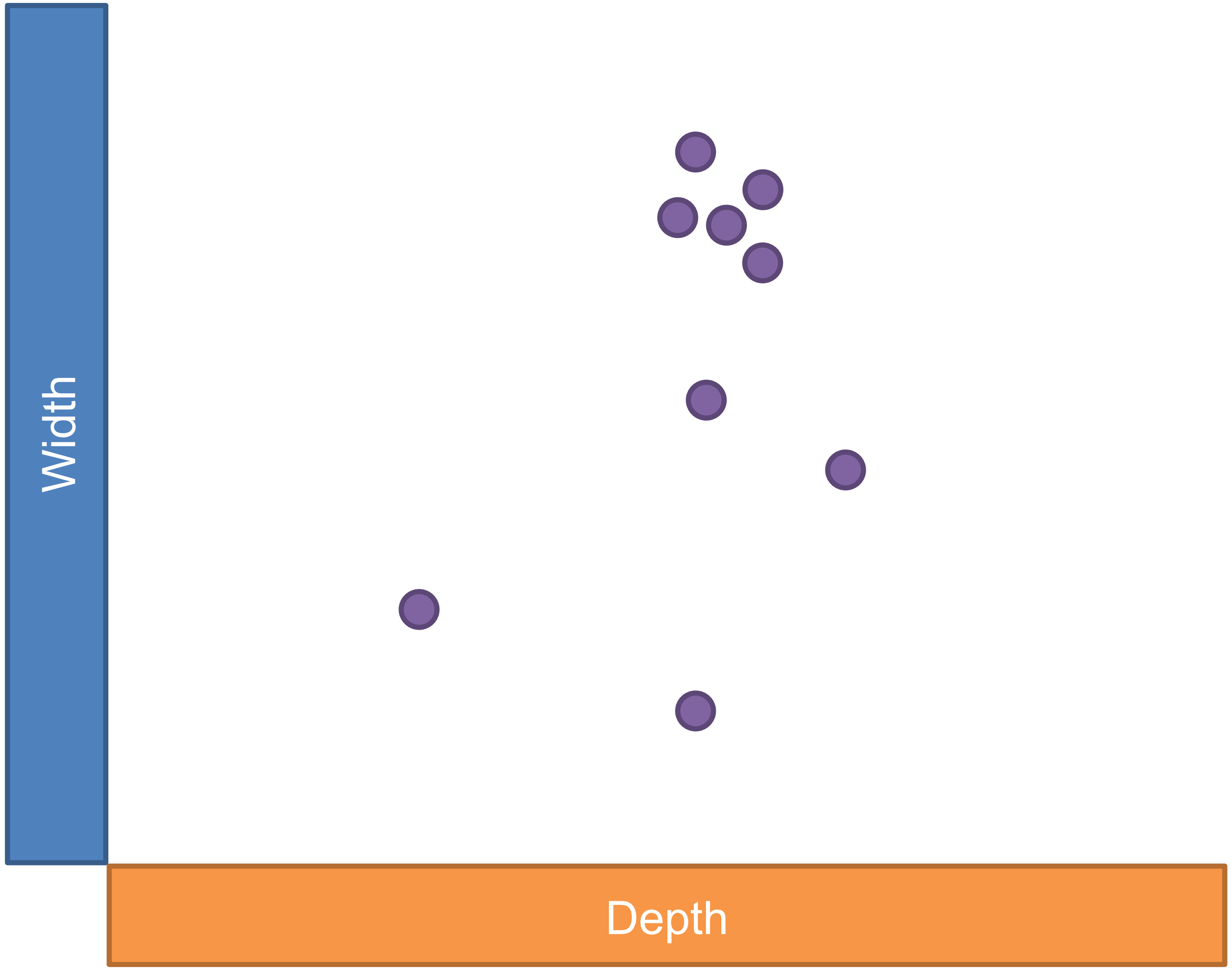
```python
from sklearn.model_selection import RandomizedSearchCV
from scipy.stats import uniform

distributions = {'width': uniform(5, 15),
                 'depth': uniform(1, 5),
                'activation':['tanh', 'relu']}

randomcv = RandomizedSearchCV(neural_network, distributions)

randomcv.fit(X_train, y_train)
```

Bayesian search



- Search based on former parameters
- Bayesian Optimization
- Converges to a minimum
- Adding Parameters adds complexity
- Unimportant parameters complicate optimization significantly

```python
from skopt import BayesSearchCV

distributions = {'width': (5, 20, 'uniform'),
                 'depth': (1, 6, 'uniform'),
                 'activation':['tanh', 'relu']}

randomcv = BayesSearchCV(neural_network, distributions)

randomcv.fit(X_train, y_train)
```

Conclusion

What we Learned

- AI and Machine Learning are related but distinct
- Open-source software makes ML easier
- Types of machine learning model:
  - Un-, Semi-, Supervised learning
  - Reinforcement Learning
- Other relevant “Learning”
  - Deep Learning
  - Transfer Learning
- Generalization and Overfitting
- Data-Preprocessing
- Hyperparameter tuning