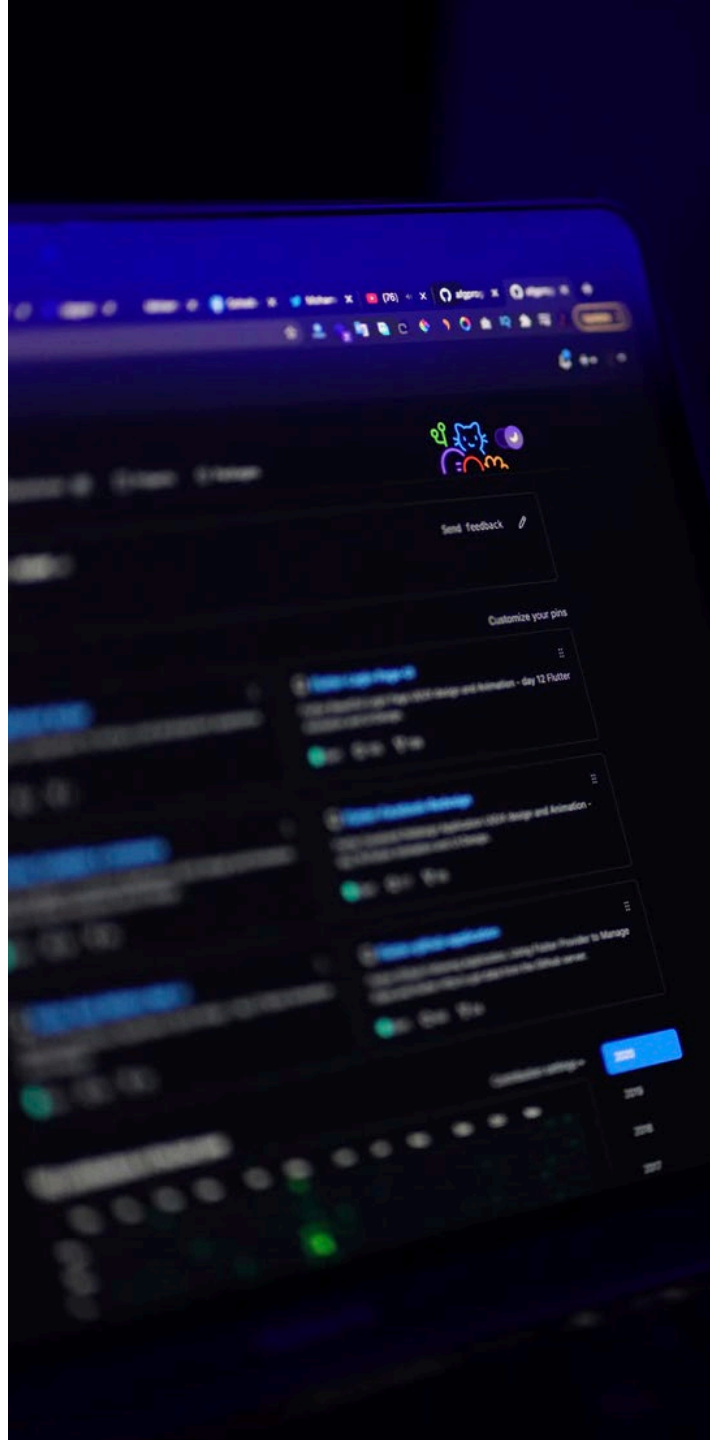
UK Data Service

Introduction to Machine Learning:
Session 1



Nadia Kennar, Research Associate at the UK Data Service (Cathie Marsh Institute, UoM)

UKRI Economic and Social Research Council

Session 1: 24 October

- Context:
  - What is machine learning?
  - What is a model?
- Machine learning methods
- How is ML different from classic statistics
- 7 Steps to Machine Learning?

Session 2: 26 October

- Exploring a specific unsupervised method, clustering
- Centroid based: specifically, k-means algorithm
- Hierarchical-based: divisive (top-down) and agglomerative (bottom-up)

Session 3: 1 November

- Live Code Demo
- Using a dataset to explore different clustering methods
- First hour in Python
- Second hour in R

What is Machine Learning ?

- Machine Learning is a subfield of *artificial intelligence (AI)*
- i.e. the capability of a machine to imitate intelligent human behaviour

” the field of study that gives computers the ability to learn without explicitly being programmed.”

Arthur Samual, 1950

Functions of a Machine Learning System

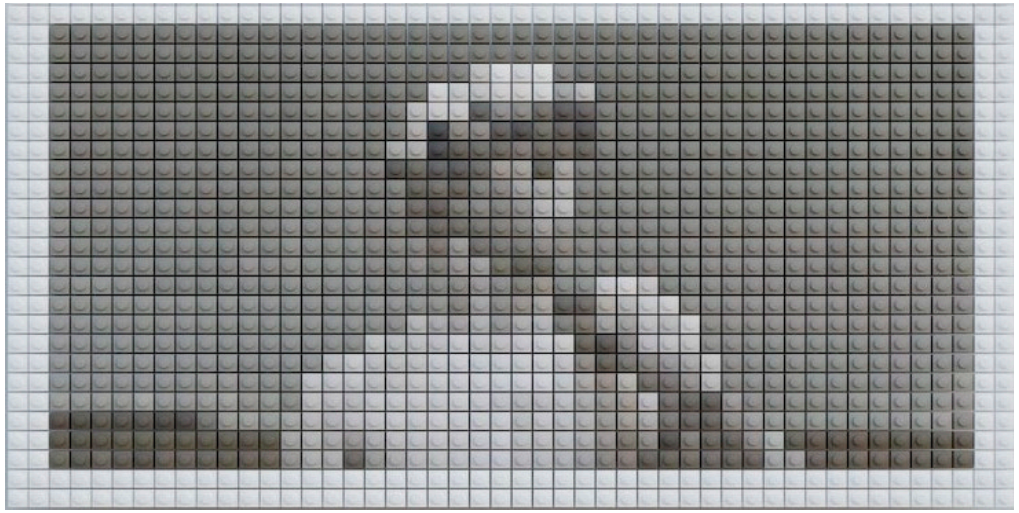
- **Descriptive:** uses the data to explain what happened
- **Predictive:** the system use the data to predict what will happen
- **Prescriptive:** the system will use the data to make suggestions about what action to take.

Credit: T.Molone, D.Russ, R.Laubacher

Why is machine learning important?

- Increase in volume and variety of data
- The access and affordability of computational power
- Mitigate risks and improve overall quality of life

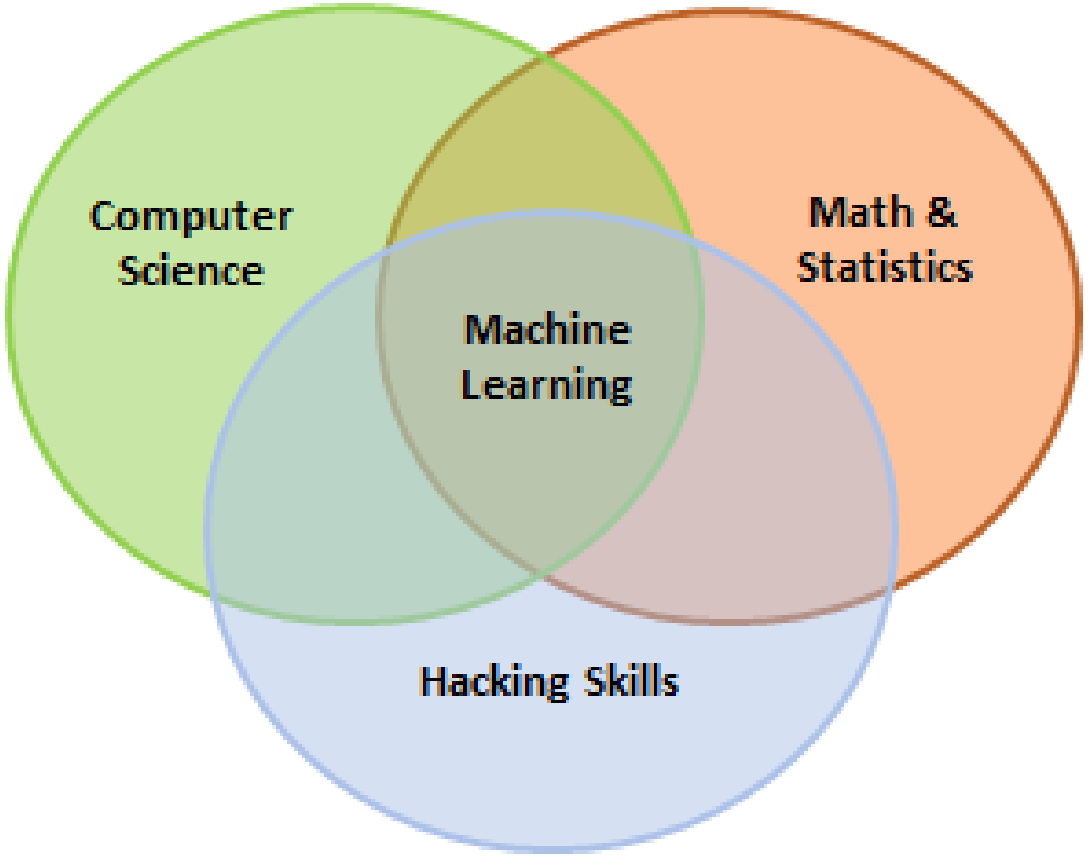
Why is machine learning important?



Machine Learning vs Classic Statistics

	ML	Statistics
Approach	Data driven (limit assumptions)	Algorithmic model
Driver	Fitting Data	Math, theory
Focus	Predictive Accuracy	Hypothesis Testing
Inference	Predictions	Parameter Estimation

The ‘gray area’

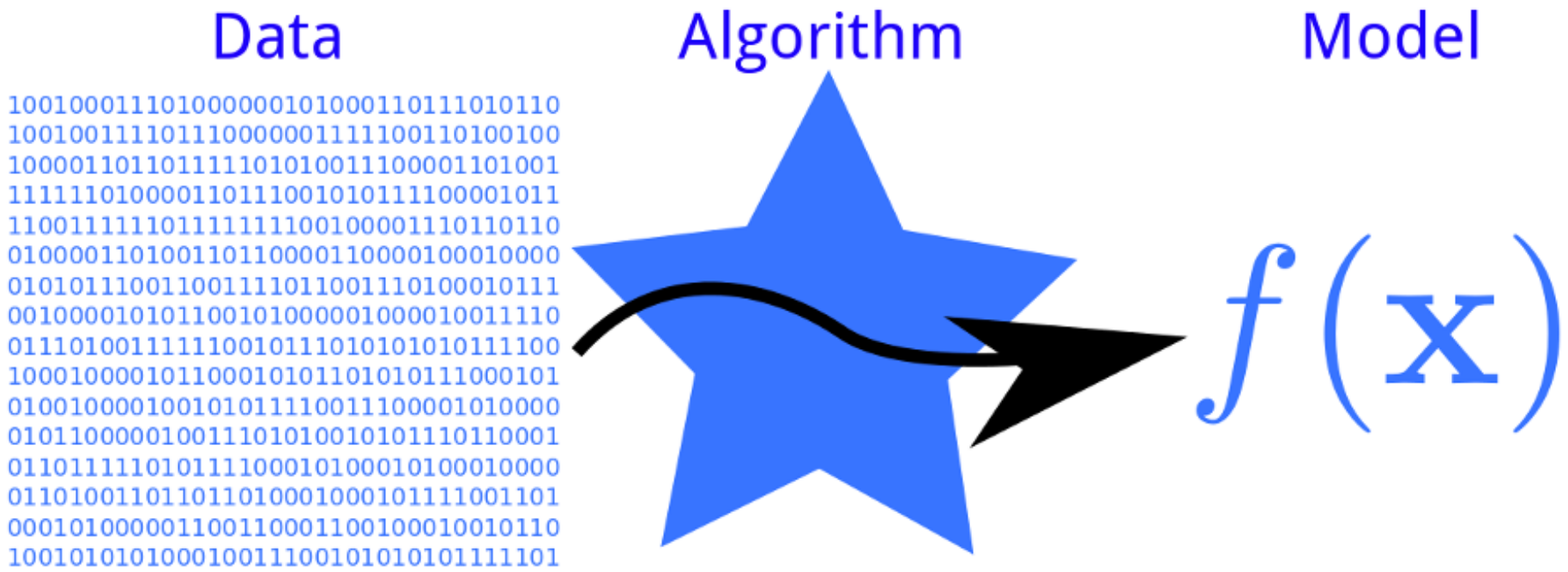


= Regression

Machine Learning Methods

- **Supervised Learning**
- **Unsupervised Learning**

Model vs Algorithm?



How does a Machine Learning Algorithm Work ?

- **A Decision Process** = the steps that takes in the data and guesses what kind of pattern your algorithm is looking to find.

- **A Loss/Error Function =** measure how good the guess was by comparing it other examples

- **A Model Optimization Process =** A method in which the algorithm looks at the miss and then updates the decision process

More about ‘models’

- Inductive nature
- Sample size
- Inductive bias and generalisation

What is Supervised Learning?

Supervise = “observe and direct the execution of (a task or activity)”

- How do you supervise a machine learning model?
- “Teach the model”
  - Load the model with knowledge then have it predict future instances

Continued…

- We teach the model by training it with a labelled dataset

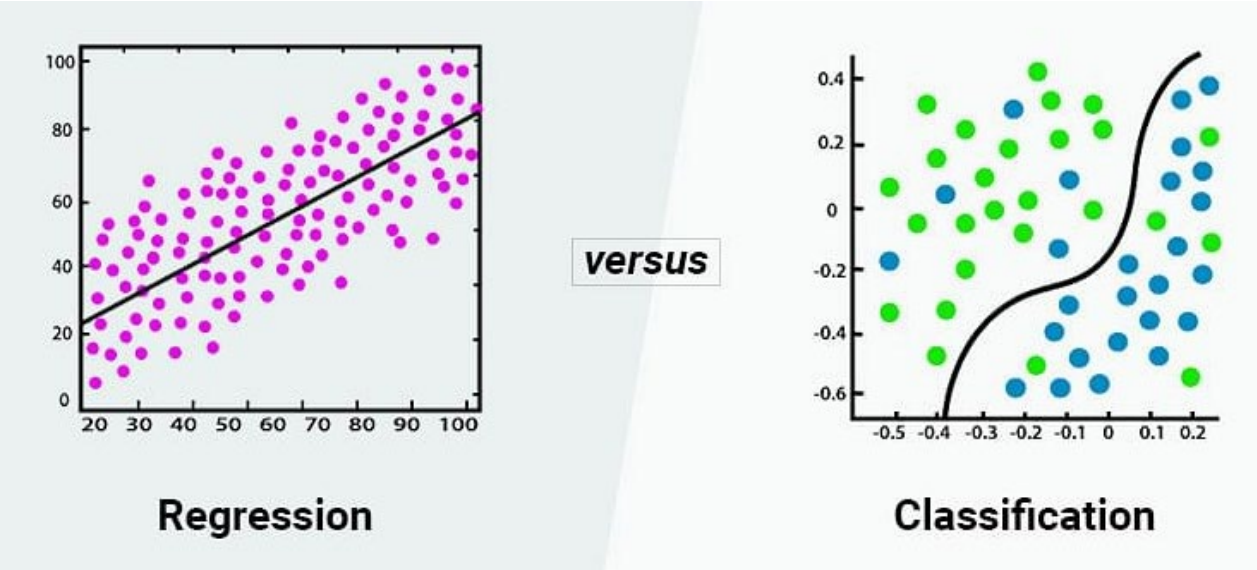
Features

Sepal Length	Sepal Width	Petal Length	Petal Width	Class
5.10	3.50	1.40	0.20	Iris-setosa
4.90	3.00	1.40	0.20	Iris-setosa
7.00	3.20	4.70	1.40	Iris-versicolor
6.40	3.20	4.50	1.50	Iris-versicolor
6.90	3.10	4.90	1.50	Iris-versicolor
6.30	3.30	6.00	2.50	Iris-virginica
5.80	2.70	5.10	1.90	Iris-virginica
7.10	3.00	5.90	2.10	Iris-virginica

Attributes

Continued…

- There are two types of supervised learning



i.e. predicting if sepal length increase does sepal width also increase

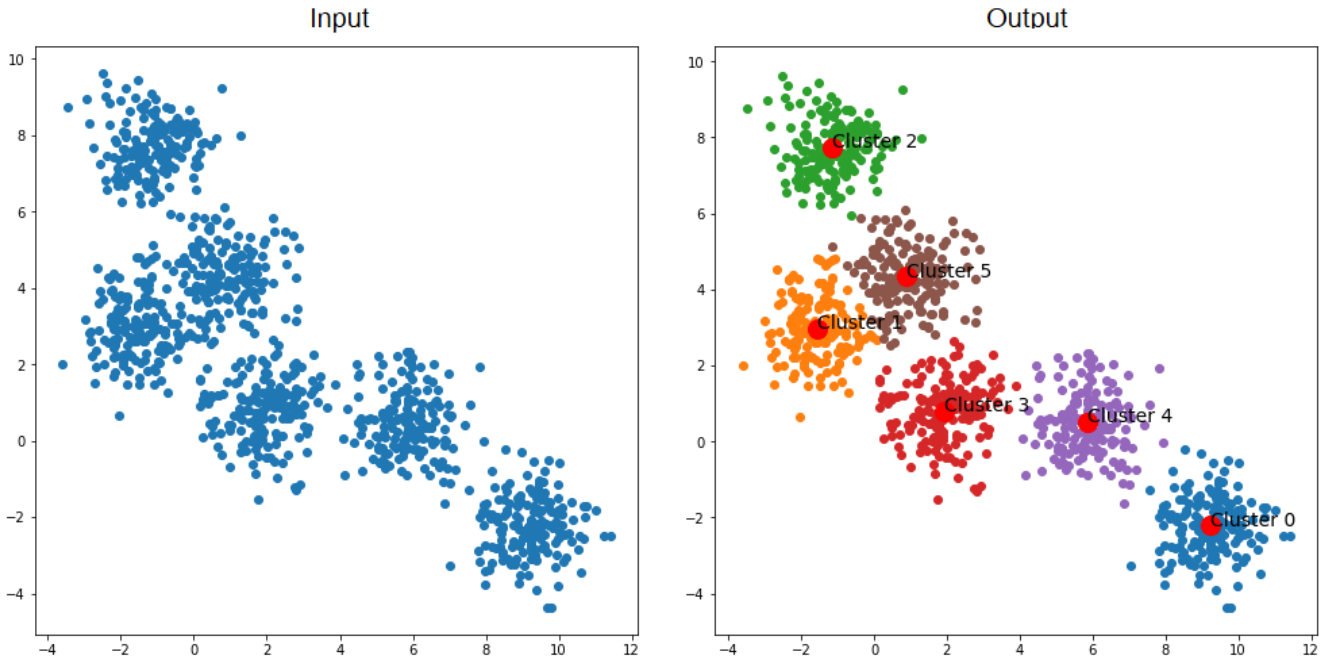
i.e. predicting class against sepal width

Supervised Algorithm Examples

Classification	Regression
Decision Tree	Simple linear regression
Random Forecast	Multiple Linear regression
K-Nearest Neighbor	Polynomial regression

What is unsupervised learning?

- We do not supervise the model but instead let the model work on its own to discover information that may not be visible



Unsupervised Algorithm Examples

Clustering	Association	Dimensionality Rediction
K-Means	Apriori	Principle component analysis
Hierarchical	Eclat	Singular Value Decomposition
Probabilistic	FP-Growth	Autoencoders

Other Machine Learning Methods

- Semi-supervised learning
- Reinforcement Learning
- Deep Learning

7 Steps of Machine Learning

1. Collect Data
2. Data Preparation
3. Choosing a model
4. Training
5. Evaluation
6. Parameter Tuning
7. Prediction

Credit: Yufeng G, 2017

Case Study:

- Let’s say we have been asked to create a system that answers the question of whether a food item can be classed as an ‘apple’ or as an ‘orange’

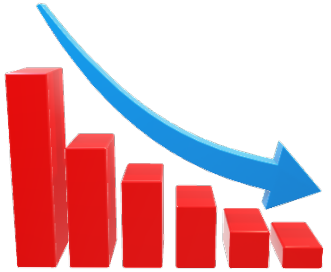
Step 1 – Gather data

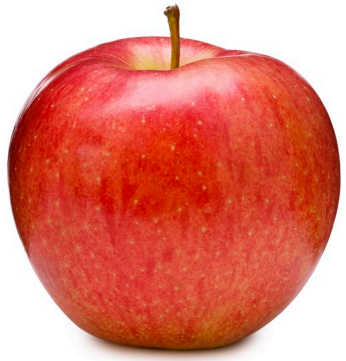
Colour	Shape	Apple or Orange
Green	Round	Apple
Orange	Round	Orange

Step 2 – Prepare Data

- Randomise the order of your data
- Examine for any skewness
- Split data into two parts
  - ~80% would be used for training
  - ~20% would be used for the evaluation process

Step 3 – Choosing a Model





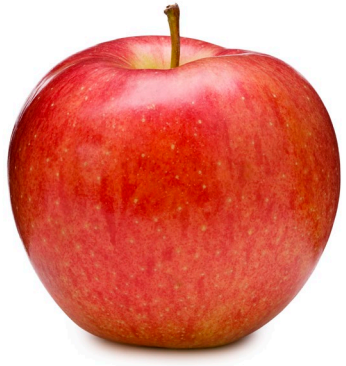
The colour red = Apple

The colour orange = Orange

Step 4 – Training

$$y = m * x + b$$

OUTPUT SLOPE INPUT Y-INTERCEPT



The colour red = Apple

The colour orange = Orange

And again…

- **A Decision Process** = the steps that takes in the data and guesses what kind of pattern your algorithm is looking to find.

- **A Loss/Error Function =** measure how good the guess was by comparing it other examples

- **A Model Optimization Process =** A method in which the algorithm looks at the miss and then updates the decision process

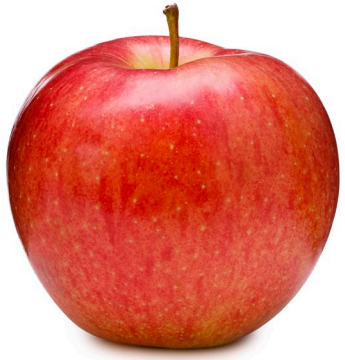
Step 5 – Evaluation

Step 6: Parameter Tuning

You could either:

- Revisit the training step and use multiple sweeps of the training data set for training the model
- Refining the initial values given to the model

Step 7 – Prediction



Apple ✓

Orange

Lets recap

Supervised learning	Unsupervised learning
Input data is labeled	Input data is unlabeled
Has a feedback mechanism	Has no feedback mechanism
Data is classified based on the training dataset	Assigns properties of given data to classify it
Divided into Regression & Classification	Divided into Clustering & Association
Used for prediction	Used for analysis
Algorithms include: decision trees, logistic regressions, support vector machine	Algorithms include: k-means clustering, hierarchical clustering, apriori algorithm
A known number of classes	A unknown number of classes

Let’s explore some other examples

Q.1 What is the best model for ML?

- We need to predict the number of vehicle purchases in a city for historical data?

- Would we need supervised vs unsupervised learning methods?

Q.2 What is the best model for ML?

- We need to identify if a potential customer in that city would purchase a vehicle, given their income and community history

- Would we need supervised vs unsupervised learning methods?

Q.3 What is the best model for ML?

- We want to determine different segments of customers (i.e. gender, age, income bracket, education etc)

- Would we need supervised vs unsupervised learning methods?

Round Off

- The first half
  - What a ML is
  - Its' functions
  - How ML works
  - How ML is different to classic stats
- The second half
  - Different ML methods
  - Identified different models within each method
  - 7 steps to Machine Learning

UK Data Service

References

- Functions of machine learning
- Machine Learning Algorithims
- IBM
- Classificaiton vs regression image
- Machine Learning vs Classic Stats
- Supervised vs unsupervised
- 7 Steps to Machine Learning

UK Data Service

Material for Tuesday 1 November

GitHub:

https://github.com/UKDataServiceOpen/ML_Workshop

UK Data Service

Thank You.

Email:
nadia.kennar@manchester.ac.uk,
louise.capener@Manchester.ac.uk

Twitter:
@NadiaKennar
@CapenerLouise