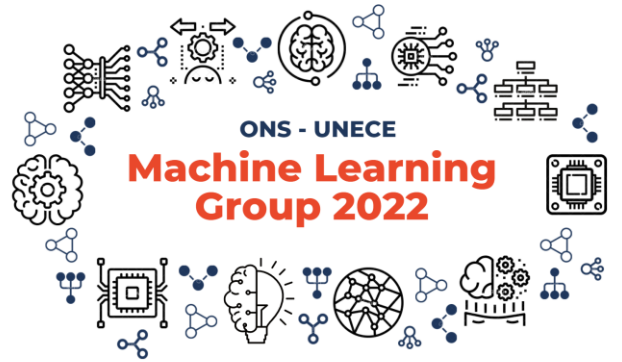
Data Science Campus

Welcome



ONS-UNECE ML Group

Coffee and Coding

28 April 2022

Data Science Campus

UNECE

Housekeeping

Recording

Today's webinar will be recorded. The recording will be available on the statswiki.unece.org/display/ML/Machine+Learning+Group+2022 after the event.

Feedback

At the end of the seminar, you will be asked to take part in a short survey.

Foundations of Machine Learning for non-programmers

Thomas Wise
Data Science Graduate
Consumer Price Method Transformation
@thomasj_wise

28 April 2022

Who am I?

- MSci in Psychology (University of Reading, UK), with a focus on clinical psychology and research methods
- MSc in Methodology and Statistics (Utrecht University, NL), with a focus on machine learning comparisons for predicting PTSD

- I’m passionate about data literacy & education, generative art and data visualization
- I’m an avid cook, indoor gardener and lover of musical theatre!

Session Goals & Aims

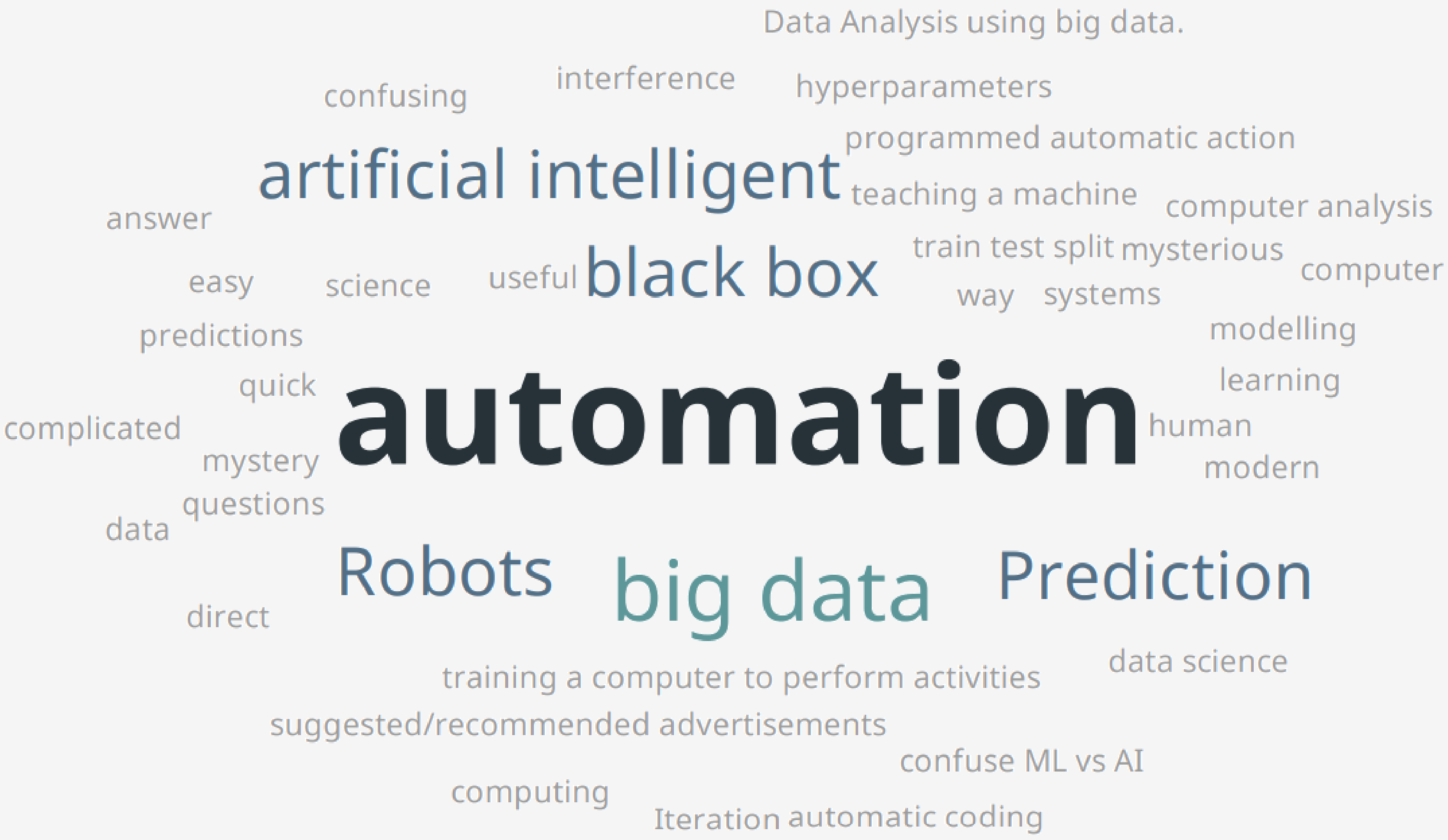
- Begin to understand:
  - The utility and purpose of Machine Learning
  - The problems which can be tackled with Machine Learning
  - The steps involved in addressing Machine Learning problems
  - How to interpret and evaluate Machine Learning problems

Session Schedule

- Introduction
- Part 1a: What is Machine Learning?
- Part 1b: When to use Machine Learning?
- Part 1c: Reproducibility & Machine Learning
- Break: 10 minutes
- Part 2a: How to use Machine Learning
- Part 2b: How to evaluate Machine Learning Models
- Part 2c: Interpreting Machine Learning Models

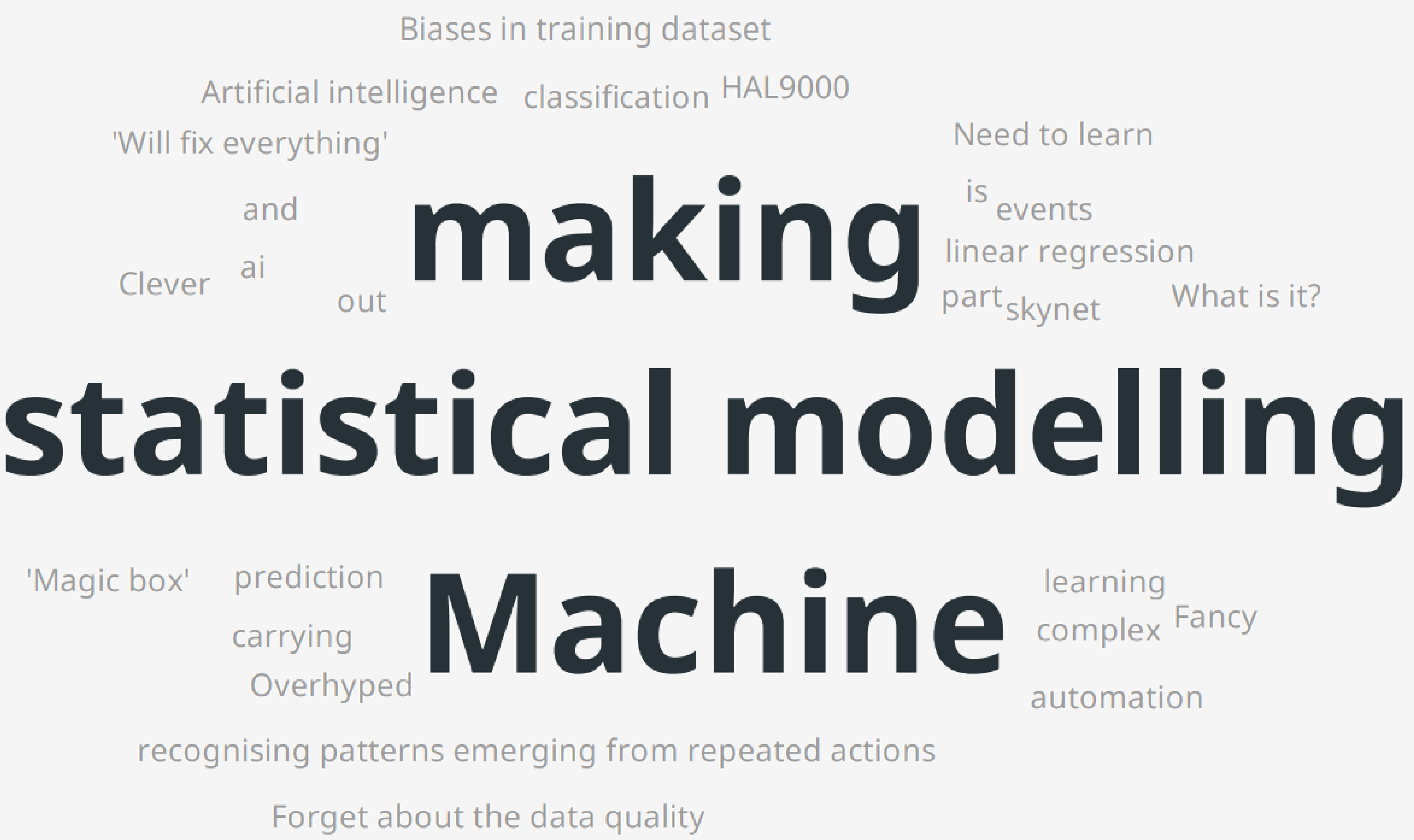
What do you think when you hear Machine Learning?

AM Session



What do you think when you hear Machine Learning?

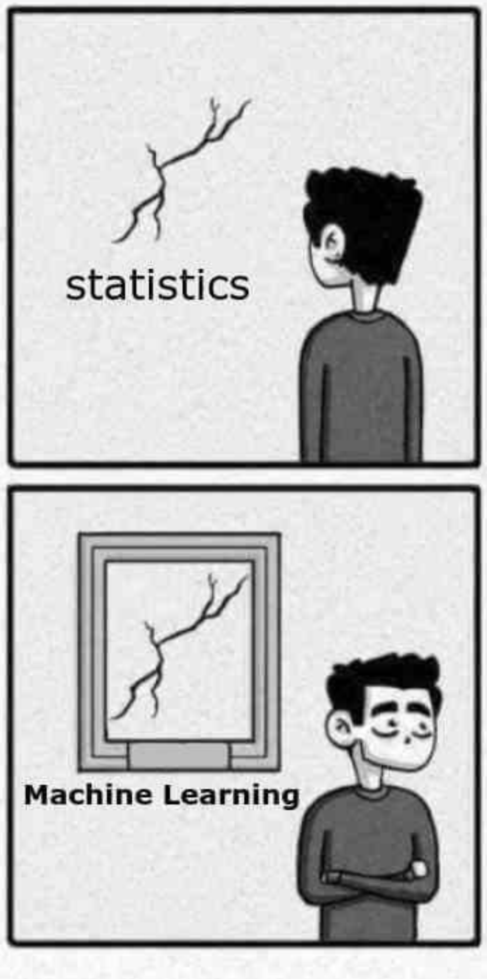
PM Session

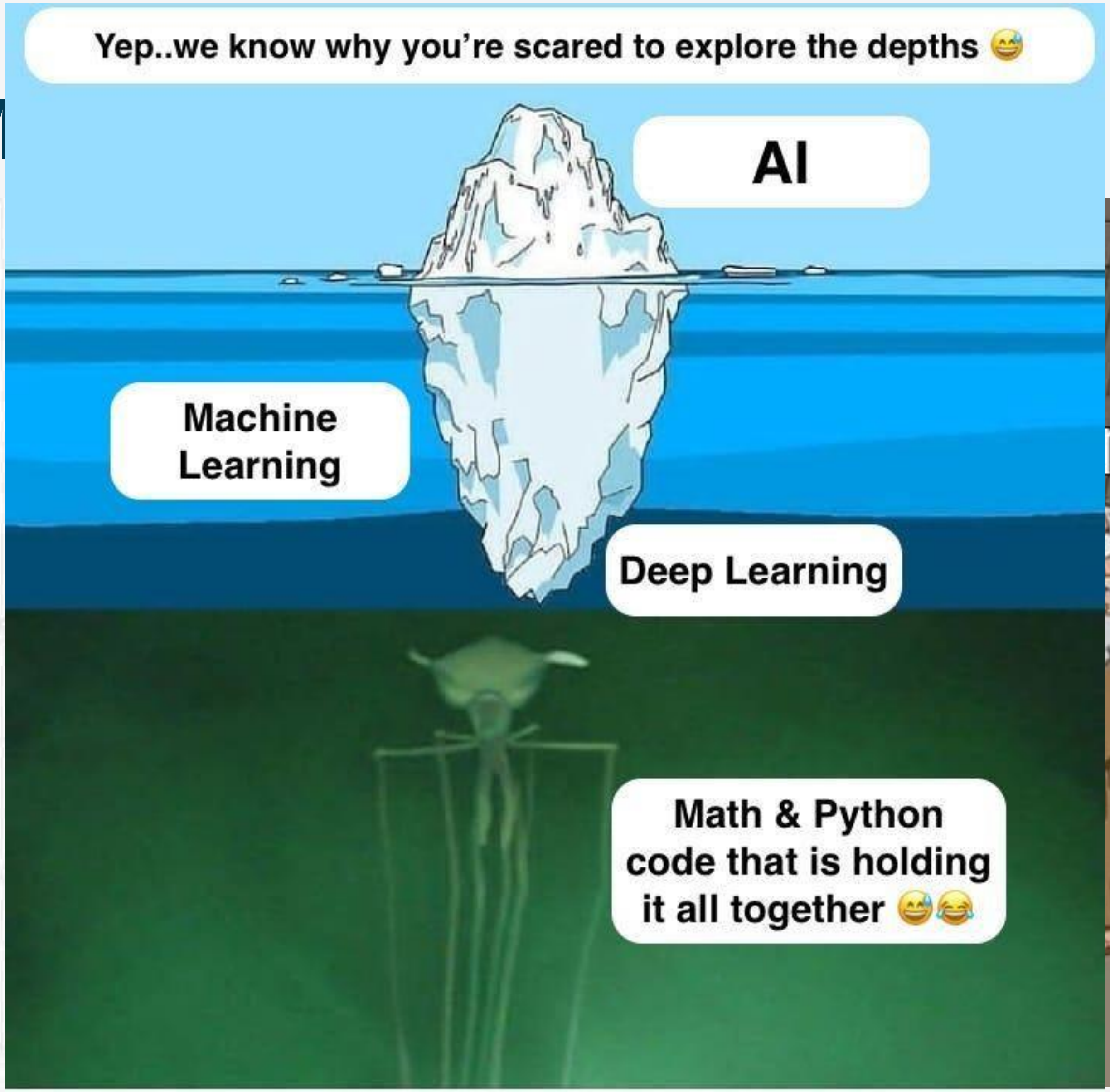


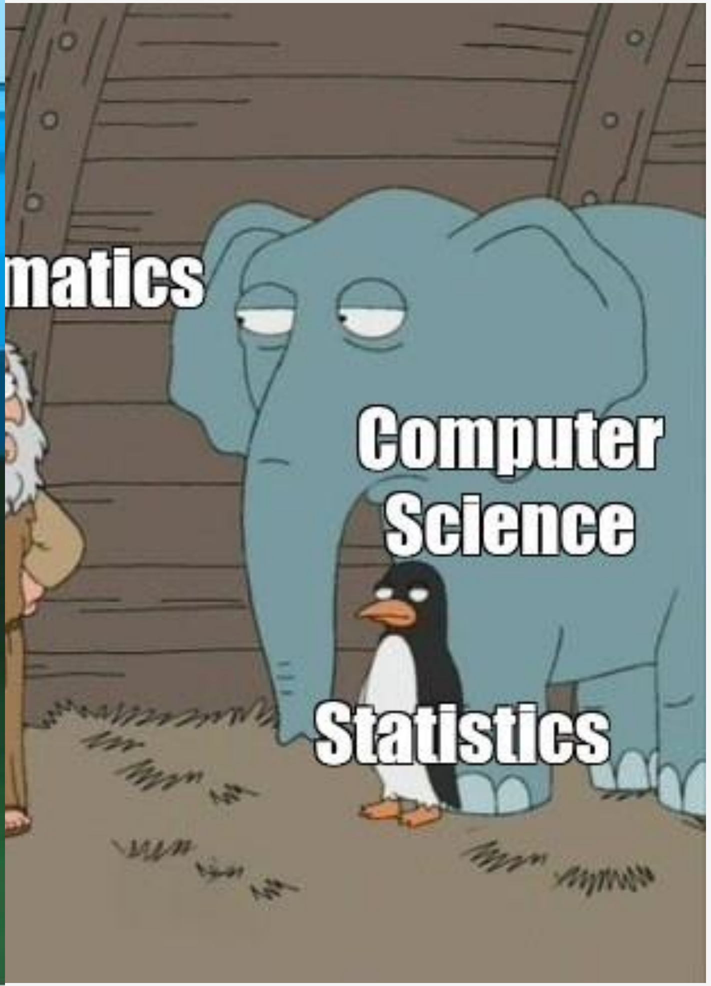
Jargon Buster

- Algorithm: *A finite sequence of well-defined instructions*
- Metric: *A quantitative assessment of performance*
- Aggregates: *Combining of multiple lower level statistics or layers of information.*
- Hyperparameter: *A Parameter used to control the learning process, not derived through training*

What is M


statistics
Machine Learning


Yep..we know why you're scared to explore the depths 😅
AI
Machine Learning
Deep Learning
Math & Python code that is holding it all together 😅😂


matics
Computer Science
Statistics

What is Machine Learning

- Wikipedia defines machine learning as:
  - *Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data.*[1]
  - *Machine learning (ML) is a branch of* ***applied computer science****, focusing upon mathematical processes or equations, which improve automatically through experience and by the use of data.*

1: Mitchell, Tom (1997). *Machine Learning*. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892.

What methods are captured by Machine Learning?

- Traditional Statistical Methods:
  - Regression
  - Principal Component Analysis (PCA)
  - *Correlations (debatable)*
- Computer Vision
- Natural Language Processing
- Supervised Learning
- Unsupervised Learning

What Official Statistical Problems could be addressed using Machine Learning?

AM Session

predict socioeconomic

missing

rates of engagement E&I text classification area

validity

Automating data collection processes with NLP

predictions values data linkage

small area level estimates Imputation

What Official Statistical Problems could be addressed using Machine Learning?

PM Session

Regression and Classification problems.

natural language processing

categorical coding

administrative

backcasting

data

checking

clustering

Unsure

quality

health outcomes

nowcasting

image recognition

means

demand forecasts

gambling

It

incoming

Forecasting

Classification

semantic analysis

Data linkage

increased automation of processes?

Types of Machine Learning Technique

- Supervised Learning
  - Those which use *tagged* or *labelled* data to learn.
- Methods include:
  - Regression
  - Classification

- Unsupervised Learning
  - Those which use *untagged* or *unlabeled* data to learn.
- Methods include:
  - Clustering

Types of Machine Learning Technique

- Supervised Learning
  - Regression
    - Problem aim: predict or project
    - Outcome variable: Continuous
  - Classification
    - Problem aim: classify
    - Outcome variable: Categorical
- Unsupervised Learning
  - Clustering
    - Problem aim: group or cluster
    - No outcome variable

Types of Machine Learning Technique

- Supervised Learning
  - Regression
    - Linear Regression
    - General Linear Models
  - Classification
    - Decision Trees
    - Support Vector Machines
- Unsupervised Learning
  - Clustering
    - K-Means Clustering

Some Real World Examples

Example	Data Information	Technique Category
Imputation in the sample survey on participation of Polish residents in trips: *Statistical Office in Rzeszów, Statistics Poland; Sebastian Wójcik.*	Multiple input variables, **outcome variable is continuous**	Supervised Learning - Regression
An ML application to automate an existing manual process through the use of aerial imagery. *Australian Bureau of Statistics; Daniel Merkas and Debbie Goodwin*	Multiple input variables, **outcome variable is categorical**	Supervised Learning - Classification
No UNECE Examples known, but can include Recommender Systems and Customer Segmentation	Multiple variables, **without outcome variable**	Unsupervised Learning - Clustering

Example 1: Supervised Learning, Regression

- *Imputation in the sample survey on participation of Polish residents in trips:*
  - *Statistical Office in Rzeszów, Statistics Poland*
  - *Sebastian Wójcik*
- Machine Learning Models:
  - Classification and Regression Trees *(CART)*
  - Random Forest
  - Optimal Weighted Nearest Neighbor
  - Support Vector Machines

Example 2: Supervised Learning, Classification

- *An ML application to automate an existing manual process through the use of aerial imagery.*
  - *Australian Bureau of Statistics (ABS)*
  - *Daniel Merkas and Debbie Goodwin*
- Machine Learning Techniques
  - Computer Vision

What could we do to ensure the reproducibility of our machine learning projects?

AM Session

library of methods
make code open source
code
Share metadata
Use version control - Git
github
ensure
official
transparent data and techniques
make
use
Repository
share the code
RAP
unittests
programming
readily
stats
goo
modules
documentation
conventions
practice
current
available
it
sense checking
RAPs
sharing codes
set seed
open codes
a stable system
release
packages
reproducible analytical pipelines
keep a snapshot of the dataset
open data and code
Use good standard coding

What could we do to ensure the reproducibility of our machine learning projects?

PM Session

consistent recommended software

Functions

Data centric

Train on relevant and high-quality data

publish the training set

Peer review

data cleaning

simplicity

always give same inputs

clear documentation

Clarity

use open source ML pakcages/models

code availability

set the seed!

consistent methodology

data input and tranformation automation

clearly labelled code

Reproducibility: Core Elements

Core Element	Definition	Challenges	Solutions
Code	The foundation of any project, which defines the ML algorithm.	• Inconsistent Style • (Pseudo) Randomness • Untracked Development	• Experiment tracking and logging • Version Control • Style Guides • Randomization management
Data	What the ML algorithm was trained, validated and tested upon.	• Changes in Data	• Data change tracking and logging • Data Versioning • Artifact & Model storage
Environment	The environment in which the ML algorithm was built, developed and run	• Hyperparameter Inconsistency • Library, Framework and Package changes, updates or revisions	• Environment recording & dependency management • Model Versioning • Version Control • Model Registry

Adapted from: https://neptune.ai/blog/how-to-solve-reproducibility-in-ml

Reproducibly in Practice

Solutions	Practical Tools
Change tracking and logging	• Git, Gitlab or equivalent • Comprehensive ReadMe or Diary based document
Version control	• Git, GitLab or equivalent
Style Guides	• Python: PEP8 • R: Tidyverse or Google Style Guide
Environment Management	• Git, Gitlab or equivalent • Comprehensive ReadMe or Diary based document • Requirements File (listing all packages and versions)
Randomness Management	• Setting randomness parameter (seeds)

What country or organisation do you work for?

AM Session

Forest Research (part of Forestry Commission)

UN Women, New York

Urban Foresight

Central Statistical Bureau of Latvia

UNESCWA

Australia

ONS

UK

Latvia

Canada /Statistics Canada

UK - ONS

University of Queensland

ABS

Istat

Statistics Poland

TurkStat

Central Statistics Office Ireland

national statistics office

Ireland

CSO (Ireland)

Lebanon, UN ESCWA now ; Canada, Statistcis Canada then !

Central Statistical Bureau Republic of Latvia

What country or organisation do you work for?

PM Session

Ecuador, in the National Statistics Office
National Institute of Statistics Ecuador
Canada;
Statistics
US - UNSD
Office for National Statistics
Brazilian
Canada
Service
incomelow
ONS
ppl
delivery
Institute
Brazil / IBGE
underserved
Org
Food
racialized
UK
Geography
United Nations
new
seniors
Frontline
University
uk ons
UK - Stockport Council
immigrants
StatCan
homeless
NISRA - NI
NSO)
ONS - UK
Statistical Center of Iran
(Brazilian
Ireland's NSI - the Central Statistics Office.
Department of Health & Social Care

What is your role in your organisation?

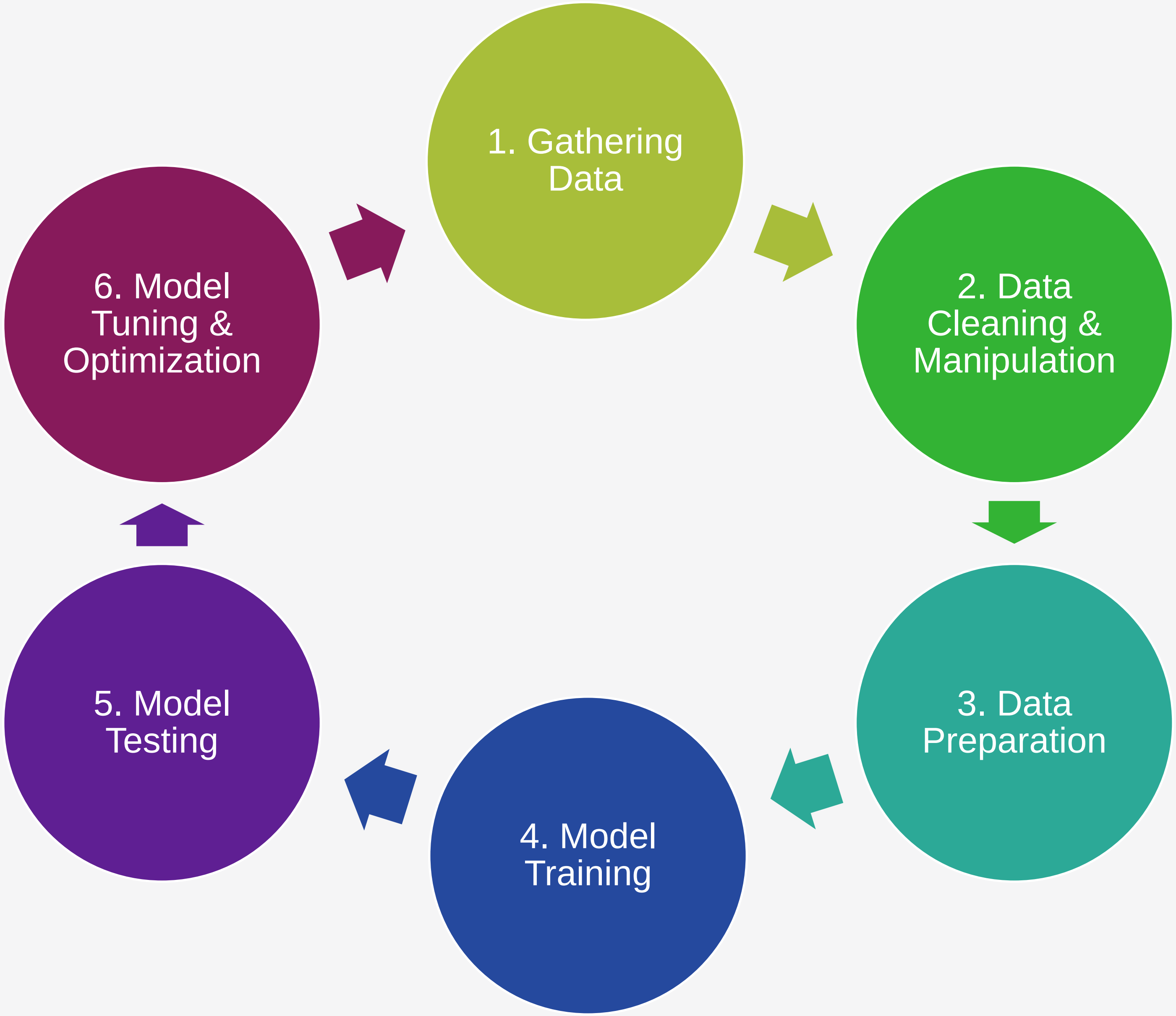
AM Session

What is your role in your organisation?

PM Session

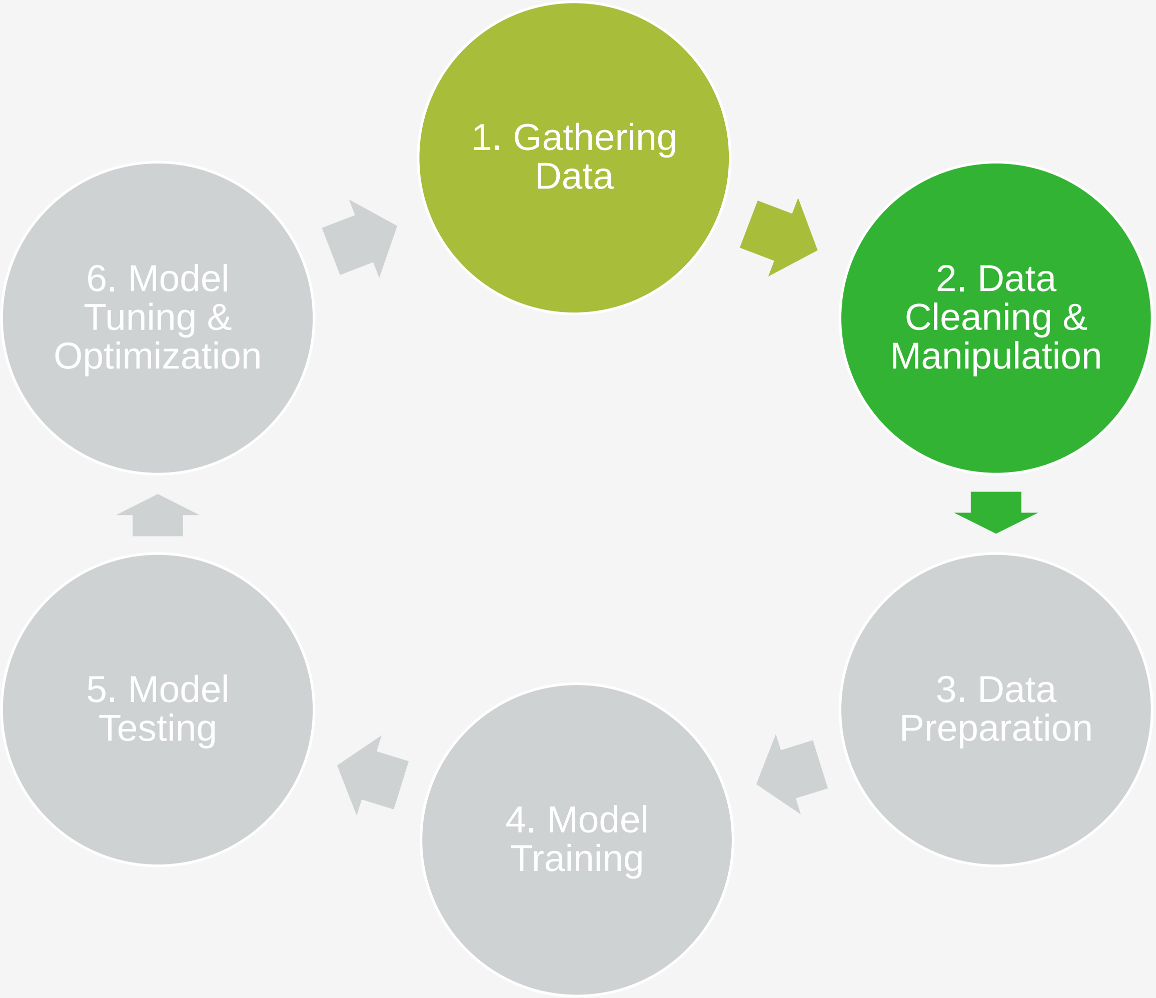
Senior social researcher
Systems Analyst
Officer
official
production
researcher
innovation
Team
visualisation
Analyst
concept
checking
role
data
Statistical
oil
Ronald
engineering
sector
working
projects
statistical
gas
and
specialist
Statistician
statistics
the
statistical officer
manager
Publishing
proof
Data Analyst
project
ONS
ways
Product
new
Research analyst
reasearch officer
Research officer
science
Support
improve
engineer
PhD Student
Research Officer - COVID

Machine Learning Lifecycle



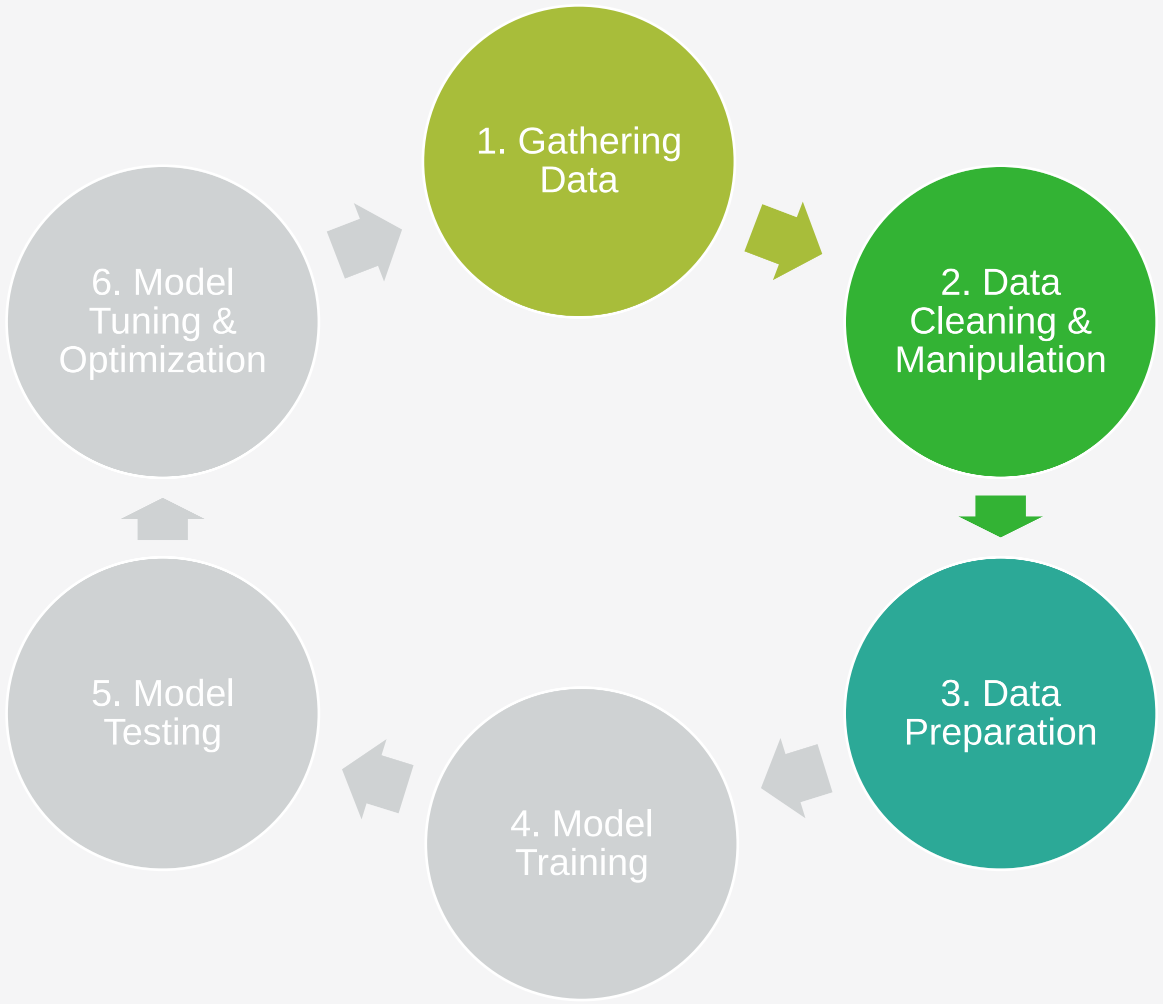
Data Cleaning & Manipulation

- Handling Missing Data
- Handling of Outliers
- Handling of incorrect data
- Feature Engineering
- Feature Selection
- Quality assurance



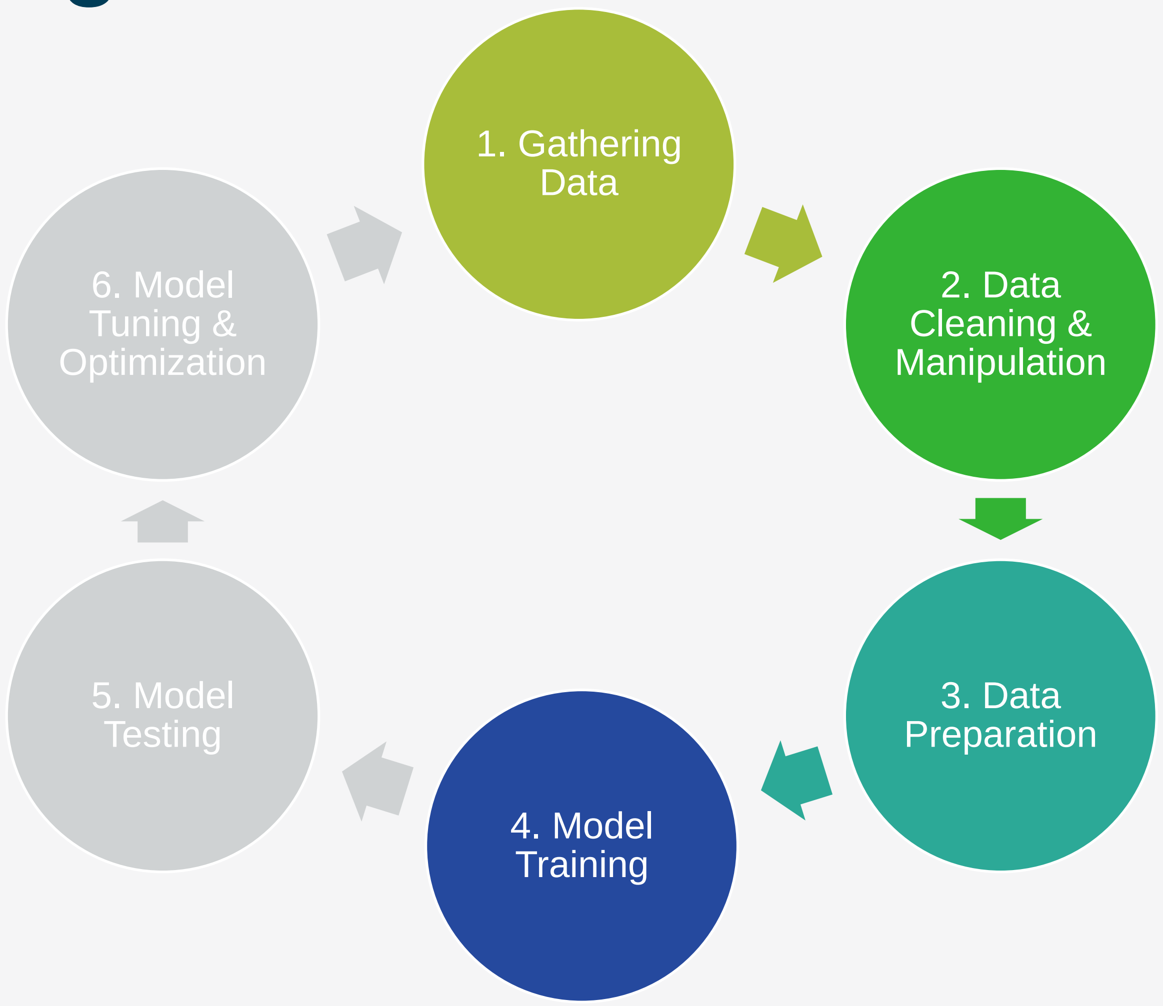
Data Preparation

- Split data into:
  - Training and Testing Sets
  - Training and Validation Sets
- Specific Model Preparation
- Determine K-Folds Cross Validations

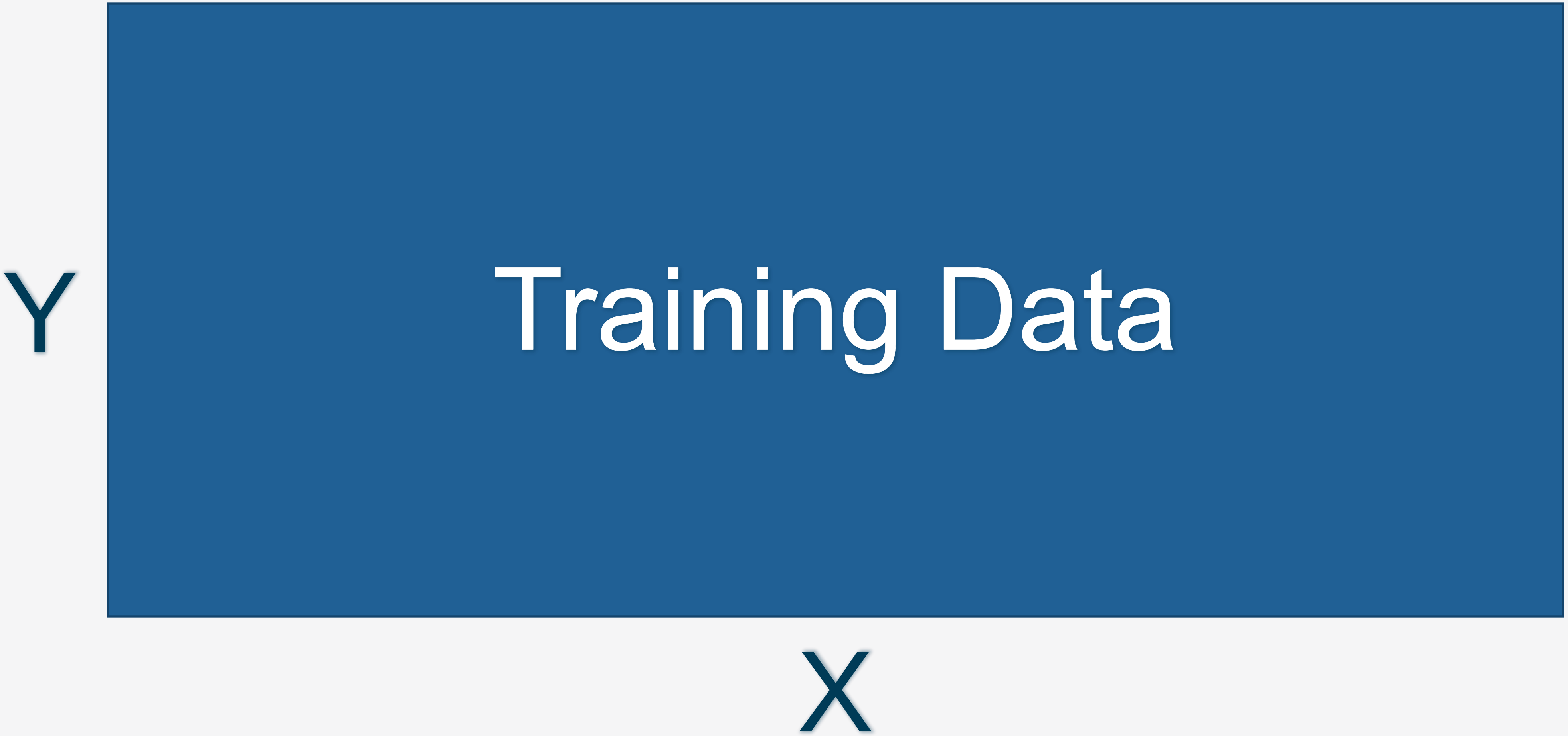


Model Application: Training

- Once cleaned data can be trained to a model
- Usually begin with your default model



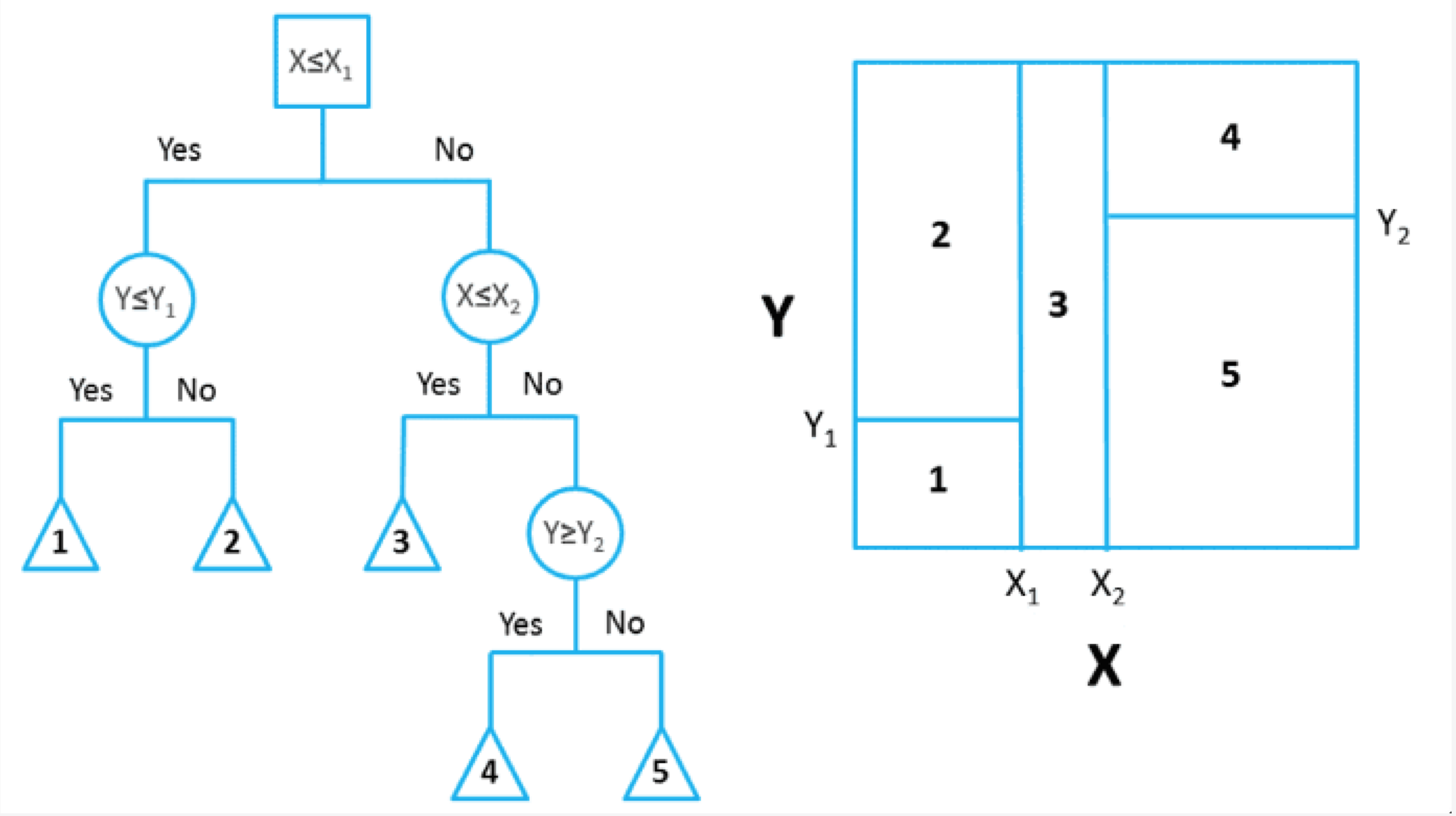
Model Application: Training Decision Trees



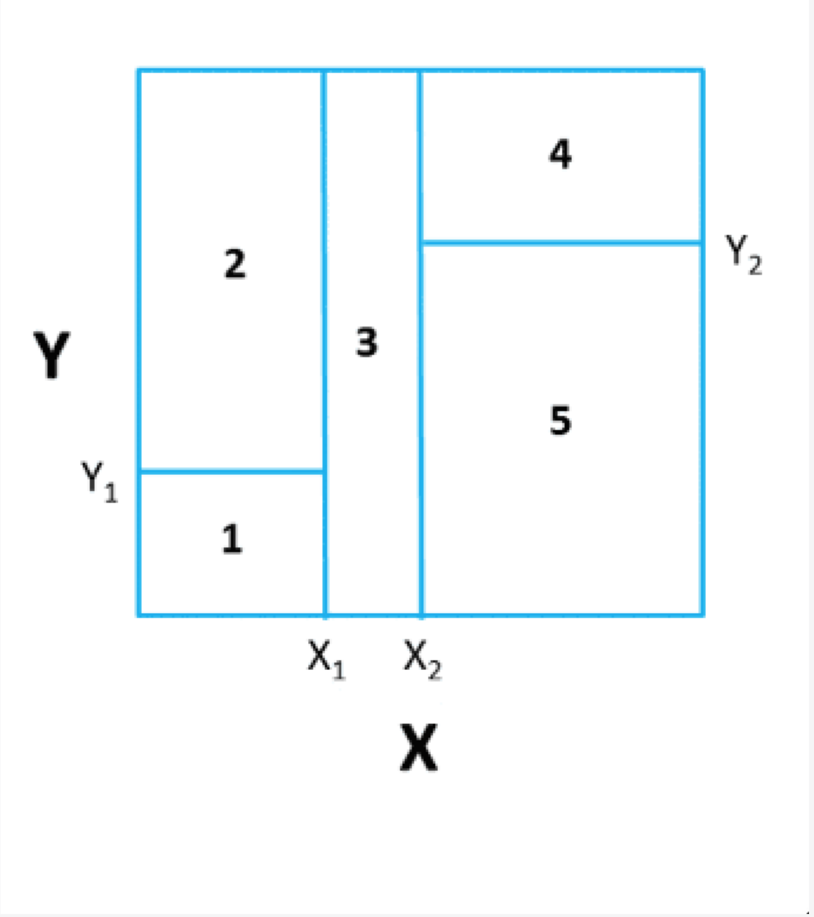
Model Application: Training Decision Trees

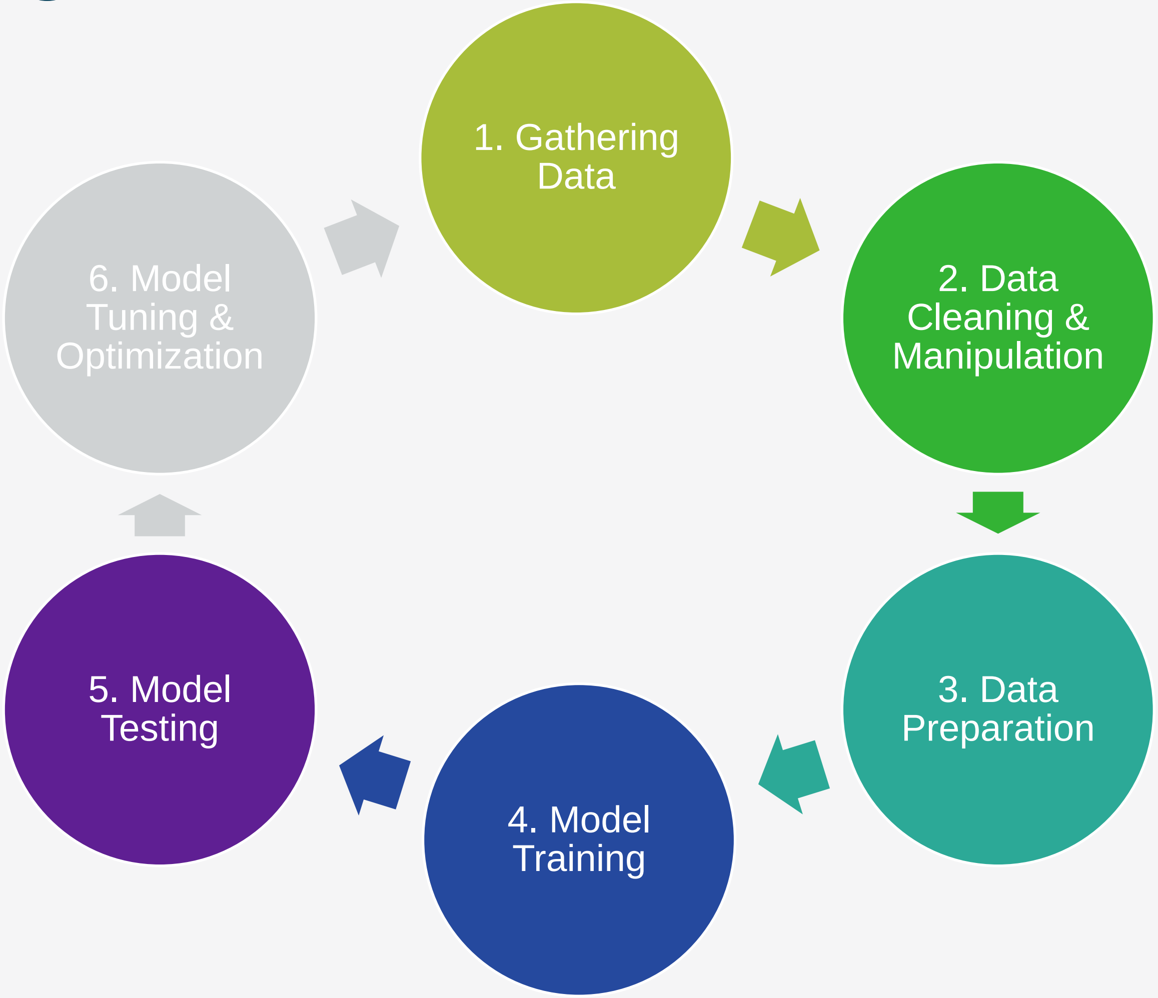
Model Application: Training Decision Trees

Model Application: Training Decision Trees



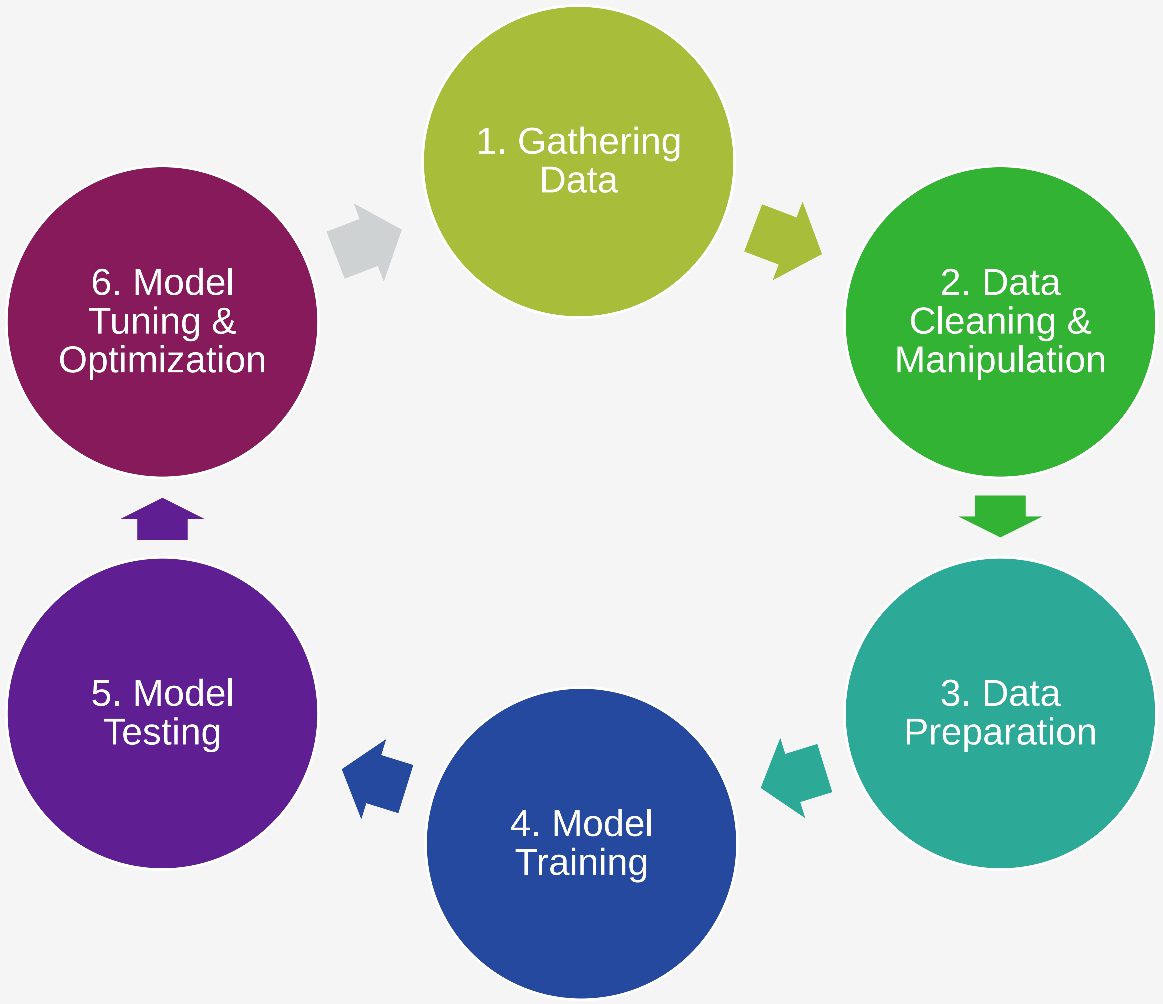
Model Application: Testing





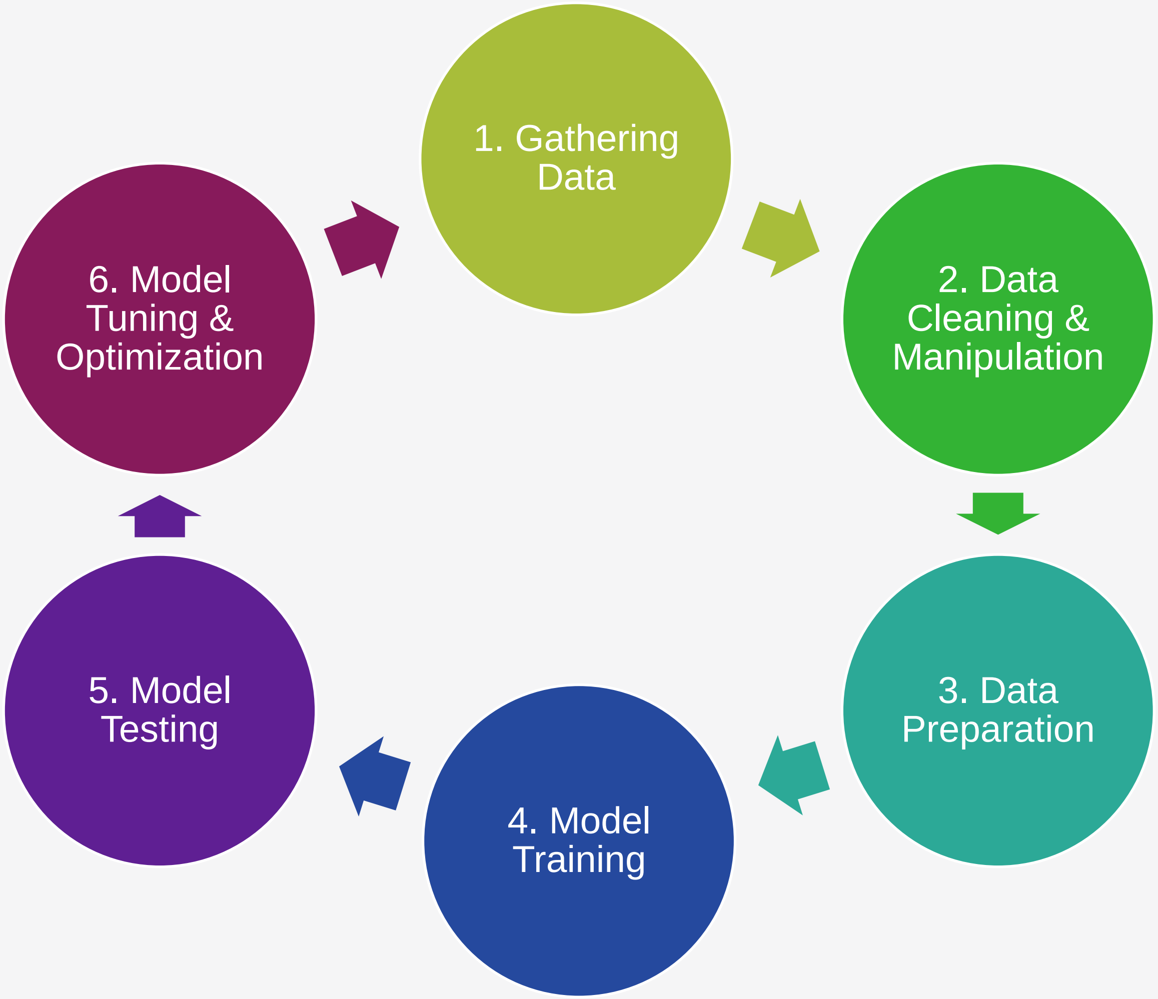
Model Tuning and Optimization

- Grid Search
  - Build models for every combination of hyperparameter value
- Random Search
  - Build models for a random set of hyperparameters based on a distribution
- Bayesian Optimization
  - Build models based on Bayesian methodology with the previous hyperparameter influencing the next.



Model Deployment and Future

- Model Deployment:
  - As web applications
  - Batch production
  - Model Embedding



Evaluating ML Models: Regression

Metric	Description
R-Squared	**Amount of variance explained in the model**
Adjusted R-Squared	Amount of variance explained in the model
Mean Absolute Error (MAE)	**Absolute Difference between actual and predicted values**
Mean Absolute Percentage Error (MAPE)	**Percentage variation of the Mean Absolute Error**
Mean Squared Error (MSE)	Squared Difference between actual and predicted values
Root Mean Squared Error (RMSE)	**Square Rooted of Mean Squared Error**
Root Mean Squared Log Error (RMSLE)	Log of the Root Mean Squared Error

Evaluating ML Models: Regression

Metric	What does it show?	Interpreting the score	Advantages	Disadvantages
R-Squared	Proportion of explained variance	Range: 0-1 0 = No explanation 1 = Total explanation	Highly interpretable	Works in isolation, not across model comparable
Mean Absolute Error (MAE)	Measure of prediction error	Range: 0 to infinity 0 = No prediction error Inf = Large prediction error	Robust to outliers	Not graphically differentiable
Mean Absolute Percentage Error (MAPE)	Percentage based measure of prediction error	Range: 0-100 (%) 0 = No prediction error 100 = Large prediction error	Highly interpretable	Does not handle zero as an actual value well
Root Mean Squared Error (RMSE)	Measure of prediction error	Range: 0 to infinity 0 = No prediction error Inf = Large prediction error	Highly interpretable	Not as robust to outliers

Evaluating ML Models: Regression

1. Draw N samples without replacement, amounting to 90% of the dataset size
2. Train model based on selected hyperparameters
3. Make predictions based upon the remaining 10% of cases
4. Calculate metrics (R^2, RMSE, MAPE & MAE)
5. Average metrics over all N draws.

Evaluating ML Models: Regression

III Dataset pertaining to expenditures for transport

Method (R function)	MAE	MAPE	RMSE	R2
Linear Model OLS (lm)	440.05	1.483	853.00	0.418
General Linear Model GLS (glm)	440.05	1.483	853.00	0.418
Robust Linear Model (rlm)	413.35	1.208	874.08	0.402
LARS (lar)	448.23	1.559	892.36	0.368
Predictive Mean Matching (pmm)	434.59	1.666	841.59	0.434
CART (rpart)	373.62	1.149	696.27	0.612
Random Forest (randomForest)	407.10	1.259	820.22	0.462
Optimal Weighted Nearest Neighbour (kknn)	393.03	1.244	774.25	0.533
Support Vector Machine (svm) radial kernel	519.22	1.384	1149.6	0.057
Support Vector Machine (svm) linear kernel	458.63	0.912	1043.7	0.272

Evaluating ML Models: Classification

		Predicted Condition	
	Total Population = P+N	Predicted Positive (PP)	Predicted Negative (PN)
Actual Condition	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

- Correct results:
  - True Positive (TP)
  - True Negative (TN)
- Incorrect results:
  - False Negative (FN)
  - False Positive (FP)

Evaluating ML Models: Classification

Predicted Condition

Actual Condition

Total Population = P+N	Predicted Positive (PP)	Predicted Negative (PN)	
Positive (P)	True Positive (TP)	False Negative (FN)	Sensitivity, True Positive Rate, Recall = TP/P
Negative (N)	False Positive (FP)	True Negative (TN)	Specificity, True Negative Rate, Selectivity = TN/N
Accuracy = TP+TN/P+N			

- Sensitivity or True Positive Rate
  - True Positive / Positive
- Specificity or True Negative Rate
  - True Negative / Negative
- Accuracy
  - (True Positive + True Negative) / Total Population

Evaluating ML Models: Classification

Predicted Condition

Actual Condition

Total Population = P+N	Predicted Positive (PP)	Predicted Negative (PN)	
Positive (P)	True Positive (TP)	False Negative (FN)	Sensitivity, True Positive Rate, Recall = TP/P
Negative (N)	False Positive (FP)	True Negative (TN)	Specificity, True Negative Rate, Selectivity = TN/N
Accuracy = TP+TN/P+N	Positive Predictive Value, Precision = TP/PP	Negative Predictive Value = TN/PN	

- Precision or Positive Predictive value
- Negative Predictive value
- F1-score
- Area under the Curve (AUC)

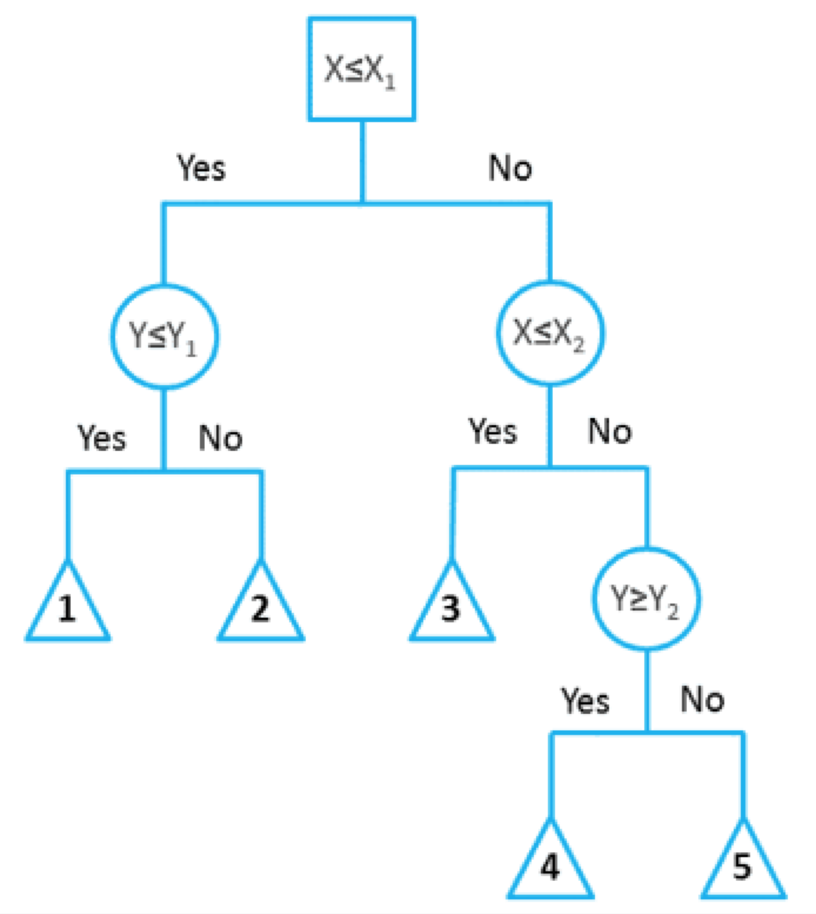
Evaluating ML Models: Classification

	Actual						
	Commercial	High Density	Residential	Under-Construction	Vacant Land	Poor Geocode	Total
Commercial	201	16	2	3	4	3	229
High Density	34	229	4	5	2	1	275
Residential	3	5	238	7	2	4	259
Under-Construction			5	230	8	1	244
Vacant Land	9			4	232	3	248
Poor Geocode	3		1	1	2	238	245
Total	250	250	250	250	250	250	1,500

	Commercial	High Density	Residential	Under-Construction	Vacant Land	Poor Geocode	Total
True Positive	201	229	238	230	232	238	1,368
True Negative	1,167	1,139	1,130	1,138	1,136	1,130	6,840
False Positive	28	46	21	14	16	7	132
False Negative	49	21	12	20	18	12	132
Total	1,445	1,435	1,401	1,402	1,402	1,387	8,472

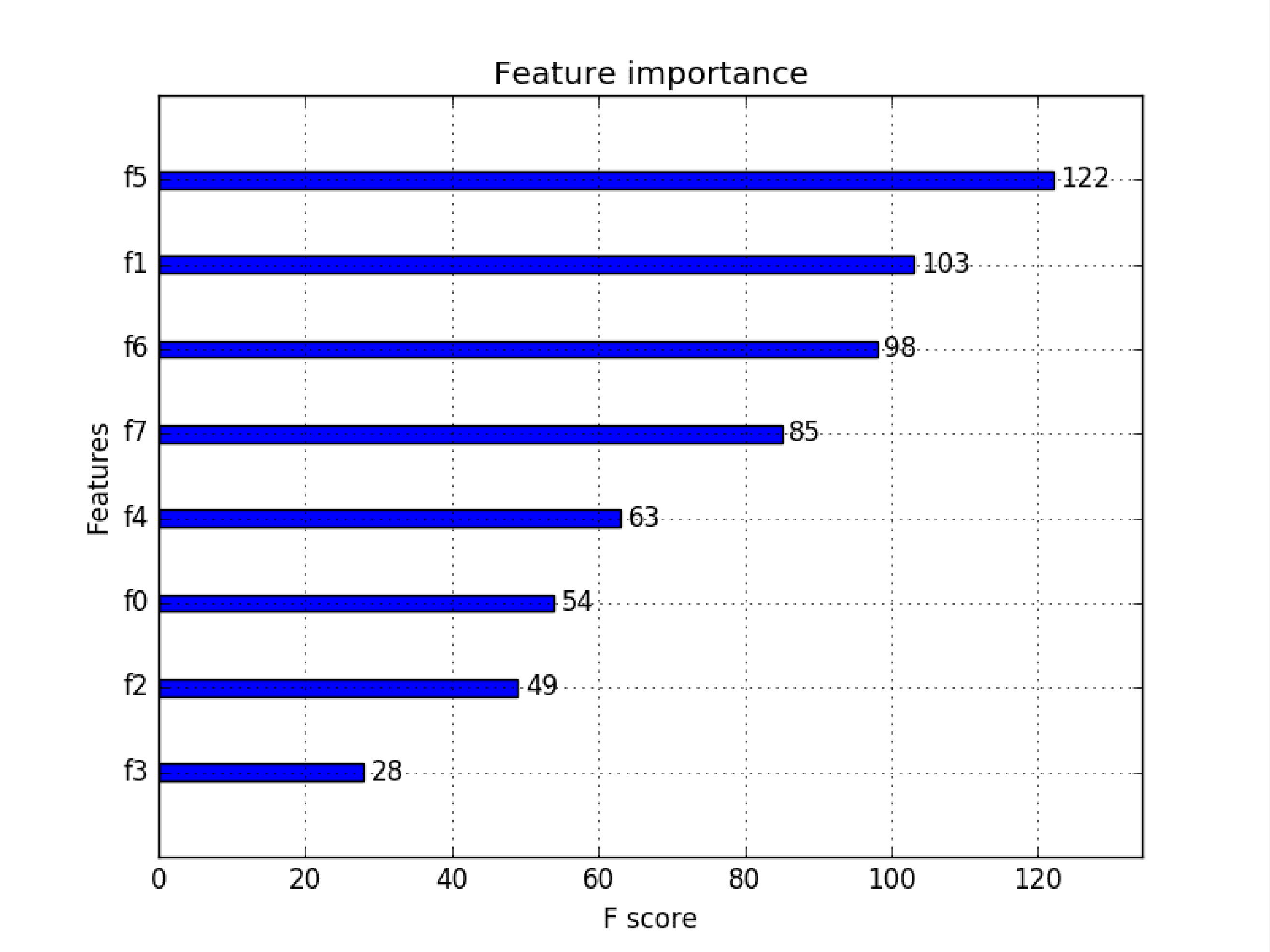
	Commercial	High Density	Residential	Under-Construction	Vacant Land	Poor Geocode
Proportional Accuracy	94.7%	95.3%	97.6%	97.6%	97.6%	98.6%
Proportional Accuracy [95% CI]	94.7% [93.5%, 95.8%]	95.3% [94.2%, 96.4%]	97.6% [96.9%, 98.4%]	97.6% [96.8%, 98.4%]	97.6% [96.8%, 98.4%]	98.6% [98%, 99.2%]
Precision	87.8%	83.3%	91.9%	94.3%	93.5%	97.1%
Recall	80.4%	91.6%	95.2%	92.0%	92.8%	95.2%
F1	83.9%	87.2%	93.5%	93.1%	93.2%	96.2%

Interpretability of Machine Learning Models



Layer Type	Details
2D Conv Layer	[Filters = 32, Kernel = 3x3, Activation = Relu]
Max Pooling Layer	[Pool = 2x2]
2D Conv Layer	[Filters = 32, Kernel = 3x3, Activation = Relu]
Max Pooling Layer	[Pool = 2x2]
2D Conv Layer	[Filters = 32, Kernel = 3x3, Activation = Relu]
Max Pooling Layer	[Pool = 2x2]
2D Conv Layer	[Filters = 32, Kernel = 3x3, Activation = Relu]
Max Pooling Layer	[Pool = 2x2]
Flatten Layer	
Drop Out Layer	[Rate = 0.5]
Dense Layer	[Units = 512, Activation = Relu]
Dense Layer	[Units = 6, Activation = Softmax]

Interpretability of Machine Learning Models



Session Goals & Aims

- Begin to understand:
  - The utility and purpose of Machine Learning
  - The problems which can be tackled with Machine Learning
  - The steps involved in addressing Machine Learning problems
  - How to interpret and evaluate Machine Learning problems

Things I've not covered / Next Steps for you

- Model Over/Under fitting, and its impact
- Impact of Bias in model selection and application
- Hyperparameter details & tuning techniques
- Benefits and Curse of Ensemble Methodologies
- Calculating Multi-category confusion matrix metrics
- Discussing the role of sample & group size in performance

Further Reading:

- UNECE Website: https://statswiki.unece.org/display/ML/Studies+and+Codes
- UNECE Machine Learning for Official Statistics guide:
  - https://unece.org/statistics/publications/machine-learning-official-statistics
- An introduction to Statistical Learning;
  - Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
- Pattern Recognition and Machine Learning
  - Chistopher M. Bishop
- R for Data Science
  - Hadley Wickham and Garrett Grolemund
- Python Data Science Handbook
  - Jake VanderPlas

Join the Machine Learning 2022 Group!

- Machine Learning Group 2022:
  - https://statswiki.unece.org/display/ML/Machine+Learning+Group+2022

- To join the mailing list or group, drop the team an email:
  - ML2022@ons.gov.uk

Examples

- Statistics Poland:
  - https://statswiki.unece.org/display/ML/Studies+and+Codes?preview=/285216428/290358687/ML_WP1_EI_Poland.pdf

- ABS
  - https://statswiki.unece.org/display/ML/Studies+and+Codes?preview=/285216428/290358690/ML_WP1_Imagery_Australia.pdf

Thank you for coming