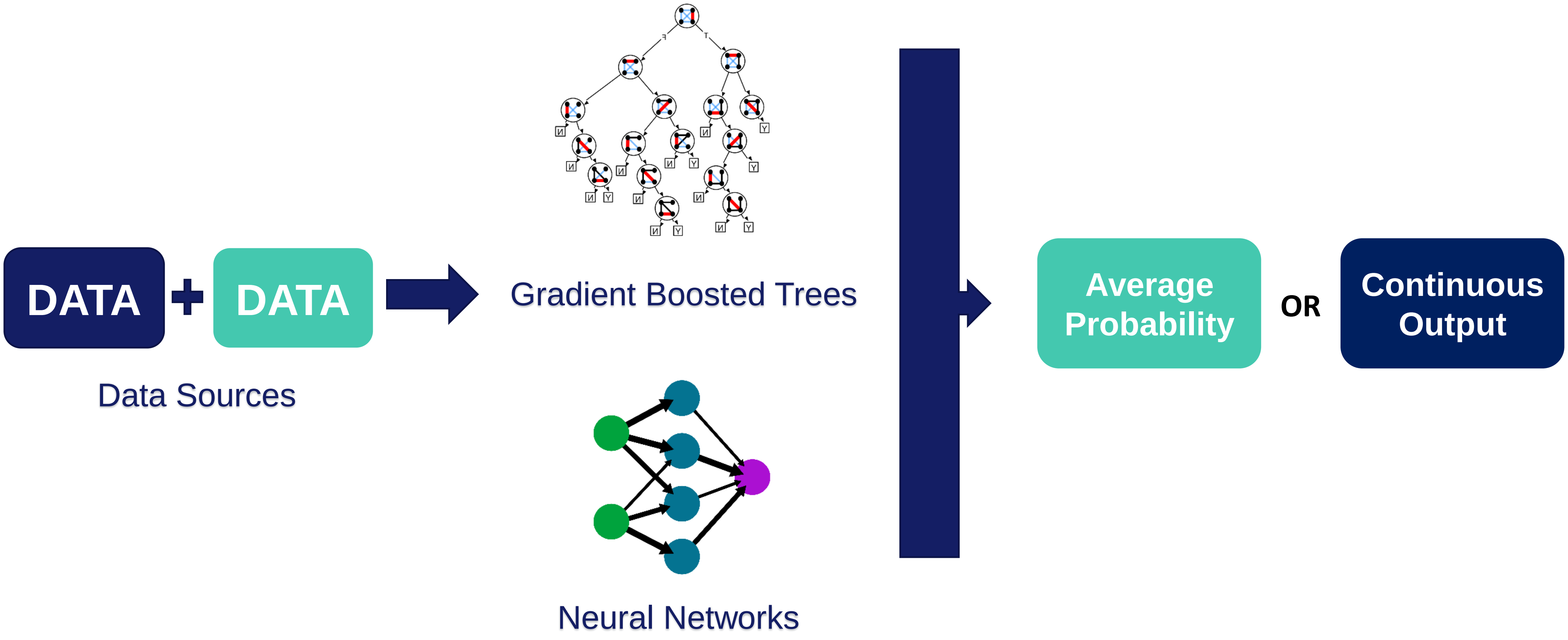
Engineering and selecting features for machine learning

Soledad Galli, PhD

DSF meetup with Busuu

London, 16th October 2018

Machine Learning Pipeline



Machine Learning Finance and Insurance

Claims

Fraud

Credit Risk

Marketing

Pricing

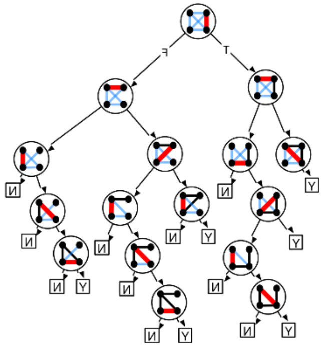
Machine Learning Pipeline

DATA

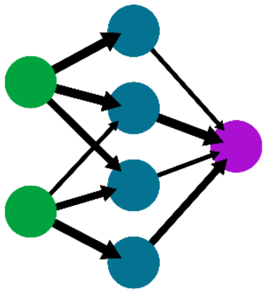
+

DATA

Data Sources



Gradient Boosted Trees



Neural Networks

Average Probability

Continuous Output

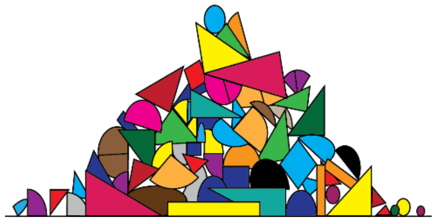
Machine Learning Pipeline

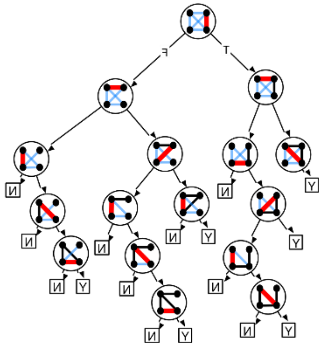
DATA

+

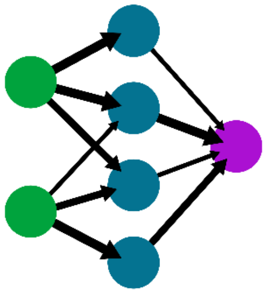
DATA

Data Sources





Gradient Boosted Trees

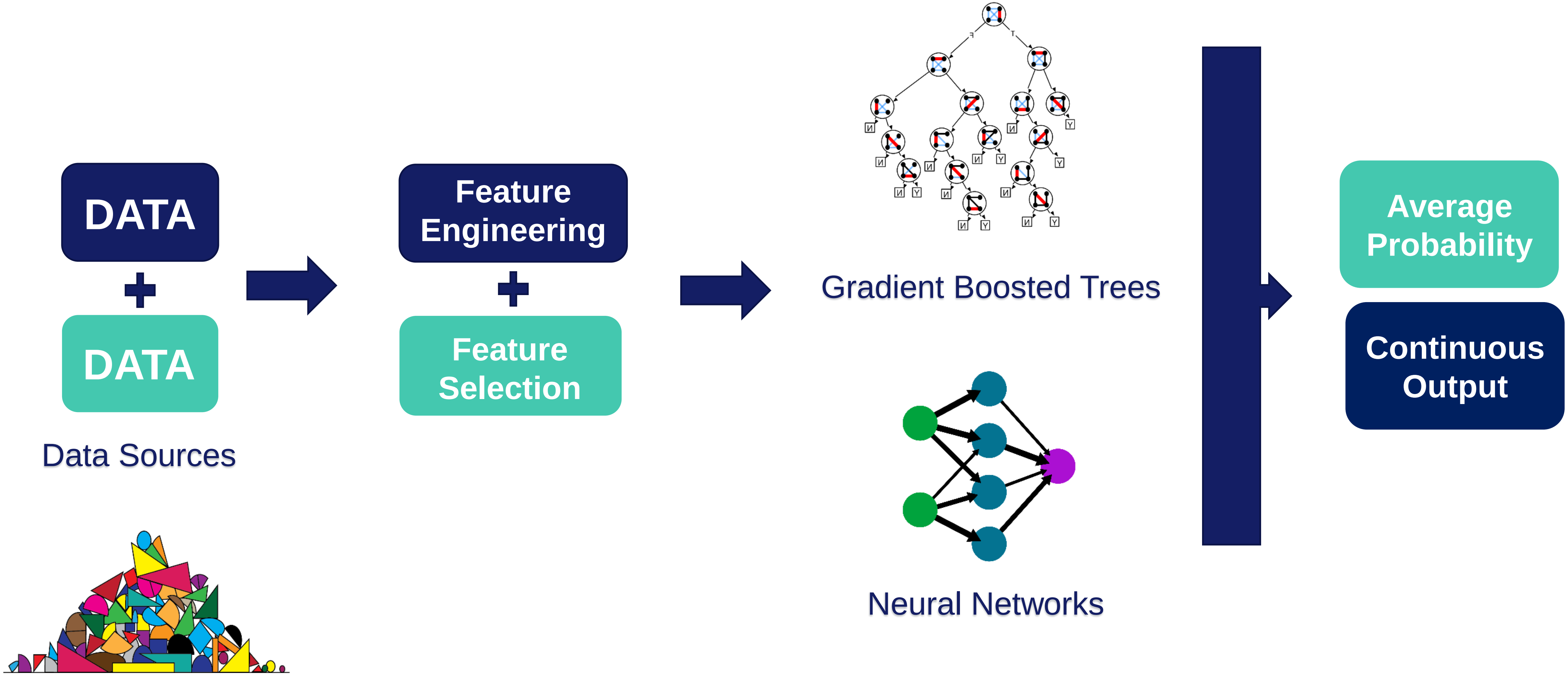


Neural Networks

Average Probability

Continuous Output

Machine Learning Pipeline



Data Pre-processing Journey

- Common issues found in variables
- Feature / Variable engineering: solutions to the data issues
- Feature selection: do we need to select features?
- Feature / Variable selection methods
- Overview and knowledge sources

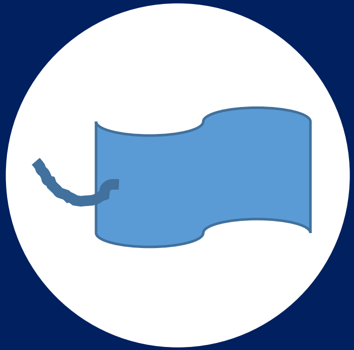
Data Pre-processing Journey

- **Common issues found in variables**
- Feature / Variable engineering: solutions to the data issues
- Feature selection: do we need to select features?
- Feature / Variable selection methods
- Overview and knowledge sources

Problems in Variables

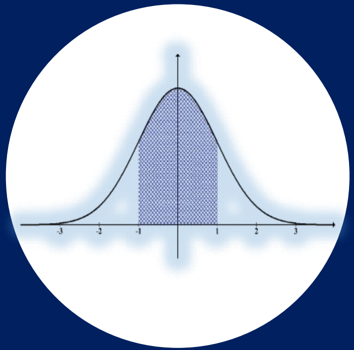
Missing data

Missing values within a variable



Labels

Strings in categorical variables



Distribution

Normal vs skewed

Outliers

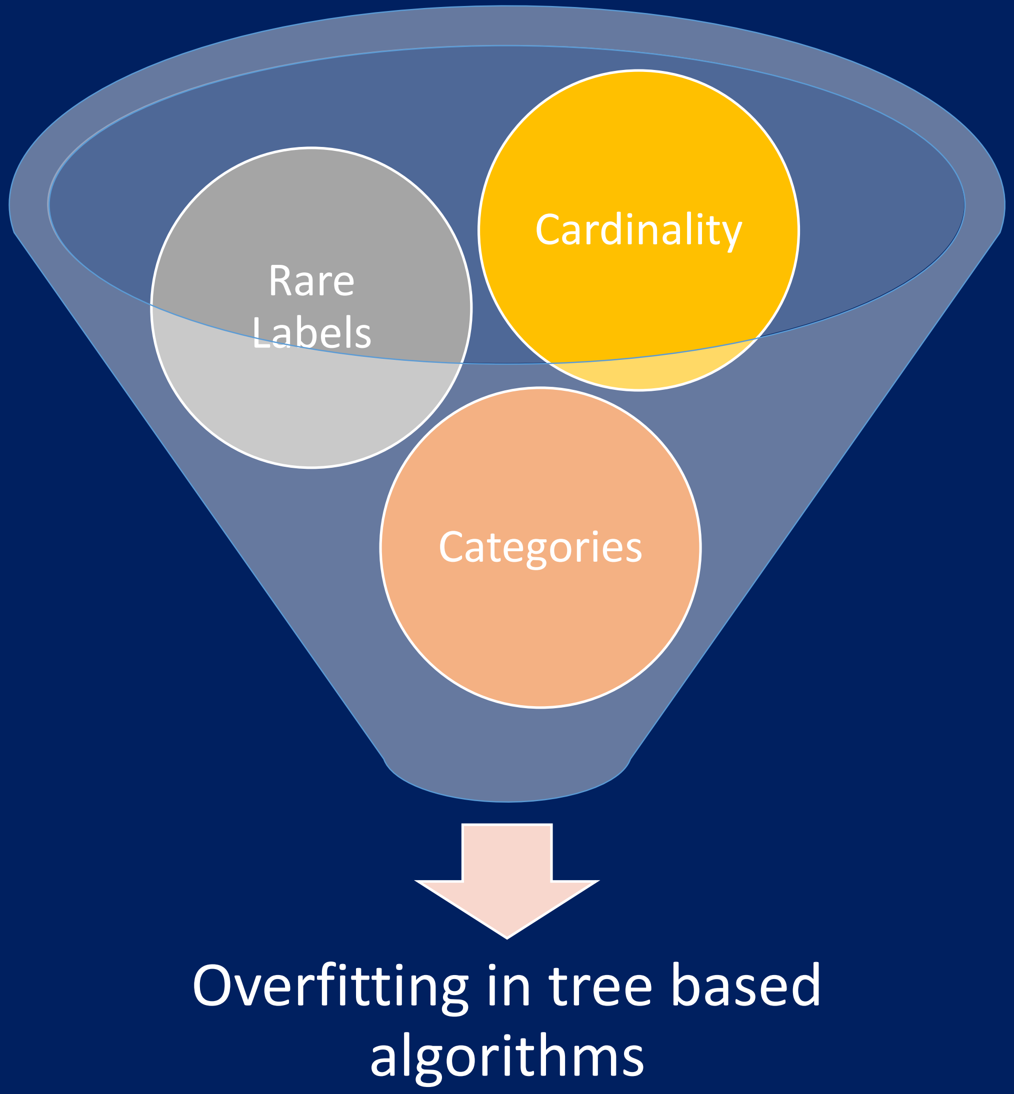
Unusual or unexpected values

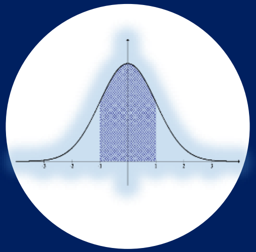
Missing Data

- Missing values for certain observations
- Affects all machine learning models
    - Scikit-learn

Labels in categorical variables

- Cardinality: high number of labels
- Rare Labels: infrequent categories
- Categories: strings
  - Scikit-learn

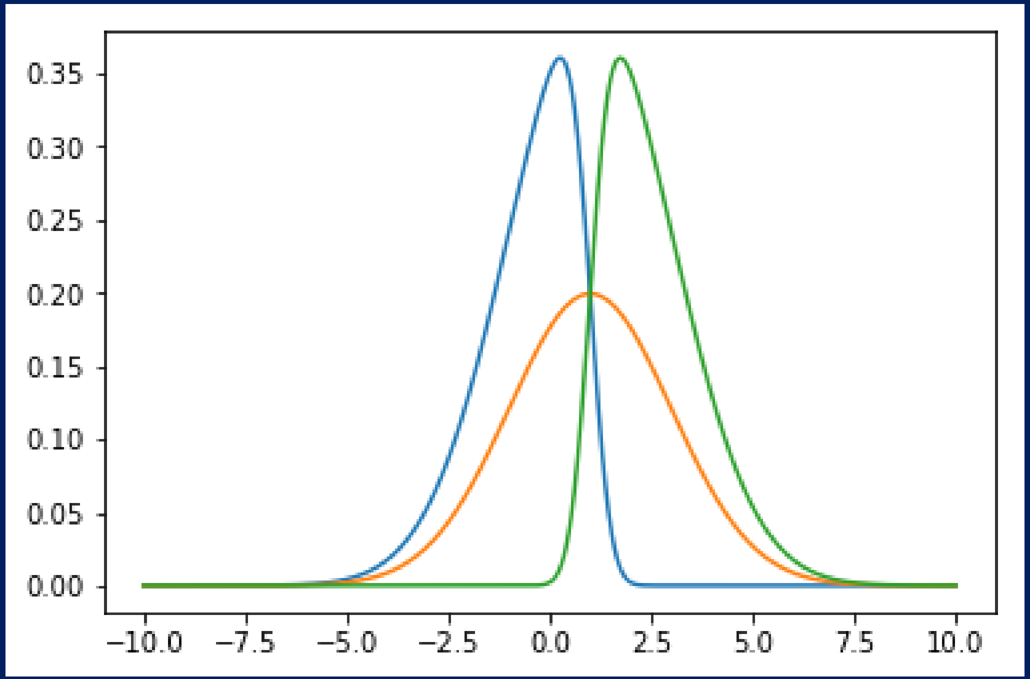




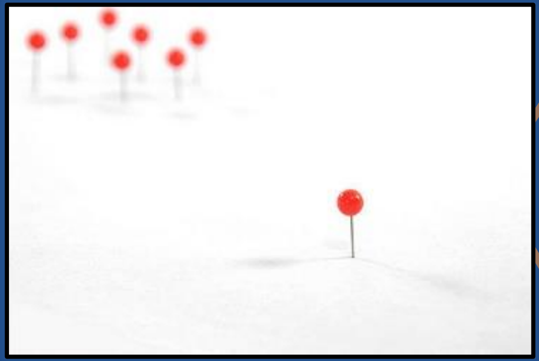
Distributions

- Linear model assumptions:
    - Variables follow a Gaussian distribution
- Other models: no assumption
    - Better spread of values may benefit performance

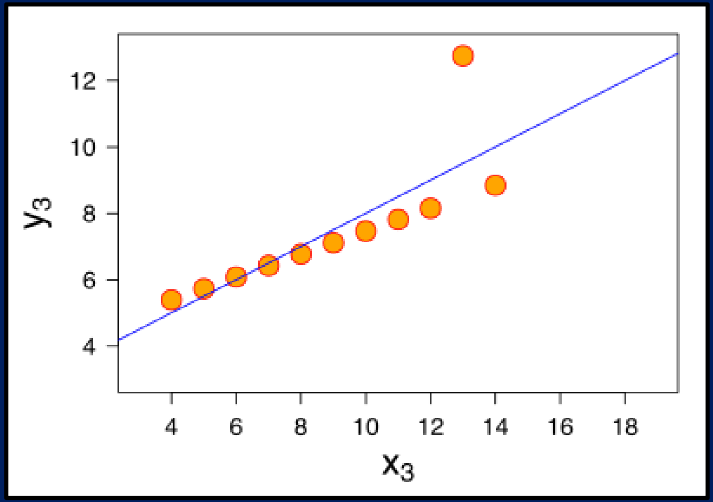
Gaussian vs Skewed



Outliers



Linear models

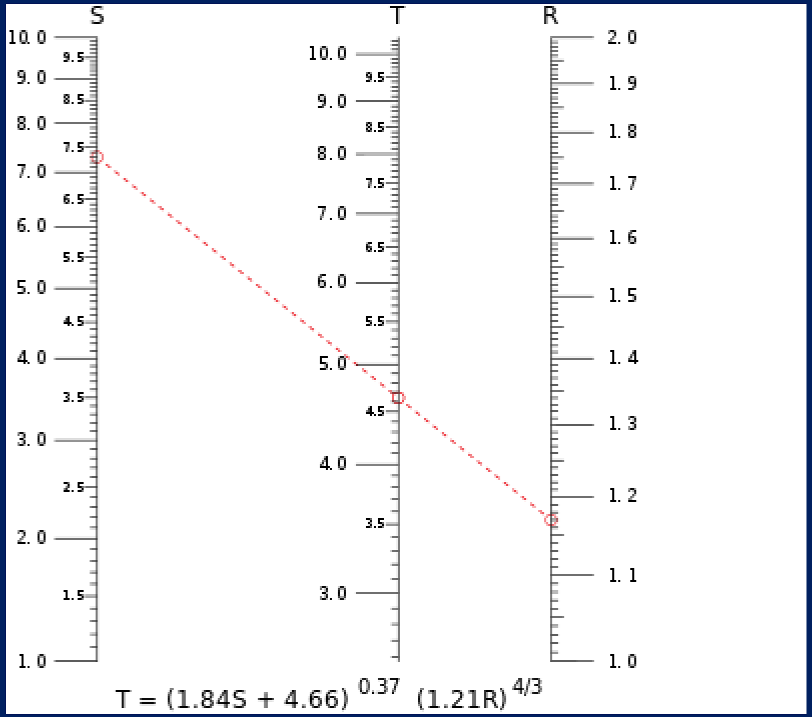


Adaboost

Tremendous weights

Bad generalisation

Feature Magnitude - Scale



Machine learning models sensitive to feature scale:

- Linear and Logistic Regression
- Neural Networks
- Support Vector Machines
- KNN
- K-means clustering
- Linear Discriminant Analysis (LDA)
- Principal Component Analysis (PCA)

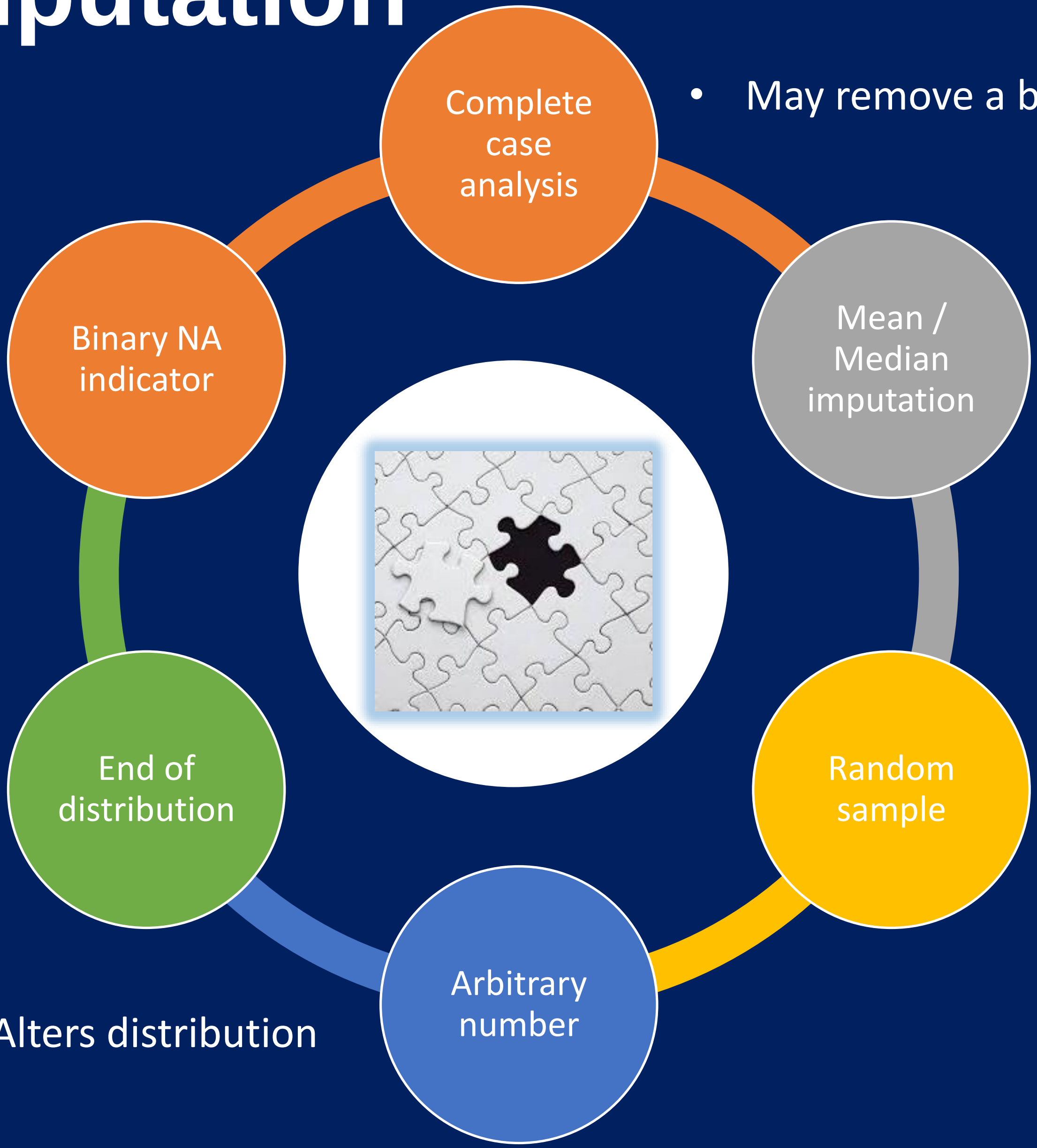
Tree based ML models insensitive to feature scale:

- Classification and Regression Trees
- Random Forests
- Gradient Boosted Trees

Data Pre-processing Journey

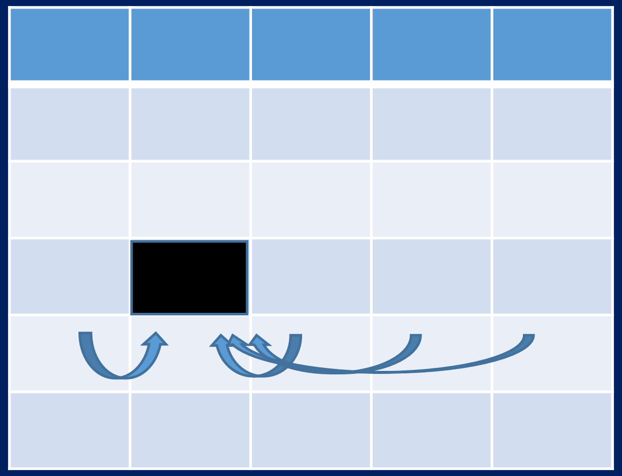
- Common issues found in variables
- **Feature / Variable engineering: solutions to the data issues**
- Feature selection: do we need to select features?
- Feature / Variable selection methods
- Overview and knowledge sources

Missing Data Imputation



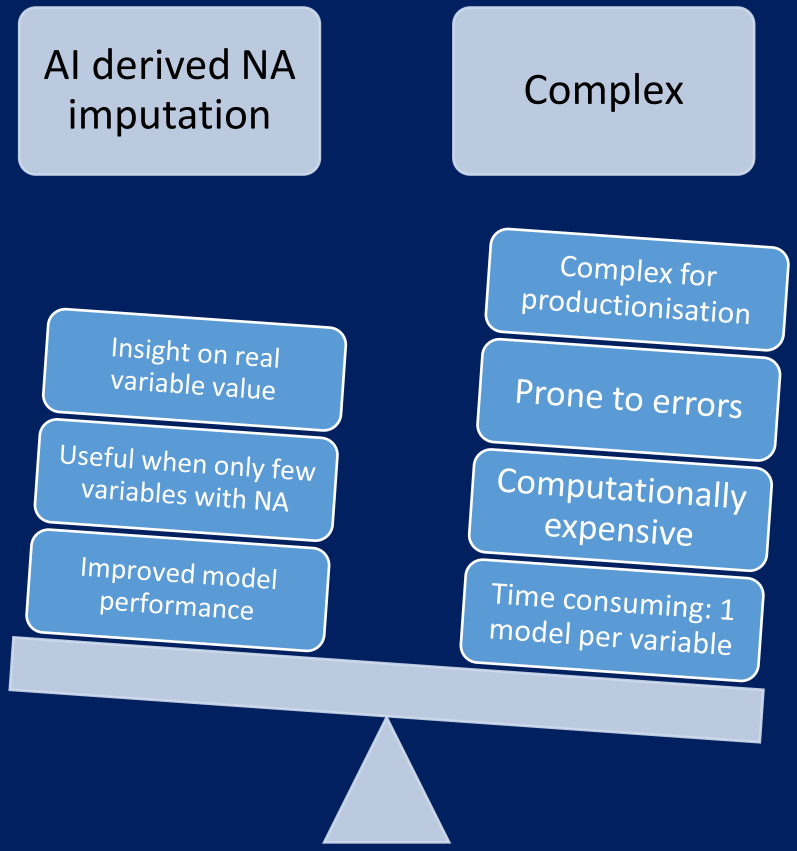
- May remove a big chunk of dataset
- Alters distribution
- Element of randomness
- Alters distribution
- Still need to fill in the NA

More on Missing Data Imputation



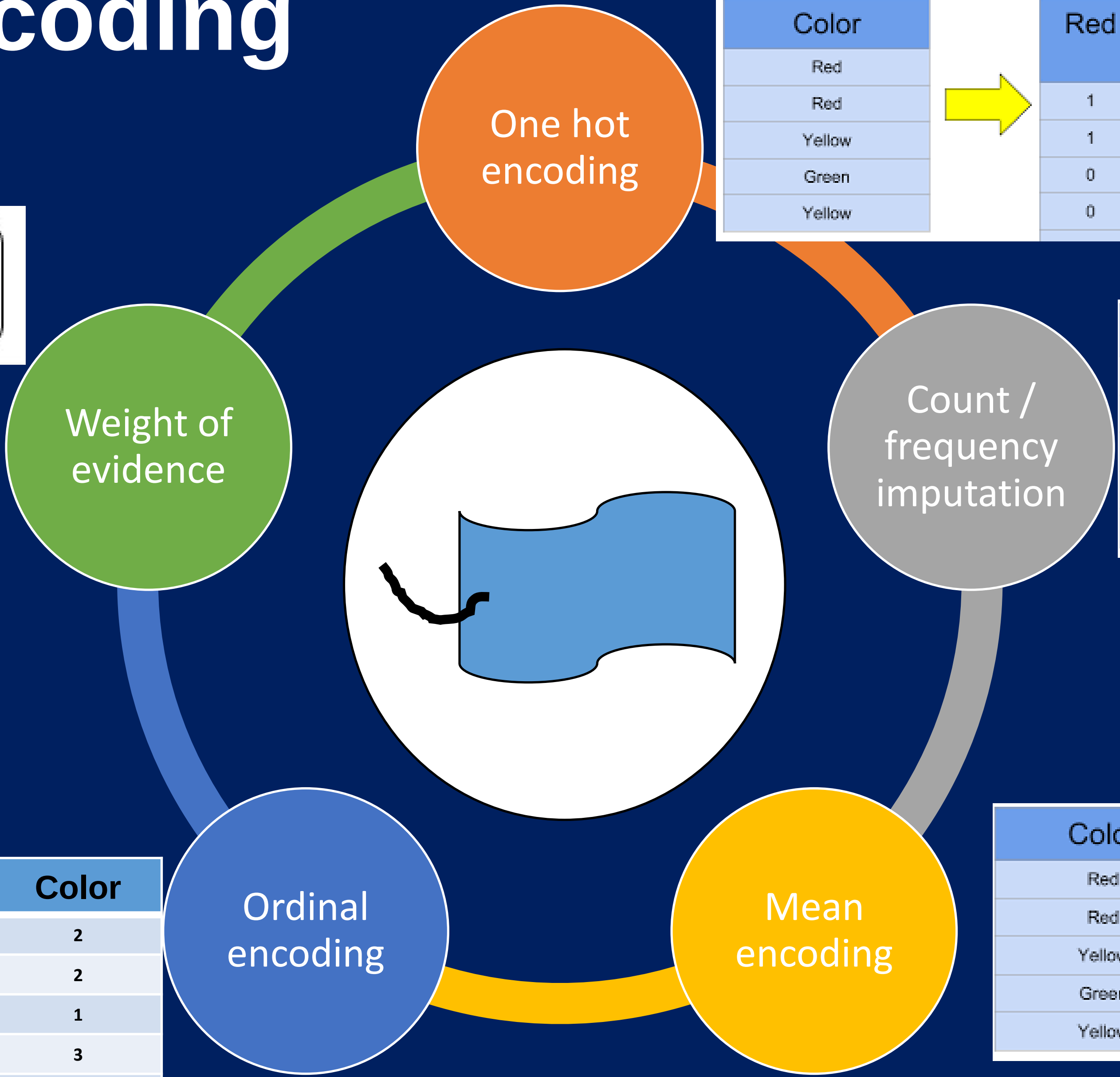
Use neighbouring variables to predict the missing value

- KNN
- Regression



Label Encoding

$$\text{WOE} = \ln\left(\frac{\%\text{ of non-events}}{\%\text{ of events}}\right)$$



Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

Color
Red
Red
Yellow
Green
Yellow

Color
2
2
2
1
2

Color
Red
Red
Yellow
Green
Yellow

Target	Color
0	0.5
1	0.5
1	1
0	0
1	1

Color
Red
Red
Yellow
Green
Yellow

Target	Color
0	2
1	2
1	1
0	3
1	1

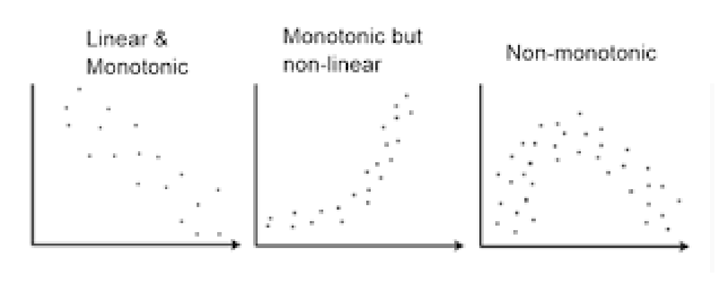
Label Encoding

One hot encoding

- Expands the feature space

Count / frequency imputation

- No monotonic relationship



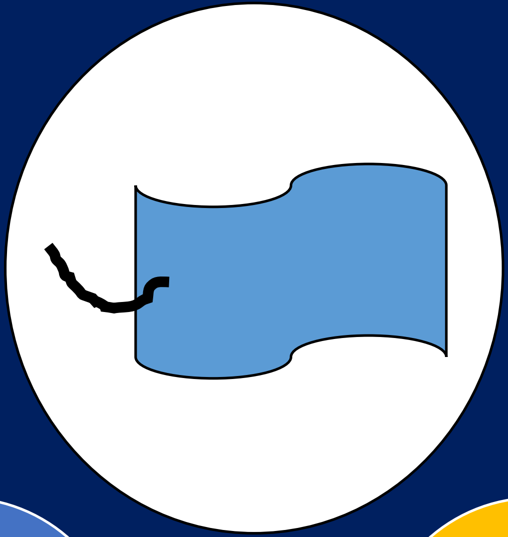
Mean encoding

Ordinal encoding

- Prone to overfitting

Weight of evidence

- Account for zero values as it uses logarithm



Rare Labels

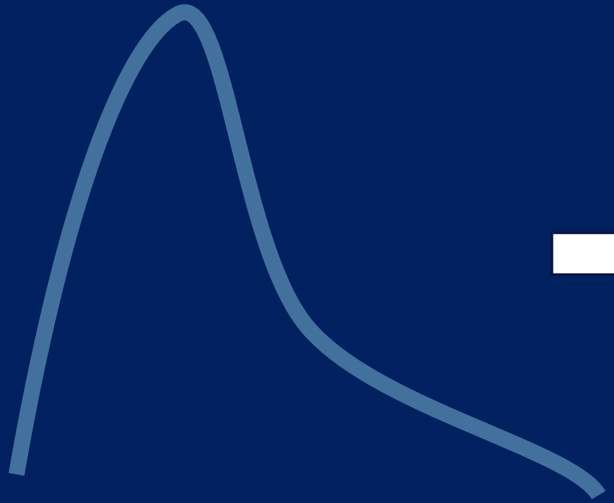
Distribution: Gaussian Transformation

Skewed Gaussian

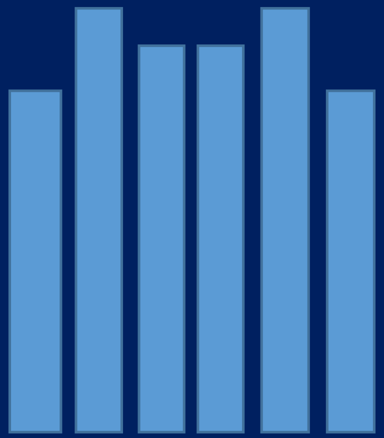
Variable transformation

- Logarithmic → ln(x)
- Exponential → x Exp (any power)
- Reciprocal → (1 / x)
- Box-Cox → (x Exp (λ) – 1) / λ
    - λ varies from -5 to 5

Distribution: Discretisation



Skewed



Improved value spread

Discretisation

- Equal width bins
  - Bins ➔ (max – min) / n bins
  - Generally does not improve the spread
- Equal frequency bins
  - Bins determined by quantiles
  - Equal number of observations per bin
  - Generally improves spread

Outliers

Trimming	• Remove the observations from dataset
Top | bottom coding	• Cap top and bottom values
Discretisation	• Equal bin / equal width

Data Pre-processing Journey

- Common issues found in variables
- Feature / Variable engineering: solutions to the data issues
- **Feature selection: do we need to select features?**
- Feature / Variable selection methods
- Overview and knowledge sources

Why Do We Select Features?

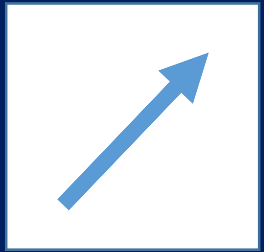
- Simple models are easier to interpret
- Shorter training times
- Enhanced generalisation by reducing overfitting
- Easier to implement by software developers → Model production
- Reduced risk of data errors during model use
- Data redundancy

Variable Redundancy

Constant variables
Only 1 value per variable

Quasi – constant Variables
> 99% of observations show same value

Duplication
Same variable multiple times in the dataset

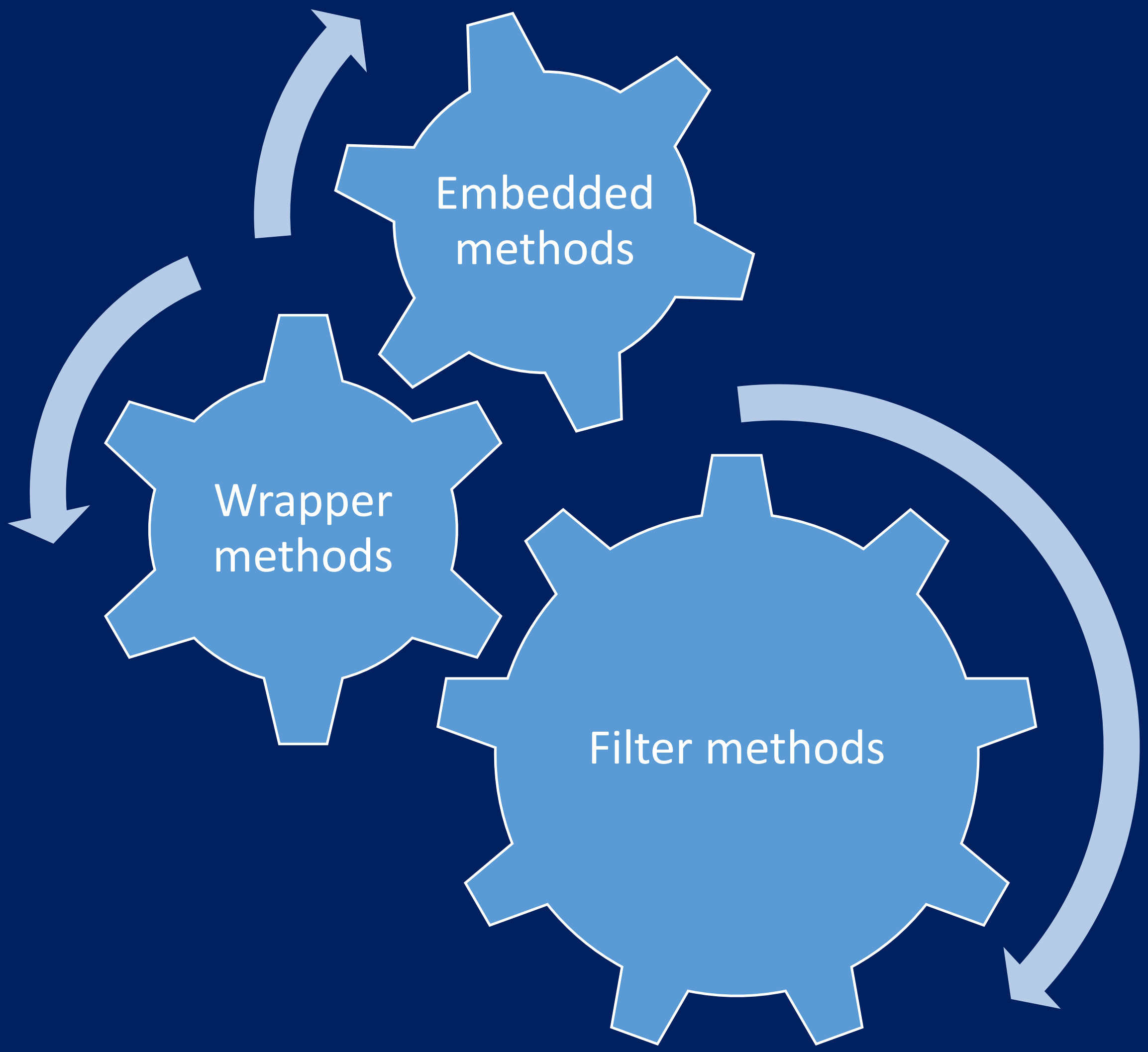


Correlation
Correlated variables provide the same information

Data Pre-processing Journey

- Common issues found in variables
- Feature / Variable engineering: solutions to the data issues
- Feature selection: do we need to select features?
- **Feature / Variable selection methods**
- Overview and knowledge sources

Feature Selection Methods

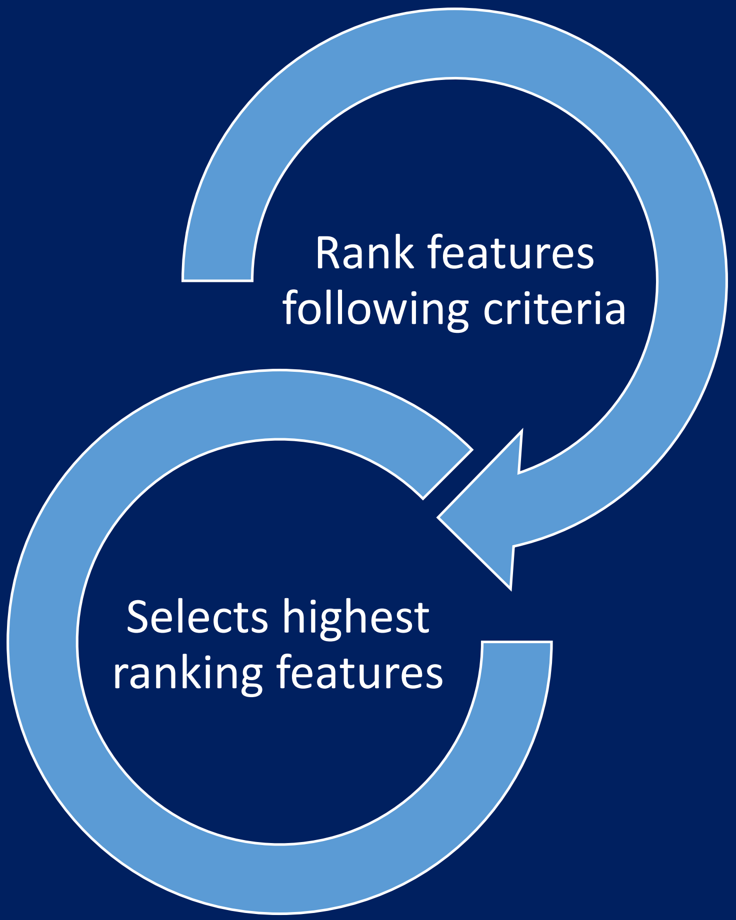


Filter methods

- Independent of ML algorithm
- Based only on variable characteristics

Pros	Cons
Quick feature removal	Does not capture redundancy
Model agnostic	Does not capture feature interaction
Fast computation	Poor model performance

Filter methods



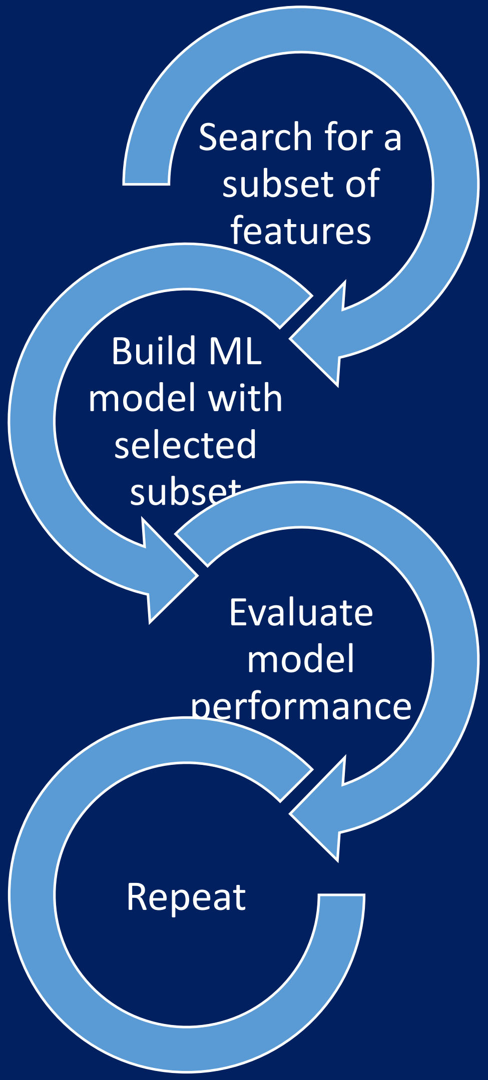
- Chi-square | Fisher Score
- Univariate parametric tests (anova)
- Mutual information
- Variance

Wrapper methods

- Considers ML algorithm
- Evaluates subsets of features

Pros	Cons
Considers feature interaction	Not model agnostic
Best performance	Computation expensive
Best feature subset for a given algorithm	Often impracticable

Wrapper methods



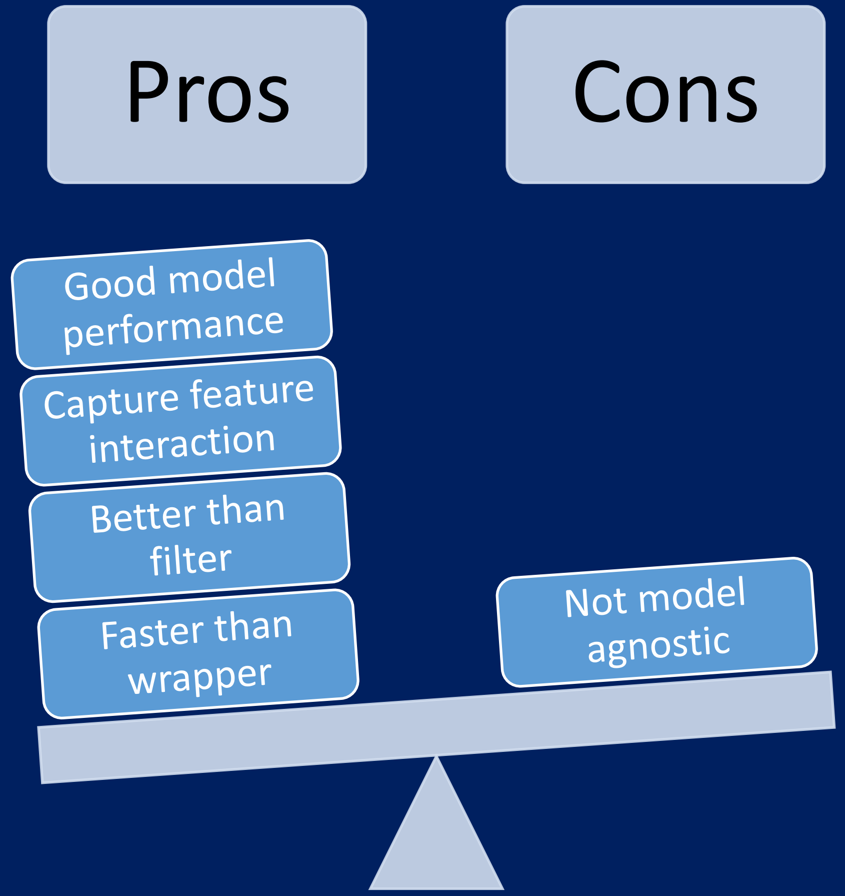
Forward feature selection
- Adds 1 feature at a time

Backward feature elimination
- Removes 1 feature at a time

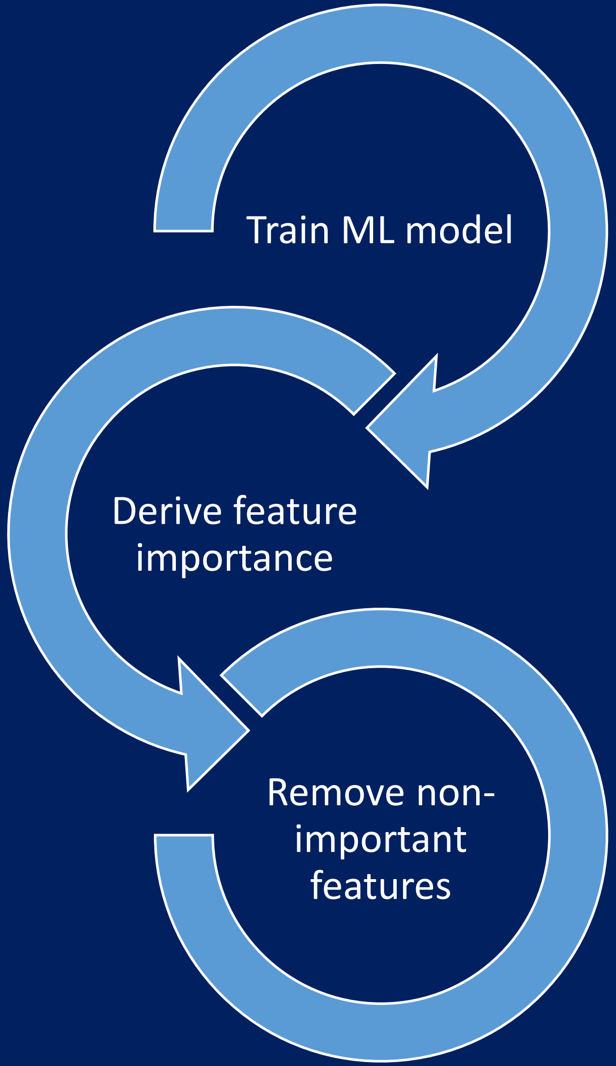
Exhaustive feature search
- Searches across all possible feature combinations

Embedded methods

Feature selection during training of ML algorithm



Embedded methods

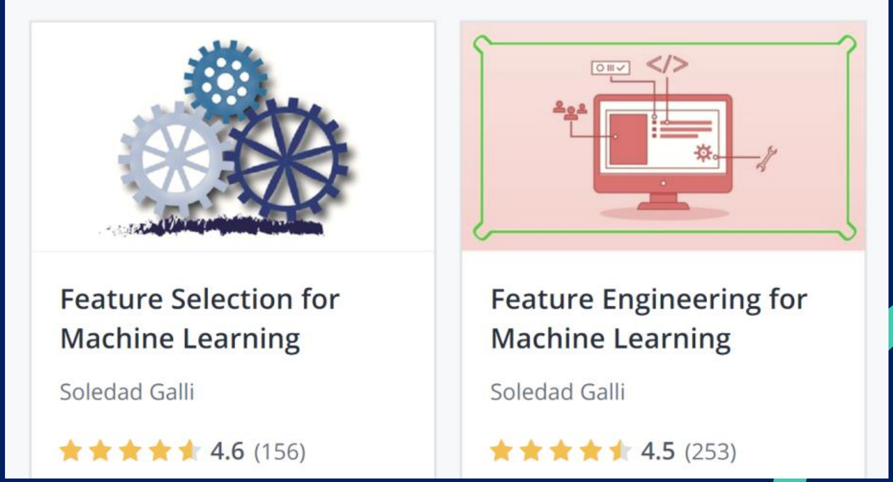


- LASSO
- Tree derived feature importance
- Regression coefficients

Data Pre-processing Journey

- Common issues found in variables
- Feature / Variable engineering: solutions to the data issues
- Feature selection: do we need to select features?
- Feature / Variable selection methods
- Overview and knowledge sources

Knowledge Resources



Udemy.com, includes code

The 2009 Knowledge Discovery in Data Competition (KDD Cup 2009)
Challenges in Machine Learning, Volume 3

Gideon Dror, Marc Boullé, Isabelle Guyon,
Vincent Lemaire, and David Vogel, editors

Summary of learnings from the winners

Feature Engineering + Selection

scikit learn

Feature Engine

Python package for feature engineering
Work in progress

MLXTEND