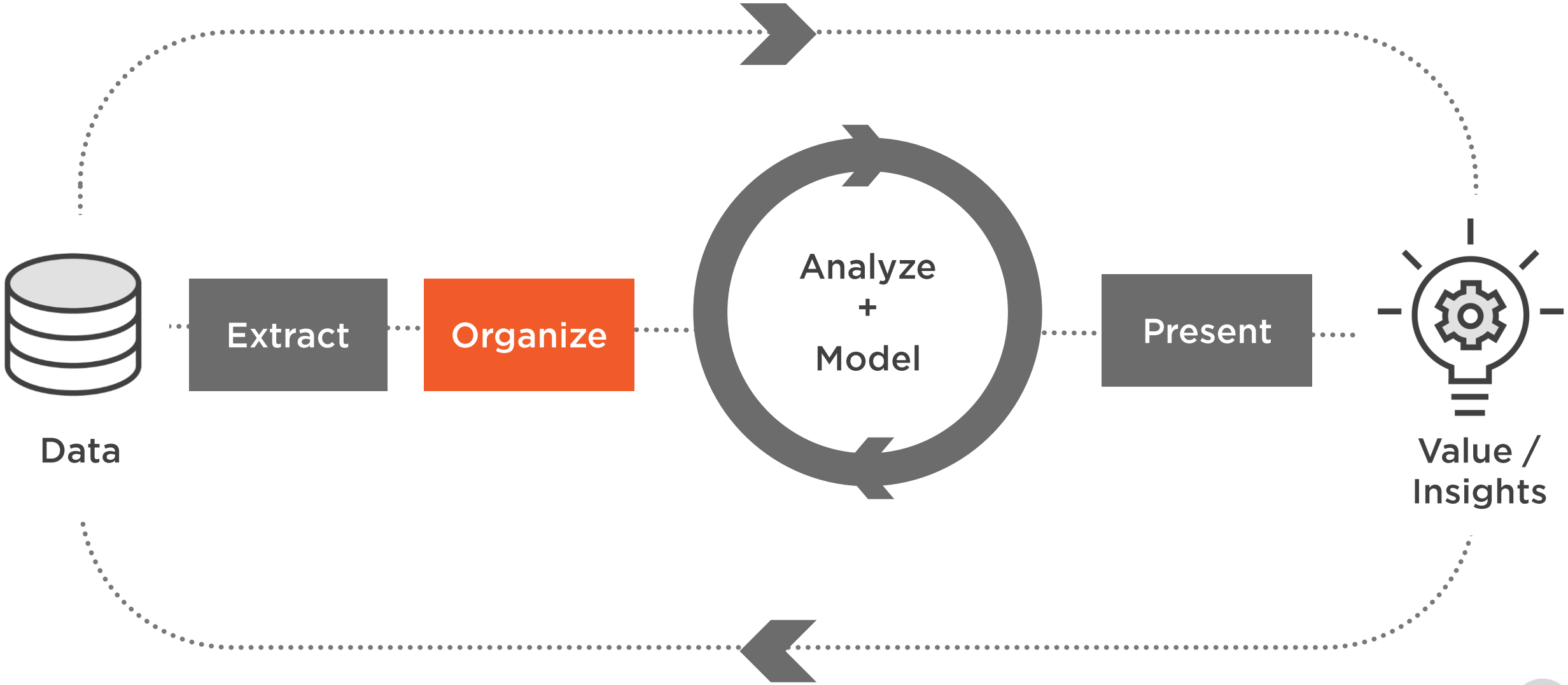
Exploring and Processing Data – Part 3

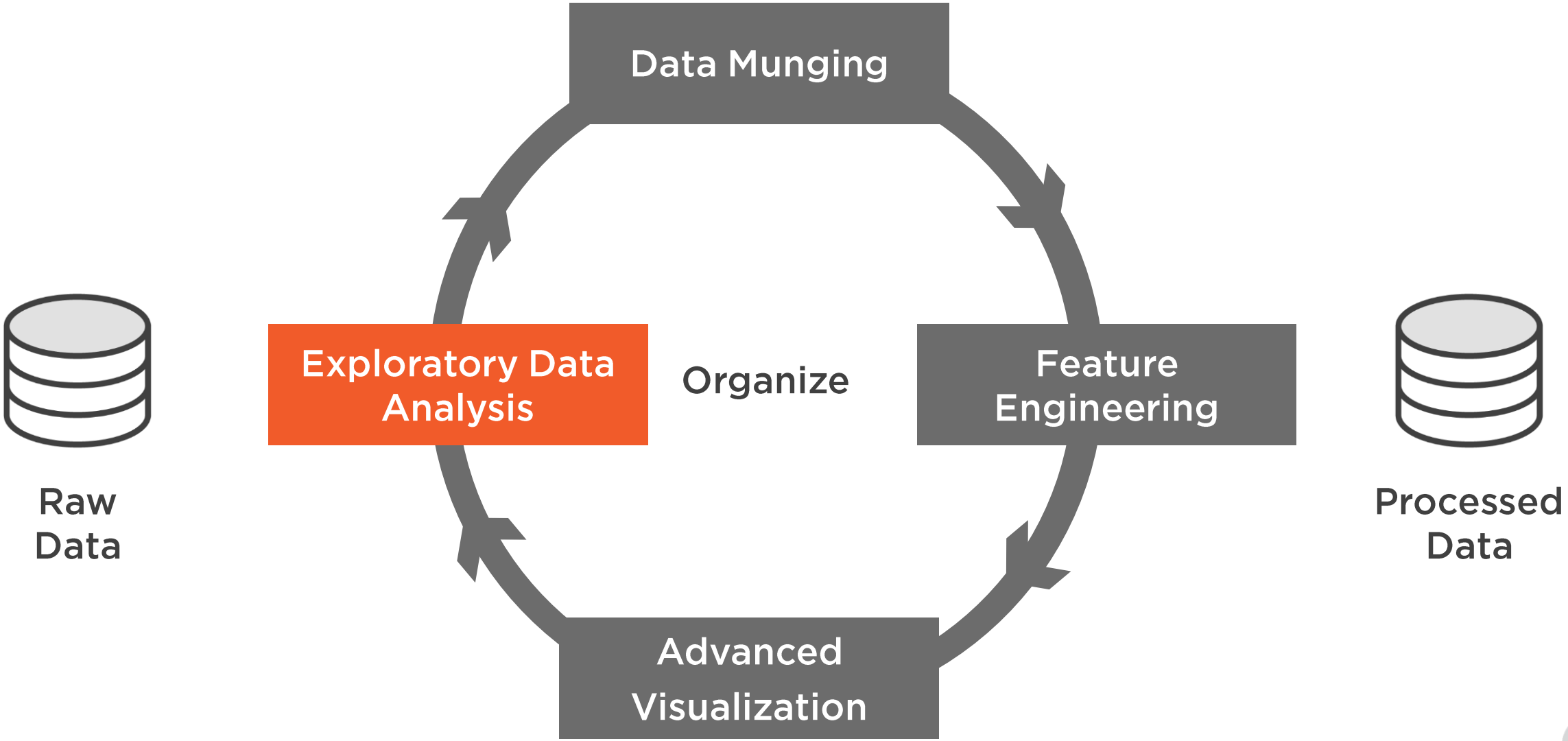
Abhishek Kumar
AUTHOR

@meabhishekkumar

Data Science Project Cycle



Organize


Data Munging
Exploratory Data Analysis
Organize
Feature Engineering
Advanced Visualization
Raw Data
Processed Data

Overview (Concepts)

Data munging

- Treating missing values
- Working with outliers

Feature engineering

- Derived features
- Categorical Feature encoding

Reproducible script

Visualization

Overview (Tools)

Python

- NumPy
- Pandas
- Matplotlib

Data Munging

Data Issues

Missing values

Extreme values (outliers)

Erroneous values

Missing Value

Value not known

Very common in real world

Reasons

- Non availability
- Manual data entry process
- Equipment error

Missing Value

Issue

- Inaccurate analysis
- Modeling won’t work in many cases

Solution

- Deletion
- Imputation

Mean Imputation

Age	10	10	11	9	10	8	12	12	8	?

Mean = 90 / 9 = 10

Age	10	10	11	9	10	8	12	12	8	10

Median Imputation

Age	10	10	11	9	10	8	12	12	8	?

Median = 10

Age	10	10	11	9	10	8	12	12	8	10

Mode Imputation

Class	1	1	2	2	3	3	?	1	1	3

Mode = 1

Class	1	1	2	2	3	3	1	1	1	3

Forward / Backward Fill

Forward Fill

Data	1	1	1	2	2	2	?	3	3	3

Backward Fill

Data	1	1	1	2	2	2	?	3	3	3

Predictive Model

x1	x.	xn
2.0	?	1
?	70.0	2
3.0	78.0	1

Predictive Model

x1	x.	xn
2.0	60.0	1
1.5	70.0	2
3.0	78.0	3

Demo

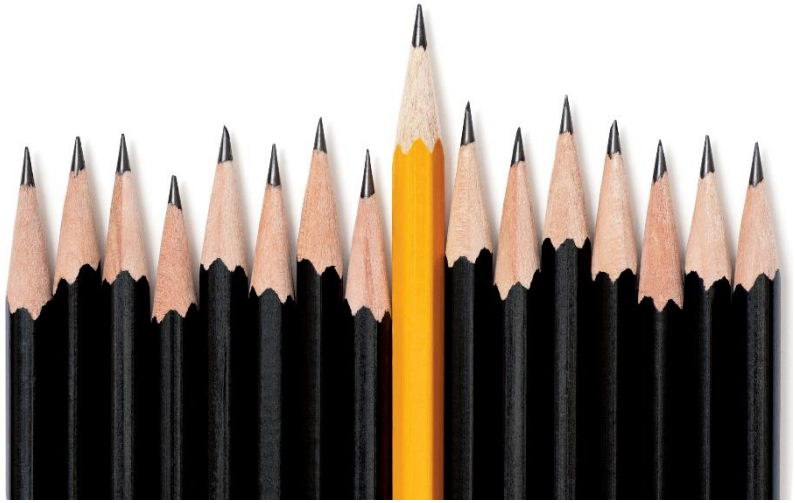
Treating missing values using Pandas - Part 1

Demo

Treating missing values using Pandas – Part 2

Demo

Treating missing values using Pandas – Part 3



Outliers

Different from normal

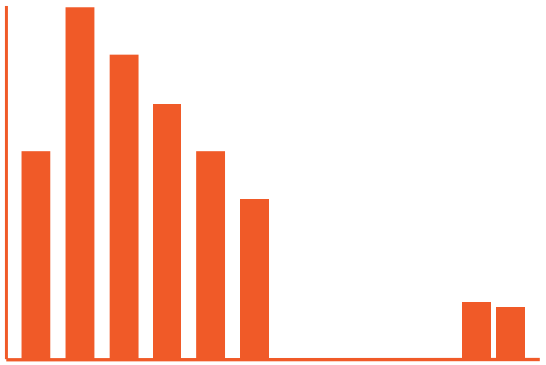
Multiple source

- Data entry
- Data processing
- Natural

Issue

- Biased analysis
- Biased models

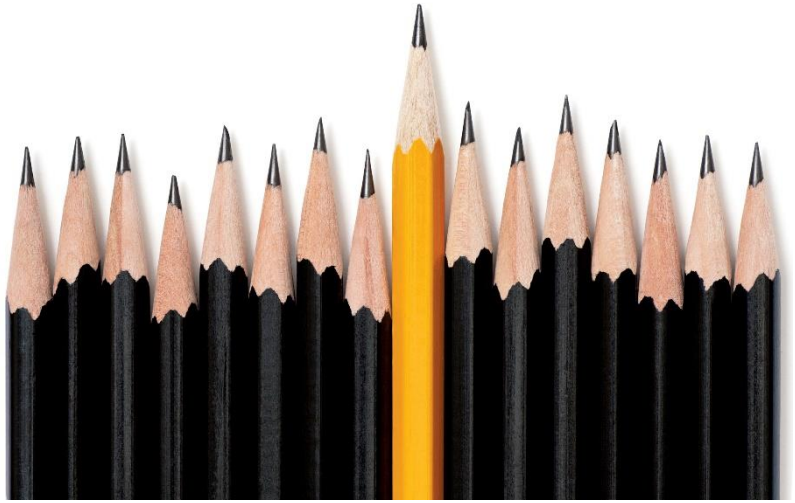
Outlier Detection



Histogram

Boxplot

Scatterplot



Outlier Treatment

- Removal
- Transformation
- Binning
- Imputation

Demo

Detecting and treating outliers using Pandas and NumPy

Feature Engineering

Feature Engineering

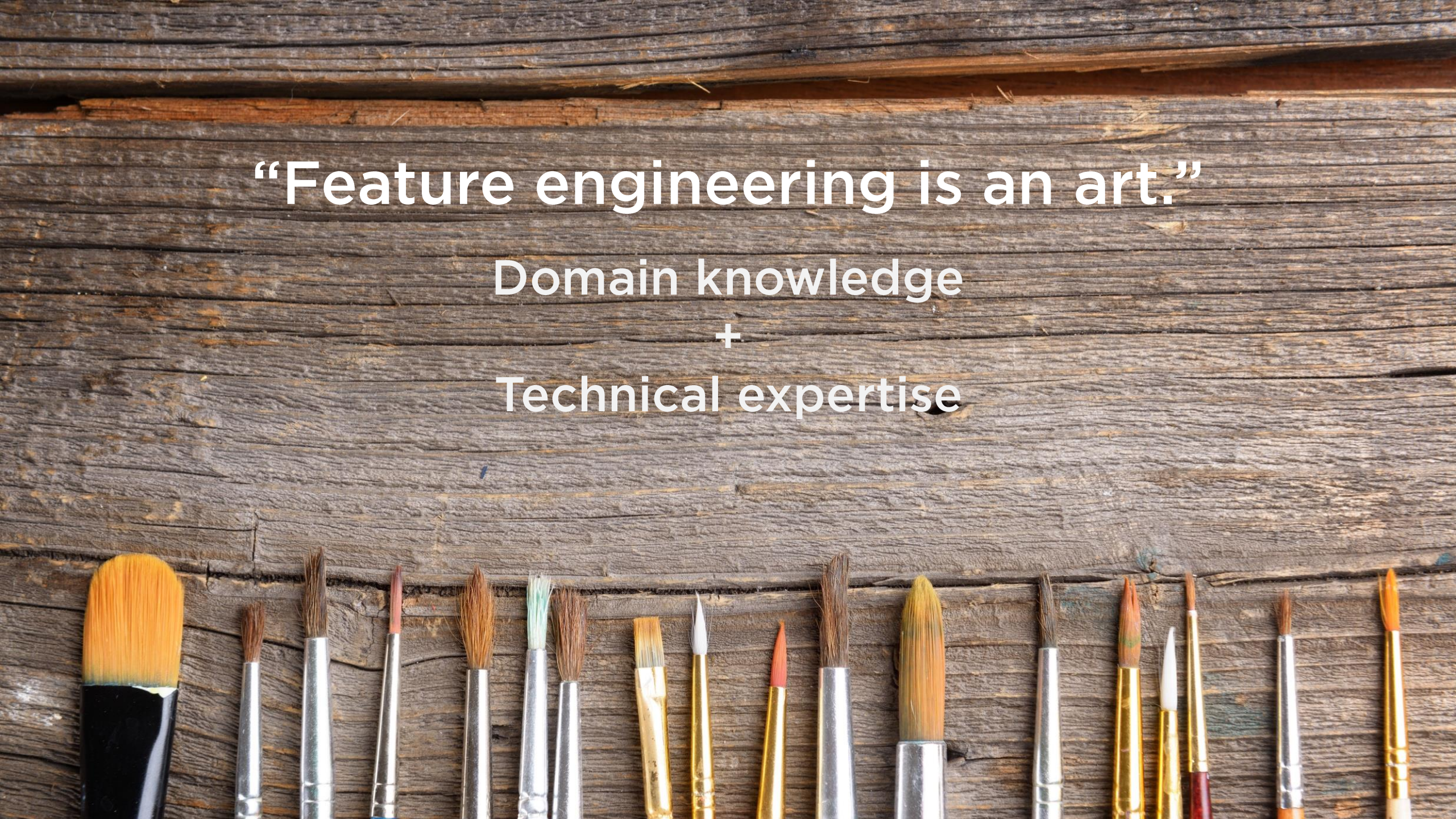
Process of transforming raw data to better representative features in order to create better predictive models

Feature Engineering

Transformation

Creation (using domain expertise)

Selection


“Feature engineering is an art.”
Domain knowledge
+
Technical expertise

Demo

Feature creation using Pandas and NumPy – Part 1

Demo

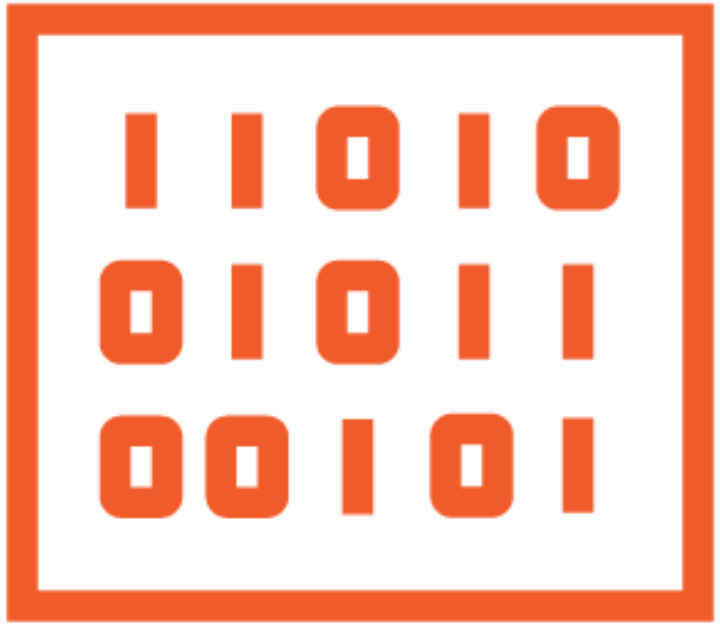
Feature creation using Pandas and NumPy – Part 2

Demo

Feature creation using Pandas and NumPy – Part 3

Demo

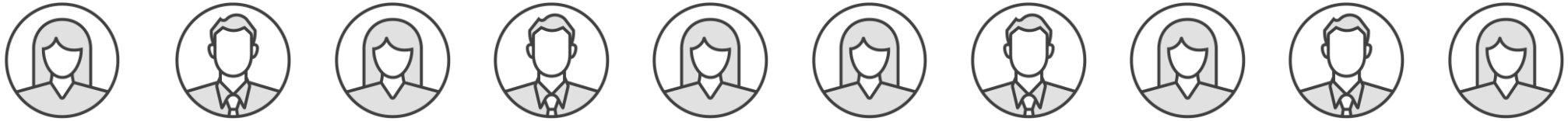
Feature creation using Pandas and Numpy – Part 4



Categorical Feature Encoding

Converting categorical feature to numerical feature

Binary Encoding



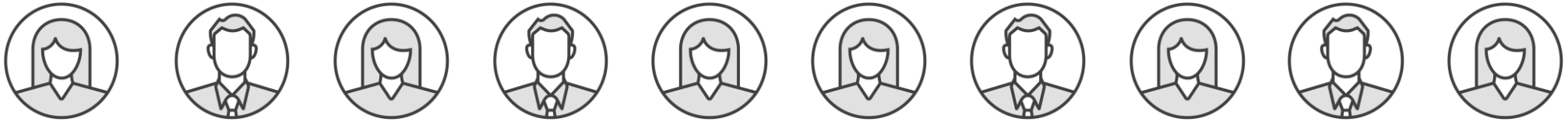
Gender	F	M	F	M	F	F	M	F	M	F
Is_Male	0	1	0	1	0	0	1	0	1	0
Is_Female	1	0	1	0	1	1	0	1	0	1

Label Encoding

Fare	Low	Medium	High	High	Low	Medium	High	Low	High	Low
Fare	1	2	3	3	1	2	3	1	3	1

Label	Encoded Value
Low	1
Medium	2
High	3

One-Hot Encoding



Embarked	A	B	A	B	C	A	B	C	A	B
Is_A	1	0	1	0	0	1	0	0	1	0
Is_B	0	1	0	1	0	0	1	0	0	1
Is_C	0	0	0	0	1	0	0	1	0	0

Label	Is_A	Is_B	Is_C
A	1	0	0
B	0	1	0
C	0	0	1

Demo

Categorical feature encoding using Pandas

Demo

Drop and reorder columns using Pandas

Demo

Save dataframe to file using Pandas

Demo

Reproducible script for data processing using Pandas and NumPy

Demo

Creating visualization using Matplotlib

Demo

Committing changes to git

Summary

Data munging

Feature engineering

Matplotlib