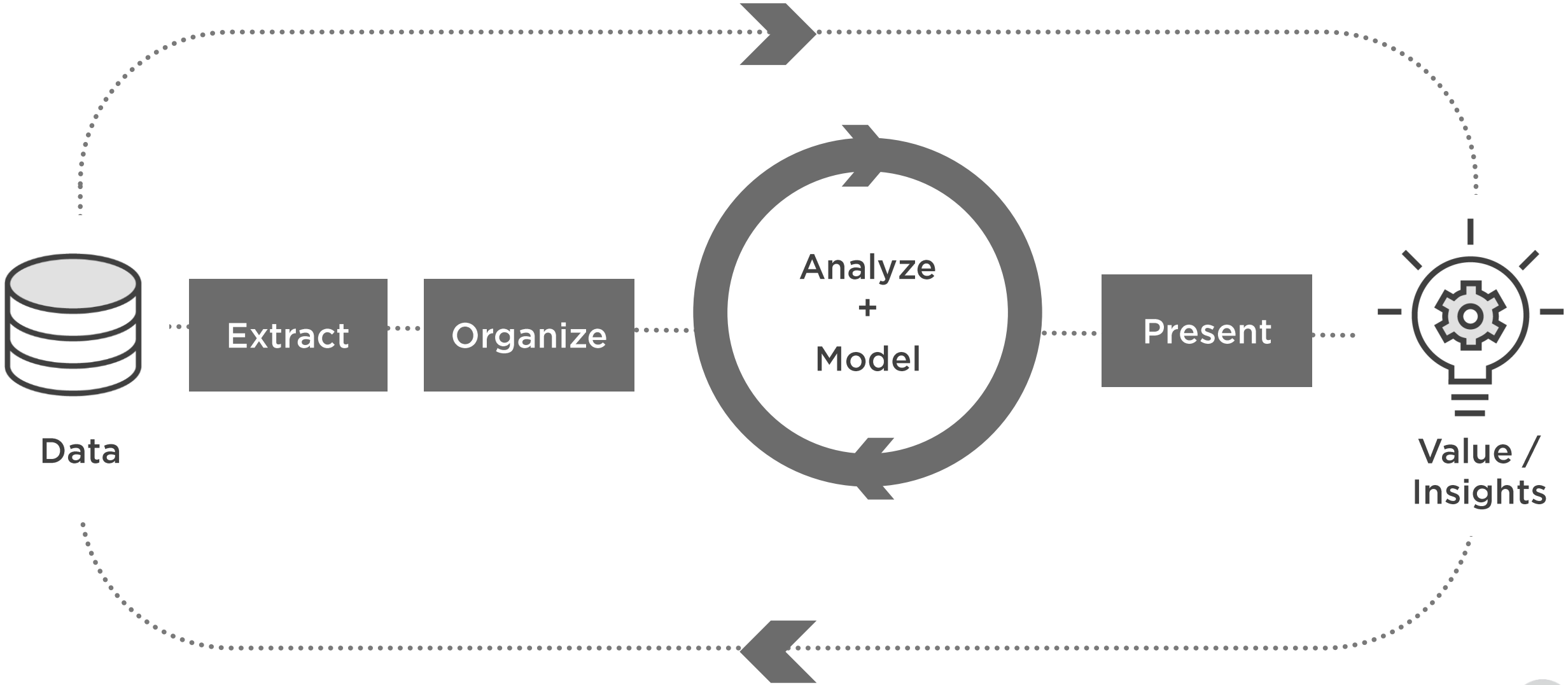
Exploring and Processing Data – Part 1

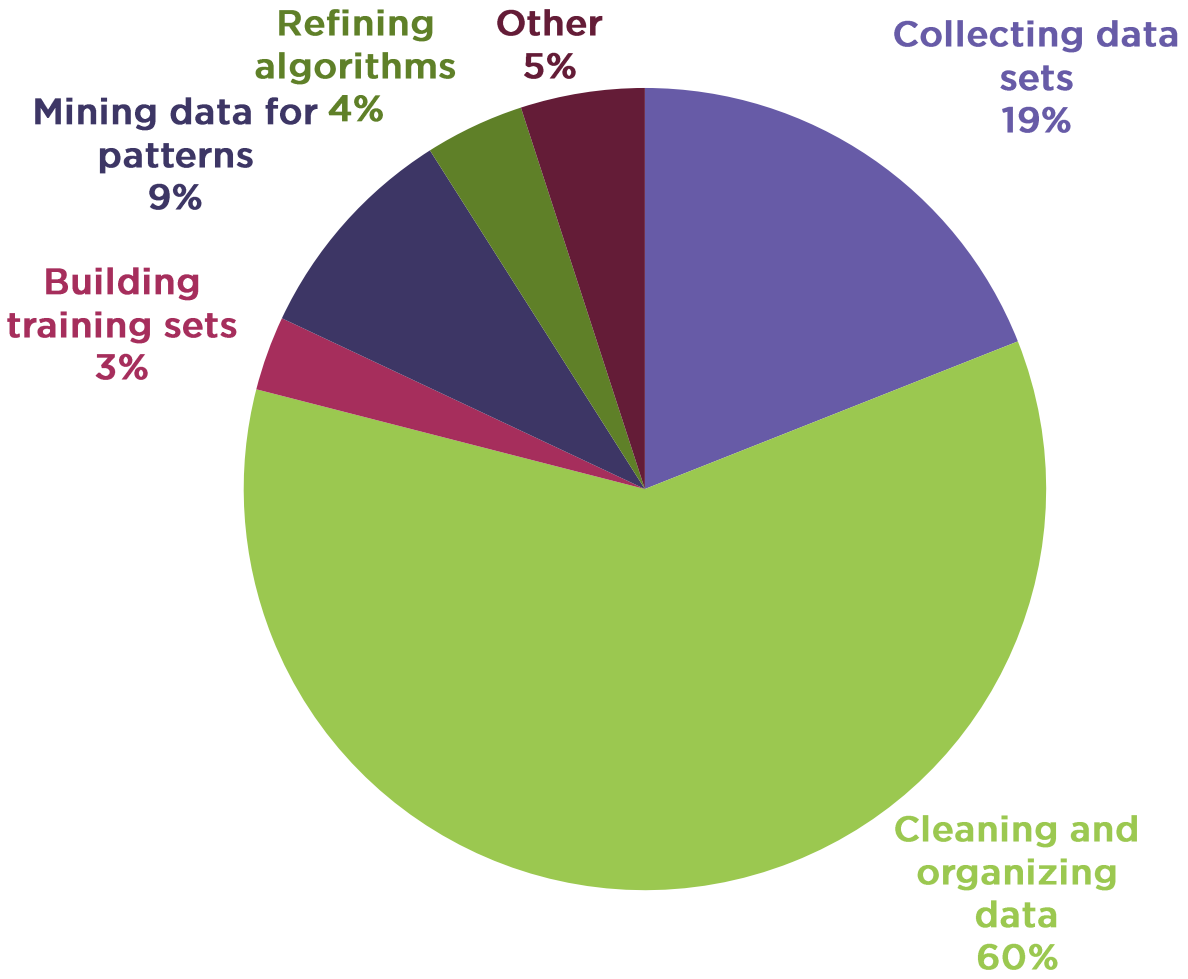
Abhishek Kumar
AUTHOR

@meabhishekkumar

Data Science Project Cycle

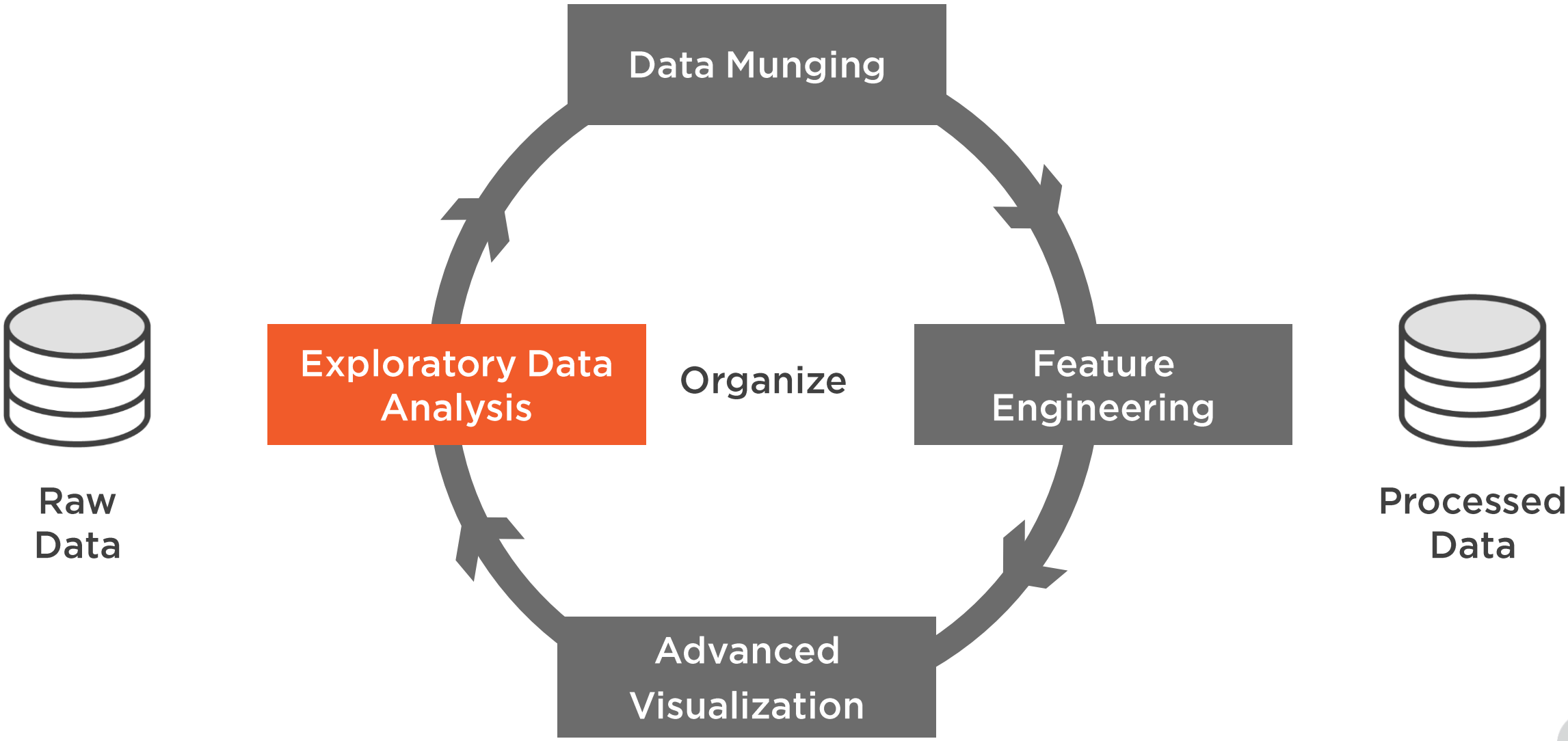


Where Data Scientists Spend Their Time?



Source : Crowdflower data science report 2016 (link)

Organize


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

Exploratory Data Analysis

Basic structure

Summary statistics

Distributions

Grouping

Crosstabs, Pivots

Overview

(Concepts)

Import Data

Exploratory data analysis

- Basic structure
- Summary statistics
- Distributions
- Grouping
- Crosstabs
- Pivots

Overview (Tools)

Python

- NumPy
- Pandas

NumPy

- Fundamental tool for scientific computing
- Very efficient array operations
- Work on multi-dimensional arrays and matrices
- High level mathematical functions

Pandas

Built on top of NumPy

Data structure and operations on tabular data (Pandas dataframe)

Data visualization using Matplotlib

	Column - 1	...	Column - n
Row 1	...	...	...
Row ..	...	...	...
Row m	...	...	...

Exploratory Data Analysis

Exploratory Data Analysis

Basic structure

Summary statistics

Distributions

Grouping

Crosstabs, Pivots

Basic Structure

- How many rows or observations?
- How many columns or features?
- Column data types
- Explore head or tail

Demo

Investigating basic structure using Pandas

PassengerId	◂ Passenger ID
Survived	◂ If Survived (1 – yes, 0 – no)
Pclass	◂ Passenger class (1 – 1st class, 2 – 2nd class, 3 – 3rd class)
Name	◂ Name
Sex	◂ Gender
Age	◂ Age

SibSp	Number of siblings / spouses aboard
Parch	Number of parents / children aboard
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin
Embarked	Point of embarkment (C = Cherbourg; Q = Queenstown; S = Southampton)

Demo

Selection, indexing and filtering using Pandas

Exploratory Data Analysis

Basic structure

Summary statistics

Distributions

Grouping

Crosstabs, Pivots

Summary Statistics

Numerical

- Centrality measure (mean, median)
- Dispersion measure (range, percentiles, variance , standard deviation)

Categorical

- Total count
- Unique count
- Category Counts and proportions
- Per category statistics

Centrality Measure

One number to represent entire set of values

Number central to the data

Central tendency

Mean / Average

Average behavior

Centrality Measure : Mean or Average

10 10 11 9 10 8 12 12 8 10

Mean age : sum of ages / count = 100 / 10 = 10

Problem : Affected by extreme values

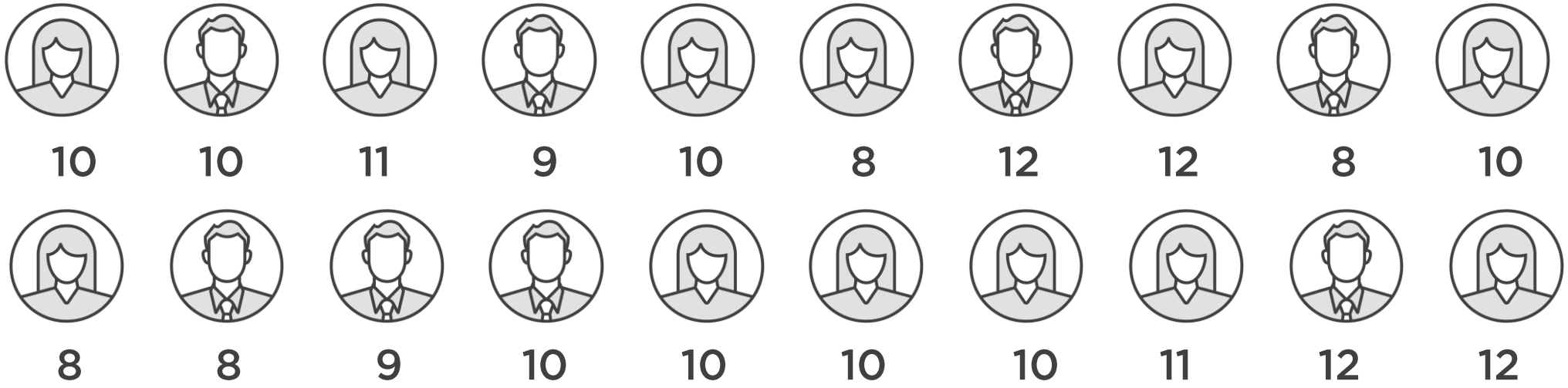
10 10 11 9 10 8 12 12 8 10 98

Mean age : sum of ages / count = 198 / 11 = 18

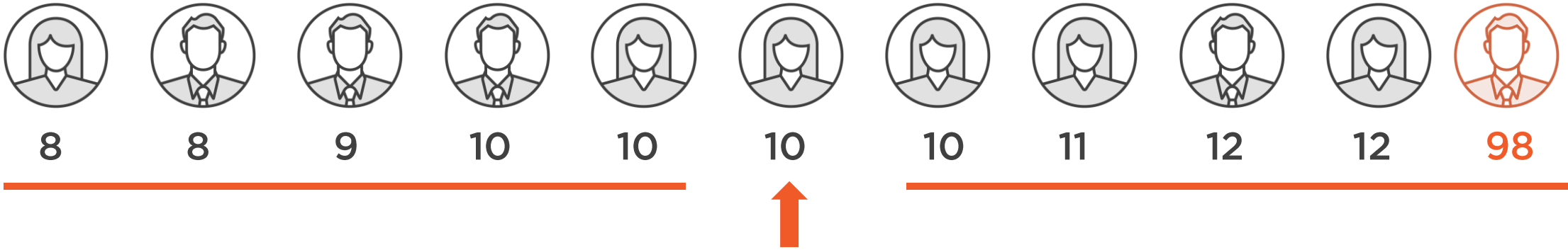
Median

Middle value in the sorted list

Centrality Measure : Median



Median = (10 + 10) / 2 = 10



Median = 10

Spread / Dispersion Measure

How spread out values are from central value

Variability

Range

Difference between maximum and minimum

Spread : Range

10 10 11 9 10 8 12 12 8 10

Age range : max – min = 12 – 8 = 4

Problem : Affected by extreme values

10 10 11 9 10 8 12 12 8 10 98

Age range : max – min = 98 – 8 = 90

Percentiles

x percentile is y means x% of values are below y

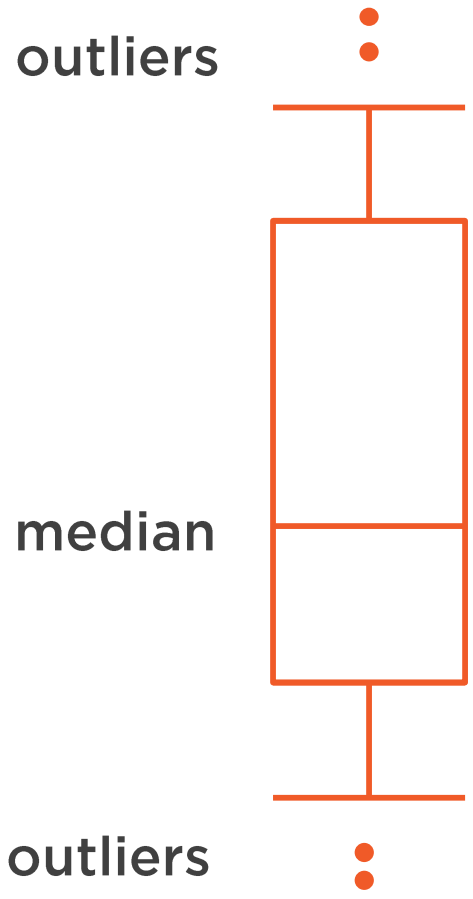
50 percentile is 10 means 50% of values are below 10

25^{th}, 50^{th} , 75^{th}

- Bucket – 1 : Below 25^{th}
- Bucket – 2 : 25^{th} – 50^{th}
- Bucket – 3 : 50^{th} - 75^{th}
- Bucket – 4 : above 75^{th}

Quartiles

Box-Whisker Plot



Max (excluding extreme values)

Bucket – 4 : Above 75th

75th

Bucket – 3 : 50th – 75th

50th

Bucket – 2 : 25th – 50th

25th

Bucket – 1 : Below 25th

Min (excluding extreme values)

Inter-quartile range (IQR)

Outliers = Beyond 1.5 x IQR

Variance

Measure of variability

How far each value in list from mean value

Small variance = less spread

High variance = large spread

Variance = $\frac{sum((value - mean)^2)}{count}$

Affected by extreme values

Unit is not clear

Standard Deviation

Standard deviation = $\sqrt{variance}$

Unit is same as that of the feature

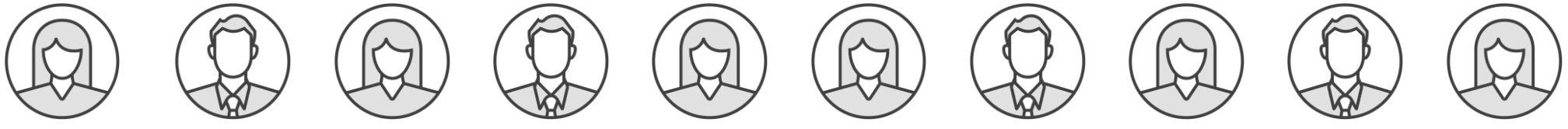
Low standard deviation = less spread

High standard deviation = large spread

Demo

Getting summary statistics for numerical features using Pandas and NumPy

Counts and Proportions



Total count : 10

Unique count : 2

Gender	Count	Proportion
M	4	4 / 10 = 0.4
F	6	6 / 10 = 0.6

Demo

Summary statistics for categorical feature using Pandas and NumPy

Summary

- Import data
- Basic structure
- Summary statistics