Representing, Processing, and Preparing Data

UNDERSTANDING DATA CLEANING AND PREPARATION TECHNIQUES

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Identifying problems that hinder analytics

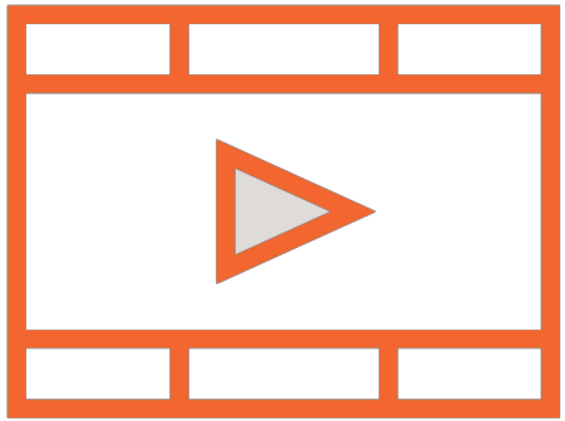
Common technology tools to work with data

Dealing with missing data

Dealing with outliers and erroneous data

Prerequisites and Course Outline

Prerequisites



- **Basic Python programming**
- **Basic Excel spreadsheets**
- **Basic SQL for relational databases**
- **High school math**

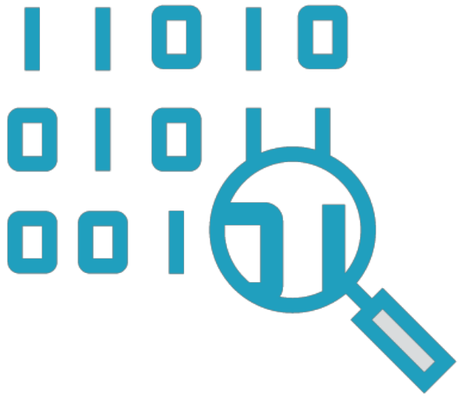
Course Outline

- Data cleaning and preparation techniques
- Processing data using spreadsheets and Python
- Collecting data to extract insights
- Processing data using relational databases
- Representing insights from data

“My mind is made up. Don’t confuse me with the facts.”

Some powerful person

Thoughtful, Fact-based Point of View



Fact-based

Built with painstakingly collected data

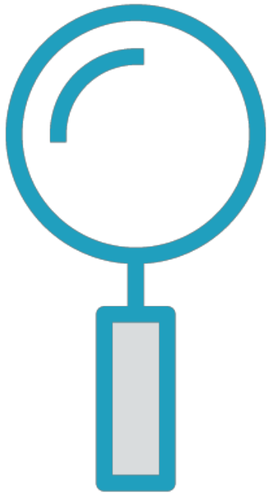
Thoughtful

Balanced, weighing pros and cons

Point of View

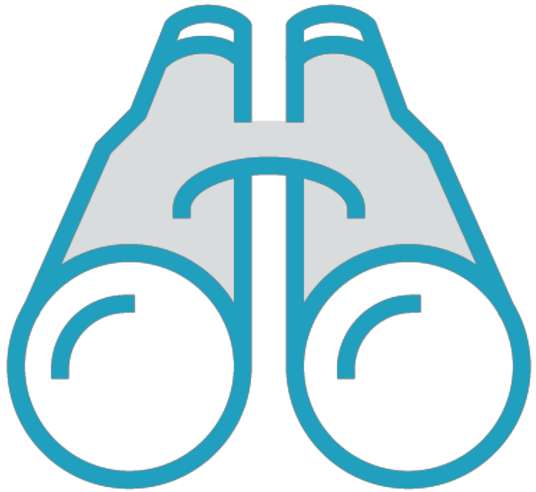
Prediction, recommendation, call to action

Two Sets of Statistical Tools



Descriptive Statistics

Identify important elements in a dataset



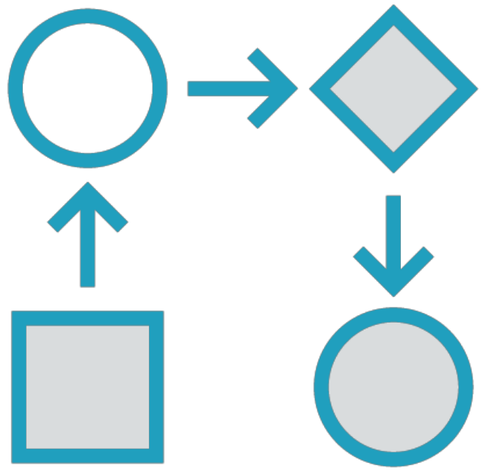
Inferential Statistics

Explain those elements via relationships with other elements

Two Hats of a Data Professional

Find the Dots

Identify important elements in a dataset



Connect the Dots

Explain those elements via relationships with other elements

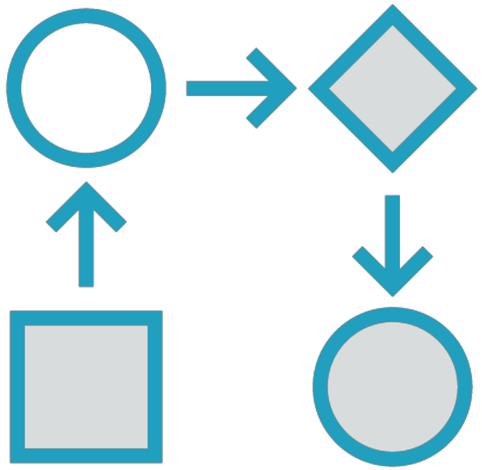
Finding the Dots

Data is more and more plentiful

However careful handling is needed

- Missing values
- Outliers
    - Genuine outliers
    - Erroneously measured points

Connecting the Dots



Spreadsheets

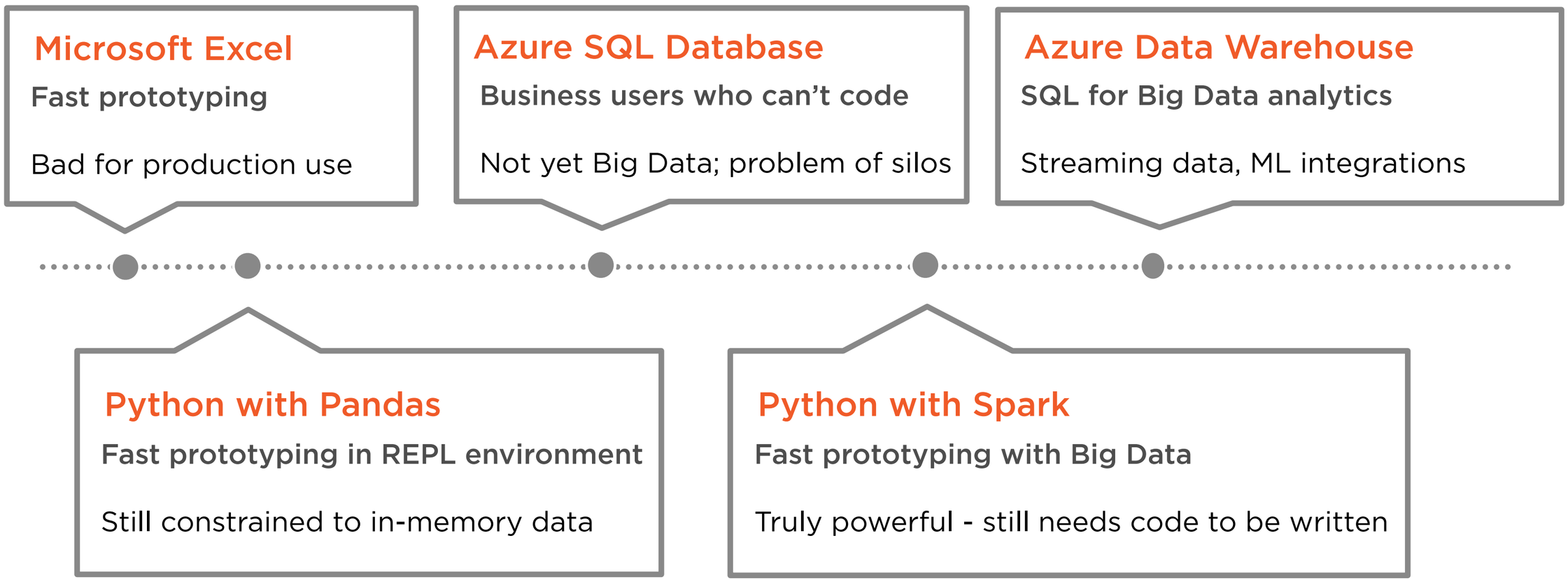
Programming languages

- In-memory processing
- Distributed processing

SQL

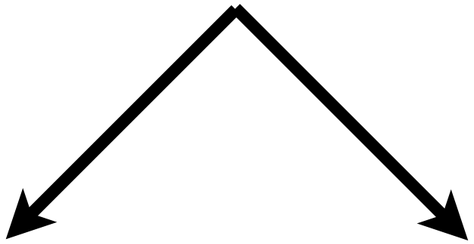
- Relational databases
- Data warehouses

Choices of Technology



Missing Data

Missing Data



Deletion

Imputation

Deletion a.k.a. Listwise Deletion

Delete an entire record (row) if a single value (column) is missing. Simple but can lead to bias.

Listwise Deletion

Most common method in practice

Can reduce sample size significantly

If values are not missing at random, can introduce significant bias

Imputation

Fill in missing column values, rather than deleting records with missing values.

Imputation

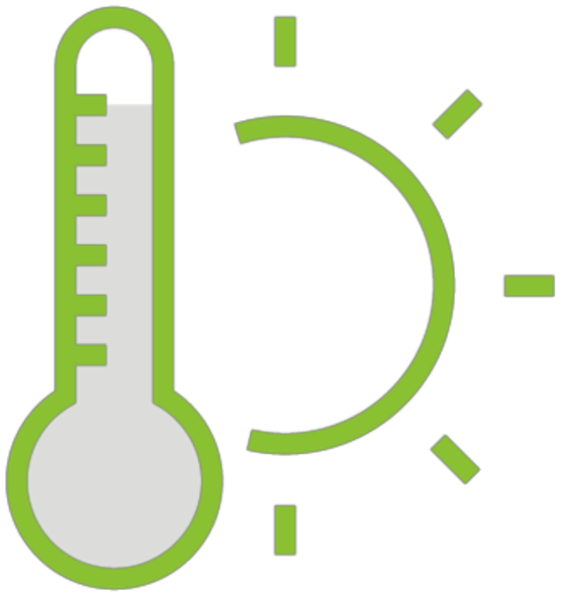
Methods range from very simple to very complex

Simplest method: Use column average

Can interpolate from nearby values

Can even build model to predict missing values

Hot-deck Imputation



Sort records based on any criteria

For each missing value, use immediately prior available value

“Last Observation Carried Forward”

For time series, equivalent to assuming no change since last measurement

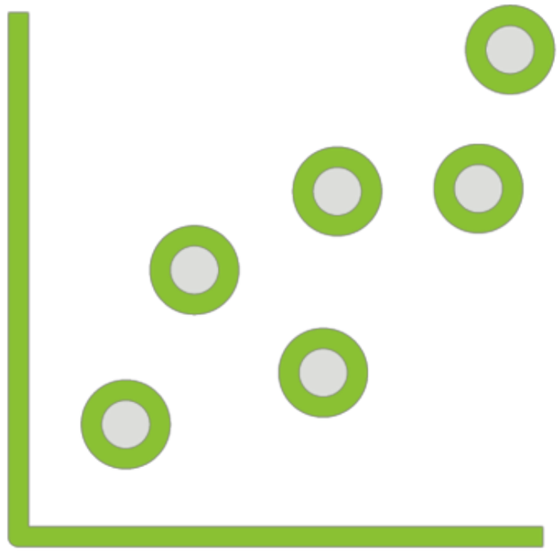
Mean Substitution

For each missing value, substitute mean of all available values

Has effect of weakening correlations between columns

Can be problematic when bivariate analysis required

Regression



Fit model to predict missing column based on other column values

Tends to strengthen correlations

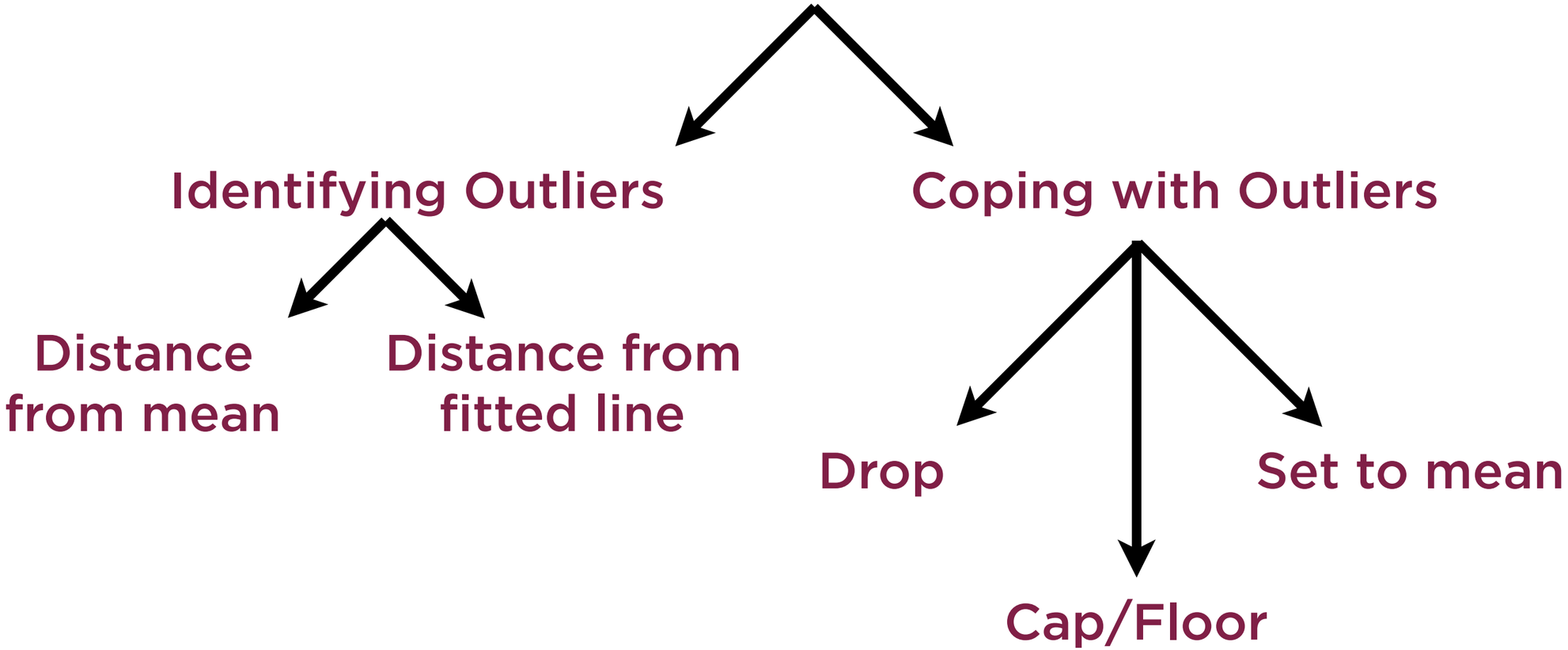
Regression and mean substitution have complementary strengths

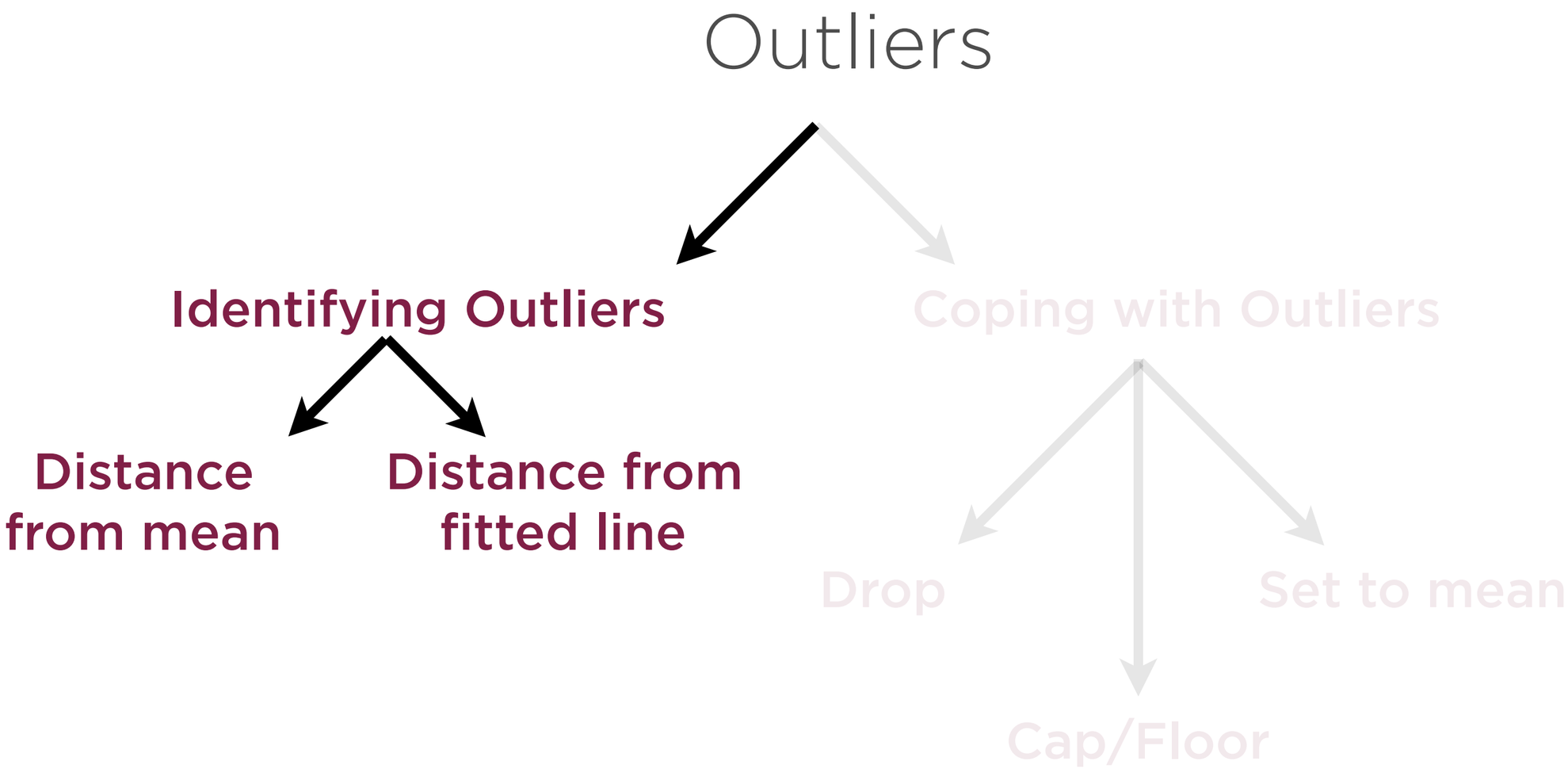
Outliers

Outlier

A data point that differs significantly from other data points in the same data set.

Outliers


Identifying Outliers
Coping with Outliers
Distance from mean
Distance from fitted line
Drop
Cap/Floor
Set to mean


Outliers
Identifying Outliers
Coping with Outliers
Distance from mean
Distance from fitted line
Drop
Set to mean
Cap/Floor

Identifying Outliers

Distance from mean

Distance from fitted line

Identifying Outliers

Distance from mean

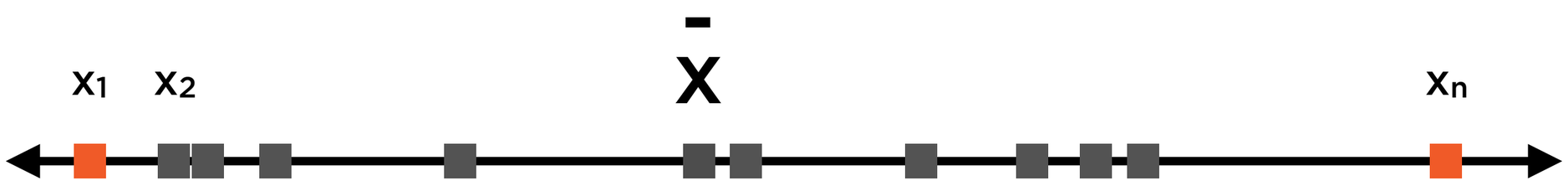
Distance from fitted line

Mean as Headline

The mean, or average, is the one number that best represents all of these data points

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Variation Is Important Too

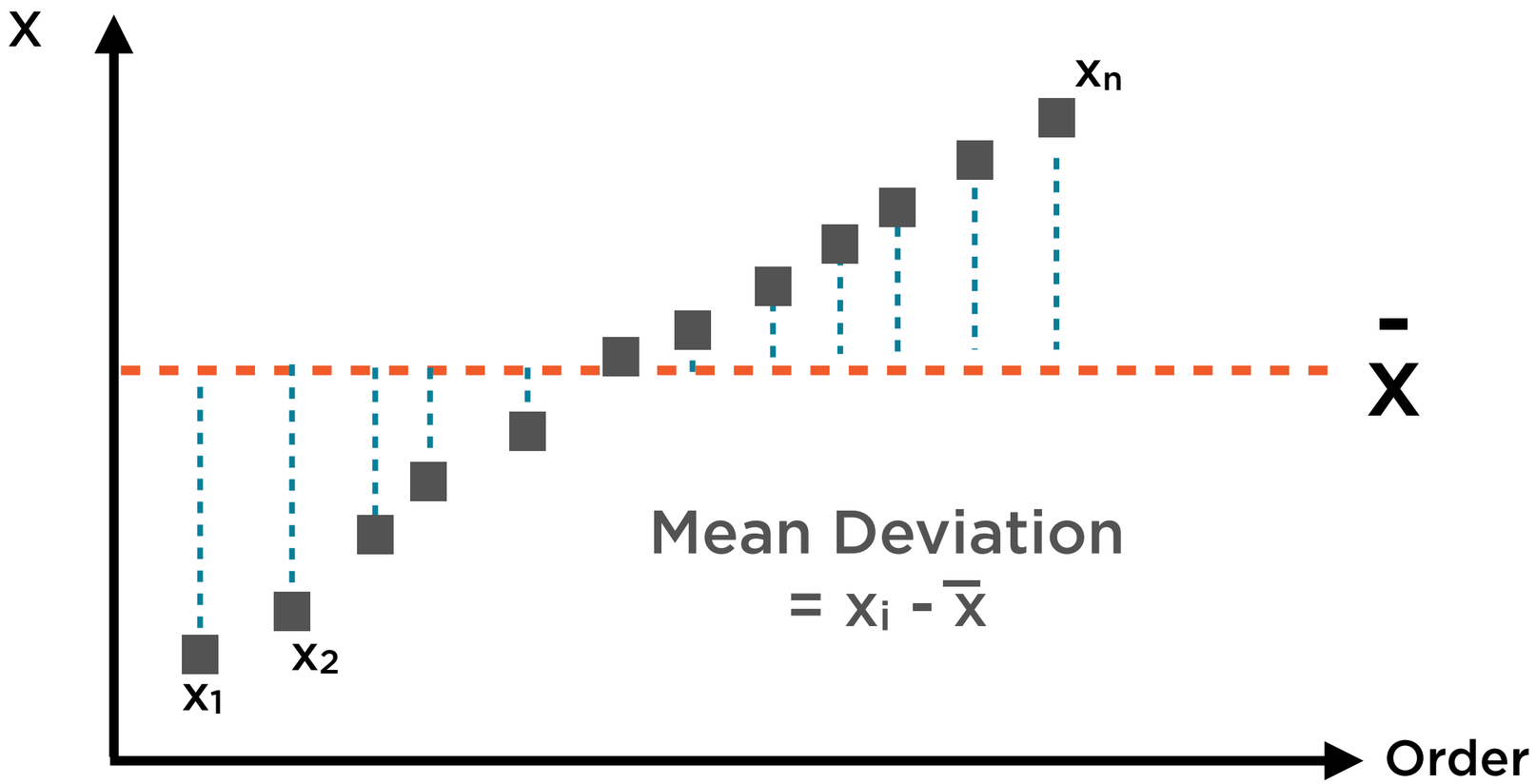


"Do the numbers jump around?"

$$\text{Range} = X_{max} - X_{min}$$

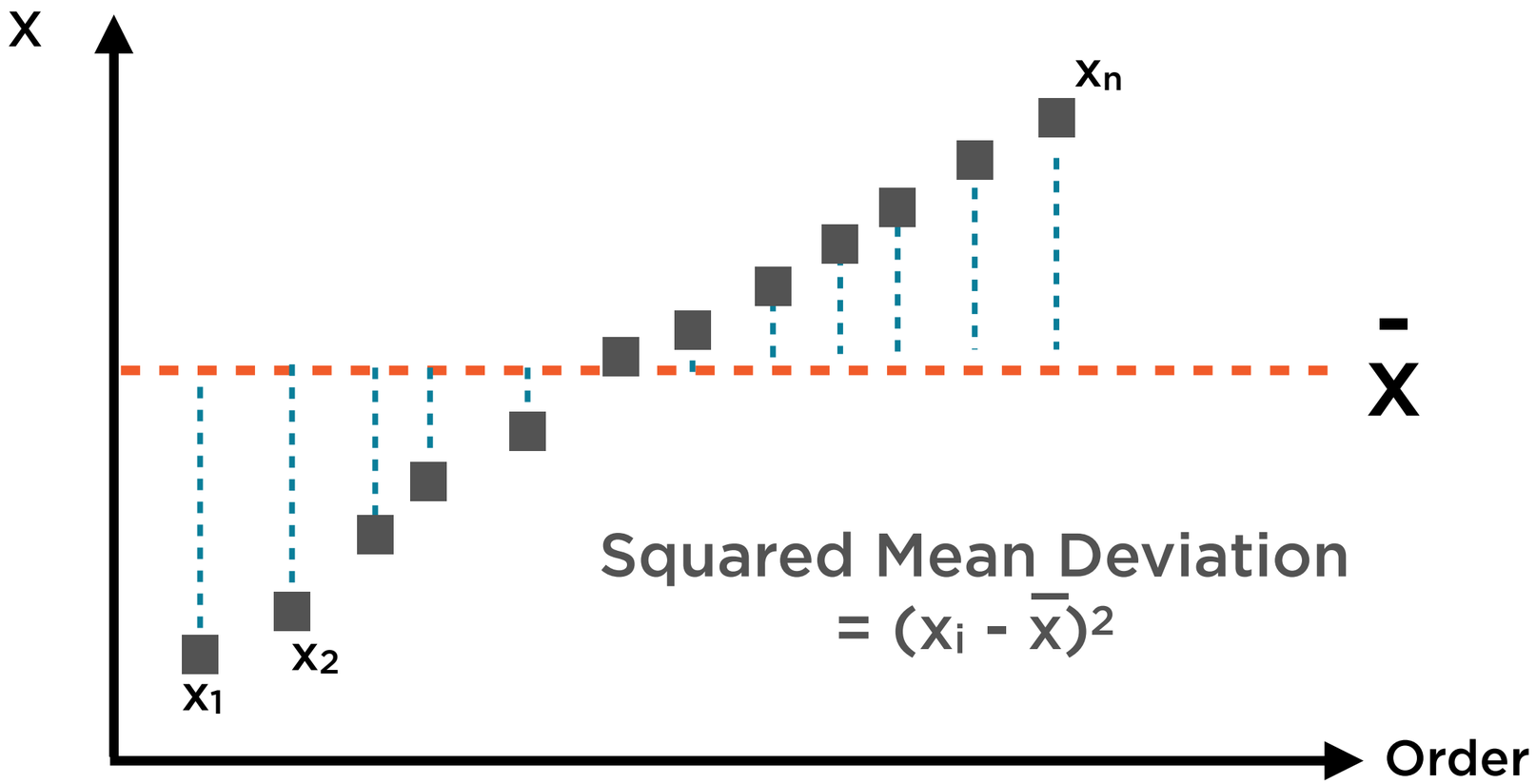
The range ignores the mean, and is swayed by outliers - that's where variance comes in

Variance as Asterisk



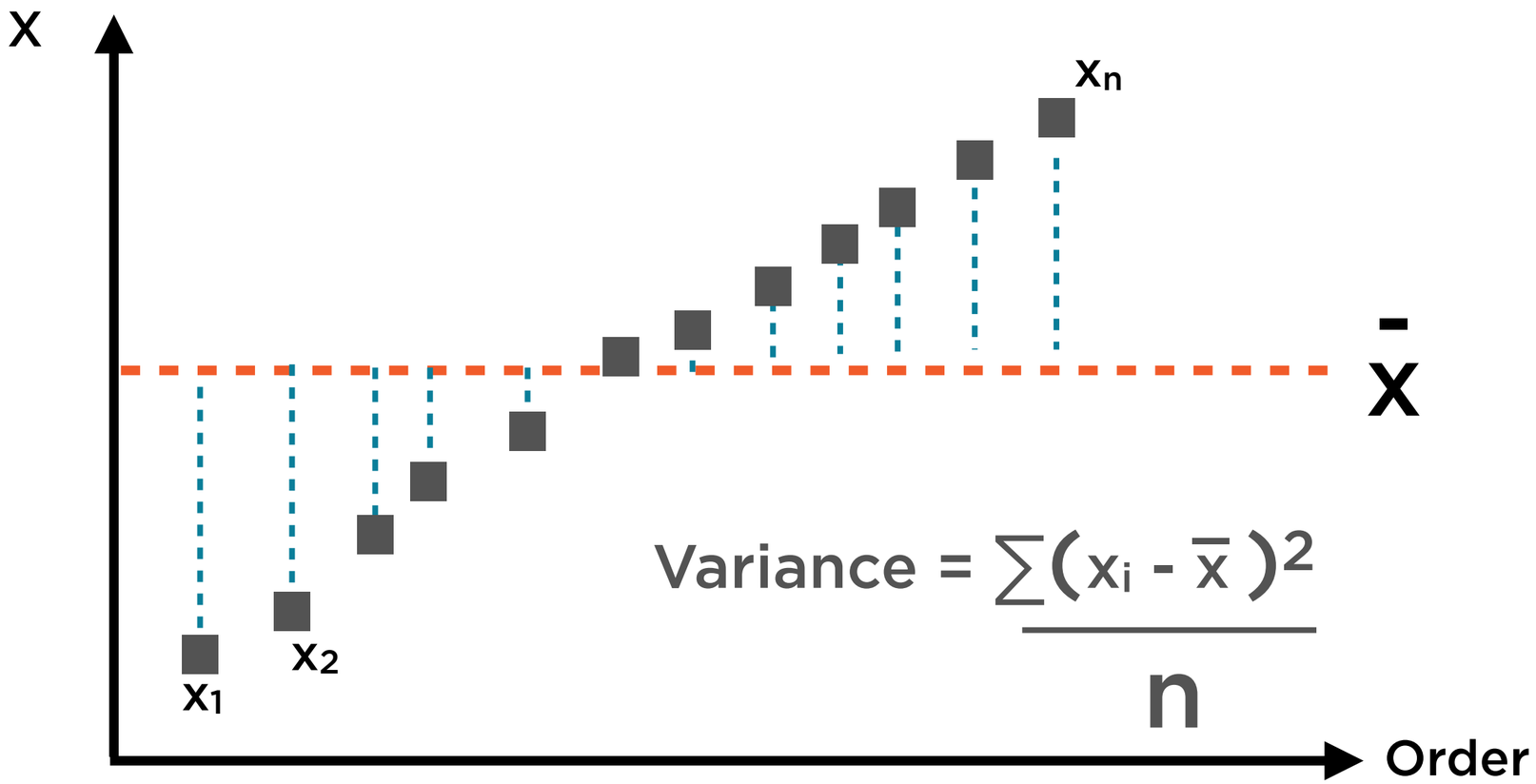
Variance is the second-most important number to summarize this set of data points

Variance as Asterisk



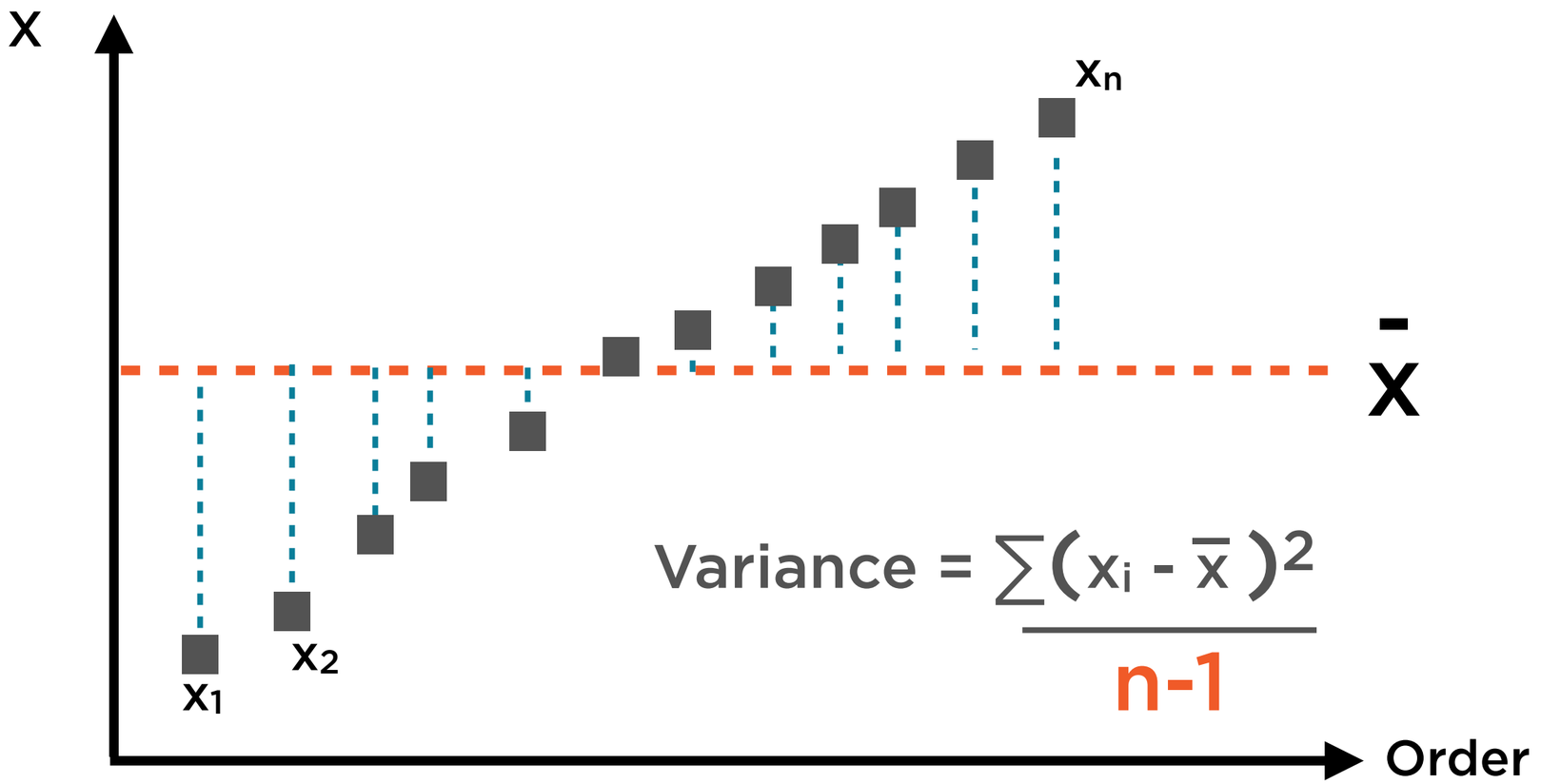
Variance is the second-most important number to summarize this set of data points

Variance as Asterisk



Variance is the second-most important number to summarize this set of data points

Variance as Asterisk



$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

We can improve our estimate of the variance by tweaking the denominator - this is called Bessel's Correction

Mean and Variance

Mean and variance succinctly summarize a set of numbers

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Variance and Standard Deviation

Standard deviation is the square root of variance

$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$\text{Std Dev} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Outliers

Points that lie more than 3 standard deviations from the mean are often considered outliers

Outliers

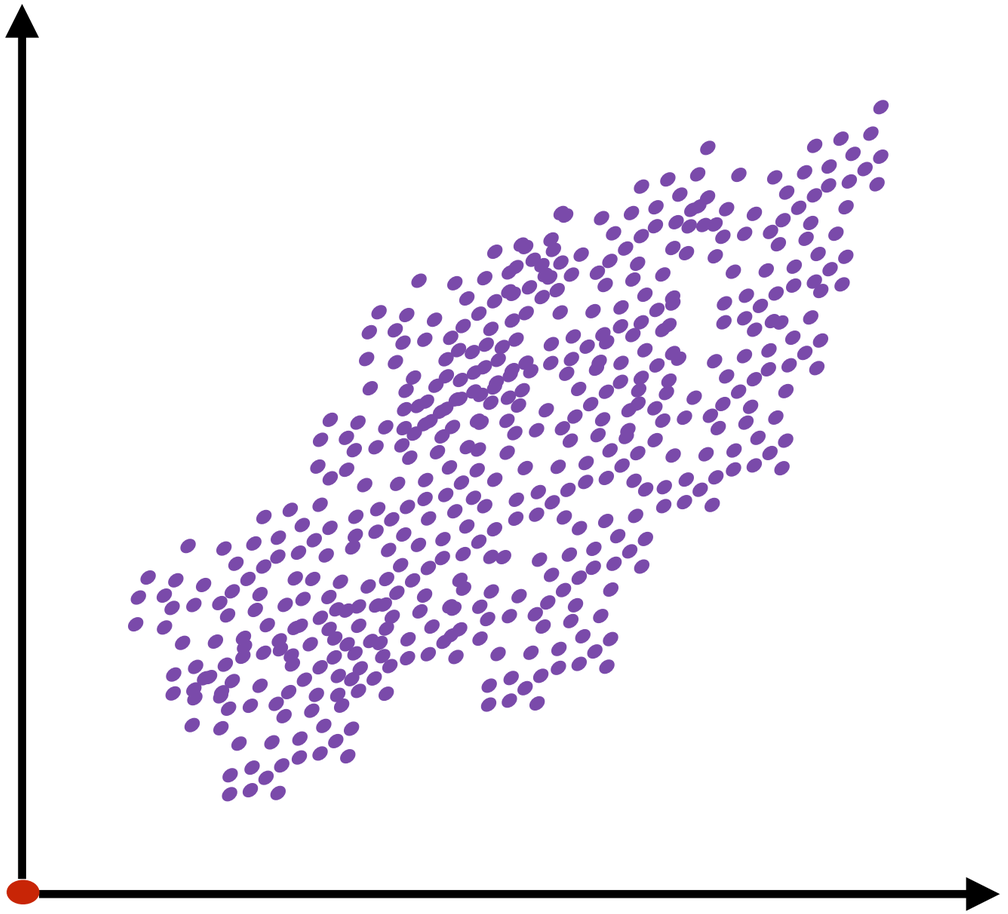
Points that lie more than 3 standard deviations from the mean are often considered outliers

Identifying Outliers

Distance from mean

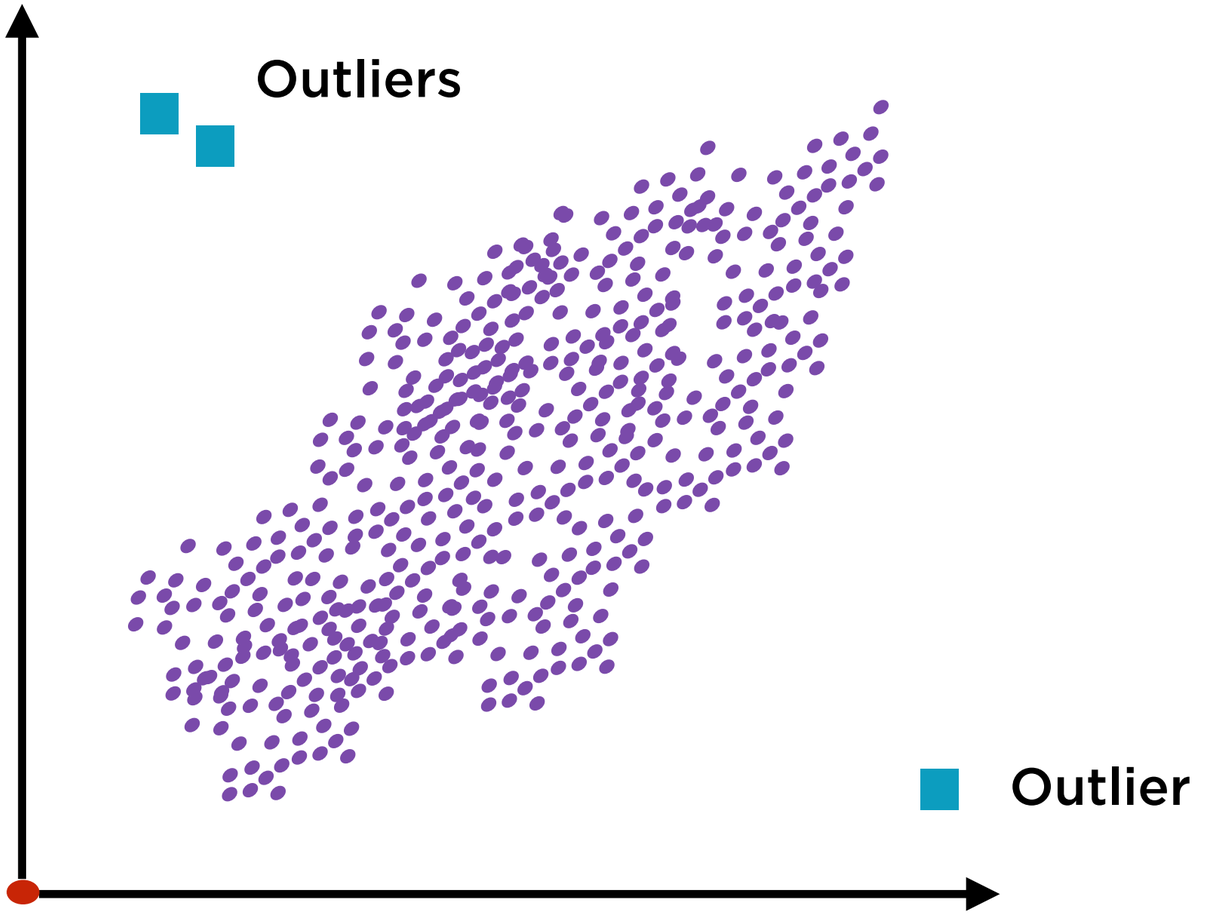
Distance from fitted line

Outliers



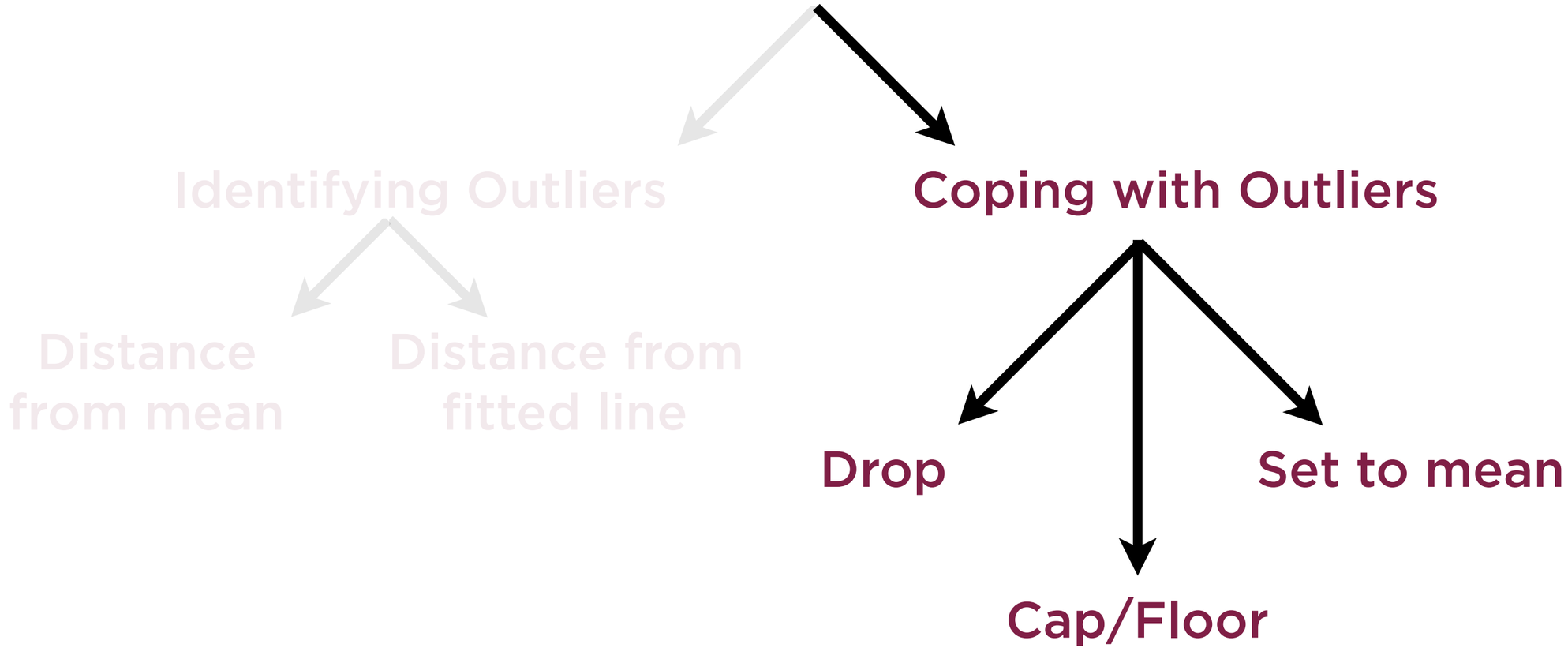
Outliers might also be data points that do not fit into the same relationship as the rest of the data

Outliers



Outliers might also be data points that do not fit into the same relationship as the rest of the data

Outliers


Identifying Outliers
Distance from mean
Distance from fitted line
Coping with Outliers
Drop
Cap/Floor
Set to mean

Coping with Outliers

Always start by scrutinizing outliers

If erroneous observation

- Drop if all attributes of that point are erroneous
- Set to mean if only one attribute is erroneous

Coping with Outliers

If genuine, legitimate outlier

- Leave as-is if model not distorted
- Cap/Floor if model is distorted
    - Need to first standardize data
    - Cap positive outliers to +3
    - Floor negative outliers to -3

Summary

- Identifying problems that hinder analytics
- Common technology tools to work with data
- Dealing with missing data
- Dealing with outliers and erroneous data