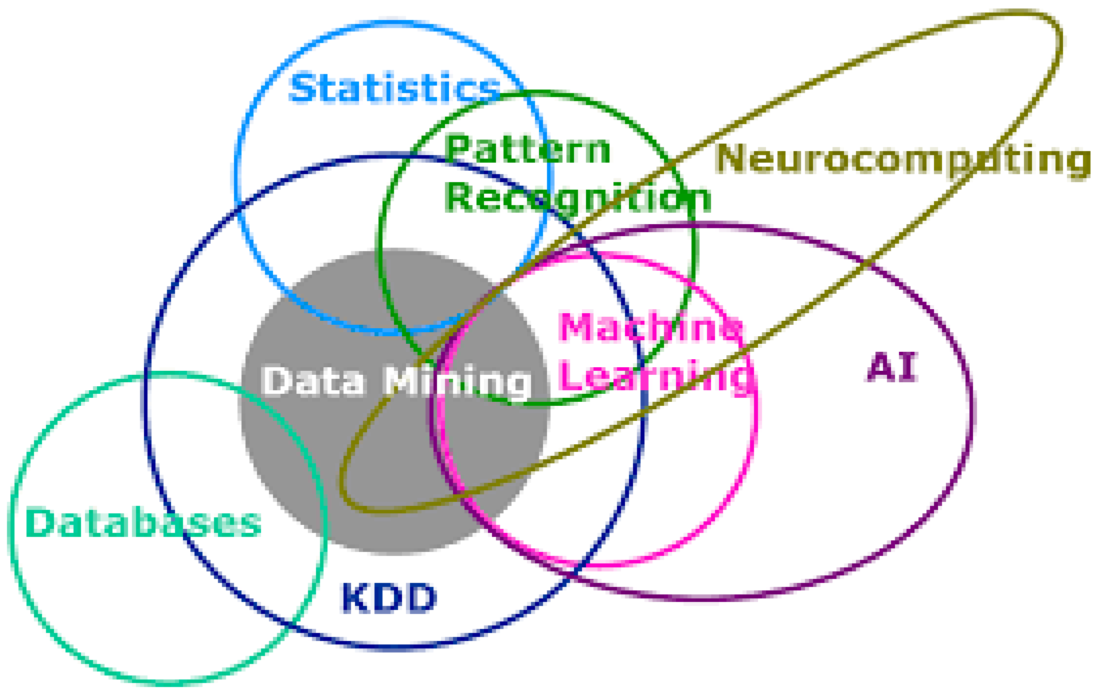
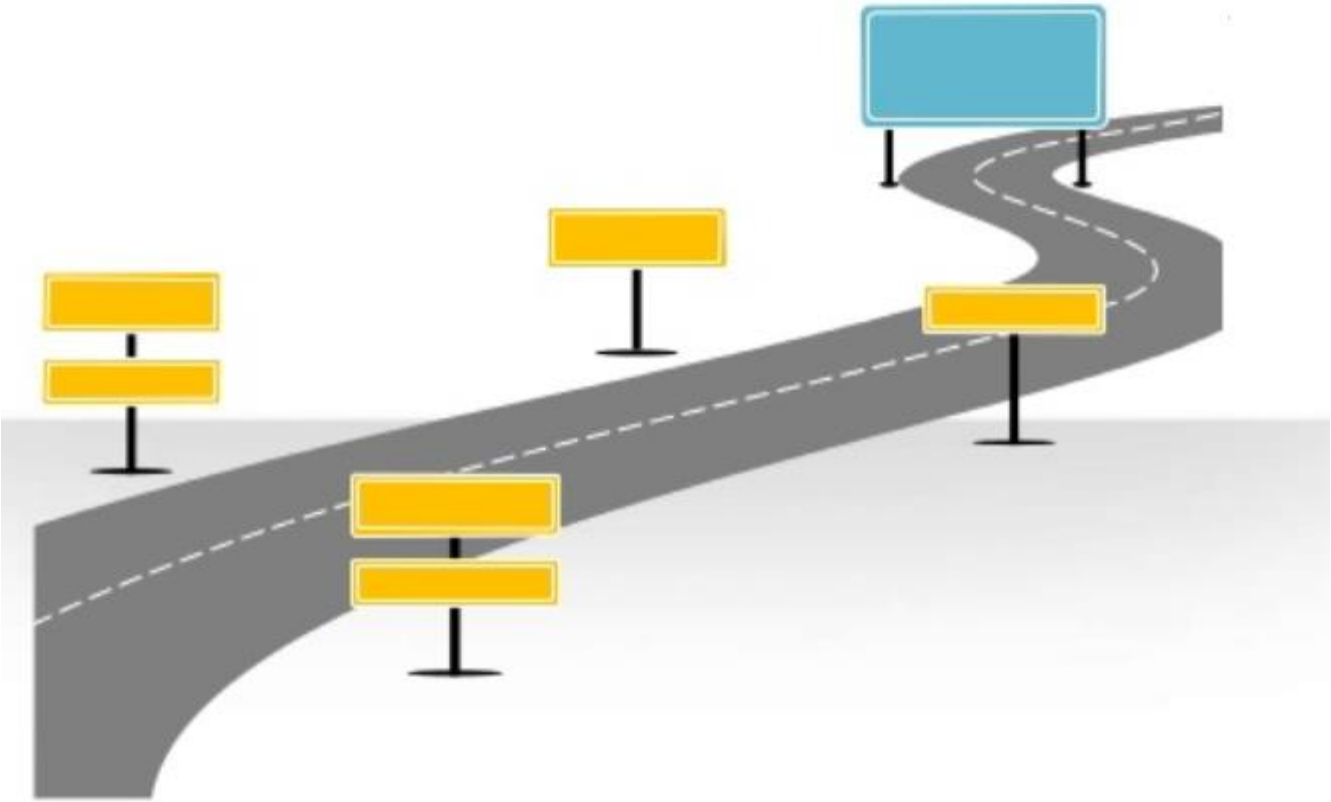
In the name of Allah the most Beneficial ever merciful

بسم الله الرحمن الرحيم



# Today's Agenda

Find Strong Connection between SE and AI



# *Artificial Intelligence (AI) in Software Engineering*

## ANOVA Table



*Department of Computer Science , Univeristy of Karachi (DCS-UBIT)*
*25th May 2021*

# Week 11 Analysis of Variance Table

# Motivation

We need ANOVA Test in Artificial Intelligence for Feature Selection

# 1 1995: Intuitive Statistics, Oxford Press, New York

| Measurement type | Continuous, parametric | Nominal/ordinal/ nonparametric | Dichotomous (two possible outcomes) | Survival (time to event) (*not clear*) |
|---|---|---|---|---|
| Describe one group | Mean, SD | Median, percentiles | Proportion | Kaplan–Meier survival curve, median survival |
| Compare one group to a hypothetical value | One sample *t* -test | Wilcoxon test | Chi-squared or binominal test | |
| Compare two unpaired groups | Unpaired *t* -test | Mann–Whitney | Fisher's or Chi-squared | Log rank or Mantel–Haenszel |
| Compare two paired groups | Paired *t* -test | Wilcoxon | McNamara's | Conditional proportional hazards regression |
| Compare three or more unmatched groups | One way ANOVA | Kruskal–Wallis | Chi-squared | Cox proportional hazards regression |
| Compare three or more matched groups | Repeated-measured ANOVA | Friedman | Cochrane Q | Conditional proportional hazards regression |
| Quantify relationship between two variables | Pearson correlation | Spearman correlation | Contingency coefficients | |
| Predict value from another variable | Linear (or nonlinear) regression | Nonparametric regression | Simple logistic regression | Cox proportional hazards regression |
| Predict values from several measured or binominal variables | Multiple linear (or nonlinear) regression | | Multiple logistic regression | Cox proportional hazards regression |

*Source*: This table is derived from Mikulski, H. (1995). *Intuitive Statistics*. Oxford Press. New York.

UNIVERSITY OF KARACHI

# ANOVA TABLE Format

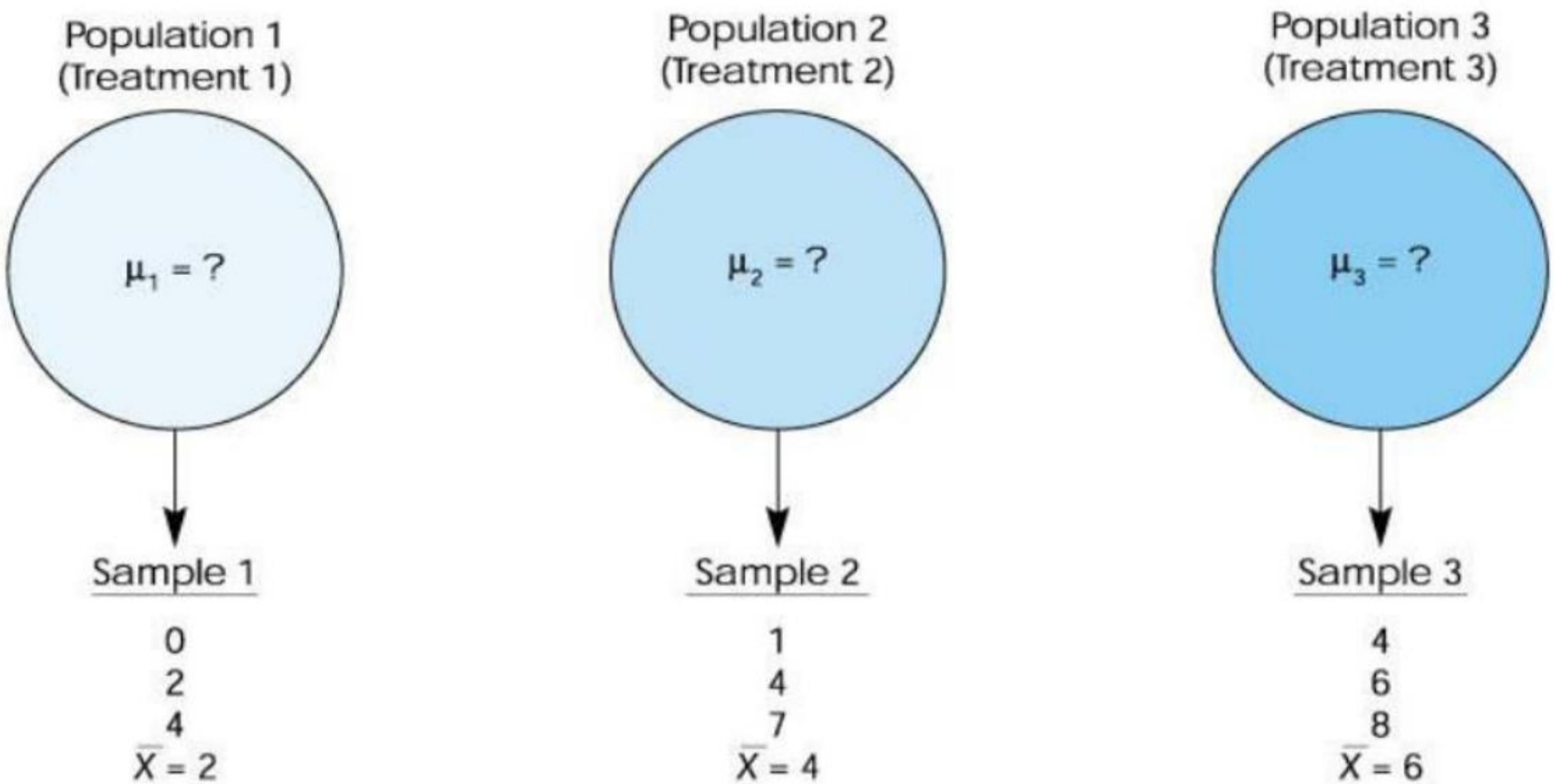
The ANOVA (analysis of variance) table splits the sum of squares into its components.

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | | | | |
| Residual | 6 | | | | |
| Total | 7 | | | | |

# Problem Design for ANOVA

| Bus | Motorbike | Car |
|---|---|---|
| Time | Time | Time |




UNIVERSITY OF KARACHI

# How to fill ANOVA TABLE?

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Squares (MS) | F |
|---|---|---|---|---|
| Within<br>Error/Residual SS<br>Beyond Control | $SSW = \sum_{j=1}^{k}\sum_{j=1}^{l}(X-\bar{X}_j)^2$ | $df_w = k-1$ | $MSW = \frac{SSW}{df_w}$ | $F = \frac{MSB}{MSW}$ |
| Between<br>Regression/Explained SS | $SSB = \sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2$ | $df_b = n-k$ | $MSB = \frac{SSB}{df_b}$ | |
| Total SS | $SST = \sum_{j=1}^{n}(\bar{X}_j - \bar{X})^2$ | $df_t = n-1$ | | |

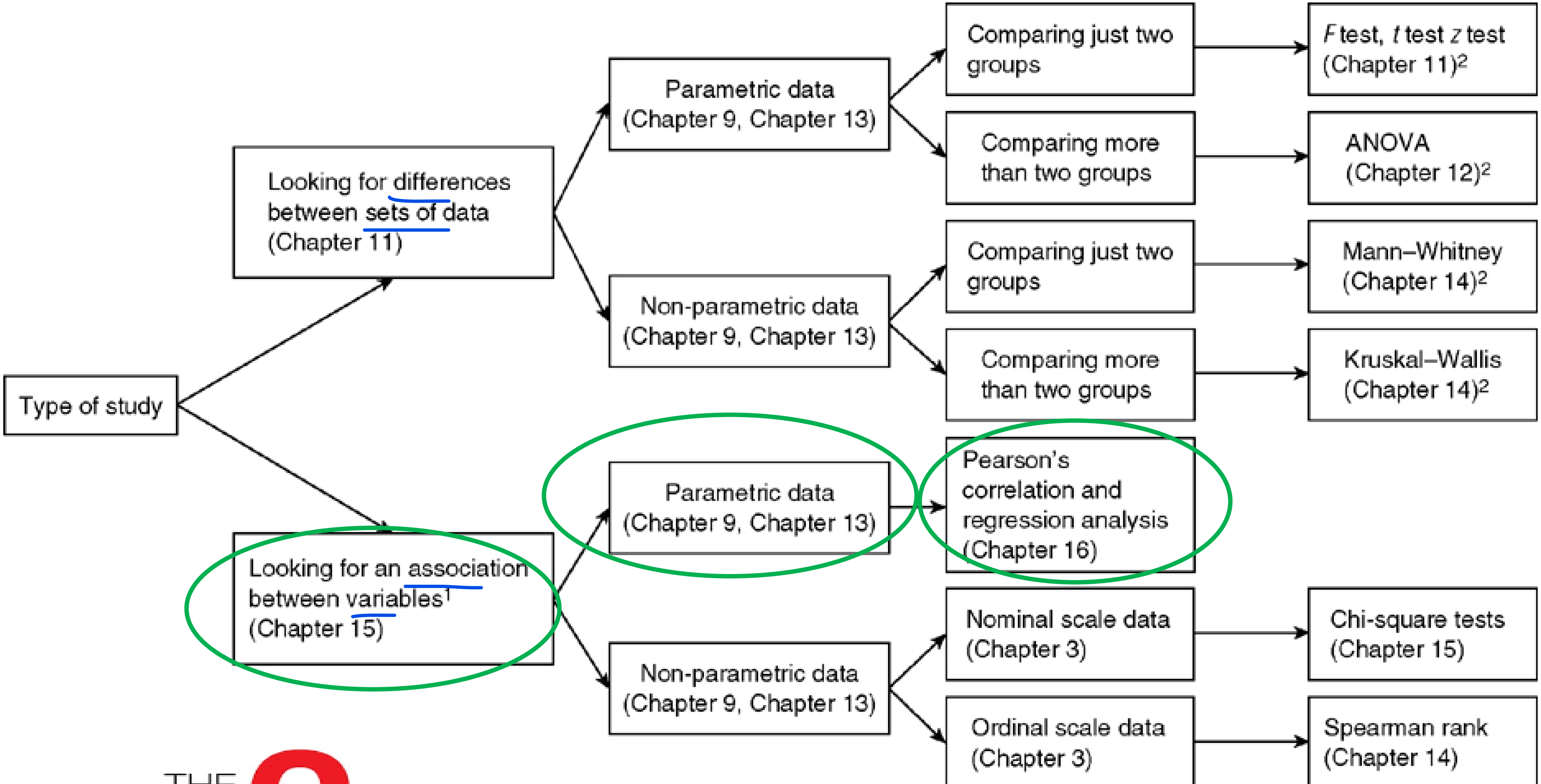
k: no. of groups/classes/subjects

n: no. of samples/observations/

df: degree of freedom

UNIVERSITY OF KARACHI

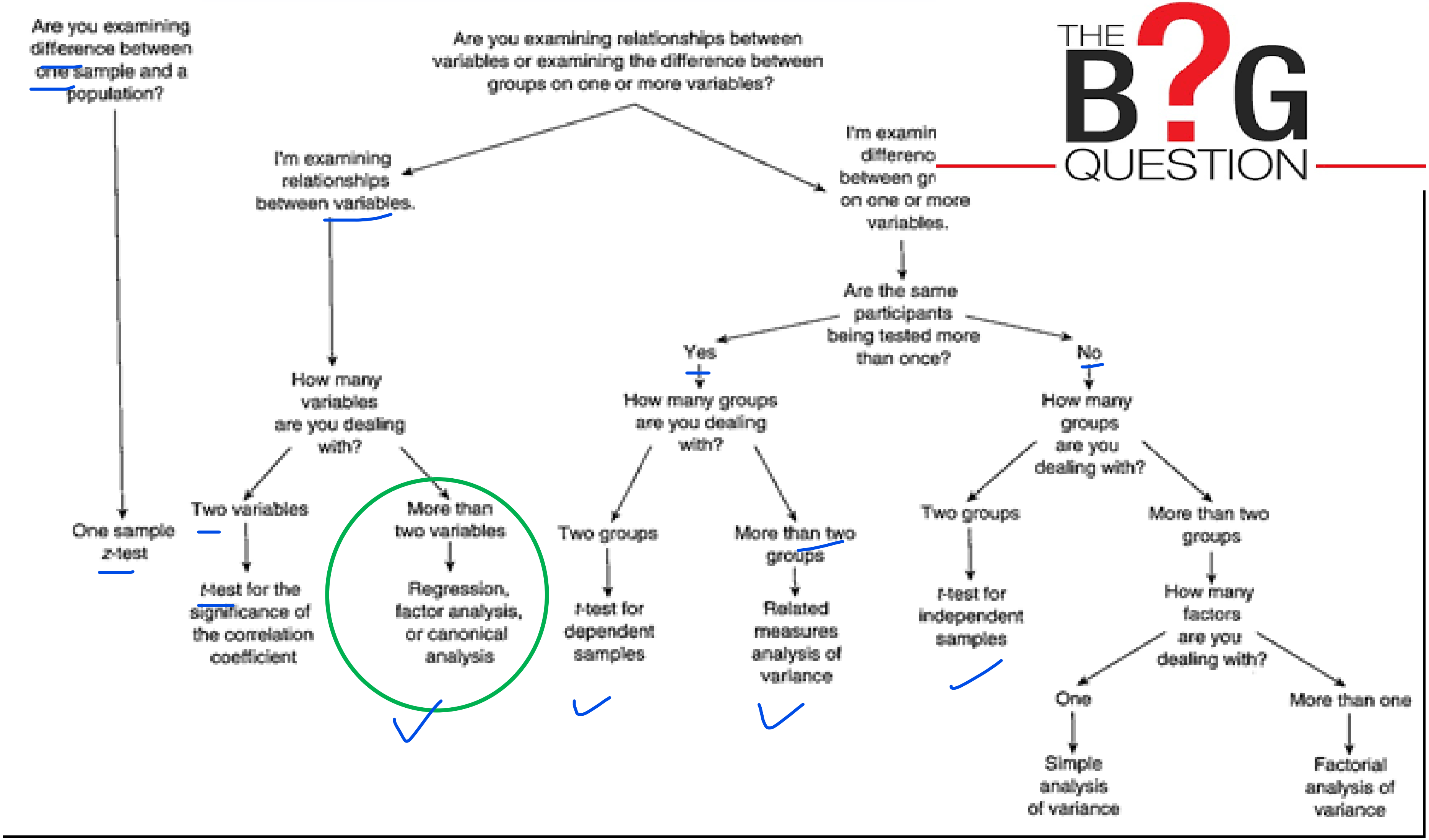
# Do we understand our type of Study + Data ?




THE B?G QUESTION

UNIVERSITY OF KARACHI

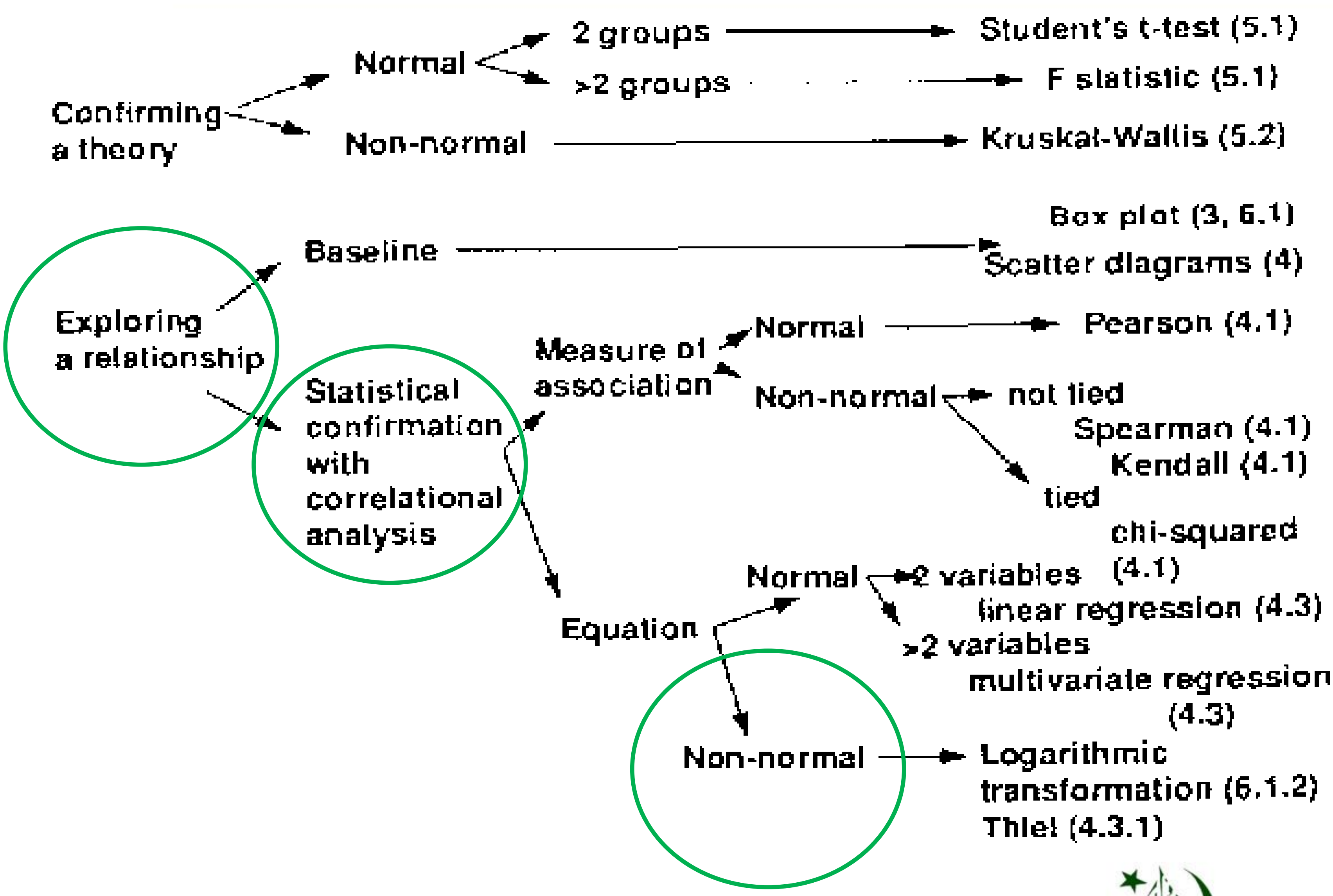
# Build understanding of Variable Types




THE B?G QUESTION

UNIVERSITY OF KARACHI

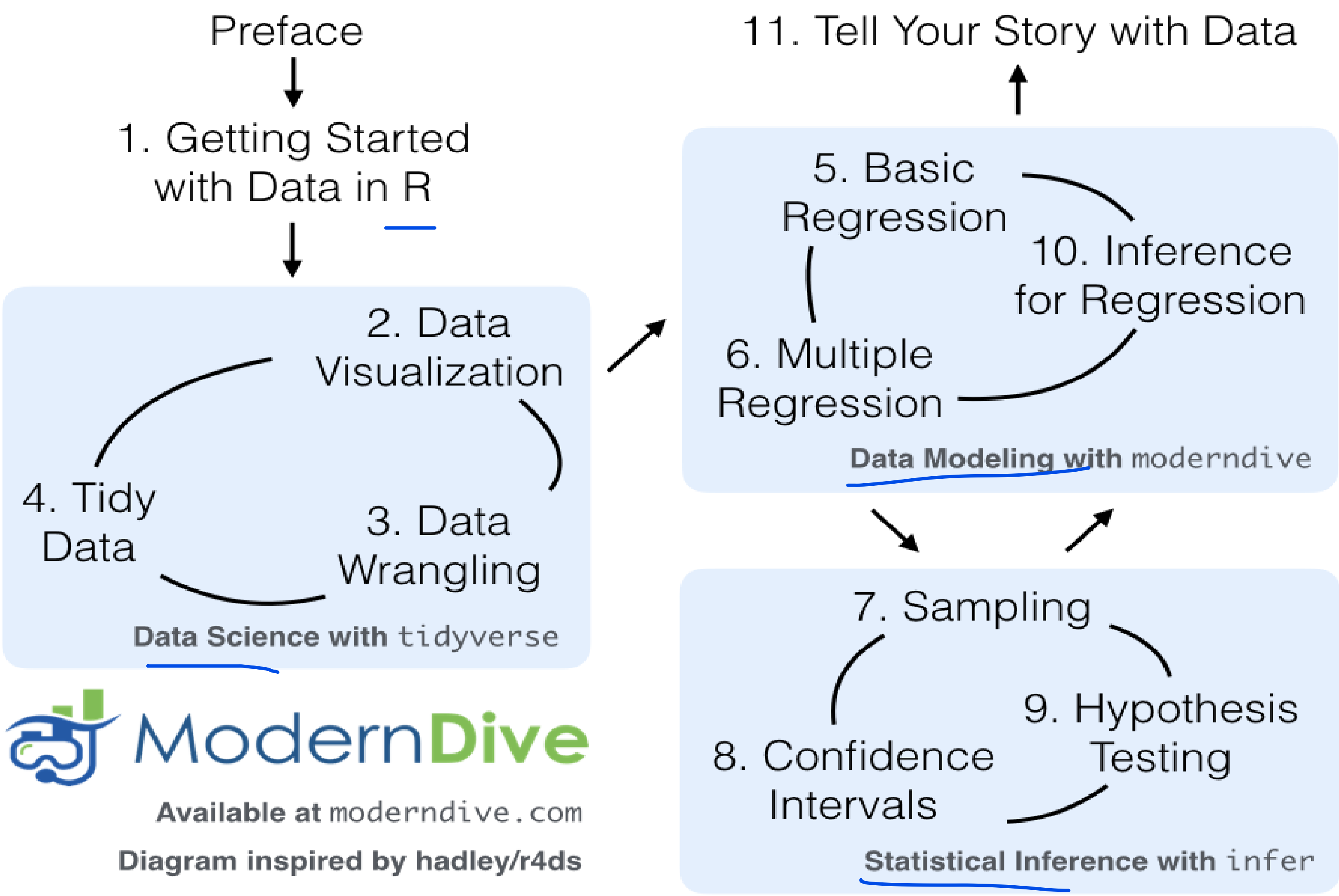
# Build understanding of Problem type




UNIVERSITY OF KARACHI

# Modern Artificial Intelligence

# Modern AI




ModernDive

Available at moderndive.com

Diagram inspired by hadley/r4ds

NIST Explainable AI Draft NISTIR8312.pdf
SSB 2021 May 18th at 8:01 PM

National Institute of Standards and Technology
U.S. Department of Commerce

# Four Principles of Explainable Artificial Intelligence

https://www.nist.gov/topics/artificial-intelligence/ai-foundational-research-explainability.

UNIVERSITY OF KARACHI

NIST Explainable AI Draft NISTIR8312.pdf
SSB 2021 May 18th at 8:01 PM

# Table of Contents

What we will do !..........

Software Defect Prediction

JM1 Dataset

# 28 October 20th , 25th February 2020

Open Access Article

## Software Defect Prediction Using Heterogeneous Ensemble Classification Based on Segmented Patterns

by Hamad Alsawalqah [1], Neveen Hijazi [1], Mohammed Eshtay [2], Hossam Faris [1,*], Ahmed Al Radaideh [3], Ibrahim Aljarah [1] and Yazan Alshamaileh [1]

1 King Abdullah II School for Information Technology, The University of Jordan, Amman 11942, Jordan
2 Department of Computer Science, Faculty of Information Technology, Zarqa University, Zarq 13132, Jordan
3 College of Humanities and Sciences, University of Science and Technology of Fujairah, Fujairah 2202, UAE
* Author to whom correspondence should be addressed.

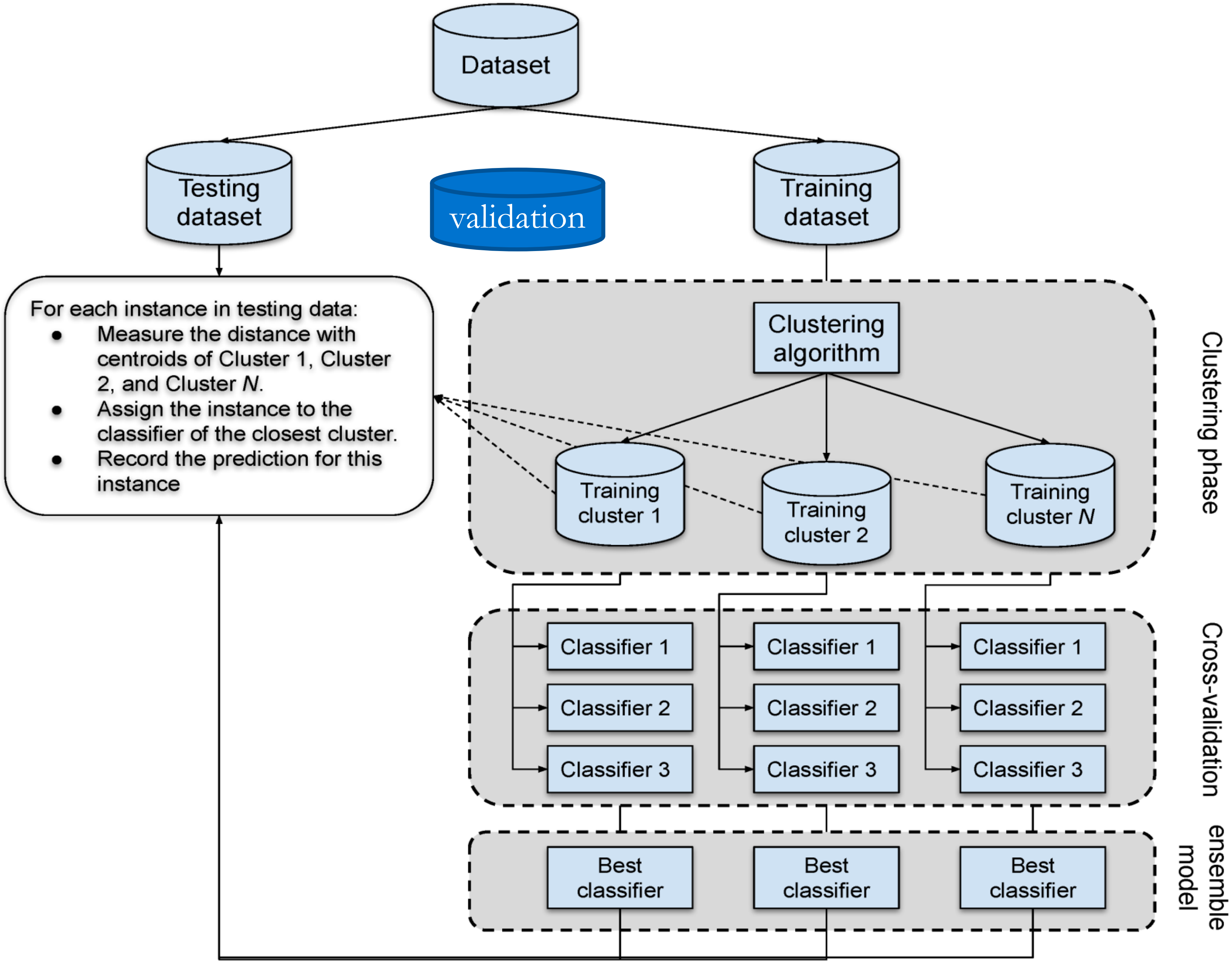
*Appl. Sci.* **2020**, *10*(5), 1745; **https://doi.org/10.3390/app10051745**

**Received: 28 October 2019 / Revised: 20 February 2020 / Accepted: 25 February 2020 / Published: 3 March 2020**

(This article belongs to the Section **Computing and Artificial Intelligence**)

UNIVERSITY OF KARACHI

# Proposed methodology for software defect prediction




RSITY OF
ACHI

3

# JM1 is listed as highly imbalanced datasets in experimental study

SN Computer Science (2020) 1:108
https://doi.org/10.1007/s42979-020-0119-4

Check for updates

ORIGINAL RESEARCH

SN

## Evaluation of Sampling-Based Ensembles of Classifiers on Imbalanced Data for Software Defect Prediction Problems

Thanh Tung Khuat[1] · My Hanh Le[2]

Received: 21 September 2019 / Accepted: 11 March 2020


UNIVERSITY OF KARACHI

# 4 Software Prediction, Imbalance data, Ensemble Learning

**Abstract**
Defect prediction in software projects plays a crucial role to reduce quality-based risk and increase the capability of detecting faulty program modules. Hence, classification approaches to anticipate software defect proneness based on static code characteristics have become a hot topic with a great deal of attention in recent years. While several novel studies show that the use of a single classifier causes the performance bottleneck, ensembles of classifiers might effectively enhance classification performance compared to a single classifier. However, the class imbalance property of software defect data severely hinders the classification efficiency of ensemble learning. To cope with this problem, resampling methods are usually combined into ensemble models. This paper empirically assesses the importance of sampling with regard to ensembles of various classifiers on imbalanced data in software defect prediction problems. Extensive experiments with the combination of seven different kinds of classification algorithms, three sampling methods, and two balanced data learning schemata were conducted over ten datasets. Empirical results indicated the positive effects of combining sampling techniques and the ensemble learning model on the performance of defect prediction regarding datasets with imbalanced class distributions.

**Keywords** Software defect prediction · Random undersampling · Random oversampling · SMOTE · Data balancing · Ensemble learning · Imbalanced data

UNIVERSITY OF KARACHI

# Imbalance Data → Biased Classifier

- ✓ Software quality datasets are usually imbalanced, in which most defects of the software system may be found in a small ratio of modules. Hence, the number of faulty records in a software dataset is much smaller than that of non-defective examples

- ✓ Nearest neighbor (KNN), support vector machines (SVM), neural networks (NN), Bayesian network (BN), and decision tree (DT)

UNIVERSITY OF KARACHI

# Summary of 10 highly imbalance data

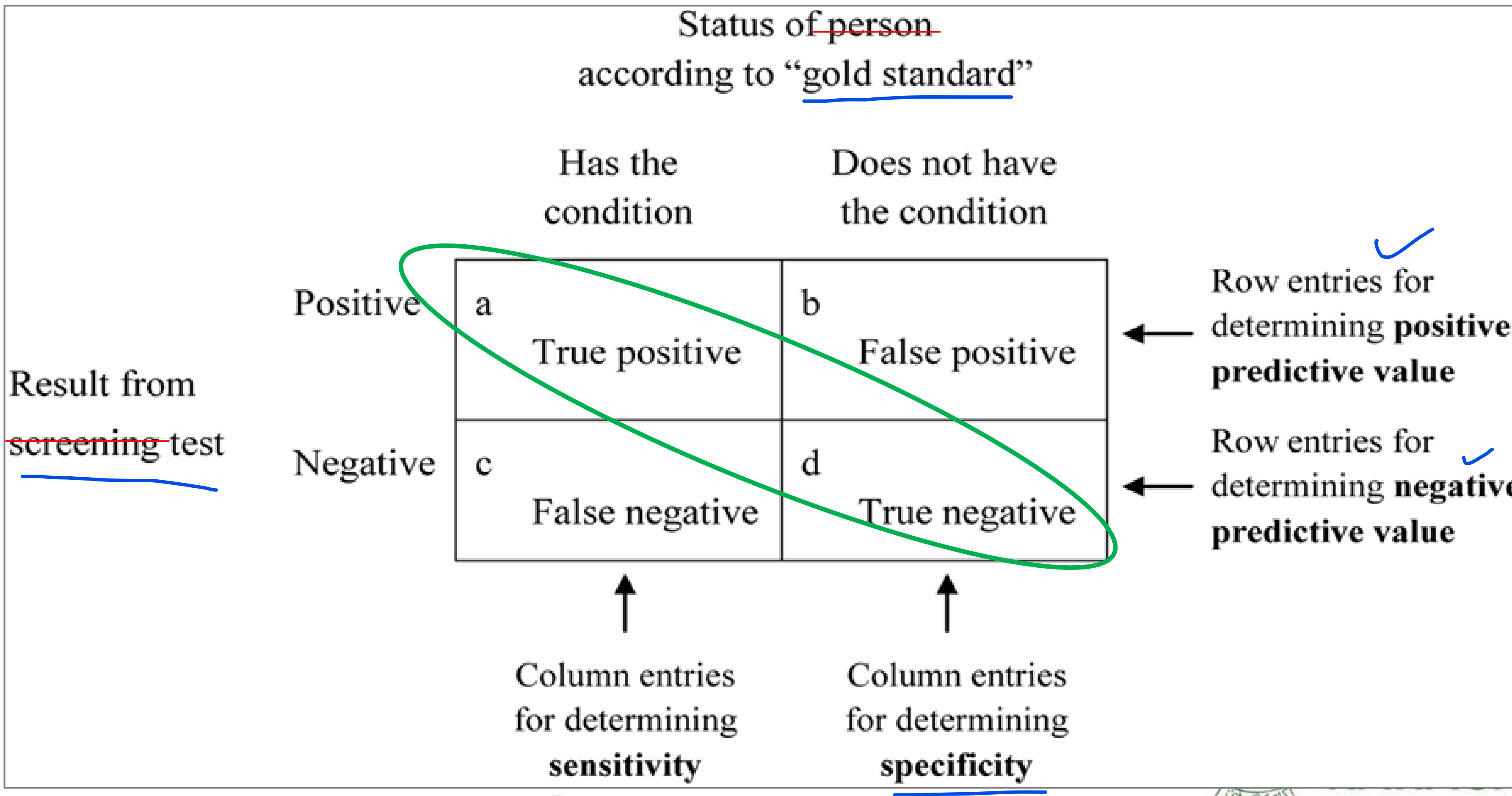
SN Computer Science (2020) 1:108 Page 9 of 16 108

**Table 1** Summary of the 10 highly imbalanced datasets in experimental study

| Dataset | Language | # Attr. | # Ins. | # Defect | # Non-defect | % Defect |
|---|---|---|---|---|---|---|
| JM1 | C | 21 | 7782 | 1162 | 6110 | 14.93 |
| KC3 | Java | 39 | 194 | 36 | 158 | 18.56 |
| PC1 | C | 37 | 705 | 61 | 644 | 8.65 |
| ant 1.7 | Java | 20 | 745 | 166 | 579 | 22.28 |
| camel 1.6 | Java | 20 | 965 | 188 | 777 | 19.48 |
| ivy 2.0 | Java | 20 | 352 | 40 | 312 | 11.36 |
| poi 2.0 | Java | 20 | 314 | 37 | 277 | 11.78 |
| tomcat | Java | 20 | 858 | 77 | 781 | 8.97 |
| xalan 2.4 | Java | 20 | 723 | 110 | 613 | 15.21 |
| synapse 1.2 | Java | 20 | 256 | 86 | 170 | 33.59 |

# Imbalance Data → Biased Classifier→ High False Negative Rate

These classifiers bias toward the dominant class and tend to disregard the minority class, which results in high false-negative rates



# COVID Example

Actually, Not COVID Patient but Algorithm/classifier declared the patient with COVID

| | | Actual Values 1 | |
|---|---|---|---|
| | | Defective (Y) | Non-defective (N) |
| Predicted Values | 2 Defective (Y) | TP | FP 3 |
| | Non-defective (N) | FN | TN |

UNIVERSITY OF KARACHI

# Weekly Assignment

# Practice Data partitioning and imbalance

- ✓ Find relevant Code/Dataset/ ????
- ✓ Run or work on excel sheet
- ✓ Submit code/Analysis with feedback

☐ I don't understand a single word about partitioning and imbalance

☐ If not, What is the main hindrance to start work?

☐ If Yes, share your experience/practice/work

UNIVERSITY OF KARACHI

## Practice Confusion Matrix

- ✓ Find a toy example on confusion matrix.
- ✓ Solve/Redo it in your own handwriting
- ✓ Tick following Check box and submit

☐ I don’t understand a single word about confusion matrix

☐ I can solve toy problems like this

☐ I understand and explain/present it to my friends.

2

## Practice Data partitioning and imbalance

- ✓ Find relevant Code/Dataset/ ????
- ✓ Run or work on excel sheet
- ✓ Submit code/Analysis with feedback

☐ I don't understand a single word about partitioning and imbalance

☐ If not, What is the main hindrance to start work?

☐ If Yes, share your experience/practice/wrok

UNIVERSITY OF KARACHI

# Practice a Toy Example of ANOVA Table

Use ctrl+click to access link

[One-way ANOVA - Test Procedure, Merits and Demerits, Example Solved Problems | Analysis of Variance | Statistics (brainkart.com)]

UNIVERSITY OF KARACHI