Introduction to Data Science

Dr. P. ANANDKUMAR

Founder and Director

ROOT - IT LEARNING CENTRE, TRICHY

www.therootlearning.com

Agenda

- Data Science and Decision Making
- Asking Great Questions
- Data Analysis
- Data Analytics
- Data Science - Job Profiles
- Types of Analytics
- Importance of Probability and Data Visualization
- Evolution of ML and AI Systems

"Data is the **new oil**. It's valuable, but if unrefined, it cannot be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity. **So, data must be broken down and analyzed for it to have value**."

- Clive Robert Humbly

Data Science and Decision Making

Experience informs intuition…

The Rise of Scientific Approach…

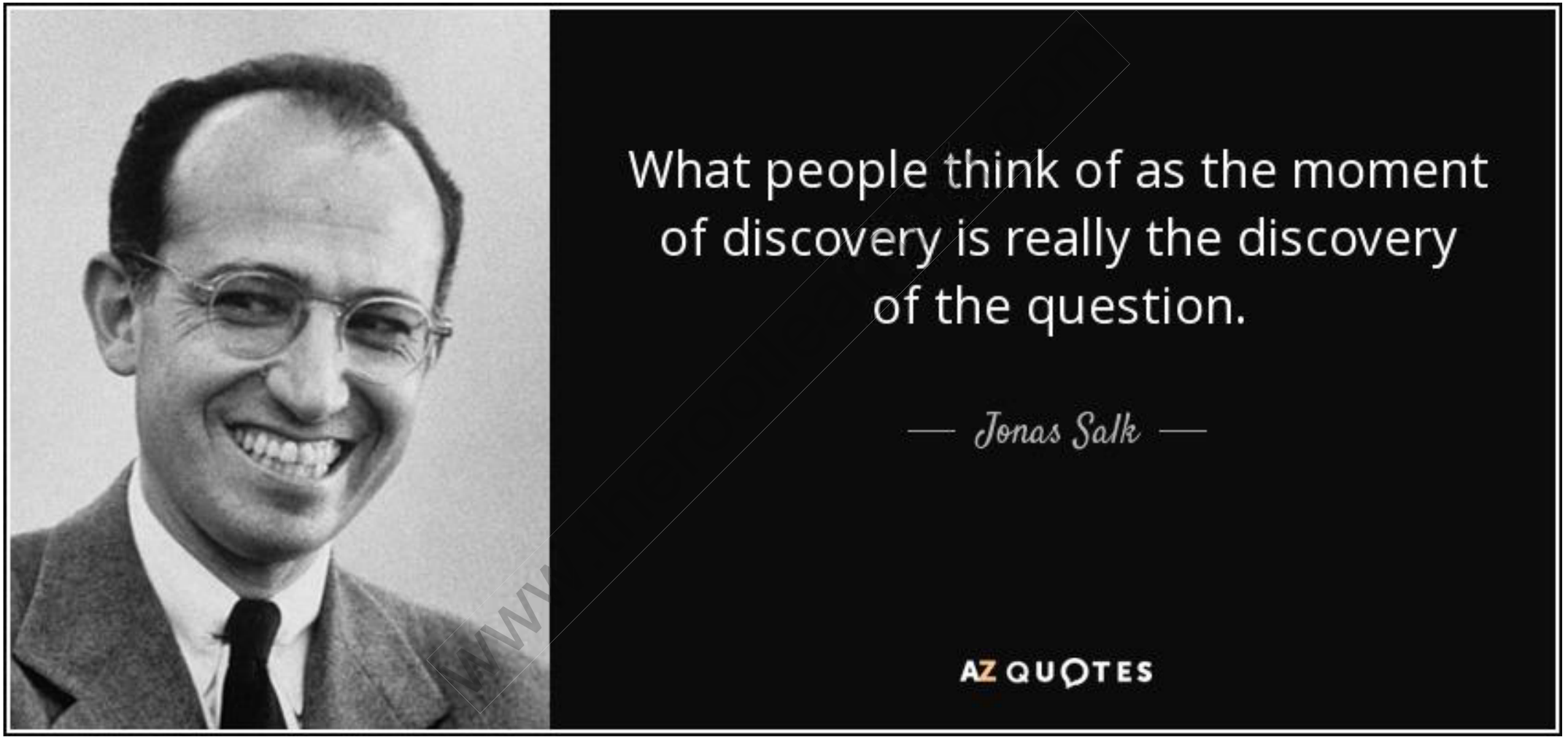
Turning from Qualitative to Quantitative Investigations…

Data Science Transformed Our World…

Fundamental Transformations…

- Alchemy into Chemistry
- Natural Philosophy into Physics and Biology
- Folk Remedies into Medicine

Asking Great Questions


What people think of as the moment of discovery is really the discovery of the question.
— Jonas Salk —
AZ QUOTES

It’s not about being hostile or disapproving

It’s about finding the **critical questions…**

A good question will challenge your thinking

It cannot be easily dismissed or ignored…

Bottled Mango Juice vs Fresh Mango Juice?

Which one will you choose?

Data Science Everywhere

Data Science: Why Now?

- The development of mathematical/computational tools and techniques – Open Source Communities
- The access to extensive and affordable computing power – Cloud and Cheap Computing Devices
- The availability of data – Open Data Networks

Data Analysis

A combination of applied mathematics and computer science

The Beginning of Data Analysis – WW2

- Armed forces adopted it first, for optimizing both men and machines
- P.M.S. Blackett, Nobel Prize Winner for Study on Cosmic Rays was a pioneer in this field
- Reduced the Anti-Aircraft Ammunition required to shoot one German Plane from 20000 rounds to 4000 rounds

Data analysis is a set of tools, existing and ever developing…

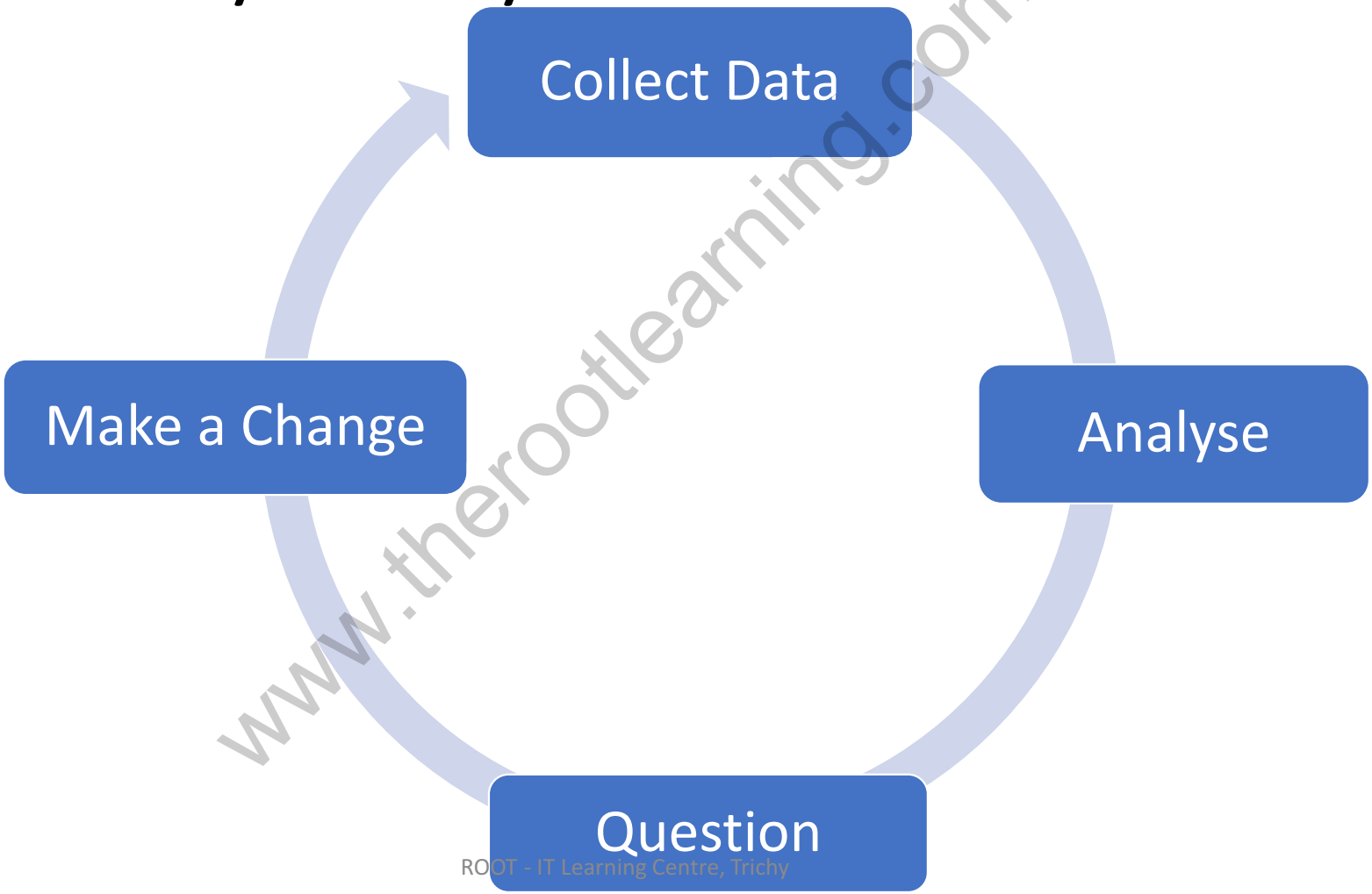
Data analysis is also a mind-set

It’s a way of improving our ability to ask questions and an expectation that data can **make possible** new answers…

Armor for Fighter Jets?

- Looking at jets that returned with damage, people suggested and went for armouring the damaged portions
- Blackett and his team had a different thought
- Armor the areas that were not affected or not damaged
- Why?
- They made it back… Right?

Data Analysis Cycle



Data analysis enables us to predict with **better accuracy** and **better probability**, many aspects of the future…

Case Study: Linear Regression

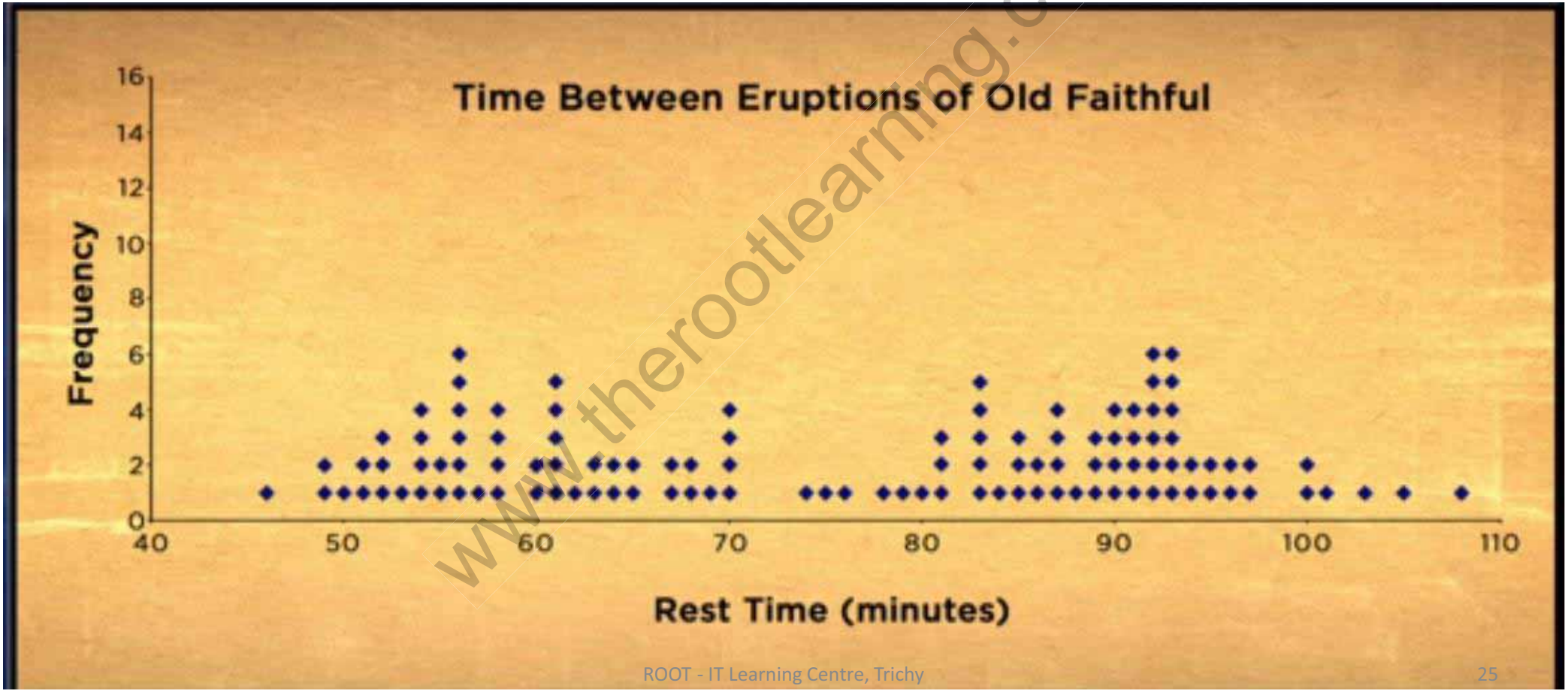
Let’s start with an example…

Yellow Stone National Park, US

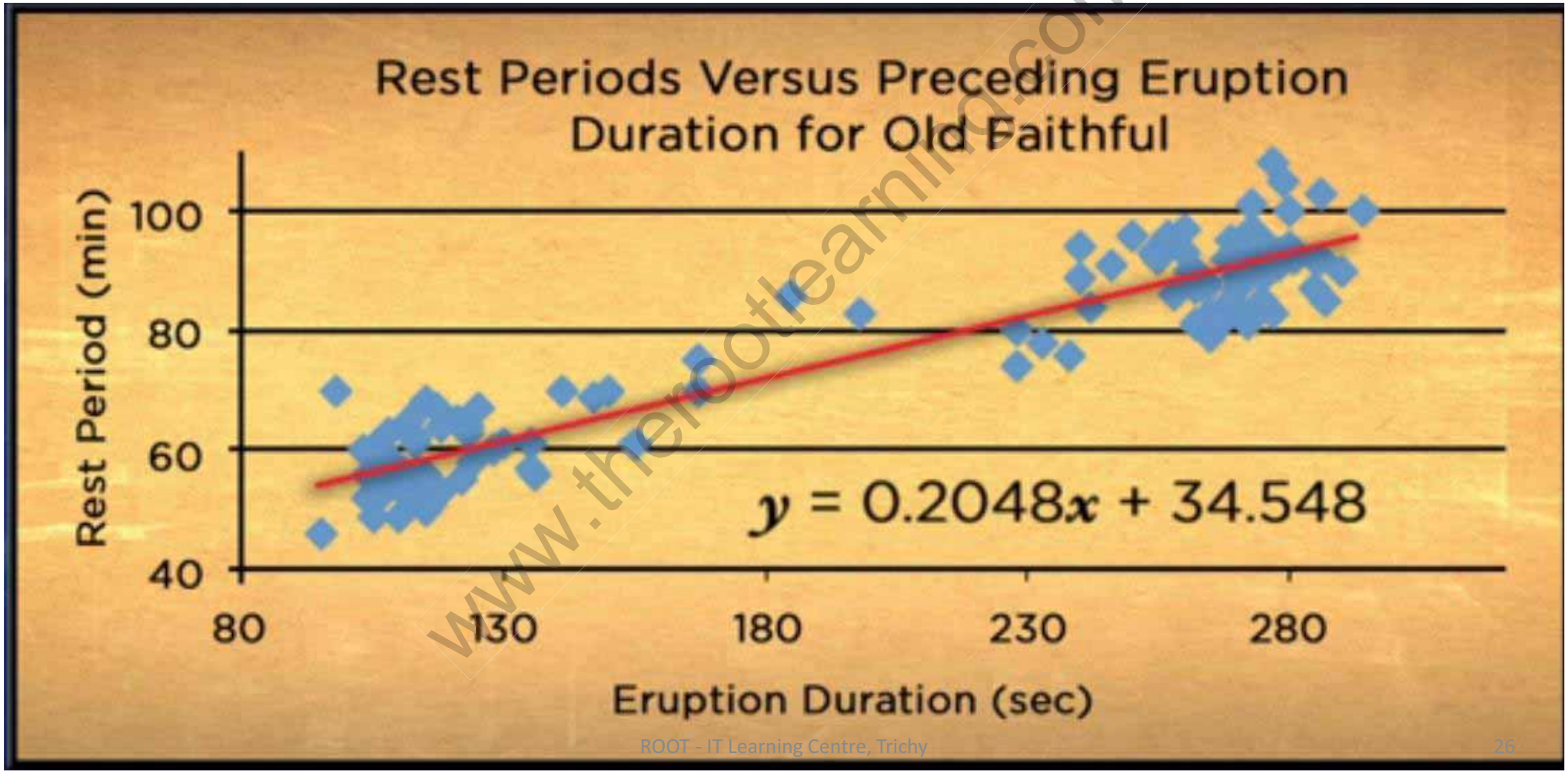
- Beneath the park is an active, super volcano
- Contains half of world's geothermal features
- Contains more than half of world's geysers
- Famous of these is **"Old Faithful"**
- It's the biggest, regular geyser in the park

Geysers – Old Faithful

When will it erupt? See the Data (112 instances)

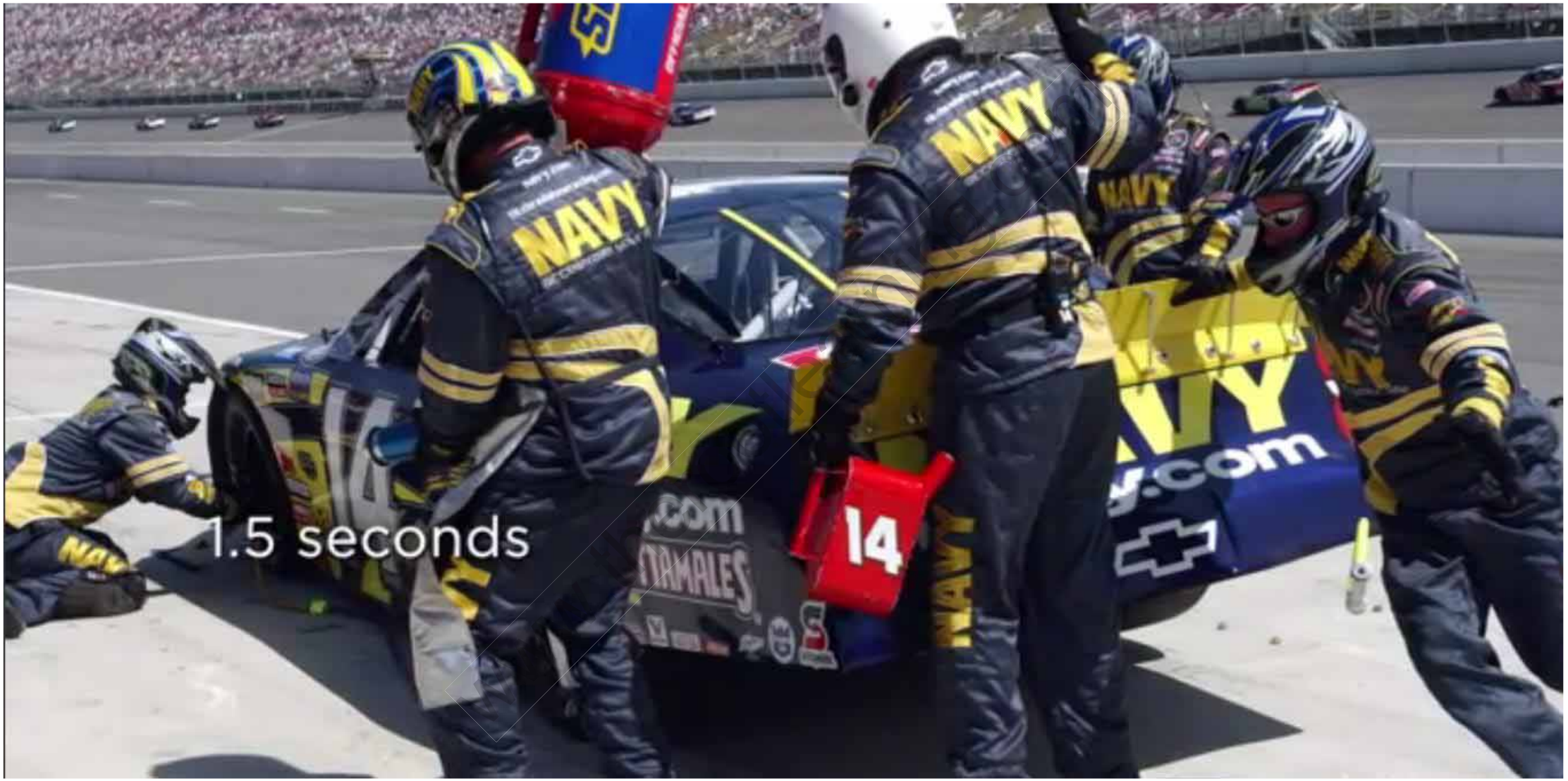


Can we predict it? Better?



To apply quantitative methods to help people, find ways **to do what they do better…**

Case Study: F1 Wheel Replacement


1.5 seconds
NAVY
14

Bolts are Tight Enough? 200 KMPH

Data Analytics

A Structured and Rigorous Approach Towards Examining Data

Data Analytics is the Process of

- Collecting Data
- Inspecting Data
- Cleaning Data
- Transforming Data
- Modelling Data and
- Using Data to Inform Decision Making

Data Analyst

- A person whose job is to examine information in order to find something out, or to help with making decisions (Cambridge Dictionary)
- Data Analysts work with data in all shapes, forms and fashions
- They make data more meaningful than just lines on a spreadsheet
- They create data stories to convey information in a visual form

Business Analytics

Continuous and iterative exploration of past business performance to gain insights, to make data driven decisions…

Data Science – Job Profiles

Business Analytics

Business Intelligence

Data Engineering

Data Science

Compare and Understand

Process	Data Engineering	Business Intelligence	Business Analytics	Data Science
Integrate Data Sources				
Build Data Pipelines				
Process and Transform Data				
Store Data				
Dashboards/Reports				
Exploratory Analytics				
Statistical Modeling				
Machine Learning				
Business Recommendation				
Business Action				

Data Engineering

Process	Data Engineering	Business Intelligence	Business Analytics	Data Science
Integrate Data Sources	X			
Build Data Pipelines	X			
Process and Transform Data	X			
Store Data	X			
Dashboards/Reports				
Exploratory Analytics				
Statistical Modeling				
Machine Learning				
Business Recommendation				
Business Action				

Business Intelligence

Process	Data Engineering	Business Intelligence	Business Analytics	Data Science
Integrate Data Sources	X			
Build Data Pipelines	X			
Process and Transform Data	X			
Store Data	X			
Dashboards/Reports		X		
Exploratory Analytics		X		
Statistical Modeling				
Machine Learning				
Business Recommendation		X		
Business Action				

Business Analytics

Process	Data Engineering	Business Intelligence	Business Analytics	Data Science
Integrate Data Sources	X			
Build Data Pipelines	X			
Process and Transform Data	X			
Store Data	X			
Dashboards/Reports		X	X	
Exploratory Analytics		X	X	
Statistical Modeling			X	
Machine Learning			X	
Business Recommendation		X	X	
Business Action			X	

Data Science

Process	Data Engineering	Business Intelligence	Business Analytics	Data Science
Integrate Data Sources	X			X
Build Data Pipelines	X			X
Process and Transform Data	X			X
Store Data	X			X
Dashboards/Reports		X	X	X
Exploratory Analytics		X	X	X
Statistical Modeling			X	X
Machine Learning			X	X
Business Recommendation		X	X	X
Business Action			X	X

Data Engineering: Toolkit

1. Linux/Unix
2. BASH Scripting
3. Python Scripting
4. Programming and Problem Solving
5. Data Structures and Algorithms
6. SQL - Basic Queries, Advanced Queries for Data Science, Stored Procedures, Triggers
7. Data Modelling, Fundamentals of Data Warehousing, ETL, RDBMS, Normalization, De-normalization
8. NoSQL Databases, Types and Choice of NoSQL Databases
9. Hadoop - Distributed Systems, HDFS, Map Reduce, Hive and Pig
10. Data Pre-Processing
11. Establishing Data Pipelines
12. Machine Learning Fundamentals
13. PySpark and SparkQL
14. Web Development Fundamentals for Ingesting Data from APIs, Web Services and JSON Documents
15. Cloud Platforms and Tools (Basic Knowledge)
16. Data Visualization using Excel, Tableau, D3.js

Types of Analytics

Types of Analytics – Categories

Category	Purpose
Descriptive Analytics	Tell me what happened, and why. Tell me what is happening right now, and why.
Predictive Analytics	Tell me what is likely to happen, and why.
Discovery Analytics	Tell me something important… Even without me asking specific questions.
Prescriptive Analytics	Tell me what my options are. Tell me what I should do.

Types of Analytics – Stages

Analytics Stage	Question
Descriptive	What happened?
Exploratory	What is going on?
Explanatory	Why did it happen? (Root Cause)
Predictive	What will happen?
Prescriptive	How do I take advantage?
Experimental	How well will it work?

Descriptive Analytics

What Happened?

The goal of descriptive analytics is to present numerical and summarized facts about the performance of a business.

Exploratory Analytics

What Is Going On?

The goal of exploratory data analytics is to deep dive into the data to understand patterns and confirm hypothesis.

Explanatory Analytics

Why Did It Happen?

The goal of explanatory data analytics is to find reasons for business results.

Predictive Analytics

What Will Happen?

The goal of predictive analytics is to identify the likelihood of future outcomes based on historical data, statistics, and machine learning (ML).

Prescriptive Analytics

How Do I Take Advantage of It?

The goal of prescriptive analytics is to identify ways and means to take advantage of the findings and predictions provided by earlier stages of analytics.

Experimental Analytics

How Well Will It Work?

The goal of experimental analytics is to test a hypothesis or an alternative to understand actual performance on the field.

Importance of Probability and Data Visualization

What does Probability Convey?

There is Uncertainty in Statistics

Example

- There are 10000 people
- 1% have a rare disease
- There's a test that's 99% effective
  - 99% of sick patients test positive
  - 99% of healthy patients test negative
  - **There is 1% Error**

Example

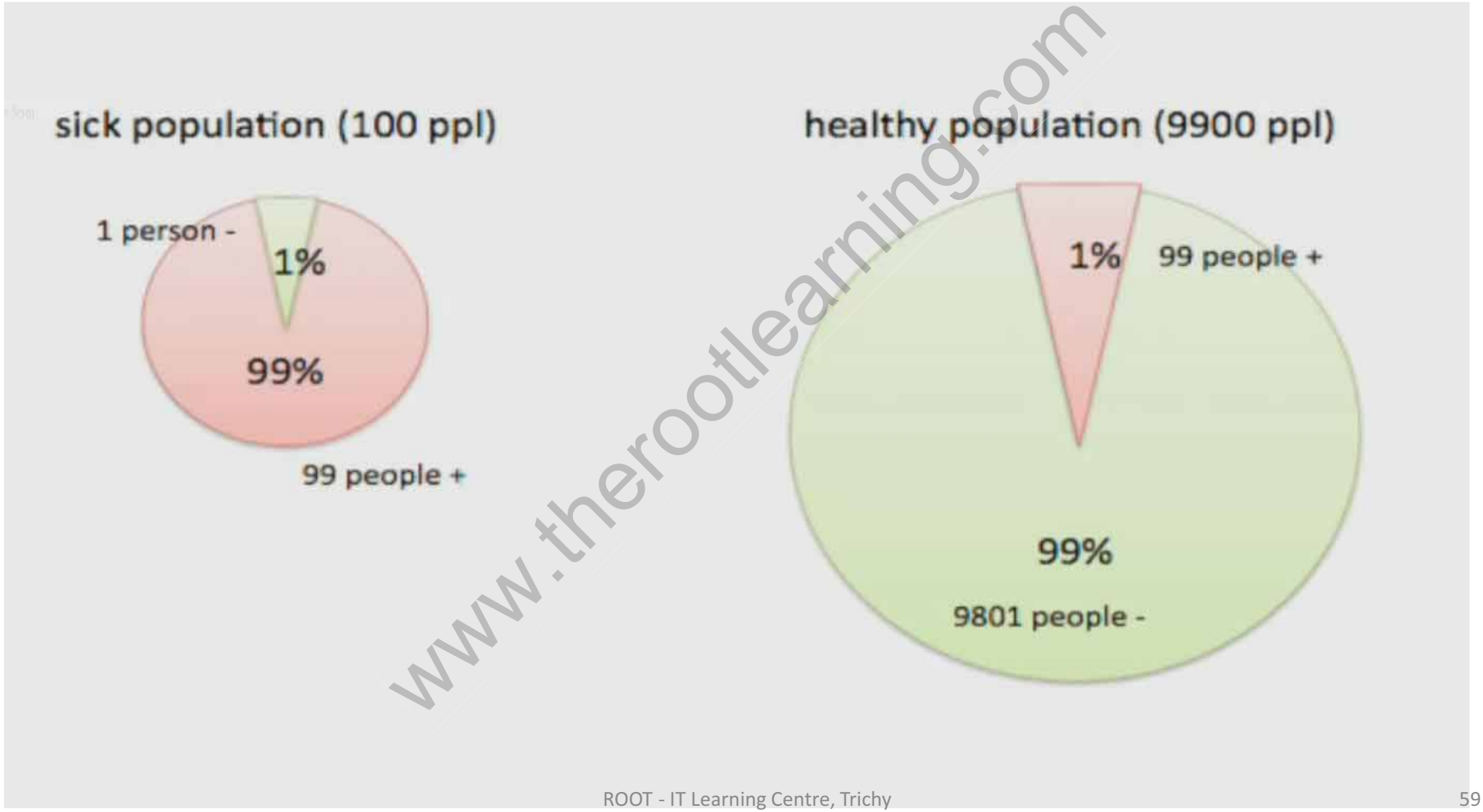
- Assume that we are testing everybody in the 10000 people population

- **Given a positive result, what is the probability that the patient is sick?**

Disease Diagnosis

- 99 sick patients test positive and 99 healthy patients test positive
- Given a positive test result, there is a 50% probability that the patient is sick…

Makes Sense….



Reliability of Metrics: Accuracy

Imbalance in Real Datasets

- Reevaluate how you think of probability…
- Credit Card Fraud Detection…
  - **9 out of every 10 transactions** has to be analyzed for fraud…
  - Why?
    - Imbalance is very high
    - Prediction Accuracy is Low (**Even 90% is too bad**)
- **Outlier Analysis** in general is affected by such imbalance

Data Visualization

Data Visualization is as important as Statistical Analysis

Anscombe's quartet

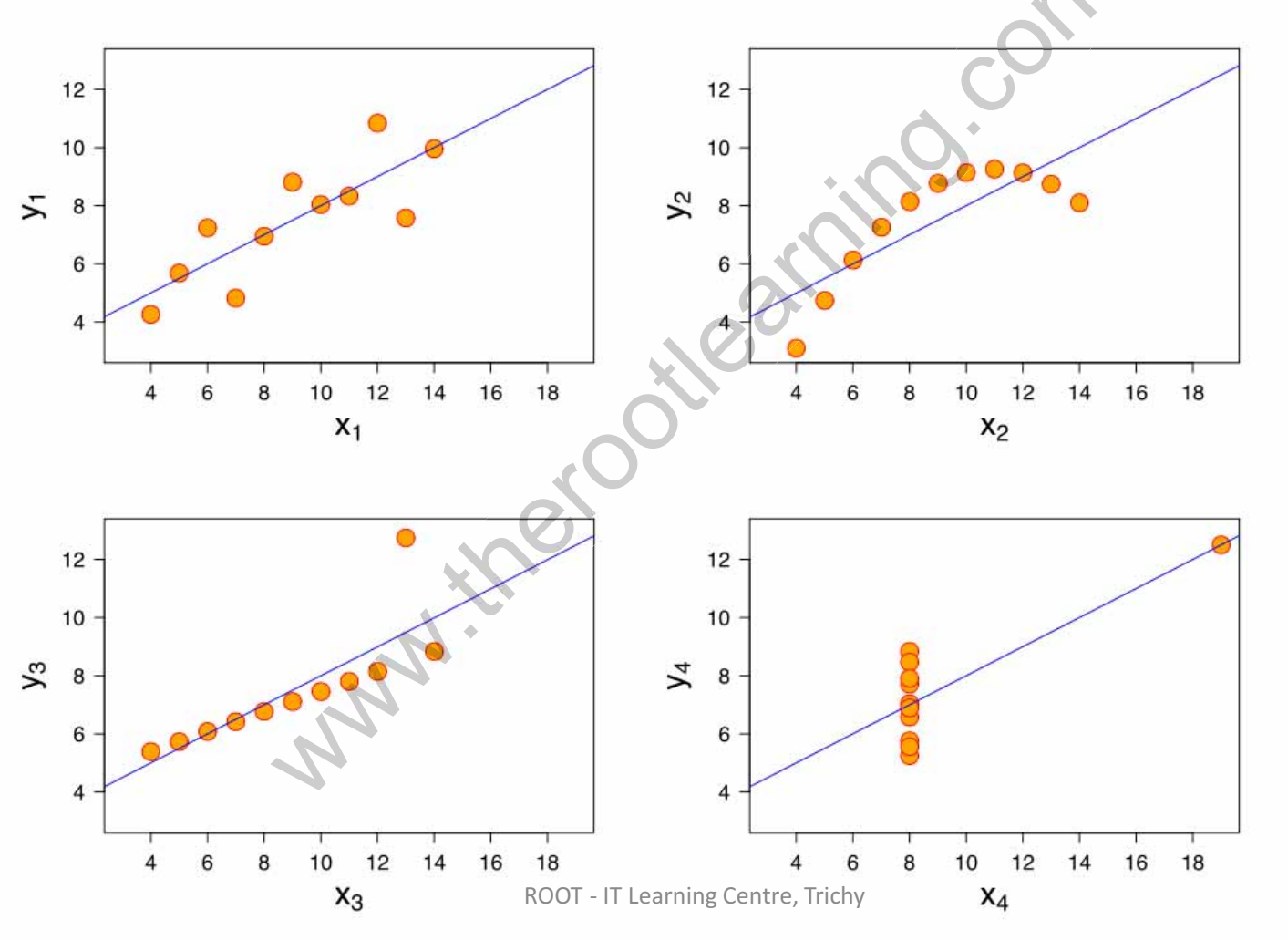
Statistics: Anscombe's quartet (1973)

Anscombe's quartet							
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

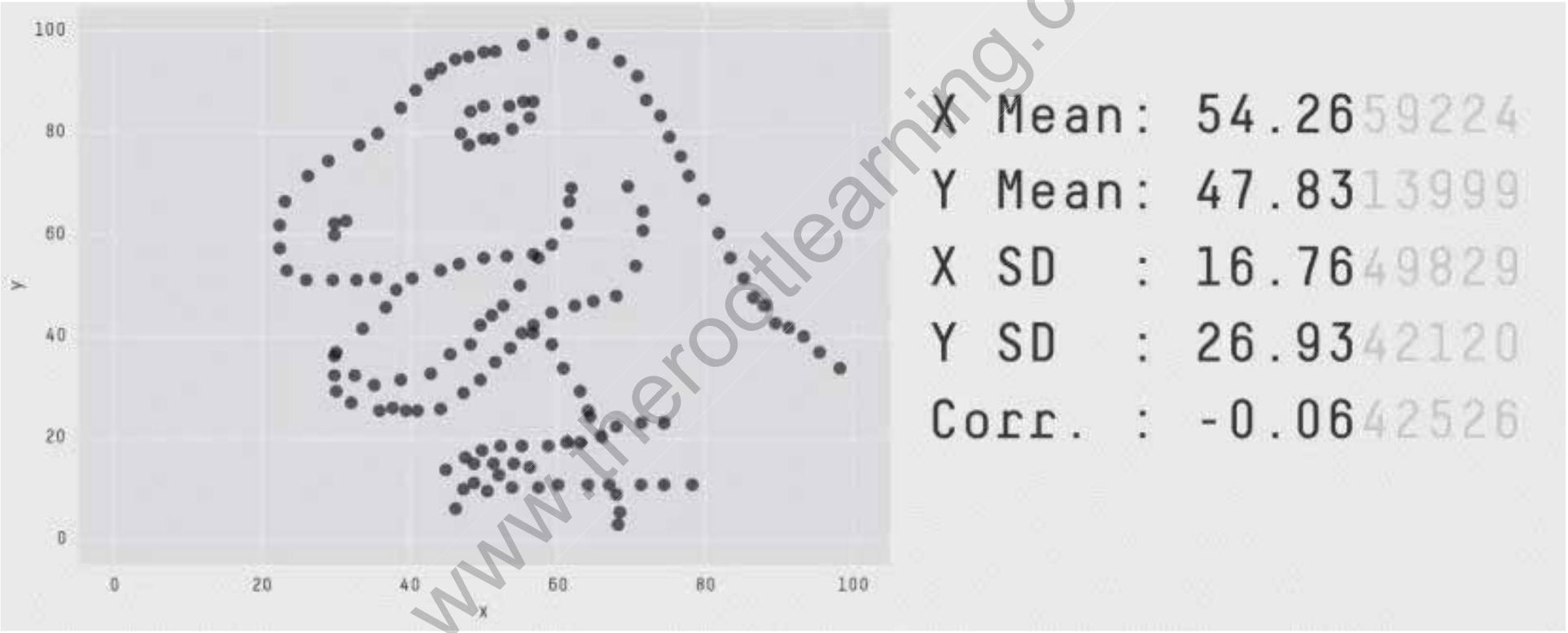
Statistical Values are the same for all four datasets

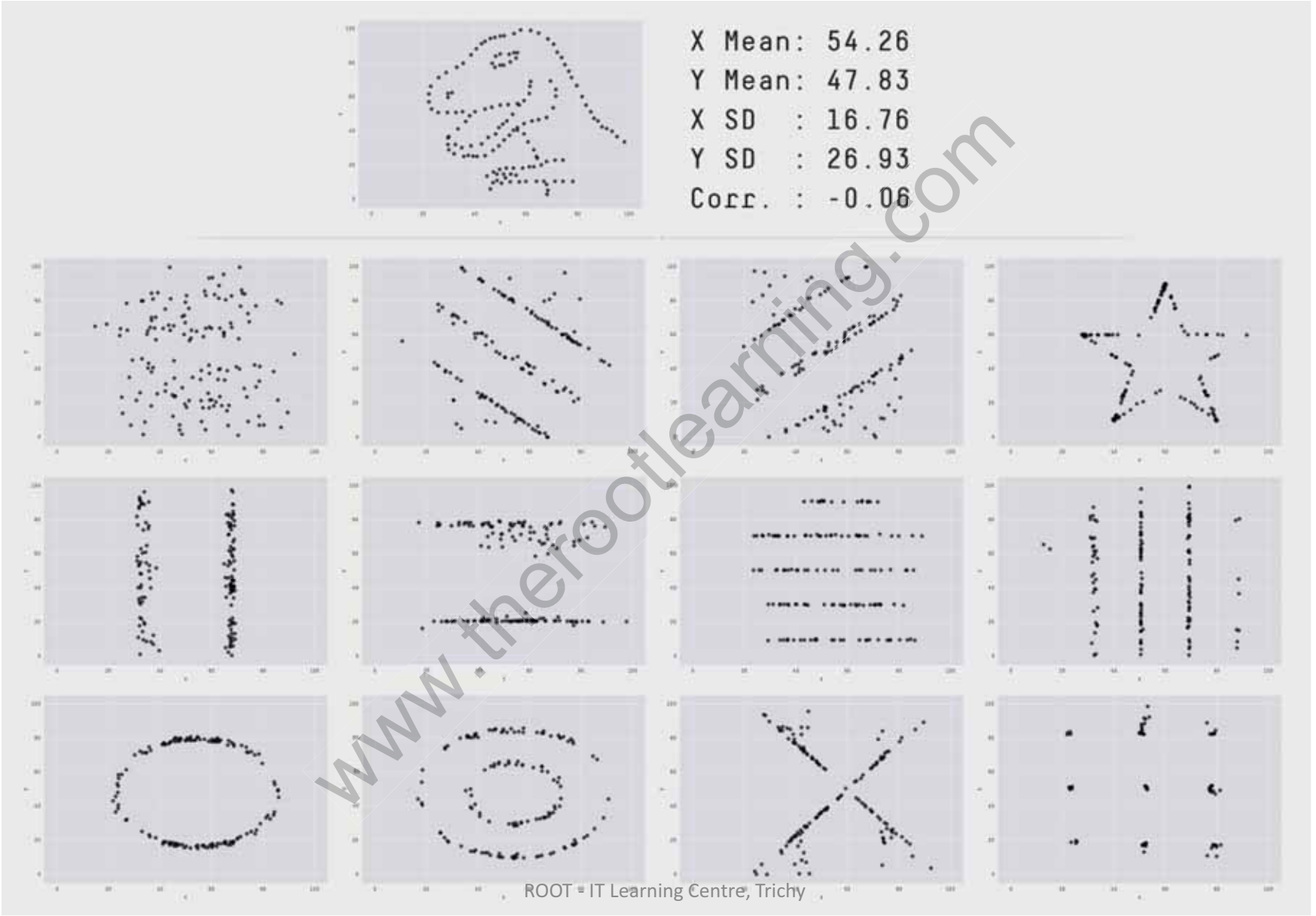
Property	Value
Mean of *x* in each case	9 (exact)
Sample variance of *x* in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Sample variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

But Visualization Differs



Datasaurus Dozen





Installation of Anaconda for Lab Session

- Visit the GitHub Link given below to download the installation instructions for Anaconda and Jupyter Notebook
- https://github.com/rootanand/CARE-FDP
- Follow the instructions and keep it ready for the final lab session
- Additional materials (if any) will be shared using the same GitHub repository

Thank You

Mobile : **9790636324**

Mail : **root.anand@gmail.com**

Web : **www.therootlearning.com**