Modul 3

Design Thinking in Data Science

Data Science Program

Purwadhika
Startup and Coding School

A small bit of review on Linear Regression

Quiz:

- *What is linear regression?*
- *What are the assumptions of linear regression?*
- *How can you measure the significance of the model?*
- *What is underfitting? How to solve it?*
- *What is overfitting? How to solve it?*

A small bit of review on Linear Regression

- Tool for finding existence of an association relationship between a dependent variable (Y) and one or more independent variables (X_1, X_2, …, X_n)
- Relationship can be linear or non-linear

A small bit of review on Linear Regression

- Use of co-efficient of determination to check the goodness of fit of regression
- Analysis of Variance (ANOVA) and F-test to check the overall fitness of the regression model
- T-test to validate relationship between dependent and independent variables
- Residual analysis to check the model adequacies

A small bit of review on Linear Regression

UNDERFITTING

- High Bias and Low Variance
- Simple Model
- Does not capture the underlying trend of the data

A small bit of review on Linear Regression

UNDERFITTING (Solutions)

- Train a more complex model
- Obtain more features

A small bit of review on Linear Regression

OVERFITTING

- High Variance and Low Bias
- Too complex
- Fits the data too well
- Learns the noise in the training data which impacts the performance on test data

A small bit of review on Linear Regression

OVERFITTING (Solutions)

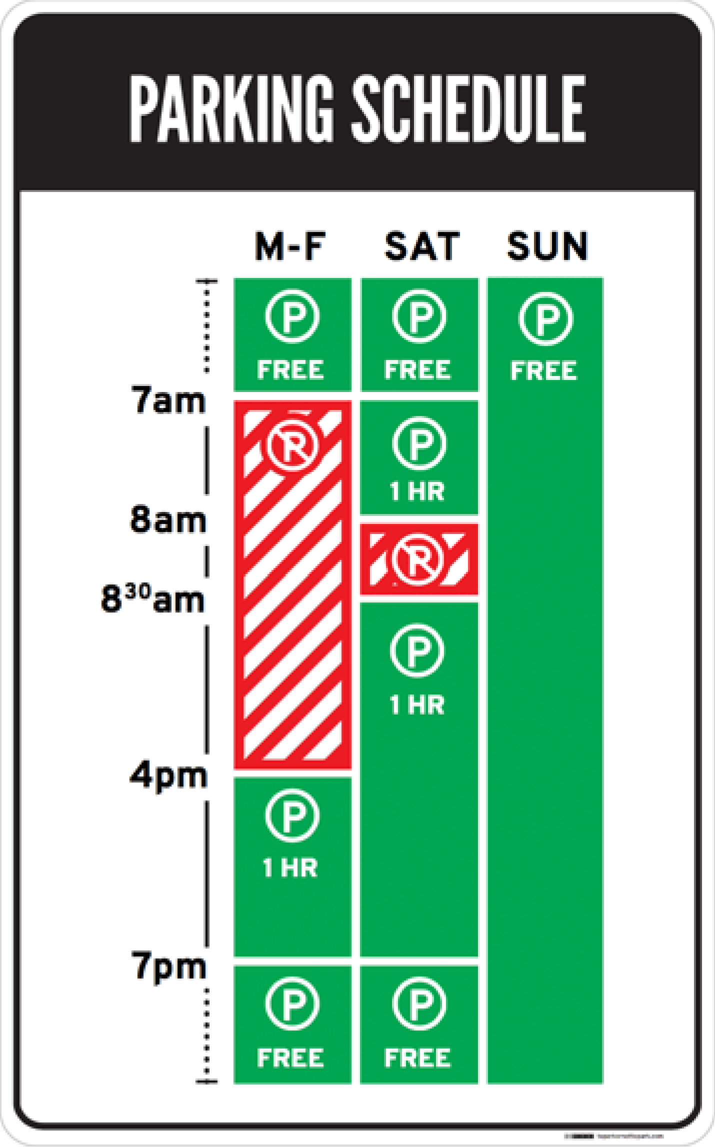
- Decrease the number of features
- Increasing the number of training examples

Roadmap

DESIGN THINKING IN DATA SCIENCE

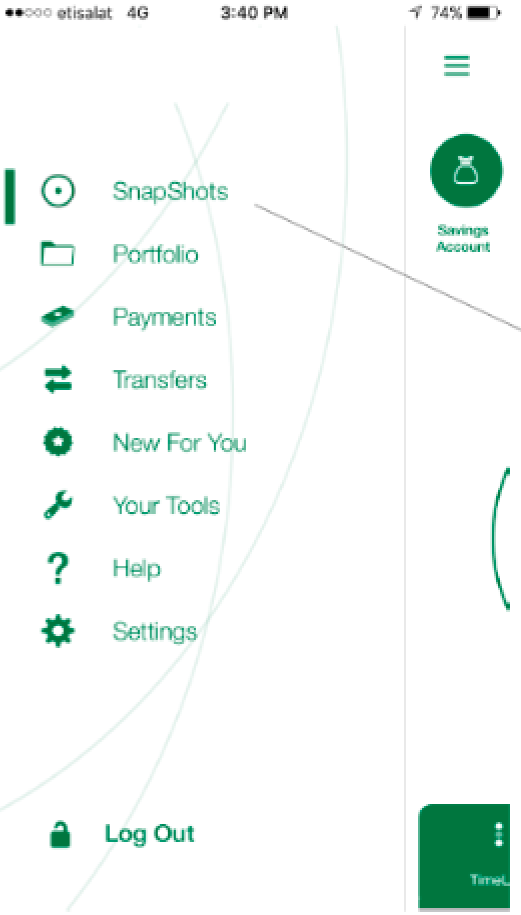
- Background
- A bit story about Design Thinking
- Design Thinking
- A bit story about CRISP-DM
- CRISP-DM
- Selecting Analytical Methodology
- Case Study

A bit of background and motivation



rwadhika
p and Coding School

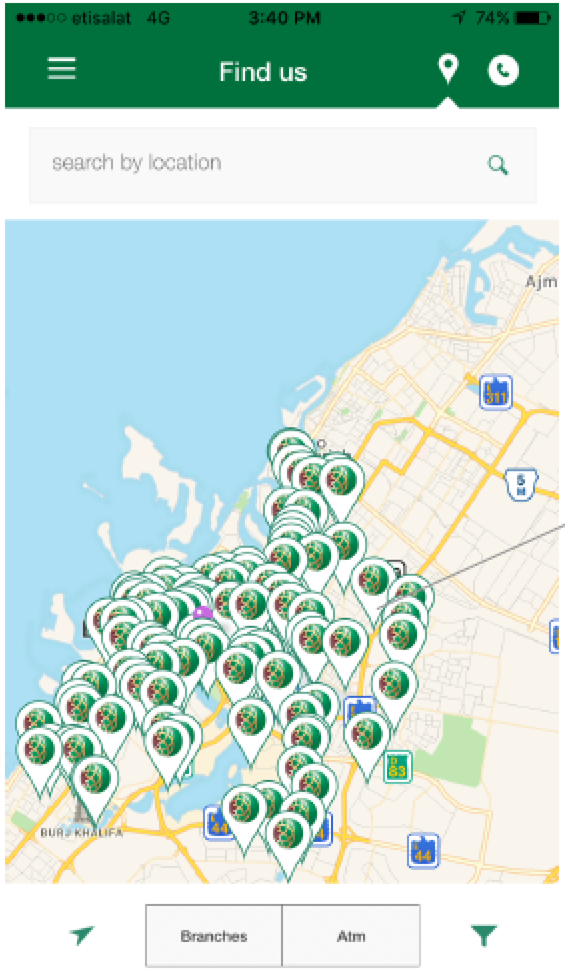
A bit of background and motivation



SnapShots meant as Dashboard/Home

App Navigation

A bit of background and motivation



Location Labels show all branches at one time

A bit of background and motivation

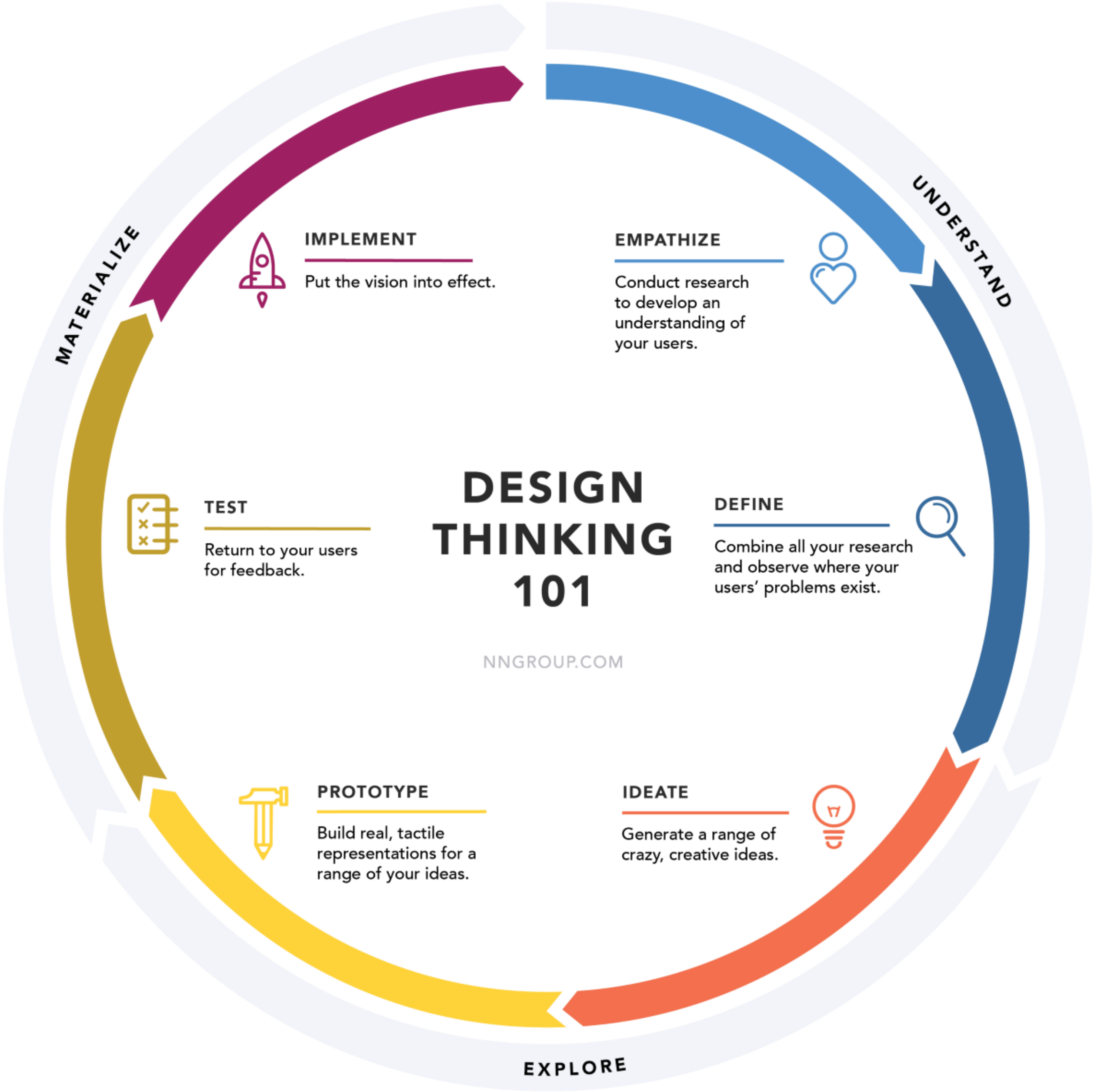
We need a framework to help us design better products

A bit of background and motivation

We need HUMAN-CENTERED DESIGN

Once upon a time ….

- 1951: John E. Arnold began teaching Creativity at MIT
- 1962: Conference on Systematic and Intuitive Methods in Engineering, Industrial Design, Architecture and Communications, London, UK, started interest studying design processes and developing new design methods.
- 1987: Peter Rowe publishes *Design Thinking*, focused on architecture and planning.
- 1991: The first symposium on Research in Design Thinking is held at Delft University, The Netherlands. IDEO design consultancy formed by combining three industrial design companies. They are one of the first design companies to showcase their design process, which draws heavily on the Stanford University curriculum.
- 2005: Stanford University's d.school begins to teach engineering students design thinking as a formal method.



DESIGN THINKING 101

NNGROUP.COM

UNDERSTAND

EMPATHIZE

Conduct research to develop an understanding of your users.

DEFINE

Combine all your research and observe where your users' problems exist.

EXPLORE

IDEATE

Generate a range of crazy, creative ideas.

PROTOTYPE

Build real, tactile representations for a range of your ideas.

MATERIALIZE

TEST

Return to your users for feedback.

IMPLEMENT

Put the vision into effect.

urwadhika
ırtup and Coding School

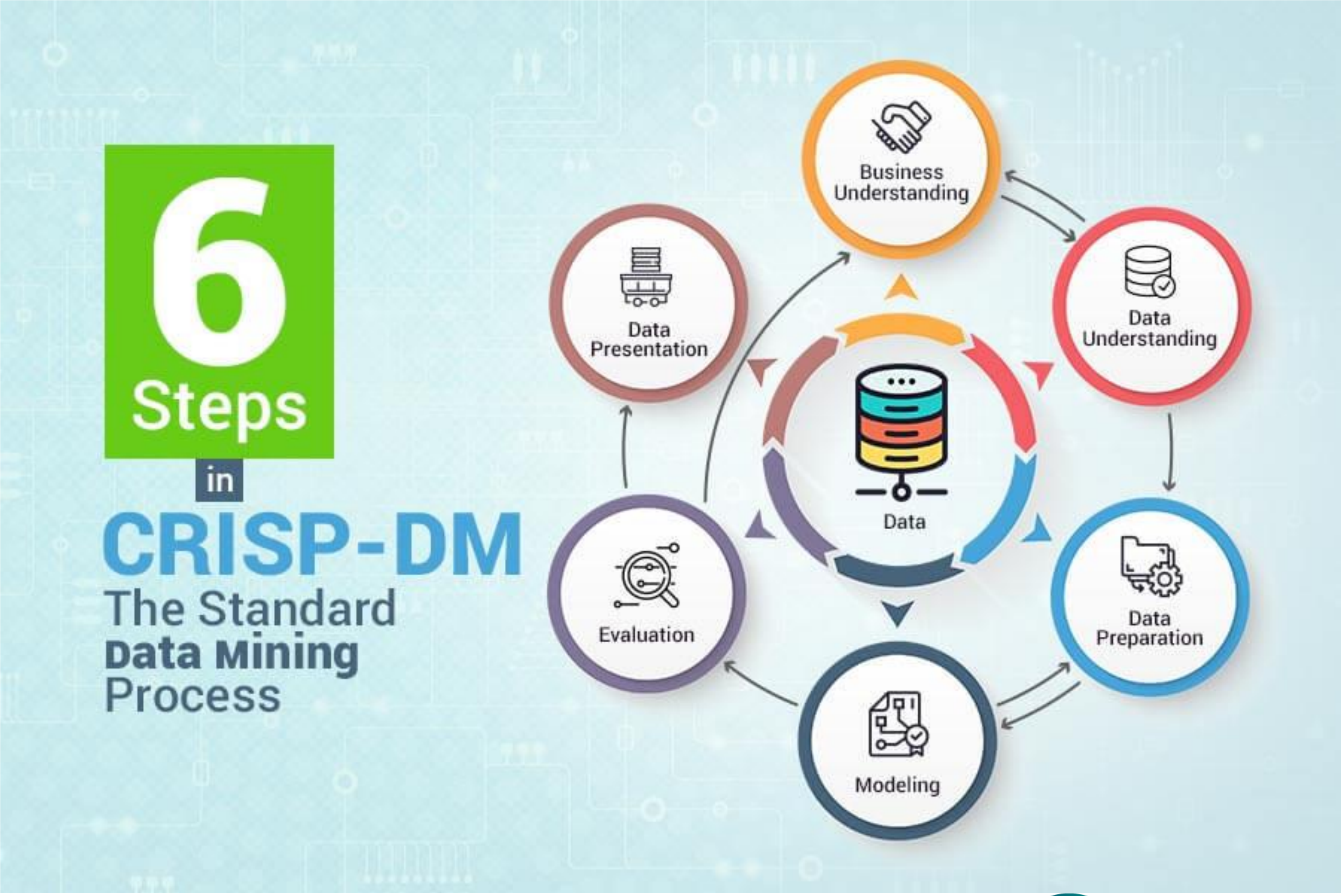
Case Study (Design Thinking on the spot)

Once upon a time ….

- 1996: DamilerChrysler, SPSS, NCR
- 1997: CRISP-DM got funding from European Comission
  - intended to be industry-, tool-, and application-neutral (CRISP-DM SIG-- Special Interest Group, Amsterdam day-workshop)
- 1997-1998: CRISP-DM SIG evolved into more than 200 members. Conferences (New York, London, and Brussels). Cases on Mercedes-Benz and OHRA
- 1999: The end of EC-funded project. CRIPS-DM 1.0
- 2006: Trial to move forward to CRIPS-DM 2.0 but the progress is left unknown until now.

Once upon a time ….

“We need to be aware that we are standing on the shoulders of the giants, each one of them with their own struggles.”



Purwadhika
Startup and Coding School

Business Understanding

"This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used."

Key Questions:

- What decisions needs to be made?
- What information is needed to inform those decisions?
- What type of analysis can provide the information needed to inform those decisions?

Purwadhika
Startup and Coding School

Case

“Berapa banyak beras yang harus disediakan oleh pemerintah untuk konsumsi di daerah DKI Jakarta pada setiap bulan di tahun 2019?”

~Business understanding?~

Data Understanding

"The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information."

Key Questions:

- What data is needed?
- What data is available?
- What are the important characteristics of the data?

Case

“Berapa banyak beras yang harus disediakan oleh pemerintah untuk konsumsi di daerah DKI Jakarta pada setiap bulan di tahun 2019?”

~Data understanding?~

Data Preparation

"The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools."

Data Preparation

Common Steps Used in Data Preparation

- Gathering: When gathering data - you may need to collect data from multiple sources within your organization.
- Cleansing: The data set you are working with may have issues that you want to resolve prior to your analysis. This can be in the form of incorrect or missing data.
- Formatting: You may need to format the data by changing the way a date field appears, renaming a field, or even rotating the data, similar to using a pivot table.
- Blending: You may want to blend, or combine, your data with other datasets to enrich it with additional variables, similar to using the vlookup function in Excel.
- Sampling: Lastly, you may want to sample the dataset and work with a more manageable number of records.

Analysis and Modelling

"In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed."

Important Steps

- Determine what methodology to use to solve the problem
- Determine the important factors or variables that will help solve the problem
- Build a model to solve the problem
- Run the model and move to the evaluation phase

Case

“Berapa banyak beras yang harus disediakan oleh pemerintah untuk konsumsi di daerah DKI Jakarta pada setiap bulan di tahun 2019?”

~Data understanding?~

Evaluation

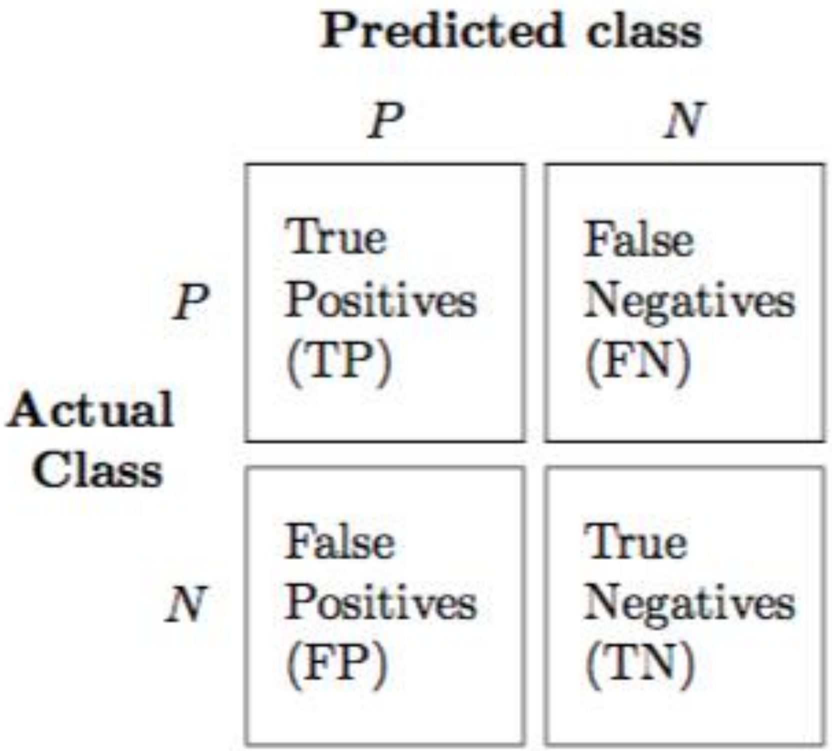
"At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached."

Evaluation

Important Steps

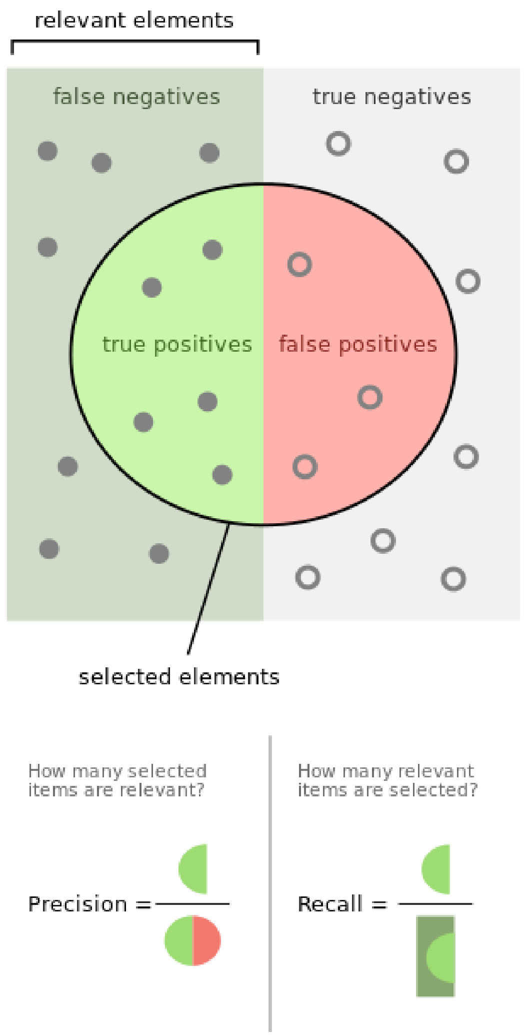
- Observe the key results on the model
- Ensure the results make sense within the content of the business problem
- Determine whether to proceed to the next step or return to a previous phase
- Repeat as many times as necessary

Confusion Matrix

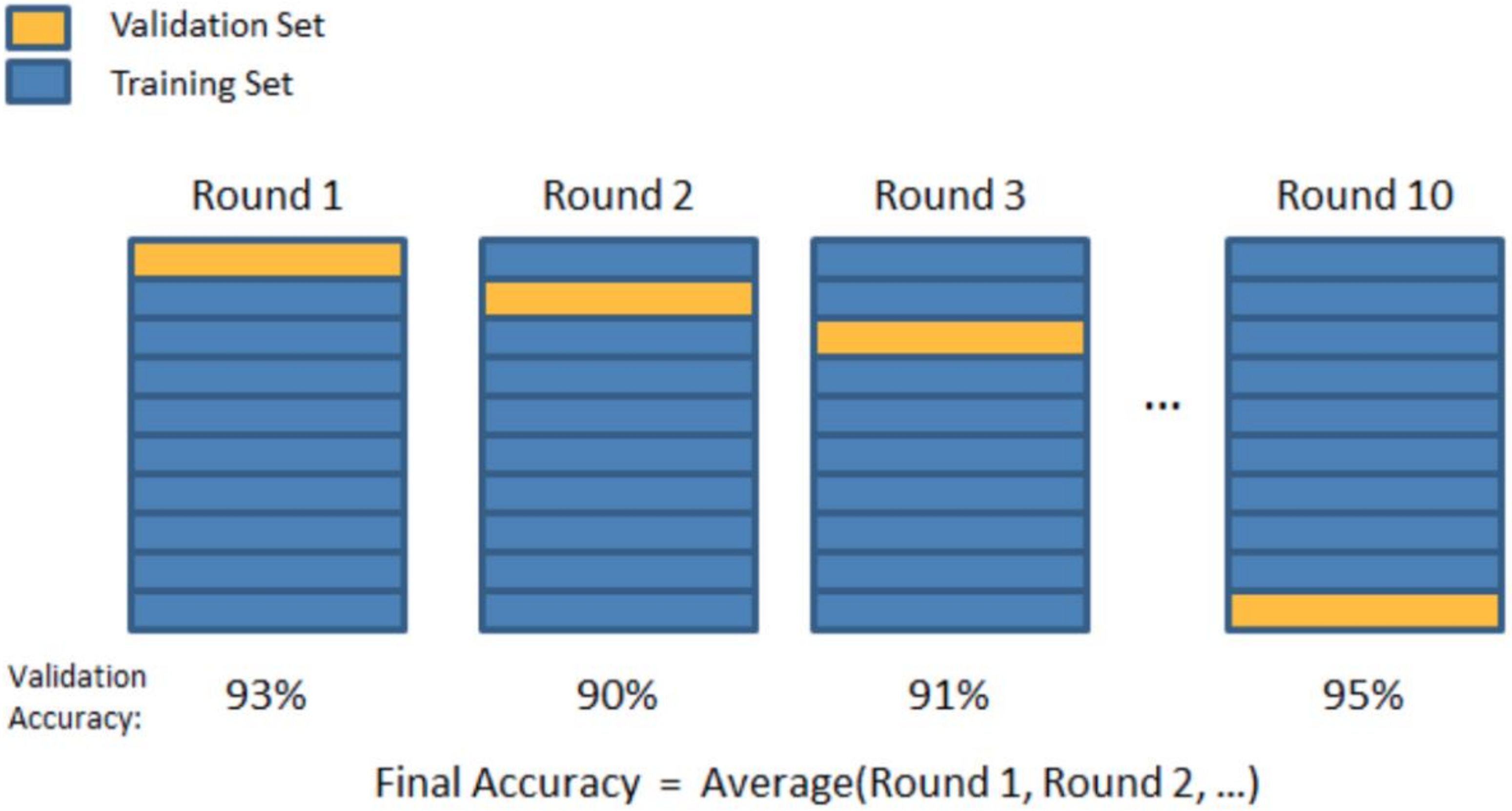


		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Precision & Recall



k-fold Cross Validation



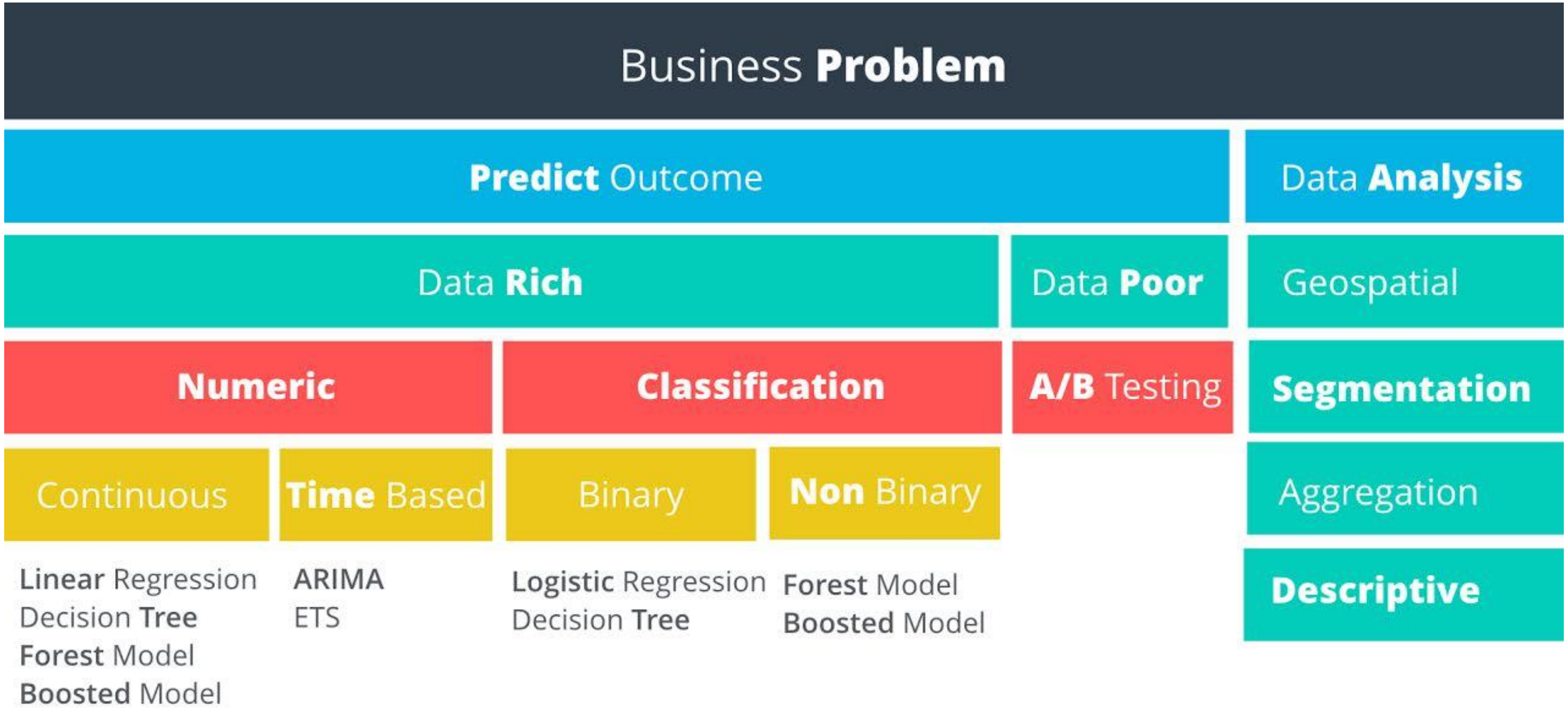
Presentation and Visualization

"Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. Even if the analyst deploys the model it is important for the customer to understand up front the actions which will need to be carried out in order to actually make use of the created models."

Presentation and Visualization

Key Considerations

- Determine the best method of presenting insights based on the analysis
- Determine the best method of presenting insights based on the audience
- Make sure the amount of information shared is not overwhelming
- Use the results to tell a story to the audience
- For more complex analyses, you may want to walk the audience through the analytical problem solving process
- Always reference the data sources used
- Make sure your analysis supports the decisions that need to be made



Business **Problem**

- **Predict** Outcome
  - Data **Rich**
    - **Numeric**
      - Continuous
        - **Linear** Regression
        - Decision **Tree**
        - **Forest** Model
        - **Boosted** Model
      - **Time** Based
        - **ARIMA**
        - ETS
    - **Classification**
      - Binary
        - **Logistic** Regression
        - Decision **Tree**
      - **Non** Binary
        - **Forest** Model
        - **Boosted** Model
  - Data **Poor**
    - **A/B** Testing
- Data **Analysis**
  - Geospatial
  - **Segmentation**
  - Aggregation
  - **Descriptive**

Non Predictive Analysis

- Geospatial
- Segmentation
- Aggregation
- Descriptive

Purwadhika
Startup and Coding School

Geospatial

This type of analysis uses location based data to help drive your conclusions. Examples include identifying customers by a geographic region, calculating the distance store locations or creating a trade area based upon customer locations.

Segmentation

Segmentation is the process of grouping data together. Groups can be simple, such as customers who have purchased different items, to more complex segmentation techniques where you identify stores that are similar based upon the demographics of their customers.

Aggregation

This methodology simply means calculating a value across a group or dimension and is commonly used in data analysis. For example, you may want to aggregate sales data for a salesperson by month - adding all of the sales closed for each month. Then, you may want to aggregate across dimensions, such as sales by month per sales territory. Aggregation is often done in reporting to be able to "slice and dice" information to help managers make decisions and view performance.

Descriptive Analysis

Descriptive statistics provides simple summaries of a data sample. Examples could be calculating average GPA for applicants to a school, or calculating the batting average of a professional baseball player. In our electricity supply scenario, we could use descriptive statistics to calculate the average temperature per hour, per day, or per date. Some of the commonly used descriptive statistics are Mean, Median, Mode, Standard Deviation, and Interquartile range.

Predictive Analysis

Data Rich vs. Data Poor

Purwadhika
Startup and Coding School

Data Poor

If there is not sufficient usable data to solve the problem, then we need to set up an experiment to help us get the data we need. An experiment in a business context is usually referred to as an A/B Test.

Data Rich

Numeric vs. Non-Numeric Predictive Analysis

Assuming we have enough data to proceed with the analysis, our next decision is to look at the outcome we're trying to predict and determine if it's a numeric outcome or a non-numeric outcome.

Data Rich

Regression Models

Numeric outcomes are those where the outcome is simply a number. Predicting the demand for electricity or the hourly temperature are both numeric outcomes. Models predicting numeric data are called regression models.

Data Rich

Classification Models

Non-numeric outcomes are those where we’re trying to predict the category into which a case (e.g. customer) falls, such as whether a customer will pay on-time, pay late, or default on a payment. Another example is the whether an electronic device will fail before 1000 hours or not. Models predicting non-numeric data are called classification models.

Examples

Tricycle Manufacturer's Production Department

For our first example, imagine that a manufacturer wants to use historical production data to know how many tricycles they’ll need to produce over the next six months to meet expected demand.

Hot & Fresh Pizza's Marketing Department

For our second example, Hot & Fresh Pizza wants to use sales data from their existing stores and respective demographic data around those stores to predict how many pizzas they’ll sell at their new store location.

Examples

Risk Management Department at a Bank

And for our third example, a bank wants to use historical data of their clients to predict whether a new customer will default on a loan, always pay on time, or sometimes pay.

Quick Questions

Numerical Models

Target Variables

Target variables represent the outcome we are trying to predict. In order to select the right predictive model, we first determine whether the target variable is numeric or non-numeric. The type of numeric or non-numeric target variables will then help us select which model is appropriate. Let’s start with numeric variables.

Numerical Models

Types of Numeric Variables

The three most common types of numeric variables are continuous, time-based, and count.

Continuous

A continuous variable is one that can take on all values in a range. For instance your height can be measured down to many decimal places. We do not grow in even inch intervals.

Time-Based

A time-based numeric variable is one where you are trying to predict what will happen over time. This is often related to forecasting.

Count

Count variables are numbers that are discrete, positive integers. They're called count numbers because they're used to analyze variables that you can count. As modeling these type of variables is not common in business, we won't be covering this topic in this course.

Non-Numerical Models

Non-Numeric Variables

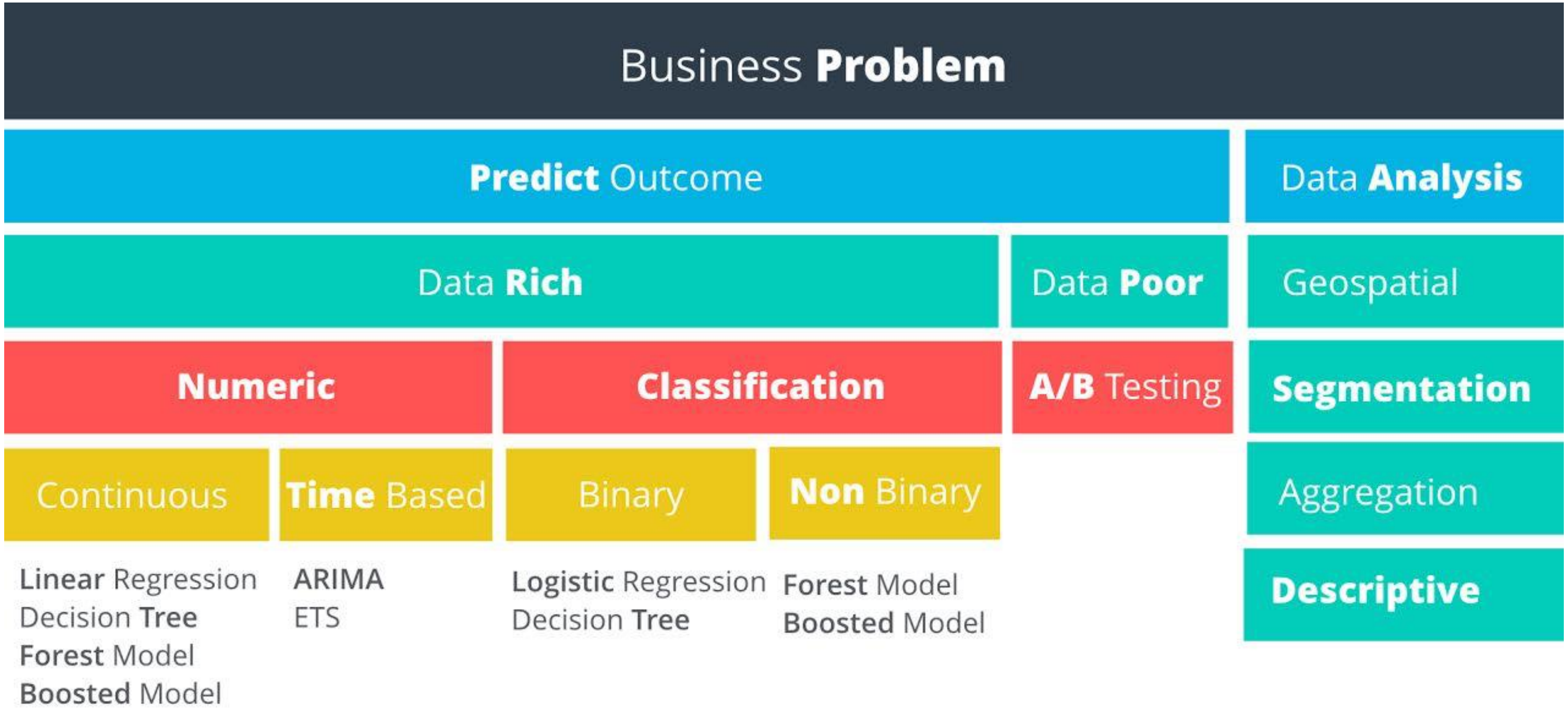
A non-numeric variable is often called categorical, because the values of the variable take on a discrete number of possible values or categories. Examples include whether an electronic device will fail before 1000 hours or not; whether a customer will pay on-time, pay late, or default on a payment, or whether a store is classified as large, medium or small.

Non-Numerical Models

Classification Models: Binary and Non-Binary

When modeling categorical variables, the number of possible outcomes is an important factor. If there are only two possible categorical outcomes, such as Yes or No, or True or False, then the variable can be described as Binary.

If there are more than two possible categorical outcomes, such as small, medium, or large, or pay on-time, pay late, or default on a payment, then the variable can be described as non-binary. The important take-away from this lesson is the ability to determine if you should use a classification model, and whether it should be a binary model or a non-binary model. Ben Burkholder will lead a course focused on classification models and will go into detail about these types of models.



Business **Problem**

- **Predict** Outcome
  - Data **Rich**
    - **Numeric**
      - Continuous
        - **Linear** Regression
        - Decision **Tree**
        - **Forest** Model
        - **Boosted** Model
      - **Time** Based
        - **ARIMA**
        - ETS
    - **Classification**
      - Binary
        - **Logistic** Regression
        - Decision **Tree**
      - **Non** Binary
        - **Forest** Model
        - **Boosted** Model
  - Data **Poor**
    - **A/B** Testing
- Data **Analysis**
  - Geospatial
  - **Segmentation**
  - Aggregation
  - **Descriptive**

Quick Questions

Case Study

Purwadhika
Startup and Coding School

Wrap Up!

- Design Thinking: Human-centered Design
- Design Thinking Steps: Empathy, Definition, Ideation, Prototype, Test, Implement.
- CRIPS-DM: Cross-Industry Standard Process for Data Mining
- CRIPS-DM Steps: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Presentation.
- Mostly used data-mining model:
  - Predictive: Regression (Numeric), Classification (Categorical)
  - Non-Predictive: Clustering, Ranking, Summarization