Data Science Project Lifecycle

AI4I-DSPM-1

Black Raven (James Ng)

07 Mar 2021 · 21 min read

This is a memo to share what I have learnt in Data Science Project Lifecycle, capturing the learning objectives as well as my personal notes. The course is by Infocomm Media Development Authority (IMDA) AI For Industry (AI4I) with 24 slides compiled by LIM Tern Poh.

AI SINGAPORE

AI for Professionals (AI4P) Course
Lesson 5: Data Science Project Lifecycle

LIM Tern Poh
AI Engineer | AI Industry Innovation

Typical Steps of Data Science Project Workflow

1. Who are my team?
Who are the crucial team members in a data science project? What roles they play?

2. Has The Wheel Been Invented?
Has there been similar model done? How others solved it?

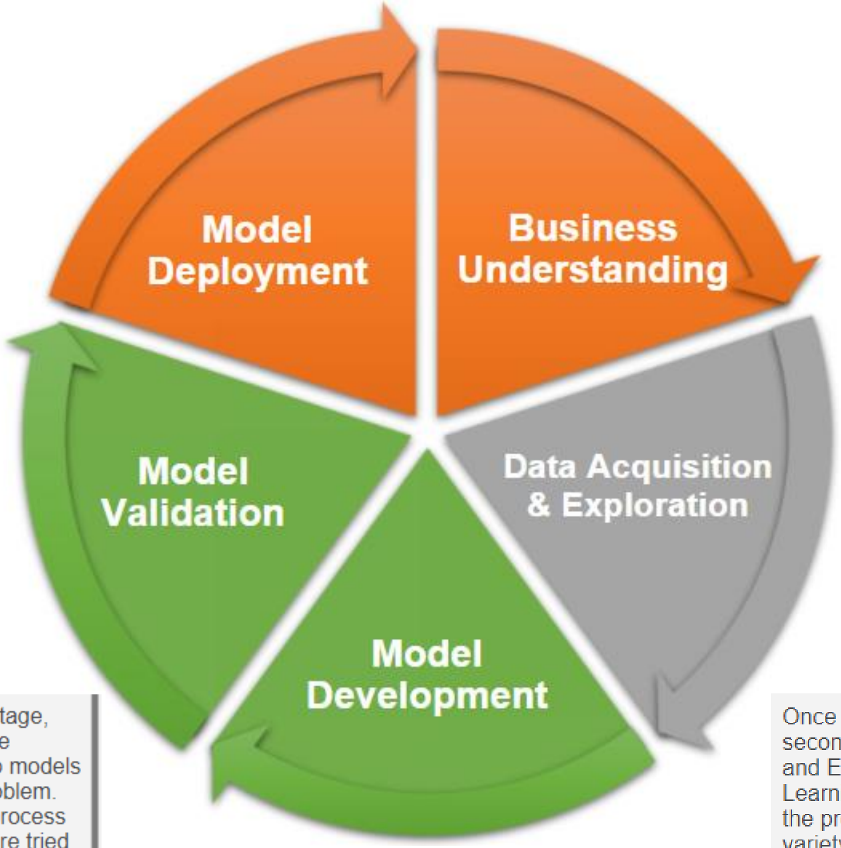
3. How To Manage Data Science Project?
Introduction to commonly used SCRUM methodology to maximise project's success rate

4. How To Analyse Model's Performance?
Using right metrics to measure and improve model's performance

5. What Are The Ways To Integrate The Model Into It System?
Deploying the model to make it accessible to others

5 Stages of Data Science Project

There are five main stages in a typical data science project lifecycle; each stage builds on previous stages to lead to a successful data science project. The stages are in an iterative loop. For instance, after the model is deployed, there could be feedbacks from users or new business requirements. This leads to another round of data science project lifecycle to improve existing or create new data science model.



The first stage of the data science project starts with Business Understanding; this is perhaps the most important stage. It is at this stage where the problem is defined and scoped; if the problem is poorly-defined, it will lead to wasted time and investment. The model developed from the data science project will not be able to solve an actual business problem and deliver business value to users.

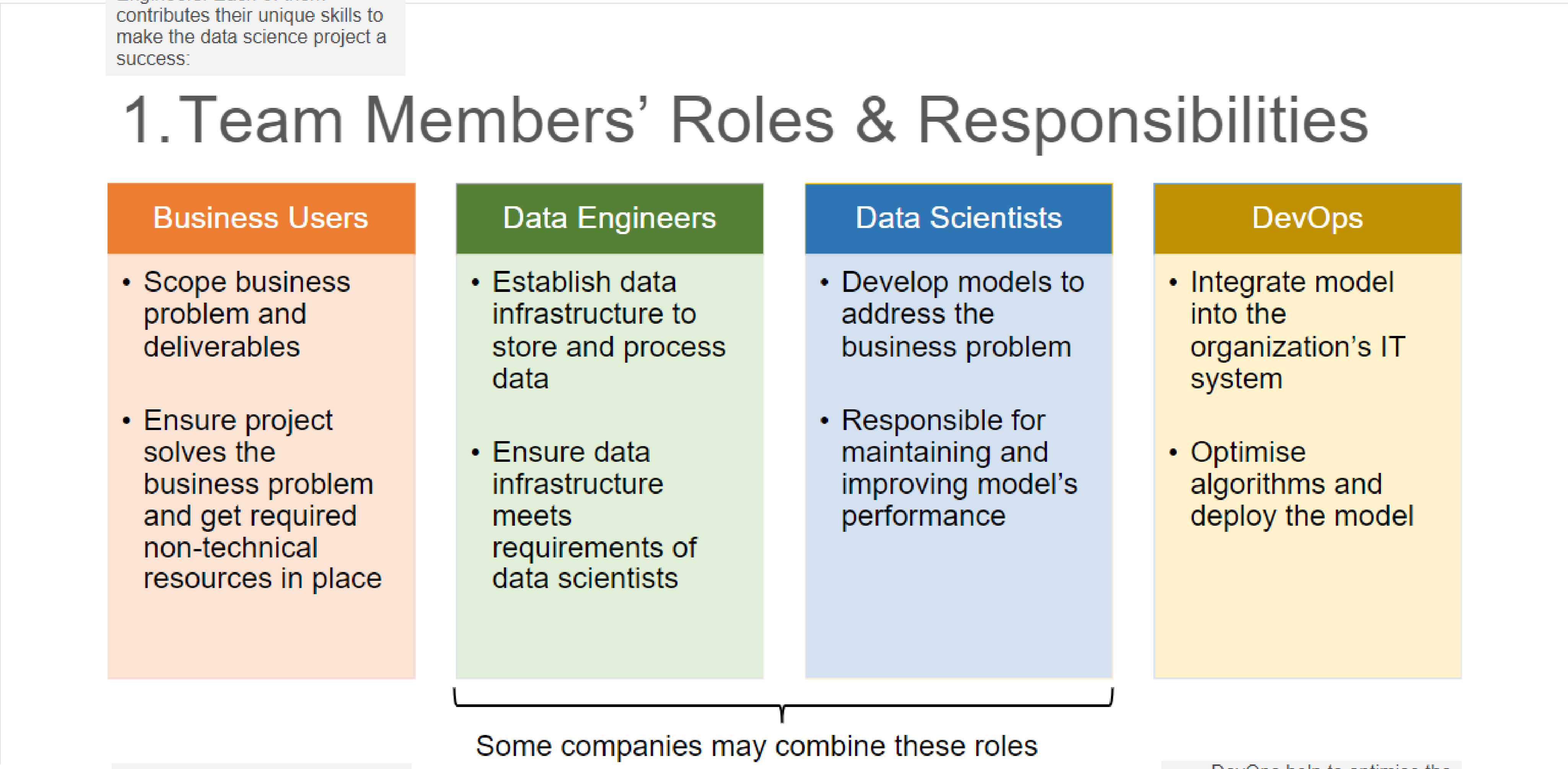
Once the problem is defined, the second stage is Data Acquisition and Exploration. Machine Learning requires data relevant to the problem, lots of it and a variety of it, to train a model. If the company already has the required data, great. Otherwise, the company will have to determine how to best collect the required data. It could be done by purchasing data from third-party or setting up a new data pipeline to collect data from new sources. Exploratory Data Analysis is then done to understand the structure of the data collected; it could be as simple as plotting graphs to advanced statistical methods.

At Model Development stage, Data Scientist will use the collected data to develop models to solve the business problem. This will be an iterative process where different models are tried and tested. The best performing model is then Validated by using appropriate metrics, such as accuracy score, to gauge the model's performance. If the model's performance meets expectation, it will go into the Model Deployment stage. DevOps or Software Engineers will integrate the model into the company's IT system to make it easily accessible to business users.

Steps of Data Science Project Workflow

1. Who are the crucial team members in a data science project? What roles they play?
2. Has the wheel been invented? How others solve similar problems?
3. How to manage data science project with SCRUM methodology?
4. How to analyse model’s performance and improve it?
5. What are the ways to integrate model into IT system?

A data science team is a team sport. It involves not just Data Scientists, but also Business Users, DevOps, and Data Engineers. Each of them contributes their unique skills to make the data science project a success:



1. Team Members' Roles & Responsibilities

Business Users	Data Engineers	Data Scientists	DevOps
• Scope business problem and deliverables • Ensure project solves the business problem and get required non-technical resources in place	• Establish data infrastructure to store and process data • Ensure data infrastructure meets requirements of data scientists	• Develop models to address the business problem • Responsible for maintaining and improving model's performance	• Integrate model into the organization's IT system • Optimise algorithms and deploy the model

Some companies may combine these roles

- Business Users have business domain knowledge and are usually proficient in business intelligence and data visualisation. They help to identify what problems are worth solving for the organisation;

- Data Engineers prepare the necessary data infrastructure for data scientists to perform their modelling work;

- Data Scientists experiment and develop models to address the problem;

- DevOps help to optimise the algorithms and deploy the model to make it accessible in the organisation.

Occasionally, some organisations combine DevOps and Data Scientists into a single position.

Steps of Data Science Project Workflow

1. Who are the crucial team members in a data science project? What roles they play?
2. Has the wheel been invented? How others solve similar problems?
3. How to manage data science project with SCRUM methodology?
4. How to analyse model's performance and improve it?
5. What are the ways to integrate model into IT system?

AI SINGAPORE

2. AI Technology Landscape Scan (examples)

AI SINGAPORE MAKERSPACE

https://makerspace.aisingapore.org/

GitHub

https://github.com/

Towards Data Science

https://towardsdatascience.com/

Google Scholar

https://scholar.google.com/

AI technology landscape scan is about researching what are the available solutions and how others have applied the solutions.

Given the pace of data science development, it is likely that someone who has solved the same problem you are facing. A robust technology landscape scan could help you accelerate your data science project by either re-applying or adapting others' solutions to your problems.

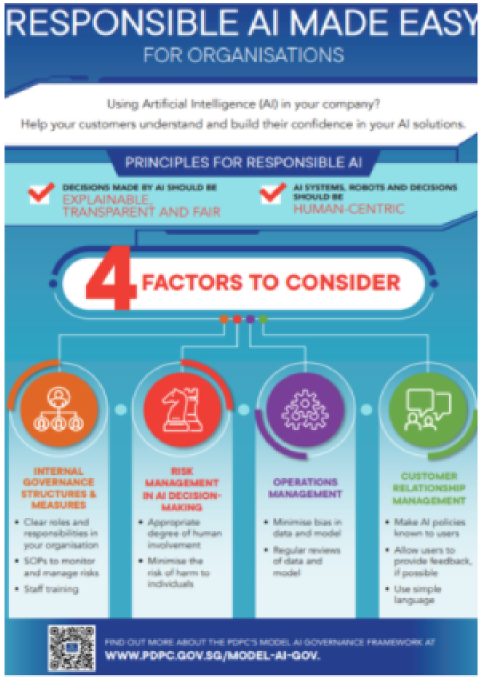
The general recommendation is to start at Share AI (https://makerspace.aisingapore.org/share-ai/), maintained by AI Singapore's engineers, apprentices and local AI community. This provides a snapshot of the latest AI techniques, codes, best practices and 100E projects done at AI Singapore. Another alternative is TowardsDataScience (https://towardsdatascience.com/); most of the articles here are well-written and easy to follow.

If you are unable to find what you need, the next step is to search on Google Scholar (https://scholar.google.com/) for the latest research journals. Another alternative is GitHub (https://github.com/); AI practitioners could have worked on a similar problem and shared their code publicly there.

2. AI Governance Framework

- Singapore's PDPC (Personal Data Protection Commission) Singapore has an AI Governance Framework
- Framework frames discussions around the challenges and possible solutions to harness AI in a responsible way
- Full report is accessible via

  https://www.pdpc.gov.sg/MODEL-AI-GOV



RESPONSIBLE AI MADE EASY

FOR ORGANISATIONS

Using Artificial Intelligence (AI) in your company?

Help your customers understand and build their confidence in your AI solutions.

PRINCIPLES FOR RESPONSIBLE AI

✓ DECISIONS MADE BY AI SHOULD BE EXPLAINABLE, TRANSPARENT AND FAIR

✓ AI SYSTEMS, ROBOTS AND DECISIONS SHOULD BE HUMAN-CENTRIC

4 FACTORS TO CONSIDER

INTERNAL GOVERNANCE STRUCTURES & MEASURES

- Clear roles and responsibilities in your organisation
- SOPs to monitor and manage risks
- Staff training

RISK MANAGEMENT IN AI DECISION-MAKING

- Appropriate degree of human involvement
- Minimise the risk of harm to individuals

OPERATIONS MANAGEMENT

- Minimise bias in data and model
- Regular reviews of data and model

CUSTOMER RELATIONSHIP MANAGEMENT

- Make AI policies known to users
- Allow users to provide feedback, if possible
- Use simple language

FIND OUT MORE ABOUT THE PDPC'S MODEL AI GOVERNANCE FRAMEWORK AT WWW.PDPC.GOV.SG/MODEL-AI-GOV.

AI SINGAPORE

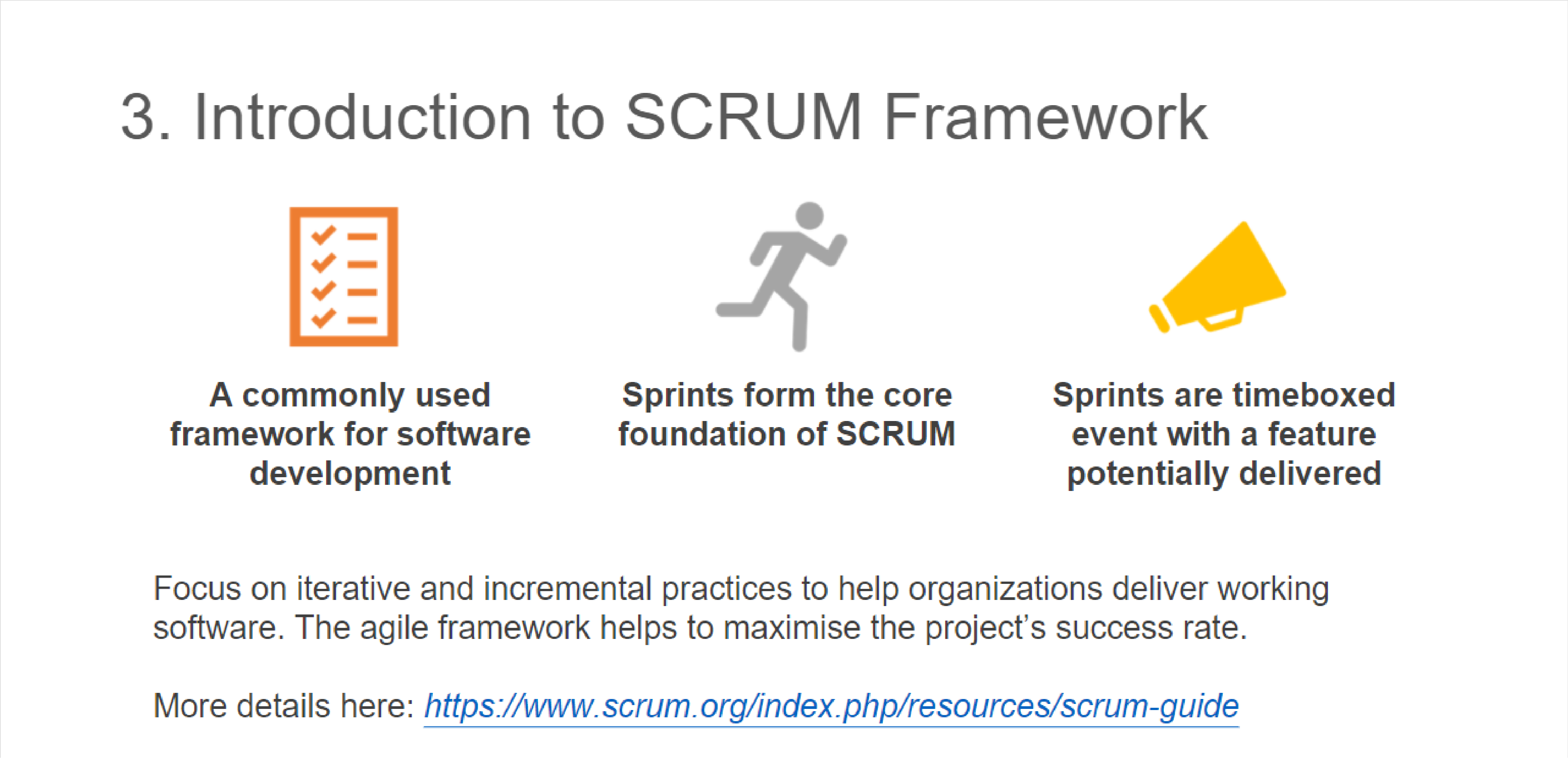
While conducting AI technology landscape scan, do keep a lookout for applicable regulation to your model.

In Singapore, for instance, PDPC has developed the first edition of a Model AI Governance Framework (Model Framework) - an accountability-based framework to help chart the language and frame the discussions around harnessing AI in a responsible way. The Model Framework translates ethical principles into practical measures that can be implemented by organisations deploying AI solutions at scale. Through the Model Framework, PDPC aims to promote AI adoption while building consumer confidence and trust in providing their personal data for AI.

While every country has a different set of regulations for AI usage, this Model Framework could serve as a general framework when discussing potential AI project with your stakeholders.

Steps of Data Science Project Workflow

1. Who are the crucial team members in a data science project? What roles they play?
2. Has the wheel been invented? How others solve similar problems?
3. How to manage data science project with SCRUM methodology?
4. How to analyse model's performance and improve it?
5. What are the ways to integrate model into IT system?



3. Introduction to SCRUM Framework

A commonly used framework for software development

Sprints form the core foundation of SCRUM

Sprints are timeboxed event with a feature potentially delivered

Focus on iterative and incremental practices to help organizations deliver working software. The agile framework helps to maximise the project's success rate.

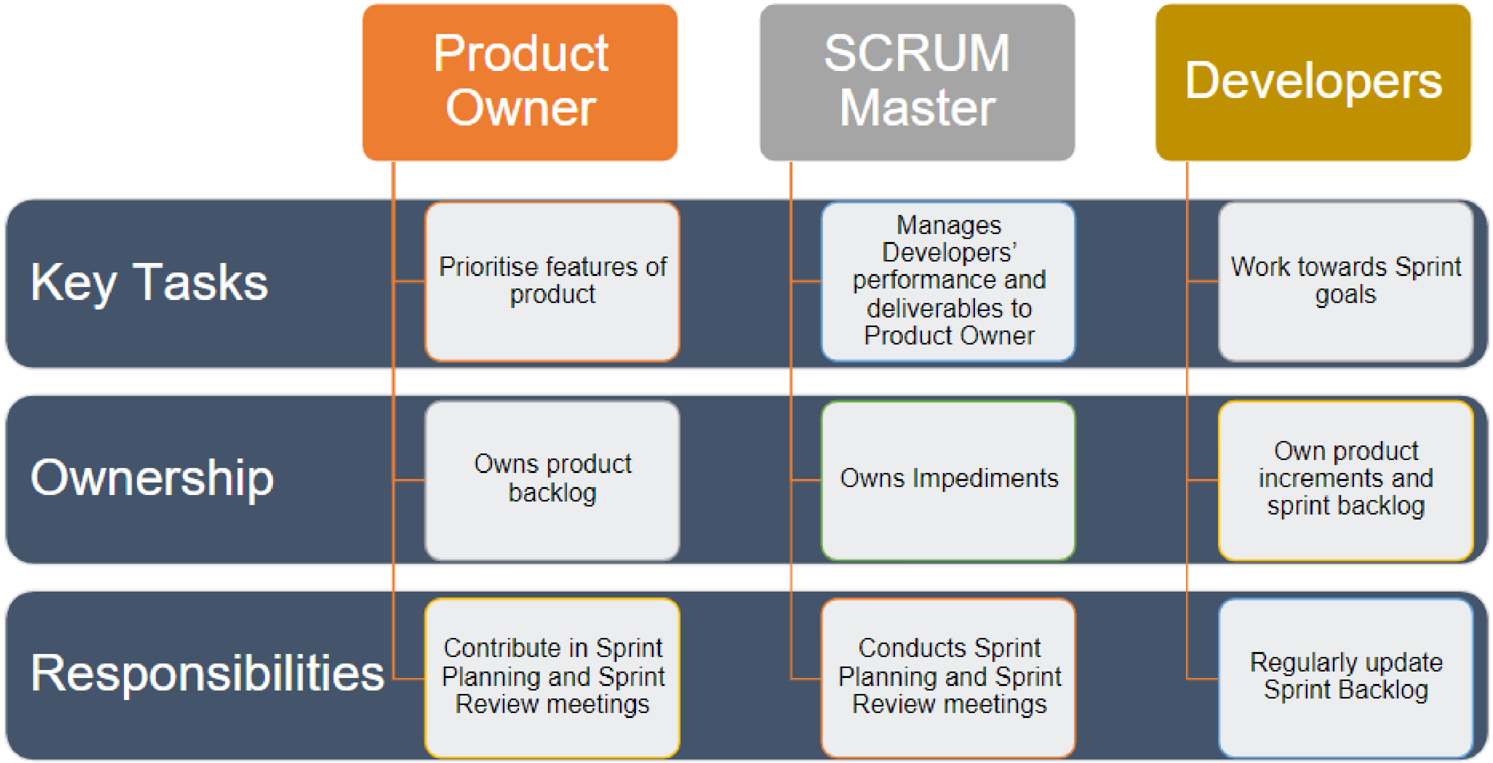
More details here: *https://www.scrum.org/index.php/resources/scrum-guide*

SCRUM is a commonly used agile framework for software development. It has a loose set of guidelines; they focus on iterative and incremental practices to help organizations deliver working software more frequently. This agile framework helps to maximise the project's success rate.

Sprints form the core foundation of SCRUM; each Sprint is a time box event typically lasting 3 weeks. At the end of each Sprint, a potentially releasable product or product's feature is created. This iterative approach with frequent feedback from business helps to ensure whatever software features developed to address the business problem.

Scrum.Org has published an official SCRUM guide; it contains all the key activities and principles of SCRUM. You could read through it if you wish to understand SCRUM better.

3. Key Members In A SCRUM Team

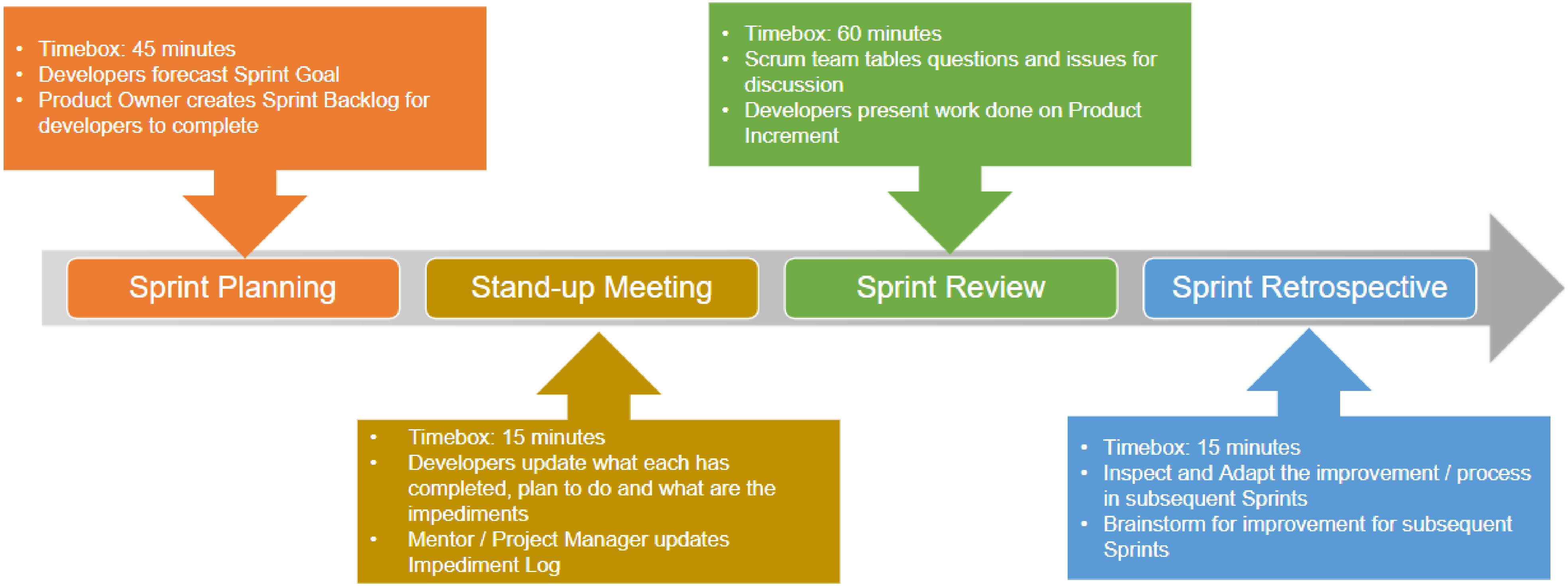


	Product Owner	SCRUM Master	Developers
Key Tasks	Prioritise features of product	Manages Developers' performance and deliverables to Product Owner	Work towards Sprint goals
Ownership	Owns product backlog	Owns Impediments	Own product increments and sprint backlog
Responsibilities	Contribute in Sprint Planning and Sprint Review meetings	Conducts Sprint Planning and Sprint Review meetings	Regularly update Sprint Backlog

The three main functions of a SCRUM team - product owner, SCRUM master, and developers - work together to ensure the successful delivery of the model.

The product owner usually comes from the business side; they have done their market research and understand what users want. Based on their business understanding, the product owner creates a list of 'to-do', or product backlog, to guide the work of the developers. The developers will prioritise their work based on this backlog and are expected to update their progress regularly during Sprint Reviews. SCRUM master helps facilitate meetings and remove any impediments to the developers' work. In other words, Product Owner gives direction, Developers built, and SCRUM Master facilitate.

3. Key Activities of SCRUM



Sprint Planning

- Timebox: 45 minutes
- Developers forecast Sprint Goal
- Product Owner creates Sprint Backlog for developers to complete

Stand-up Meeting

- Timebox: 15 minutes
- Developers update what each has completed, plan to do and what are the impediments
- Mentor / Project Manager updates Impediment Log

Sprint Review

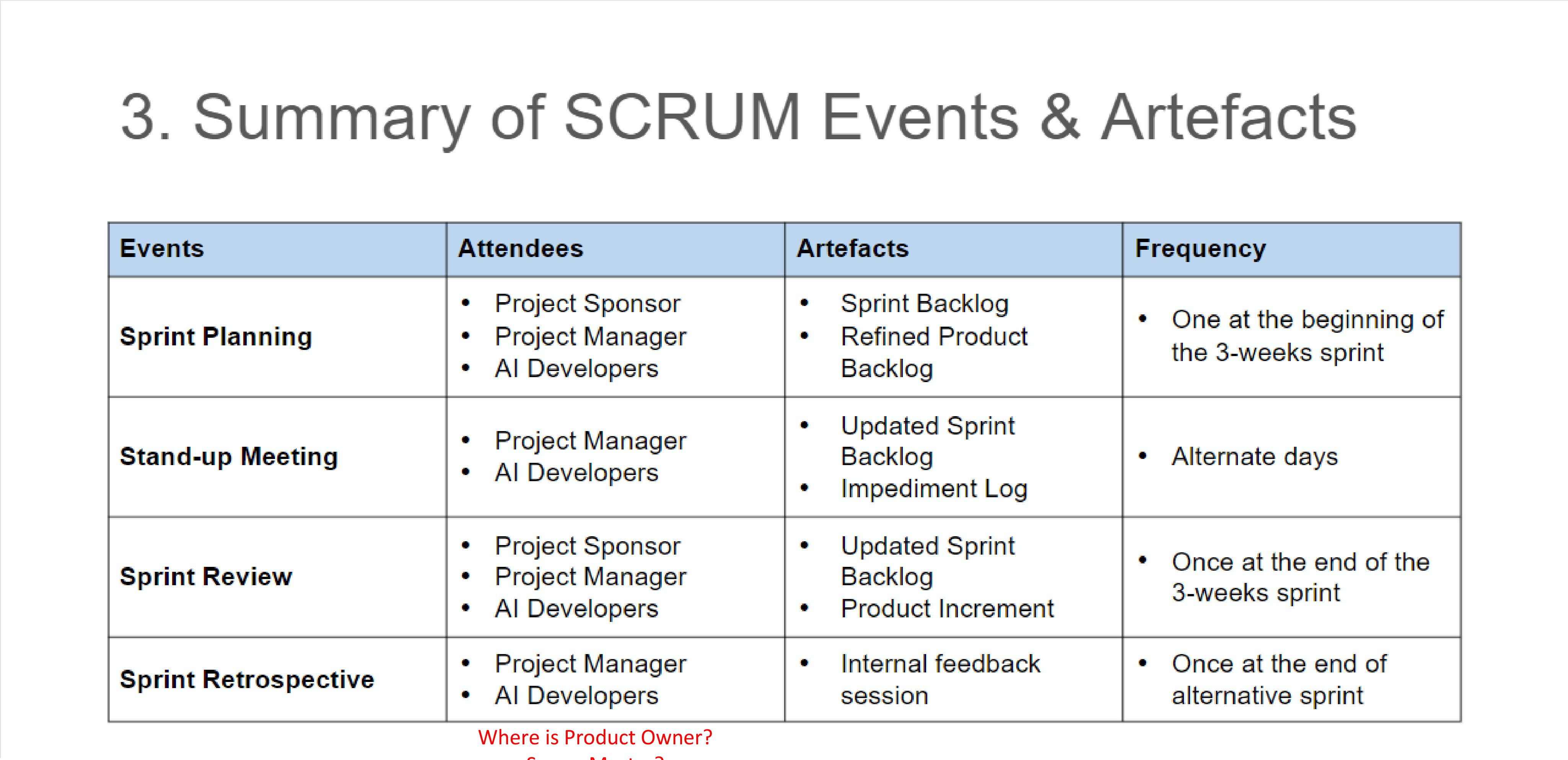
- Timebox: 60 minutes
- Scrum team tables questions and issues for discussion
- Developers present work done on Product Increment

Sprint Retrospective

- Timebox: 15 minutes
- Inspect and Adapt the improvement / process in subsequent Sprints
- Brainstorm for improvement for subsequent Sprints

There are four main activities of SCRUM: Sprint Planning, Stand-Up Meeting, Sprint Review, and Sprint Retrospective.

All activities are timeboxed so that team members will work together to concretely define ambiguous tasks. This also reinforce the agile principle of SCRUM: frequent short (time-boxed) updates to get users' feedbacks instead of infrequent long updates.



3. Summary of SCRUM Events & Artefacts

Events	Attendees	Artefacts	Frequency
Sprint Planning	• Project Sponsor • Project Manager • AI Developers	• Sprint Backlog • Refined Product Backlog	• One at the beginning of the 3-weeks sprint
Stand-up Meeting	• Project Manager • AI Developers	• Updated Sprint Backlog • Impediment Log	• Alternate days
Sprint Review	• Project Sponsor • Project Manager • AI Developers	• Updated Sprint Backlog • Product Increment	• Once at the end of the 3-weeks sprint
Sprint Retrospective	• Project Manager • AI Developers	• Internal feedback session	• Once at the end of alternative sprint

Where is Product Owner?
Scrum Master?

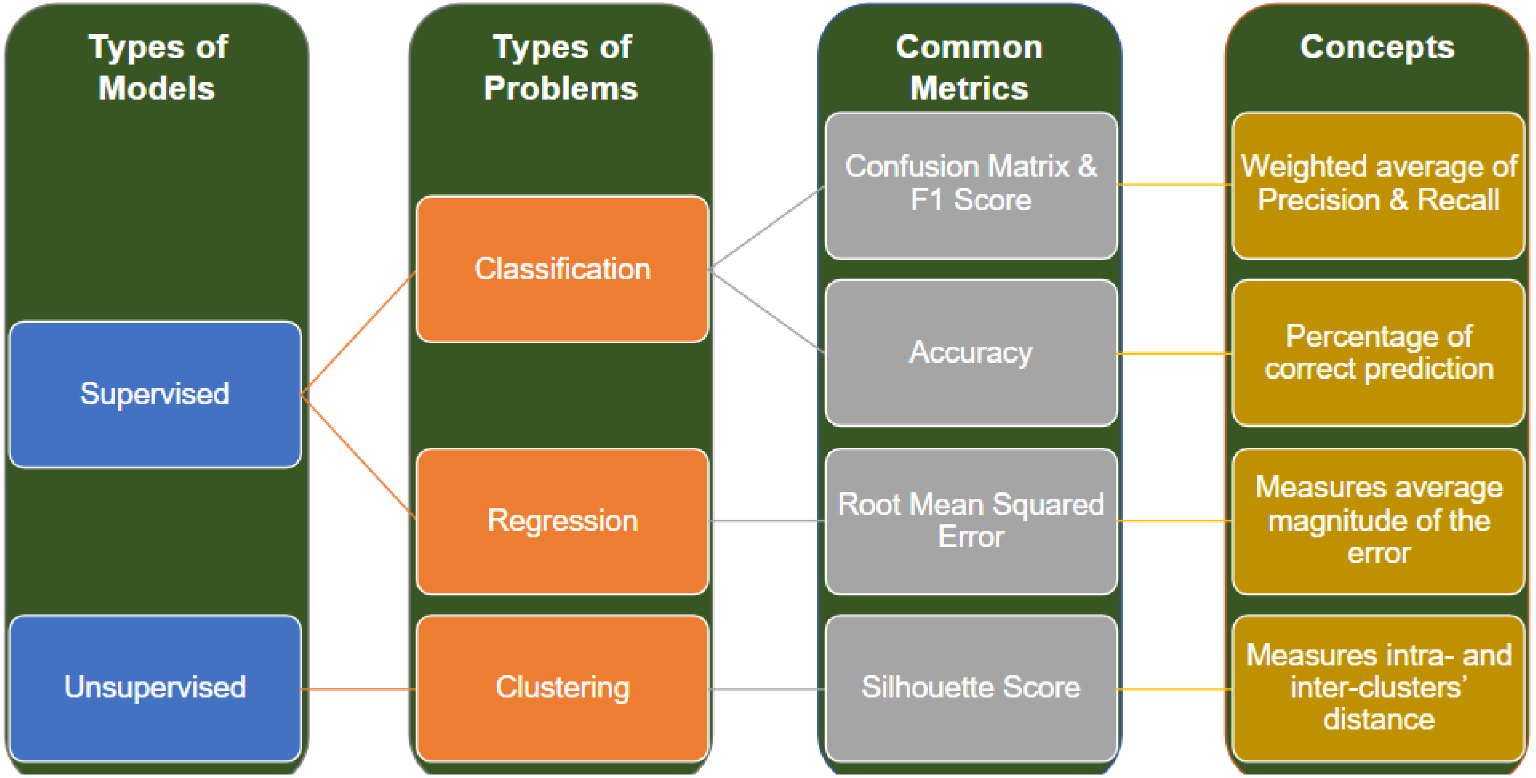
The table summarises the main events of a SCRUM.

Compared to traditional project management methodology, where meetings are held infrequently and meetings are for an update on project's progress, SCRUM focuses on frequent meetings (stand-up meeting is held on alternate days) and each meeting is expected to deliver an artefact (or an outcome). This frequent interaction with business users helps to ensure that the product that the engineering team is building meets the expectation of business users. There is no point delivering a project on time and within budget if the product does not solve the business problem.

Steps of Data Science Project Workflow

1. Who are the crucial team members in a data science project? What roles they play?
2. Has the wheel been invented? How others solve similar problems?
3. How to manage data science project with SCRUM methodology?
4. How to analyse model's performance and improve it?
5. What are the ways to integrate model into IT system?

4. Common Metrics for Model Validation



Types of Models

Supervised

Unsupervised

Types of Problems

Classification

Regression

Clustering

Common Metrics

Confusion Matrix & F1 Score

Accuracy

Root Mean Squared Error

Silhouette Score

Concepts

Weighted average of Precision & Recall

Percentage of correct prediction

Measures average magnitude of the error

Measures intra- and inter-clusters' distance

The appropriate metrics to use are dependent on the nature of the problem. The best practice should be to select the metrics before any calculations are ever made. In that way, we will not be tempted to select the model based on our biases or preferences and then justify our choice by selecting only the favourable metrics.

For classification problem (i.e. identifying whether the email is spam), both accuracy and F1 score could be used. F1 score is preferred over accuracy in situations where the dataset is imbalance (data or occurrence for a certain prediction label is rare).

For regression problem (i.e. predicting the price of a car), Root Mean Squared Error could be used. It measures the distance between predicted and actual value; bigger error is penalised more compared to smaller error.

For clustering problem (i.e. identifying the number of potential groups of data), Silhouette Score could be used. It measures both the 'tightness' of data in same the cluster and 'separation' of data in different clusters.

4. Metrics - Confusion Matrix, Precision, Recall, and F1 Score

Confusion Matrix

	Predicted False	**Predicted True**	**Total**
Actual: False	50 (True Negative)	10 (False Positive)	60
Actual: True	5 (False Negative)	100 (True Positive)	105
Total	55	110	165

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

= 0.91

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

= 0.95

$$F1\ Score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

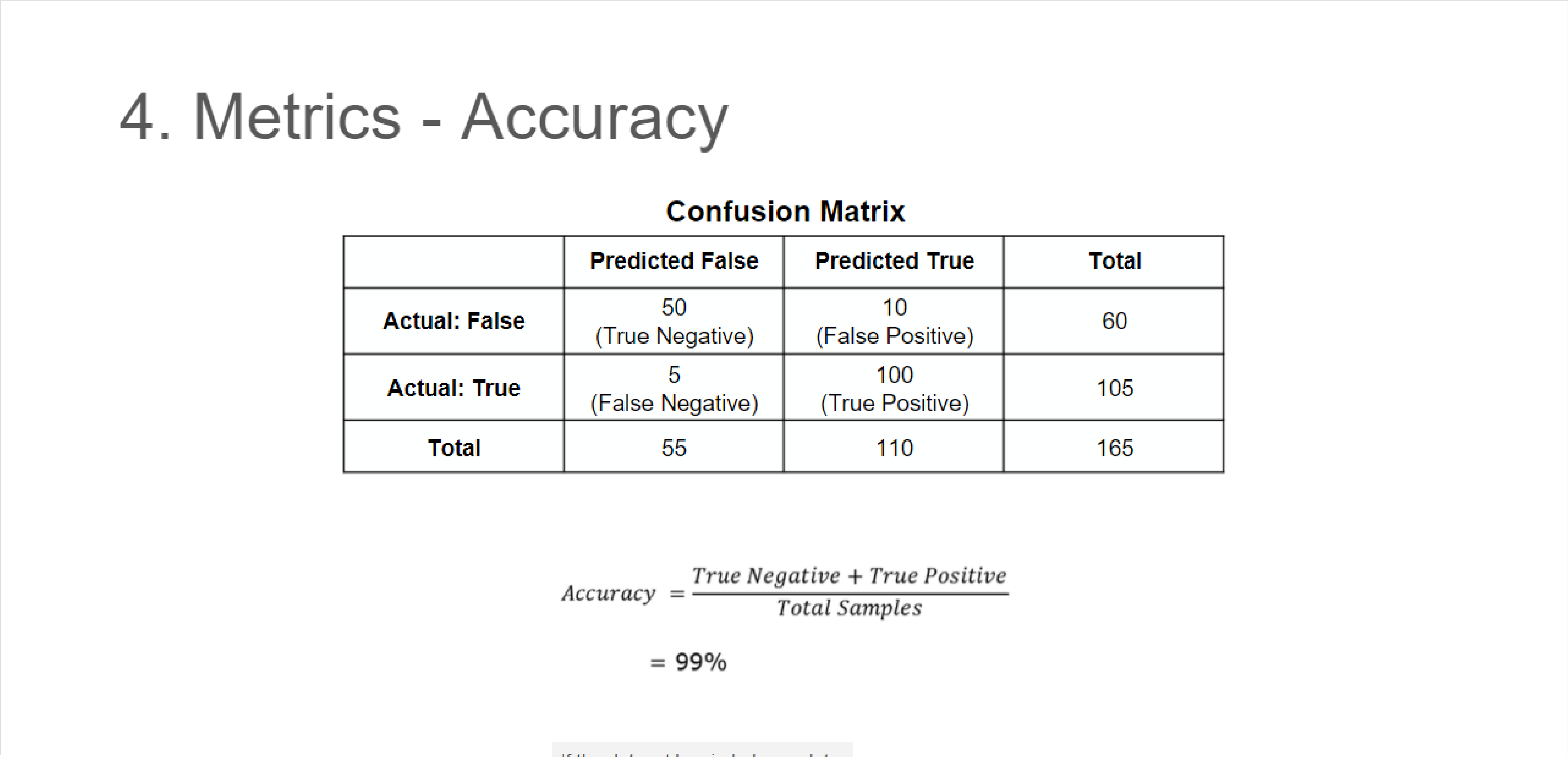
= 0.93

The confusion matrix shows the actual and predicted results in a single table. It highlights where the model is performing well or poorly. For instance, if there is high false positive, further investigation can be done to understand the root cause behind it. It could be due to 'dirty' data or insufficient data given to the model for certain classes of predictions. The ideal outcome is to have as many True Negative and True Positive as possible.

In addition to showing how well the model is performing in different classification category, the confusion matrix allows the calculation of Precision, Recall, and F1 Score.

Precision measures the model's ability not to label a negative sample as positive; a low precision score means there is a lot of false positive predicted by the model. Recall measures the ability of the model to find all positive samples; a low recall score means the model is unable to identify only relevant samples. F1 score is the weighted average of Precision and Recall. The higher the score for Precision, Recall, and F1, the better the model.

When comparing the performance of different models, f1 score is commonly used as it factors in both precision and recall.



4. Metrics - Accuracy

Confusion Matrix

	Predicted False	**Predicted True**	**Total**
Actual: False	50 (True Negative)	10 (False Positive)	60
Actual: True	5 (False Negative)	100 (True Positive)	105
Total	55	110	165

$$Accuracy = \frac{True\ Negative + True\ Positive}{Total\ Samples}$$

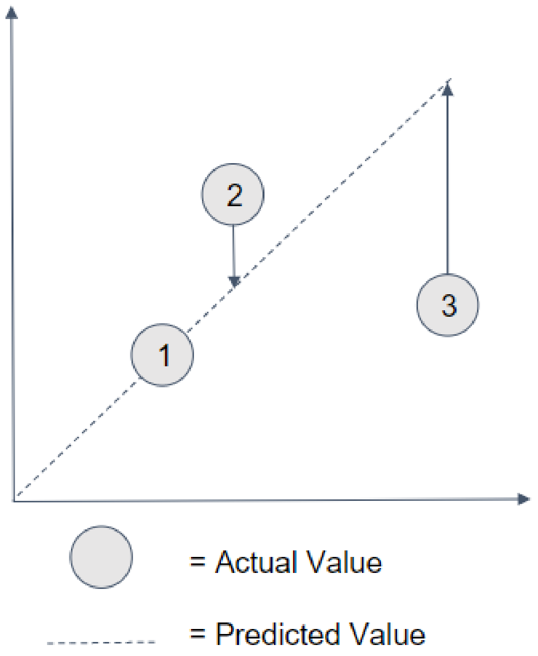
= 99%

The accuracy of the model shows how likely the model can predict the right classification. A model with higher accuracy is considered a better model. However, given a choice, it is generally better to measure the model's performance in terms of precision, recall, and f1 score. These metrics give further details about the model's performance compared to just accuracy.

If the dataset has imbalance data (i.e. one specific label constitutes only a minor percentage of the entire dataset), accuracy will not be a good metrics to use to measure the model performance. Why? Let's discuss this further using an example of a spam email filter. In the universe, let's assume that 99% of emails are not spam; only 1% of email is spam. Our model would have achieved an accuracy of 99% by simply labelling all emails as not spam!

4. Metrics - Root Mean Squared Error

- RMSE measures the difference between predicted and actual value.
- The bigger the difference, the higher the error.

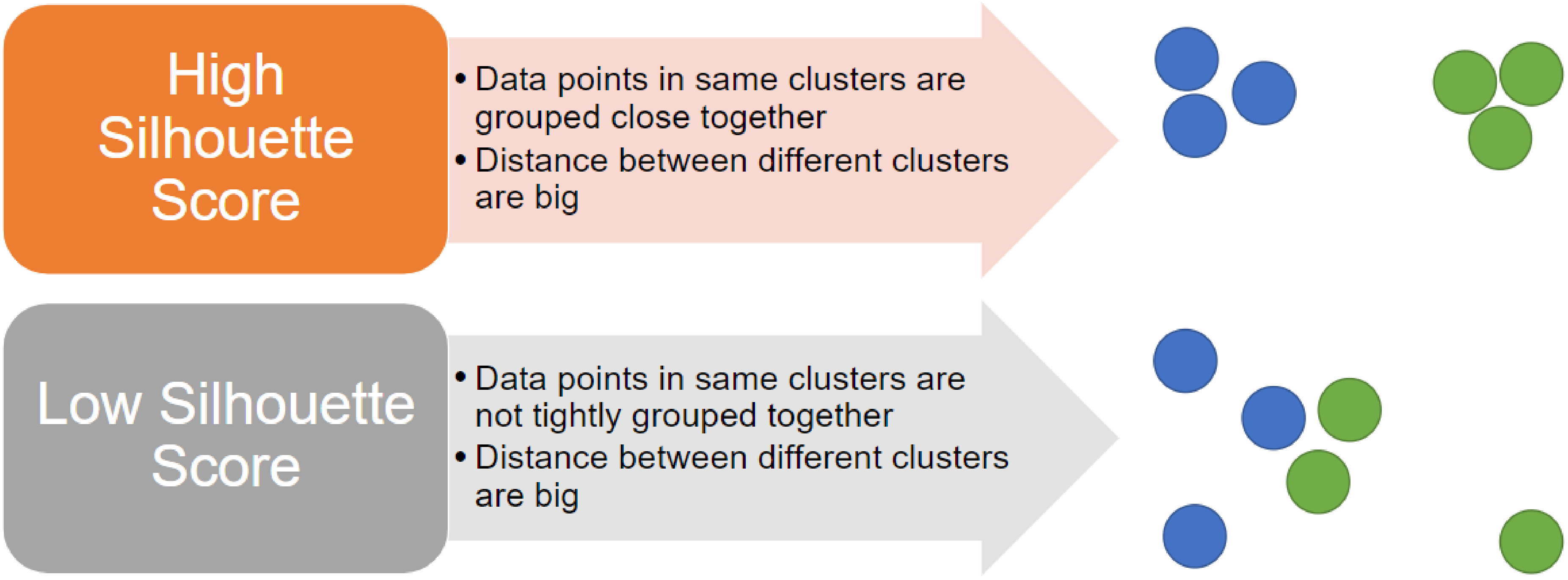


The following type of questions are solved using regression method:

- How much will the price be?
- What is the time of arrival?
- What is the height of someone based on his weight?

RMSE measures the difference between predicted and actual value. For instance, if the actual price is $2, RMSE will show the same magnitude of error if the model predicts $4 ($2 over actual) or $0 ($2 below actual). The difference in value matters more than the difference in the direction of value.

4. Metrics - Silhouette Score



High Silhouette Score

- Data points in same clusters are grouped close together
- Distance between different clusters are big

Low Silhouette Score

- Data points in same clusters are not tightly grouped together
- Distance between different clusters are big

The metrics used for unsupervised learning models typically measure some variations of two distances: distances between data points in the same clusters and distances between data points in different clusters. Different metrics for unsupervised learning take different forms of these two distances to come up with a score.

Higher scores are given when data points in same clusters are close together (tightness of data in same the cluster) and when data points from different clusters are far from each other (separation of data points from other clusters).

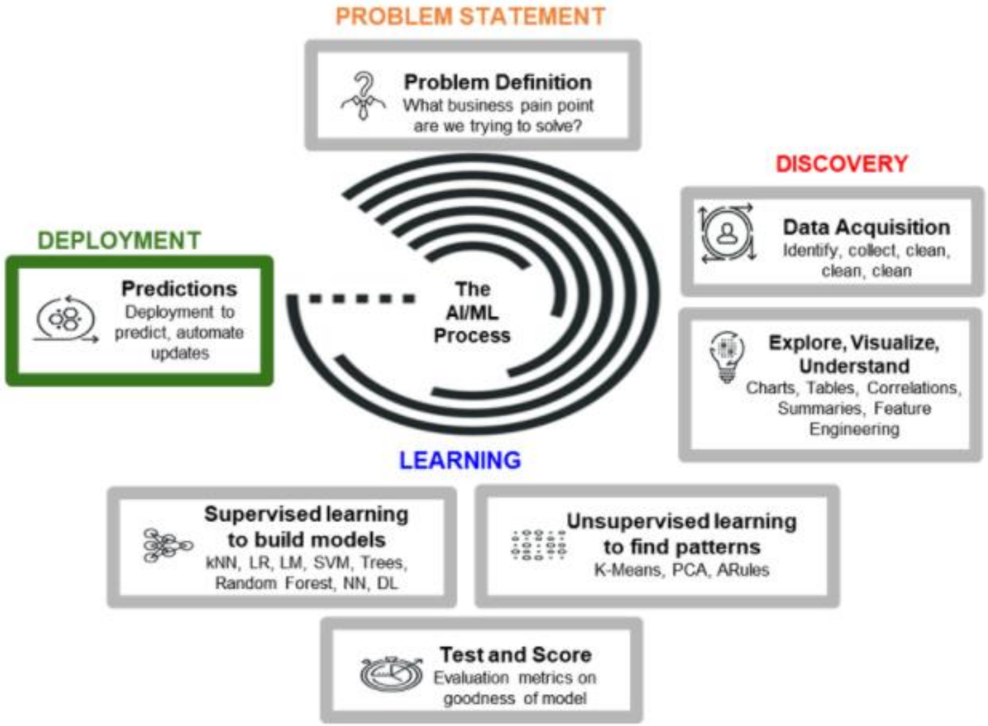
Steps of Data Science Project Workflow

1. Who are the crucial team members in a data science project? What roles they play?
2. Has the wheel been invented? How others solve similar problems?
3. How to manage data science project with SCRUM methodology?
4. How to analyse model's performance and improve it?
5. What are the ways to integrate model into IT system?

AI SINGAPORE

5. Deployment of Model

- Deployment is the last stage of machine learning workflow.
- Deployment integrates the ML model into IT system.



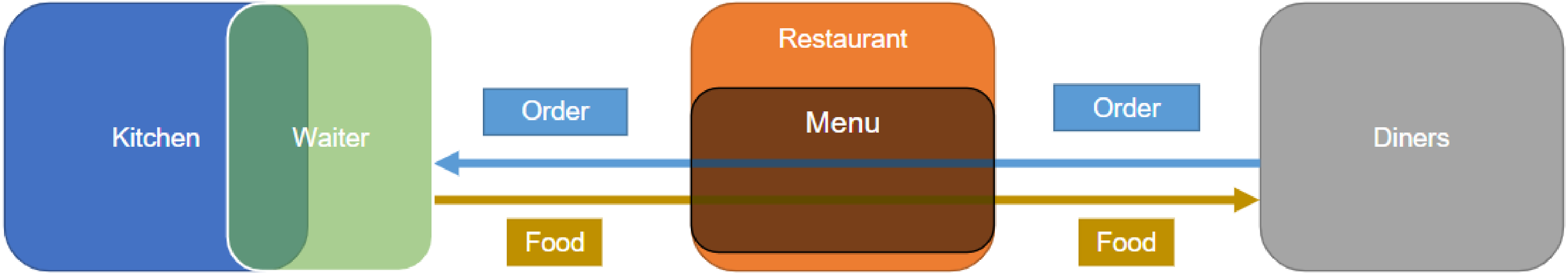
PROBLEM STATEMENT
Problem Definition
What business pain point are we trying to solve?
DISCOVERY
Data Acquisition
Identify, collect, clean, clean, clean
Explore, Visualize, Understand
Charts, Tables, Correlations, Summaries, Feature Engineering
DEPLOYMENT
Predictions
Deployment to predict, automate updates
The AI/ML Process
LEARNING
Supervised learning to build models
kNN, LR, LM, SVM, Trees, Random Forest, NN, DL
Unsupervised learning to find patterns
K-Means, PCA, ARules
Test and Score
Evaluation metrics on goodness of model

Deployment is the last stage of building an ML/AI model. A trained model residing on your local computer can only be accessible by you; this limits its accessibility. Therefore, after the model is built, it is necessary to deploy it to make it accessible to others.

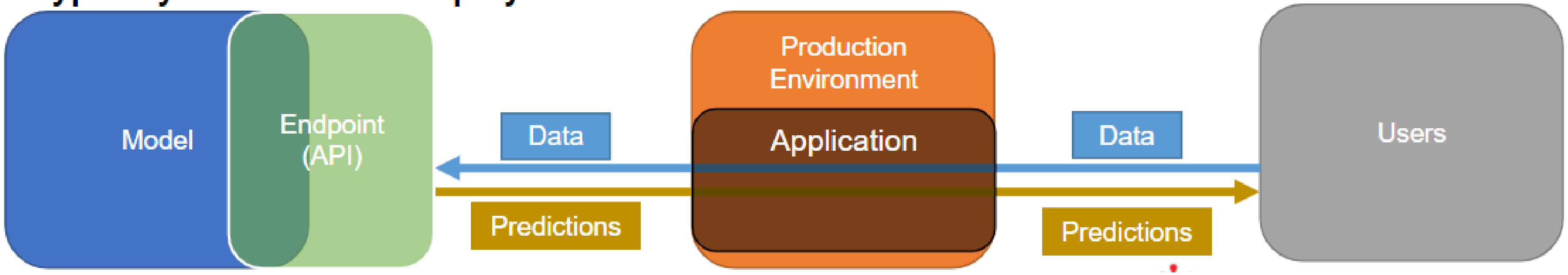
Before reaching this stage, the model needs to undergo rigorous checks and tests to ensure major bugs (or errors in codes) are corrected. In an organisation, the trained model is typically passed to a DevOps or a Software Engineer for deployment.

5. Data Flow of A Deployed Model

Analogy of Model Deployment: Dining at a restaurant



Typically data flow of a deployed model



A model is typically deployed by making an endpoint or API available. This endpoint will be responsible for the data communication between the model and software application that access the model. Essentially, the endpoint serves as the interface between a model and an application or software.

An analogy of model deployment is like going for dinner. The diners (users) will look at the menu (application) and pass the order (data) to the waiter. The waiter (endpoint) will take the order to the kitchen (model) to have your food (prediction) prepared. Once done, the waiter will serve you your food (prediction).

5. Platforms for Model Deployment

Deploy via REST API

Flask

Deploy via Commercial Cloud Providers

Microsoft Azure

Google Cloud

amazon web services

Deploying on Microsoft Azure: https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-deploy-and-where

AI SINGAPORE

The complexity of model deployment depends on the production environment (IT system) and the model's complexity (the type and amount of data it requires to generate the prediction as well as the frequency of the prediction output).

Deployment is commonly done by converting the Python model into an intermediate standard format, such as Open Neural Network Exchange (ONNX). This is then converted into a software compatible with the production environment. You can read up more about ONNX via this link (https://onnx.ai/).

The exact steps for model deployment vary based on the method of deployment. The two main categories of model deployment are via Flask on or on commercial Cloud:

- To learn more about model deployment using Flask, check this article (https://www.datacamp.com/community/tutorials/machine-learning-models-api-python).
- For model deployment via a commercial cloud, you have to check the respective cloud provider's documentation. Each commercial cloud provider requires different steps to deploy your model. As an example, you could check Microsoft Azure's documentation on model deployment.

AI SINGAPORE

Thank you

www.aisingapore.org

Course completed!

Happy learning!