



COS10022 – DATA SCIENCE PRINCIPLES

Dr Pei-Wei Tsai (Lecturer, Unit Convenor)
ptsai@swin.edu.au, EN508d

SWIN
BUR
NE

SWINBURNE
UNIVERSITY OF
TECHNOLOGY

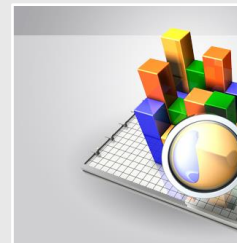


Week 07



**Problem
Transformation**

**Data Analytics
Lifecycle**



**Data Store,
Ethics, and
Security**



PROBLEM TRANSFORMATION

Using Proper Techniques to Solve the Problems

The Problem to Solve	The Category of Techniques
I want to group items by similarity. I want to find structure (commonalities) in the data	Clustering
I want to discover relationships between actions or items	Association Rules
I want to determine the relationship between the outcome and the input variables	Regression
I want to assign (known) labels to objects	Classification
I want to find the structure in a temporal process I want to forecast the behavior of a temporal process	Time Series Analysis
I want to analyze my text data	Text Analysis

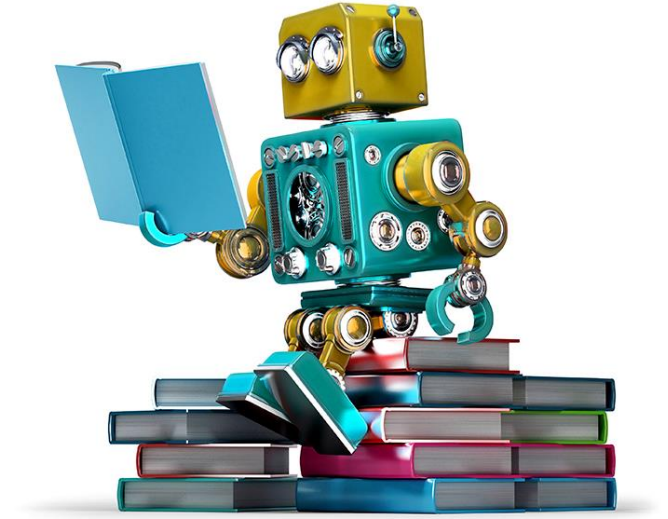


Key Questions

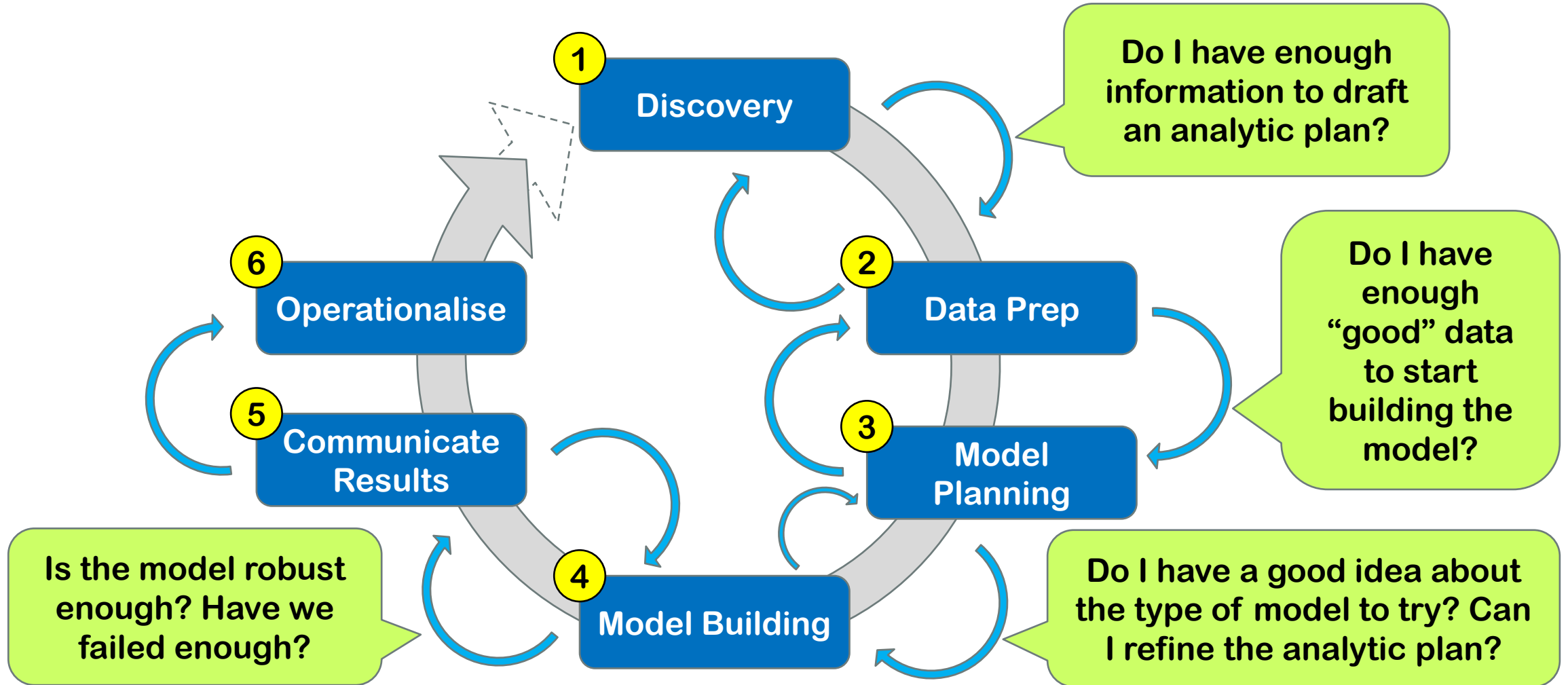
1. **What is Data Analytics Lifecycle?**
2. **Why do we need the Data Analytics Lifecycle?**
3. **What are the components in the Data Analytics Lifecycle?**
4. **For each phase in the Data Analytics Lifecycle:**
 - 1) **What is the phase about?**
 - 2) **What key activities are included in each phase?**
 - 3) **What resource and tools can I use?**
 - 4) **Examples**
5. **What are the key outputs of a successful data science project?**

Learning Outcomes

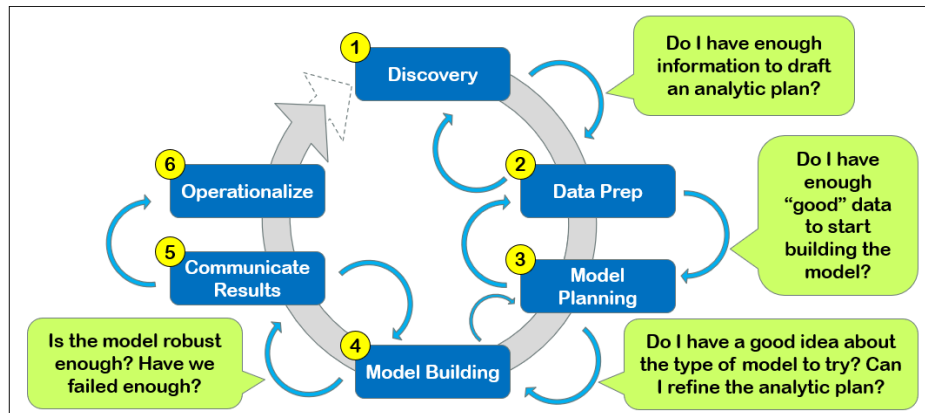
- This lecture supports the achievement of the following learning outcomes:
- 1. Appreciate the roles of data science and Big Data analytics in organisational contexts.
- 2. **Compare and analyse the key concepts, techniques and tools for discovering, analysing, visualising and presenting data.**
- 3. **Describe the processes within the Data Analytics Lifecycle.**
- 4. Analyse organisational problems and formulate them into data science tasks.
- 5. Evaluate suitable techniques and tools for specific data science tasks.
- 6. Develop and execute an analytics plan for a given case study.



Data Analytics Lifecycle



Data Analytics Lifecycle



- Data Analytics Lifecycle defines analytics process **best practices** that span discovery to project completion.
- A data science framework advocated by DELL EMC.
- Provides guidance to carry out a data science project.
- Comprises of six different phases.



Data Analytics Lifecycle

- Emphasises the following principles of data science best practices:
 - Data Science project is iterative.
 - It is possible to move *forward* or *backward* between most phases in the lifecycle.
 - A project work can occur in several phases *at once*.
 - The best gauge of advancing to the next phase is to ask key questions to test whether the data science team has accomplished enough to move forward.
 - Ensure teams do the appropriate work both up front, and at the end of the projects, in order to succeed. Too often teams focus on Phases 2 to 4, and want to jump into modelling work before they are ready.

Engage #1: The Genesis of EMC's Data Analytics Lifecycle

- Read David Dietrich's [blog](#).
- What data sources did the author rely on in formulating the Data Analytics Lifecycle?
- Identify five existing analytics approaches/methods that contributed to various phases in Data Analytics Lifecycle.

The screenshot shows a Dell Technologies blog page. The header includes the Dell Technologies logo, social media icons, a 'SUBSCRIBE TO INFOCUS' button, and a search icon. The navigation bar lists categories: AI/IOT/ANALYTICS, APPLICATIONS/DEVOPS, DELL IT, LEARNING, MULTI-CLOUD, SERVICE EXCELLENCE, and WORKFORCE. The article title is 'The Genesis of EMC's Data Analytics Lifecycle' by David Dietrich, dated November 1, 2013. The article text begins with 'When I developed a new Data Analytics Lifecycle for EMC's Data Science & Big Data Analytics course in 2011, I had no idea the attention it would receive. Although I have been doing analytical work for most of my career, I needed to do considerable research to create a solid process for others to follow. After some preliminary research, I realized that there were surprisingly few existing frameworks for conducting data analytics. The best sources that I came across were these:'. A list of sources follows, including CRISP-DM, Tom Davenport's DELTA framework, 'MAD Skills: New Analysis Practices for Big Data', Doug Hubbard's Applied Information Economics (AIE) approach, and The Scientific Method. The page also features 'Recent Comments' and 'Recent Posts' sections on the right side.

Dell Technologies

AI/IOT/ANALYTICS APPLICATIONS/DEVOPS DELL IT LEARNING MULTI-CLOUD SERVICE EXCELLENCE WORKFORCE

AI/IOT/ANALYTICS

The Genesis of EMC's Data Analytics Lifecycle

By David Dietrich
November 1, 2013

When I developed a new Data Analytics Lifecycle for EMC's Data Science & Big Data Analytics course in 2011, I had no idea the attention it would receive. Although I have been doing analytical work for most of my career, I needed to do considerable research to create a solid process for others to follow. After some preliminary research, I realized that there were surprisingly few existing frameworks for conducting data analytics.

The best sources that I came across were these:

- **CRISP-DM**, which provides useful inputs on ways to frame analytics problems and is probably the most popular approach for data mining that I found.
- Tom Davenport's **DELTA** framework from his text "Analytics at Work."
- "MAD Skills: New Analysis Practices for Big Data" provided inputs for several of the techniques mentioned in Phases three to five of my Data Analytics Lifecycle which focus on model planning, execution, and key findings.
- Doug Hubbard's **Applied Information Economics (AIE)** approach from his work "How to Measure Anything." The focus of this work differs a bit from a classic data mining approach. Hubbard's approach emphasizes estimating and measuring for the purpose of making better decisions. It has some very useful ideas, and helps one understand how to approach analytics challenges from a unique angle and treat them more like decision science problems.
- **The Scientific Method**. Although it has been in use for centuries, it still provides a solid framework for

Recent Comments

- Bob Feiner on 2B Pounds of Electronics Recovered--And We're Just Getting Started!
- J Mark Scott-CTO/Partner on 2B Pounds of Electronics Recovered--And We're Just Getting Started!
- michaelcorsole04 on 10 Considerations to Optimize Microsoft Dynamics 365 Worker Experiences (Infographic)
- Derek on How Containers Are Making Way for the 5G and Edge-centric World -- Part 1
- Daniel on How to Modernize Your PC Management Approach

Recent Posts

- > The Talent Gap: Building a Bridge to the Other Side
- > 2B Pounds of Electronics Recovered--And We're Just Getting Started!
- > The Undiscovered Country: A Trekkie's View of the Future of Networking
- > 10 Most Read InFocus Blogs of 2019: Laying the Foundation for What's Possible in the Next Decade!
- > How Containers Are Making Way for the

Why Do We Need the Data Analytics Lifecycle?

- The real world is chaotic, and so are its problems and data.
- Often, being a data scientist in an organisation is also a chaotic experience.
- You may (quite easily) feel lost.
- Accumulation of data science “*know how*” and best practices from industries and academia.
- Provides order and sense-making.
- Provides measures of success.



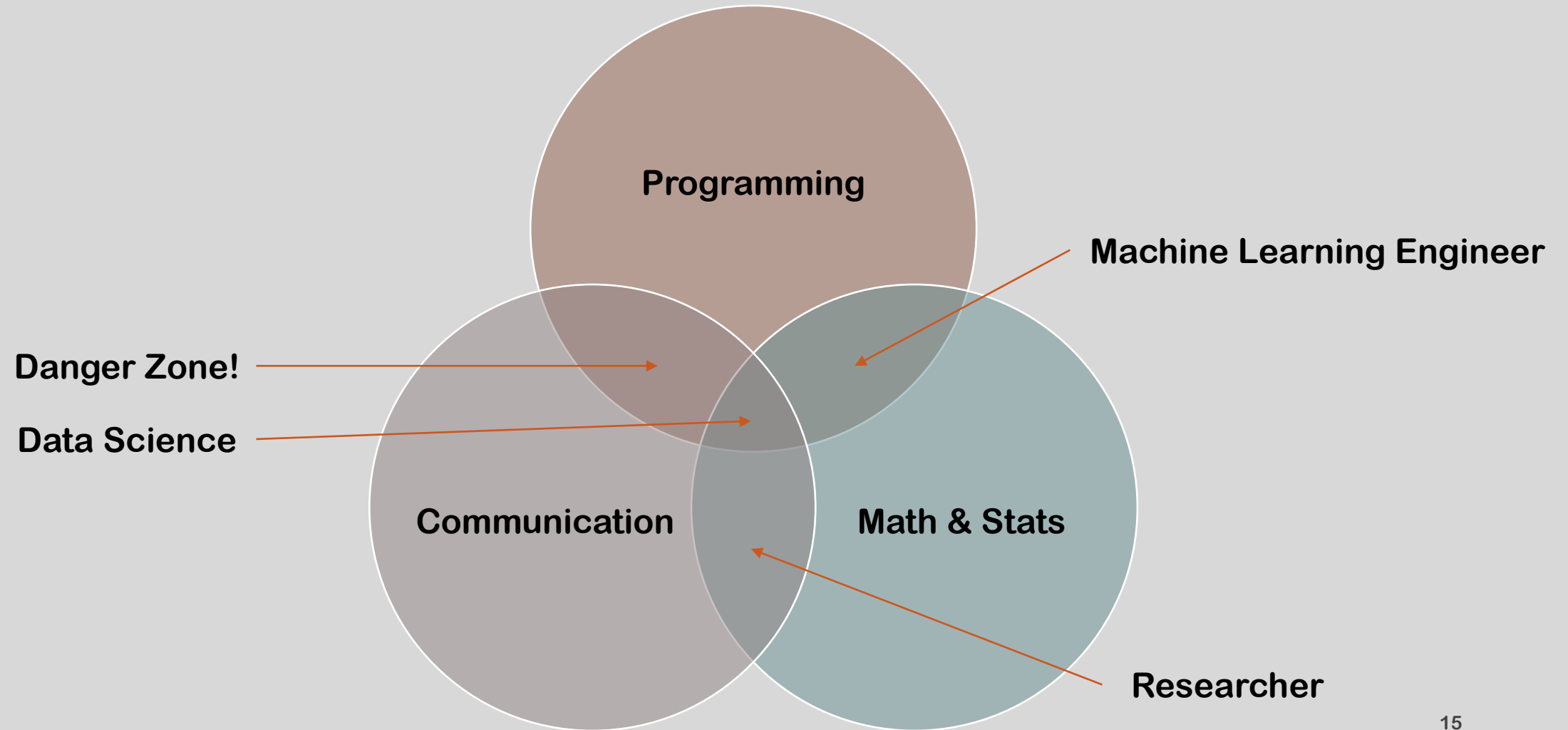
Engage #2: Introducing the Innovation Analytics Case Study at EMC Corp.

- Read this page on Steve Todd's blog:
- https://stevetodd.typepad.com/my_weblog/2012/03/a-strategy-for-innovation-analytics.html
- We will return to this case study throughout this lecture to help illustrate how each phase in the Data Analytic Lifecycle can be actually implemented in a real industrial setting.
- Steve Todd is currently a Fellow at Dell Technologies. He was previously the Distinguished Engineer and Director of Global Research and Innovation, as well as Corporate VP of Strategy and Innovation at EMC Corporation. More about Steve: <https://www.linkedin.com/in/stevejtodd/>



RELATED INFORMATION

Data Science Venn Diagram



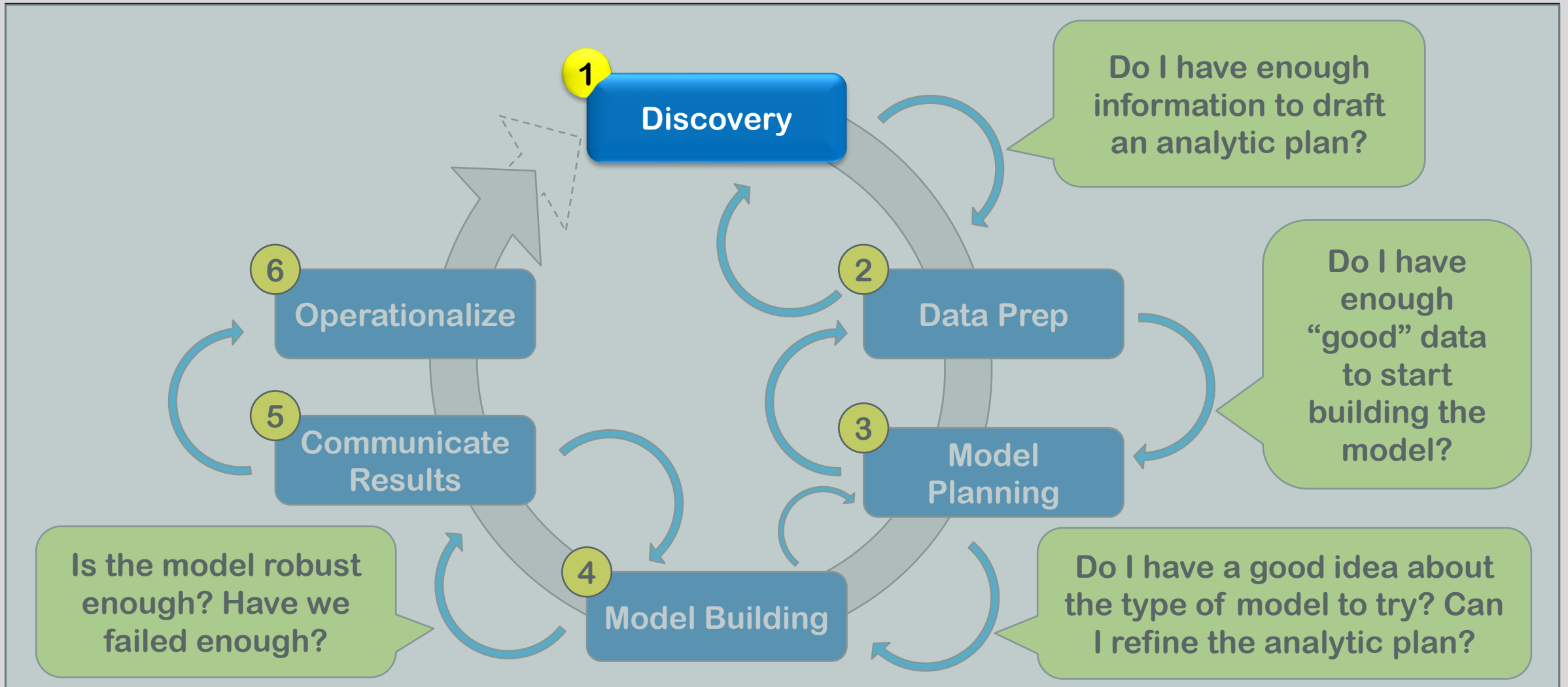
Let the Data Speaks for You but not the Human Bias.

- Simply put, HiPPO (Highest Paid Person's Opinion) effect states that the authority figure's suggestions are interpreted as the final truth, and promptly implemented, even if the findings from the data are contrary.

HiPPO Effect



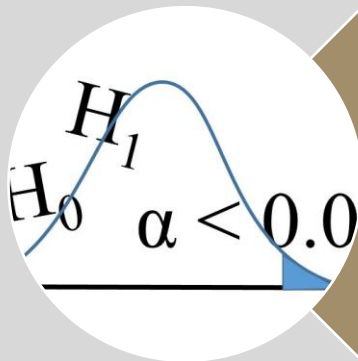
Data Analytics Lifecycle – Phase 1



Phase 1 - Discovery

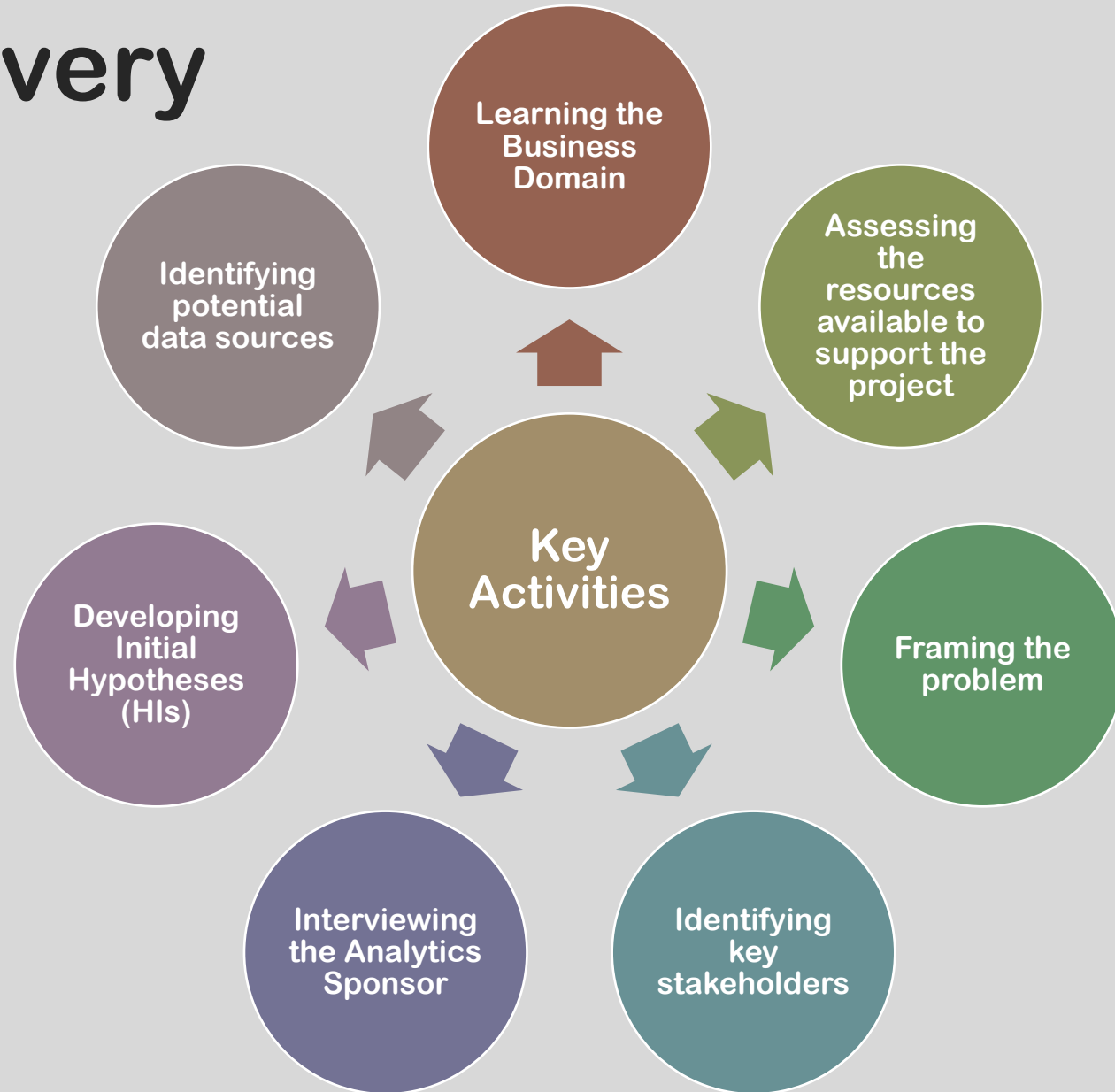


Data science team learns the business domain, and assesses the resources available to support the project in terms of people, technology, time, and data.



Important activities include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

Phase 1 - Discovery



Phase 1 - Discovery

1. Learning the Business Domain

- **Key Activities**

Determine how much business or domain knowledge that a data scientist needs in order to develop models in Phase 3 and 4.

To decide the resources needed for the project team;

To ensure that the team has the right balance of domain knowledge and technical expertise.

Phase 1 - Discovery

2. Assessing the resources available to support the project

- **Key Activities**

Consider the available tools and technology the team will be using and the types of systems needed for later phases. Resources include technology, tools, systems, data, and people.

Take inventory of the types of data available to the team for the project.

Ensure the data science team has the right mix of domain experts, customers, analytic talent, and project management.

Phase 1 - Discovery

- Key Activities

3. Framing the problem

Framing is the process of stating the analytics problem to be solved.

Establish *failure criteria*.

Write down the problem statement and share it with the key stakeholders.

Identify the main objectives of the project, identify what needs to be achieved in business terms, and identify what needs to be done to meet the needs.

Phase 1 - Discovery

4. Identifying key stakeholders

- **Key Activities**

Stakeholders include anyone who will benefit from the project, or will be significantly impacted by the project.

Articulate the “pain points” to be addressed.

Outline the type of activity and participation expected from each stakeholder.

Phase 1 - Discovery

5. Interviewing the Analytics Sponsor

◦ Key Activities

What is the desired outcome of the project?

What business problem is the team trying to solve?

What data sources are available?

What industry issues may impact the analysis?

What timelines need to be considered?

Who could provide insight into the project?

Who has the final decision-making authority on the project?

How will the focus and scope of the problem change if time, people, risk, resources, or the size and attributes of the data change?

Phase 1 - Discovery

6. Developing Initial Hypotheses (HIs)

- **Key Activities**

Form ideas that can be tested with data. Start with just a few primary hypotheses/ideas, then develop several more.

Compare their answers with the outcome of an experiment.

Gather and assess hypotheses from stakeholders and domain experts.

Phase 1 - Discovery

7. Identifying potential data sources

Identify data
sources

Capture aggregate
data sources

Review the raw
data

Evaluate the data
structures and
tools needed

- **Key Activities**

Scope the sort of
data infrastructure
needed for this
type of problem

Engage #2-1: Introducing the Innovation Analytics Case Study at EMC Corp.

Explore an example of Phase 1 in the Innovation Analytics project at EMC:

http://stevetodd.typepad.com/my_weblog/2012/03/phase-1-innovation-analytics.html

Engage #2-1: Introducing the Innovation Analytics Case Study at EMC Corp.

- Key Activities

Frame the business problem as an analytic challenge that can be solved in phases.

Understand what's been done in the past.

Assess the resources supporting the project (people, technology, time, and data).

Form initial hypotheses.

Determine readiness to move to the next phase.

Moving from Phase 1 to Phase 2

- Data Analytics Lifecycle is intended to accommodate ambiguity. This reflects most real-life situations.

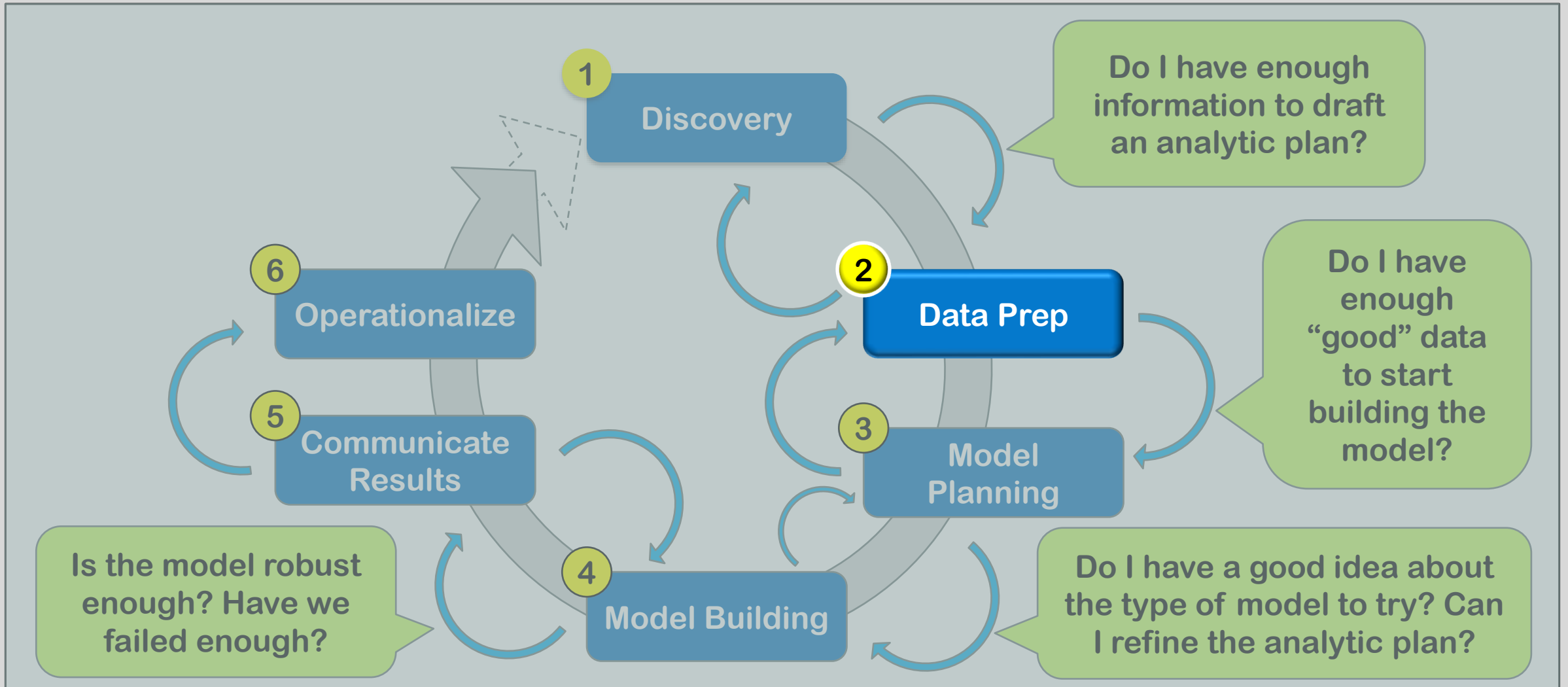


The data science team can move to the next phase if it has enough information to draft an *analytics plan* and share it for peer review.



Do I have enough information to draft an analytics plan and share for peer review?

Data Analytics Lifecycle – Phase 2



Phase 2 – Data Preparation



Phase 2 requires the presence of an analytics sandbox, in which the data science team work with data and perform analytics for the duration of the project.

The team performs ETLT to get the data into the sandbox, and familiarize themselves with the data thoroughly.



ETL + ELT = ETLT

(Extraction, Transform, and Load)

When to use ETL or ELT?



Order_ID	Price	Date
30017	AUD38.99	Apr 15, 2020
30429	AUD92.03	12/12/2020



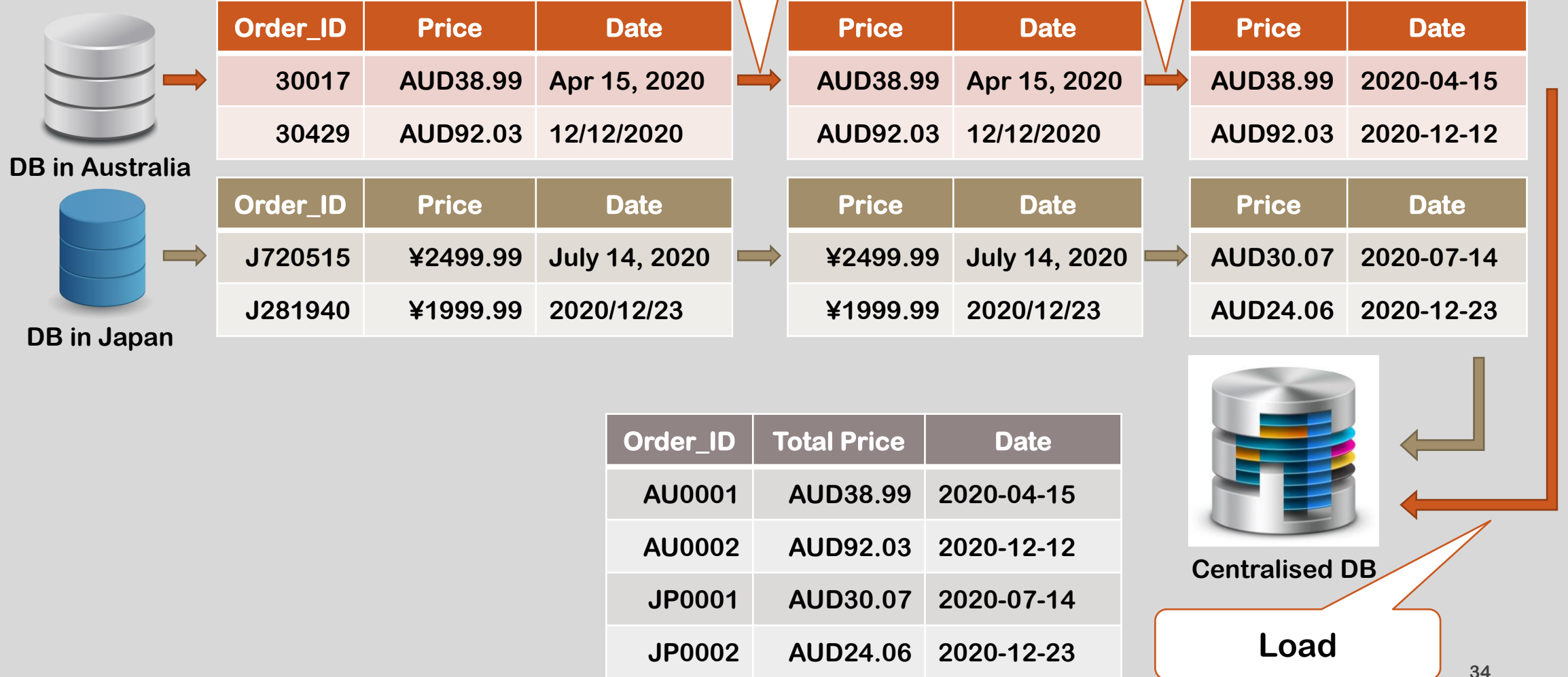
Order_ID	Price	Date
J720515	¥2499.99	July 14, 2020
J281940	¥1999.99	2020/12/23



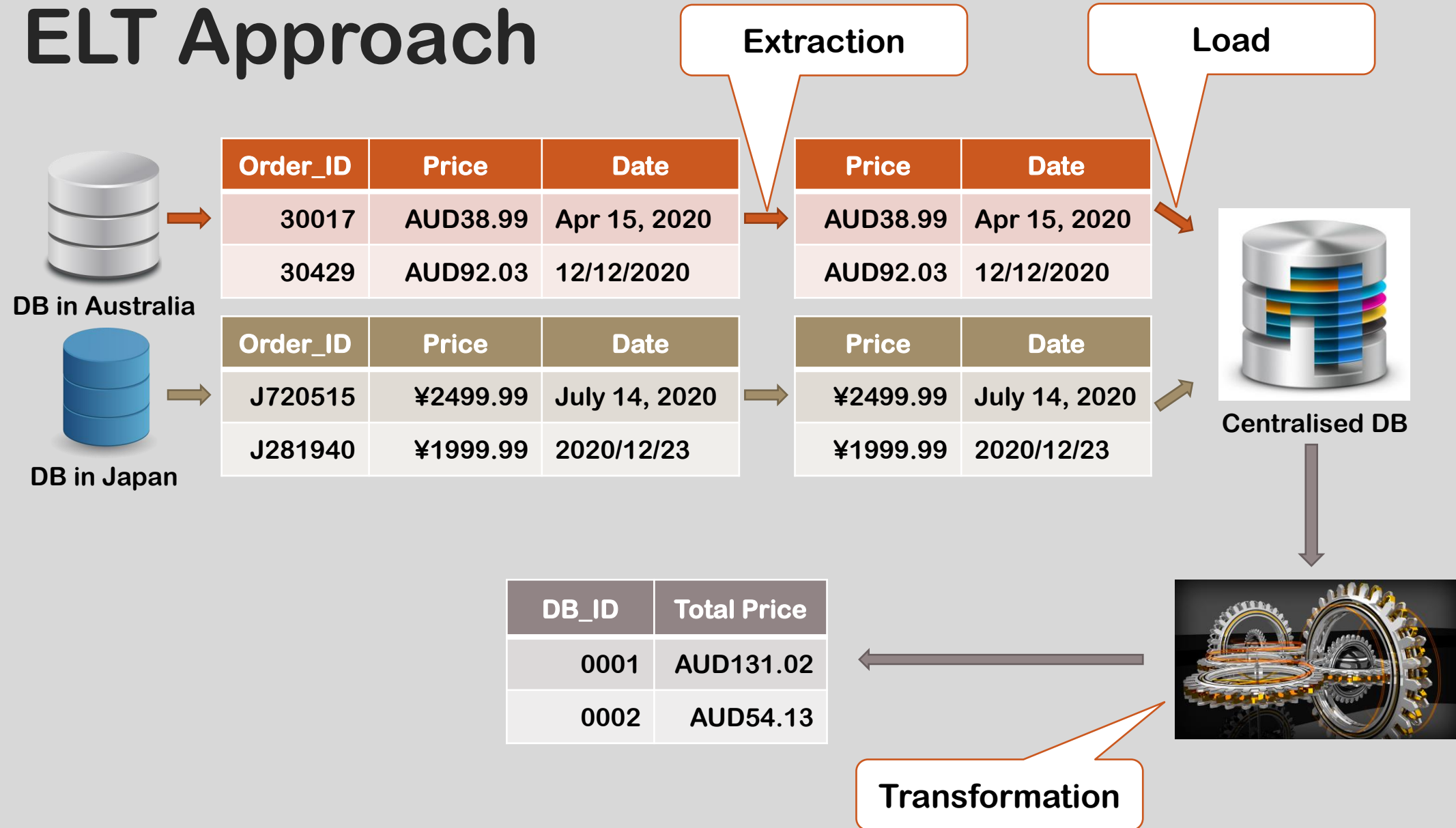
ETL Approach

Extraction

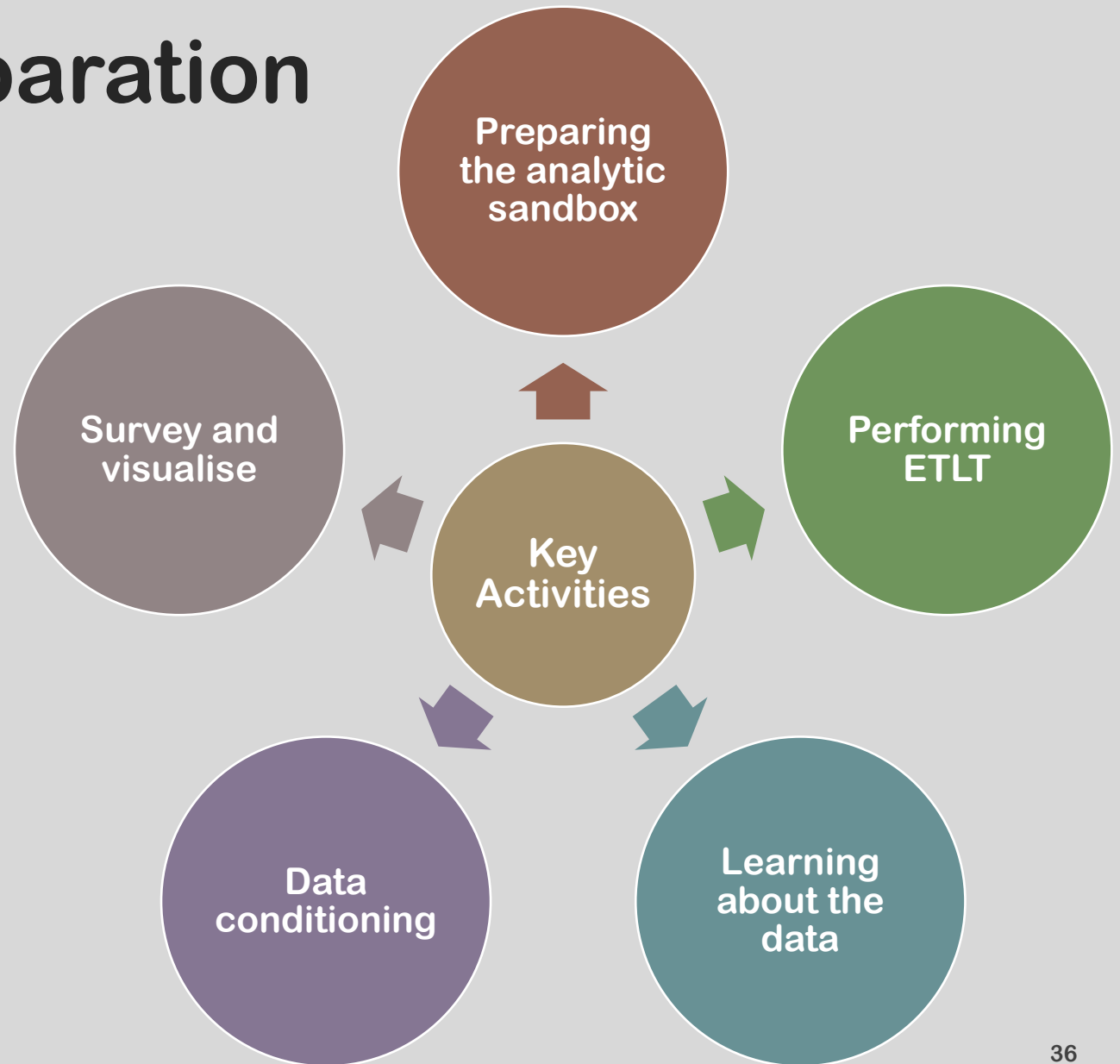
Transformation



ELT Approach



Phase 2 – Data Preparation



Phase 2 - Data Preparation

1. Preparing the analytic sandbox

- Key Activities

An analytic sandbox (or, *workspace*) allows the data science team to explore the data without interfering with live production databases.

Expect the sandbox to be **LARGE** (5 to 10 times greater than the original datasets).

Collect all kinds of data into the sandbox, ranging from summary-level aggregated data, structured data, raw data feeds, and unstructured text data from call logs/web logs.

IT group may require justification to develop an analytic sandbox.

Phase 2 - Data Preparation

- **Key Activities**

2. Performing ETLT

- ETLT is a combined approach where the team may choose to perform either ETL or ELT when populating the sandbox.
- This choice depends on the team's specific goals.

**Perform
data gap
analysis.**

Consider how to parallelise the movement of big datasets into the sandbox (Big ETL), e.g., Hadoop, MapReduce, Twitter API.

Prior to data movement, determine the kinds of transformation to be performed on the data.

Phase 2 - Data Preparation

3. Learning about the data

- Key Activities

Understand the acceptable range of values, expected output, and data entry errors.

Identify additional data sources that the team can leverage but currently unavailable.

It is advisable to build a *dataset inventory*.

A Sample of Dataset Inventory

Dataset	Data Available and Accessible	Data Available, but not Accessible	Data to Collect	Data to Obtain from Third Party Sources
Products shipped	●			
Product Financials		●		
Product Call Centre Data		●		
Live Product Feedback Surveys			●	
Product Sentiment from Social Media				●

- Source: Dietrich, D. ed., 2015. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualising and Presenting Data*. EMC Education Services.

Phase 2 - Data Preparation

- Key Activities

4. Data conditioning

The process of cleaning data, normalising datasets, and performing transformations on the data.

Join or merge different datasets to be ready for analyses.

Decide which aspects of datasets will be useful to analyse in later steps, i.e., which one to keep or discard.

Phase 2 - 4. Data Conditioning

Questions to ask

- What are the data sources?
- What are the target fields, columns, and attributes?
- How clean is the data?
- How consistent are the contents and fields

Additional considerations

- Assess the consistency of the data types.
- Review the content of data columns or other inputs, and ensure they make sense.
- Look for any evidence of systematic error, e.g. data feeds from sensors breaking without anyone noticing, leading to invalid, incorrect, or missing data values.

Phase 2 - Data Preparation

- Key Activities

5. Survey and visualise

Leverage data visualisation tools to gain overview of the data and detect outliers/skewness.

Shneiderman's "Overview first, zoom and filter, then details-on-demand" visual analytics paradigm

- Enables user to find areas of interest, zoom and filter to find more detailed information about a particular area to the data, and then find the detailed data behind a particular area.

Phase 2 - 5. Survey and Visualise

Questions to ask

- Does the data distribution stay consistent over all the data?
- Does the data represent the population of interest?
- For time-related variables, are the measurements daily, weekly, or monthly? Is that good enough?
- Is the data standardized or normalised? Are the scales consistent?

Additional considerations

- Review data to ensure the calculations remained consistent within columns or across tables for a given data field.
- Assess the granularity of the data, the range of values, and the level of aggregation of the data.
- For geospatial datasets, examine the consistency of the state or country abbreviations used.
- Assess the units used, e.g. English or Metric units.

Analytical Sandbox



- Readymade environment for customers to start building PoCs
- Ready analytical plug-ins to expedite analytical development (Fraud detection, sentiment analysis etc.)

Engage #2-2: Introducing the Innovation Analytics Case Study at EMC Corp.

Explore an example of Phase 2 in the Innovation Analytics project at EMC:

http://stevetodd.typepad.com/my_weblog/2012/03/phase-2-innovation-analytics-data-preparation-1.html

Several tools are commonly used for this phase. Click the icons below to explore some of these tools.



Engage #2-2: Introducing the Innovation Analytics Case Study at EMC Corp.

- Key Activities

Prepare strong bandwidth and network connections to your sandbox.

Collect as much data as you can, including summary data, structured/unstructured, raw data feeds, call logs, web logs, etc.

Determine the type of transformations you will need to assess data quality and derive statistically useful measures.

Transform the data *after* it is in the sandbox.

Acquire the right set of tools for the transformation.

Moving from Phase 2 to Phase 3

- Data Analytics Lifecycle is intended to accommodate ambiguity. This reflects most real-life situations.

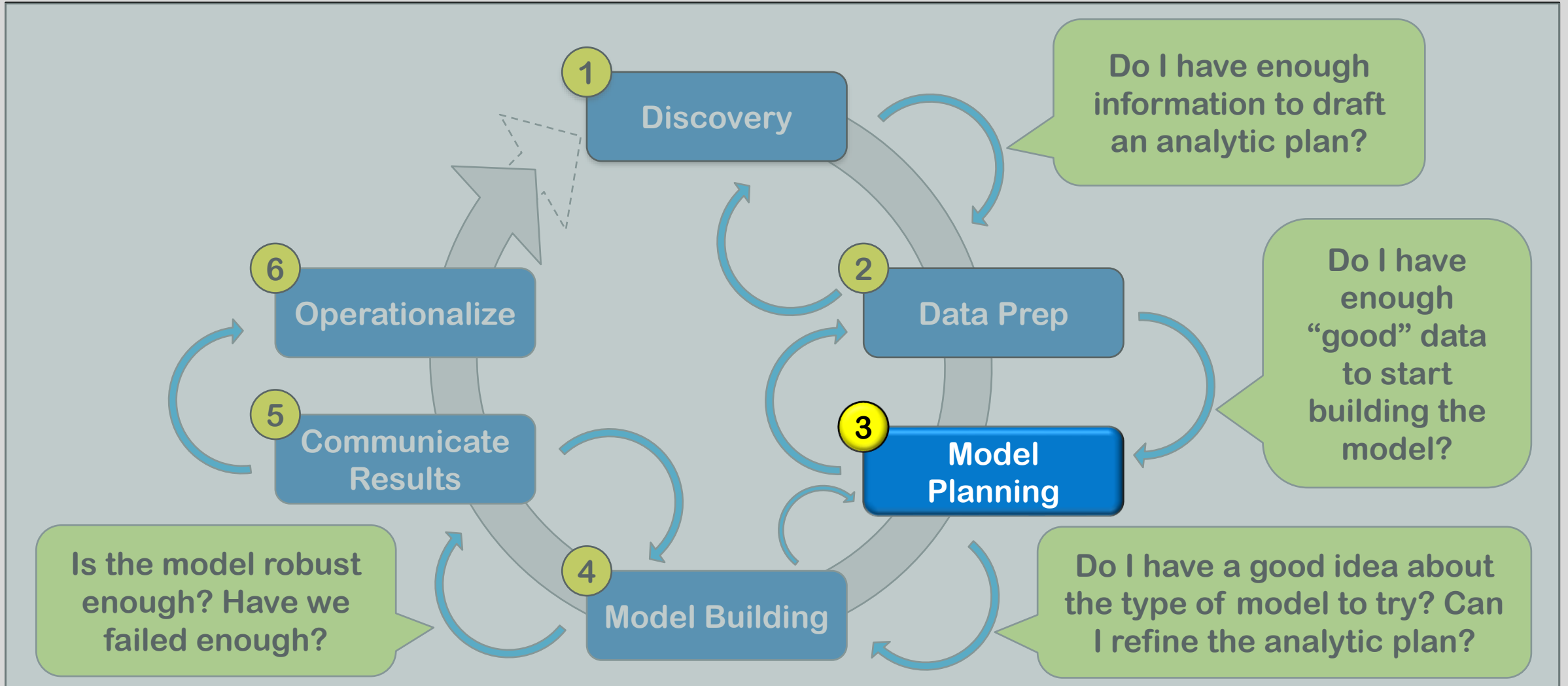


The data science team can move to the next phase if it has deeply knowledgeable about the data.

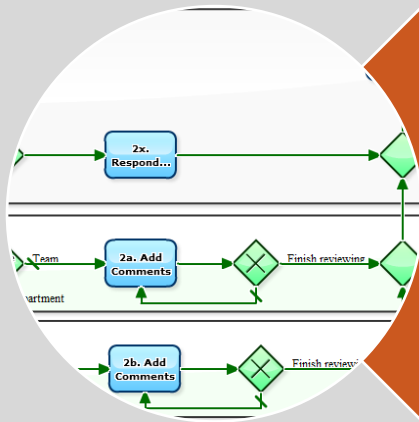


Do I have enough good quality data to start building the model?

Data Analytics Lifecycle – Phase 3



Phase 3 – Model Planning

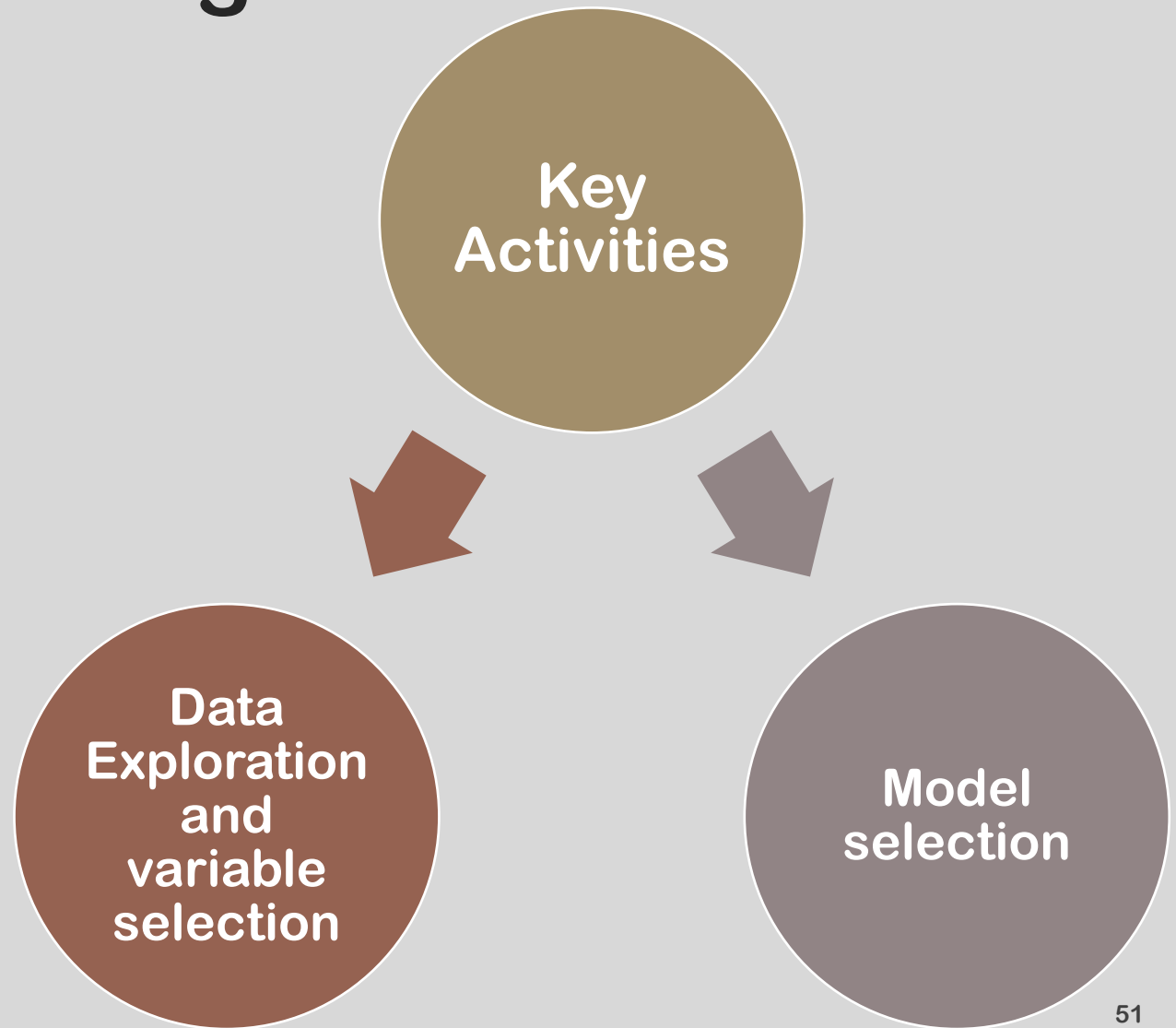


The data science team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase.



The team also explores the data to learn about the relationships between variables, and subsequently selects key variables and the most suitable models.

Phase 3 – Model Planning



Phase 3 – Model Planning

- Key Activities

1. Data Exploration and variable selection

Objective: to understand the relationships among the variables so as to inform selection of the variables and methods, leading to a better understanding of the problem domain.

Stakeholders and domain experts may have pre-existing yet flawed assumptions about the data. Here, the data science team's role is to objectively question these assumptions and correct any bias.

Phase 3 – Model Planning

1. Data Exploration and variable selection



- Capture the most essential predictors and variables rather than considering every possible variable that domain experts think may influence the outcome.

Phase 3 – Model Planning

2. Model Selection

- Key Activities

Objective: to shortlist candidate analytical techniques (i.e. models) based on the end goal of the project.

For Big Data, determine techniques best suited for structured, or unstructured data, or whether a hybrid approach is the best.

Identify and document the modelling assumptions that have informed the selection and construction of preliminary models.

Phase 3 – Model Planning

- Common tools used in the phase:



R

- It has a complete set of modelling capabilities and provides a good environment for building interpretive models with high-quality code.



SQL Analysis Service

- It perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models.



SAS/Access

- It provides integration between SAS and the analytic sandbox via data connectors such as ODBC, JDBC, and OLE DB.

Engage #3: Introducing the Innovation Analytics Case Study at EMC Corp.

Explore an example of Phase 3 in the Innovation Analytics project at EMC:

https://stevetodd.typepad.com/my_weblog/2012/05/phase-3-innovation-analytics-.html

Engage #3: Introducing the Innovation Analytics Case Study at EMC Corp.

- Key Activities

The structure of the data will dictate what tools and analytic techniques can be used in Phase 4. Is textual data being analysed? If so, then maybe Sentiment Analysis using Hadoop is the right approach. Does the sandbox contain structured financial data? Perhaps regression via the R analytics platform is the right method to use.

The analytical technique that is chosen must map back to the business objectives. The objectives are met when the working hypotheses are proved or disproved. This condition clearly highlights why the generation of an Analytic Plan is so important.

Determine whether or not the situation warrants a series of tests, or only one test. If a series of techniques must be used as part of a larger analytic workflow, then the team may benefit from an analytic workflow tool such as Alpine Miner.

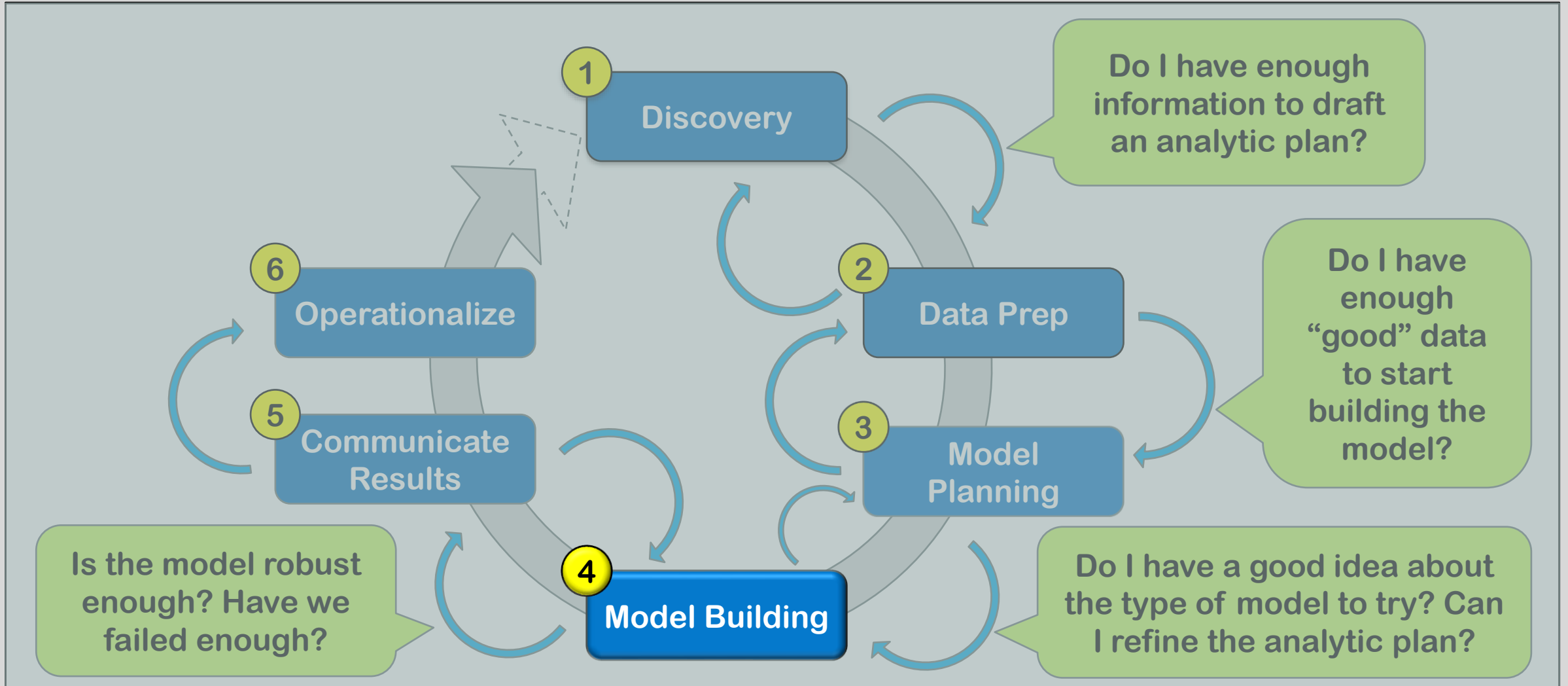
Moving from Phase 3 to Phase 4

- Data Analytics Lifecycle is intended to accommodate ambiguity. This reflects most real-life situations.

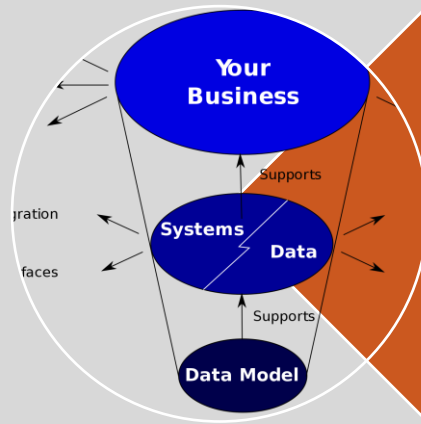
The data science team can move to the next phase once it has a good idea about the type of model to try and the team has gained enough knowledge to refine the analytic plan.

Do I have enough good idea about the
type of model to try?
Can I refine the analytic plan?

Data Analytics Lifecycle – Phase 4



Phase 4 – Model Building



The data science team develops datasets from testing, training, and production purposes. The team builds and executes models based on the work done in the model planning phase.



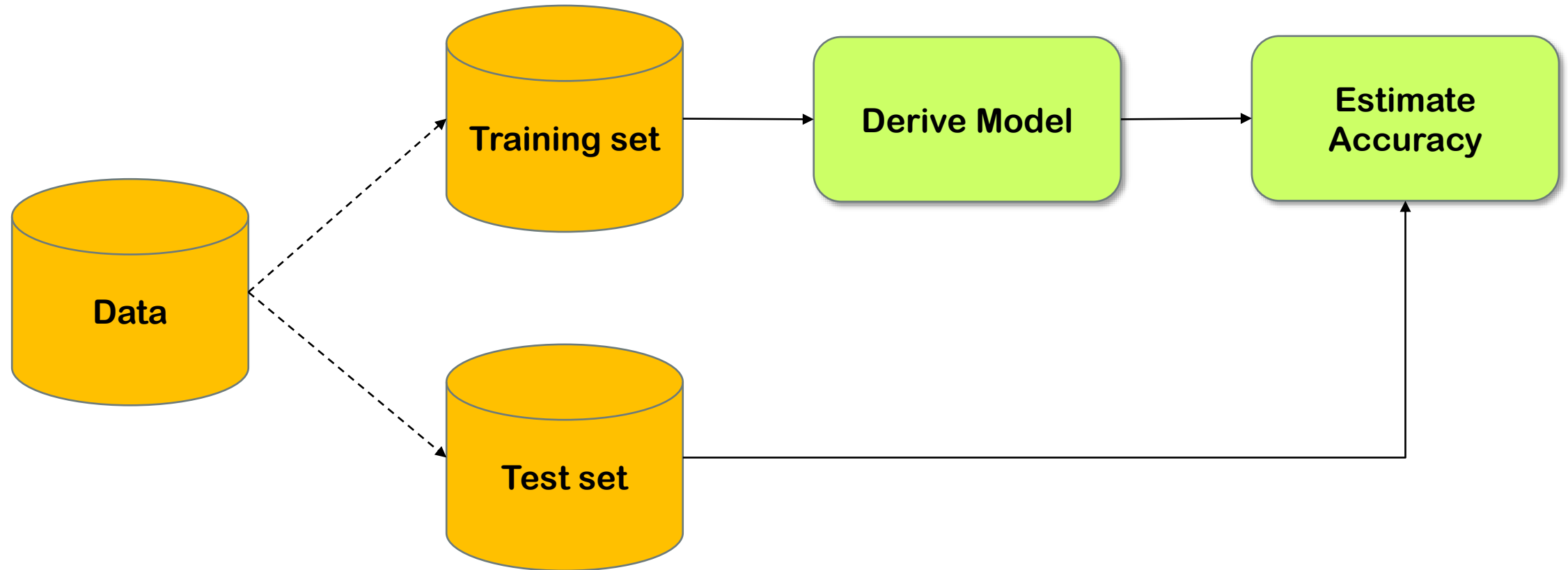
The team also considers the sufficiency of the existing tools to run the models, or whether a more robust environment for executing the models is needed (e.g. fast hardware, parallel processing, etc.).

Phase 4 – Model Building

Key Activities

- **Develop analytical model, fit it on the training data, and evaluate its performance on the test data.**
- **Training Data:** the portion of data used to discover a predictive relationship.
- **Test Data:** the portion of data used to assess the strength and utility of a predictive relationship.
- **The training and test datasets are usually independent from each other (non-overlapping).**

Phase 4 – Model Building



Phase 4 – Model Building

Questions to ask

- Does the model appear valid and accurate on test data?
- Does the model's output/behaviour make sense to the domain experts?
- Do the parameter values of the fitted model make sense in the context of the domain?
- Is the model sufficiently accurate to meet the goal?
- Does the model avoid intolerable mistakes?
- Are more data or inputs needed?
- Will the kind of model chosen support the runtime requirements?
- Is a different form of the model required to address the business problem?

Phase 4 – Model Building

- Commercial tools used in this phase:

- Allows users to run predictive and descriptive models based on large volumes of data from across the enterprise.

SAS
Enterprise
Miner



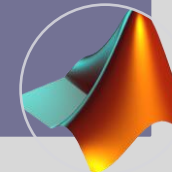
- Offers methods to explore and analyse data through GUI.

IBM SPSS
Modeler



- Provides a high-level language for performing a variety of data analytics, algorithms, and data exploration.

Matlab



- Provides a GUI front end for users to develop analytic workflows and interact with Big data tools and platforms on the back end.

Chorus 6



... and many other well-regarded data mining tools, e.g., STATA, STATISTICA, Mathematica, etc.

Phase 4 – Model Building

- Open source tools:

- PL/R is a procedural language for PostgreSQL with R which allows R commands to be executed in database.

R and
PL/R



- Programming language for computational modelling, with some of the functionalities of Matlab.

Octave



- Data mining package with an analytic workbench and rich Java API.

WEKA



- Offers rich machine learning and data visualisation packages: Scikit-learn, NumPy, SciPy, Pandas, and Matplotlib.

Python



- Provides machine learning library of algorithms that can be executed in-database, for PostgreSQL or Greenplum.

MADlib



Engage #4: Introducing the Innovation Analytics Case Study at EMC Corp.

Explore an example of Phase 4 in the Innovation Analytics project at EMC:

http://stevetodd.typepad.com/my_weblog/2012/06/finding-boundary-spanners.html

Engage #4: Introducing the Innovation Analytics Case Study at EMC Corp.

- Key Discoveries

After visualising the submission record of EMC's Innovation Showcase, it is easier to discover relationships between different submitters.

By drilling down into the identified clusters, further relationships can be discovered.

The visualisation can also give the analyst clues of identifiable factors in a dataset.

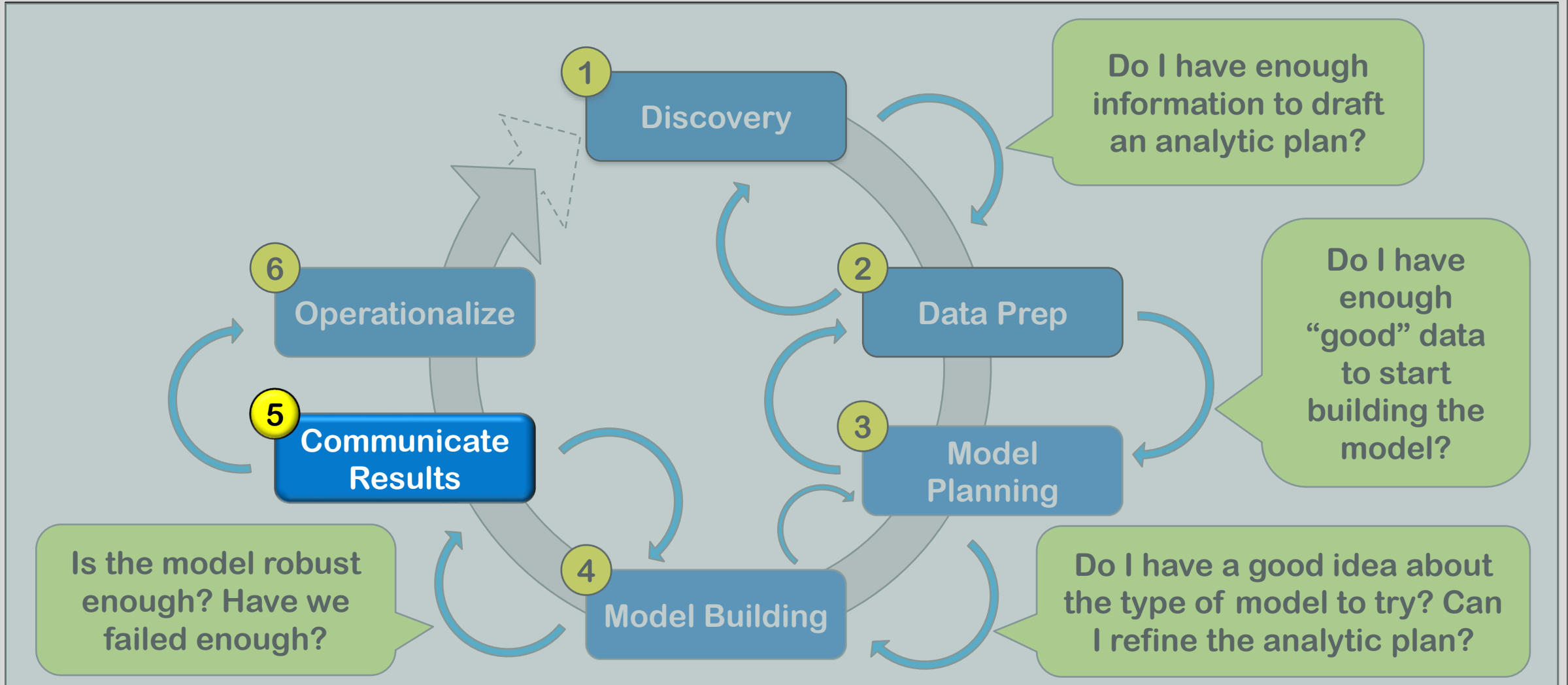
Moving from Phase 4 to Phase 5

- Data Analytics Lifecycle is intended to accommodate ambiguity. This reflects most real-life situations.



The data science team can move to the next phase if the model is sufficiently robust to solve the problem, or if the team has failed.

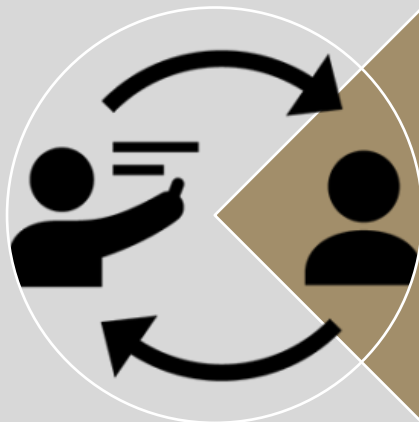
Data Analytics Lifecycle – Phase 5



Phase 5 – Communicate Results

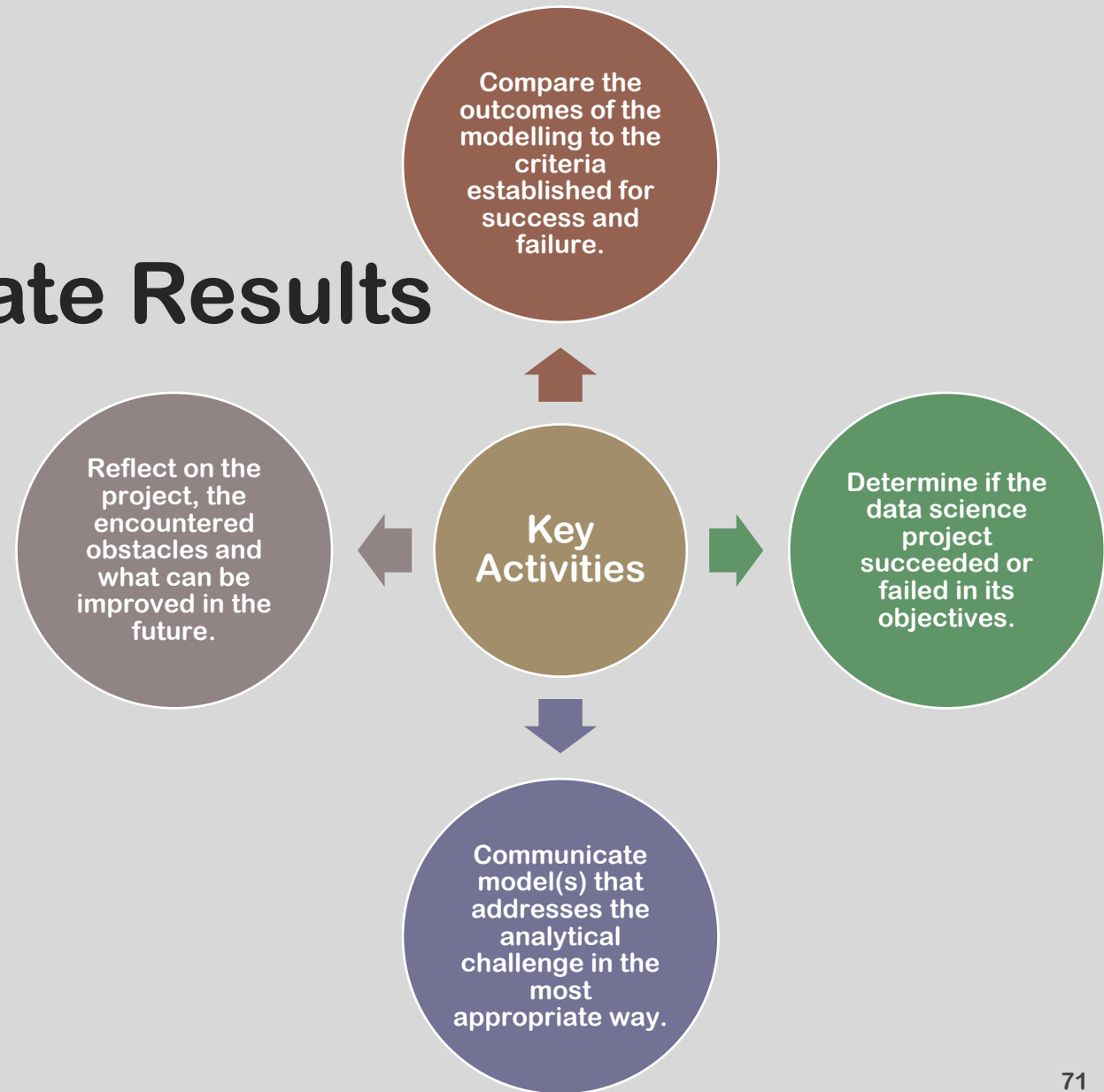


The data science team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1.



The team also considers the sufficiency of the existing tools to run the models, or whether a more robust environment for executing the models is needed (e.g. fast hardware, parallel processing, etc.).

Phase 5 – Communicate Results



Phase 5 – Communicate Results

- Key Activities

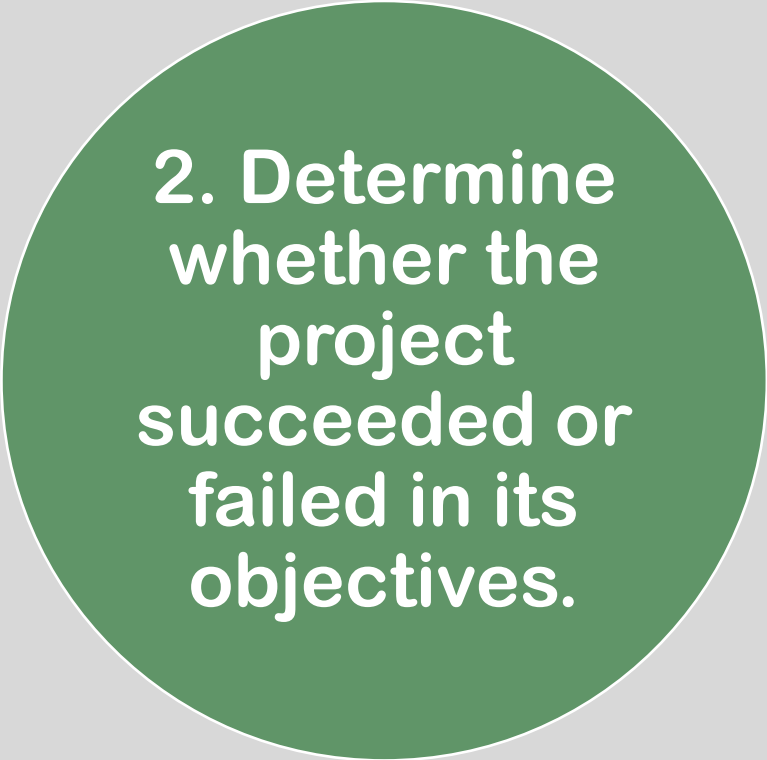
1.
Compare
outcomes
to judge
whether
success or
failure.

Consider how to best articulate the findings and outcomes to various stakeholders.

Take into account *caveats*, assumptions, and any limitation of the results.

Phase 5 – Communicate Results

- Key Activities



2. Determine whether the project succeeded or failed in its objectives.

Failure should not be considered as a true failure. Rather, it should be viewed as a failure of the data to accept or reject a given hypothesis adequately.

Determine if the results are statistically significant and valid.

Phase 5 – Communicate Results

- Key Activities

3. Communicate model(s) that addresses the analytical challenge in the most appropriate way.

Record all the findings and select three most significant ones that can be shared with the stakeholders.

Reflect on the implications of these findings and measure the business value.

Phase 5 – Communicate Results

- Key Activities

4. Reflect on the project, the encountered obstacles and what can be improved in the future.

Make recommendations for future work or improvements to the existing processes.

Consider what each of the team members and stakeholders must do next in order to fulfill his or her responsibilities.

Engage #5: Introducing the Innovation Analytics Case Study at EMC Corp.

Explore an example of Phase 5 in the Innovation Analytics project at EMC:

http://stevetodd.typepad.com/my_weblog/2012/06/phase-5-innovation-analytics-global-knowledge-flight-patterns.html

Engage #5: Introducing the Innovation Analytics Case Study at EMC Corp.

- Key Notes

The vague, fuzzy notions of earlier phases should be replaced by quantifiable conclusions.

What are the three most significant findings in the observation of the data?

The template is an excellent stimulus for communication, and ties back directly to the "plea for resources" made before Phase 1.

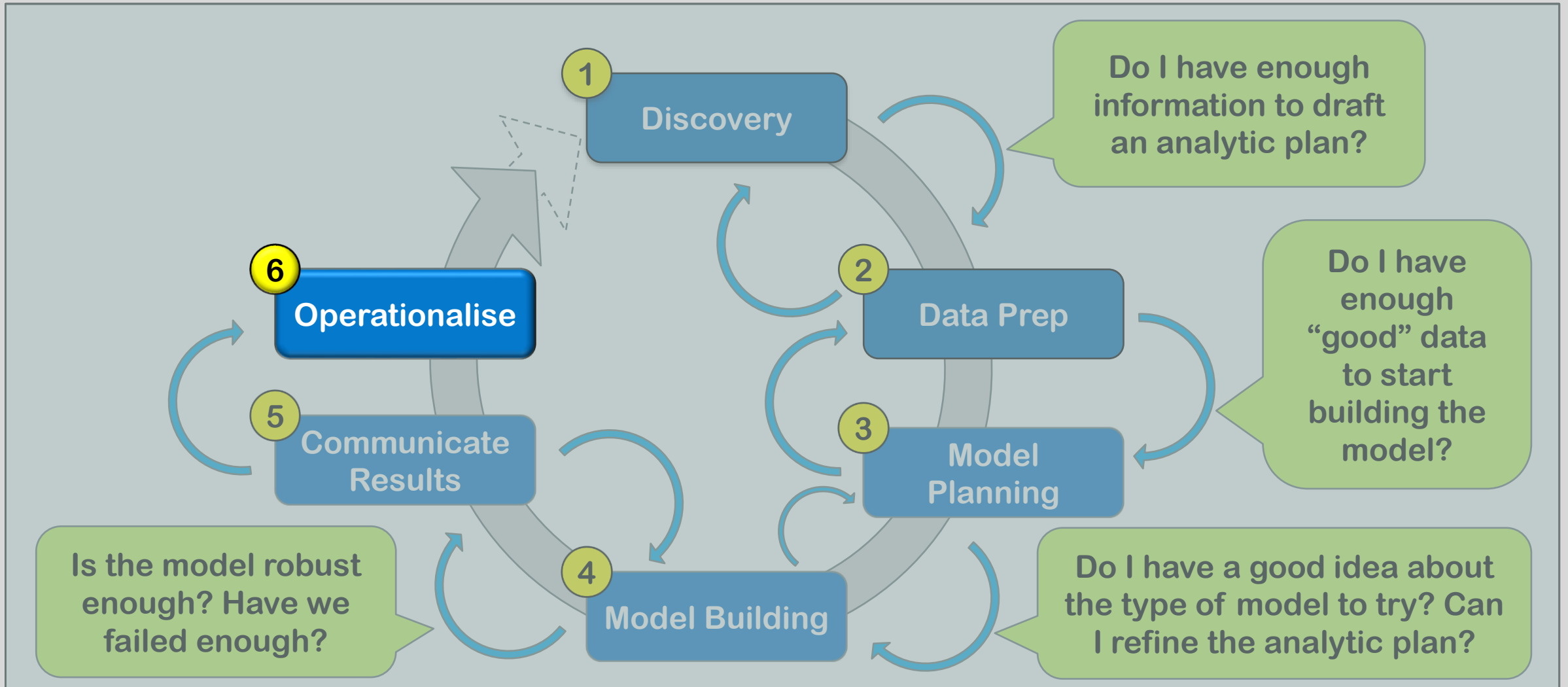
Moving from Phase 5 to Phase 6

- Data Analytics Lifecycle is intended to accommodate ambiguity. This reflects most real-life situations.



The data science team can move to the next phase if the team has documented and reported the key findings, and major insights have been derived from the analysis.

Data Analytics Lifecycle – Phase 6



Phase 6 – Operationalise

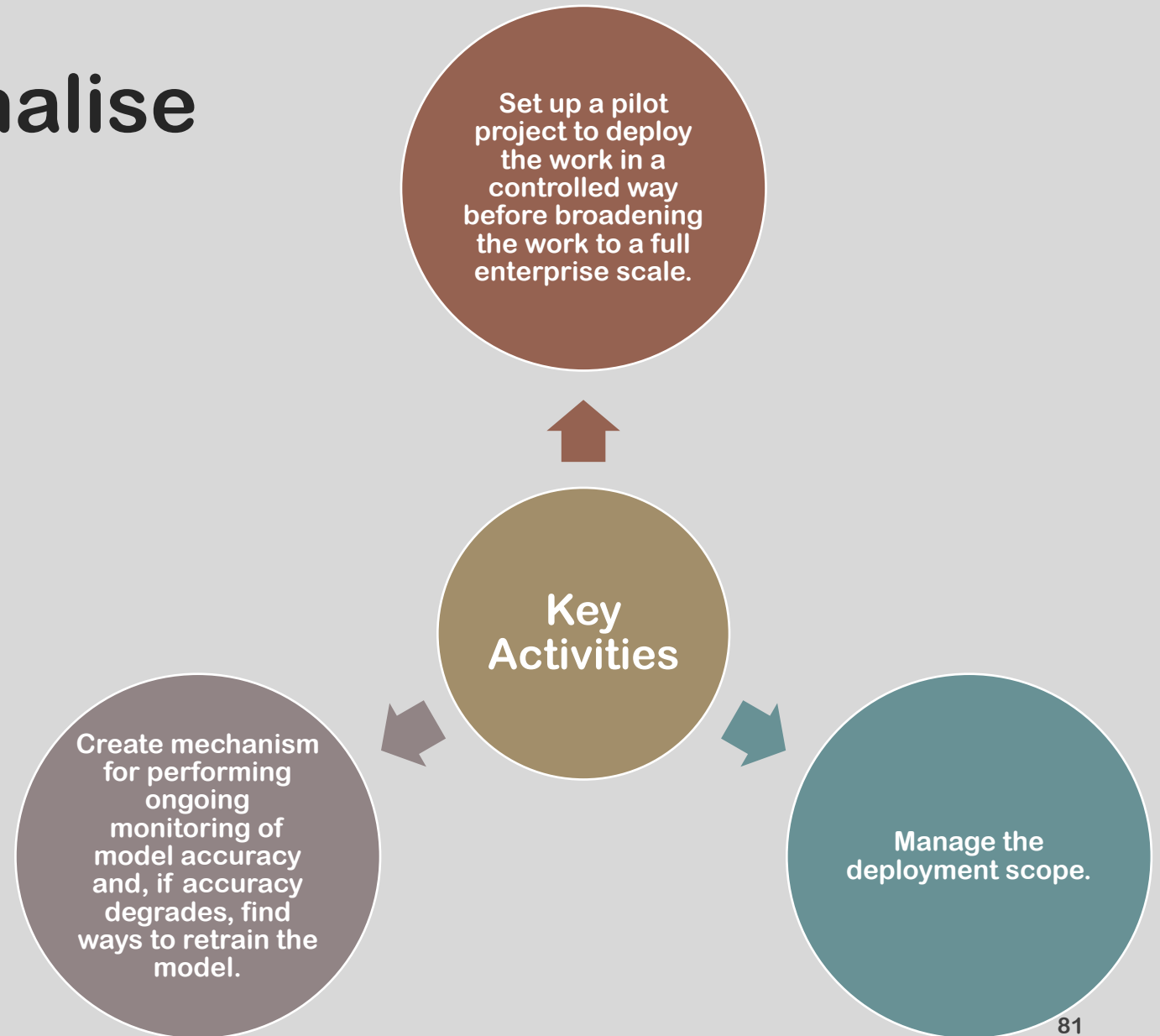


The data science team delivers final reports, briefings, code, and technical documents.



The team may run a pilot project to implement the models in a production environment.

Phase 6 – Operationalise



Phase 6 – Operationalise

- Key Activities

1. Set up a pilot project to deploy the work in a controlled way before broadening the work to a full enterprise scale.


Risk can be better managed.

Further learning the performance and/or constraints of the model.

Adjustments can still be made before a full deployment.

Phase 6 – Operationalise

- Key Activities



2. Manage
the
deployment
scope.

Consider running a model in a production environment for a discrete set of products or a single line of business.

Minimising risks, while allowing model fine-tuning.

Phase 6 – Operationalise

- Key Activities

3. Create mechanism for performing ongoing monitoring of model accuracy and, if accuracy degrades, find ways to retrain the model.

Design “alerts” to detect when the model is operating “out-of-bounds,” e.g., inputs are beyond the allowable range and may affect the accuracy of the model.

Engage #6: Introducing the Innovation Analytics Case Study at EMC Corp.

Explore an example of Phase 5 in the Innovation Analytics project at EMC:

http://stevetodd.typepad.com/my_weblog/2012/07/phase-6-innovation-analytics-operationalize.html

Engage #6: Introducing the Innovation Analytics Case Study at EMC Corp.

- Key Notes

Phase 6 moves the analytic models out of the sandbox and into production. It is advised that a production "pilot" be run first (as opposed to deploying the model on a wide-scale). This approach minimises risk. Smaller-scale deployment allows the team to learn about the performance and make adjustments before a full deployment.

Phase 6 may require a new team of people to join the initiative (the people that are responsible for running the production environment).

The team needs to assess whether the model is meeting goals and expectations, and if desired changes are actually occurring. The data may change over time, or live data may morph to the point where the model needs to be updated or retrained.

Expected Key Outputs

Stakeholders and their expected outputs from a data science project

Business User	Benefits and implications of the findings to the business.
Project Sponsor	Business impacts of the project, the risks, ROI, and the way the project can be evangelised within the organisation and beyond.
Project Manager	The extent to which the project is completed on time and within budget; how well the goals were met.
Business Intelligence Analyst	The extent to which the reports and dashboards he manages will be impacted and need to change.
Data Engineer and DBA	Available code from the analytic project and technical documents on how to implement it.
Data Scientist	Sharable code and explanations of the model to peers, managers, and other stakeholders.

Engage #7: Visual Information – Seeking Mantra

Watch the introductory video on the link below of Tableau visualisation software:

<https://public.tableau.com/s/>

Explore further Shneiderman's visual information-seeking mantra at:

[http://www.infovis-wiki.net/index.php/Visual Information-Seeking Mantra](http://www.infovis-wiki.net/index.php/Visual_Information-Seeking_Mantra)





DATA STORE, ETHICS, AND SECURITY

The Basic Unit in Computer Systems



- The basic unit in IBM PC is “bit.”
- 1 bit is capable of representing “0” and “1,” which corresponds to two states.
→ $2^1 = 2$
- Using 2 bits together, it is capable of representing four states including “00,” “01,” “10,” and “11.”
→ $2^2 = 4$
- 8 bits = 1 Byte.
→ $2^8 = 256$

Data Types

- Before storing the data, let's take a look at what kinds of data are commonly seen in our daily life.



Text Data



Image Data



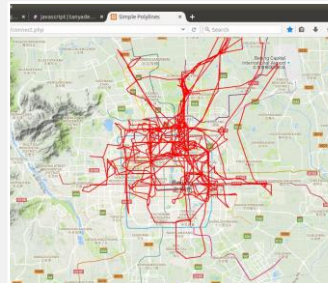
Audio Data



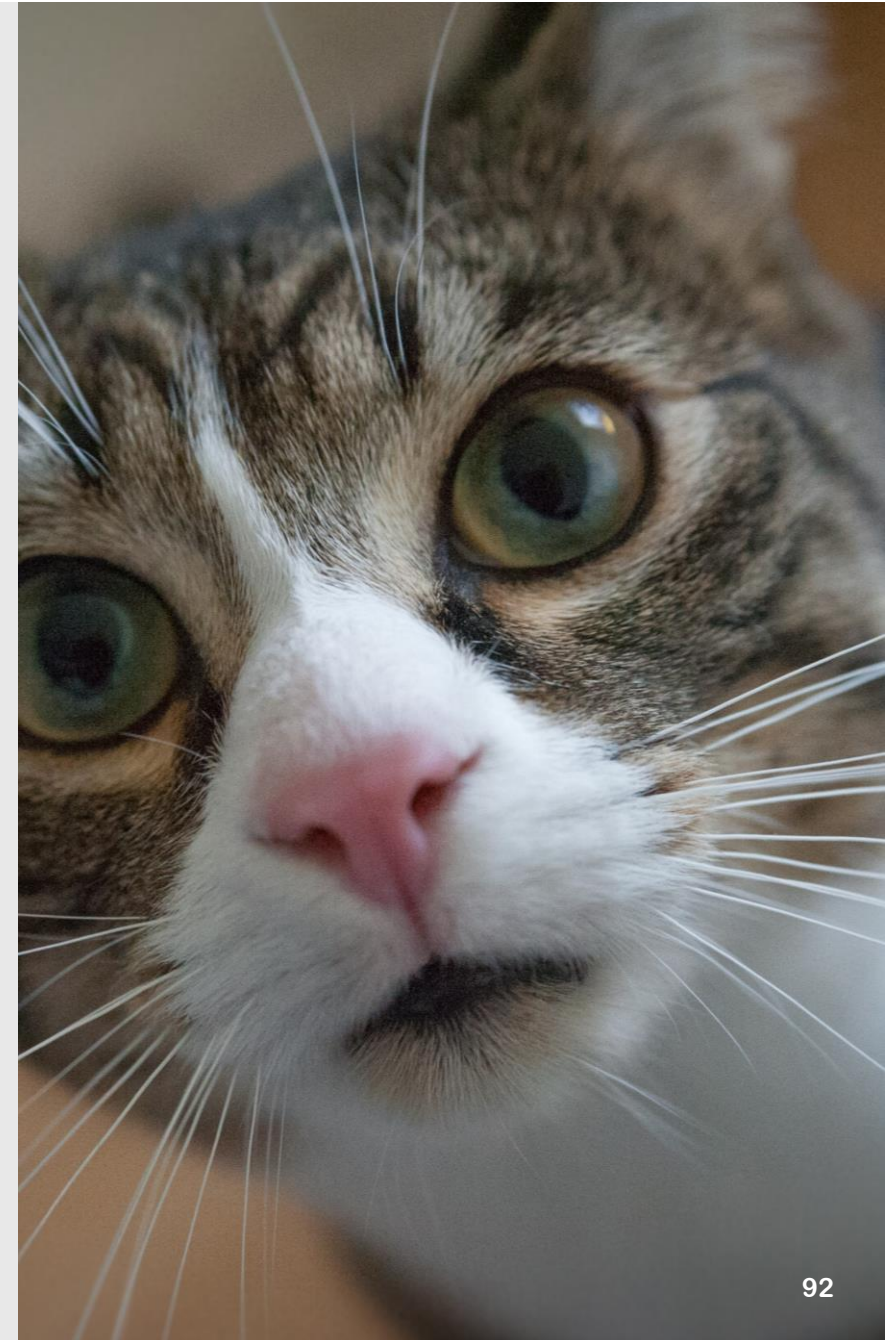
Streaming/Video Data

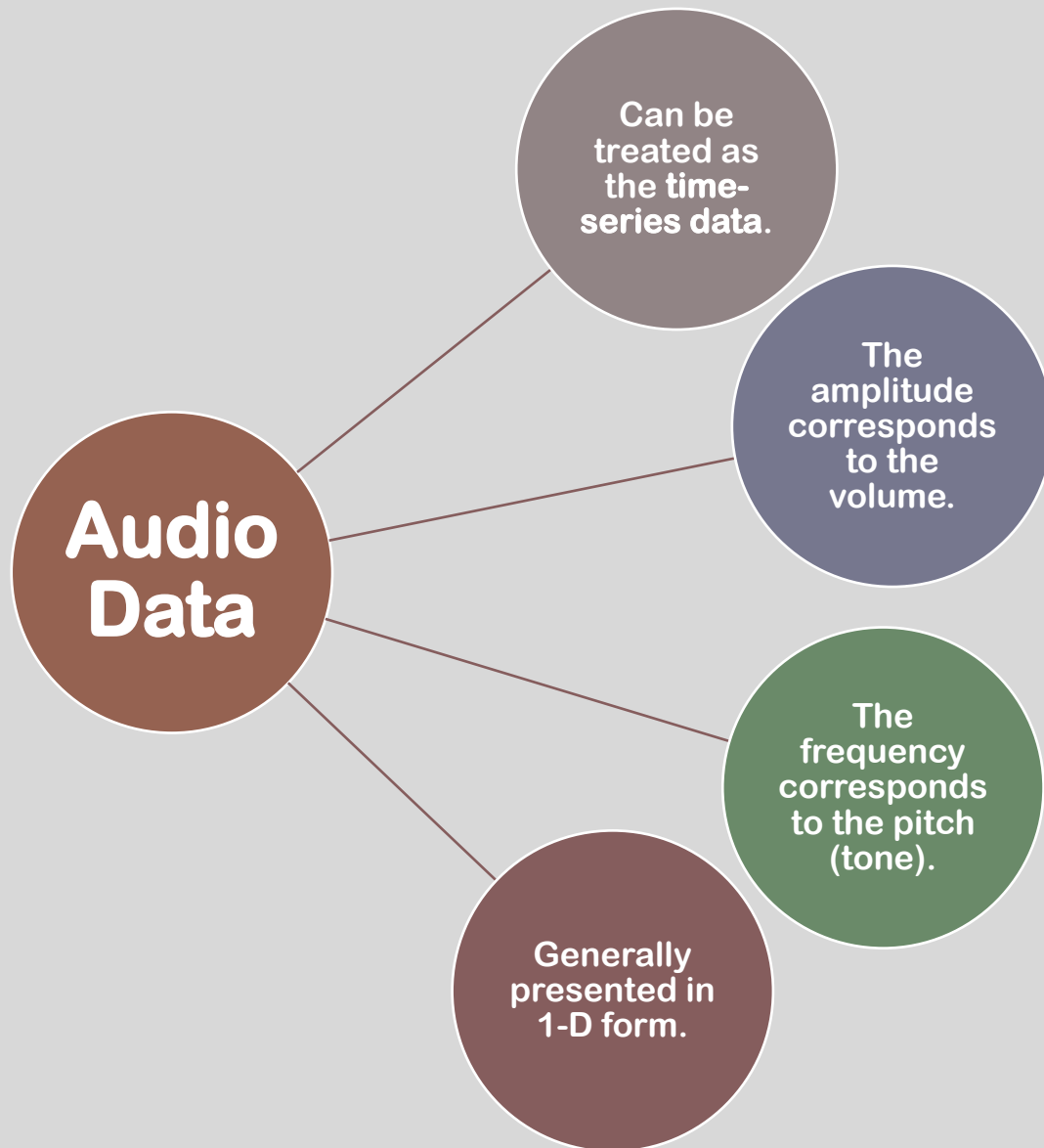


3-D Image



Trajectory Data





- In fact, the audio signal is composed of two parts: the dc component and the ac component.
- The DC part serves as a bias for raising the level of volume. It is usually removed before analysing the signal.

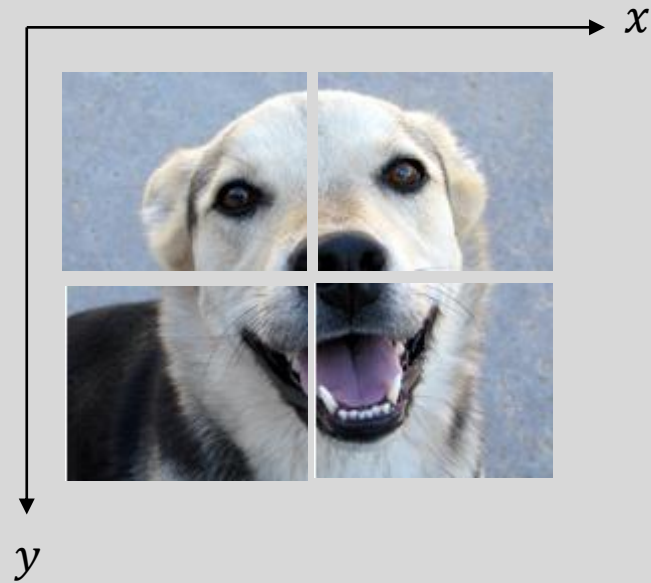
Image Data

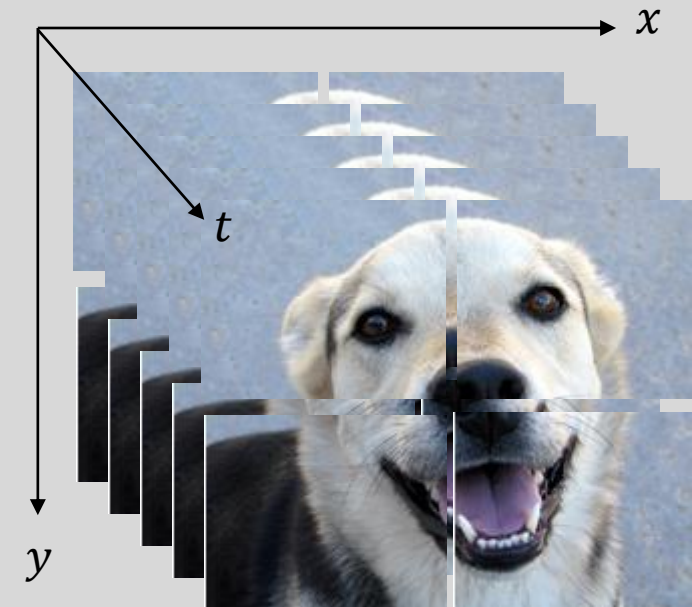
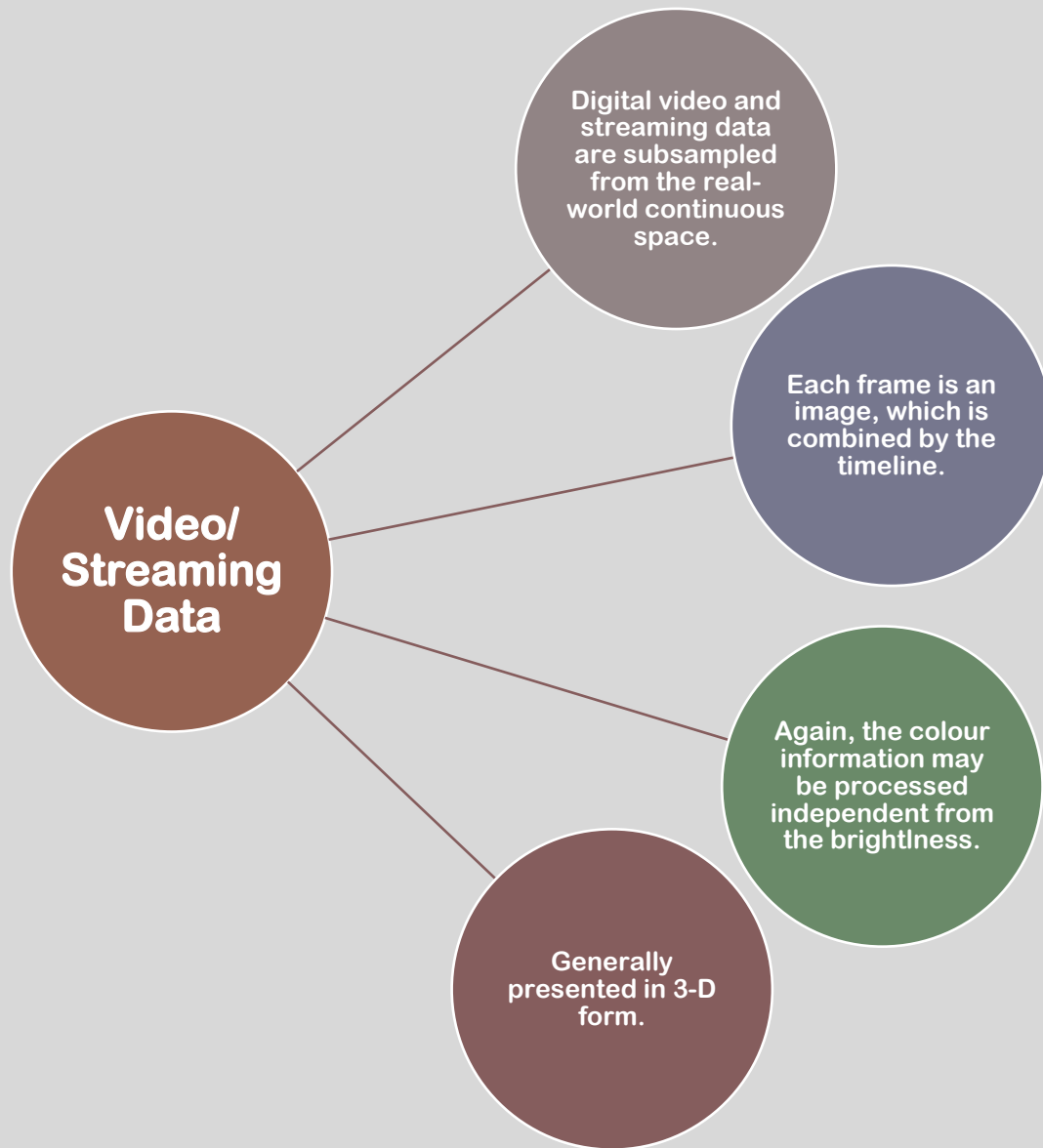
Digital image is subsampled from the real-world continuous space.

Some Models used for storing images isolate the brightness from the colour channels while some processes the colour with the brightness.

Pixel is the basic unit for representing a digital image.

Generally presented in 2-D form.





3-D Image

Instead of using "Pixel" to represent a point, the 3D image uses "Voxel" to represent a point.

Different layers can be retrieved by slicing the 3-D image along different axes.

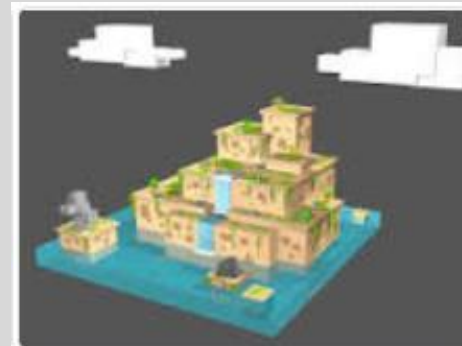
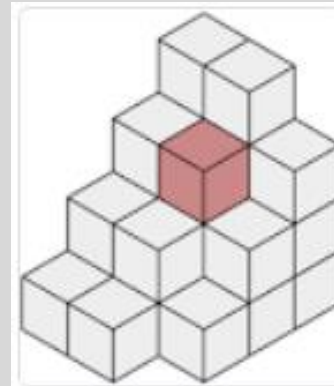
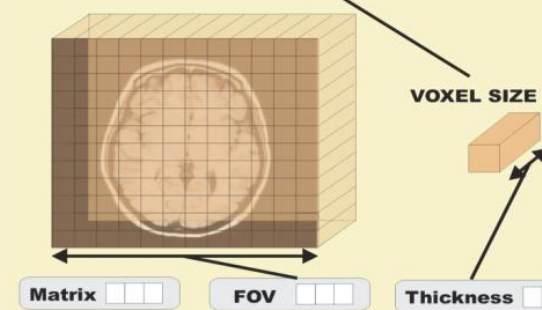
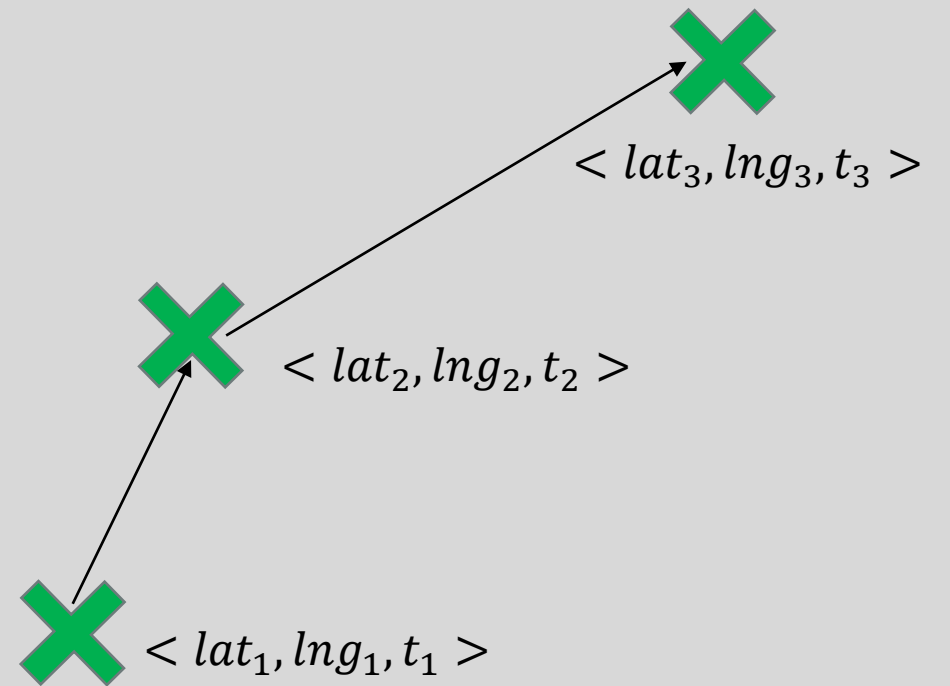
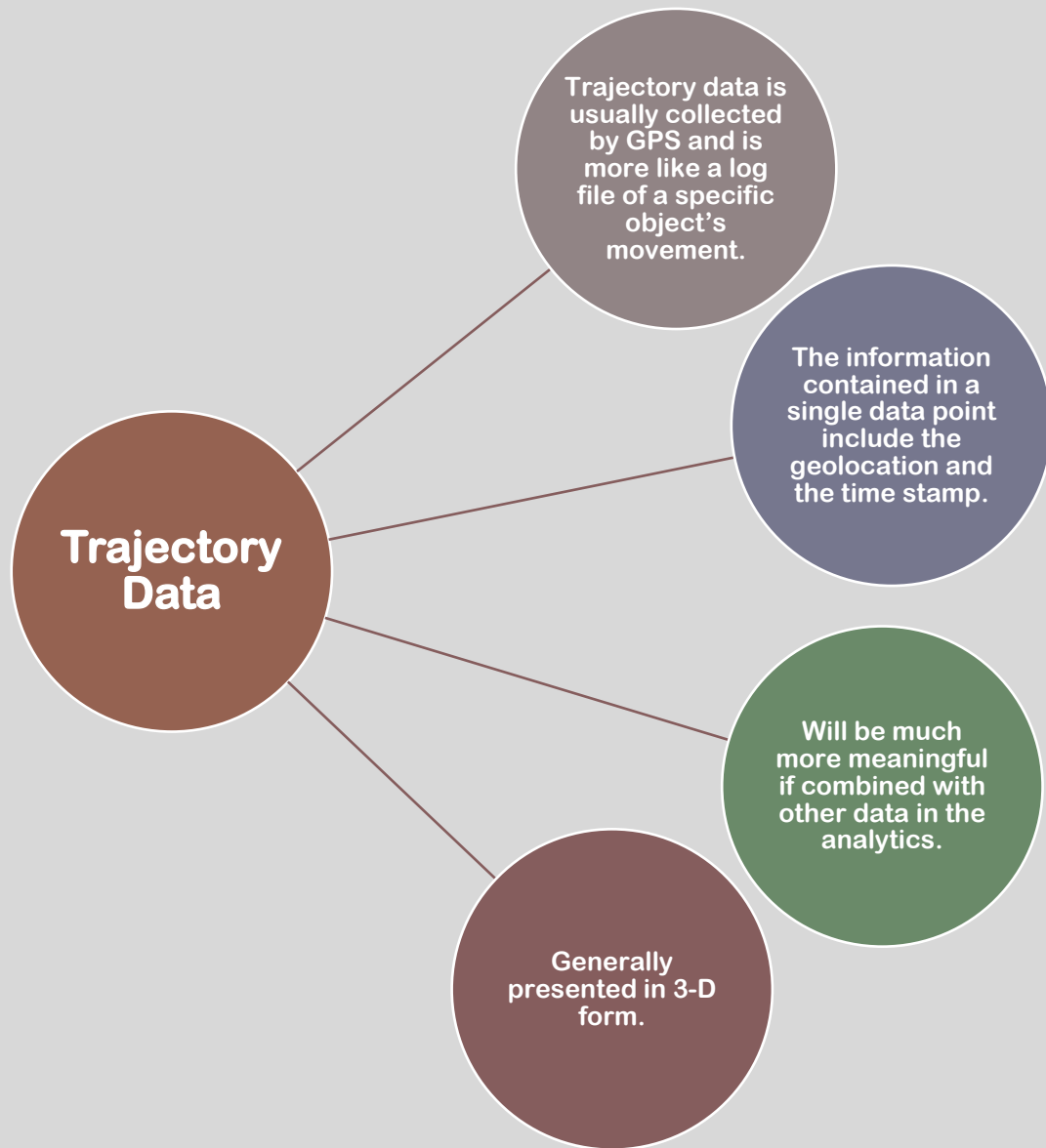


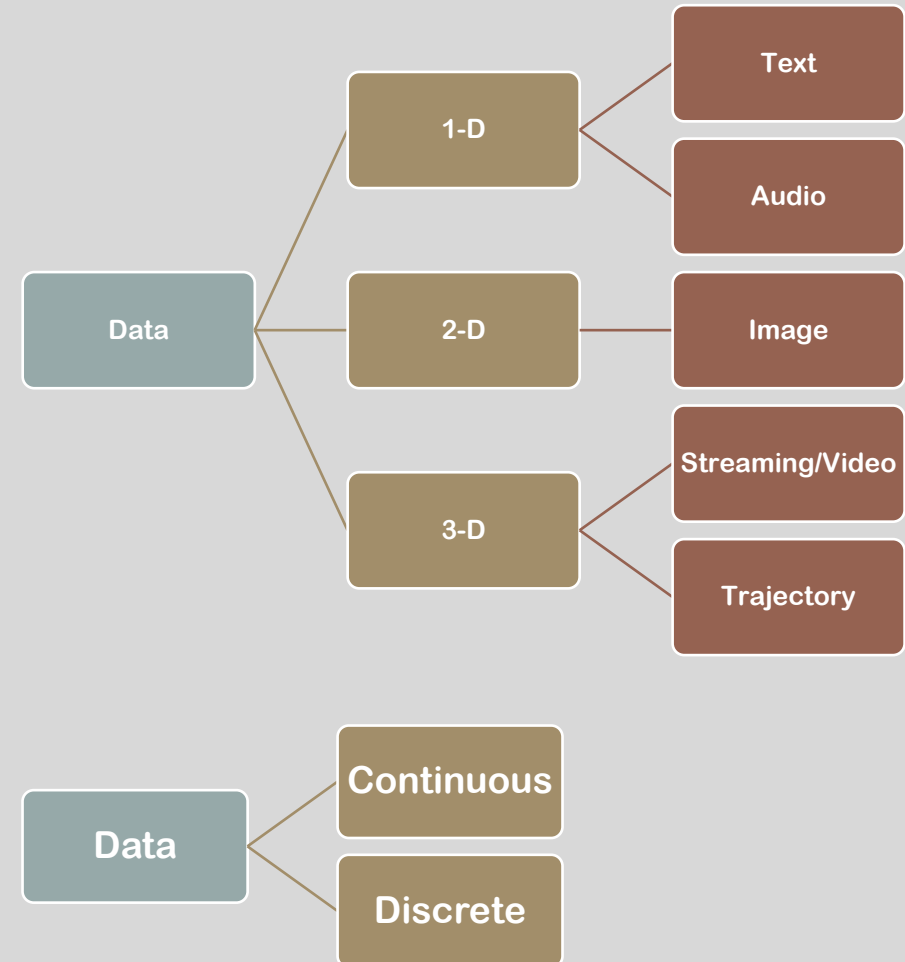
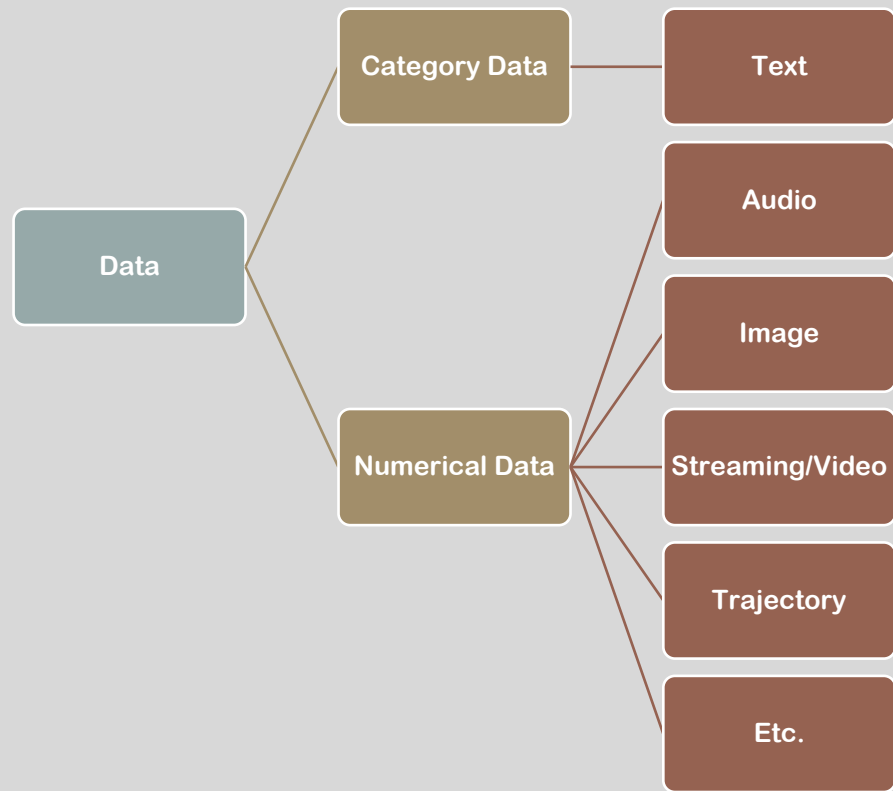
IMAGE DETAIL



Sprawl



Categories



Structured or Unstructured Data?

Structured Data

- Structured data is arranged in a specific manner in tables.
- It is more suitable for structured database such as MySQL, PostgreSQL, etc.

Semi-structured Data

- Semi-structured data is a form of structured data that does not obey the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.

Unstructured Data

- Unstructured data is information that either does not have a predefined data model or is not organised in a pre-defined manner.
- Unstructured data is typically text-heavy.
- Unstructured database such as MongoDB is generally used to store the unstructured data.

Structured or Unstructured Data?

Structured Data

Order_ID	Price	Date
30017	AUD38.99	Apr 15, 2020
30429	AUD92.03	12/12/2020

Semi-structured Data

- `<xml><Order_ID>30017</Order_ID><Price>AUD38.99</Price><Date>Apr 15, 2020</Date><Order_ID>30429</Order_ID><Price>AUD92.03</Price><Date>12/12/2020</Date></xml>`

Unstructured Data

- Order 30017 was made on Apr 15, 2020 with the total amount of AUD38.99.
- Order ID 30429 was created on the twelfth of December in 2020 and the price is AUD92.03.



How to Store Data For Your Data Science Process?

- **Identify Your Goals**
 - Identifying the goals is the first step in process of data storage because all following steps will depend on this.
- **Big Data or Small Data**
 - After having a clear cut goal, you need to decide which type of data you need. It's totally depending on your goal and the available resources.
- **Avoid Data Fatigue**
 - Try not to over storage or measurement of data which is useless for you or doesn't align with your data collection goals.
- **Data Management**
 - Chose either SQL or NoSQL way to manage the stored data.



ETHICS



ETHICS IN DATA SCIENCE



What is Ethics?

Data ethics refers to the moral principles and values that govern the collection, use, and dissemination of data. It involves considering the impact of data on individuals, society, and the environment, and making decisions that are consistent with ethical standards.

Data ethics involves a range of considerations, including:

1. **Privacy:** protecting the personal information of individuals and preventing its misuse.
2. **Consent:** obtaining informed consent from individuals before collecting or using their data.
3. **Transparency:** being transparent about how data is collected, used, and shared.
4. **Bias:** identifying and mitigating bias in data collection and analysis.
5. **Security:** protecting data from unauthorized access or misuse.
6. **Accountability:** taking responsibility for the ethical use of data and ensuring that any negative impacts are addressed.

ETHICS



What is Ethics?

Data ethics is increasingly important as the amount of data collected and used by organizations continues to grow. Ensuring that data is collected and used in an ethical manner is essential for maintaining trust with stakeholders and protecting the rights of individuals.

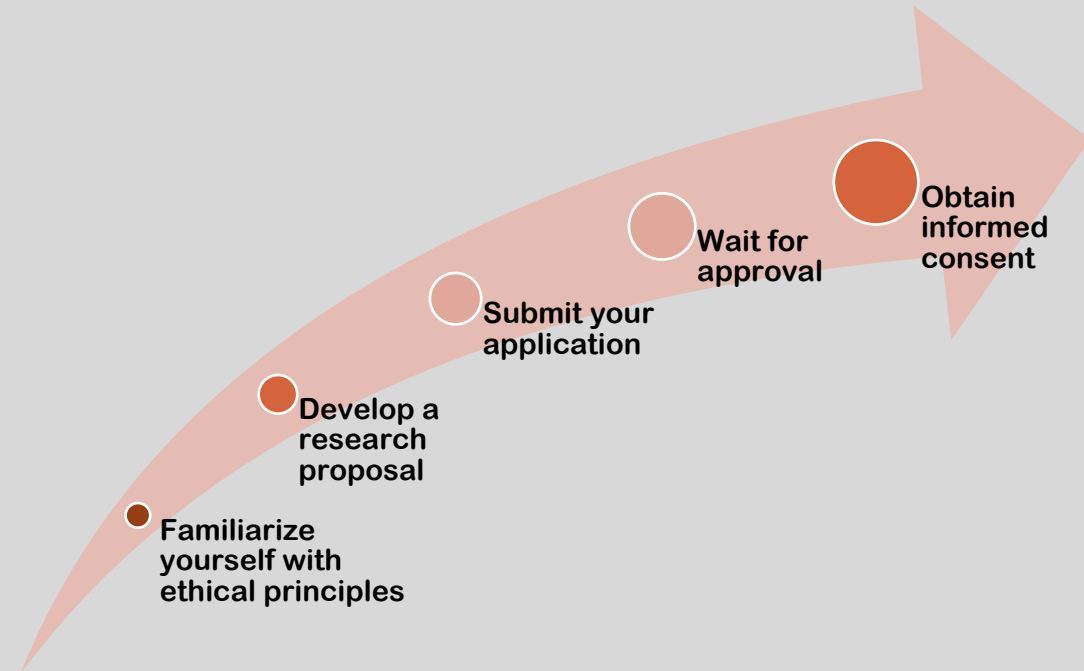
Nowadays, doing any project that involves data from human or user behaviours, ethics approval is a must before the project can go into the execution phase.

ETHICS



How to Apply for Ethics Approval?

The procedure to obtain ethics approval for research can vary depending on the specific requirements of the research institution, as well as the nature of the research being conducted. However, some common steps involved in the process are:





- 1. Familiarize yourself with ethical principles:** Before beginning the application process, it is important to understand the ethical principles that govern research, such as respect for autonomy, beneficence, non-maleficence, and justice. These principles will guide your research design and help you identify potential ethical issues that may arise.
- 2. Develop a research proposal:** Your research proposal should provide a detailed description of the research question, methods, and potential risks and benefits to participants. It should also include information on how you plan to obtain informed consent, protect participant privacy, and minimize potential harms.
- 3. Submit your application:** Once you have developed your research proposal, you will need to submit an ethics application to the appropriate research ethics committee or institutional review board (IRB). This application will typically include a description of your research question, methods, data collection and analysis procedures, and any potential risks and benefits to participants. You may also be required to provide a copy of your informed consent form.

A close-up photograph of a person's hand holding a blue pen. The hand is positioned on the right side of the frame, with the thumb and index finger gripping the pen. A single, continuous, horizontal blue line has been drawn across the upper portion of the image, extending from the pen towards the left edge. The background is a plain, light gray.



How to Apply for Ethics Approval?

4. **Wait for approval:** The ethics committee or IRB will review your application and may request additional information or revisions before granting approval. The review process can take several weeks or even months, depending on the complexity of the research and the workload of the ethics committee or IRB.
5. **Obtain informed consent:** Once you have received ethics approval, you can begin recruiting participants and obtaining informed consent. Informed consent involves providing participants with a clear understanding of the nature of the research, the risks and benefits of participating, and their right to withdraw at any time.
6. **Conduct your research:** With ethics approval and informed consent in place, you can begin conducting your research in accordance with the approved protocol.



How to Apply for Ethics Approval?

It is important to note that the process for obtaining ethics approval can be complex and time-consuming, but it is necessary to ensure that research is conducted in an ethical manner and respects the rights and welfare of participants.



DATA SECURITY

Data Security

Data security refers to the protection of digital data, such as files, databases, and other sensitive information, from unauthorized access, use, disclosure, modification, or destruction. It involves implementing measures to prevent data breaches, theft, and other forms of cyber-attacks that can compromise the confidentiality, integrity, and availability of data.

Data security measures may include:

Access controls

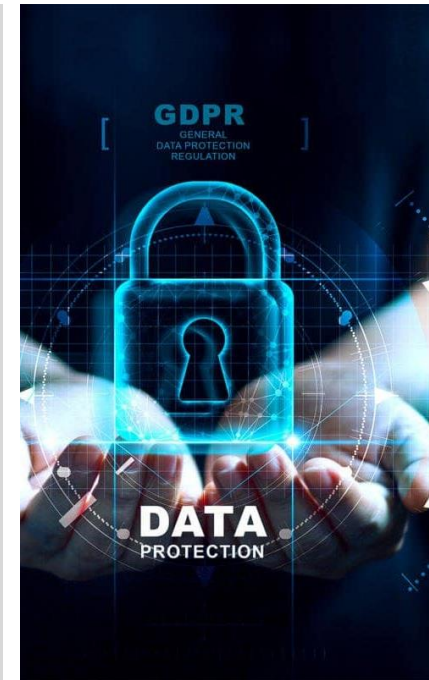
Encryption

Network security

Physical security

Data backup and recovery

Data retention policies



Data Security

Access controls

- Limiting access to sensitive data by requiring authentication, such as passwords or biometric identification.

Encryption

- Using encryption algorithms to encode sensitive data, making it unreadable to unauthorized users.

Network security

- Securing networks with firewalls, intrusion detection and prevention systems, and other security technologies.

Physical security

- Protecting physical devices that store or process sensitive data, such as servers, hard drives, and mobile devices, with physical access controls and security measures.

Data backup and recovery

- Regularly backing up data to secure locations to ensure that it can be restored in the event of a data breach or system failure.

Data retention policies

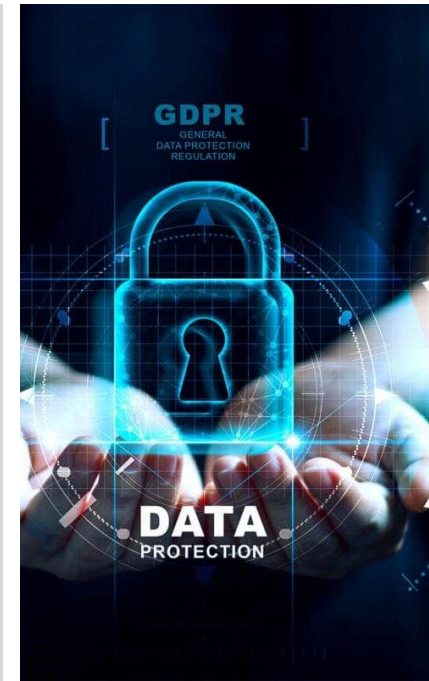
- Establishing policies and procedures for retaining and disposing of data in accordance with legal and regulatory requirements.



Data Security

Data security is essential for maintaining the confidentiality, integrity, and availability of sensitive information, such as personal information, financial data, and intellectual property.

Failure to implement adequate data security measures can result in serious consequences, such as loss of business reputation, financial loss, legal liability, and damage to individual privacy and rights.





Problem Transformation

- Using Proper Techniques to Solve the Problems.
- Convert the business problem into a data problem.



Data Analytics Lifecycle

- The Genesis of EMC's Data Analytics Lifecycle
- Steps for completing a data science project.

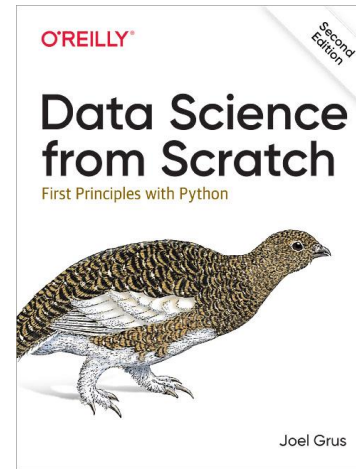
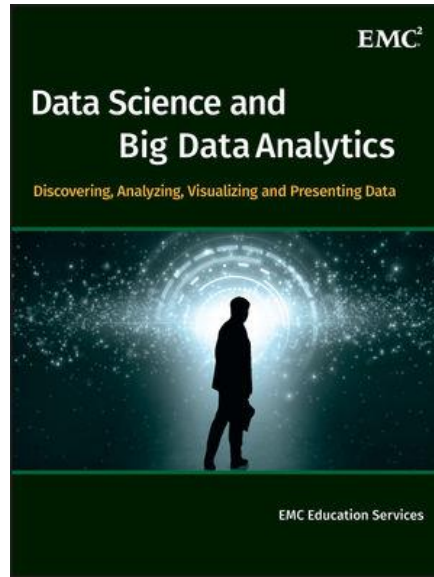
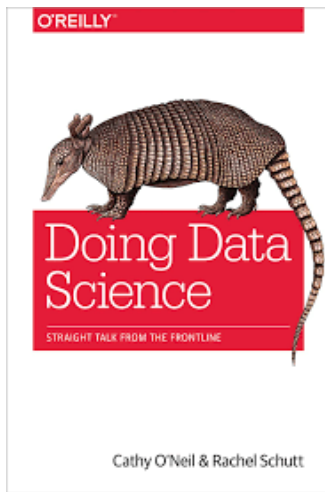


Data Store

- Identify the goal, first.
- Based on the goal and the available resources, choose the data to be stored.
- Chose the most suitable tool for storing and managing the data.

Summary





Texts and Resources

- Unless stated otherwise, the materials presented in this lecture are taken from:
 - Dietrich, D. ed., 2015. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. EMC Education Services.
 - Schutt, R. and O'Neil, C., 2013. *Doing data science: Straight talk from the frontline*. O'Reilly Media, Inc.
 - Pierson, L., *Data Science for Dummies, 2nd Edition*, John Wiley & Sons © 2017
 - Mueller, J. P. and Massaron, L., *Python for Data Science for Dummies, 2nd Edition*, John Wiley & Sons © 2019 (432 pages), ISBN:9781119547624
 - Joel Grus, 2019. *Data Science from Scratch – First Principles with Python, 2nd Edition*, O'Reilly Media, Inc.

