Society of Actuaries in Ireland

Institute of Technology Carlow

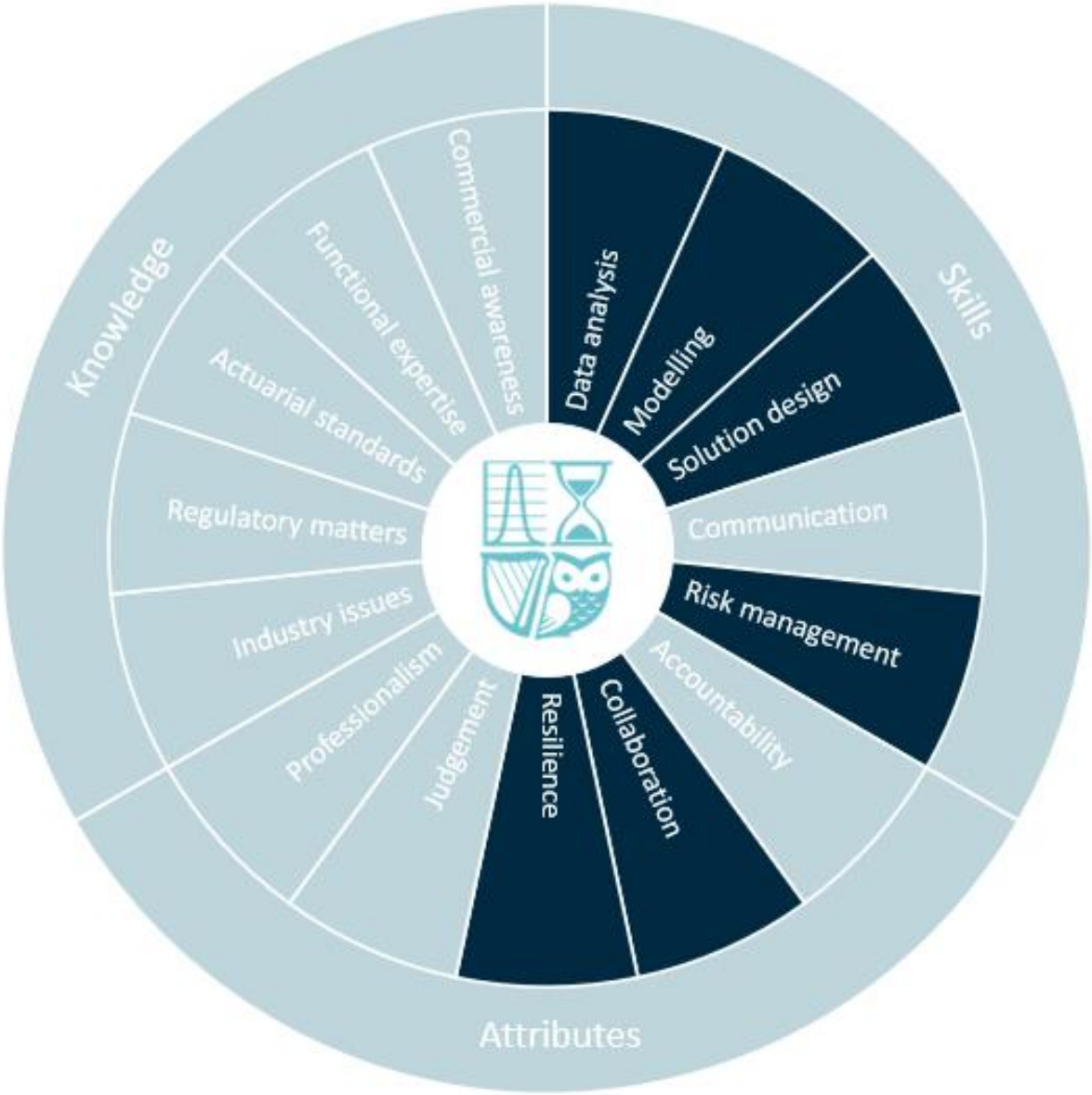
CRISP-DM - a structured approach to planning a data analytics project.

October 5th 2021

Disclaimer

The views expressed in this presentation are those of the presenter(s) and not necessarily those of the Society of Actuaries in Ireland or their employers.

SAI Competency Framework Wheel



Data, data everywhere…

- Greg Doyle B.Sc. M.Sc. PhD

- My observations come from:

  - Personal experiences from teaching, research and professional consultancy work

  - Advisory engagements with various industries, organisations and SME's

  - Discussions with colleagues and company executives

  - **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

Aims of this talk

- The aim is to provide
  - a thorough understanding of process models in data science using CRISP-DM as an exemplar
  - briefly showcasing some previous projects (within the bounds of client confidentiality)
  - This will include the
    - main phases
    - strengths
    - weaknesses

  of CRISP-DM weaved throughout, and
    - alternative process models/methodologies

Agenda

- Introduction – business case for analytics
- Framework/process model – need repeatability and reliability
- Main phases of CRISP-DM
  - Business understanding
  - Data understanding
  - Data preparation
  - Modelling
  - Evaluation
  - Deployment
- Some real world examples – using CRISP-DM
- Alternatives/modifications/add-ons to CRISP-DM
- Technologies & tools for data scientists

Business case for analytics

- Optimise people
- Optimise processes
- Optimise material management
- Fraud reduction
- Data based decision making
- Learning etc.

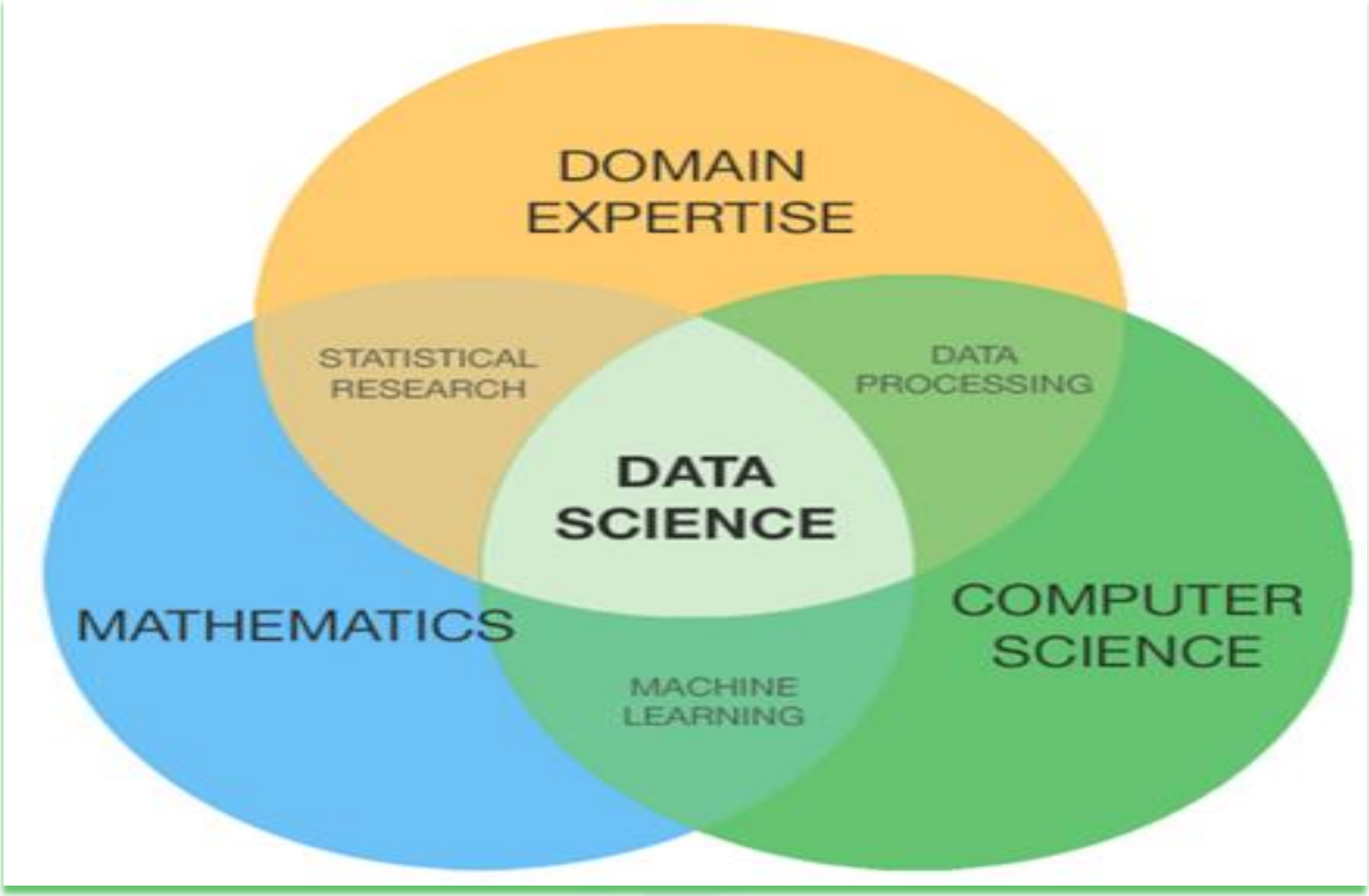
Business case for analytics

- Short term easy wins
- Address business drivers for the company/leaders
- Analytics to help decision makers (perf. v obj.)
- Connect data & analytics governance to business outcomes/objectives
- **Data quality is key**

Industries hiring data scientists 2021

Finance - JPMorgan Chase, ICIC Bank, HDFC, HSBC, BNP Paribas, Citi Group
Media - Dish Network, Netflix, Time Warner, Fox, Viacom, NDTV
Healthcare - GSK, GE Healthcare, and Sonofi.
Retail - Amazon, Walmart, Flipkart
Telecoms - Vodafone-IDEA
Automotive - General Motors, Volkswagen, Maruti Suzuki, Hyundai and Honda
Digital Marketing - Amazon, Google, Facebook, Flipkart, Walmart
Cyber Security - Accenture, Cisco, IBM, Microsoft, McAfee
…

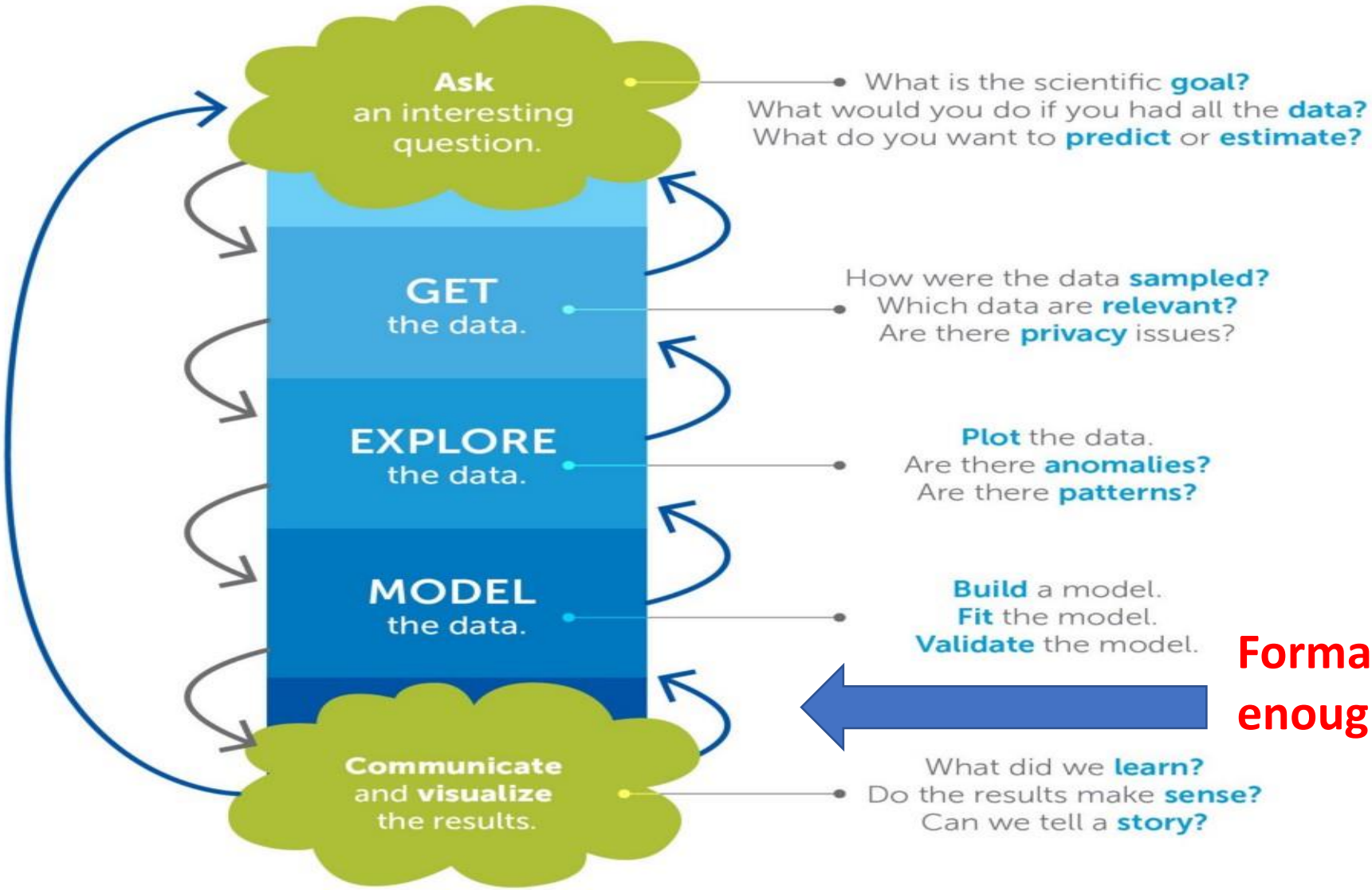
Data science & analytics skills



Analytics frameworks - process models

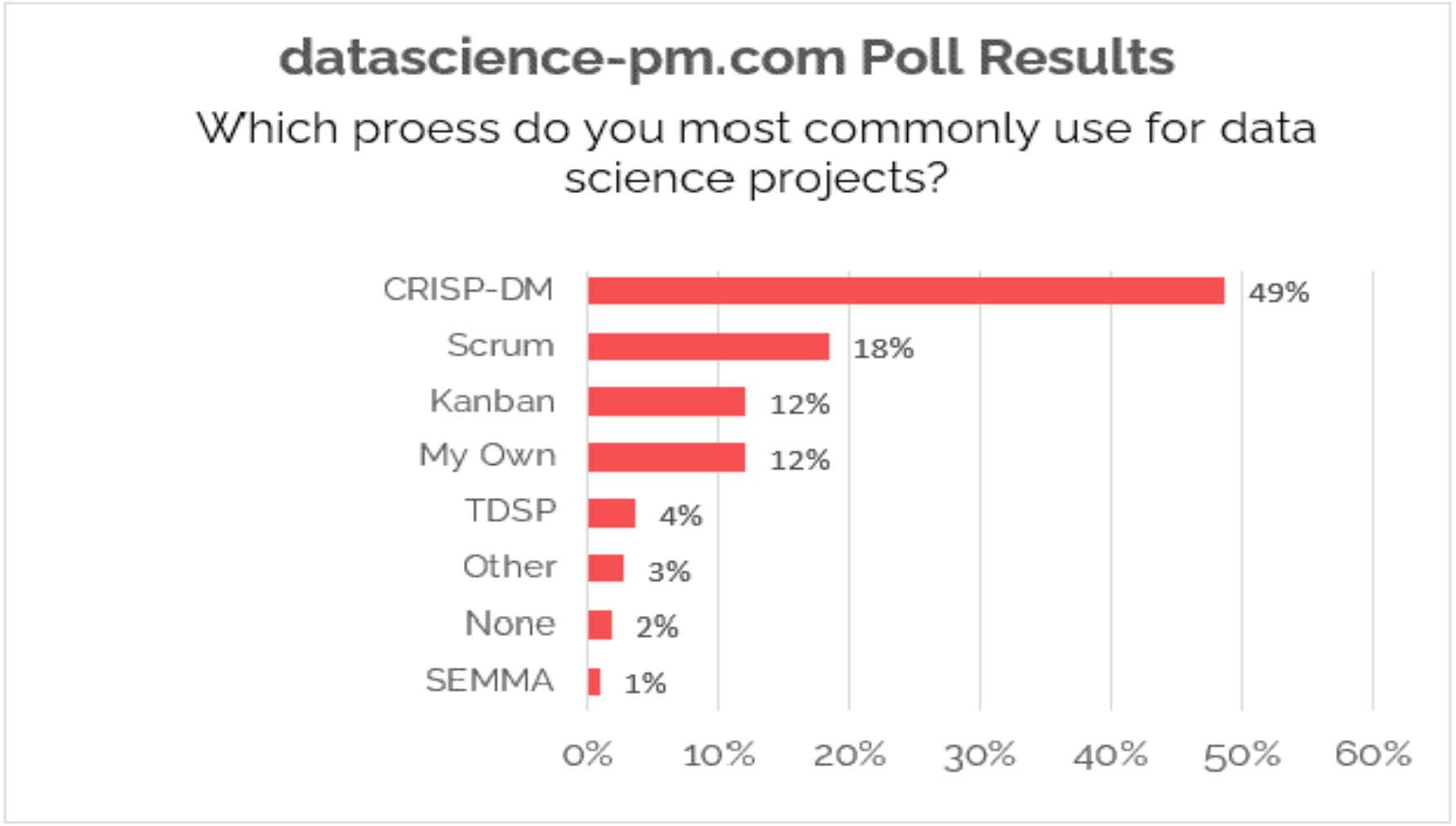
- To do data projects well the process must be reliable and repeatable
- Framework for recording experience
  - Allows projects to be replicated, scientific
- Aids and drives project planning and management
- Comfort factor for new adopters
  - Demonstrates maturity of data analytics
  - Reduces dependency on specific data analytics experts

The Data Science Process



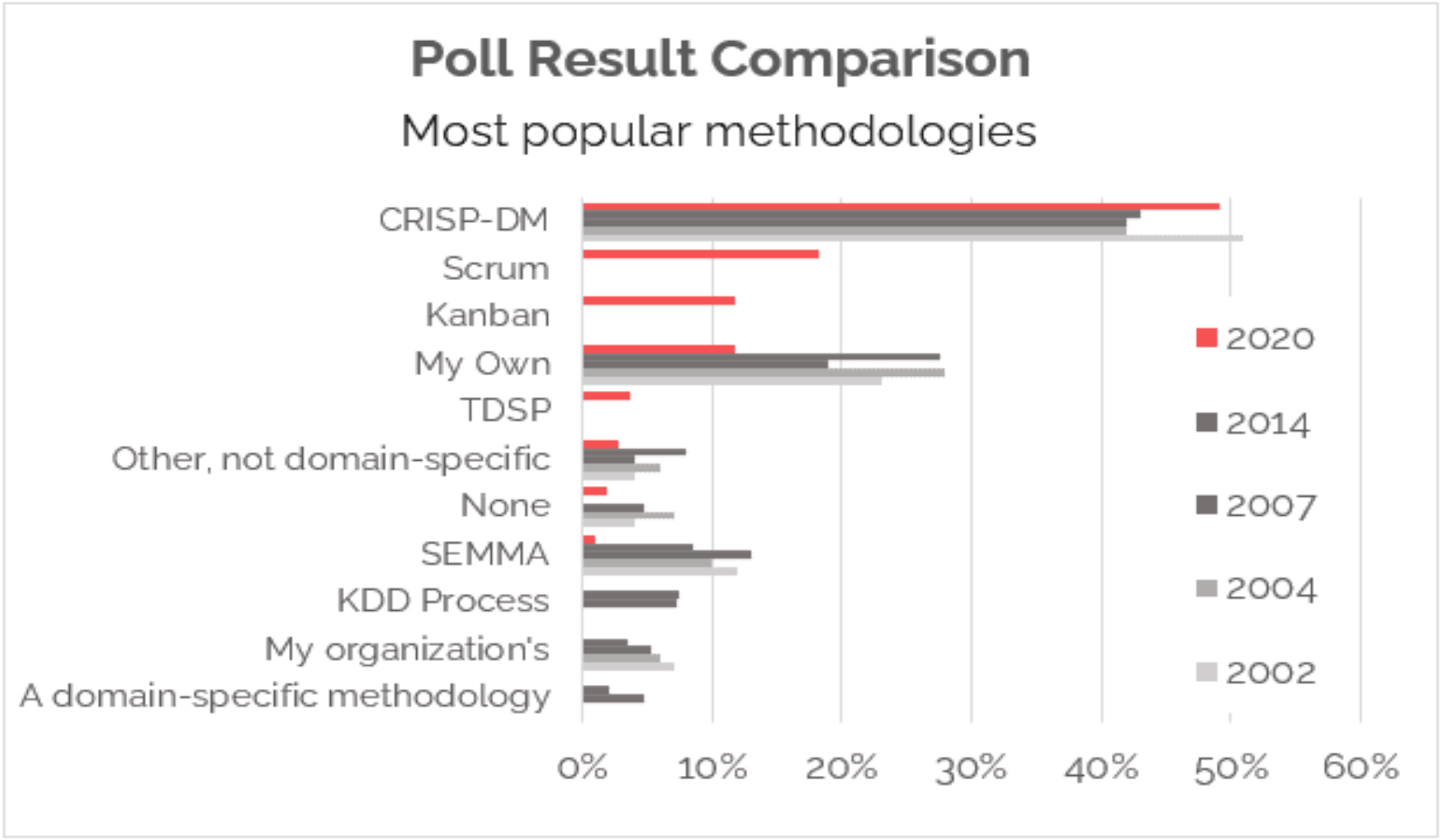
Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course http://cs109.org/.

Frameworks – most common process models



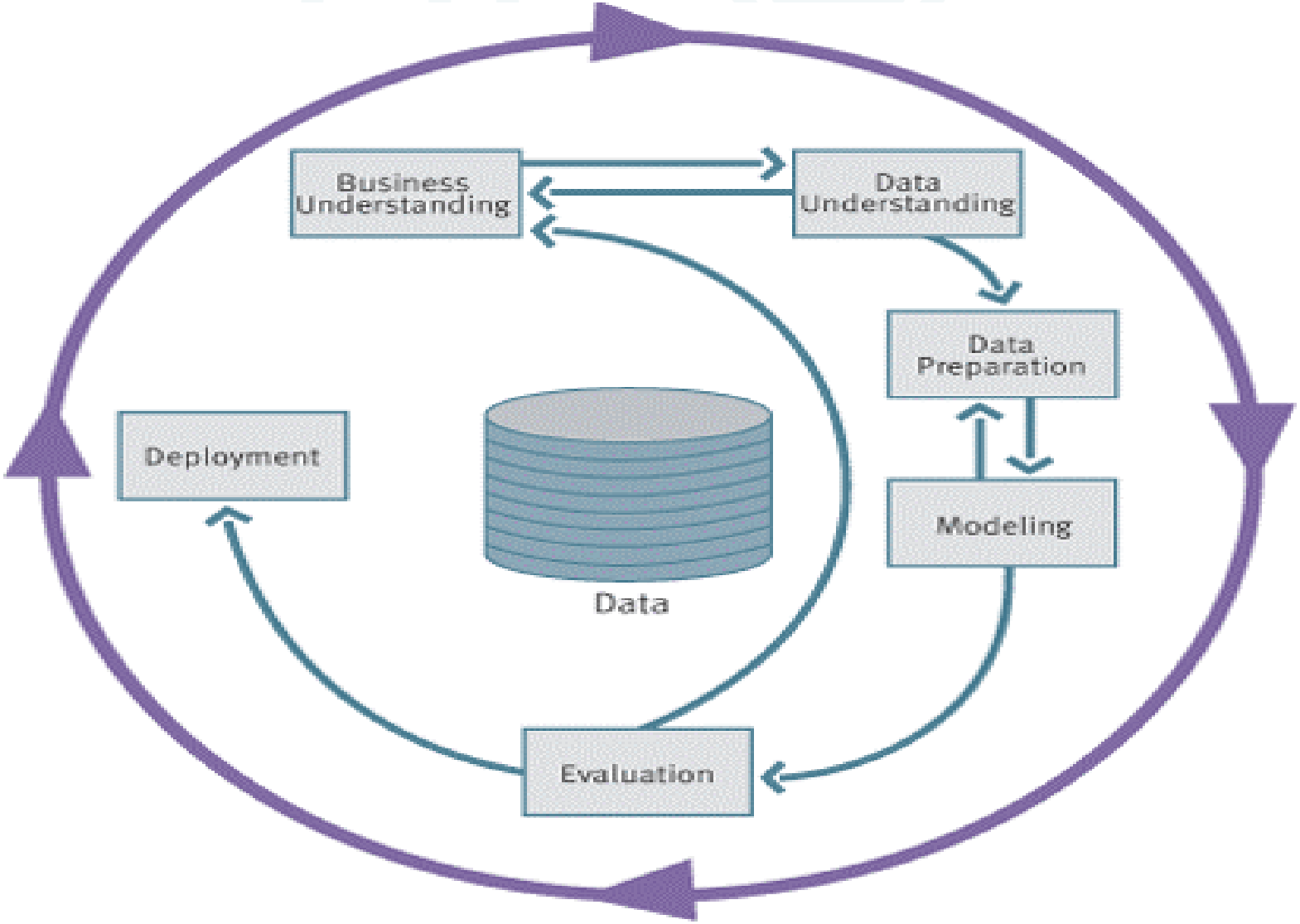
Sources https://www.datascience-pm.com/crisp-dm-still-most-popular/ and https://www.kdnuggets.com/ - November 2020

Frameworks – most common process models

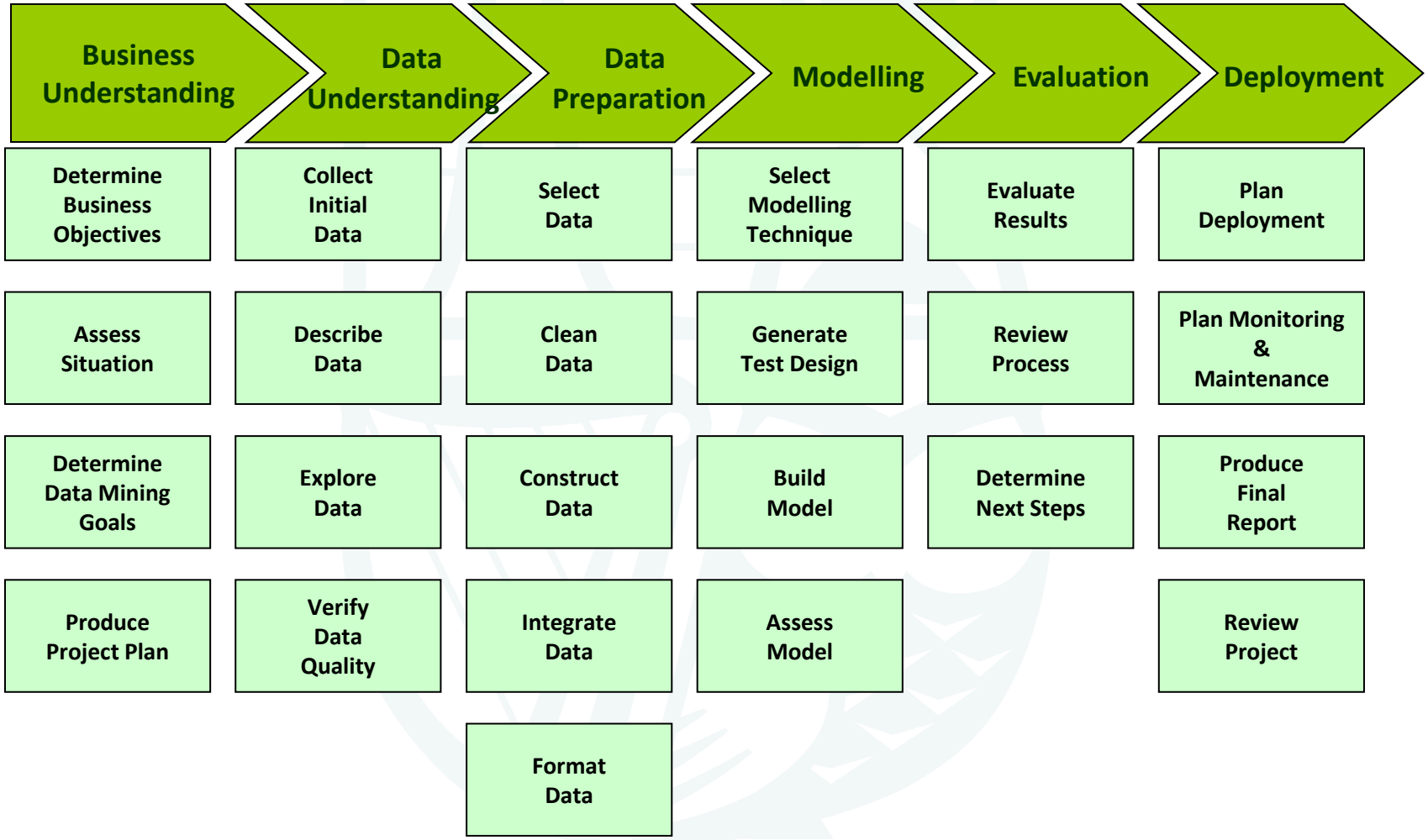


Sources https://www.datascience-pm.com/crisp-dm-still-most-popular/ and https://www.kdnuggets.com/

CRISP-DM – Process model/framework



Framework – CRISP-DM phases & tasks



CRISP-DM – Hierarchical model

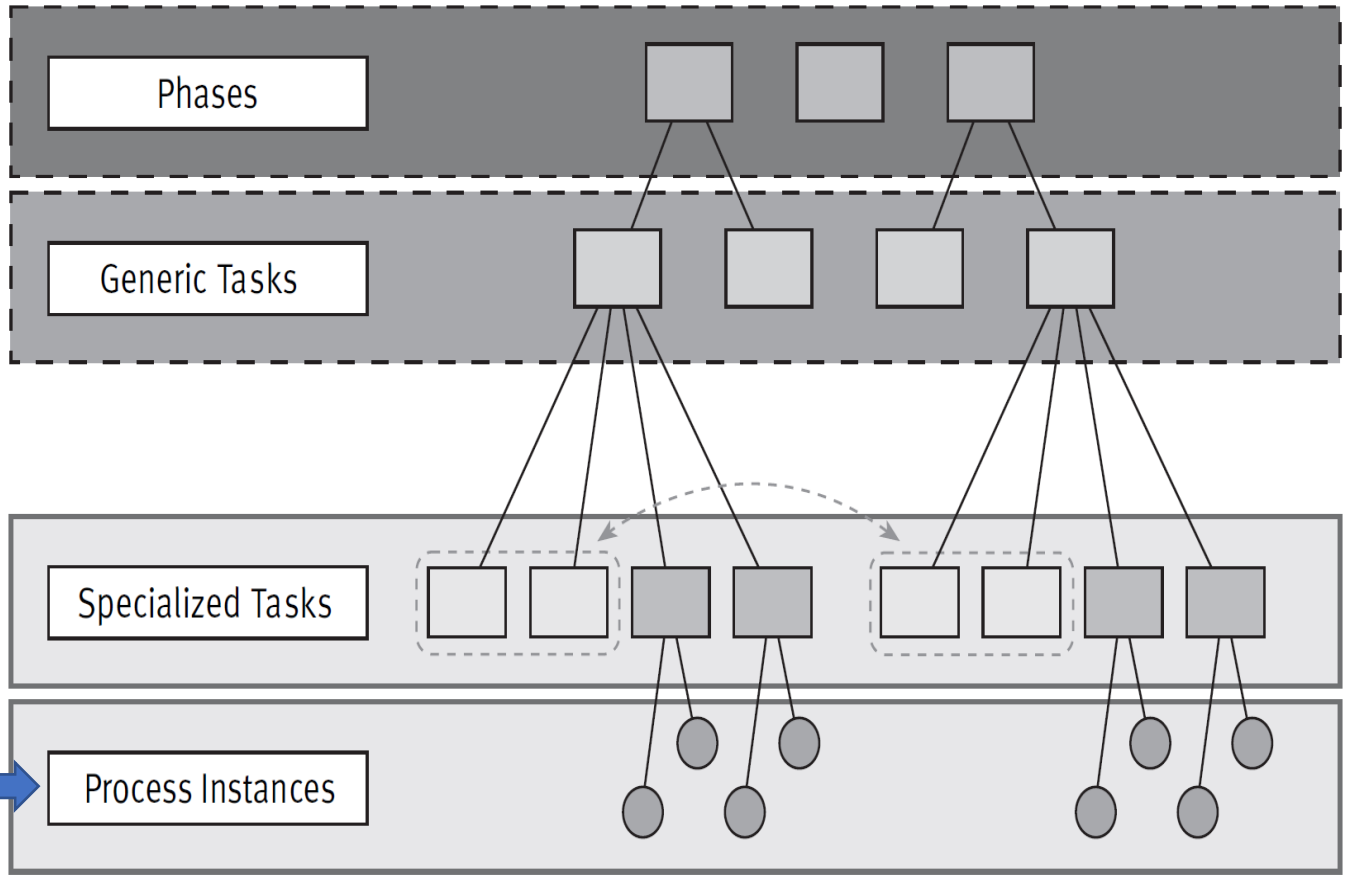


Figure 1: Four level breakdown of the CRISP-DM methodology

Abstract

CRISP Process Model

Mapping

CRISP Process

Specific

Record of actions, decisions, results of an actual data mining engagement

Framework – CRISP-DM tasks & outputs

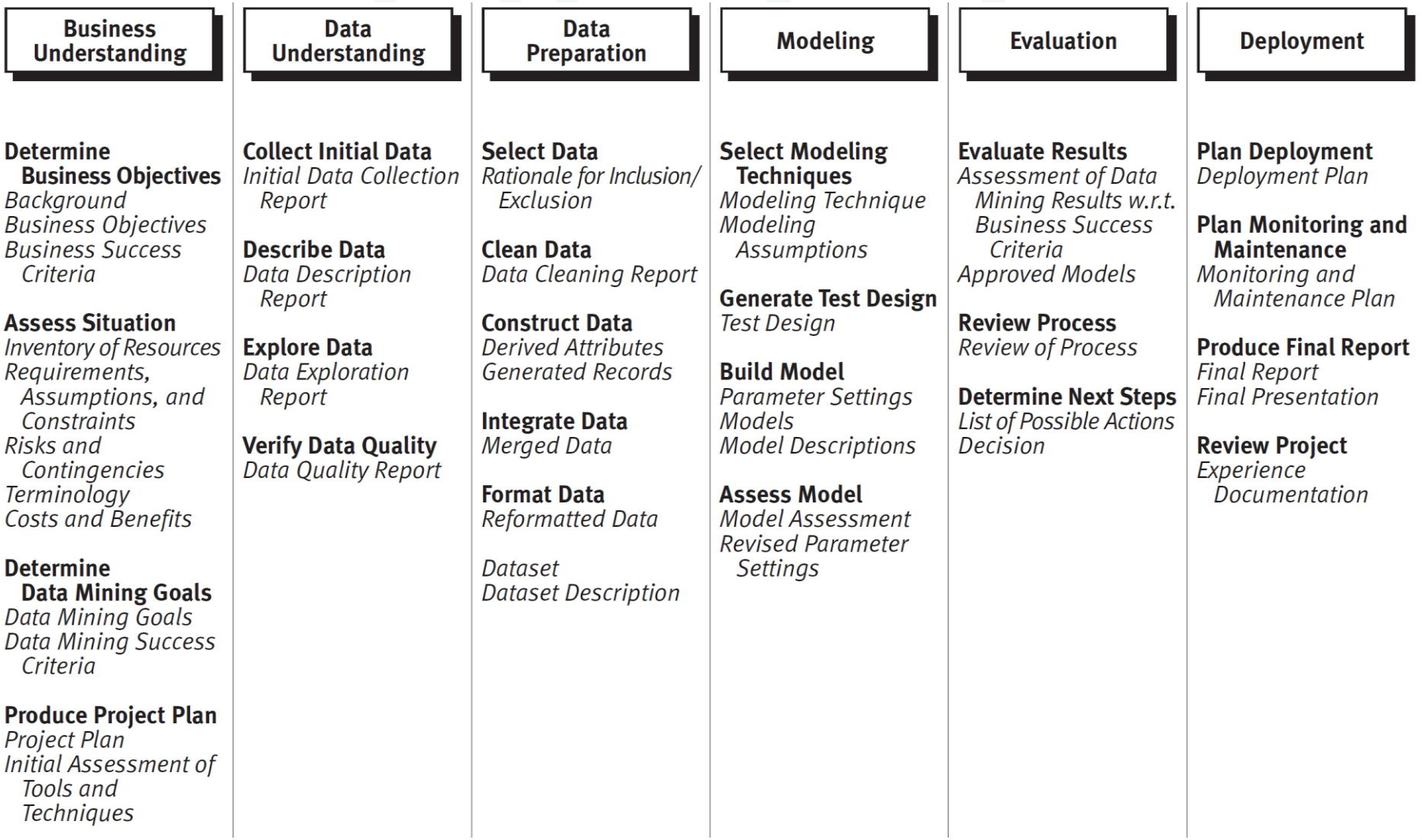
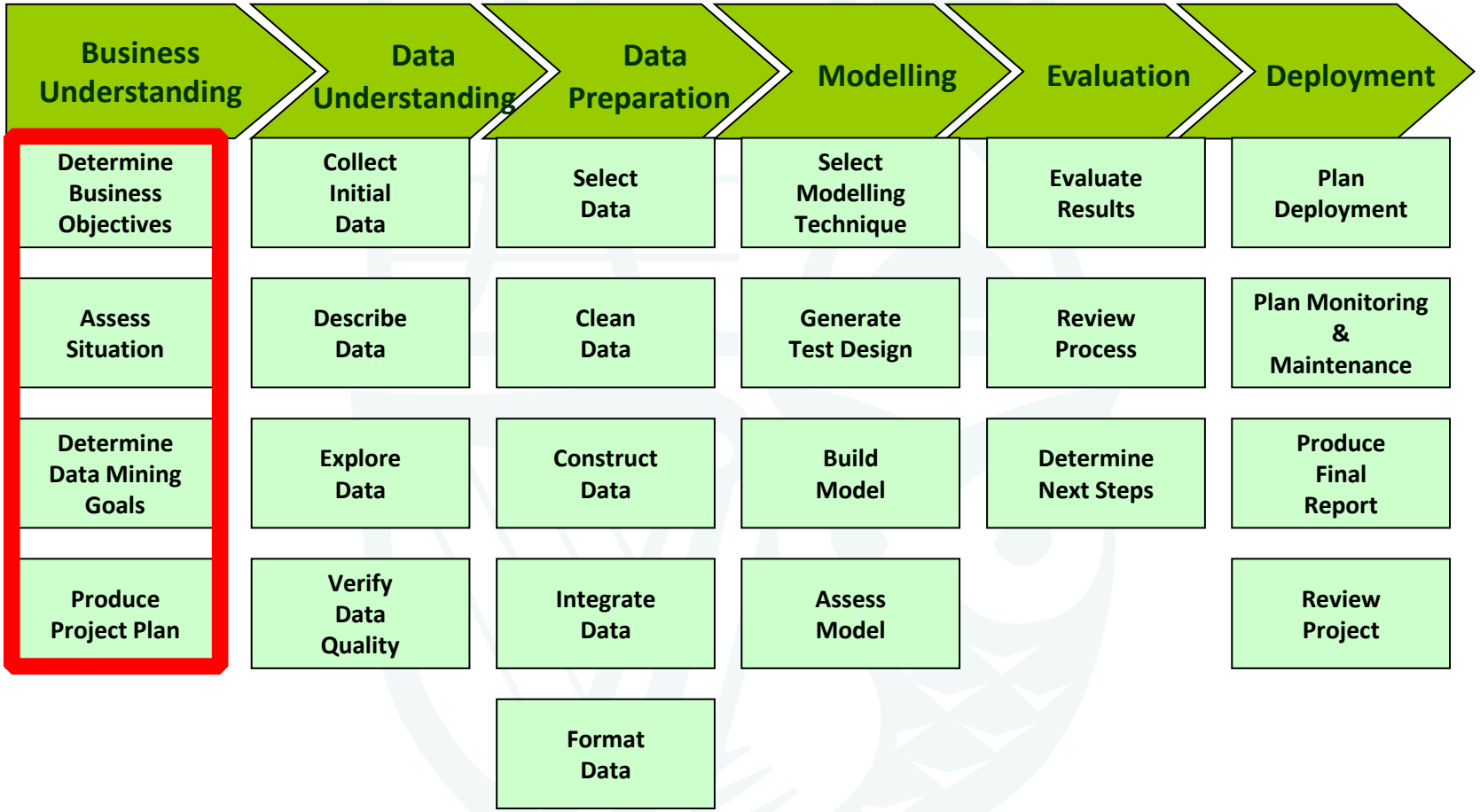
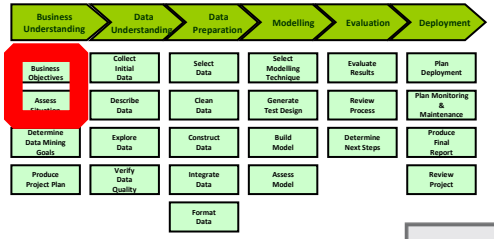


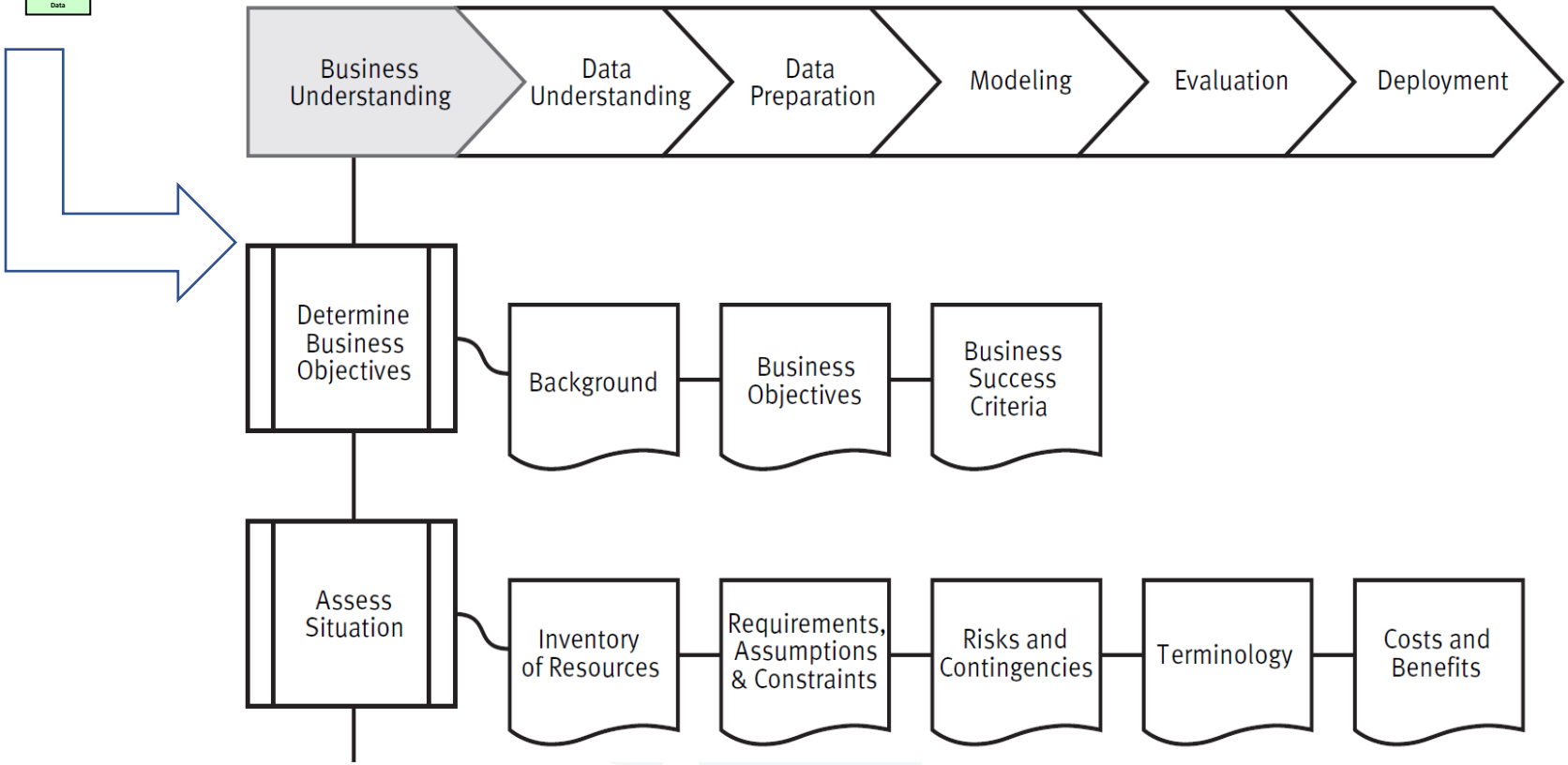
Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

CRISP-DM – BU tasks



CRISP-DM – BU tasks & outputs





CRISP-DM – BU ref. model task description

1.1 Determine business objectives

Task

Determine business objectives

The first objective of the data analyst is to thoroughly understand, from a business perspective, what the customer really wants to accomplish. Often the customer has many competing objectives and constraints that must be properly balanced. The analyst's goal is to uncover important factors, at the beginning, that can influence the outcome of the project. A possible consequence of neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions.

Outputs

Background

Record the information that is known about the organization's business situation at the beginning of the project.

Business objectives

Describe the customer's primary objective, from a business perspective. In addition to the primary business objective, there are typically other related business questions that the customer would like to address. For example, the primary business goal might be to keep current customers by predicting when they are prone to move to a competitor. Examples of related business questions are "How does the primary channel used (e.g., ATM, branch visit, Internet) affect whether customers stay or go?" or "Will lower ATM fees significantly reduce the number of high-value customers who leave?"

Business success criteria

Describe the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific and able to be measured objectively, for example, reduction of customer churn to a certain level, or it might be general and subjective, such as "give useful insights into the relationships." In the latter case, it should be indicated who makes the subjective judgment.

1.2 Assess situation

Task

Assess situation

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and project plan. In the previous task, your objective is to quickly get to the crux of the situation. Here, you want to expand upon

CRISP-DM – BU user guide task details

1.1 Determine business objectives

Task

Determine business objectives

The first objective of the analyst is to thoroughly understand, from a business perspective, what the customer really wants to accomplish. Often the customer has many competing objectives and constraints that must be properly balanced. The analyst's goal is to uncover important factors at the beginning of the project that can influence the final outcome. A likely consequence of neglecting this step would be to expend a great deal of effort producing the correct answers to the wrong questions.

Output

Background

Collate the information that is known about the organization's business situation at the start of the project. These details not only serve to more closely identify the business goals to be achieved but also serve to identify resources, both human and material, that may be used or needed during the course of the project.

Activities

Organization

- Develop organizational charts identifying divisions, departments, and project groups. The chart should also identify managers' names and responsibilities
- Identify key persons in the business and their roles
- Identify an internal sponsor (financial sponsor and primary user/domain expert)
- Indicate if there is a steering committee and list members
- Identify the business units which are affected by the data mining project (e.g., Marketing, Sales, Finance)

Problem area

- Identify the problem area (e.g., marketing, customer care, business development, etc.)
- Describe the problem in general terms
- Check the current status of the project (e.g., Check if it is already clear within the business unit that a data mining project is to be performed, or whether data mining needs to be promoted as a key technology in the business)

CRISP-DM – BU user guide task details

- Identify target groups for the project result (e.g., Are we expected to deliver a report for top management or an operational system to be used by naive end users?)
- Identify the users' needs and expectations

Current solution

- Describe any solution currently used to address the problem
- Describe the advantages and disadvantages of the current solution and the level to which it is accepted by the users

Output

Business objectives

Describe the customer's primary objective, from a business perspective. In addition to the primary business objective, there are typically a large number of related business questions that the customer would like to address. For example, the primary business goal might be to keep current customers by predicting when they are prone to move to a competitor, while a secondary business objective might be to determine whether lower fees affect only one particular segment of customers.

Activities

- Informally describe the problem to be solved
- Specify all business questions as precisely as possible
- Specify any other business requirements (e.g., the business does not want to lose any customers)
- Specify expected benefits in business terms

Beware!

- Beware of setting unattainable goals—make them as realistic as possible.

Output

Business success criteria

Describe the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific and readily measurable, such as reduction of customer churn to a certain level, or general and subjective, such as "give useful insights into the relationships." In the latter case, be sure to indicate who would make the subjective judgment.

Activities

- Specify business success criteria (e.g., Improve response rate in a mailing campaign by 10 percent and

CRISP-DM – BU user guide task details

Remember! Each of the success criteria should relate to at least one of the specified business objectives.

Good idea! Before starting the situation assessment, you might analyze previous experiences of this problem—either internally, using CRISP-DM, or externally, using pre-packaged solutions.

1.2 Assess situation

Task **Assess situation**

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and in developing the project plan.

Output **Inventory of resources**

List the resources available to the project, including personnel (business and data experts, technical support, data mining experts), data (fixed extracts, access to live warehoused or operational data), computing resources (hardware platforms), and software (data mining tools, other relevant software).

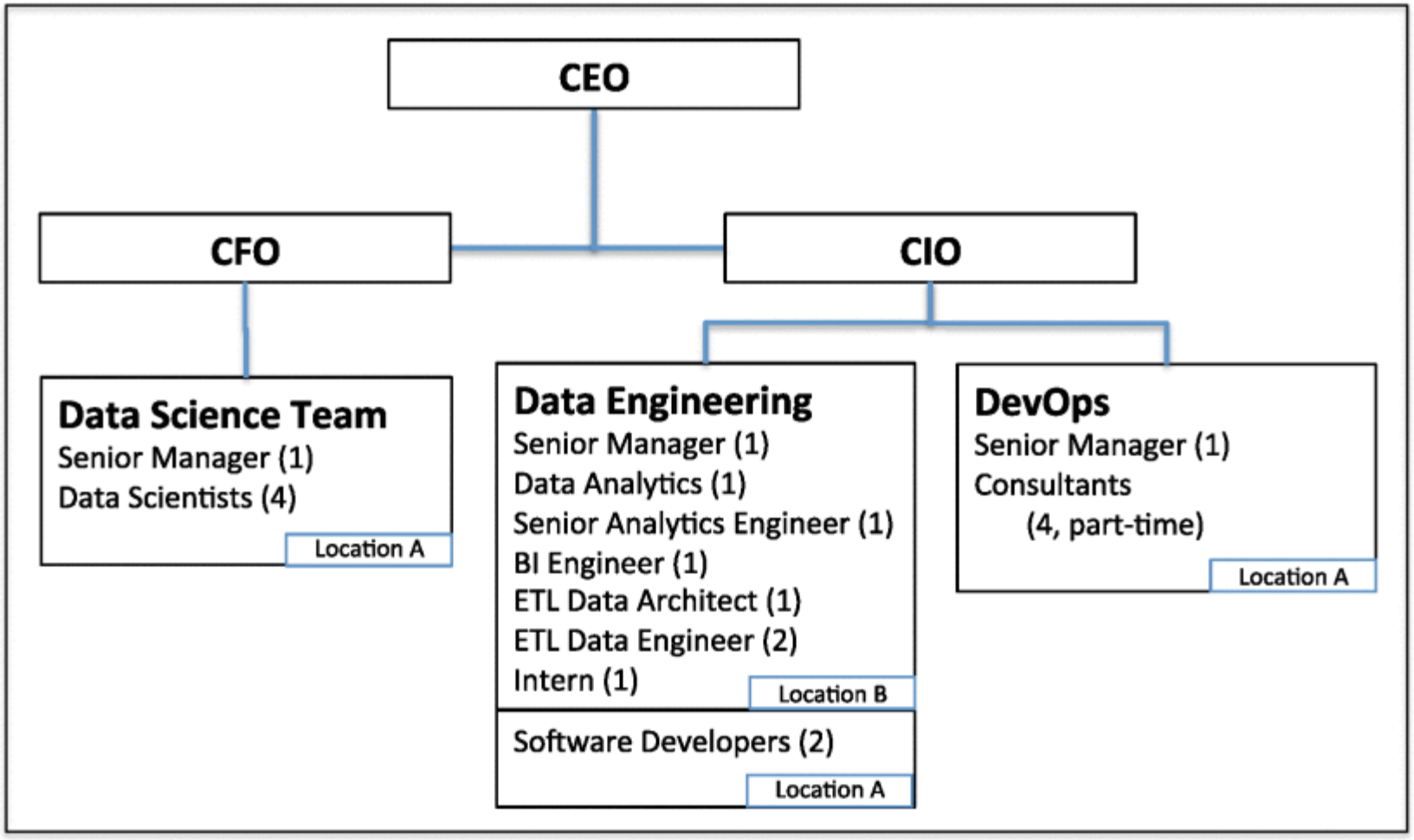
Activities **Hardware resources**

- Identify the base hardware
- Establish the availability of the base hardware for the data mining project
- Check if the hardware maintenance schedule conflicts with the availability of the hardware for the data mining project
- Identify the hardware available for the data mining tool to be used (if the tool is known at this stage)

Sources of data and knowledge

- Identify data sources
- Identify type of data sources (online sources, experts, written documentation, etc.)
- Identify knowledge sources
- Identify type of knowledge sources (online sources, experts, written documentation, etc.)
- Check available tools and techniques
- Describe the relevant background knowledge (informally or formally)

CRISP-DM - Structure of the Big Data Science Team

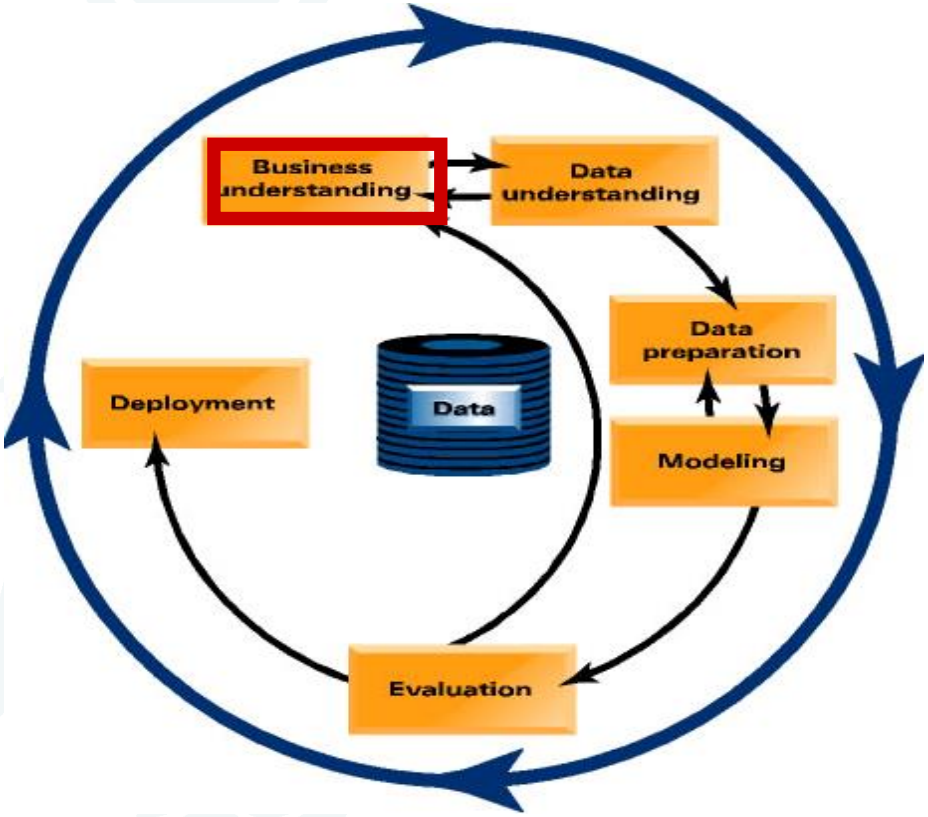


Source: J. S. Saltz and I. Shamshurin, "Achieving Agile Big Data Science: The Evolution of a Team's Agile Process Methodology," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 3477-3485, doi: 10.1109/BigData47090.2019.9005493.

DM Process - Phase 1 BU

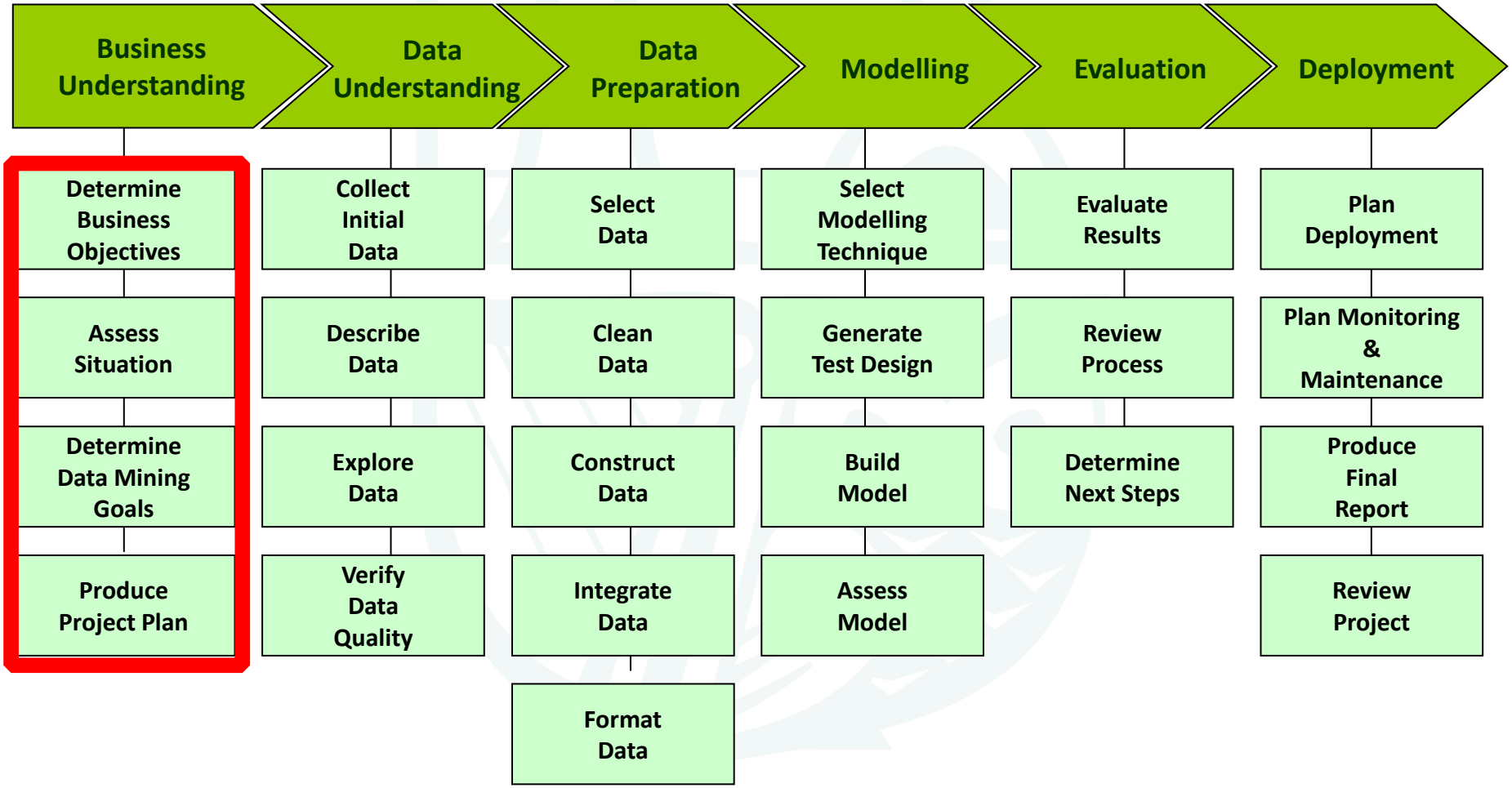
1. Business Understanding

- Statement of business objective
- Statement of data mining objective
- Statement of success criteria



Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives

CRISP-DM: Phase 1 Business Understanding



DM Process - Phase 1 BU

1. Determine business objectives

- Understand in detail, from a business perspective, what the client(coach) wants to achieve
- Discover important factors, at the start, that can influence the outcome of the project
- Ignoring this step wastes a huge amount of effort producing the **correct answers** to the **wrong questions**

2. Assess situation

- Detailed fact-finding about all of the resources, constraints, assumptions and other factors that should be considered
- Elaborate on the specific details

Remember your business!!!

DM Process - BU "Assess situation" reference model

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

DM Process - BU “Assess situation” reference model

1.2 Assess situation

Task	**Assess situation**
	This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and project plan. In the previous task, your objective is to quickly get to the crux of the situation. Here, you want to expand upon the details.

DM Process - BU "Assess situation" reference model

Outputs

Inventory of resources

List the resources available to the project, including personnel (business experts, data experts, technical support, data mining experts), data (fixed extracts, access to live, warehoused, or operational data), computing resources (hardware platforms), and software (data mining tools, other relevant software).

Requirements, assumptions, and constraints

List all requirements of the project, including schedule of completion, comprehensibility and quality of results, and security, as well as legal issues. As part of this output, make sure that you are allowed to use the data.

List the assumptions made by the project. These may be assumptions about the data that can be verified during data mining, but may also include non-verifiable assumptions about the business related to the project. It is particularly important to list the latter if it will affect the validity of the results.

List the constraints on the project. These may be constraints on the availability of resources, but may also include technological constraints such as the size of dataset that it is practical to use for modeling.

Risks and contingencies

List the risks or events that might delay the project or cause it to fail. List the corresponding contingency plans, what action will be taken if these risks or events take place.

Terminology

Compile a glossary of terminology relevant to the project. This may include two components:

(1) A glossary of relevant business terminology, which forms part of the business understanding available to the project. Constructing this glossary is a useful "knowledge elicitation" and education exercise.

(2) A glossary of data mining terminology, illustrated with examples relevant to the business problem in question

Costs and benefits

Construct a cost-benefit analysis for the project, which compares the costs of the project with the potential benefits to the business if it is successful. The comparison should be as specific as possible. For example, use monetary measures in a commercial situation.

DM Process - BU "Assess situation" user guide

1.2 Assess situation

Task

Assess situation

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data analysis goal and in developing the project plan.

Output

Inventory of resources

List the resources available to the project, including personnel (business and data experts, technical support, data mining experts), data (fixed extracts, access to live warehoused or operational data), computing resources (hardware platforms), and software (data mining tools, other relevant software).

Activities

Hardware resources

- Identify the base hardware
- Establish the availability of the base hardware for the data mining project
- Check if the hardware maintenance schedule conflicts with the availability of the hardware for the data mining project
- Identify the hardware available for the data mining tool to be used (if the tool is known at this stage)

Sources of data and knowledge

- Identify data sources
- Identify type of data sources (online sources, experts, written documentation, etc.)
- Identify knowledge sources
- Identify type of knowledge sources (online sources, experts, written documentation, etc.)
- Check available tools and techniques
- Describe the relevant background knowledge (informally or formally)

Personnel sources

- Identify project sponsor (if different from internal sponsor as in Section 1.1.1)
- Identify system administrator, database administrator, and technical support staff for further questions
- Identify market analysts, data mining experts, and statisticians, and check their availability
- Check availability of domain experts for later phases

DM Process - BU "Assess situation" user guide

Remember!

Remember that the project may need technical staff at odd times throughout the project, for example during data transformation.

Output

Requirements, assumptions, and constraints

List all requirements of the project, including schedule of completion, comprehensibility, and quality of results and security, as well as legal issues. As part of this output, make sure that you are allowed to use the data.

List the assumptions made by the project. These may be assumptions about the data, which can be verified during data mining, but may also include non-verifiable assumptions related to the project. It is particularly important to list the latter if they will affect the validity of the results.

List the constraints made on the project. These constraints might involve lack of resources to carry out some of the tasks in the project in the time required, or there may be legal or ethical constraints on the use of the data or the solution needed to carry out the data mining task.

Activities

Requirements

- Specify target group profile
- Capture all requirements on scheduling
- Capture requirements on comprehensibility, accuracy, deploy ability, maintainability, and repeatability of the data mining project and the resulting model(s)
- Capture requirements on security, legal restrictions, privacy, reporting, and project schedule

Assumptions

- Clarify all assumptions (including implicit ones) and make them explicit (e.g., to address the business question, a minimum number of customers with age above 50 is necessary)
- List assumptions on data quality (e.g., accuracy, availability)
- List assumptions on external factors (e.g., economic issues, competitive products, technical advances)
- Clarify assumptions that lead to any of the estimates (e.g., the price of a specific tool is assumed to be lower than $1,000)
- List all assumptions regarding whether it is necessary to understand and describe or explain the model (e.g., how should the model and results be presented to senior management/sponsor)

DM Process - Phase 1 BU

3. Determine data mining goals

- a business goal states objectives in business terminology
- a data mining goal states project objectives in technical terms

- Business goal - "Increase catalog sales to existing customers."

- Data mining goal - "Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city) and the price of the item."

- **Exercise** – identify business and data mining goals for your business/area of interest

DM Process - Phase 1 BU

- Sales made or the likelihood of a sale being made?
- Sales lead identified/followed or an easy sale made?
- Do salespersons vie for stats yes/no? Why? Is this good/bad?
- Can individual stats help the sales team, yes/no?
- Identify some important sales team stats
- **Reflection**

DM Process - Phase 1 BU

- Business goal - “Increase points scored in trips to the Red Zone.”

- Data mining goal - “Predict how many lineouts a player will win, based on their lineout wins over the past three years, location information (lineouts won in each zone) and which team’s lineout it is.”

- **Consider** Are these appropriate goals?

- **Reflection/discussion**

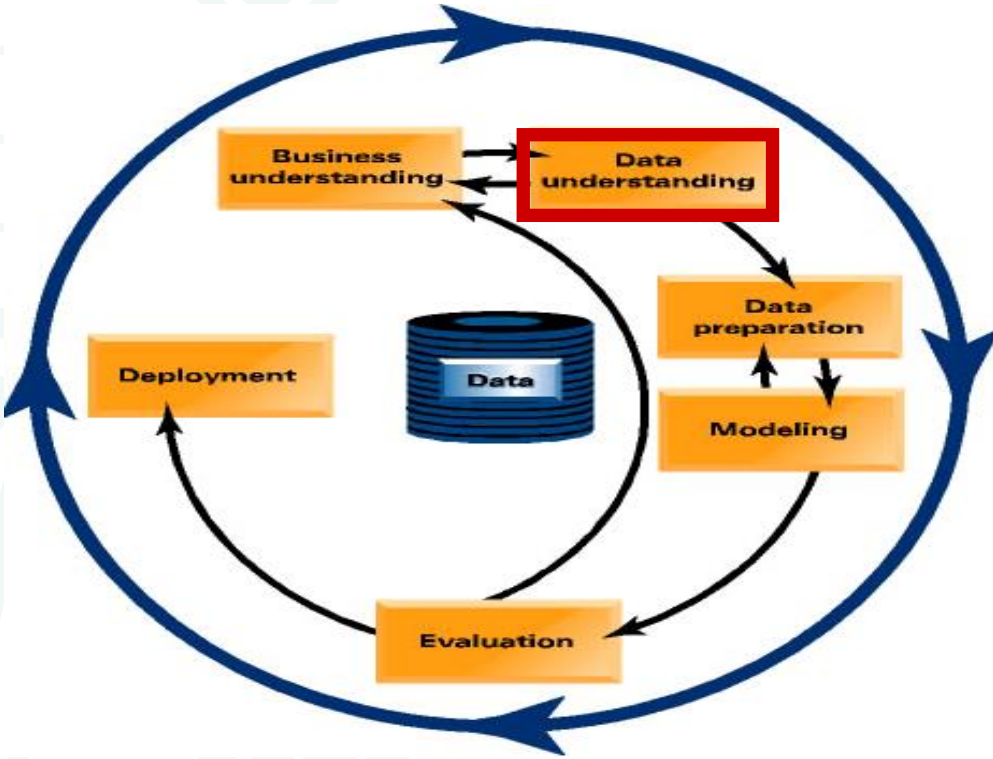
DM Process - Phase 1 BU

- **4. Produce project plan**
  - Describe the intended plan for achieving the data mining goals and the business goals
  - In the plan specify the set of steps to be performed during the rest of the project including an initial selection of tools and techniques

DM Process - Phase 2 DU

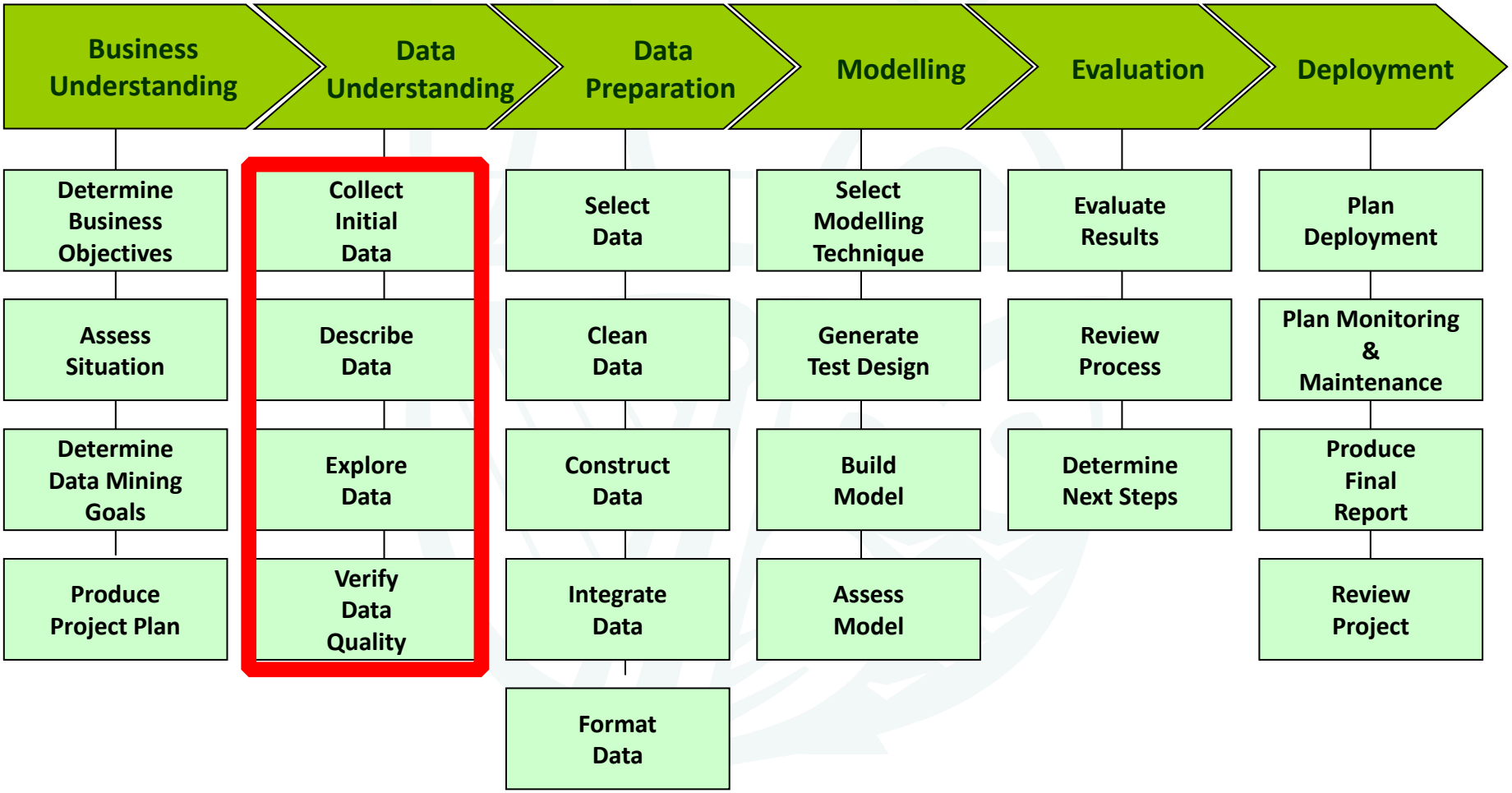
2. Data Understanding

- Collect data
- Describe data
- Explore the data
- Verify the quality and identify outliers



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

CRISP-DM: Phase 2 Data Understanding



DM Process - Phase 2 DU

- **1. Collect initial data**
  - Acquire within the project the data listed in the project resources
  - Includes data loading if necessary for data understanding
  - Possible that this leads to initial data preparation steps
  - If acquiring multiple data sources, integration is an additional issue, either here or in the later data preparation phase

- **2. Describe data**
  - Examine the “gross” or “surface” **properties** of the acquired data
  - Report on the results
  - **Exercise –** what other properties might the data have?

DM Process - Phase 2 DU

- **3. Explore data** - tackles the possible data mining questions, which can be addressed using querying, visualisation and reporting including:
  - Distribution of key attributes, results of simple aggregations relations between pairs or small numbers of attributes properties of significant sub-populations, simple statistical analyses
  - May address directly the data mining goals
  - May contribute to, or refine the data description and quality reports
  - May feed into the transformation and other data preparation needed

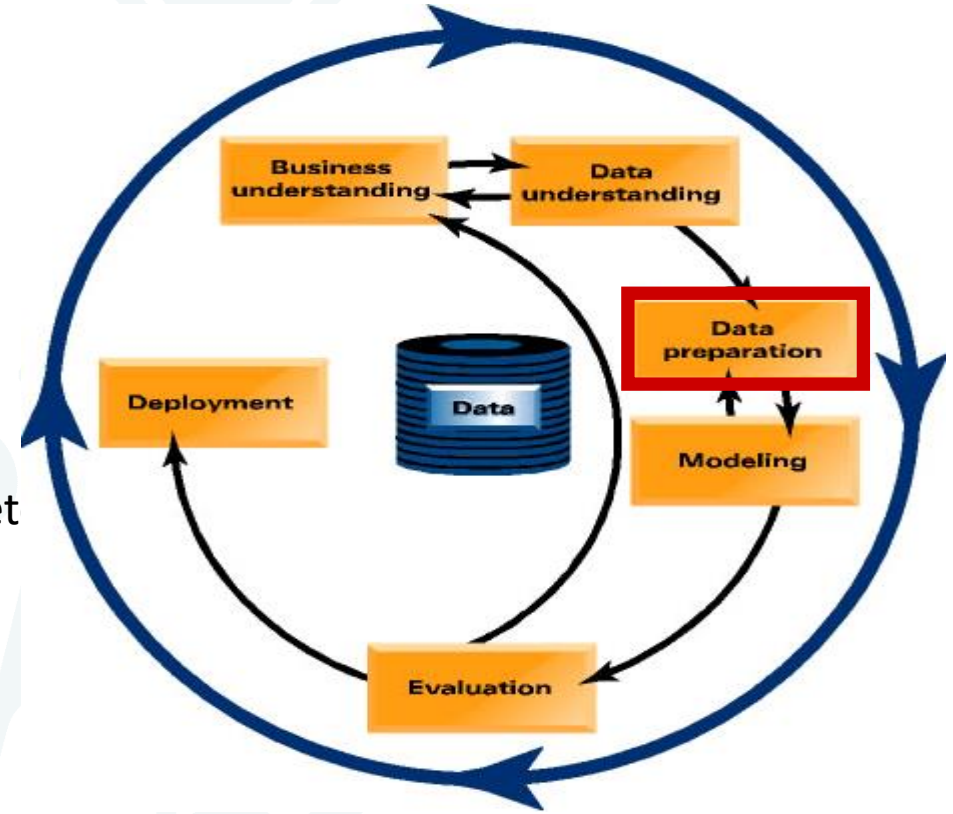
- **4. Verify data quality** - examine the quality of the data, including:
  - Is the data complete?
  - Are there missing/obviously incorrect values in the data?
  - **Exercise** - Identify other potential data issues

DM Process - Phase 3 DP

3. Data Preparation

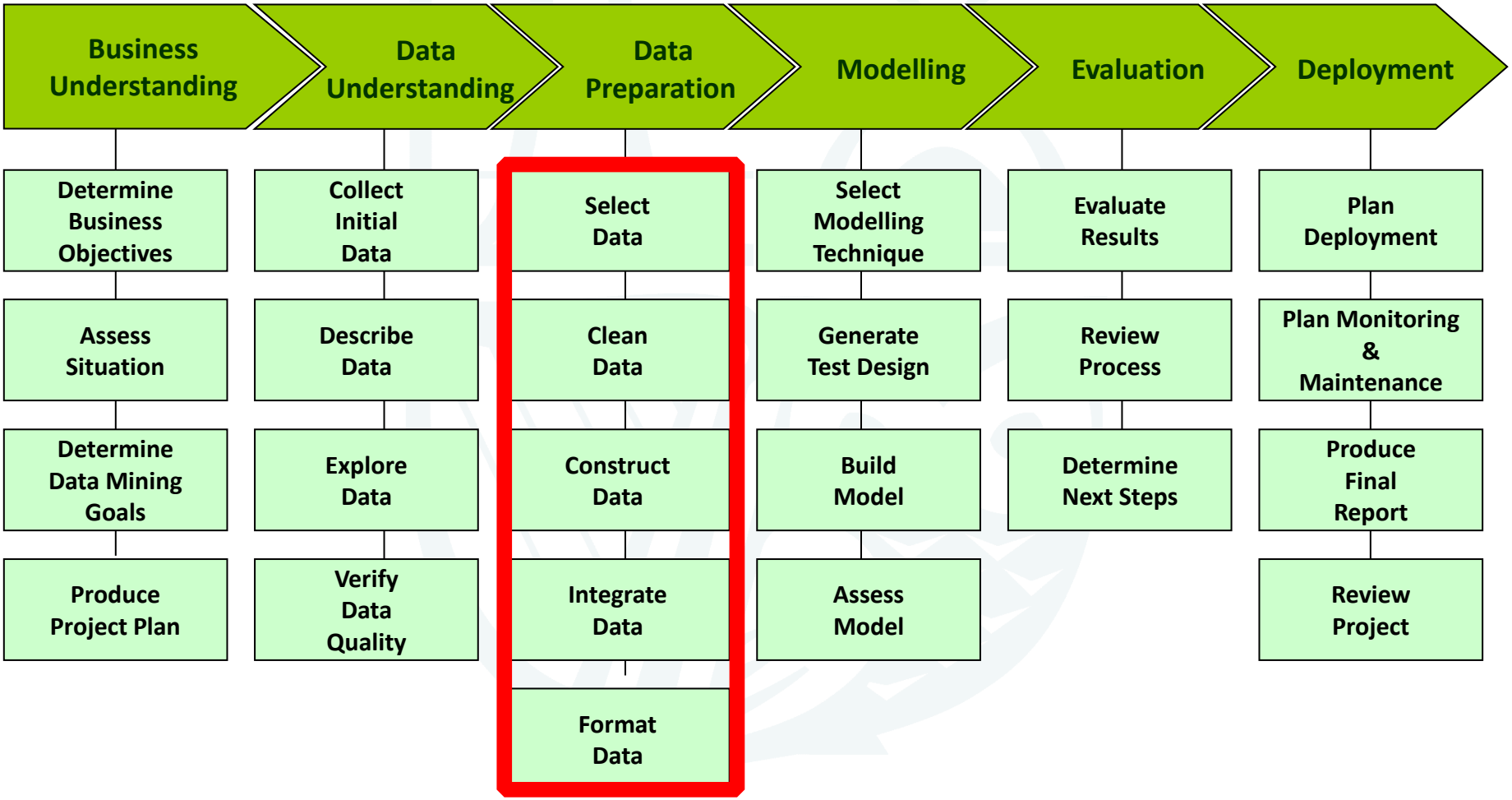
Typically 90% of time taken on this phase

- Collection
- Assessment
- Consolidation and Cleaning
- Data selection
  - Remove “noisy” data, repetitions, et
  - Remove outliers?
  - Select samples
  - visualisation tools
- Transformations - variables, formats



Covers all activities to construct the final dataset 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 modelling tools.

CRISP-DM: Phase 3 Data Preparation



DM Process - Phase 3 DP

- **1. Select data**
  - Decide on the data to be used for analysis
  - Criteria include relevance to the data mining goals, quality and technical constraints such as limits on data volume or data types
  - Covers selection of attributes as well as selection of records in a table

- **2. Clean data**
  - Raise the data quality to the level required by the selected analysis techniques
  - May involve selection of clean subsets of the data, the insertion of suitable defaults or more ambitious techniques such as the estimation of missing data by modelling

DM Process - Phase 3 DP

- **3. Construct data**
  - Constructive data preparation operations such as the production of derived attributes, entire new records or transformed values for existing attributes
- **4. Integrate data**
  - Methods where information is combined from multiple tables or records to create new records or values
- **5. Format data**
  - Formatting transformations refer to primarily syntactic modifications made to the data that do not change its meaning, but might be required by the modelling tool

DM Process - Phase 3 DP

- **Exercise -** Outline a complete example of data preparation that you have completed. Identify an example that includes each step described here, briefly describing each step:

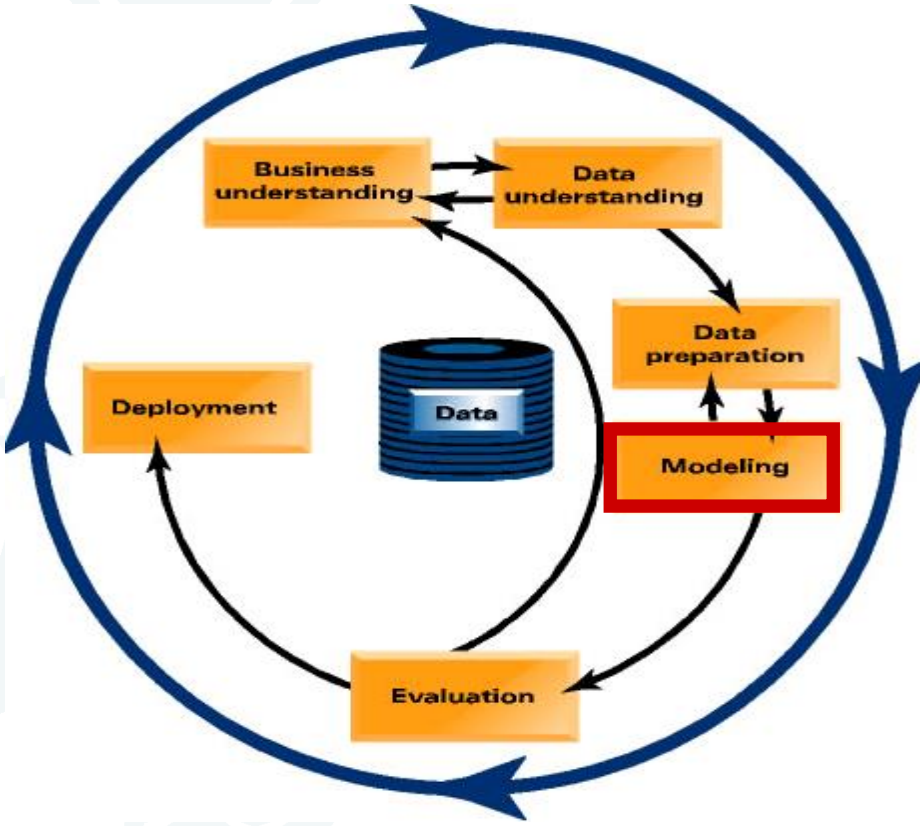
1. Select data
2. Clean data
3. Construct data
4. Integrate data
5. Format data

Discussion

DM Process - Phase 4 MB

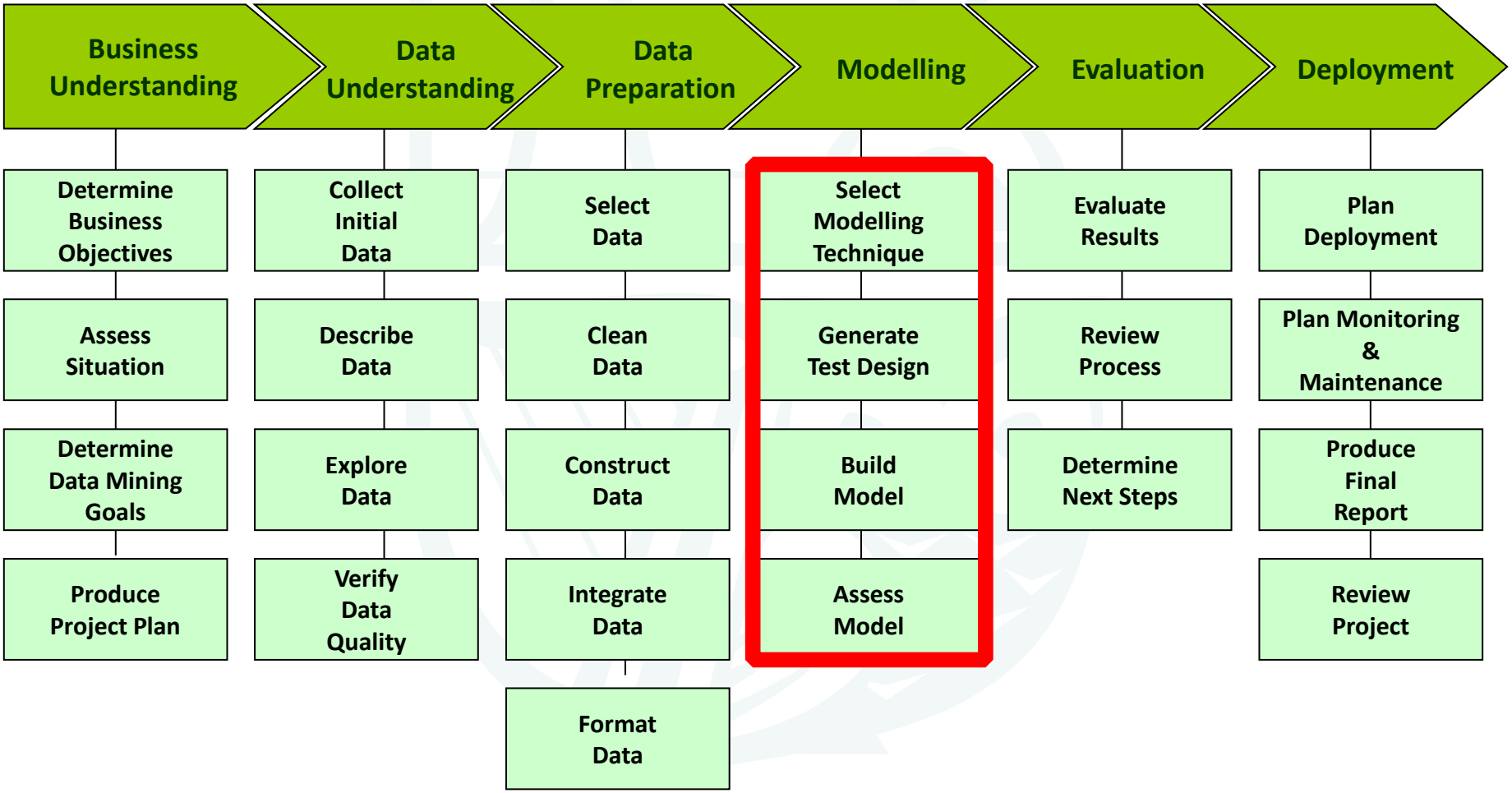
4. Model Building

- Selection of the modelling techniques
  - Based upon the data mining objective(s)
- Build Model
  - Modelling can be an iterative process
  - Model for description or prediction
- Assess Model
  - Rank different models applied



Modelling techniques are selected and applied and their parameters are calibrated to optimal values. Some techniques have specific requirements on the form of data. Reiteration of the data preparation phase is thus often necessary

CRISP-DM: Phase 4 Model Building

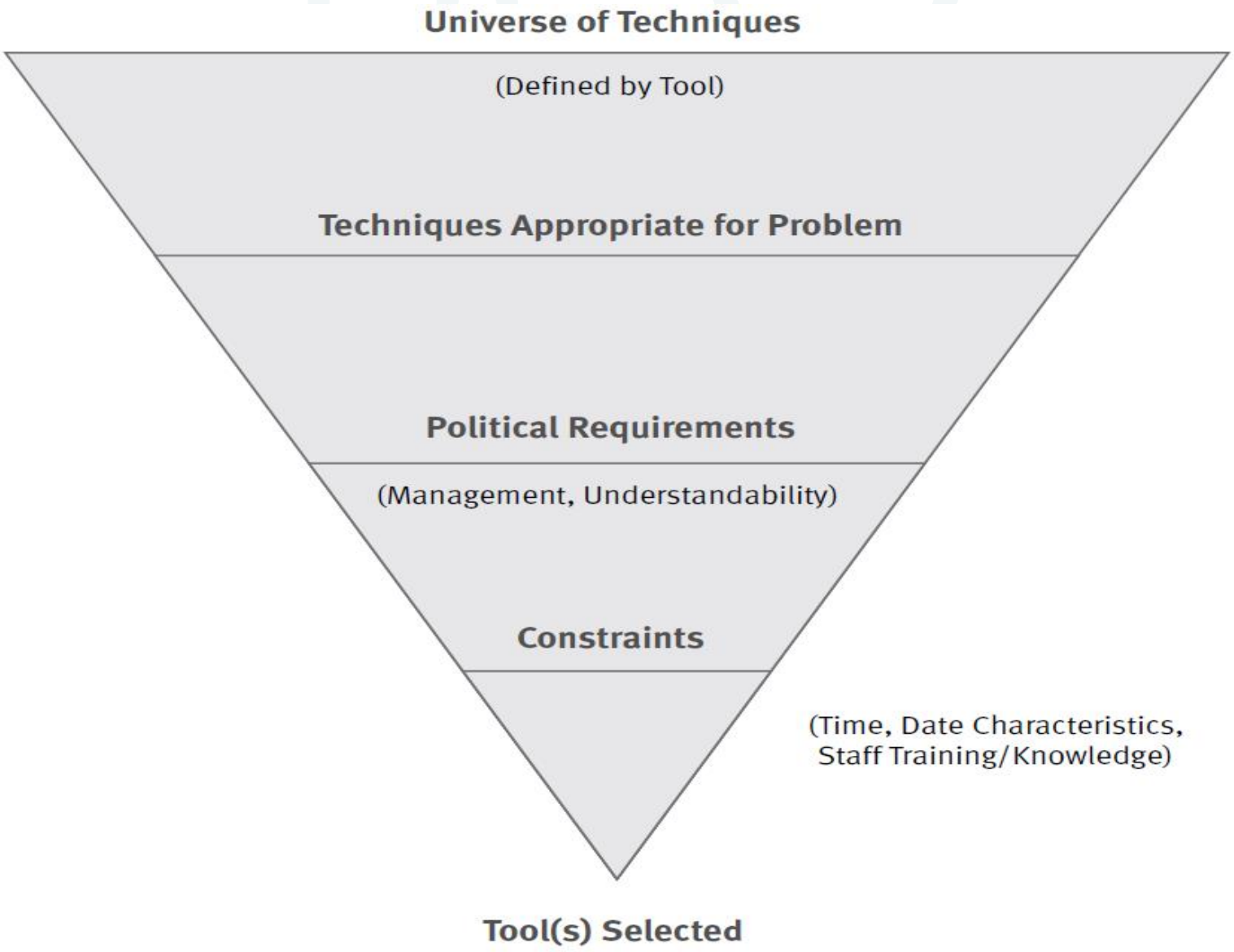


DM Process - Phase 4 Model Building

- **1. Select modelling technique**
  - Select the appropriate modelling technique to be used, for example, decision tree, neural network, simple mathematical/statistics model
  - Where multiple techniques are applied, perform this task for each data mining objective separately

Exercise – Consider/outline common modelling techniques applied in your business

DM Process - Phase 4 Model Building



DM Process - Phase 4 MB

- **Select modelling technique**

As the first step in modelling, select the actual initial modeling technique. If multiple techniques are to be applied, perform this task separately for each technique.

- **Output - Modelling technique** Record the actual modeling technique that is used.
- **Activities -** Decide on appropriate technique for exercise, bearing in mind the tool selected.
- **Output - Modelling assumptions -** Many modeling techniques make specific assumptions about the data.

Activities

- Define any built-in assumptions made by the technique about the data (e.g., quality, format, distribution)
- Compare these assumptions with those in the Data Description Report
- Make sure that these assumptions hold and go back to the Data Preparation Phase, if necessary

DM Process - Phase 4 MB

- **2. Generate test design**
  - Before building a model, generate a procedure or mechanism to test the quality and validity of the model
  - For example, in classification, it is common to use error rates as quality measures for data mining models. Typically we would separate the dataset into training and test data sets and build the model on the training set and estimate/test the model quality on the test set

Exercise – Outline how you test the quality and validity of your data models

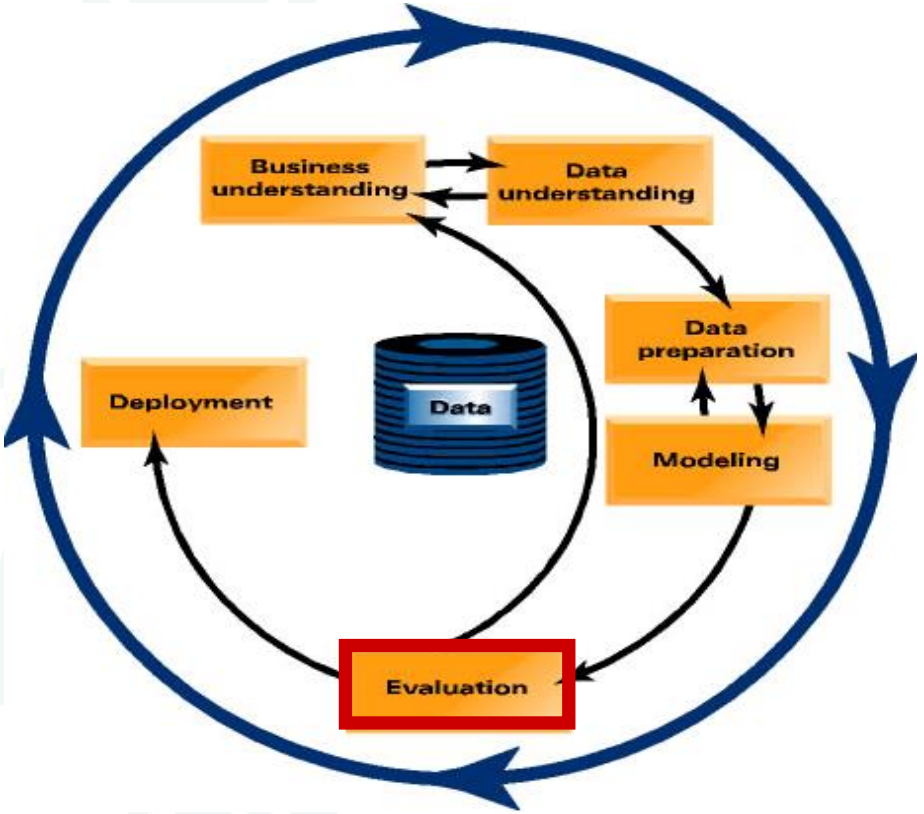
DM Process - Phase 4 MB

- **3. Build model**
  - Run the modelling tool on the prepared dataset to create one or more models that are based on the data mining objective(s)
  - **Exercise** - Identify tools that you use for building your models
- **4. Assess model**
  - Interpret the model according to domain knowledge, the data mining success criteria and the desired test design
  - Technically assess the success of the application of modelling and discovery techniques
  - Discuss the data mining results in the business context (with analysts and domain experts)
  - Consideration of models only (later evaluation phase considers all other results that were produced in the course of the data analytics project)
  - **Exercise** - Consider how you perform the model assessment step

DM Process - Phase 5 ME

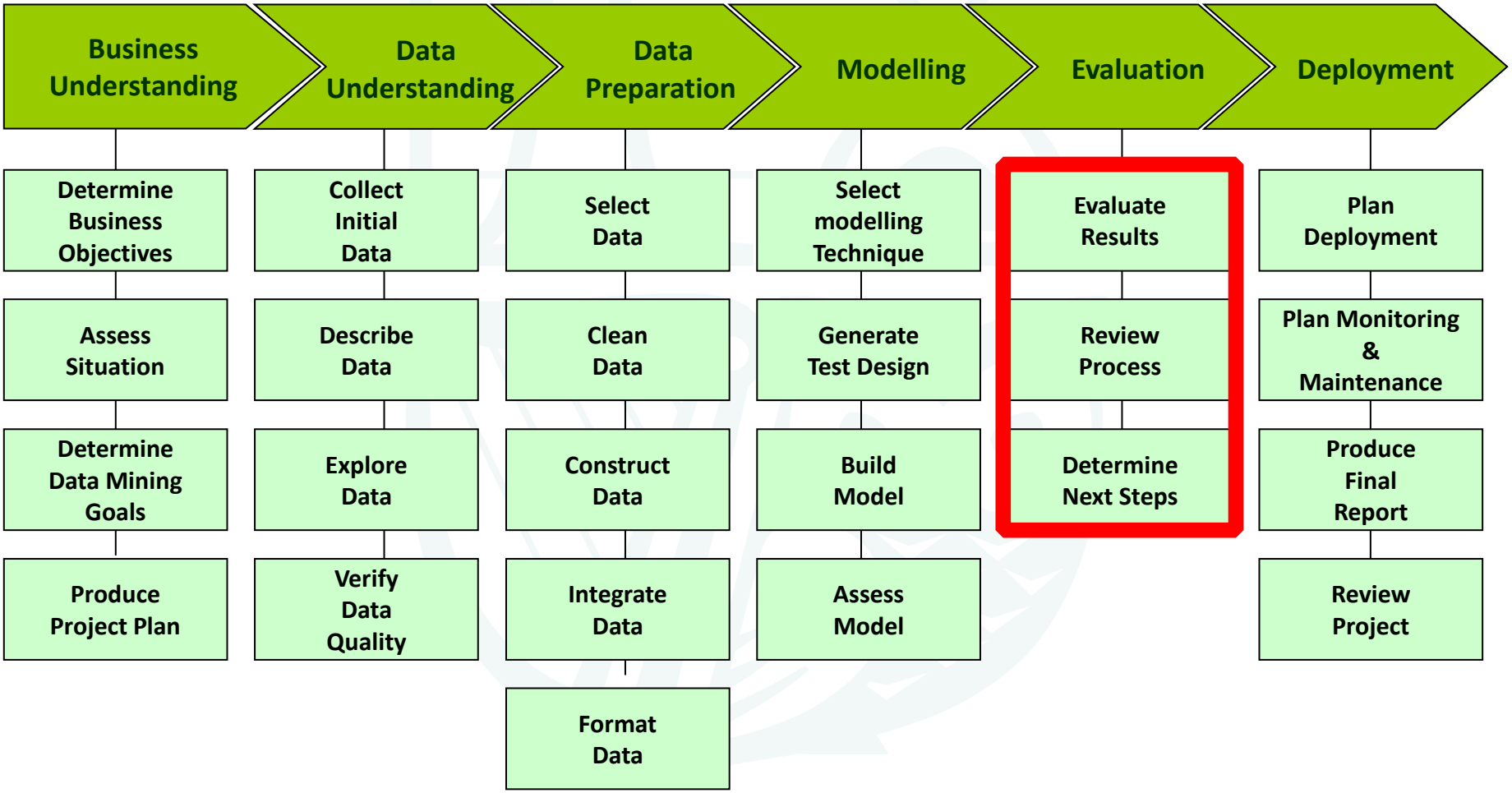
5. Model Evaluation

- Evaluation of model
  - How well the model performed on test data, meets business needs
- Methods and criteria
  - Depends on model type, e.g., confusion matrix with classification models, mean error rate with regression models
- Interpretation of model
  - Important or not, difficulty depends on algorithm, discover reasons why



Evaluate model and review steps executed to construct the model to ensure it properly achieves business objectives. Key objective - determine important business issue(s) 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.

CRISP-DM: Phase 5 Model Evaluation



DM Process - Phase 5 ME

- **1. Evaluate Results**
  - Assess the degree to which the model meets the business objective(s)
  - Determine if there is/are some business reason(s) why the chosen model is deficient
  - Test the model(s) on test application data and on the real application if time and budget constraints permit
  - Assess any additional data mining results generated
  - Identify additional challenges, information or suggestions for future directions

Interpretation - explain results rather than just present them

DM Process - Phase 5 ME

- **2. Review process**
  - Perform thorough review data analytics process in order to determine if there is any important factor or task that has been overlooked
  - Review quality assurance issues, for example, data quality appropriateness of model, model building

- **3. Determine next steps**
  - Decide how to proceed at this stage:

    1) finish the current project and move on to deployment or

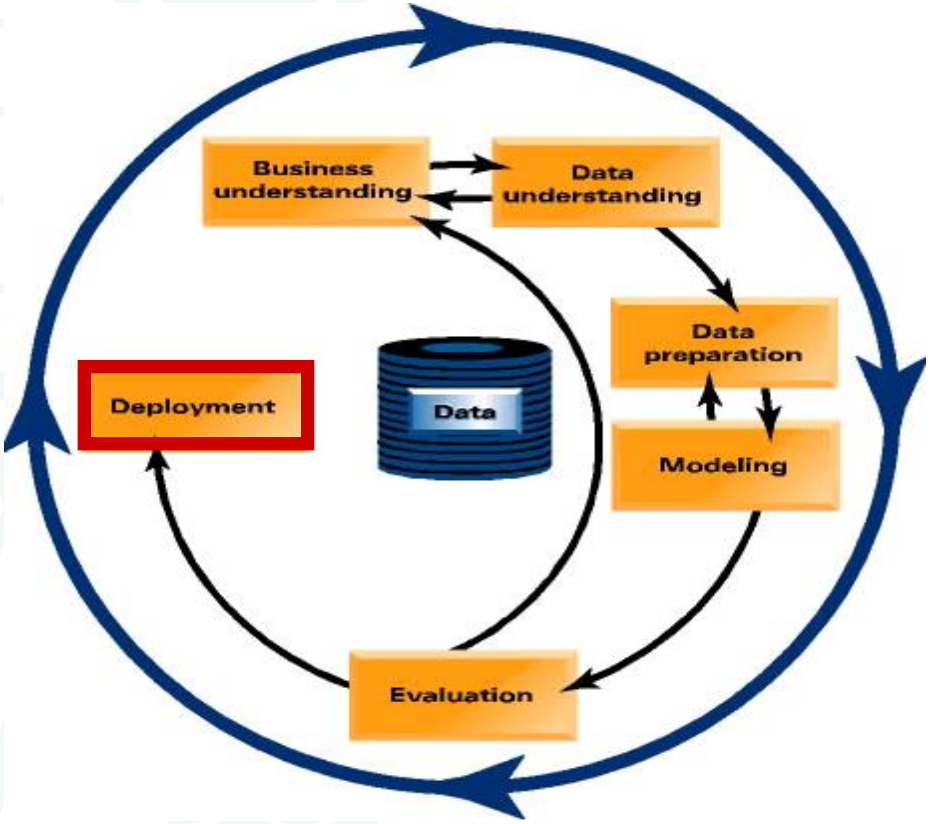
    2) initiate further iterations or

    3) set up new data mining project(s)
  - Include analyses of remaining resources and budget that influences the decisions

DM Process - Phase 6 MD

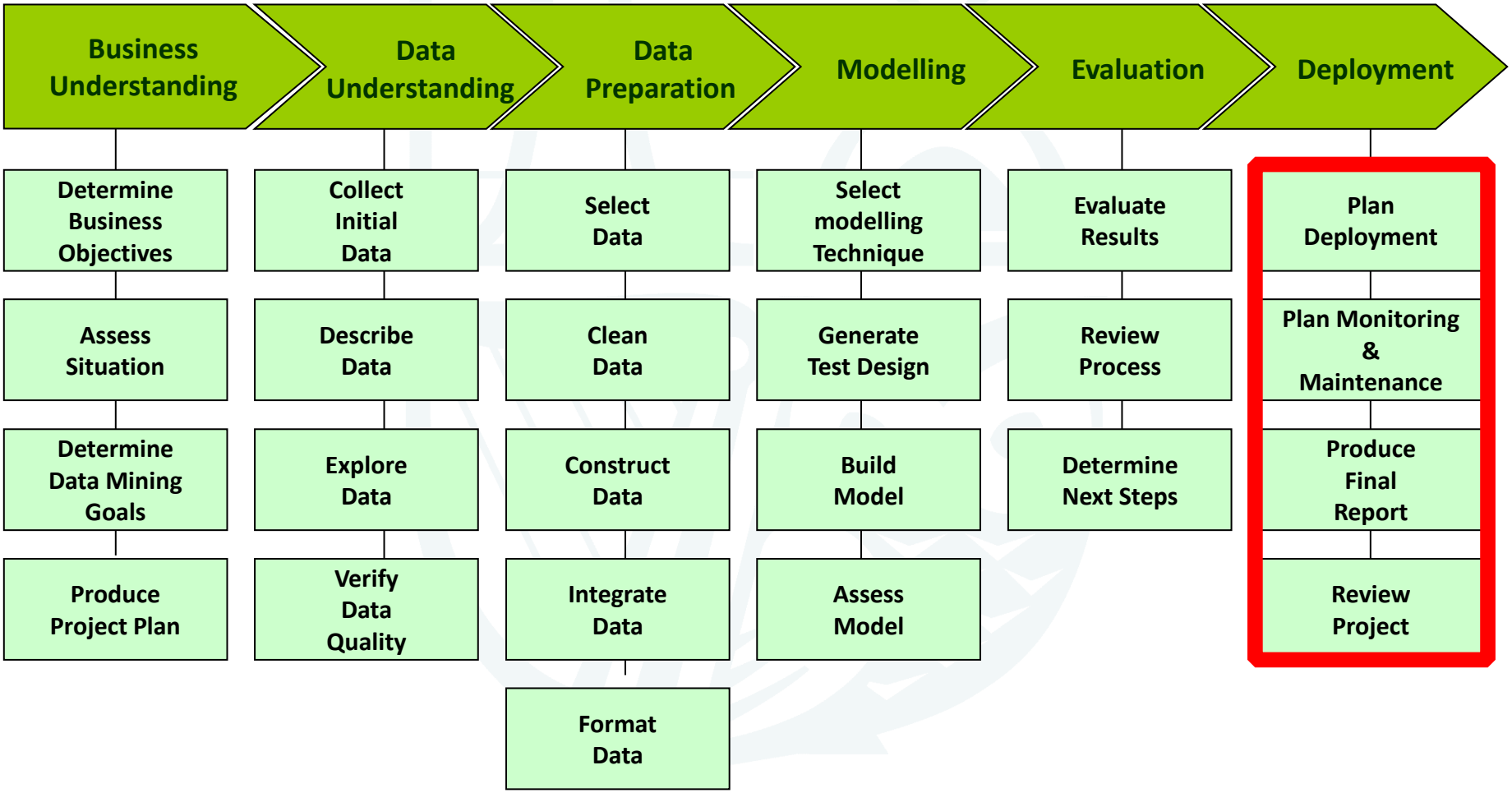
6. Deployment

- Determine how the results need to be utilised
- Determine who needs to use the results
- Determine how often do the results need to be used
- Deploy data mining results by
  - Reports to decision makers, using results as business rules, interactive information feeds etc.



The knowledge gained will need to be organised and presented in a way that the consumer can effectively use. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

CRISP-DM: Phase 6 Model Deployment



DM Process - Phase 6 MD

- **1. Plan deployment**
  - To deploy the data mining result(s) into the business, use the evaluation results and conclude a strategy for deployment
    How does this happen in your organisation?
  - Document the procedure for later deployment

- **2. Plan monitoring and maintenance**
  - Important if the data mining results are to become integral to the day-to-day business and environment
  - Avoid long periods of incorrect usage of data mining results
  - A detailed monitoring process is required
  - Cognisant of the specific type of deployment

- **Discussion** - Does this happen in your organisation?

DM Process - Phase 6 MD

- **3. Produce final report**
  - Project leader and team create a final project report
  - A summary of the project and its experiences or a final comprehensive presentation of the data mining result(s)
- **4. Review project**
  - Assess what went right, what went wrong, what worked well and what needs to be improved
- **Discussion** - Does this happen in your organisation?

CRISP-DM: Summary

1. Business Understanding

1. Understanding project objectives and requirements
2. Data mining problem definition

2. Data Understanding

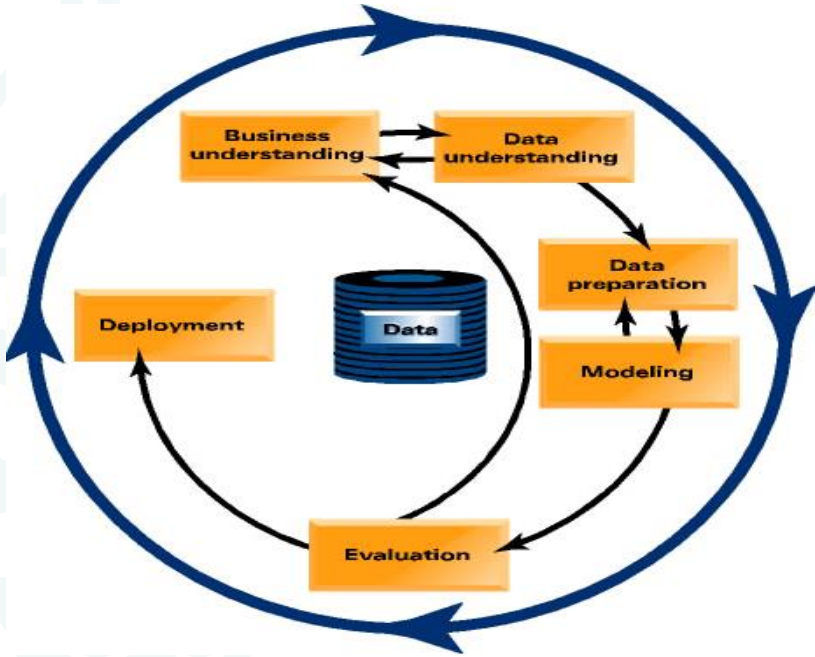
1. Initial data collection and familiarisation
2. Identify data quality issues
3. Initial, obvious results

3. Data Preparation

1. Record and attribute selection
2. Data cleansing

4. Modelling

1. Run the data mining tools



5. Evaluation

1. Determine if results meet business objectives
2. Identify business issues that should have been addressed earlier

6. Deployment

1. Put the resulting models into practice
2. Set up for repeated/continuous mining of the data

CRISP-DM Strengths

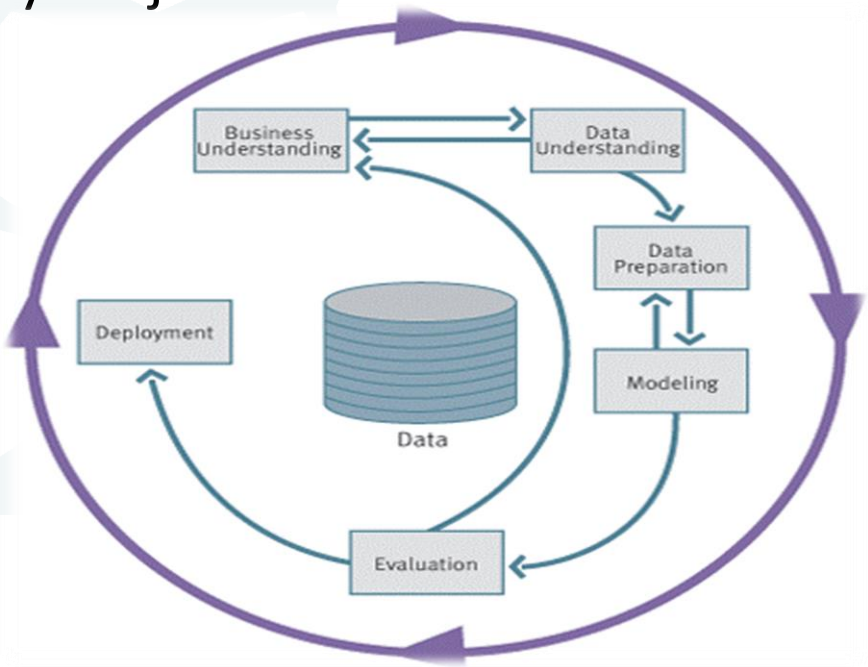
- The data mining process must be reliable and repeatable by people with little data mining skills
- CRISP-DM provides a uniform framework for
  - Data mining guidelines
  - Documentation of data mining experiences
- CRISP-DM is flexible to account for differences
  - Different business problems
  - Different goals
  - Different data

CRISP-DM Weaknesses

- CRISP-DM provides a waterfall framework unless vertical slicing is used.
- CRISP-DM does not cover deployment in newer environments
- CRISP-DM does not cover the application scenario where an ML model is maintained as an application
- CRISP-DM lacks guidance on quality assurance methodology.
- CRISP-ML and newer versions are trying to address the weakness around deployment

CRISP-DM - On our projects

- Enterprise Ireland Innovation Vouchers (EI IV) & privately funded projects
- Enterprise Ireland Innovation Partnership Project (EI IPP)
- M.Sc. In Data Science(DS) Industry Projects
- Research projects



EI Innovation Vouchers (6-8 weeks)

- BU – easy to get from domain experts as projects smaller in scale
- DU – subject to readily available data (often MS Excel, csv), hard to assess due to lack of company knowledge
- DP – we prepare what is provided, we often need/ask for more
- Modeling – straightforward/often more descriptive or visual in nature
- Evaluation – dependent on models required, descriptive statistics, regression
- Deployment – variable, assessment of where company currently is/recommendations

M.Sc. In DS industry projects (6-8 months)

- BU –from domain experts, projects medium size, longer term
- DU – much more time for EDA, additional data gathering
- DP – much fuller undertaking based on company requirements
- Modeling – more advanced predictive modelling, usually ML, nice to solve business problem
- Evaluation – more time to assess the models, iteration, testing, validation stronger
- Deployment – variable, but may serve as a basis for further work/POC

EI IPP (18 months)

- BU – from domain experts, projects much large in scale
- DU – subject to & limited by the equipment/need
- DP – we gather what we need and iterate early & often
- Modeling – much more advanced and exploratory, fail fast and iterate
- Evaluation – dependent on client needs and accuracy and precision required
- Deployment – variable but may serve as a basis for further work/POC

Technologies on our/other projects

EI IV, EI IPP, M.Sc. In DS

- Infrastructure – Microsoft Azure, Amazon Web Services AWS, Hadoop cluster, local machines, GPUs
- Data manipulation – SQL, NoSQL, NewSQL
- Databases/datasets – MySQL, MS Excel, SQL Server, flat files, csv
- Programming – **Python**, **R**, Julia

Technologies on our/other projects

EI IV, EI IPP, M.Sc. In DS

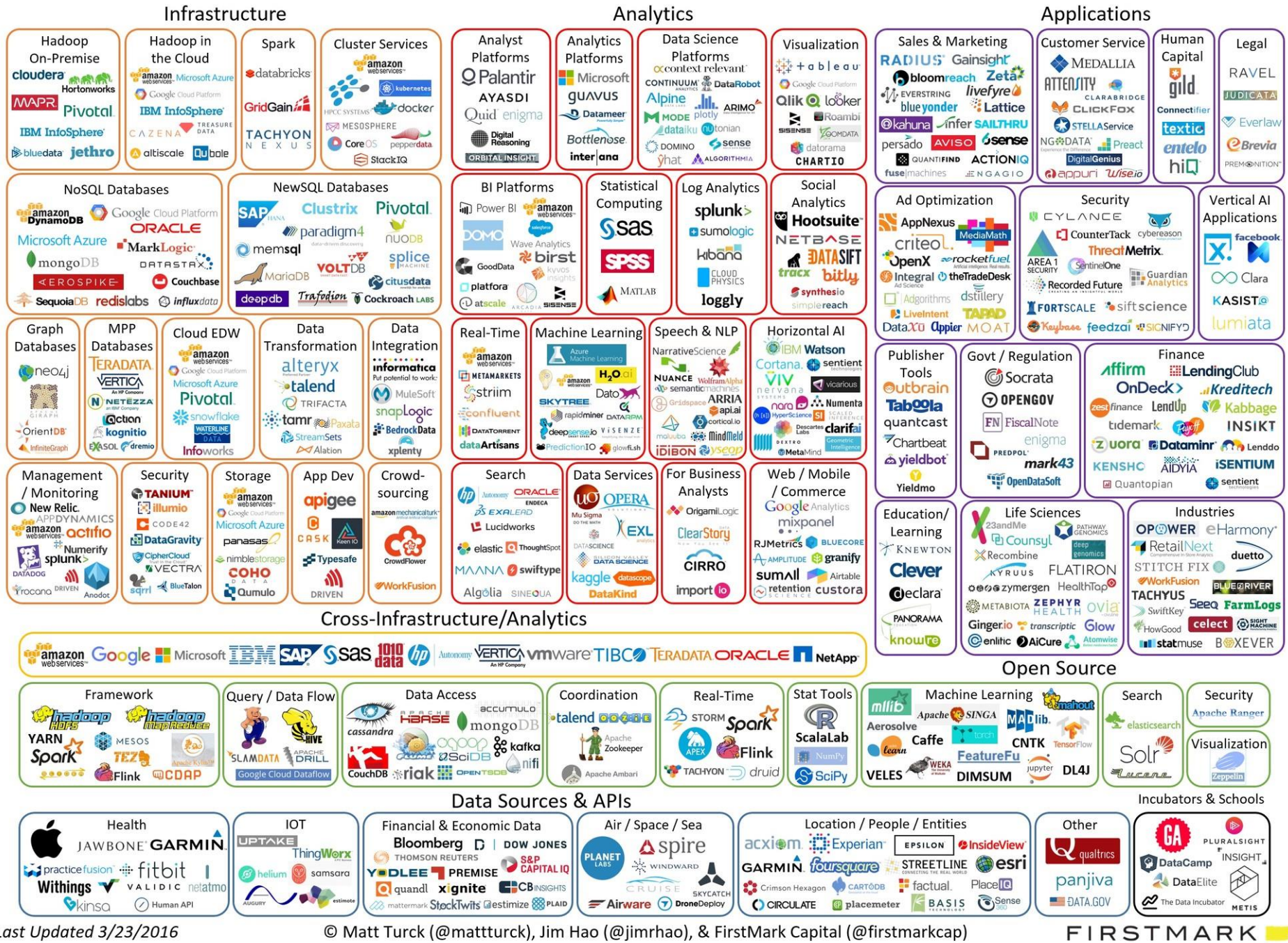
- Visualisation –Matplotlib, Dash & Plotly, ggplot, R Shiny apps
- Data science/machine learning platforms– RapidMiner & Weka, KNIME, Azure ML, MATLAB, SPSS
  https://www.gartner.com/reviews/home
- Other – Jupyter notebooks https://jupyter.org/, GitHub

Analytics and ML software tools/platforms

- Table 1: Top Analytics/Data Science/ML Software in 2019 KDnuggets Poll

Software	2019 % share	2018 % share	2017 % share
Python	65.8%	65.6%	59.0%
RapidMiner	51.2%	52.7%	31.9%
R	46.6%	48.5%	56.6%
Excel	34.8%	39.1%	31.5%
Anaconda	33.9%	33.4%	24.3%
SQL	32.8%	39.6%	39.2%
Tensorflow	31.7%	29.9%	22.7%
Keras	26.6%	22.2%	10.7%
scikit-learn	25.5%	24.4%	21.9%
Tableau	22.1%	26.4%	21.8%
Apache Spark	21.0%	21.5%	25.5%

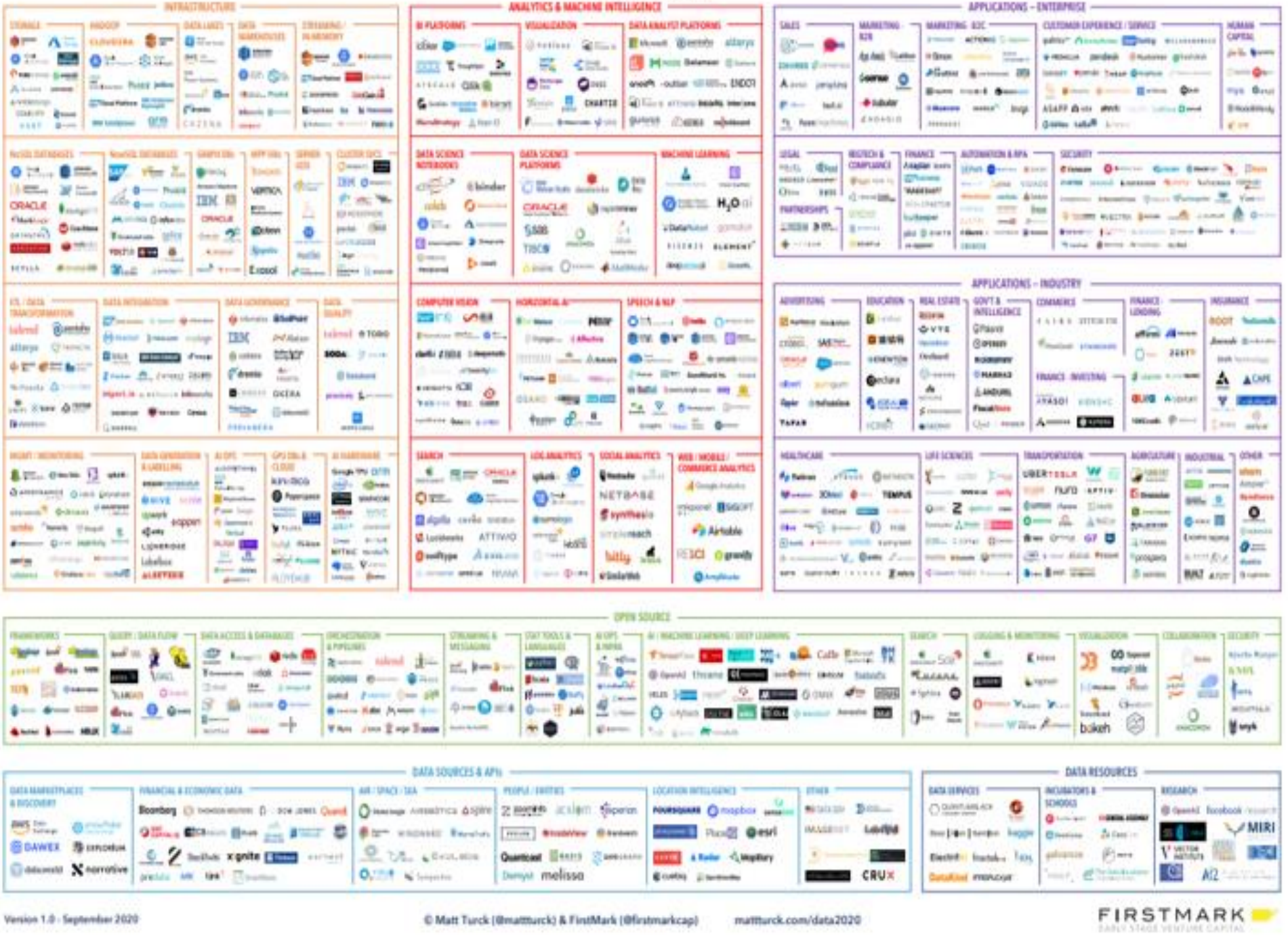
Big Data Landscape 2016 (Version 3.0)



Last Updated 3/23/2016

FIRSTMARK

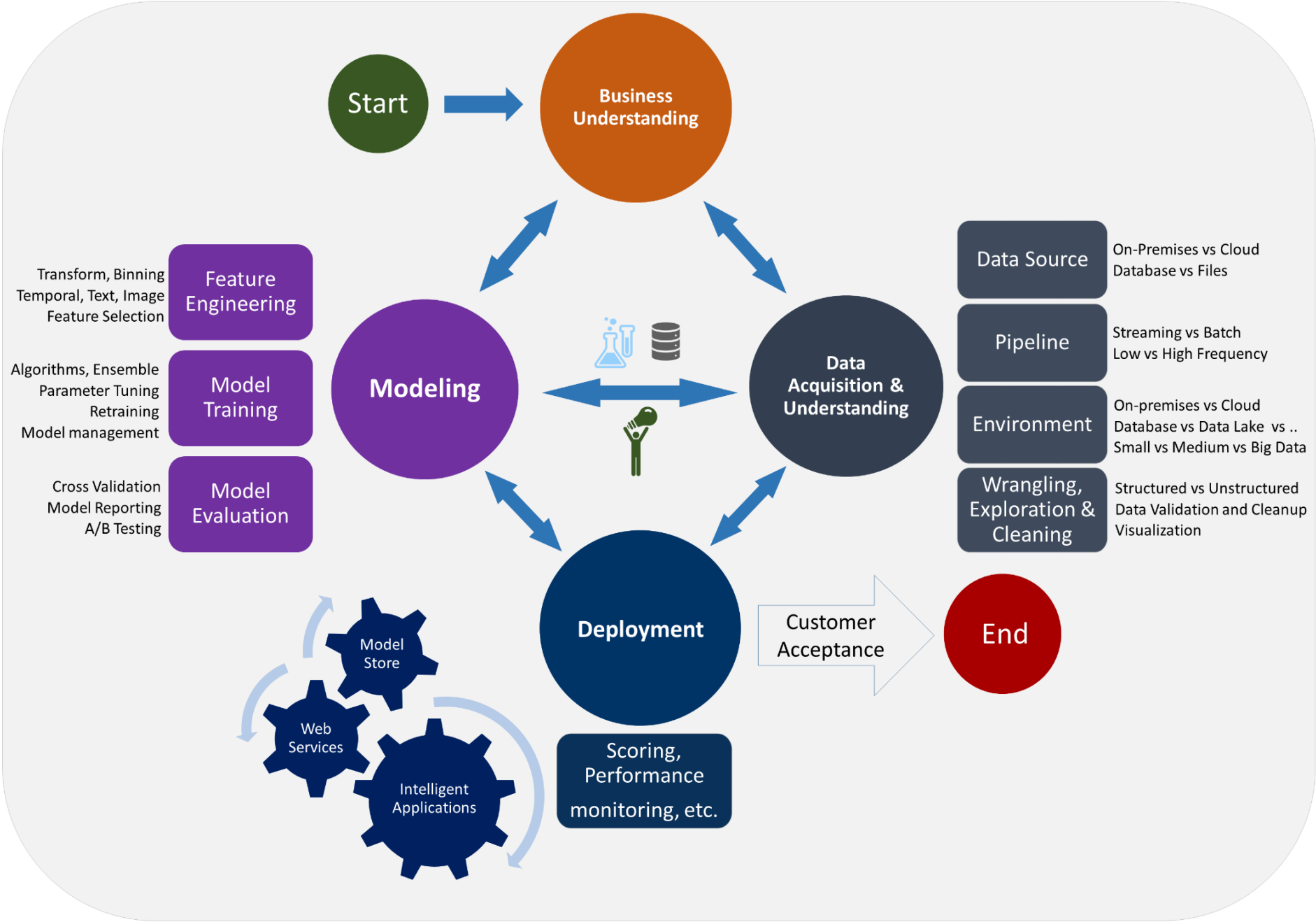
DATA & AI LANDSCAPE 2020



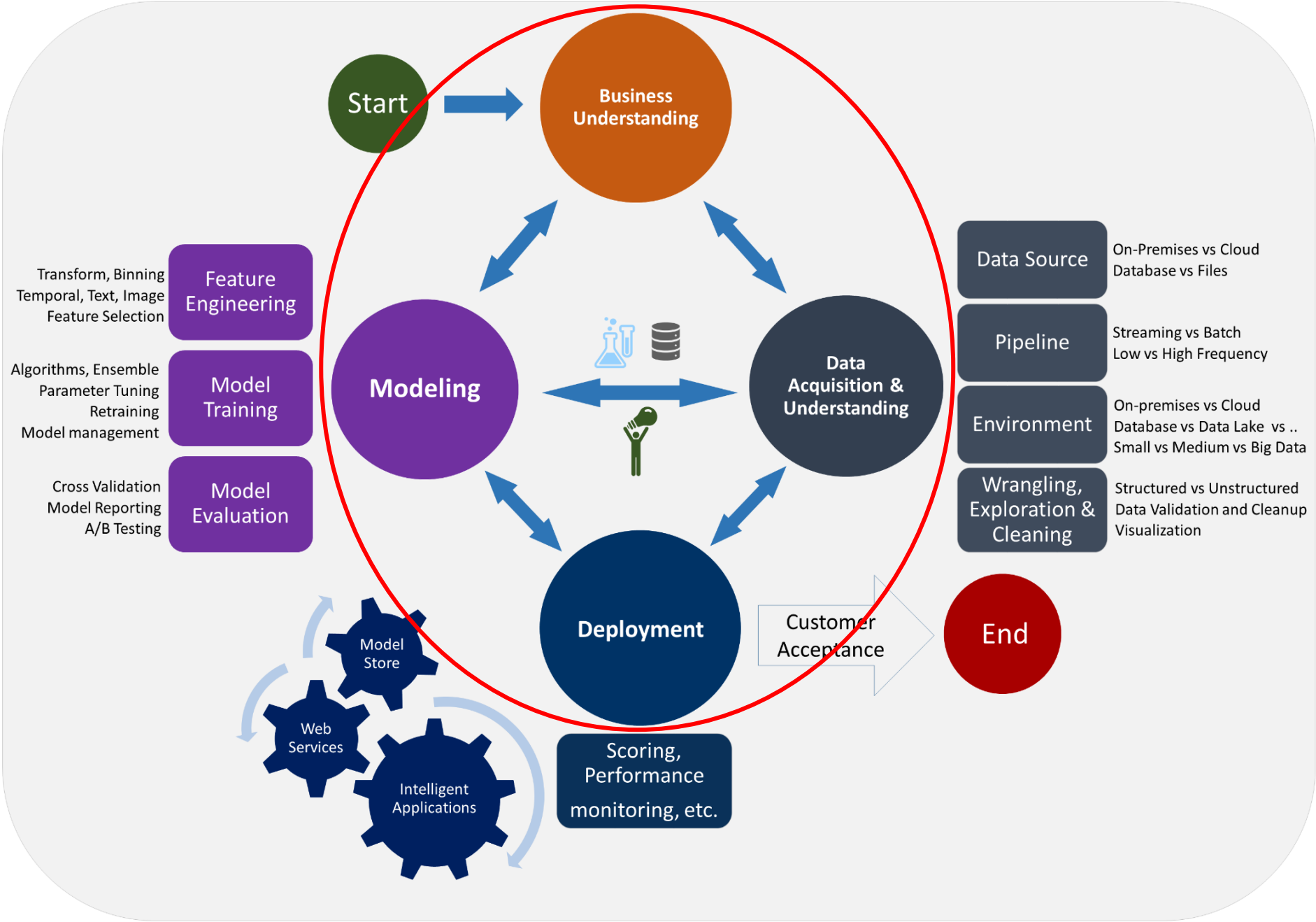
mattturck.com/data2020

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

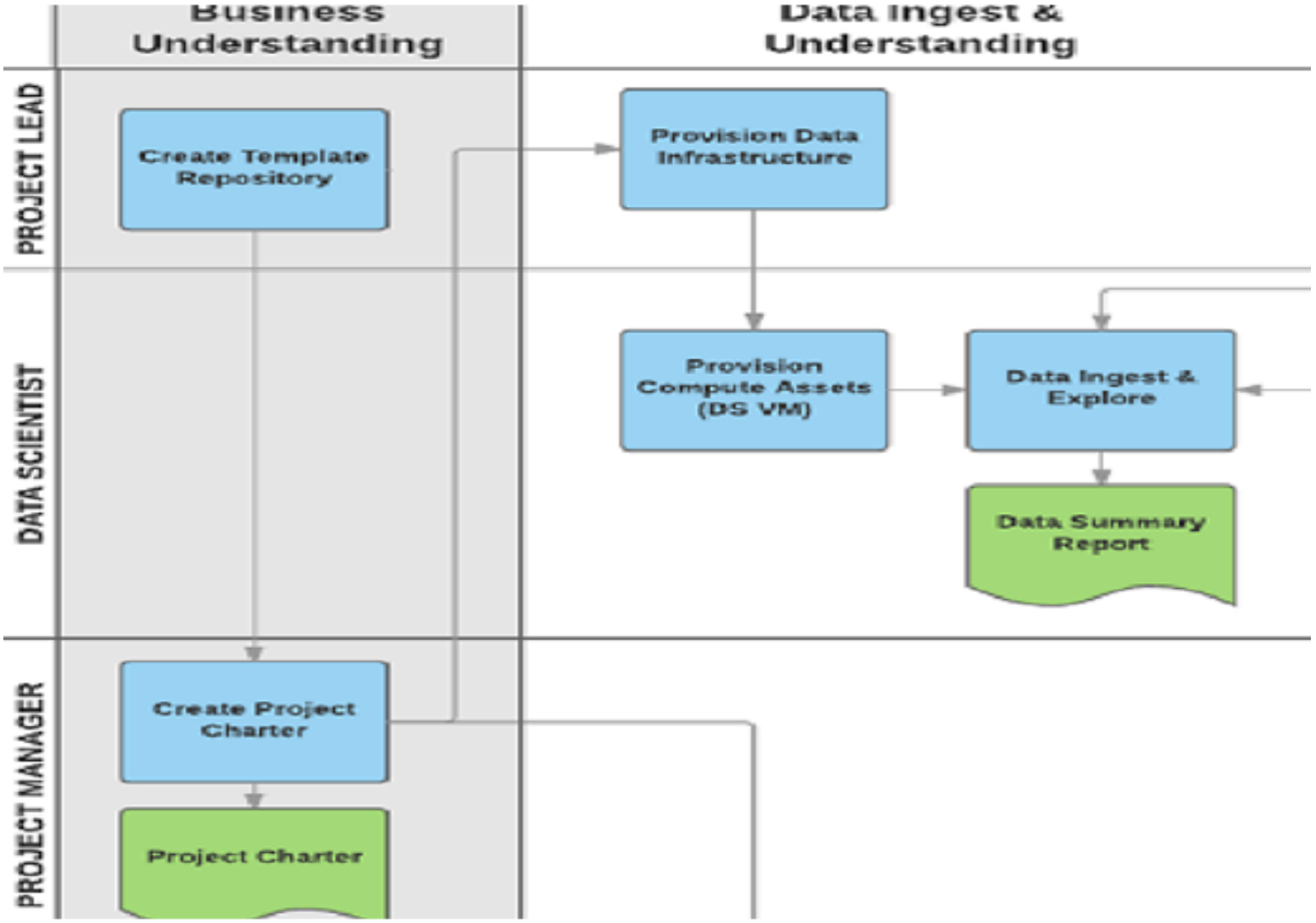
Data Science Lifecycle



Data Science Lifecycle



Microsoft TDSP - Tasks & artefacts



Source: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

Microsoft TDSP - Key components

- A **data science agile, iterative lifecycle** definition
- A **standardized collaborative (team) project structure**
- **Infrastructure and resources** for data projects – on-site/cloud datasets/DB, big data (SQL or spark) clusters, ML services (Azure Machine Learning)
- **Tools and utilities** recommended for project execution

- Source: https://docs.Microsoft.Com/en-us/azure/machine-learning/team-data-science-process/overview

Conclusion & recommendations

- CRISP-DM - the de facto industry leader, traditional or agile
- Makes data analytics process more reliable & repeatable
- Learn more about CRISP-DM (TDSP or another process model/framework)
- Learn Python and/or R
- Learn how to use Jupyter notebooks
- Undertake/review some of Andrew Ng ML courses

Questions

Please click on the ‘Raise Hand’ icon

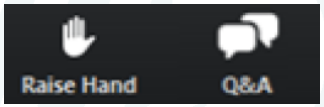
to ask a question

and

wait to be unmuted

or

Use the Q&A function



References & resources

- CRISP-DM model documentation is available here: https://www.the-modeling-agency.com/crisp-dm.pdf

- ASUM DM available here:

http://gforge.icesi.edu.co/ASUM-DM_External/index.htm#cognos.external.asum-DM_Teaser/deliveryprocesses/ASUM-DM_8A5C87D5.html and here: https://www.Researchgate.Net/publication/321944704_combining_process_guidance_and_industrial_feedback_for_successfully_deploying_big_data_projects

- Team Data Science Process https://docs.microsoft.com/en-Us/azure/architecture/data-science-process/overview and https://github.com/Azure/Azure-TDSP-ProjectTemplate

- Angée S., Lozano-Argel S.I., Montoya-Munera E.N., Ospina-Arango JD., Tabares-Betancur M.S. (2018) Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects. In: Uden L., Hadzima B., Ting IH. (eds) Knowledge Management in Organizations. KMO 2018. Communications in Computer and Information Science, vol 877. Springer, Cham. https://doi.org/10.1007/978-3-319-95204-8_51

References & resources

- Studer, S.; Bui, T.B.; Drescher, C.; Hanuschkin, A.; Winkler, L.; Peters, S.; Müller, K.-R. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. Mach. Learn. Knowl. Extr. 2021, 3, 392–413. https://doi.org/10.3390/make3020020

- Schröerab, C; Kruseb, F; GómezbA, JM; Systematic Literature Review on Applying CRISP-DM Process 2020 Model DOI: 10.1016/j.procs.2021.01.199

- J. S. Saltz and N. Hotz, "Identifying the most Common Frameworks Data Science Teams Use to Structure and Coordinate their Projects," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 2038-2042, doi: 10.1109/BigData50022.2020.9377813.

- Grady, Nancy W.. “KDD meets Big Data.” 2016 IEEE International Conference on Big Data (Big Data) (2016): 1603-1608.

- Volk, Matthias et al. “Approaching the (Big) Data Science Engineering Process.” IoTBDS (2020).

Thank you for listening!