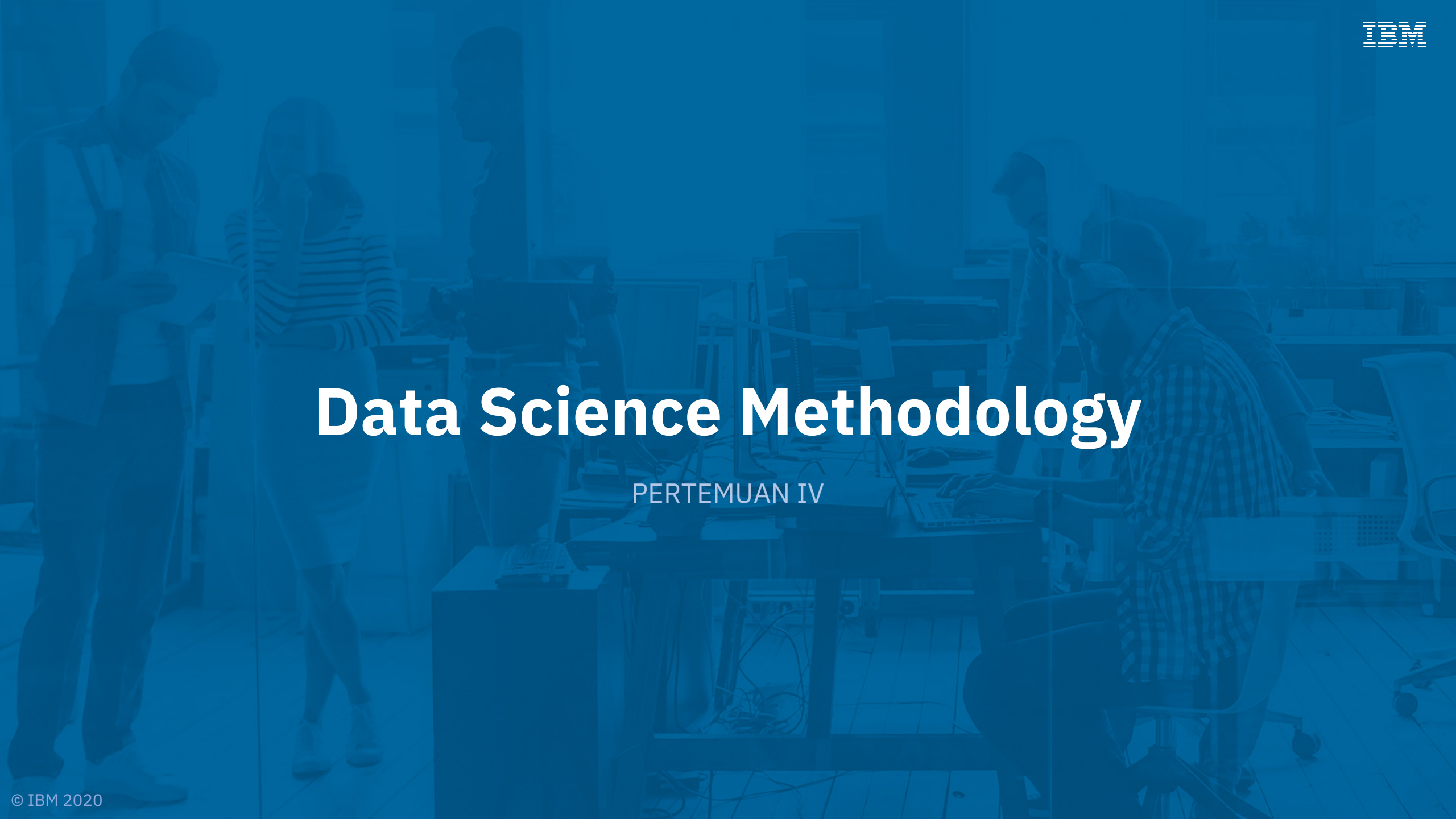


Data Science Methodology

PERTEMUAN IV

Learning Objectives

In this course you will learn about:

- The major steps involved in tackling a data science problem.
- The major steps involved in practicing data science, from forming a concrete business or research problem, to collecting and analyzing data, to building a model, and understanding the feedback after model deployment.
- How data scientists think through tackling interesting real-world examples.

Data Science Methodology

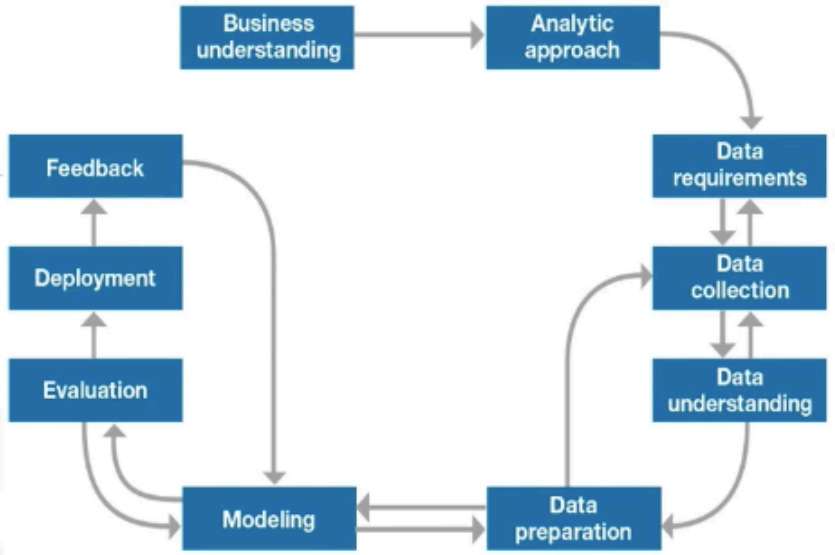
Data Science Methodology's purpose is **to share a methodology that can be used within data science, to ensure that the data used in problem solving is relevant and properly manipulated** to address the question at hand.

The data science methodology discussed in this course has been outlined by John Rollins, a seasoned and senior data scientist currently practicing at IBM.

This course is built on his experience and expresses his position on the importance of following a methodology to be successful.

Data Science Methodology Stages

In a nutshell…



The **Data Science Methodology** aims to answer the following 10 questions in this prescribed sequence:

From problem to approach:

1. *What is the problem that you are trying to solve?*
2. *How can you use data to answer the question?*

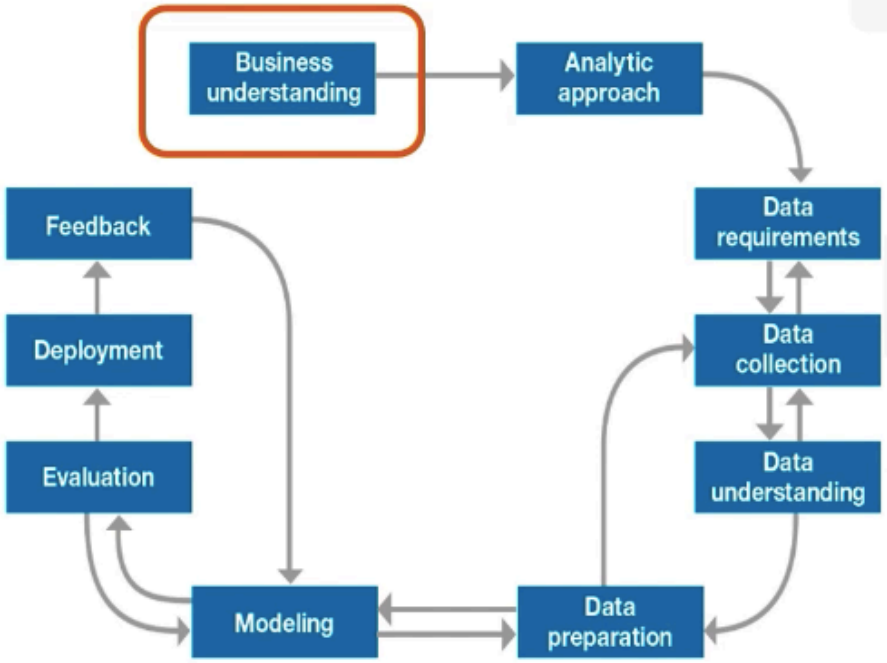
Working with the data:

3. *What data do you need to answer the question?*
4. *Where is the data coming from (identify all sources) and how will you get it?*
5. *Is the data that you collected representative of the problem to be solved?*
6. *What additional work is required to manipulate and work with the data?*

Deriving the answer:

7. *In what way can the data be visualized to get to the answer that is required?*
8. *Does the model used really answer the initial question or does it need to be adjusted?*
9. *Can you put the model into practice?*
10. *Can you get constructive feedback into answering the question?*

From Understanding to Approach



Business understanding

- *What is the problem that you are trying to solve?*

Business Understanding (Example)

You've been called into a meeting by your boss, who gave an important task one with a very tight deadline that absolutely has to be met.

You ensure that all aspects of the task have been considered and the meeting ends with both of you confident that things are on track.

Later after you've spent some time examining the various issues at play, you realize that you need to ask several additional questions in order to truly accomplish the task.

Unfortunately, the boss won't be available again until tomorrow.

Now, with the tight deadline still ringing in your ears, you start feeling a sense of uneasiness. So, what do you do? Do you risk moving forward or do you stop and seek clarification.

Business Understanding

Data science methodology **begins with spending the time to seek clarification, to attain what can be referred to as a business understanding.**

Having this understanding is placed at the beginning of the methodology because getting clarity around the problem to be solved, allows you to determine which data will be used to answer the core question.

Business Understanding

Rollins suggests that **having a clearly defined question is vital because it ultimately directs** the analytic approach that will be needed to address the question.

All too often, much effort is p**ut into answering what people THINK is the question, and while the methods used to address that question might be sound**, they don't help to solve the actual problem.

Establishing a clearly defined question starts with understanding the GOAL of the person who is asking the question.

Business Understanding

For example, if a business owner asks: "**How can we reduce the costs of performing an activity?**"

We need to understand, **is the goal to improve the efficiency of the activity? Or is it to increase the businesses profitability?**

Once the goal is clarified, the next piece of the puzzle is to figure out **the objectives that are in support of the goal.**

Business Understanding

By breaking down the objectives, structured discussions can take place where priorities can be identified in a way that can lead to organizing and planning on how to tackle the problem.

Depending on the problem, different stakeholders will need to be engaged in the discussion to help determine requirements and clarify questions.

Business Understanding

The key business sponsors involvement throughout the project was critical, in that the sponsor:

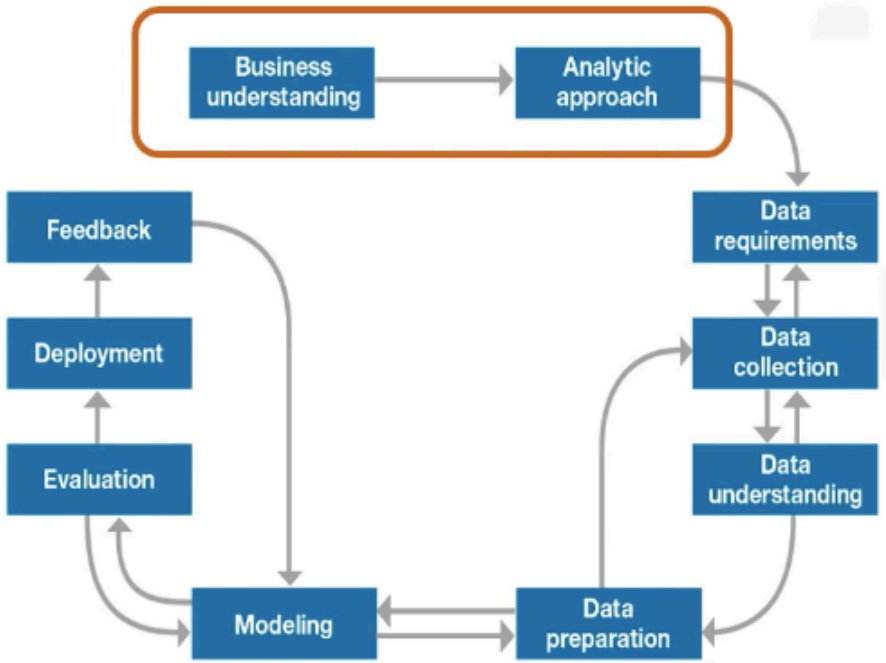
Set overall direction

Remained engaged and provided guidance.

Ensured necessary support, where needed.

Getting stakeholder “buy-in” and support

From Understanding to Approach



Business understanding

- *What is the problem that you are trying to solve?*

Analytic approach

- *How can you use data to answer the question?*

Analytic Approach

Selecting the right analytic approach depends on the question being asked.

The approach involves seeking clarification from the person who is asking the question, so as to be able to pick the most appropriate path or approach.

Once the problem to be addressed is defined, the appropriate analytic approach for the problem is selected in the context of the business requirements.

This is the second stage of the data science methodology.

Analytic Approach

Once a strong understanding of the question is established, the analytic approach can be selected.

This means identifying what type of patterns will be needed to address the question most effectively.

If the question is to determine probabilities of an action, then a predictive model might be used.

If the question is to show relationships, a descriptive approach maybe be required.

Pick analytic approach based on type of question

Descriptive
- Current status

Diagnostic (Statistical Analysis)
- What happened?
- Why is this happening?

Predictive (Forecasting)
- What if these trends continue?
- What will happen next?

Prescriptive
- How do we solve it?

What are the types of questions?

If the question is to determine probabilities of an action

- Use a Predictive model

If the question is to show relationships

- Use a descriptive model

If the question requires a yes/no answer

- Use a classification model

Analytic approach

- *How can you use data to answer the question?*

- The correct approach depends on business requirements for the model

Analytic Approach

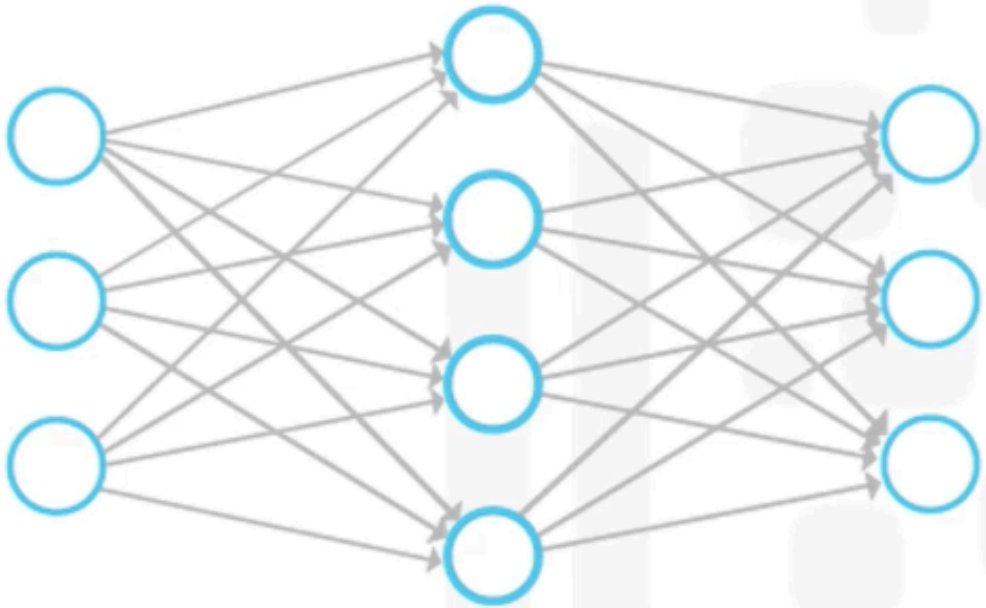
Statistical analysis applies to problems that require counts.

For example if the question requires a yes/ no answer, then a classification approach to predicting a response would be suitable.

Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed.

Machine Learning can be used to identify relationships and trends in data that might otherwise not be accessible or identified.

Will machine learning be utilized?



Machine Learning

- Learning without being explicitly programmed
- Identifies relationships and trends in data that might otherwise not be accessible or identified
- Uses clustering association approaches

Analytic Approach

In the case where the question is to learn about human behavior, then an appropriate response would be to use Clustering Association approaches.

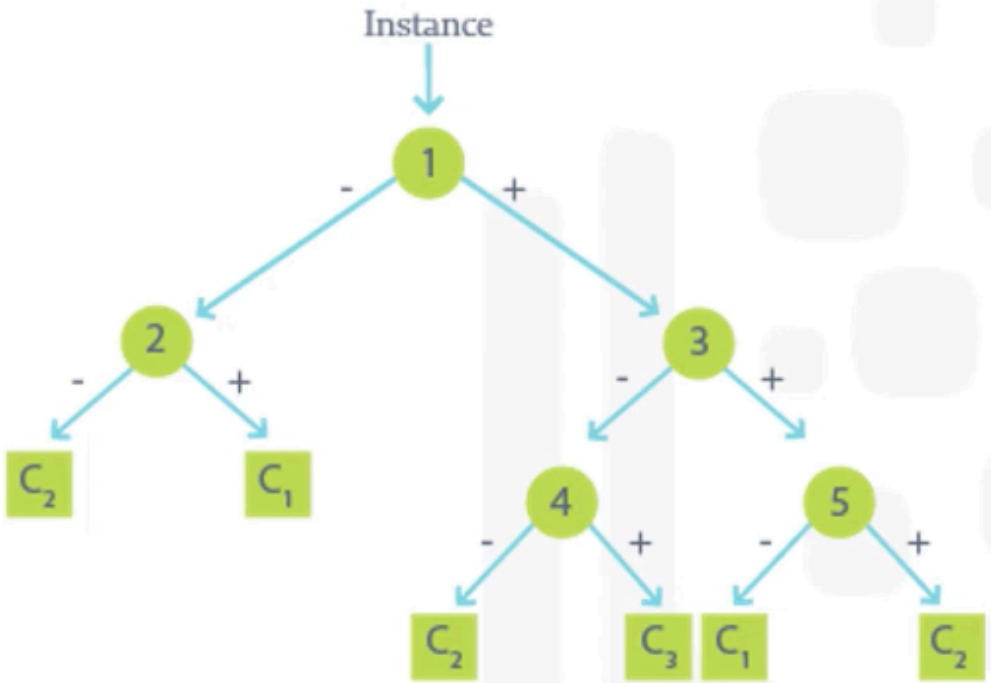
So now, let's look at the case study related to applying Analytic Approach.

For the case study, a decision tree classification model was used to identify the combination of conditions leading to each patient's outcome.

In this approach, examining the variables in each of the nodes along each path to a leaf, led to a respective threshold value.

This means the decision tree classifier provides both the predicted outcome, as well as the likelihood of that outcome, based on the proportion at the dominant outcome, yes or no, in each group.

Case Study – Decision tree classification selected!



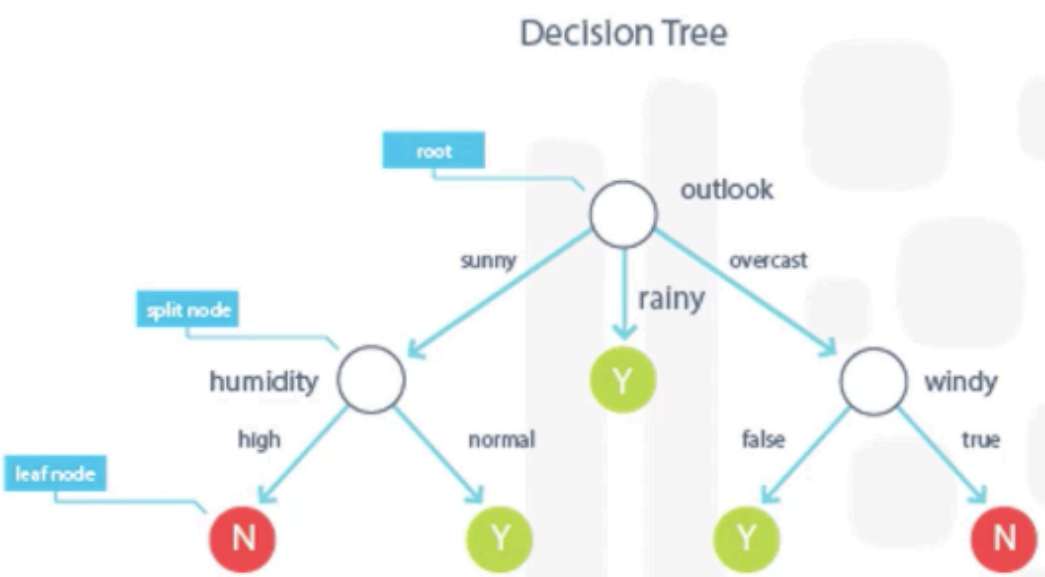
Predictive model

- To predict an outcome

Decision tree classification

- Categorical outcome
- Explicit “decision path” showing conditions leading to high risk
- Likelihood of classified outcome
- Easy to understand and apply

Case Study – Example of decision tree classification



Predictive model

- To predict an outcome

Decision tree classification

- Categorical outcome
- Explicit “decision path” showing conditions leading to high risk
- Likelihood of classified outcome
- Easy to understand and apply

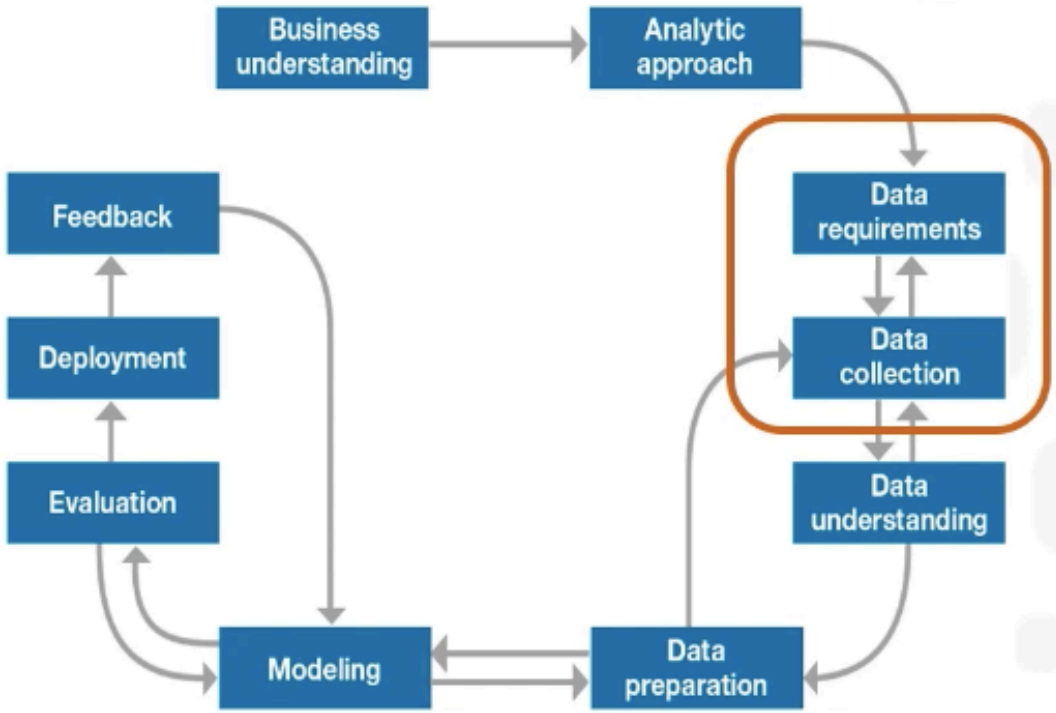
Analytic Approach

From this information, the analysts can obtain the readmission risk, or the likelihood of a yes for each patient. If the dominant outcome is yes, then the risk is simply the proportion of yes patients in the leaf.

If it is no, then the risk is 1 minus the proportion of no patients in the leaf.

A decision tree classification model is easy for non-data scientists to understand and apply, to score new patients for their risk of readmission.

From Requirements to Collection



Data Requirements

- *What are data requirements?*

Data Collection

- *What occurs during data collection?*

Data Requirement

If your goal is to make a spaghetti dinner but you don't have the right ingredients to make this dish, then your success will be compromised.

Think of this section of the data science methodology as cooking with data. Each step is critical in making the meal.

So, if the problem that needs to be resolved is the recipe, so to speak, and data is an ingredient, then the data scientist needs to identify:

- **which ingredients are required,**
- **how to source or the collect them,**
- **how to understand or work with them, and**
- **how to prepare the data to meet the desired outcome.**

Data Requirement

Building on the understanding of the problem at hand, and then using the analytical approach selected, **the Data Scientist is ready to get started.**

Now let's look at some examples of the data requirements within the data science methodology.

This includes identifying the necessary data content, formats and sources for initial data collection.

Data Requirement

So now, let's look at the case study related to applying "Data Requirements".

In the case study, the first task was to define the data requirements for the decision tree classification approach that was selected.

This included selecting a suitable patient cohort from the health insurance providers member base.

Data Requirement (Case)

In order to compile the complete clinical histories, three criteria were identified for inclusion in the cohort.

First, a patient needed to be admitted as in-patient within the provider service area, so they'd have access to the necessary information.

Second, they focused on patients with a primary diagnosis of congestive heart failure during one full year.

Third, a patient must have had continuous enrollment for at least six months, prior to the primary admission for congestive heart failure, so that complete medical history could be compiled.

Data Requirement (Case)

Congestive heart failure patients who also had been diagnosed as having other significant medical conditions, were excluded from the cohort because those conditions would cause higher-than-average re-admission rates and, thus, could skew the results.

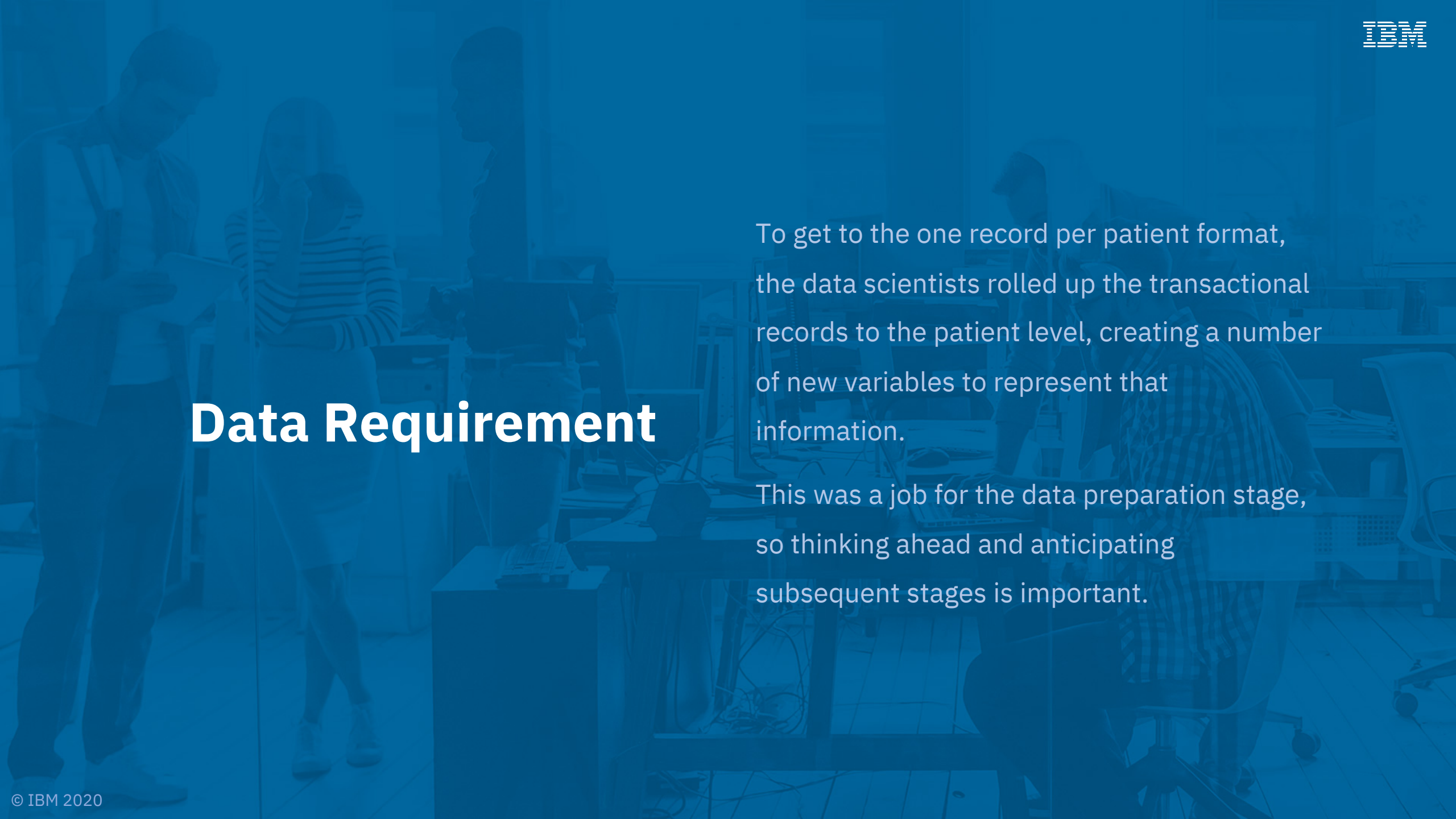
Then the content, format, and representations of the data needed for decision tree classification were defined.

Data Requirement (Case)

This modeling technique requires one record per patient, with columns representing the variables in the model.

To model the readmission outcome, there needed to be data covering all aspects of the patient's clinical history.

This content would include admissions, primary, secondary, and tertiary diagnoses, procedures, prescriptions, and other services provided either during hospitalization or throughout patient/doctor visits.



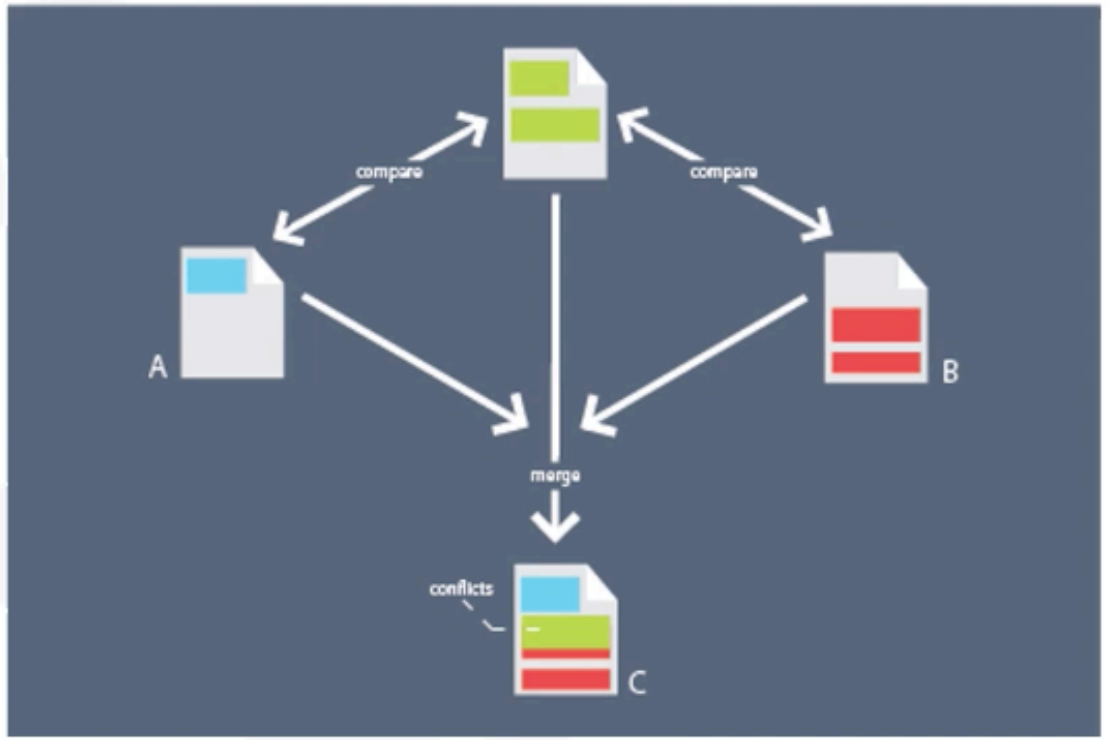
Data Requirement

To get to the one record per patient format, the data scientists rolled up the transactional records to the patient level, creating a number of new variables to represent that information.

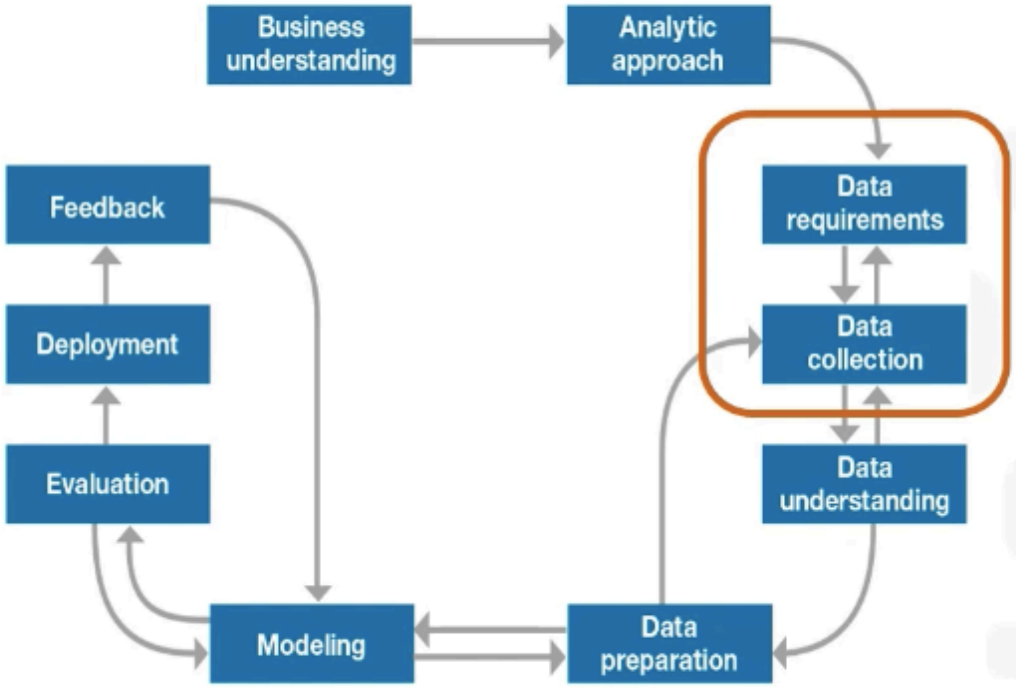
This was a job for the data preparation stage, so thinking ahead and anticipating subsequent stages is important.

Case Study – Defining the data

- Content, formats, representations suitable for decision tree classifier
  - One record per patient with columns representing variables (dependent variable and predictors)
  - Content covering all aspects of each patient's clinical history
    - Transactional format
    - Transformations required



From Requirements to Collection



Data Requirements

- *What are data requirements?*

Data Collection

- *What occurs during data collection?*

Data Collection

After the initial data collection is performed, **an assessment by the data scientist takes place** to determine whether or not they have what they need.

As is the case when shopping for ingredients to make a meal, some ingredients might be out of season and more difficult to obtain or cost more than initially thought.

In this phase the data requirements are revised and decisions are made as to whether or not the collection requires more or less data.

Case Study – Gathering available data

- Available data sources
  - Corporate data warehouse (single source of medical & claims, eligibility, provider and member information)
  - In-patient record system
  - Claim payment system
  - Disease management program information

Data Collection

Once the data ingredients are collected, then in the data collection stage, the data scientist will have a good understanding of what they will be working with.

Techniques such as descriptive statistics and visualization can be applied to the data set, to assess the content, quality, and initial insights about the data.

Gaps in data will be identified and plans to either fill or make substitutions will have to be made.

Data Collection

In essence, the ingredients are now sitting on the cutting board.

Collecting data requires that you know the source or, know where to find the data elements that are needed.

DBAs and programmers often work together to extract data from various sources, and then merge it.

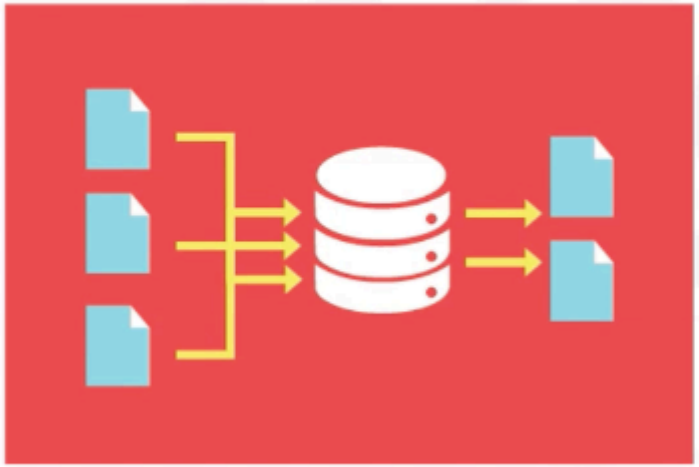
Data Collection

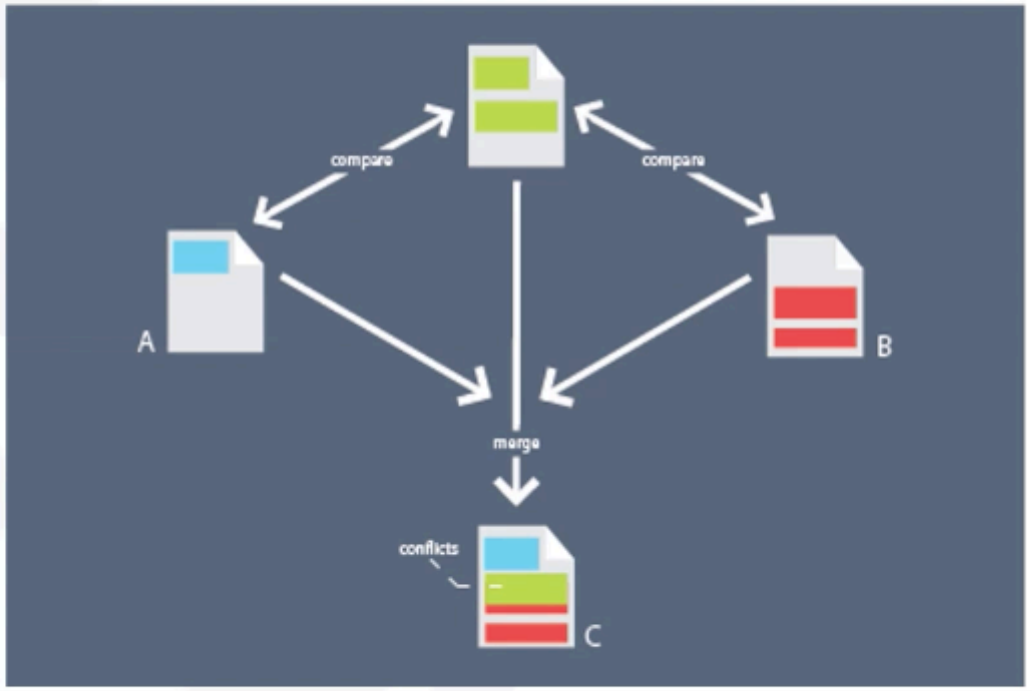
This allows for removing redundant data, making it available for the next stage of the methodology, which is data understanding.

At this stage, if necessary, data scientists and analytics team members can discuss various ways to better manage their data, including automating certain processes in the database, so that data collection is easier and faster.

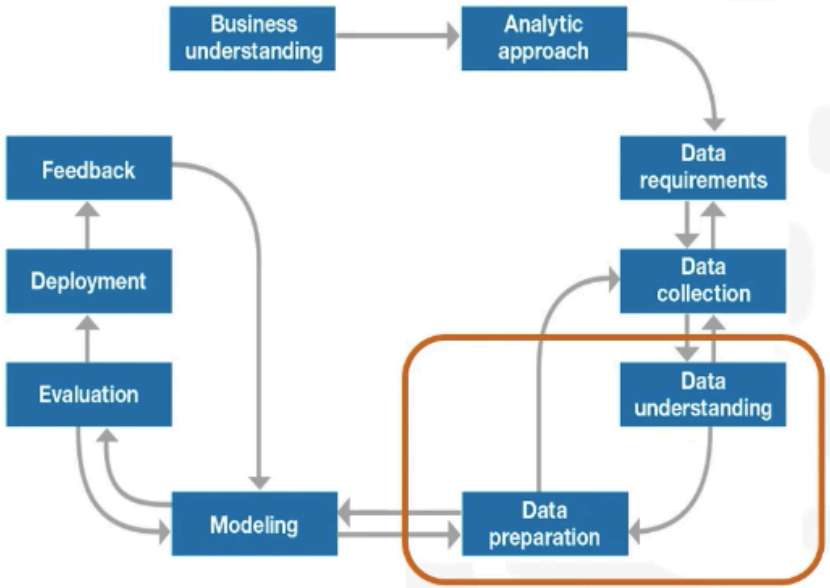
Case Study – Merging data

- Eliminate redundant data





From Understanding to Preparation



Data understanding

- *What does it mean to "prepare" or "clean" data?*

Data preparation

- *What are ways in which data is prepared?*

BIG DATA UNIVERSITY

Data Understanding

Data understanding encompasses **all activities related to constructing the data set**.

Essentially, the data understanding section of the data science methodology answers the question: **Is the data that you collected representative of the problem to be solved?**

Statistics needed to be run against the data columns that would become variables in the model.

The more one works with the problem and the data, the more one learns and therefore the more refinement that can be done within the model, ultimately leading to a better solution to the problem.

Data Understanding

From the information provided, certain values can be re-coded or perhaps even dropped if necessary, such as when a certain variable has many missing values.

The question then becomes, does "missing" mean anything?

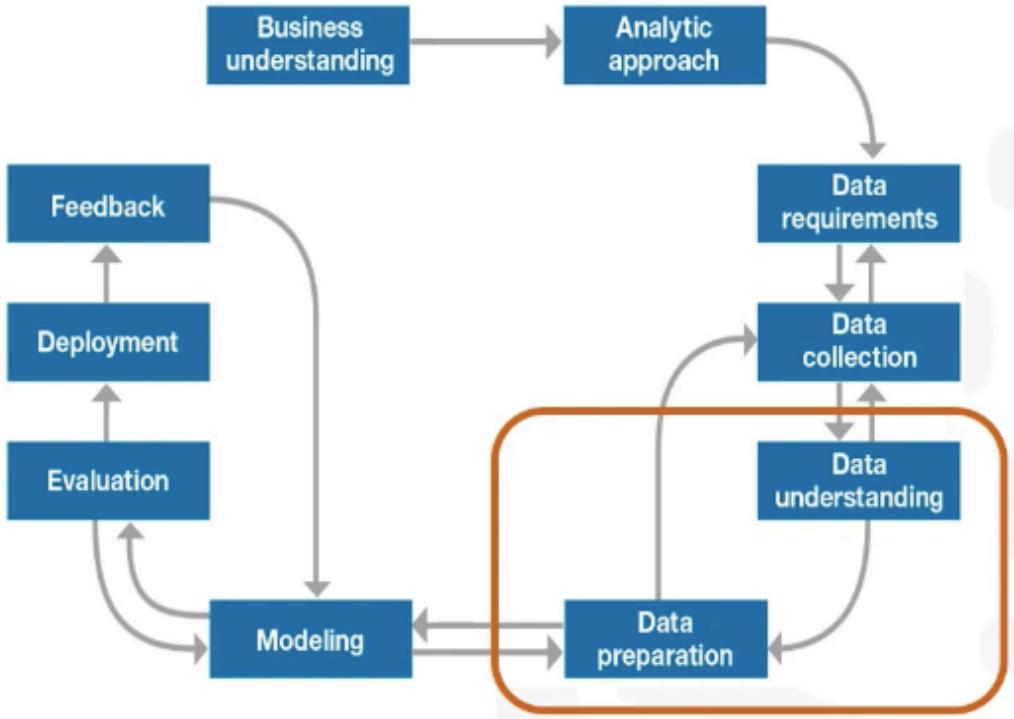
Sometimes a missing value might mean "no", or "0" (zero), or at other times it simply means "we don't know".

Or, if a variable contains invalid or misleading values, such as a numeric variable called "age" that contains 0 to 100 and also 999, where that "triple-9" actually means "missing", but would be treated as a valid value unless we corrected it.

Case study – Looking at data quality

- Data quality
  - Missing values
  - Invalid or misleading values

From Understanding to Preparation



Data understanding

- *What does it mean to "prepare" or "clean" data?*

Data preparation

- *What are ways in which data is prepared?*

Data Preparation

In a sense, data preparation is similar **to washing freshly picked vegetables in so far as unwanted elements,** such as dirt or imperfections, are removed.

Together with **data collection and data understanding, data preparation is the most time-consuming phase** of a data science project, typically taking seventy percent and even up to even ninety percent of the overall project time.

Automating some of the data collection and preparation processes in the database, can reduce this time to as little as 50 percent.

Examples of data cleansing

Name	Date	Age	Location	Country
John Doe	2012 02 20	32	ON	CAN
May Lag	2013 02 33	2	ON	CA
Henry Oon	30-Sep-12	35	Ontario	CANADA
Kelly, Tom	2015 02 20	65	ON	CA
John Kell	2016 02 20		AB	CA
Henry Oon	30-Sep-12	35	Ontario	CANADA

Data Preparation

To continue with our cooking metaphor, we know that the process of chopping onions to a finer state will allow for its flavors to spread through a sauce more easily than that would be the case if we were to drop the whole onion into the sauce pot.

Similarly, **transforming data in the data preparation phase is the process of getting the data into a state where it may be easier to work with.**

Data Preparation

Specifically, the data preparation stage of the methodology answers the question: **What are the ways in which data is prepared?**

To work effectively with the data, it must be prepared in a way that addresses missing or invalid values and removes duplicates, toward ensuring that everything is properly formatted.

The data scientist needs to know what they're looking for within their dataset to address the question.

Data Preparation

The data preparation phase sets the stage for the next steps in addressing the question.

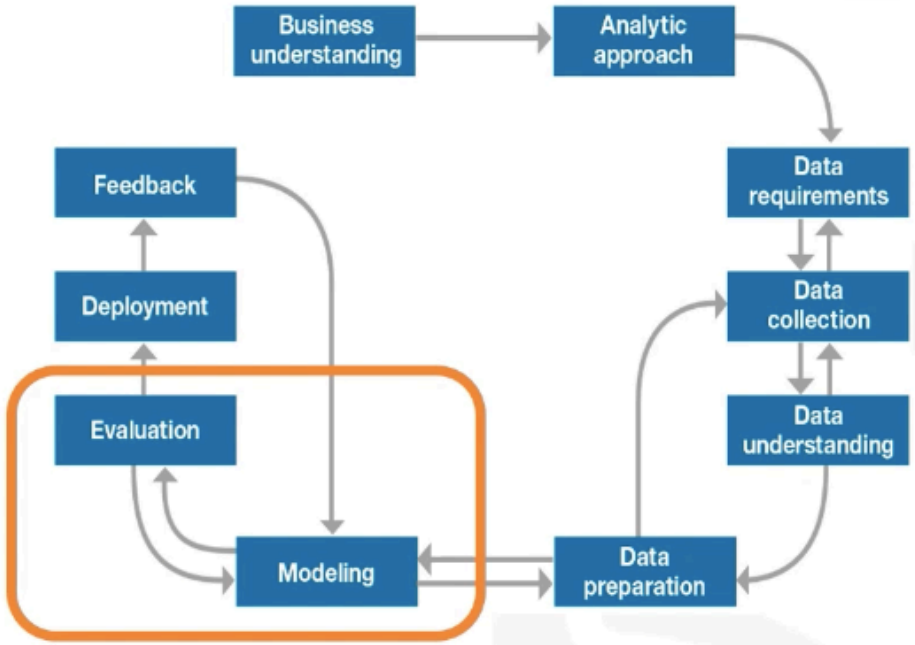
While this phase may take a while to do, **if done right the results will support the project**.

If this is skipped over, then the outcome will not be up to par and may have you back at the drawing board.

It is vital to take your time in this area, and use the tools available to automate common steps to accelerate data preparation.

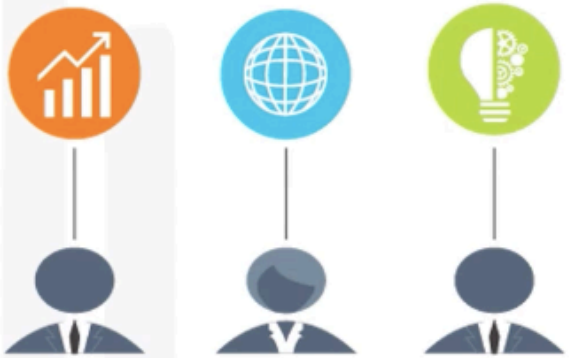
Make sure to pay attention to the detail in this area. After all, it takes just one bad ingredient to ruin a fine meal.

From Modeling to Evaluation



Modeling

- *In what way can the data be visualized to get to the answer that is required?*



Evaluation

- *Does the model used really answer the initial question or does it need to be adjusted?*

BIG DATA UNIVERSITY

Modelling

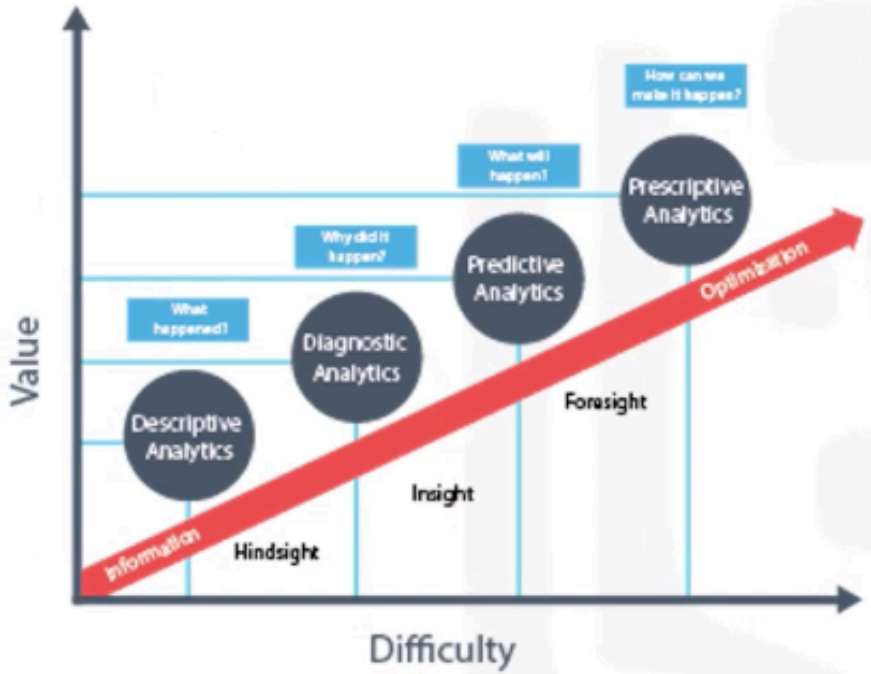
Modelling is the stage in the data science methodology **where the data scientist has the chance to sample** the sauce and determine if it's bang on or in need of more seasoning!

Data Modelling focuses on developing models that are either descriptive or predictive.

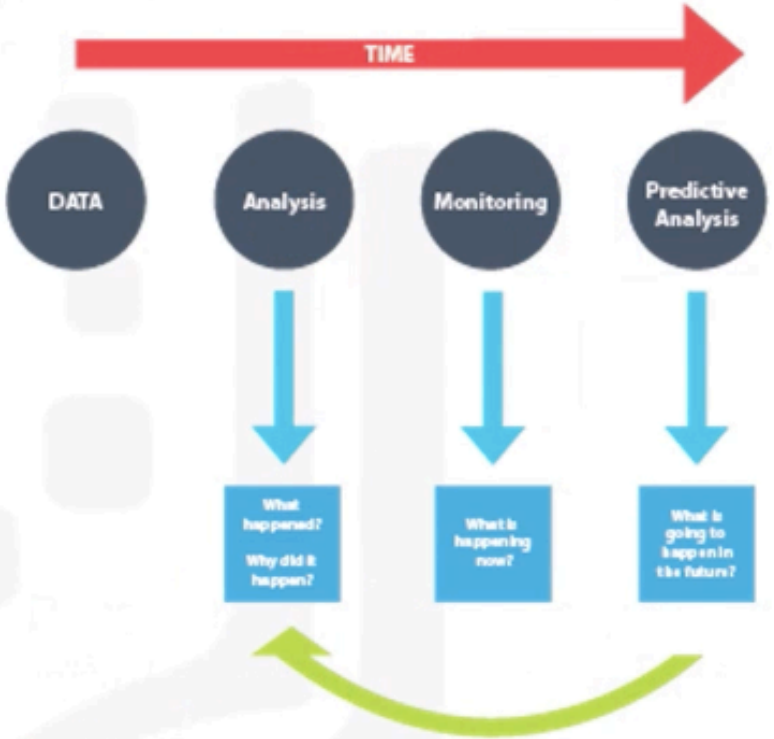
An example of a descriptive model might examine things like: **if a person did this, then they're likely to prefer that.**

Data Modeling – Using Predictive or Descriptive?

Descriptive Analytics



Predictive Analytics



Modelling

- A predictive model **tries to yield yes/no, or stop/go type outcomes**.
- These models are based on the analytic approach that was taken, either statistically driven or machine learning driven.
- The data scientist will use **a training set** for predictive modelling.
- **A training set is a set of historical data in which the outcomes are already known**.
- The training set acts like a gauge to determine if the model needs to be calibrated.
- In this stage, **the data scientist will play around with different algorithms** to ensure that the variables in play are actually required.

Modelling

The success of data compilation, preparation and modelling, depends on the understanding of the problem at hand, and the appropriate analytical approach being taken.

The data supports the answering of the question, and like the quality of the ingredients in cooking, sets the stage for the outcome.

Constant refinement, adjustments and tweaking are necessary within each step to ensure the outcome is one that is solid.

Modelling

In John Rollins' descriptive Data Science Methodology, the framework is geared to do 3 things:

First, **understand the question at hand.**

Second, **select an analytic approach or method to solve the problem**, and

third, **obtain, understand, prepare, and model the data**.

The end goal is to move the data scientist to a point where a data model can be built to answer the question.

Understanding the question

1. Understand the question at hand
2. Select an analytic approach or method to solve the problem
3. Obtain, understand, prepare, and model the data

Modelling

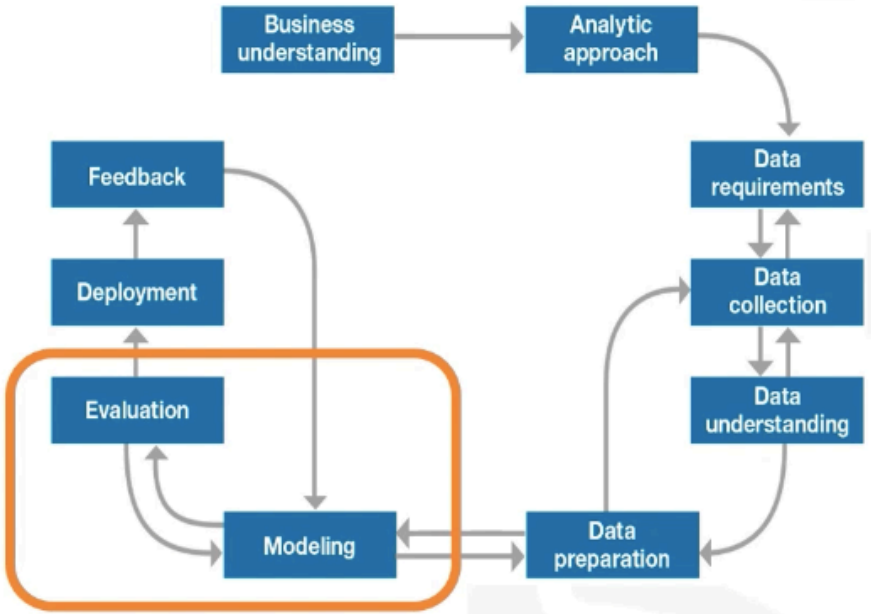
This relevance is critical to the data science field overall, as it is a fairly new field of study, and we are interested in the possibilities it has to offer.

The more people that benefit from the outcomes of this practice, the further the field will develop.

With dinner just about to be served and a hungry guest at the table, the key question is: **Have I made enough to eat? Well, let's hope so.**

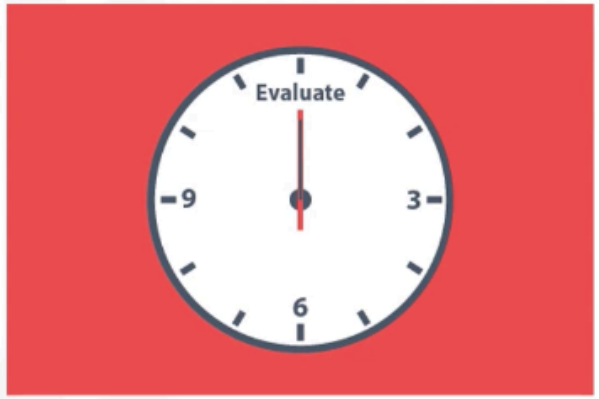
In this stage of the methodology, model evaluation, deployment, and feedback **loops** ensure that the answer is near and relevant.

From Modeling to Evaluation



Modeling

- *In what way can the data be visualized to get to the answer that is required?*



Evaluation

- *Does the model used really answer the initial question or does it need to be adjusted?*

BIG DATA UNIVERSITY

Evaluation

A model evaluation goes hand-in-hand with model building as such, the modeling and evaluation stages are **done iteratively.**

Model evaluation **is performed during model development** and before the model is deployed.

Evaluation **allows the quality of the model to be assessed** but it's **also an opportunity to see if it meets the initial request.**

Evaluation answers the question: **Does the model used really answer the initial question or does it need to be adjusted?**

Evaluation

Model evaluation can have two main phases.

The first is **the diagnostic measures phase**, which is used to ensure the model is working as intended.

If the model is a predictive model, a decision tree can be used to evaluate if the answer the model can output, is aligned to the initial design.

It can be used to see where there are areas that require adjustments.

Evaluation

If the model is a descriptive model, one in which relationships are being assessed, **then a testing set with known outcomes can be applied, and the model can be refined as needed.**

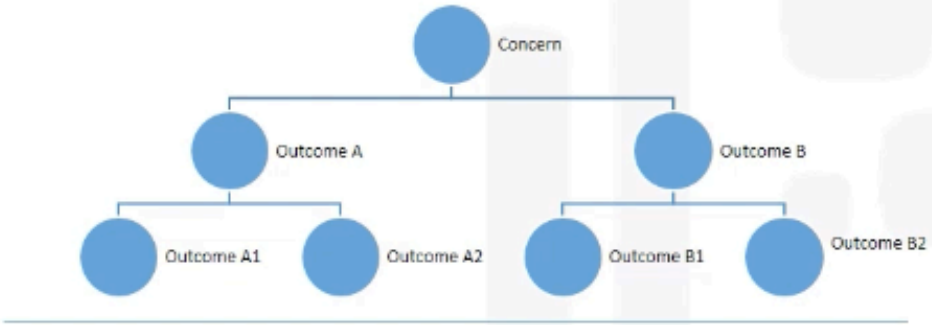
The second phase of evaluation that may be used is statistical significance testing.

This type of evaluation can be applied to the model to ensure that the data is being properly handled and interpreted within the model.

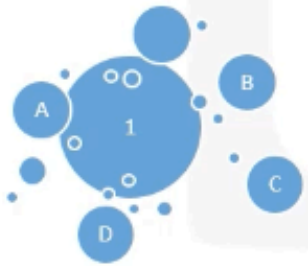
When and how to adjust the model?

Diagnostic measures

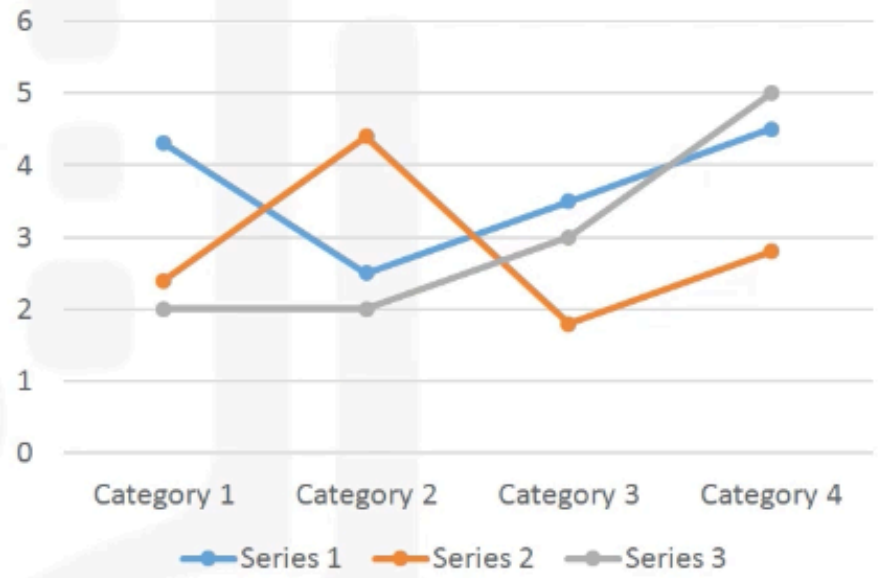
Predictive Model



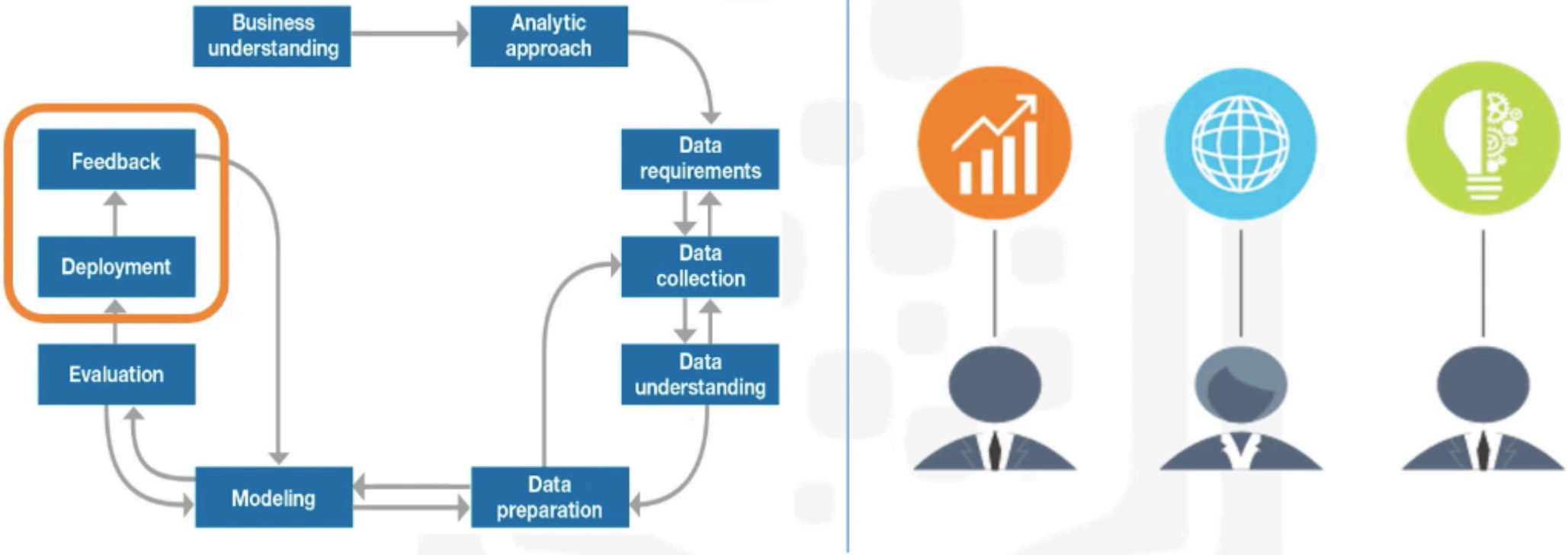
Descriptive Model



Statistical Significance



From Deployment to Feedback



Deployment

While a data science model will provide an answer, **the key to making the answer relevant and useful to address the initial question,** involves getting the stakeholders familiar with the tool produced.

In a business scenario, **stakeholders have different specialties that will help make this happen,** such as the solution owner, marketing, application developers, and IT administration.

Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test.

Deployment

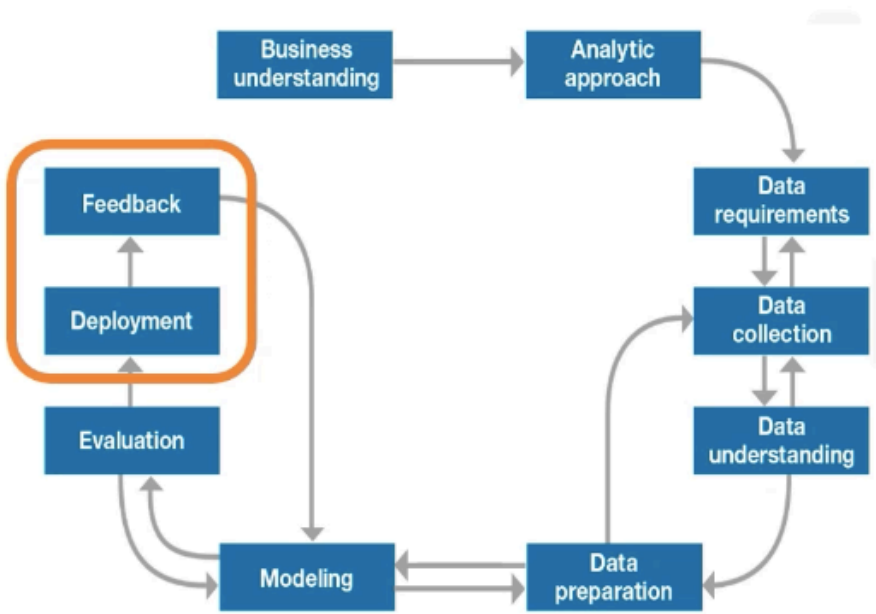
Depending on the purpose of the model, it may be rolled out to a limited group of users or in a test environment, to build up confidence in applying the outcome for use across the board.

So now, let's look at the case study related to applying Deployment"

In preparation for solution deployment, **the next step was to assimilate the knowledge for the business group** who would be designing and managing the intervention program to reduce readmission risk.

In this scenario, the business people **translated the model results** so that the clinical staff could understand how to identify high-risk patients and design suitable intervention actions.

Feedback – Problem solved? Question answered?



Feedback

Once in play, **feedback from the users will help to refine the model** and assess it for performance and impact.

The value of the model will be dependent on successfully incorporating feedback and making adjustments for as long as the solution is required.

Throughout the Data Science Methodology, **each step sets the stage for the next.**

Making the methodology cyclical, ensures refinement at each stage in the game.

The feedback **process is rooted in the notion that, the more you know, the more that you'll want to know**.

Feedback

Once the model is evaluated and the data scientist is confident it'll work, it is deployed and put to the ultimate test: **actual, real-time use in the field.**

So now, let's look at our case study again, to see how the Feedback portion of the methodology is applied.

From Deployment to Feedback

Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test

- Actual real-time use in the field

BIG DATA UNIVERSITY

Feedback

After the deployment and feedback stages, the impact of the program would be reviewed after period its implementation.

Then the model would be refined, based on all of the data compiled after model implementation and the knowledge gained throughout these stages.

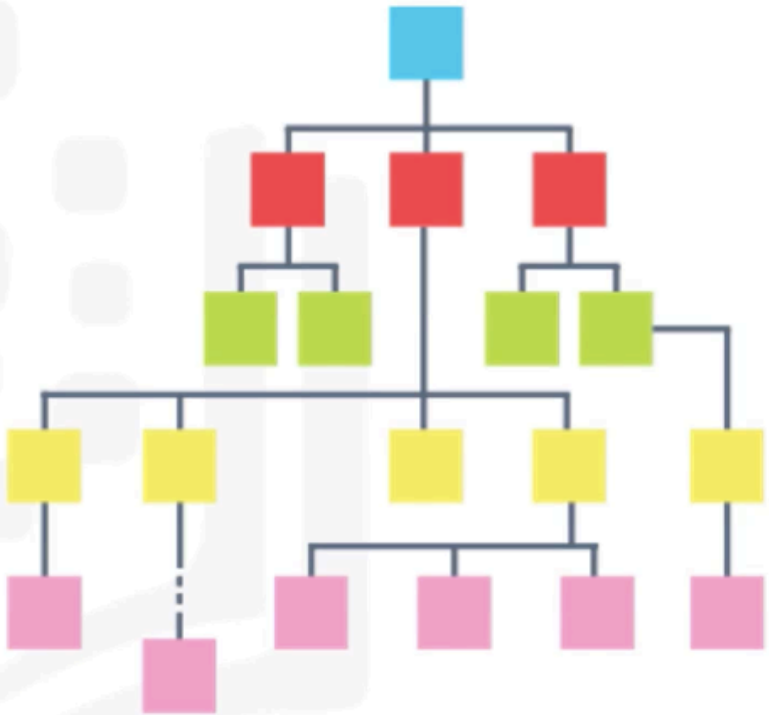
Other refinements included: Incorporating information about participation in the intervention program, and possibly refining the model to incorporate detailed pharmaceutical data.

Case Study – Redeployment

Review and refine intervention actions

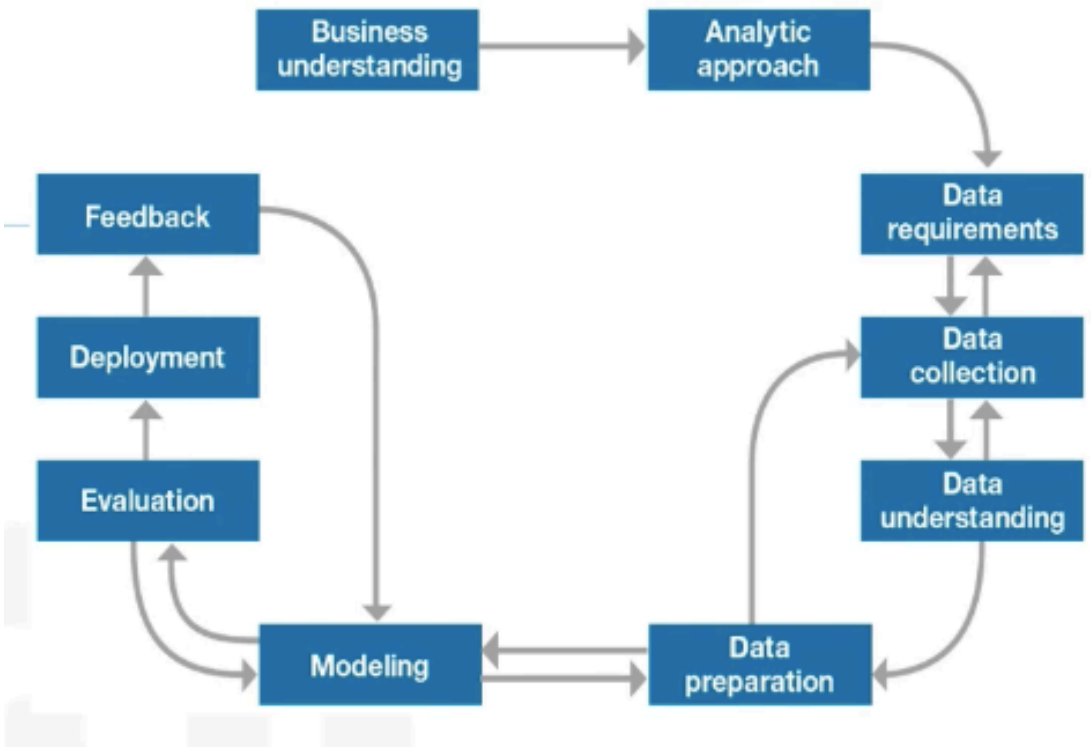
Redeploy

- Continue modeling, deployment, feedback, and refinement throughout the life of the intervention program



BIG DATA UNIVERSITY

Summary



Data Science Methodology's purpose is **to share a methodology that can be used within data science, to ensure that the data used in problem solving is relevant and properly manipulated** to address the question at hand.

Summary

Data science methodology **begins with spending the time to seek clarification, to attain what can be referred to as a business understanding.**

Having this understanding is placed at the beginning of the methodology because getting clarity around the problem to be solved, allows you to determine which data will be used to answer the core question.

Once a strong understanding of the question is established, the analytic approach can be selected.

This means identifying what type of patterns will be needed to address the question most effectively.

Building on the understanding of the problem at hand, and then using the analytical approach selected, **the Data Scientist is ready to get started.**

This includes identifying the necessary data content, formats and sources for initial data collection.

After the initial data collection is performed, **an assessment by the data scientist takes place** to determine whether or not they have what they need.

Summary

Data understanding encompasses **all activities related to constructing the data set**. Essentially, the data understanding section of the data science methodology answers the question: **Is the data that you collected representative of the problem to be solved?**

Together with **data collection and data understanding, data preparation is the most time-consuming phase** of a data science project, typically taking seventy percent and even up to even ninety percent of the overall project time.

Summary

Specifically, the data preparation stage of the methodology answers the question: **What are the ways in which data is prepared?**

To work effectively with the data, it must be prepared in a way that addresses missing or invalid values and removes duplicates, toward ensuring that everything is properly formatted.

Modelling is the stage in the data science methodology **where the data scientist has the chance to sample** the sauce and determine if it's bang on or in need of more seasoning!

Data Modelling focuses on developing models that are either descriptive or predictive.

Summary

A model evaluation goes hand-in-hand with model building as such, the modeling and evaluation stages are **done iteratively.**

Model evaluation **is performed during model development** and before the model is deployed.

Evaluation **allows the quality of the model to be assesse**d but it's **also an opportunity to see if it meets the initial request.**

Summary

While a data science model will provide an answer, **the key to making the answer relevant and useful to address the initial question,** involves getting the stakeholders familiar with the tool produced.

In a business scenario, **stakeholders have different specialties that will help make this happen,** such as the solution owner, marketing, application developers, and IT administration.

Once in play, **feedback from the users will help to refine the model** and assess it for performance and impact.

The value of the model will be dependent on successfully incorporating feedback and making adjustments for as long as the solution is required.