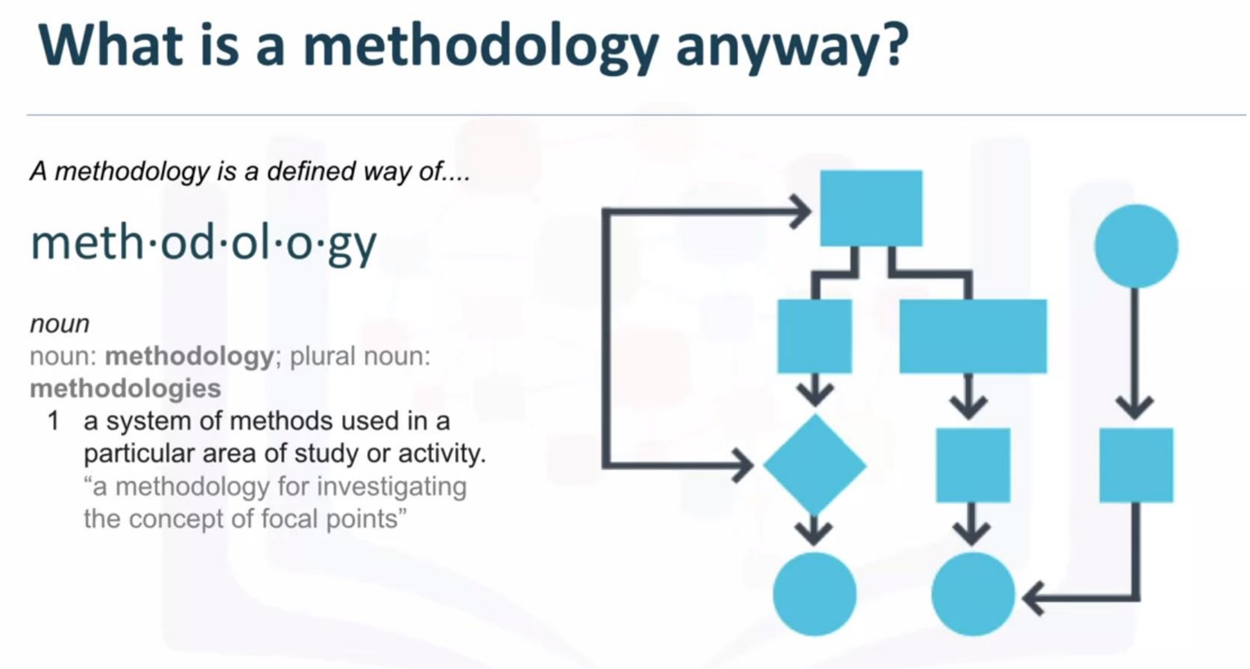
Data science methodology



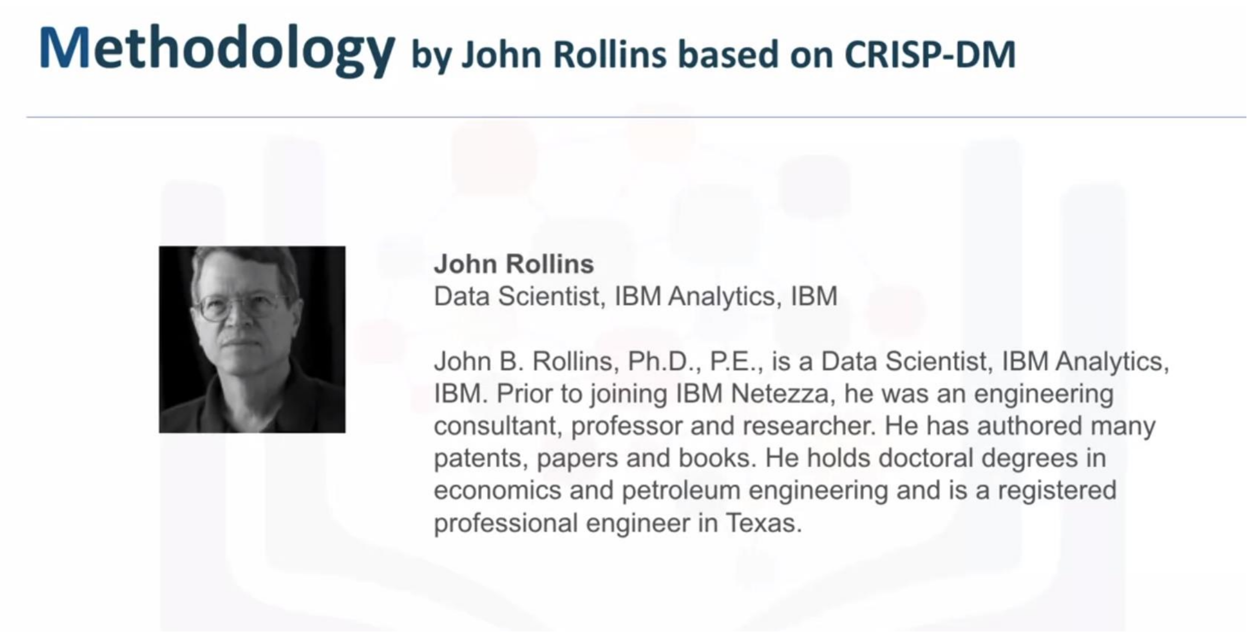
What is a methodology anyway?

A methodology is a defined way of....

meth·od·ol·o·gy

noun
noun: **methodology**; plural noun: **methodologies**

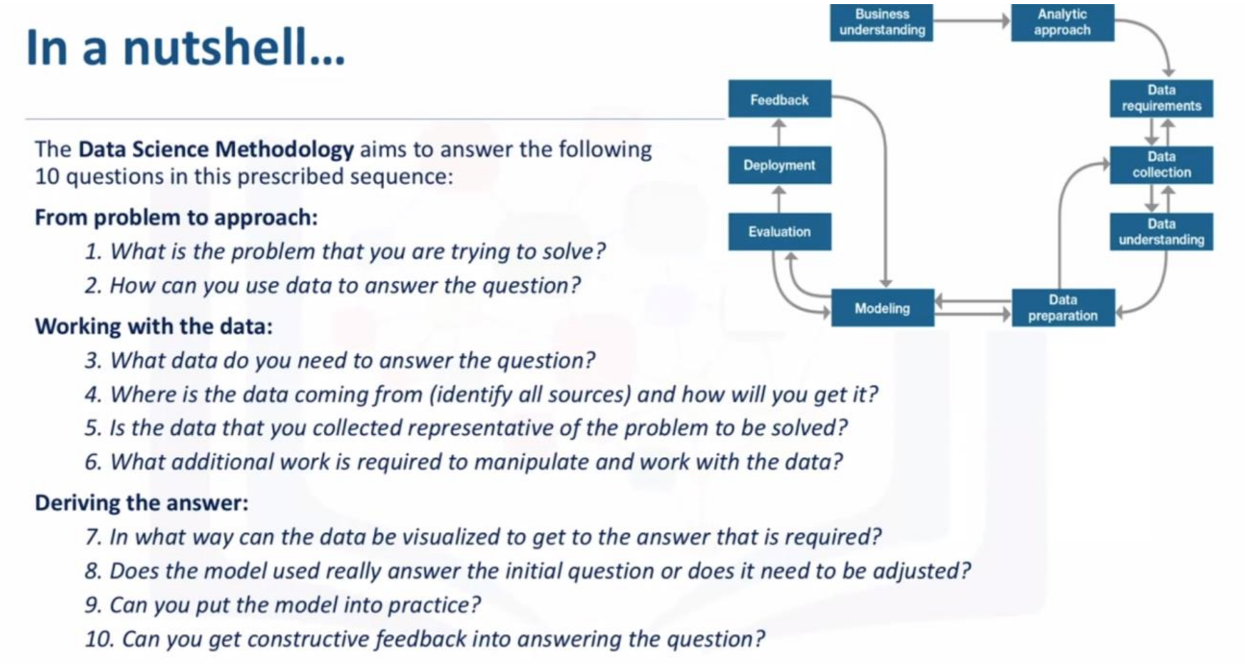
1 a system of methods used in a particular area of study or activity.
"a methodology for investigating the concept of focal points"



Methodology by John Rollins based on CRISP-DM

John Rollins
Data Scientist, IBM Analytics, IBM

John B. Rollins, Ph.D., P.E., is a Data Scientist, IBM Analytics, IBM. Prior to joining IBM Netezza, he was an engineering consultant, professor and researcher. He has authored many patents, papers and books. He holds doctoral degrees in economics and petroleum engineering and is a registered professional engineer in Texas.



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?*

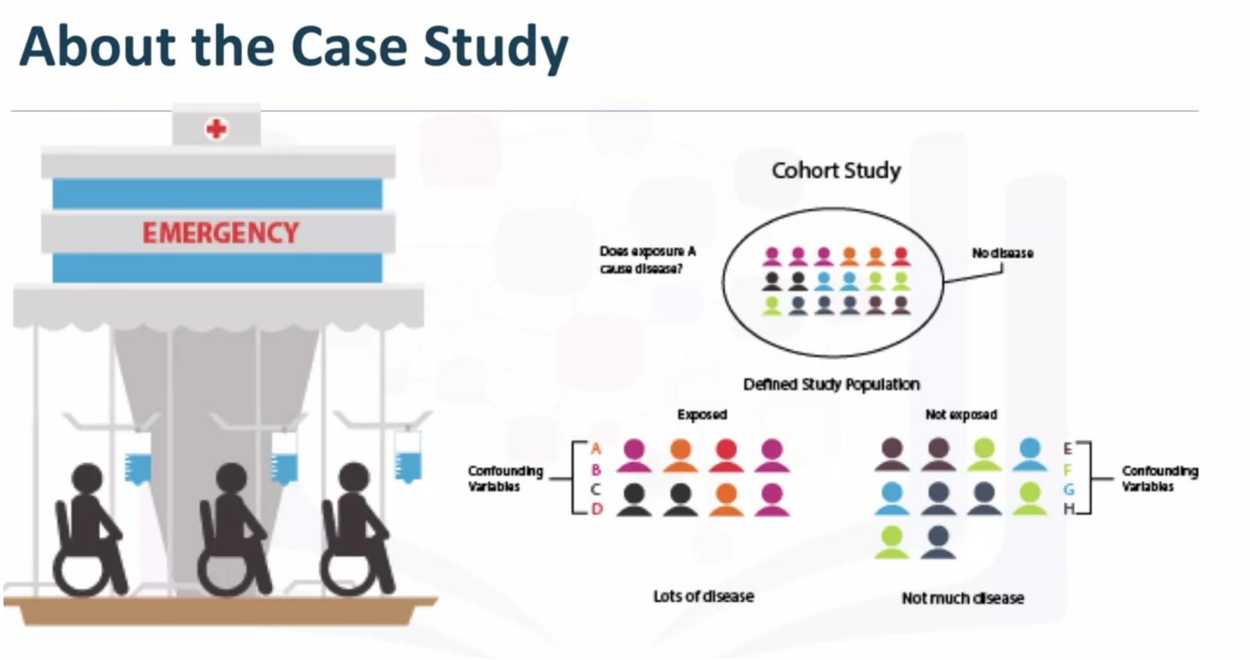
Working with the data:

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?*

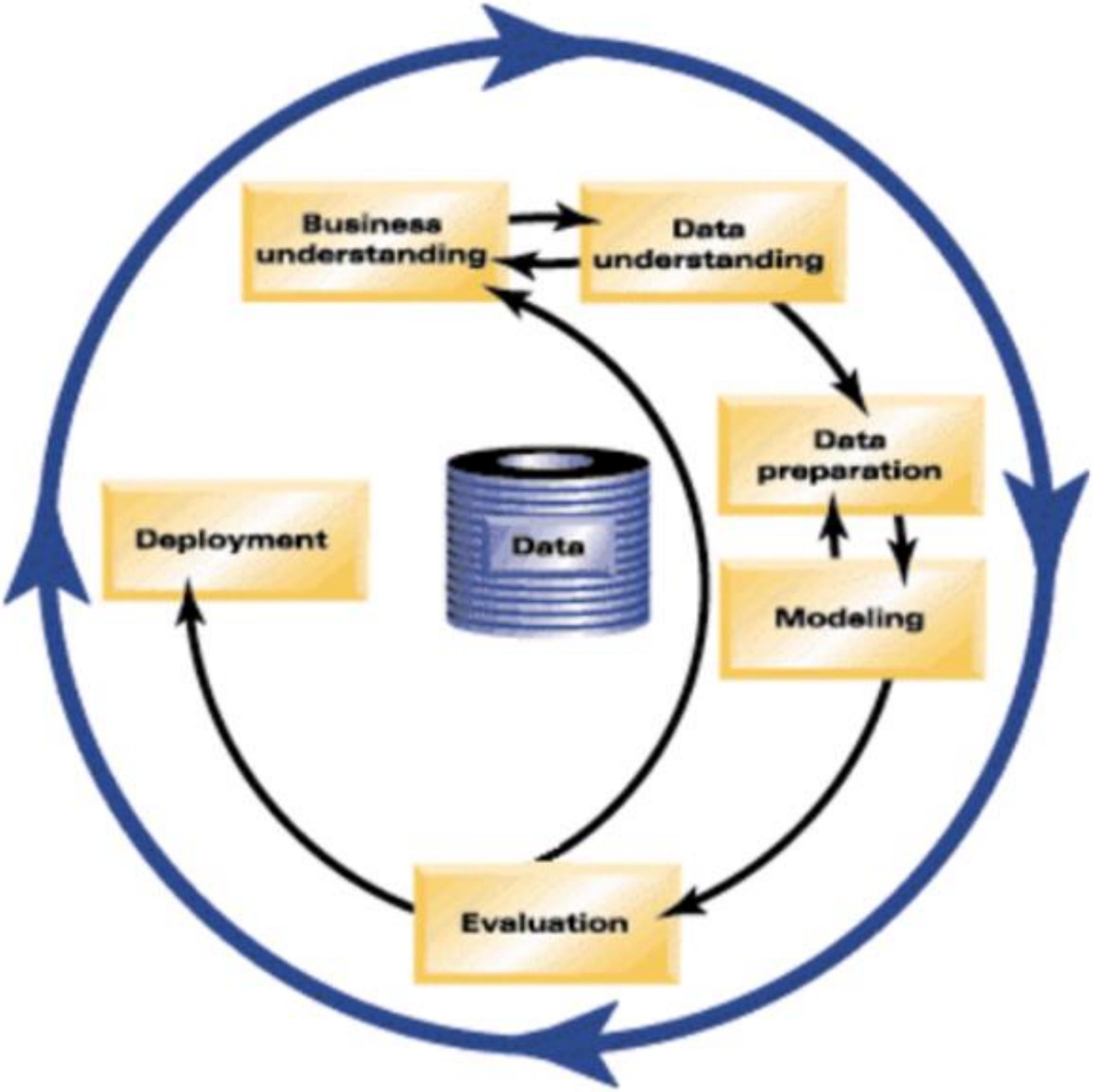
The case study included in the course, highlights how the data science methodology can be applied in context.
It revolves around the following scenario
It revolves around the following scenario: There is a limited budget for providing healthcare
in the system to properly address the patient condition prior to the initial patient discharge.
The core question is: What is the best way to allocate these funds to maximize their
use in providing quality care? As you'll see, if the new data science pilot
program is successful, it will deliver better patient care by giving physicians new tools
to incorporate timely, data-driven information into patient care decisions.



About the Case Study

What is CRISP-DM?

The CRISP-DM methodology is a process aimed at increasing the use of data mining over a wide variety of business applications and industries. The intent is to take case specific scenarios and general behaviors to make them domain neutral.



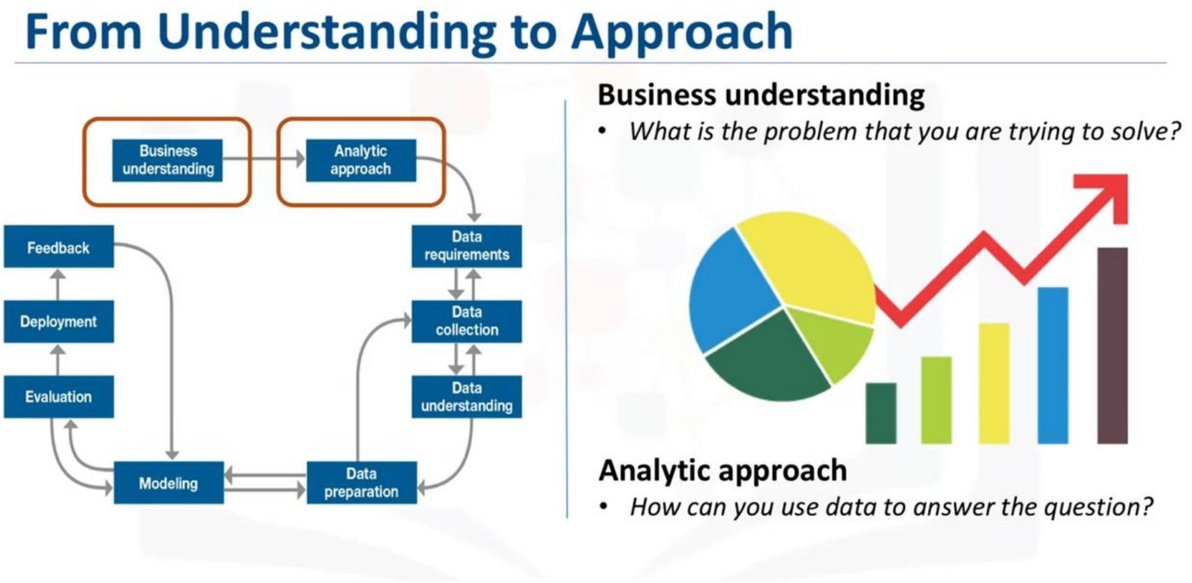
1. **Business Understanding** This stage is the most important because this is where the intention of the project is outlined. Foundational Methodology and CRISP-DM are aligned here. It requires communication and clarity. The difficulty here is that stakeholders have different objectives, biases, and modalities of relating information. They don’t all see the same things or in the same manner. Without clear, concise,

and complete perspective of what the project goals are resources will be needlessly expended.

2. **Data Understanding** Data understanding relies on business understanding. Data is collected at this stage of the process. The understanding of what the business wants and needs will determine what data is collected, from what sources, and by what methods. CRISP-DM combines the stages of Data Requirements, Data Collection, and Data Understanding from the Foundational Methodology outline.
3. **Data Preparation** Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. Data Preparation is common to CRISP-DM and Foundational Methodology.
4. **Modeling** Once prepared for use, the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge. This is the purpose of data mining: to create knowledge information that has meaning and utility. The use of models reveals patterns and structures within the data that provide insight into the features of interest. Models are selected on a portion of the data and adjustments are made if necessary. Model selection is an art and science. Both Foundational Methodology and CRISP-DM are required for the subsequent stage.
5. **Evaluation** The selected model must be tested. This is usually done by having a pre-selected test, set to run the trained model on. This will allow you to see the effectiveness of the model on a set it sees as new. Results from this are

used to determine efficacy of the model and foreshadows its role in the next and final stage.

6. **Deployment** In the deployment step, the model is used on new data outside of the scope of the dataset and by new stakeholders. The new interactions at this phase might reveal the new variables and needs for the dataset and model. These new challenges could initiate revision of either business needs and actions, or the model and data, or both.



From Understanding to Approach

Business understanding

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

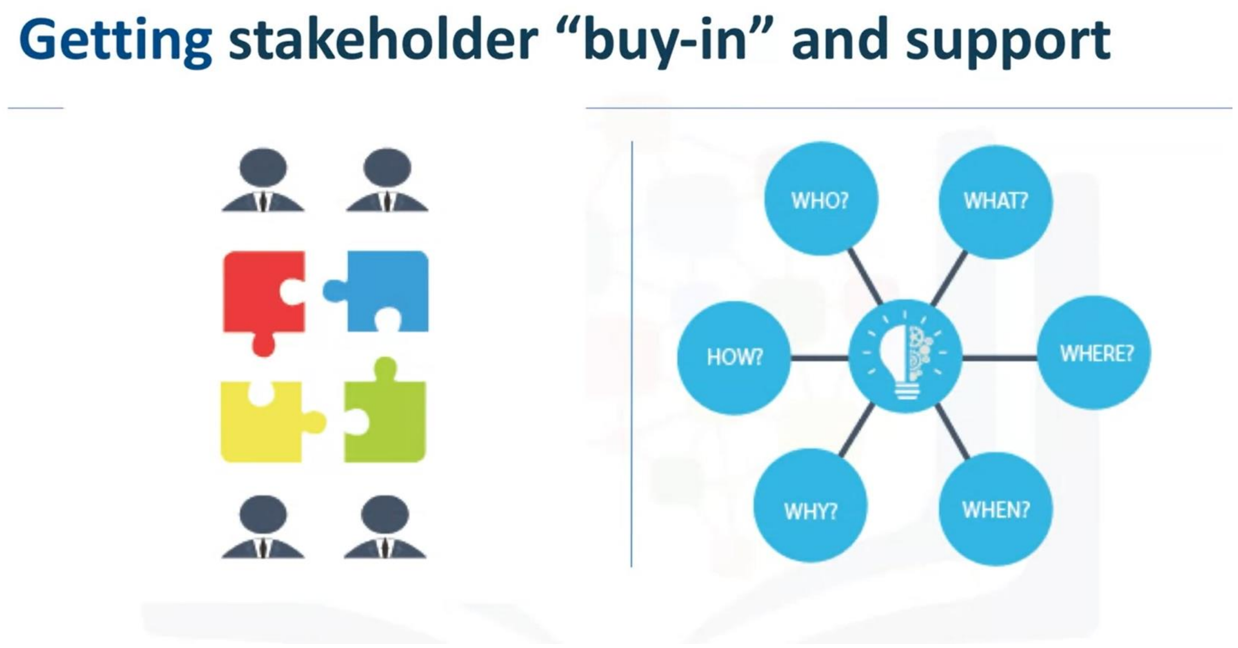
Analytic approach

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

Seeking clarification – What's the goal?

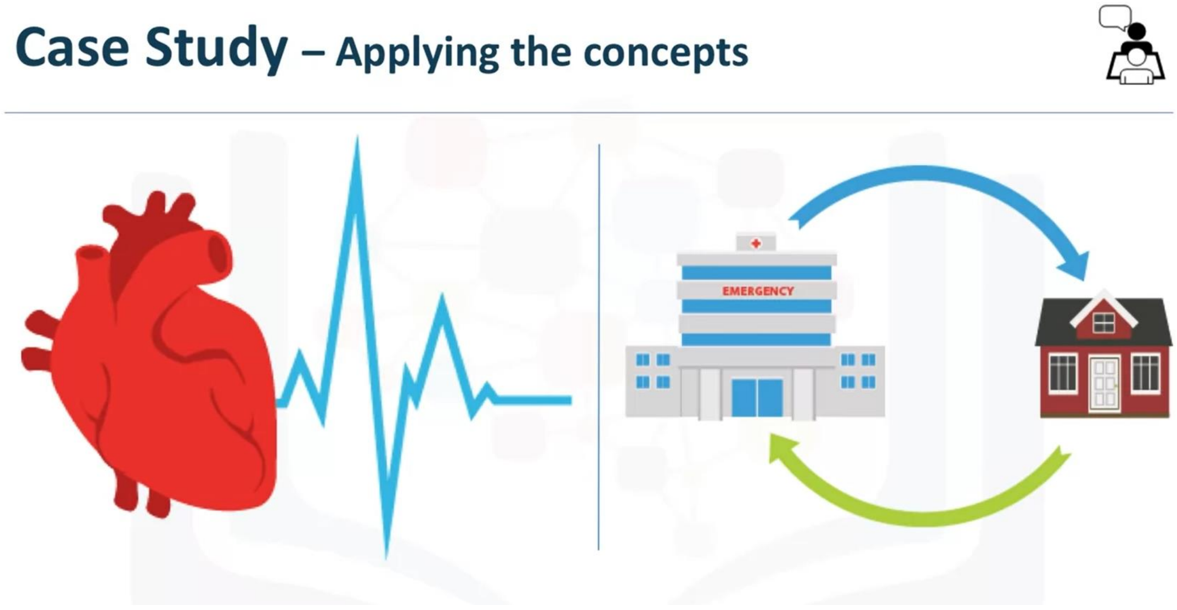
if a business owner asks: "How can we reduce the costs of performing an activity?"

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.



Getting stakeholder “buy-in” and support

Stakeholders should be used in order to clarify questions and to determine requirments



Case Study – Applying the concepts

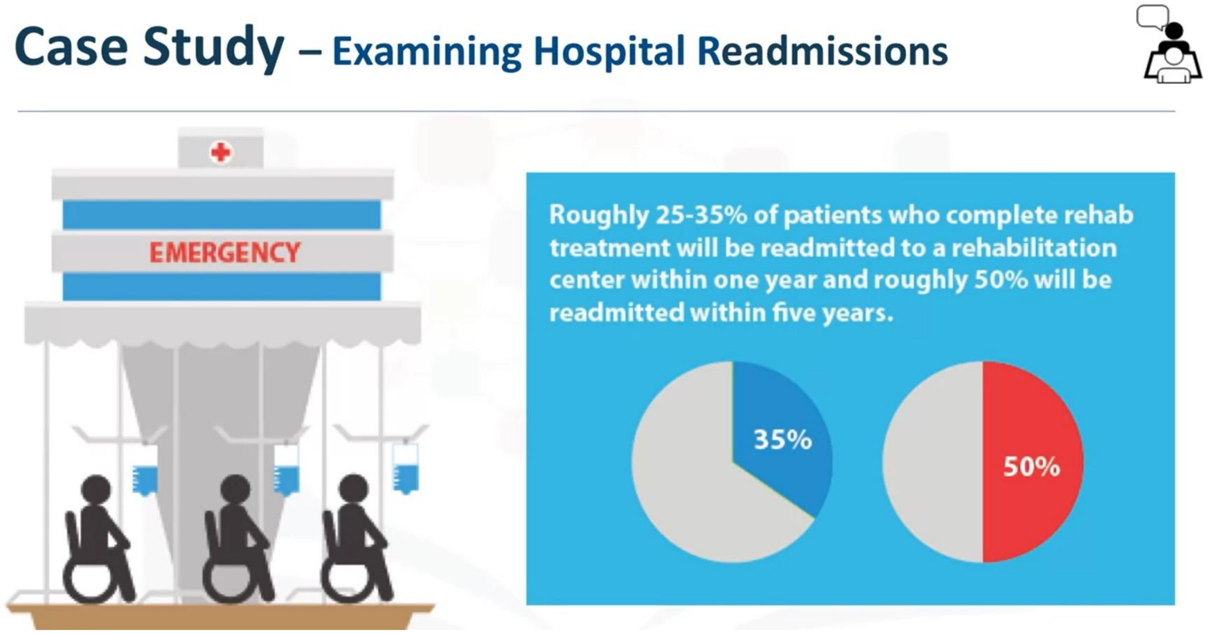
In the case study, the question being asked is: What is the best way to allocate the limited
healthcare budget to maximize its use in providing quality care?

This question is one that became a hot topic for an American healthcare insurance provider.
As public funding for readmissions was decreasing, this insurance company was at risk of having
to make up for the cost difference,which could potentially increase rates for its customers.
how data science could be applied to the question at hand.

Case Study – What are the goals & objectives?

Define the GOALS
- To provide quality care without increasing costs

Define the OBJECTIVES
- To review the process to identify inefficiencies

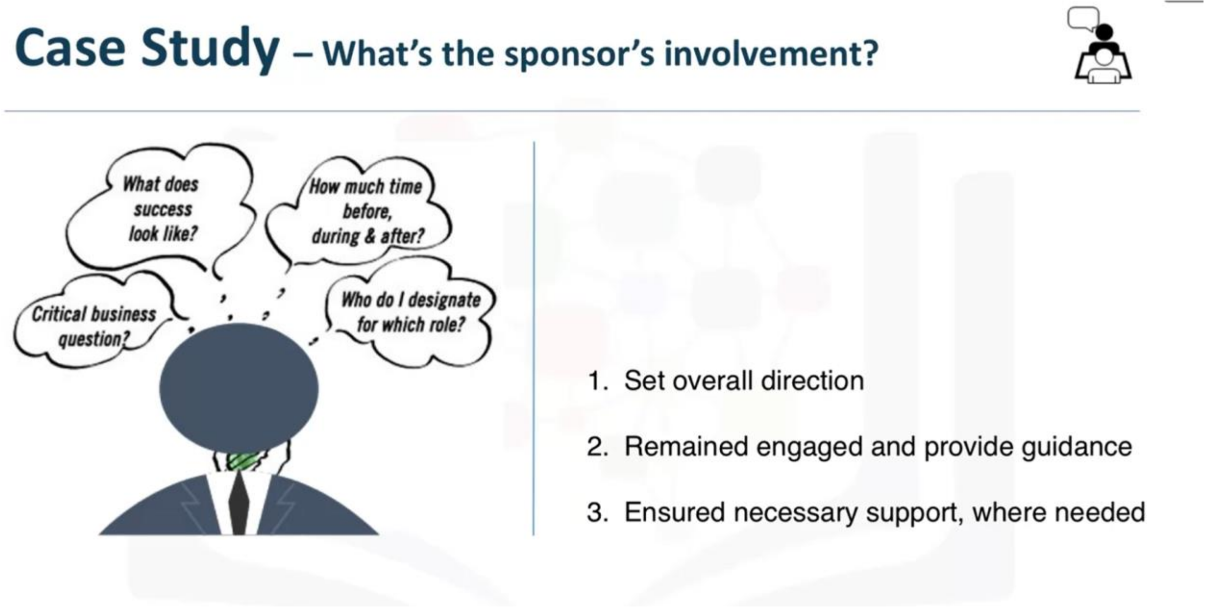


Case Study – Examining Hospital Readmissions

Roughly 25-35% of patients who complete rehab treatment will be readmitted to a rehabilitation center within one year and roughly 50% will be readmitted within five years.

After reviewing some records, it was discovered that the patients with congestive heart failure
were at the top of the readmission list

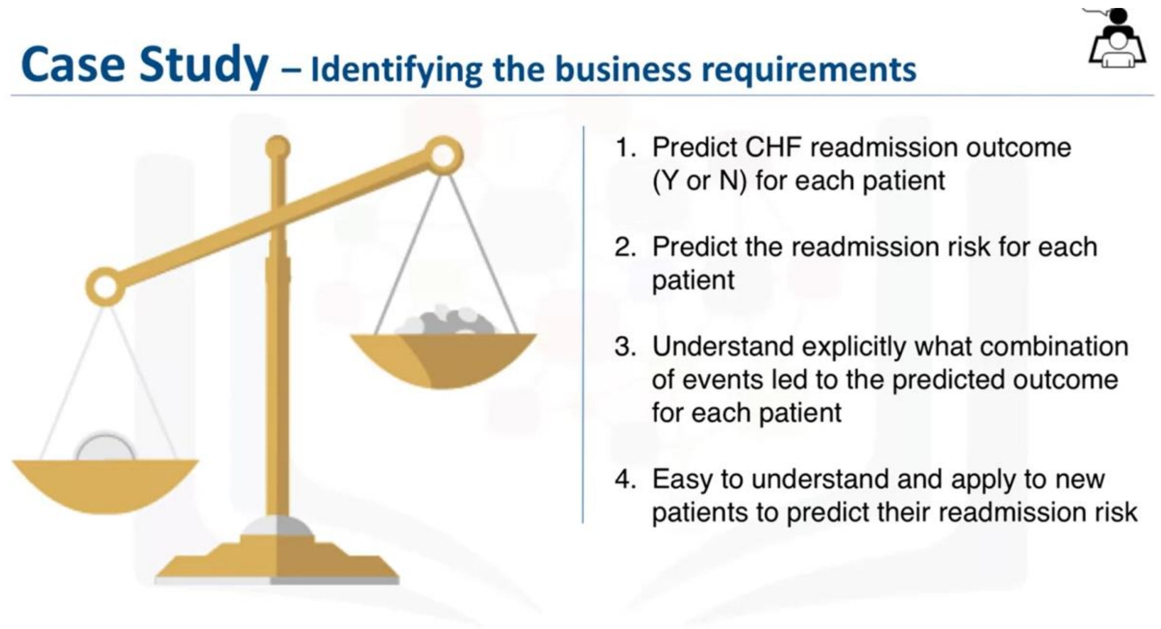
It was further determined that a decision-tree model could be applied to review this scenario,
to determine why this was occurring.



Case Study – What's the sponsor's involvement?

What does success look like?
How much time before, during & after?
Critical business question?
Who do I designate for which role?

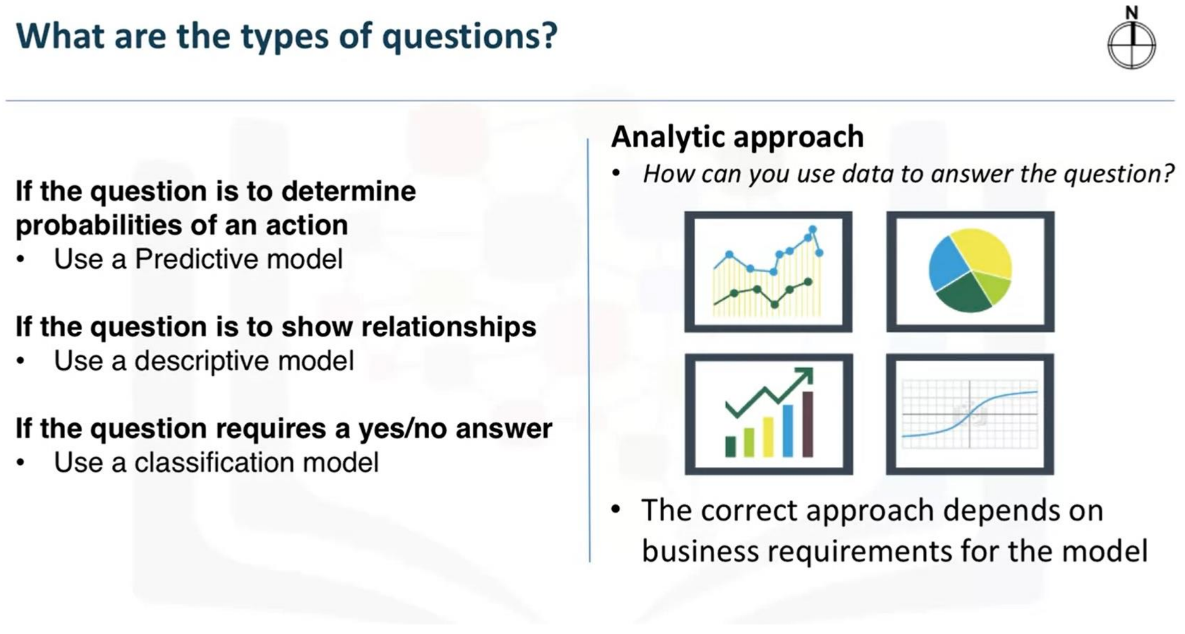
1. Set overall direction
2. Remained engaged and provide guidance
3. Ensured necessary support, where needed



Case Study – Identifying the business requirements

1. Predict CHF readmission outcome (Y or N) for each patient
2. Predict the readmission risk for each patient
3. Understand explicitly what combination of events led to the predicted outcome for each patient
4. Easy to understand and apply to new patients to predict their readmission risk

Analytic approach



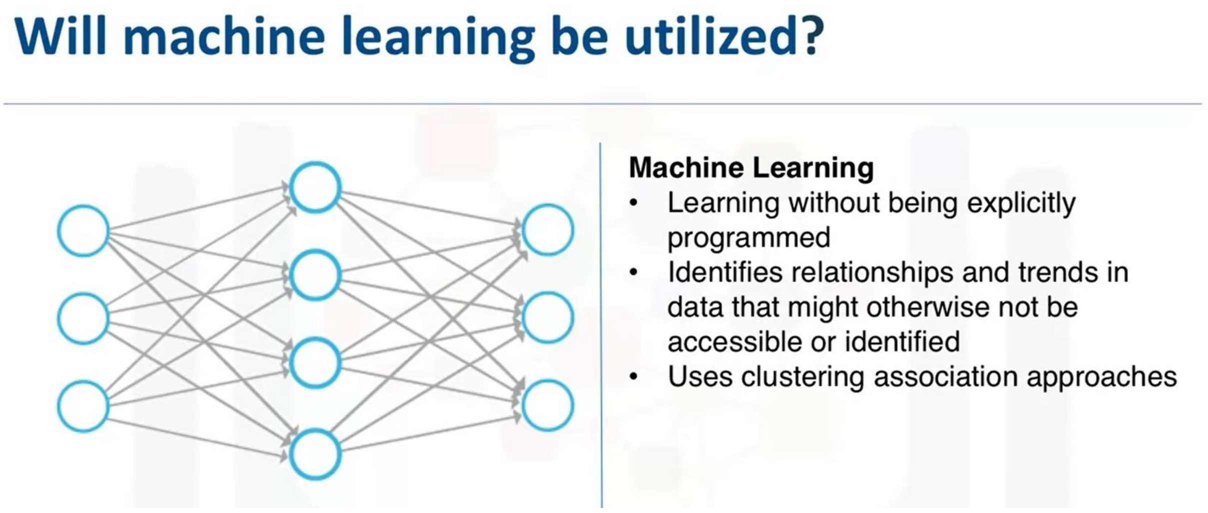
What are the types of questions?

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

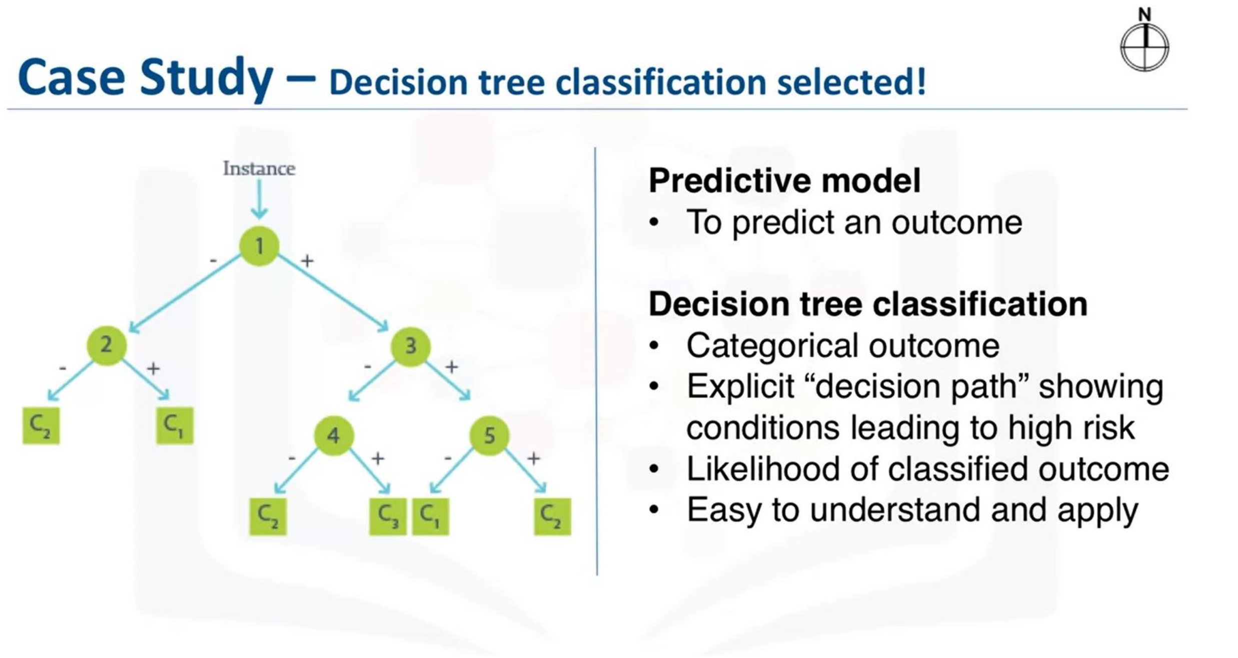
Analytic approach
- *How can you use data to answer the question?*
- The correct approach depends on business requirements for the model



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



Case Study – Decision tree classification selected!

Predictive model

- To predict an outcome

Decision tree classification

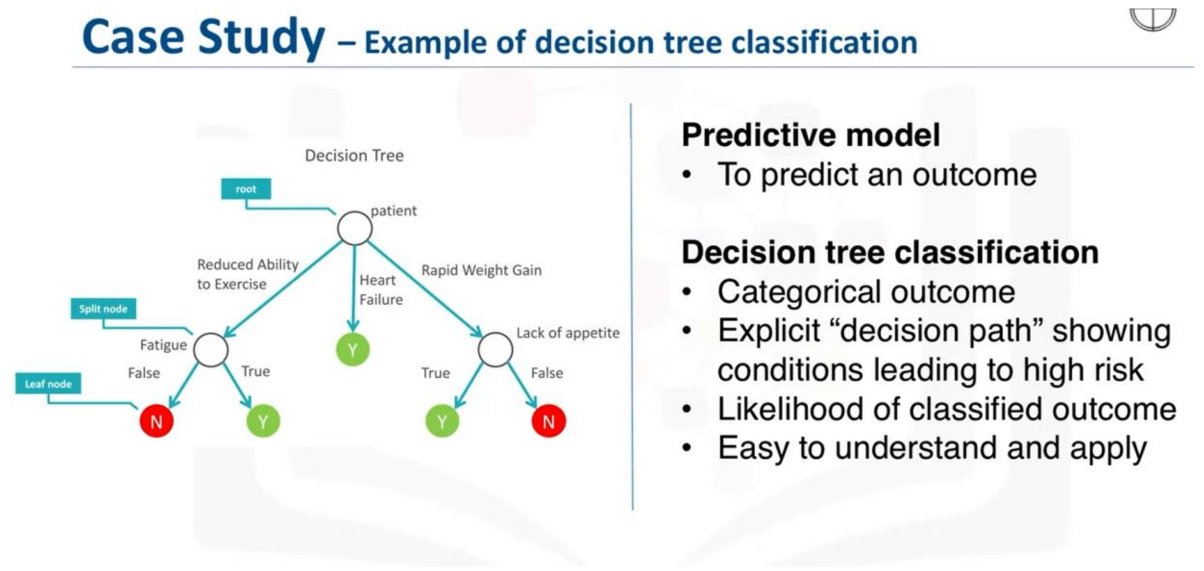
- Categorical outcome
- Explicit "decision path" showing conditions leading to high risk
- Likelihood of classified outcome
- Easy to understand and apply

Why is the analytic approach stage important?

Because it helps identify what type of patterns will be needed to address the question most effectively.

Why is the business understanding stage important?

It helps clarify the goal of the entity asking the question.



Case Study – Example of decision tree classification

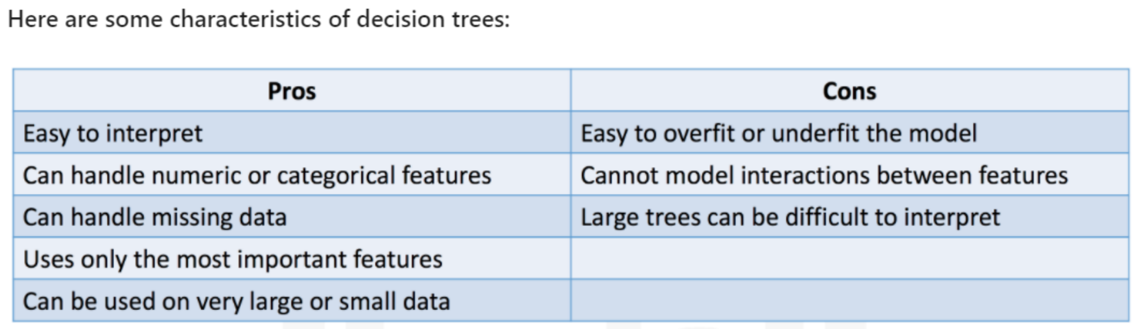
Predictive model

- 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

Decision tree



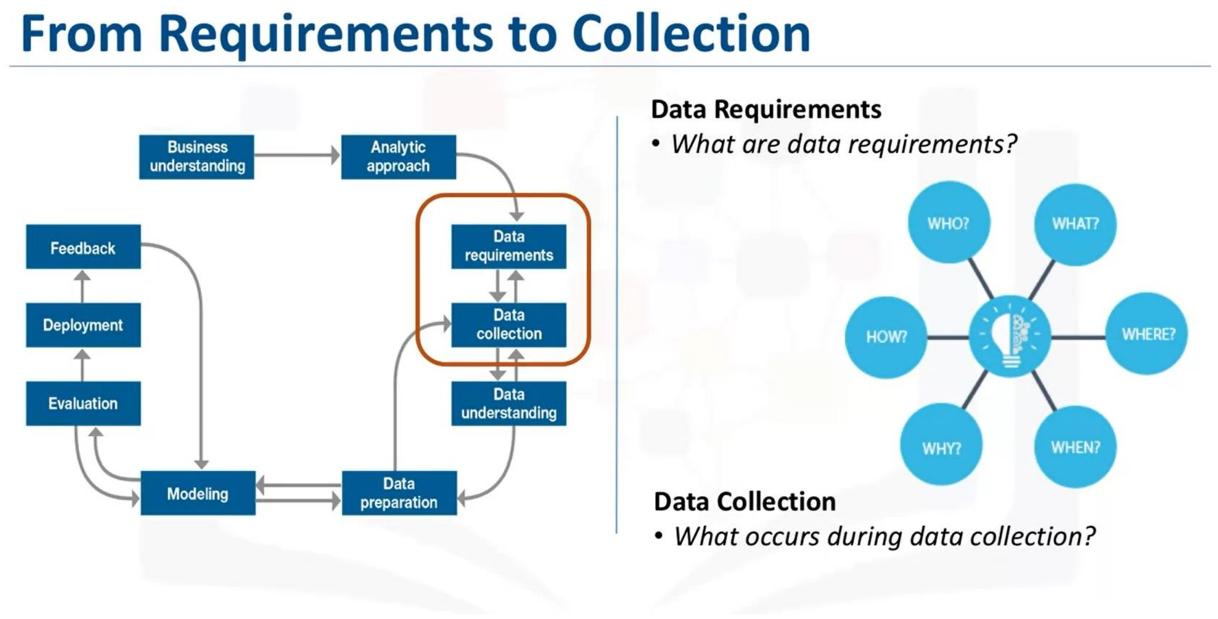
Here are some characteristics of decision trees:

Pros	Cons
Easy to interpret	Easy to overfit or underfit the model
Can handle numeric or categorical features	Cannot model interactions between features
Can handle missing data	Large trees can be difficult to interpret
Uses only the most important features	
Can be used on very large or small data	

, 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 to collect them, how to understand or work with them, and how to prepare the data to meet the desired
outcome.

it's vital to define the data requirements for decision-tree classification.
This includes identifying the necessary data content, formats and sources for initial data

collection.



From Requirements to Collection

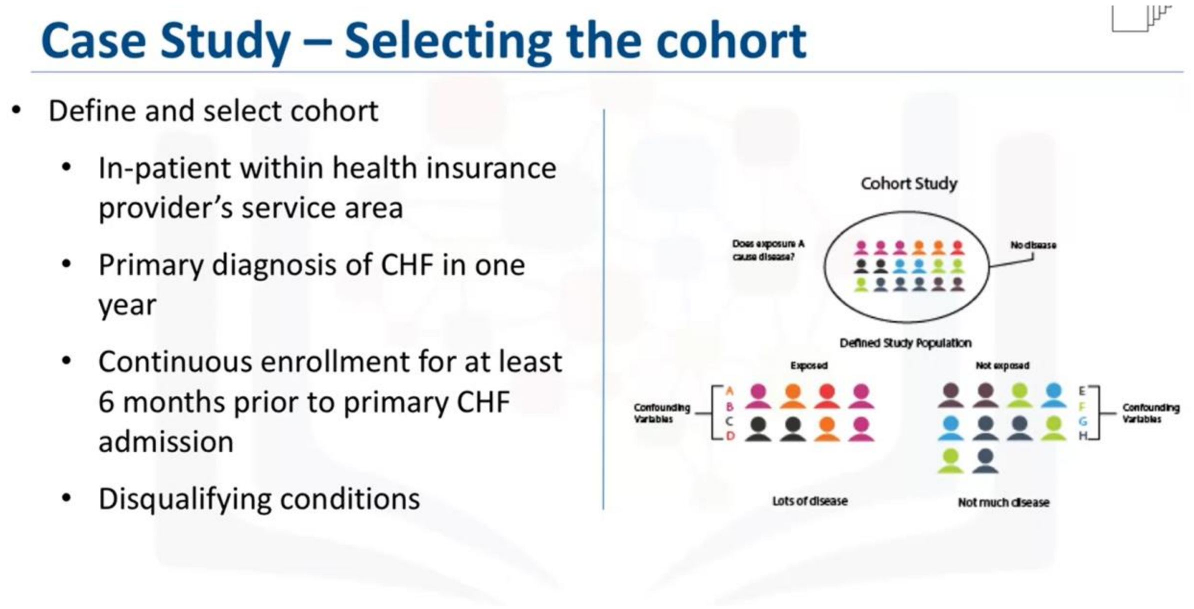
Data Requirements

- *What are data requirements?*

Data Collection

- *What occurs during data collection?*

it's vital to define the data requirements for decision-tree classification.
This includes identifying the necessary data content, formats and sources for initial data
collection.

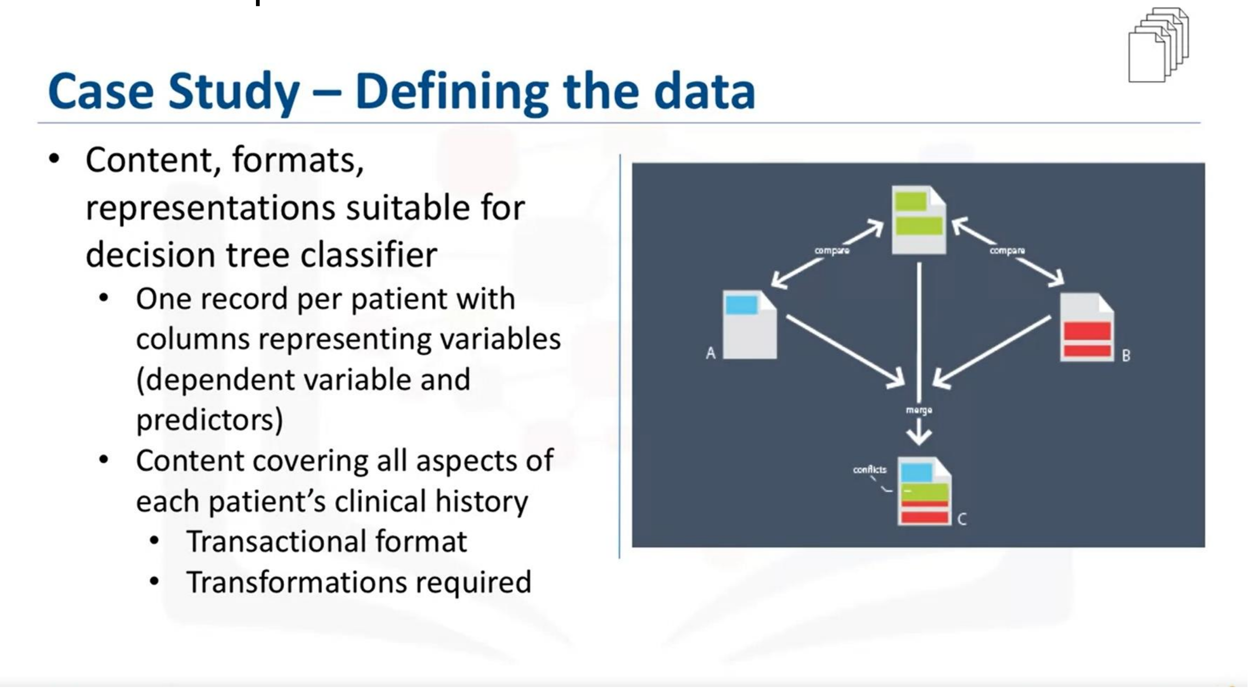


Case Study – Selecting the cohort

- Define and select cohort
  - In-patient within health insurance provider's service area
  - Primary diagnosis of CHF in one year
  - Continuous enrollment for at least 6 months prior to primary CHF admission
  - Disqualifying conditions

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.



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

Data collection

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

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
its case study : Collecting data requires that you know the source or, know where to find the data elements

that are needed.
In the context of our case study, these can include:
demographic, clinical and coverage information of patients, provider information, claims records, as well as
pharmaceutical and other information related to all the diagnoses of the congestive heart
failure patients
When collecting data, it is alright to defer decisions about unavailable data, and attempt to acquire it at a later stage.

DBAs and programmers often work together to extract data from various sources, and then
merge it.
This allows for removing redundant data, making it available for the next stage of the methodology,
which is data understanding.

In the initial data collection stage, data scientists identify and gather the available data resources. These can be in the form of structured, unstructured, and even semi-structured data relevant to the problem domain.

Web Scraping of Online Food Recipes

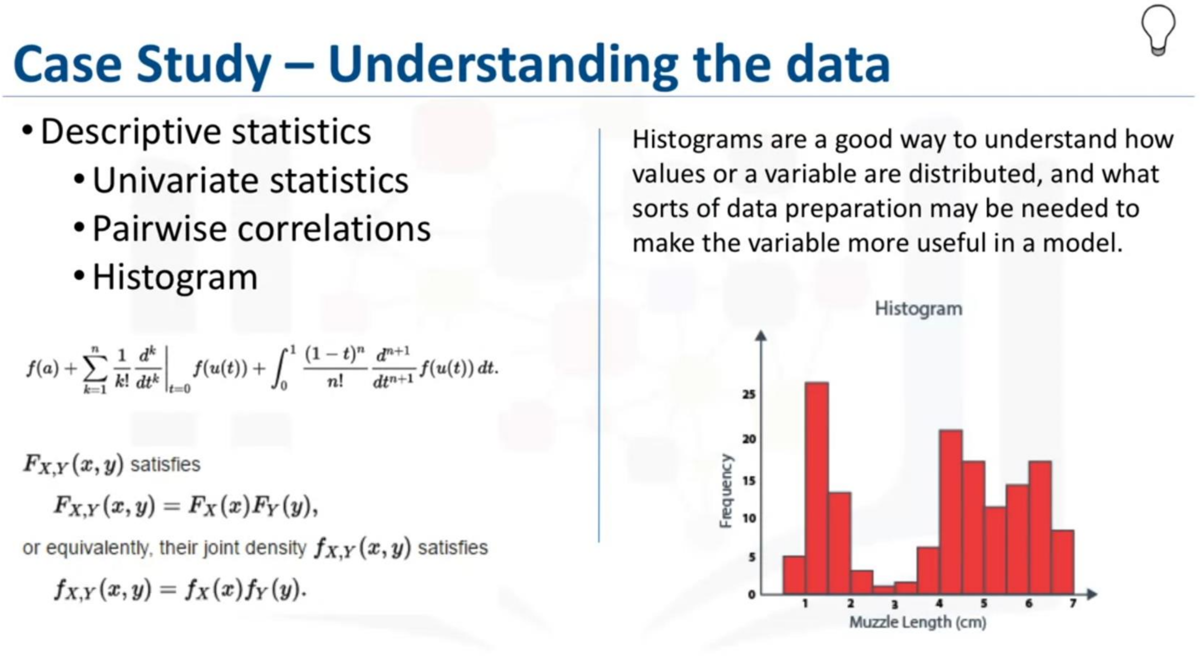
Yong yeol an

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

Is the data that you collected representative of the problem to be solved?

Case study

statistics needed to be run against the data columns that would become variables in the model.
First, these statistics included Hearst, univariates, and statistics on each variable, such as mean, median, minimum, maximum, and standard deviation
Second, pairwise correlations were used, to see how closely certain variables were related, and which ones, if any, were very highly correlated, meaning that they would be essentially redundant,
thus making only one relevant for modeling.
Third, histograms of the variables were examined to understand their distributions.
Histograms are a good way to understand how values or a variable are distributed, and which sorts of data preparation may be needed to make the variable more useful in a model.
For example, for a categorical variable that has too many distinct values to be informative in a model, the histogram would help them decide how to consolidate those values.



Case Study – Understanding the data

- Descriptive statistics
  - Univariate statistics
  - Pairwise correlations
  - Histogram

$$f(a)+\sum_{k=1}^{n}\frac{1}{k!}\frac{d^k}{dt^k}\bigg|_{t=0}f(u(t))+\int_0^1\frac{(1-t)^n}{n!}\frac{d^{n+1}}{dt^{n+1}}f(u(t))\,dt.$$

$F_{X,Y}(x,y)$ satisfies

$$F_{X,Y}(x,y)=F_X(x)F_Y(y),$$

or equivalently, their joint density $f_{X,Y}(x,y)$ satisfies

$$f_{X,Y}(x,y)=f_X(x)f_Y(y).$$

Histograms are a good way to understand how values or a variable are distributed, and what sorts of data preparation may be needed to make the variable more useful in a model.