Data Science Methodologies

Data Science: Data science combines statistics, technology, and doman expertise to extract insights from vast data,

Challenges:

Resolve problems of misunderstanding the business quewstions

Not Knowing how to apply the data to resolve the business problem correctly.

Using a methodology can help address these issues,

What is a methodology?

A methodology is a system of method

A guideline for decision-making during the scientific process.

Data Scence Methodology is a structured approach that guides data scientists in solving complex problems and making data driven decisions.

It involves:

Perform data collection

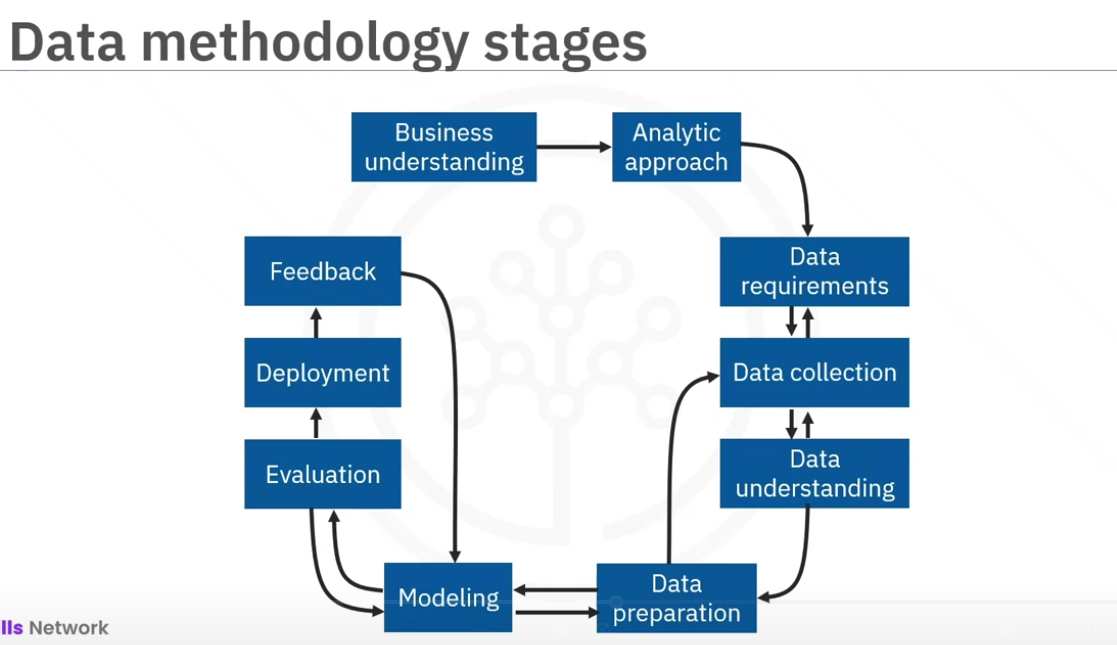
Creation of measurement strategies

Comparisions of data analysis methods.

Using data science methodology provides practical guidance to conduct scientific research efficiently.

Avoid jumping to solutions before the analysis.

Stages in Data Science methodology:



Asking question is the cornorstone in data science. Data science aims to answer the following 10 questions.

Define the issue

Determine your approach

In Business Understanding:1. What is the problem that you’re trying to solve?

In Analytic approach : 2. How can you use data to answer the business question?

Next 4 question are organized around data.

In data requirements: 3. What data do you need to answer the question,

In data collection: 4. Where is the data sourced from and how will you receive the data

In data Understanding: 5. Does the data you collected reppresent the problem to be solved?

In data preperation 6. What other work is required to manipualte and work with the data?

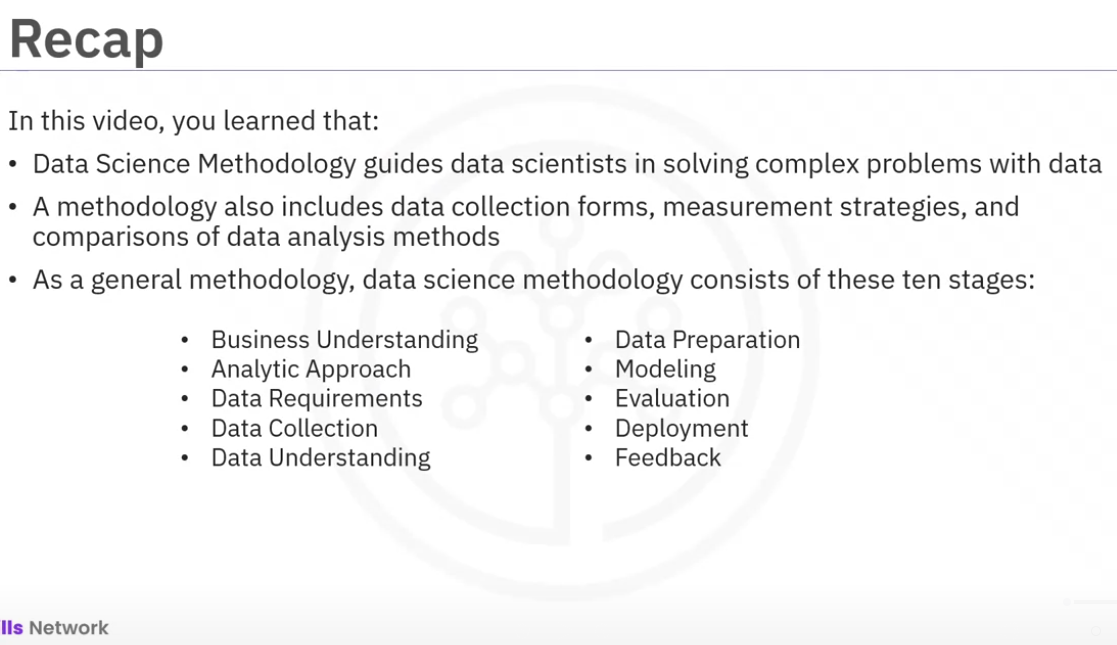
The final 4 questions are designed to validate the final design of the data.

In Modelling: 7. When you apply data visualization do you see answers to address the business problem?

In evaluation: 8. Does the data model answer initial business question ir must you adjest the data>

In deployment: 9. Can you put the data into practice?

In Feedback: 10. Can you get constructive feedback from the data and the stakeholder to answerthe business question?



Data Science Methodologies 101

Question:

Understanding the question changes a lot.

Data?

If we dont understand the question the way we process data is lost.

Methodology Definition:

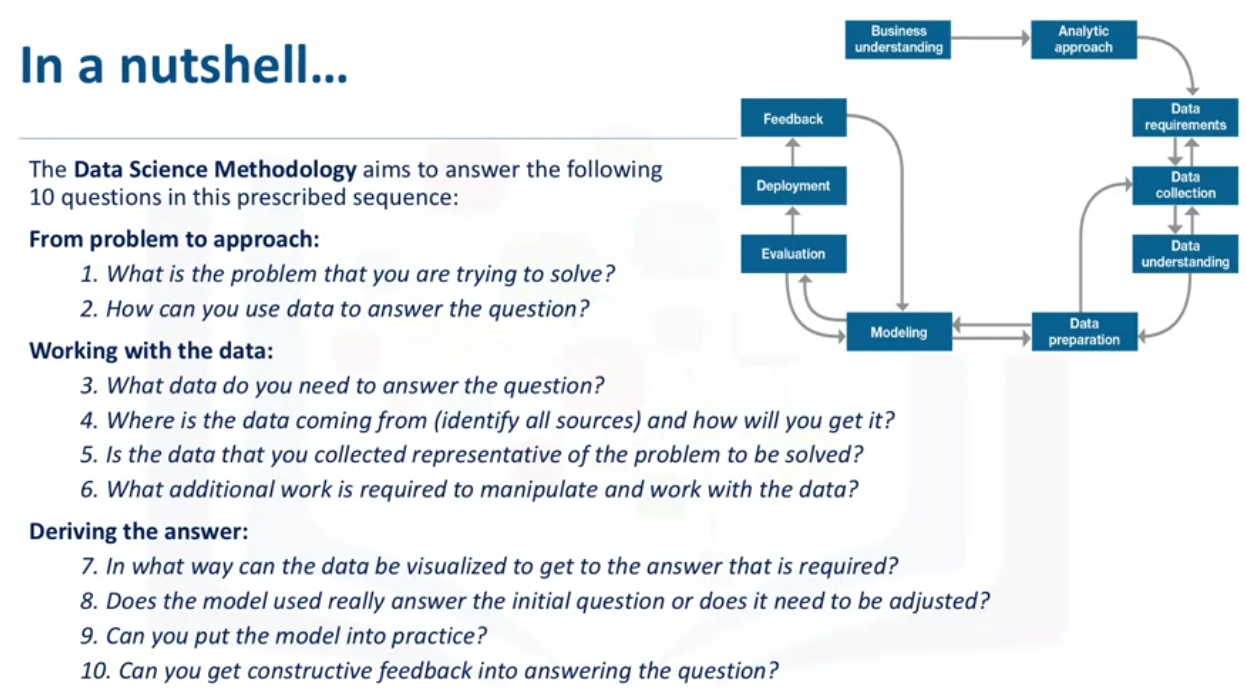
A system of methods used in a particular are od study or activity.

We should not ignore the methodology. Otfen people igore the methodology.

Goal of the course is to share proper methodology to make sure darta is relevant and properly used in the data science steps to address the question at hand.

Definition used was given by John Rolluns Data Scientist at IBM.

Methodology Aims to answer 10 questions



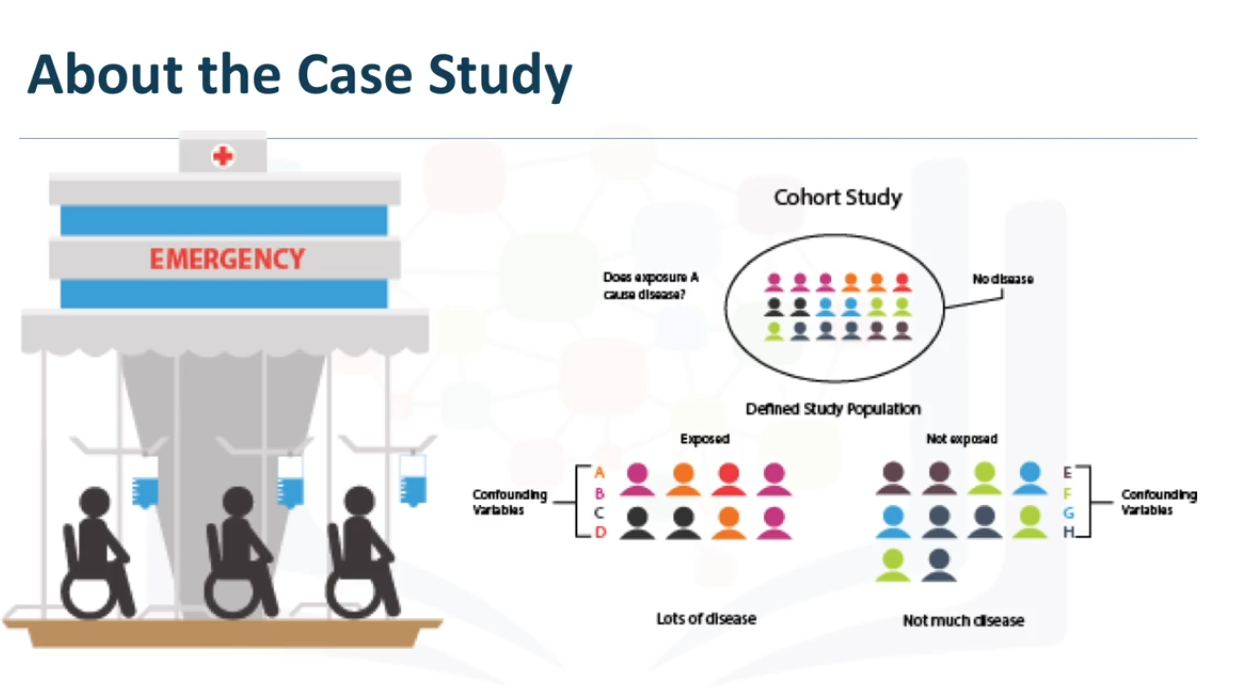
Business Understanding

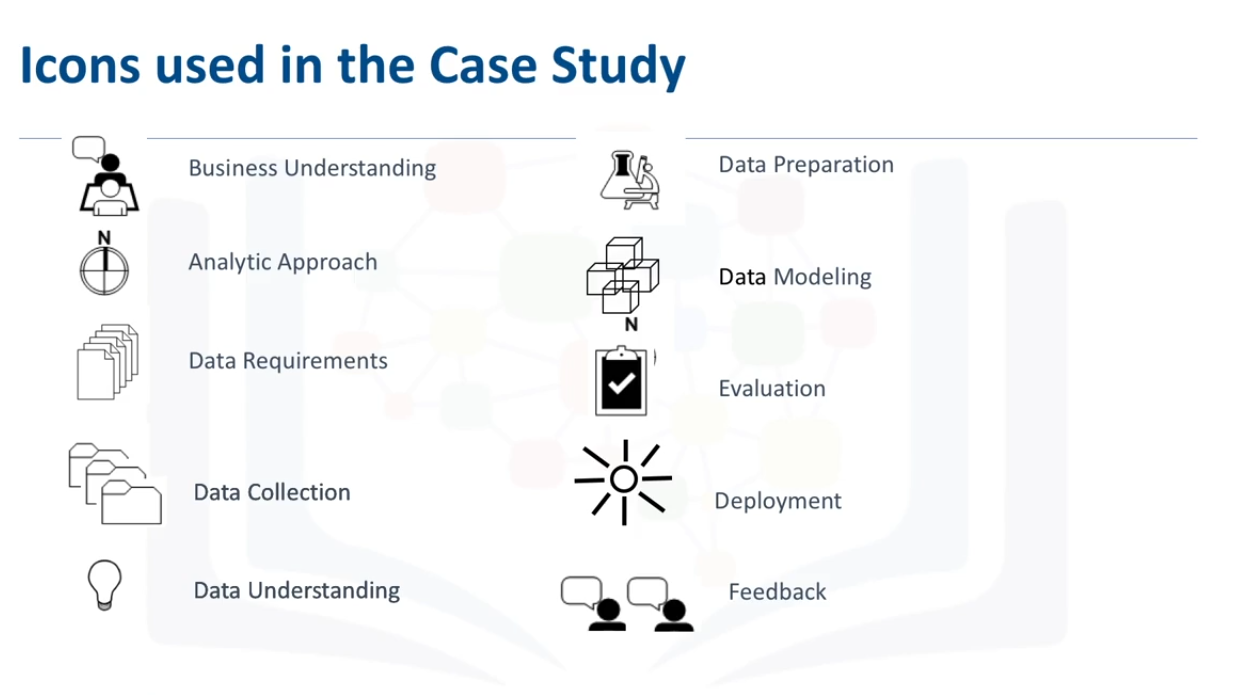
The Importance of Questions:

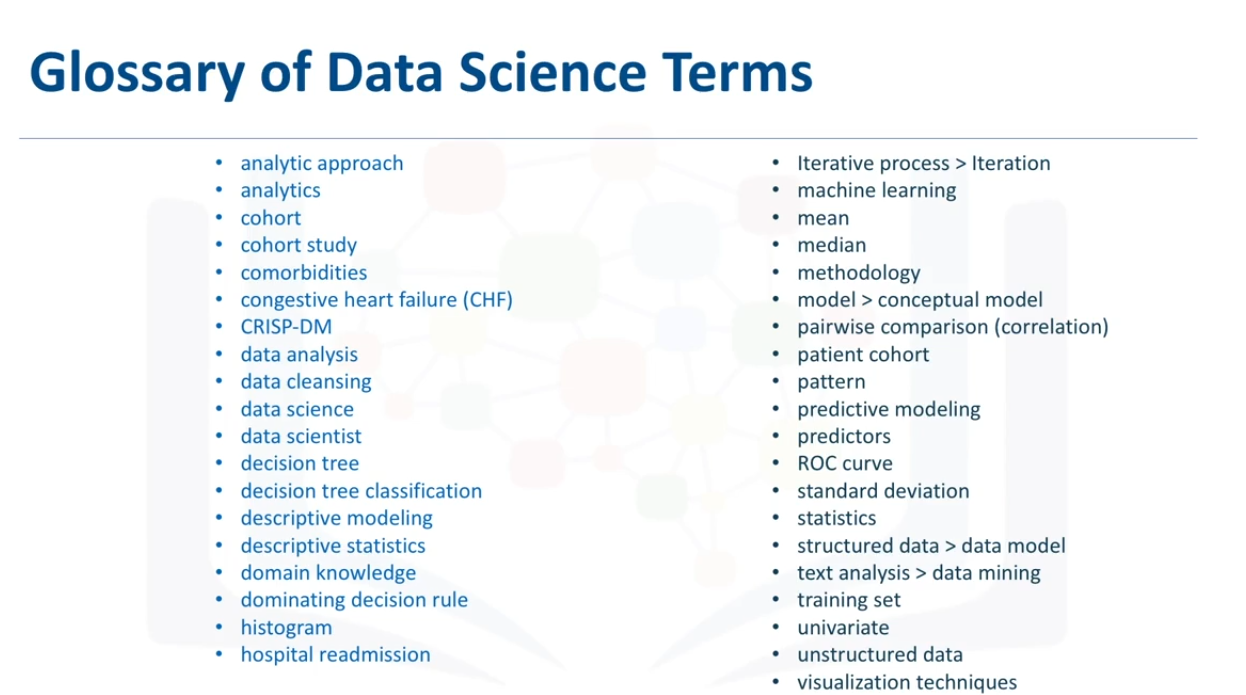
Boss gives a task

has a meeting and decides all aspects of the task

After examining various issues, you realize you need more







<https://author-ide.skills.network/render?token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJtZF9pbnN0cnVjdGlvbnNfdXJsIjoiaHR0cHM6Ly9jZi1jb3Vyc2VzLWRhdGEuczMudXMuY2xvdWQtb2JqZWN0LXN0b3JhZ2UuYXBwZG9tYWluLmNsb3VkL0lCTURldmVsb3BlclNraWxsc05ldHdvcmstRFMwMTAzRU4tU2tpbGxzTmV0d29yay9sYWJzL01vZHVsZSUyMDEvQ1JJU1AtRE0ubWQiLCJ0b29sX3R5cGUiOiJ0aGVpYSIsImFkbWluIjpmYWxzZSwiaWF0IjoxNjgzNzQ5OTY2fQ.dn1nYO7SmJSnAr_5nSy1qJ63HDAgs3MOKX97SuVu4fE>

Business Understanding

Boss Gives a task.

Has a meeting and you two decide on all aspects task.

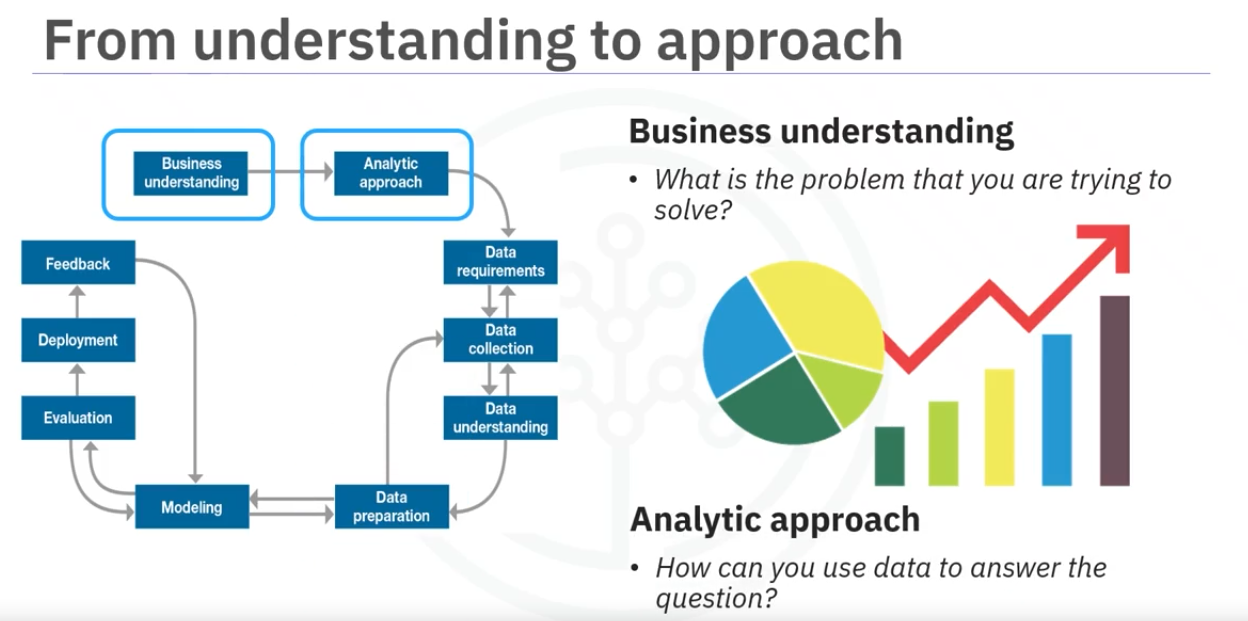
The meeting ends.

Later that day you realize after poking around the project that some more questions are needed to be answered.

But the boss is not available,

What do you do? The deadline is short.

Continue working or clarify the doubts?



From Understanding to Approach (The 10 steps)

Data Science Metodology begins with spending time to seek clarification to attain what can be refered to Business Understanding.

Business Understanding:

What is the problem that you are trying to solve?

This is the first step as it allows to know which data we have to use in order to get quality insights later.

Having a clearly defined question is vital. This directs the analytical approach that we will make use of.

If we confuse ourself with the question, then the answel will not be useful. So defining correct question is very important.

Much effort is put into answering what people think is the question. While the methods use to answer those questions might be sound, they dont help to solve the actual problem.

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

Example:

If a business owner asks how to reduce the cost to perform an activity:

We need to understand is the goal to improve the effieciency of the activity or to improve the business profitability,

Once the goal is determined, find the objectives that are in support of the Goal.

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

Depending on the problem different stake holders will be involved in the discussion.

Analytical Approah

How can you use data to answer the question?

Case study in applying business understanding:

What is the best way to use the limited budget of healthcare to maximize money’s use in provide quality care?

Its a hot topic for an American Insurance Provider as public funding for re admissions was decreasing. The company was at risk of making cost difference and could potentially increase the rates for their customers. Which is not good.

They sat down with healthcare authorities and brough in data scientist at IBM to see how data science could be applied to the task at hand.

Before starting the process, goals and objectives need to be defines.

Define the goals

To provide quality care without increase costs

Define the objectives

To review the process to identify inefficiencies

The team then prioritised patient re admission as an area for review.

They found that roughly 25-35% of patients who complete rehab treatment will be readmitted to rehabilitatio center within one year and roughly 50% will be readmitted within five years

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.

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

Finally, four business requirements were identified for whatever model would be built.

Namely:

1.Predicting readmission outcomes for those patients with Congestive Heart Failure (Y or N ) for each patient

2.Predicting readmission risk.

3.Understanding the combination of events that led to the predicted outcome

4.Applying an easy-to-understand process to new patients, regarding their readmission

Analytics Approach

Which analytic approach to pick?

Depends on question being asked.

Once the problem is defined we select the analytic approach.

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

Types:

Descriptive Analysis:

Current status

Diagnostic (Statistical Analysis)

What Happened?

Why is this happening>

Predictive Analysis (Forcasting)

What if these trends continue?

What will happen next?

Helps to predict Someting.

Prescription

How do we solve it?

if the question is to determine probabilities of an action

Use Predictive model

if the question is to show relationships

Use descriptive model

One which looks at clusters and similar activities based on events and preferences.

if you need counts:

Use statistical analysis.

If yes/ not answer is needed

Use classification

ML:

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.

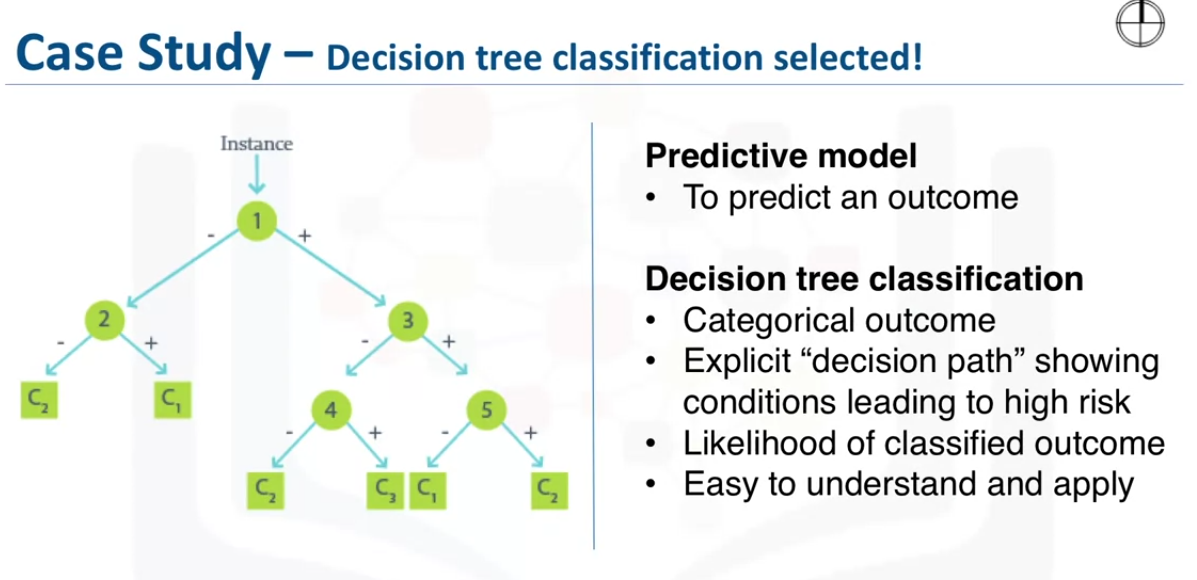
In the case where the question is to learn about human behaviour, then an appropriate response would be to use Clustering Association approaches.

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

Case study:

to identify the combination of conditions leading to each patient's outcome.

A Decision Tree Algorithm is used.



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.

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.

Clinicians can readily see what conditions are causing a patient to be scored as high-risk and multiple models can be built and applied at various points during hospital stay. This gives a moving picture of the patient's risk and how it is evolving with the various treatments being applied.

read analytical approach based on question type document.

Do this exercise: <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMSkillsNetwork-DS0103EN-Coursera/labs/labactivity/M1_LA1_Business_Understanding.html?origin=www.coursera.org>

Do this exercise: <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMSkillsNetwork-DS0103EN-Coursera/labs/labactivity/M1_LA1_Business_Understanding.html?origin=www.coursera.org>

Read Week1\_lesson1\_summary file.

Read Week1\_glossary

Week 1 Lab1

Data Requirements

Think of it like cooking with data.

Each step is important.

Data is Ingredient to solve a problem.

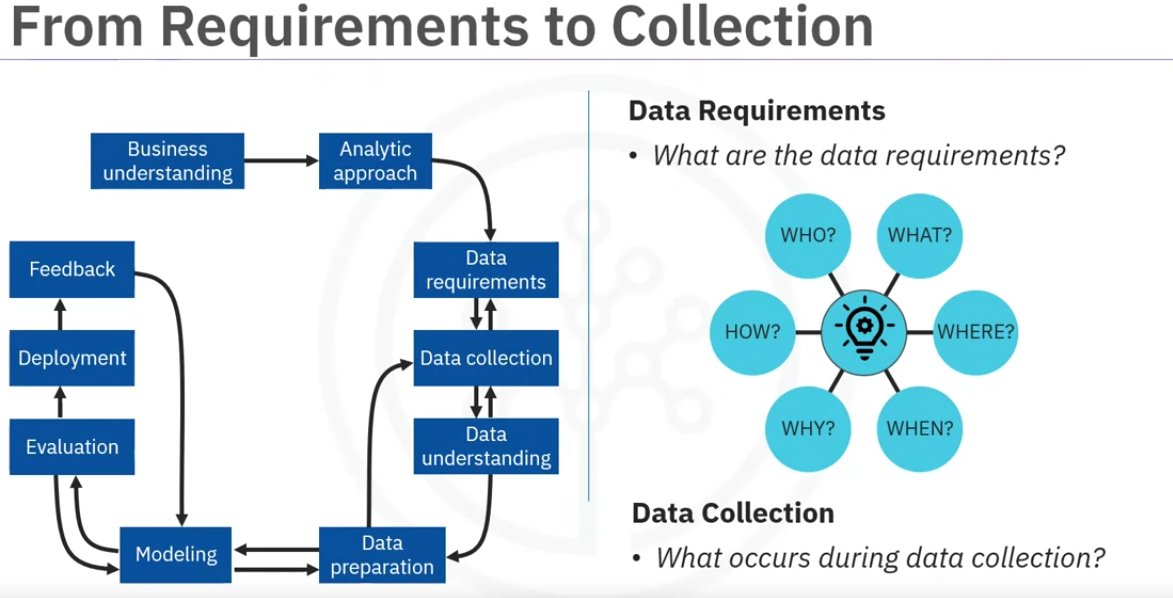
Data scientist neeeds to think about how to

Which ingredeients(data) is required.

how to source or collect them

How to understand of work with them

How to prepare the data to meet the requirements of the problem at hand.



Data Requirements Stage:

What are the data requirements?

Prior to data collection and preperation stages its important to define the data for decision tree clalssification.

Identifying the data content, format and sources.

Case Study:

Collecting patient cohort for Decision tree from the insurance providers memberbase

3 criteria was included that the patient shoud have to get into the cohort.

1. Patient had to be addmitted as in patient un the provider’s service area.
2. Primary diagnosis for CHF in one year
3. Continuos enrollment for atleast 6 months prior to primary CHF admission so complete medical history can be compiled.
4. CHF patients with other Major problems were not included.

Then content, format and representations for decision tree classification are defined.

1 record per patient with columns representing variables(dependent variables and predictors) things

Content covering all aspects of each patient’s clinical history. This could include primary, secondary, and tertiary diagnises, procedures etc.

One patient can have thousands of rows depending on the different attributes. So, transformation is required to make it into 1 row per patient.

The data transformation will be done in the data preperation stage but its good to plan it ahead.

Data Collection

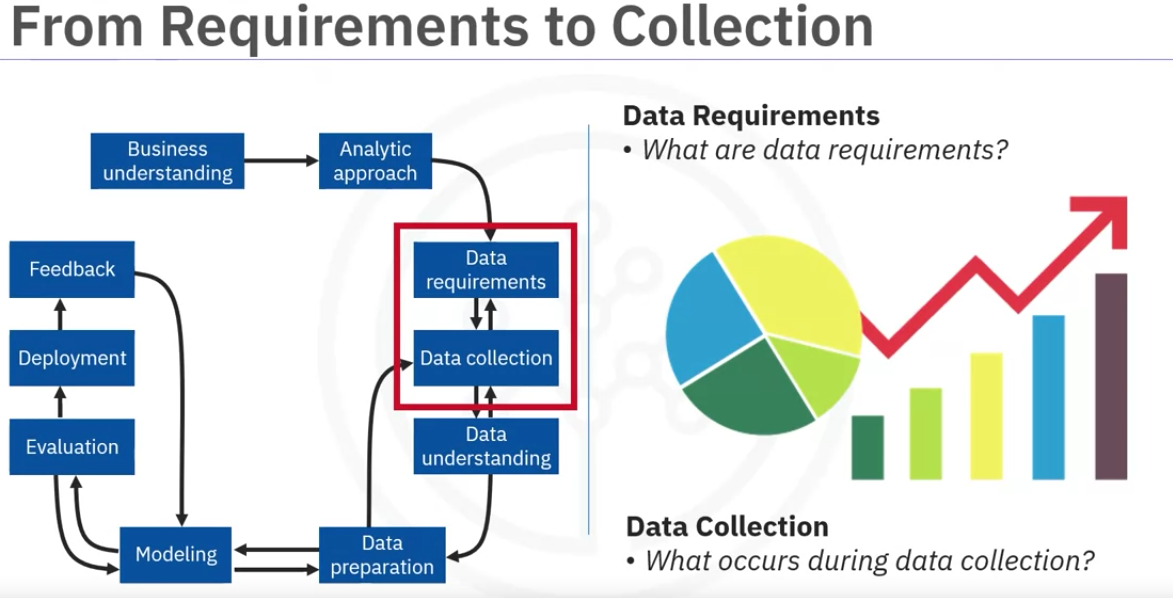
What occurs during data collection?

After planning is done in this stage decisions are made if the data is enough or we need more.

Once data is collected the Data scientist will have a good understanding what data is available to be working with.

Descriptive stats and visualization can be applied to get initial insights about the data.

Gaps in data is identified and removed or filled in with other data.



Case Study

Need to know the source .

Corporate Warehouses(single source of medical and claims, eligibility, provider and member information)

In-patient record system

claim payment system

Disease management program information.

Dat wanted but not available

Pharmaceutical records

Decided to defer(its ok not to have it. Aquire at a later stage.)

In this case intermediate data that is generated was also helpful

DBA and programmer work together to gather data from databases and merge them to eleminate redundant data.

Can also automate some processes.

Week 1 Lab2

Week1\_lesson2\_summary

Week1\_Lesson2\_glossary