Case study:

**The case study included in the course, highlights how the data science methodology can be applied in context.**

There is a limited budget for providing healthcare to the public.

Hospital re-admissions for re-occurring problems can be seen as a sign of failure 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.

Before even starting to collect data, the goals and objectives needed to be defined.

After spending time to determine the goals and objectives, the team prioritized "patient readmissions" as an effective area for review.

With the goals and objectives in mind, it was found that approximately 30% of individuals who finish rehab treatment would be readmitted to a rehab center within one year; and that 50% would 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.

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

Namely:

* Predicting readmission outcomes for those patients with Congestive Heart Failure
* Predicting readmission risk.
* Understanding the combination of events that led to the predicted outcome
* Applying an easy-to-understand process to new patients, regarding their readmission risk.

Business Understanding

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

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

Namely:

Predicting readmission outcomes for those patients with Congestive Heart Failure

Predicting readmission risk.

Understanding the combination of events that led to the predicted outcome

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

risk.

Analytic Approach

Although the analytics approach is the second stage of the data science methodology, it is still independent of the business understanding stage.

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.

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.

This would be one that would look at clusters of similar activities based on events and

preferences.

* 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

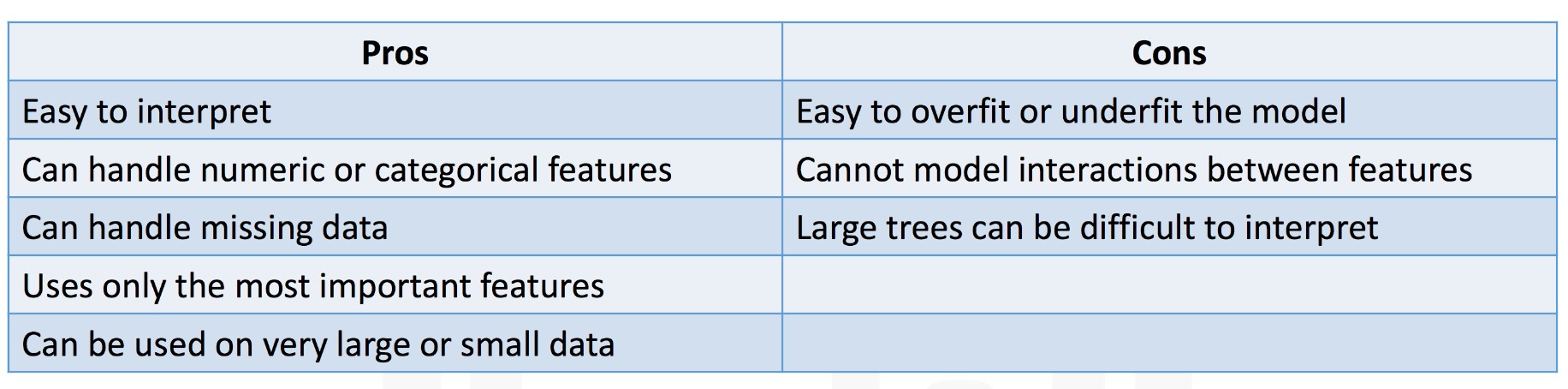
being explicitly programmed.

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

be accessible or identified.

* Decision trees are built using recursive partitioning to classify the data.
* When partitioning the data, decision trees use the most predictive feature (ingredient in this case) to split the data.
* Predictiveness is based on decrease in entropy - gain in information, or *impurity*.

**Here are some characteristics of decision trees:**



**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.

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 Requirements

**Identifying the required data fulfills the data requirements stage of the data science methodology.**

Data Collection

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.

data scientists typically use descriptive statistics and visualization techniques to better understand the data and get acquainted with it. Data scientists, essentially, explore the data to:

* understand its content,
* assess its quality,
* discover any interesting preliminary insights, and,
* determine whether additional data is necessary to fill any gaps in the data.

Data Understanding

In order to understand the data related to congestive heart failure admissions, descriptive 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.

The univariates, statistics, and histograms are also used to assess data quality.

Initially, the meaning of congestive heart failure admission was decided on the basis

of a primary diagnosis of congestive heart failure.

But working through the data understanding stage revealed that the initial definition

was not capturing all of the congestive heart failure admissions that were expected, based

on clinical experience This meant looping back to the data collection stage and adding secondary and tertiary diagnoses, and building a more comprehensive definition of congestive heart failure admission.

#### Data Preparation

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.

Feature engineering is also part of data preparation.

It is the process of using domain knowledge of the data to create features that make the

machine learning algorithms work.

A feature is a characteristic that might help when solving a problem.

Features within the data are important to predictive models and will influence the results

you want to achieve.

Feature engineering is critical when machine learning tools are being applied to analyze

the data.

Specifically, the data preparation stage of the methodology answers the question: What

are the ways in which data is prepared?

Case Study:

In the case study, an important first step in the data preparation stage was to actually define congestive heart failure.

* First, the set of diagnosis-related group codes needed to be identified, as congestive heart failure implies certain kinds of fluid buildup.
* We also needed to consider that congestive heart failure is only one type of heart failure.
* Clinical guidance was needed to get the right codes for congestive heart failure.

The next step involved defining the re-admission criteria for the same condition.

The timing of events needed to be evaluated in order to define whether a particular congestive heart failure admission was an initial event, which is called an index admission, or a congestive heart failure-related re-admission.

Based on clinical expertise, a time period of 30 days was set as the window for readmission relevant for congestive heart failure patients, following the discharge from the initial admission.

Next, the records that were in transactional format were aggregated, meaning that the data included multiple records for each patient.

Transactional records included professional provider facility claims submitted for physician, laboratory, hospital, and clinical services.

Also included were records describing all the diagnoses, procedures, prescriptions, and other information about in-patients and out-patients.

A given patient could easily have hundreds or even thousands of these records, depending on their clinical history.

Then, all the transactional records were aggregated to the patient level, yielding a single record for each patient, as required for the decision-tree classification method that would be used for modeling.

As part of the aggregation process, many new columns were created representing the information in the transactions.

For example, frequency and most recent visits to doctors, clinics and hospitals with diagnoses, procedures, prescriptions, and so forth.

Co-morbidities with congestive heart failure were also considered, such as diabetes, hypertension, and many other diseases and chronic conditions that could impact the risk of re-admission

for congestive heart failure.

Aggregating the transactional data at the patient level, meant merging it with the other patient data, including their demographic information, such as age, gender, type of insurance, and so forth.

The result was the creation of one table containing a single record per patient, with many columns representing the attributes about the patient in his or her clinical history.

These columns would be used as variables in the predictive modeling.

Here is a list of the variables that were ultimately used in building the model.

The dependent variable, or target, was congestive heart failure readmission within 30 days following discharge from a hospitalization for congestive heart failure, with an outcome of either yes

or no.

The cohort was then split into training and testing sets for building and validating the

model, respectively.

#### Modeling - Concepts

This portion of the course is geared toward answering two key questions:

First, what is the purpose of data modeling, and

second, what are some characteristics of this process?

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.

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.

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.

Case study:

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?

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.

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.

This is designed to avoid unnecessary second guessing when the answer is revealed.