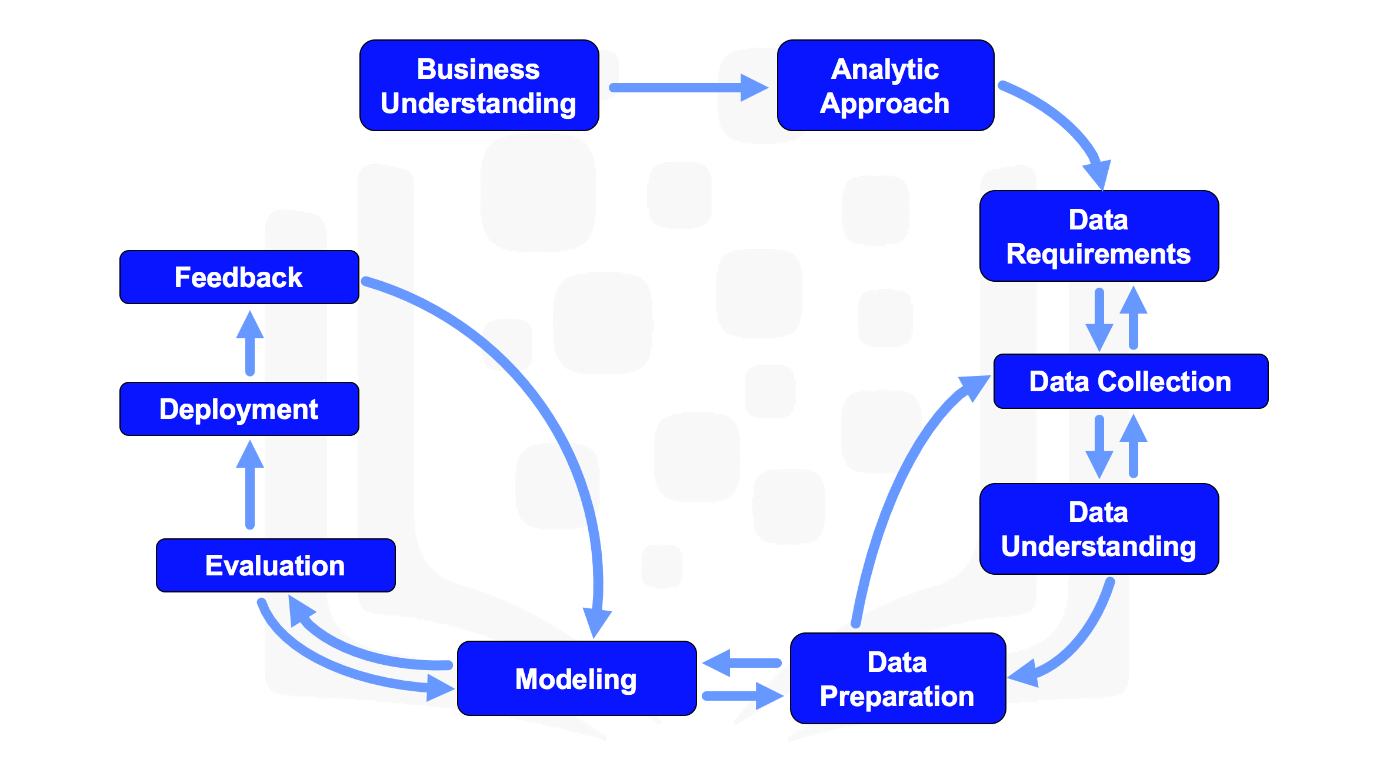
Methodology is a system of methods used in particular area of study or activity. Course is based on the John Rollins experience

**The Data Science Methodology:**

* 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:
  1. What data do you need to answer the question?
  2. Where is the data coming from (identify all sources) and how will you get it?
  3. Is the data that you collected representative of the problem to be solved?
  4. What additional work is required to manipulate and work with the data?
* Deriving the answer:
  1. In what way can the data be visualized to get the answer that is required?
  2. Does the model used really answer the initial question or does it need to be adjusted?
  3. Can you put the model into pracrice?
  4. Can you get constructive feedback into answering the question?



**Buisiness side of question (Q 1.)**

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

Example: what is the best way to allocate limited healthcare to maximize effort?

After defining the goals and objectives researchers found that ~30% of ppl who finished rehab treatment would be readmitted to rehab in 1 year. ~50 of ppl will be readmitted in 5 years.

Researchers found that patients with congestive heart failure were at the top of readmission list.

Bcs of the decision tree researchers also found why it was occuring.

* Predict readmission outcome for each patient
* Predict readmission risk
* Understand the combination of events that led to the predicted outcome
* Apply the approach to new patients to predict their readmission risk

Next IBM mentors proposed an on-site workshop to solve the things and guide analytics team

**Analytic side of question (Q 2.)**

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

Picking analytic approach based on type of the question:

* Descriptive:
  + Current status
* Diagnostic (Statistical Analysis)
  + What happened?
  + Why is it happening?
* Predictive (Forecasting)
  + What is these trend continue?
  + What will happen next?
* Perspective
  + How do we solve it?

If the question is to determine probabilities of the 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

ML approach is related to:

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

Case Study. Selection of predictive model. To illustrate analytics approach lets visualize the decision tree of the congestive heart failure:

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

**Data requirements (Q 3.)**

Cooking parallel: It is needed to identify which ingridients are required, how to collect them and work with them

* Identifying necessary data content
* Formats and sources for initial data

Case study. Defining data requirements for the selected method - decision tree classification.

1. Selecting the cohort based on the congestive heart failure patients from the health insurance providers:
   * Selected patients must be admitted as in-patient within the provider service area, so they would have access to the necessary information
   * Selecting patients with primary diagnosis of CHF in one year
   * Continuous enrollment for at least 6 month prior to primary CHF admisiions
   * Don't taking into account people with other significat medical conditions except CHF (because it would cause higher re admission rates and might skew the results).
2. Content, formats, representation 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(Initialy was in a transactional format and required transformations)

**Data collection (Q 4.)**

Cooking parallel: do I have all ingridients which I need? Are the ingridients good enough?  
Descriptive statistics and visualization can be applied to the dataset to asses the content, quality and initial insights about the data. Gaps in the data will be identified and plans to either fill or make substitutions will have to be made.

After revising the ingridients the decision is made: is it needed to gain more data or there is no need for it.

Case study. Gathering avilable 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 mgmt program information
* Wanted but not available:
  + Pharmaceutical records
  + Decided to defer

**Data understanding (Q 5.)**

Data understanding encompasses all activities related to constructing the dataset.

Data understanding section of the data science methodology answers the question is the collected data representative of the problem to be solved?

Case study. Understanding the data

In order to understand the data related to the congestive heart failure admissions, descriptive statistics needed to be run against the data columns, that would become variables in the model.

1. Descriptive statistics included Hearst, univariates, and statistics on each variable such as mean, median, minimum, maximum and std deviation.
2. 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.
3. Histograms of the variables were examined to understand their distributions. They are used because it would give an idea about what sorts of data preparation may be needed to make the variable more useful in a model. 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 quality.

Univariates, statistics and histograms are also used to assess data quality. 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. It is important to check missing data format and type. It can be Null, 0 or, for example 999.

Case study. Iterative process.  
Initially, the meaning of congestive heart failure admission was decided on the basis of a primary diagnosis of congestive heart failure. While working on the data understanding stage it revealed, that the initial definition was not capturing all of the congestive heart failure admissions, that were expected, based on clinical experience. It sended the researchers back to the data collection stage and adding secondary and tertiary diagnoses in order to build more comprehesive definition of congestive heart failure admission. (Can be used as exmple of the interactive processes in the methodology). The more one works with the problem and the data, the more one learns and therefore more refinement can be done within the model, ultimately leading to a better solution.

**Data Preparation (Q 6.)**

Data preparation analogy - cleaning vegetables to remove unwanted elements. Data preparation takes around 70-90% of all time  
Automation and preparation processes in the data base can reduce this time to as little as 50 percent.

Data preparation - also can be associated with cutting onions  
Data preparation is needed to deal with following problems:

* Removing duplicates
* Dealing with missing data
* Dealing with invalid values
* Fixes formatting

Feature engineering is the process of using domain knowledge of the data to create features, that make the machine learning algorithms work.

Feature engineering is critical when ML tools are being applied to analyze the data.

Feature - characteristic, that might help when solving a problem

*Case Study*

Defining congestive heart failure. Here is the reason why it is not as simple as it sounds like:

* Identifying set of diagnosis-related group codes because congestive heart failure implies certain kinds of fluid buildup
* Taking into consideration that congestive heart failure is just a certain heart failure type

Records in transactional format were aggregated, meaning that the data included multiple records for each patient Records were rolled up to 1 record per patient with new columns, which represented the transaction(outpatient visits, frequency, recency, length of stay ...)

Then, it was checked which factors affected readmission more, and which factors affected readmission less

Then a data set was merged into one table with taking into account demographic information(age, gender,..)

The final dataset looked like rows with different patients and columns with different attributes of patient clinical history.

* Target: CHF readmission with 30 days (yes/no) following discharge from CHF hospitalization
* Measures: Gender, age, primary DRG, length of stay, prior admissions, line of buisiness, CHF diagnosis importance(primary, secondary, tretiary)
* Diagnosis flags(Y/N) CHF< Diabets, Atrial fibrilation, pneumonia, renal failure, hypertension

Chohort: 2343 patients 70/30 training test set splits

**Modeling (Q 7.)**

Analogy: sample the food and determine if it is bang on or in need of more seasoning  
Data modelling focuses on devloping models that are either descriptive or predictive.

The key question is: have I made enough to eat

*Case study*

Overall accuracy (% correct Y & N), Sensitivity (Y accuracy), Specificity (N accuracy)

1. st decision tree classification model:
   * Low accuracy on "Yes" outcome (Bad Yes accuracy, best overall accuracy). Overall accuracy = 85%, sensitivity = 45%, specificity = 97%
2. nd decision tree classification model:
   * Relative cost Y:N = 9:1 (Good Yes accuracy but bad overall accuracy because of large number of false-positives). Overall accuracy = 49%, sensitivity = 97%, specificity = 35%
3. rd decision tree classification model:
   * Relative cost Y:N = 4:1 (Good balance between Y accuracy and N accuracy). Overall accuracy = 81%, sensitivity = 68%, specificity = 85%

**Evaluation (Q 8.)**

When and how to adjust the model? Also, ensures that the data are properly handled and interpreted

1st phase:  
Diagnostic measures:

* predictive model:
  + In case of a deicision tree decision tree itself can be used to evaluate if the answer is aligned to the initial design. It can be used to see where there are areas that require adjustments.
* descriptive model:
  + In case of descriptive model then a testing set with known outcomes can be applied and the model can be refined as needed.

2nd phase:

* statistical significance testing:
  + Needed to ensure that data is properly handled and interpreted within the model.

*Case study*

MIsclassification cost tuning:

* Tune the relative misclassification costs (Y:N)
* Balance TP rate and FP rate for the best model

Applying receiver operating characteristic curve or ROC curve:

* Classification model performance
* TPr vs FPr
* Optimal model at maximum separation   
  Curve quantifies how werll a binary classification model performs declassifying the yes and no based on criteria (in that case it is misclassification cost)

**Deployment (Q 9.)**

Gettin the stakeholders familiar with the tool produced Firstly the initial model could be rolled out to a limited group of users or in a test envronment.

*case study*

Assimilate knowledge for business

* Practical understanding of the meaning of model results
* Implications of model results for designing intevention actions

Congestive heart failure application requirements:

* Automated, near-real-time risk assessments of CHF inpatients
* Easy to use
* Automated data preparation and scoring
* Up-to-date risk assessment to help clinicians target high-risk patients

Additional requirements:

* Training for clinical staff
* Tracking/ monitoring processes

Examples of deployed solutions:

* Hospitalization risk for juvenile diabetes patients depending on state
* Risk summary report by decision tree model note
* Individual patient risk report

**Feedback (Q 10.)**

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.

Feedback process might be rooted because more you know - more you want to know. That's the way John Rollins sees it (and hopefully you too).

*Case study*

Assessing model performance:

* Measure results of applying the risk model to the CHF patient population
* Track patients who received intervention
* Actual readmission outcomes
* Measure effectovemess off intervention
* Compare readmission rates before & after model implementation

Model accuracy consists of:

* Interpretation
* Data quality
* Time

Refine model:

* Initial review after the first year of implementation
* Based on feedback data and knowledge gained
* Participation in intervention program
* Possibly incorporate detailed pharmaceutical data originally deferred
* Other possible refinements as yet unknown

Review and refine intervention actions: Redeploy:

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

The data science methodology is highly iterative, ensuring the refinement at each stage in the game.

**Summary**

Data scientist is thinking about:

* Forming a concrete buisiness or research problem
* Collecting and analyzing data
* Building a model
* Understanding a feedback after model deployment

It is also important to:

* Understand the question
* Pick most effective analytic approach

Working with the data includes:

* Determining the data requirements
* Collecting the appropriate data
* Understanding the data
* Preparing the data for modeling

Evaluating and deploying the model:

* Getting feedback on it
* Using the feedback constructively so as to improve model

**Quizes**

**W1\_From Problem to Approach**

1. Question 1 Select the correct statement:
   * The data science methodology described in this course is outlined by John Rollins from IBM.
   * None of the above statements are correct.
   * The data science methodology described in this course is limited to IBM.
   * The data science methodology described in this course is only used by certified data scientists.
2. Question 2 The first stage of the data science methodology is Data Understanding.
   * False (Buisiness understanding)
3. Question 3 Business Understanding is an important stage in the data science methodology. Why?
   * Because it involves domain expertise. (True)
   * Because it clearly defines the problem and the needs from a business - perspective. (True)
   * Because it ensures that the work generates the intended solution. (True)
   * Because it shapes the rest of the methodological steps. (True)
   * None of the above.
4. Question 4 Which of the following statements about the analytic approach are - correct?
   * If the question defined in the business understanding deals with exploring relationships between different factors, then a classification approach would be the right analytic method.
   * If the question defined in the business understanding stage can be answered by determining probabilities of an action, then a predictive model would be the right analytic approach. (True)
   * If the question defined in the business understanding deals with exploring relationships between different factors, then a predictive model would be the right analytic approach.
   * If the question defined in the business understanding deals with exploring relationships between different factors, then a descriptive approach, where clusters of similar activities based on events and preferences are examined, would be the right analytic method. (True)
5. Question 5 For the case study, a decision tree classification model was used to identify the combination of conditions leading to each patient's outcome.
   * True

**W1\_From Requirements to Collection**

1. Question 1 Which of the following analogies is used in the videos to explain the Data Requirements and Data Collection stages of the data science methodology?
   * You can think of the Data Requirements and Data Collection stages as building an outpatient clinic for patients with congestive heart failure, where the medical condition is the data and the patients are the ingredients.
   * You can think of the Data Requirements and Data Collection stages as a cooking task, where the problem at hand is a recipe, and the data to answer the question is the ingredients. (True)
2. Question 2 The Data Requirements stage of the data science methodology involves identifying the necessary data content, formats and sources for initial data collection.
   * True
3. Question 3 Which of the following statements are correct?
   * Data scientists determine how to collect the data.(True)
   * Data scientists identify the data that is required for data modeling.(True)
   * Data scientists determine how to prepare the data.(True)
   * None of the above.
4. Question 4 In the Data Collection stage, the data requirements are revised and decisions are made as to whether or not more data is needed.
   * True
5. Question 5 In the Data Collection stage, 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
   * True

**W2\_From Understanding to Preparation**

1. Question 1 Select the correct statement about the Data Understanding stage.
   * The Data Understanding stage encompasses sorting the data.
   * The Data Understanding stage encompasses all activities related to constructing the dataset. (True)
   * The Data Understanding stage encompasses removing redundant data.
   * All of the above statements about the Data Understanding stage are correct.
2. Question 2 In the case study, working through the Data Understanding stage, it was revealed that the initial definition was not capturing all of the congestive heart failure admissions that were expected, based on clinical experience.
   * True
3. Question 3 Select the correct statement about the Data Preparation stage.
   * The Data Preparation stage involves addressing missing values.
   * The Data Preparation stage involves correcting invalid values and addressing outliers.
   * The Data Preparation stage involves removing duplicate data.
   * The Data Preparation stage involves properly formatting the data.
   * All of the above statements are correct. (True)
4. Question 4 Select the correct statement about what data scientists and database administrators (DBAs) do during the Data Preparation stage.
   * During the Data Preparation stage, data scientists and DBAs define the variables to be used in the model.
   * During the Data Preparation stage, data scientists and DBAs determine the timing of events.
   * During the Data Preparation stage, data scientists and DBAs aggregate the data and merge them from different sources.
   * During the Data Preparation stage, data scientists and DBAs identify missing data.
   * All of the above statements are correct. (True)
5. Question 5 Select the correct statement about the Data Preparation stage of the data science methodology.
   * Data Preparation is typically the least time-consuming methodological step.
   * Data Preparation involves dealing with missing improperly coded data and can include using text analysis to structure unstructured or semi-structured text data. (True)
   * Data Preparation cannot be accelerated through automation.
   * None of the above statements are correct.

**W2\_From Modeling to Evaluation**

1. Question 1 Select the correct statement.
   * A training set is used for statistical analysis.
   * A training set is used for predictive modeling.(True)
   * A training set is used for descriptive modeling.
   * A training set is used for data visualization.
2. Question 2 A statistician calls a false-negative, a type I error, and a false-positive, a type II error.
   * False
3. Question 3 Which statement best describes the Modeling Stage of the data science methodology?
   * The Modeling stage is followed by the Analytic Approach stage.
   * Modeling always uses training and test sets.
   * Modeling is always based on predictive models.
   * Modeling may require testing multiple algorithms and parameters. (True)
4. Question 4 Model Evaluation includes ensuring that the data are properly handled and interpreted.
   * True
5. Question 5 Select the correct statements about the ROC curve? -
   * ROC stands for Receiver Operating Characteristic curve, which was originally developed to detect enemy aircrafts on radar. (True)
   * The ROC curve is a useful diagnostic tool for determining the optimal classification model. (True)
   * By plotting the true-positive rate against the false-positive rate for different values of the relative misclassification cost, the ROC curve can be used to select the optimal model. (True)
   * The ROC curve was originally developed to optimize healthcare and detect congestive heart failure readmission rate.

**W3\_From Deployment to Feedback**

1. Question 1 The final stages of the data science methodology are an iterative cycle between Modelling, Evaluation, Deployment, and Feedback.
   * True
2. Question 2 Feedback is not required once the model is deployed because the Model Evaluation stage would have assessed the model and made sure that it performed well.
   * False
3. Question 3 Deploying a model into production represents the end of the iterative process that includes Feedback, Model Refinement, and Redeployment.
   * False
4. Question 4 Select the correct sentence about the data science methodology explained in the course
   * The data science methodology depends on a specific set of technologies or tools.
   * The data science methodology always starts with data collection.
   * The data science methodology is not an iterative process – one does not go back and forth between methodological steps.
   * The data science methodology provides the data scientist with a framework on how to proceed to obtain answers. (True)
5. Question 5 A data scientist determines that building a recommender system is the solution for a particular business problem at hand. What stage of the data science methodology does this represent?
   * Modeling.
   * Model Evaluation.
   * Deployment.
   * Analytic Approach. (True)
6. Question 6 A data scientist, John, was asked to help reduce readmission rates at a local hospital. After some time, John provided a model that predicted which patients were more likely to be readmitted to the hospital and declared that his work was done. Which of the following best describes this scenario?
   * Even though John only submitted one solution, it might be a good one. However, John needed feedback on his model from the hospital to confirm that his model was able to address the problem appropriately and sufficiently. (True)
   * John only provided one model as a solution and he should have provided multiple models.
   * John’s mistake is that he lied in the Analytic Approach step of the data science methodology.
   * John still needed to collect more data.
7. Question 7 Data scientists typically use descriptive statistics and data visualization techniques for exploratory analysis of data and to get acquainted with it.
   * True
8. Question 8 Which of the following represent the two important characteristics of the data science methodology?
   * It immediately ends when the model is deployed because no feedback is required.
   * It is a highly iterative process and immediately ends when the model is deployed.
   * It has no endpoint because data collection occurs before identifying the data requirements.
   * It is a highly iterative process and it never ends. (True)
9. Question 9 Data scientists may use either a “top-down” approach or a “bottom-up” approach to data science. These two approaches refer to:
   * "Top-down” approach – models are fit before the data is explored. “Bottom-up” approach – data is explored, and then a model is fit.
   * “Top-down” approach – using massively parallel, warehouses with huge data volumes as the data source. “Bottom-up” approach – using a sample of small data before using large data.
   * “Top-down” approach – the data, when sorted, is modeled from the “top” of the data towards the “bottom”. “Bottom-up” approach – the data is modeled from the “bottom” of the data to the “top”.
   * “Top-down” approach – first defining a business problem then analyzing the data to find a solution. “Bottom-up” approach – starting with the data, and then coming up with a business problem based on the data. (True)
10. Question 10 Data scientists should maintain continuous communication with stakeholders throughout a project so that business stakeholders can ensure the work remains on track to generate the intended solution.
    * True