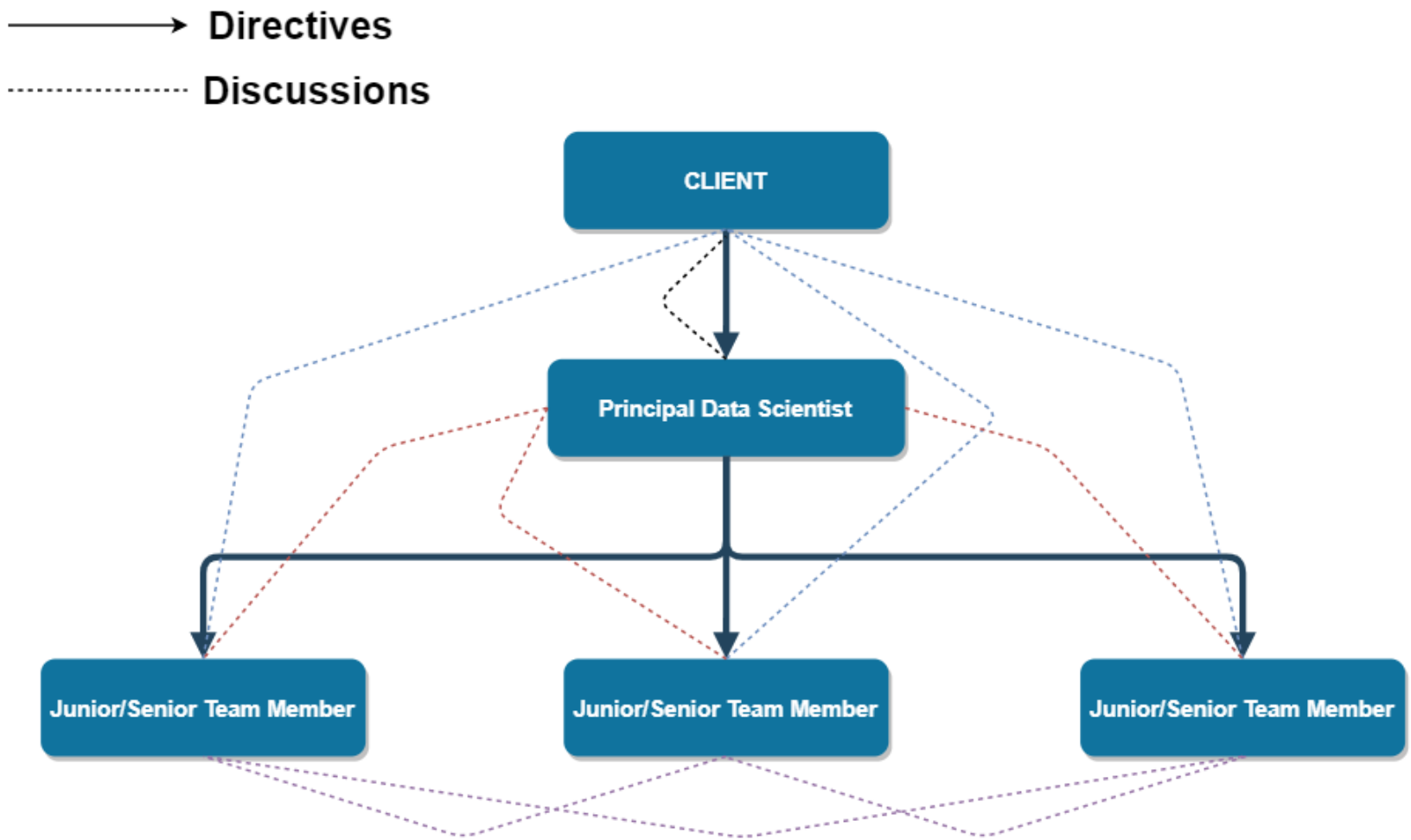
Preparing for a Project

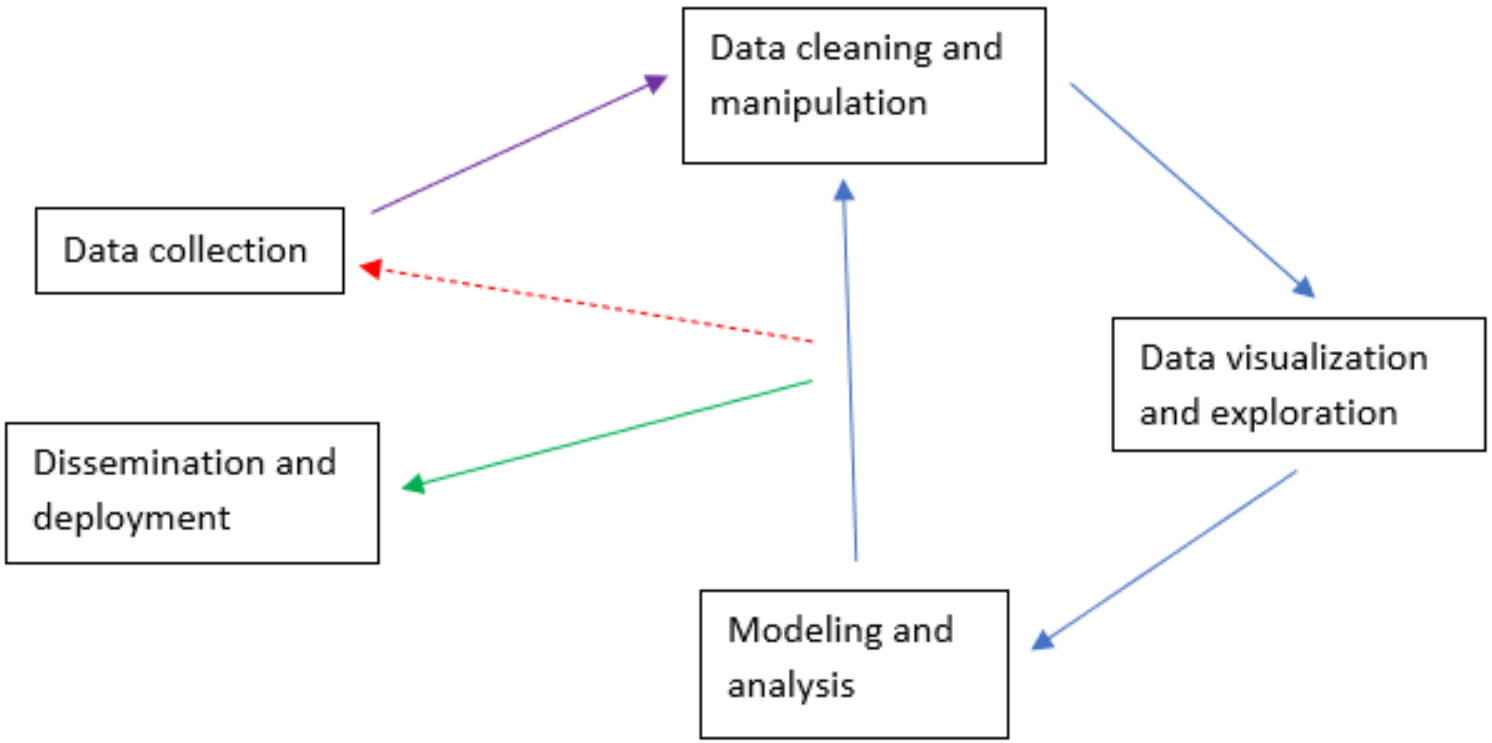
Ryan Miller

Organizational Structure



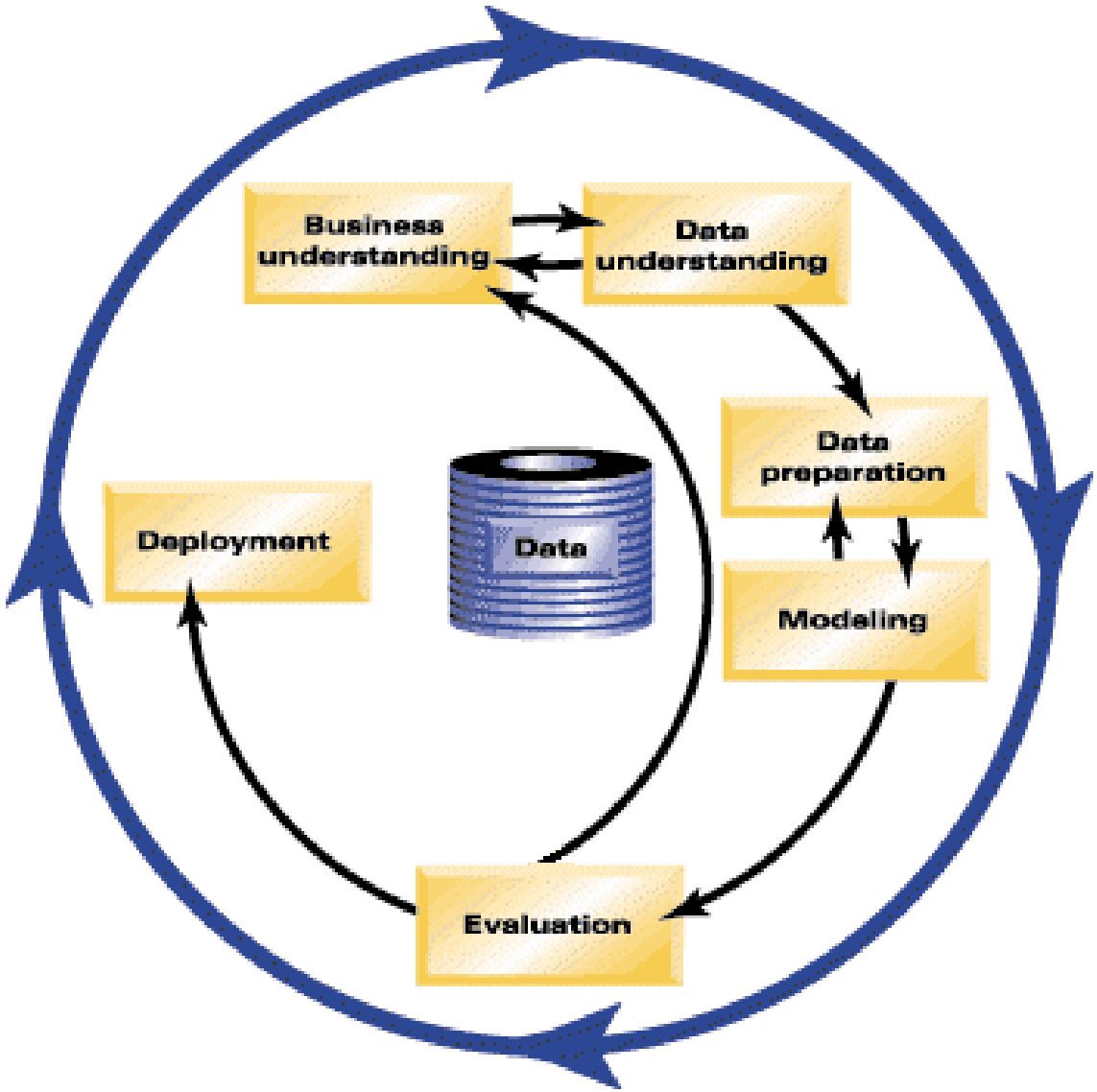
Project Life Cycle

Previously, we considered the following cycle:



This framework is based upon a model known as the Cross-Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM


Business understanding
Data understanding
Data preparation
Modeling
Evaluation
Deployment
Data

Group #1 - Business Understanding

Components:

1) Gather necessary background information
2) Document specific specific objectives
3) Determine success criteria for the project

These should be coordinated with the client, with special focus given to establishing common definitions and terminology.

Objective vs. Subjective Success Criteria

You may opt for a mixture of objective and subjective success criteria:

- Objective = "Increase the time visitors spend on the landing page by 10%"
- Subjective = "Identify customer clusters for targeted marketing"

What advantages/disadvantages are there to each?

Group #2 - Data understanding

Components:

1) Understanding and acquiring the data
2) Describing the data
3) Exploring the data
4) Verifying data quality

These should be carried out at the team level (and cross-referenced with the principal and client if necessary)

Group #3 - Data Preparation

Components:

- Merging/joining (ie: `left_join`)
- Selecting relevant subsets (ie: `filter`)
- Aggregating records (ie: `group_by` and `summarize`)
- Deriving new attributes (ie: `mutate`)
- Handling missing data (ie: `complete.cases` or `knnImput`/`rfImpute`)

These should be carried out at the team level and cross-referenced by the principal (they are seldom relevant to the client at this point)

Group #4 - Modeling and Analysis

Components:

- Selecting a model
- Evaluating the "goodness" of a model
- Building the model
- Note: you may replace "model" with "product" in some applications

The client seldom has expertise in modeling/analysis (otherwise they'd likely do this stage of the project themselves)

Group #5 - Evaluation

Components:

1) Consider your model/product in regards to the business success criteria you came up with in Phase 1
2) Formalize your findings such that they can be easily understood by the client

These tasks should be undertaken in coordination with the client and principal

Review #1

For the following scenario determine:

A) Where in the data science life cycle this event would likely occur (which "box" and which "group" of tasks).

B) How you'd address the situation (what you'd do and where you'd go next).
A project is using medical records to build a model to predict A1c levels using more readily available measures such as blood pressure, age, weight, and waist circumference. Using the `is.na` function in `R`, it is discovered that 88% of the available records do not have an A1c measurement.

Review #1 - Possible Answers

- This scenario occurred during **data understanding** tasks performed during **data visualization and exploration**.
- It is possible that the 12% of medical records with A1c provides a reasonable data set. In this case you'd progress to data preparation or modeling and analysis (filtering out the missing data and preparing the other variables)
- It is also possible that relying on only 12% of the available records results in too small of a sample, or introduces bias into the analysis. In this case you'd return to business understand and data collection (reviewing the original goals and making adjustments, or trying to obtain additional data)

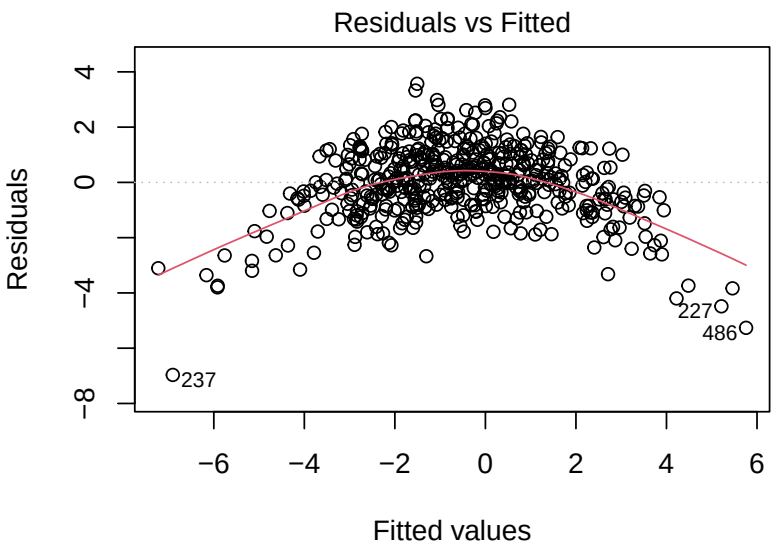
Review #2

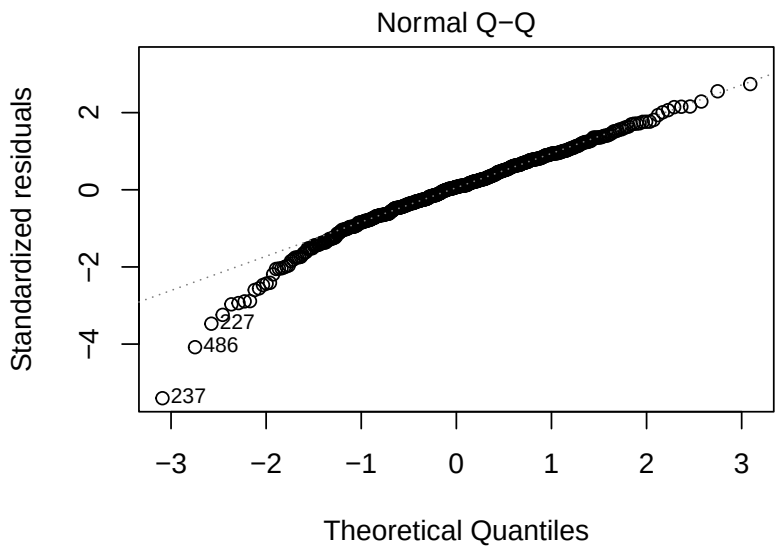
For the following scenario determine:

A) Where in the data science life cycle this event would likely occur (which "box" and which "group" of tasks).

B) How you'd address the situation (what you'd do and where you'd go next).

In the aforementioned project you fit a linear regression model containing several variables, you receive the following model diagnostics from your software





Review #2 - Possible Answers

- This scenario occurred during the **Modeling** tasks performed during **modeling and analysis**.
- Linear regression doesn't appear to be an appropriate model based upon these diagnostics. The true relationship might be quadratic or "U-shaped" (evidenced by the large negative residuals for high/low fitted values). The QQ-plot also calls into question whether the residuals are normally distributed (not a disaster for model fitting, but a problem for statistical inference)
  - For these reasons the logical next step is return to data cleaning and manipulation and carry out additional data preparation tasks (variable transformations).

Review #3

For the following scenario determine:

A) Where in the data science life cycle this event would likely occur (which "box" and which "group" of tasks).

B) How you'd address the situation (what you'd do and where you'd go next).
After revising the linear regression model in the previous example, a final model is chosen and is applied a "test set" of 100 new records that occurred after the original dataset was finalized. The model predicts A1c within 10% of the actual value for 86% of these new records.

Review #3 - Possible Answers

- This scenario occurred in the **Evaluation phase**
- Where to go next *depends upon the project's business goals*. If predicting A1c within 10% of the actual value for 86% of cases satisfies the previously established goals, the model is suitable for the Deployment phase. Otherwise the project might need to return to square one.

Next Steps

Read each project description. Next, I'd like you to discuss the following for each project:

1. The business understanding steps necessary to successfully work on the project (ie: what do you need to discuss with the client, what do you need to learn, what do you currently view "success" as for the project)
2. The data understanding steps necessary to start the project (ie: what do you anticipate the data looking like, what will you need to explore, what will you need to clarify with your client)

I'd like each group to submit a document summarizing their thoughts (for both of the above bullets for each project). You may record your thoughts as bullet points, or in paragraphs.