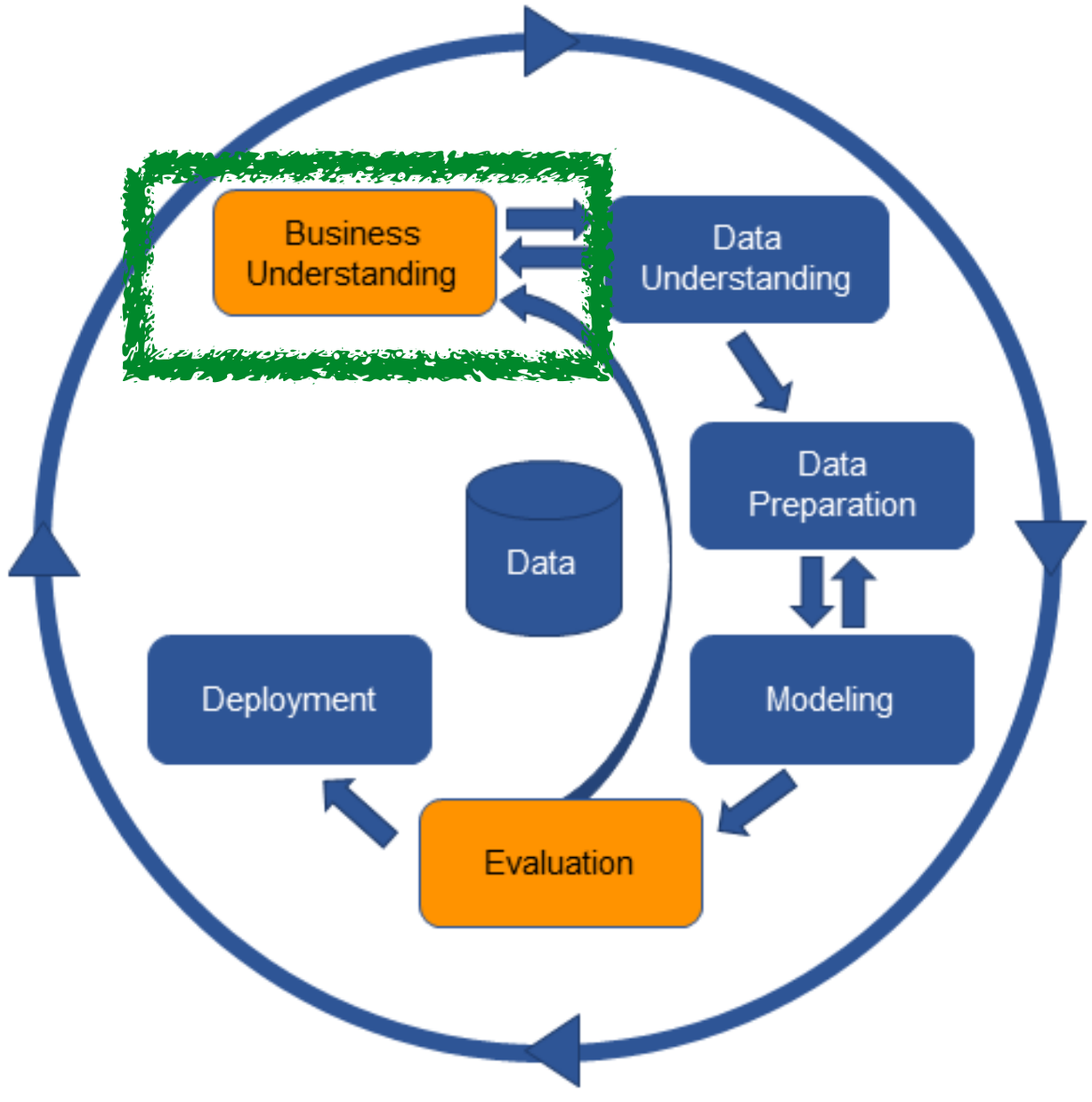
CRISP-DM Methodology

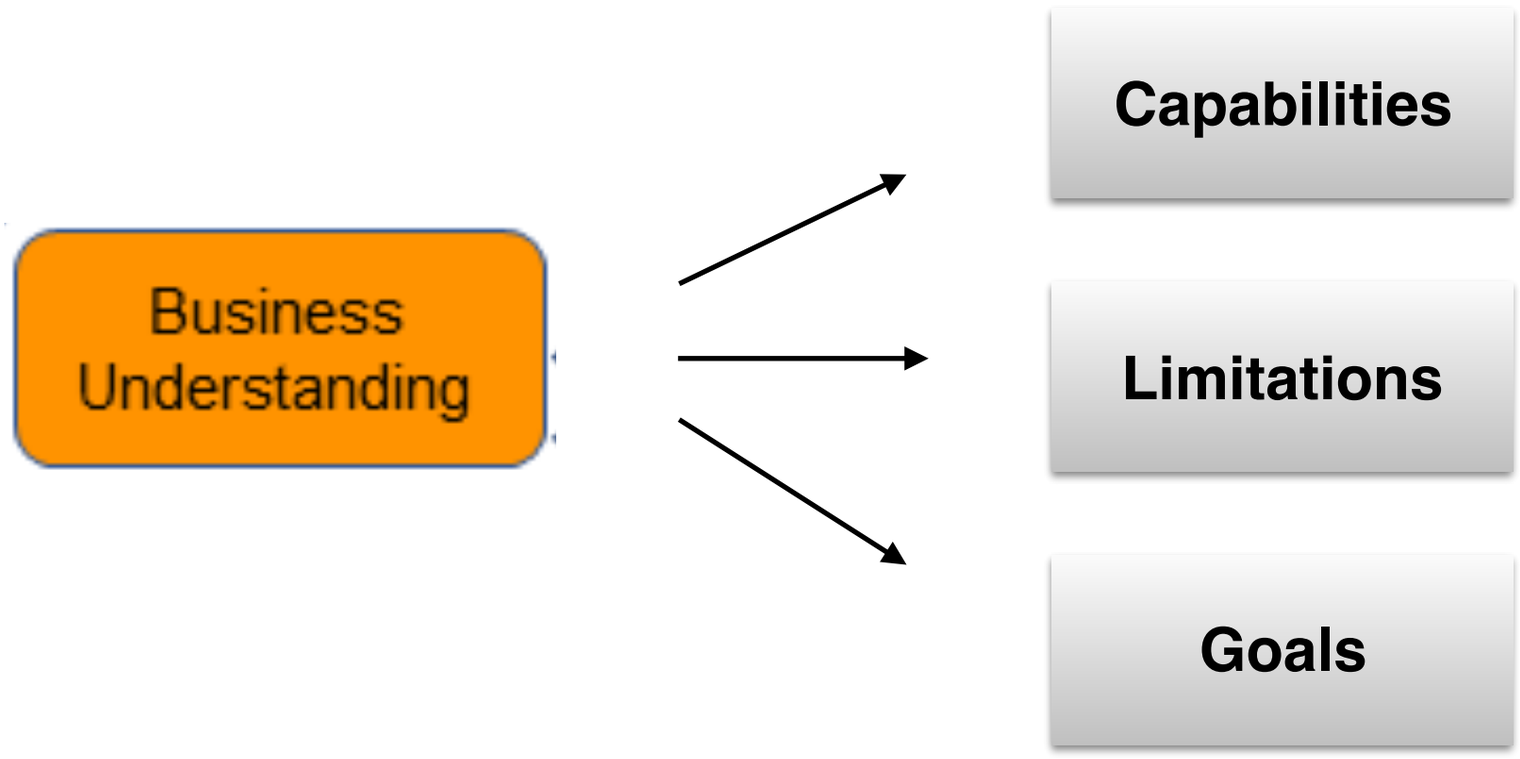
An overview of the data mining-analysis-science lifecycle

CRISP-DM

Cross-Industry Standard Process for Data Mining

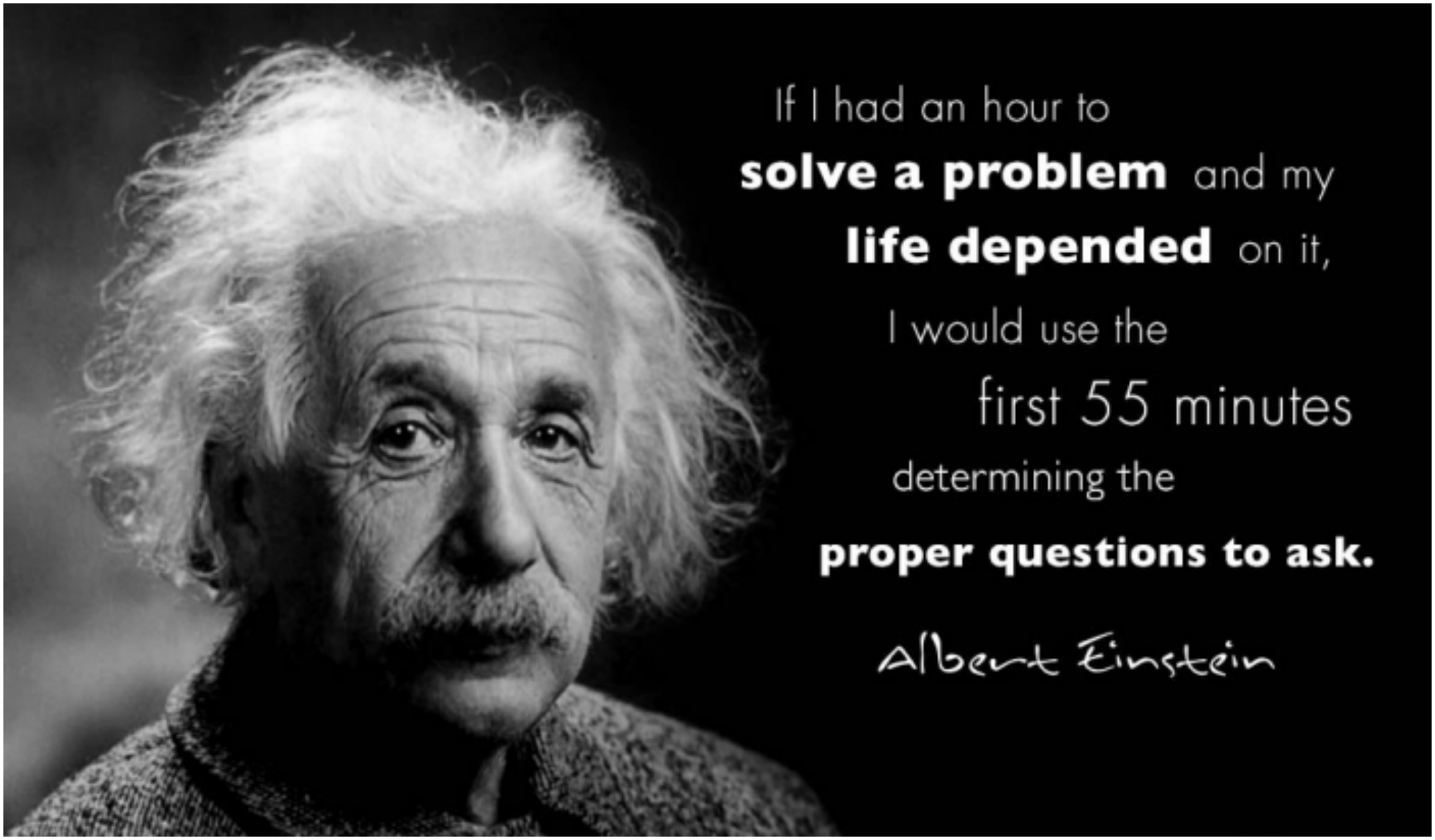


1. Business Understanding



…purely mathematical perspective?

- Can data alone represent the optimal representation of the world?
- many models, features, parameters…
- need for human abilities to streamline the process
- Interpretability

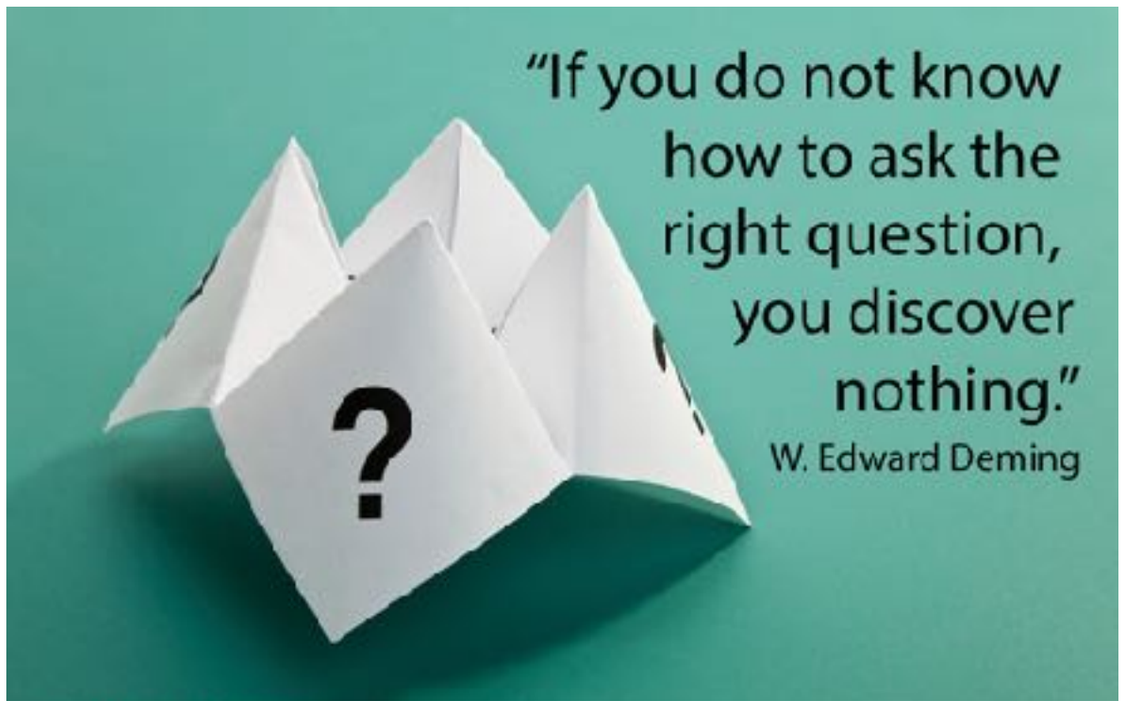

If I had an hour to
solve a problem and my
life depended on it,
I would use the
first 55 minutes
determining the
proper questions to ask.
Albert Einstein

Ask the *right* questions

1. How can we increase profit?

2. How can we increase customer satisfaction?

3. How can we increase market share?



High-Level Questions

- What is the company's goal?
- Who are their users?
- Who are their stakeholders?
- What is the company's structure?
- How does the company make money?

Project-Level Questions

- **What is the goal of this project?**
  - To make *forecast*?
  - To *create* user profiles?
  - To *predict* whether a user will take a certain action?
  - To *test* the success of a new product?
  - To make *recommendations* to users?

- **What data exists and how was it collected?**
    - Can I collect more data?
    - Do I have a budget (of time, money. and/or people power) for this?
    - What is my timeframe?
    - How will I evaluate my findings?
    - How will my findings be used?
    - Is this a supervised or unsupervised learning problem? If supervised, do I currently have access to training data with a labelled target?
    - What are the potential costs and benefits of taking an action based on my findings?

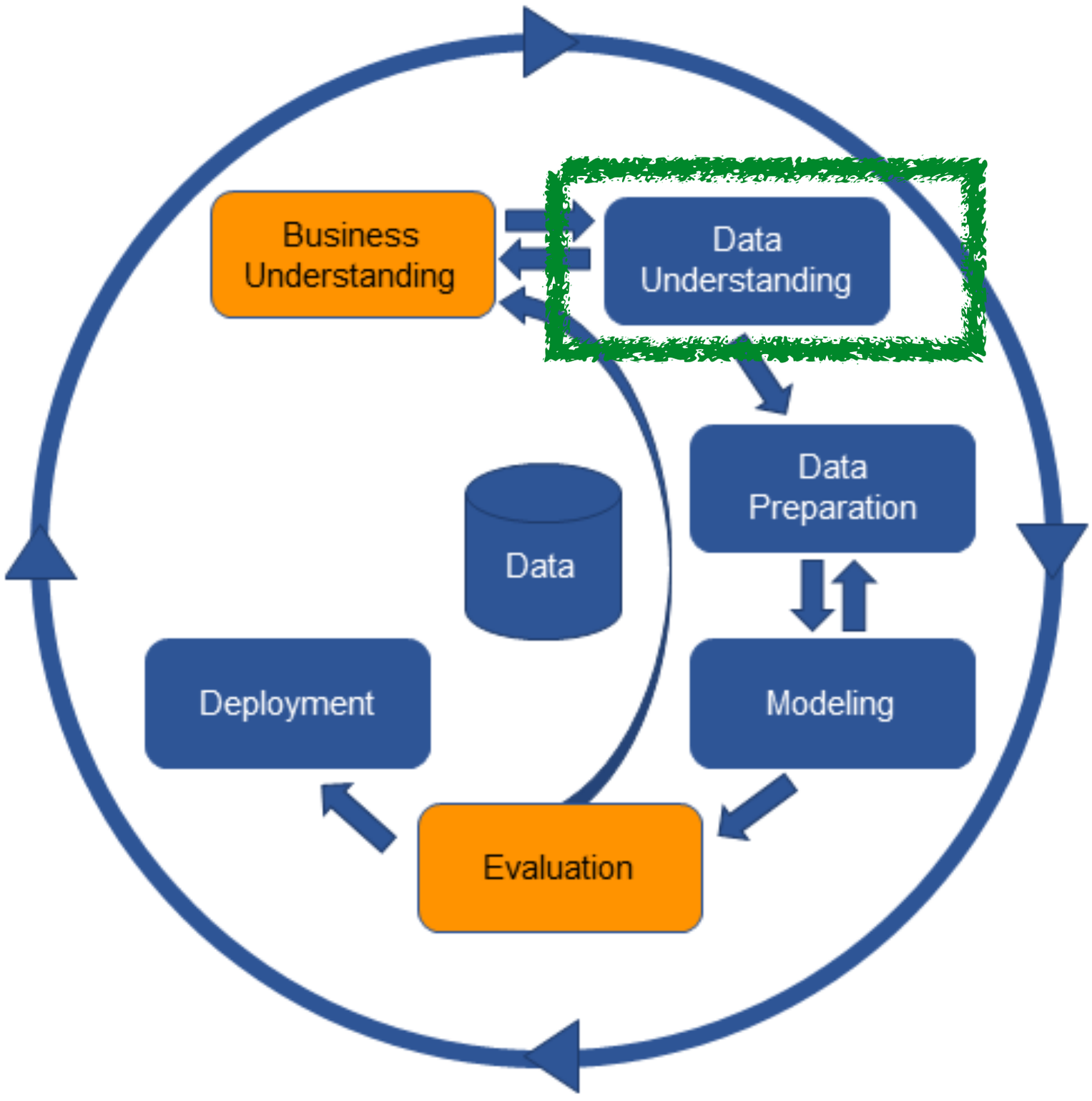
Brainstorming Questions

- What are some of the **features** I predict will be major drivers of my target?

- Are these features easy to collect? Impossible?
    - If hard or impossible, are there other features I could use as **proxies**?

- What will I do with **missing data**?

- How will I **validate** my work?


Business Understanding
Data Understanding
Data Preparation
Data
Modeling
Deployment
Evaluation

2. Data Understanding

Look at the big picture

- What is your goal?
- What type of data modeling does that goal suggest?
    - Supervised classification or regression?
    - Unsupervised clustering or similarity scoring?

Real business problems are complex and will probably require a combination of approaches

Conceptualize an ideal dataset

- In an *ideal world*, what data would you require to solve your business problem?

- Does that data exist?
- If not:
    - Could it be collected?
    - What would be the cost of collecting it?
    - Could a proxy be used?

Think about your data sources

- **How** was this data collected?
- **When** was this data collected?
- Has the way it was collected changed over time?
- Who, where, or what is the data being collected **from**?

- Does all the data of interest exist in a single dataset?
  - If not, how can it be combined?
    - How will you identify the matches?
    - How will you deal with differing formats?

Think about data types

- How clean is my data? Ideally, a data set wiki
  - have only quantitative values
  - have no missing values
  - have no improbable or impossible values .

Are you ever going to encounter a dataset like this outside of academics? NOPE!

- Do you have **categorical data**?
  - How many categories?
  - Can the categories be clustered?
  - How could you represent these categories quantitatively?

- Do you have **time-related data**?
  - What does it represent?
  - Why might it be valuable?
  - At what level of granularity?

- Do you have **strings as data**?
  - What do they represent?
  - Is each string unique?
  - Is uniformity enforced?

- Do you have other **media types** (images, music) **as data**?
- How do these data related to your goal?
- How will you represent them quantitatively?


Business Understanding
Data Understanding
Data Preparation
Data
Modeling
Deployment
Evaluation

3. Data Preparation

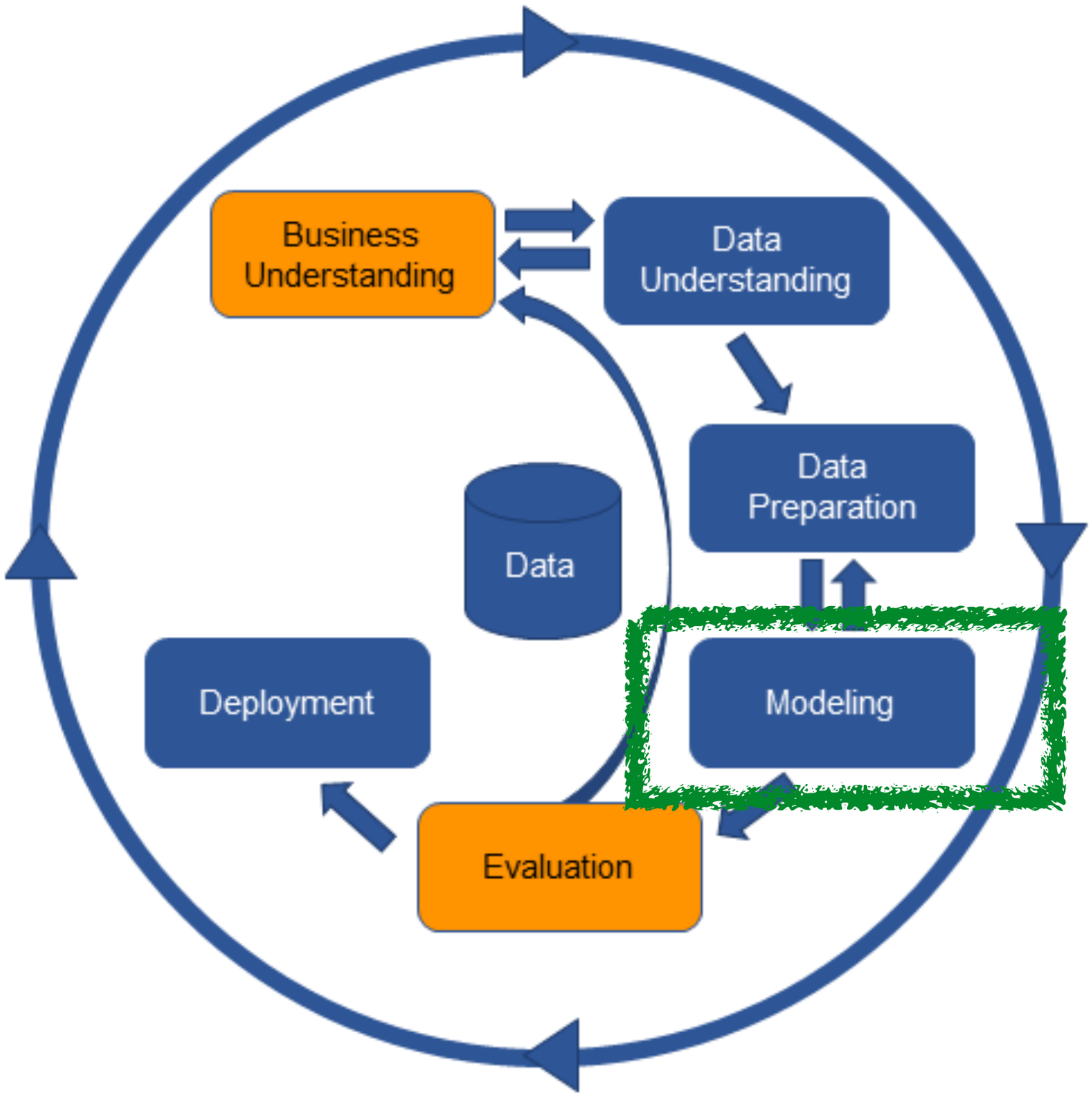
- **Data is messy.**
    - Feature Engineering

- What are some problems we might encounter with our data?

- **We might have missing values. What are some options of dealing with missing values?**
  - Find the mean of the column and replace all missing values with the median.
  - Do the above, except use the median.
    - Drawback? Imputing. We end up skewing our data.
    - Collinear with another feature?

- If your data is binary/class-based, fill missing values with the majority class.
- Drop all data points with missing values.
- Drop the column(s) with missing values.
- Encode the fact that the data is missing.

- **Important Note**: Any feature engineering you perform:
    - Should only be performed on training data, never validation data or test data.
        - It must be *applied* to the test data
            - If you inputed the mean for some missing values, you **must** *use the same mean* to fill in missing values in your test data. DO NOT take the mean of your test data. Similarly, if you standardized the data in one or more columns, you need to scale the test data according to those rules.


Business Understanding
Data Understanding
Data Preparation
Data
Modeling
Deployment
Evaluation

4. Modeling

- Without the steps surrounding modeling, our models are worthless.

- Discuss the *questions you should ask* to lead you to the models that are appropriate for your business problem

Supervised Learning Problem

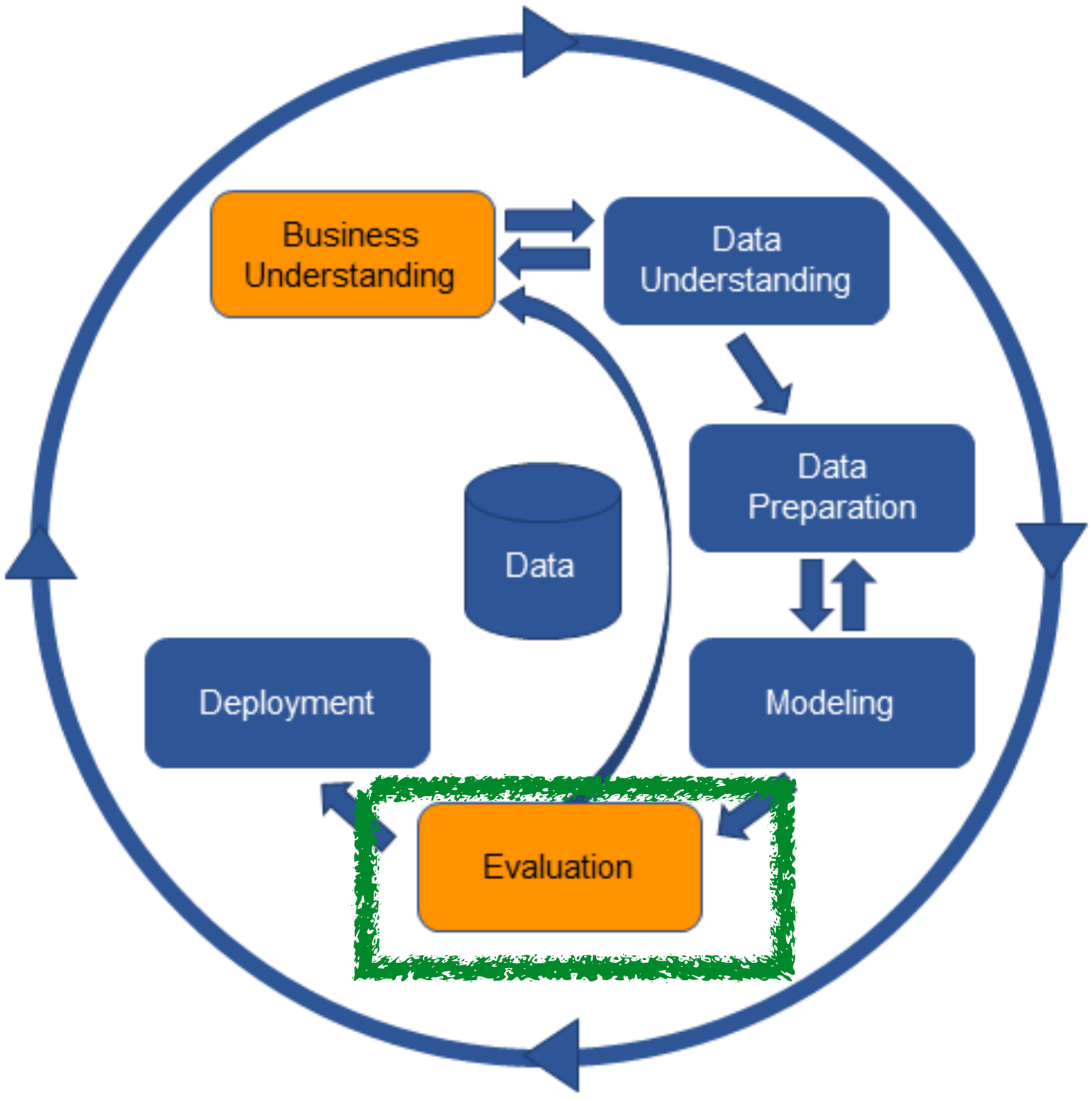
- Always start with a **baseline model.**
  - **Regression Problems**
    - Predict the mean or median of the target
    - Predict a random value between the min and max of the target
  - **Classification Problem**
    - Predict the majority class
    - Compute the probability of each class and predict a random class based on this breakdown

Error Metric(s)

- Regression Problems
  - (Root) Mean Squared Error
  - Coefficient of Determination (usually called R2)
- Classification
  - Accuracy
  - Recall
  - Log loss

Takeaway

- You should establish an **error metric** *at the beginning of the process* (and have a good reason for choosing it), so that you have a way of comparing one model to the next to assess its performance relative to your goal. Different business goals will suggest different error metrics.


Business Understanding
Data Understanding
Data Preparation
Data
Modeling
Deployment
Evaluation

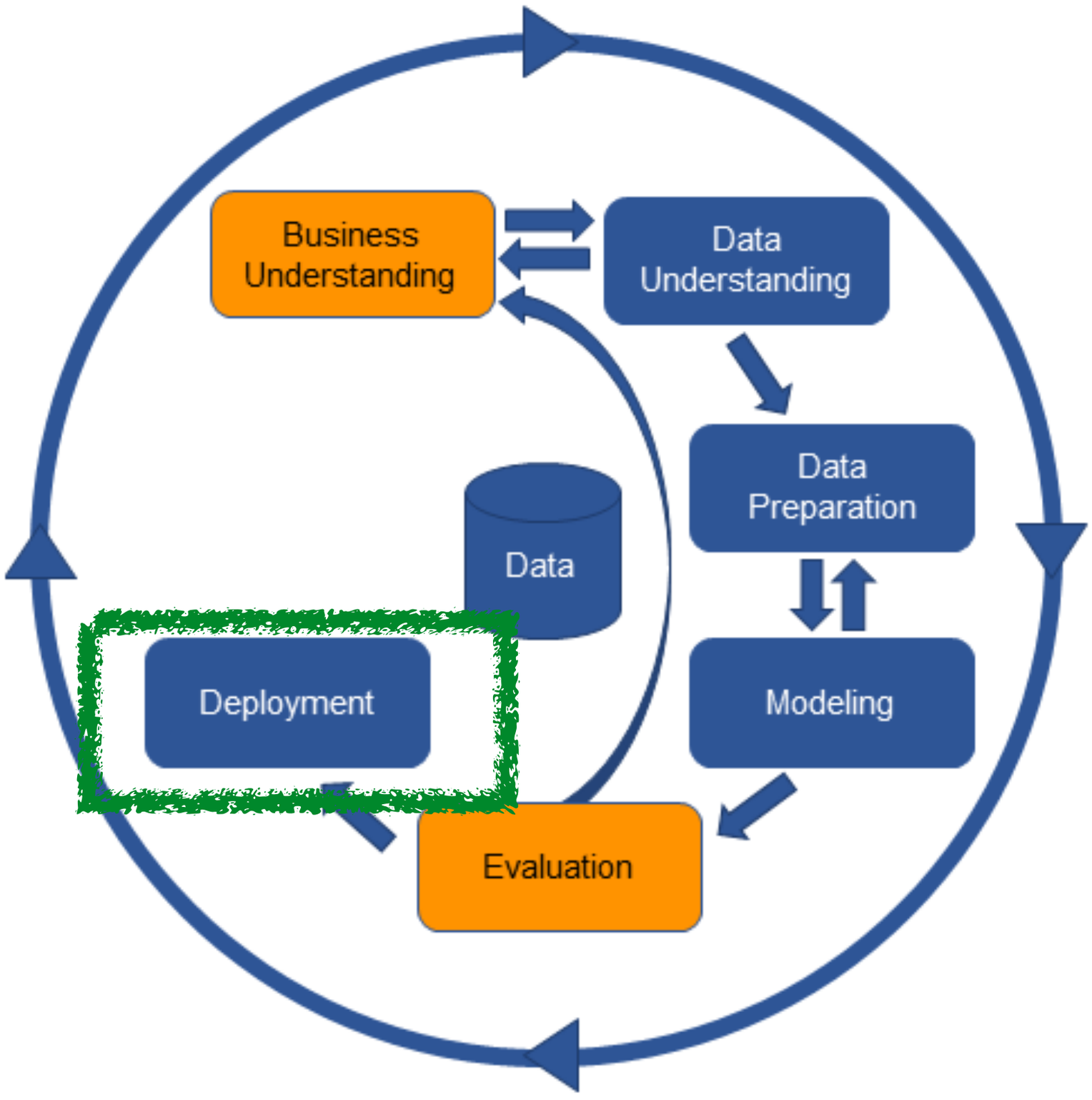
5. Evaluation

- **Without the ability to evaluate models, models themselves are useless.**

- There should be a feedback loop between evaluation and modeling.

  - Cross-validation

  - Reality checks (how robust our model will be in the real world?)


Business Understanding
Data Understanding
Data Preparation
Data
Modeling
Deployment
Evaluation

6. Deployment

- This part can mean a lot of things
  - Applying predictions to **new data** (filtering spam, offering loans of a certain amount)
  - Deploy the model itself in a **dynamic fashion** such that it adapts to user behavior.

Putting It All Together

- You can find a short case here.
- **Individually**. Please answer each question twice:
  1. The answer you'd provide if your counterpart were a data scientist
  2. The answer you'd provide if your counterpart were the CEO of the company
- Working in teams, prepare a short presentation to motivate and illustrate your answers.