The Machine Learning Process

Machine Learning

- **Supervised Learning:** Supervised Learning is a class of Machine Learning in which a model is "trained" using a set of pre-existing labeled data.

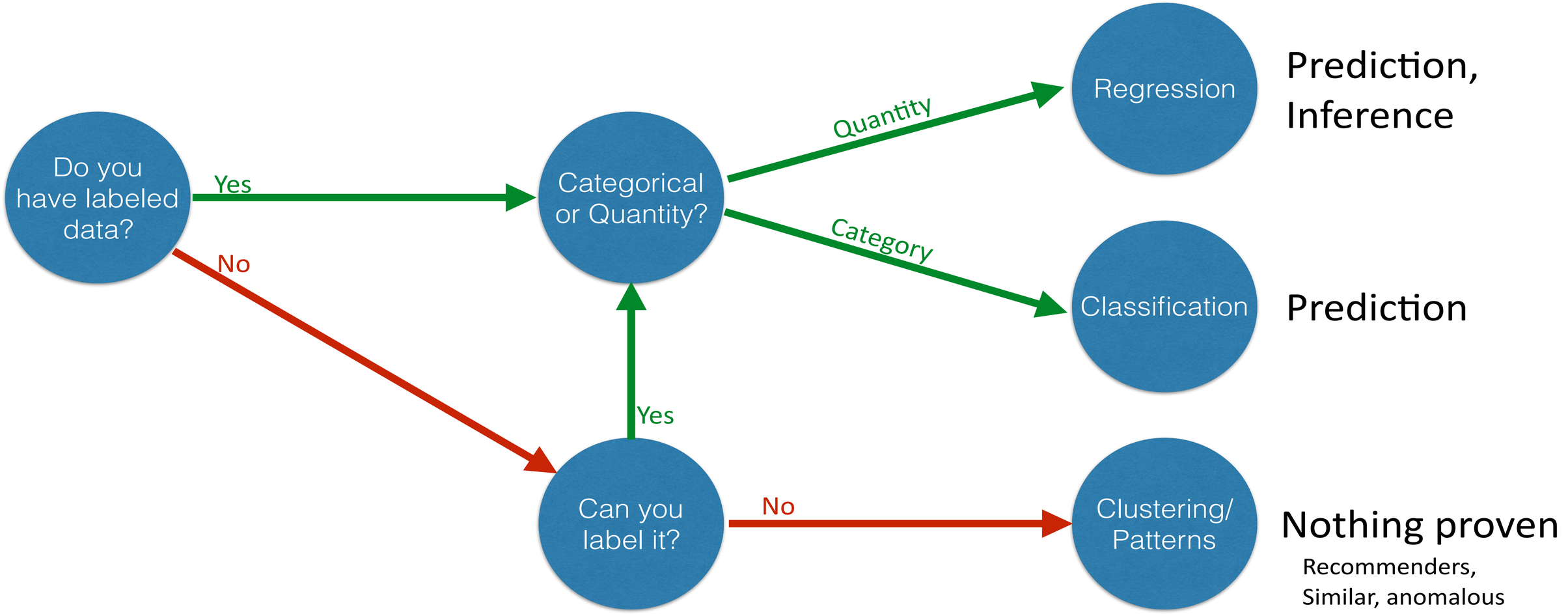
- **Unsupervised Learning:** A class of Machine Learning algorithms in which a model is built without the use of labeled data.

Machine Learning Problem Types

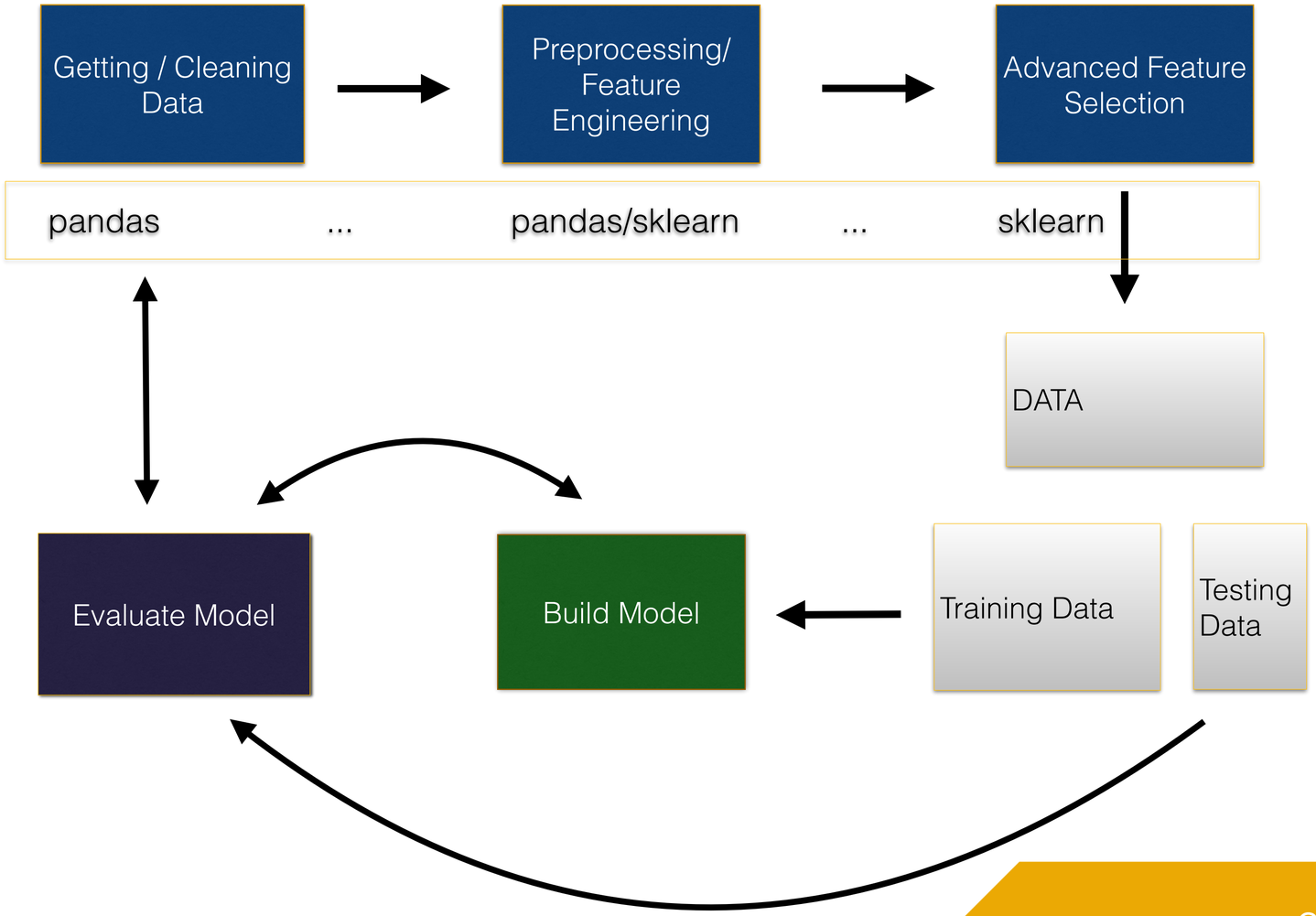
- **Classification**: Assigning or predicting a observation's membership in discrete class
- **Regression:** Predicting a continuous value based on the observations' features

- **Clustering:** Identifying groupings within a dataset
- **Dimensionality Reduction:** Reducing the number of variables in a feature set

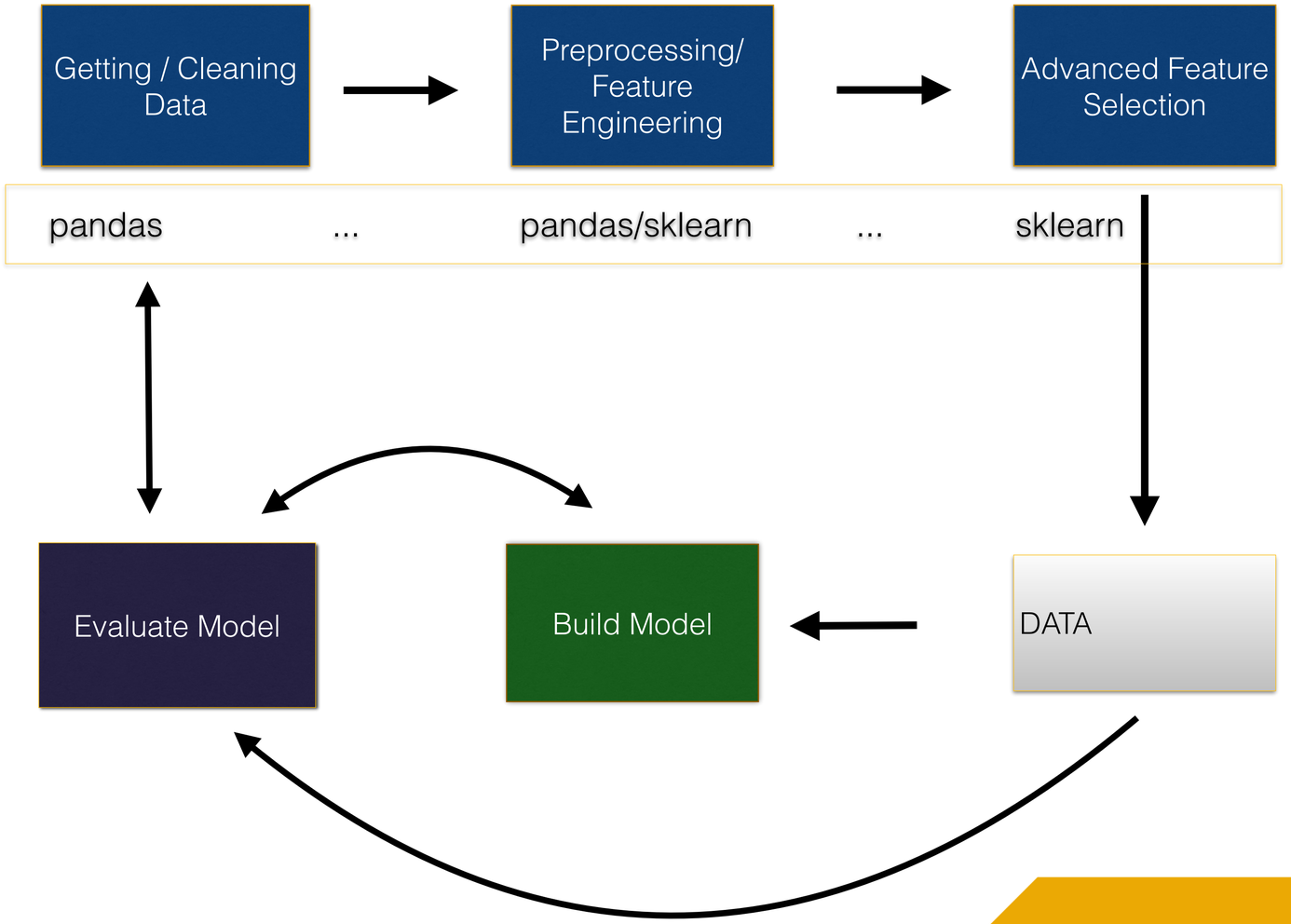
What Problem am I solving?



Supervised Machine Learning Process



Unsupervised Machine Learning Process



First, define your analytic question.

What are you trying to do?

How do you define success?
What are you measuring?

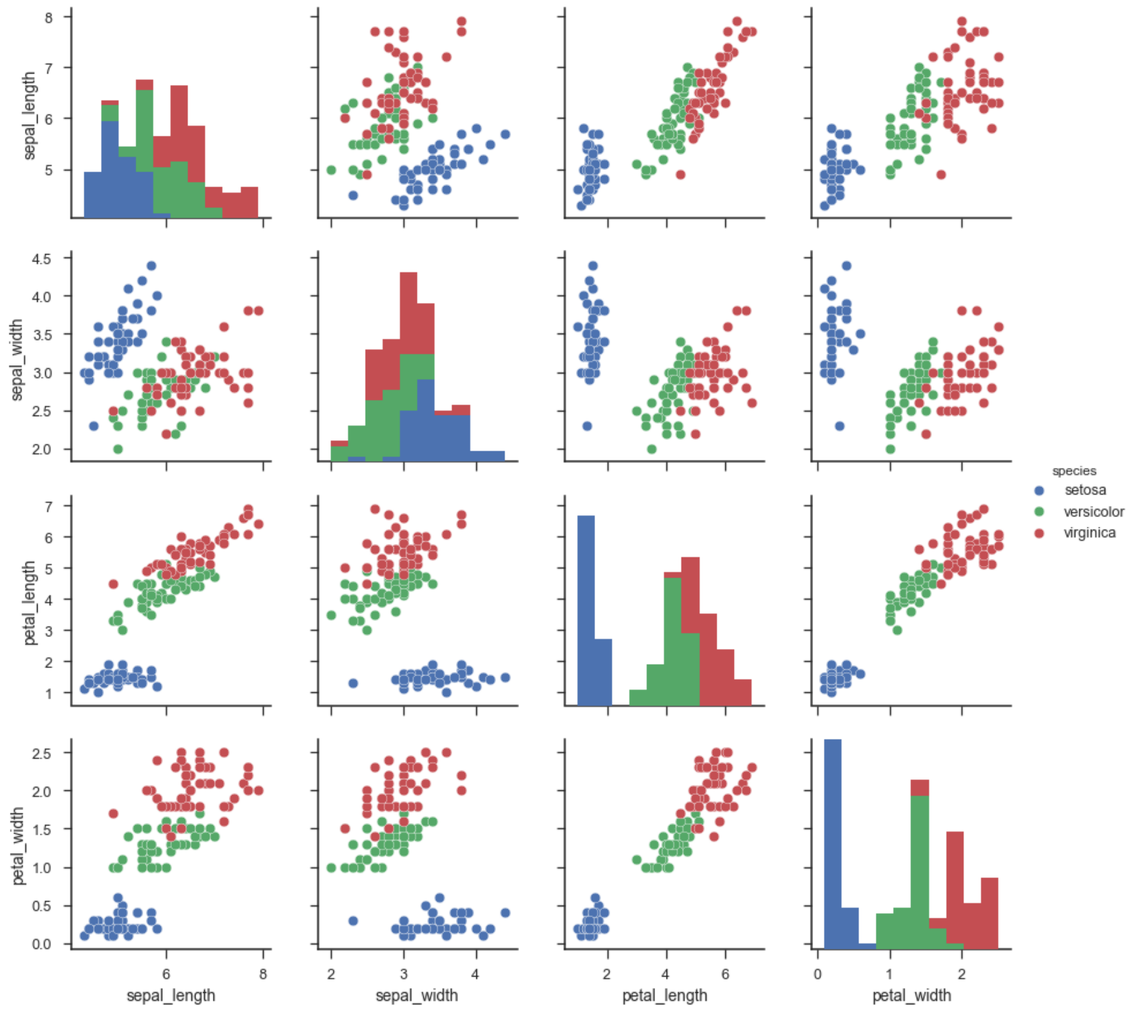
Choose data sources

- What is available?
- Is it enough?
- Is the data reliable/clean/consistent?
- What other data could you use?

Other Considerations

- Policies
- Legal contraints
- Biases in Data
- Latency
- Data size

Gather and Explore Your Data



Is the data good enough?

What are the rules governing its use?

Do I have enough?

Do problems or biases exist in the data that could cause problems?

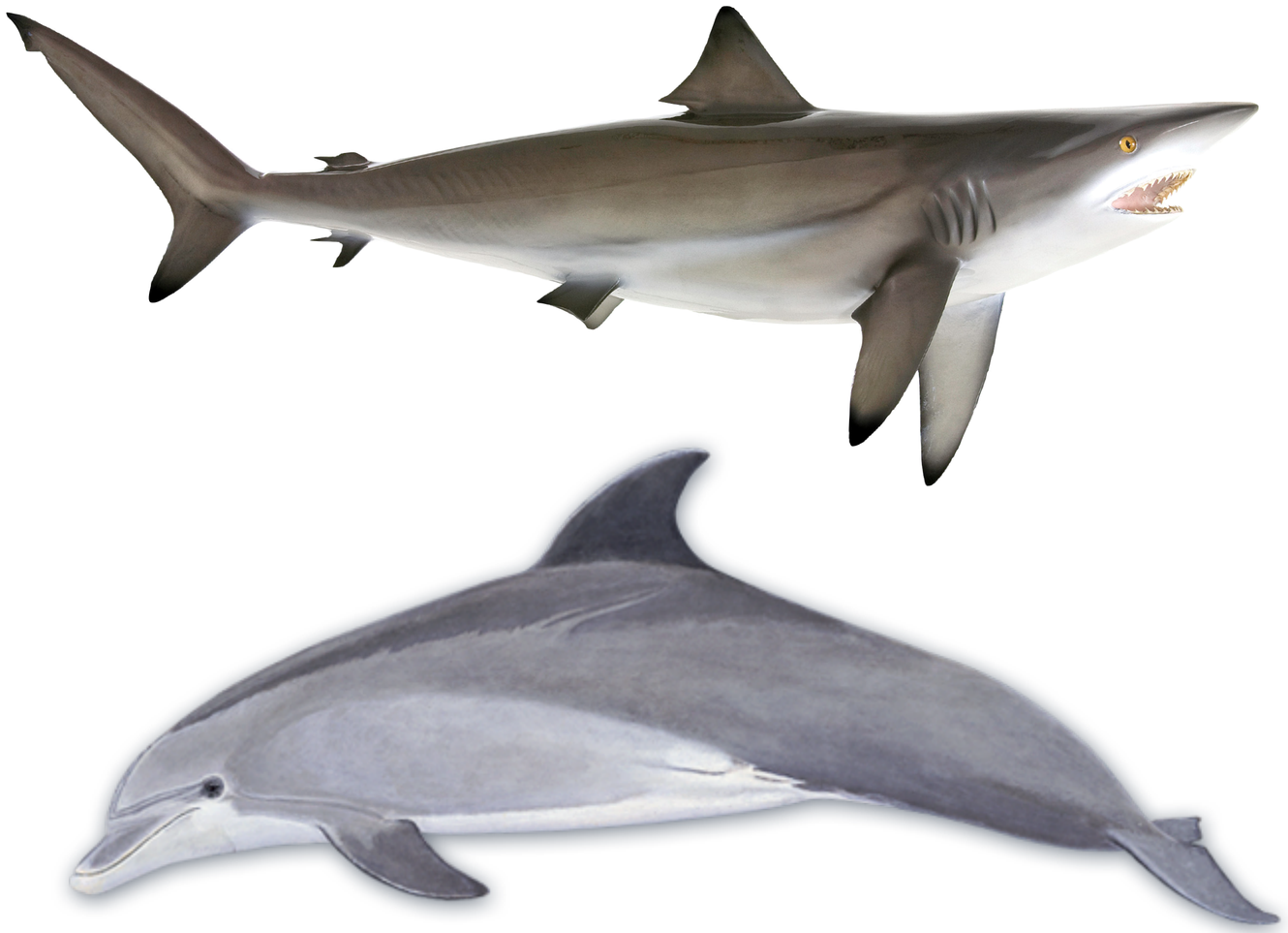
Feature Engineering

- Define what you are trying to measure. These will become the **observations** or rows of your final dataset
- Define how you will mathematically represent your data. This will be come the **features** or columns of your final dataset.

Feature Engineering

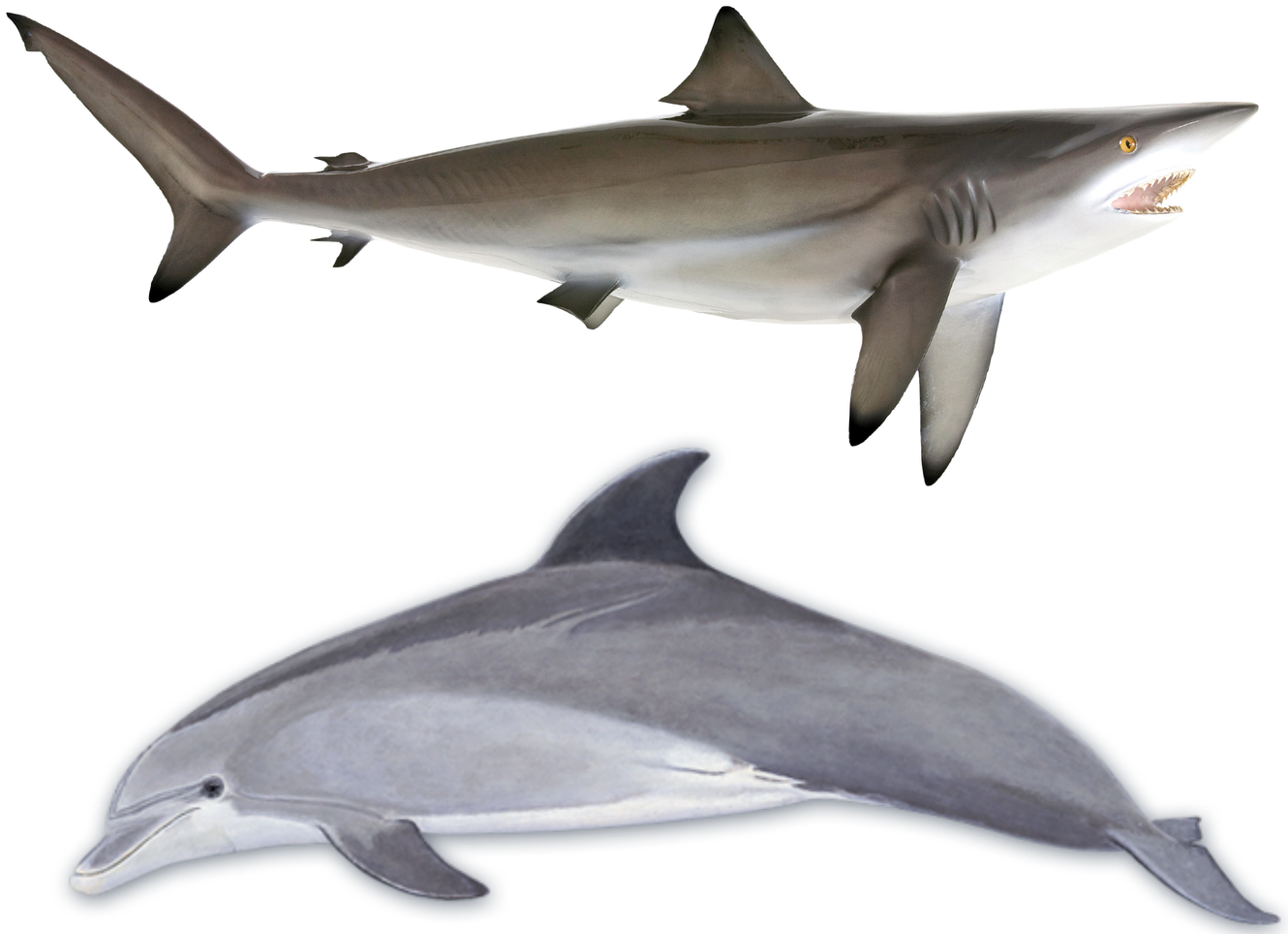
Feature	Value
Color	Gray
Fins	7
Predator	TRUE

Feature Engineering



Feature	Value
Color	Gray
Fins	7
Predator	TRUE

Feature Engineering



Feature	Value
Color	Gray
Fins	7
Predator	TRUE
Mammal	**TRUE**

Build and Tune your Model

- Believe it or not, this is the easy part.
- Most of this is **done using libraries** like scikit-learn, mllib, tensorflow, caret or keras, and **many steps can be automated**.
- You can even do it in Splunk or Elasticsearch.

Evaluate Performance

- Use various scoring methods, or write your own to determine model performance.
- Go back to step 1 and repeat! (Do not pass go, do not collect $200)

Group Discussion

Consider that you are building a system to identify fraudulent credit card transactions. In your groups, try to answer the following questions:

1. What are some features that you would want to capture?
2. What data sets will you need?
3. What legal and policy challenges might you face?
4. What other challenges you could foresee in this problem?
5. How will you define success?
6. How can you articulate the value of this model to stakeholders?

The Python Data Science Ecosystem

Machine Learning Ecosystem

- **Data Gathering:** Pandas, Drill, BeautifulSoup, PyDBAPI, PyDAL, Boto3
- **Feature Extraction:** Pandas, NumPy, Featuretools
- **Machine Learning**
  - **"Regular" ML:** Scikit-learn (sklearn), h2o, mllib (PySpark)
  - **Deep Learning:** Tensorflow, Keras, Theano, Caffe, PyTorch, HuggingFace
- **Visualization**: Matplotlib, Seaborn, LIME, plotly, Streamlit

Data Scientists spend 50-90% of their time being…

Data Janitors