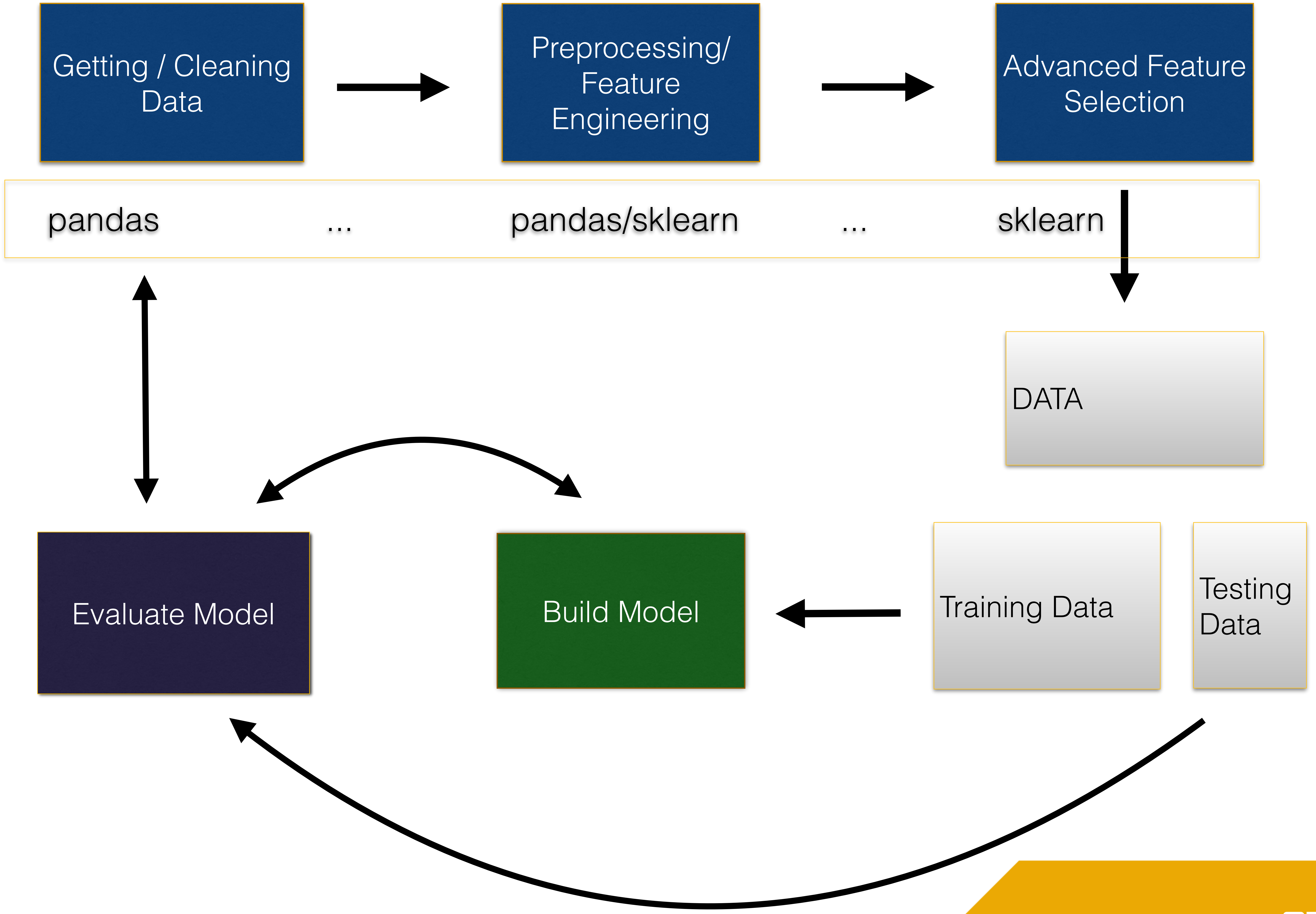
The Machine Learning Process

Supervised Machine Learning Process


Getting / Cleaning Data

Preprocessing/ Feature Engineering

Advanced Feature Selection

pandas ... pandas/sklearn ... sklearn

DATA

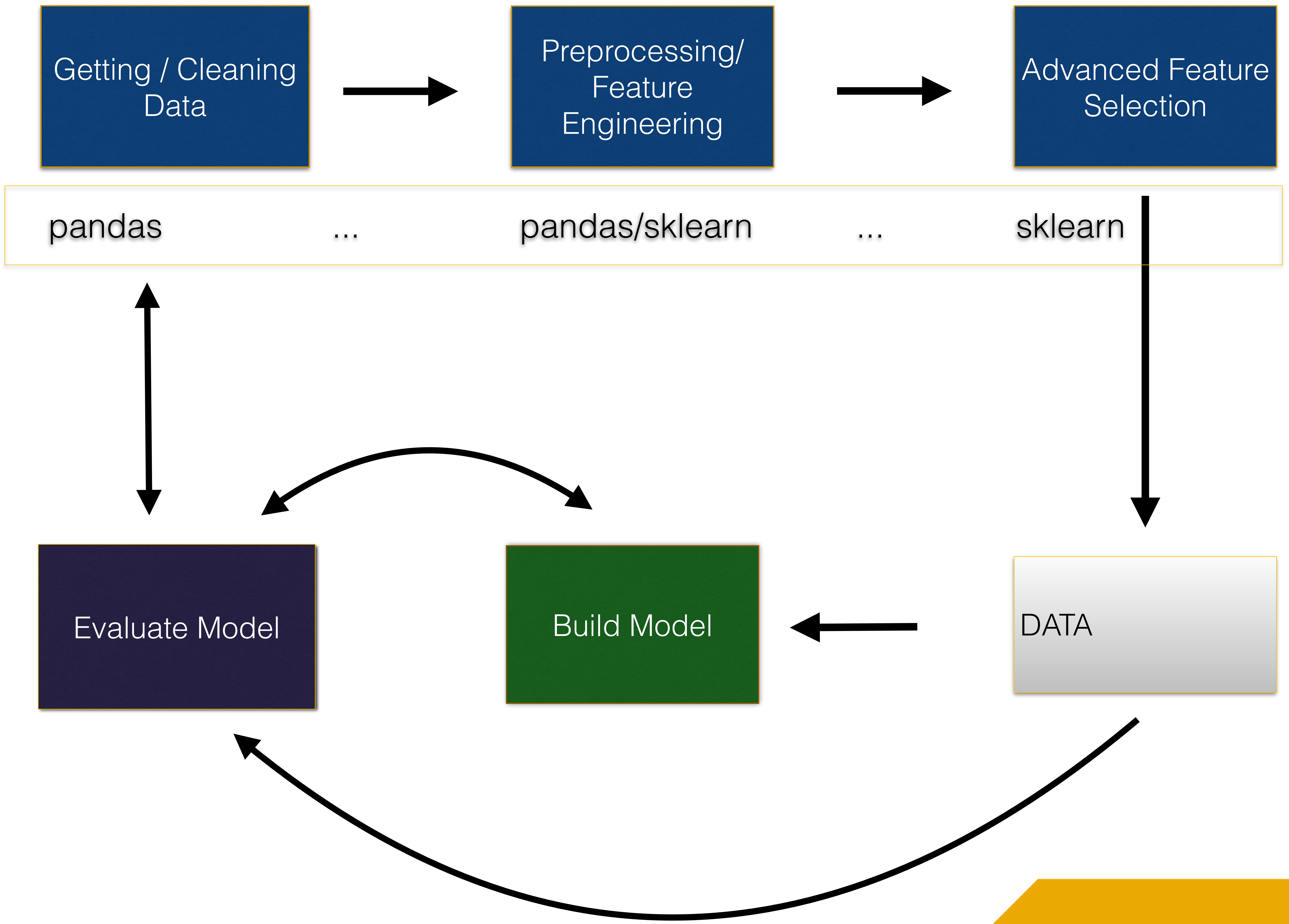
Training Data

Testing Data

Build Model

Evaluate Model

Unsupervised Machine Learning Process



First, define your analytic question.

What are you trying to do?

How do you define success?
What are you measuring?

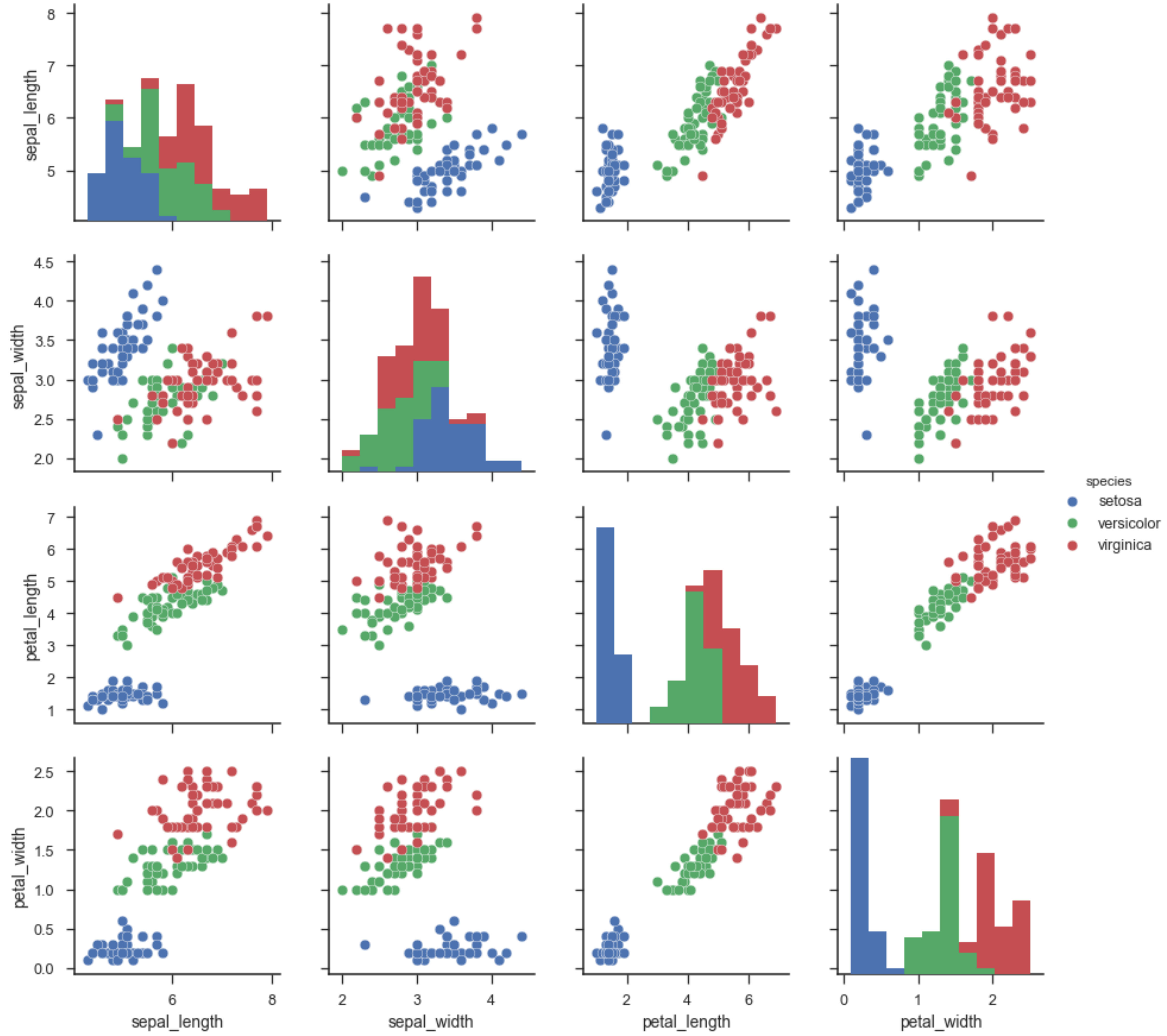
Choose data sources

- 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?

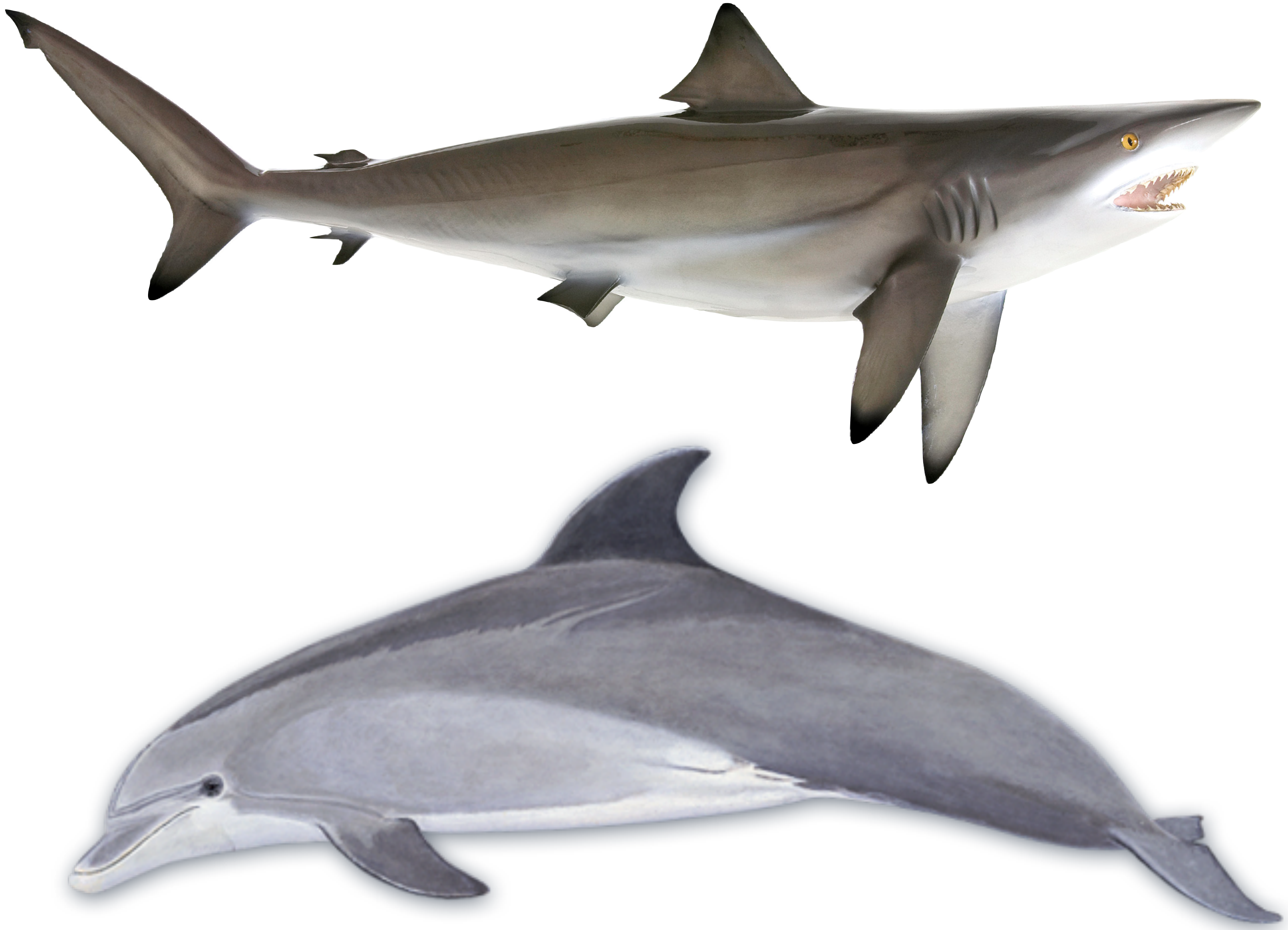
Feature Engineering

- 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

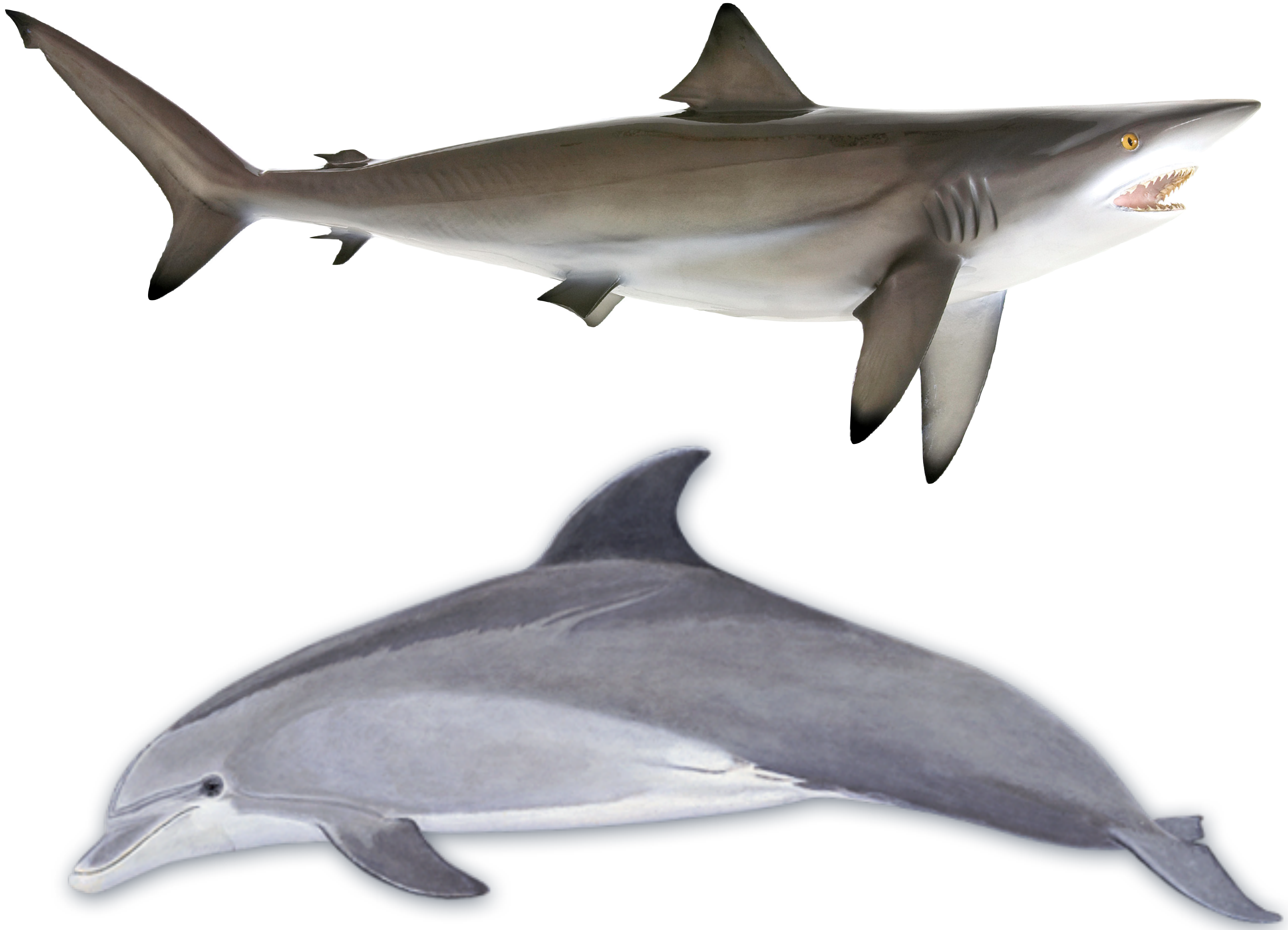
Feature	Value
Color	Gray
Fins	7
Predator	TRUE

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.

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