



Attacking Artificial Intelligence Introduction

Charles S. Givre CISSP

Course Agenda

Day 1: Introduction & Machine Learning

- Overview of Machine Learning
- Red Teaming ML models

• Day 2: Generative AI

- AI Theory
- AI Architecture

• Day 3: Red Teaming Generative AI

Our Lawyers Make Us Say This



All materials presented in this training and those provided as an adjunct to the program are copyrighted 2020 by GTK Cyber LLC.

They are intended solely for the use of registered program participants and may not be reproduced or redistributed in any manner for any other reason.

Charles Givre, CISSP

- Sr. Tech Lead @ RTX
- Startup Founder
- Ex Deutsche Bank, JP Morgan
- PMC Chair for Apache Drill
- Senior Lead Data Scientist @ Booz Allen
- 5 Years @ CIA
- Undergraduate in Comp.Sci & Music

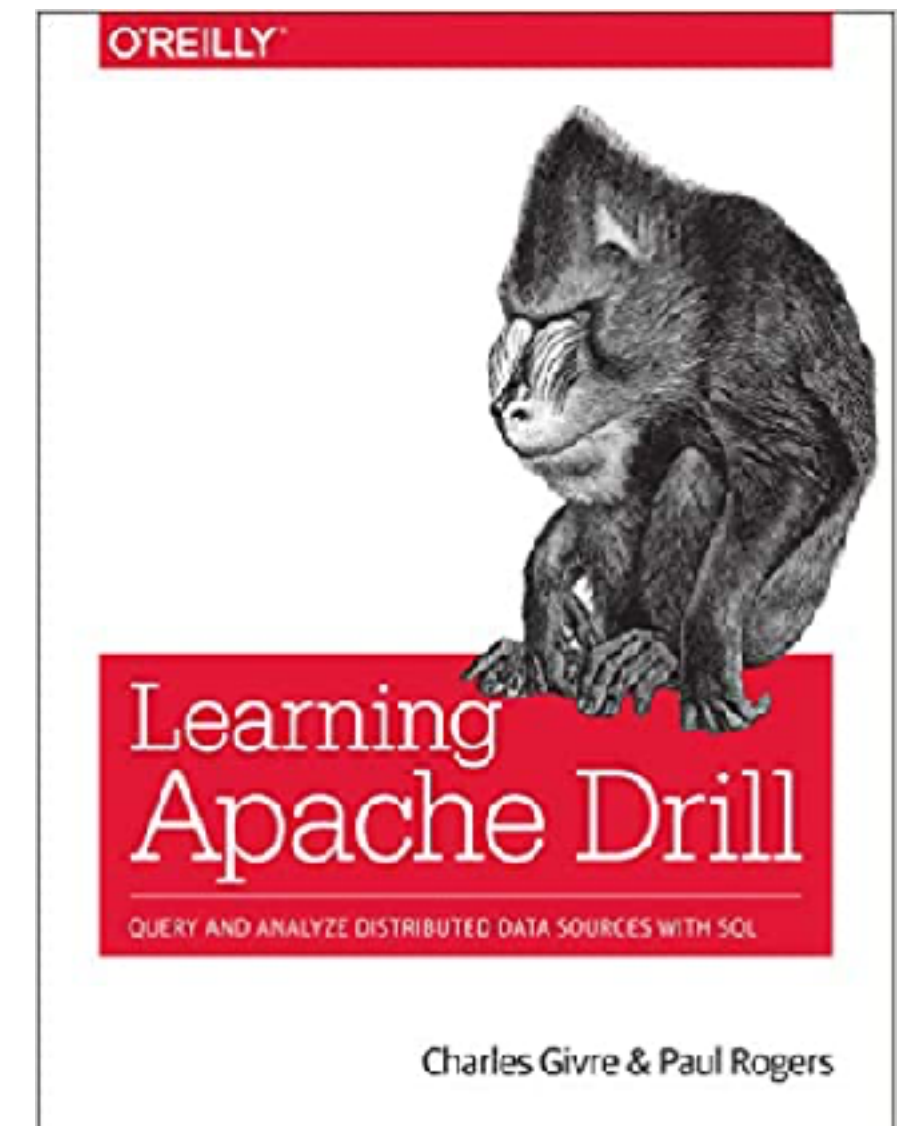


JPMORGAN
CHASE & CO.




Booz | Allen | Hamilton

100 YEARS




Who are you?

- Your name (or what you want us to call you)
 - Your job role
 - What you hope to get out of this class
 - Your level of experience with coding
- 
- A solid yellow geometric shape in the bottom right corner of the slide, consisting of a rectangle with a diagonal cut-off corner.

Why Attack?

A Brief History...

- In 2016-2019 time frame, many research papers were published on this topic, yet there are few known attacks using these techniques
 - 2023: The Year of Generative AI:
 - 2024: People realize that Generative AI is really hard to secure
 - 2025: Let's throw AI agents into the mix!
- 
- A solid yellow decorative shape in the bottom right corner of the slide, consisting of a trapezoid with a diagonal cut on its left side.

What Can You Do?

What is Machine Learning (ML) Artificial Intelligence (AI)

“Machine Learning is the science of getting computers to act without being explicitly programmed.”

– <https://www.coursera.org/course/ml>

"A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E ."

–Tom Mitchell, Carnegie Mellon University

“Machine learning explores the construction and study of algorithms that can learn from and **make predictions on data**. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, **rather than following strictly static program instructions.**”







- Blacklists
- Simple keyword matching
- Naive Bayesian Classifiers
- Deep Learning

Artificial Intelligence

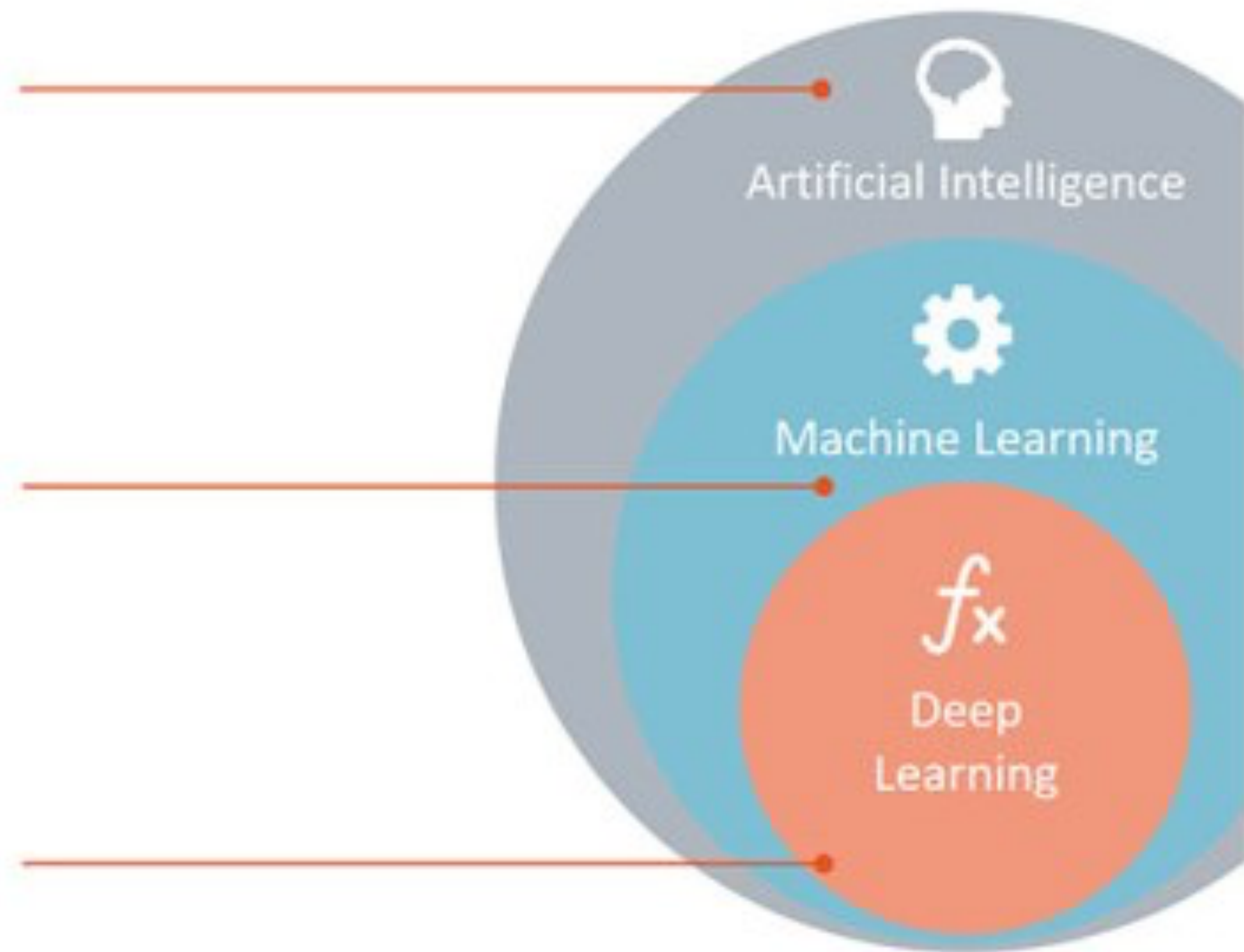
Any technique which enables computers to mimic human behavior.

Machine Learning

Subset of AI techniques which use statistical methods to enable machines to improve with experiences.


Deep Learning

Subset of ML which make the computation of multi-layer neural networks feasible.




[@katherinebailey](#) Because marketing? Every time someone calls simple linear regression “AI” Gauss turns over in his grave.

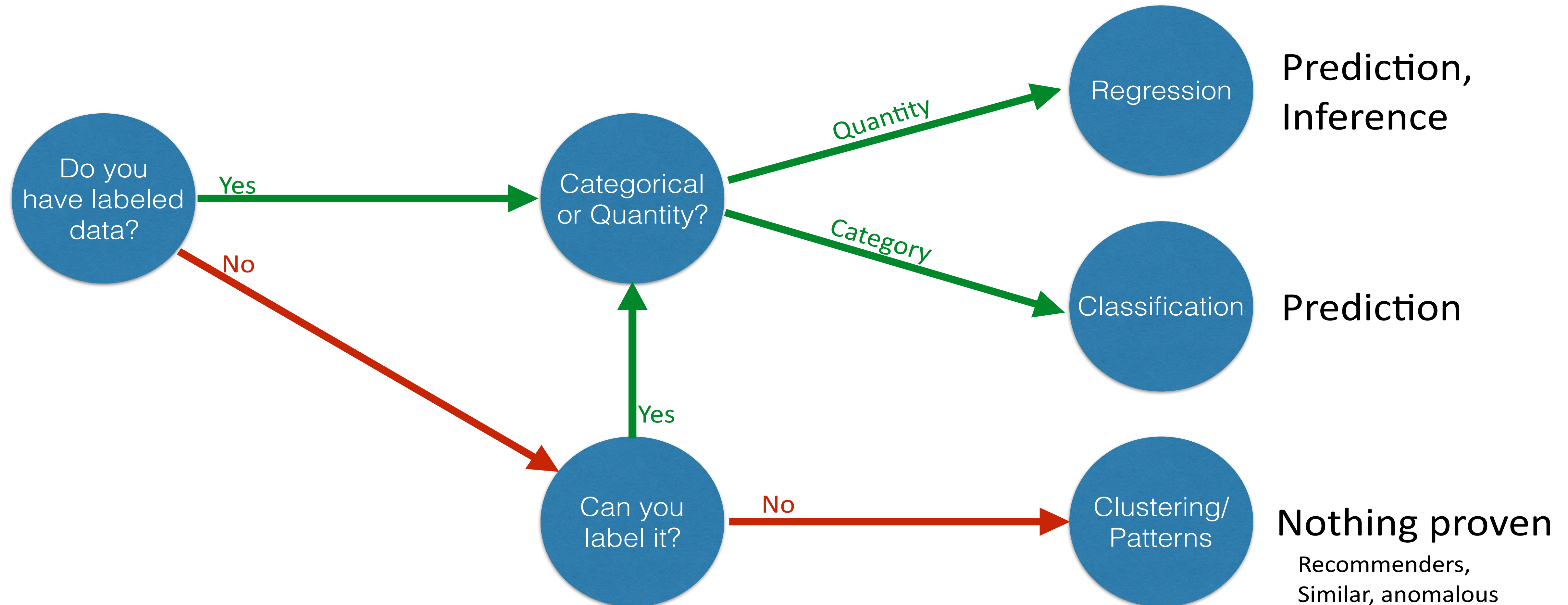
Machine Learning Problems

- **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.
- 
- A decorative orange geometric shape, resembling a stylized arrow or a corner piece, is located in the bottom right corner of the slide.

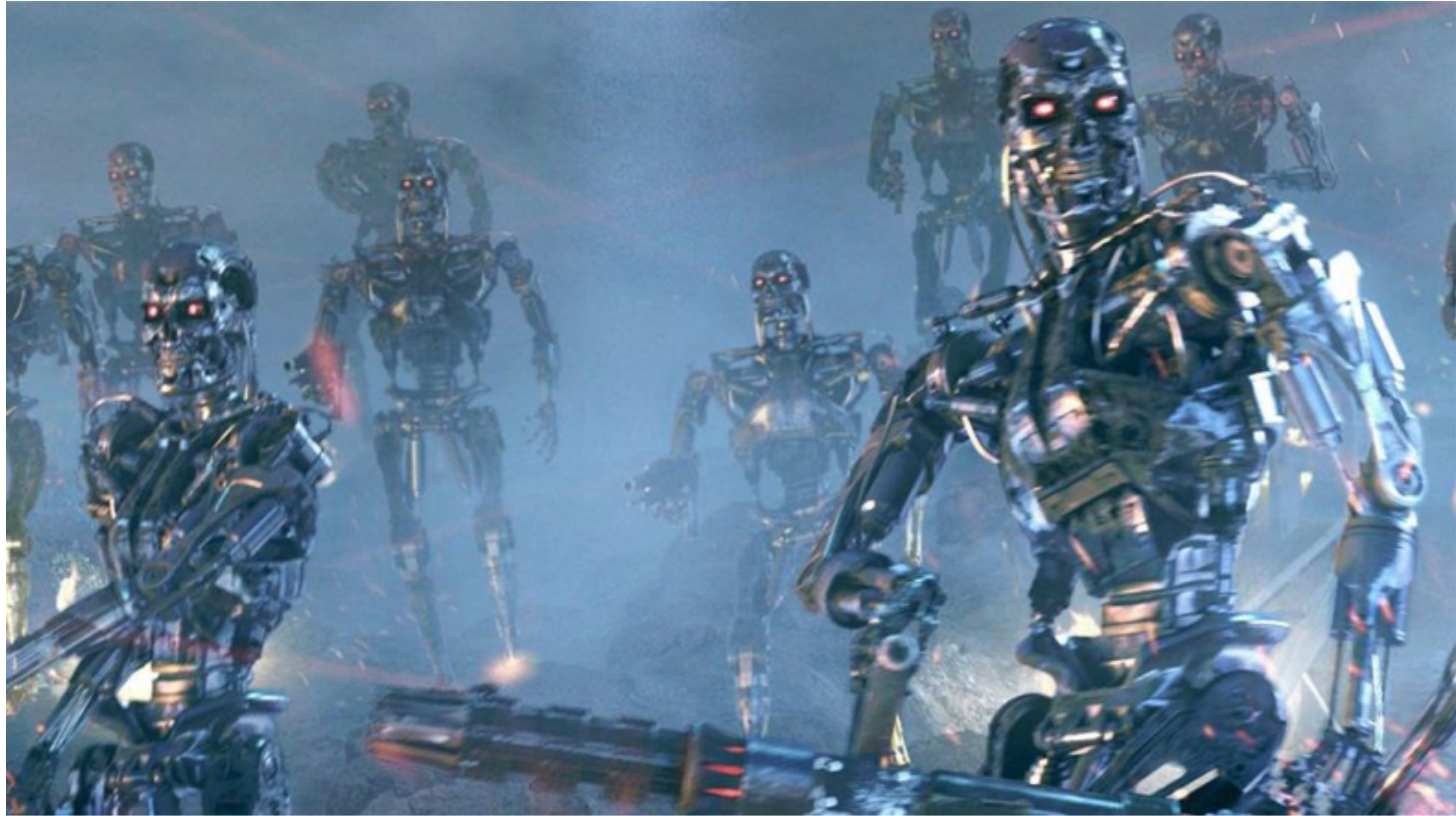
Machine Learning Problem Types

- **Classification:** Assigning or predicting an 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
- 
- A decorative orange geometric shape, resembling a stylized arrow or a corner piece, is located in the bottom right corner of the slide.

What Problem am I solving?

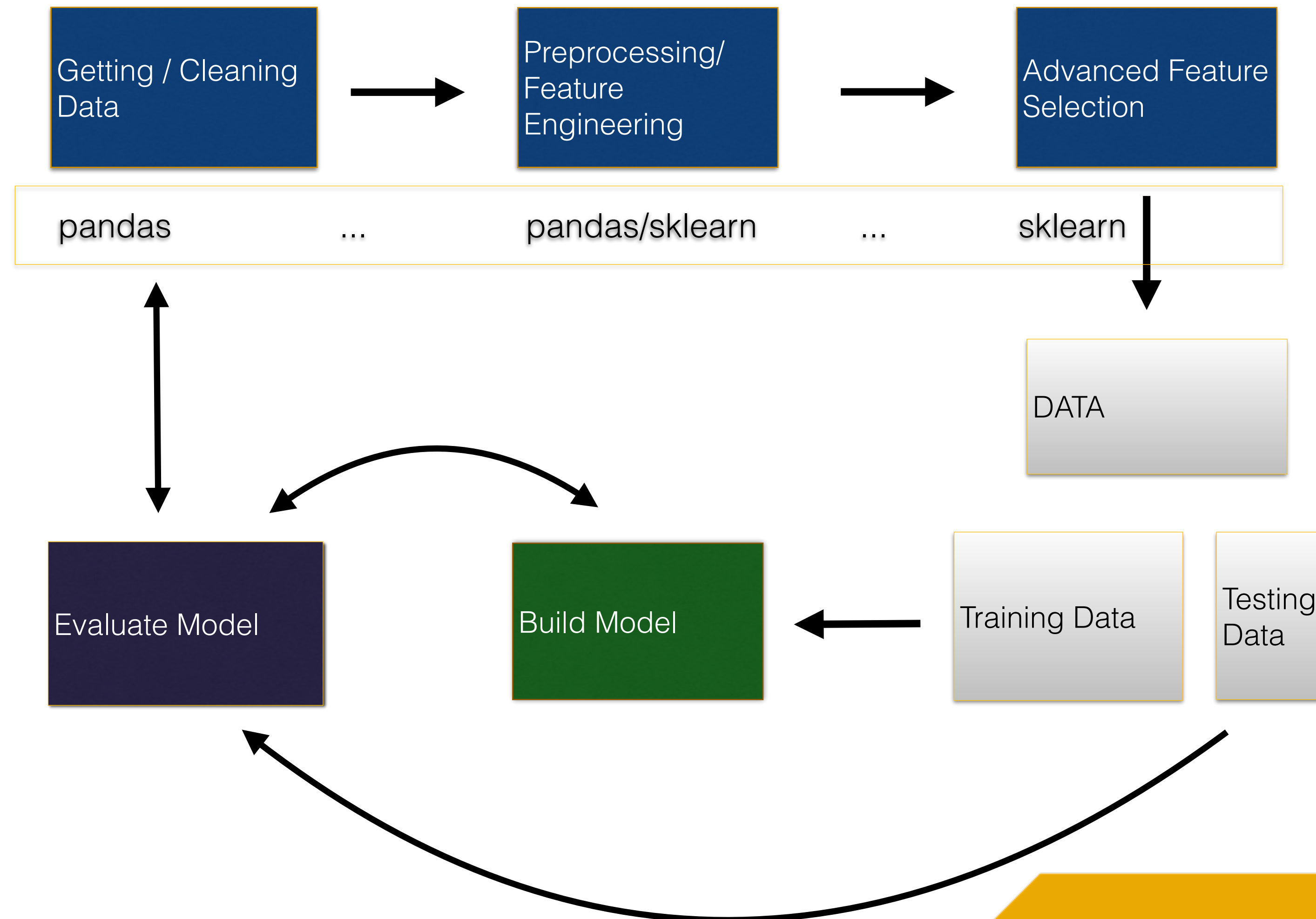


What it is Not

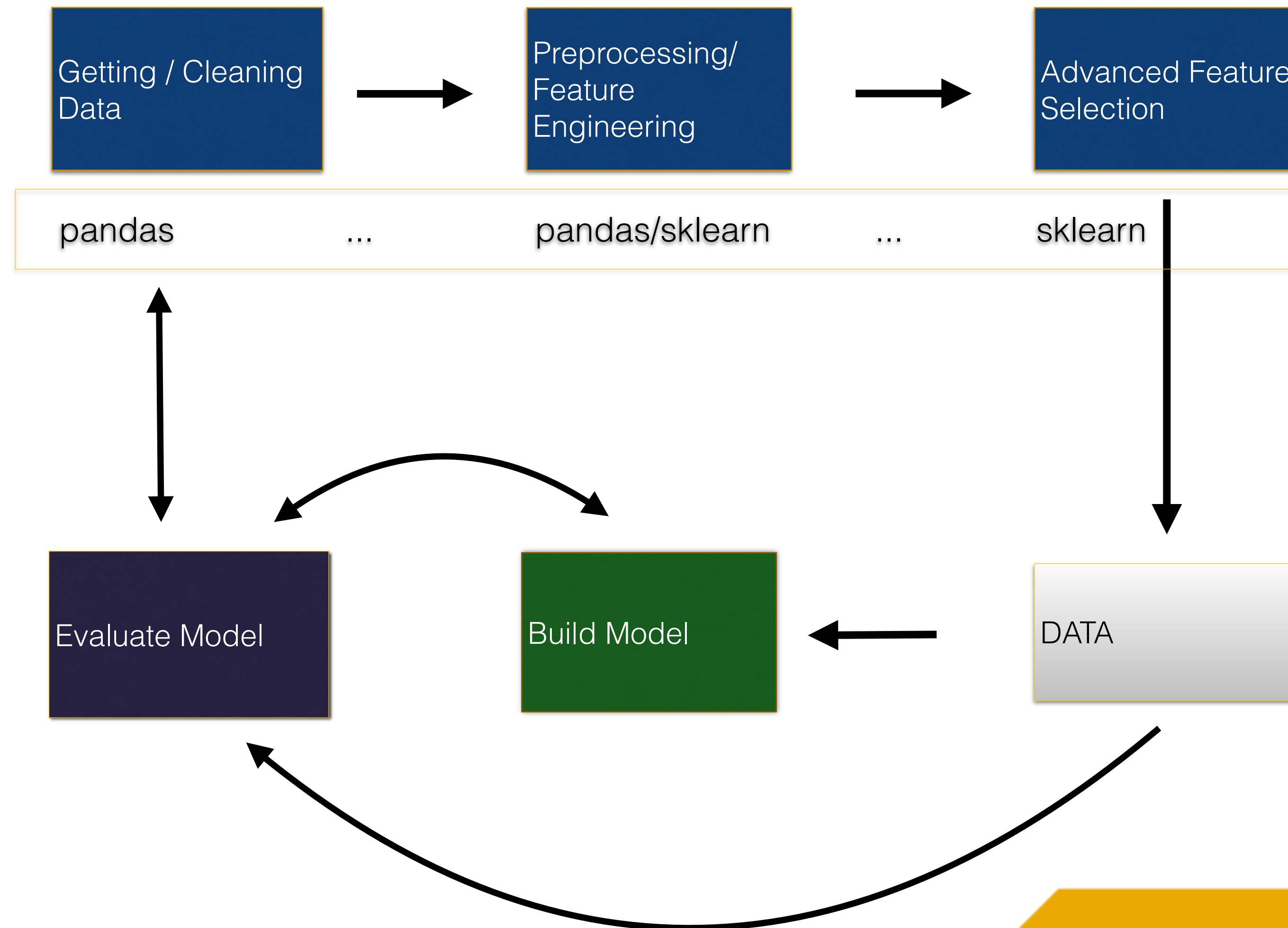


The Machine Learning Process

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?**

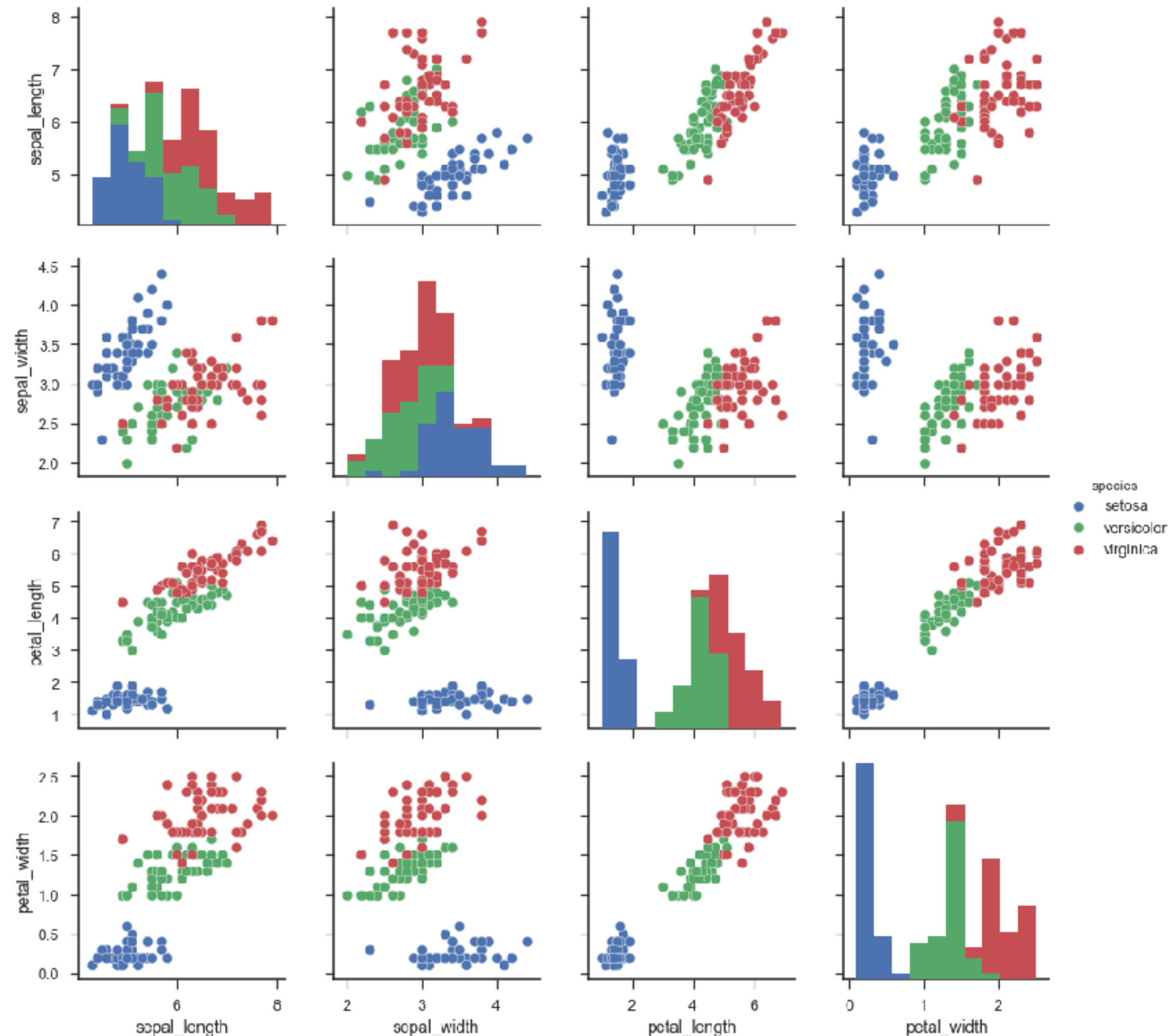
Choose data sources

- What is available?
 - Is it enough?
 - Is the data reliable/clean/consistent?
 - What other data could you use?
- 
- A decorative orange geometric shape is located in the bottom right corner of the slide, consisting of a large triangle pointing upwards and to the right, with a smaller triangle attached to its left side, pointing upwards and to the left.

Other Considerations

- Policies
 - Legal constraints
 - Biases in Data
 - Latency
 - Data size
- 
- A decorative orange geometric shape is located in the bottom right corner of the slide, consisting of a trapezoid and a triangle.

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

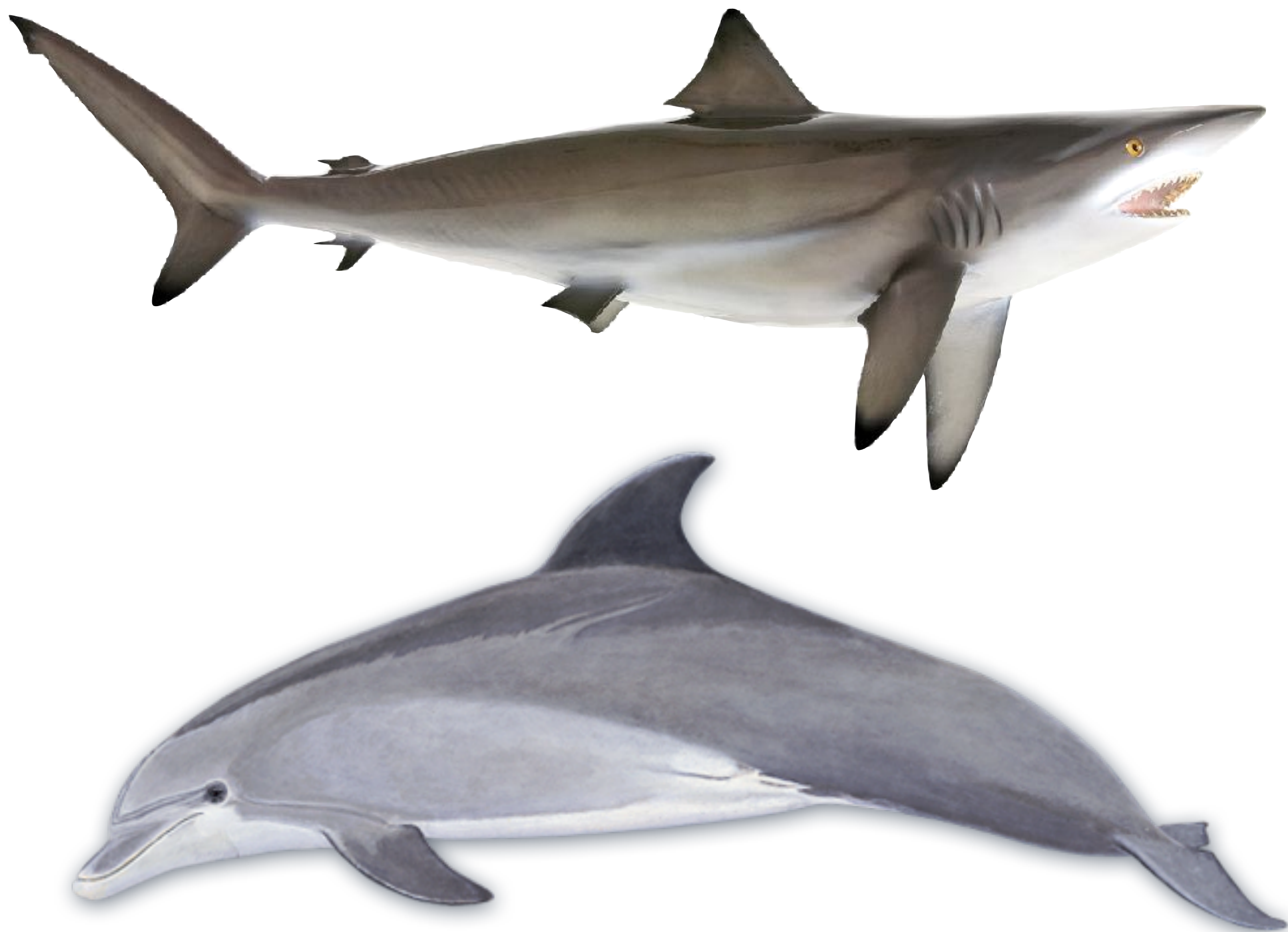
- Define what you are trying to measure. These will become the **observations** or rows of your final dataset
 - Define how you will mathematically represent your data. This will become the **features** or columns of your final dataset.
- 
- A decorative orange geometric shape, resembling a stylized 'L' or a corner piece, is located in the bottom right corner of the slide.

Feature Engineering



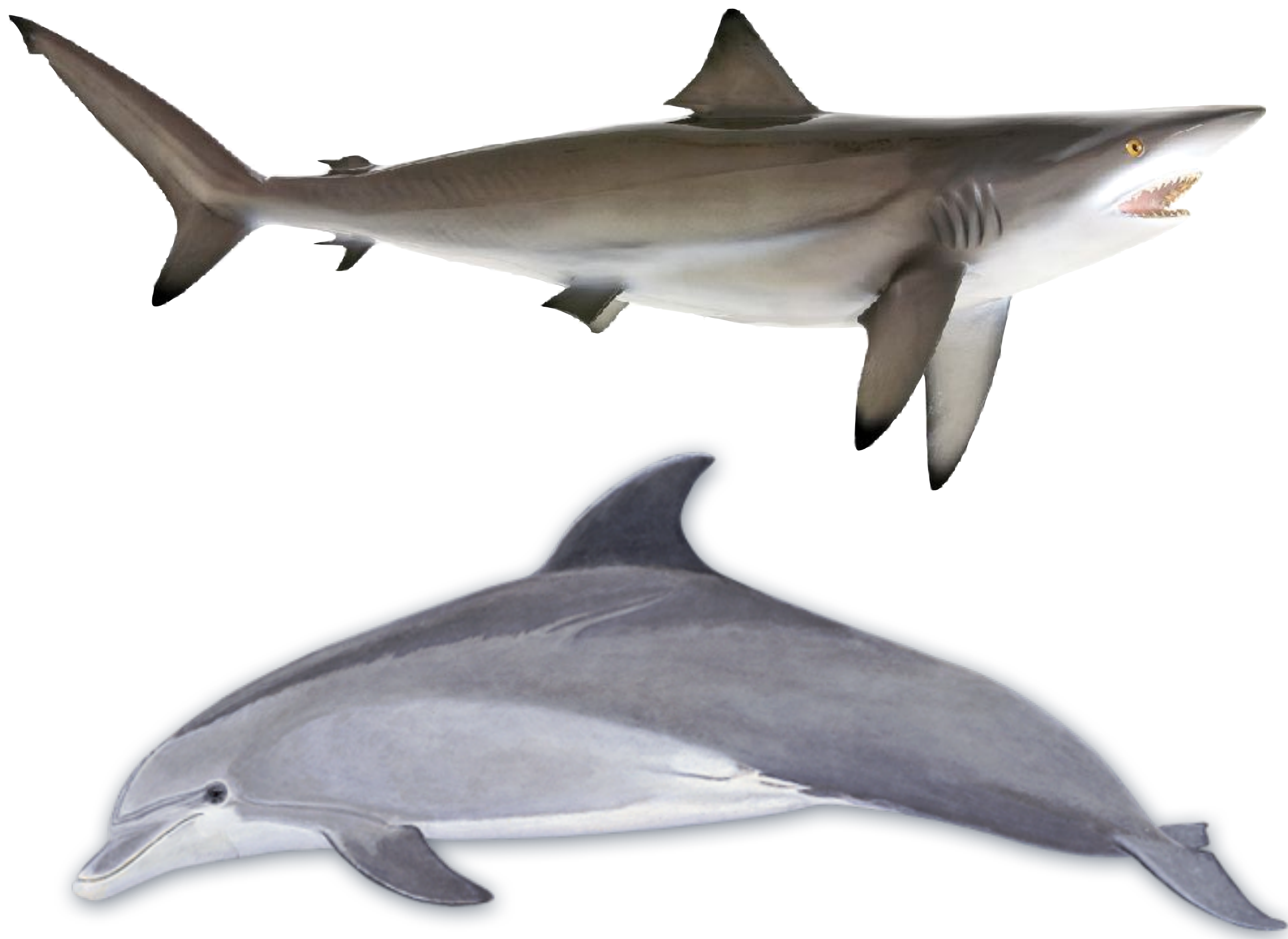
Feature	Value
Color	Gray
Fins	7
Predator	TRUE

Feature Engineering




Feature	Value
Color	Gray
Fins	7
Predator	TRUE

Feature Engineering



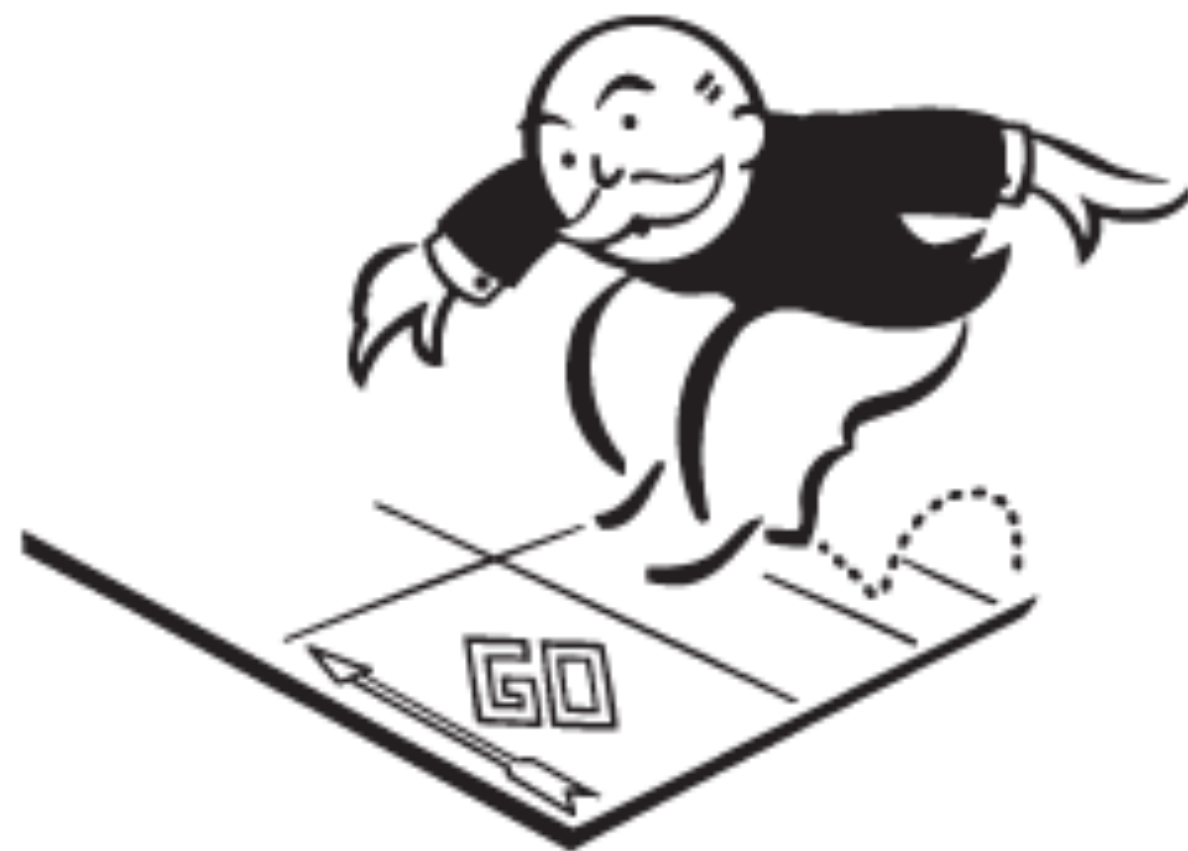
Feature	Value
Color	Gray
Fins	7
Predator	TRUE
Mammal	TRUE

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.
- 
- A decorative orange geometric shape, resembling a stylized 'L' or a corner piece, is located in the bottom right corner of the slide.

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:

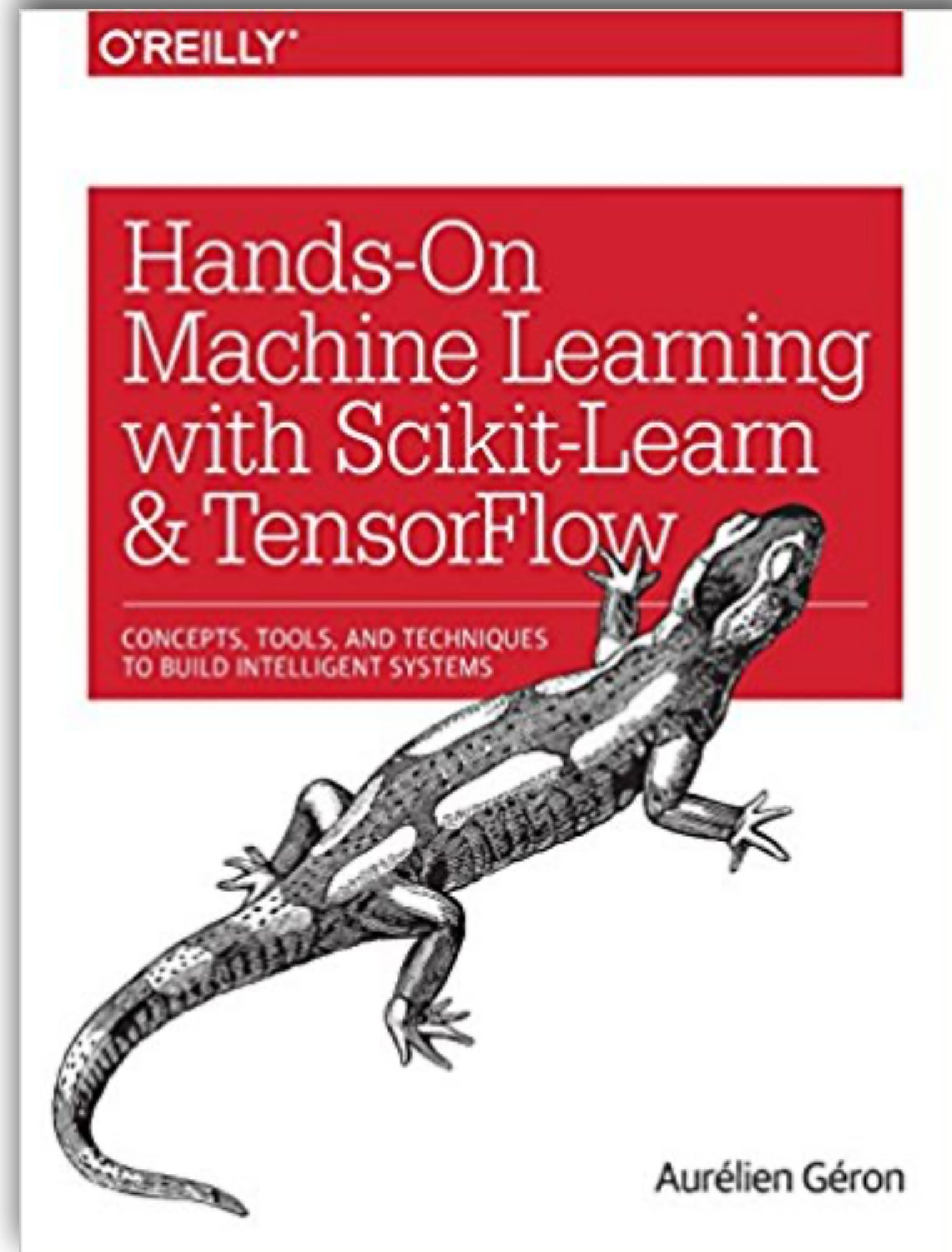
1. What are some features that you would want to capture?
2. What data sets will you need?
3. What legal and policy challenges might you face?
4. What other challenges you could foresee in this problem?
5. How will you define success?
6. How can you articulate the value of this model to stakeholders?

The Python Data Science Ecosystem

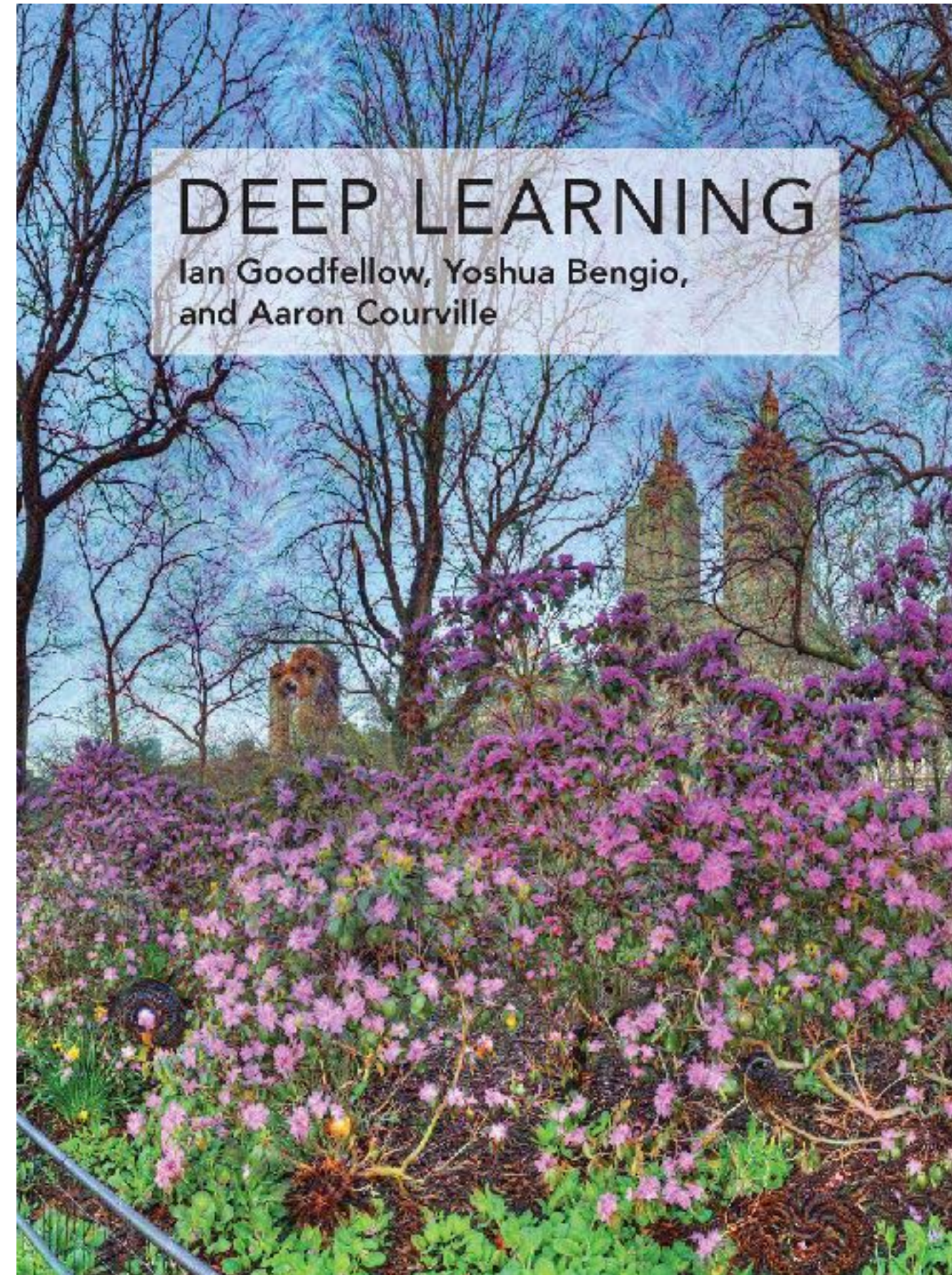
Machine Learning Ecosystem

- **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
- **Visualization:** Matplotlib, Seaborn, Yellowbrick, LIME, ggplot, plot.ly,

Recommended Reading



Recommended Reading



<http://www.deeplearningbook.org/>

O'REILLY

Machine Learning & Security

PROTECTING SYSTEMS WITH DATA AND ALGORITHMS



Clarence Chio & David Freeman

O'REILLY®



Feature Engineering for Machine Learning

PRINCIPLES AND TECHNIQUES FOR DATA SCIENTISTS

Alice Zheng & Amanda Casari

O'REILLY



Learning Apache Drill

QUERY AND ANALYZE STRUCTURED DATA

Charles Givre & Paul Rogers