Machine Learning on Big Data

Part One

Agenda

- Machine Learning Overview
- Model Categories & Types of Output
- Operationalizing Machine Learning
- Threats to Machine Learning

Machine Learning Overview

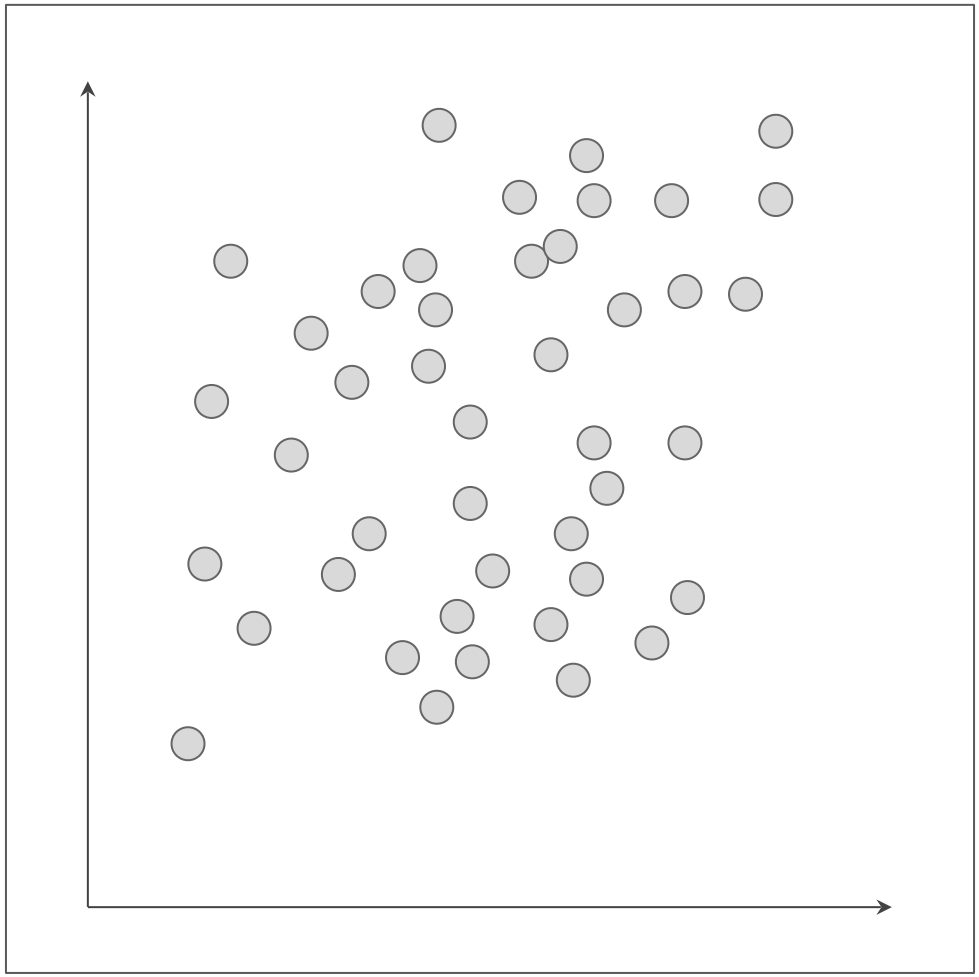
Learning by Example

Given a bunch of examples (data) *extract* a meaningful pattern upon which to *act*.

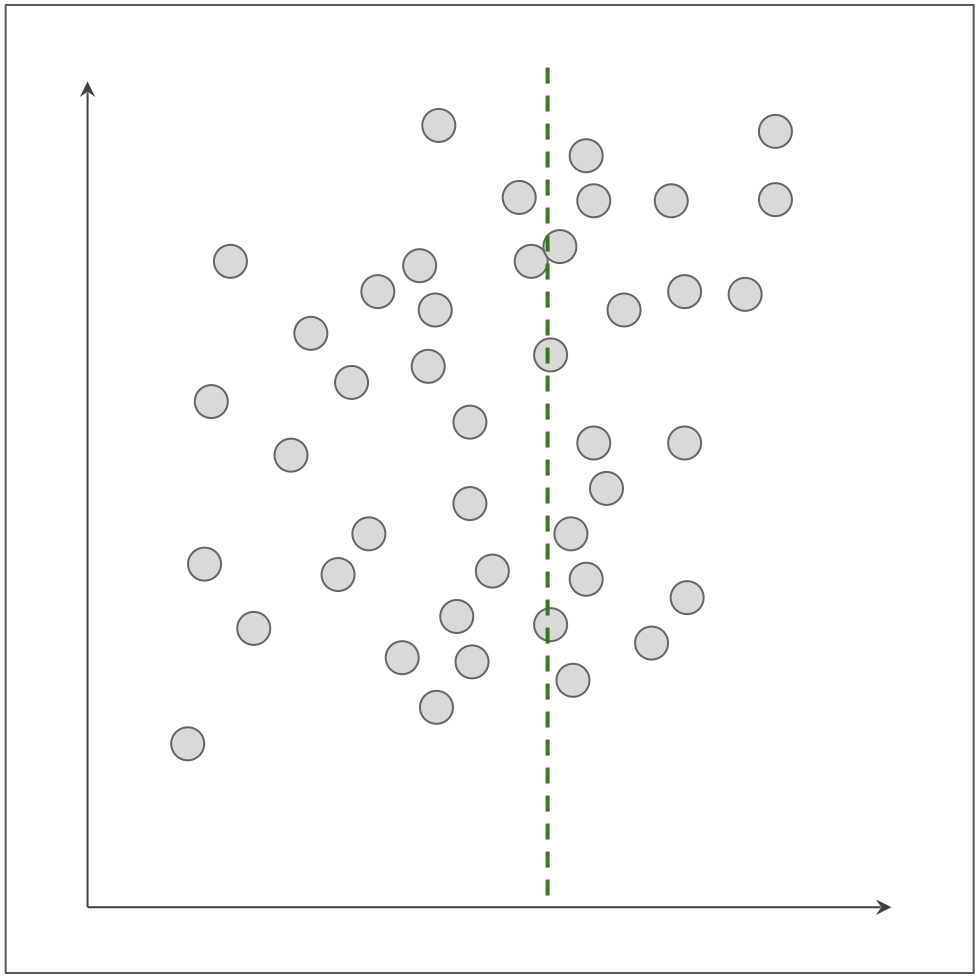
Problem Domain	Machine Learning Class
Infer a function from labeled data	Supervised learning
Find structure of data without feedback	Unsupervised learning
Interact with environment towards goal	Reinforcement learning

How do you make predictions from data?

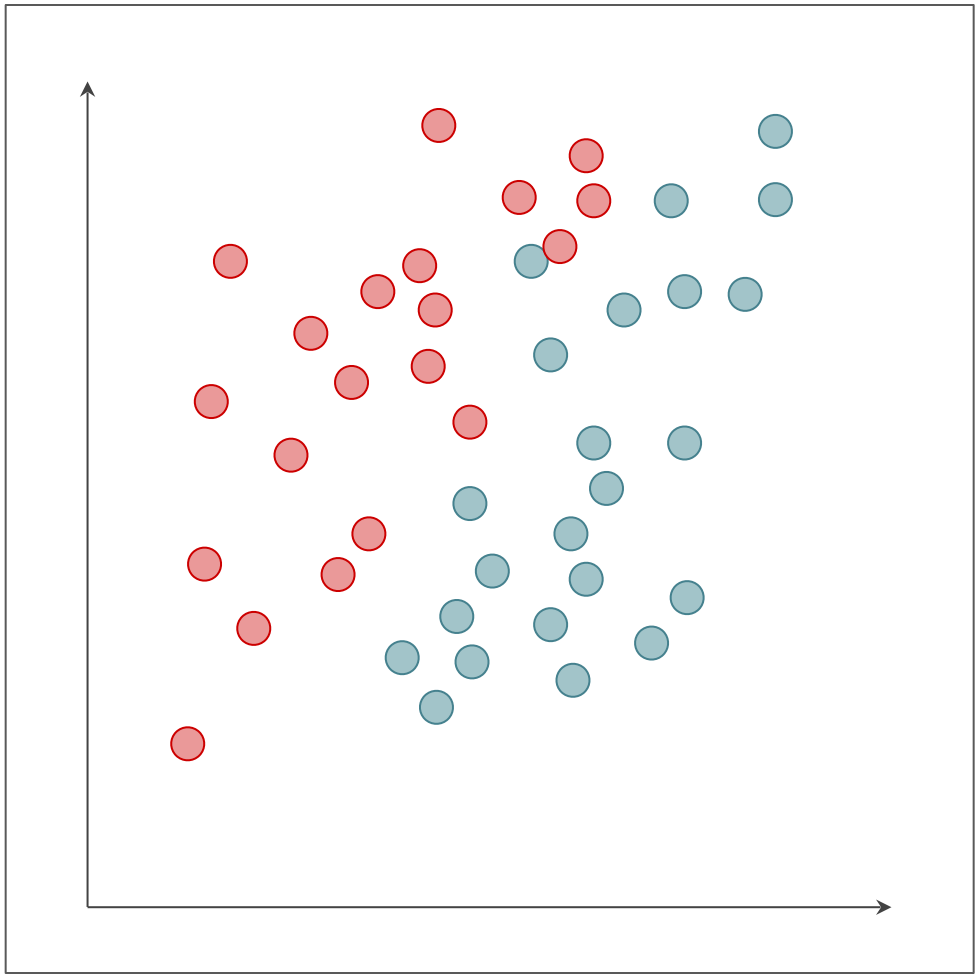
What Patterns Do You See?



What is the Y Value?

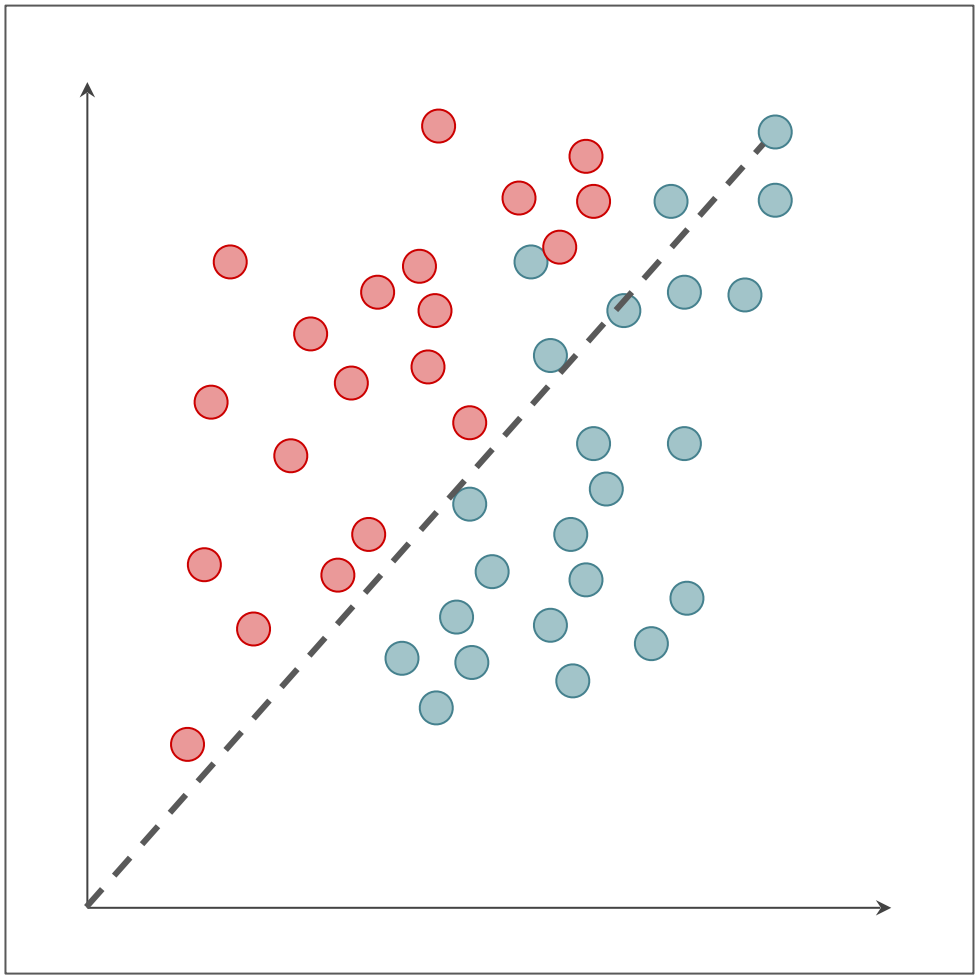


How Do You Determine Red from Blue?

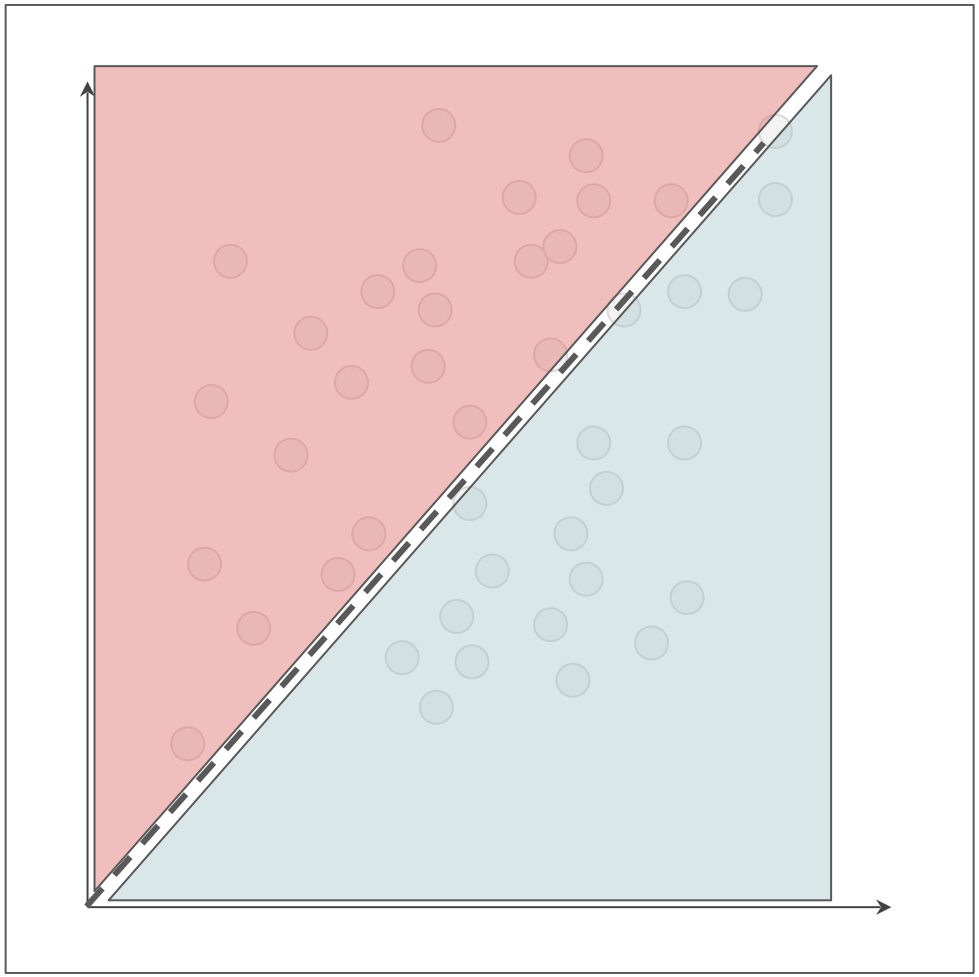


Machine learning efficacy relies on separability and generalizability

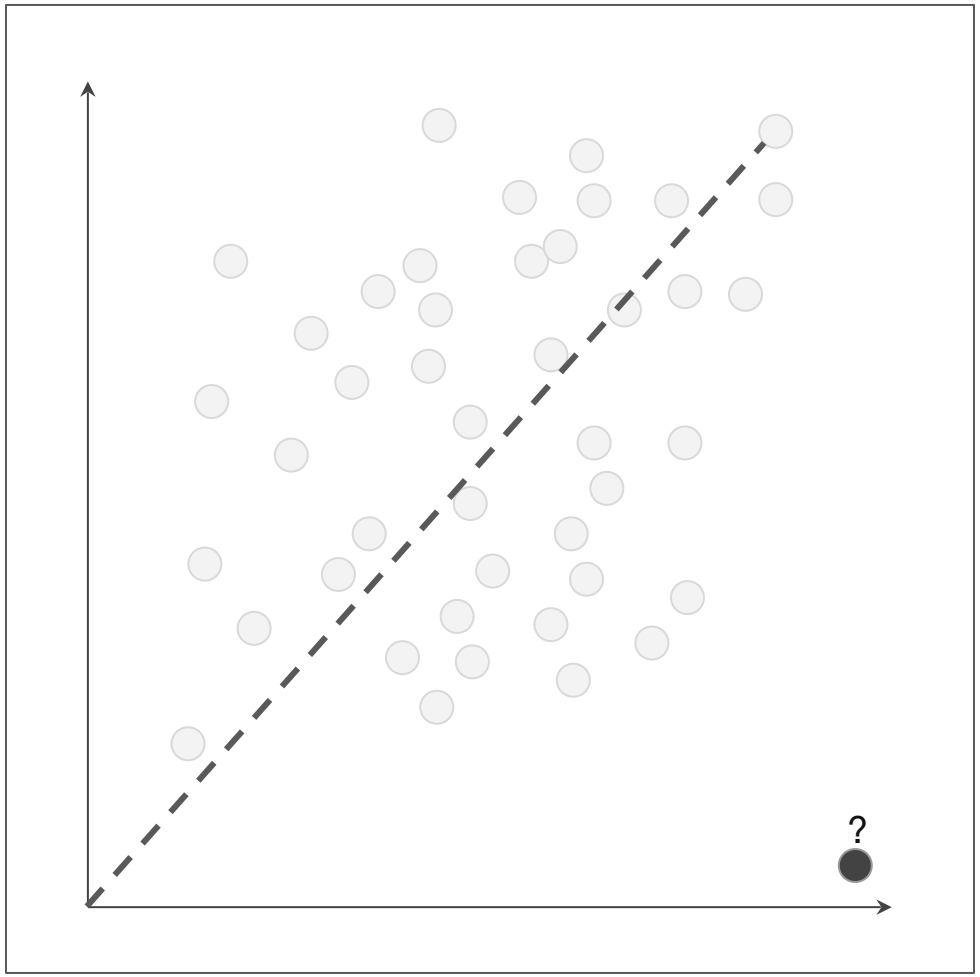
Separability



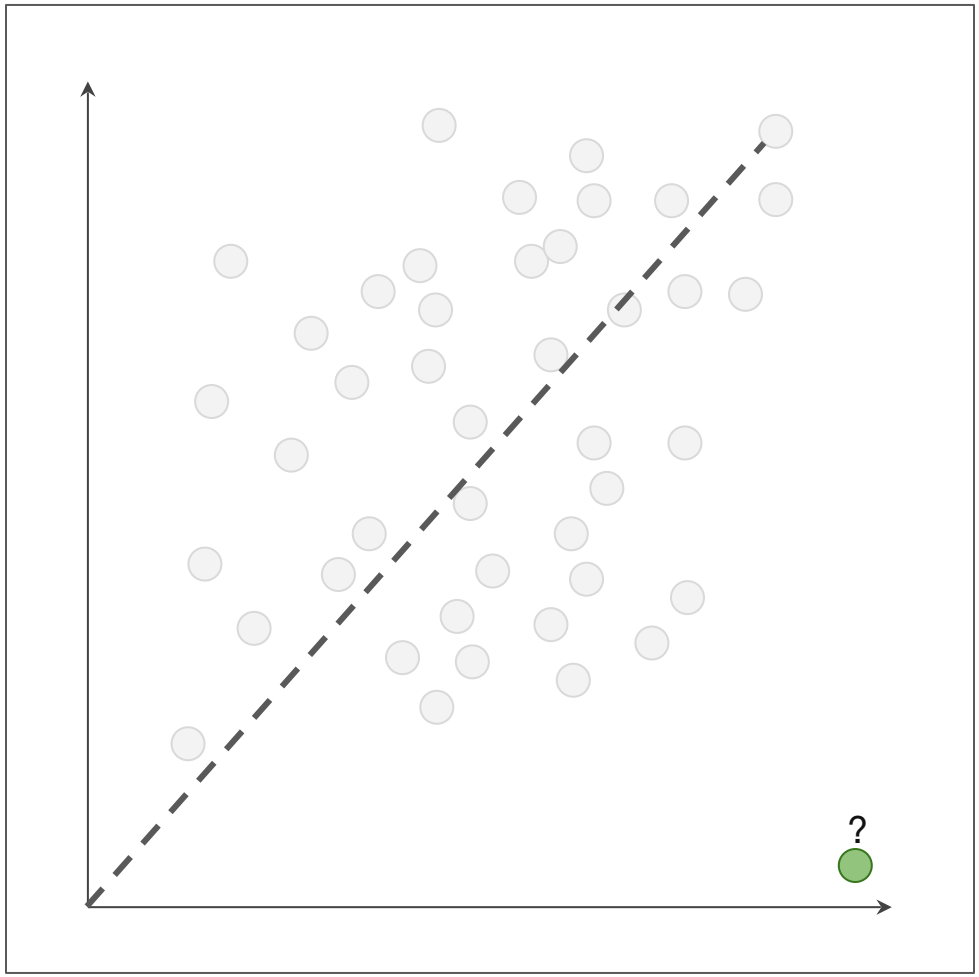
Separability



Generalizability



Generalizability Relative to Training Data



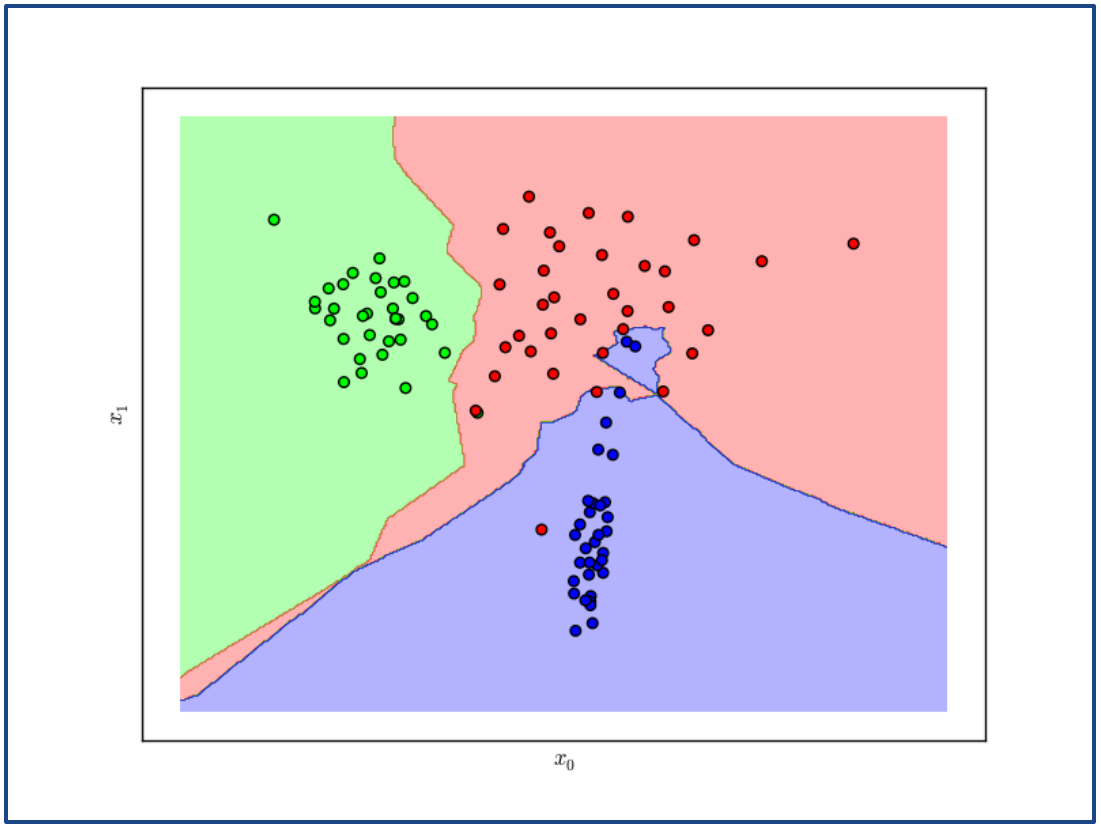
Model Categories & Types of Output

Types of Algorithms by Output

Input *training* data to *fit* a model which is then used to *predict* incoming inputs into ...

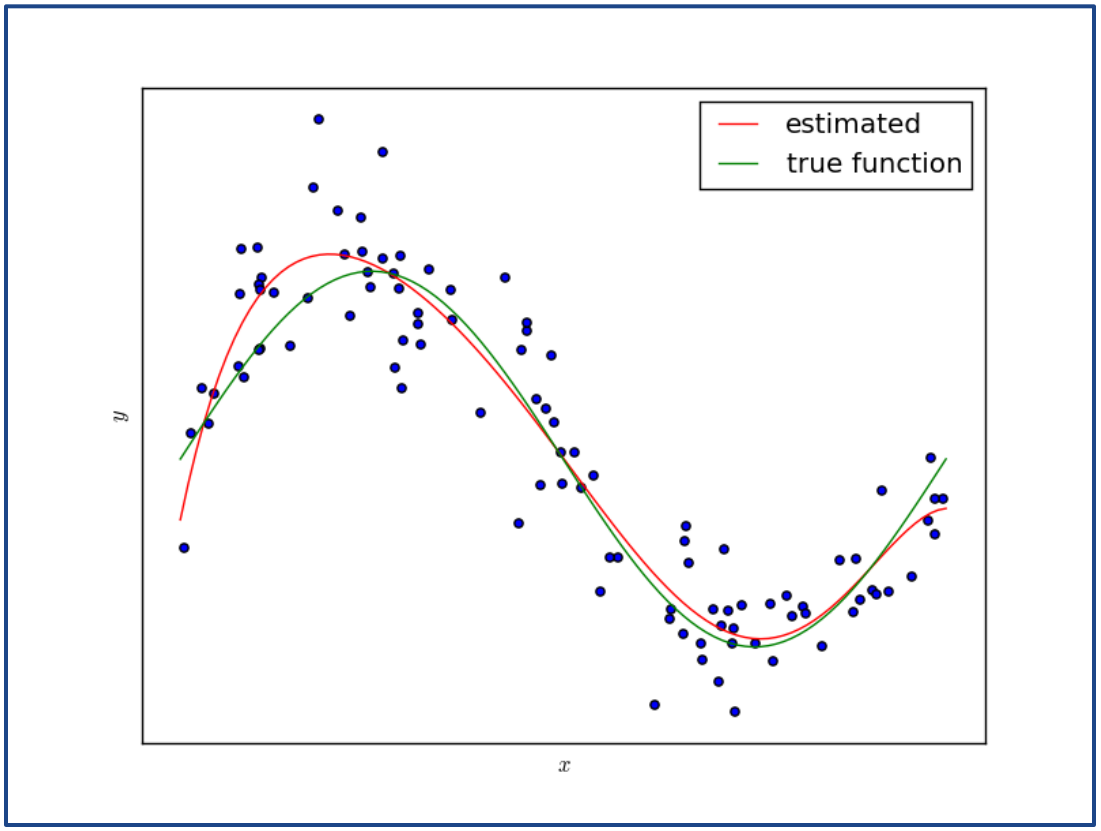
Type of Output	Algorithm Category
Output is one or more discrete classes	**Classification (supervised)**
Output is continuous	**Regression (supervised)**
Output is membership in a similar group	**Clustering (unsupervised)**
Output is the distribution of inputs	Density Estimation
Output is simplified from higher dimensions	Dimensionality Reduction

Classification



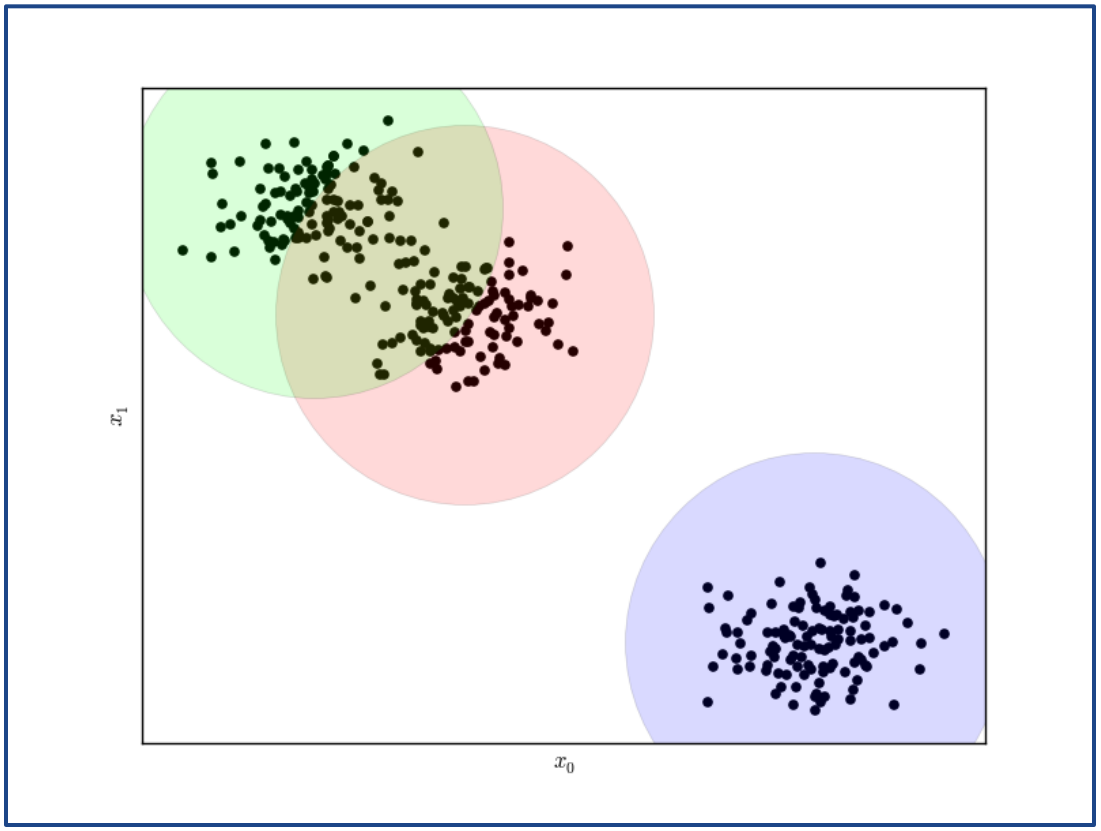
Given labeled input data (with two or more labels), fit a function that can determine for any input, what the label is.

Regression



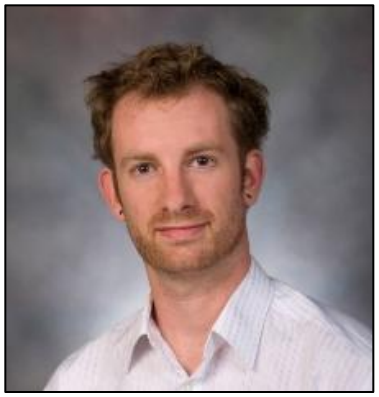
Given continuous input data fit a function that is able to predict the continuous value of input given other data.

Clustering



Given data, determine a pattern of associated data points or clusters via their similarity or distance from one another.

Hadley Wickham (2015)



"Model" is an overloaded term.

- **Model family** describes, at the broadest possible level, the connection between the variables of interest.
- **Model form** specifies exactly how the variables of interest are connected within the framework of the model family.
- A **fitted model** is a concrete instance of the model form where all parameters have been estimated from data, and the model can be used to generate predictions.

http://had.co.nz/stat645/model-vis.pdf

Dimensions and Features

In order to do machine learning you need a data set containing instances (examples) that are composed of features from which you compose dimensions.

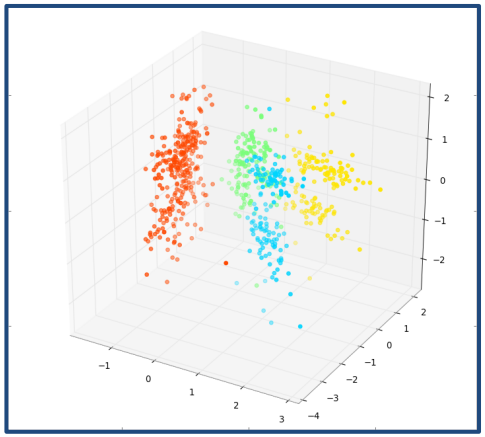
Instance: a single data point or example composed of fields

Feature: a quantity describing an instance

Dimension: one or more attributes that describe a property

Feature Space

Feature space refers to the n-dimensions where your variables live (not including a target variable or class). The term is used often in ML literature because in ML all variables are features (usually) and feature extraction is the art of creating a space with decision boundaries.

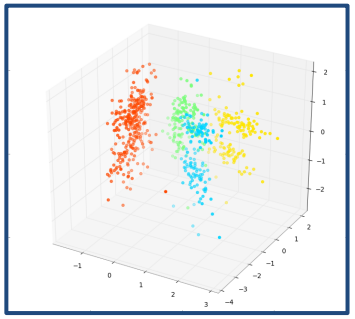


http://stats.stackexchange.com/questions/46425/what-is-feature-space

Feature Space

Target

1. Y ≡ Thickness of car tires after some testing
2. period



Variables

1. X_1 ≡ distance traveled in test
2. X_2 ≡ time duration of test
3. X_3 ≡ amount of chemical C in tires

The feature space is R^3, or more accurately, the positive quadrant in R^3 as all the X variables can only be positive quantities.

Mappings

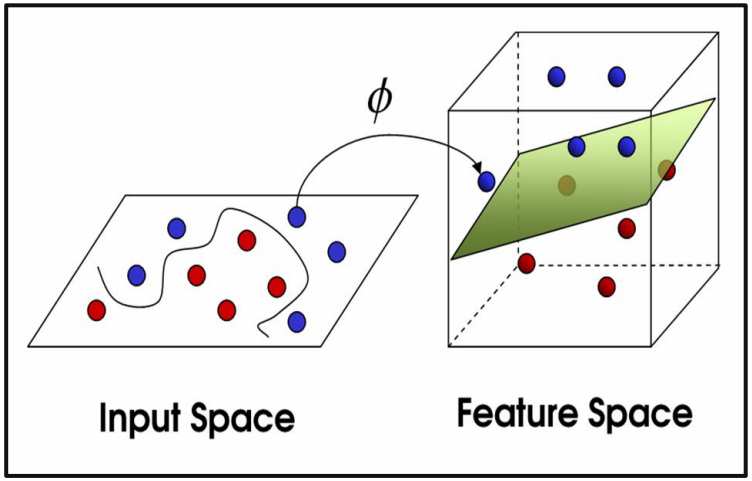
Domain knowledge about tires might suggest that the speed the vehicle was moving at is important, hence we generate another variable, X_4 (this is the feature extraction part):

$X_4 = X_1 * X_2 \equiv$ the speed of the vehicle during testing.

This extends our old feature space into a new one, the positive part of R^4.

A mapping is a function, ϕ, from R^3 to R^4:

$\phi(x_1,x_2,x_3) = (x_1,x_2,x_3,x_1x_2)$



http://stats.stackexchange.com/questions/46425/what-is-feature-space

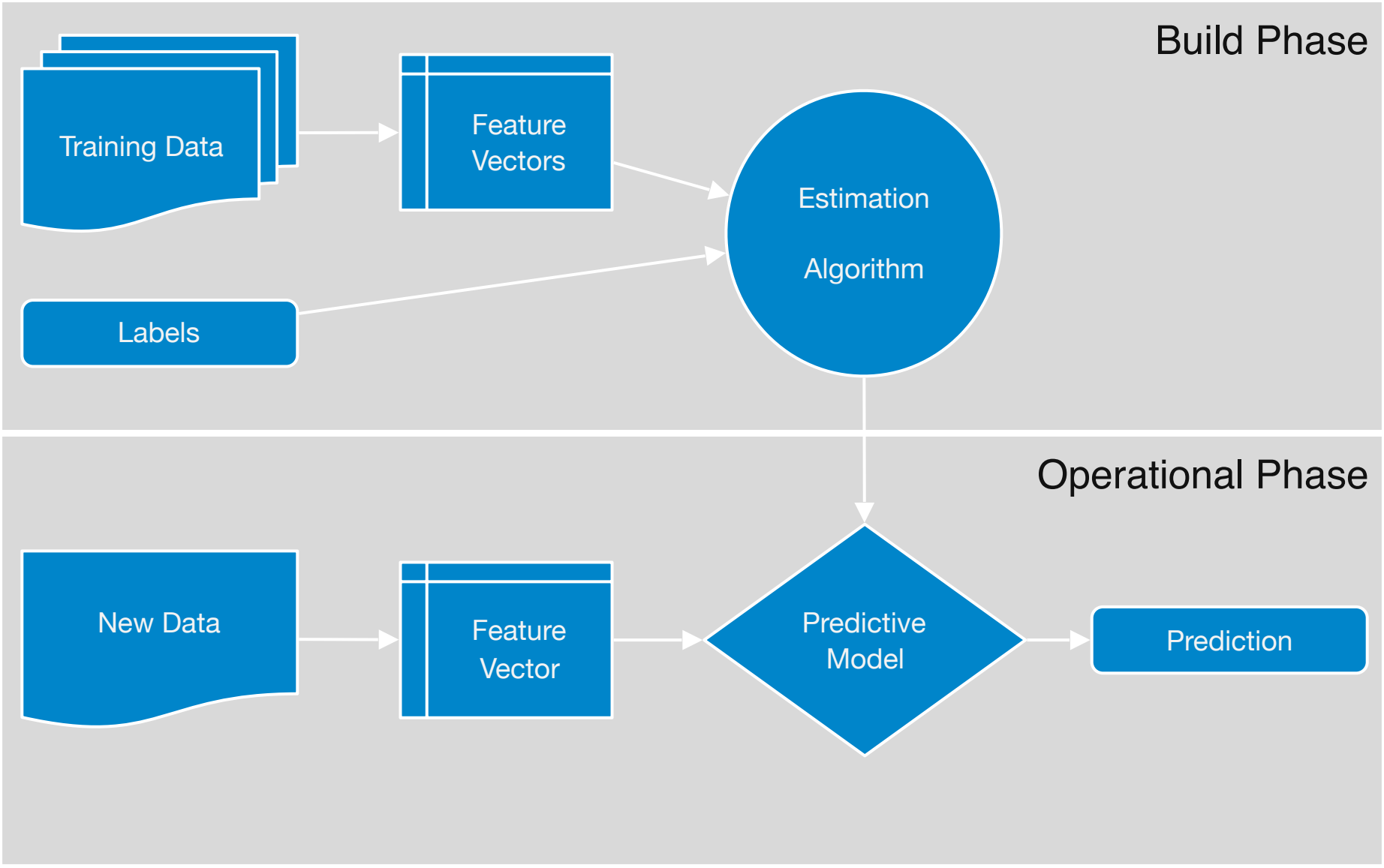
Your Task

Given a data set of instances of size N, create a model that is fit from the data (built) by extracting features and dimensions. Then use that model to predict outcomes …

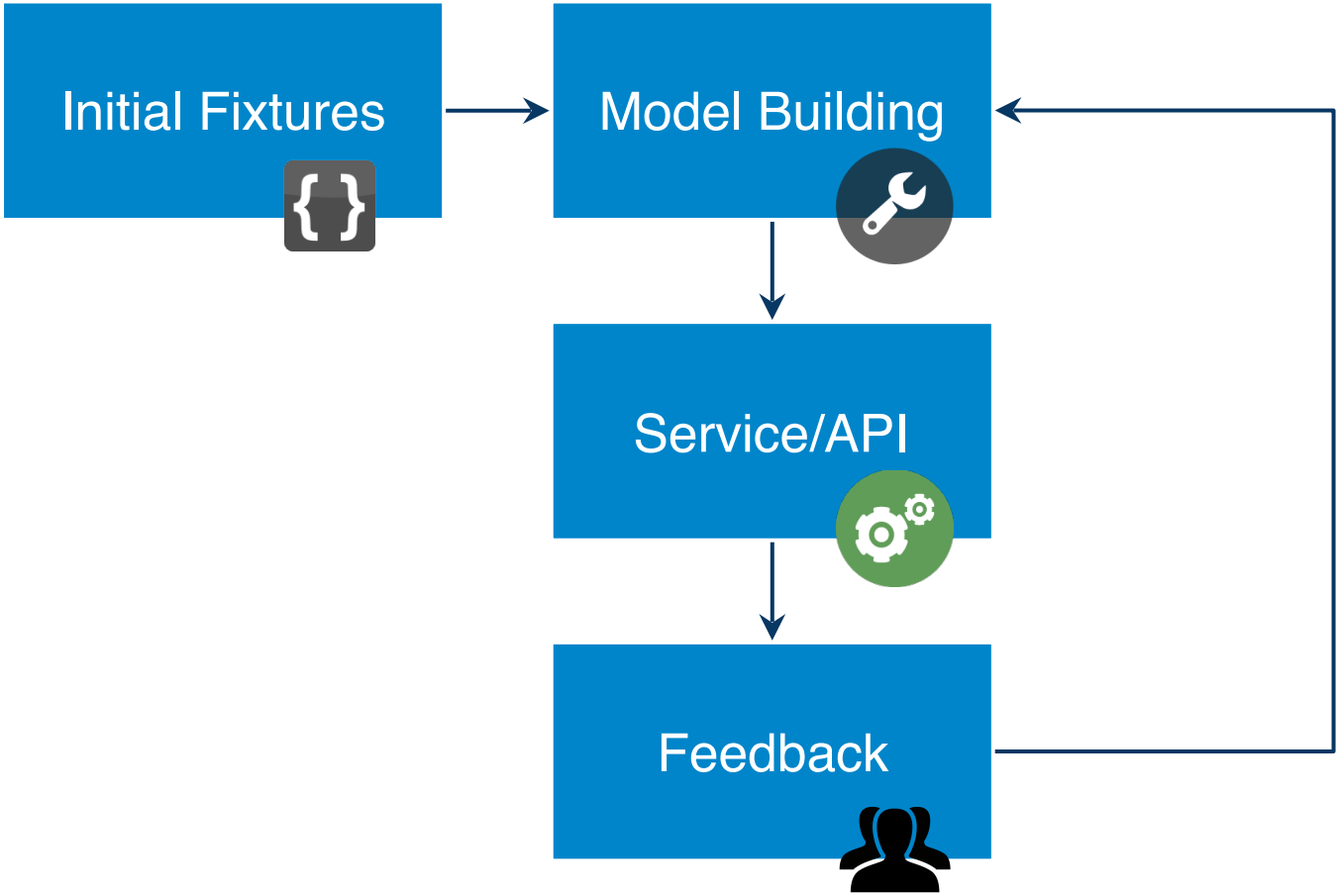
1. Data Wrangling (normalization, standardization, imputing)
2. Feature Analysis/Extraction
3. Model Selection/Building
4. Model Evaluation
5. Operationalize Model

Operationalizing Machine Learning

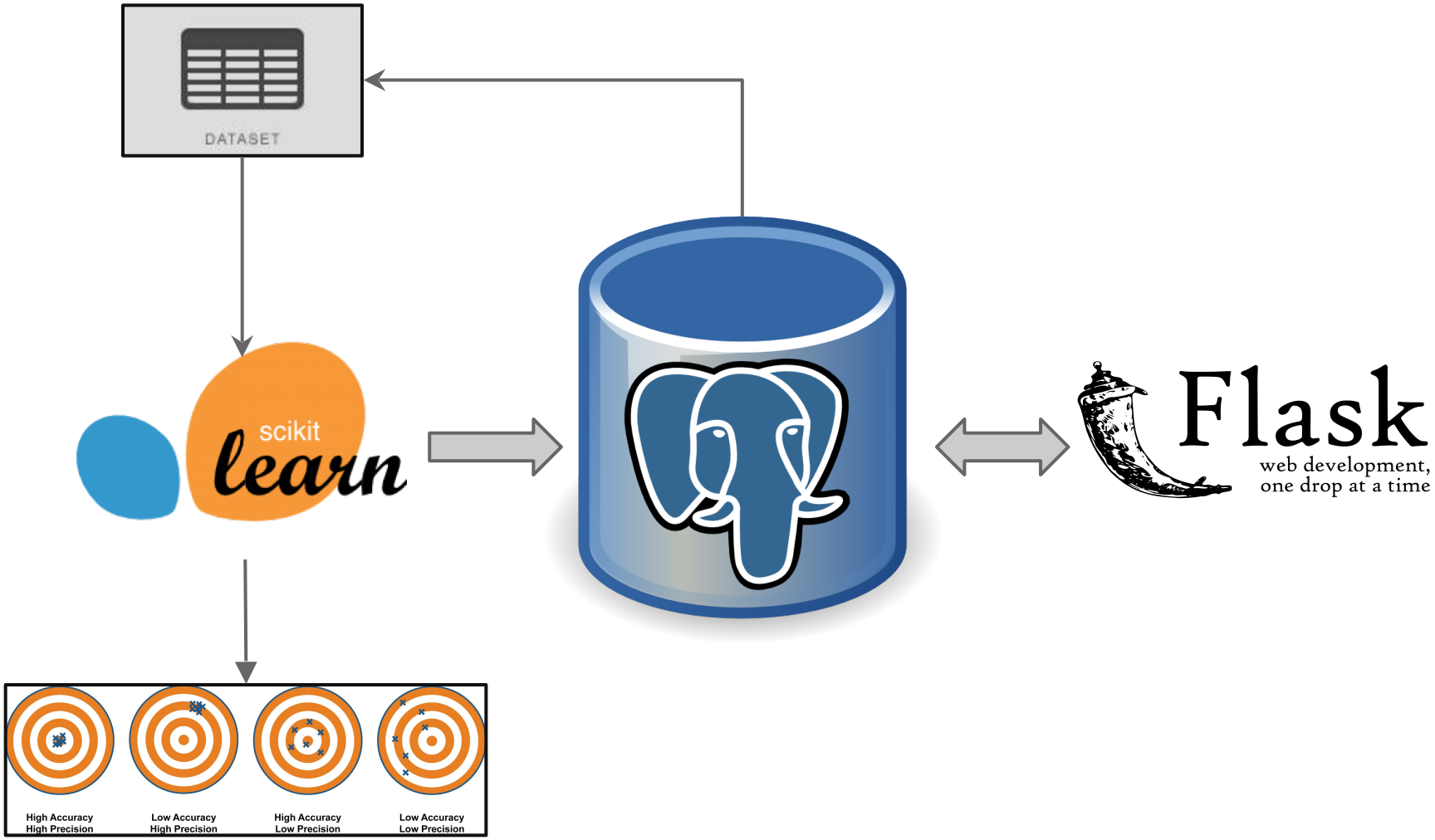
The Machine Learning Lifecycle



The Learning Part of Machine Learning

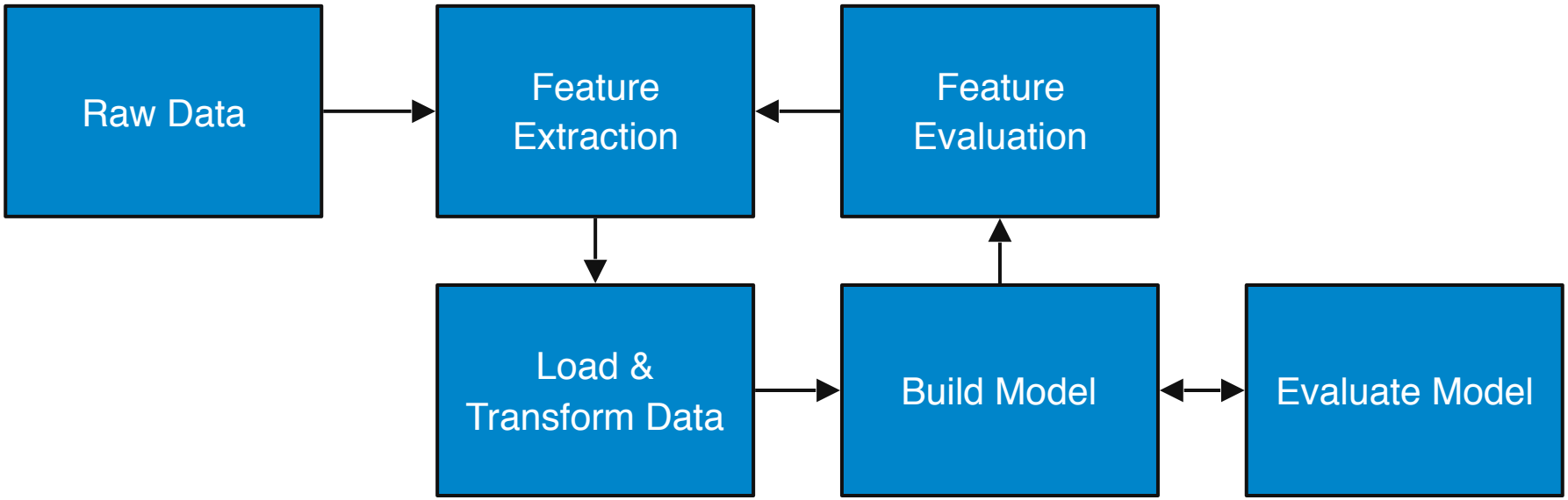


Simple Deployment



Development Operations

Development follows a similar pattern to the operational life cycle, but focuses more on *feature engineering*.



The Model Selection Triple

Selecting models to use in practice is associated with the model selection triple:

- **Feature Engineering**
- **Algorithm Selection**
- **Parameter Tuning**

Selection can be automated (using cross-validation for evaluation, and search for parameter tuning and feature optimization) or human steered. The best approach is usually a combination.

http://pages.cs.wisc.edu/~arun/vision/

Feature Engineering

The process of using domain knowledge of the data to create features that make machine learning algorithms work.

Done manually it can be difficult and expensive.

Automated feature learning obviates the necessity for manual feature engineering.

Feature learning allows a system to automatically discover the representations needed for feature detection or classification from raw data.

Algorithm Selection

Look at the mindmap

Parameter Tuning

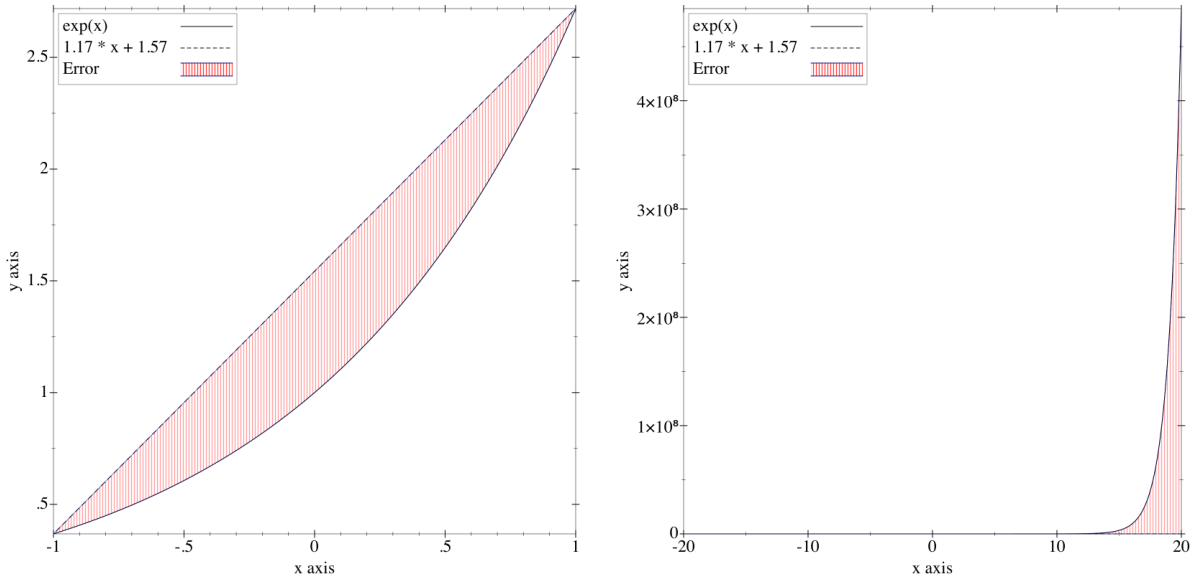
Hyperparameter: parameters whose values are set prior to the commencement of the learning process; values of other parameters are derived via training.

Methods: grid search, random search (both in scikit-learn).

Threats to Machine Learning

Underfitting

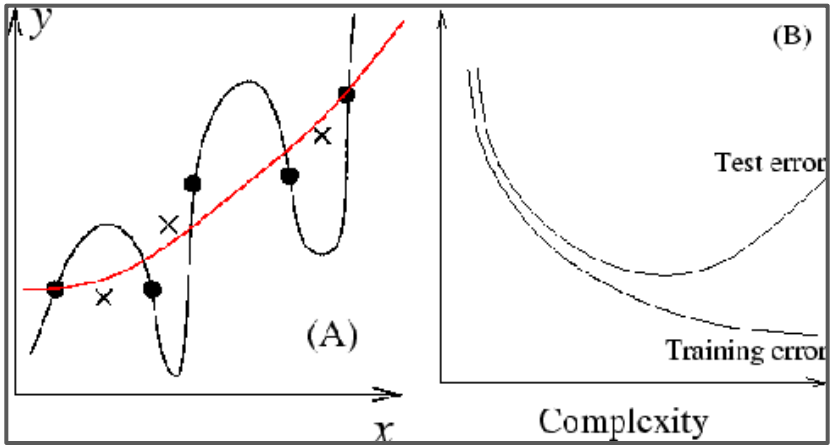
Not enough information to accurately model real life. Can be due to high bias, or just a too simplistic model.



Solution: Cross Validation

Overfitting

Create a model with too many parameters or is too complex. “Memorization of the data” - and the model can’t generalize very well.

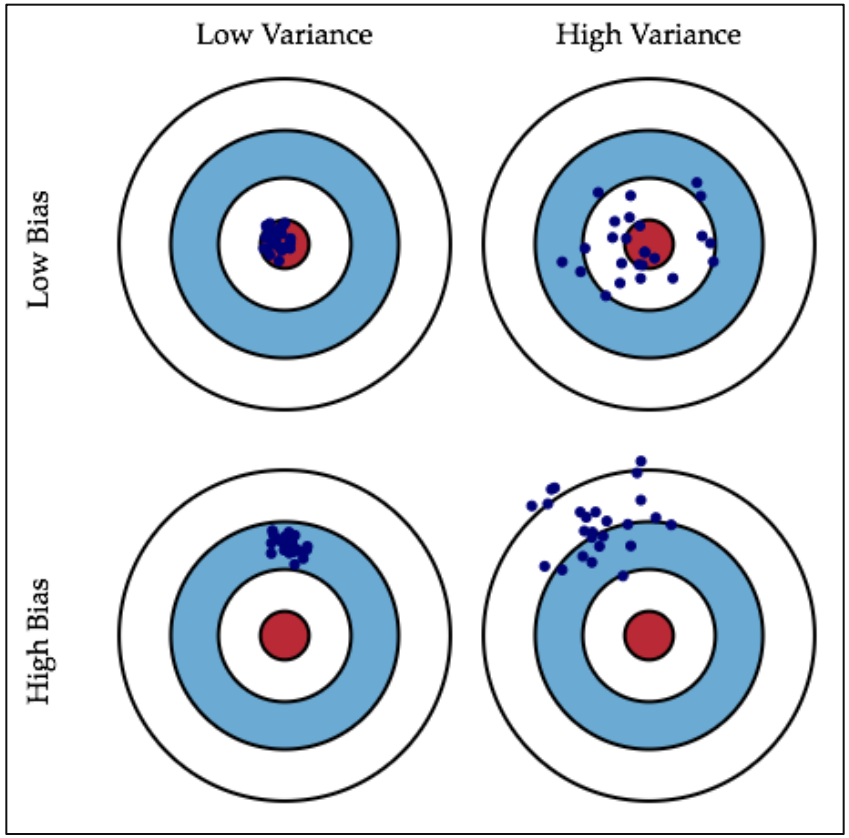


Solution: Benchmark Testing, Ridge Regression, Feature Analyses, Dimensionality Reduction

Error: Bias vs Variance

Bias: the difference between expected (average) prediction of the model and the correct value.

Variance: how the predictions for a given point vary between different realizations for the model.

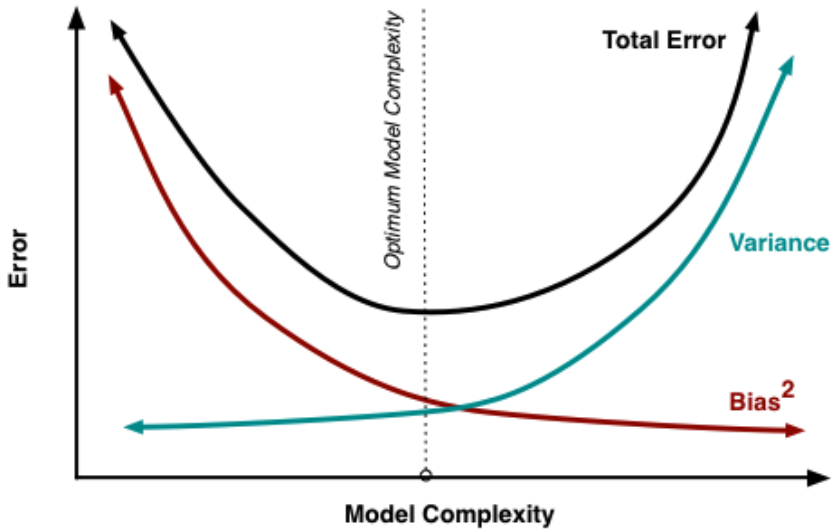


http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias vs. Variance Trade-Off

Related to model complexity:

The more parameters added to the model (the more complex), Bias is reduced, and variance increased.



Sources of complexity:

- k (nearest neighbors)
- epochs (neural nets)
- # of features
- learning rate

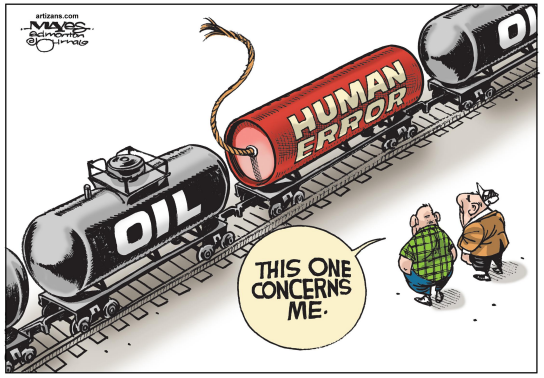
http://scott.fortmann-roe.com/docs/BiasVariance.html

Unstable Data

Randomness is a significant part of data in the real world but problems with data can significantly affect results:

- outliers
- skew
- missing information
- incorrect data

Solution: seam testing/integration testing



Unpredictable Future

Machine learning models attempt to predict the future as new inputs come in - but human systems and processes are subject to change.

Solution: Precision/Recall tracking over time

Hands-On Lab

Task: Case Study

Design a machine learning system that answers the question.

1. What problem are you trying to solve?
2. What kind of ML problem is it?
    1. What is your target variable?
    2. What are your independent variables?
3. Who will be affected by the results of your analysis?
4. What are the technical and non-technical risks?
5. Who do you need on your team to build the system?
6. What does the ultimate result look like?