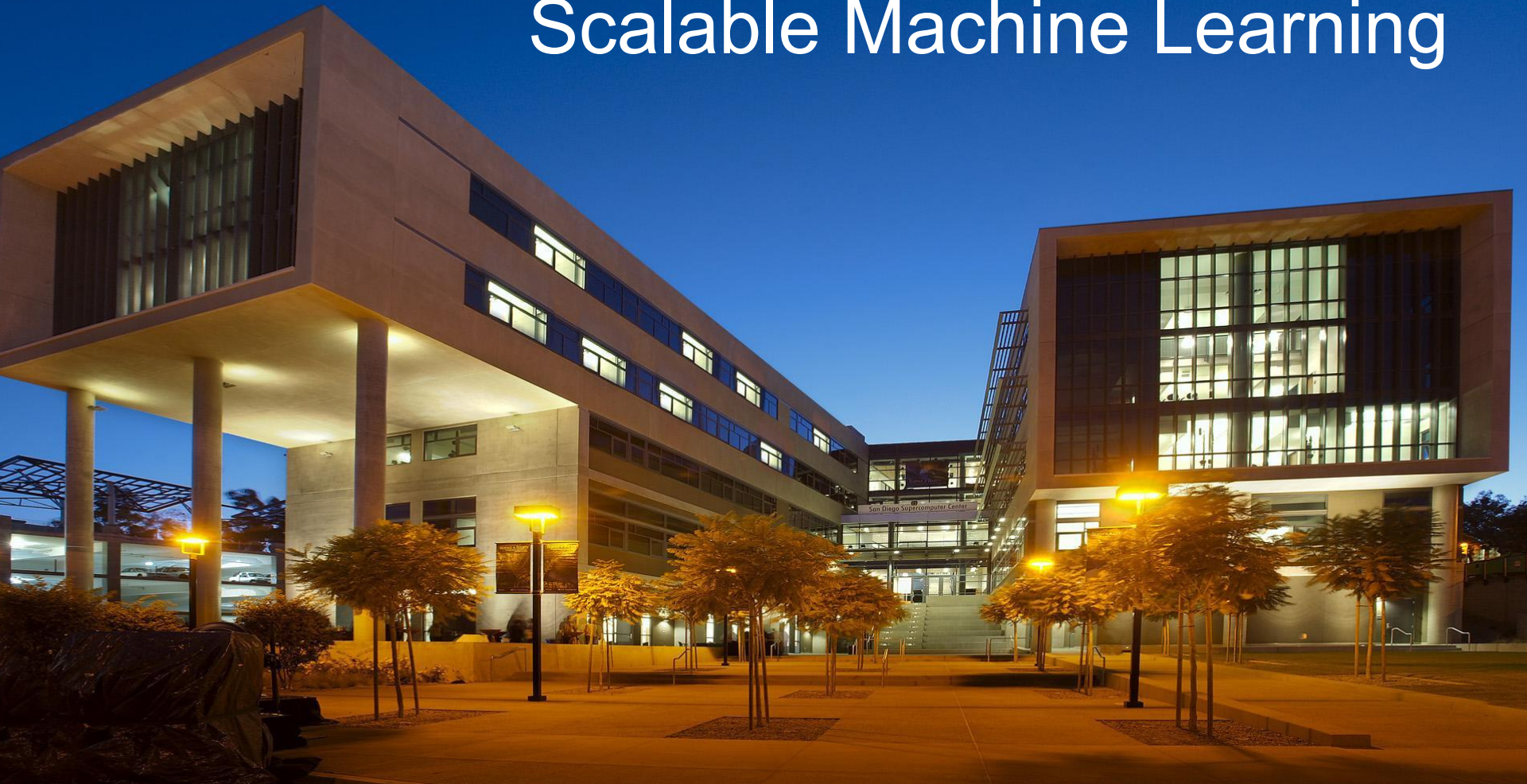


CIML Summer Institute 2022

Scalable Machine Learning



Scalable Machine Learning Agenda

- 8:00 - 8:05 – Welcome**
- 8:05 - 9:05 – Introduction to Singularity**
- 9:05 - 10:50 – CONDA & Jupyter on Expanse**
- 10:50 - 11:20 – Break/Lunch**
- 11:20 - 11:35 – Machine Learning Overview**
- 11:35 - 12:15 – R on HPC**
- 12:15 - 12:25 – Break**
- 12:25 - 2:05 – Spark**

Machine Learning Overview

Mai H. Nguyen, Ph.D.

What is Machine Learning?

- **Machine learning is ...**
 - “... a subfield of computer science that ... explores the study and construction of algorithms that can learn from and make predictions on data.” ([wikipedia.org](https://en.wikipedia.org))
 - “... a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed.” (whatis.techtarget.com)
 - “... a method of data analysis that automates analytical model building and ... allows computers to find hidden insights to produce ... predictions that can guide better decisions and smart actions...” (www.sas.com)

What is Machine Learning?

learning from data

no explicit programming

discover hidden patterns

data-driven decisions

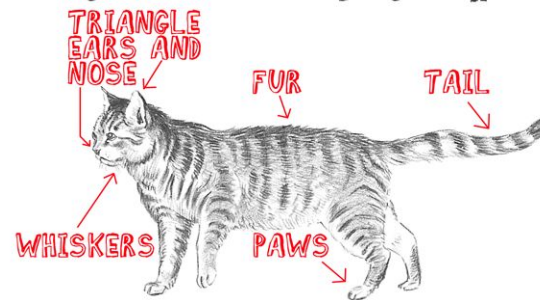
What is Machine Learning?

learning from data

no explicit programming



What Characteristics Do Cats Have



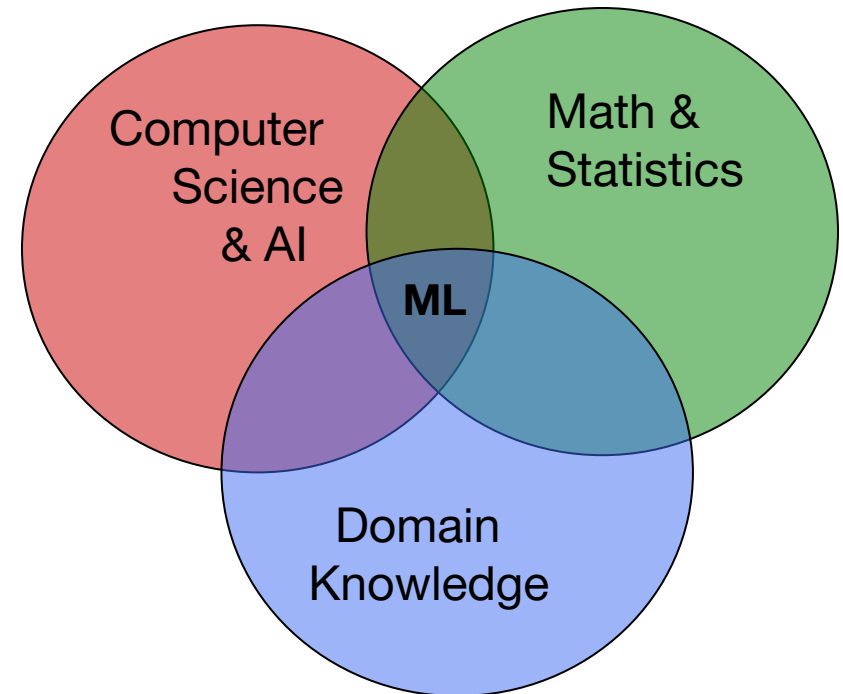
What is Machine Learning?

- **Working Definition**

- The field of machine learning focuses on the study and construction of computer systems that can learn from data without being explicitly programmed. Machine learning algorithms and techniques are used to build models to discover hidden patterns and trends in the data, allowing for data-driven decisions to be made.

Machine Learning as Interdisciplinary Field

- **ML combines concepts & methods from many disciplines:**
 - Mathematics, statistics, computer science, artificial intelligence, etc.
- **ML is being used in various fields:**
 - Science, engineering, business, medical, law enforcement, etc.



Why the Increased Interest in ML?

- **Advances in processing power, storage capacity, mobile computing, and interconnectivity are creating unprecedented data:**
 - User preferences and purchasing history on websites
 - Scientific data from remote sensors and instruments
 - Personal health data from wearable devices
 - Medical data from drug trials, treatment options, patient population
 - Social media data related to customer satisfaction, political trends, health epidemics, law enforcement, terrorist activities

MACHINE LEARNING APPLICATIONS

Best Sellers based on your browsing history



Apple AirPods with Charging Case (Wired)
★★★★☆ 153,701
\$129.00



Apple AirPods Pro
★★★★★ 54,773
\$219.00



Apple EarPods with Lightning Connector - White
★★★★★ 38,539
\$19.98



Apple AirPods with Wireless Charging Case
★★★★☆ 24,208
\$159.99



TOZO T10 Bluetooth 5.0 Wireless Earbuds with Wireless Charging Case IPX8 Waterproof TWS...
★★★★☆ 107,951
\$29.98

Inspired by your browsing history



AirPods Case Cover with Keychain, Full Protective Silicone AirPods Accessories Skin Cover...
★★★★☆ 18,919
\$5.99



Apple Watch Series 3 (GPS, 38mm) - Space Gray Aluminum Case with Black Sport Band
★★★★★ 49,269
\$169.00



AirPods Case, GMYLE Silicone Protective Shockproof Case Cover Skins with Keychain...
★★★★☆ 15,592
\$5.98



Apple 5W USB Power Adapter
★★★★☆ 3,627
\$16.99



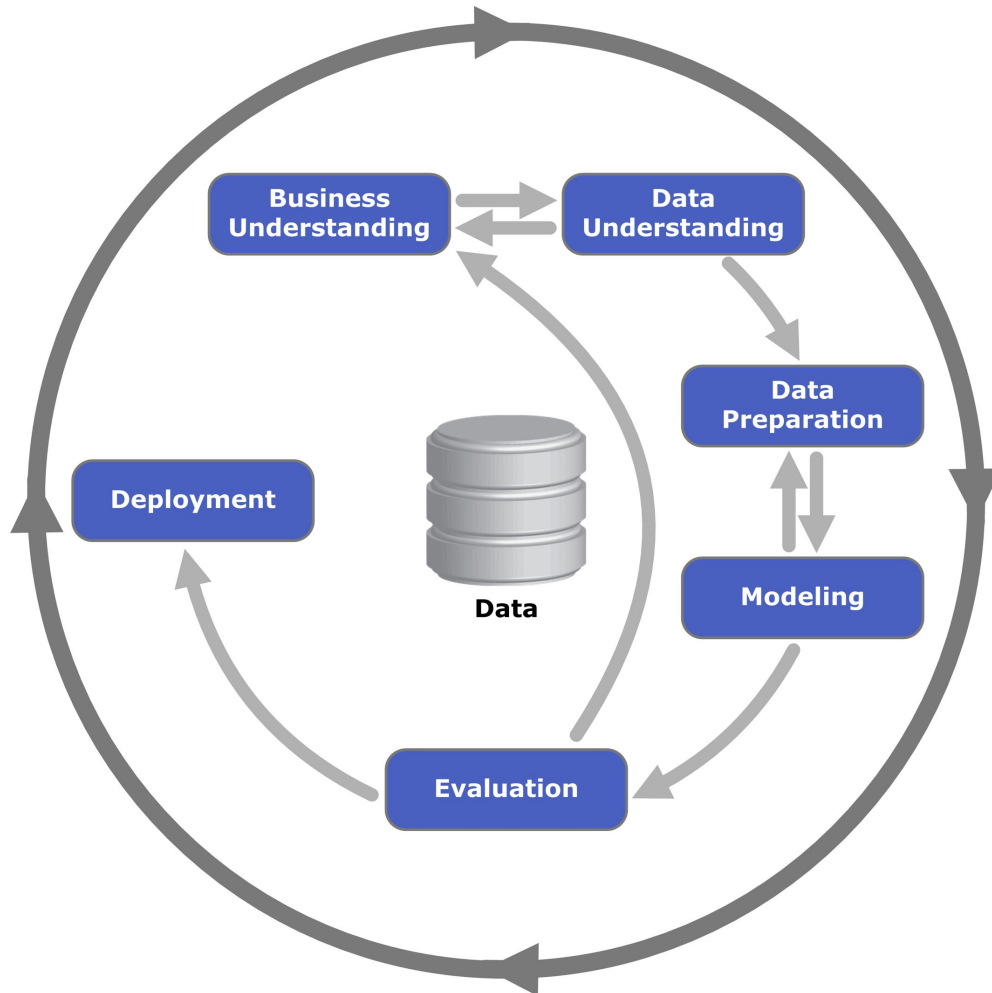
AmazonBasics Premium AirPods Case - Compatible with Apple AirPods 1 & 2, Pink
★★★★☆ 78
\$6.77



SENTIMENT ANALYSIS



MACHINE LEARNING PROCESS



**CRoss Industry
Standard Process
for Data Mining**

https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Phase 1: Business Understanding

- **Define problem or opportunity**
 - What is the problem of interest? Why is it interesting?
- **Assess situation**
 - Resources
 - Requirements, assumptions, and constraints
 - Risks and contingencies; costs and benefits
- **Formulate goals and objectives**
 - Goals and objectives
 - Success criteria
- **Create project plan**
 - Steps to achieve goals

Phase 2: Data Understanding

- **Data Acquisition**

- Collect available data related to problem
- Consider all sources: flat files, databases, sensors, websites, etc.
- Integrate data from multiple sources

- **Exploratory Data Analysis**

- Preliminary exploration of data
- To become familiar with data



<http://www.greenbookblog.org/2013/08/04/50-new-tools-democratizing-data-analysis-visualization/>

Phase 3: Data Preparation

- **Goal:**

- Prepare data to make it suitable for modeling
- Also referred to as 'data preprocessing', 'data munging', 'data wrangling'

- **Activities:**

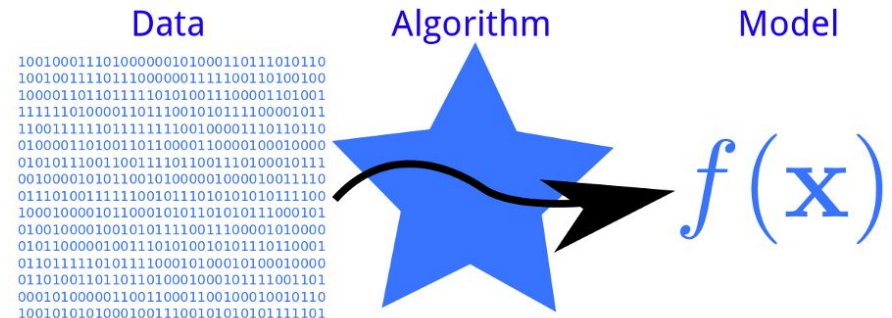
- Identify and address quality issues
- Select features to use
- Create data for modeling



<http://www.datasciencecentral.com/profiles/blogs/5-data-cleansing-tools>

Phase 4: Modeling

- **Determine type of problem**
 - Classification
 - Regression
 - Cluster analysis
- **Build model(s)**
 - Select modeling technique(s) to use
 - Construct model(s)
 - Train model(s)



<http://phdp.github.io/posts/2013-07-05-dtl.html>

Phase 5: Evaluation

- **Assess model performance**

- Determine metrics & methods to assess model results
 - Accuracy measures, confusion matrix, etc.
- Evaluate model results w.r.t. success criteria
 - Does model's performance meet success criteria?
 - Have all requirements been met?

- **Make Go/No-Go decision**

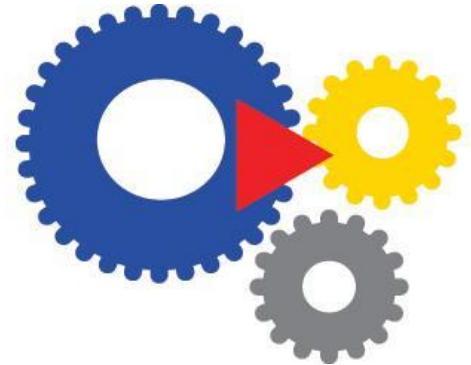
- Go: Deploy model
- No-Go: Determine next steps



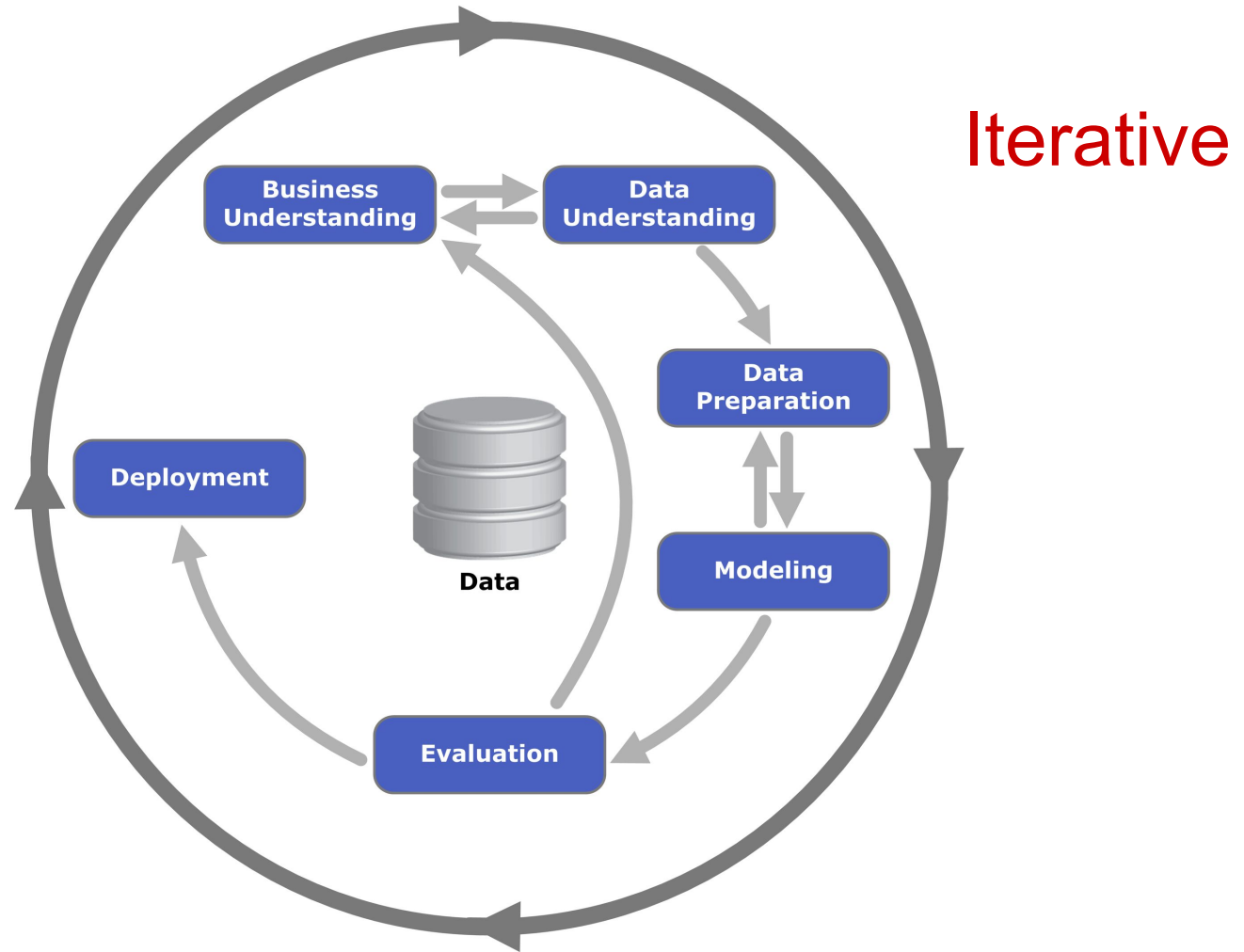
<http://www.impactptac.com/?id=10>

Phase 6: Deployment

- **Documentation**
 - Summarize findings and recommend uses
 - Document code, create user's guide, etc.
- **Packaging**
 - Modularize code
 - Containerize code
- **Model deployment**
 - Integrate model into decision-making process in production
 - Inference serving
- **Model monitoring & maintenance**
 - Monitor model performance
 - Plan for updating/correcting model
- **Versioning**
 - code, model, data, environment, configuration, etc.

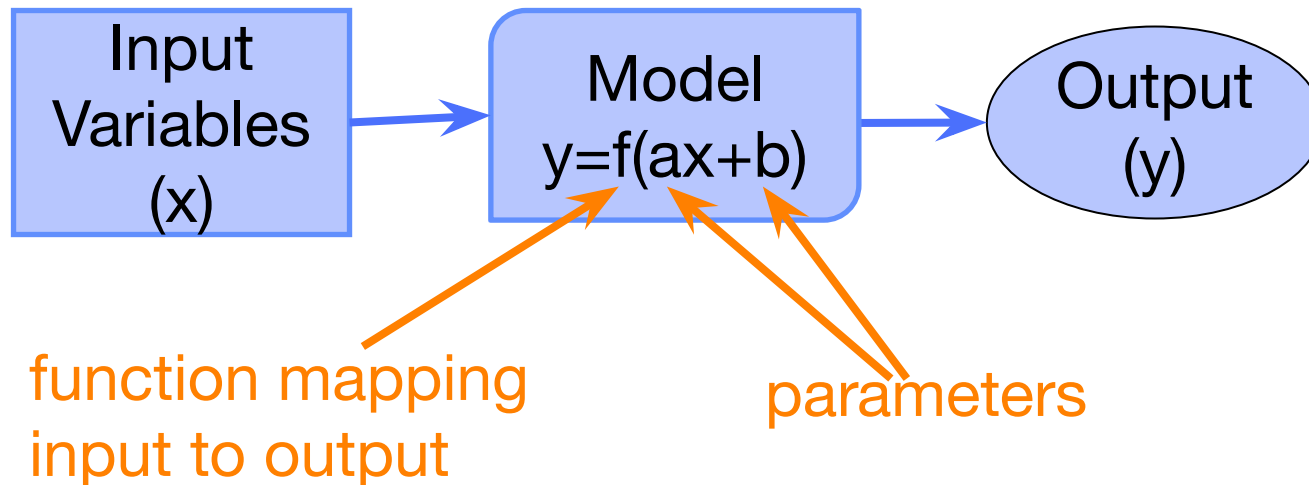


Machine Learning Process

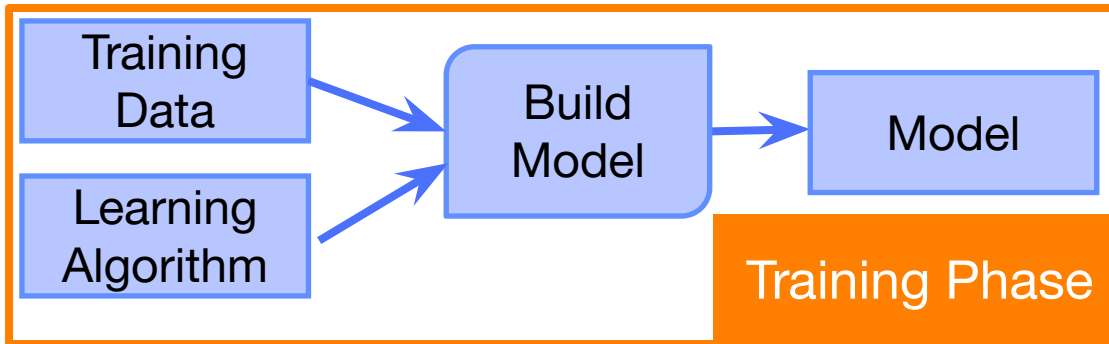


MACHINE LEARNING MODEL

- ML model = mathematical model with parameters that maps input to output
- Model parameters are adjusted during model training to change input-output mapping
- Parameters are learned or estimated from data
 - “fitting the model”, “training the model”, “building the model”
- Goal: Minimize some error function

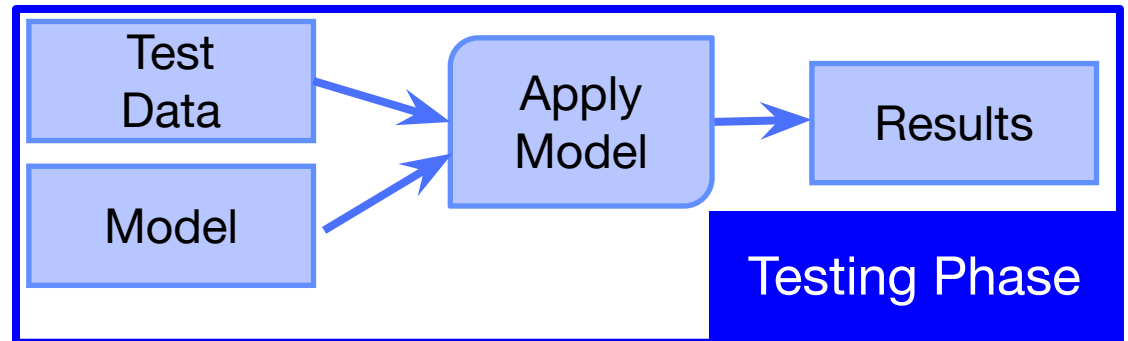


BUILDING VS APPLYING MODEL

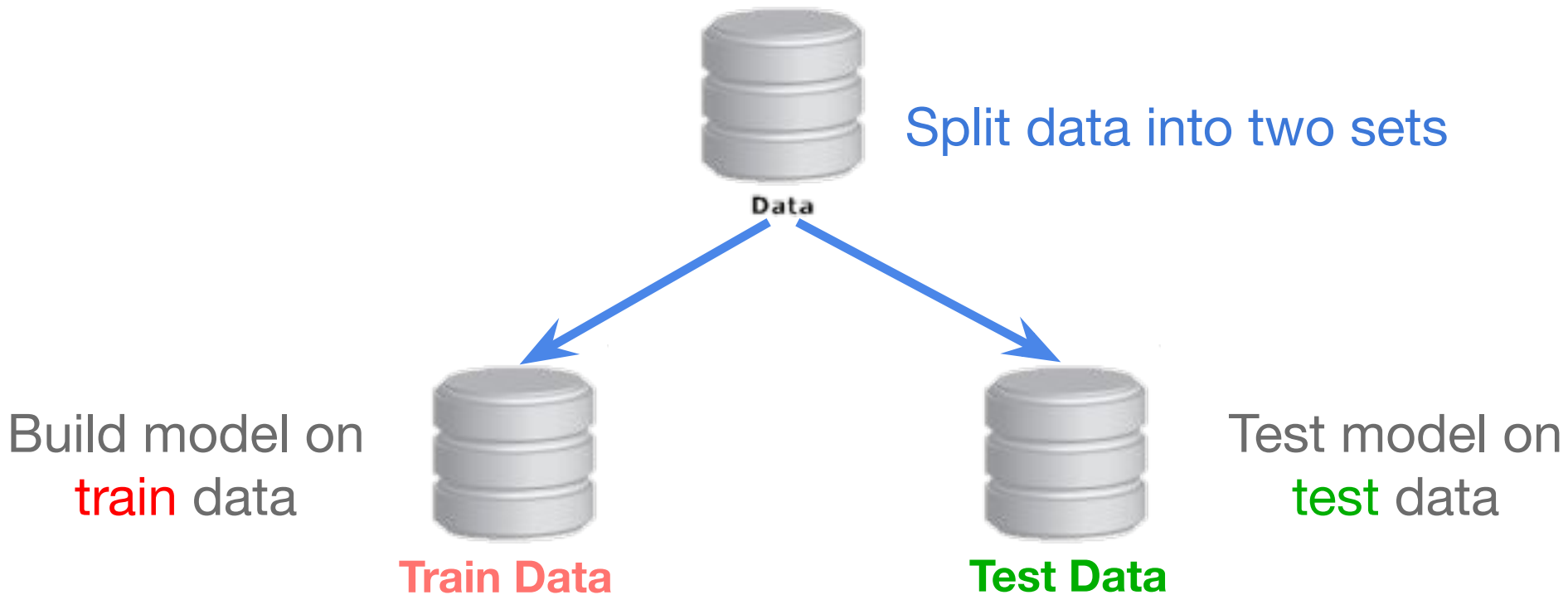


Adjust model
parameters
“Train”

Test model on
new data
“Inference”



GENERALIZATION

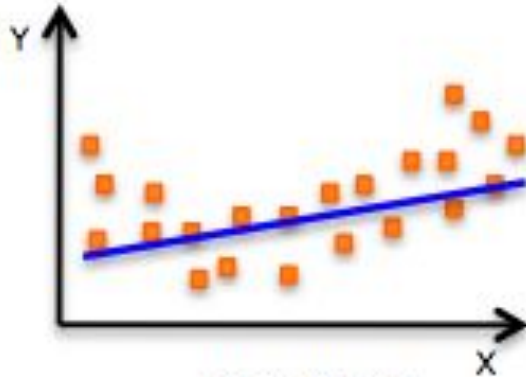


Goal: Want model to perform well on data it was not trained on, i.e., to **generalize** well to unseen data

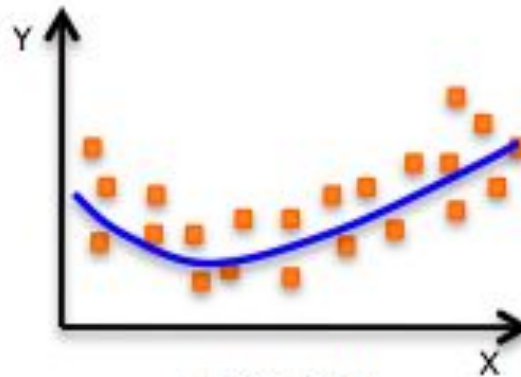
OVERFITTING & GENERALIZATION

- **Overfitting**
 - Model is fitting to noise in data instead of to underlying distribution of data
- **Reasons for overfitting**
 - Training set is too small
 - Model is too complex, i.e., has too many parameters
- **Overfitting leads to poor generalization**
 - Model that overfits will not generalize well to new data

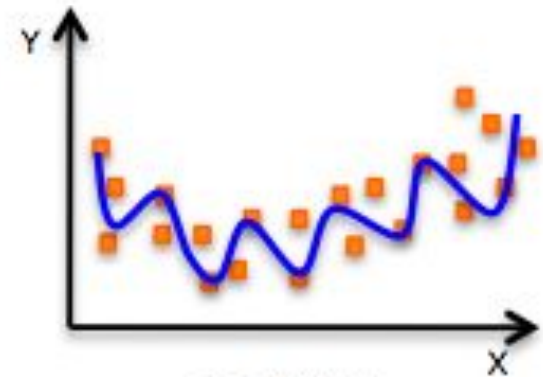
OVERFITTING



Underfitting



Just right!



overfitting

<http://stats.stackexchange.com/questions/192007/what-measures-you-look-at-the-determine-over-fitting-in-linear-regression>

Underfitting

Model has not learned
structure of data

High training error
High test error

Just Right

Model has learned
distribution of data

Low training error
Low test error

Overfitting

Model is fitting to
noise in data

Low training error
High test error

ADDRESSING OVERFITTING

- **Model complexity**
 - Number of parameters in model
 - Chance of overfitting increases with model complexity
- **Validation set**
 - Monitor error on training and validation data
 - To determine when to stop training
- **Regularization**
 - Constrain or shrink (“regularize”) model parameters
 - Add penalty term to error function used to train model
 - e.g., Add L1-norm and/or L2-norm regularization to linear regression model

Scalable Machine Learning

- What is scalable machine learning?
- Applying machine learning to 'big data'



https://infocus.emc.com/scott_burgess/15350/

Big Data

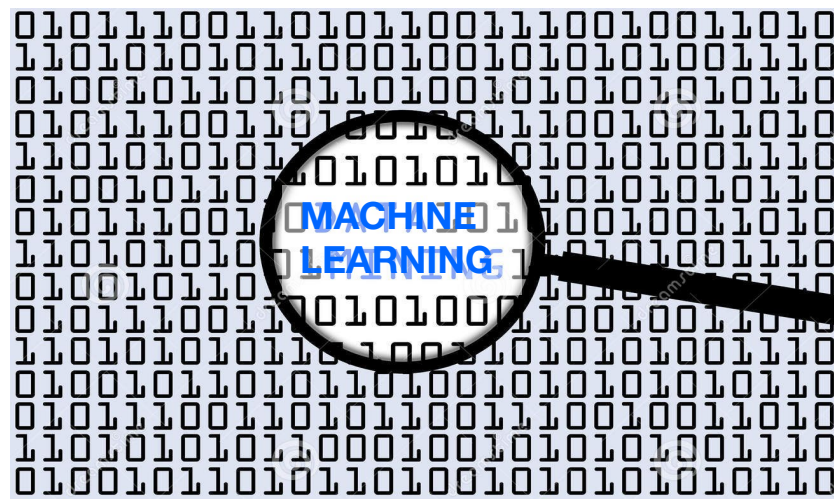
- **V's of Big Data (Doug Laney of Gartner)**
 - **Volume**
 - Vast amounts of data being generated
 - Petabytes (10^{15} bytes), exabytes (10^{18} bytes), and even more
 - **Velocity**
 - Speed at which data is being generated
 - Data is being generated continuously
 - **Variety**
 - Different forms of data
 - Numeric, text, images, voice, geospatial, etc.
 - **Veracity**
 - Quality of data

Fifth 'V' of Big Data: Value

- **Goal of processing Big Data is to extract value from data**
 - Fifth 'V' of Big Data: Value
- **Not sufficient to collect Big Data**
- **Need to analyze data to gain insights for decision-making**

Scalable Machine Learning

- **Extracting value is at the heart of analyzing any data**
 - This is done using machine learning
- **New technologies and approaches needed to address challenges (the V's) of Big Data**
 - Parallel processing
 - Scalable algorithms
 - Distributed platforms



Download from
Dreamstime.com

This watermarked comp image is for previewing purposes only.

<http://www.dreamstime.com/stock-photos-data-mining-image35154223>

35154223

Alain Lacroix | Dreamstime.com