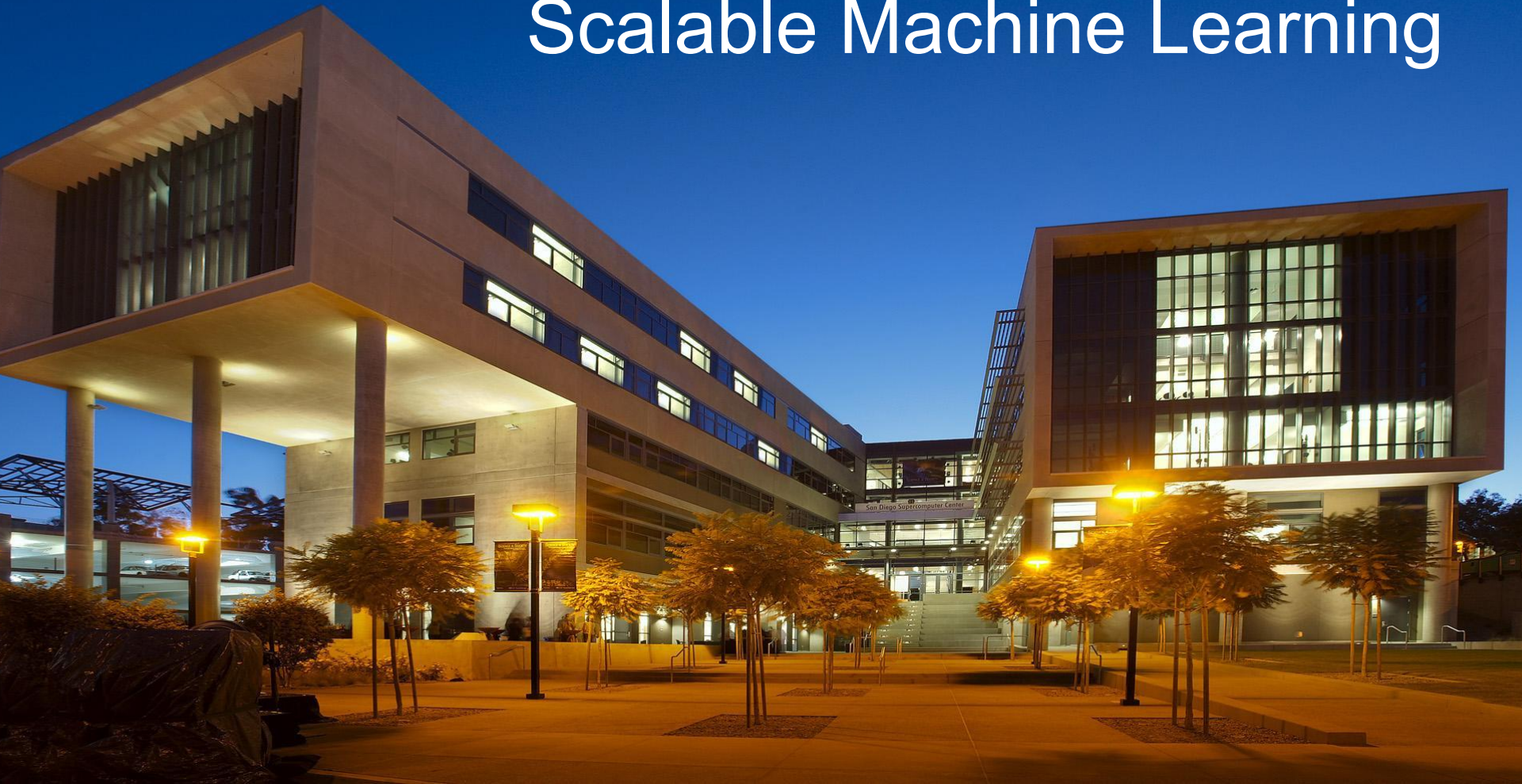


SDSC Summer Institute 2022

Scalable Machine Learning



Introductions

- **Mai H. Nguyen, Ph.D.**
 - Lead for Data Analytics
- **Paul Rodriguez, Ph.D.**
 - Research Analyst

Scalable Machine Learning Agenda

12:00 - 12:10 – Machine Learning Overview

12:10 - 12:45 – R on HPC

12:45 - 12:55 – Break

12:55 - 2:25 – Spark

2:25 - 2:30 – Wrap-up

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](https://www.techtarget.com/whatis/definition/machine-learning))
 - “... 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

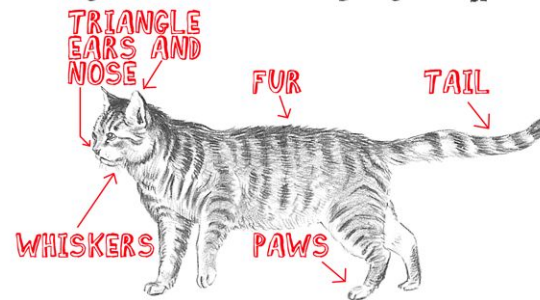
What is Machine Learning?

learning from data

no explicit programming



What Characteristics Do Cats Have



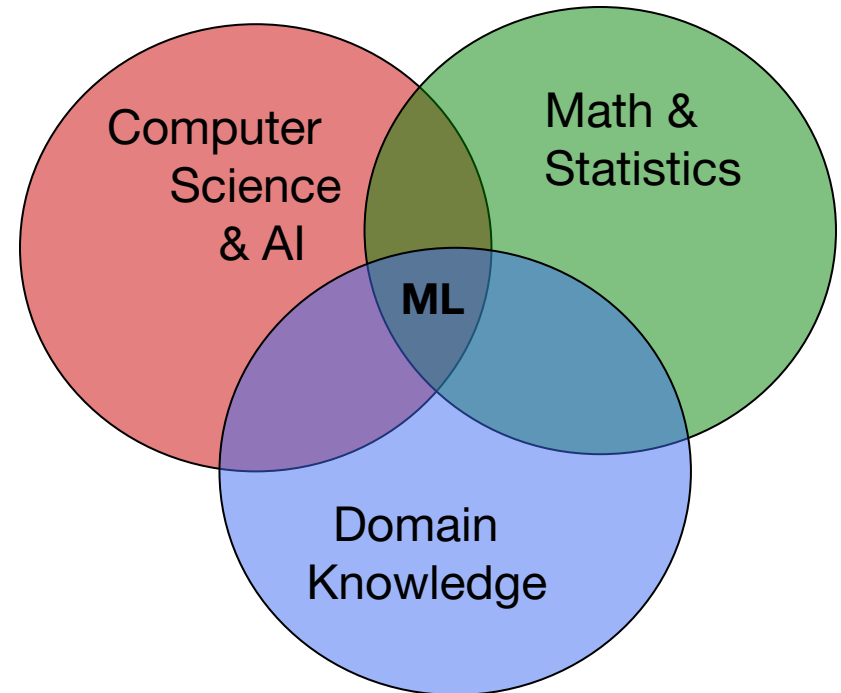
What is Machine Learning?

- **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.



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

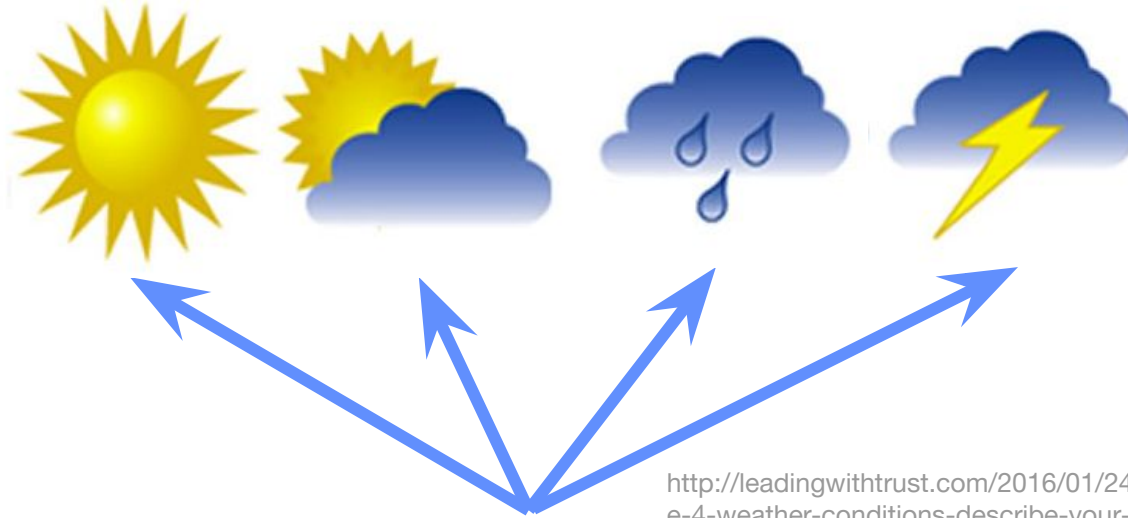


Main Machine Learning Approaches

- **Classification**
- **Regression**
- **Cluster Analysis**

CLASSIFICATION

- **Goal: Predict category given input data**
 - Target is categorical variable



- **Examples**

- Classify tumor as benign or malignant
- Determine if credit card transaction is legitimate or fraudulent
- Identify customer as residential, commercial, public
- Predict if weather will be sunny, cloudy, windy, or rainy

REGRESSION

- **Goal: Predict numeric value given input data**
 - Target is numeric variable



www.wallstreetpoint.com

- **Examples**
 - Predict price of stock
 - Estimate demand for a product based on time of year
 - Determine risk of loan application
 - Predict amount of rain

CLUSTER ANALYSIS

- **Goal:** Organize similar items into groups



<http://www.bostonlogic.com/blog/2014/01/segment-your-leads-to-get-better-results/>

- **Examples**

- Group customer base into segments for effective targeted marketing
- Identify areas of similar topography (desert, grass, etc.)
- Categorize different types of tissues from medical images
- Discover crime hot spots

Supervised vs. Unsupervised

- **Supervised Approaches**

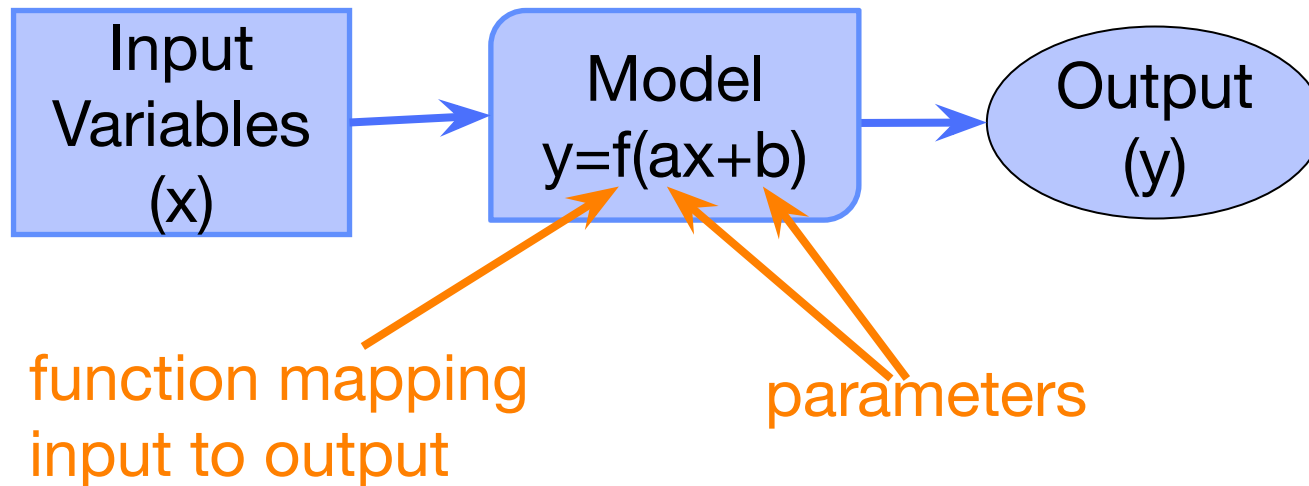
- Target (what you're trying to predict) is provided
 - 'Labeled' data
- Classification and regression approaches are supervised

- **Unsupervised Approaches**

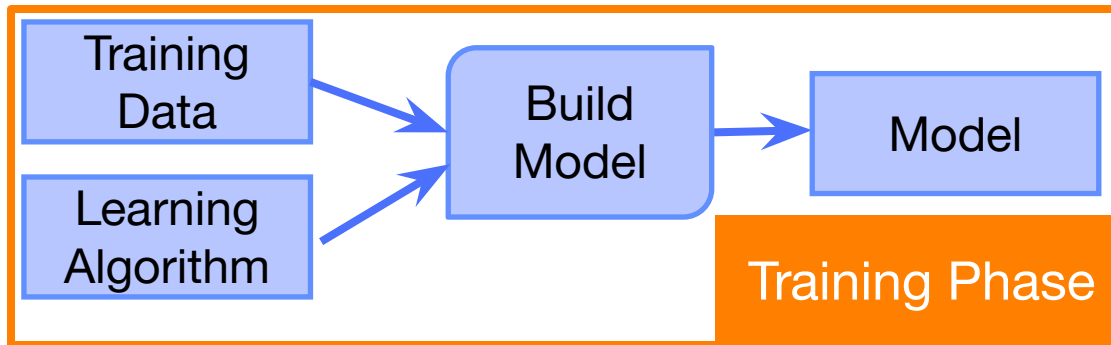
- Target is unknown or unavailable
 - 'Unlabeled' data
- Cluster analysis is unsupervised

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

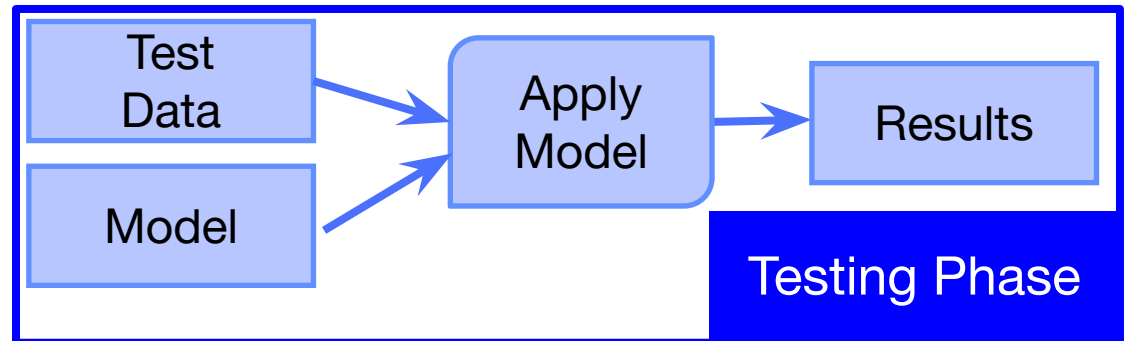


BUILDING VS APPLYING MODEL



Adjust model
parameters
“Train”

Test model on
new data
“Inference”



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

Scalable Machine Learning

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



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

Big Data

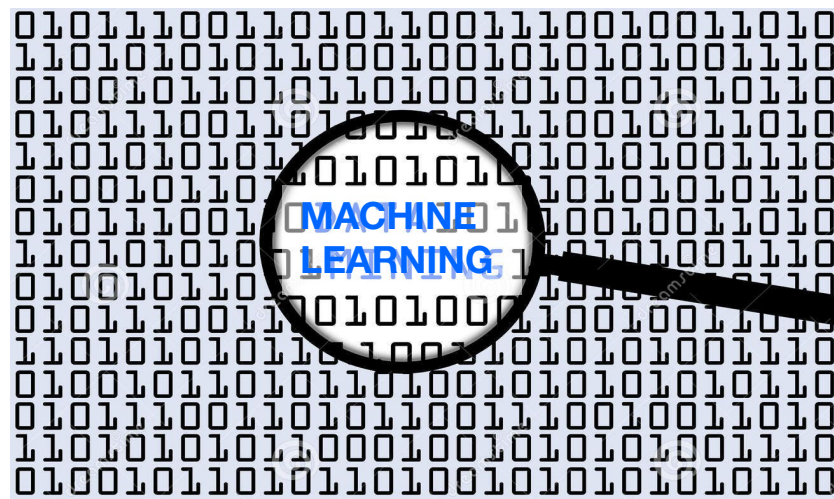
- **V's of Big Data (Doug Laney of Gartner)**
 - **Volume**
 - Vast amounts of data being generated
 - Petabytes (10^{15} bytes), exabytes (10^{18} bytes), and even more
 - **Velocity**
 - Speed at which data is being generated
 - Data is being generated continuously
 - **Variety**
 - Different forms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



35154223

Alain Lacroix | Dreamstime.com

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