# CIML Summer Institute 2023
# Scalable Machine Learning


SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

# Scalable Machine Learning Agenda

8:00 - 8:30 – Breakfast
8:30 - 8:40 – Welcome
8:40 - 10:00 – Introduction to Singularity
10:00 - 10:10 – Break
10:10 - 12:10 – CONDA & Jupyter on Expanse
12:10 - 1:10 – Lunch
1:10 - 1:30 – Machine Learning Overview
1:30 - 2:25 – R on HPC
2:25 - 2:35 – Break
2:35 - 4:35 – 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)
  - "... 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*

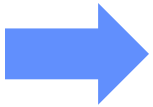
*discover hidden patterns*

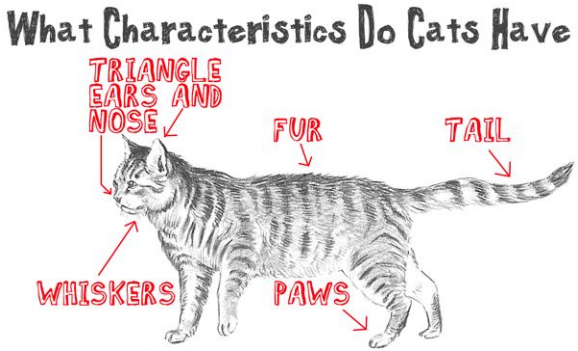
*data-driven decisions*

# What is Machine Learning?

*learning from data*

*no explicit programming*



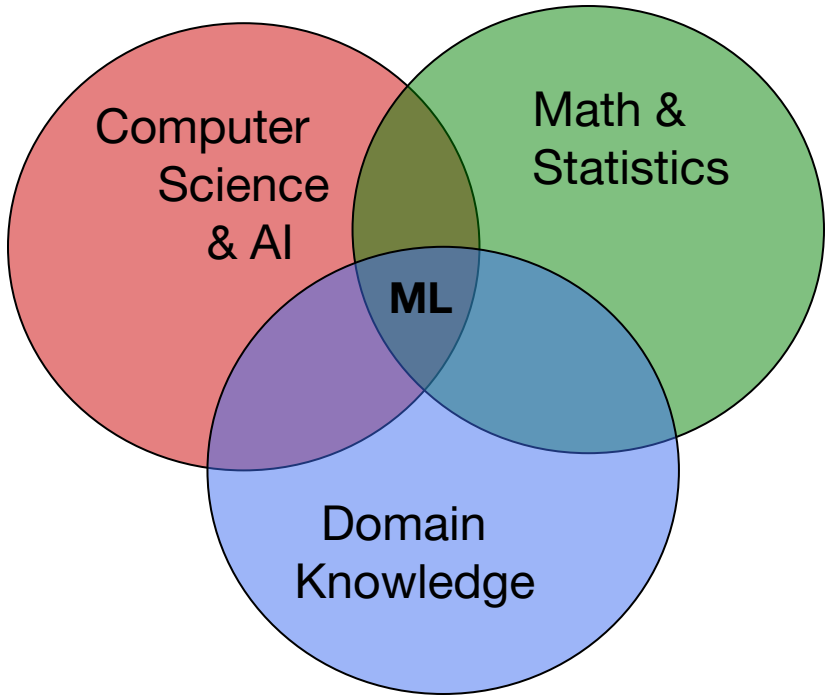


# What is Machine Learning?

- **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**
  - Create unprecedented data
  - Can store and process more data

- **Data-driven applications in many areas**
  - Science:  bioinformatics, image analysis, remote sensing
  - Personal health data from wearable devices
  - Medicine:  drug design, healthcare, data from wearable devices
  - Retail:  targeted advertisement, dynamic pricing
  - Finance:  fraud detection, risk analysis
  - Manufacturing:  preventive maintenance, supply chain management
  - 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)
★★★★½ 153,701
$129.00

Apple AirPods Pro
★★★★½ 54,773
$219.00

Apple EarPods with Lightning Connector - White
★★★★½ 38,539
$19.98

Apple AirPods with Wireless Charging Case
★★★★½ 24,208
$159.99

TOZO T10 Bluetooth 5.0 Wireless Earbuds with Wireless Charging Case IPX8 Waterproof TWS...
★★★★½ 107,951
$29.98

**Inspired by your browsing history**

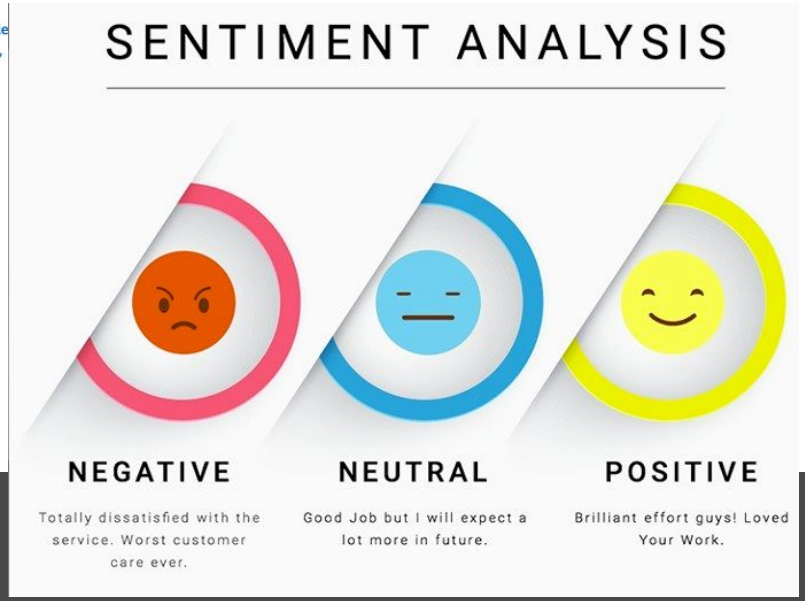
AirPods Case Cover with Keychain, Full Protective Silicone AirPods Accessories Skin Cover...
★★★★½ 18,919
$5.99

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

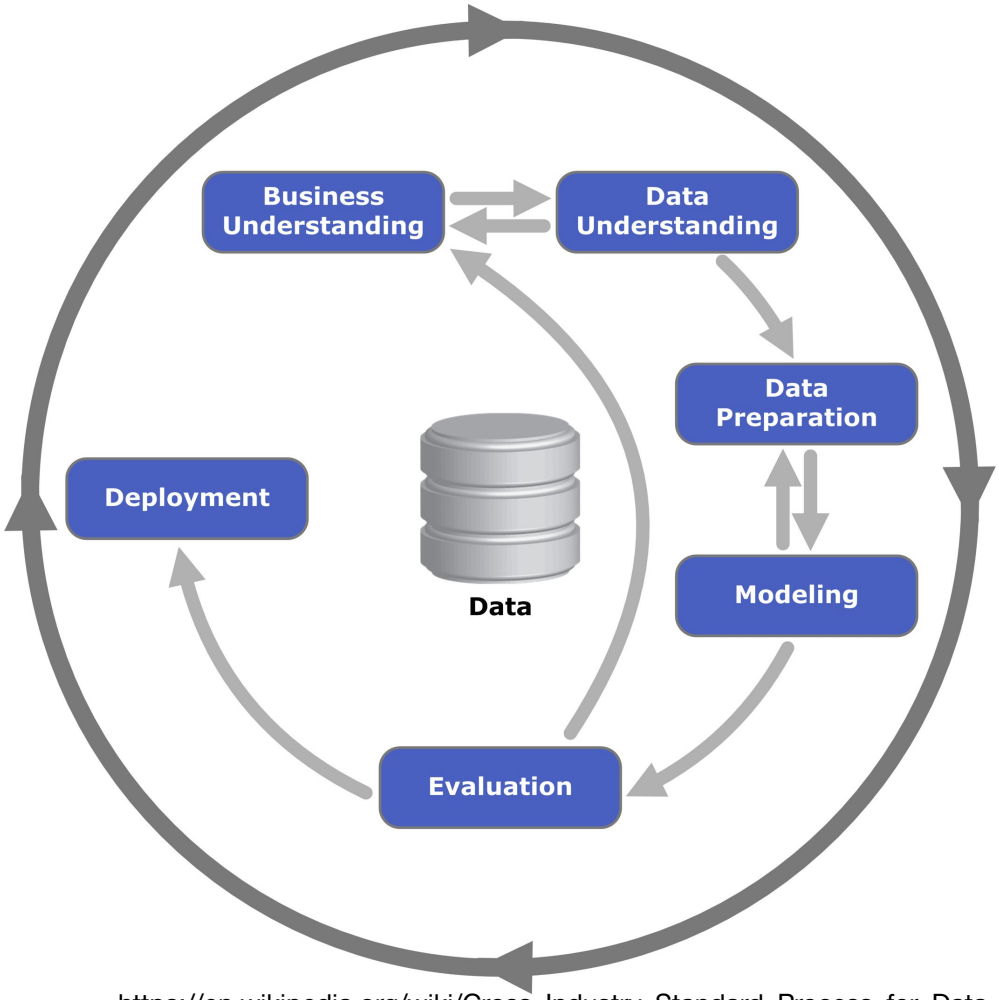
AirPods Case, GMYLE Silicone Protective Shockproof Case Cover Skins with Keychain...
★★★★½ 15,592
$5.98

Apple 5W USB Power Adapter
★★★★½ 3,627
$16.99

AmazonBasics Premium AirPods Case - Compatible with Apple AirPods 1 & 2, Pink
★★★★½ 78
$6.77



# MACHINE LEARNING PROCESS




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

## CRoss Industry Standard Process for Data Mining

ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

# 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-ew-tools-democratizing-data-analysis-visualization/

# Phase 3: Data Preparation

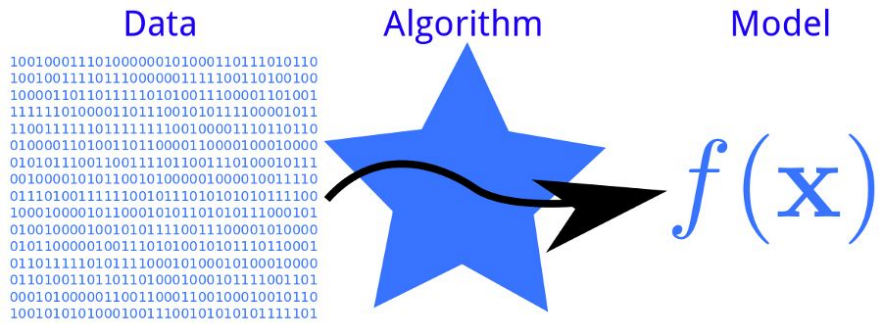
- **Goal:**
  - Prepare data to make it suitable for modeling
  - Also referred to as ‘data preprocessing’, ‘data munging’, ‘data wrangling’

- **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

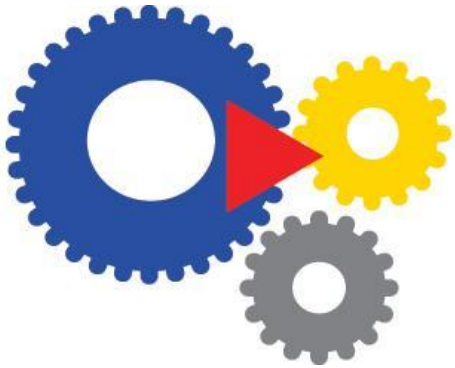
# Phase 5: Evaluation

- **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

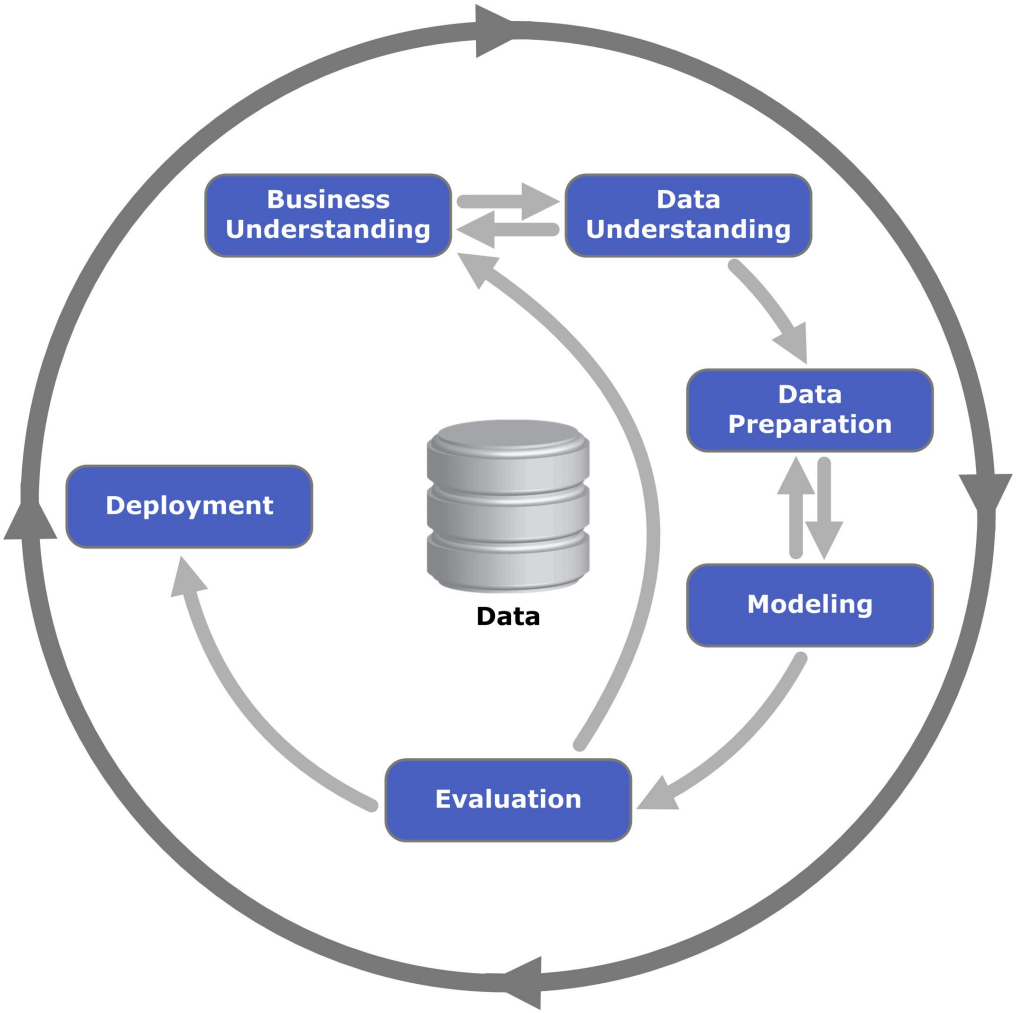
http://www.impactptac.com/?id=10

# Phase 6: Deployment

- **Documentation**
  - Summarize findings and recommend uses
- **Model Deployment**
  - Optimize model for inference
  - Integrate model into decision-making process in production
  - Package model
  - Make model available for inference
- **Model monitoring & maintenance**
  - Monitor model performance
  - Plan for updating/correcting model



# Machine Learning Process



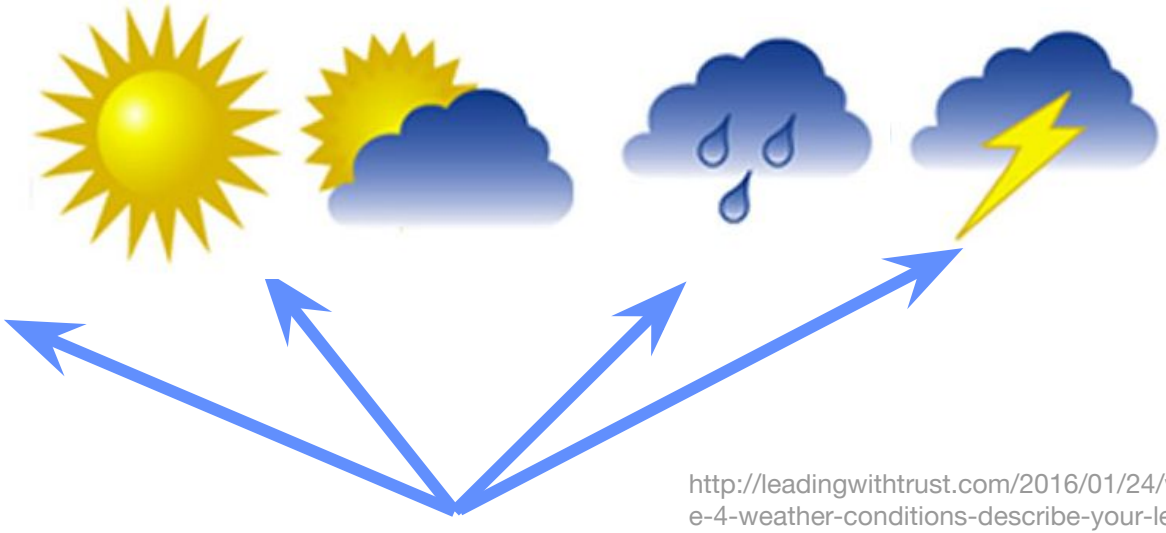

Iterative

# Main Machine Learning Approaches

- Classification
- Regression
- Cluster Analysis

# CLASSIFICATION

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



http://leadingwithtrust.com/2016/01/24/which-of-these-4-weather-conditions-describe-your-leadership/

- **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/

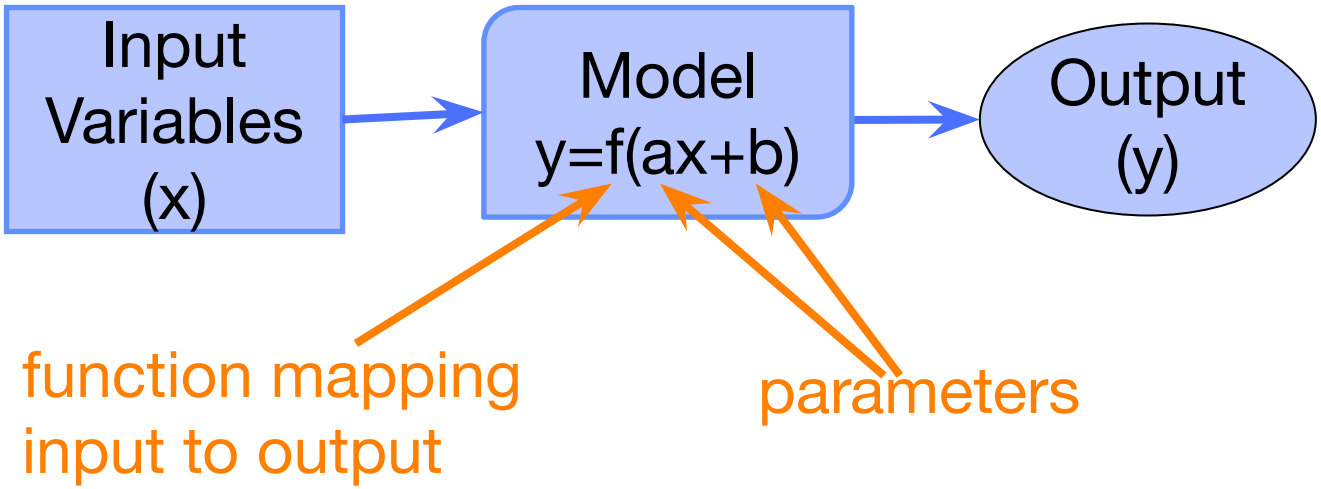
- **Examples**
  - Group customer base into segments for effective targeted marketing
  - Identify areas of similar topography (desert, grass, etc.)
  - Categorize different types of tissues from medical images
  - Discover crime hot spots

# Supervised vs. Unsupervised

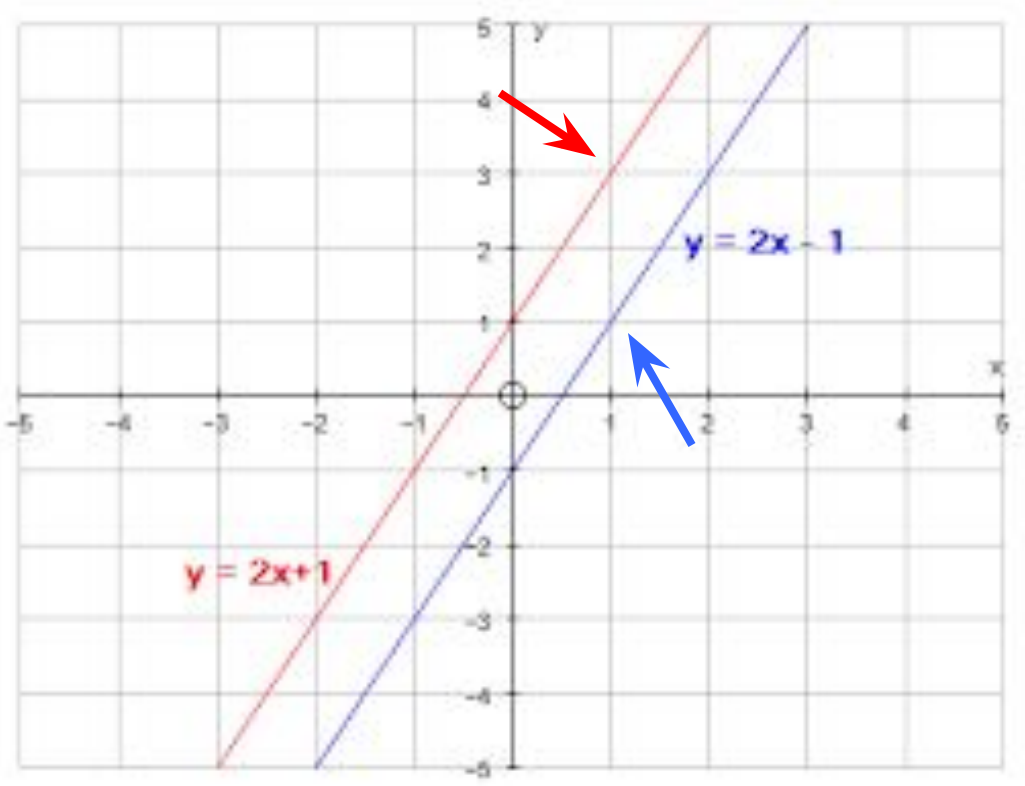
- **Supervised Approaches**
  - Target (what you're trying to predict) is provided
    - 'Labeled' data
  - Classification and regression approaches are supervised
- **Unsupervised Approaches**
  - Target is unknown or unavailable
    - 'Unlabeled' data
  - Cluster analysis is unsupervised

# 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



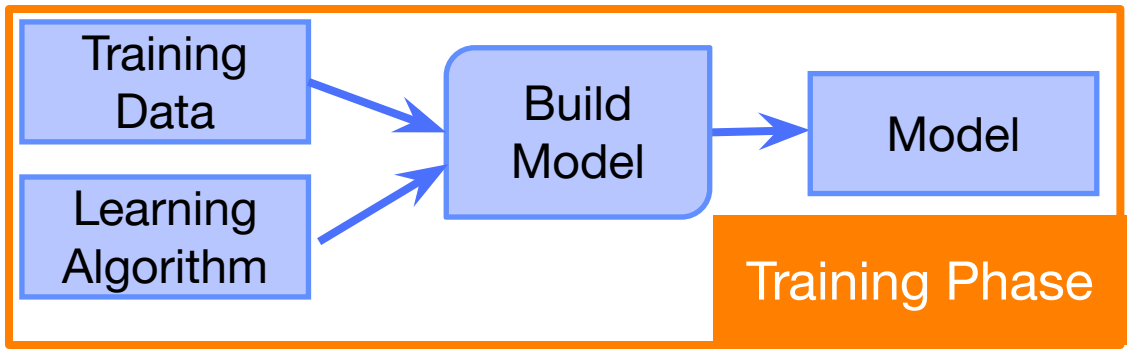
# ADJUSTING MODEL PARAMETERS




slope m = 2
y-intercept b = -1
x=1 => y=2*1-1=1

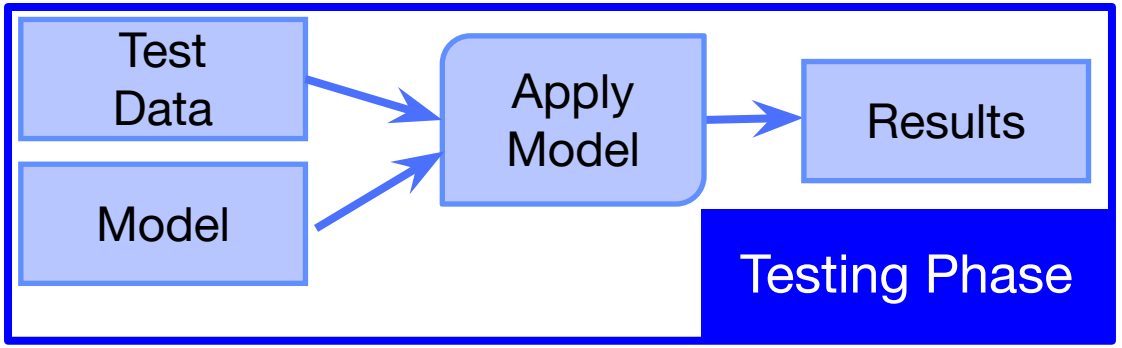
slope m = 2
y-intercept b = +1
x=1 => y=2*1+1= 3

# BUILDING VS APPLYING MODEL




Adjust model parameters
"Train"

Test model on new data
"Inference"



# GENERALIZATION


Split data into two sets

Build model on train data

Train Data

Test Data

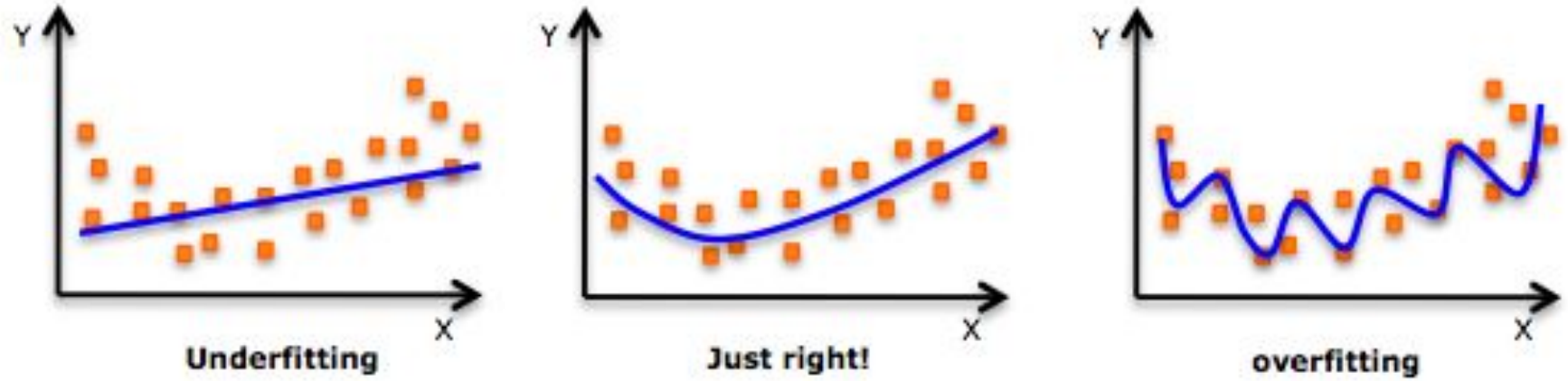
Test model on test data

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




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

| Underfitting | Just Right | Overfitting |
|---|---|---|
| Model has not learned structure of data | Model has learned distribution of data | Model is fitting to noise in data |
| High training error<br>High test error | Low training error<br>Low test error | Low training error<br>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’

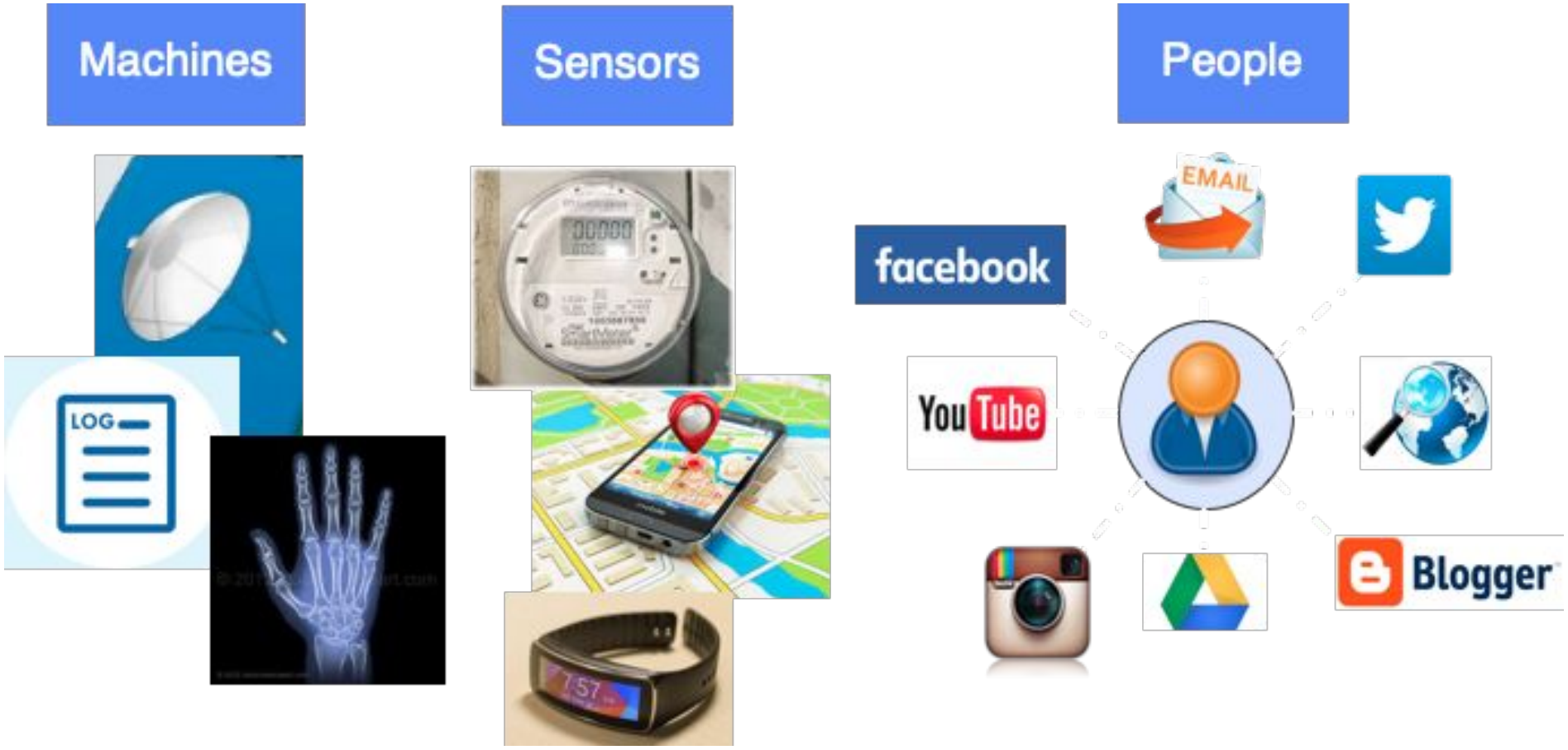

https://infocus.emc.com/scott_burgess/15350/

# Scalable Machine Learning

- **"Growing torrent" of data**
- **Data**
  - Comes in large volumes
  - Continuous
  - Complex

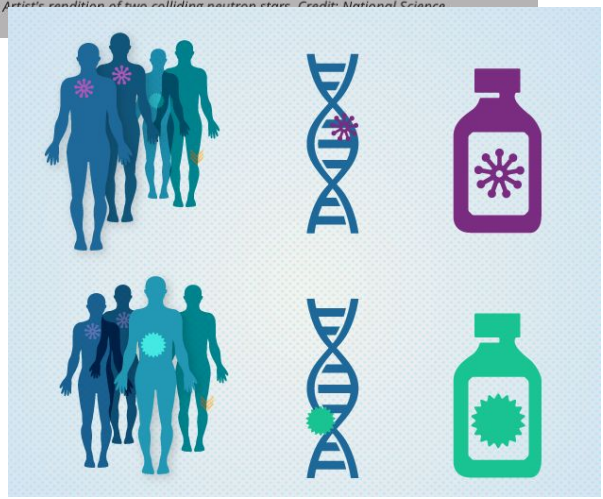
http://www.digitalzenway.com/2011/12/data-diet-a-resolution-you-can-stick-to/

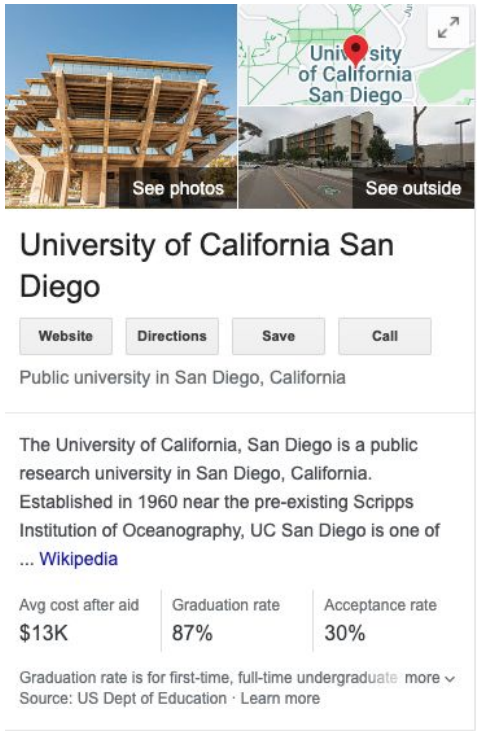
# Where Does Big Data Come From?

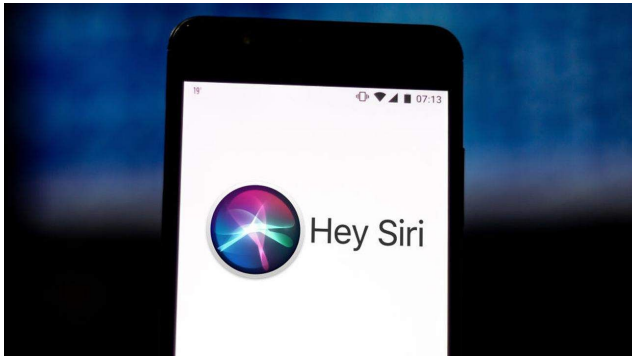


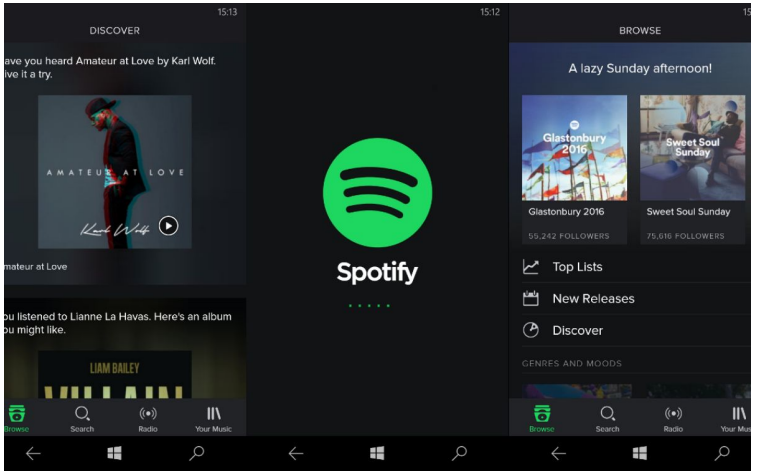
# How is Big Data Used?

Artist's rendition of two colliding neutron stars. Credit: National Science





# V's of 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 continously
  - **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


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

# Machine Learning Overview

- **Machine learning**
  - Definition, applications
- **Machine learning process**
  - Business understanding, data understanding, data preparation, modeling, evaluation, deployment
- **Machine learning approaches**
  - Classification, regression, cluster analysis
  - Supervised vs. unsupervised
- **Machine learning model**
  - Training vs. applying model
  - Overfitting & generalization
- **Scalable machine learning**
  - V's of Big data
  - New approaches needed to scale to big data

# Questions?