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

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**

Introductions

- **Mai Nguyen, Ph.D.**
  - Lead for Data Analytics
- **Paul Rodriguez, Ph.D.**
  - Computational Data Scientist

Machine Learning Overview

Mai H. Nguyen, Ph.D.

What is Machine Learning?

- How would you define machine learning?

Source:
http://halalfocus.net/uk-will-people-pay-more-to-ensure-their-meat-is-not-halal/question-mark-nothing/

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

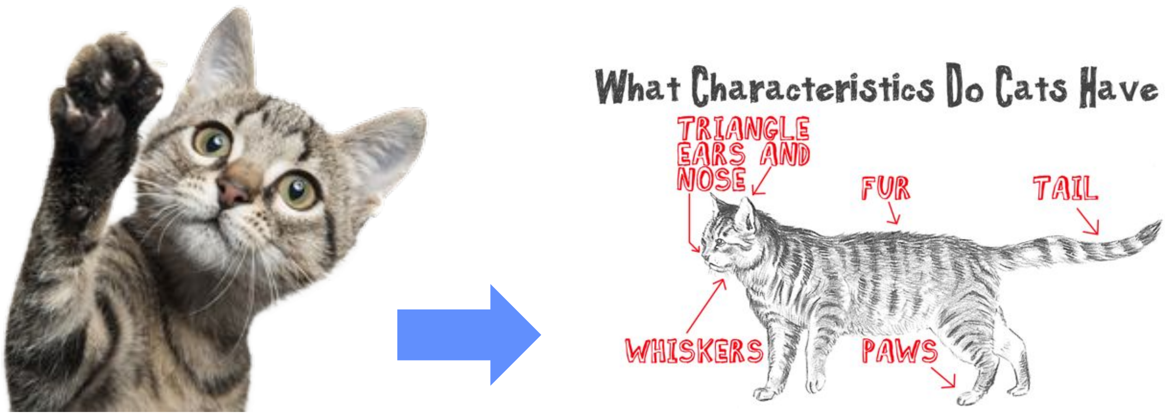
discover hidden patterns

data-driven decisions

What is Machine Learning?

learning from data

no explicit programming

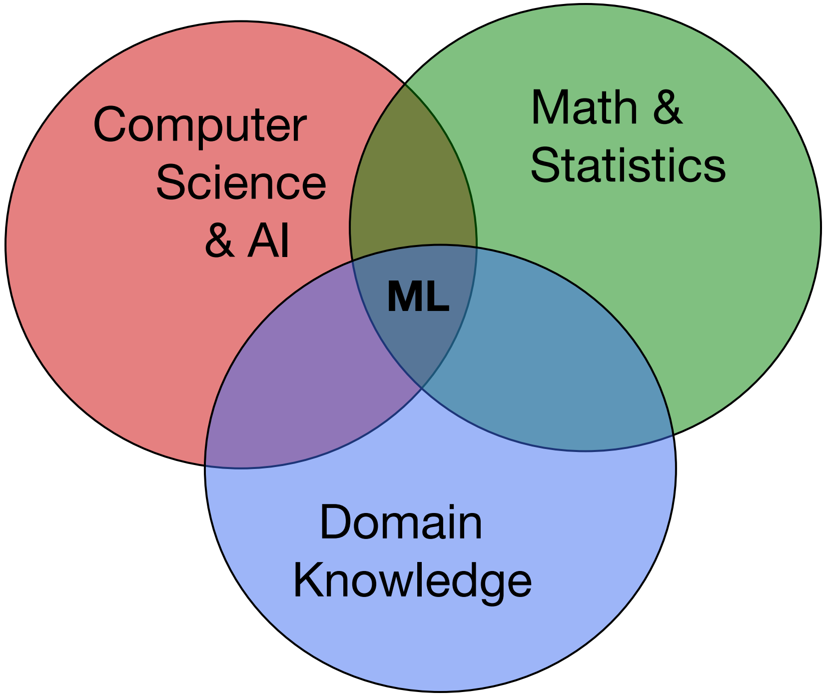


What is Machine Learning?

- **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?

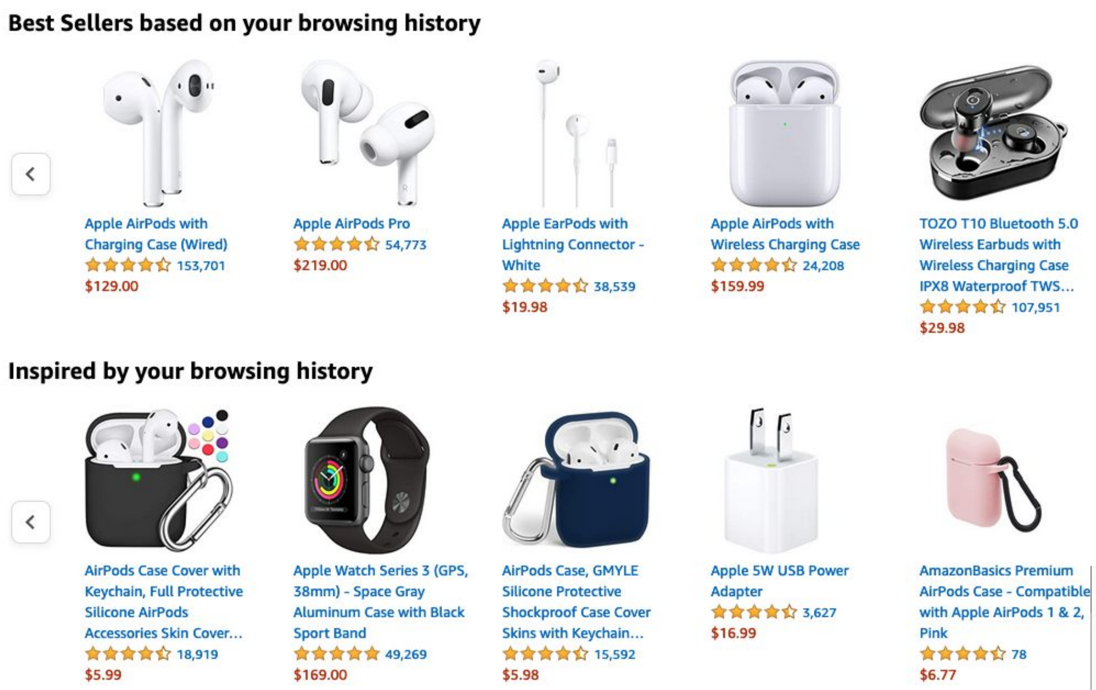
- **Advances in processing power, storage capacity, mobile computing, and interconnectivity**
  - Create unprecedented data
  - Can store and process more data

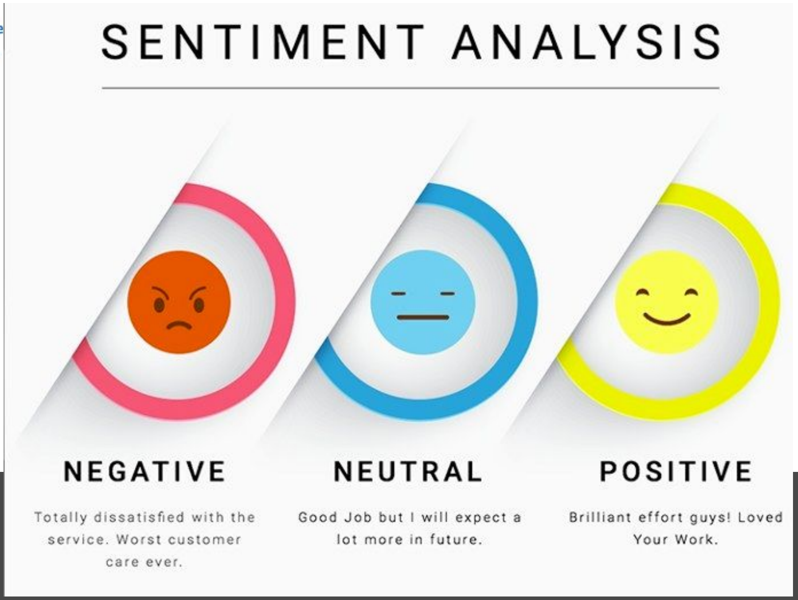
- **Data-driven applications in many areas**
  - Science: bioinformatics, image analysis, remote sensing
  - Personal health data from wearable devices
  - Medicine: drug design, healthcare, data from wearable devices
  - Retail: targeted advertisement, dynamic pricing
  - Finance: fraud detection, risk analysis
  - Manufacturing: preventive maintenance, supply chain management
  - Social media data related to customer satisfaction, political trends, health epidemics, law enforcement, terrorist activities

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

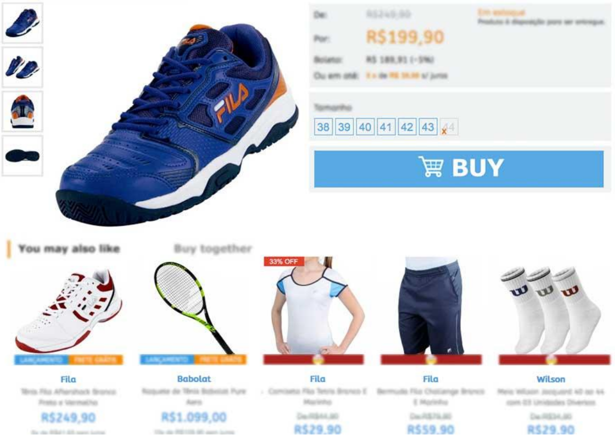
MACHINE LEARNING APPLICATIONS





Applications of Machine Learning

- Recommendations on websites
- Targeted ads on mobile apps
- Handwriting recognition
- Fraud detection
- Sentiment analysis
- Network intrusion detection
- Drug effectiveness analysis
- Crime pattern detection
- Self-driving cars

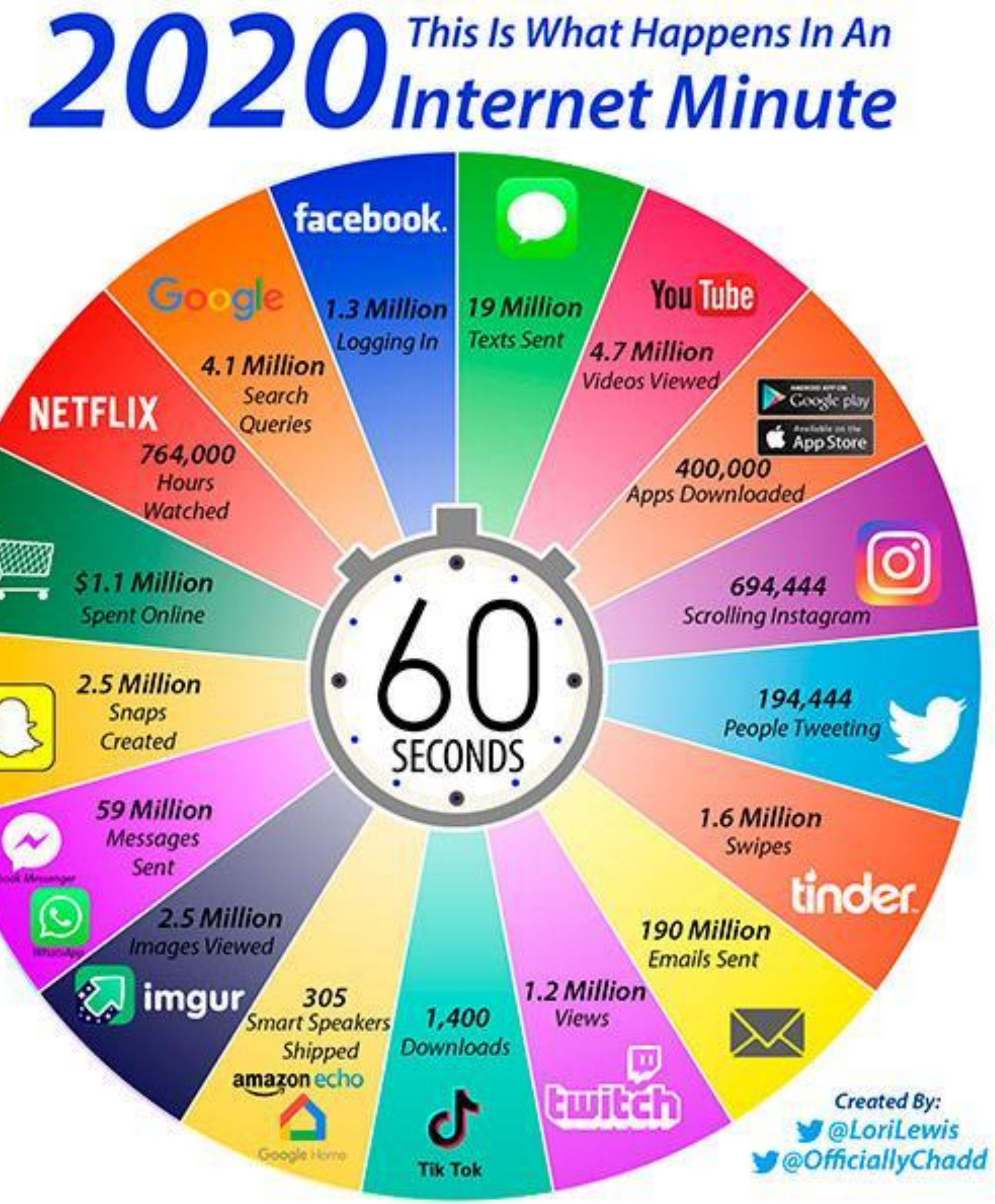


Scientific Data Analysis

- **HPWREN – High Performance Wireless Research and Education Network**
  - 30 TB of data: sensor and imagery data from weather stations in San Diego county per year (hpwren.ucsd.edu)
- **MODIS – Moderate Resolution Imaging Spectroradiometer**
  - 219 TB of data: moderate resolution satellite imagery covering Earth's surface per year (modis.gsfc.nasa.gov)
- **Precision Medicine**
  - 4 EB (10^{18} bytes) of data: genome sequences of people diagnosed with cancer (www.fastcompany.com)
- **LIGO, Deep Space Network, Protein Data Bank, …**

How much data is generated every minute on the Internet?

https://www.allaccess.com/merge/archive/31294/infographic-what-happens-in-an-internet-minute



Data Deluge

- **Data Deluge:**
  - Rapid growth in amount of digital data, and problems of managing this data.
  - "We are drowning in information and starving for knowledge"
    – John Naisbitt

    Source: Megatrends, 1982

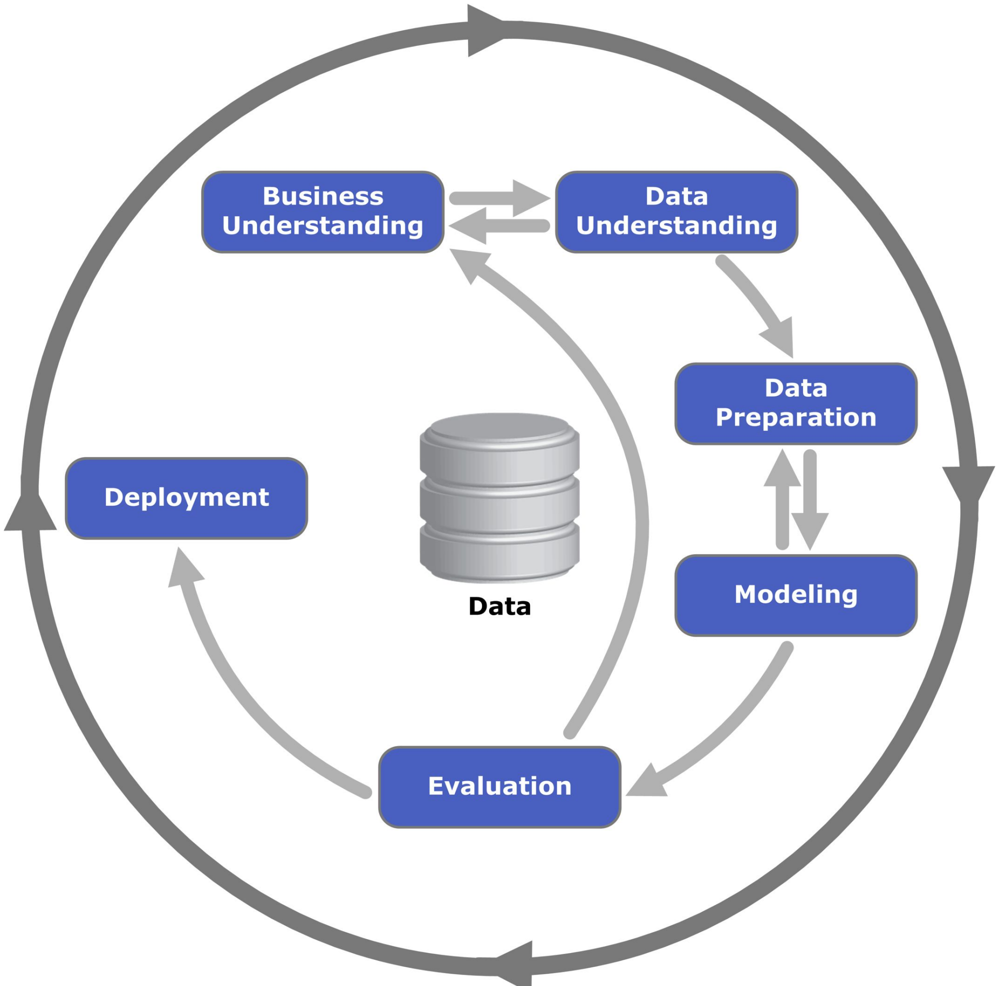
http://www.digitalzenway.com/2011/12/data-diet-a-resolution-you-can-stick-to/

Why do Machine Learning?

- **How can all of this data be turned into useful information?**
- **Answer:**
  - Apply machine learning!

http://www.kdnuggets.com/2015/03/all-machine-learning-models-have-flaws.html

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-visualiza on/

Exploratory Data Analysis

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

- **Goal:**
  - Exploratory data analysis -> data understanding -> informed analysis
  - Also referred to as ‘data profiling’.
- **Techniques:**
  - Summary statistics
    - Mean, frequency, mode, range, variance, standard deviation, etc.
  - Visualization
    - Histograms, scatter plots, line graphs, etc.
  - Look for:
    - Correlations, general trends, outliers, etc.

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

Data Quality

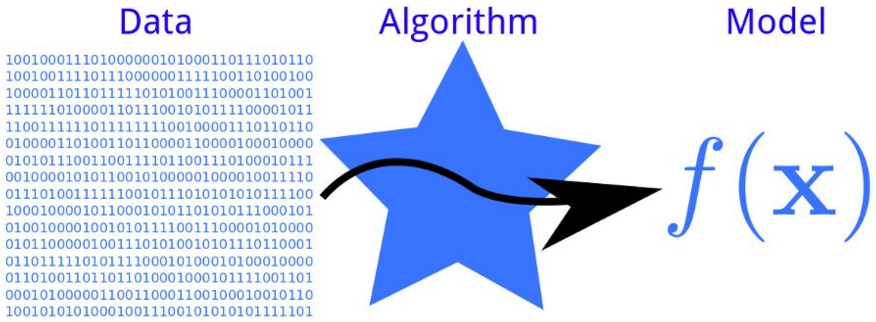
- **Data Quality Issues**
  - Missing Values
  - Duplicate Data
  - Inconsistent Data
  - Noise
  - Outliers

Source: *http://www.datasciencecentral.com/profiles/blogs/5-data-cleansing-tools*

- **Addressing data quality**
  - Also referred to as ‘data cleansing’ or ‘data cleaning’.
- **Important: Garbage in = Garbage out!**
  - Proper data preparation is crucial to machine learning process.

Phase 4: Modeling

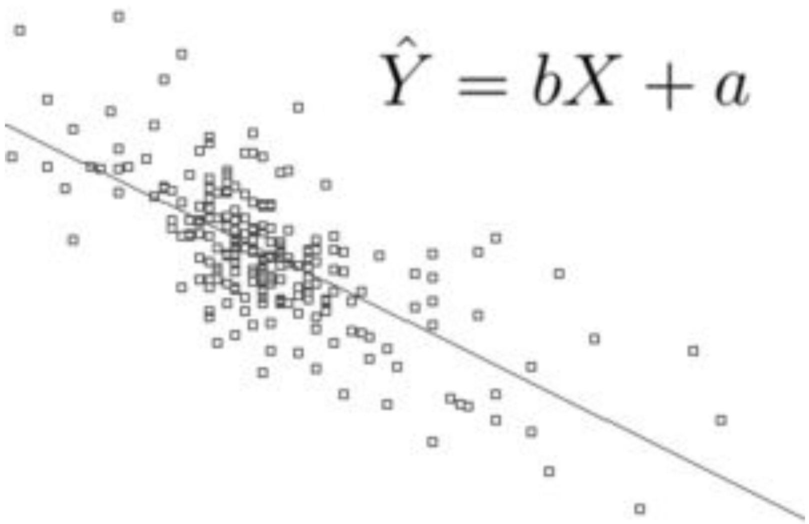
- **Determine type of problem**
  - Classification
  - Regression
  - Cluster analysis
- **Build model(s)**
  - Select modeling technique(s) to use
  - Construct model(s)
  - Train model(s)



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

Building Model

- **Goal:**
  - Construct model that accurately predicts targets of training data as well as of new data.
  - This is called "generalization".
- **Process:**
  - Adjust model's parameters to minimize error using a learning algorithm.



Source: https://en.wikiversity.org/wiki/Linear_regression

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

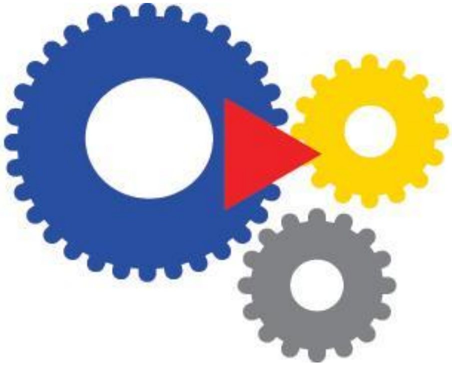
Evaluation Outcome

- **Determine next steps**
  - Go/No-go decision
  - Go:
    - Proceed to Model Deployment to apply model.
  - No-Go:
    - List of possible actions
      - Different modeling technique?
      - More data cleansing?
      - More data?

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

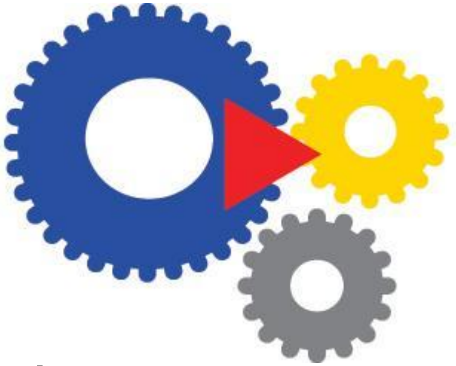
Phase 6: Deployment

- **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



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.

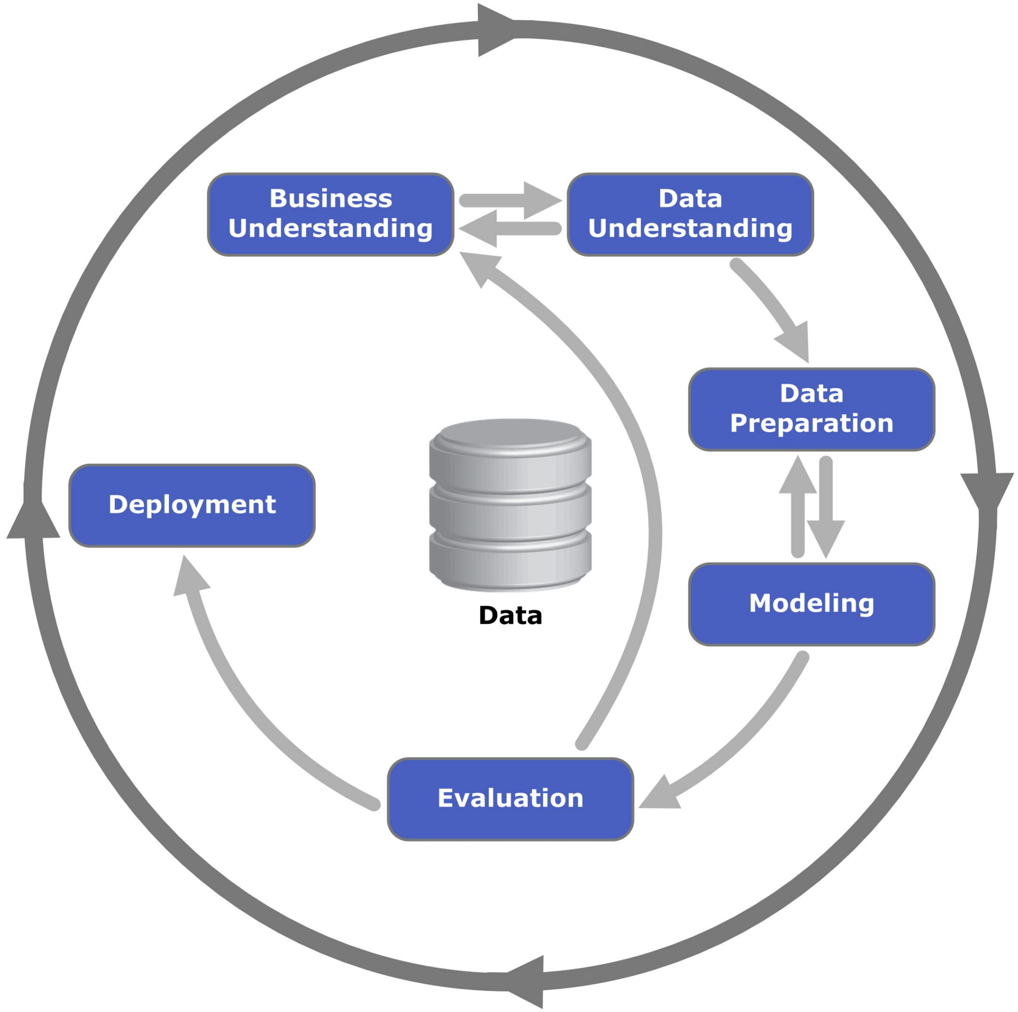


Model Deployment

- **Considerations**
  - Think system-level: ML model is part of a larger system
  - How quickly and often does model need to make predictions?
  - How large is the data that needs to be processed?
  - Plan for monitoring, updates, etc.
- **Tools**
  - Cloud service (PaaS, IaaS, SaaS)
  - Containers (Docker, Kubernetes)
  - Distributed web app architecture
  - Microservices
  - Others …

Machine Learning Process

Iterative



DM Process – Key Points

- **CRISP-DM**
  - Process model that describes phases in data mining process
- **Phases**
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

References

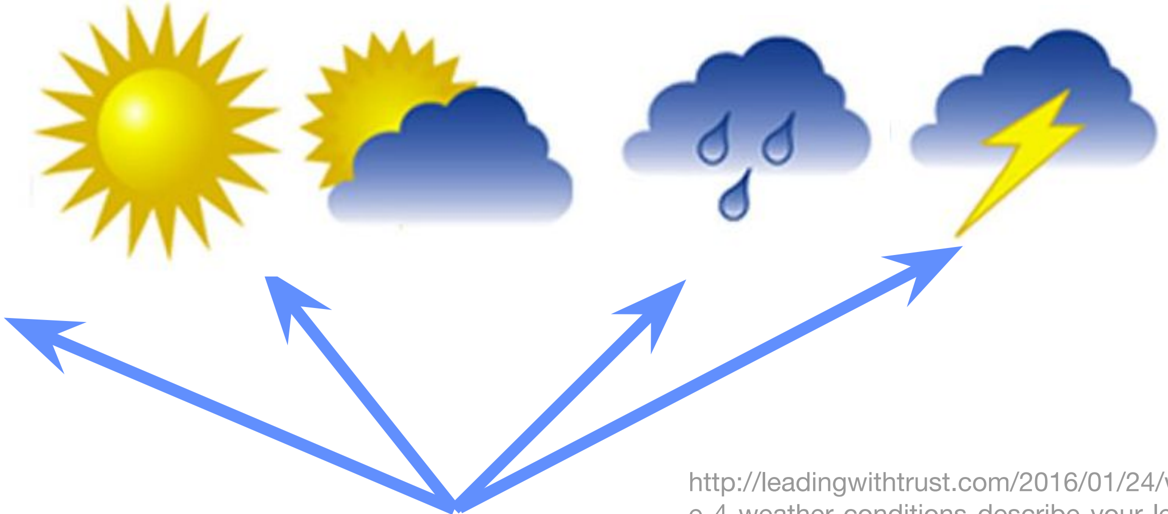
- SPSS. (2000). CRISP-DM 1.0. Retrieved from ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

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

Association Analysis

- **Goal: Find rules to capture co-occurrence relationships between items**

Customers who bought this:

Source: http://www.supercouponlady.com/best-diaper-deals-this-week/

Also bought:

Source: http://www.bizjournals.com/triangle/news/2012/06/21/new-craft-beer-store-opening-in-north.html

Association Analysis Examples

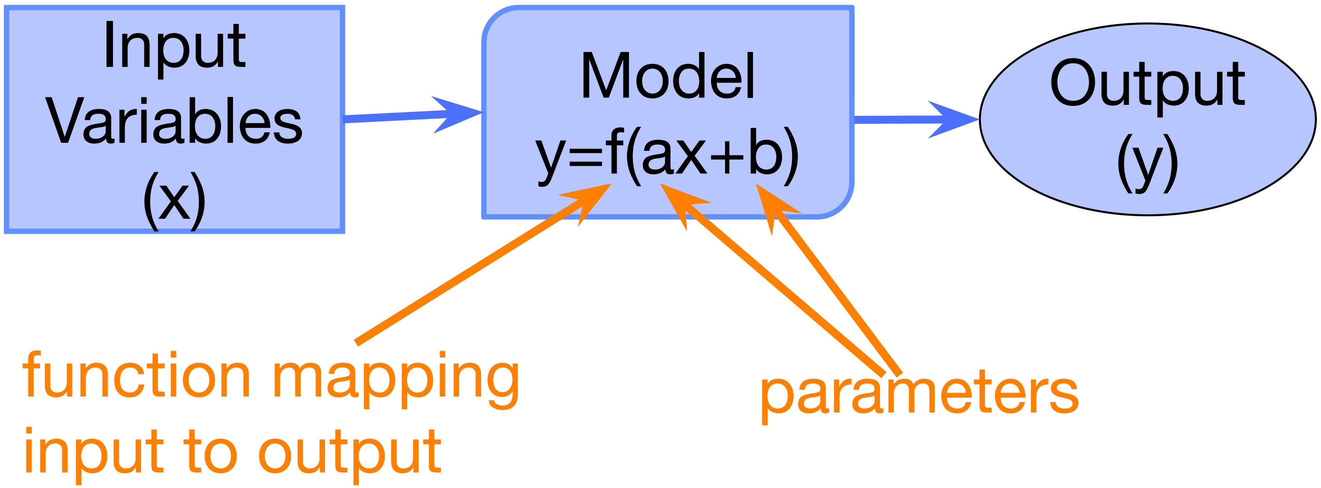
- **Cross-selling**
  - Recommended items based on your purchase/browsing history
- **Sales promotions**
  - Have sales on garden hose and potting soil at same time since people tend to buy these items together
- **Product placement**
  - Place diapers close to beer aisle to drive sales of both products.

Supervised vs. Unsupervised

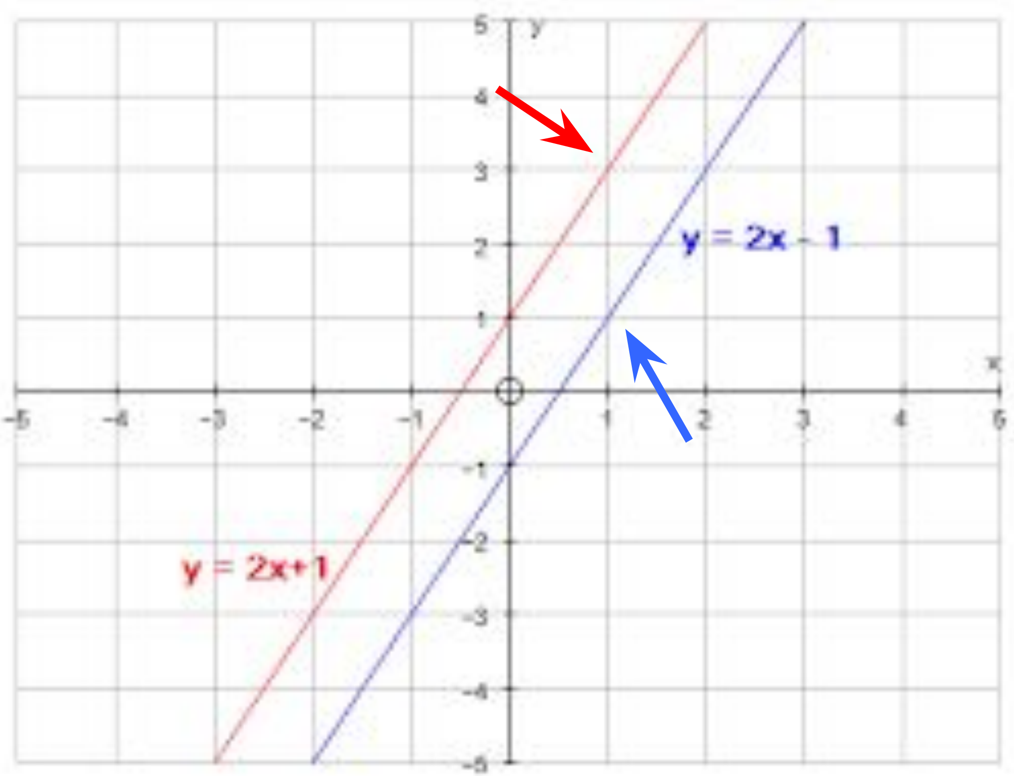
- **Supervised Approaches**
  - Target (what you’re trying to predict) is provided
    - ‘Labeled’ data
  - Classification and regression approaches are supervised
- **Unsupervised Approaches**
  - Target is unknown or unavailable
    - ‘Unlabeled’ data
  - Cluster analysis is unsupervised

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



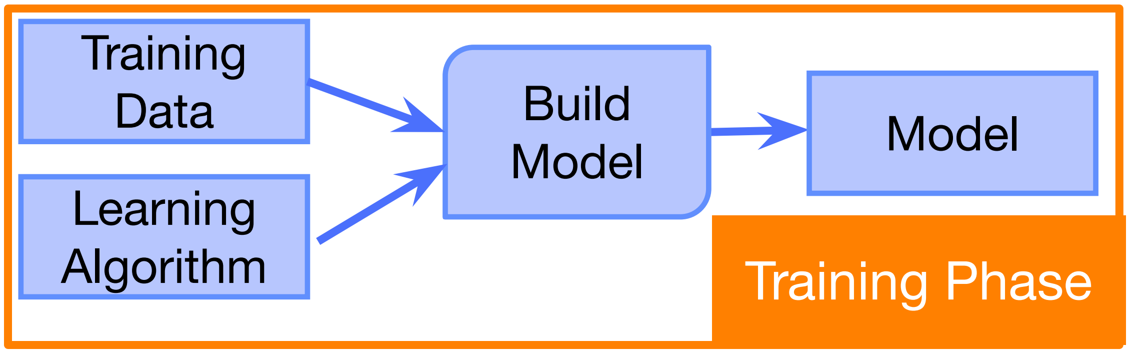
ADJUSTING MODEL PARAMETERS



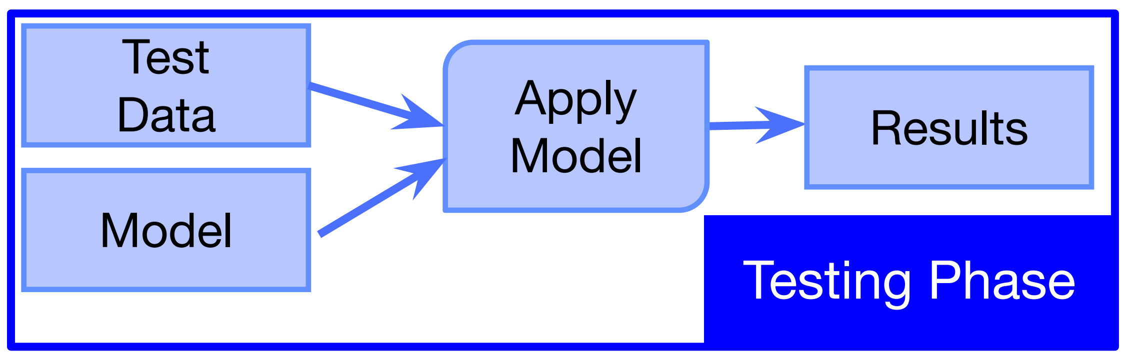
slope m = 2
y-intercept b = -1
x=1 => y=2*1-1=1

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

BUILDING VS APPLYING MODEL



Adjust model parameters
“Train”

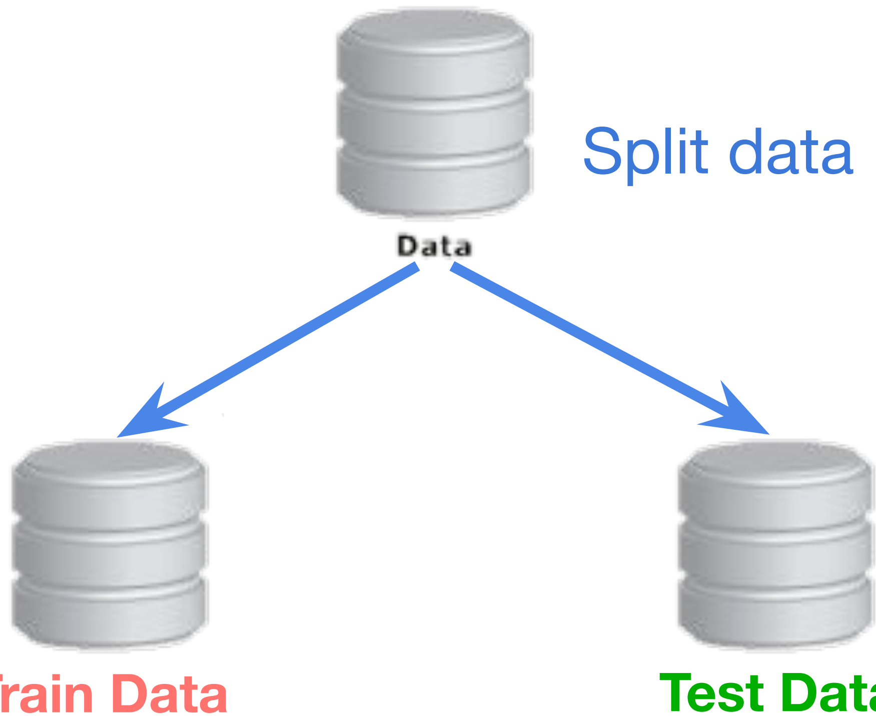


Test model on new data
“Inference”

GENERALIZATION

Split data into two sets

Build model on train data



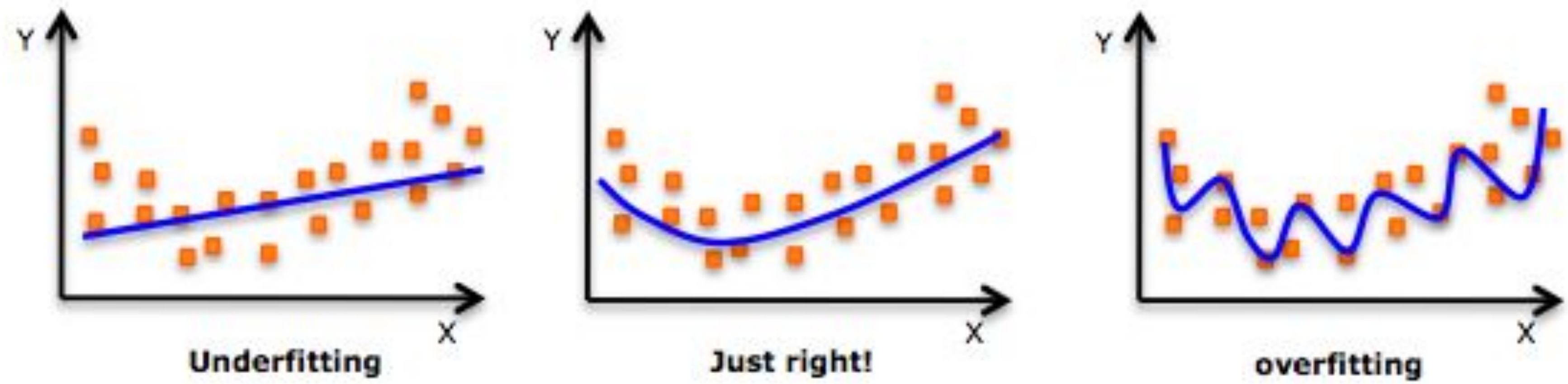
Test model on test data

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



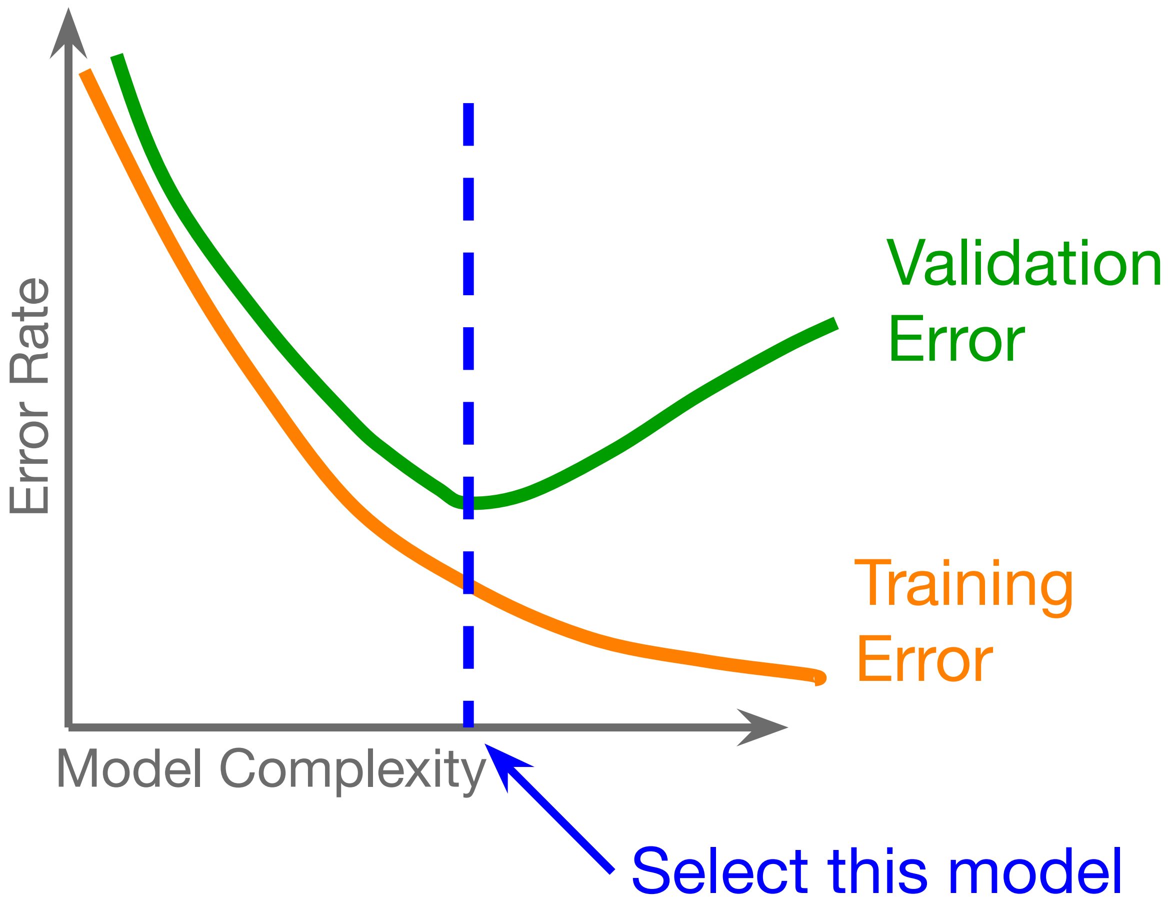
http://stats.stackexchange.com/questions/192007/what-measures-you-look-at-the-determine-over-fitting-in-linear-regression

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 High test error	Low training error Low test error	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

VALIDATION SET



Scalable Machine Learning

- What is scalable machine learning?
- Applying machine learning to ‘big data’

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

Scalable Machine Learning

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

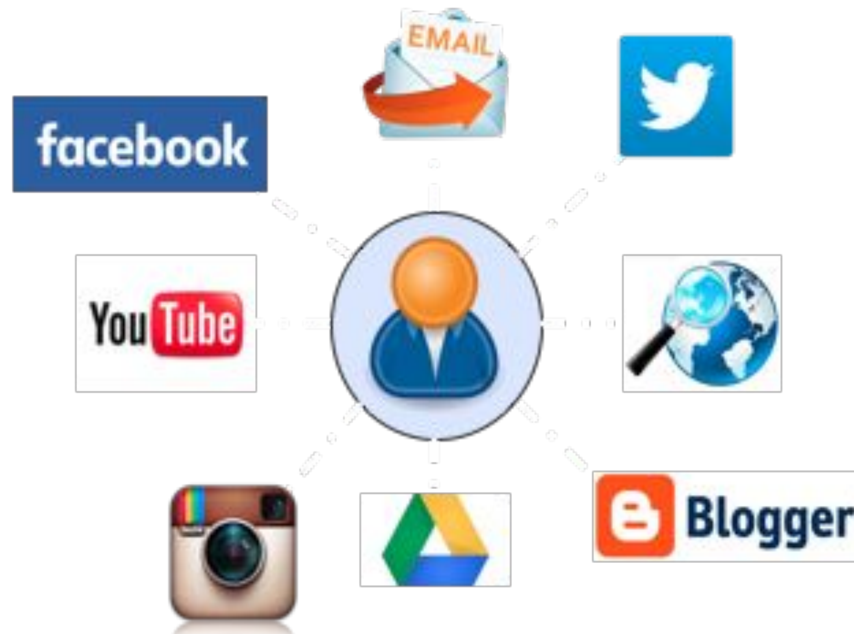
- **"Growing torrent" of data**
- **Data**
  - Comes in large volumes
  - Continuous
  - Complex

Where Does Big Data Come From?

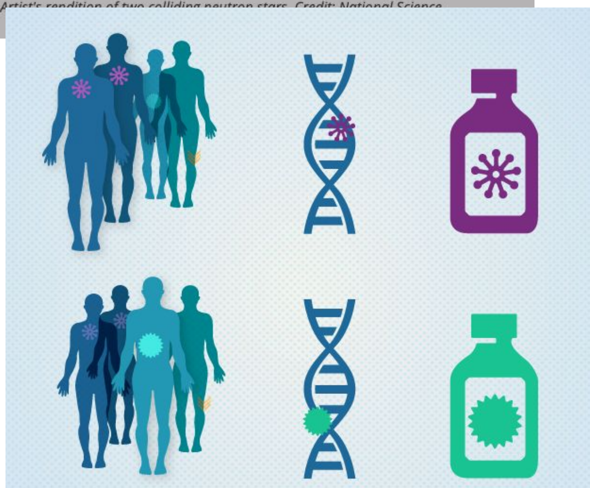
Machines

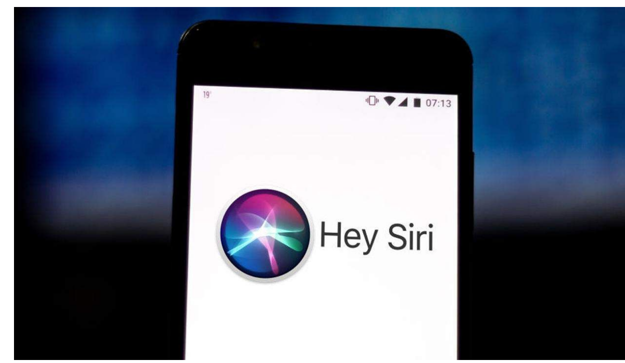
Sensors

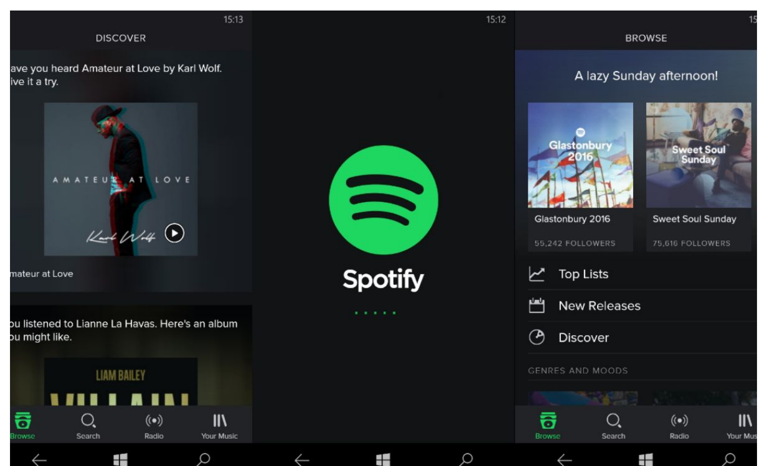
People



How is Big Data Used?







V’s of Big Data

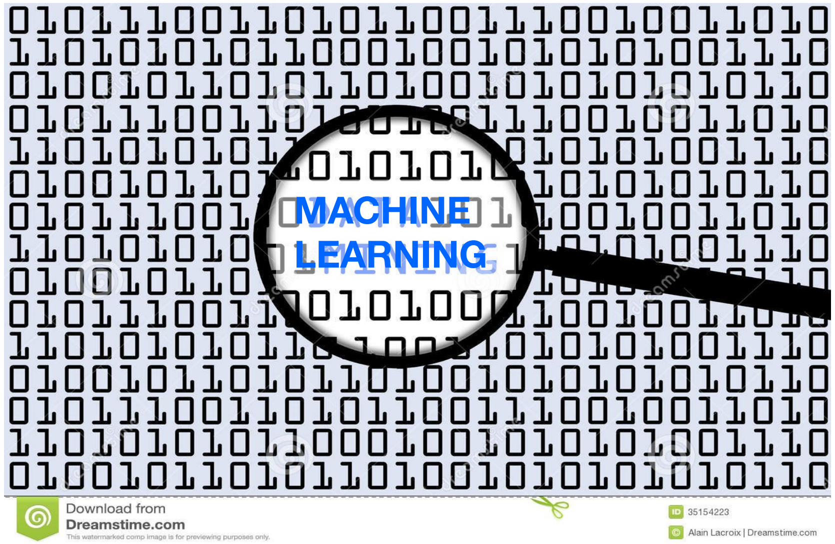
- **V’s of Big Data (Doug Laney of Gartner)**
  - **Volume**
    - Vast amounts of data being generated
    - Petabytes (10^{15} bytes), exabytes (10^{18} bytes), and even more
  - **Velocity**
    - Speed at which data is being generated
    - Data is being generated continously
  - **Variety**
   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

ML Introduction – Key Points

- **Definition of machine learning (What)**
- **Reasons for doing machine learning (Why)**
- **Machine learning approaches (How)**
  - Classification
  - Regression
  - Cluster analysis
  - Supervised vs. unsupervised
- **Machine learning process**
  - Business understanding, data understanding, data preparation, modeling, evaluation, deployment

Questions?