2018 SDSC Summer Institute
Machine Learning Overview

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

Machine Learning Overview Agenda

8:30 - 9:00 – Machine Learning Introduction

9:00 - 9:30 – Data Exploration

9:30 - 10:00 – Data Preparation

10:00 - 10:15 – Break

10:15 - 11:00 – Classification

11:00 - 11:45 – Clustering

11:45 - 12:00 – Wrap-Up

Introduction to Machine Learning

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?

data-driven decisions

discover hidden patterns

no explicit programming

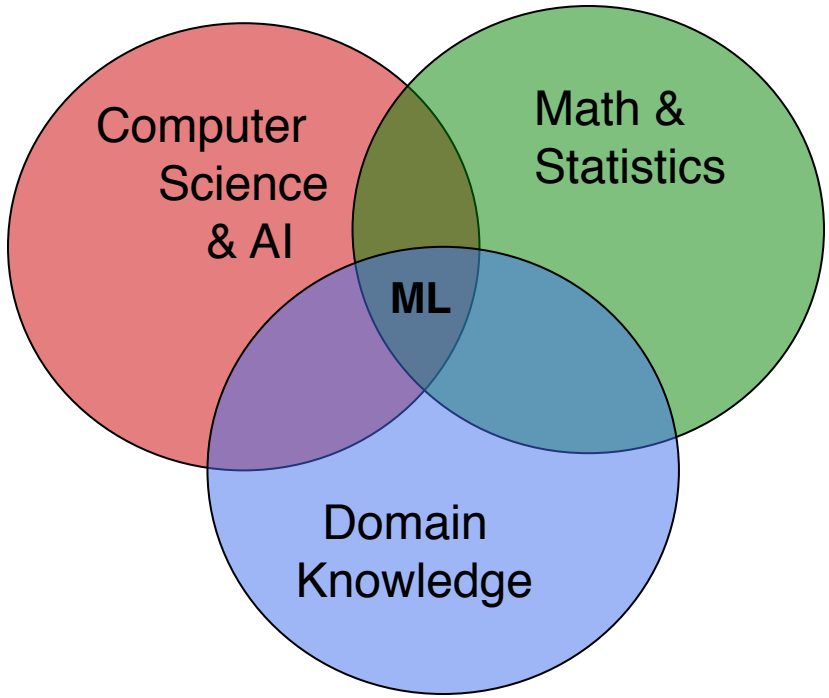
learning from data

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 has been used in various applications:**
  - Science, engineering, business, medical, law enforcement, etc.



Why the Increased Interest in ML?

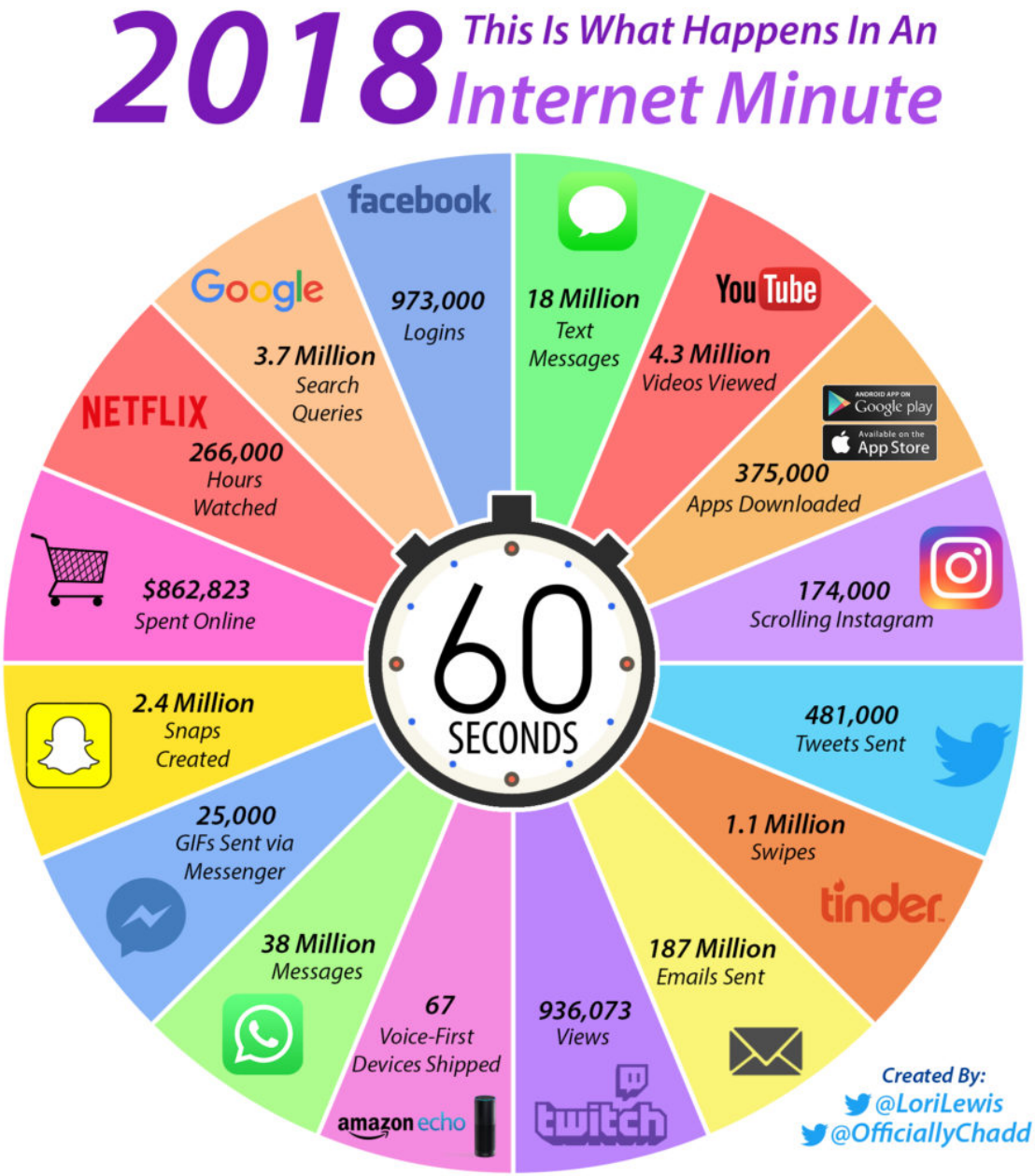
- **Advances in processing power, storage capacity, mobile computing, and interconnectivity are creating unprecedented data:**
  - User preferences and purchasing history on websites
  - Scientific data from remote sensors and instruments
  - Personal health data from wearable devices
  - Medical data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

Scientific Data Analysis

- **HPWREN**
  - 30 TB of data: sensor and imagery data from weather stations in San Diego county per year (hpwren.ucsd.edu)
- **MODIS**
  - 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 who will be diagnosed with cancer in 2016 (www.fastcompany.com)
- **LIGO, Deep Space Network, Protein Data Bank, …**

How much data is generated every minute on the Internet?

http://www.visualcapitalist.com/internet-minute-2018/



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

Source:
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!

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

Applications of Machine Learning

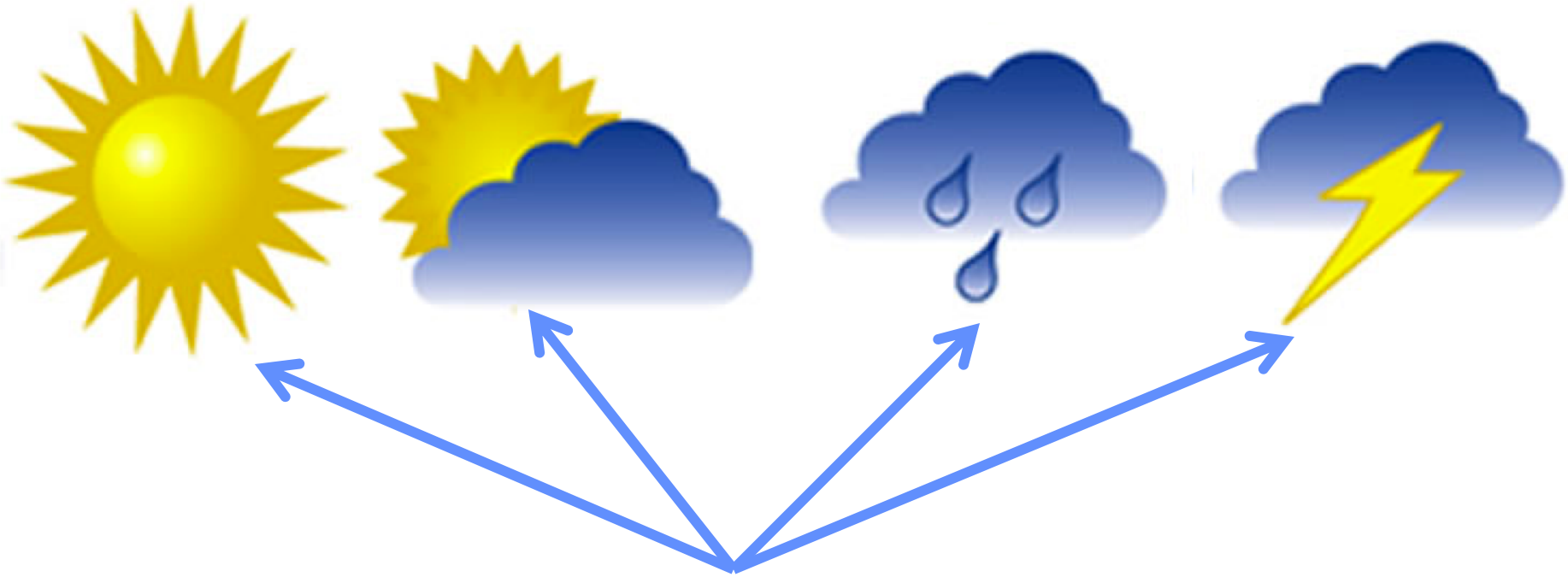
- Recommendations on websites
- Targeted ads on mobile apps
- Handwriting recognition
- Fraud detection
- Sentiment analysis
- Network intrusion detection
- Drug effectiveness analysis
- Crime pattern detection
- Self-driving cars

Machine Learning Approaches

- Classification
- Regression
- Cluster Analysis
- Association Analysis

Classification

- **Goal: Predict category given input data.**



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

Classification Examples

- Classify tumor as benign or malignant
- Determine if credit card transaction is legitimate or fraudulent
- Identify handwritten digits (0..9)
- Predict if weather will be sunny, cloudy, windy, or rainy.
- Identify hospital patients at high risk of re-admission.

Regression

- **Goal: Predict numeric value given input data.**

Source: www.wallstreetpoint.com

Regression Examples

- Predict price of a stock
- Estimate likelihood of drug effectiveness for a patient
- Determine risk of loan application
- Predict amount of rain for a particular region
- Estimate demand for a product based on time of year

Cluster Analysis

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

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

Cluster Analysis Examples

- Group customer base into segments for more effective targeted marketing.
- Identify areas of similar topography (e.g., mountains, desert, plains) for land use applications.
- Categorize genes with similar functionalities.
- Identify 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
  - Association analysis and cluster analysis are unsupervised.

Where's the Big Data?

- How is all of this related to Big Data?
- First, let's define 'Big Data'.

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

Big Data

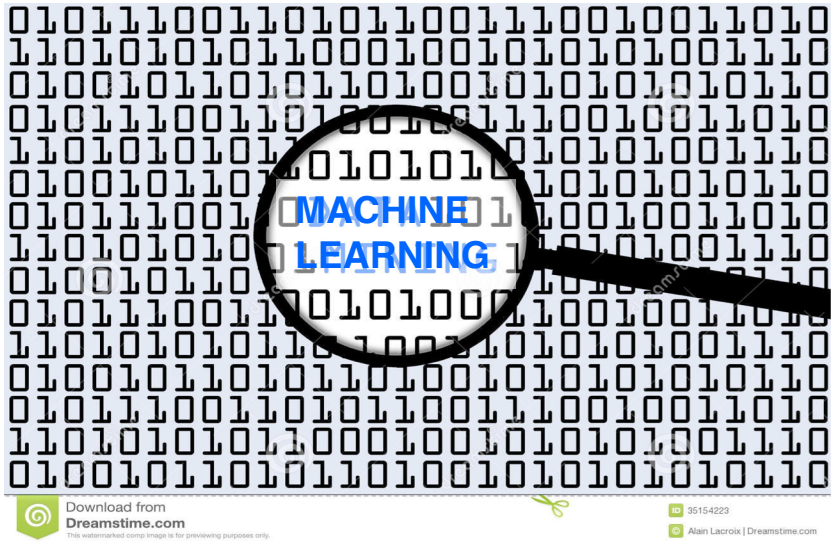
- **V's of Big Data (Doug Laney of Gartner)**
  - Volume
    - Vast amounts of data generated every second/minute/hour/day in digitized world
    - Petabytes (10^{15} bytes), exabytes (10^{18} bytes), and even more
  - Velocity
    - Speed at which data is being generated
    - Streaming data vs. static data
  - Variety
    - Different forms that data can be in
    - 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 and process Big Data.**
- **Need to analyze data to gain insights for decision-making.**

Machine Learning & Big Data

- **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.**
  - Distributed computing platforms, scalable algorithms, non-conventional data stores



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

ML Introduction – Key Points

- **Definition of machine learning (What)**
- **Reasons for doing machine learning (Why)**
- **Machine learning approaches (How)**
  - Classification
  - Regression
  - Cluster analysis
  - Association analysis
  - Supervised vs. unsupervised

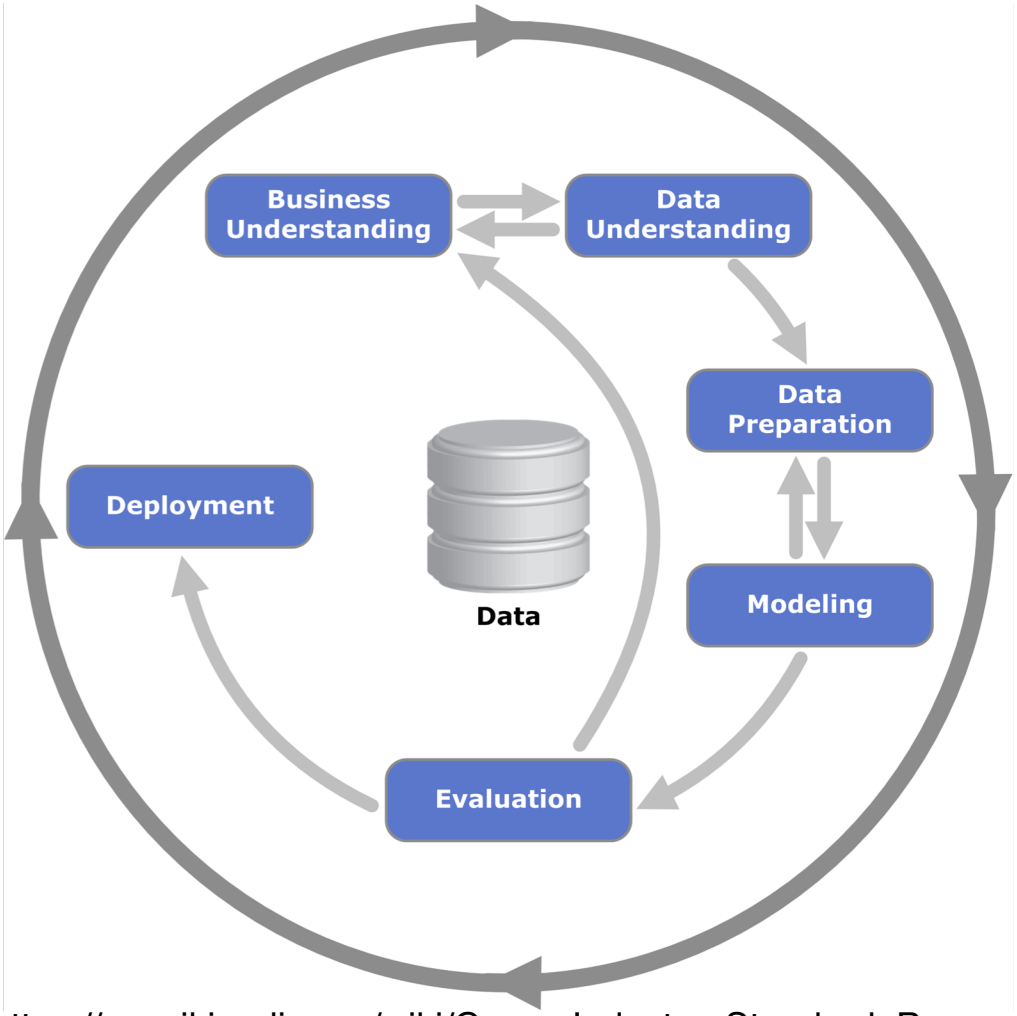
Questions?

Machine Learning Process

CRISP-DM

- **CRoss Industry Standard Process for Data Mining**
  - Process model describing steps in data mining process
- **Phases**
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment

CRISP-DM Diagram



Source: : 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

Exploratory Data Analysis

Source:*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 attributes to use
  - Create data for modeling

Data Quality

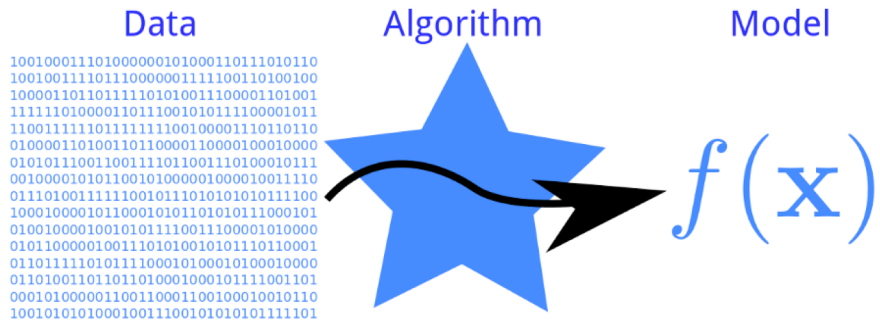
- **Data Quality Issues**
  - Missing Values
  - Duplicate Data
  - Inconsistent Data
  - Noise
  - Outliers

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

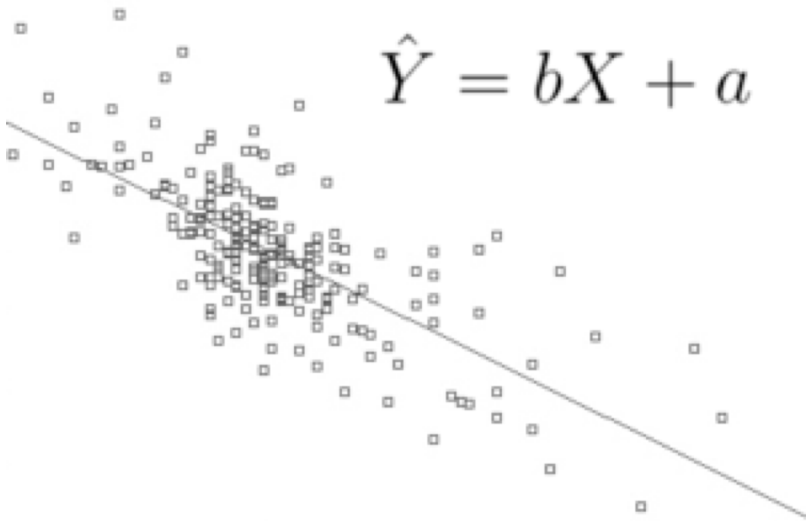
- **Determine type of problem**
  - Classification
  - Regression
  - Cluster analysis
  - Associative analysis
- **Select modeling technique(s) to use**
  - Decision tree
  - Linear regression
  - k-Means
  - etc.



Source: 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 measure
    - Confusion matrix
    - ROC chart
    - etc.
  - Evaluate model results w.r.t. success criteria.
    - Does model's performance meet success criteria?
    - Have all requirements been met?

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

- **Produce final report**
  - Summarize findings and recommend uses.
- **Deploy model**
  - Migrate model to production environment.
  - To integrate model into decision-making process.
- **Create plan for model monitoring & maintenance**
  - Monitoring model performance.
  - Plan for updating model.
- **Review and document project**

Model Deployment

- **Approaches**
  - Use machine learning tool for scoring
  - Generate model in Java, C, …
  - Generate model in SQL for database use
  - Use cloud-based service (SaaS)
- **PMML**
  - Predictive Model Markup Language
  - Used to share & migrate model between applications and platforms
- **Also referred to as “operationalization”.**

CRISP-DM: Iterative Process

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

Questions?