SDSC Summer Institute 2020

Machine Learning Overview Agenda

8:00 - 8:45 – Machine Learning Introduction & Data Exploration

8:45 - 9:10 – Data Preparation

9:10 - 9:25 – Break

9:25 - 10:00 – Classification

10:00 - 10:40 – Clustering

10:40 - 10:45 – Wrap-Up

Introduction to Machine Learning

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

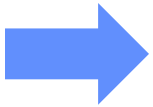
discover hidden patterns

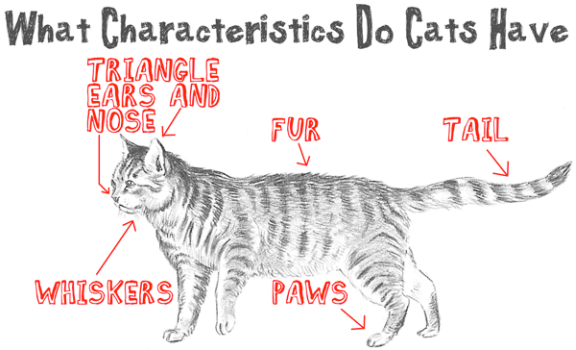
data-driven decisions

What is Machine Learning?

learning from data

no explicit programming



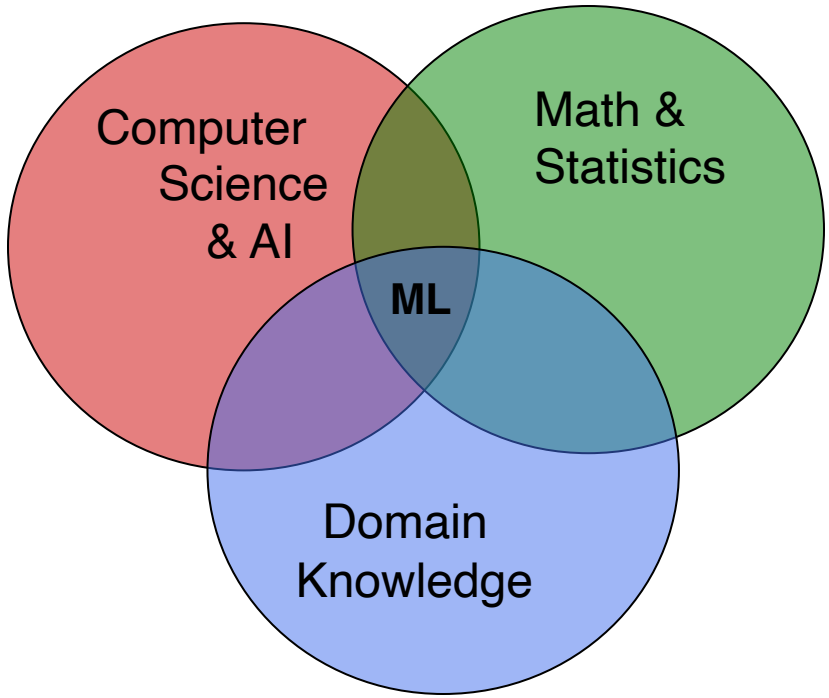


What is Machine Learning?

- **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?

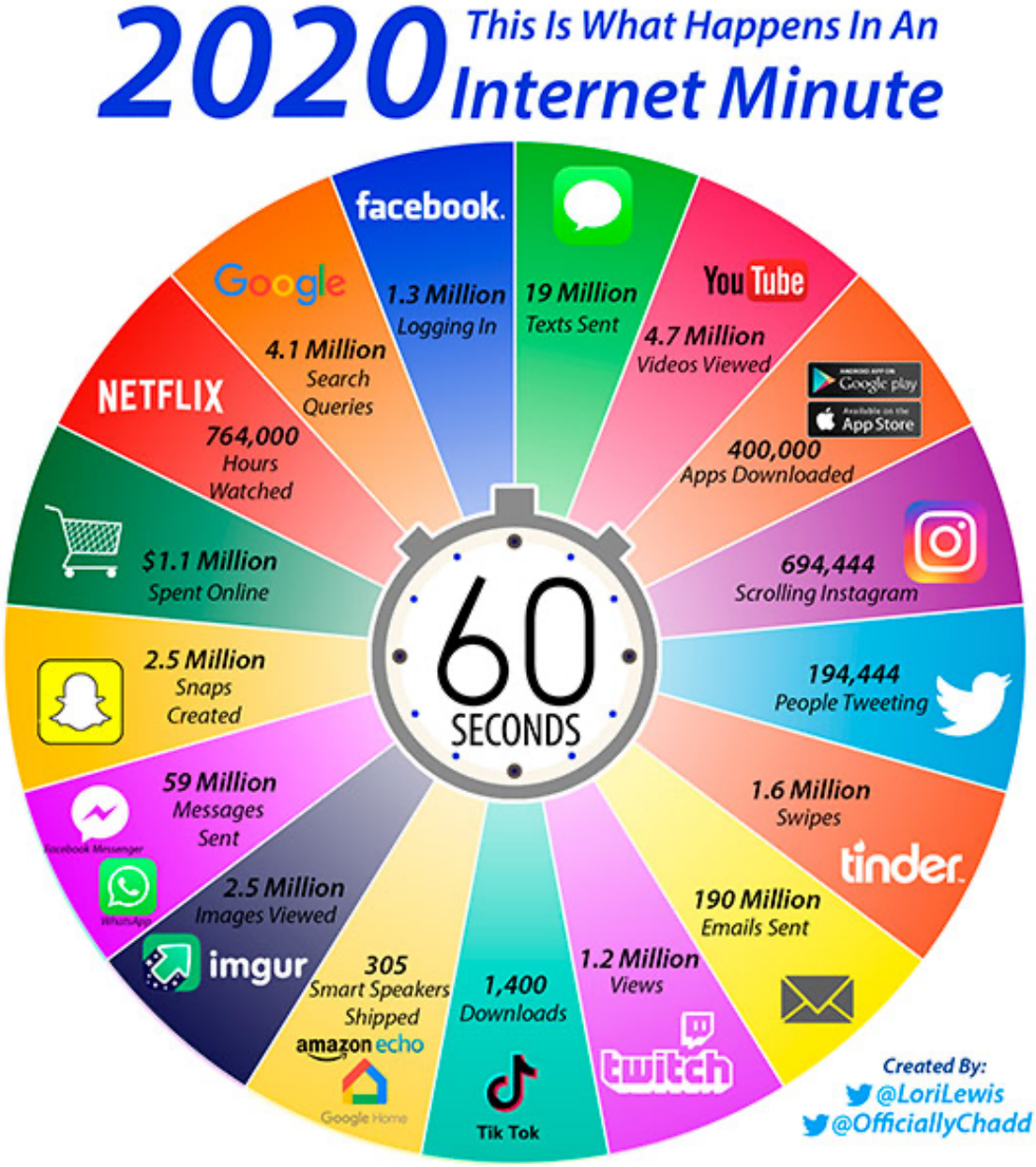
- **Advances in processing power, storage capacity, mobile computing, and interconnectivity are creating unprecedented data:**
  - User preferences and purchasing history on websites
  - Scientific data from remote sensors and instruments
  - Personal health data from wearable devices
  - Medical data 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 – 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/

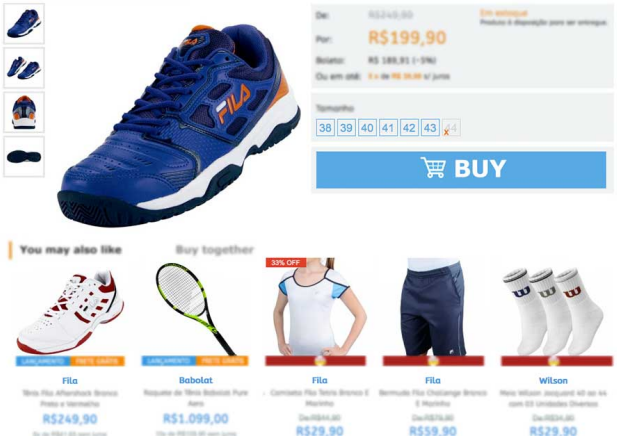
Why do Machine Learning?

- **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

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

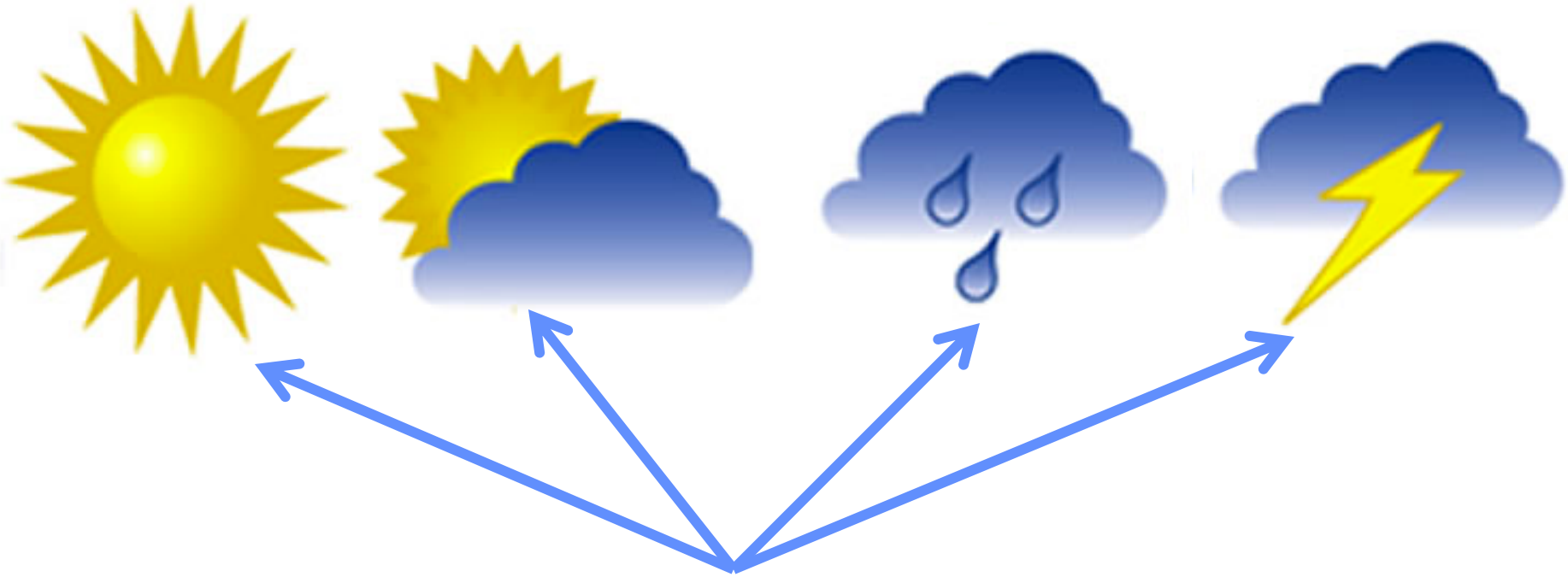


Machine Learning Approaches

- Classification
- Regression
- Cluster Analysis

Classification

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



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
- Estimate demand for a product based on time of year
- Predict amount of rain for a particular region

Cluster Analysis

- Goal: Organize similar items into groups

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

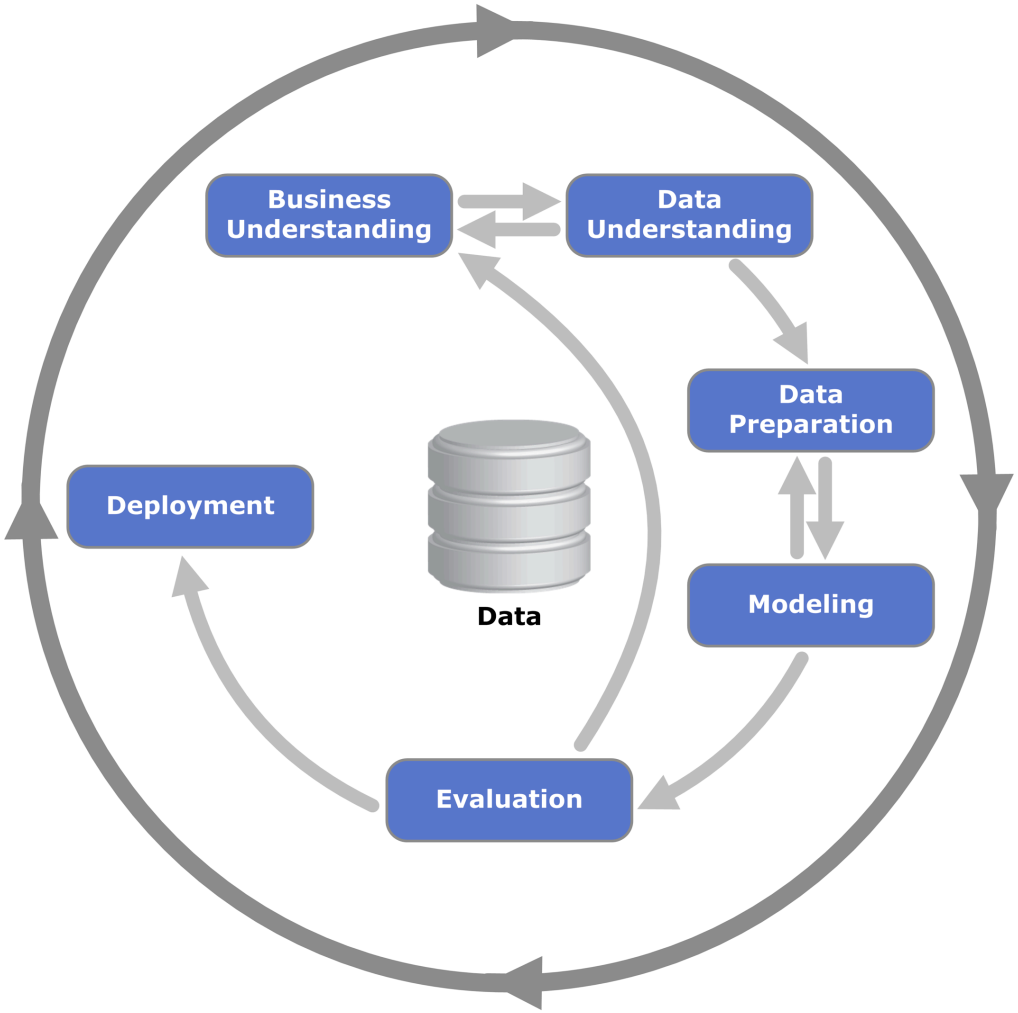
Supervised vs. Unsupervised

- **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 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



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

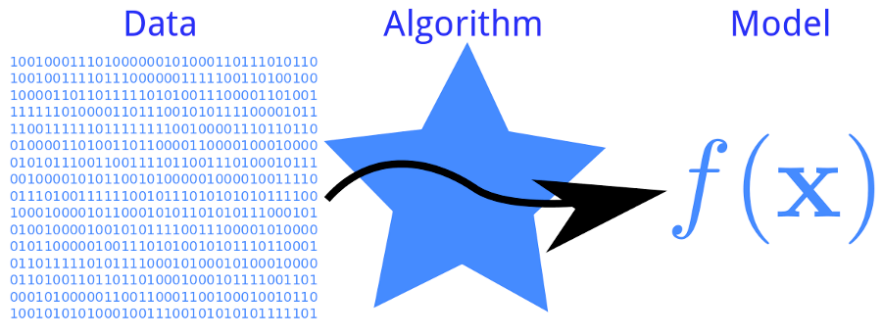
- **Goal:**
  - Prepare data to make it suitable for modeling
  - Also referred to as ‘data preprocessing’, ‘data munging’, ‘data wrangling’

- **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

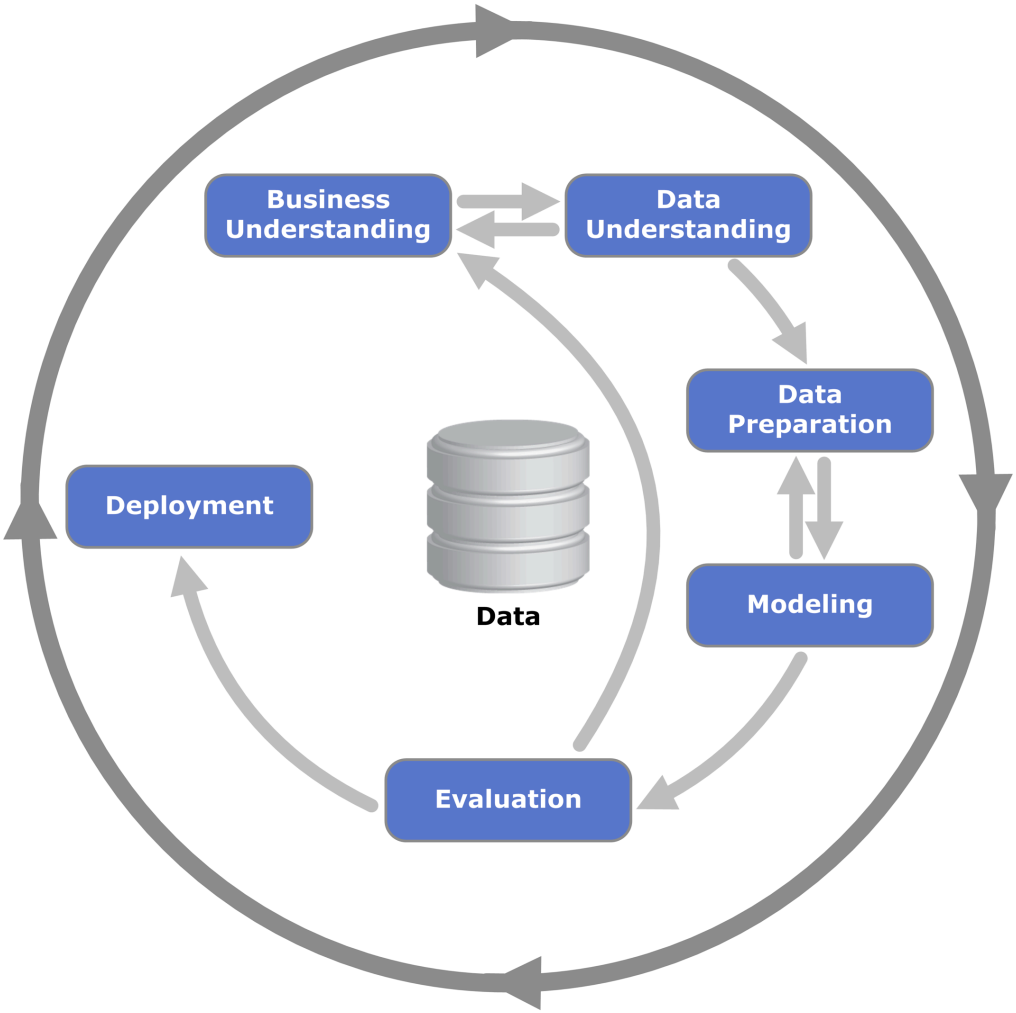
- **Assess model performance**
  - Determine metrics & methods to assess model results
    - Accuracy measures, confusion matrix, etc.
  - Evaluate model results w.r.t. success criteria
    - Does model's performance meet success criteria?
    - Have all requirements been met?
- **Make Go/No-Go decision**
  - Go: Deploy model
  - No-Go: Determine next steps

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

Phase 6: Deployment

- **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**

CRISP-DM: Iterative Process



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?