

2016 SDSC Summer Institute Machine Learning



Machine Learning Outline

AM Session

- Introduction
- Process
- R & RStudio
- Data Exploration
- Data Preparation

PM Session

- Classification
- Cluster Analysis
- Scaling R
- Spark MLlib

Introduction to Machine Learning

Mai H. Nguyen

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](https://en.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

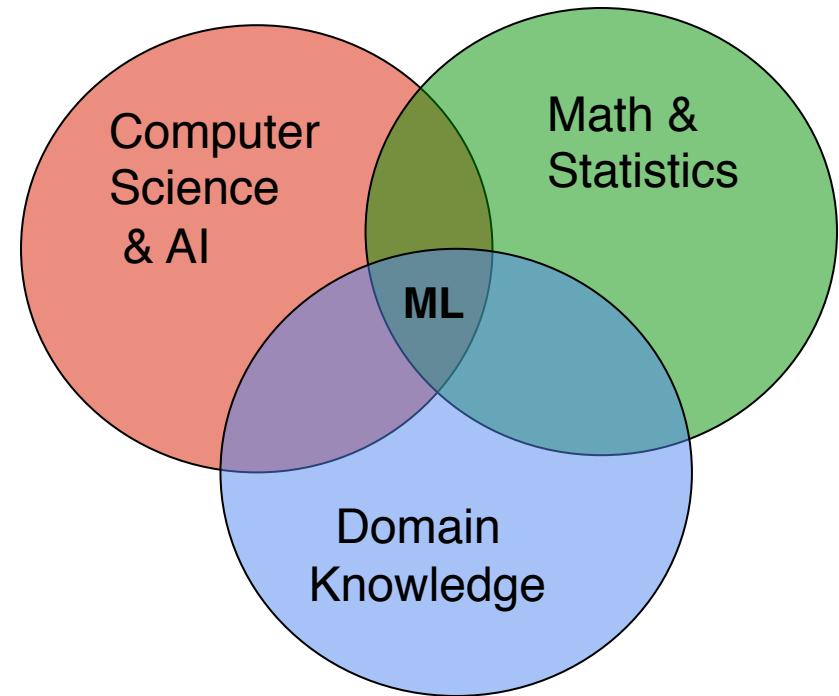
What is Machine Learning?

- **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?

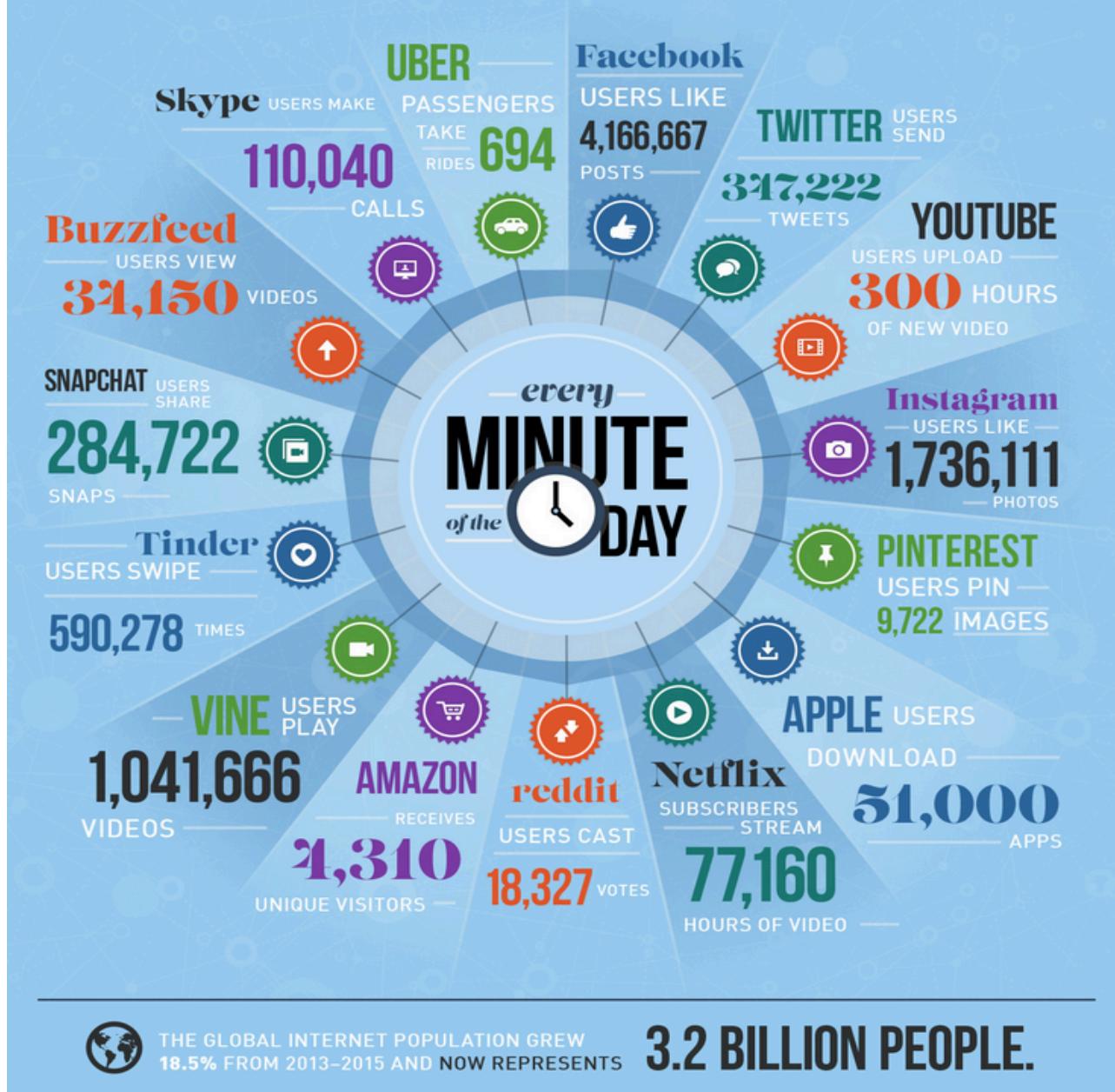
- **Advances in processing power, storage capacity, mobile computing, and interconnectivity are creating unprecedented data:**
 - User preferences and purchasing history on websites
 - Scientific data from remote sensors and instruments
 - Personal health data from wearable devices
 - Medical data from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

Source: www.domo.com



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/](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

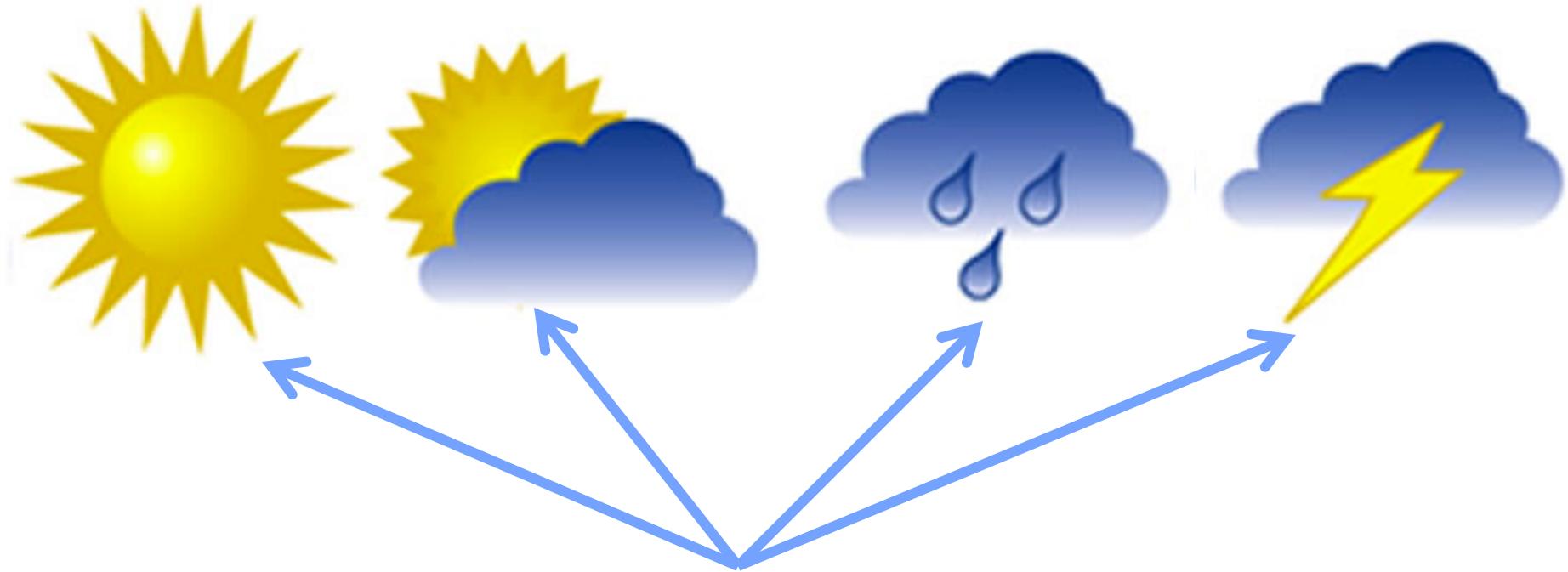
- Recommendations on websites
- Targeted ads on mobile apps
- Handwriting recognition
- Fraud detection
- Sentiment analysis
- Network intrusion detection
- Drug effectiveness analysis
- Crime pattern detection
- Self-driving cars

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/](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](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?

