# Introduction

# Machine Learning

- In the early stages, computers were taught to play simple games of tic-tac-toe and chess.
- Later, machines were given control of traffic lights, followed by military drones and missiles.
- Now, computers have become responsive and learn how to teach themselves.
- ***The goal of today's machine learning is not to create an artificial brain, but rather to assist us in making sense of the world's massive data stores.***

# Machine Learning

- ***The field of machine learning provides a set of algorithms that transform data into actionable knowledge.***
- Between databases and sensors, many aspects of our lives are recorded.
- Governments, businesses, and individuals are recording and reporting information, from the monumental to the mundane.
- Weather sensors record temperature and pressure data, surveillance cameras watch sidewalks and subway tunnels, and all manner of electronic behaviors are monitored: transactions, communications, friendships, and many others

# Machine Learning

- This deluge of data has led some to state that we have entered an era of **Big Data,** but this may be a bit of a misnomer. Human beings have always been surrounded by large amounts of data. What makes the current era unique is that we have vast amounts of *recorded data, much of which can be directly accessed by computers.*
- Larger and more interesting data sets are increasingly accessible at the tips of our fingers, only a web search away. This wealth of information has the potential to inform action, given a systematic way of making sense from it all.

# Machine Learning

- The field of study interested in the development of computer algorithms to transform data into intelligent action is known as **machine learning.**

# Machine Learning vs. Data Mining

- A closely related sibling of machine learning, **data mining, is concerned with the** generation of novel insights from large databases. As the implies, data mining involves a systematic hunt for nuggets of actionable intelligence.
- Although there is some disagreement over how widely machine learning and data mining overlap, a potential point of distinction is that machine learning focuses on teaching computers how to use data to solve a problem, while data mining focuses on teaching computers to identify patterns that humans then use to solve a problem.

# Machine Learning vs. Data Mining

- Virtually all data mining involves the use of machine learning, but not all machine learning involves data mining.
- For example, you might apply machine learning to data mine automobile traffic data for patterns related to accident rates; on the other hand, if the computer is learning how to drive the car itself, this is purely machine learning without data mining.

# Uses and abuses of machine learning

- Some people have speculated that computer intelligence will replace humans in many information technology occupations, just as machines replaced humans in the fields, and robots replaced humans on the assembly line.
- The truth is that even as machines reach such impressive milestones, they are still relatively limited in their ability to thoroughly understand a problem. They are pure intellectual horsepower without direction. A computer may be more capable than a human of finding subtle patterns in large databases, but it still needs a human to motivate the analysis and turn the result into meaningful action.

# Uses and abuses of machine learning

- Machines are not good at asking questions, or even knowing what questions to ask. They are much better at answering them, provided the question is stated in a way the computer can comprehend.
- Present-day machine learning algorithms partner with people much like a bloodhound partners with its trainer; the dog's sense of smell may be many times stronger than its master's, but without being carefully directed, the hound may end up chasing its tail.

# Machine learning successes

- Machine learning is most successful when it augments rather than replaces the specialized knowledge of a subject-matter expert.

# Machine learning successes

- A survey of recent success stories includes several prominent applications:
  - Identification of unwanted spam messages in e-mail
  - Segmentation of customer behavior for targeted advertising
  - Forecasts of weather behavior and long-term climate changes
  - Reduction of fraudulent credit card transactions
  - Actuarial estimates of financial damage of storms and natural disasters
  - Prediction of popular election outcomes
  - Development of algorithms for auto-piloting drones and self-driving cars
  - Optimization of energy use in homes and office buildings
  - Projection of areas where criminal activity is most likely
  - Discovery of genetic sequences linked to diseases

# The limits of machine learning

- Although machine learning is used widely and has tremendous potential, it is important to understand its limits.
- Machine learning, at this time, is not in any way a substitute for a human brain. It has very little flexibility to extrapolate outside of the strict parameters it learned and knows no common sense.

# The limits of machine learning

- Without a lifetime of past experiences to build upon, computers are also limited in their ability to make simple common sense inferences about logical next steps.
- Take, for instance, the banner advertisements seen on many web sites. These may be served, based on the patterns learned by data mining the browsing history of millions of users. According to this data, someone who views the websites selling shoes should see advertisements for shoes, and those viewing websites for mattresses should see advertisements for mattresses.
- The problem is that this becomes a never-ending cycle in which additional shoe or mattress advertisements are served rather than advertisements for shoelaces and shoe polish, or bed sheets and blankets.

# Machine learning ethics

- Due to the relative youth of machine learning as a discipline and the speed at which it is progressing, the associated legal issues and social norms are often quite uncertain and constantly in flux.
- Caution should be exercised while obtaining or analyzing data in order to avoid breaking laws, violating terms of service or data use agreements, and abusing the trust or violating the privacy of customers or the public.

# Machine learning ethics

- Retailers routinely use machine learning for advertising, targeted promotions, inventory management, or the layout of the items in the store. Many have even equipped checkout lanes with devices that print coupons for promotions based on the customer's buying history. In exchange for a bit of personal data, the customer receives discounts on the specific products he or she wants to buy. At first, this appears relatively harmless. But consider what happens when this practice is taken a little bit further.

# Machine learning ethics -Story

- One possibly apocryphal tale concerns a large retailer in the U.S. that employed machine learning to identify expectant mothers for coupon mailings. The retailer hoped that if these mothers-to-be received substantial discounts, they would become loyal customers, who would later purchase profitable items like diapers, baby formula, and toys.
- Equipped with machine learning methods, the retailer identified items in the customer purchase history that could be used to predict with a high degree of certainty, not only whether a woman was pregnant, but also the approximate timing for when the baby was due.

# Machine learning ethics -Story

- After the retailer used this data for a promotional mailing, an angry man contacted the chain and demanded to know why his teenage daughter received coupons for maternity items. He was furious that the retailer seemed to be encouraging teenage pregnancy! As the story goes, when the retail chain's manager called to offer an apology, it was the father that ultimately apologized because, after confronting his daughter, he discovered that she was indeed pregnant!
- Whether completely true or not, the lesson learned from the preceding tale is that common sense should be applied before blindly applying the results of a machine learning analysis. This is particularly true in cases where sensitive information such as health data is concerned. With a bit more care, the retailer could have foreseen this scenario, and used greater discretion while choosing how to reveal the pattern its machine learning analysis had discovered.

- Certain jurisdictions may prevent you from using racial, ethnic, religious, or other protected class data for business reasons. Keep in mind that excluding this data from your analysis may not be enough, because machine learning algorithms might inadvertently learn this information independently.
- For instance, if a certain segment of people generally live in a certain region, buy a certain product, or otherwise behave in a way that uniquely identifies them as a group, some machine learning algorithms can infer the protected information from these other factors.
- In such cases, you may need to fully "de-identify" these people by excluding any *potentially* identifying data in addition to the protected information. Apart from the legal consequences, using data inappropriately may hurt the bottom line. Customers may feel uncomfortable or become spooked if the aspects of their lives they consider private are made public.
- In recent years, several high-profile web applications have experienced a mass exodus of users who felt exploited when the applications' terms of service agreements changed, and their data was used for purposes beyond what the users had originally agreed upon.
- The fact that privacy expectations differ by context, age cohort, and locale adds complexity in deciding the appropriate use of personal data. It would be wise to consider the cultural implications of your work before you begin your project.

# How machines learn?

- A formal definition of machine learning proposed by computer scientist Tom M. Mitchell states that

  - A Computer program is said to learn from experience E with respect to some class of tasks T and the performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

- a machine learns whenever it is able to utilize its experience such that its performance improves on similar experiences in the future.

# How machines learn?

- Regardless of whether the learner is a human or machine, the basic learning process is similar. It can be divided into four interrelated components:
  - **Data storage** utilizes observation, memory, and recall to provide a factual basis for further reasoning.
  - **Abstraction** involves the translation of stored data into broader representations and concepts.
  - **Generalization** uses abstracted data to create knowledge and inferences that drive action in new contexts.
  - **Evaluation** provides a feedback mechanism to measure the utility of learned knowledge and inform potential improvements.

# Machine learning in practice

- Regardless of the task at hand, any machine learning algorithm can be deployed by following these steps:

1. **Data collection:** The data collection step involves gathering the learning material an algorithm will use to generate actionable knowledge. In most cases, the data will need to be combined into a single source like a text file, spreadsheet, or database.
2. **Data exploration and preparation:** The quality of any machine learning project is based largely on the quality of its input data. Thus, it is important to learn more about the data and its nuances during a practice called data exploration. Additional work is required to prepare the data for the learning process. This involves fixing or cleaning so-called "messy" data, eliminating unnecessary data, and recoding the data to conform to the learner's expected inputs.

# Machine learning in practice

3. **Model training:** By the time the data has been prepared for analysis, you are likely to have a sense of what you are capable of learning from the data. The specific machine learning task chosen will inform the selection of an appropriate algorithm, and the algorithm will represent the data in the form of a model.
4. **Model evaluation:** Because each machine learning model results in a biased solution to the learning problem, it is important to evaluate how well the algorithm learns from its experience. Depending on the type of model used, you might be able to evaluate the accuracy of the model using a test dataset or you may need to develop me asures of performance specific to the intended application.
5. **Model improvement:** If better performance is needed, it becomes necessary to utilize more advanced strategies to augment the performance of the model. Sometimes, it may be necessary to switch to a different type of model altogether. You may need to supplement your data with additional data or perform additional preparatory work as in step two of this process

# Types of input data

- If a feature represents a characteristic measured in numbers, it is unsurprisingly called **numeric.**
- Alternatively, if a feature is an attribute that consists of a set of categories, the feature is called **categorical or nominal.**
- A special case of categorical variables is called **ordinal,** which designates a nominal variable with categories falling in an ordered list.

# Types of machine learning algorithms

- Predictive Model
- Descriptive Model

# Predictive Model

- A **predictive model** is used for tasks that involve, as the name implies, the prediction of one value using other values in the dataset.
- The learning algorithm attempts to discover and model the relationship between the **target feature (the feature being** predicted) and the other features.

# Supervised Learning

- Because predictive models are given clear instruction on what they need to learn and how they are intended to learn it, the process of training a predictive model is known as **supervised learning.**
- The supervision does not refer to human involvement, but rather to the fact that the *target values* provide a way for the learner to know how well it has learned the desired task.
- Stated more formally, given a set of data, a supervised learning algorithm attempts to optimize a function (the model) to find the combination of feature values that result in the target output.

# Classification

- The often used supervised machine learning task of predicting which category an example belongs to is known as **classification.**
- Using a classifier one can predict whether:
  - An e-mail message is spam
  - A person has cancer
  - A football team will win or lose
  - An applicant will default on a loan

# Classification

- In classification, the target feature to be predicted is a categorical feature known as the **class,** and is divided into categories called **levels.**
- A class can have two or more levels, and the levels may or may not be ordinal.
- Because classification is so widely used in machine learning, there are many types of classification algorithms, with strengths and weaknesses suited for different types of input data.

# Classification

- Supervised learners can also be used to **predict numeric data** such as income, laboratory values, test scores, or counts of items.
- To predict such numeric values, a common form of numeric prediction fits linear regression models to the input data.

# Descriptive Model

- A **descriptive model** is used for tasks that would benefit from the insight gained from summarizing data in new and interesting ways
- Because there is no target to learn, the process of training a descriptive model is called **unsupervised learning**

# Descriptive Model

- The descriptive modeling task called **pattern discovery** is used to identify useful associations within data.
- Pattern discovery is often used for **market basket analysis** on retailers' transactional purchase data. Here, the goal is to identify items that are frequently purchased together, such that the learned information can be used to refine marketing tactics.
- For instance, if a retailer learns that swimming trunks are commonly purchased at the same time as sunglasses, the retailer might reposition the items more closely in the store or run a promotion to "up-sell“ customers on associated items.

# Descriptive Model

- The descriptive modeling task of dividing a dataset into homogeneous groups is called **clustering.**
- This is sometimes used for segmentation analysis that identifies groups of individuals with similar behavior or demographic information, so that advertising campaigns could be tailored for particular audiences.
- Although the machine is capable of identifying the clusters, human intervention is required to interpret them.
- For example, given five different clusters of shoppers at a grocery store, the marketing team will need to understand the differences among the groups in order to create a promotion that best suits each group

| Model | Learning task |
|---|---|
| **Supervised Learning Algorithms** | |
| Nearest Neighbor | Classification |
| Naive Bayes | Classification |
| Decision Trees | Classification |
| Classification Rule Learners | Classification |
| Linear Regression | Numeric prediction |
| Regression Trees | Numeric prediction |
| Model Trees | Numeric prediction |
| Neural Networks | Dual use |
| Support Vector Machines | Dual use |
| **Unsupervised Learning Algorithms** | |
| Association Rules | Pattern detection |
| k-means clustering | Clustering |
| **Meta-Learning Algorithms** | |
| Bagging | Dual use |
| Boosting | Dual use |
| Random Forests | Dual use |

# Summary

- Using the self-learning algorithms from the field of machine learning, we can turn this data into knowledge.
- Instead of requiring humans to manually derive rules and build models from analyzing large amounts of data, machine learning offers a more efficient alternative for capturing the knowledge in data to gradually improve the performance of predictive models, and make data-driven decisions

# Summary

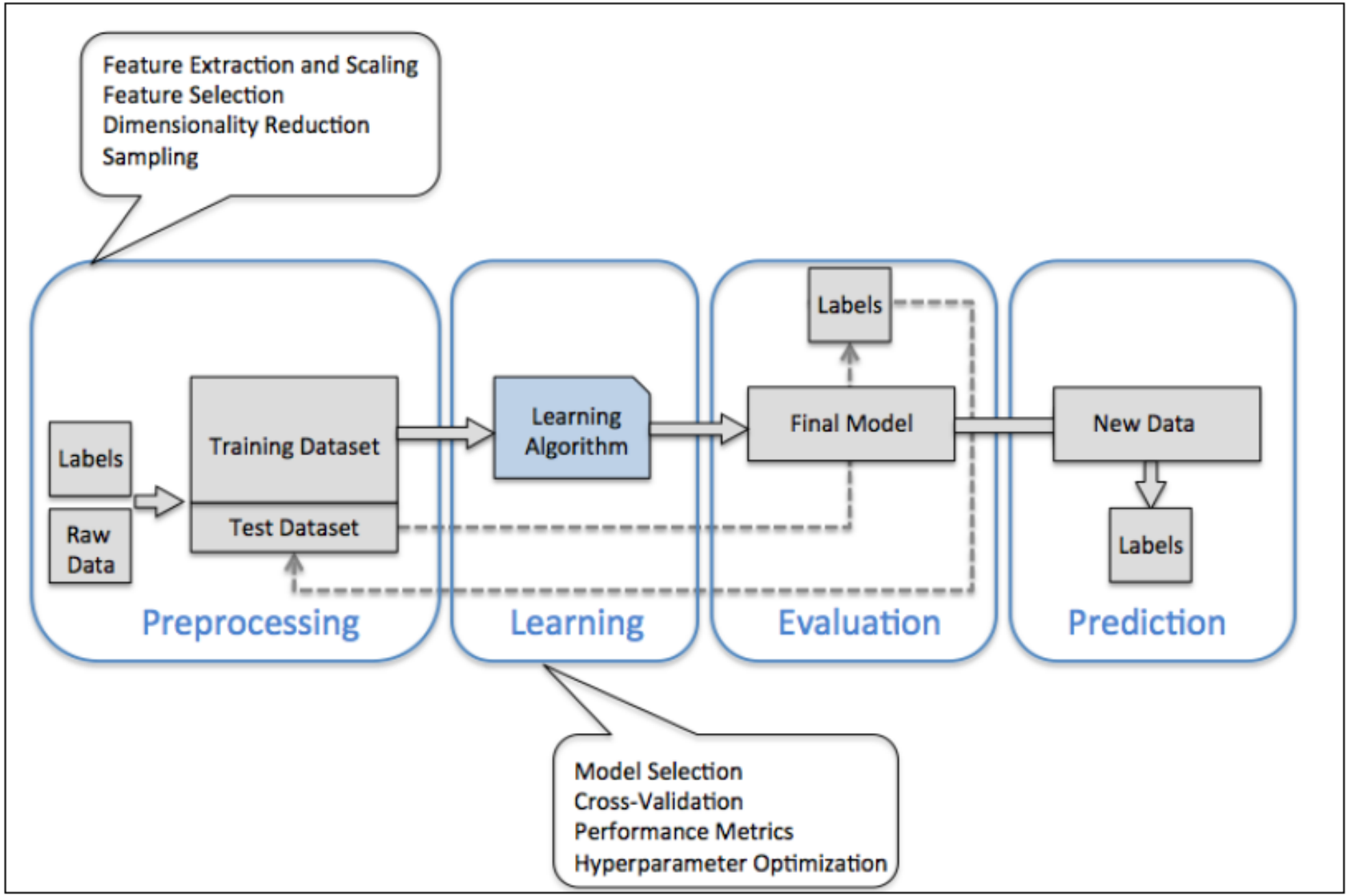
- The main goal in **supervised learning** is to learn a model from labeled training data that allows us to make predictions about unseen or future data. Here, the term supervised refers to a set of samples where the desired output signals (labels) are already known.
- Classification is a subcategory of supervised learning where the goal is to predict the categorical class labels of new instances based on past observations. Those class labels are discrete and unordered values .
- Another subcategory of supervised learning is *regression,* where the outcome signal is a continuous value.

# Summary

- In unsupervised learning, we deal with unlabeled data or data of unknown structure.
- Using unsupervised learning techniques, we are able to explore the structure of our data to extract meaningful information without the guidance of a known outcome variable.

# Summary

- Clustering is an exploratory data analysis technique that allows us to organize a pile of information into meaningful subgroups (clusters) without having any prior knowledge of their group memberships.
- Each cluster that may arise during the analysis defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters, which is why clustering is also sometimes called "unsupervised classification."


Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Learning Algorithm
Labels
Final Model
New Data
Labels
Preprocessing
Learning
Evaluation
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization