Introduction to Machine Learning in Geosciences
Introduction II
GEO371T/GEO391

Mrinal K. Sen
Geosciences
UT Austin

September 2, 2021

Content

- Motivation
- Human Learning/How do we learn to learn?
- Artificial intelligence
- Data Analytics
- What is Machine Learning?
- History of Machine Learning
- Types of ML: Supervised, Unsupervised, Semi-supervised and Reinforced Learning
- Classification and Regression
- ML background linear algebra, statistics, computing
- Learning under uncertainty!
- shallow and deep learning

Human Learning

"The truth is that a human is just a brief algorithm — 10,247 lines. They are deceptively simple. Once you know them, their behavior is quite predictable."
— Westworld

What is intelligence?

- Intelligence is an umbrella term used to describe a property of the mind that encompasses many related abilities, such as the capacities
  - to reason,
  - to plan,
  - to solve problems,
  - to think abstractly,
  - to comprehend ideas,
  - to use language, and
  - **to learn and to solve problems**

[from Wikipedia]

Intelligence

- Problem solving is to find the "best" solution in the problem space.
- Reasoning is to interpret or justify solutions or subsolutions.
- Planning is to find ways for solving the problem.
- Thinking abstractly is to simulate the problem solving process inside the system (brain).
- Idea/language comprehension is a way (or means) for data/problem/knowledge representation;
- Learning is the process to find better ways for solving a problem (or a class of problems).

[from Wikipedia]

Human Learning

- We learn from the things that happen to us - our experiences. For example, we learned that lightning is followed by thunder, we learned not to tell lies because it can cause us to lose our credibility and to lose our friends, or that we learned how to dance by watching others demonstrate dance steps to us.
- We can say that we have learned these things because we have acquired appropriate responses for them - we cover our ears when lightning strikes, we try to avoid telling lies, and we dance.
- Learning is acquiring relatively permanent change in behavior through experience. We experience things and learn to modify our behaviors based on what we know.
- Learning applies not just to humans, but also to animals.

https://general-psychology.weebly.com/how-do-we-learn.html

Human Learning

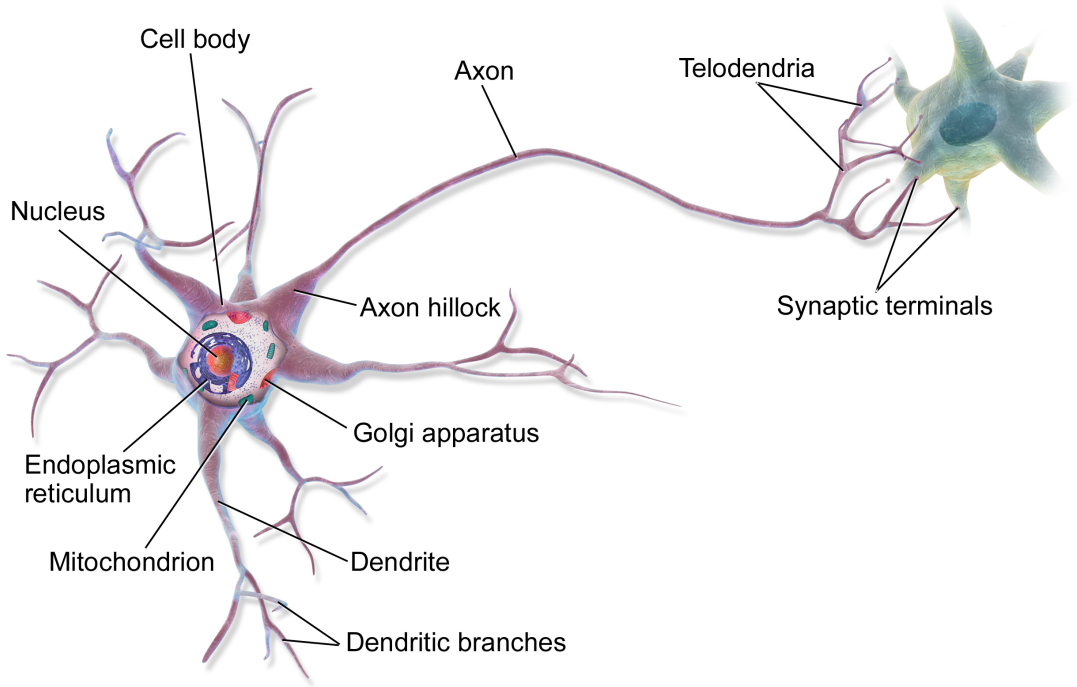
- **Observational Learning** is learning by watching others engage in different behaviors. You probably have learned to dance by watching your teacher demonstrate some dance steps to you.
- **Associative Learning** is learning by establishing connections between events. Conditioning is the method for teaching associations: classical and operant conditioning.
- **Classical conditioning** is the method of teaching associations between two different stimuli. We learn the connection between lightning and thunder because they almost always occur together. Because of this, whenever we see lightning, we cover our ears in anticipation of thunder.
- **Operant conditioning** is the method of teaching associations between behaviors and consequences. Operant conditioning uses rewards and punishments to strengthen or weaken behaviors. For example - the connection between telling lies and losing credibility and friends.

https://general-psychology.weebly.com/how-do-we-learn.html

Human Learning

Human Learning: Physiology

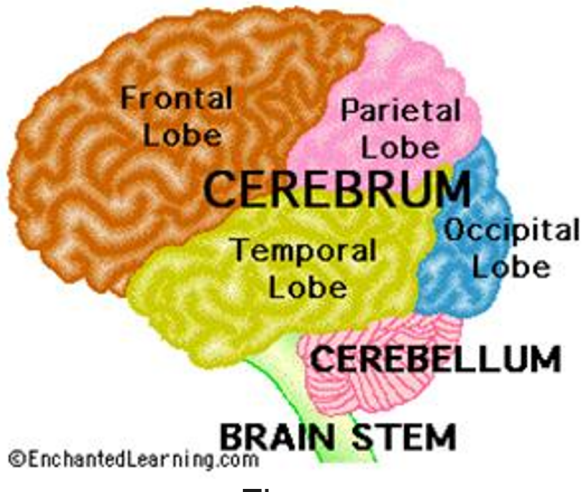
- Neuron, neurone (old British spelling) or nerve cell, is an electrically excitable cell[1] that communicates with other cells via specialized connections called synapses. It is the main component of nervous tissue in all animals except sponges and placozoa. Plants and fungi do not have nerve cells.
- A group of connected neurons is called a neural circuit.



https://trainingindustry.com/articles/content-development/how-the-brain-learns/

Human Learning: Physiology

- The brain acts as a dense network of fiber pathways consisting of approximately 100 billion 10^{10} **neurons**.
- Three principal parts – **stem**, **cerebellum** and **cerebrum**
- the cerebrum is most important in learning, since this is where higher-ordered functions like memory and reasoning occur. Each area of the cerebrum specializes in a function – sight, hearing, speech, touch, short-term memory, long-term memory, language and reasoning abilities are the most important for learning.



https://trainingindustry.com/articles/content-development/how-the-brain-learns/

Artificial Intelligence (AI)

- Artificial intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing (NLP), speech recognition and machine vision.
- Artificial Intelligence (AI) is the branch of computer sciences that emphasizes the development of intelligence machines, thinking and working like humans.
- Artificial Intelligence (AI) involves using computers to do things that traditionally require human intelligence. This means creating algorithms to classify, analyze, and draw predictions from data. It also involves acting on data, learning from new data, and improving over time.

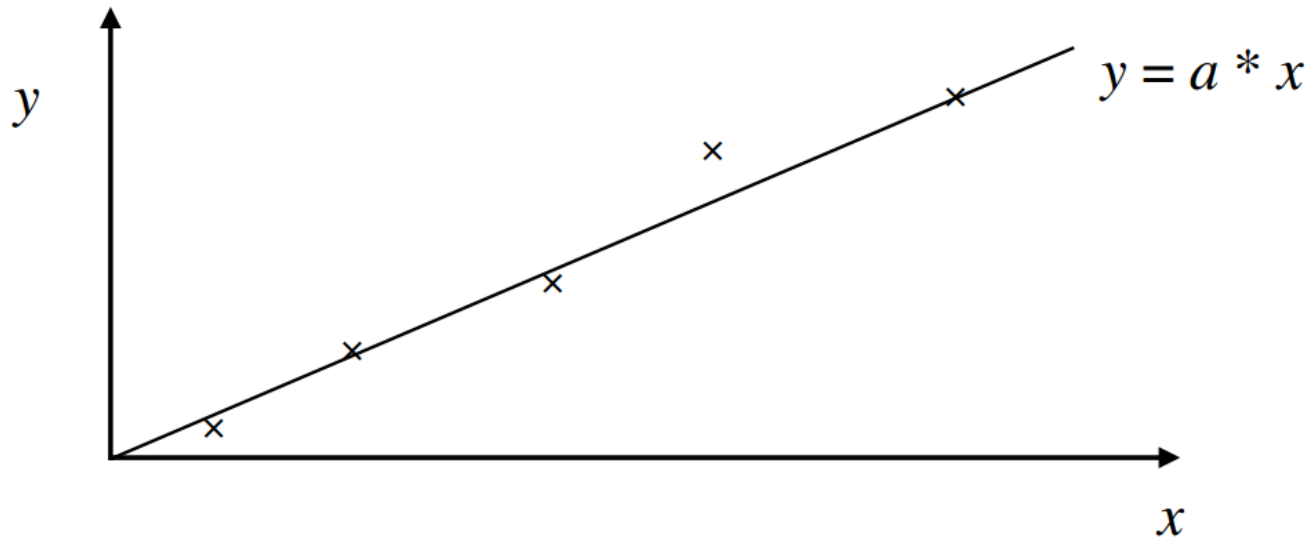
https://medium.com/mytake/artificial-intelligence-explained-in-simple-english-part-1-2-1b28c1f762cf

Data Analytics and Data Analysis

- The primary difference between analytics and analysis is a matter of scale, as data analytics is a broader term of which data analysis is a subcomponent.
- **Data analysis** refers to the process of examining, transforming and arranging a given data set in specific ways in order to study its individual parts and extract useful information.
- **Data analytics** is an overarching science or discipline that encompasses the complete management of data. This not only includes analysis, but also data collection, organisation, storage, and all the tools and techniques used.
- While data analysts and data scientists both work with data, the main difference lies in what they do with it. Data analysts examine large data sets to identify trends, develop charts, and create visual presentations to help businesses make more strategic decisions.
- **Data scientists**, on the other hand, design and construct new processes for data modeling and production using prototypes, algorithms, predictive models, and custom analysis.

Empirical Inference

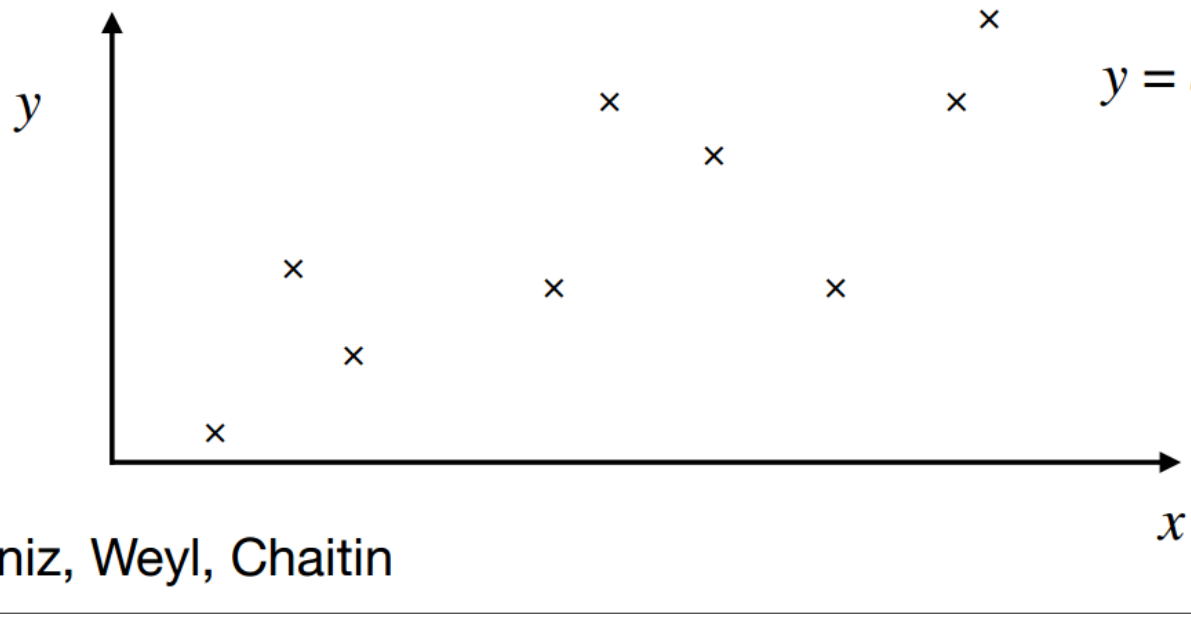
- Drawing conclusions from empirical data (observations, measurements)
- Example 1: scientific inference



slide by Bernhard Schölkopf

Empirical Inference

- Drawing conclusions from empirical data (observations, measurements)
- Example 1: scientific inference

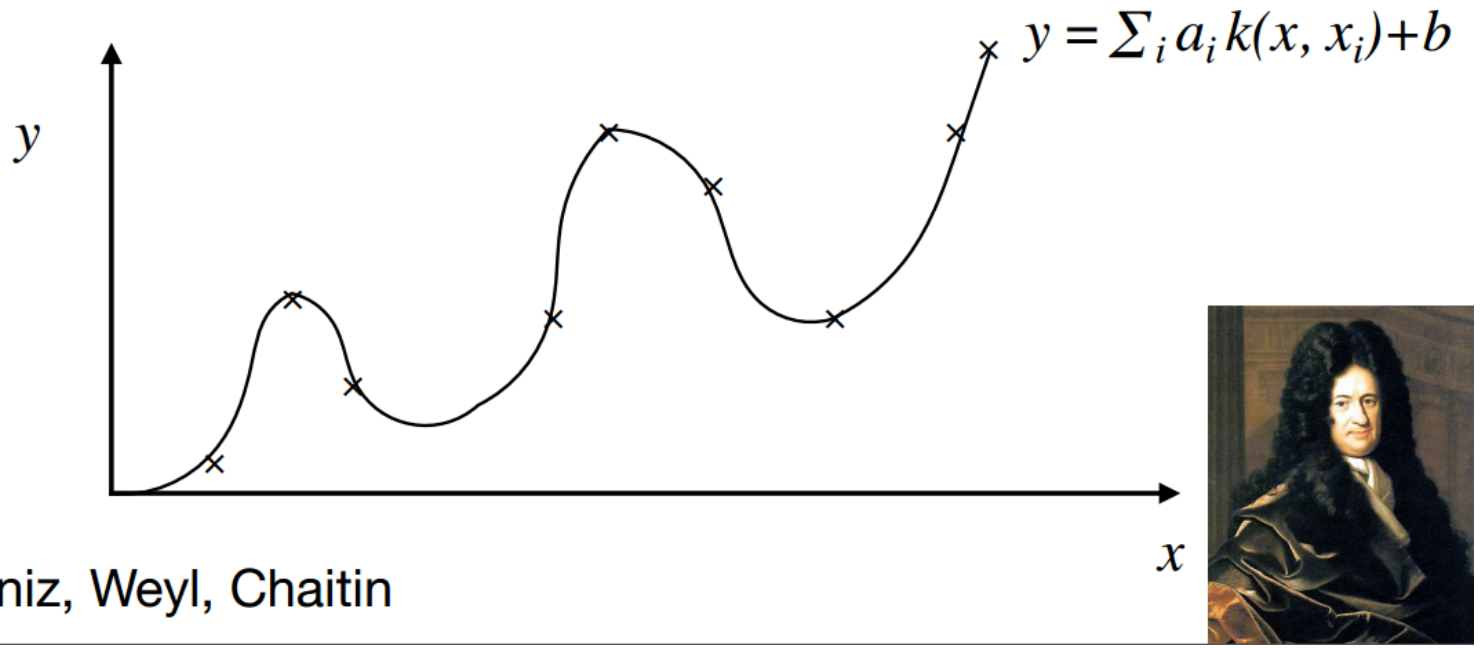


slide by Bernhard Schölkopf

Leibniz, Weyl, Chaitin

Empirical Inference

- Drawing conclusions from empirical data (observations, measurements)
- Example 1: scientific inference



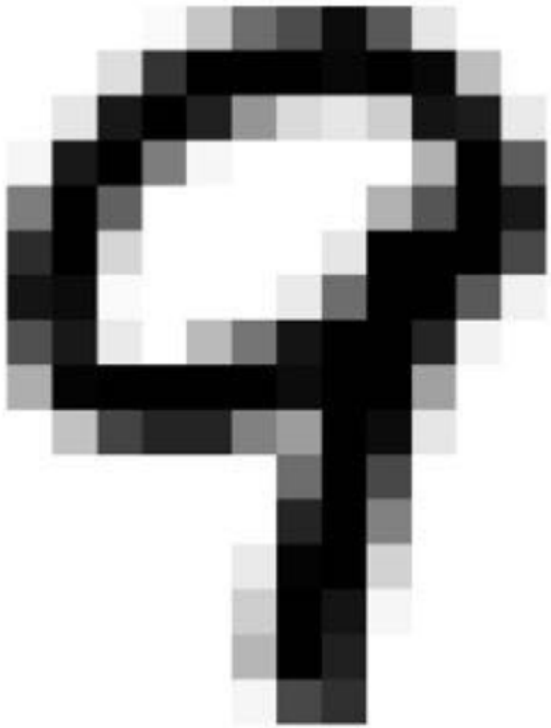
Leibniz, Weyl, Chaitin

slide by Bernhard Schölkopf

Empirical Inference

- Example 2: perception

slide by Bernhard Schölkopf

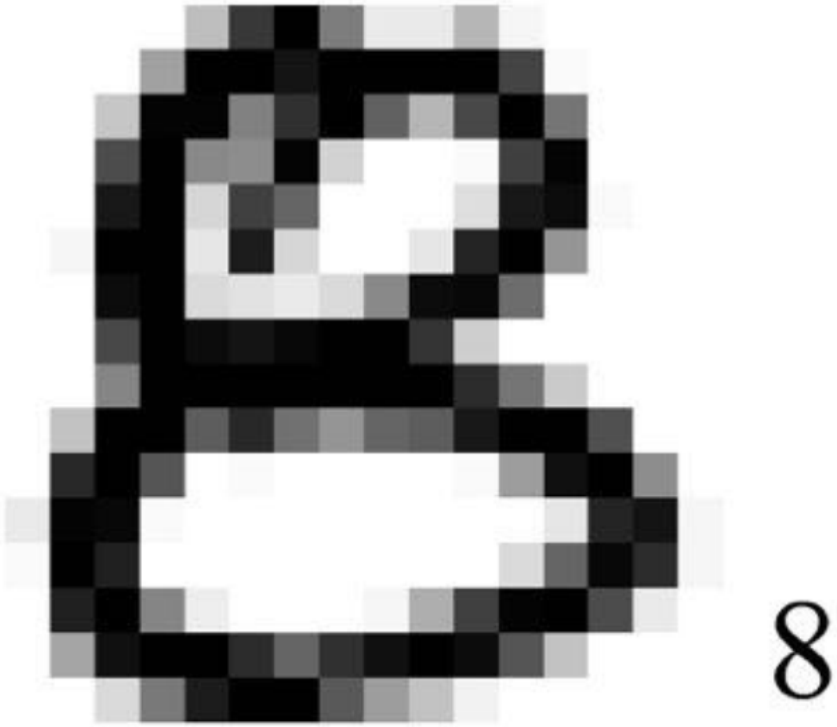


9

slide by Bernhard Schölkopf

8

slide by Bernhard Schölkopf



8

slide by Bernhard Schölkopf

slide by Bernhard Schölkopf

Empirical Inference

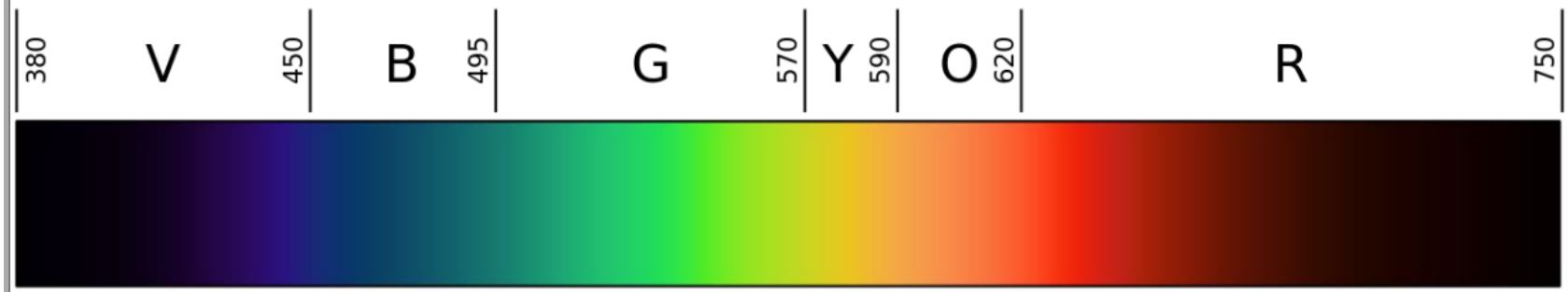
- Example2: perception

"The brain is nothing but a statistical decision organ"

H. Barlow

slide by Bernhard Schölkopf

Color Perception



Hard Inference Problems

- High dimensionality — consider many factors simultaneously to find regularity
- Complex regularities — nonlinear; nonstationary, etc.
- Little prior knowledge — e.g. no mechanistic models for the data
- Need large data sets — processing requires computers and automatic inference methods

slide by Bernhard Schölk

What is machine learning?

Example: Netflix Challenge

- Goal: Predict how a viewer will rate a movie
- 10% improvement = 1 million dollars

slide by Yaser Abu-Mostapha

Example: Netflix Challenge

- Goal: Predict how a viewer will rate a movie
- 10% improvement = 1 million dollars
- Essence of Machine Learning:
  - A pattern exists
  - We cannot pin it down mathematically
  - We have data on it

slide by Yaser Abu-Mostapha

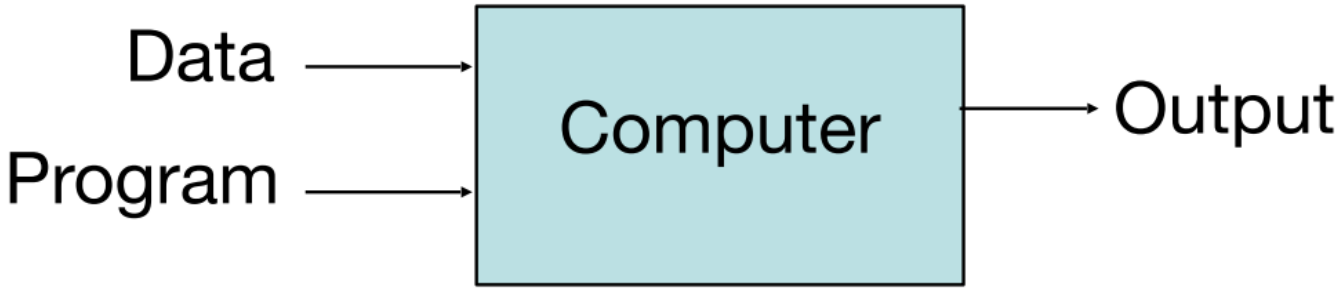
What is Machine Learning?

- [Arthur Samuel, 1959]
  - Field of study that gives computers
  - the ability to learn without being explicitly programmed
- [Kevin Murphy] algorithms that
  - automatically detect patterns in data
  - use the uncovered patterns to predict future data or other outcomes of interest
- [Tom Mitchell] algorithms that
  - improve their performance (P)
  - at some task (T)
  - with experience (E)

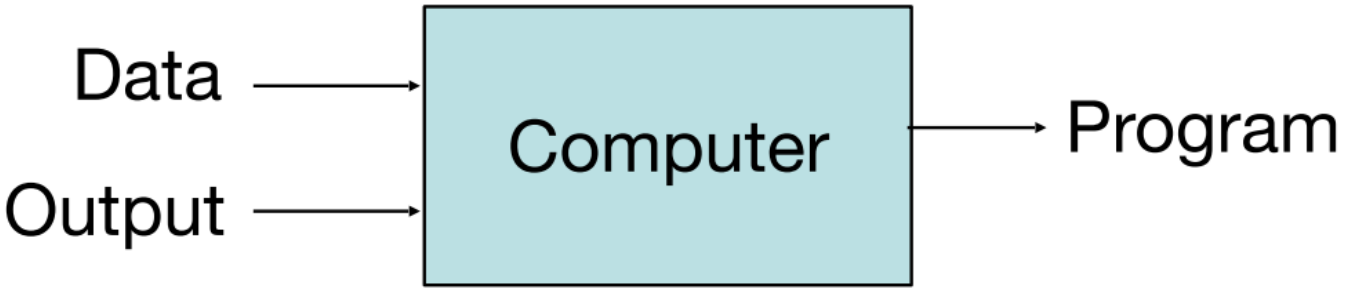
slide by Dhruv Batra

Comparison

- **Traditional Programming**



- **Machine Learning**



slide by Pedro Domingos, Tom Mitchel, Tom Dietterich

About 2,670,000,000 results (0.65 seconds)

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can access data and use it learn for themselves. Jun 18, 2020

expertsystem.com › machine-learning-definition ▾

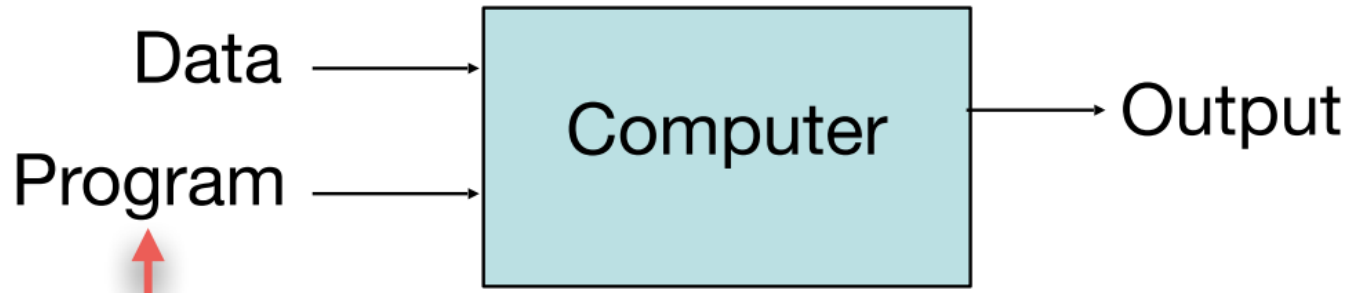
What is Machine Learning? A definition - Expert System

Machine Learning

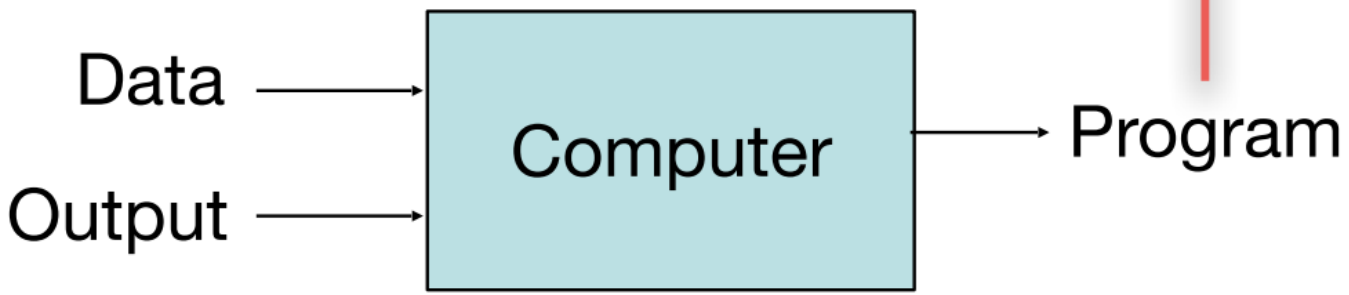
- Machine Learning is a branch of computer science
- It gives computers the ability to learn without being explicitly programmed. (Samuel 1959)
- Evolved from the study of pattern recognition and computer learning theory in artificial intelligence
- *Machine learning explores the study and construction of algorithms that can learn from and make predictions on data*

Comparison

- **Traditional Programming**



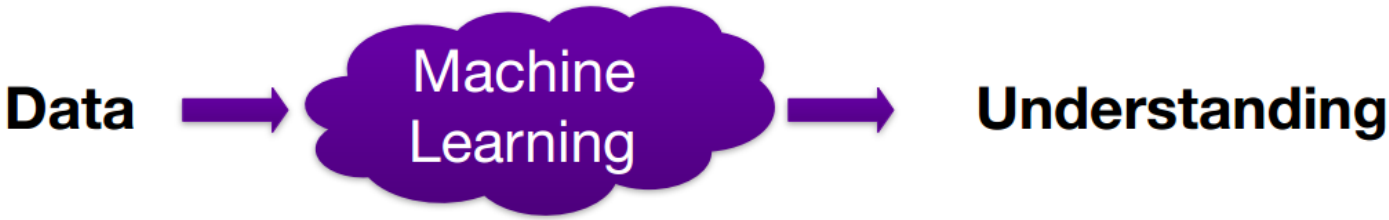
- **Machine Learning**



slide by Pedro Domingos, Tom Mitchel, Tom Dietterich

What is Machine Learning?

- If you are a Scientist

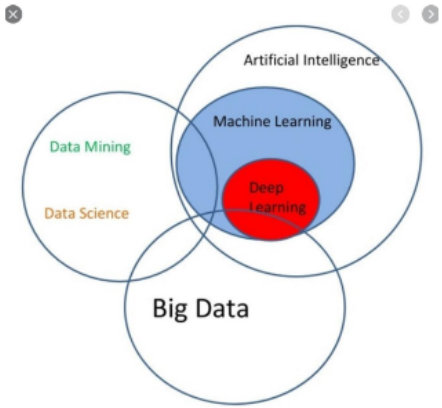


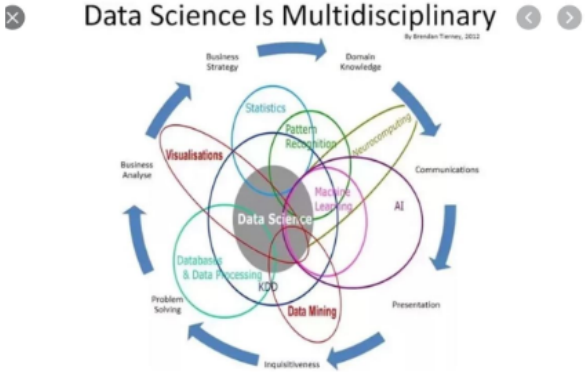
- If you are an Engineer / Entrepreneur
  - Get lots of data
  - Machine Learning
  - ???
  - Profit!

slide by Dhruv Batra

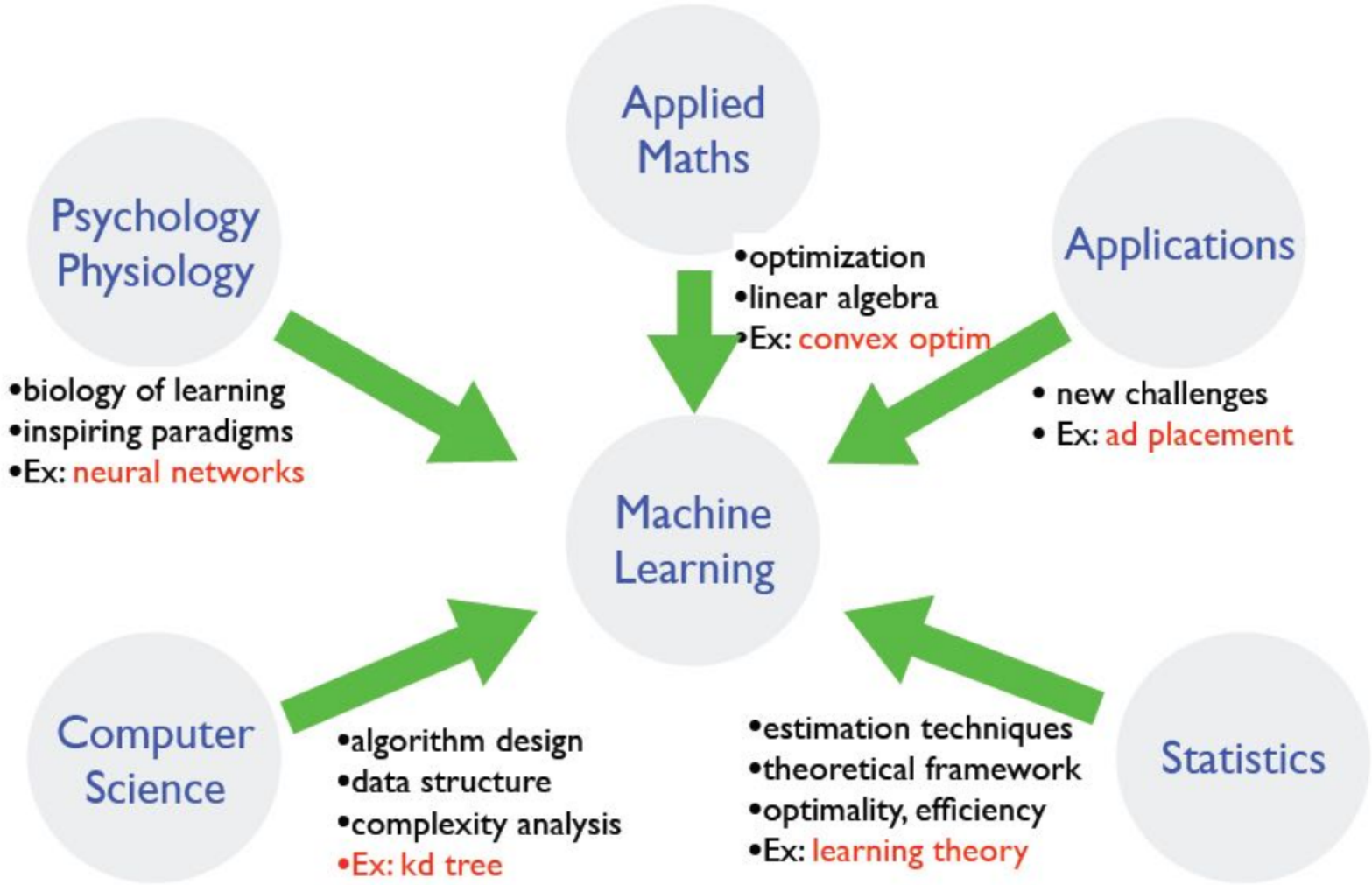
Data Analytics

Data analytics is the science of analyzing raw **data** in order to make conclusions about that information.

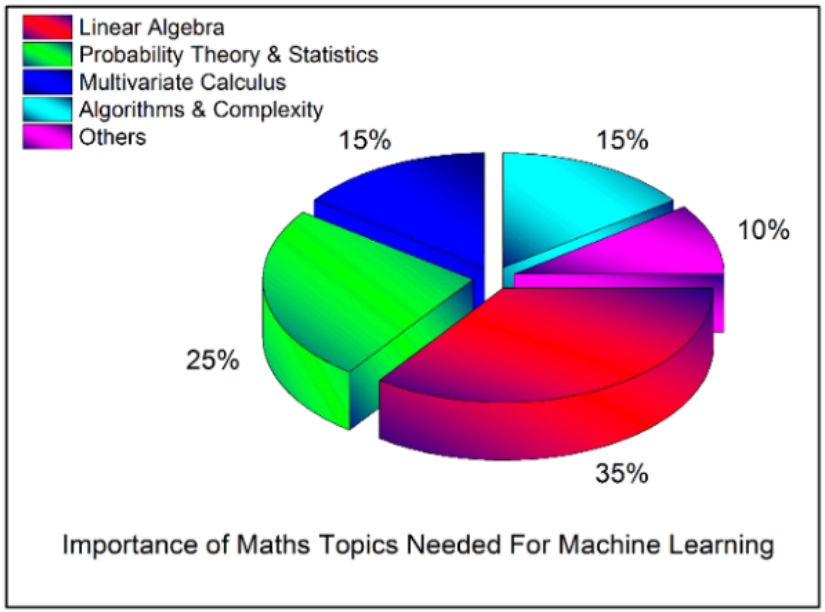




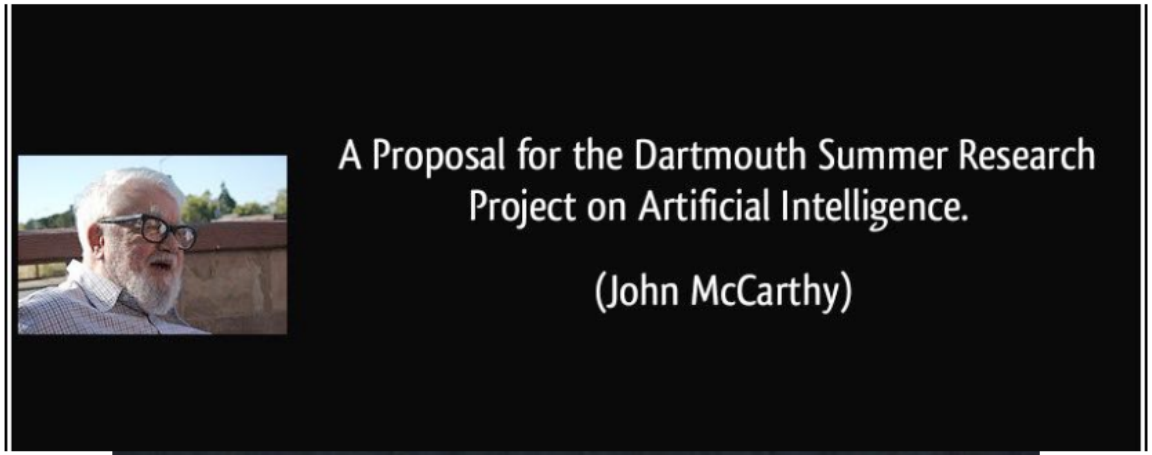
Where does ML fit in?



slide by Fei Sha


Linear Algebra
Probability Theory & Statistics
Multivariate Calculus
Algorithms & Complexity
Others
15%
15%
10%
25%
35%
Importance of Maths Topics Needed For Machine Learning

A Brief History of AI



slide by Dhruv Batra

A Proposal for the

DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

June 17 - Aug. 16

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1) Automatic Computers

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

2) How Can a Computer be Programmed to Use a Language

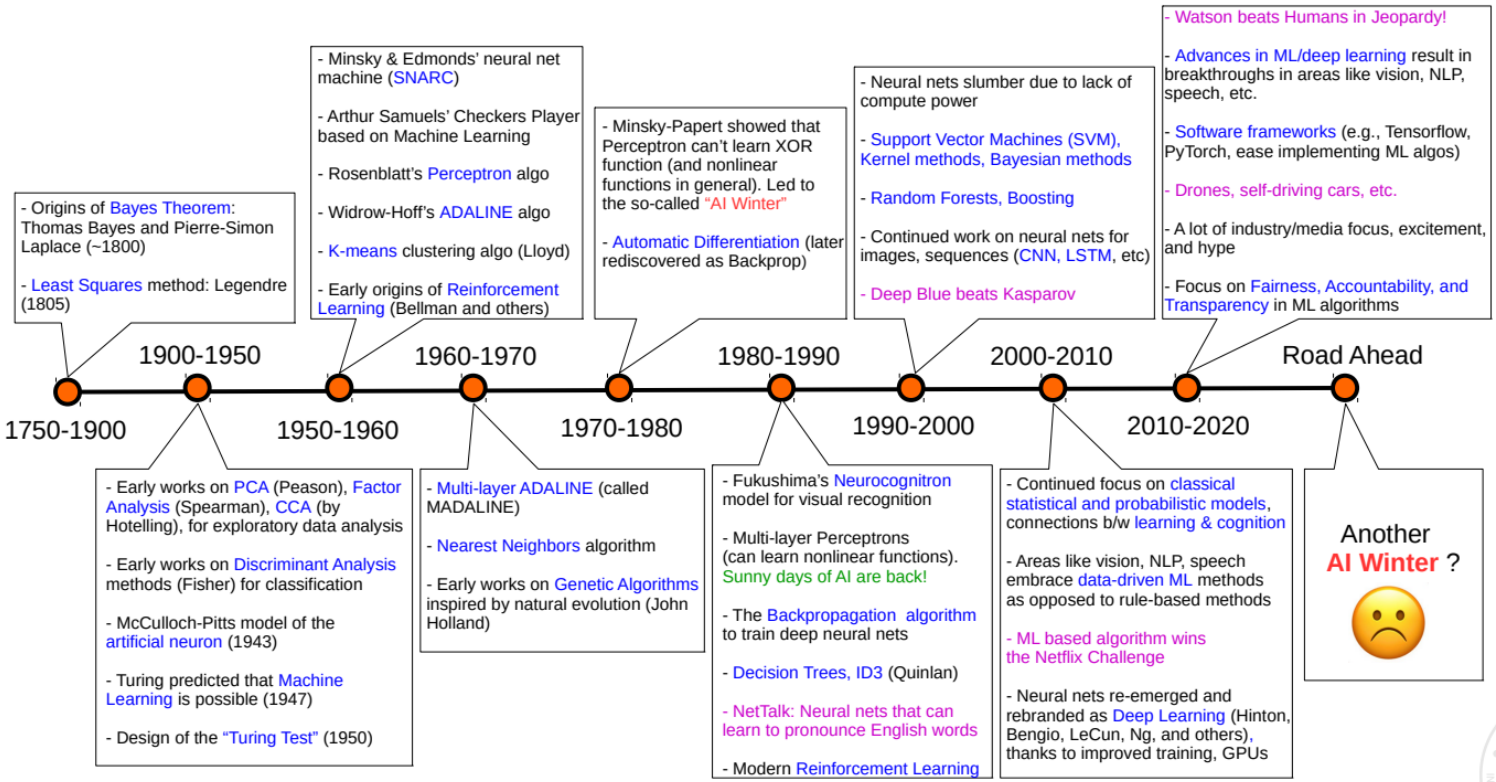
It may be speculated that a large part of human thought con-

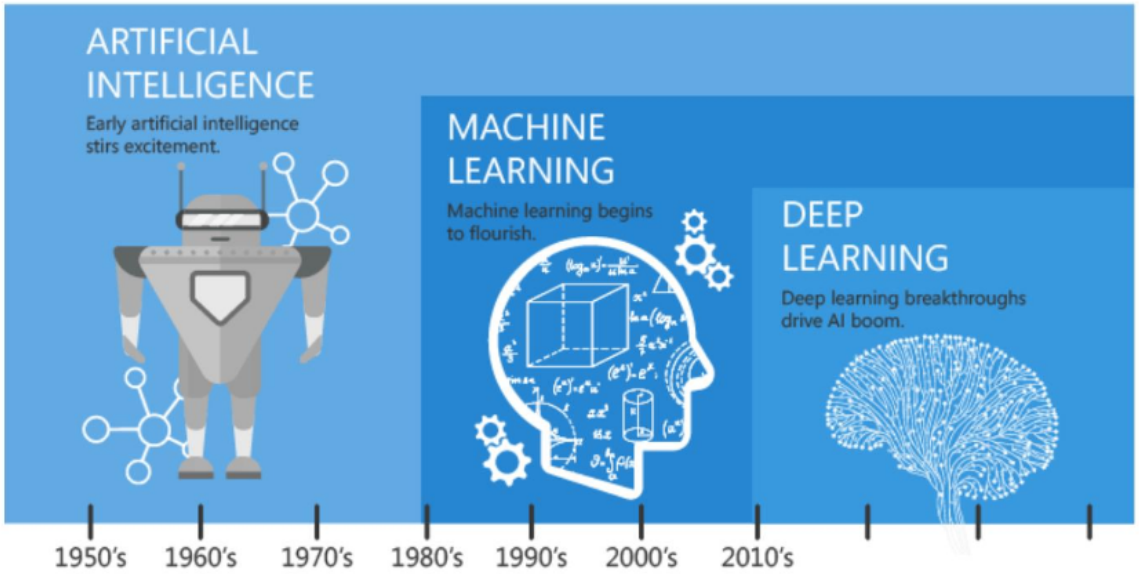
adopted from Dhruv B

History of Machine Learning

- Neural Networks (1960)
- Multi-layer Perceptions (1985)
- Restricted Boltzman Machines (1986)
- Support Vector Machine (1995)
- Deep Belief Networks – New interest in deep learning (2005) CNN
- Deep Recurrent Neural Network (2009)
- Convolutional DBN (2010)
- Max Pooling CDBN (2011)

Machine Learning: A Brief Timeline and Some Milestones





Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

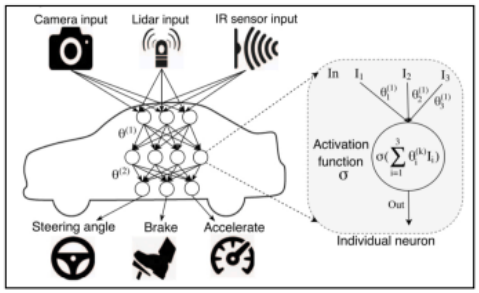
MIT Technology Review, April 23rd, 2013

Today AI is ubiquitous

- Automate routine labor
  - Search

- Understand speech
  - SIRI, Alexa

- Autonomous Vehicles



Why are things working today?

- More compute power
- More data
- Better algorithms/ models

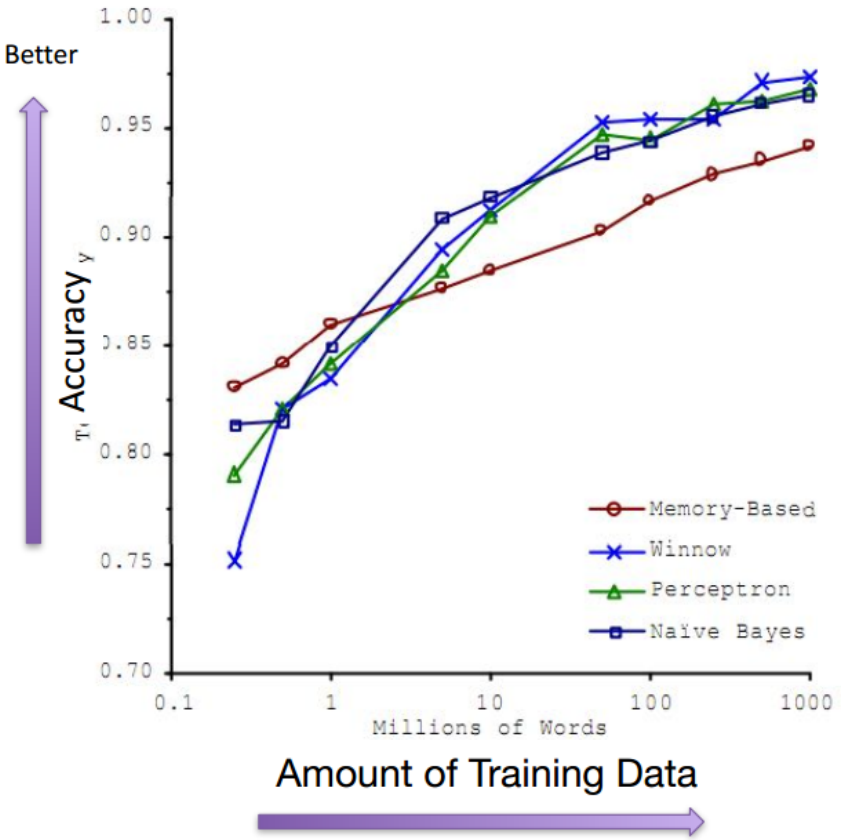


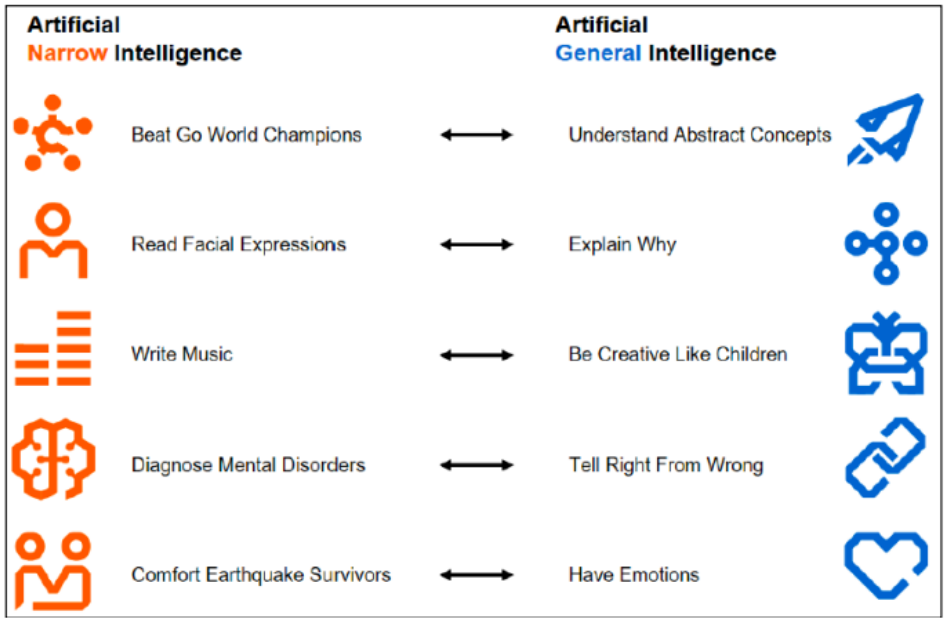
Figure Credit: Banko & Brill, 2011

slide by Dhruv Batra

AI Paradox

- Hard problems for people are easy for AI
- Easy problems are hard for AI
  - Narrow Intelligence General Intelligence

People easy tasks:

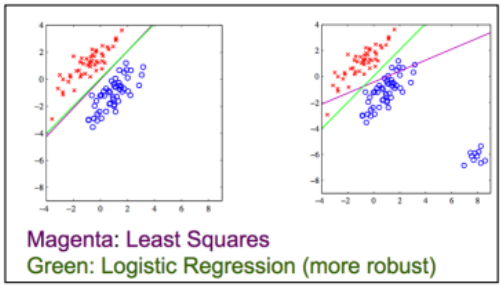


6

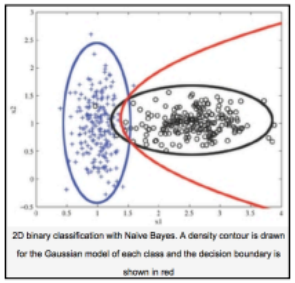
The Machine Learning approach

- Difficulties of hard-coded approach suggests:
  - Allow computers to learn from experience
- First determine what features to use
- Learn to map the features to outputs

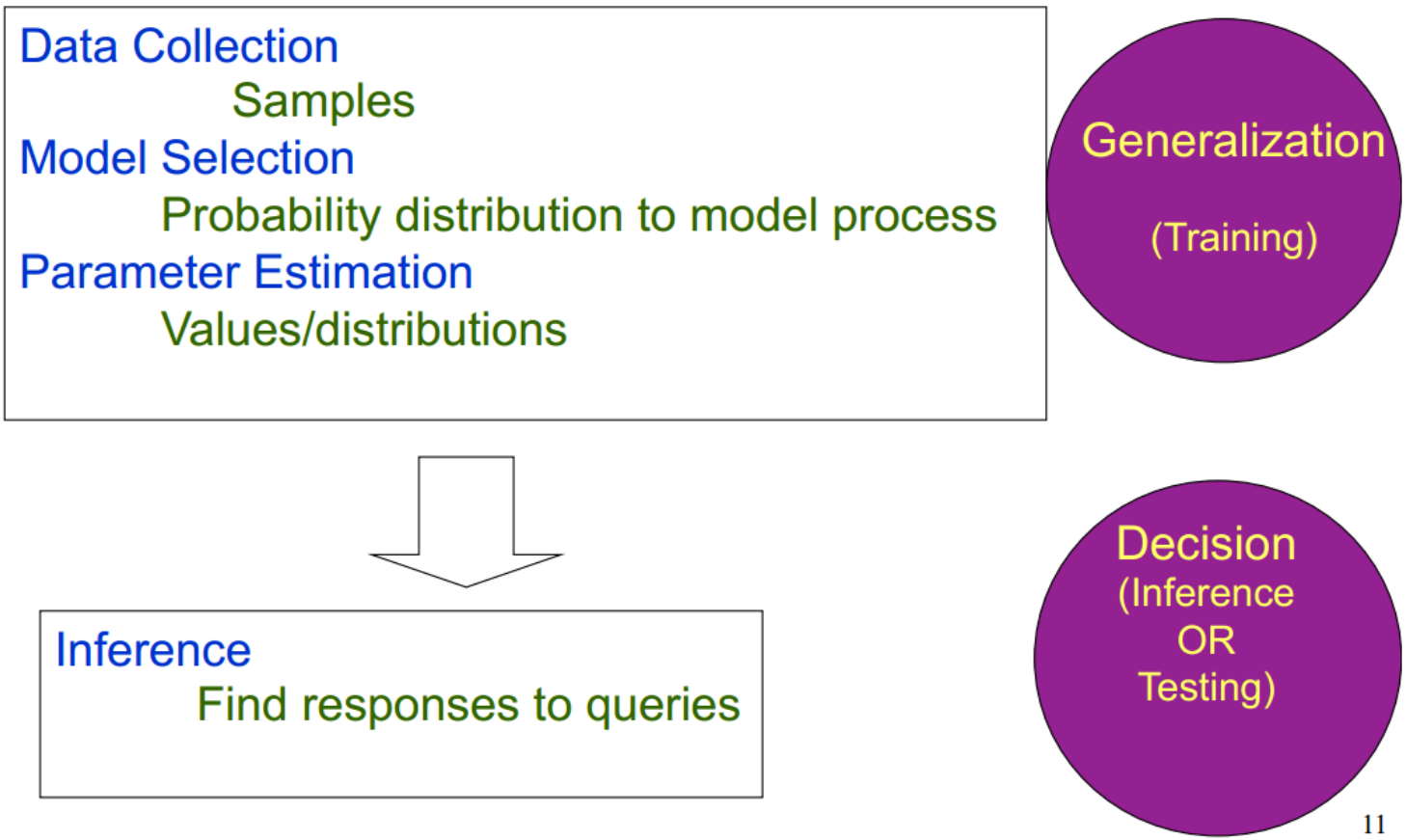
Linear classifier



Quadratic classifier



The ML Approach



11

Learning Problem Definition

- Improving some measure of performance P when executing some task T through some type of training experience E
- Example: Learning to detect credit card fraud

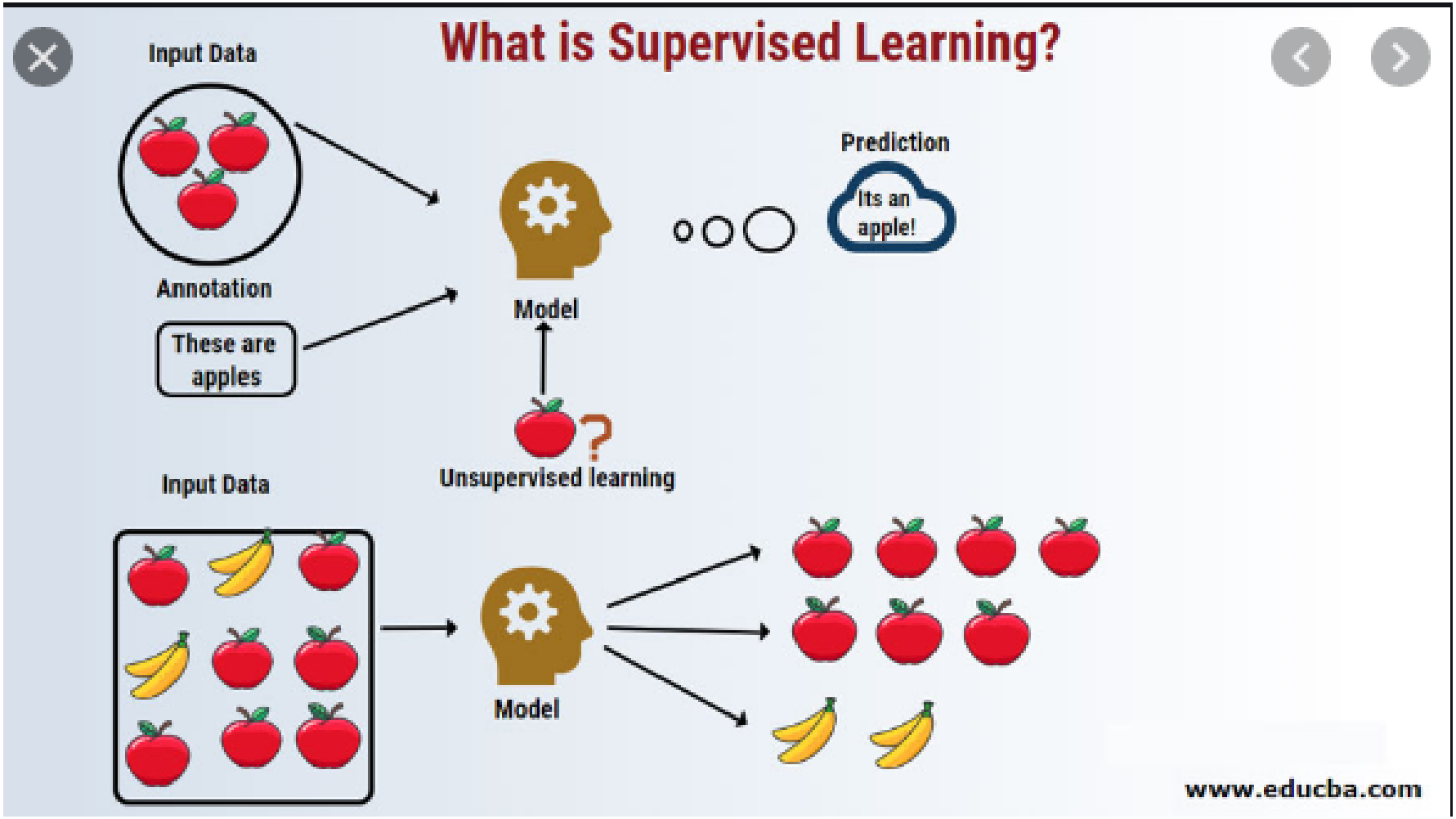
- **Task T**
  - Assign label of fraud or not fraud to credit card transaction
- **Performance measure P**
  - Accuracy of fraud classifier

  With higher penalty when fraud is labeled as not fraud
- **Training experience E**
  - Historical credit card transactions labeled as fraud or not

12

ML Problem Types

1. Based on Type of Data
   1. Supervised, Unsupervised, Semi-supervised
   2. Reinforcement Learning
2. Based on Type of Output
   - Regression, Classification
3. Based on Type of Model
   - Generative, Discriminative

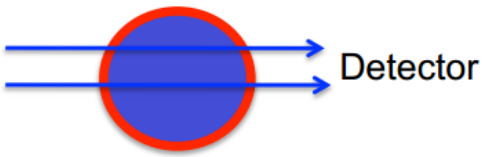

What is Supervised Learning?
Input Data
Annotation
These are apples
Model
Prediction
Its an apple!
Unsupervised learning
Input Data
Model
www.educba.com

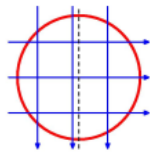
Supervised Learning

- Most widely used methods of ML, e.g.,
  - Spam classification of email
  - Face recognizers over images
  - Medical diagnosis systems
- Inputs $\boldsymbol{x}$ are vectors or more complex objects
  - documents, DNA sequences or graphs
- Outputs are binary, multiclass(K),
  - Multi-label (more than one class), ranking,
  - Structured:
    - $\boldsymbol{y}$ is a graph satisfying constraints, e.g., POS tagging
  - Real-valued or mixture of discrete and real-valued

Supervised Classification Example

- Off-shore oil transfer pipelines
  - Non-invasive measurement of *proportion* of oil,water, gas
    - Called Three-phase Oil/Water/Gas Flow
- Input data: Dual-energy gamma densitometry
  - Beam of gamma rays passed through pipe
  - Attenuation in intensity indicates density of material
  - Single beam insufficient
    - Two degrees of freedom: fraction of oil, fraction of water
    - One beam of Gamma rays of two energies (frequencies)

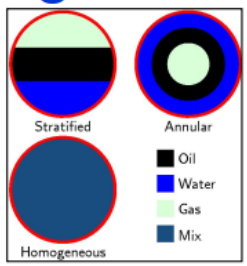




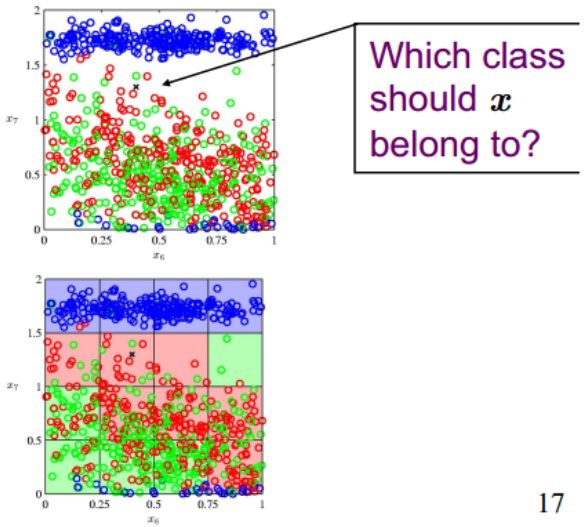
- Six Beams
- 12 measurements
  - attenuation

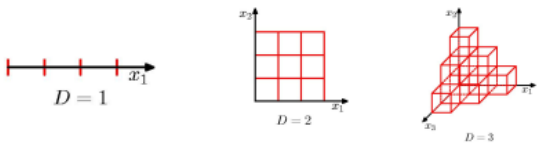
Prediction Problems

1. Predict Volume Fractions of oil/water/gas
2. Predict configuration (one of three)
   - Twelve Features
     - Three classes
     - Two variables, 100 points shown



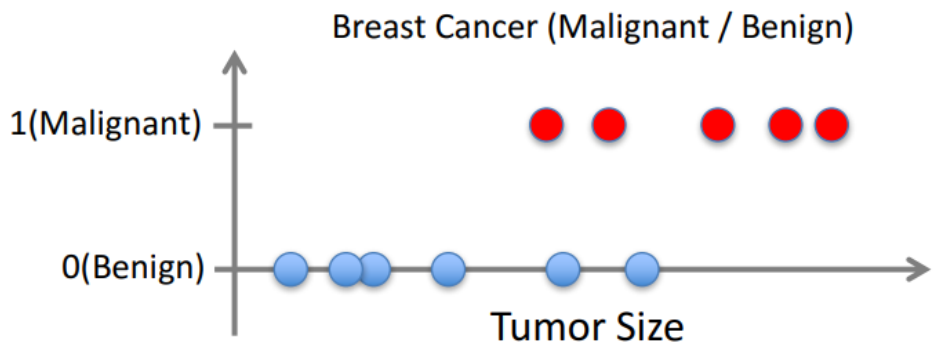
- Naïve cell based voting fails
  - exponential growth of cells with dimensionality
  - 12 dimensions discretized into 6 gives 3 million cells
- Hardly any points in each cell





Supervised Learning: Classification

- Given (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)
- Learn a function $f(x)$ to predict y given x
  - y is categorical == classification

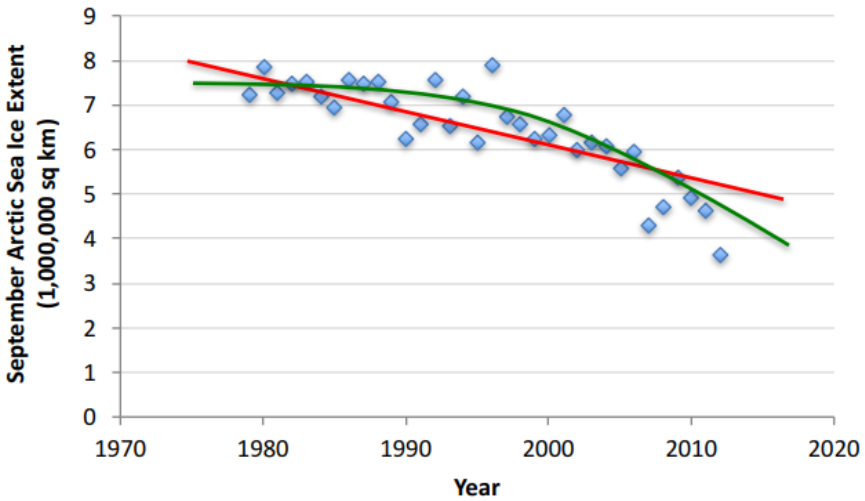


Based on example by Andrew Ng

27

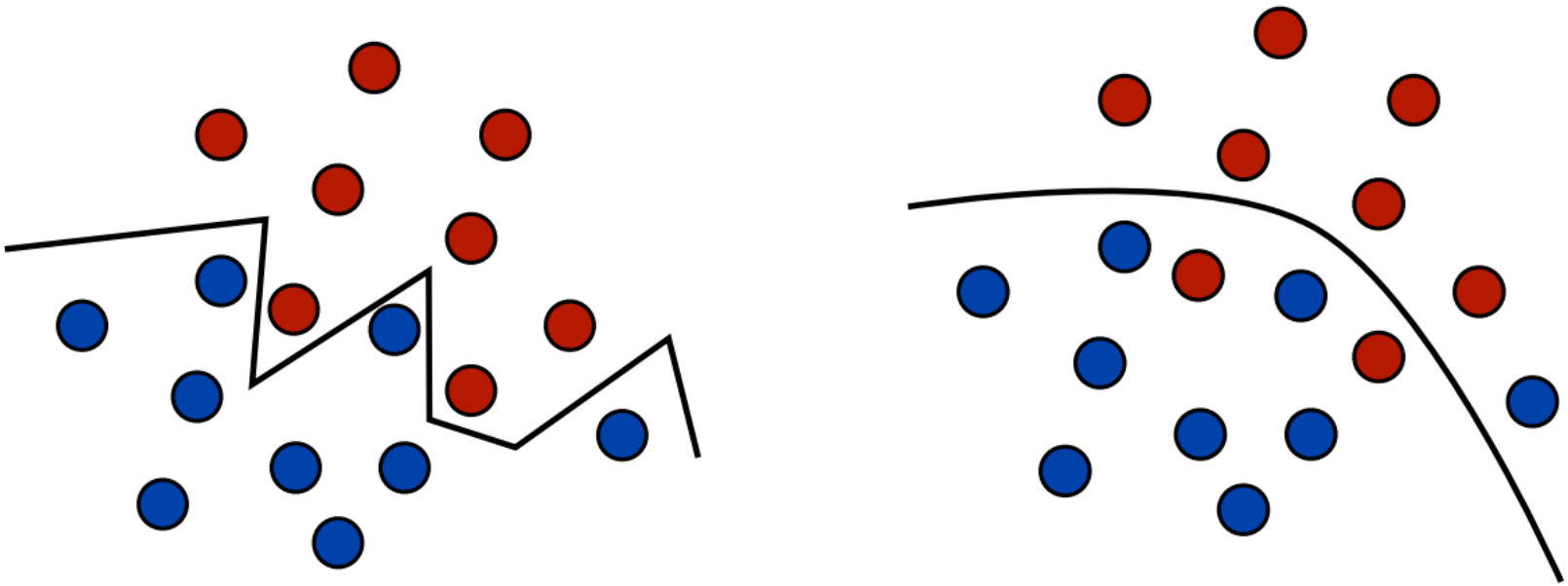
Supervised Learning: Regression

- Given $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
  - y is real-valued == regression



Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)

Learning ≠ Fitting



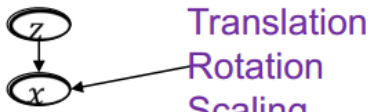
Notion of simplicity/complexity.

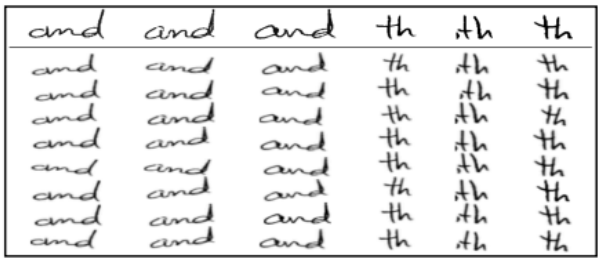
→ How do we define complexity?

Ability to Generalize

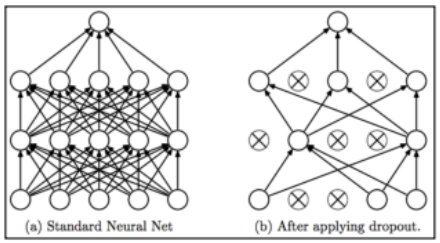
- ML algorithms need to perform well not just on training data but on new inputs as well

1. Parameter Norm Penalties (L^2- and L^1- regularization)
2. Data Set Augmentation



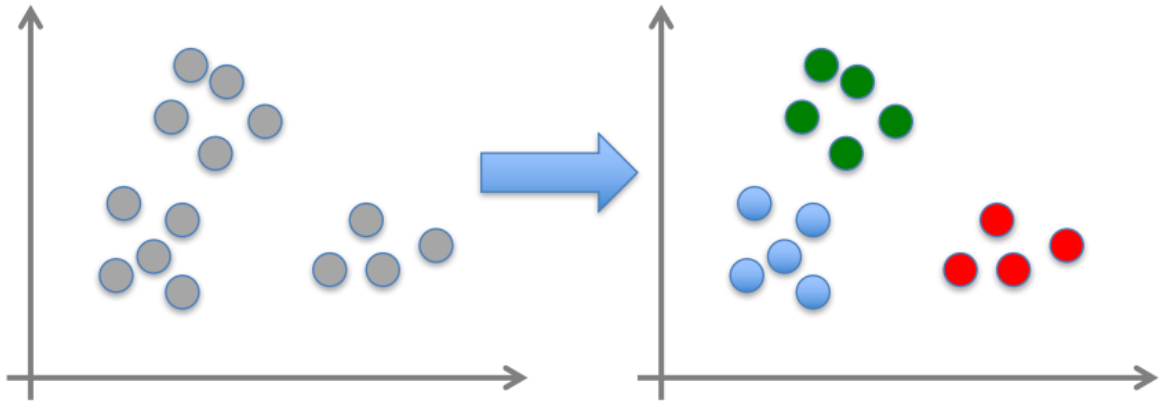


3. Early Stopping
4. Dropout



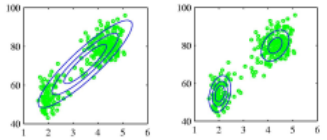
Unsupervised Learning

- Given $x_1, x_2, ..., x_n$ (without labels)
- Output hidden structure behind the x's
  - E.g., clustering

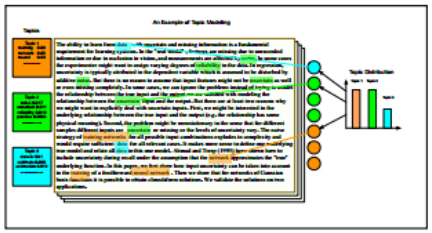


Unsupervised Learning

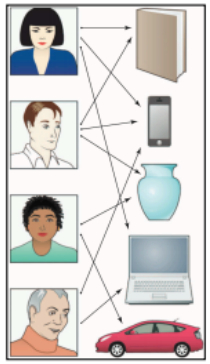
- Unlabeled data assuming underlying structure



1. Clustering to find partition of data
2. Identify a low-dimensional manifold
   - PCA, Autoencoder

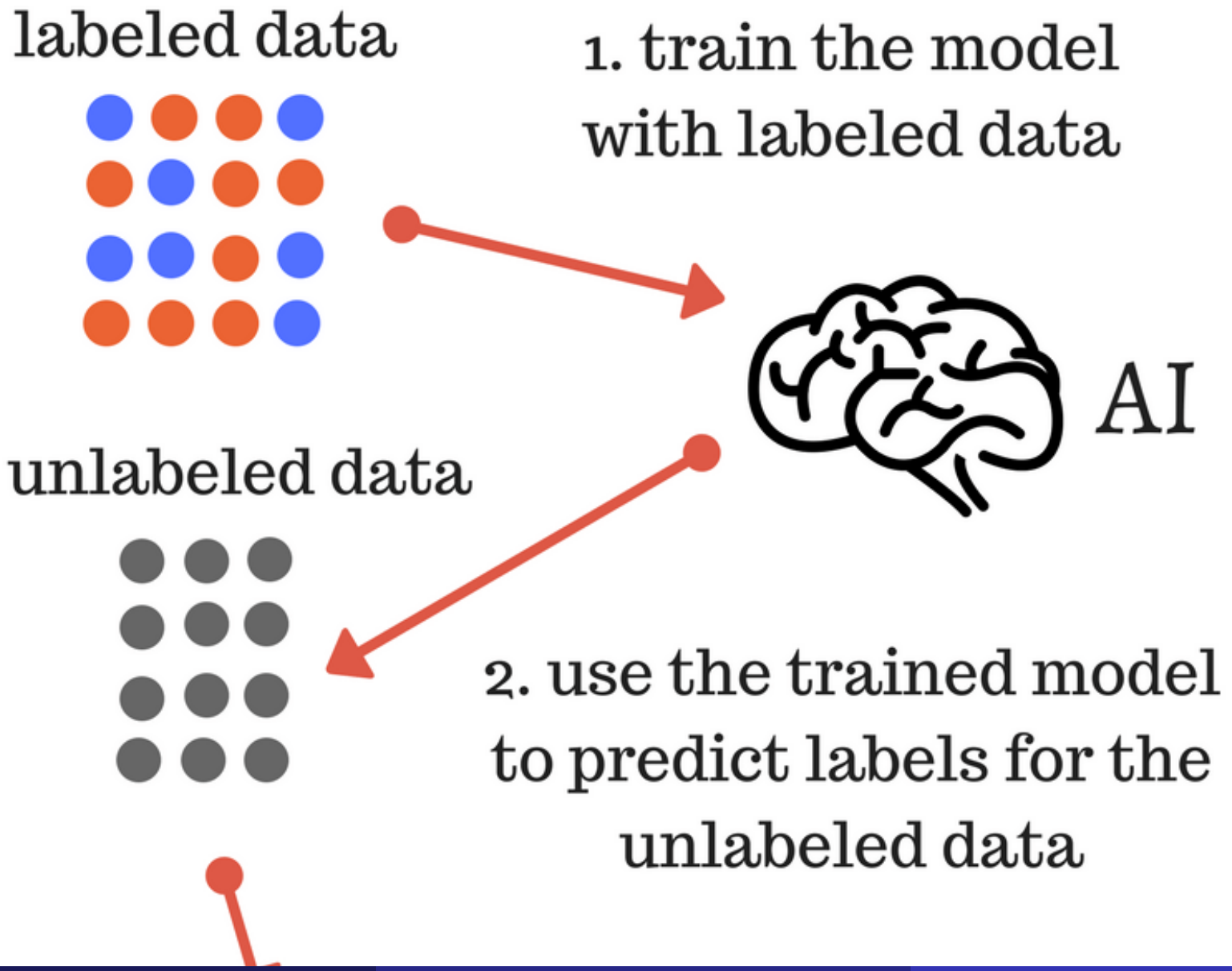


3. Topic modeling
   - Topics are distributions over words
   - Document: a distribution across topics
     - Methods: SVD, Collaborative Filtering
4. Recommendation Systems



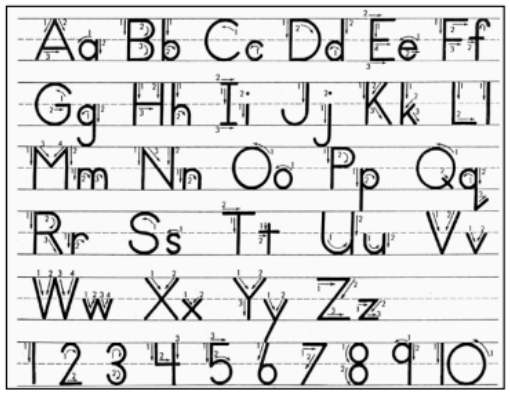
– Data links between users and items
– Suggest other items to user
– Solution: SVD, Collaborative Filtering

semi-supervised

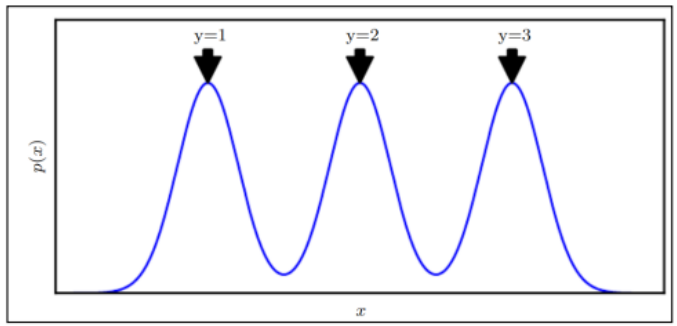


How semi-supervised can succeed

- Ex: density over $\boldsymbol{x}$ is a mixture over three components, one per value of $\mathbf{y}=$ cap/small/digit
- If components well-separated:
  - modeling $p(\mathbf{x})$ reveals where each component is
    - A single labeled example per class enough to learn $p(\mathbf{y}|\mathbf{x})$



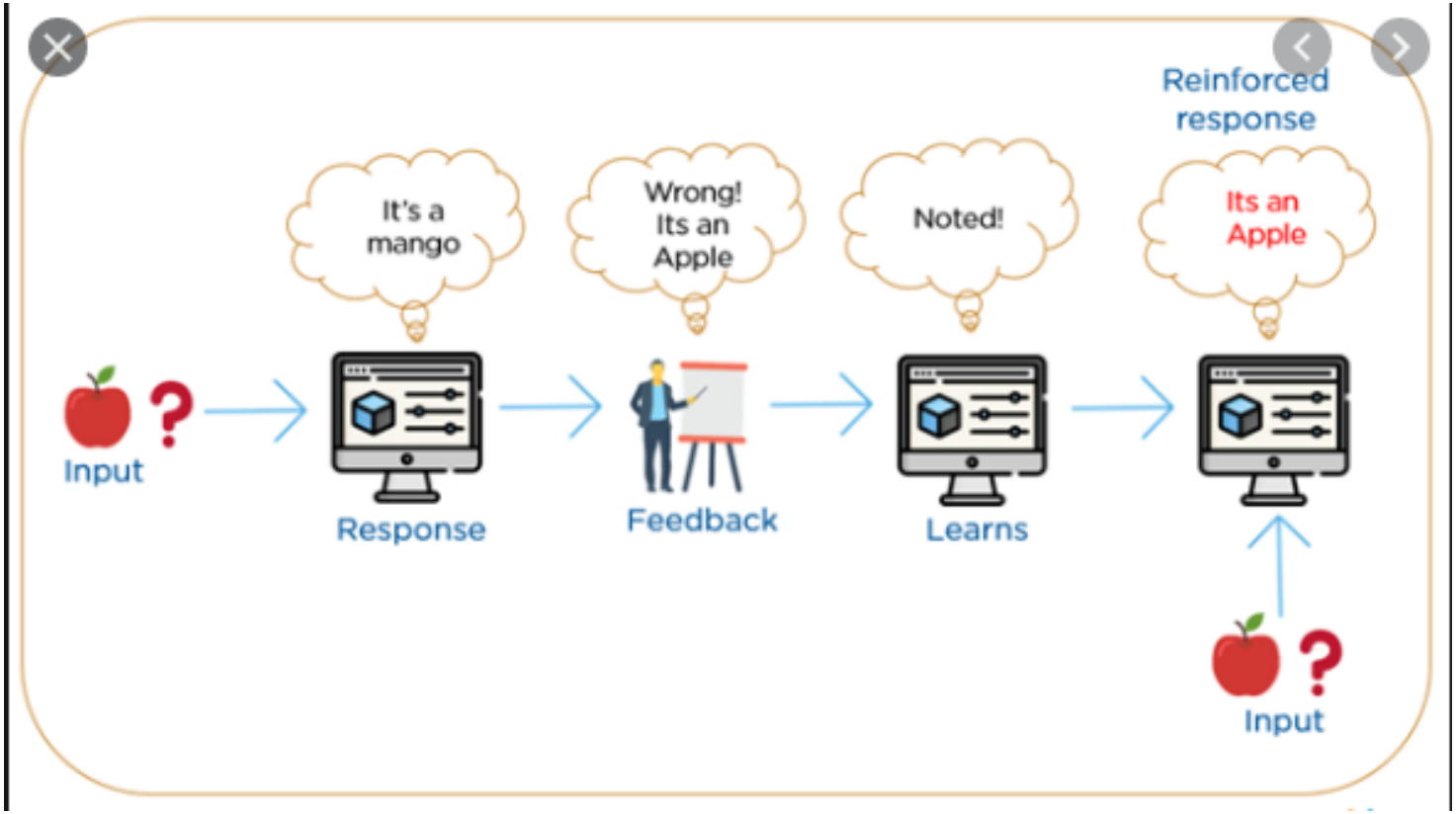
$\boldsymbol{x}$ = no. of black pixels



In this case $p(\mathbf{y}|\mathbf{x})$ is a univariate Gaussian for $\mathbf{y}=1,2,3$

21

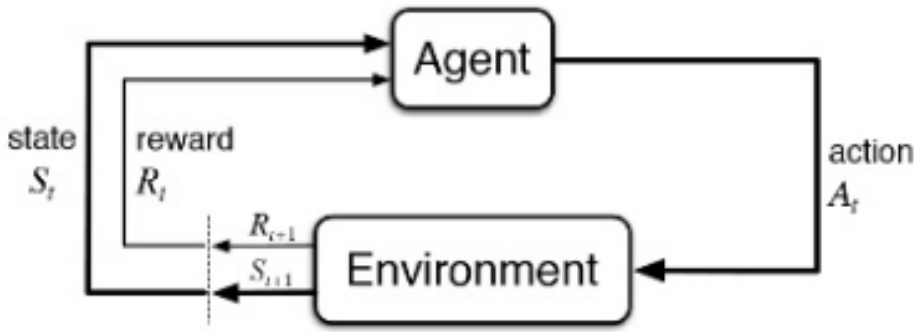
Reinforced Learning



Reinforced Learning

What is reinforcement learning?

- No explicit training data set.
- Nature provides reward for each of the learners actions.
- At each time,
  - Learner has a state and choses an action.
  - Nature responds with new state and a reward.
  - Learner learns from reward and makes better decisions.



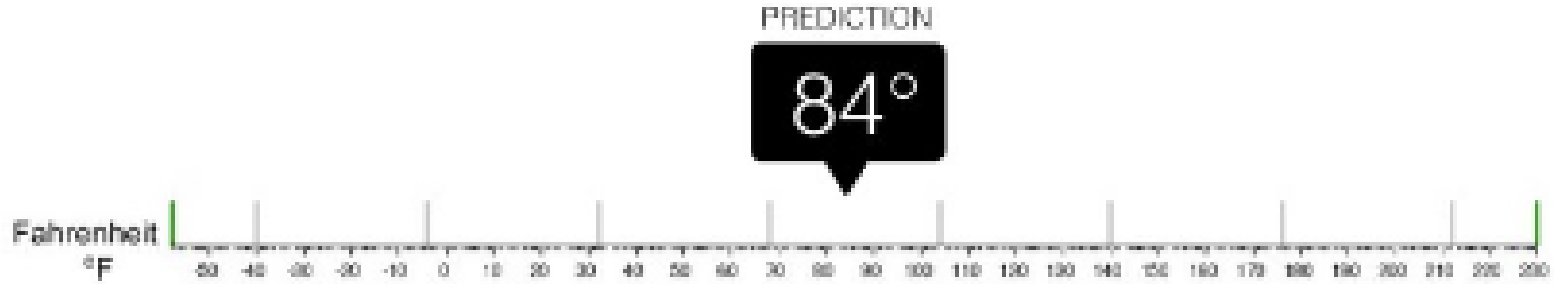
6

Reinforcement Learning

- Given a sequence of states and actions with (delayed) rewards, output a policy
  - Policy is a mapping from states → actions that tells you what to do in a given state
- Examples:
  - Credit assignment problem
  - Game playing
  - Robot in a maze
  - Balance a pole on your hand

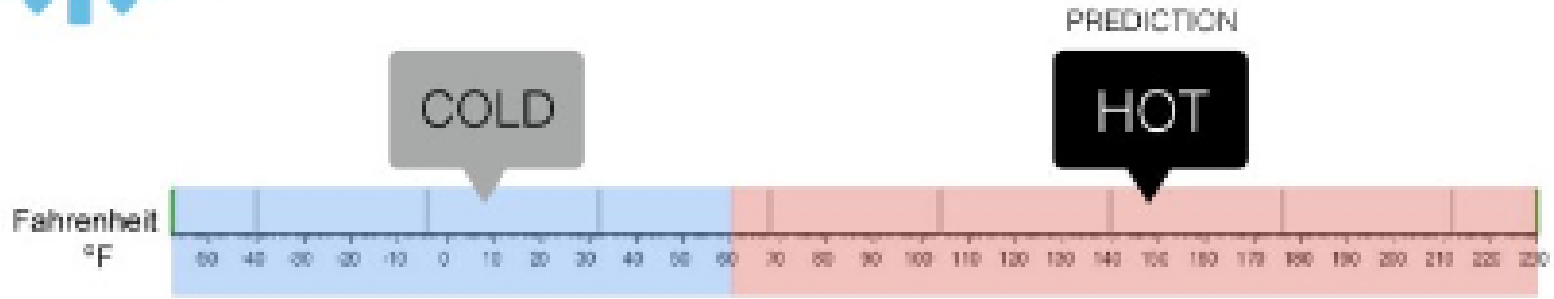
Regression

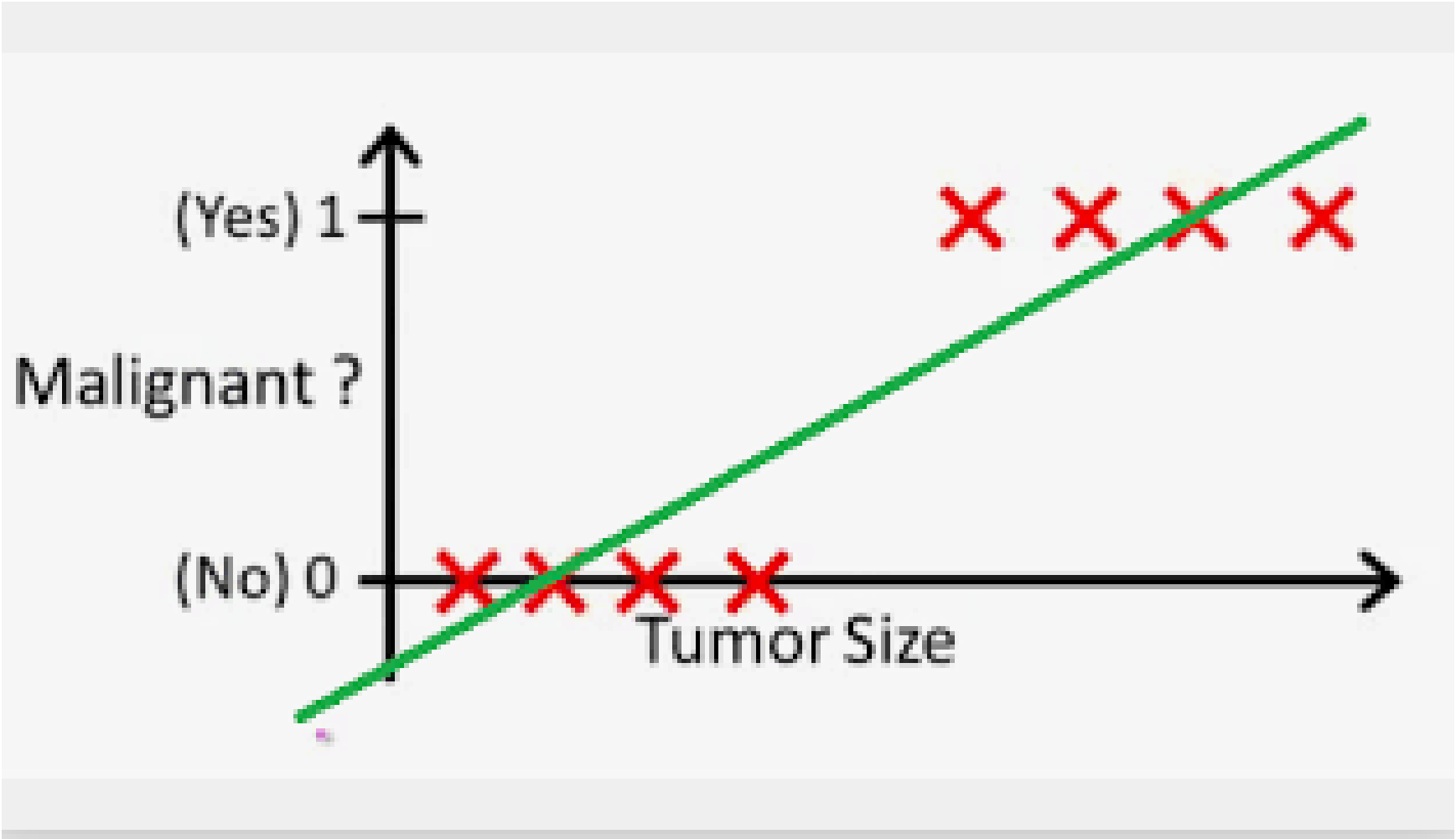
What is the temperature going to be tomorrow?



Classification

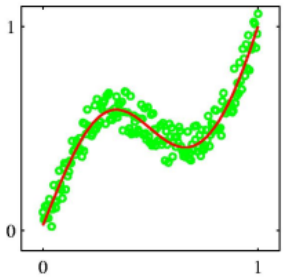
Will it be Cold or Hot tomorrow?




(Yes) 1
Malignant ?
(No) 0
Tumor Size

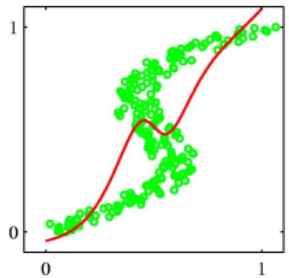
Regression

Problem data set



Red curve is result of fitting a two-layer neural network by minimizing squared error

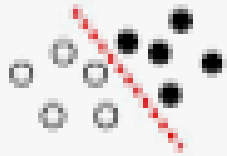
Corresponding inverse problem by reversing x and t



Very poor fit to data: GMMs used here

Generative vs. Discriminative

- Generative:
  - probabilistic "model" of each class
  - decision boundary:
    - where one model becomes more likely
  - natural use of unlabeled data
- Discriminative:
  - focus on the decision boundary
  - more powerful with lots of examples
  - not designed to use unlabeled data
  - only supervised tasks



Generative classifiers

- Assume some functional form for $\mathbf{P(Y)}$, $\mathbf{P(X|Y)}$
- Estimate parameters of $\mathbf{P(X|Y)}$, $\mathbf{P(Y)}$ directly from training data
- Use Bayes rule to calculate $\mathbf{P(Y\,|X)}$

Discriminative Classifiers

- Assume some functional form for $\mathbf{P(Y|X)}$
- Estimate parameters of $\mathbf{P(Y|X)}$ directly from training data

Examples:

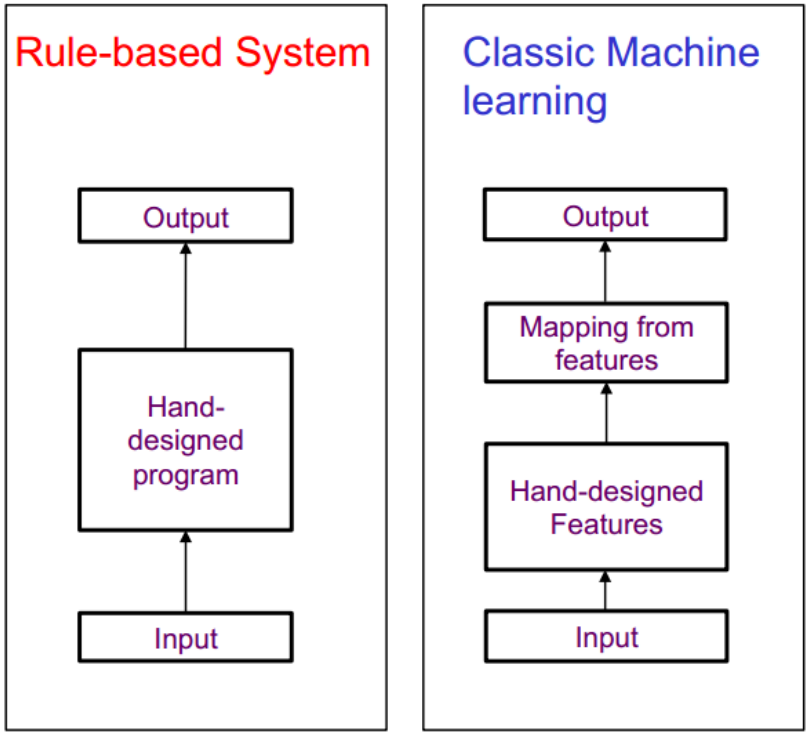
Generative classifiers

- Naïve Bayes
- Bayesian networks
- Markov random fields
- Hidden Markov Models (HMM)

Discriminative Classifiers

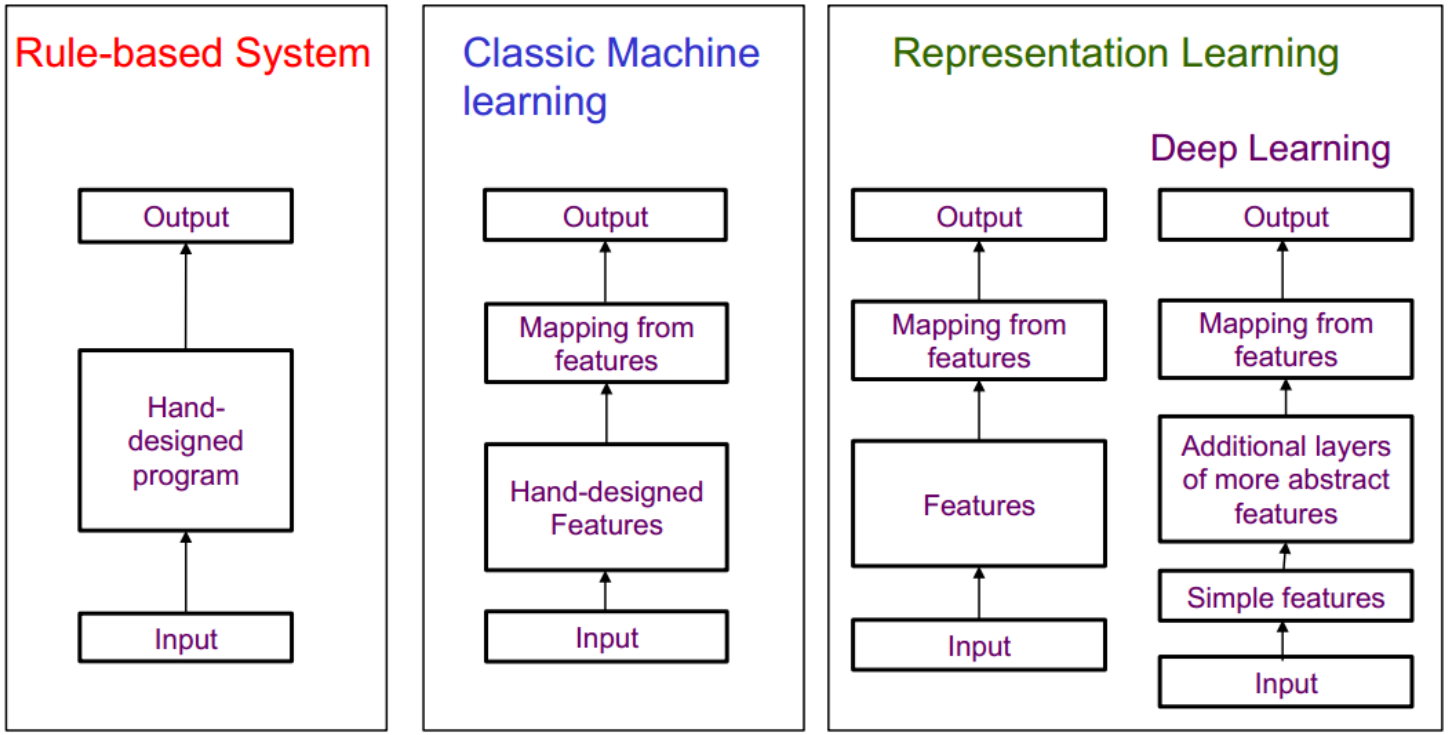
- Logistic regression
- Scalar Vector Machine
- Traditional neural networks
- Nearest neighbour
- Conditional Random Fields (CRF)s

Two paradigms in AI



Shaded boxes indicate components that can learn from data

Summary of AI Models



Shaded boxes indicate components that can learn from data

ML Steps

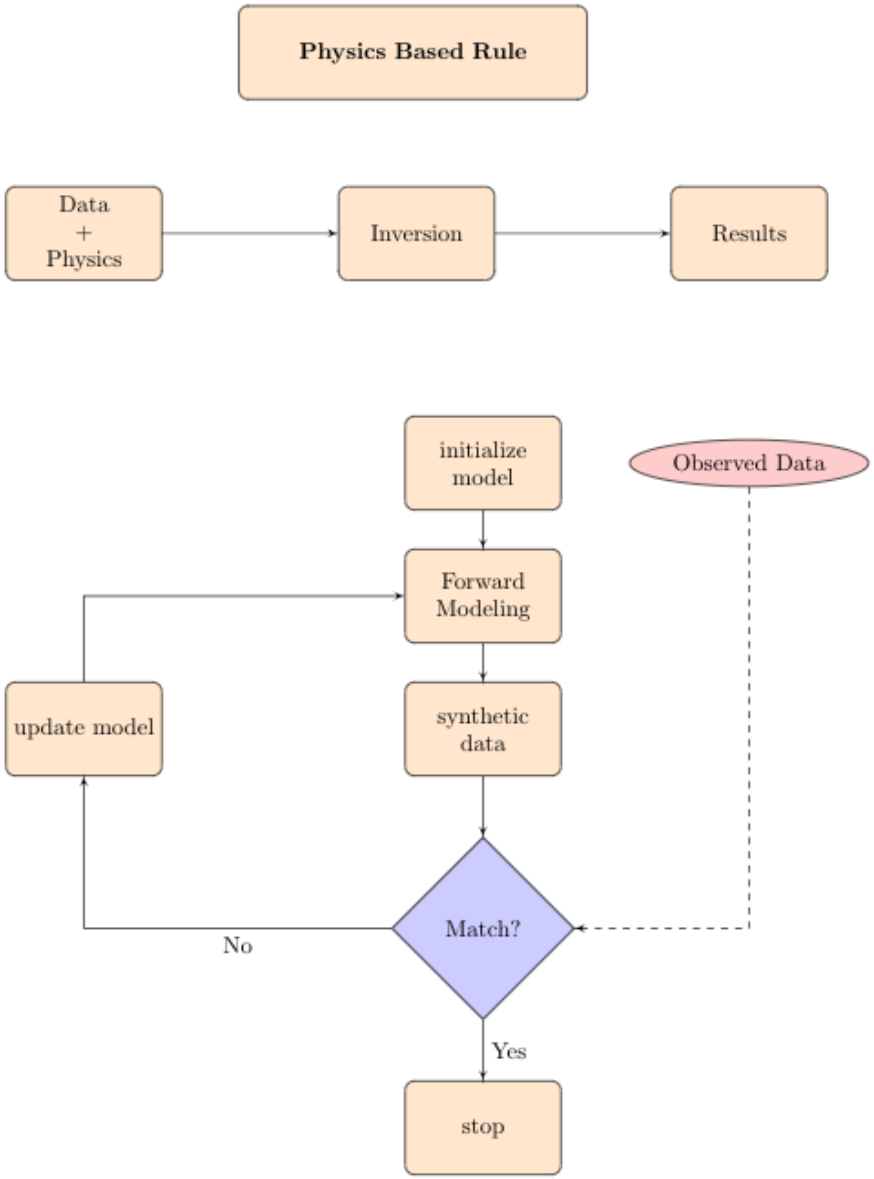
- Spaces: input space X, output space Y,
- Loss Function: $L : YxY \rightarrow \mathbb{R}$.
  - $L(y, \hat{y})$:
  - binary classification: 0-1 loss, $L(y, \hat{y}) = 1_{y \neq \hat{y}}$.
  - regression: $l(y, \hat{y}) = (y - \hat{y})^2$.
- Hypothesis Set: subset of functions out of which the learner selects his hypothesis.
  - depends on features.
  - represents prior knowledge about task.

Supervised Learning Set-Up

- Training Data: $S = ((x_1, y_1), (x_2, y_2), ..., (x_m, y_m)$; x_i: data; y_i: Label data
- Problem: Find hypothesis h with small generalization error (OPTIMIZATION),
  - Deterministic case: output label deterministic function of input, $y = f(x)$.
  - Stochastic case: output probabilistic function of input.

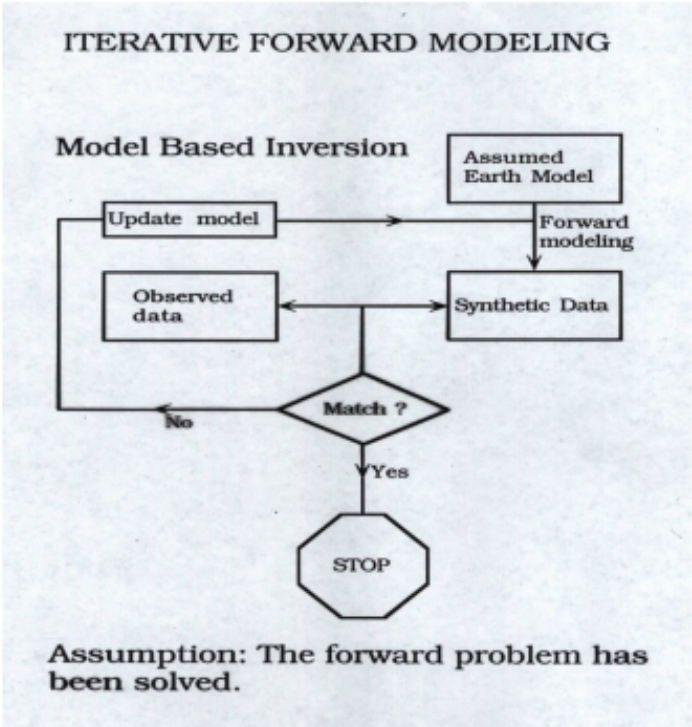
Notice any similarities with the way we solve our inverse problems?

Physics-Based Approach



Inversion

- Direct Inversion: Reverse physics – VERY UNSTABLE for most applications
- Model Based Inversion: Iterative Model fitting



INVERSION = OPTIMIZATION

Inversion Formulation

- **Data Vector**

$$\mathbf{d} = [d_1, d_2, .., d_N]^T$$

- **Model Vector**

$$\mathbf{m} = [m_1, m_2, .., m_M]^T$$

- **In general**

$$N \neq M.$$

- **Forward Modeling**

$$\mathbf{d} = g(\mathbf{m})$$

- **Linear Problem**

$$\mathbf{d} = \mathbf{Gm}$$

Inversion Formulation

Error/Cost/Objective/Misfit Function

measures differences or similarities between observed and synthetic (numerically computed) data

$$E(\mathbf{m}) = \| \mathbf{d} - g(\mathbf{m}) \|$$

$\| . \|$ is called a 'norm' – most commonly used norm is an L_2 norm or a least squares error measure

$$E(\mathbf{m}) = (\mathbf{d} - g(\mathbf{m}))^T (\mathbf{d} - g(\mathbf{m}))$$

$$E(\mathbf{m}) = \sum_{i=1}^{ND} \left(d^i_{obs} - d^i_{syn} \right)^2$$

Machine Learning

- Forward Modeling operator is unknown!
- Goal: Find an operator f that can be applied to the data to estimate models.
- Find the operator by systematic examination of a series of observed data and their known answers. Learn from experience! TRAINING

$$\mathbf{m}_{est} = f\,\mathbf{d}$$

.

When to Use Machine Learning

It is important to remember that ML is not a solution for every type of problem. There are certain cases where robust solutions can be developed without using ML techniques. For example, you don't need ML if you can determine a target value by using simple rules, computations, or predetermined steps that can be programmed without needing any data-driven learning.

Use machine learning for the following situations:

- *You cannot code the rules:* Many human tasks (such as recognizing whether an email is spam or not spam) cannot be adequately solved using a simple (deterministic), rule-based solution. A large number of factors could influence the answer. When rules depend on too many factors and many of these rules overlap or need to be tuned very finely, it soon becomes difficult for a human to accurately code the rules. You can use ML to effectively solve this problem.
- *You cannot scale:* You might be able to manually recognize a few hundred emails and decide whether they are spam or not. However, this task becomes tedious for millions of emails. ML solutions are effective at handling large-scale problems.

ANAZON-machinelearning-dg.pdf

Building a Machine Learning Application

Building ML applications is an iterative process that involves a sequence of steps. To build an ML application, follow these general steps:

1. Frame the core ML problem(s) in terms of what is observed and what answer you want the model to predict.
2. Collect, clean, and prepare data to make it suitable for consumption by ML model training algorithms. Visualize and analyze the data to run sanity checks to validate the quality of the data and to understand the data.
3. Often, the raw data (input variables) and answer (target) are not represented in a way that can be used to train a highly predictive model. Therefore, you typically should attempt to construct more predictive input representations or features from the raw variables.
4. Feed the resulting features to the learning algorithm to build models and evaluate the quality of the models on data that was held out from model building.
5. Use the model to generate predictions of the target answer for new data instances.

ANAZON-machinelearning-dg.pdf

ML Steps

Formulating the Problem

The first step in machine learning is to decide what you want to predict, which is known as the label or target answer. Imagine a scenario in which you want to manufacture products, but your decision to manufacture each product depends on its number of potential sales. In this scenario, you want to predict how many times each product will be purchased (predict number of sales). There are multiple ways to define this problem by using machine learning. Choosing how to define the problem depends on your use case or business need.

ANAZON-machinelearning-dg.pdf

ML Steps

It is important to avoid over-complicating the problem and to frame the simplest solution that meets your needs. However, it is also important to avoid losing information, especially information in the historical answers. Here, converting an actual past sales number into a binary variable “over 10” versus “fewer” would lose valuable information. Investing time in deciding which target makes most sense for you to predict will save you from building models that don’t answer your question.

ANAZON-machinelearning-dg.pdf

Collecting Labeled Data

ML problems start with data—preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called *labeled data*. In supervised ML, the algorithm teaches itself to learn from the labeled examples that we provide.

Each example/observation in your data must contain two elements:

- The target – The answer that you want to predict. You provide data that is labeled with the target (correct answer) to the ML algorithm to learn from. Then, you will use the trained ML model to predict this answer on data for which you do not know the target answer.
- Variables/features – These are attributes of the example that can be used to identify patterns to predict the target answer.

ANAZON-machinelearning-dg.pdf

Analyzing Your Data

Before feeding your labeled data to an ML algorithm, it is a good practice to inspect your data to identify issues and gain insights about the data you are using. The predictive power of your model will only be as good as the data you feed it.

When analyzing your data, you should keep the following considerations in mind:

- Variable and target data summaries – It is useful to understand the values that your variables take and which values are dominant in your data. You could run these summaries by a subject matter expert for the problem that you want to solve. Ask yourself or the subject matter expert: Does the data match your expectations? Does it look like you have a data collection problem? Is one class in your target more frequent than the other classes? Are there more missing values or invalid data than you expect?
- Variable-target correlations – Knowing the correlation between each variable and the target class is helpful because a high correlation implies that there is a relationship between the variable and the

ANAZON-machinelearning-dg.pdf

ML Steps

Feature Processing

After getting to know your data through data summaries and visualizations, you might want to transform your variables further to make them more meaningful. This is known as *feature processing*. For example, say you have a variable that captures the date and time at which an event occurred. This date and time will never occur again and hence won't be useful to predict your target. However, if this variable is transformed into features that represent the hour of the day, the day of the week, and the month, these variables could be useful to learn if the event tends to happen at a particular hour, weekday, or month. Such feature processing to form more generalizable data points to learn from can provide significant improvements to the predictive models.

ANAZON-machinelearning-dg.pdf

Splitting the Data into Training and Evaluation Data

The fundamental goal of ML is to *generalize* beyond the data instances used to train models. We want to evaluate the model to estimate the quality of its pattern generalization for data the model has not been trained on. However, because future instances have unknown target values and we cannot check the accuracy of our predictions for future instances now, we need to use some of the data that we already know the answer for as a proxy for future data. Evaluating the model with the same data that was used for training is not useful, because it rewards models that can "remember" the training data, as opposed to generalizing from it.

A common strategy is to take all available labeled data, and split it into training and evaluation subsets, usually with a ratio of 70-80 percent for training and 20-30 percent for evaluation. The ML system uses the training data to train models to see patterns, and uses the evaluation data to evaluate the predictive quality of the trained model. The ML system evaluates predictive performance by comparing predictions on the evaluation data set with true values (known as ground truth) using a variety of metrics. Usually, you use the "best" model on the evaluation subset to make predictions on future instances for which you do not know the target answer.

ANAZON-machinelearning-dg.pdf

ML Steps

Training the Model

You are now ready to provide the ML algorithm (that is, the *learning algorithm*) with the training data. The algorithm will learn from the training data patterns that map the variables to the target, and it will output a model that captures these relationships. The ML model can then be used to get predictions on new data for which you do not know the target answer.

ANAZON-machinelearning-dg.pdf

ML Steps

Learning Algorithm

The learning algorithm's task is to learn the weights for the model. The weights describe the likelihood that the patterns that the model is learning reflect actual relationships in the data. A learning algorithm consists of a loss function and an optimization technique. The loss is the penalty that is incurred when the estimate of the target provided by the ML model does not equal the target exactly. A loss function quantifies this penalty as a single value. An optimization technique seeks to minimize the loss. In Amazon Machine Learning, we use three loss functions, one for each of the three types of prediction problems. The optimization technique used in Amazon ML is online Stochastic Gradient Descent (SGD). SGD makes sequential passes over the training data, and during each pass, updates feature weights one example at a time with the aim of approaching the optimal weights that minimize the loss.

ANAZON-machinelearning-dg.pdf

ML Steps

Training Parameters

The Amazon ML learning algorithm accepts parameters, called hyperparameters or training parameters, that allow you to control the quality of the resulting model. Depending on the hyperparameter, Amazon ML auto-selects settings or provides static defaults for the hyperparameters. Although default hyperparameter settings generally produce useful models, you might be able to improve the predictive performance of your models by changing hyperparameter values. The following sections describe common hyperparameters associated with learning algorithms for linear models, such as those created by Amazon ML.

Learning Rate

The learning rate is a constant value used in the Stochastic Gradient Descent (SGD) algorithm. Learning rate affects the speed at which the algorithm reaches (converges to) the optimal weights. The SGD algorithm makes updates to the weights of the linear model for every data example it sees. The size of these updates is controlled by the learning rate. Too large a learning rate might prevent the weights from approaching the optimal solution. Too small a value results in the algorithm requiring many passes to approach the optimal weights.

In Amazon ML, the learning rate is auto-selected based on your data.

ANAZON-machinelearning-dg.pdf

ML Steps

Model Size

If you have many input features, the number of possible patterns in the data can result in a large model. Large models have practical implications, such as requiring more RAM to hold the model while training and when generating predictions. In Amazon ML, you can reduce the model size by using L1 regularization or by specifically restricting the model size by specifying the maximum size. Note that if you reduce the model size too much, you could reduce your model's predictive power.

For information about the default model size, see Training Parameters: Types and Default Values (p. 75). For more information about regularization, see Regularization (p. 14).

Number of Passes

The SGD algorithm makes sequential passes over the training data. The `Number of passes` parameter controls the number of passes that the algorithm makes over the training data. More passes result in a model that fits the data better (if the learning rate is not too large), but the benefit diminishes with an increasing number of passes. For smaller data sets, you can significantly increase the number of passes, which allows the learning algorithm to effectively fit the data more closely. For extremely large datasets, a single pass might suffice.

For information about the default number of passes, see Training Parameters: Types and Default Values (p. 75).

Data Shuffling

In Amazon ML, you must shuffle your data because the SGD algorithm is influenced by the order of the rows in the training data. Shuffling your training data results in better ML models because it helps

ANAZON-machinelearning-dg.pdf

ML Steps

Regularization

Regularization helps prevent linear models from overfitting training data examples (that is, memorizing patterns instead of generalizing them) by penalizing extreme weight values. L1 regularization has the effect of reducing the number of features used in the model by pushing to zero the weights of features that would otherwise have small weights. As a result, L1 regularization results in sparse models and reduces the amount of noise in the model. L2 regularization results in smaller overall weight values, and stabilizes the weights when there is high correlation between the input features. You control the amount of L1 or L2 regularization applied by using the `Regularization type` and `Regularization amount` parameters. An extremely large regularization value could result in all features having zero weights, preventing a model from learning patterns.

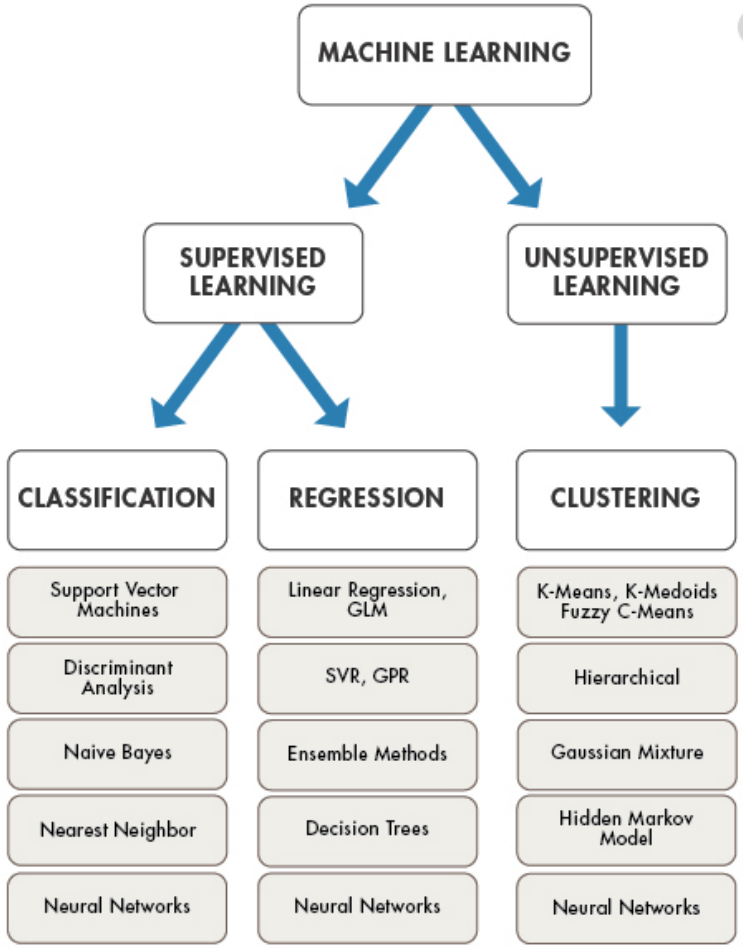
For information about the default regularization values, see Training Parameters: Types and Default Values (p. 75).

Evaluating Model Accuracy

The goal of the ML model is to learn patterns that generalize well for unseen data instead of just memorizing the data that it was shown during training. Once you have a model, it is important to check if your model is performing well on unseen examples that you have not used for training the model.

ANAZON-machinelearning-dg.pdf

Summary


MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISED LEARNING
CLASSIFICATION
REGRESSION
CLUSTERING
Support Vector Machines
Discriminant Analysis
Naive Bayes
Nearest Neighbor
Neural Networks
Linear Regression, GLM
SVR, GPR
Ensemble Methods
Decision Trees
Neural Networks
K-Means, K-Medoids Fuzzy C-Means
Hierarchical
Gaussian Mixture
Hidden Markov Model
Neural Networks