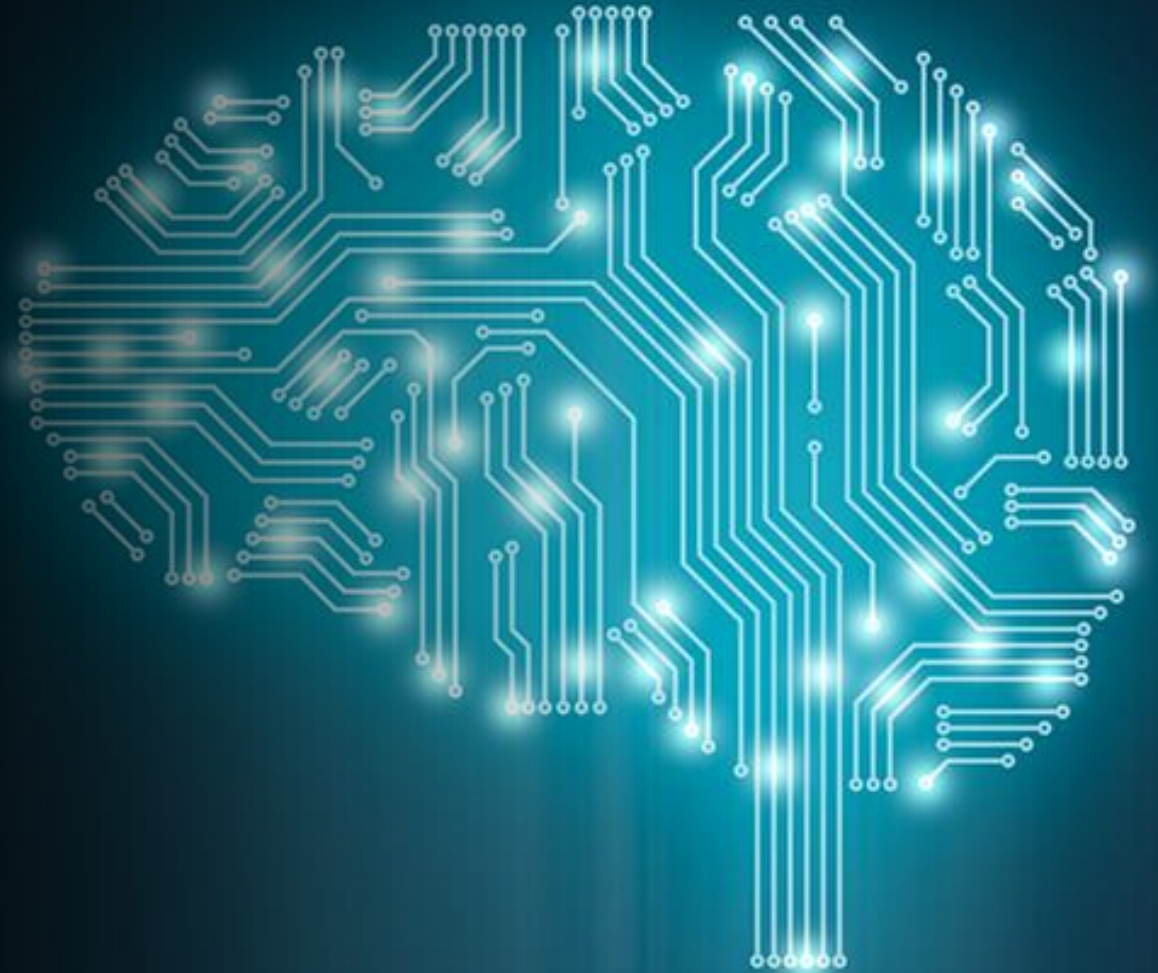


INTRODUCTION TO MACHINE LEARNING

DATA 602

Lecture 1



Why Becoming a Data Scientist?

Sources: Zip Recruiter, Glassdoor

Top 10 Best Jobs in America for 2021

Rank	Job Title	Median Base Salary	Job Satisfaction	Job Openings
1	Java Developer	\$90,830	4.2	10,103
2	Data Scientist	\$113,736	4.1	5,971
3	Product Manager	\$121,107	3.9	14,515
4	Enterprise Architect	\$131,361	4.0	10,069
5	Devops Engineer	\$110,003	4.0	6,904
6	Information Security Engineer	\$110,000	4.0	5,621
7	Business Development Manager	\$82,182	4.1	8,827
8	Mobile Engineer	\$94,301	4.1	4,631
9	Software Engineer	\$110,245	3.8	40,564
10	Dentist	\$134,122	4.0	4,315

Source: Glassdoor Economic Research (Glassdoor.com/research)



Data Scientist Salary in Washington, DC Comparison by Location

Nationwide United States	\$119,413
Washington, DC DC	\$116,118

Artificial Intelligence:

- A branch of computer science dealing with the simulation of intelligent behavior in computers
- The capability of a machine to imitate intelligent human behavior (Webster Merriam)

Machine Learning:

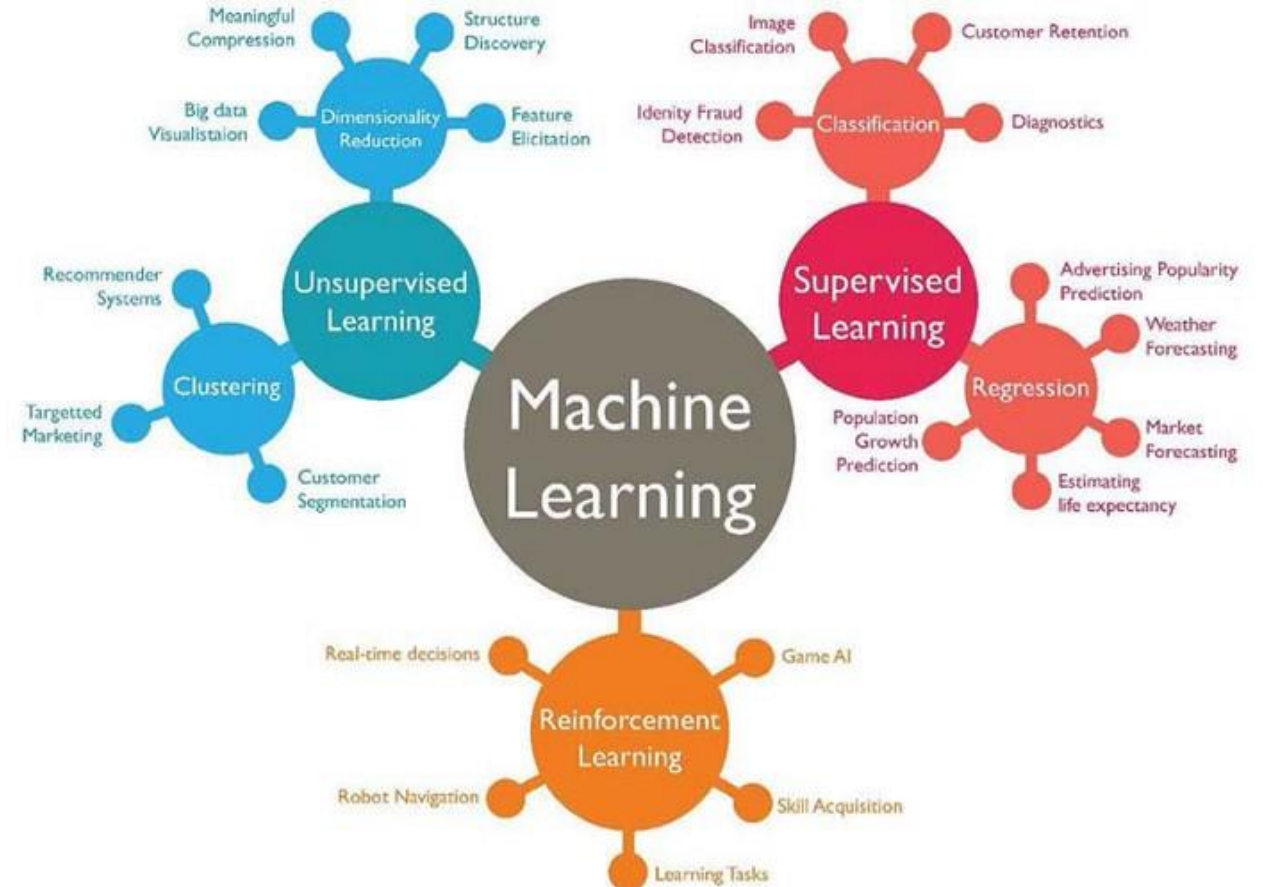
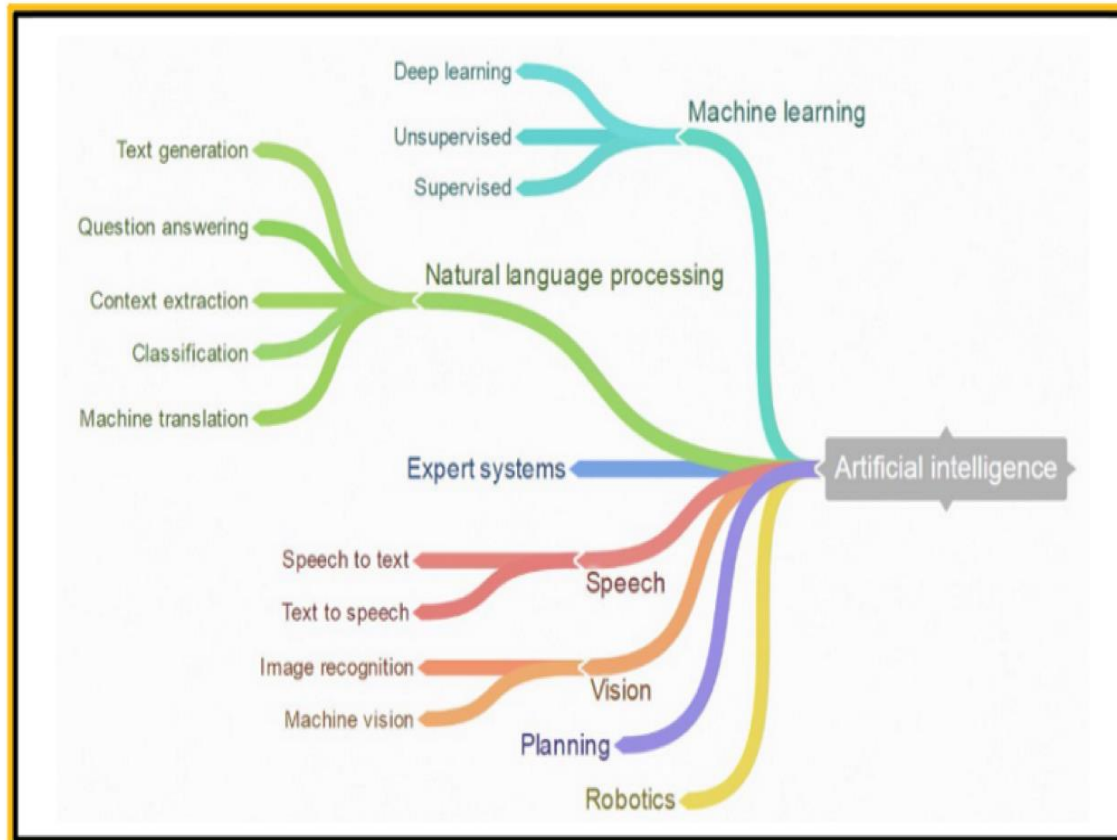
- Machine learning is a method of data analysis that automates analytical model building
- It is a branch of Artificial Intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention (SAS Institute)

Statistics:

- A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data (Webster Merriam)

Data Science:



- Data science is a multidisciplinary blend of **data inference, algorithm development, and technology** in order to solve analytically complex problems (DataJobs)





Source: wordstream

- The goal of machine learning is to take a data sample, and divide it into a training and testing dataset
- The machine learns from the **training** dataset by applying an **algorithm**
 - An algorithm is a **step procedure** to solve logical and mathematical problems
- Once the machine has learned from the **training** dataset, the next step is to compare how well the algorithm will perform based on data the computer has never processed before (different from the training data)
- The comparison between the outcomes from the **training** and **testing** dataset determines the **accuracy** of **predictability**
- **Machine learning** models are designed to make the most accurate predictions possible
- **Statistical models** are designed for inference about the relationship between variables

A Comparison of Machine Learning with AI and Statistics


Attributes	 Machine learning	 AI
Data-driven decisions	✓	
Mimicking humans		✓
Learning	✓	✓
Rule-based agent		✓
Actuation and mechanics		✓

Attributes	 Machine learning	 Statistics
Data-driven decisions	✓	✓
Predictions or decisions	✓	
Patterns	✓	
Computation	✓	
Model learning		✓

Source: MIT

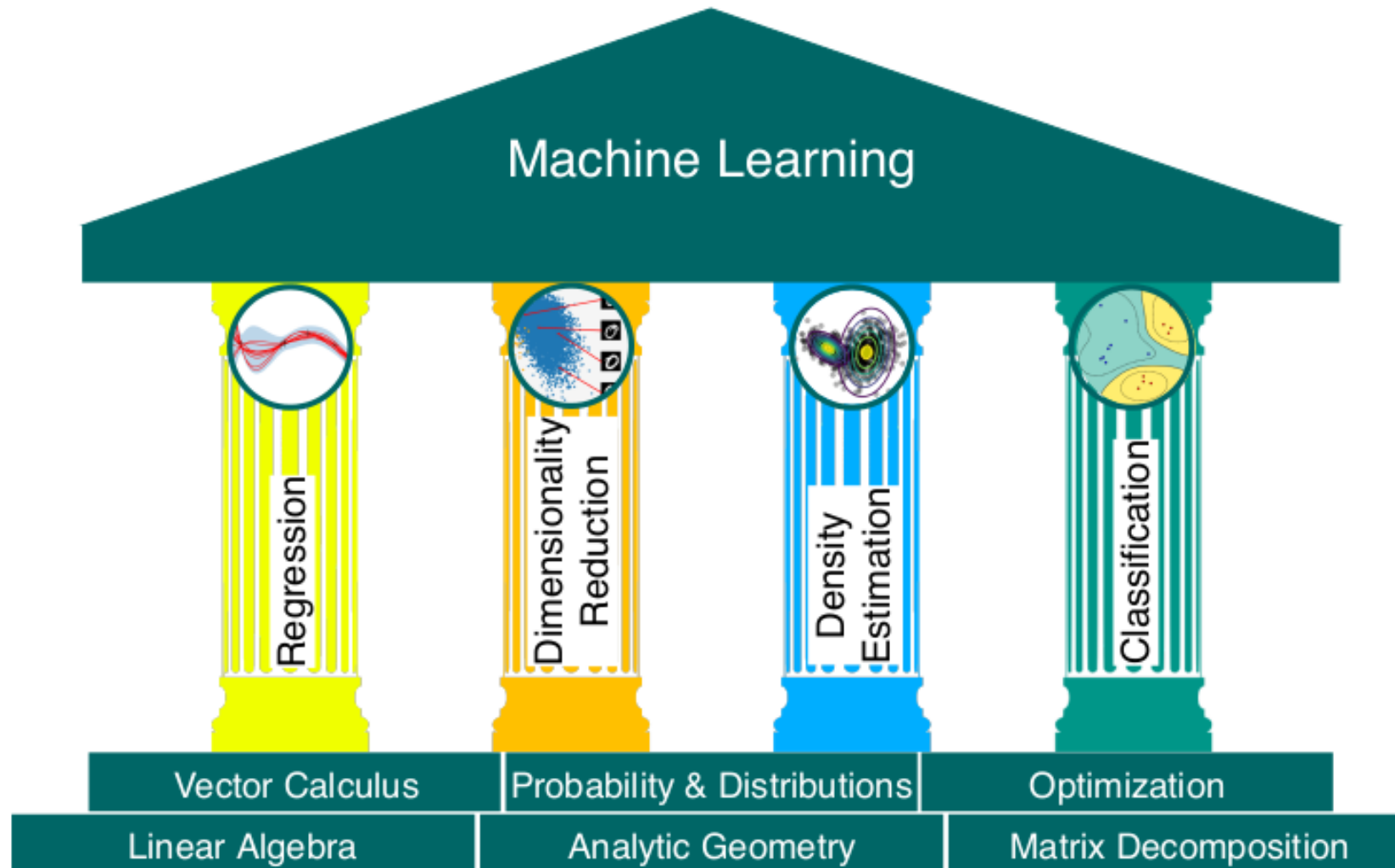
- Statistics is concerned with developing models that characterize, explain, and describe phenomena
- Machine learning is concerned with prediction

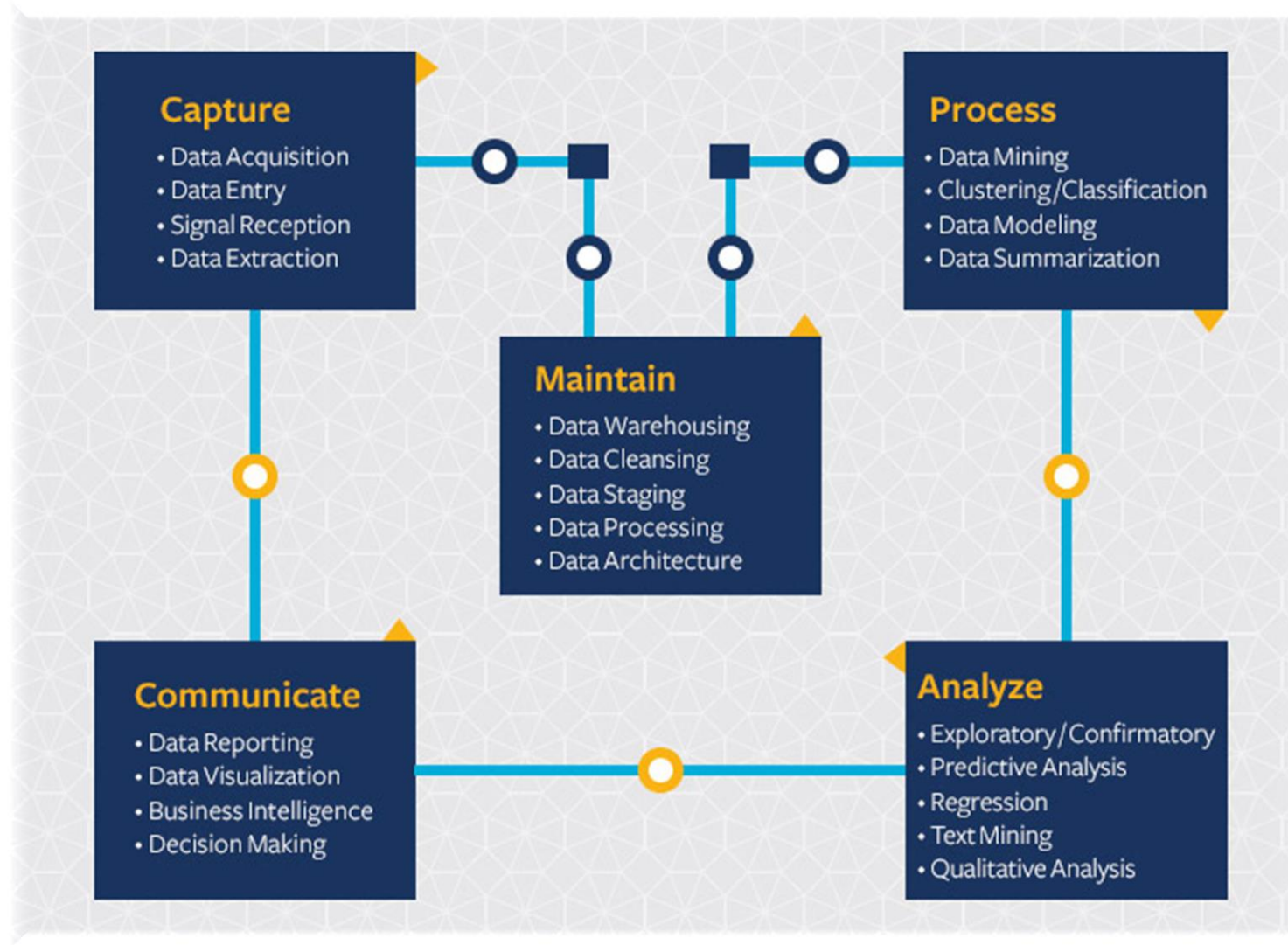
MACHINE LEARNING **VERSUS** **ARTIFICIAL INTELLIGENCE**

Artificial Intelligence	Machine Learning
AI is human intelligence demonstrated by machines to perform simple to complex tasks.	It provides machines the ability to learn and understand without being explicitly programmed.
The idea behind AI is to program machines to carry out tasks in more human ways or smart ways.	The key to teaching computers to think and understand like we do is machine learning.
It is based on characteristics of human intelligence.	It is based on the system of probability.
It is used in healthcare, finance, transportation, aviation, marketing, media, education, etc.	It is used for optical character recognition, web security, imitation learning, etc.
	

	Statistics	Machine Learning
Approach	Data Generating Process	Algorithmic Model
Driver	Math, Theory	Fitting Data
Focus	Hypothesis Testing, Interpretability	Predictive Accuracy
Data Size	Any Reasonable Set	Big Data
Dimensions	Used Mostly for Low Dimensions	High Dimensional Data
Inference	Parameter Estimation, Predictions, Estimating Error Bars	Prediction
Model Choice	Parameter Significance, In-sample Goodness of Fit	Cross-validation of Predictive Accuracy on Partitions of Data
Popular Tools	R	Python
Interpretability	High	Low

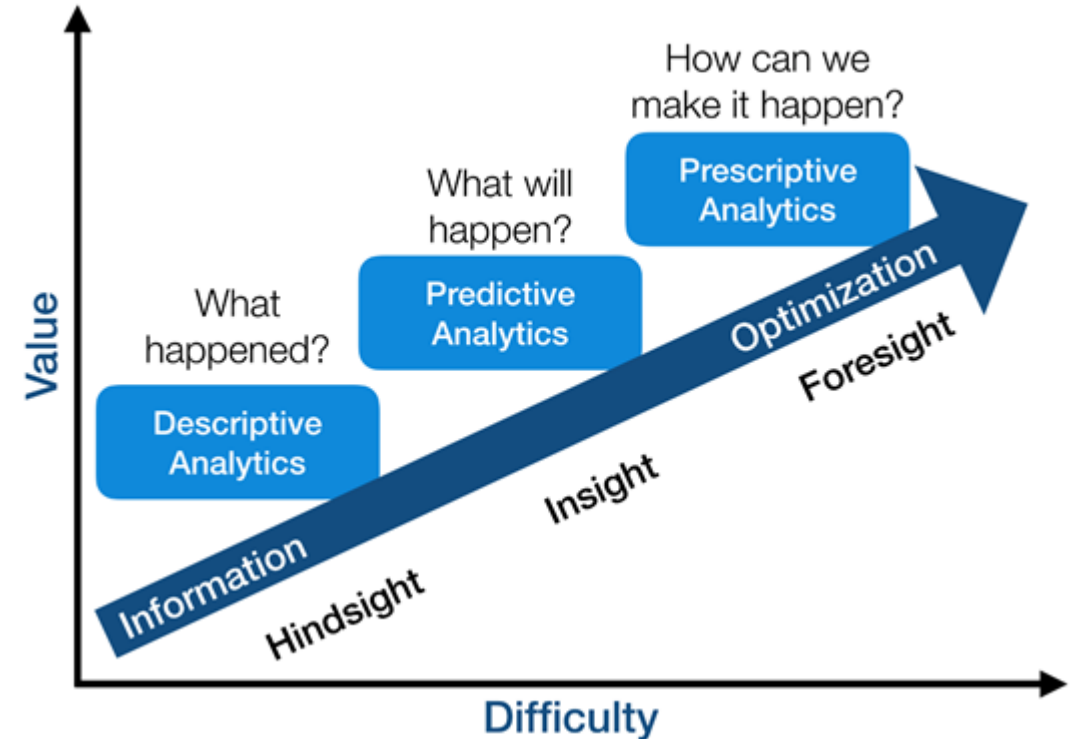
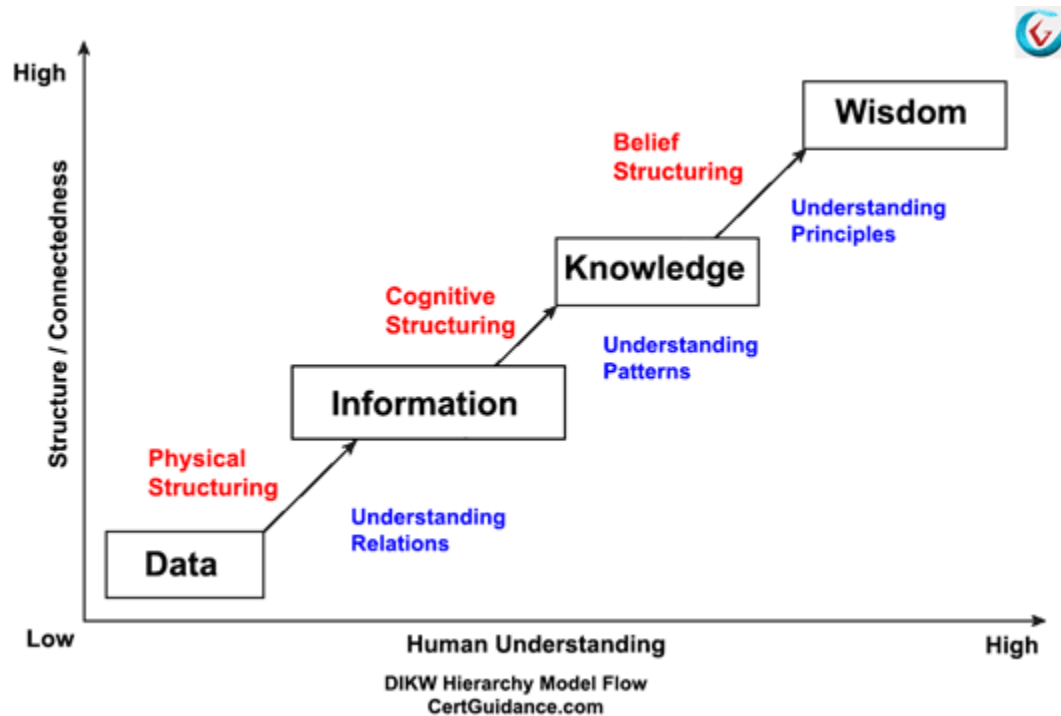
Statistics	Machine Learning
Data Point	Instance
Covariate	Feature
Parameters	Weights
Estimation / Fitting	Learning
Regression / Classification	Supervised Learning
Clustering / Density Estimation	Unsupervised Learning
Response	Label
Test set performance	Generalization





Source: UC Berkley

From Data to Wisdom / From Information to Optimization



- First, we need to **understand the data**
- Second, we make **predictions** using regression, classification, and/or neural networks
- Third, we make decisions under **uncertainty**
- Finally, we determine causal inference to **decide**

Source: DIKW, eduCBA

Supervised Learning: the goal is to learn a specified model from labeled data for training purposes so that predictions can be made about future data or unseen data

- 'Supervised' implies that the labels, or the output signals we want, are already known

There are two types of supervised learning:

- **Classification**→ when the output variable is a category such 'red', 'disease', 'not religious', etc. (**class prediction**)
- **Regression**→ when the output variable is a real value such as 'dollars', 'weight', 'zero or one', or 'interval value' (**continuous outcome prediction**)

Unsupervised Learning: the goal is to explore the data structure to identify patterns or to reduce data dimensions to a few meaningful components that explain most of the variations. Association is another area of unsupervised learning especially in Natural Language Processing with recommender systems

There are two main types of unsupervised learning:

- **Clustering/Dimensionality Reduction**→ involves the discovery of inherent groups in the data, such as grouping by customer type
- **Association**→ involves rules that describe large portions of data, such as recommendations

Some talk about **semi-supervised learning** models including Support Vector Machines, Neural Networks, and Principal Component Analysis (PCA)

Machine Learning Categories (Cont.)

Reinforcement learning is the third type of Machine Learning algorithms. The goal is to develop an agent of system that can learn from experience and improve its performance

- An **agent** learns by interacting with its **environment**
- The **agent** receives **rewards** by performing correctly and **penalties** otherwise
- The agent learns without human intervention how to maximize its rewards and minimize penalties
- It is a type of **dynamic programming** that uses a system of rewards and punishments
- **Reinforcement learning** involves the balance between **exploration** and **exploitation**
- It is important to note that supervised and unsupervised models also apply to the analysis of texts, videos, language, and photographs
- Sentiment analysis, spam detection, image classifications, Natural Language Processing are forms of supervised learning

The two main challenges are **overfitting** and **underfitting** or the **variance/bias trade-off**

- In the case of **overfitting**, the model includes **random error** or **noise** instead of the **underlying relationship**
 - An overfit model features **high variance** and **low bias**. This leads to **poor generalization**
- In the case of **underfitting**, the model has **poor predictive** capabilities because it poorly takes into consideration **non-linearities**
 - An **underfit model** features **high bias** and **low variance**
- A model's generalization error can be expressed as the sum of three very different errors: **bias, variance, and irreducible errors**

Variance

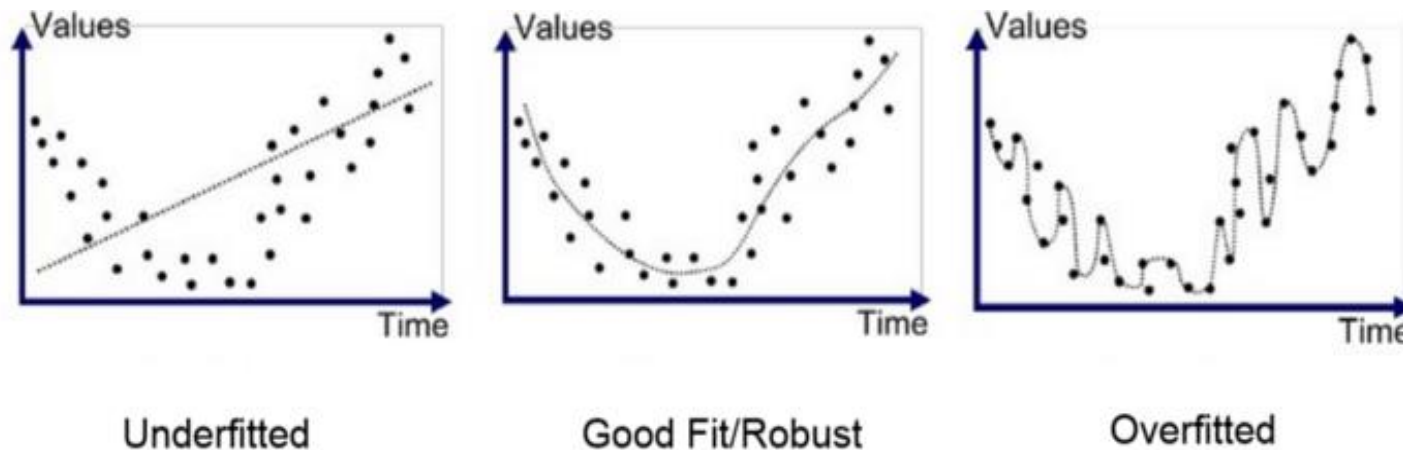
- **Variance** measures how much a model changes in response to the training data
- **Variance** can originate from the model's excessive sensitivity to small variations in the training data
- A model with many degrees of freedom (such as a high-degree polynomial model) is likely to have high variance, and thus, to overfit the training data

Bias

- **Bias** is the other side of variance as it represents the **strength of assumptions about data**
- **Bias** is due to **wrong assumptions**. If we assume the data is linear when it is quadratic, then the model will then underfit the data
- **Bias** refers to **how much we ignore data**, while **variance** implies **how dependent a model is on the data**

Irreducible Errors

- **Irreducible errors** are due to noise in the data itself
- The only way to reduce irreducible errors is to make sure the data is clean i.e., check outliers, missing data
- **Model complexity** will generally increase its variance and reduce its bias and vice versa, hence the trade-off
- A **validation set** for model tuning can prevent underfitting and overfitting
- A program that **generalizes well** will be able to **effectively perform** a task with new data
- In contrast, a program that **memorizes the training data by learning an overly-complex model** could predict the values of the response variable for the training set accurately but will **fail to predict the value of the response variable** for new examples
- **Memorizing** the **training set** is called **overfitting**



Tables

- Data are organized in a table format
- The basic table is a **two-dimensional** grid in which the rows represent the **individual elements** and the columns the **quantities related to each of the elements**
- The **feature matrix** includes two dimensional shapes: the n_{samples} and the n_{features}
- The n_{samples} are the rows of data, while the n_{features} are the columns of data
- The feature matrix is generally stored in a matrix named X
- The one-dimensional **target** or **label array** represents the y variable
- In some cases, it is also possible to have a n_{samples} and n_{targets} array
- The target array is the quantity we would like to predict from the data

Process to Predict the Labels and Measure Model Accuracy in the Case of a Linear Regression Model

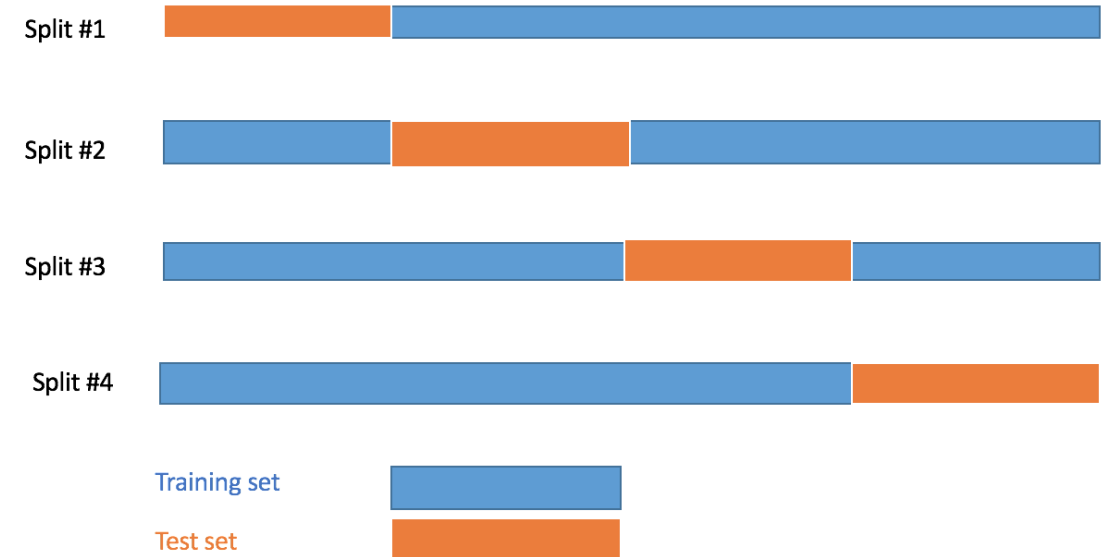
1. We **choose a model class** → `from sklearn.linear_model import LinearRegression`
2. We **instantiate the model** → `model=LinearRegression()`
3. We **split the sample into a training and a test dataset** →
 1. `from sklearn.model_selection import train_test_split`
 2. `X_train, X_test, y_train, y_test=train_test_split(X,y, random_state=1234)`
4. We **fit the model to the data** → `model.fit(X_train, y_train)`
5. We **predict on new data** → `y_model=model.predict(X_test)`
6. We determine the **fraction of the predicted labels** that match **their true value** →
 1. `from sklearn.metrics import accuracy_score`
 2. `print('Accuracy Score: \n', accuracy_score(y_test, y_model))`

- The **test set** is used to:
 - Evaluate the performance of the model using some performance metric
 - It is important that no observation from the training set be included in the test set
- The program is still evaluated on the **test set** to provide an estimate of its **performance in the real world**
- In addition to the training and test data, a third set of observations, called a **validation** or **hold-out set**, is sometimes required
 - The **validation set** is used to tune variables called **hyperparameters** that control how the algorithm learns from the training data
 - The **validation set** should not be used to estimate real-world performance because the program has been tuned to learn from the training data in a way that optimizes its score on the validation data

Cross-Validation

- **What is Cross-Validation?**
- It can be used to train and validate a model on the same data
- It is a statistical method used to estimate the skill of machine learning models
- It is a resampling procedure used to **evaluate** machine learning models on a limited data sample
- The procedure has a single parameter called k that refers to the number of splits in the data sample
 - As such, the procedure is often called k -fold cross-validation
- In cross-validation, the training data is partitioned into k folds
 - The model is trained using all but one of the partitions or folds and tested on the remaining partitions
 - The partitions are then rotated several times so that the model is trained and evaluated on all the data
 - The mean of the model's scores on each of the partitions is a better estimate of performance in the real world than an evaluation using a single training/testing split
- **Cross-Validation** consists in randomly dividing the set of observations into k groups, or folds, of approximately equal size
 - The first fold is treated as a **validation set**, and the method is fit on the remaining $k - 1$ folds

4-fold cross-validation



- **Use Machine Learning to validate your judgments**
 - If a GPS drives you to the wrong location, do you question the judgment of the GPS?
- **What happens if a self-driving car kills someone on the road?**
 - The machine does not know anything about normative values
- **The results are just as good as the model and the data provided**
- **It is difficult to apply Machine Learning in stochastic and deterministic models**
 - The model does not know when it is violating physical laws
- **Statistical modeling is *confirmatory*. Machine Learning modeling is *exploratory***
- **Interpretability** may often be an issue with Machine Learning models
 - **Neural Networks** are often difficult to explain to those who have had no exposure to them
 - If you cannot convince your clients that you understand how the algorithm came to a specific outcome, how likely are they to trust you and your expertise?



- We defined machine learning as the **design of programs** that can **improve their performance** at a task by **learning from experience**
- We discussed the different types of supervision
 - **Supervised learning**: a program learns from inputs that are labeled with their corresponding outputs (**regression** and **classification**)
 - **Unsupervised learning**: a program must discover structure in only unlabeled inputs
 - Unsupervised learning tasks include **clustering**, in which observations are organized into groups according to some similarity measure, **dimensionality reduction**, which reduces a set of explanatory variables to a smaller set of synthetic features that retain as much information as possible, and **association**
 - **Semi-supervised approaches** make use of both labeled and unlabeled training data
- We discussed common types of machine learning tasks
 - **In classification tasks**, the program predicts the value of a discrete response variable from the observed explanatory variables
 - **In regression tasks**, the program must predict the value of a continuous response variable from the explanatory variables
- We also reviewed the **bias-variance trade-off** and discussed **common performance measures** for different machine learning tasks
- Balancing **generalization** and **memorization** is a problem common to many machine learning algorithms
- We will discuss **regularization** or **penalization**, which can be applied to many models **to reduce overfitting**