Artificial Intelligence III:
Artificial Intelligence and Deep Learning

*Lecture 2*

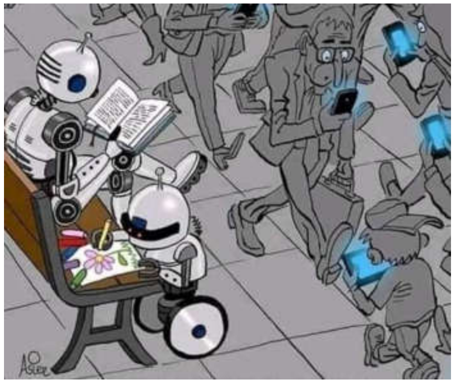
# Fundamental Principles

Dr. Patrick Chan
patrickchan@ieee.org
South China University of Technology, China



## Machine Learning

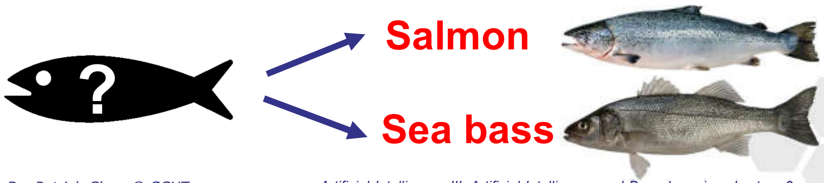
- Design an algorithm is not easy
- What if a machine can learn…





## Machine Learning: Fish Packing Plant Example
## Salmon / Sea Bass

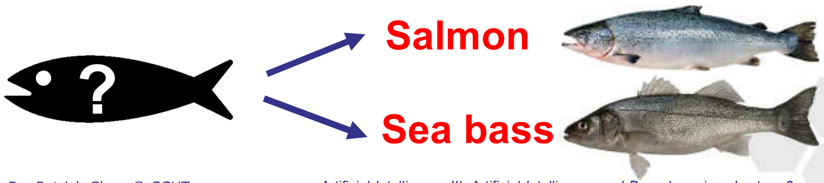
- **Fish Packing Plant** wants to automate the process of sorting incoming fishes (**Salmon** / **Sea Bass**) on a belt according to species
- How to design a system?





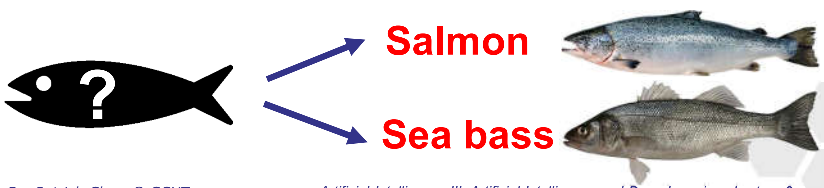

## Machine Learning: Fish Packing Plant Example
## Salmon / Sea Bass

- Rule-base system is commonly used
  - Example: For a Fish
    - If Length > 10 and Fin Area > 10, Sea Bass
    - If Weight > 4.3 or Length < 4, Salmon
  - Factor of a rule:
    - Characteristic (Feature) — Length, Weight, Color, Shape of fin / head,
    - Quantification (Threshold) — Fin area > 10, Sea Bass; Fin area < 10, Salmon



# Salmon / Sea Bass

- Difficult to determine a rule manually, even for an expert
  - How to define the rules (Feature and Threshold)?
  - E.g. If Weight > 4.3 or Length < 4, Salmon
- Machine Learning can help
  - You do not know how but implement an algorithm which can learn from data








# Process

Fish

↓ *Object*

Sensing (camera)

↓ *Image*

Preprocessing (Isolate Fish, reduce noise…)

↓ *Refined Image*

Feature Extraction (Take Measurement)

↓ *Input Features*

Classification

↓ *Output*

Class (Salmon / Sea Bass)





# Process

- **Sensing**
  - **Digitize** the object to the format which can be handled by machines
  - Example
    - Type of Device
      Camera? Depth Camera? Infra-red? Ultrasound? Movement Sense? Combination?
    - Setting of Device
      Number? Angle? Overlap shooting range?
    - Background
      Lighting? Background simplicity?

Object → **Sensing** → Preprocessing → Feature Extraction → Decision Making → Class





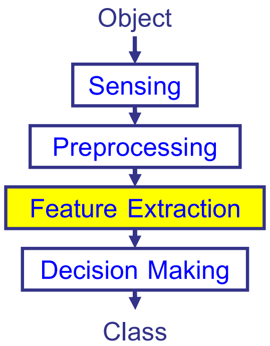
# Process

- **Preprocessing**
  - **Refine** the data
  - Example
    - Lighting conditions
    - Position of fish
    - Angle of fish
    - Noise
    - Blurriness
    - Segmentation (remove object from background)

Object → Sensing → **Preprocessing** → Feature Extraction → Decision Making → Class

# Process

- **Feature Extraction**
  - Decide which **information** is able to distinguish classes
  - Example
    - Length, width, weight, number and shape of fins, tail shape, etc.
  - Rely on technical background and common sense
    - Experts may help

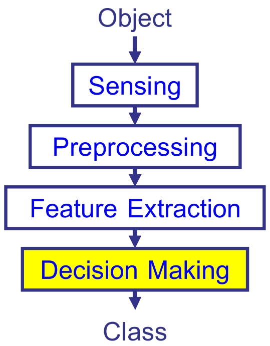







# Process

- **Decision Making**
  - Decision Type:
    - Class (Classification)
    - Value (Regression, Value Prediction)
    - Rank (Ranking)
    - Action (Reinforcement Learning)
    - Region (Segmentation)
  - Many machine learning techniques are available

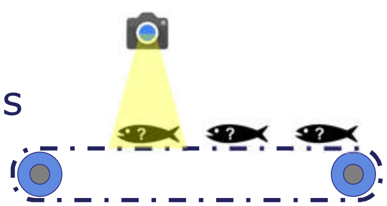




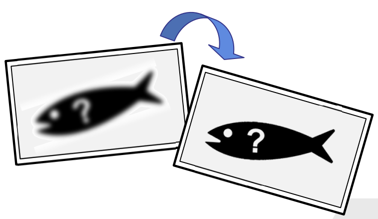



# Sensing & Preprocessing

- **Sensing**
  Assume a fish is put on a belt and a single camera is installed to take a photo on each fish



- **Preprocessing**
  Remove the blueness and noise







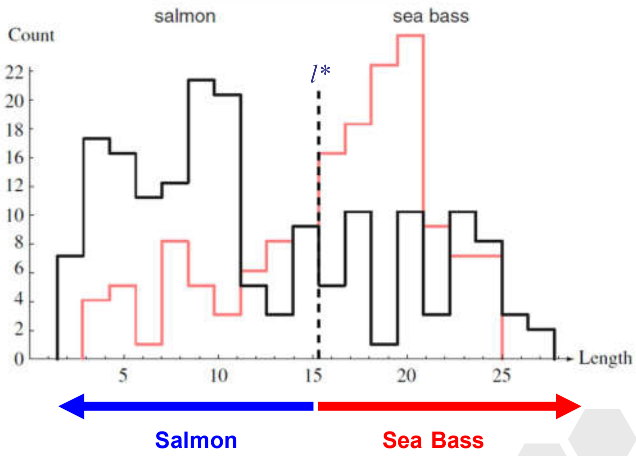
# Feature Extraction

- The expert (e.g. Fisherman) suggests salmon is usually shorter than sea bass
- Length is chosen (as a feature) as a decision criterion

# Feature Extraction

- 15 is selected as the threshold
- Although sea bass is longer in general, there are many exceptions
- The experts "may be" wrong!
- How about other features?
- E.g. lightness

Histograms of the length feature for sea bass and salmon



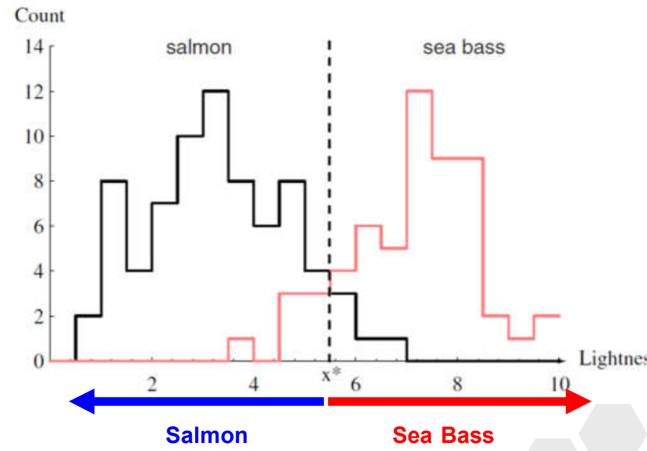





# Feature Extraction

- Try another feature "Lightness"
- 5.5 is selected as the threshold
- "lightness" is better than "length"

Histograms for the lightness feature for sea bass and salmon








# Cost Consideration

- Besides accuracy, "**costs of different errors**" can be considered
- **Case 1: Company's view**

  Salmon \$ > Seabass \$

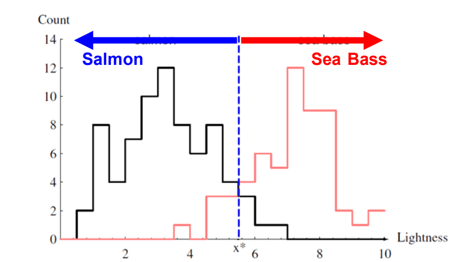
  - Salmon is more expensive than sea bass. Selling Salmon with the price of sea bass will be a loss
    - If salmon is classified as sea bass : **HIGH** cost
    - If sea bass is classified as salmon : **LOW** cost
- **Case 2: Customer's view**
  - Customers who buy salmon will be upset if they get sea bass; Customers who buy sea bass will not be upset if they get the more expensive salmon
    - If salmon is classified as sea bass : **LOW** cost
    - If sea bass is classified as salmon : **HIGH** cost

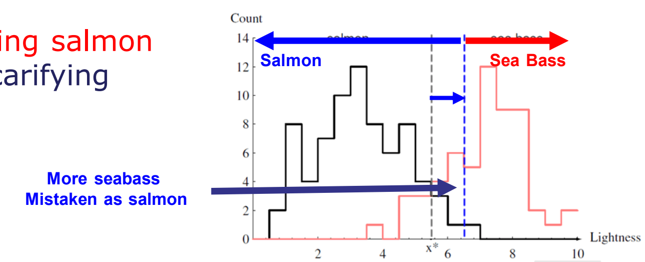




# Cost Consideration

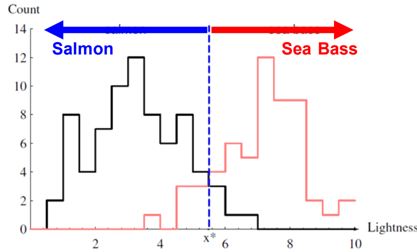
- **Case 1: Company's view**
  - **HIGH** cost
    Salmon is classified as sea bass
  - **LOW** cost
    Sea bass is classified as salmon
  - Avoid classifying salmon wrongly by scarifying sea bass

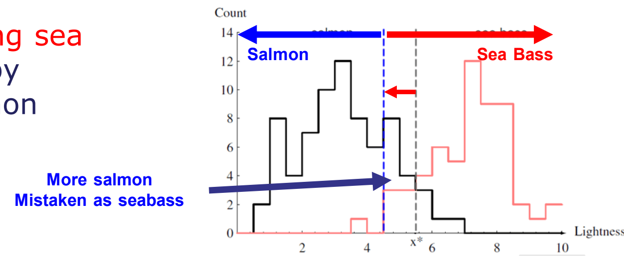


# Cost Consideration

- **Case 2: Customer's view**
  - **LOW** cost
    Salmon is classified as sea bass
  - **HIGH** cost
    Sea bass is classified as salmon
  - Avoid classifying sea bass wrongly by scarifying salmon

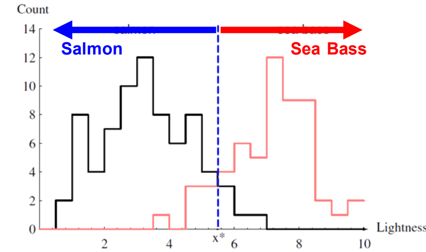


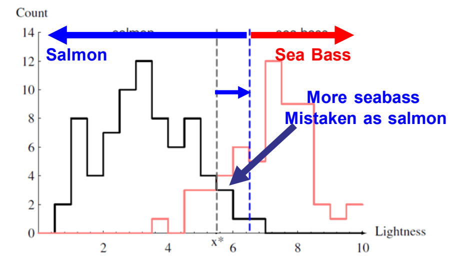




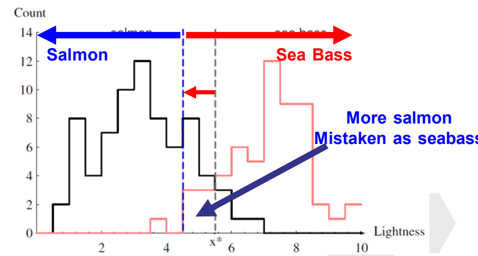

# Cost Consideration




Case 1




Case 2





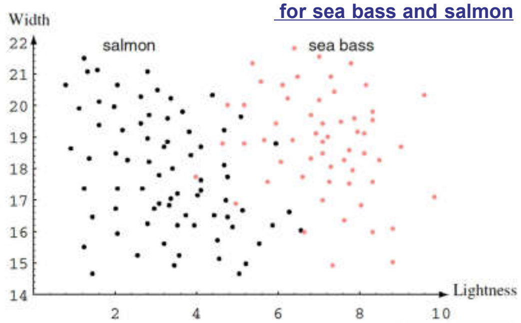



# Multiple Features

- Only ONE feature may not be good enough
- **More features** should be considered
- Two features: Lightness ($x_1$), Width ($x_2$)
- A fish is represented by a point in a 2D feature space:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

The two features (lightness and width) for sea bass and salmon



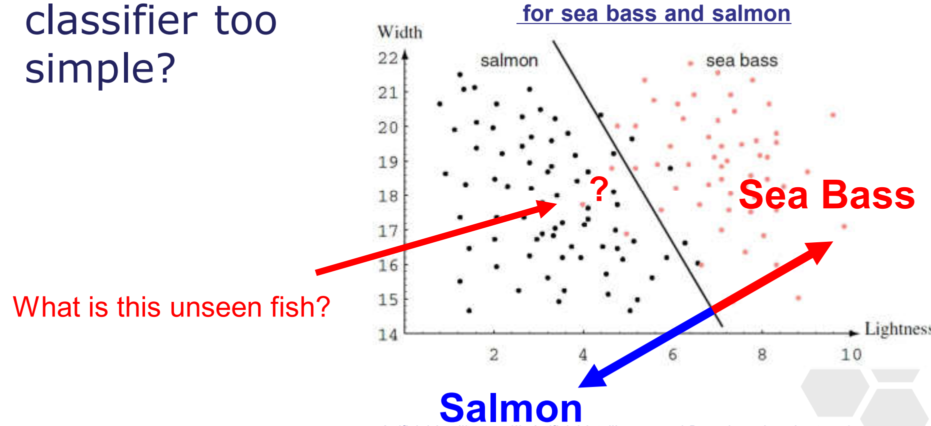





# Classifier

- A **decision boundary** can be drawn to divide the feature space into two regions
- Is it a linear classifier too simple?

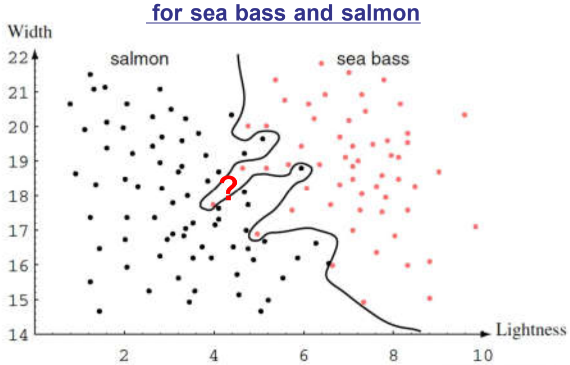
The two features (lightness and width) for sea bass and salmon



## Classifier

- Will other classifiers be better?
  - More complex classifier
- Perfectly classify training samples
- Ultimate objective is to classify unseen samples correctly
- Can it be generalized to unseen sample?

The two features (lightness and width) for sea bass and salmon



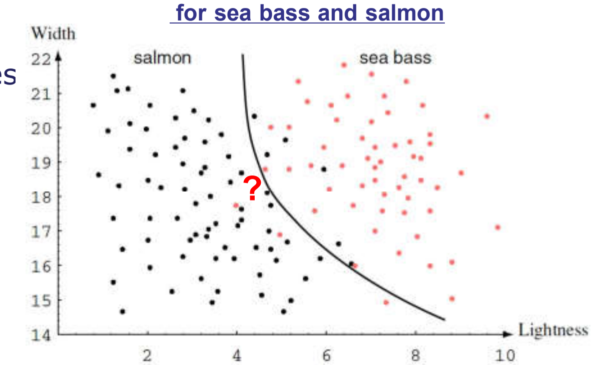





## Classifier

- **Tradeoff** between accuracy of training samples and complexity
- Look more reasonable
  - Not too complex
  - Good in classifying the training samples

The two features (lightness and width) for sea bass and salmon








## Summary

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

- Task T : Separate Salmon and Sea Bass
- Performance P : Accuracy on identification
- Experience E : Caught Salmon and Sea Bass



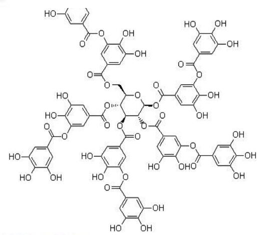
## Key Concepts

- Data Type
- Dataset (Collected Samples)
- Data Cleaning
- Model Requirement
- Evaluation
- Comparison
- Terminology

# Key Concepts: Data Type

- Record-based data
  - Data matrix
  - Document data
  - Transaction data
- Graph-based data
  - World wide web
  - Molecular structure
  - Map data
- Ordered data
  - Spatial data
  - Temporal data
  - Sequential data
  - Genetic/Genomic sequence data

| Sale ID | Time | Customer | Product ID | Quantity |
|---|---|---|---|---|
| S00001 | 12/1/2012 9:00:00 AM | C0001 | P025 | 1 |
| S00002 | 12/1/2012 9:05:58 AM | C0025 | P025 | 3 |
| S00003 | 12/1/2012 9:11:33 AM | C0010 | P001 | 2 |
| S00004 | 12/1/2012 9:17:16 AM | C0017 | P023 | 4 |
| S00005 | 12/1/2012 9:23:04 AM | C0018 | P016 | 5 |
| S00006 | 12/1/2012 9:28:43 AM | C0011 | P018 | 4 |
| S00007 | [illegible] | [illegible] | [illegible] | [illegible] |







# Record-based Data

- Common illustration of Data
  - Excel
  - Traditional Database
- Properties / features of a specific object
- For example, human
  - Eye size (mm unit)
  - Eye color
  - Skin color
  - Height (cm unit)
  - Wear glasses or not
  - Gender
  - Age
  - Length of finger (cm unit)
  - ...

Attributes, Characteristics, Features, Variables, ..., etc

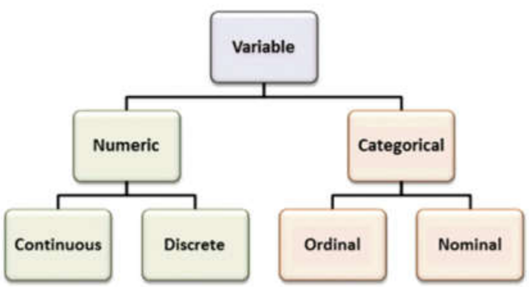
Object, Instance, Individual, Sample, Subject ...

| *Tid* | Refund | Marital Status | Taxable Income |
|---|---|---|---|
| 1 | Yes | Single | 125K |
| 2 | No | Married | 100K |
| 3 | No | Single | 70K |
| 4 | Yes | Married | 120K |
| 5 | No | Divorced | 95K |
| 6 | No | Married | 60K |
| 7 | Yes | Divorced | 220K |
| 8 | No | Single | 85K |
| 9 | No | Married | 75K |
| 10 | No | Single | 90K |





# Record-based Data: Variable



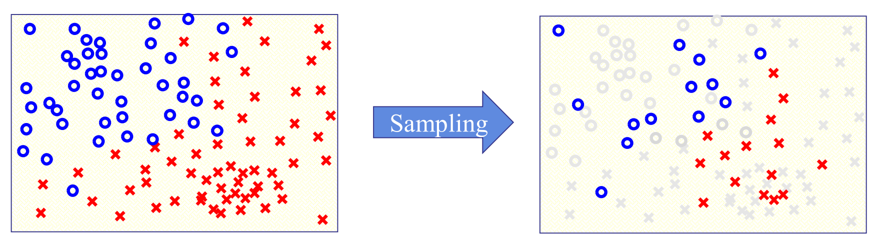

- Continuous: real number, e.g. height = 167.23cm
- Discrete: integer, e.g. age = 1
- Nominal: a natural order or rank, e.g. High Low
- Ordinal: no order, e.g. Red Blue Yellow





# Dataset (Collected Samples)

- Aim of machine learning
  Perform well on all samples
- What information we have? Collected samples




**Population**

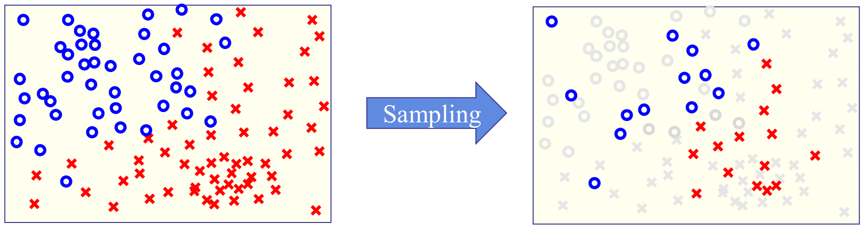
- All possible samples (usually infinite)
- Usually represented by a distribution

**Collected Samples**

- Limited number
- Subset of population

Key Concepts

# Dataset (Collected Samples)



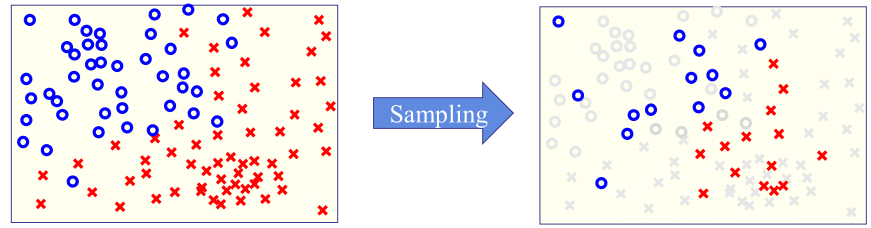

**Population** **Collected Samples**

- Any samples can be chosen?
  - Distribution of collected samples should be similar to the population's one
  - Independent and identically distributed (iid) is assumed
    - Randomly choose without any bias



Key Concepts

# Dataset (Collected Samples)



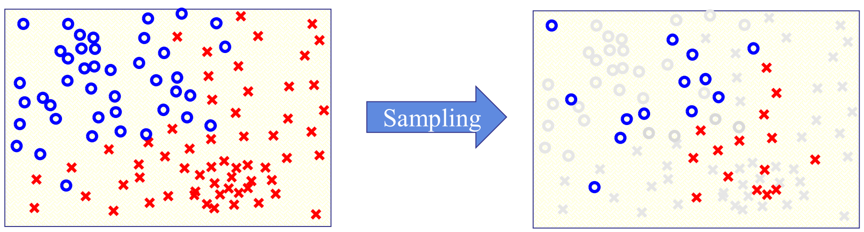

**Population** **Collected Samples**

- More samples are better?
  - Usually yes when the sampling method is fair



Key Concepts

# Dataset (Collected Samples)




**Population** **Collected Samples**

- How many samples are enough?
  - Depend on the complexity of the problem
  - Rely on the performance of the trained model
  - General answer: more is better



Key Concepts

# Data Cleaning

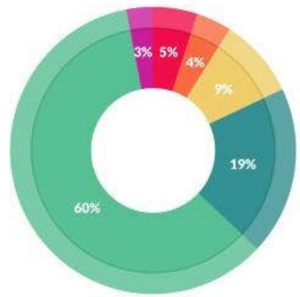
- Once you have finished collecting samples, can we start the learning immediately?

- Garbage in Garbage out

Key Concepts

# Data Cleaning

- What data scientists spend the most time doing?
  - Building training sets 3%
  - **Cleaning and organizing data 60%**
  - Collecting data set 19%
  - Mining data 9%
  - Refining algorithms 4%
  - Other 5%






Key Concepts: Data Cleaning

# Bad Quality Data

- Identifying damaged or inaccurate, incomplete, incorrect, or irrelevant parts within data
- Replacing, modifying, or deleting the dirty or rough data
- **Quality**
  - Noise
  - Outliers
  - Missing values
  - Duplicate data




Key Concepts: Data Cleaning

# Noise

- Noise can refer to any random fluctuation in a signal
- Attribute noise
  Errors in features
  - **Value errors**
    Incorrect or erroneous values
  - **Missing values**
    Data entries that are absent
  - **Irrelevant values**
    Does not contribute
- Class noise
  Variability or errors in the labels
  - **Contradictory examples**
    Identical features but different labels.
  - **Mislabeled examples**
    label is incorrect

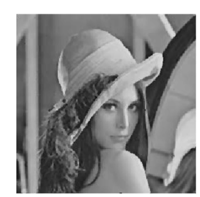
| Feature 1 | Feature 2 | Class |
|---|---|---|
| 0.25 | Red | + |
| 0.25 | Red | - |
| 0.99 | Yellow | - |
| 1.02 | Yellow | + |
| 2.05 | Yellow | - |

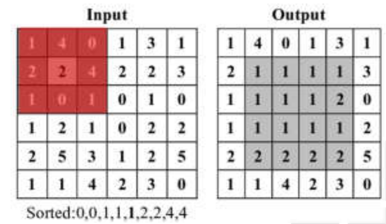


Key Concepts: Data Cleaning

# Noise

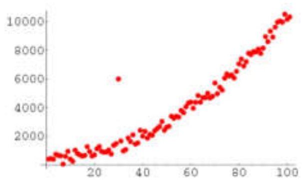
- Median/mean noise filter
  - Apply a sliding window to an image
    - Sort the pixel values within the window
    - Calculate the median/average value from the sorted values as the new value for the current pixel
  - Repeat until all the pixels are processed

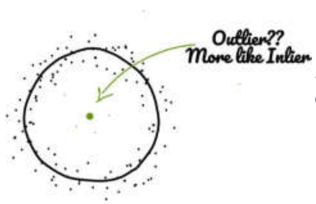


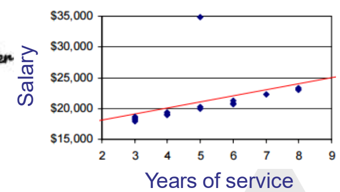


# Outlier

- An outlier is a data point that differs significantly from other observations
- Identification:
  - **Subjective**: visualization
  - **Objective**: statistical way, i.e. Cook's D value








# Outlier

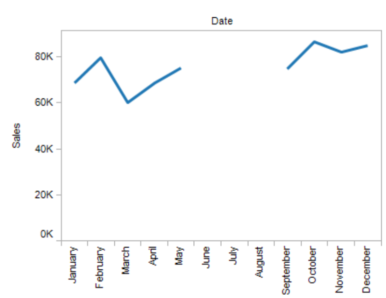
- Deletion
  - If the outlier is caused by human error (i.e. typo, unrealistic response)
- Replacement
  - Replace observations with other values (mean, etc.)
- Handle Differently
  - Analyze outliners separately from the rest

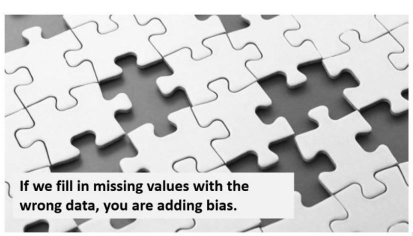




# Missing Value

- No data value is stored
- Missing data are a common occurrence and can have a significant effect on the conclusions








# Missing Value

- Deletion
  - Delete the whole observation with missing values
  - Partial deletion (delete the part of missing values in downstream modeling)
- Replace with other values
  - mean, mode, median
  - Possibility that the model will be distorted
- Insert predicted values
  - Imputation by using statistical or machine learning methods

Key Concepts: Data Cleaning

# Duplicate Data

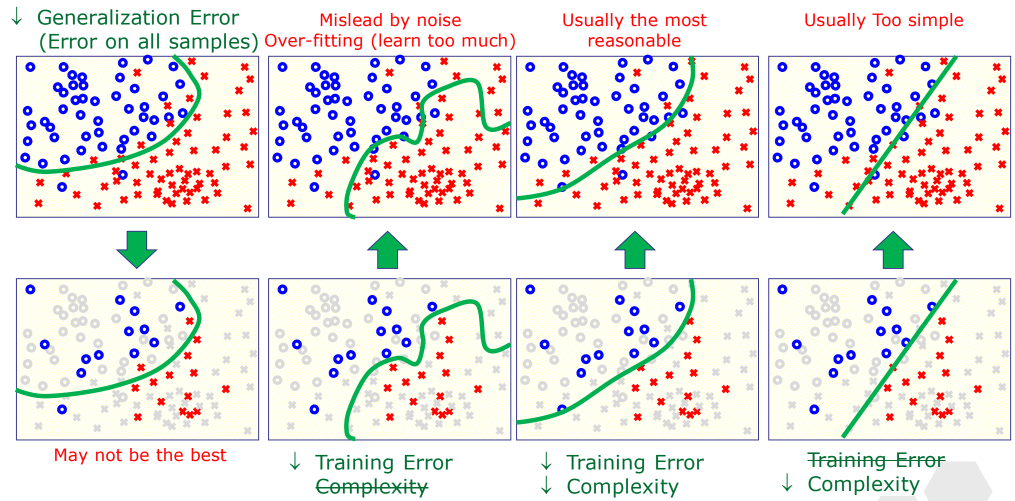
- Objects that are duplicates, or almost duplicates of one another
- Common issue when collecting data from heterogeneous sources
  - E.g. If a person has multiple email addresses
- Deletion
  - Which record should be deleted?
    - Time
    - Source Trustworthy
    - Majority



Key Concepts

# Model Requirements

**Population**

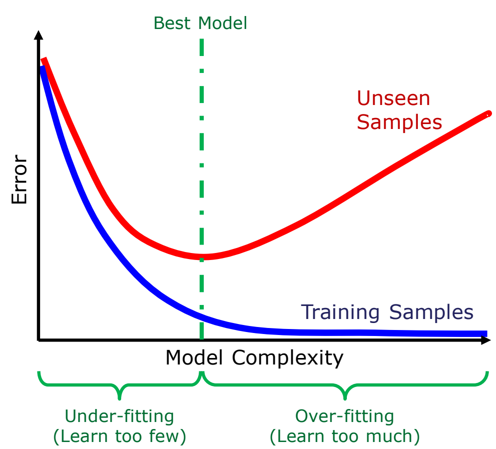



**Collected Samples**



Key Concepts

# Model Requirements




- Be noted that this is just a common situation in machine learning, where each application varies
  - E.g. Having more training samples reduces the variance between two lines.



Key Concepts

# Evaluation

- How to evaluate our trained model?
  - Evaluate by training samples?
    No, already see in training
  - What information we have?
    Collected samples
  - Some collected samples should not be used in training
  - Separate into two non-overlapping sets:
    - Training set : For training
    - Test set : For evaluation

Key Concepts

# Comparing Classifiers

- For a classification problem, given
  - Dataset $D$
  - Classifiers $A$ and $B$
- How can we measure which classifier, A or B, is better for $D$?



Key Concepts

# Comparing Classifiers

- **Method**
  - Randomly separate $D$ into training and test sets
  - Use Training Set to train $A$ and $B$
  - Use Test Set to measure the performances of the trained $A$ and $B$
  - Select the better performing classifier
- Is it ok?
  - The winner may just be lucky in performing better for that particular test set.
  - No guarantee for different test sets



Key Concepts

# Comparing Classifiers

- The bias of test set should be reduced
- Two re-sampling techniques
  - Independent Run
  - Cross-Validation



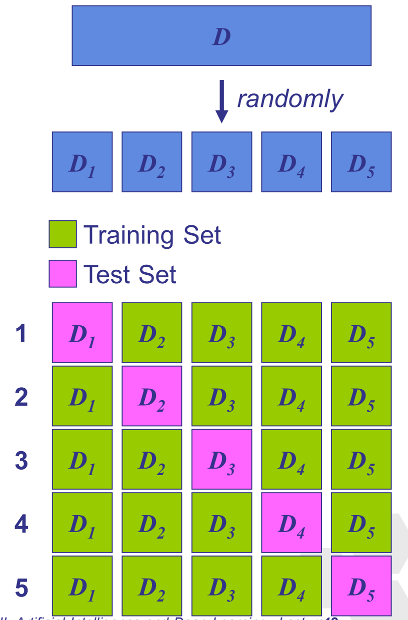
Key Concepts: Comparing Classifiers

# Independent Run

- Statistical method
- Also called Bootstrap and Jackknifing
- Repeat the experiment "n" times independently
  - Repeat $n$ times
    - $i$ is the number of running time
    - Randomly separate $D$ into Training Set$_i$ and Test Set$_i$
    - Use Training Set$_i$ to train $A_i$ and $B_i$
    - Use Test Set$_i$ to evaluate the trained $A_i$ and $B_i$
  - Select the classifier with higher average accuracy

# Cross-Validation

- M-fold Cross-Validation
- Dataset D is randomly divided into $m$ disjoint sets $D_i$ of equal size $n / m$, where $n$ is the number of samples in dataset
- Repeat $m$ times
  - Trained by $D_j$
  - Evaluated by all $D_i$ except $D_j$
- Select the classifier with higher average accuracy








# Terminology

- **Instance / Sample**
  Observations from an application
- **Feature / Attribute**
  Property or characteristic of a sample
- **Dimensionality**
  The number of features





# Terminology

- **Training Set**
  A set of samples used to train a model
- **Test Set**
  A set of samples used to evaluate the performance of the trained model. Usually separate from the training set.
- **Unseen Samples**
  Any samples not in training set





# Terminology

- **Training Error**
  Error on training samples
- **Test Error**
  Error on test samples
- **Generalization Error**
  The ability of a model to perform well on unseen samples

  In some discussion,
  Test Error = Generalization Error

# Terminology

- **Objective Function / Error Function**

  A mathematical function used to quantify error made by a model, closely related to the objective

  Can be more than error on samples, may include any other concepts

  E.g. complexity of a model