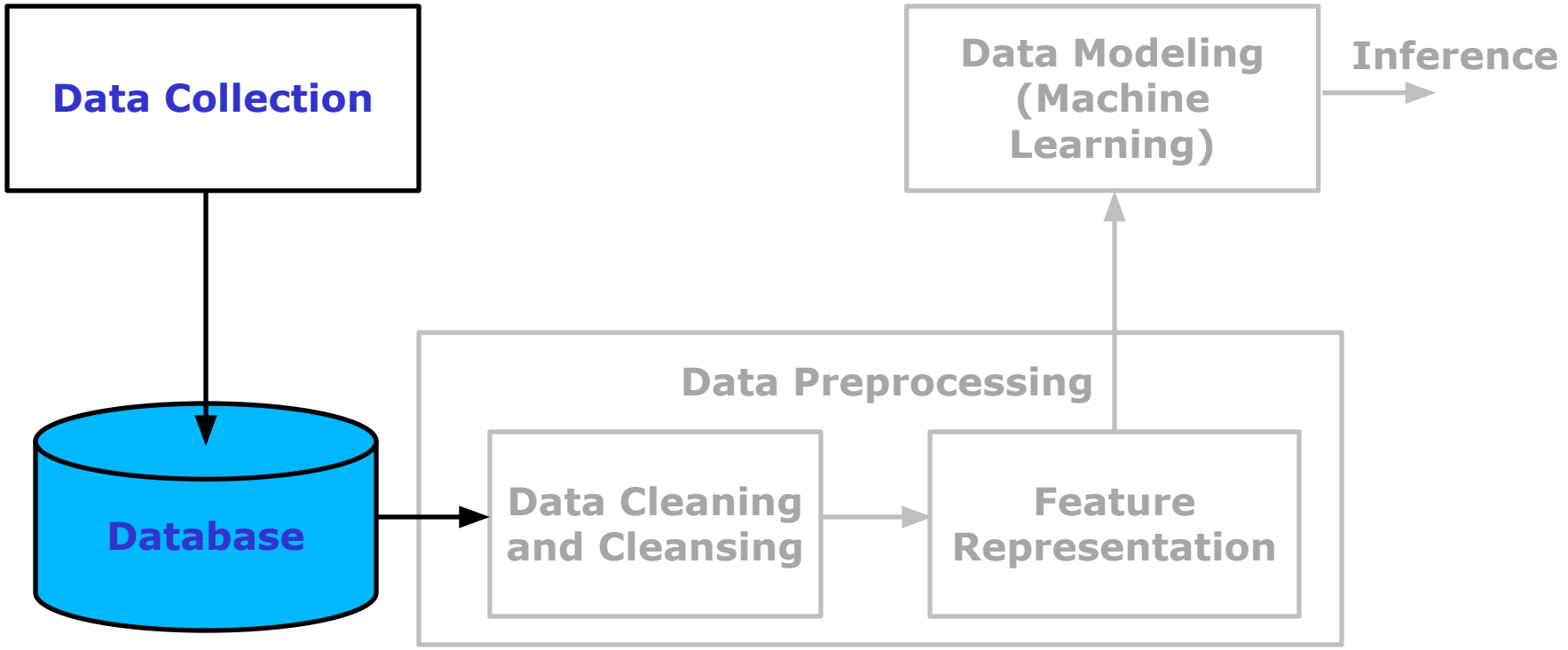
Introduction to Machine Learning: Data Modeling

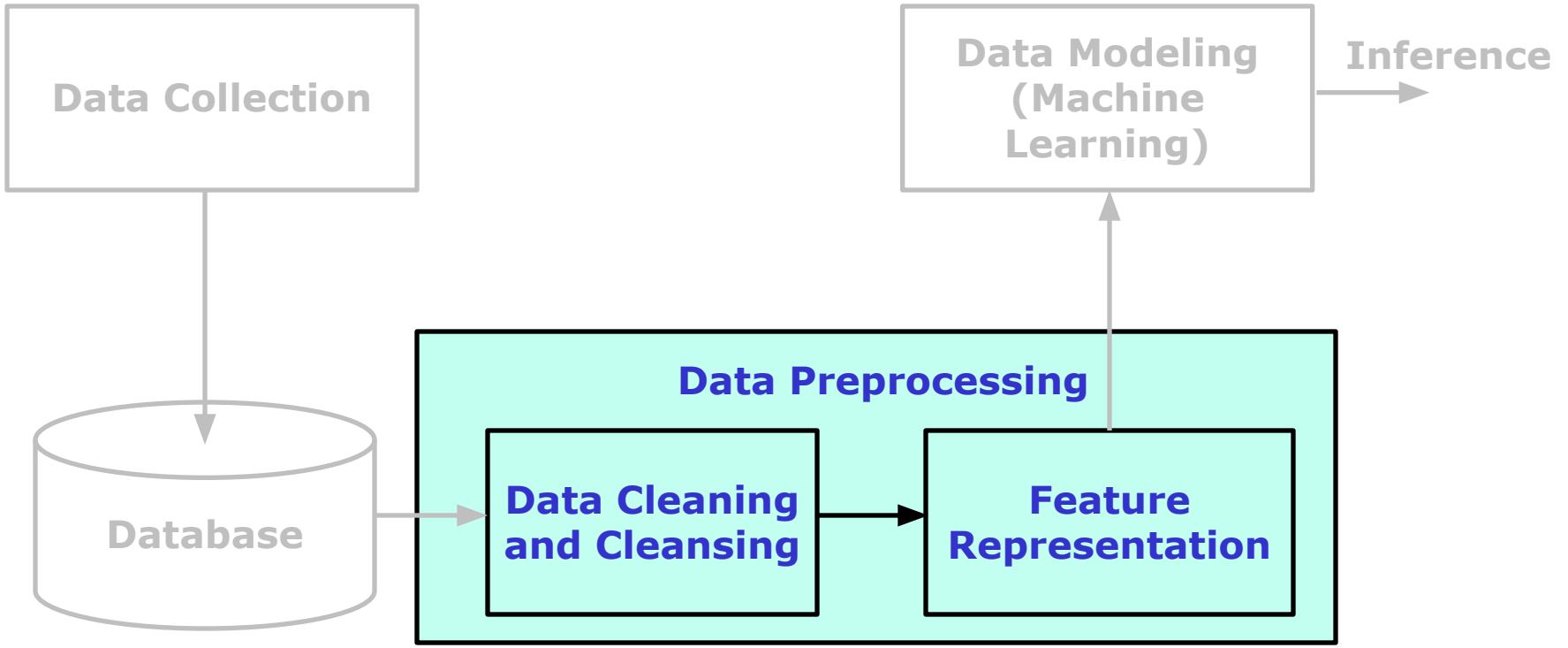
Data Science

- Multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insight from structured and unstructured data
- Central concept is gaining insight from data



Data Science

- Multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insight from structured and unstructured data
- Central concept is gaining insight from data

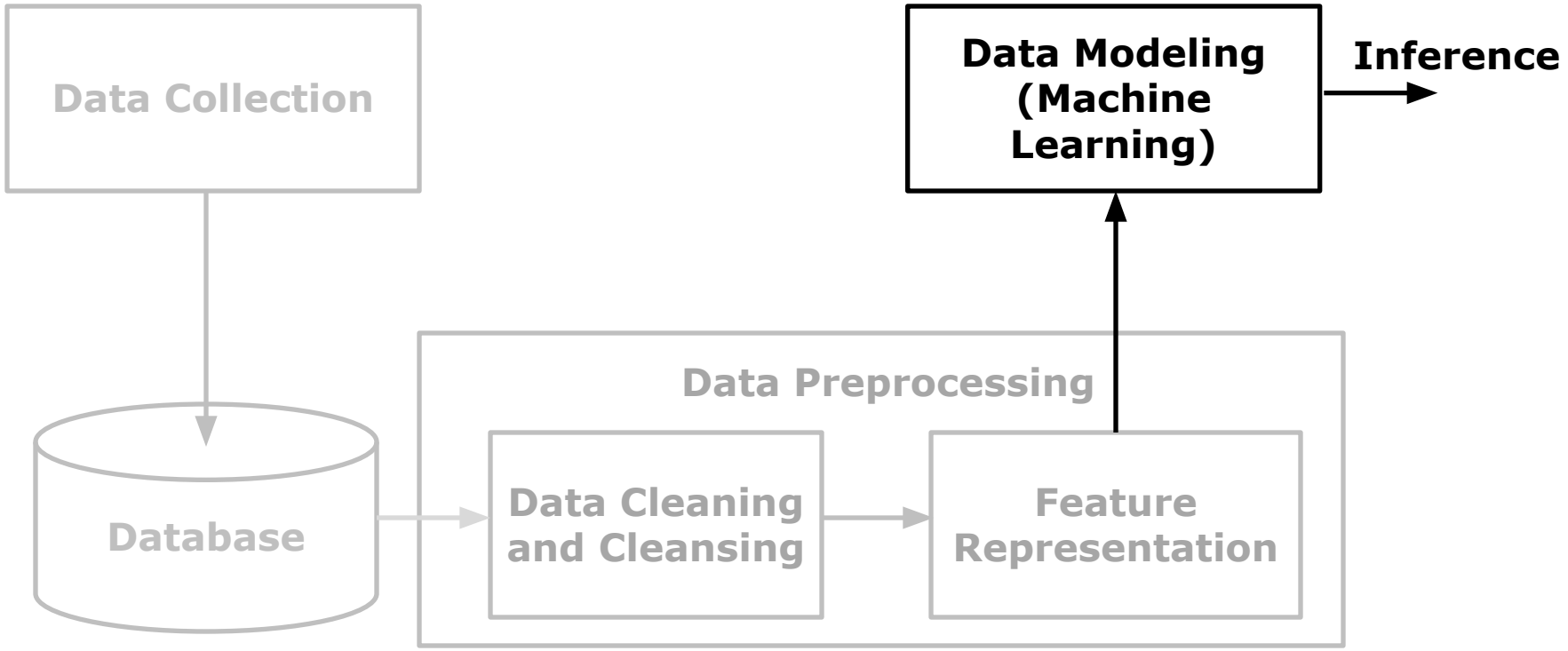


Data Preprocessing and Descriptive Data Analytics

- Data preprocessing involve:
  - Data cleaning, Data integration, Data transformation, Data reduction
- Descriptive data analytics serves as a foundation for data preprocessing
- It helps us to study the general characteristics of data and identify the presence of noise or outliers
- Data characteristics:
  - Central tendency of data
    - Centre of the data
    - Measuring mean, median and mode
  - Dispersion of data
    - The degree to which numerical data tend to spread
    - Measuring range, quartiles, interquartile range (IQR), the five-number summary and standard deviation
- Descriptive analytics are the backbone of reporting

Data Science

- Multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insight from structured and unstructured data
- Central concept is gaining insight from data
- Machine learning uses data to extract knowledge – predictive analytics



Predictive Data Analytics

- It is used to identify the trends, correlations and causation by learning the patterns from data
- Study and construction of algorithms that can learn from data and make predictions on data
- It involve tasks like
  - Classification: Categorical label prediction
    - E.g.: predicting the presence or absence of disease or
    - predicting the category of the disease according to symptoms
  - Regression: Numeric prediction
    - E.g.: predicting the amount of landslide or
    - predicting the amount of rainfall
  - Clustering: Grouping of similar patterns
    - E.g.: grouping the similar items to be sold or
    - grouping the people from the same region
- Learning from data

Machine Learning:
Learning from Data

- 1, 2, 3, 4, 5, ?, …, 24, 25, 26, 27, ?
- 1, 3, 5, 7, 9, ?, …, 25, 27, 29, 31, ?
- 2, 3, 5, 7, 11, ?, …, 29, 31, 37, 41, ?
- 1, 4, 9, 16, 25, ?, …, 121, 144, 169, ?
- 1, 2, 4, 8, 16, 32, ?,…, 1024, 2048, 4096, ?
- 1, 1, 2, 3, 5, 8, ?, …, 55, 89, 144, 233, ?
- 1, 1, 2, 4, 7, 13, ?, 44, 81, 149, 274, 504, ?
- 3, 5, 12, 24, 41, ?, …., 201, 248, 300, 357, ?
- 1, 6, 19, 42, 59, ?, …, 95, 117, 156, 191, ?

- 1, 2, 3, 4, 5, 6, …, 24, 25, 26, 27, 28
- 1, 3, 5, 7, 9, 11, …, 25, 27, 29, 31, 33
- 2, 3, 5, 7, 11, 13, …, 29, 31, 37, 41, 43
- 1, 4, 9, 16, 25, 36, …, 121, 144, 169, 196
- 1, 2, 4, 8, 16, 32, 64,…, 1024, 2048, 4096, 8192
- 1, 1, 2, 3, 5, 8, 13, …, 55, 89, 144, 233, 377
- 1, 1, 2, 4, 7, 13, 24, 44, 81, 149, 274, 504, 927
- 3, 5, 12, 24, 41, 63, ….., 201, 248, 300, 357, 419

  (2, 7, 12, 17, 22, 27, 32, 37, 42, 47, 52, 57, 62)
- 1, 6, 19, 42, 59, ?, …, 95, 117, 156, 191, ?

- **Pattern: Any regularity or structure in data or source of data**
- **Pattern Analysis: Automatic discovery of patterns in data**

Image Classification

Tiger

Giraffe

Horse

Bear

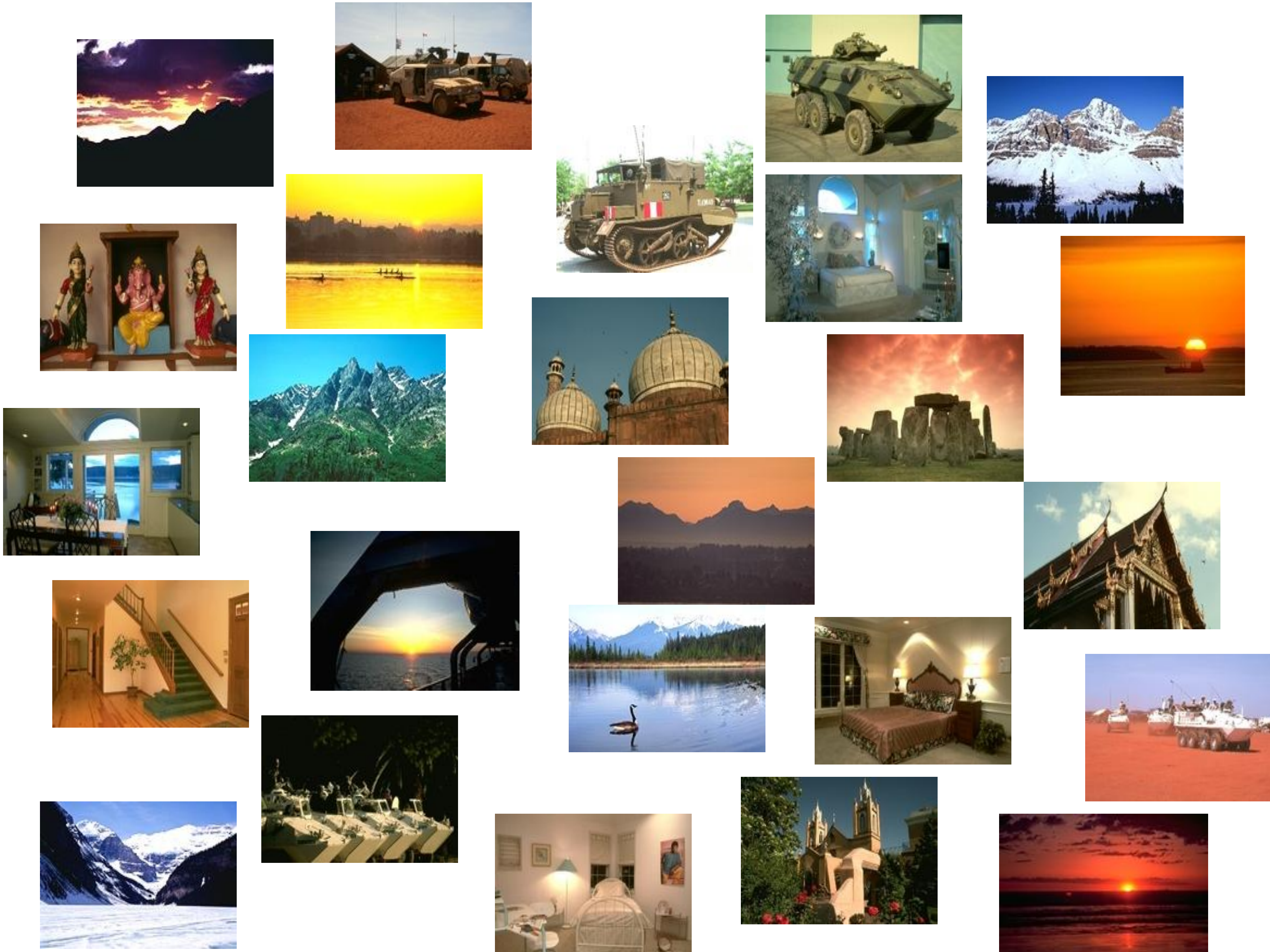
Intraclass variability

Scene Image Classification

Interclass similarity

Tall building	Inside city	Street	Highway	Coast	Open country	Mountain	Forest

Scene Image Clustering



Scene Image Clustering

Residential Interiors

Mountains

Military Vehicles

Sacred Places

Sunsets & Sunrises

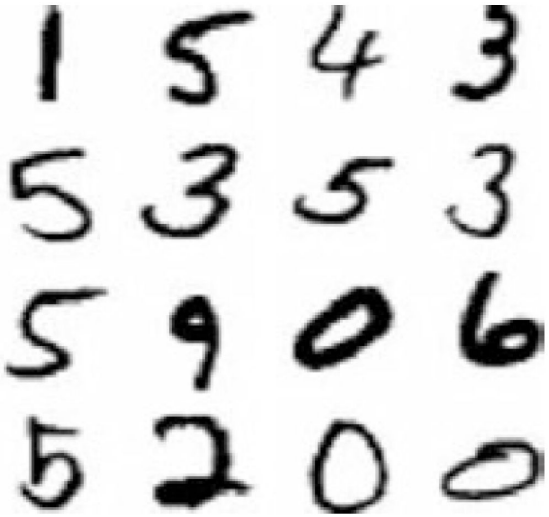
Machine Learning for Pattern Recognition

- Learning: Acquiring new knowledge or modifying the existing knowledge
- Knowledge: Familiarity with information present in data
- Learning by machines for pattern analysis: Acquisition of knowledge from data to discover patterns in data
- Data-driven techniques for learning by machines: Learning from examples (Training of models)
- Generalization ability of learning machines: Performance of trained models on new (test) data
- Target of learning techniques: Good generalization ability
- Learning techniques: Estimation of parameters of models
- Learning machines and Learning techniques for pattern analysis:
  - Statistical Models (Maximum likelihood)
  - Artificial Neural Networks (Error correction learning)
  - Kernel Methods (Learning optimal linear relationships)

Machine Learning Definition

- Arthur Samuel (1959)
  - Field of study that gives computers the ability to learn without being explicitly programmed

- Tom Mitchel (1998)
  - A computer program is said to learn from experience with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience (example) E



- T: Recognizing and classifying handwritten digits presented as image
- P: Percentage of digits correctly classified
- E: Database of handwritten digits images

Machine Learning Definition

- Arthur Samuel (1959)
  - Field of study that gives computers the ability to learn without being explicitly programmed

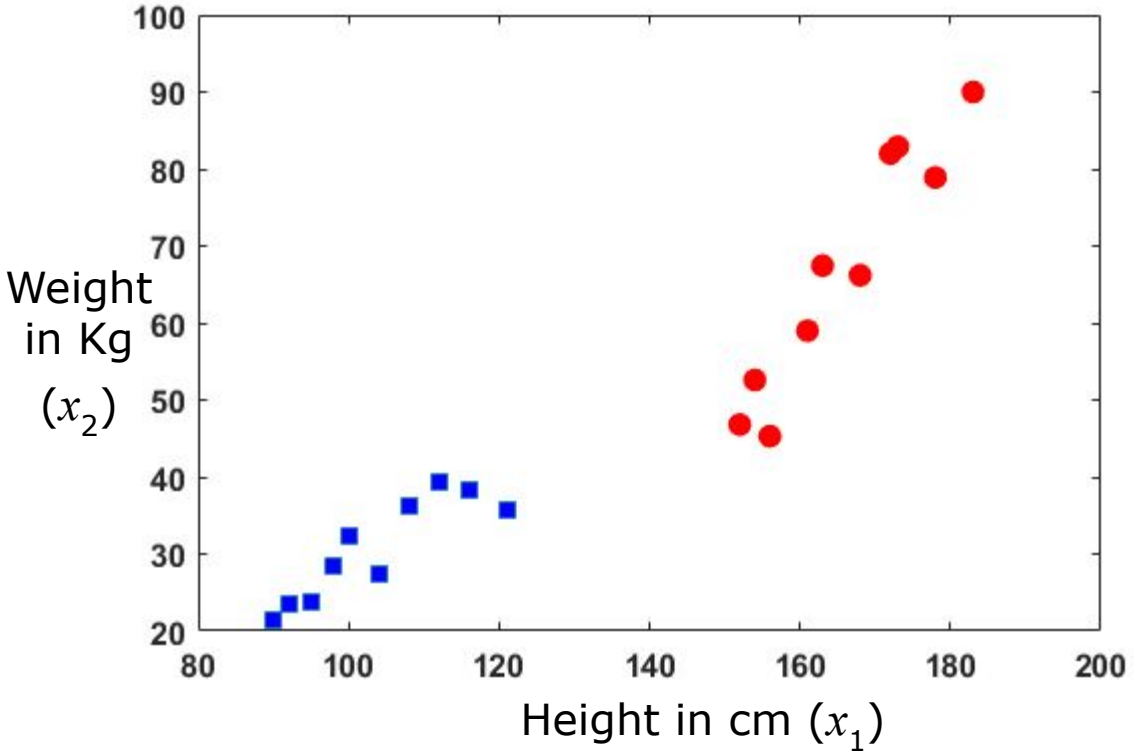
- Tom Mitchel(1998)
  - A computer program is said to learn from experience with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E

- Mapping from input to output

Illustration - Data1: Representing a Person

Height	Weight
90	21.5
95	23.67
100	32.45
116	38.21
98	28.43
108	36.32
104	27.38
112	39.28
121	35.8
92	23.56
152	46.8
178	78.9
163	67.45
173	82.9
154	52.6
168	66.2
183	90
172	82
156	45.3
161	59

- A person is represented using two attributes:
  - Height
  - Weight



$$\mathbf{x} = [x_1 \;\; x_2]^\top$$

Illustration – Data2: Iris (Flower) Data [1]

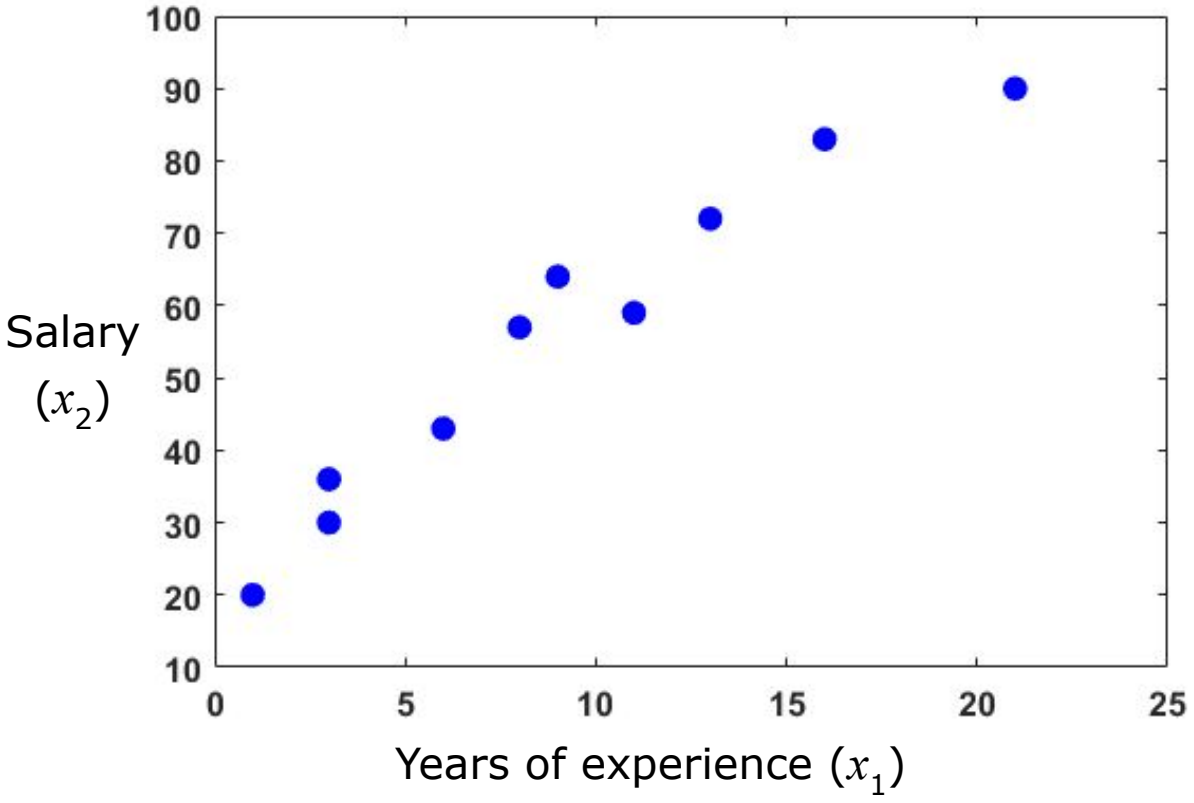
Sepal_Length	Sepal_Width	Petal_Length	Petal_Width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
7	3.2	4.7	1.4
6.4	3.2	4.5	1.5
6.9	3.1	4.9	1.5
6.3	3.3	6	2.5
5.8	2.7	5.1	1.9
7.1	3	5.9	2.1
5.7	2.8	4.1	1.3
7.3	2.9	6.3	1.8
7.3	2.9	6.3	1.8
5.3	3.7	1.5	0.2
4.9	2.4	3.3	1
5	3.5	1.6	0.6
6.3	3.3	4.7	1.6
5.8	2.7	3.9	1.2
5.8	2.8	5.1	2.4
4.4	3	1.3	0.2
6.2	3.4	5.4	2.3

$$\mathbf{x} = [x_1 \;\; x_2 \;\; x_3 \;\; x_4]^\top$$

[1] R. A. Fisher, "*The use of multiple measurements in taxonomic problems*" Annual Eugenics, 7, Part II, pp. 179-188, 1936.

Illustration – Data3: Years of Experience and Salary

Years of experience (x_1)	Salary (in Rs 1000) (x_2)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83



$$\mathbf{x} = [x_1 \;\; x_2]^\top$$

Illustration – Data4: Environmental Data

Temperature	Humidity	Pressure	Rain
25.47	82.19	1036.35	6.75
26.19	83.15	1037.60	1761.75
25.17	85.34	1037.89	652.50
24.30	87.69	1036.86	963.00
24.07	87.65	1027.83	254.25
21.21	95.95	1006.92	339.75
23.49	96.17	1006.57	38.25
21.79	98.59	1009.42	29.25
25.09	88.33	991.65	4.50
25.39	90.43	1009.66	112.50
23.89	94.54	1009.27	735.75
22.51	99.00	1009.80	607.50
22.90	98.00	1009.90	717.75
21.72	99.00	996.29	513.00
23.18	98.97	800.00	195.75
21.24	99.00	1009.21	474.75
21.63	99.00	1008.89	409.50
20.91	99.00	1008.89	1161.00
23.67	97.80	1009.38	0.00
24.53	92.90	1008.66	0.00

$$\mathbf{x} = [x_1 \;\; x_2 \;\; x_3 \;\; x_4]^\top$$

Supervised and Unsupervised Learning

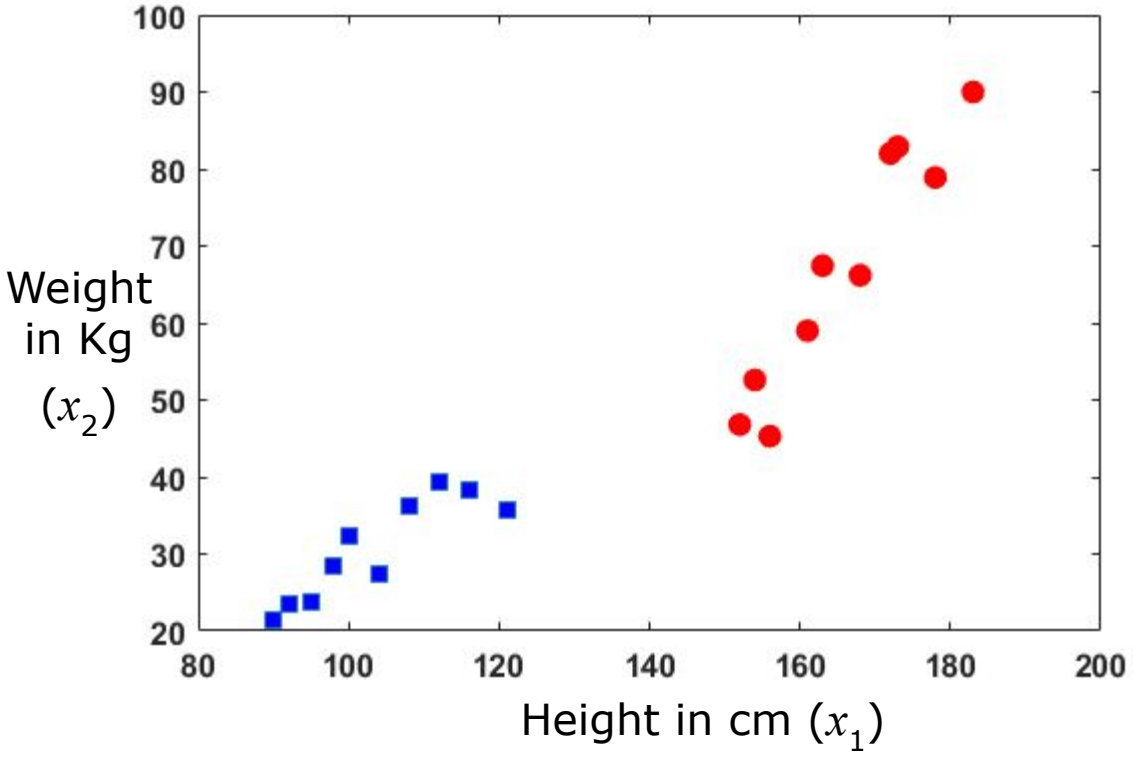
Supervised Learning

- Learning under the supervision
  - Student learning from teacher
  - Child learning to recognize objects/animals
- In the context of machine learning, data used for learning (Train data) is labeled
- Labeled data: Data for which the target value is already known

Labeled Data – Illustration: Data1 - Representing a Person

Height	Weight	Class
90	21.5	0
95	23.67	0
100	32.45	0
116	38.21	0
98	28.43	0
108	36.32	0
104	27.38	0
112	39.28	0
121	35.8	0
92	23.56	0
152	46.8	1
178	78.9	1
163	67.45	1
173	82.9	1
154	52.6	1
168	66.2	1
183	90	1
172	82	1
156	45.3	1
161	59	1

- A person is represented using two attributes:
  - Height
  - Weight
- Class (y):
  - Child (0)
  - Adult (1)



$\mathbf{x} = [x_1 \;\; x_2]^\top$ Target/Output : $y \in \{0,1\}$

Labeled Data – Illustration: Data2 - Iris (Flower) Data

Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Class
5.1	3.5	1.4	0.2	1
4.9	3	1.4	0.2	1
4.7	3.2	1.3	0.2	1
7	3.2	4.7	1.4	2
6.4	3.2	4.5	1.5	2
6.9	3.1	4.9	1.5	2
6.3	3.3	6	2.5	3
5.8	2.7	5.1	1.9	3
7.1	3	5.9	2.1	3
5.7	2.8	4.1	1.3	2
7.3	2.9	6.3	1.8	3
7.3	2.9	6.3	1.8	3
5.3	3.7	1.5	0.2	1
4.9	2.4	3.3	1	2
5	3.5	1.6	0.6	1
6.3	3.3	4.7	1.6	2
5.8	2.7	3.9	1.2	2
5.8	2.8	5.1	2.4	3
4.4	3	1.3	0.2	1
6.2	3.4	5.4	2.3	3

- Class (y):
  - Iris Setosa (1)
  - Iris Versicolour (2)
  - Iris Virginica (3)

$$\mathbf{x} = [x_1 \;\; x_2 \;\; x_3 \;\; x_4]^\top$$

Target/Output:

$$y \in \{1, 2, 3\}$$

Labeled Data – Illustration: Data3 - Years of Experience and Salary

Years of experience (x_1)	Salary (in Rs 1000) (x_2)	Raise (y)
3	30	1
8	57	0
9	64	1
13	72	1
3	36	1
6	43	0
11	59	1
21	90	1
1	20	0
16	83	0

- Class – Raise in Salary (y):
  - Yes(1)
  - No (0)

$\mathbf{x} = [x_1 \;\; x_2]^T$ Target/Output : $y \in \{0,1\}$

Labeled Data – Illustration: Data3 - Years of Experience and Salary

Years of experience (x)	Salary (in Rs 1000) (y)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

- Input variable: Years of experience
- Output variable: Salary

Illustration – Data4: Environmental Data

Temperature	Humidity	Pressure	Rain
25.47	82.19	1036.35	6.75
26.19	83.15	1037.60	1761.75
25.17	85.34	1037.89	652.50
24.30	87.69	1036.86	963.00
24.07	87.65	1027.83	254.25
21.21	95.95	1006.92	339.75
23.49	96.17	1006.57	38.25
21.79	98.59	1009.42	29.25
25.09	88.33	991.65	4.50
25.39	90.43	1009.66	112.50
23.89	94.54	1009.27	735.75
22.51	99.00	1009.80	607.50
22.90	98.00	1009.90	717.75
21.72	99.00	996.29	513.00
23.18	98.97	800.00	195.75
21.24	99.00	1009.21	474.75
21.63	99.00	1008.89	409.50
20.91	99.00	1008.89	1161.00
23.67	97.80	1009.38	0.00
24.53	92.90	1008.66	0.00

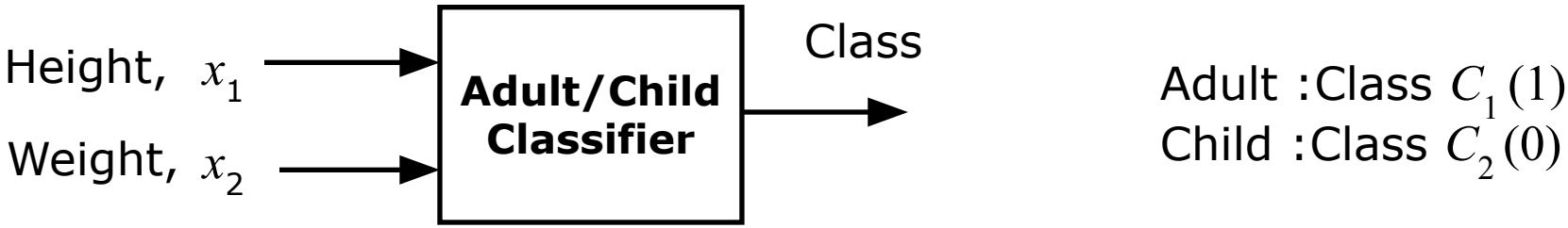
- Predicting `Rain` (target attribute) based on `Temperature`, `Humidity` and `Pressure`
- Input variable: `Temperature`, `Humidity` and `Pressure`
- Output variable: `Rain`

Supervised Learning

- In supervised learning, each example (data sample) is a pair consisting of an input example (typically a vector) and a desired output value (also called the target)
- Task of learning a function that maps an input to an output based on example input-output pairs

$$y_n = f(\mathbf{x}_n)$$

- A supervised learning algorithm
  - analyzes the training data and
  - produces an inferred function, which can be used for predicting the output of a new examples
- One of the scenario will be the algorithm to determine the class labels for unseen instances



Supervised Learning

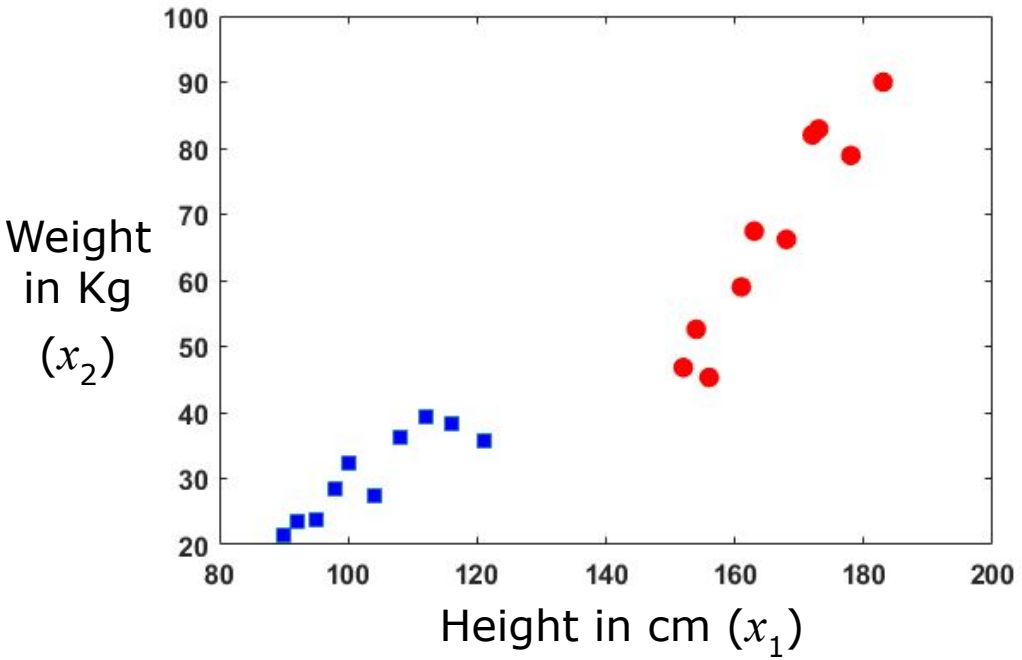
- Supervised learning is grouped into
  - Classification
  - Regression
- Classification:
  - Output variable is categorical
  - Categorical label prediction
  - Example:
    - Predicting a person as adult or child (2-class)
    - Predicting the raise in salary based on the year of experience and salary (2-class)
    - Identify an email as spam or not (2-class)
    - Predicting the presence or absence of disease (2-class)
    - Categorising the disease according to symptoms (Multi-class)
    - Categorizing the Iris flowers (Multi-class)

Supervised Learning

- Supervised learning is grouped into
  - Classification
  - Regression
- Regression:
  - Output variable is real or continuous value
  - Numeric prediction
  - Example:
    - predicting the salary based on the experience
    - predicting the amount of rainfall based on atmospheric temperature, humidity, pressure, amount of sunlight etc.

Unsupervised Learning

- Learning without a supervision
- In the context of machine learning, data used for learning (Train data) is unlabeled
- Given these unlabeled data machine tries to identify the pattern and give the response
- Example:
  - A person is represented using two attributes: Height and Weight
  - No label is given
  - Machine try to learn the patterns from the given set and groups them based on the similarity



Unsupervised Learning

- Unsupervised learning is grouped into
  - Clustering
  - Association
- Clustering:
  - Partitioning the data into cohesive groups such that the data samples in a group are similar
  - Example:
    - Grouping the persons based on their height and weight
    - Given the customer and their purchase data:
      - Grouping the customers based on the similar products purchased
- Association:
  - It is a rule-based machine learning to discover the interesting variables in a data set
  - Example:
    - Given the customer and their purchase data:
      - Finding the products purchased together

Summary

- Machine learning: Learning from data
- Supervised machine learning
  - Data used for learning (Train data) is labeled
  - Each example (data sample) is a pair consisting of an input example (typically a vector) and a desired output value (also called the target)
  - Task of learning a function that maps an input to an output based on example input-output pairs
  - Classification and Regression
- Unsupervised machine learning
  - Data used for learning (Train data) is unlabeled
  - Given these unlabeled data machine tries to identify the pattern based on similarity
  - Clustering and Association

Text Books

1. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Third Edition, Morgan Kaufmann Publishers, 2011.

2. S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 2009.

3. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.