SIT720 - Machine Learning

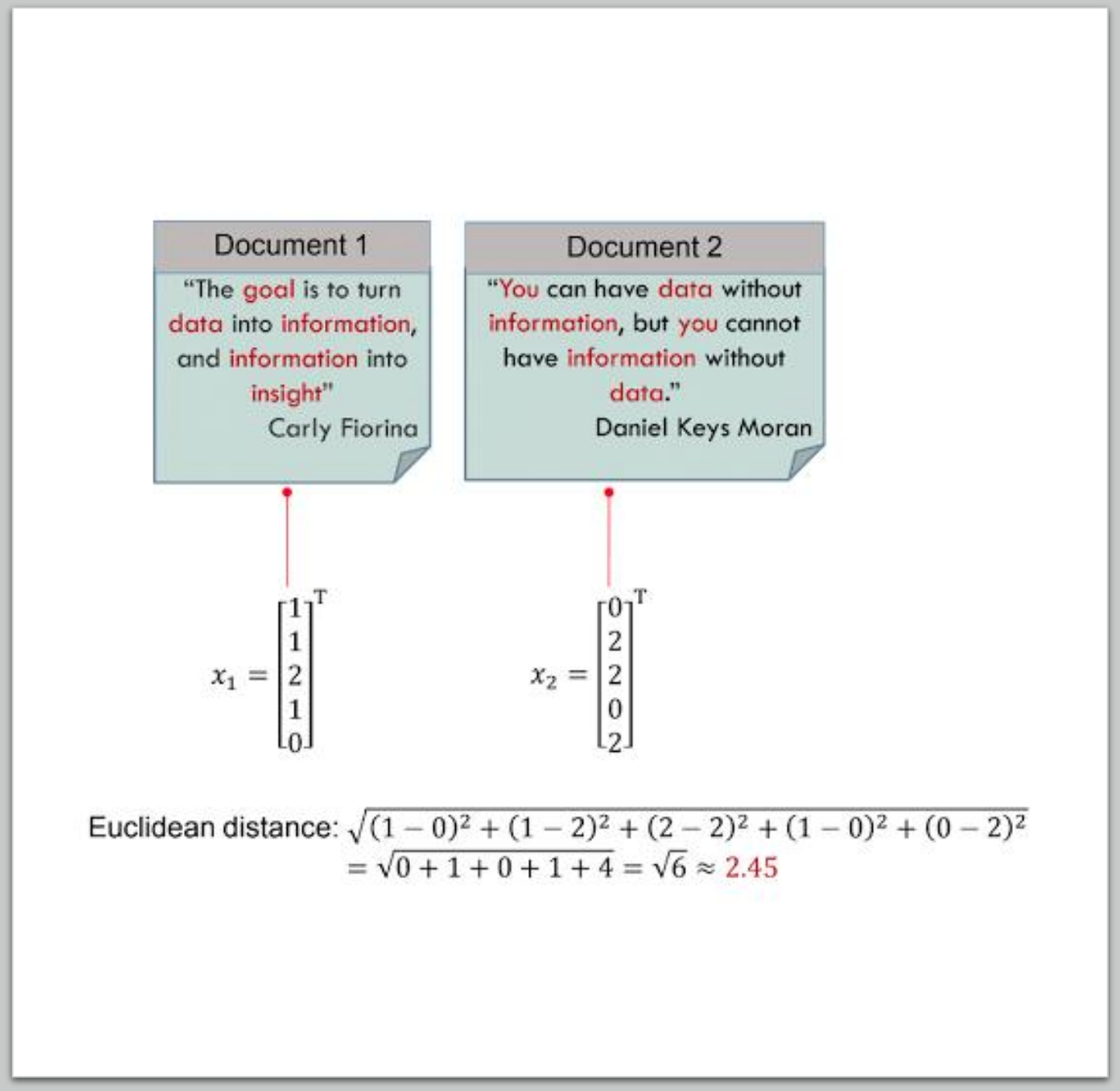
Week 2 – Foundation of ML

Revising knowledge of Linear Algebra and Probability

- Feature vectors and matrices
- Probability Concepts
  - Random experiment & Event
  - Joint probability
  - Conditional probability
  - Bayes Rules
- Random variable
  - Distribution of random variables
- Data wrangling
  - Missing value replacement
  - Scale or normalisation
  - Non-numeric data encoding

Feature Vectors

- Vector space model is representation of set of documents as vectors.
- It is a fundamental step in information retrieval operations
- Text data representation as **Feature Vectors**

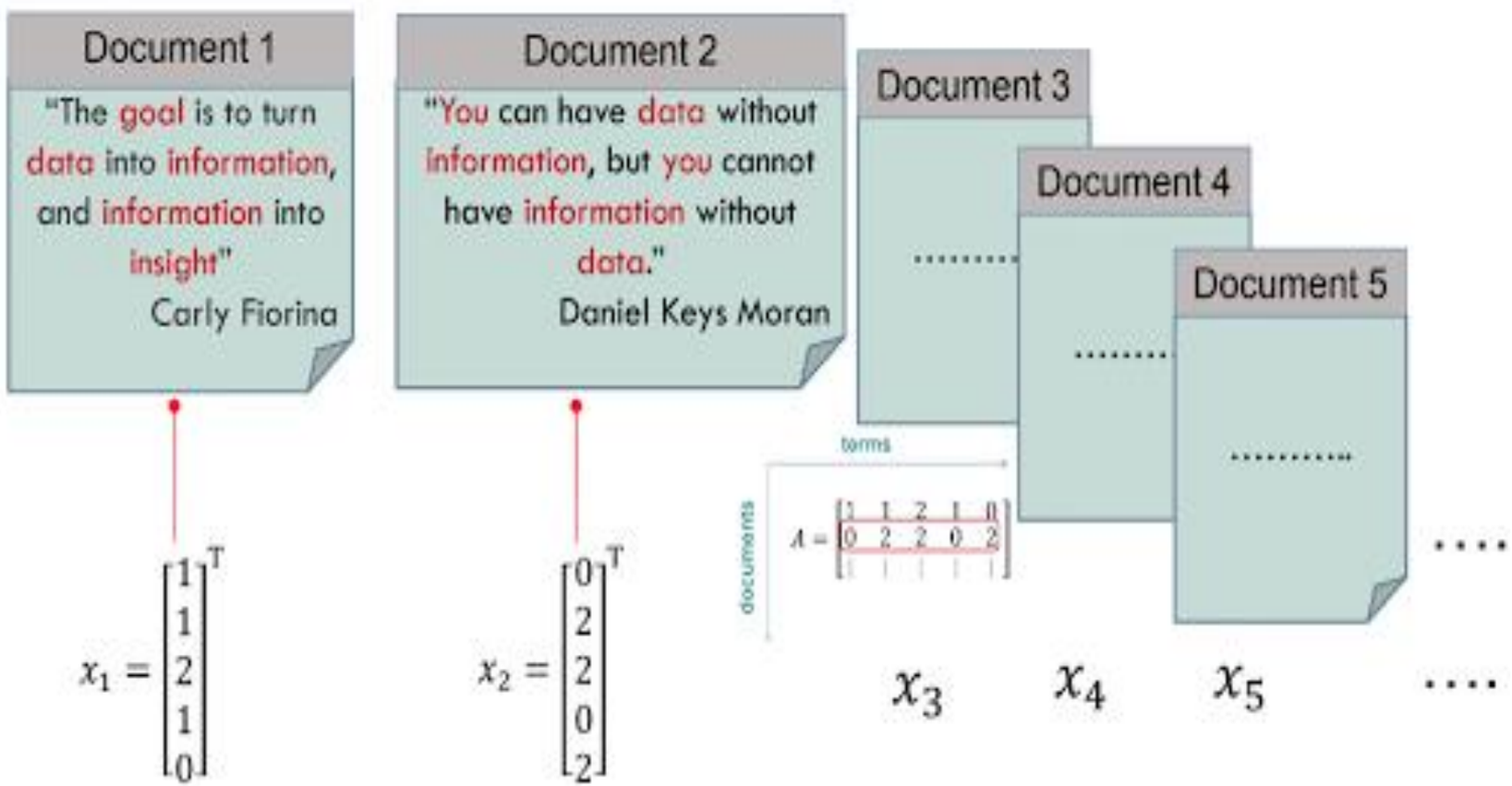


Feature matrix

- We can extend the concept of the feature vector towards a feature matrix by stacking feature vectors as a matrix X
  - We create a vocabulary of features for all the instances in the dataset
  - Represent each instance as a vector on features listed in the vocabulary
  - If our dataset has N instances, we create N vectors x1,x2,,...,xN
  - Each of these vectors is called a feature vector
  - We stack these vectors as a matrix X and call it a feature matrix

Feature matrix

- Example of steps mentioned earlier:



Revising knowledge of Linear Algebra and Probability

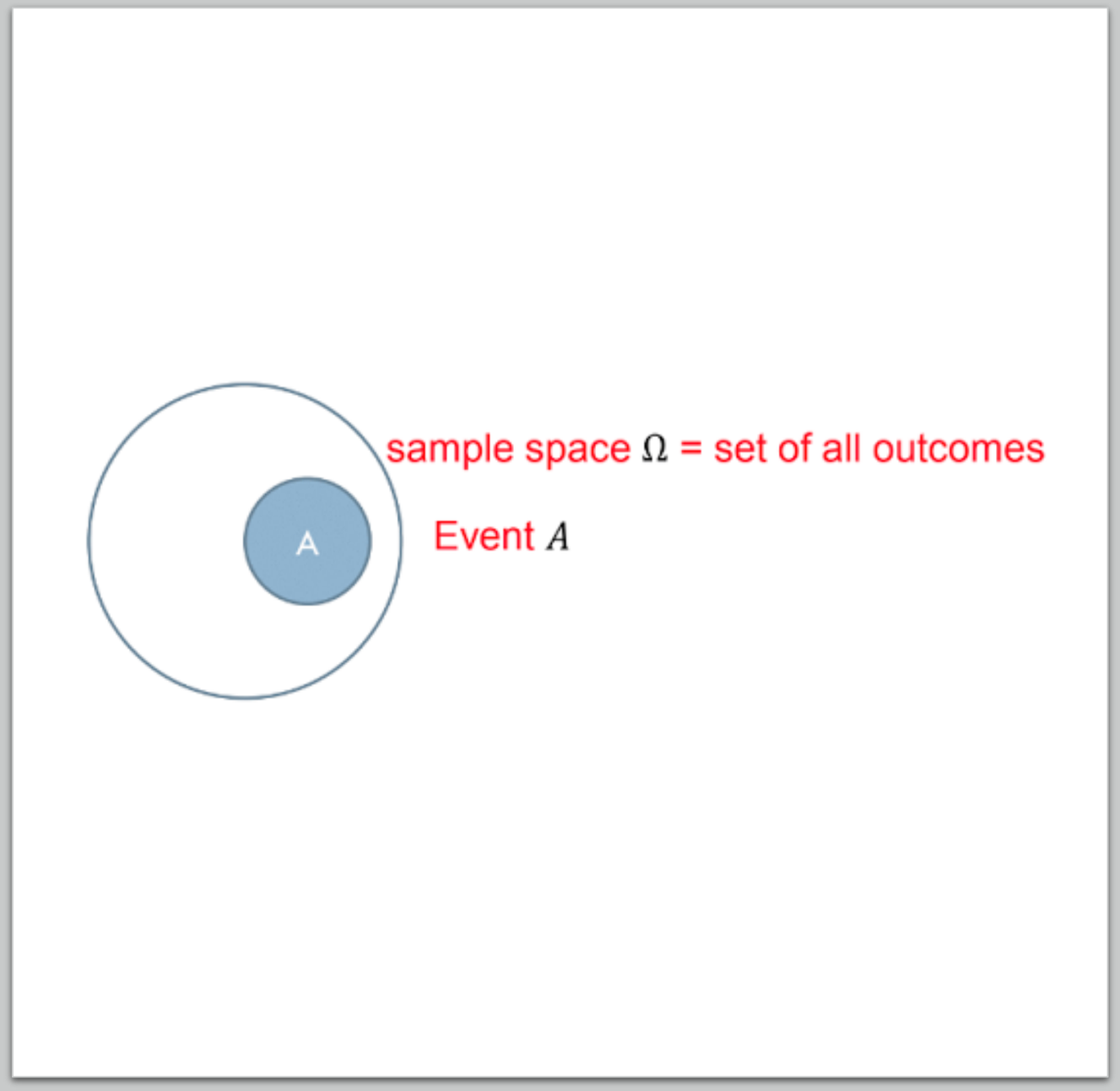
- Feature vectors and matrices
- Probability Concepts
  - Random experiment & Event
  - Joint probability
  - Conditional probability
  - Bayes Rules
- Random variable
  - Distribution of random variables
- Data wrangling
  - Missing value replacement
  - Scale or normalisation
  - Non-numeric data encoding

Random experiment

- An experiment or a process for which the outcome cannot be predicted with certainty.
  - toss of a coin
  - roll of a dice
  - counting the number of phone calls received on a mobile phone in a given duration
  - daily temperature
  - how many times a specific word appears in each document of a corpus

Event & Sample Space

- **Event**: a set of outcomes of a random experiment
  - For a coin toss experiment, sample space Ω={head,tail}. And event A could be either {head} or {tail}
  - For a dice roll experiment, sample space Ω={1,2,3,4,5,6} and event A={{1},{2},{3},{4},{5},{6}}



Probability

- **Probability** is defined for an event and is the **measure of the likelihood that an event** will occur. It is quantified as a number between 0 and 1
- The probability of an event A occurring is denoted as P(A)
- The probability of an event A not occurring is denoted as

$$P(\bar{A}) = 1 - P(A)$$

Joint Probability

- Probability can be defined **jointly for more than one event**. Consider a random experiment where we **toss two coins**.

- In this case the probability of seeing “head for coin-1” and “head for coin-2” is an example of two events. If two events, **A and B are independent** then the joint probability is

$$P(A \text{ and } B) = P(A)P(B)$$

- Assuming fair coins with probability of head as 1/2. The probability of head for the first coin and head for the second coin is:

$$P(\{head - first\} \text{and} \{head - second\}) = \frac{1}{2} * \frac{1}{2}$$

Conditional Probability

- What is the probability of some event A, given the occurrence of another event B.
- Condition probability P(A|B), read as the probability of A given B is defined as

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

- Provided P(B) is not zero

Byes Rule

- Essence of most of Bayesian approaches
- Mathematical rule explaining how you should change your existing beliefs in the light of new occurrence

- Bayes rule describes the probability of an event A based on another event B that is related to A

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ in which, } P(B) \neq 0$$

- If cancer is related to age, using Bayes' rule information about a person's age can be used to more accurately assess the probability that the person has cancer.

Revising knowledge of Linear Algebra and Probability

- Linear Algebra
  - Vector & their operations
  - Matrix & their operations
  - Feature vectors and matrices
- Probability Concepts
  - Random experiment & Event
  - Joint probability
  - Conditional probability
  - Bayes Rules
- Random variable
  - Distribution of random variables
- Data wrangling
  - Missing value replacement
  - Scale or normalisation
  - Non-numeric data encoding

Random Variable

- Is a variable whose possible values are the generated outcomes of a random phenomenon
- Is a function that can assign probabilities to events of interest in a random experiment

- if we toss a coin the possible outcomes are head or tail. Let us define a random variable X so that X=1 means head and X=0 means tail.
  - *The function is nothing but the mapping X=1 to head or X=0 to tail*.
  - Let's say P(head)=0.6, P(tail)=0.4. Then we can say P(X=1)=0.6, P(X=0)=0.4

Types of Random Variable

- Discrete Random Variable
  - countable number of values (i.e., faces of a dice, number of emails received in an hour)

- Continuous Random Variable
  - can take values on a infinite continuum (i.e., height of a person, time to failure)

Discrete Random Variable

- Defined using a **Probability Mass Functions (PMF)**, denoted as π(x)

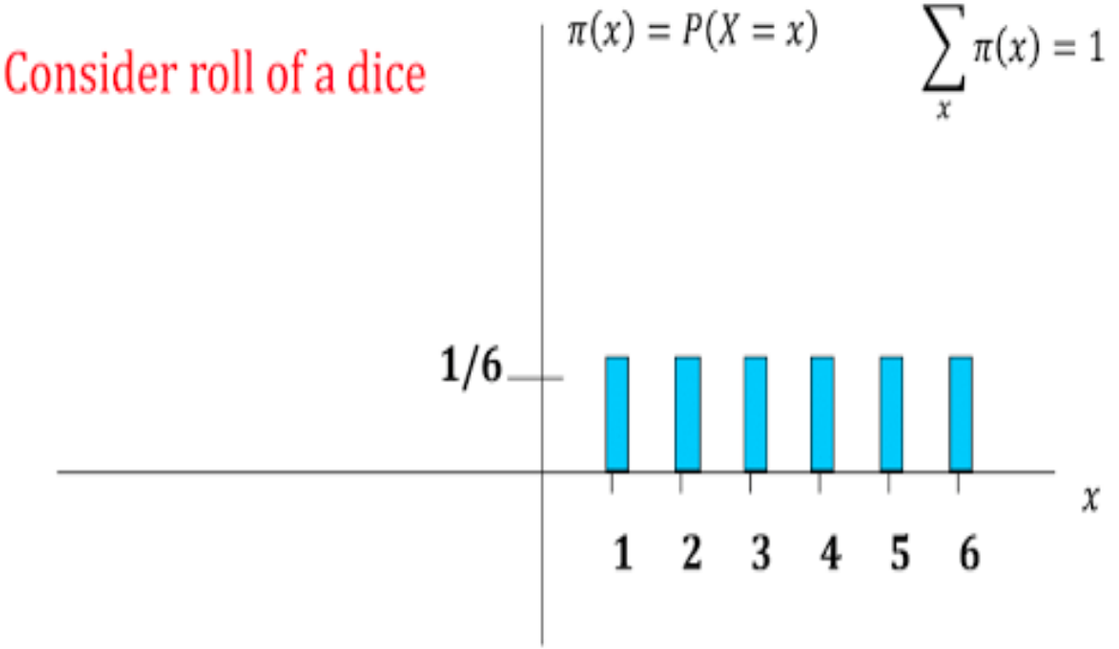
- The PMF assigns a probability to each possible value of the random variable as π(x)=P(X=x) summing them to 1, i.e.

$$\sum_{x} \Pi\,(x) = 1$$

- Rolling a dice is a perfect example of random variables. But what if someone asks about the probability of rolling a dice and getting a number **less than 5**?

Discrete Random Variable...

- Cumulative Distribution Function (CDF).
- The cumulative distribution function gives us the cumulative probability associated with a function. it is defined as:



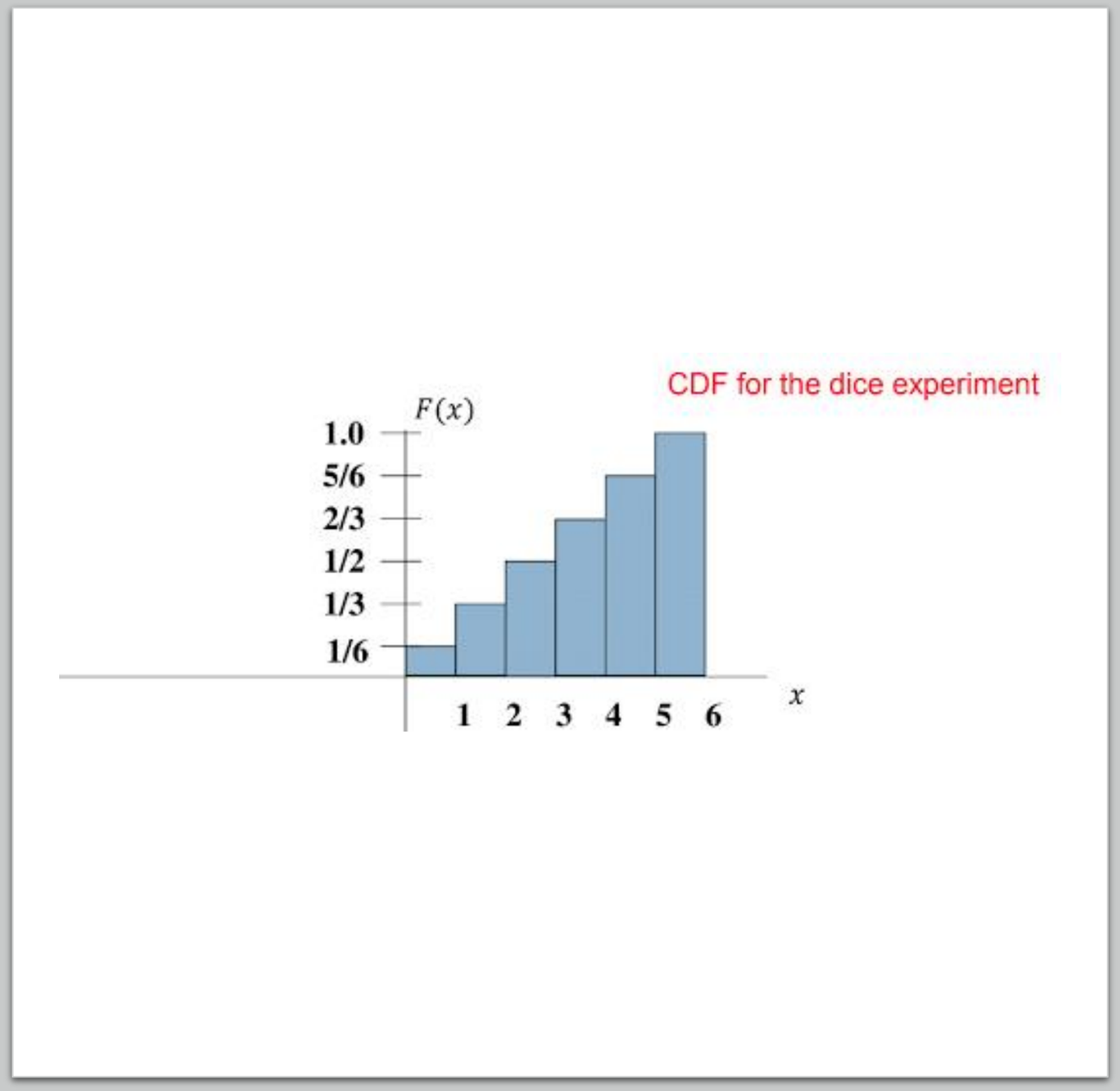
Discrete Random Variable...

- In the figure, it is discontinuous at points x_i's.
- The probability of seeing a number equal or less than five is

$$P(X \leq 5) = \frac{5}{6}$$

- probability of seeing a number greater than five is

$$P(X > 5) = 1 - \frac{5}{6} = \frac{1}{6}$$



Continuous Random Variable

- Continuous random variables are defined using Probability Density Functions (PDF, *statistical expression*) , denoted as f(x).

- PDF assigns a probability to a range of values of the random variable as

$$f(x)d(x) = P(x \leq X \leq x + dx) \quad \text{integrating to 1}$$

- So, we can say:

$$\int_{-\infty}^{+\infty} f\,(x)dx = 1$$

- Probability assigned at any exact value is zero (in the continues space). But we can talk about probabilities over a range such as

$$P(X \geq a), P(X < b) \text{ or } P(c \leq X \leq d)$$

Distribution of Random Variable

- Probability distribution is a function that links each outcome of a statistical experiment with its probability of occurrence
- **Bernoulli distribution**
  - discrete distribution and defined for a binary random variable with values X=0 and X=1
  - so $\pi(0) = P(X = 0) = p$ and $\pi(1) = P(X = 1) = 1 - p$
  - B means Bernoulli in notation $\pi(x) = B(x||p)$ or $x \sim B(x||p)$
  - For example
    - we can say a distribution over the outcome of an exam is Bernoulli. We may pass (x=1) or fail (x=0)

Distribution of Random Variable...

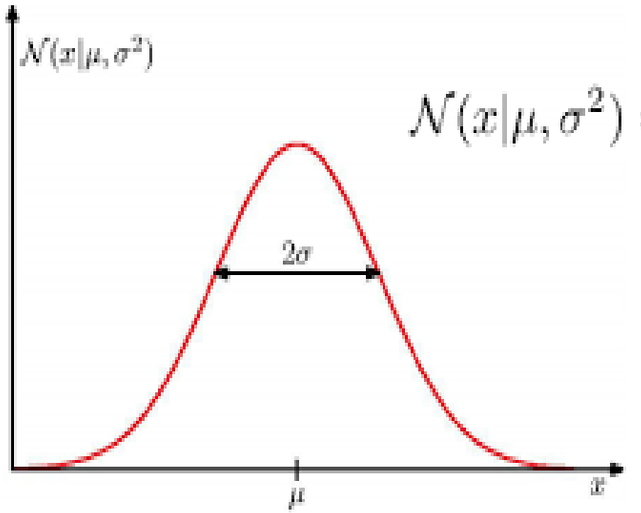
- **Uniform distribution**
  - can be defined for both discrete and continuous random variables. For a discrete random variable
  - For discrete $\pi(x_i) = P(X = x_i) = \frac{1}{N}, i = 1..N$
  - U means uniform in notation $\pi(x) = U(x||N) \vee x \sim U(x||N)$
  - For a continuous random variable $f(x) = \frac{1}{b-a}, a \leq x \leq b$
  - U means uniform in notation $f(x) = U(x||a,b) \vee x \sim U(x||a,b)$
  - Rolling a fair dice follows a uniform distribution (discrete space)

Distribution of Random Variable...

- **Normal distribution**
  - is defined for continuous random variables
  - most popular distribution

$$N(x||\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

  - where N means normal
  - popular because natural phenomena are approximately following a normal distribution



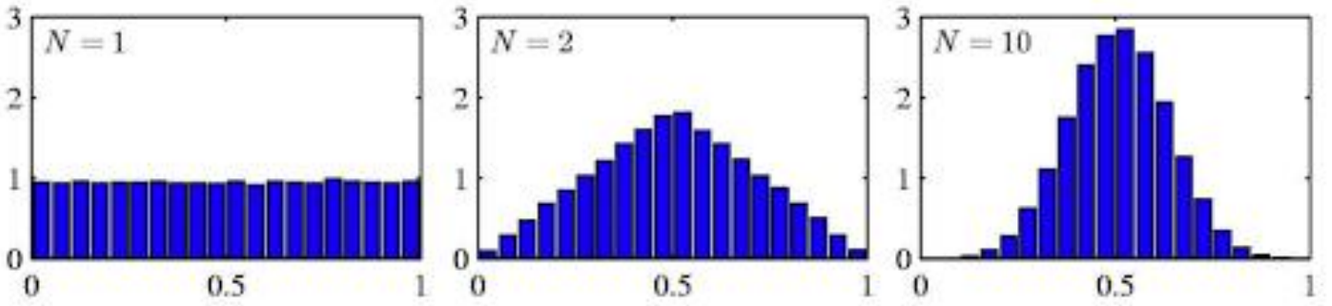
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Distribution of Random Variable...

- **Central limit theorem**
  - if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the existing population, then the distribution of the sample means will be approximately normally distributed
  - So, the distribution of the sum of *N* i.i.d.(independent and identically distributed) random variables becomes increasingly normal (Gaussian) as *N* grows

$$Y = X_1 + X_2 + \ldots + X_N$$

  - N uniform [0,1] random variables, following central limit theorem



Revising knowledge of Linear Algebra and Probability

- Linear Algebra
  - Vector & their operations
  - Matrix & their operations
  - Feature vectors and matrices
- Probability Concepts
  - Random experiment & Event
  - Joint probability
  - Conditional probability
  - Bayes Rules
- Random variable
  - Distribution of random variables
- Data wrangling
  - Missing value replacement
  - Scale or normalisation
  - Non-numeric data encoding

Data wrangling...

- **Data wrangling**, also known as data munging, is the process of **cleaning, transforming, and organizing a dataset** to make it suitable for analysis.

- This often involves a combination of manual and automated processes, and it is a crucial step in the data science pipeline.

- **Missing value replacement** is one of the essential parts of data wrangling.
  - The machine learning model cannot process null values, and null values misguide the machine learning model.
  - Null or missing value can be replace by the following methods:
    - Replace with immediate value
    - Replace with mean or median value of the row

Immediate, Mean and Median value...

Mean value

$$mean(TV) = \frac{\sum_{i=1}^{n} TV_i}{n}$$

	TV	Radio	Newspaper	Sales	Types
25	262.9	3.5	19.5	12.0	newspaper
26	142.9	29.3	12.6	NaN	electronics
27	240.1	16.7	22.9	15.9	electronics
28	248.8	27.1	22.9	18.9	newspaper
29	70.6	16.0	40.8	10.5	newspaper
30	292.9	28.3	43.2	21.4	newspaper
31	112.9	17.4	38.6	11.9	electronics
32	97.2	1.5	30.0	9.6	electronics
33	265.6	20.0	0.3	17.4	newspaper
34	95.7	1.4	7.4	9.5	newspaper
35	290.7	4.1	8.5	12.8	electronics
36	266.9	43.8	5.0	25.4	electronics
37	74.7	49.4	45.7	14.7	newspaper
38	NaN	26.7	35.1	10.1	newspaper
39	228.0	37.7	32.0	21.5	electronics

Immediate value

$median(TV) = TV_{n/2}$, Where n is number of samples and even

Median value

Example - replace missing value with immediate value...

Given datasets contain with N/A value as follows:

	TV	Radio	Newspaper	Sales	Types
25	262.9	3.5	19.5	12.0	newspaper
26	142.9	29.3	12.6	NaN	electronics
27	240.1	16.7	22.9	15.9	electronics
28	248.8	27.1	22.9	18.9	newspaper
29	70.6	16.0	40.8	10.5	newspaper
30	292.9	28.3	43.2	21.4	newspaper
31	112.9	17.4	38.6	11.9	electronics
32	97.2	1.5	30.0	9.6	electronics
33	265.6	20.0	0.3	17.4	newspaper
34	95.7	1.4	7.4	9.5	newspaper
35	290.7	4.1	8.5	12.8	electronics
36	266.9	43.8	5.0	25.4	electronics
37	74.7	49.4	45.7	14.7	newspaper
38	NaN	26.7	35.1	10.1	newspaper
39	228.0	37.7	32.0	21.5	electronics

After replacing NaN value with immediate value:

	TV	Radio	Newspaper	Sales	Types
25	262.9	3.5	19.5	12.0	newspaper
26	142.9	29.3	12.6	15.9	electronics
27	240.1	16.7	22.9	15.9	electronics
28	248.8	27.1	22.9	18.9	newspaper
29	70.6	16.0	40.8	10.5	newspaper
30	292.9	28.3	43.2	21.4	newspaper
31	112.9	17.4	38.6	11.9	electronics
32	97.2	1.5	30.0	9.6	electronics
33	265.6	20.0	0.3	17.4	newspaper
34	95.7	1.4	7.4	9.5	newspaper
35	290.7	4.1	8.5	12.8	electronics
36	266.9	43.8	5.0	25.4	electronics
37	74.7	49.4	45.7	14.7	newspaper
38	228.0	26.7	35.1	10.1	newspaper
39	228.0	37.7	32.0	21.5	electronics

Revising knowledge of Linear Algebra and Probability

- Linear Algebra
  - Vector & their operations
  - Matrix & their operations
  - Feature vectors and matrices
- Probability Concepts
  - Random experiment & Event
  - Joint probability
  - Conditional probability
  - Bayes Rules
- Random variable
  - Distribution of random variables
- Data wrangling
  - Missing value replacement
  - Scale or normalisation
  - Non-numeric data encoding

Scaling or normalisation...

- **Normalization** is a scaling technique used to transform the values of a dataset into a common range.
- This is often done to improve the performance of machine learning algorithms, as many algorithms operate better when the data is in a standardized range.
- Min-max normalization:
  - Scales data to a range of 0 to 1, where 0 is the minimum value in the dataset and 1 is the maximum value.
  - Formula: $v' = \frac{v - \min(v)}{\max(v) - \min(v)}$, where $\max(v)$ is maximum and $\min(v)$ is minimum value for vector v.

student	height	weight
1	6.149041	181.9149
2	4.337117	193.9508
3	5.686612	158.2599
4	4.659973	182.3271
5	5.5731	182.5691
6	6.336751	199.5378
7	5.881856	169.0076
8	5.732311	185.0168
9	6.71516	175.5523
10	4.930869	167.0906

student	height	weight
1	0.449207	0.660173
2	0.201498	0
3	0.780809	0.551649
4	0.930324	0.429412
5	0.376887	1
6	0.3889	0.475258
7	1	0.700081
8	0	0.835615
9	0.11599	0.795638
10	0.539802	0.34183

Revising knowledge of Linear Algebra and Probability

- Linear Algebra
  - Vector & their operations
  - Matrix & their operations
  - Feature vectors and matrices
- Probability Concepts
  - Random experiment & Event
  - Joint probability
  - Conditional probability
  - Bayes Rules
- Random variable
  - Distribution of random variables
- Data wrangling
  - Missing value replacement
  - Scale or normalisation
  - Non-numeric data encoding

Non-numerical data encoding...

- Unlike features with quantitative value, some features contain categorical values which the machine cannot understand.
  - To solve this, encoding techniques are used to convert to integer values.
- There are several well-known techniques of encoding such as:
  - OrdinalEncoder
  - One-Hot Encodings
  - LabelEncoder.

Non-numerical values in “Types” column

	TV	Radio	Newspaper	Sales	Types
0	230.1	37.8	69.2	22.1	newspaper
1	44.5	39.3	45.1	10.4	newspaper
2	17.2	45.9	69.3	9.3	electronics
3	151.5	41.3	58.5	18.5	electronics
4	180.8	10.8	58.4	12.9	newspaper

After “label encoding” newspaper==1 and electronics==0

	TV	Radio	Newspaper	Sales	Types	interger label
0	230.1	37.8	69.2	22.1	newspaper	1
1	44.5	39.3	45.1	10.4	newspaper	1
2	17.2	45.9	69.3	9.3	electronics	0
3	151.5	41.3	58.5	18.5	electronics	0
4	180.8	10.8	58.4	12.9	newspaper	1

Thank You