IMMUNE CODING INSTITUTE

Probability and Statistic for ML

Module 1.2

carlos.chavez@immune.institute

Outline

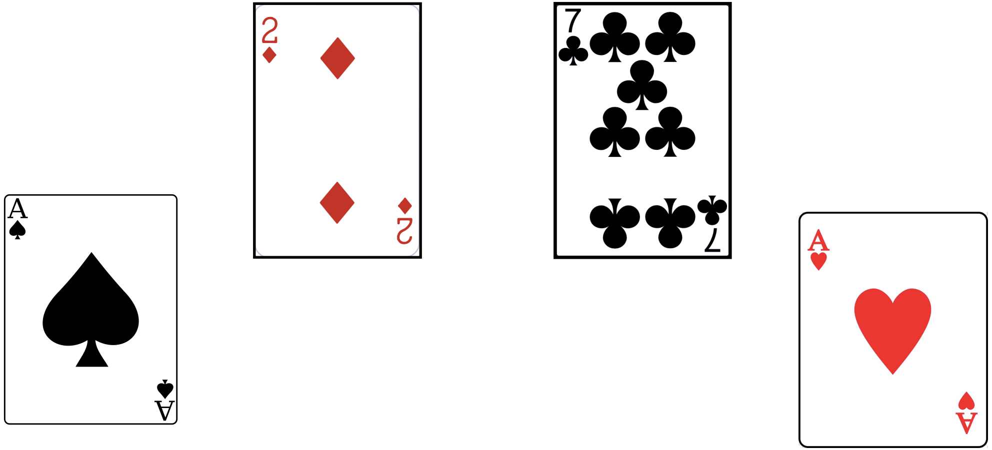
- Introduction to Probability and Combinatorics
    - Definition and Notation
    - Do we know how to count?
- Conditional probability
    - Bayes' Rule
    - Independence
- Random Variables
    - Cumulative distribution functions
- Main distributions
- Central Limit Theorem

Let’s start from the beginning… Do we know how to count?

Example: Pick up a card, any card..!

Pick a card from a standard deck of 52 cards and lets count the following <u>events</u>:

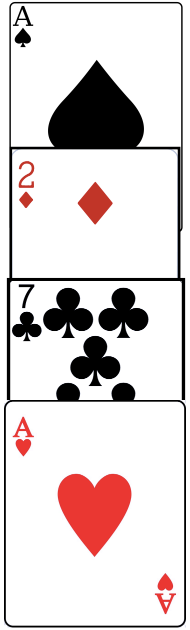
- **A**: card is an ace
- **B**: card is a black suit
- **D**: card is a diamond
- **H**: card is heart



Example: Pick up a card, any card..!

- **A**: card is an ace
- **B**: card is a black suit
- **D**: card is a diamond
- **H**: card is heart

Events:



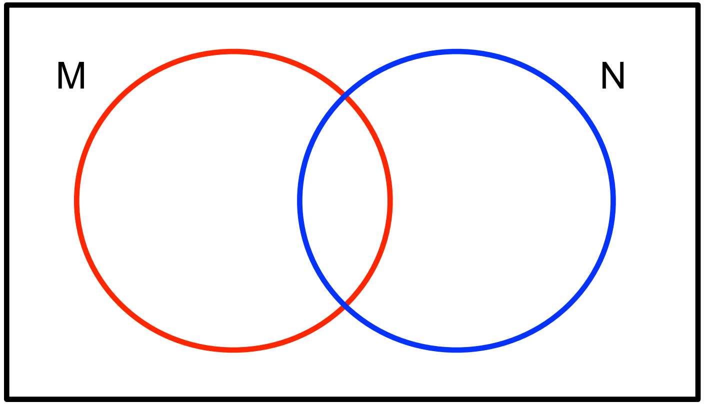
1) $A \cap H$

2) $A \cap B$

3) $A \cup D \cup H$

4) $A^c \cap B^c$

5) $(A \cup B)^c$



IMMUNE
CODING INSTITUTE

Probability

Definition (Naive): If all outcomes are equally likely, the probability of an event A is happening is:

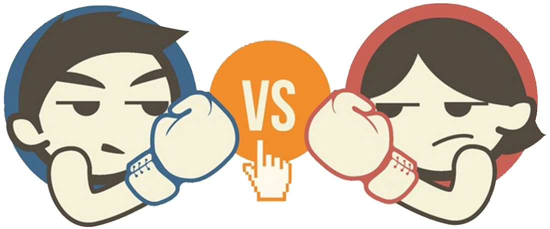
$$\boxed{P_{naive}(A) = \frac{\text{number of outcomes favorable to A}}{\text{number of outcomes}}}$$

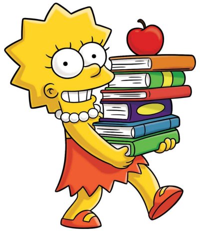
Properties:

1) $\boxed{0 \leq P(A) \leq 1}$

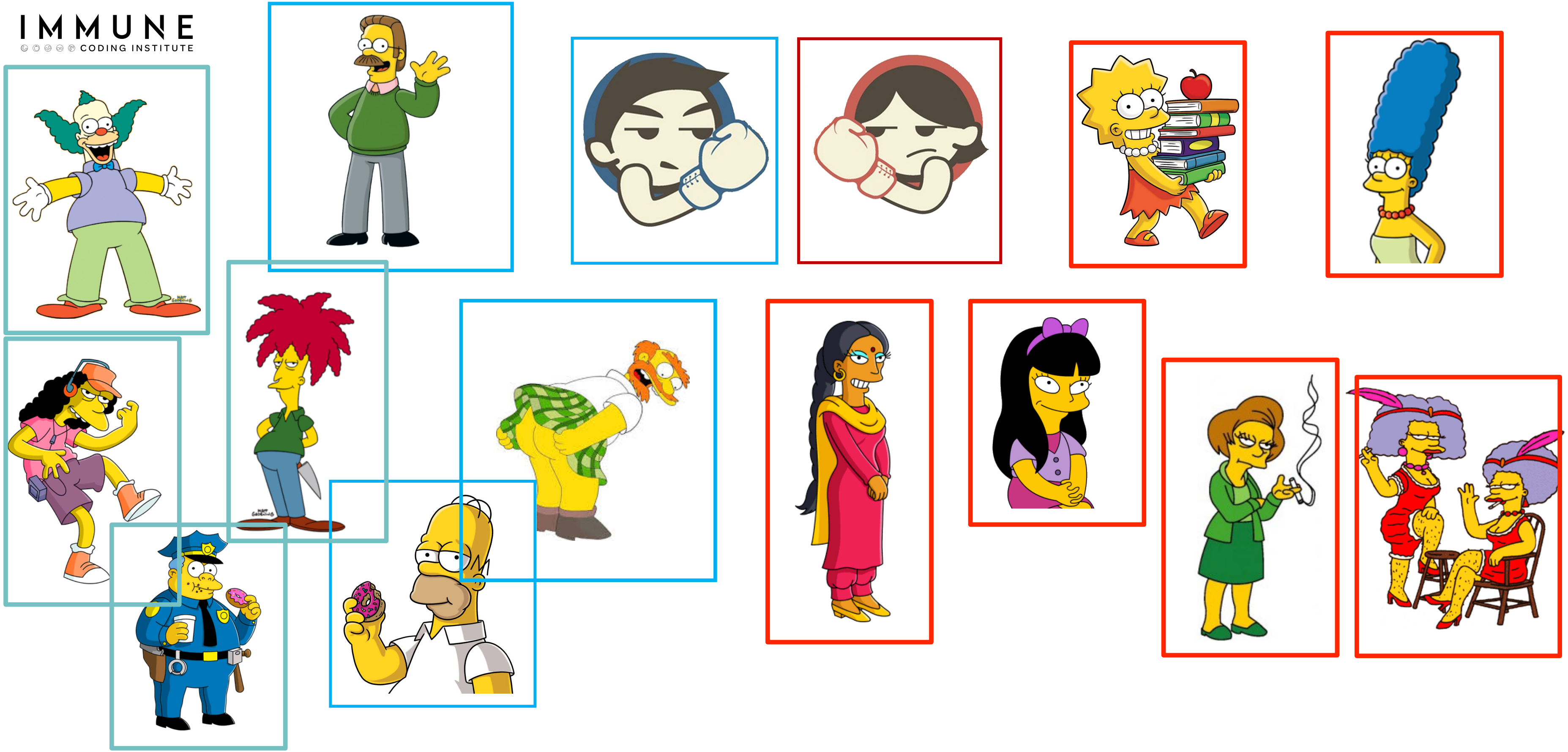
2) $\boxed{P(S) = 1}$

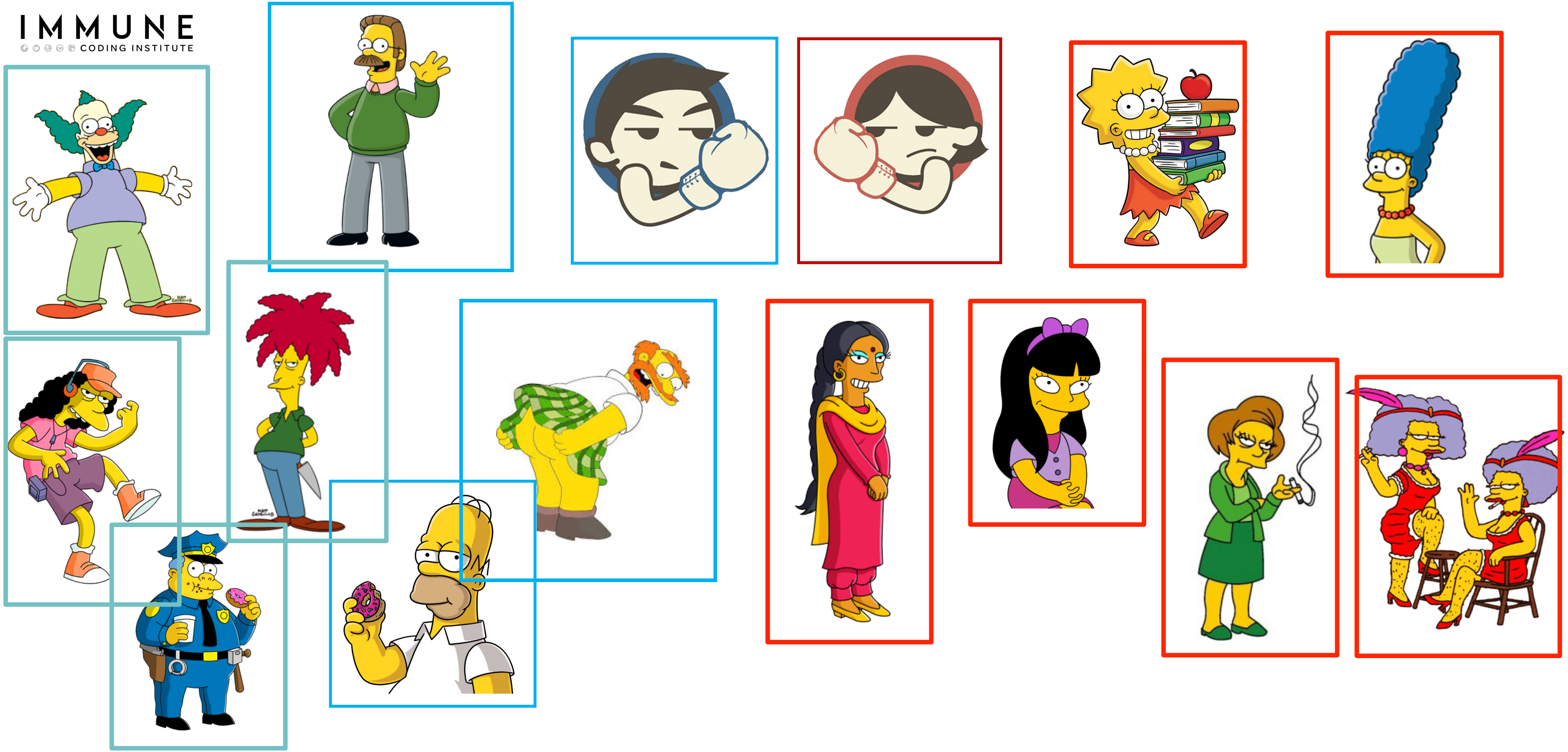
3) $\boxed{P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)}$


VS

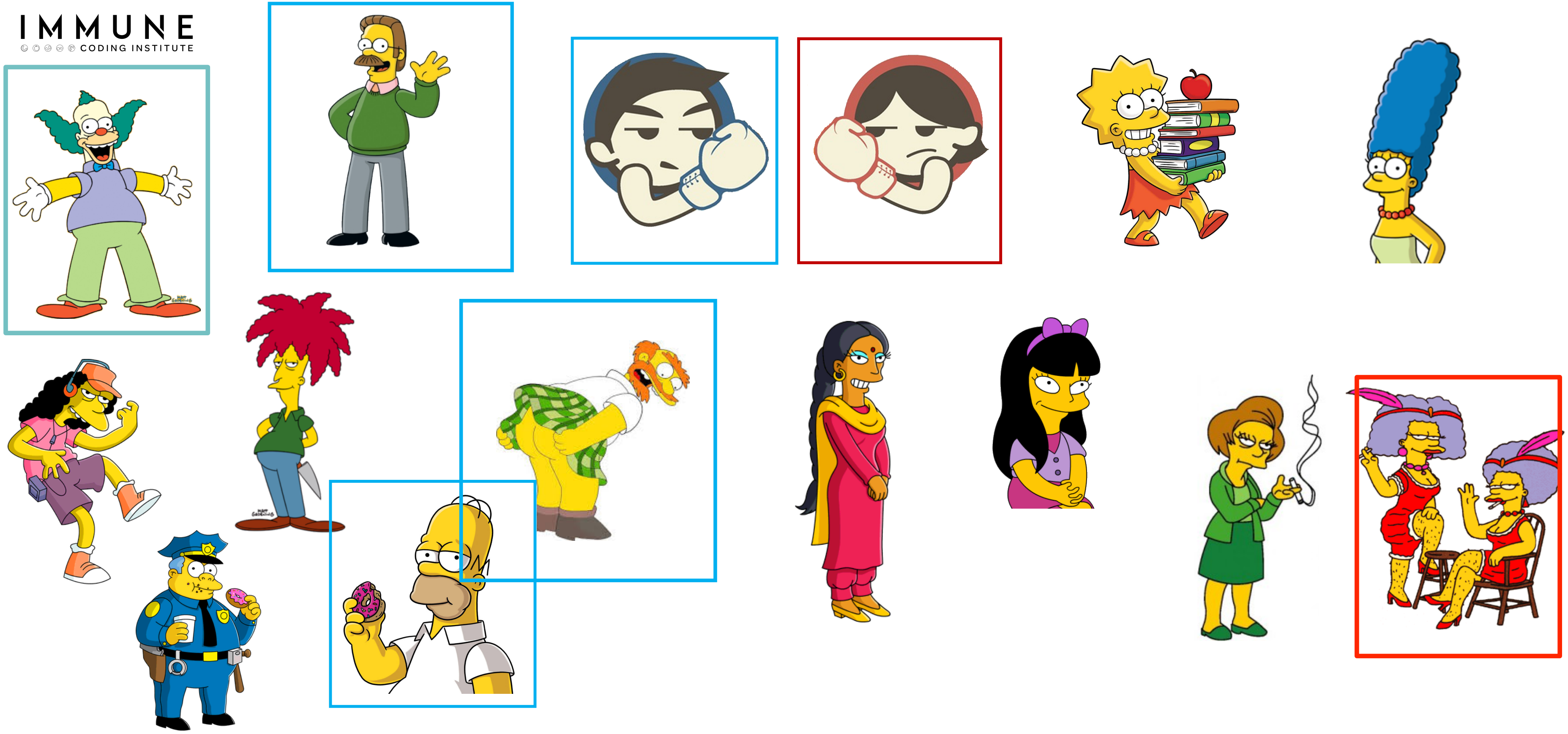


IMMUNE CODING INSTITUTE



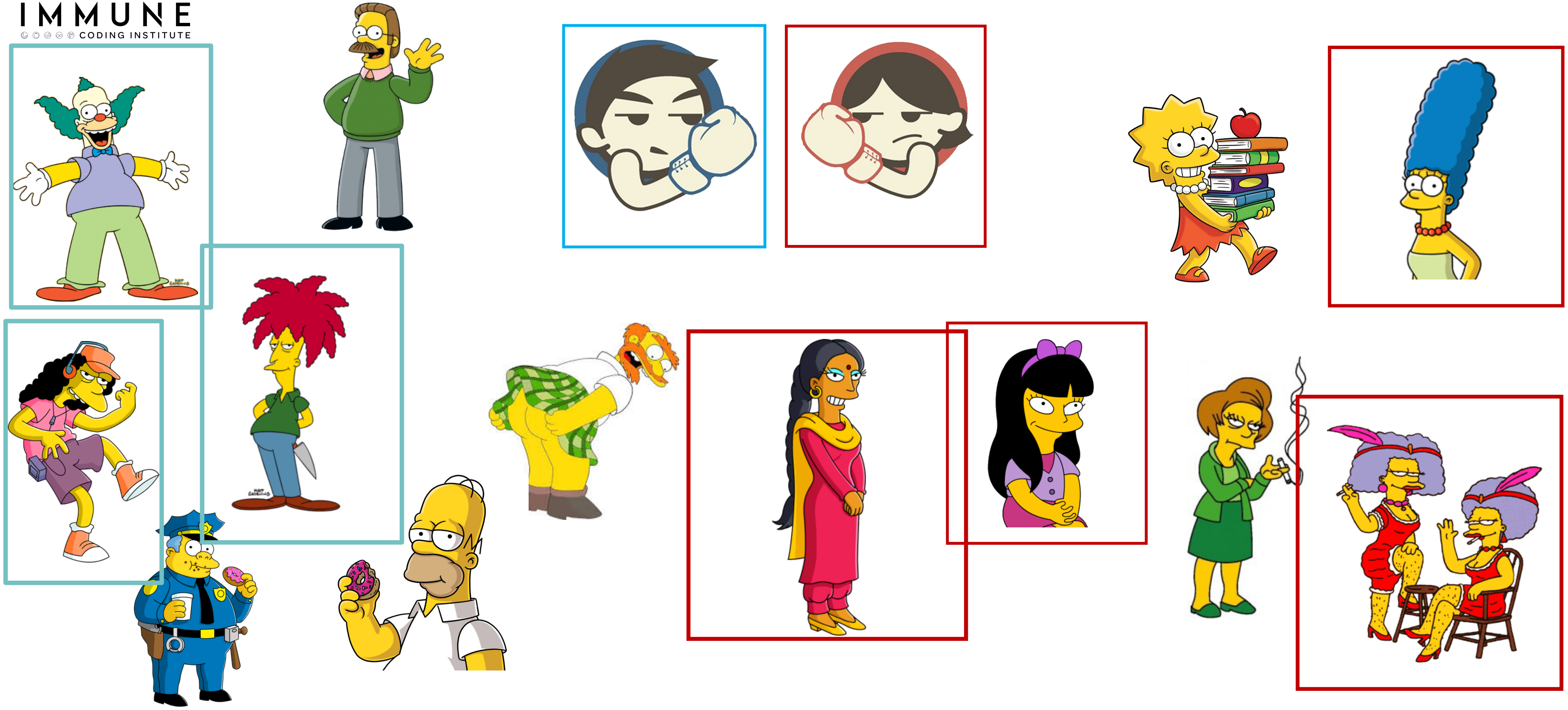


$$P(man) = 7/14 = 0.50 \qquad P(woman) = 7/14 = 0.50$$



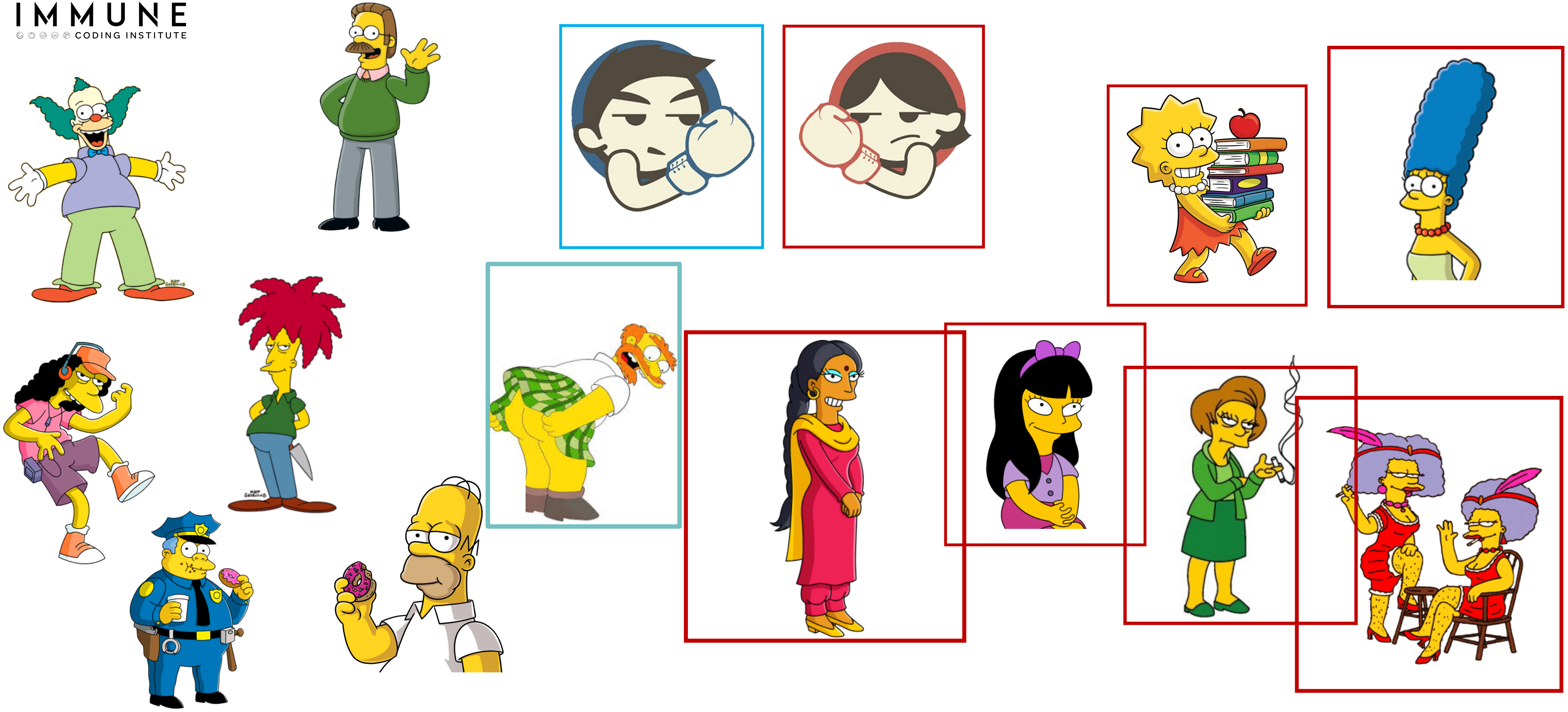
$$P(moustache|man) = 4/7 = 0.57 \qquad P(moustache|woman) = 2/7 = 0.28$$

IMMUNE CODING INSTITUTE

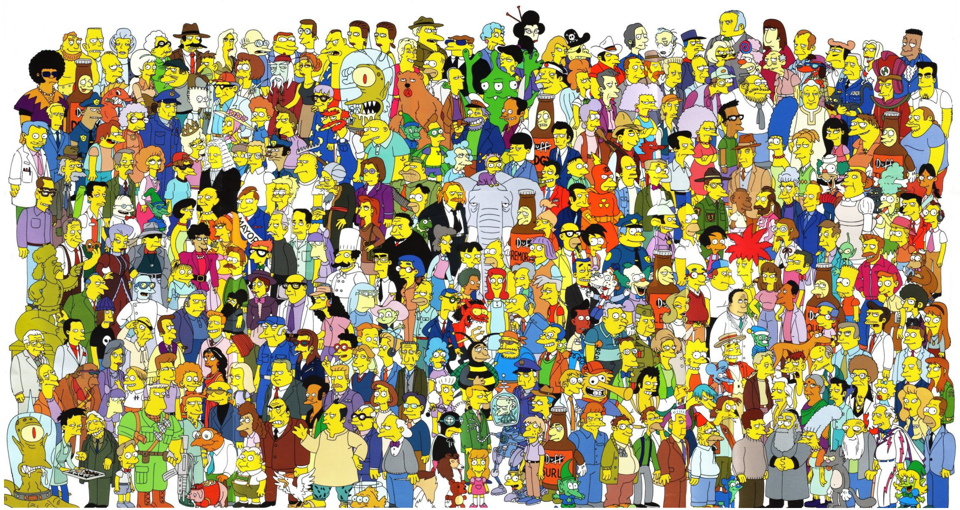


$$\boxed{P(long\ hair|man) = 3/7 = 0.42 \qquad P(long\ hair|woman) = 5/7 = 0.42}$$

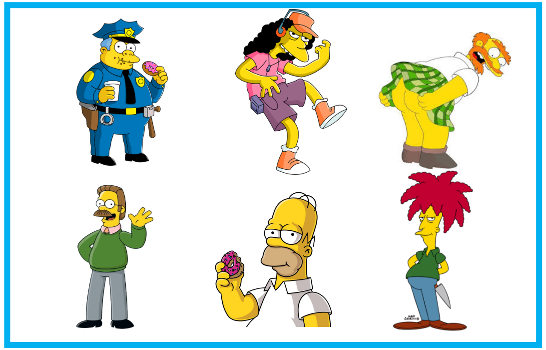
IMMUNE
CODING INSTITUTE

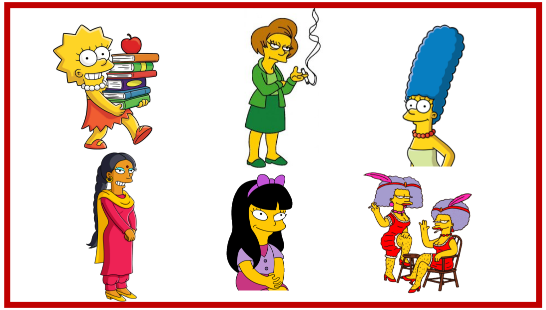


$$P(skirt|man) = 1/7 = 1 \qquad P(skirt|woman) = 7/7 = 1$$



Representative population

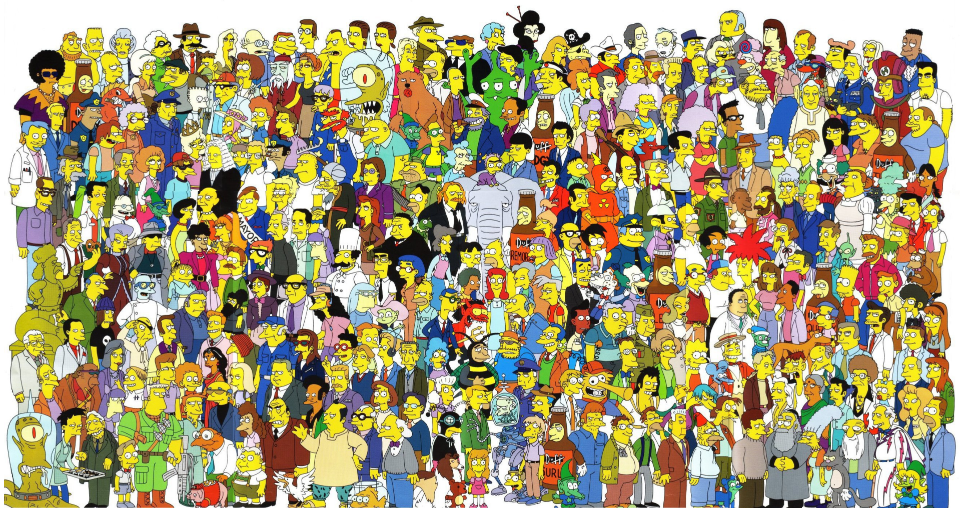




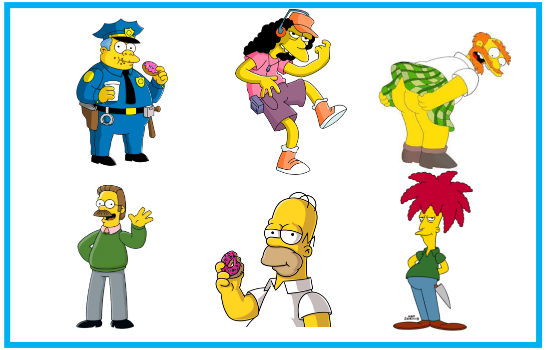
Representative population

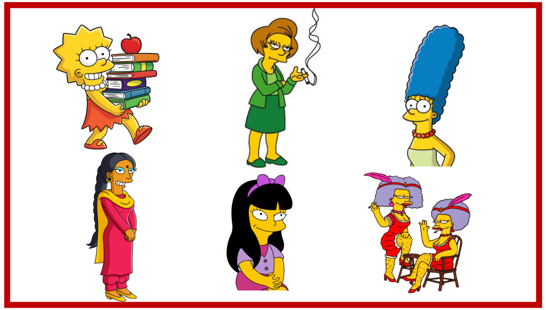
There will be determining characteristics for the male sex and others for the female

Are there cases where it is not so clear to which group the individual belongs?



Representative population





Representative population

There will be determining characteristics for the male sex and others for the female

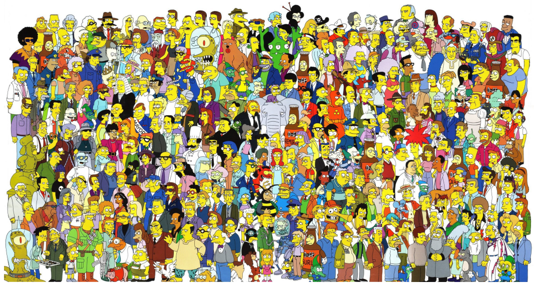
Are there cases where it is not so clear to which group the individual belongs?

Thinking Conditionally

Independent Events

A and B are independent if knowing whether A occurred gives no information about whether B occurred

$$P(A \cap B) = P(A)P(B)$$
$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$



Conditional independence

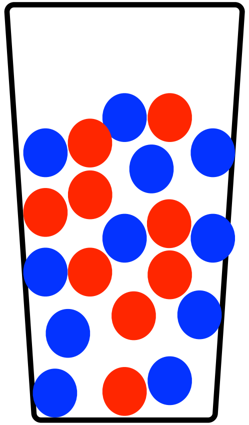
A and B are conditional independent given C if

$$P(A \cap B|C) = P(A|C)P(B|C)$$

Conditional independence does not imply independence, and independence does not imply conditional independence,

Thinking Conditionally

Let's say we have a glass with 20 balls in it:

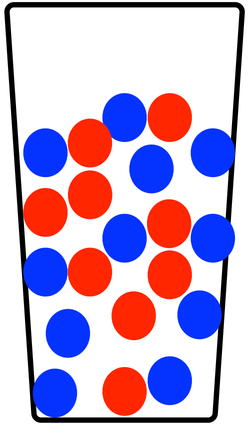


9 red balls A

11 blue balls B

Thinking Conditionally

Let's say we have a glass with 20 balls in it:



9 red balls A

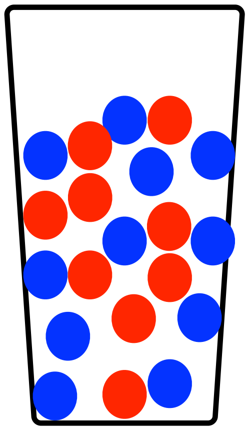
11 blue balls B

$$P(A) =$$

$$P(B) =$$

Thinking Conditionally

Let's say we have a glass with 20 balls in it:



9 red balls A

11 blue balls B

$$P(A) = \frac{9}{20} \qquad P(A|E_1) =$$

$$P(B) = \frac{11}{20} \qquad P(B|E_1) =$$

Bayes' Rules

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

is a way of finding a **probability** when we know certain other probabilities.

Which tells us: how often A happens given that B happens —> $P(A|B)$

When we know: how often B happens *given that A happens* —> $P(B|A)$

and how likely A is on its own —> $P(A)$

and how likely B is on its own —> $P(B)$

Bayes' Rules

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

Let's say $P(Fire|Smoke)$ means how often there is a fire when we can see smoke

Let's say $P(Smoke|Fire)$ means how often there is a smoke when there is fire

Example:

Fires are less probable: 1%

Smoke is faily common due to barbecues: 10%

Almost all fires produce smoke: 90%

IMMUNE CODING INSTITUTE

Bayes' Rules

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

Let's say $P(Fire|Smoke)$ means how often there is a fire when we can see smoke

Let's say $P(Smoke|Fire)$ means how often there is a smoke when there is fire

Example:

Fires are less probable: 1%

Smoke is faily common due to barbecues: 10%

Almost all fires produce smoke: 90%

$$P(Fire|Smoke) = \frac{P(Smoke|Fire)P(Fire)}{P(Smoke)}$$

$$P(Fire|Smoke) = \frac{0.9 * 0.1}{0.10}$$

Bayes' Rules

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

Example:

Let's imagine 100 people at a party, and you count how many wear pink or not, and if a man or not, and get these numbers:

	Pink	notPink
Man	5	35
notMan	20	40

$$P(Man|Pink) = ?$$

Bayes' Rules

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

One of most common uses for Bayes Theorem is the **False Positive** and **False Negatives**

Example: medical test

For the people that **really do** have the sickness, the test says **YES** 85% of the time

For the people that do not have the sickness, the test says **YES** 10% of the time. (FP)

If 12% of the population have the sickness, and the test says **Yes**, what are the chances that someone really hast the sickness?

Bayes' Rules

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

For the people that **really do** have the sickness, the test says **YES** 85% of the time

For the people that do not have the sickness, the test says **YES** 10% of the time. (FP)

If 12% of the population have the sickness, and the test says **Yes**, what are the chances that someone really hast the sickness?

$$P(Sick|Yes) = ?$$

Bayes' Rules

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

For the people that **really do** have the sickness, the test says **YES** 85% of the time

For the people that do not have the sickness, the test says **YES** 10% of the time. (FP)

If 12% of the population have the sickness, and the test says **Yes**, what are the chances that someone really hast the sickness?

$$P(Sick|Yes) = ?$$

$$P(Sick|Yes) = \frac{P(Yes|Sick)P(Sick)}{P(Yes)}$$

Bayes' Rules

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

For the people that **really do** have the sickness, the test says **YES** 85% of the time

For the people that do not have the sickness, the test says **YES** 10% of the time. (FP)

If 12% of the population have the sickness, and the test says **Yes**, what are the chances that someone really hast the sickness?

$$P(Sick|Yes) = ? \qquad P(Yes) = 0.12 * 0.85 + 0.10 * 0.88$$

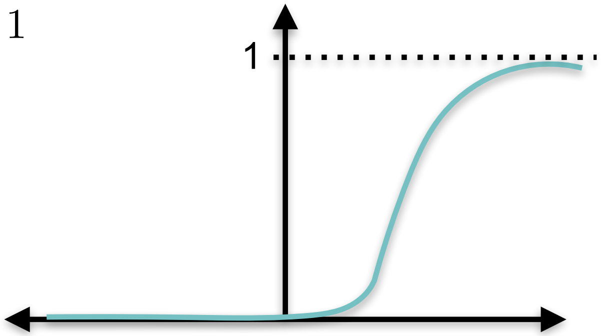
$$P(Sick|Yes) = \frac{P(Yes|Sick)P(Sick)}{P(Yes)} = \frac{0.85 * 0.12}{0.19}$$

Random Variables

Random Variables: a random variable, noted by X, is a function maps every element in a sample space to real line.

Cumulative distribution function: The cumulative distribution F, which is monotonically non-decreasing and such as

$$\lim_{x \to -\inf} F(x) = 0 \qquad \lim_{x \to \inf} F(x) = 1$$



is defined as: $$\boxed{F(x) = P(X \leq x)}$$

Remark: $P(a < X \leq b) = F(a) - F(b)$

Random Variables

Probability density function: is the probability that takes X on values between two adjacent realizations of random variables

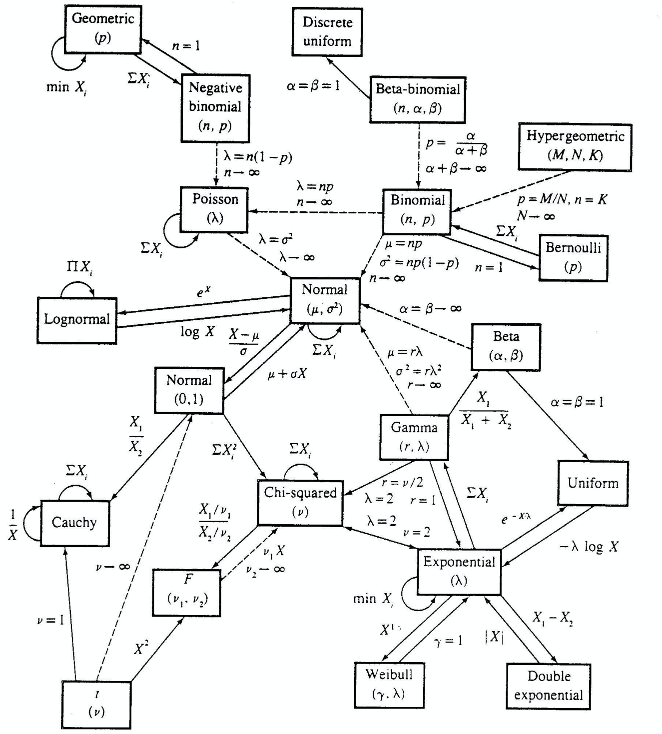
Relationship between Distribution and Density functions:

Case	CDF F	PDF f	Properties of PDF
(D)	$F(x) = \sum_{x_i \leqslant x} P(X = x_i)$	$f(x_j) = P(X = x_j)$	$0 \leqslant f(x_j) \leqslant 1$ and $\sum_j f(x_j) = 1$
(C)	$F(x) = \int_{-\infty}^{x} f(y)dy$	$f(x) = \frac{dF}{dx}$	$f(x) \geqslant 0$ and $\int_{-\infty}^{+\infty} f(x)dx = 1$

IMMUNE
CODING INSTITUTE

Main Distributions

630 *Table of Common Distributions*



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

Continuous Distributions: A probability distribution in which the random variable X can take on any value (is continuous). Example: The normal distribution

Discrete Distributions: A discrete distribution means that X can assume one of a countable (usually finite) number of values. Example: roll a fair dice.

python™ Let's code!

Central Limit Theorem

Let us have a random sample $X_1, X_2, \ldots, X_n$ following a given distribution with mean μ and σ^2 then we have:

$$\boxed{\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)}$$

“The central limit theorem states that the sum of a number of independent and identically distributed random variables with finite variances will tend to a normal distribution as the number of variables grows”

Let’s code!

IMMUNE CODING INSTITUTE

IMMUNE

CODING INSTITUTE