# Probability in Real-Life

Example Applications from Visual Neuroscience, Colour Blindness Detection and Covid-19 Outbreak Modelling

---

Karthik PN
Department of ECE,
IISc, Bangalore
Contact: periyapatna@iisc.ac.in, pnkarthik1992@gmail.com

September 28, 2020

# Dedication

Prof. Calyampudi Radhakrisha Rao (C. R. Rao)

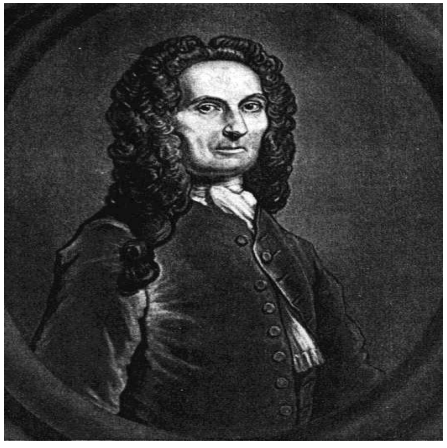
Sep 10, 2020 - present

Pic courtesy: Google

# The Beginnings

# Humble Beginnings: The Frequentist Approach

Jacob Bernoulli

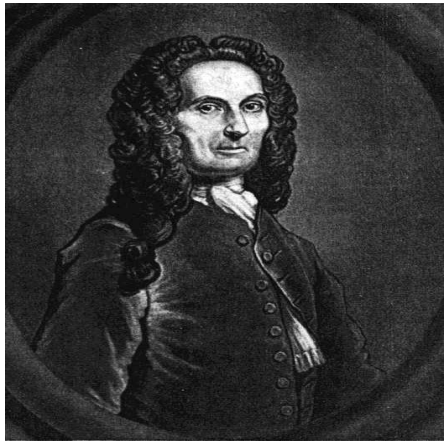


Abraham de Moivre

- Era: 1718

# Humble Beginnings: The Frequentist Approach

Jacob Bernoulli

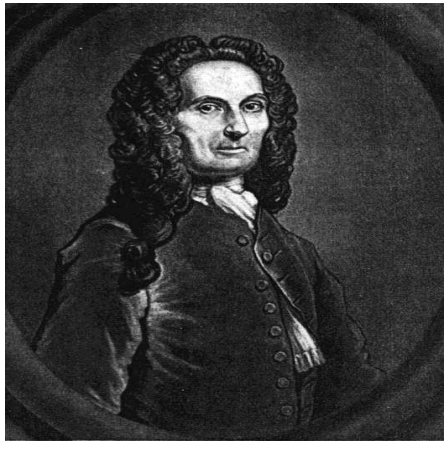


Abraham de Moivre

- Era: 1718
- Probability – the "chance" of occurrence

# Humble Beginnings: The Frequentist Approach

Jacob Bernoulli

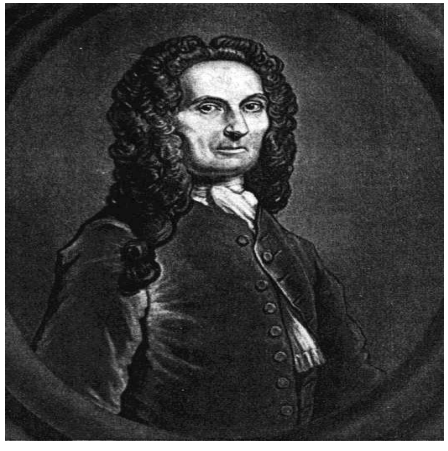


Abraham de Moivre

- Era: 1718
- Probability – the "chance" of occurrence
- Consider an experiment $E$ that is repeated $N$ times; e.g., $N \sim 10^5$

# Humble Beginnings: The Frequentist Approach

Jacob Bernoulli



Abraham de Moivre

- Era: 1718
- Probability – the "chance" of occurrence
- Consider an experiment $E$ that is repeated $N$ times; e.g., $N \sim 10^5$
- Probability of the event $E$ is then given by

$$P(E) = \frac{\#\text{ times event } E \text{ occurred}}{N}$$

- Frequentist notion of probability

# Criticisms to the Frequentist Approach

- The German and English mathematicians were opposed to the idea of computing likelihood of occurrence of events via the frequentist approach
- Some experiments cannot be repeated multiple times.
  E.g., appearance of the Halley's comet, a pandemic breakout

# Criticisms to the Frequentist Approach

- The German and English mathematicians were opposed to the idea of computing likelihood of occurrence of events via the frequentist approach
- Some experiments cannot be repeated multiple times.
  E.g., appearance of the Halley's comet, a pandemic breakout
- There arose the need to come up with a watertight theory that enables carrying out analysis

# Criticisms to the Frequentist Approach

- The German and English mathematicians were opposed to the idea of computing likelihood of occurrence of events via the frequentist approach
- Some experiments cannot be repeated multiple times.
  E.g., appearance of the Halley's comet, a pandemic breakout
- There arose the need to come up with a watertight theory that enables carrying out analysis

In enter Borel and Lebesgue!

# Borel and Lebesgue's Measure Theory

Émile Borel

Henri Lebesgue

- Era: 1894

# Borel and Lebesgue's Measure Theory

Émile Borel

Henri Lebesgue

- Era: 1894
- Émile Borel and his then PhD student Henri Lebesgue discovered the subject of Measure Theory

# Borel and Lebesgue's Measure Theory

Émile Borel

Henri Lebesgue

- Era: 1894
- Émile Borel and his then PhD student Henri Lebesgue discovered the subject of Measure Theory

# Our Third Hero

Andrey Kolmogorov

- In 1933, Kolmogorov borrowed the key concepts from measure theory and formulated an axiomatic theory of probability, known today as Probability Theory

# Our Third Hero

Andrey Kolmogorov

- In 1933, Kolmogorov borrowed the key concepts from measure theory and formulated an axiomatic theory of probability, known today as Probability Theory
- Eliminated the need to think of probabilities in terms of relative frequencies

# Our Third Hero

Andrey Kolmogorov

- In 1933, Kolmogorov borrowed the key concepts from measure theory and formulated an axiomatic theory of probability, known today as Probability Theory
- Eliminated the need to think of probabilities in terms of relative frequencies

"The theory of probability as a mathematical discipline can and should be developed from axioms in exactly the same way as Geometry and Algebra."

# Probability Theory: An Overview

# Random Variables and CDF

- Random variables: quantities whose values are uncertain and/or not known in advance
  - The number of spikes generated in the optic nerve when looking at an image
  - Incubation period of a patient exposed to the covid-19 virus
  - Strength of the signal received at the receiver antenna

# Random Variables and CDF

- Random variables: quantities whose values are uncertain and/or not known in advance
  - The number of spikes generated in the optic nerve when looking at an image
  - Incubation period of a patient exposed to the covid-19 virus
  - Strength of the signal received at the receiver antenna
- Suppose $X$ is a random variable. Its cumulative distribution function (CDF), denoted $F_X$, is a function

$$F_X : \mathbb{R} \longrightarrow [0, 1],$$

defined as

$$F_X(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

# Discrete Random Variables and PMF

- A random variable taking finite or countably infinitely many values is known as a discrete random variable

# Discrete Random Variables and PMF

- A random variable taking finite or countably infinitely many values is known as a discrete random variable
  - The number of people who turn up for a live seminar
  - The number of tosses of a fair coin until the pattern HTH appears for the first time

# Discrete Random Variables and PMF

- A random variable taking finite or countably infinitely many values is known as a discrete random variable
  - The number of people who turn up for a live seminar
  - The number of tosses of a fair coin until the pattern HTH appears for the first time
- Suppose $X$ is a discrete random variable. Its **probability mass function (PMF)**, denoted $p_X$, is a function

$$p_X : \mathbb{R} \longrightarrow [0, 1],$$

defined as

$$p_X(x) = P(X = x), \quad x \in \mathbb{R}.$$

# Some Important PMFs

- Bernoulli distribution
  - $X \in \{0, 1\}$
  - Characterised by a single parameter $p \in [0, 1]$
  - The PMF is given by

$$p_X(1) = p, \quad p_X(0) = 1 - p.$$

# Some Important PMFs

- Bernoulli distribution
  - $X \in \{0, 1\}$
  - Characterised by a single parameter $p \in [0, 1]$
  - The PMF is given by

$$p_X(1) = p, \quad p_X(0) = 1 - p.$$

- Binomial distribution
  - Characterised by two parameters $n \geq 1$ and $p \in [0, 1]$
  - $X \in \{0, 1, \ldots, n\}$
  - The PMF is given by

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \ldots, n\}.$$

# Some Important PMFs

- Multinomial distribution
  - Characterised by the parameters $n \geq 1$, $L \geq 2$ and $p_1, \ldots, p_L \in [0, 1]$
  - $X_1, \ldots, X_L \in \{0, \ldots, n\}$, $\sum_{i=1}^{L} X_i = n$
  - The joint PMF of $X_1, \ldots, X_L$ is given by

$$p_{X_1,\ldots,X_L}(k_1, \ldots, k_L) = \frac{n!}{k_1! \cdots k_L!} p_1^{k_1} \cdots p_L^{k_L}, \quad \sum_{i=1}^{L} k_i = n.$$

# Some Important PMFs

- Multinomial distribution
  - Characterised by the parameters $n \geq 1$, $L \geq 2$ and $p_1, \ldots, p_L \in [0, 1]$
  - $X_1, \ldots, X_L \in \{0, \ldots, n\}$, $\sum_{i=1}^{L} X_i = n$
  - The joint PMF of $X_1, \ldots, X_L$ is given by

$$p_{X_1,\ldots,X_L}(k_1, \ldots, k_L) = \frac{n!}{k_1! \cdots k_L!} p_1^{k_1} \cdots p_L^{k_L}, \quad \sum_{i=1}^{L} k_i = n.$$

- Poisson distribution
  - Characterised by a parameter $\lambda > 0$
  - $X \in \{0, 1, 2, \ldots\}$
  - The PMF is given by

$$p_X(k) = \frac{e^{-\lambda}\ \lambda^k}{k!}, \quad k \in \{0, 1, 2, \ldots\}.$$

# Continuous Random Variables and PDF

- A random variable is said to be a continuous random variable if its CDF is differentiable[1]

[1] For the mathematically inclined reader, a random variable is said to be continuous if its CDF is *absolutely continuous*.

# Continuous Random Variables and PDF

- A random variable is said to be a continuous random variable if its CDF is differentiable[1]
- Suppose $X$ is a continuous random variable. Its probability density function (PDF), denoted $f_X$, is a function

$$f_X : \mathbb{R} \longrightarrow [0, \infty),$$

defined as

$$f_X(x) = \frac{d}{dt} F_X(t)\Big|_{t=x}, \quad x \in \mathbb{R}.$$

[1] For the mathematically inclined reader, a random variable is said to be continuous if its CDF is *absolutely continuous*.

# Some Important PDFs

- Gaussian distribution
  - Characterised by two parameters $\mu \in \mathbb{R}$ and $\sigma > 0$
  - $X \in \mathbb{R}$
  - The PDF is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

# Some Important PDFs

- Gaussian distribution
  - Characterised by two parameters $\mu \in \mathbb{R}$ and $\sigma > 0$
  - $X \in \mathbb{R}$
  - The PDF is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

- Exponential Distribution
  - Characterised by a parameter $\nu > 0$
  - $X \in (0, \infty)$
  - The PDF is given by

$$f_X(x) = \nu e^{-\nu x}, \quad x \in (0, \infty).$$

# Some Important PDFs

- Gamma distribution
  - Characterised by two parameters $\alpha > 0$ (shape) and $\beta > 0$ (rate)
  - $X \in (0, \infty)$
  - The PDF is given by

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x \in (0, \infty).$$

# Some Important PDFs

- Gamma distribution
  - Characterised by two parameters $\alpha > 0$ (shape) and $\beta > 0$ (rate)
  - $X \in (0, \infty)$
  - The PDF is given by

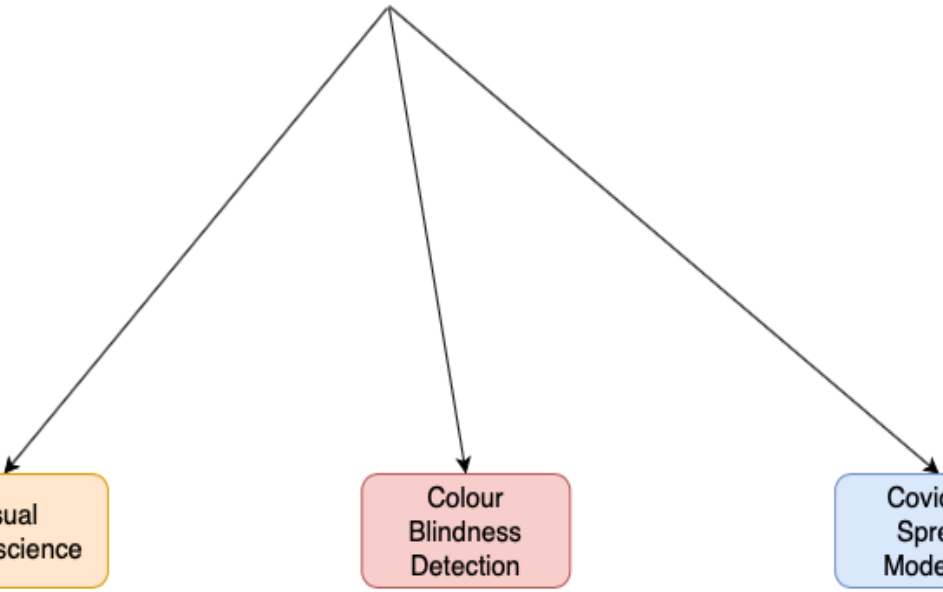
$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x \in (0, \infty).$$

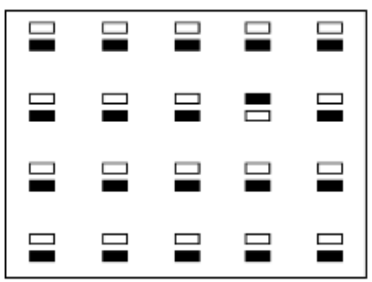
- Uniform distribution
  - Characterised by two parameters $a, b \in \mathbb{R}$, $a < b$
  - $X \in [a, b]$
  - The PDF is given by

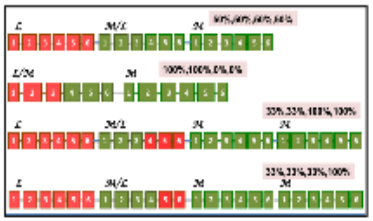
$$f_X(x) = \frac{1}{b-a}, \quad x \in [a, b]$$

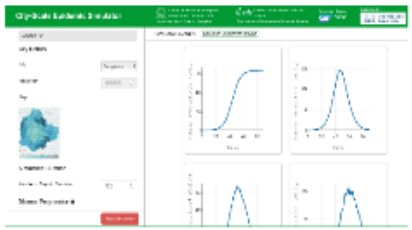
# Example Applications

# Rest of the Talk



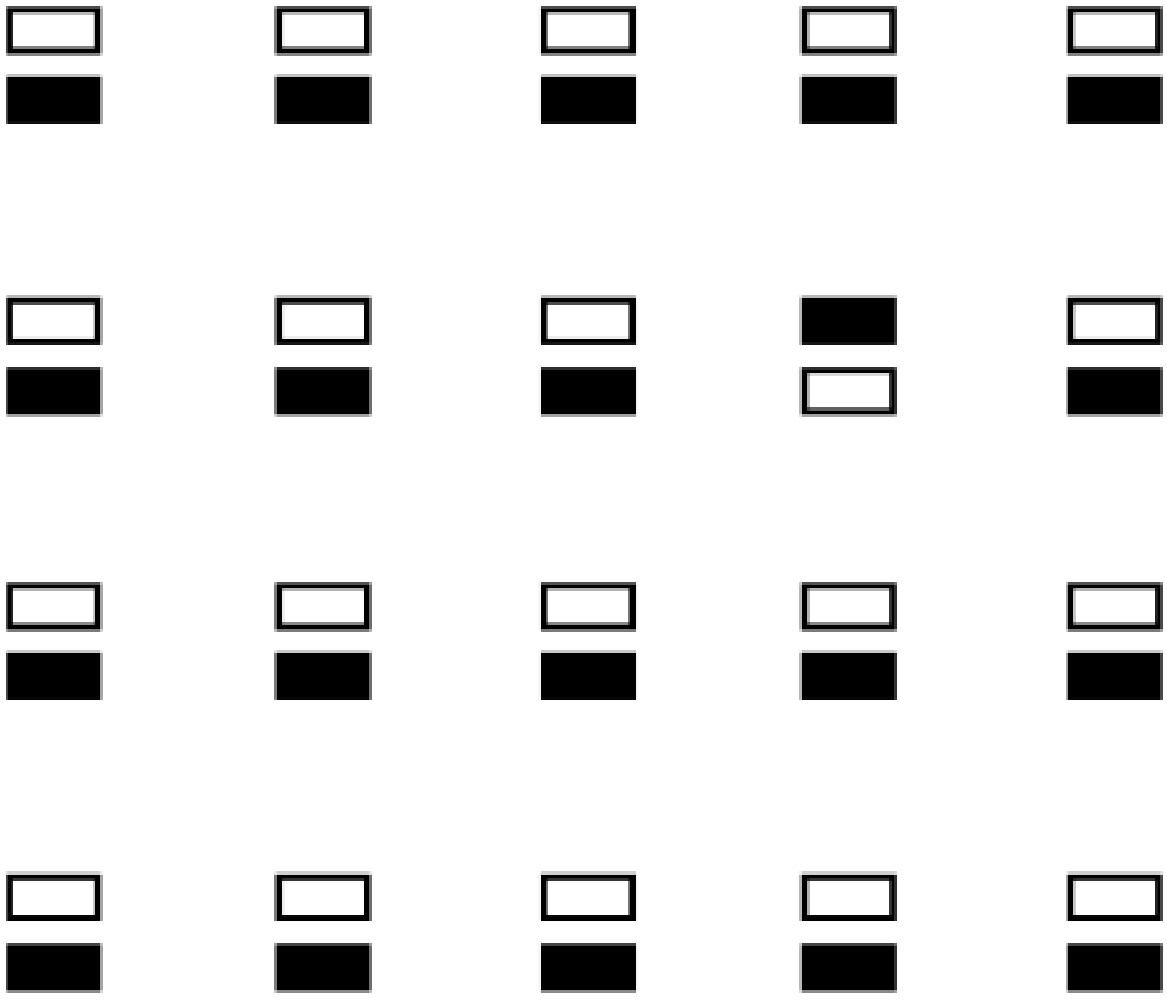





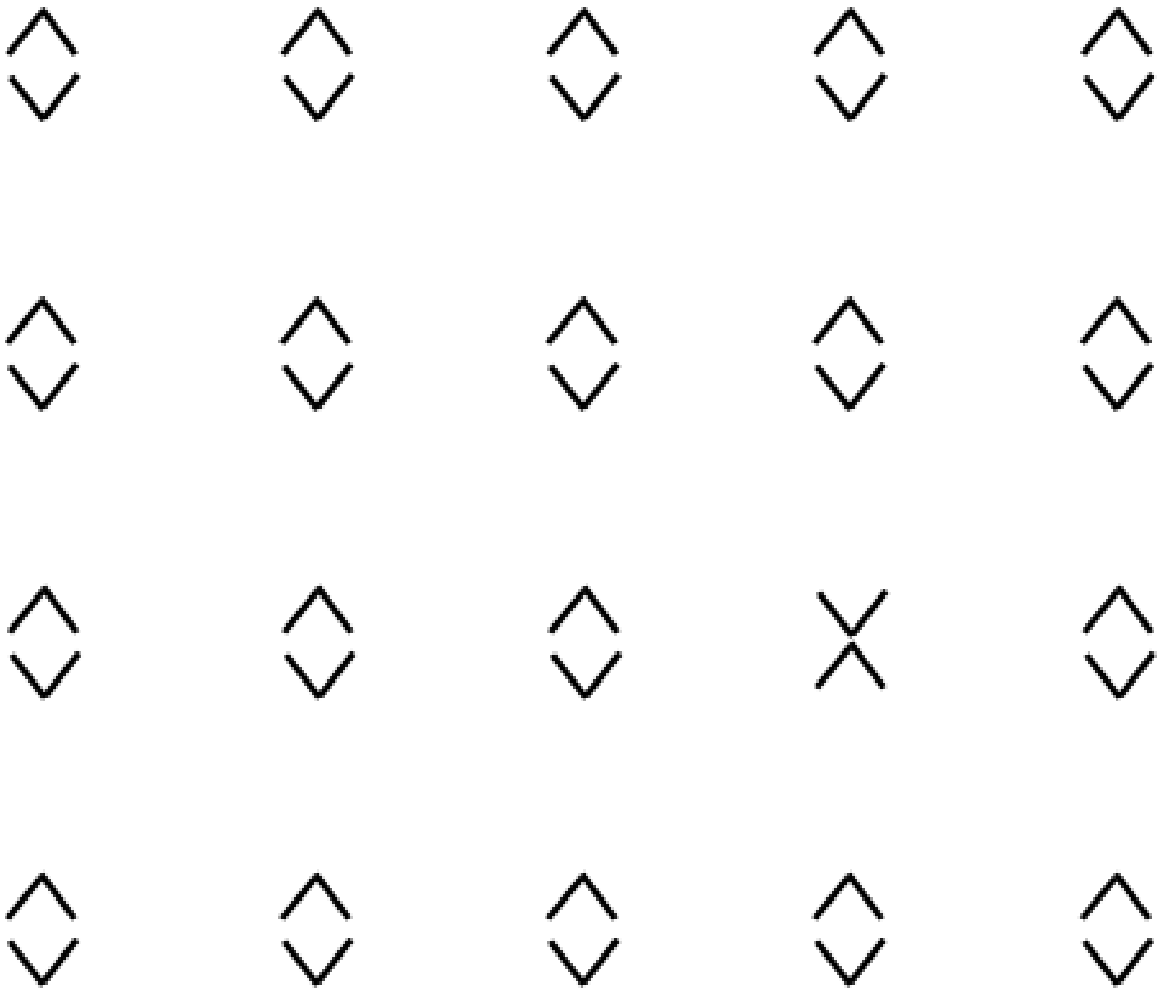
# Probability and Visual Neuroscience

# Identify the Odd Image - 1



# Identify the Odd Image - 2

# Identify the Odd Image - 3



# Search Time Distribution

- The time taken to search for the odd image is a function of the "oddball" and "distracter" images that are displayed

# Search Time Distribution

- The time taken to search for the odd image is a function of the "oddball" and "distracter" images that are displayed

- The search time is a non-negative, real-valued random variable whose distribution is not known

# Search Time Distribution

- The time taken to search for the odd image is a function of the "oddball" and "distracter" images that are displayed
- The search time is a non-negative, real-valued random variable whose distribution is not known
- The distribution of search time can be determined with the help of real-world data

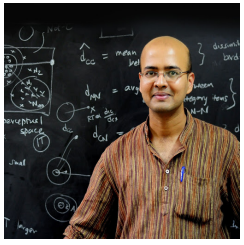
# Some Real-World Data

| 2 | Colour | Colour | Pattern | Pattern | Chevron | Chevron |
|---|---|---|---|---|---|---|
| 3 | Oddball L | Oddball R | Oddball L | Oddball R | Oddball L | Oddball R |
| 4 | 375 | 5025 | 771 | 1319 | 485 | 501 |
| 5 | 1425 | 1146 | 1490 | 1149 | 554 | 655 |
| 6 | 2088 | 1540 | 1532 | 1431 | 820 | 932 |
| 7 | 875 | 1422 | 994 | 542 | 875 | 1263 |
| 8 | 1373 | 1646 | 1590 | 1160 | 490 | 658 |
| 9 | 2036 | 1866 | 1917 | 3182 | 875 | 1156 |
| 10 | 875 | 985 | 1161 | 651 | 381 | 490 |
| 11 | 1095 | 1260 | 1146 | 1037 | 435 | 490 |
| 12 | 1531 | 1147 | 1435 | 983 | 546 | 545 |
| 13 | 1480 | 656 | 831 | 709 | 820 | 656 |
| 14 | 930 | 2196 | 2687 | 1765 | 655 | 491 |
| 15 | 1536 | 1155 | 2480 | 1698 | 711 | 490 |
| 16 | 710 | 927 | 762 | 610 | 445 | 441 |
| 17 | 1095 | 881 | 775 | 831 | 380 | 545 |


Color


Pattern


Chevron


Arun Sripati


Carl R. Olson

[2] Sripati, A.P., and Olson, C.R., 2010, Responses to compound objects in monkey inferotemporal cortex: the whole is equal to the sum of the discrete parts. J. Neurosci. 30: 7948-7960.

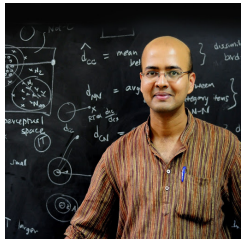
# Some Real-World Data

| 2 | Colour | Colour | Pattern | Pattern | Chevron | Chevron |
|---|---|---|---|---|---|---|
| 3 | Oddball L | Oddball R | Oddball L | Oddball R | Oddball L | Oddball R |
| 4 | 375 | 5025 | 771 | 1319 | 485 | 501 |
| 5 | 1425 | 1146 | 1490 | 1149 | 554 | 655 |
| 6 | 2088 | 1540 | 1532 | 1431 | 820 | 932 |
| 7 | 875 | 1422 | 994 | 542 | 875 | 1263 |
| 8 | 1373 | 1646 | 1590 | 1160 | 490 | 658 |
| 9 | 2036 | 1866 | 1917 | 3182 | 875 | 1156 |
| 10 | 875 | 985 | 1161 | 651 | 381 | 490 |
| 11 | 1095 | 1260 | 1146 | 1037 | 435 | 490 |
| 12 | 1531 | 1147 | 1435 | 983 | 546 | 545 |
| 13 | 1480 | 656 | 831 | 709 | 820 | 656 |
| 14 | 930 | 2196 | 2687 | 1765 | 655 | 491 |
| 15 | 1536 | 1155 | 2480 | 1698 | 711 | 490 |
| 16 | 710 | 927 | 762 | 610 | 445 | 441 |
| 17 | 1095 | 881 | 775 | 831 | 380 | 545 |


Color


Pattern


Chevron


Arun Sripati

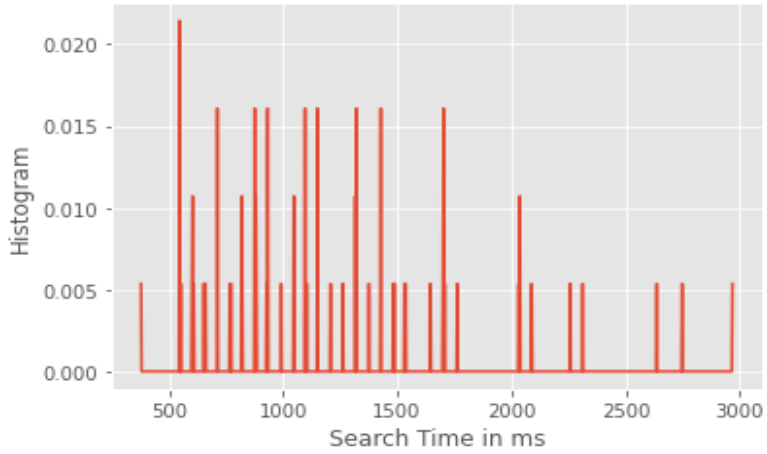

Carl R. Olson

The above data[2] shows the search times (in ms) measured for each of the image pairs shown on the right. This data was collected for 6 individuals, with 12 measurements for each individual.

[2] Sripati, A.P., and Olson, C.R., 2010, Responses to compound objects in monkey inferotemporal cortex: the whole is equal to the sum of the discrete parts. J. Neurosci. 30: 7948-7960.

# Determining the Best-Fit Distribution[3]

```python
import numpy
import pylab
import pandas as pd
import random
from scipy import stats


data = pd.read_csv('search_times_oddball_L.csv', header=None)

# indices for sampling the training data
len_data = len(data)
train_indices = random.sample(range(len_data), int(len_data/2))
training_data = data.loc[train_indices]
test_indices = [x for x in range(len_data) if x not in train_indices]
testing_data = data.loc[test_indices]

num_bins = 100
counts, bin_edges = numpy.histogram(training_data, bins=num_bins, normed=True)
cdf = numpy.cumsum(counts/sum(counts))
pylab.plot(bin_edges[1:len(bin_edges)], counts)
pylab.xlabel('Search Time in ms')
pylab.ylabel('Histogram of training data')

# sample mean and variance
sample_mean = numpy.mean(testing_data.to_numpy())
sample_var = numpy.var(testing_data.to_numpy())

# estimate of shape and rate parameters
rate_param = sample_mean / sample_var
shape_param = sample_mean * rate_param

# KS test to report goodness of fit of Gamma distribution

D, p = stats.kstest(training_data, 'gamma', args=(shape_param, 0, 1./rate_param))
```




Kolmogorov-Smirnoff test results:

- $D = 0.973$
- $p_{\text{value}} = p = 6.57 \times 10^{-57}$

[3] Carried out for the "Oddball L" data under the "Colour" column.

# Determining the Best-Fit Distribution[3]

```python
import numpy
import pylab
import pandas as pd
import random
from scipy import stats


data = pd.read_csv('search_times_oddball_L.csv', header=None)

# indices for sampling the training data
len_data = len(data)
train_indices = random.sample(range(len_data), int(len_data/2))
training_data = data.loc[train_indices]
test_indices = [x for x in range(len_data) if x not in train_indices]
testing_data = data.loc[test_indices]

num_bins = 100
counts, bin_edges = numpy.histogram(training_data, bins=num_bins, normed=True)
cdf = numpy.cumsum(counts/sum(counts))
pylab.plot(bin_edges[1:len(bin_edges)], counts)
pylab.xlabel('Search Time in ms')
pylab.ylabel('Histogram of training data')

# sample mean and variance
sample_mean = numpy.mean(testing_data.to_numpy())
sample_var = numpy.var(testing_data.to_numpy())

# estimate of shape and rate parameters
rate_param = sample_mean / sample_var
shape_param = sample_mean * rate_param

# KS test to report goodness of fit of Gamma distribution

D, p = stats.kstest(training_data, 'gamma', args=(shape_param, 0, 1./rate_param))
```


Kolmogorov-Smirnoff test results:

- $D = 0.973$
- $p_{\text{value}} = p = 6.57 \times 10^{-57}$

A Gamma distribution is fit to the training dataset, with the parameters of the Gamma distribution estimated from the testing dataset

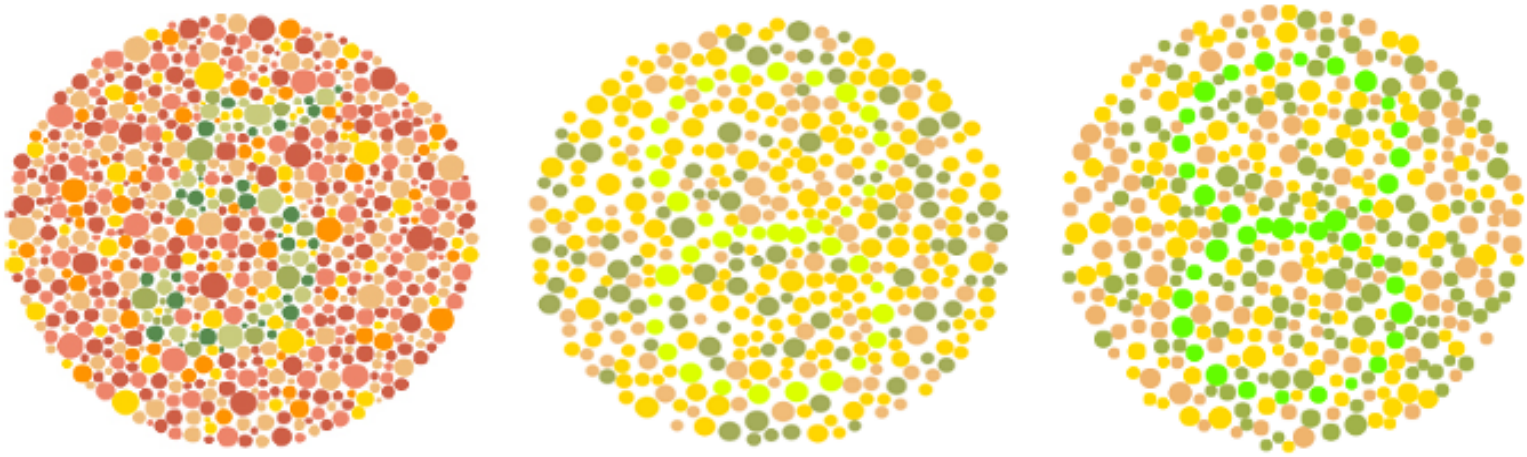
[3] Carried out for the "Oddball L" data under the "Colour" column.

# Probability and Colour Blindness Detection

# Colour Blindness

[4]Known as *Ishihara* cards.

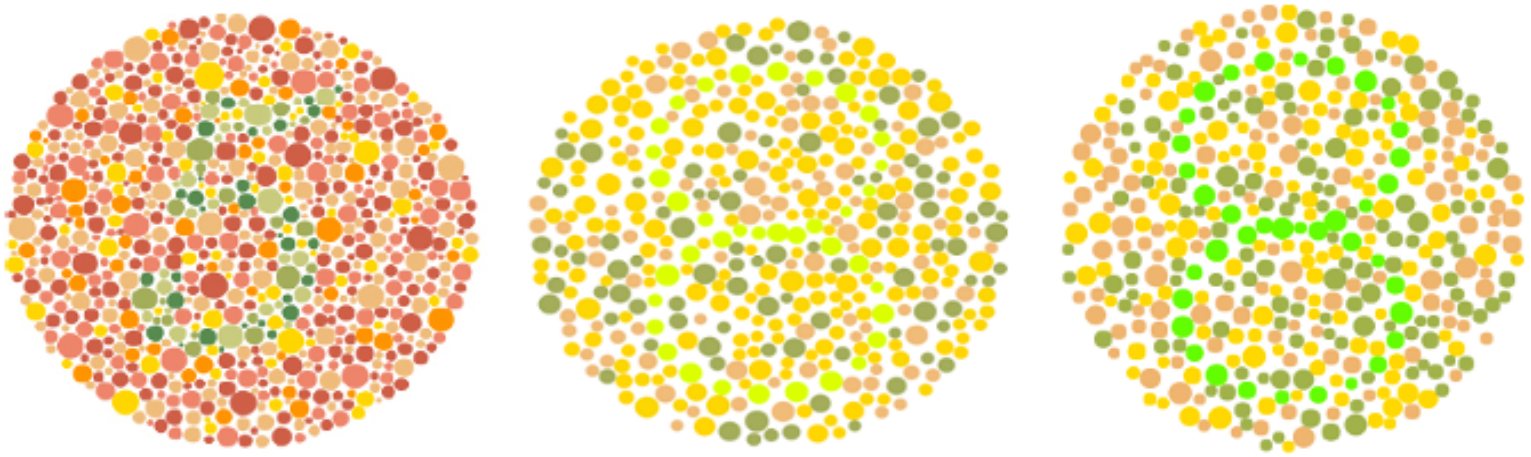
# Colour Blindness



- About 5% of the people cannot spot the numbers in these cards[4]

[4] Known as *Ishihara* cards.

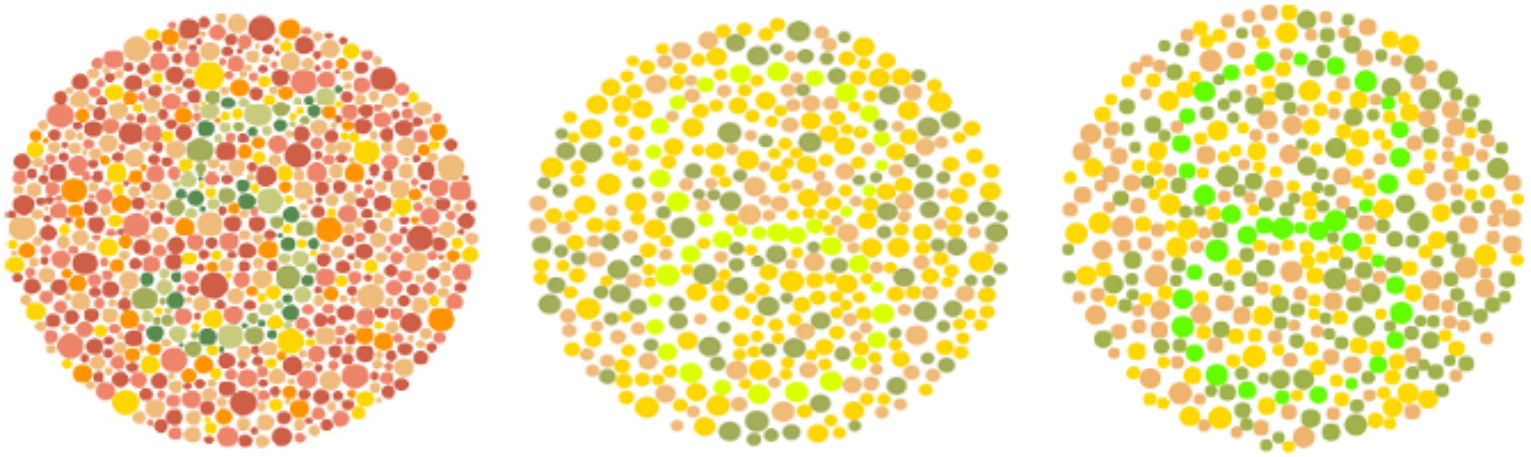
# Colour Blindness



- About 5% of the people cannot spot the numbers in these cards[4]
- These 5% are predominantly males!

[4] Known as *Ishihara* cards.

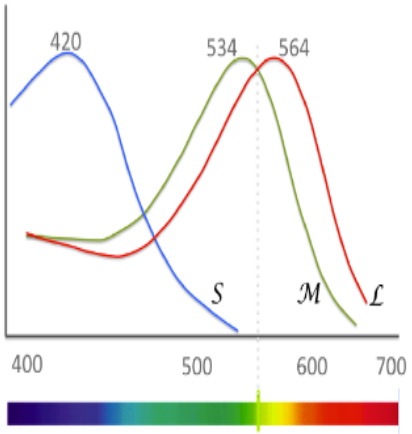
# Colour Blindness



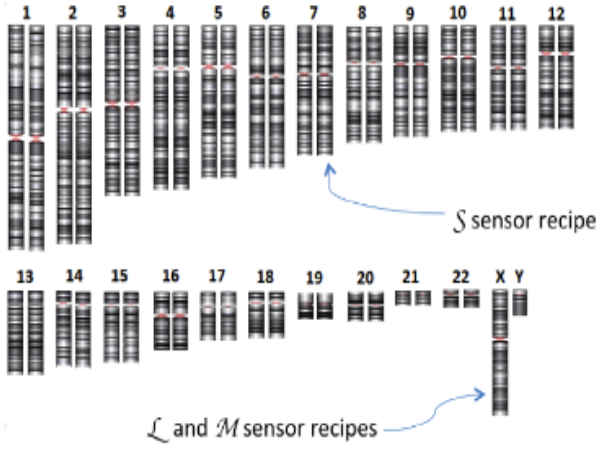
- About 5% of the people cannot spot the numbers in these cards[4]
- These 5% are predominantly males!
- What is the cause of this problem?

[4] Known as *Ishihara* cards.

# Rods and Cones: The Colour Sensors in the Eyes



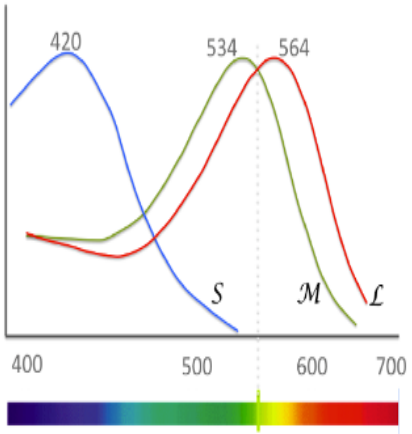

Red, green and blue wavelengths



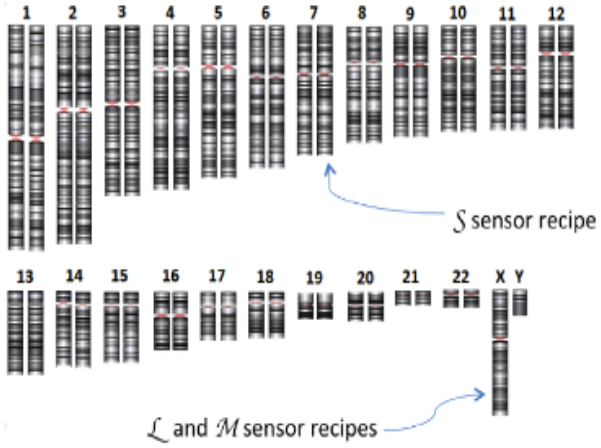

A representation of genomic information in a typical cell of males

- Rods and cones are responsible for colour perception

# Rods and Cones: The Colour Sensors in the Eyes



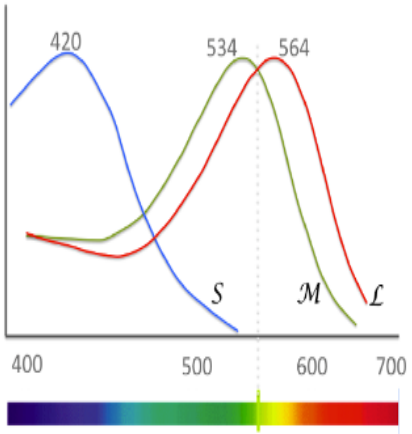

Red, green and blue wavelengths



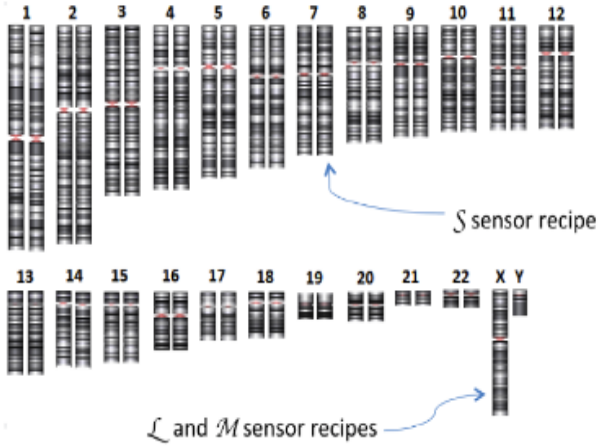

A representation of genomic information in a typical cell of males

- Rods and cones are responsible for colour perception
- These cells (and all other cells) have 23 pairs of chromosomes. The last pair is XX for females and XY for males

# Rods and Cones: The Colour Sensors in the Eyes



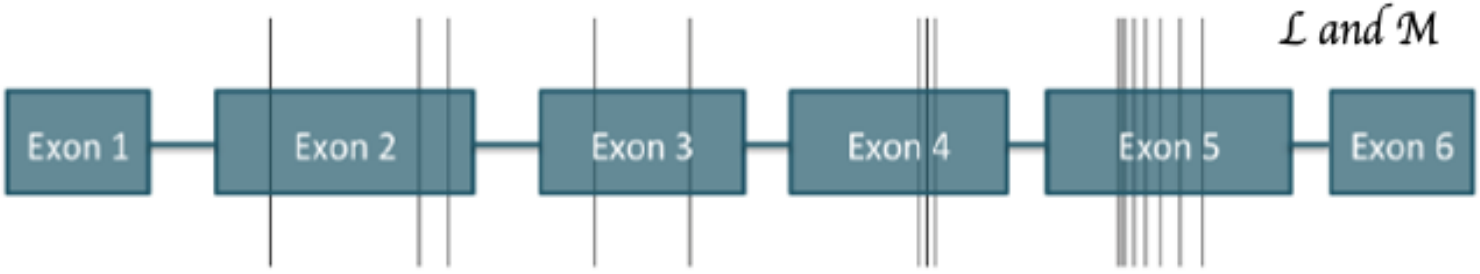

Red, green and blue wavelengths



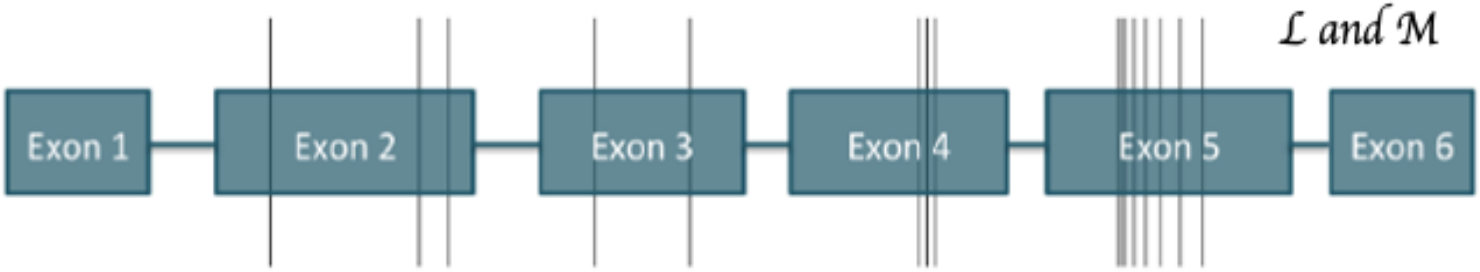

A representation of genomic information in a typical cell of males

- Rods and cones are responsible for colour perception
- These cells (and all other cells) have 23 pairs of chromosomes. The last pair is XX for females and XY for males
- The recipe for identifying red and green colours are in the X chromosome of the 23rd pair

# Exons: The Carriers of Genetic Information

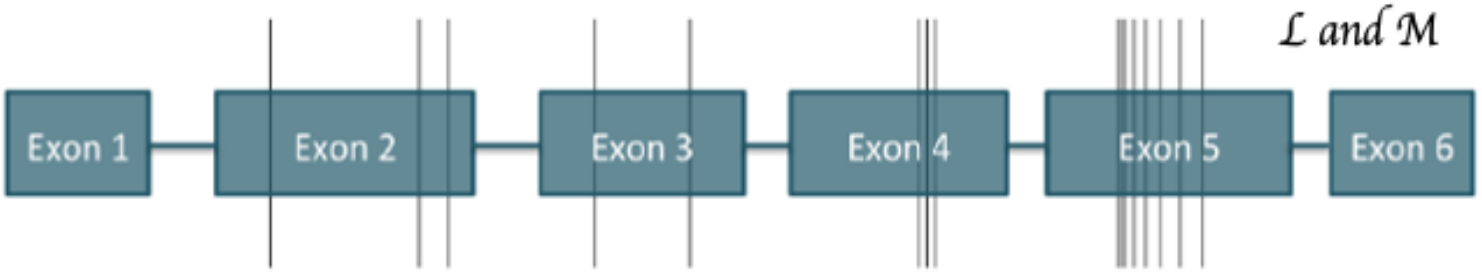


# Exons: The Carriers of Genetic Information




- The recipes for green and red colours have only 15 differences in their ACGT sequences, confined to exons 2, 3, 4 and 5

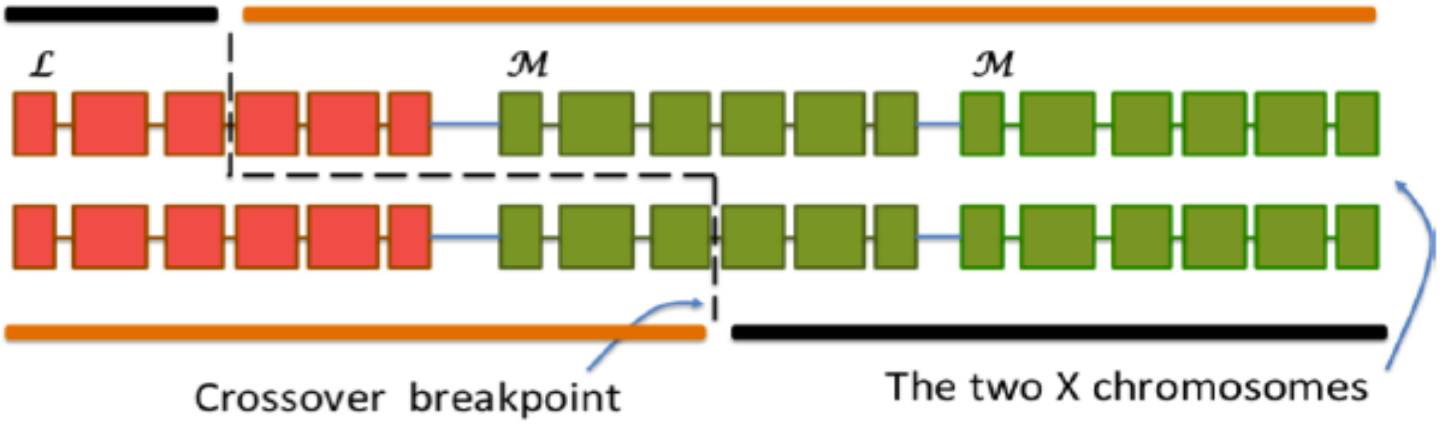
# Exons: The Carriers of Genetic Information



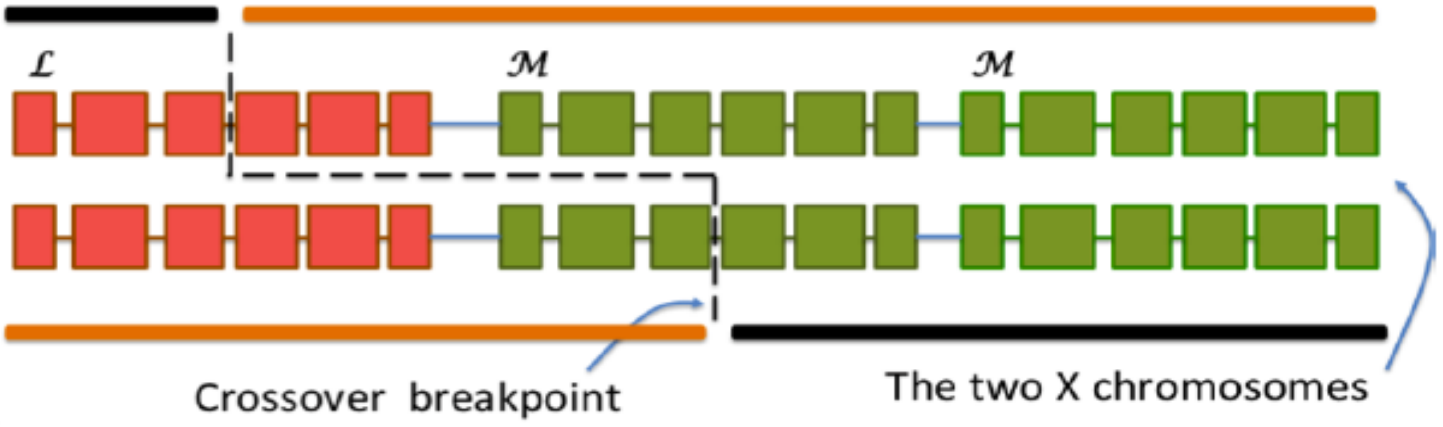

- The recipes for green and red colours have only 15 differences in their ACGT sequences, confined to exons 2, 3, 4 and 5

- The problem arises when the red and green exons of the two X chromosomes in females crossover with one another!

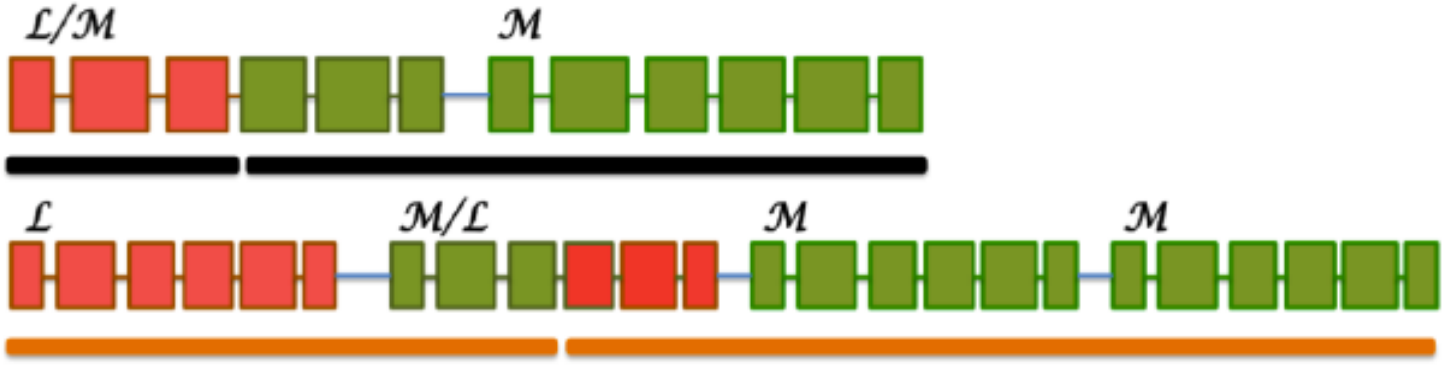
# Crossing Over



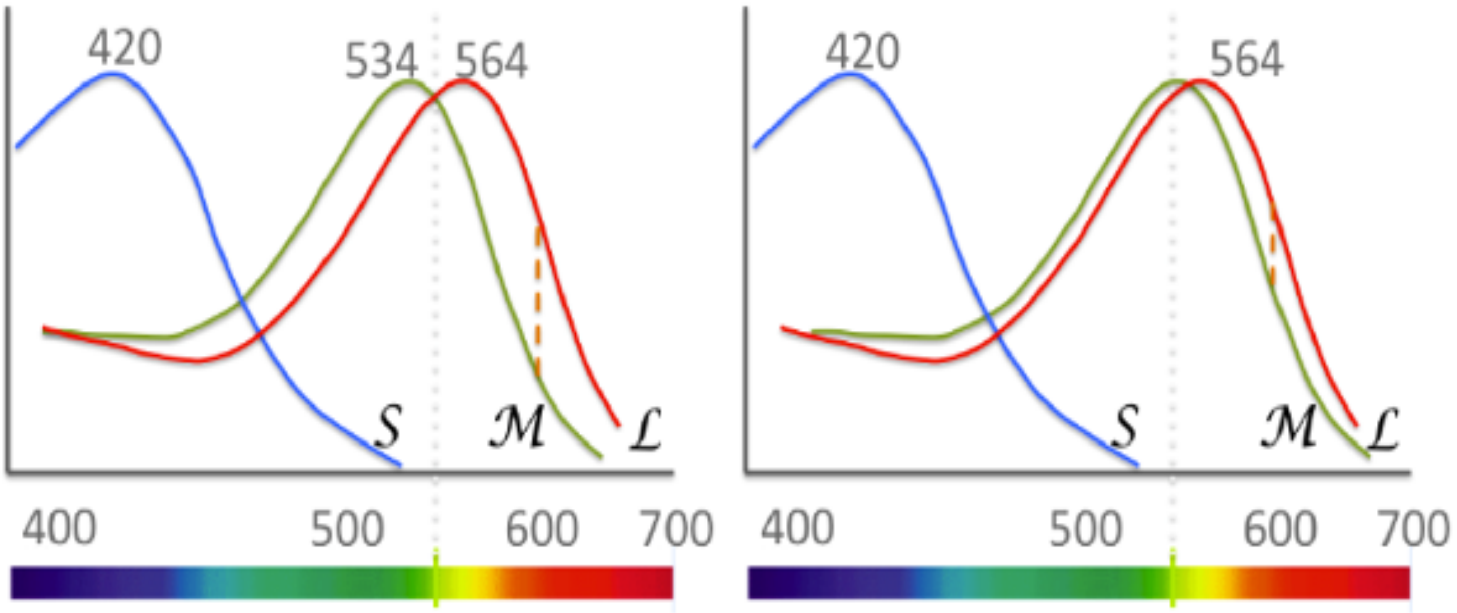
# Crossing Over




The end result of crossing over is the following sets of new genes:

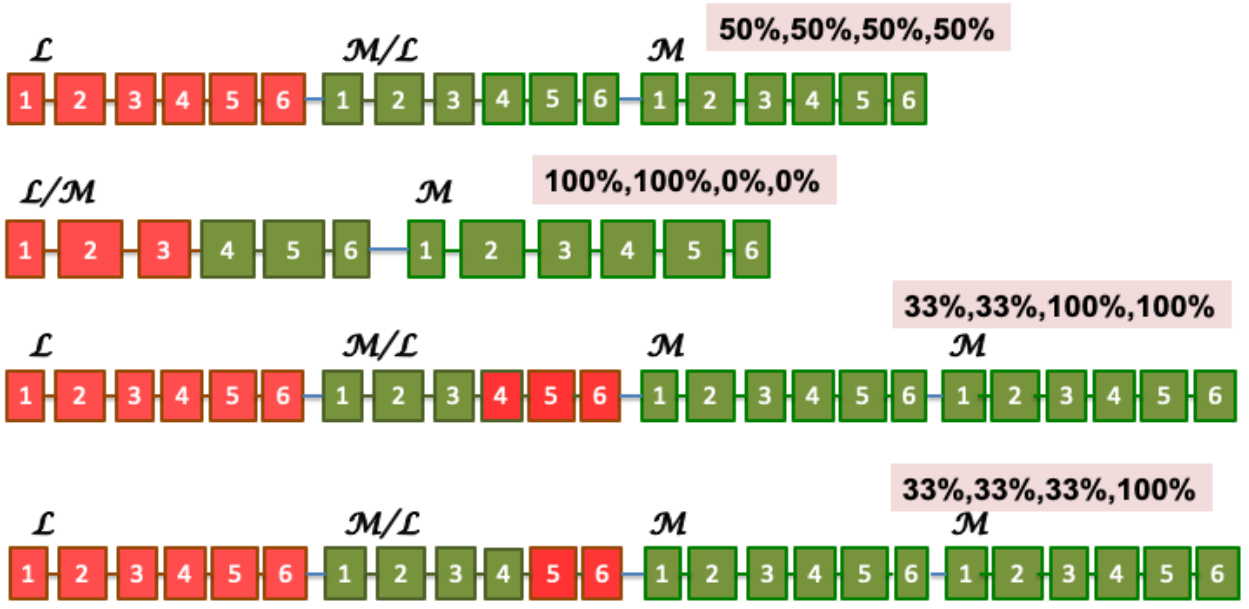


# Crossing Over




The red and green wavelength peaks coming may come closer to one another after crossing over!

# Likely Exon Configurations Leading to Colour Blindness




The most likely configuration leading to colour blindness in a colour blind person can be identified through genetic sequencing[5]

[5] Also known as Next-Generation DNA Sequencing (NGS).

# Genetic Sequencing

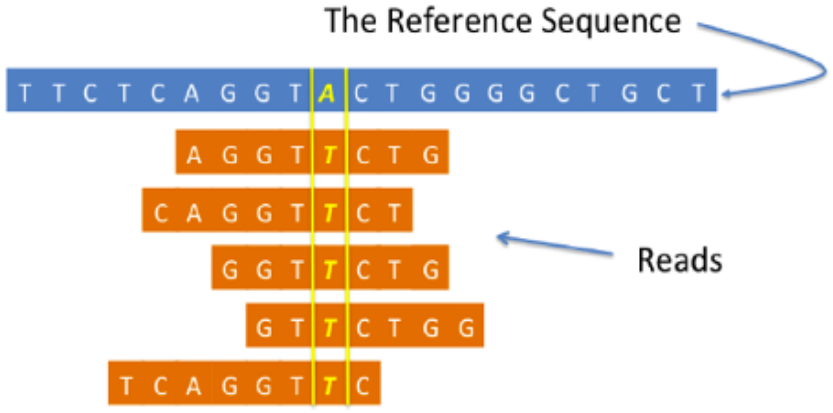
# Genetic Sequencing


- State-of-the-art genetic sequencing machines cannot handle the complete genetic sequence from all the cells at once

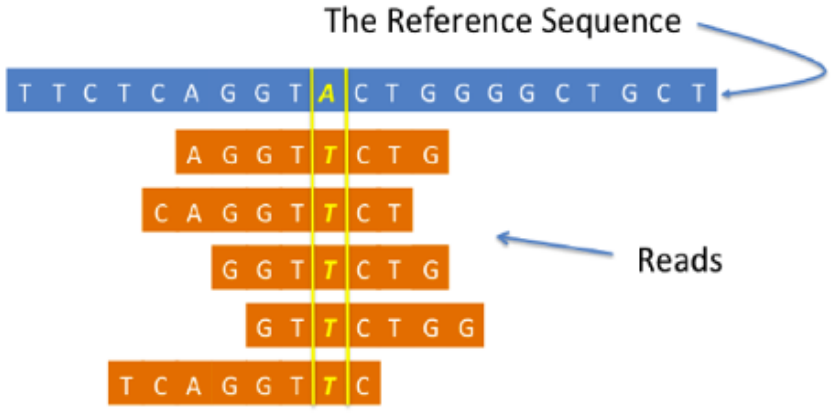
# Genetic Sequencing


- State-of-the-art genetic sequencing machines cannot handle the complete genetic sequence from all the cells at once
- The genetic sequence from each cell is broken into smaller chunks called reads. Typically, these reads are sampled from the original sequence at uniformly randomly chosen locations

# Completing the Jigsaw Puzzle

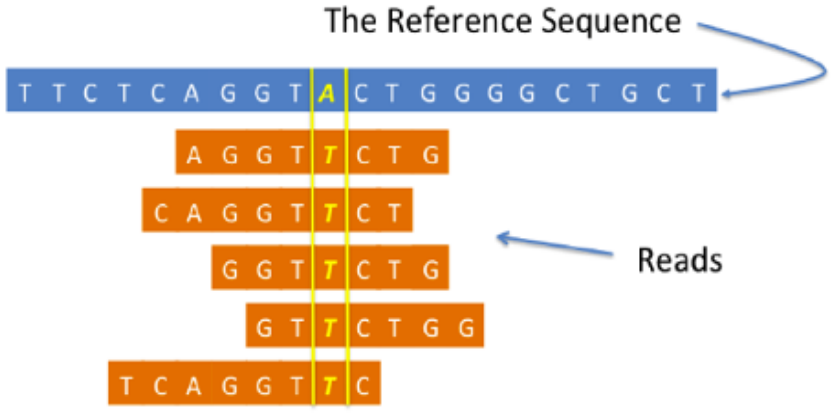


# Completing the Jigsaw Puzzle



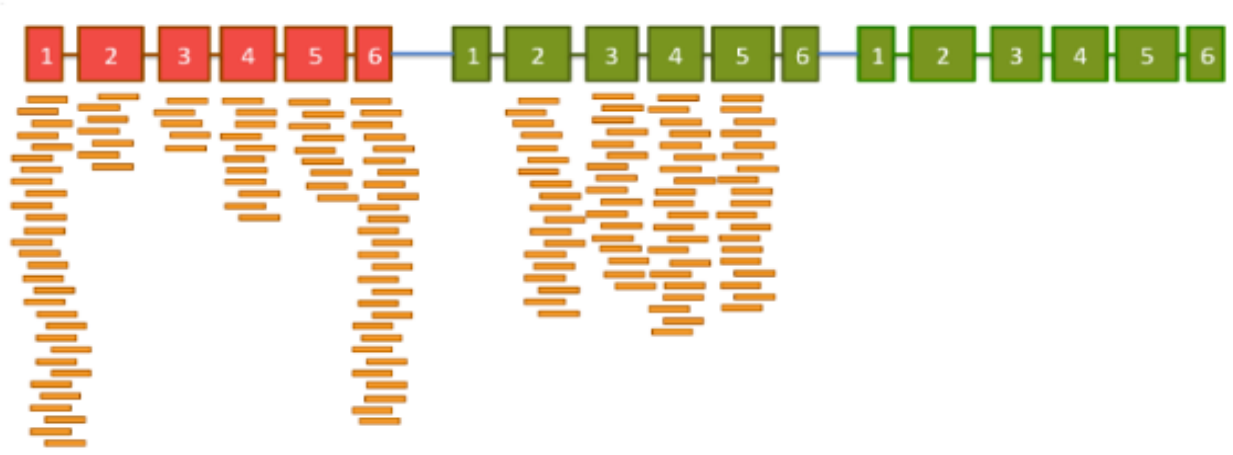

- Reads from a colour-blind person's genetic sequence are aligned with a reference genetic sequence of a healthy individual

# Completing the Jigsaw Puzzle

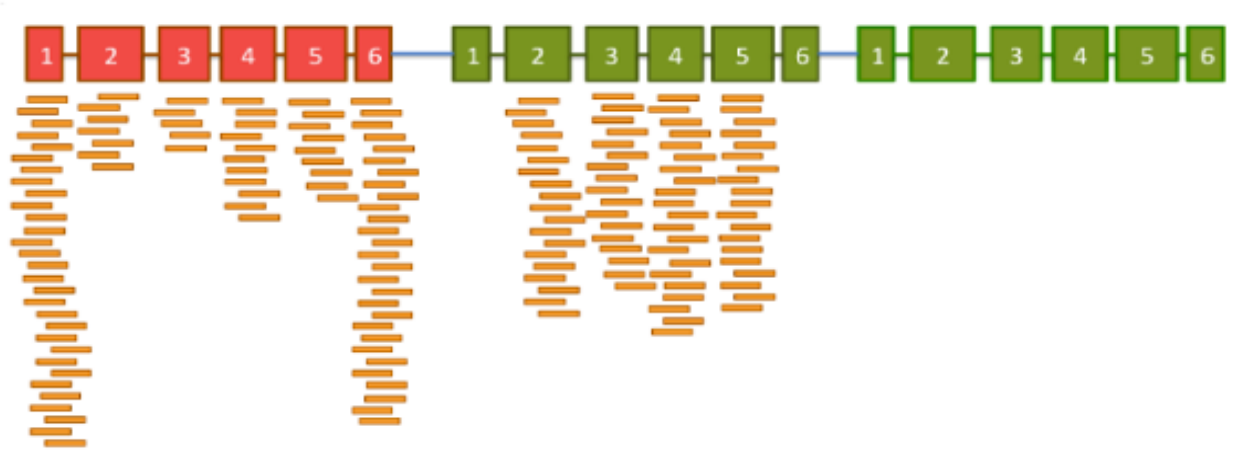



- Reads from a colour-blind person's genetic sequence are aligned with a reference genetic sequence of a healthy individual
- Small mismatches in alignment are allowed in practice

# Accumulating the Read Alignment Counts

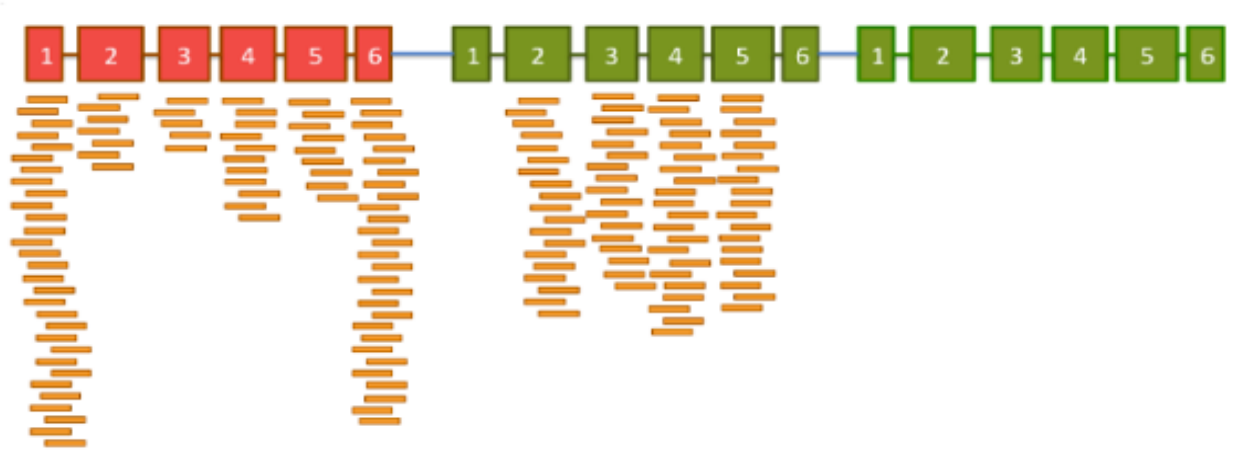


# Accumulating the Read Alignment Counts




- For each exon, the number of the reads whose genetic sequence finds a match within the genetic sequence of the exon is noted

# Accumulating the Read Alignment Counts




- For each exon, the number of the reads whose genetic sequence finds a match within the genetic sequence of the exon is noted
- The most likely configuration causing colour blindness can be identified probabilistically

# Identifying the Most Likely Configuration

- Assume that the total number of reads is $n$

# Identifying the Most Likely Configuration

- Assume that the total number of reads is $n$
- Six red exons and six green exons - 12 exons in all

# Identifying the Most Likely Configuration

- Assume that the total number of reads is $n$
- Six red exons and six green exons - 12 exons in all
- The number of read matchings for each of the above 12 exons is a random variable. Let these random variables be denoted $X_1, \ldots, X_{12}$

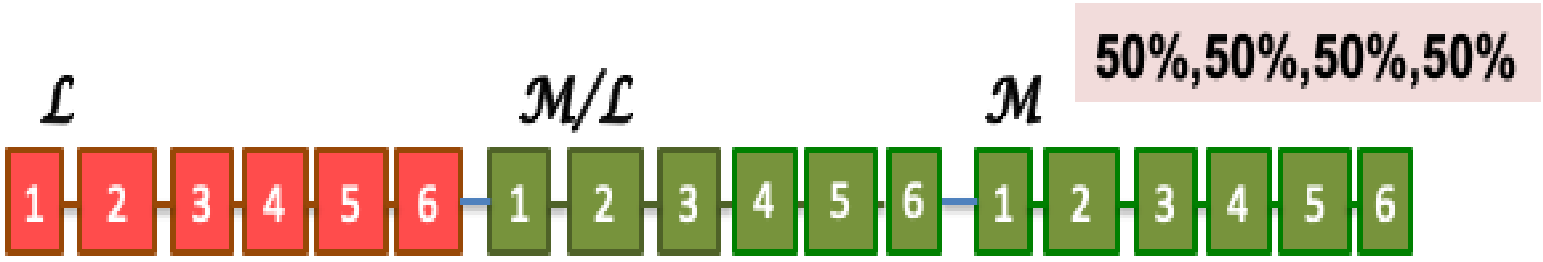
# Identifying the Most Likely Configuration

- Assume that the total number of reads is $n$
- Six red exons and six green exons - 12 exons in all
- The number of read matchings for each of the above 12 exons is a random variable. Let these random variables be denoted $X_1, \ldots, X_{12}$
- From the read matching exercise, we get the values of these random variables. Let these values be denoted $k_1, \ldots, k_{12}$; note that $\sum_{i=1}^{12} k_i = n$

# Identifying the Most Likely Configuration

- Let the lengths of the 12 exons in any given configuration be $\ell_1, \ldots, \ell_{12}$
- Let $L = \sum_{i=1}^{12} \ell_i$

# Configuration 1 ($C_1$)

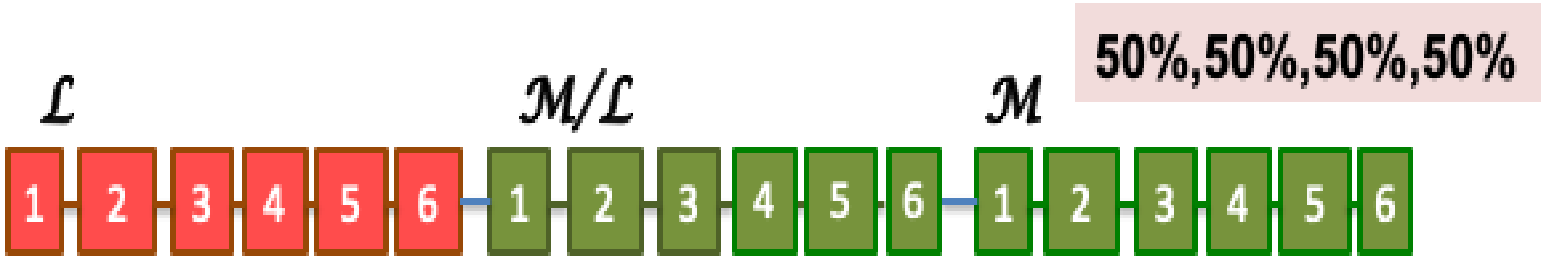



- If the reads are sampled uniformly randomly from the genetic sequence of a colour blind person, then[6]

$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_1) =$$

[6] We use here the fact that the length of a red exon is equal to that of the corresponding green exon. Thus, for e.g., exon 2 of red and green have the same length.

# Configuration 1 ($C_1$)



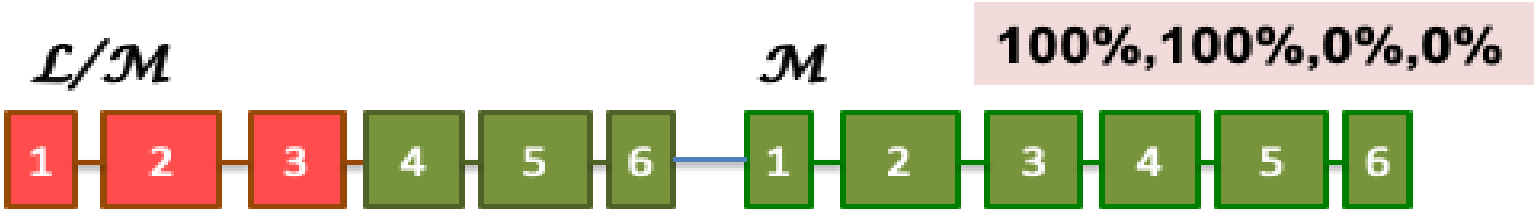

- If the reads are sampled uniformly randomly from the genetic sequence of a colour blind person, then[6]

$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_1) = \frac{n!}{k_1! \cdots k_{12}!} \left(\frac{\ell_1}{L}\right)^{k_1} \cdots \left(\frac{\ell_{12}}{L}\right)^{k_{12}}$$

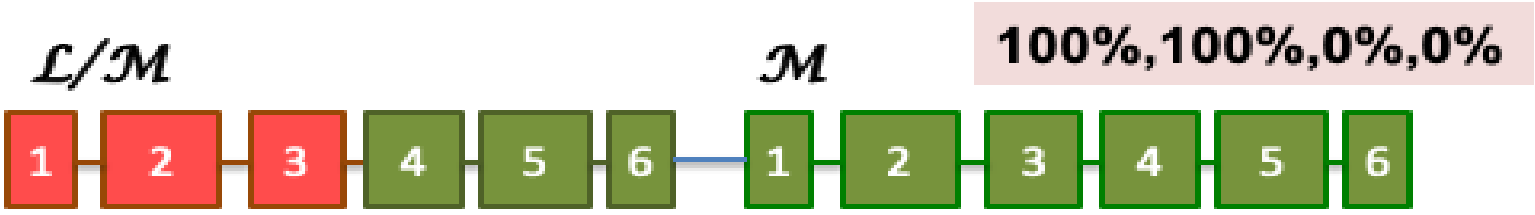
A multinomial distribution!

[6] We use here the fact that the length of a red exon is equal to that of the corresponding green exon. Thus, for e.g., exon 2 of red and green have the same length.

# Configuration 2 ($C_2$)

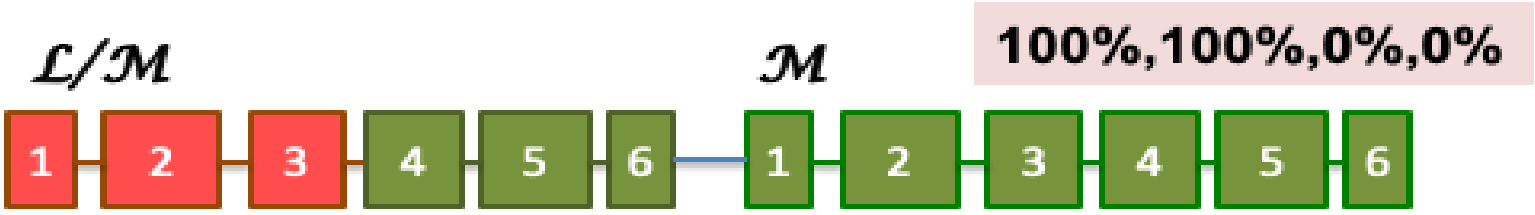


# Configuration 2 ($C_2$)



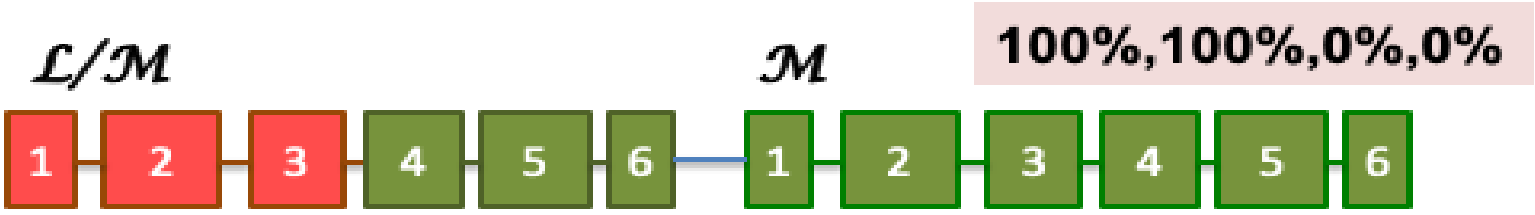

- Here, $X_4 = 0 = X_5 = X_6 = k_4 = k_5 = k_6 = \ell_4 = \ell_5 = \ell_6$

# Configuration 2 ($C_2$)




- Here, $X_4 = 0 = X_5 = X_6 = k_4 = k_5 = k_6 = \ell_4 = \ell_5 = \ell_6$
- $L = \ell_1 + \ell_2 + \ell_3 + \ell_7 + \cdots + \ell_{12}$

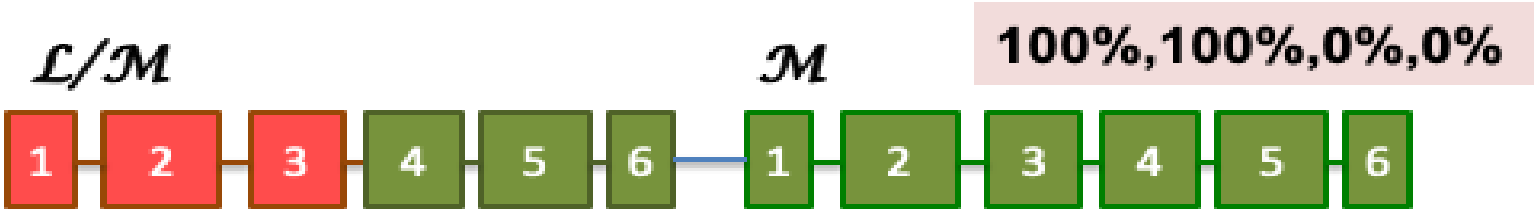
# Configuration 2 ($C_2$)




- Here, $X_4 = 0 = X_5 = X_6 = k_4 = k_5 = k_6 = \ell_4 = \ell_5 = \ell_6$
- $L = \ell_1 + \ell_2 + \ell_3 + \ell_7 + \cdots + \ell_{12}$
- If the reads are sampled uniformly randomly from the genetic sequence of a colour blind person, then

$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_2) =$$

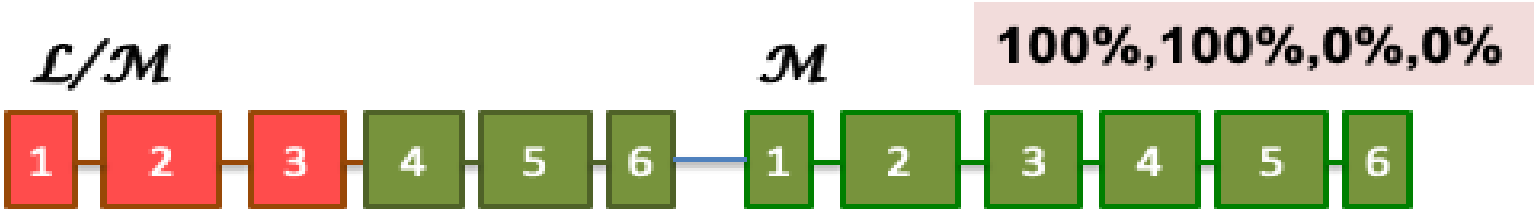
# Configuration 2 ($C_2$)




- Here, $X_4 = 0 = X_5 = X_6 = k_4 = k_5 = k_6 = \ell_4 = \ell_5 = \ell_6$
- $L = \ell_1 + \ell_2 + \ell_3 + \ell_7 + \cdots + \ell_{12}$
- If the reads are sampled uniformly randomly from the genetic sequence of a colour blind person, then

$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_2) = \frac{n!}{k_1! \cdots k_{12}!} \left(\frac{\ell_1}{L}\right)^{k_1} \cdots \left(\frac{\ell_{12}}{L}\right)^{k_{12}}$$

# Configuration 2 ($C_2$)




- Here, $X_4 = 0 = X_5 = X_6 = k_4 = k_5 = k_6 = \ell_4 = \ell_5 = \ell_6$
- $L = \ell_1 + \ell_2 + \ell_3 + \ell_7 + \cdots + \ell_{12}$
- If the reads are sampled uniformly randomly from the genetic sequence of a colour blind person, then

$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_2) = \frac{n!}{k_1! \cdots k_{12}!} \left(\frac{\ell_1}{L}\right)^{k_1} \cdots \left(\frac{\ell_{12}}{L}\right)^{k_{12}}$$

- Similarly, the probabilities can be computed for the remaining two exon configurations (left as exercise)

# Most Likely Configuration

- Compare the values

$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_1),$$
$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_2),$$
$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_3),$$
$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_4).$$

# Most Likely Configuration

- Compare the values

$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_1),$$
$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_2),$$
$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_3),$$
$$P(X_1 = k_1, \ldots, X_{12} = k_{12} \mid C_4).$$

- The most likely configuration is the one for which the value of the corresponding probability term is the largest! Ties can be resolved according to a pre-assigned rule!

# Probability and Covid-19 Spread Modelling

# The Covid-19 Epidemic

- An ongoing epidemic that began in Dec 2019. The first case in India was reported on Jan 30, 2020

[7] As taken from `www.covid19india.org`.

# The Covid-19 Epidemic

- An ongoing epidemic that began in Dec 2019. The first case in India was reported on Jan 30, 2020

- As of Sep 27, 2020, the number of cases is 60,50,975[7]

[7] As taken from `www.covid19india.org`.

# The Covid-19 Epidemic

- An ongoing epidemic that began in Dec 2019. The first case in India was reported on Jan 30, 2020
- As of Sep 27, 2020, the number of cases is 60,50,975[7]
- Probabilistic models – mathematical models to predict the number of cases in the future. The predicted numbers must be practically comparable with the actual numbers

[7] As taken from `www.covid19india.org`.

# Models

- Use the observed data to fit curves via regression and estimate parameters of the fitted curve

# Models

- Use the observed data to fit curves via regression and estimate parameters of the fitted curve

- Model the physical dynamics of epidemic spread at a macroscopic level – compartmentalise the population into susceptible, exposed, infected, recovered (SEIR) groups

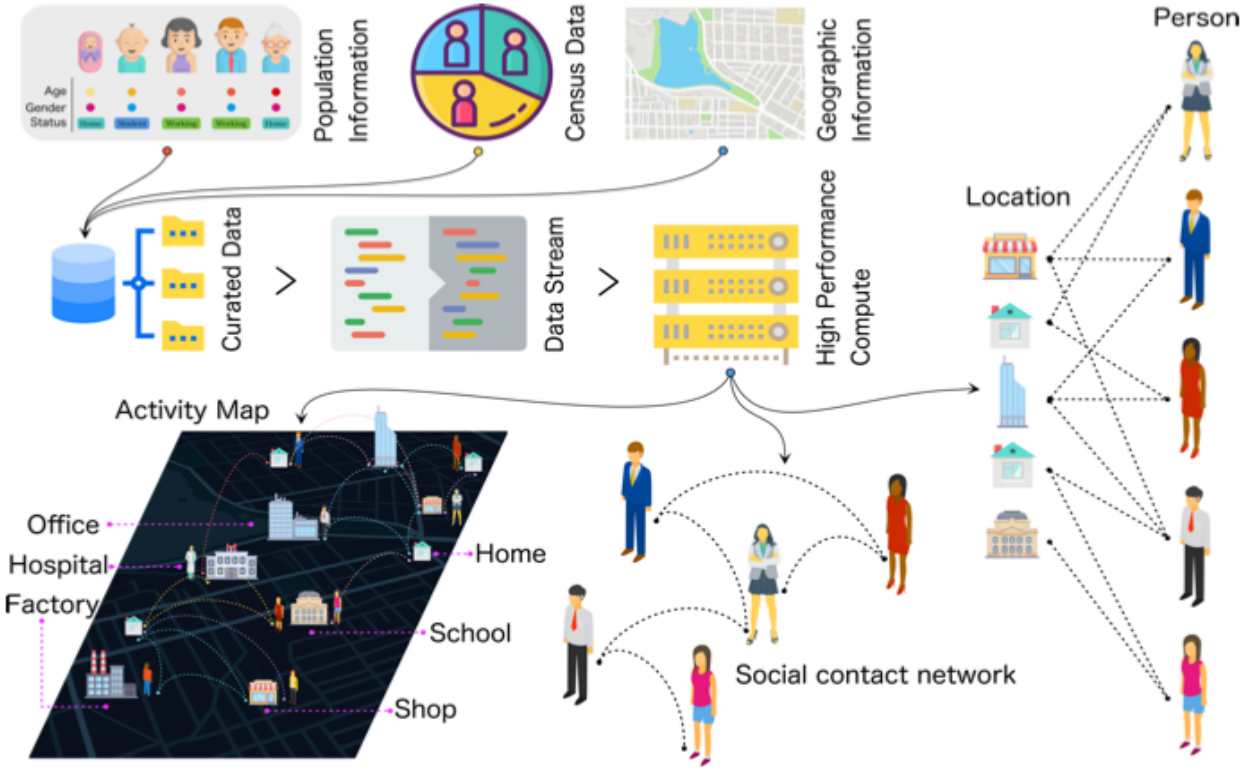
# Models

- Use the observed data to fit curves via regression and estimate parameters of the fitted curve

- Model the physical dynamics of epidemic spread at a macroscopic level – compartmentalise the population into susceptible, exposed, infected, recovered (SEIR) groups

- **Agent-based models** – microscopic models which take into account the population distribution based on census data, the distribution of households in each ward, age distribution in every household, etc

# Agent-based Models




A schematic representation of an agent-based model

# A City-Scale Simulator

- We shall look at a specific city-scale simulator[8] built by a team from IISc Bangalore and TIFR Mumbai

[8] https://cni-iisc.github.io/epidemic-simulator/

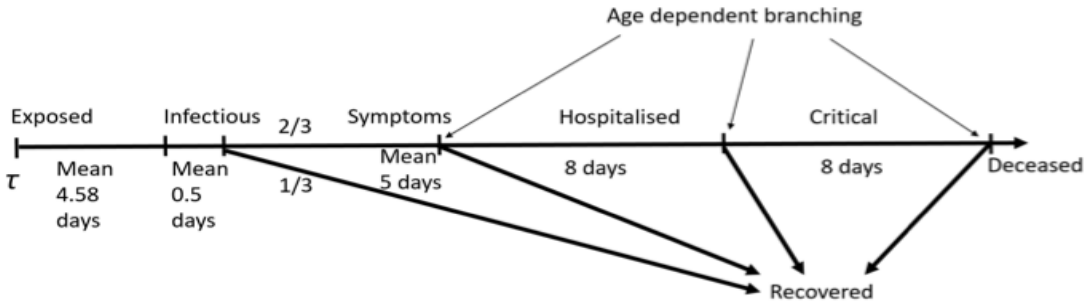
# A City-Scale Simulator

- We shall look at a specific city-scale simulator[8] built by a team from IISc Bangalore and TIFR Mumbai

- This simulator is based on a synthetic city created through simulation (respecting the specifics of the actual city). The disease progression in this city is modelled probabilistically

[8] https://cni-iisc.github.io/epidemic-simulator/

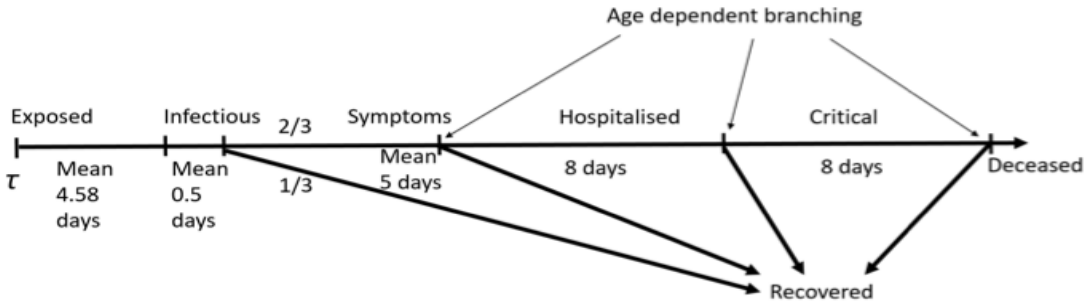
# Disease Progression[9] [10]




[9] K. Prem, Y. Liu, T. W. Russell, A. J. Kucharski, R. M. Eggo, N. Davies, S. Flasche, S. Clifford, C. A. Pearson, J. D. Munday et al., "The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study," The Lancet Public Health, 2020.

[10] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg et al., "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand," Tech. Report, 2020.

# Disease Progression[9] [10]

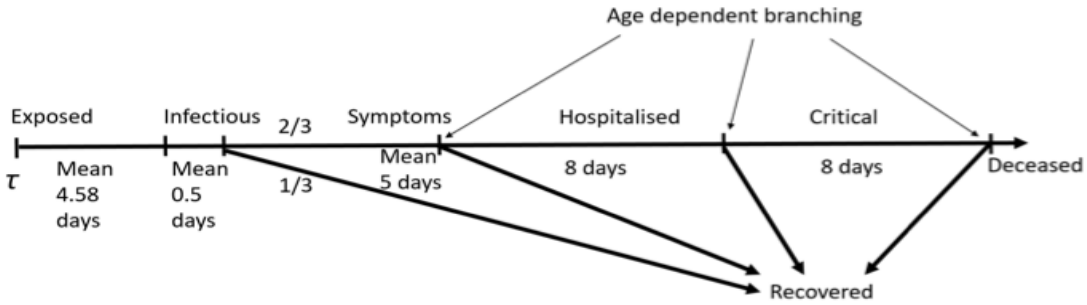



- The incubation period is modelled to have Gamma distribution with shape parameter 2 and scale parameter 2.29 (thus, mean = 4.58)

[9] K. Prem, Y. Liu, T. W. Russell, A. J. Kucharski, R. M. Eggo, N. Davies, S. Flasche, S. Clifford, C. A. Pearson, J. D. Munday et al., "The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study," The Lancet Public Health, 2020.

[10] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg et al., "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand," Tech. Report, 2020.

# Disease Progression[9] [10]

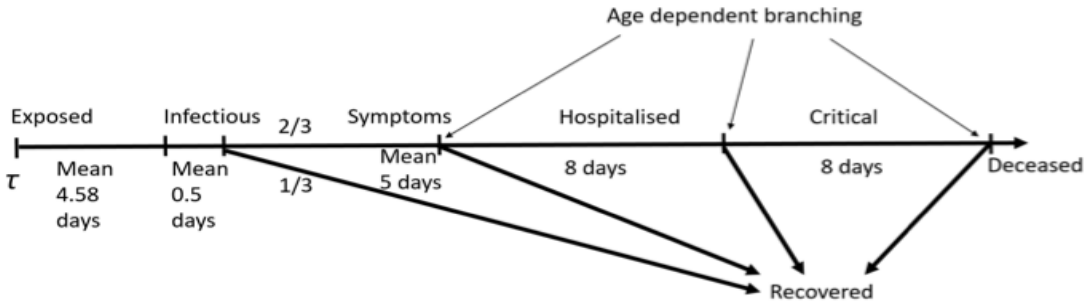



- The incubation period is modelled to have Gamma distribution with shape parameter 2 and scale parameter 2.29 (thus, mean = 4.58)
- Individuals are infectious for an exponentially distributed period of 0.5 days

[9] K. Prem, Y. Liu, T. W. Russell, A. J. Kucharski, R. M. Eggo, N. Davies, S. Flasche, S. Clifford, C. A. Pearson, J. D. Munday et al., "The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study," The Lancet Public Health, 2020.

[10] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg et al., "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand," Tech. Report, 2020.

# Disease Progression[9] [10]



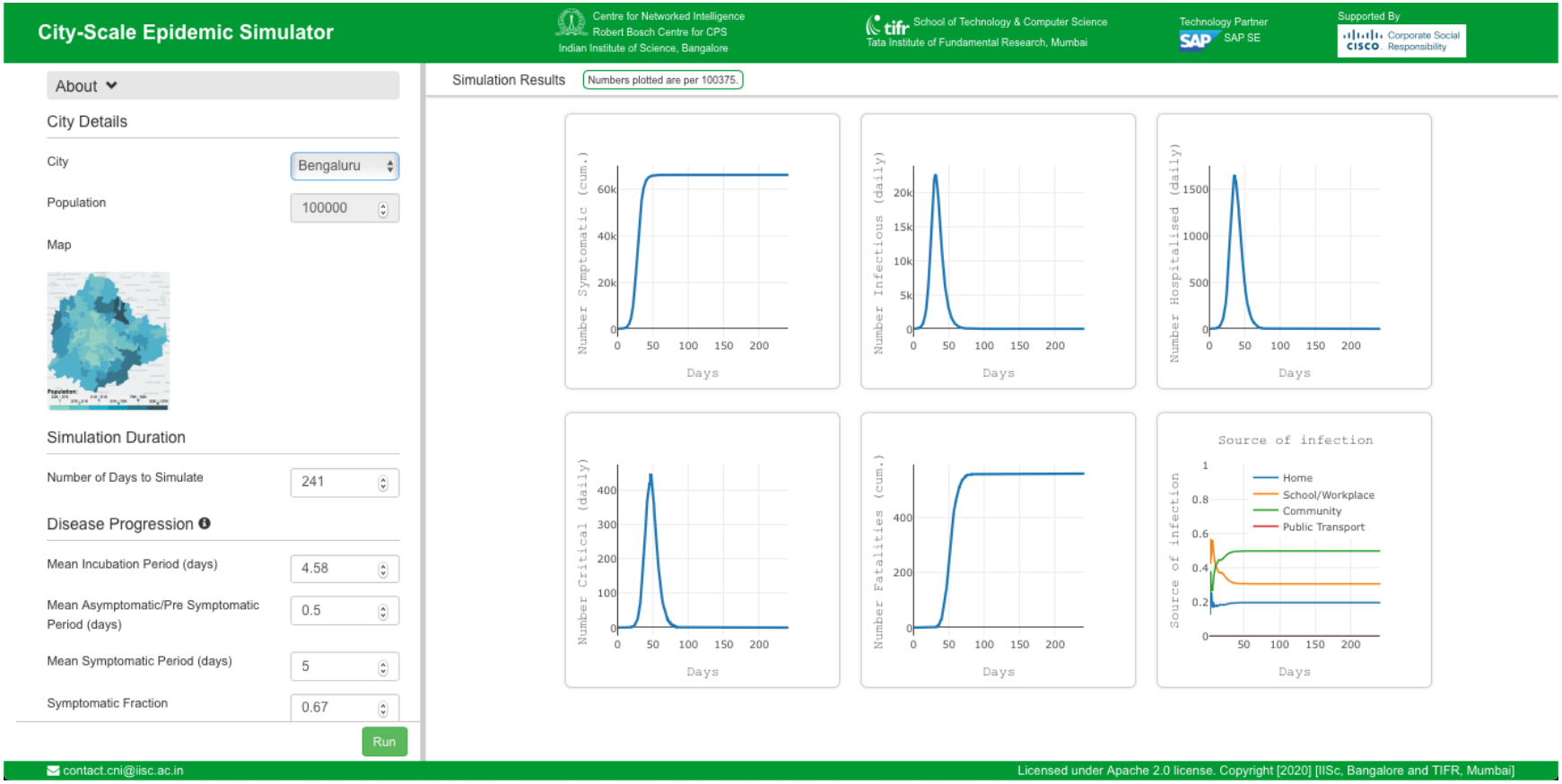

- The incubation period is modelled to have Gamma distribution with shape parameter 2 and scale parameter 2.29 (thus, mean = 4.58)
- Individuals are infectious for an exponentially distributed period of 0.5 days
- Suppose 2/3rd of the infectious patients end up showing symptoms. The mean time to show symptoms is **exponentially distributed** with mean 5 days

---

[9] K. Prem, Y. Liu, T. W. Russell, A. J. Kucharski, R. M. Eggo, N. Davies, S. Flasche, S. Clifford, C. A. Pearson, J. D. Munday et al., "The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study," The Lancet Public Health, 2020.

[10] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg et al., "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand," Tech. Report, 2020.

# The City Scale Simulator




The full simulator is accessible at

`https://cni-iisc.github.io/epidemic-simulator/`

# Caveats While Using the Simulator

- The simulator is mainly to help understand the importance of implementing NPIs in containing the spread of the epidemic. It should not be used for medical diagnostic or treatment purposes

- The numbers reflected by the simulator may not be close to what is seen in reality (as can be seen in the image on the previous slide). This is because the parameters used in the simulation are not updated based on the observed data

# Bibliography

- The images of Bernoulli, de Moivre, Borel, Lebesgue and Kolmogorov are from their respective Wikipedia pages
- The images of Arun Sripati and Carl Olson are from their respective institute websites[11] [12]
- The image on slide 12 was created on `draw.io`
- The images on slide 18 were generated on my laptop

[11] `http://www.cns.iisc.ac.in/home/people/sp-arun/`
[12] `http://www.cnbc.cmu.edu/colson/`

# Bibliography

- The images used in the slides on visual neuroscience are from the lecture notes and the dataset made available on the webpage[13] of the Data Analytics course taught at IISc
- The images used in the slides on colour blindness are from the lecture slides made available on the webpage[14] of the Data Analytics course taught at IISc
- The images used in the slides on covid-19 spread modelling are from the report available at `https://arxiv.org/pdf/2008.04849.pdf`

[13] `https://ece.iisc.ac.in/~rajeshs/E0259/02_data_visual_neuroscience.htm`
[14] `https://ece.iisc.ac.in/~rajeshs/E0259/05_data_colour_blindness.htm`

Any mistake anywhere is a result of my (un)mindfulness at the time of preparing these slides..

Any mistake anywhere is a result of my (un)mindfulness at the time of preparing these slides..

Thank You!

https://www.karthikpn.com