

# Medi**chal**



---

Using medical data  
to create  
challenges

# Summary

## Introduction

### I. Medical data: issue & answers

#### A. Issue: Data privacy

#### B. Answers

### II. Challenges

#### A. Survival analysis

#### B. Binary classification

## Conclusion

# Introduction

- ▷ Create challenges
  - Codalab
  - University courses
- ▷ Domain: Healthcare
  - Using medical records



# Medical data: issues & answers

## I. Medical data: issue & answer

### A. Issue: Data privacy

Data privacy is the relationship between the collection of data and its dissemination.

#### Example:

Name	Age	Gender	Localisation	HIV
Martine	23	M	Saclay	1

## I. Medical data: issue & answer

### B. Answers

- ▷ De-identification
  - Replace name by an identifier

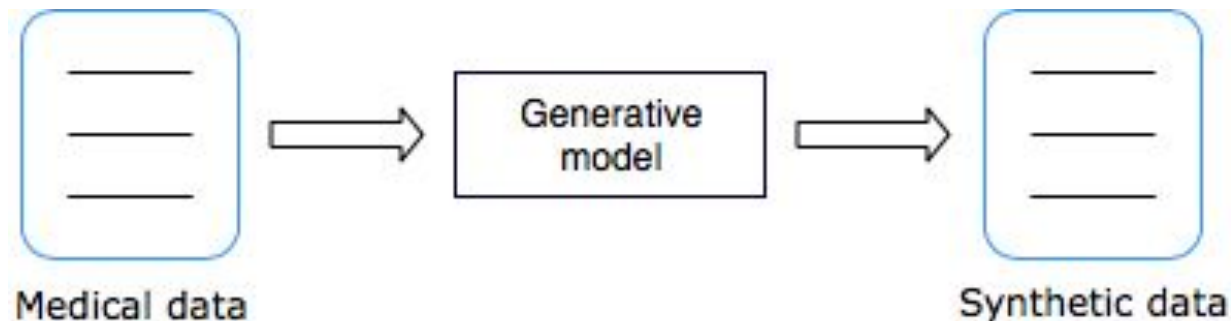
ID	Age	Gender	Localisation	HIV
140203	23	M	Saclay	1

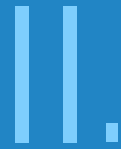
Major default : Other variables may contain information to directly retrieve the patient.

## I. Medical data: issue & answer

### B. Answers

- ▷ Data generation
  - **Learn the data distribution** using a **generative model** (e.g. GAN, RandomForest) that ensures privacy
  - **Generate** samples according to this distribution





# Challenges



## II. Challenges

### Context

These challenges are for the moment aimed at students of **RPI**, Rensselaer Polytechnic Institute as part of a statistical course they follow.

The objective of the challenge is

- for the participant to **acquire knowledge** around the topic of the challenge.
- for the teacher to have a tool to **assess students' performance**.

## II. Challenges

### A. Binary classification

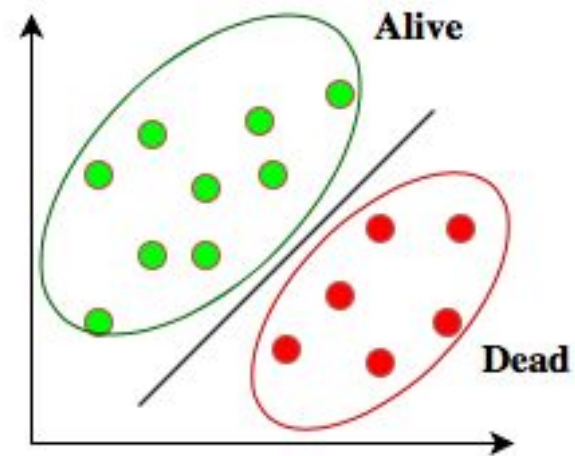
#### Objective :

Classify data where the outcome variable is the **class  $p$**  (0 or 1).

#### Example:

Predict life or death of hospitalized patients.

**class  $p = 0$**     if alive  
                  1    if dead



## II. Challenges

### B. Survival analysis

#### Objective :

Analyze data where the outcome variable is the **time**  $t$  until the occurrence of an **event**  $e$ .

#### Example:

What is the probability that a patient dies 10 days after hospitalization?

**time**  $t$  = 10 days

**event**  $e$  = death

## II. Challenges

### B. Survival analysis

#### Objective :

Analyze data where the outcome variable is the **time**  $t$  until the occurrence of an **event**  $e$ .

#### Example:

What is the probability that a patient dies 10 days after its hospitalization?

**time**  $t$  = 10 days

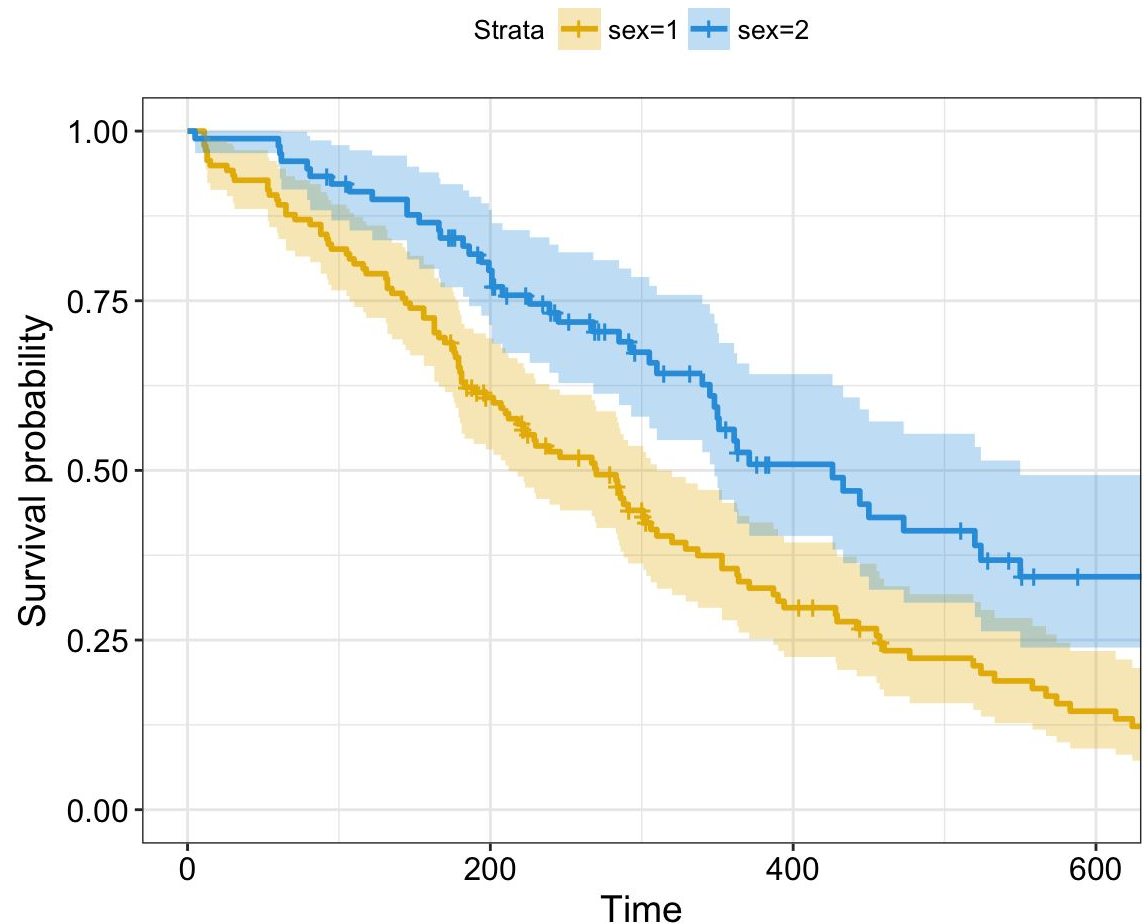
**event**  $e$  = death

## II. Challenges

### B. Survival analysis

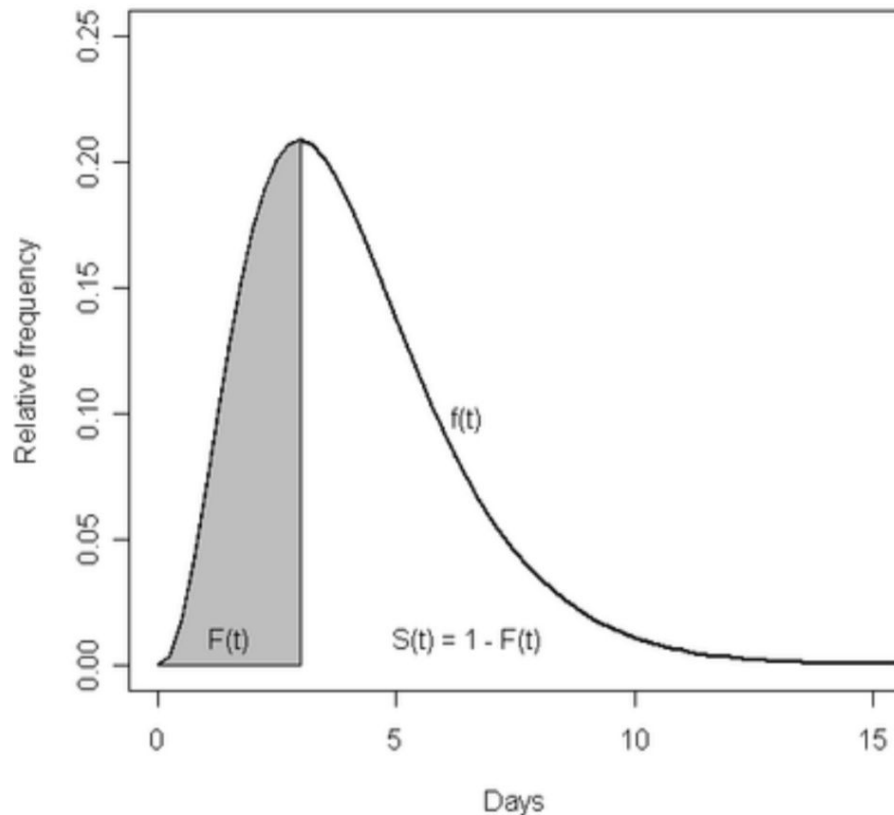
#### Example:

What is the probability that an individual dies of a cancer in function of the time and according to gender?



## II. Challenges

### B. Survival analysis



$f(t)$  = distribution  
function

$F(t)$  = cumulative  
distribution function

$S(t)$  = survival  
function



# Presentation of the challenge

## Survival Analysis - Jupyter Notebook

# Conclusion: Objectives

- ▷ Data generation
  - Good quality
  - Ensure privacy (theoretically and practically)
- ▷ Challenges
  - Build fully functional challenges
  - Auto-grading system for the survival challenge
  - ...



Thanks!

**Any questions?**