

# When machines talk. A data challenge by SNCB-NMBS

Engineering Department, B.TC-451

contact: [georges.tod@sncb.be](mailto:georges.tod@sncb.be)  
October 2024



# Outline

- 1 Abstract
- 2 Introduction
- 3 The challenge
  - Real life problem
  - Main columns
  - Extra columns
- 4 Insights



# Outline

## 1 Abstract

## 2 Introduction

## 3 The challenge

- Real life problem
- Main columns
- Extra columns

## 4 Insights



The challenge proposed here, shares some real life data and problem to students. By carefully using statistics/ML methods and carefully handcrafting algorithms, you have the opportunity to contribute to the maintenance processes of *your*<sup>1</sup> railway vehicles. The impact on the long term is expected to be on the reliability of our service - at nationwide scale.

---

<sup>1</sup>as of today, we are still a public company



# Outline

## 1 Abstract

## 2 Introduction

## 3 The challenge

- Real life problem
- Main columns
- Extra columns


## 4 Insights



# from raw sensor data to states

Recent railway vehicles are equipped with sensors on most of their subsystems such as pantographs, traction converters, doors, heating, ventilation and air-conditioning (HVAC), European Train Control System (ETCS)<sup>2</sup>, etc. which report some states via a wired network to a central on-board computer. The latter typically transmits some of these states or combinations of these states as tokens to a cloud service over a cellular network.

---

<sup>2</sup>the ETCS helps drivers to operate the trains safely 

# from states to tokens and events

The tokenized states are often nominated as *information codes* (e.g. door 1 is open) or *fault codes* (e.g. battery temperature is too high) and represent *events* that happen on-board the vehicles.



# machines talk

A loose analogy with human language can be made. Information and fault codes could be seen as the vocabulary of railway vehicles (*or any machine*) and the more there are, the more expressive the vehicle language will be. However, this language is far from being trivial and requires time for humans to learn it.





# Outline

## 1 Abstract

## 2 Introduction

## 3 The challenge

- Real life problem
- Main columns
- Extra columns

## 4 Insights



# A real life problem

On a daily basis, we receive thousands of sequences of events from thousands of vehicles. Since machines tend to degrade, sometimes technical failures appear. In this challenge, we provide a labeled dataset of sequence of events with technical incident types. The main challenges are to,

- find sub sequences of events (scenarios) that seem to be highly associated to some types of incidents
- automatically suggest incident types based on new sequences of events

Both results are useful for railway vehicles maintenance technicians!



# Main columns of the dataset

The sequences-lists are synchronized,

- **incident id**: this is the ID of an incident
- **vehicle sequence**: this list contains a sequence of vehicle IDs that have reported an event. When more than one unique vehicle appears in this list, it means the vehicles were coupled together.
- **events sequence**: this list contains the sequence of events reported by each vehicle in the **vehicle sequence**.
- **seconds to incident sequence**: this list contains the time in seconds to the incident for each event in **events sequence**. Negative numbers are events that happen before the incident and positive numbers represent events that happen after and incident.



## Extra columns of the dataset

Some extra context is provided. It is not mandatory to leverage this data, but interesting to see how you could use it,

- **approx lat, approx lon**: mean event sequence GPS position.
- **train kph sequence**: this list contains the train speed (in kilometers per hour) estimation for each event of **events sequence**.
- **dj dc state sequence**: this list contains the state of the DC switch (boolean) for each event of **events sequence**.
- **dj ac state sequence**: this list contains the state of the AC switch (boolean) for each event of **events sequence**.

Infrastructure power lines in Belgium are typically either 3kV DC or 25kV AC. The state of the switches tell us wheter a given vehicle is in a state that can capture a DC or an AC component. When none are active, the vehicle is powered on batteries only: some communication and light systems can be powered but no traction can be supplied.



# Outline

## 1 Abstract

## 2 Introduction

## 3 The challenge

- Real life problem
- Main columns
- Extra columns

## 4 Insights



# Insights

At SNCB-NMBS, we casted the problem as a classification task using events happening 10min after incident and up to 4 hours before an incident. Our algorithm reaches an F1-score of around 85%. Added value for us does not come only from trying to do better - we are as much interested **in the methods and approaches you imagine to solve the problem.**

