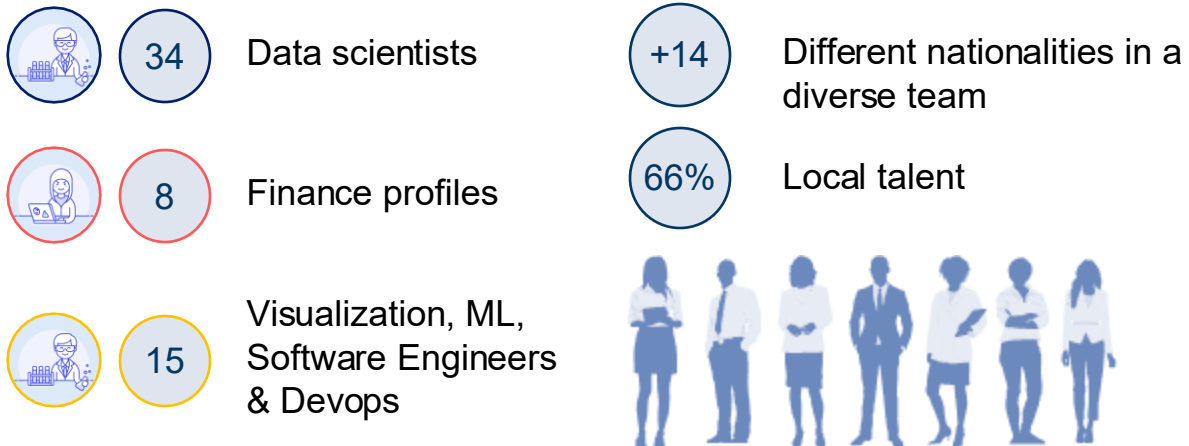




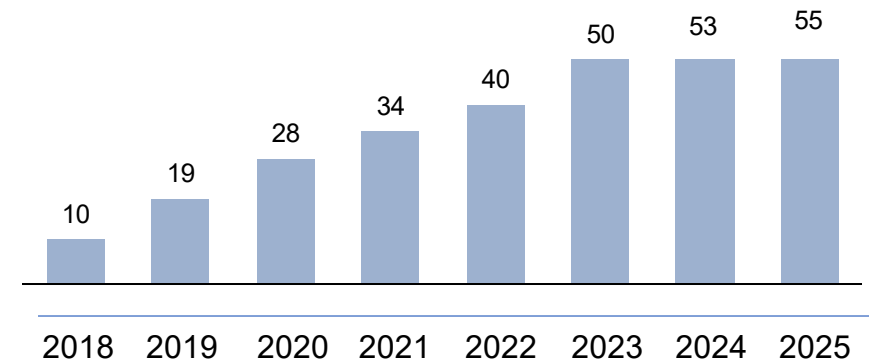
<sup>TH</sup> the last edition  
The **financial challenge** of the year  
**NOVARTISDATATHON**  
online

# The BCN Digital Finance Hub

## The team



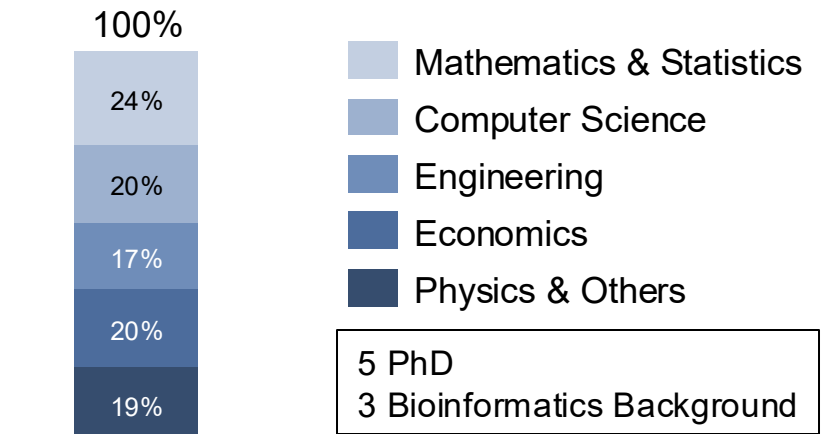
## Evolution of the hub



## Barcelona: European pole for tech and AI talent

- **Tech cluster:** In the last decade, many companies (Amazon, Microsoft, AstraZeneca,...) located their global AI hubs.
- **Forefront centers & infrastructures:** Bcn supercomputing center, Quantum Computer, Synchrotron,...
- Proficiency **local universities** promoting Data Science, Mathematics or Statistics.
- **Attractive city** for international talent to reallocate.

## Background

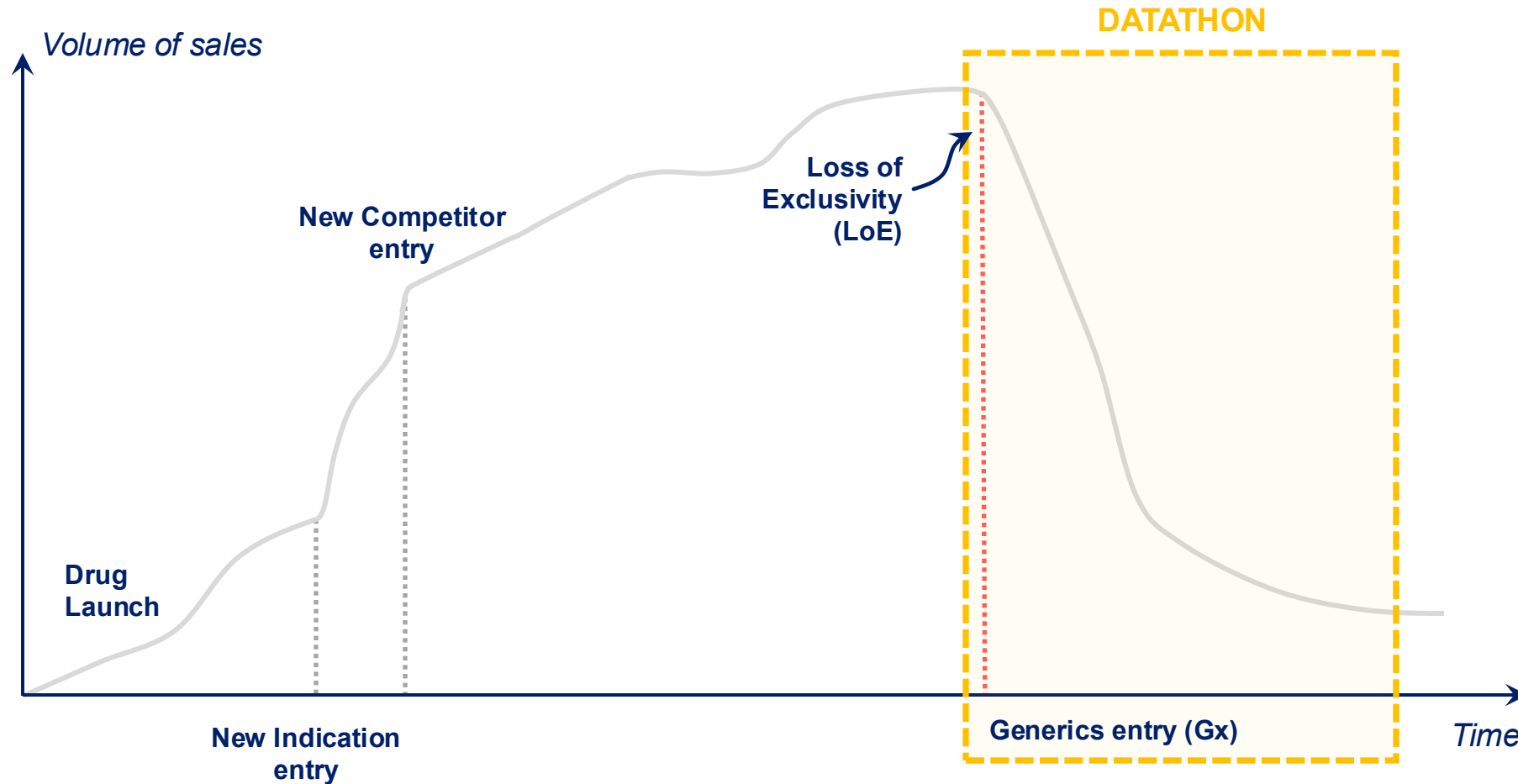




# **Datathon 8th edition**

**Generic impact**

# Lifecycle of a drug



# Generic erosion

## 1. Formula

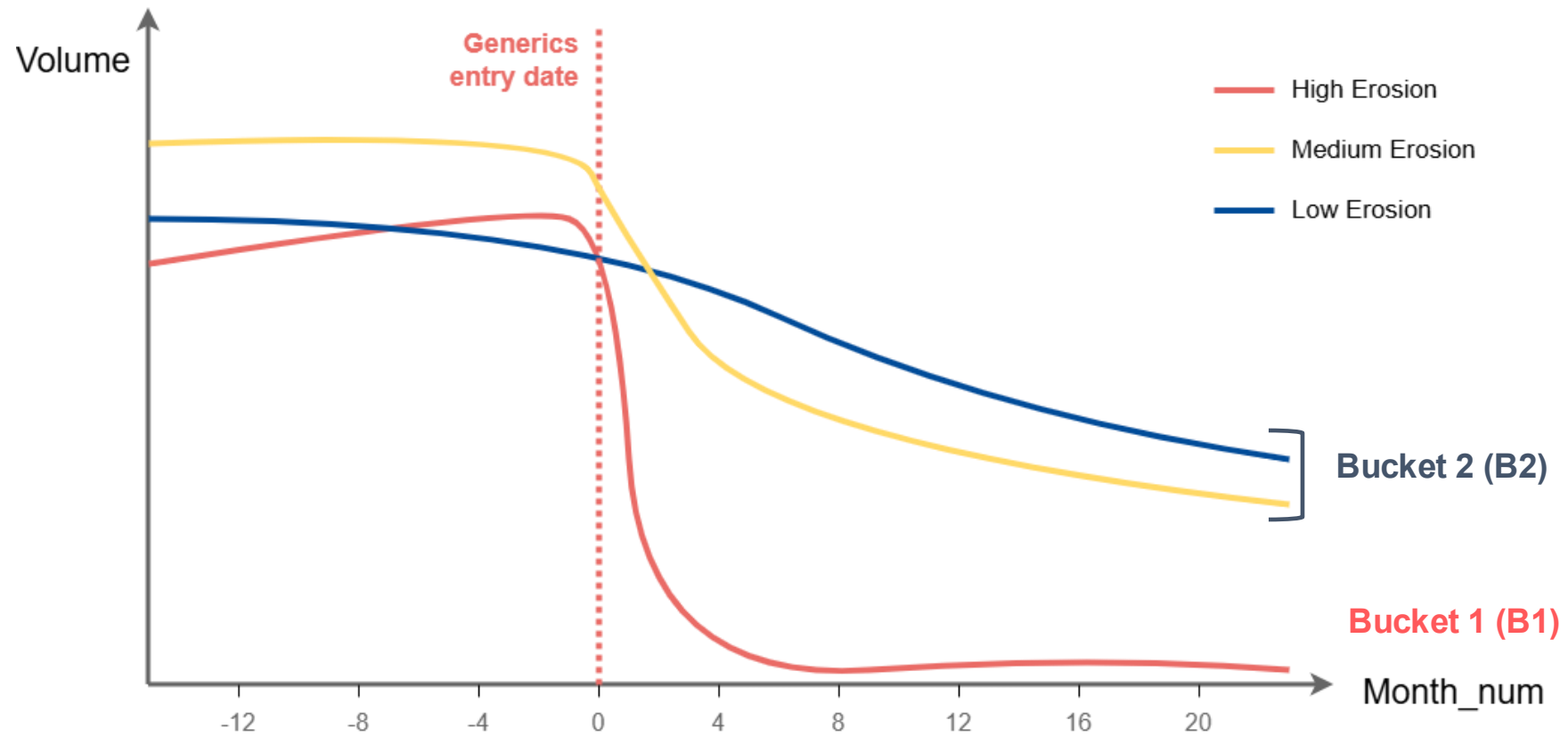
**Mean Generic Erosion** is defined as the mean of the normalized volumes after the generic entry considering a 24-month horizon. Volumes after generic entry are normalized by the average monthly volume of the last 12 months before the generic entry.

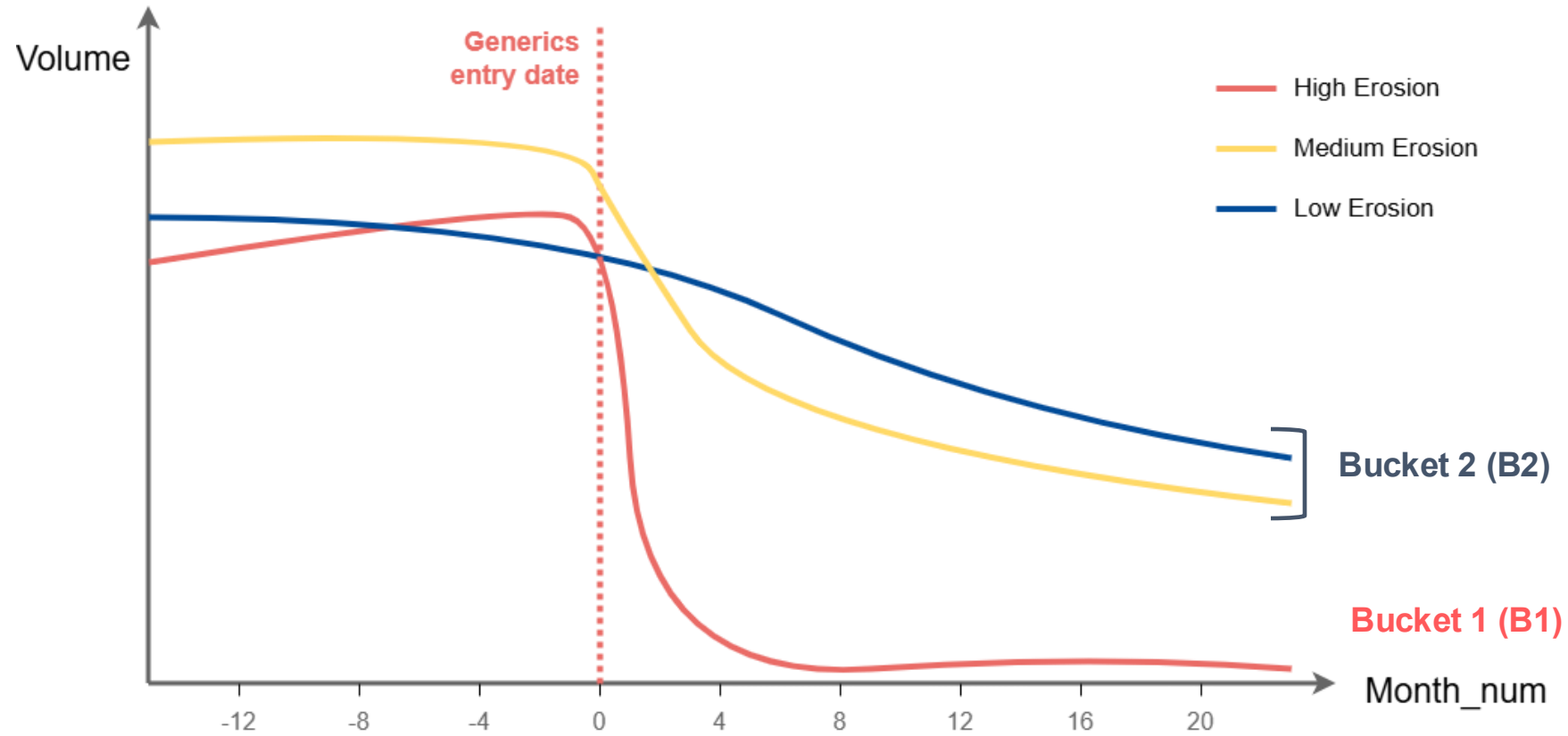
$$\text{Mean Generic Erosion} = \frac{\sum_{i=0}^{23} Vol_{norm,i}}{24}$$

$$Vol\ Norm_i = \frac{Vol_i}{Avg_j}$$

$$Avg_j = \frac{\sum_{i=-12}^{-1} Y_{j,i}^{act}}{12}$$







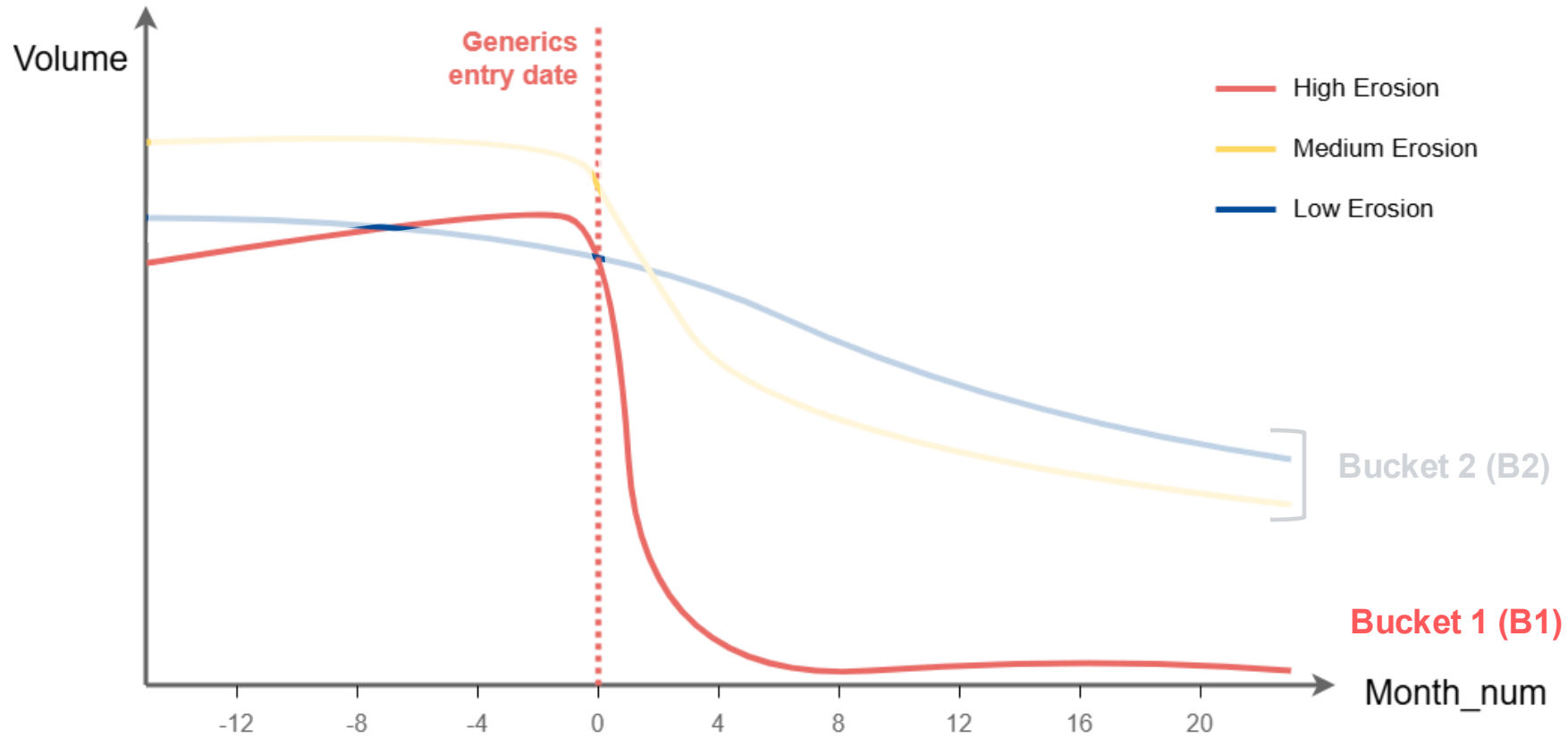
Bucket 2 (B2)

Mean Erosion  $\in (0.25, 1]$

Bucket 1 (B1)

Mean Erosion  $\in [0, 0.25]$





Bucket 2 (B2)

Mean Erosion  $\in (0.25, 1]$

Bucket 1 (B1)

Mean Erosion  $\in [0, 0.25]$

**Datathon Focus!!!**





# Datathon Challenge

## 1. Data Science Challenge

Participants are asked to **forecast the volume erosion following generic entry** over a **24-month horizon** from the generic entry date. Forecasting should be performed at two distinct time points:

- **Scenario 1: Right after the generic entry date:** No actuals post generic entry. *Forecast from month 0 to month 23.*
- **Scenario 2: 6 months after the generic entry date:** *Forecast from month 6 to month 23.*

## 2. Business Challenge

All the teams that present in front of the Jury will be asked to provide a **deep exploratory analysis** on the preprocess carried out with **focus on the high-erosion cases**. We encourage the participants to use visualization tools.



# Datathon Winner Selection Process

The winner selection will take place in **two evaluation phases**:

**Phase 1 – Model Evaluation:** Participants must submit volume predictions for the entire test dataset, which includes both Scenario 1 and Scenario 2 cases.

- **Phase 1-a: Scenario 1 Evaluation**

- All teams will be evaluated on Scenario 1 prediction accuracy.
- The top 10 teams with the lowest prediction error will advance to Phase 1-b.

- **Phase 1-b: Scenario 2 Evaluation**

- Only the top 10 teams from Scenario 1 will be evaluated on Scenario 2 prediction accuracy.
- Among these, the top 5 teams with the lowest Scenario 2 prediction error will advance to the Final Phase.

**Phase 2 – Jury Evaluation**

- The final 5 teams will present their methodology, insights, and conclusions to a Jury composed of both technical and business experts.
- After reviewing the presentations, the Jury will select the top 3 winning teams.



# Challenge: Data Provided (1/2)

**Target Variable:** Monthly volume for 2,293 country–brand combinations that experienced generic entry.

- **Training Set** (1,953 observations):
  - Includes up to 24 months of volume data *before* generic entry
  - And up to 24 months *after* generic entry
- **Test Set** (340 observations):
  - Two forecasting scenarios are evaluated:
    - Scenario 1 (~2/3 of test set; 228 observations): Forecast required from **Month 0 to Month 23**
    - Scenario 2 (~1/3 of test set; 112 observations): Forecast required from **Month 6 to Month 23**

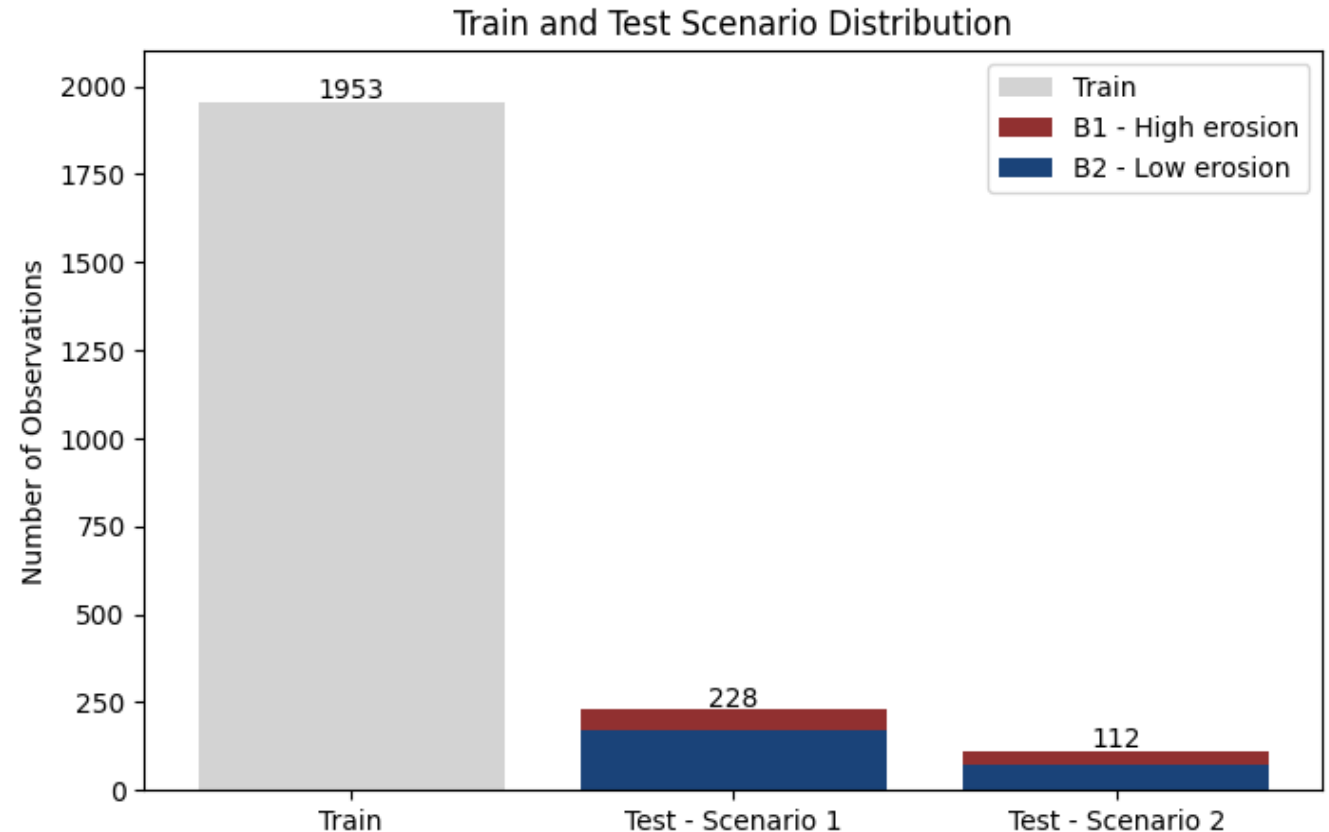
# Challenge: Data Provided (2/2)

## Erosion Buckets in Test Set:

The 340 test observations are distributed in the following way across the scenarios and erosion levels:

- 0–0.25 (Bucket 1, B1: High erosion)
- 0.25–1 (Bucket 2, B2: Low erosion)

This structure holds *in both forecasting scenarios (scenario 1 and scenario 2)*.



# Data available: Volume

**df\_volume.csv**: includes information regarding the volume of sales before and after the generic entry:

- **country**: country name
- **brand\_name**: brand name
- **month**: name of the month
- **months\_postgx**: number of month after generic entry. Month\_postgx equal to 0 denotes the generics entry. Negative values refer to months before the generics entry (eg. Month\_postgx = -3 denotes three months before the generics entry)
- **volume**: number of drugs sold

country	brand_name	month	months_postgx	volume
COUNTRY_B6AE	BRAND_1C1E	Jul	-24	272594.39
COUNTRY_B6AE	BRAND_1C1E	Aug	-23	351859.31
COUNTRY_B6AE	BRAND_1C1E	Sep	-22	447953.48
COUNTRY_B6AE	BRAND_1C1E	Oct	-21	411543.29
COUNTRY_B6AE	BRAND_1C1E	Nov	-20	774594.45
COUNTRY_B6AE	BRAND_1C1E	Dec	-19	442279.18
COUNTRY_B6AE	BRAND_1C1E	Jan	-18	485069.49
COUNTRY_B6AE	BRAND_1C1E	Feb	-17	549902.7

# Data available: Generics features

**df\_generics.csv**: includes information about the country, the drug and the number of generics existing for that specific brand in that country:

- **country**: country name
- **brand\_name**: brand name
- **months\_postgx**: number of months after generic entry. Month\_postgx equal to 0 denotes the generics entry
- **n\_gxs**: number of generics. Note that the number of generics might vary along time

country	brand_name	months_postgx	n_gxs
COUNTRY_B6AE	BRAND_DF2E	0	0.0
COUNTRY_B6AE	BRAND_DF2E	1	0.0
COUNTRY_B6AE	BRAND_DF2E	2	1.0
COUNTRY_B6AE	BRAND_DF2E	3	2.0
COUNTRY_B6AE	BRAND_DF2E	4	2.0
COUNTRY_B6AE	BRAND_DF2E	5	2.0
COUNTRY_B6AE	BRAND_DF2E	6	2.0
COUNTRY_B6AE	BRAND_DF2E	7	2.0

# Data available: Drug-related features

**df\_medicine\_info.csv**: includes information regarding each drug and administration within a country:

- **country**: country name
- **brand\_name**: brand name
- **ther\_area**: refers to the drugs' therapeutical area
- **hospital\_rate**: percentage of the drug being delivered in hospitals
- **main\_package**: most common format in which the drug is dispensed (eg. PILL)
- **biological**: boolean indicating whether the drug is derived from a living organism (eg. proteins, antibodies, nucleic acids)
- **small\_molecule**: boolean indicating whether the drug is a low molecular weight compound (typically synthesized chemically)

country	brand_name	ther_area	hospital_rate	main_package	biological	small_molecule
COUNTRY_0024	BRAND_1143	Sensory_organs	0.09	EYE DROP	False	True
COUNTRY_0024	BRAND_1865	Muscoskeletal_Rheu..	92.36	INJECTION	False	False
COUNTRY_0024	BRAND_240F	Antineoplastic_and...	36.94	PILL	False	True
COUNTRY_0024	BRAND_2F6C	Antineoplastic_and...	0.01	INJECTION	True	False
COUNTRY_0024	BRAND_3A67	Nervous_system	nan	PILL	False	False
COUNTRY_0024	BRAND_3CB9	Antineoplastic_and...	1.42	PILL	False	True
COUNTRY_0024	BRAND_3E0C	Antineoplastic_and...	47.06	INJECTION	True	False
COUNTRY_0024	BRAND_41B7	Nervous_system	0.02	PILL	False	True

# Metric: Prediction Error (Phase 1-a)

In this first scenario (Scenario 1), participants will provide predictions without knowing **any actual data** after the generic entry date. To compute the prediction error for this phase (Phase 1-a), we will evaluate the difference between the predicted values and the actual volumes in four different ways, weighted as follows:

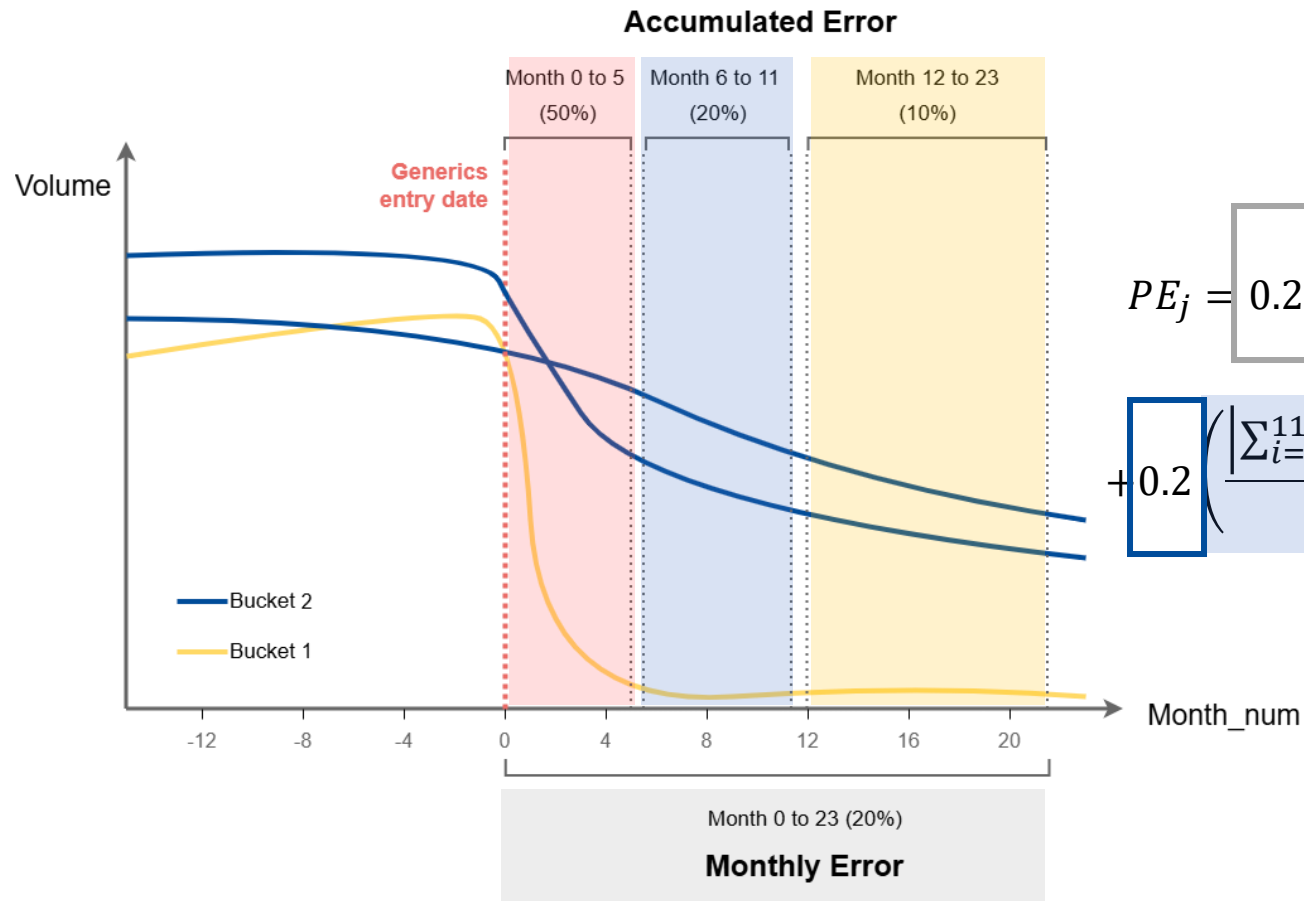
1. Absolute **monthly** error of all 24 months (20%)
2. Absolute **accumulated** error of months 0 to 5 (50%)
3. Absolute **accumulated** error of months 6 to 11 (20%)
4. Absolute **accumulated** error of months 12 to 23 (10%)

All the 4 items will be normalized by the average ( $Avg_j$ ) monthly volume of the last 12 months before the generic entry to consider the magnitude of the brand.



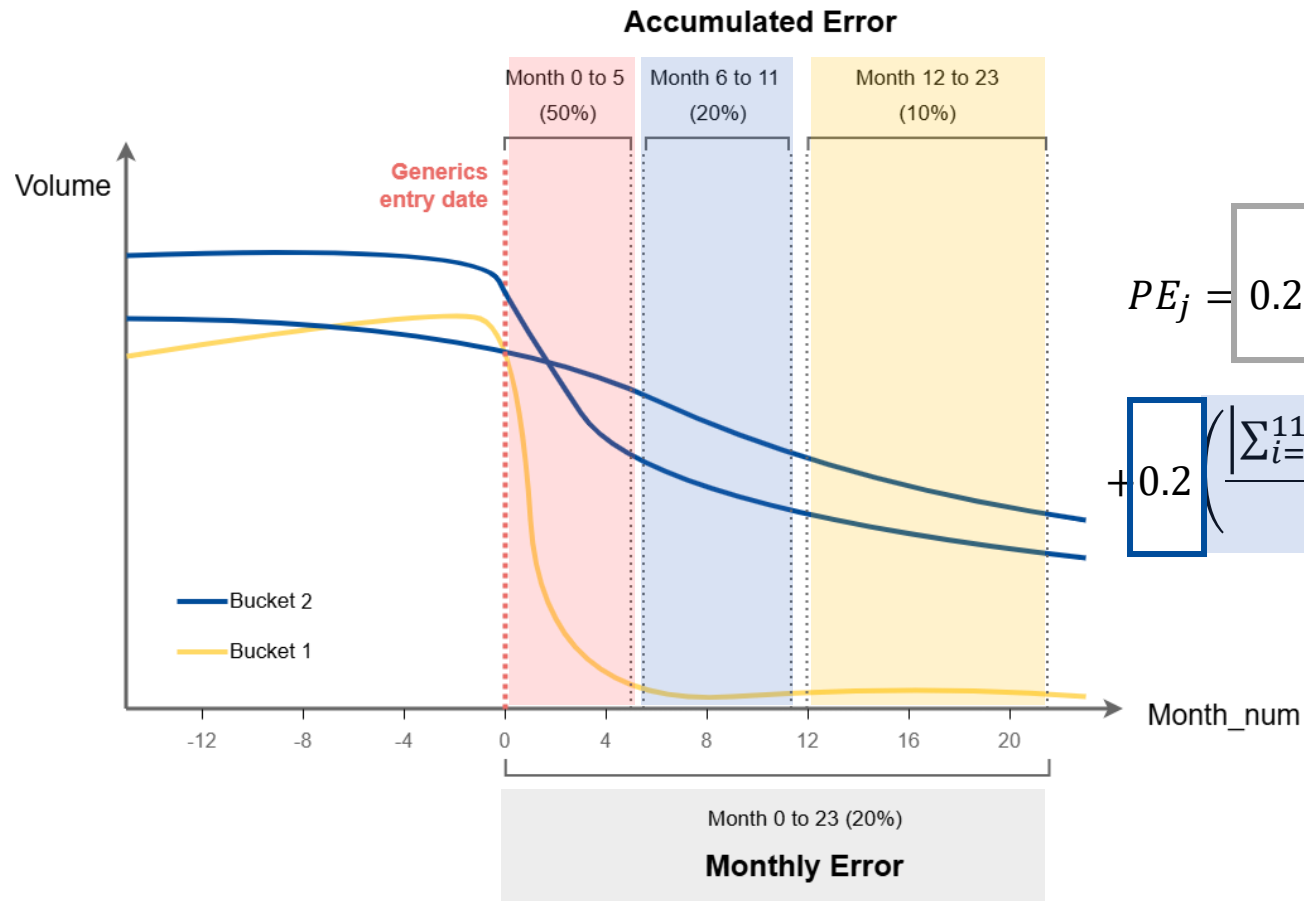


# Metric: Prediction Error (Phase 1-a)



$$PE_j = 0.2 \left( \frac{\sum_{i=0}^{23} |Y_{j,i}^{act} - Y_{j,i}^{pred}|}{24 \cdot Avg_j} \right) + 0.5 \left( \frac{|\sum_{i=0}^5 Y_{j,i}^{act} - \sum_{i=0}^5 Y_{j,i}^{pred}|}{6 \cdot Avg_j} \right) + 0.2 \left( \frac{|\sum_{i=6}^{11} Y_{j,i}^{act} - \sum_{i=6}^{11} Y_{j,i}^{pred}|}{6 \cdot Avg_j} \right) + 0.1 \left( \frac{|\sum_{i=12}^{23} Y_{j,i}^{act} - \sum_{i=12}^{23} Y_{j,i}^{pred}|}{12 \cdot Avg_j} \right)$$

# Metric: Prediction Error (Phase 1-a)



$$PE_j = 0.2 \left( \frac{\sum_{i=0}^{23} |Y_{j,i}^{act} - Y_{j,i}^{pred}|}{24 \cdot Avg_j} \right) + 0.5 \left( \frac{|\sum_{i=0}^5 Y_{j,i}^{act} - \sum_{i=0}^5 Y_{j,i}^{pred}|}{6 \cdot Avg_j} \right) + 0.2 \left( \frac{|\sum_{i=6}^{11} Y_{j,i}^{act} - \sum_{i=6}^{11} Y_{j,i}^{pred}|}{6 \cdot Avg_j} \right) + 0.1 \left( \frac{|\sum_{i=12}^{23} Y_{j,i}^{act} - \sum_{i=12}^{23} Y_{j,i}^{pred}|}{12 \cdot Avg_j} \right)$$

$$Avg_j = \frac{\sum_{i=-12}^{-1} Y_{j,i}^{act}}{12}$$

# Metric: Prediction Error (Phase 1-a)

Prediction Error per country brand ( $PE_j$ ):

$$PE_j = 0.2 \left( \frac{\sum_{i=0}^{23} |Y_{j,i}^{act} - Y_{j,i}^{pred}|}{24 \cdot Avg_j} \right) + 0.5 \left( \frac{|\sum_{i=0}^5 Y_{j,i}^{act} - \sum_{i=0}^5 Y_{j,i}^{pred}|}{6 \cdot Avg_j} \right) \\ + 0.2 \left( \frac{|\sum_{i=6}^{11} Y_{j,i}^{act} - \sum_{i=6}^{11} Y_{j,i}^{pred}|}{6 \cdot Avg_j} \right) + 0.1 \left( \frac{|\sum_{i=12}^{23} Y_{j,i}^{act} - \sum_{i=12}^{23} Y_{j,i}^{pred}|}{12 \cdot Avg_j} \right)$$

The final Prediction Error ( $PE$ ) will be the weighted sum of the averages of all the individual prediction errors ( $PE_j$ ) across the two buckets. Note that the average of Bucket 1 is weighted twice as much as that from Bucket 2.

$$PE = \frac{2}{n_{B1}} \sum_{j=1}^{n_{B1}} PE_{j,B1} + \frac{1}{n_{B2}} \sum_{j=1}^{n_{B2}} PE_{j,B2}$$

# Metric: Prediction Error (Phase 1-b)

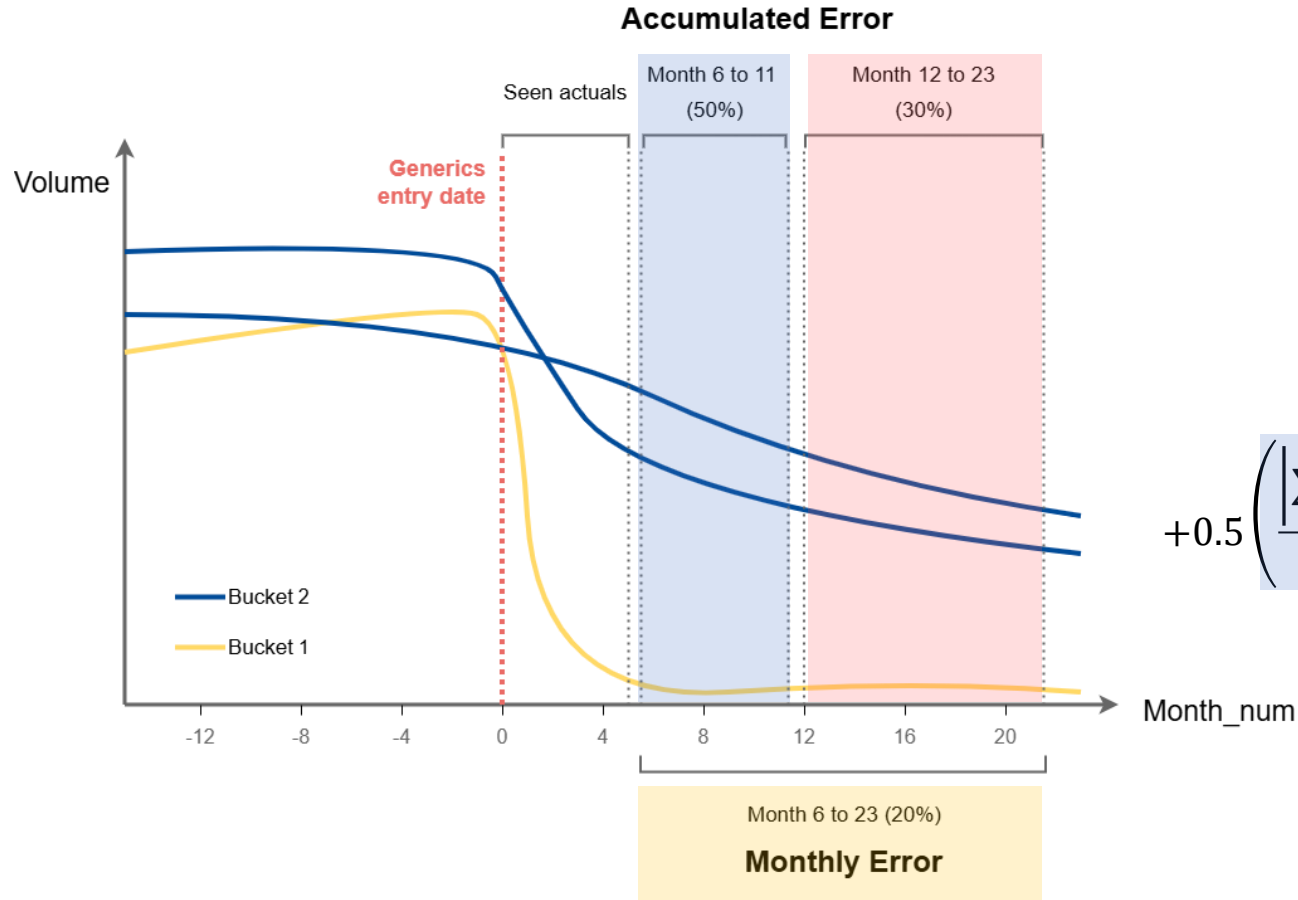
In this second scenario (Scenario 2), participants will provide predictions based on **6 actual data points available** after the generic entry date. To compute the prediction error of Phase 1-b, we will evaluate the difference between the predicted values vs the actual volume in three different ways weighted as follows:

1. Absolute **monthly** error of months 6 to 23 (20%)
2. Absolute **accumulated** error of months 6 to 11 (50%)
3. Absolute **accumulated** error of months 12 to 23 (30%)

All the 3 items will be normalized by the average ( $Avg_j$ ) monthly volume of the last 12 months before the generic entry to consider the magnitude of the brand.



# Metric: Prediction Error (Phase 1-b)



$$PE_j = 0.2 \left( \frac{\sum_{i=6}^{23} |y_{j,i}^{\text{act}} - y_{j,i}^{\text{pred}}|}{18 \cdot Avg_j} \right)$$

$$+ 0.5 \left( \frac{|\sum_{i=6}^{11} y_{j,i}^{\text{act}} - \sum_{i=6}^{11} y_{j,i}^{\text{pred}}|}{6 \cdot Avg_j} \right) + 0.3 \left( \frac{|\sum_{i=12}^{23} y_{j,i}^{\text{act}} - \sum_{i=12}^{23} y_{j,i}^{\text{pred}}|}{12 \cdot Avg_j} \right)$$

# Metric: Prediction Error (Phase 1-b)

Prediction Error per country brand ( $PE_j$ ):

$$PE_j = 0.2 \left( \frac{\sum_{i=6}^{23} |y_{j,i}^{\text{act}} - y_{j,i}^{\text{pred}}|}{18 \cdot \text{Avg}_j} \right) + 0.5 \left( \frac{|\sum_{i=6}^{11} y_{j,i}^{\text{act}} - \sum_{i=6}^{11} y_{j,i}^{\text{pred}}|}{6 \cdot \text{Avg}_j} \right) + 0.3 \left( \frac{|\sum_{i=12}^{23} y_{j,i}^{\text{act}} - \sum_{i=12}^{23} y_{j,i}^{\text{pred}}|}{12 \cdot \text{Avg}_j} \right)$$

The final Prediction Error ( $PE$ ) will be the weighted sum of the averages of all the individual prediction errors ( $PE_j$ ) across the two buckets. Note that the average of Bucket 1 is weighted twice as much as that from Bucket 2.

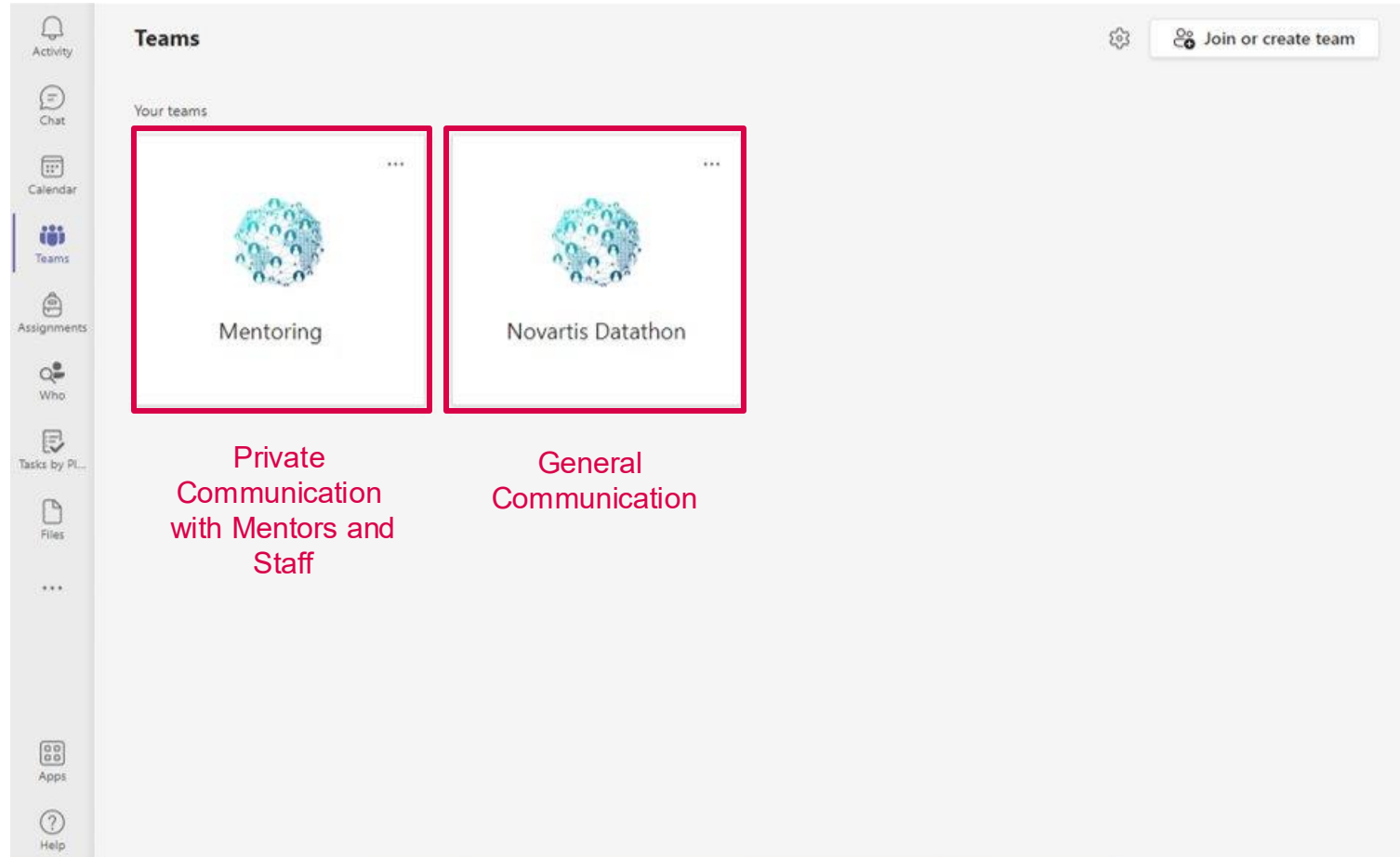
$$PE = \frac{2}{n_{B1}} \sum_{j=1}^{n_{B1}} PE_{j,B1} + \frac{1}{n_{B2}} \sum_{j=1}^{n_{B2}} PE_{j,B2}$$

# Technical part

Communication: **Microsoft Teams**  
Upload the results: **Datathon Platform**

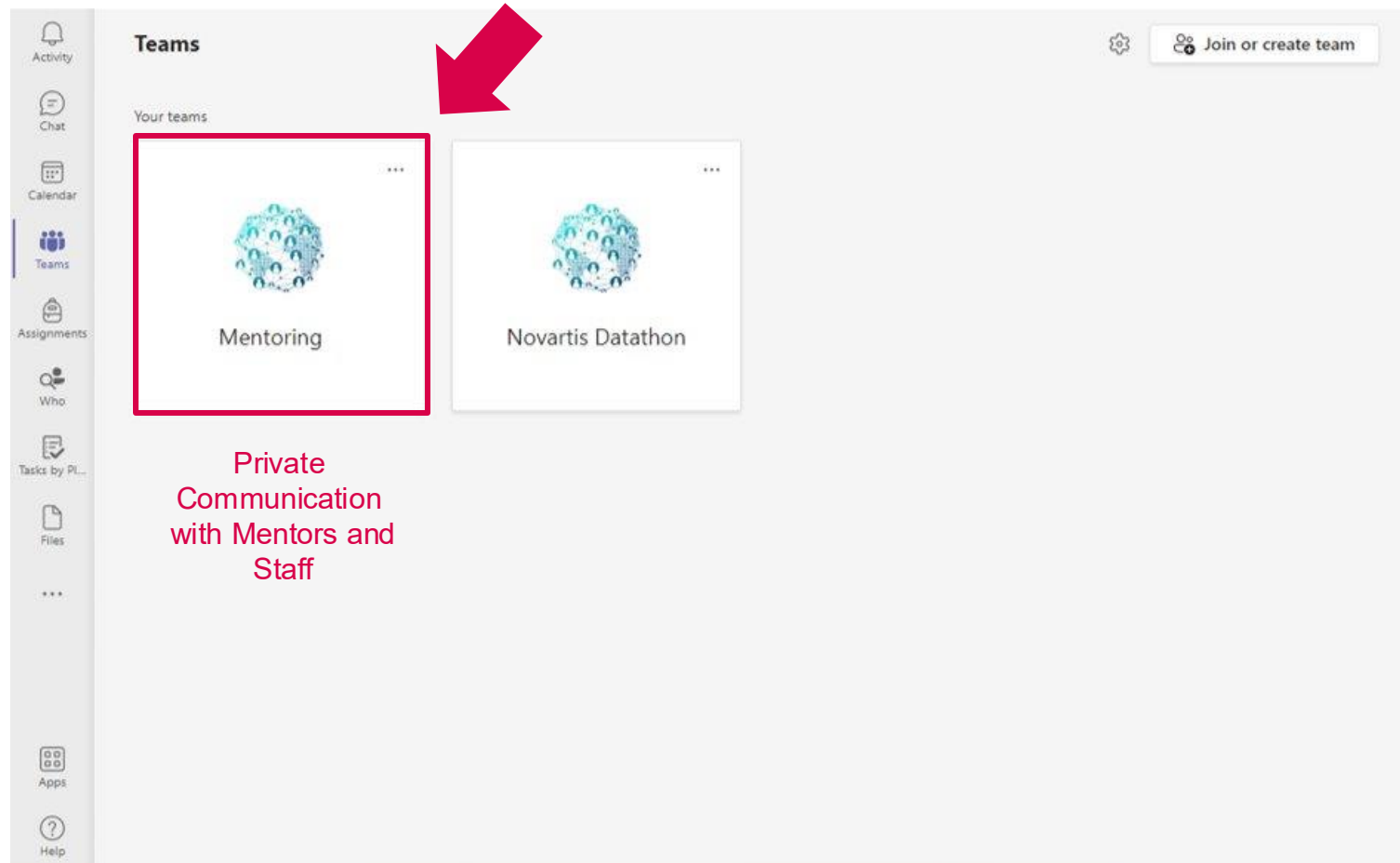


# Communication Channel

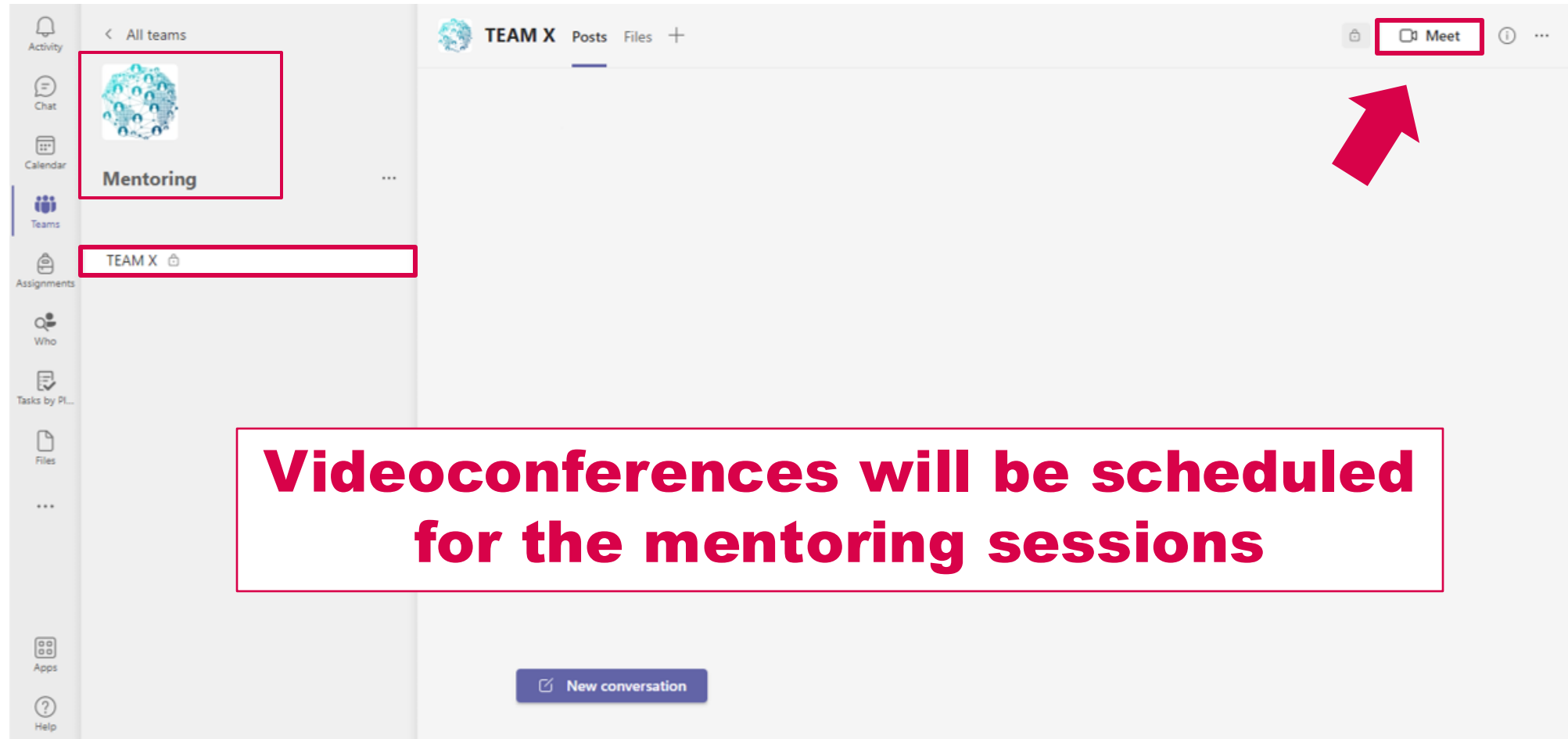




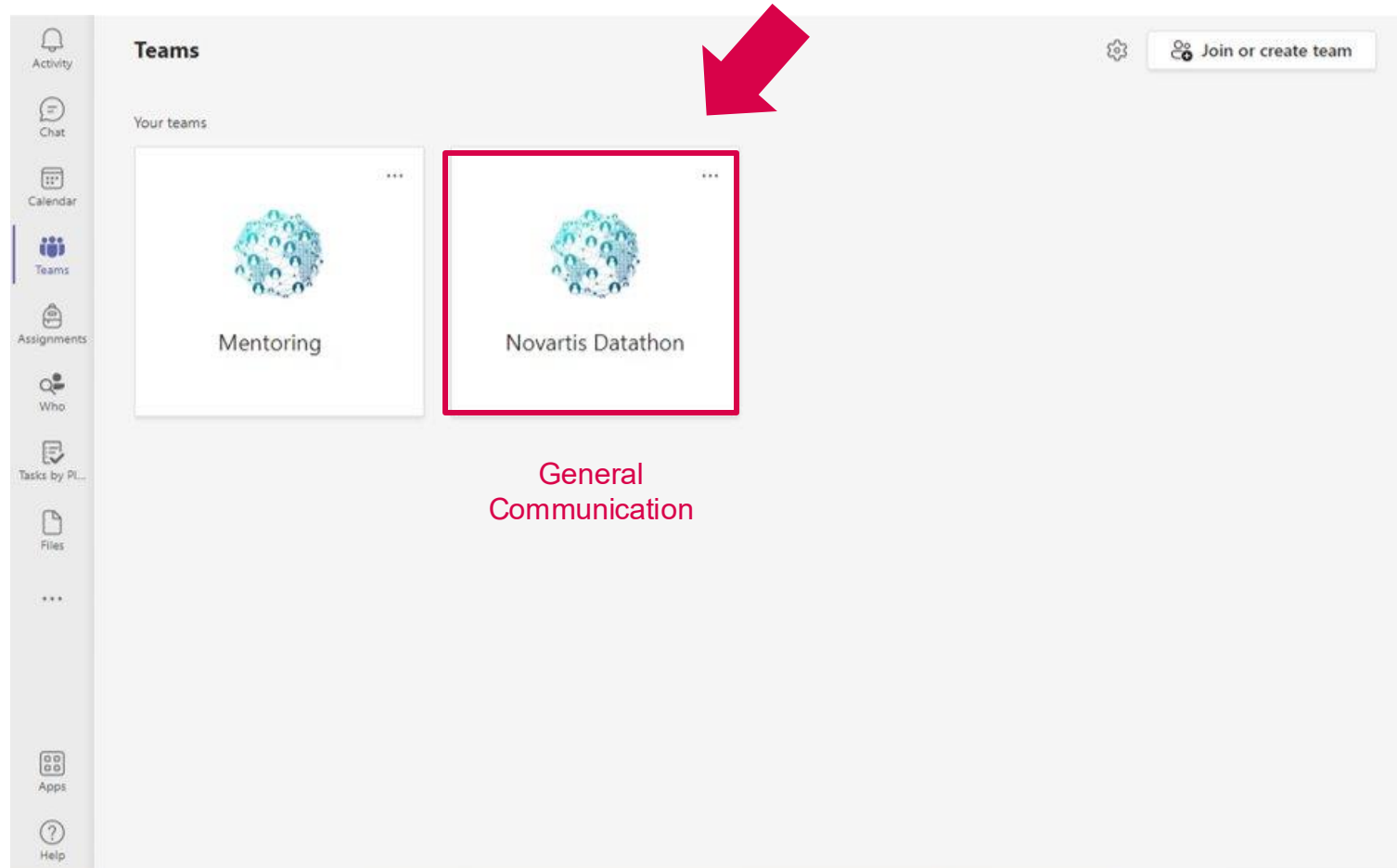
# Communication Channel



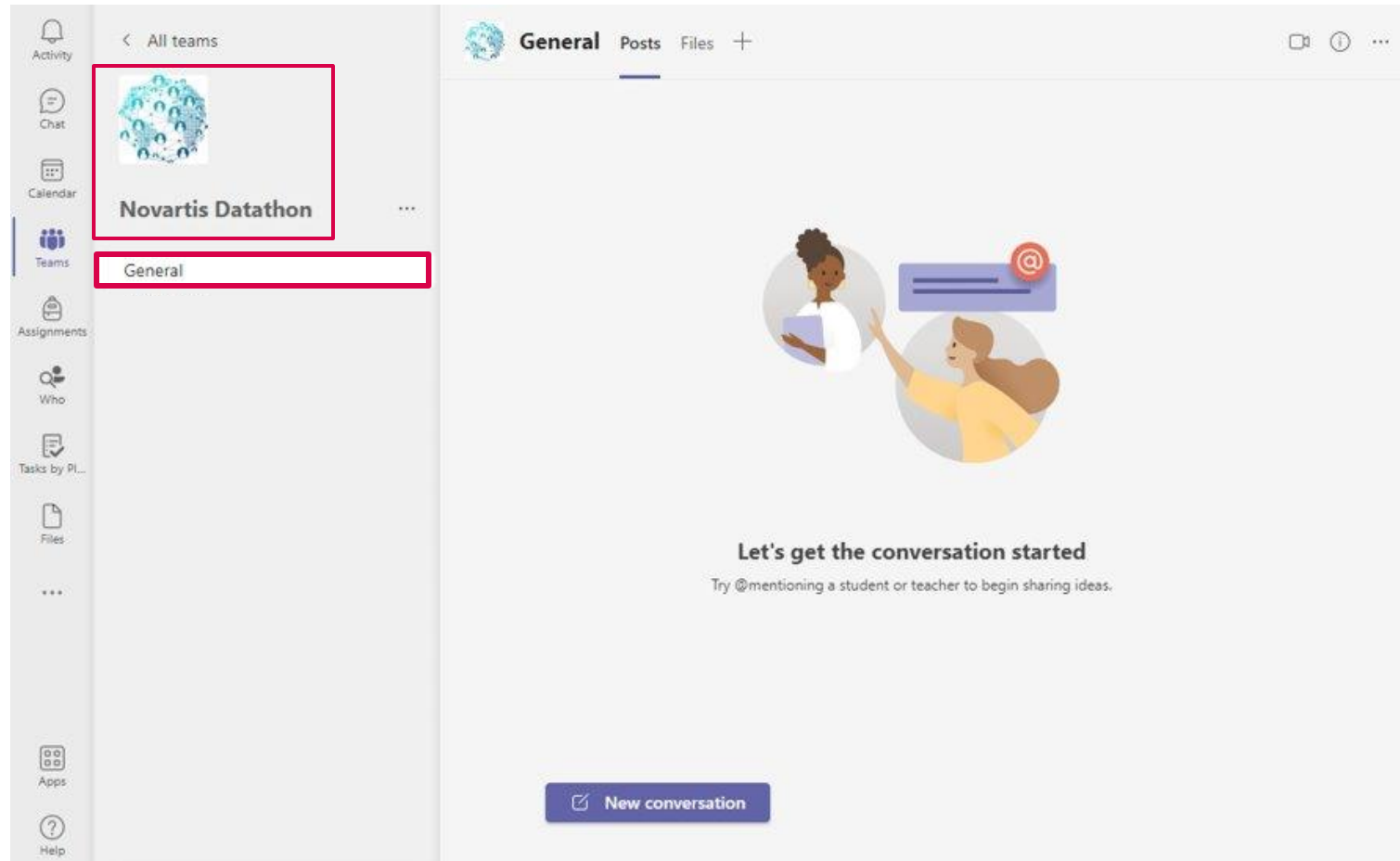
# Communication Channel – Private communication



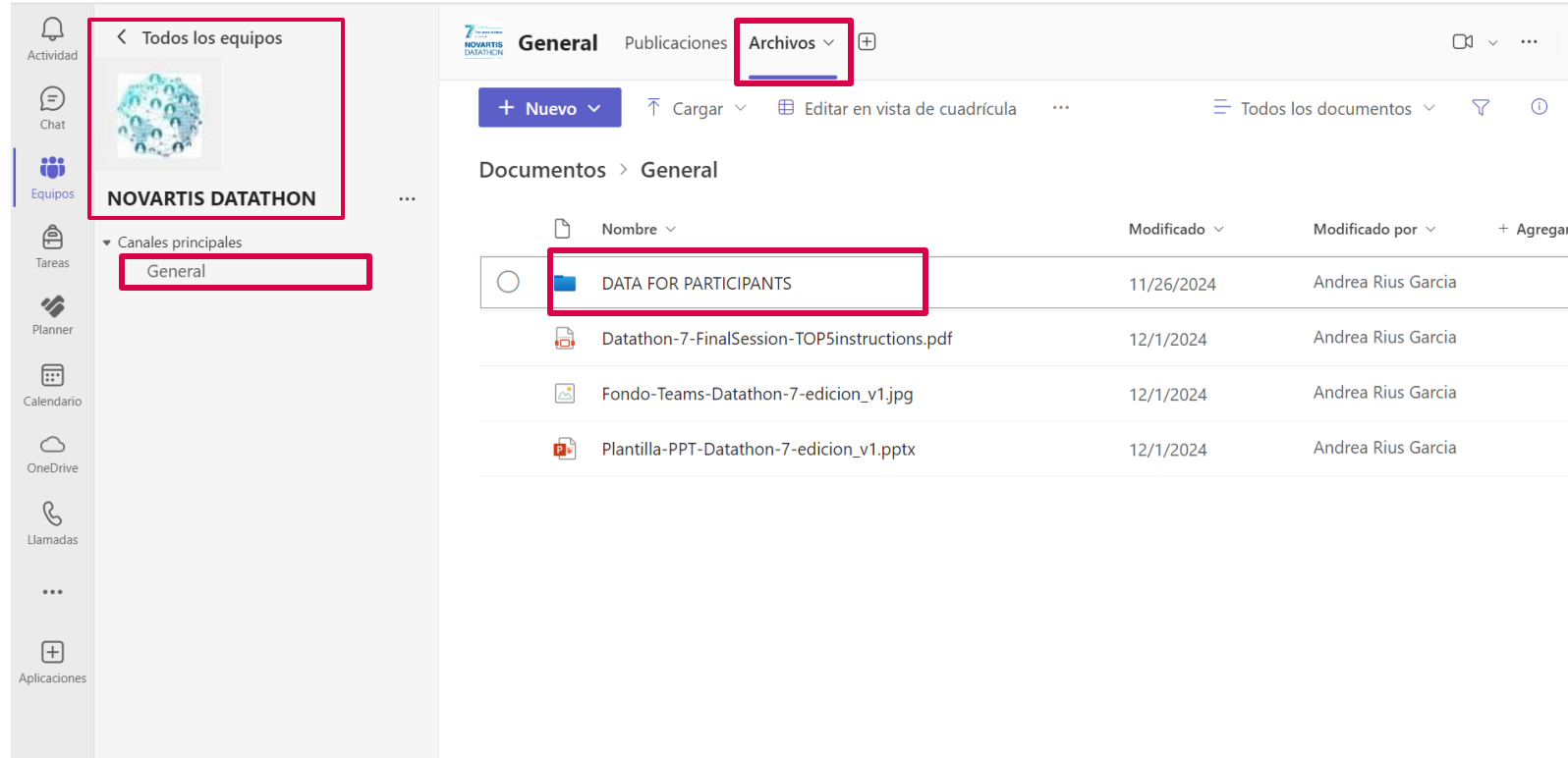
# Communication Channel



# Communication Channel – General communication



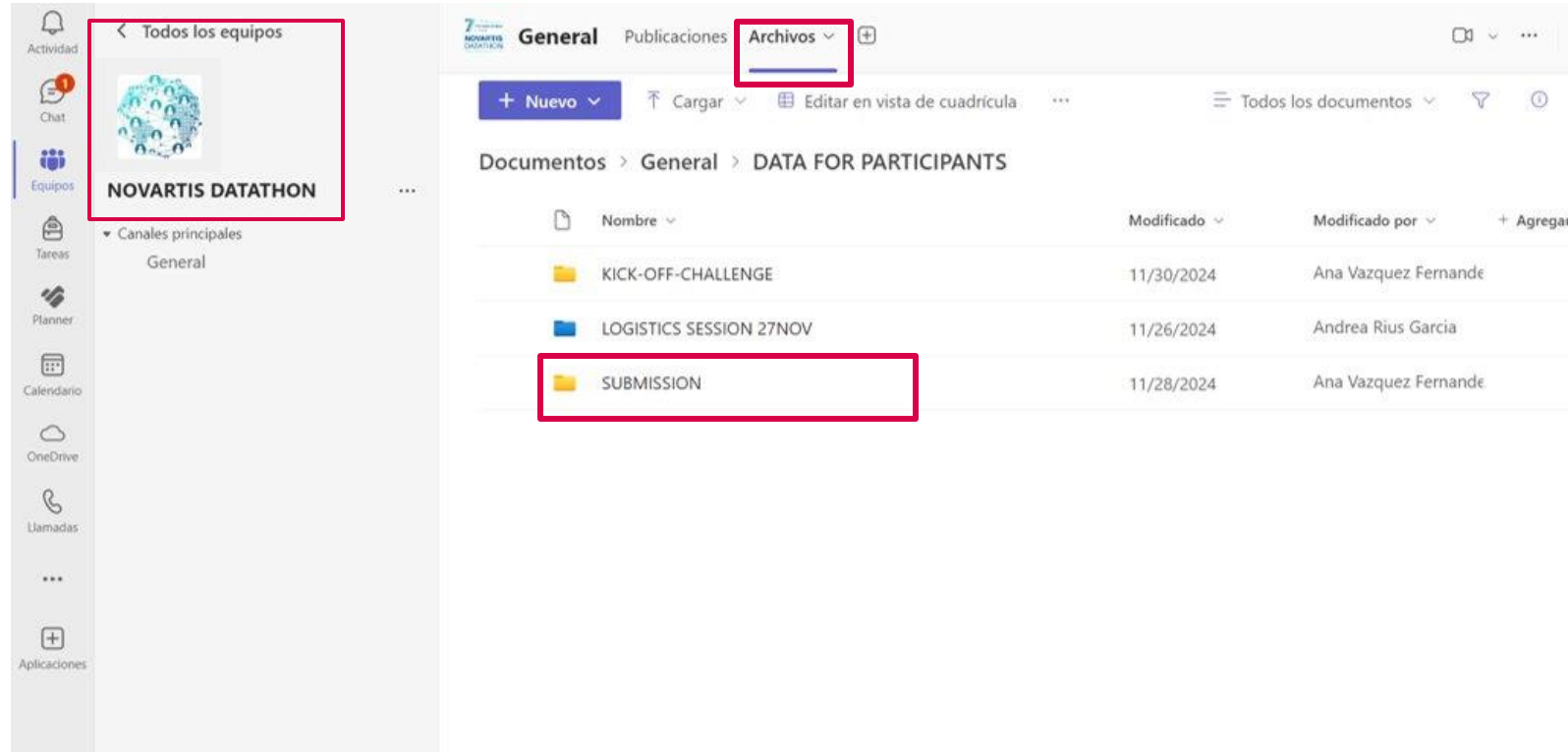
# Download data



The screenshot shows the Novartis Datathon Teams interface. On the left sidebar, the 'Equipos' (Teams) section is expanded, showing 'Todos los equipos' (All teams) and the 'NOVARTIS DATATHON' team. Under 'Canales principales' (Main channels), the 'General' channel is selected. The main content area shows the 'Archivos' (Files) tab, which displays a list of documents in the 'General' channel. The 'DATA FOR PARTICIPANTS' folder is highlighted.

Nombre	Modificado	Modificado por	+ Agregar
DATA FOR PARTICIPANTS	11/26/2024	Andrea Rius Garcia	
Datathon-7-FinalSession-TOP5instructions.pdf	12/1/2024	Andrea Rius Garcia	
Fondo-Teams-Datathon-7-edicion_v1.jpg	12/1/2024	Andrea Rius Garcia	
Plantilla-PPT-Datathon-7-edicion_v1.pptx	12/1/2024	Andrea Rius Garcia	

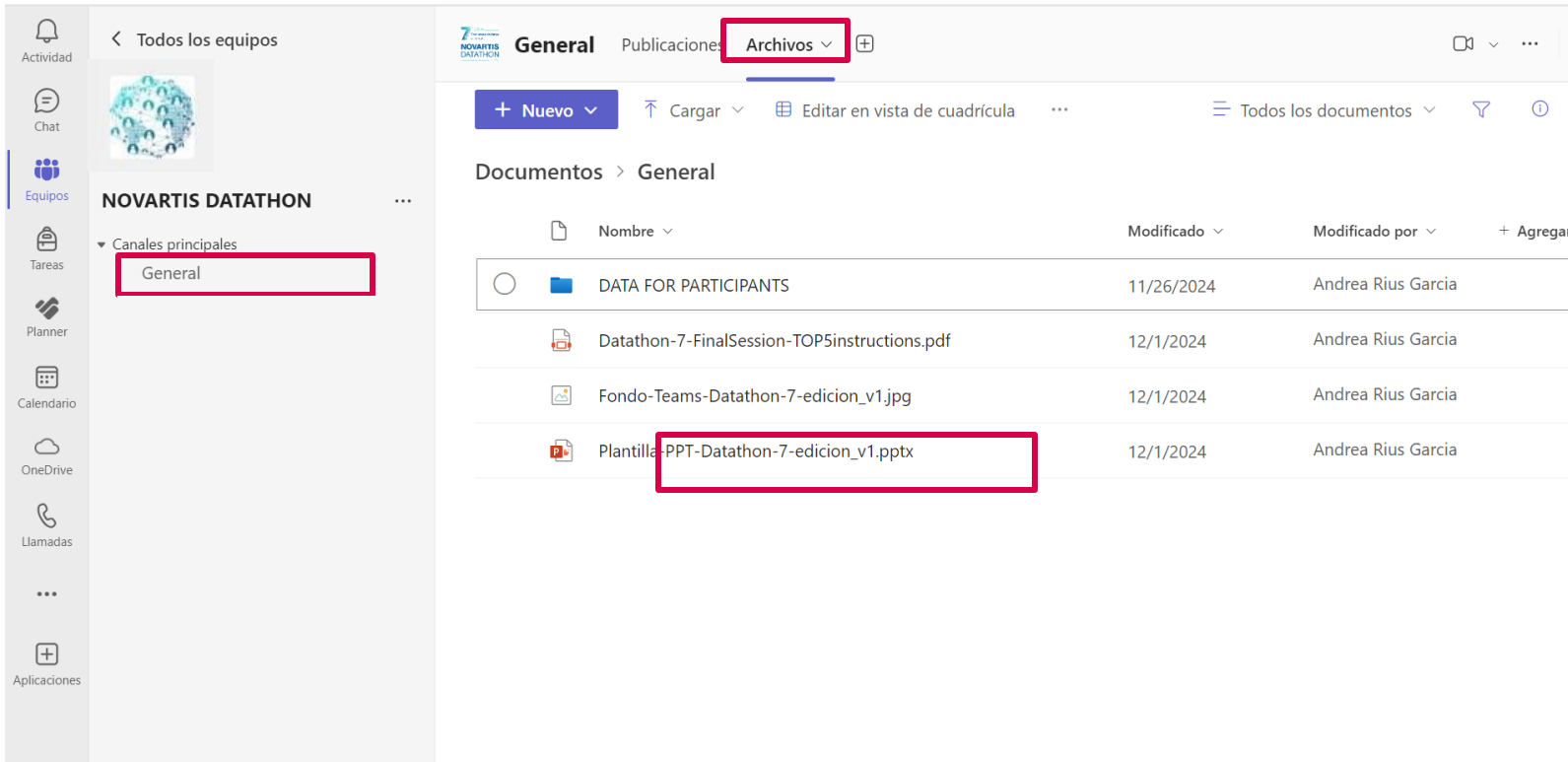
# Download data



The screenshot shows the Novartis Datathon interface. On the left sidebar, the 'Equipos' section is highlighted with a red box, showing the 'NOVARTIS DATATHON' team. The main content area is divided into three tabs: 'General', 'Publicaciones', and 'Archivos'. The 'Archivos' tab is selected and highlighted with a red box. Below the tabs, there is a table of documents. The table has columns for 'Nombre', 'Modificado', and 'Modificado por'. The 'SUBMISSION' folder is highlighted with a red box.

Nombre	Modificado	Modificado por
KICK-OFF-CHALLENGE	11/30/2024	Ana Vazquez Fernande
LOGISTICS SESSION 27NOV	11/26/2024	Andrea Rius Garcia
<b>SUBMISSION</b>	11/28/2024	Ana Vazquez Fernande

# Submit presentation & code TOP 5



The screenshot shows the Novartis Datathon interface. On the left sidebar, the 'Equipos' section is expanded, and the 'General' channel is selected. The main content area shows the 'Archivos' (Files) tab, which contains a list of documents. The document 'Plantilla-PPT-Datathon-7-edicion\_v1.pptx' is highlighted with a red box.

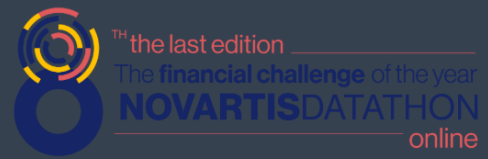
Nombre	Modificado	Modificado por	+ Agregar
DATA FOR PARTICIPANTS	11/26/2024	Andrea Rius Garcia	
Datathon-7-FinalSession-TOP5instructions.pdf	12/1/2024	Andrea Rius Garcia	
Fondo-Teams-Datathon-7-edicion_v1.jpg	12/1/2024	Andrea Rius Garcia	
Plantilla-PPT-Datathon-7-edicion_v1.pptx	12/1/2024	Andrea Rius Garcia	

# Platform login

## Credentials

user: teamX@novartisdatathon

password: pwdteamX



8<sup>TH</sup> the last edition  
The financial challenge of the year  
**NOVARTISDATATHON**  
online

LOG IN

Email

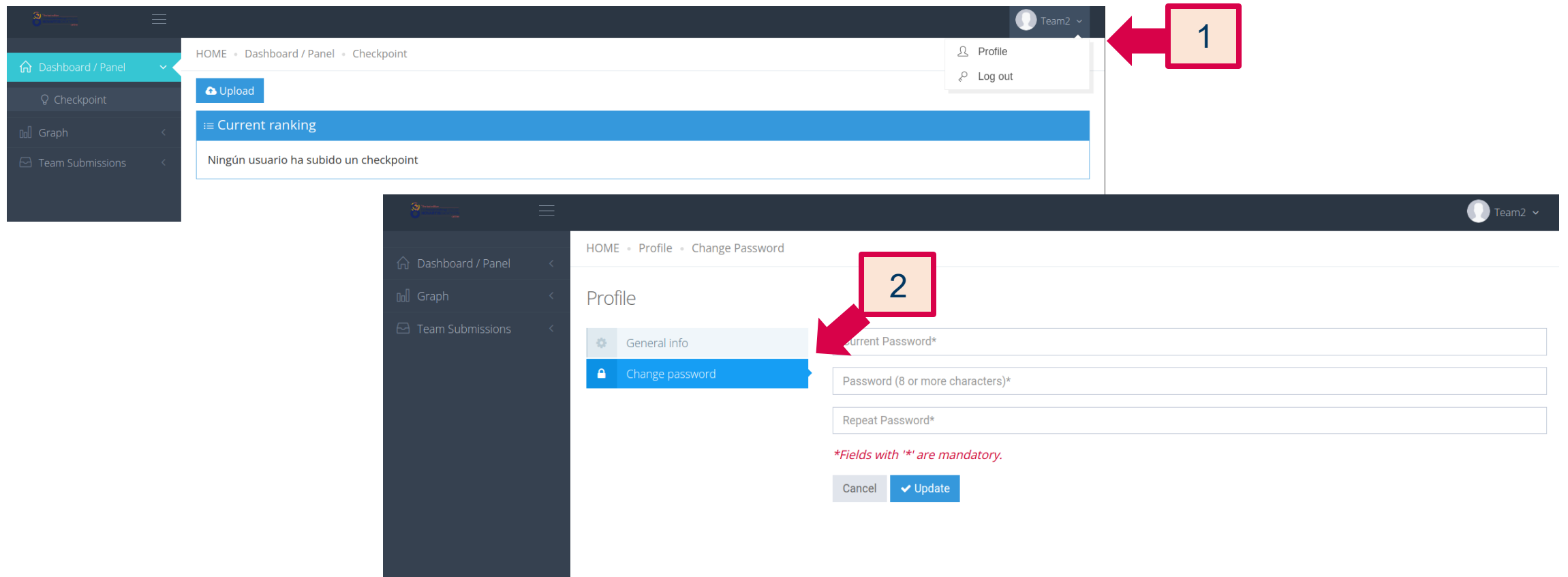
Password

Sign in

© Euratec. All rights reserved. [Privacy politics](#)



# How to submit results – Change the password



The image shows two screenshots of a web application interface. The top screenshot shows the 'Dashboard / Panel' page with a user profile dropdown menu open, labeled with a red box and the number '1'. The bottom screenshot shows the 'Profile' page with the 'Change password' option selected, labeled with a red box and the number '2'.

**Step 1:** The user profile dropdown menu is open, showing options for 'Profile' and 'Log out'.

**Step 2:** The 'Change password' option is selected in the profile menu.

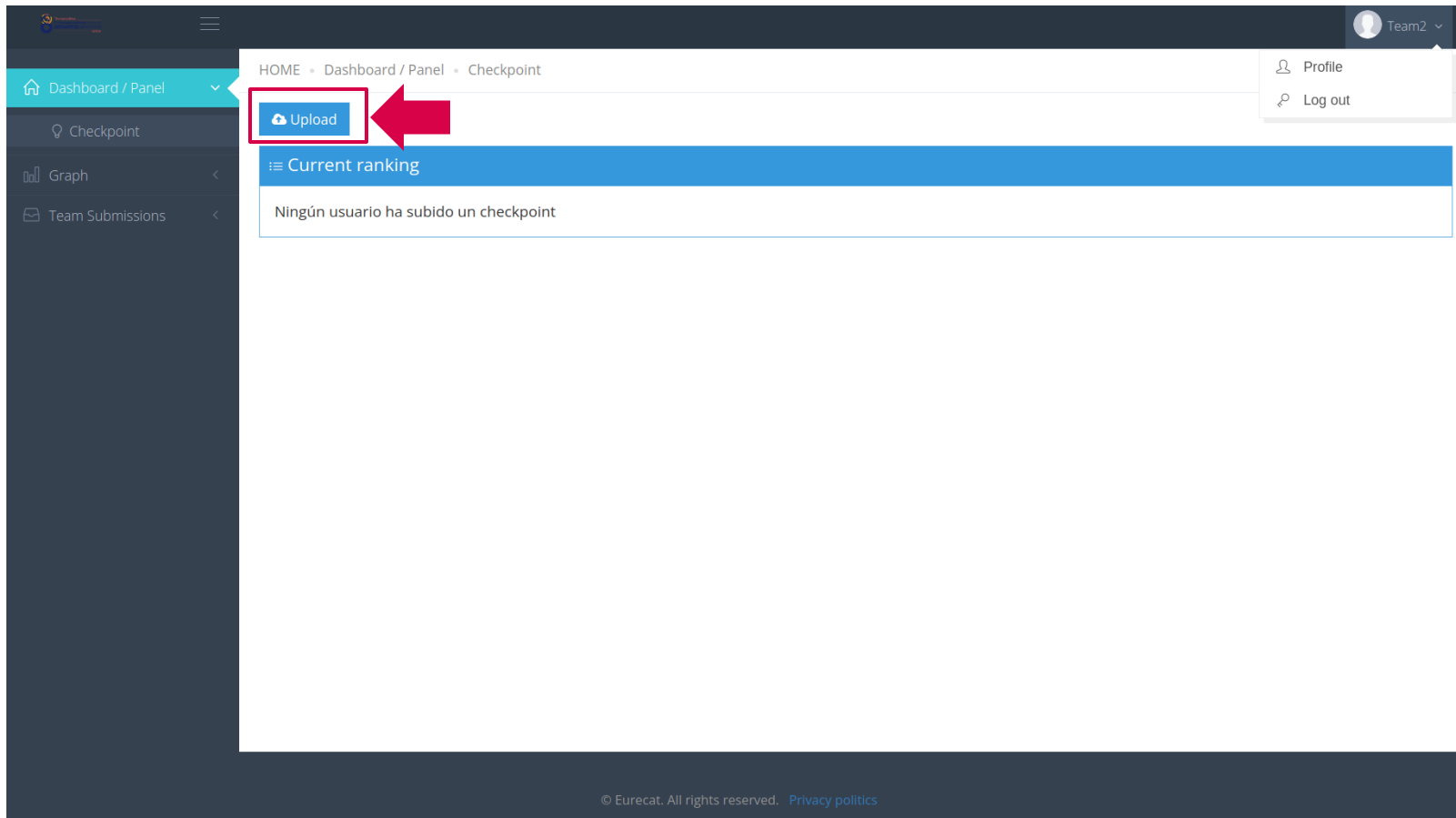
The 'Change Password' form includes the following fields:

- Current Password\*
- Password (8 or more characters)\*
- Repeat Password\*

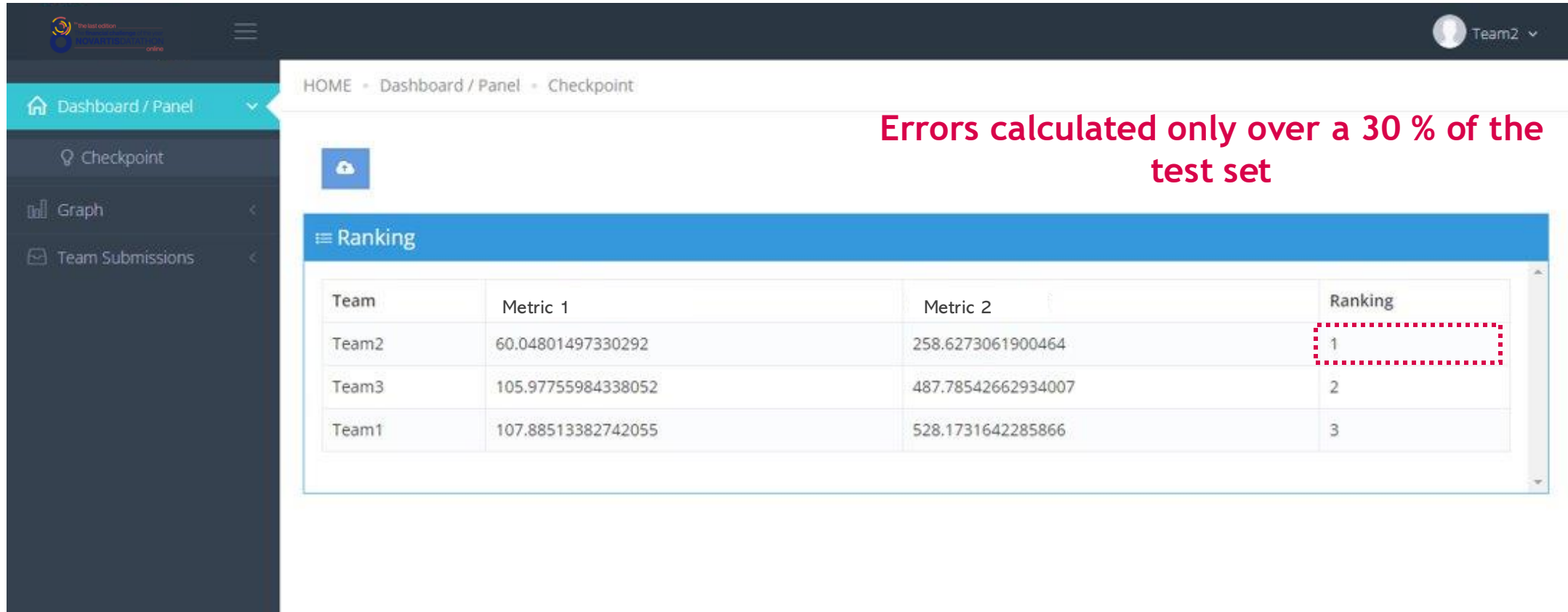
*\*Fields with '\*' are mandatory.*

Buttons: Cancel, Update

# How to submit results – Submission



# How to submit results – Ranking checkpoints



HOME • Dashboard / Panel • Checkpoint

Dashboard / Panel

Checkpoint

Graph

Team Submissions

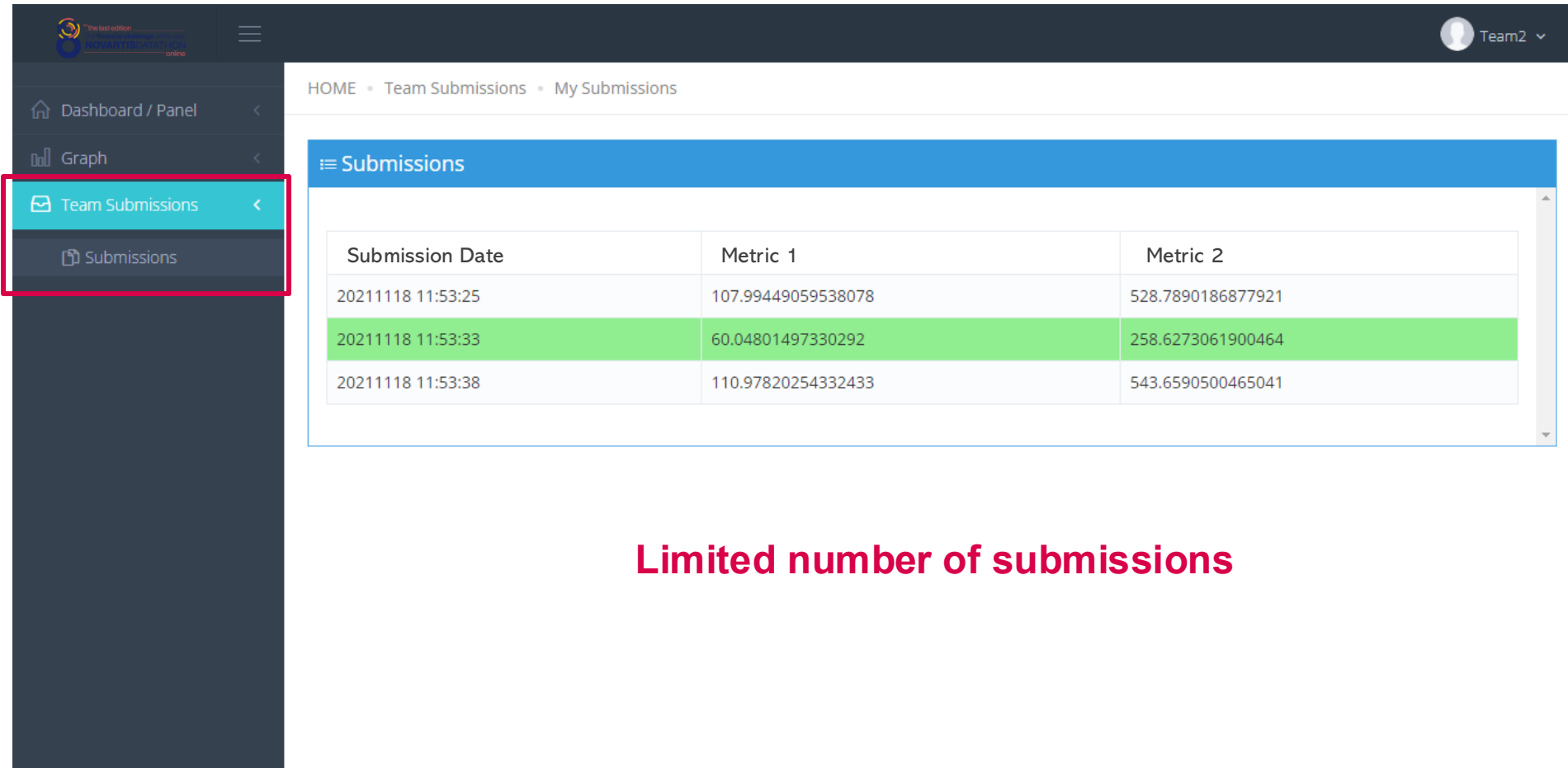
Ranking

Team	Metric 1	Metric 2	Ranking
Team2	60.04801497330292	258.6273061900464	1
Team3	105.97755984338052	487.78542662934007	2
Team1	107.88513382742055	528.1731642285866	3

Errors calculated only over a 30 % of the test set

Your best submission is shown

# How to submit results – History of submissions



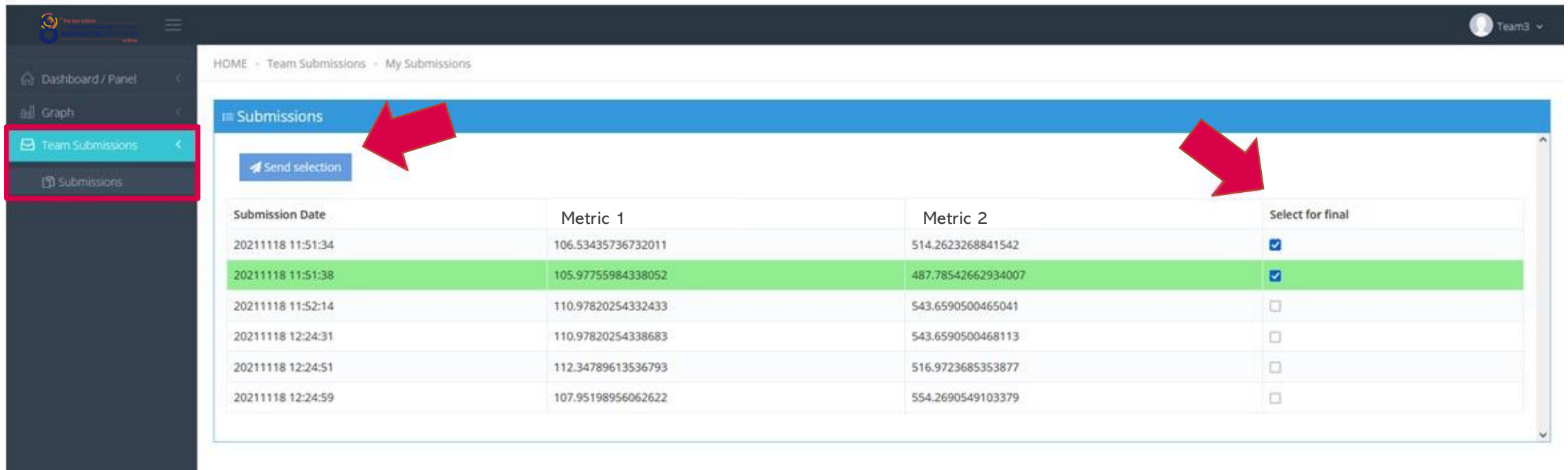
HOME • Team Submissions • My Submissions

### Submissions

Submission Date	Metric 1	Metric 2
20211118 11:53:25	107.99449059538078	528.7890186877921
20211118 11:53:33	60.04801497330292	258.6273061900464
20211118 11:53:38	110.97820254332433	543.6590500465041

**Limited number of submissions**

# How to submit results – Final submission (last hour)



The screenshot shows the Novartis submission interface. On the left, a sidebar menu has 'Team Submissions' highlighted with a red box. The main area displays a table of submissions under the 'Submissions' header. A red arrow points to the 'Send selection' button, and another red arrow points to the 'Select for final' checkbox in the table.

Submission Date	Metric 1	Metric 2	Select for final
20211118 11:51:34	106.53435736732011	514.2623268841542	<input checked="" type="checkbox"/>
20211118 11:51:38	105.97755984338052	487.78542662934007	<input checked="" type="checkbox"/>
20211118 11:52:14	110.97820254332433	543.6590500465041	<input type="checkbox"/>
20211118 12:24:31	110.97820254338683	543.6590500468113	<input type="checkbox"/>
20211118 12:24:51	112.34789613536793	516.9723685353877	<input type="checkbox"/>
20211118 12:24:59	107.95198956062622	554.2690549103379	<input type="checkbox"/>

**30<sup>th</sup> Nov between 9:30am and 10:30am \*:**  
Select a **maximum** number of submissions


\*Central European Time – Barcelona, UTC +1h

# How to submit results

FINAL results calculated over the  
100% of the test set  
once the submission deadline is  
over  
(30<sup>th</sup> Nov 10:30am)



# Show results



Dashboard / Panel

Checkpoint

Final - TOP 10

Final - TOP 5

Graph

Team Submissions

HOME

Dashboard / Panel

Checkpoint

Team3

Ranking

Team	Metric 1	Metric 2	Ranking
Team2		3.6273061900464	1
Team3	105.97755984338052	487.78542662934007	2
Team4		4.2623268841542	3
Team1		3.1731642285866	4

Top 10 on Metric 1

Top 5 on Metric 2

# Submit presentation & code TOP 5

Activity Chat Calendar Teams Assignments Who Tasks by Pl... Apps Help

< All teams

**Mentoring 1**

**TEAM X**

TEAM X

Upload

Sync Copy link Download Add cloud storage All Documents

TEAM X

Name Modified Modified By

Drag files here

- 1 Data\_Novartis\_Datathon-Results\_Presentation\_TeamX
- 2 Data\_Novartis\_Datathon-Final\_Code\_TeamX



# Agenda



## THU 27 November

17:00h - 18:00h | Kick-off



## FRI 28 November

09:00h - 18:00h | Attendance of questions

09:00h - 12:00h | Mentoring

16:00h - 18:00h | Mentoring



## SAT 29 November

09:00h - 18:00h | Attendance of questions

09:00h - 12:00h | Mentoring

16:00h - 18:00h | Mentoring



## SUN 30 November

09:00h | Welcome and Jury introduction

09:30h - 10:30h | Final submissions

10:30h | Deadline Submit final csv

11:30h | Show Results

12:00h | Deadline to upload TOP5 presentation

13:00h - 14:30h | Finalists' presentations TOP 5

14:30h - 15:00h | Jury deliberates

15:00h | Announcement of the Winners

\*Central European Time - Barcelona, UTC +1h

Several thick, curved lines in dark blue, yellow, and red are positioned on the left side of the slide, partially overlapping the text.

**Thank you  
and good luck!**

Several thick, curved lines in dark blue, yellow, and red are positioned in the bottom right corner of the slide.