NOVARTIS

8TH the last edition

The **financial challenge** of the year

NOVARTISDATATHON

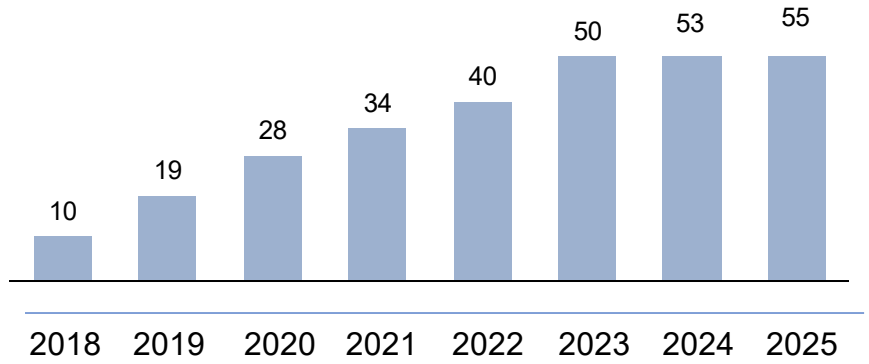
online

The BCN Digital Finance Hub

The team

- 34 Data scientists
- 8 Finance profiles
- 15 Visualization, ML, Software Engineers & Devops
- +14 Different nationalities in a diverse team
- 66% Local talent

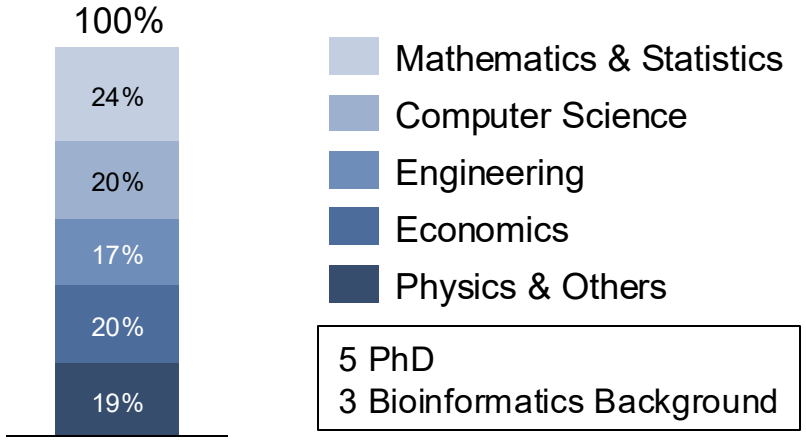
Evolution of the hub



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



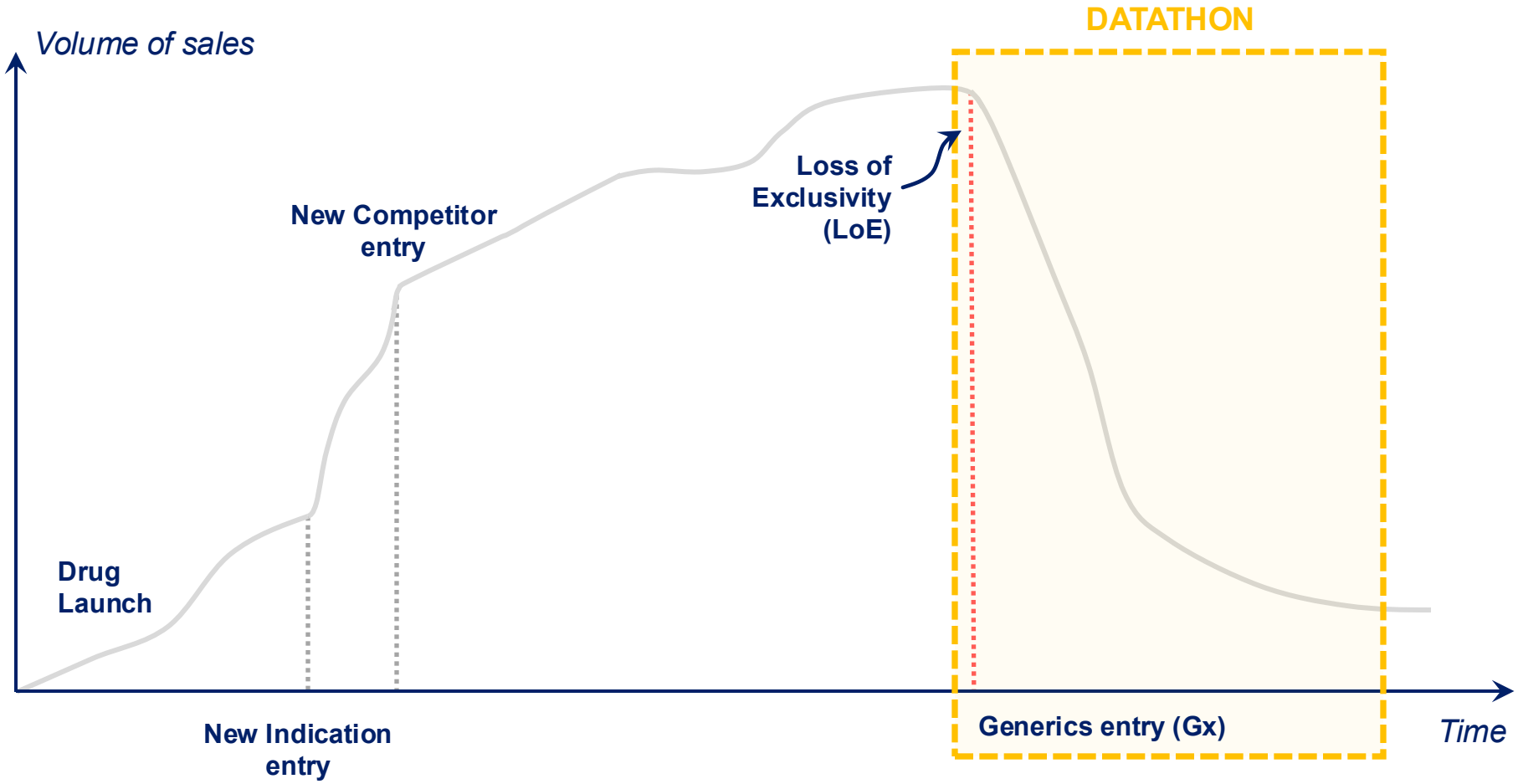
5 PhD
3 Bioinformatics Background

NOVARTIS

Datathon 8th edition

Generic impact

Lifecycle of a drug



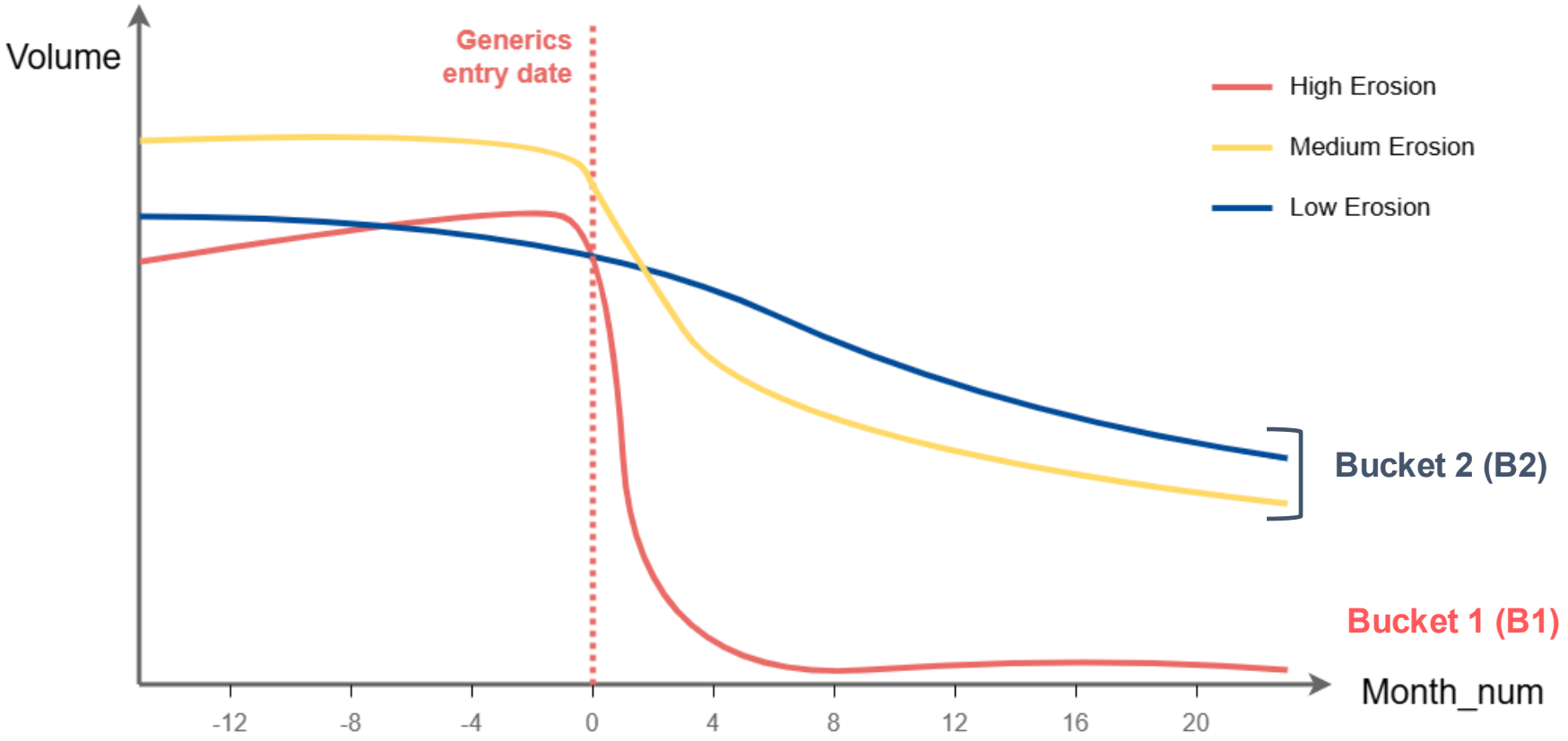
Generic erosion

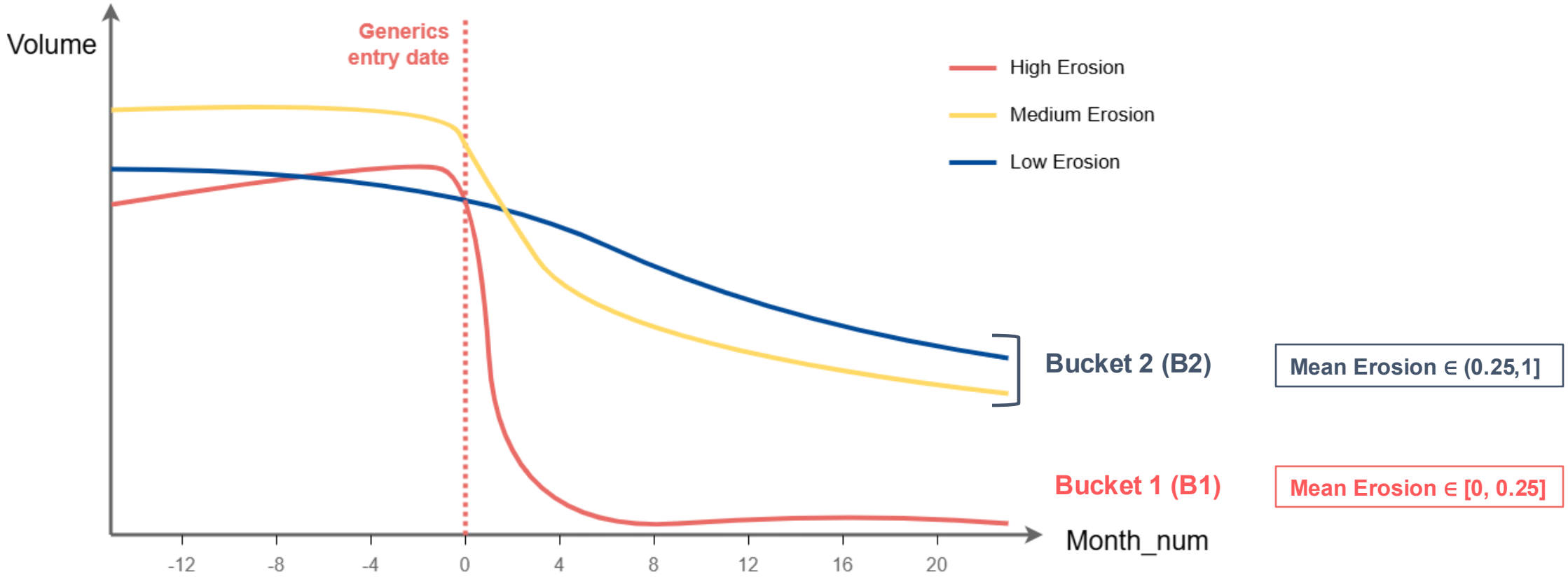
1. Formula

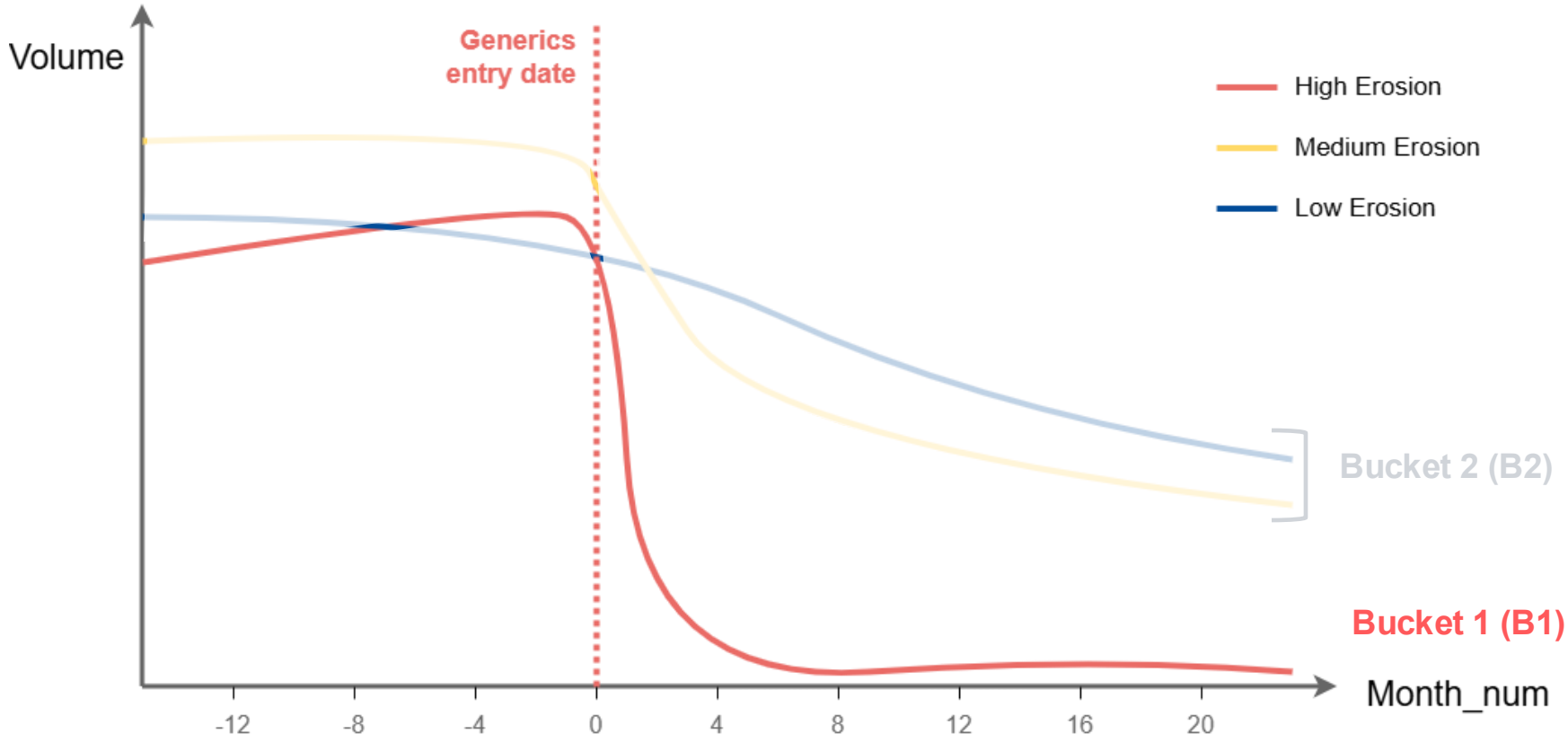
Mean Generic Erosion is defined as the mean of the normalized volumes after the generic entry considering a 24-month horizon. Volumes after generic entry are normalized by the average monthly volume of the last 12 months before the generic entry.

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

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


Volume
Generics entry date
High Erosion
Medium Erosion
Low Erosion
Bucket 2 (B2)
Bucket 1 (B1)
Month_num
-12
-8
-4
0
4
8
12
16
20


Volume
Generics entry date
High Erosion
Medium Erosion
Low Erosion
Bucket 2 (B2)
Mean Erosion ∈ (0.25,1]
Bucket 1 (B1)
Mean Erosion ∈ [0, 0.25]
Month_num
-12
-8
-4
0
4
8
12
16
20


Volume
Generics entry date
High Erosion
Medium Erosion
Low Erosion
Bucket 2 (B2)
Mean Erosion ∈ (0.25,1]
Datathon Focus!!!
Bucket 1 (B1)
Mean Erosion ∈ [0, 0.25]
Month_num
-12
-8
-4
0
4
8
12
16
20

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**

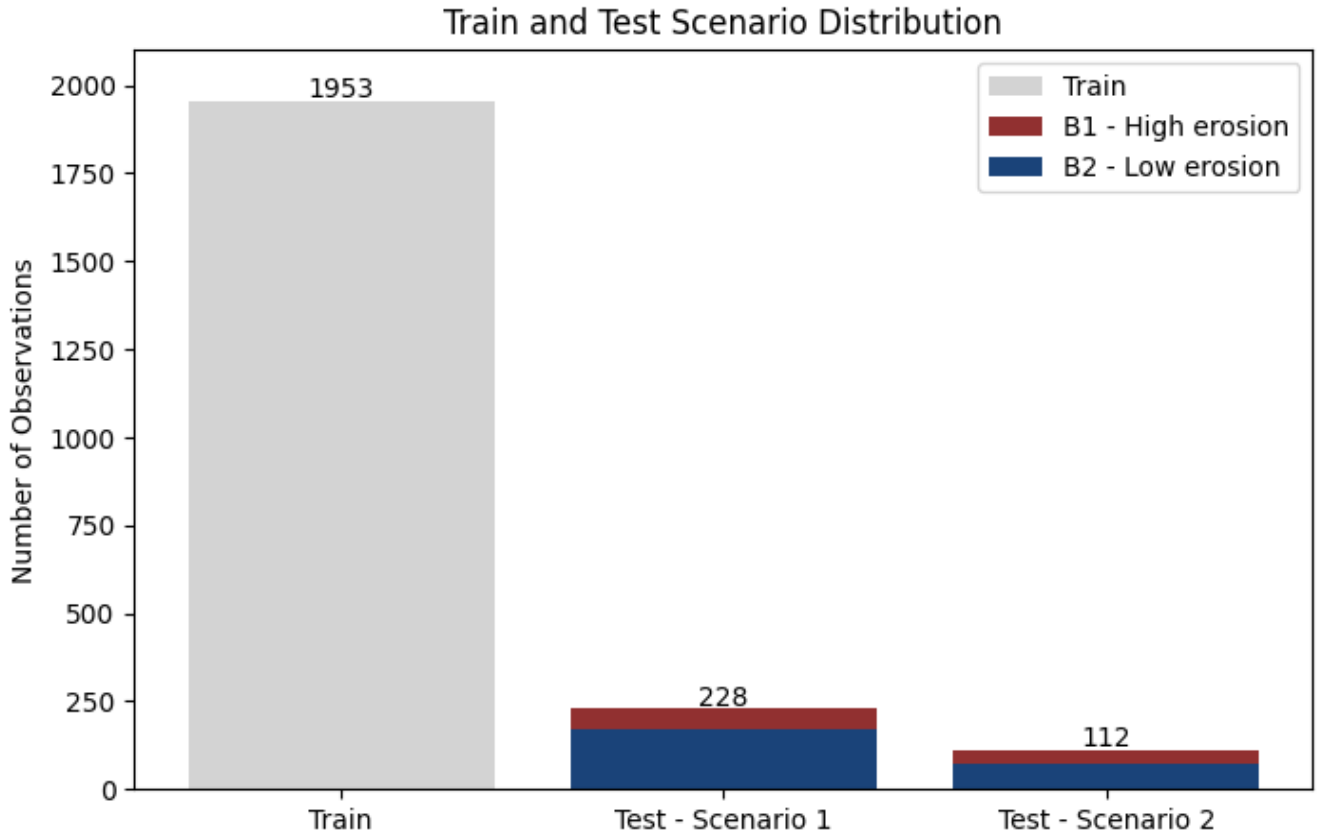
Challenge: Data Provided (2/2)

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

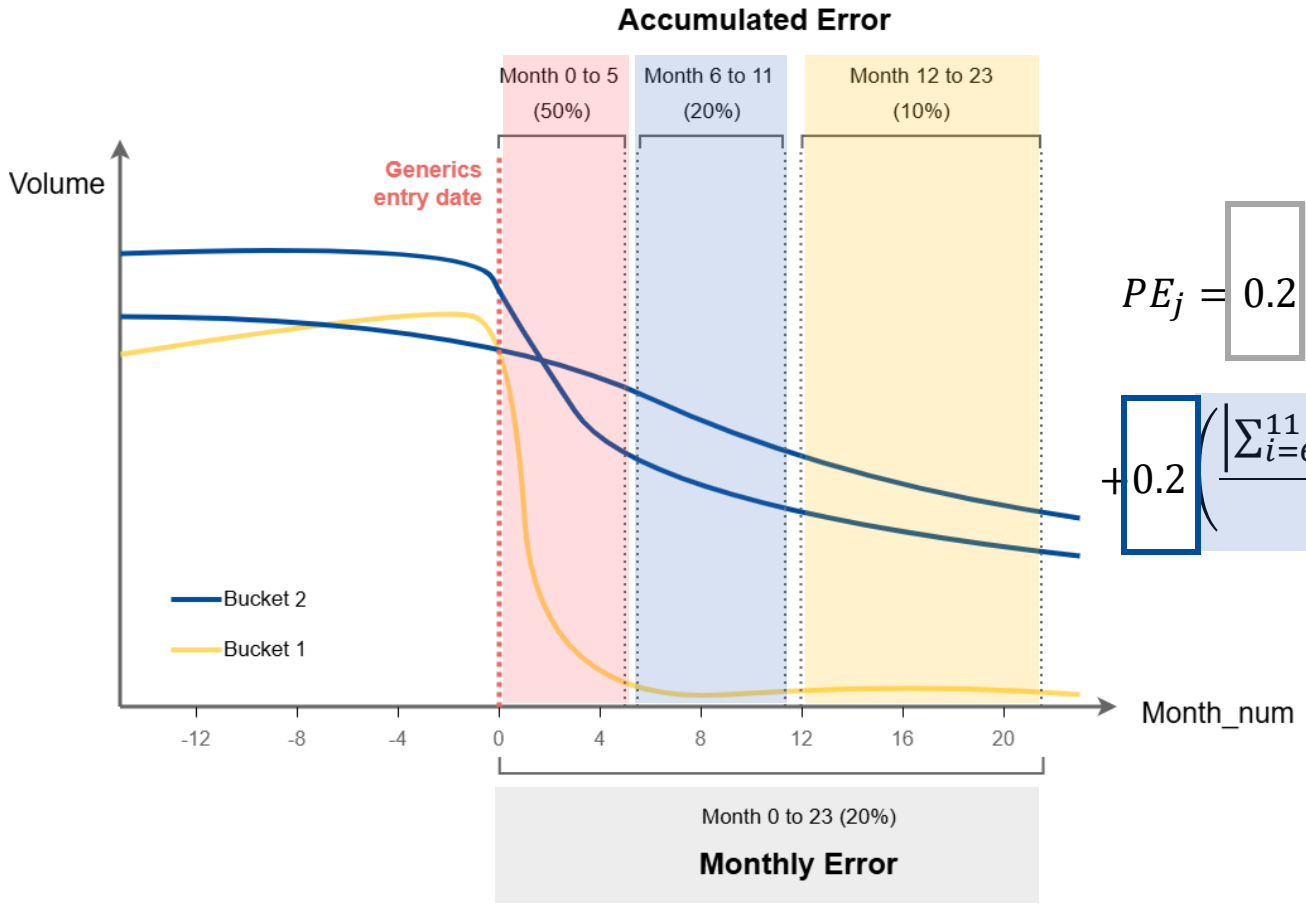
Metric: Prediction Error (Phase 1-a)

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.

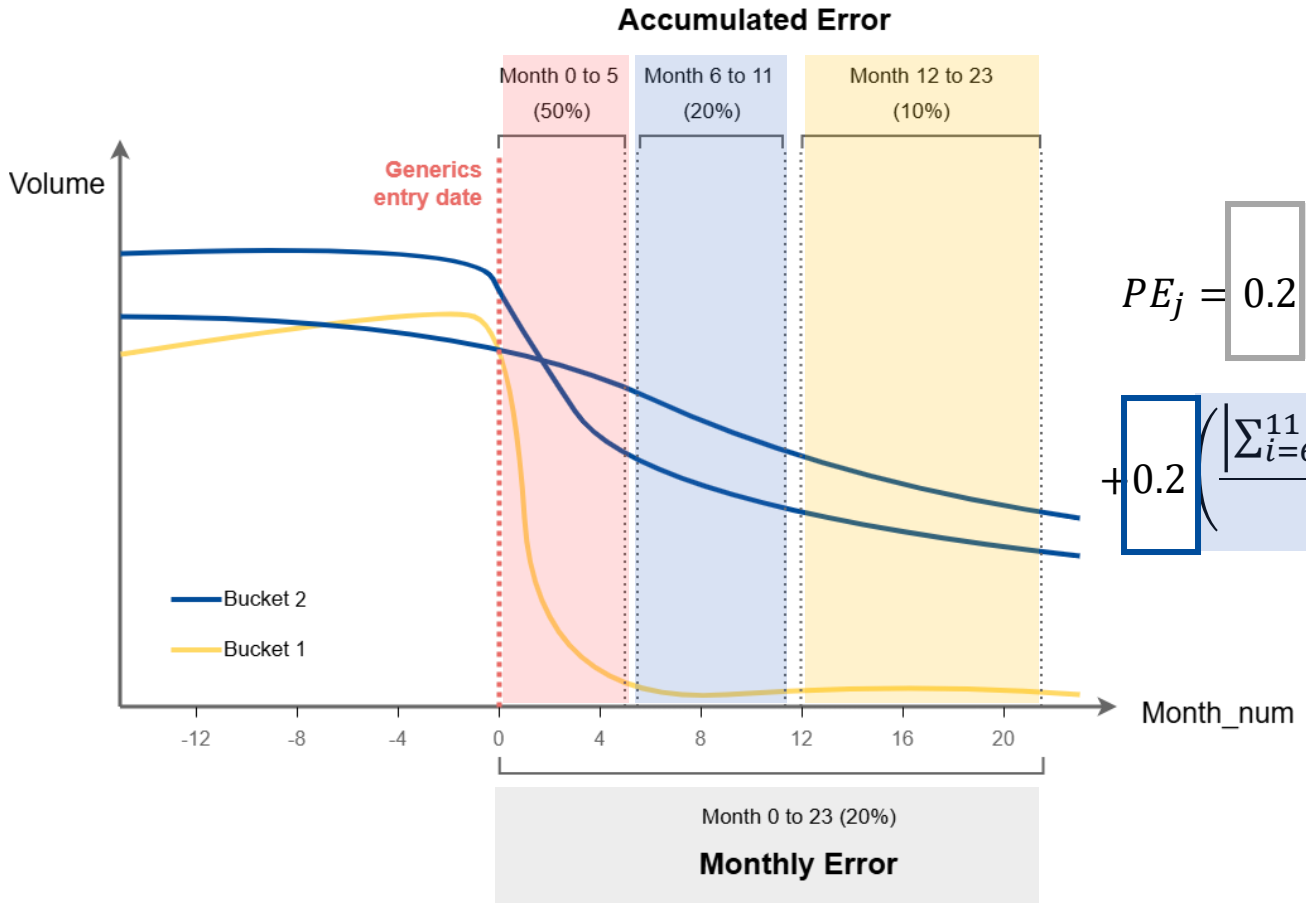
Metric: Prediction Error (Phase 1-a)



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

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

Metric: Prediction Error (Phase 1-a)



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

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

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

Metric: Prediction Error (Phase 1-a)

Prediction Error per country brand (PE_j):

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

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

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.

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

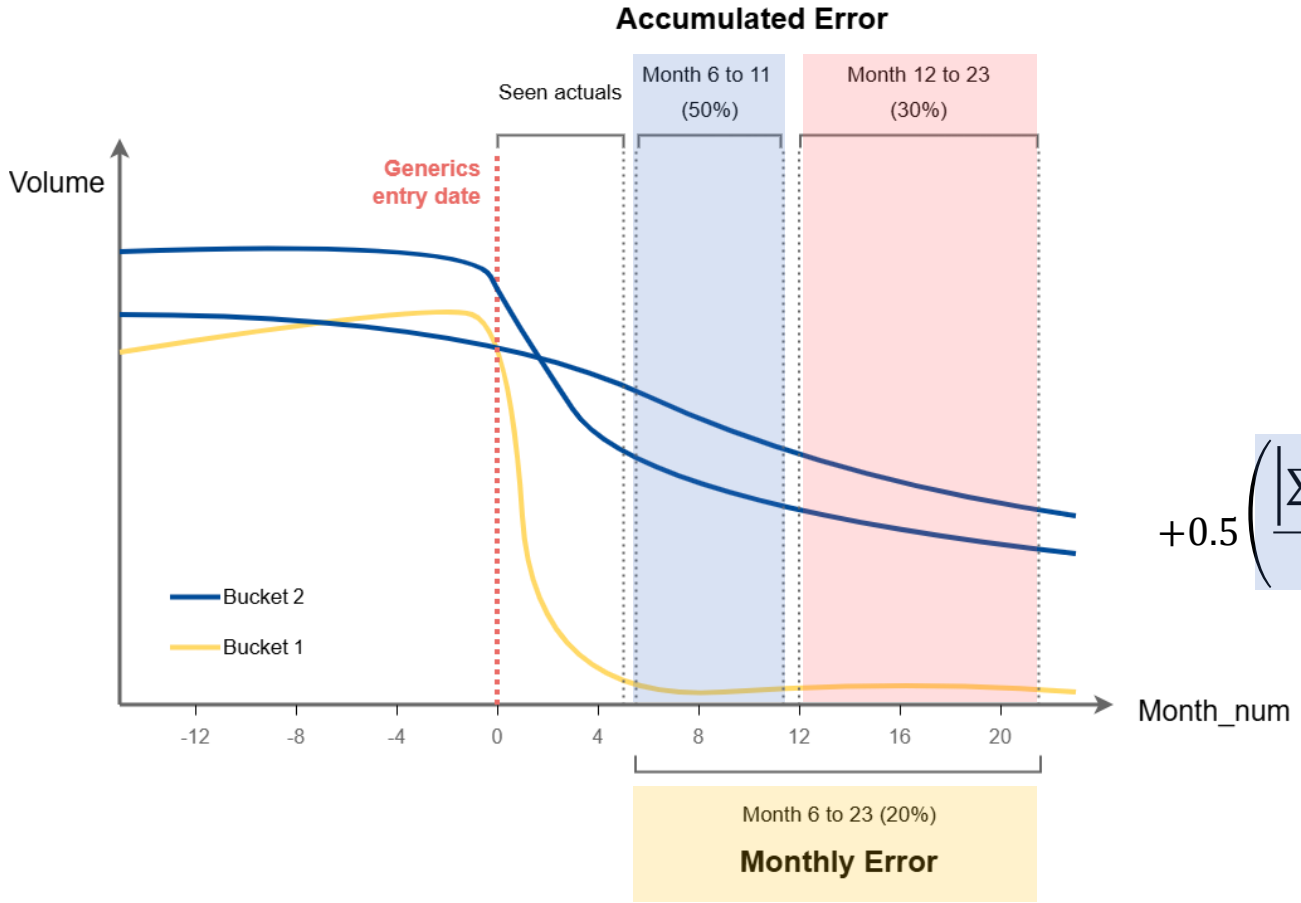
Metric: Prediction Error (Phase 1-b)

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)



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

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

Metric: Prediction Error (Phase 1-b)

Prediction Error per country brand (PE_j):

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

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.

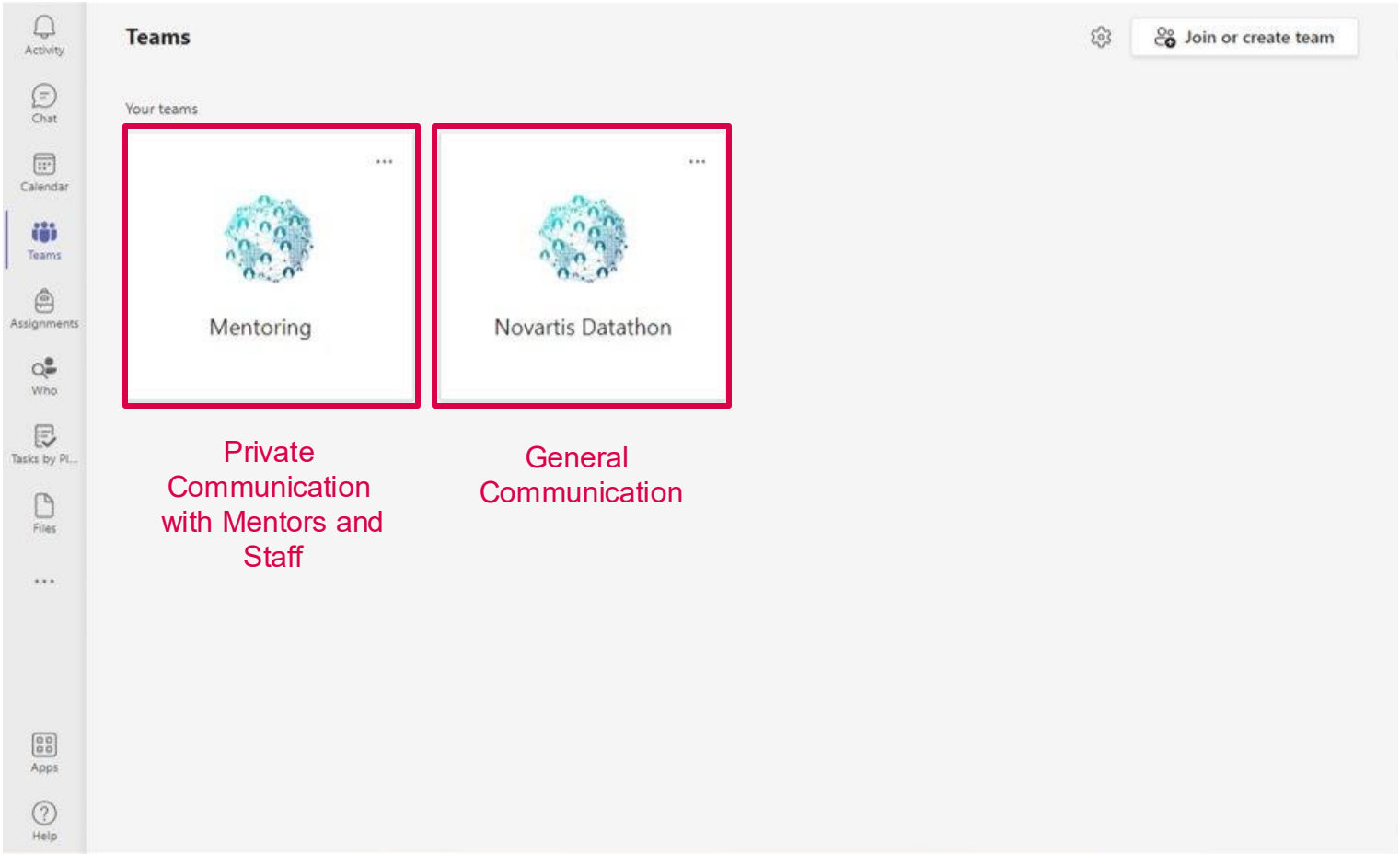
$$PE = \frac{2}{n_{B1}}\sum_{j=1}^{n_{B1}} PE_{j,B1} + \frac{1}{n_{B2}}\sum_{j=1}^{n_{B2}} PE_{j,B2}$$

Technical part

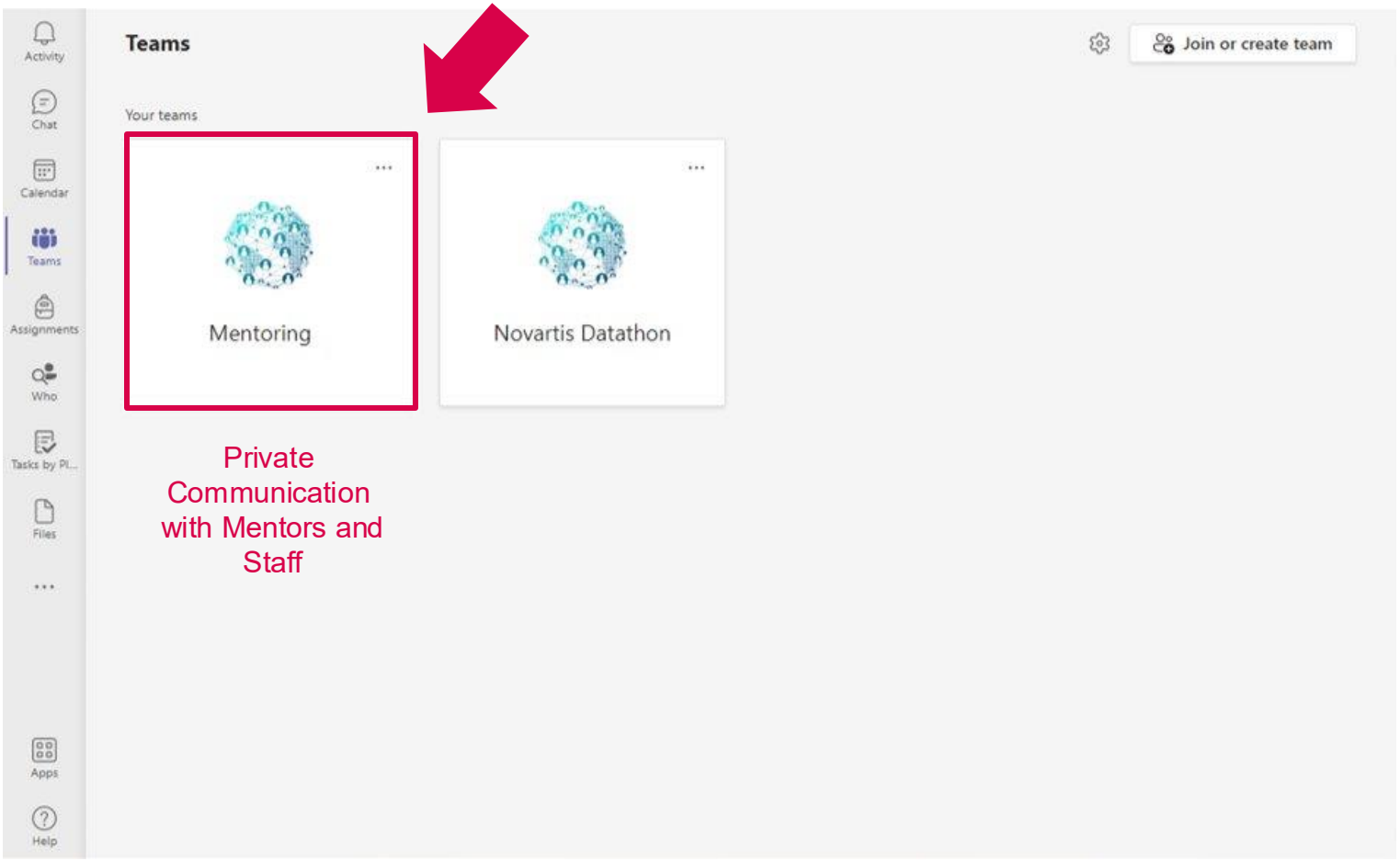
Communication: **Microsoft Teams**

Upload the results: **Datathon Platform**

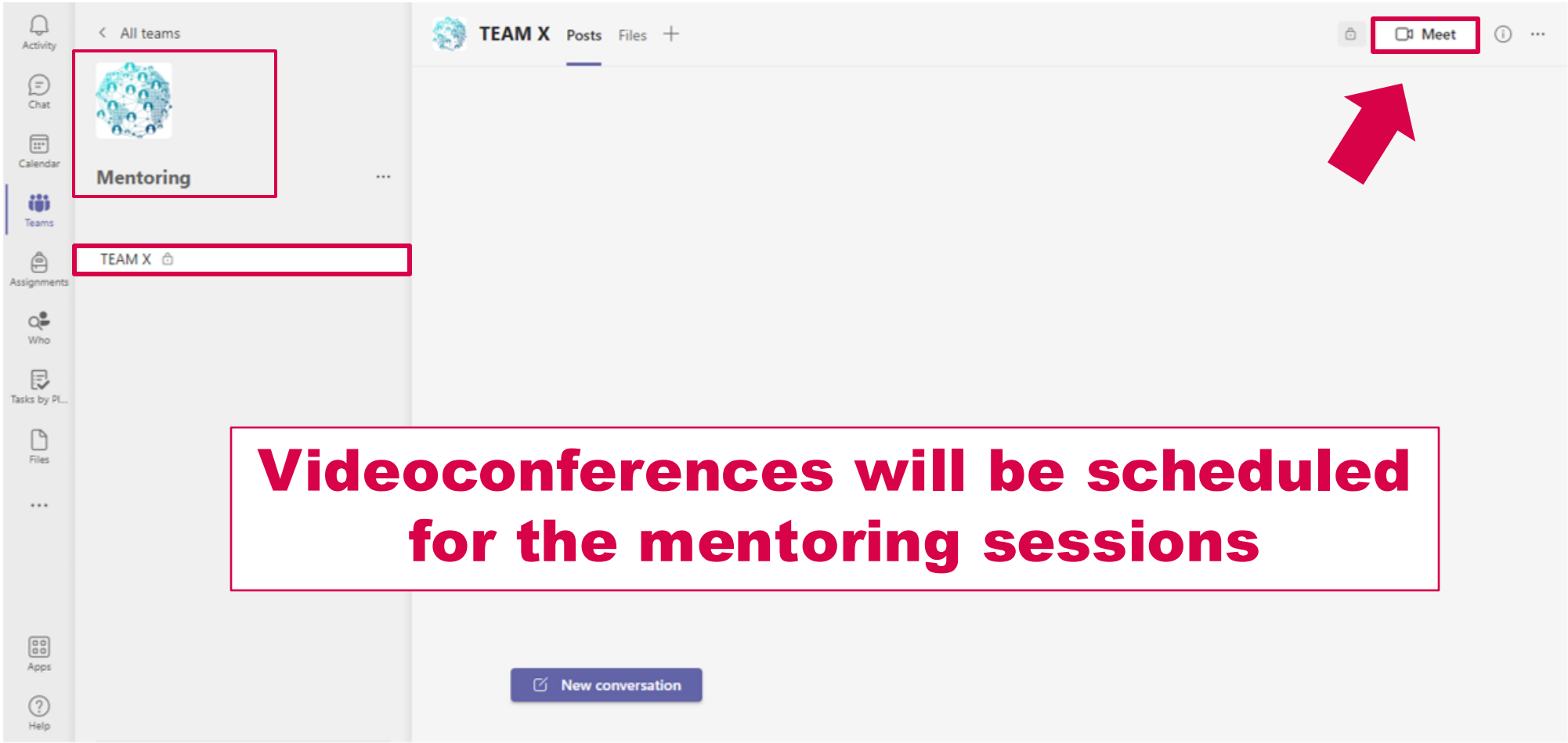
Communication Channel



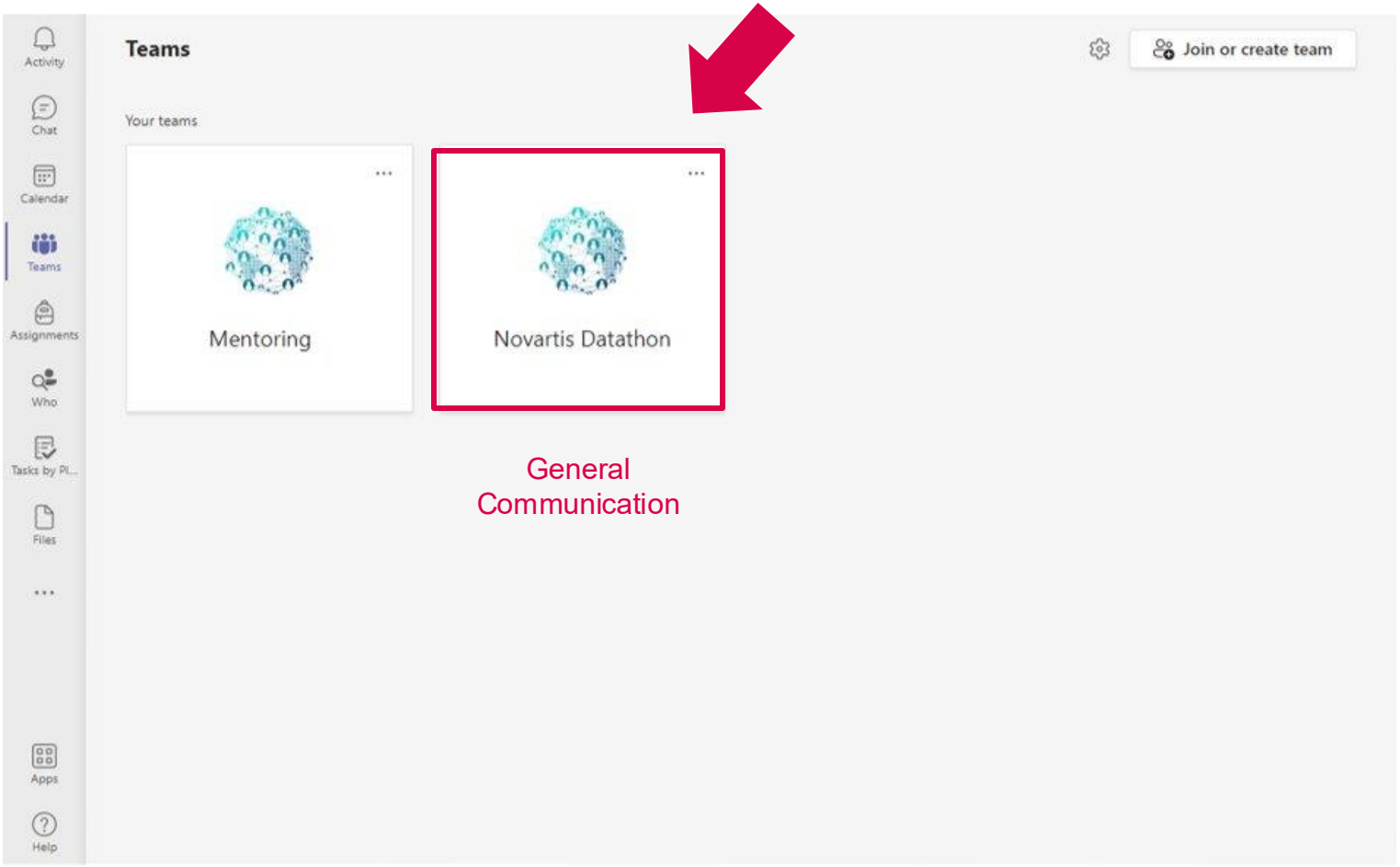
Communication Channel



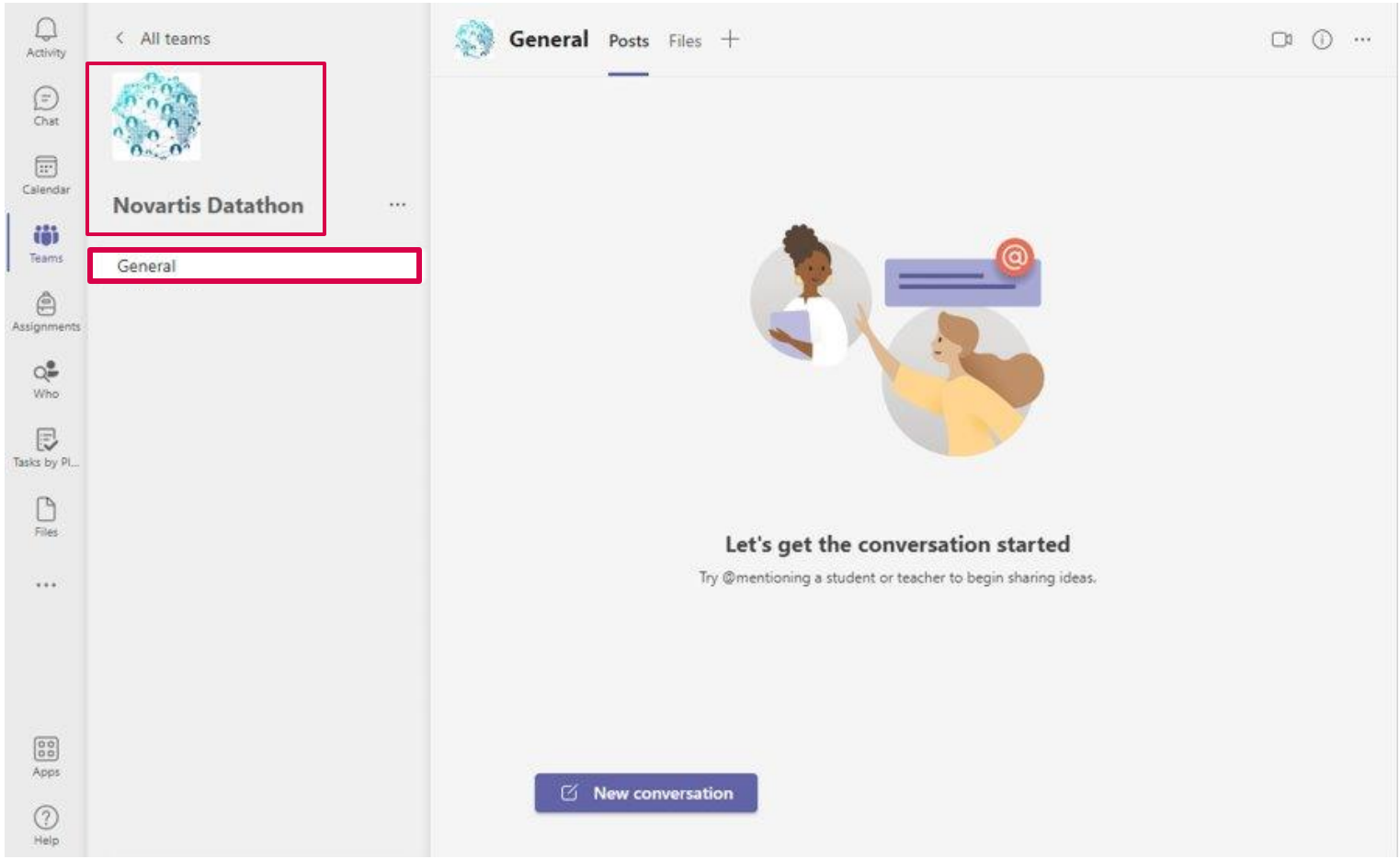
Communication Channel – Private communication



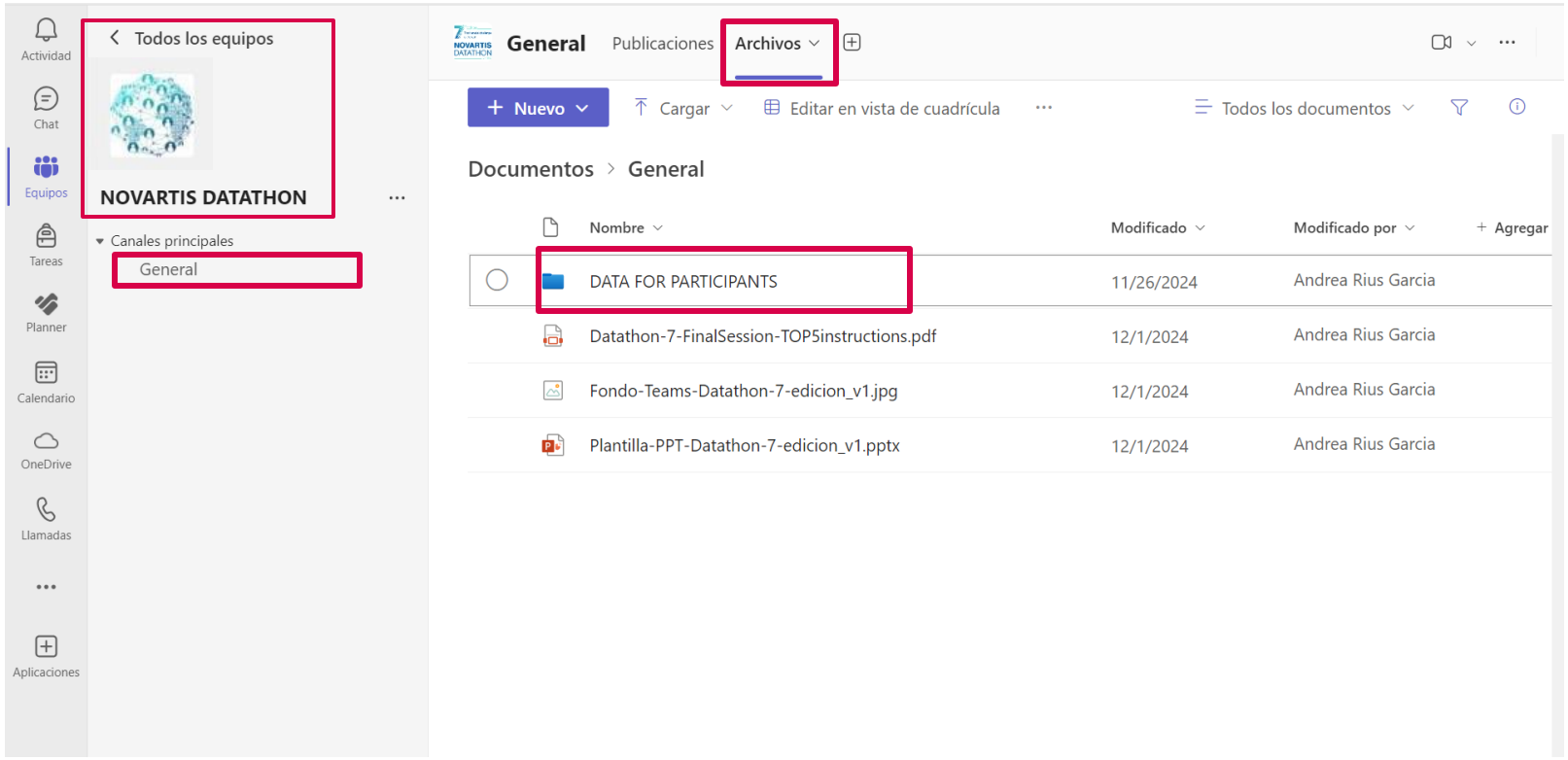
Communication Channel



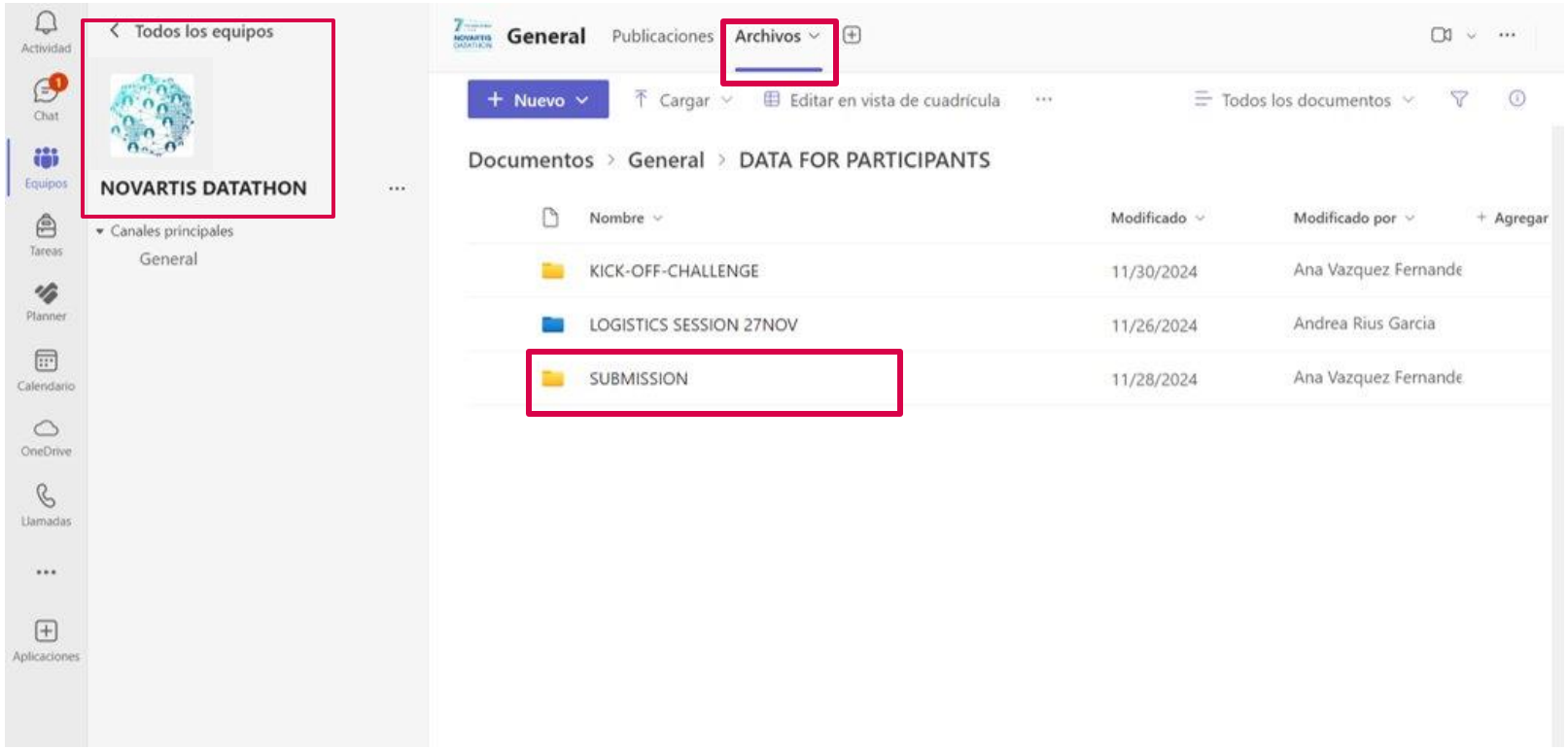
Communication Channel – General communication



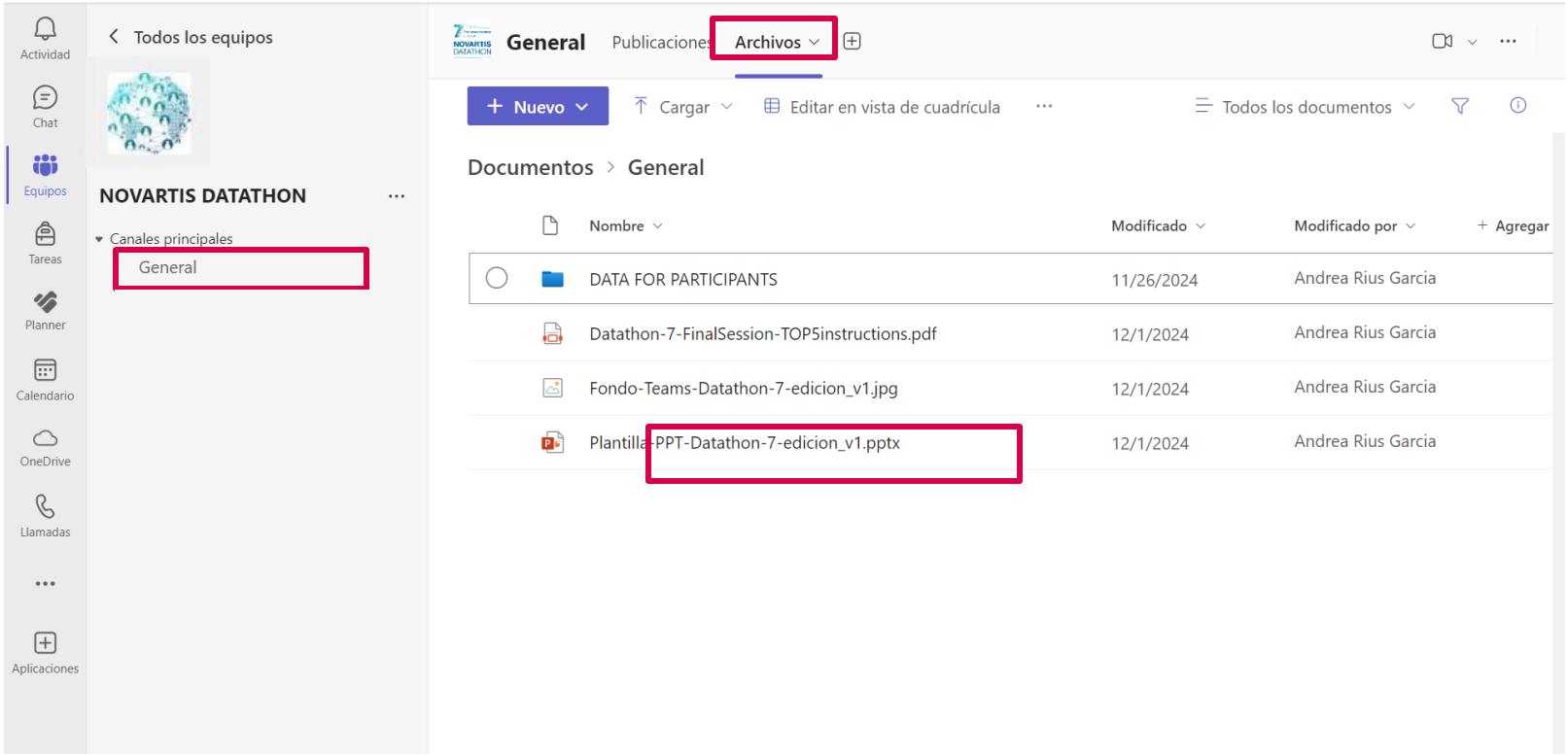
Download data



Download data



Submit presentation & code TOP 5

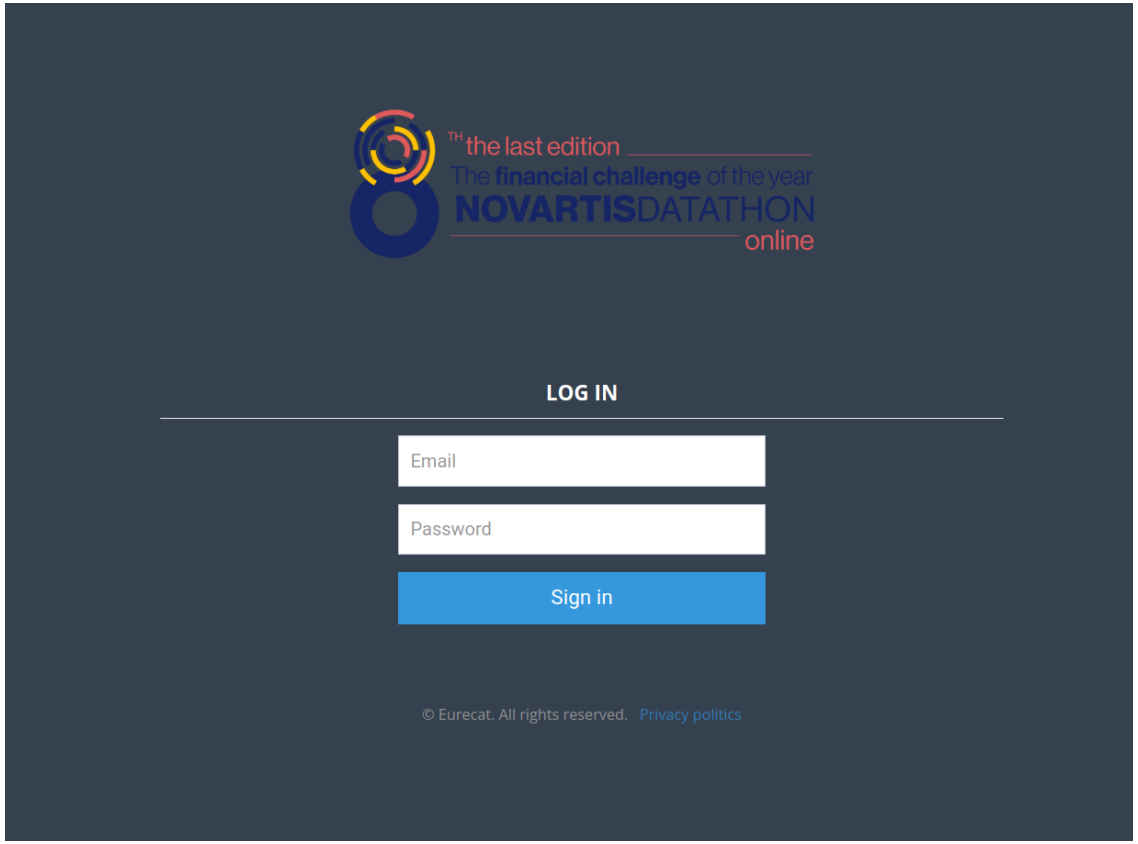


Platform login

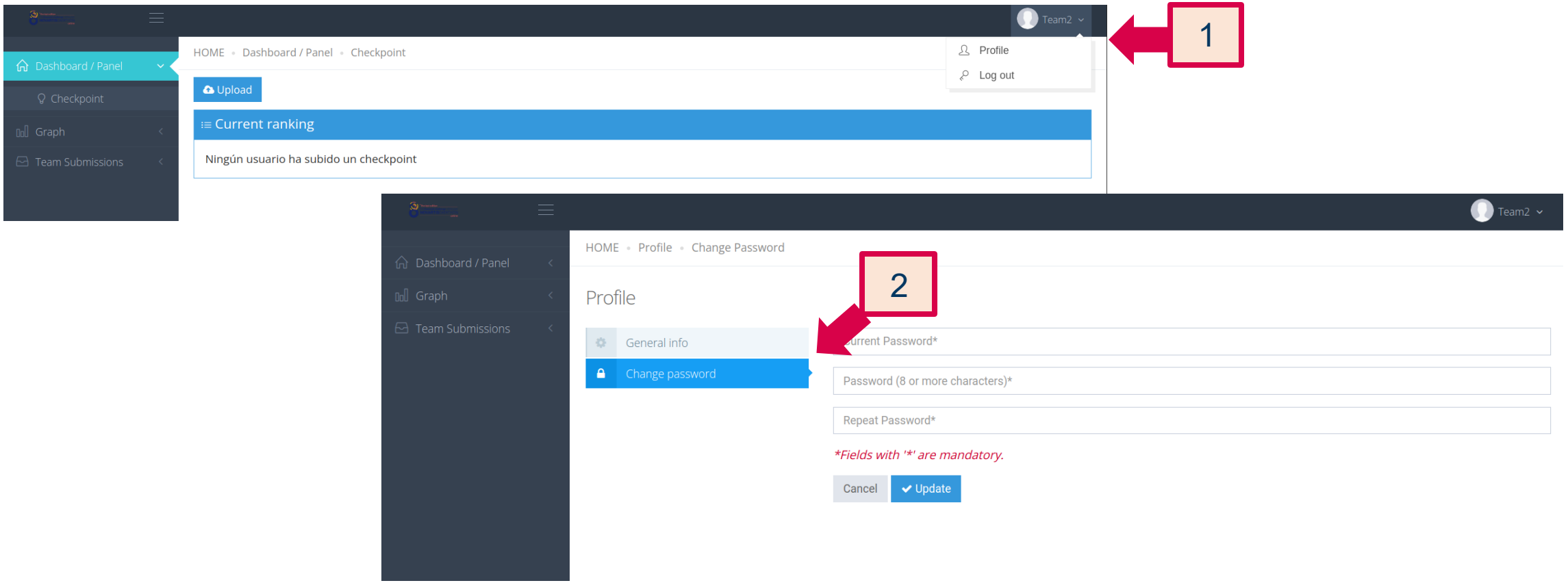
Credentials

user: teamX@novartisdatathon

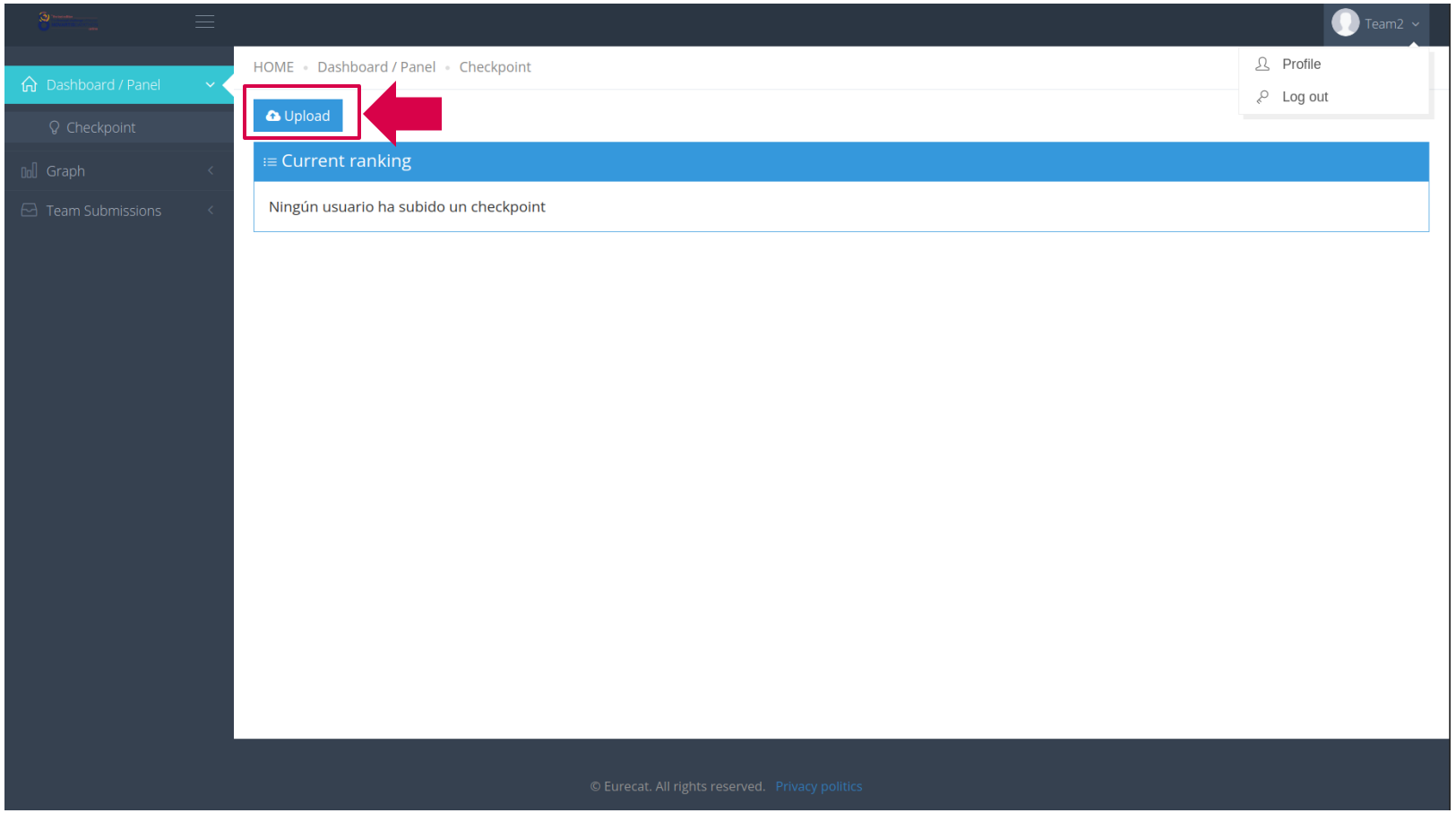
password: pwdteamX



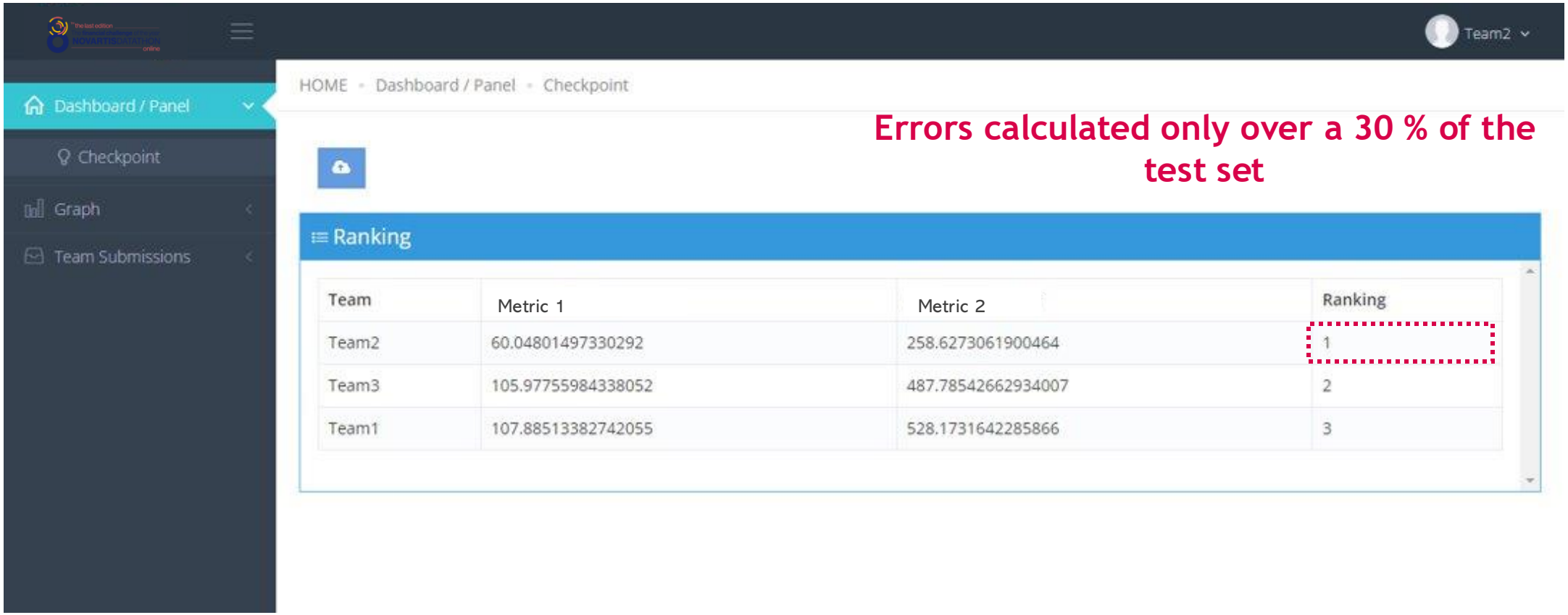
How to submit results – Change the password



How to submit results – Submission



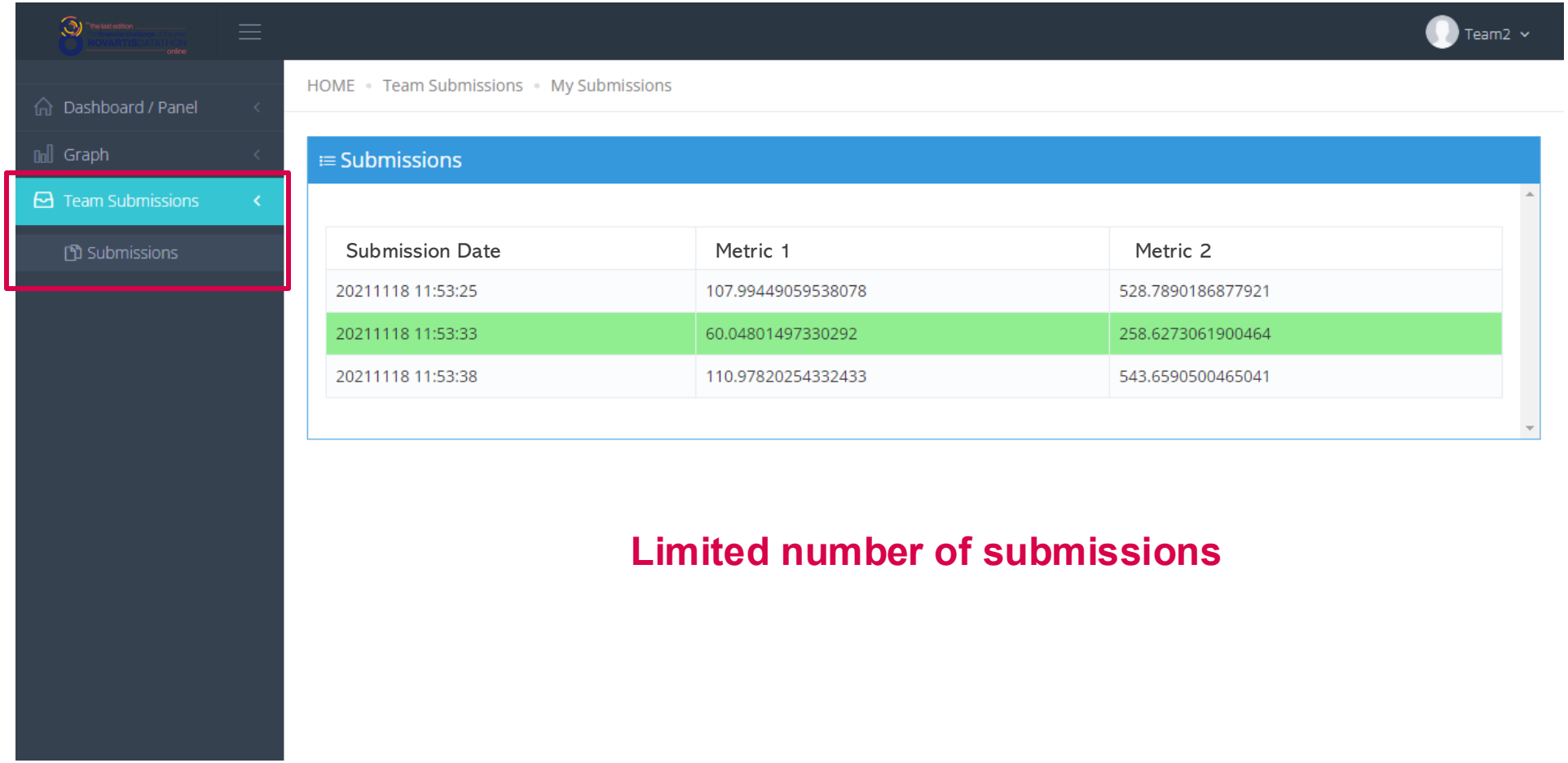
How to submit results – Ranking checkpoints



Errors calculated only over a 30 % of the test set

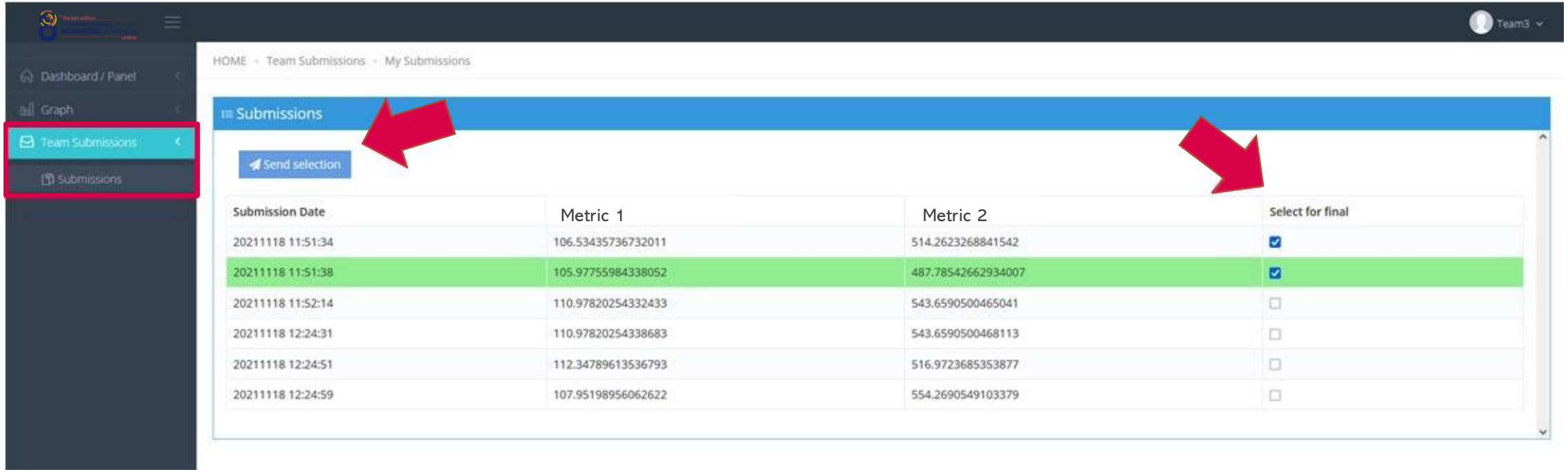
Your best submission is shown

How to submit results – History of submissions



Limited number of submissions

How to submit results – Final submission (last hour)



30th Nov between **9:30am** and **10:30am** *:

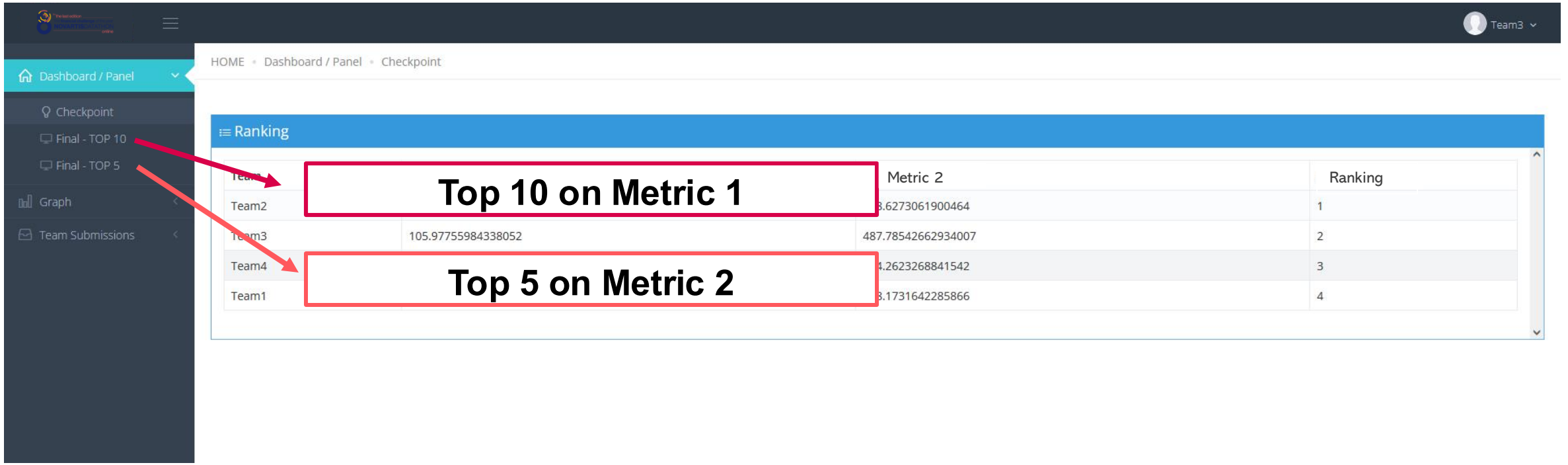
Select a **maximum** number of submissions

*Central European Time – Barcelona, UTC +1h

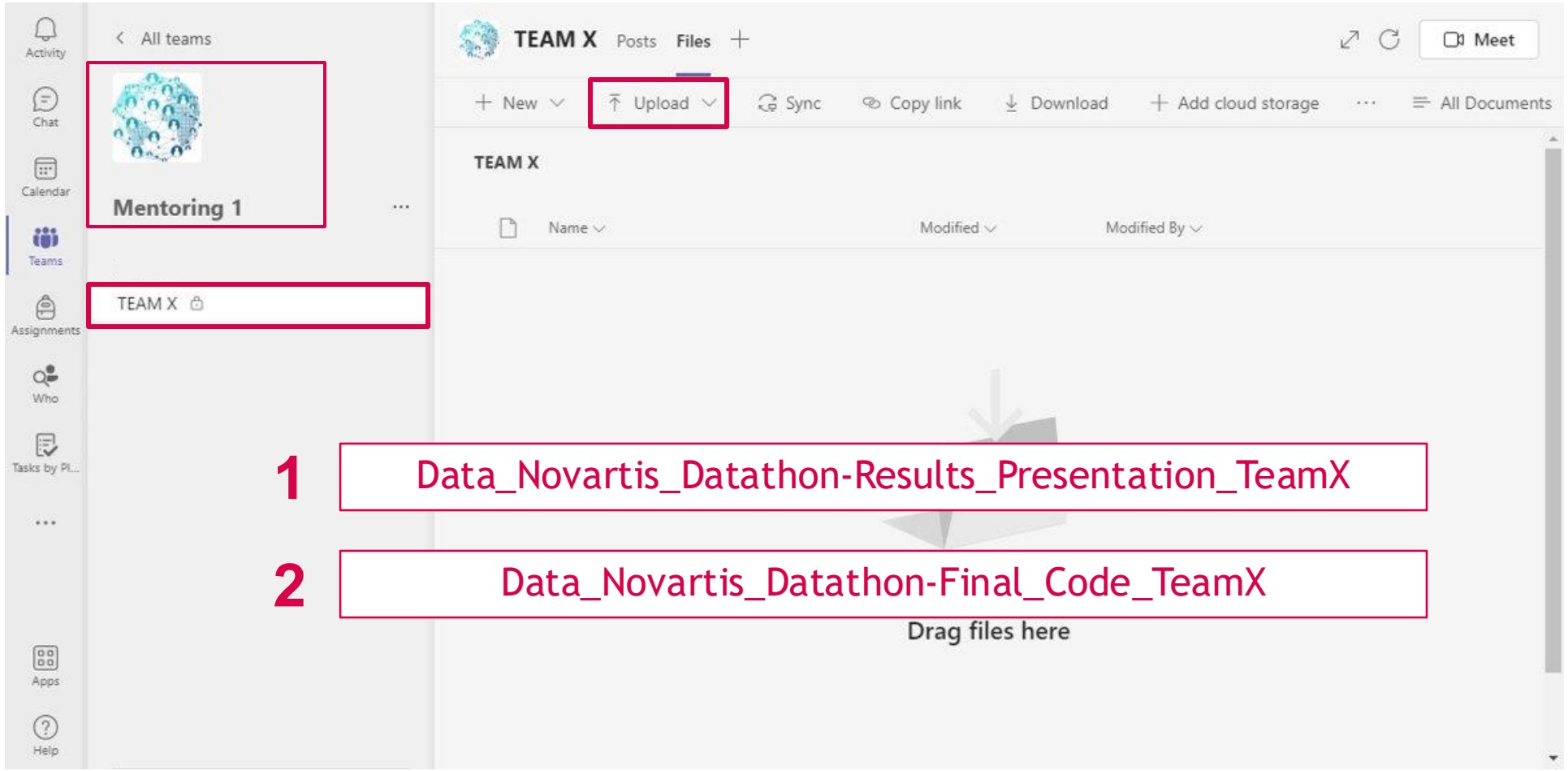
How to submit results

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

Show results



Submit presentation & code TOP 5



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

NOVARTIS

Thank you and good luck!