The **financial challenge** of the year

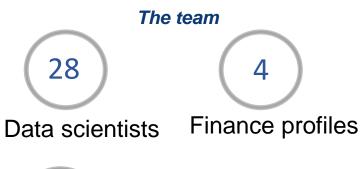
NOVARTISDATATHON

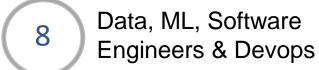


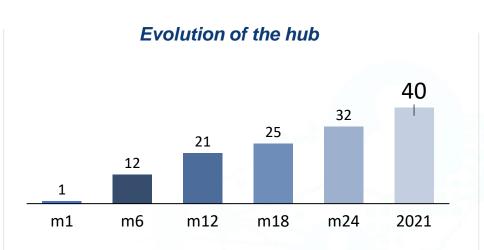


The BCN Digital Finance Hub

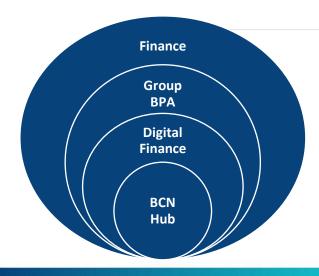












Our position at Novartis

- Innovative Medicine
- Global Drug Development
- Technical Operations

Background

22% Mathematics & Statistics

23% **Computer Science**

17% Engineering

19% **Economics**

19% Physics & others

4th edition _______ The financial challenge of the year NOVARTISDATATHON online

Our mission

Reimagine medicine to improve people's lives

- We are a medicine-focused company powered by advanced therapy platforms and artificial intelligence.
- R&D of new products of our Innovative Medicine Division can take 10 to 15 years from discovery to commercial product launch.

New launches



- The first months after a product is released on the market are crucial.
- We aim to reach as many patients as possible in a short period of time.
- To reach many patients, we need to reach many HealthCare Professionals (HCPs).

Our field force is the key.

Datathon question



What is the impact of field force on the sales of new launches?

Datathon challenge



1. Technical challenge

For the two products in scope, participants are required to provide **14 months of sales forecast** after the launch date together with **80% confidence intervals** for the given predictions.

2. Business challenge

All teams that present in front of the Jury will be asked to provide a detailed analysis exploring the impact of provided features on the sales growth, with a **focus** on the **field force** activities.

Datathon criteria



The winner will be selected according to three phases with different criteria.

1. Accuracy: a quantitative metric is adopted to compute the accuracy of the sales forecasting you are going to provide.

This metric will be used to select the **top 10 teams** with the lowest error in the sales forecast.

2. Confidence: a second quantitative metric is used to measure the accuracy of the provided confidence intervals (upper and lower bounds for the sales).

The **5 teams** with the smallest error in this metric will be selected to present in front of the Jury.

3. Jury's criteria: the members of the Jury have both technical and business background, and they can ask questions about any aspects of the solution.

At the end, the Jury decides the 3 winners of the Datathon.



- month
- region
- brand: anonymized name of the brand
- total sales corresponding to the given month, region and brand

month	region	brand	sales
2020-01	region_0	brand_1	0.00
2020-01	region_0	brand_2	0.00
2020-01	region_0	brand_3	65007.49
2020-01	region_0	brand_12_market	509023.69
2020-01	region_0	brand_3_market	940469.05



- month
- region
- brand: anonymized name of the brand
 - brand_1, brand_2: two newly-launched brands

month	region	brand	sales
2020-01	region_0	brand_1	0.00
2020-01	region_0	brand_2	0.00
2020-01	region_0	brand_3	65007.49
2020-01	region_0	brand_12_market	509023.69
2020-01	region_0	brand_3_market	940469.05



- month
- region
- brand: anonymized name of the brand
 - brand_3: another Novartis brand that is no longer promoted

month	region	brand	sales
2020-01	region_0	brand_1	0.00
2020-01	region_0	brand_2	0.00
2020-01	region_0	brand_3	65007.49
2020-01	region_0	brand_12_market	509023.69
2020-01	region_0	brand_3_market	940469.05



- month
- region
- brand: anonymized name of the brand
 - brand_12_market: market potential of brand_1 and brand 2
 - brand_3_market: market potential of brand_3

month	region	brand	sales
2020-01	region_0	brand_1	0.00
2020-01	region_0	brand_2	0.00
2020-01	region_0	brand_3	65007.49
2020-01	region_0	brand_12_market	509023.69
2020-01	region_0	brand_3_market	940469.05



Sales data are split to train (public) and test (private) part.

Train data

- Regions 0, 1, ..., 150:
 - sales information for brand_1, brand_2, brand_3, brand_12_market and brand_3_market



Sales data are split to train (public) and test (private) part.

Train data

- Regions 0, 1, ..., 150:
 - sales information for brand_1, brand_2, brand_3, brand_12_market and brand_3_market
- Regions 151, 152, ..., 200:
 - sales information only for brand_3, brand_12_market and brand_3_market



Sales data are split to train (public) and test (private) part.

Train data

- Regions 0, 1, ..., 150:
 - sales information for brand_1, brand_2, brand_3, brand_12_market and brand_3_market
- Regions 151, 152, ..., 200:
 - sales information only for brand_3, brand_12_market and brand_3_market

Test data

- Information on sales for brand_1 and brand_2 in regions 151, 152, ..., 200.
- These are the sales that you are required to predict; these data are not publicly available.

Data provided – HCPs



- HCP = **H**ealth **C**are **P**rofessional
- 5 specializations can prescribe brand_1 and brand_2:

	region	Internal medicine	Internal medicine / pneumology	General practicioner	Internal medicine and general practicioner	Pediatrician
0	region_0	315	18	264	2	82
1	region_1	239	7	273	1	42
2	region_2	153	5	182	2	47
3	region_3	181	11	206	0	30
4	region_4	151	9	199	0	33

Data provided – prioritized HCPs



- Tier denotes the priority level, 1 has highest priority
- Non-prioritized HCPs are not included in this table.

	hcp	tier	region	specialty
0	83	2	region_59	Internal medicine
1	112	1	region_103	Internal medicine
2	307	2	region_33	Internal medicine / pneumology
3	387	2	region_97	Internal medicine / pneumology
4	426	1	region_118	General practicioner

Data provided – glossary

Brands: product names

- **brand_1** and **brand_2** are the two newly-launched brands. Both have the same indication, that is both are targeting patients with the same respiratory disease.
- brand_3 is another Novartis brand that has indication different from brand_1 and brand_2. It has
 stable position at the market and its sales are not affected by the promotional activities relevant to
 brand_1 and brand_2.

Market potential: total sales of all brands by Novartis and competition targeting patients with the same disease.

- **brand_12_market** is the market potential for brand_1 and brand_2. Both brands have the same indication, which means that they have the same market potential.
- brand_3_market is the market potential for brand_3.

Note that market potentials contain the sales of the brand itself, for instance brand_12_market covers the sales of brand_1 and brand_2.

Data provided – glossary

Prioritized HCPs

- Each HCP is assigned a tier this is internal Novartis code denoting the priority of the HCP. Tier 1
 means higher priority than tier 2 and tier 2 is higher priority than tier 3.
- The numbers of HCPs in the high priority table grouped by region and specialty are smaller than in the previous table – this is because the high priority table does not contain detailed information on the non-prioritized HCPs.

Data provided - Activity



Data on promotional activities (visits) performed by Novartis field-force representatives (reps).

- **hcp**: anonymized code of the visited HCP
- specialty of the HCP
- **region** where the activity has been done (few HCPs are active in more than one region)
- **brand** discussed during the visit
- month of the visit
- communication channel: f2f (face-to-face), phone, video or other (mainly emails with personalized content)
- count: number of visits of this type

hcp	specialty	region	brand	month	channel	count
39972	Internal medicine / pneumology	region_61	brand_1	2020-05	video	1
64026	General practicioner	region_168	brand_2	2020-05	phone	1
27892	Internal medicine	region_26	brand_2	2020-06	phone	1
33500	General practicioner	region_26	brand_2	2020-06	phone	1

Data provided - RTEs



Information about Rep-Triggered Emails

Rep Triggered Emails are emails with a pre-defined content prepared by the marketing team. They are triggered by the rep who decides which content is relevant to each HCP and when those emails are sent.

- hcp, specialty, region
- **Emails** are related to brand_1 and brand_2
- email_type: general (general information about the respiratory disease not mentioning the product name) or product_related (directly mentions the product name)

hcp	specialty	region	brand	email_type	content_id	no. openings	no. clicks	time_sent	time_last_opened	time_last_clicked
171	General practicioner	region_132	brand_2	product_related	a3D6f0000019fLWEAY	3	0	2020-07- 02 12:07:00	2020-07-02 18:53:00	NaT
171	General practicioner	region_132	brand_1	product_related	a3D6f0000019jDTEAY	1	0	2020-08- 17 12:33:00	2020-08-17 13:50:00	NaT
171	General practicioner	region_132	brand_1	product_related	a3D6f000000MtbrEAC	2	0	2020-09- 08 10:55:00	2020-09-08 20:40:00	NaT
171	General practicioner	region_132	brand_1	general	a3D6f000000Mpy8EAC,a3D6f0000019ivnEAA	1	0	2020-08- 12 17:31:00	2020-08-12 17:49:00	NaT
173	Internal medicine	region_46	brand_1	general	a3D6f000001HndwEAC	0	0	2021-02- 23 13:45:00	NaT	NaT

Data provided - RTEs



- **content_id**: each type of content designed by the marketing team is assigned a unique id.
- no. openings: number of times the email has been opened by the HCP
- **no. clicks:** number of times the HCP clicked on the link in the email
- time_sent: time stamp of when the email has been sent
- time_last_opened: last time the link in the email has been opened
- time_last_clicked: last time the content of the link has been clicked

hcp	specialty	region	brand	email_type	content_id	no. openings	no. clicks	time_sent	time_last_opened	time_last_clicked
171	General practicioner	region_132	brand_2	product_related	a3D6f0000019fLWEAY	3	0	2020-07- 02 12:07:00	2020-07-02 18:53:00	NaT
171	General practicioner	region_132	brand_1	product_related	a3D6f0000019jDTEAY	1	0	2020-08- 17 12:33:00	2020-08-17 13:50:00	NaT
171	General practicioner	region_132	brand_1	product_related	a3D6f000000MtbrEAC	2	0	2020-09- 08 10:55:00	2020-09-08 20:40:00	NaT
171	General practicioner	region_132	brand_1	general	a3D6f000000Mpy8EAC,a3D6f0000019ivnEAA	1	0	2020-08- 12 17:31:00	2020-08-12 17:49:00	NaT
173	Internal medicine	region_46	brand_1	general	a3D6f000001HndwEAC	0	0	2021-02- 23 13:45:00	NaT	NaT

Data provided - Regions



Information about the different regions

- population in the region
- area of the region
- average per capita income in 2016 and 2018

	region	population	area	pci16	pci18
0	region_0	381958	1372.440545	38799.972381	40214.895762
1	region_1	431383	4701.726213	45377.708041	47137.854237
2	region_2	318026	2574.943849	42578.018793	44456.389595
3	region_3	343861	3639.373478	46856.492170	48564.228841
4	region_4	330916	3926.381089	44462.421679	46388.545070

Insights and hints



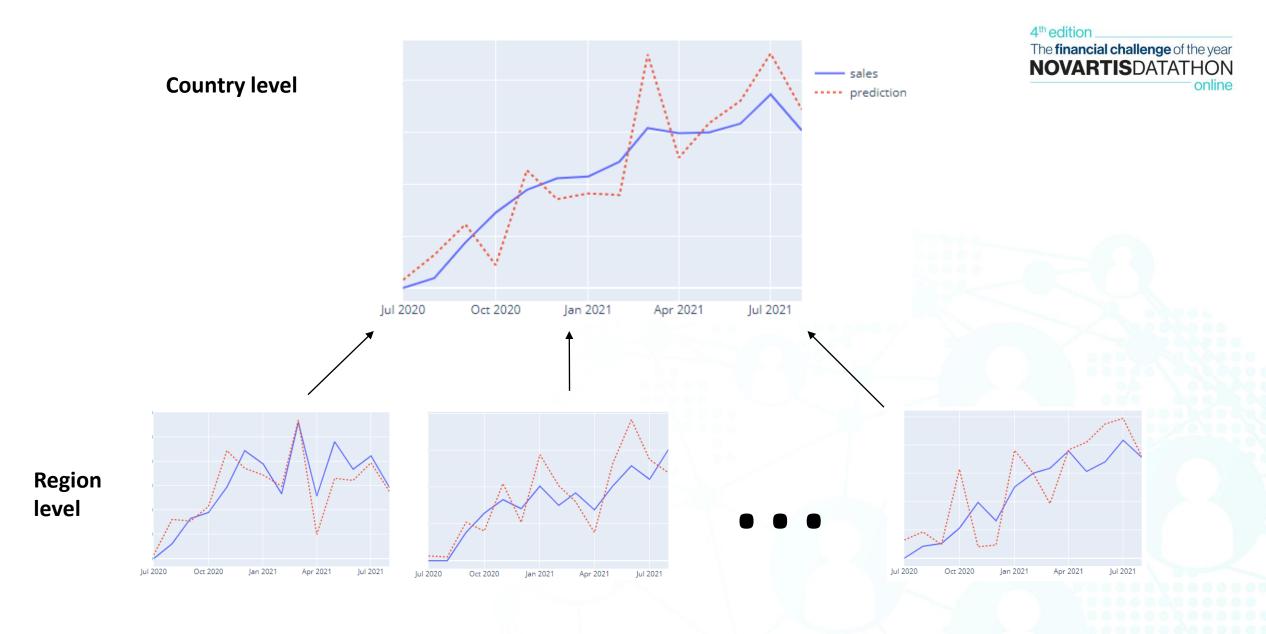
- Among all the activities we have, how should sales representatives distribute their efforts? Should they communicate by f2f or digital channel?
- The regions are different and in some of them, Novartis historically had stronger position than in others.

Accuracy metric



To compute the prediction error, the absolute difference between the monthly predicted values and the actual sales is evaluated for the two distinct brands.

- 1. Both errors of Brand_1 and Brand_2 are weighted equally.
- 2. For each brand, the metric is computed both at regional and country level.



Accuracy metric

Region level

 $y_{r,t}^b$ = true sales of brand b in region r month t $\hat{y}_{r,t}^b = \text{model predictions}$

Mean absolute error: region level

$$MAE_r^b = \frac{1}{14} \sum_{t} |y_{r,t}^b - \hat{y}_{r,t}^b|$$

Mean absolute error: aggregated

$$MAE^{b} = \frac{MAE_{c}^{b}}{\langle m_{c} \rangle} + \frac{1}{50} \sum_{r} \frac{MAE_{r}^{b}}{\langle m_{r} \rangle}$$

 $\langle m_c \rangle$, $\langle m_r \rangle$ = brand_12_market at country / region level averaged over the time period 01/2020 - 08/2021



Country level (country is represented by the regions present in the test dataset)

$$y_{c,t}^b = \sum_r y_{r,t}^b$$
 total sales of brand b
 $\hat{y}_{c,t}^b = \sum_r \hat{y}_{r,t}^b$ total sales predictions of brand b

Mean absolute error: country level

$$MAE_c^b = \frac{1}{14} \sum_{t} |y_{c,t}^b - \hat{y}_{c,t}^b|$$

Final metric:

$$MAE = 10000 \left(\frac{1}{2} MAE^{(1)} + \frac{1}{2} MAE^{(2)} \right)$$

Confidence interval metric



 $\hat{u}_{r,t}^b$ = upper prediction bound for region r, brand b, month t $\hat{\ell}_{r,t}^b$ = lower prediction bound for region r, brand b, month t

Given the prediction interval $\{\hat{\ell}_{r,t}^b, \hat{u}_{r,t}^b\}$, the confidence interval metric has two components:

1. Coverage – do the actual values fall inside the interval?

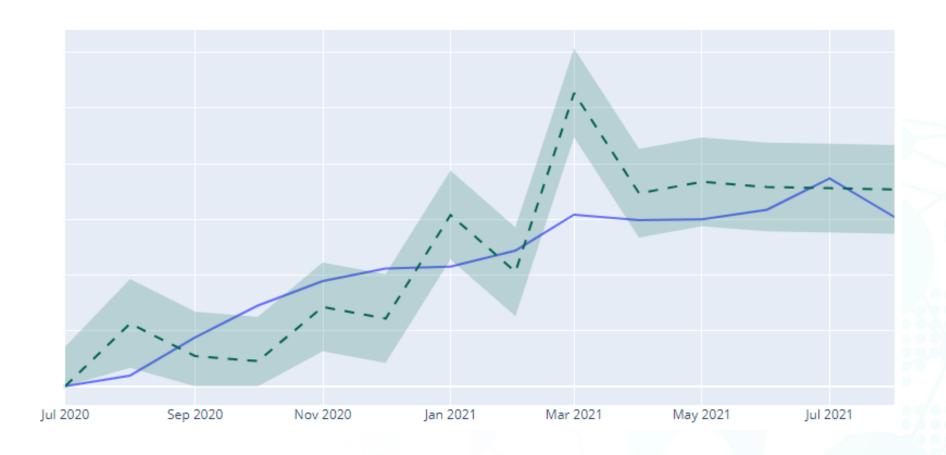
$$\hat{\ell}^b_{r,t} < y^b_{r,t} < \hat{u}^b_{r,t}$$

2. Width – wide intervals are penalized.

$$\left|\hat{u}_{r,t}^{b}-\hat{\ell}_{r,t}^{b}\right|$$

Confidence metric





Confidence metric formula



Confidence error for given region, brand and time

Interval width penalization

Real point above upper interval

Real point below lower interval

$$\Delta^{b}_{r,t} = \hat{u}^{b}_{r,t} - \hat{\ell}^{b}_{r,t} + \frac{2}{\alpha} \left[(y^{b}_{r,t} - \hat{u}^{b}_{r,t}) \Theta (y^{b}_{r,t} - \hat{u}^{b}_{r,t}) + (\hat{\ell}^{b}_{r,t} - y^{b}_{r,t}) \Theta (\hat{\ell}^{b}_{r,t} - y^{b}_{r,t}) \right]$$

Confidence error for given region and brand

$$\Delta_r^b = \frac{1}{14} \sum_t \Delta_{r,t}^b$$

Confidence error for given brand

$$\Delta^b = \frac{1}{50} \sum_r \frac{\Delta_r^b}{\langle m_r \rangle}$$

 α = confidence level; you are required to predict 80% confidence intervals, so $\alpha = 0.2$

Final confidence error:

$$\Delta = 10000 \left(\frac{1}{2} \Delta^{(1)} + \frac{1}{2} \Delta^{(2)} \right)$$

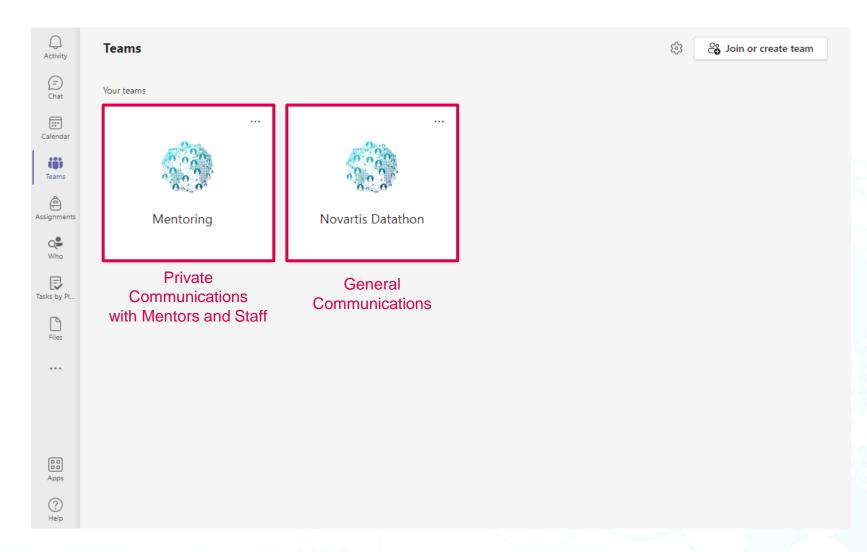
Technical part



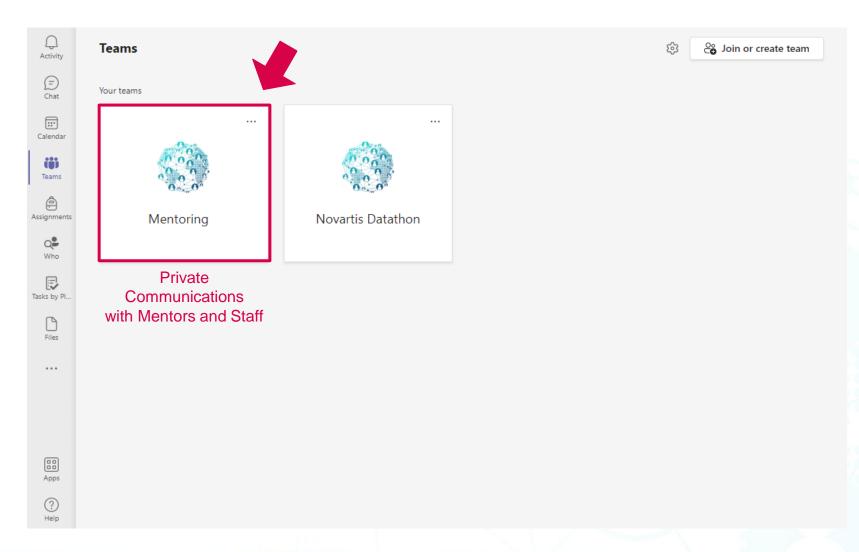
Communication: Microsoft Teams

Upload the results: Datathon Platform



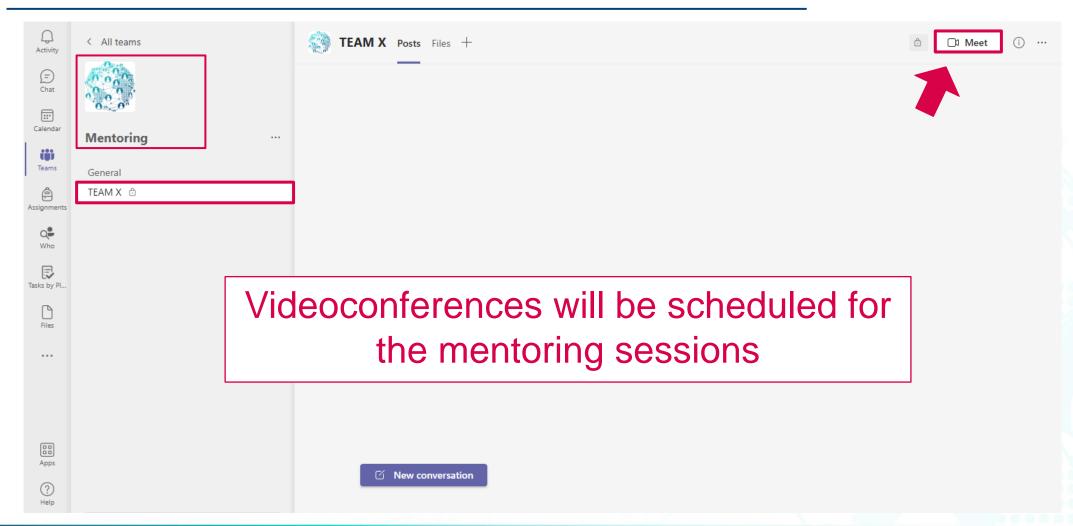




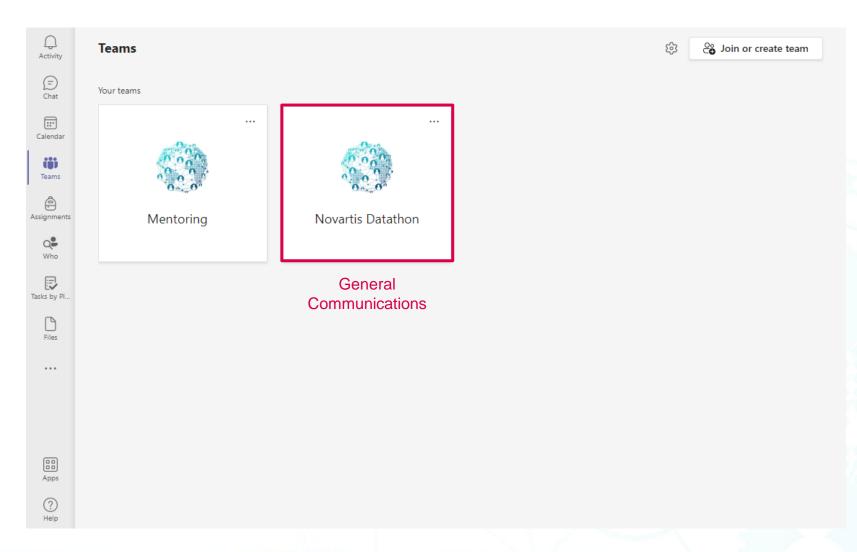




Private communications

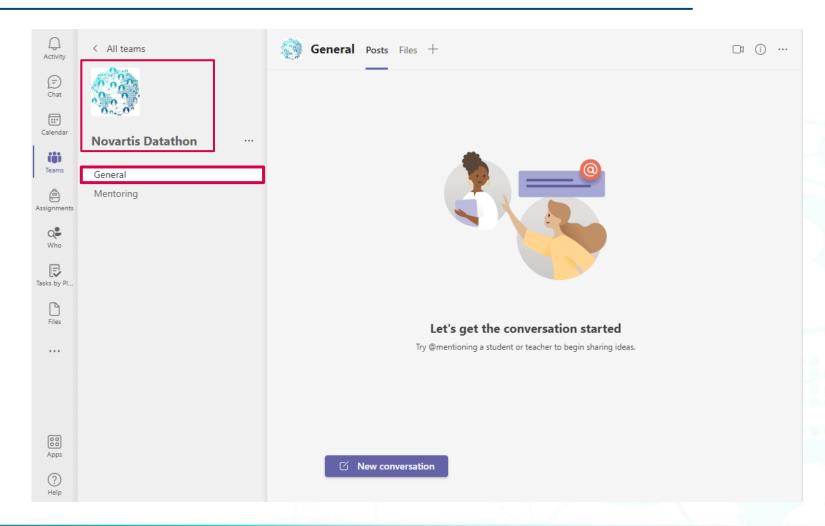






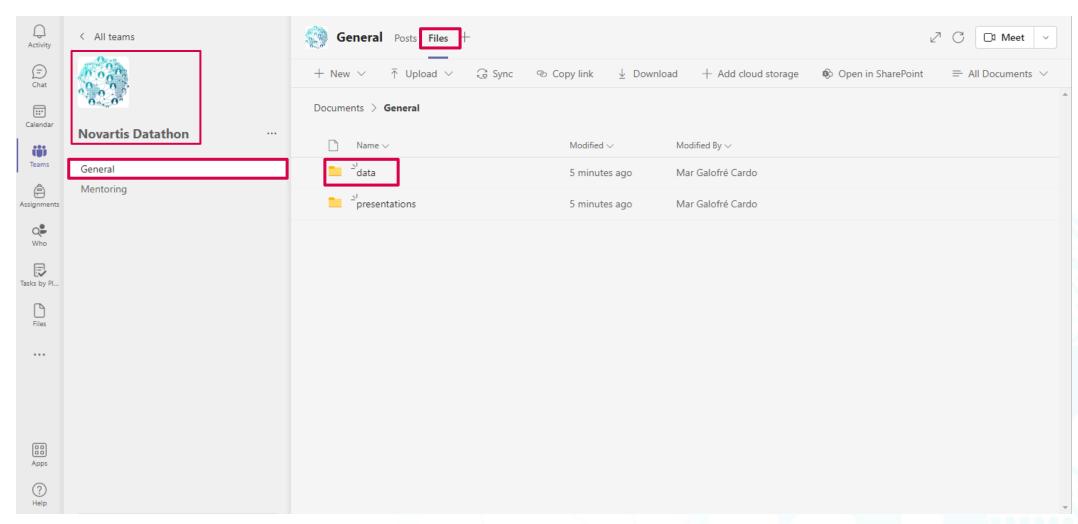
4th edition The **financial challenge** of the year **NOVARTIS**DATATHON

General communications



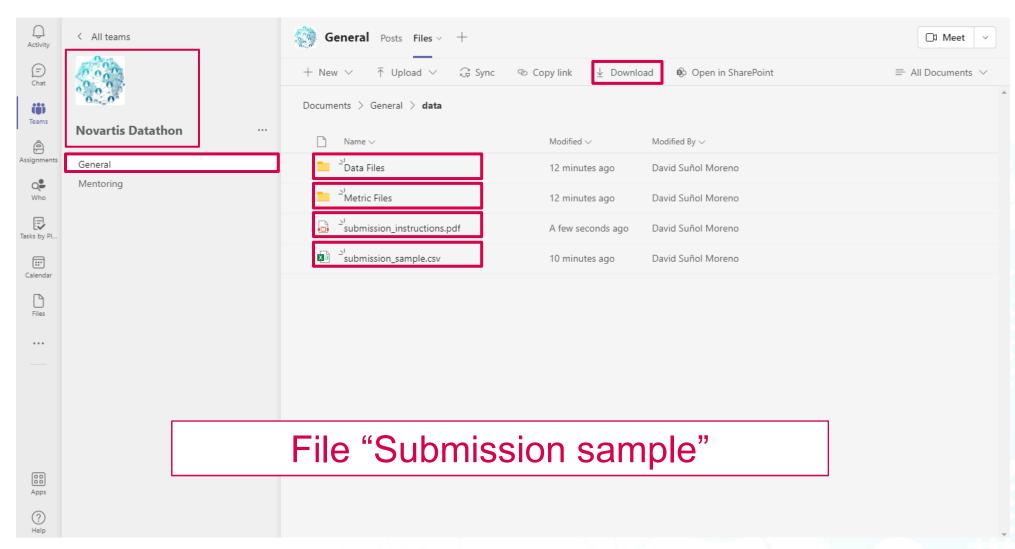
Download data





Download data

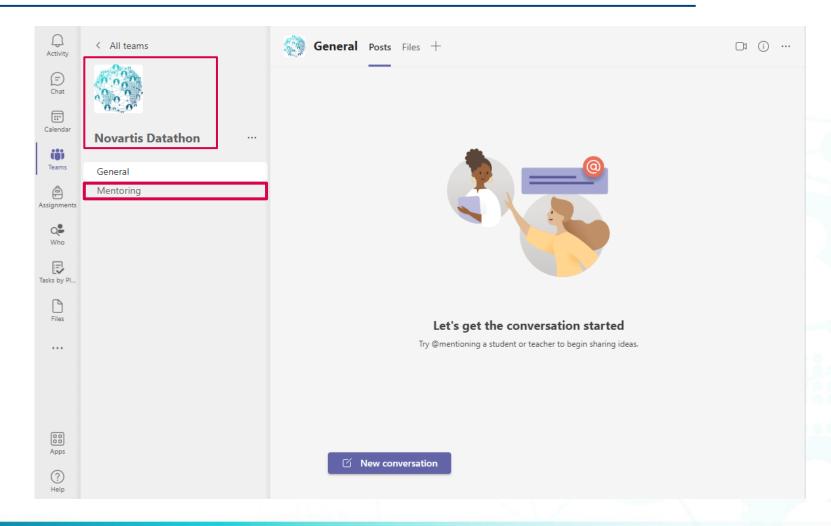




Communication Channel

4th edition The financial challenge of the year NOVARTISDATATHON online

General communications



Login in the platform

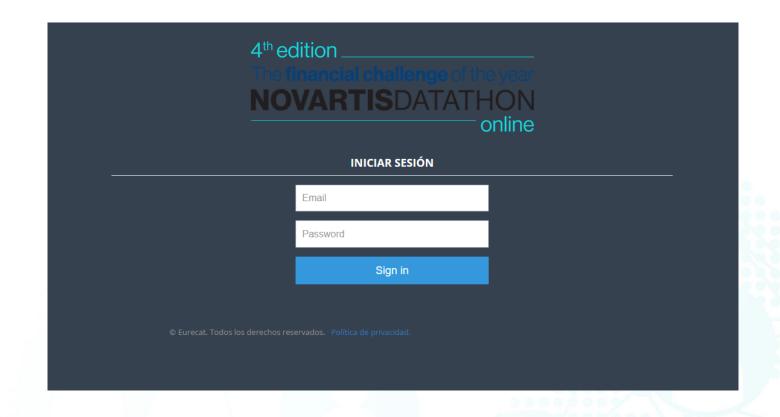


URL: http://84.88.76.50/

Credentials

user: teamX@novartisdatathon

password: pwdteamX



The **financial challenge** of the year **NOVARTIS**DATATHON

Submission structure

The **csv** you submit **must have**:

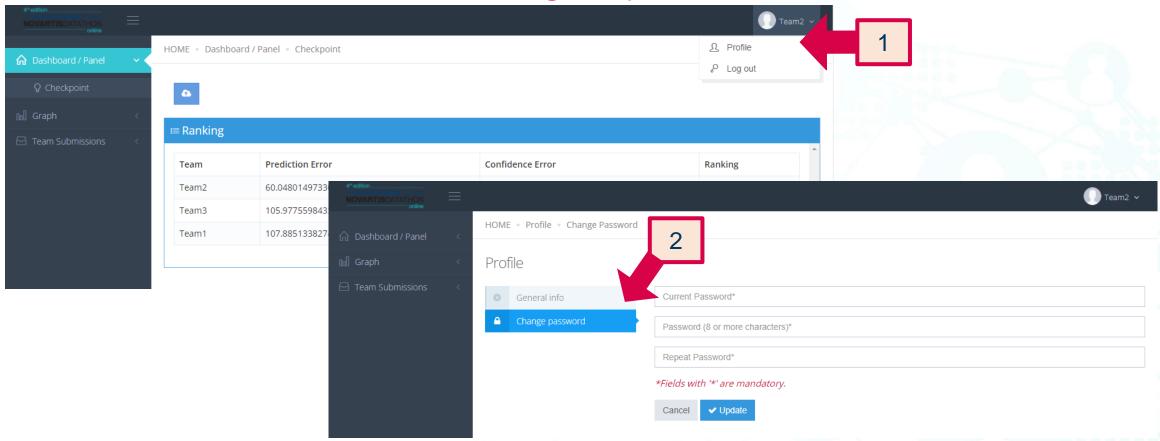
- a header
- same number of rows and columns as the test dataset / template
- all values (no NaNs allowed)
- comma-separated values
- points representing decimals

month	region	brand	sales	lower	upper
2020-07	region_151	brand_1	50.00	0.00	100.00
2020-07	region_151	brand_2	50.00	0.00	100.00
2020-07	region_152	brand_1	50.00	0.00	100.00
2020-07	region_152	brand_2	50.00	0.00	100.00
2020-07	region_153	brand_1	50.00	0.00	100.00
2020-07	region_153	brand_2	50.00	0.00	100.00
2020-07	region_154	brand_1	50.00	0.00	100.00
2020-07	region_154	brand_2	50.00	0.00	100.00
2020-07	region_155	brand_1	50.00	0.00	100.00
2020-07	region_155	brand_2	50.00	0.00	100.00
2020-07	region_156	brand_1	50.00	0.00	100.00
2020-07	region_156	brand_2	50.00	0.00	100.00
2020-07	region_157	brand_1	50.00	0.00	100.00
2020-07	region_157	brand_2	50.00	0.00	100.00
2020-07	region_158	brand_1	50.00	0.00	100.00
2020-07	region_158	brand_2	50.00	0.00	100.00
2020-07	region_159	brand_1	50.00	0.00	100.00
2020-07	region_159	brand_2	50.00	0.00	100.00
2020-07	region_160	brand_1	50.00	0.00	100.00
2020-07	region_160	brand_2	50.00	0.00	100.00
2020-07	region_161	brand_1	50.00	0.00	100.00
2020-07	region_161	brand_2	50.00	0.00	100.00
2020-07	region_162	brand_1	50.00	0.00	100.00
2020-07	region_162	brand_2	50.00	0.00	100.00

4th edition
The financial challenge of the year
NOVARTISDATATHON
online

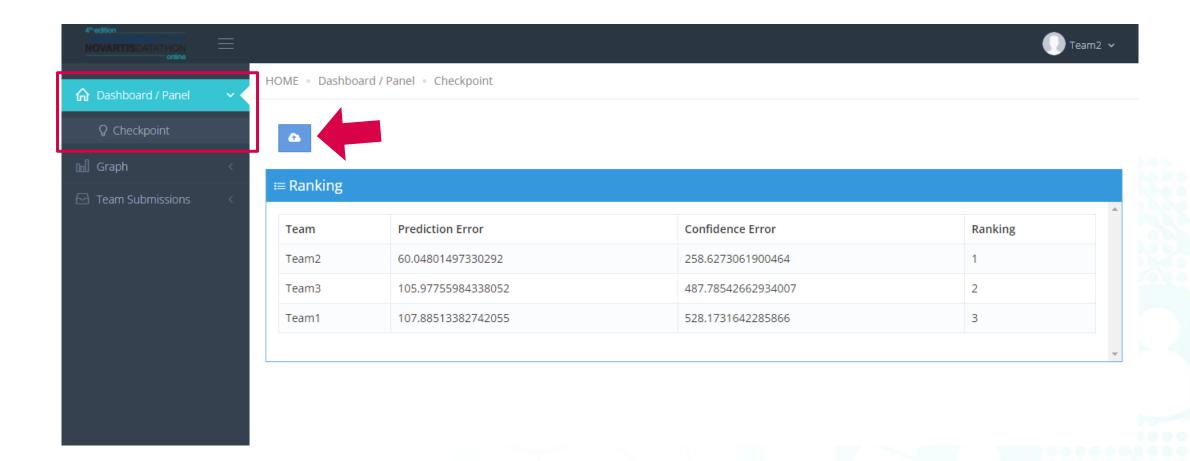
Change the password

Please **change** the password



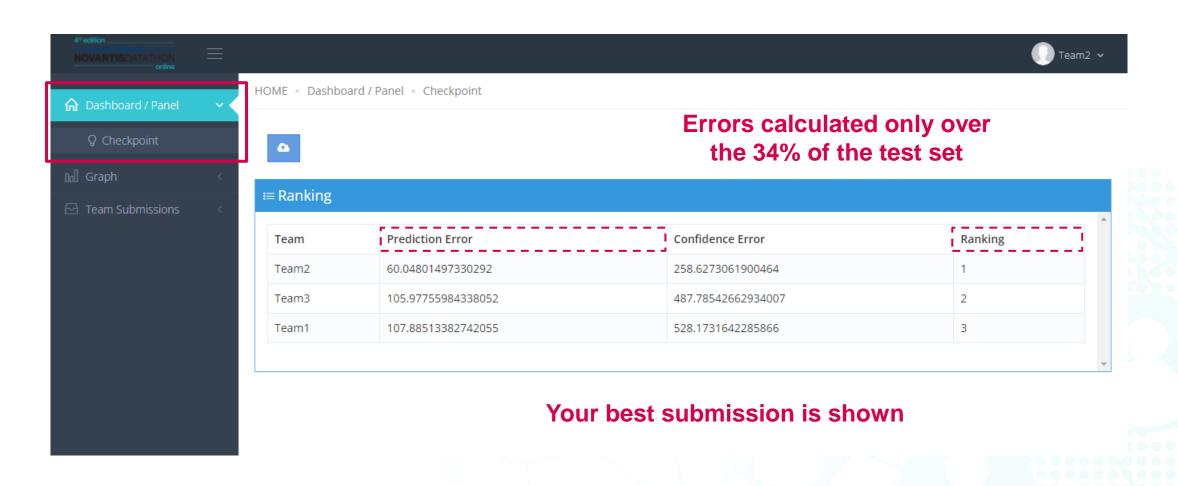
4th edition The financial challenge of the year NOVARTISDATATHON online

Submission



4th edition The financial challenge of the year NOVARTISDATATHON online

Ranking checkpoints



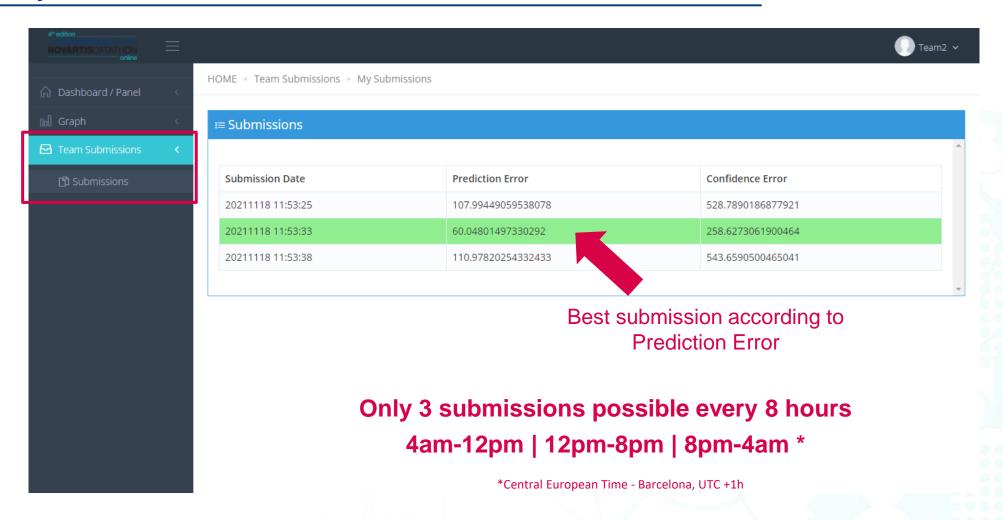
4th edition ______ The financial challenge of the year NOVARTISDATATHON online

Ranking checkpoints



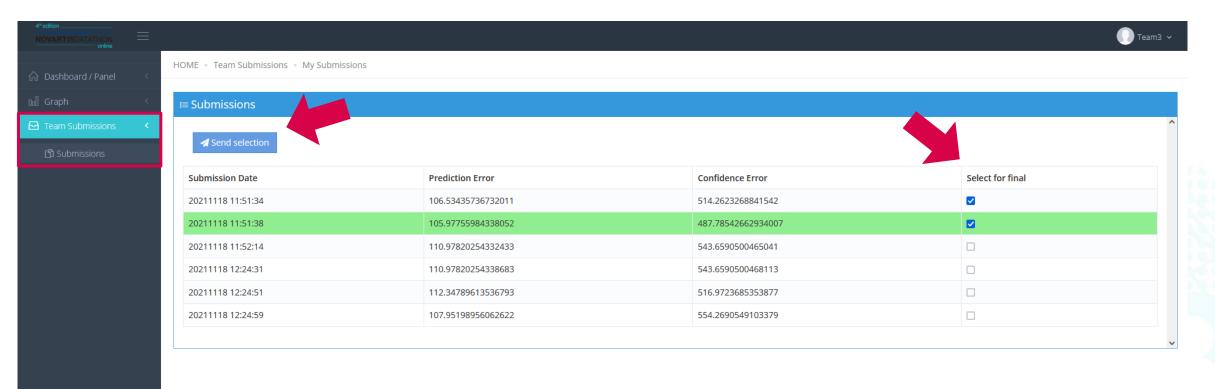
4th edition The financial challenge of the year NOVARTISDATATHON online

History of submissions



4th edition The financial challenge of the year **NOVARTIS**DATATHON

Final Submission (last hour)



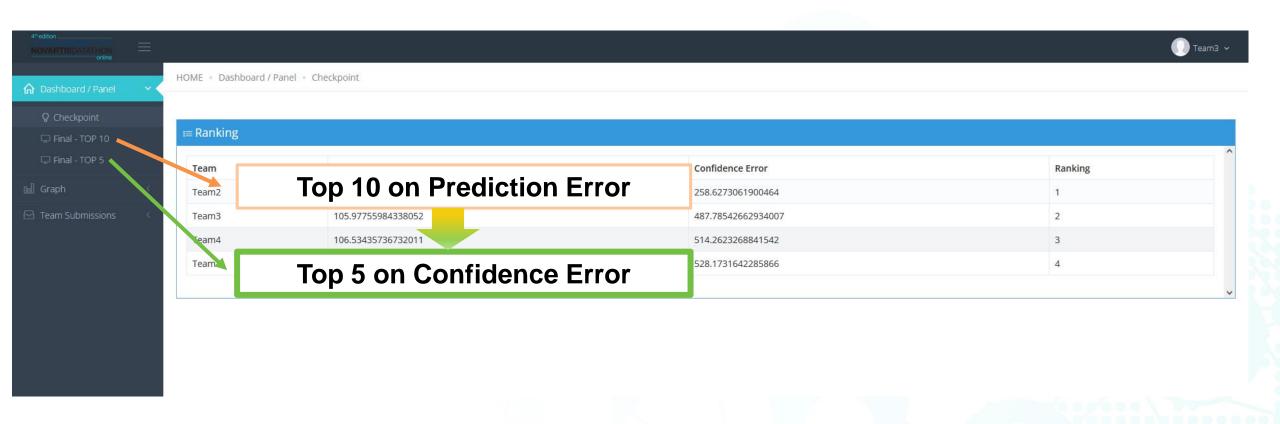
28th Nov between 10am and 11am *: select maximum two submissions

*Central European Time - Barcelona, UTC +1h

FINAL results calculated over the 100% of the test set once the datathon is over (28th Nov 11am)

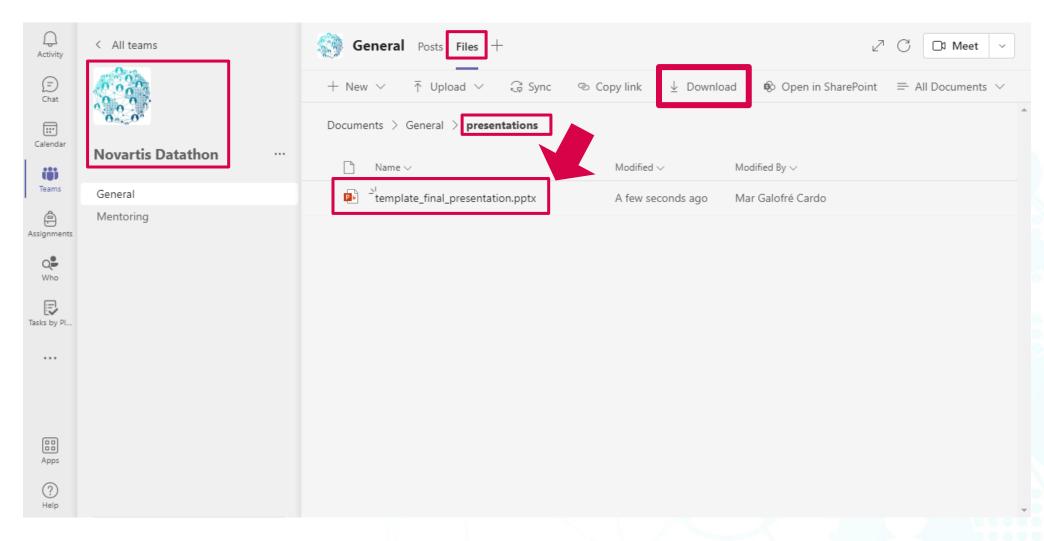


Final results: Deadline 11am on Sunday



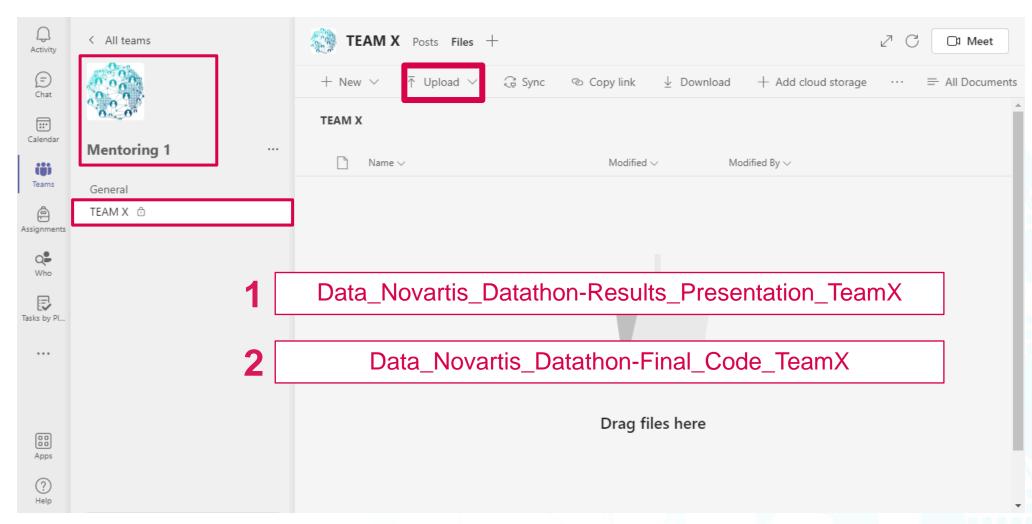
Submit presentation & code TOP5





Submit presentation & code TOP5





Agenda





THU 25 November

17:00h – 18:00h | Kick-off

18:00h – ... | Case work



FRI 26 November

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

09:00h - 12:00h | Mentoring

16:00h - 18:00h | Mentoring



SAT 27 November

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

09:00h – 12:00h | Mentoring

16:00h – 18:00h | Mentoring



SUN 28 November

09:00h | Welcome and Jury introduction

09:00h – 11:00h | Mentoring, if necessary

10:00h - 11:00h | Final submissions

11:00h | Deadline Submit final csv

11:30h | Results

12:00h | Deadline to upload TOP5 presentation

13:00h – 14:30h | Finalists' presentations

14:30h – 15:00h | Jury deliberates

15:00h | Announcement of the Winners

*Central European Time - Barcelona, UTC +1h