Welcome, data enthusiasts, innovators, and problem solvers from around the world. We are thrilled to host the eighth annual Datathon at the Barcelona Digital Finance Hub, where participants come together to tackle real-world financial challenges using data science.

My name is Pere, and I am a data scientist at the Digital Finance Hub. Let me start by introducing you to our team and our journey.

We are the Barcelona Digital Finance Hub, and over the past years we have been on an incredible path of growth and innovation. Our mission is crystal clear: to harness the power of data science and technology to revolutionize financial processes globally, pushing big data and AI to the forefront of financial innovation.

At the heart of our work stands Novartis' unwavering mission: to reimagine medicine to improve people's lives. We are committed to using innovative science and technology to address healthcare challenges, discover and develop breakthrough treatments, and find new ways to deliver them to as many people as possible.

At Novartis, we play a pivotal role in innovative medicine, global drug development, and technical operations. At the Hub, we are the link between finance and digital innovation, propelling the pharmaceutical industry into the future through data, automation, and AI.

Our evolution has been remarkable. What started with just 10 members in 2018 has grown into a powerhouse of 55 people in 2025, all passionate about data and finance. Today, our team consists of 34 data scientists, 8 finance professionals, and 15 experts in visualization, machine learning, software engineering, and DevOps.

Our strength lies in our diversity. Our backgrounds span mathematics, statistics, computer science, economics, physics, engineering, and more. This diverse expertise fuels our passion for data-driven innovation.

Barcelona itself plays a key role in our story. As a European hub for tech and AI talent, the city offers a thriving tech cluster, world-class research infrastructure such as the Barcelona Supercomputing Center, and strong university programs in data science and mathematics. It is also an attractive city for international talent, which makes it the perfect home for our growing digital finance community.

Now, it is time for the moment you have all been waiting for. Get ready to embrace the thrill, creativity, and innovation as we kick off the eighth annual Datathon at the Barcelona Digital Finance Hub. Good luck to all of you.

Thank you, Pere. Now, let me take a moment to introduce the journey of a drug from its launch to its eventual decline in sales, and to connect it with the focus of this year's Datathon.

When a drug is first introduced to the market, sales volume is usually low, as awareness and adoption are just beginning. Then, during the growth phase, volume increases rapidly as the product gains acceptance in the market, until the drug reaches its maturity phase, where sales tend to stabilize.

After that, a new competitor may enter the market, slowing down growth or even causing volumes to decline. Eventually, when the drug's patent expires – what we call the loss of exclusivity – generic products start entering the market. This is where we typically observe a sharp drop in sales volume, a phenomenon known as generic erosion.

This generic erosion is exactly the focus of this year's Datathon.

From a business perspective, it is crucial to anticipate how a drug will behave once generics enter the market, as this has a direct impact on revenue forecasts, product planning, and strategic decisions. Being able to predict this erosion allows countries and companies not only to plan ahead and minimize losses, but also to adapt their strategies and remain competitive in the post-patent period.

We define the mean generic erosion as the mean of the normalized volume after generic entry, calculated over the 24 months following that event. To do this, we normalize post-entry volumes using the average monthly sales during the 12 months before generic entry.

You may now ask: how do we classify this mean erosion?

We identify three types of erosion patterns:

- **Low erosion drugs**: these drugs experience minimal impact after generic entry. In this case, volume remains relatively stable, so the mean normalized erosion stays close to 1.
- **Medium erosion drugs**: these drugs show a moderate decline in sales after generic entry, with mean erosion levels between 0 and 1.
- **High erosion drugs**: these are drugs that experience a sharp drop in volume, resulting in a mean normalized erosion close to 0.

Using these patterns, we have classified brands into two erosion buckets:

- **Bucket 1**: high-erosion drugs, characterized by a mean erosion between 0 and 0.25.
- **Bucket 2**: drugs with a mean erosion higher than 0.25, including medium and low erosion cases.

The brands in Bucket 1 are the main focus of this year’s Datathon. Stay tuned – we will soon explain why they are so critical.

Now that we have seen the context and business needs, let us move on to the challenge proposed in this year’s Datathon.

From a technical point of view, you will be asked to forecast volume erosion following generic entry over a 24-month horizon from the generic entry date. This forecast must be produced at two different points in time:

1. **Scenario 1 – Right after generic entry**
    - You will have no actual data after the generic entry date.
    - Your task is to forecast monthly volumes from month 0 to month 23.
2. **Scenario 2 – Six months after generic entry**
    - You will already have the first 6 months of actual data after generic entry.
    - Your goal is to forecast monthly volumes from month 6 to month 23.

These two scenarios simulate real-world business situations, where the amount of post-entry information can vary and models must adapt accordingly.

Even though this Datathon has a strong technical component, there is also a crucial business dimension driving the question we are asking you to address. It is not only about how you solve the problem, but also why you choose a specific approach.

Therefore, all teams presenting in front of the jury will be asked to deliver a deep exploratory analysis of their data preprocessing, with a special focus on high-erosion cases – that is, drugs whose sales experience a sharp drop after generic entry. We strongly encourage you to use visualization tools to make your findings clear, interpretable, and business-oriented.

Thank you, Luisa, for setting the stage so clearly. Now, let us move on to how the Datathon winners will be selected.

The evaluation process will take place in two main phases.

Phase 1 – Model Evaluation

In the first phase, all participants will be required to submit their volume predictions for the entire test dataset. This test set includes both Scenario 1 and Scenario 2 cases.

Within this first phase, there are two steps:

- **Step 1 – Scenario 1 evaluation** All teams will be evaluated based on their prediction accuracy for Scenario 1. The teams with the lowest prediction errors will advance. Specifically, the top 10 teams will move on to the next step.
- **Step 2 – Scenario 2 evaluation** Only those 10 teams will then be evaluated again, this time based on their prediction accuracy for Scenario 2. Among these, the top 5 teams with the lowest prediction errors will advance to the final phase of the competition.

Phase 2 – Jury Evaluation

In the second and final phase, the 5 finalist 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.

Now that we have covered the evaluation process, let us take a closer look at the data you will be working with in this year's Datathon.

The **target variable** in this challenge is the **monthly volume** for 2,293 country–brand combinations that have experienced a generic entry.

- The **training set** includes 1,953 such observations. For each of these, you will have access to:
  - up to 24 months of volume data **before** generic entry, and
  - up to 24 months of volume data **after** generic entry.

This structure is designed to help you understand historical patterns and how volumes evolve once a generic competitor enters the market.

- The **test set** contains 340 observations and is used to evaluate your forecasts across the two scenarios:
  - In **Scenario 1** (about two-thirds of the test set, 228 observations), you will need to forecast from month 0 to month 23, predicting volumes right after the generic entry.

- In **Scenario 2** (the remaining one-third of the test set, 112 observations), you will forecast from month 6 to month 23. In this case, you will already have the first 6 months of post-entry actuals available.

Understanding the structure of these datasets is essential, as it will guide your model building, your feature engineering, and your strategy for handling the two different forecasting scenarios.

Continuing with the data challenge, let us briefly discuss how the test observations are distributed across the different erosion levels.

The test set (340 observations) is divided across:

- the two forecasting scenarios (Scenario 1 and Scenario 2), and
- the two erosion buckets (Bucket 1: high erosion, 0–0.25; Bucket 2: 0.25–1, mid and low erosion).

This same structure is maintained in both scenarios, ensuring consistent evaluation and that your models are tested under comparable conditions.

On the training side, the almost 2,000 country–brand combinations represent the full set of time series available for modeling. In the test set, both scenarios follow the same proportions across the two erosion buckets, although the total number of observations in Scenario 2 is smaller, reflecting the division between the forecasting tasks.

In the next section, we will walk you through the datasets that you will be working with. You will receive **three different data files**, each providing complementary information that will allow you to explore sales behavior, market dynamics, and product characteristics within the pharmaceutical domain.

1. **Sales Volume Dataset** This dataset contains information on sales volume before and after the entry of generic drugs into the market. Each row includes:

    - **country** – the market of reference,
    - **brand_name** – the product of interest,
    - **month** – the calendar month for each observation,
    - **months_post_gx** – the number of months relative to generic entry (0 is the month of entry; negative values are months before; positive values are months after),
    - **volume** – the number of units (drugs) sold.

    The variable **volume** is your main target throughout the analysis.

2. **DF_Generics Dataset** This dataset complements the first one by providing information about the presence and evolution of generics in each country and for each brand. It includes:

    - **country** and **brand_name** for identification,
    - **months_post_gx** (same definition as above),
    - **number_of_gx** – the number of generics available at that point in time.

    Note that the number of generics may vary over time as new products enter or leave the market.

3. **Drug Characteristics Dataset** The third dataset provides contextual and descriptive information about each drug. Each row includes:

  - **country** and **brand_name**,
  - **therapeutic_area** – the therapeutic area of the drug,
  - **hospital_rate** – the percentage of units delivered through hospitals,
  - **main_package** – the most common format (e.g., pills, vials),
  - **biological** – a Boolean variable indicating whether the drug is derived from living organisms (e.g., proteins, antibodies),
  - **small_molecule** – a Boolean variable indicating whether it is a chemically synthesized compound of low molecular weight.

This third dataset will help you characterize products and better interpret variations in sales behavior. You can assume that the features in this dataset do not change over time.

Now, let us define the **metrics** that we will use to evaluate each team. Both metrics for Phase 1A and Phase 1B (corresponding to Scenario 1 and Scenario 2) are based on prediction errors, but they are defined over different intervals.

Metric for Phase 1A – Scenario 1

In Phase 1A, participants must provide 24 months of predictions without knowing any actual data after the generic entry date.

To compute the prediction error for this phase, we will compare the predicted volumes with the actual volumes in four ways, giving more weight to the early months after generic entry:

- the monthly error across all 24 months,
- the accumulated error for months 0–4 (the first 5 months),
- the accumulated error for months 6–11,
- the accumulated error for months 12–23.

These components are combined using weights so that:

- the first 5 months have the highest impact (where most of the erosion occurs),
- followed by the middle horizon,
- and finally the full 24-month error.

Since each brand can have very different magnitudes, all errors are normalized by the **average monthly volume of the 12 months before generic entry**.

Once we compute the prediction error for each country–brand combination (denoted by *j*), the overall prediction error for a team is calculated as a **weighted sum of the average prediction errors across the two buckets**. Prediction errors in Bucket 1 (high erosion) are weighted **twice as much** as those in Bucket 2 in the final score. Each bucket is also divided by the number of country–brand pairs within that bucket, so that the resulting magnitude is independent of bucket size.

Metric for Phase 1B – Scenario 2

For Phase 1B, we use a similar approach, but now we evaluate predictions for Scenario 2, where actual data is available up to month 5 and predictions are required from month 6 to month 23.

To compute the prediction error for this phase, we compare predicted volumes with actual volumes in three ways:

- the monthly error across months 6–23,
- the accumulated error for months 6–11,
- the accumulated error for months 12–23.

Again, all errors are normalized by the average monthly volume of the 12 months before generic entry.

For each country–brand combination *j*, the prediction error is calculated as:

- 20% of the accumulated monthly error (months 6–23),
- 50% of the accumulated error for months 6–11,
- 30% of the accumulated error for months 12–23.

Once we compute these errors for each country–brand pair, the overall prediction error for a team is obtained as the weighted sum of the average errors across the two buckets. As before, prediction errors in Bucket 1 are weighted twice as much as those in Bucket 2 in the final score.

Now, let us move on to the **logistics and technical platform** that will support the Datathon.

Communication between teams and mentors will take place through **Microsoft Teams**. Each participant will be able to log in with their user account.

Once on the platform, you will find two main channels:

1. **Mentoring**

   - This is a private channel used for communication between each team and its mentors.
   - Only team members and their mentors have access to that chat.
   - Mentoring meetings will take place here: mentors will start the meetings, and you will join them by clicking the "Meet" button at the agreed time.

2. **Novartis Datathon**

   - This is the general channel for open communication among all participants.

   - Within it, there is a **General** sub-channel where only mentors can post. It will provide general information and announcements about the Datathon.

   - In the **Files** tab of this channel, you will find a folder called **Data for Participants**. Inside it, there is a **Submissions** folder containing:

     - the metrics for cross-validation,
     - instructions for submitting results,
     - examples of the required data formats.

At the end of the challenge, the five finalist teams will have to make a public presentation. In the Files tab, you will also find templates to prepare your slides.

Now let us talk about the **submission platform** that will be used to upload your results, calculate metrics, and generate the ranking.

To log into the platform, use your team's username and password. The access link is provided in the submission instructions document.

Once you enter the platform, the first thing you should do is **change your password**. To do this:

- go to the options on the right,
- click on **Profile**,
- then click **Change password**, and fill in the required fields.

After changing the password, you can start uploading your submissions:

- Go to the left-hand menu,
- click on **Dashboard/Panel**,
- then click on **Checkpoint**,
- and use the upload button to submit your file.

If an error message appears, it means that the file structure is not correct.

As soon as the first valid file is submitted, your team will appear in the ranking. Each team will appear only once, showing its **best solution**. The ranking is updated every time a team uploads a new submission.

Both metrics are calculated using a test set that is divided into:

- a **public test set** with 30% of the data, and
- a **private test set** with the remaining 70%.

During the competition, only the public test set will be used to update the leaderboard.

We strongly recommend making a **test submission** within the first hours of the challenge to ensure that you fully understand how the platform works.

You can review your past submissions using the **Team Submissions** section. Keep in mind that the **number of submissions is limited**: you may submit up to **three times every eight hours**, so make the most of each attempt.

On Sunday at 9:30 AM, the **"Select your final option"** feature will be activated. You will have until 10:30 AM to select which submission you want to use for the final evaluation. At 10:30 AM, the Datathon will end, and no further changes will be allowed. From that point, the final results will be calculated using the private test set.

Once the final submissions are checked and evaluated on the complete test set, the **top 10 results** will be published, ordered by their Scenario 1 score. Then, the **top 5 results** will be published, ordered by their Scenario 2 score. These 5 teams will be the **finalists**.

Only these finalist teams will have to prepare a presentation summarizing the methodology and results of their algorithms. You will have until 12:00 to upload your presentation to your private channel, following the specified naming convention. In addition, finalist teams are required to upload the **code** used to obtain the results of their final submissions.

At 1:00 PM, the finalist presentations will take place. At 2:30 PM, the jury will deliberate on the presentations and announce the winners.

We wish you all a productive, insightful, and enjoyable Datathon. Good luck, and we look forward to seeing your ideas and solutions.