Mercury Challenge FAQ

What is the goal of the Mercury Challenge?

The Mercury Challenge seeks innovative solutions and methods for the automated production of forecasts for a series of event classes across Egypt, Saudi Arabia, Iraq, Syria, Qatar, Lebanon, Jordan, and Bahrain. Participants of the challenge are provided a dataset of Gold Standard Reports (GSR) that are comprised of IARPA-vetted and researched news sources for their solutions.

How do you know that the Gold Standard Sources (GSS) cited in the Gold Standard Reports (GSR) provide reliable training data?

The GSS production process has been underway for nearly three years. Every month the IARPA team reviews these GSR and compares them to many other sources of information. The provided training data, while not perfect, is accurate enough to reflect the true state of events in the subject countries within the range of the challenge scoring criteria.

Are you looking for people who are good forecasters for the challenge, or are you looking for data scientists who can create solutions?

The Mercury Challenge is looking for data science / machine learning solutions capable of interacting with an API and processing data sources. The output of this challenge should be a reproducible code-base that can be run by machines and/or be fully automated.

Why does it take so long for the GSR to be resolved / leaderboard to be scored?

To ensure product quality, the GSR creation process involves multiple quality control phases requiring human intervention. The GSR are published on a monthly basis, and it takes about two weeks to finalize the GSR from the preceding month. This process has worked very well over the past few years for other programs and the challenge will benefit from maintaining the GSR production process as is.

Am I allowed to use other data sources in the challenge beyond the provided GSR file?

Yes, you can use any data source in the challenge as long as you have the rights to use the data. You are responsible for ensuring that you have obtained the appropriate permissions, data rights for the challenge. Please note that all data sources need to be documented as part of the solution

review process in order to be scored.

Am I allowed to use subscription-based or fee-based data sources in the challenge?

Yes, as long as you have paid for, or have legal access to the data sources, you may use them in the challenge. Please note that all data sources need to be documented as part of the solution review process to be scored.

There are a very low number of MA events for Bahrain, Qatar, Jordan and even Saudi Arabia data in the training data. How do you expect a participant to effectively train an algorithm with such a small data set?

Correct.   That is why the awards for successful forecasts for these low-activity countries are higher than those for countries with higher levels of activity. We encourage all participants to attempt a solution despite this and are looking forward to seeing innovative solutions to overcome those limitations.

There seems to be a large number of unspecified actors and targets at times for some of the GSR countries, why is this the case?

The GSR creation process has very strict rules to ensure actor and targets that are listed can be explicitly verified in the GSS. When local issues restrict the ability of the GSS to precisely describe the actor or target, the GSR will default to "unspecified." But this is good news for the challenge participants in that "unspecified" actors or targets are "wildcards" and will match to any warning actor or target.

Some GSR contain a very long string of actors or targets, why is this the case?

The GSR creation process has very strict rules to ensure actor and targets that are listed can be explicitly verified in the GSS. If the GSS is not explicit about a single actor or target, or contains multiple actors, then the GSR will include all the possible actors or targets. One of the more extreme examples is this actor set within the Syrian Military Activity (MA) training set:

National Defence Forces Syria / NDF / Quwat Ad-Difa Al-Watani / Popular Committees / Lijan Militias / Al-Lijan Al-Sha'Biyah / Peoples Committees; Syrian Arab Military

Please note that this long string only occurs in 0.44% of the Syria MA GSR. The good news is, if a participant's warning has any one of those actors in a warning, that warning will match to the GSR.

The Anglicized State or other city name in the GSR is not one I can find in any gazetteer and city names where I am certain there is Military Activity or Civil Unrest are not listed anywhere in the GSR training data. Why is this the case?

The IARPA team used a specific gazetteer for GSR production that can cause some confusion on place names. However, the challenge scoring will be based on the latitude and longitude of participant warnings with respect to the latitude and longitude in the GSR, so these discrepancies should not affect participant scores. You may reference the GeoNames Search database for assistance with geographic names.

The challenge Press Release says the challenge is focused on "non-violent" civil unrest events. How does a participant determine if an event is non-violent or not?

> Please refer to the Mercury Challenge Handbook that details how each event is defined. For "non-violent Civil Unrest" the definition includes the following guidelines:
>
> The following examples are considered <u>non-violent</u> Civil Unrest:
>
> - Threats of violence, e.g., yelling, cursing
> - Police arresting protestors
> - Hunger strikes
>
> The following are examples of events of a violent nature that would also be considered <u>violent</u> Civil Unrest and should not be forecasted:
>
> - Clashes with police, e.g., police using tear gas or high pressure water hoses
> - Clashes between opposing groups resulting in injuries
> - Pushing or being forcibly removed from an area

Q&A - Session 1 Responses

Will I retain intellectual property rights to my solution?

Solvers will retain intellectual property rights at the end of the Mercury Challenge, regardless if they are a winner. However, a solution diagram will be required for all prize winners so it is recommended that Solvers consult the Prize Eligibility information to better understand if they will be able to produce the materials required for eligibility and for the solution diagram.

Will there be a RESTful API? And is one forecast a day used for scoring or do we have to do multiple forecasts a day to receive timeliness points on a breaking situation?

All solvers will be given a URL to which JSON formatted warnings are submitted. The Challenge will test validity of submissions using Docker containers. Solvers will be given a full version of the scoring system. Express Score is available to test the scoring system for warnings for those who do not wish to install Docker.

Once the API is ready, documentation on how to connect to the API will be made available to Solvers.

What languages and software can challenge participants use?

Any programming language can be used by challenge participants. However, all warnings will be submitted in JSON format, as specified in the Mercury Challenge Handbook.

Can you elaborate more on Lead Time? Is this saying that there's a benefit to submitting early forecasts?

Mercury scoring considers two time-based metrics. Lead Time measures the number of days between warning submission and first report of the event. Date Score measures the number of days difference between the predicted Event_Date in the warning and the actual Event_Date in the matched warning. For example, suppose a warning is submitted on (2018-07-16) for MA that is predicted will occur on Friday (2018-07-20). Futhermore, suppose that this warning is matched to an event that actually occurs on (2018-07-21) and is first reported on (2018-07-22). The Lead Time for this warning is 6 days (22 - 16) and the Date Score is based on a difference of 1 day (21 - 20) which would yield DS = 0.75. The requirement for positive Lead Time is to ensure that your warning is about the future. The requirement for positive Date Score is to ensure that warnings and events too widely separated in date are not matched.

What will the F Score be used for in the course of scoring?

The F Score is used as a means of balancing scores. F Score is designed to prevent flooding methods (i.e. submitting many warnings without regard for accuracy in order to cover all possible outcomes). For more information on how F Score is calculated, please consult the Mercury Challenge Handbook, under the section "Military Activity (MA) Judging Metrics."

Do we need to submit multiple forecasts per day? Is intra-day time measured or is there one time per day where we need to get a forecast in?

For Military Activity events, each warning involves a specific, discrete event with an actor or location associated with. Multiple warnings are not required, but it is possible to batch submit warnings. However, all warnings are scored on an individual basis.

Submissions are only analyzed at the date level, not below that. Solvers can submit at any time during the day, which is defined as ending at 21:00 UTC.

When is the ground truth generated? Is the ground truth also generated continuously?

Ground truth will be updated on a bi-weekly schedule and will be placed in the GitHub Repo when it is ready. For more details, please consult the Mercury Challenge Handbook, under the section "Non-Violent Civil Unrest Definition."

Event counts for disease and ACLED events get revised over time. Typically, there is an undercount that gets retroactively updated months later. How will the Mercury Challenge handle this?

Mercury Challenge does not have a strict ground truth threshold, so it is possible there will be retroactive revisions to the sources. However, in our analysis these revisions tend to stabilize quickly and we do not anticipate it being a problem for the Mercury Challenge. Due to the nature of the challenge, when our final scoring for a Period is complete, the counts that exist on the site at that time will be used for the challenge. Counts updated in the future will not be factored into the challenge.

If you register are you required to submit to each challenge?

No, registration does not commit you to participating and you are not required to participate in all three periods. However, the longer you participate in the challenge, the more likely you are to qualify for prizes.

Can you discuss the solution diagram required at the end of the challenge? Is the same level of scientific rigor that you expect from a BAA Solver expected from the solution diagram?

The solution diagram is a document that outlines the technical underpinnings of your solutions, to include what scientific techniques were applied, what data sources were involved, and what code was developed, among other things. The intent to inform readers of how the solution operates. The diagram should be high-level and understandable but should still contain enough technical information to inform the reader of the solution works. The goal is for a technical expert to be able to understand what a Solver did without having to include the technical details.

What kind of follow-on can we expect for if we have a great solution (i.e. Other Transaction Authority)?

Solvers should not expect any kind of formal follow-on (i.e. Other Transaction Authority). However, winners will be recognized by IARPA post-challenge, in the form of press releases and newsletters. Additionally, challenge winners have gone on to use their solutions in past challenges and as a basis for white papers and other academic involvement. Participants will retain ownership of intellectual property for their solutions.