ITU AI/ML in 5G Challenge Management Guidelines

Version 01, 03 June 2021

Contents

PART I – DATA SHARING GUIDELINES

The success of the ITU AI/ML in 5G Challenge depends on the availability of data and whether data owners are able, and willing, to share data with others. Rapid and unrestricted sharing of data and resources is essential for advancing the Challenge. However, there are cases where unrestricted data sharing is not possible. This document therefore addresses measures that can be taken to ensure that data providers are able to share relevant data with problem solvers or researchers under specific agreements to ensure data privacy and integrity. Consequently, having an institutional data sharing guideline is the first step towards encouraging companies, data providers, collaborators, researchers and professionals to share relevant data for the challenge.

NOTE - Data providers/owners: defined as entities who have data to share for specific problem statements of the ITU AI/ML in 5G Challenge. This data may be useful for training and testing of AI/ML models.

This document outlines a data management and sharing guideline. This guideline would help data owners to derive maximum value from their data while protecting the interests of their institution and its members.

Data classification categories

For the purposes of the ITU AI/ML challenge, we consider the data classification categories[1] below:

Data Category	Description
Public/Open Data	Data that can be made publicly available because disclosure is associated with little or minimal privacy impact to individuals and/or organizations. This includes data that is anonymous, aggregated or non-sensitive. NOTE - This kind of data can be shared without any restrictions.
Restricted data	Some data are moderately sensitive and cannot be shared publicly because disclosure can cause minor privacy impact for an individual, put an individual or community at risk of a privacy incident, or negatively impact upon an organization's capacity to compete in the market or carry out its activities. Example: measurement data obtained per access network or access network site. NOTE - This kind of data needs to be pre-processed with an intention of removing the privacy impact before being shared. Restricted data may be available only under certain conditions set forth by the data provider. *Example 1*: Restricted data may be made available after signing a non-disclosure agreement (NDA). *Example 2*: Restricted data may be available only for use within the hosted platform and not for moving out of the hosted platform (i.e. no downloading of data may be allowed). *Example 3*: Restricted data may be available to citizens of a particular country or region, e.g. under data privacy regulations of EU or China.
Secret	Also known as "personal, or confidential", this is composed of highly sensitive information that may cause serious distress or increase risk to an individual's safety, or violate an individual's privacy, or impact the compliance to privacy regulations by organizations. This includes personal data that could identify an individual (either on their own or if combined with other data sets), and protection incident management information. NOTE - This kind of data must not be shared.

[1] IOM, "Guideline for DTM Coordinators: Identifying Sensitive Data and Inter-Organizational Data Sharing Pathways"

In order to determine the sensitivity level of a dataset/information type, it is recommended that the data owner perform a classification of the data and a risk-assessment on the potential impact that disclosure of each dataset/information type might have.

For the ITU AI/ML Challenge, we are interested in data that is classified as open or restricted.

Options for hosting “restricted data” for AI/ML in 5G Challenge

Data providers who would like to share data under the “restricted data” category have the following options to choose from:

Option 1: Self hosted

- The data provider hosts an ML sandbox, including toolsets, e.g. for training and data handling. The ML sandbox will be in-premise of the data provider.

  NOTE - The ML Sandbox is defined in ITU-T Y.3172 “Architectural framework for machine learning in future networks including IMT-2020”.

- According to step 7 of the “Data Sharing Guidelines” (see below), user agreements are drafted for access to this ML sandbox. For example, no download of data may be allowed.
- The ITU Secretariat will provide data to interested participants for the data provider’s problem statement, based on preferences by the participants during registration and subsequent discussions.
- The data provider shortlists the candidates who can access the restricted data.
- A user agreement is signed which makes the participants eligible to compete in the ITU AI/ML in 5G Challenge using the restricted data.

Option 2: ITU hosted

- The data provider instantiates an ML sandbox, including toolsets (e.g. for training and data handling). This ML Sandbox will be in-premise of ITU (Geneva).
- All other steps remain the same as in Option 1.

  NOTE- In this option, ML sandbox maintenance is taken care of by ITU.

  NOTE - ITU-hosted ML sandbox may be reused in future editions of such Challenges.

NOTE - With ITU facilitating the sharing of data between data providers and eligible participants, this may eliminate the need for each participant or team to negotiate with the data provider individually.

Risk assessment

Risk assessment with respect to the data classification (see above) must be carried out by the data provider, considering that data sensitivity is:

- **Contextual**: What may not constitute sensitive data and information in one context, may be sensitive in another.
- **Temporal**: Data may not be sensitive now, but may become sensitive in the future due to changes in context, such as shifts in policies and/or safety of specific populations.
- **Relational**: One dataset on its own may not be sensitive, however it could become sensitive if analyzed in combination with other dataset(s).

Standards, metadata and documentation

For data sharing to be a success it is important that data are prepared in such a way that those using the dataset have a clear understanding of what the data mean so that they can be used appropriately. To enable this, data providers are encouraged to include with the dataset all the necessary information (metadata) describing the data and their format. This information should include such information as:

- the methodology used to collect data
- definitions of variables
- units of measurement
- data format
- file type of the data
- any assumptions made.

Data sharing guidelines

The figure below shows the steps to be considered when an entity (data owner) is planning to share data for the ITU AI/ML in 5G Challenge.

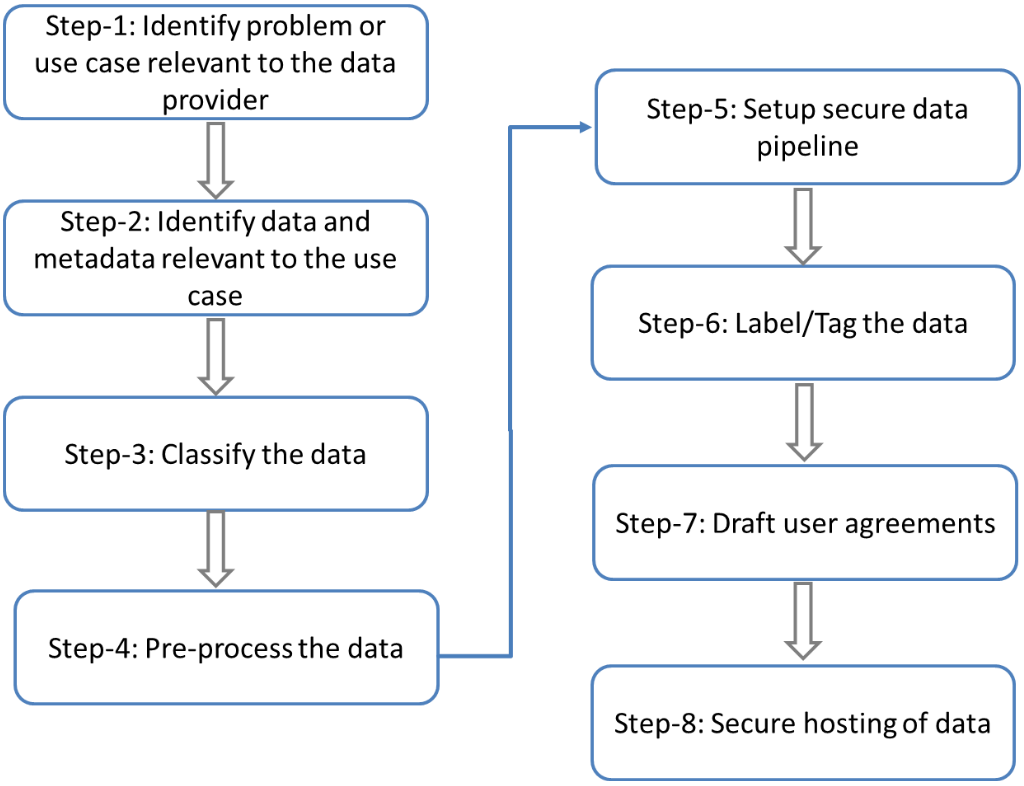
Step-1: Identify problem or use case relevant to the data provider
Step-2: Identify data and metadata relevant to the use case
Step-3: Classify the data
Step-4: Pre-process the data
Step-5: Setup secure data pipeline
Step-6: Label/Tag the data
Step-7: Draft user agreements
Step-8: Secure hosting of data

Figure 1: Data Sharing Guidelines for the ITU AI/ML in 5G Challenge

Step 1: *Identify problem or use case relevant to the data provider*. In this context, the data owner should choose what type of problem they would like to purse or consider during the Challenge. This will help determine the data relevant for the problem.

Step 2: ***Identify data and metadata relevant to the use case.*** The problem and/or data owner determines what type of data they would provide to solve the problem identified in Step 1. In this step, the dataset identified should also contain all the necessary information (metadata) describing the data and their format.

NOTE - ITU can offer expertise to identify data to be collected based on metadata relevant to the use case.

Step 3: ***Classify the data.*** In this step, the data is classified whether it is open (publicly available) or restricted (i.e. provided to Challenge participants after certain transformations, under certain rules or user agreements) or secret (not shared at all). This may depend on the internal risk assessment of the data provider.

Step 4: ***Preprocess the data (optional step).*** This is an optional step based on the output of step 3 above. Data anonymization is a type of preprocessing whose intent is privacy protection. It is the process of either encrypting or removing personally identifiable information from data sets. The data provider should decide which information to keep for data to be useful and which to anonymize or to transform or to omit.

Step 5: ***Setup secure data pipeline.*** A data pipeline is a series of data processing steps. It enables a smooth, automated flow of data from one station to the next. It starts by defining what, where, and how data is collected. It automates the processes involved in extracting, transforming, combining, validating, and loading data for further analysis and visualization. Data pipelines consist of three key elements: a source, a processing step or steps, and a destination. Data pipelines enable the flow of data from an application to a data warehouse, from a data lake to an analytics database, or into a ML pipeline system, for example.

Step 6: ***Label/Tag the data (optional step).*** Data labelling is the process of detecting and tagging data samples. The process can be manual but is usually performed or assisted by software. Labelled data is a group of samples that have been tagged with one or more labels. In machine learning, if you have labelled data, that means your data is marked up, or annotated, to show the target, which is the answer you want your machine learning model to predict. In general, data labelling can refer to tasks that include data tagging, annotation, classification, moderation, transcription, or processing. Labelled data highlights data features - or properties, characteristics, or classifications - that can be analyzed for patterns that help predict the target.

Step 7: ***Draft user agreements.*** A user agreement is an agreement made between the owner, administrator or provider of a service (data owner) and the user of such a service (Challenge participants), that defines the rights and responsibilities of both the parties. Privacy policies, or terms and conditions are examples of a user agreement.

Step 8: ***Secure hosting of data.*** In this step, the data owner or ITU provides a platform to store restricted data in a manner compliant with the data provider’s data sharing policy. The Challenge participants can access the restricted secure data hosted on the platform by signing non-disclosure agreement or user agreements. This data can be accessed by using password or tokens.

PART II – WORKING MODEL FOR JUDGES PANEL

Judges Panel

The Judges Panel is a collection of individuals from across the world who will evaluate, on an ongoing basis, the progress and merit of the solutions proposed by the participants. The Judges Panel will monitor entries during the competition phase and evaluate best solutions shortlisted for the Grand Challenge Finale. The hosts of problem statements will provide a score for each participant or team at the end of the competition phase. Individuals in the Judges Panel will be selected by the Challenge Management Board (CMB).

Working model for the Judges panel:

1) CMB will call for and work on the draft of judgment criteria and review it along with the Judges Panel. The judgement criteria will contain a reference to (a) individual evaluation criteria mentioned in the problem statements and (b) a uniform set of criteria, which when applied, helps to select the top few entries from each host of a problem statement. The final reviewed judgment criteria will be baselined in the Challenge website.

   NOTE - The final judgment criteria will make sure that the selection of top few entries from each host is using a uniform approach world-wide.

2) The Judges panel (or its subset) will periodically meet with hosts to review the progress of the leaderboard maintained by the regional host. Any best practices or problems which arise out of such meetings will be discussed in the CMB.

   NOTE - This will make sure that the selection of competition phase winners by hosts are applying the uniform criteria laid forth by the CMB and the Judges Panel.

3) The Judges Panel will recommend the invitees from competition phase for each of the problem statements supplied by the hosts to the final event.
4) The Judges Panel will review, along with the CMB, the scoring criteria for the entries to the Grand Challenge Finale. This scoring criteria will be baselined on the Challenge website.
5) The Judges Panel will recommend the final winners from the Grand Challenge Finale (according to the baselined scoring criteria).

Confidentiality terms for Judges

Judges have to accept the terms and conditions laid down by the CMB when joining the Judges Panel. This includes the following parts:

- By joining the Judges Panel, the judges accept to abide to confidentiality terms. This means they cannot disclose any information about a submission to any third parties.
- Conflict of interest:
  In case a judge finds conflict of interest in judging some of the submissions, she/he should declare that well in advance and abstain from judging those submissions. The following scenarios provide guidelines for finding conflict of interest:

  Scenario 1: If a judge is a member of the same entity (e.g. a company, university) as an employee or a team who submits a solution to one of the problem statements, the judge must declare a conflict of interest when the submission has been made (during evaluation/scoring of the solution) and abstain from taking part in the judgment of this submission.

Scenario 2: A person is listed as the contact for a problem statement, but she also wants to submit a solution to that problem statement. In this scenario, the host and contact person will declare a conflict of interest during the registration or submission. The team is allowed to make a reference solution which will not be considered for any prize for this particular submission, but will be given due consideration of a solution, e.g. an invited presentation can be made by the team based on the submission.

Draft scoring criteria

The scoring criteria for each of the problem statement is provided in the Problem statements and data resources document. For the Grand Challenge Finale, the criteria are laid out on the challenge website in the Participation Guidelines document.

PART III – HOST ON-BOARDING GUIDELINES

This section is intended for new hosts of problem statements of the competition phase for the ITU AI/ML in 5G Challenge. To have your problem statement ready and accepted for the Challenge, please make sure that the following checklist is satisfied:

Check Point 1 - Coordinator: Make sure that you have a person who can run the Challenge problem statement and represent your entity, in coordination with ITU, for the duration of the Challenge (June–December 2021).

Check Point 2 - **Problem statement(s) description:** Make sure that you submit your problem statement(s) to ITU using the template provided on the Challenge website.

Check Point 3 - **Nominee for the Judges Panel**: Participate and evaluate submissions (as part of the Judges Panel).

Check Point 4 - **Web-admin**: Setup a local website or coordinate with ITU to publish the problem statement on the ITU Challenge Management Platform with problem description and a link to the dataset. A website and logo should be provided so that participants can access the site where the description of the dataset and other resources are provided.

- If your problem statement requires data, you are required to provide the dataset that participants are going to use in the Challenge. We encourage that you provide information or toy examples for your problem statement.
- Some problems might require a “Sandbox” to test the submissions [August-October 2021] – at this point, we may need a setup with simulators.
- Maintain the leaderboard for your problem statement [July-December 2021].

Check Point 5 - **Evaluation**: Make sure that you have a committee (or a person) who can review the code submissions and evaluate the results. Description of what the participants are required to submit, the evaluation criteria and possible deadlines which follows the ITU Challenge timelines.

Check Point 6 - **Funding for Prizes by host:** If you offer prizes for your problem statement (money, winner-certificates), please inform ITU. Please describe the prizes on your hosted website, apart from the prizes listed on the ITU Challenge website.

Check Point 7 - **Presentations/webinar** to participants: The host of a problem statement will have to make a presentation (webinar) to describe the problem and evaluation criteria, resources available and incentives, act as a mentor to their problem statement to guide the participants, and participate in the Slack channel discussions.

Check Point 8 - **CMB**: Challenge Management Board (CMB) nominee: participate once a month in CMB discussion to understand the decisions in the CMB.

Check Point 9 - **Marketing**: The host publishes the website of its problem statement (in English and local languages) and invite colleagues to participate in the Challenge. The hosts markets (alongside ITU) the Challenge to attract participants to the problem statement.

Solutions from the Challenge are recommended to be shared with the community as open-source. An example of the 2020 repository can be accessed here: https://github.com/ITU-AI-ML-in-5G-Challenge

Please see examples below for 2021 Challenge website:

https://bnn.upc.edu/challenge/gnnet2021/

NOTE **-** Furthermore, apart from these requirements, we encourage open-source submissions and data sets to be open to all participants (if you cannot open to everyone, please do let us know).

____