How can data and AI be a lifeline for a vulnerable coastline?

The better the question. The better the answer.
The better the world works.

Microsoft

Guidance and Suggestions for Participants of the 2024 EY Open Science Data Challenge

Welcome to the 2024 EY Open Science Data Challenge! Held annually, the EY Open Science Data Challenge gives thousands of early-career professionals and university students the opportunity to use data, artificial intelligence (AI) and technology to help build a sustainable future for society and the planet.

The 2024 challenge focuses on coastal resilience and climate change. Participants will use high-resolution satellite datasets to build predictive models to help vulnerable coastal communities adapt to evolving conditions and recover from extreme climate events. These solutions will result in new and innovative ideas designed to increase the impact of data for societal benefit.

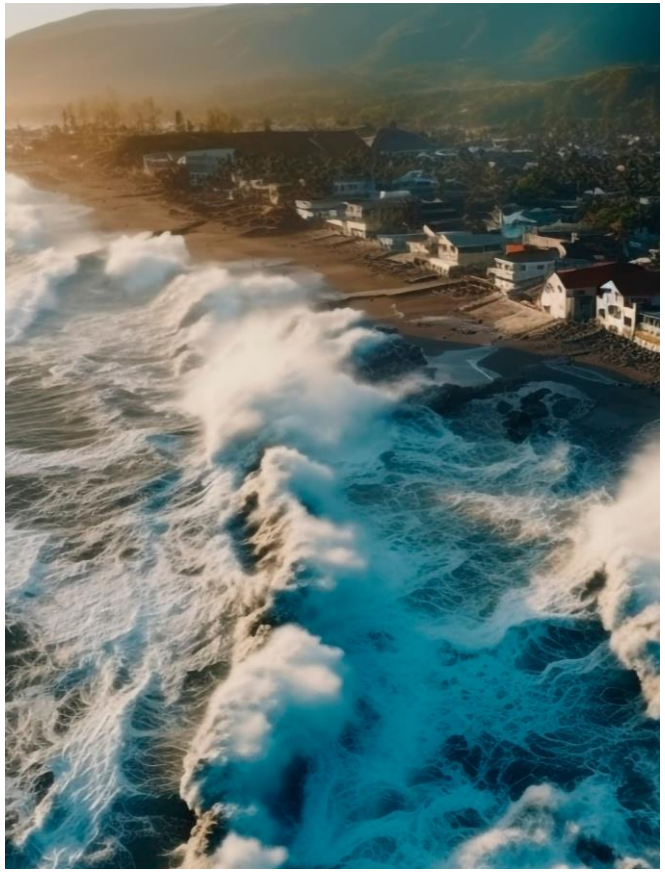
Why coastal resilience?

For centuries, people have settled in large numbers where the land meets the sea, drawn by an abundance of resources and opportunities. Today, nearly 75% of the world's population lives within 50 kilometers of the ocean.[1]

The exposure is especially acute for developing countries and small island states, where climate risk is layered over existing vulnerabilities.[2] The International Monetary Fund (IMF) estimates that adaptation costs for small island nations exposed to tropical cyclones and rising seas could reach up to 20% of gross domestic product (GDP).[3]

Tropical cyclones have enormous destructive potential. They account for the highest losses of all natural hazards:

- Caused the most damage of all-weather disasters between 1980 and 2021, with over $1.1 trillion total, and an average cost of $20.5 billion per event.
- In the past decade alone, economic losses from tropical cyclones exceeded 573.2 billion U.S. dollars globally.
- They are also responsible for the highest number of deaths (6,697 deaths between 1980 and 2021).



[1] https://www.weforum.org/agenda/2022/02/ecologically-intact-coastlines-rare-study
[2] https://www.oecd.org/env/cc/2502872.pdf, Poverty and Climate Change: Reducing the Vulnerability of the Poor through Adaptation
[3] https://www.imf.org/en/Blogs/Articles/2022/03/23/blog032322-poor-and-vulnerable-countris-need-support-to-adapt-to-climate-change

WIIFM&O (what's in it for me and others?)

Consistent with previous challenges, the 2024 competition is rooted in action and informed by altruism.

It isn't a competition for competition's sake; but rather a purpose-driven mission to address the acute problem of coastal vulnerability due to climate change. Accordingly, the benefits of participating in the challenge are great.

For you

It's a chance to connect your **personal** drive to help others and build a better world with your talent and skills in data science and AI. You'll also improve your competency in science, machine learning and managing large volumes of data, boosting your **professional** profile and recruitment potential.

For the best and brightest ideas, rewards await – in recognition, cash prizes and attendance at the IGARRS conference in July 2024 in **Athens, Greece**. The conference is the leading meeting of more than 2.500 esteemed scientists and professionals in the Remote Sensing field worldwide. The 2024 conference theme, "Acting for Sustainability and Resilience," is well-aligned with the data challenge topic and promises to deliver exciting opportunities for learning, professional growth and networking. As part of the conference experience, challenge finalists will orally present papers on their solutions in a dedicated session, allowing them to showcase their capabilities and strengths.

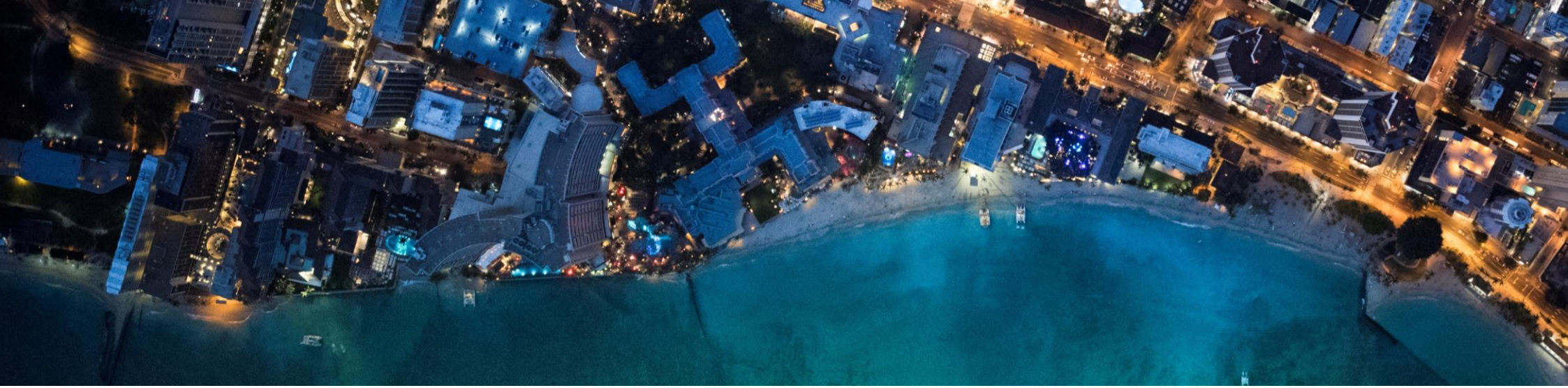
Cash prizes are as follows: The challenge winner receives **$10,000 USD**, the first runner up receives **$5,000 USD** and the second runner up is awarded **$2,500 USD**. If a team wins the challenge, the prize will be shared among all team members.

For others

The solutions you'll formulate through challenge involvement will provide tangible benefits to those suffering in vulnerable, data-poor areas. Due to insufficient mapping of infrastructure and ecosystems, governments and institutions lack a clear and complete picture of the natural and material world around them. The good news? A wealth of data is out there. Through the challenge, you can apply leading technology innovations to accelerate coastal resilience and enable solutions that scale globally.

For our world

Because we connect each year's challenge to the United Nations Sustainable Development Goals, you're not working in a vacuum. Rather, there is alignment with ongoing efforts centered on the SDG (Sustainable Development Goals).



What you need to know before you get started

This challenge consists of two phases.

The Phase 1 goal will be to develop machine learning and AI models using high-resolution satellite data to assess storm damage and support disaster response and recovery efforts in data-poor coastal communities. Participants will be provided with high-resolution 30-cm analysis-ready satellite data from Maxar's GEO-1 mission. These unique datasets from before and after a tropical storm provide a unique view of the storm damage and allow the use of object detection techniques to identify building damage. In addition, participants will receive sample notebooks and algorithms that demonstrate machine learning approaches.

The Phase 2 goal is to develop a practical "business plan" that describes how your Phase 1 model could be applied by local beneficiaries to assess coastal infrastructure damage, vulnerability, socioeconomic impact, and climate change risk for future storms.

This document is meant to provide background and dataset information, and references for participants. The guidance and suggestions presented here should help participants understand how the satellite data can be used to identify different building types and storm damage and build corresponding machine learning models.

The use of satellite data to support disaster response and recovery is not new. Researchers and government organizations routinely use satellite data to identify the location of storm damage for the purpose of recovery efforts, estimations of socioeconomic impact, and insurance claims.

Unfortunately, this data is often used to manually search for storm damage rather than using automated tools. Open-source solutions that automate the search and identification of storm damage using high-resolution satellite data do not exist. But this data challenge gives you the chance to use cloud computing and machine learning to make contributions toward coastal resilience.

Analysis Region

For this challenge, we have selected a region over Puerto Rico which was devastated by hurricane Maria on September 20, 2017. The storm center crossed the southeast coast of Puerto Rico near Yabucoa with maximum winds at 135 knots. The hurricane's center crossed the island, roughly diagonally from southeast to northwest, for about 8 hours before moving into the Atlantic Ocean.

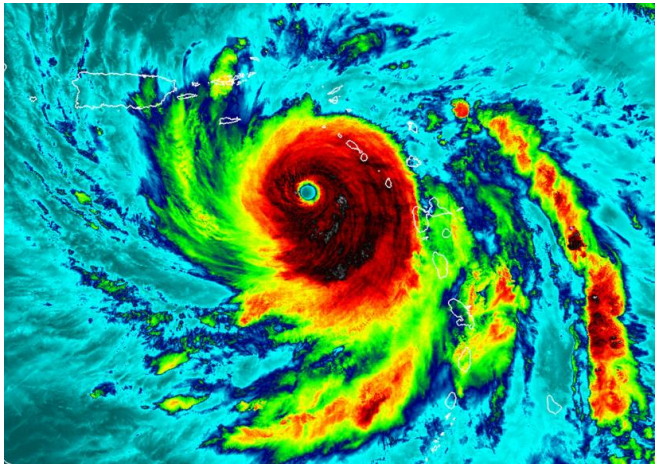


Figure 1. *VIIRS satellite image of Hurricane Maria nearing peak intensity on September 19 prior to landfall in Puerto Rico. Image courtesy of UW-CIMSS.*

Figure 2: *Figure 2. Storm damage from Hurricane Maria in Puerto Rico. Credit: Mario Tama, Getty Images.*

Puerto Rico was devastated by winds and floods. The National Oceanic and Atmospheric Administration (NOAA) estimate of damage in Puerto Rico and the U.S. Virgin Islands due to Maria is $90 billion USD, which makes Maria the third costliest hurricane in U.S. history, behind Katrina (2005) and Harvey (2017). In addition to widespread building damage (Figure 2 and 3) from winds (Figure 4), there was also up to 38 inches of rain and 40,000 landslides. More importantly, there was an estimated 3,000 deaths due to the storm.

Figure 3: *The region of interest selected for the data challenge contains significant building damage in the area near San Juan, Puerto Rico and to the south.*
Credit: Google Earth Pro and Brian Killough, EY.

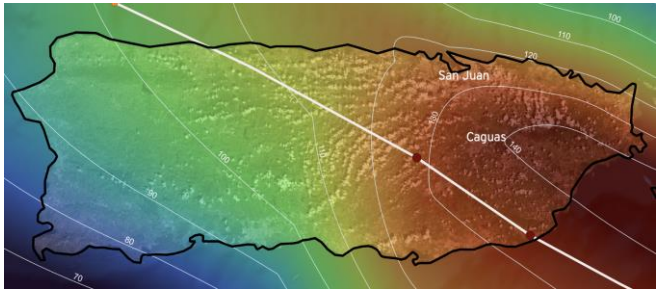


Figure 4. *Estimated peak wind gusts (mph) for Hurricane Maria over Puerto Rico. Track positions (6-hour) are linearly interpolated, and wind speeds (600-sec gusts) simulated based on Emanuel and Rotunno (2011), using CLIMADA open-source risk software. Note: Wind speeds correspond to model results and have not been validated with observations.*
Credit: Borja Reguero, University of California Santa Cruz.

Satellite Datasets

Maxar High-Resolution Data

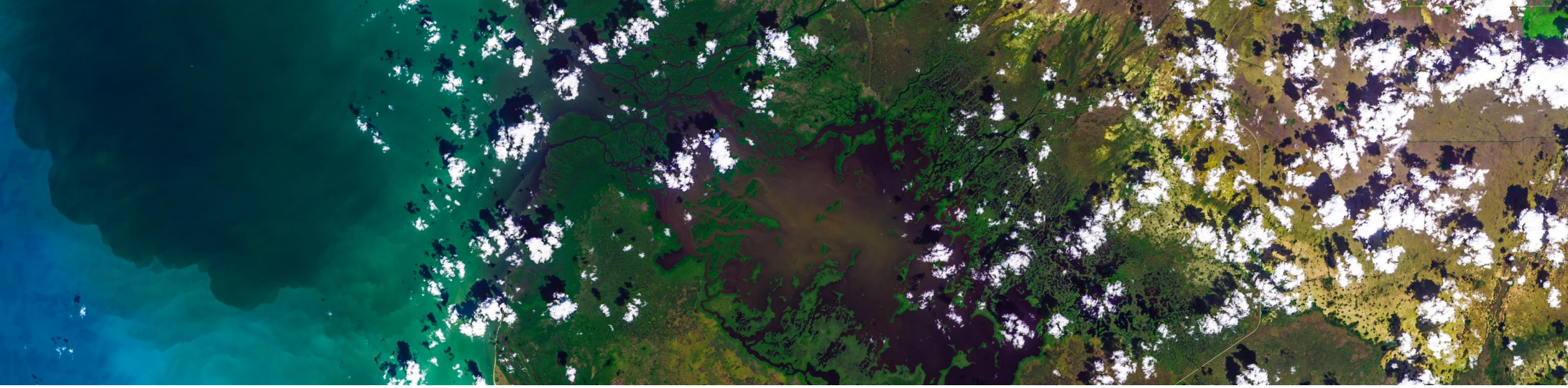
The satellite data used for this challenge has been provided by Maxar's Open Data Program. This high-value (estimated value:$30,000 USD) commercial data is the best available data on the market to view detailed storm damage from space. Maxar's Open Data program was created in January of 2017 to provide imagery and related data in support of humanitarian crises, with a focus on disaster response. The objectives of the Open Data Program are twofold: (1) humanitarian response efforts with timely, actionable information and (2) foster a community of practice around satellite imagery and Earth intelligence for disaster management. Maxar's Open Data Program provides its data free of charge and allows public use according to the Creative Commons 4.0 license. The analysis-ready imagery is processed to include orthorectification (ground position), atmospheric compensation, and pan-sharpening. Images are typically provided for before and after an event and are in common cloud optimized GeoTIFF (COG) format.

For our data challenge, we have selected two datasets (Figure 5) from Maxar's GEO-1 mission before (August 29, 2017) and after (October 12, 2017) the primary storm (September 20, 2017).

Figure 5: *The selected region of interest (328 sq.km.) covers an area to the south of San Juan, Puerto Rico. Data from Maxar's GEO-1 mission is provided for participants to develop AI object detection models. The data is shown here for before (Left - August 29, 2017) and after (Right - October 12, 2017) the storm. Even at these large scales, it is possible to see the impacts of the storm as many areas lost considerable vegetation (green on the left, brown on the right).*

This visual panchromatic data is 30-cm spatial resolution and provides an exceptional detailed view (Figure 6) of storm damage for the entire region of interest (328 sq.km.). Typical panchromatic data is created using a single grey-scale band, but Maxar has an algorithm that colorizes the panchromatic data to create an RGB (Red-Green-Blue) image. Using this data, it will be possible to develop object detection algorithms to identify structure types (houses vs. commercial buildings) and storm damage (e.g., roof loss).

Figure 6: *Maxar's GEO-1 satellite data (30-cm resolution) allows the detection of damaged and undamaged buildings from space. This example visual panchromatic product is from October 12, 2017, about one month after the storm had passed. Credit: Maxar Open Data Program.*



ESA's Sentinel-2 Moderate-Resolution Data

The launch of the European Copernicus Sentinel-2 missions in 2015 and 2017 provides optical data at 10-meter spatial resolution and a revisit every 10 days with one mission and every 5 days with two missions. This free and open data is readily available from the Microsoft Planetary Computer. But the issue with optical data is that it cannot penetrate clouds. So, if a cloud is over any given location, the data is not useable. In the case of Puerto Rico, clouds are persistent and often cause a significant loss in data. In addition, our ability to identify these clouds in the data is not perfect, so we often have issues with data that is contaminated by clouds which impacts the results of models. Though we are getting better at using satellite data, it is not perfect. So, for this challenge, we have provided a sample Sentinel-2 notebook that filters clouds and provides details on vegetation extent and its change due to storm damage.

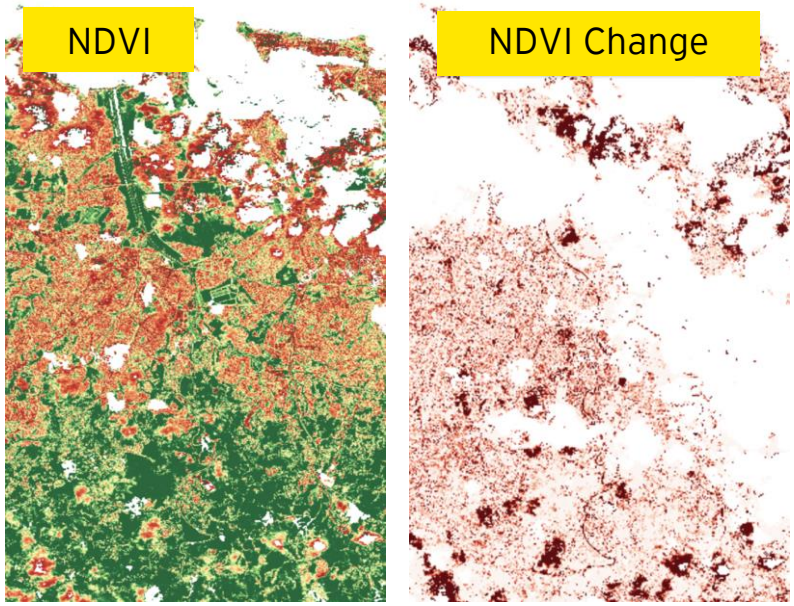


Figure 7: *Sentinel-2 post-storm vegetation index (left) and vegetation index change (right) over the Puerto Rico data challenge region. Red regions (right) are coincident with significant vegetation loss due to storm winds and rain. Large white regions in the center and to the north are either cloud or water and filtered from the change product. Credit: Brian Killough, EY.*

NASA's Landsat Moderate Resolution Data

Landsat data has been around since the early-1980s with a spatial resolution of 30-meters per pixel and a revisit of 16 days for one mission. We currently have two operational missions (Landsat-8 and Landsat-9) which yield 8-day revisits at any given location. Like Sentinel-2, this free and open data is readily available from the Microsoft Planetary Computer. In addition, this data is also "optical" so it cannot penetrate clouds. Therefore, we have developed a sample Landsat notebook that filters clouds and provides details on vegetation extent and its change due to storm damage. Participants will find the land change results from the Landsat data are like those from the Sentinel-2 data. Extensive vegetation damage is seen across the region and some areas were significantly impacted.

Other Datasets

In addition to satellite data, participants will be given building footprint data (Figure 8) over the region of interest. This data can be used to filter the large analysis region to find areas where there is a mix of residential and commercial buildings. These areas would be ideal to build training datasets.

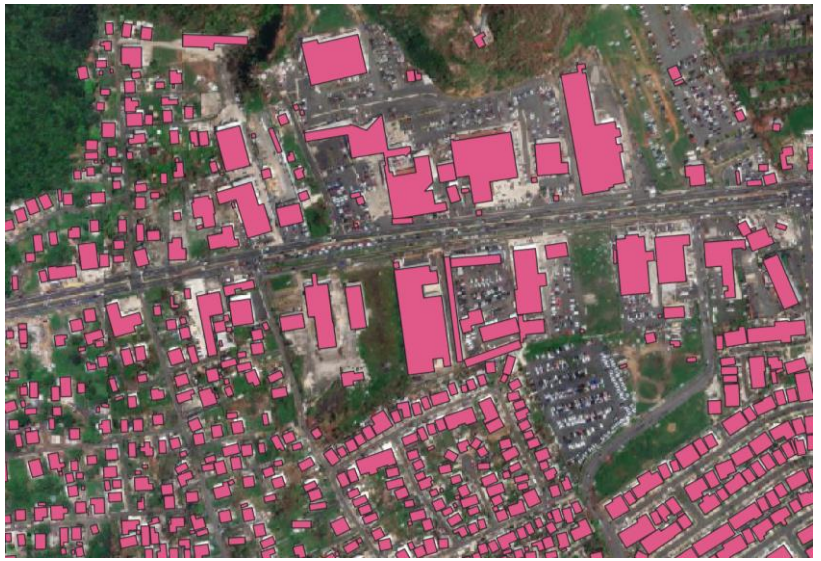


Figure 8. *A building footprint map can quickly identify regions where there is a mix of residential (smaller) and commercial (larger) buildings to support labeling of training data. In the image here, it is easy to see the uniform and smaller residential buildings on the bottom half of the image and the larger and non-uniform commercial buildings in the top half of the image. Credit: Bing Maps (https://github.com/microsoft/GlobalMLBuildingFootprints)*

Phase 1: Model

In Phase 1 of the challenge, participants will develop a machine learning / artificial intelligence (AI) model that can accurately detect damaged infrastructure using the satellite imagery of a cyclone-affected area. These would include residential buildings (both undamaged and damaged) and commercial buildings (both undamaged and damaged). To get started, participants are provided with a sample benchmark notebook that will demonstrate a simple damage assessment object detection model. This sample model is designed to detect both "damaged" and "undamaged" infrastructure within the realm of residential and commercial buildings.

In the sample model, we have utilized just the post-event image from Maxar's GEO-1 mission, which was captured on October 12, 2017. However, participants can also include pre-event images to expand the training dataset and heighten the number of class instances. In transitioning to the discussion of image processing and data management, it is crucial to keep in mind the potential challenges when dealing with notably large areas of analysis. The broad scope of the area for analysis may cause image processing to be quite demanding. A useful strategy for efficiently dealing with the data involves the creation of small grids. As demonstrated in our sample model notebook, we've created grids sized 512x512 pixels, and for compatibility with annotation tools, these were transformed to .JPG files, since GeoTIFF images are not supported.

Having ensured the image data is preprocessed efficiently, we now can turn our attention to the integral aspect of annotating this data to build an effective object detection model. Annotations play a pivotal role in enabling successful object detection tasks. Through marking or annotating objects within a particular selection of images, you are effectively showing the model what to find. Several annotation tools, including LabelMe, the VGG Image Annotator, and LabelImg are at your disposal. However, you need to exercise judgement as each annotation tool creates a distinct image annotation file format, which may or may not align with the needs of your model. As such, the appropriate tool for annotation should be determined by the specific requirements of the model you are aiming to develop. In the sample model notebook we have provided, we have used the YOLOv8 object detection model as our base model. YOLOv8 requires the annotations to be in a specific format i.e., with one *.txt file per image. Participants can refer to the sample notebook to understand the detailed format of the annotations which is required to build the model and generate a score on the challenge platform.

The final step involves the construction of an object detection model. Although the sample model displays a decent mean average precision (mAP) score of 0.34, you have the opportunity to improve the model's performance by focusing on classes where its object detection is failing and providing more labels for instances of that class. This could potentially enhance the mAP score. Furthermore, creating synthetic data is another strategy participants could adopt to boost their results. This artificial data creation has the potential to enrich a model's learning experience and subsequently escalate its predictive performance. In addition, you are encouraged to explore various other generative AI image models that could generate synthetic images. This strategy can enhance the training data quantity, which may subsequently result in improved outcomes.

Furthermore, we are more than open to participants venturing beyond the bounds of currently used models and exploring alternative object detection architectures or even devising completely unique strategies for assessing and distinguishing between damaged and undamaged buildings. Participants may use additional datasets to build the model, if the source of such data is publicly available and referenced.

How you'll be evaluated (Phase 1)

An out-of-sample validation dataset will be provided to participants. Your submission/prediction (.zip file), will be compared with the ground truth file and a metric mAP (Mean Average Precision) will be generated to evaluate the performance of the model.

We will conduct eligibility screening of the prospective global semi-finalists to confirm eligibility to proceed to Phase 2 of the competition (and thus become a global semi-finalist) and ask them to submit a content package to support their final submitted Phase 1 solution on the platform, which will be used to confirm their eligibility to proceed to Phase 2 of the competition. The content package consists of:

1. The code and model used in the Phase 1 submission.
2. A document to describe the model development approach.
3. A Curriculum Vitae (CV) and most recent academic transcript of the individual or each individual that is part of a team.

We will select the **top 10 performers** based on model accuracy and design in achieving the objective of the challenge, which is to produce a machine learning/AI data model with the ability to detect key infrastructure and land features and their associated damage due to a tropical storm.

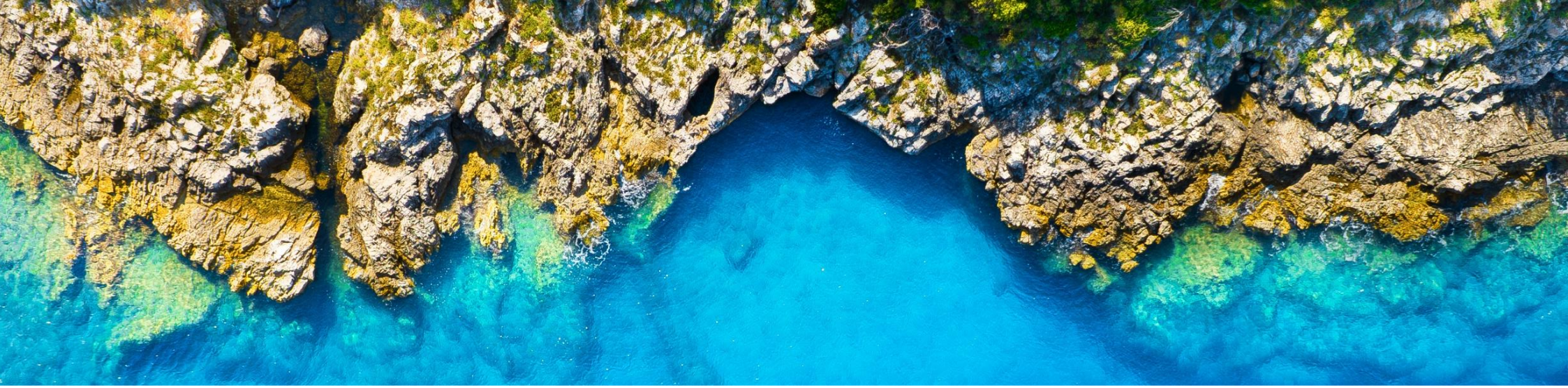
Phase 2

If selected as a semi-finalist, you will move forward to participation in Phase 2 of the data challenge. At this stage, you will develop a written document (4 pages or less) and a video (less than 5 minutes). The document should describe how the proposed solution from Phase 1 could be practically applied by local beneficiaries to assess coastal infrastructure damage and socioeconomic impact, support post-disaster recovery, and address coastal resilience plans and climate change risk for future storms. It should also include suggestions for the use of other datasets and proposed steps for developing additional models and data science solutions. Participants should follow the provided template and use a strategic and well-structured approach while infusing creativity and considering generative-AI tools for completeness and enhanced impact.

In the end, this 4-page document will form the basis for a paper to be submitted by the finalists to the International Geoscience and Remote Sensing Symposium (IGARSS) conference in Athens, Greece in July 2024. Finalists will be invited to attend the IGARSS conference and present their paper in a dedicated oral session. In addition, finalists will be asked to upload their winning models and reports to a designated University of California Santa Cruz open-source software repository for further dissemination to support potential beneficiaries and Maxar's Open Data Program.

How you'll be evaluated (Phase 2)

Global semi-finalists must submit their Phase 2 content package by **21 April 2024**. Five global finalists will be evaluated based on the following selection criteria: critical thinking, methodology and communication. The judging panel will select one global winner and two runners-up.



Conclusions

The 2024 EY Open Science Data Challenge is an excellent opportunity for university students and early career professionals to develop much needed open-source solutions to support disaster response and recovery in coastal communities. Today, automated solutions using high-resolution commercial data are not available for common public use but can provide a critical lifeline to vulnerable regions. When combined with data from Maxar's Open Data Program, your innovative AI models may end up being a significant contribution to local beneficiaries around the world. We look forward to seeing your results and wish you the best of luck.

Questions?
Contact us at **datachallenge@ey.com**

Can't get enough information?

Here are some references that can help you with the challenge:

1. *Transformation and Innovation in the Wake of Devastation – An Economic and Disaster Recovery Plan for Puerto Rico*, Puerto Rice Science, Technology and Research Trust (https://prsciencetrust.org/), 2018.
2. *National Hurricane Center Tropical Cyclone Report*, Hurricane Maria (AL152017), https://www.nhc.noaa.gov/data/tcr/AL152017_Maria.pdf
3. Jean et.al., *Combining satellite imagery and machine learning to predict poverty*. Science 353, 790–794, 2016.
4. Kim et.al., *Disaster assessment using computer vision and satellite imagery: Applications in detecting water-related building damages*, Front. Environ. Sci. 10:969758, 2022.
5. Polina Berezina and Desheng Liu, *Hurricane damage assessment using coupled convolutional neural networks: a case study of hurricane Michael*, Geomatics, Natural Hazards and Risk, 2022.
6. Quoc Dung Cao and Youngjun Choe, *Post-Hurricane Damage Assessment Using Satellite Imagery and Geolocation Features*, 2020.

EY | Building a better working world

EY exists to build a better working world, helping to create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

ey.com