
How can AI and data help cool a steaming city?
Join the EY Open Science AI & Data Challenge at challenge.ey.com/2025
The better the question. The better the answer. The better the world works.
EY
Shape the future with confidence

Guidance and Suggestions for Participants of the 2025 EY Open Science AI and Data Challenge

Welcome to the 2025 EY Open Science AI & Data Challenge! Held annually, the challenge gives thousands of early-career professionals and university students the opportunity to use data, artificial intelligence (AI) and computing technology to create solutions that address critical climate issues, building a more sustainable future for society and the planet. By joining this challenge, you will become part of an important community who has decided to engage in activism using space tech and AI to solve an important global sustainability issue at scale.

Background

The 2025 challenge focuses on a phenomenon known as the Urban Heat Island (UHI) effect. This issue can result in temperature variations between rural and urban environments that exceed 10-degrees Celsius in some cases, and can cause significant health-, social-, and energy-related issues. In many industrialized countries, heat events account for more than all other natural hazards combined. [References 1,2] Urban areas are most susceptible to heat stress due to the high density of buildings, lack of vegetation (green space), lack of water bodies, and waste heat from industry and transportation. Those particularly vulnerable to heat-related problems include young children, older adults, outdoor workers and low-income populations. According to the World Health Organization (WHO), over 55% of the world's population lives in urban areas with that proportion expected to increase to 68% by 2050. According to the Intergovernmental Panel on Climate Change (IPCC), climate change is expected to cause increasing temperatures which, when combined with population increases, will put more citizens in danger of negative health effects related to extreme heat.

Challenge Goals

Though this topic has been widely studied and documented, it is still unknown to many people and there is a need for increased awareness and accurate open-source models. The primary goal of the data challenge will be to develop a digital model to predict the locations and severity of the UHI effect and to understand the drivers of this phenomenon. This micro-scale machine learning and AI model will be developed using near-surface air temperatures, building footprint data, weather data, and optical satellite data to identify vegetation and water proximity and surface state. A secondary goal of the challenge is to address the practical application of the output models for local decision-makers, scaling such solutions to other cities around the world, and considerations for additional datasets that could improve model accuracy and evaluate socioeconomic impact.

The UHI effect is typically modeled using coarse satellite-based measurements of surface temperatures such as those produced by National Aeronautics Space Agency's (NASA) Landsat mission. Such models do not reflect the near-surface micro-climate air temperature that most impacts the human population. Therefore, there is a need for simplified, yet accurate, UHI models to allow urban planners and city managers to better understand the location and severity of UHI issues in their city and the drivers of urban heating. Such models will bring attention to this important global sustainability problem and may force decisions to alter existing urban plans or influence future ones. In addition, such models can be used to understand the impact of natural areas and support their preservation and future expansion.

Challenge Approach

Participants will be provided with many datasets to consider for their models. Their ability to determine which datasets and parameters are the most important for model accuracy will determine the finalists. These datasets include:

- Ground-based air temperatures
- Building footprints
- European Sentinel-2 optical satellite data (multispectral)
- NASA Landsat optical satellite data (surface temperature)
- Local weather (temperature, relative humidity, wind speed and direction, solar flux)

These datasets will be used to develop an open-source machine learning model to predict UHI hotspots at micro-scales (meters) across the city.

Additionally, the model should be designed to discern and highlight the key factors that contribute significantly to the existence of these UHI hotspots within city environments.

Participants will be given ground-level air temperature data in an index format, which was collected on July 24, 2021 by Climate, Adaptation, Planning, Analytics (CAPA) Strategies using automobile traverses across the Bronx and Manhattan regions of New York City in the United States. The data was collected in the afternoon between 3:00 pm and 4:00 pm. This dataset includes time stamps, traverse points (latitude and longitude) and the corresponding UHI Index values for 11,229 data points. These UHI Index values comprise the target parameter dataset for your model. Participants will then consider the use of building footprint data, Sentinel-2 satellite data, Landsat satellite data, and local weather data as feature weather data as feature datasets within their models. As a note, participants are allowed to use additional datasets for their models, provided those datasets are open and available to all public users, and the source of such datasets are referenced in the model.

Required Skills

Participants in this challenge can benefit from a basic understanding of statistics and Python programming, but there are no prerequisites for participation. Participating in this challenge can improve skills in machine learning, AI, data science, and working with satellite datasets. The data challenge has been designed to attract beginners and those less familiar with AI and Python programming.

Computing Requirements

This data challenge was designed to run on a local computer with common computing resources (e.g., 4 cores, 32 GB memory). The configuration should include a Python programming environment and a code development tool (e.g., Jupyter). It is also possible to participate in this challenge using common cloud-based environments, such as those available from Microsoft (Azure), Google (Google Cloud, Earth Engine) or GitHub (Codespaces).

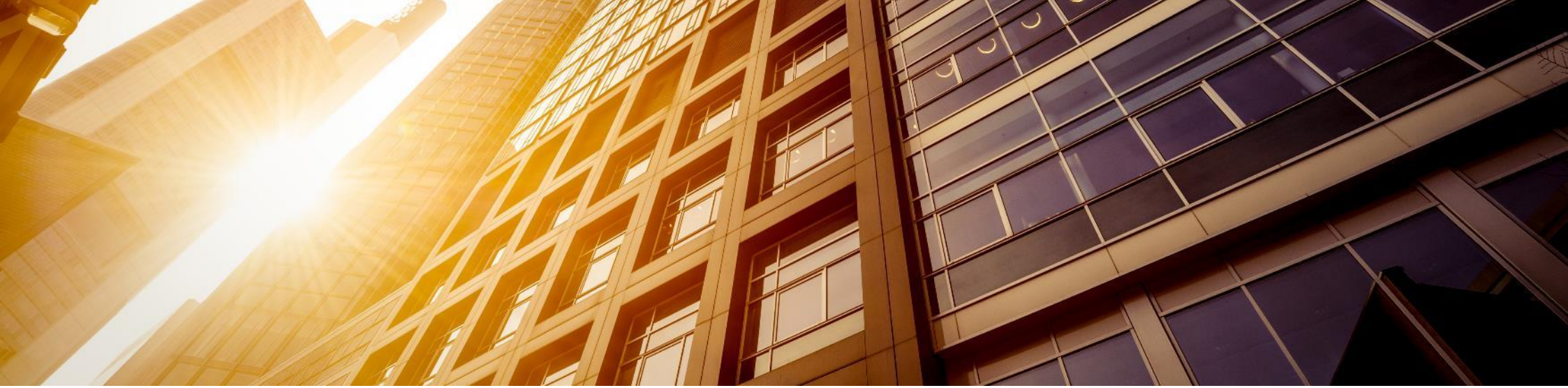
Dataset Summary

Temperature Data

Data was collected by CAPA Strategies using a ground traverse (Figures 1 and 2) with vehicles and bicycles on a single day in the summer of 2021. This data collection effort resulted in 11,229 data points which will be the focus for this data challenge.

Figure 1. *Ground-level temperature data was collected by CAPA Strategies and community volunteers using temperature recording devices mounted to cars and bikes. This data collection campaign was part of the international "Heat Watch" program.*
Credit: CAPA Strategies, LLC.

Figure 2. *Data was collected across Manhattan and the Bronx in New York City on July 24, 2021, between 3:00 pm and 4:00 pm. The data (11,229 points) was converted to a UHI Index for the purpose of this data challenge. The image above shows areas of lower UHI index values (cool spots) in yellow and areas of higher UHI index values (hot spots) in dark red.*
Credit: Brian Killough, EY



For this challenge, we have created a unique UHI index for every data point location. This index reflects the local temperature at the data point location compared to the city's average temperature across all data points during the time window of the data collection. Though this is not a perfect approach to modelling the complex urban heating dynamics of a city, it will provide a reasonably accurate model of urban heat islands in the city at the time of day consistent with the data collection. In an ideal situation, time series data would be collected at thousands of locations across the city and weather data (e.g., wind speed, wind direction, solar flux) would be added to the model to yield more accuracy and allow for consideration of natural variability.

UHI Index = (Temperature at a given location) / (Mean temperature for all locations)
Note: UHI index calculations only used data within a 1-hour data collection window

The chosen UHI index serves as a crucial metric for assessing the intensity of heat within different urban zones of the city. For comparison, most literature calculates a UHI index based on temperature differences between inner city locations and rural locations far outside of the city. Since we did not have data from rural locations, we created a unique UHI index that reflects the variability of temperatures within our collected dataset and time collection window. As an example, a UHI index value of 1.0 suggests the local temperature is the same as the mean temperature of all collected data points. UHI index values above 1.0 are consistent with hotspots above mean temperature values and UHI index values below 1.0 are consistent with cooler locations in the city. Participants will use their models to predict these UHI values across the city. Figure 3 shows a histogram of UHI values for the data challenge.

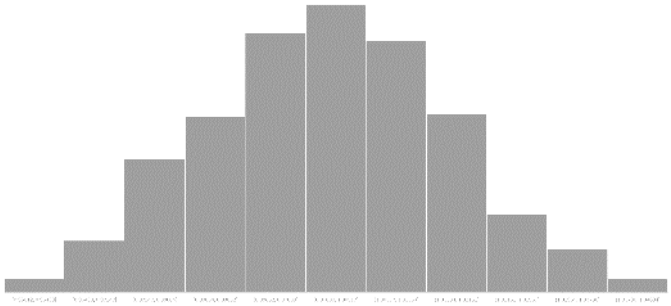


Figure 3. *Histogram of UHI values (12,309 total) representing the target dataset for the challenge. Most of the data is close to the mean temperature (UHI=1.0) but there is variability suggesting cooler regions (UHI=0.956, minimum) and hotter regions (UHI=1.046, maximum) within the bounds of the data collection region. Credit: Brian Killough, EY.*

The range of collected temperatures had a maximum difference of 7.5 Degrees-Fahrenheit (4.2 Degrees-Celsius). Though this is much lower than known global extremes (10 Degrees-Celsius difference), the collected data does allow for the identification of urban heat islands. When converted to UHI index values (range of 0.956 to 1.046), this yielded a 9% variation in UHI values across the data collection region.

Building Footprints

Many studies suggest that the density of buildings in a city influences ground temperatures and ultimately contribute to the UHI issue. [Reference 3] This effect is typically driven by buildings blocking the flow of air and adding waste heat. For this challenge, we have provided a building footprint dataset. Such information could be used in your digital model as a feature that drives local urban heating.

Figure 4. *Higher building density is known to increase local air temperatures and contribute to the UHI effect. Data challenge participants will be given building footprint data for consideration in their digital models.*
Credit: Cornell University Geospatial Information Repository (https://cugir.library.cornell.edu/)

Satellite Data

Satellite data can be quite beneficial for understanding the location and severity of urban heat islands. Missions such as NASA's Landsat and European Space Agency's (ESA) Sentinel-2 provide continuous global coverage at regular revisit rates and open and free access to the datasets on cloud computing frameworks. These datasets provide a unique look at urban areas and often measure regions not covered by in-situ measurements. But these datasets also have limitations that must be considered when using them for UHI models. For example, these optical satellites cannot penetrate clouds or vegetative cover, so this often leads to data gaps and decreased data quality. In addition, these satellites have consistent acquisition times near the middle of the day which does not usually coincide with collected ground data or the times of maximum urban heating. So, some of these factors should be considered when using satellite data in UHI models.

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 (https://planetarycomputer.microsoft.com/catalog). But optical data cannot penetrate clouds, so it is necessary to filter out clouds or select scenes that have very low levels of cloud cover. For this challenge, we have provided a sample Sentinel-2 Python notebook that selects low-cloud scenes and creates a median mosaic without cloud contamination. This product can be used to assess the impacts of vegetation extent, water, or urban density on urban heating. It is well known [Reference 4] that proximity to vegetation (green space), proximity to water, and local urban density contribute to the effects of urban heating. Figure 5 shows an example of a Sentinel-2 mosaic product illustrating the spatial variation of the Normalized Difference Vegetation Index (NDVI) over our data challenge region (shown in white).

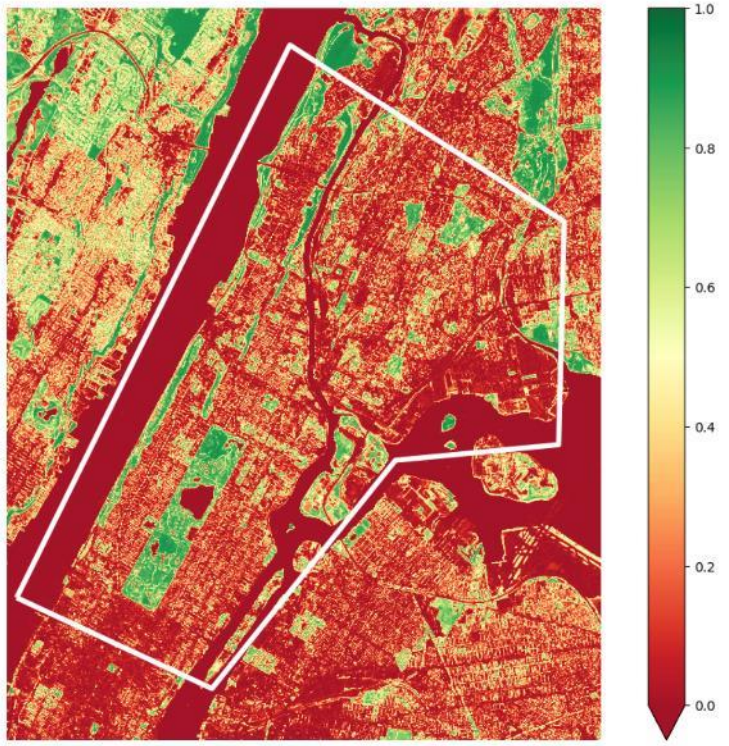


Figure 5. *Sentinel-2 Normalized Difference Vegetation Index (NDVI) over the data challenge region (New York City). This product is based on a median mosaic from June-August 2021 which removes cloud contamination. Areas of light green or dark green are consistent with the presence of vegetation. Areas of dark red are consistent with dense urban environments or water. This information can be used in your digital model as vegetation, urban density and water can impact local urban heating. Credit: Brian Killough, EY (using data from ESA's Sentinel-2 mission on the Microsoft Planetary Computer).*

NASA's Landsat mission has been in continuous operation since 1984. This mission includes a unique thermal infrared sensor that can measure Land Surface Temperature (LST). With two missions available at any given time, this data is freely available every 8 days for the entire world from the Microsoft Planetary Computer (see link above). This data has been widely used to assess UHI variations in many studies, but its limitations should be considered. For example, LST only measures the surface temperature of what can be seen from space. That includes tops of buildings, open streets or open land areas. It does not reflect the air temperature conditions near the surface where humans exist. The coarse resolution of the LST product (100 meters) often mixes surface responses from tops of buildings and open roads or land which adds error to the data. In addition, the accuracy of the LST data depends strongly on corrections for atmospheric effects and an accurate estimate of surface emissivity. Finally, one needs to consider the timing of the data product. For our data challenge case, no Landsat data is available on the same day as the ground data collection. Also, the Landsat data acquisition time is about 11:30 am which does not exactly match the time of the day when ground traverse data was collected (3:00 pm to 4:00 pm).

For this challenge, we have provided a sample Landsat LST Python notebook that selects a clear scene (June 16) close to the date of the ground data collection (July 24). This product can be used to assess building surface and ground surface temperatures which may contribute to local air heating in a UHI model. Figure 6 shows an example of a Landsat LST product illustrating the spatial variation of surface temperatures over our data challenge region.

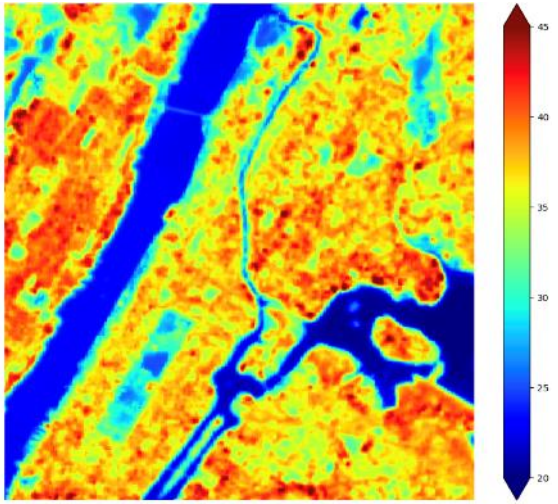


Figure 6. *Landsat Land Surface Temperature (LST) over the data challenge region (New York City) on June 16, 2021. Areas of red are consistent with high surface temperatures (above 40C) such as the tops of buildings with dark coatings and black pavement. Areas of blue are consistent with areas of lower surface temperature (below 30C) such as water and vegetation. The locations and severity of these extremes may be an important feature in a UHI model. Credit: Brian Killough, EY (using data from NASA's Landsat mission on the Microsoft Planetary Computer).*

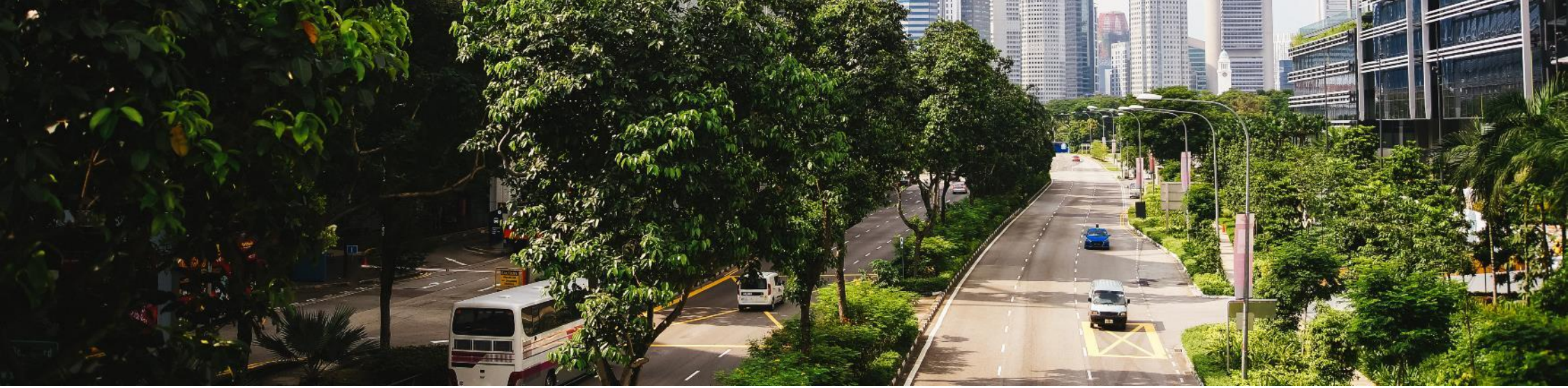
Local Weather Data

Local weather data can also be an important feature in a UHI model. For example, high wind speeds can cause mixing of near-surface air and reduce UHI intensity and variability. Wind direction can also impact localized heating as building configurations can block local mixing of near-surface air and cause increased urban heating. Finally, solar flux can be blocked by clouds or building configurations which could reduce local surface heating and near-surface air heating. For this challenge, we have provided a local weather dataset from the New York State Mesonet (https://nysmesonet.org). Two weather stations, located at the southern and northern ends of the ground traverse data region, collected near-surface data (2-meters height) every 5 minutes during the day. This data includes temperature, relative humidity, wind speed, wind direction, and solar flux. This data may be useful for your AI model.

Model Development and Evaluation

Participants will develop a machine learning / AI model [example in References 4 and 5] that can accurately predict UHI index values at specific locations in New York City. To get started, participants are provided with a sample benchmark Python notebook that will demonstrate a simple UHI prediction model. This sample model is designed to use UHI index data from the target dataset that was compiled from ground traverse data collected on July 24, 2021. Sentinel-2 satellite data spectral bands are used as the feature dataset in the sample model. The model uses a common 70/30 training and testing split to evaluate model performance. The sample model produces modest results to allow significant improvements by data challenge participants. Some suggestions for improved model performance include: consideration of additional datasets (Landsat LST, building footprints, weather), consideration of additional Sentinel-2 bands or spectral indices, proximity of satellite data features to data collection points, building density in proximity to data collection points, regression algorithms, and hyperparameter tuning. As a note, participants are allowed to use additional datasets for their model, provided those datasets are open and available to all public users and the source of such datasets are referenced in the model.

In the end, participant models will be tested against known UHI index values (validation dataset) for a portion of the region that is not included in the target dataset. Predictions on the validation dataset shall be saved in a Comma-Separated Values (CSV) file and uploaded to the challenge platform to get a score on the ranking board, which you can improve over the course of the challenge with subsequent model revisions and submissions. Up to ten semi-finalists will be selected based on the highest overall accuracy in the form of a coefficient of determination (R-squared). These semi-finalist models will be further reviewed for compliance, innovation and model efficiency to select up to 5 finalist teams.



Business Plan Development and Evaluation

Up to five finalist teams (or individuals) will be asked to develop a practical business plan that describes how their AI model could be applied by local beneficiaries to address the impacts and concerns of urban heating. Finalists will be required to submit a written document (4 pages or less) and a video (less than 5 minutes) that includes the following: their analysis approach, considerations for scaling such solutions to other cities or other portions of New York City, additional datasets that could improve model accuracy if given more time and resources, socioeconomic impact on vulnerable communities, impact on energy demand, and practical application for local governments or urban planning decision-makers. 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.

Conclusions

The 2025 EY Open Science AI & Data Challenge is an excellent opportunity for young professionals to develop open-source solutions that can help bring cooling relief to vulnerable communities. Such comprehensive digital models are not available for common public use and innovative AI models may end up lending a significant contribution to beneficiaries around the world. Entrants with top phase 1 scores and exceptional phase 2 business plans will take home cash prizes and receive an invitation to an exciting awards celebration. We look forward to seeing your results and wish you the best of luck.

References

1. Poumadère M, Mays C, Le Mer S, Blong R. The 2003 Heat Wave in France: Dangerous Climate Change Here and Now. Risk Analysis, Vol. 25, Issue 6, Dec 2005, pp. 1483-1494. https://doi.org/10.1111/j.1539-6924.2005.00694.x
2. Borden, K.A., Cutter, S.L. Spatial patterns of natural hazards mortality in the United States. International Journal of Health Geographics, 7, Article 64 (2008). https://doi.org/10.1186/1476-072X-7-64
3. Lee S, Kim D. Multidisciplinary Understanding of the Urban Heating Problem and Mitigation: A Conceptual Framework for Urban Planning. Int J Environ Res Public Health. 2022 Aug 18;19(16):10249. https://doi.org/10.3390/ijerph191610249
4. Shandas, V., Voelkel, J., Williams, J., & Hoffman, J., (2019). Integrating Satellite and Ground Measurements for Predicting Locations of Extreme Urban Heat. Climate, 7(1), 5. https://doi.org/10.3390/cli7010005
5. Voelkel, J., & Shandas, V. (2017). Towards Systematic Prediction of Urban Heat Islands: Grounding Measurements, Assessing Modeling Techniques. Climate, 5(2), 41. https://doi.org/10.3390/cli5020041

Questions?

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

EY | Building a better working world

EY is building a better working world by creating new value for clients, people, society and the planet, while building trust in capital markets.

Enabled by data, AI and advanced technology, EY teams help clients shape the future with confidence and develop answers for the most pressing issues of today and tomorrow.

EY teams work across a full spectrum of services in assurance, consulting, tax, strategy and transactions. Fueled by sector insights, a globally connected, multi-disciplinary network and diverse ecosystem partners, EY teams can provide services in more than 150 countries and territories.

All in to shape the future with confidence.

EYG no. 010461-24Gbl
ED None

ey.com