The Open Cities AI Challenge

As urban populations grow, more people are exposed to the benefits and hazards of city life. One challenge for cities is managing the risk of disasters in a constantly changing built environment. Buildings, roads, and critical infrastructure need to be mapped frequently, accurately, and in enough detail to represent assets important to every community. Knowing where and how assets are vulnerable to damage or disruption by natural hazards is key to disaster risk management (DRM).

The Global Facility for Disaster Reduction and Recovery (GFDRR) is a global partnership that provides knowledge, funding, and technical assistance towards achieving the vision of a world where resilient societies manage and adapt to ever-changing disaster and climate risk, and where the human and economic impact of disasters is reduced. Recognizing the significant potential of machine learning and geospatial technologies for DRM, GFDRR Labs is engaging the expert community through initiatives like this challenge to advance the creation of global public goods for improving disaster resilience.

The goal of this challenge is to accelerate the development of more accurate, relevant, and usable open-source AI models to support mapping for disaster risk management in African cities.

Better performing and responsibly used AI systems can provide more accurate, faster, and lower-cost approaches to assessing risk and protecting lives and property.

Learn more about the Challenge which ran from Dec 2019 - Mar 2020 and see the open-sourced winning results at the DrivenData competition site.

Problem and Dataset Description

In this challenge, you will be segmenting building footprints from aerial imagery. The data consists of drone imagery from 10 different cities and regions across Africa. Your goal is to classify the presence or absence of a building on a pixel-by-pixel basis.

The features in this dataset

The features in this dataset are the images themselves and the building footprints in the GeoJSONs, which can be used to train a building segmentation model. All training data (with the exception of the labels for images of Zanzibar) are pulled from OpenStreetMap.

Images

The images are stored as large Cloud Optimized GeoTiffs (COG). Spatial resolution varies from region to region. All images include 4 bands: red, green, blue and alpha. The alpha band can be used to mask out NoData values.

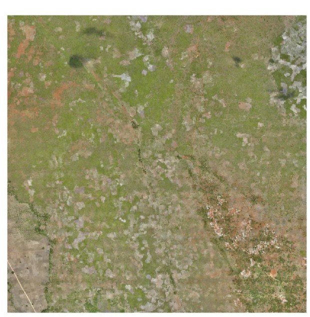
Given that the labels vary in quality (e.g. how exhaustively an image is labeled, how accurate the building footprints are), the training data have been divided up into `tier 1` and `tier 2` subsets. The tier 1 images have more complete labels than tier 2. **We encourage you to begin training models on the tier 1 data before trying to incorporate tier 2.**

All training images have been reprojected to the appropriate UTM zone projection for the region that they represent.

TRAINING DATA SUMMARY

City	Data class	Scene count	AOI area (sq km)	Building count	Total building size (sq km)	Average building size (sq m)	Building ratio (portion of area covered by bldgs)
acc	train_tier_1	4	7.86	33585	2.85	84.84	0.36
dar	train_tier_1	6	42.90	121171	12.02	99.20	0.28
dar	train_tier_2	31	223.28	571047	53.77	94.16	0.24
gao	train_tier_2	2	12.54	15792	1.28	81.05	0.10
kam	train_tier_1	1	1.14	4056	0.22	53.14	0.19
kin	train_tier_2	2	1.01	2357	0.17	71.29	0.17
mah	train_tier_2	4	19.40	7313	1.51	206.48	0.08
mon	train_tier_1	4	2.90	6947	1.05	150.71	0.36
nia	train_tier_1	1	0.68	634	0.03	47.43	0.04
nia	train_tier_2	2	2.46	7444	0.47	62.76	0.19
ptn	train_tier_1	2	1.87	8731	0.64	72.73	0.34
znz	train_tier_1	13	102.61	13407	1.62	120.83	0.02

TIER 1 SAMPLE

Area (abbreviation)	Scene ID	Thumbnail	Resolution	Pixel width x height
Accra (acc)	665946		2 cm	84466 x 150147
Kampala (kam)	4e7c7f		4 cm	39270 x 40024
Pointe-Noire (ptn)	f49f31		20 cm	6605 x 4185
Zanzibar (znz)	aee7fd		7 cm	40551 x 40592

TIER 2 SAMPLE

Area (abbreviation)	Scene ID	Thumbnail	Resolution	Pixel width x height
Ngaoundere (gao)	4f38e1		5 cm	56883 x 59802
Mahe Island (mah)	71e6c2		7 cm	52517 x 91616
Dar es Salaam (dar)	ef8f27		7 cm	50259 x 48185

Labels

Each image in the train set corresponds to a GeoJSON, where labels are encoded as FeatureCollections. `geometry` provides the outline of each building in the image. There are additional fields like `building:material` which you are free to use in training your models, but keep in mind none of this metadata will be provided for the test chips. Your goal is only to classify the presence (or lack thereof) of a building on a pixel-by-pixel basis.

`train_metadata.csv` links the each image in the train set with its corresponding GeoJSON label file. This csv also includes the region and tier of the image. Note that region information is not provided for the test set.

Label GeoJSON files have been clipped to the extents of the non-NoData portions of the images, all building geometries will overlap with image data.

Example training data image and labels

As previously mentioned, tier 1 labels are generally more accurate those in tier 2.

Tier 1 label example (Kampala)	Tier 2 label example (Dar es Salaam)
	

Data format

The metadata for the competition datasets are stored in SpatioTemporal Asset Catalogs (STACs). A STAC is a standardized specification that allows you to easily query geospatial imagery and labels. STACs are comprised of a series of JSON files that reference each other as well as the geospatial assets (e.g. imagery, labels) that they reference. PySTAC is a simple Python library for manipulating working with STAC objects.

TRAIN STACS

Each of these catalogs contains a collection for each of the regions that are included in that subset of training data. Within each region are the COGs and GeoJSON files that are represented as STAC Item and LabelItems, respectively. The JSON files (e.g. `catalog.json`, `collection.json`, `b15fce.json`) include spatial and temporal information about the assets that objects and assets included below them. They also reference their 'child' and 'parent' objects, enabling you to easily traverse the file tree.

`train_tier_1` and `train_tier_2` STACs have identical formats but describe different data (tier 1 and 2 training data, respectively).

Sample file structure

```
train_tier_1/
├── catalog.json
├── acc
│    ├── collection.json
│    ├── 665946
│    │    ├── 665946.json
│    │    └── 665946.tif
│    ├── 665946-labels
│    │    ├── 665946-labels.json
│    │    └── 665946.geojson
│    ├── a42435
│    │    ├── a42435.json
│    │    └── a42435.tif
│    ├── a42435-labels
│    │    ├── a42435-labels.json
│    │    └── a42435.geojson
│    ├── ...
|
├── dar
│    ├── collection.json
│    ├── b15fce
│    │    ├── b15fce.json
│    │    └── b15fce.tif
│    ├── b15fce-labels
│    │    ├── b15fce-labels.json
│    │    └── b15fce.geojson
│    ├── ...
|
├── ...
```

Niamey tier 1 collection json

```
{
    "id": "nia",
    "stac_version": "0.8.1",
    "description": "Tier 1 training data from nia",
    "links": [
        {
            "rel": "item",
            "href": "./825a50/825a50.json",
            "type": "application/json"
```

```
        },
        {
            "rel": "item",
            "href": "./825a50-labels/825a50-labels.json",
            "type": "application/json"
        },
        {
            "rel": "root",
            "href": "../catalog.json",
            "type": "application/json"
        },
        {
            "rel": "parent",
            "href": "../catalog.json",
            "type": "application/json"
        }
    ],
    "extent": {
        "spatial": {
            "bbox": [
                [
                    2.000607112710697,
                    13.570445755015827,
                    2.0089069498293726,
                    13.583969811536608
                ]
            ]
        },
        "temporal": {
            "interval": [
                [
                    "2019-10-29T00:00:00Z",
                    null
                ]
            ]
        }
    },
    "license": "various",
    "stac_extensions": [
        "label"
    ]
}
```


STAC objects must include a valid datetime string but temporal was not available for all images. The date "2019-10-29" was used as a placeholder for all scenes for which date of capture was unavailable.

A peek inside the Niamey tier 1 training data JSON shows how STAC encodes the metadata for this pair of image and label. The links with all STACs for this competition are relative and self contained. For ease of use, the assets are located alongside the STAC JSON files in the file tree. This means that a COG or GeoJSON file will always be located in the same directory as its corresponding item JSON file.

About ML for Disaster Risk Management (DRM)

As urban populations grow, managing this growth in a way that fosters cities' resilience to natural hazards and climate change becomes a greater challenge that requires detailed, up-to-date geographic data of the built environment. Buildings, roads, and critical infrastructure all need to be mapped frequently, accurately, and in enough detail to represent assets important to every community.

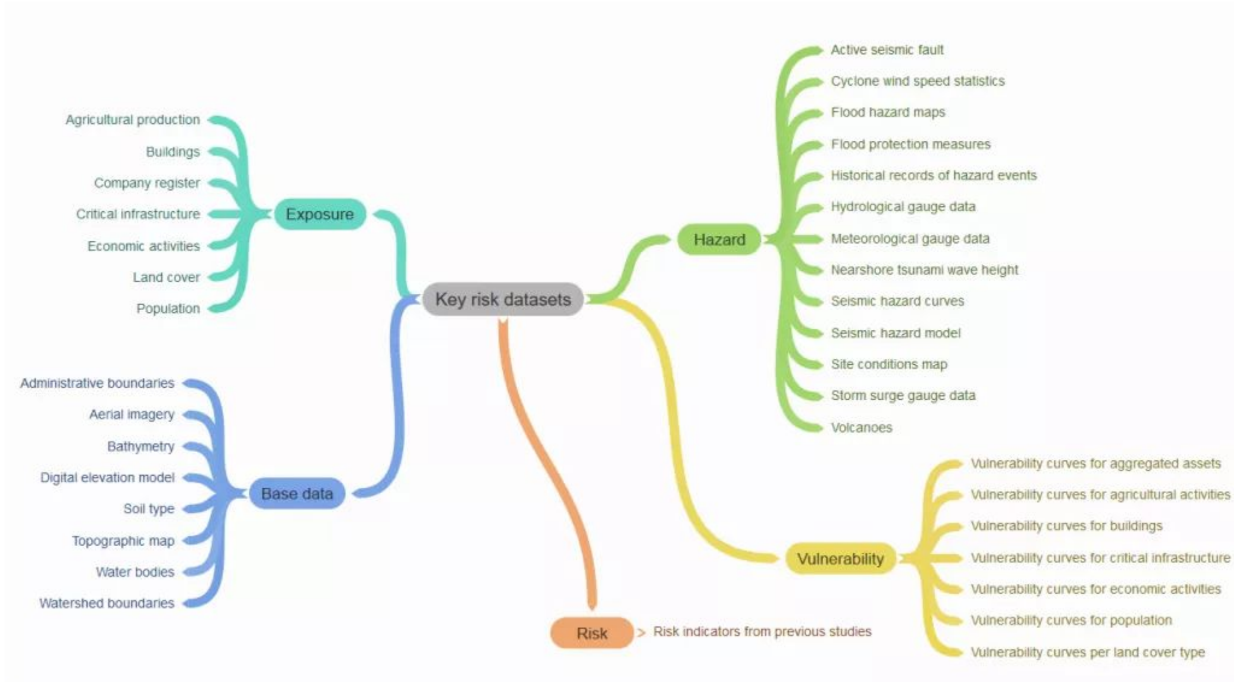
Knowing where and how assets are exposed and vulnerable to damage or disruption by natural hazards is key to disaster risk management (DRM). For example, new building construction into floodplains exposes their inhabitants to flood risk which may be increasing due to sea level rise and extreme weather events. A building's location, shape, and construction materials may make it more vulnerable to earthquake or wind damage than nearby buildings. Managing disaster risk is most effective with an accurate, detailed, and current understanding of the building stock of a city.

WHAT IS DISASTER RISK?

Disaster risk is a combination of three components: hazard, exposure, and vulnerability. Data from each of these categories can be used to paint a picture of risk in a certain location and over time.

- **Hazard**: a potentially destructive physical phenomenon (e.g., earthquake, windstorm, flood).
- **Exposure**: the location, attributes, and value of assets that are important to communities (people, buildings, factories, farmland, etc.) and that could be affected by a hazard.
- **Vulnerability**: the likelihood that assets will be damaged/destroyed/affected when exposed to a hazard. For example, a building with multiple floors may be more vulnerable to shaking from an earthquake and more likely to collapse than a one-story building. Another example, an elderly person may be more vulnerable to the impacts of flooding because s/he has a harder time evacuating or moving quickly.

TYPES OF RISK DATA



Source: understandrisk.org/vizrisk/what-is-risk/

Keeping up-to-date about buildings and their attributes (see the GED4ALL exposure data schema for more on what is used in DRM) is a particularly difficult task across Africa where populations are expected to double in the next 25 years. Cities are rapidly growing both denser and more spread-out with a diverse mix of formal and informal building construction. Addressing this challenge requires innovative, open, and dynamic data collection and mapping processes.

Open Cities Africa

The Open Cities Africa program, an initiative of the Global Facility for Disaster Reduction and Recovery (GFDRR), has: 1) established new information infrastructure for disaster risk management and urban resilience planning, 2) fostered local OpenStreetMap (OSM) communities, and 3) collected up-to-date open spatial data related to disaster risk in 11 cities across Africa.

Similar programs like Dar Ramani Huria engaged local communities and students in Dar es Salaam, Tanzania to create highly accurate maps of buildings, roads, drainage networks of the most flood-prone areas that now serve as foundational tools for the city's planning and growth beyond flood resilience. The Zanzibar Mapping Initiative was the world's largest aerial mapping exercise and used consumer drones and local mappers to update the base map of Zanzibar. The data is now openly available for all purposes related to the island's conservation and development.

Each project starts by assessing goals and existing resources, engaging government, community, and other partners, and scoping the mapping work to be done. Updated overhead images of the areas of interest are obtained from consumer drones and satellites. This high resolution imagery is manually inspected and features like buildings, roads, and drainage networks are digitally mapped in a participatory manner. Fieldwork is conducted to map other features and add detailed attributes that may not be clearly visible from overhead imagery.

The collected data is used to design tools and products that support decision-making by partners and stakeholders. These digitized maps are published to OpenStreetMap and the imagery to OpenAerialMap where they serve as data public goods that can be used and improved by all. Training, community engagement, and collaboration are emphasized throughout the process to foster local networks of talent in digital cartography, robotics, software development, and data science.

This competition

Machine learning for visual tasks could improve mapping quality, speed, and cost. Recent advancements in ML for mapping include Facebook's AI-assisted road mapping tool for OSM, Microsoft's country-scale automated building footprint extraction (in USA, Canada, Tanzania and Uganda), and competitions like SpaceNet for better solutions for road and building mapping and xView2 for post-disaster building damage assessment.

These applications all feature the computer vision task of semantic segmentation: classifying every pixel in an image into categories like building, road, tree, background. Semantic segmentation is useful for mapping because its pixel-level outputs are relatively easy to visually interpret, verify, and use as-is (e.g. calculation of built-up surface area) or as inputs to downstream steps (e.g. segment building footprints first and then classify building attributes in finer detail).

With more and more OSM data labeled on high resolution drone imagery through participatory mapping in diverse urban environments across Africa, how might we use these to develop better open-source building segmentation models to keep up in our understanding of rapidly growing cities? How might we create and apply machine learning systems in the most responsible ways for disaster risk management?

The unique challenges and opportunities of this competition include:

- **Better mapping of diverse urban environments**: Machine learning models for building segmentation have mostly been trained on satellite imagery at spatial resolutions of 30 cm/pixel or lower of geographies outside of Africa. With new training data comprised of drone imagery routinely collected at much higher resolutions (3-20cm/pixel) and buildings labeled by local OSM communities across many African cities, we have the potential to develop models that can better map these diverse and densely built-up urban environments.
- **Making the most of imperfect training data for more pixel-perfect mapping**: The training data imagery represents a diverse range of geographies, spatial resolutions, sensors, and aerial surveying conditions. Their labels (OSM building footprint tracings) are inconsistent: some images have pixel-perfect

tracings of every building while others have many missing or misaligned labels. New techniques to make better use of this diverse, noisy data could unlock the potential of OpenStreetMap as ML training data for many new geographies and imagery sources.

- **Testing model robustness and generalizability to new data**: Test imagery may come from areas that are not present in the training set. Participants will need to develop models that perform best on new, unseen data. Doing so will increase the usefulness of ML for mapping on imagery created through similar processes but in new geographies and diverse conditions.
- **Integrating ML into participatory mapping and open data efforts**: Building on efforts like ML-Enabler by the Humanitarian OpenStreetMap Team, what are innovative ways to integrate high-performing open source ML solutions to enhance mapper experience and OSM data quality? On the flip side, what are novel data pre-processing and clean-up techniques that could increase the value of OSM data for geospatial machine learning?
- **Responsibly using ML to support disaster risk management and urban resilience planning**: See the Responsible AI track for more information.

Sample drone imagery and OSM building footprint labels from 10 cities

Project partners

GFDRR Labs

The Global Facility for Disaster Reduction and Recovery (GFDRR) is a partnership of the World Bank, United Nations, major donors and recipient countries under the International Strategy for Disaster Reduction (ISDR) system to support the implementation of the Hyogo Framework for Action (HFA). Launched in September 2006, GFDRR provides technical and financial assistance to help disaster-prone countries decrease their vulnerability and adapt to climate change. GFDRR works closely with UN agencies, client governments, World Bank regional offices, and other partners.

To meet the needs of a rapidly changing world, GFDRR Labs supports the use of innovative approaches to science, technology, communication and design in promoting new ideas and the development of original tools to empower decision-makers in vulnerable countries to strengthen their resilience. Recent innovations in the field have enabled better access to disaster and climate risk information and a greater capacity to create, manage, and use this information. Lab activities are designed and implemented in partnership with government institutions and key international and local partners, ensuring that all activities add value in planning, operational, and recovery activities.

In 2011, GFDRR launched the Open Data for Resilience Initiative (OpenDRI) to apply the concepts of the global open data movement to the challenges of reducing vulnerability to natural hazards and the impacts of climate change. OpenDRI supports efforts to build capacity and long-term ownership of open data projects that are tailored to meet specific needs. OpenDRI is guided by nine core principles, and engages with client governments in three main areas:

- Sharing data through open data platforms
- Collecting data through community mapping and crowdsourcing
- Using data through risk visualization and communication

Azavea

Azavea is a software company that focuses on products and professional services for turning geospatial data into actionable insights. Azavea is a B corporation that operates with a mission to advance the state of the art in geospatial technology and apply it for civic, social, and environmental impact.

DrivenData

DrivenData runs online machine learning competitions where data scientists and quantitative experts from all around the world compete to build the best algorithms for social good. The DrivenData Labs team also works on data science projects directly with mission-driven organizations.