# Open Machine Learning Challenge for Urban Resilience

## Concluding Report

Prepared for: Global Facility for Disaster Reduction and Recovery (GFDRR)

Prepared by: DrivenData and Azavea

Date shared: April 29, 2020

# Objectives

- Summarize key elements and results of the competition
- Reflect on processes, lessons learned, opportunities, and recommendations
- Provide ongoing reference for the challenge

# Contents

# Overview and Quick Links


Open Cities AI Challenge

- Permanent website: link
- Launch date: Dec 19, 2020
- Submissions close: March 16, 2020

Project Partners

- GFDRR Labs:
  Project organizer and sponsor
- Azavea:
  Geospatial data preparation
- DrivenData:
  Challenge design and hosting

# Overview

## Machine Learning for DRM

The goal of the challenge was to **accelerate the development of more accurate, representative, and usable open-source machine learning models for disaster risk management (DRM) in African cities**, starting by mapping where buildings are present.

*Comparing hand-labeled building footprints overlaid on drone imagery for 10 African urban areas included in the Challenge training dataset*

# Motivation

## Resilient Urban Planning

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 dynamic built environment.

**Buildings, roads, etc. 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).


*A field mapper from Open Cities Accra observes standing water and refuse in a flood-prone neighborhood of Accra, Ghana. Photo courtesy of Gabriel Joe Amuzu, Amuzujoe Photography.*

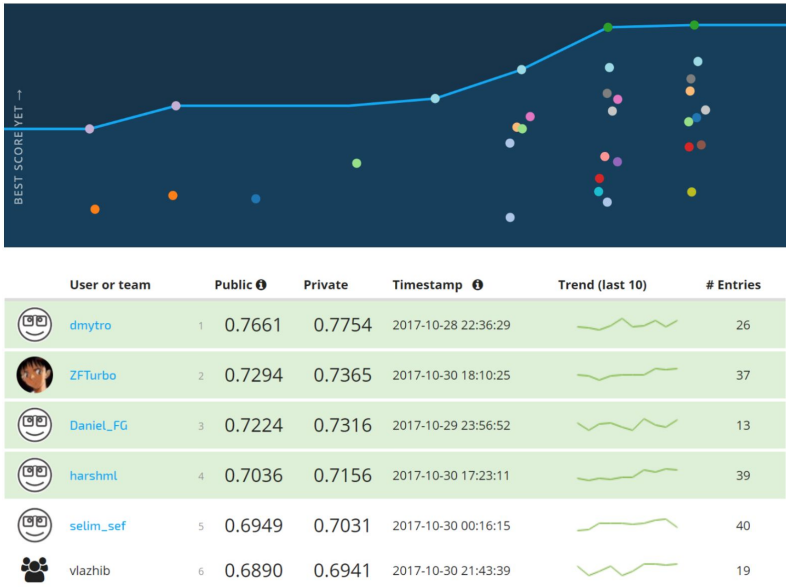
# Approach

## Why an ML Challenge?

This is a hard problem where the best approaches are not evident at the outset. ML challenges have been proven to:

- Engage a large, global data community
- Bring a wide diversity of backgrounds and skills
- Test hundreds or thousands of models quickly and cost effectively
- Elevate the best-performing solutions automatically to the top of the leaderboard

Joy's Law: *No matter who you are, most of the smartest people work for someone else.*




*Example leaderboard from DrivenData's ML challenge platform showing user-specific performance and trends over time.*

# Data

## Building on Open Mapping Efforts

The Open Cities Africa (OCA) project creates open resilience data to inform disaster risk management (DRM) and urban planning through participatory mapping.

Digitized maps are published to OpenStreetMap (OSM) and aerial imagery to OpenAerialMap (OAM) where they serve as data public goods that can be used and improved by all.

This competition featured drone imagery from 12 different cities and regions across Africa. Images were provided as 4-band GeoTiffs, with accompanying annotations indicating the pixel-wise presence of buildings.

*Example of an aerial image in Kampala, Uganda, annotated with the presence of buildings. This example was part of the competition training data.*

# Competition

## Segmenting Buildings for Disaster Resilience

### THE CHALLENGE

This competition featured two tracks:

- **Semantic Segmentation track**: Build computer vision models to identify building footprints from aerial imagery across diverse African cities and regions.
- **Responsible AI track**: Apply an ethical lens to the design and use of AI systems for DRM. Submission required for prize eligibility.

### THE DATASET

The final data provided to participants included:

- 12 African cities/regions across 11 countries
- 4 cities/regions in test set, including 2 exclusively (to encourage generalizability)
- Data of varying quality, split into Tier 1 (higher quality) and Tier 2 (variable quality)
- 878 GB of aerial imagery covering 715,974 buildings across 421 square kms

# Implications

## Unique Challenges and Opportunities

- ★ **Better mapping of diverse urban environments** (resolution, building density, etc.)
- ★ **Making the most of imperfect training data** for more pixel-perfect mapping
- ★ **Testing model robustness and generalizability** to new data
- ★ **Integrating open ML models into participatory mapping** and open data efforts
- ★ **Responsibly using ML** to support disaster risk management and urban resilience planning

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 |

*Samples from Tier 1 challenge training data.*

# Results

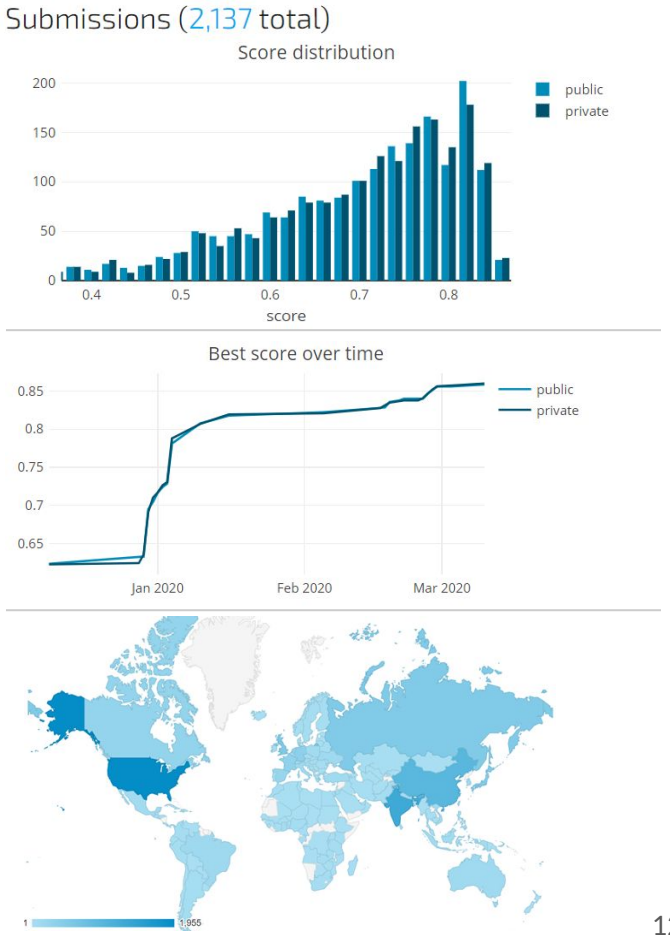
## What happened?

Extensive engagement

- **9,951 visitors** to the challenge site from 147 countries
- **1,106 participants** joined the challenge
- **2,137 submissions** generated for Segmentation evaluation and 26 final submissions entered in Responsible AI track

Boundary-pushing performance

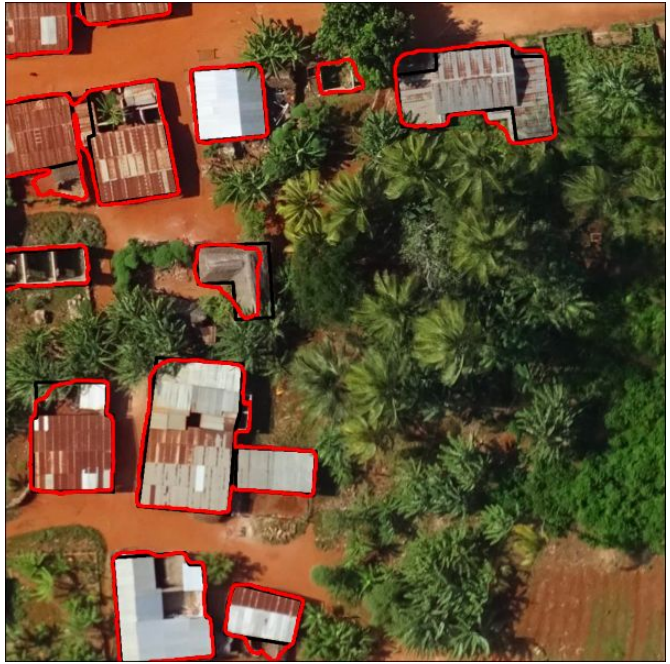
- **0.8598 Jaccard score** (intersection over union, or IoU)
- **92% precision** (proportion of predictions in ground truth)
- **93% recall** (proportion of ground truth in predictions)




*Score distribution of 2,000+ submissions (top), best-performing submission throughout the challenge with end result surpassing 85% IoU (middle), and 147 countries represented by challenge visitors (bottom).*

# Sample Outputs

*Geography: Lusaka*
*Jaccard: 0.94*



*Geography: Zanzibar*
*Jaccard: 0.88*

Predicted

Ground truth

# Meet the Competition Winners

## SEGMENTATION

**Pavel Iakubovskii**
Chebarkul, Russia
Geospatial Computer Vision Engineer
1st Place ($6,000)

**Kirill Brodt**
Almaty, Kazakhstan
Computer Graphics Researcher
2nd Place ($4,000)

**Michal Busta**
Prague, Czech Republic
Software Engineer
3rd Place ($2,000)

## RESPONSIBLE AI

**Catherine Inness**
London, UK
Data Science Master's Student at UCL
Prize 1 ($1,000)

**Chris Arderne**
Cape Town, South Africa
Data Consultant in Energy and Climate
Prize 2 ($1,000)

**Thomas Kavanagh & Alex Weston**
Brooklyn, USA
Data Scientists
Prize 3 ($1,000)

# Quick Links

**COMPETITION PAGES**

Competition Home
Includes summary description, participation and submission totals, prize pool and winner

Problem Description
Details about the available data, semantic segmentation task, metric and submission format

Responsible AI Description
Background, task, and submission specifications

About the Project
Background on DRM, Open Cities Africa, competition context and project partners

Leaderboard
Final rankings of all modeling submissions evaluated against the private test set

Data
Ongoing link to open challenge data, with STACs

# Quick Links

**RESULTS AND RESOURCES**

Repository of winning solutions
Open source code and documentation from all prize-winning solutions

Results + winners blog post
Blog post announcing competition results, winner interviews, and links to their solutions

Admin dashboard (restricted)
Summary of challenge stats for administrators

GFDRR challenge post on Towards Data Science
Medium post summarizing the competition and motivating context, mapping needs, and data

Benchmark tutorial blog post
Includes summary description, participation and submission totals, prize pool and winner

Benchmark repository
Benchmark model code using Raster Vision

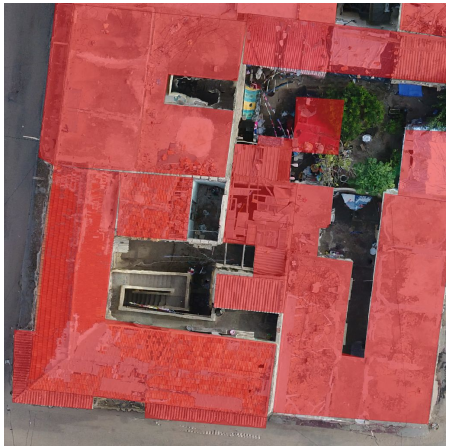
# Data and Problem Exploration

# Data Overview

- Approximately 80 images from many African cities that varied in:
  - Size
  - Resolution
  - Public accessibility status
- Private building labels
  - For a few cities
- Any labels we could pull from OSM
  - Publicly accessible
  - Drainage + building features
  - Label quality varies

# Many different competition options

- Buildings
  - Building segmentation
  - Roof material classification
  - Building material classification
  - Building quality/completeness
- Drainage
  - Drainage line segmentation
  - Drainage type/cover type classification
  - Drainage chip classification



*Building features*

*Drainage feature*

# Challenges

- Public/private data
- Overlapping bounding boxes
- Incomplete labels
- Incomplete attributes
- Attributes/features not discernible from above
- Imbalanced categorical datasets
- Tightly clustered buildings in dense urban areas

# Drainage

*Accurately labeled*

*Mislabelled or underground*

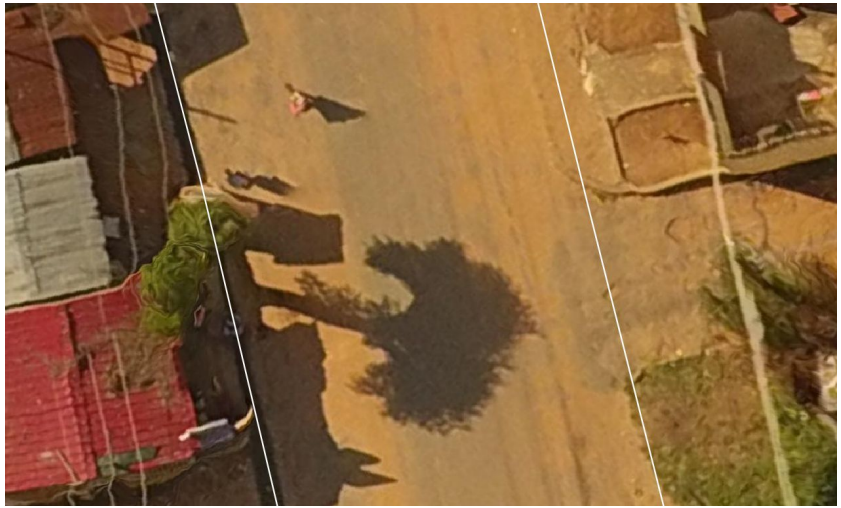
*“Ditch” rather than “drain”*

# Drainage Attribute Completeness

| | acc | dar | gao | kam | kin | mah | mon | nia | ptn | stl | znz |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | 901 | 19,323 | 0 | 347 | 0 | 0 | 220 | 208 | 48 | 99 | 9 |
| Cover type | 7.9% | 0.57% | - | 7.2% | - | - | 32.7% | 0% | 0% | 0% | 0% |
| Material | 22.4% | 9.2% | - | 48.7% | - | - | 60.9% | 0% | 45.8% | 87.9% | 0% |
| Width | 16.3% | 75.6% | - | 25.6% | - | - | 77.7% | 72.6% | 100% | 98% | 0% |

# Drainage

- Segmentation not viable
- Labels either incorrect or some features underground
- Chip classification could work but didn't add enough value



# Building Attribute Completeness

| | acc | dar | gao | kam | kin | mah | mon | nia | ptn | stl | znz |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Total bldgs | 26,428 | 682,981 | 20,527 | 4,606 | 2,668 | 10,479 | 13,470 | 128,585 | 20,010 | 40,916 | 36,742 |
| Bldg material | 24.1% | 66.7% | 0.0% | 86.3% | 1.8% | 1.73% | 66.3% | 0.0% | 42.0% | 69.8% | 3.6% |
| Roof material | 23.7% | 0.51% | 0.0% | 0.0% | 0.0% | 1.69% | 35.0% | 0.0% | 0.0% | 0.0% | 0.12% |

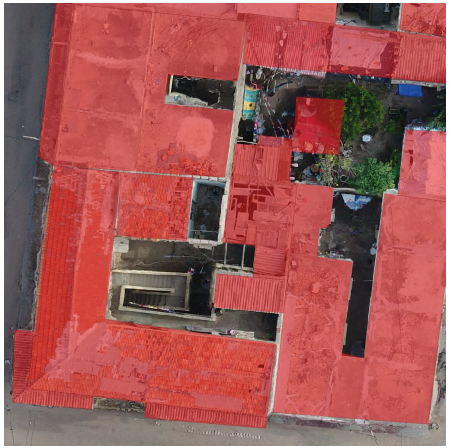
# Inconsistent Key/Value Pairs

```
"roof_material": {
    "other": [
        "{'roof:material': 'cgi'}",
        "{'roof:material': 'No roof'}",
        "{'roof:material': 'corrugated'}",
        "{'roof:material': 'No Roof'}",
        "{'roof:material': 'no roof'}",
        "{'roof:material': 'asbestos'}",
        "{'roof:material': 'metal;tile'}",
        "{'roof:material': 'slate'}",
        "{'roof:material': 'thatch'}",
        "{'roof:material': 'corrugated iron'}",
        "{'roof:material': 'wood;metal'}",
        "{'roof:material': 'canvas;metal'}",
        "{'roof:material': 'metal;wood'}",
        "{'roof:material': 'glass'}",
        "{'roof:material': 'wood;tile'}",
        "{'roof:material': 'corrugated_iron;wood'}",
```

```
"wood": [
    "{'roof:material': 'wood'}",
    "{'building:roof': 'wood'}"
],
"concrete": [
    "{'roof:material': 'concrete'}",
    "{'roof:material': 'cement'}",
    "{'building:roof': 'concrete'}",
    "{'building:roof': 'Concrete'}",
    "{'building:roof': 'ceramic'}",
    "{'building:roof': 'cement'}"
],
"metal": [
    "{'roof:material': 'metal'}",
    "{'roof:material': 'industrial_troughed_sheet'}",
    "{'roof:material': 'corrugated_iron_sheets'}",
    "{'roof:material': 'corrugated_iron'}",
    "{'roof:material': 'corrugatted_iron'}",
```

# Many different competition options

- Buildings
  - **Building segmentation**
  - Roof material classification
  - Building material classification
  - Building quality/completeness
- Drainage
  - Drainage line segmentation
  - Drainage type/cover type classification
  - Drainage chip classification

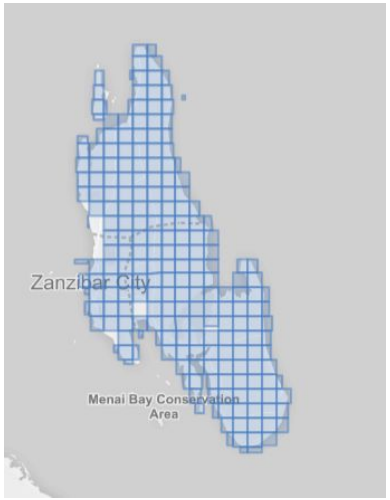


*Building features*

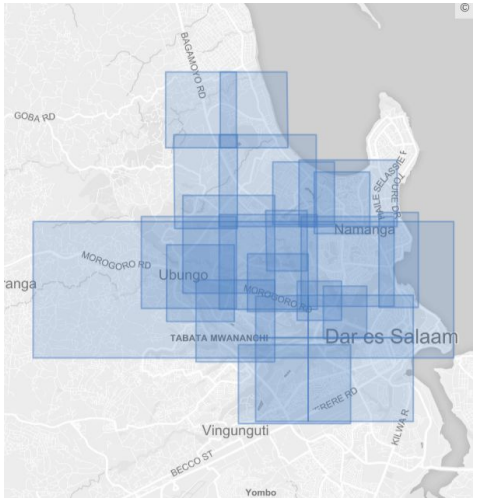
*Drainage feature*

# Imagery Extents

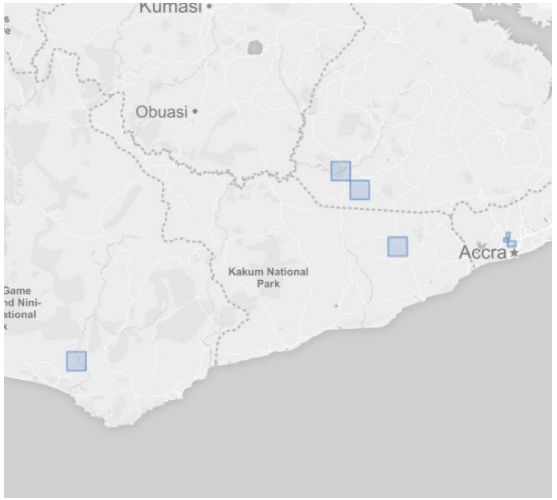
## Zanzibar




## Dar Es Salaam




## Ghana



# Label Completeness

# Label Accuracy

*Tier 1 data*

*Tier 2 data*

# Label Accuracy

- Inconsistent building labels among scenes
- Tier 1 data proved to be much more useful than Tier 2
- Design competition that could viably use only Tier 1 data
- Provide participants with Tier 2 data as an additional resource, not the primary training set
- Difficult to scale validation of building labels

*Tier 1 data*

*Tier 2 data*

# Landscape Diversity

Models must be able to generalize to buildings in various distinct landscapes

*Dense urban*

*Mixed-density urban*

*Rural*

# Train / Test Split

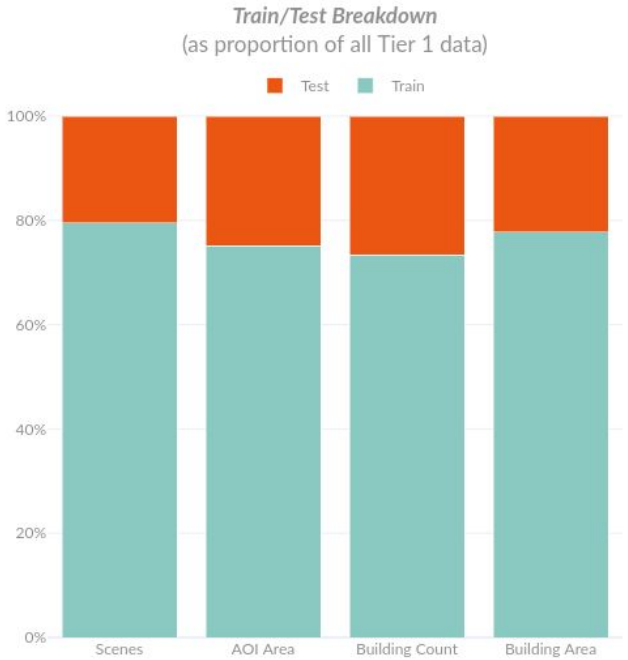
**Objective:** Create a test set that was accurately labeled and had similar characteristics to training set (most importantly Tier 1).

**Considerations:**

- Geography (avoiding over overfitting to specific specific cities/countries and enabling generalization to others)
- Landscape (including both urban and rural scenes in each dataset)
- Building density (some rural scenes had almost no buildings)
- Quantity (training/test sets with appropriate quantities of scenes, buildings, and pixels)

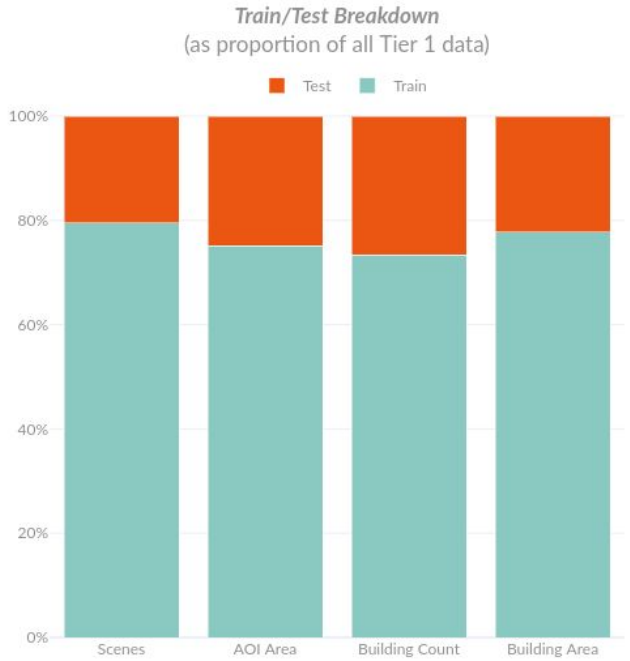
# Train / Test Split

- Data organized into scenes (aerial images covering a contiguous area)
- Roughly 80/20 % split among tier 1 scenes
  - 31 Tier 1 training scenes
  - 8 Tier 1 test scenes



# Train / Test Split

- Test images split among four cities
    - Scenes from two of four cities/regions were also present in the training set (Zanzibar, Niamey) while two were not (Saint Louis, Lusaka)
- AOI Area, Building Count and Total building area were roughly on par with the 80/20 split in scenes



# Anonymized Test Set

Test images were anonymized to prevent contestants from downloading building labels from OSM

- 1024 x 1024 “chips,” small enough to not give away their specific location
- Number of chips in a scene varied according to its size and shape
    - Chips per scene ranged from 794 (Saint Louis) to 3,104 (Zanzibar)

*Example test chip*

# Anonymized Test Set

Test images were anonymized to prevent contestants from downloading building labels from OSM

- Geospatial information was obscured by resetting coordinates of all chips to (0.000, 0.000)
- Created (0,1) raster masks of building polygon labels
- Broke test set into 50/50 stratified samples (based on the same criteria as train/test) for public and private leaderboards.

*Example test chip*

# Data Processing Obstacles

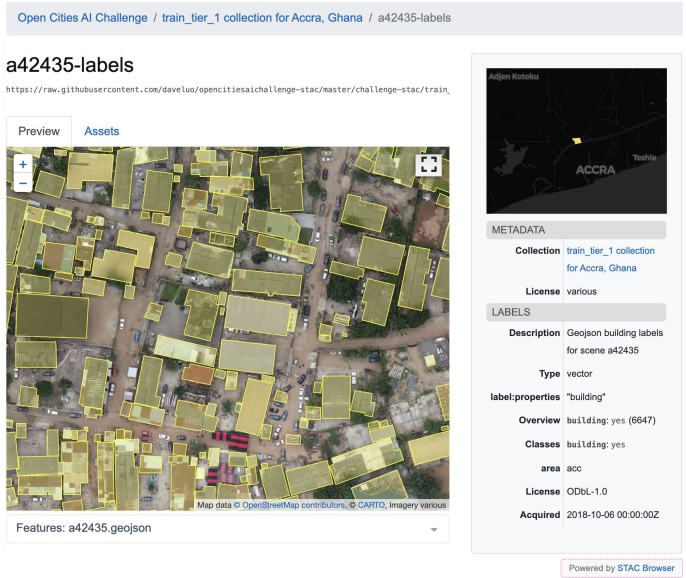
- Inconsistent ground-surface resolutions across scenes from different regions
- Large imagery and geojson datasets
    - Tried to reduce data size by downsampling images (where possible), simplifying polygons and compressing TIFFs
- Cloud-optimizing entire dataset
- Inconsistent encoding of NoData across all images
- Cropping labels to AOI boundaries

*AOI polygons were not available for most scenes, so cropping building labels required vectorization of non-NoData portions of scenes.*

# STAC

*Spatio-Temporal Asset Catalog*

- Need for a consistent system for encoding metadata
- Include spatial (bounding box, geojson boundary) and temporal (capture date/time) of images
- System for exposing metadata on train/test imagery and training labels to contestants
- Tools built on top of it (i.e. STAC Browser, PySTAC)
- Opportunity to expose more people to spec




*STAC Browser*

# Semantic Segmentation Track

# Overview

- Segment building footprints from aerial imagery, classifying the presence or absence of buildings on a pixel-by-pixel basis
- 3 prizes awarded ($6,000, $4000, $2,000 for 1st, 2nd, and 3rd, respectively)
- Evaluation based on objective statistical metric (Jaccard)
- 2,137 total submissions made across 1,106 participants
- Top score on private test set: 0.8598

*Sample building footprints in two locations provided in the challenge dataset.*

# Evaluation

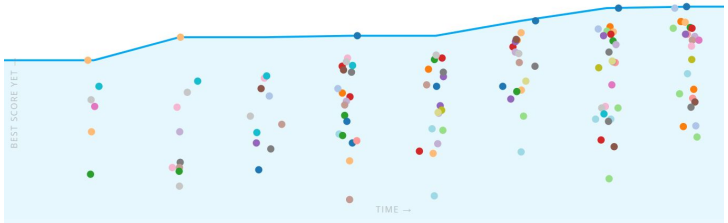
- Submissions made in the form of `.tar` or `.zip` file with a single-band 1024 x 1024 TIFF mask for each chip in the test set
- Limit to 3 submissions per day allowed for each team
- Evaluated using similarity measure called Jaccard index (i.e. intersection over union):

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

*Example test chip image (left) and submission chip file (right). The submission chip reflects a single-band with predicted building pixels (white) and non-building pixels (black).*

# Winning Solutions

- 1st place: Pavel Iakubovskii (@qubvel)
  Jaccard score: 0.8598
  open-cities-ai-challenge/1st Place
- 2nd place: Kirill Brodt (@kbrodt)
  Jaccard score: 0.8575
  open-cities-ai-challenge/2nd Place
- 3rd place: Michal Busta (@MichalBusta)
  Jaccard score: 0.8401
  open-cities-ai-challenge/3rd Place



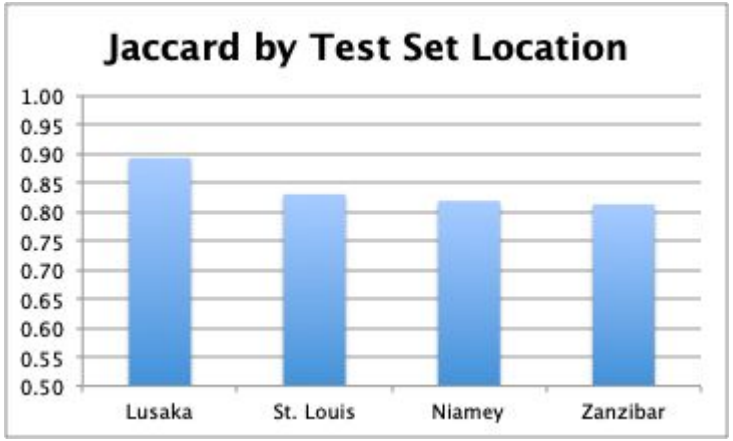

| User or team | | Best private score | Timestamp | Trend (last 10) | # Entries |
|---|---|---|---|---|---|
| qubvel | 1 | 0.8598 | 2020-03-10 14:35:28 | | 53 |
| kbrodt | 2 | 0.8575 | 2020-03-16 23:27:18 | | 39 |
| MichalBusta | 3 | 0.8401 | 2020-03-15 23:57:03 | | 110 |
| akldf | 4 | 0.8393 | 2020-03-16 10:33:16 | | 64 |
| bonlime | 5 | 0.8322 | 2020-03-16 15:45:09 | | 39 |

*Challenge leaderboard displaying final scores achieved by top participants. The top graph provides an interactive sample of scores achieved throughout the competition (higher is better).*

# Results

- Top solution: 0.8598 Jaccard score
- Balanced precision and recall
  - 92.1% precision (true positive / predictions)
  - 92.8% recall (true positive / ground truth)
- Jaccard score by test set location (see graph)
  - Lusaka: 0.892
  - St. Louis: 0.830
  - Niamey: 0.819
  - Zanzibar: 0.813

Locations not included in training set, suggesting greater generalizability to new sites




*Jaccard scores across the 4 locations in the test set. Lusaka and St. Louis were only represented in the test set, while data from Niamey and Zanzibar was also included in the training set.*

# Sample outputs

*Geography: Lusaka*
*Jaccard: 0.93*

*Geography: Zanzibar*
*Jaccard: 0.88*

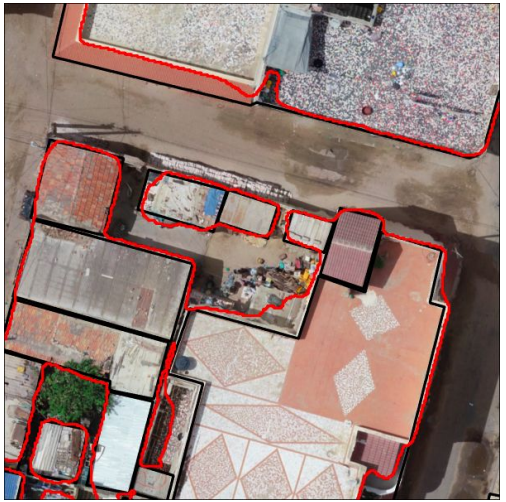
*Geography: Niamey*
*Jaccard: 0.88*

Predicted  Ground truth

# Sample outputs

*Geography: Zanzibar*
*Jaccard: 0.90*

*Geography: Niamey*
*Jaccard: 0.89*



*Geography: St. Louis*
*Jaccard: 0.88*

Predicted  Ground truth

# Sample outputs

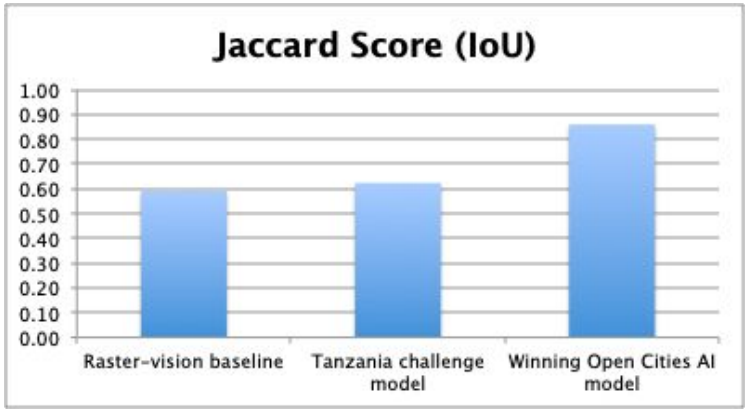
*Geography: St. Louis*
*Jaccard: 0.52*

*Geography: Niamey*
*Jaccard: 0.69*

*Geography: Zanzibar*
*Jaccard: 0.78*

Predicted  Ground truth

# Comparison with baseline models




*Jaccard scores measured on the challenge test set for three models. The two left models represent what was possible with little effort before the challenge; the right is the winning model.*

Two solutions were also available to gauge what relatively low-effort models would have produced before the challenge

- **Tanzania challenge model**: Based on a top model in the Open AI Tanzania Challenge, fine-tuned on a subset of Tier 1 training data
  **Jaccard score = 0.6235**
- **Raster-vision baseline**: Benchmark created by Azavea team as point-of-departure tutorial for the Open Cities AI Challenge
  **Jaccard score = 0.5915**

**Note**: these solutions were not developed to be competitive in this challenge, and were not trained on the full dataset. They should only be considered as representing low-effort alternatives available before the challenge began.

# Lessons Learned

**What worked well**

- Submission acceptance and scoring were smooth, enabled 2,000+ models to be evaluated
- Sharing data using STACs, visual examples, pre-trained models and other open tools
- Tier 1 and Tier 2 training data provided good balance of scale and reliability
- Connecting communities in pipeline through open data, annotation, and modeling

**Opportunities**

- Keeping sufficient test data private during workflow for fair evaluation (affected timing)
- Ground truth on building delineation
  - Pixel-based metric chosen because delineation was not sufficiently reliable
  - Results in over-weighting of bigger, more well-defined structures and people who live in them
  - Opportunity to reflect mapper confidence in delineation during collection

The winning models achieved high levels of performance building on the diversity of training and test data.

These models can be used on datasets outside of the competition as long as the user understands the limitations and has building footprints to validate accuracy.

# Responsible AI Track

*With great computational power comes great responsibility.*

*As ML experts who directly develop and apply algorithmic systems,* ***this Responsible AI track presents an opportunity to examine the practical ethics and appropriate use of our work applied to the field of disaster risk management.***

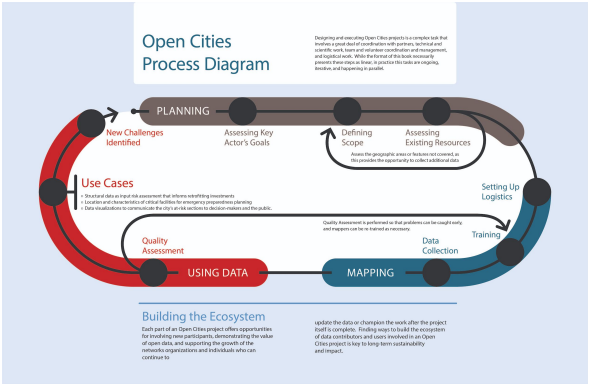
# Overview

- Apply an ethical lens to the design and use of AI systems for DRM
- 3 prizes awarded ($1,000 each)
- Open to all (no need to also submit to Segmentation track), and submission was *required* to be eligible for any prize
- Flexible format (Jupyter notebooks, slides, blogs, essays, demos, product mockups, speculative fiction, art work, synthesis of research papers or original research, etc.)
- Evaluated by panel of judges based on a pre-released rubric
- Novel—this is the first time we've seen an ethics track included in an ML challenge

# Prompt: Scenario Description

Bring an applied ethics lens to the design and use of **ML pipelines for improved mapping (i.e. via semantic segmentation) in DRM applications.**

- Data collection and annotation: OSM, OAM, OpenDRI's 9 principles for DRM and open data projects
- Data curation and management: processing and preparing drone imagery and building footprints for the challenge
- Model development and evaluation: semantic segmentation pixel-based classification task and Jaccard metric
- Applications: downstream uses such as new construction, aid delivery, retrofits and inspections, etc.




*Open Cities process diagram for planning, mapping, and using data, shared with participants. Source: opencitiesproject.org/about/*

# Prompt: Creating A Submission

**How might we improve the creation and application of ML solutions** to mitigate biases, promote fair and ethical use, communicate insights clearly, and make safeguards to protect users and end-beneficiaries?

Submissions could focus on any (or all) of the following areas:

- Framework: What approach or principles would you use to examine the ethical considerations of using ML in this scenario?
- Identification: What are the potential harms at play in this scenario?
- Mitigation: How can these ethical issues be mitigated? What technical approaches or tools would you use?


*Case studies aggregated by GFDRR were shared for additional reference and inspiration. Source: Machine Learning for Disaster Risk Management*

# Judging Rubric

Submissions were evaluated by a panel of judges based on the following rubric:

- Thoughtfulness (40%): Depth of inquiry, synthesis of ideas, managing trade-offs
- Relevance (20%): Ethical lense applied to DRM, consider challenge data sources
- Innovation (20%): Novel approaches, takeaways are insightful, thought-provoking, and actionable
- Clarity (20%): Communicated clearly, understandable to the non-technical layperson

| | | | Scores (1-10) - for each Judge to complete | | | | Calculated |
|---|---|---|---|---|---|---|---|
| Finalist | | Link to submission | Thoughfulness | Relevance | Innovation | Clarity | Score |
| A | How to improve the use of AI for building segmentation | /13nnh7jhFqaLcuYQdK6Ax2 | | | | | 0.0 |
| B | Open Cities AI Challenge: Segmenting Buildings for DisasterResilience | /1OzAUOxQsXvB_j5zkKLn6 | | | | | 0.0 |
| C | Fairness in Machine Learning - How a Model Trained on Aerial Imagery Can Contain Bias | n?id=17AiHergAfk4uLho3FY- | | | | | 0.0 |
| D | Responsible AI for Disaster Risk Management | /1Nz3m00PDR_59XmZFz86s | | | | | 0.0 |
| E | Evidence-based ethics for AI in Disaster Resilience | ument/d/1LI8yb7ryRAqUabqi | | | | | 0.0 |
| F | Ethical analysis of applying ML model to disaster risk management | n?id=1cWZpq1k9Ap- | | | | | 0.0 |
| G | Ethical Machine Learning for Disaster Risk Management | ument/d/1FgjaYhooWeN2cp | | | | | 0.0 |

*Screenshot of judging spreadsheet provided to each judge to score finalist submissions according to pre-released rubric*

# Judging Process

- Step 1: Narrow submissions down to 10 finalists (Dave @ GFDRR)
- Step 2: Score finalists according to four judging criteria to produce individual rankings (judging panel: Caroline Gevaert, Dennis Wagenaar, Nuala Cowan)
- Step 3: Aggregate rankings using rank choice voting to determine the winners!

| | | | *Judge rankings* | | | | |
|---|---|---|---|---|---|---|---|
| **Finalist** | | **Link to submission** | **Judge A** | **Judge B** | **Judge C** | **Combined** | **Final rankings** |
| C | Fairness in Machine Learning - How a Model Trained on Aerial Imagery Can Contain Bias | n?id=17AiHergAfk4uLho3FY- | 1 | 3 | 2 | 6 | 1 |
| I | Stop pretending technology is value neutral | https://rdrn.me/ethics-ai/ | 2 | 2 | 3 | 7 | 2 |
| J | Contributed Geographic Information: Gray Zones in Collection and Usage | ument/d/11eHxEc6BZwYbgx1 | 3 | 5 | 1 | 9 | 3 |
| D | Responsible AI for Disaster Risk Management | /1Nz3m00PDR_59XmZFz86s | 5 | 1 | 4 | 10 | 4 |
| E | Evidence-based ethics for AI in Disaster Resilience | ument/d/1LI8yb7ryRAqUabqi | 4 | 3 | 9 | 16 | 5 |
| G | Ethical Machine Learning for Disaster Risk Management | ument/d/1FgjaYhooWeN2cp | 6 | 8 | 5 | 19 | 6 |
| F | Ethical analysis of applying ML model to disaster risk management | n?id=1cWZpq1k9Ap- | 7 | 6 | 6 | 19 | 7 |
| B | Open Cities AI Challenge: Segmenting Buildings for DisasterResilience | /1OzAUOxQsXvB__j5zkKLn6 | 8 | 6 | 6 | 20 | 8 |
| H | AI Ethics Self-Assessment | /1WNVIOOgejMBulUp0pEqv | 9 | 8 | 8 | 25 | 9 |
| A | How to improve the use of AI for building segmentation | /13nnh7jhFqaLcuYQdK6Ax2 | 10 | 10 | 10 | 30 | 10 |

*Judging spreadsheet displaying final aggregation of judge-provided rankings to determine winners*

# Winning Submissions

- Prize 1: Catherine Inness (@Catherine_I)

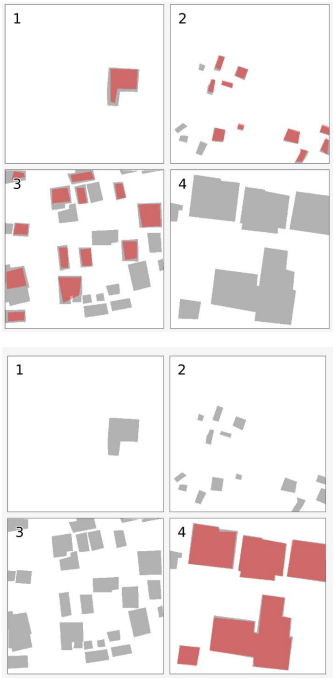
  *Fairness in Machine Learning: How Can a Model Trained on Aerial Imagery Contain Bias?*

- Prize 2: Chris Arderne (@chrisjames)

  *Stop pretending technology is value neutral*

- Prize 3: Thomas Kavanagh (@thomkav) and Alex Weston (@alweston)

  *Contributed Geographic Information: Gray Zones in Collection and Usage*




*Screenshots from 2nd winning submission, illustrating bias considerations in pixel-level eval metric (building size matters; top IoU = 22%, bottom IoU = 47%)*

# What's Next

- Invited Responsible AI winners to AI for DRM working group
- Update and make winning submissions public
- Announce winners to 40K+ challenge community
- Blog posts
    - Looking to consolidate into two posts for sharing (est. May/June)
    - Responsible AI themes through GFDRR
    - Technical results and open solutions on Medium

# Lessons Learned

**What worked well**

- Useful to make the track mandatory
- Concrete prompts and context
- “Office hours” provided avenue to submit early and expand thinking (may do  sooner or more)
- Process flow from submission acceptance to judging and awards
- Showed that tracks like this can work and members of the community want to engage

**Opportunities**

- Encourage more code-based or interactive submissions
  - Narrow format that people chose to submit (expository writing)
  - Could reflect in judging criteria
  - More concrete examples that resemble range of formats
- Promote teaming and connection-building
- Additional outreach to ethical ML circles

Open data facilitates open discussions about data ethics and privacy (as opposed to private data which can be collected and analyzed without local stakeholders awareness).

These conversations benefit from the perspective of data scientists, and this thinking is not optional. There should be an active feedback loop between ethical discussion and the way ML in DRM is carried out.

# Parting Thoughts

# Highlights

- Open data → open ML models for DRM
- Data diversity in locations and sensors
- 1,100 participants generating 2,100 submissions, resulting in 0.86 IoU of winning solution
- Global engagement: challenge visitors from 147 countries, winners from 4 continents
- Integrating ethical thinking into ML challenges

# Sample Applications

The challenge datasets had enough diversity in locations and sensors to make the winning classifier useful for a range of urban mapping projects in Africa.

- Risk assessment and mitigation (buildings in flood zones, fire risk zones, steep slopes, etc.)
- Monitoring change in building coverage (growth, disaster damage, etc.)
- Integrating with building attributes to prioritize retrofitting or aid (e.g. roof material, “soft story” buildings, informal settlements, etc.)
- Cadastre and land rights (urban built-up coverage, building counts, etc.)

# Recommendations

- Focus on **scene completeness** to ensure positive and negative coverage (e.g., drainage)
- More consistent **delineation of buildings** (especially relevant in dense urban environments and informal settlements)
- **Withhold label data** for competitions before ultimately contributing them to OSM
- Splitting data into **quality "tiers"** can help balance scale and reliability in ML
- More **concrete prompts for building ethical thinking into data projects** at relevant points across the AI pipeline (e.g., collection, annotation, modeling, application)
- Track **performance benchmarks**, e.g., open "model zoo" and human comparisons, and ideally how they impact applications (manage what you measure)

# Future Opportunities

- Consistent and exhaustive labeling of **additional building characteristics** to facilitate computer vision models for completion level, materials, or quality
- Additional exhaustive labeling of **urban landscape characteristics**, e.g., drainage
- Continue to gather drone imagery and building labels from **increasingly diverse set of landscapes**
- **Incorporating bias considerations/disclosures** into machine learning projects
- **Tools for ML and open mapping communities working together** (e.g., prioritizing uncertain or changing areas for annotation, scenes for quality checking, etc.)

# Thank You!

This challenge was made possible by the dedication of this amazing group.

- **GFDRR/World Bank**: Dave Luo, Vivien Deparday, Nick Jones, Caroline Gevaert, Cristiano Giovando, Axel Blanc, Grace Doherty, Nuala Cowan, Robert Soden
- **DrivenData team and advisors**: Emily Miller, Greg Lipstein, Peter Bull, Robert Gibboni, Isaac Slavitt, Joseph Muhlhausen (WeRobotics)
- **Azavea team**: Simon Kassel, Rob Emanuele, Esther Needham, Ross Bernet
- **Judges**: Caroline Gevaert, Dennis Wagenaar, Nuala Cowan
- **Open mapping communities**: OSM, OAM, OpenDRI, Open Cities Africa
- **Data science communities**: Special thanks to everyone who participated!