Analytics

Pro Bono Data Consulting for the Social Sector

- Introduction to Delta Analytics
- Technical Deep dive - Rainforest Connection

Sara Hooker
Executive Director
Delta Analytics

sara@deltanalytics.org
twitter: sarahookr

Sean McPherson
Delta Data Science
Fellow

sean@deltanalytics.org
twitter: PhaseAnalytics

Delta Analytics

Delta Analytics fills the technical skill gap and enables nonprofits to **accelerate their impact**.

Acute need for skills based volunteering.

- Data revolution
- Complex problems are now within reach
- Skills gap is larger than ever

Nonprofits are pushing the hardest for data

- focus on accountability
- desire to understand impact

Organizations bridging the skill gap. And most importantly… you.

Eric & Wendy Schmidt
Data Science For Social Good
Summer Fellowship
The University of Chicago

kaggle

BAYES IMPACT

DRIVENDATA

DataKind SF Bay Area

Big Data for Social Good Challenge
Powered by IBM + Hadoop

Our Impact

27 projects with nonprofits and social impact organizations

90+ Fellows volunteering part-time over 3 years

Over 15,000 hours donated

$0.00 charged for services

17 US and 10 International projects (Tanzania, UK, Kenya, and more)

Some of our grant recipients:

Community Engagement

- United Way — United Way of the Bay Area
- Tenants Together
- fsd foundation for sustainable development
- Ashby Village
- SJC social justice collaborative

Education

- Roots of Success
- eneza education — spreading education everywhere
- KUOW.ORG
- City Year

Economic Development

- Village Enterprise
- TechnoServe — Business Solutions to Poverty
- BUILD — In Business to Learn
- sama

Environmental

- LivelyHoods
- BioLite
- Karibu
- BGCI — Plants for the Planet
- Rainforest Connection

Where do Delta Fellows work?

facebook

STITCH FIX

AON

udemy

NETFLIX

twitter

Google

CORNERSTONE RESEARCH

3Q DIGITAL

Brattle

THE POWER OF ECONOMICS

... and many more

Delta Data Fellows

Partner with a non-profit grant recipient for 6 months.

&

Delta Teaching Fellows

Democratize access to machine learning education by building technical capacity around the world.

- full time professionals
- donate time for 6 month projects
- cohort of 30 fellows, January-July

Delta teaching fellows believe that data is powerful, and that anybody should be able to harness it for change in their community.

moringa SCHOOL

BRAVE academy

Andela

2017 Grant Recipients

RAINFOREST CONNECTION

sama SOURCE

CITY YEAR

BGCI
Plants for the Planet

TECHNOSERVE
BUSINESS SOLUTIONS TO POVERTY

BUILD
www.build.org

SJC social justice collaborative

Deep dive into our work with Rainforest Connection

Delta Team

Sara Hooker
Executive Director
Delta Analytics

Sean McPherson
Data Analyst
Northrop Grumman

Steven Troxler
Data Scientist
Stitch Fix

Cassandra Jacobs
Data Scientist
Stitch Fix

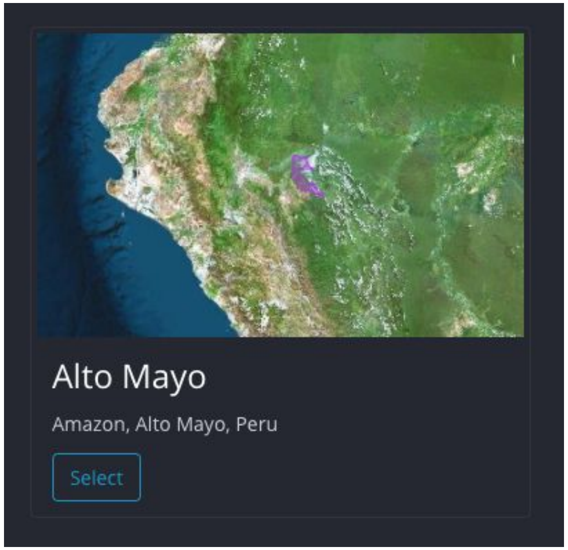
Rainforest Team

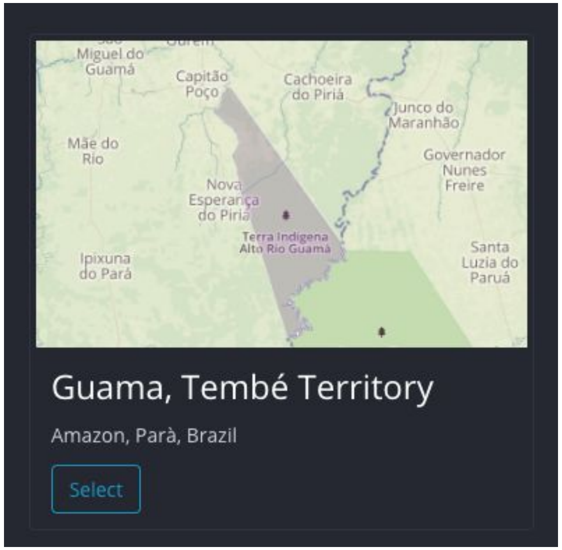
Close partnership with Rainforest Eng team
Stefan Zapf and Christopher Kaushaar

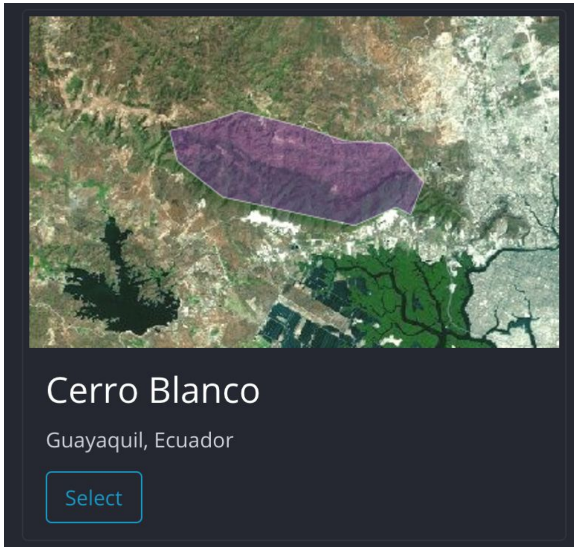
Rainforest Connection collects and deploys recycled cells to rainforests around the world. The cellphones stream audio **24/7**, and advanced deep learning models detect when the sound of a chainsaw occurs.

Within minutes, an alert is sent to conservationists.

Streaming audio data from rainforests in Ecuador, Peru and Brazil.







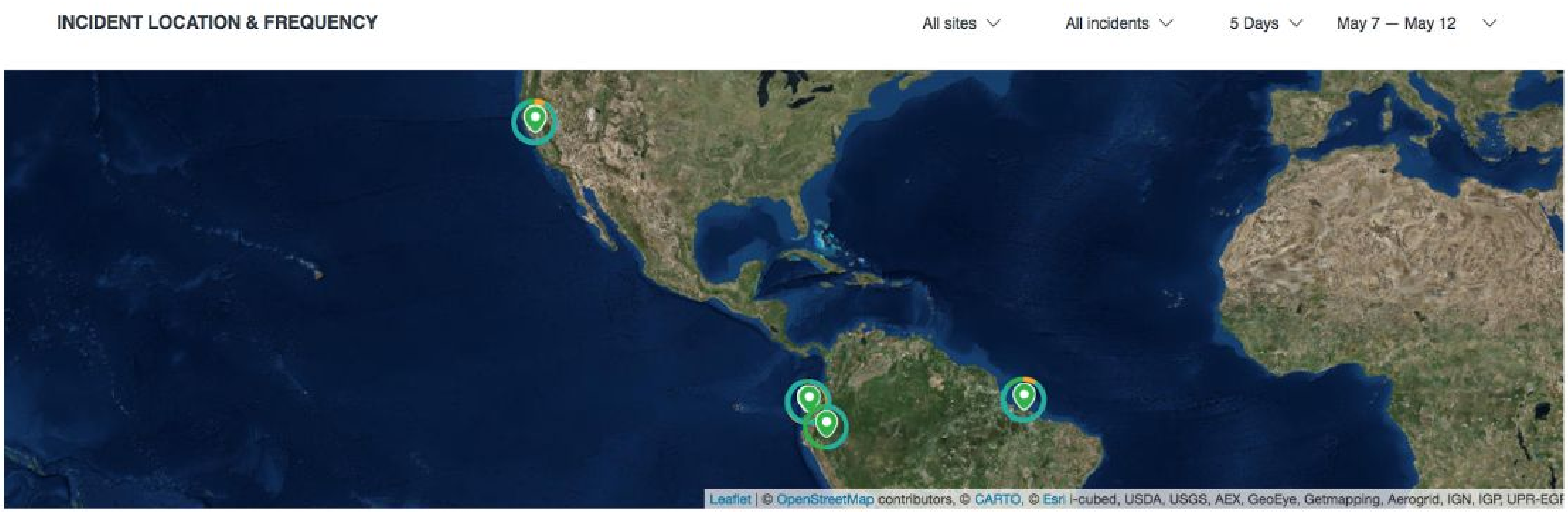
Additional deployments are made available as conservationists and research partners request them.

Guardians require little maintenance

Recycled cell phones are powered by solar panels, they are deployed by Rainforest Connection high in the rainforest canopies to avoid detection.

- protected from rain and weather elements by a small waterproof case
- rely on local cell phone networks to send data

Conservation partners are immediately alerted when a chainsaw is detected, and can monitor aggregate behavior in their region.



3 key goals:

1. Chainsaw detection model:
    a. Accurately detect chainsaws
    b. Data augmentation
    c. Enhance data quality

2. Provide engineering capacity
    a. Code improvement
    b. More efficient data processing

3. Model the direction and distance of chainsaw sound.

1. Accurately detect chainsaws

How is the training data set collected?

Collection from local experiment sites.

- **tree cuttings**
- **planned chainsaw data collection days**

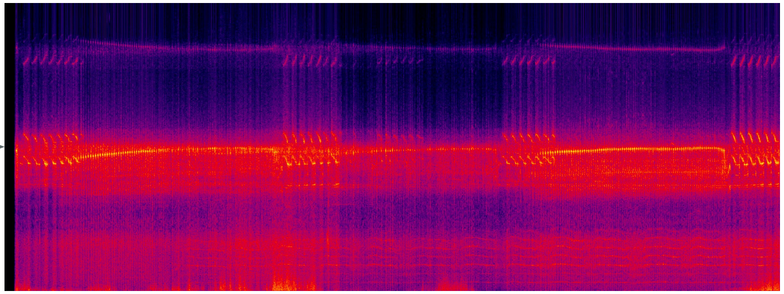
date	chainsaw
4/18/2017	31
4/19/2017	45
4/20/2017	73
4/21/2017	40
4/22/2017	19
4/23/2017	3
4/24/2017	21
4/25/2017	25
4/26/2017	82
4/27/2017	109
4/28/2017	32
4/29/2017	52
4/30/2017	31
5/1/2017	13
5/2/2017	53

Fold in positive test audio from our rainforest sites to gain additional examples (pseudo labelling).

- **confirm whether model labelling is correct using human validation.**

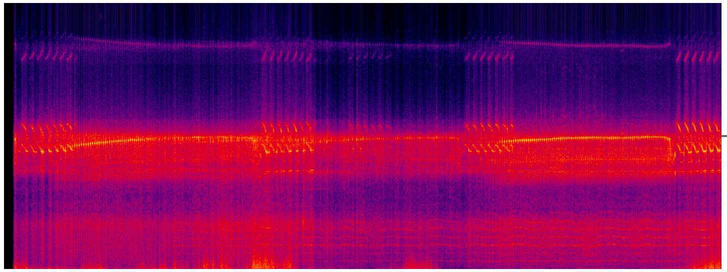
approach: turn audio detection into an image classification problem. apply a deep learning architecture.

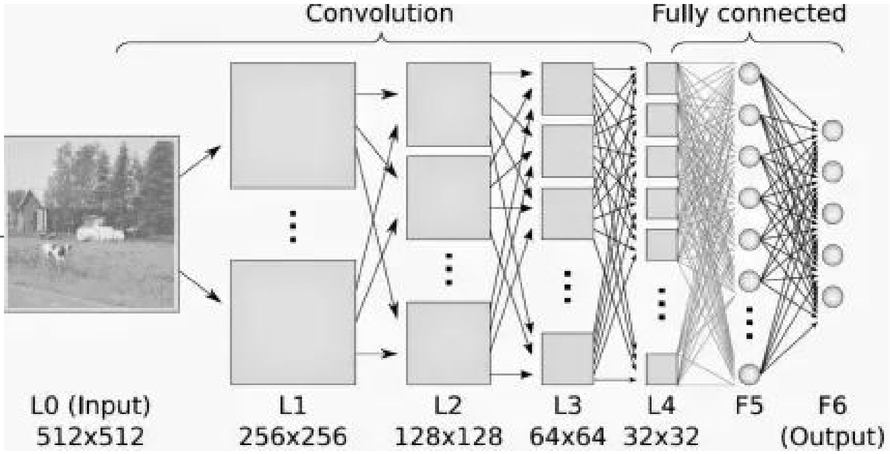
Audio streamed from conservation sites in Ecuador, Peru and Brazil.



Convert audio to a spectrogram (visual way to represent the signal strength of a sound).

Convolutional neural network architecture for image classification



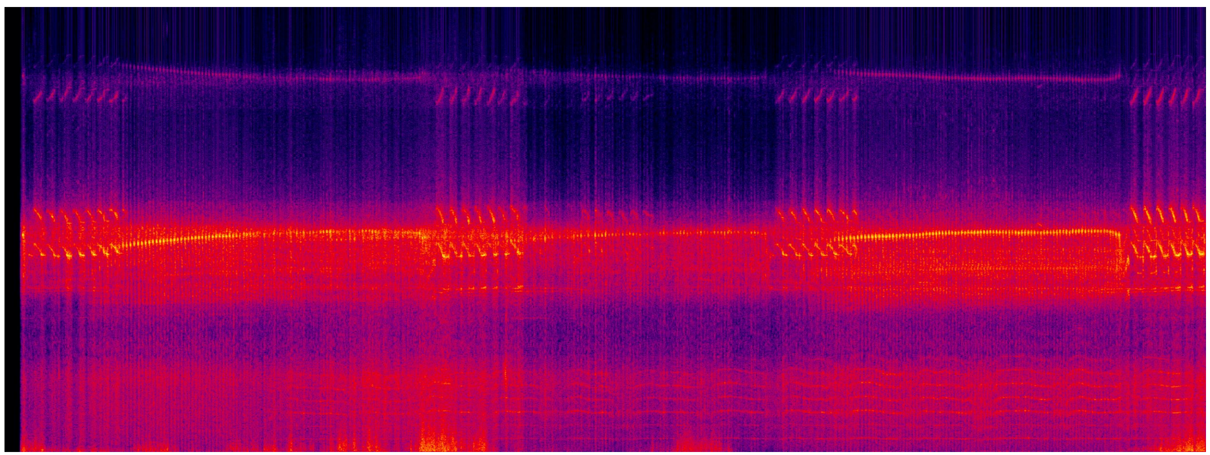


Balanced 50/50 labelled spectrogram is our training set.

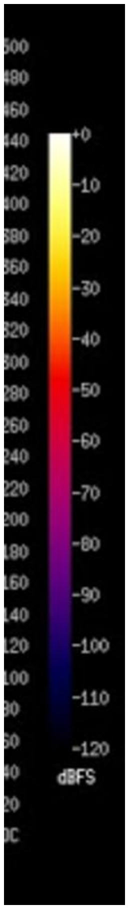
Convolutional neural network is a special type of deep learning architecture used to detect objects in images.

Spectrogram is a graphic visualization of frequency and intensity of a sound with respect to time.

frequency (Hz)

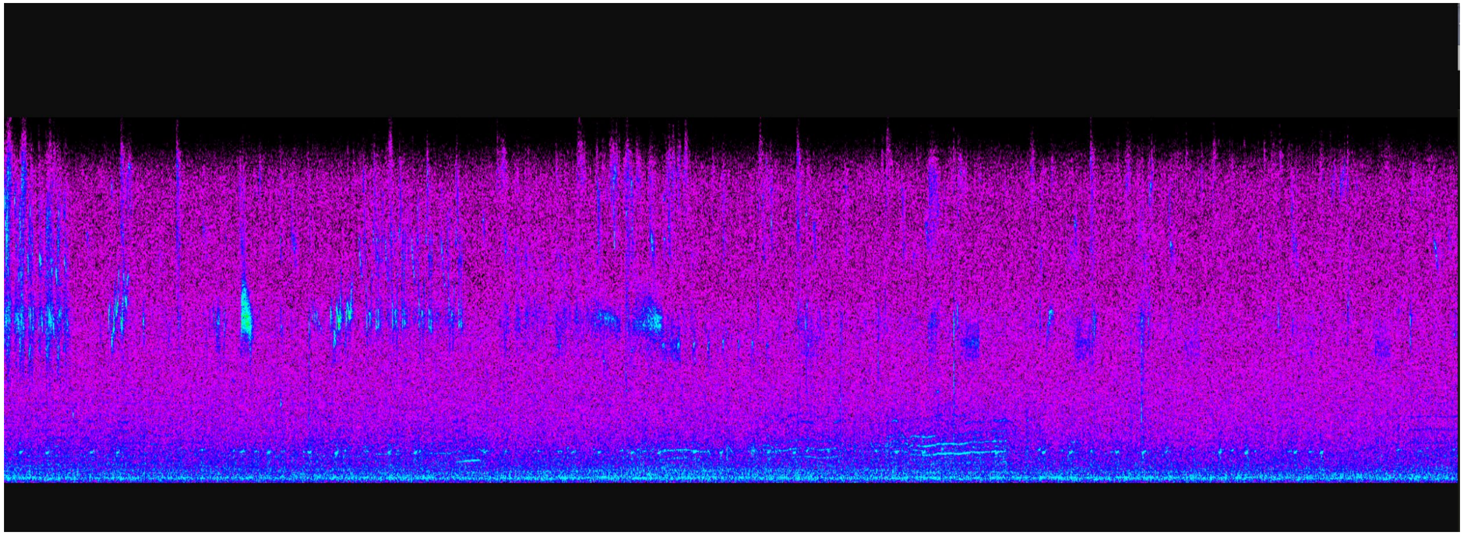


time(s)

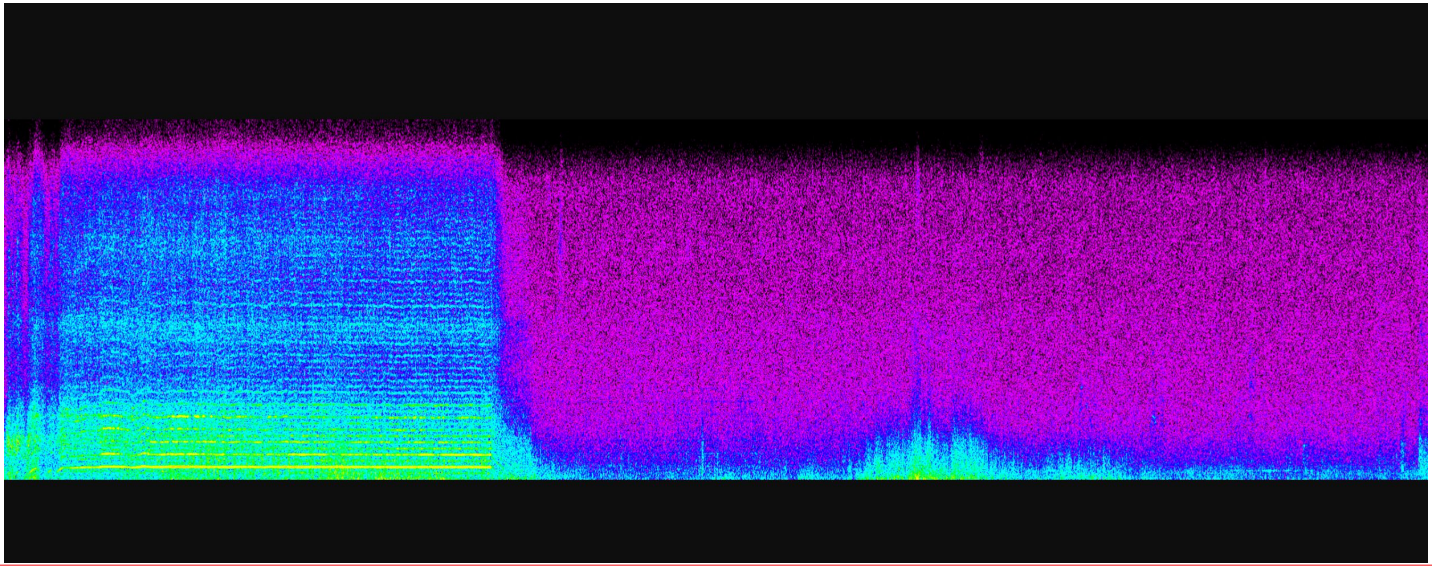


amplitude is captured by the intensity of color

weak audio example here

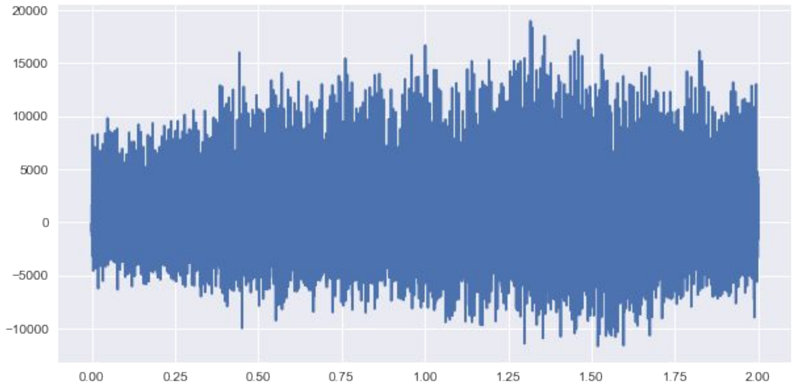


strong audio example here

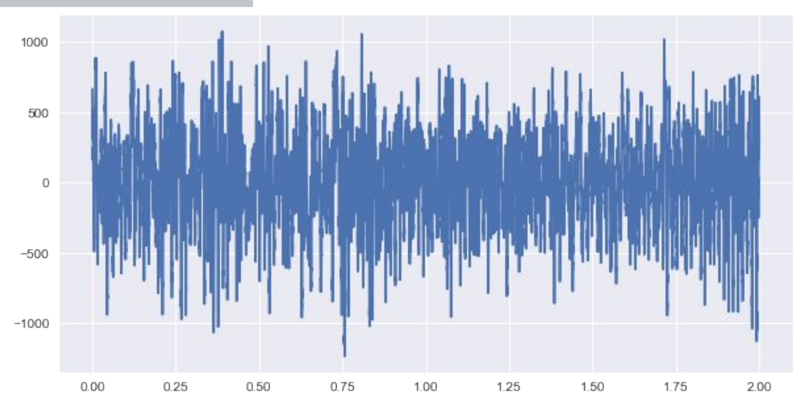


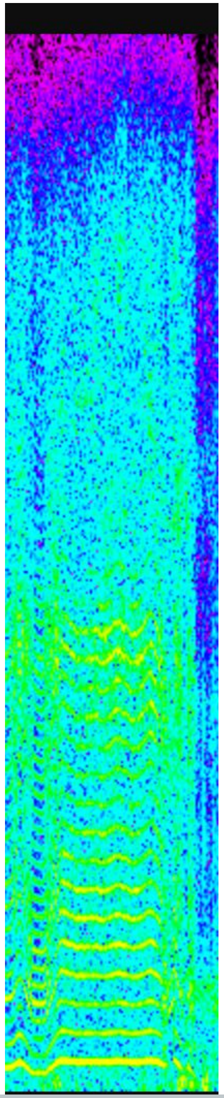
Why Use spectrograms?

Chainsaw Audio

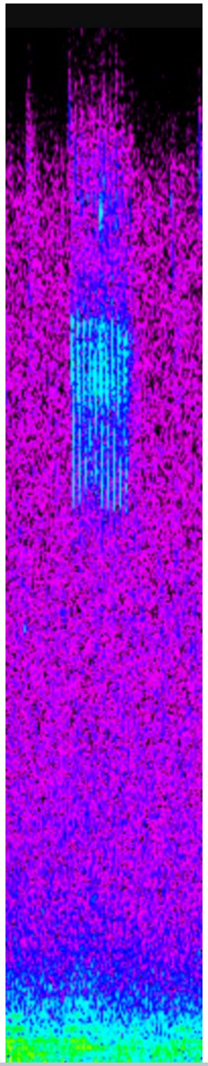


Birdsong Audio



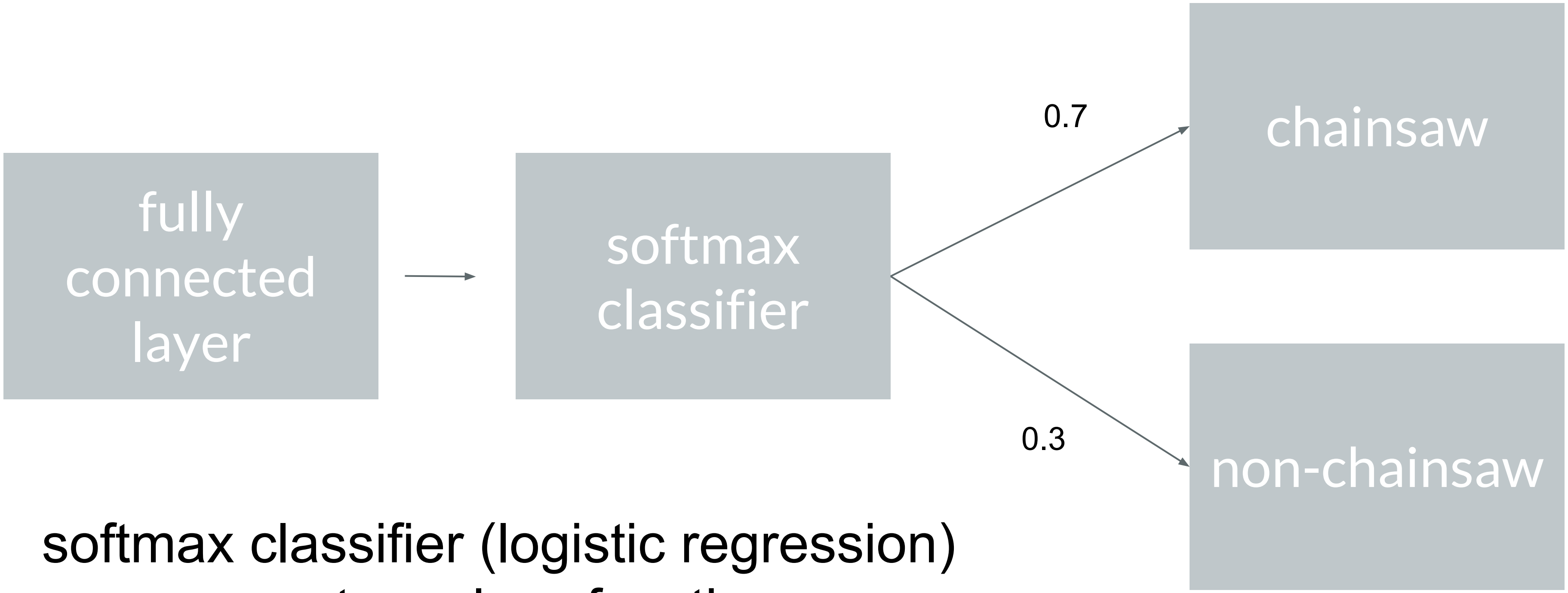


Chainsaw Spectrogram



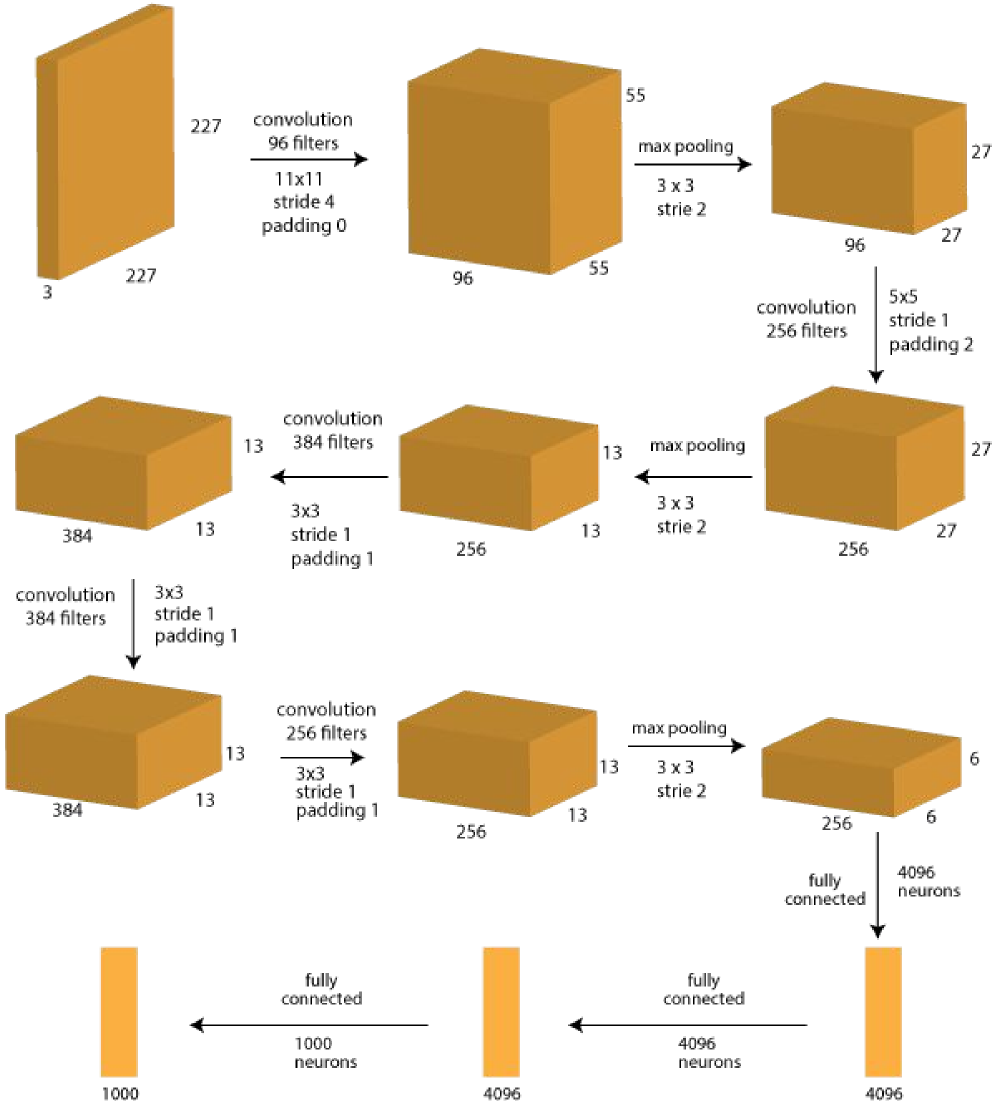
Birdsong Spectrogram

Binary classification problem: model outputs probability that a spectrogram is a chainsaw.



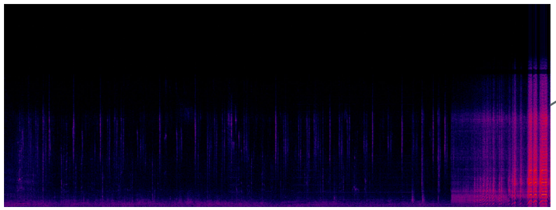
- softmax classifier (logistic regression)
  - cross-entropy loss function
- minimize loss function through backpropagation

Architecture very simple (variant of AlexNet)



- added batch-normalization
- RELU activation
- experimented with depth (number of hidden layers)
- adam is used as optimizer.
- black and white spectrogram (no amplitude dimension)

Threshold for classification is set at 0.9. False positives pose high cost.



Positive example **Negative example**

```
[[ 0.09271322  0.90728676]
 [ 0.12525368  0.87474632]
 [ 0.12527105  0.87472898]
 ...,
 [ 0.10612514  0.89387488]
 [ 0.1254736   0.87452638]
 [ 0.12120702  0.878793  ]]
```

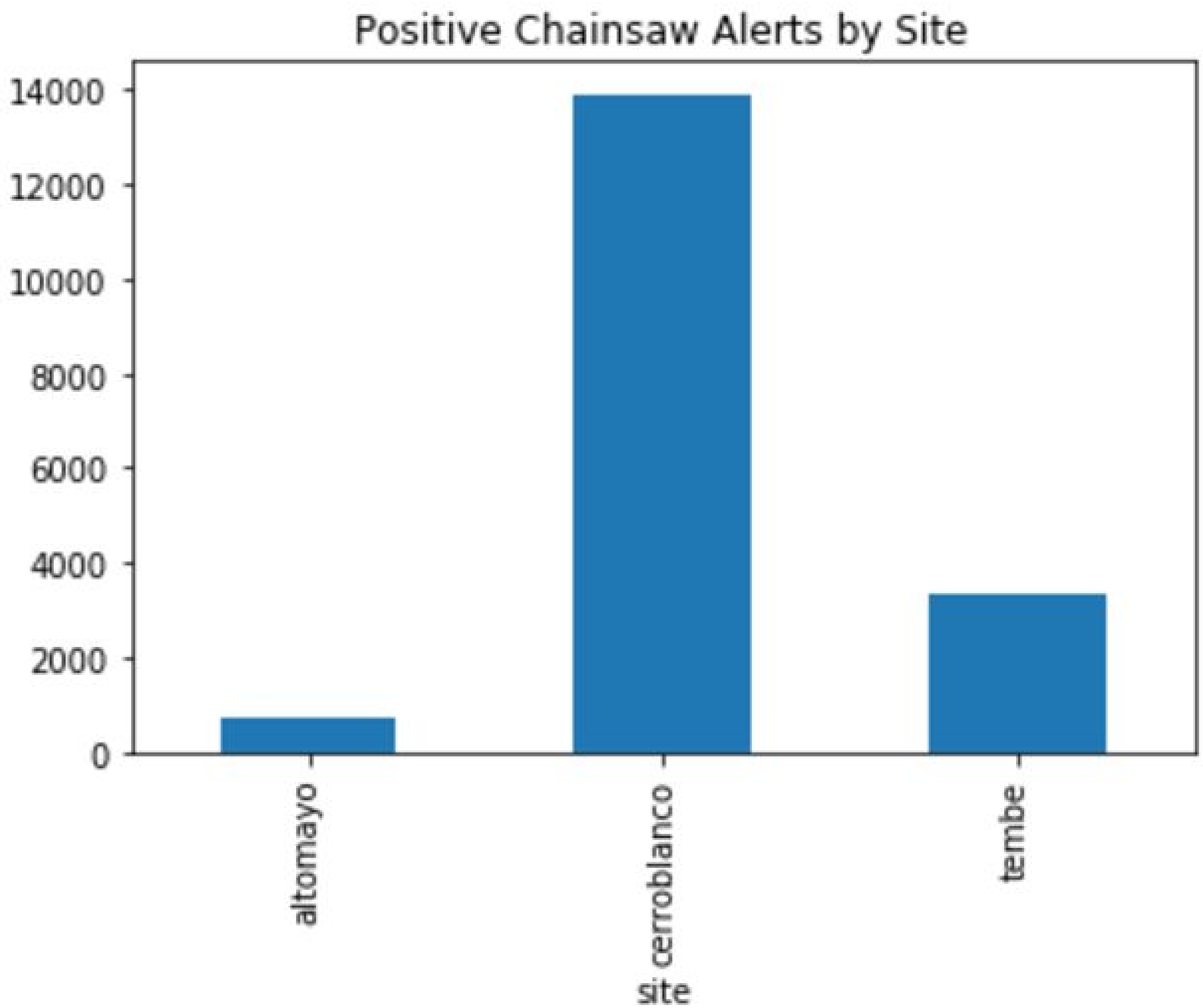
Performance

All strong variants of model architecture had accuracy of over 95%, with recall & precision>98%.

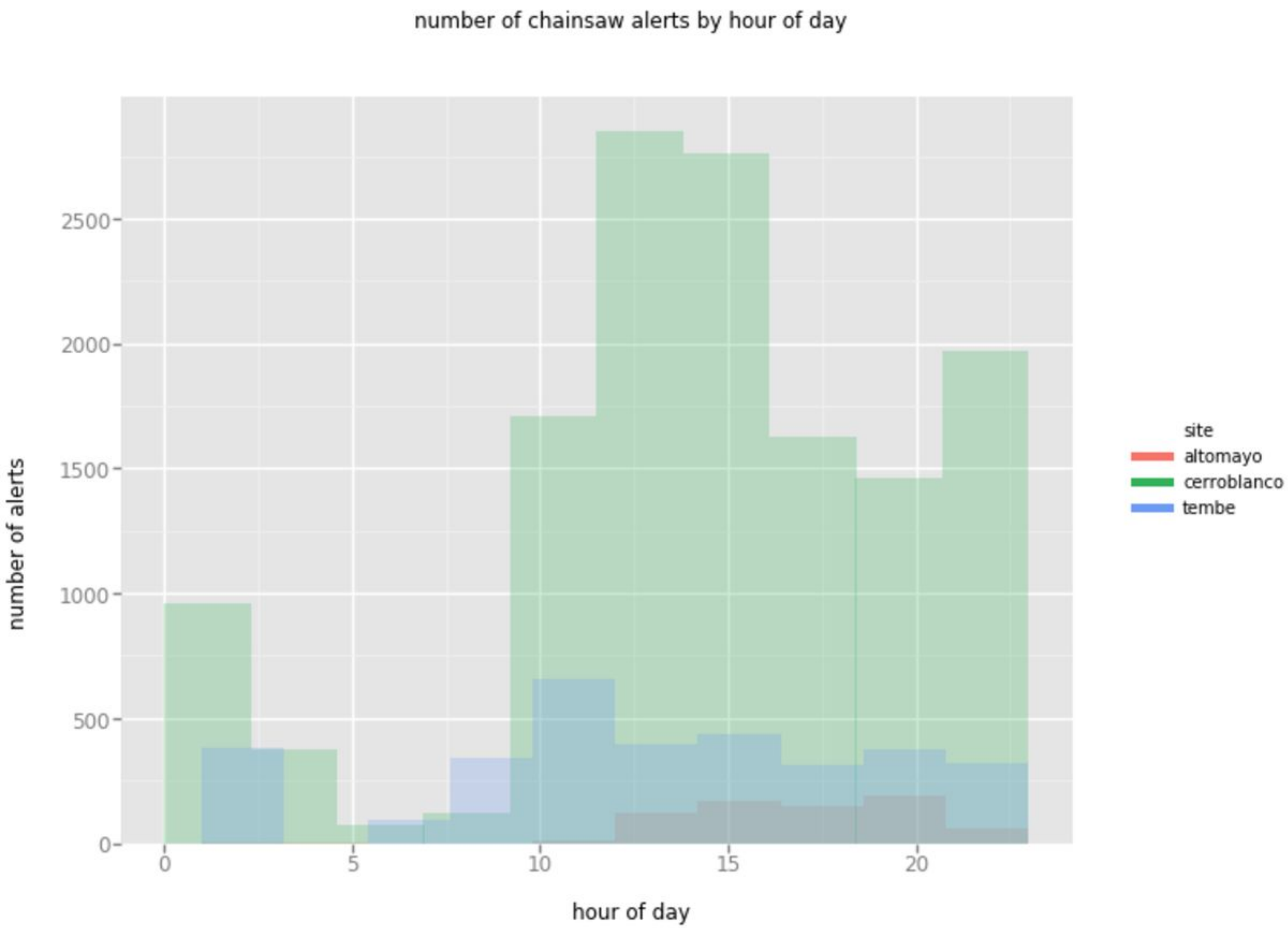
Key issue: performance actually not indicative of test time accuracy. Our limited test set does not have the same data distribution at the rainforest site.

Data gap: we don not know about the chainsaws the model did not catch.

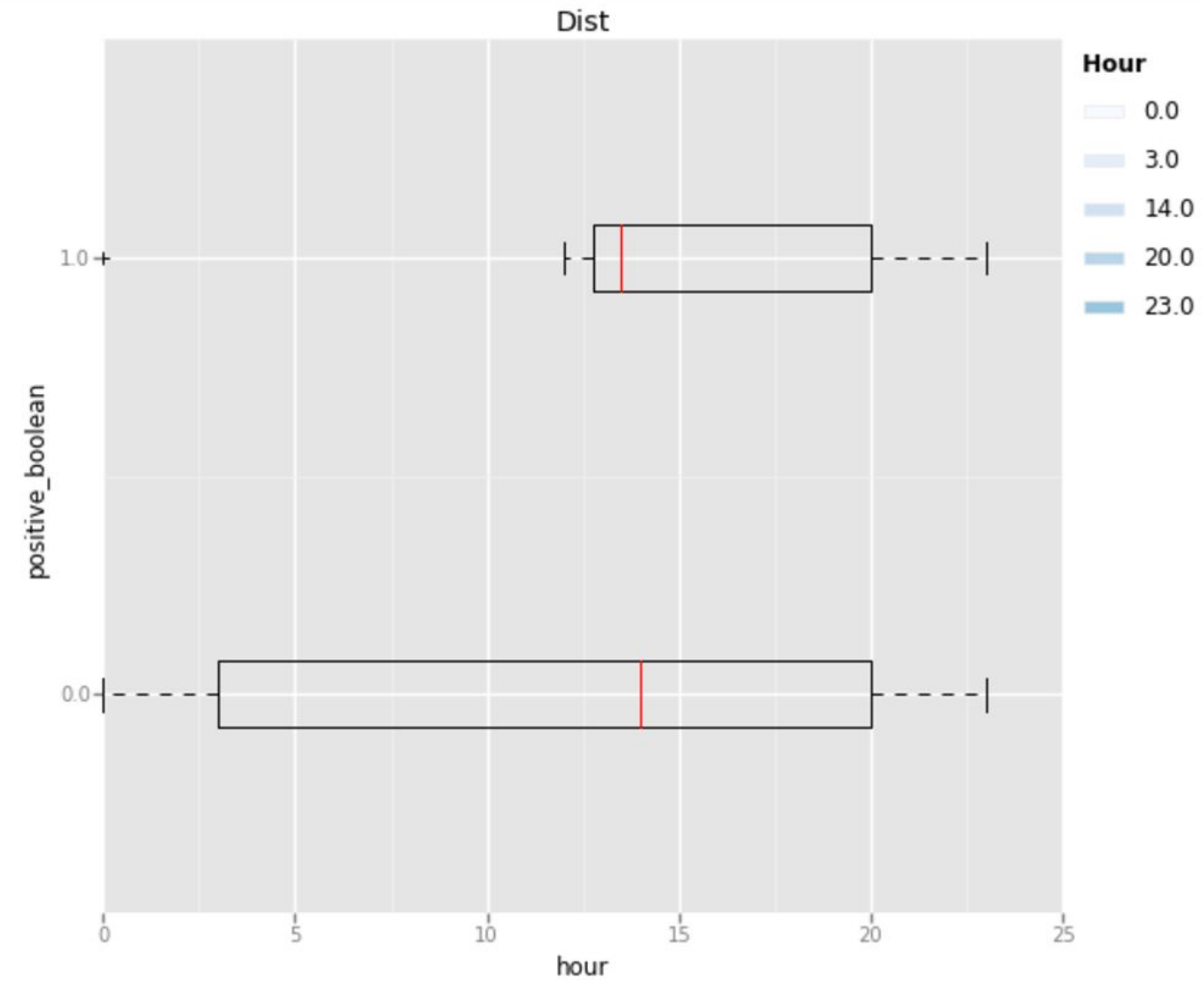
Positive alerts by test site



Seasonality & Time



Seasonality & Time



Performance and lack of mature data set raises key questions:

- true ground truth
- limited labelled data set
- different data distributions at each test site
- accounting for different test conditions (rain, mosquito season)

Ways to address this:

- use ensembles of models (models in production are ensembles)
- incorporate metadata

Data Augmentation

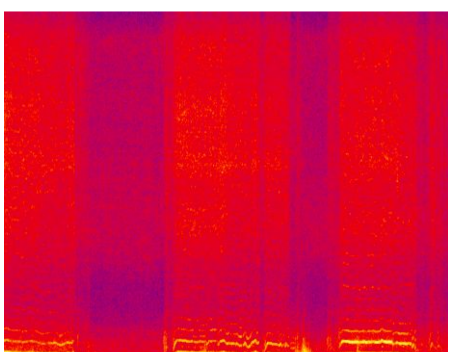
Limited labelled data set:

- initial data set had ~2,400 2s spectrograms. These were built off 45 positively labelled sound clips.

AudioSet

data augmentation

transformation of images



pseudo-labelling

Labelled common background noises provide varied negative class examples.

Negative Class:

	index	id	display_name
111	111	/m/015p6	bird
128	128	/m/09f96	mosquito
130	130	/m/07pjwq1	buzz
283	283	/m/03m9d0z	wind
284	284	/m/09t49	rustling leaves
289	289	/m/06mb1	rain
290	290	/m/07r10fb	raindrop
291	291	/t/dd00038	rain on surface

Positive class:

	index	id	display_name
347	347	/m/01j4z9	chainsaw
424	424	/m/0_ksk	power tool
425	425	/m/01d380	drill
438	438	/m/07pczhz	chop
439	439	/m/07pl1bw	splinter

Bridges the gap between our experiment data and data distribution at test site.

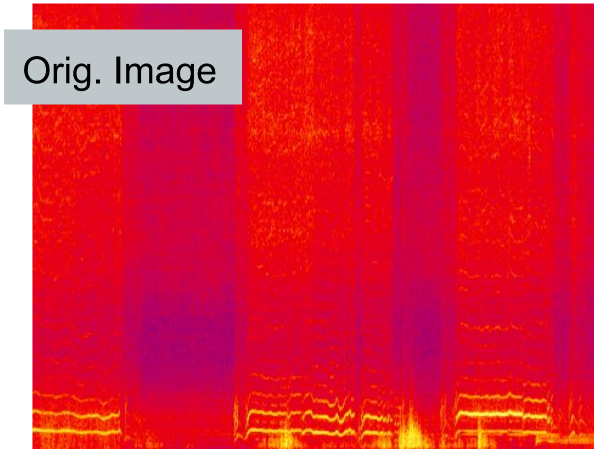
10 sec labelled clips.

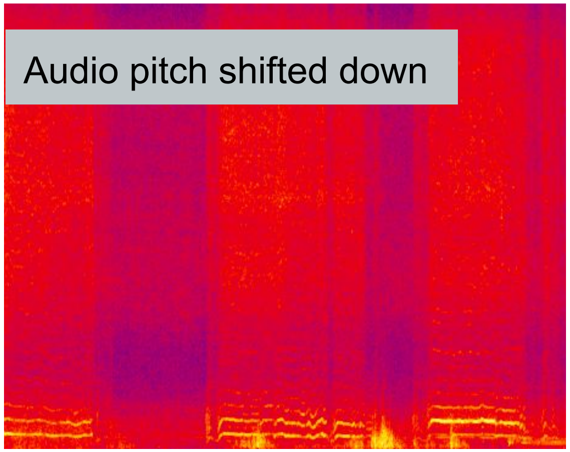
extremely valuable: chainsaw
less valuable: rustling leaves

Overall, at this stage of research quantity of data is more important. Particularly for providing stratified examples of the negative class (for example mosquito).

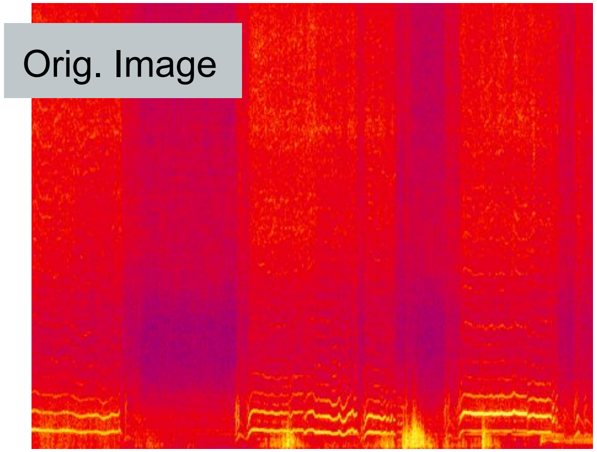
Transformations of labelled data.

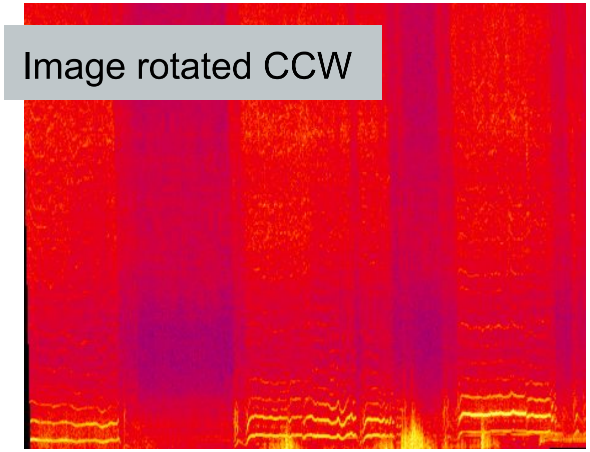
Pitch Shift





Rotate Image

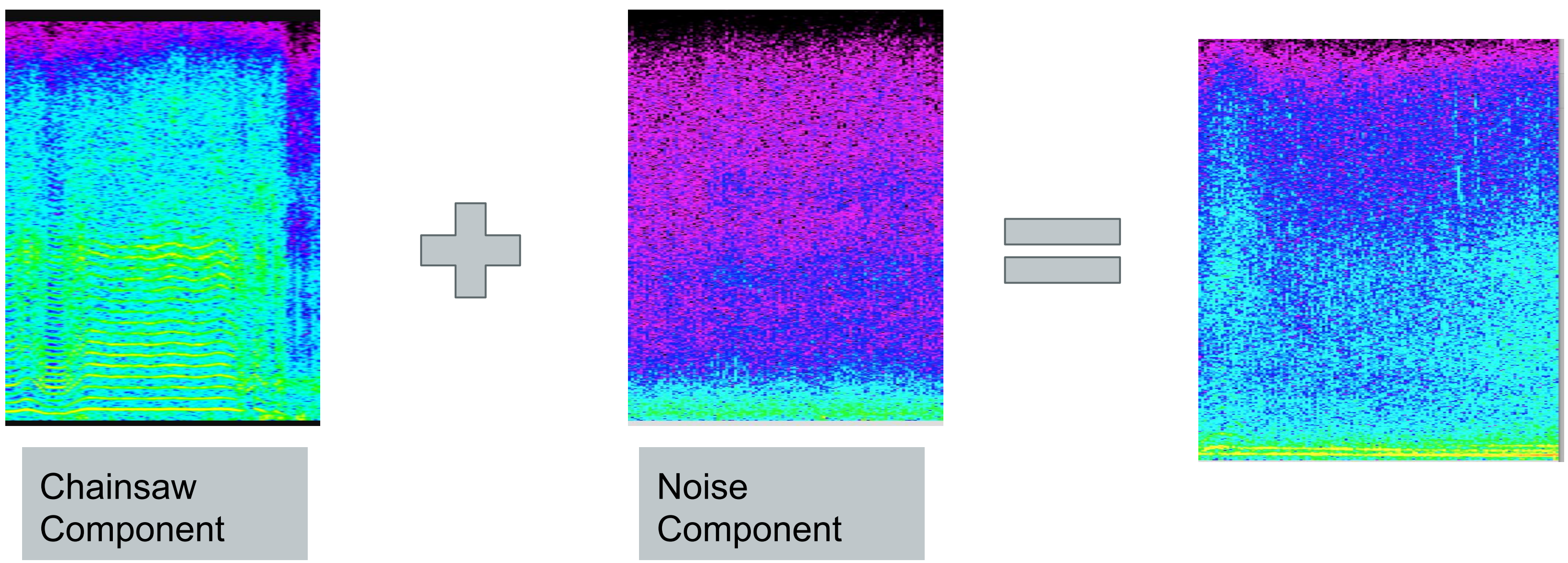




Improve Data Preprocessing

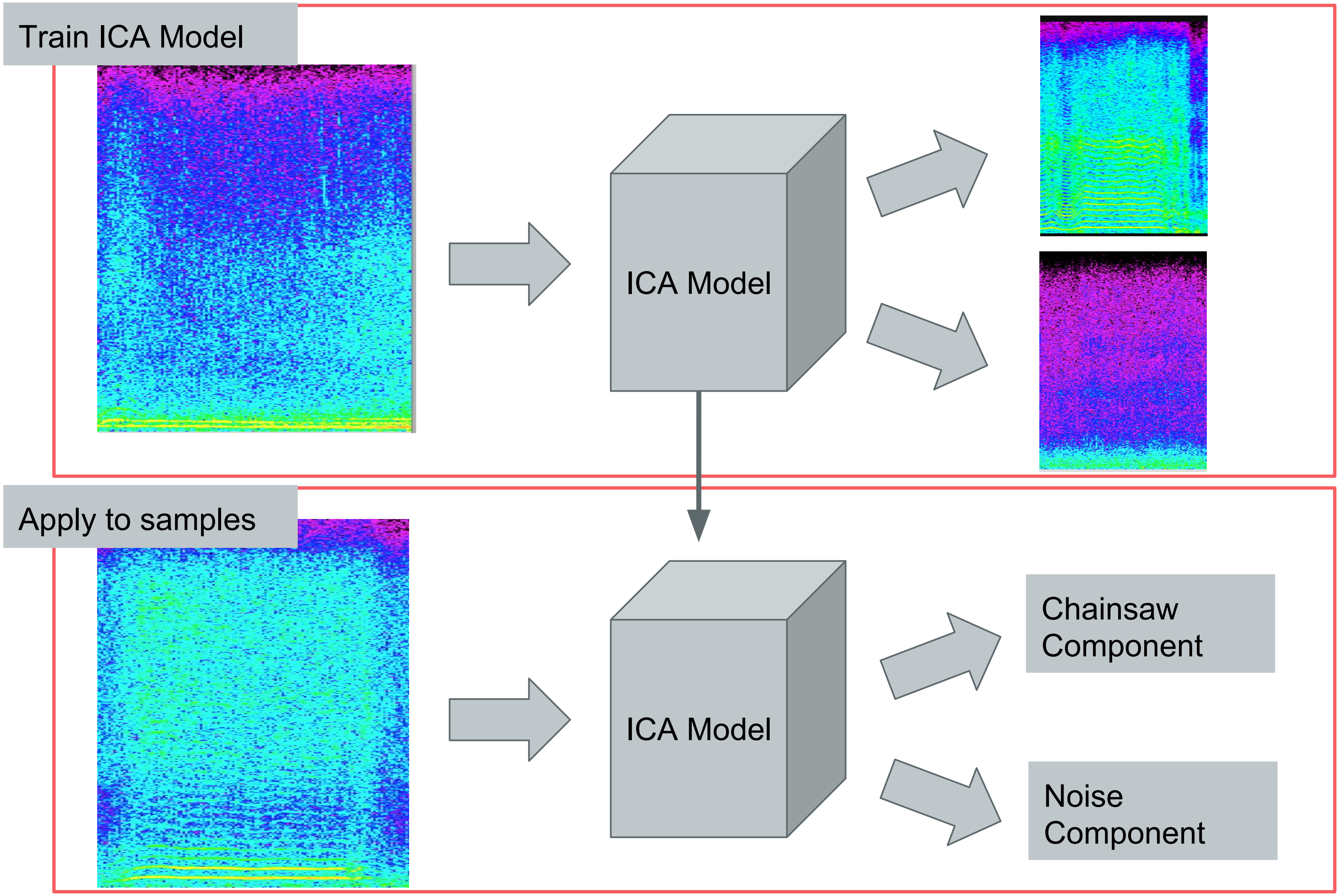
Independent Component Analysis

- Neural networks attempt to find structure
- Goal: reduce the computational overhead during training
- Simple assumption:
  - Two components present in positive examples
  - Chainsaw-like component
  - Background noise component



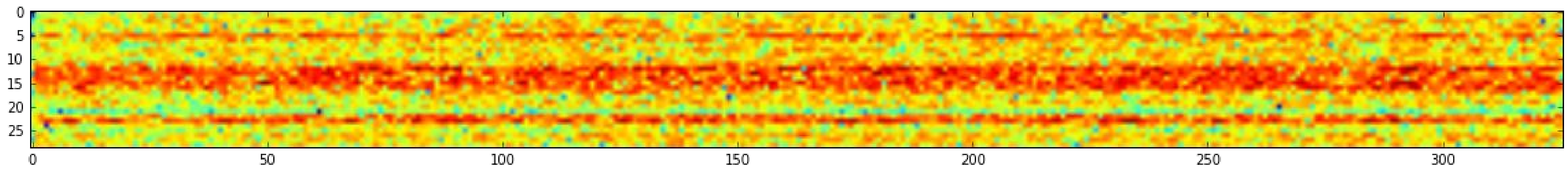
ICA allows us to isolate chainsaw component

- Take positive examples and decompose them into two components, then apply this model to all samples

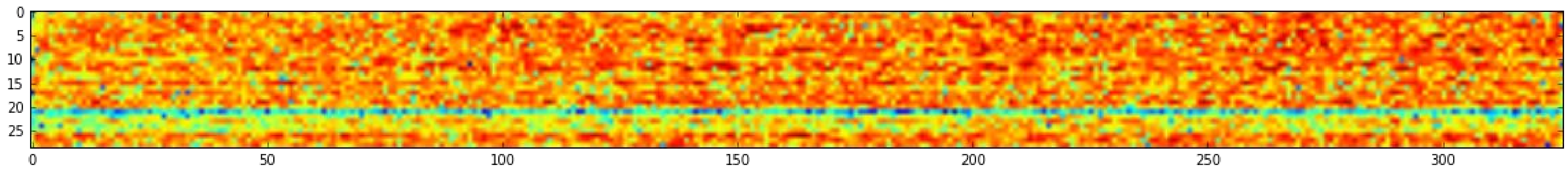


Independent Components Analysis

- Possible to learn two distinct signals
- Current impact on classification accuracy unknown -- unclear how to integrate two spectrograms during model training



Component 1



Component 2

2. Model Directionality of Sound

Model direction and distance of sound

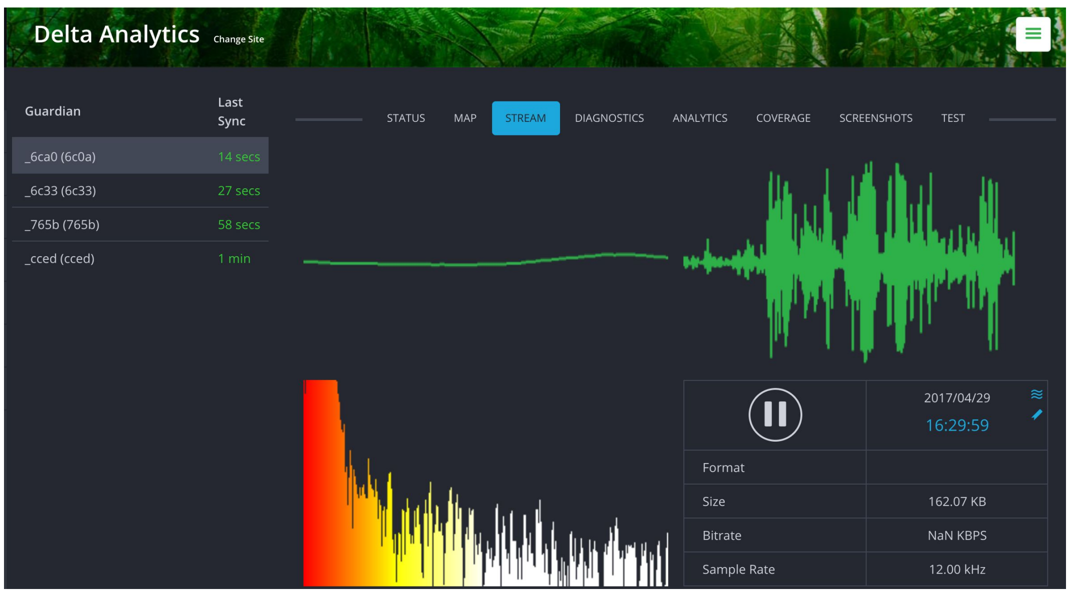
Providing direction and distance of sound reduces the area that rainforests conservationists have to search when they receive an alert.

Delta Analytics Test Site

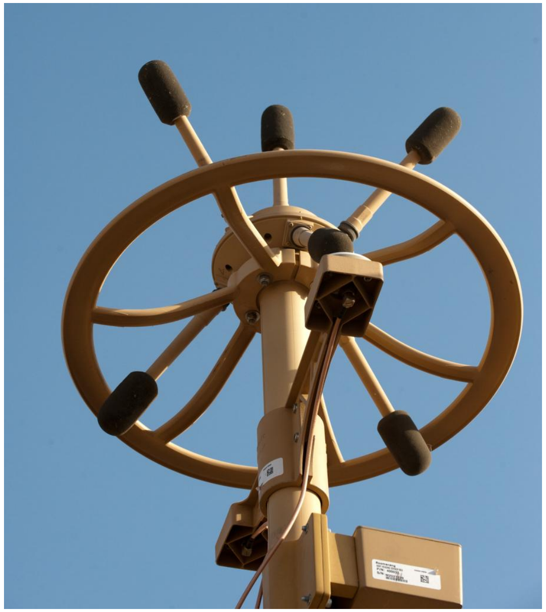
We have been given four Guardians and our own site

Goal of Delta site is to generate training data to support our research areas

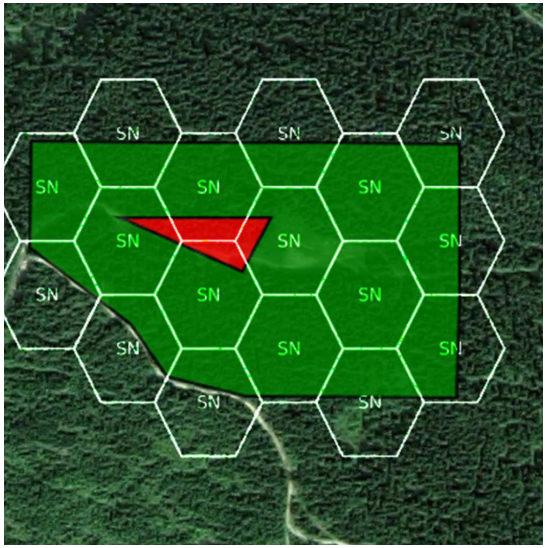
Guardians are not fixed, vary configuration



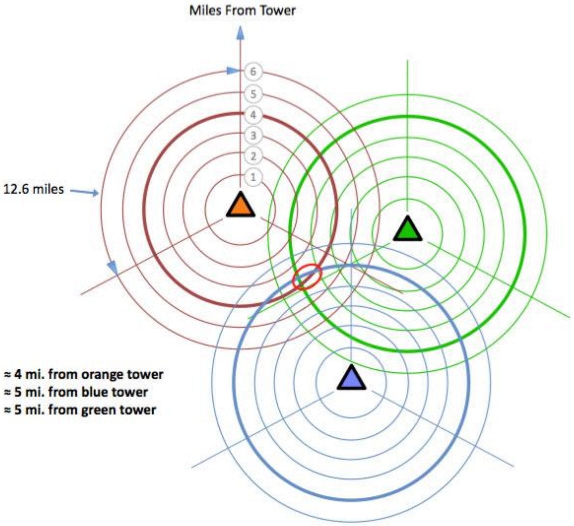
Approach: Source localization based on time delay of arrival



Gunshot Detection



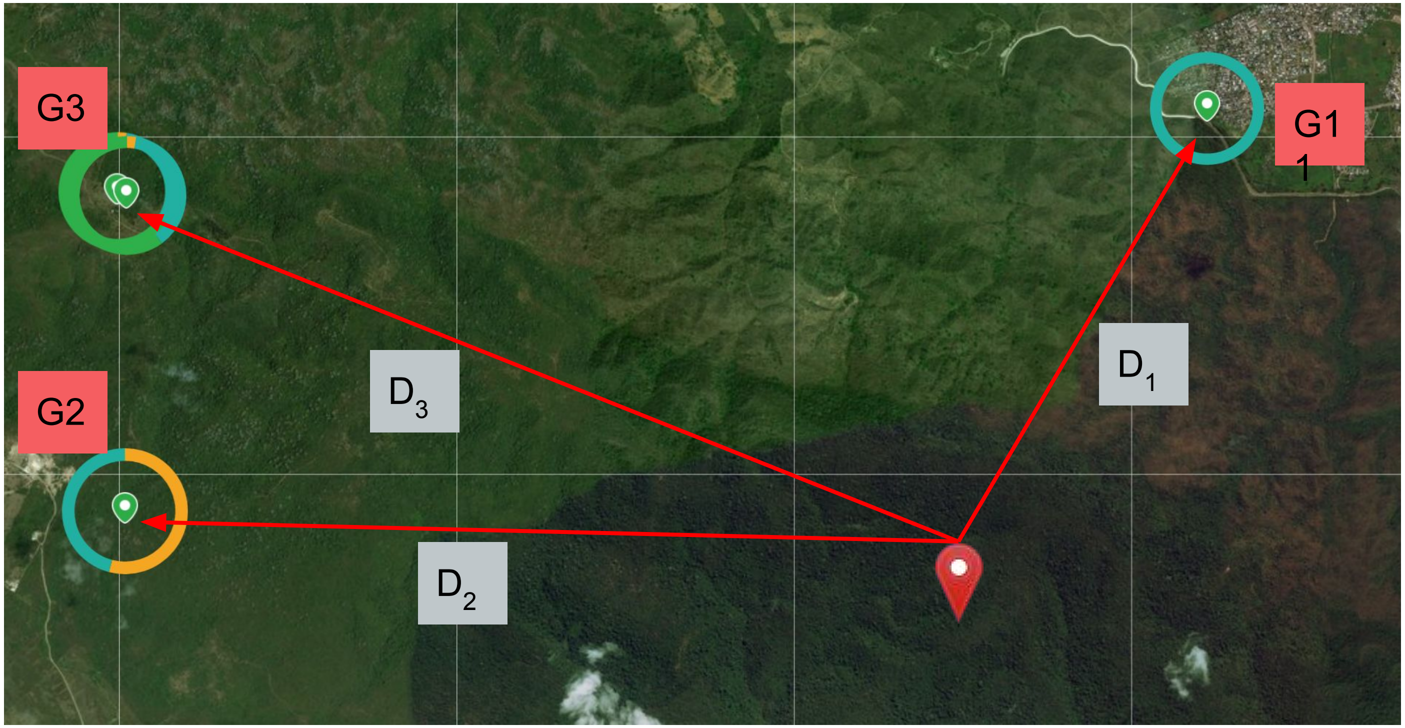
Localization in wireless sensor network



Cell tower triangulation

http://www.defenceimagery.mod.uk
http://article.sciencepublishinggroup.com/html/10.11648.j.ijssn.20150303.12.html
https://wrongfulconvictionsblog.org/2012/06/01/cell-tower-triangulation-how-it-works/

We locate the sound using the time delay between sensors.



Time Delay

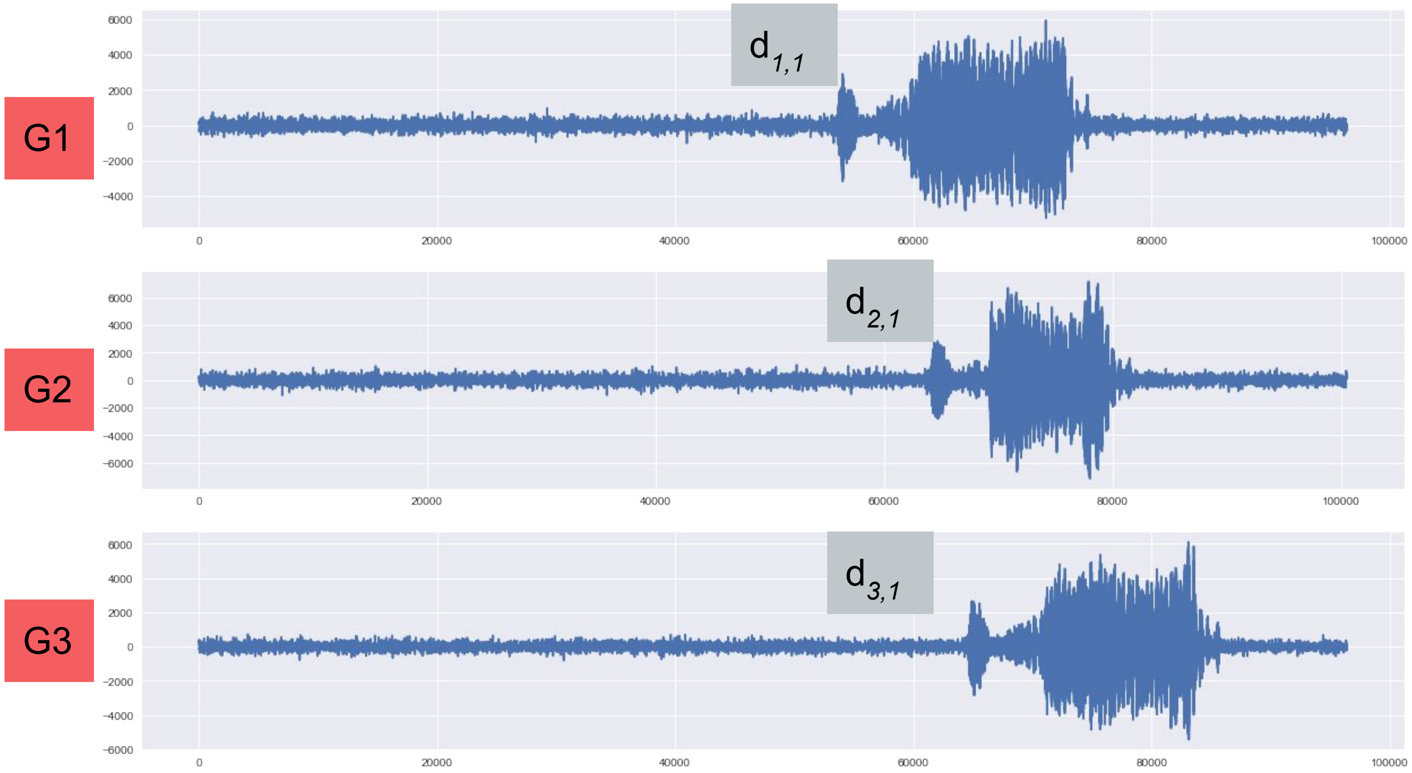
Sound reaches further away sensor later than closer sensors

Detection time ($d_{i,j}$) at Guardian *i* of sound *j* (time t_j)

$$d_{i,j} = t_j + \frac{D_i}{S_{sound}}$$

Don't know t_j or D_i! But assume location of sensor is known

Determine relative delay for all guardian pairs.

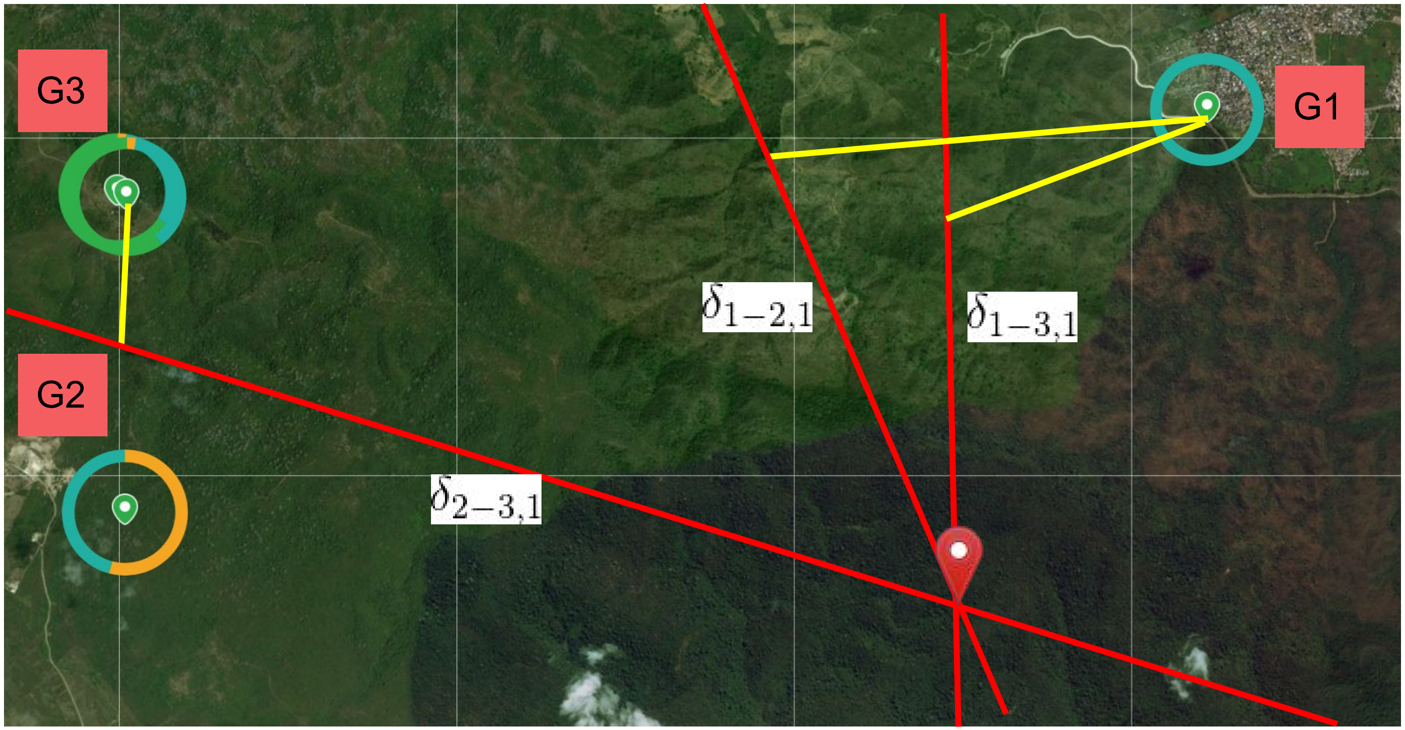


This can be formalized as:

$$\delta_{a-b,j} = d_{a,j} - d_{b,j}$$

Actual delays are determined using cross correlation

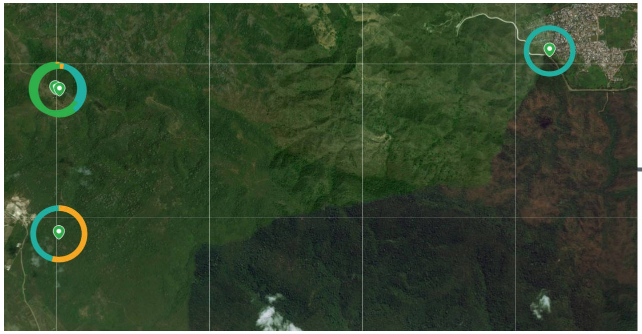
Triangulation



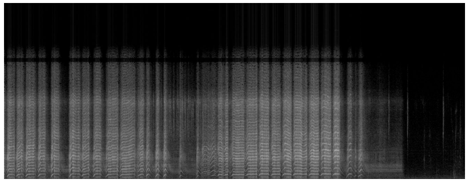
Each $\delta_{a-b,j}$ determines a line in two dimensions along which the source can be.

Multiple lines determine the point of the source.

System End-to-End

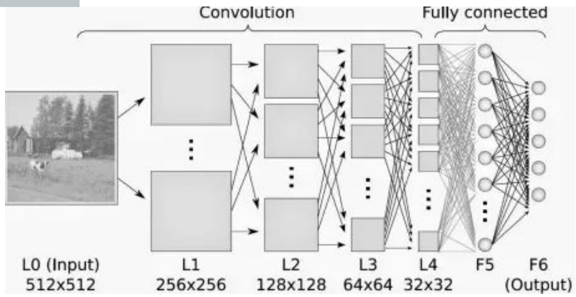


Record audio at sites

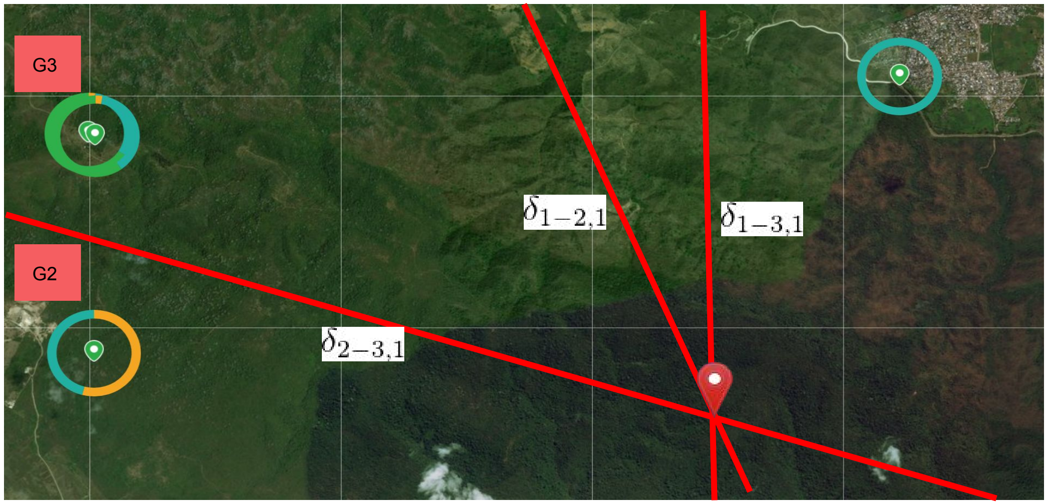


$$Metadata = \begin{Bmatrix} Site \\ Time \\ Weather \end{Bmatrix}$$

Transmit audio and metadata



Classify audio using CNN



Provide alert and direction estimate

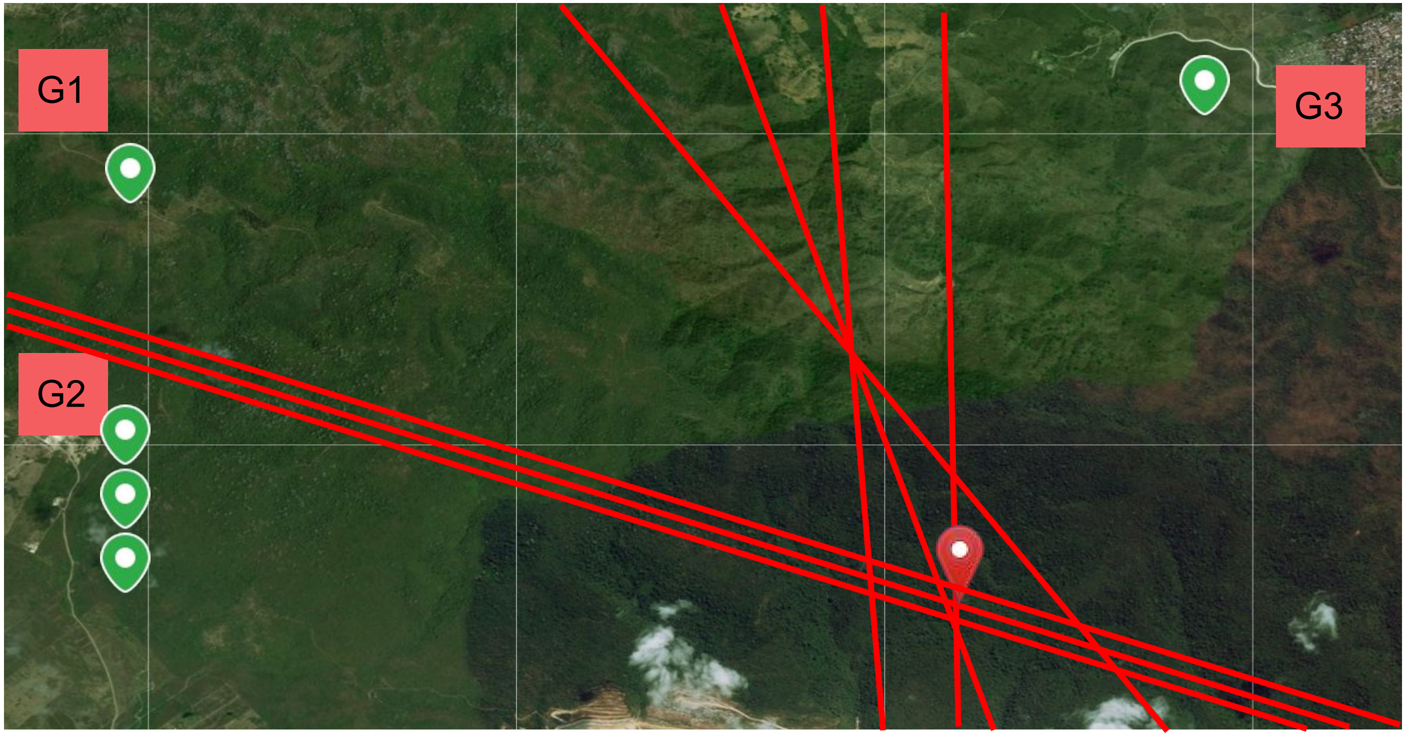
Positive Detections Only
-> Direction estimate
-> Alert ranger

Unique Challenges Posed by Our Environment

Unique challenges posed by rainforest deployment environment

1. Relative position of guardians unknown
2. Guardians lose sync in time with each other and drop data
3. Large coverage area

1. Guardian position is unknown



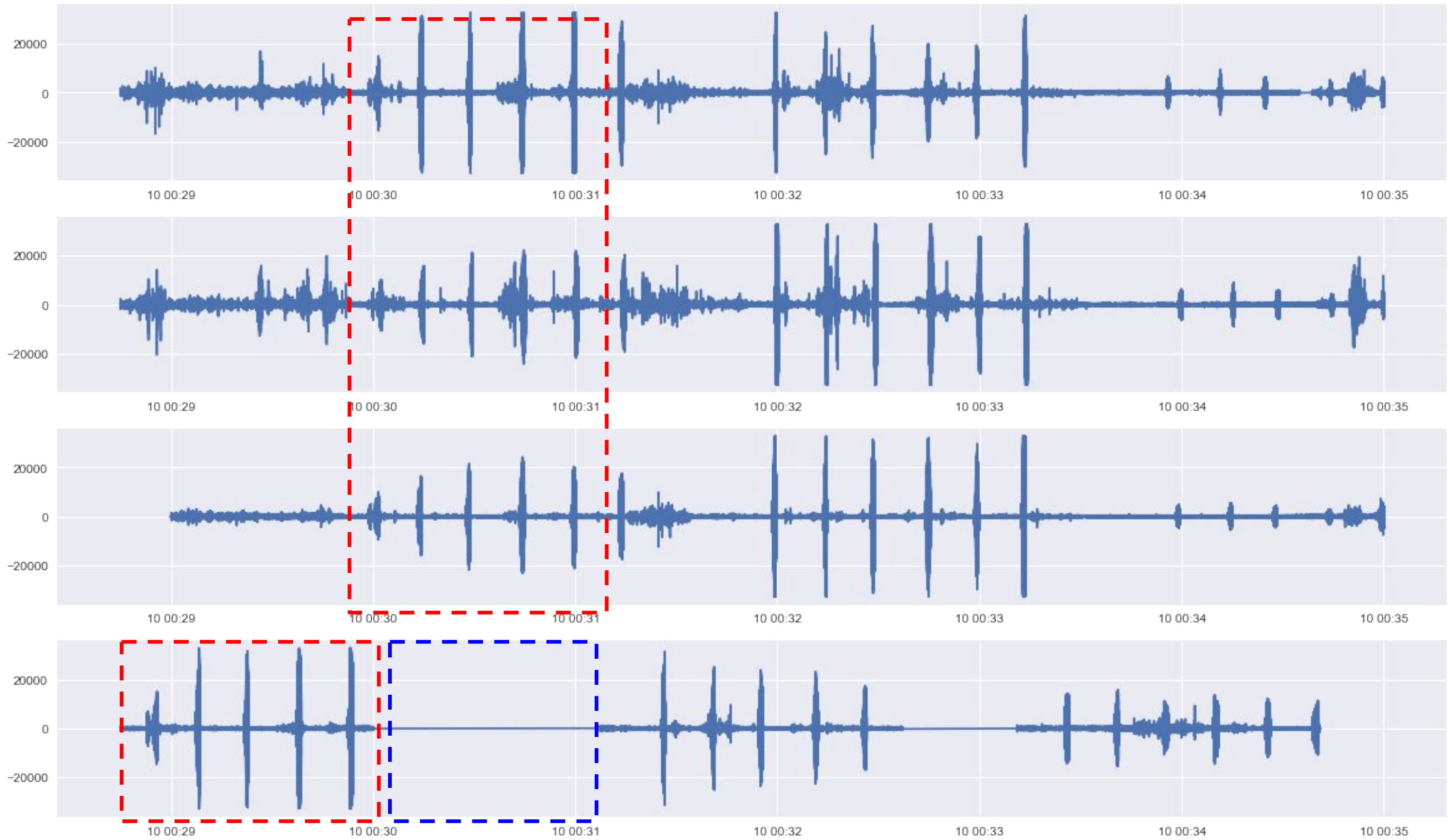
Source + sensor localization

Two-stage solution:

1. Determine sensor locations
2. Determine source locations

Both can be solved using gradient descent optimization.

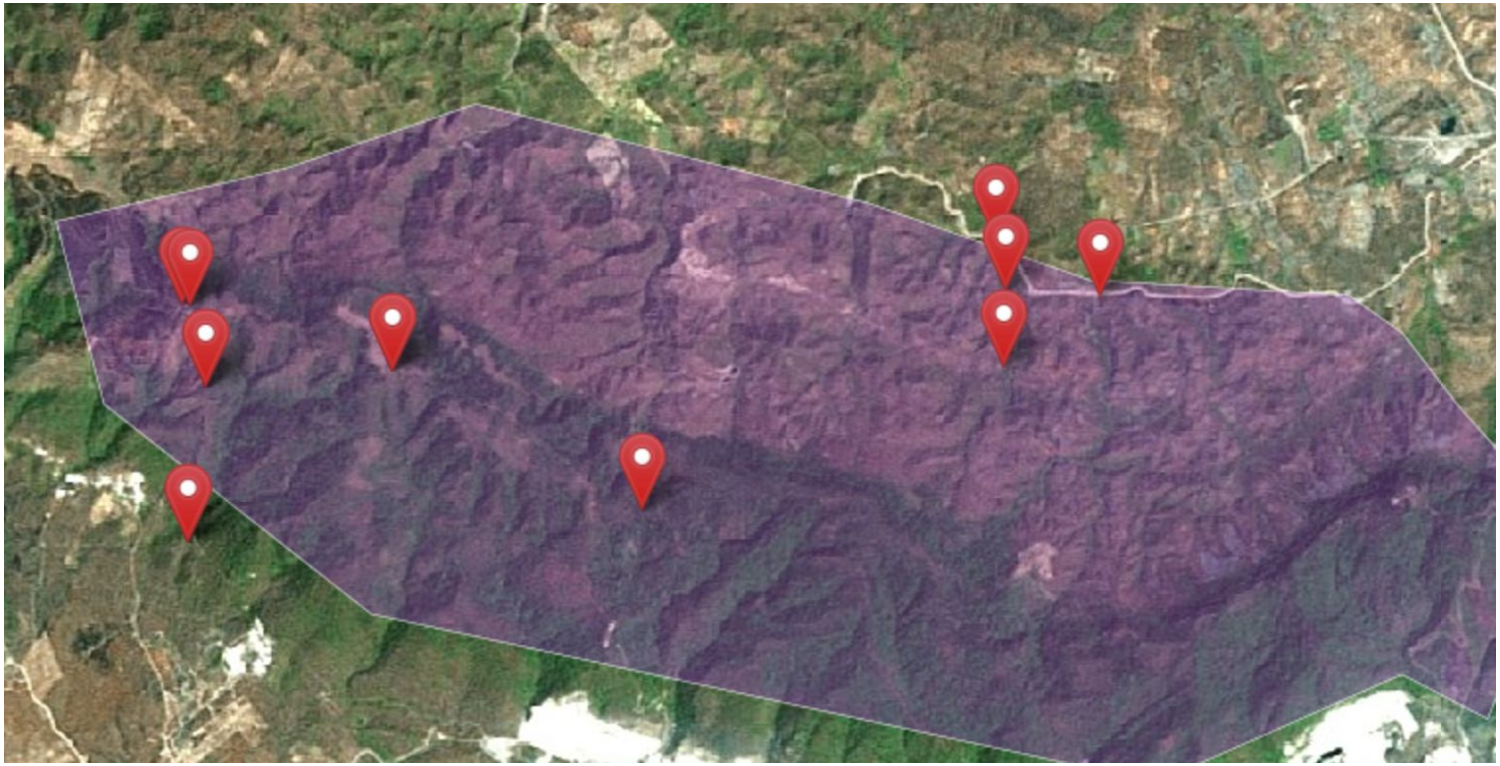
2. Guardians lose sync in time and drop data



Guardians can lose time sync or completely drop data which results in incorrect estimation of Guardian location and source location

Algorithm must be robust to these types of errors.

3. Large coverage area



Optimal guardian placement is very important. We desire large coverage area with minimal number of Guardians

Increase distance between Guardians:

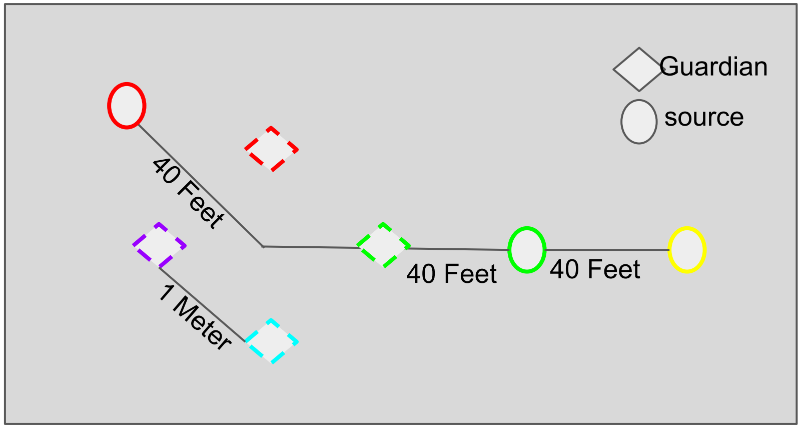
- Fewer Guardians detect each incident
- Decrease accuracy of direction algorithm

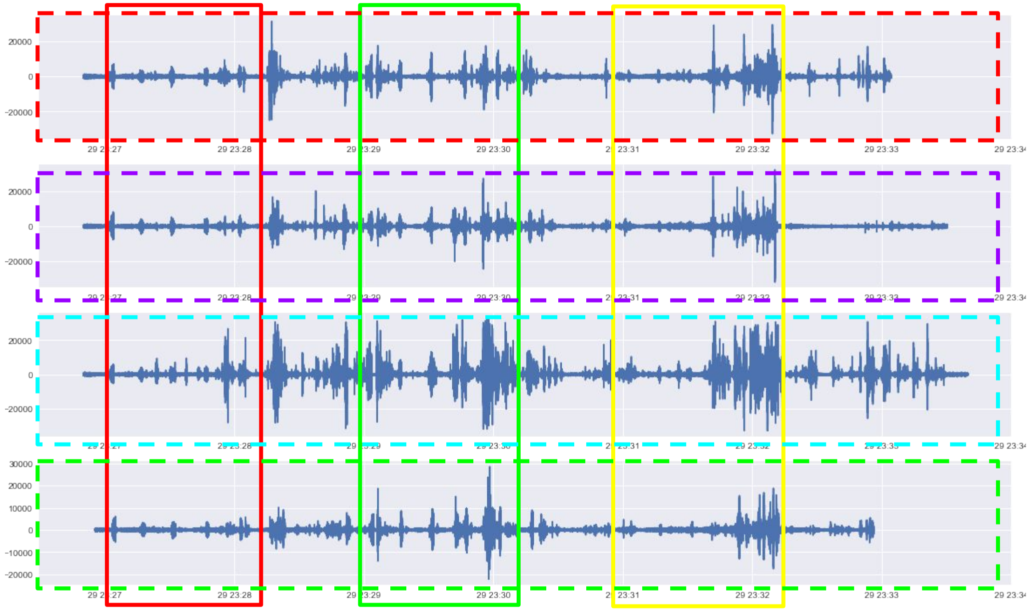
Directional Experiments

Initial Experiments

- Walk circularly around Guardian array and generate multiple impulsive sounds (e.g., whistle) and/or harmonic sounds (e.g., chainsaw, birdsong) at given locations (i.e., hours on a clock-face)

- Walk in straight line away from Guardian array and generate multiple impulsive sounds every N feet





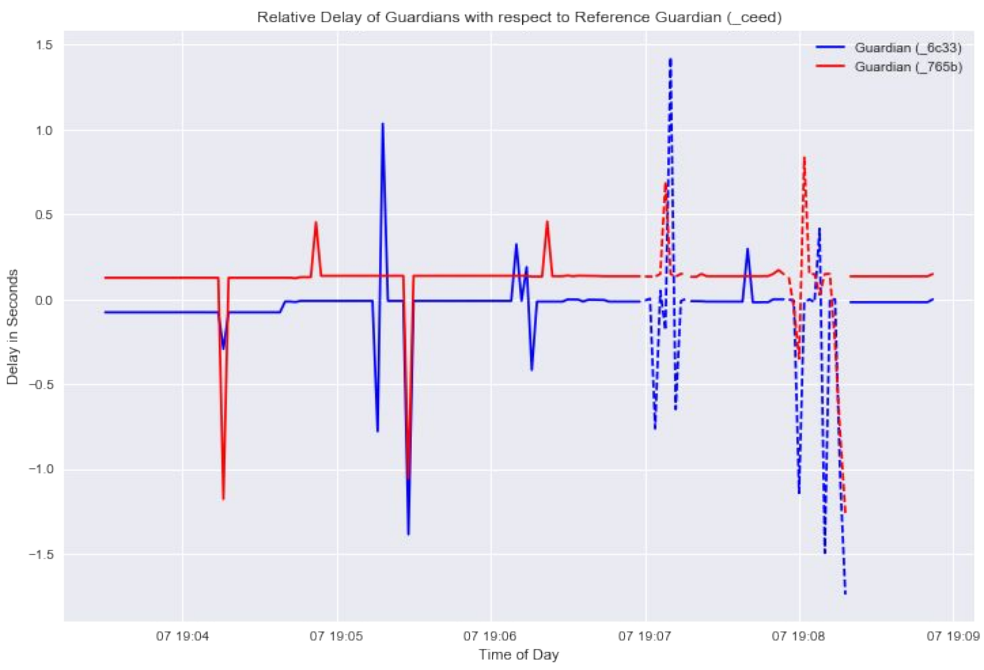
Experiments with Chainsaw

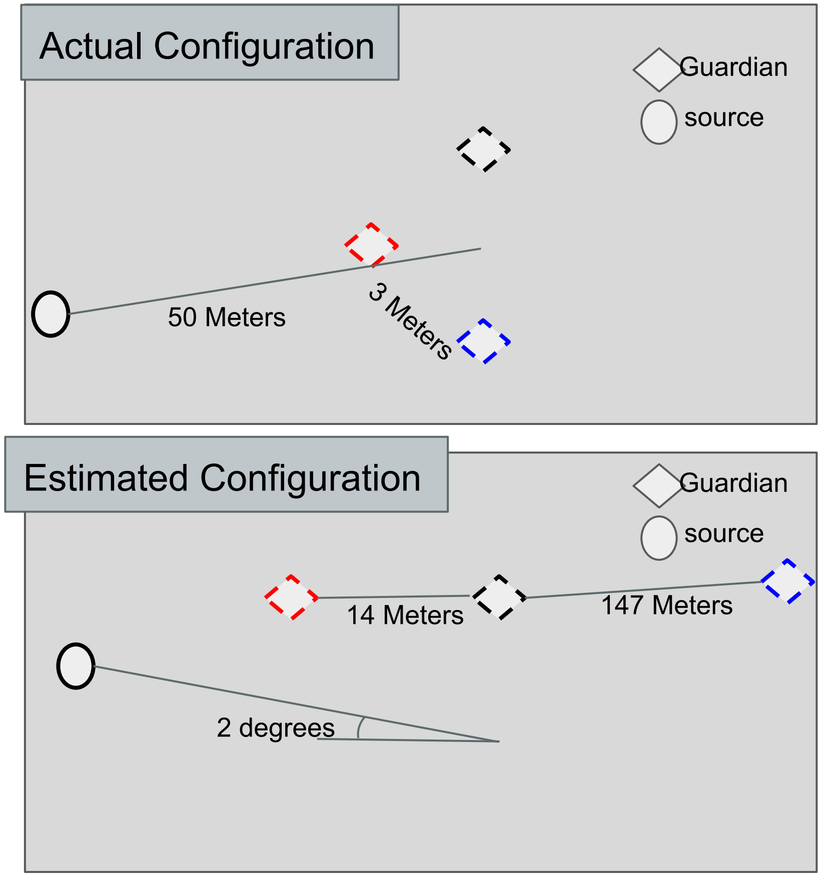
Captured three hours of chainsaw sounds on four Guardians

Chainsaw Experiments

Estimated delays are inaccurate at audio sample boundaries, and during periods of ambient noise

Excessive delays flatten estimated sensor array and source angle





Next steps!

RAINFOREST CONNECTION

- Deliver final project in July. Research poster session at University of San Francisco.
- Make code public as a resource to researchers
- Collaborate on research paper on estimating directionality

We will start accepting fellow applications for the 2018 cohort in July.

Note: fellows are all based in the Bay Area.

http://www.deltanalytics.org/

Questions?

Appendix

Model Training Interface

Why?

- Use a standard interface for model training, so that researchers can spend time on features and models rather than plumbing
- Help enable a transition to automated model training in the cloud
- Create a feature datastore with a standard format, where there's a record of how each feature was created

How?

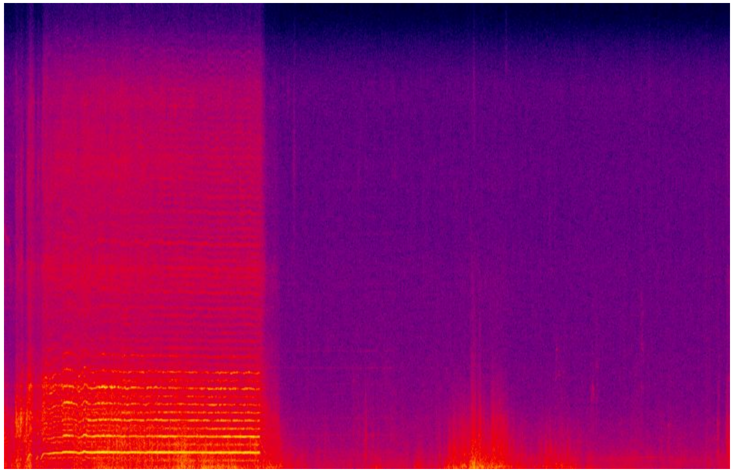
- Write an api client to automate downloads
- Develop an interface for feature writing and reading
    - so that researchers can create new features to plug into the model
- Create an interface for data selection
    - so that researchers can play with different training and data augmentation techniques
- Save a full record of the parameters of every action (data download, feature creation, model training)
    - so that analysis is repeatable

CNN architecture with a few hidden layers (similar to AlexNet)

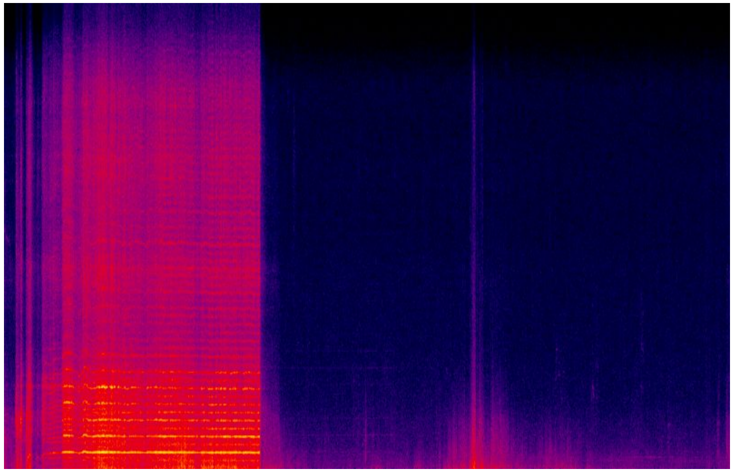
- RELU activation function
- Multiple convolutional layers interspersed with pooling layers
- Multiple fully connected layers
- Dropout used to avoid overfitting

Pre-processing steps in production

Orig. Image



1. Agate Filter



2. Subtracting the mean