astræa

Mapping Deforestation Trends in the Brazilian Amazon Using Machine Learning

Jason Brown, Senior Data Scientist

Courtney Whalen, Data Scientist

Agenda

- Overview
- Data
- Software
- Model
- Results
- Conclusions

Extracurricular

- Machine learning happy hour 7pm in Lindbergh Room
- Visit us in booth 14 all week!
- "Using Deep Learning to Derive 3D Cities from Satellite Imagery" Wednesday at 2pm in Gateway II

astræa

Overview

Goal

- Identify areas of the planet where deforestation is occurring
  - Monitoring deforestation is important for understanding impacts on climate change

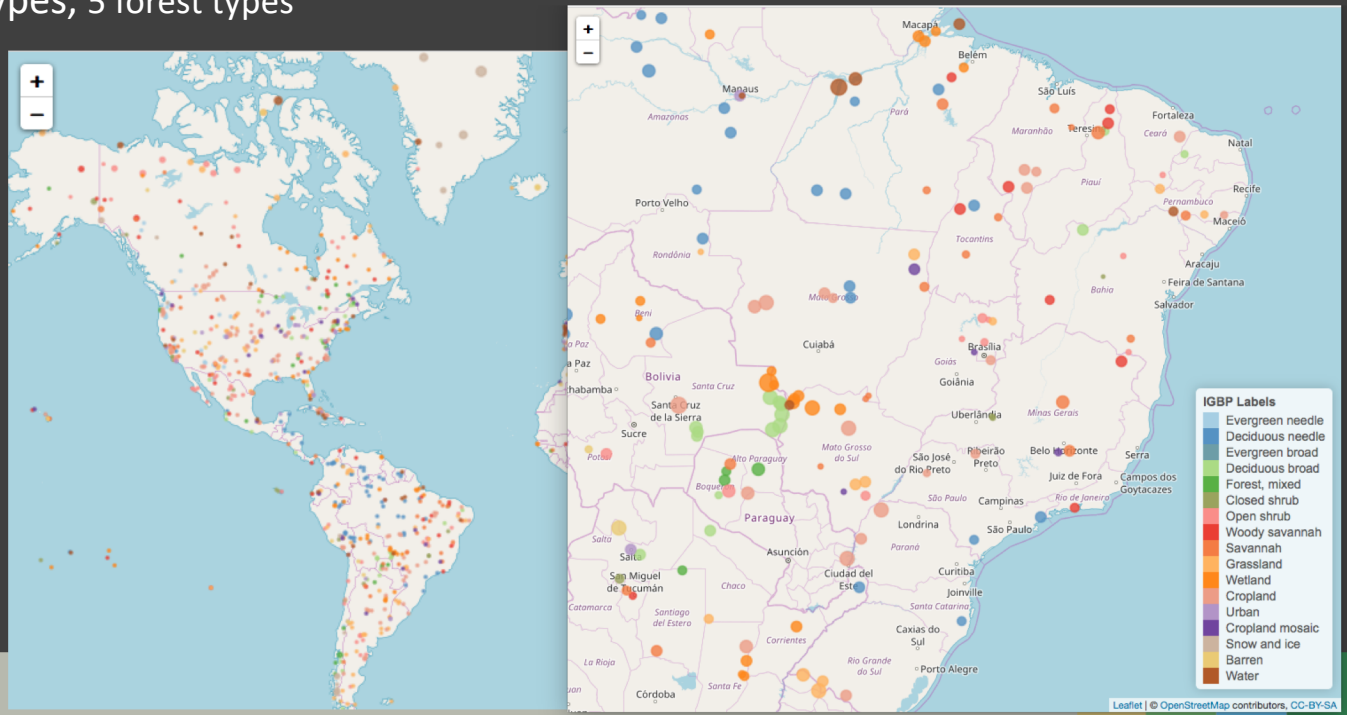
Methodology

- Supervised classification model using remote sensing data
  - Binary target
- Post-classification change detection

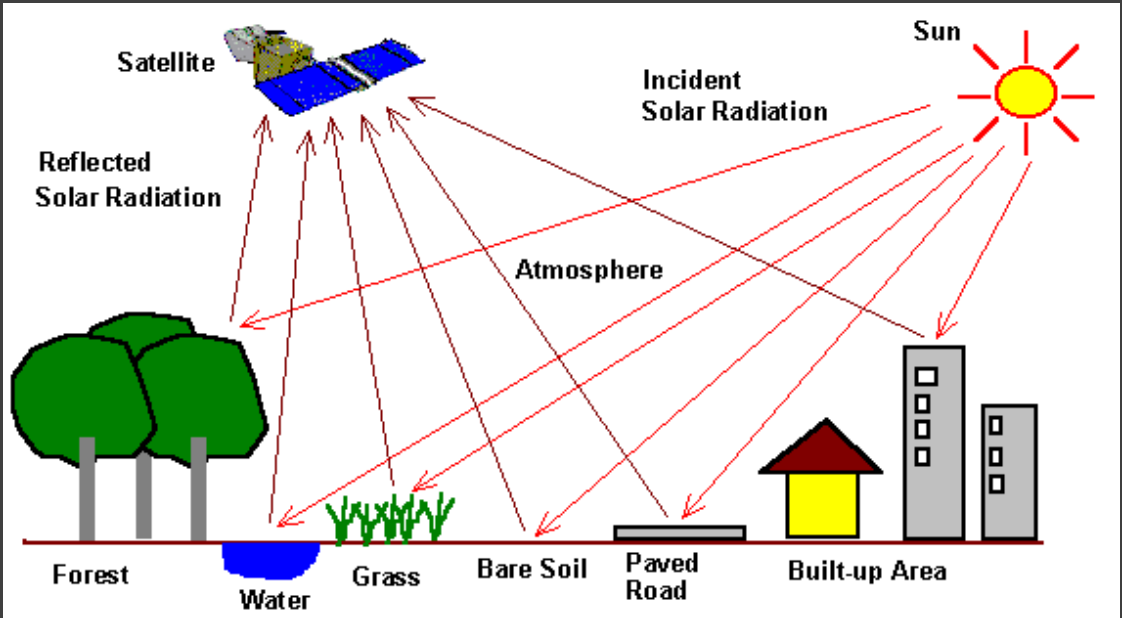
Data

Labeled Data

- System for Terrestrial Ecosystem Parameterization (STEP) reference data
  - ~2,000 hand-labeled sites scattered across every continent; 10,000 km^2 or 4000 mi^2
  - 17 land cover types, 5 forest types



Remote Sensing

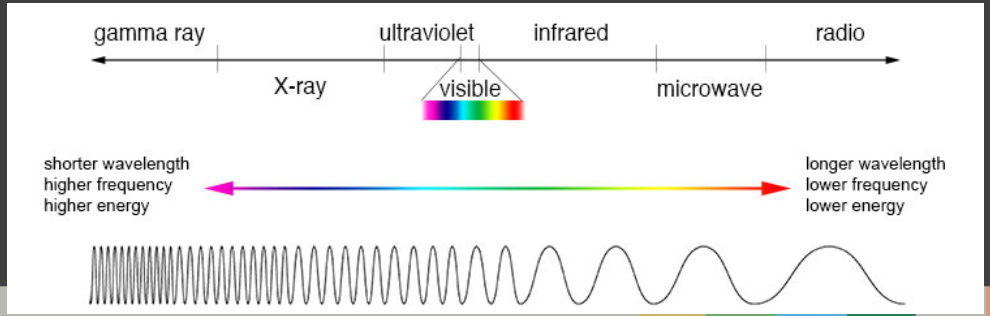


astræa

Feature Data

- Source:
  - Moderate Resolution Imaging Spectroradiometer (MODIS)
  - Surface reflectance
  - Revisit rate: 1-2 days
  - Spatial resolution: 500 m
- Features:
  - 7 spectral bands & normalized difference vegetation index (NDVI)
  - Monthly mean
  - Yearly aggregates: min, max, mean, variance
  - 128 features in all

Band #	Spectral band	Bandwidth (nm)
1	Red	620 - 670
2	NIR	841 - 876
3	Blue	459 - 479
4	Green	545 - 565
5	SWIR	1230 - 1250
6	SWIR	1628 - 1652
7	SWIR	2105 - 2155



Software

RasterFrames

- Why Spark?
  - Distributed computation allows global scale processing
  - SQL and ML Pipeline APIs
  - Top level Apache project
- Why RasterFrames?
  - Spark SQL interface to geospatial raster data
  - Available in Python and Scala
  - LocationTech incubator project
- Apache 2.0 licensed

Learn more and try it at **RasterFrames.io**

Label and Feature Rasters

- Label data: Land cover polygons -> binary forest attribute -> bit type raster
- Feature data: Monthly and yearly surface reflectance and NDVI summaries

spatial_key	bounds	forest
[208,46]	POLYGON ((3113461.455346264 4781387.23499605, 3113461.455346264 4892582.286972702, 3224656.5073229186 4892582.286972702, 3224656.5073229186 4781387.23499605, 3113461.455346264 4781387.23499605))	
[208,45]	POLYGON ((3113461.455346264 4892582.286972702, 3113461.455346264 5003777.338949354, 3224656.5073229186 5003777.338949354, 3224656.5073229186 4892582.286972702, 3113461.455346264 4892582.286972702))	

Feature Engineering with RasterFrames

```
SELECT spatial_key,
    rf_localAggMin(red) as red_min,
    rf_localAggMax(red) as red_max,
    rf_localAggMean(red) as red_mean
FROM df
GROUP BY spatial_key
```

```
df.groupBy(df.spatial_key).agg( \
    localAggMin(df.red).alias('red_min'), \
    localAggMax(df.red).alias('red_max'), \
    localAggMean(df.red).alias('red_mean'))
```

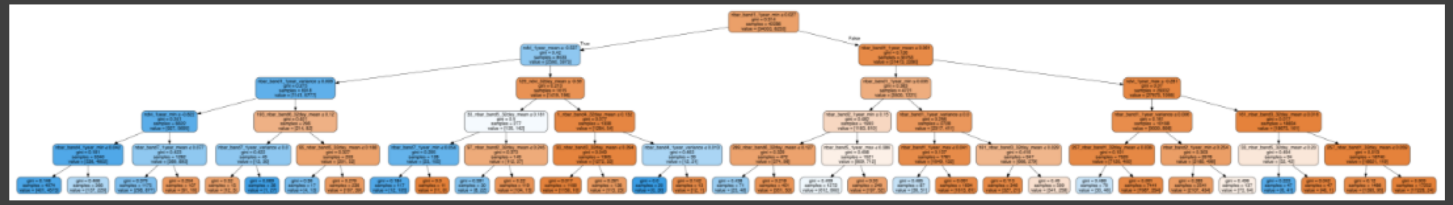
astræa

```
df.groupBy("spatial_key").agg(
    localAggMin($"red") as "red_min",
    localAggMax($"red") as "red_max",
    localAggMean($"red") as "red_mean")
```

Model

Model Training

- Trained on MODIS 2012 data at pixel level
- 80% train, 20% test
  - Train/test split strategy:
    - All pixels from each STEP site kept within the same set
    - Balanced forest/not forest pixels within each set



astræa

Model Selection

Model	Parameters	Accuracy
Decision Tree	Depth = 10	0.939
	Depth = 15	0.926
	Depth = 20	0.914
Random Forest	Depth = 10	0.943
	Depth = 15	0.951
	Depth = 20	0.952
Gradient Boosted Trees	Depth = 10	0.941
	Depth = 15	0.928
	Depth = 20	0.912
Neural Net	5 layers	0.935
Logistic Regression	Elastic net = 0.5	0.918

Final model: Random Forest

- 1000 trees
- Depth of 15
- Minimum of 10 instances per node

Predicted

Actual: Forest?	True	False
True	1642	452
False	70	8496

Precision = 0.96
Recall = 0.78

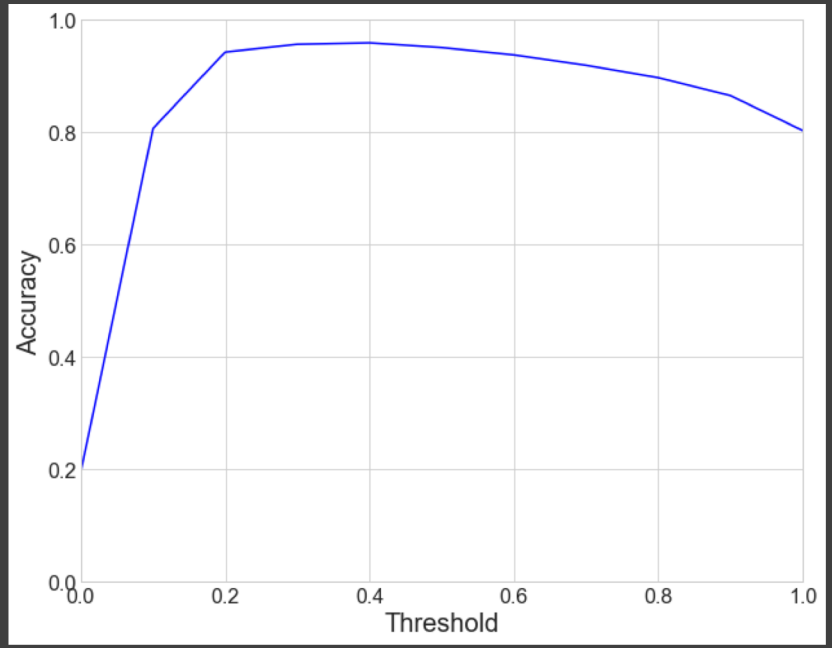
astræa

Feature Importance

1. Red band - year aggregate minimum
2. Red band - year aggregate mean
3. NDVI - year aggregate mean
4. Green band - year aggregate mean
5. NDVI - year aggregate maximum
6. Blue band - year aggregate minimum
7. Blue band - year aggregate mean
8. Green band - year aggregate minimum
9. Green band - year aggregate maximum
10. NDVI - June aggregate mean
11. Red band - August aggregate mean
12. NDVI - year aggregate minimum
13. Blue band - year aggregate minimum
14. Red band - year aggregate minimum
15. Red band - June aggregate mean

Forest Threshold

Threshold	Accuracy
0	0.196
0.1	0.807
0.2	0.943
0.3	0.957
0.4	0.959
0.5	0.951
0.6	0.938
0.7	0.920
0.8	0.897
0.9	0.866
1	0.804

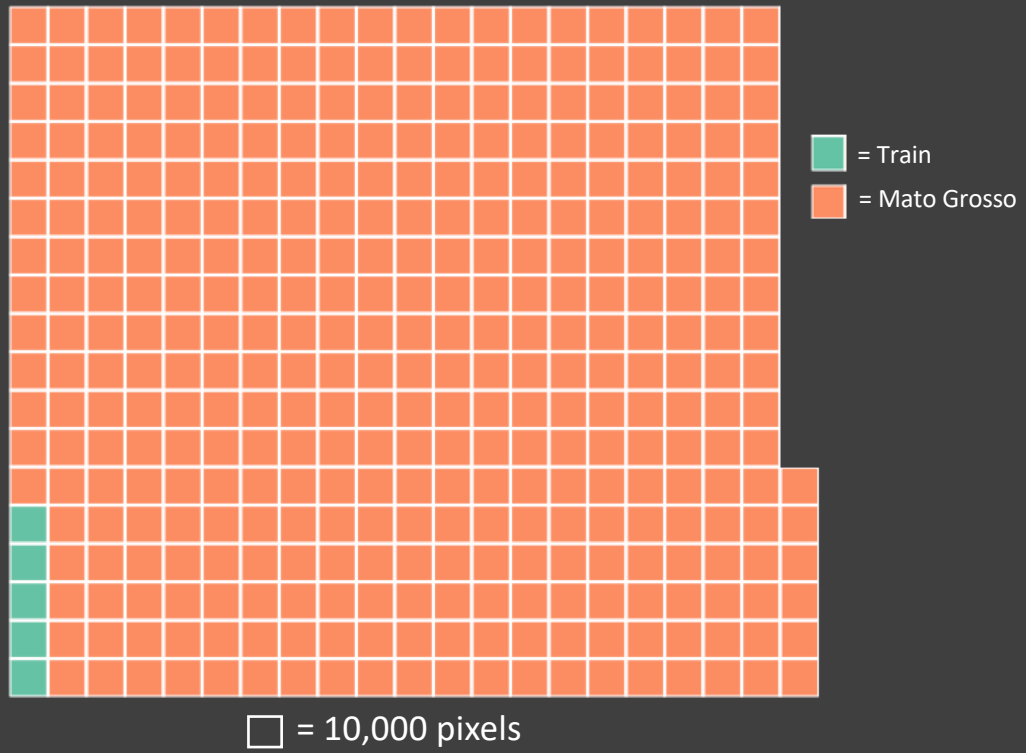


Mato Grosso

- Intense deforestation in Mato Grasso
  - Area: 348,788 mi² or 903,357 km²

Training vs Scoring

- Training:
  - 52,956 pixels
- Mato Grosso:
  - 3,613,428 pixels
  - 61,428,276 pixels over 17 years



Scoring the Model

- Score 17 years
  - Enables Post Classification Change Detection
- Compute time: 6-7 hours per year
  - Spark Standalone cluster: 6 workers, 48 cores, 24 GB per executor
- Write GeoTIFF

```
import geotrellis.raster.io.geotiff._

val raster = scoredRf.toRaster($"pForest", 1800, 1440)
GeoTiff(raster).write("scored.tiff")
```

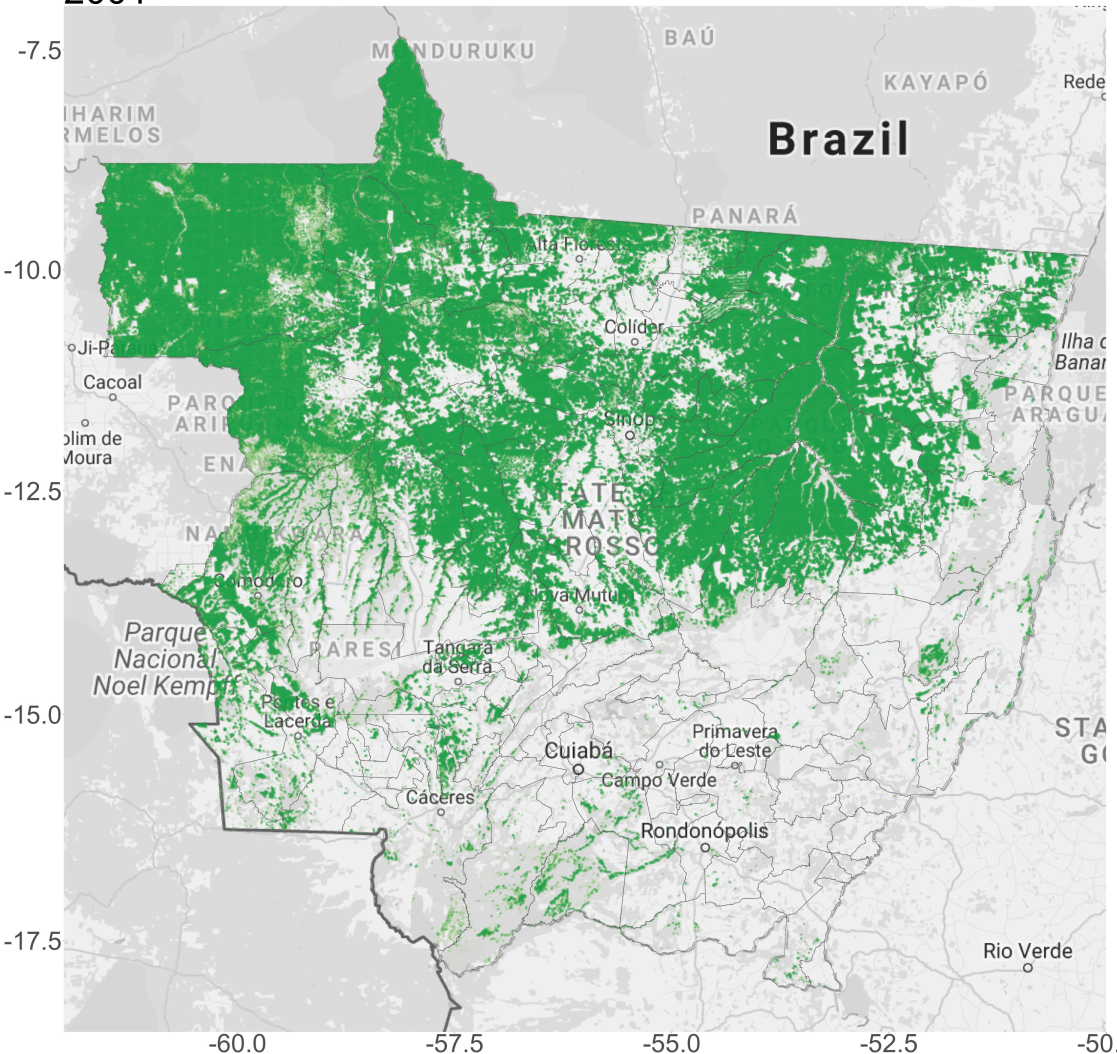
astræa

Results

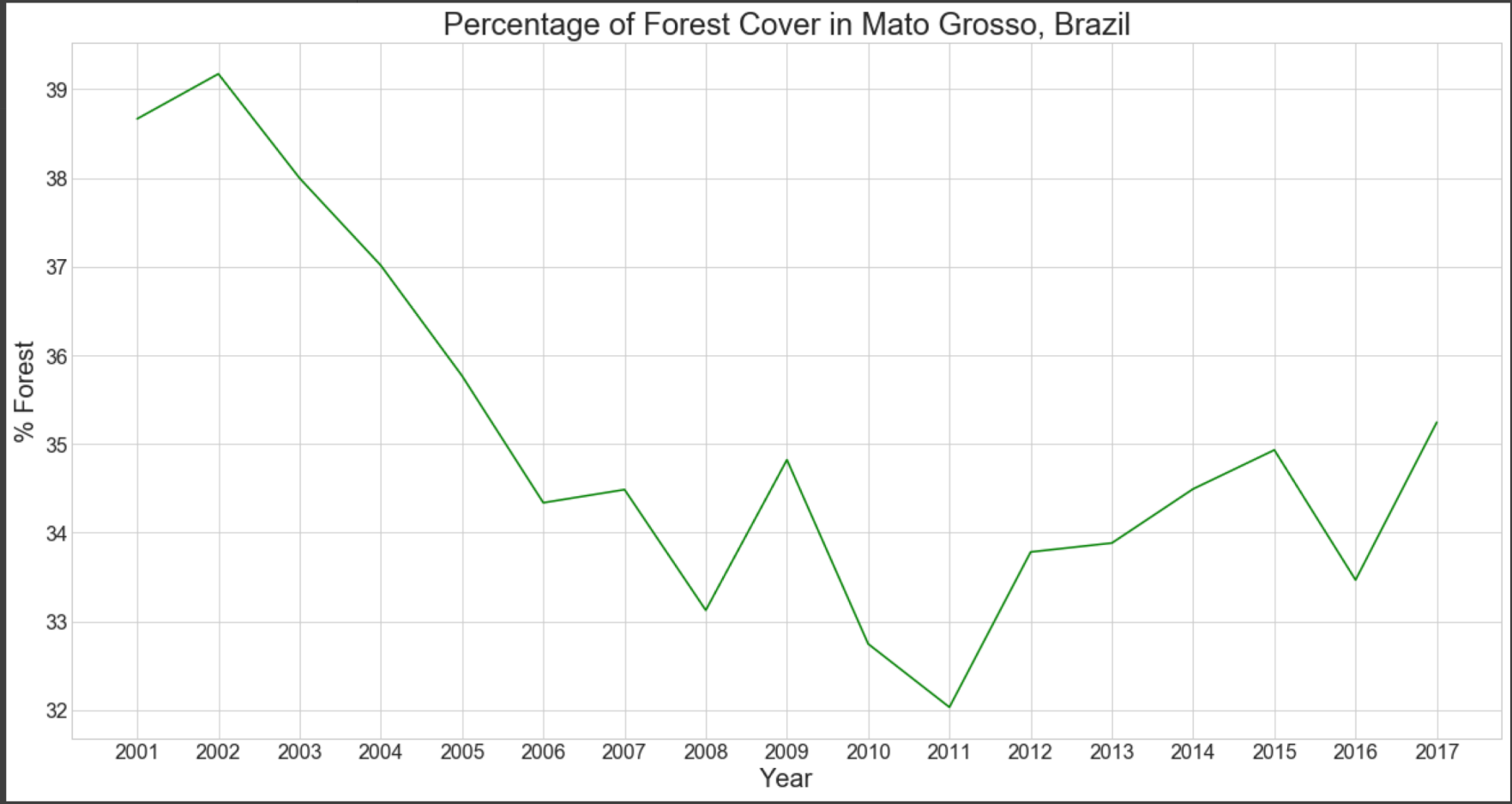
2001-Jul-30 Landsat 7

2014-Aug-11 Landsat 8

Full size interactive at http://forest.astraea.earth/



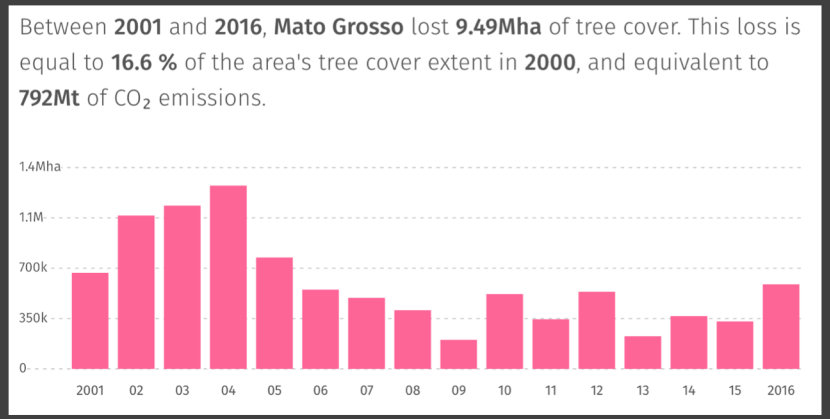
Full size interactive at
http://forest.astraea.earth/

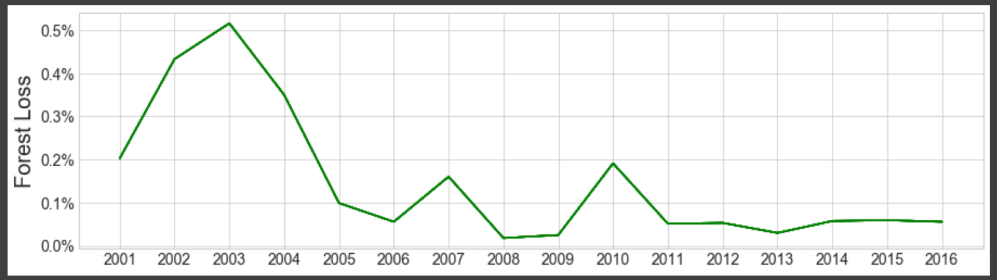

Percentage of Forest Cover in Mato Grosso, Brazil
% Forest
39
38
37
36
35
34
33
32
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year

Forest Loss Comparison

GLOBAL FOREST WATCH globalforestwatch.org

astræa Forest Model





Source: Tree cover loss: Hansen/UMD/Google/USGS/NASA

0.88 correlation

astræa

Conclusion

Next Steps

Improving model:

- Add more features
- Higher spatial resolution
- Better handling of seasonal differences across world

Applying model:

- Shorten time window for change detection
- Score the whole world

Questions?

Come visit Astraea at booth 14!

astraea.earth

rasterframes.io

forest.astraea.earth

Happy hour 7pm in the Lindbergh Room

Appendix

NDVI

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$$

astræa

Tile table with colors but no label

spatial_key	temporal_key	timestamp	bounds	MODIS_MCD43A4_nbar_band1	MODIS_MCD43A4_nbar_band3	MODIS_MCD43A4_nbar_band4
[50,40]	[1501632000000]	2017-08-02 00:00:00.0	POLYGON ((-6115727.858715877 -1389938.1497081555 , -6115727.858715877 ...			
[49,40]	[1501632000000]	2017-08-02 00:00:00.0	POLYGON ((-6393715.488657508 -1389938.1497081555 , -6393715.488657508 ...			
[50,39]	[1501632000000]	2017-08-02 00:00:00.0	POLYGON ((-6115727.858715877 -1111950.5197665244 , -6115727.858715877 ...			

IGBP Land Cover Classification System

Class	Class name	Description
1	Evergreen needleleaf forests	Lands dominated by needleleaf woody vegetation with a percent cover >60% and height exceeding 2 m. Almost all trees remain green all year. Canopy is never without green foliage.
2	Evergreen broadleaf forests	Lands dominated by broadleaf woody vegetation with a percent cover >60% and height exceeding 2 m. Almost all trees and shrubs remain green year round. Canopy is never without green foliage.
3	Deciduous needleleaf forests	Lands dominated by woody vegetation with a percent cover >60% and height exceeding 2 m. Consists of seasonal needleleaf tree communities with an annual cycle of leaf-on and leaf-off periods.
4	Deciduous broadleaf forests	Lands dominated by woody vegetation with a percent cover >60% and height exceeding 2 m. Consists of broadleaf tree communities with an annual cycle of leaf-on and leaf-off periods.
5	Mixed forests	Lands dominated by trees with a percent cover >60% and height exceeding 2 m. Consists of tree communities with interspersed mixtures or mosaics of the other four forest types. None of the forest types exceeds 60% of landscape.
6	Closed shrublands	Lands with woody vegetation less than 2 m tall and with shrub canopy cover >60%. The shrub foliage can be either evergreen or deciduous.
7	Open shrublands	Lands with woody vegetation less than 2 m tall and with shrub canopy cover between 10% and 60%. The shrub foliage can be either evergreen or deciduous.
8	Woody savannas	Lands with herbaceous and other understory systems, and with forest canopy cover between 30% and 60%. The forest cover height exceeds 2 m.
9	Savannas	Lands with herbaceous and other understory systems, and with forest canopy cover between 10% and 30%. The forest cover height exceeds 2 m.
10	Grasslands	Lands with herbaceous types of cover. Tree and shrub cover is less than 10%.
11	Permanent wetlands	Lands with a permanent mixture of water and herbaceous or woody vegetation. The vegetation can be present either in salt, brackish, or fresh water
12	Croplands	Lands covered with temporary crops followed by harvest and a bare soil period (e.g., single and multiple cropping systems). Note that perennial woody crops will be classified as the appropriate forest or shrub land cover type
13	Urban and built-up lands	Land covered by buildings and other man-made structures.
14	Cropland/natural vegetation mosaics	Lands with a mosaic of croplands, forests, shrubland, and grasslands in which no one component comprises more than 60% of the landscape.
15	Snow and ice	Lands under snow/ice cover throughout the year.
16	Barren	Lands with exposed soil, sand, rocks, or snow and never have more than 10% vegetated cover during any time of the year.
17	Water bodies	Oceans, seas, lakes, reservoirs, and rivers. Can be either fresh or saltwater bodies.