

# Why Is It So Hard to Scale Development Data Projects? *(Answer: Build a Data Lake First)*

Jonathan Hersh (Chapman)

*Private Sector Development Research Network Workshop on Firm-level Data*

4/6/2021

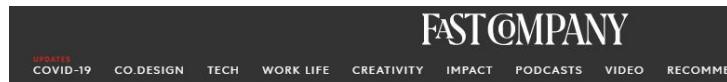
# About: Jonathan Hersh, PhD

- Assistant Professor Economics and Management Science Chapman University
- PhD in economics, Boston University
- **Research Fields:**
  - Applications of artificial intelligence (computer vision)
  - Economics of information systems
  - Development economics
  - Digitization strategy
- **Teaching Fields:**
  - Machine learning
  - Applications of artificial intelligence



# Formerly: Data Scientist @ World Bank

- Satellite Imagery + Computer Vision + Machine Learning
- Advised World Bank/IDB on COVID poverty transfers in Belize, Togo, Guinea



## How Satellite Data And Artificial Intelligence Could Help Us Understand Poverty Better

New technology lets computers understand what they see in an image—or a million images.



[PHOTO: FLICKR USER RODRIGO CARVALHO]

BY MAYA CRAIG 3 MINUTE READ

Data analytics firm Orbital Insight is partnering with the World Bank to [test technology that could help measure global poverty](#) using satellite imagery and artificial intelligence.

Bloomberg

Economics

## Poverty Surveyors in Sri Lanka Get Some Help From Satellites Orbiting the Earth

The World Bank is teaming with a Silicon Valley startup to test whether poverty can be measured using satellite images.

By Adam Satariano

November 6, 2015, 7:00 AM PST Updated on November 6, 2015, 1:57 PM PST

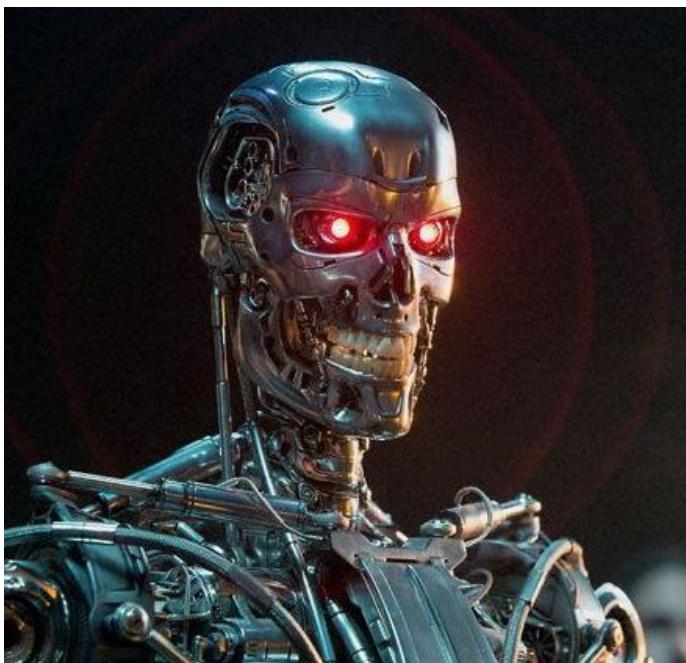
In mountainous areas of Pakistan or far-flung villages in Sri Lanka, finding reliable economic information is extremely difficult. The World Bank's solution has been to send surveyors to study the conditions on the ground, which is an expensive, time-consuming, and imprecise task. The resulting dearth of data leaves governments, aid groups, and researchers unsure of where to put resources that can be critical to helping the world's most impoverished areas.

# Why Do Development Data Projects Fail?

1. Models can only be as good as your data pipeline (Bad data = bad model)
2. Build a consistent data lake first (excel files are bad)
3. All data and models should be an API (i.e. Application Programming Interface or structured way to access data)

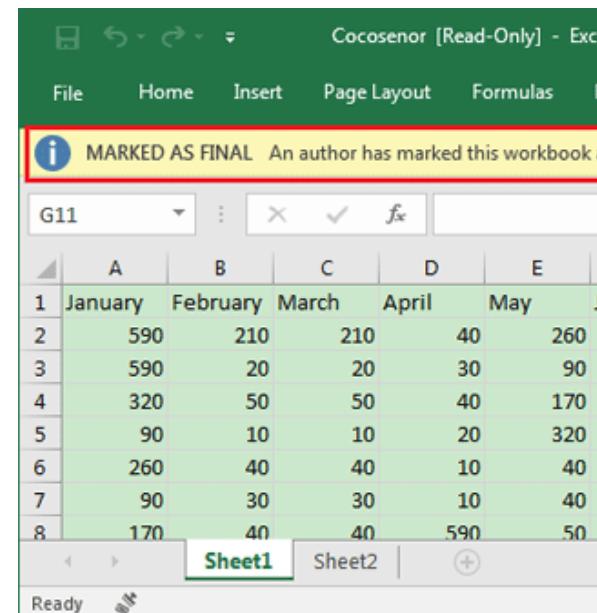


# Perceptions of capabilities of Artificial Intelligence



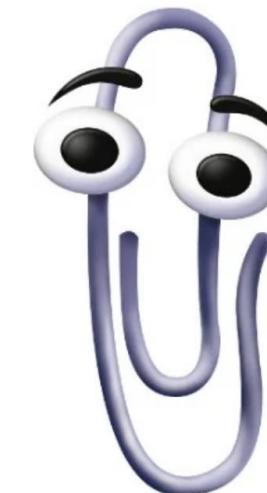
+

# Existing data infrastructure



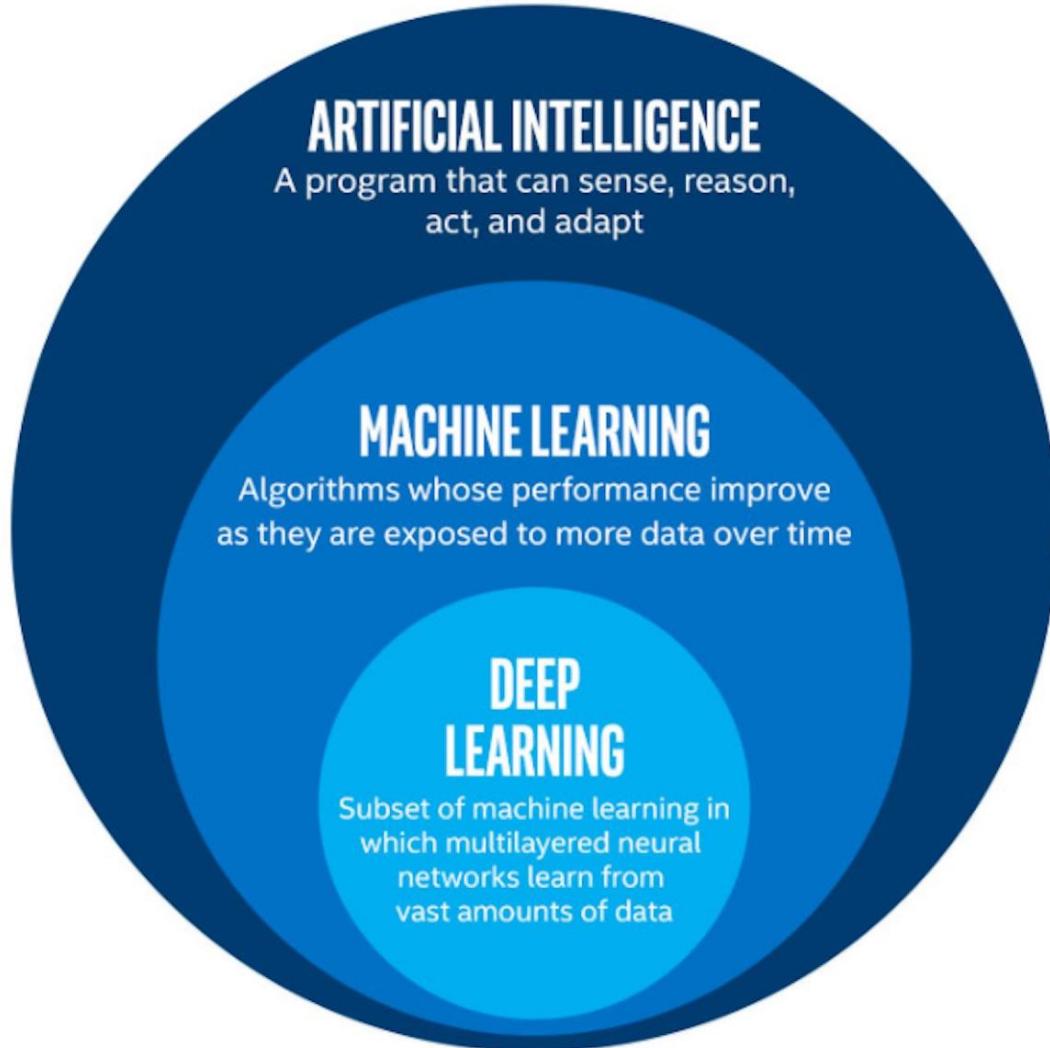
	A	B	C	D	E	J
1	January	February	March	April	May	
2	590	210	210	40	260	
3	590	20	20	30	90	
4	320	50	50	40	170	
5	90	10	10	20	320	
6	260	40	40	10	40	
7	90	30	30	10	40	
8	170	40	40	590	50	

=



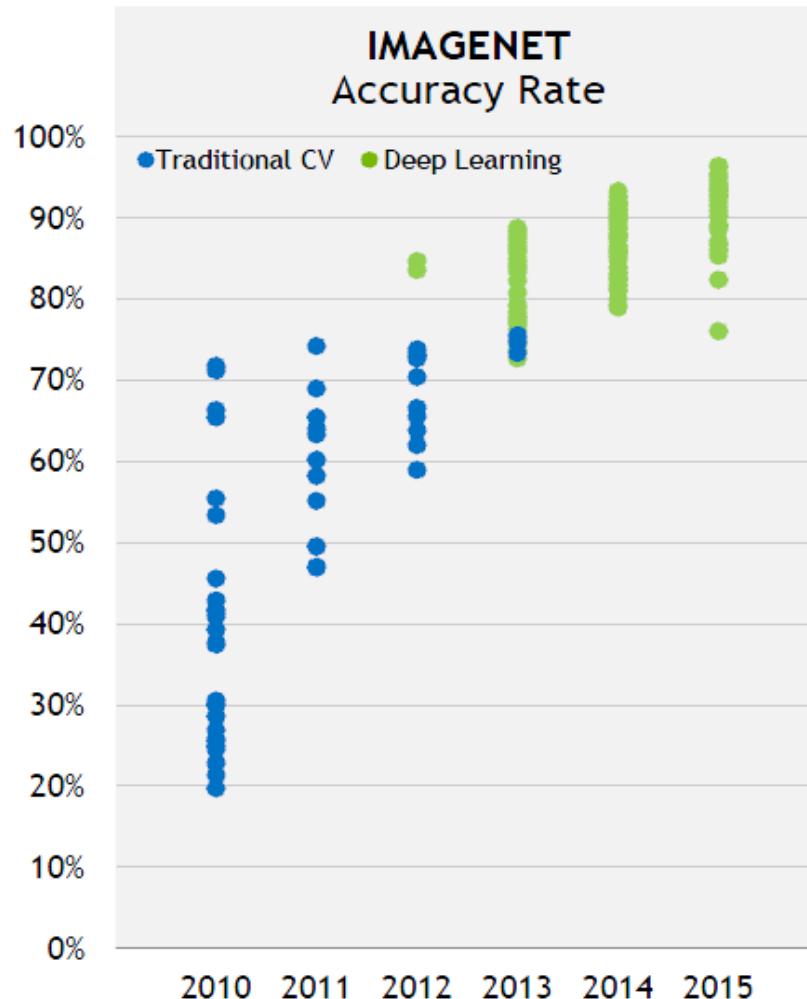
Sometimes I just popup for no reason at all. Like now.

# Artificial Intelligence vs Machine Learning vs Deep Learning



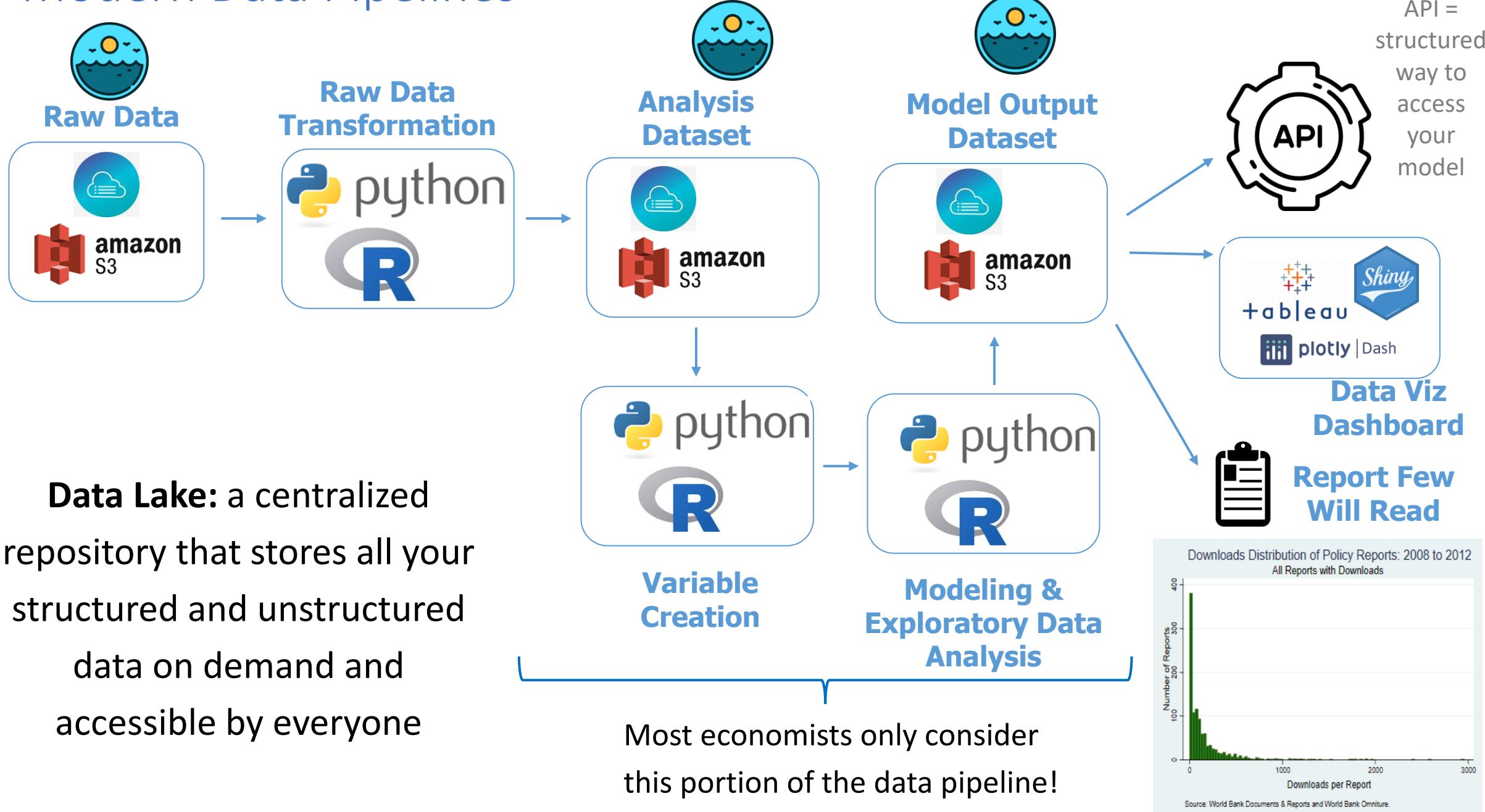
**“The deep in deep learning isn’t a reference to any kind of deeper understanding achieved by the approach; rather, it stands for this idea of successive layers of representations” – Francois Chollet**

# One Reason for The Deep Learning Hype: Images

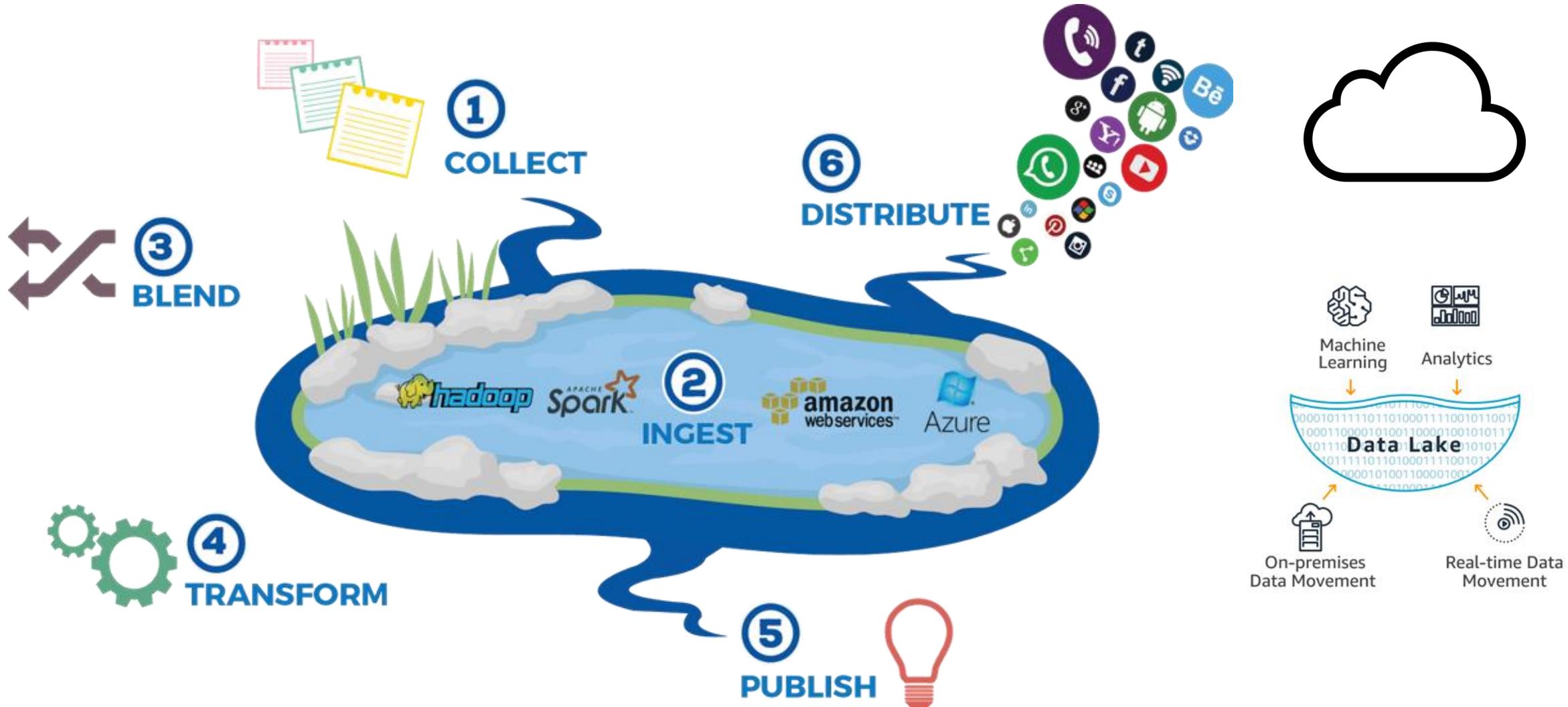


- Before deep learning, computers couldn't recognize objects from images.
- In 2012 researchers first used a deep learning model on the ImageNet competition, which tests an AI to recognize objects in images
- But ImageNet had 1M+ images!
- Translation: models and data are close complements

# Modern Data Pipelines

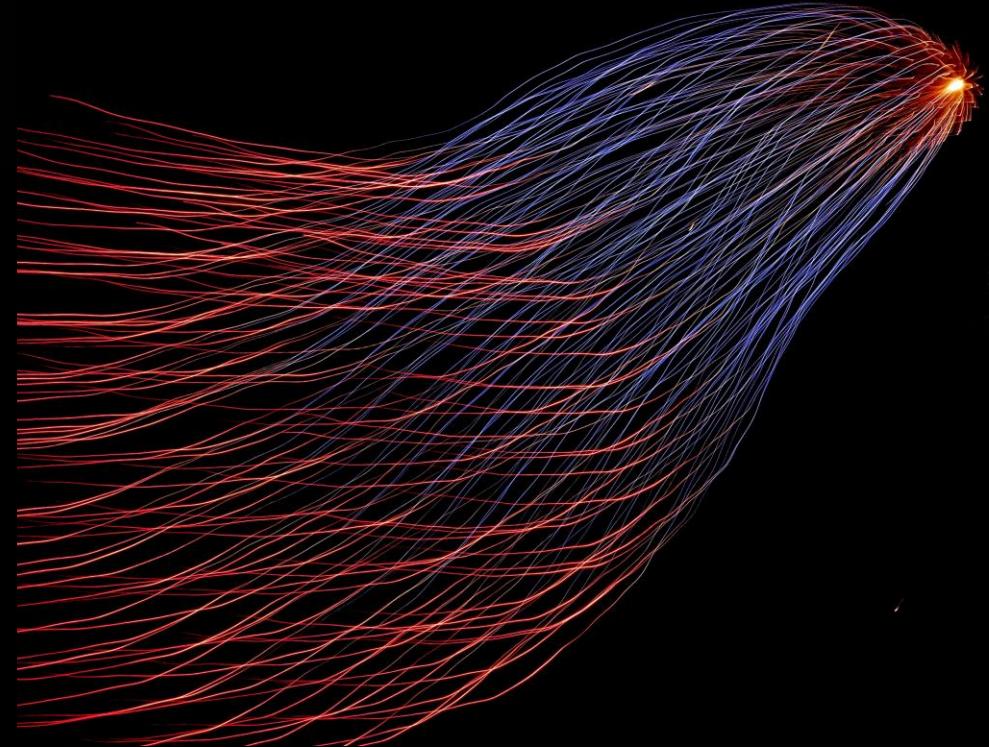


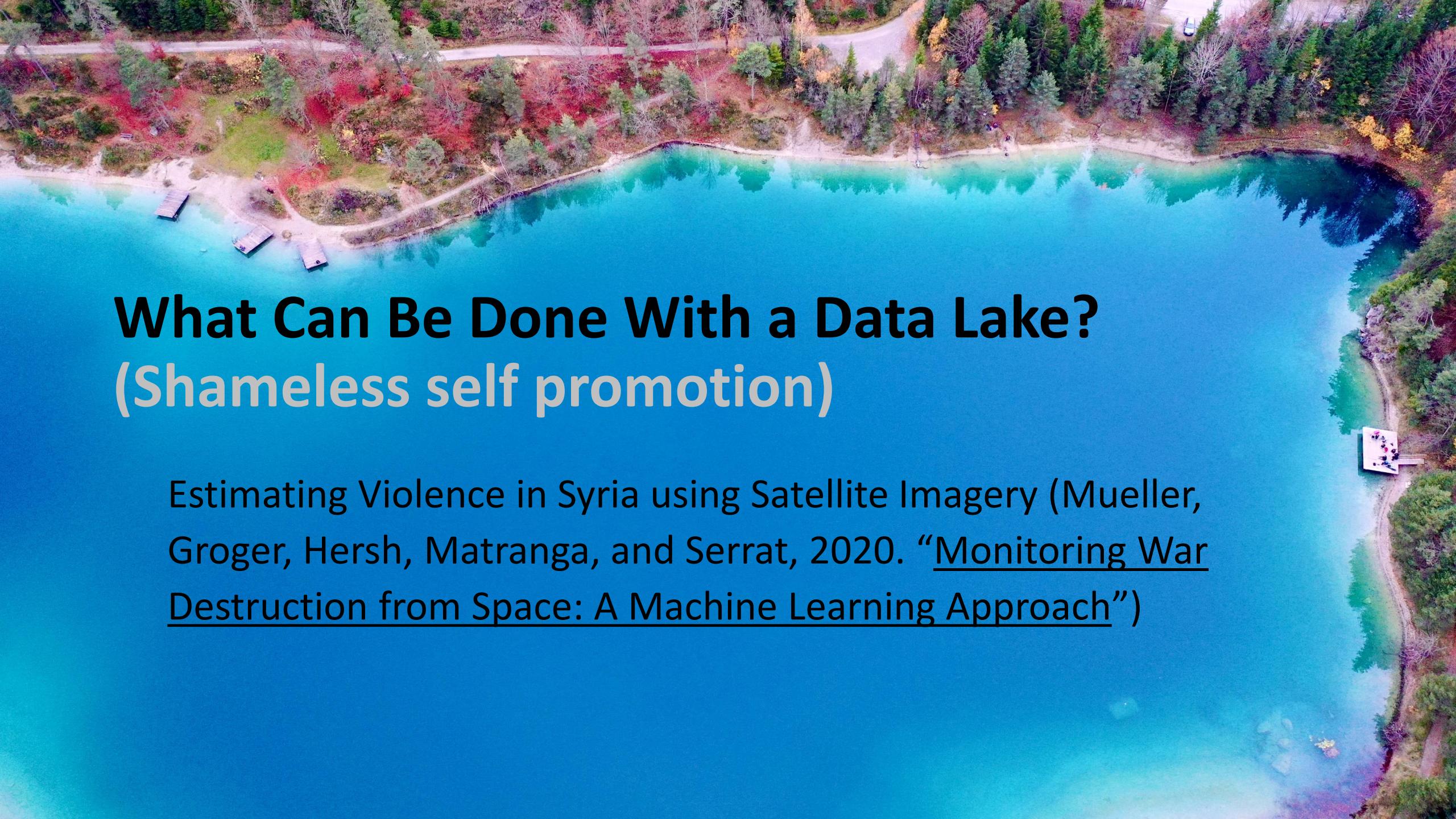
# Many Options of Data Lakes to Choose From



# Questions at This Point

- What's best data lake?
  - What's the best food when you're starving?
  - Cloud for sure, open source if you can
- What programming languages do I need to learn?
  - None really. Version control (git). R and Python will make your life easier. SQL variants (or just [dbplyr](#))
- **Theme: Have a consistent data strategy and ensure everyone abides by it**



The background of the slide is a high-angle aerial photograph of a lake. The water is a vibrant turquoise color, reflecting the surrounding environment. The shoreline is lined with trees displaying autumn foliage in shades of red, orange, and yellow. A paved path or road follows the curve of the lake. In the top right corner, a small white boat is visible on the water.

# What Can Be Done With a Data Lake? (Shameless self promotion)

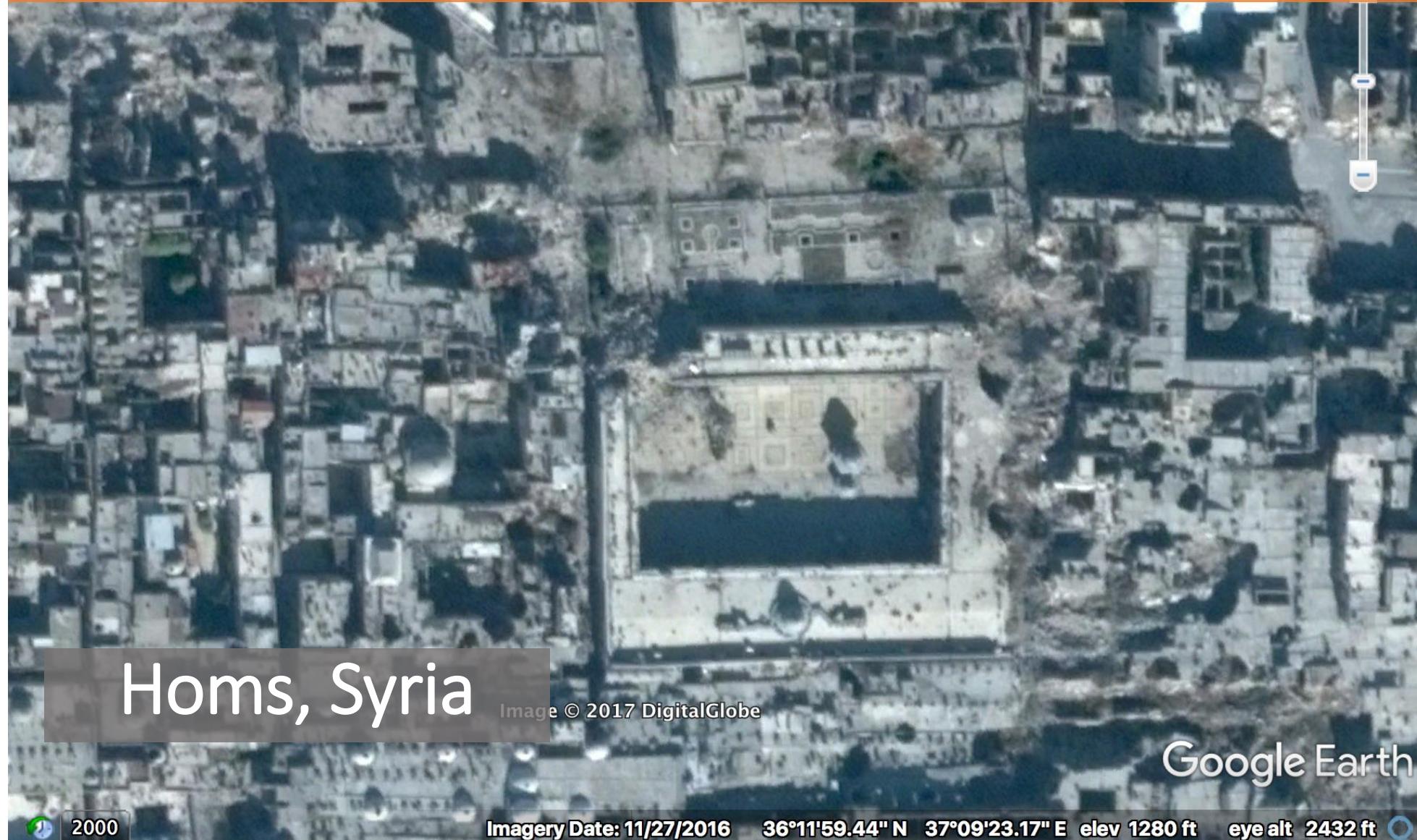
Estimating Violence in Syria using Satellite Imagery (Mueller, Groger, Hersh, Matranga, and Serrat, 2020. “[Monitoring War Destruction from Space: A Machine Learning Approach](#)”)

# Information on ongoing violence may be biased, outdated, and imprecise

## Syria



# Detecting Conflict from Satellite Imagery



Homs, Syria

Image © 2017 DigitalGlobe

Google Earth

2000

Imagery Date: 11/27/2016 36°11'59.44" N 37°09'23.17" E elev 1280 ft

eye alt 2432 ft

# The Dream: Teach a Computer to Spot Building Destruction Automatically



Homs, Syria

Image © 2017 DigitalGlobe

Google Earth

2000

Imagery Date: 11/27/2016 36°11'59.44" N 37°09'23.17" E elev 1280 ft

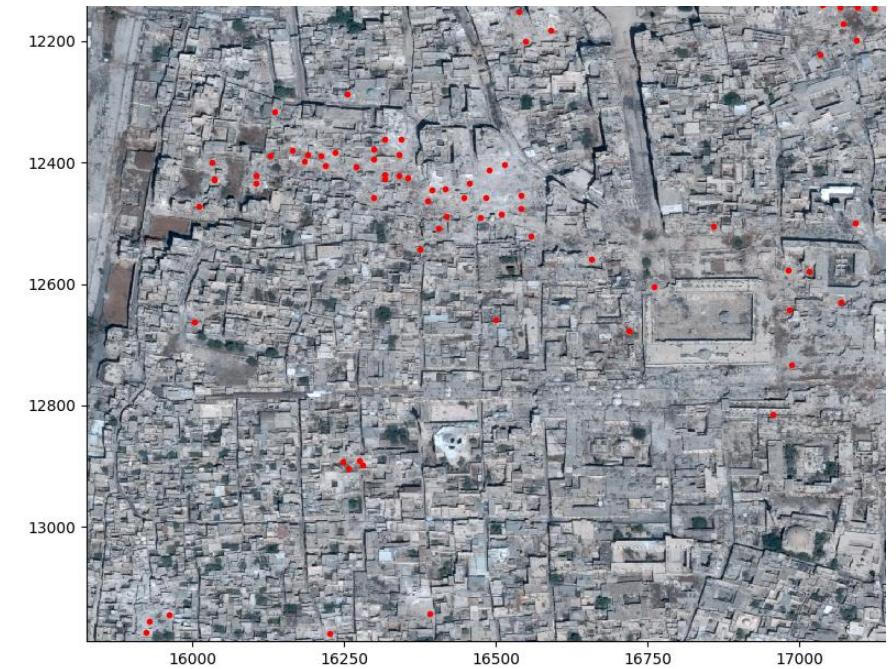
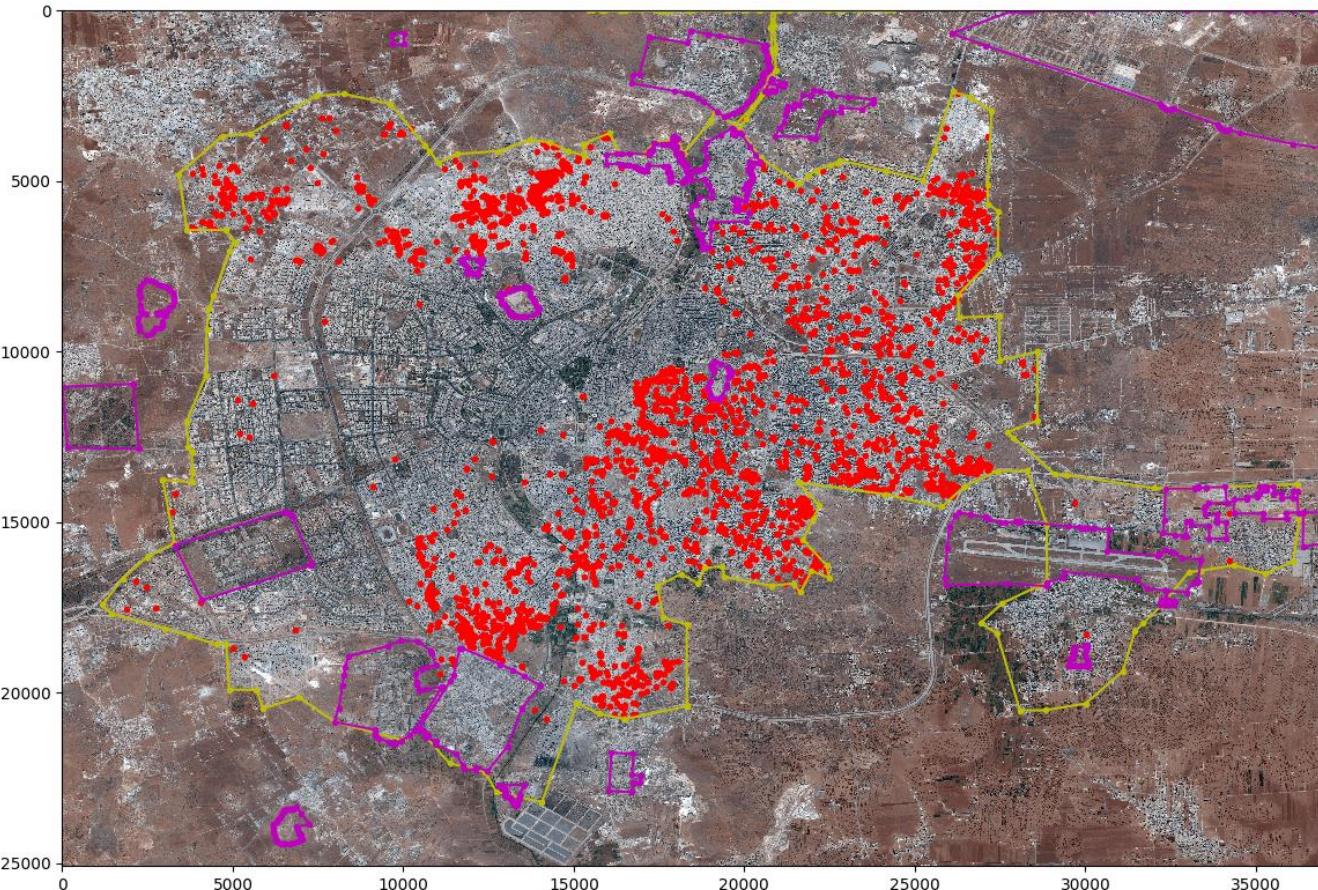
eye alt 2432 ft

# This Paper:

- Train CNN deep learning architecture to recognize building destruction from Google Maps imagery
- Training data: UNOSAT hand-labeled building destruction
  - Novel data augmentation approach to expand training labels to 2.2m
- Novel second machine learning stage that uses temporal and contextual information to increase precision due to imbalanced data
- Validate using external bombing event data



# Why is this hard? Destruction is Sparse Relative to Non-Destruction



Only 2.3% of images show any destruction in our sample of six Syrian cities

# Illustration of Unbalanced Data Problem

- **Accuracy:**  $\frac{TP + TN}{TP + FN + FP + TN}$
- **Precision** (share of positives predicted correctly):  $\frac{TP}{TP + FP}$
- **True Positive Rate / Recall** (share of actual positives predicted that are correct):  $\frac{TP}{TP + FN}$
- Suppose you have 100,000 images, but only 1000 (1%) are destroyed
- 12% FPR means your model produces:  $99,000 [FP + TN] * 0.12[FPR] = 11,880$  FPs
- 80% TPR/recall means your model produces  $1000 * 0.8$  TP = 800 TPs
- Precision for that model is:  $\frac{TP}{TP + FP} = \frac{800}{11880 + 800} = 0.063$
- Probability of being correct if you find destruction is only 6.3%!

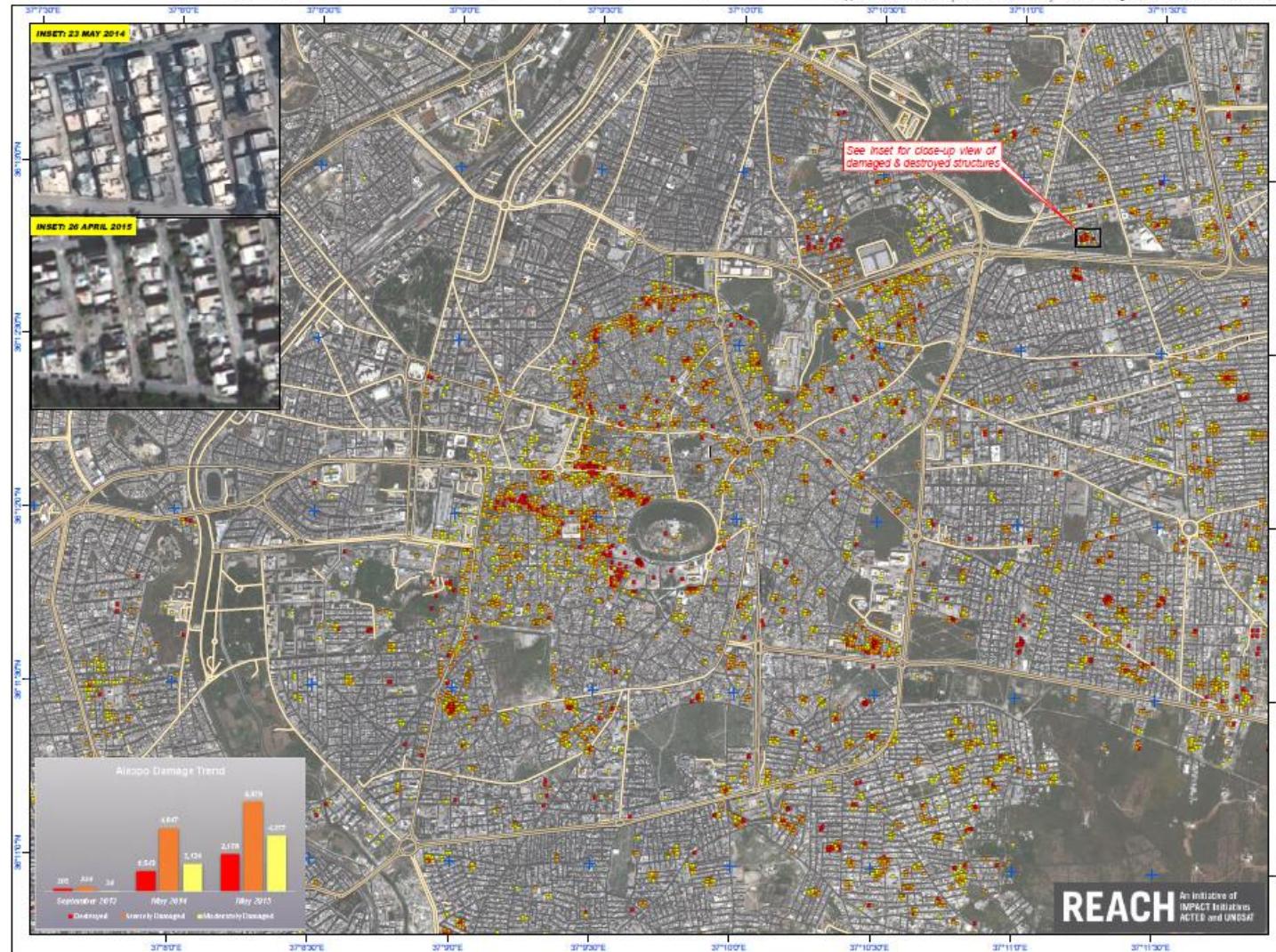
		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative



# Ground Truth: Building Destruction Annotations from UNOSAT

## DAMAGE ASSESSMENT OF ALEPOO, ALEPPO GOVERNORATE, SYRIA

Analysis with Pleiades Data Acquired 01 May 2015, 26 April 2015 and WorldView-2 Data Acquired 23 May 2014, 23 September 2013, and 21 November 2010



This map illustrates satellite-detected damage in a portion of the city of Aleppo, Syria. The analysis was based on satellite imagery acquired 01 May 2015, 26 April 2015, 23 May 2014, 23 September 2013, and 21 November 2010. UNTAR - UNOSAT identified a total of 6,177 affected structures within the extent of the map. Among them, 2,040 structures were destroyed, 2,641 severely damaged and 1,496 moderately damaged. The city-wide analysis of Aleppo revealed a total of 14,034 affected structures, of which 2,079 were destroyed, 6,079 severely damaged and 5,876 moderately damaged. Much of the city was damaged by 23 May 2014. 6,887 structures were newly damaged and 90 structures experienced an increase in damage between that date and 01 May 2015. This analysis does not include areas outside the city. This is a preliminary analysis and has not yet been validated in the field. Please send ground feedback to UNTAR - UNOSAT.

Complex Emergency  
7/10/2015

Version 1.0

Activation Number:  
CE2013004SYR



**LEGEND**

- Destroyed
- Severely Damaged
- Moderately Damaged
- Highway / Primary Road
- Secondary Road
- Local / Urban Road

Map Scale for A3: 1:20,000  
Meters

Satellite Data (1): Pleiades  
Acquisition Date: 01 May 2015 & 26 April 2015  
Resolution: 60 cm  
Copyright: © CNES (2015), Distribution AIRBUS DS

Source: Airbus Defense and Space  
Source: CNES (2015), Distribution AIRBUS DS

Imagery Date: 23 May 2014, 23 September 2013 & 21 November 2010

Resolution: 60 cm  
Copyright: European Space Imaging

Road Data: Google Map Maker / OSM / ESRI

Other Data: UNHCR, UNICEF, NASA, NOAA

Analysis: UNTAR - UNOSAT

Production: UNTAR / UNOSAT

Analysts conducted with ArcGIS v10.3

Coordinate System: WGS 1984 UTM Zone 37N

Projection: Transverse Mercator

Datum: WGS 1984

The depiction and use of boundaries, geographic names and related data shown here are not warranted or intended to do so. Implied political boundaries or positions expressed in this map do not imply official endorsement or acceptance by the United Nations. UNOSAT is a program of the United Nations Institute for Training and Research (UNITAR), providing satellite imagery and related geospatial information, research and analysis to UN organizations and development agencies and their implementing partners.

This work by UNTAR/UNOSAT is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

# Label Augmentation Method

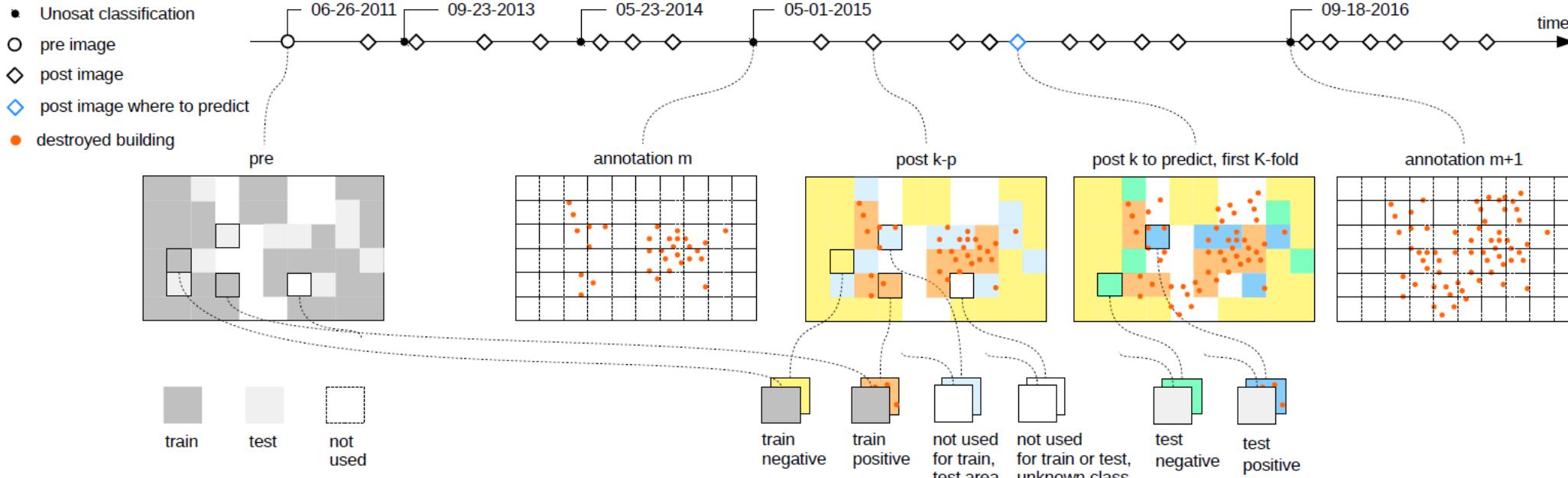
- To increase the number of training labels, we apply a novel data augmentation method based on assumptions of the data generating process
- **No reconstruction =>** label destroyed at time t, remain destroyed t + 1
- **No reconstruction =>** label not destroyed at t, is also not destroyed t - 1

T=1	T = 2	T = 3	T=4	
				 Label
				 Augmented label

# Images By City Over Time

City	Total images	Temporal periods (post period)	Labeled images (with augmentation)	Share of images with destruction
Aleppo	2,106,412	22	1,626,920	1.82%
Daraa	202,462	13	125,231	1.00%
Deir-Ez-Zor	98,602	7	84,723	2.86%
Hama	285,057	9	224,365	3.73%
Homs	200,035	5	83,941	8.26%
Raqqa	180,184	8	112,481	2.14%
All	3,072,752	64	2,257,661	2.37%

# Sampling for Testing and Training



- Four temporal periods with annotations (at patch level)
- Training and testing samples are from distinct spatial areas

# Neural Network Architecture

- We use a Convolutional Neural Network (CNN), which is a series of neural network filters, where the filters have been optimized for the prediction task
- We tried many architectures. Standard (ResNet, VGG16), and boutique (U-Net)
- In the end we use a simple but flexible 2 convolutional layer network
- We use a random search algorithm to sample hyperparameters:
  - Number of convolutional layers
  - Number of neurons of the fully connected layer
  - Filters and kernels of the convolutional layer and activation functions (relu vs sigmoid)
  - Pooling size and max pooling layer
  - Varying dropout
  - Different epochs, batch sizes, class weight

# Second Stage Machine Learning Smoothing

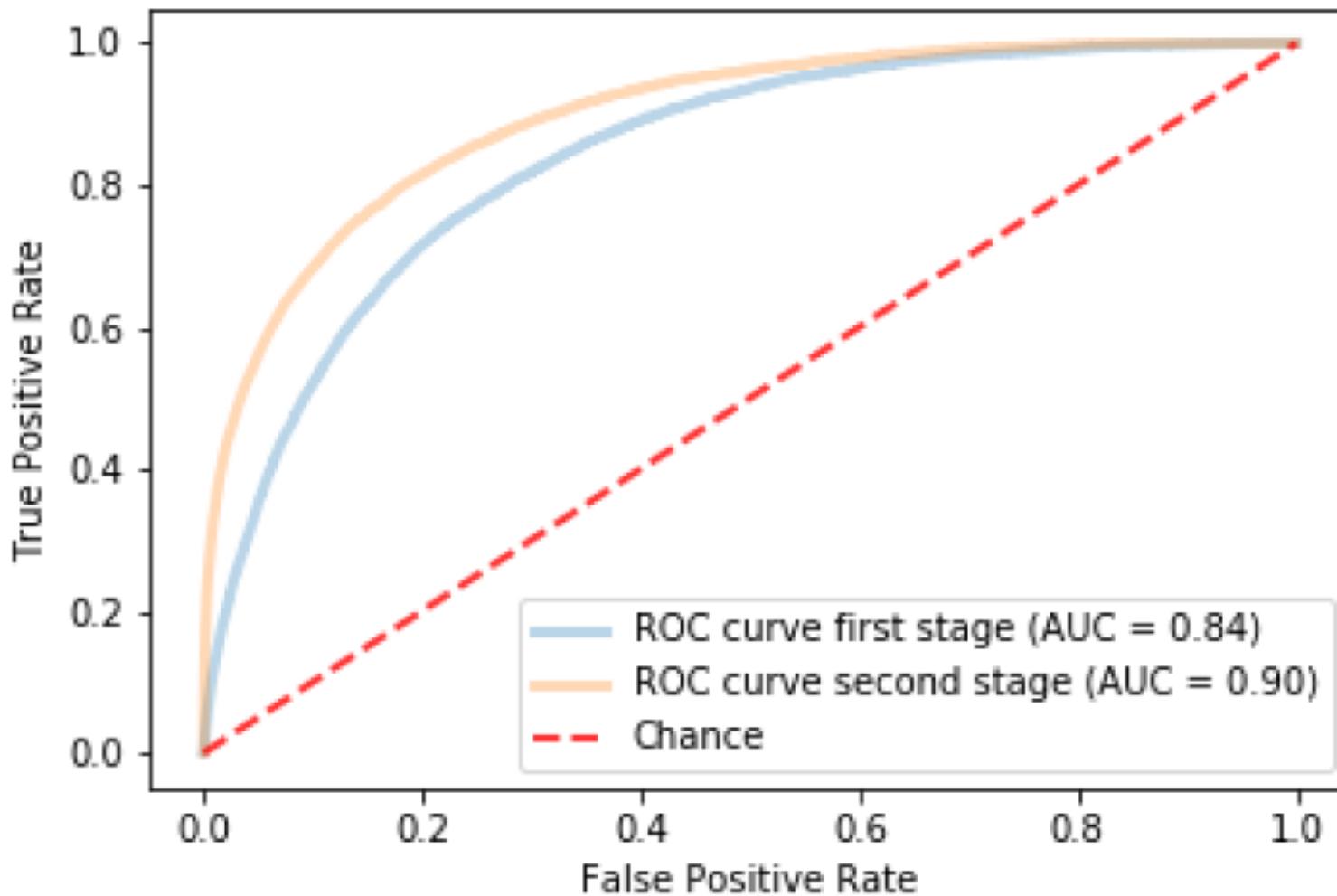
- Idea: destruction is correlated in space and time
- We train a random forest model on the CNN predictions using two spatial lags and two temporal lags
- This stage separated from CNN stage for maximum flexibility and modularity

Table 2. Model precision varying second stage

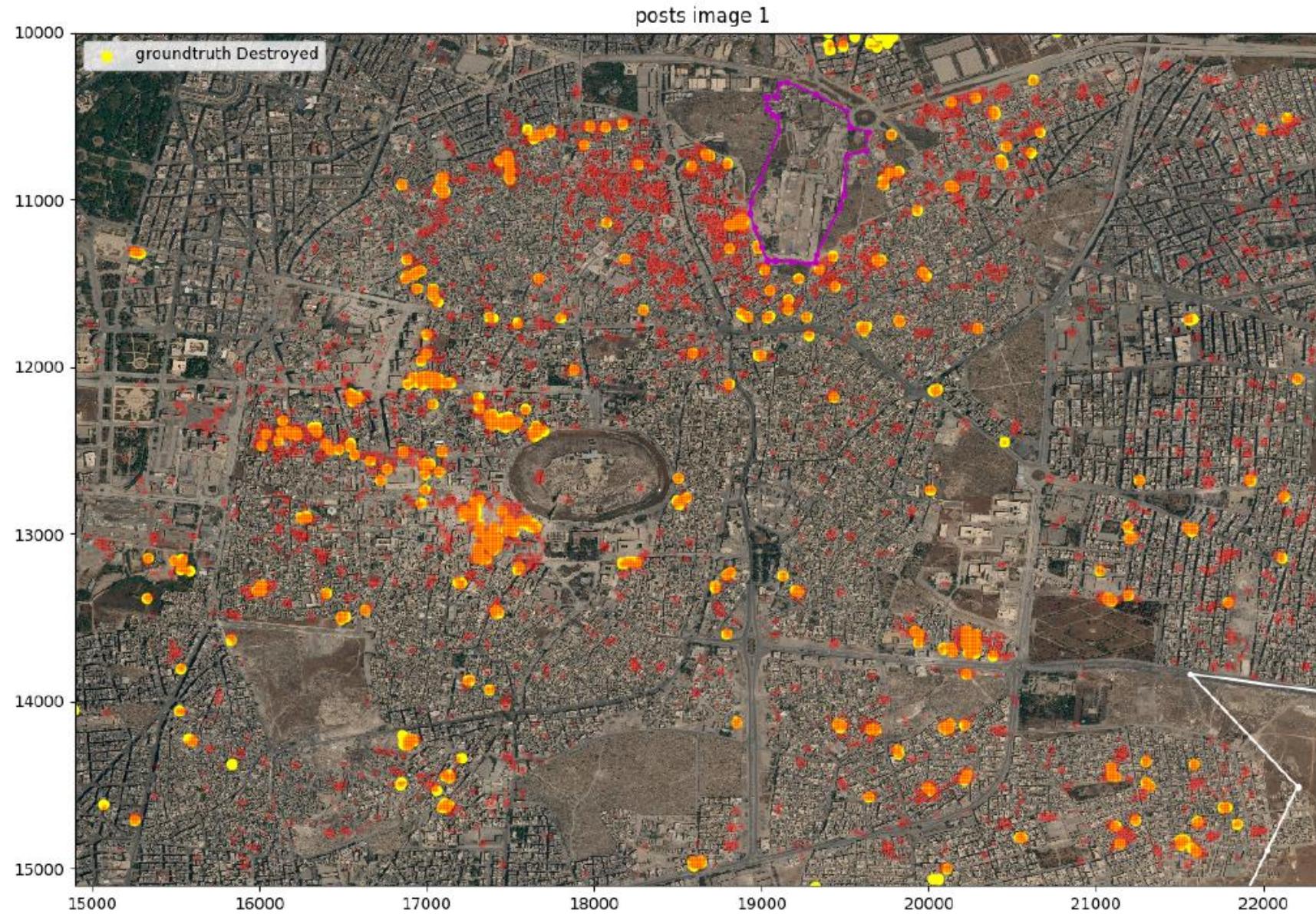
City	(1)	(2)	(3)	(4)
	First stage	CNN + Spatial	CNN + Spatial & 1 temporal lag/lead	CNN + Spatial & 2 temporal lag/leads
Aleppo	16.1%	16.8%	28.2%	35.7%
Daraa	4.2%	4.6%	9.5%	12.1%
Deir-Ez-Zor	11.0%	12.0%	18.6%	21.9%
Hama	54.5%	65.3%	67.5%	68.0%
Homs	25.8%	35.3%	44.6%	56.1%
Raqqa	12.8%	17.8%	20.9%	31.8%
All	24.5%	28.7%	37.4%	42.7%

Sources: Author calculations, UNITAR/UNOSAT damage annotations for Syria.

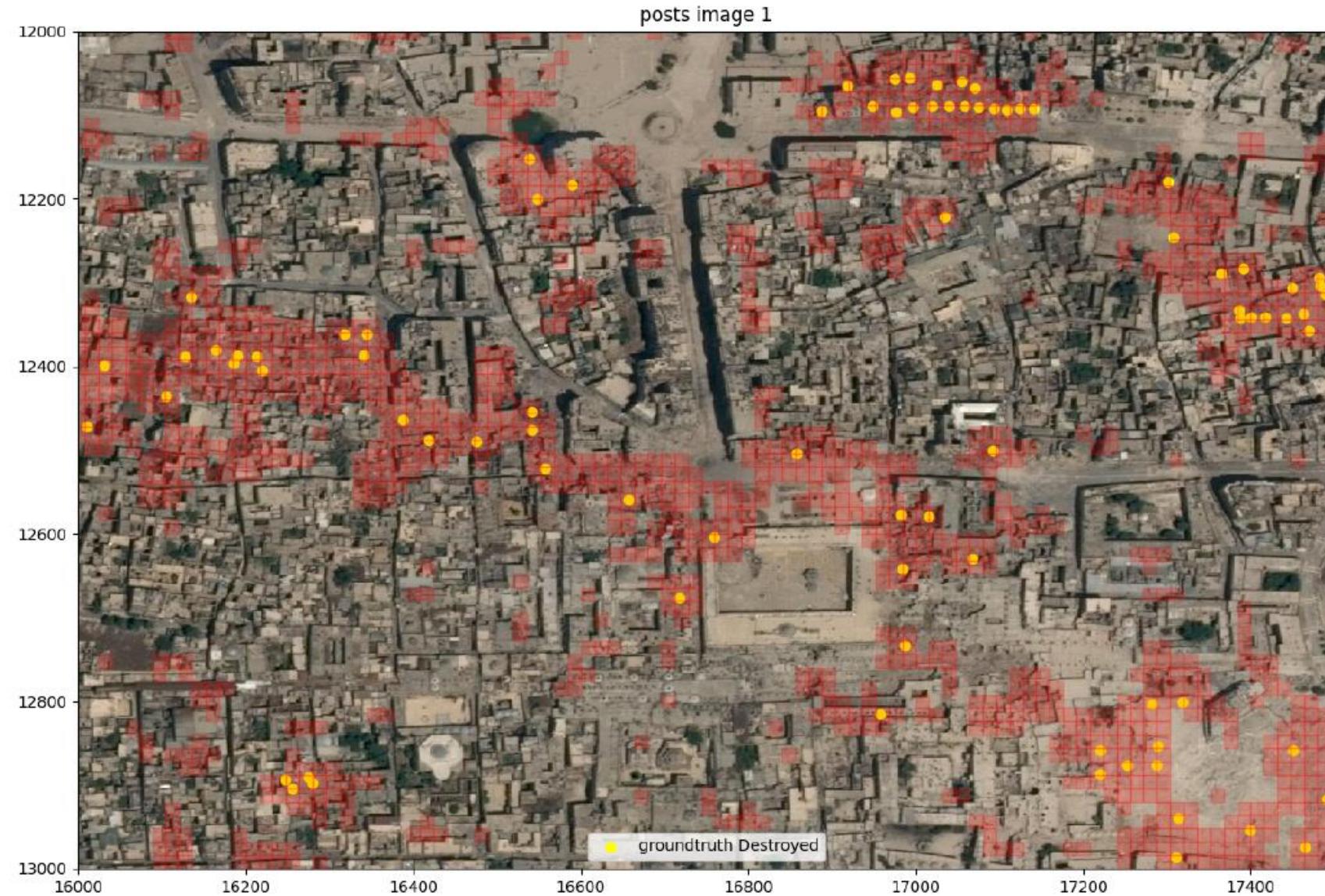
# Aleppo ROC Curve (Test Sample)



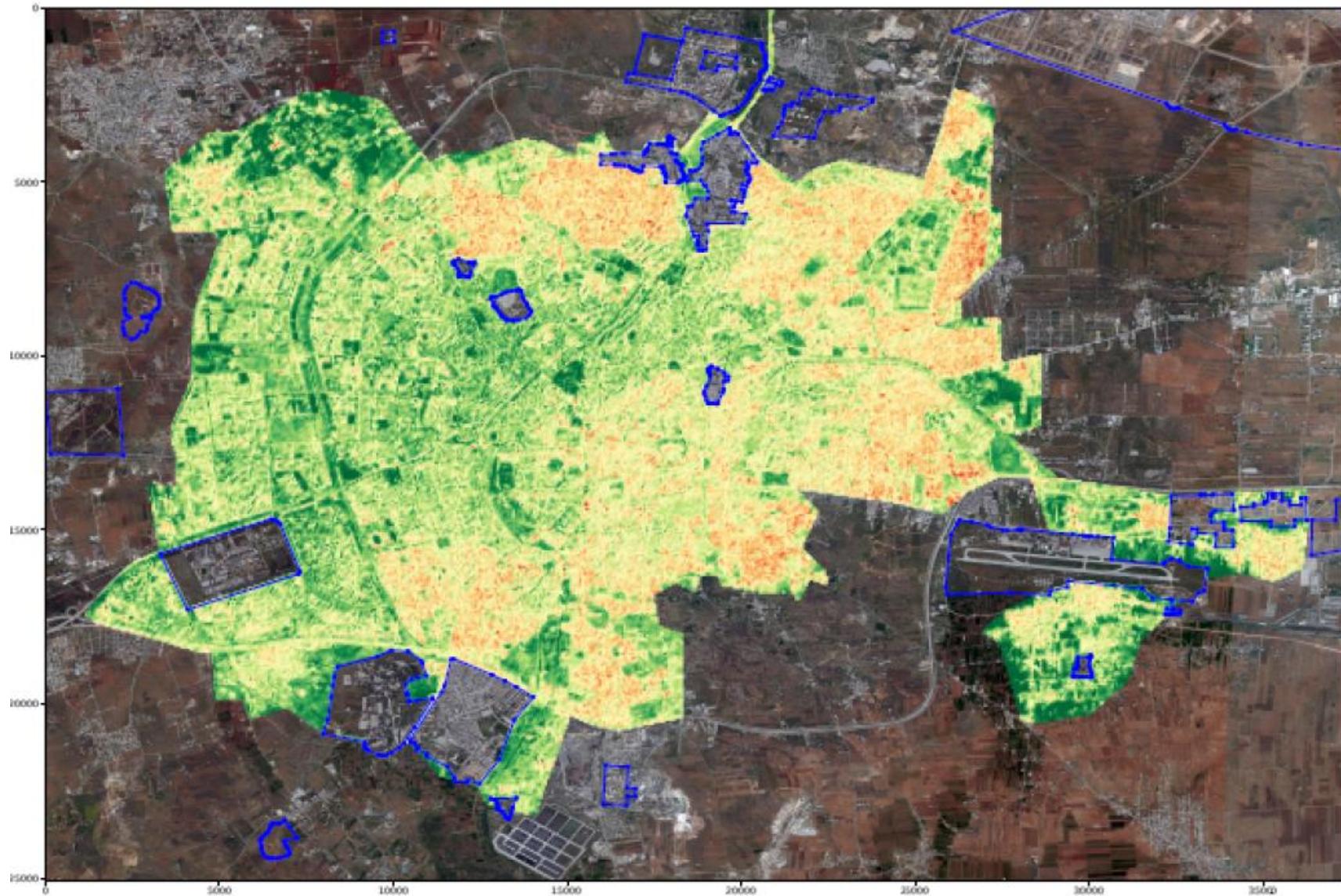
# Destruction Predictions Aleppo



# Destruction Predictions Aleppo



# Destruction Predictions Aleppo



# Conclusion

- We propose a novel data augmentation and second stage machine learning stage to increase precision of building destruction
- 2<sup>nd</sup> ML stage doubles precision, to 42%, good enough for automated detection
- Predictions of building destruction are validated using external bombing event data
- Next steps:
  - Other cities?
  - Severity of damage?

# Comments/suggestions appreciated!

Jonathan Hersh

Assistant Professor

Argyros School of Business, Chapman University

[hersh@chapman.edu](mailto:hersh@chapman.edu)