Get Your Hands Dirty: You Should Model with Unstructured Data

Jonathan Hersh (Chapman)

University of Michigan, School of Information

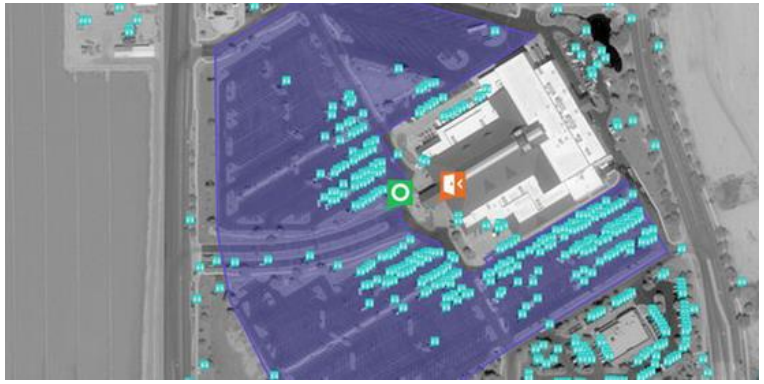
3/30/2021

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

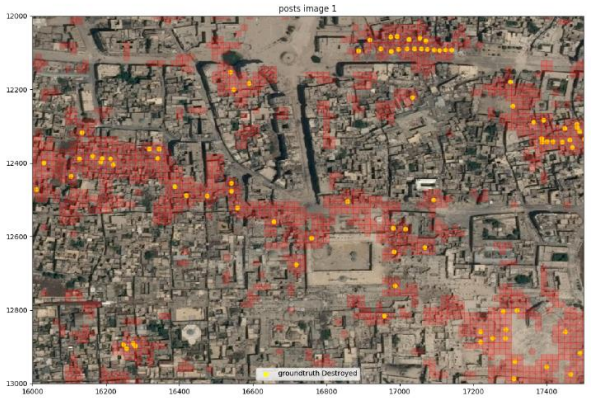
My Research

- Satellite Imagery + Computer Vision + Machine Learning



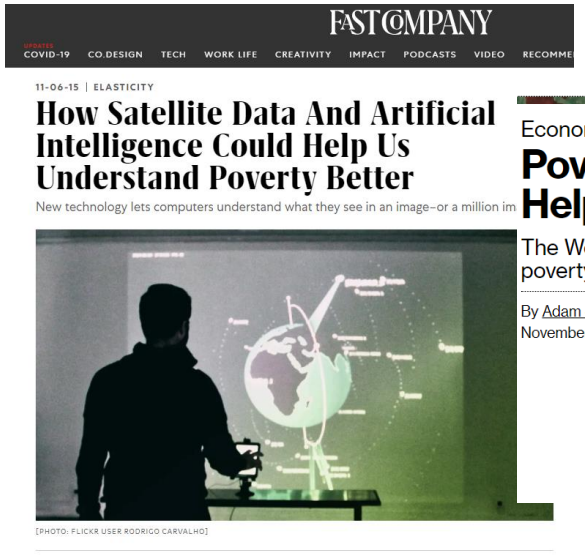
Count cars in parking lots!

Dense Prediction: Scanning Aleppo



Damaged buildings in Syria!

- Advised World Bank/IDB on COVID poverty transfers in Belize, Togo, Guinea


FAST COMPANY

COVID-19 CO.DESIGN TECH WORK LIFE CREATIVITY IMPACT PODCASTS VIDEO

11-06-15 | ELASTICITY

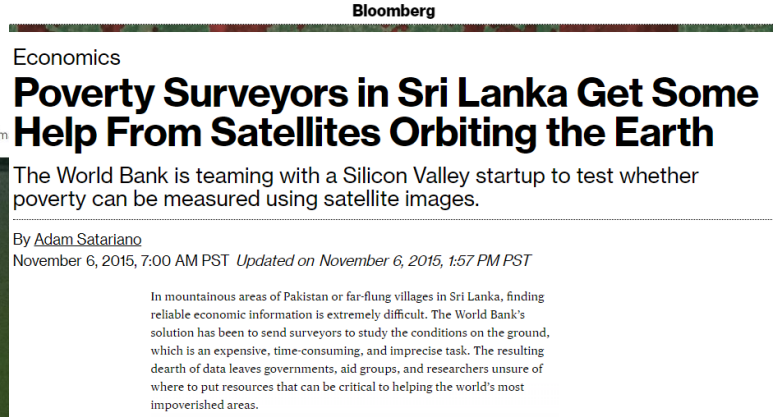
How Satellite Data And Artificial Intelligence Could Help Us Understand Poverty Better

New technology lets computers understand what they see in an image–or a million im

[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.

More “Business” Research

- Online Media Piracy

Forbes

There's Hope To Combat Piracy If Hollywood, Industry, and Government Unite

Nelson Granados Contributor ⓘ
Hollywood & Entertainment
I cover digital trends in travel, media and entertainment.

This article is more than 5 years old.

Several studies have shown that piracy hurts the revenues of content owners, and instead pirate sites are reaping hundreds of millions of dollars in online advertising. Yet theft of movies and TV content seems to be as rampant today as ever. The Motion Picture Association of America (MPAA) reports that in 2014, just in the U.S. alone, 710 million movies and TV shows were shared via BitTorrent sites. Extrapolating to a global scale (the U.S. is less than 5% of the world's population) and adding streaming and other piracy methods, losses were likely in the billions of dollars. The staggering order of magnitude may lead some to wonder if it's even worth fighting the battle, or if it has been lost already. Can the battle against piracy be won? If so, how?

- IT Strategy

The Paradox of Openness: Exposure vs. Efficiency of APIs

Seth G. Benzell*
Guillermo Lagarda†
Jonathan Hersh‡
Marshall Van Alstyne§

August 3, 2019

ABSTRACT

APIs are the building blocks of digital platforms, yet there is little quantitative evidence on their use. Do API adopting firms do better? Do such firms change their operating procedures? Using proprietary data from a major API tools provider, we explore the impact of API use on firm value and operations. We find evidence that API use increases market capitalization and lowers R&D expenditures. We then document an important downside. API adoption increases the risk of data breaches, a risk that rises when APIs are more open or place less emphasis on security. Firms reduce API data flows in the month before a hack announcement, consistent with a conscious attempt to limit breach scope. In the same period, however, the variance of API data flows increases, consistent with heterogeneity in firms' ability to detect and shut down unauthorized data access. Our findings highlight a fundamental paradox of openness: It increases upside value and downside risk at the same time. We document that firms respond to these trade-offs in logical ways and conclude that the benefits of opening APIs exceed the risks for firms situated to adopt a platform strategy.

Keywords: Platforms, APIs, Information Security, Technology Strategy, Market Capitalization

Most Proud of: Cited on the Wikipedia Page for “Waffle”

Not logged in Talk Contributions Create account Log in

Article Talk | Read View source View history | Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Waffle

From Wikipedia, the free encyclopedia

This article is about the batter/dough-based food. For other uses, see Waffle (disambiguation).

A **waffle** is a dish made from leavened batter or dough that is cooked between two plates that are patterned to give a characteristic size, shape, and surface impression. There are many variations based on the type of waffle iron and recipe used. Waffles are eaten throughout the world, particularly in Belgium, which has over a dozen regional varieties.[1] Waffles may be made fresh or simply heated after having been commercially cooked and frozen.

Waffle

Place of origin	France, Belgium
Main ingredients	Batter or dough
Variations	Liège waffle, Brussels Waffle, Flemish Waffle, Bergische waffle, Stroopwafel and others

Cookbook: Waffle
Media: Waffle

Contents [hide]

1 Etymology
2 History
  2.1 Medieval origins
  2.2 14th–16th centuries
  2.3 17th–18th centuries
  2.4 19th–21st centuries
3 Varieties
4 Toppings
5 Consistency
6 Shelf stability and staling
7 See also

References

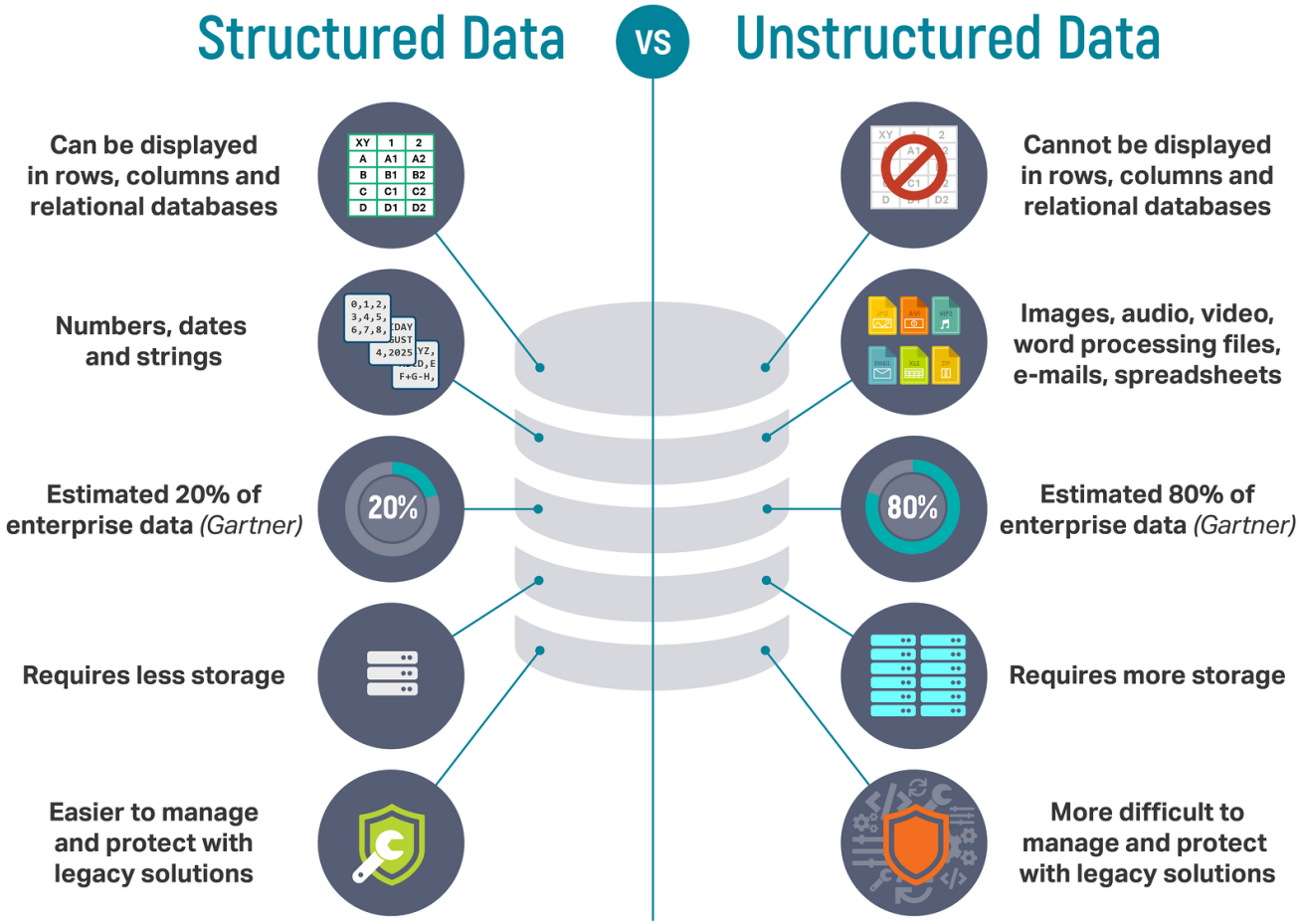
1. ^ "Les Gaufres Belges" Archived 2012-08-20 at the Wayback Machine. Gaufresbelges.com. Retrieved on 2013-04-07.
2. ^ Robert Smith (1725). *Court Cookery*. p. 176.
3. ^ "Waffle" Archived 2013-04-07 at the Wayback Machine, The Merriam-Webster Unabridged Dictionary

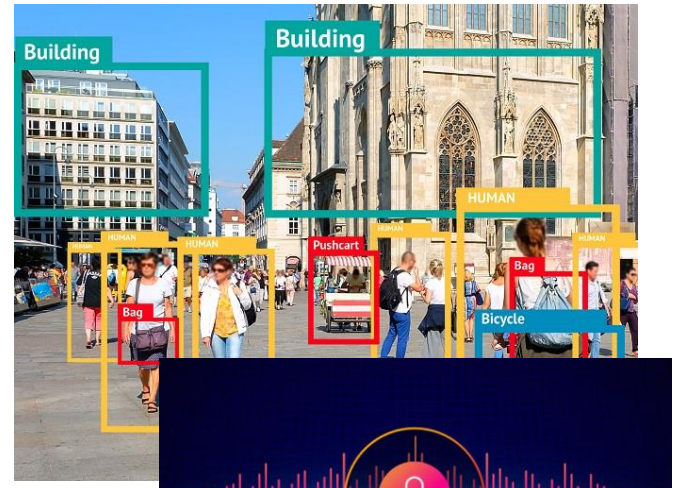
52. ^ [a] [b] "Sweet Diversity: Overseas Trade and Gains from Variety after 1492" Archived 2013-07-26 at the Wayback Machine, Jonathan Hersh, Hans-Joachim Voth, Real Sugar Prices and Sugar Consumption Per Capita in England, 1600–1850, p.42

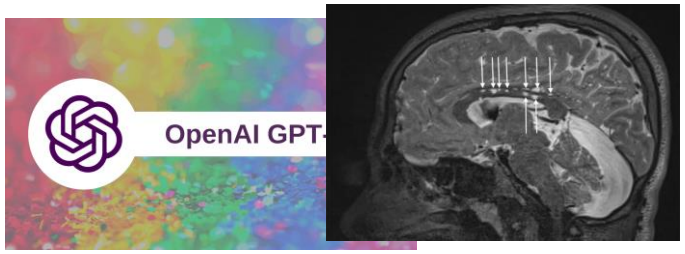
Key Takeaways

1. Use unstructured (non-tabular) data if you want to have a predictive edge
2. Be a full stack data scientist (models are APIs)

What is Unstructured Data?

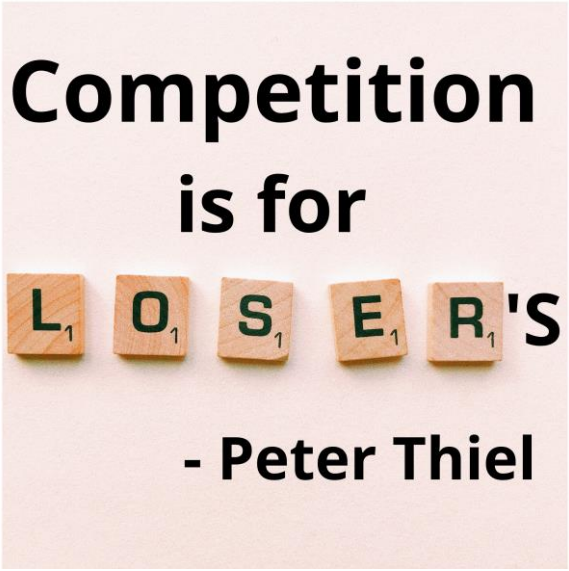




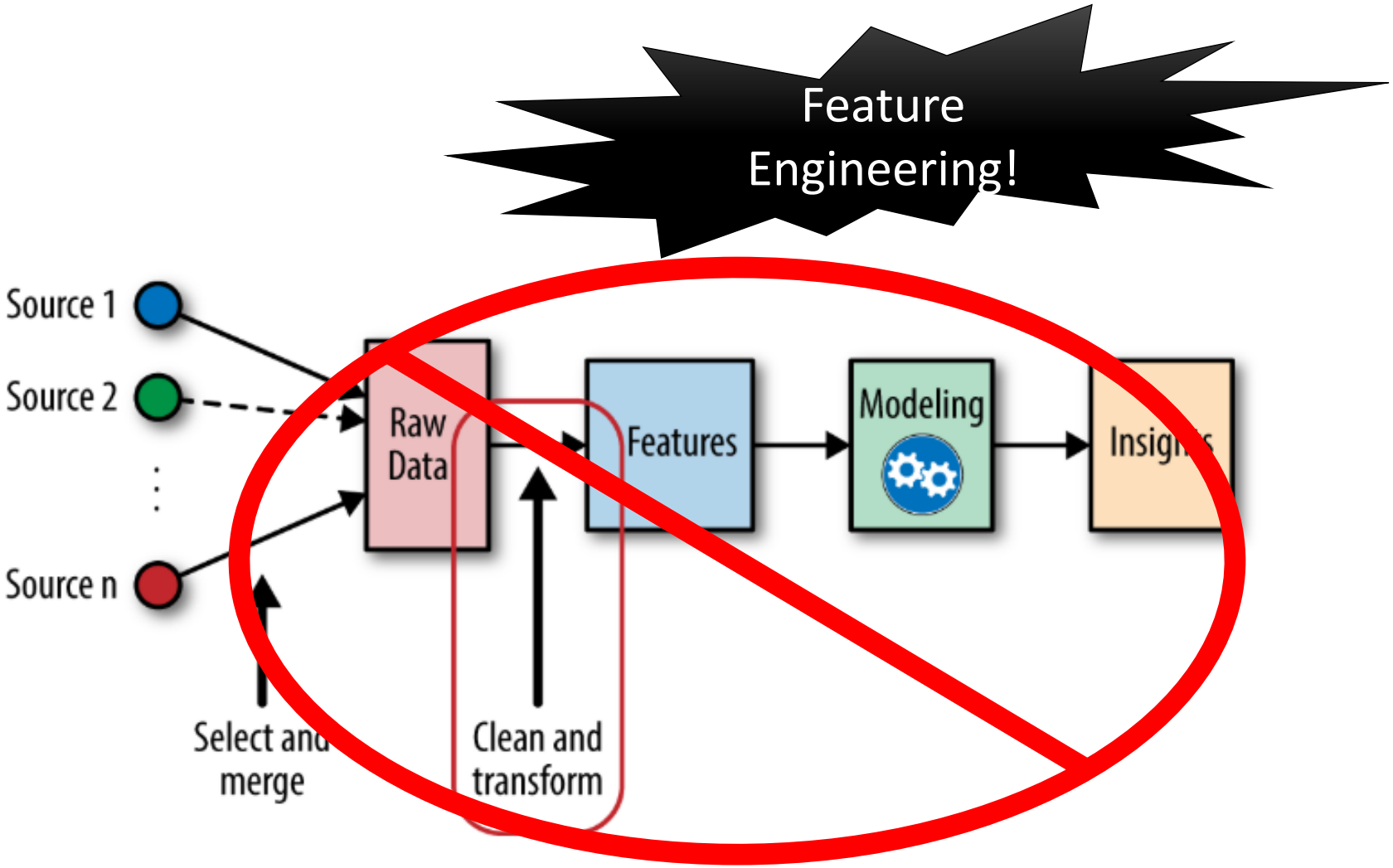


Why Doesn’t Everyone Use Unstructured Data?

- Unstructured = hard to work with!
  - Annoying to process, store, transform
- Difficulty to work with is an advantage if you’re starting out
- Hard = less competition



How Do We Bring Structure To Unstructured Data?



A Better Data Pipeline: Models Are APIs!

API = structured way to access your model

Raw Data

Raw Data Transformation

Analysis Dataset

Model Output Dataset

Data Viz Dashboard

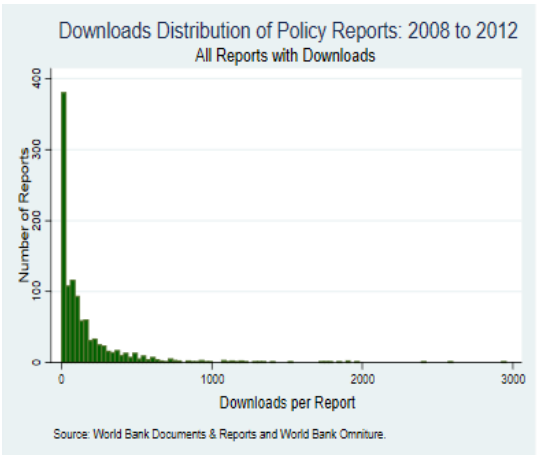
Report No One Will Read

Feature Engineering

Modeling & Exploratory Data Analysis

It's impossible to be an expert at all levels of the pipeline, but too many students starting out *only* focus on the middle.
Spread out!

Most classes only cover this portion of the data pipeline!



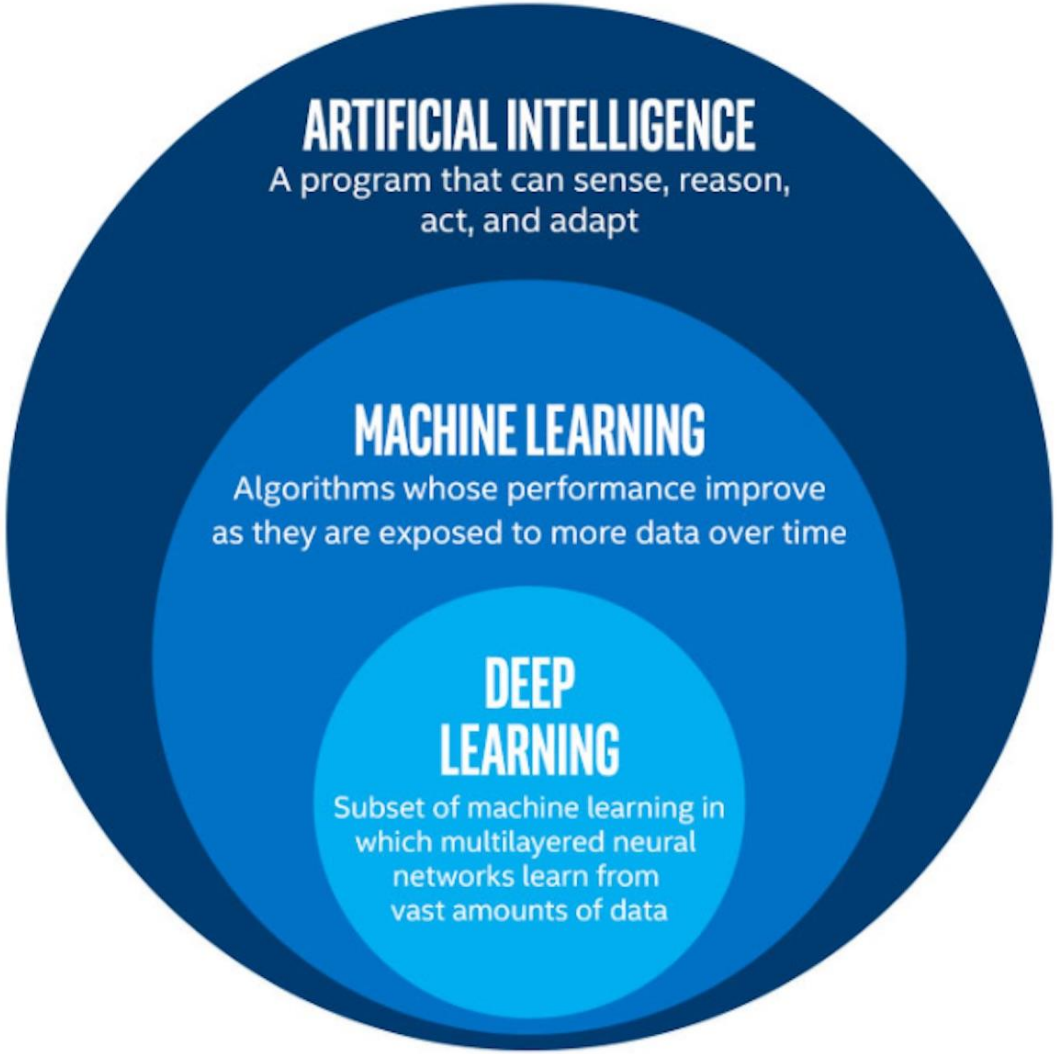
Questions at This Point

- What unstructured data do you think might be valuable in X years?
- How can you ensure safety/privacy and still take advantage of this data?
- What tools/libraries do you need to get good at to use this data?

Examples of Building Unstructured Data Pipelines for Insight

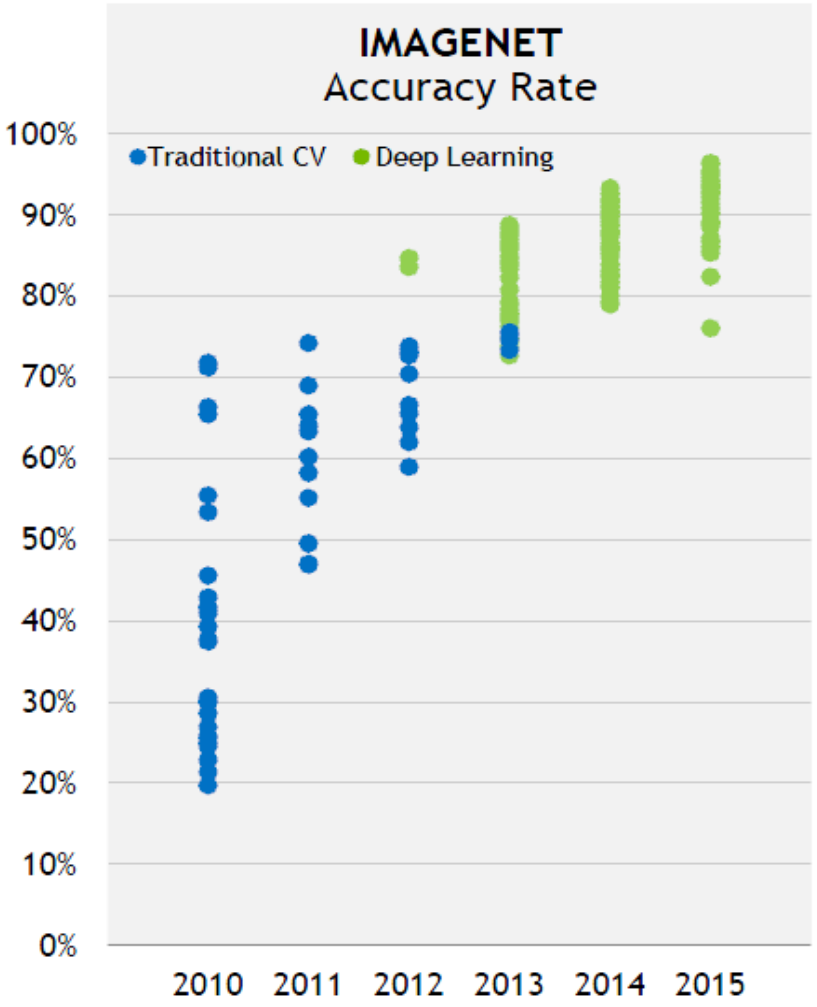
1. Convolutional Neural Networks as Feature Engineers
2. Estimating Violence in Syria using Satellite Imagery (Mueller et al., 2020. “Monitoring War Destruction from Space: A Machine Learning Approach”)

AI vs ML vs Deep Learning

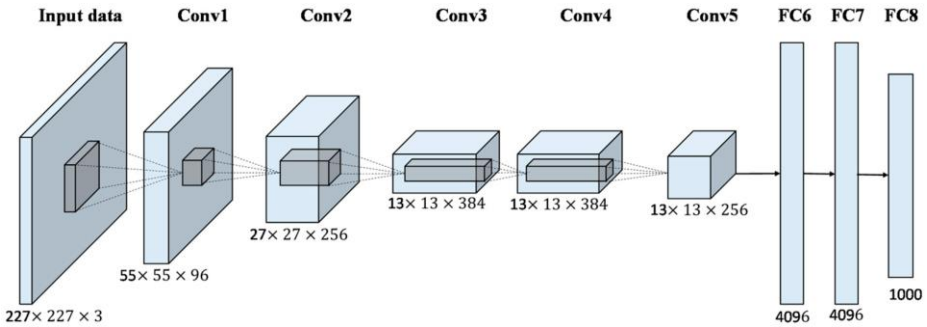


"**Deep learning is a specific subfield of machine learning…The deep in deep learning isn't a reference to any kind of deeper understanding achieved by the approach**; rather, it stands for this idea of successive layers of representations" – Francois Chollet

Deep Learning on Images: Convolutional Neural Networks



- Before deep learning, computers couldn't recognize objects from images.
- In 2012, Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton applied AlexNet, a CNN trained off of Graphical Processing Units (GPUs) to the ImageNet competition, designed to test computer's ability to see objects in images
- This kicked off a deep learning boom!

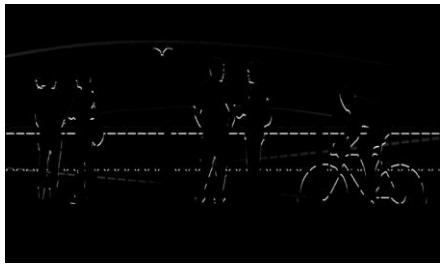


Convolutional Layers

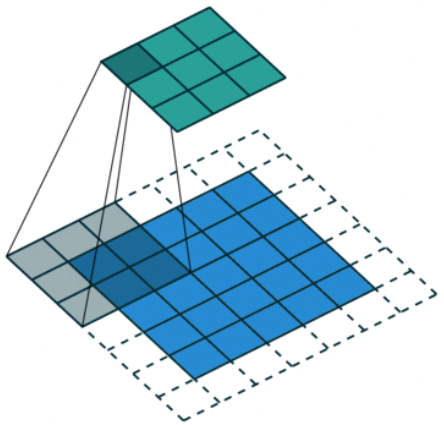
- A common operation in computer vision for processing images is to **convolve** the image with a **filter** or **kernel**

- These filters extract only certain salient features from the image, which vary depending on the filter used.

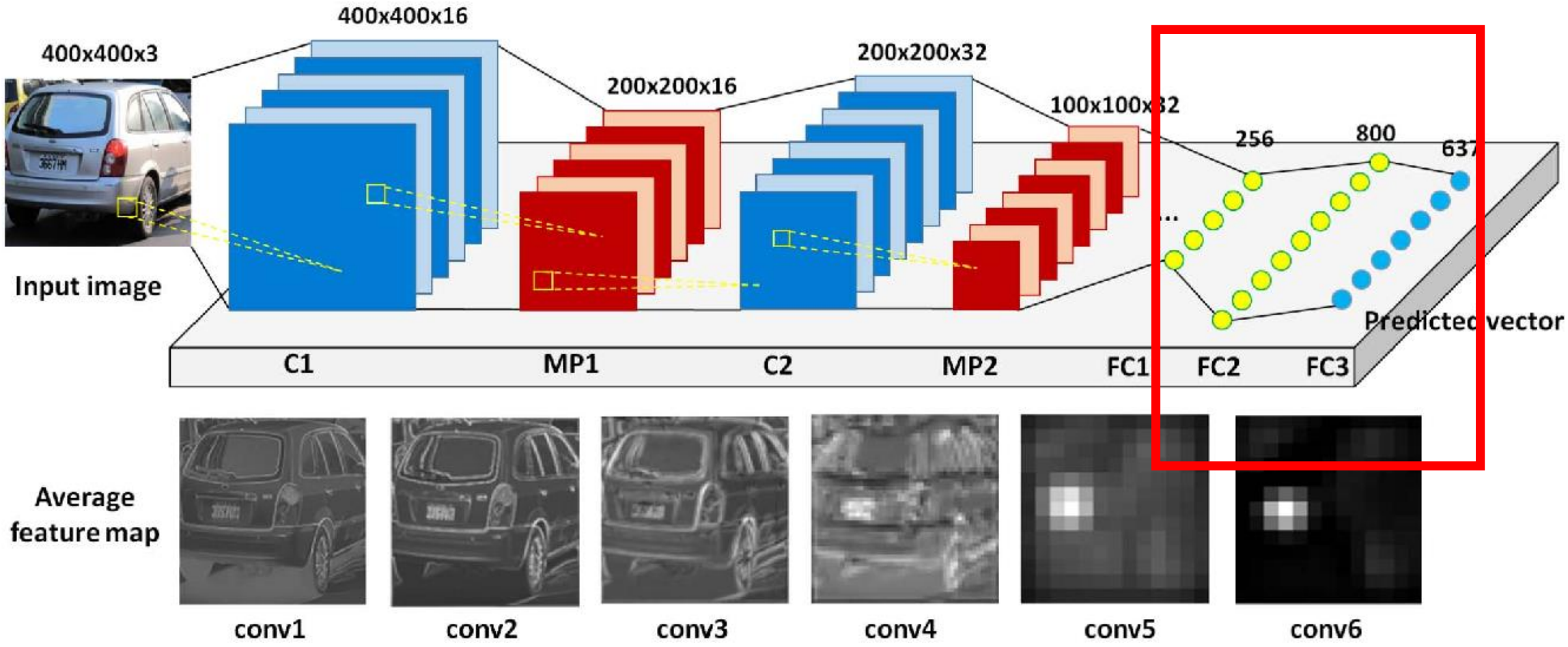
Vertical Edge Filter



Horizontal Edge Filter



From Filters to Feature Maps To Predictions!

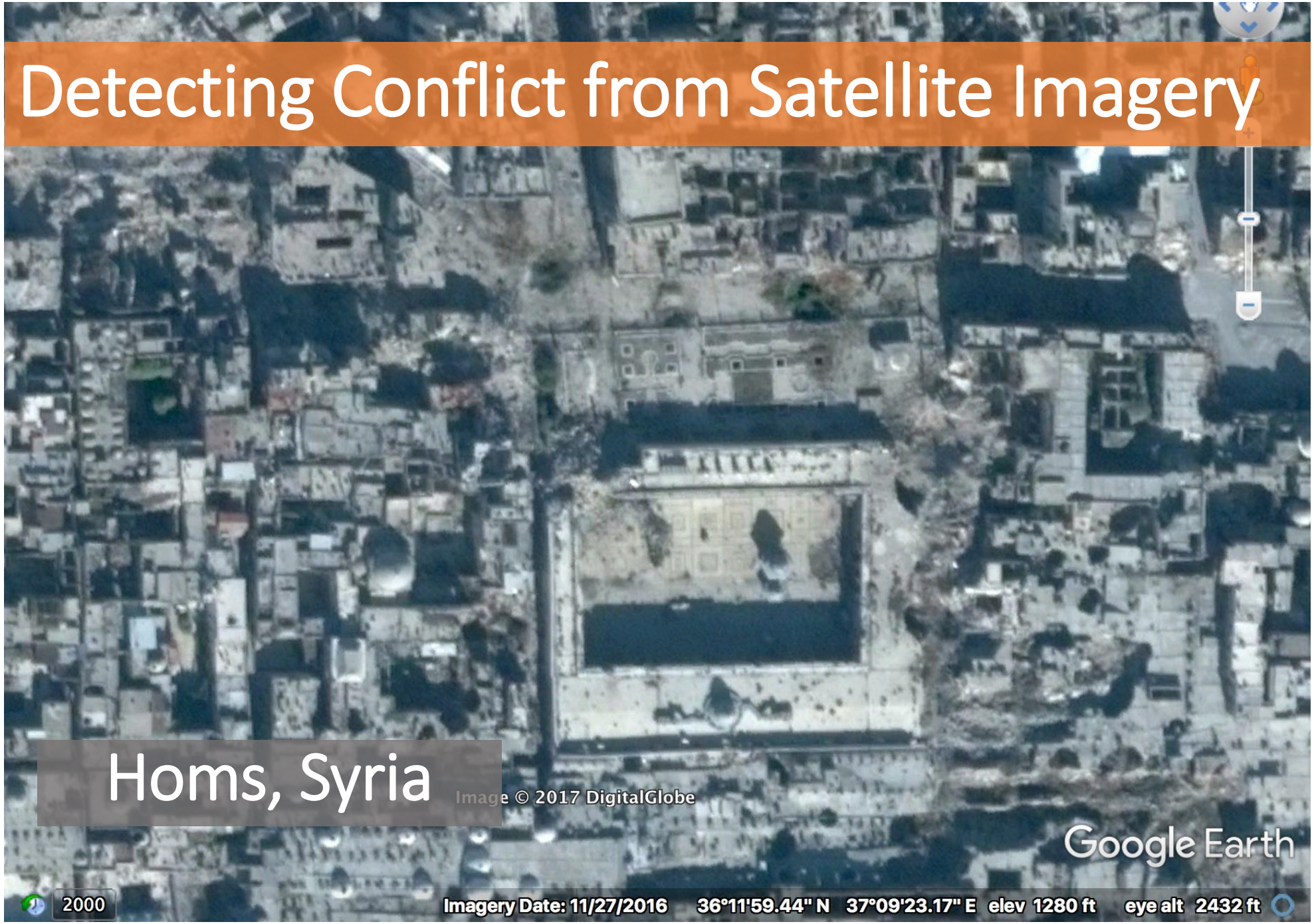


Examples of Building Unstructured Data Pipelines for Insight

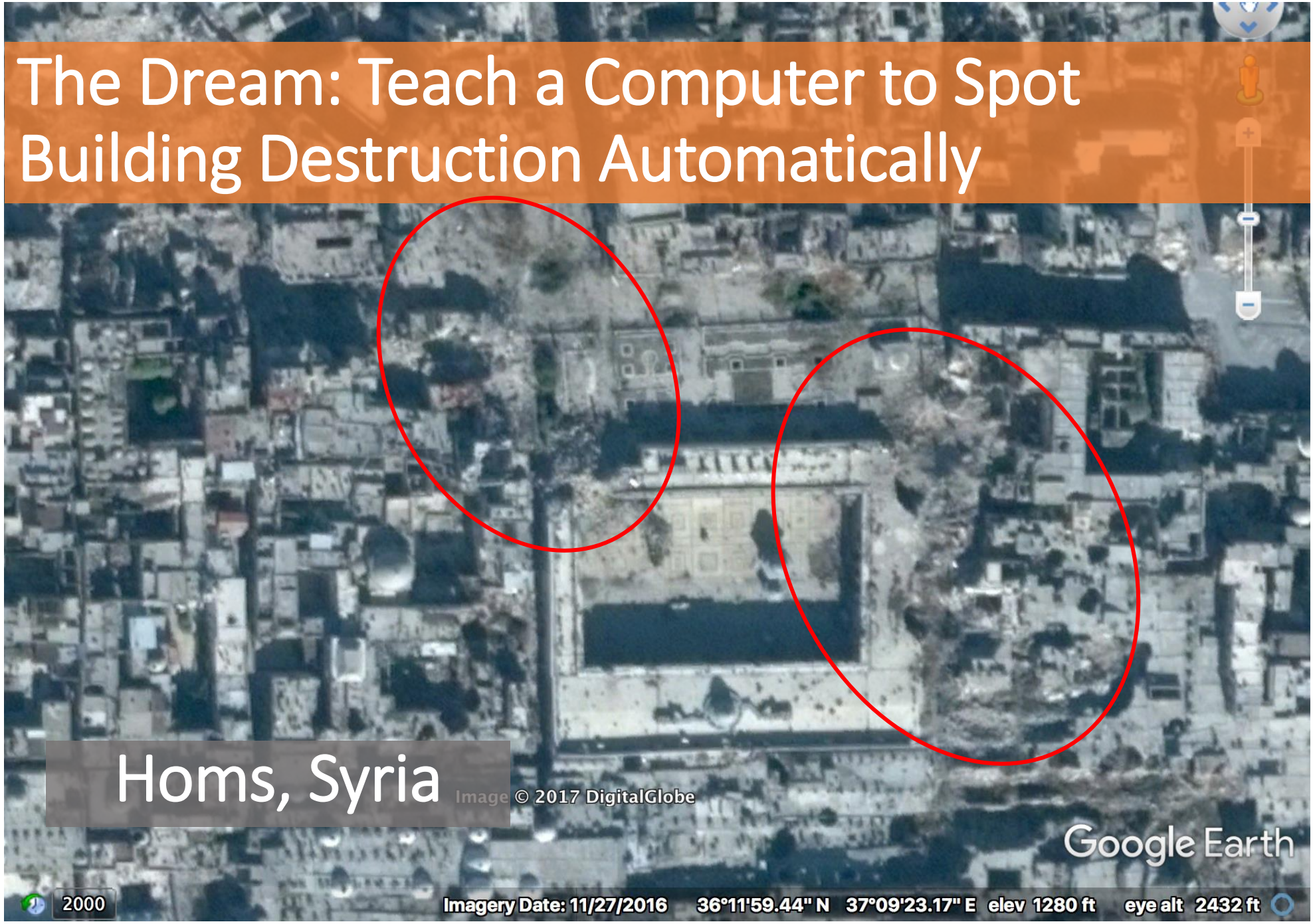
1. Convolutional Neural Networks as Feature Engineers
2. **Estimating Violence in Syria using Satellite Imagery (Mueller et al., 2020. "Monitoring War Destruction from Space: A Machine Learning Approach")**

Motivation Facts for Talk Today
Information on ongoing violence may be biased, outdated, and imprecise

Detecting Conflict from Satellite Imagery



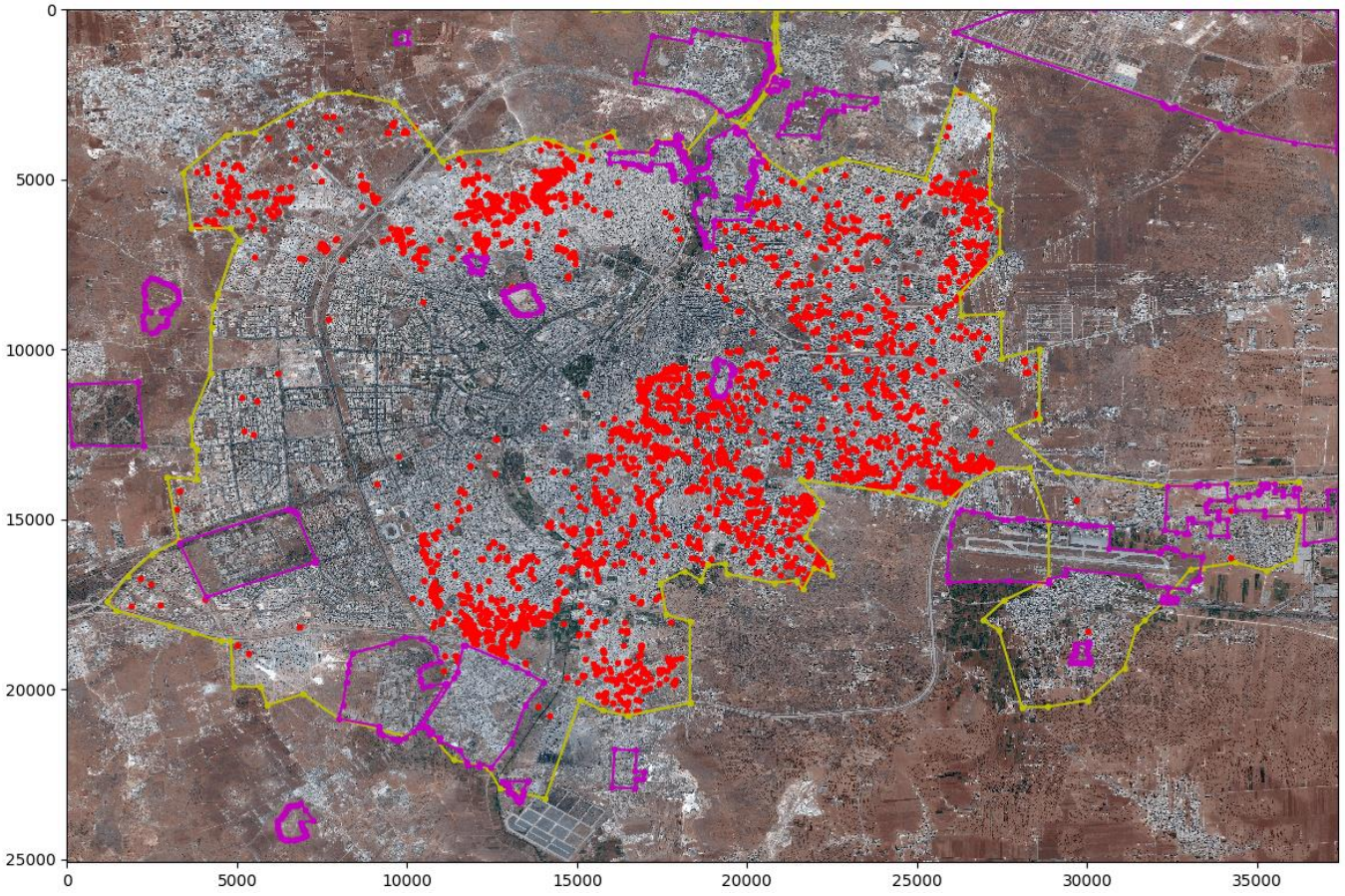
The Dream: Teach a Computer to Spot Building Destruction Automatically

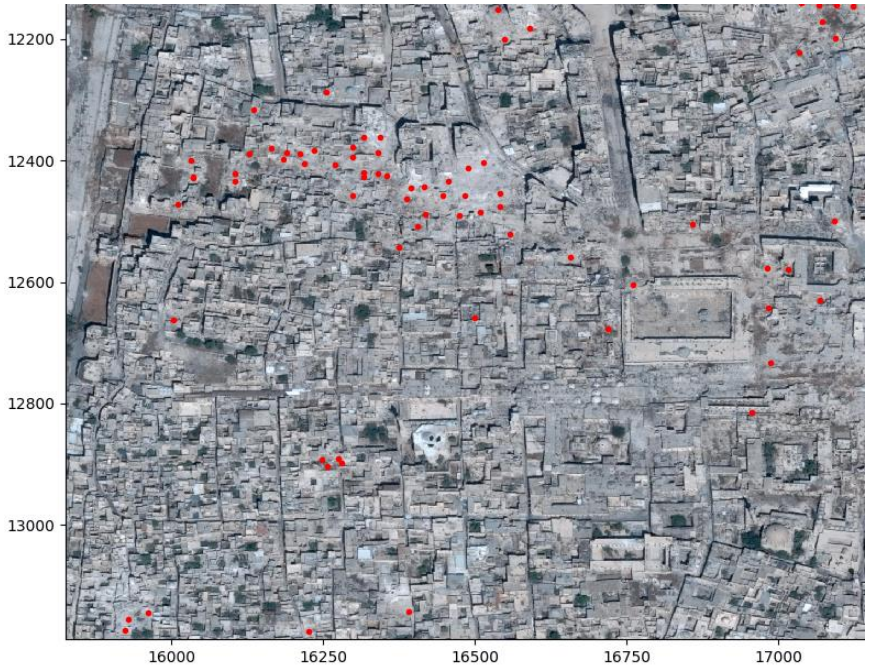


This Paper:

- 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

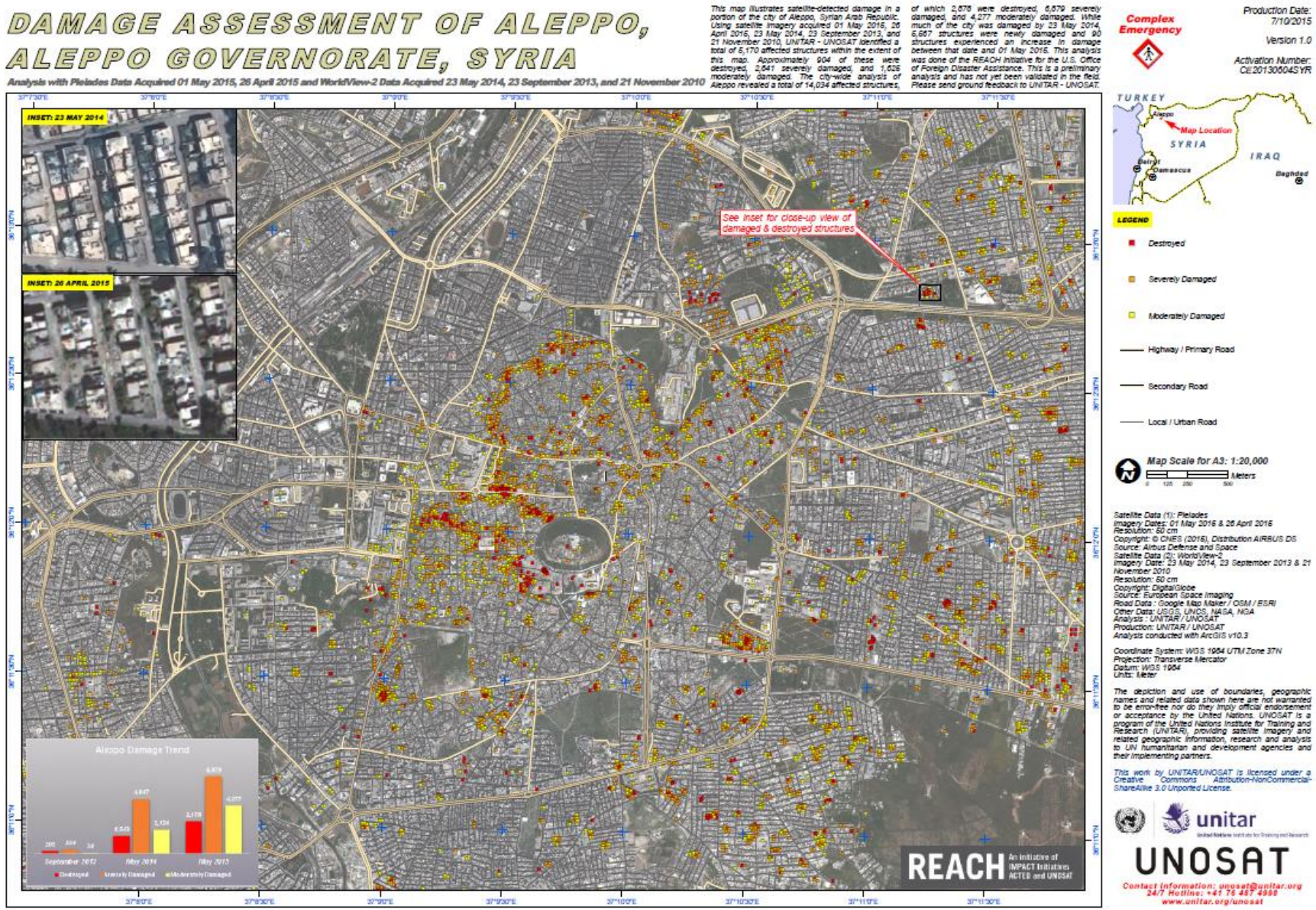
Illustration of Unbalanced Data Problem

- **Accuracy**: $\frac{\mathrm{TP}+\mathrm{TN}}{TP+FN+FP+TN}$
- **Precision** (share of positives predicted correctly): $\frac{\mathrm{TP}}{TP+FP}$
- **True Positive Rate / Recall** (share of actual positives predicted that are correct): $\frac{\mathrm{TP}}{TP+FN}$

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

- Suppose you have 100,000 images, but only 1000 (1%) are destroyed
- 12% FRP 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: $\frac{\mathrm{TP}}{TP+FP} = \frac{800}{11880+800} = 0.063$
- Probability of being correct if you find destruction is only 6.3%!

Ground Truth: Building Destruction Annotations from UNOSAT



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
			
			

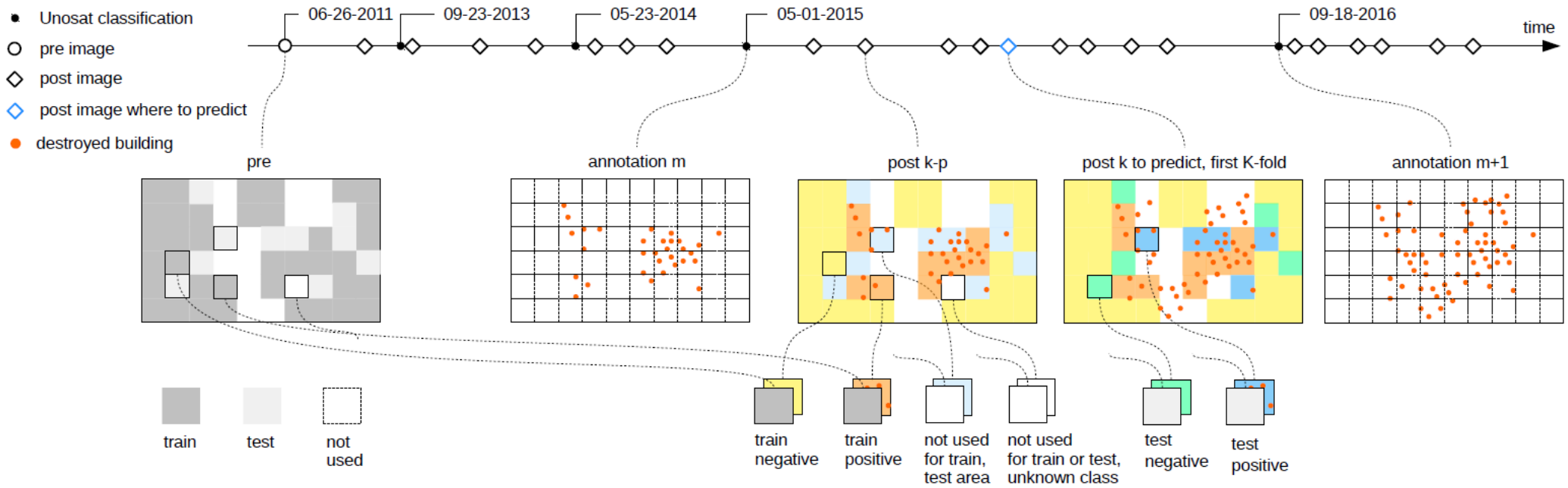

Label


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

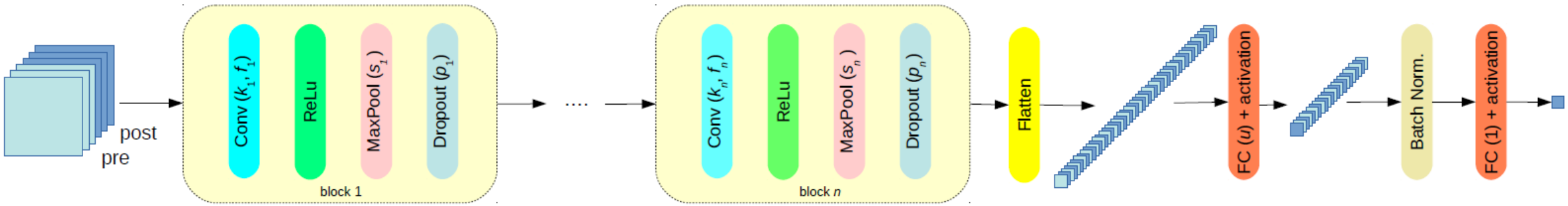


- Four temporal periods with annotations (at patch level)
- Training and testing samples are from distinct spatial areas

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

Neural Network Architecture



Each convolutional block composed of a convolutional layer, a ReLu, a MaxPool, and Dropout with parameters cross-validated

Key: flexible network where hyper parameters optimized given validation data performance

Second Stage Machine Learning Smoothing

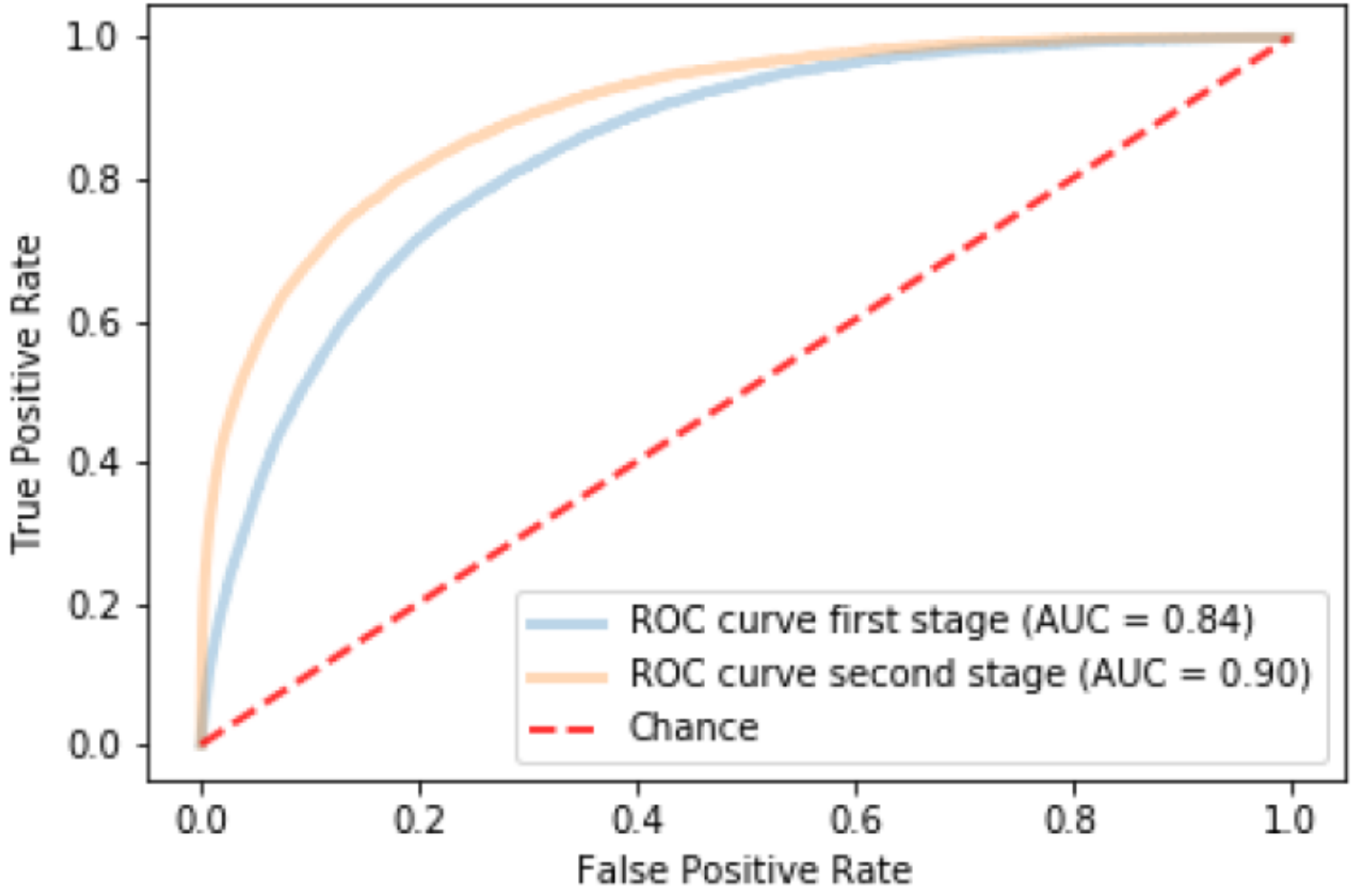
- Idea: destruction is correlated in space and time
- We train a random forest model on the CNN predictions using two spatial lags and two temporal lags
- This stage separated from CNN stage for maximum flexibility and modularity

Table 2. Model precision varying second stage

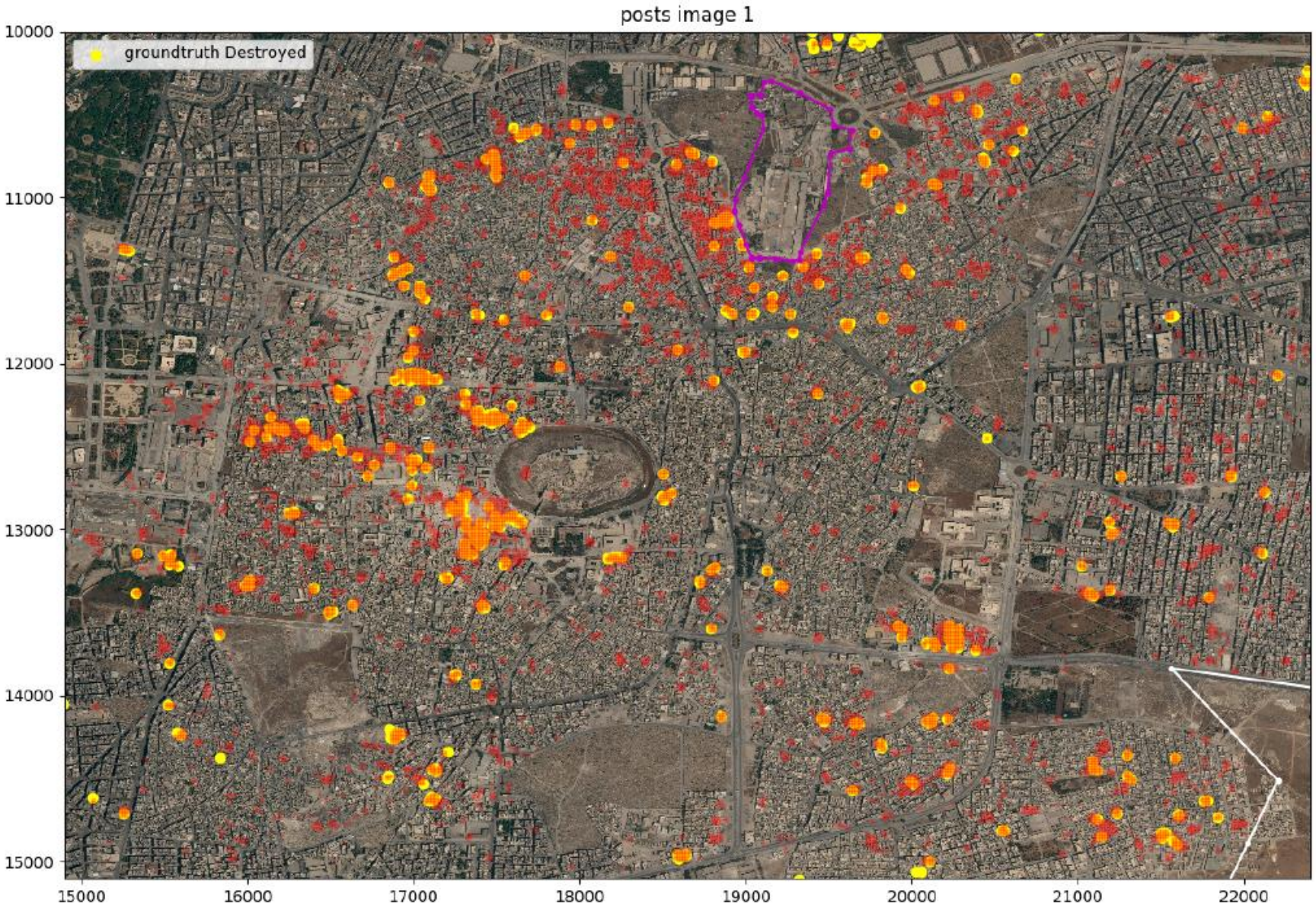
City	(1) First stage	(2)	(3) Second stage	(4)
	CNN	CNN + Spatial	CNN + Spatial & 1 temporal lag/lead	CNN + Spatial & 2 temporal lag/leads
Aleppo	16.1%	16.8%	28.2%	35.7%
Daraa	4.2%	4.6%	9.5%	12.1%
Deir-Ez-Zor	11.0%	12.0%	18.6%	21.9%
Hama	54.5%	65.3%	67.5%	68.0%
Homs	25.8%	35.3%	44.6%	56.1%
Raqqa	12.8%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

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

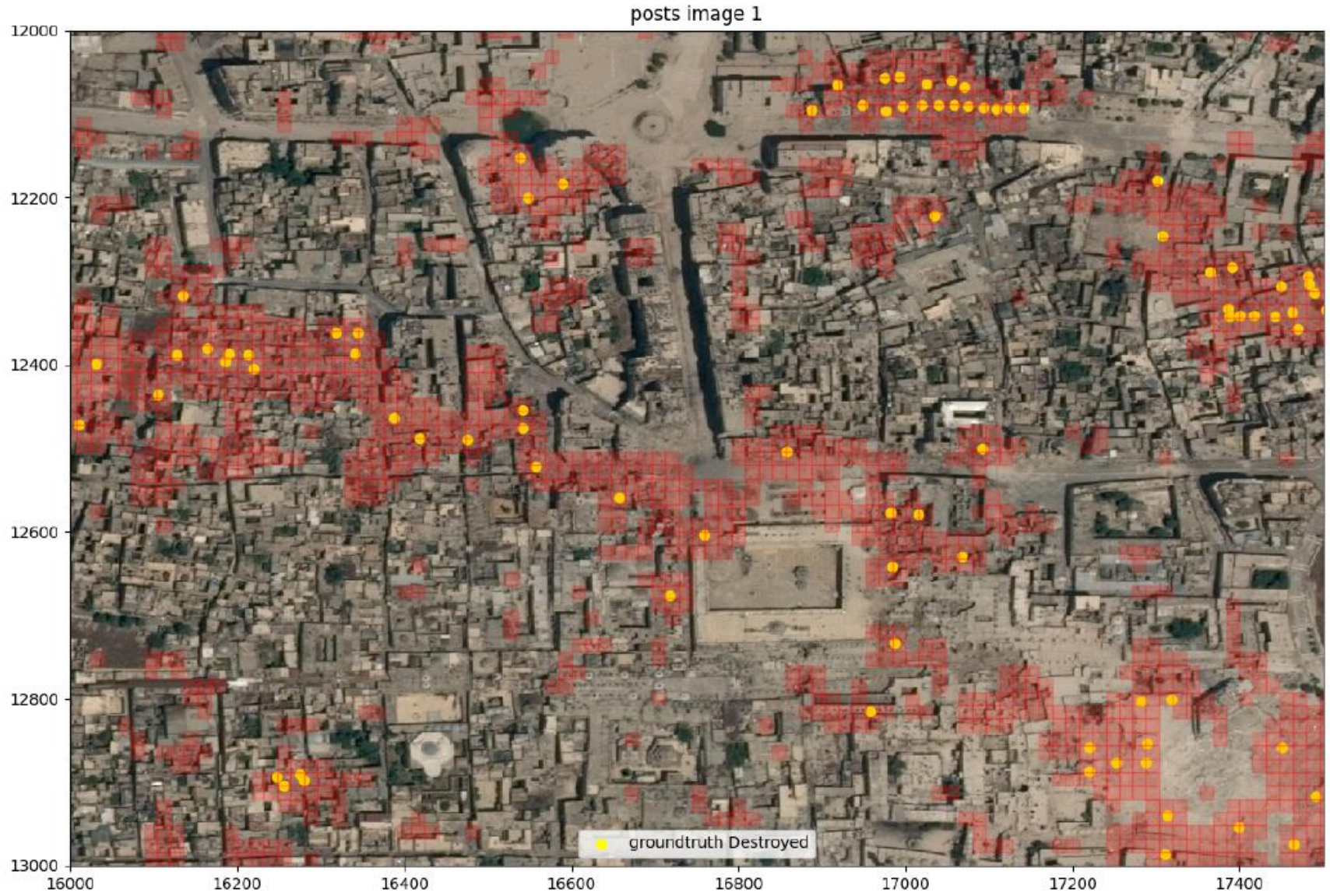
Aleppo ROC Curve (Test Sample)



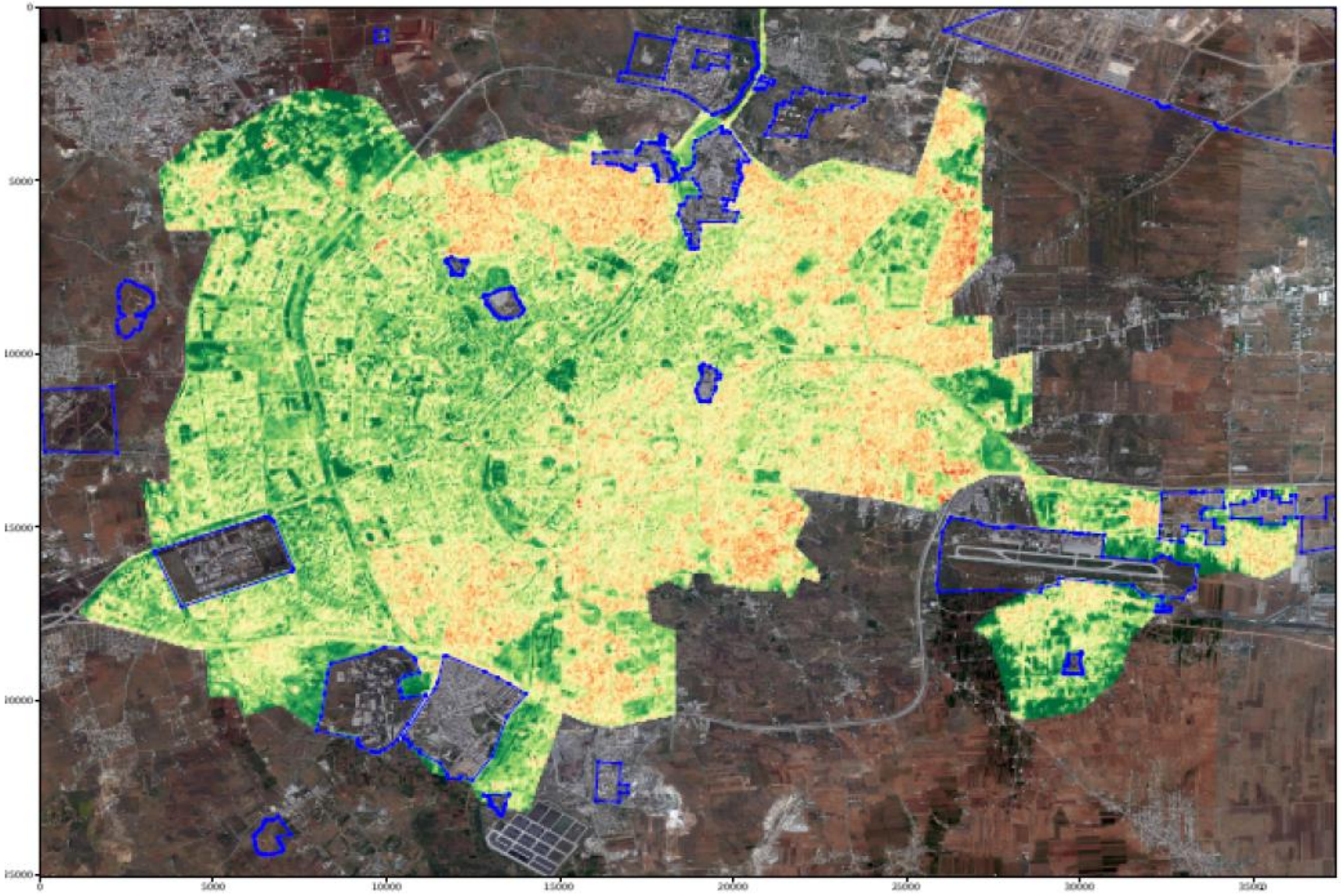
Dense Predictions Aleppo



Dense Predictions Aleppo



Dense Predictions Aleppo



Conclusion

- We propose a novel data augmentation and second stage machine learning stage to increase precision of building destruction
- 2nd 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