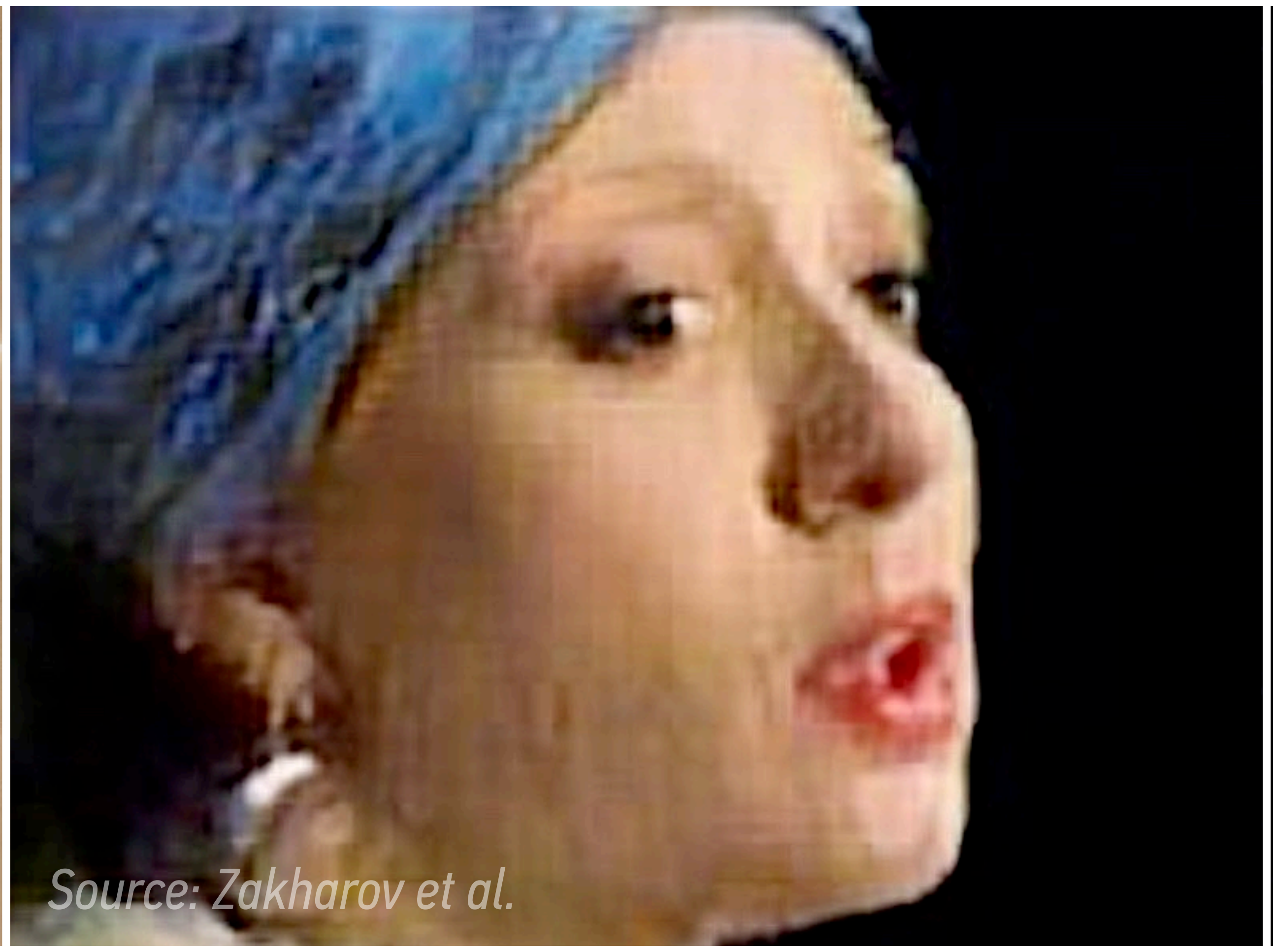
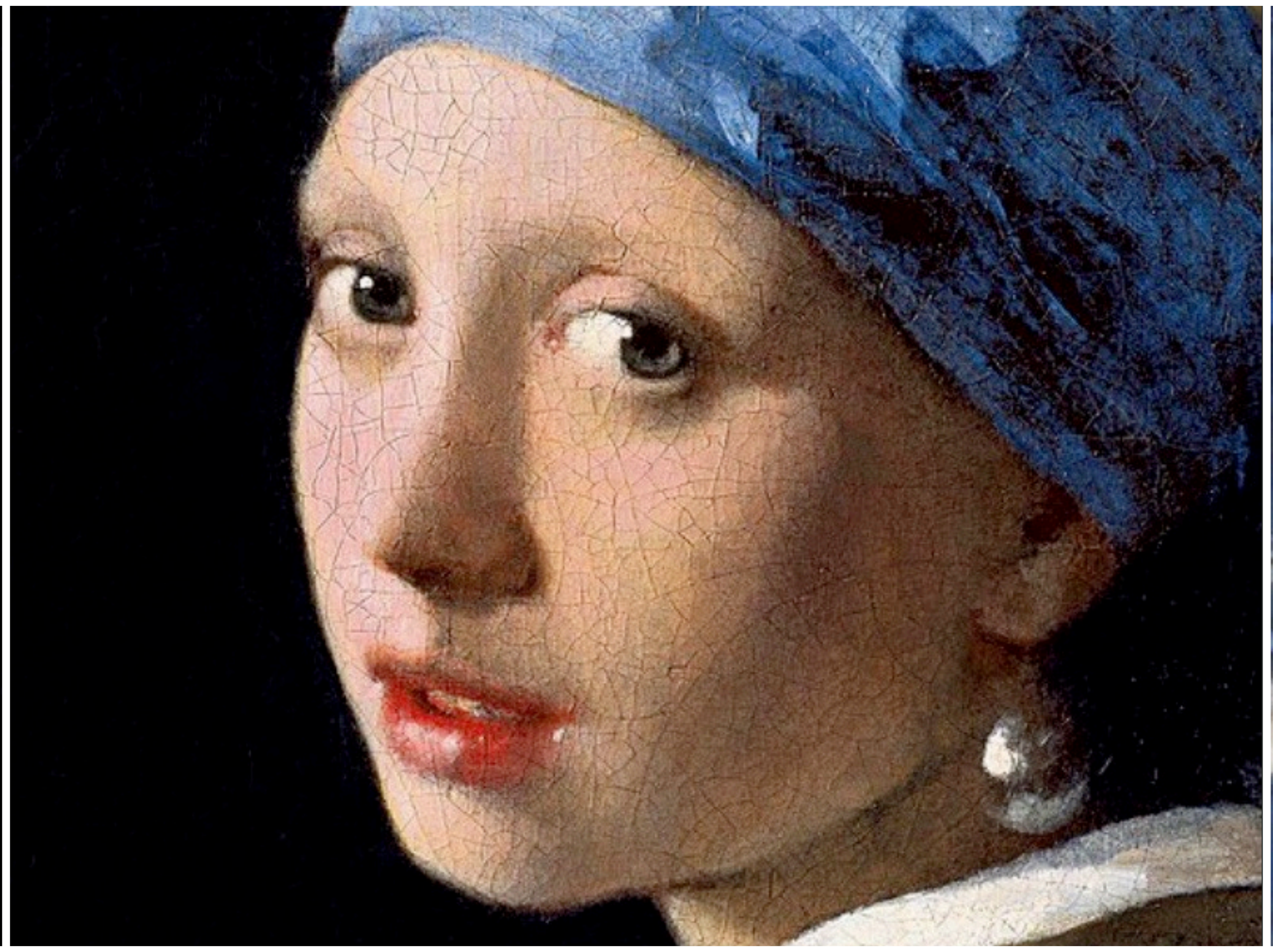





# CITIZEN-BASED MONITORING FOR PEACE & SECURITY

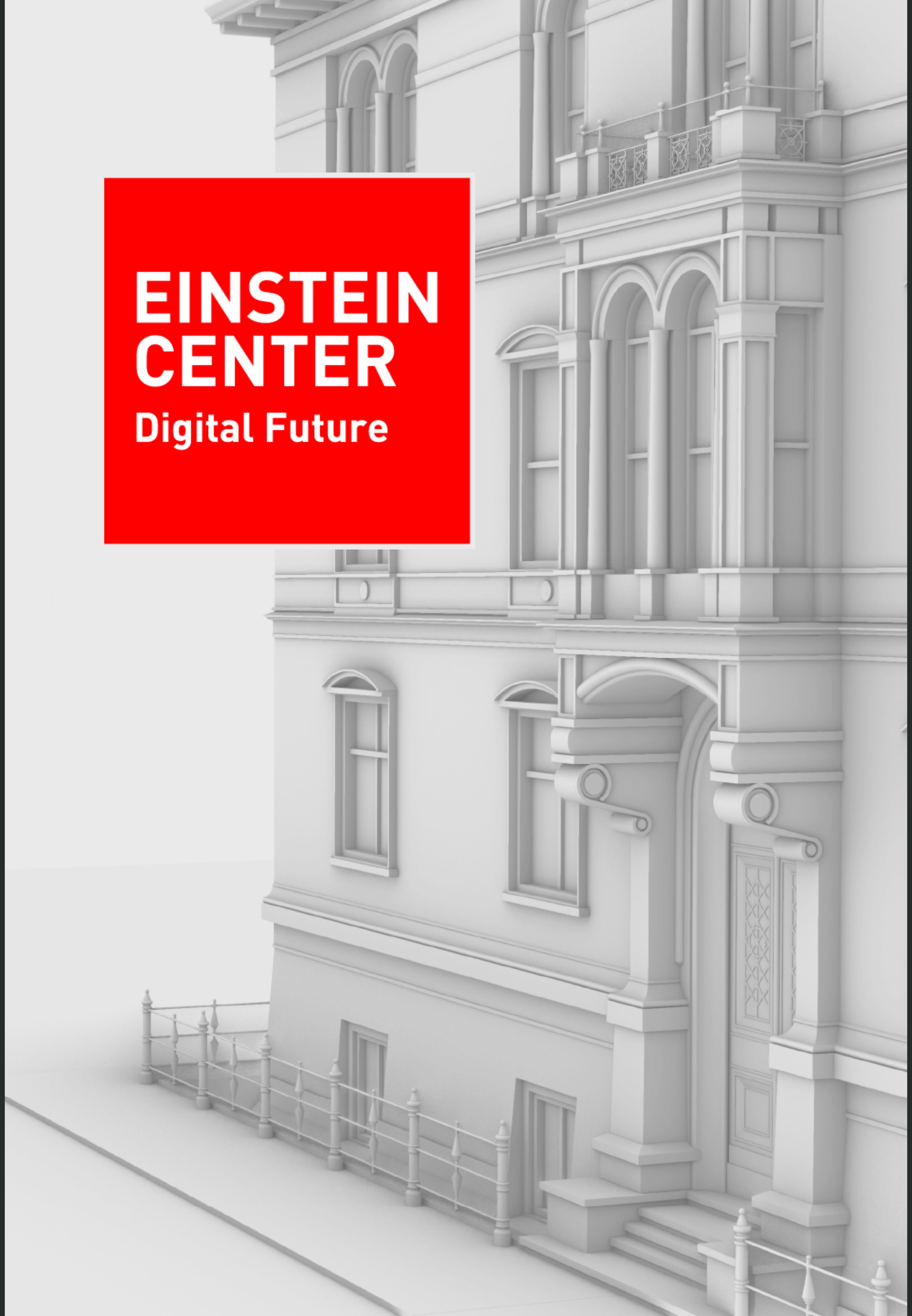
## IN THE ERA OF SYNTHETIC MEDIA AND DEEPFAKES

Alex Glaser and Vy Nguyen
Princeton University | Berliner Hochschule für Technik
Einstein Center Digital Future, Berlin

Helmholtz Einstein International Berlin Research School in Data Science
Berlin, July 12, 2023

Revision 3

BHT
Berliner Hochschule
für Technik



EINSTEIN
CENTER
Digital Future

SCIENCE &
GLOBAL SECURITY
PRINCETON UNIVERSITY

# PROJECT TEAM

Vy Nguyen
Berliner Hochschule für Technik

Felix Biessmann
Berliner Hochschule für Technik

Rebecca D. Frank
University of Tennessee, Knoxville

Sara Al-Sayed
Princeton University

Igor Moric
Princeton University

Kristian Hildebrand
Berliner Hochschule für Technik

Alex Glaser
Princeton University

Johannes Hoster
Berliner Hochschule für Technik

Supported by the German Foundation for Peace Research

About | Princeton Science & Gl

sgs.princeton.edu/about

Incognito



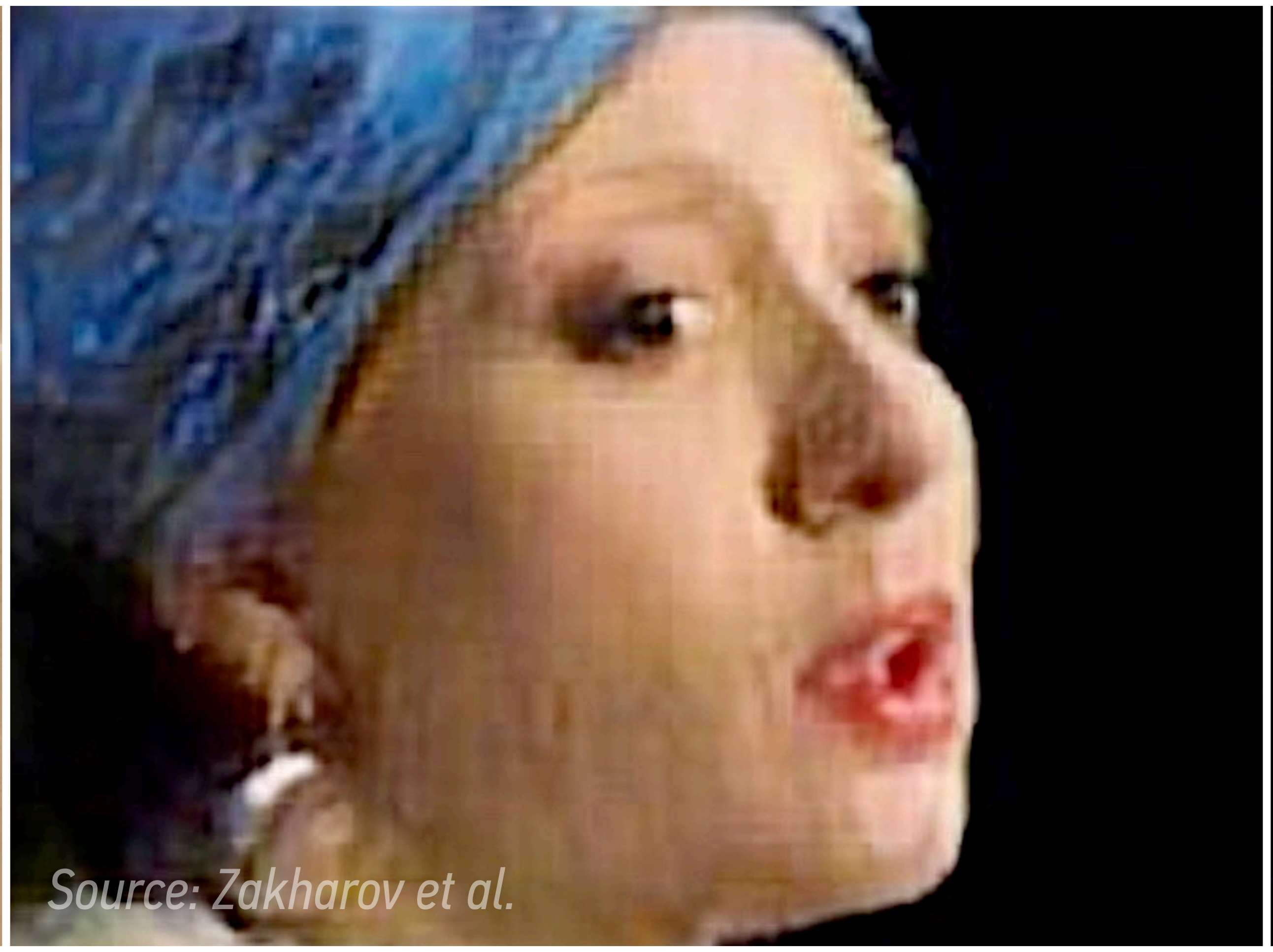
ABOUT US

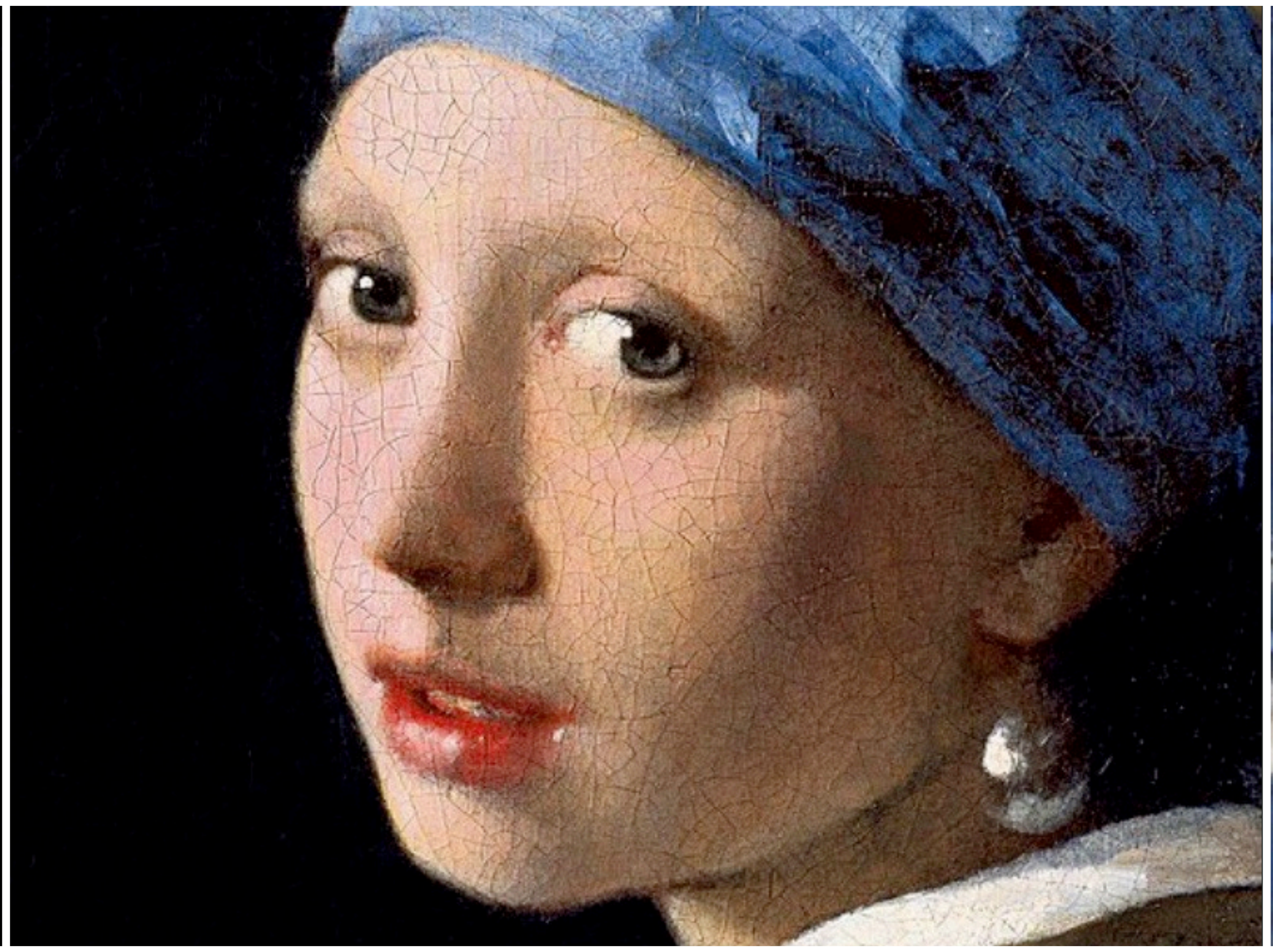
# Science, technology, and policy for a safer and more peaceful world

NUCLEAR | VERIFICATION | FISSILE MATERIALS | REGIONS | SPACE | EMERGING TECHNOLOGIES | BIOTECHNOLOGY

sgs.princeton.edu


Source: Planet Labs, Inc.


Source: Zakharov et al.



# CITIZEN-BASED MONITORING FOR PEACE & SECURITY

## IN THE ERA OF SYNTHETIC MEDIA AND DEEPFAKES

Alex Glaser and Vy Nguyen

Princeton University | Berliner Hochschule für Technik

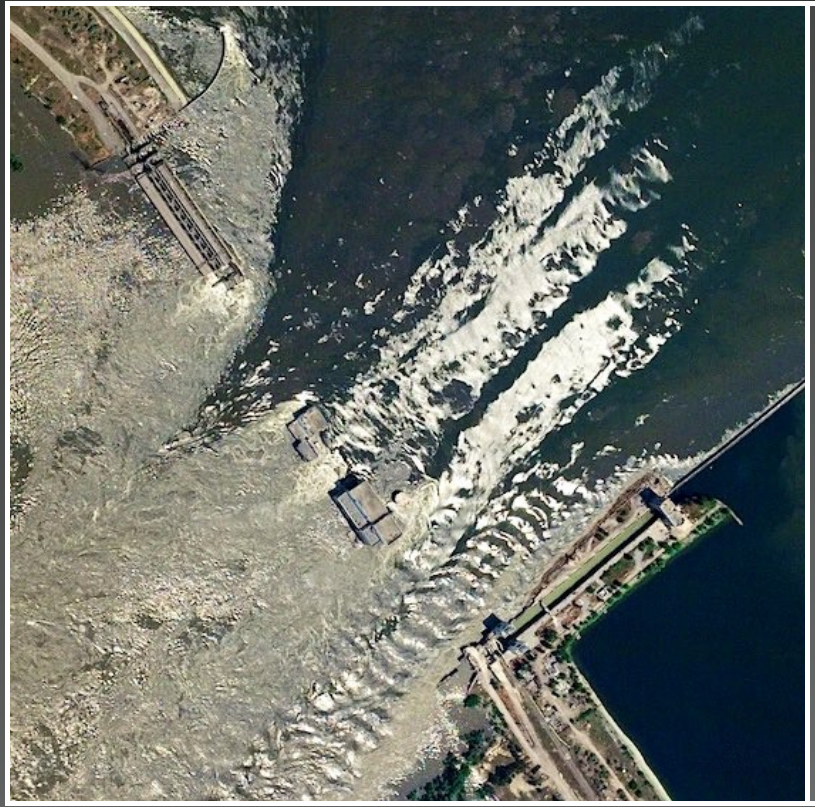
Einstein Center Digital Future, Berlin

Helmholtz Einstein International Berlin Research School in Data Science

Berlin, July 12, 2023

independent

# TWO MAJOR DEVELOPMENTS



## ABILITY TO MONITOR THE PLANET IN NEAR REAL-TIME

Evolving “megaconstellations” of optical imaging (and other) satellites with revisit times as short as 20 minutes; even high-resolution imagery becoming commercially available at scale

Relevant for many communities with an interest in Earth Observation (EO)

## ABILITY TO GENERATE SYNTHETIC MEDIA THAT ARE INDISTINGUISHABLE FROM REAL MEDIA

With the advent of Generative AI (such as Stable Diffusion or DALL·E 2), it is becoming easier to generate realistic synthetic media and deepfakes — posing a range of challenges for society and policy

Dilemma to avoid: “When everything is possible, nothing really matters”

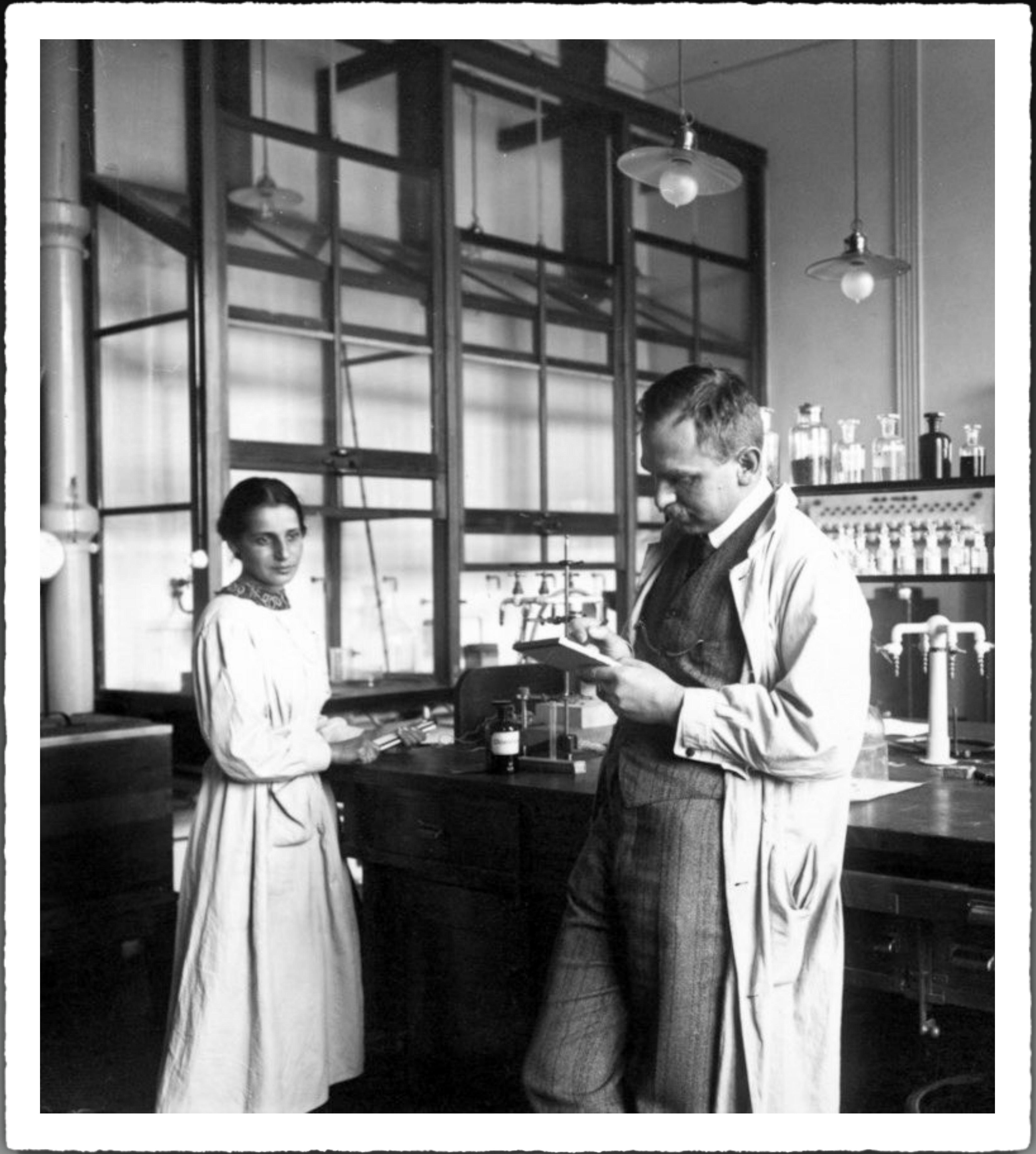
*Source: Planet Labs (top) and Pablo Xavier, www.reddit.com/r/midjourney (bottom)*

*“Historically, it will turn out that there was this weird time when people just assumed that photography and videography were true. And now that very short little period is fading.”*

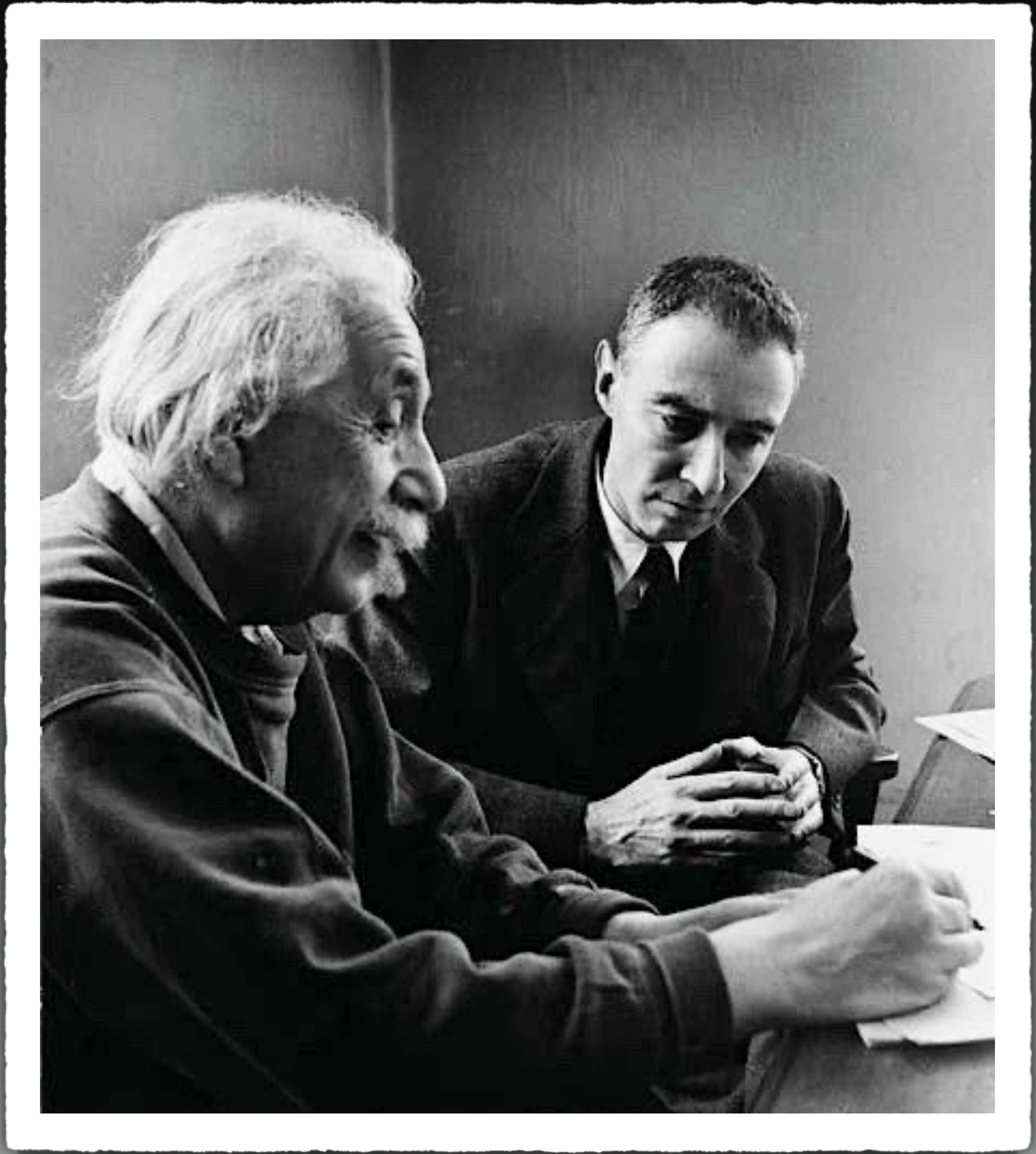
Alexei A. Efros

November 2018

# BACKGROUND

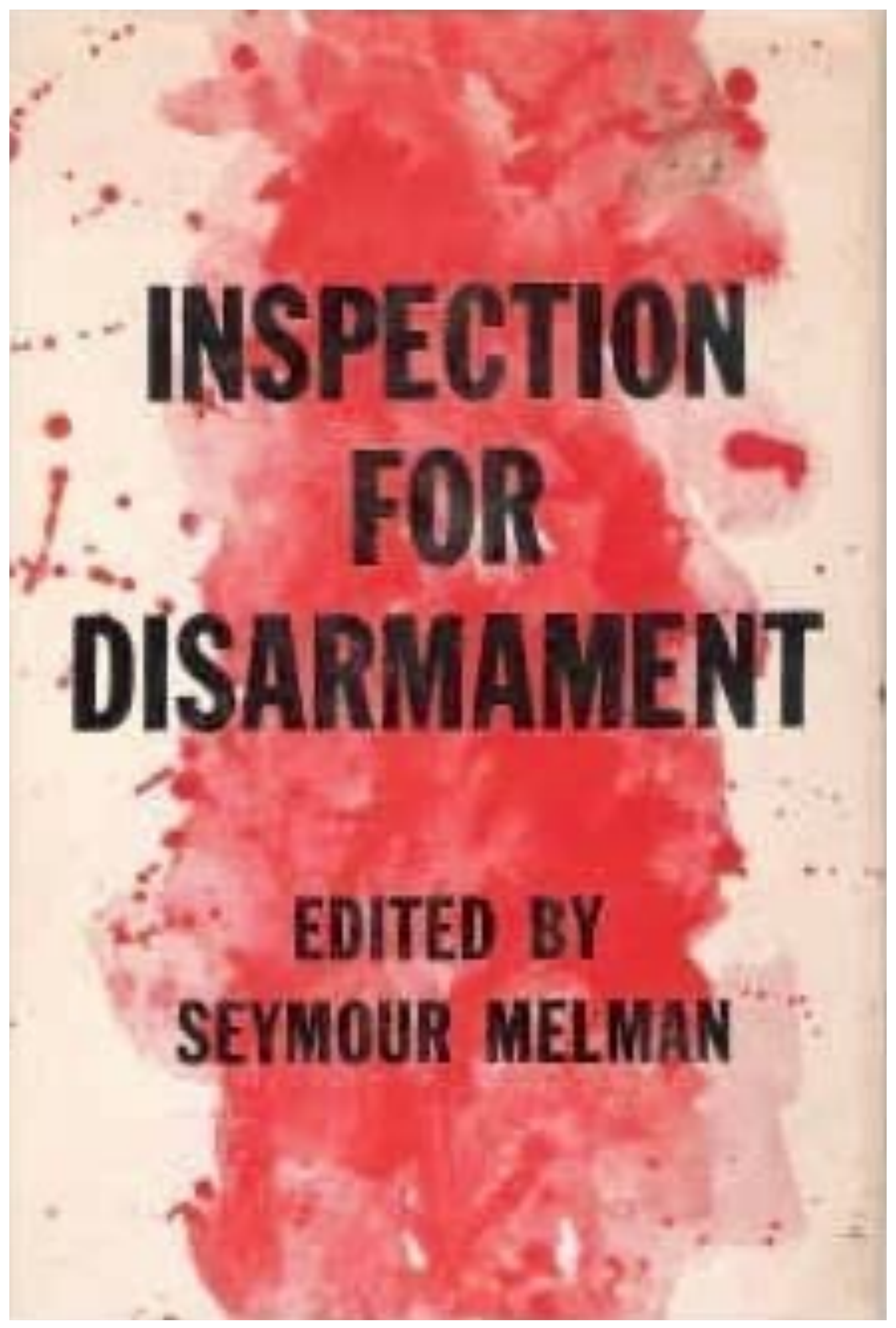


Lise Meitner and Otto Hahn, Berlin, c. 1925
They would discover nuclear fission in 1938/1939



Albert Einstein and J. Robert Oppenheimer, Princeton, 1947
Photo by Alfred Eisenstaedt

# “INSPECTION BY THE PEOPLE”




From this viewpoint the problem may be posed: How can the manpower requirements for a major clandestine production effort be used to strengthen the possibilities of inspection for disarmament?

Inspection by the people is a method that would serve this purpose. In addition to the specific monitoring activities of the inspectorate, it would be invaluable to have a randomly distributed network of inspection that is based upon public support for inspection for disarmament. Such public support could reinforce the work of the inspectorate and could help to undercut evasion efforts that require substantial organizations and widespread production systems. The operation of effective world-wide inspection by the people would be facilitated if the disarmament agreements included provisions which made it a duty, an explicit obligation, of the citizens of participating countries to report violations to the international inspectorate.

Seymour Melman (ed.), *Inspection for Disarmament,* Columbia University Press, New York, 1958
see in particular: “Inspection by the People: Mobilization of Public Support” (pp. 38–44)

For a similar discussion, see Jerome B. Wiesner, ”Inspection for Disarmament,” Chapter 4 in *Arms Control: Issues for the Public,* Prentice-Hall, 1961



**Briefing** Open-source intelligence

The Economist August 7th 2021

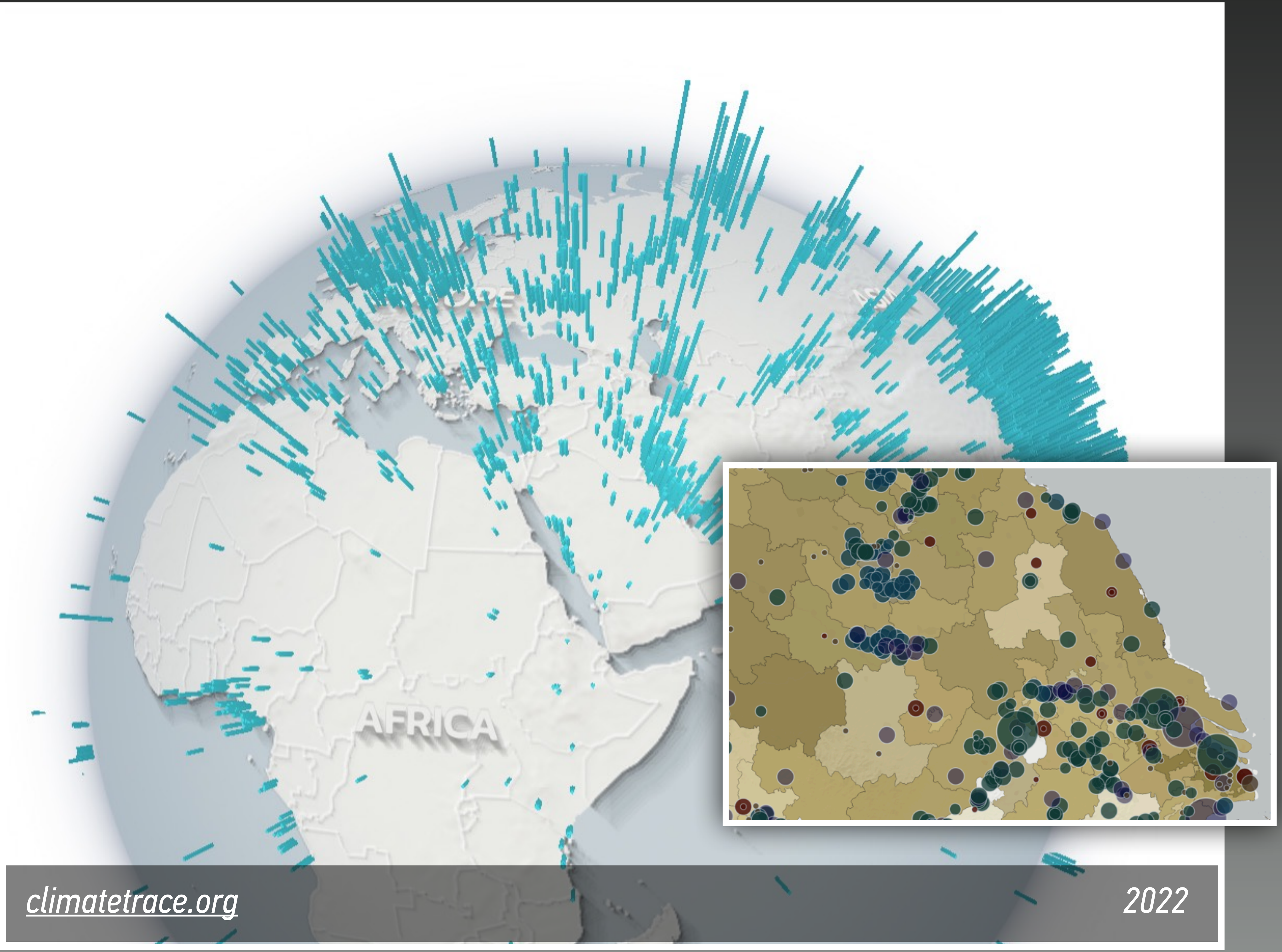
## Trainspotting, with nukes

Geo4Nonpro, a crowdsourced project which let budding hobbyists and seasoned experts collaborate to annotate satellite pictures of everything from uranium mines in India to chemical-weapon facilities in Syria. "It's fun," says Mr Eveleth.

www.economist.com/weeklyedition/2021-08-07

# ENVIRONMENTAL MONITORING


*Skybox Imaging, www.youtube.com/watch?v=fCrB1t8MncY* *2013*




*climatetrace.org* *2022*

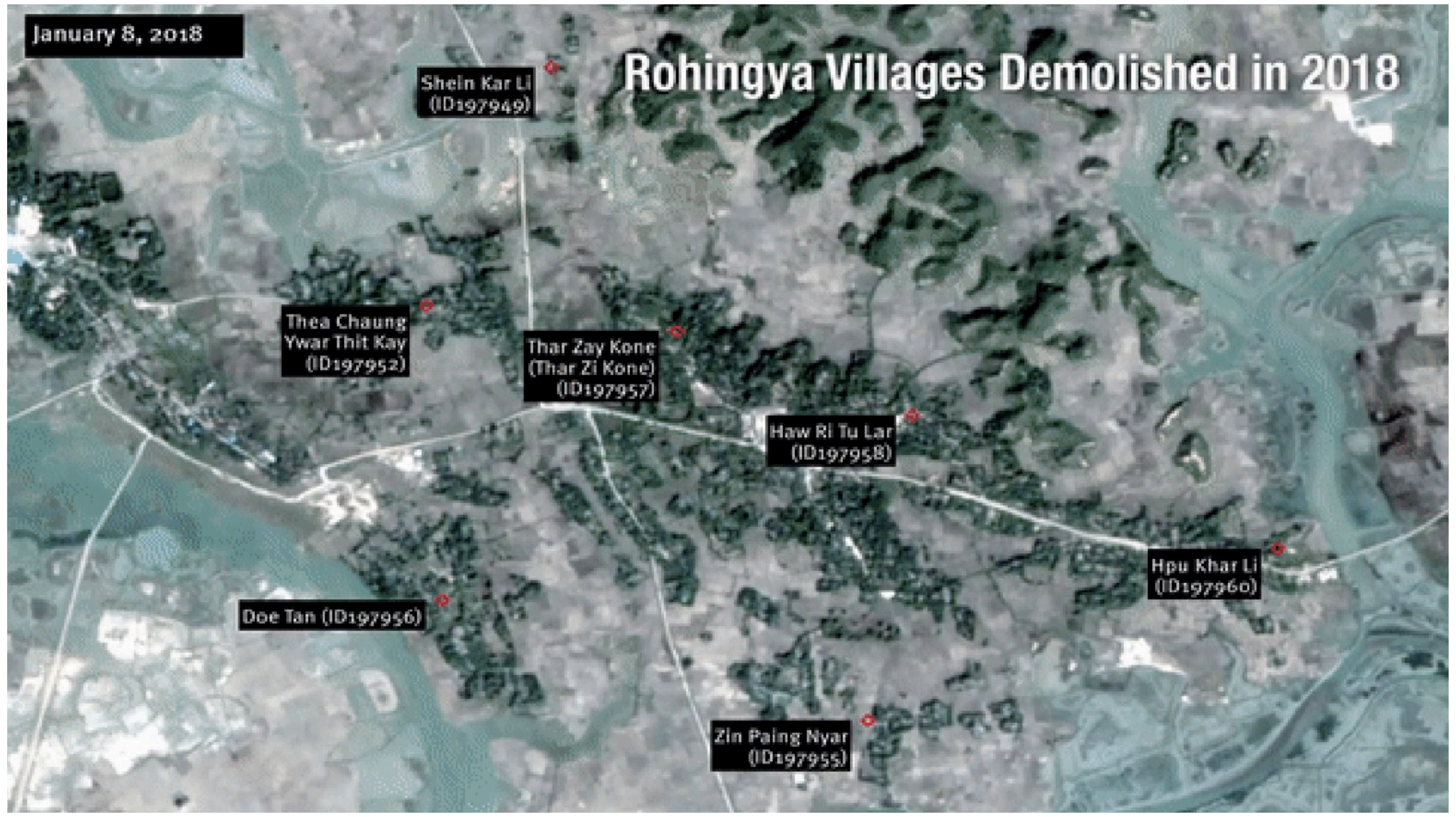
spectrum.ieee.org/how-to-track-the-emissions-of-every-power-plant-on-the-planet-from-space

# ARCHAEOLOGICAL SITE MONITORING

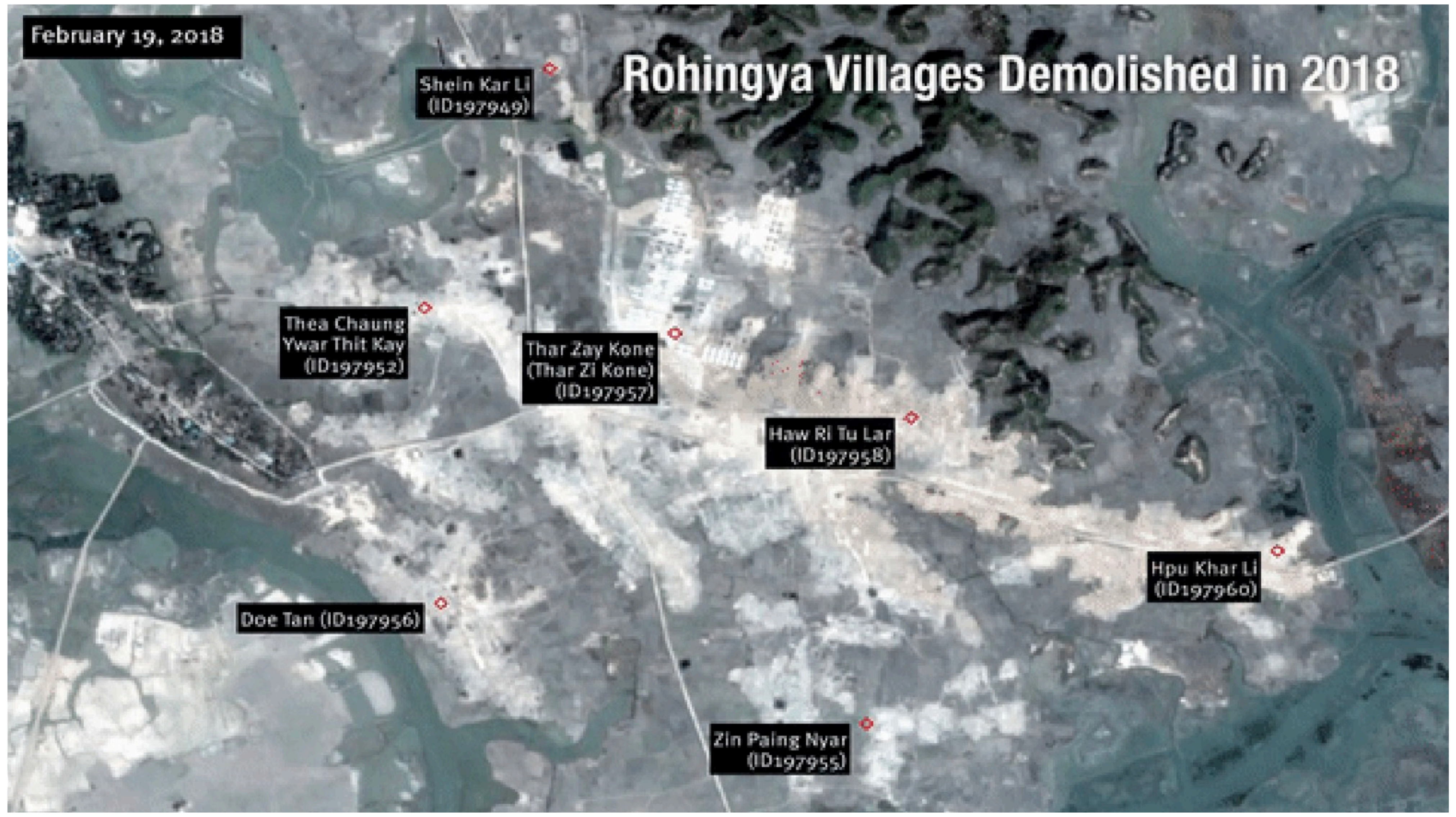


Jesse Casana and Elise Jakoby Laugier, "Satellite Imagery-based Monitoring of Archaeological Site Damage in the Syrian Civil War"
*PLOS One,* 12 (11), November 30, 2017, doi.org/10.1371/journal.pone.0188589

# HUMAN RIGHTS MONITORING




*Myanmar (20.900, 92.366)* *Prior to destruction, Jan. 2018*




*After destruction, Feb. 2018*

Burma: Scores of Rohingya Villages Bulldozed, New Satellite Images Show Destruction Indicating Obstruction of Justice, February 2018
www.hrw.org/news/2018/02/23/burma-scores-rohingya-villages-bulldozed and www.hrw.org/tag/rohingya

*ICBM silo field, under construction; Copernicus Sentinel Data, January 2, 2023 (42.273 N, 92.682 E)*
*fas.org/blogs/security/2021/07/china-is-building-a-second-nuclear-missile-silo-field/*

# ISSUES & CHALLENGES

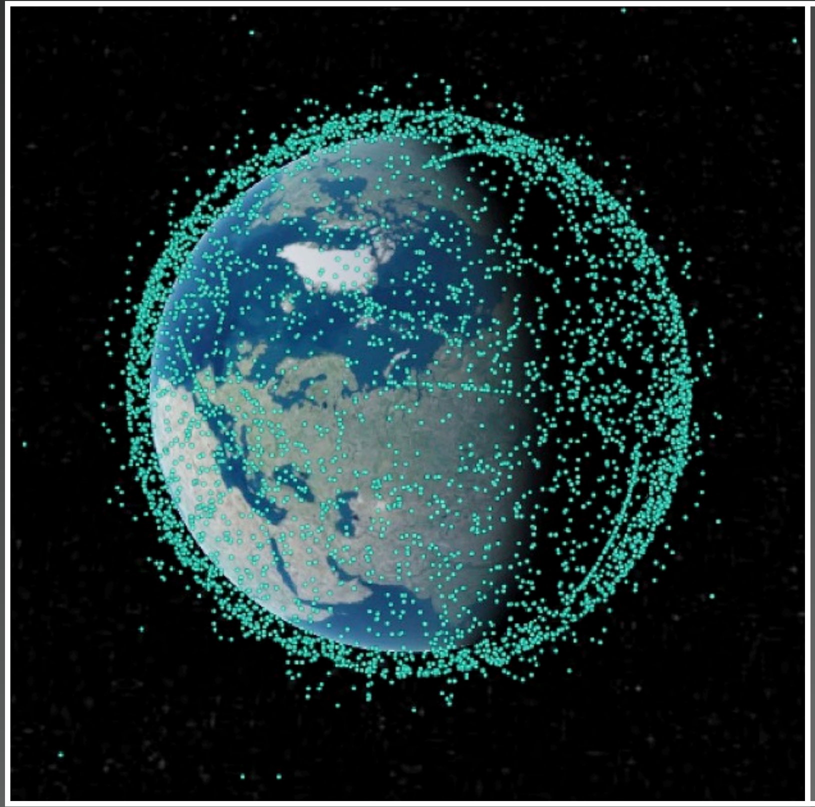
# LACK OF ACCESS TO IMAGERY

*“Analyzing the planet at scale with satellite imagery and machine learning is a dream that has been constantly hindered by the cost of difficult-to-access highly-representative high-resolution imagery.”*

*Julien Cornebise, Ivan Oršolić, and Freddie Kalaitzis, Open High-Resolution Satellite Imagery: The WorldStrat Dataset — With Application to Super-Resolution, July 2022,* arxiv.org/abs/2207.06418

citizenevidence.org/2020/07/06/using-artificial-intelligence-to-scale-up-human-rights-research-a-case-study-on-darfur/

# OVERABUNDANT … BUT ALSO SCARCE

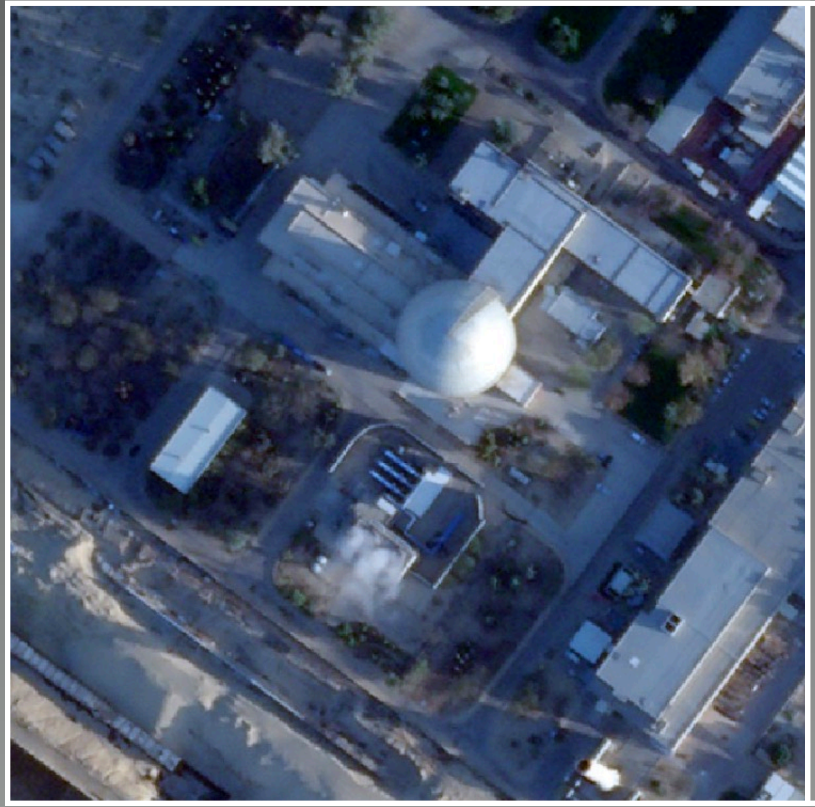


## DATA ARE OVERABUNDANT

Increasing number of vendors, more sensors and bands (optical and radar)

Some efforts underway to index and search this growing archive (for example, bigearth.eu)

Data can only be processed using (machine-learning) algorithms



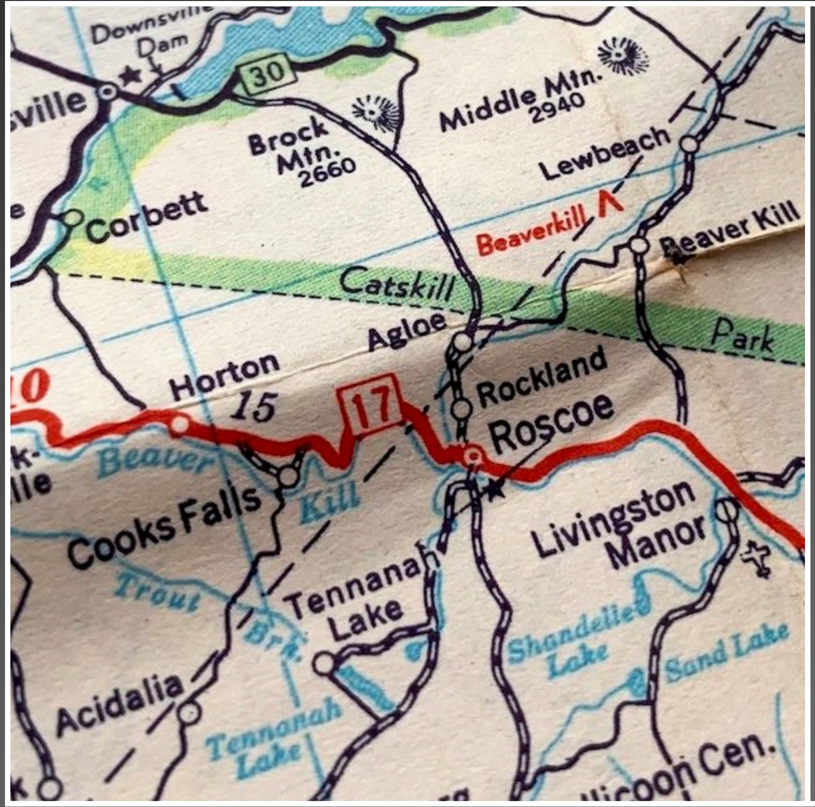
## REPRESENTATIVE DATA CAN BE SCARCE

Depending on use case, very few representative sites/scenes that could be used for training of detection algorithms; high false-positive or false-negative rates likely

Deception efforts possible (unlike in most other use cases)

*Source: wayfinder.privateer.com (top) and Planet Labs (bottom)* mis

# GEOSPATIAL INFORMATION




## GEOSPATIAL MISINFORMATION (THEN)

An old problem; fake locations and other inaccuracies have been part of mapmaking for centuries; including “copyright traps” and “paper towns” as a strategy to thwart plagiarism

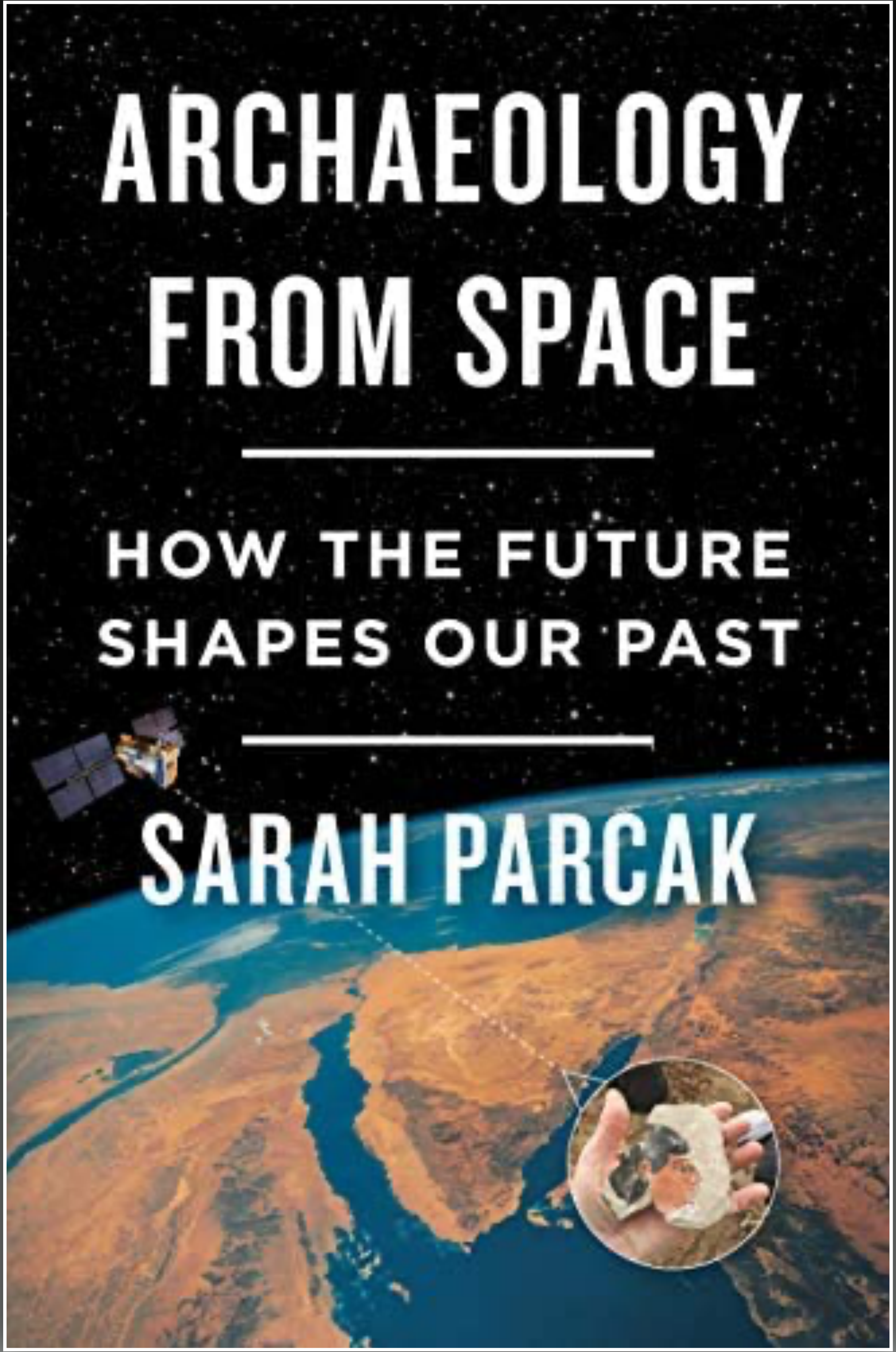
Mark Monmonier, *How To Lie With Maps,* University of Chicago Press, 1996

## GEOSPATIAL MISINFORMATION IN THE AGE OF AI

Few known examples, but circumstantial evidence suggests that AI has been used to manipulate scenes and pixels to create artifacts on satellite imagery for malicious purposes

Bo Zhao, Shaozeng Zhang, Chunxue Xu, Yifan Sun, and Chengbin Deng, “Deep Fake Geography? When Geospatial Data Encounter Artificial Intelligence,” *Cartography and Geographic Information Science,* 2021

*Source: Esso Map, 1956 (top) and Pierre Markuse (medium.com, bottom)*




The New York Times

Opinion | THE PRIVACY PROJECT

# Are We Ready for Satellites That See Our Every Move?

We should consider the ethical implications of satellites that can identify us, and our license plates, from space.

Oct. 15, 2019 4 MIN READ

**By Sarah Parcak**
Dr. Parcak is an archaeologist and remote-sensing expert.

www.nytimes.com/2019/10/15/opinion/satellite-image-surveillance-that-could-see-you-and-your-coffee-mug.html

## QUESTION 1

# Can we generate & use synthetic satellite imagery to improve detection (or other) algorithms?

(when applied to real-world problems/imagery)

## QUESTION 2

# Can we use synthetic imagery to assess the "true" potential of satellites for monitoring & verification?

QUESTION 3

# Can we help support efforts to confirm the authenticity of digital media?

(and, in particular, the provenance & authenticity of satellite imagery)

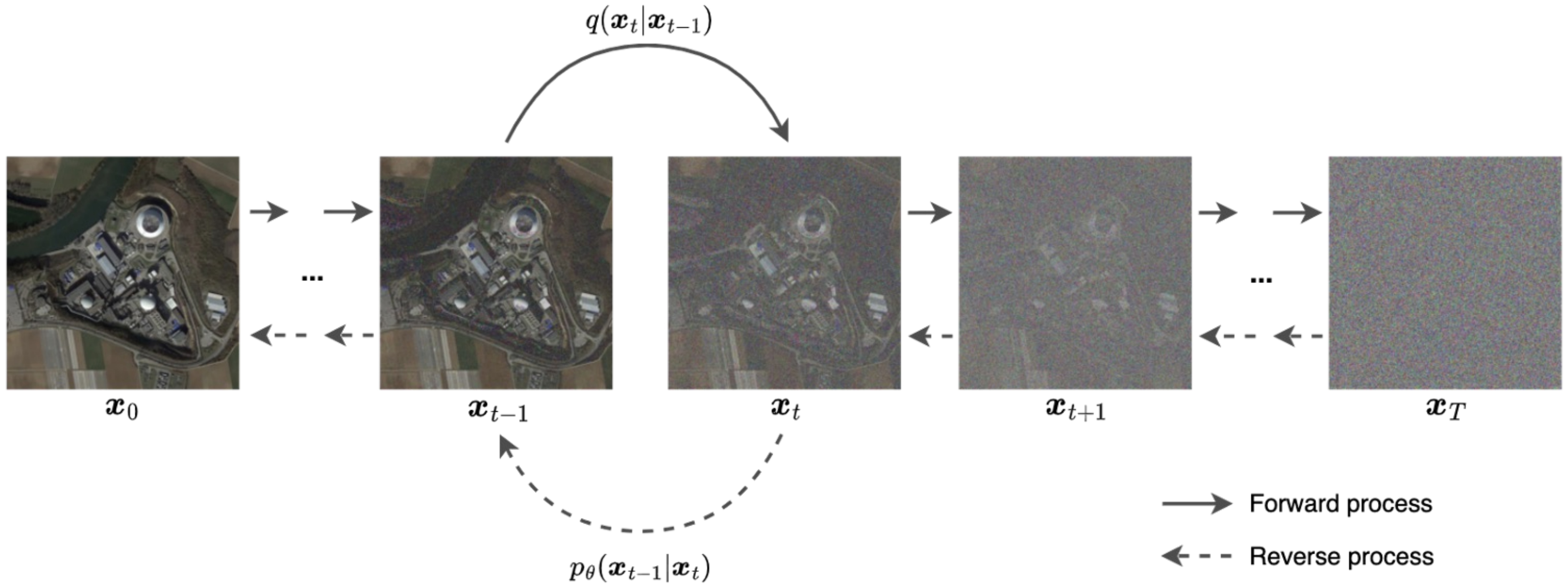
## QUESTION 1

# Can we generate & use synthetic satellite imagery to improve detection (or other) algorithms?

(when applied to real-world problems/imagery)

# Background.

- Recently, **Diffusion Models** have surpassed state-of-the-art GANs in several tasks, notably in image generation.
  The noising process consists of two stages, the forward diffusion and the reverse process.
- **Text-to-image** is an increasingly popular and intuitive approach for conditional image synthesis.
- In the **remote sensing domain**: There are several works on image-to-image translation tasks, but few regarding the generation of novel imagery.




Berliner Hochschule für Technik
Studiere Zukunft

# Background.

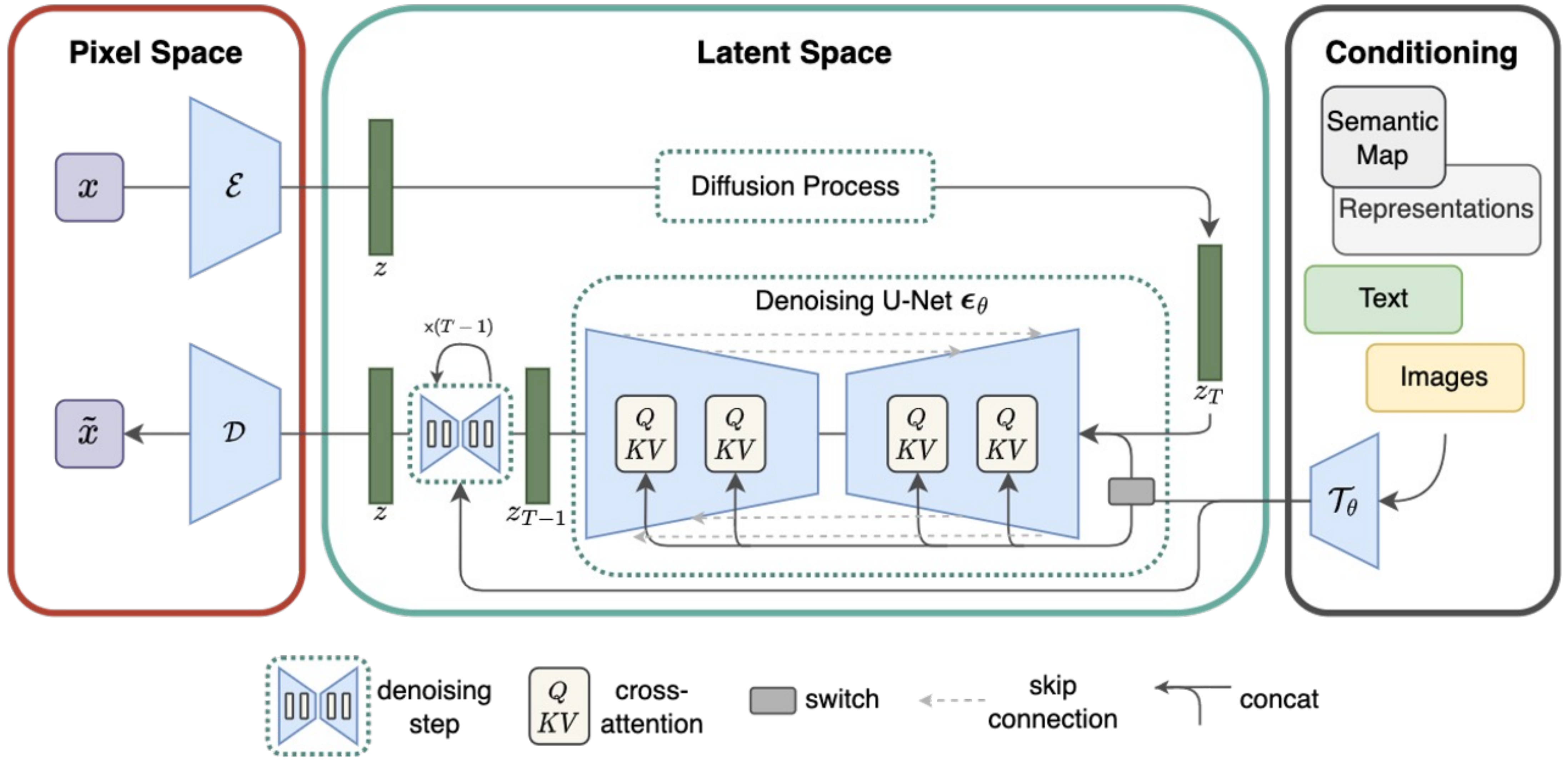
Stable Diffusion:

- Pre-trained text-to-image model based on **Latent Diffusion Models** (LDM)

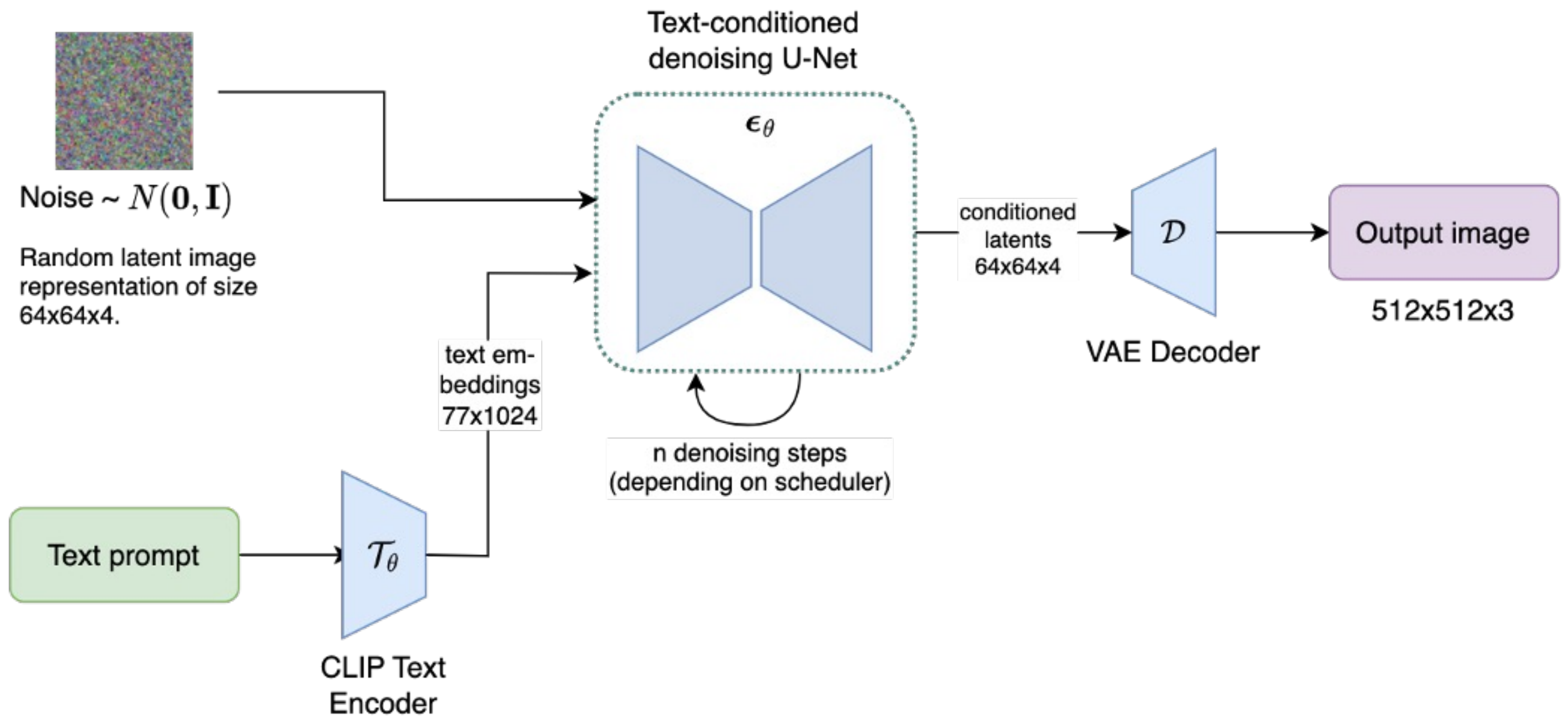
$$L_{LDM} = \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon,t}[||\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))||^2]$$

- Consists of three main components: **VAE**, **CLIP text encoder** and **conditional U-Net**

- Advantages over other models like DALL-E:
  - Code and model weights are **open-source**
  - Relatively **low resource** and memory requirements




Model Architecture of a Latent Diffusion Model.




Inference Process of Stable Diffusion.


**Berliner Hochschule für Technik**
Studiere Zukunft

# Methodology.

Single nuclear power plant in Neckarwestheim.

Six sample images of different nuclear power plants.

Data/Target objects:

- **Nuclear power plants**
    - Scraped using Google Earth Engine (EE), resolution of approx. ~0.8m
    - **Six** training images of a **single site** in Neckarwestheim
    - **202** training images of **185 nuclear power plants** from all over the world

- General land-use classes seen in the **UC Merced (UCM)** benchmark dataset
    - **2100** images (21 classes with 100 images each) with **corresponding text captions**
    - Resolution of approx. ~0.3m


Overview of the 21 different land-use classes in the UCM dataset.

**Berliner Hochschule für Technik**
Studiere Zukunft

# Methodology.

Implementation:

**Baseline**: Unmodified/vanilla Stable Diffusion, using prompt engineering to improve results.

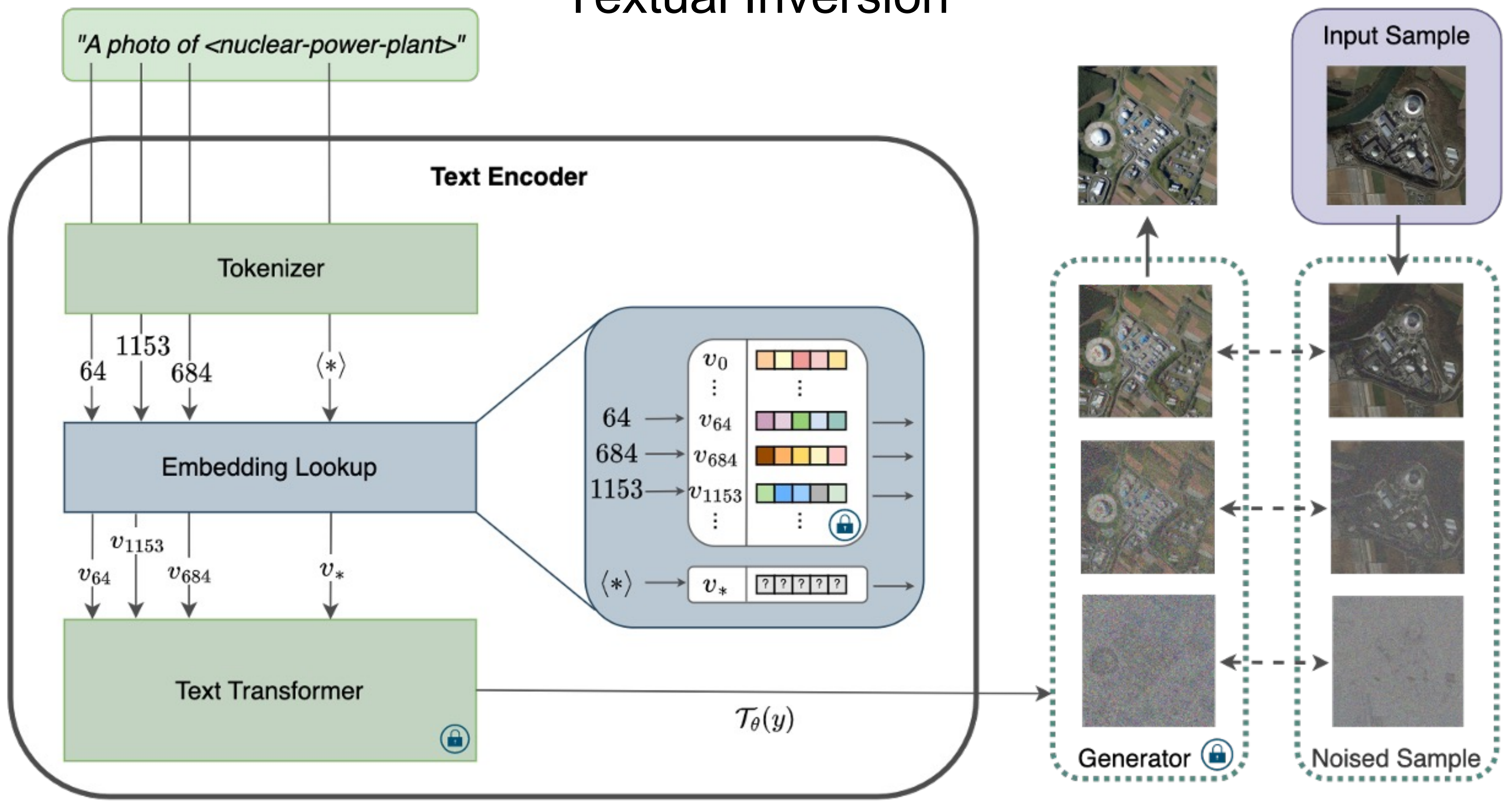
**Fine-tuning**: Implementation of several fine-tuning approaches, namely

- Textual Inversion
- DreamBooth
- Text-to-image fine-tuning

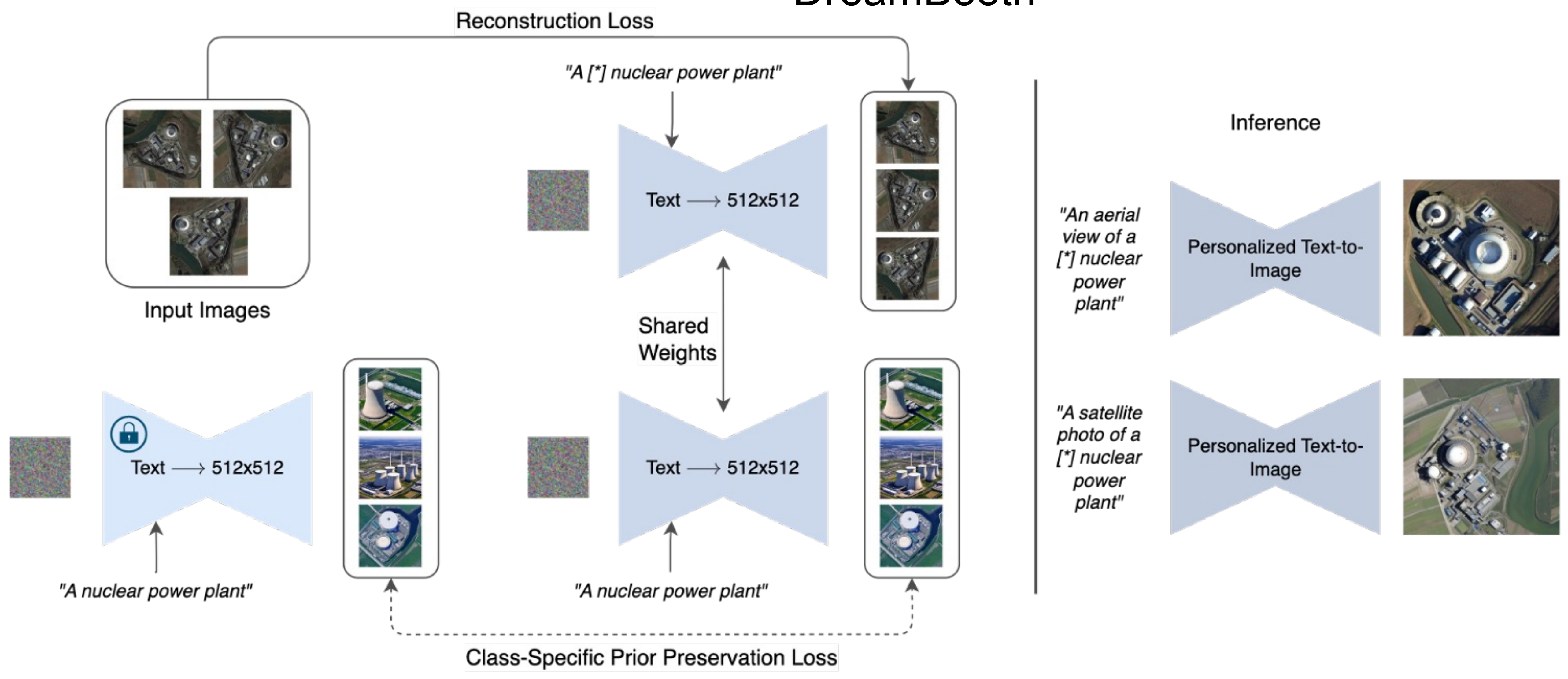
Generated image of a nuclear power plant.

Berliner Hochschule für Technik
Studiere Zukunft

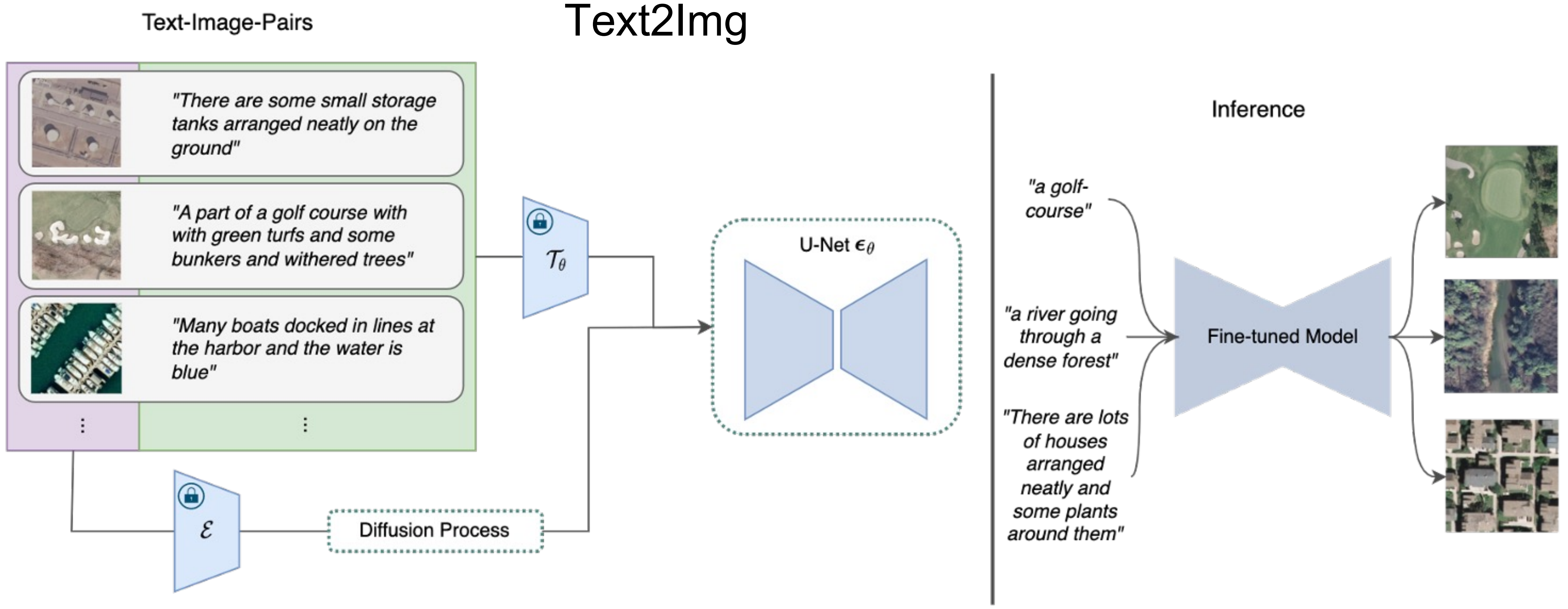
## Textual Inversion




## DreamBooth




## Text2Img




**Berliner Hochschule für Technik**
Studiere Zukunft

# Methodology Details.

# Evaluation.

1. Qualitative evaluation based on **visual assessment**.

2. Quantitative evaluation by applying a variety of state-of-the-art metrics:
   a. Inception Score (**IS**)

   $$\mathrm{IS} = \exp( \mathbb{E}_{x\sim p_\theta}[ D_{KL}(p(y|\boldsymbol{x}) \| p(y))] )$$

   b. Fréchet Inception Distance (**FID**) and
      Fréchet Clip Distance (**FCD**)

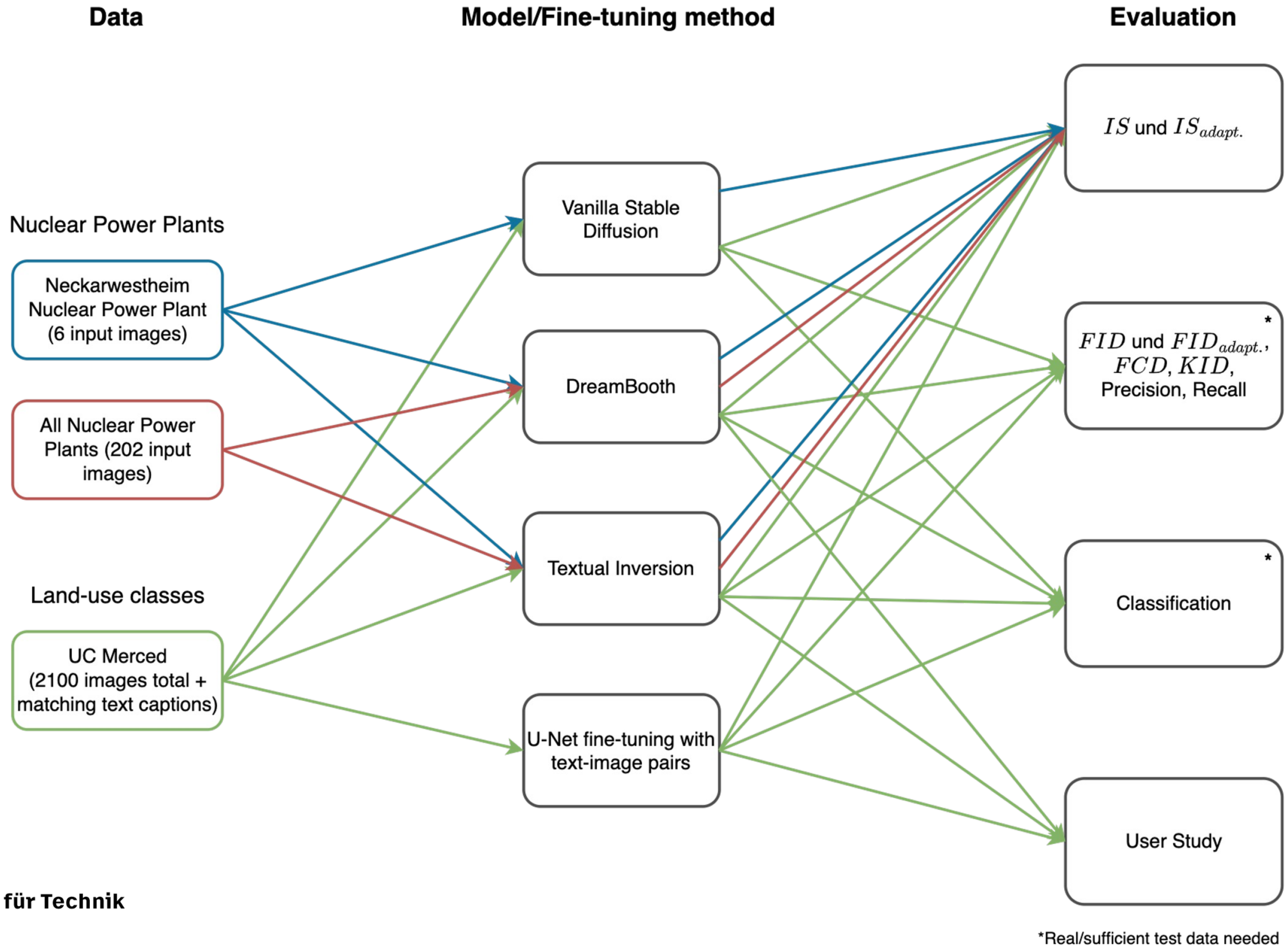
   $$\mathrm{FID} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|^2 + Tr(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{1/2})$$

   c. Modified $\mathbf{IS}_{\mathbf{adapt.}}$ and $\mathbf{FID}_{\mathbf{adapt.}}$ using a different underlying model
   d. Kernel Inception Distance (**KID**)
   **e.** **Precision** and **Recall**

3. For UCM: Conducting an additional **user study** (classify a shown image as real or fake) and applying data in a downstream **classification task** (classifier trained on real UCM data and tested on generated images).

Berliner Hochschule für Technik
Studiere Zukunft

# Overview.



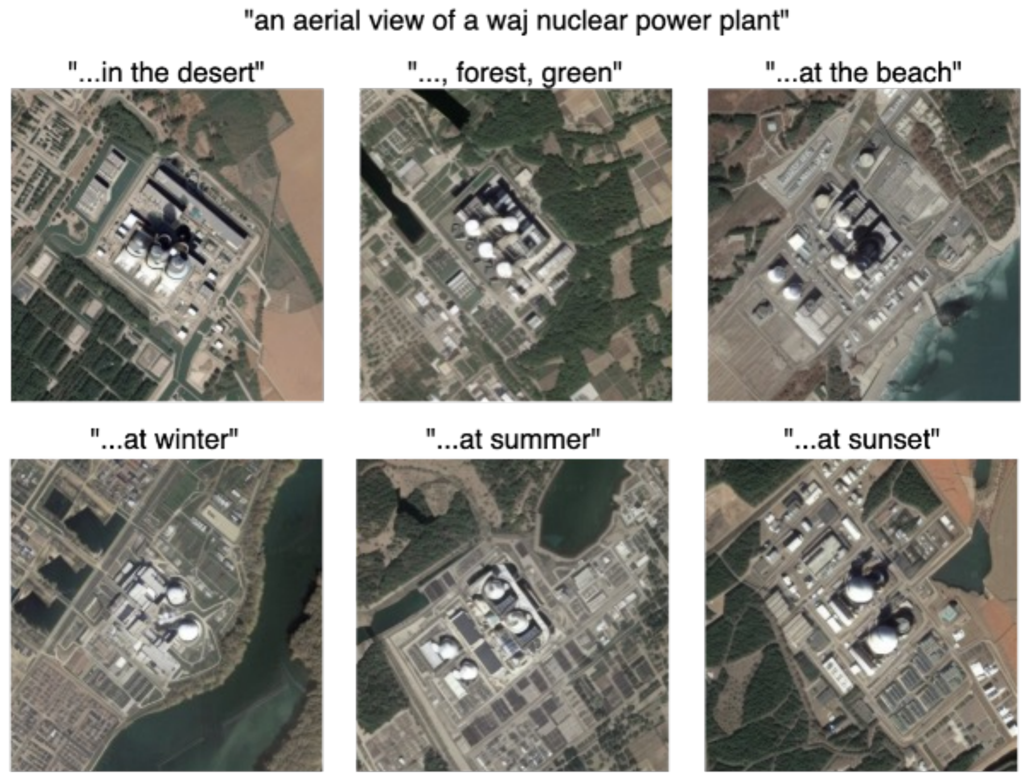

**Berliner Hochschule für Technik**
Studiere Zukunft

# Results and Discussion - Nuclear Power Plants.

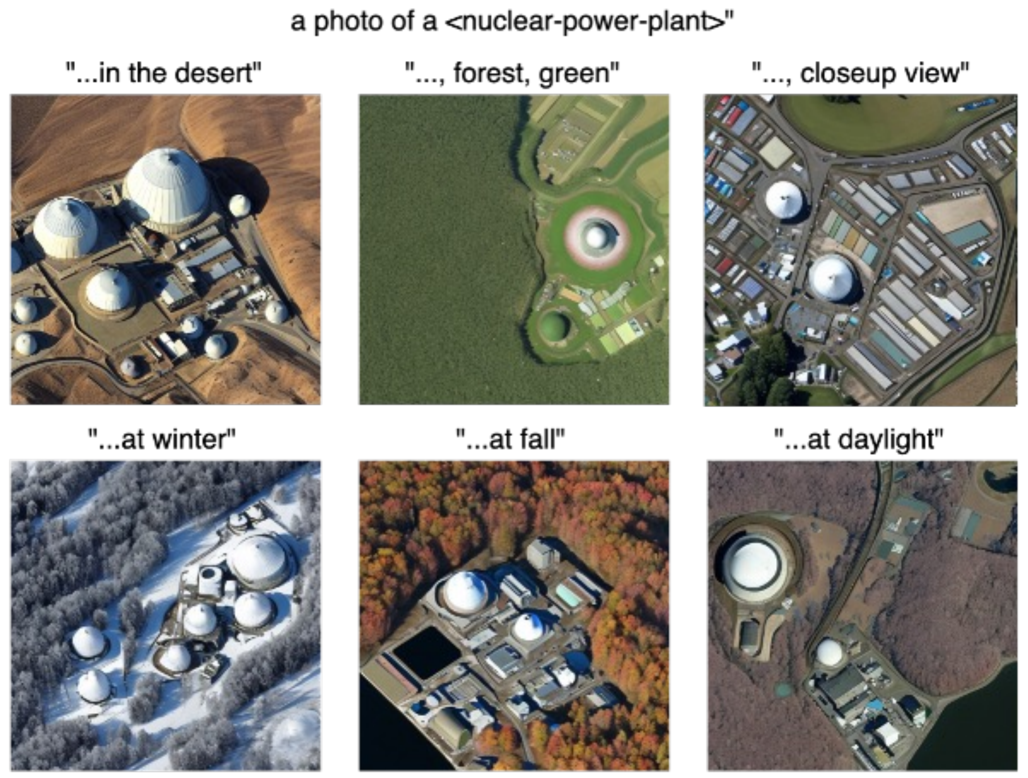
Visual assessment:

- Difficult to achieve satisfactory results using prompt engineering alone
- **Additional conditioning** seems to work better with Textual Inversion (TI)
- DreamBooth (DB) is better at preserving **image fidelity**
- Characteristics like the **viewing angle** can be influenced through the selection of input images


DB Neckarwestheim.




DB All.




TI Neckarwestheim.


TI All.


Vanilla SDiff.

Berliner Hochschule für Technik
Studiere Zukunft

# Results and Discussion - Nuclear Power Plants.

Quantitative evaluation:

**Textual Inversion** models seem to perform the best, DreamBooth trained models the worst.

But: **Difficult to draw conclusions** from the IS and $IS_{adapt.}$ alone:

- **Comparison** to real data is lacking
- Even real images don't achieve best scores
- No indication on how well **text prompts** align with the generated image
- **User study** possibly needed to validate or disprove findings

| Model/Data | $IS_{202} \uparrow$ | $IS_{adapt._{202}} \uparrow$ | $IS_{6000} \uparrow$ | $IS_{adapt._{6000}} \uparrow$ |
|---|---|---|---|---|
| Real train images | 3.12±0.44 | 3.80±0.43 | - | - |
| DB Neckar | 2.84±0.38 | 2.13±0.21 | 3.10±0.08 | 2.26±0.08 |
| TI Neckar | **4.06±0.42** | 3.52±0.93 | **5.53±0.11** | 4.13±0.11 |
| DB All | 2.30±0.26 | 2.49±0.28 | 2.60±0.07 | 2.76±0.05 |
| TI All | 3.36±0.36 | **4.26±0.66** | 4.97±0.13 | **5.39±0.15** |
| Van. SDiff. | 2.99±0.32 | 3.25±0.59 | 3.75±0.09 | 4.03±0.12 |

Berliner Hochschule für Technik
Studiere Zukunft

# Results and Discussion - UCM.

Visual assessment:

- Notable **domain gap** present for vanilla SDiff (SD)
- General idea of the **concepts can be reproduced** for all three fine-tuning methods.
- Image quality depends on the class, **structure** of objects plays a vital role.

- **Prior knowledge** can be leveraged to make changes regarding e.g. seasonality.

T2I.

DB.

TI.

SD.

Real.

# Results and Discussion - UCM: Quantitative evaluation.

- **U-Net component** seems to be the most effective space to further train
- Underlying models regarding metrics might not be suitable
- The similarity between **feature spaces** doesn't necessarily align with human perception
- Synthetic images don't perform as well as the real images. But …
  - … User study: All tested approaches can generate imagery that is, partly, able to **fool the untrained human eye**
  - … Classification task: 80% of generated images can be correctly classified, showing the **potential** of synthetic data
  - … the obtained **robust ranking** regarding model performances aligns with the ranking from the conducted user study

| Model/Data | IS ↑ | $IS_{adapt.}$ ↑ | FID ↓ | $FID_{adapt.}$ ↓ | FCD ↓ | KID ↓ | Precision ↑ | Recall ↑ |
|---|---|---|---|---|---|---|---|---|
| UCM Test ($n = 210$) | 4.90±0.52 | 11.14±1.14 | - | - | - | - | - | - |
| UCM Val | 4.95±0.52 | **10.70±1.28** | **147.05** | **20.09** | **9.95** | **0.00±0.02** | **0.70** | **0.75** |
| Text2Img | 6.23±0.92 | 10.57±1.02 | 175.38 | 27.16 | 16.50 | 0.01±0.02 | 0.57 | 0.43 |
| DreamBooth | 5.10±0.80 | 8.35±0.74 | 185.83 | 37.93 | 17.98 | 0.02±0.02 | 0.42 | 0.54 |
| Textual Inversion | 5.20±0.92 | 7.03±0.89 | 191.77 | 25.53 | 18.93 | 0.02±0.02 | 0.31 | 0.44 |
| Vanilla Stable Diffusion | **6.27±0.60** | 6.43±0.87 | 243.24 | 79.61 | 45.61 | 0.05±0.02 | 0.03 | 0.48 |
| UCM Val+Test ($n = 420$) | 5.85±0.68 | 13.78±1.30 | - | - | - | - | - | - |
| Text2Img | 6.86±0.80 | **13.01±0.85** | **139.65** | 23.32 | **12.62** | **0.01±0.02** | **0.56** | 0.38 |
| DreamBooth | 5.99±0.44 | 10.40±0.71 | 171.69 | 35.67 | 15.99 | 0.01±0.02 | 0.34 | 0.27 |
| Textual Inversion | 6.08±0.73 | 8.10±0.70 | 177.61 | **22.13** | 16.75 | 0.02±0.02 | 0.33 | 0.17 |
| Vanilla Stable Diffusion | **7.55±1.23** | 7.27±0.71 | 207.41 | 65.55 | 41.15 | 0.05±0.02 | 0.02 | **0.48** |

Berliner Hochschule für Technik
Studiere Zukunft

# Conclusion.

A large pre-trained vision-language model can be **fine-tuned** to fit a specific domain, the **prior knowledge** allows for additional conditioning.

Synthetic data can obtain evaluation scores of the same order of magnitude as real data and able to **fool the human eye**.

Reliable **quantitative measures** are important and require further research.

Berliner Hochschule für Technik
Studiere Zukunft

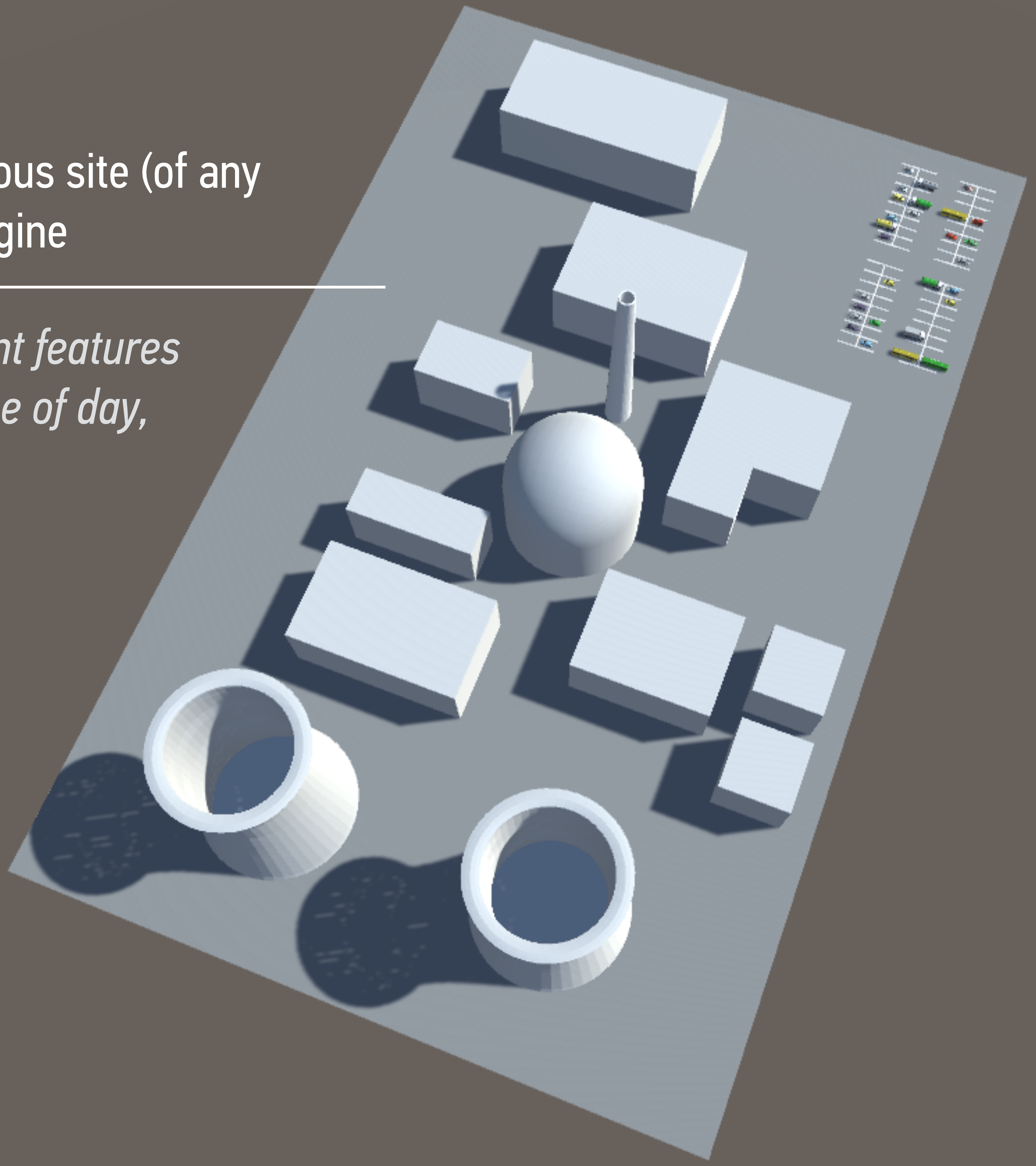
# ANOTHER APPROACH

(Game Engines)

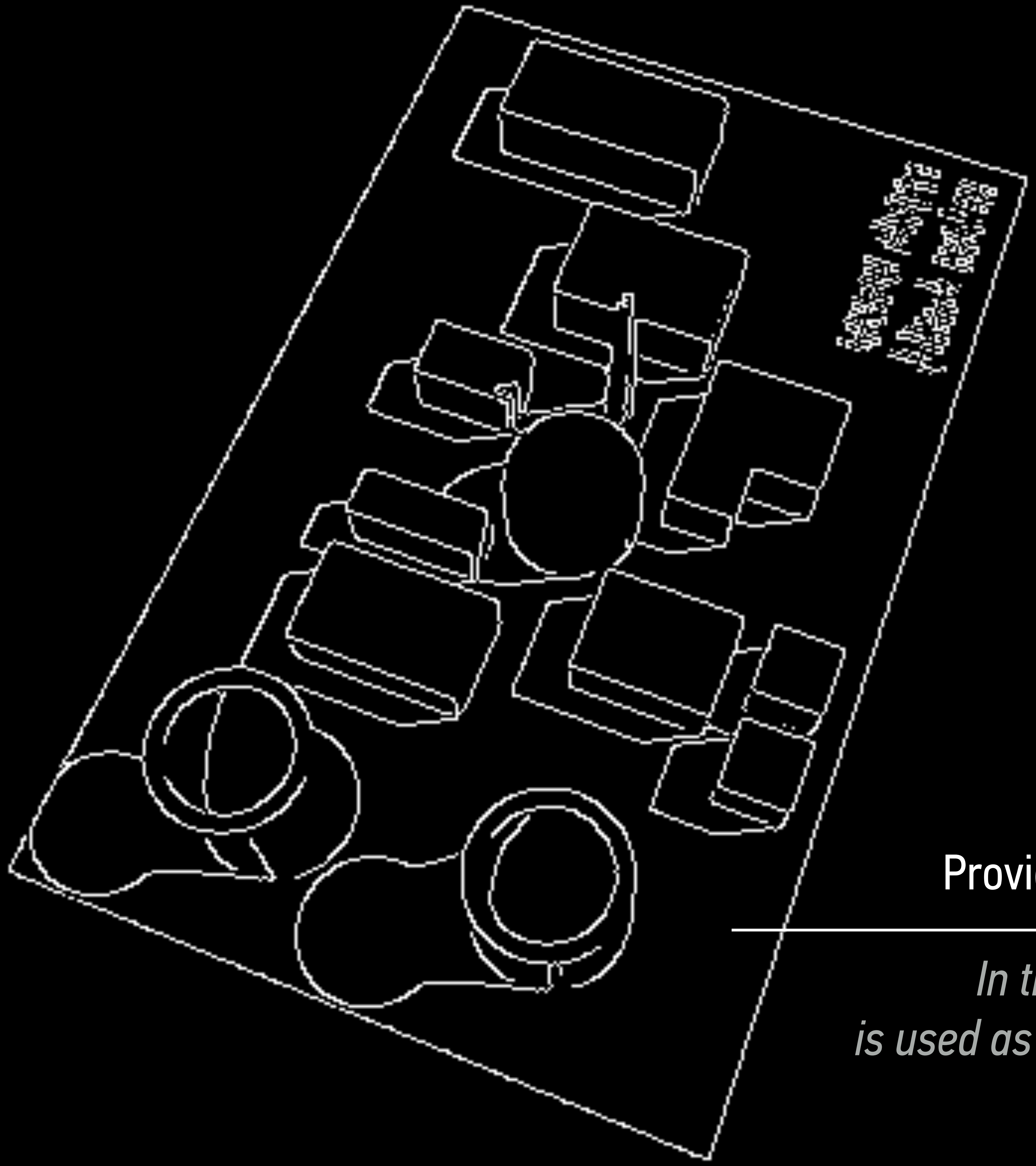
(led by Johannes Hoster and Kristian Hildebrand)

Procedurally generate layout of a fictitious site (of any desired type) using a modern Game Engine

*Game Engine enables control of relevant features of scene, including: level of activity, time of day, cloud coverage, off-nadir angle, etc.*



*Courtesy: Johannes Hoster, BHT*



Provide input modalities for structural guidance

*In this example, the "canny edge" of the scene is used as an additional modality for a text-to-image composable adapter ("T2I CoAdapter")*

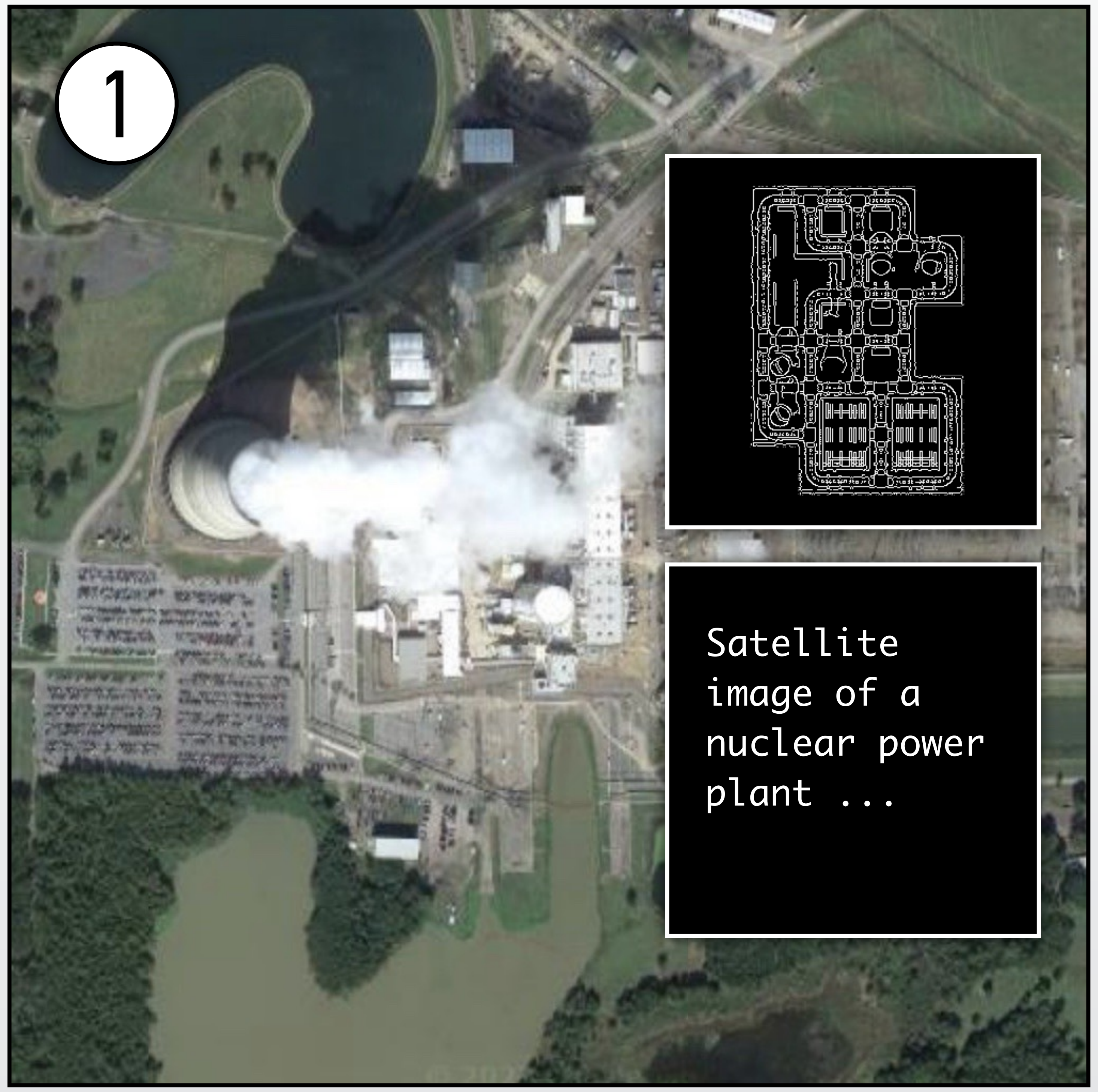
*The canny edge complements the style image and the text prompt provided to the diffusion model*

*Courtesy: Johannes Hoster, BHT*


Courtesy: Johannes Hoster, BHT

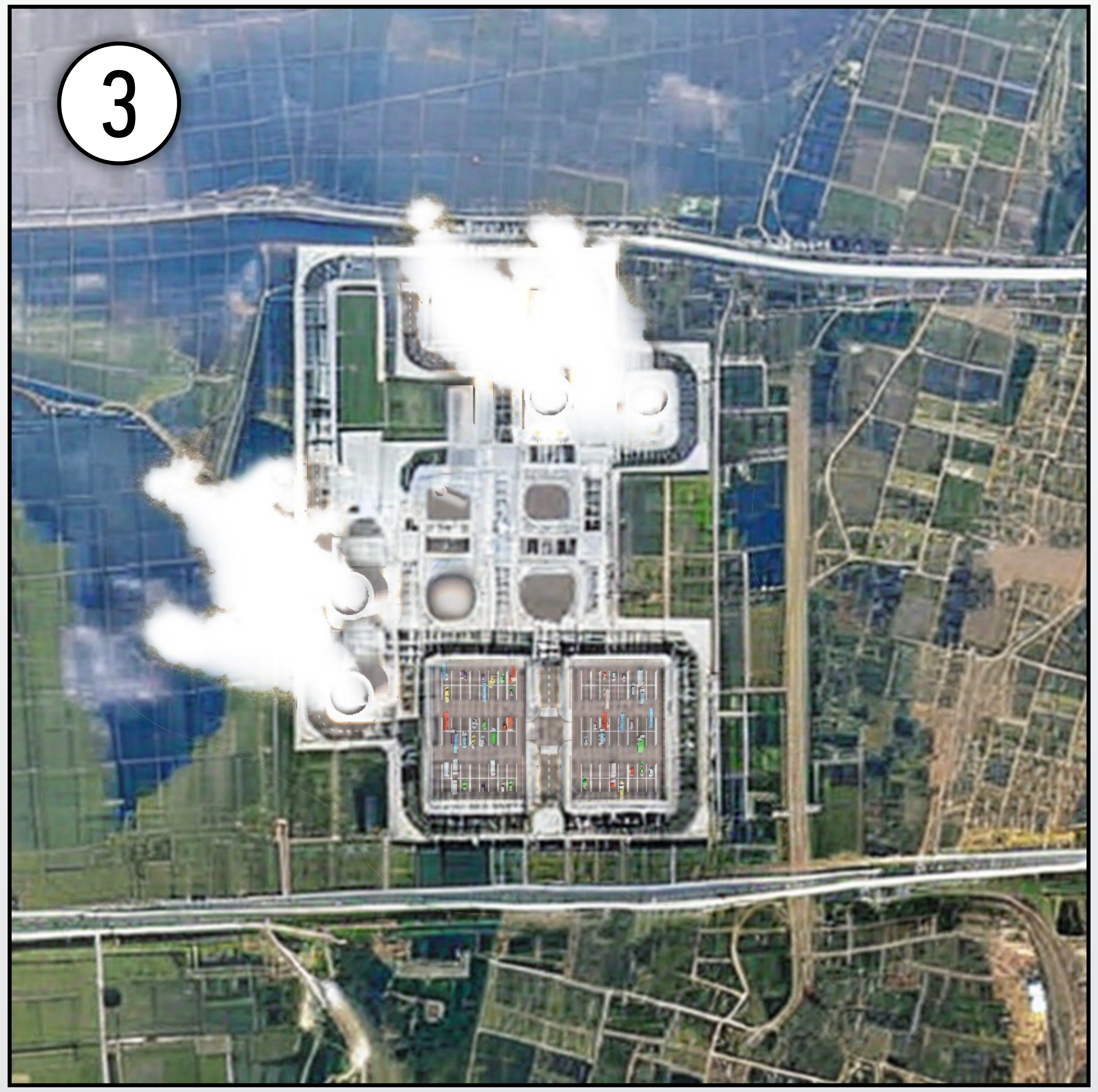
# USING GAME ENGINES & MACHINE LEARNING

## TO CREATE SYNTHETIC SATELLITE IMAGERY




Satellite imagery of real
nuclear power plant


Synthesized image
(with colormap of reference imagery)




Final image with details
from game-engine render included

J. Hoster, S. Al-Sayed, F. Biessmann, A. Glaser, K. Hildebrand, I. Moric, and Vy Nguyen, *INMM & ESARDA Joint Annual Meeting,* Vienna, May 2023

## QUESTION 2

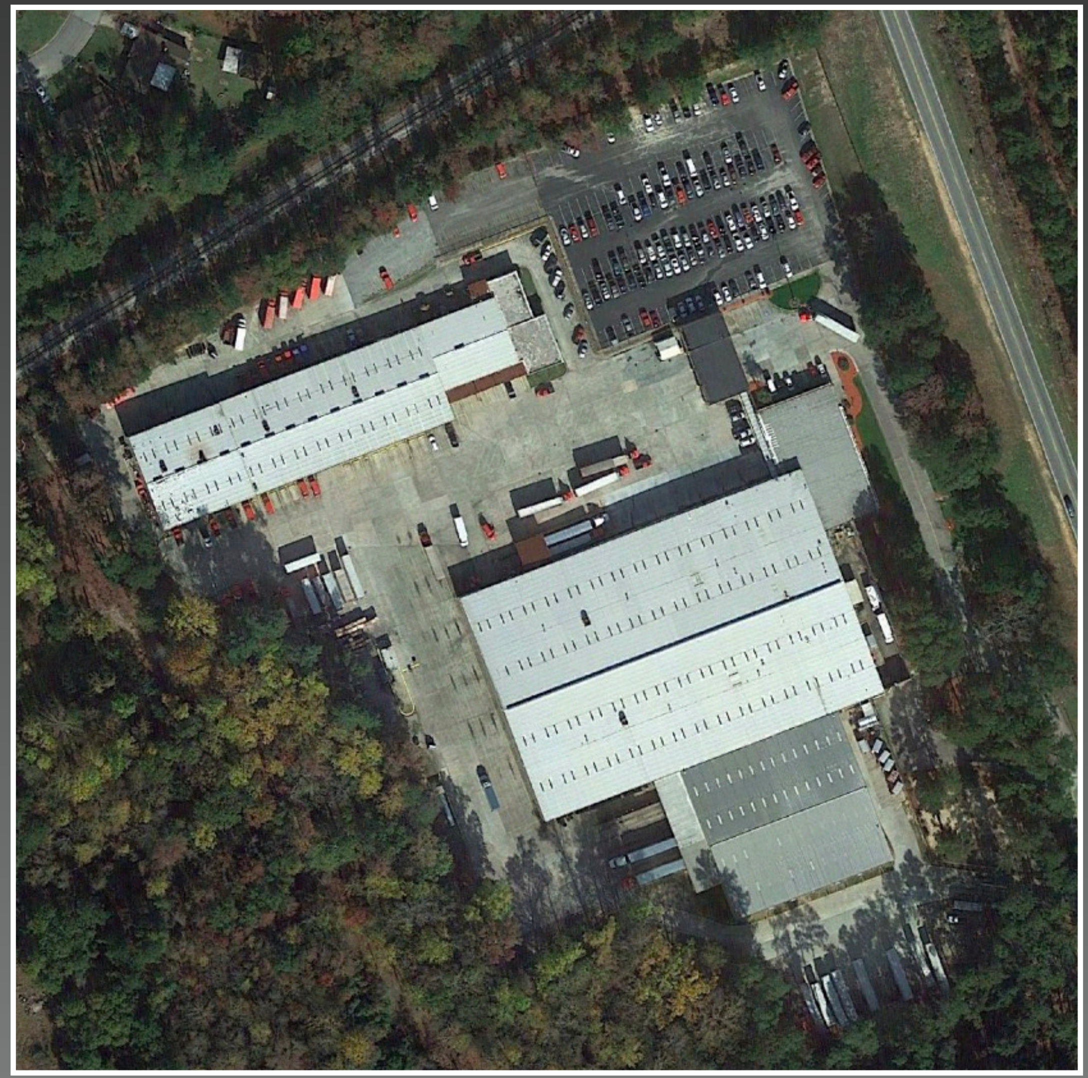
# Can we use synthetic imagery to assess the "true" potential of satellites for monitoring & verification?

*Fordow Enrichment Plant, Iran, in January 2016 (34.885 N, 50.996 E)*
*Iran's second enrichment plant was disclosed in September 2009; the plant itself is underground*

# "PATTERN OF LIFE ANALYSIS"

## UNDERSTANDING A SITE'S "BEHAVIOR" AND ITS RELATIONSHIP TO OTHER SITES






Beverage (bottling) facility, Atlanta, Georgia (33.4582 N, 82.0686 W)

*Source: Google Earth; see also: www.planet.com/pulse/what-is-rapid-revisit-and-why-does-it-matter*

## QUESTION 3

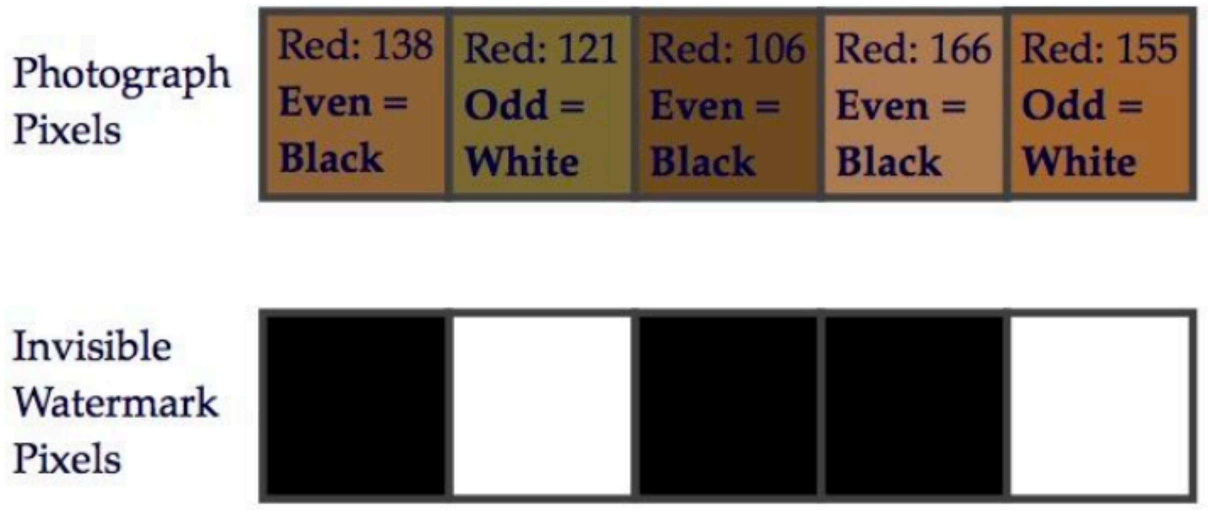
# Can we help support efforts to confirm the authenticity of digital media?

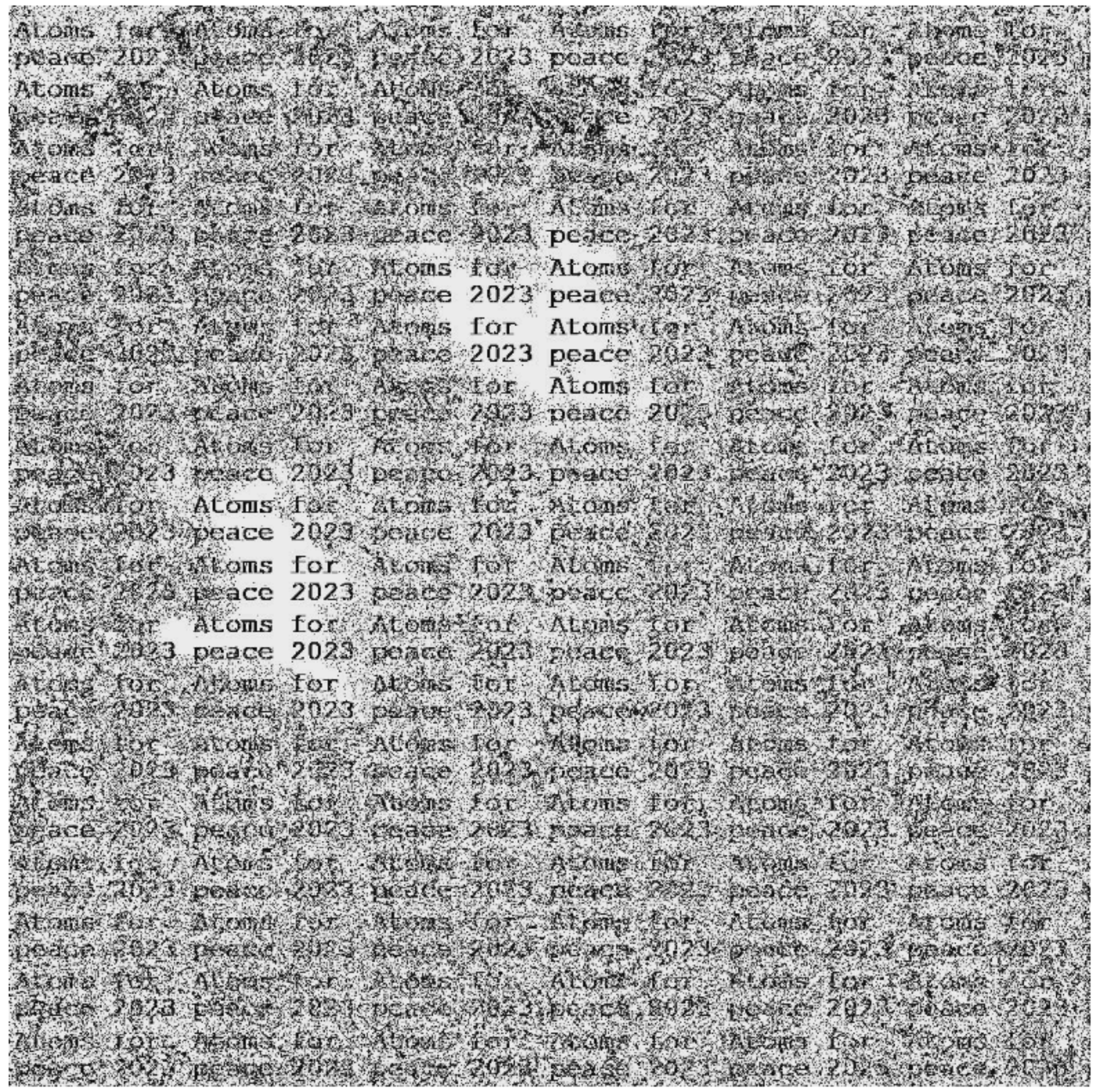
(and, in particular, the provenance & authenticity of satellite imagery)

# WATERMARKING SYNTHETIC MEDIA IS “EASY”

## BUT IT DOES NOT REALLY ADDRESS (SOME) KEY CONCERNS ABOUT MISINFORMATION

Image with invisible watermark



Retrieved watermark
“Atoms for peace 2023”

*Source: invisiblewatermark.net (courtesy Johannes Hoster)*

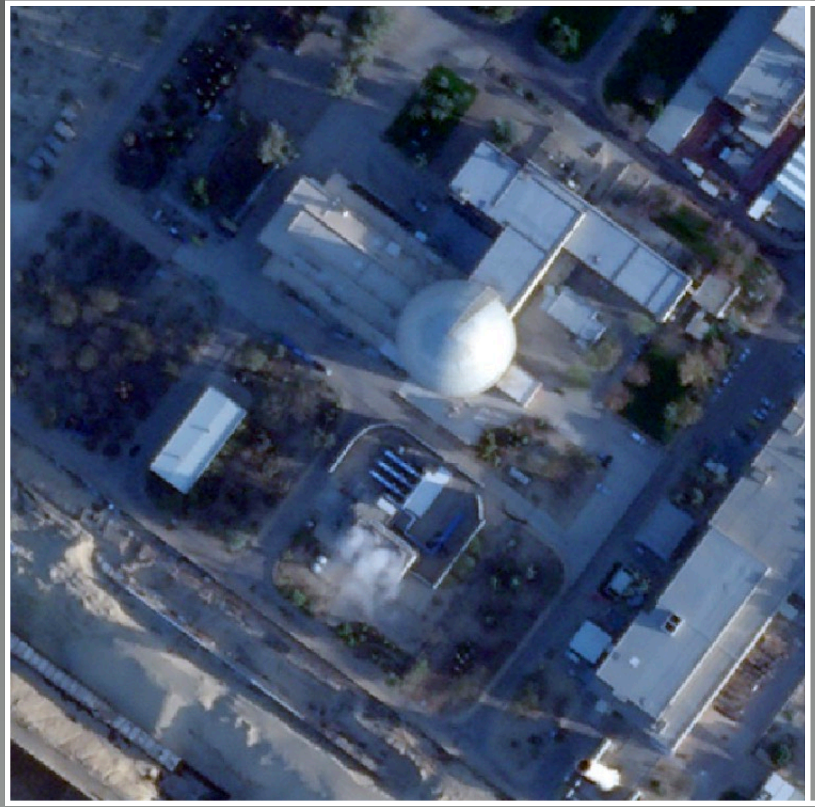
# DIGITAL CONTENT PROVENANCE & AUTHENTICITY


## WHAT TO WATERMARK: SYNTHETIC AND/OR AUTHENTIC MEDIA?

Ideally, watermark all authentic media; harder for some types of media than for others

Some industry efforts underway

- Coalition for Content Provenance and Authenticity (C2PA, c2pa.org)
  Led by Adobe; members include Microsoft, Intel, Arm, but also Canon, Nikon, and many others



## SOME PRINCIPLES & CRITERIA FOR WATERMARKING OF DIGITAL MEDIA

- Security and robustness, i.e., watermarks that are resilient to manipulation
- Privacy, i.e., ability to control the privacy of information, including the identity of the source
- Scalability and flexibility, i.e., standards ought to be applicable to all common and future media types
- Universality and accessibility

*See also: c2pa.org/principles*

*Source: www.natezeman.com (top) and Planet Labs (bottom)*

# CONCLUDING THOUGHTS

## A NEW ERA OF GLOBAL TRANSPARENCY?

There is a widely shared expectation—or hope—that broad access to open-source information will enable the timely detection of non-compliance with relevant international agreements

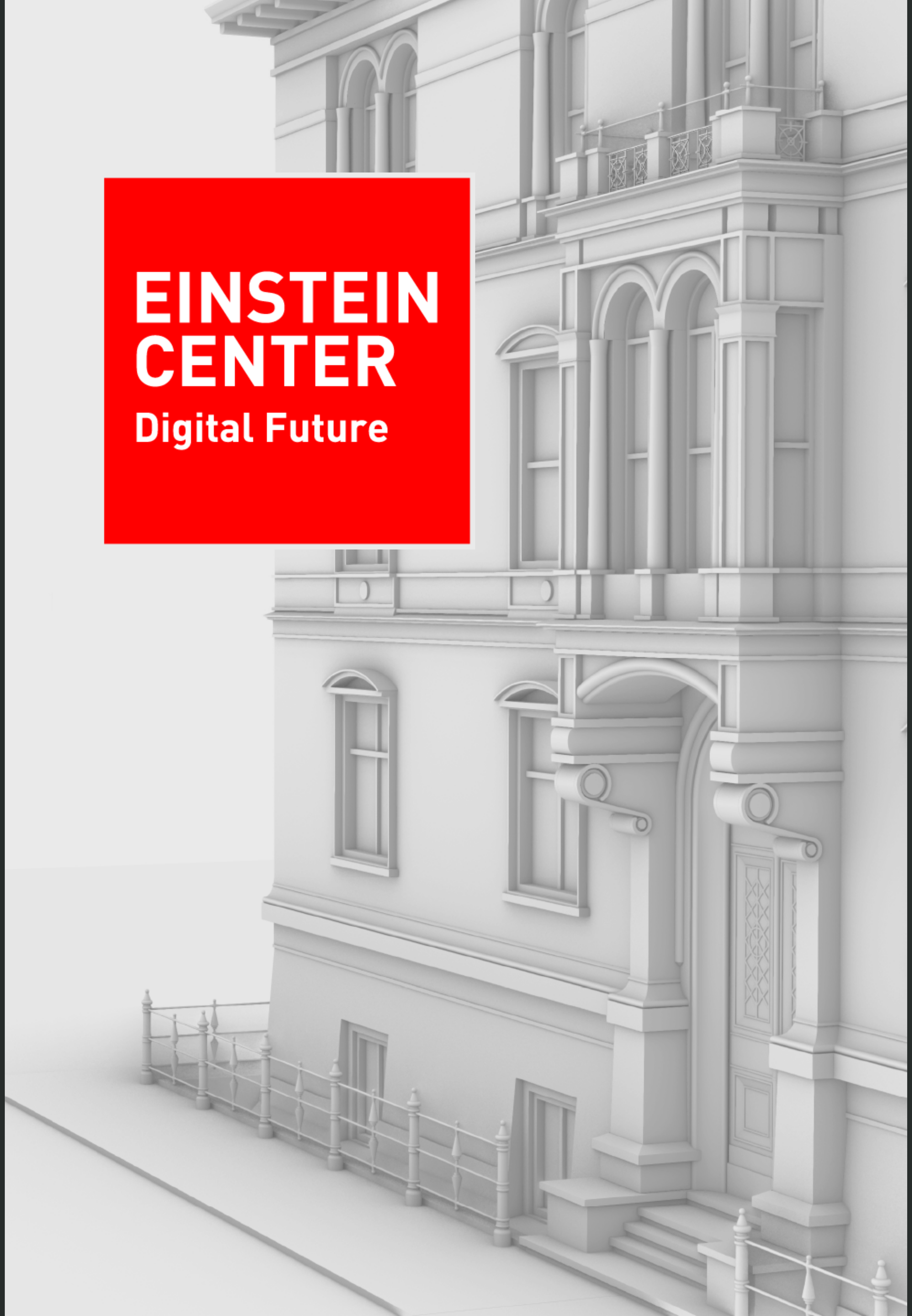
In reality, there are major obstacles to overcome to achieve this vision

## SYNTHETIC MEDIA ARE HERE TO STAY

Just like in the case of spam, malware, or phishing, “we should prepare ourselves for an equally protracted battle to defend against various forms of abuse perpetrated using generative AI.” (Hany Farid, The Conversation, March 2023)

*Source: Google Earth (top) and Chris Umé (bottom)*

BHT
Berliner Hochschule für Technik



EINSTEIN CENTER
Digital Future

SCIENCE & GLOBAL SECURITY
PRINCETON UNIVERSITY