# Machine Learning based Turbulence Prediction in Urban Areas

## CIMPA Summer School 2026 | Course Part 1

**Maximilian Dauner**
**maximilian.dauner0@hm.edu**
Munich University of Applied Sciences
Munich Center for Digital Sciences and AI (MUC.DAI)

MUC.DAI
Munich Center for
Digital Sciences and AI

## Agenda – 22. May

Uses cases & my research topic

Introduction to AI

AI-Topic: Dataset

Lab 1
From PDE Simulations to Training Data

## Agenda – 23. May

Short recap

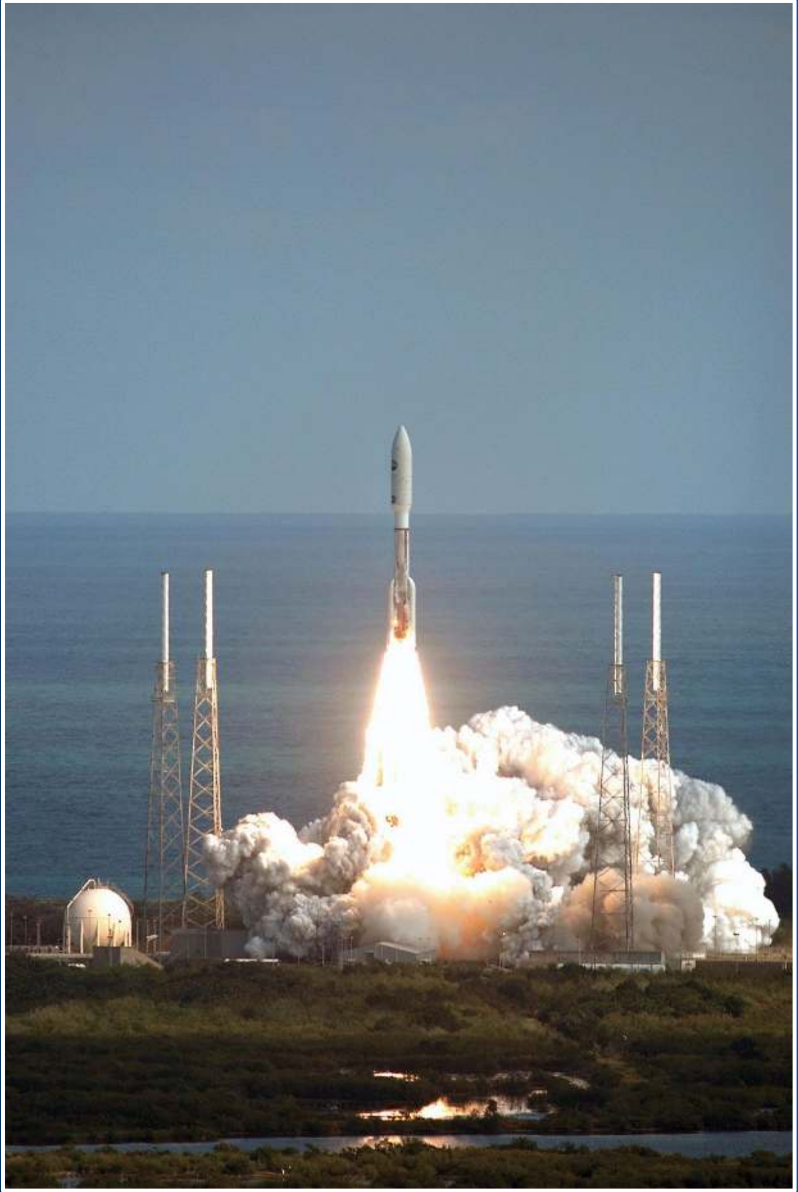
Introduction to neural network and training

Topic: (Fourier) Neural Operator Setup

Lab 2
Train your own Neural Operator


A 1413


Vittorio

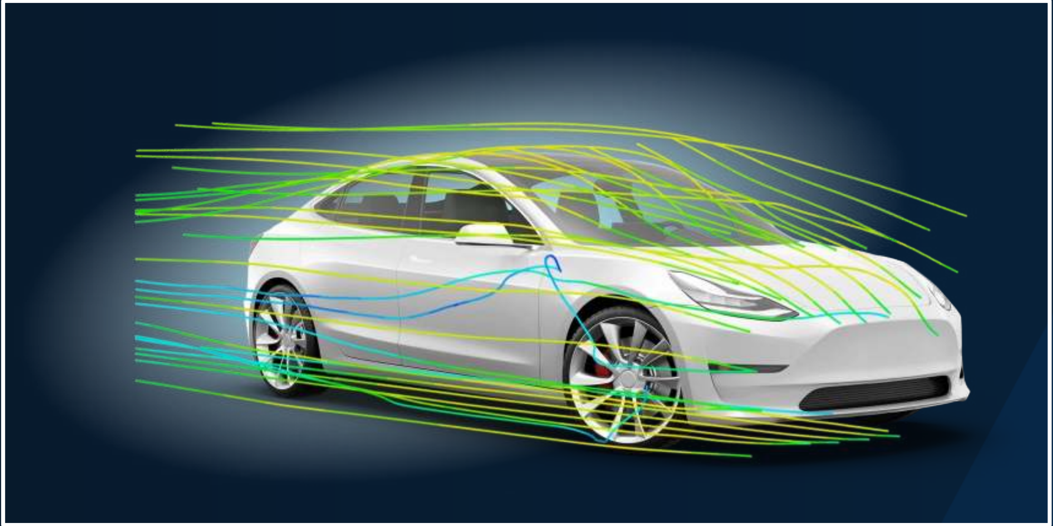


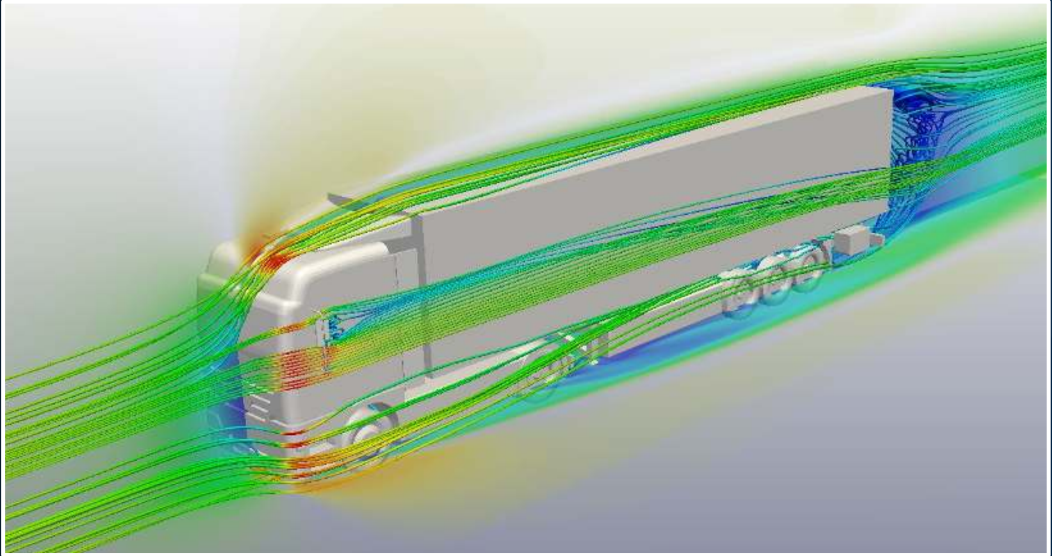
# How do we currently simulate turbulent flows?

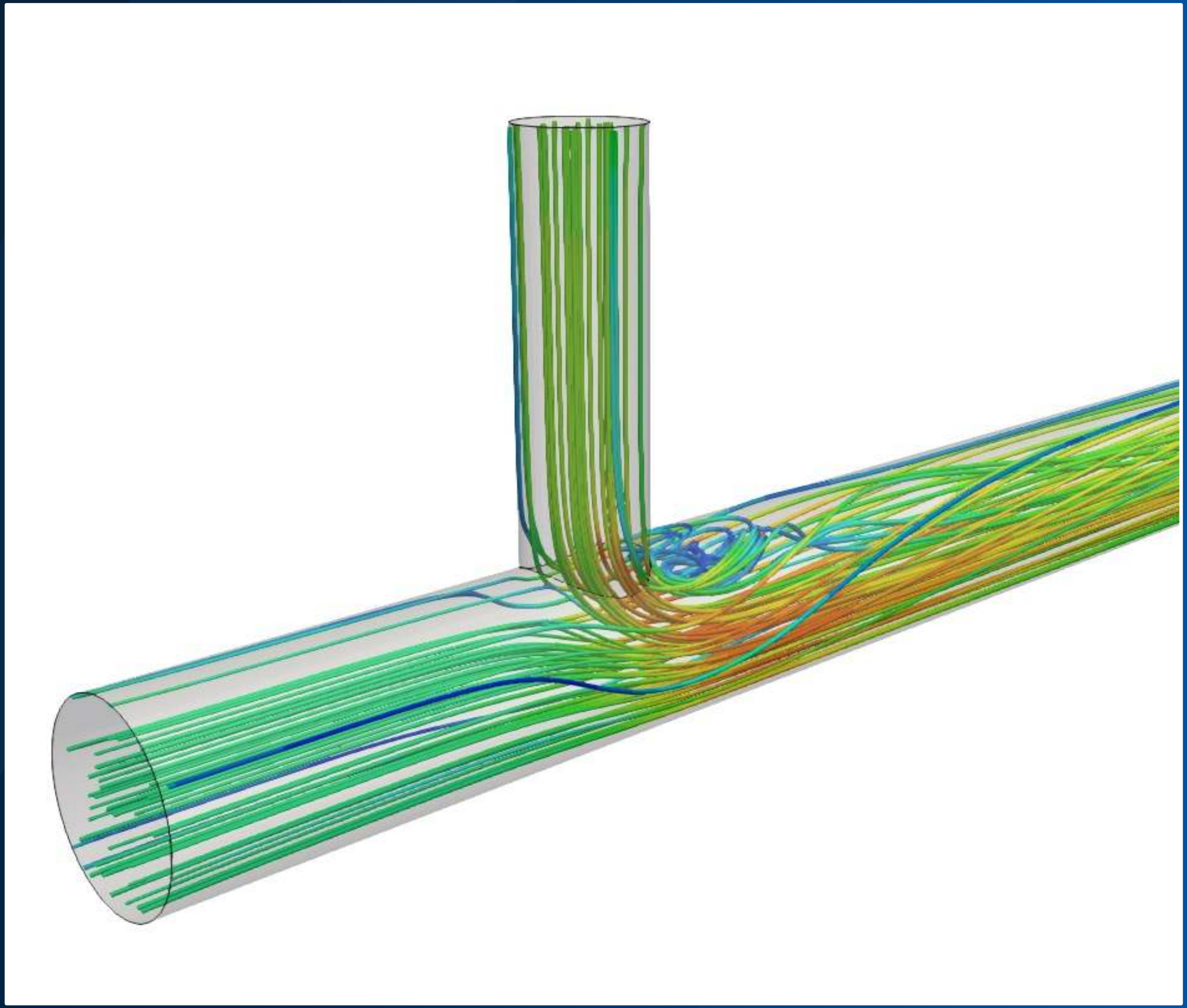
First, we need to understand how physical systems are traditionally simulated.

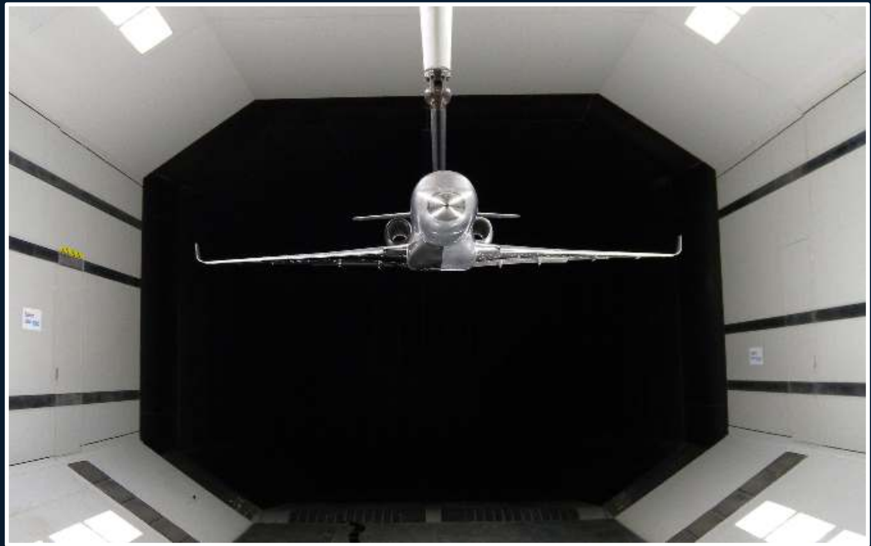
Today, turbulent flows are mainly studied through **computer simulations** and **physical experiments**. In simulations, we describe the flow with mathematical equations and solve them step by step. In experiments, we measure real flow behavior, for example in wind tunnels or test environments.

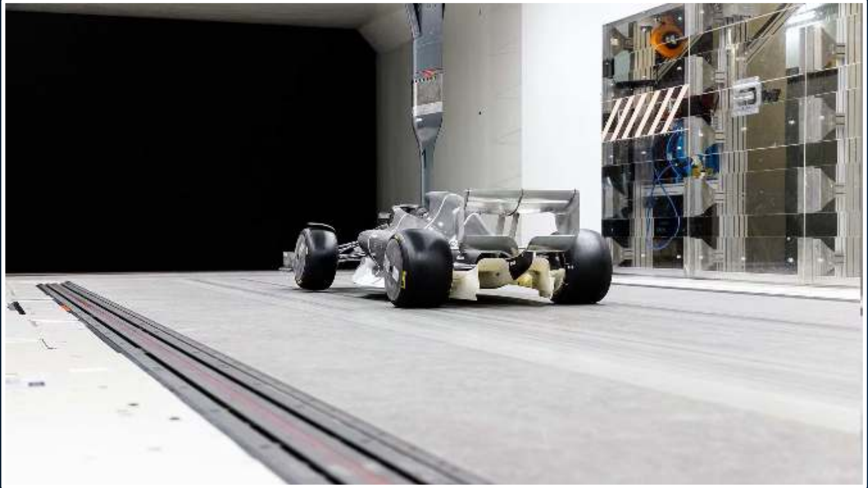
Modern turbulence simulation is always a trade-off between **physical accuracy, computational cost, geometry complexity and boundary conditions**.












CAE
CAE7000XR

# How could AI improve this process?

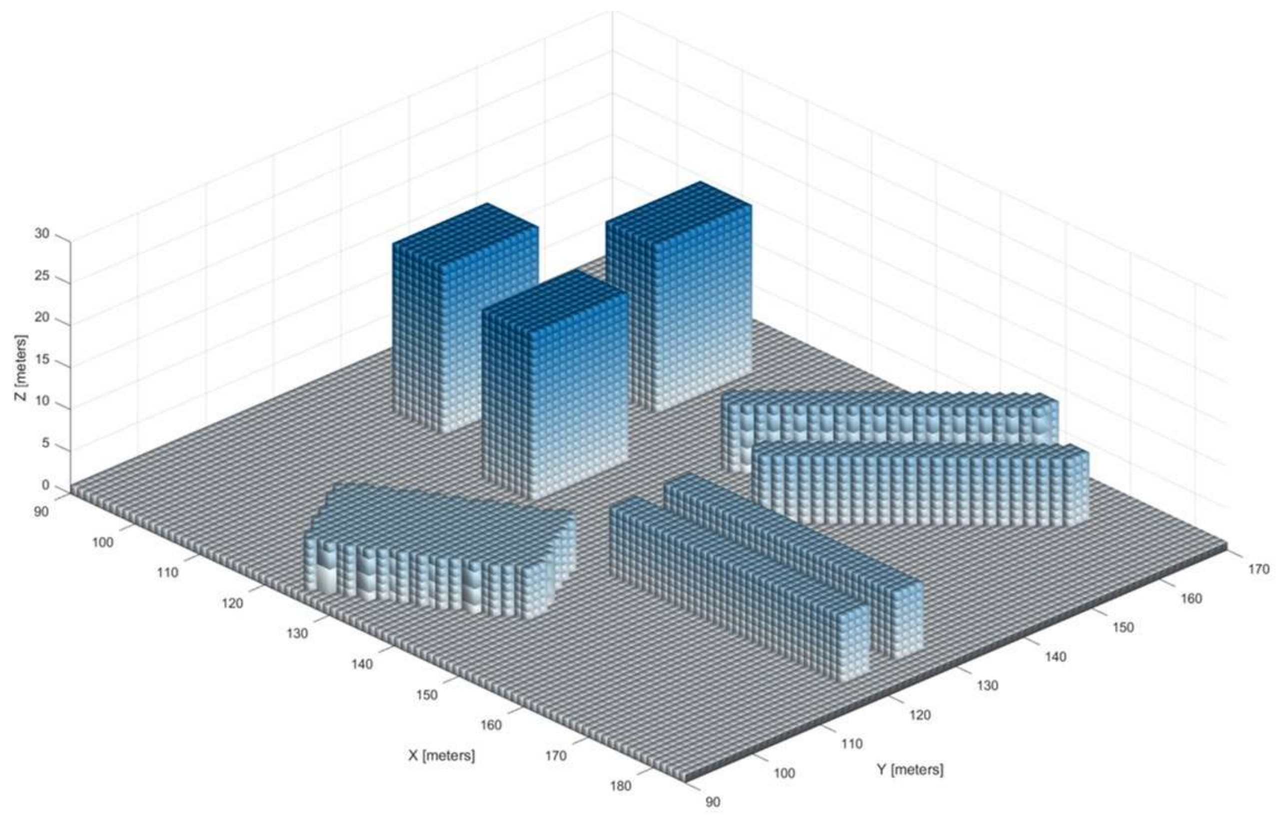
AI can support existing simulation workflows.

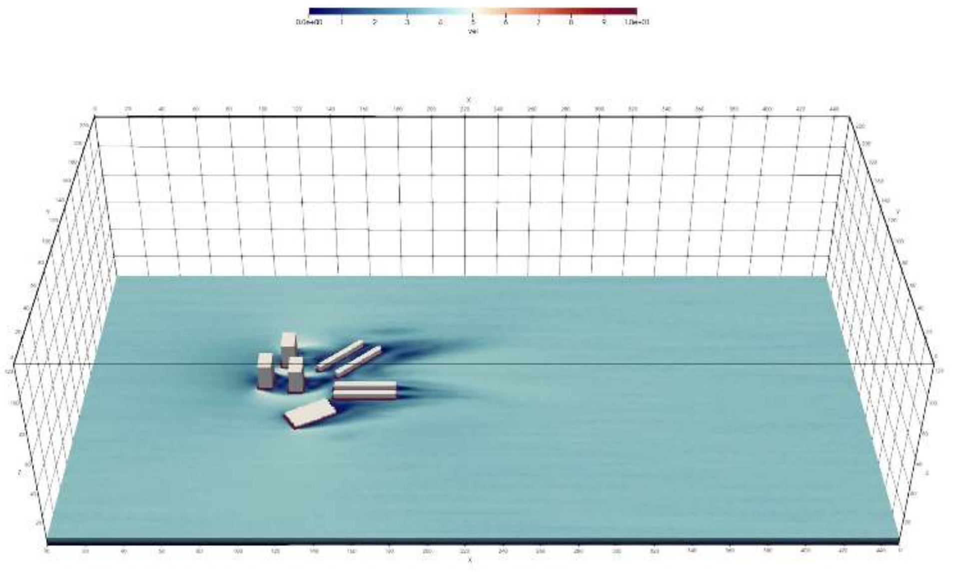
AI can learn from existing simulation and measurement data to create **fast surrogate models** that approximate parts of the flow prediction process much faster than running a full CFD simulation every time.

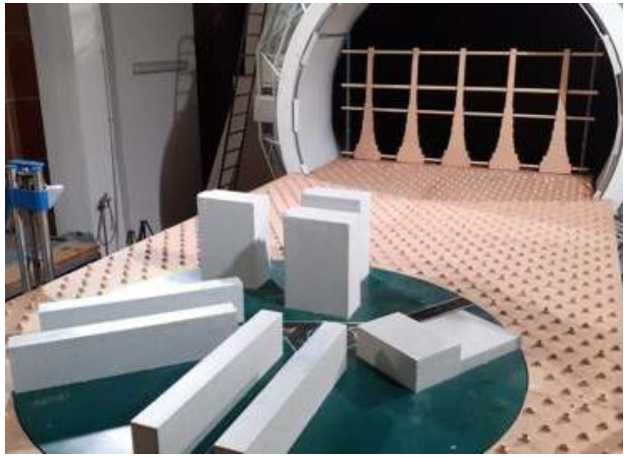
The main advantage is not that AI replaces physics or CFD simulations, but that it can **accelerate repeated predictions, support design exploration, and make complex flow information available faster**.

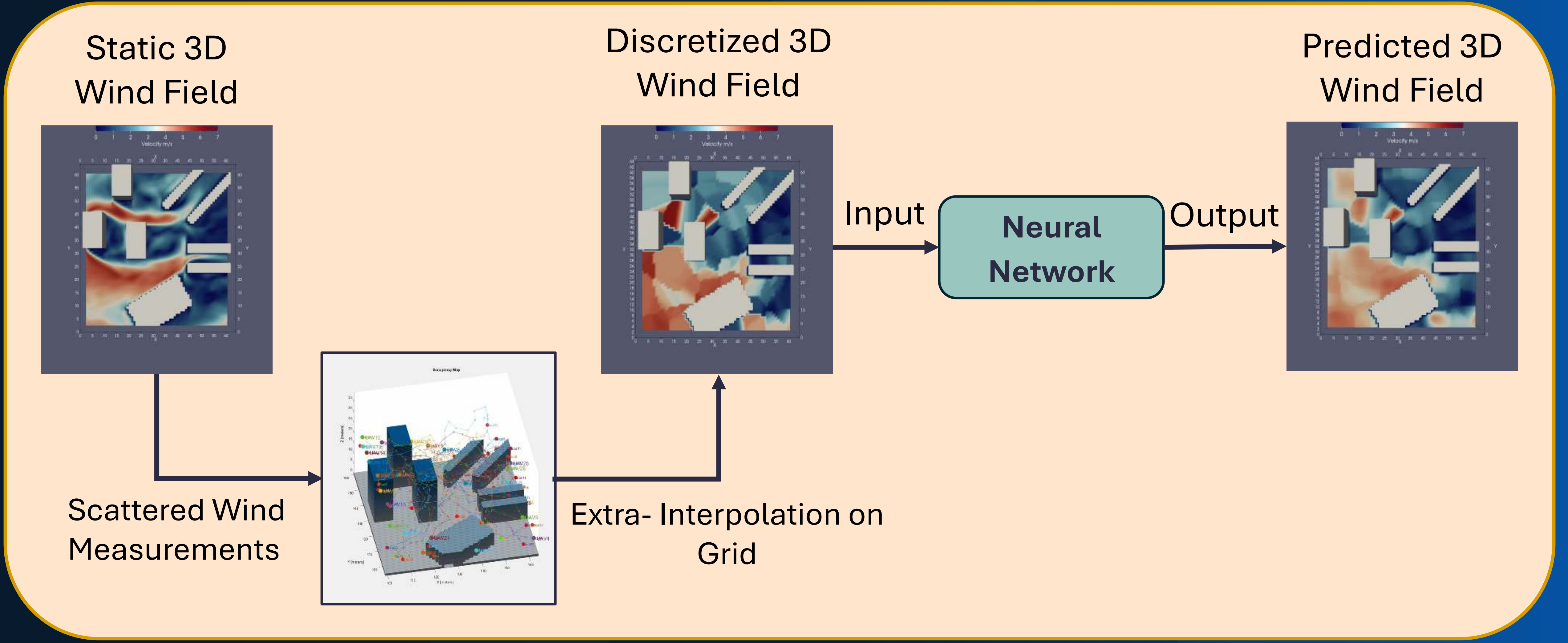
# Where does my research fit in?

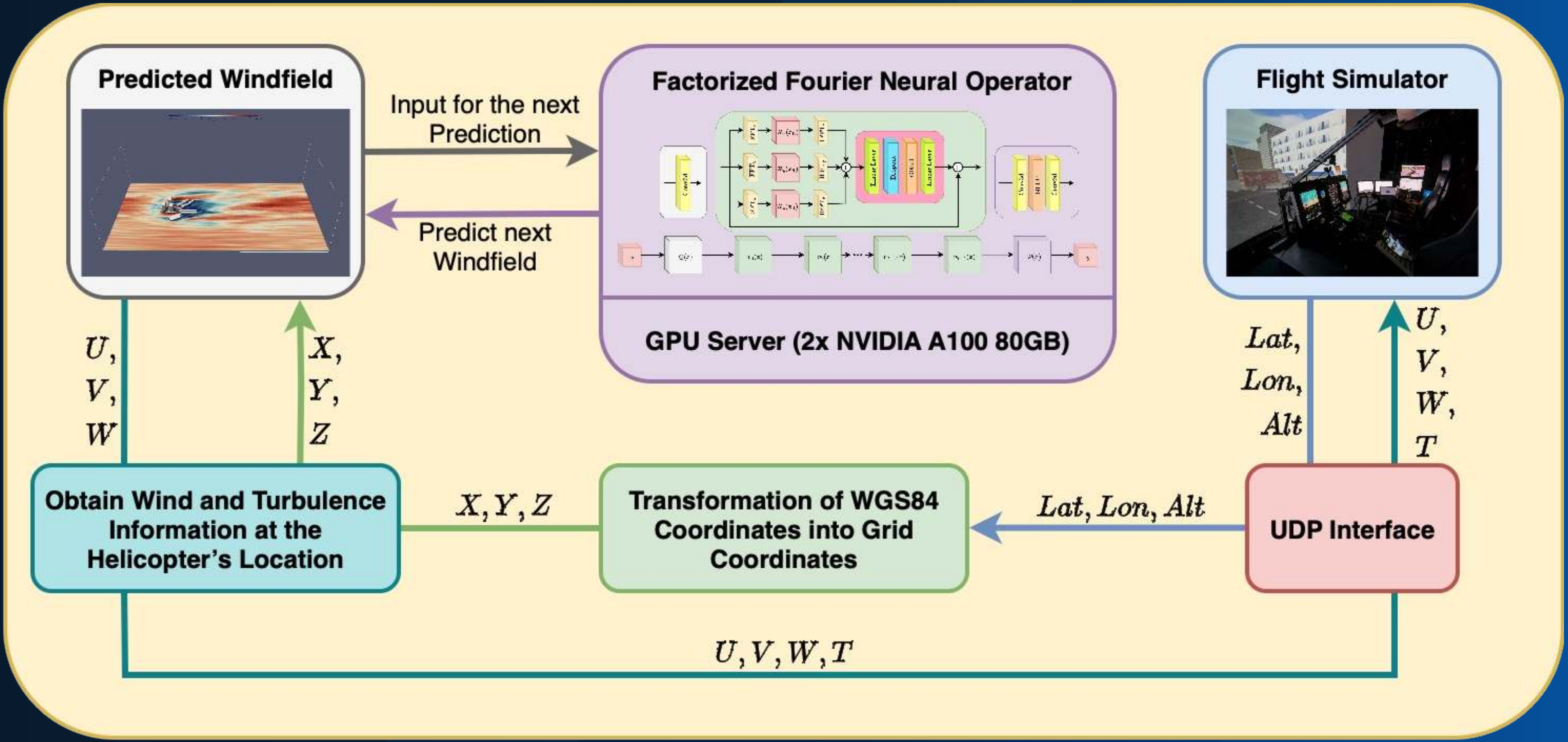
My research explores AI-based methods for faster turbulence prediction.

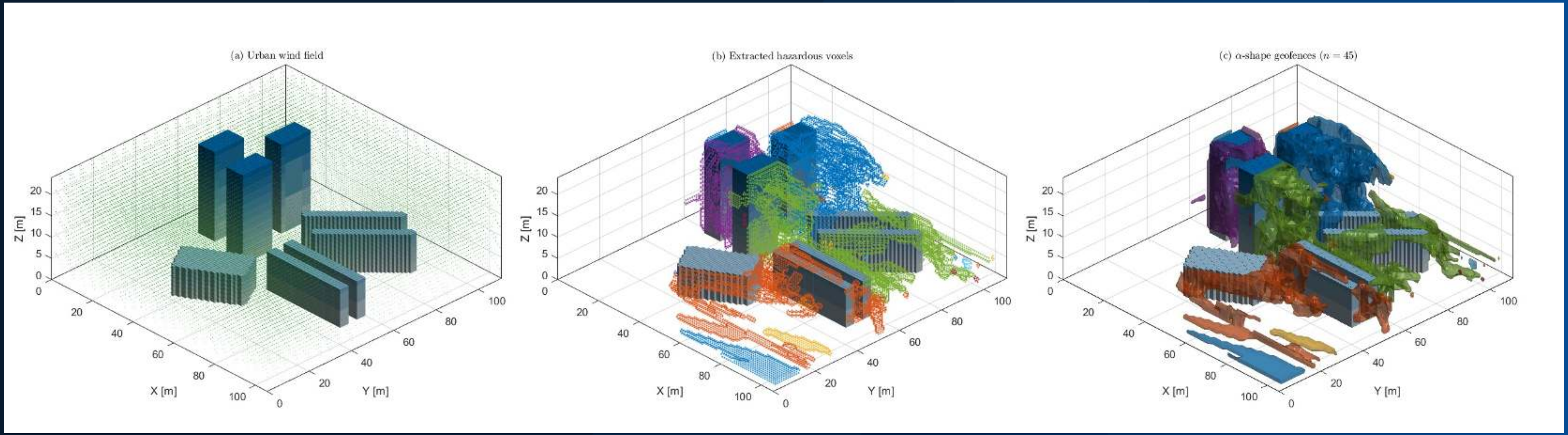

Z [meters]
X [meters]
Y [meters]



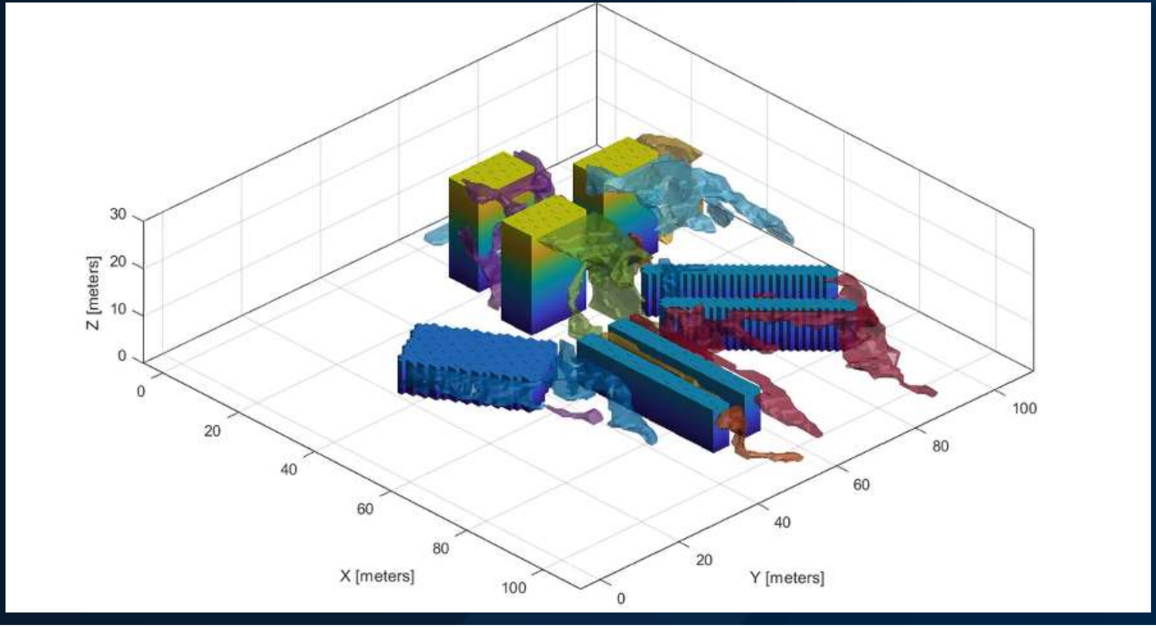



Static 3D Wind Field
Discretized 3D Wind Field
Predicted 3D Wind Field
Input
Neural Network
Output
Scattered Wind Measurements
Extra- Interpolation on Grid

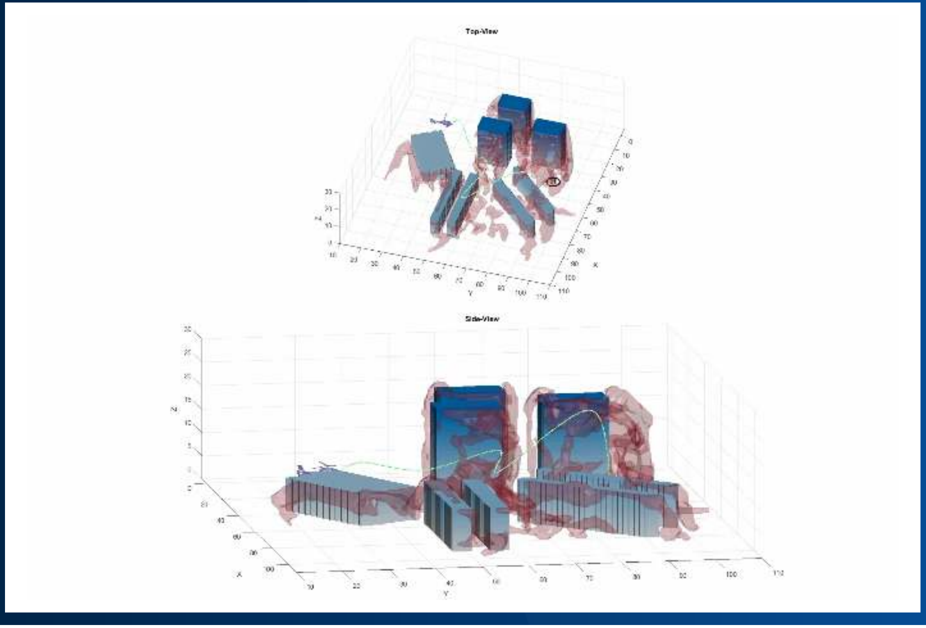

Predicted Windfield
Input for the next Prediction
Predict next Windfield
Factorized Fourier Neural Operator
GPU Server (2x NVIDIA A100 80GB)
Flight Simulator
U, V, W
X, Y, Z
Lat, Lon, Alt
U, V, W, T
Obtain Wind and Turbulence Information at the Helicopter's Location
X, Y, Z
Transformation of WGS84 Coordinates into Grid Coordinates
Lat, Lon, Alt
UDP Interface
U, V, W, T



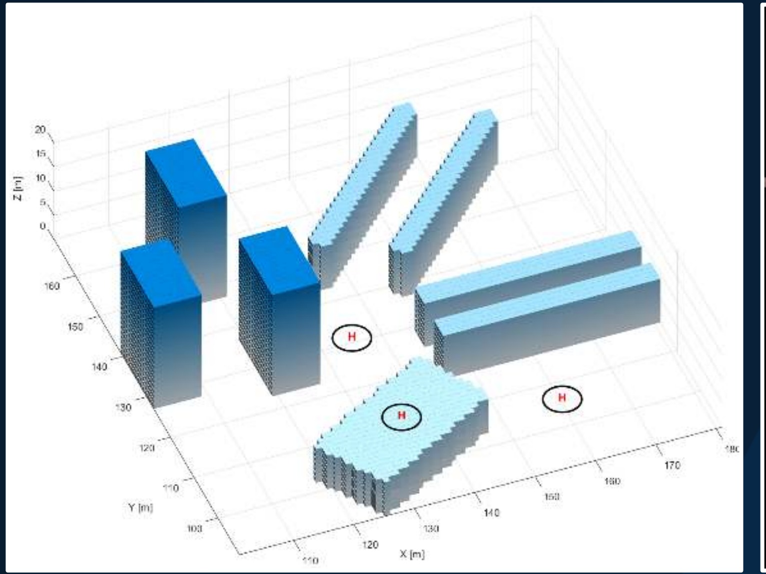
(a) Urban wind field
(b) Extracted hazardous voxels
(c) α-shape geofences (n = 45)
Z [m]
X [m]
Y [m]
0
5
10
15
20
40
60
80
100

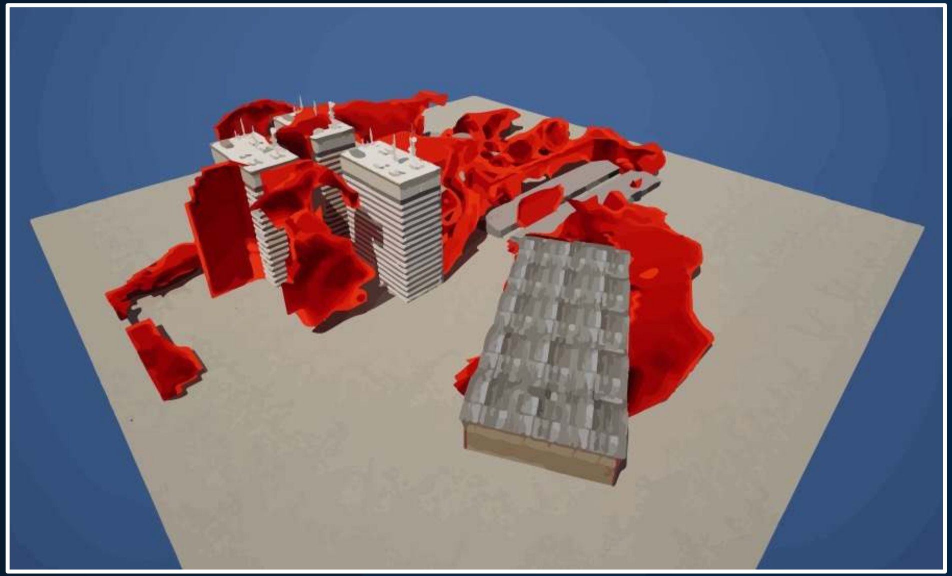
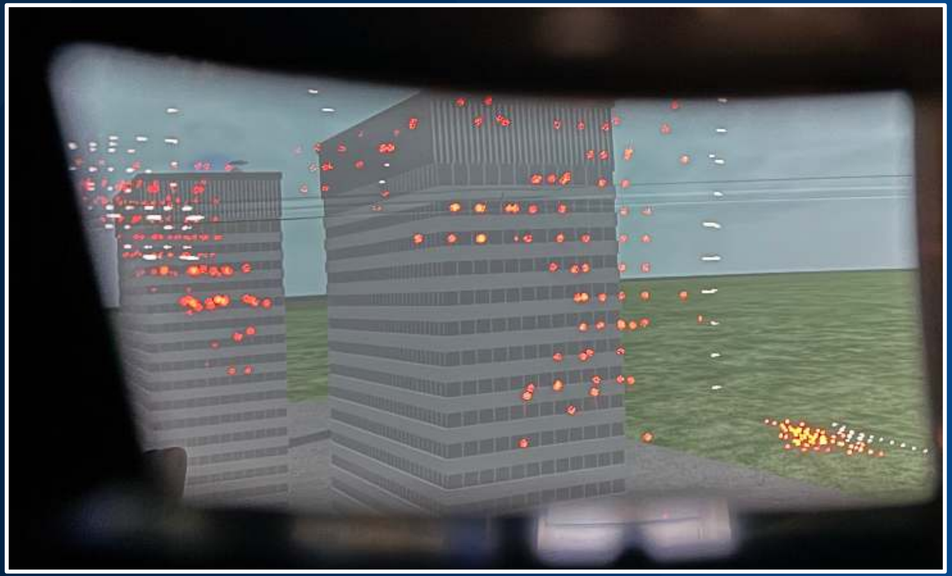



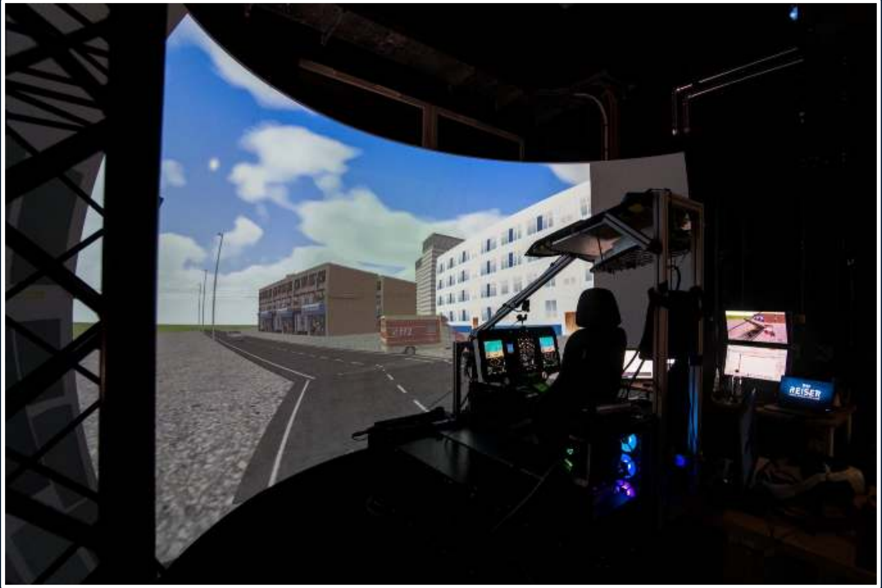
Z [meters]
X [meters]
Y [meters]
0
10
20
30
40
60
80
100

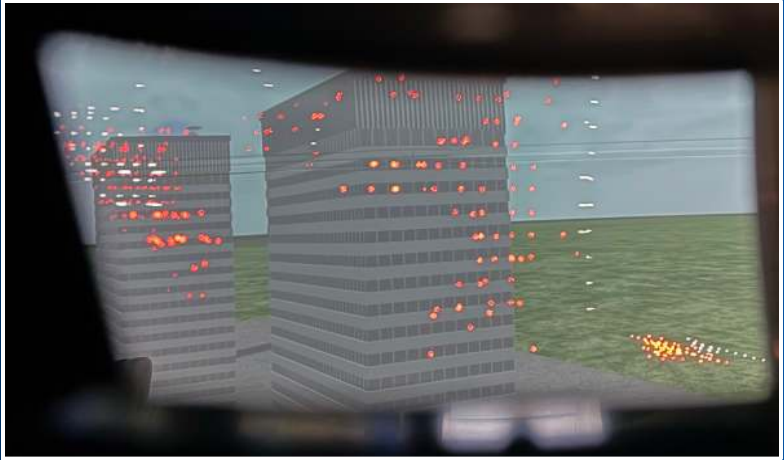
Top-View
Side-View


Z [m]
20
15
10
5
0
160
150
140
130
120
110
100
Y [m]
110
120
130
140
150
160
170
180
X [m]
H
H
H

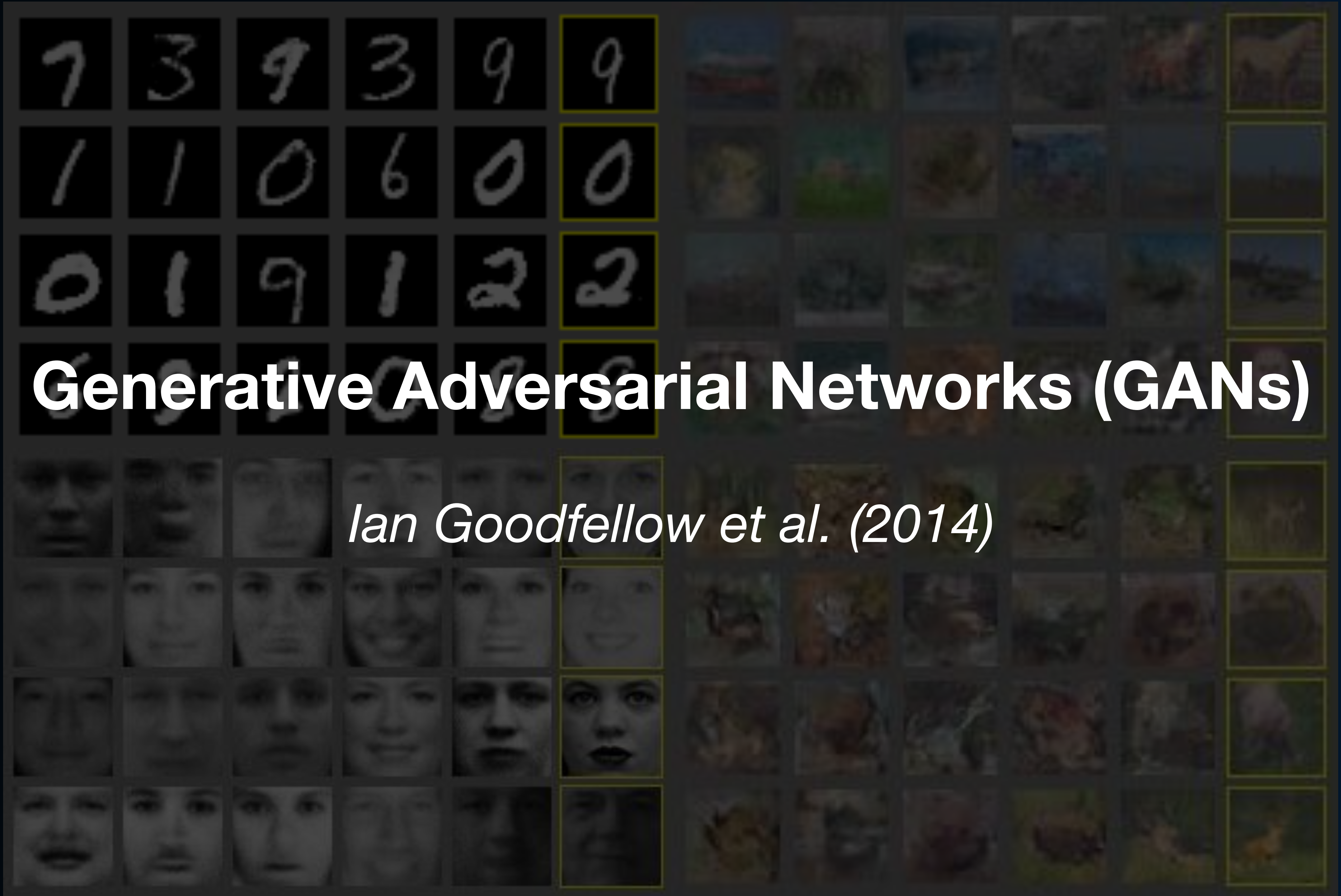

REISER


REISER



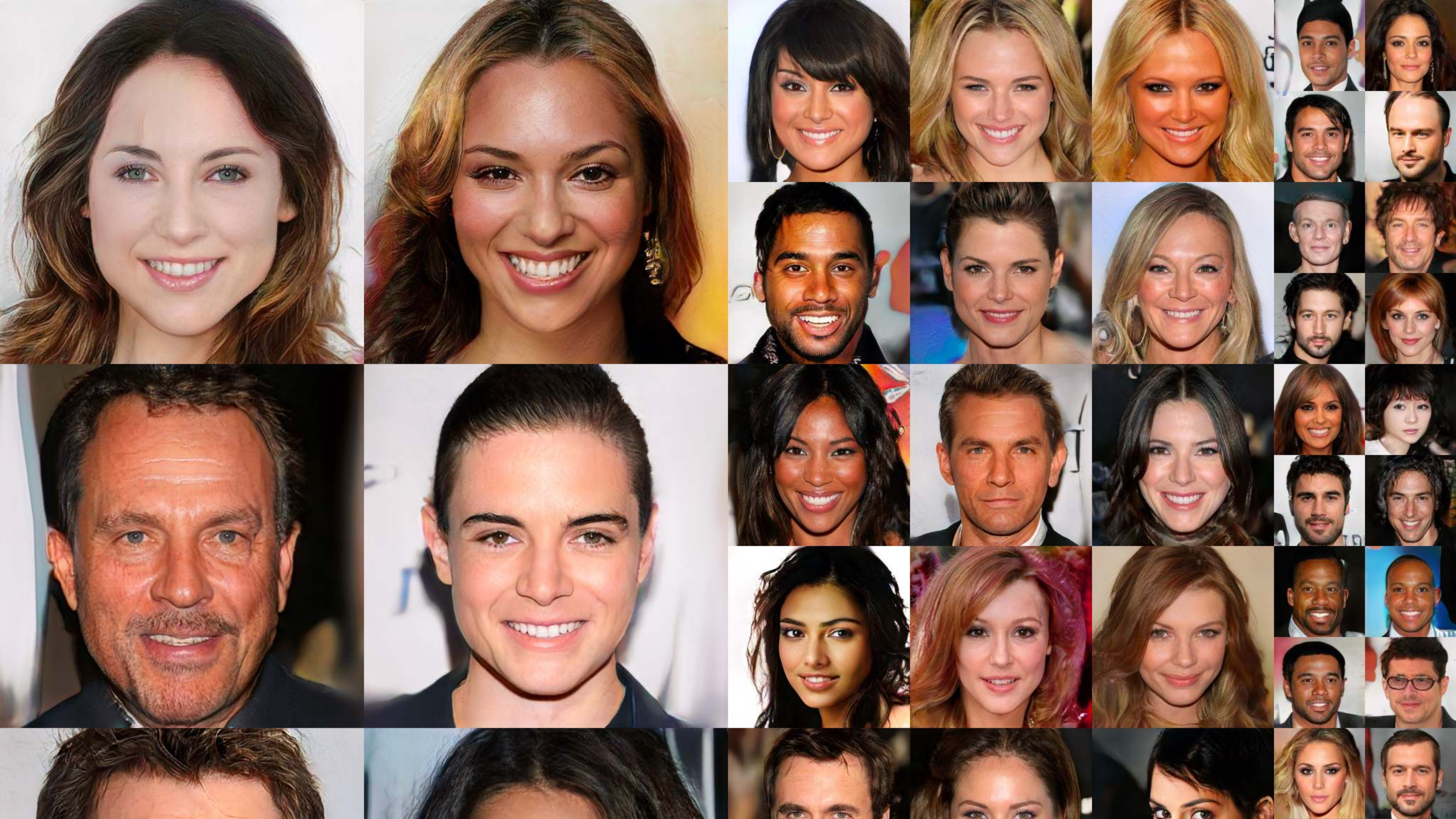
# Introduction to Artificial Intelligence

A first look at the concepts behind AI systems.



# Generative Adversarial Networks (GANs)

*Ian Goodfellow et al. (2014)*



# Progressive Growing GANs (ProGAN)

*Kerras et al. (2017)*

*a hedgehog using a calculator*

*a corgi wearing a red bowtie and purple party hat*

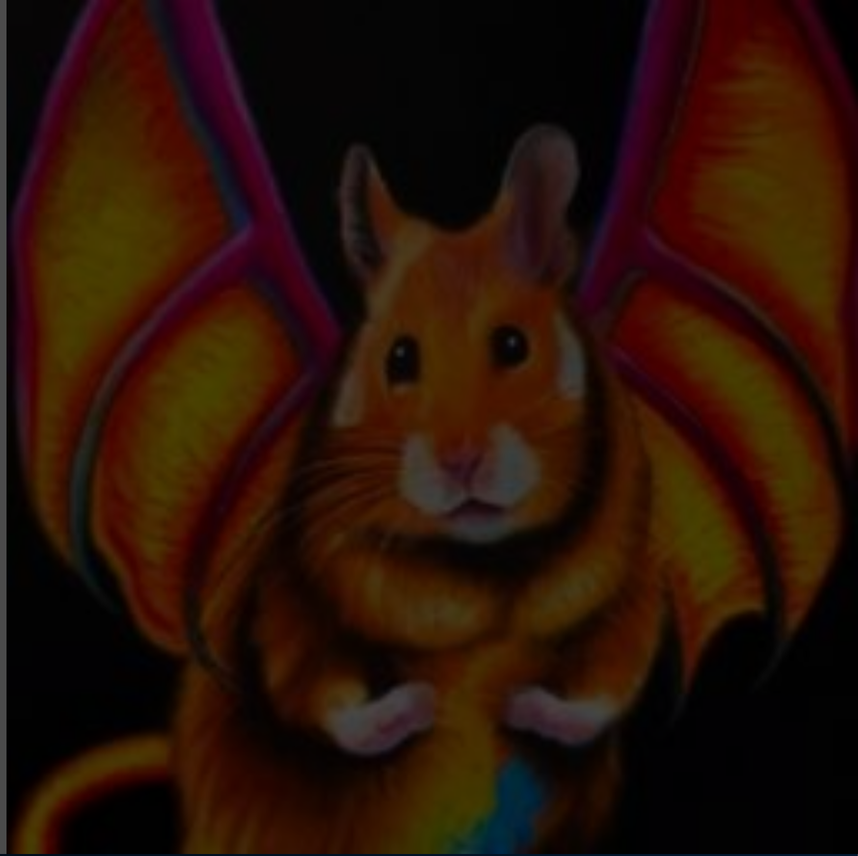
*a high-quality oil painting of a psychedelic hamster dragon*

# GLIDE

*Nichol et al. (2021)*

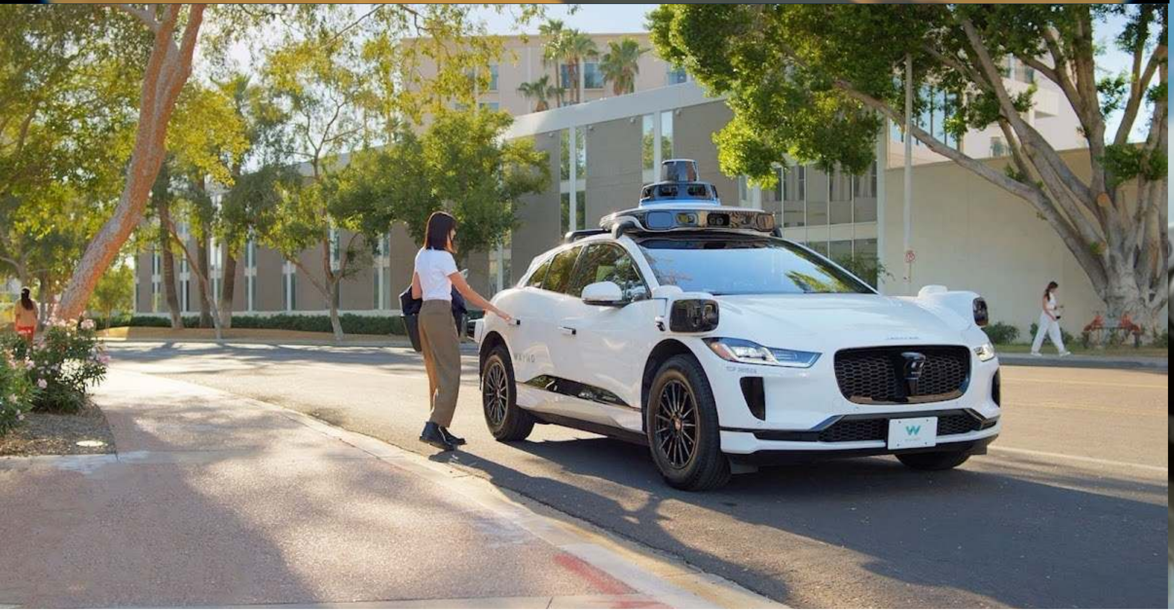
*a hedgehog using a calculator*

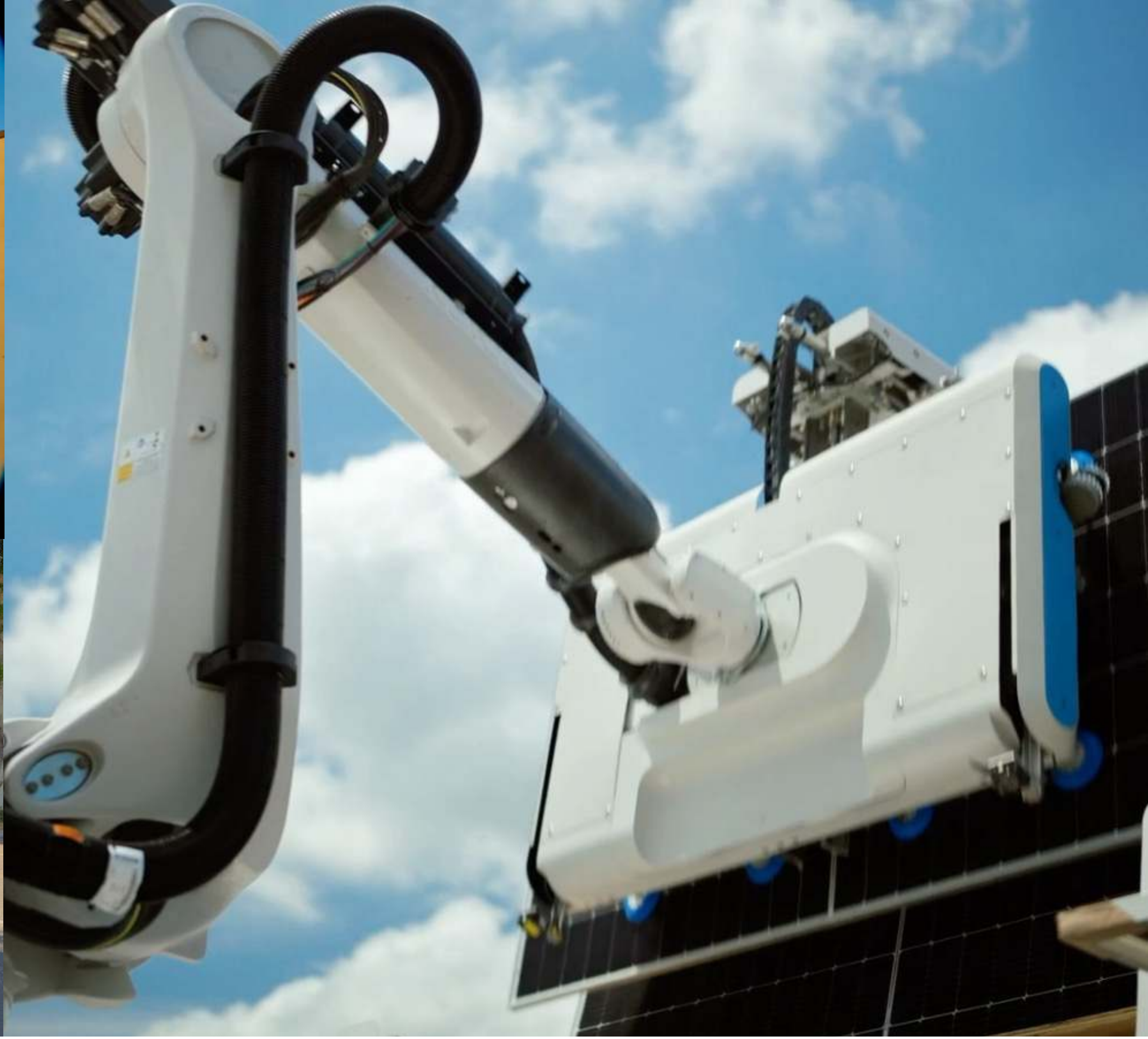
*a corgi wearing a red bowtie and purple party hat*



*a high-quality oil painting of a psychedelic hamster dragon*







# NEO
Home Robot

Product FAQ

**Standard** — **$499/mo** Subscription
- Monthly Subscription
- Starter Productivity Package
- Standard Delivery

**Early Access** — **$20,000** Ownership
- Ownership with 3-Year Warranty
- Premium Support
- Priority Delivery

**$200 Deposit Due Today**
Fully Refundable
US Deliveries start 2026

Order Now

## Artificial Intelligence (AI)

Systems that simulate aspects of human intelligence, such as reasoning, learning, problem-solving, and decision-making.

## Machine Learning (ML)

A subset of AI in which systems learn from data and improve their performance through training without being explicitly programmed for every task.

### Neural Network (NN)

Computational models inspired by the human brain that recognize patterns and relationships in data through interconnected layers of artificial neurons.

#### Deep Learning

A subset of machine learning that uses deep neural networks with multiple layers to process complex data and solve advanced tasks.

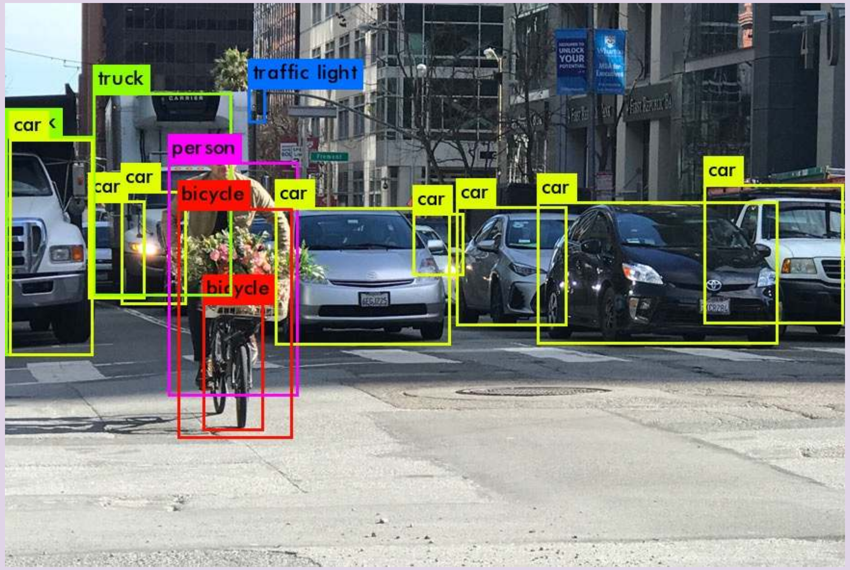
Gemini

Claude

## Discriminative AI

Models that learn to assign data to predefined categories or classes.

They make decisions by identifying differences between classes and determining which category an input most likely belongs to.




## Generative AI

Models that can create new content independently, such as text, images, audio, or code.

They learn patterns from existing data and use them to generate new, similar, but original examples.

# Time Travel: Before the AI Boom

Before machine learning, many intelligent-looking programs relied on manually defined rules and features.


SINGAPORE
SMASH
2025
DHS
EXPERIENCE A

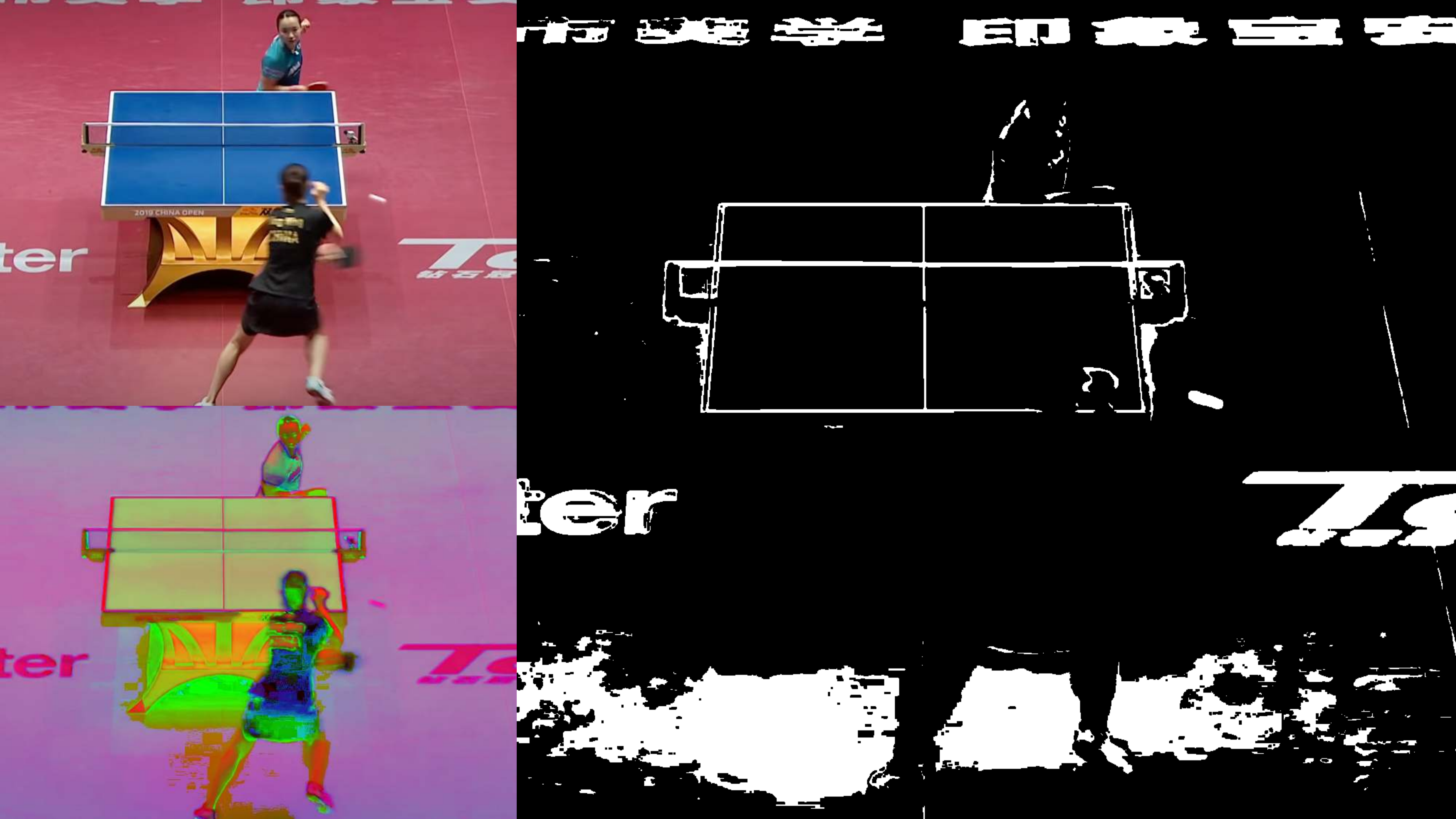

城市
2019 CHINA OPEN
双鱼

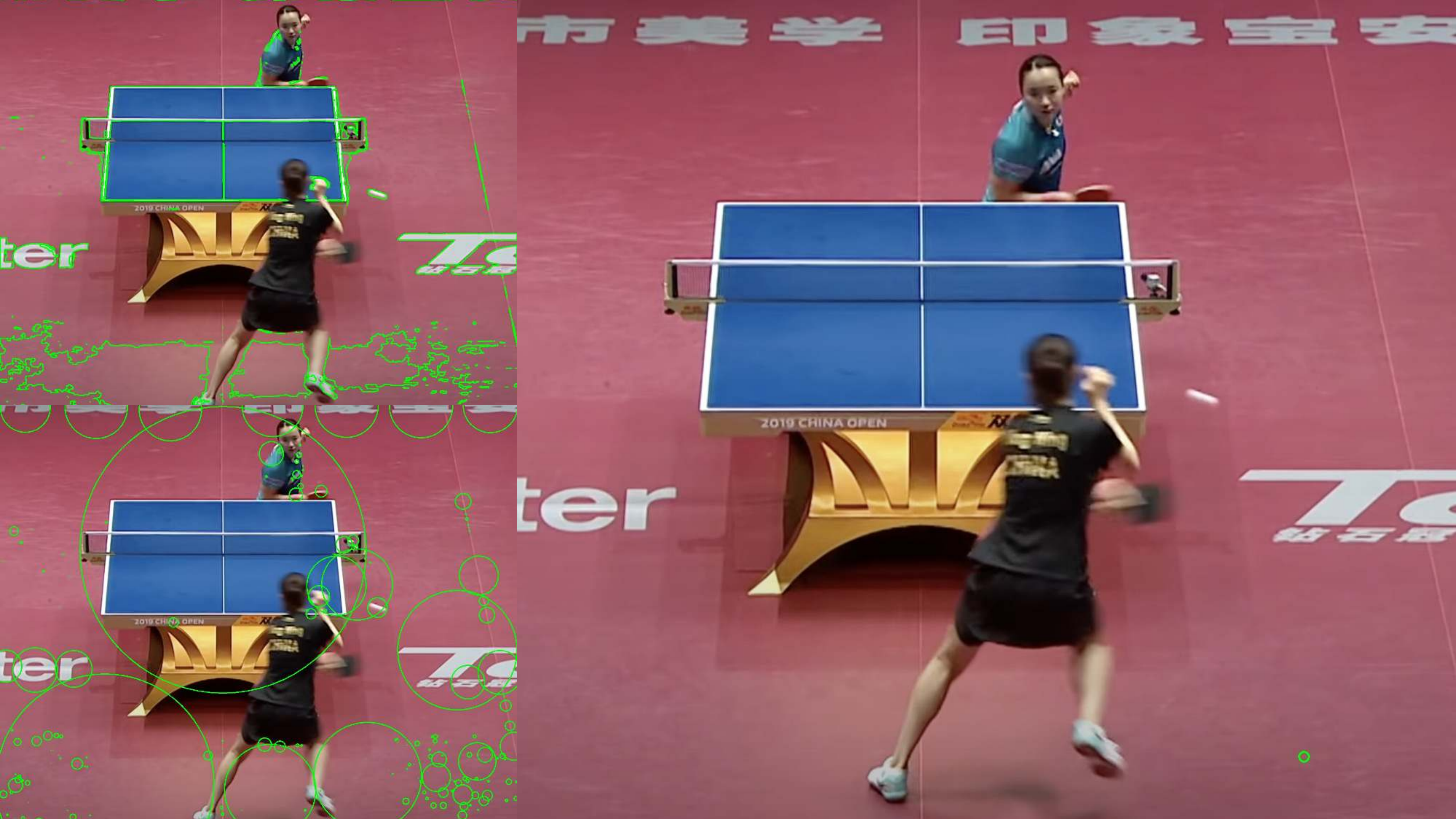
# When do rules start to break?

A rule-based vision system only detects what we explicitly tell it to look for.


DHS
SINGAPORE
SMASH
2025


DHS
DHS
SINGAPORE
SMASH
2025

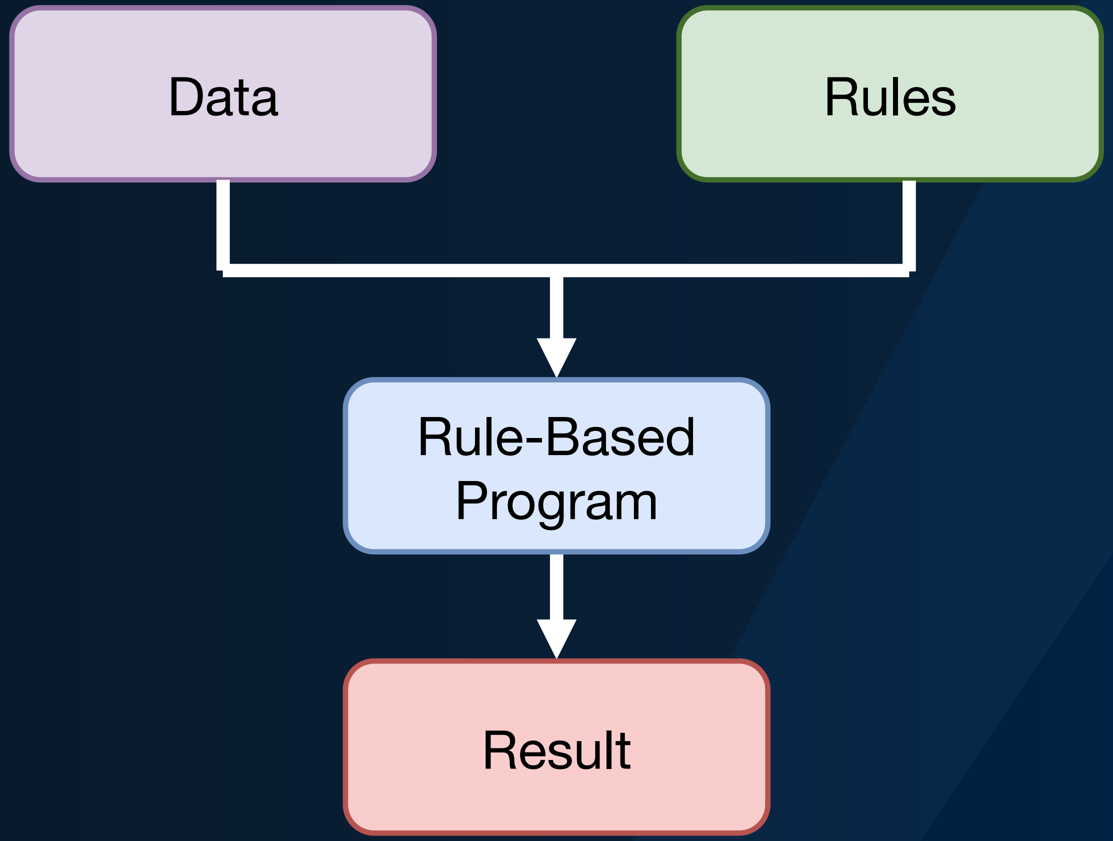

2019 CHINA OPEN


2019 CHINA OPEN
ter

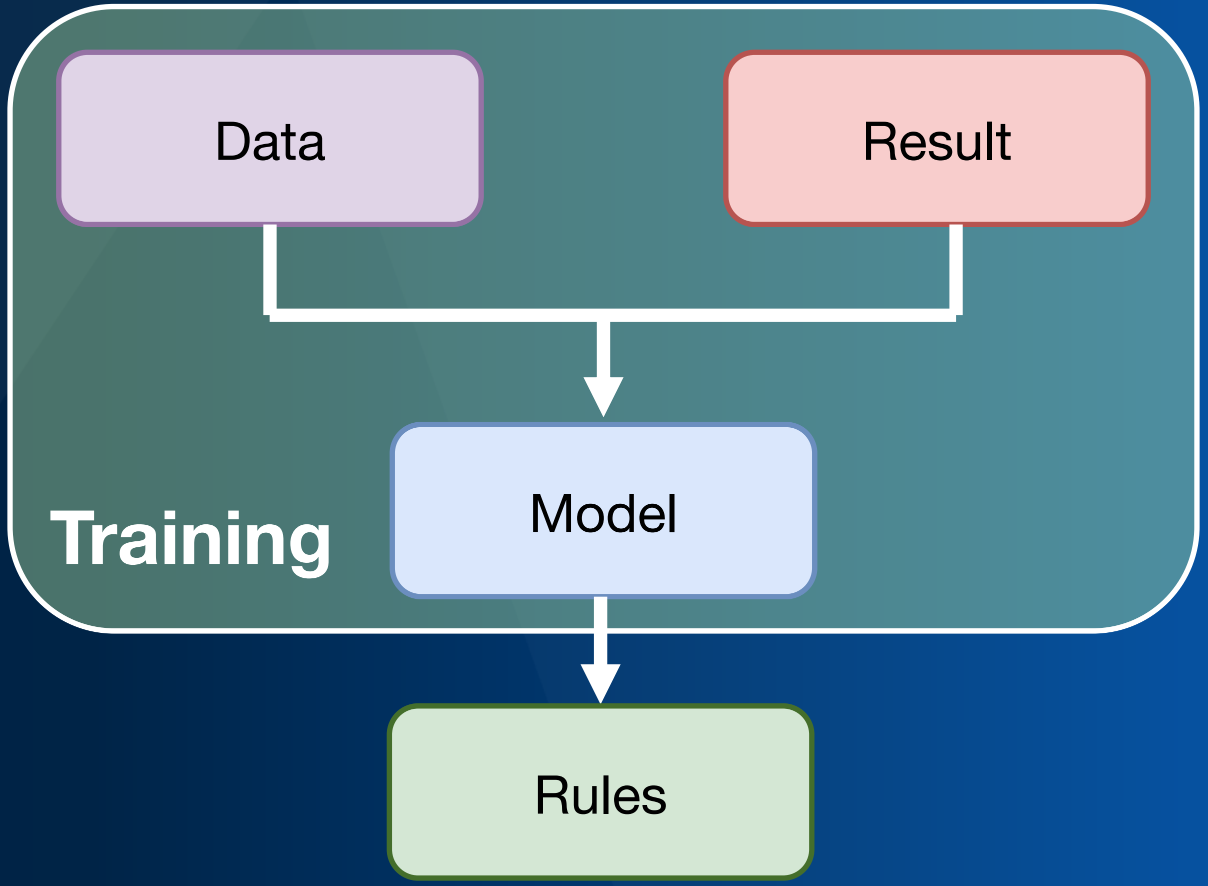
# From handwritten rules to learning from data

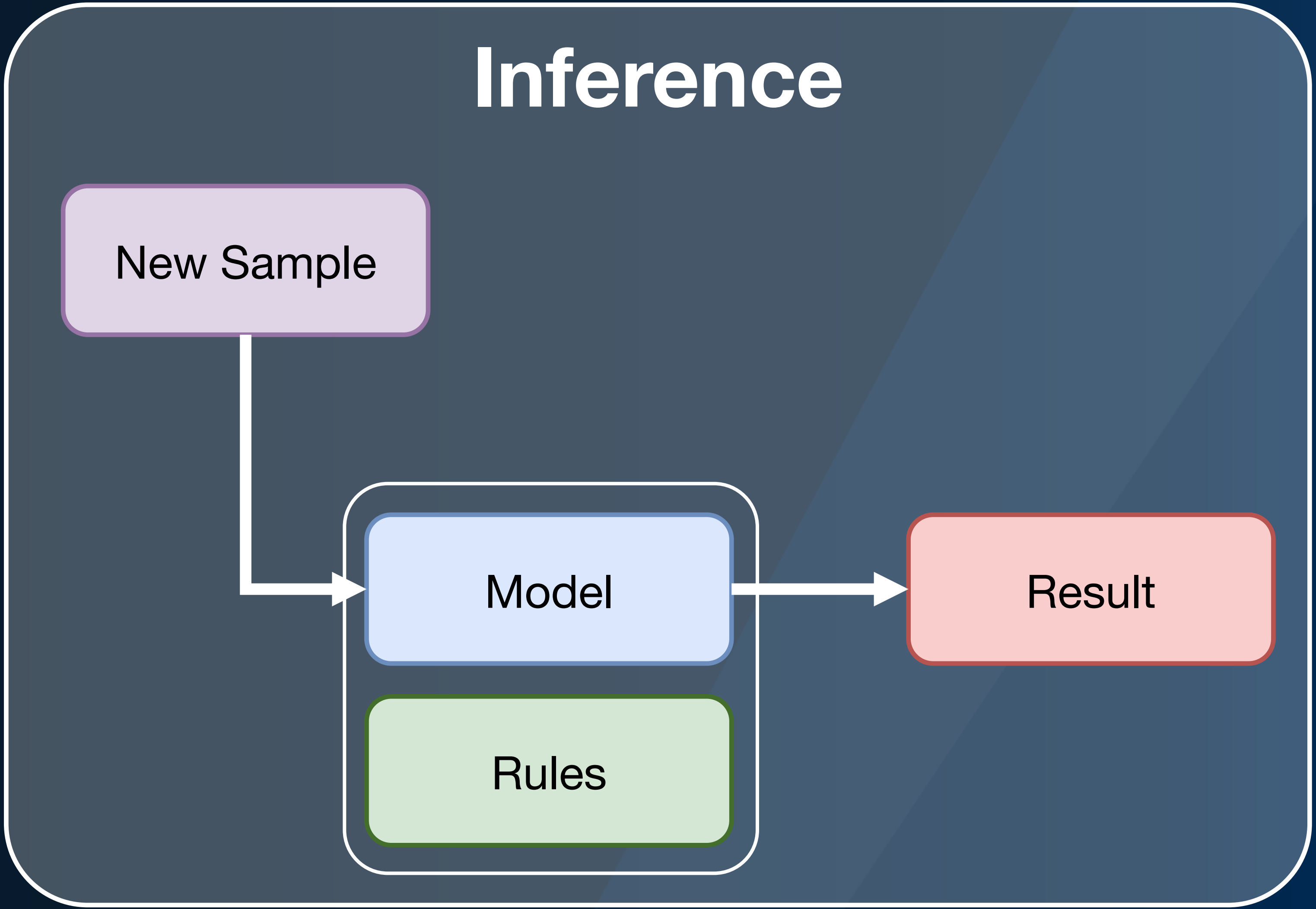
Instead of manually defining every rule, learning-based systems use examples to discover the relevant patterns directly from data.

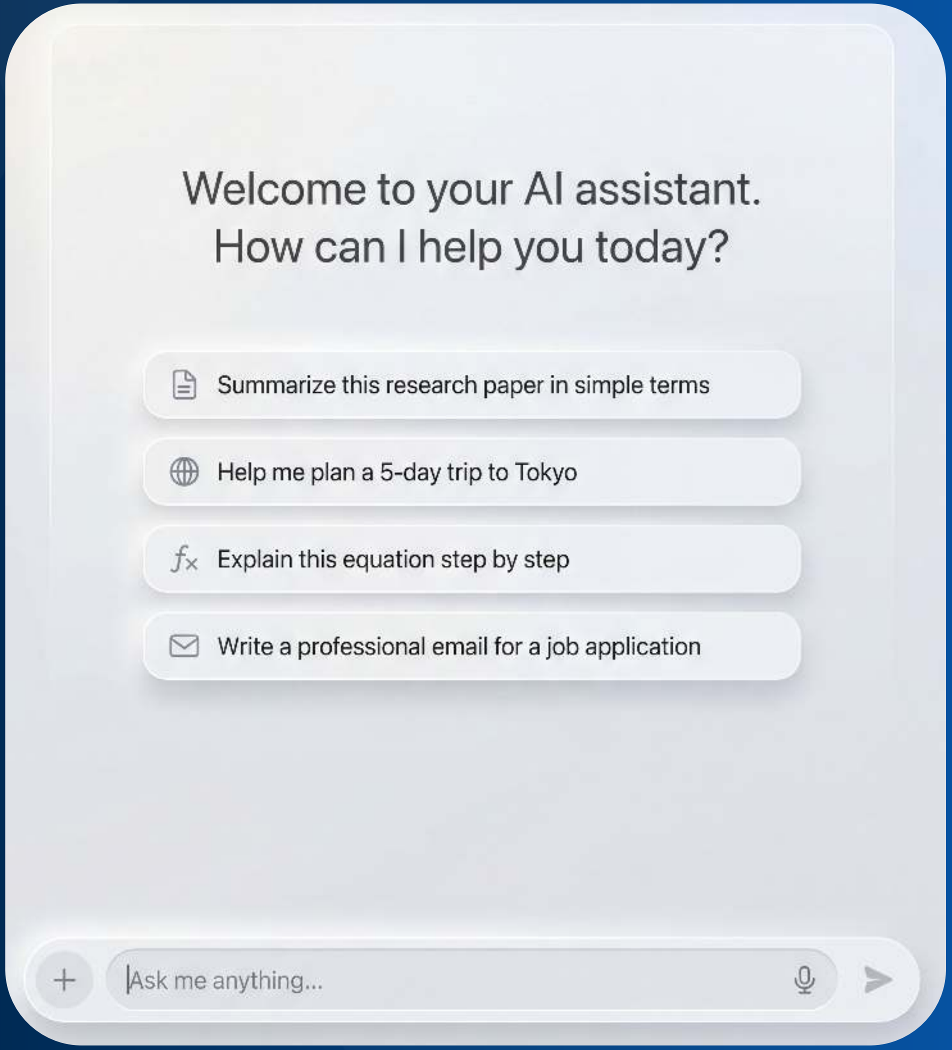
# Rule-Based System




# Learning-Based System




Inference
New Sample
Model
Rules
Result


Welcome to your AI assistant.
How can I help you today?
Summarize this research paper in simple terms
Help me plan a 5-day trip to Tokyo
Explain this equation step by step
Write a professional email for a job application
Ask me anything...

## Dataset

A collection of examples that form the model's knowledge base.

## Model

A neural network with a specific architecture adapted on the data or use case.

## Training

The process where a neural network learns patterns from data.

## Inference

After training, the model uses what it has learned to make predictions or generate content from unseen inputs.

## Dataset

A collection of examples that form the model's knowledge base.

**Model**

Goal: Human Face

## Dataset

A collection of examples that form the model's knowledge base.

**Model**

Goal: Human Face

# Dataset

A collection of examples that form the model's knowledge base.

**Model**

Goal: Human Face

A Green Monkey

**Model**

## Dataset

Goal: Text to Image

A collection of examples that form the model's knowledge base.

A Green Monkey

# Dataset

Goal: Text to Image

A collection of examples that form the model's knowledge base.

**Model**

A snowy suburban street with a red car parked on the side and a green truck driving down the road. The street is lined with houses and trees, and there is a sign on the left side of the image

A Green Monkey

**Model**

## Dataset

Goal: Text to Image

A collection of examples that form the model's knowledge base.

A red helmet with a black chin strap and a black and red logo on the side is displayed against a white background.

A Green Monkey

# Dataset

Goal: Text to Image

A collection of examples that form the model's knowledge base.

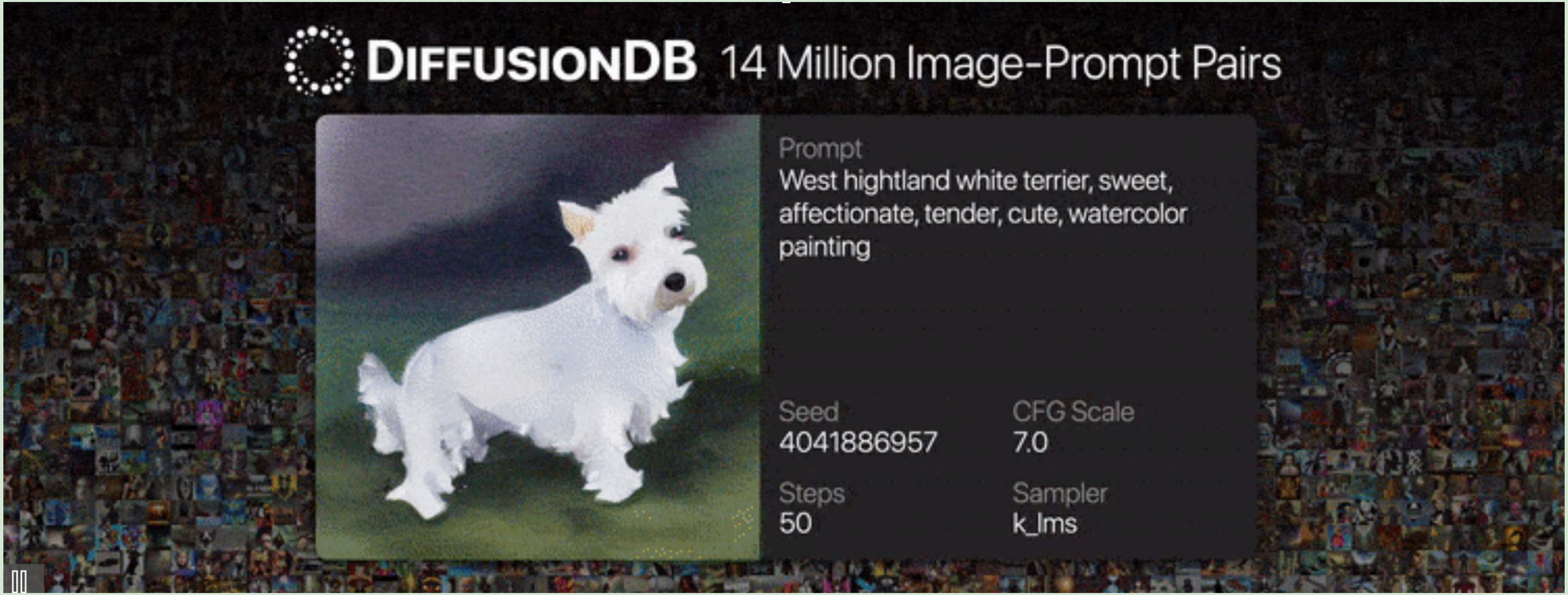
**Model**

A man in a suit is standing at a podium with a microphone in front of him. He is wearing a light blue tie with a subtle pattern. The background shows a blurred view of trees and a clear sky.

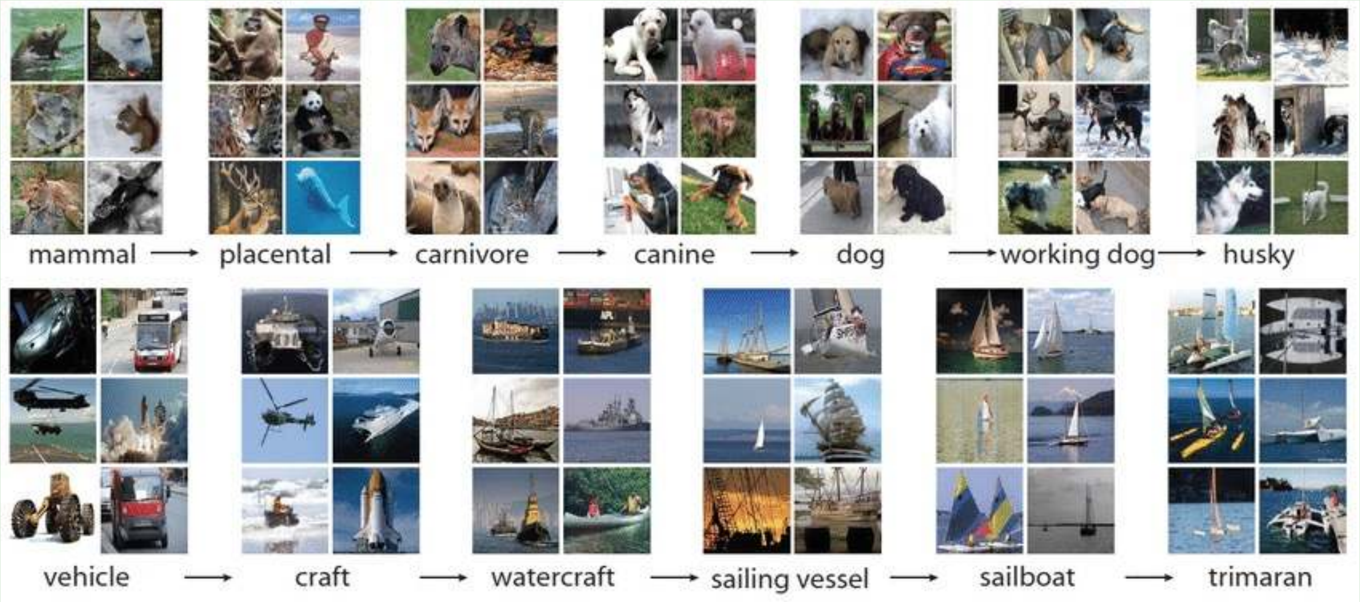
# Dataset

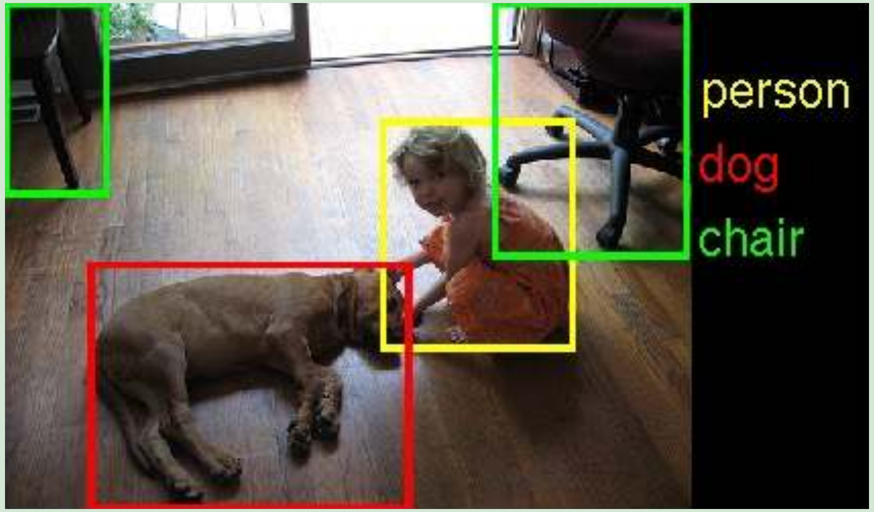
| jpg<br>dict | json<br>dict |
|---|---|
|  | {"prompt": "A promotional image for a crampon storage bag with a group of hikers in the background. The bag is prominently displayed in the foreground, with a large, detailed image of the bag's mesh design an… |
|  | {"prompt": "A small, shiny copper-colored object is placed on a white surface."} |
|  | {"prompt": "A stack of pancakes with a glossy, caramel-like syrup on top is presented. The pancakes are golden brown and appear to be freshly cooked. The syrup is dripping down the sides of the pancakes,… |
|  | {"prompt": "The first image shows a close-up of a motorcycle's rear fender with the Harley-Davidson logo prominently displayed. The logo is centered on the fender, which is a dark color with a glossy finish.… |
|  | {"prompt": "Aerial view of a tropical island with a long, curved row of traditional thatched-roof huts on stilts over the water. The huts are connected by a wooden walkway. The island has a sandy beach with a… |
|  | {"prompt": "Aerial view of a lush green mountainous landscape with a winding path leading through the hills. The terrain is covered with dense vegetation, and the sky is partly cloudy."} |

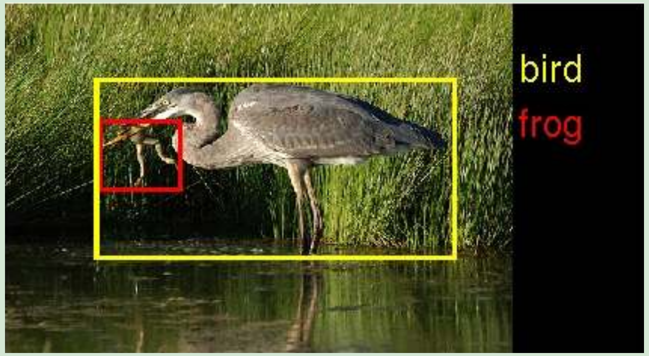
# Dataset

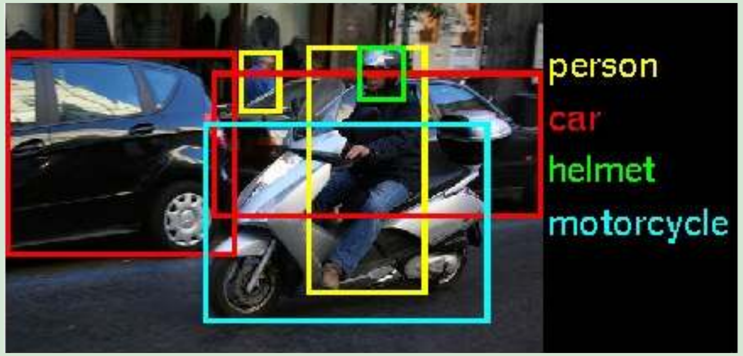

DIFFUSIONDB 14 Million Image-Prompt Pairs
Prompt
West hightland white terrier, sweet, affectionate, tender, cute, watercolor painting
Seed
4041886957
CFG Scale
7.0
Steps
50
Sampler
k_lms

# Dataset


mammal → placental → carnivore → canine → dog → working dog → husky
vehicle → craft → watercraft → sailing vessel → sailboat → trimaran


person
dog
chair


bird
frog


person
car
helmet
motorcycle

Can you help me?

Large Language Model

I'll do my best! What exactly can I help you with?

## Dataset

Goal: Text to Text

A collection of examples that form the model's knowledge base.

COMMON CRAWL

### Common Crawl February 2025 Crawl Archive (CC-MAIN-2025-08)

The February 2025 crawl archive contains 2.67 billion pages, see the announcement for details.

#### Data Size and File Listings

| Data Type | File List | #Files | Total Size Compressed (TiB) |
|---|---|---|---|
| Segments | segment.paths.gz | 100 | |
| WARC | warc.paths.gz | 90000 | 82.17 |
| WAT | wat.paths.gz | 90000 | 18.98 |
| WET | wet.paths.gz | 90000 | 7.40 |
| Robots.txt files | robotstxt.paths.gz | 90000 | 0.15 |
| Non-200 responses | non200responses.paths.gz | 90000 | 3.09 |
| URL index files | cc-index.paths.gz | 302 | 0.20 |
| Columnar URL index files | cc-index-table.paths.gz | 900 | 0.23 |

# What does the model never see?

A model can only learn from the examples, patterns, and perspectives contained in its dataset.


NOMOS
GLASHÜTTE

Breitling
TOP TIME
TACHYMETER

STERNGLAS
ZEITMESSER
AUTOMATIK


CONTRACT

Goal: Generate Human Faces

Model

What kinds of faces can the model generate from this dataset?

## Dataset

**Goal: Generate Human Faces**

**Model**

**What kinds of faces can the model generate from this dataset?**

**Dataset**

**Now the dataset is no longer based on text or images. It is a physical field on space, time, geometry and boundary conditions.**

# How does a dataset for turbulence prediction look like?

Think first. What information would you need to predict turbulences?

## Task

Brainstorm for 5 minutes

What is included in such a dataset?

# Define the specific use case and the needed spatial and temperol resolution first

Coordinates: $\mathbf{x} = (x, y, z)$ $\mathbf{x} \in \mathbb{R}^3$

Discrete time: $t_i = t_0 + i\Delta t$

Velocity vector: $\mathbf{u}(\mathbf{x}, t) \in \mathbb{R}^3$

Velocity components: $\mathbf{u}(\mathbf{x}, t) = \big(u(\mathbf{x}, t), v(\mathbf{x}, t), w(\mathbf{x}, t)\big)$

Wind-speed magnitude: $U(\mathbf{x}, t) = \sqrt{u(\mathbf{x}, t)^2 + v(\mathbf{x}, t)^2 + w(\mathbf{x}, t)^2}$

# There is no universal turbulence dataset

A turbulence dataset is only useful if its scale, resolution, variables, and time steps match the physical problem we want to solve.

# Meteorological Mast

Wind shears at fixed horizontal locations.

Vertical line in physical space: $\Omega_{1D} = \{(x_0, y_0, z) \in \mathbb{R}^3 \mid z \in [z_{\min}, z_{\max}]\}$

Vertical grid points: $z_j = z_{\min} + j\Delta z, \quad j = 0, \dots, N_z - 1$

Sample location: $\mathbf{x}_j = \left(x_0, y_0, z_j\right) \in \mathbb{R}^3$

Velocity sample: $\mathbf{u}(\mathbf{x}_j, t_i) = \left(u(\mathbf{x}_j, t_i), v(\mathbf{x}_j, t_i), w(\mathbf{x}_j, t_i)\right) \in \mathbb{R}^3$

# Meteorological Mast

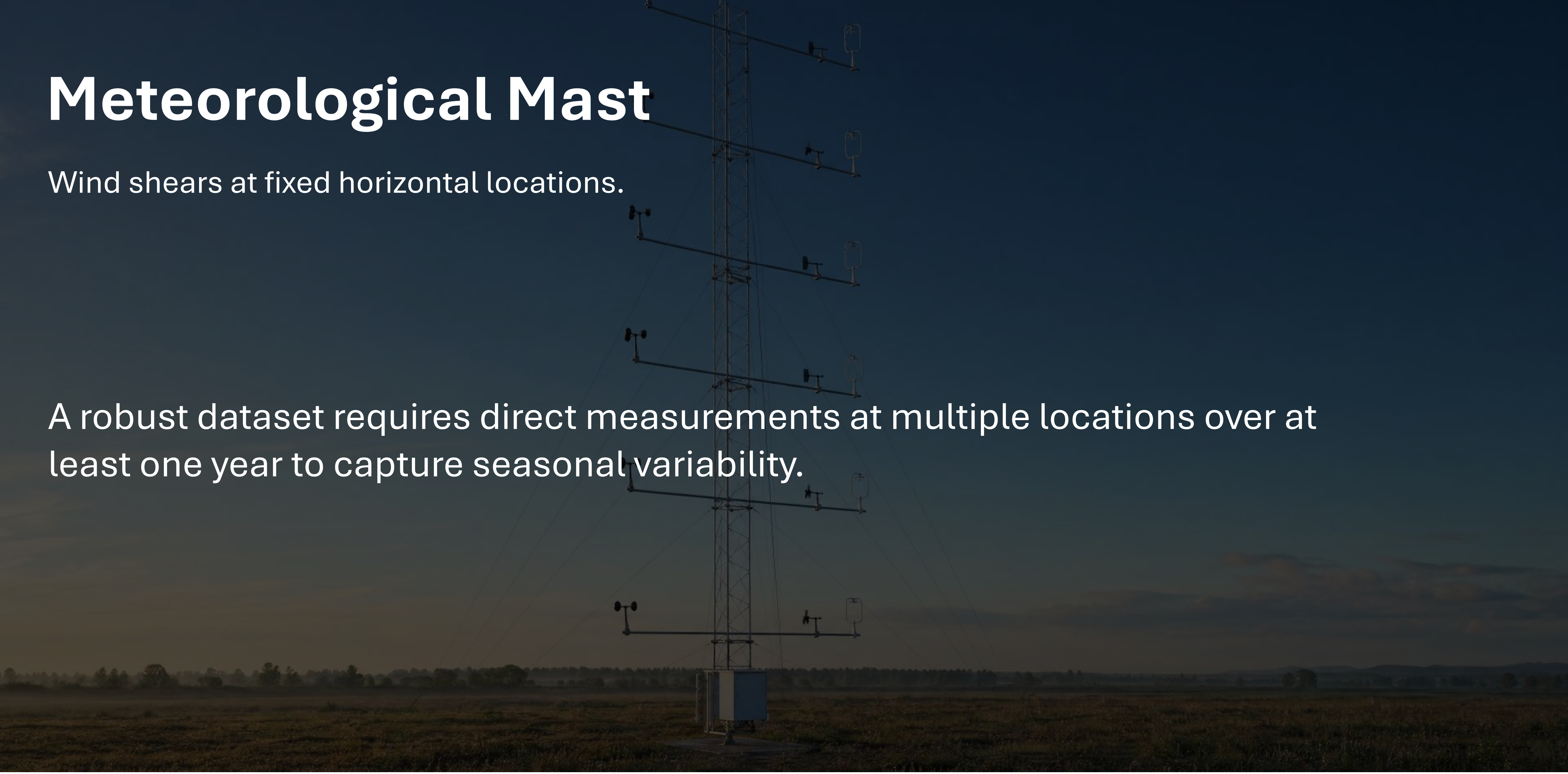
Wind shears at fixed horizontal locations.

$$\Omega_{1D}^{\text{mast}} = \{(x_0, y_0, z) \in \mathbb{R}^3 \mid z \in [10{,}200]\ \text{m}\}$$

$$z_j = 10\ \text{m} + j\Delta z, \qquad \Delta z \approx 10\text{–}50\ \text{m} \qquad j = 0, \dots, N_z - 1$$

$t_i = t_0 + i\Delta t, \qquad \Delta t \approx 10\ \text{min}$ for averaged wind statistics

$$\mathbf{X}_{1D}^{\text{mast}} \in \mathbb{R}^{N_t \times N_z \times C}, \qquad C = (U, \theta, T, p, TI)$$



# Meteorological Mast

Wind shears at fixed horizontal locations.

A robust dataset requires direct measurements at multiple locations over at least one year to capture seasonal variability.

# Pedestiran Level Wind Prediction

Wind comfort or wind safety at human height.

2D spatial domain: $\Omega_{2D} = \{(x, y, z_p) \in \mathbb{R}^3 \mid x \in [x_{\min}, x_{\max}],\ y \in [y_{\min}, y_{\max}]\}$

2D grid points: $x_j = x_{\min} + j\Delta x, \quad y_k = y_{\min} + k\Delta y$

Sample location: $\mathbf{x}_{j,k} = (x_j, y_k, z_p) \in \mathbb{R}^3$

Velocity sample: $\mathbf{u}(\mathbf{x}_{j,k}, t_i) = \big(u(\mathbf{x}_{j,k}, t_i), v(\mathbf{x}_{j,k}, t_i), w(\mathbf{x}_{j,k}, t_i)\big) \in \mathbb{R}^3$

# Pedestiran Level Wind Prediction

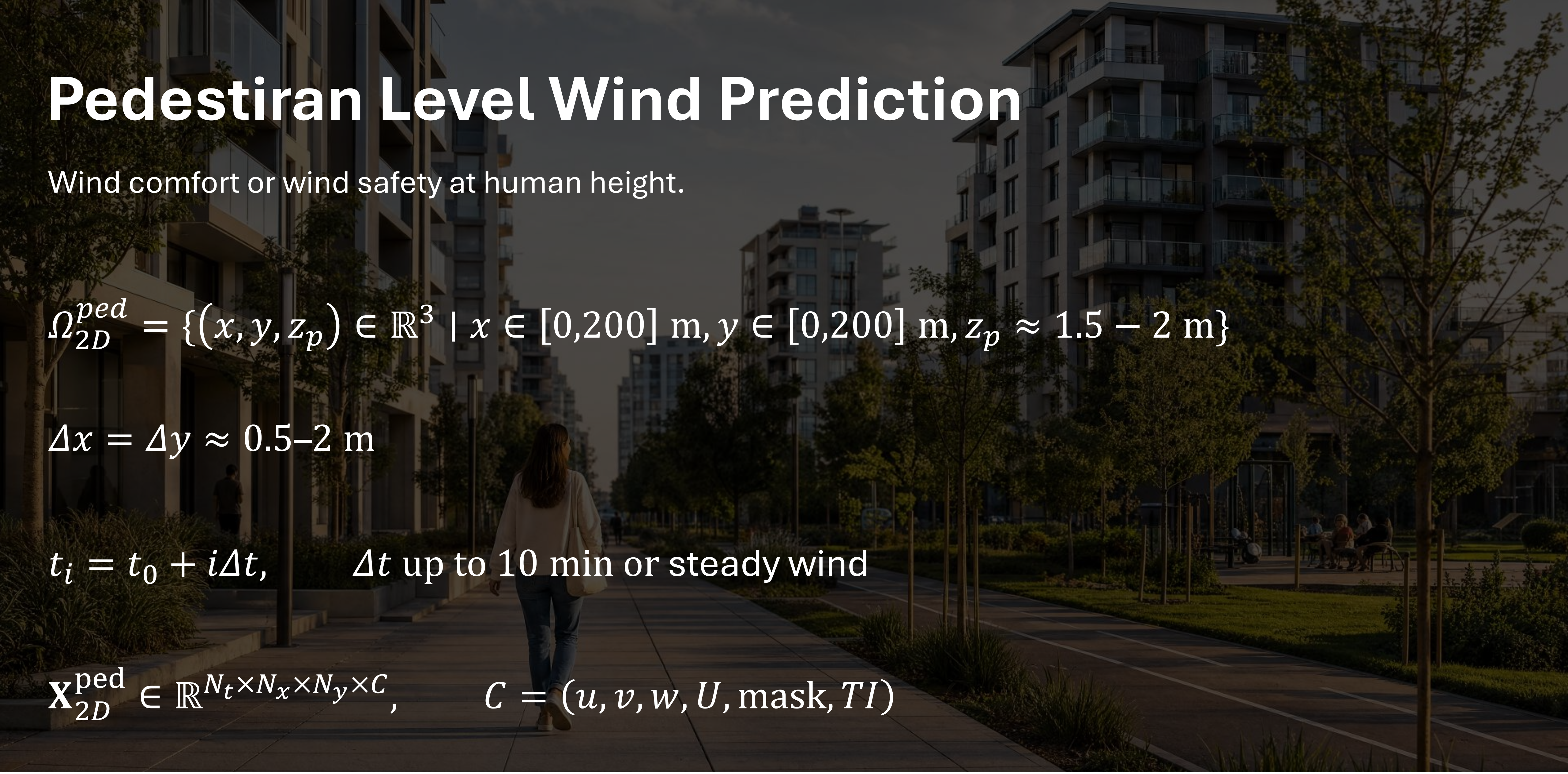
Wind comfort or wind safety at human height.

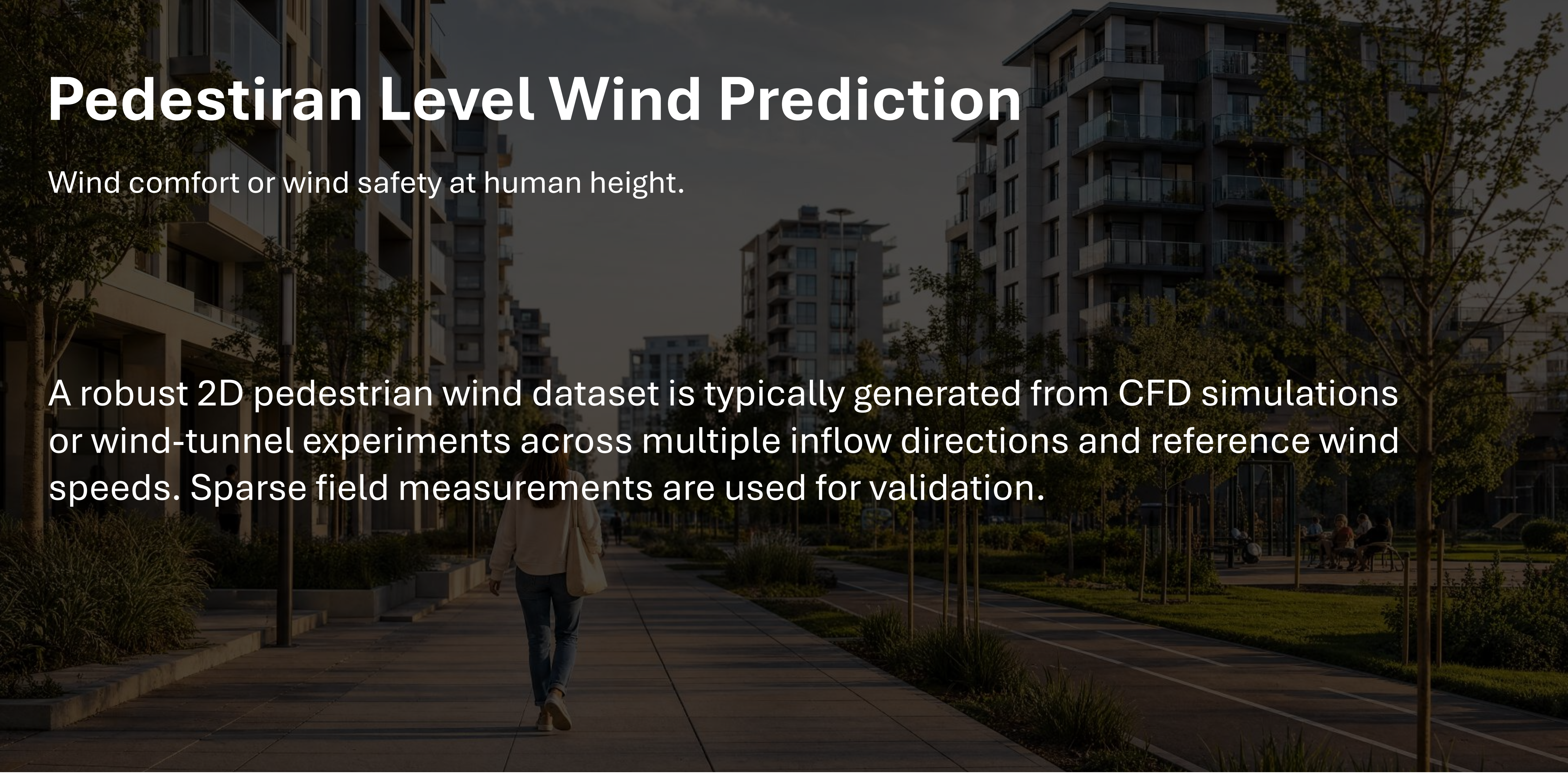
$$\Omega_{2D}^{ped} = \{(x, y, z_p) \in \mathbb{R}^3 \mid x \in [0{,}200]\ \mathrm{m}, y \in [0{,}200]\ \mathrm{m}, z_p \approx 1.5 - 2\ \mathrm{m}\}$$

$$\Delta x = \Delta y \approx 0.5\text{–}2\ \mathrm{m}$$

$$t_i = t_0 + i\Delta t, \qquad \Delta t \text{ up to 10 min or steady wind}$$

$$\mathbf{X}_{2D}^{\mathrm{ped}} \in \mathbb{R}^{N_t \times N_x \times N_y \times C}, \qquad C = (u, v, w, U, \mathrm{mask}, TI)$$





# Pedestiran Level Wind Prediction

Wind comfort or wind safety at human height.

A robust 2D pedestrian wind dataset is typically generated from CFD simulations or wind-tunnel experiments across multiple inflow directions and reference wind speeds. Sparse field measurements are used for validation.

# Urban Area Wind Prediction

Local heat, ventilation, comfort, vegetation effects, urban planning scenario.

3D spatial domain: $\Omega_{3D} = [x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}] \times [z_{\min}, z_{\max}] \subset \mathbb{R}^3$

3D grid points: $x_j = x_{\min} + j\Delta x, \quad y_k = y_{\min} + k\Delta y, \quad z_l = z_{\min} + l\Delta z$

Sample location: $\mathbf{x}_{j,k,l} = (x_j, y_k, z_l) \in \mathbb{R}^3$

Velocity sample: $\mathbf{u}(\mathbf{x}_{j,k,l}, t_i) = \big(u(\mathbf{x}_{j,k,l}, t_i), v(\mathbf{x}_{j,k,l}, t_i), w(\mathbf{x}_{j,k,l}, t_i)\big) \in \mathbb{R}^3$

# Urban Area Wind Prediction

Local heat, ventilation, comfort, vegetation effects, urban planning scenario.

$$\Omega_{3D}^{urban} = \{(x, y, z\ ) \in \mathbb{R}^3 \mid x \in [0{,}750]\ \mathrm{m}, y \in [0{,}750]\ \mathrm{m}, z \in [0{,}200]\ \mathrm{m}\}$$

$$\Delta x \approx \Delta y \approx \Delta z \approx 1\text{–}10\ \mathrm{m}$$

$$t_i = t_0 + i\Delta t, \qquad \Delta t \text{ ranging from seconds to hours}$$

$$\mathbf{X}_{3D}^{\mathrm{urban}} \in \mathbb{R}^{N_t \times N_x \times N_y \times N_z \times C}, \qquad C = (u, v, w, T, p, q, \text{building mask, vegetation})$$

# Urban Area Wind Prediction

Local heat, ventilation, comfort, vegetation effects, urban planning scenario.

Dense urban wind and microclimate datasets are typically generated from CFD simulations or wind-tunnel experiments over realistic urban geometries, while field measurements and long-term meteorological data are used to define boundary conditions, evaluate seasonal relevance, and validate the predicted flow patterns.

# Global Weather Simulation

Extreme weather forecasting and digital twins.

$$\Omega_{3D}^{global} = \{(\lambda, \varphi, \ell) \mid \lambda \in [-180^\circ, 180^\circ],\ \varphi \in [-90^\circ, 90^\circ],\ \ell = 1, \ldots, N_z\}$$

$\Delta x, \Delta y \approx 9 - 31\ \mathrm{km}$ for high-resolution global forecasts

$t_i = t_0 + i\Delta t,\qquad \Delta t$ typical 1, 3, 6, 12 hours

$$\mathbf{X}_{3D}^{\mathrm{global}} \in \mathbb{R}^{N_t \times N_\lambda \times N_\varphi \times N_z \times C}$$

# Global Weather Simulation

Extreme weather forecasting and digital twins.

A robust dataset for global weather prediction is typically derived from reanalysis data, where satellite, weather-station, radiosonde, aircraft, buoy, and radar observations are assimilated into a numerical weather model to create a globally complete, physically consistent atmospheric state over many decades.

# Predicting turbulent flow in urban areas

Urban flow predictions could help drones and air taxis avoid hazardous turbulence and plan safer flight paths.

# City Geography Markup Language (CityGML)

Open, standardized data model and exchange format used to store and share digital 3D models of cities and landscapes.

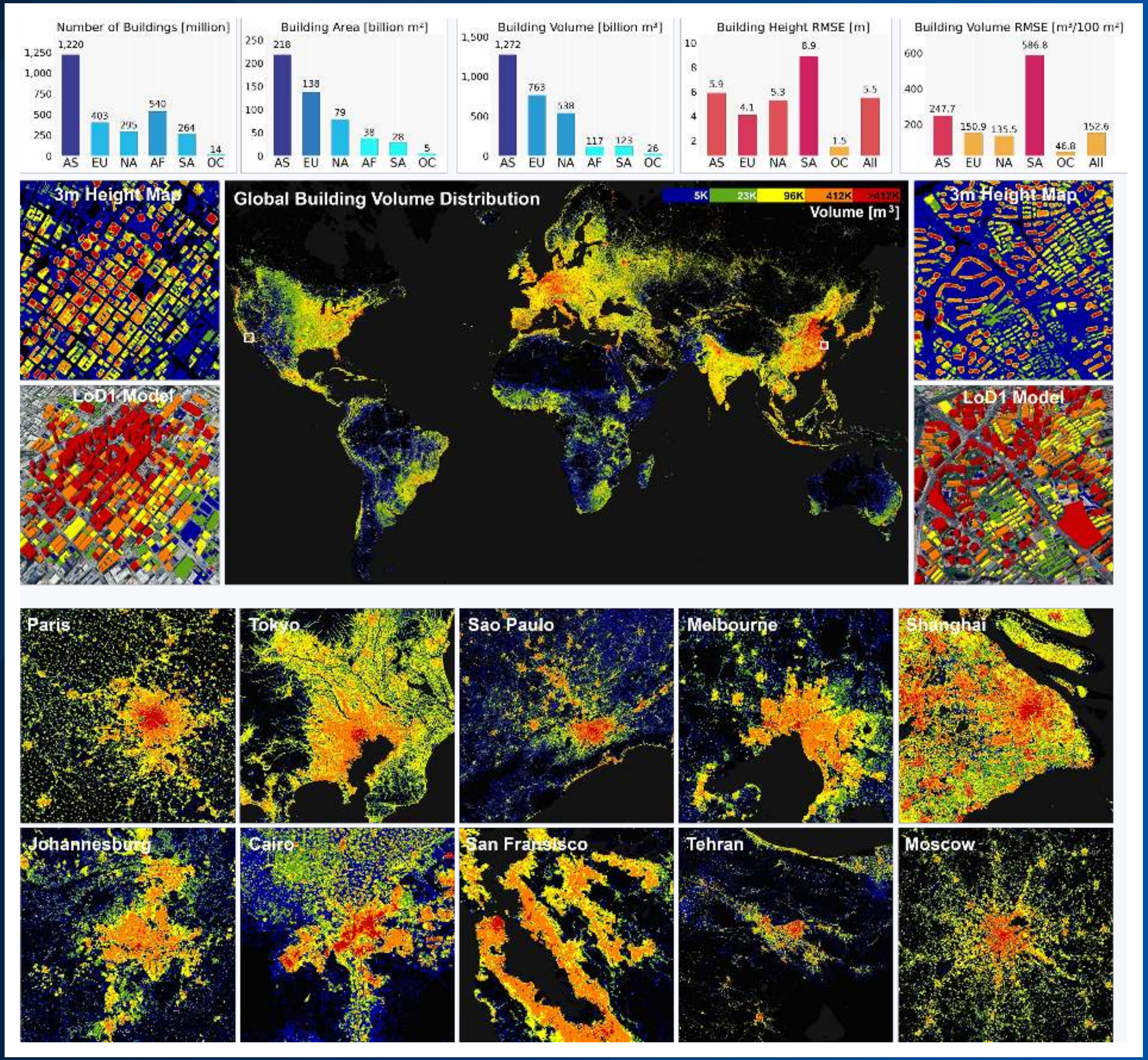
| | |
|---|---|
| LoD0 | Building footprints, roof prints, roads, terrain surfaces |
| LoD1 | Buildings are vertical boxes with flat roofs |
| LoD2 | Roof shapes, roof planes, sometimes building parts |
| LoD3 | Windows, doors, façade details, balconies, roof/facade structures |

**Urban Scene = Buildings + Terrain + Vegetation + Land Cover + Infrastructure**

# Urban Geometry Data Sources

Overview of common urban data sources for creating simplified or realistic CFD domains.

| Source | Typical content | Strength | Watch out |
|---|---|---|---|
| Official LoD2 / CityGML / CityJSON | buildings with roof shapes, IDs, attributes | good provenance, semantic structure | licensing, missing terrain/vegetation |
| National datasets: 3DBAG, LoD2-DE/Bavaria | large-scale building stock at LoD1/LoD2 | consistent regional coverage | varying attributes and update cycles |
| City portals: Helsinki, Berlin, NYC, etc. | city model, terrain, mesh, sometimes textures | often easy to access per tile | format conversion required |
| OSM + footprints + heights | global coverage, variable tags | good fallback for early prototypes | quality varies; OSM license obligations |
| LiDAR / photogrammetry reconstruction | terrain + DSM/point clouds | can recover missing geometry | high processing cost |
| Manual/parametric modeling | idealized blocks/canyons | controlled experiments | less realistic morphology |



Number of Buildings [million]
Building Area [billion m²]
Building Volume [billion m³]
Building Height RMSE [m]
Building Volume RMSE [m³/100 m²]
AS EU NA AF SA OC
AS EU NA SA OC All
3m Height Map
LoD1 Model
Global Building Volume Distribution
5K 23K 96K 412K >412K
Volume [m³]
3m Height Map
LoD1 Model
Paris
Tokyo
Sao Paulo
Melbourne
Shanghai
Johannesburg
Cairo
San Fransisco
Tehran
Moscow

# What is the problem when using such constructed urban areas as basis for a dataset?

# Sim-to-Real Gap

Even a high-fidelity static urban model with LoD3 buildings, vegetation, terrain, and surface information cannot fully reproduce real urban turbulence, because dynamic elements such as traffic-induced turbulence, pedestrians, temporary objects, heat emissions, and time-varying surface states are often missing or simplified.

# Spacing for urban turbulences

Use the target physics to decide how much of the city must be resolved.

### Coarse urban climate

$\Delta \approx 10\text{–}40\ \mathrm{m}$

City-scale thermal or climate screening. Buildings may be parameterized or partly resolved.

### Building-resolving flow

$\Delta \approx 1\text{–}5\ \mathrm{m}$

Urban wind, ventilation, pedestrian or drones. Often the relevant regime for turbulence datasets.

### High-detail turbulence

$\Delta \approx 0.5\text{–}2\ \mathrm{m}$

Research-grade LES around buildings. Expensive but better for shear layers and near-ground statistics.

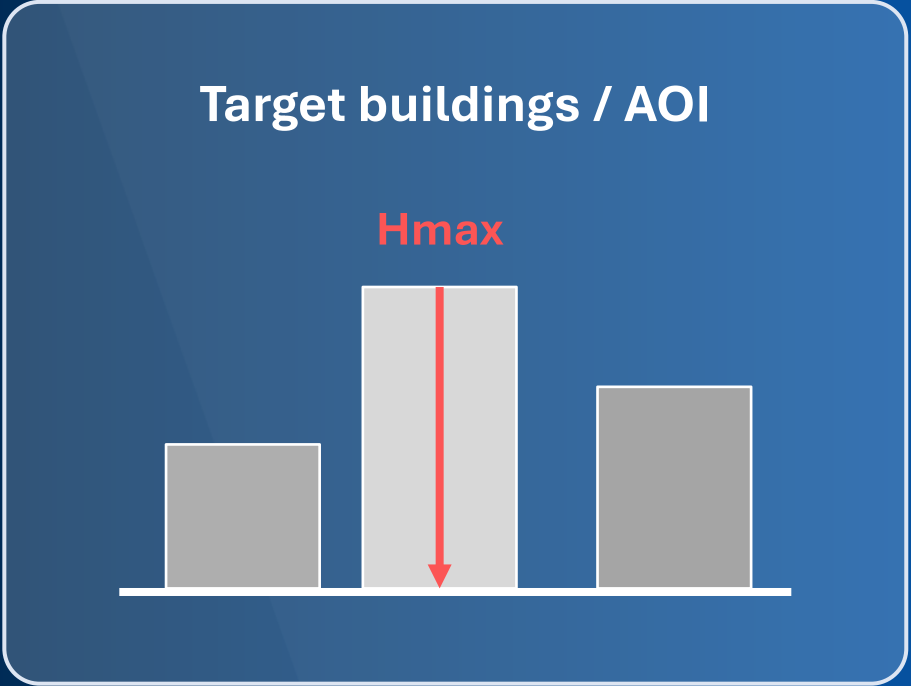
# Define the reference height first

Boundary-distance rules scale with a characteristic building height H.

H = max building height in the target / assessment area

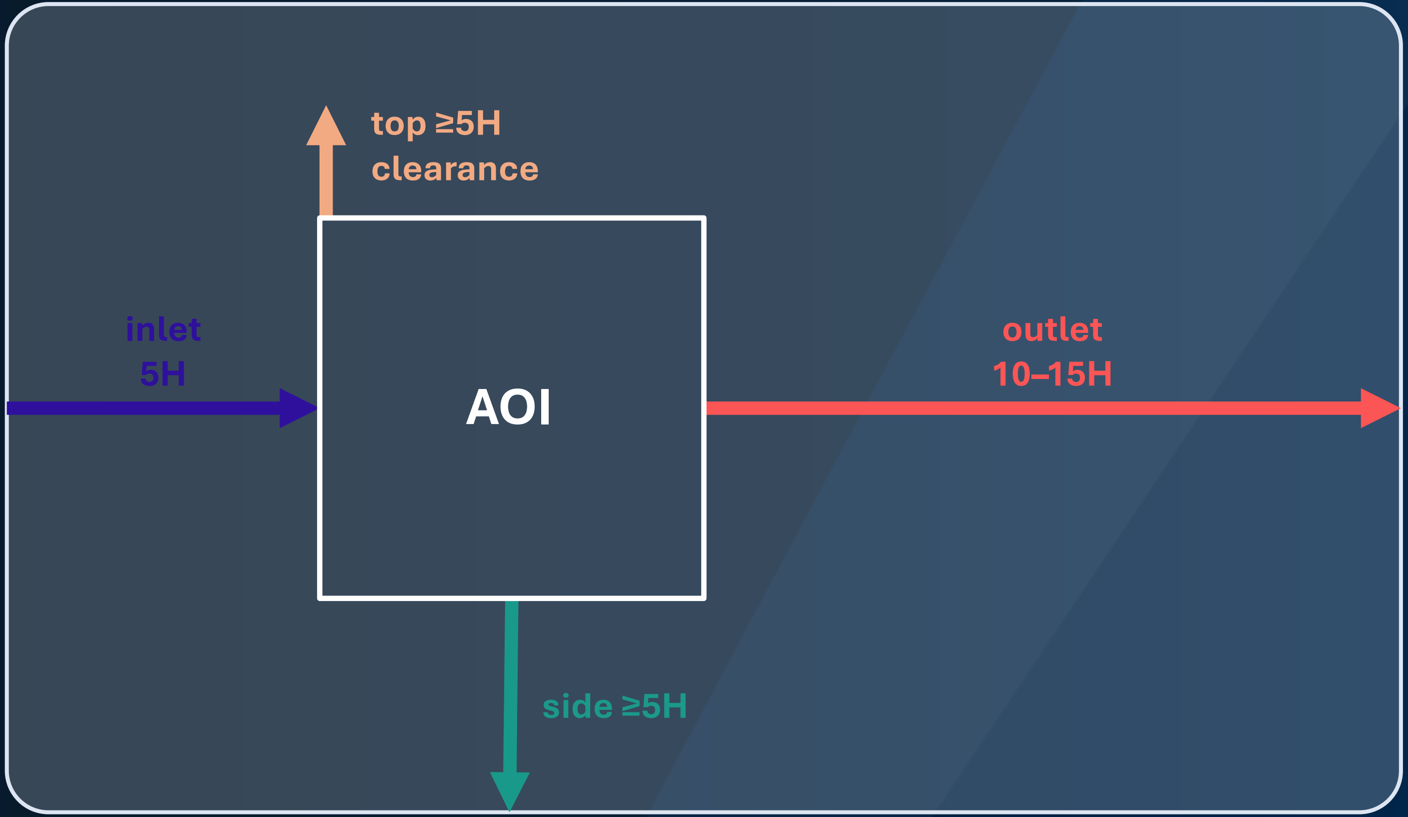
AOI = area where the CFD results will become training labels

Use Hmax for conservative domain sizing and for a district, H should represent the tallest relevant obstacle, not the average building height.



# Prominent boundary-distance

AIJ/COST-style guidelines provide starting values. Sensitivity tests remain necessary.




L_upstream ≥ 5H

L_downstream ≥ 10/15 H (AIJ/COST)

L_lateral ≥ 5H

L_top clearance ≥ 5H

# Urban Area Wind Prediction

Local heat, ventilation, comfort, vegetation effects, urban planning scenario.

$H = 40\ \mathrm{m}$

$AOI = 500\ \mathrm{m} \times 500\ \mathrm{m}$

$\Delta x = \Delta y = \Delta z = 2\ \mathrm{m}$

$\mathrm{Lup} = 5\mathrm{H} = 200\ \mathrm{m}$

$\mathrm{Ldown} = 15\mathrm{H} = 600\ \mathrm{m}$

$\mathrm{Lside} = 5\mathrm{H} = 200\ \mathrm{m}$

$\mathrm{ztop} \geq \mathrm{H} + 5\mathrm{H} = 240\ \mathrm{m}$

$\mathrm{L}x \approx 200 + 500 + 600 = 1300\ \mathrm{m}$

$\mathrm{L}y \approx 200 + 500 + 200 = 900\ \mathrm{m}$

$\mathrm{N}z \approx 240/2 = 120$ grid levels

$\mathrm{N}cells \approx 650 \cdot 450 \cdot 120 = 35.1$ million

# Urban Area Wind Prediction

Local heat, ventilation, comfort, vegetation effects, urban planning scenario.

$\mathrm{L}x \approx 200 + 500 + 600 = 1300\ \mathrm{m}$

$\mathrm{L}y \approx 200 + 500 + 200 = 900\ \mathrm{m}$

$\mathrm{N}z \approx 240/2 = 120\ \text{grid levels}$

$\mathrm{N}cells \approx 650 \cdot 450 \cdot 120 = 35.1\ \text{million}$

$C = (u, v, w, p, q,\ \text{mask}, \text{SDF})$

$Nt = 60$

$60 \cdot 31.5\ \text{million} \cdot 7 \cdot 4\ \textbf{bytes}$

$= 52{,}920{,}000{,}000\ \textbf{bytes} = 52.92\ \text{GB}$

$$\text{Memory snapshot} \approx Nt \cdot \mathrm{N}cells \cdot C \cdot 4\ \text{bytes} \quad (\text{float32})$$

# Domain nesting: coarse context, fine target region

Nested grids let you include city-scale context without forcing 2 m cells everywhere.

**Outer Context**
$\Delta \approx 8 - 16$ m

**Intermediate**
$\Delta \approx 4 - 8$ m

**Target**
$\Delta \approx 2 - 4$ m

- outer domain provides topography and larger-scale boundary-layer context
- inner domain resolves flow-building interactions
- training labels can be exported only from the high-resolution AOI

# Which design is better for a generalizable neural operator?

## Option A

1 neighborhood × 64 directions × 5 speeds × 3 seeds = 960 cases

## Option B

40 neighborhoods × 8 directions × 3 speeds × 1 seed = 960 cases

What should the test set contain if we claim “generalization to unseen urban neighborhoods”?

# CFD Tools for Urban Flow Data Generation

Comparing simulation workflows by use case, setup effort, and suitability for machine learning datasets.

| Tool | Best use case | Input burden | Output suitability for ML |
|---|---|---|---|
| PALM / PALM-4U | building-resolving urban atmosphere and LES | static/dynamic drivers, urban input data | netCDF-oriented, good for array workflows |
| uDALES | urban LES, heat transfer, pollutant dispersion | triangulated urban surfaces / IBM | research-grade, requires custom export |
| OpenFOAM | flexible RANS/LES CFD and custom physics | surface cleanup + meshing | strong but needs conversion layer |
| WRF / WRF4PALM | mesoscale forcing for microscale domains | WRF outputs + PALM domain setup | useful for time-varying boundaries |
| ENVI-met | planning-scale microclimate and vegetation/thermal scenarios | model setup + surface/plant parameters | good for thermal scenario datasets |
| Commercial CFD | industrial automation and support | license + workflow automation | good internally; less open/reproducible |

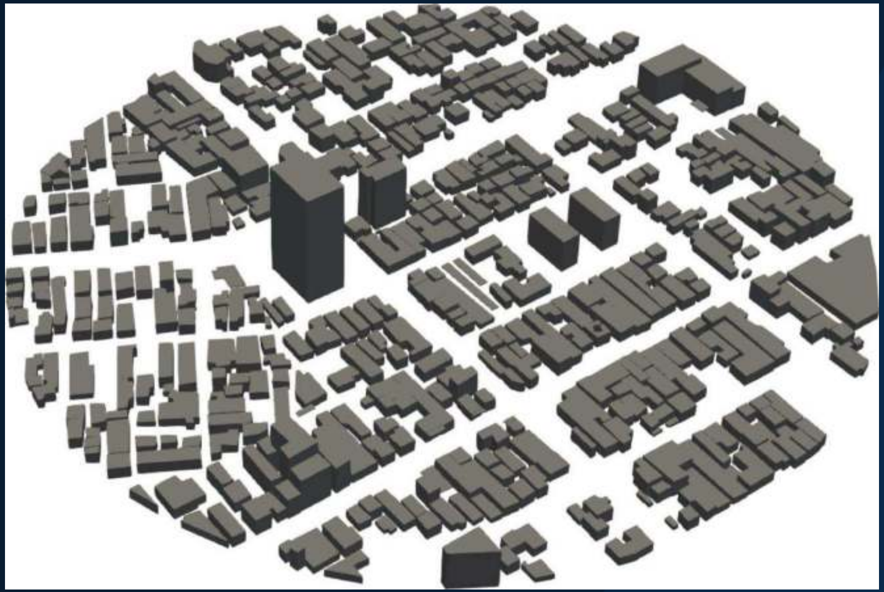
# Validation Strategy for Urban CFD Data

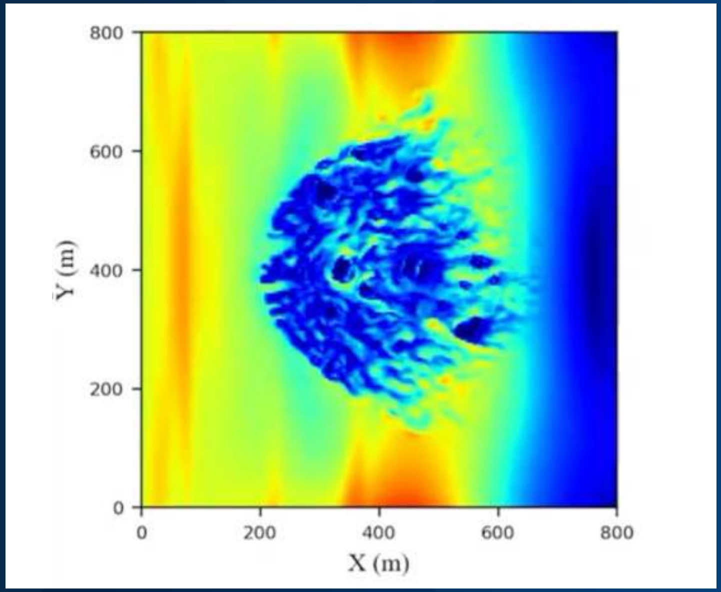
Overview of physical benchmarks and numerical checks used to assess simulation quality.

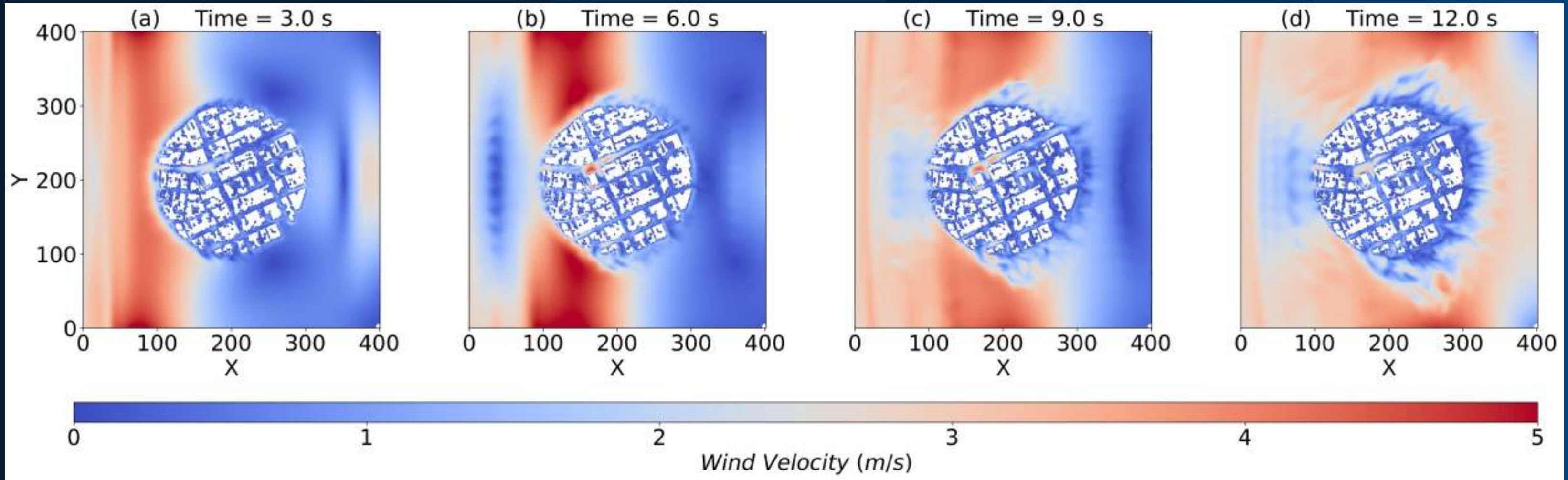
| Validation layer | What to compare | Role |
|---|---|---|
| Field measurements | wind speed/direction, temperature, humidity, surface temperature | real-world anchor; usually sparse |
| Wind tunnel benchmarks | pedestrian-level velocities, pressures, tracer data | controlled, repeatable reference |
| AIJ / COST-style cases | standardized urban wind environments | solver and setup validation |
| Internal CFD QA | ABL profile preservation, mass balance, convergence, grid sensitivity | detect numerical artefacts |
| Application metrics | comfort class, hot-spot detection, wake error, spectra | links dataset to target use |

# From urban models to CFD datasets

Convert geometric representations into simulation-ready flow data that can be used for analysis and machine learning.

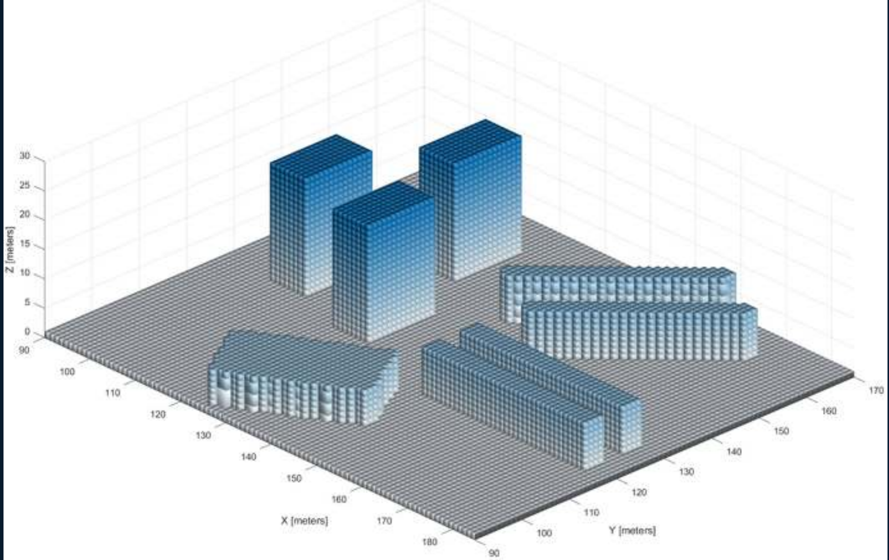


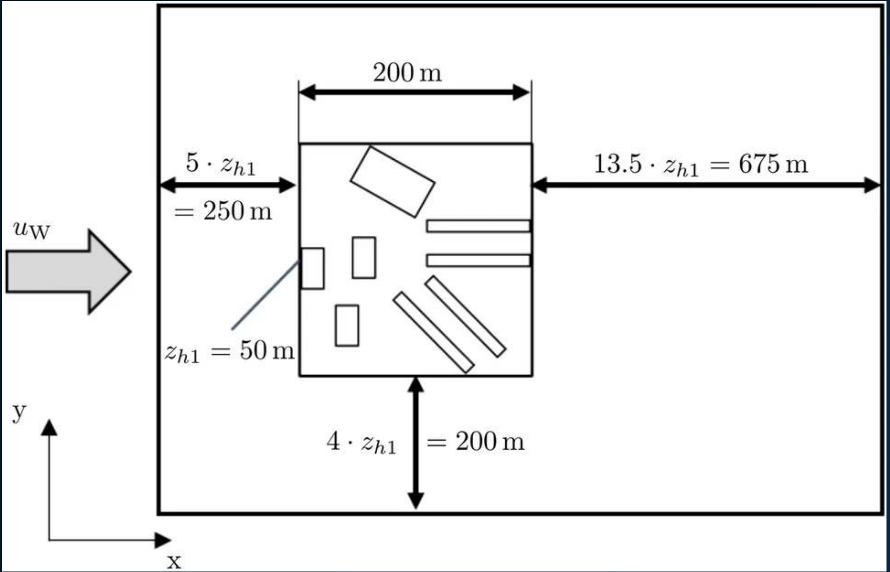


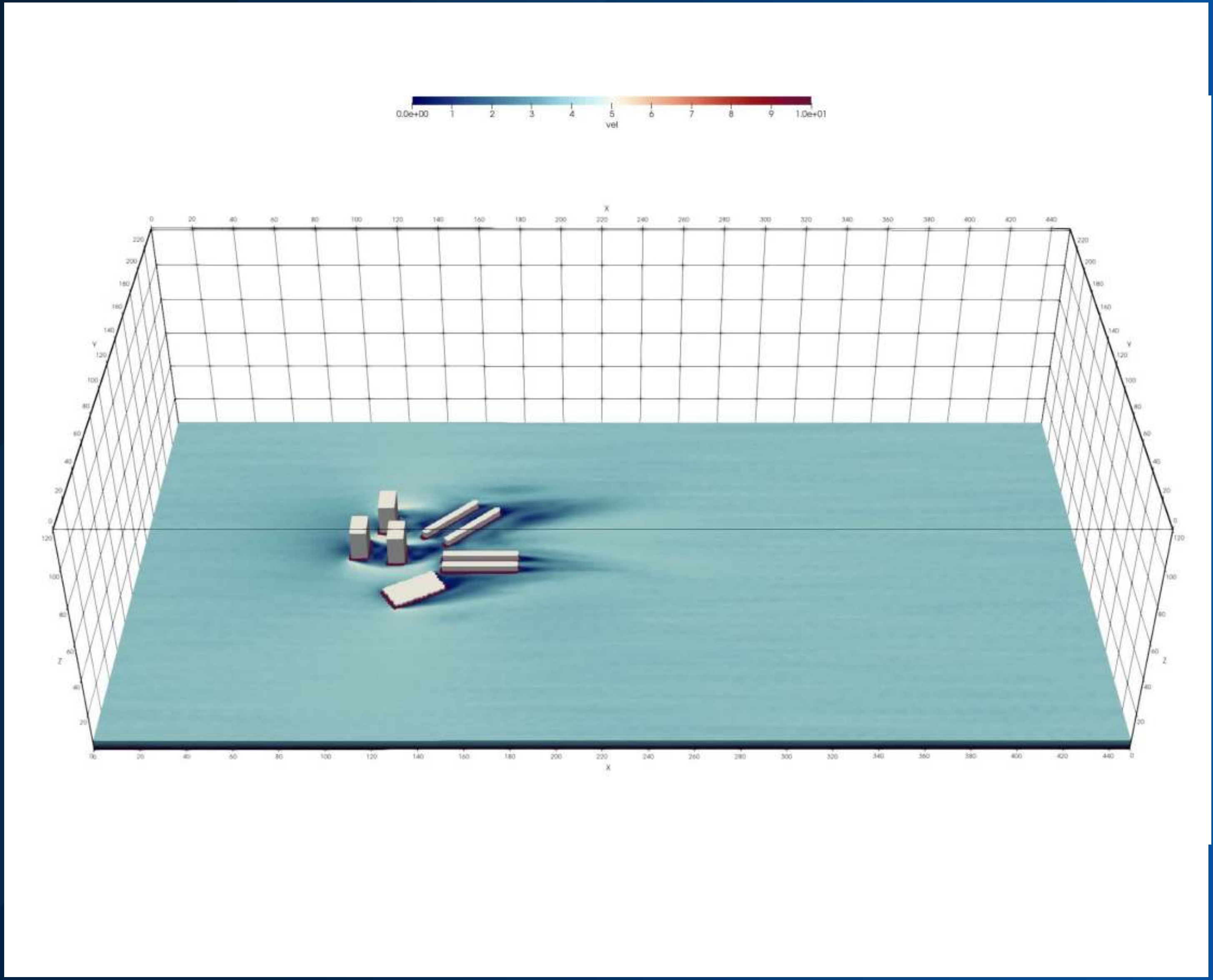



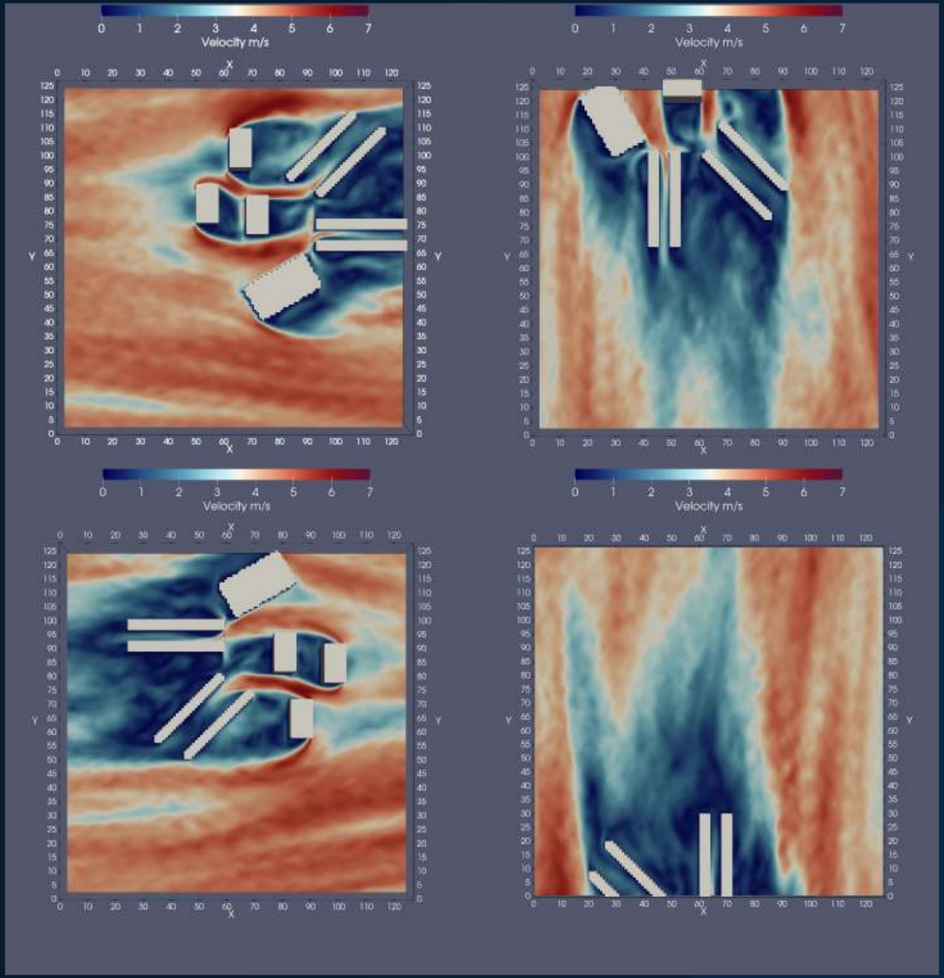

HM MUC.DAI
Munich Center for
Digital Sciences and AI

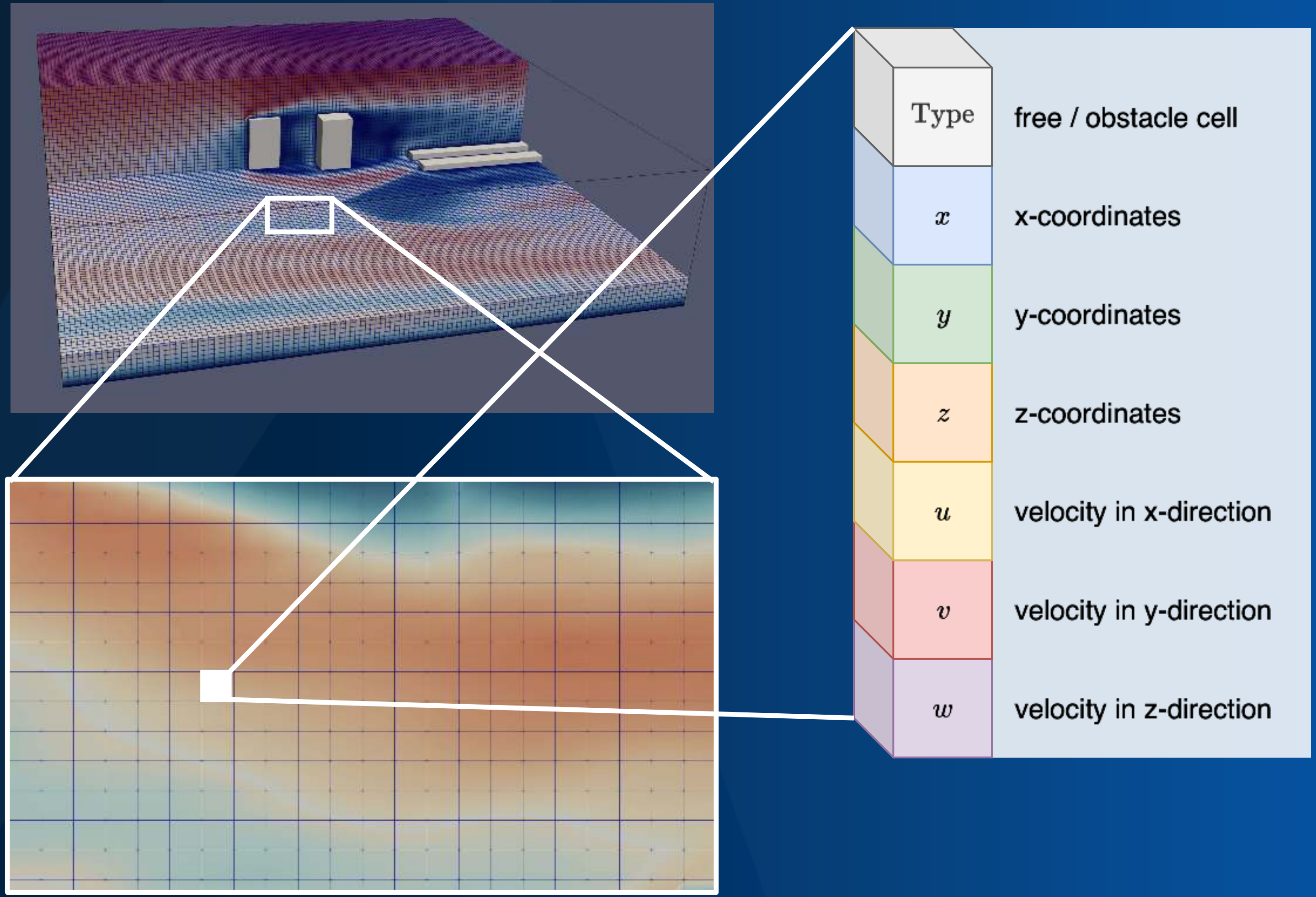
CIMPA


Z [meters]
X [meters]
Y [meters]


200 m
$5 \cdot z_{h1}$
$= 250\,\mathrm{m}$
$13.5 \cdot z_{h1} = 675\,\mathrm{m}$
$u_W$
$z_{h1} = 50\,\mathrm{m}$
$4 \cdot z_{h1}$
$= 200\,\mathrm{m}$
y
x


0.0e+00
1
2
3
4
5
6
7
8
9
1.0e+01
vel


Velocity m/s
Velocity m/s
Velocity m/s
Velocity m/s


Type
free / obstacle cell
$x$
x-coordinates
$y$
y-coordinates
$z$
z-coordinates
$u$
velocity in x-direction
$v$
velocity in y-direction
$w$
velocity in z-direction

# Global Weather Simulation

| | | |
|---|---|---|
| 1 | **Define task** | mean comfort map, dynamic wind forecast, thermal microclimate, pollutant spread |
| 2 | **Acquire geometry** | LoD2/CityGML/CityJSON, terrain, vegetation, surfaces, licenses |
| 3 | **Prepare domain** | clean, simplify, mesh/voxelize, set boundaries, define grid |
| 4 | **Design cases** | directions, speeds, stability, weather episodes, surface scenarios |
| 5 | **Run + validate** | CFD/LES/microclimate model, monitoring, field/wind-tunnel checks |
| 6 | **Export ML dataset** | Zarr/HDF5/NetCDF, masks/SDF, metadata, splits, documentation |

# UrbanTales Case Study

A realistic dataset pipeline from city geometry to PALM LES labels.

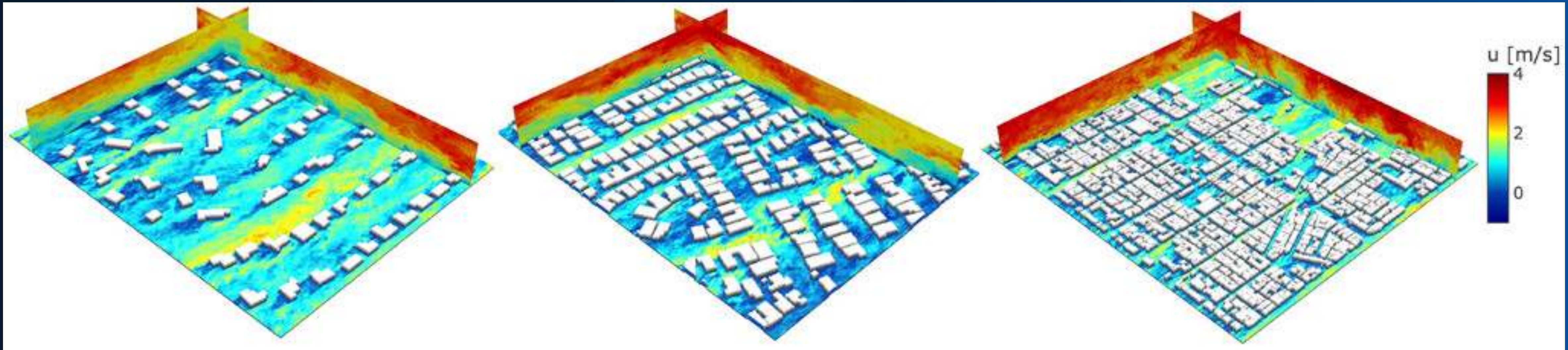
# UrbanTALES makes the dataset pipeline explicit

The dataset combines controlled idealized arrays and realistic neighborhoods under one LES setup.

538 LES cases
3,000,000+ CPU hours
35 TB of storage

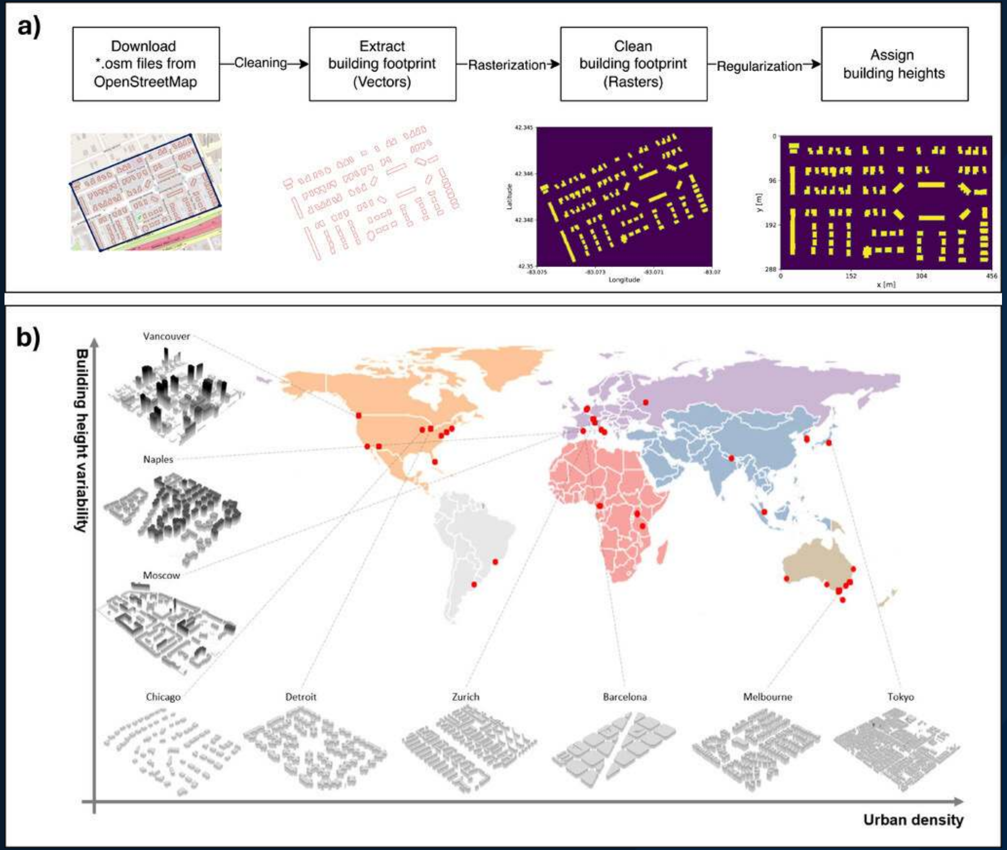
314 realistic cases
160 uniform-height
154 variable-height

224 idealized arrays
aligned / staggered
height + density sweeps



# Everything starts from real city geometry

UrbanTALES regularizes heterogeneous map data into the raster topography PALM can simulate.




1 Select an AOI
Bounding box defines one neighborhood block.

2 Gather footprints
OSM is primary; Microsoft Open Buildings fills gaps.

3 Rasterize for PALM
Vector footprints become 1 m building-height rasters.

4 Assign heights
OSM height, WSF3D, or prescribed defaults.

# Everything starts from real city geometry

UrbanTALES regularizes heterogeneous map data into the raster topography PALM can simulate.

**1. Download** — *.osm files for bounding box → **2. Extract** — building footprints as vectors → **3. Rasterize** — convert polygons to grid → **4. Regularize** — rotate/crop clean domain → **5. Assign heights** — fill each building cell

$$\text{OSM polygons} \xrightarrow[\Delta x = \Delta y = 1\,\mathrm{m}]{\text{OSM2LES}} h_b(x_j, y_k)$$

# Rasterized urban topography

UrbanTALES regularizes heterogeneous map data into the raster topography PALM can simulate.

$$x_j = x_{\min} + j\Delta x, \quad y_k = y_{\min} + k\Delta y, \quad \Delta x = \Delta y = 1\,\mathrm{m}$$

$$h_{j,k} = h_b\left(x_j, y_k\right), \quad h_{j,k} = 0 \text{ for nonbuilding cells}$$

$$\mathbf{H} = \left[h_{j,k}\right] \in \mathbb{R}^{N_x \times N_y}$$

### Interpretation

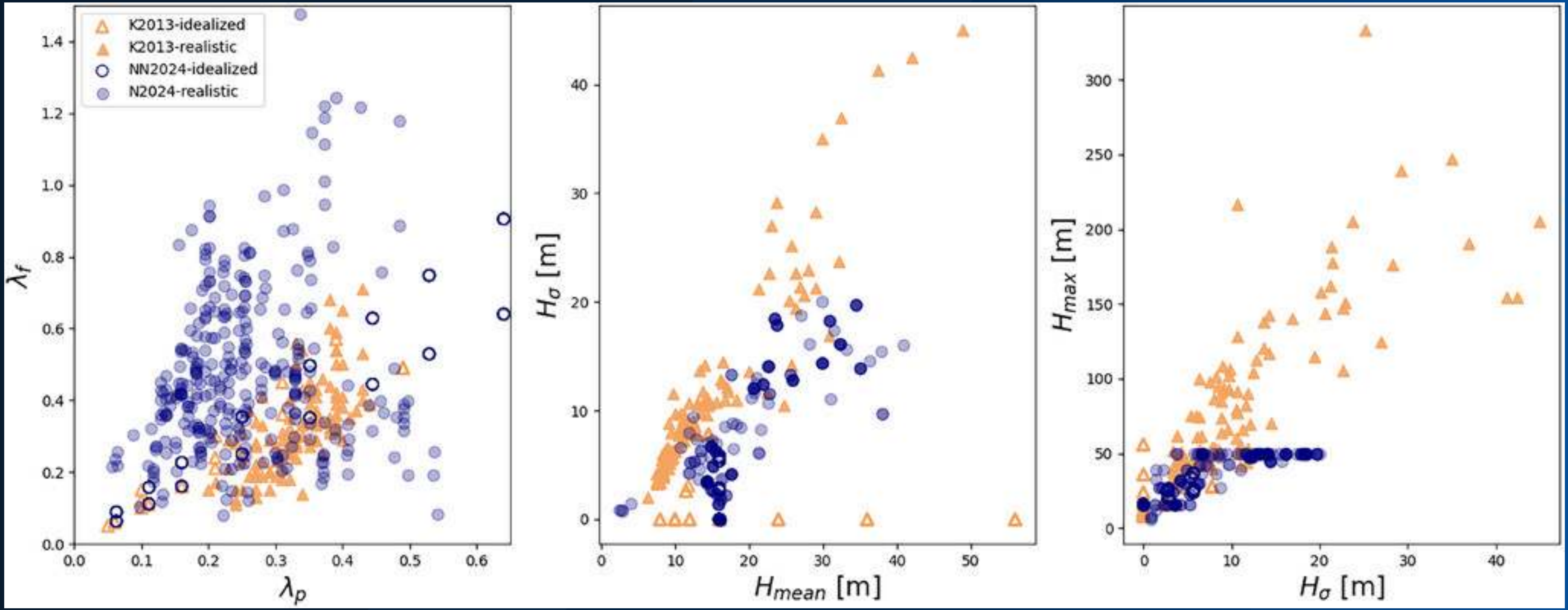
The static urban scene is stored as a 2D height field. A grid cell is either air/ground (0 m) or a building column with a prescribed height.

### Why this matters

This makes the geometry compatible with PALM while enabling systematic morphology statistics and reproducibility across hundreds of cases.

# Areas sampled across density, height and context.

The realistic cases cover global urban morphology rather than repeating a single neighborhood.



# Areas sampled across density, height and context.

The realistic cases cover global urban morphology rather than repeating a single neighborhood.

$$H_T = 5\text{–}8\, H_{\text{mean}}, \qquad H_T \approx 130\ \text{m}$$

$$\Delta x = \Delta y = 1\ \text{m} \quad (\text{realistic}), \qquad \Delta x = \Delta y = 0.5\ \text{m} \quad (\text{idealized})$$

$$\Delta z = \text{uniform below 50 m}, \qquad \Delta z_{n+1} = 1.1 \Delta z_n \quad (z > 50\ \text{m})$$

# Areas sampled across density, height and context.

The realistic cases cover global urban morphology rather than repeating a single neighborhood.

| | |
|---|---|
| **Turbulence closure** | 1.5-order closure |
| **Vertical grid** | Arakawa staggered C grid |
| **Horizontal discretization** | Second-order central differences |
| **Time discretization** | Minimal storage scheme |
| **Pressure solver** | FFTW |
| **Lateral boundaries** | Periodic |
| **Surface boundary** | Neumann following log law, $z_0$ = 0.01 m |
| **Top boundary** | Dirichlet, no slip |
| **Topography** | 2D building-height ASCII field |

# Areas sampled across density, height and context.

The realistic cases cover global urban morphology rather than repeating a single neighborhood.

$$u_i(\mathbf{x}, t) = \overline{u_i}(\mathbf{x}) + u_i'(\mathbf{x}, t)$$

First moments: $\overline{u}, \overline{v}, \overline{w}, \overline{U}, \overline{p}, \overline{c}$

Second moments: $\overline{u_i' u_j'}$

Third moments: $\overline{u_i' u_j' u_k'}$

**Why moments?**

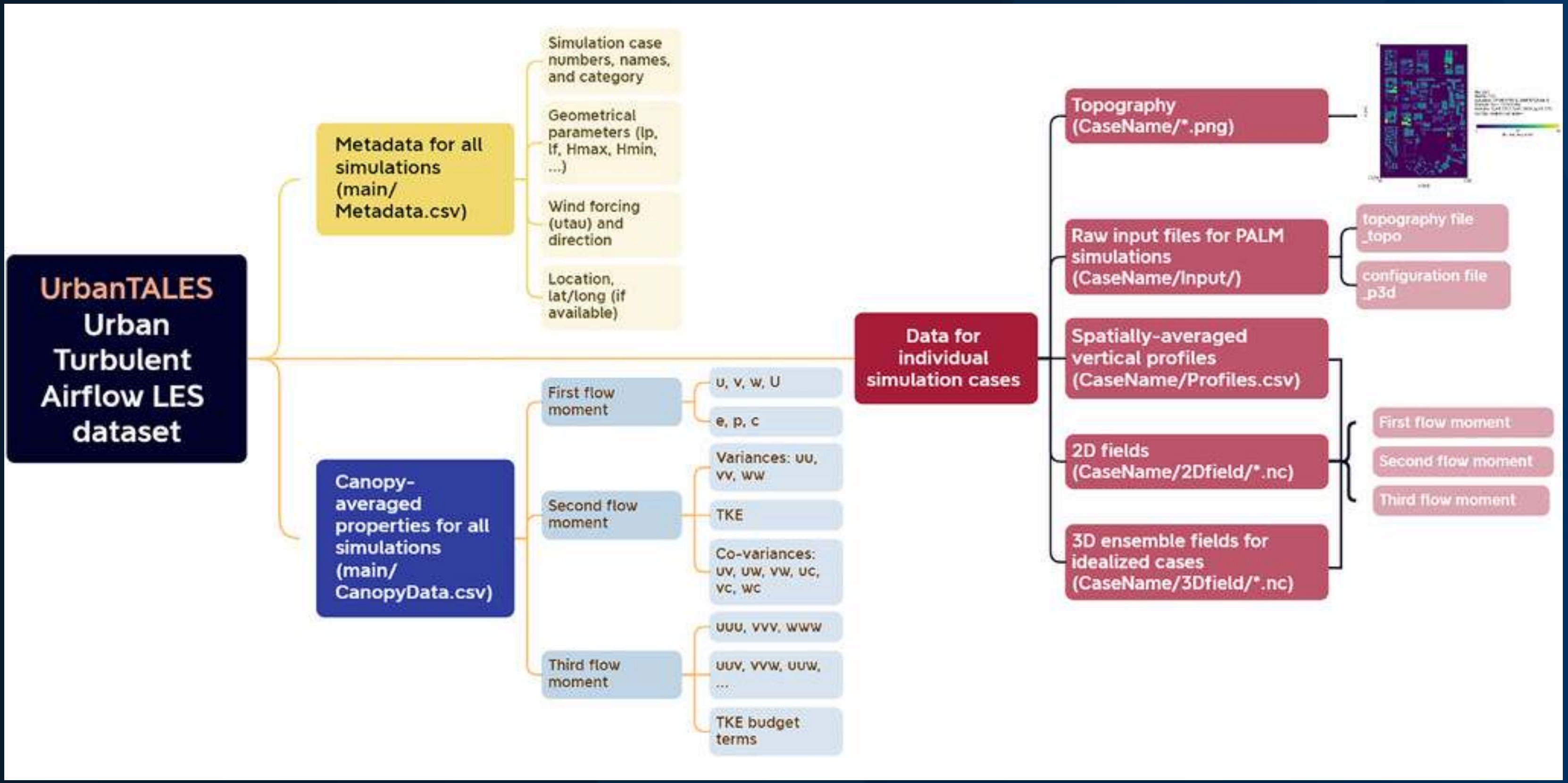
Mean fields describe average ventilation; second moments quantify turbulence intensity and fluxes; third moments support budget terms.

**ML implication**

For FNO/PINO training, these are alternative targets: mean surrogate, turbulence-statistics surrogate, or budget-aware closure target.

# Areas sampled across density, height and context.

The realistic cases cover global urban morphology rather than repeating a single neighborhood.




Dataset-level: Metadata.csv and CanopyData.csv summarize all simulations.

Case-level: topography images, PALM input files, vertical profiles and 2D fields.

Labels: first, second and third flow moments, including TKE and covariance terms.

3D fields are provided for selected idealized cases; additional 3D data are request-based.

# Areas sampled across density, height and context.

The realistic cases cover global urban morphology rather than repeating a single neighborhood.

Realistic neighborhoods
$\Delta x = \Delta y = 1$ m
vertical grid: 1 m up to 50 m
stretching ratio 1.1 above 50 m

Idealized arrays
$\Delta x = \Delta y = \Delta z = 0.5$ m
cheaper geometry enables finer cells
used for systematic block sweeps

Domain height
HT ≈ 130 m
≈ 5–8 Hmean
Hmax capped at 50 m

Spinup
4 h
~140 eddy-turnover times

Average
8 h
~375 eddy-turnover times

Store
30 min
mean-statistics cadence

# From data to flow intelligence

AI can only learn what the dataset represents, from basic examples and rule-based thinking to complex urban CFD flow fields.

# Thank you for your attention!

## CIMPA Summer School 2026 | Course Part 1

**Maximilian Dauner**

**maximilian.dauner0@hm.edu**

Munich University of Applied Sciences

Munich Center for Digital Sciences and AI (MUC.DAI)

HM MUC.DAI Munich Center for Digital Sciences and AI

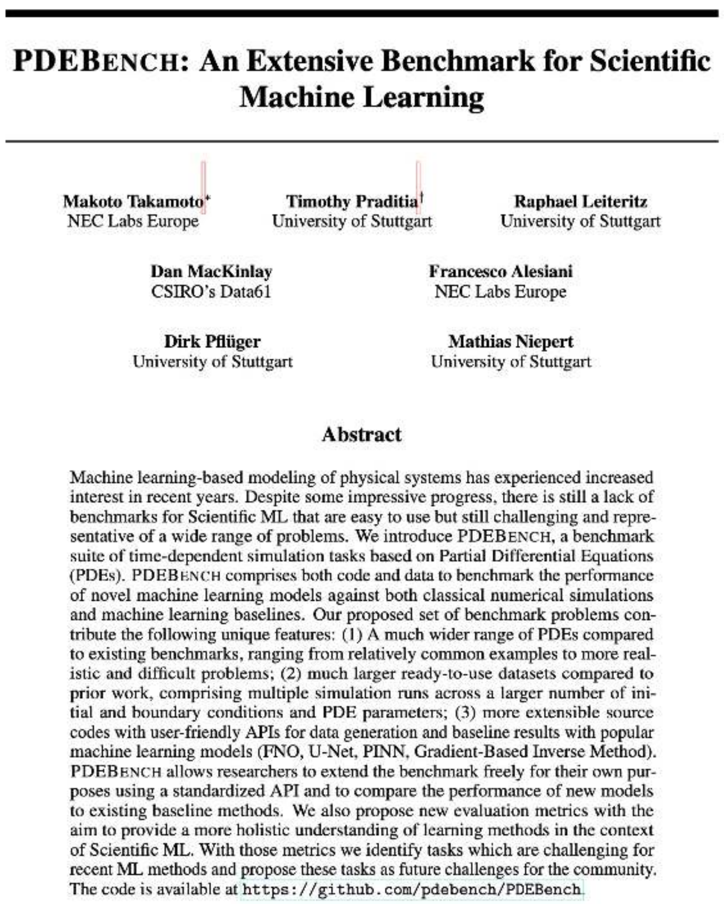
CIMPA

# Lab 1

## From PDE Simulations to Training Data

In this lab, you will explore a 1D PDEBench dataset and learn how physical simulation data is structured, visualized, analyzed, and transformed into supervised learning pairs.

By computing trajectory statistics, studying distributions, checking data splits, and testing simple baselines, you will understand why dataset quality and preprocessing are essential for reliable scientific machine learning.



**PDEBENCH: An Extensive Benchmark for Scientific Machine Learning**


Makoto Takamoto* NEC Labs Europe
Timothy Praditia† University of Stuttgart
Raphael Leiteritz University of Stuttgart
Dan MacKinlay CSIRO's Data61
Francesco Alesiani NEC Labs Europe
Dirk Pflüger University of Stuttgart
Mathias Niepert University of Stuttgart



**Abstract**

Machine learning-based modeling of physical systems has experienced increased interest in recent years. Despite some impressive progress, there is still a lack of benchmarks for Scientific ML that are easy to use but still challenging and representative of a wide range of problems. We introduce PDEBENCH, a benchmark suite of time-dependent simulation tasks based on Partial Differential Equations (PDEs). PDEBENCH comprises both code and data to benchmark the performance of novel machine learning models against both classical numerical simulations and machine learning baselines. Our proposed set of benchmark problems contribute the following unique features: (1) A much wider range of PDEs compared to existing benchmarks, ranging from relatively common examples to more realistic and difficult problems; (2) much larger ready-to-use datasets compared to prior work, comprising multiple simulation runs across a larger number of initial and boundary conditions and PDE parameters; (3) more extensible source codes with user-friendly APIs for data generation and baseline results with popular machine learning models (FNO, U-Net, PINN, Gradient-Based Inverse Method). PDEBENCH allows researchers to extend the benchmark freely for their own purposes using a standardized API and to compare the performance of new models to existing baseline methods. We also propose new evaluation metrics with the aim to provide a more holistic understanding of learning methods in the context of Scientific ML. With those metrics we identify tasks which are challenging for recent ML methods and propose these tasks as future challenges for the community. The code is available at https://github.com/pdebench/PDEBench.


HM

MUC.DAI
Munich Center for
Digital Sciences and AI

CIMPA

# Lab 1

## From PDE Simulations to Training Data

In this lab, you will explore a 1D PDEBench dataset and learn how physical simulation data is structured, visualized, analyzed, and transformed into supervised learning pairs.

By computing trajectory statistics, studying distributions, checking data splits, and testing simple baselines, you will understand why dataset quality and preprocessing are essential for reliable scientific machine learning.

https://syncandshare.lrz.de/getlink/fiVMAZBH1ha1h4Lrn4a98C/public_data


HM
MUC.DAI
Munich Center for Digital Sciences and AI
CIMPA