The importance of reliable datasets for AI algorithms from an emergency response preparedness perspective

Anna Wawrzyńczak-Szaban (a,b)

(a)University of Siedlce, Poland

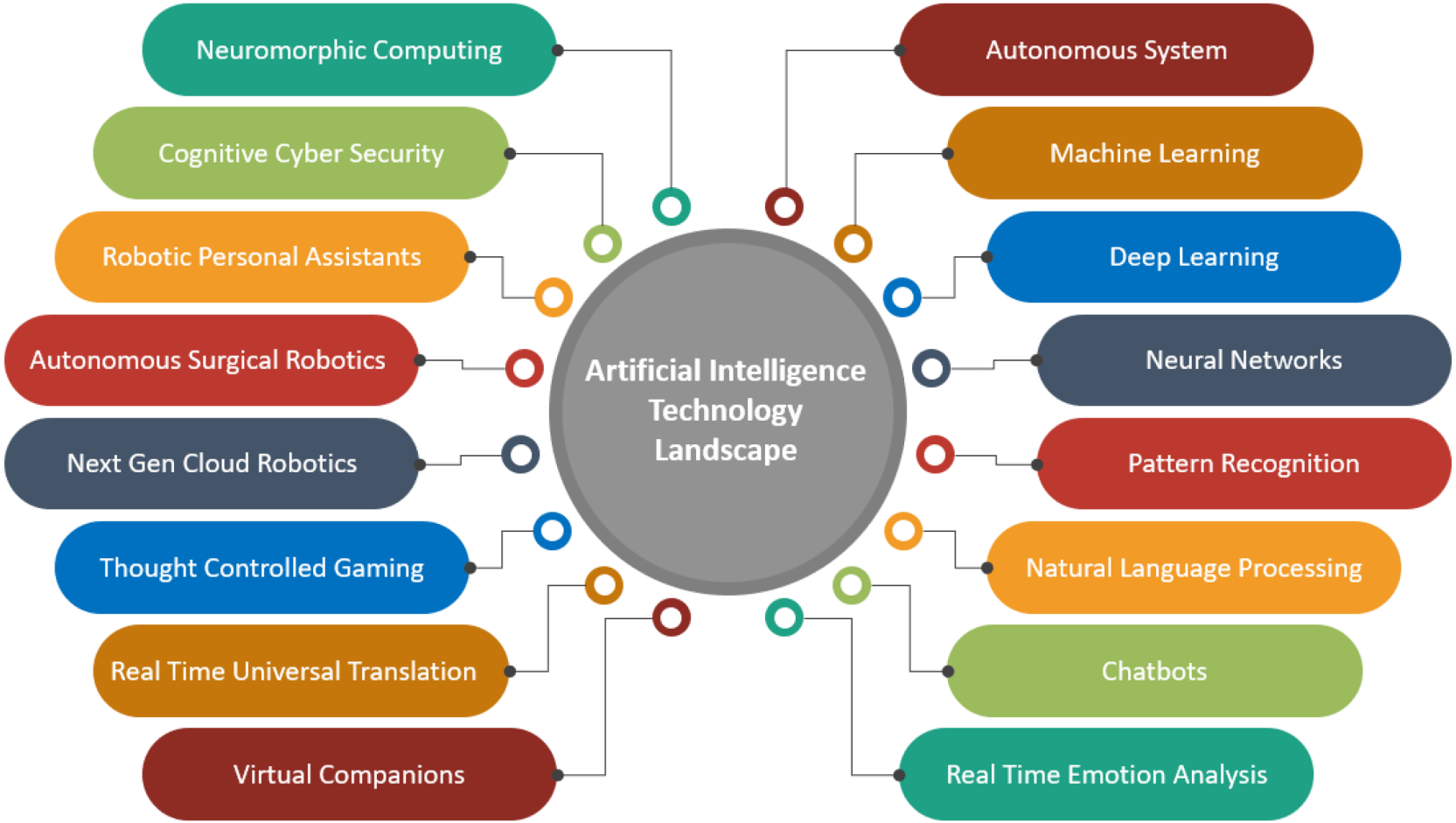
(b)National Centre for Nuclear Research, Poland

PITHIA-NRF Third Innovation Day
12 June 2024, Warsaw, Poland

Outline

- Importance of data for AI systems.
- Elements of the release source localization system.
- Why artificial neural networks?
- ANN models for urbanized area
- Conclusions

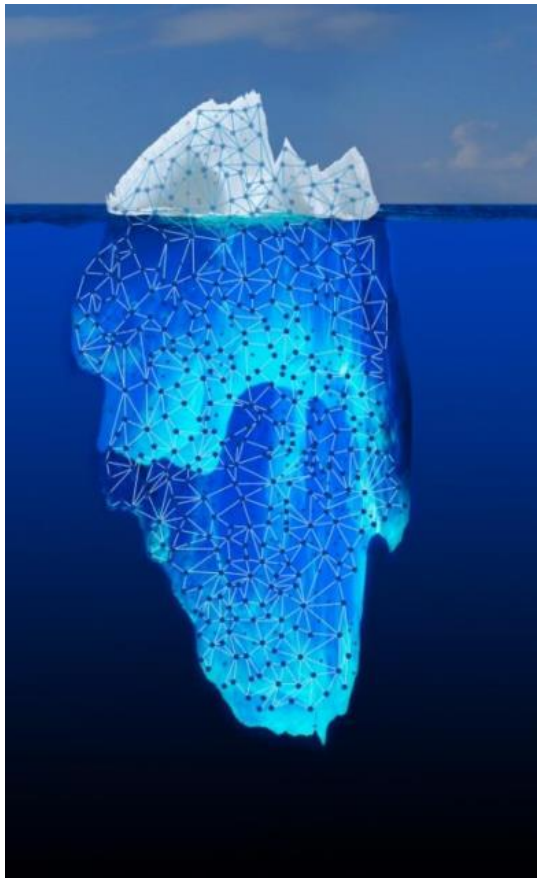
AI Lanscape



Aldoseri, A.; Al-Khalifa, K.N.; Hamouda, A.M. Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. *Appl. Sci.* **2023**, *13*, 7082. https://doi.org/10.3390/app13127082

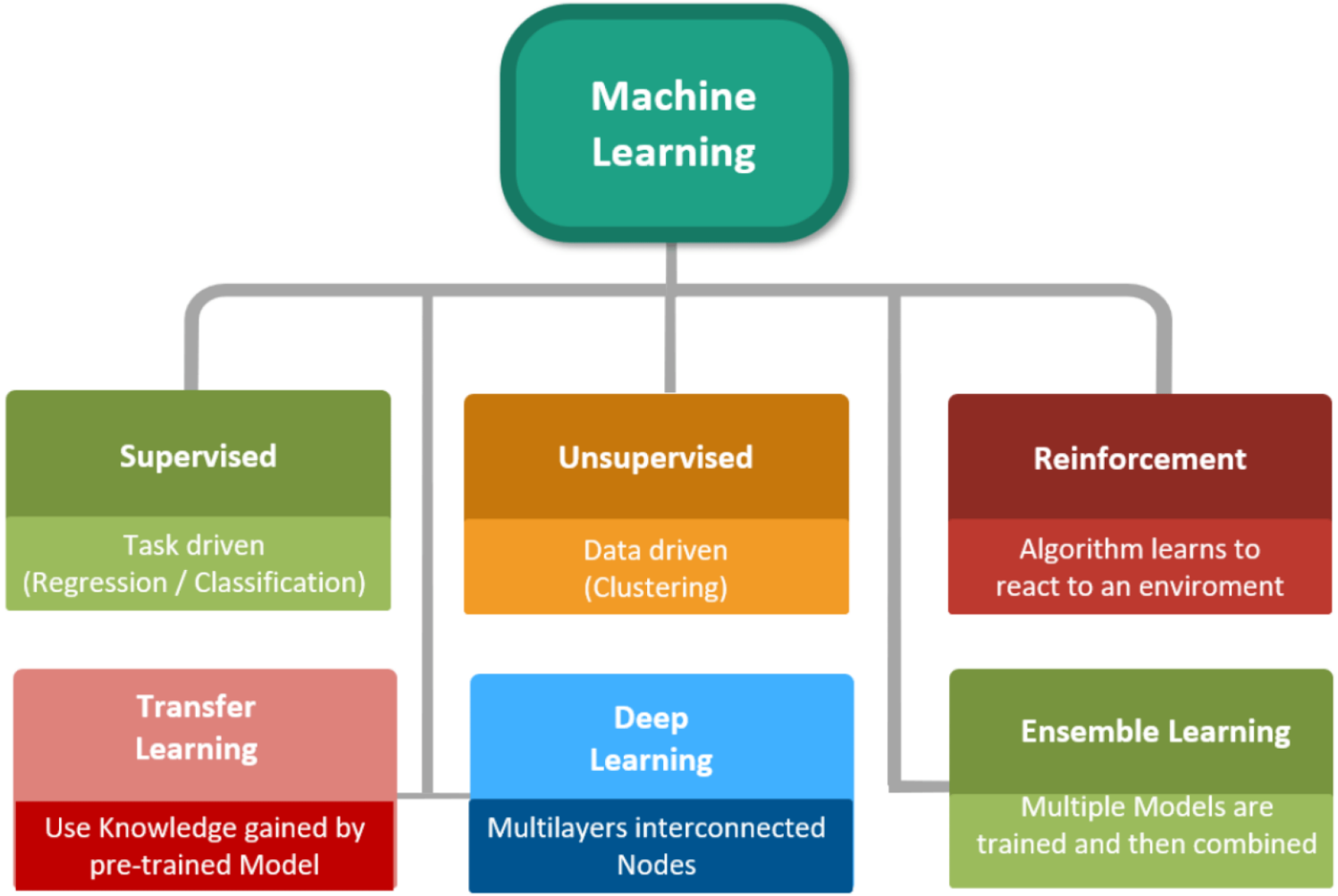
Data for AI

- Data are critical for AI because they are the foundation upon which machine learning algorithms learn, make predictions, and improve their performance over time.
- To train an AI model, large amounts of data are required to enable the model to recognize patterns, make predictions, and improve its performance over time.



Huge Datasets lead to better AI models

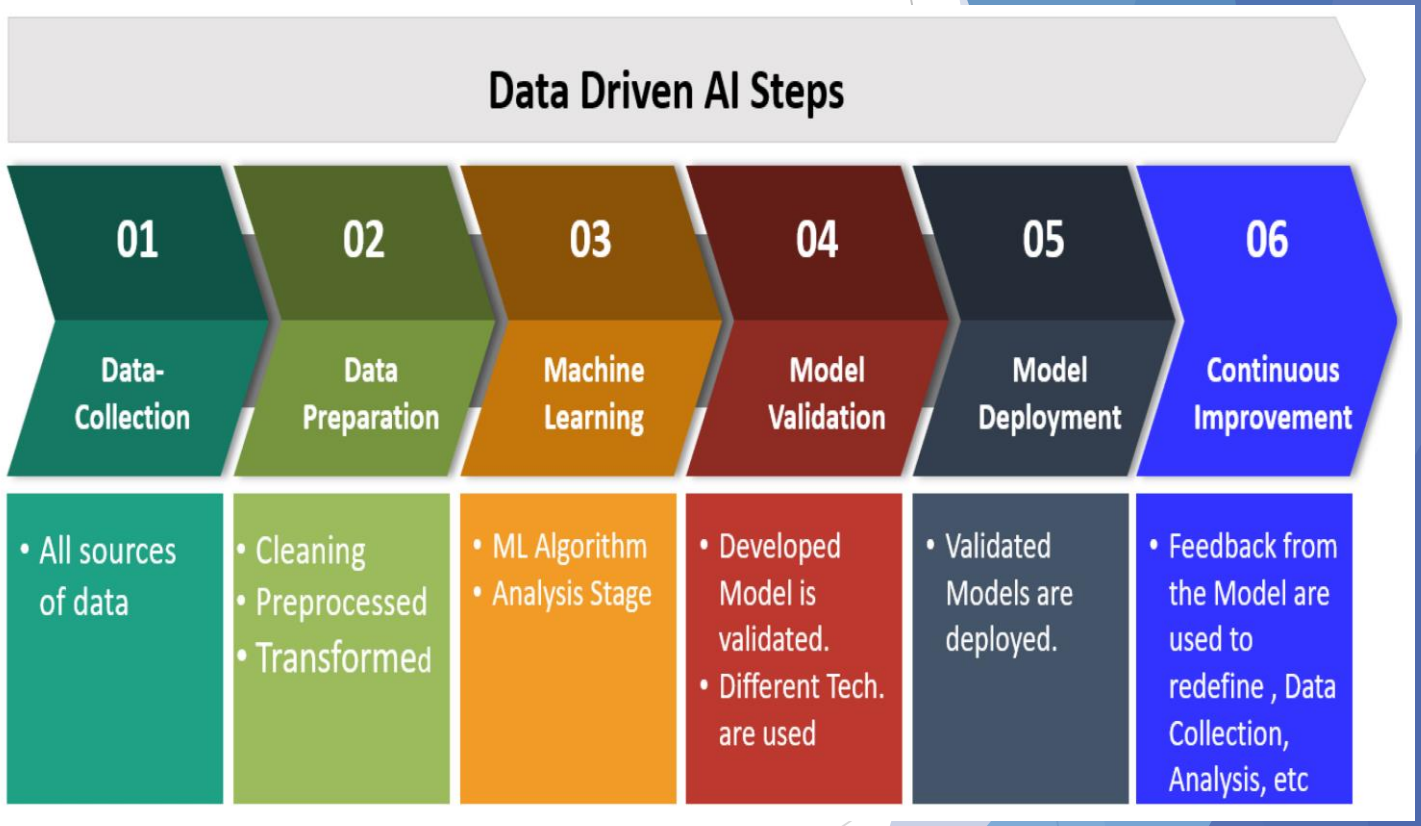
Data Learning Approaches



Aldoseri, A.; Al-Khalifa, K.N.; Hamouda, A.M. Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. *Appl. Sci.* **2023**, *13*, 7082. https://doi.org/10.3390/app13127082

Data-Driven Approach

- This focuses on building AI models that are specifically designed to make predictions or decisions based on data.
- This approach emphasizes the selection, processing, and analysis of data to identify patterns, relationships, and insights that can be used to improve the accuracy and performance of an AI model.
- The goal is to develop an AI model that can learn and adapt to new data without being constrained by a predefined set of rules or assumptions.
- This approach is particularly useful when data are relatively homogeneous or when the goal is to automate a specific decision-making process



Impact of Data Quality on AI

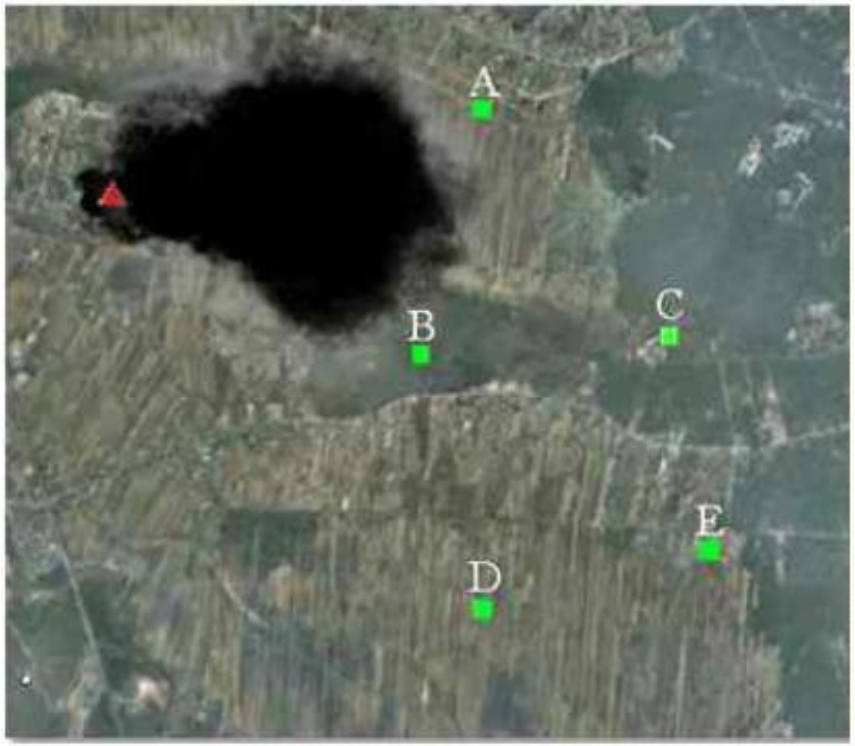
In the world of artificial intelligence, data is king.

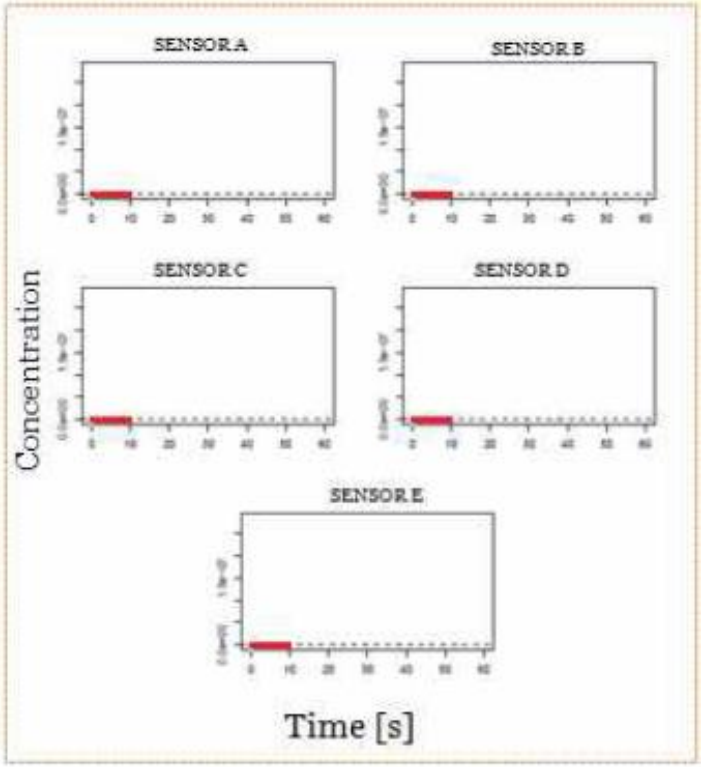
- The impact of data quality on AI is profound.
- AI models trained on poor-quality data will likely perform inadequately, as the insights they generate will be based on flawed or misleading information.
- **High-quality data**, on the other hand, increase the **accuracy** and performance of AI systems.

- Companies like Google, Amazon and Facebook dominated their industries because they were the first to begin building data sets. Their data sets have become so large, and their data collection and analysis so sophisticated that they are able to grow their competitive advantage.

Process of hazardous release

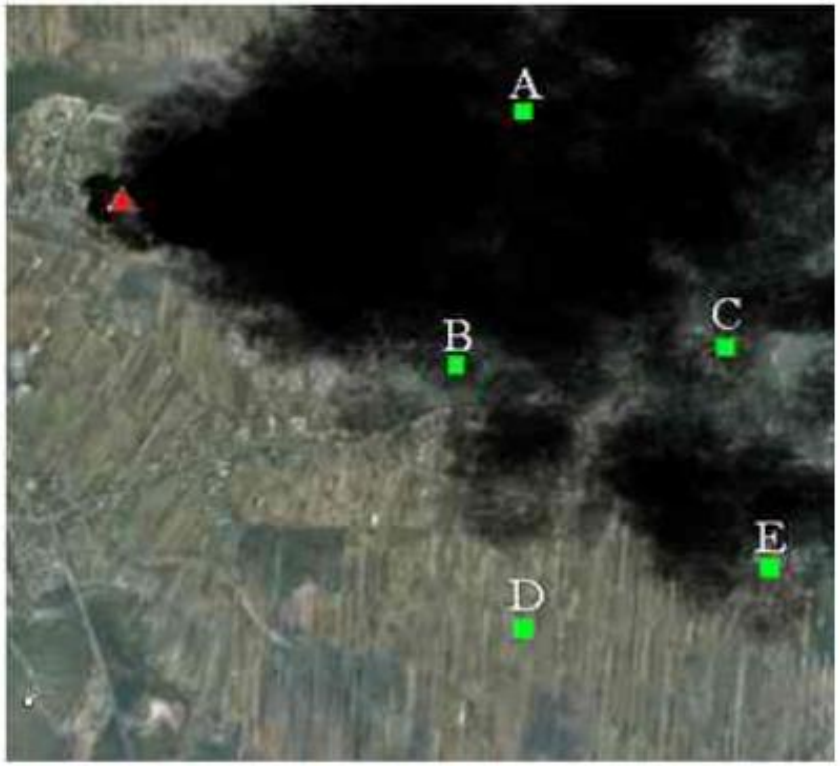
- Assume that we have some toxic substance accidentally released.
- We have the substance point-concentrations with some time interval.

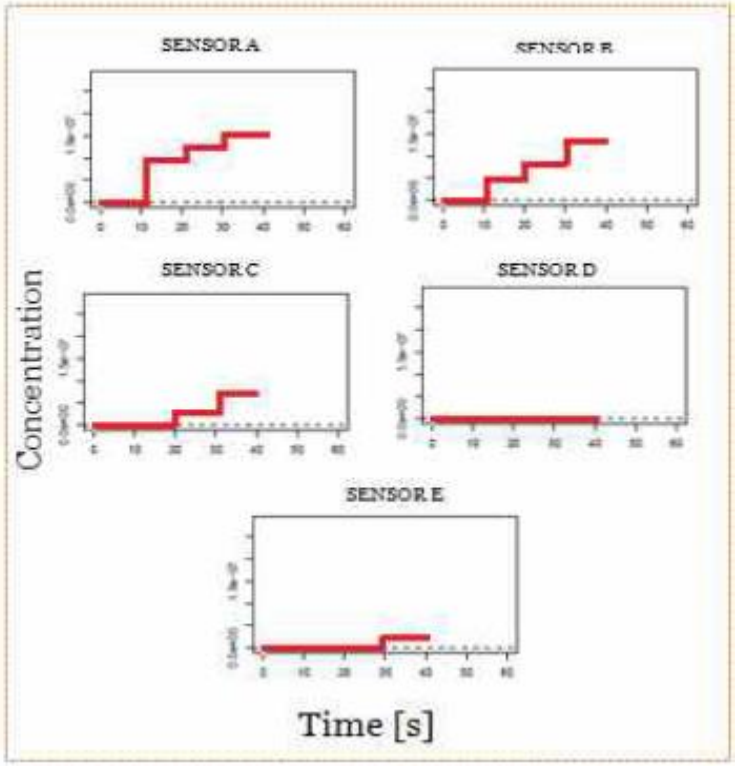




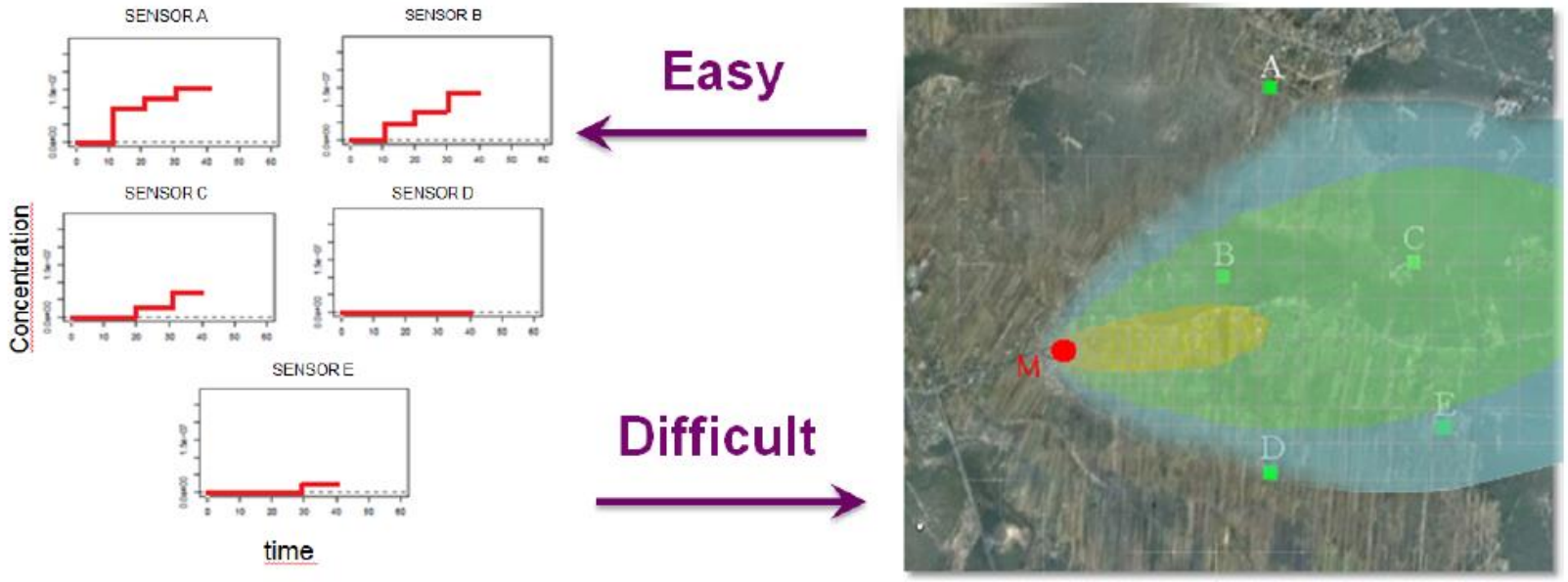
Process of hazardous release

- Assume that we have some toxic substance accidentally released.
- We have the substance point-concentrations with some time interval.





BUT...



Based on sparse-point substance concentrations we have to answer the questions

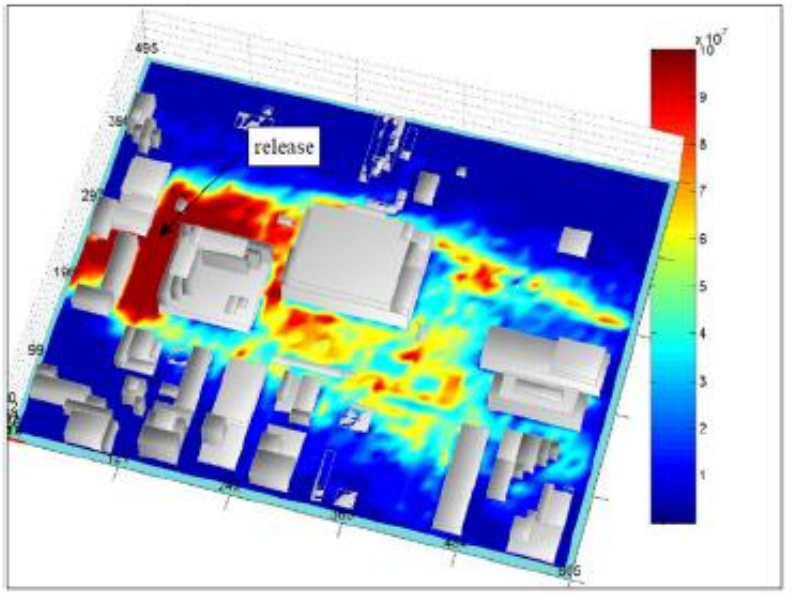
Where? How much?

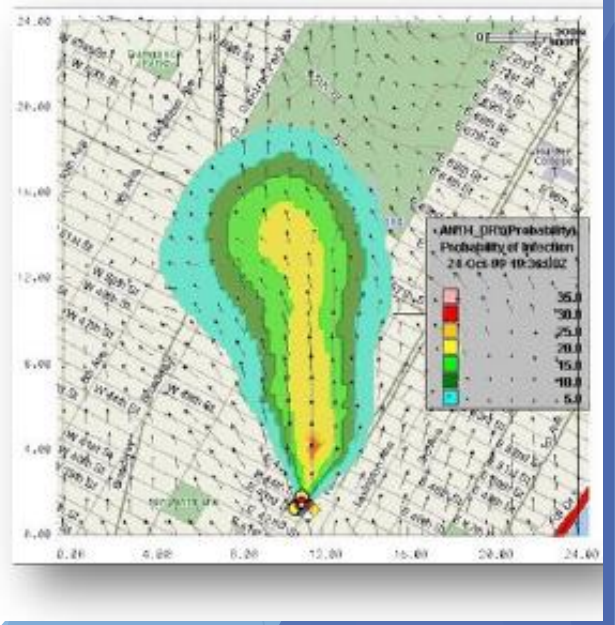
substance was released.

Time of giving the answer to this questions is crucial!

How to find a contamination source ?

- Build a model of contaminant transport in the atmosphere and compare point concentrations derived from the model with the measured data obtained from sensor networks
- Problem: Find the values of the dispersion model, (such as the location of the source of release) which will the best „fitted" our model to the observational data.
- Sampling of the model parameter space by e.g.
  - Sequential Monte Carlo
    [e.g. Wawrzynczak, Kopka, Lecture Notes in Computer Science,2014]
  - Aproximate Bayesian computation
    [e.g. Kopka, Wawrzynczak, Atmosph. Env. 2018],
  - Genetic algorithm
    [e.g. Wawrzynczak et. Al., Studies in Computational Inteligence, 2016]
  - Cellular automata
    [e.g. Szaban, Wawrzynczak, *J.* of Parallel and Distrib.d Computing, 2022*]*
  - …





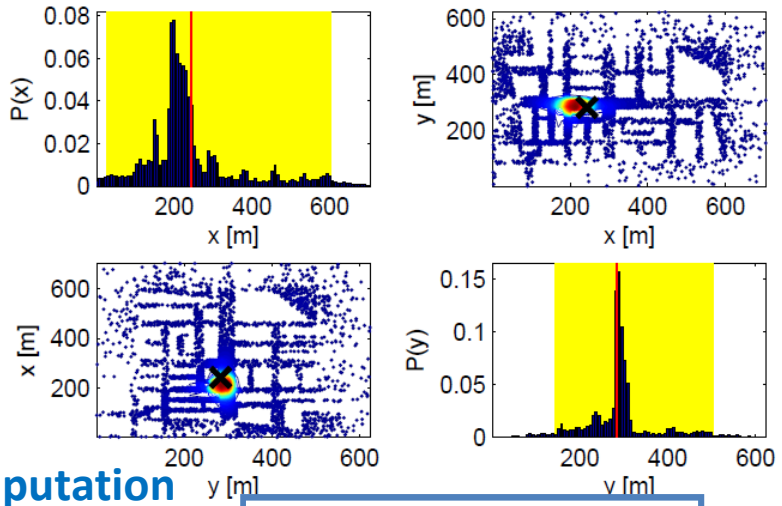
Elements of the contaminant source localization system

- Contaminant data at sensors location
- Parameters space scanning algorithms (MCMC, SMC, ABC, GA etc.)
- Dispersion model –run as a „forward” model
  - open areas - most common is Gaussian plume model
  - urbanized areas - CFD, QUIC (Los Alamos)
    - **PROBLEM- computationally expensive, single run at least 3 minutes in small domain, usually required minimum 50000 runs =>minimum 100 days on single machine**
    - **Thus not possible to apply in the real-time contaminant localization emergency system**
  - IDEA - train the ANN neural network to predict the contaminant dispersion on given terrain
    - difficult and long training but after ANN has a very quick response time

Localization System

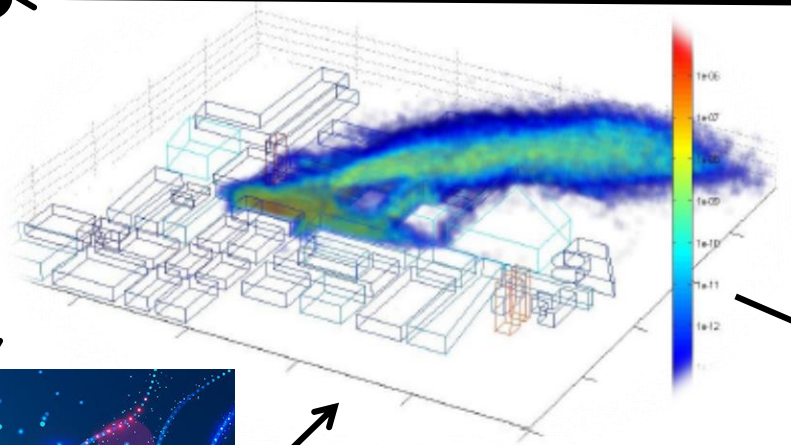
A priori: $\pi(\theta)$

$\theta - parameters\ set$



Approximate Bayesian Computation

$t = t + 1$

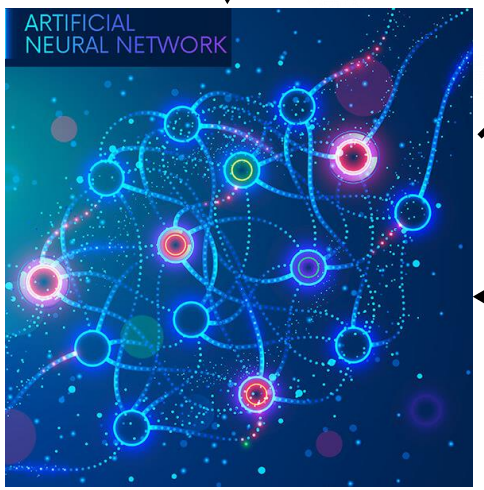


A Posteriori

$\pi(\theta|d_{obs}^{1:t})$

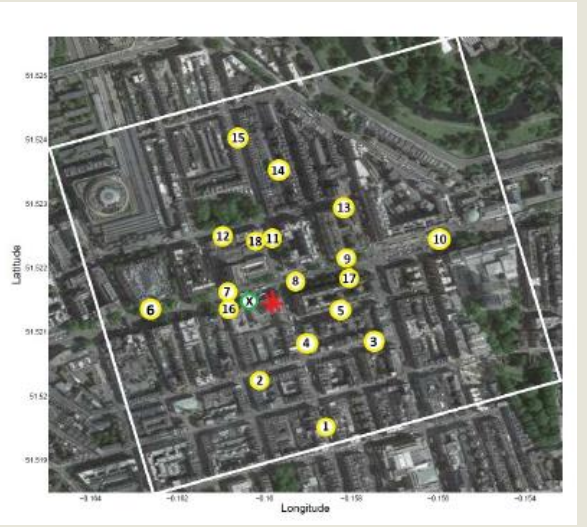
Distance Measure
Fractional Bias

$d^{1:t}$ $d_{obs}^{1:t}$



I

Event



Sensor Network

S1 S2 S3 S4

e_{obs} S5 ... S18

Meteorological station

WCC

ANN Model, City domain- DAPPLE site

- We decided to check the possibility to train the ANN to simulate the airborne toxin transport in the area of central London where the DAPPLE experiment was conducted [www.dapple.org.uk].
- Even though the DAPPLE field experiment was quite extensive the data available from its Trials are very limited from a point of view of the ANN training requirement.
- From four Trials run in March and Jun 2007, we have concentrations at 15 receptor positions for 30 minutes with 3-minutes intervals. This gives us, in sum, about **600 point-concentrations** for four various source positions and release rates. This number of data is **not enough** to properly train the ANN.
- **The only solution is to use the verified and well-recognized dispersion model to generate the data-set utilized to train, test, and validate the ANN**

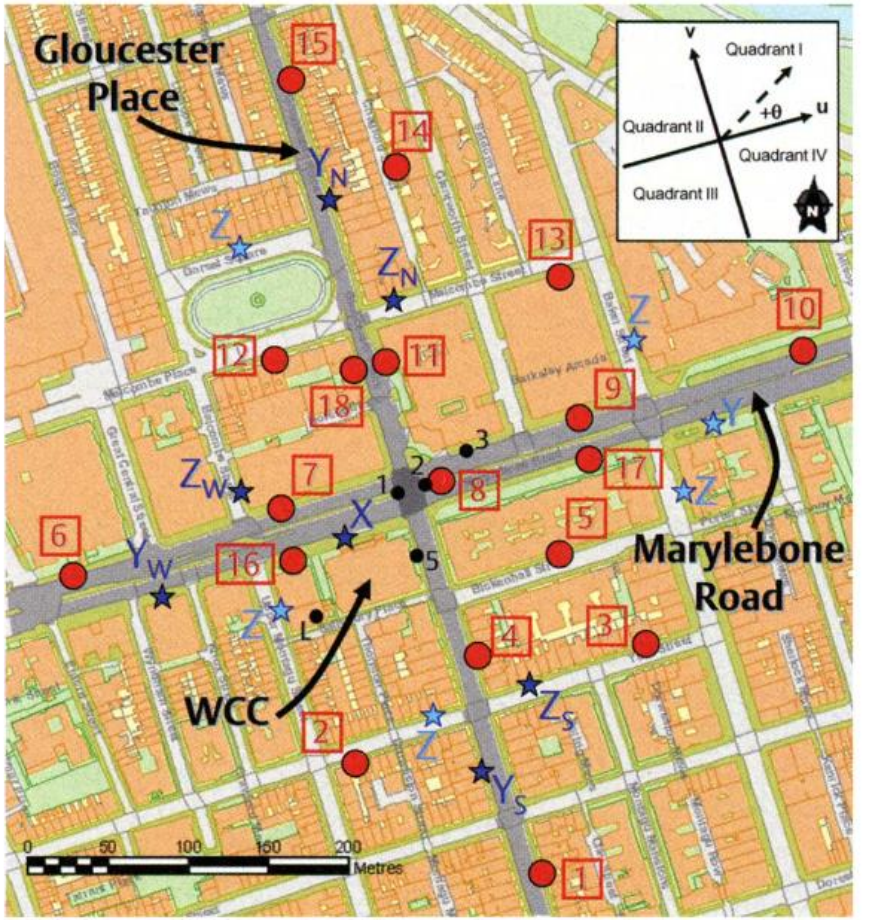
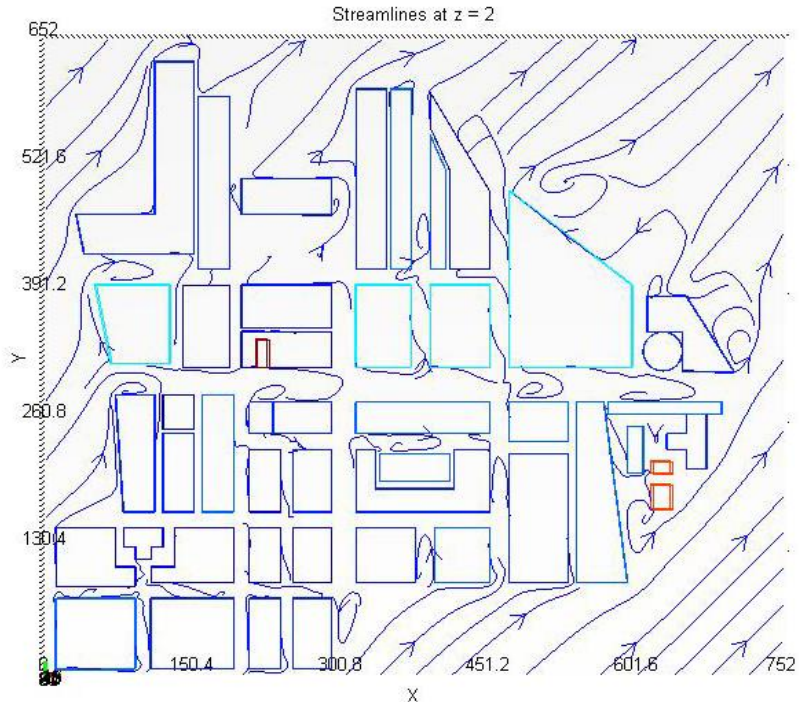
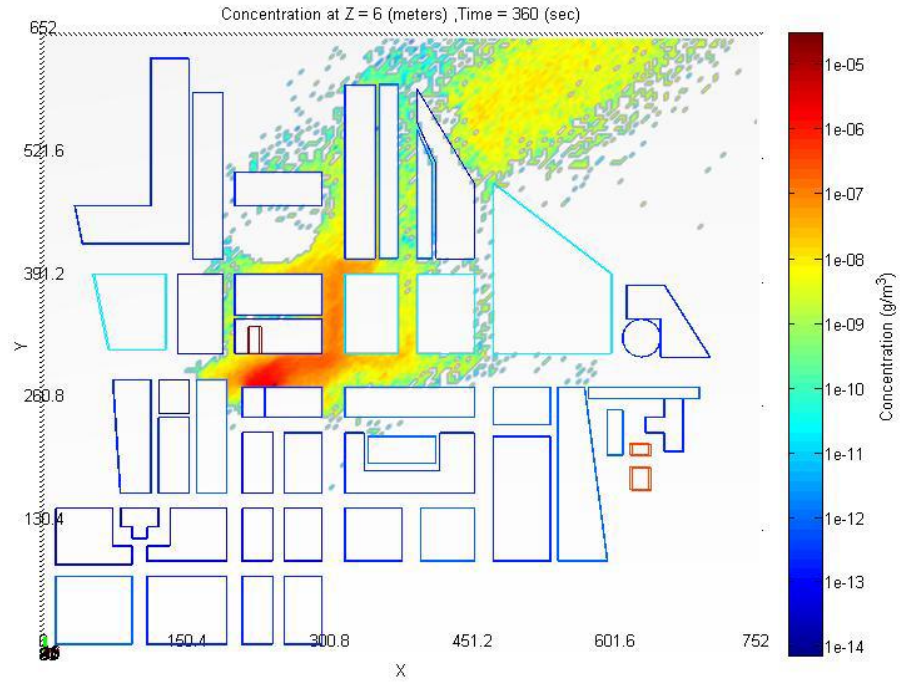


FIG. 1. The map shows the DAPPLE area of central London, and is centered at the focal intersection, that of Marylebone Road and Gloucester Place (at 51.5218°N, 0.1597°W). Also shown are the locations used in the summer 2007 campaign.

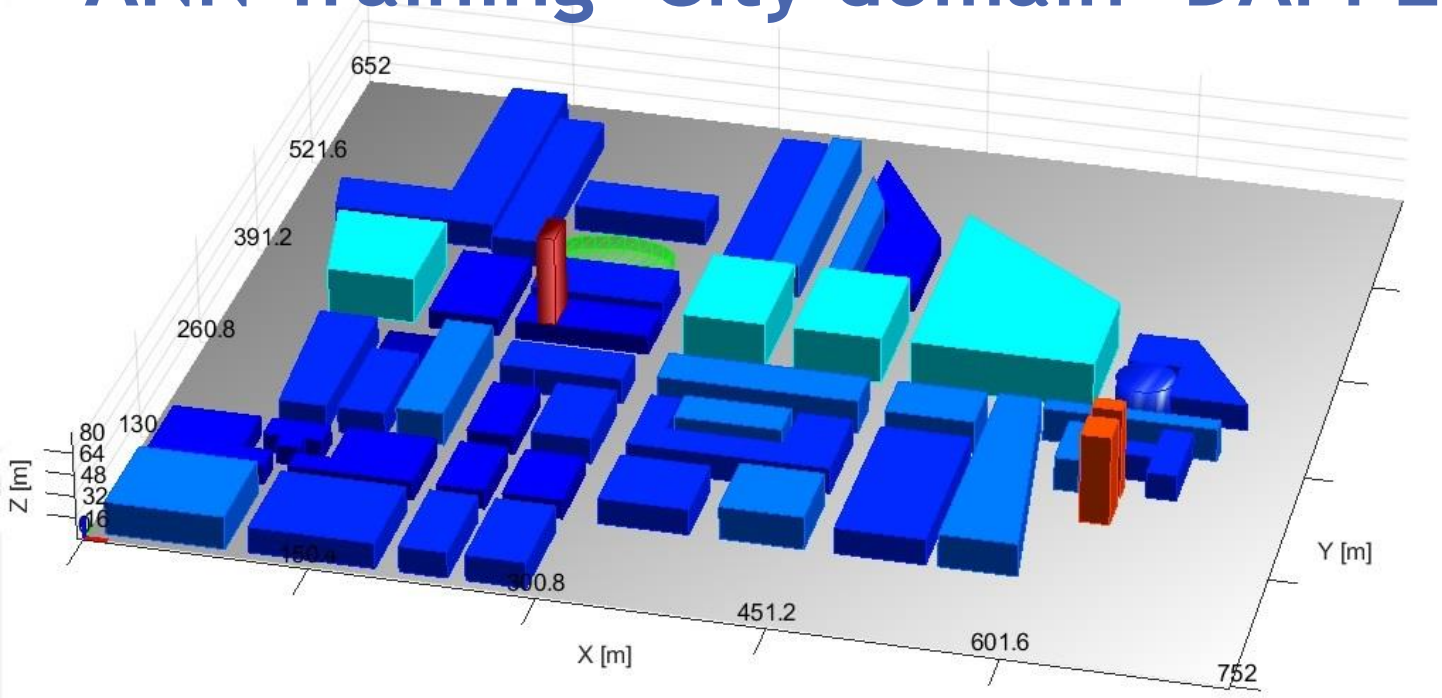
Quick Urban Industrial Complex (QUIC) Dispersion Modeling System, Los Alamos National Laboratory

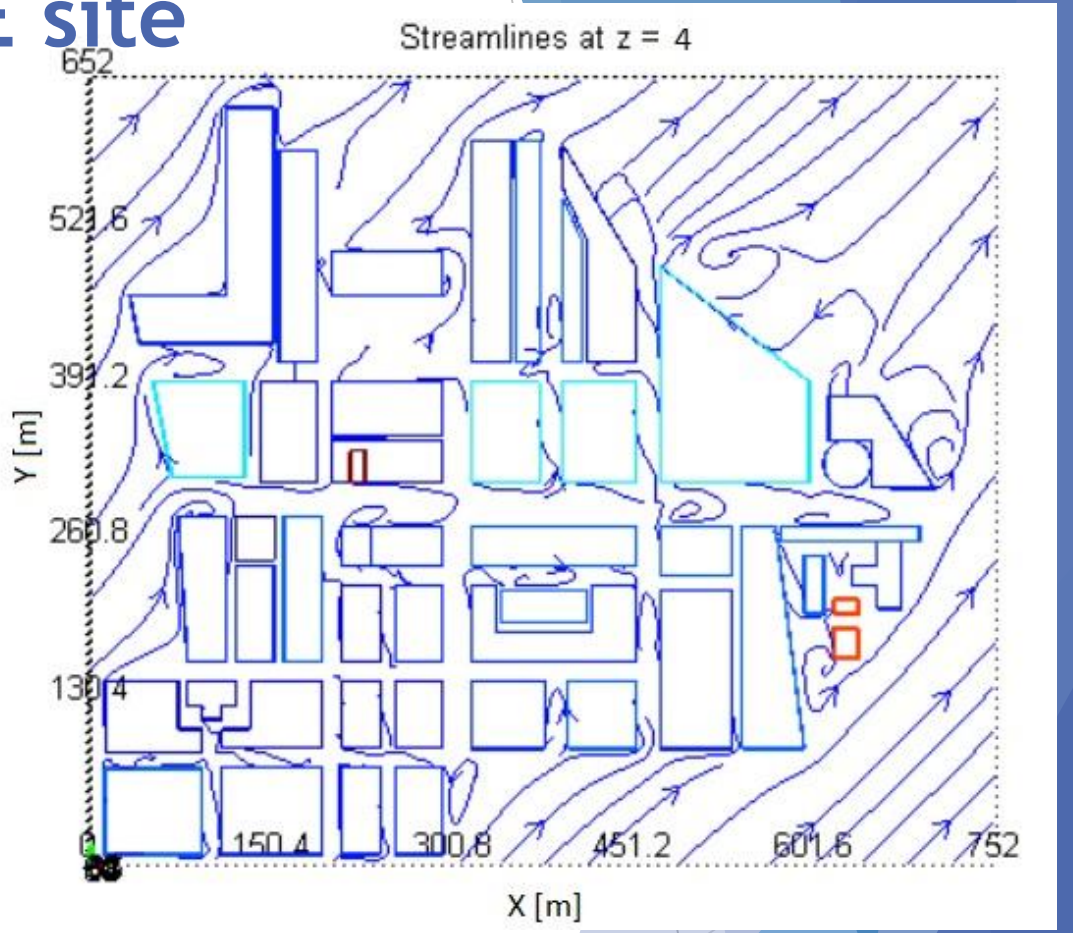
- QUIC-URB uses a 3D mass-consistent wind model to combine adequately resolved time-averaged wind fields around buildings
- QUIC-PLUME is a Lagrangian particle model which describes gas dispersion by simulating the release of particles and moving them with an instantaneous wind composed of mean and turbulent components [M. D. Williams, M. J. Brown, B. Singh, D. Boswell, Quic-plume theory guide, LANL (2004) 43]





ANN Training -City domain- DAPPLE site

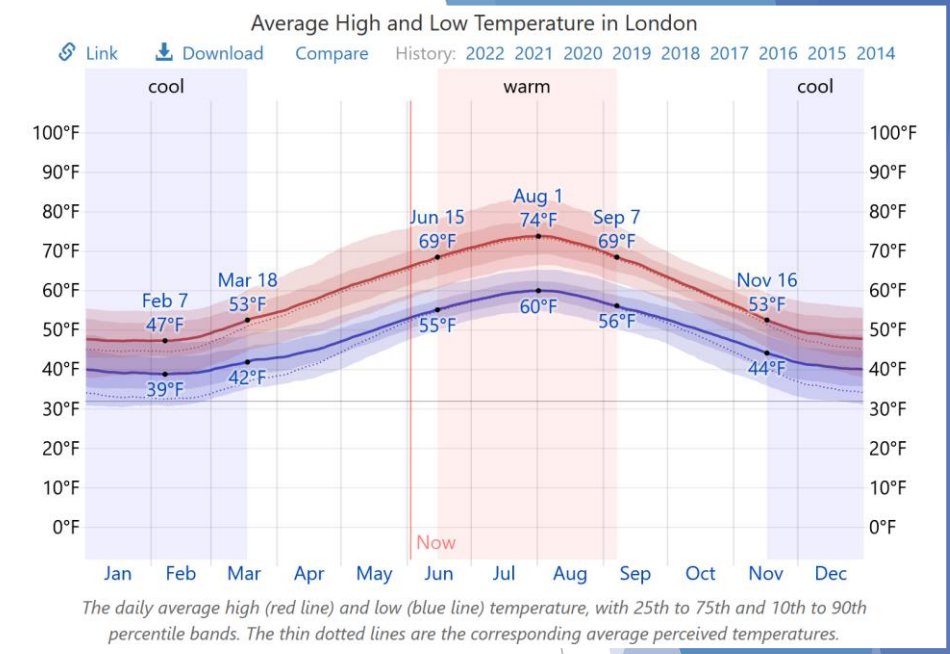




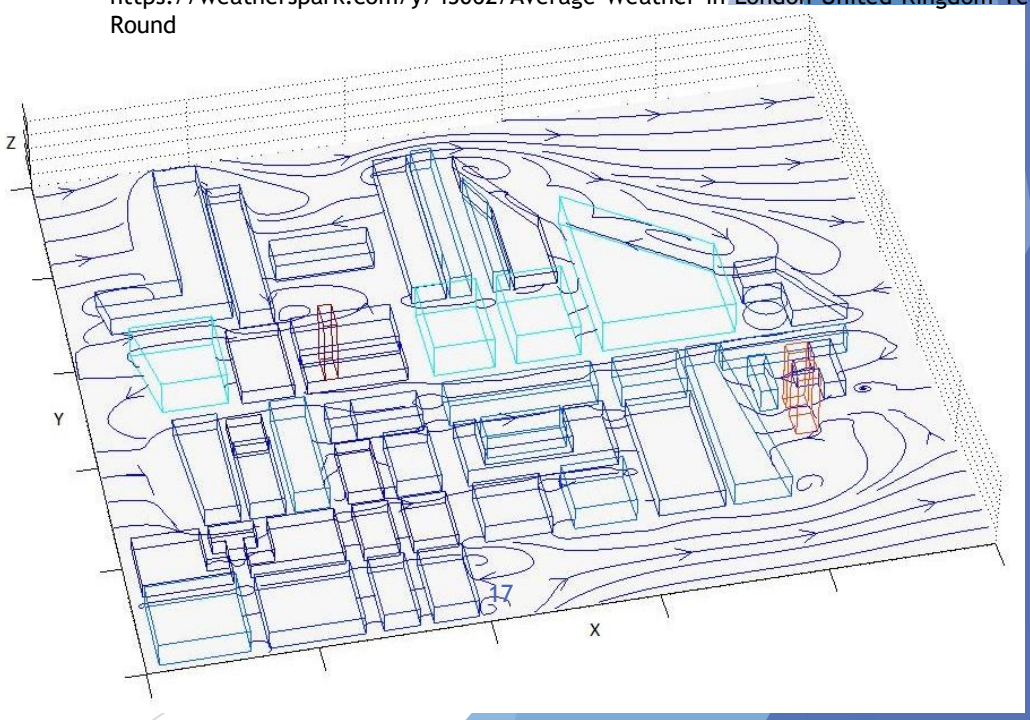
- We have prepared the domain of size 752 m × 652 m × 80 m in which we have placed representations of the original buildings.
- The average building height in the area is 21.6m (range 10 to 64m).
- The whole considered domain and the estimated by the QUIC-URB sample wind field around the buildings are presented in above Figs.

ANN Model, City domain- DAPPLE site

- The meteorological data were included accordingly to the statistics
observed for the summer season in London during the years 2014-2021.
- Consequently, the average temperature in the selected season of the year was ~18 °C, and the average wind speed was 5m/s.
- Additionally, the percentage share
of each of the eight main wind directions was taken into account (N - 10%, S - 10%, W - 20%, E - 10%, NW - 10%, NE - 11%, SE - 4%, SW - 25%).



https://weatherspark.com/y/45062/Average-Weather-in-London-United-Kingdom-Year-Round



ANN Model, City domain- DAPPLE site

- We have set the simulations of an ideal gas continuous release and registered its concentration for thirty minutes using the QUIC-PLUME
- To reflect the real measurement conditions we have randomly drawn the 600 contamination source locations, release rate within the interval Q $\in$< 100–999 > mg and its duration within interval < 2min; 30min > and 19616 registration points (representing the sensor locations) per single release.
- ~445 simulations, resulted in training dataset of size about 5×10^7 vectors

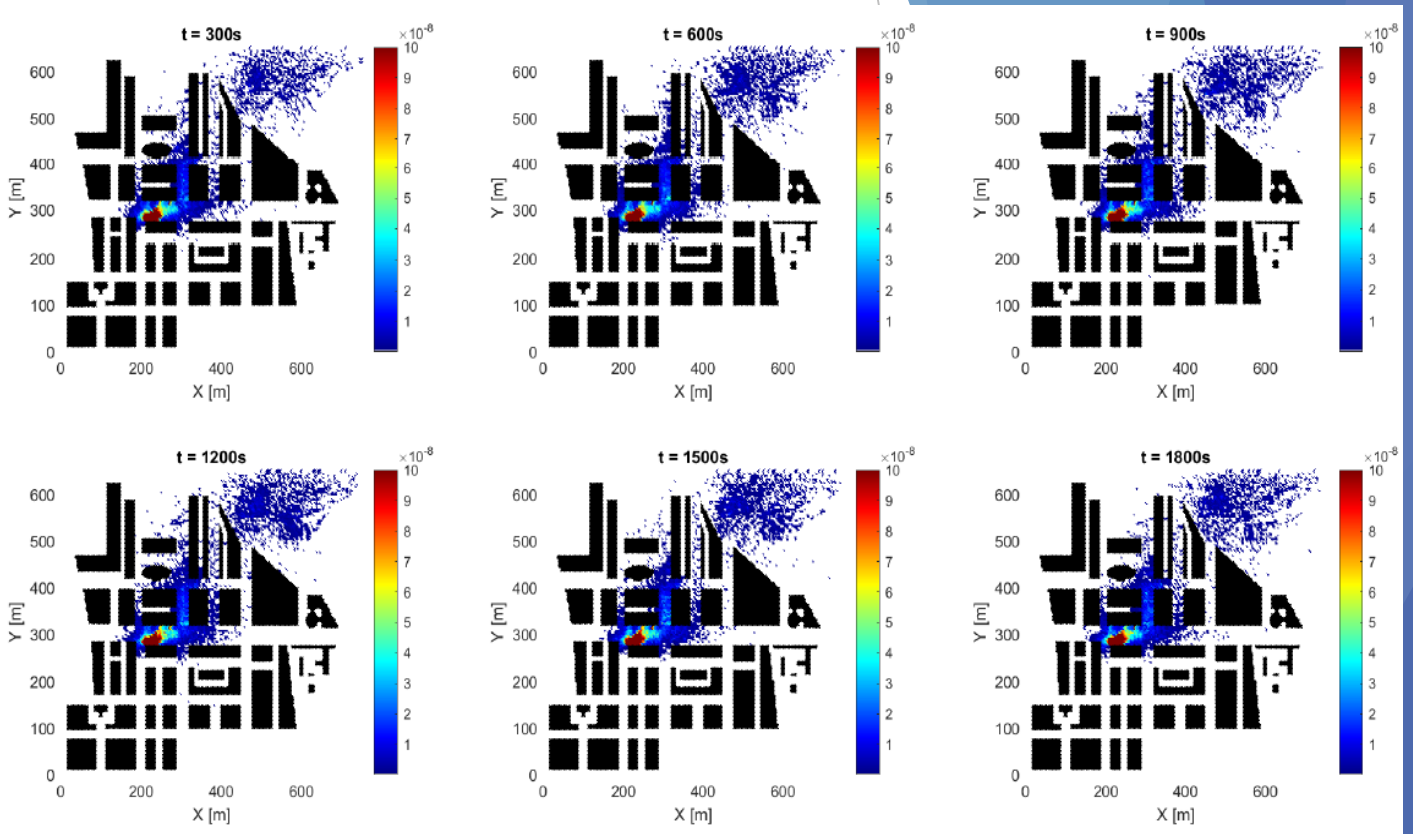
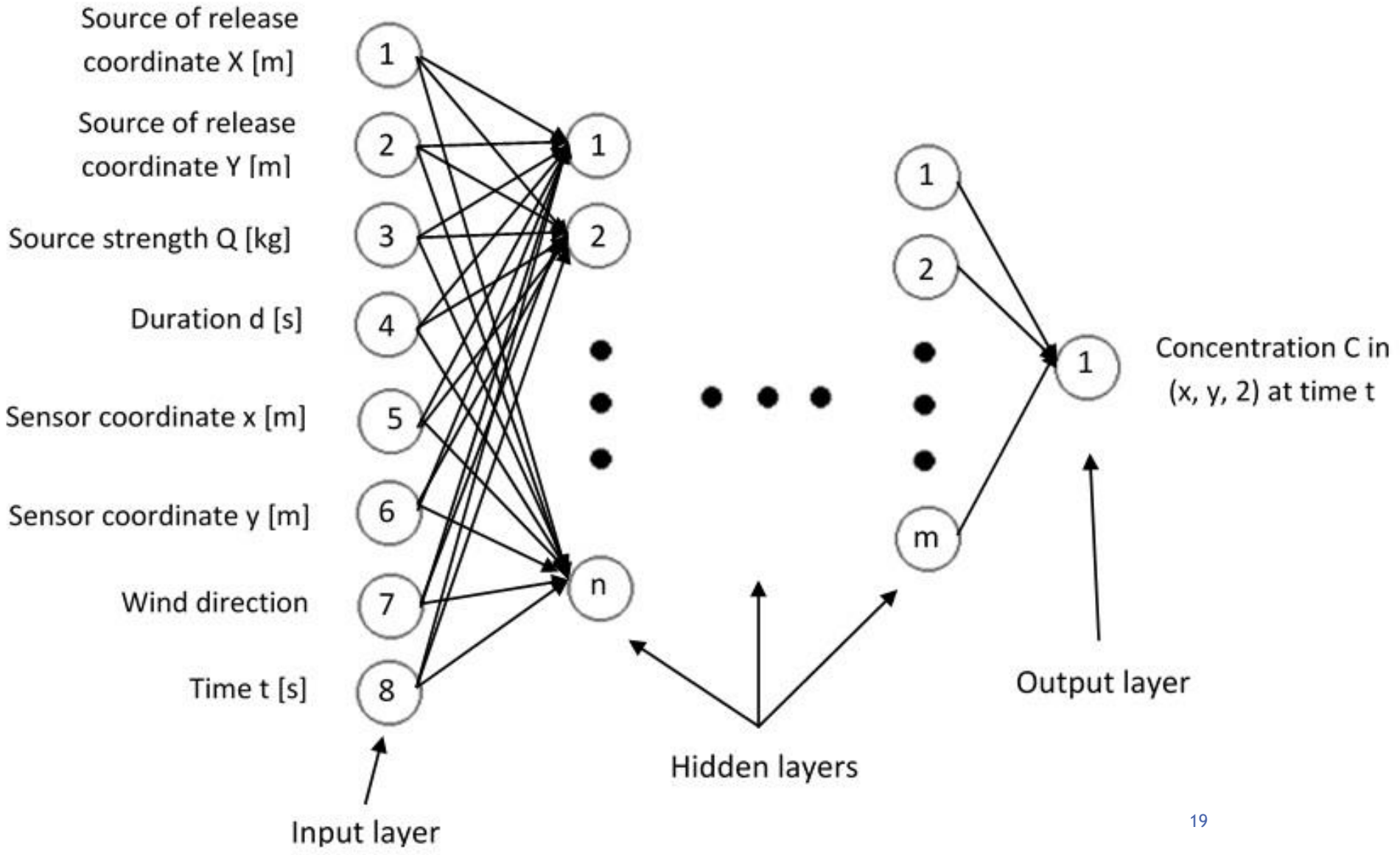


Fig. The normalized concentration of the gas during thirty minutes after the 15 minute release of gas from the source located at x = 300m, y = 240m, z = 7m within the considered domain as simulated by the QUIC model.

ANN topology



ANN Training Results, City domain- DAPPLE site

- Taking into account the complexity of the transport of the airborne contaminant in the turbulent wind around the buildings, the quality of the trained ANN is quite good.
- These R-value values indicate a signicant relationship between the outputs and targets.
- The regression lines show that ANN slightly underestimates the higher concentrations

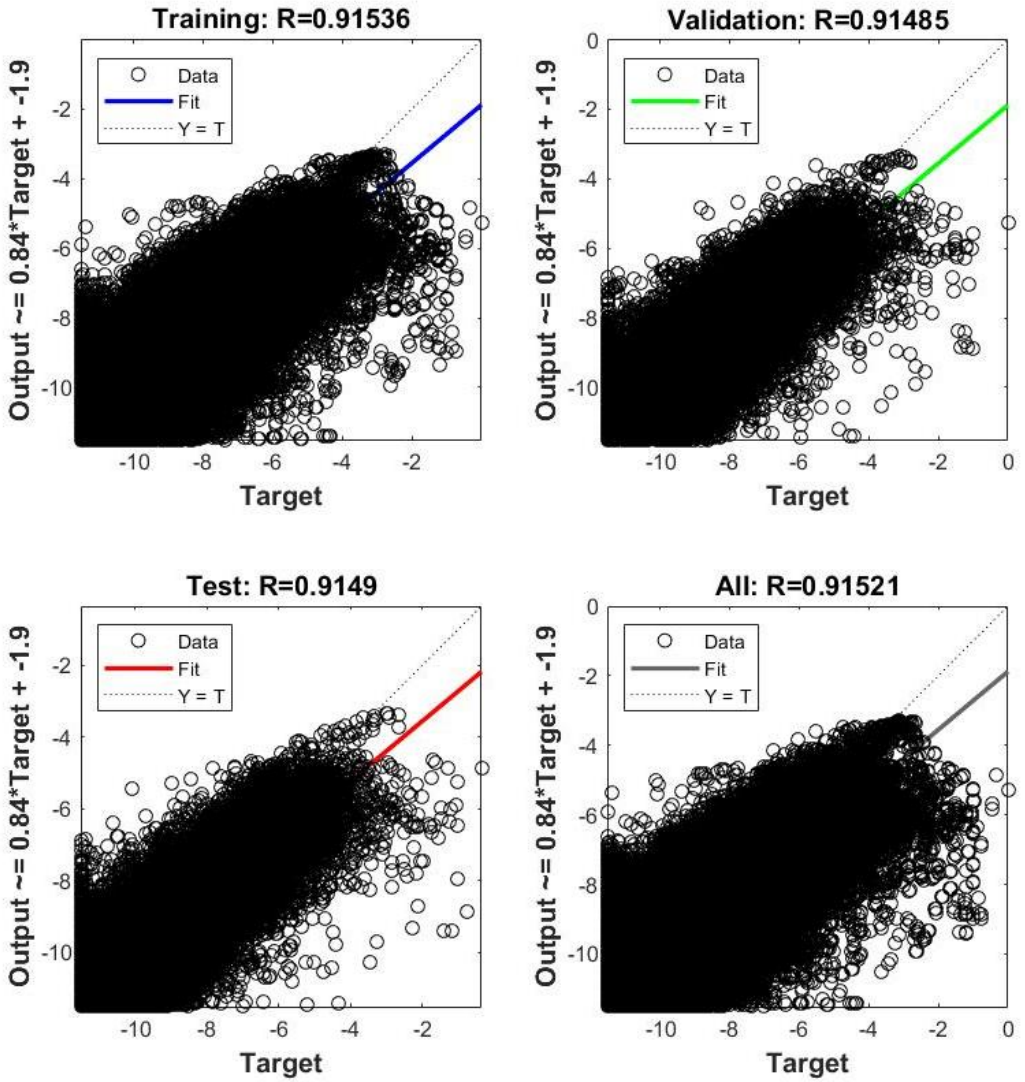


Fig. The scatter plots representing results of training, testing, and validation process. The dashed line represents the ideal fit.

Importance of the noise in the training dataset

$$MSSDLE_{\vec{x}}(t=j) = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N_L}\sum_{L=1}^{N_L}\left[\frac{1}{N_K-1}\sum_{K=2}^{N_K}\left(\frac{\ln\left(\frac{(C_j^{m(K,L)}-C_j^{m(K-1,L)})}{\Delta K}\right)-\ln\left(\frac{(\hat{C}_j^{m(K,L)}-\hat{C}_j^{m(K-1,L)})}{\Delta K}\right)}{\ln\left(\frac{(C_j^{m(K,L)}-C_j^{m(K-1,L)})}{\Delta K}\right)+\ln\left(\frac{(\hat{C}_j^{m(K,L)}-\hat{C}_j^{m(K-1,L)})}{\Delta K}\right)}\right)^2\right]\right] \quad (3)$$

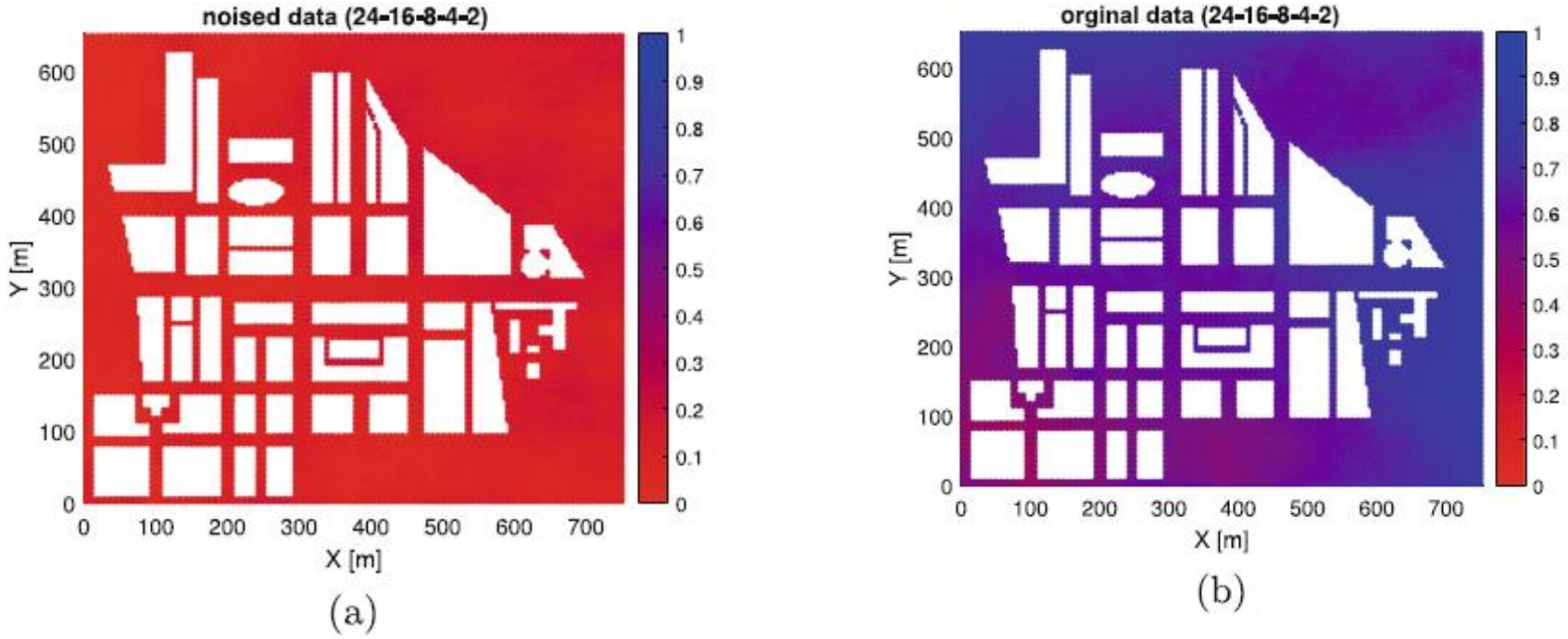
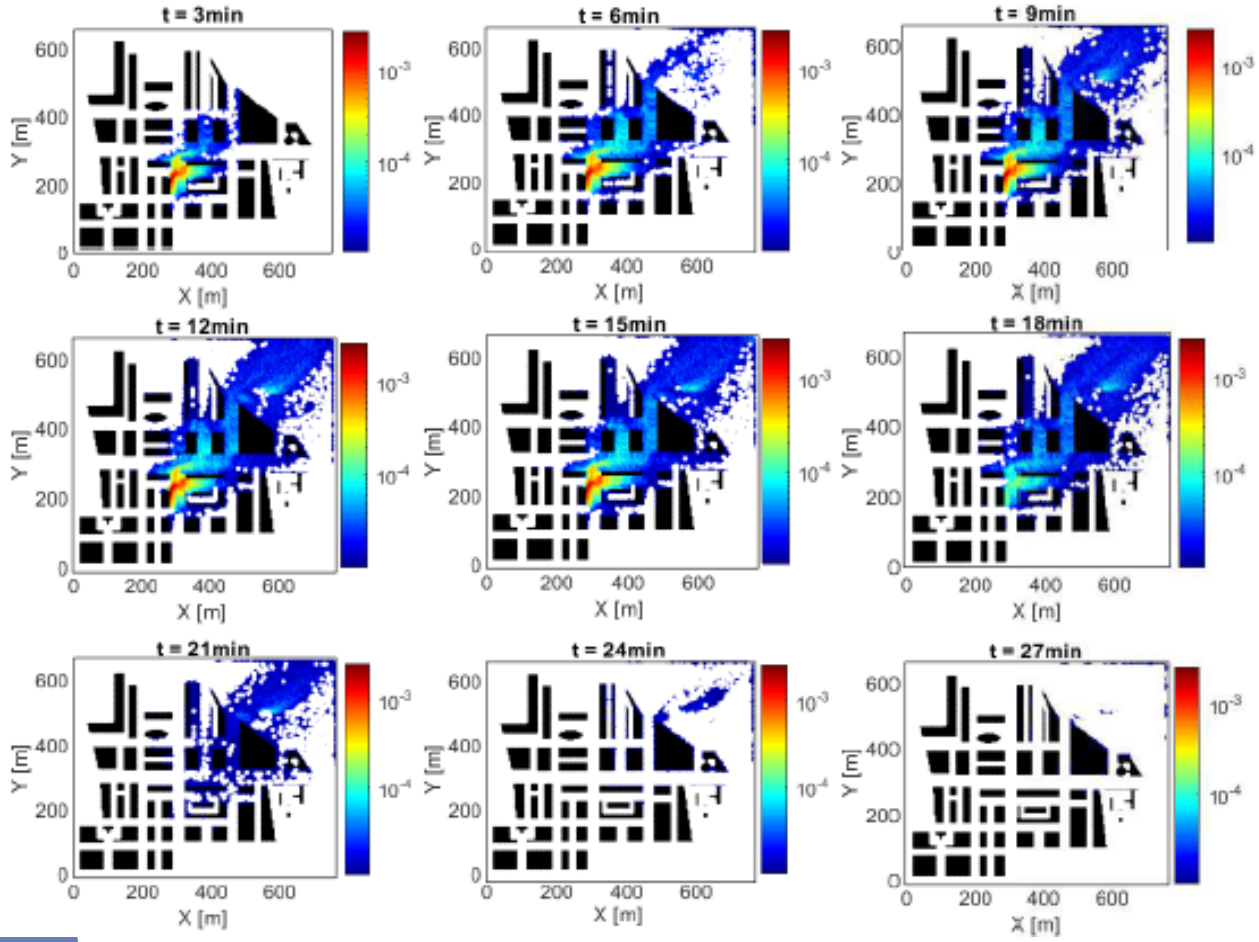


Fig. 2. The measure $MSSDLE$ (Eq. 2) distribution in the 2D city domain for the ANN with hidden layers $24-16-8-4-2$ trained on (a) noised and (b) original data.

Does the ANN learned the physics standing behind the gas dispersion over the highly urbanized area?



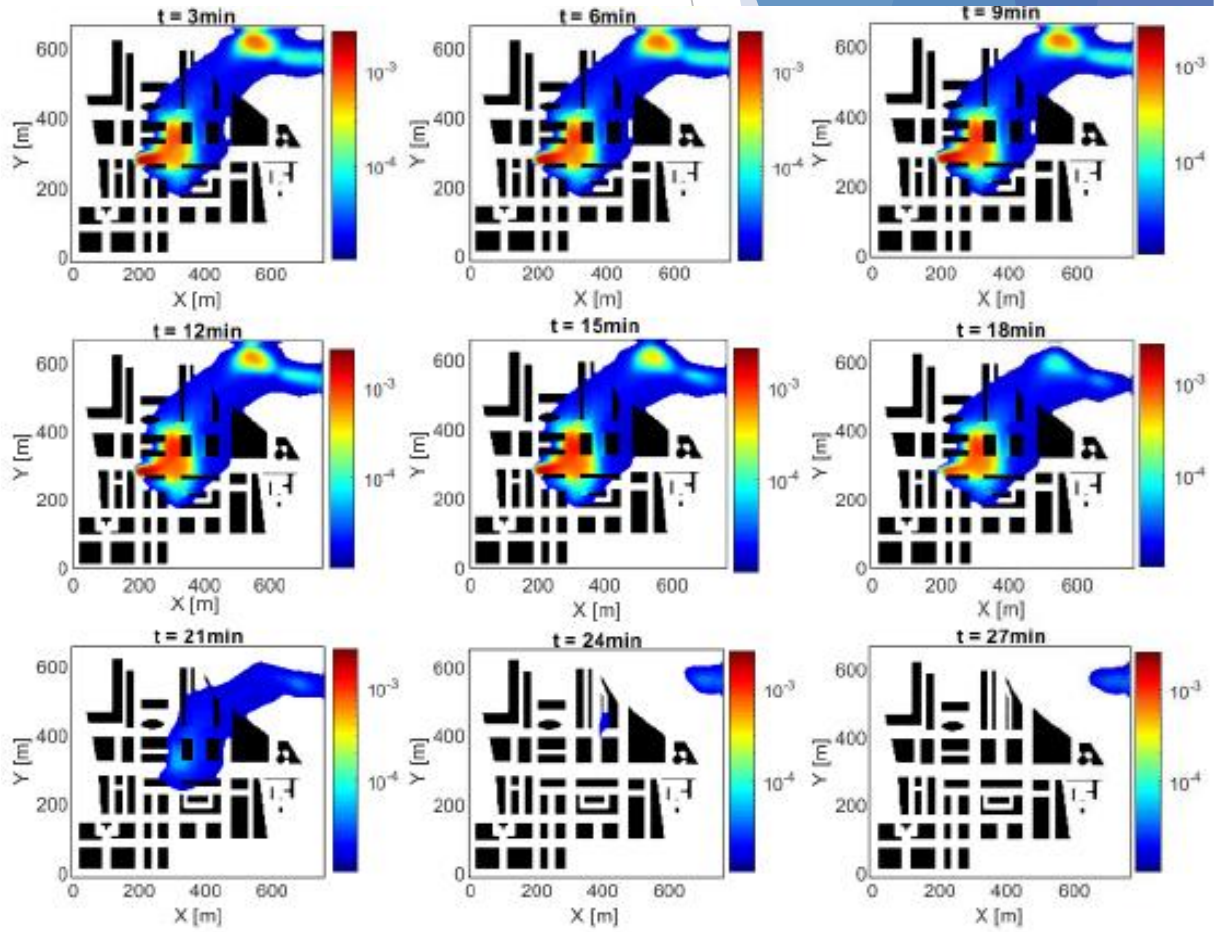
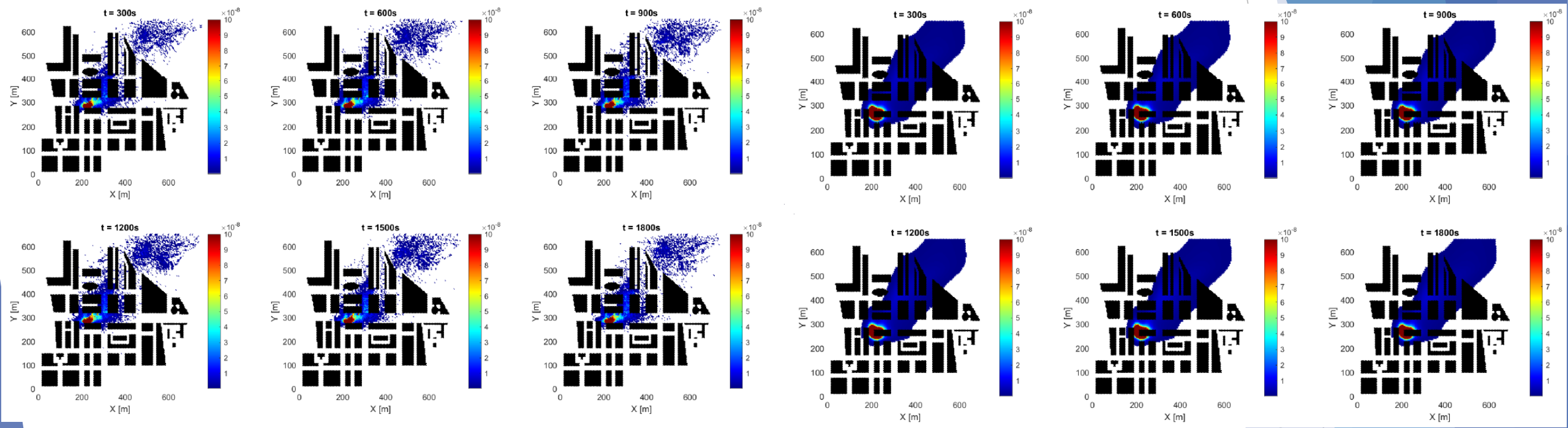


Fig. The dispersion of the contaminant simulated by the QUIC (left) vs. ANN (right) , Scenario 1

Does the ANN learned the physics standing behind the gas dispersion over the highly urbanized area?



QUIC

(~300 seconds for 1 point concentration estimation)

Fig. The dispersion of the contaminant simulated by QUIC (left) vs. ANN (right) , Scenario 2

ANN

(~3 seconds for 1 point concentration estimation)

Does ANN work well in the emergency localization case?

- We have tested our localization system using the DAPPLE experiment.
- The successful localization for that case with use of the Approximate Bayesian Computation (ABC) as the sampling algorithm and QUIC model as a forward model was published in *Kopka, Wawrzynczak, Atmosph. Enviroment, 2018*,
  but ……. the reconstruction time was quite long – not applicable in the real-time working system
- So… we have used the trained ANN in the place of the QUIC dispersion model used up to now in our reconstruction

Localization for the DAPPLE experiment by ANN model

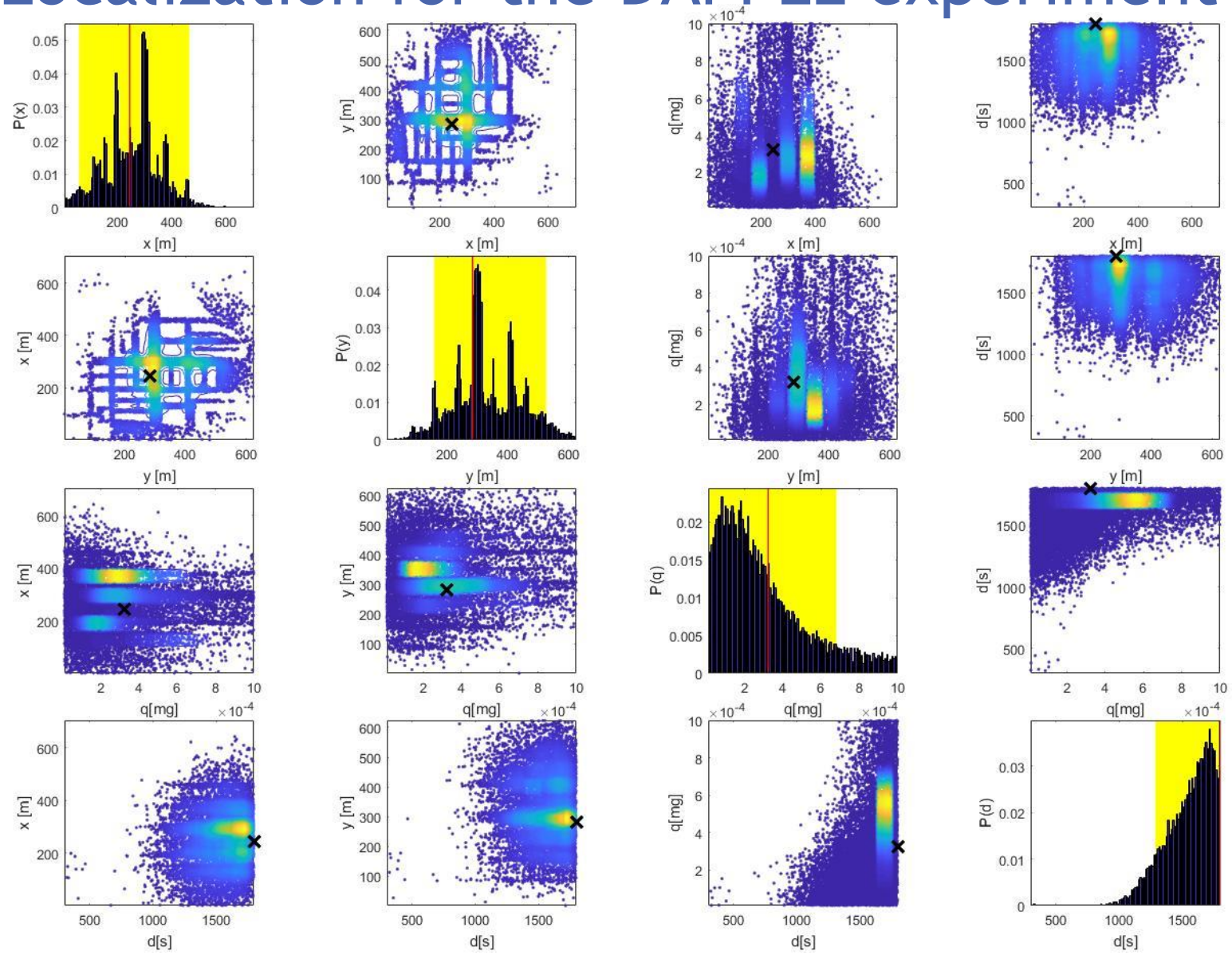


Fig. The bivariate and marginal posterior distributions for searched parameters obtain localization with use of the ANN and ABC. Red vertical line marks the target value.

Conclusions

- The ANN work well as surrogate models for the atmospheric dispersion
  - … when answer time is more important than precision
- Application of the ANN model in the emergency localization system make possible to achieve the real-time response, but…
  - it requires site-specific well trained ANN ->
    - lots of data both from models and field experiments (problematic) in various weather conditions.

Conclusions

Collecting experimental data and making it available in the form of a homogeneous database is invaluable in the AI era.

Thank You