On Machine Learning from Environmental Data

Mikhail Kanevski

IDYST, University of Lausanne, Mikhail.Kanevski@unil.ch

UNIL | Université de Lausanne

Acknowledgements

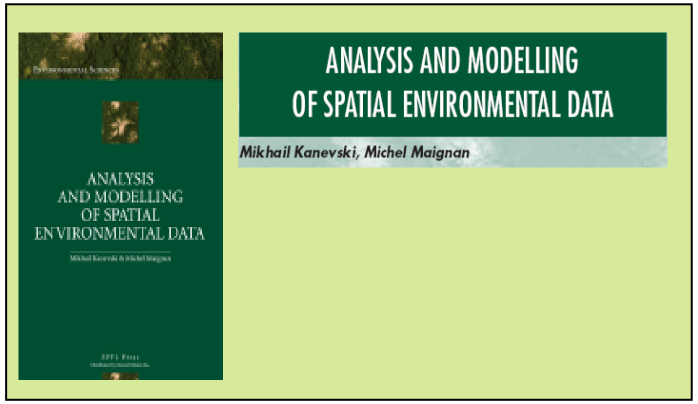
I would like to thank many colleagues, PhD students and friends for very useful and fruitful intellectual collaboration on different scientific topics, bureaucratic questions, project management tasks, etc.

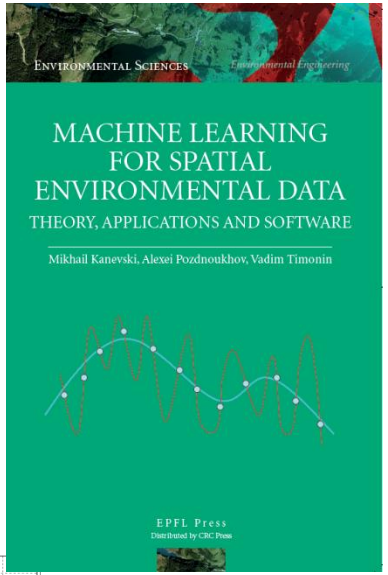
I would like thank University of Lausanne and SNSF for their support of my research during many years

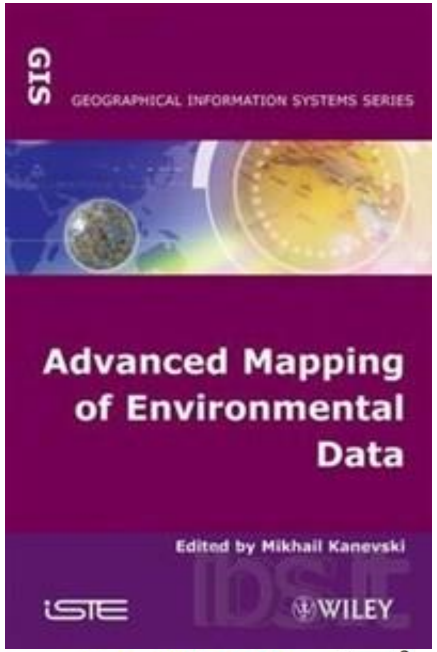
EGU, especially the division of "Earth and Space Science Informatics" : it is a real honor for me to get this award!

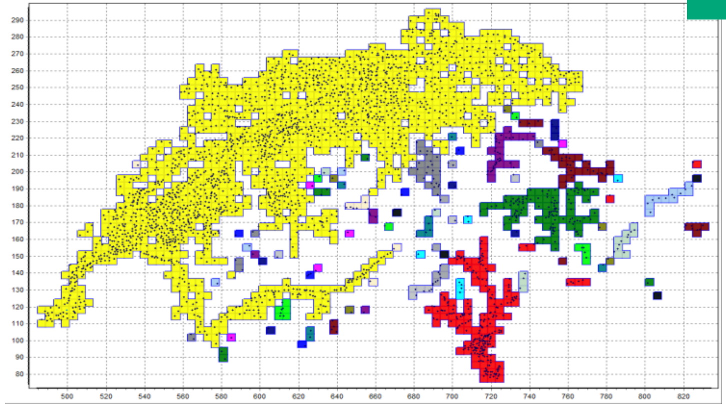
My Research

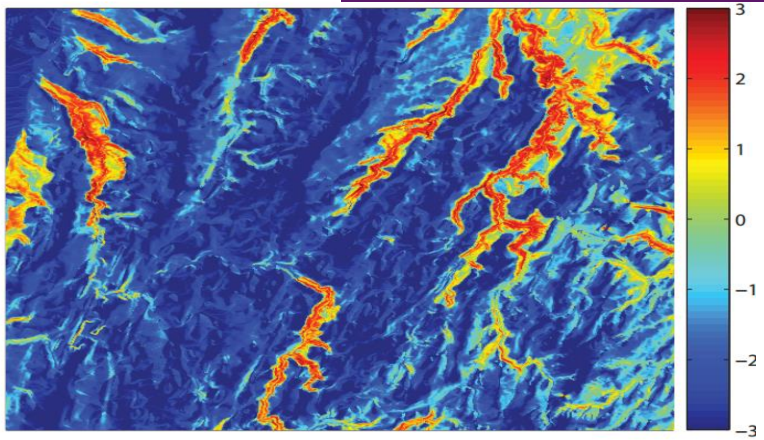
- *Environmental modelling*
- *Geomatics & Geocomputing*
- *Geostatistics*
- *Machine learning* & *Environmental data mining*
- *Time series analysis and forecasting*
- *Space-time point pattern analysis*
- *Applications: natural hazards (forest fires, landslides, avalanches), environmental risks (air, water and soil pollution), renewable energy resources, socio-economic and financial data, geodemography,…*

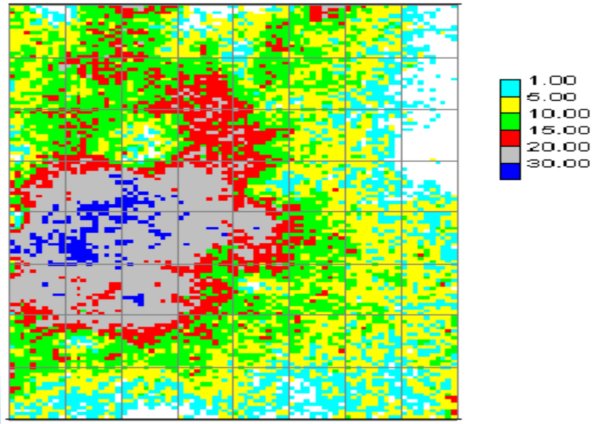








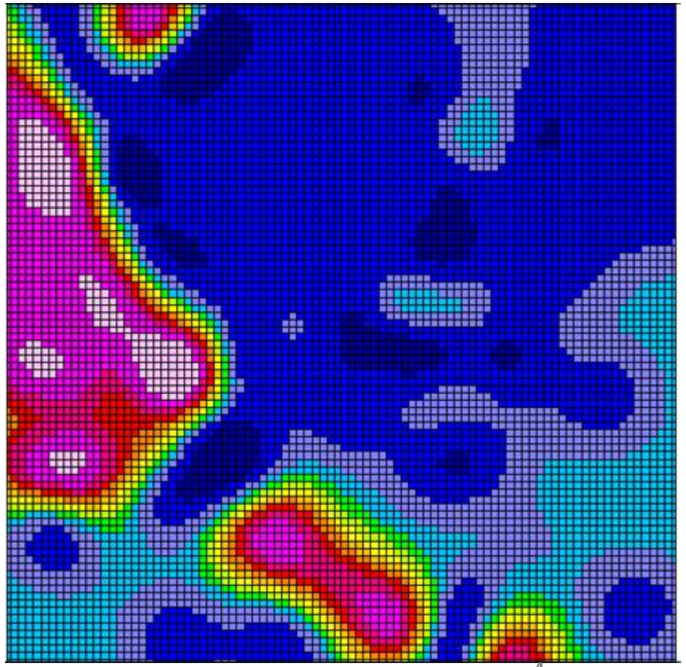




Environmental data: challenges

- wide variety of data
- small, medium and big data
- multi-scale
- multivariate
- uncertain
- nonhomogeneous
- high dimensional
- nonlinear
- complex
- data and science-based models

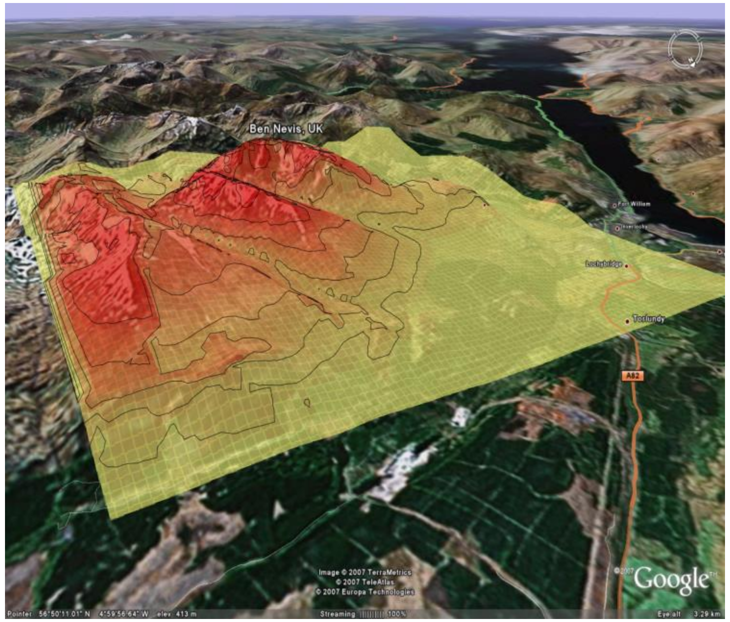
  ….
- Ill-posed (ill-defined) problems

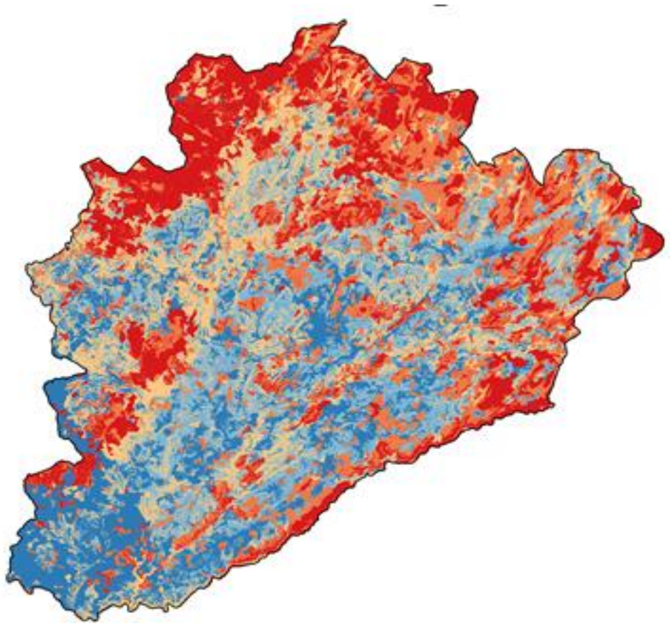


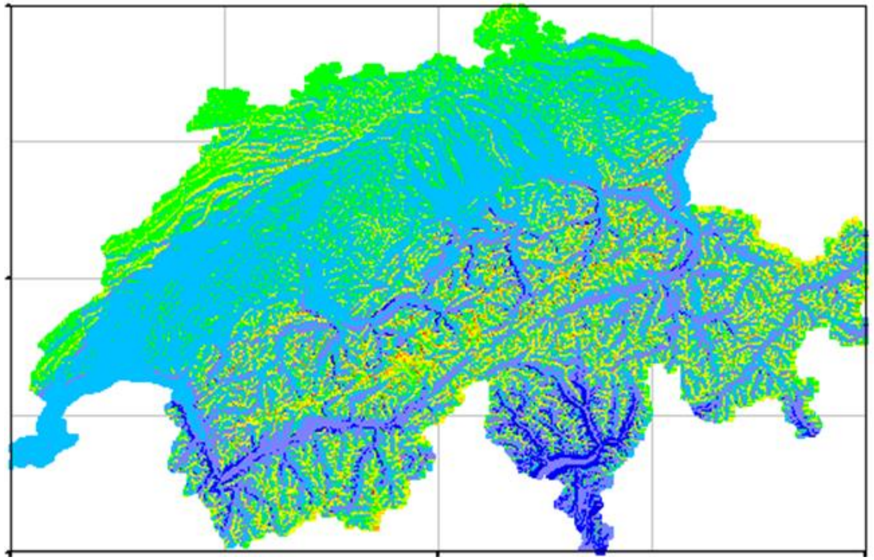
Case studies and Dimensionality

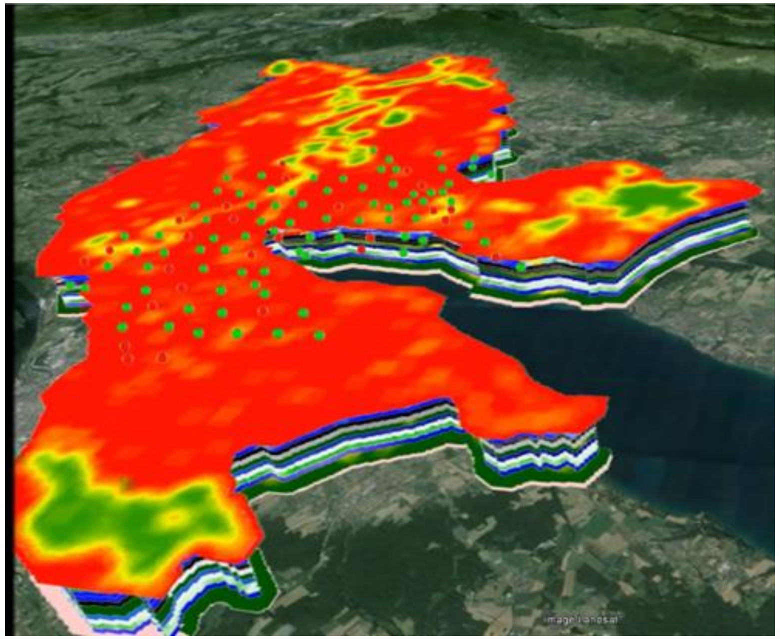
Wind fields >13d
Avalanches > 40d
Landslides >18d
Permafrost - >20d
City pollution >50
remote sensing >100
Wildfires > 25
Swiss population distribution > 5

...









Wildfire (Source: foresttech)

Air pollution in London. (Photo: Mike Hewitt/Getty Images)

Avalanche (Source: National Geographic)

A landslide in the Cusco region of Peru destroyed more than 100 houses in March 2018. (Wikipedia)

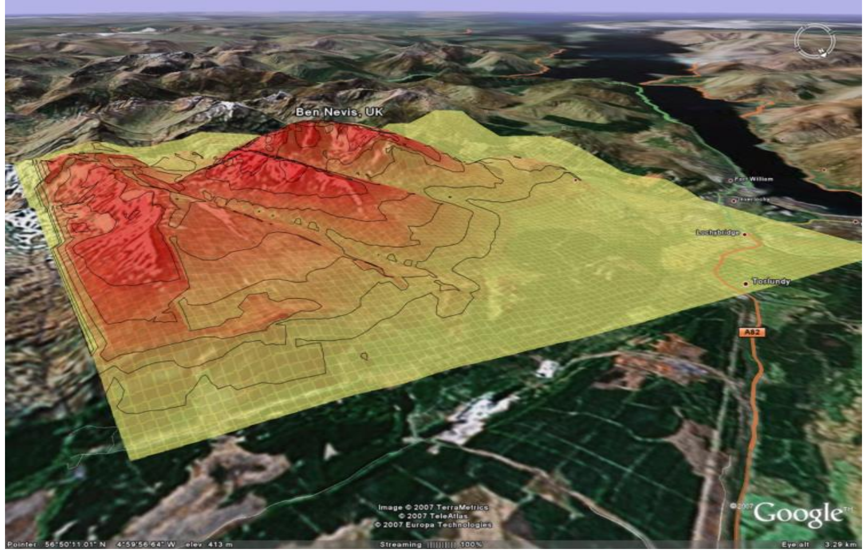
“I have a problem…”

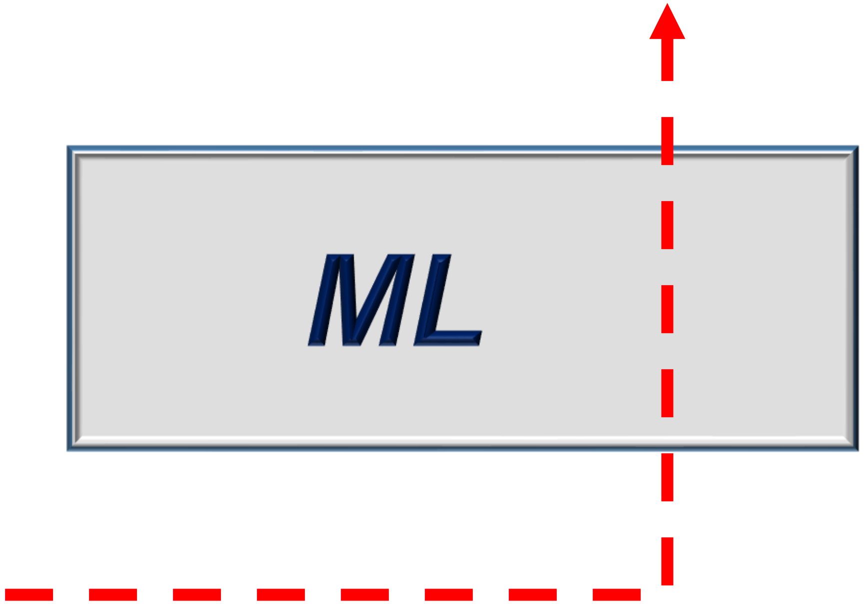
https://www.tutorialspoint.com/renewable_energy/wind_energy_introduction.htm

ML Modelling:

From original geospace to feature space

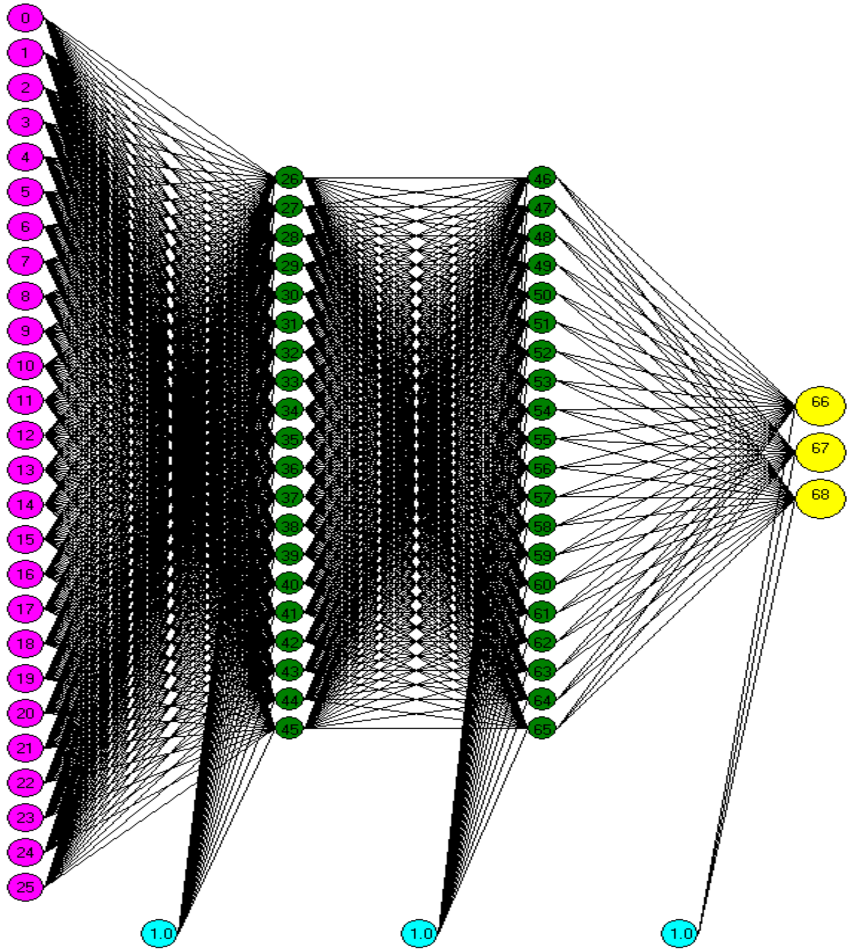
(problem formulation and data quality and quantity!)





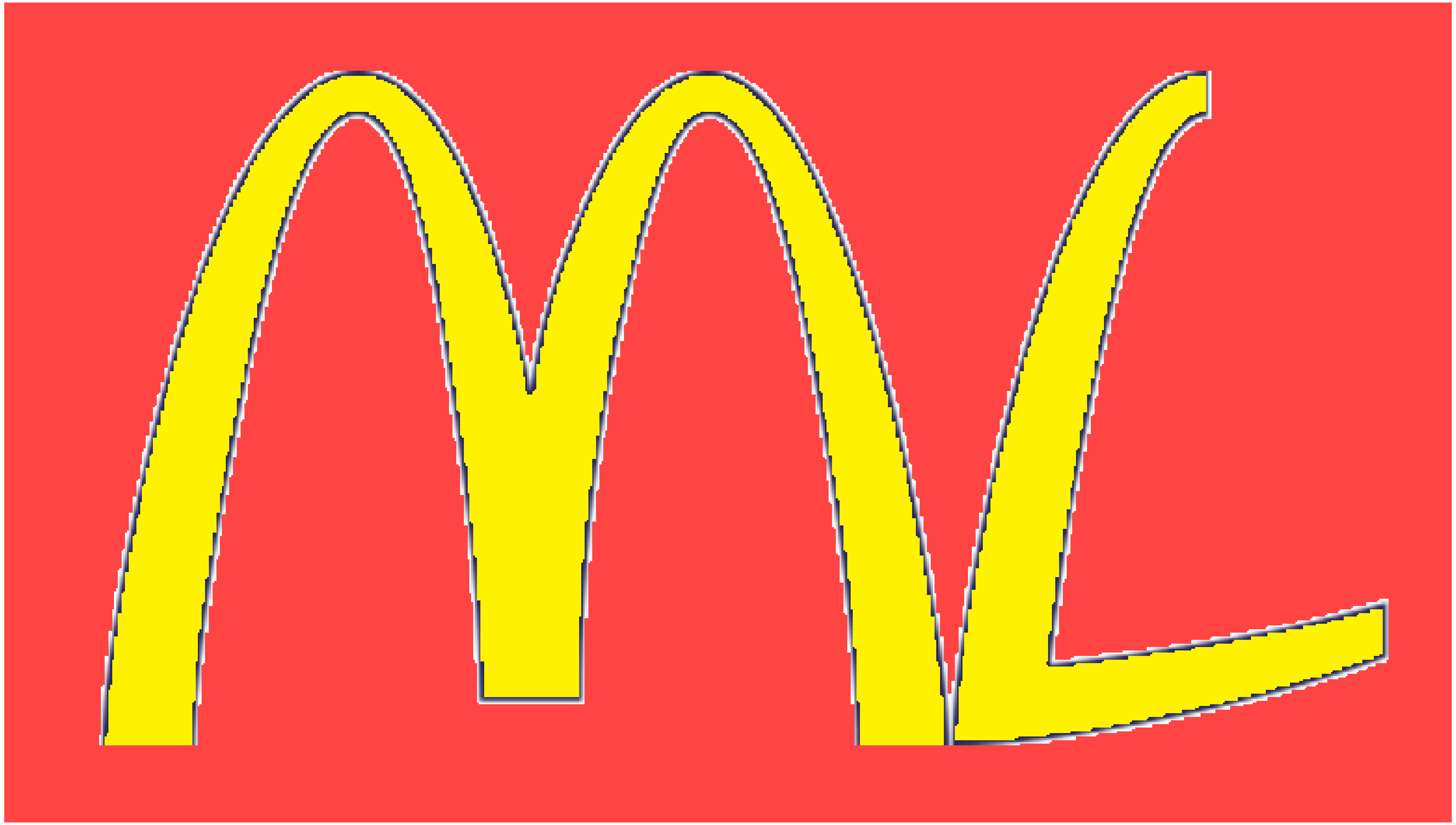
Why Machine Learning

- Universal
- Nonlinear
- Robust
- Data adapted, data driven
- Easy data and knowledge integration
- Good for high dimensional spaces
- Good generalization properties
- …



ML = very interesting and useful analysis/modelling/visualization tool!

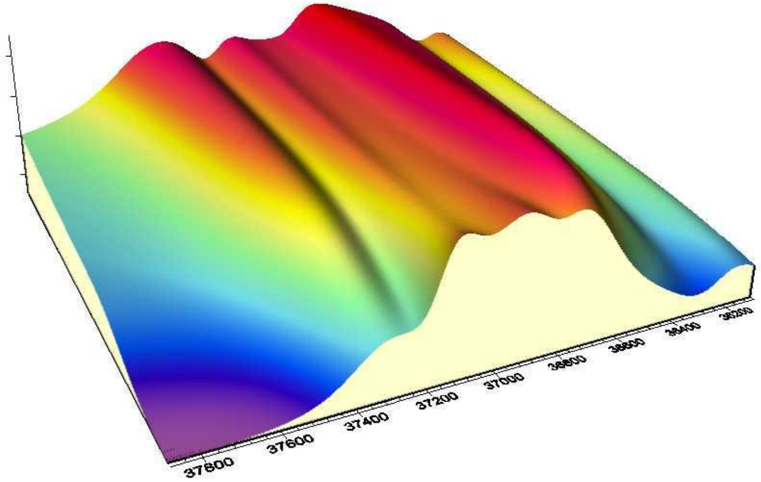
MACHINE LEARNING TODAY



M. Kanevski

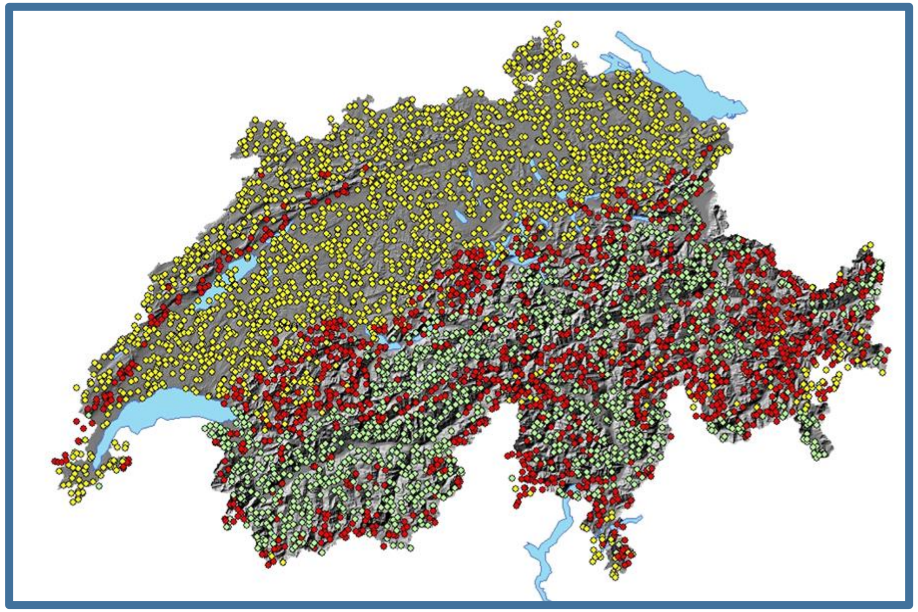
Learning from spatio-temporal data in terms of patterns/structures:

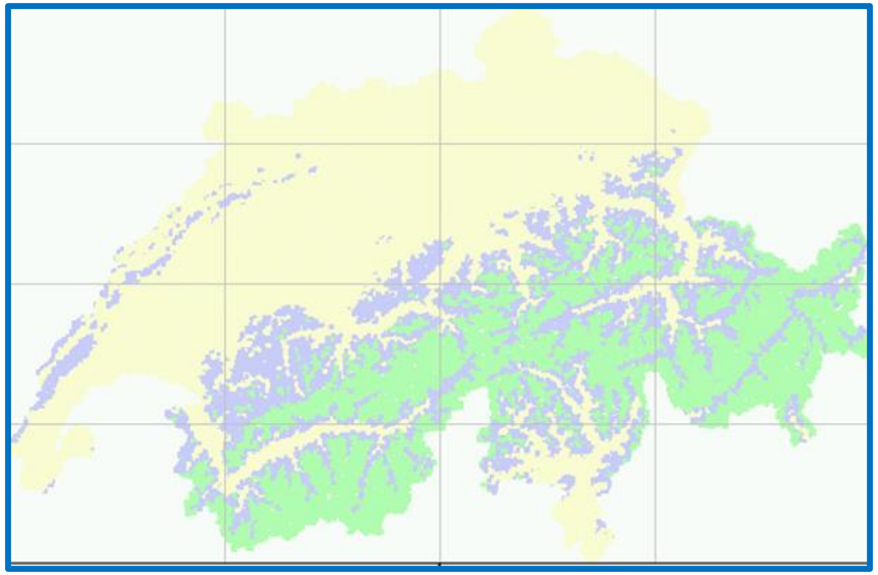
- pattern recognition,
- pattern modelling,
- pattern prediction



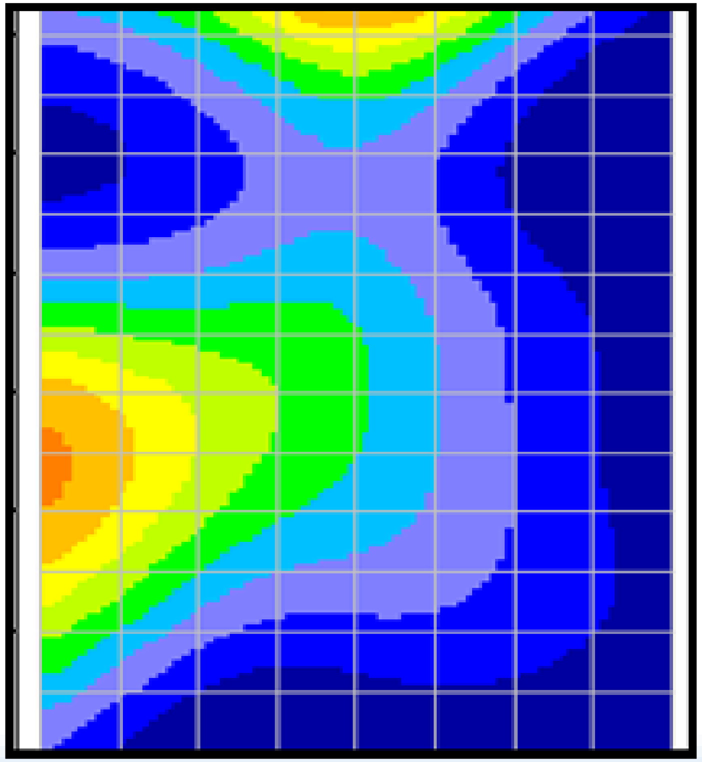
In general, it is three different problems

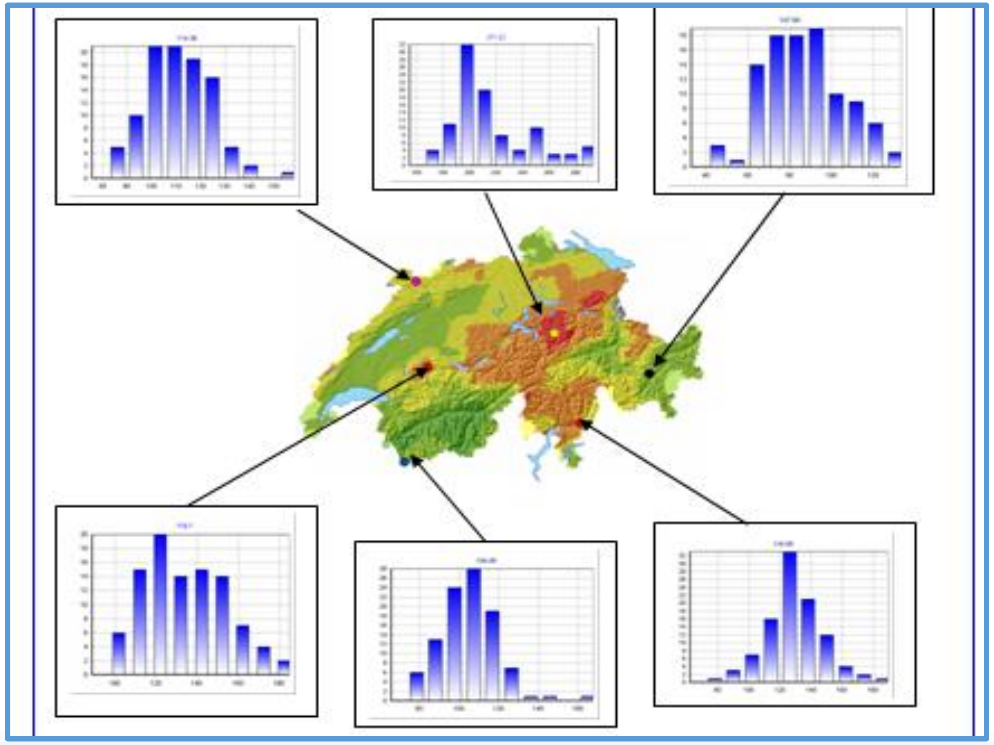
Major fundamental tasks in learning from data





Clustering
Classification
Regression
Pdf modeling

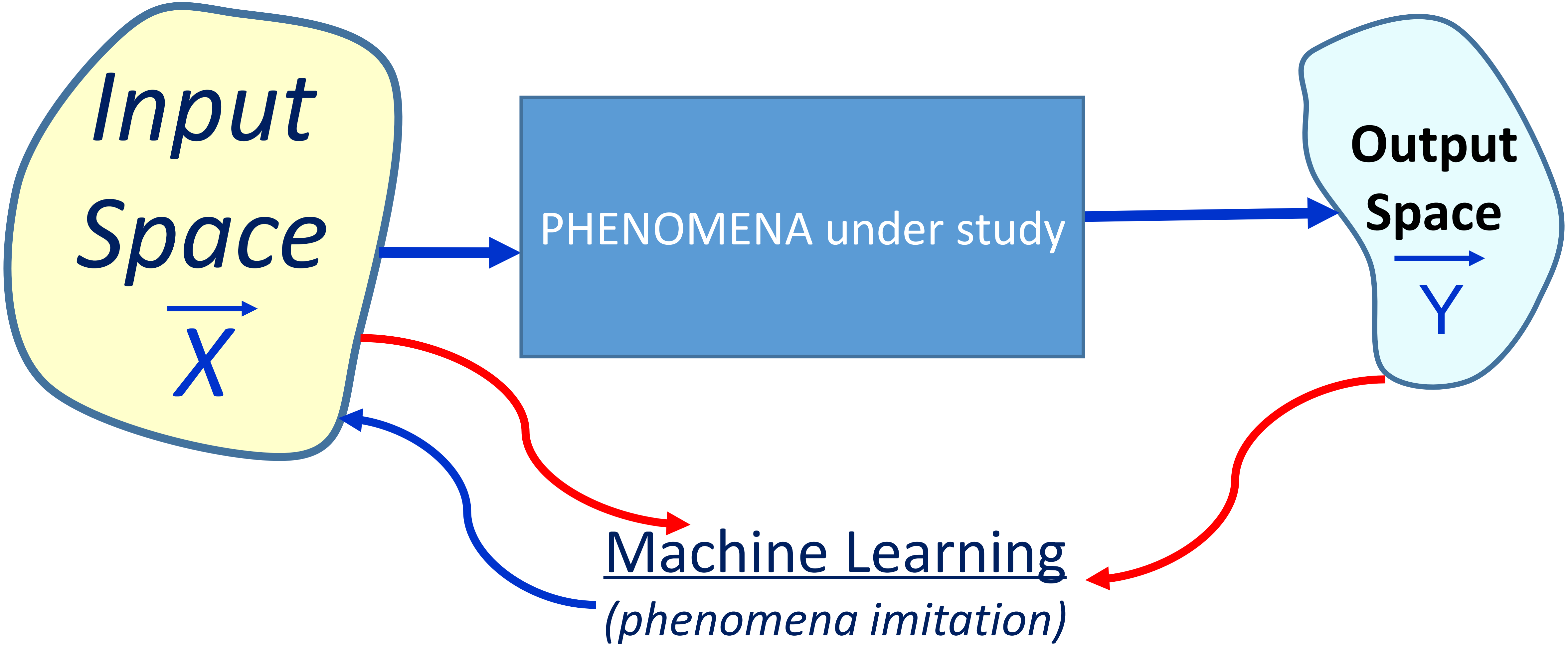




Predictive Learning

(The algorithmic modelling culture, L. Breiman)

(minimisation of the generalisation/testing error)

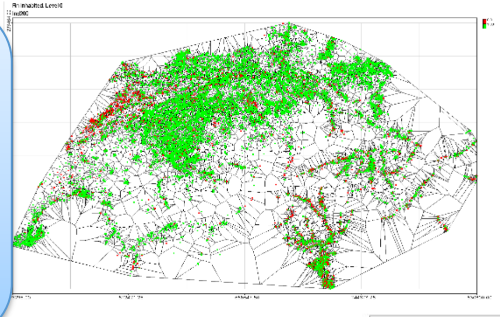


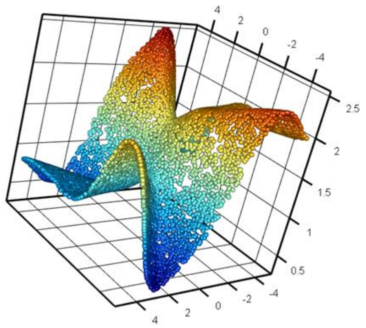
Intelligent exploratory data analysis (IEDA)

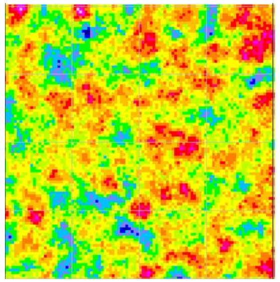
Monitoring networks design & redesign, Sampling, Clustering Validity domain

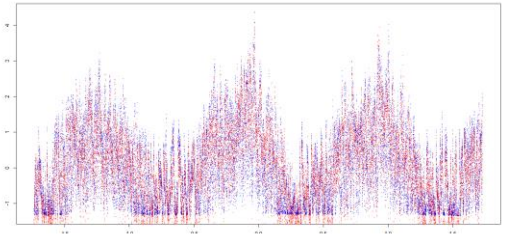
Predictability, noise Patterns Yes/No

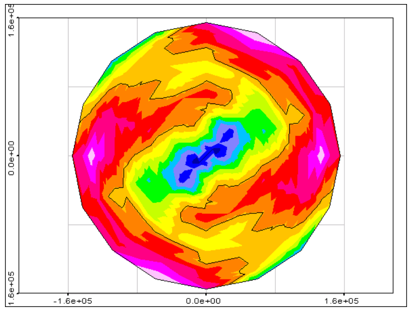
Visual Analytics











Statistics, Traditional EDA

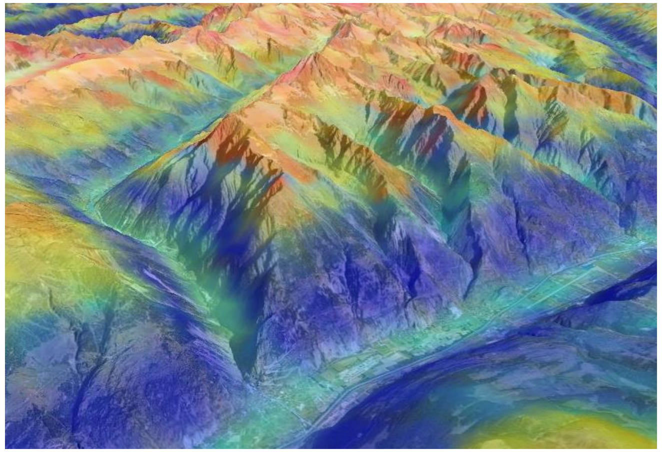
Data cleaning (data quality) Pre-processing

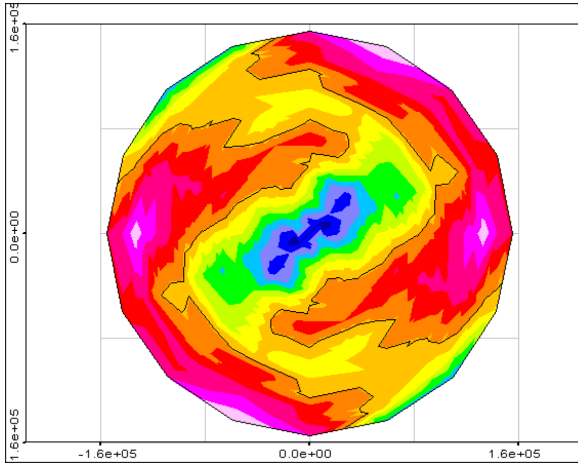
Feature construction (Featurerization)

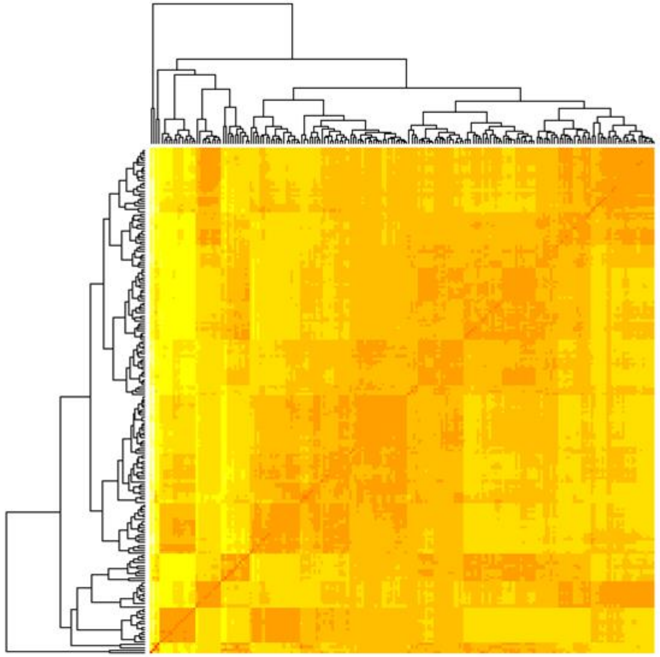
Correlations/Dependencies Spatial, Temporal, Spatio-temporal

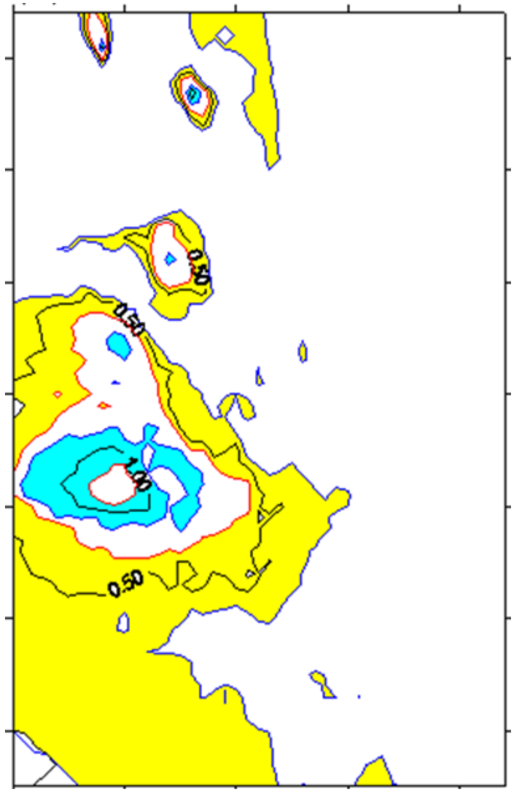
V I S U A L I Z A T I O N

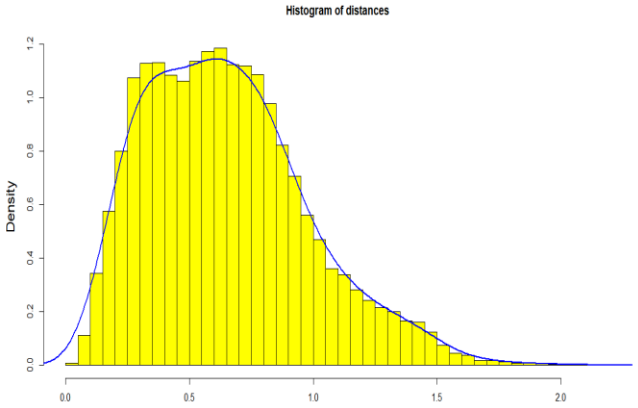
Visualise ASAP and AMAP!

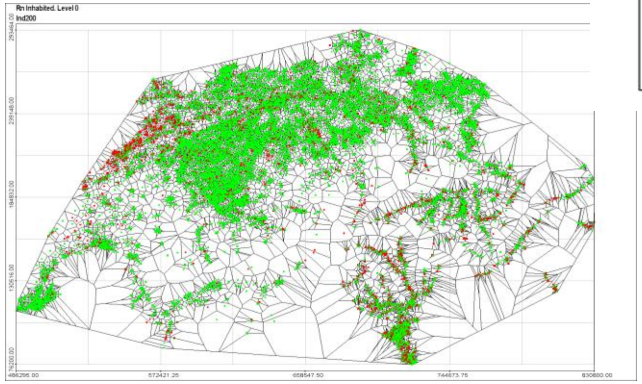


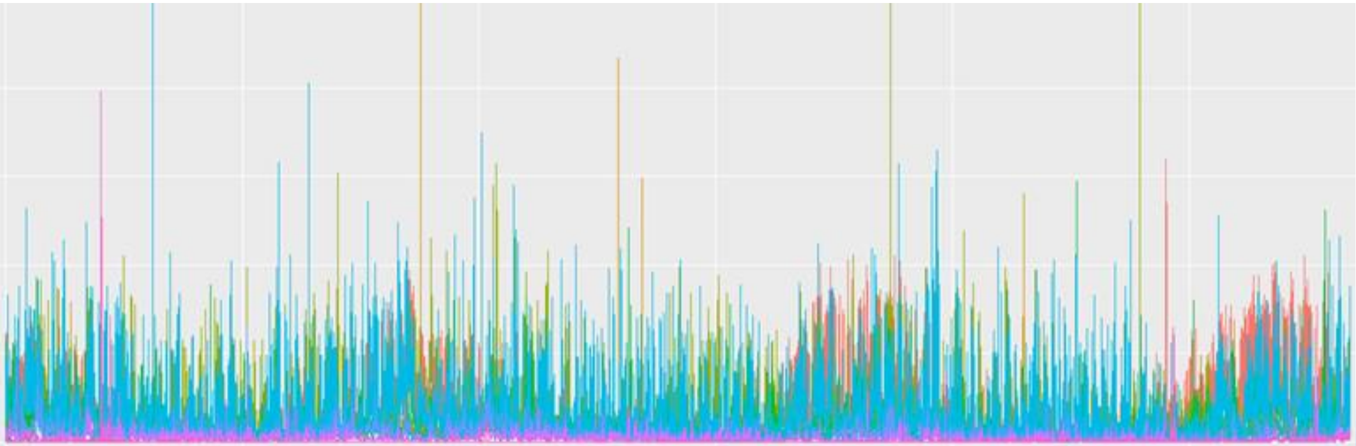


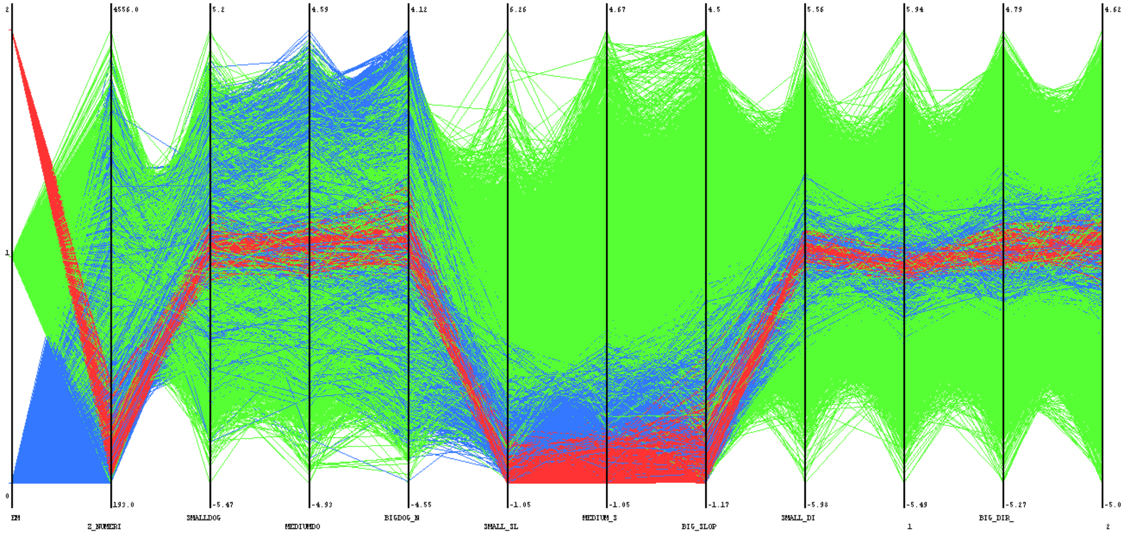






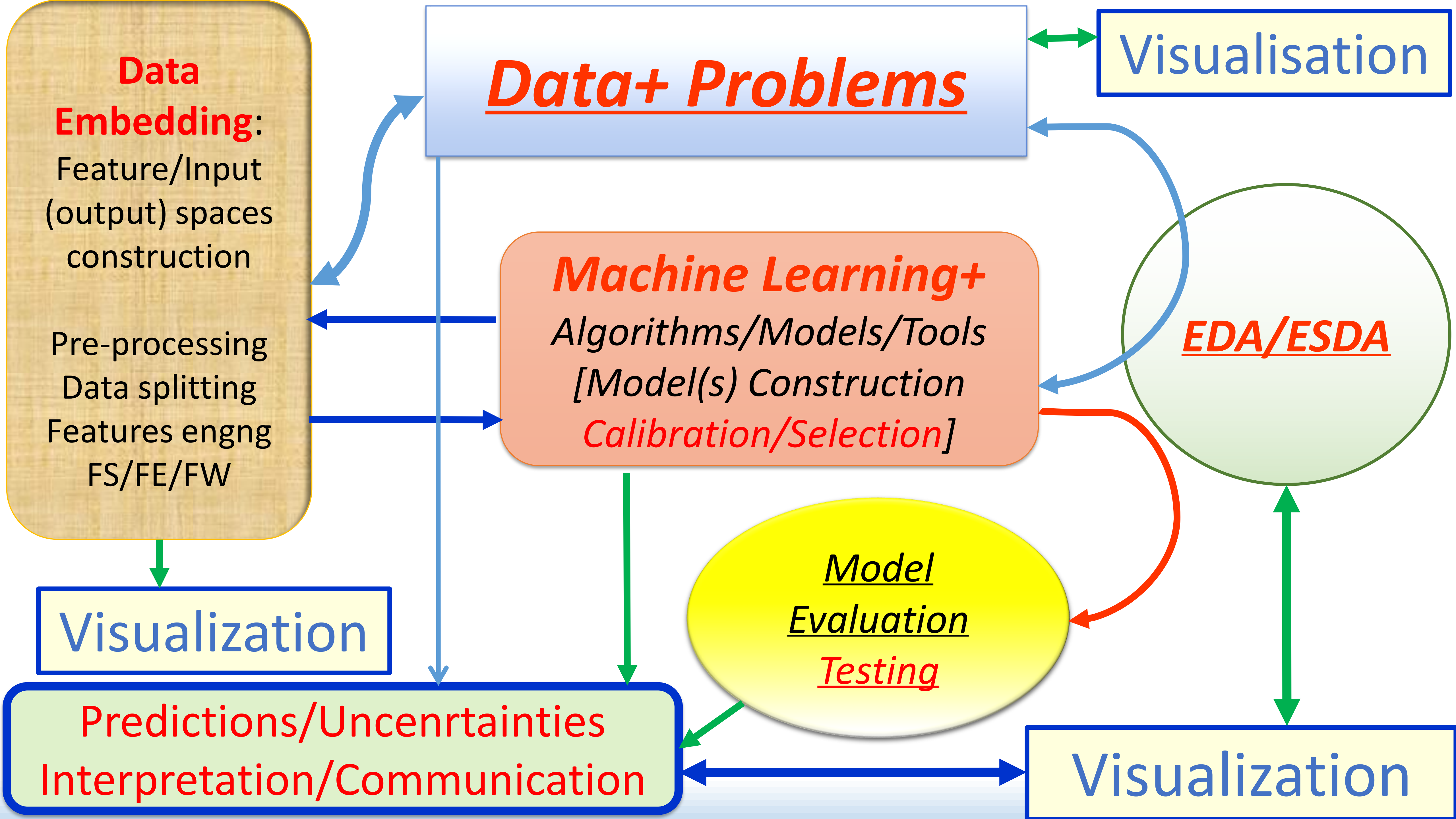






IEDA conclusions

1. Intelligent EDA (IEDA) often is $\geq$ 70-80% of the success -> quality of data
2. IEDA helps in developing interpretable ML and selecting relevant modelling tools (not necessarily ML)
3. IEDA often applies ML tools to better understand data and phenomena under study
4. Visualisation: important at all steps of the study. Visual analytics and data mining
5. IEDA for data, results and residuals!



Data Embedding:
Feature/Input (output) spaces construction
Pre-processing
Data splitting
Features engng
FS/FE/FW
Data+ Problems
Visualisation
Machine Learning+
Algorithms/Models/Tools
[Model(s) Construction
Calibration/Selection]
EDA/ESDA
Model
Evaluation
Testing
Visualization
Predictions/Uncenrtainties
Interpretation/Communication
Visualization

Construction of the input/feature space

(Data Embedding)

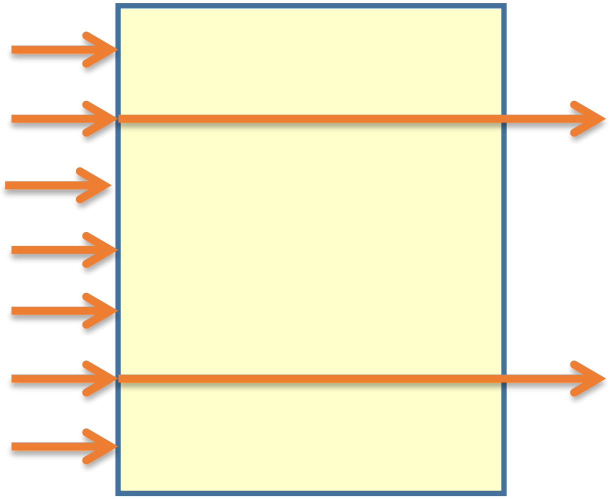
Input space in real complex geo- environmental data case studies is rarely known.

Construction of the input feature space: Expert knowledge, publications, previous studies, experimentation, feature engineering

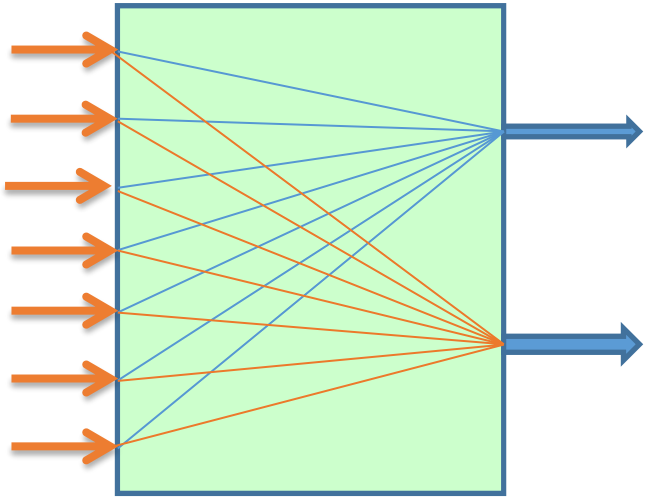
Features can be: *Relevant (RL), Redundant (RD) or Irrelevant (IR)*

Therefore: feature selection/extraction phase in ML modelling is very important. This process can be dynamic.

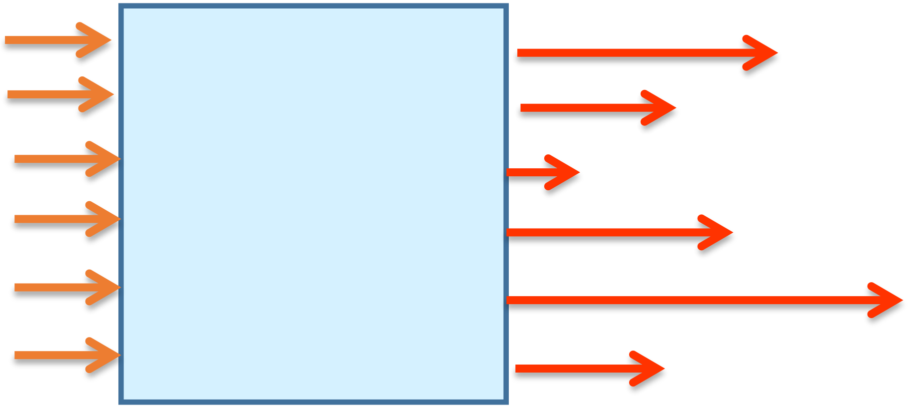
Feature Selection



Feature Extraction



Feature Weighting

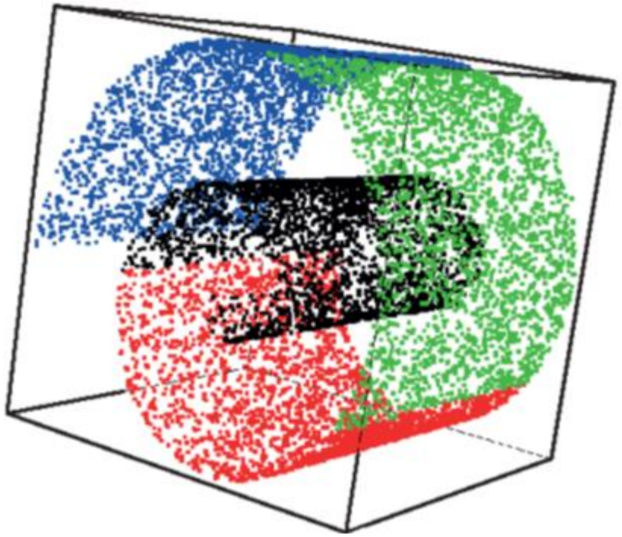


Filter

Wrapper

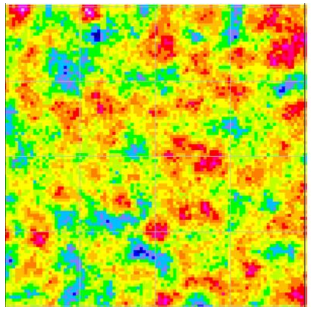
Embedded Method

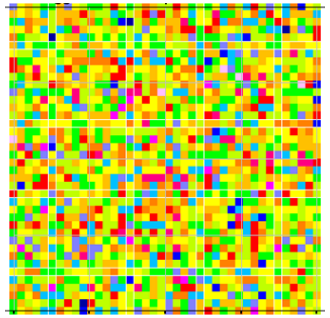
Comments on FS
& dimensionality reduction:



- FS helps in understanding of data and phenomena under study
- Improves modelling without the loss of the quality
- Computational efficiency
- Curse of dimensionality.
- FS & active learning
- FS can be used to develop relevant embedding for time series
- Recently Deep Learning has pushed frontiers in this domain via automatization of feature generation and selection.

DATA decomposition





PATTERNS (information) + **NOISE (residuals)**

If you know/estimate a noise level (unexplainable variability) in data – you know a lot!
There are quite good algorithms to do it: delta test, gamma test, nonparametric estimates,…

1-NN noise estimator

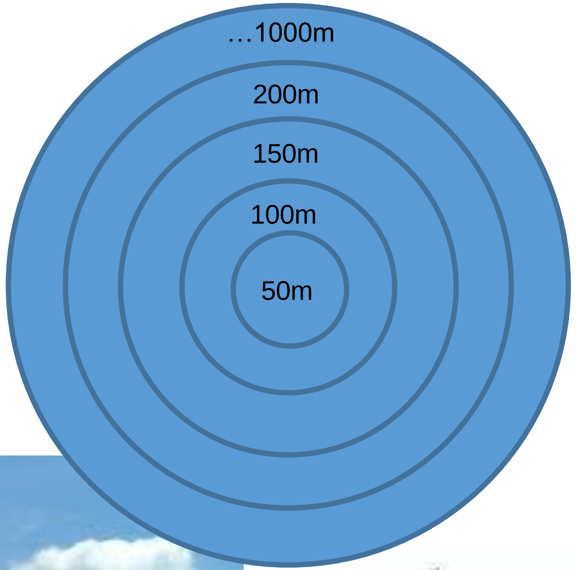
$$Y_i = m(X_i) + r_i$$

$$\mathrm{Res}\,Var_M^{1NN} = \frac{1}{2M} \sum_{i=1}^{M} (Y_i - Y_{N[i,1]})^2$$

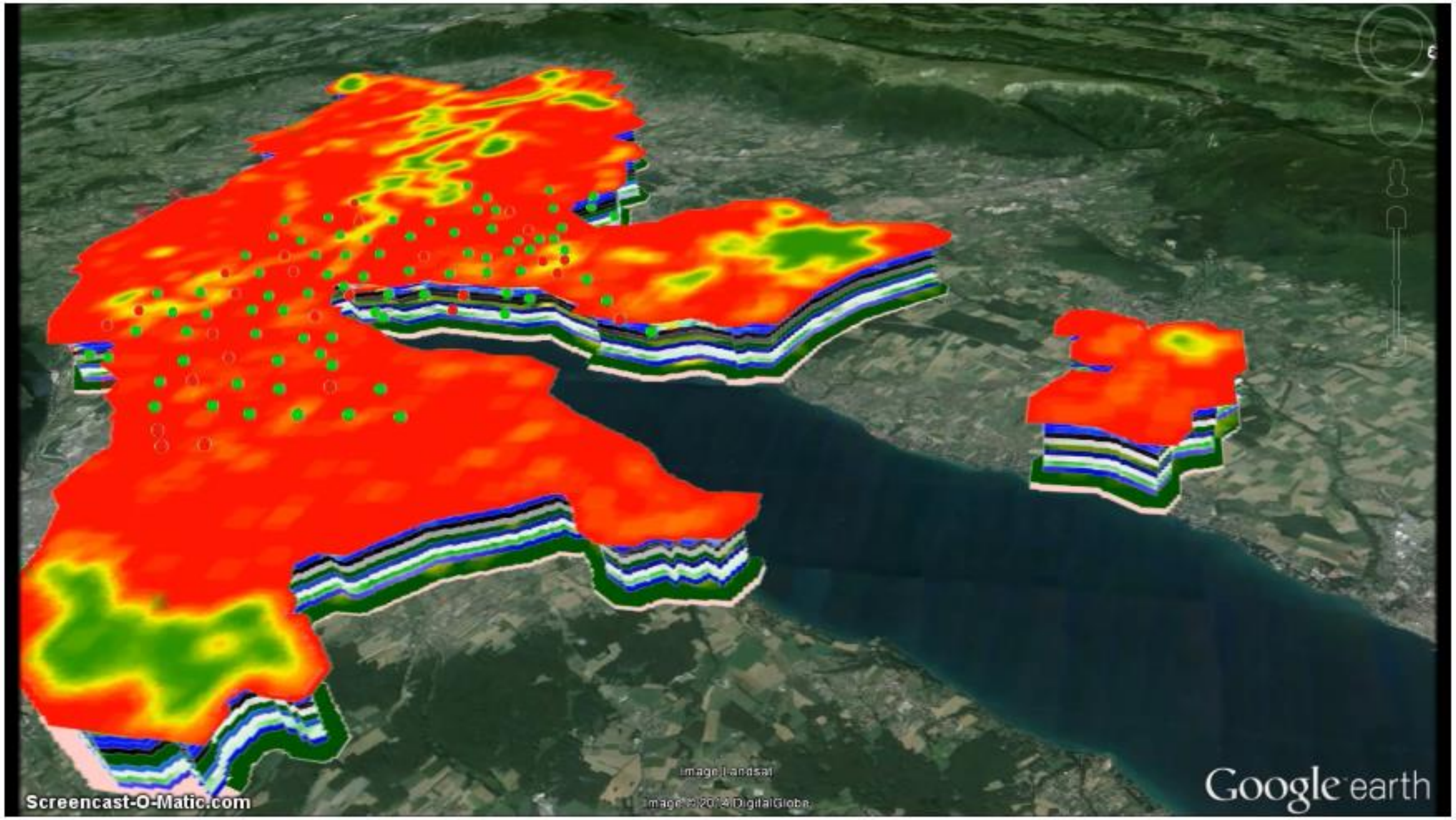
Pollution in a city using ML-based Land Use Regression (LUR) Models

(A. Champendal et al., Air Pollution Mapping Using Nonlinear Land Use Regression Models, 2014)

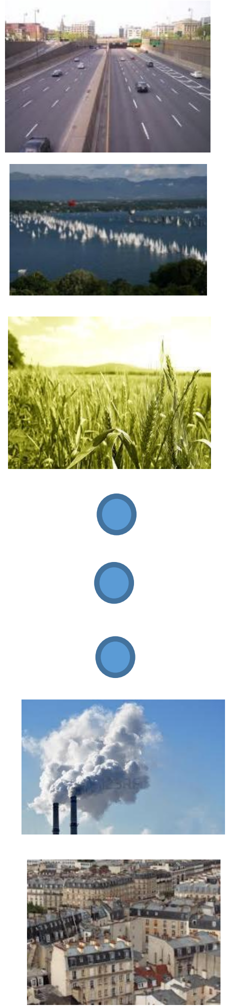
Pollution in a city: Input space construction

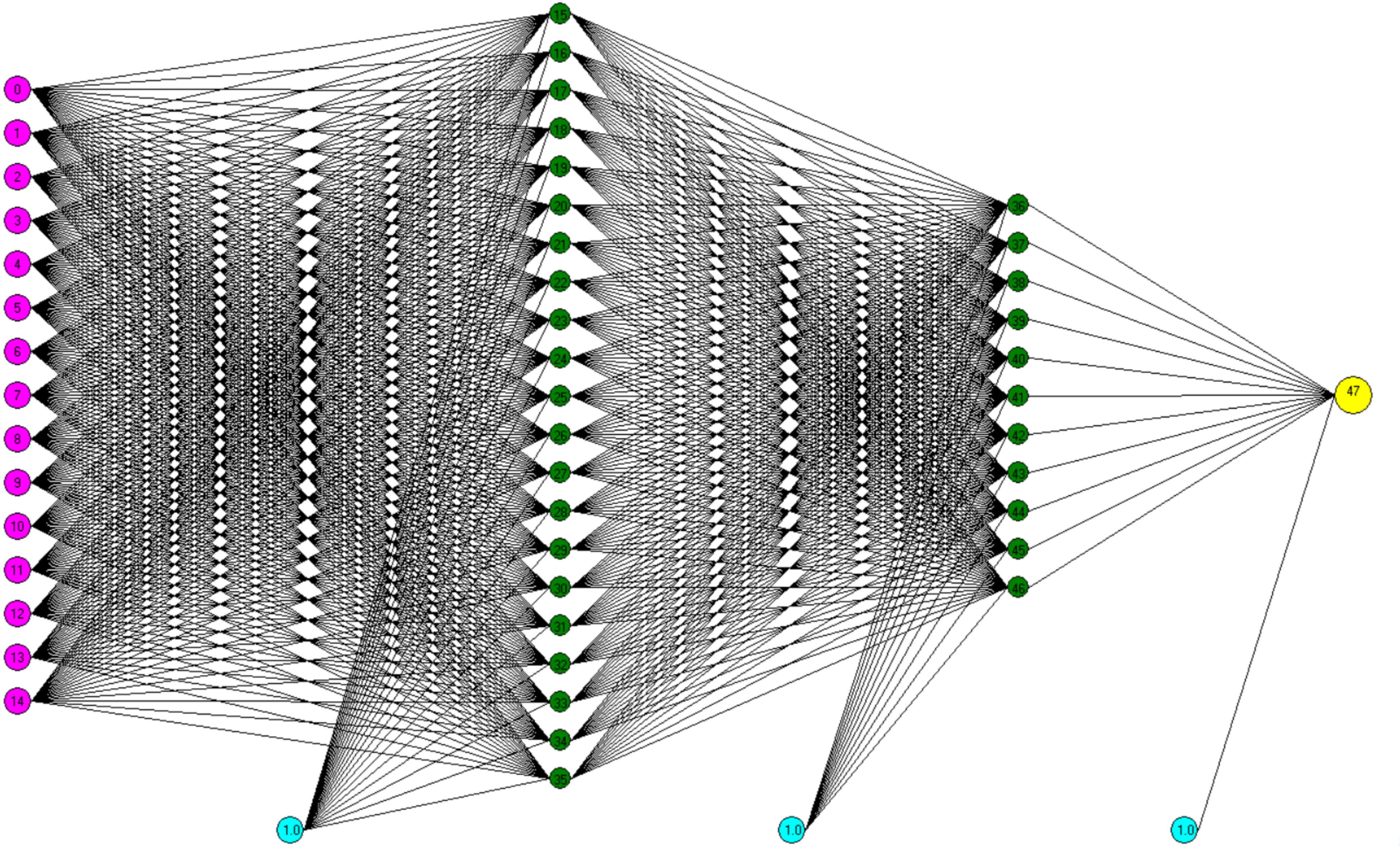


“Sandwich” of the input features

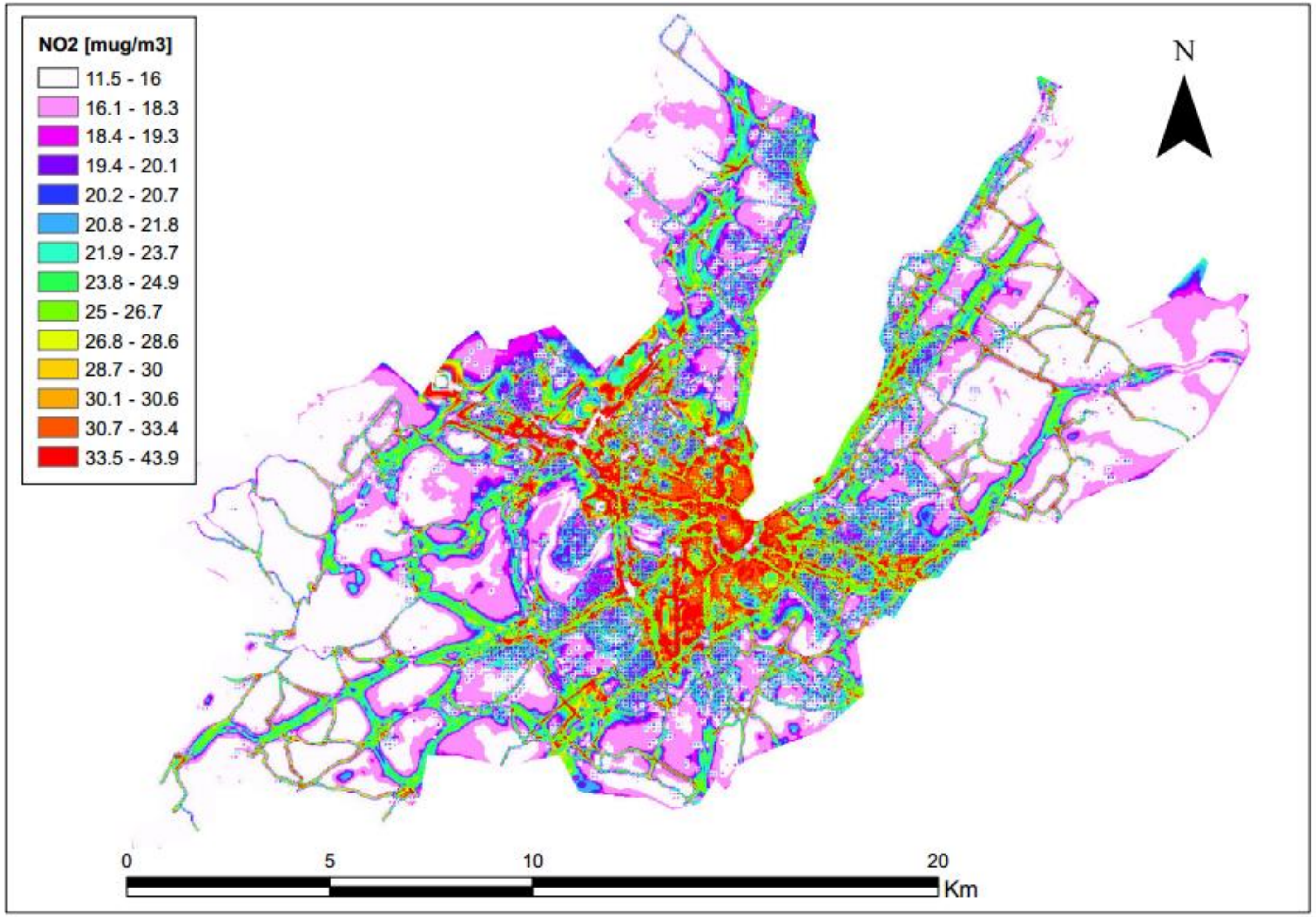


Pollution mapping using Multilayer Perceptron





Nonlinear (MLP, RF) Land Use Regression Modelling



Classical Linear LUR Testing:

R2 = 0.56
RMSE train = 4.51
RMSE test = 4.78

R2 = 0.78. RMSE Train = 2.93. RMSE Test = 3.88

Problem	MLA Used
Clustering, dimensionality reduction	k-means, kernel k-means, EM, MDN, GMM, ELM, SOM, manifold learning, Sftools, IDmining, kPCA Autoencoders,…
Classification	kNN, kernel kNN, MLP, RBF, PNN, GP, SVM, ELM , RF, DL,…
Regression (mapping), Forecasting	kNN, kernelkNN, MLP, SVR, RBF, AGRNN, GRNN, GP, ELM, RF, DL
Advanced topics	Active learning (data collection, MNO), Multi-task learning, Multi-kernel learning, Semi-supervised learning, Transfer learning, Uncertainties quantification, Hybrid models,…

Some conclusions

- ML is a very good exploratory and modelling approach having many useful tools and instruments. ML is data-driven and significantly depends on data quality and quantity
- Learn and use different ML models (use simulated, shuffled and benchmark data, learn from previous studies,…). There is no “free lunch”
- Do not forget hypotheses and conditions behind. Visualise, regularise, validate and test, explain, discuss and communicate.

Challenges and current trends

- Data collection and Intelligent EDA (*data centric approach*)
- From dependencies to cause-effect relationships
- Wider application of active learning, multi-task learning, transfer learning, ensemble learning, etc.
- Uncertainties. Risks and extremes. Model evaluation criteria
- Visualization and visual analytics
- Science-based and data-driven models: *physics-aware ML,…*
- *Interpretability/Explainability of models, results and decisions*
- ***Education**: new curricula + DATA thinking & intuition.*
- *New generation of researchers: excellent domain knowledge integrated with deep understanding of ML*

Thank you for you attention!