ML-PREP
Training Session
PROGRAM

Mark Asch - ML-PREP

2025

Outline

Day	Topic	Links
1	Machine Learning	Lectures Examples
2	Geostatistics	Lectures Examples
3	Wave propagation	Lectures Examples
4	Research projects	Directory

1. PLEASE CLICK ON THE LINKS

2. The mornings will be dedicated to lectures.

3. The afternoons will concentrate on practical examples and exercises.

DAY 1:

MACHINE LEARNING

Advanced ML

- Pre-requisites
  - ⇒ Basics of Machine Learning: lecture notes and examples are here.
- Theory:
  - ⇒ how to choose a method?
  - ⇒ cross-validation and tuning
  - ⇒ evaluation and performance metrics
  - ⇒ causality and correlation
  - ⇒ features and model selection
  - ⇒ PINN
- Examples and Exercises:

DAY 2:

GEOSTATISTICS
and
MACHINE LEARNING

Geostatistics & ML

- Pre-requisites:

  $\Rightarrow$ Basic course lecture notes

- Theory:

  $\Rightarrow$ Geostatistics
    $\rightarrow$ probability and stochastic processes
    $\rightarrow$ variograms
    $\rightarrow$ kriging

  $\Rightarrow$ Geospatial data analysis and machine learning
    $\rightarrow$ model evaluation
    $\rightarrow$ spatial cross-validation

- Examples and Exercises

DAY 3:

WAVE PROPAGATION

Wave Propagation

- Pre-requisites
- Theory:
  - ⇒ basics of seismic wave propagation: harmonic waves, acoustic waves, seismic waves
  - ⇒ finite difference method
  - ⇒ finite element method
  - ⇒ spectral element method
- Examples and Exercises

DAY 4:

RESEARCH PROJECTS

Propositions

1. Landslide inventory using ensemble, tree-based machine learning.

2. Extensive machine learning study of parameters in the Factor of Safety formula of Newmark.

3. Coupling seismic wave propagation with landslide triggering (Newmark equation).

TRAINING OVERVIEW

Where do we begin?

- Different starting points and prior experience:
  - ⇒ mathematics, statistics, probability theory
  - ⇒ data science and machine learning
  - ⇒ GIS and geospatial data

Where do we arrive?

- Unique end point that ensures everyone is at the same level of knowledge of:
  - $\Rightarrow$ machine learning,
  - $\Rightarrow$ geospatial data analysis,
  - $\Rightarrow$ basic seismic wave propagation.

How do we get there?

- Presentation of tools in a big toolbox.
- Understanding why, and not just how!
- Ethical, reproducible and responsible science...

ML for Science

- Motivation: see initial lecture.

- Objective: find the mapping (function, pattern) f that relates outcomes Y (observations, measurements) to explanatory variables (inputs, features, causes) X, such that
$$Y = f(X) + \epsilon,$$
where ϵ represents the intrinsic uncertainty (noise) of the underlying phenomenon.

- In practice, we will seek an approximation $\hat{f}$ to f such that the resulting
$$\hat{Y} = \hat{f}(X)$$
is as close as possible to Y.

- This is done by minimizing a suitable loss function
$$\mathcal{L} = \left\| Y - \hat{Y} \right\|.$$

The PTP method

- PTP = Please the Professor... (or the Boss)
- In Machine Learning we are strongly tempted to do our best by minimizing the model error:

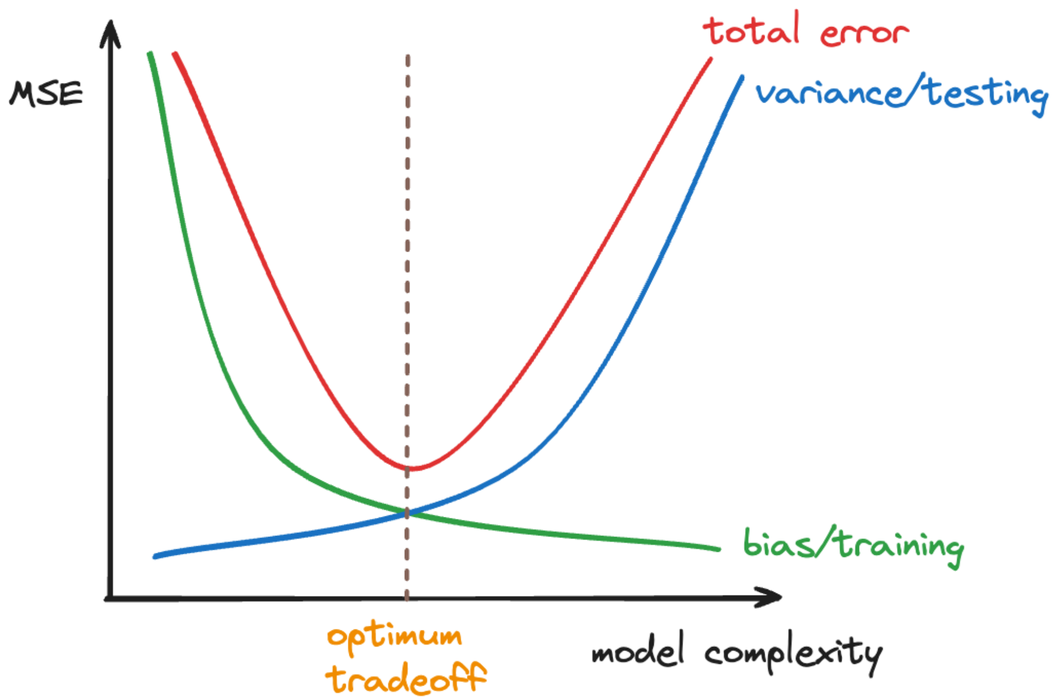
  $\Rightarrow$ RMSE (root mean-squared error) for regression problems,
  $\Rightarrow$ CEE (cross-entropy error) for classification problems.

- Recall the XKCD cartoon: "stir the pile until the answer looks right"

  $\Rightarrow$ This is NOT reproducible, NOT responsible, NOT ethical.

What goes wrong here?

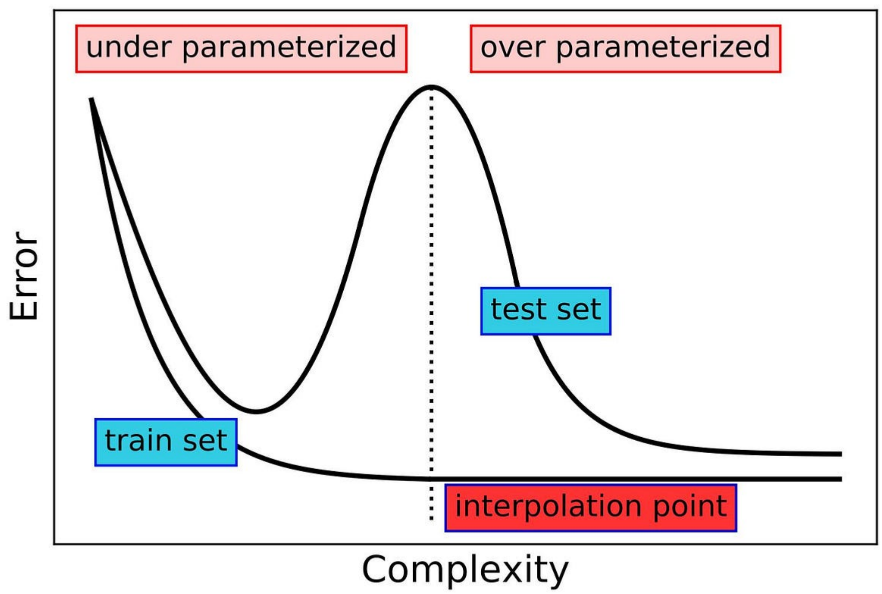
- The bias-variance tradeoff:



- Option 1: fit the noise ⇒ high accuracy (low bias), but large uncertainty (high variance)—overfitting.

- Option 2: compromise ⇒ lower accuracy (higher bias), but lower uncertainty (lower variance)

- Double-descent phenomenon (for DNNs):



What is the objective?

- The principal objective is good predictive performance, NOT good training accuracy.
- Mathematically, recall the trained ML model for $Y = f(X)$, f unknown,

$$\hat{f}\colon X \to \hat{Y}$$

- Prediction: for $(X^*, Y^*) \notin (X, Y)$, how accurate is $\hat{f}(X^*)$?
- In other words, if

$$\hat{f}\colon X^* \to \hat{\hat{Y}},$$

then how big is the error

$$\left\|\hat{\hat{Y}} - Y^*\right\|?$$

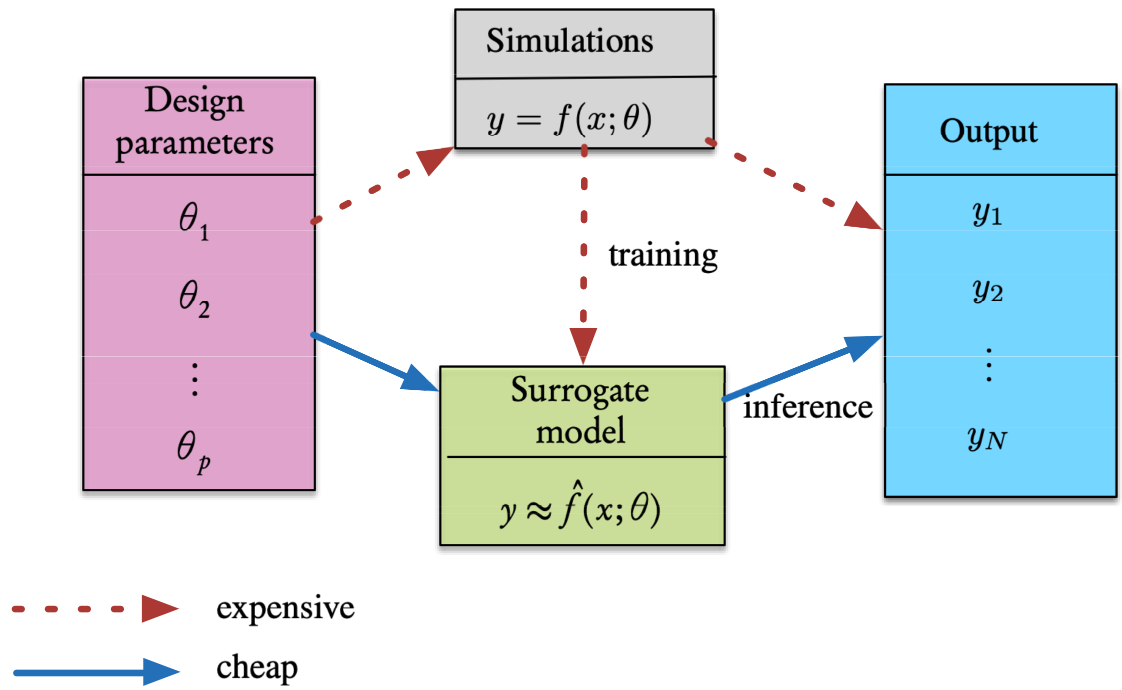
How do we avoid the trap?

- By definition, unseen data has not been seen in the training...
- Best effort in this case is to use

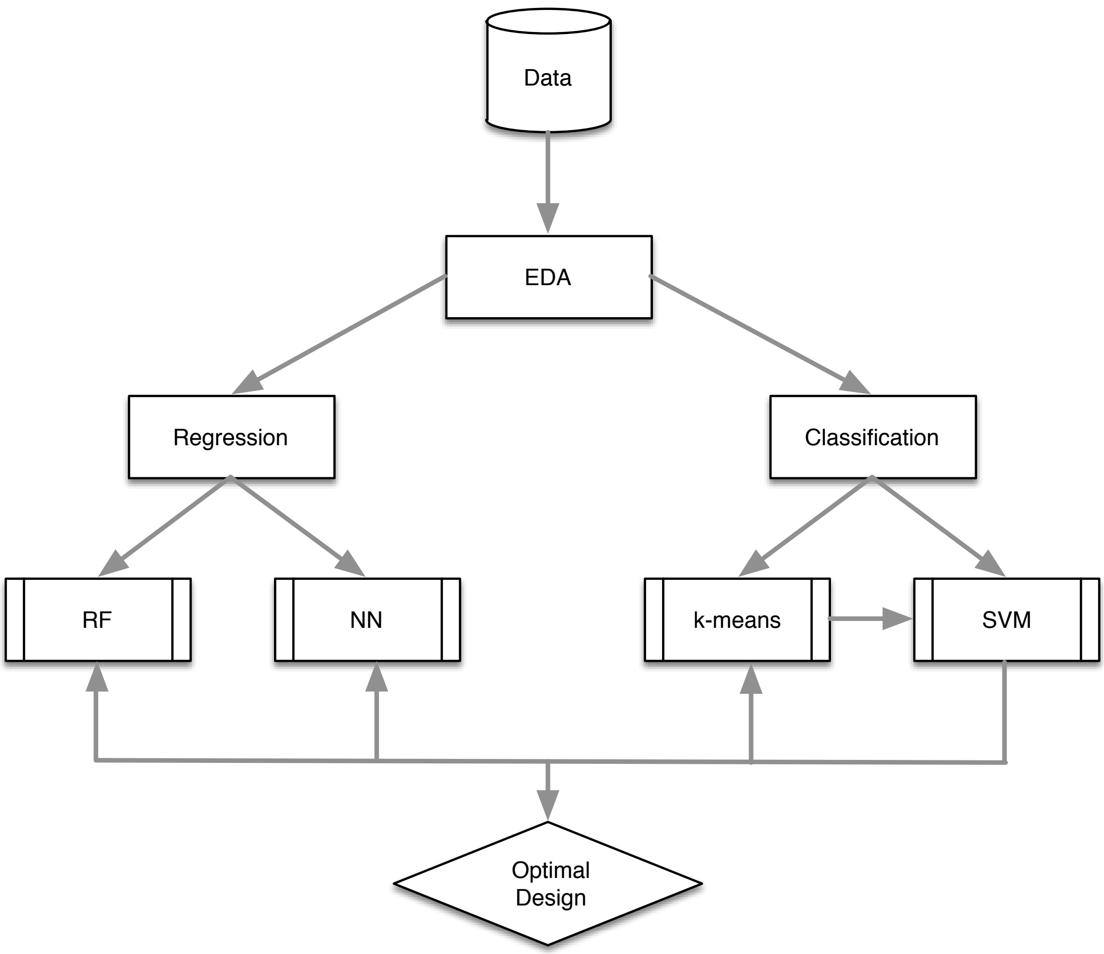
  $\Rightarrow$ nested cross-validation for tuning and testing
  $\Rightarrow$ repeated cross-validation for training and testing

- Then to report confidence intervals, taking uncertainty into account—not just the "best" result (see PTP metod above).

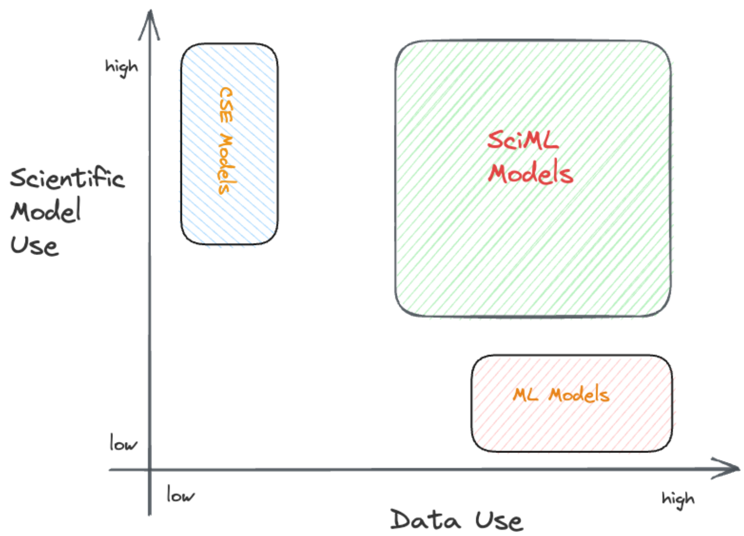
SUMO

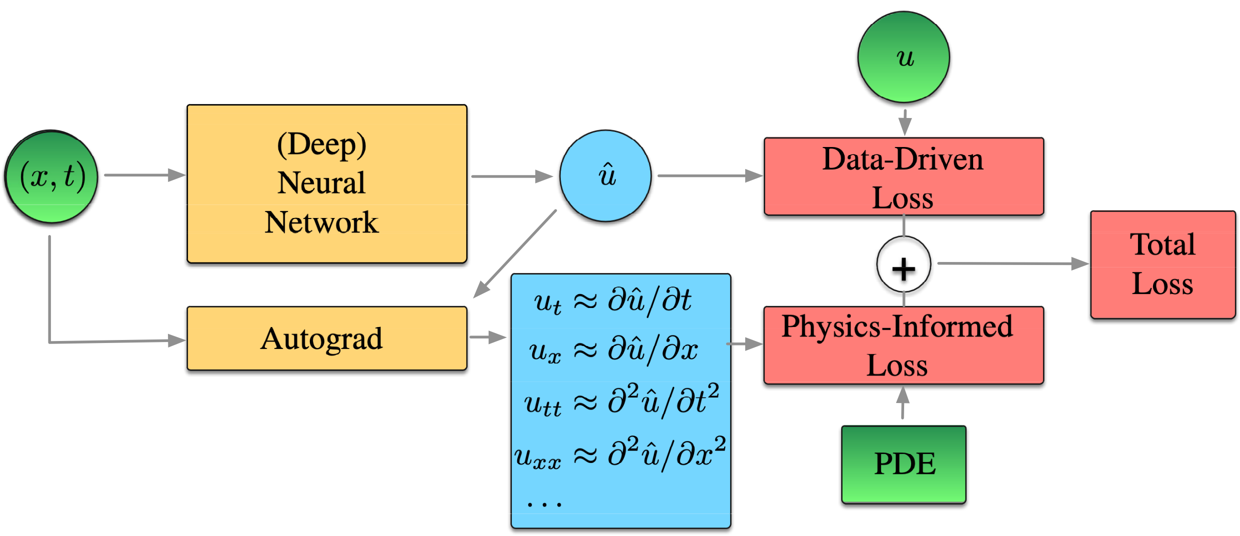

Simulations
$y = f(x;\theta)$
Design parameters
θ_1
θ_2
$\vdots$
θ_p
training
Output
y_1
y_2
$\vdots$
y_N
Surrogate model
$y \approx \hat{f}(x;\theta)$
inference
expensive
cheap

EDA


Data
EDA
Regression
Classification
RF
NN
k-means
SVM
Optimal
Design

SciML & PINN


high
Scientific
Model
Use
low
CSE Models
SciML
Models
ML Models
low
Data Use
high


(x,t)
(Deep)
Neural
Network
$\hat{u}$
u
Data-Driven
Loss
Autograd
$u_t \approx \partial \hat{u}/\partial t$
$u_x \approx \partial \hat{u}/\partial x$
$u_{tt} \approx \partial^2 \hat{u}/\partial t^2$
$u_{xx} \approx \partial^2 \hat{u}/\partial x^2$
$\dots$
+
Physics-Informed
Loss
Total
Loss
PDE

WARNING: ML is not everything!

- ML depends on data availability and data quality—without these we cannot obtain good models and reliable predictions.

- ML should be coupled with classical modeling approaches

  ⇒ statistics
  ⇒ differential equations
  ⇒ empirical knowledge.

- ML (IMHO) will never replace human researchers, and we should not be fooled by apparent fluency as exhibited by LLMs, where we forgo explainability, transparency and reproducibility.

References

1. G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor. *An Introduction to Statistical Learning*. Springer. 2023. ISL Site

2. C. Wikle, A. Zammit-Mangion, N. Cressie. *Spatio-Temporal Statistics with R*. CRC Press. 2019. `https://spacetimewithr.org/`

3. H.P. Langtangen. *Finite Difference Computing with PDEs: A Modern Software Approach.* Springer Cham. 2017. Download pdf

4. H. Igel. *Computational Seismology: A Practical Introduction.* Oxford University Press, 2017. website