Physics-Informed Machine Learning of Geothermal, Geomechanical, Geochemical Process

Velimir V. Vesselinov (monty) (monty@envitrace.com)
Tracy L. Kliphuis (trace) (trace@envitrace.com)

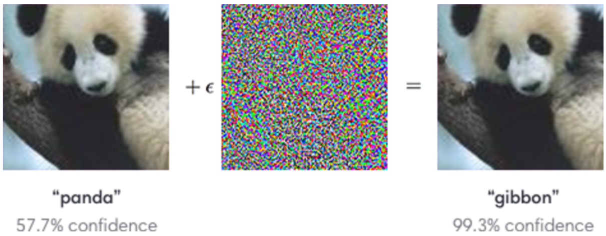
EnviTrace LLC, Santa Fe, NM, USA
https://EnviTrace.com

Machine Learning (ML) Methods

- Supervised
- Unsupervised (Self-supervised)
- Physics (Science) Informed

Supervised Machine Learning (ML)

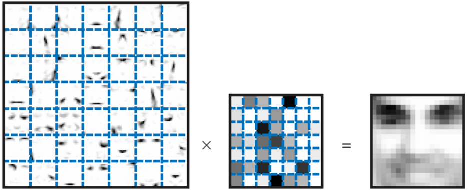
- requires labeling and huge training (labeled) datasets
- introduces subjectivity through the labeling process
- labeling for science applications is challenging
- black box: we do not know why it works
- cannot discover something that we do not know already
- can be severely impacted by data noise: **adversarial examples**



$\Rightarrow$ major limitations of the **supervised** ML methods for **science** applications

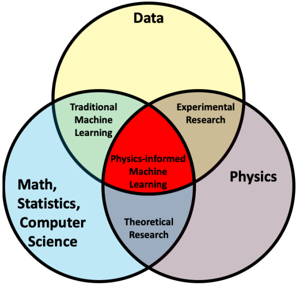
Unsupervised (Self-supervised) Machine Learning (ML)

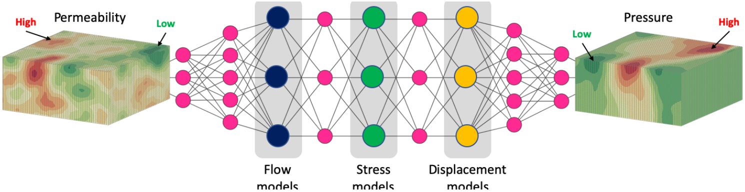
- extracts hidden features (signals, signatures) in the processed data automatically without any prior information
- applicable for both categorization and prediction
- produces unbiased analyses not impacted by data labeling, subject-matter-expert (SME) opinions, and physics assumptions
- identifies features that distinguish images of animals (e.g., cats, dogs, horses, etc.) or geothermal features
- categorizes data and SME can identify (“label”) animals (or geologic features)
- SME needed after ML is performed



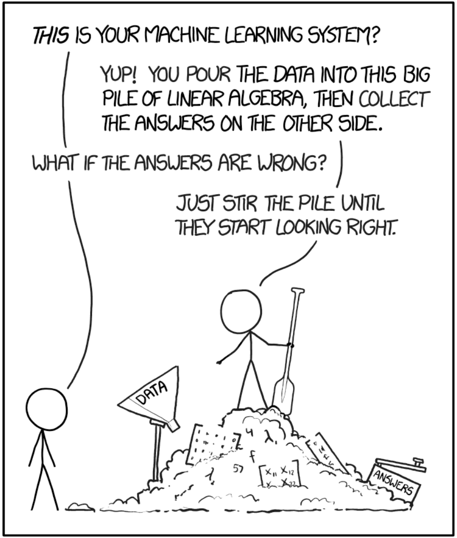
Physics Informed Machine Learning (PIML)

- learns from data but includes preconceived science knowledge
- physics/science information embedded in the ML framework or added as penalties
- PIML neural networks are problem specific
- needs SME inputs related to the analyzed problem
- SME needed before and after ML is performed
- increases efficiency, accuracy, and robustness
- accelerated by differentiable programming (julia)





Machine Learning according to xkcd



SmartTensors

- Novel patented, open-source, unsupervised and physics-informed ML methods and computational techniques
- Based on matrix/tensor factorization coupled with custom k-means clustering and nonnegativity/sparsity/physics-informed constraints
- Developed in julia
  - `http://smarttensors.com`
  - `https://github.com/SmartTensors`
- Capable of efficiently processing large datasets (GB/TB's)
- Applied in numerous geoscience projects (SBIR, DOE, ARPA-E, etc.)

NMF*k* SmartTensors NTF*k*

- **ChemML** (DOE SBIR):
  - `https://envitrace.com/projects/chemml.html`
  - Physics-informed AI/ML for characterization, parameterization, and prediction of contaminant transport and remediation processes
  -
- **GeoTGO** (DOE SBIR):
  - `https://envitrace.com/projects/geothermal.html`
  - Equitable and inclusive tool for community-based geothermal development
- **GeoThermalCloud** (DOE GTO):
  - `https://github.com/SmartTensors/GeoThermalCloud.jl`
  - Cloud Fusion of Big Data and Multi-Physics Models using ML for Discovery, Exploration and Development of Hidden Geothermal Resources
- **ML4Geo** (ARPA-E):
  - `https://github.com/ML4Geo`
  - ML-based Well Design to Enhance Unconventional Energy Production

- Feature extraction (**FE**)
- Blind source separation (**BSS**)
- Detection of disruptions / anomalies
- Separate physics processes
- Discover unknown dependencies and phenomena
- Develop reduced-order / surrogate models
- Identify interrelationships between model inputs and outputs
- Guide development of physics models representing the data
- Optimize data acquisition
- Make predictions

Applications

- **Field Data**:
  - Contamination
  - Climate
  - Seismic
  - Geothermal
  - Oil/gas production
  - CO_2 sequestration
  - Wildfires
  - COVID-19
- **Lab Data**:
  - X-ray Spectroscopy
  - UV Fluorescence Spectroscopy
  - Microbial population growth
  - Fracture development
  - Isotope fractionation
- **Operational Data**:
  - Neutron Accelerator (LANSCE)
  - Oil/gas production
  - Geothermal energy production
  - CO_2 sequestration
- **Model Outputs**:
  - Geothermal
  - Watershed
  - CO_2 sequestration
  - Oil/gas production
  - Climate
  - Reactive mixing $A + B \to C$
  - Co-polymers Phase separation
  - Protein Molecular Dynamics

Publications

- Fleming et al., ML in Earth and Environmental Science Requires Education and Research Policy Reforms, **Nature Geoscience**, 10.1038/s41561-021-00865-3, 2021.
- Siler et al., ML to identify geologic factors associated with production in geothermal fields: A case-study using 3D geologic data, Brady geothermal field, Nevada, **Geothermal Energy**, 10.1186/s40517-021-00199-8, 2021.
- Ahmed et al., ML to Discover Mineral Trapping Signatures due to CO2 Injection, **Journal of Greenhouse Gas Control**, 10.1016/j.ijggc.2021.103382, 2021.
- Ahmed et al., A comparative study of ML models for predicting the state of reactive mixing, **Journal of Computational Physics** 10.1016/j.jcp.2021.110147, 2021.
- Mehana et al., Machine-Learning Predictions of the Shale Wells? Performance, **Journal of Natural Gas Science and Engineering**, 10.1016/j.jngse.2021.103819, 2021.
- Vesselinov et al., Unsupervised ML Based on Non-Negative Tensor Factorization for Analyzing Reactive-Mixing, **Journal of Computational Physics**, Special issue: ML, 2019.
- Stanev et al., Unsupervised Phase Mapping of X-ray Diffraction Data by Nonnegative Matrix Factorization Integrated with Custom Clustering, **Nature Computational Materials**, 2018.
- Vesselinov et al., Nonnegative Tensor Factorization for Contaminant Source Identification, **Journal of Contaminant Hydrology**, 2018.
- O'Malley et al., Nonnegative/binary matrix factorization with a D-Wave quantum annealer, **PLOS ONE**, 2018.
- Vesselinov et al., Contaminant source identification using semi-supervised ML, **Journal of Contaminant Hydrology**, 2017.
- Alexandrov, Vesselinov, Blind source separation for groundwater level analysis based on nonnegative matrix factorization, **Water Resources Research**, 2014.

ChemML: Motivation

- Characterization and remediation of contaminants in groundwater is challenging
- Geologic and geochemical data are massive, complex, and difficult to interpret
- Data are often not fully used
- Building geochemical models representing these data and predicting future behavior is also challenging
- These models also rely on numerous assumptions and are time-consuming to build and execute
- Assumptions can to lead errors
- **ChemML** provides an alternative which develops models that are:
  - fast and simple
  - data driven, robust and minimize assumptions
  - defensible, easy to use and understand

- Example: 4 buckets representing 4 different water types
- Buckets have different geochemical concentrations and contaminants

ChemML: Machine learning for water forensics

- Water from the buckets is mixed in unknown way in the subsurface
- Mixing is caused by many poorly understood subsurface processes

- Water compositions of the original water types (buckets) are typically unknown
- Only groundwater mixtures observed in monitoring wells are known

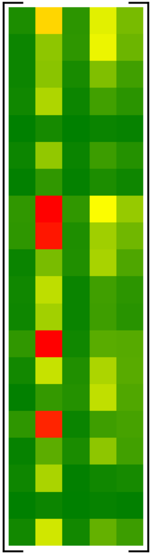
ChemML: Machine learning for water forensics

- **ChemML** can estimate bucket composition using only observed mixtures
- **ChemML** can estimate uncertainties and make predictions

ChemML: How it works?

- **ChemML** uses Machine Learning (ML) to estimate bucket compositions
- Machine Learning is a form of Artificial Intelligence (AI) that works by exploring data and finding patterns with minimum human intervention

ChemML: data matrix



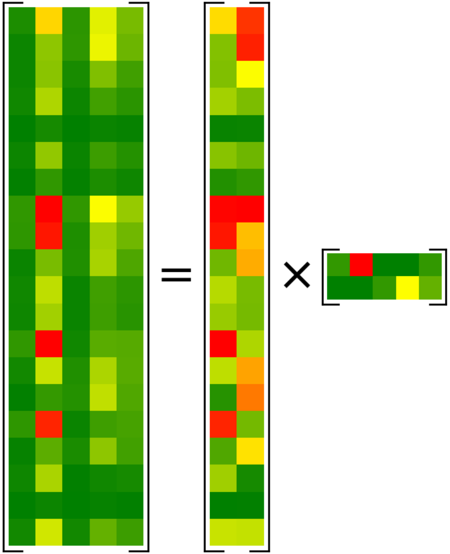
$\boldsymbol{X}$

$[20 \times 5]$

$\boldsymbol{X}$ – **data** matrix
[**geochemical species** $\times$ **monitoring wells**]

$\boldsymbol{X}$ may have empty cells (data gaps)

ChemML: matrix factorization to charecterize water mixing and reactions



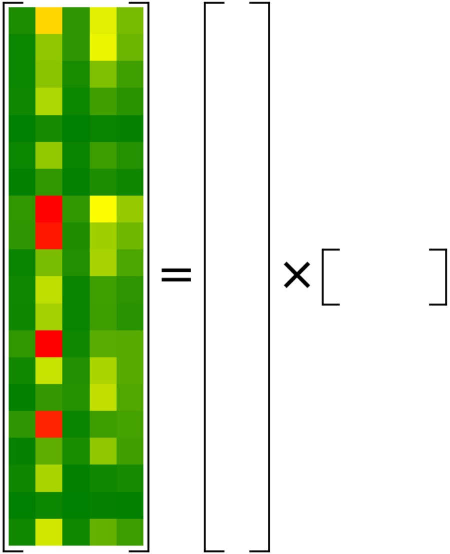
$$\boldsymbol{X} = \boldsymbol{W} \times \boldsymbol{H}$$
$$[20 \times 5] = [20 \times 2] \times [2 \times 5]$$

$\boldsymbol{X}$ – **data** matrix
[**geochemical species** $\times$ **monitoring wells**]

$\boldsymbol{W}$ – **bucket** matrix
[**geochemical species** $\times$ **buckets**]

$\boldsymbol{H}$ – **mixing** matrix
[**buckets** $\times$ **monitoring wells**]

ChemML: matrix factorization to charecterize water mixing and reactions



$\boldsymbol{X} = \boldsymbol{W} \times \boldsymbol{H}$

$[20 \times 5] = [20 \times \textbf{?}] \times [\textbf{?} \times 5]$

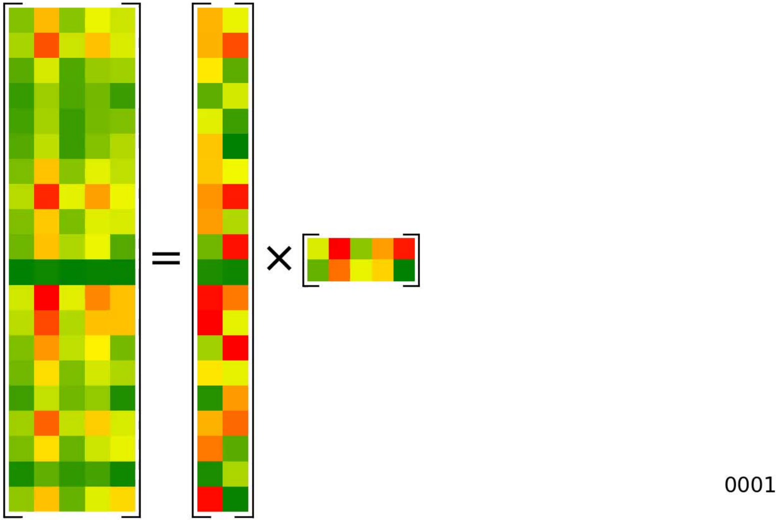
$\Rightarrow$ 100 **knowns** (if there are no data gaps)

$\Rightarrow$ **unknown** number of buckets (2 or more)

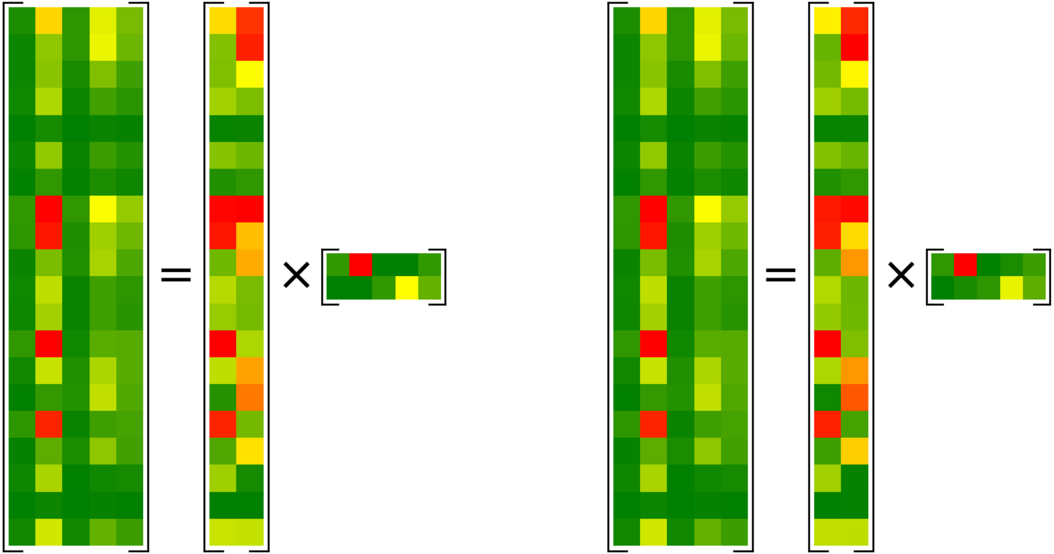
$\Rightarrow$ **unknown** matrix elements of $\boldsymbol{W}$ and $\boldsymbol{H}$ (50 or more)

$\Rightarrow$ **Physics constraints**:

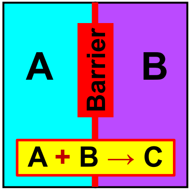
- all elements of $\boldsymbol{W}$ and $\boldsymbol{H} \geq 0$
- $\sum_{k=1}^{K} H_{k,j} = 1 \; \forall j$


0001

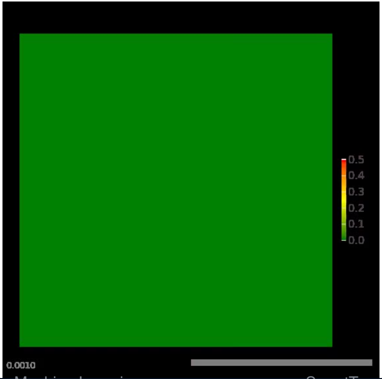
ChemML: true vs. estimated matrix factorization

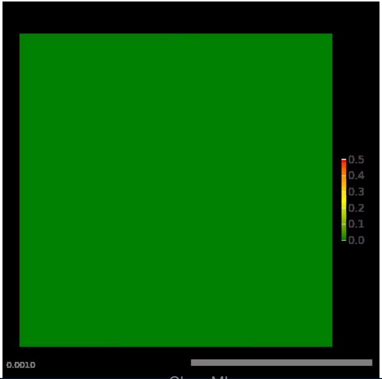


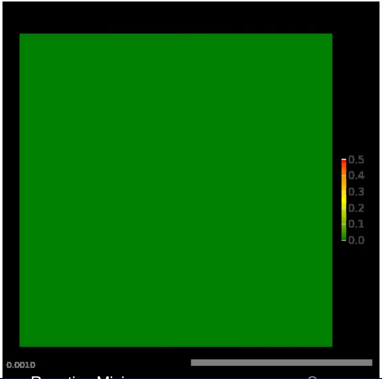
Reactive mixing



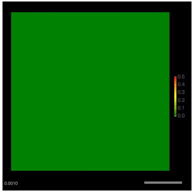
- > 2000 simulations of C concentrations in time/space with varying model inputs representing reactive mixing (5 input model parameters)
- **NTFk** identifies physics processes impacting C concentrations and their relationship to model inputs



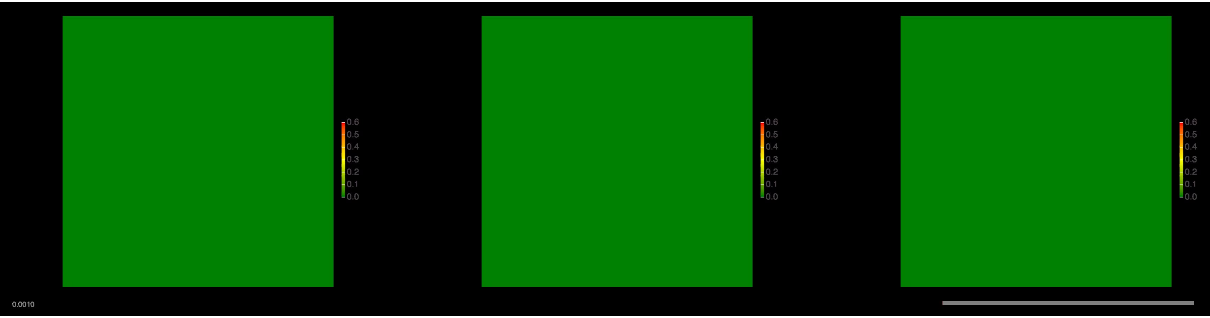




Reactive mixing: SmartTensors results

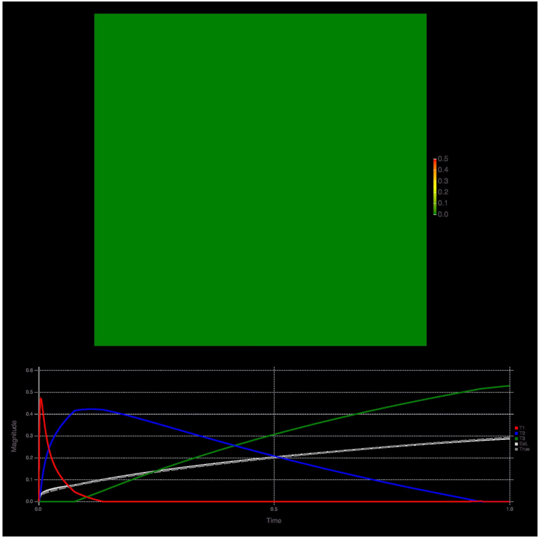


- **NTFk** extracts the dominant time/space features (**processes** / **vortices**) and compresses the model outputs
- Compression: > 200GB $\rightarrow \sim 70$MB (ratio ~ 3000
  Here, $(1000 \times 81 \times 81) \rightarrow (3 \times 12 \times 13)$ $(time \times rows \times columns)$

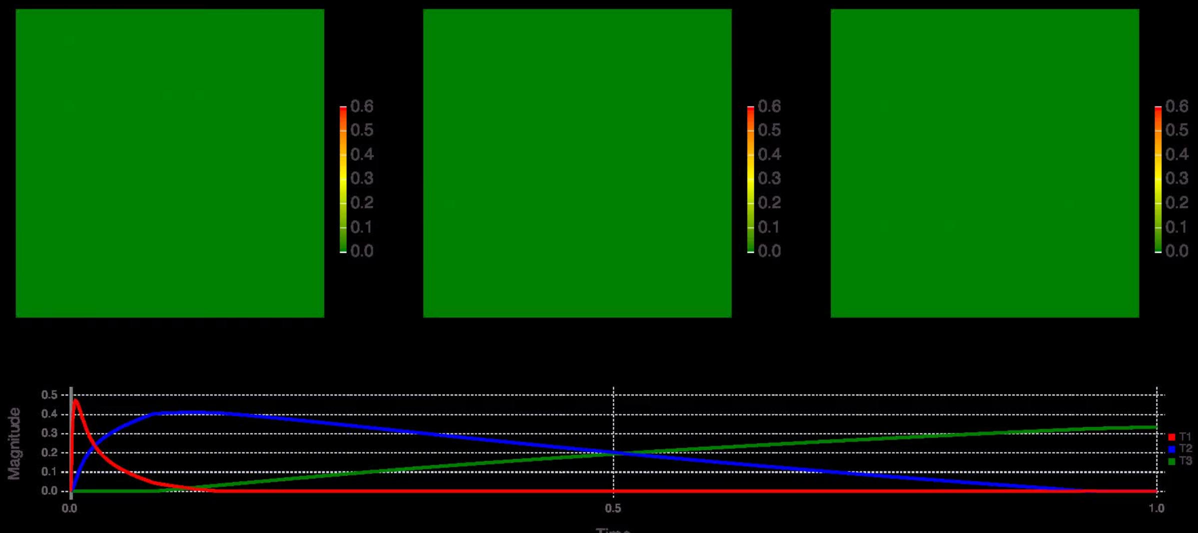


Advection **Dispersion** **Diffusion**

- T1: Advection
- T2: Dispersion
- T3: Diffusion



Reactive mixing: NTFk results



ChemML: Summary

- All the models are wrong, but some are useful
- **ChemML** provides effective **data mining solutions**
- **ChemML** applies novel ML methods developed by our team

SmartTensors: Summary

- Developed **novel** unsupervised and physics-informed ML methods
- Implemented in **cutting-edge** open-source computational framework called **SmartTensors** based on Nonnegative Factorization (Matrices/Tensors)
- **SmartTensors** has been used to solve various real-world problems
- **SmartTensors** deployment as a service on `https://JuliaHub.com` is coming soon
- **SmartTensors** just received 2 2021 R&D100 awards

Machine Learning (ML) Algorithms / Codes developed by our team

- Codes:

NMFk | MADS | NTFk

- Projects, Notebooks, Examples, Tutorials:

  `http://tensors.lanl.gov`
  `http://SmartTensors.github.io`
  `https://github.com/SmartTensors`
  `https://github.com/SmartTensors/NMFk.jl/tree/master/notebooks`
  `https://github.com/SmartTensors/GeoThermalCloud.jl`
  `https://github.com/SmartTensors/SmartTensorsTutorials.jl`