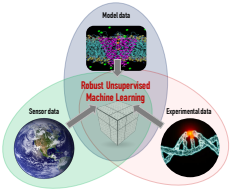
Unsupervised Machine Learning based on Nonnegative Factorization

Velimir V. Vesselinov (monty), Boian S. Alexandrov, Daniel O'Malley
Maruti K. Mudunuru, Satish Karra

Earth and Environmental Sciences Division — Theoretical Division
Los Alamos National Laboratory, NM 87545, USA



Machine Learning (ML)

- **Supervised** Machine Learning: requires prior categorization of the processed data (can introduce subjectivity)
- **Unsupervised** Machine Learning: discovers hidden features in the processed data without any prior information
- **Deep** Machine Learning: ... coupled supervised and unsupervised techniques

Unsupervised Machine Learning Applications

- **Data analytics**:
  - Feature extraction (**FE**)
  - Blind source separation (**BSS**)
  - Image recognition
  - Detection of disruptions / anomalies
  - Guide development of physics / reduced-order models representing the data
- **Analyses of model outputs**:
  - Identify dominant processes (features) in the model outputs
  - Guide development of reduced-order models

Nonnegative Factorization + custom clustering

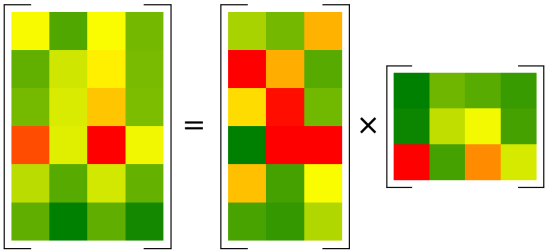
- We have developed a series of novel Unsupervised Machine Learning methods based on Nonnegative Factorization + custom clustering
  - identify **the number of robust features** in the data
  - extract **robust features** representing the data
  - extracted features are parts of the data allowing for **intuitive** interpretations
- Selected publications:
  - Vesselinov, O'Malley, Alexandrov, Contaminant source identification using semi-supervised machine learning, Journal of Contaminant Hydrology, 10.1016/j.jconhyd.2017.11.002, 2017.
  - Alexandrov, Vesselinov, Blind source separation for groundwater level analysis based on nonnegative matrix factorization, Water Resources Research, 10.1002/2013WR015037, 2014
- LANL Unsupervised Machine Learning (ML) **Patent**:
  Alexandrov, Vesselinov, Alexandrov, Iliev, Stanev, Source Identification by Non-Negative Matrix Factorization Combined with Semi-Supervised Clustering, LANS Ref. No. S133364.000, KS Ref. No. 8472-97415-01, August 2017

Nonnegative Factorization + custom clustering

- **NMFk**: matrix-based (two-dimensional) algorithm (well-tested; widely used)
  - Extract barometric and pumping effects in pressure data
  - Identify and predict processes for optimal control of the LANSCE particle accelerator
  - Characterize materials using X-ray
  - Analyze model predictions of molecular dynamics trajectories
  - Characterize influenza epidemics
  - Extract image features using Quantum Computing (**D-Wave**)
  - Identify cancer signatures in human genomes (**30+** papers in Nature/Science/Cell)
- **NTFk**: tensor-based (high-dimensional) algorithm (actively developed at the moment)
- Here, we present **NTFk** applications related to contaminant transport characterization

NMF: Nonnegative Matrix Factorization

X = W × H

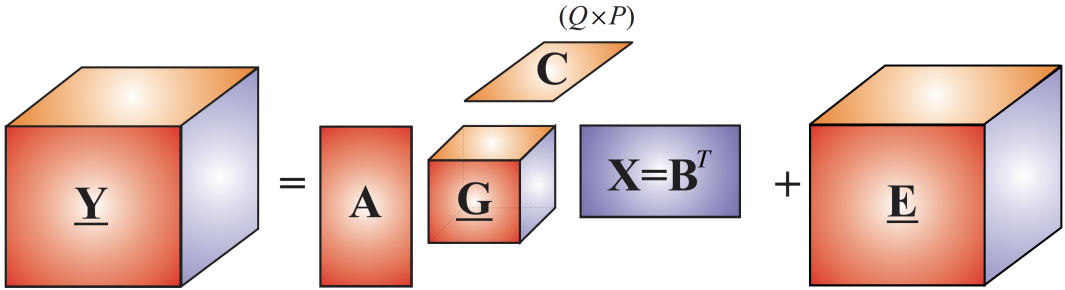


$(I \times Q)$ $(I \times K)$ $(K \times Q)$

- **X**: data matrix
- **W**: mixing matrix (**unknown**)
- **H**: source matrix (**unknown**)

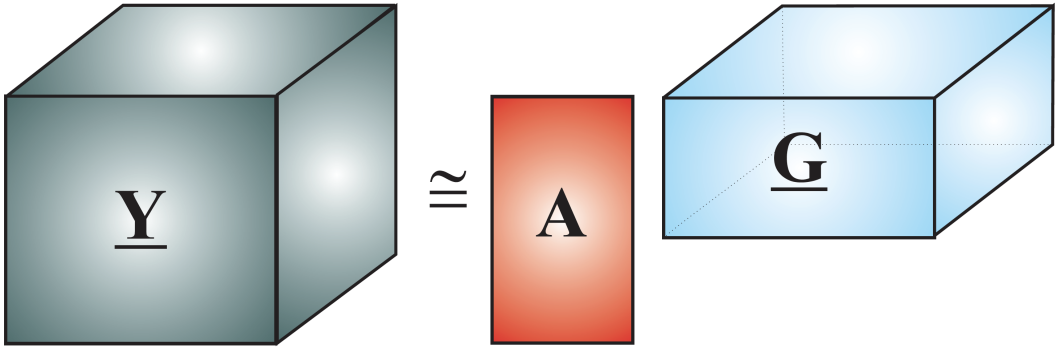
- **I**: number of observation points (wells)
- **Q**: number of geochemical species observed (e.g., Cr^{6+}, SO_4^{2+}, NO_3^-, etc.)
- **K**: number of **unknown** groundwater types mixed at each well
- **Constraints**:
  all matrix elements ≥ 0
  $$\sum_{k=1}^{K} W_{i,k} = 1 \; \forall i$$

Tucker tensor factorization (3D case)



Factorizing all **3** dimensions ($I \to J, T \to R, Q \to P$)

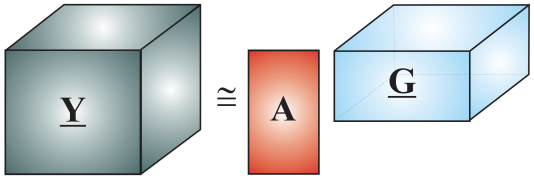
Tucker-1 tensor factorization (3D case)



$(I \times Q \times T)$ $(Q \times K)$ $(I \times K \times T)$

Factorizing only **1** of the dimensions ($Q \to K$)

NTF: Nonnegative Tensor Factorization based on Tucker-1 decomposition



$(I \times Q \times T)$ $(Q \times K)$ $(I \times K \times T)$

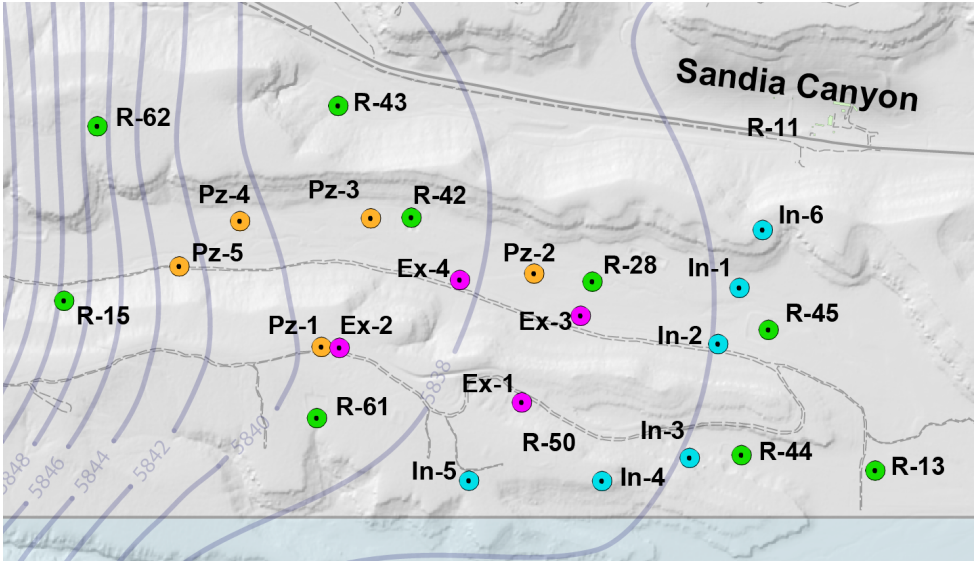
- **Y**: data tensor
- **A**: source (groundwater type) matrix (**unknown**)
- **G**: mixing tensor (**unknown**)

- **I**: number of observation points (wells)
- **Q**: number of geochemical species observed (e.g., Cr^{6+}, SO_4^{2+}, NO_3^-, etc.)
- **T**: number of observation times (e.g, 2001, 2002, ..., 2017)
- **K**: number of **unknown** groundwater types mixed at each well
- **Constraints**:
  all tensor/matrix elements ≥ 0
  $$\sum_{k=1}^{K} G_{i,k,t} = 1 \; \forall i, t$$

Nonnegative Factorization (NTF) Analyses

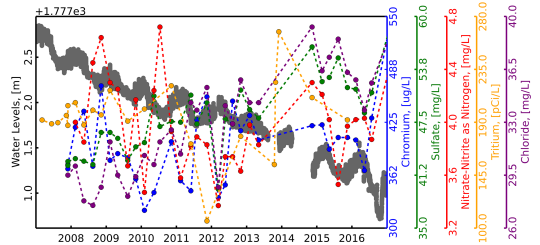
- Major challenges for both **NMFk** and **NTFk**
  - identifying the number of unknown features (groundwater types) **K**
    (in **NMFk**, resolved using custom clustering; based on the Frobenius norm and cluster Silhouettes;
    identification under **NTFk** is much more challenging)
  - solving the constraint optimization problem to estimate matrix/tensor elements
  - dealing with large high-dimensional datasets (high-performance computing)
  - ...
- We apply (demonstrate) **NTFk** to two datasets
  - **Field Data**: time-dependent mixing of groundwater types (contaminant sources)
  - **Simulation Data**: fluid mixing impacts on a geochemical reaction $A + B \rightarrow C$

LANL site

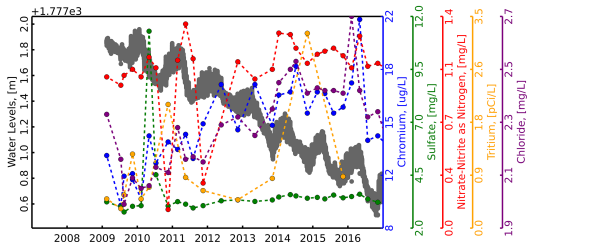

Sandia Canyon
R-62
R-43
R-11
Pz-4
Pz-3
R-42
In-6
Pz-5
Ex-4
Pz-2
R-28
In-1
R-15
Pz-1
Ex-2
Ex-3
In-2
R-45
Ex-1
R-61
R-50
In-3
R-44
R-13
In-5
In-4
5838
5840
5842
5844
5846
5848

LANL hydrogeochemical datasets (high-dimensional)

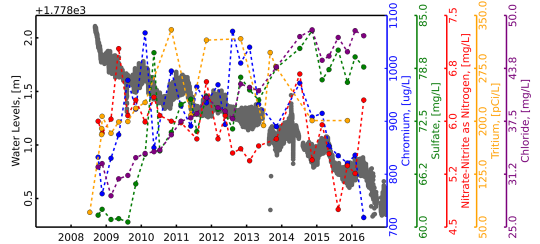
R-28



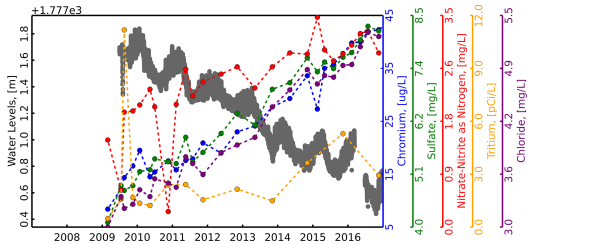
R-44#1



R-42

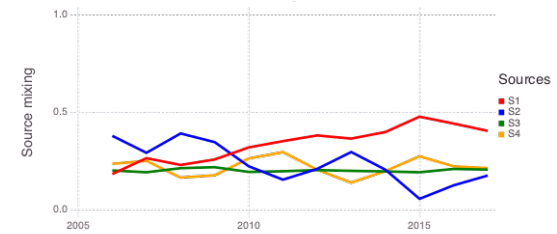


R-45#1

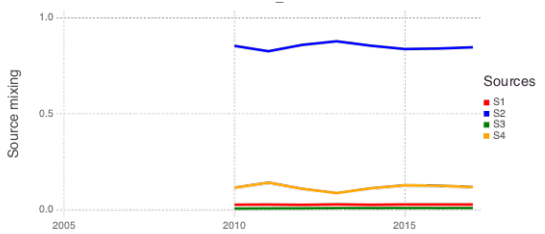


NTFk estimated time-dependent mixing of 4 groundwater types at various wells

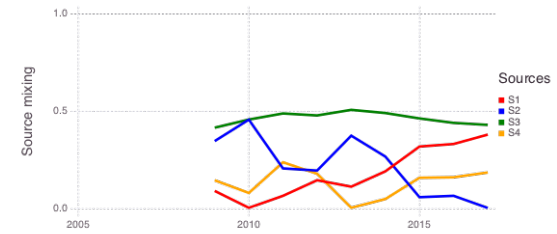
R-28



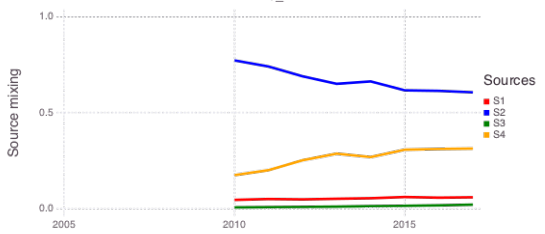
R-44#1



R-42

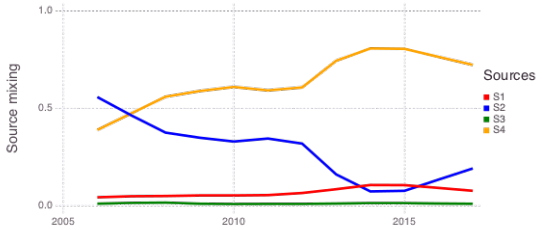


R-45#1

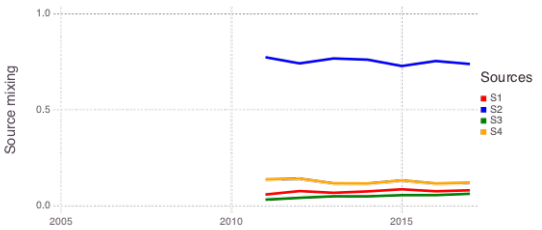


NTFk estimated time-dependent mixing of 4 groundwater types at various wells

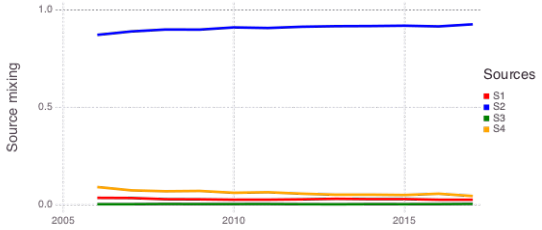
R-11



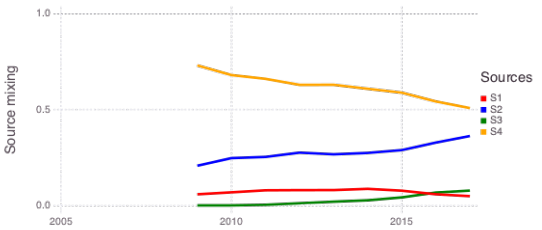
R-50#1



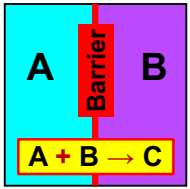
R-1



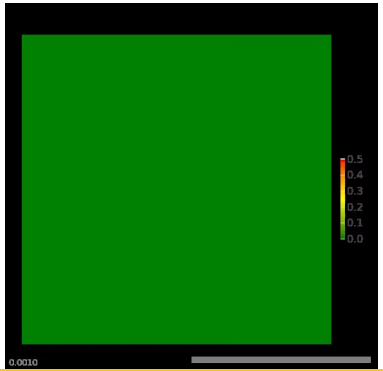
R-43#1

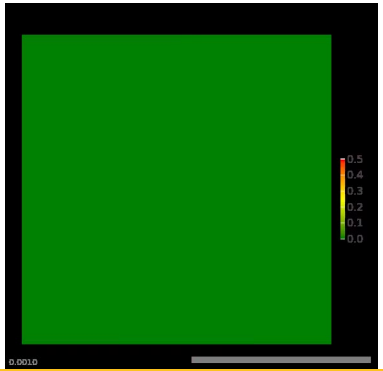


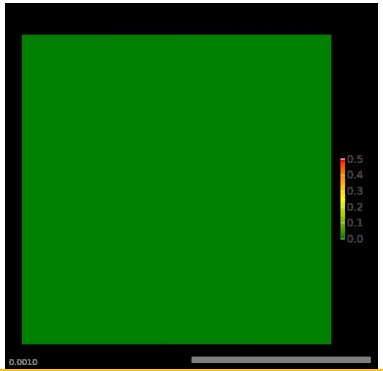
Fluid mixing



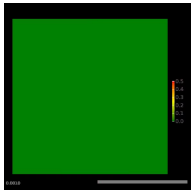
- We want to find how time/space behavior of C concentrations is controlled by the simulated physics processes
- > 2000 simulations of C concentrations in time/space for a series of model parameters impacting fluid mixing; 3 example predictions:



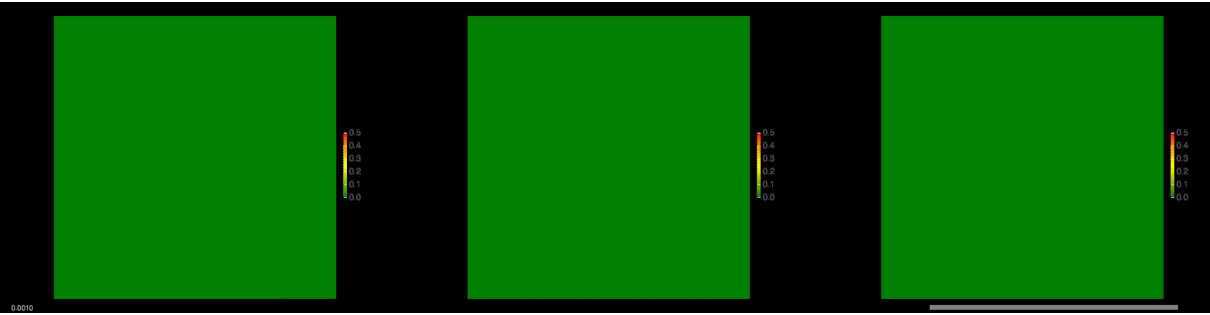




NTFk results



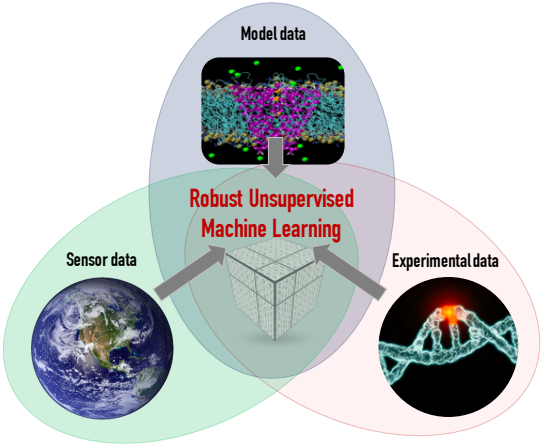
- > 200 GB simulation data compressed to ≈ 70 MB (compression $\approx 4 \times 10^{-4}$)
  Here, $(1000 \times 81 \times 81) \rightarrow (3 \times 8 \times 9)$
- NTFk processed all the data and extracted the dominant time/space features (**processes** / **vortices**)



Advection **Dispersion** **Diffusion**

Summary

- We have developed a series of novel unsupervised ML methods based on Nonnegative Factorization (Matrices/Tensors)
- These ML methods have been used to solve various real-world problems
- Some of our ML analyses brought breakthrough discoveries (especially related to human cancer research)
- We have developed a series of ML computational tools for solving big-data problems using high-performance computing (HPC)



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

- NMFk + ShiftNMFk + GreenNMFk **(patent)**
- NTFk (prototype; work in progress)
- NBMF: Quantum machine learning using **D-Wave**
- SVR/SVM: Support Vector Regression/Machine
  `http://github.com/madsjulia/SVR.jl`
- MADS: Model-Analyses & Decision Support
  open-source, version-controlled, high-performance computational framework
  `http://mads.lanl.gov` `http://madsjulia.github.io/Mads.jl`

`http://madsjulia.github.io/Mads.jl/Examples/blind_source_separation`