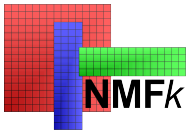
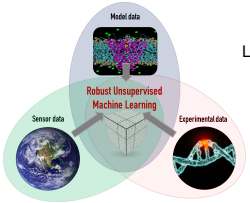
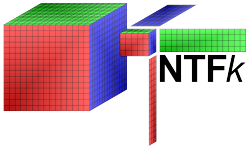
Novel Unsupervised Machine Learning Methods for Data Analytics and Model Diagnostics

Velimir V. Vesselinov (monty) (vvv@lanl.gov)
Earth and Environmental Sciences Division, Los Alamos National Laboratory, NM, USA
http://tensors.lanl.gov





LA-UR-19-22579



Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires “labeling” (prior categorization (knowledge) about the processed data)

  Example: Recognizes images of cats and dogs after extensive training; but cannot recognize horses if not trained

  Cannot discover something that we do not know already

- **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis**; **data-driven science**)

  Example: Identifies features that distinguish images of animals (e.g., cats, dogs, horses, etc.)

Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires “labeling” (prior categorization (knowledge) about the processed data)

  Example: Recognizes images of cats and dogs after extensive training; but cannot recognize horses if not trained

  Cannot discover something that we do not know already

- **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis**; **data-driven science**)

  Example: Identifies features that distinguish images of animals (e.g., cats, dogs, horses, etc.)

Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires “labeling” (prior categorization (knowledge) about the processed data)

  Example: Recognizes images of cats and dogs after extensive training; but cannot recognize horses if not trained

  Cannot discover something that we do not know already

- **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis**; **data-driven science**)

  Example: Identifies features that distinguish images of animals (e.g., cats, dogs, horses, etc.)

Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires “labeling” (prior categorization (knowledge) about the processed data)

  Example: Recognizes images of cats and dogs after extensive training; but cannot recognize horses if not trained

  Cannot discover something that we do not know already

- **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis**; **data-driven science**)

  Example: Identifies features that distinguish images of animals (e.g., cats, dogs, horses, etc.)

Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires “labeling” (prior categorization (knowledge) about the processed data)

  Example: Recognizes images of cats and dogs after extensive training; but cannot recognize horses if not trained

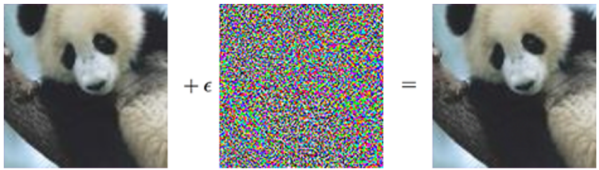
  Cannot discover something that we do not know already

- **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis**; **data-driven science**)

  Example: Identifies features that distinguish images of animals (e.g., cats, dogs, horses, etc.)

Why not supervised Machine Learning (ML)

- **Supervised** ML
  - introduces subjectivity (through the labeling process)
  - does not provide insights why horses are different from dogs / cats
  - cannot make predictions (that we do not know already)
  - requires huge training (labeled) datasets
  - we do not know why it works
  - is impacted by “adversarial examples”



$\Rightarrow$ major limitations of the **supervised** methods for **data-driven science** applications

- **Data Analytics**: Identify signals (features) in datasets
- **Model Diagnostics**: Identify processes (features) in model outputs
- **Physics-Informed Machine Learning**: Coupled Data/Model Analytics

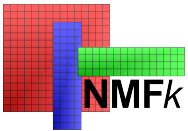
- **Data Analytics**: Identify signals (features) in datasets
  - Feature extraction (**FE**)
  - Blind source separation (**BSS**)
  - Detection of disruptions / anomalies
  - Image recognition
  - Discover unknown dependencies and phenomena
  - Develop reduced-order/surrogate models
  - Guide development of physics models representing the data
  - Make predictions
  - Optimize data acquisition
  - “Label” datasets for supervised ML analyses

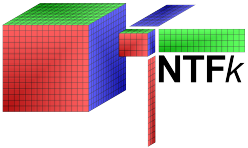
- **Model Analytics/Diagnostics**: Identify processes (features) in model outputs
  - Separate processes (inseparable during modeling)
  - Model reduction
  - Identify dependencies between model inputs and outputs
  - Discover unknown dependencies and phenomena
  - Make predictions

- **Physics-Informed Machine Learning**: Coupled Data/Model Analytics

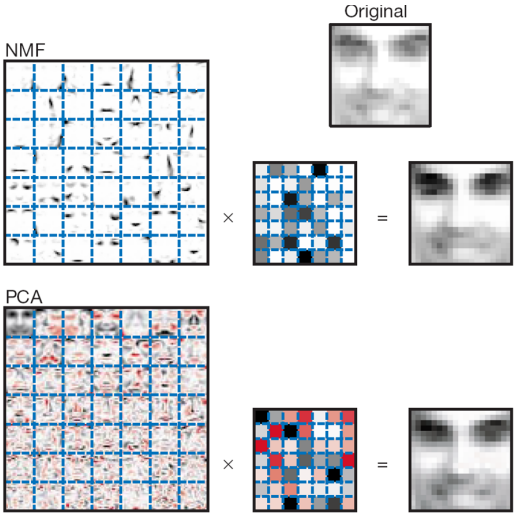
  Simultaneous ML analyses of data (observations/experiments) and model outputs with physics constraints

- We have developed a series of novel unsupervised Machine Learning (ML) methods and computational techniques
- Our methods are based in matrix/tensor factorization coupled with custom k-means clustering and nonnegativity/sparsity constraints:
  - NMFk: Nonnegative **Matrix** Factorization
  - NTFk: Nonnegative **Tensor** Factorization
- NMFk/NTFk are capable to efficiently process large datasets (GB/TB's) utilizing GPU's & TPU's (TensorFlow, PyTorch, MXNet)





Why nonnegativity?



Nonnegativity constraints provide meaningful and interpretable results (+sparsity)

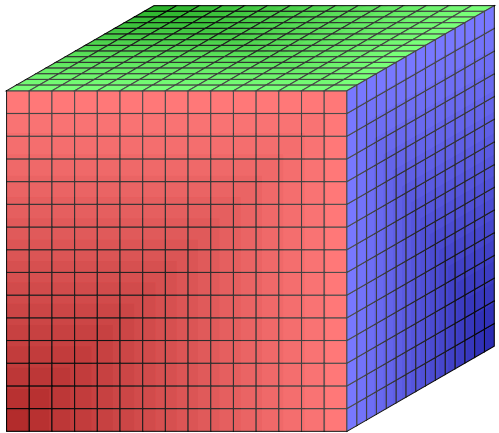
Why tensors?

- **Tensors** (multi-dimensional/multi-modal/multi-way datasets) are everywhere:
  - observational data are typically a 5-D tensor (x, y, z, t, attributes)
  - model outputs are typically a 5-D tensor (x, y, z, t, attributes)
  - data dependency to N parameters will form a $(N + 5)$-D tensor

Why not SVD (PCA)?

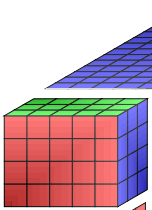
- **SVD** cannot be applied on multi-dimensional (tensor) datasets
- **HOSVD** cannot tell us the number of features (signals)

Tucker Tensor Factorization (3D case): Tucker-3

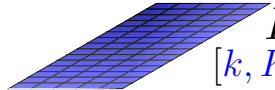


X
$[K, M, N]$

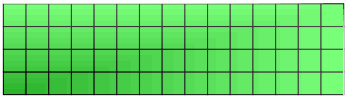
$\approx$



G
$[k, m, n]$



H
$[k, K]$

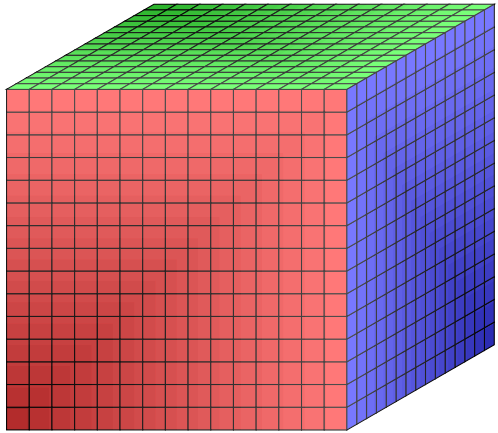


W
$[n, N]$

V
$[m, M]$

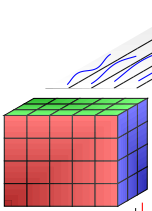
$$X \approx G \otimes H \otimes W \otimes V$$

Tucker Tensor Decomposition: Feature extraction



X
$[K, M, N]$

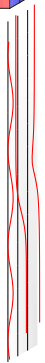
$\approx$



G
$[k, m, n]$

H
$[k, K]$

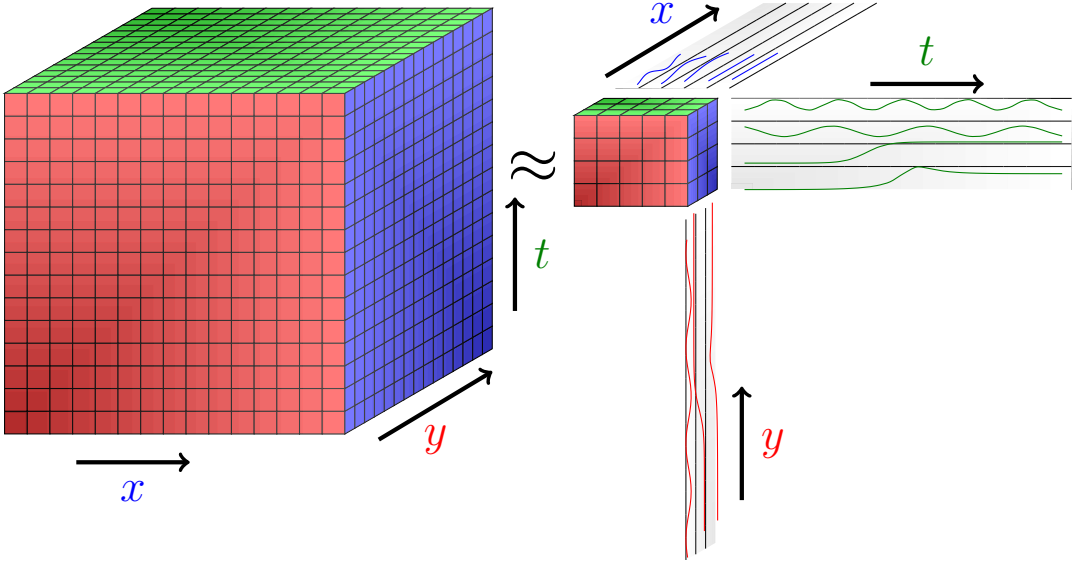
W
$[n, N]$



V
$[m, M]$

$$X \approx G \otimes H \otimes W \otimes V$$

Tucker Tensor Decomposition: Feature extraction



- **Identifying the number of unknown features**:
  - resolved using custom k-means clustering and sparsity constraints on the core tensor
  - number of features identified based on the reconstruction quality (e.g., Frobenius norm) and cluster Silhouettes
- **Solving a non-unique optimization problem**:
  - addressed through multistarts, regularization and nonnegativity constraints
  - applying diverse optimization techniques (Multiplicative/Alternating Least Squares algorithms, NLopt, Ipopt, Gurobi, MOSEK, GLPK, Clp, Cbc, ...)
- **Processing Big Data**:
  - GPU's / TPU's / Distributed computing
  - Account for data sparsity and structure
  - Nonnegative Tensor Trains
- **Dealing with Noisy Data**:
  - Random noise impacts accuracy but it is accountable
  - Systematic noise is identified as separate signals

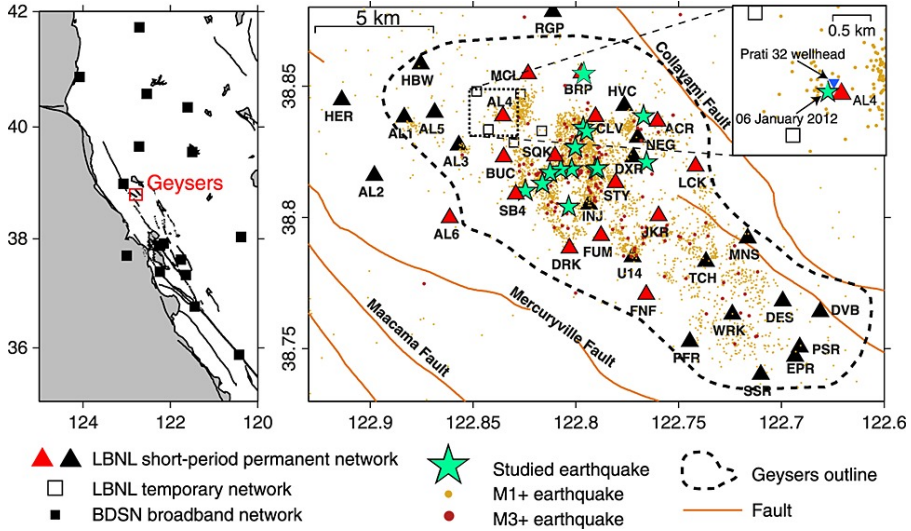
4GB Tensor ($1000 \times 1000 \times 1000$)

Framework	Execution time (seconds)
MATLAB	2634
NumPy	881
MXNet	644
PyTorch	121
TensorFlow	119
Julia	109

- **Field Data**:
  - Groundwater contamination
  - US Climate data
  - Geothermal data
  - Seismic data
- **Lab Data**:
  - X-ray Spectroscopy
  - UV Fluorescence Spectroscopy
  - Microbial population analyses
- **Operational Data**:
  - LANSCE: Los Alamos Neutron Accelerator
  - Oil/gas production
- **Model Outputs**:
  - Reactive mixing $A + B \rightarrow C$
  - Phase separation of co-polymers
  - Molecular Dynamics of proteins
  - EU Climate modeling (Helmholtz Institute, Germany)

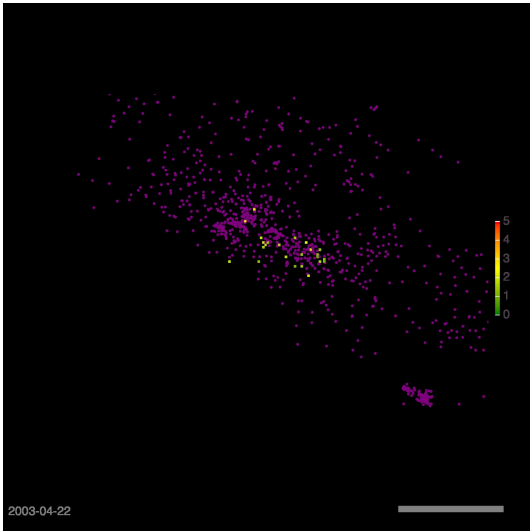
- Vesselinov, Munuduru, Karra, O'Maley, Alexandrov, Unsupervised Machine Learning Based on Non-Negative Tensor Factorization for Analyzing Reactive-Mixing, **Journal of Computational Physics**, (in review), 2019.
- Stanev, Vesselinov, Kusne, Antoszewski, Takeuchi, Alexandrov, Unsupervised Phase Mapping of X-ray Diffraction Data by Nonnegative Matrix Factorization Integrated with Custom Clustering, **Nature Computational Materials**, 2018.
- Vesselinov, O'Malley, Alexandrov, Nonnegative Tensor Factorization for Contaminant Source Identification, **Journal of Contaminant Hydrology**, 2018.
- O'Malley, Vesselinov, Alexandrov, Alexandrov, Nonnegative/binary matrix factorization with a D-Wave quantum annealer, **PLOS ONE**, 2018.
- 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, **WRR**, 10.1002/2013WR015037, 2014.

Geysers Geothermal Field

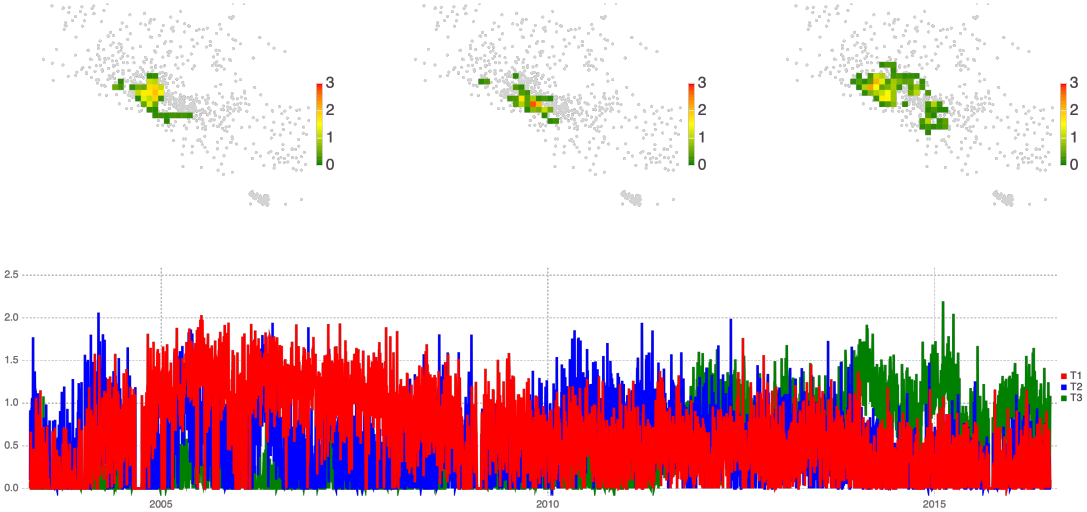


Geysers seismicity

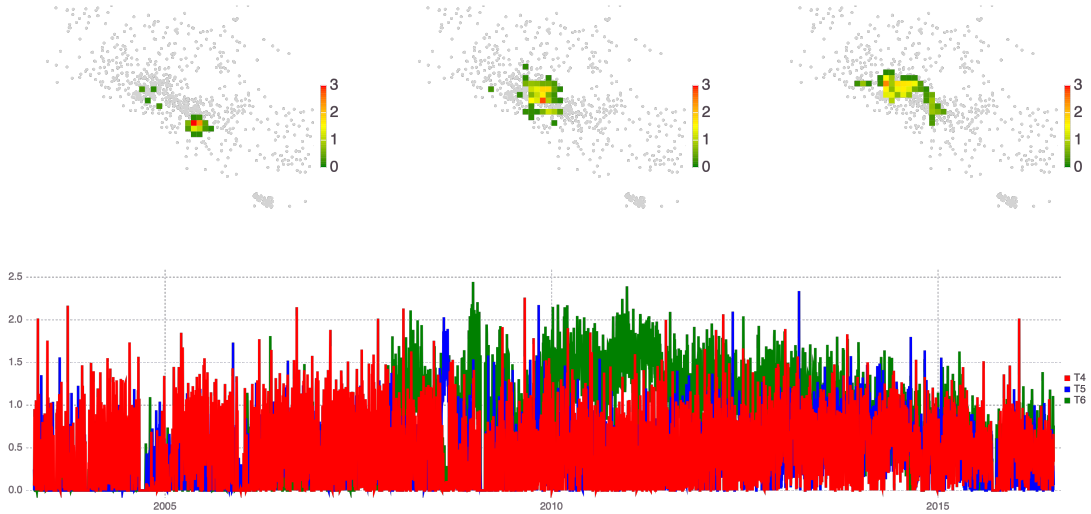
- 470,263 seismic events have been identified between 2003 and 2016
- Tensor: total energy of events over a discretized domain
- **NTF**k extracts spatial footprints and temporal patterns of dominant hidden (latent) features related to:
  - Total water injection
  - EGS injection (starting November 6th, 2011)
  - ...



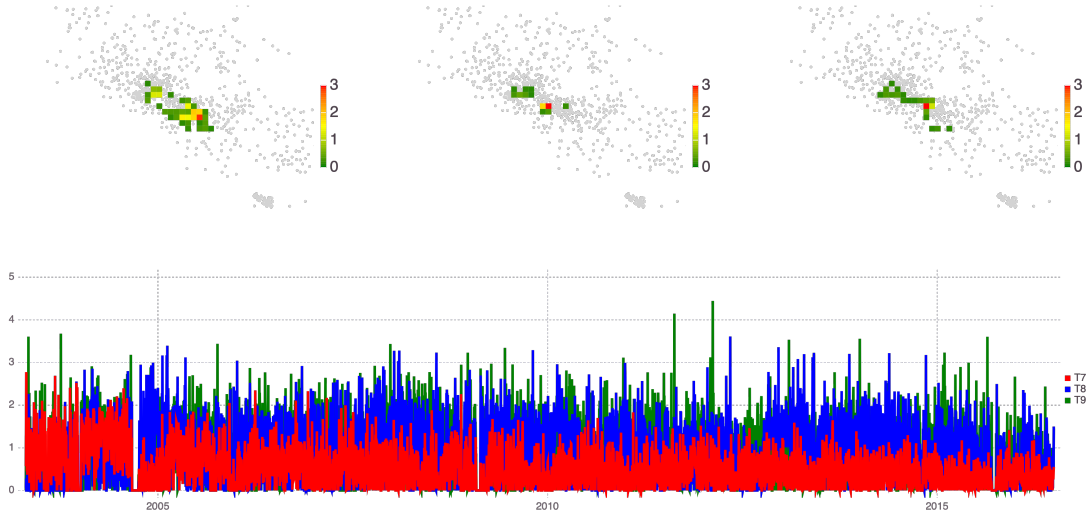
Geysers seismicity: NTFk results



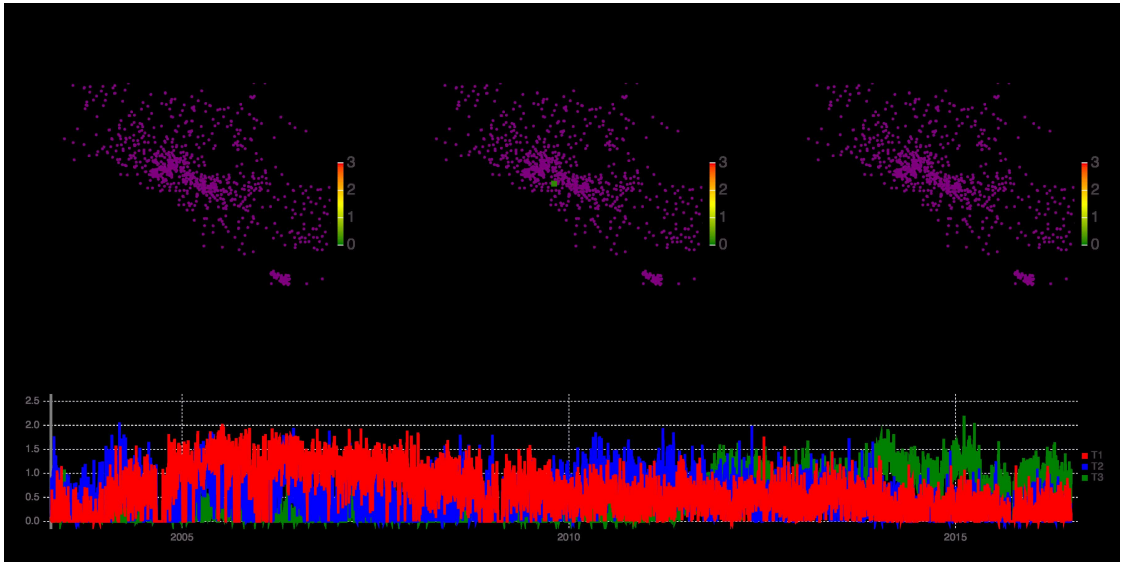
Geysers seismicity: NTFk results



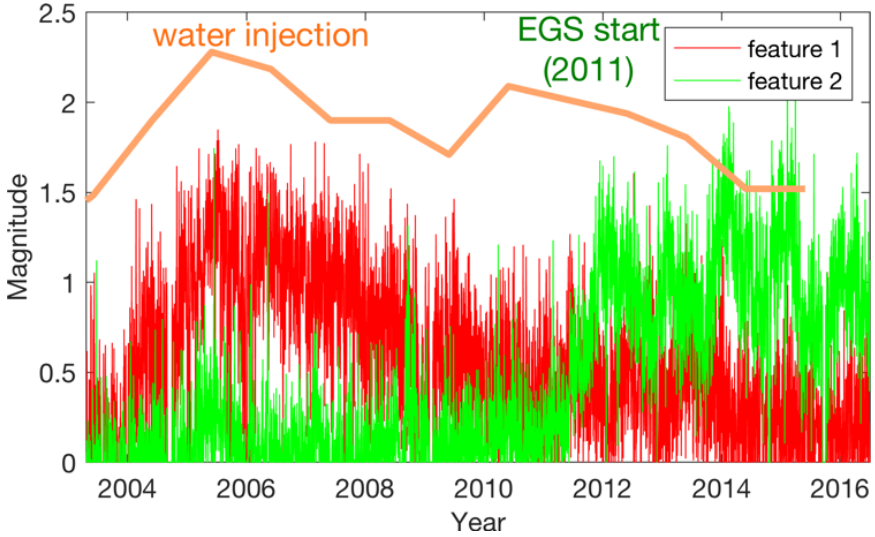
Geysers seismicity: NTFk results



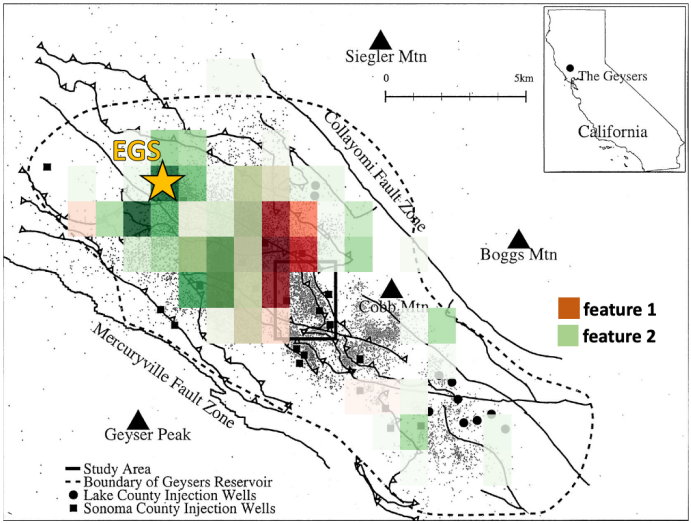
Geysers seismicity: NTFk results



Geysers seismicity: NTFk results

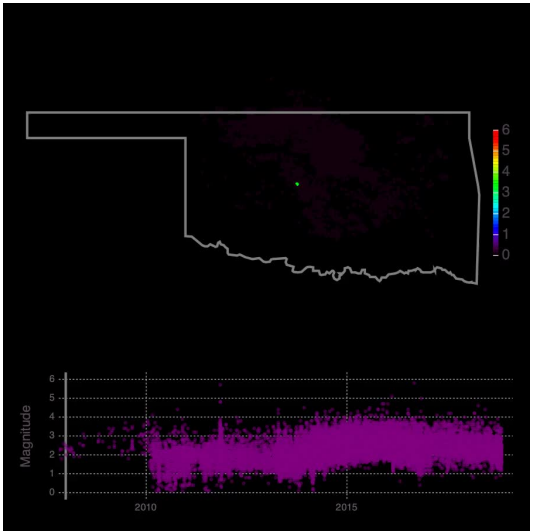


Geysers seismicity: NTFk results

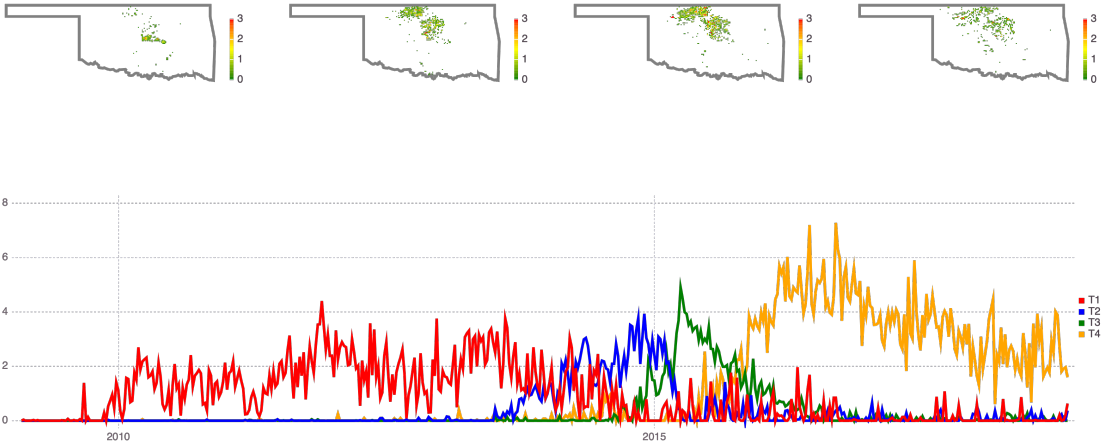


Oklahoma seismicity

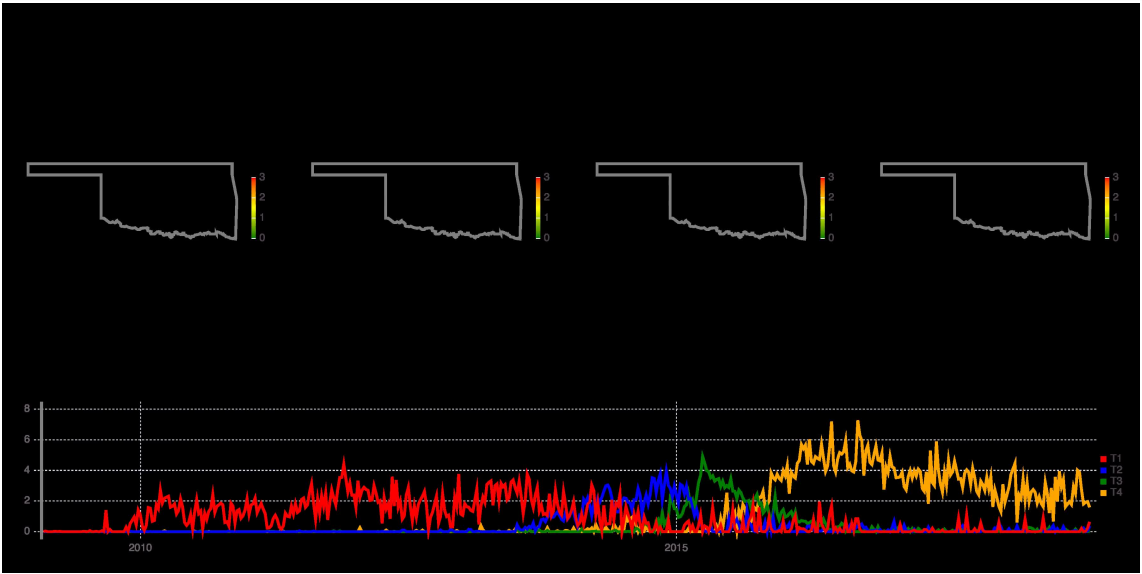
- 32,251 seismic events from 1989 to 2017
- Tensor: total energy of events over a discretized domain
- **NTF**k extracts spatial footprints and temporal patterns of dominant hidden (latent) features



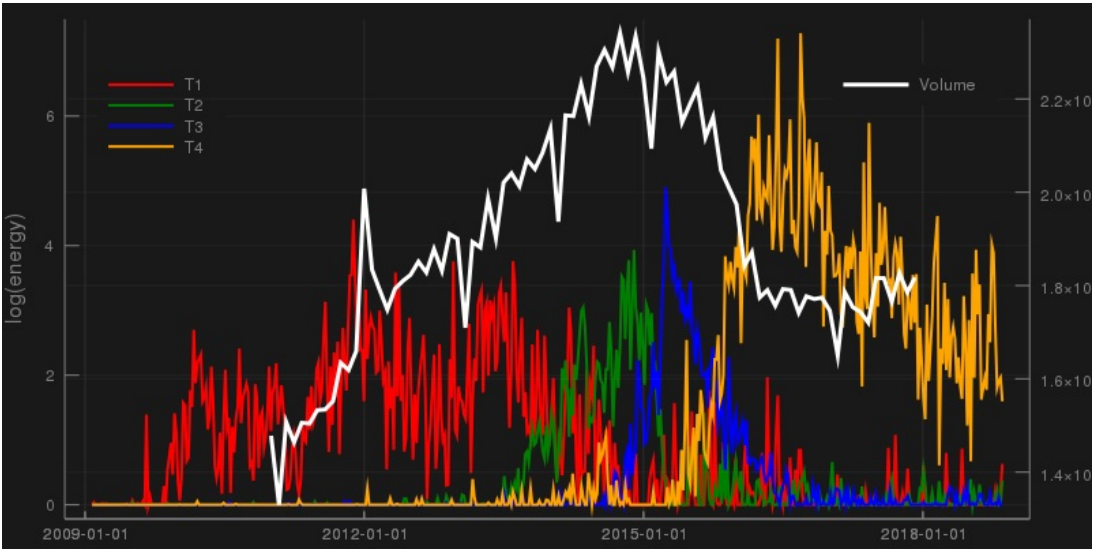
Oklahoma seismicity: reconstruction by 4 features (signals)



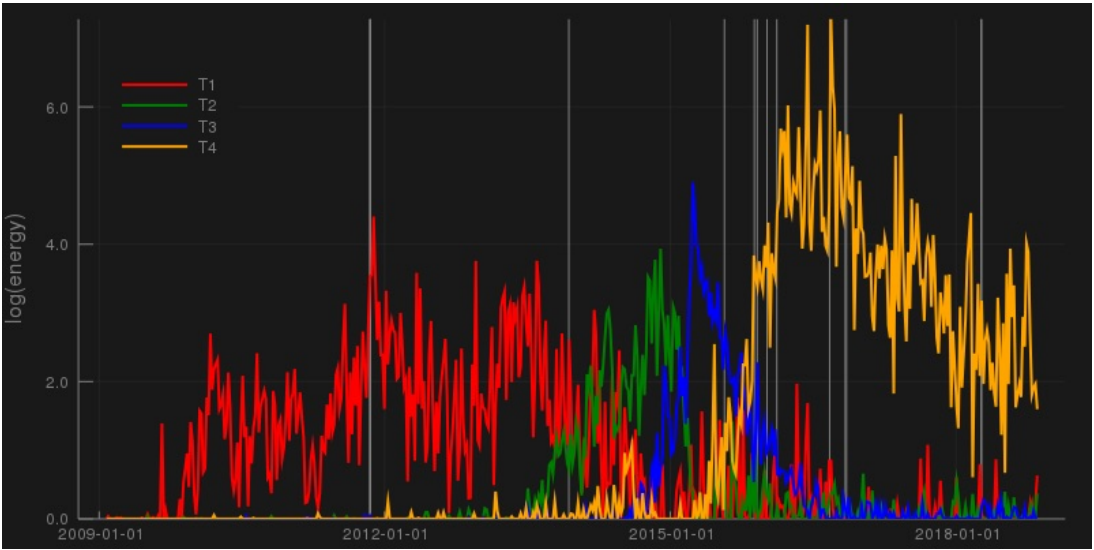
Oklahoma seismicity: reconstruction by 4 features (signals)



Oklahoma seismicity: extracted signals vs. injected volumes

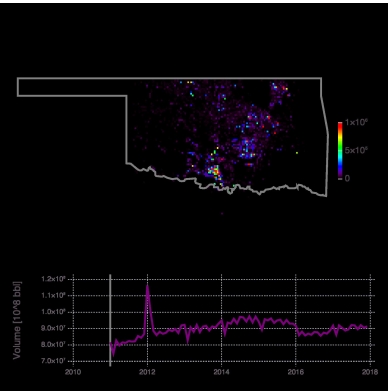


Oklahoma seismicity: extracted signals vs. majors seismic events

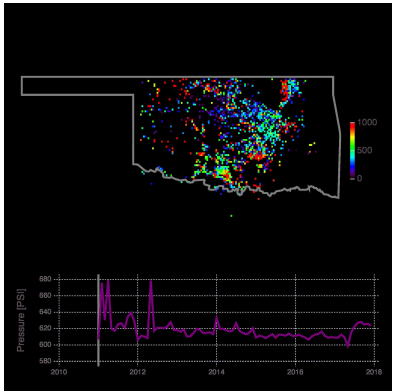


Oklahoma seismicity

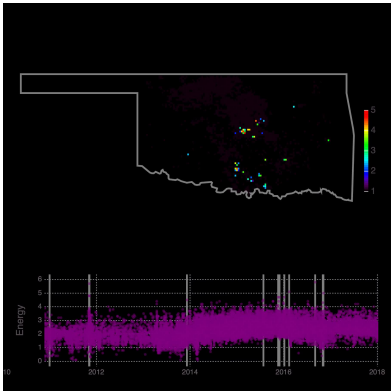
volume



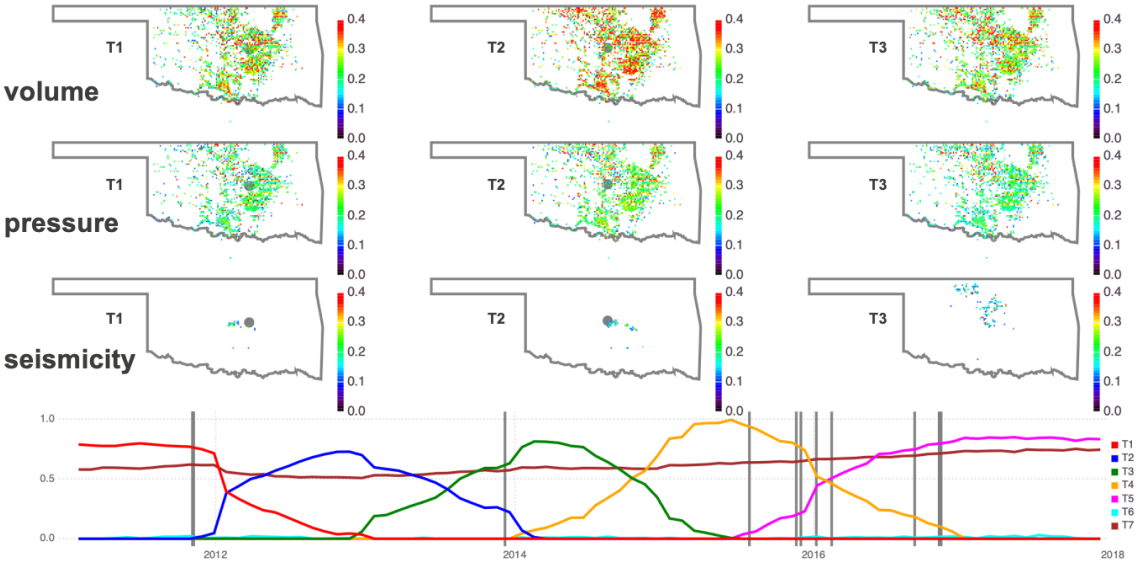
pressure



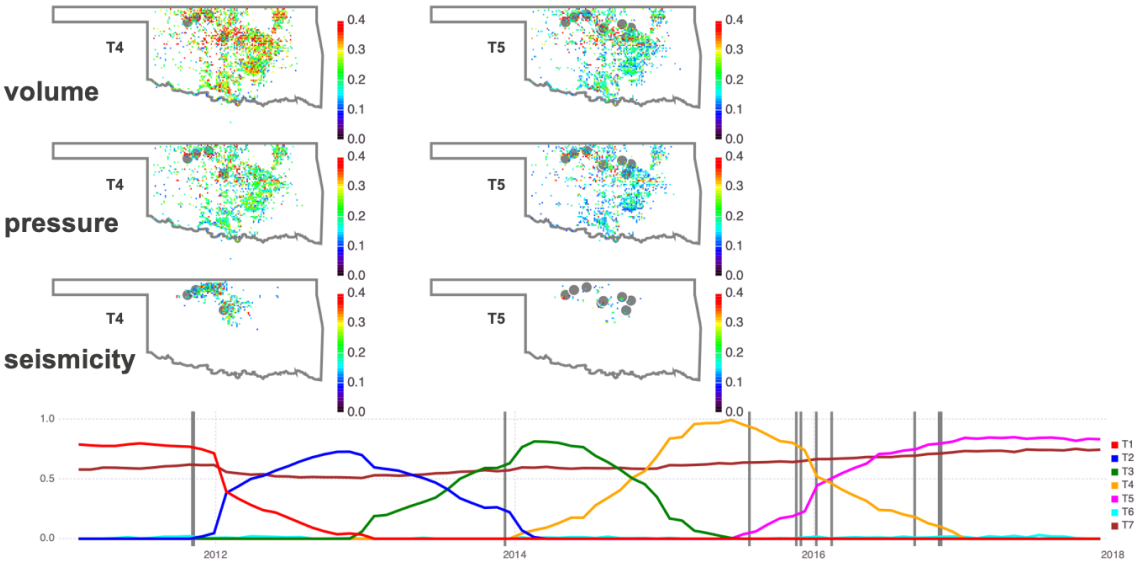
seismicity



Oklahoma seismicity: 5 volume/pressure/seismicity features

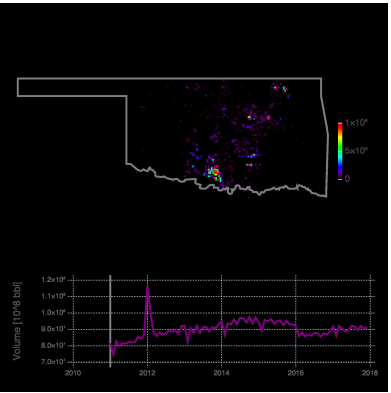


Oklahoma seismicity: 5 volume/pressure/seismicity features

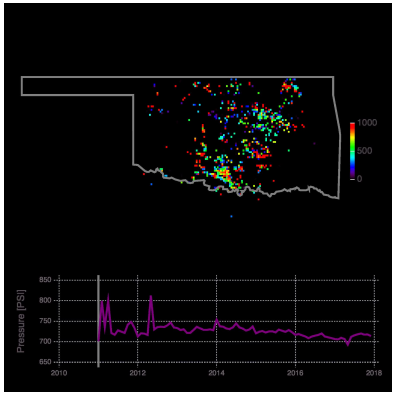


Oklahoma seismicity: Recovery injection

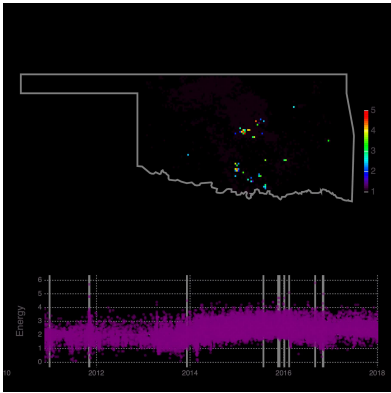
volume recovery



pressure recovery



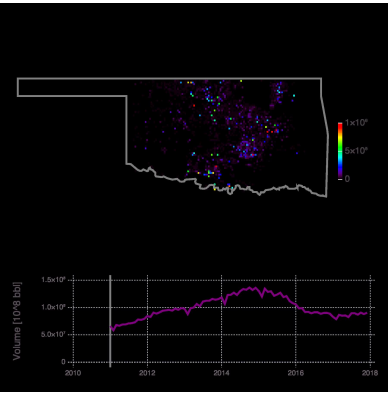
seismicity



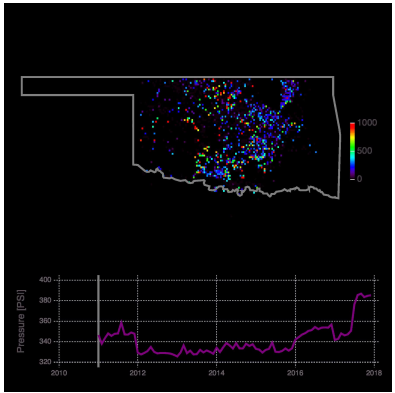
Recovery injection has limited impact on seismicity

Oklahoma seismicity: Disposal injection

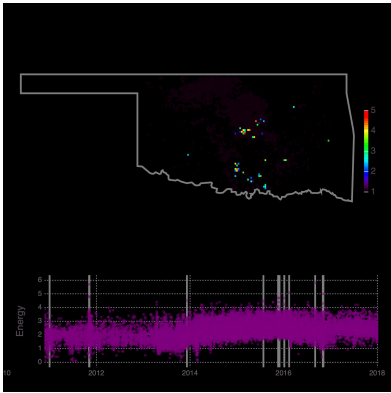
volume disposal



pressure disposal



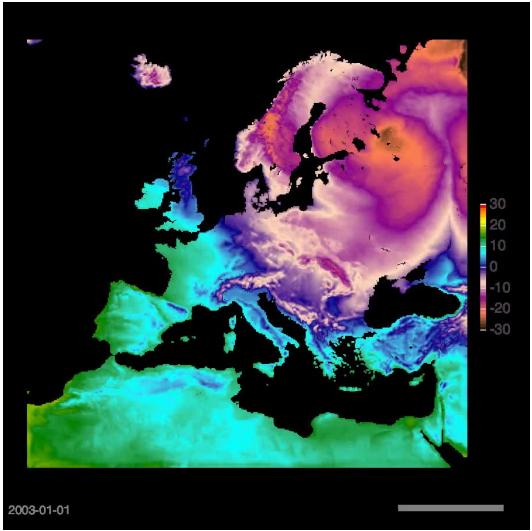
seismicity



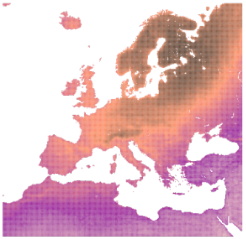
Disposal injection has impact on seismicity

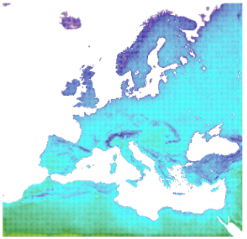
Climate model of Europe: air temperature

- fluctuations in the air temperature [$^{\circ}C$]
- **NTF**k extracts spatial footprints and temporal patterns of dominant hidden (latent) features related to :
  - storm signal
  - winter seasonal signal
  - summer seasonal signal
  - ....



Climate model of Europe: air temperature fluctuations represented by 3 signals

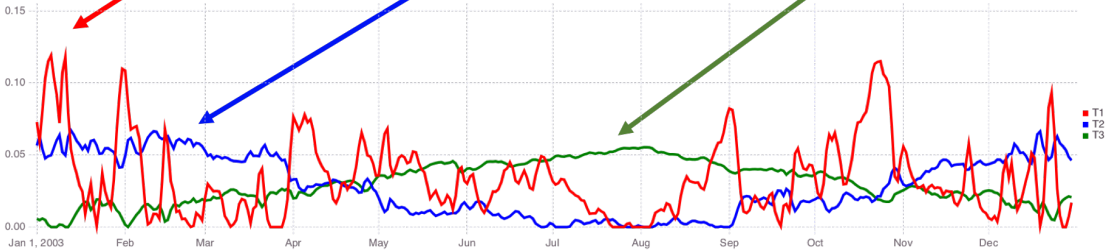




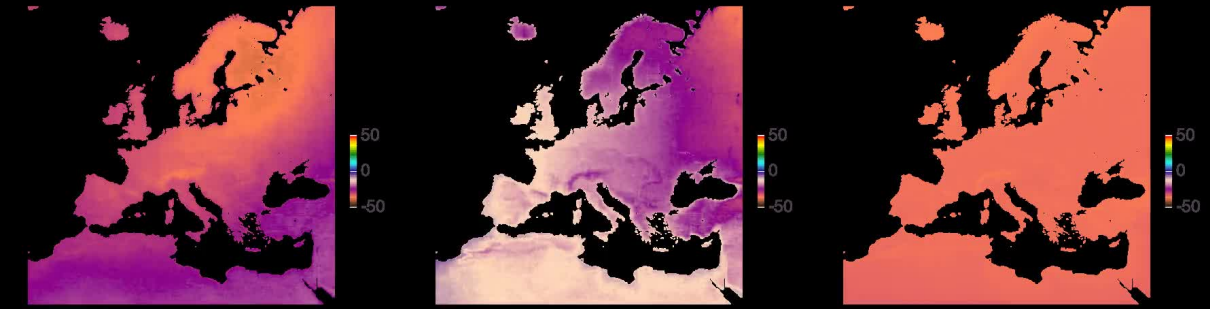
Storm signal

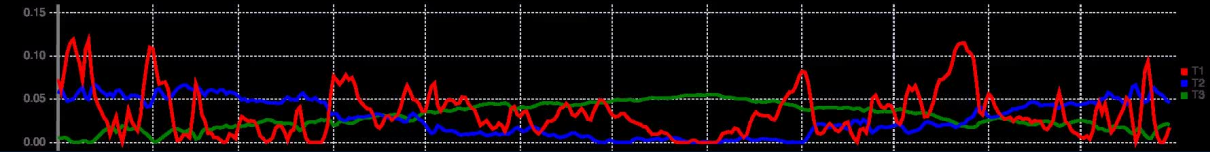
Winter signal

Summer signal



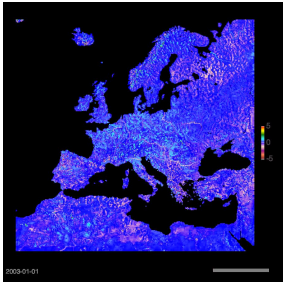
Climate model of Europe: air temperature fluctuations represented by 3 signals



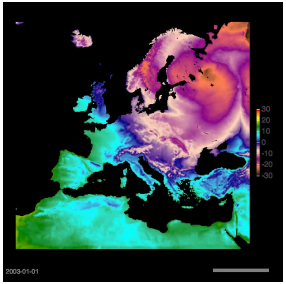


Climate model of Europe: analyze all model outputs (>40) simultaneously

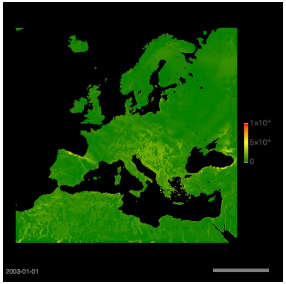
Water-table depth



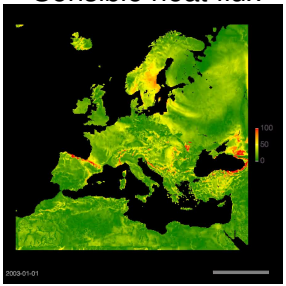
Temperature



Evaporation



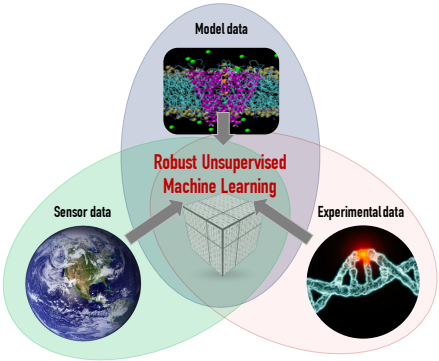
Sensible heat flux



- Find interconnections among model outputs
- Evaluate impacts of different model setups
- Find dominant processes impacting model predictions
  (e.g., occurrence of heat waves, climate impacts on groundwater resources, impacts of subsurface processes on atmospheric conditions)

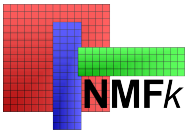
Summary

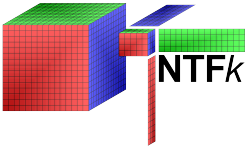
- Developed **novel** unsupervised ML methods and computational tools based on Nonnegative Factorization (Matrices/Tensors)
- Our ML methods have been used to solve various real-world problems (brought breakthrough discoveries related to human cancer research)



- Codes:

NMFk MADS NTFk





- Examples:

`http://madsjulia.github.io/Mads.jl/Examples/blind_source_separation`
`http://tensors.lanl.gov`
`http://tensordecompositions.github.io`