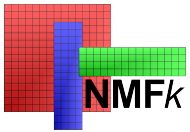
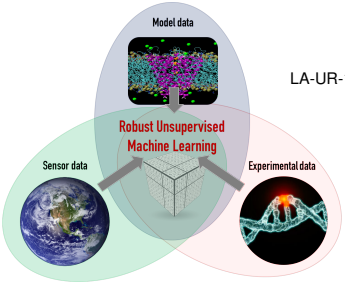
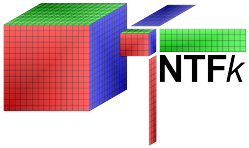
Unsupervised Machine Learning Methods for Feature Extraction

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





LA-UR-19-21450



Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires prior categorization (knowledge) of the processed data **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

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

  Cannot find something that we do not already know

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

  Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

- **AI + Unsupervised** ML: ... coupled AI and unsupervised techniques ... currently pursued by many

Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires prior categorization (knowledge) of the processed data **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

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

  Cannot find something that we do not already know

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

  Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

- **AI + Unsupervised** ML: ... coupled AI and unsupervised techniques ... currently pursued by many

Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires prior categorization (knowledge) of the processed data **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

  Example: Recognize images of cats and dogs after extensive training; but cannot recognize horses if not trained
  Cannot find something that we do not already know

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

  Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

- **AI + Unsupervised** ML: ... coupled AI and unsupervised techniques ... currently pursued by many

Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires prior categorization (knowledge) of the processed data **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

  Example: Recognize images of cats and dogs after extensive training; but cannot recognize horses if not trained
  Cannot find something that we do not already know

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

  Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

- **AI + Unsupervised** ML: ... coupled AI and unsupervised techniques ... currently pursued by many

Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires prior categorization (knowledge) of the processed data **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

  Example: Recognize images of cats and dogs after extensive training; but cannot recognize horses if not trained
  Cannot find something that we do not already know

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

  Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

- **AI + Unsupervised** ML: ... coupled AI and unsupervised techniques ... currently pursued by many

Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires prior categorization (knowledge) of the processed data **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

  Example: Recognize images of cats and dogs after extensive training; but cannot recognize horses if not trained
  Cannot find something that we do not already know

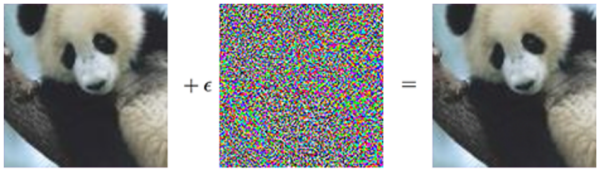
- **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis** for **data-driven science**)

  Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

- **AI + Unsupervised** ML: ... coupled AI and unsupervised techniques ... currently pursued by many

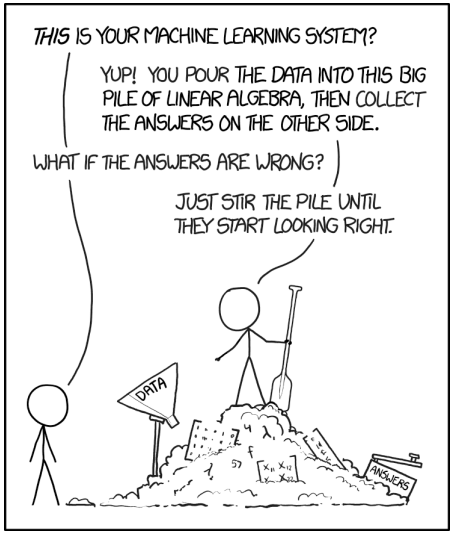
Why not supervised Machine Learning (ML)

- **Supervised** ML
  - can introduce subjectivity (through the labeling process)
  - does not provide insights why horses are different than dogs / cats
  - cannot make predictions
  - requires huge training (labeled) datasets
  - is impacted by “adversarial examples”



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

Supervised Machine Learning (xkcd)



Unsupervised Machine Learning Applications

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

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

Unsupervised Machine Learning Applications

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

Unsupervised Machine Learning Applications

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

Unsupervised Machine Learning Applications

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

Unsupervised Machine Learning Applications

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

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

Unsupervised Machine Learning Applications

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

Unsupervised Machine Learning Applications

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

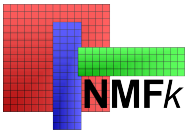
Unsupervised Machine Learning Applications

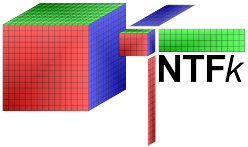
- **Model Analytics/Diagnostics**: Identify processes (features) in model outputs
  - Separate processes (inseparable during modeling)
  - Model reduction (develop reduced-order/surrogate models)
  - Identify dependencies between model inputs and outputs
  - Discover unknown dependencies and phenomena
  - Make predictions
  - Optimize data acquisition

- **Physics-Informed Machine Learning**: Coupled Data/Model Analytics (fusion)
  Simultaneous analyses of data and model outputs

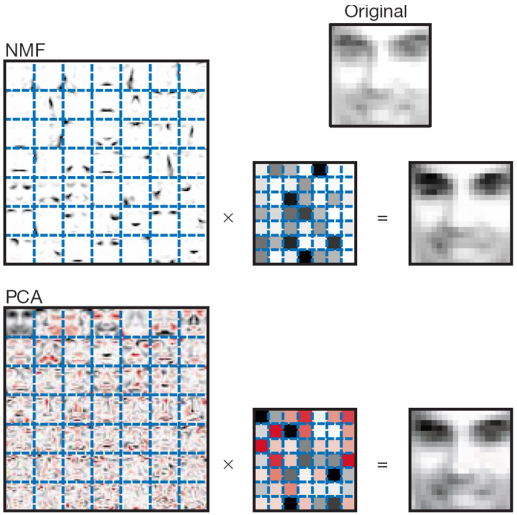
Nonnegative Matrix/Tensor Factorization

- 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)

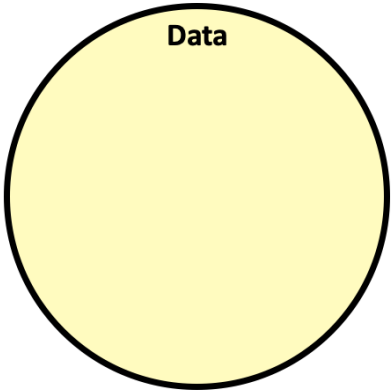
Why tensors?

- **Tensors** (multi-dimensional/multi-modal/multi-way datasets) are everywhere:
  - monitoring 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)

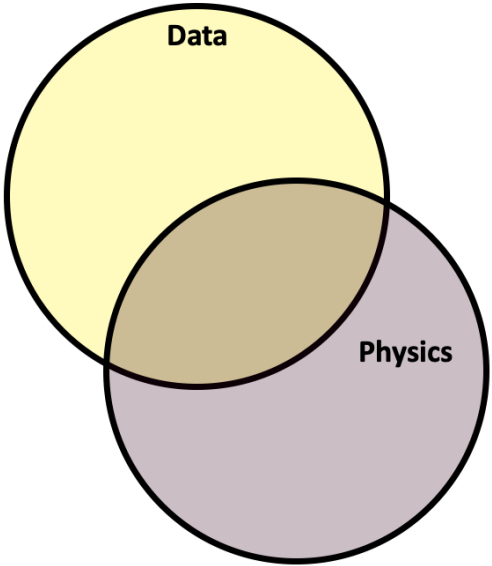
Why tensors-based unsupervised machine learning?

- **SVD** cannot be applied on multi-way (tensor datasets)
- **Tensor**-based unsupervised machine learning can be applied to
  - resolve interactions between dimensions (modes)
    - identify how controls (e.g., pressure, temperature, volume, etc.) are impacting outcomes (e.g. oil production)
    - identify how observables (e.g., pressure, temperature, volume, etc.) are impacting other observables (e.g. concentration)
    - identify how model inputs (e.g., conductivity, storage, etc.) are impacting model outputs (e.g. pressure)
  - simultaneous analyses of data and model outputs (data/model fusion; PIML)

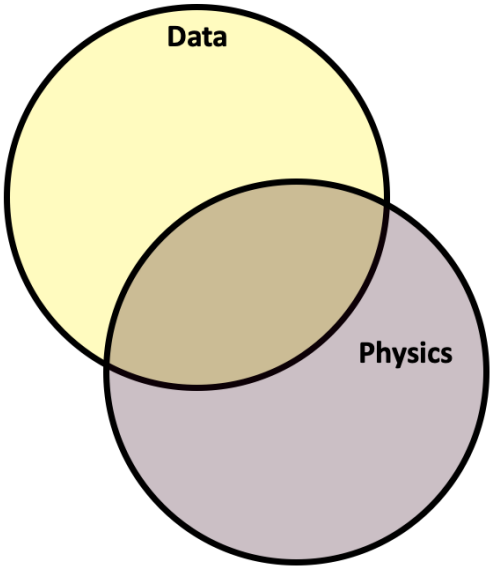
Physics-Informed Machine Learning



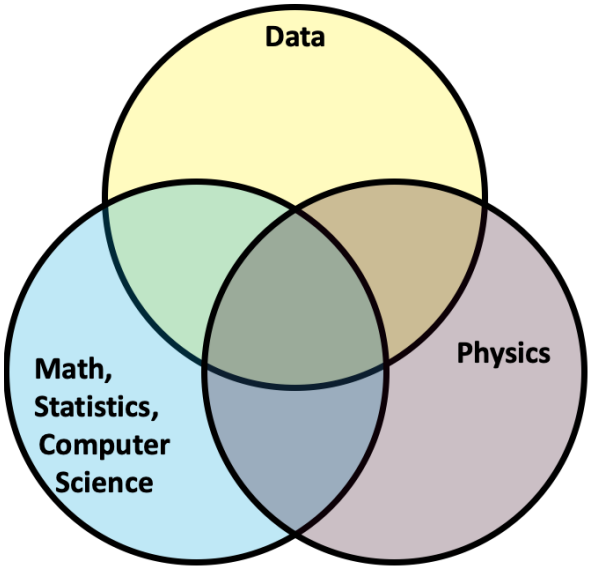
Physics-Informed Machine Learning



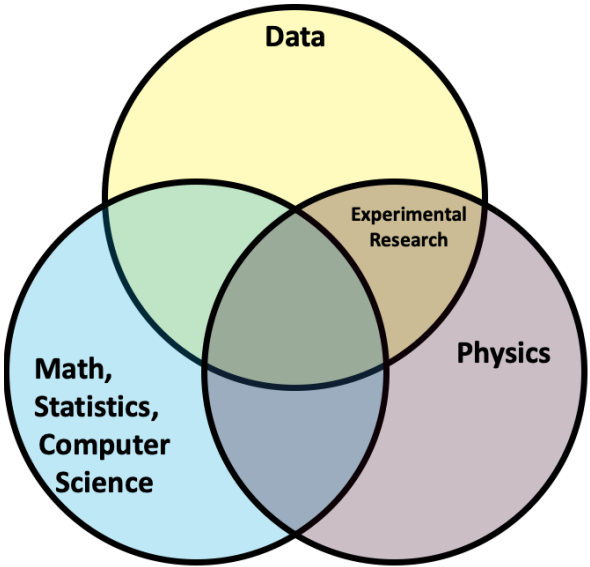
Physics-Informed Machine Learning



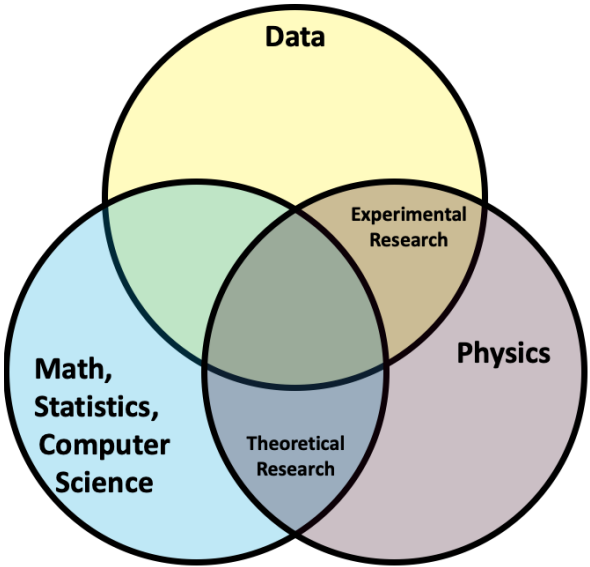
Physics-Informed Machine Learning



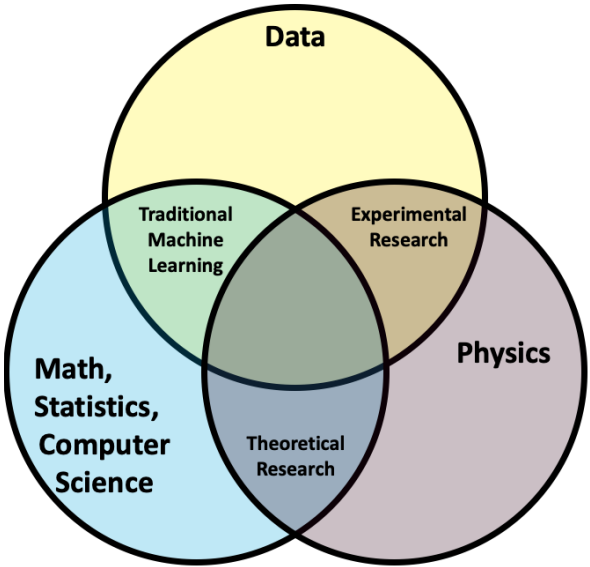
Physics-Informed Machine Learning



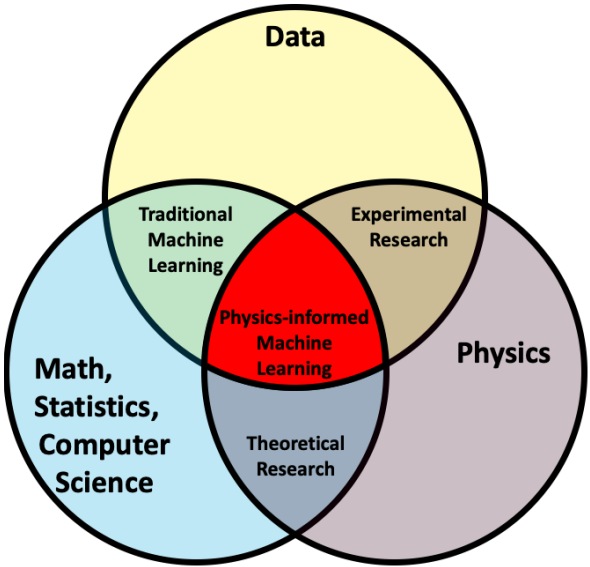
Physics-Informed Machine Learning



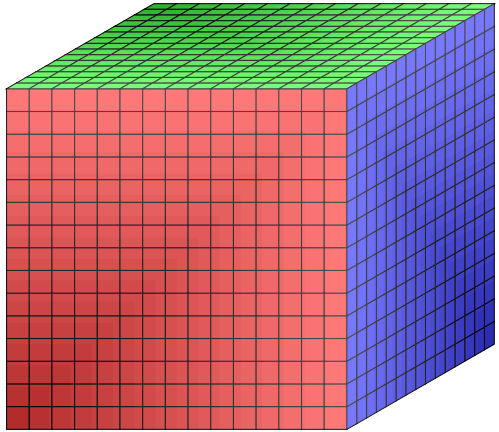
Physics-Informed Machine Learning



Physics-Informed Machine Learning

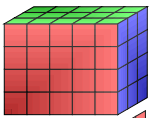


Tucker Tensor Factorization (3D case): Tucker-3

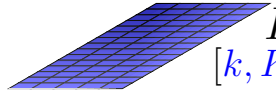


X
$[K, M, N]$

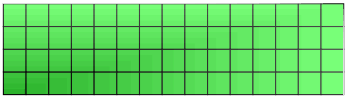
$\approx$



G
$[k, m, n]$



H
$[k, K]$

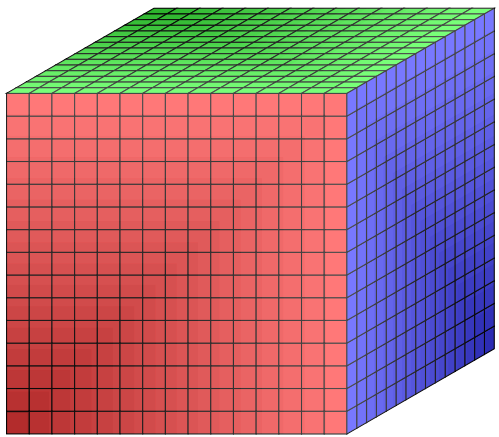


W
$[n, N]$

V
$[m, M]$

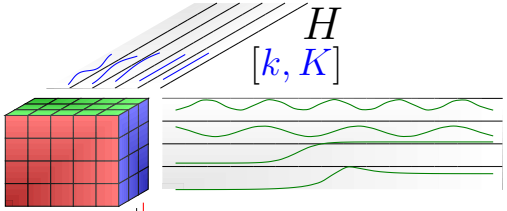
$$X \approx G \otimes H \otimes W \otimes V$$

Tucker Tensor Decomposition: Feature extraction



$$X \\ [K, M, N]$$

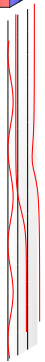
$\approx$



$$H \\ [k, K]$$

$$W \\ [n, N]$$

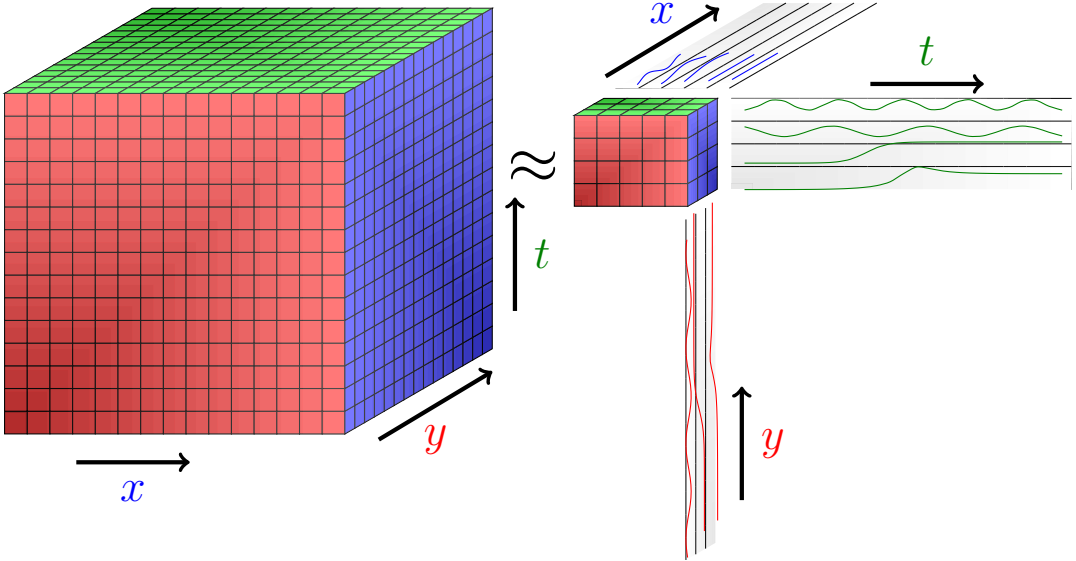
$$G \\ [k, m, n]$$



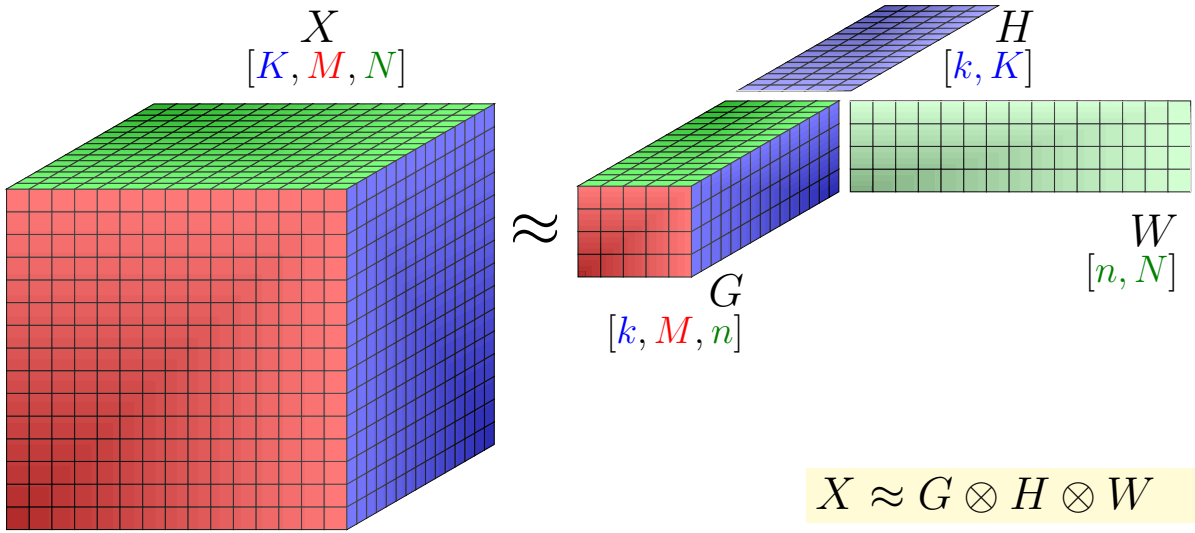
$$V \\ [m, M]$$

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

Tucker Tensor Decomposition: Feature extraction

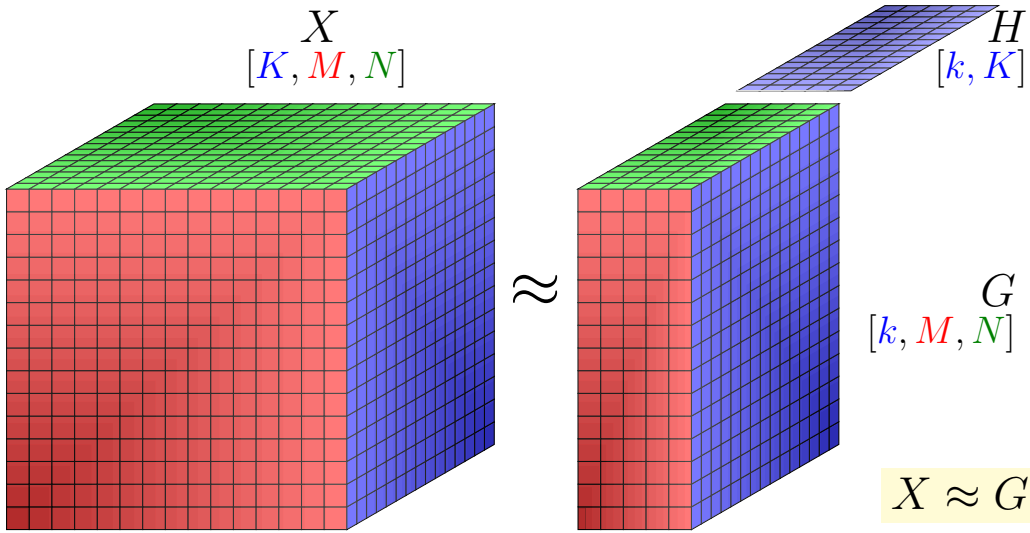


Tucker Tensor Decomposition: Tucker-2 (three possible alternatives)



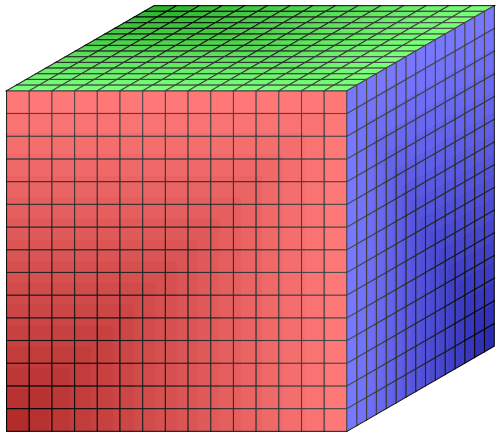
$$X \approx G \otimes H \otimes W$$

Tucker Tensor Decomposition: Tucker-1 (three possible alternatives)



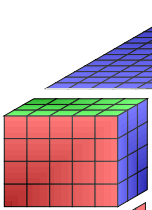
$$X \approx G \otimes H$$

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

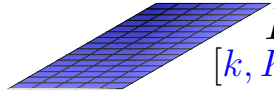


X
$[K, M, N]$

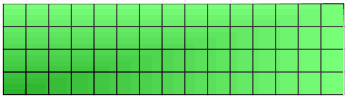
$\approx$



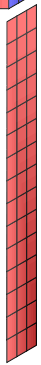
G
$[k, m, n]$



H
$[k, K]$



W
$[n, N]$

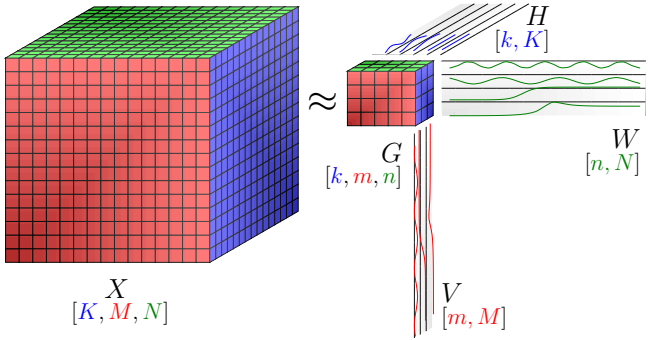


V
$[m, M]$

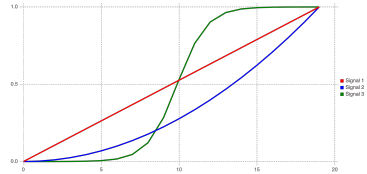
$$X \approx G \otimes H \otimes W \otimes V$$

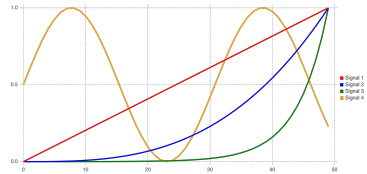
Tucker Tensor Decomposition: Feature extraction

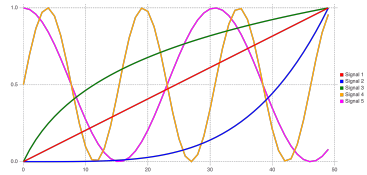
- Tucker decomposition is achieved through minimization
- Nonnegativity and sparsity constraints help the feature extraction
- Optimal number of features $[k, m, n]$ is estimated through k-means clustering of a series minimization solutions with random initial guesses



Tucker Tensor Decomposition: Example



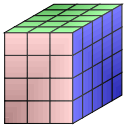




$$V = \begin{bmatrix} t \\ t^2 \\ tanh(t-10)+1 \end{bmatrix}$$

$$W = \begin{bmatrix} x \\ x^3 \\ e^x \\ sin(x)+1 \end{bmatrix}$$

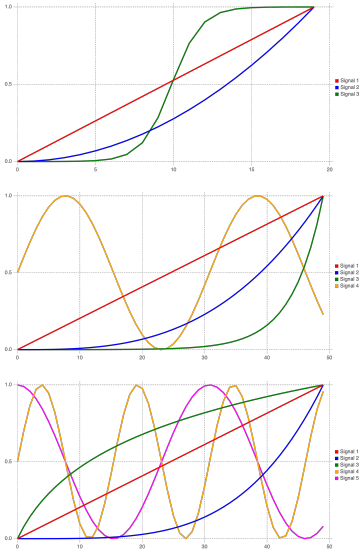
$$H = \begin{bmatrix} y \\ y^4 \\ ln(y) \\ sin(2y)+1 \\ cos(y)+1 \end{bmatrix}$$

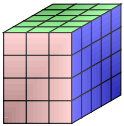


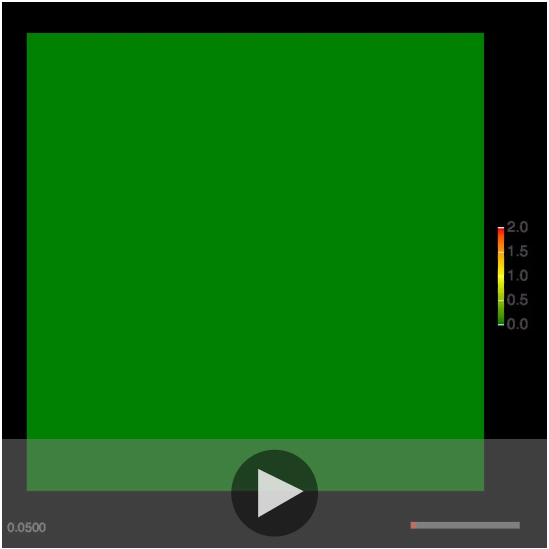
(12 elements)

$$\begin{aligned} X &= G \otimes H \otimes W \otimes V \\ &= xyt + xy^4t + xtln(y) + \\ &xt(sin(2y)+1) + x^3yt + \\ &xt(cos(y)+1) + yte^x + \\ &yt(sin(x)+1) + \\ &xyt^2 + xy(1+tanh(t-10)) + \\ &t^2e^x(sin(2y)+1) + \\ &(1+tanh(t-10))(sin(x)+1) \\ &(cos(y)+1) \end{aligned}$$

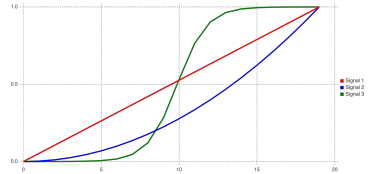
Tucker Tensor Decomposition: Example

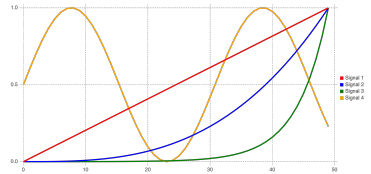


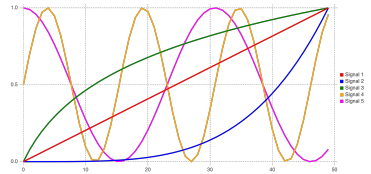




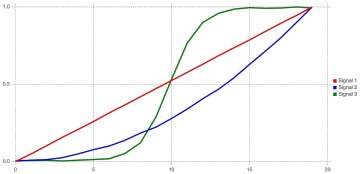
Tucker Tensor Decomposition: Example

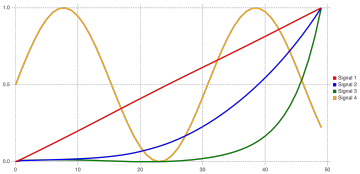


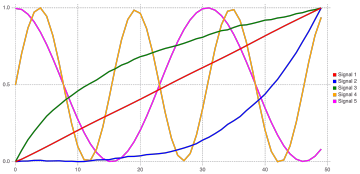




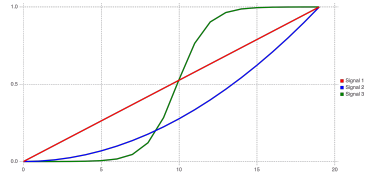
← **Truth** vs **Predictions** →



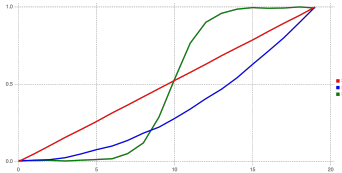




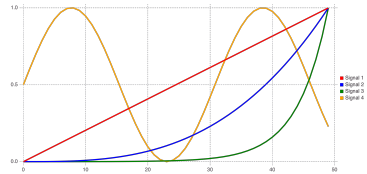
Tucker Tensor Decomposition: Example

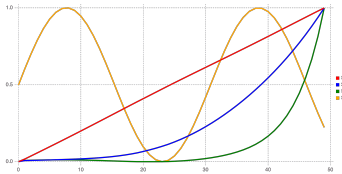


← Truth vs Predictions →

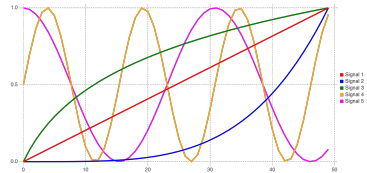


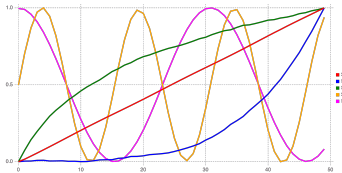
$$V = \begin{bmatrix} t \\ t^2 \\ tanh(t-10)+1 \end{bmatrix}$$





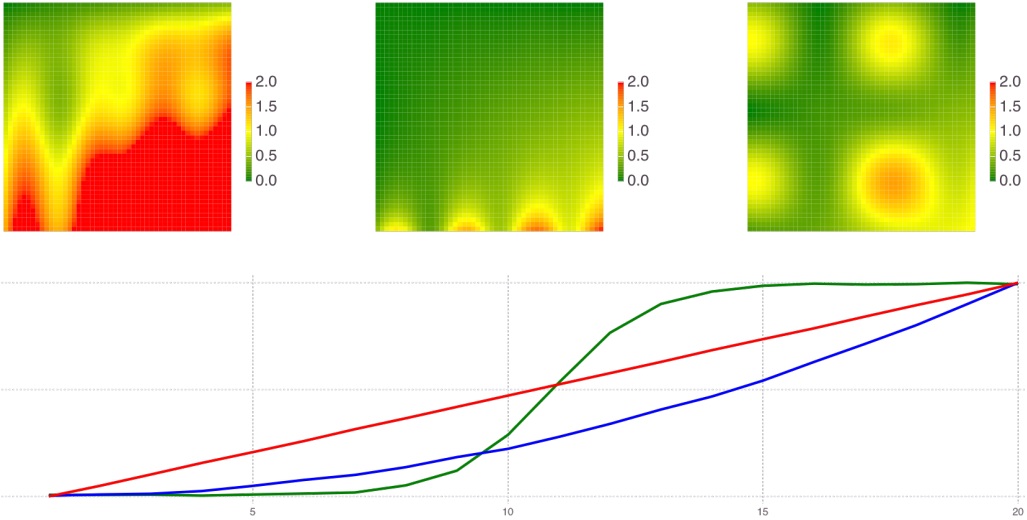
$$W = \begin{bmatrix} x \\ x^3 \\ e^x \\ sin(x)+1 \end{bmatrix}$$



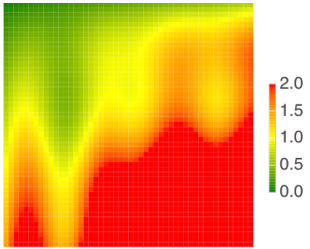


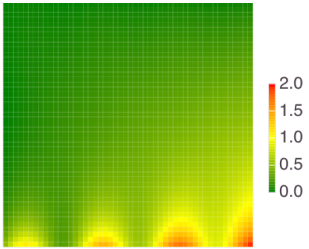
$$H = \begin{bmatrix} y \\ y^4 \\ ln(y) \\ sin(2y)+1 \\ cos(y)+1 \end{bmatrix}$$

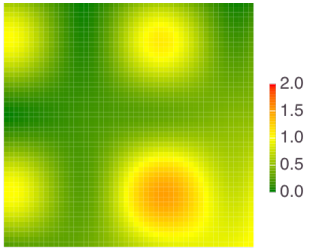
Tucker Tensor Decomposition: Example



Tucker Tensor Decomposition: Example







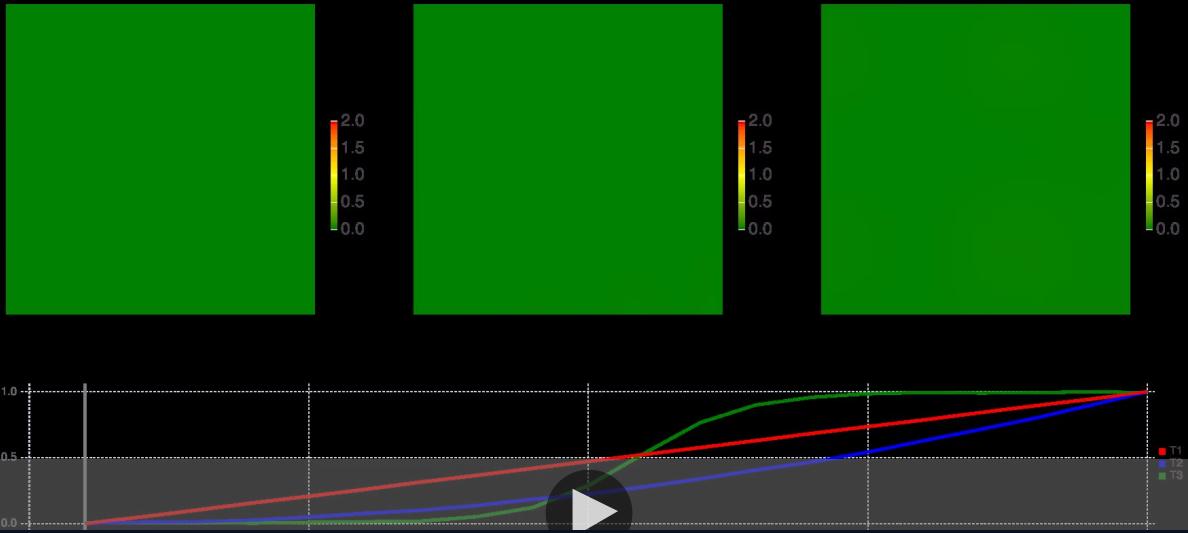
$$M = G \otimes H \otimes W$$

$$M_1 = xy + xy^4 + xlog(y+1) + x(sin(2y)+1) + ye^x + x(cos(y)+1) + x^3y + y(sin(x)+1)$$

$$M_2 = xy + e^x(cos(z)+1)$$

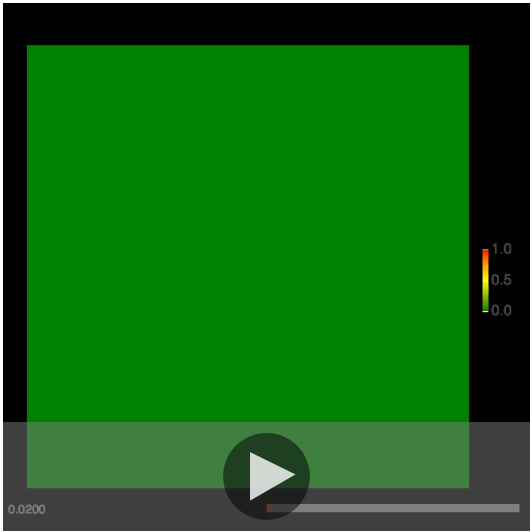
$$M_3 = xy + (sin(x)+1)(cos(y)+1)$$

Tucker Tensor Decomposition: Example

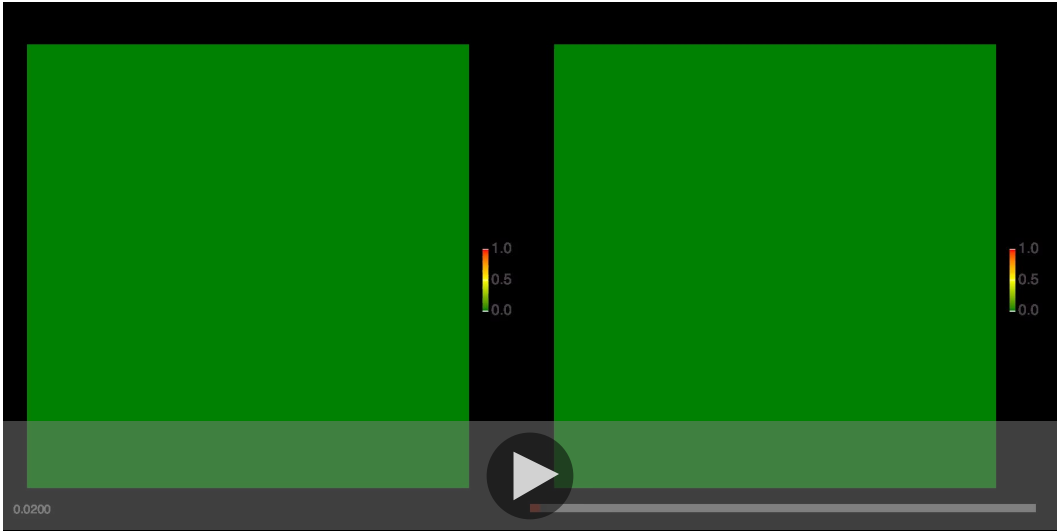


Tucker Tensor Decomposition: Example II

- $(50 \times 50 \times 50)$ tensor
- 50 columns in x
- 50 rows in y
- 50 time frames
- ‘ones‘ swimming in a sea of ‘zeros‘

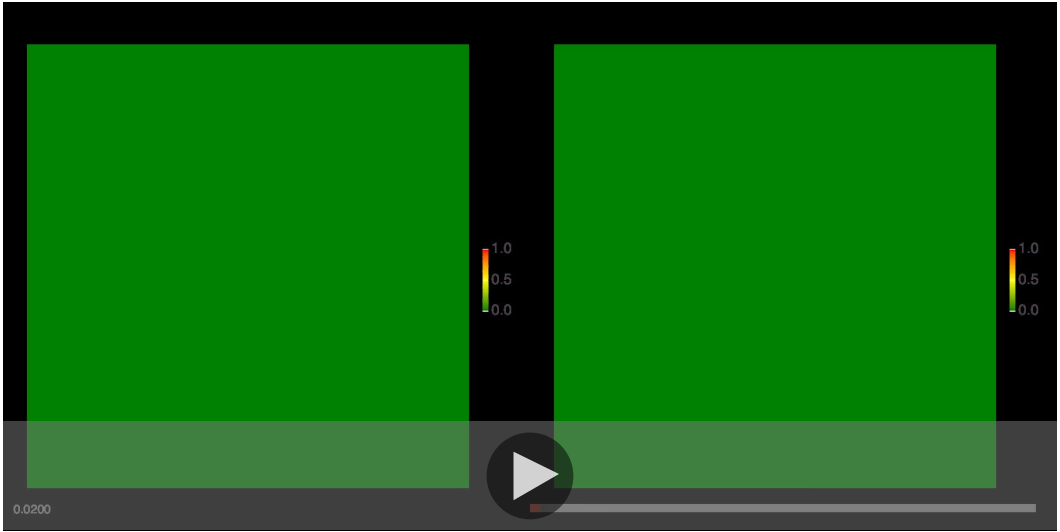


Tucker Tensor Decomposition: Example II



Factorizing all **3** dimensions $(50 \times 50 \times 50) \rightarrow (6 \times 44 \times 48)$

Tucker Tensor Decomposition: Example II



6 groups of swimmers (x); 44 lanes occupied (y); 48 time frames (first/last empty)

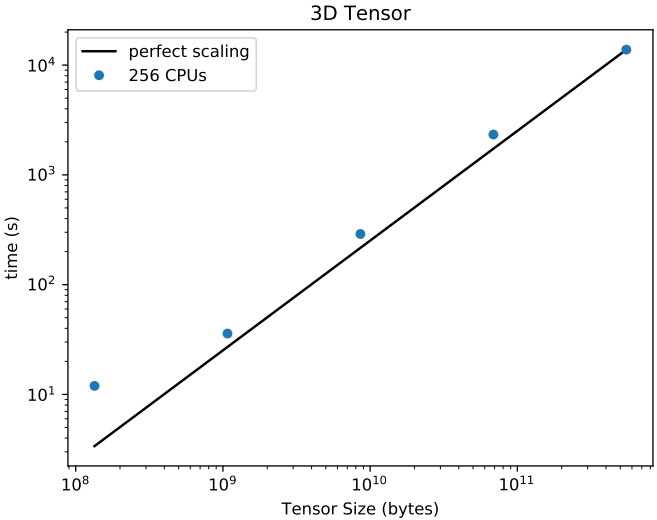
NMFk / NTFk Challenges

- **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 its 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

NMFk / NTFk Scalability



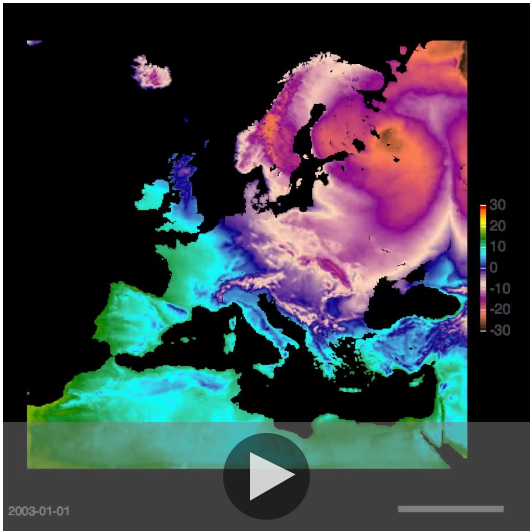
- **Field Data**:
  - Characterization of groundwater contaminant sources
  - US Climate data
  - Geothermal data
  - Seismic data
- **Lab Data**:
  - X-ray Spectroscopy
  - UV Fluorescence Spectroscopy
  - Microbial population analyses
- **Operational Data**:
  - LANSCE: Los Alamos Neutron Accelerator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

NMFk / NTFk Publications

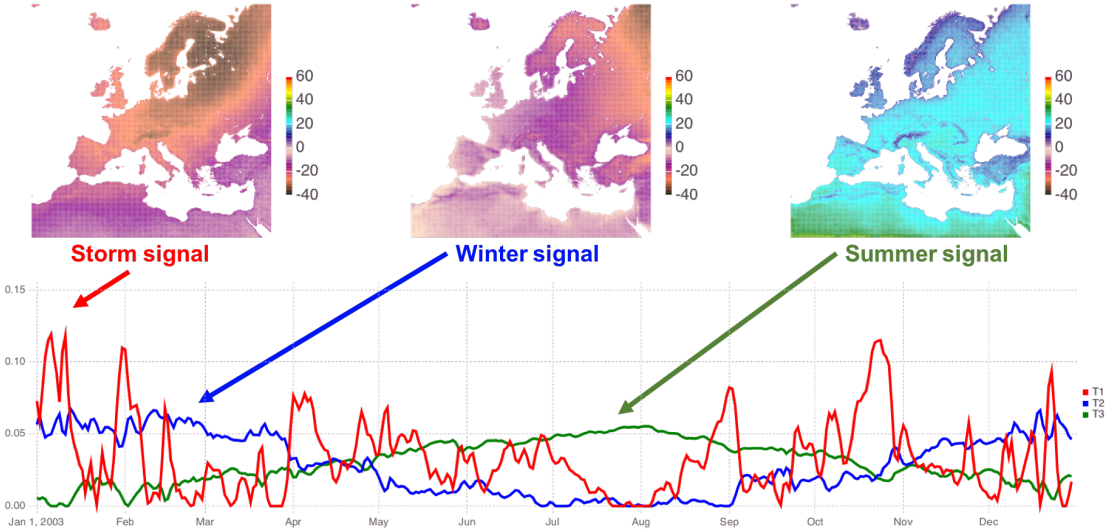
- Stanev, Vesselinov, Kusne, Antoszewski, Takeuchi, Alexandrov, Unsupervised Phase Mapping of X-ray Diffraction Data by Nonnegative Matrix Factorization Integrated with Custom Clustering, **Nature Computational Materials**, 2018.
- Vesselinov, Munuduru, Karra, O'Maley, Alexandrov, Unsupervised Machine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

Climate model of Europe: air temperature

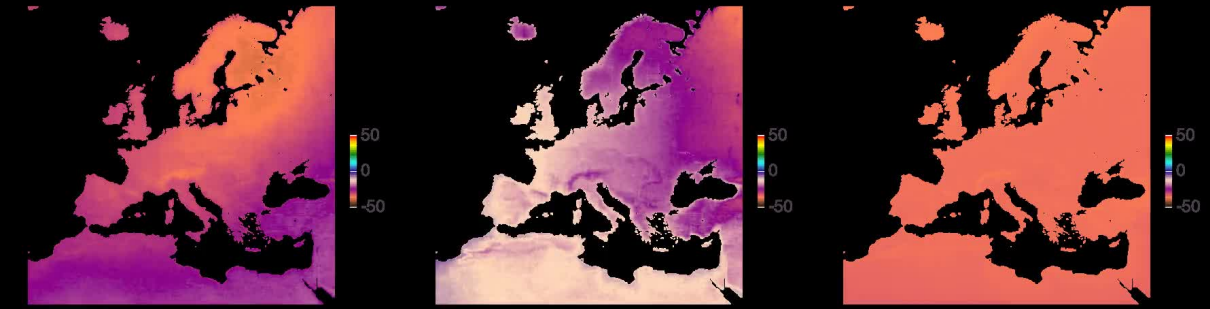
- fluctuations in the air temperature [$^{\circ}C$]
- $(424 \times 412 \times 365)$
  $(columns \times rows \times days)$
- **NTF**k extracts spatial and temporal footprints of dominant features:
  - storm signal
  - winter seasonal signal
  - summer seasonal signal
  - ....
- Data compression: ~ 4GB $\rightarrow \sim 0.5$MB
  Compression ratio: ~ 8000

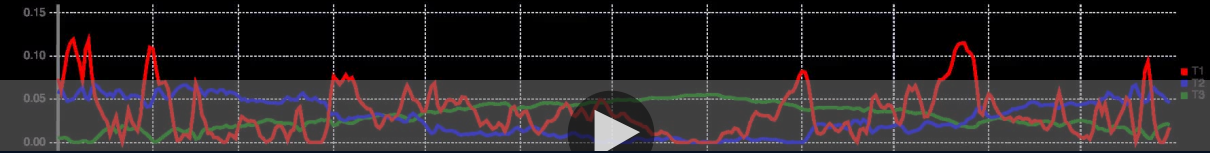


Climate model of Europe: air temperature fluctuations represented by 3 signals



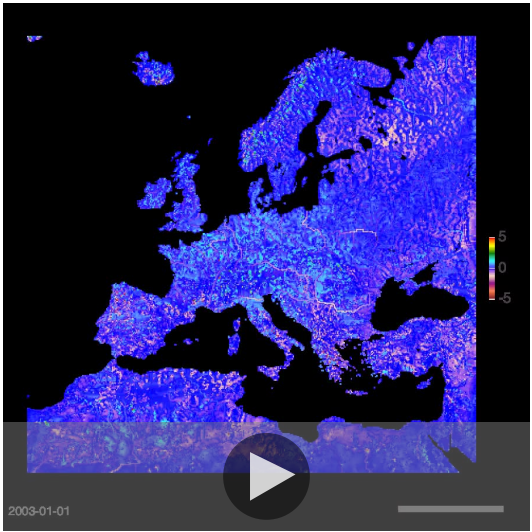
Climate model of Europe: air temperature fluctuations represented by 3 signals



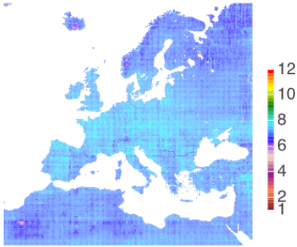


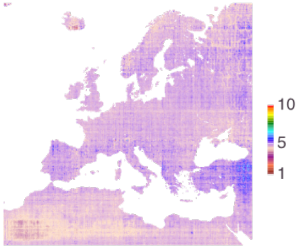
Climate model of Europe: water-table depth

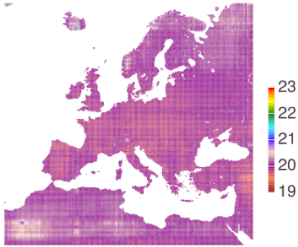
- fluctuations in the water-table depth $[m]$
- $(424 \times 412 \times 365)$
  $(columns \times rows \times days)$
- **NTF**k extracts spatial and temporal footprints of dominant signals
  - spring snowmelt signal
  - summer rainfall signal
  - seasonal signal
  - ....



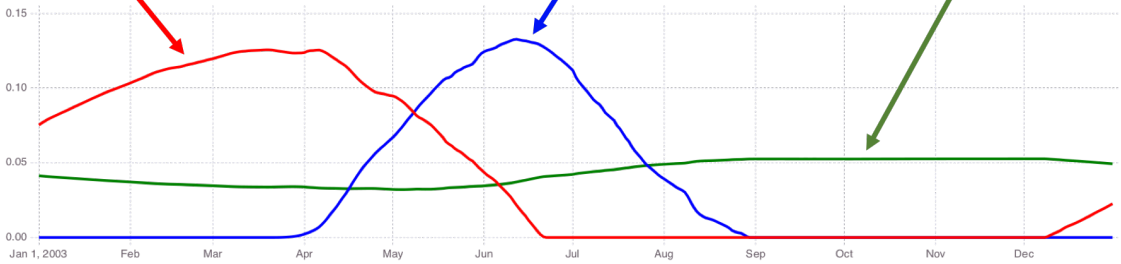
Climate model of Europe: water-table fluctuations represented by 3 signals



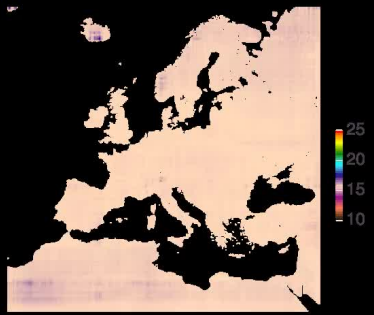


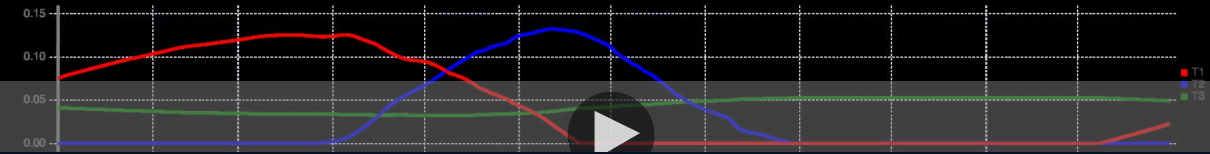


Spring snowmelt **Summer rainfalls** **Seasonality**

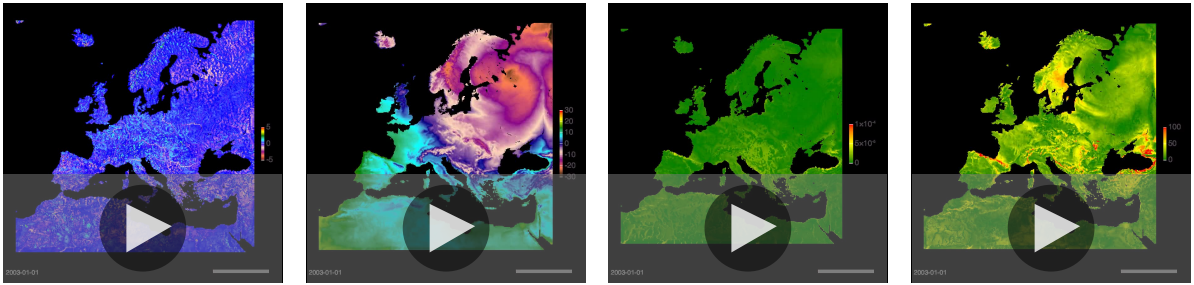


Climate model of Europe: water-table fluctuations represented by 3 signals



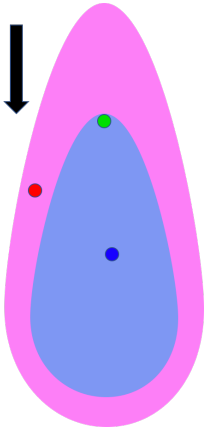


Climate model of Europe: analyze all model outputs (>40) simultaneously

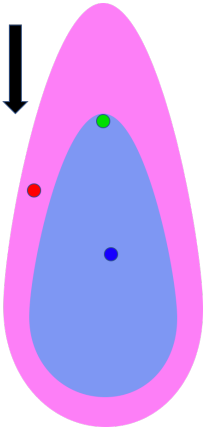


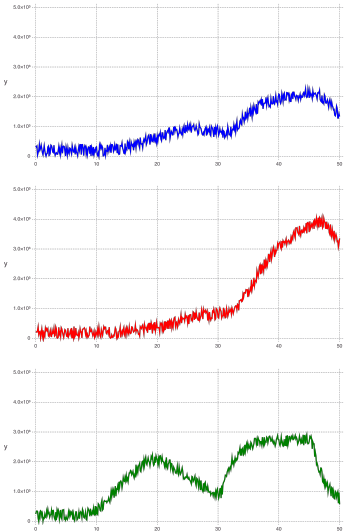
- Find interconnections among model outputs
- Evaluate impacts of different model setups
- Find dominant processes impacting model predictions
  (e.g., climate impacts on groundwater resources, impacts of subsurface processes on atmospheric conditions)

ML for extracting contaminant plumes

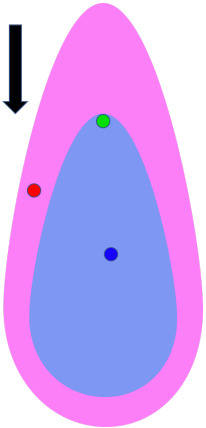


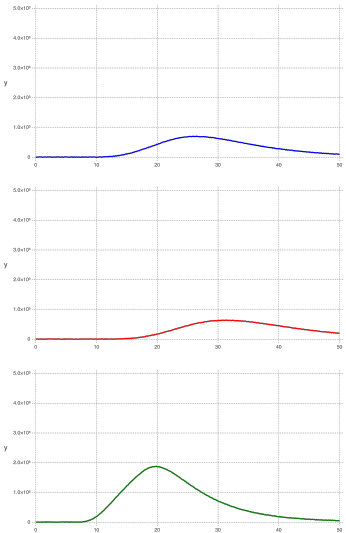
ML for extracting contaminant plumes

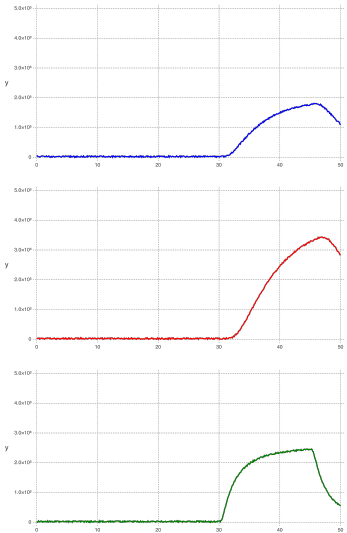




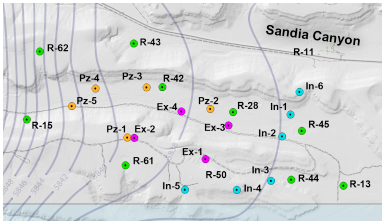
ML for extracting contaminant plumes



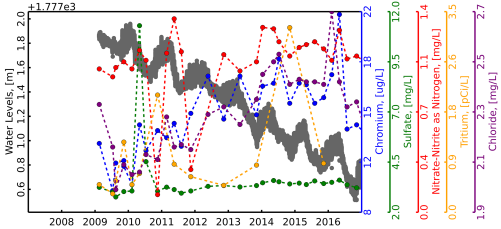




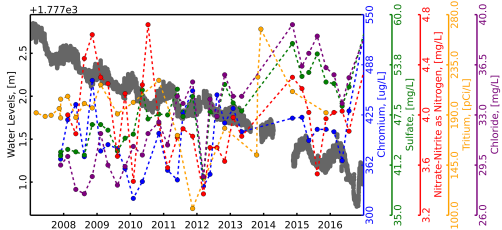
Geochemistry: LANL hydrogeochemical dataset



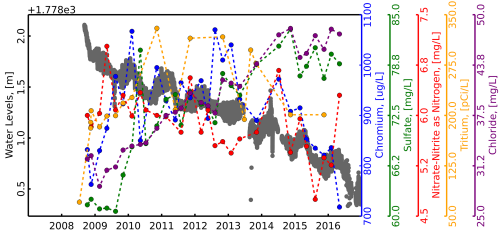
R-44#1



R-28

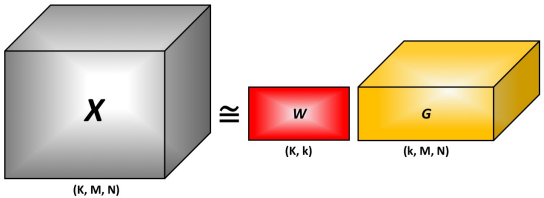


R-42



$(18 \times 8 \times 12)$ tensor $(wells \times species \times years)$

Geochemistry: Nonnegative Tensor Factorization based on Tucker-1 decomposition



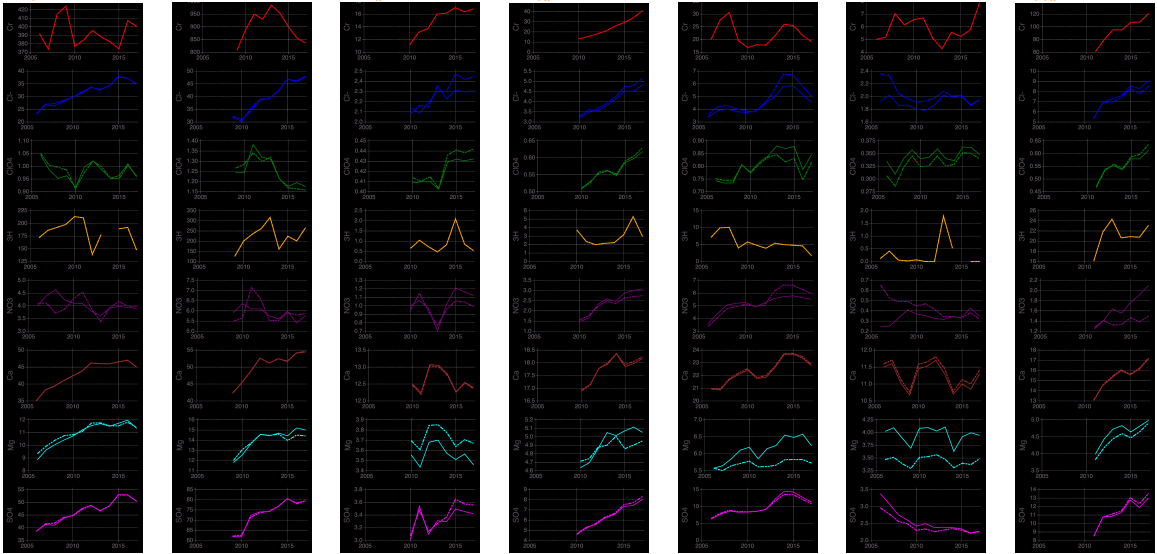
- **X**: data tensor
- **W**: source (groundwater type) matrix (**unknown**)
- **G**: mixing tensor (**unknown**)

- **M**: number of observation points (wells)
- **N**: number of observation times (e.g, 2001, 2002, ..., 2017)
- **K**: 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 tensor/matrix elements ≥ 0
  $$\sum_{i=1}^{k} G_{i,j,t} = 1 \; \forall j, t$$

NTFk analysis estimated 7 groundwater types

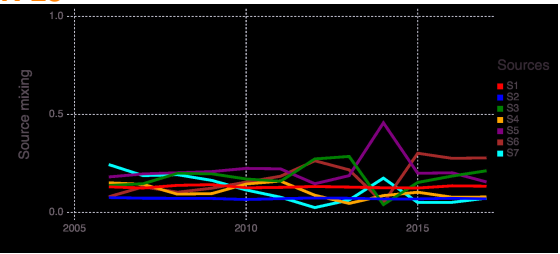
Sources	Cr $(\mu g/L)$	Cl^- (mg/L)	ClO_4 $(\mu g/L)$	^3H (pCi/L)	NO_3 (mg/L)	Ca (mg/L)	Mg (mg/L)	SO_4 (mg/L)
S1	**2970.00**	**63.00**	0.00	0.00	**14.00**	**73.00**	**25.00**	**170.00**
S5	**21.00**	**51.00**	0.00	**950.00**	2.40	**67.00**	15.00	**50.00**
S6	1.50	**64.00**	0.00	0.00	2.80	51.00	10.00	**68.00**
S2	0.79	0.35	**14.00**	0.00	0.50	5.30	1.70	0.60
S4	0.50	0.14	0.00	0.00	**10.00**	21.00	5.00	10.00
S3 (**B**)	0.25	**3.60**	0.00	0.00	0.01	**41.00**	**11.00**	0.06
S7 (**B**)	0.10	0.03	0.00	0.00	0.01	0.40	0.80	0.90

NTFk estimated concentrations at various wells

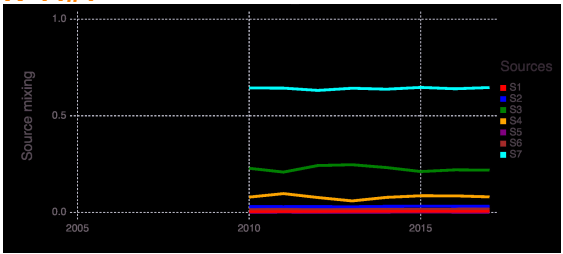


NTFk estimated time-dependent mixing of 7 groundwater types at various wells

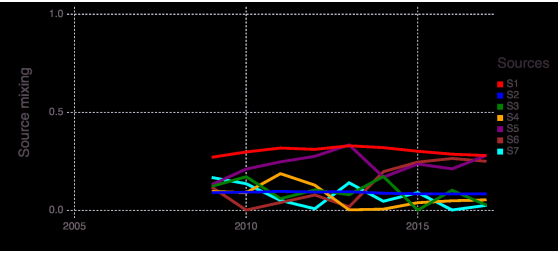
R-28



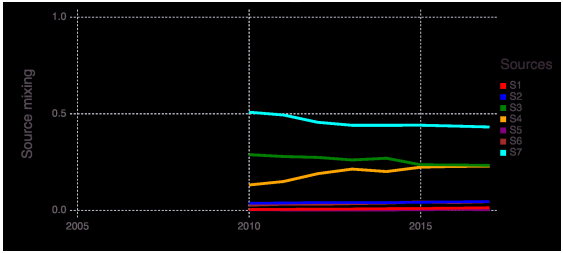
R-44#1



R-42

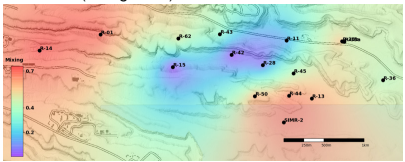


R-45#1

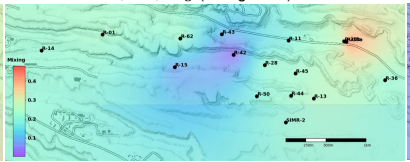


NTFk identified sources (groundwater types) Jan-Dec 2016

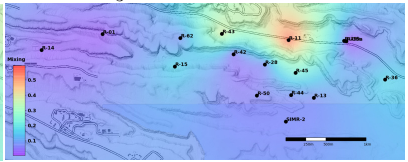
Source 7: (background)



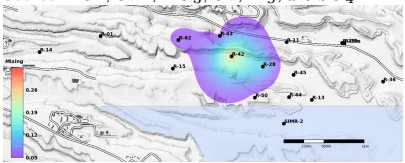
Source 3: Cl^-, Ca, Mg (background)



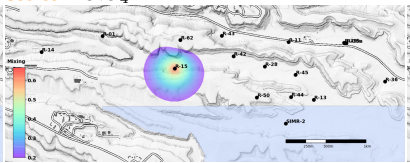
Source 4: NO_3



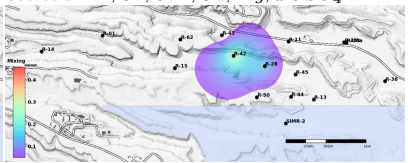
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



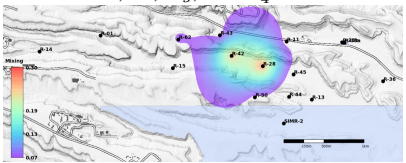
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

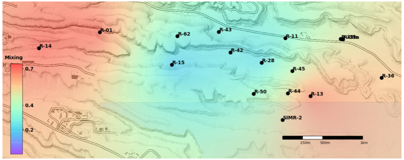


Source 6: Cl^-, Ca, Mg, and SO_4

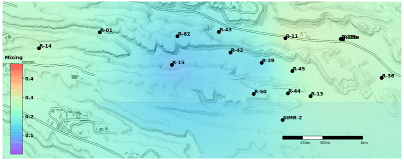


NTFk identified sources (groundwater types) Jan-Dec 2005

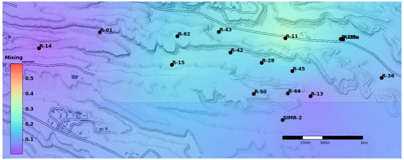
Source 7: (background)



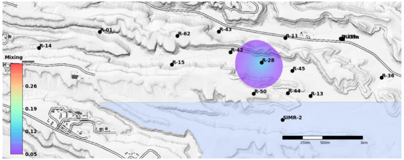
Source 3: Cl^-, Ca, Mg (background)



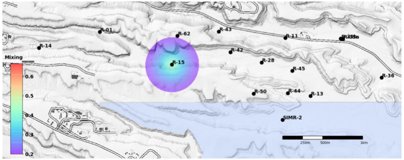
Source 4: NO_3



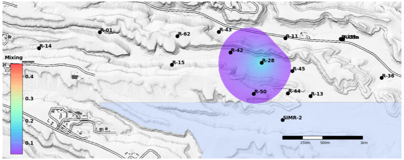
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



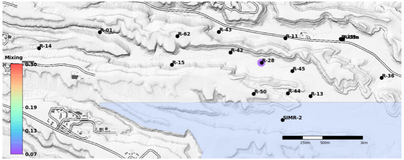
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

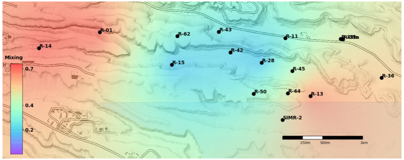


Source 6: Cl^-, Ca, Mg, and SO_4

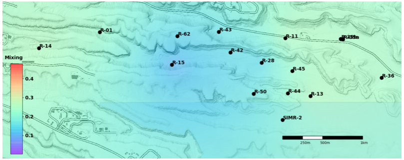


NTFk identified sources (groundwater types) Jan-Dec 2006

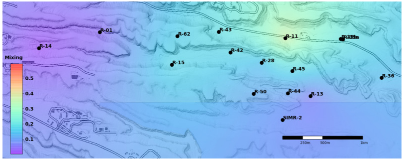
Source 7: (background)



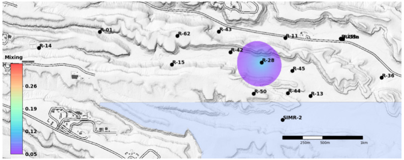
Source 3: Cl^-, Ca, Mg (background)



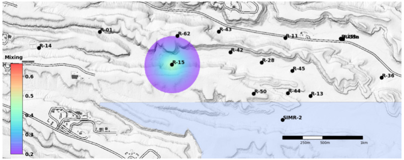
Source 4: NO_3



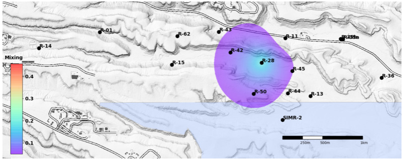
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



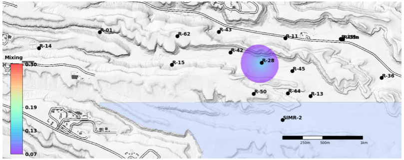
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

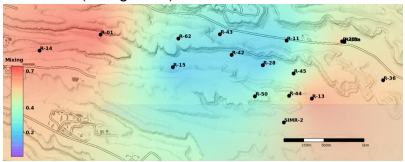


Source 6: Cl^-, Ca, Mg, and SO_4

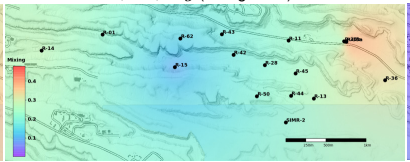


NTFk identified sources (groundwater types) Jan-Dec 2007

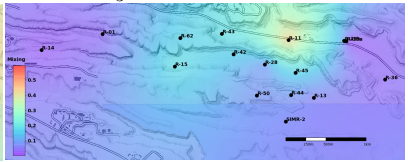
Source 7: (background)



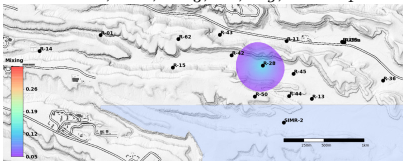
Source 3: Cl^-, Ca, Mg (background)



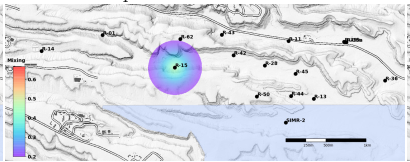
Source 4: NO_3



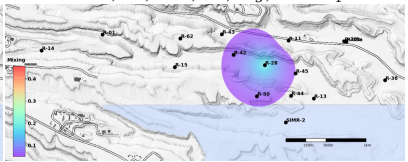
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



Source 2: ClO_4



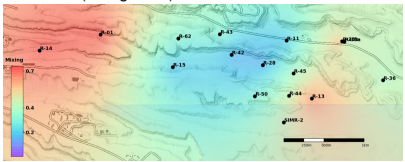
Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4



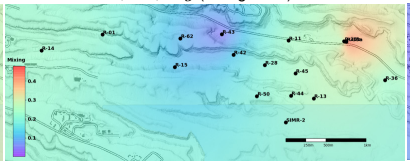
Source 6: Cl^-, Ca, Mg, and SO_4

NTFk identified sources (groundwater types) Jan-Dec 2008

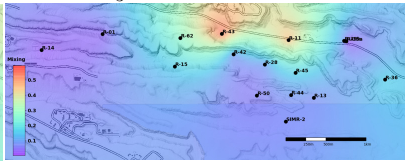
Source 7: (background)



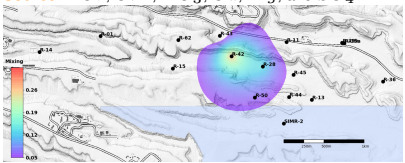
Source 3: Cl^-, Ca, Mg (background)



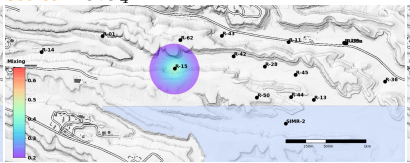
Source 4: NO_3



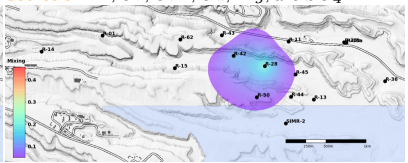
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



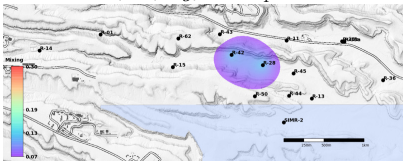
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

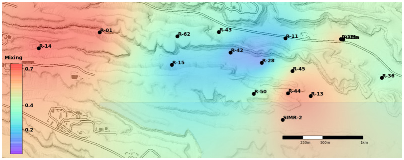


Source 6: Cl^-, Ca, Mg, and SO_4

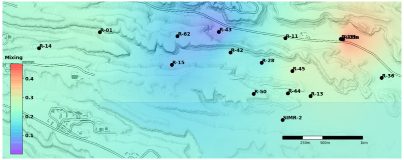


NTFk identified sources (groundwater types) Jan-Dec 2009

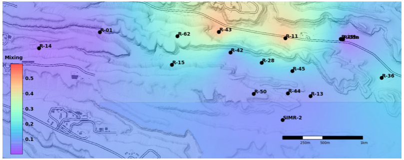
Source 7: (background)



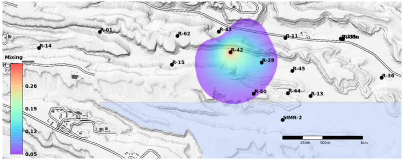
Source 3: Cl^-, Ca, Mg (background)



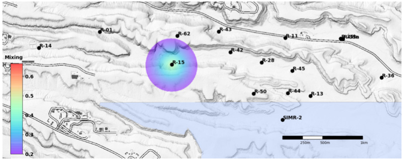
Source 4: NO_3



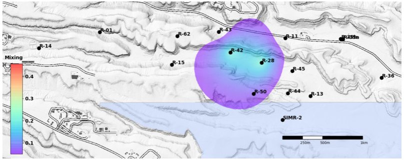
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



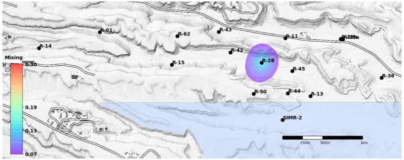
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

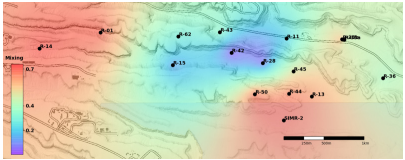


Source 6: Cl^-, Ca, Mg, and SO_4

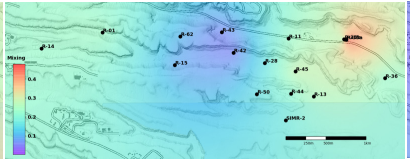


NTFk identified sources (groundwater types) Jan-Dec 2010

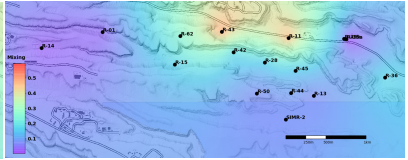
Source 7: (background)



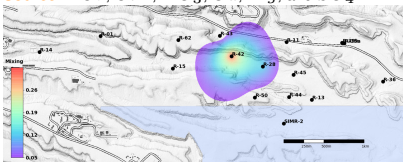
Source 3: Cl^-, Ca, Mg (background)



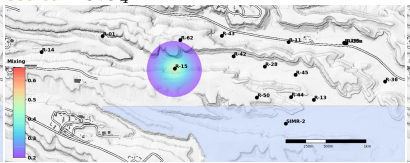
Source 4: NO_3



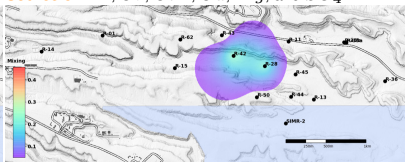
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



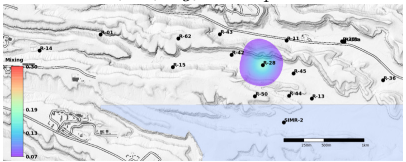
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

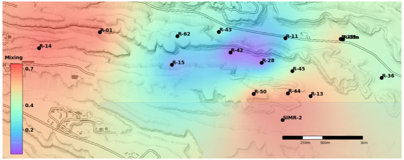


Source 6: Cl^-, Ca, Mg, and SO_4

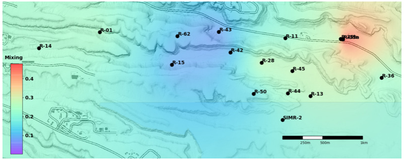


NTFk identified sources (groundwater types) Jan-Dec 2011

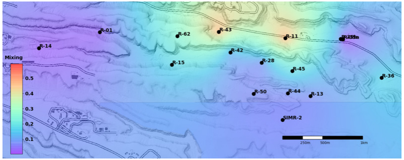
Source 7: (background)



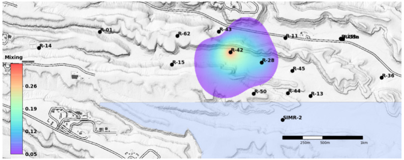
Source 3: Cl^-, Ca, Mg (background)



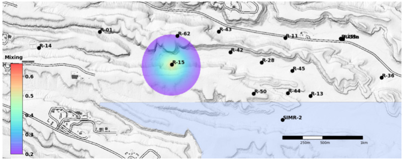
Source 4: NO_3



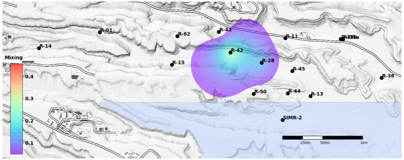
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



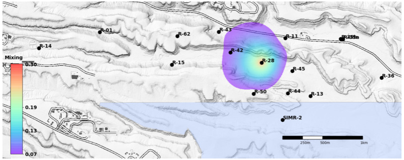
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

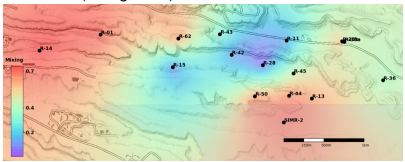


Source 6: Cl^-, Ca, Mg, and SO_4

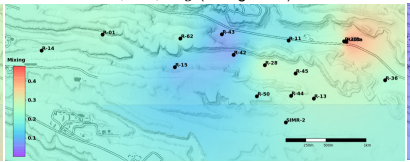


NTFk identified sources (groundwater types) Jan-Dec 2012

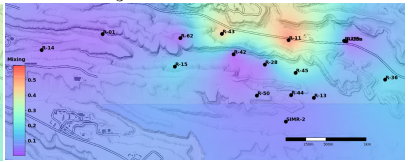
Source 7: (background)



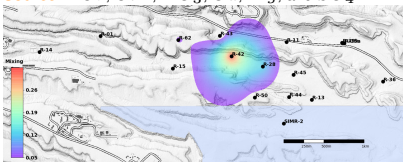
Source 3: Cl^-, Ca, Mg (background)



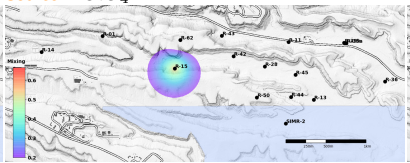
Source 4: NO_3



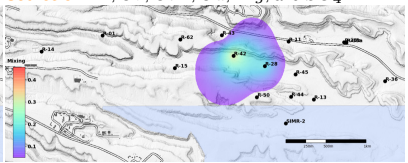
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



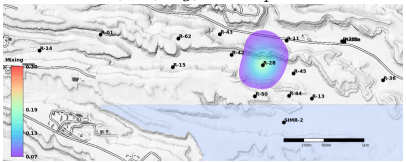
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

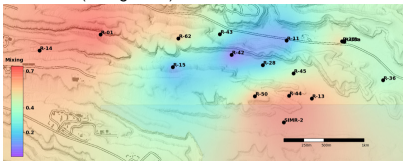


Source 6: Cl^-, Ca, Mg, and SO_4

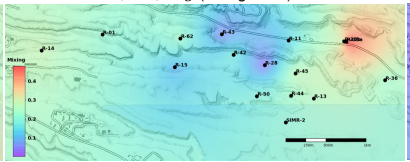


NTFk identified sources (groundwater types) Jan-Dec 2013

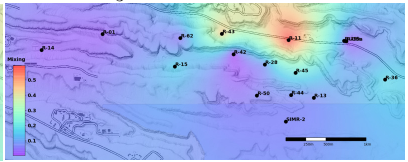
Source 7: (background)



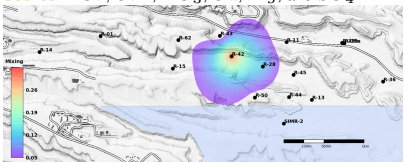
Source 3: Cl^-, Ca, Mg (background)



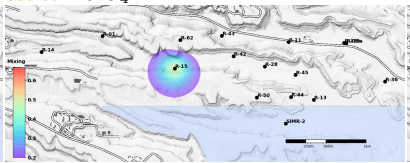
Source 4: NO_3



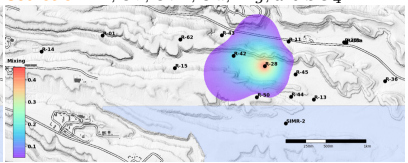
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



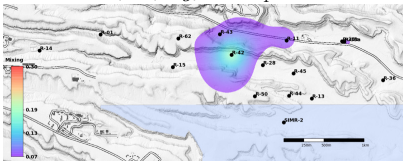
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

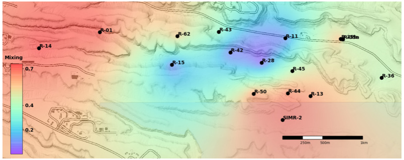


Source 6: Cl^-, Ca, Mg, and SO_4

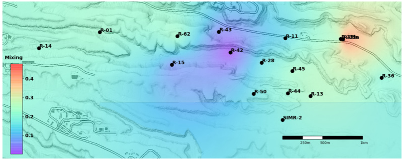


NTFk identified sources (groundwater types) Jan-Dec 2014

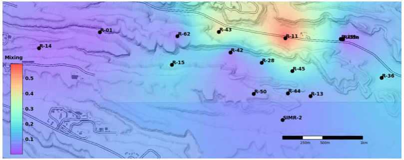
Source 7: (background)



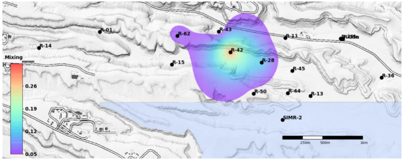
Source 3: Cl^-, Ca, Mg (background)



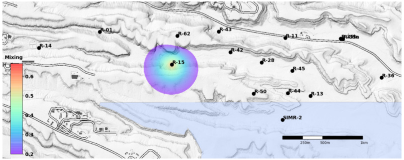
Source 4: NO_3



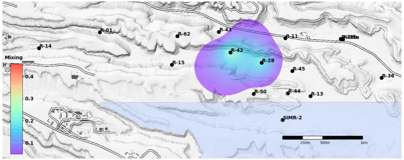
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



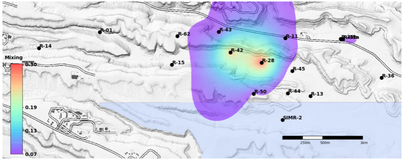
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

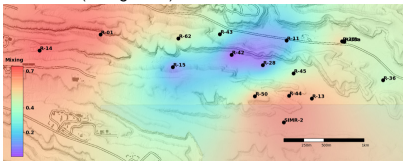


Source 6: Cl^-, Ca, Mg, and SO_4

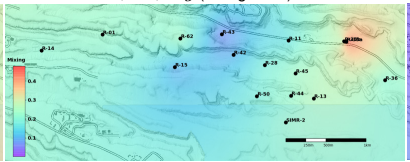


NTFk identified sources (groundwater types) Jan-Dec 2015

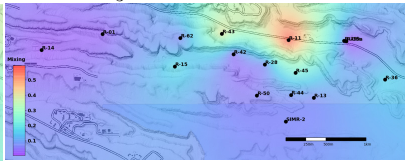
Source 7: (background)



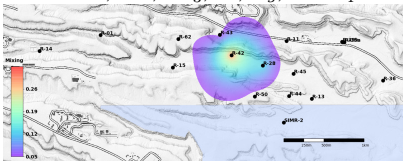
Source 3: Cl^-, Ca, Mg (background)



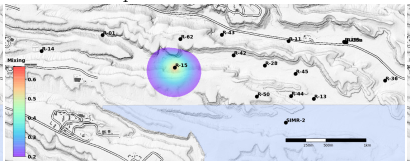
Source 4: NO_3



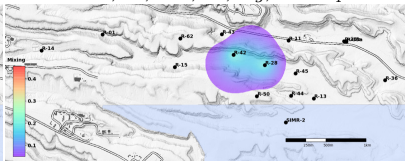
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



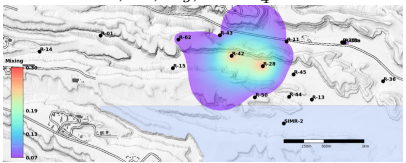
Source 2: ClO_4



Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4

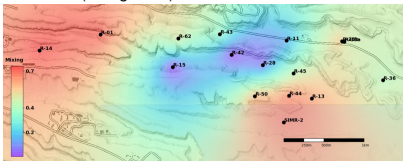


Source 6: Cl^-, Ca, Mg, and SO_4

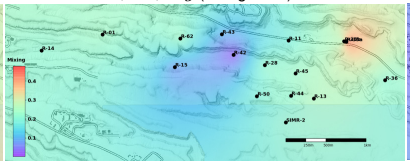


NTFk identified sources (groundwater types) Jan-Dec 2016

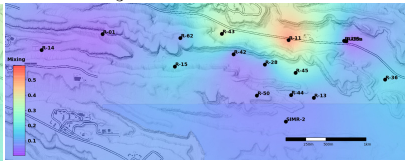
Source 7: (background)



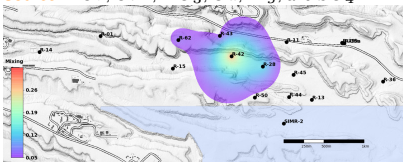
Source 3: Cl^-, Ca, Mg (background)



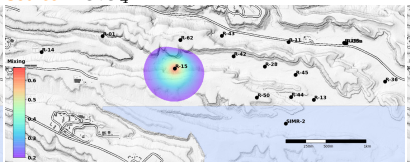
Source 4: NO_3



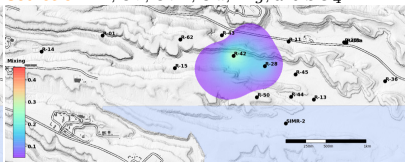
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



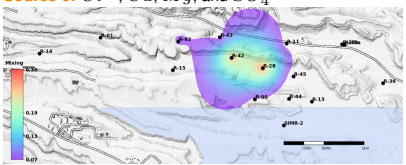
Source 2: ClO_4



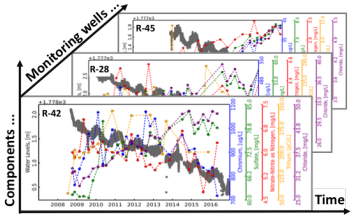
Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4



Source 6: Cl^-, Ca, Mg, and SO_4



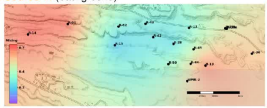
NTFk analysis of LANL hydrogeochemical datasets



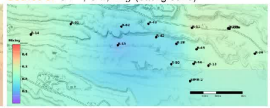
$(18 \times 8 \times 12)$ tensor
$(wells \times species \times years)$

January 2005 – December 2005

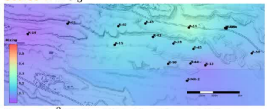
Source 7: (background)



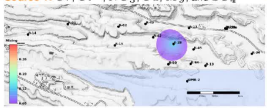
Source 3: Cl^-, Ca, Mg (background)



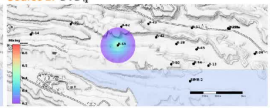
Source 4: NO_3



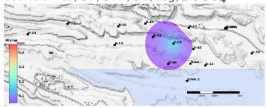
Source 1: $Cr, Cl^-, NO_3, Ca, Mg,$ and SO_4



Source 2: ClO_4



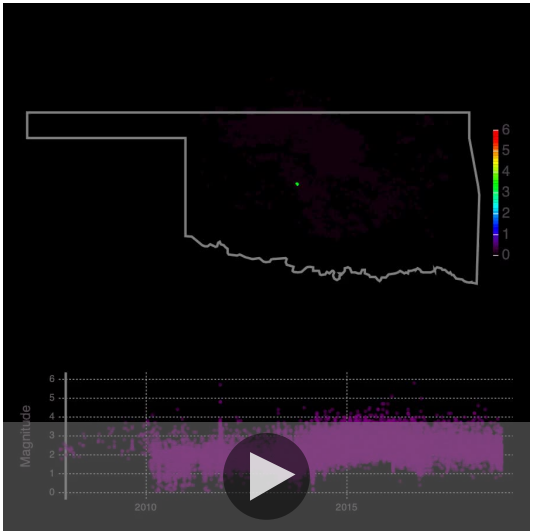
Source 5: $^3H, Cr, Cl^-, Ca, Mg,$ and SO_4



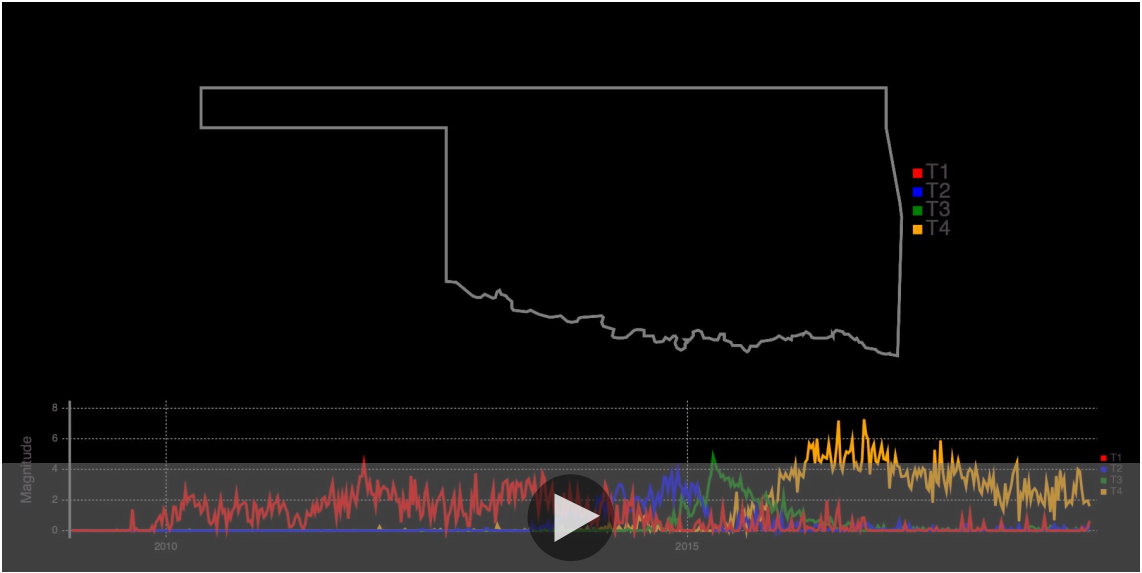
Source 6: $Cl^-, Ca, Mg,$ and SO_4

Oklahoma seismicity

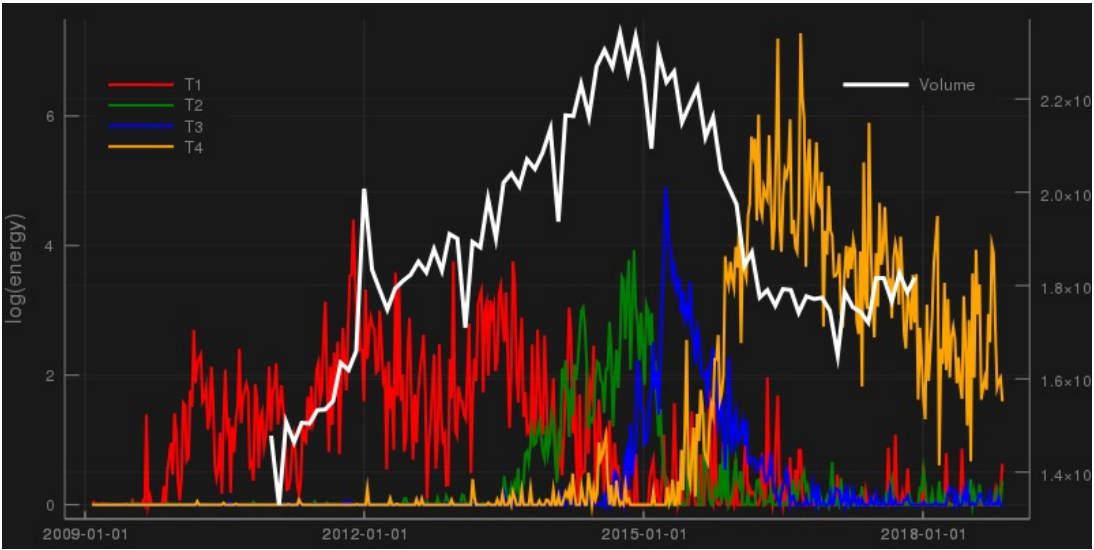
- 32,251 seismic events from 1989 to 2017
- Tensor: total energy of events over a discretized domain
  $(118 \times 97 \times 520)$
  $(columns \times rows \times weeks)$
- **NTF**k applied to extract dominant hidden (latent) features based on spatial footprints and temporal characteristics



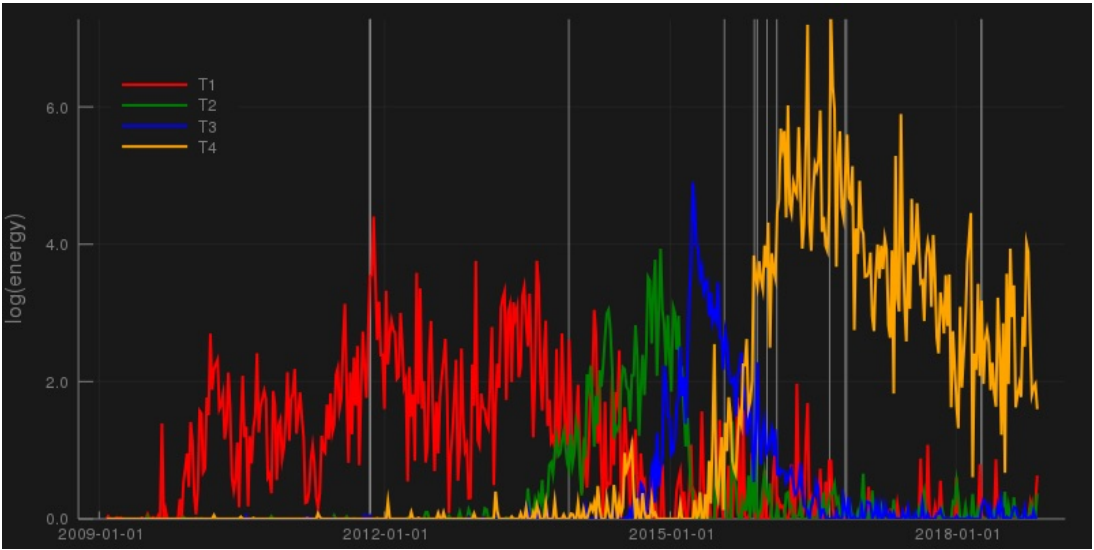
Oklahoma seismicity: reconstruction by 4 features (signals)



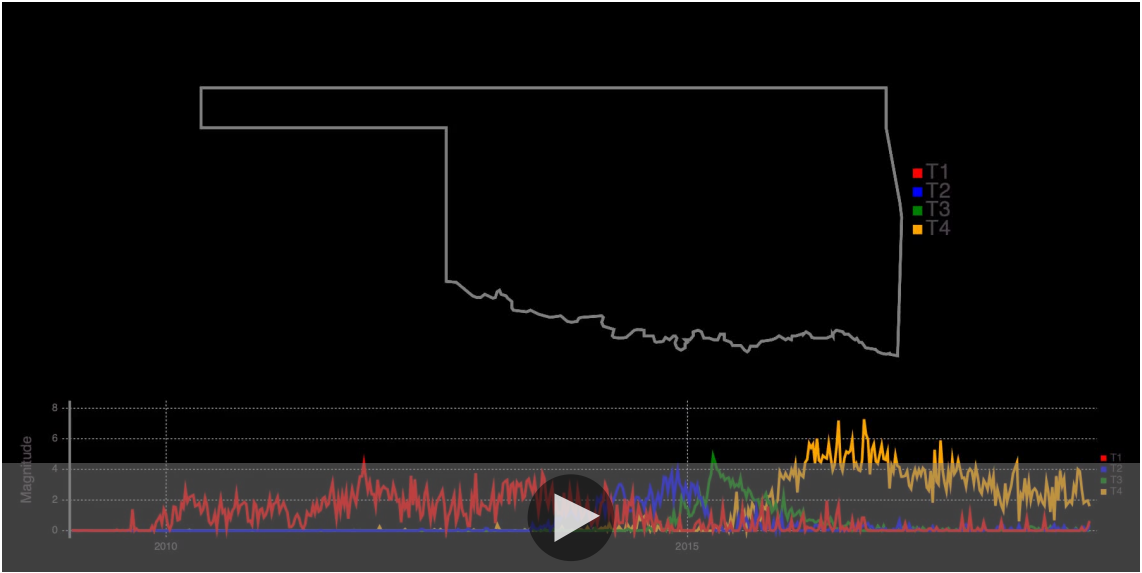
Oklahoma seismicity: extracted signals vs. injected volumes



Oklahoma seismicity: extracted signals vs. majors seismic events

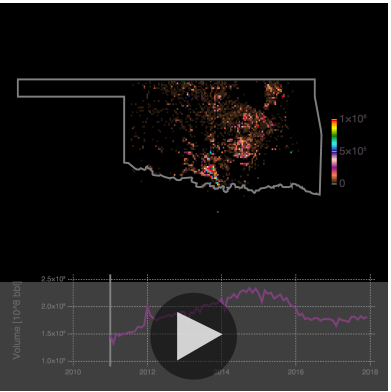


Oklahoma seismicity: 4 seismic features

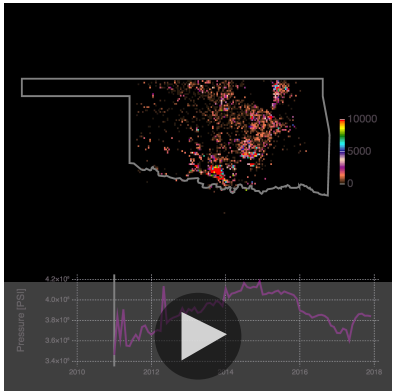


Oklahoma seismicity

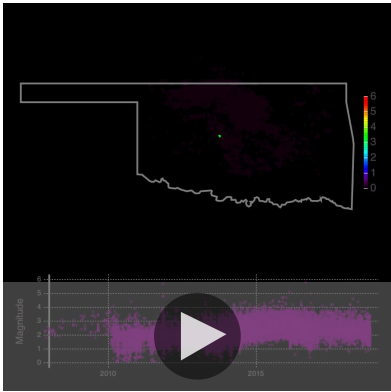
volume



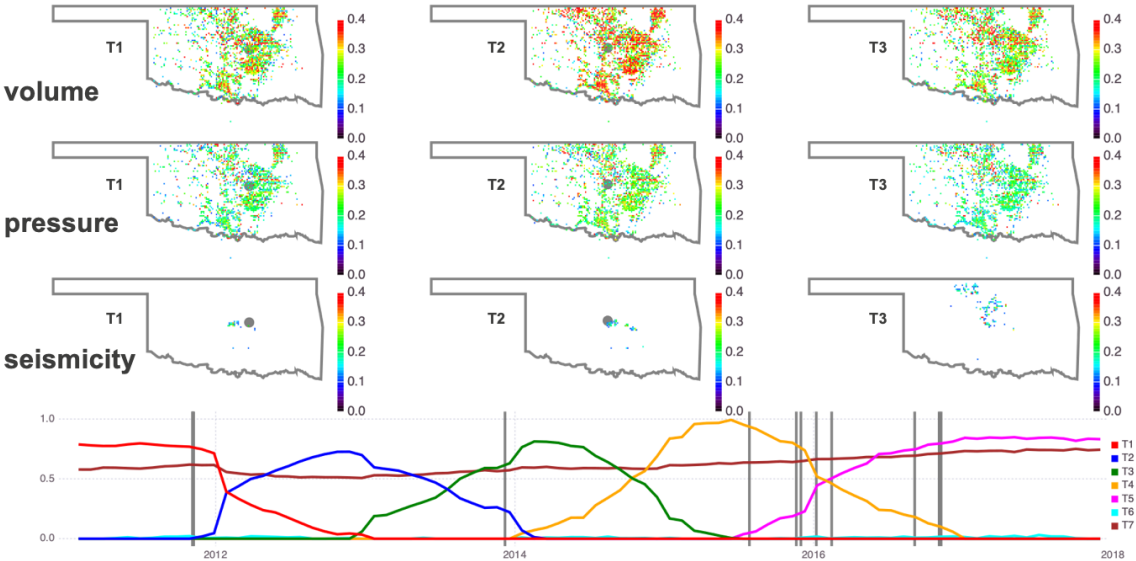
pressure



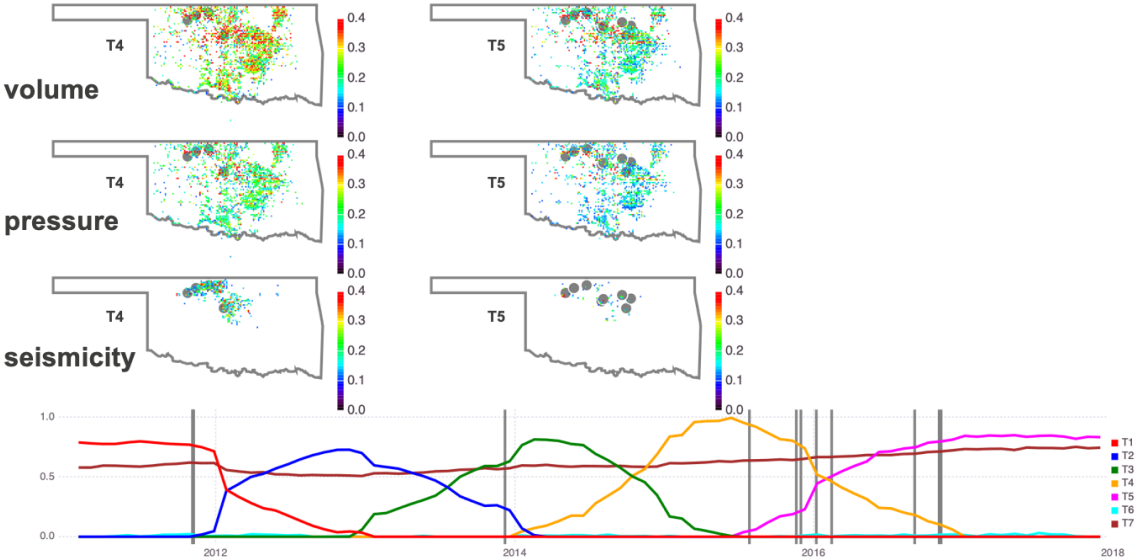
seismicity



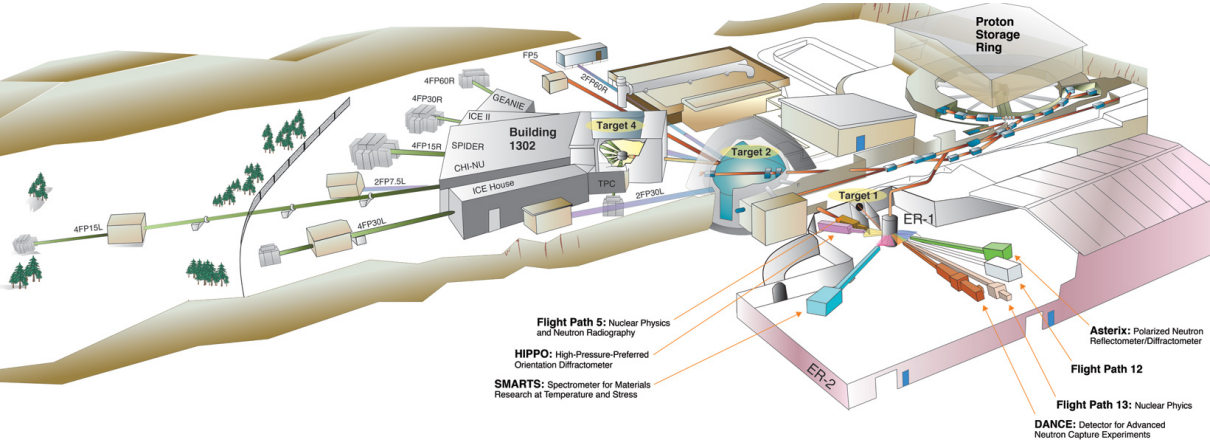
Oklahoma seismicity: 5 volume/pressure/seismicity features



Oklahoma seismicity: 5 volume/pressure/seismicity features



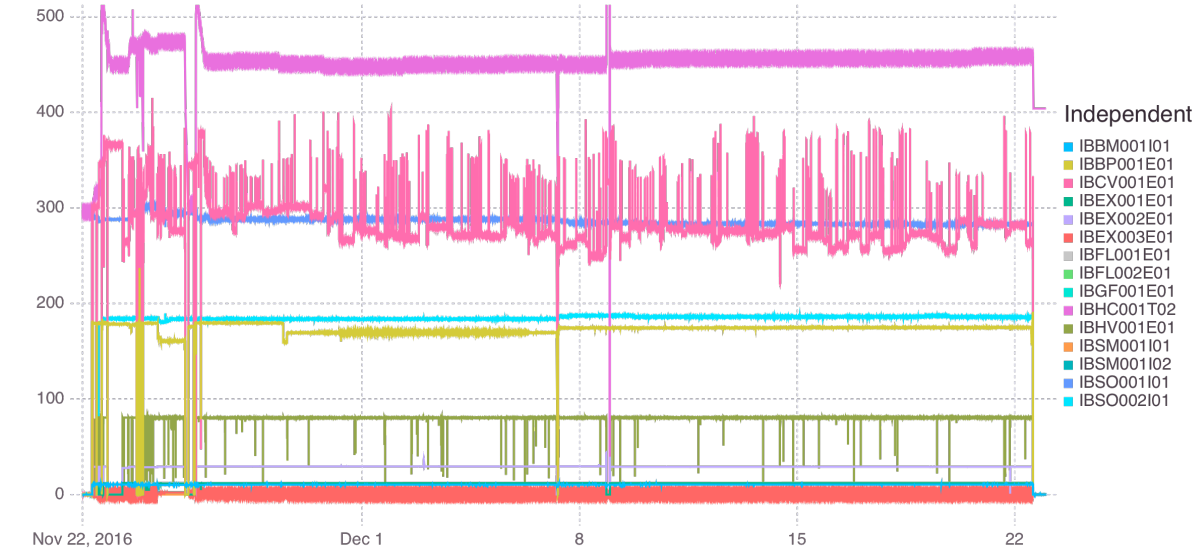
LANSCE


Proton
Storage
Ring
FP5
4FP60R
2FP60R
GEANIE
4FP30R
ICE II
Building
1302
Target 4
4FP15R
SPIDER
CHI-NU
Target 2
2FP7.5L
ICE House
TPC
2FP30L
Target 1
4FP15L
4FP30L
ER-1
Flight Path 5: Nuclear Physics
and Neutron Radiography
HIPPO: High-Pressure-Preferred
Orientation Diffractometer
SMARTS: Spectrometer for Materials
Research at Temperature and Stress
ER-2
Asterix: Polarized Neutron
Reflectometer/Diffractometer
Flight Path 12
Flight Path 13: Nuclear Phyics
DANCE: Detector for Advanced
Neutron Capture Experiments

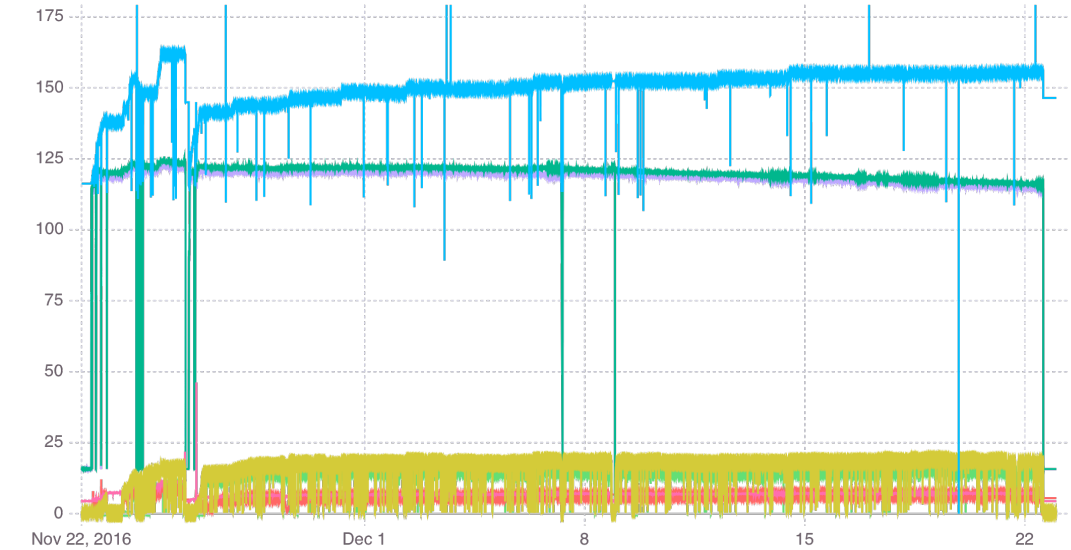
LANSCE: Problem

- Numerous independent variables ... knobs controlling the accelerator performance
  15 provided / analyzed
- Numerous dependent variables ... observations representing the accelerator performance
  8 provided / analyzed
- 312,326 * 23 = 7,183,498 $\approx$ 100 MB (per month)
  Full monthly dataset $\approx$ 2 GB; For 30 years $\approx$ 1 TB;
- Goal **#1**: ML a reduced-order model to simulate the performance (physics model does not exist and cannot be built)
- Goal **#2**: Use ML to control the knobs to optimize the performance

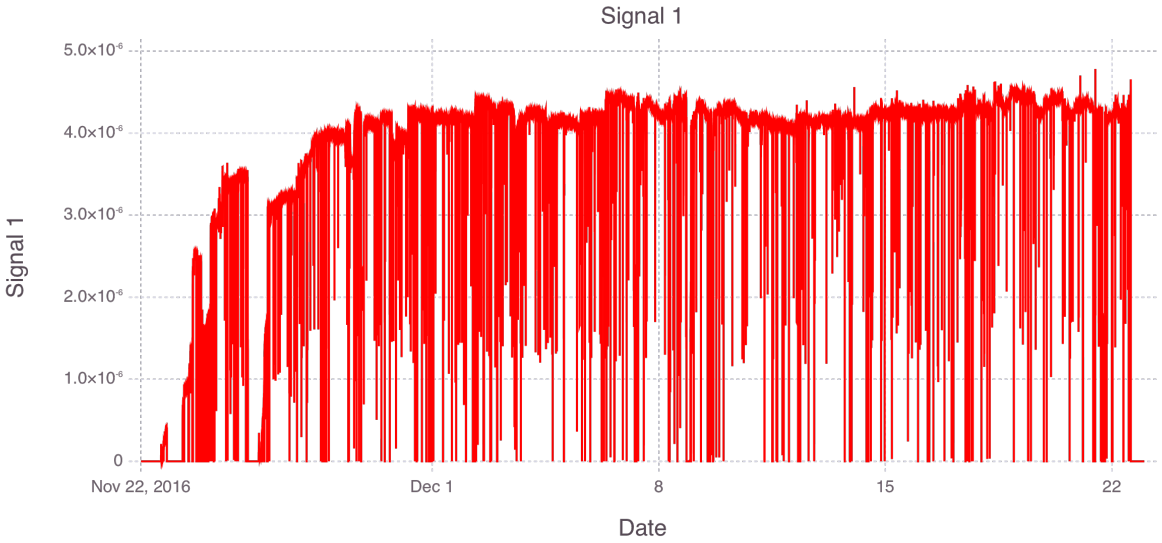
LANSCE: Independent variables = 15



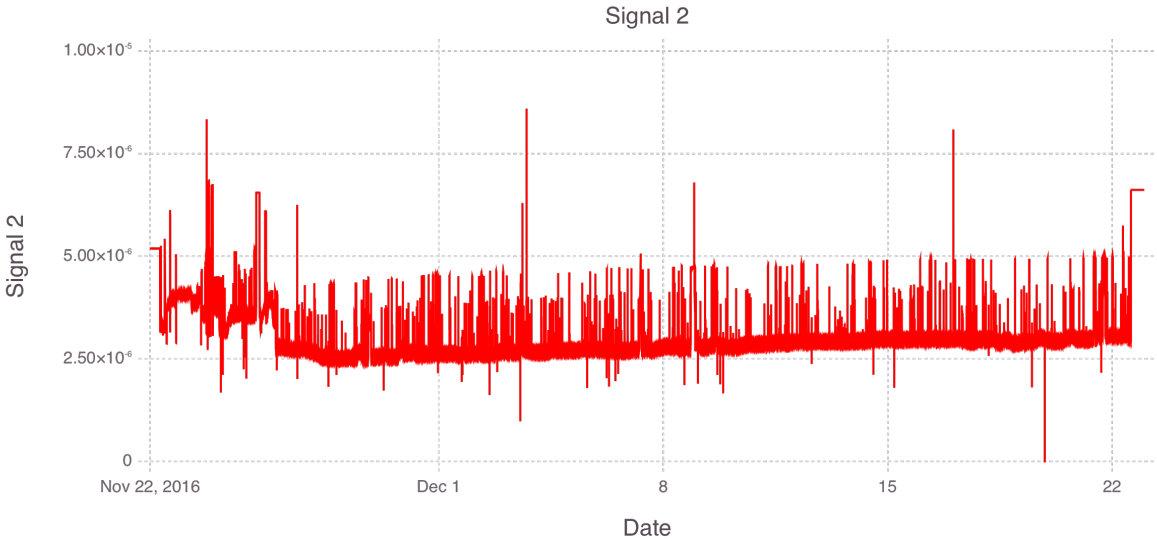
LANSCE: Dependent variables = 8



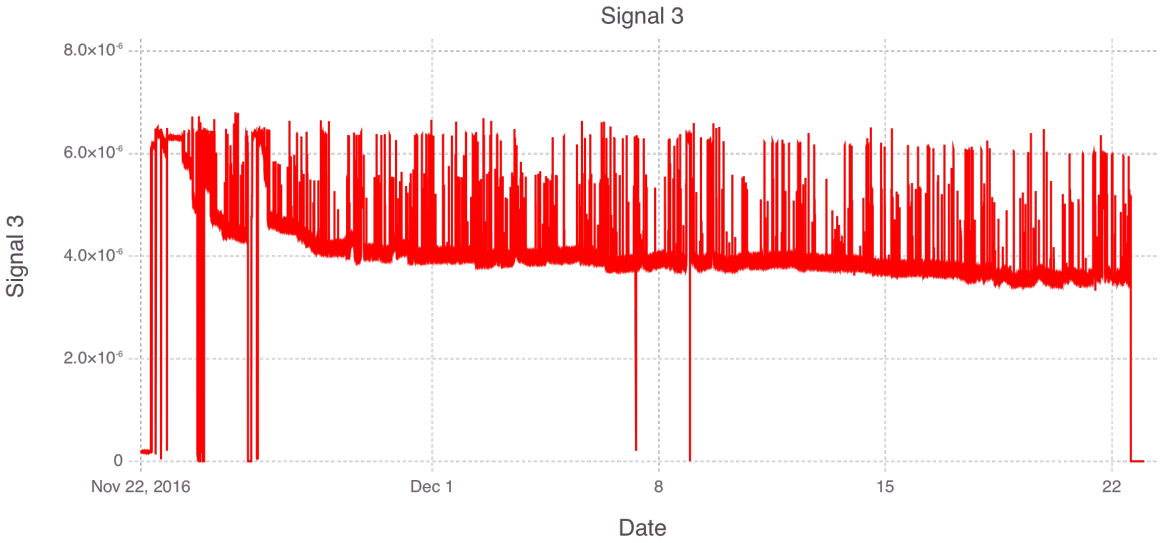
LANSCE: NMFk estimated signal #1 based on the dependent variables



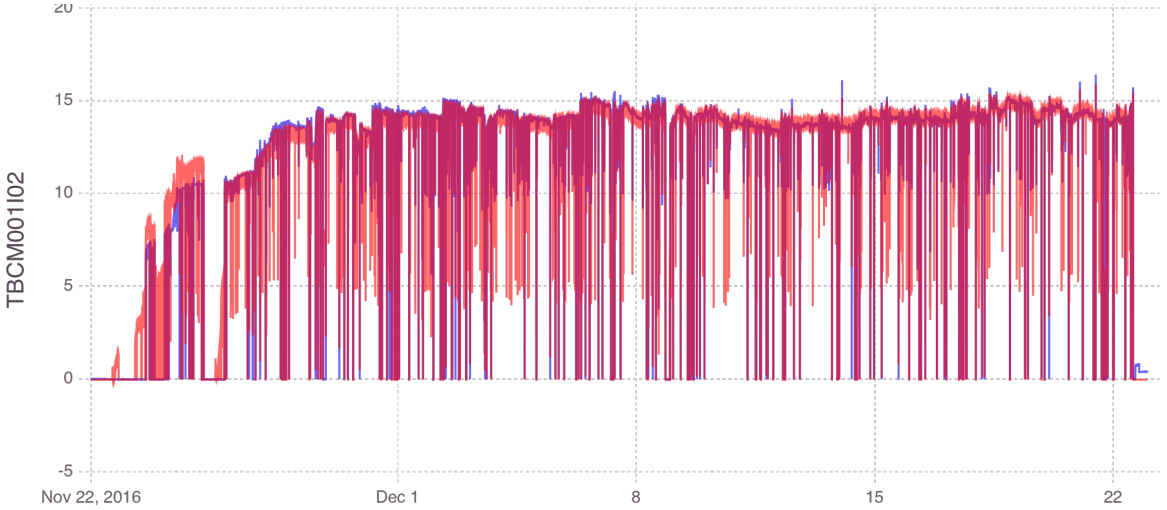
LANSCE: NMFk estimated signals #2 based on the dependent variables



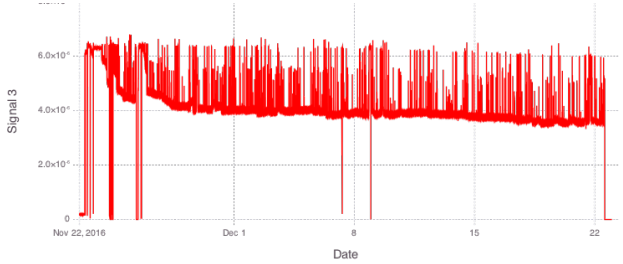
LANSCE: NMFk Estimated signal #3 based on the dependent variables

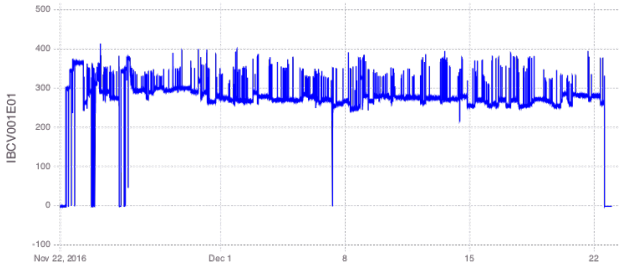


LANSCE: NMFk reconstruction of one of the dependent variables



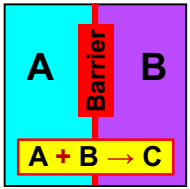
LANSCE: NMFk estimated signal #3 vs one of the independent variables



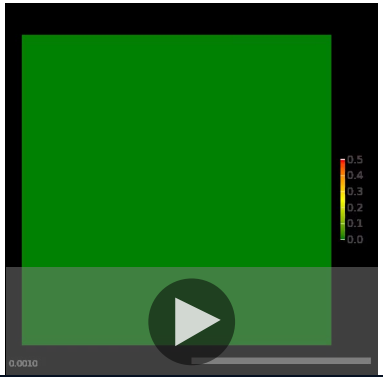


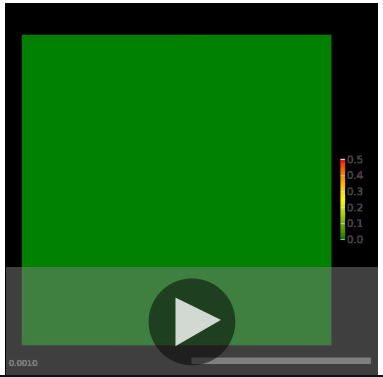
LANSCE: Reduced-order (SVR) Model

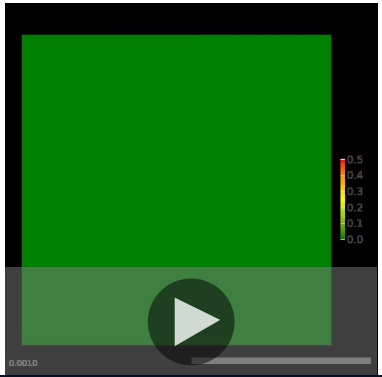
Reactive mixing



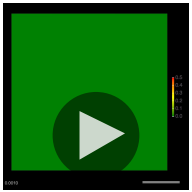
- > 2000 simulations of C concentrations in time/space with varying model inputs representing reactive mixing (5 input model parameters)
- **NTF**k identifies physics processes impacting C concentrations and their relationship to model inputs



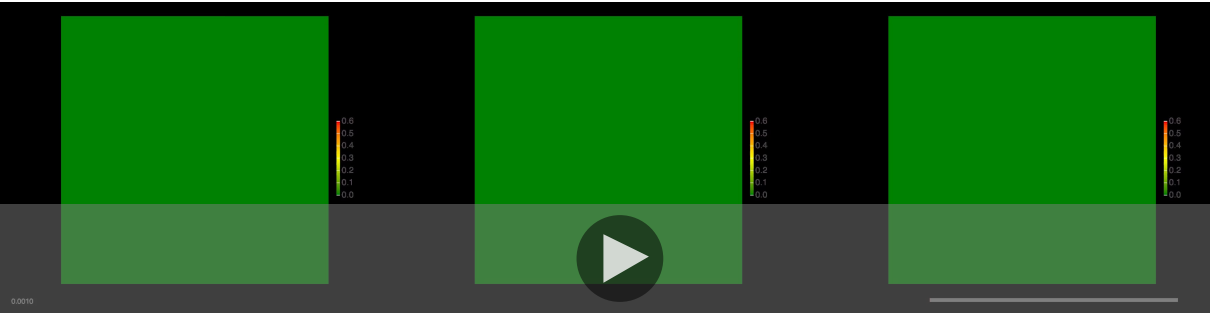




Reactive mixing: NTFk results

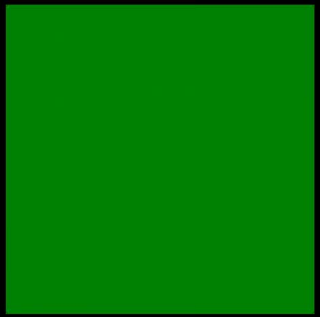


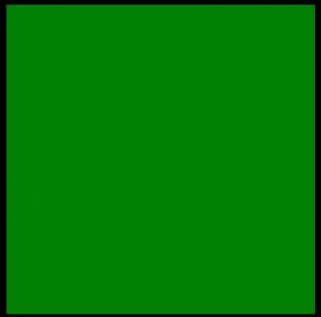
- NTFk extracts the dominant time/space features (**processes** / **vortices**) and compresses the model outputs
- Compression: > 200GB $\rightarrow \sim 70$MB (ratio ~ 3000
  Here, $(1000 \times 81 \times 81) \rightarrow (3 \times 12 \times 13)$ $(time \times rows \times columns)$

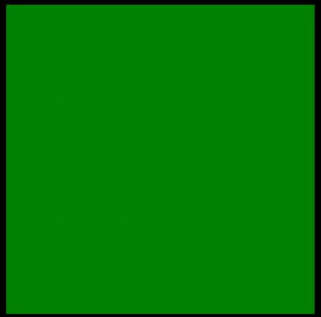


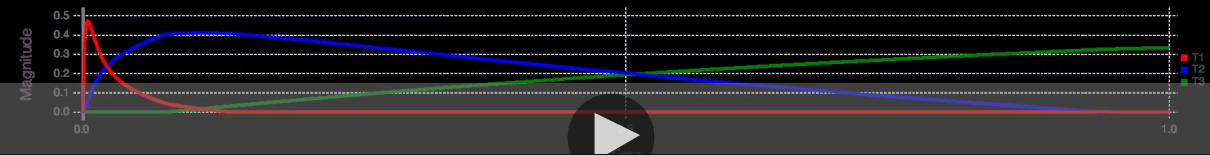
Advection **Dispersion** **Diffusion**

Reactive mixing: NTFk results

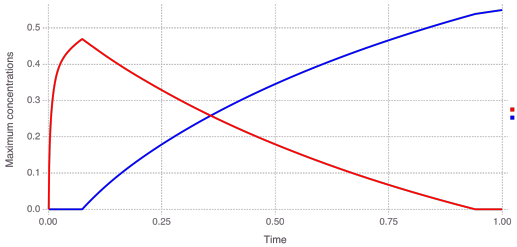


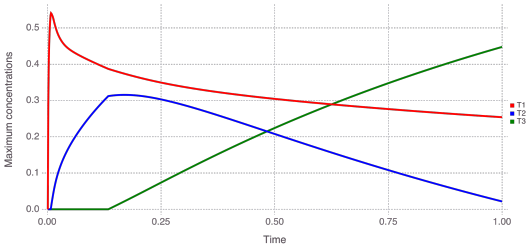


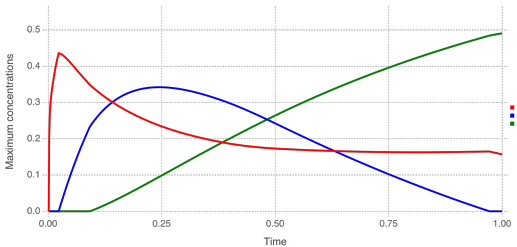




Reactive mixing: NTFk results ($\kappa_f L$ impacts)

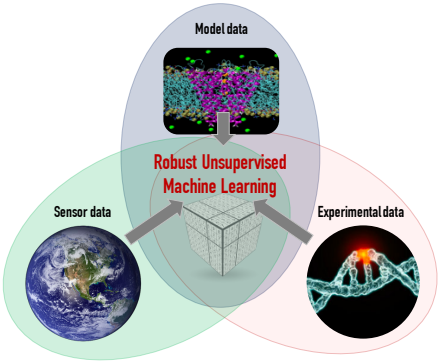






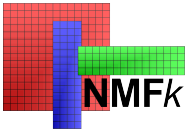
Summary

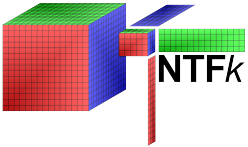
- Developed **novel** unsupervised ML methods and computational tools based on Nonnegative Factorization (Matrices/Tensors)
- Our ML methods have been used to solve various real-world problems (brought breakthrough discoveries related to human cancer research)
- Our ML work already contributes in many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



- Codes:

NMFk MADS NTFk





- Examples:

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