LA-UR-20-29279
Approved for public release; distribution is unlimited.

Title: Unsupervised and Physics-Informed Machine Learning Analyses for Characterization of Energy Production from Unconventional Reservoirs

Author(s): Vesselinov, Velimir Valentinov

Intended for: Machine Learning in Oil & Gas, 2020-11-09 (Houston, Texas, United States)
Web

Issued: 2020-11-12

Unsupervized and Physics-Informed Machine Learning Analyses for Characterization of Energy Production from Unconventional Reservoirs

Velimir V. Vesselinov (monty) (vvv@lanl.gov)

Earth and Environmental Sciences Division

Los Alamos National Laboratory, NM, USA

`http://tensors.lanl.gov`

Machine Learning (ML) methods

- **Supervised** ML: learns everything from data
  $\Rightarrow$ requires big training datasets
  $\Rightarrow$ highly impacted by noise

- **Physics-informed** ML: learns from data but includes preconceived knowledge about the governing processes
  $\Rightarrow$ requires smaller training datasets
  $\Rightarrow$ produces better predictability with lower uncertainty
  $\Rightarrow$ robust to data noise

- **Unsupervised** ML: extracts features from data that can be applied for categorization and prediction
  $\Rightarrow$ unbiased analyses not impacted by data labeling, subject-matter-expert opinions, and physics assumptions $\Rightarrow$ however, physics constraints can be added

Machine Learning (ML) methods

- **Supervised** ML: learns everything from data
  $\Rightarrow$ requires big training datasets
  $\Rightarrow$ highly impacted by noise
- **Physics-informed** ML: learns from data but includes preconceived knowledge about the governing processes
  $\Rightarrow$ requires smaller training datasets
  $\Rightarrow$ produces better predictability with lower uncertainty
  $\Rightarrow$ robust to data noise
- **Unsupervised** ML: extracts features from data that can be applied for categorization and prediction
  $\Rightarrow$ unbiased analyses not impacted by data labeling, subject-matter-expert opinions, and physics assumptions $\Rightarrow$ however, physics constraints can be added

Machine Learning (ML) methods

- **Supervised** ML: learns everything from data
  $\Rightarrow$ requires big training datasets
  $\Rightarrow$ highly impacted by noise
- **Physics-informed** ML: learns from data but includes preconceived knowledge about the governing processes
  $\Rightarrow$ requires smaller training datasets
  $\Rightarrow$ produces better predictability with lower uncertainty
  $\Rightarrow$ robust to data noise
- **Unsupervised** ML: extracts features from data that can be applied for categorization and prediction
  $\Rightarrow$ unbiased analyses not impacted by data labeling, subject-matter-expert opinions, and physics assumptions $\Rightarrow$ however, physics constraints can be added

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

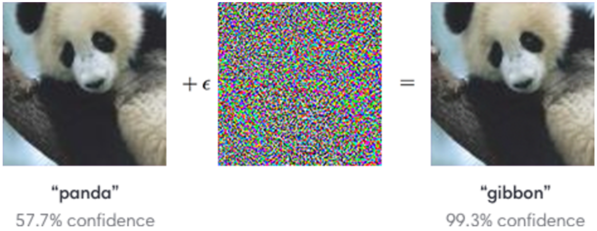
  Cannot discover something that we do not know already

- **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** ML methods for **science** applications

Unsupervised Machine Learning Applications: Data Analytics / Model Diagnostics

- Feature extraction (**FE**)
- Blind source separation (**BSS**)
- Detection of disruptions / anomalies
- Image recognition
- Separate physics processes
- Discover unknown dependencies and phenomena
- Develop reduced-order/surrogate models
- Identify dependencies between model inputs and outputs
- Guide development of physics models representing the data
- Make predictions
- Optimize data acquisition
- “Label” datasets for supervised ML analyses

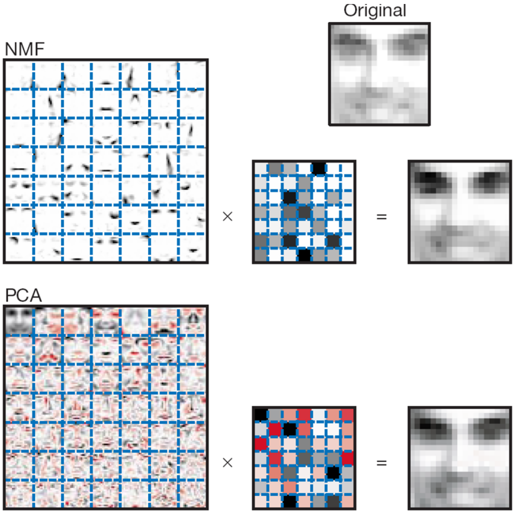
Nonnegative Matrix/Tensor Factorization

- Novel LANL-patented, open-source, unsupervised Machine Learning (ML) methods and computational techniques
- Based in matrix/tensor factorization coupled with custom k-means clustering and nonnegativity/sparsity constraints:
  - NMFk: Nonnegative **Matrix** Factorization
  - NTFk: Nonnegative **Tensor** Factorization
  - `https://github.com/TensorDecompositions`
- Capable to efficiently process large datasets (TB's) utilizing GPU's, TPU's & FPGA's

  $\Rightarrow$ julia, Flux.jl, AutoOffLoad.jl, TensorFlow, PyTorch, MXNet

Why nonnegativity?

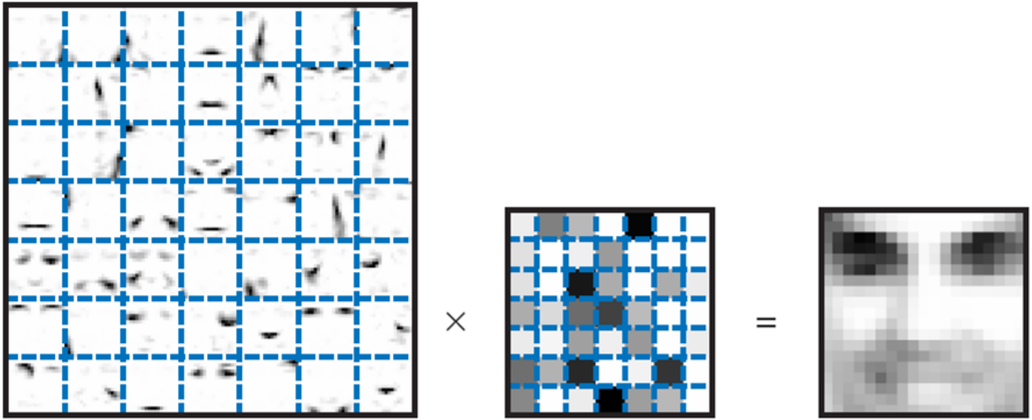
- NMF vs PCA (Lee & Seung, 1999)
- NMF: Nonnegative Matrix Factorization
- PCA: Principal Component Analysis



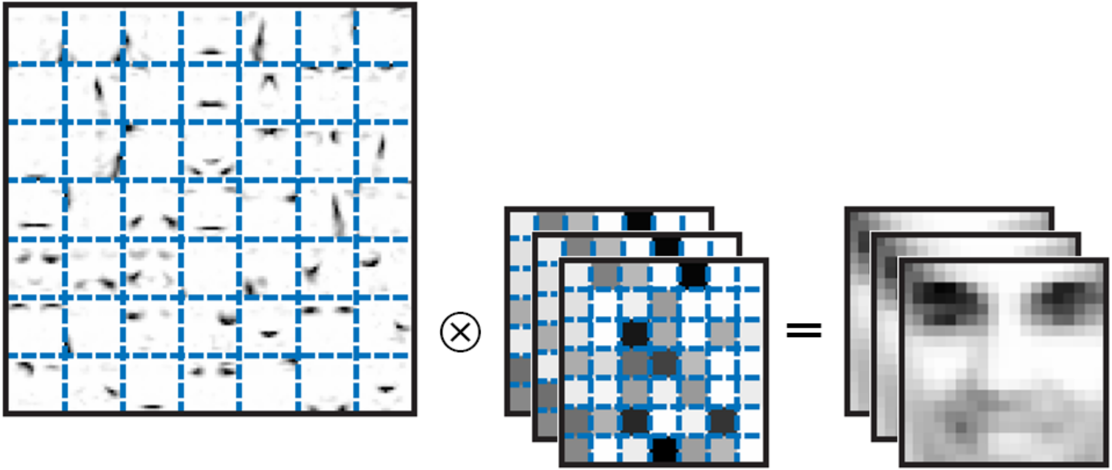
Nonnegativity constraints provide meaningful and interpretable results (+sparsity)

- **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

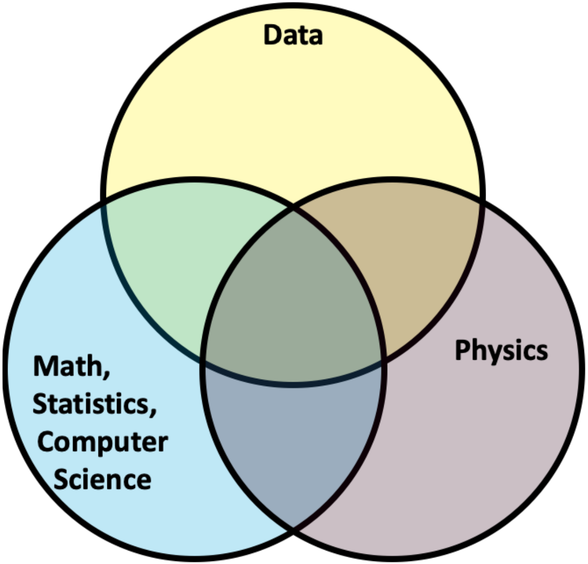
NMF: Nonnegative Matrix Factorization



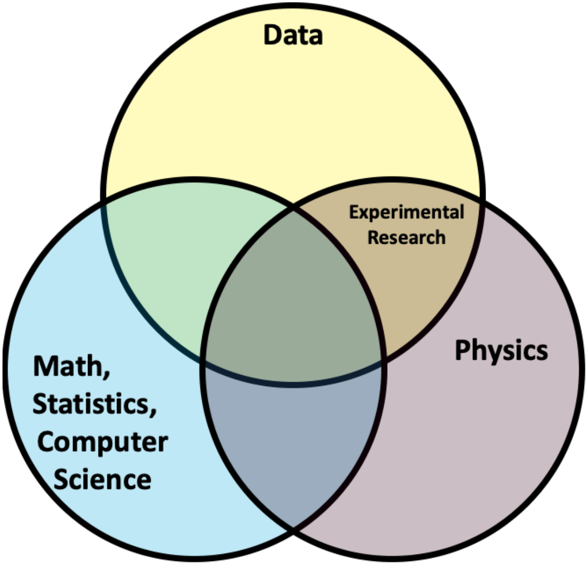
NTF: Nonnegative Tensor Factorization



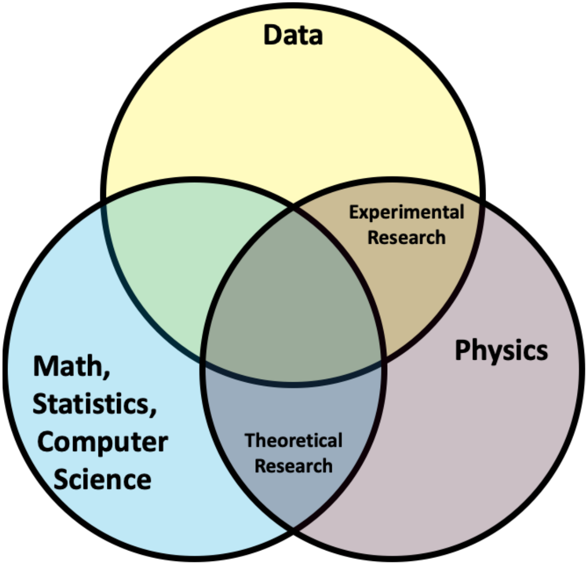
Physics-Informed Machine Learning



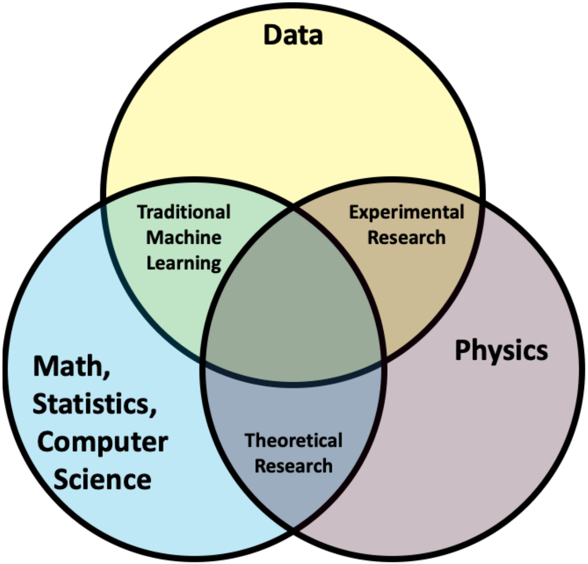
Physics-Informed Machine Learning



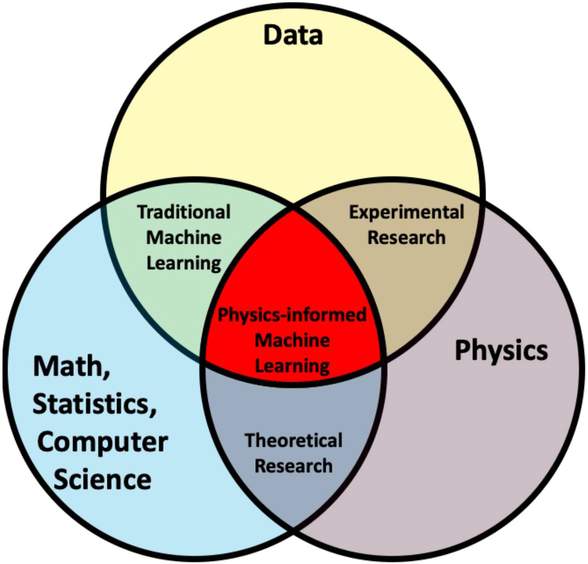
Physics-Informed Machine Learning



Physics-Informed Machine Learning

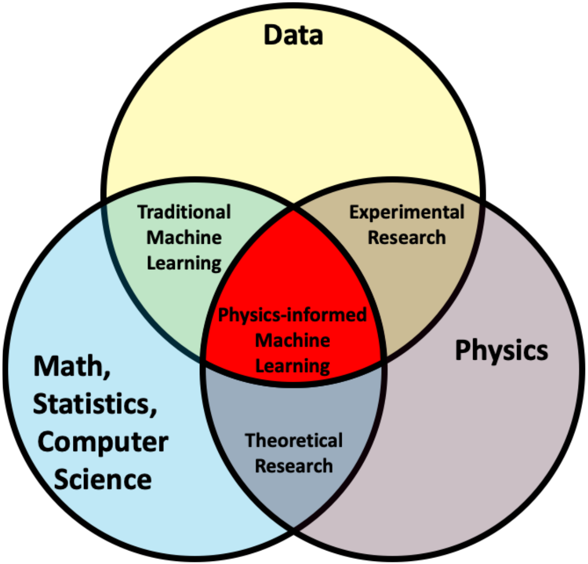


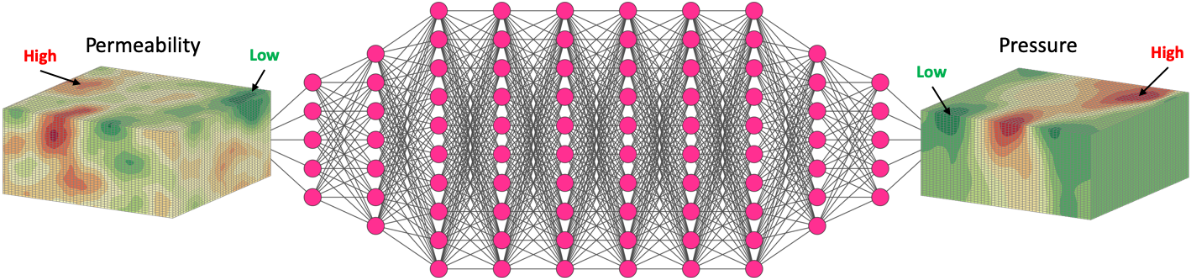
Physics-Informed Machine Learning



Science research paradigms

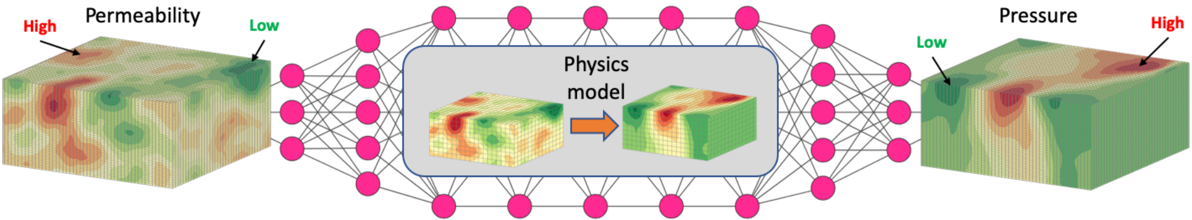
- **Empirical**: observations and experiments (since the cradle of our civilization)
- **Theoretical**: generalizations and models (since 1600's)
- **Computational**: analytical and numerical simulations (since 1950's)
- **Data-exploration**: unify data, simulations, and theory (since 2000's)





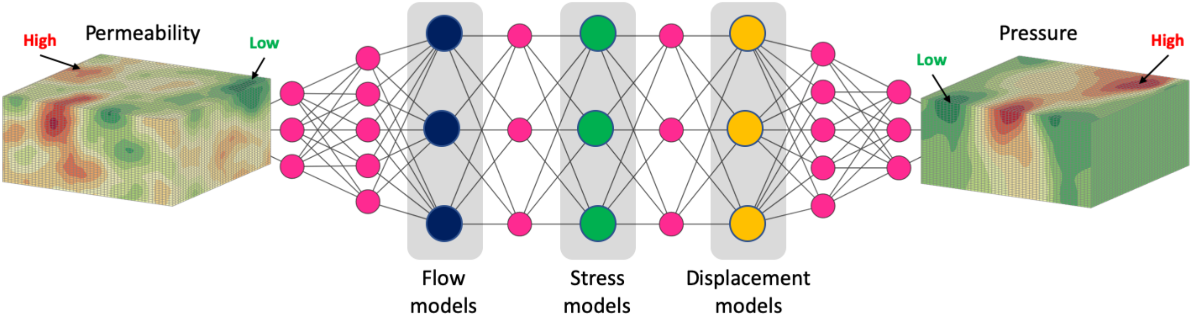
- it is a **black box**, **ad hoc** approach
- no preconceived knowledge about analyzed problem (general)
- all the neurons are $relu(Ax + b)$; A and b have no physical meaning; $relu()$ does not impose physics constraints
- neural networks needs to be very **deep** and **wide** to represent complex physics

Physics-Informed Machine Learning Neural Networks



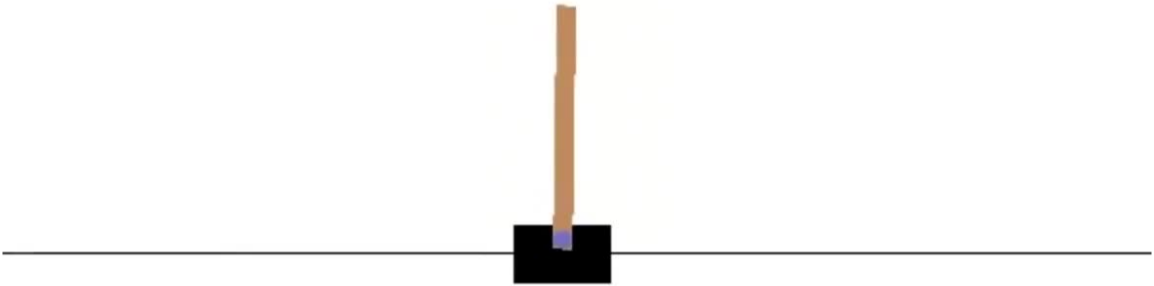
- include preconceived knowledge about analyzed problem (problem specific)
- neurons can represent $PhysicsModel(Ax + b)$; A and b have physical interpretation; $PhysicsModel()$ imposes physics constraints (e.g. conservation of mass/species)
- **PIML** models can be **trained (optimized) faster** and with **less training data**

Physics-Informed Machine Learning Neural Networks



- physics-informed layers (**“fat” neurons**) capture important governing processes (e.g., flow, stress, deformation, and displacement)
- can be done efficiently only through differentiable programming in julia

Physics-Informed Machine Learning (PIML): Cartpole

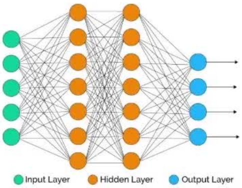


Physics-Informed Machine Learning (PIML): Cartpole

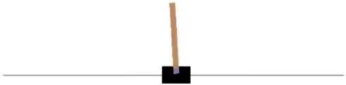
Physics-Informed Machine Learning (PIML): Cartpole

CartPole State → Neural Network → Control Parameters → Environment → Loss

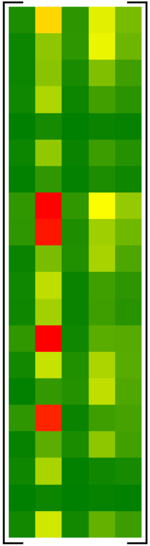
angle = -3°
velocity = 0.5°/s



{left, right}



angle^2



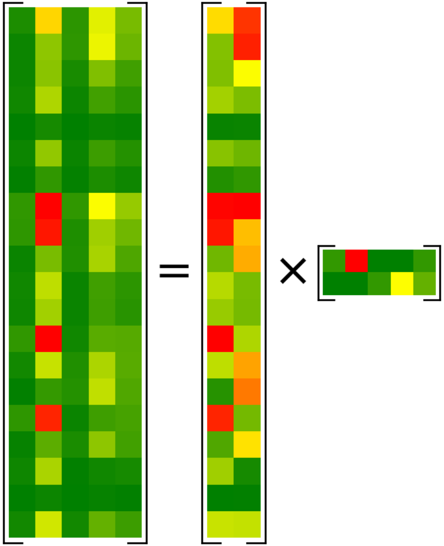
$\boldsymbol{X}$

$[20 \times 5]$

$\boldsymbol{X}$ – **data** matrix

$[$**attributes** $\times$ **observations**$]$

NMFk: matrix factorization



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

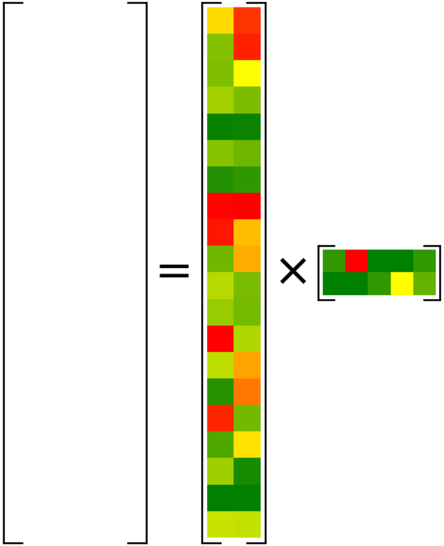
$$[20 \times 5] = [20 \times 2] \times [2 \times 5]$$

$\boldsymbol{X}$ – **data** matrix
[**attributes** $\times$ **observations**]

$\boldsymbol{W}$ – **feature** (**signal**) matrix
[**attributes** $\times$ **features**]

$\boldsymbol{H}$ – **mixing** matrix
[**features** $\times$ **observations**]

NMFk: true matrix factors



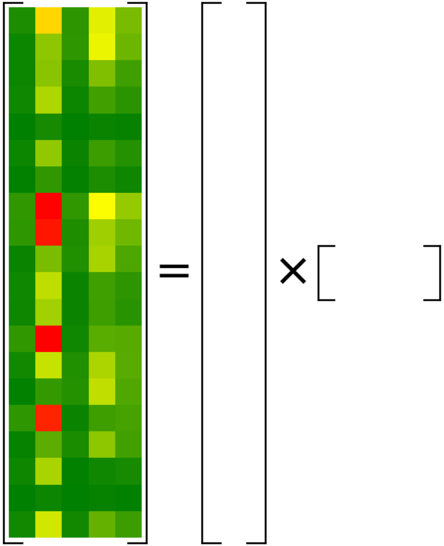
$$\boldsymbol{X} = \boldsymbol{W} \times \boldsymbol{H}$$

$$[20 \times 5] = [20 \times 2] \times [2 \times 5]$$

$\boldsymbol{X}$ – **data** matrix
[**attributes** $\times$ **observations**]

$\boldsymbol{W}$ – **feature** (**signal**) matrix
[**attributes** $\times$ **features**]

$\boldsymbol{H}$ – **mixing** matrix
[**features** $\times$ **observations**]



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

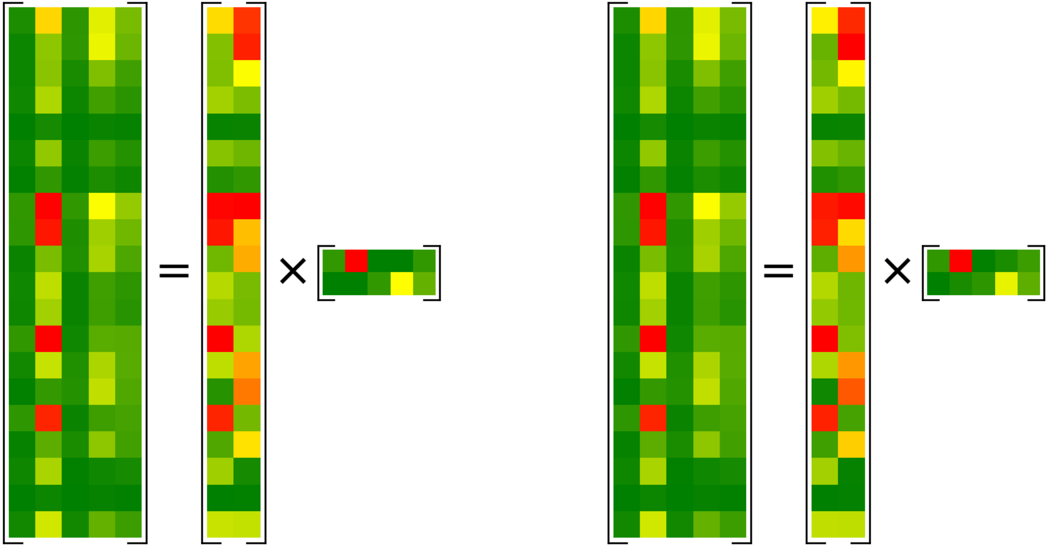
$$[20 \times 5] = [20 \times \textbf{?}] \times [\textbf{?} \times 5]$$

$\Rightarrow$ 100 **knowns**

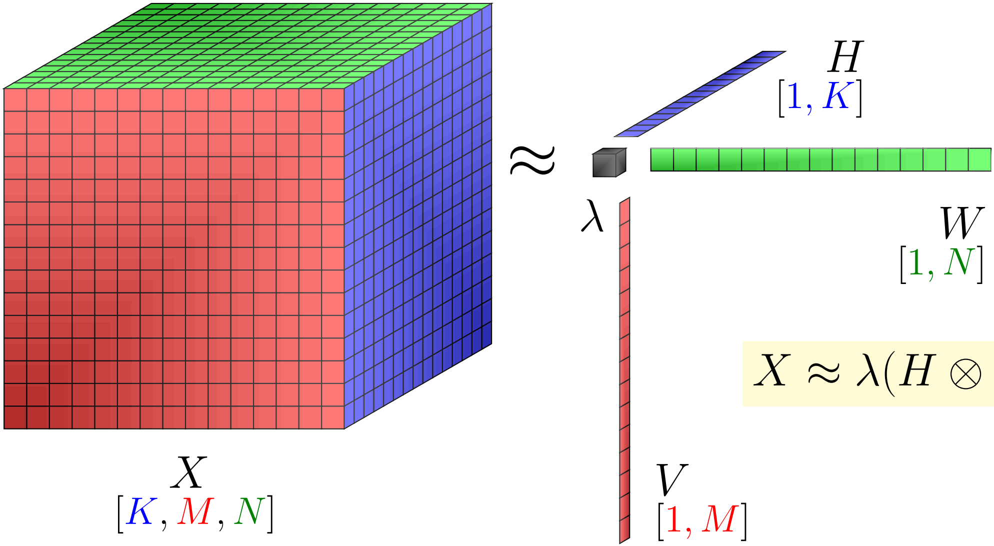
$\Rightarrow$ **unknown** number of features (2 or more)

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

NMFk: true vs. estimated matrix factorization

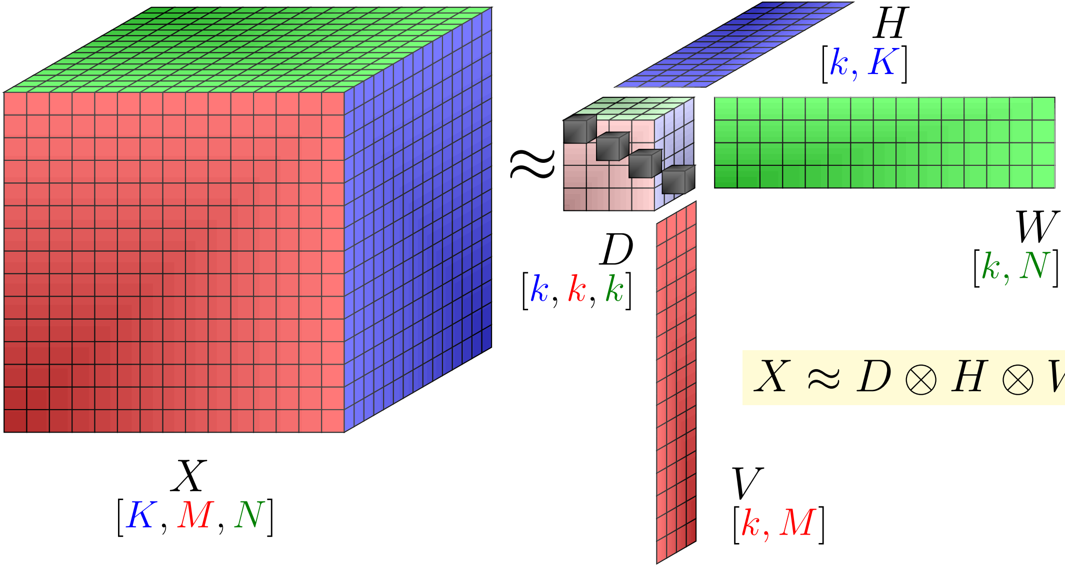


Tensor Decomposition (3D case): Rank-1 tensor



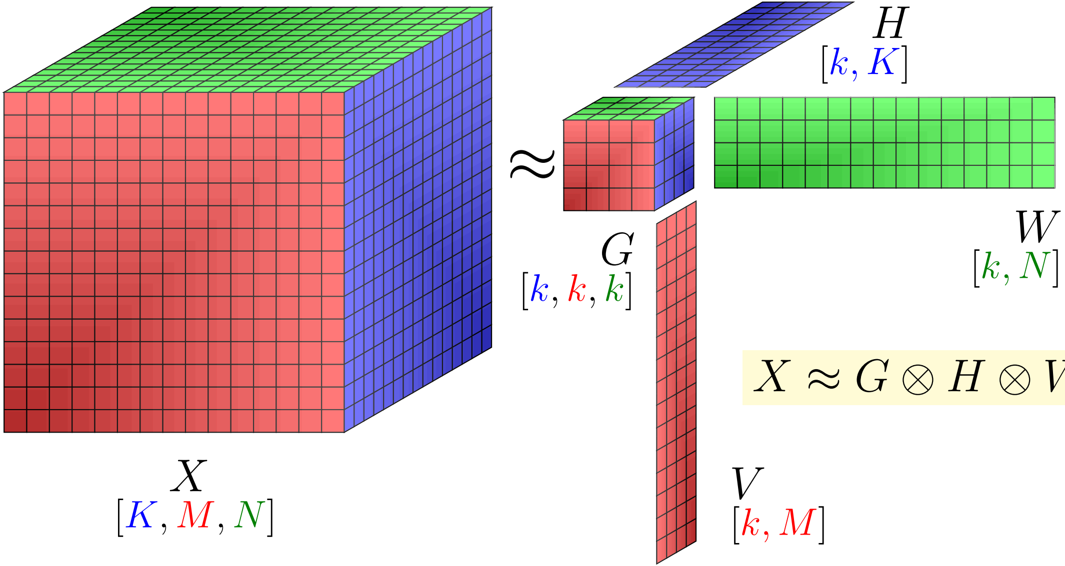
$$X \approx \lambda(H \otimes W \otimes V)$$

Tensor Decomposition (3D case): Rank-4 tensor



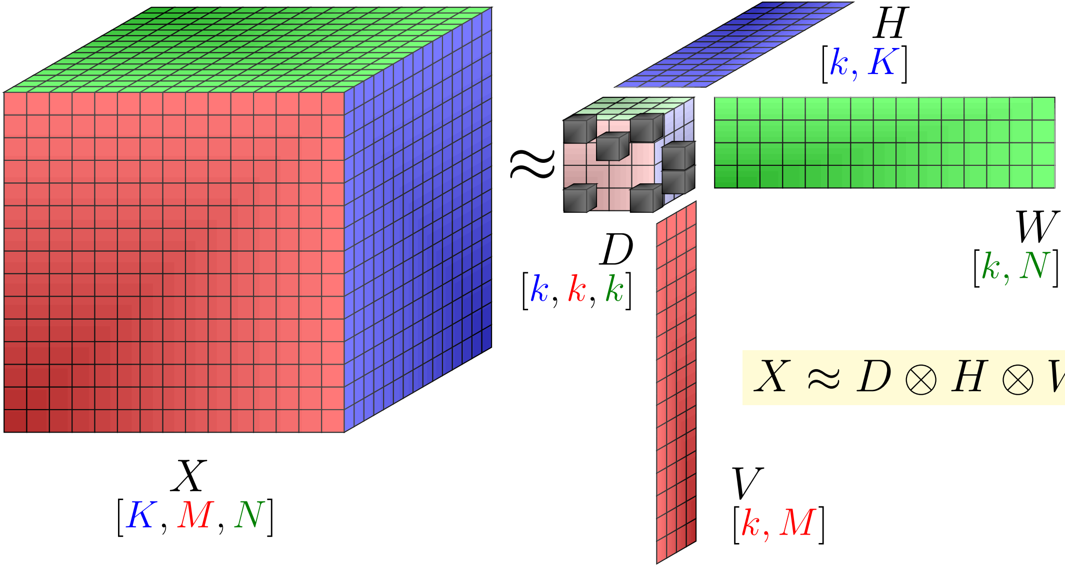
$$X \approx D \otimes H \otimes W \otimes V$$

Tensor Decomposition (3D case): Rank-64 / Multirank-(4,4,4) tensor



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

Tucker Tensor Decomposition (3D case): Rank-7 Multirank-(3,3,4)



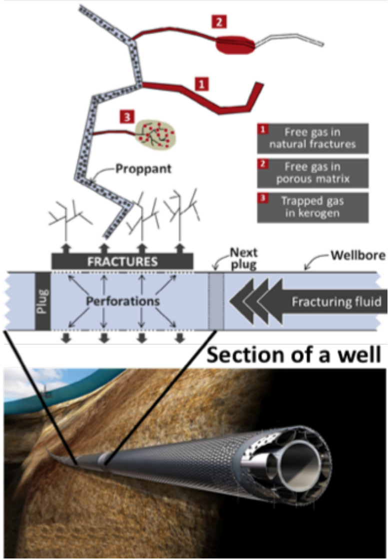
$$X \approx D \otimes H \otimes W \otimes V$$

NMFk / NTFk Analyses

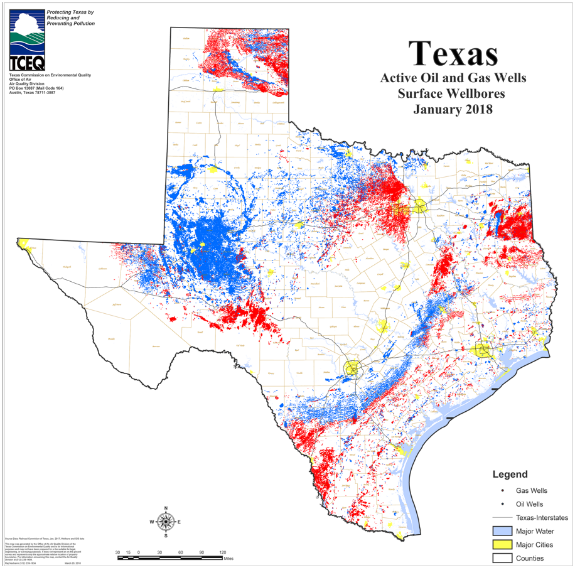
- **Field Data**:
  - Contamination
  - Climate
  - Geothermal
  - Seismic
  - Oil/gas production
- **Lab Data**:
  - X-ray Spectroscopy
  - UV Fluorescence Spectroscopy
  - Microbial population analyses
  - Isotope fractionation
- **Operational Data**:
  - LANSCE: Los Alamos Neutron Accelerator
  - Oil/gas production
- **Model Outputs**:
  - Reactive mixing $A + B \rightarrow C$
  - Phase separation of co-polymers
  - Molecular Dynamics of proteins
  - Climate modeling

Oil/Gas Production

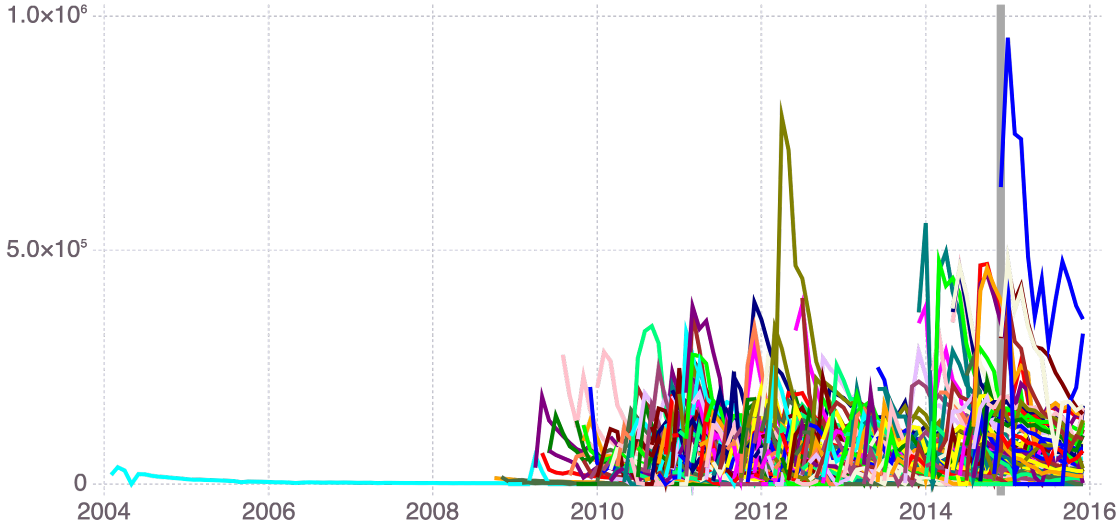
- Oil/Gas production from unconventional reservoirs extracts a small portion of the available resources (<10%)
- Oil/Gas production is challenging to predict and optimize
- Physics processes during well development (including hydrofracking) and extraction are poorly understood and challenging to simulate
- Alternative is to learn to predict system behavior based on the observed oil/gas production at existing wells



- Large public datasets are available representing unconventional oil and gas production (U.S. and world wide)
- Data represent monthly production rates (oil, gas, water) + many other well attributes
- $\sim$ 2,000,000 wells in U.S.
- $>$ 300,000 wells in Texas
- $>$ 20,000 wells in Eagle Ford Shale Play
- 327 gas wells in Eagle Ford Shale Play selected for preliminary analyses



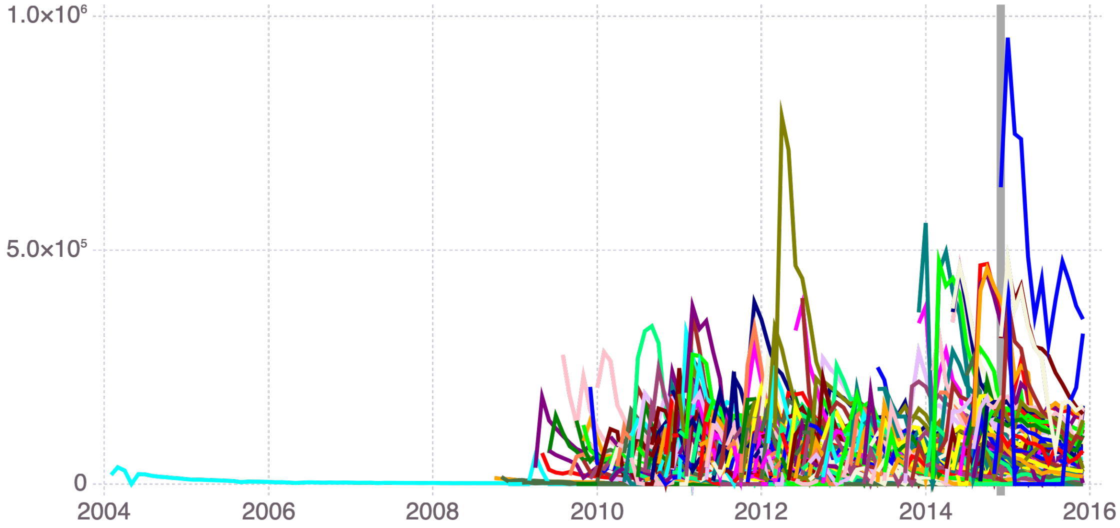
Eagle Ford Shale Play: Monthly production volumes [MCF] of 327 gas wells



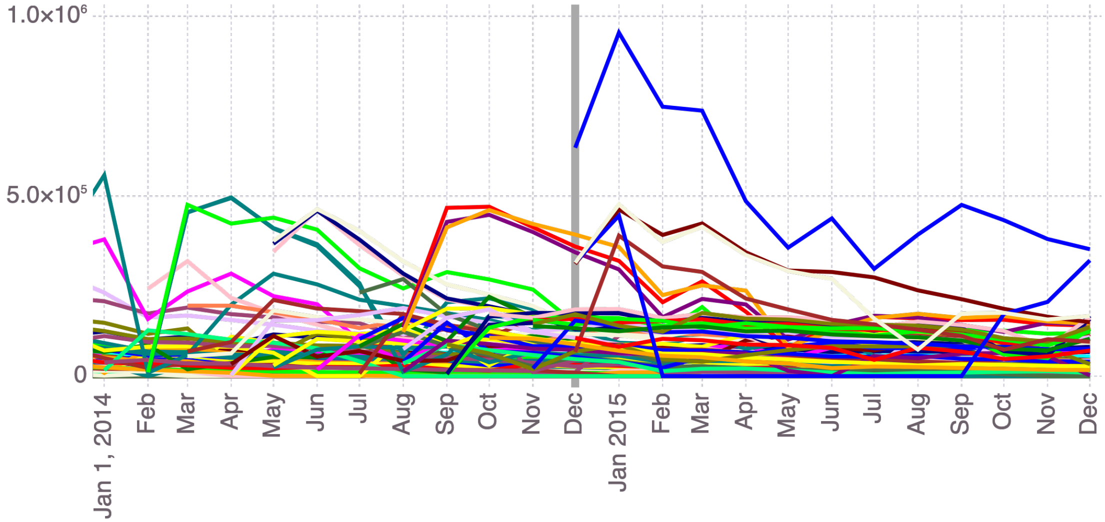
ML analysis of Oil/Gas Production Data

- Use all the data up to a given cutoff date (e.g. 2015)
- Apply ML to learn behavior of the “known“ well transients
  - Identify and group wells which behave similarly (having similar production transients)
  - Discover the optimal number of **master decline curves** required to represent the observed transients
  - **master decline curves** = production **features** or **signatures**
- Apply ML to predict **blindly** the unknown production transients beyond the cutoff
- Prediction is obtained by discovering to which type (group) the wells producing beyond the cutoff belong
- i.e., discovering what combinations of the **master decline curves** can represent the wells producing beyond the cutoff
- ML analyses performed using **NMFk**/**NTFk**

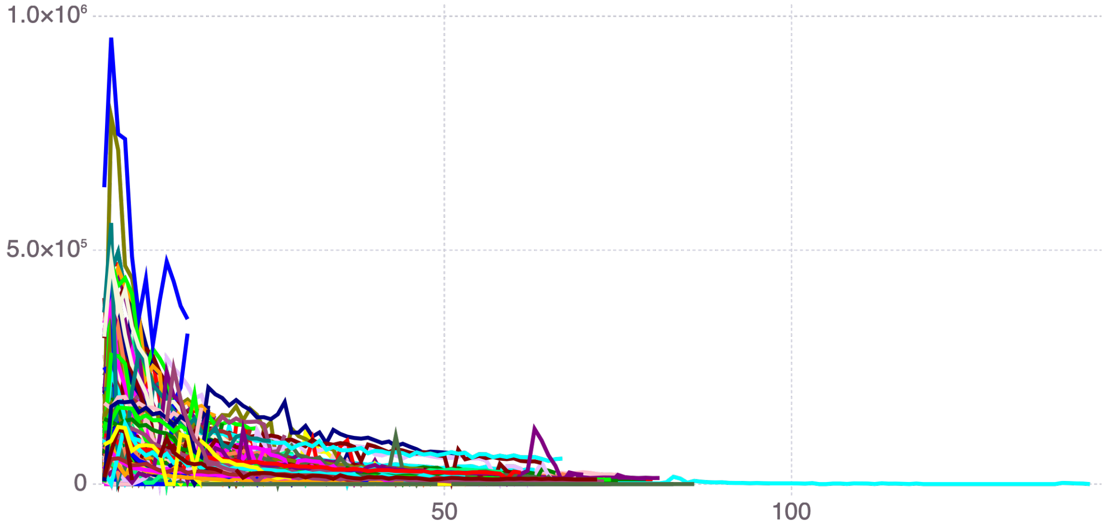
Eagle Ford Shale Play: Monthly production volumes [MCF] of 327 gas wells



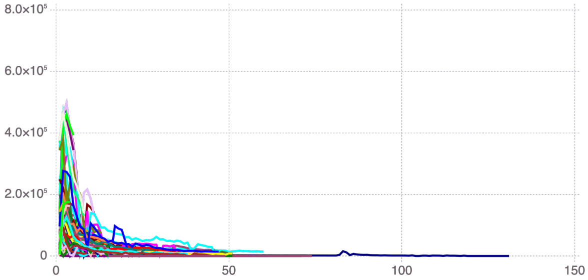
Eagle Ford Shale Play: Monthly production volumes [MCF] of 327 gas wells



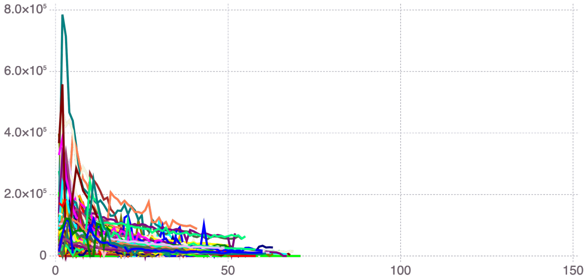
Eagle Ford Shale Play: Monthly production volumes [MCF] of 327 gas wells



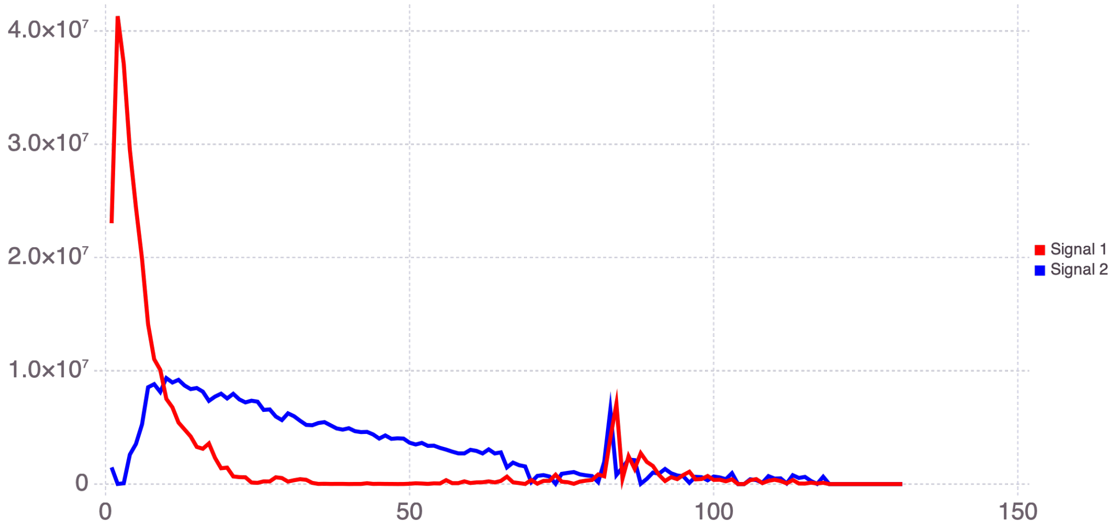
‘Fast” declining (135)



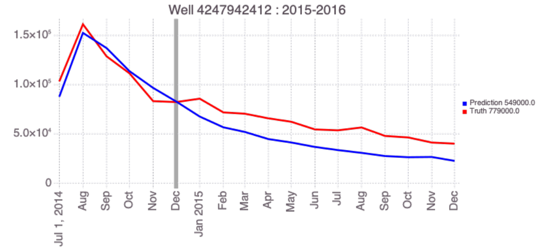
‘Slow” declining (192)

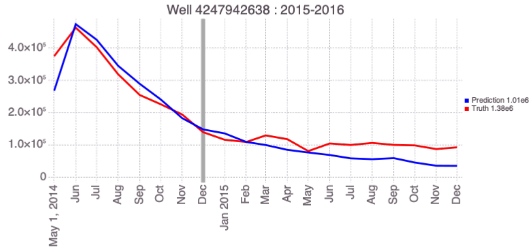


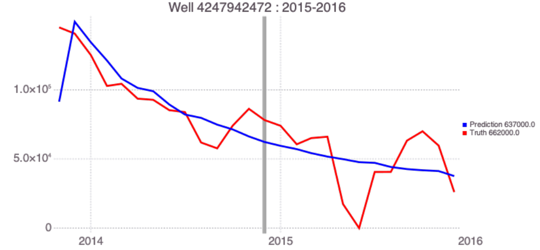
Eagle Ford Shale Play: Master Decline Curves [over months]

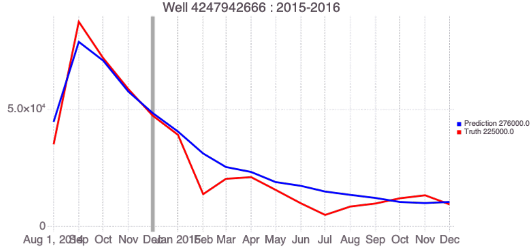


Eagle Ford Shale Play: Blind predictions beyond 2015

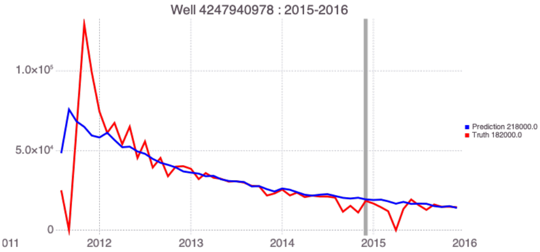


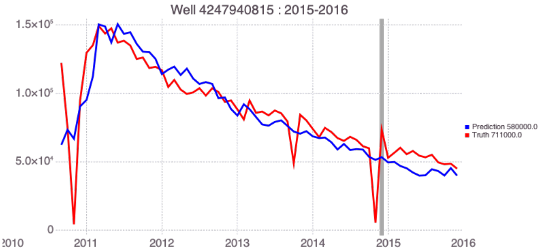


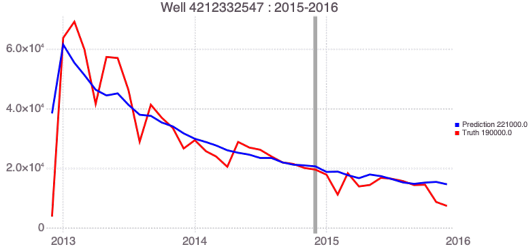


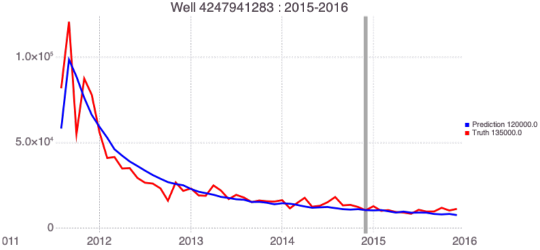


Eagle Ford Shale Play: Blind predictions beyond 2015





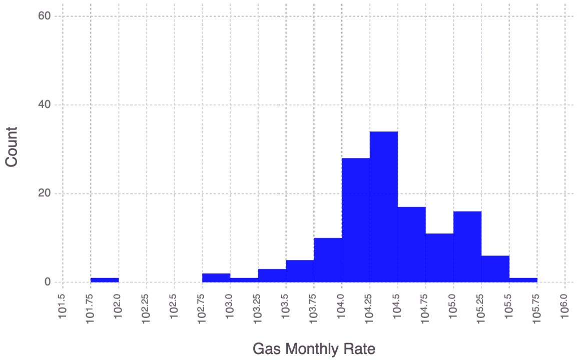




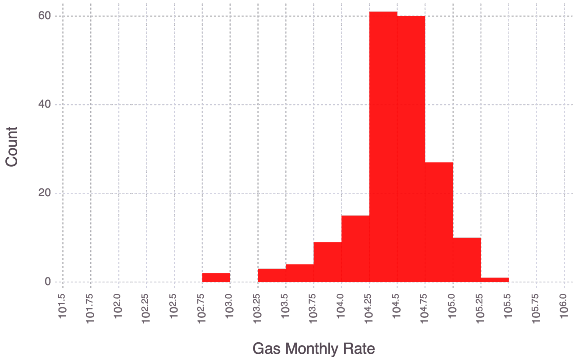
Eagle Ford Shale Play: Wells split into 2 groups

Monthly rate histograms

‘Fast” declining (135)

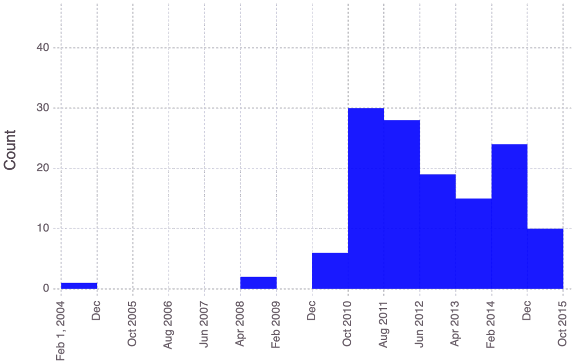


‘Slow” declining (192)

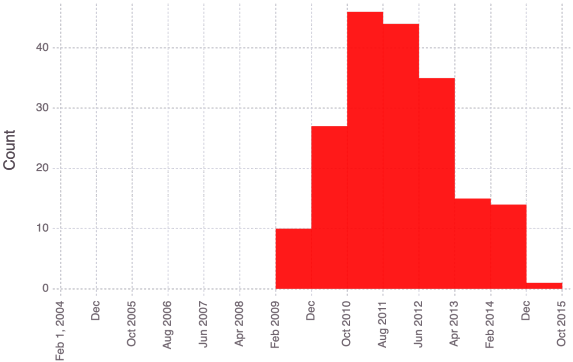


Drilling date histograms

'Fast" declining (135)



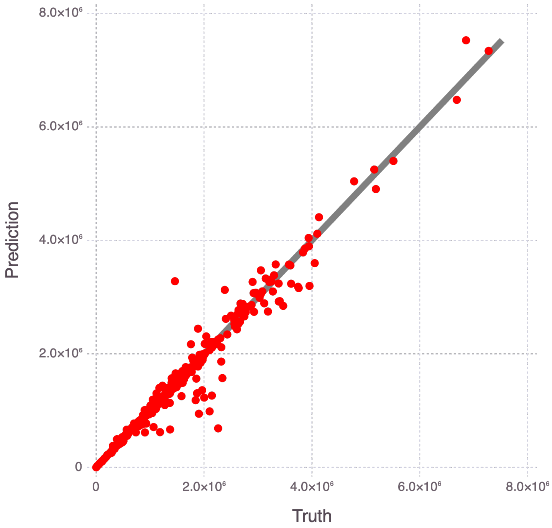
'Slow" declining (192)



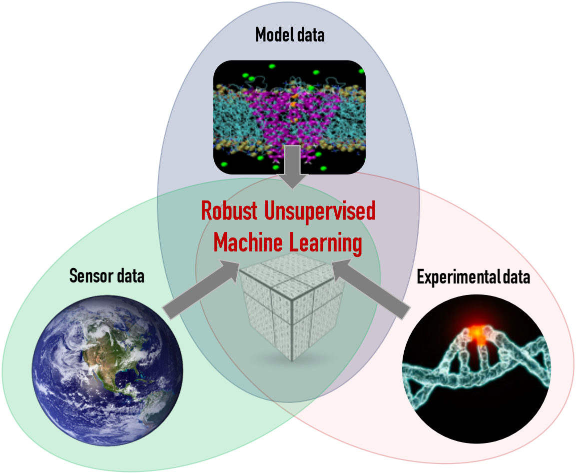
- Other well attributes also differ between the 2 groups
- For example:
  - Operators
  - Proppant mass
  - Injected fluid volumes
  - ...

Eagle Ford Shale Play: Blind predictions beyond 2015

- 300 wells continue producing beyond 2015
- $r^2 = 0.96$



- Developed **novel** unsupervised and physics-informed ML methods and computational tools
- Some of our tools have been recently patented
- Our ML methods have been used to solve various real-world problems (brought breakthrough discoveries related to human cancer research)
- Several ongoing projects (DOE, ARAP E, ...)



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

- Codes:

NMFk | MADS | NTFk

- Examples:

http://madsjulia.github.io/Mads.jl/Examples/blind_source_separation
http://tensors.lanl.gov
http://tensordecompositions.github.io
https://github.com/TensorDecompositions
https://hub.docker.com/u/montyvesselinov

Recent ML Publications

- Vesselinov, Munuduru, Karra, O'Maley, Alexandrov, Unsupervised Machine Learning Based on Non-Negative Tensor Factorization for Analyzing Reactive-Mixing, **Journal of Computational Physics**, Special issue: Machine Learning, 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**, 2017.
- Alexandrov, Vesselinov, Blind source separation for groundwater level analysis based on nonnegative matrix factorization, **WRR**, 2014.