UCLA

Towards Transparency, Fairness, and Efficiency in Machine Learning

- Deanna Needell
- Professor of Mathematics
  Executive Director, Institute for Digital Research and Education
  Dunn Family Endowed Chair in Data Theory
- UCLA

UCLA College
Mathematics

Collaborators

Erin George
UCLA

Michael Murray
UCLA

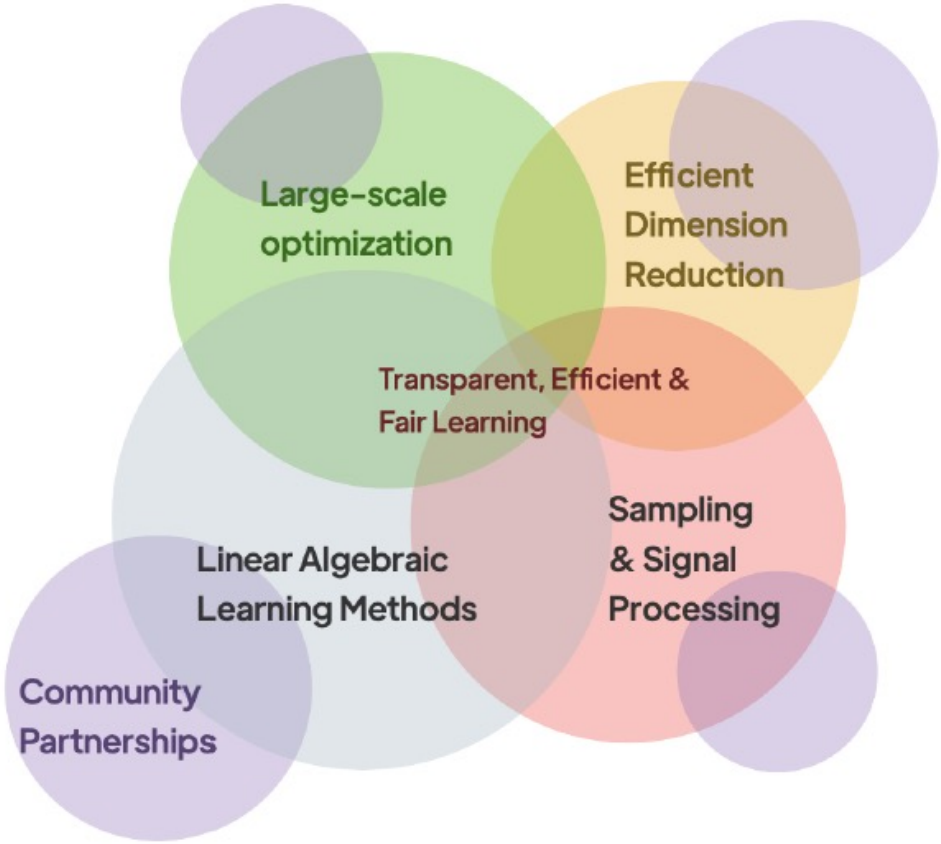
Will Swartworth
CMU

Today…

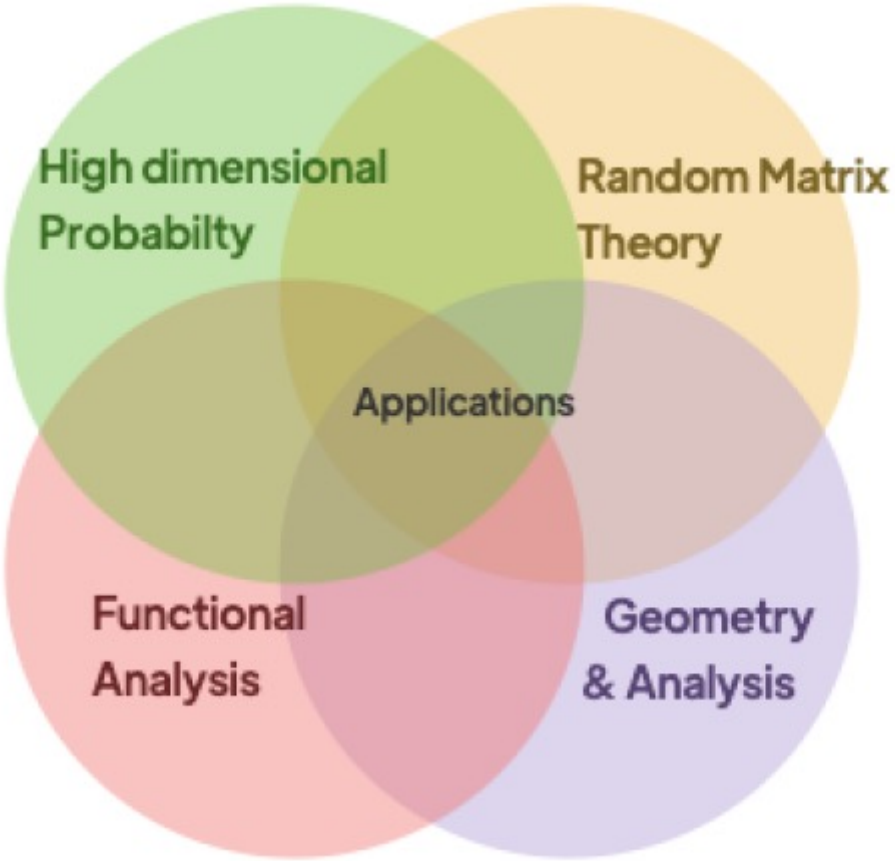
Transparency, Fairness, and Efficiency in Machine Learning

- **Transparency**:
  - Linear algebraic tools to promote transparency (NMF, CUR)
  - Understanding behavior in neural nets
- **Fairness**:
  - linear systems with latent subgroups, fair-NMF
- **Efficiency**:
  - tensorial dimension reduction with practical measurements

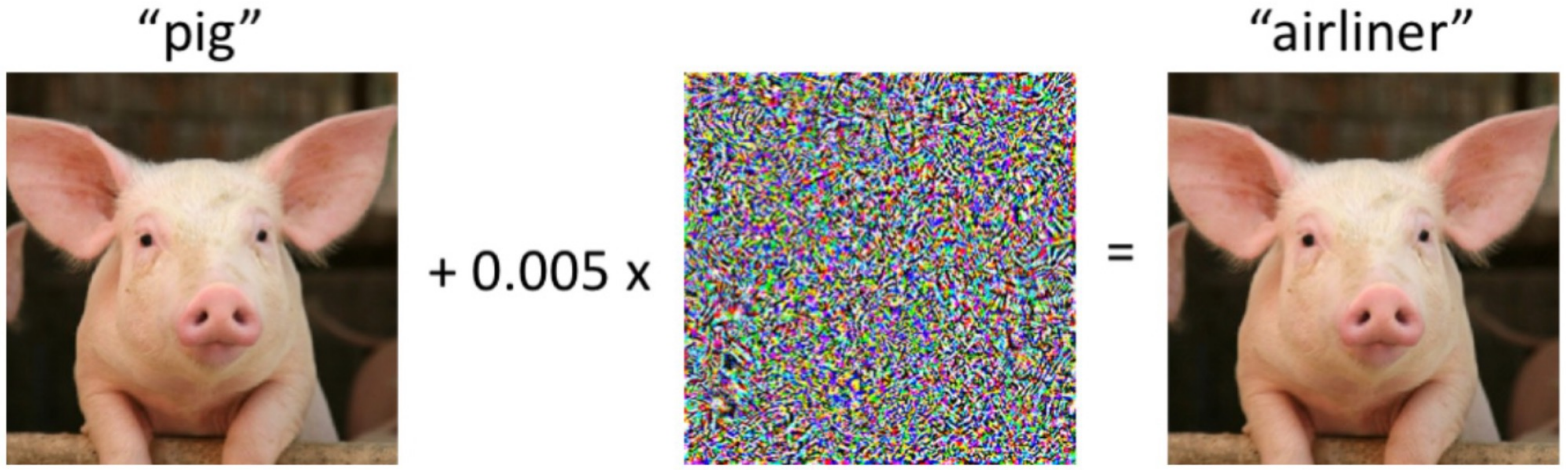
My research at a glance



My toolbelt at a glance



Towards transparency in ML



https://gradientscience.org/intro_adversarial/

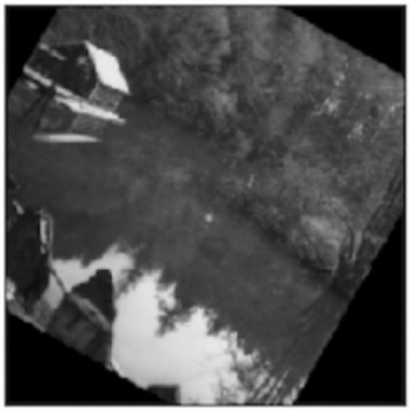
Towards transparency in ML

revolver


boathouse

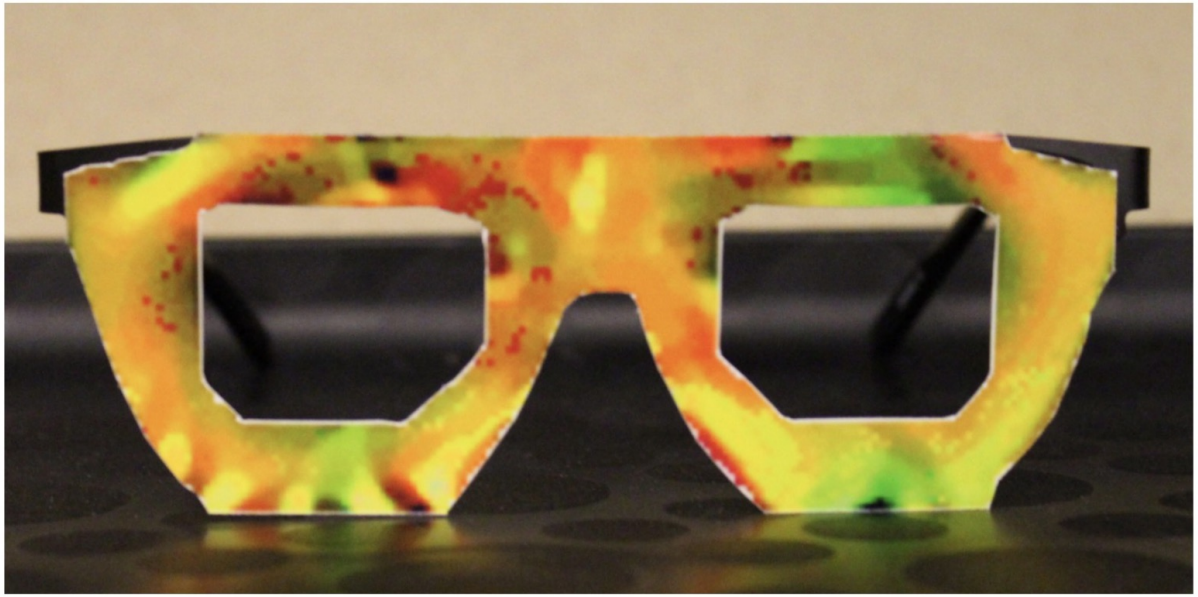

china cabinet

mousetrap


guillotine


spotlight

Towards transparency in ML



Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. Sharif et al.

Towards transparency in ML

Eykholt et al. Robust Physical-World Attacks on Deep Learning Visual Classification.

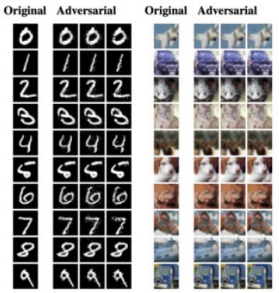
Towards transparency in ML

Getty Images, Siberian Art on Adobe Stock

Towards transparency … how to protect?

1) Training against adversaries: generate adversarial data that fools your network, then train against
   - ➢ Improves generalization but lacking robustness (new adversaries!)

2) Defensive distillation: use a second model "smoothed" in adversarial directions – trained on the primary model's output probabilities rather than thresholded decisions
   - ➢ Can be applied to any feed-forward neural network
   - ➢ Reduced prior attacks from 95% to 0.5% success
   - ➢ More robust against attacks, but fails current benchmarks [Carlini-Wagner]



Towards transparency in ML

Why do these models fail so "easily"?

- Data is very sparse in very high dimensional space → lots of room to "nudge" things around
- Models are often overconfident, especially in space they have little to go on
- Ian Goodfellow: "many of the most important problems still remain open, both in terms of theory and in terms of applications. We do not yet know whether defending against adversarial examples is a theoretically hopeless endeavor or if an optimal strategy would give the defender an upper ground. On the applied side, no one has yet designed a truly powerful defense algorithm that can resist a wide variety of adversarial example attack algorithms."

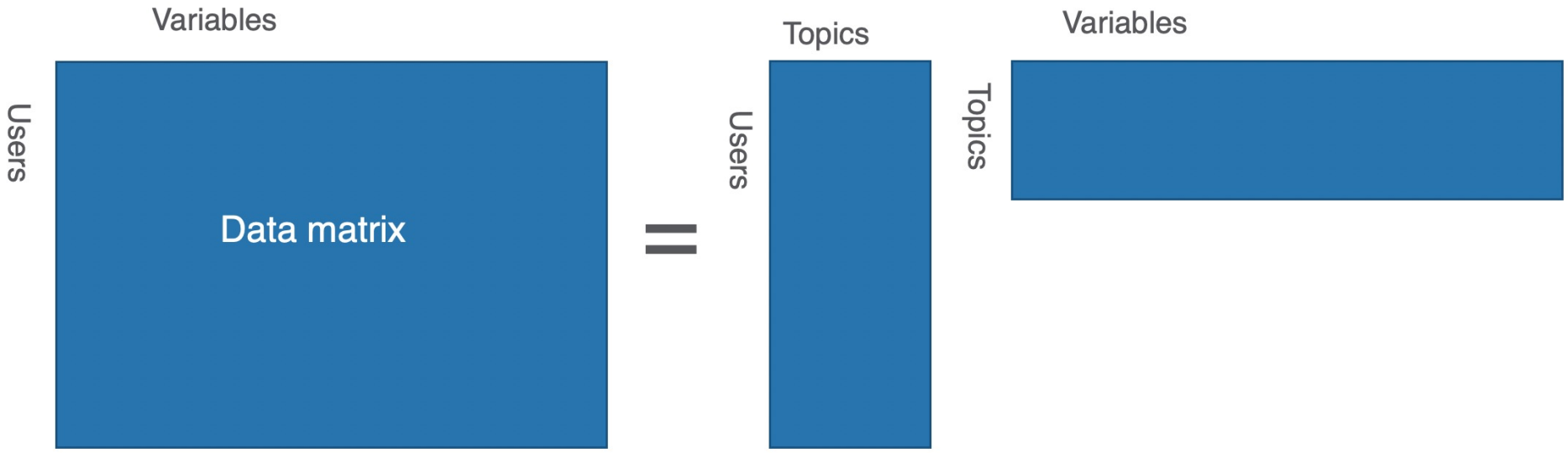
Towards transparency in ML

So how can we ever trust an output?

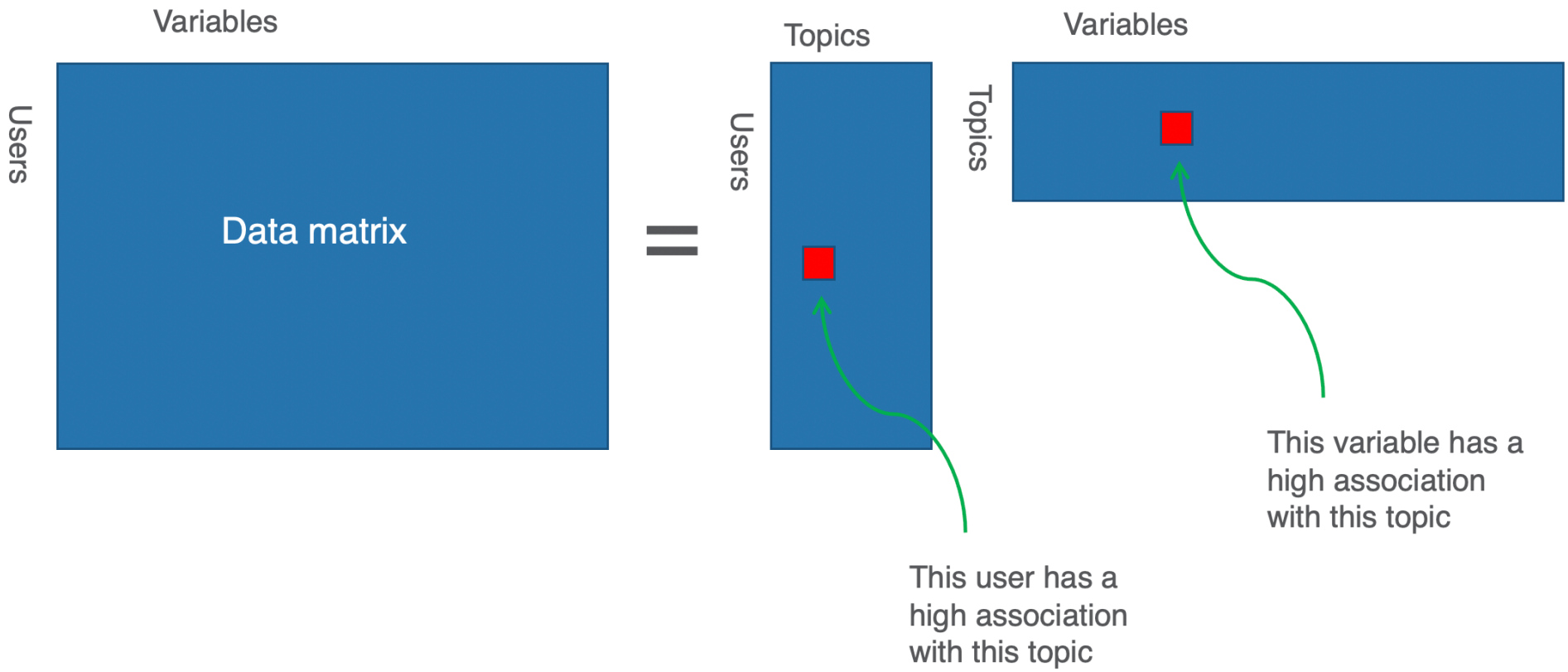
- If we could know "why" a model selects a particular output, we not only further our understanding of the method, but also begin to develop trust (cautiously)

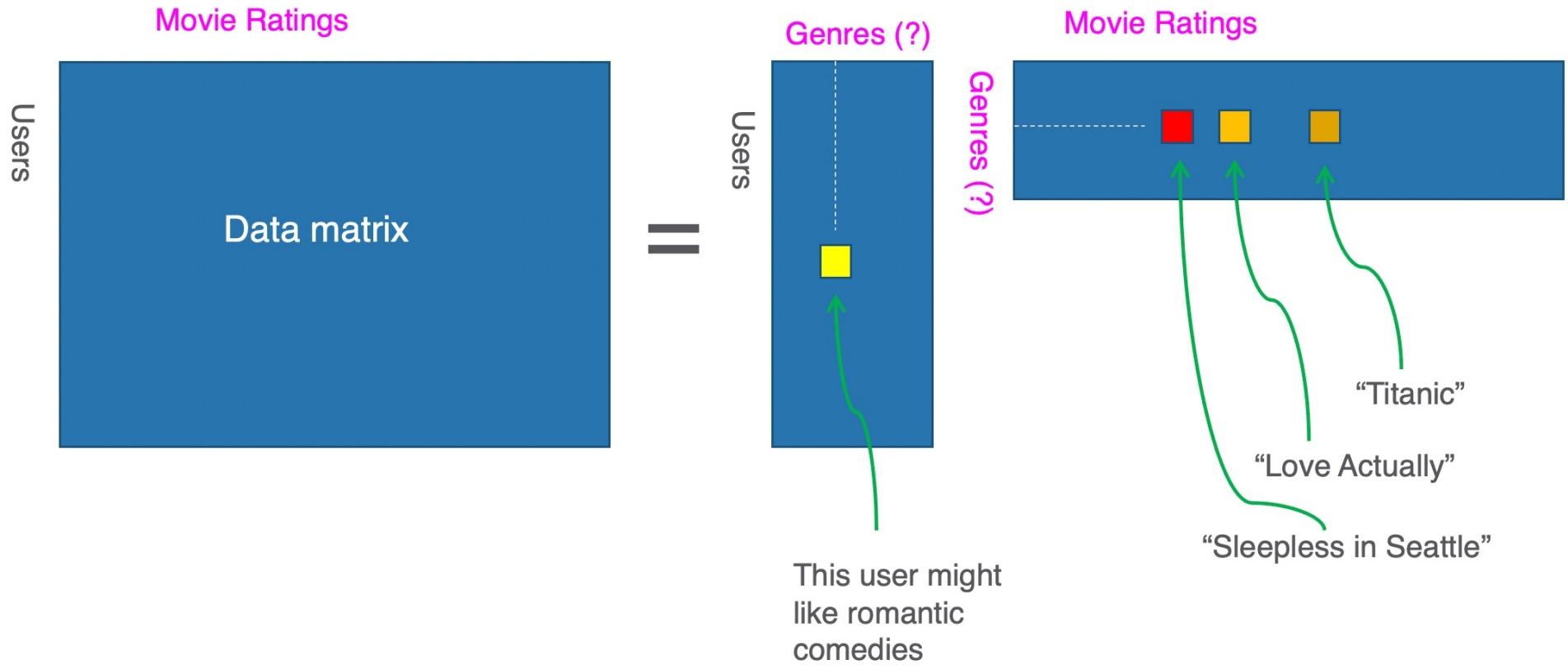
- There are other approaches that allow for this type of transparency

Non-negative matrix factorization

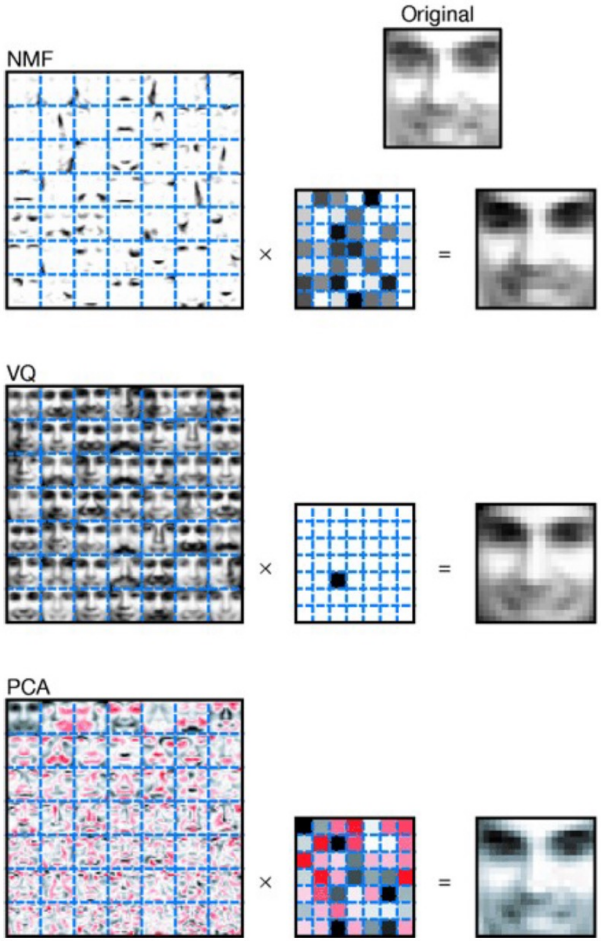


Non-negative matrix factorization




Movie Ratings
Users
Data matrix
=
Genres (?)
Users
Movie Ratings
Genres (?)
"Titanic"
"Love Actually"
"Sleepless in Seattle"
This user might like romantic comedies

Non-negative matrix factorization



- In NMF, each column of the data matrix has to be represented as a non-negative linear combination of dictionaries
- Hence the dictionaries must be "positive parts" of the columns of the data matrix
- When each column consists of a human face image, NMF learns the parts of human face (e.g., eyes, nodes, mouth)
- This is in contrast to principal component analysis and vector quantization: Due to cancellation between eigenvectors, each 'eigenface' does not have to be parts of face
- NMF was popularized by Lee and Seung in their Nature paper in 1999

Lee, D., Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999). https://doi.org/10.1038/44565

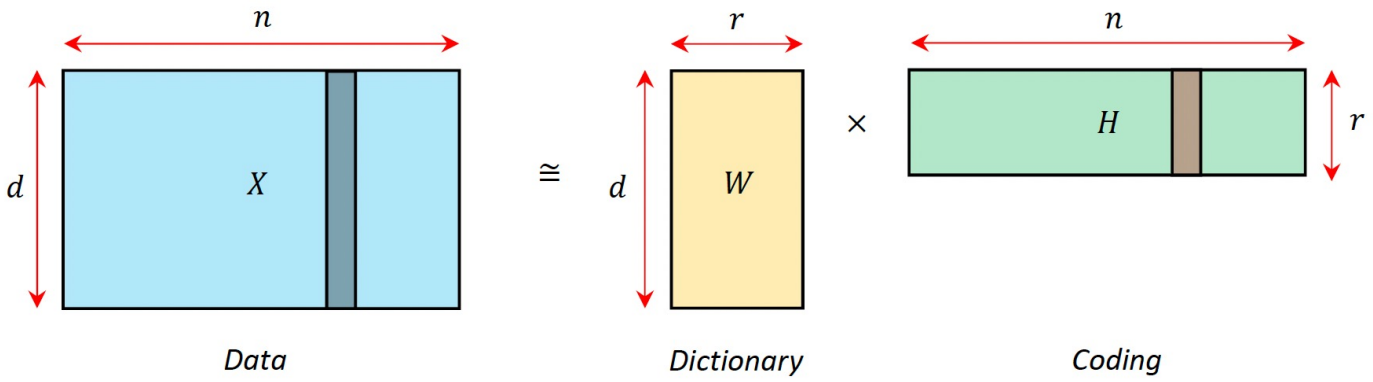
Non-negative matrix factorization

- The goal of **nonnegative matrix factorization** (NMF) is to factorize a data matrix $X \in \mathbb{R}^{d\times n}_{\geq 0}$ into a pair of low-rank nonnegative matrices $W \in \mathbb{R}^{d\times r}$ and $H \in \mathbb{R}^{r\times n}$ by solving the following optimization problem

$$\inf_{W\in\mathbb{R}^{d\times r}_{\geq 0},\, H\in\mathbb{R}^{r\times n}_{\geq 0}} \|X - WH\|_F^2,$$

  where $\|A\|_F^2 = \sum_{i,j} A_{ij}^2$ denotes the matrix Frobenius norm.

- $\texttt{Data} \approx \texttt{Dictionary} \times \texttt{Coding}$



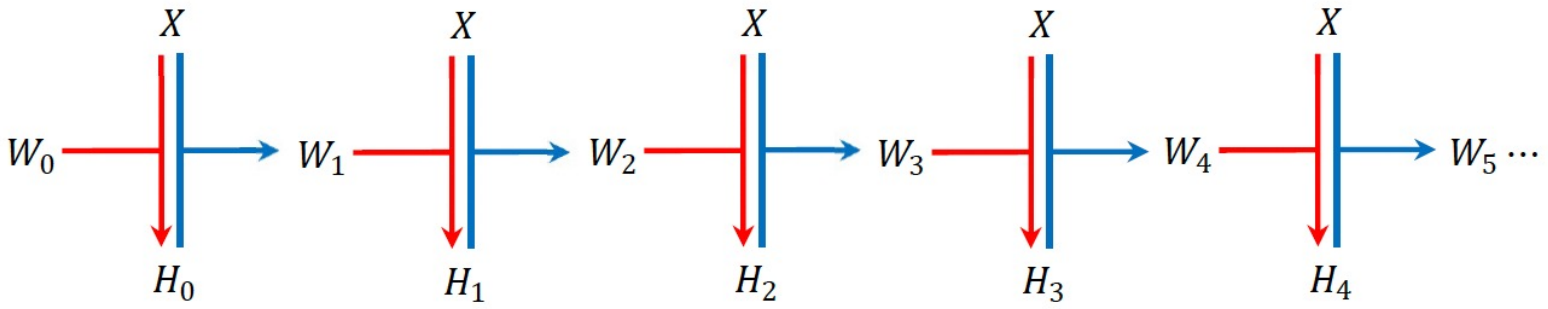
- Can be extended to tensors in a (nontrivial but) analogous way

Regularizers

- L1-regularizer for sparse topics/encodings $\|H\|_1$
- Divergences: $D(\mathbf{A}\|\mathbf{B}) = \sum_{i,j}\left(\mathbf{A}_{ij}\log\frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}} - \mathbf{A}_{ij} + \mathbf{B}_{ij}\right)$
- Classification loss: $\underset{\mathbf{W}\geqslant 0,\mathbf{H}\geqslant 0,\mathbf{C}\geqslant 0}{\operatorname{argmin}} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2 + \alpha\|\mathbf{M}\odot(\mathbf{Y}-\mathbf{C}\mathbf{H})\|_F^2$

Non-negative matrix factorization

- In order to minimize $\|X - WH\|_F$, one can use block coordinate descent, by iteratively fixing W or H and minimizing the error w.r.t. the other factor



- One of the most popular static NMF algorithm is the **Multiplicative Update** by Lee and Seung: Update all entries of H and W alternatively using the following update

$$H_{ij} \leftarrow H_{ij} \frac{[W^T X]_{ij}}{[W^T W X]_{ij}}, \qquad W_{ij} \leftarrow W_{ij} \frac{[X H^T]_{ij}}{[X H H^T]_{ij}}.$$

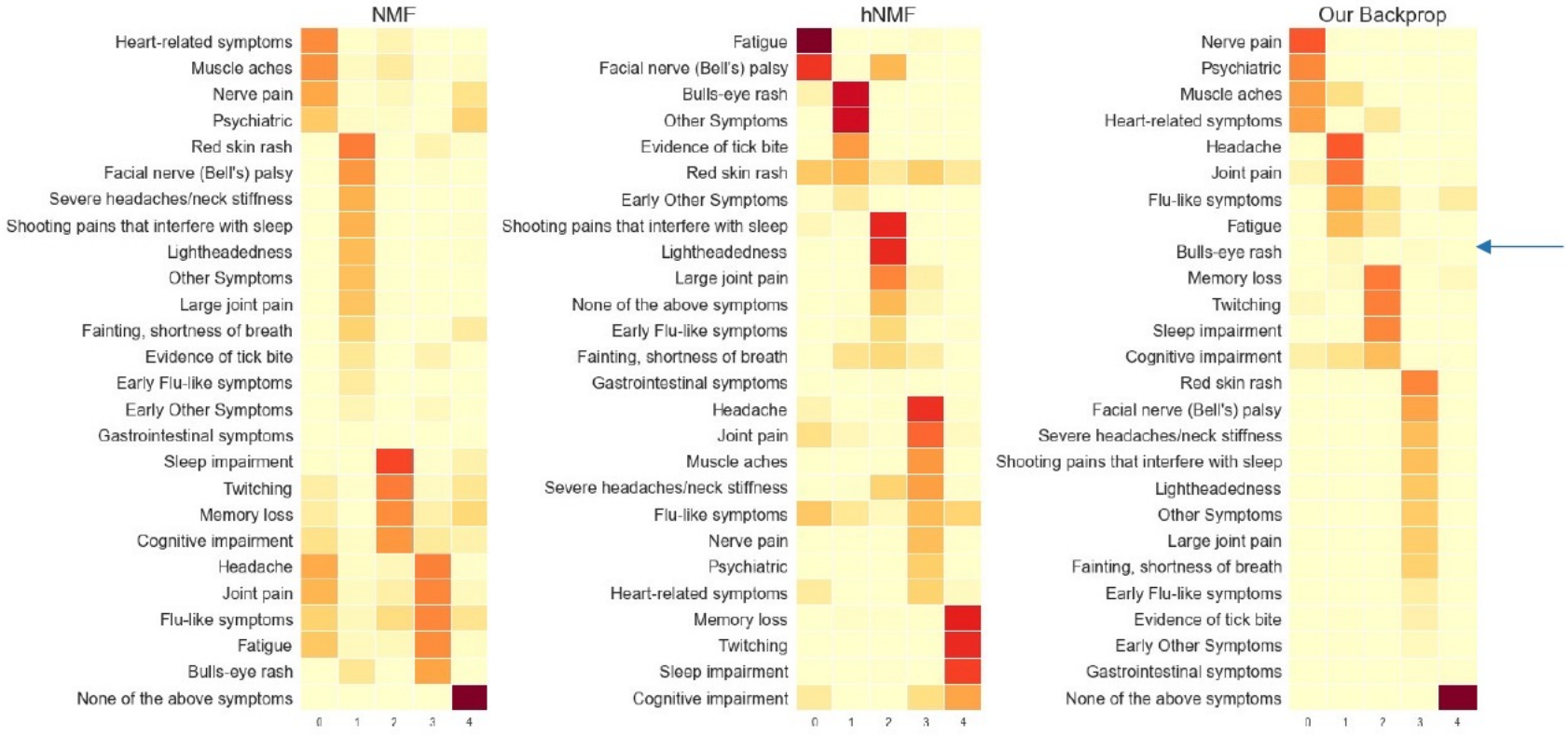
NMF-based models

- **Neural NMF with back propagation** (heuristics + applications)
- **Non-negative Tensor Factorization (NTF)** for multi-modal data (some theoretical guarantees + applications)
- **Online NMF for time series data** (theoretical guarantees + applications)
- **Guided NMF for topic seeding** (applications)
- **Stratified NMF for heterogeneous data** (some theory + applications)

MyLymeData

- Lyme disease a vector-borne disease typically transmitted by tick or insect bite or blood-blood contact
  - Symptoms often mimic those of others, e.g. MS / ALS / Parkinsons / FMA … and can become chronic
- CDC estimates 300,000 new diagnoses each year
  - Likely a grandiose underestimate
- Poorly understood, poorly funded, poorly diagnosed, poorly treated

MyLymeData



The hidden topics here may provide insight on how symptoms manifest themselves

"Feature selection from lyme disease patient survey using machine learning"
by J. Vendrow, J. Haddock, D. Needell, L. Johnson.
Algorithms, vol. 13, num. 12, pp. 334, 2020.

California Innocence Project

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
trial	dna	witness	said	help	would	que
evidence	blood	police	told	need	like	por
attorney	apartment	suspect	got	please	thank	gracias
jury	fact	trial	get	know	send	yo
defendant	items	description	would	innocence	innocence	eston
testimony	victim	interview	went	crime	questionnaire	swedes
judge	profile	also	going	years	screening	su
never	done	said	car	convicted	concern	es
sentence	could	gave	never	prove	dear	para
years	detective	detective	asked	hello	address	mucha

Table (5) The Top 10 topic keywords learned by the first layer of Bottom-up HNMF on the initial letters from both categories

CALIFORNIA INNOCENCE PROJECT

"Analysis of Legal Documents via Non-negative Matrix Factorization Methods" by R. Budahazy, L. Cheng, Y. Huang, A. Johnson, P. Li, J. Vendrow, Z. Wu, D. Molitor, E. Rebrova, D. Needell.
SIAM Undergraduate Research Online, vol. 15, 2022.

To matricize or not to matricize

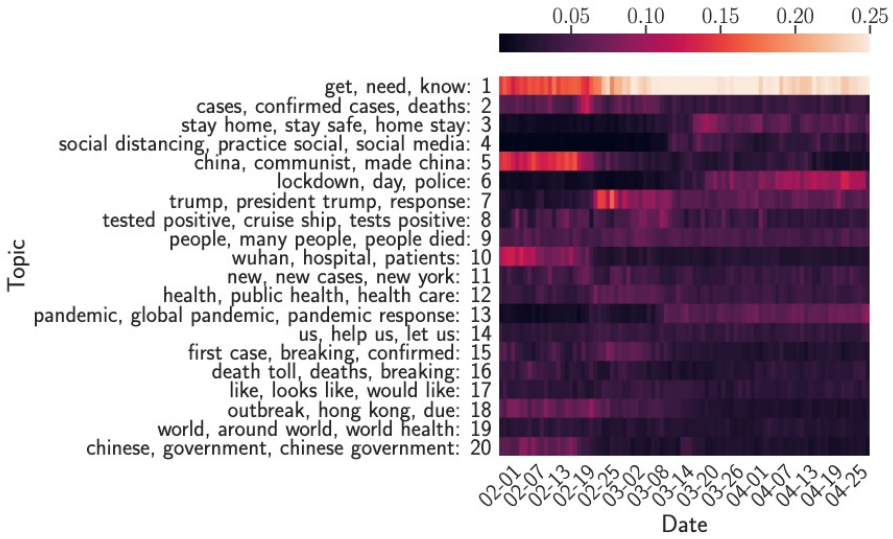


Fig. 8: The normalized mean topic representation of tweets per day learned via NMF with rank 20.

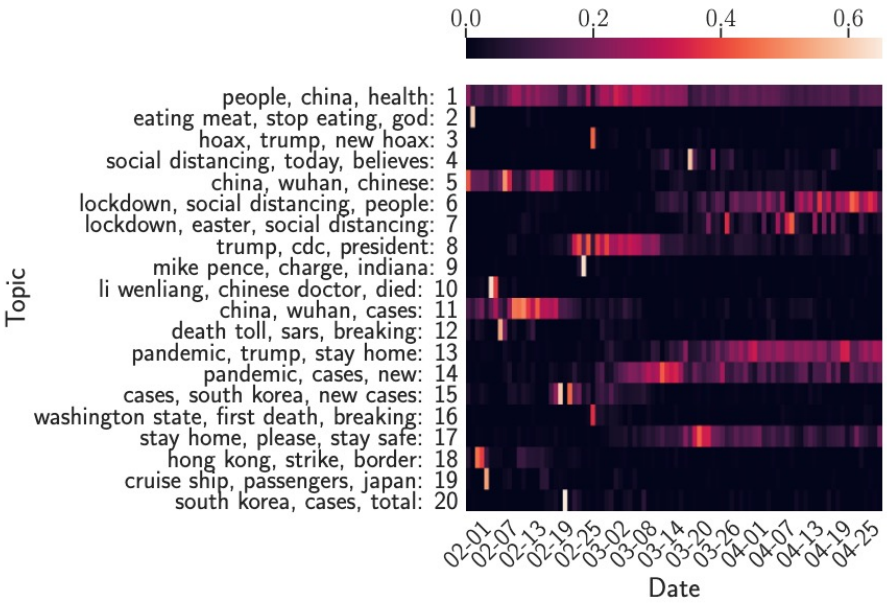


Fig. 10: The normalized factor matrix of NCPD on the tweets dataset with rank 20.

"Detecting Short-lasting Topics Using Nonnegative Tensor Decomposition" by L. Kassab, A. Kryshchenko, H. Lyu, D. Molitor, D. Needell, E. Rebrova. Submitted, 2023.

To matricize or not to matricize

	LDA	NMF	NCPD	ONCPD
(1)	would, like, one	would, like, think	space, would, like	space, would, nasa
(2)	edu, use, window	drive, sale, offer	00, sale, drive	00, sale, drive
(3)	space, launch, nasa	space, shuttle, nasa	games, game, year	games, year, game
(4)	new, sale, please	00, 20, 50	god, believe, religion	god, wrong, people
(5)	00, 50, 20	mac, hm, msu	window, widget, application	window, widget, application

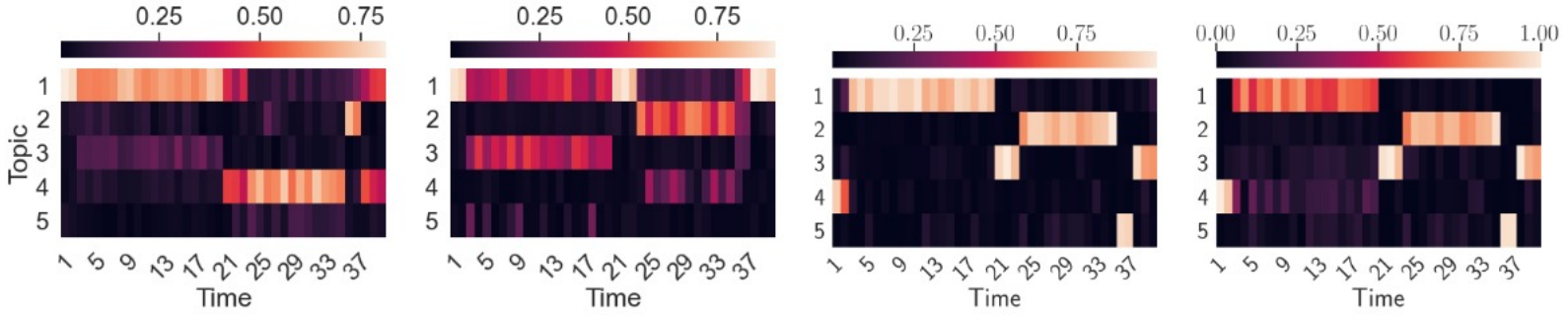
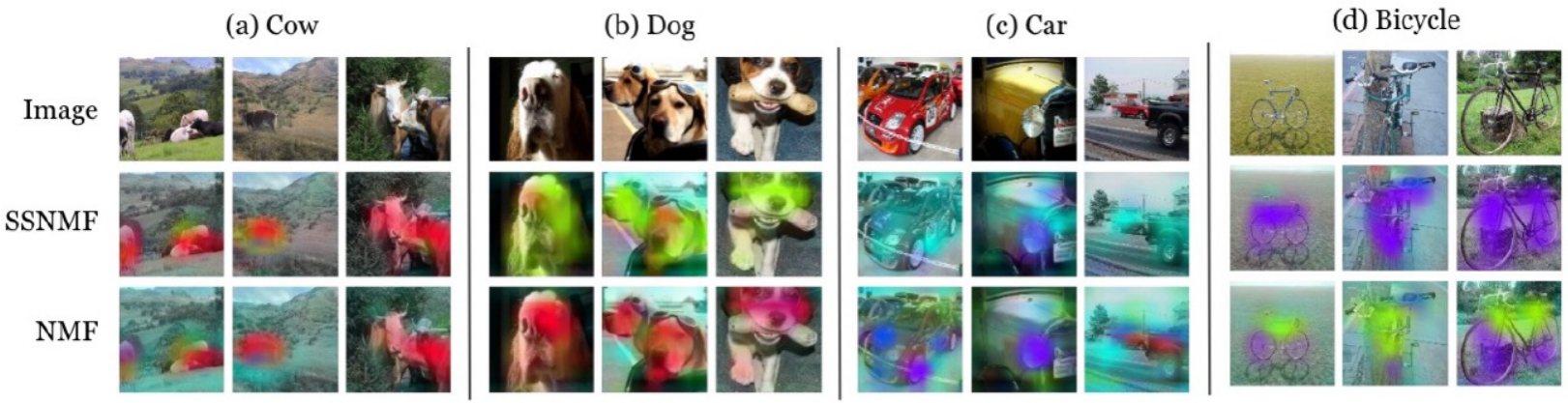


Fig. 1: The learned topics and prevalence of each extracted topic from the semi-synthetic 20 Newsgroups dataset are shown. The columns of each heatmap indicate the distribution over the extracted topics for each time slice. The top 3 keywords corresponding to each topic of the models are provided.

"Detecting Short-lasting Topics Using Nonnegative Tensor Decomposition" by L. Kassab, A. Kryshchenko, H. Lyu, D. Molitor, D. Needell, E. Rebrova. Submitted, 2023.

(O)NMF for Image co-segmentation



"Interpretability of Automatic Infectious Disease Classification Analysis with Concept Discovery"
by E. Sizikova, J. Vendrow, X. Cao, R. Grotheer, J. Haddock, L. Kassab, A. Kryshchenko, T. Merkh, R. W. M. A. Madushani, K. Moise, A. Ulichney, H. V. Vo, C. Wang, M. Coffee, K. Leonard, D. Needell.
Submitted, 2022.

ONMF for image reconstruction

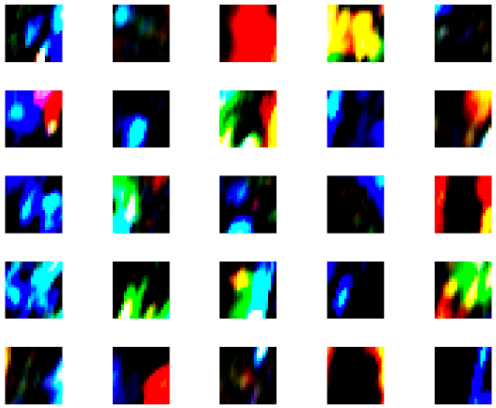


Fig. 7: Image Compression Via ONMF. (Top) uncompressed image of Leonid Afremov's famous painting "Rain's Rustle." (Middle) 25 of the 100 learned dictionary elements, reshaped from their vectorized form to color image patch form. (Bottom): Painting compressed using a dictionary of 100 vectorized 20×20 color image patches obtained from 30 data samples of ONMF, each consisting of 1000 randomly selected sample patches. We used an overlap length of 15 in the patch averaging for the construction of the compressed image.

"Online Nonnegative CP-dictionary Learning for Markovian Data"
by C. Strohmeier, H. Lyu, D. Needell.
Journal of Machine Learning Research, vol. 23, num. 148, pp. 1-50, 2022.

ONTF to learn activation patterns in mouse cortex

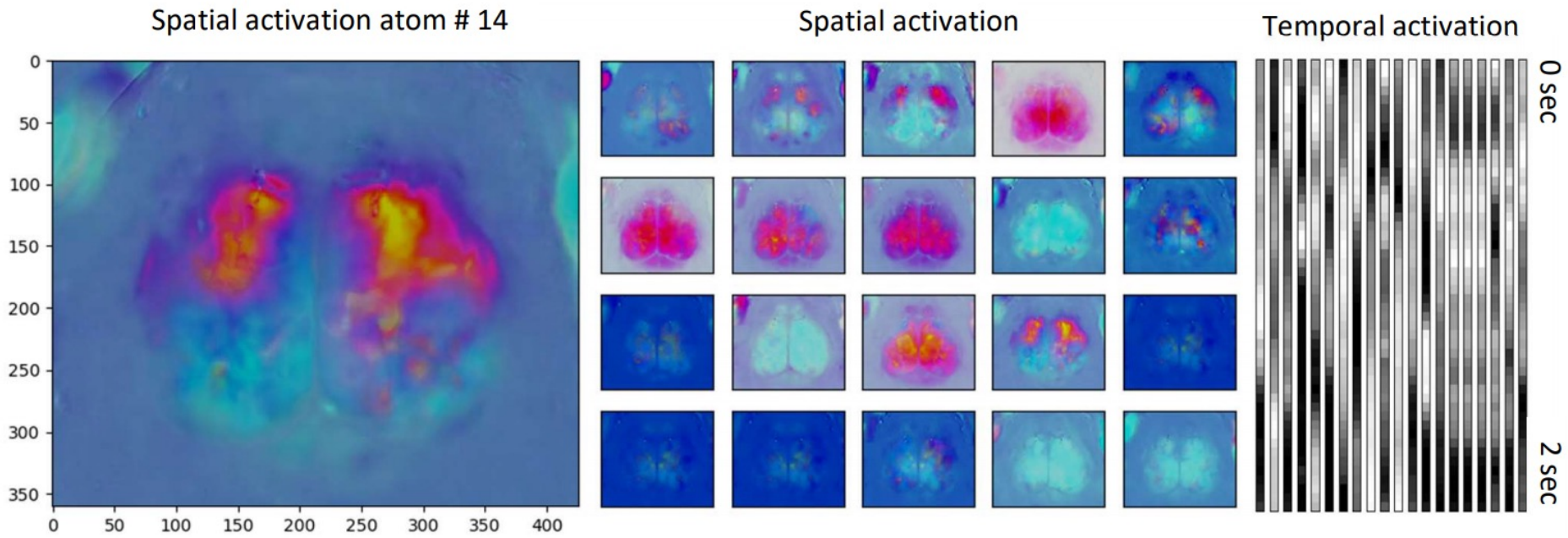


FIGURE 4. Learning 20 CP-dictionary patches from video frames on brain activity across the mouse cortex.

"Online Nonnegative CP-dictionary Learning for Markovian Data"
by C. Strohmeier, H. Lyu, D. Needell.
Journal of Machine Learning Research, vol. 23, num. 148, pp. 1-50, 2022.

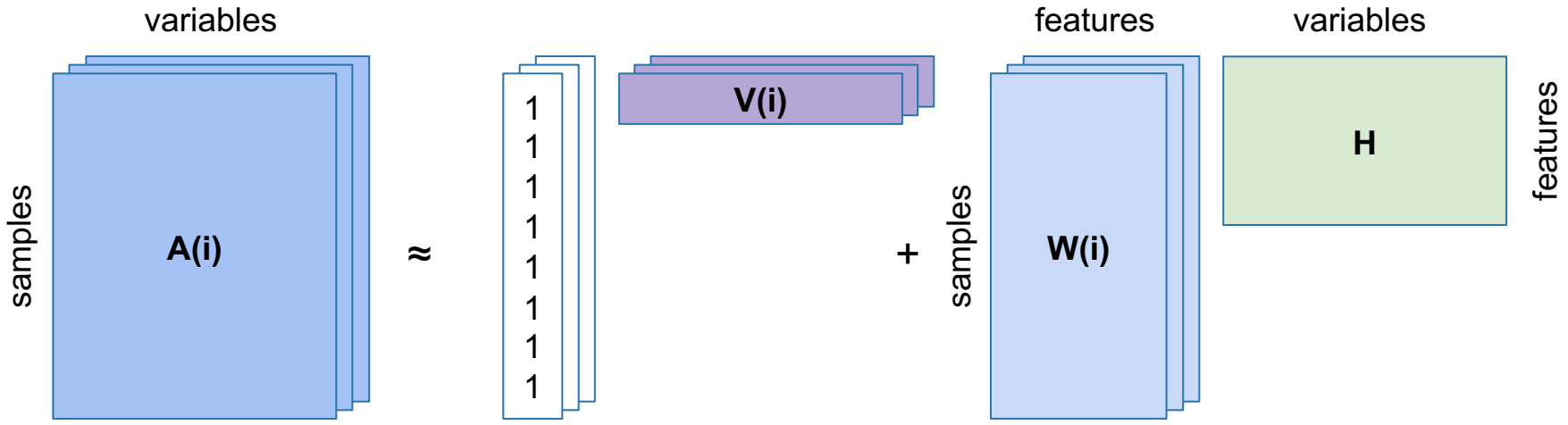
Guided NMF for topic seeding

Table 1. Topic keywords learned for a rank 4 Guided NMF on the 20 Newsgroups dataset with the seed words *pitch*, *medical*, and *space*. We see that a clear topic forms from each keyword matching one desired newsgroup class.

Topic 1	Topic 2	Topic 3	Topic 4
pitch	*medical*	*space*	people
expected	tests	nasa	know
curveball	disease	shuttle	think
stiffness	diseases	launch	time
loosen	prejudices	sci	use
shoulder	services	lunar	new
shea	graduates	orbit	see
rotation	health	earth	say
game	patients	station	us
giants	available	mission	god

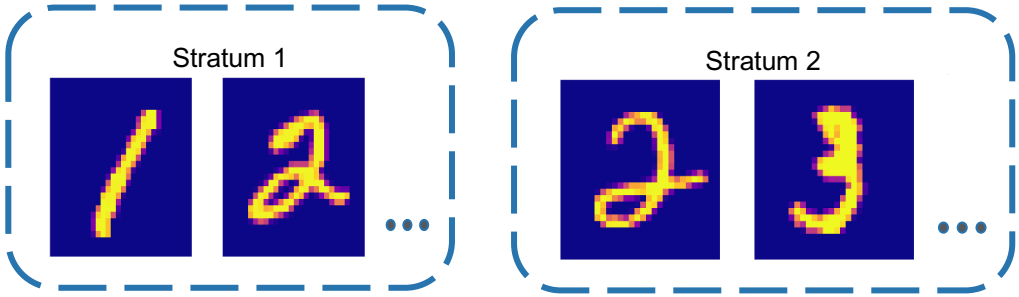
J. Vendrow, J. Haddock, E. Rebrova and D. Needell, "On a Guided Nonnegative Matrix Factorization," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 3265-32369, doi: 10.1109/ICASSP39728.2021.9413656.

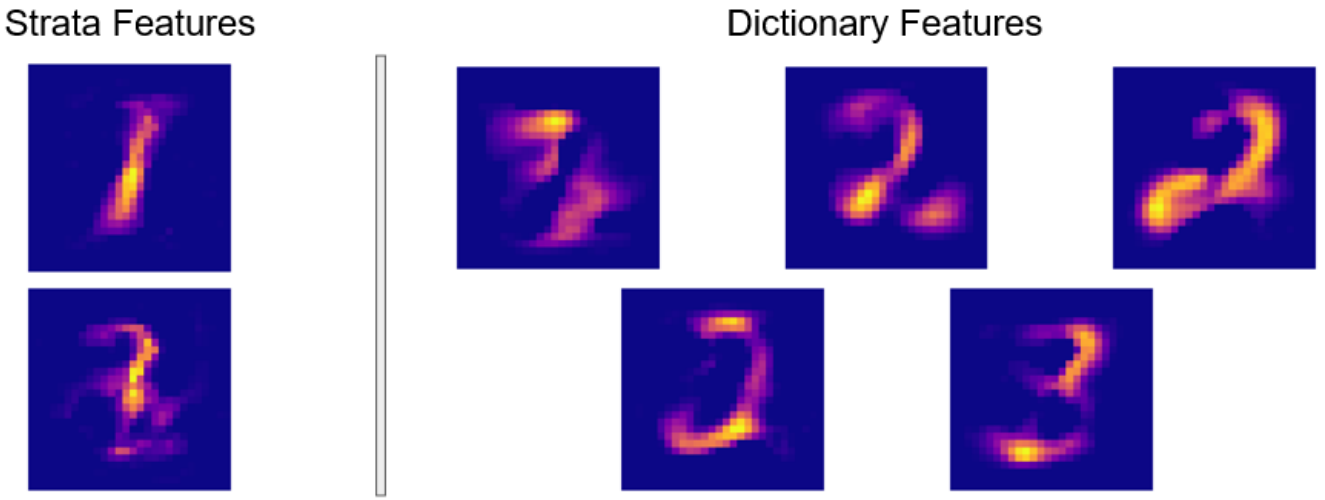
Stratified NMF for stratified (differently sourced) data



"Stratified NMF for Heterogeneous Data"
J. Chapman, Y. Yaniv, D. Needell.
Proc. 55th Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, CA, 2023.

Stratified NMF for stratified (differently sourced) data





"Stratified NMF for Heterogeneous Data"
J. Chapman, Y. Yaniv, D. Needell.
Proc. 55th Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, CA, 2023.

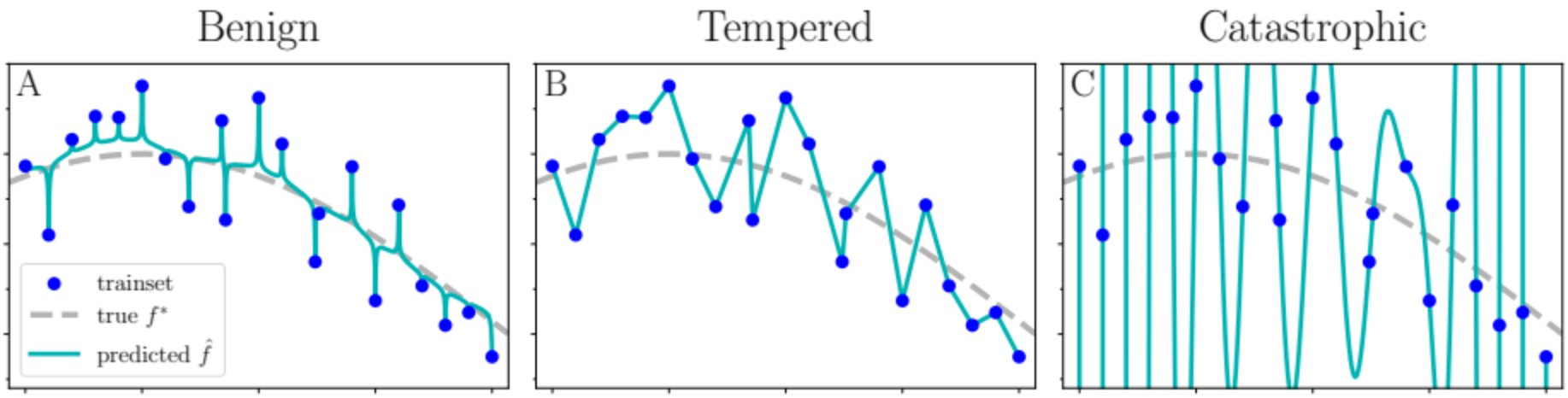
UCLA

Questions?

Next up …

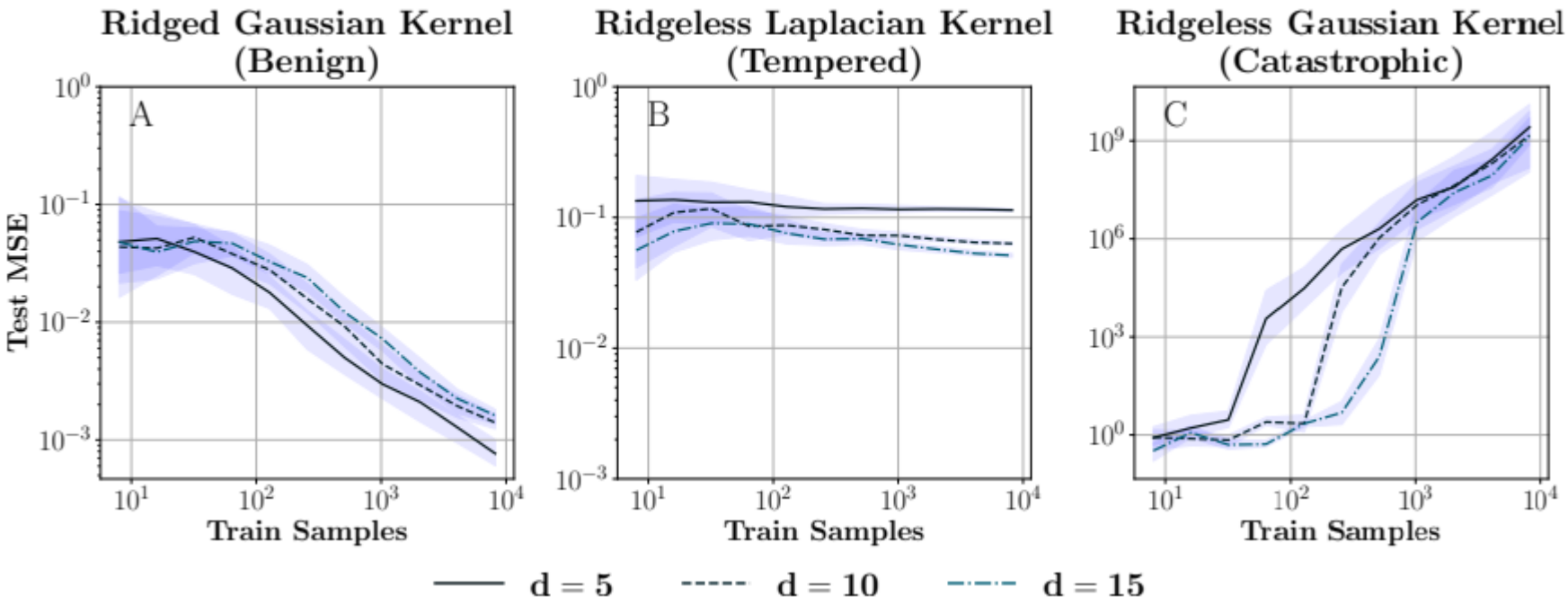
Towards transparency in neural nets

Towards understanding in ML



Mallinar, N., Simon, J. B., Abedsoltan, A., Pandit, P., Belkin, M., & Nakkiran, P. (2022). Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*.

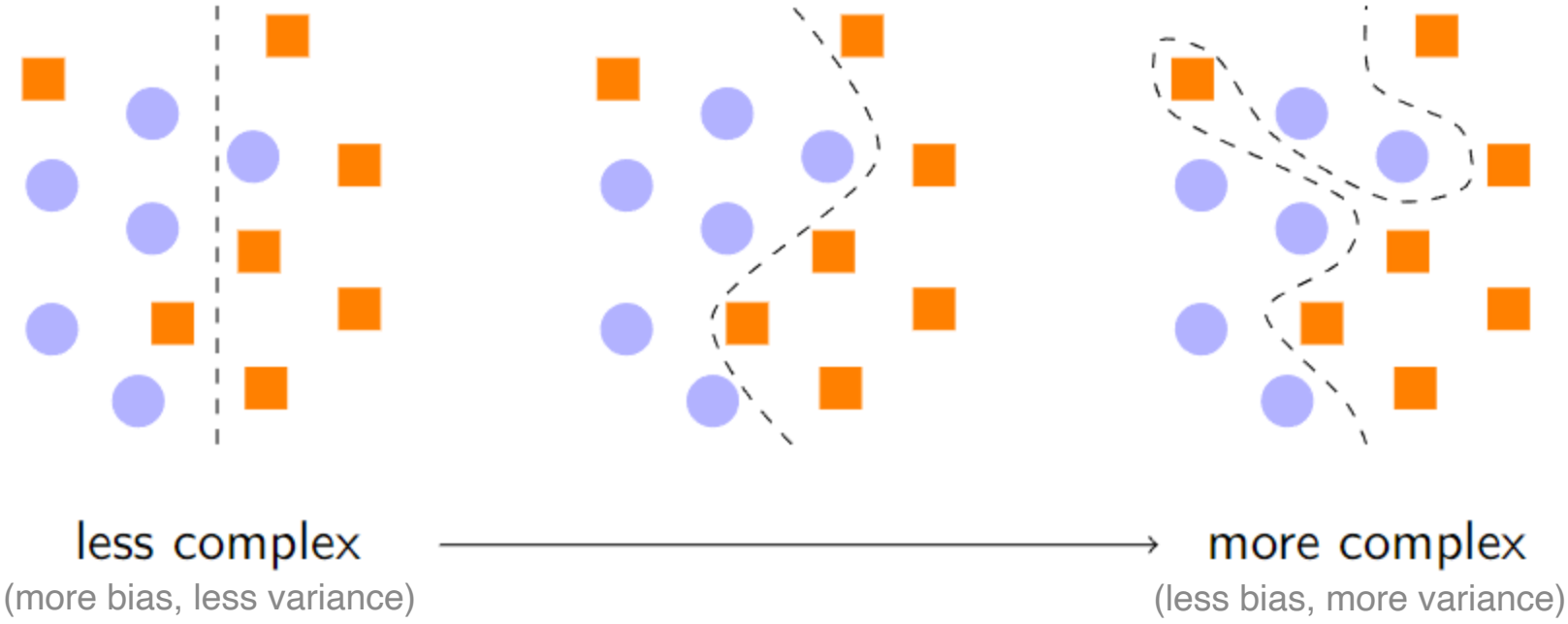
Towards understanding in ML



Kernel regression can exhibit all three (with proper choice of ridge parameter and kernel)

Mallinar, N., Simon, J. B., Abedsoltan, A., Pandit, P., Belkin, M., & Nakkiran, P. (2022). Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*.

Classical bias–variance (simplicity–sensitivity) tradeoff

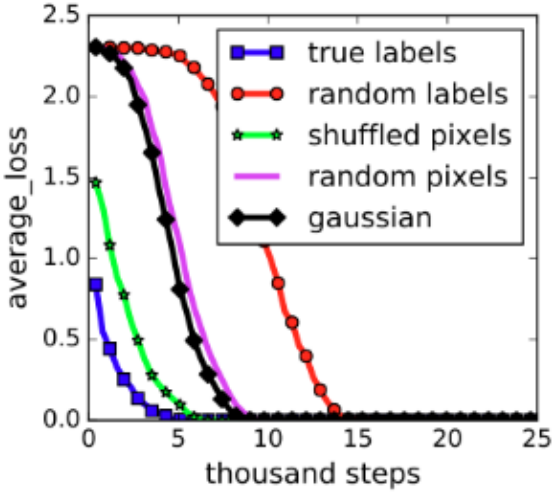


(thanks to E. George for these slides!)

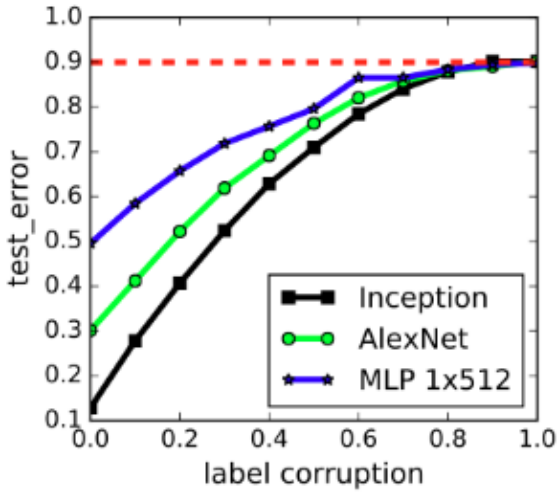
"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

And yet...

Deep learning models are highly complex and expressive, yet even when trained with no explicit regularization to perfectly interpolate noisy training data, they still generalize well



(a) learning curves



(c) generalization error growth

(CIFAR10)

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals.
Understanding deep learning requires rethinking generalization.
In International Conference on Learning Representations, 2017

Benign overfitting

A number of benign/tempered overfitting results have emerged for two layer networks trained with GD + logistic loss on noisy, linearly separable data for binary classification with near-orthogonal inputs.

- [FCB22] consider smoothed leaky ReLU activations and assume the data is drawn from a mixture of well-separated sub-Gaussian distributions.
- [XG23] extends this result to more general activation functions, including ReLU.
- [CCBG22, KCCG23] study convolutional networks where the noise and signal components lie on disjoint patches.
- [FVBS23] considers leaky ReLU and analyzes the KKT points of the max-margin problem.
- [KYS23] demonstrate benign-tempered overfitting transitions in the case of univariate inputs for ReLU networks.

"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

Benign overfitting

- Informally, we say a model exhibits **benign overfitting** if it achieves zero error on noisy training data, but still performs well on test data.
- Significant progress has been made in understanding benign overfitting in linear models, but less is known about non-linear models.
- We seek to study the dynamics of a (shallow) ReLU neural network trained using GD and hinge loss on a noisy binary classification problem.

"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

The loss

Hinge loss: $\max\{0, 1 - z\}$

- Defines a margin separating classes and penalizes points for lying within or on the incorrect side.
- Contribution of each point to overall loss driven only its network activation.
- When $y_i f(\mathbf{x}_i) \geq 1$, point no longer contributes to dynamics (switches off).

Logistic loss: $\log(1 + \exp(-z))$

- Attempts to learn log odds of point being in positive class.
- Points which are already well fitted, i.e., $y_i f(\mathbf{x}_i)$ is large, have a reduced contribution.
- A point always contributes to the dynamics of the network (never switches off).

"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

Benign overfitting

Assume inputs $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$ have a signal and noise component and let $\ghost \in [0, 1]$ control the strength of the signal component:

$$\mathbf{x}_i \approx \sqrt{\ghost} y_i \mathbf{s}_i + \sqrt{1 - \ghost} \mathbf{n}_i.$$

We show three distinct training outcomes:

1. **Benign overfitting** ($\ghost$ *small but not too small*): zero training loss and generalization error asymptotically (in dimension d) optimal.
2. **Non-benign overfitting** ($\ghost$ *very small*): zero training loss and generalization error bounded below by a constant. (note! optimal classifier exists)
3. **No overfitting** ($\ghost$ *large*): zero training loss on "clean" points but nonzero loss on "corrupted" points, and asymptotically optimal generalization error.

"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

Benign overfitting

- Training sample has $2n$ points $(\mathbf{x}_i, y_i)_{i=1}^{2n} \in (\mathbb{R}^d \times \{-1, 1\})$.
- k positive and k negative points have their output label flipped: denote $\beta(i) = -1$ if i-th point is corrupted otherwise $\beta(i) = 1$.
- Labels: $y_i = (-1)^i \beta(i)$ (clean label is $(-1)^i$)
- Inputs are of the form

$$\mathbf{x}_i = (-1)^i (\sqrt{👻}\mathbf{v} + \sqrt{1 - 👻}\beta(i)\mathbf{n}_i).$$

- **Noise vectors** $(\mathbf{n}_i)_{i=1}^{2n}$ are mutually independent and identically distributed (i.i.d.) random vectors drawn from the uniform distribution over $\mathbb{S}^{d-1} \cap \text{span}\{\mathbf{v}\}^\perp$.
- $👻 \in [0, 1]$ controls the strength of the signal versus the noise.
- Test data has same form but is assumed uncorrupted.

"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

Main results ($\gamma = \infty$)

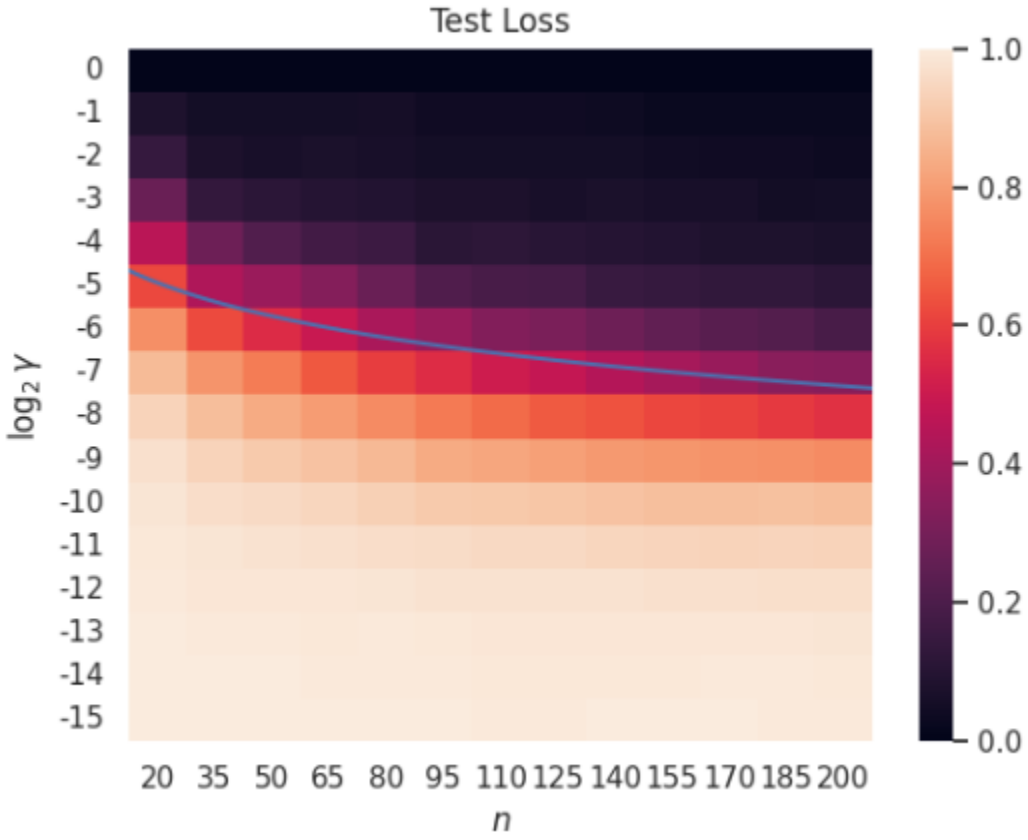
- **No overfitting (signal dominant regime):** when γ is sufficiently large relative to n, d ($\gamma \gtrsim n^{-1}$) then i) clean training points have 0 loss, ii) corrupt training points have loss one and iii) network has asymptotically optimal generalization error.
- **Benign overfitting (balanced signal-noise regime):** when γ is large but not too large relative to n, d ($d^{-1/2} \lesssim \gamma \lesssim n^{-1}$) then i) clean training points have 0 loss, ii) corrupt training points have 0 loss and iii) network has asymptotically optimal generalization error.
- **Harmful overfitting (noise dominant regime):** when γ is sufficiently small relative to n, d ($\gamma \lesssim (nd)^{-1/2}$) then i) clean training points have 0 loss, ii) corrupt training points have loss 0 and iii) network has generalization error bounded from below.

Fine print:
2n data points
2m neurons (1 layer)
d = dimension
k = # corruptions

Assume: k<cn, step size small enough, d big enough, noise nearly orthogonal

"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

Empirics match



"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

UCLA

Questions?

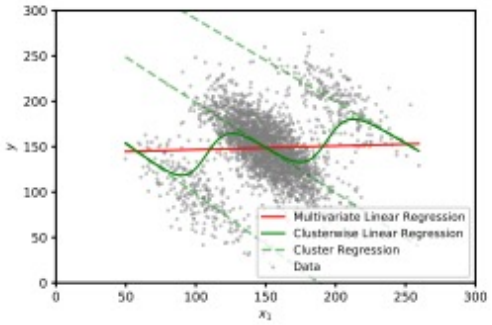
Next up …

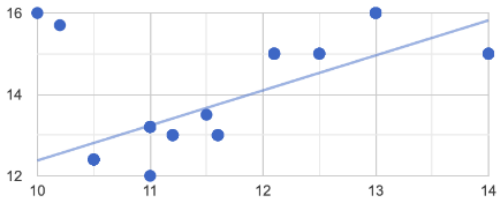
Fairness

Fairness

Often, because of objective functions over an entire population, subgroups have drastically inferior accuracy

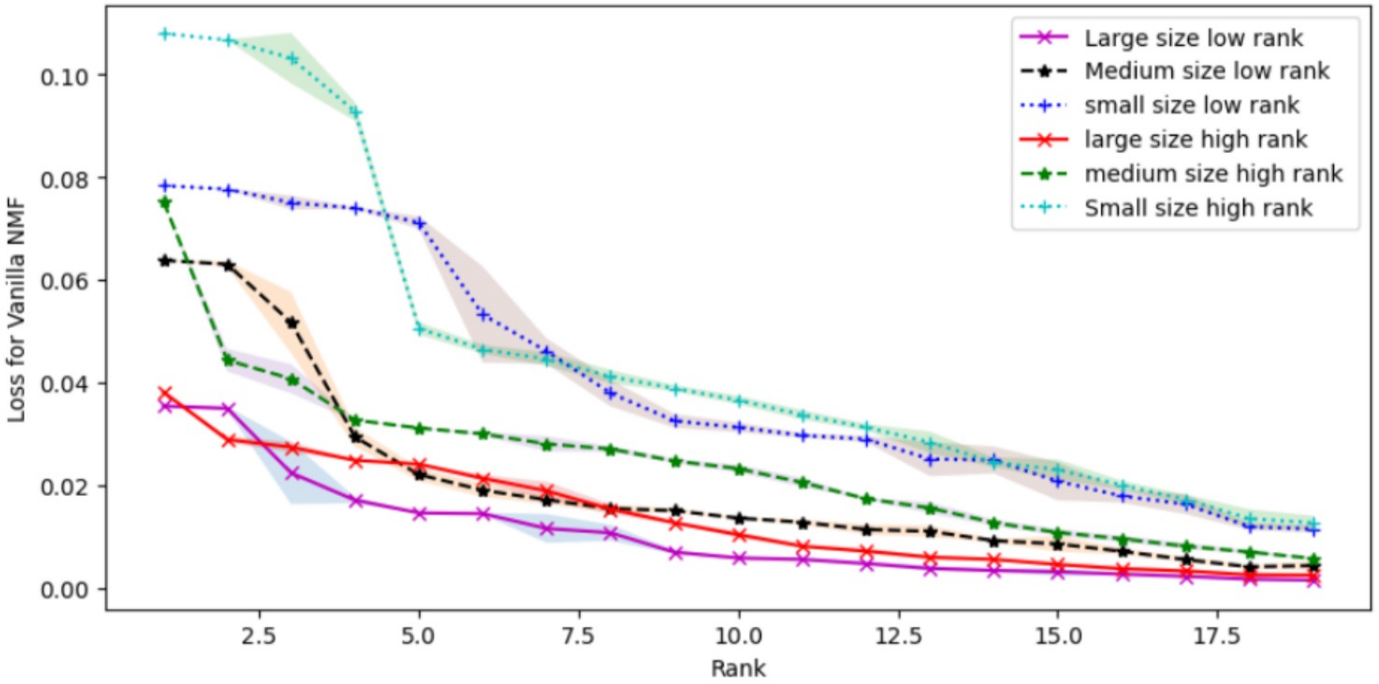
- Regression attempts to minimize ***average*** explanations
- NMF learns topics that explain the population ***overall***





Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *ACM Computing Surveys (CSUR)* 54.6 (2021): 1-35.

NMF on mixed population



Two sources of unfairness: representation and complexity

NMF on mixed population

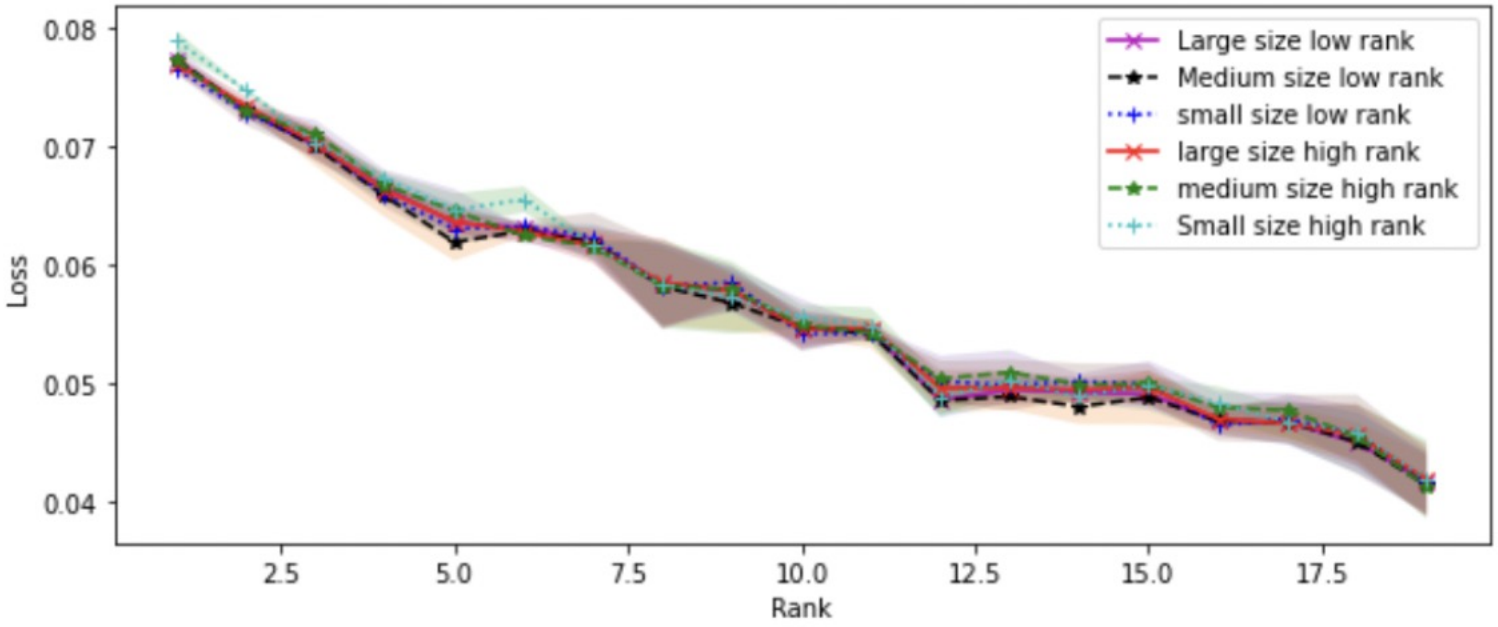
$$\min_{\mathbf{W}\in\mathbb{R}_+^{n_1\times r},\mathbf{H}\in\mathbb{R}_+^{r\times n_2}} \max_{\mathbf{A},\mathbf{B}} \left\{ \frac{\|\mathbf{A}-\mathbf{W_A H}\|_F^2 - \|\mathbf{A}-\mathbf{A}^*\|_F^2}{|A|}, \frac{\|\mathbf{B}-\mathbf{W_B H}\|_F^2 - \|\mathbf{B}-\mathbf{B}^*\|_F^2}{|B|} \right\}$$

One notion of “more fair” : Each group achieves loss equally, relative to their size and best possible loss

(many formulations)

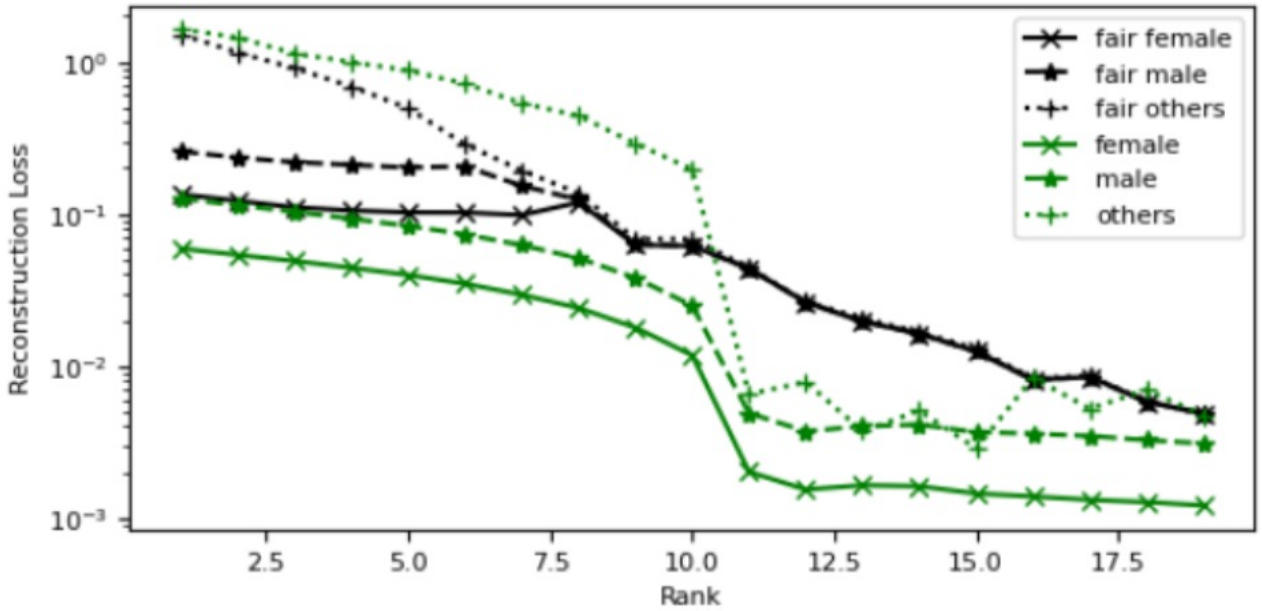
“Fairer” NMF on mixed population

One notion of “more fair” : Each group achieves loss equally, relative to their size and best possible loss

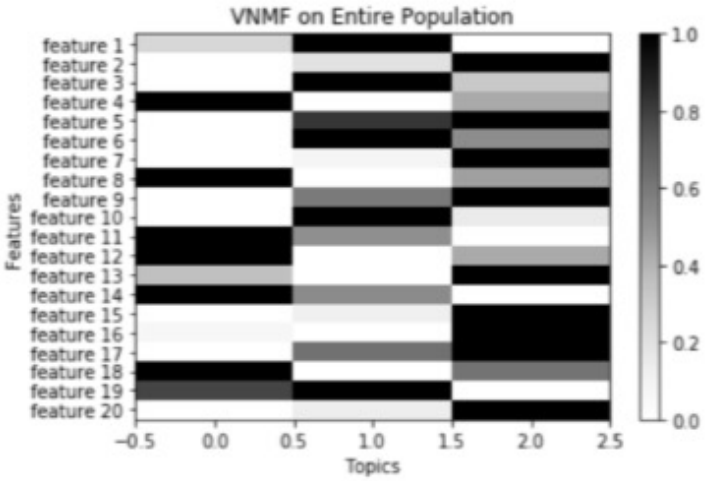


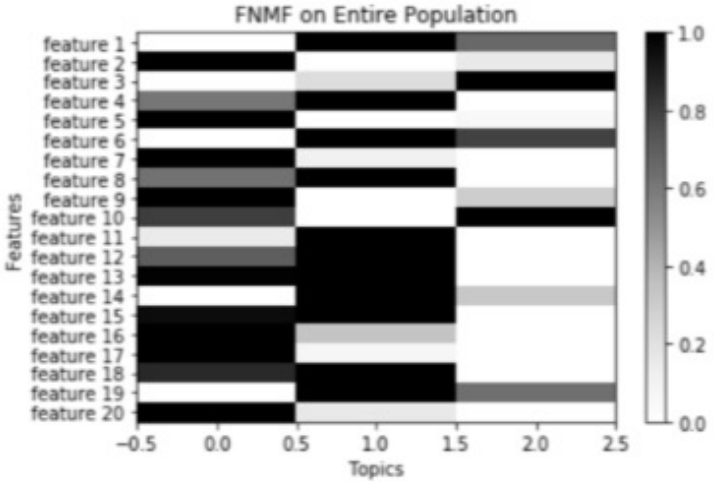
“Fairer” NMF on mixed population

One notion of “more fair” : Each group achieves loss equally, relative to their size and best possible loss



“Fairer” NMF on mixed population





UCLA

Questions?

Next up …

Practical and efficient tensor compression and reconstruction

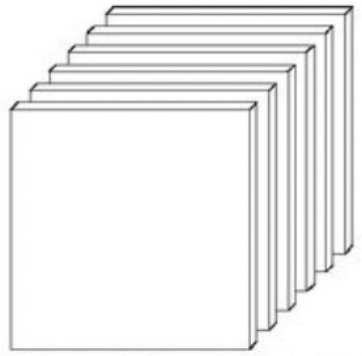
Tensor Compression and Reconstruction

Goal: Compress tensor data via linear measurements that are practical to apply and allow for efficient reconstruction

$$\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \ldots \times n_d} - d\text{-mode tensor}$$

Naturally multi-modal data is ubiquitous:

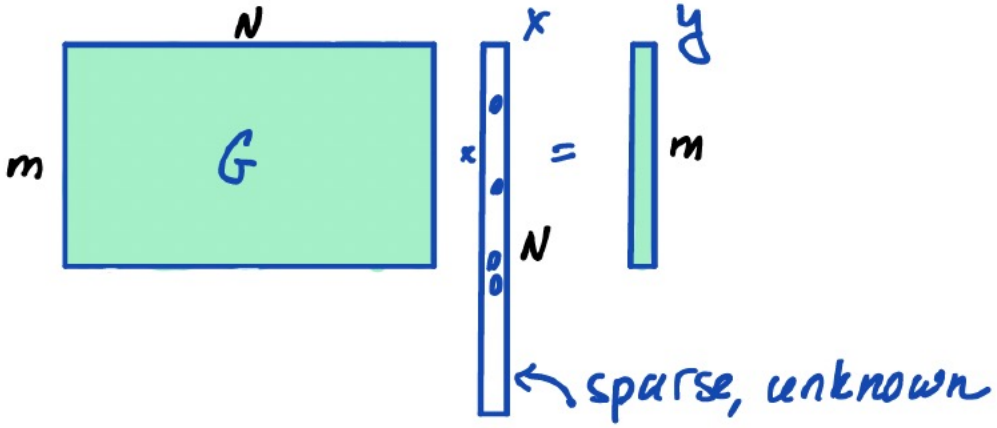
- datasets with many attributes
- datasets with temporal component
- color pictures, videos



So,

- Converting it to a vector (vectorization) or to a matrix (matricization) destroys the structure of such data!
- For x being the vectorization of a $n_1 \times n_2 \times \ldots \times n_d$-dimensional tensor, $N = \prod n_i$, resulting in a measurement matrix G of the size $m \times n^d$.

Tensor Compression and Reconstruction



Iterative methods to recover r-sparse $\mathbf{x}$: e.g., Iterative Hard Thresholding

$$\tilde{\mathbf{x}}_k = \mathbf{x}_k + \mu_k G^*(\mathbf{y} - G\mathbf{x}_k),$$
$$\mathbf{x}_{k+1} = \mathcal{H}_r\left(\tilde{\mathbf{x}}_k\right), \quad \mathcal{H}_r(.) \text{ gives the best } r\text{-sparse approximation.}$$

Tensor Compression and Reconstruction

Why does the step

$$\tilde{\mathbf{x}}_k = \mathbf{x}_k + G^*(\mathbf{y} - G\mathbf{x}_k)$$

bring us closer to the solution? Idea: re-group

$$\tilde{\mathbf{x}}_k = (I - G^*G)\mathbf{x}_k + G^*\mathbf{y}.$$

If $G^*G \sim I$ close to the identity, then $\mathbf{x} \approx G^*G\mathbf{x} = G^*\mathbf{y} \approx \tilde{\mathbf{x}}_k$ (the next iterate is close to a solution).

How to quantify $G^*G \sim I$?

RIP-property: A $m \times N$ matrix G has a (δ, r)-RIP property if

$$\left| \|G\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \leq \delta \|\mathbf{x}\|_2^2 \text{ for any } r\text{-sparse vector } x \in \mathbb{R}^N$$

→ When x is a matrix or tensor, sparsity can be substituted with (some notion of) low-rankness (tubal, Tucker or HOSVD, multi-rank, CP)

Tensor world…take a TRIP

HOSVD decomposition (Tucker rank)

$$\mathcal{X} = \mathcal{C} \times_1 U^1 \times_2 \ldots \times_d U^d$$
$$= \sum_{k_d=1}^{r_d} \ldots \sum_{k_1=1}^{r_1} \mathcal{C}(k_1, \ldots k_d) \bigcirc_{i=1}^{d} \mathbf{u}_{k_i}^i,$$

where all $\mathbf{u}_1^i, \ldots, \mathbf{u}_{r_i}^i$ are orthonormal (U^i is the $n_i \times r_i$ matrix $U^i = (\mathbf{u}_1^i, \ldots, \mathbf{u}_{r_i}^i)$).

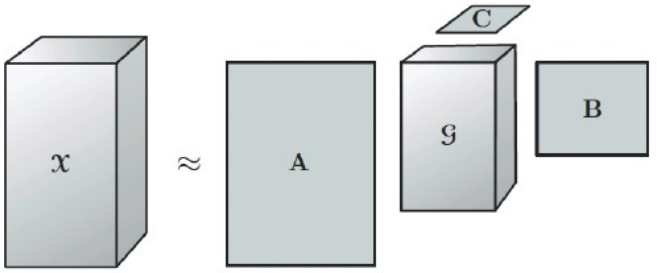


Fig. 4.1 *Tucker decomposition of a three-way array.*

TRIP($\delta, \mathbf{r}$) property We say that a linear map $\mathcal{A}$ has the TRIP($\delta, \mathbf{r}$) property if for all $\mathcal{X}$ with HOSVD rank at most $\mathbf{r} = (r_1, \ldots, r_n)$ we have

$$(1-\delta)\|\mathcal{X}\|^2 \leq \|\mathcal{A}(\mathcal{X})\|^2 \leq (1+\delta)\|\mathcal{X}\|^2$$

Tensor Compression and Reconstruction

- Methods like Tensor Iterative Hard Thresholding (TIHT) allow for efficient recovery of the tensor from TRIP measurements
  - Tucker rank (Rauhut et.al. ‘17)
  - CP rank (N et.al. ‘19)

Algorithm 1 Tensor Iterative Hard Thresholding (TIHT)

Input: operator $\mathcal{H}_r$, rank r, measurements $\boldsymbol{y}$, number of iterations T
Output: $\hat{\boldsymbol{X}} = \boldsymbol{X}^T$.
Initialize: $\boldsymbol{X}^1 = \mathbf{0}$
for $j = 0, 2, \ldots, T-1$ **do**
    $\mathbf{W}^j = \boldsymbol{X}^j + \mathcal{A}^*(\boldsymbol{y} - \mathcal{A}(\boldsymbol{X}^j))$
    $\boldsymbol{X}^{j+1} = \mathcal{H}_r(\mathbf{W}^j)$
end for

What kind of operators satisfy TRIP?

- Existing solution: ***vectorize*** the tensor and apply i.i.d. m x N subgaussian map. For HOSVD rank r, TRIP is satisfied for: (Rauhut etal. '17)

$$m \geq C\delta^{-2} \max\{(r^d + dnr) \ln d, \ln(\eta^{-1})\}$$

- Concern: A 6-mode tensor with 1000 dimensions in each mode now requires the storage of a m x 10^{18} measurement matrix!

Do we really need a huge vacuum?

For $1 \leq i \leq d'$, let A_i be an $m \times n^\kappa$ matrix, let $\mathcal{A} : \mathbb{R}^{n^\kappa \times \ldots \times n^\kappa} \to \mathbb{R}^{m \times \ldots \times m}$ be the linear map which acts modewise on d'-mode tensors by

$$\mathcal{A}(\mathcal{Y}) = \mathcal{Y} \times_1 A_1 \times_2 \ldots \times_{d'} A_{d'}.$$

For $\kappa = 2$ and subgaussian measurement matrices we have

For all $\mathcal{X}$ of rank at most $(r, r, r, \ldots, r)$ with prob $1 - \eta$

$$(1 - \delta)\|\mathcal{X}\|^2 \leq \|\mathcal{A}(\mathcal{R}(\mathcal{X})\|^2 \leq (1 + \delta)\|\mathcal{X}\|^2,$$

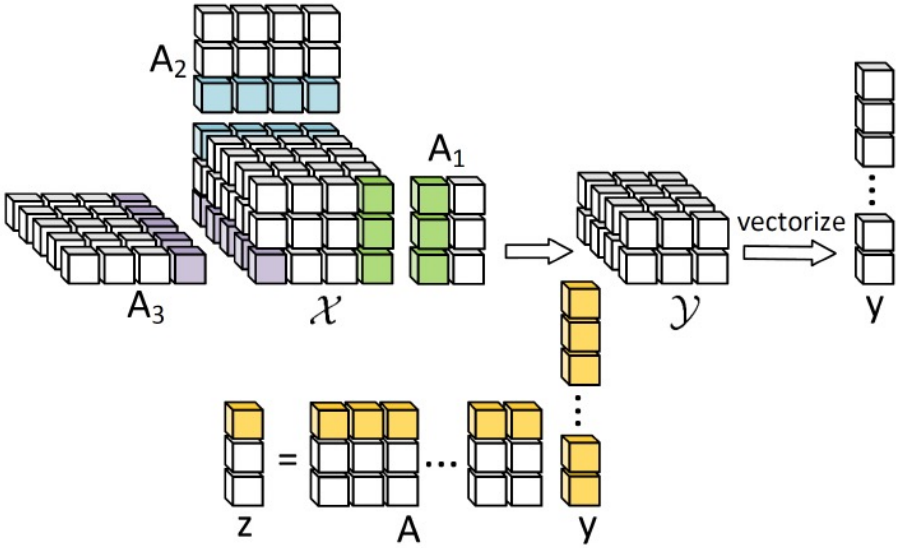
where $\mathcal{R}$ is a reshaping operator that combines pairs of the modes, for target dimensions

$$m \geq Cd^2 r^{2d} \delta^{-2} \max\{n, \log \frac{d}{\eta}\}.$$

(Iwen, N, Perlmutter, Rebrova, '22)

Tensor Compression and Reconstruction

- Memory reduction is decisive: from mn^d to $dmn^\kappa(+m^d m_2)$ where κ can be as small as 2
- Time is compatible (slightly worse than from full measurements), compression quality is compatible (slightly better than from full measurements)



- Reshaping is necessary
- The compression matrices are data-oblivious, generic and flexible (one or two stages, various RIP matrix models can be used in construction)
- Theoretical guarantees arerigrous and show the advantage, but less than experimentaly observed (room for improvement!)

UCLA

References

- **Available at math.ucla.edu/~deanna**

- "Mode-wise Tensor Decompositions: Multi-dimensional Generalizations of CUR Decompositions" by H. Cai, K. Hamm, L. Huang, D. Needell. Journal of Machine Learning Research, vol. 22, num. 185, pp.1–36, 2022.
- "Online matrix factorization for markovian data and applications to network dictionary learning" by H. Lyu, D. Needell, L. Balzano. Journal of Machine Learning Research, vol. 21, num. 251, pp. 1-49, 2020.
- "Iterative Hard Thresholding for Low CP-rank Tensor Models" by R. Grotheer, A. Ma, D. Needell, S. Li, J. Qin. Linear and Multilinear Algebra, pp 1-17, 2021.
- "Modewise Operators, the Tensor Restricted Isometry Property, and Low-Rank Tensor Recovery" by M. A. Iwen, D. Needell, M. Perlmutter, E. Rebrova. Submitted, 2022.
- Other works in preparation with: L. Kassab, E. George, L. Rebrova, M. Iwen, C. Hasselby, W. Swartworth

UCLA

Thank you for listening!

- deanna@math.ucla.edu
- math.ucla.edu/~deanna

BenOv proof idea

1. There are two phases of training driven by the relative imbalance in the number of clean versus corrupt points. Clean data dominates the dynamics early on but once fitted the corrupt points takeover.
2. In the first phase the network fits the clean data by learning a strong signal component, in particular by the end of this phase for most neurons $(-1)^j \langle \mathbf{w}_j, \mathbf{s} \rangle$ is large. Each corrupt point has some neurons of the correct output sign that activate on it throughout this phase.
3. In the second phase clean points start to switch off. The network fits the corrupt data by learning the noise components, however, only so many updates can occur before these points are fitted and thus the signal component the network has learned is not overly impacted.
4. At test time the noise component of a new input is approximately orthogonal to the noise components the network has learned, therefore it classified based on its signal component.

Assumptions

Let $\delta \in (0, 1/2)$ denote the failure probability, $\rho \in (0, 1)$ bound the magnitude of inner products of the noise and λ_w bound the norm of weight initializations. For sufficiently large and small constants $C \geq 1$ and $c \leq 1$ respectively,

1. $k \leq cn$,
2. $d \geq C\rho^{-2} \log(n/\delta)$
3. $\lambda_w \leq c\eta$
4. $\eta \leq \xi$, where ξ depends on n, m, k, 👻, and d.

η

"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

Benign overfitting

- We study a densely connected, single layer feed-forward ReLU neural network with no bias terms $f : \mathbb{R}^{2m\times d} \times \mathbb{R}^d \to \mathbb{R}$,

$$f(\mathbf{W}, \mathbf{x}) = \sum_{j=1}^{2m} (-1)^j \max\{0, \langle \mathbf{w}_j, \mathbf{x}\rangle\}.$$

- Use the hinge loss $L(t) := \sum_{i=1}^{2n} \max\{0, 1 - y_i f(t, \mathbf{x}_i)\}$.
- Inner weights trained using (sub)gradient descent. Let
  - $\mathcal{F}^{(t)} := \{i \in [2n] : \ell(t, \mathbf{x}_i) < 1\}$
  - $\mathcal{A}_j^{(t)} := \{i \in [2n] : \langle \mathbf{w}_j^{(t)}, \mathbf{x}_i\rangle > 0\}$,

  then update can be written as

$$\mathbf{w}_j^{(t+1)} = \mathbf{w}_j^{(t)} + (-1)^j \eta \sum_{l=1}^{2n} \mathbb{1}(l \in \mathcal{A}_j^{(t)} \cap \mathcal{F}^{(t)}) y_l \mathbf{x}_l.$$

"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

Benign overfitting

Theorem 1

Assume $n \geq C\log(1/\delta)$, $m \geq C\log(n/\delta)$, $\rho \leq c \cdot \text{👻}$ *and* $C\sqrt{\log(n/\delta)/d} \leq \text{👻} \leq cn^{-1}$. *Then there exists a sufficiently small step-size* η *such that with probability at least* $1-\delta$ *over the randomness of the dataset and network initialization the following hold.*

1. *The training process terminates at an iteration* $\mathcal{T}_{end} \leq \frac{Cn}{\eta}$.
2. *For all* $i \in [2n]$ *then* $\ell(\mathcal{T}_{end}, \mathbf{x}_i) = 0$.
3. *The generalization error satisfies*

$$\mathbb{P}(\text{sgn}(f(\mathcal{T}_{end}, \mathbf{x})) \neq y) \leq \exp\left(-cd \cdot \text{👻}^2\right).$$

"Training shallow ReLU networks on noisy data using hinge loss: when do we overfit and is it benign?"
E. George, M. Murray, W. Swartworth, D. Needell.
Neural Information Processing Systems (NeurIPS), Spotlight paper, 2023.

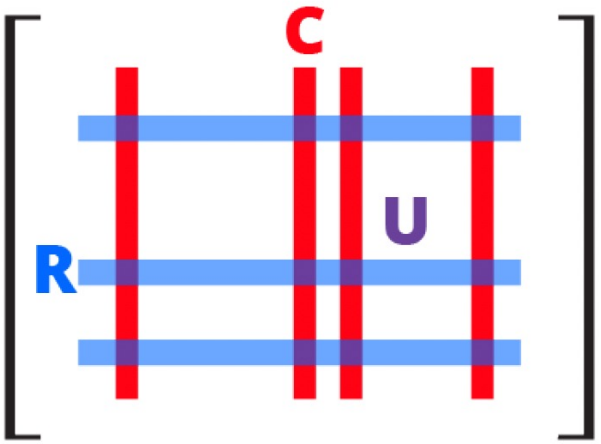
CUR Factorization

- $A \in \mathbb{R}^{d \times d}$,
- $C \in \mathbb{R}^{d \times k}$: k columns of A
- $R \in \mathbb{R}^{s \times d}$: s rows of A
- $U \in \mathbb{R}^{s \times k}$: the intersection of C and R

Theorem

If $\text{rank}(U) = \text{rank}(A)$, *then*

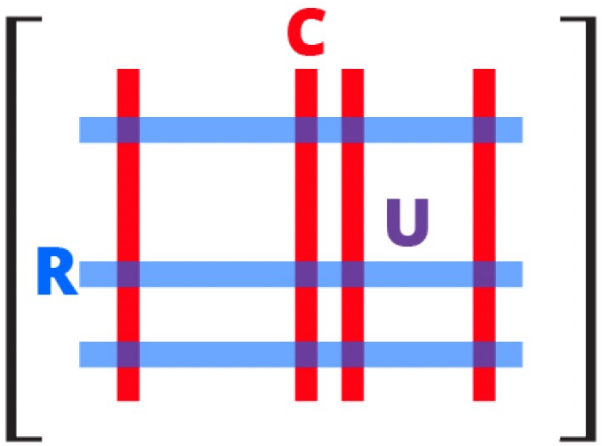
$$A = CU^{\dagger}R.$$



CUR Factorization – for tensors??

Motivation

Let $A \in \mathbb{R}^{d \times d}$ with CUR decomposition of $A = CU^\dagger R$. Then
$A = CU^\dagger R = CU^\dagger UU^\dagger R = U \times_1 (CU^\dagger) \times_2 (R^T (U^T)^\dagger)$.



CUR Factorization

Theorem (Cai–Hamm–Huang–N, 2022)

(Chidori CUR) Let $\mathcal{A} \in \mathbb{R}^{d \times \cdots \times d}$ *with* $\text{rank}(\mathcal{A}) = (r, \ldots, r)$. *Let* $I_i \subseteq [d]$. *Set* $\mathcal{R} = \mathcal{A}(I_1, \cdots, I_n)$, $C_i = \mathcal{A}_{(i)}(:, J_i := \otimes_{j \neq i} I_j)$ *and* $U_i = C_i(I_i, :)$. *Then the following are equivalent:*

1. $\text{rank}(U_i) = r$,
2. $\mathcal{A} = \underbrace{\mathcal{R} \times_1 (C_1 U_1^\dagger) \times_2 \cdots \times_n (C_n U_n^\dagger)}_{CUR}$,
3. $\text{rank}(\mathcal{R}) = (r, \cdots, r)$,
4. $\text{rank}(\mathcal{A}_{(i)}(I_i, :)) = r$ *for all* $i \in [n]$.

Moreover, if the above statements hold, then $\mathcal{A} = \mathcal{A} \times_{i=1}^{n} (C_i C_i^\dagger)$.

CUR Factorization

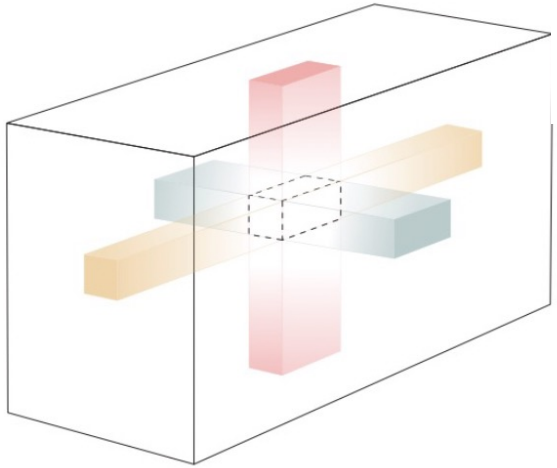


Figure 1: Illustration of Chidori CUR decomposition à la Theorem 3.1 of a 3-mode tensor in the case when the indices I_i are each an interval and $J_i = \otimes_{j \neq i} I_j$. The matrix C_1 is obtained by unfolding the red subtensor along mode 1, C_2 by unfolding the green subtensor along mode 2, and C_3 by unfolding the yellow subtensor along mode 3. The dotted line shows the boundaries of $\mathcal{R}$. In this case $U_i = \mathcal{R}_{(i)}$ for all i.

Mixon)

CUR Factorization

Theorem (Cai–Hamm–Huang–N, 2022)

(Fiber CUR): Let $\mathcal{A} \in \mathbb{R}^{d \times \cdots \times d}$ *with* $\text{rank}(\mathcal{A}) = (r, \ldots, r)$. *Let* $I_i \subseteq [d]$ *and* $J_i \subseteq [d^{n-1}]$. *Set* $\mathcal{R} = \mathcal{A}(I_1, \cdots, I_n)$, $C_i = \mathcal{A}_{(i)}(:, J_i)$ *and* $U_i = C_i(I_i, :)$. *Then the following statements are equivalent*

1. $\text{rank}(U_i) = r$,
2. $\mathcal{A} = \underbrace{\mathcal{R} \times_1 (C_1 U_1^\dagger) \times_2 \cdots \times_n (C_n U_n^\dagger)}_{CUR}$,
3. $\text{rank}(C_i) = r$ *for all* $i \in [n]$ *and* $\text{rank}(\mathcal{R}) = (r, \cdots, r)$,
4. $\text{rank}(C_i) = r$ *and* $\text{rank}(\mathcal{A}_{(i)}(I_i, :)) = r$ *for all* $i \in [n]$.

Note: We have also attained robustness results with respect to sparse corruptions.

"Robust Tensor CUR: Rapid Low-Tucker-Rank Tensor Recovery with Sparse Corruptions"
by H. Cai, Z. Chao, L. Huang, D. Needell. Submitted, 2022.

CUR Factorization

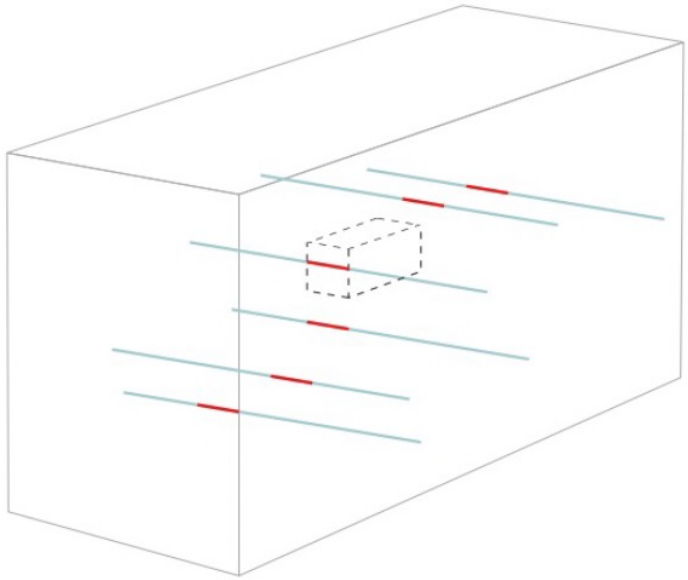


Figure 2: Illustration of the Fiber CUR Decomposition of Theorem 3.3 in which J_i is not necessarily related to I_i. The lines correspond to rows of C_2, and red indices within correspond to rows of U_2. Note that the lines may (but do not have to) pass through the core subtensor $\mathcal{R}$ outlined by dotted lines. Fibers used to form C_1 and C_3 are not shown for clarity.

CUR Factorization

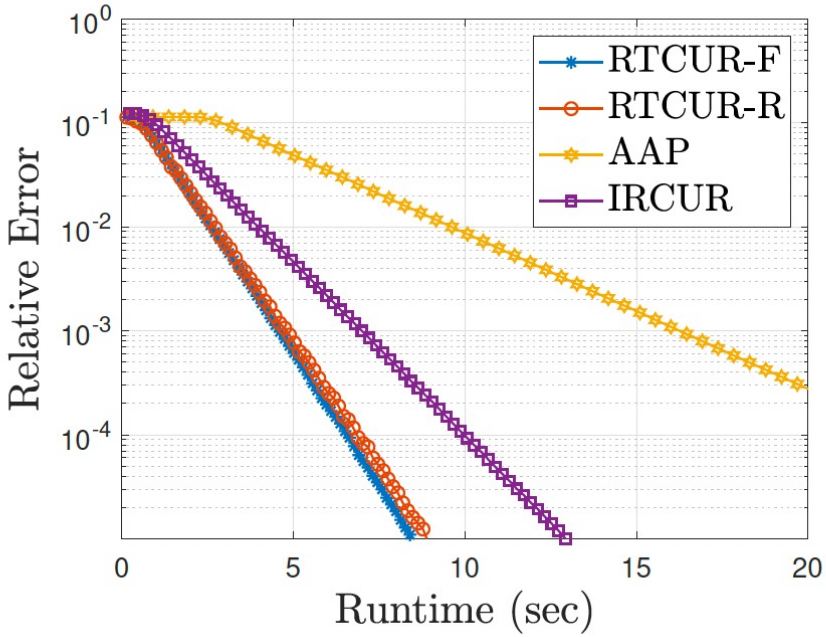
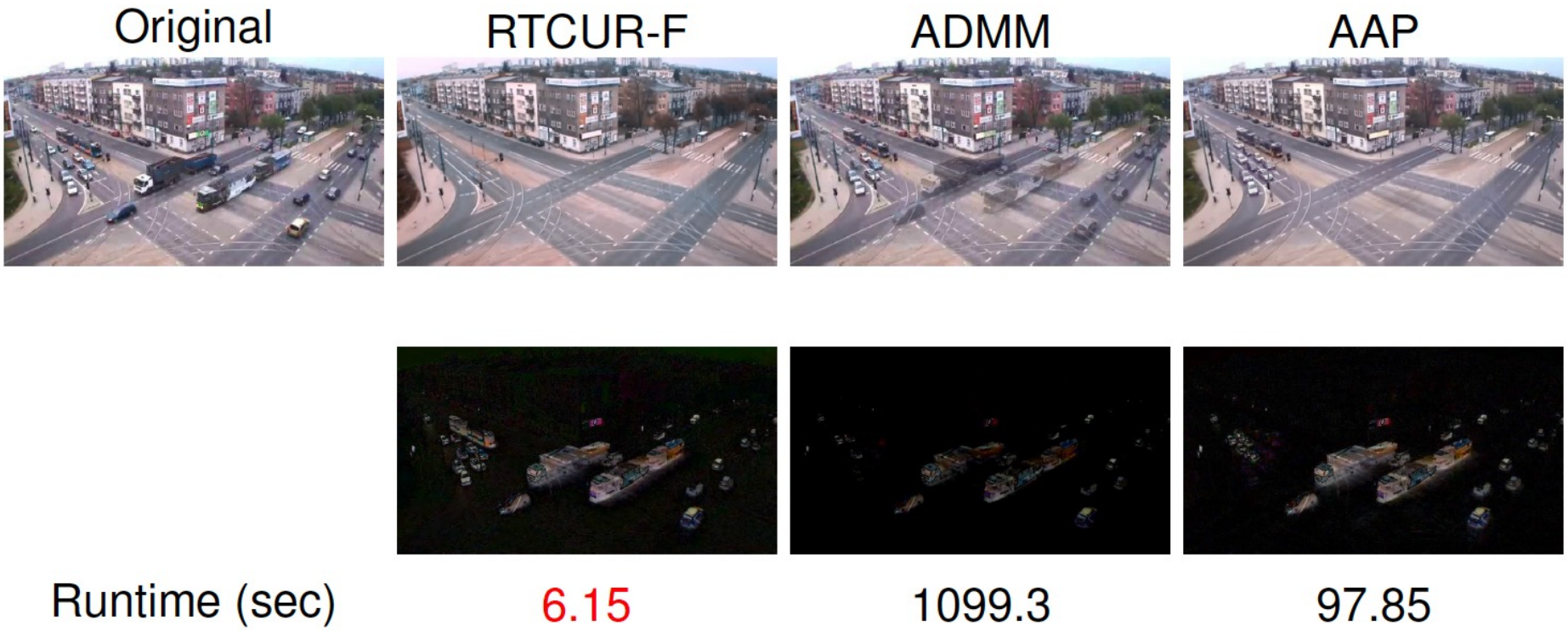


Figure: Runtime vs. relative error comparison: 3-mode tensor with $d = 500$ and multilinear rank $(3, 3, 3)$.

AAP = Accelerated Alternating Projections, IRCUR = Iterated Robust CUR for RPCA

CUR Factorization



CUR Factorization

Figure 5: Face modeling on *ExtYaleB*: Visual comparison of the outputs by RCUR and RPCA for face modeling task. The first row contains the original face images. The second and third rows are the face models and the facial occlusions outputted by RCUR, respectively. The last two rows are the face models and the facial occlusions outputted by RPCA, respectively.

Fair NMF (in progress)

- Regularizing with sup-norm not ideal (outliers, results still unfair)
- Enforcing objective function to maximize fairness across groups

$$\underset{W_S \in \mathbb{R}^{m\times d}_{\geq 0} H_S \in \mathbb{R}^{d\times |S|}_{\geq 0} \forall S \in \{M,A,B\}}{\operatorname{argmin}} \frac{1}{a^2}\|X_A - W_A H_A\|_F^2 + \frac{1}{b^2}\|X_B - W_B H_B\|_F^2$$
$$+\frac{1}{a^2}\|X_A - W_M H_A\|_F^2 + \frac{1}{b^2}\|X_B - W_M H_B\|_F^2$$
$$+\alpha\left(\frac{1}{a}\|X_A - W_M H_A\|_F - \frac{1}{b}\|X_B - W_M H_B\|_F\right)^2$$

- Iterative scheme that hones in on groups not adequately represented