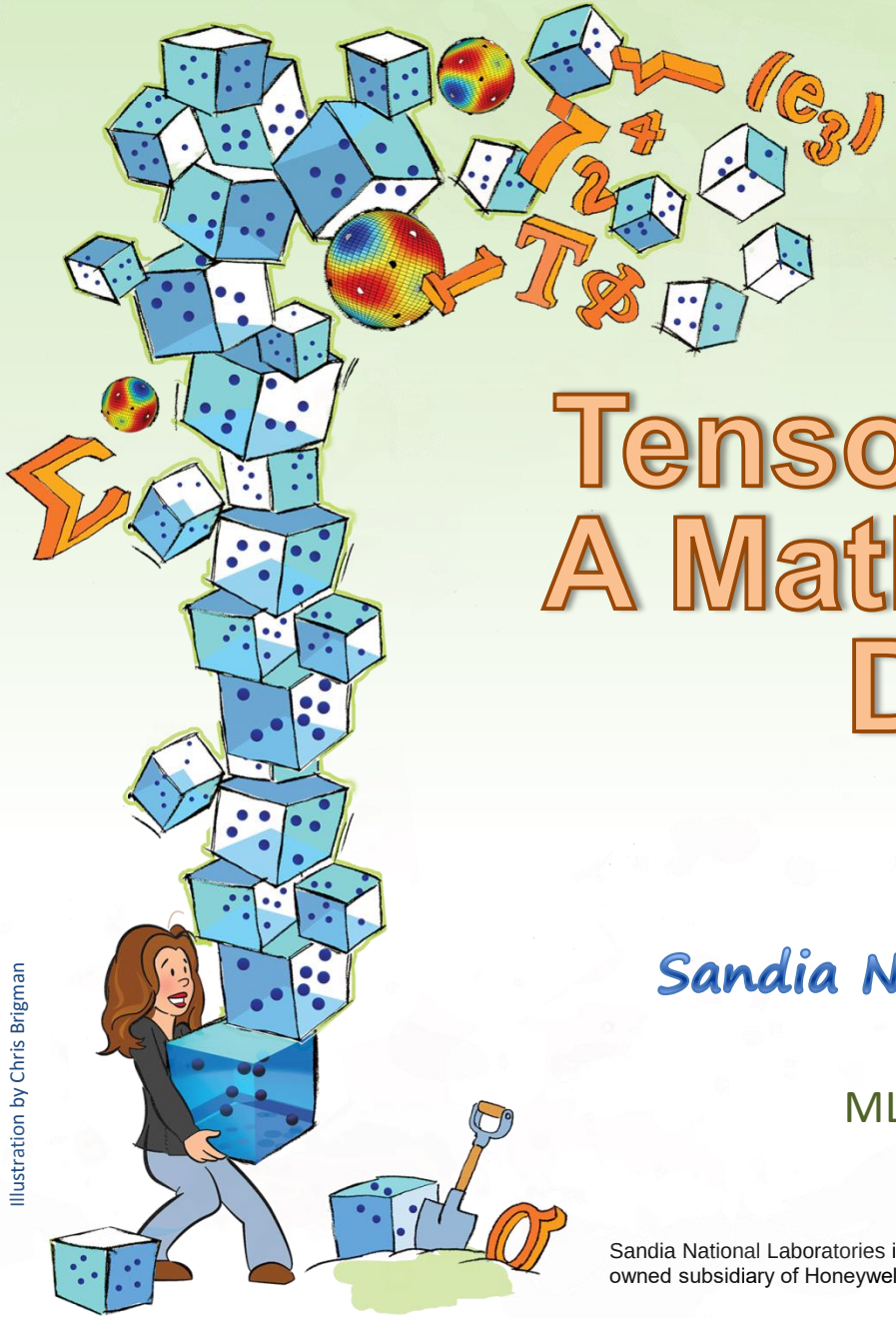


Illustration by Chris Brigman

Sandia National Laboratories

SAND2017-13021C

Tensor Decomposition: A Mathematical Tool for Data Analysis

Tamara G. Kolda

Sandia National Laboratories, Livermore, CA

MLconf: The Machine Learning Conference

San Francisco, CA, Nov 10, 2017

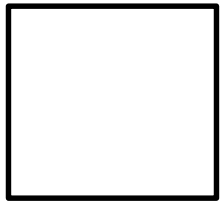
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

U.S. DEPARTMENT OF ENERGY

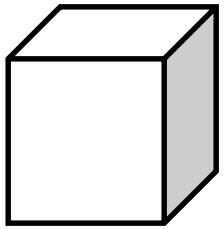
A Tensor is an d-Way Array

Vector
d = 1

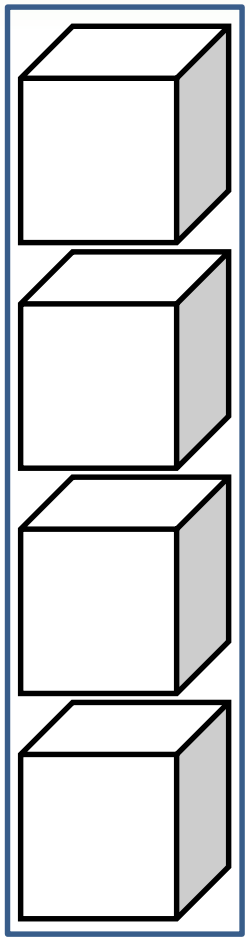
Matrix
d = 2



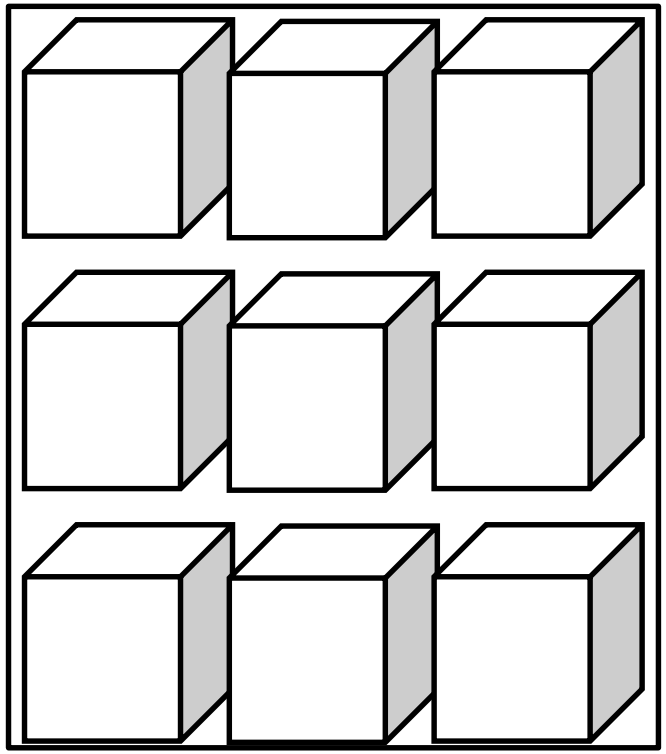
3rd-Order Tensor
d = 3



4th-Order Tensor
d = 4



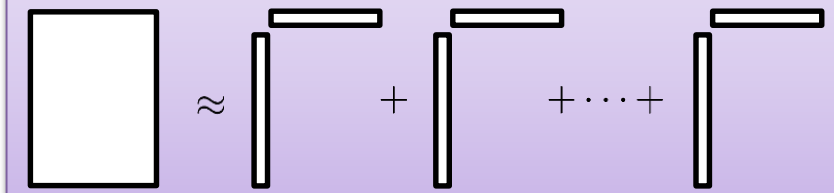
5th-Order Tensor
d = 5



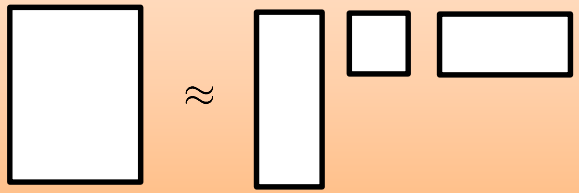
From Matrices to Tensors: Two Points of View

Singular value decomposition (SVD), eigendecomposition (EVD), nonnegative matrix factorization (NMF), sparse SVD, CUR, etc.

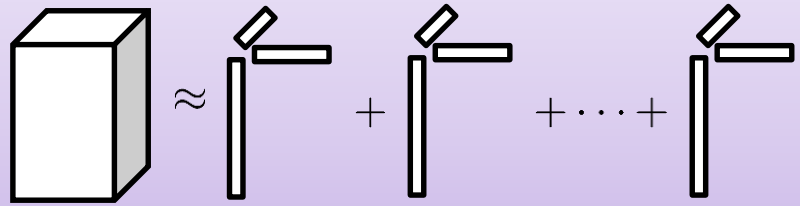
Viewpoint 1: Sum of outer products, useful for interpretation



Viewpoint 2: High-variance subspaces, useful for compression

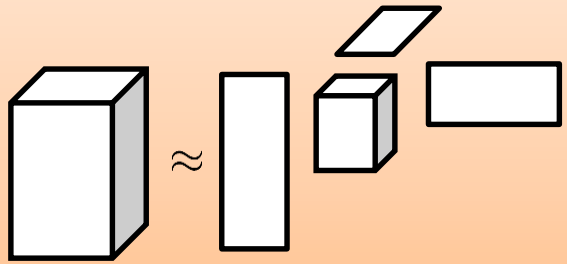


CP Model: Sum of d-way outer products, useful for interpretation



CANDECOMP, PARAFAC, Canonical Polyadic, CP

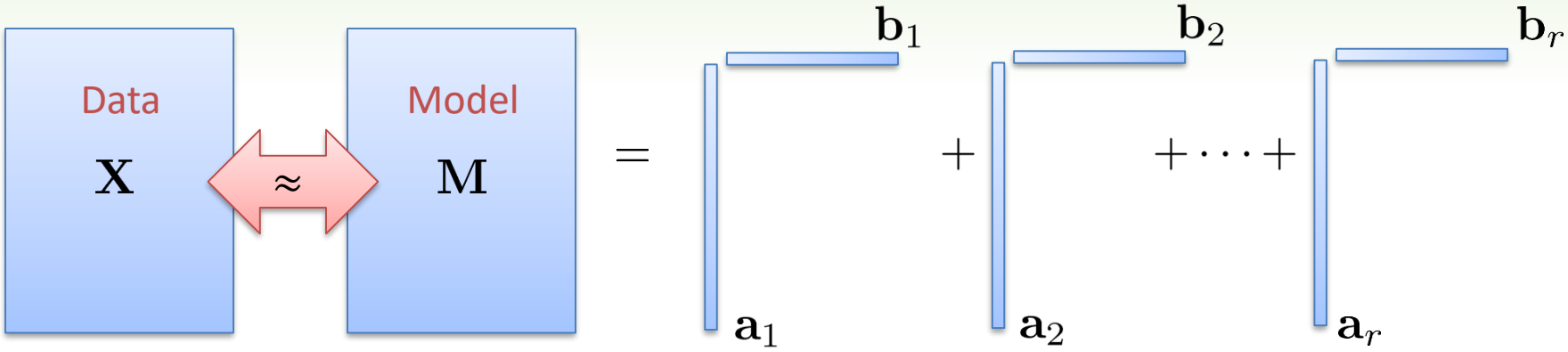
Tucker Model: Project onto high-variance subspaces to reduce dimensionality



HO-SVD, Best Rank-(R_1,R_2,…,R_d) decomposition

Other models for compression include hierarchical Tucker and tensor train.

Matrix Factorization



$$\mathbf{X} \approx \mathbf{M} = \sum_{j=1}^{r} \mathbf{a}_j \mathbf{b}_j' = \mathbf{A}\mathbf{B}'$$

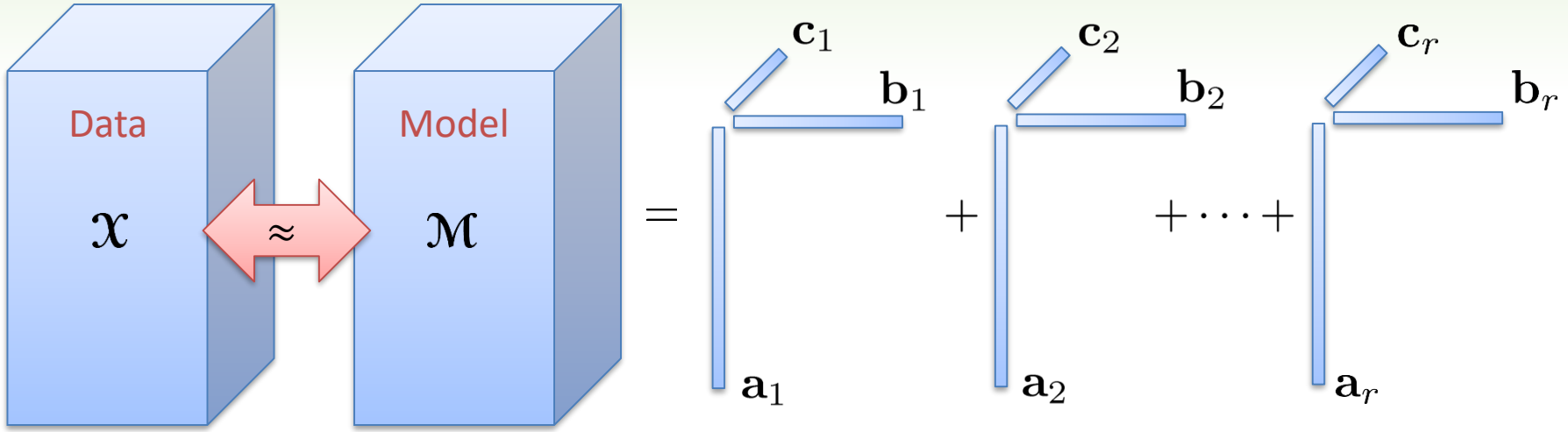
Standard Matrix Formulation

$$\mathbf{X} \approx \mathbf{M} = \sum_{j=1}^{r} \mathbf{a}_j \circ \mathbf{b}_j = [\![\mathbf{A}, \mathbf{B}]\!]$$

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{X} - \mathbf{M}\|^2 \equiv \sum_{ij} (x_{ij} - m_{ij})^2 \text{ s.t. } \mathbf{M} = [\![\mathbf{A}, \mathbf{B}]\!]$$

CP Tensor Factorization (3-way)

CP = CANDECOMP/PARAFAC or Canonical Polyadic

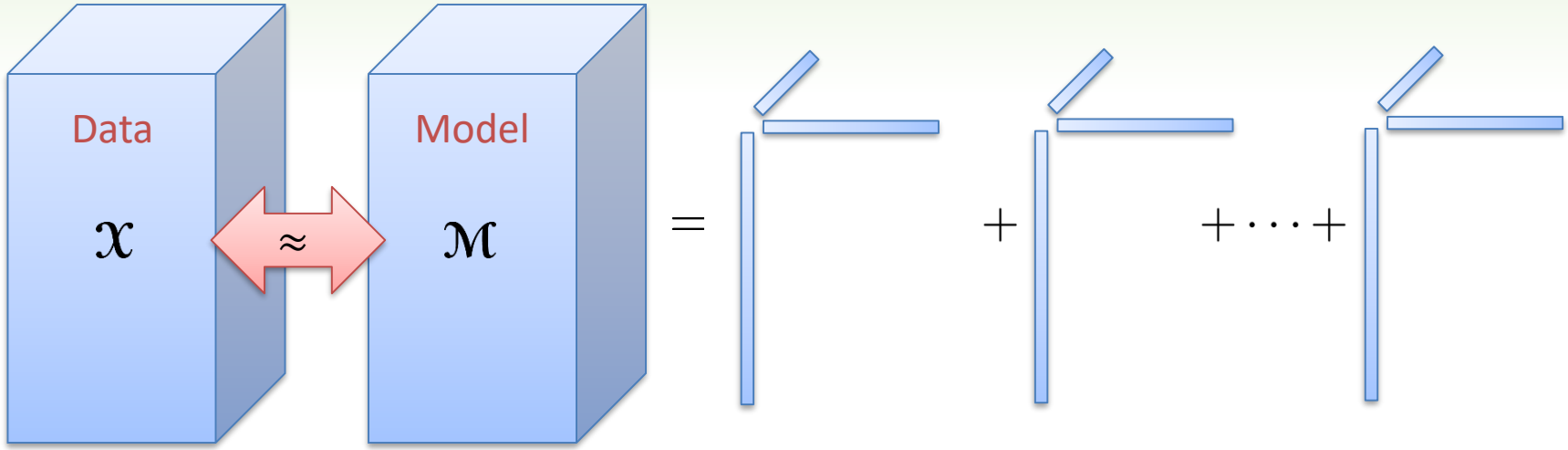


$$\mathcal{X} \approx \mathcal{M} = \sum_{j=1}^{r} \mathbf{a}_j \circ \mathbf{b}_j \circ \mathbf{c}_j = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathcal{X} - \mathcal{M}\|^2 \equiv \sum_{ijk} (x_{ijk} - m_{ijk})^2 \text{ s.t. } \mathcal{M} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

Hitchcock 1927, Harshman 1970, Carroll & Chang 1970

CP Tensor Factorization (d-way)



$$\mathcal{X} \approx \mathcal{M} = [\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d]\!]$$

$$\min_{\mathbf{A}_1,\ldots,\mathbf{A}_d} \|\mathcal{X} - \mathcal{M}\|^2 \text{ s.t. } \mathcal{M} = [\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d]\!]$$

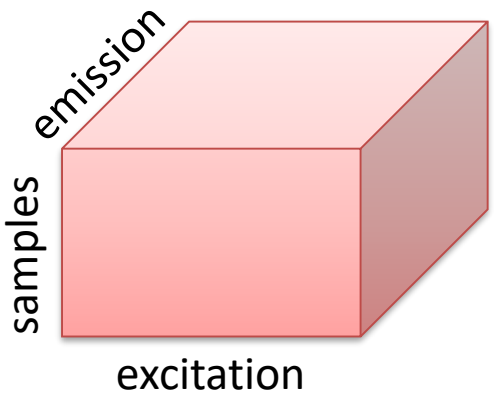
Hitchcock 1927, Harshman 1970, Carroll & Chang 1970

Example: Determining Analytes in Chemical Samples

R. Bro, **PARAFAC: Tutorial and Applications**, *Chemometrics and Intelligent Laboratory Systems*, 38:149-171, 1997

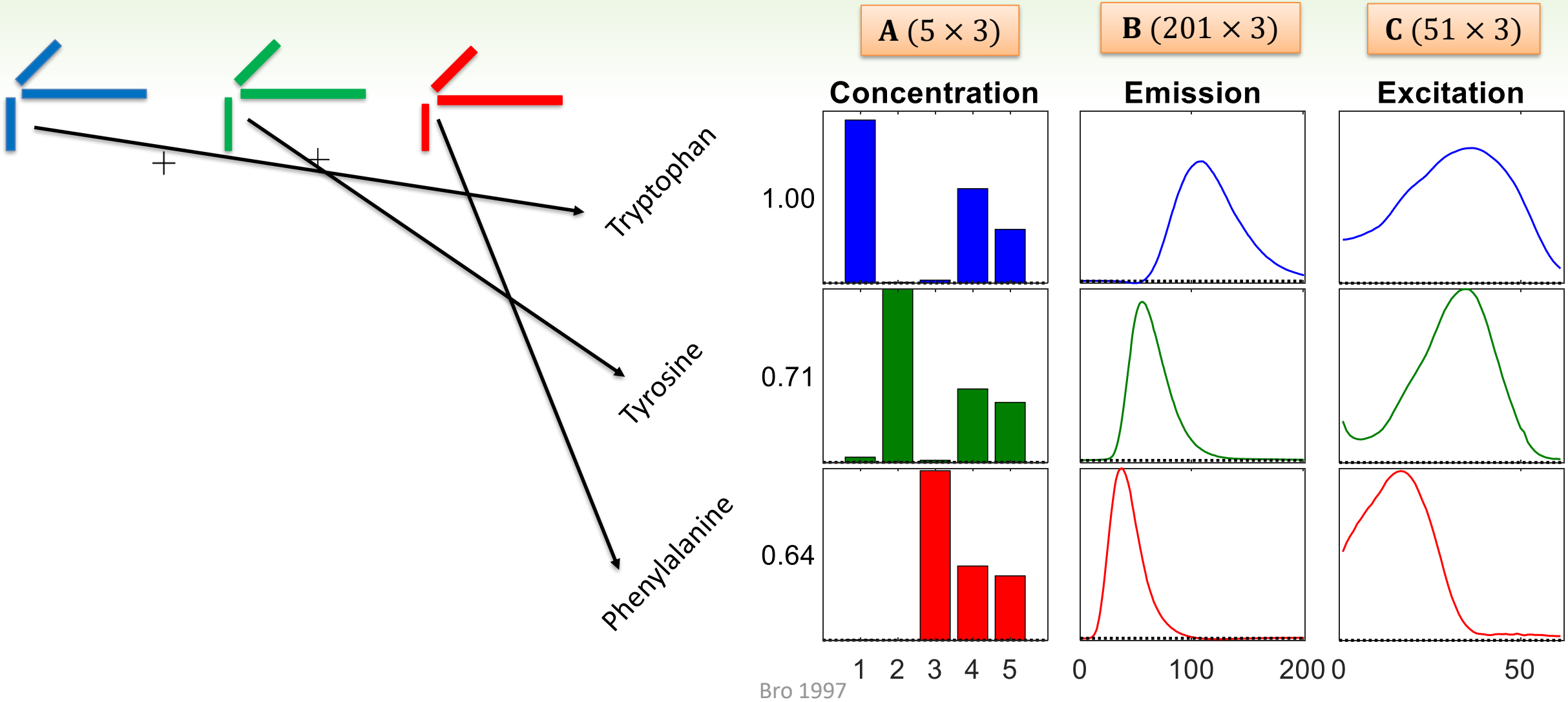
Amino Acids Fluroescence Dataset

- Fluorescence measurements of 5 samples containing 3 amino acids
  - Tryptophan
  - Tyrosine
  - Phenylalanine
- Tensor of size 5 x 51 x 201
  - 5 samples
  - 51 excitations
  - 201 emissions



R. Bro, PARAFAC: Tutorial and Applications, Chemometrics and Intelligent Laboratory Systems, 38:149-171, 1997

Rank-3 CP Factorization of Amino Acids Data



Bro 1997

Example: CP for Mouse Neural Activity

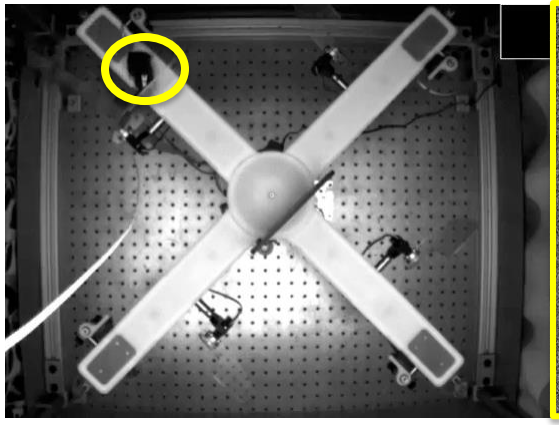
A. H. Williams, T. H. Kim, F. Wang, S. Vyas, S. I. Ryu, K. V. Shenoy, M. Schnitzer, T. G. Kolda, S. Ganguli. **Unsupervised Discovery of Demixed, Low-dimensional Neural Dynamics across Multiple Timescales through Tensor Components Analysis**. bioRxiv, 2017. https://doi.org/10.1101/211128

Motivating Example: Neuron Activity in Learning

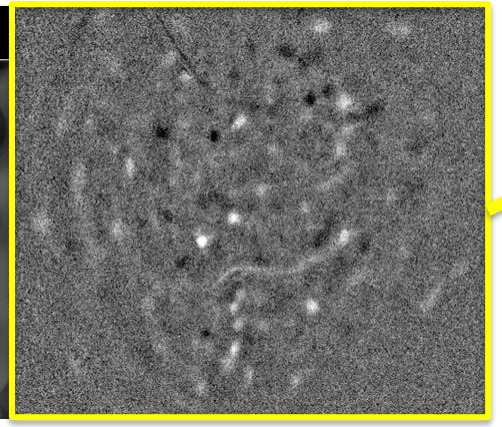
Thanks to Schnitzer Group @ Stanford
Mark Schnitzer, Fori Wang, Tony Kim

Microscope by Inscopix

mouse in "maze"



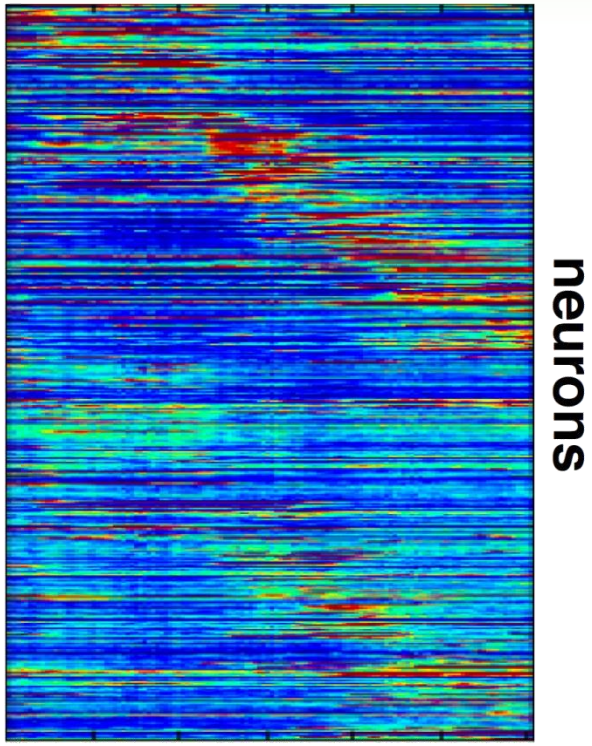
neural activity



One Column of Neuron x Time Matrix

One Trial

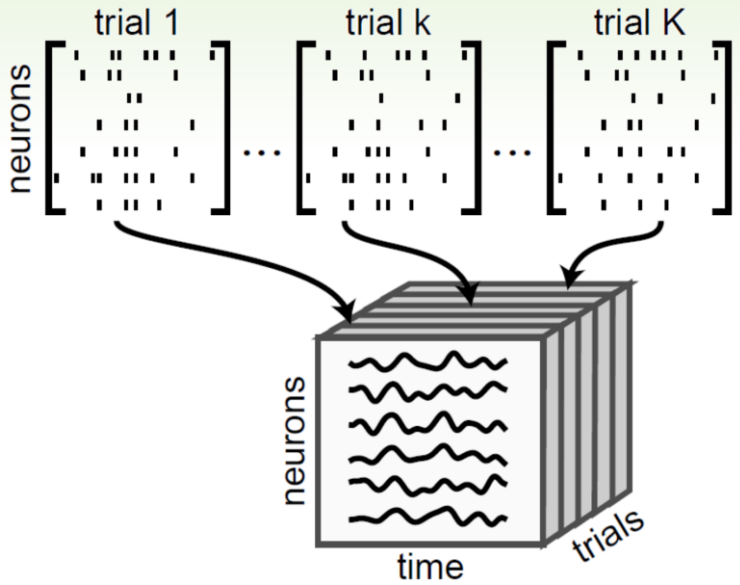
300 neurons × 120 time bins

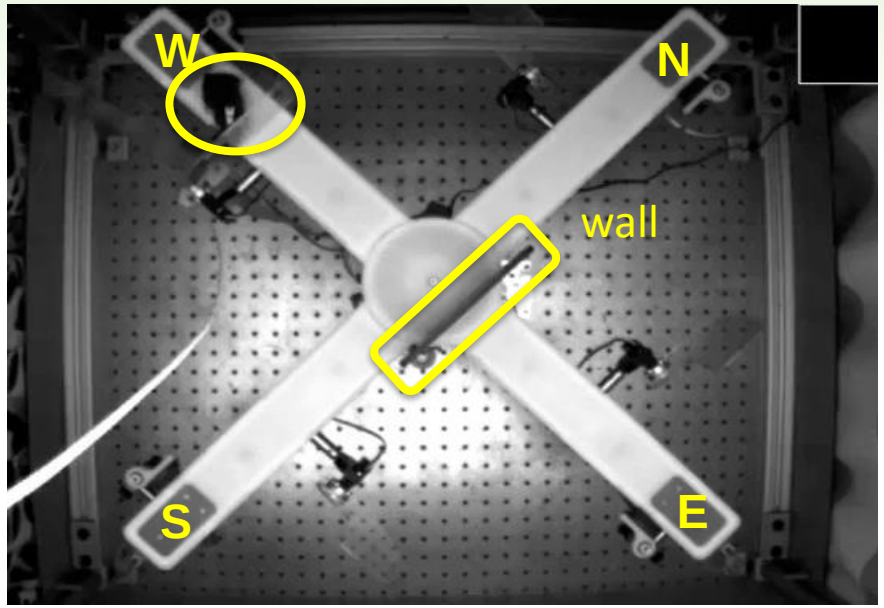


× 600 trials (over 5 days)

Williams et al., bioRxiv, 2017, DOI:10.1101/211128

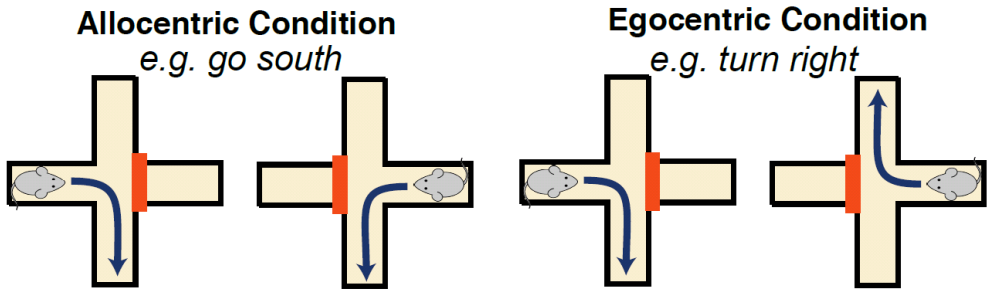
Trials Vary Start Position and Strategies





note different patterns on curtains

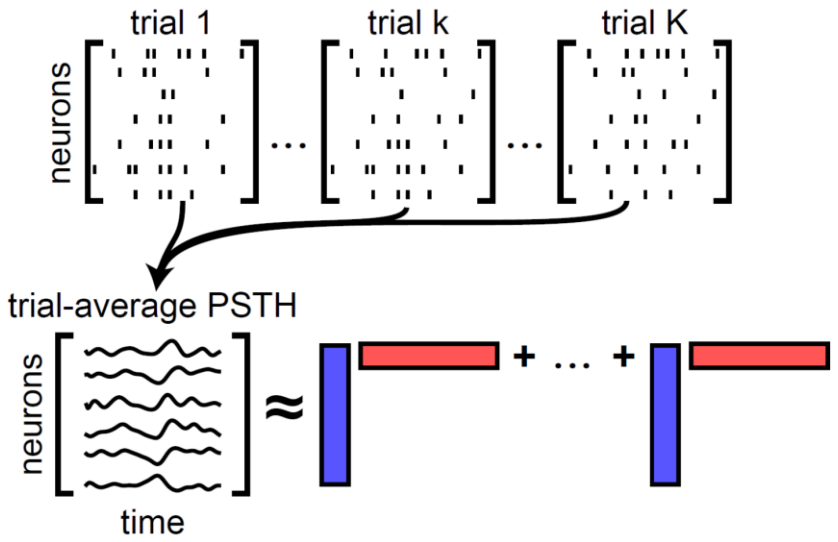
- 600 Trials over 5 Days
- Start West or East
- Conditions Swap Twice
  - ❖ Always Turn South
  - ❖ Always Turn Right
  - ❖ Always Turn South



Williams et al., bioRxiv, 2017, DOI:10.1101/211128

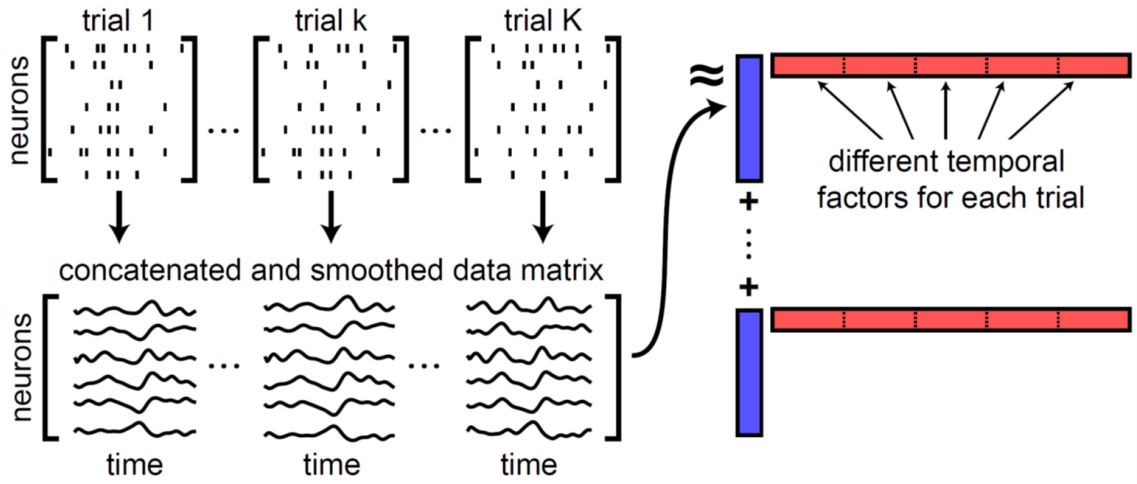
Prior Work Only Considered Two-Way Analyses

Trial Averaging + PCA



Blurs any differences across trials

Trial Concatenation + PCA



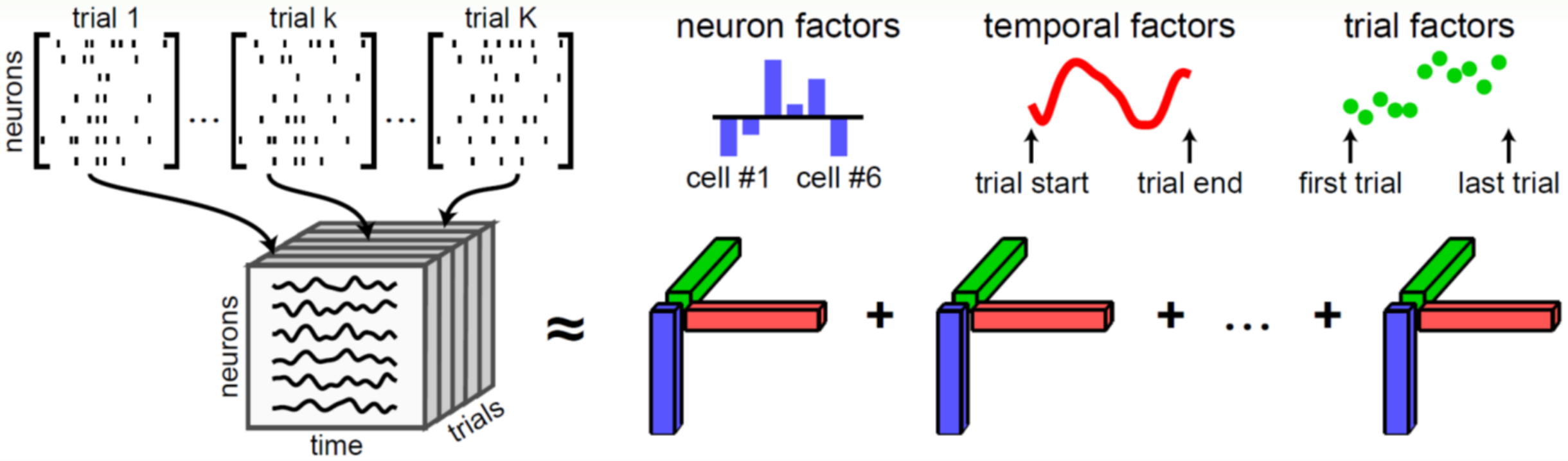
Cannot find associations between trials

Williams et al., bioRxiv, 2017, DOI:10.1101/211128

CP for Simultaneous Analysis of Neurons, Time, and Trial

Sandia National Laboratories

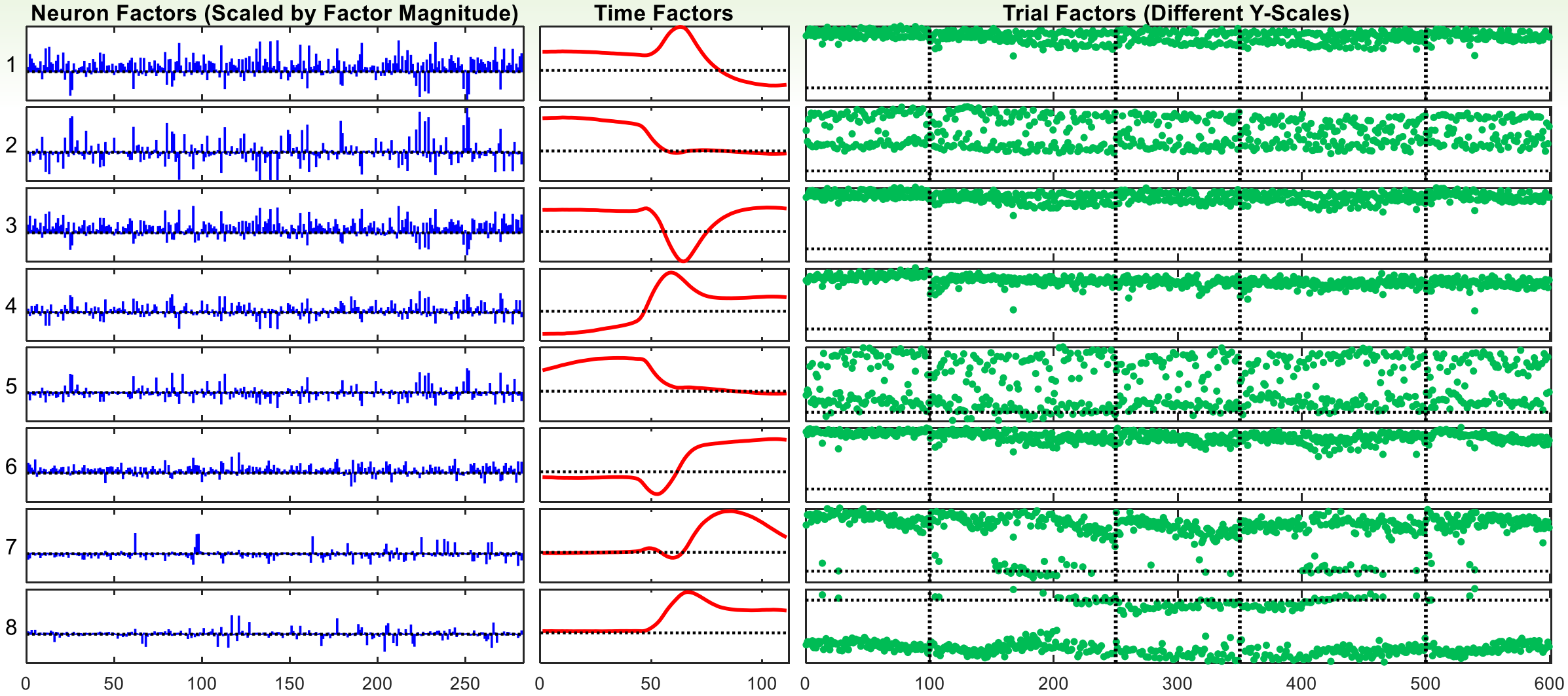
Tensor Components Analysis (TCA)



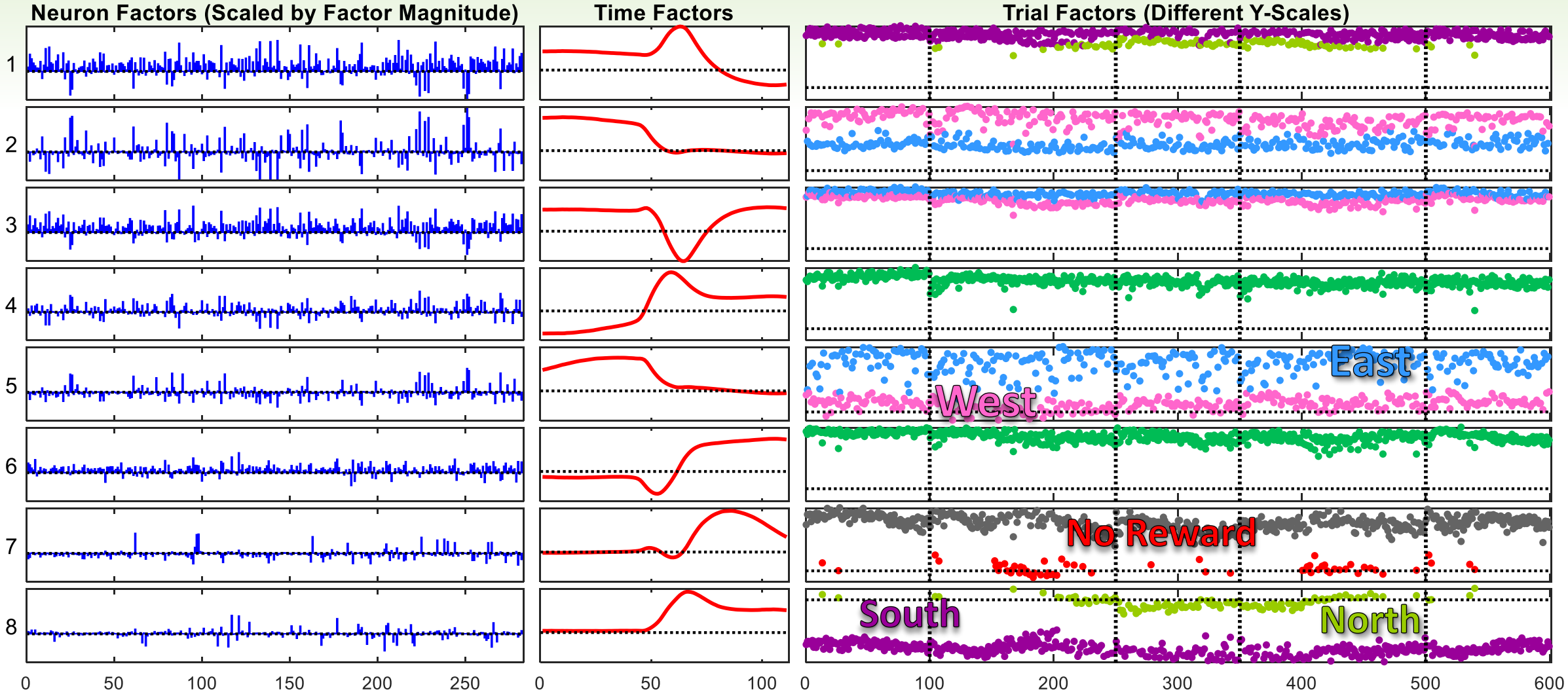
Prior tensor work in neuroscience for fMRI and EEG: Andersen and Rayens (2004), Mørup et al. (2004), Acar et al. (2007), De Vos et al. (2007), and more

Williams et al., bioRxiv, 2017, DOI:10.1101/211128

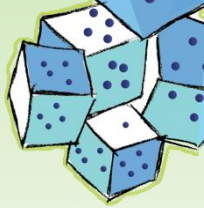
8-Component CP Decomposition of Mouse Neuron Data



Interpretation of Mouse Neuron Data



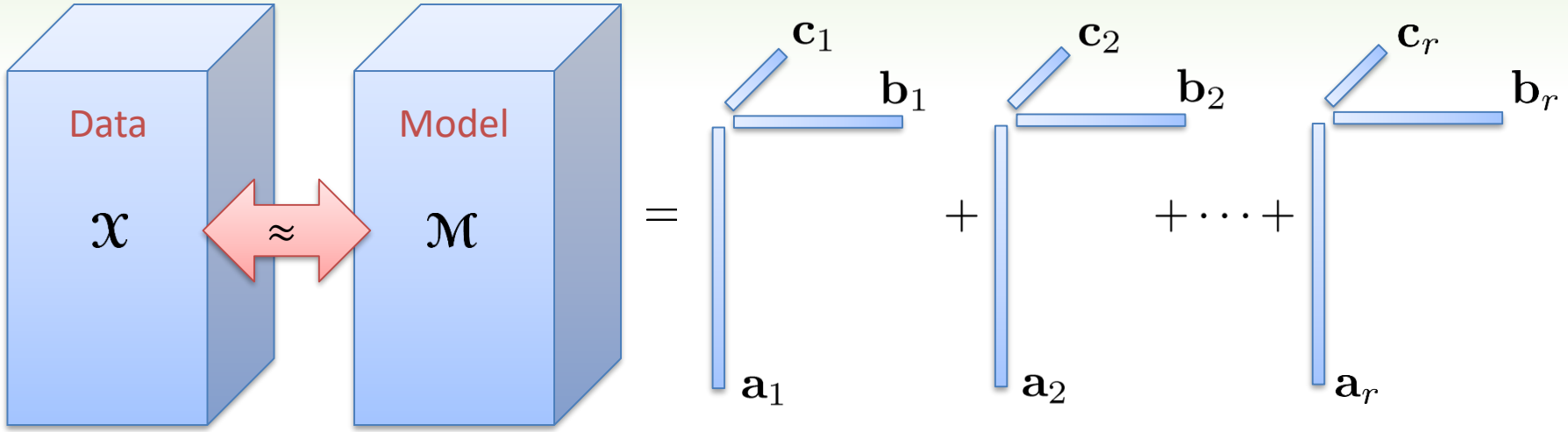
Sandia National Laboratories



Randomized Least Squares for CP Decomposition

C. Battaglino, G. Ballard, T. G. Kolda. **A Practical Randomized CP Tensor Decomposition**. arXiv:1701.06600, 2017.

Tensor Factorization (3-way)

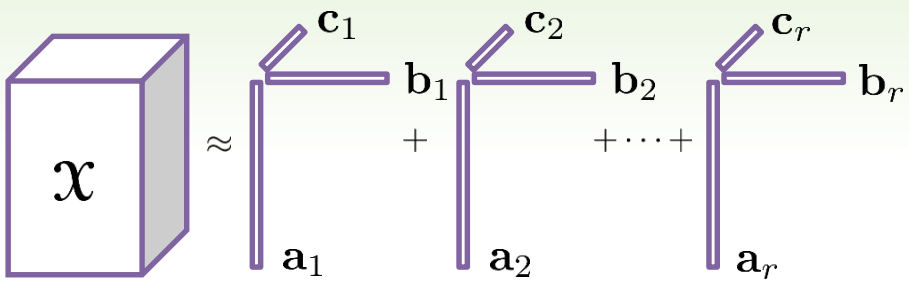


$$\mathcal{X} \approx \mathcal{M} = \sum_{j=1}^{r} \mathbf{a}_j \circ \mathbf{b}_j \circ \mathbf{c}_j = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathcal{X} - \mathcal{M}\|^2 \text{ s.t. } \mathcal{M} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

We can rewrite this as a matrix equation in **A**, **B**, or **C**.

CP-ALS: Fitting CP Model via Alternating Least Squares



$$f(\mathbf{A},\mathbf{B},\mathbf{C}) = \sum_{ik\ell}(x_{ik\ell} - \sum_{j} a_{ij}b_{kj}c_{\ell j})^2$$

Repeat until convergence:

Step 1: $\min_{\mathbf{A}} \sum_{ik\ell}\left(x_{ik\ell} - \sum_{j} a_{ij}\, b_{kj}\, c_{\ell j}\right)^2$

Step 2: $\min_{\mathbf{B}} \sum_{ik\ell}\left(x_{ik\ell} - \sum_{j} a_{ij}\, b_{kj}\, c_{\ell j}\right)^2$

Step 3: $\min_{\mathbf{C}} \sum_{ik\ell}\left(x_{ik\ell} - \sum_{j} a_{ij}\, b_{kj}\, c_{\ell j}\right)^2$

- **Rank (R) NP-Hard:** Even best low-rank solution may not exist (Håstad 1990, Silva & Lim 2006, Hillar & Lim 2009)
- **Not nested:** Best rank-(R-1) factorization may not be part of best rank-R factorization (Kolda 2001)
- **Nonconvex:** But convex linear least squares problems
- **Not orthogonal:** Factor matrices are not orthogonal and may even have linearly dependent columns
- **Essentially Unique:** Under modest conditions, CP is unique up to permutation and scaling unique (Kruskal 1977)

Harshman, 1970; Carroll & Chang, 1970

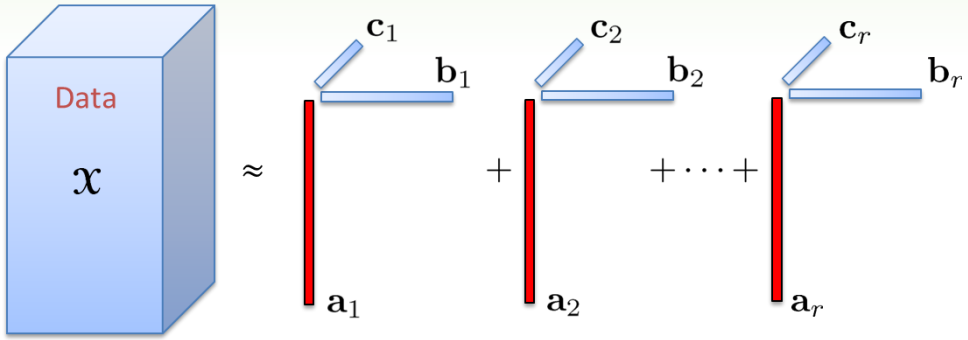
CP-ALS Least Squares Problem

Repeat until convergence…

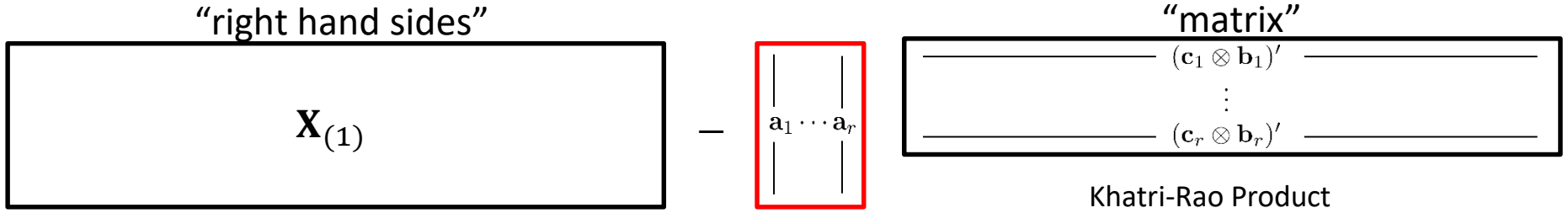
$$\min_{\mathbf{A}} \|\mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})'\|^2$$

$$\min_{\mathbf{B}} \|\mathbf{X}_{(2)} - \mathbf{B}(\mathbf{C} \odot \mathbf{A})'\|^2$$

$$\min_{\mathbf{C}} \|\mathbf{X}_{(3)} - \mathbf{C}(\mathbf{B} \odot \mathbf{A})'\|^2$$



$$\min_{\mathbf{A}} \|\mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})'\|^2$$

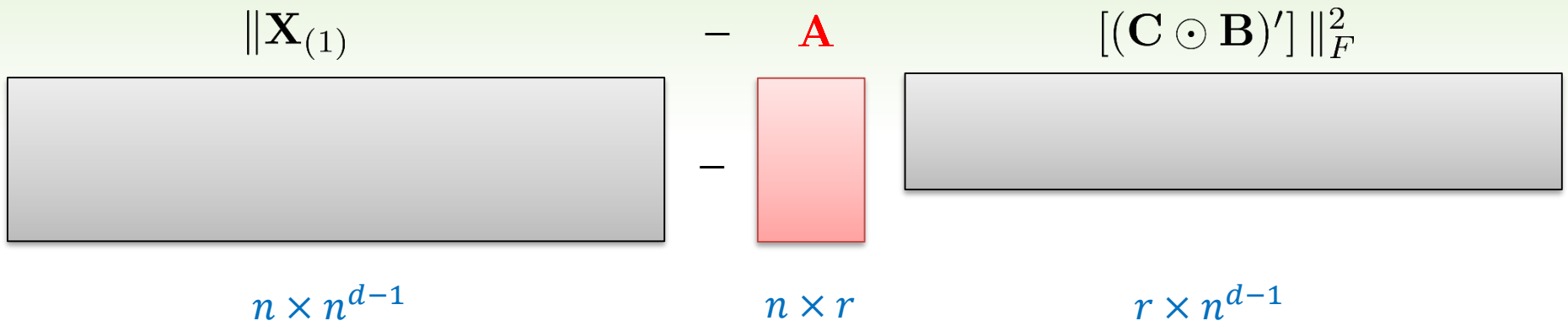


$\mathbf{A}$ $\quad$ $(\mathbf{C} \odot \mathbf{B})'$

$n \times n^2$ $\quad$ $n \times r$ $\quad$ $r \times n^2$

$n \times n^{d-1}$ $\quad$ $n \times r$ $\quad$ $r \times n^{d-1}$

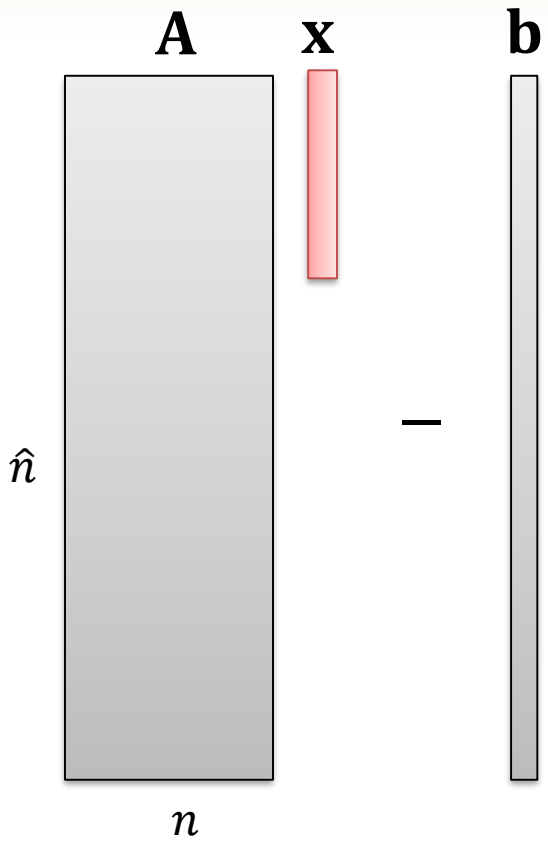
CP Least Squares Problem



How to randomize this?

Aside: Sketching for Standard Least Squares

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|$$



$$\mathcal{O}(\hat{n}n^2)$$

$$\mathbf{x} = \mathbf{A}^{\dagger}\mathbf{b}$$

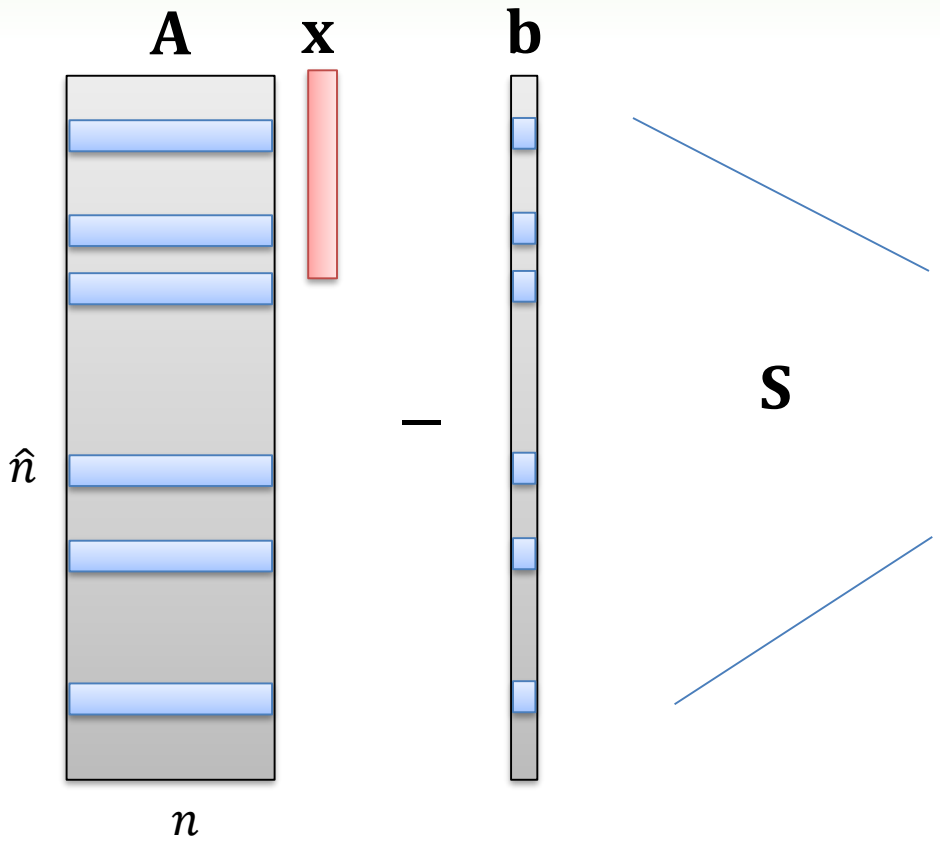
$$\mathbf{x} \leftarrow \mathbf{A} \backslash \mathbf{b}$$

Backslash causes MATLAB to automatically call the best solver (cholesky, qr, etc.)

Sarlós 2006, Woodruff 2014

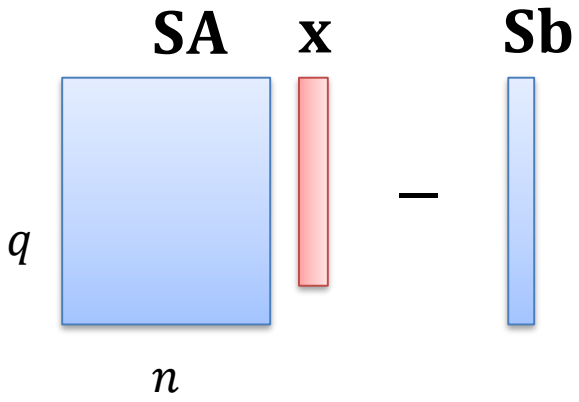
Sampled Least Squares

Choose q rows, uniformly at random



S

approximate



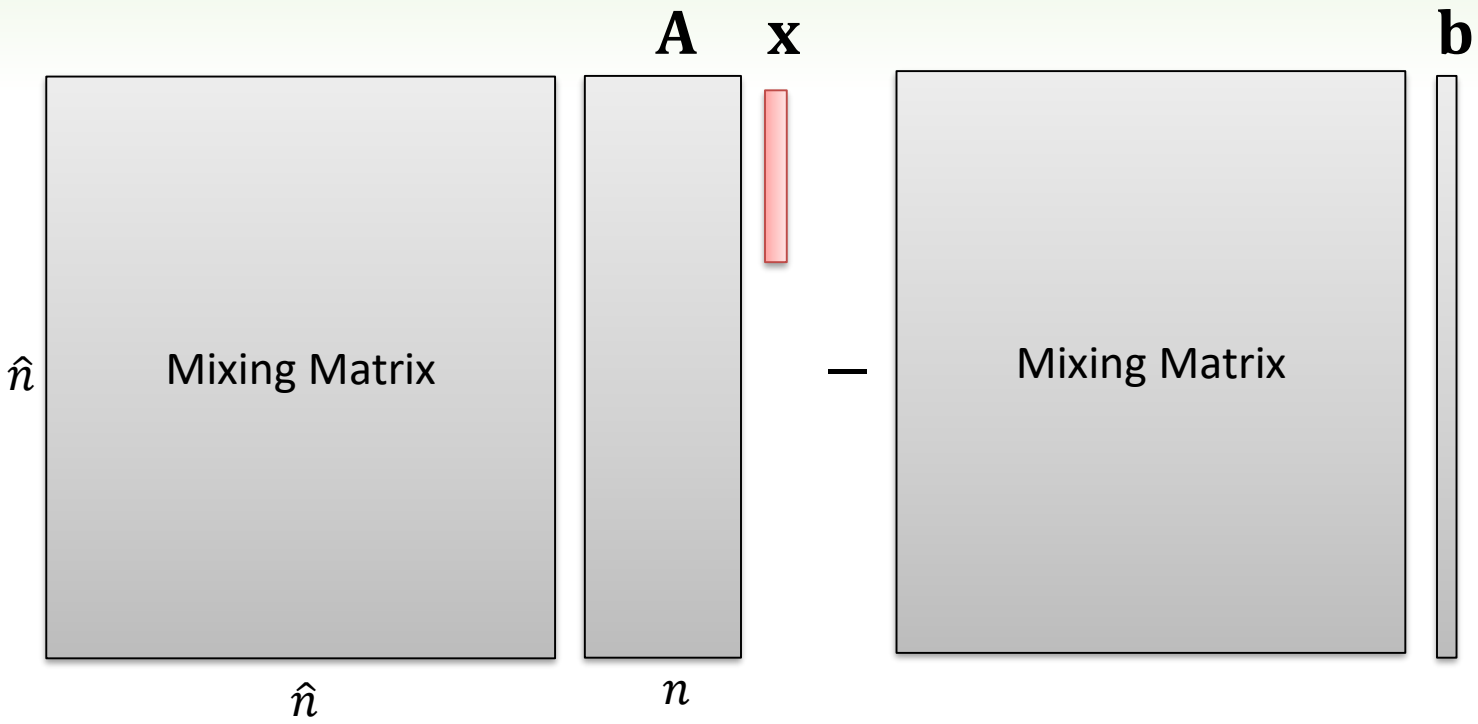
Sampling only guaranteed to "work" if the **A** *is incoherent.*

$\mathcal{O}(\hat{n}n^2)$

$\mathcal{O}(qn^2)$

Sarlós 2006, Woodruff 2014

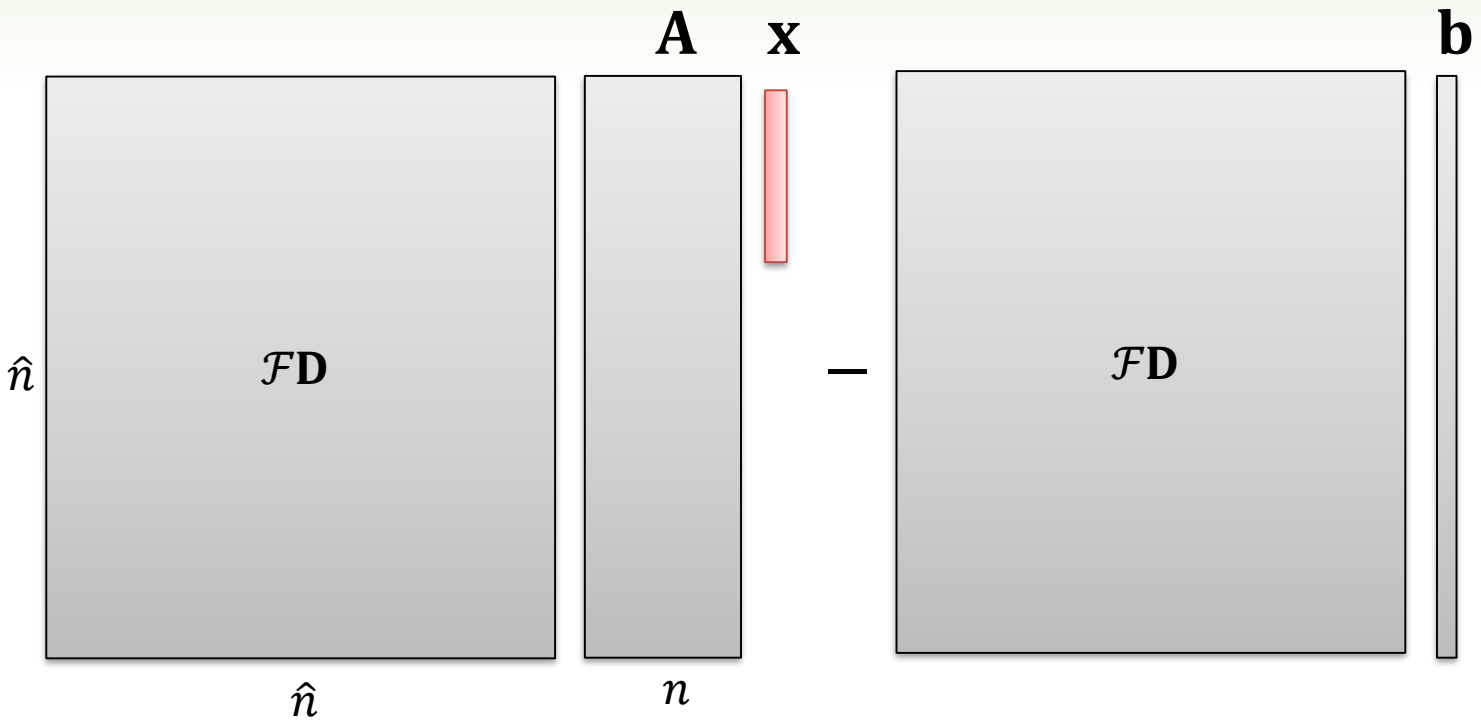
Enforcing Incoherence



- Many good choices of mixing matrix, such as a matrix with entries chosen from a uniform random distribution.
- But no reduction in cost!

Sarlós 2006, Woodruff 2014

Fast Johnson-Lindenstrauss Transform (FJLT)



- Instead, use FFT ($\mathcal{F}$) followed by random diagonal with +/- 1 entries ($\mathbf{D}$).
- Costs only $n \log \hat{n}$ to apply
- Practical application in Blendenpik, yielding ~4X speedup versus LAPACK

Sarlós 2006, Woodruff 2014, Ailon & Chazelle 2006, Avron, Maymounkov, & Toledo 2010

Sampled/Mixed Least Squares

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|$$

$$\min_{\mathbf{x}} \|\mathbf{S}\mathbf{A}\mathbf{x} - \mathbf{S}\mathbf{b}\|$$

Sampling only, No mixing.

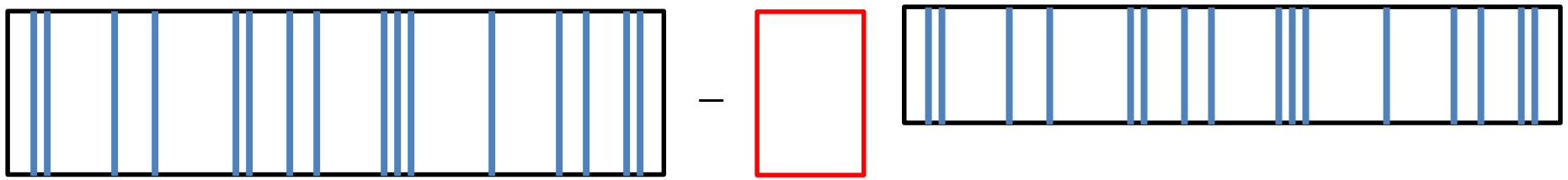
$$\min_{\mathbf{x}} \|\mathbf{S}\mathcal{F}\mathbf{D}\mathbf{A}\mathbf{x} - \mathbf{S}\mathcal{F}\mathbf{D}\mathbf{b}\|$$

Sampling + Mixing

CP-ALS-RAND

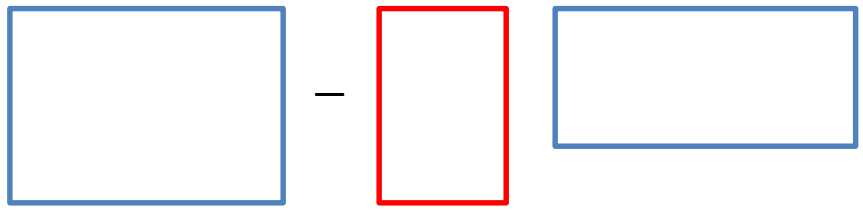
$$\min_{\mathbf{A}} \|\mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})'\|^2$$

$$\mathbf{A} \leftarrow \mathbf{X}_{(1)}/(\mathbf{C} \odot \mathbf{B})'$$



$$\|\mathbf{X}_{(1)}\mathbf{S}' - \mathbf{A}[(\mathbf{C} \odot \mathbf{B})'\mathbf{S}']\|_F^2$$

$$\mathbf{A} \leftarrow \mathbf{X}_{(1)}\mathbf{S}'/(\mathbf{C} \odot \mathbf{B})'\mathbf{S}'$$



Battaglino, Ballard, & Kolda 2017

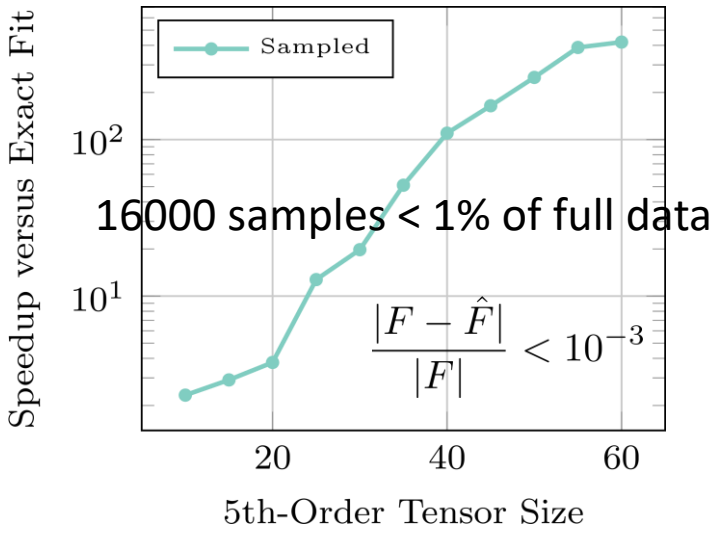
Randomizing the Convergence Check

Repeat until convergence…

$$\min_{\mathbf{A}} \|\mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})'\|^2$$

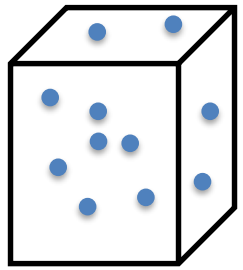
$$\min_{\mathbf{B}} \|\mathbf{X}_{(2)} - \mathbf{B}(\mathbf{C} \odot \mathbf{A})'\|^2$$

$$\min_{\mathbf{C}} \|\mathbf{X}_{(3)} - \mathbf{C}(\mathbf{B} \odot \mathbf{A})'\|^2$$



$$F(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \sum_{ijk} (x_{ijk} - m_{ijk})^2$$

$$\text{s.t. } \mathcal{M} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$



Estimate convergence of function values using small random subset of elements in function evaluation (use Chernoff-Hoeffding to bound accuracy)

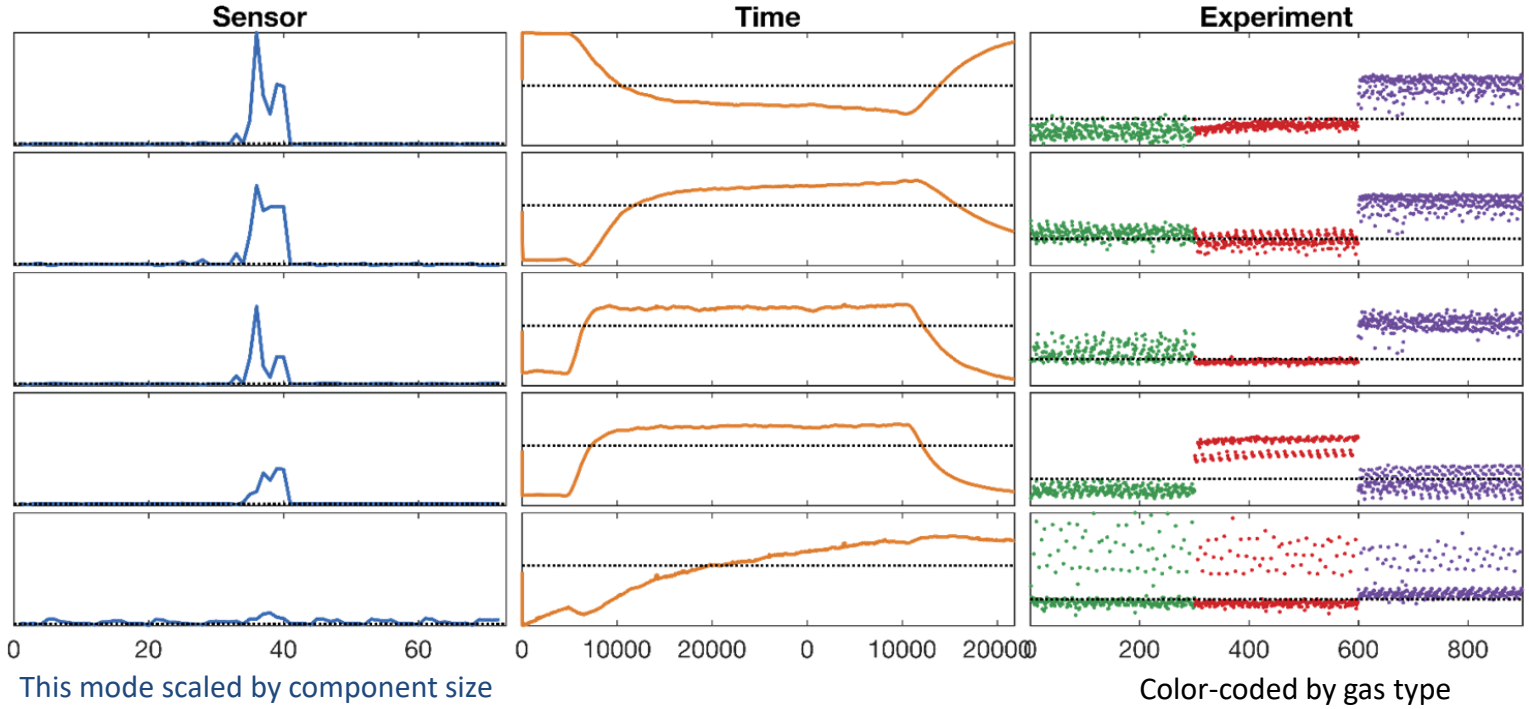
$$\hat{F}(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \frac{n^d}{|\hat{\Omega}|} \sum_{i \in \hat{\Omega}} (x_i - m_i)^2$$

Battaglino, Ballard, & Kolda 2017

Speed Advantage: Analysis of Hazardous Gas Data

900 experiments (with three different gas types) x 72 sensors x 25,900 time steps (13 GB)

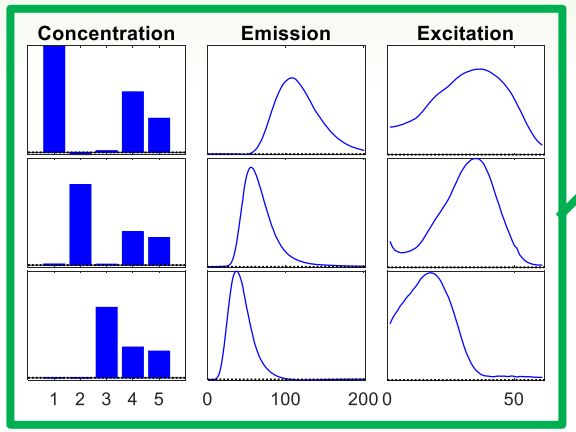
Method	Median Time (s)	Median Fit	Median Classification Error
CPRAND	53.6	0.715	0.61%
CP-ALS (H)	578.4	0.724	0.67%
CP-ALS (R)	204.7	0.724	0.67%



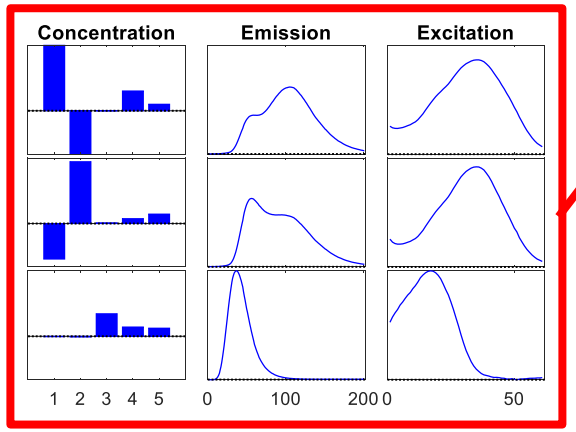
Data from Vergara et al. 2013; see also Vervliet and De Lathauwer (2016)

Battaglino, Ballard, & Kolda 2017

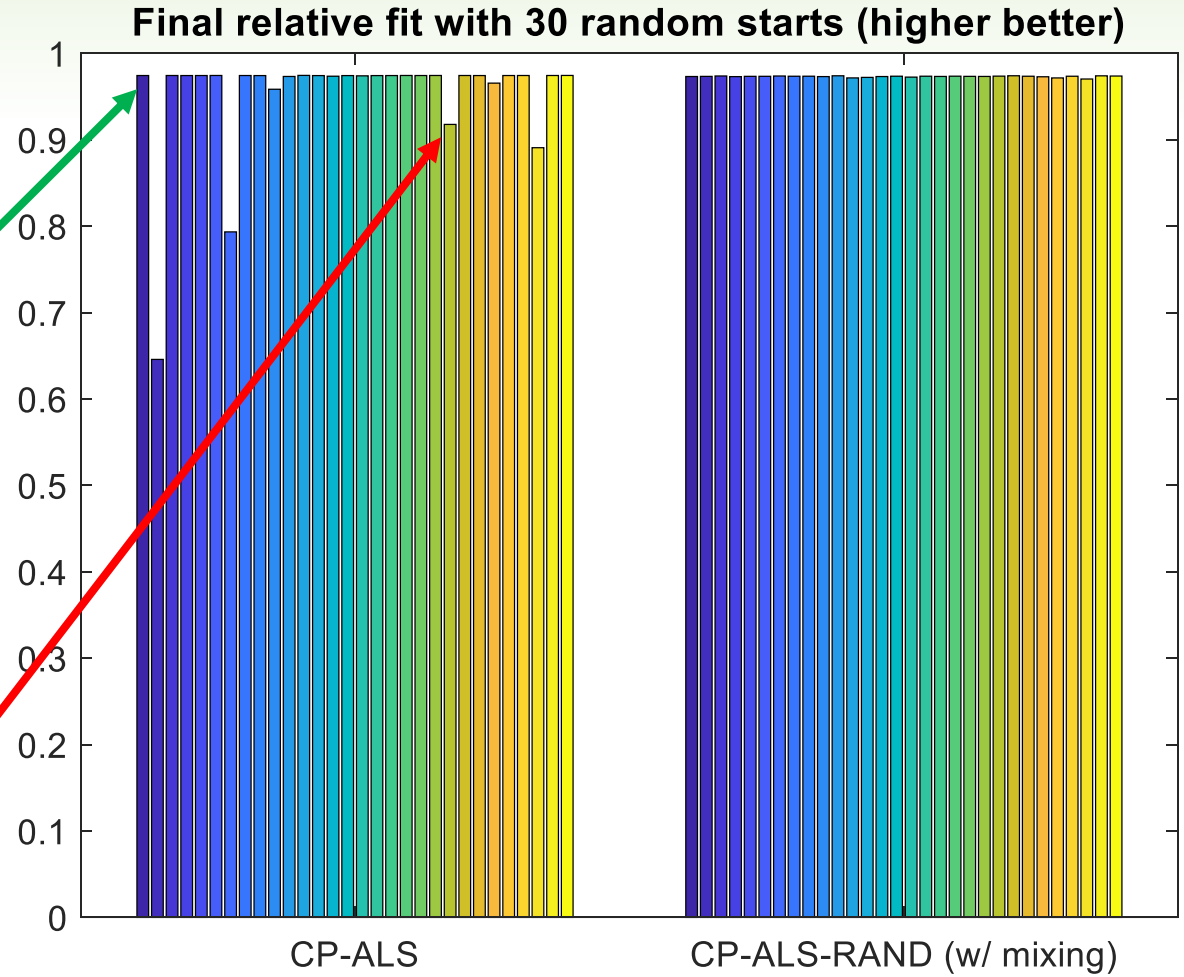
Globalization Advantage? Amino Acids Data



Fit = 0.97

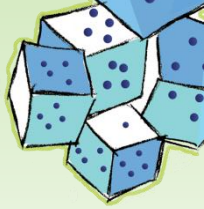


Fit = 0.92



Benefits are not as clear without mixing.

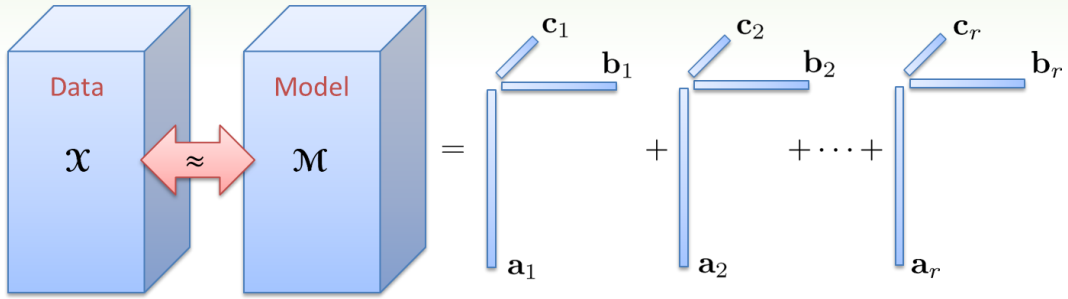
Sandia National Laboratories



Generalized CP Decomposition

Cliff Anderson-Bergman, J. Duersch, D. Hong, T. G. Kolda, **Generalized Canonical Polyadic Tensor Decomposition**, 2017 (coming soon)

Generalizing the Goodness-of-Fit Criteria



$$\mathcal{X} \approx \mathcal{M} = \sum_{j=1}^{r} \mathbf{a}_j \circ \mathbf{b}_j \circ \mathbf{c}_j = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{ijk} (x_{ijk} - m_{ijk})^2 \text{ s.t. } \mathcal{M} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

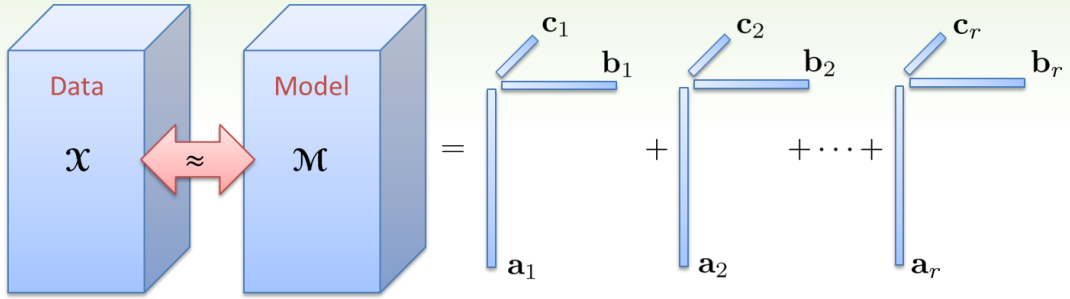
$$F(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \sum_{ijk} (x_{ijk} - m_{ijk})^2$$

$$F(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \sum_{ijk} f(x_{ijk}, m_{ijk})$$

Similar ideas have been proposed in matrix world,
e.g., Collins, Dasgupta, Schapire 2002

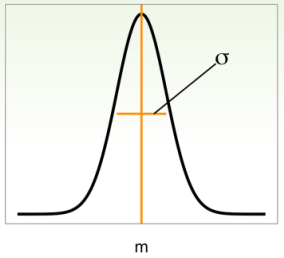
Anderson-Bergman, Duersch, Hong, Kolda 2017

“Standard” CP



Gaussian Probability Density Function (PDF)

$$\frac{e^{-(x-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$$



Typically: Consider data to be low-rank plus “white noise”

$$x_{ijk} \sim m_{ijk} + \mathcal{N}(0, \sigma)$$

Equivalently, Gaussian with mean m_{ijk}

$$x_{ijk} \sim \mathcal{N}(m_{ijk}, \sigma)$$

Link:

$$\mathbb{E}(X_{ijk}) = m_{ijk}$$

Minimize negative log likelihood:

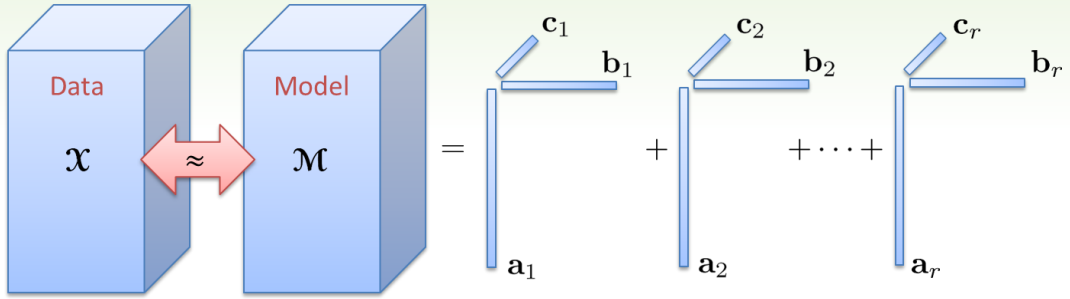
$$-\log(L(\mathcal{M})) = \sum_i \frac{(x_{ijk} - m_{ijk})^2}{2\sigma^2} + \tfrac{1}{2}\log 2\pi\sigma^2$$

Results in the “standard” objective:

$$\min F(\mathcal{X}, \mathcal{M}) = \sum_{ijk} \underbrace{(x_{ijk} - m_{ijk})^2}_{f(x_i, m_i)}$$

Anderson-Bergman, Duersch, Hong, Kolda 2017

“Boolean CP”: Odds Link



Consider data to be Bernoulli distributed with probability p_{ijk}

Convert from probability to odds:

$$p_{ijk} = \frac{m_{ijk}}{1 + m_{ijk}} \Leftrightarrow m_{ijk} = \frac{p_{ijk}}{1 - p_{ijk}}$$

$$\mathbb{E}(X_{ijk}) = p_{ijk}$$

$$x_{ijk} \sim \text{Bernoulli}(m_{ijk}/(1 + m_{ijk}))$$

Probability Mass Function (PMF):

$$p^x(1-p)^{1-x}$$

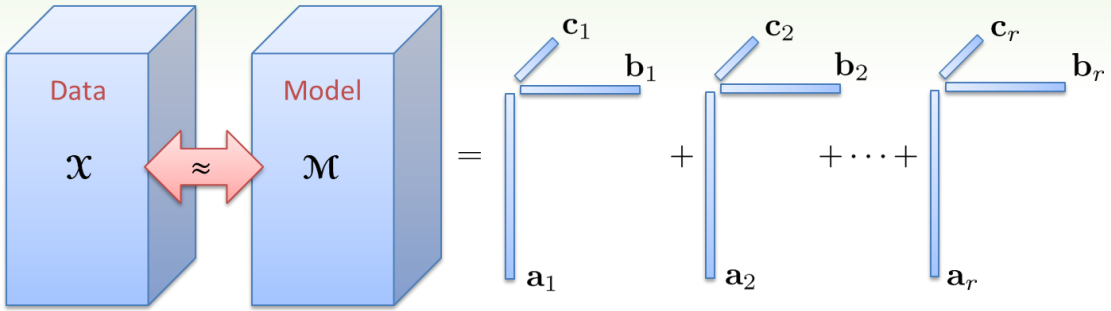
$$\left(\frac{m}{1+m}\right)^x \left(1 - \frac{m}{1+m}\right)^{(1-x)}$$

Equivalent to minimizing negative log likelihood:

$$F(\mathcal{X}, \mathcal{M}) = \sum_{i \in \mathcal{I}} \underbrace{\log(m_i + 1) - x_i \log m_i}_{f(x_i, m_i)}$$

Anderson-Bergman, Duersch, Hong, Kolda 2017

Generalized CP



$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{ijk\in\Omega} f(x_{ijk}, m_{ijk})$$

$$\text{s.t. } \mathcal{M} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

"Standard" CP uses:

$$f(x, m) = (x - m)^2$$

"Poisson" CP (Chi-Kolda 2012) uses:

$$f(x, m) = m - x \log m \quad (x \in \mathbb{N}, m \geq 0)$$

"Boolean-Odds" CP uses:

$$f(x, m) = \log(m + 1) - x \log m \quad (x \in \{0, 1\}, m \geq 0)$$

Apply favorite optimization method (including SGD) to compute the solution.

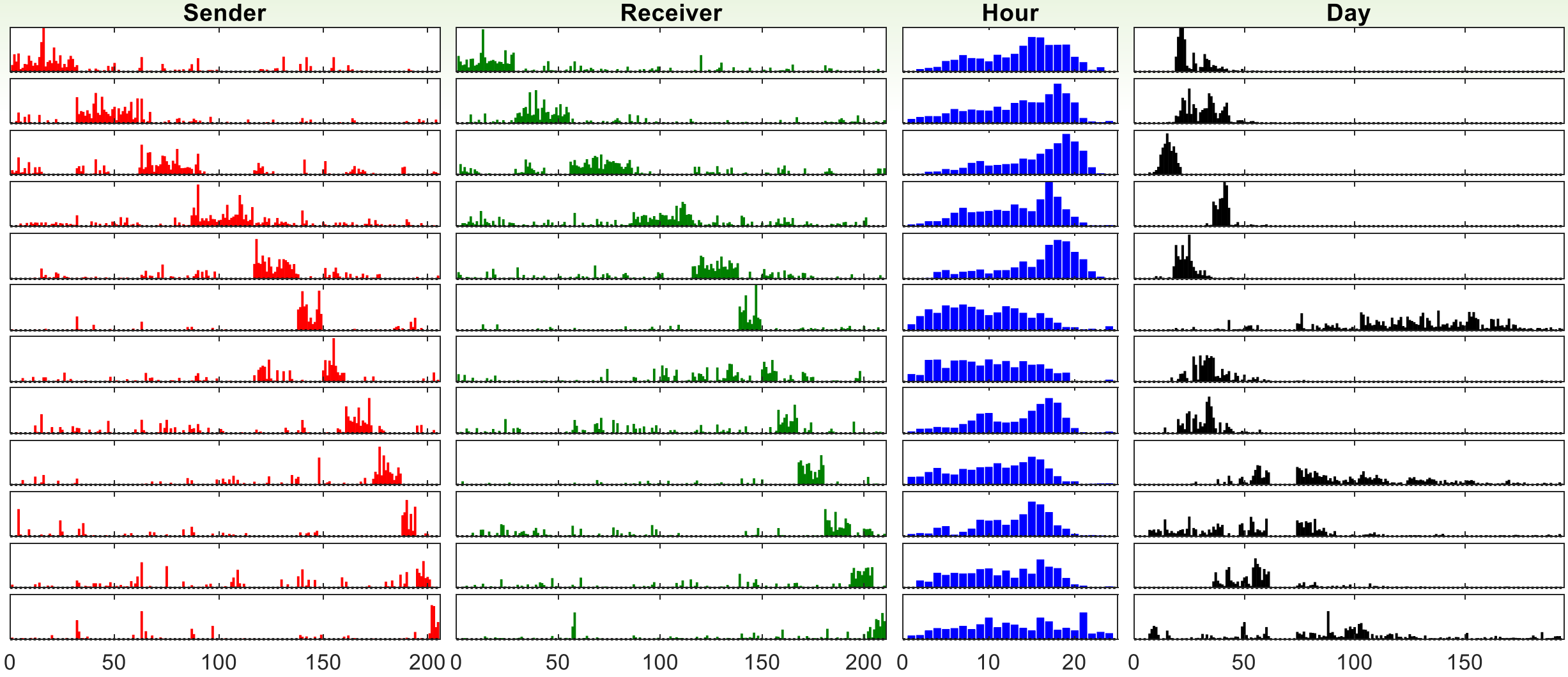
Anderson-Bergman, Duersch, Hong, Kolda 2017

A Sparse Dataset

- UC Irvine Chat Network
- 4-way binary tensor
  - Sender (205)
  - Receiver (210)
  - Hour of Day (24)
  - Day (194)
  - 14,953 nonzeros (very sparse)
- Goodness-of-fit (odds):
  $f(x, m) = \log(m + 1) - x \log m$
- Rank-12 decomposition

Opsahl, T., Panzarasa, P., 2009. Clustering in weighted networks. Social Networks 31 (2), 155-163, doi: 10.1016/j.socnet.2009.02.002

Binary Chat Data using Boolean CP



Anderson-Bergman, Duersch, Hong, Kolda 2017

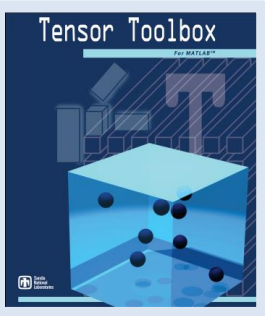
Acknowledgements

- Cliff Anderson-Bergman
- Grey Ballard
- Casey Battaglino
- Jed Duersch
- David Hong
- Alex Williams

Tensors & Data Analysis

- CP tensor decomposition is effective for unsupervised data analysis
  - Latent factor analysis
  - Dimension reduction
- CP can be generalized to alternative fit functions
  - Boolean data, count data, etc.
- Randomized techniques are open new doorways to larger datasets and more robust solutions
  - Matrix sketching
  - Stochastic gradient descent
- Other on-going & future work
  - Generalized CP (GCP) using SGD
  - Parallel CP and GCP implementations (https://gitlab.com/tensors/genten)
  - Parallel Tucker for compression (https://gitlab.com/tensors/TuckerMPI)
  - Randomized ST-HOSVD (Tucker)
  - Functional tensor factorization as surrogate for expensive functions
  - Extensions to many more applications (binary data, signals, etc.)

Kolda and Bader, Tensor Decompositions and Applications, *SIAM Review* '09



Tensor Toolbox for MATLAB: *www.tensortoolbox.org*
Bader, Kolda, Acar, Dunlavy, and others

Contact: Tammy Kolda, www.kolda.net, tgkolda@sandia.gov