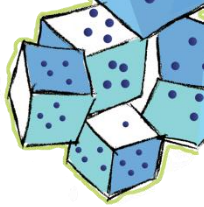


Illustration by Chris Brigman

Sandia National Laboratories

siam

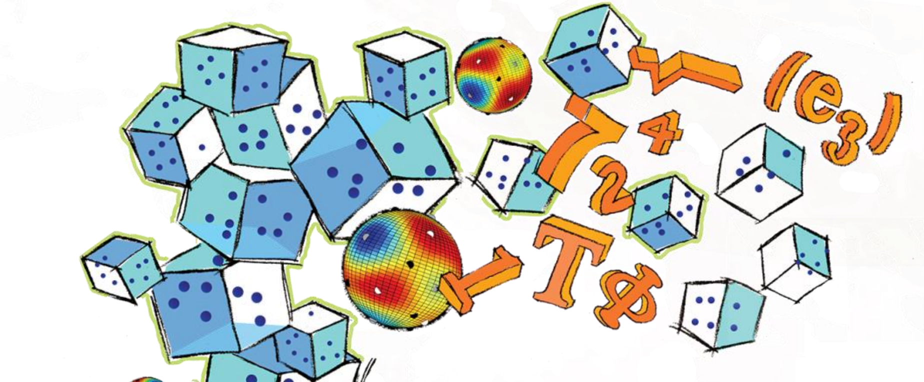
SIAM Invited Address,
JMM18,
San Diego, CA
Jan. 11, 2018

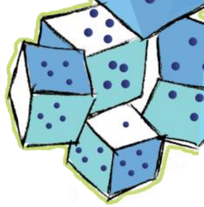
Tensor Decomposition: A Mathematical Tool for Data Analysis

Tamara G. Kolda
Sandia National Labs, Livermore, CA
www.kolda.net

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





Collaborators

Grey Ballard (Wake Forrest) &
Casey Battaglino (Georgia Tech)

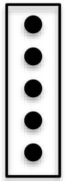
David Hong (Michigan) & Cliff Anderson-Bergman (Sandia)

Jed Duersch (Sandia)

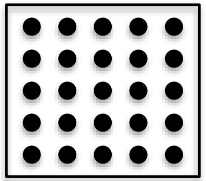
Alex Williams (Stanford)

A Tensor is an d-Way Array

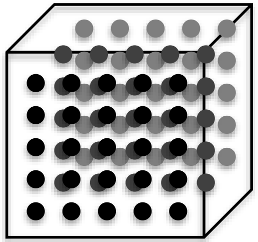
Vector
$d = 1$



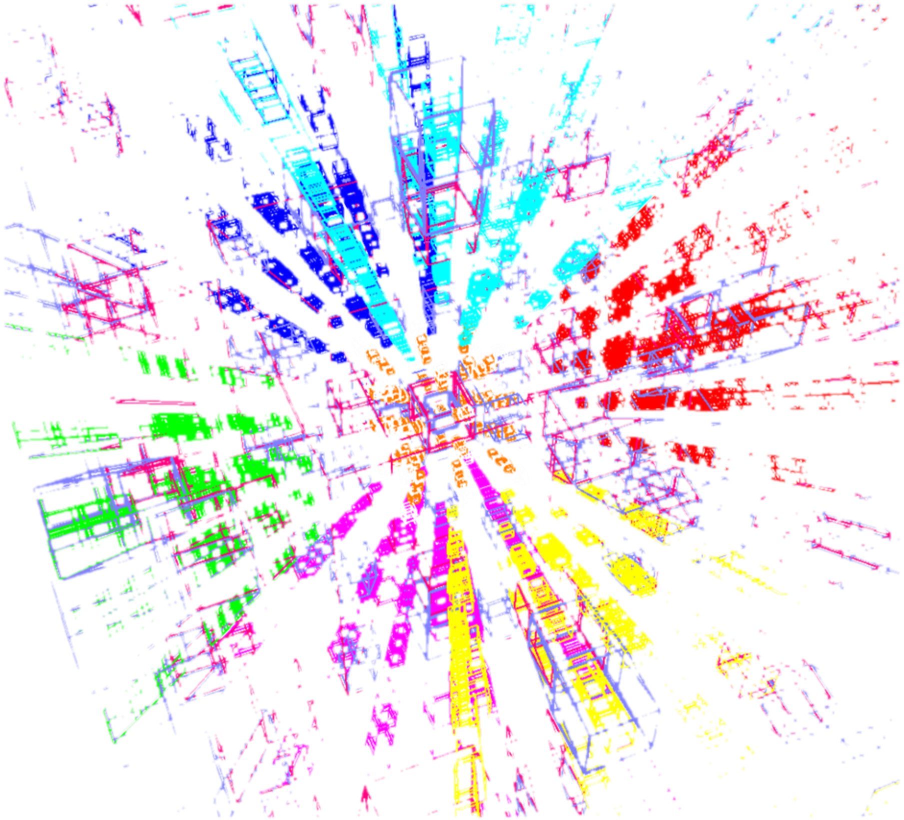
Matrix
$d = 2$



3rd-order Tensor
$d = 3$



d^{th}-order Tensor
$d > 3$

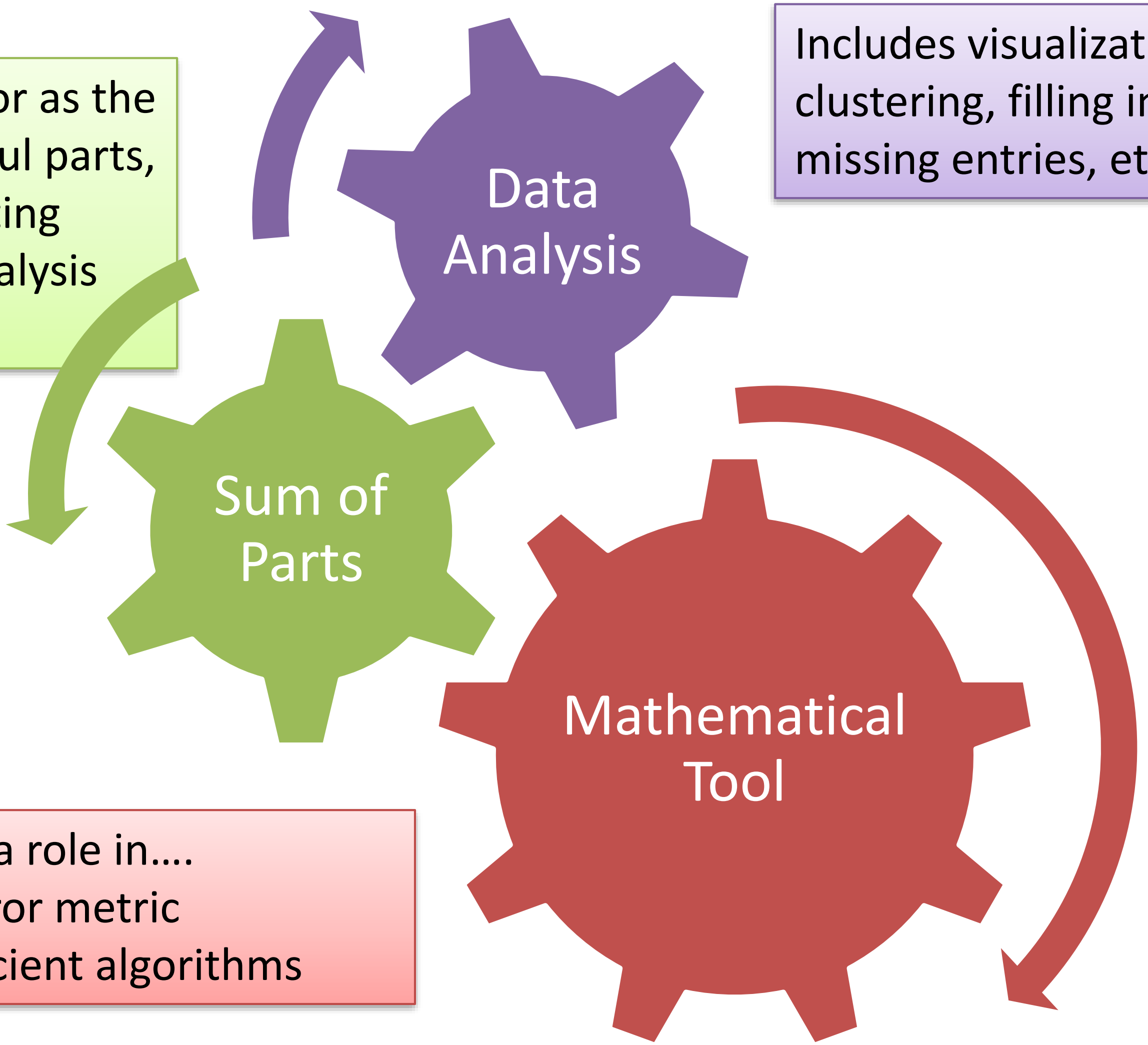


Tensor Decomposition: A Mathematical Tool for Data Analysis

Sandia National Laboratories

Express the tensor as the sum of meaningful parts, which is the starting point for data analysis activities

Includes visualization, clustering, filling in missing entries, etc.



Mathematics play a role in….
- Defining the error metric
- Developing efficient algorithms

Related Concepts for Matrices

- Principal component analysis (PCA)
- Singular value decomposition (SVD)
- Independent component analysis (ICA)
- Nonnegative matrix factorization (NMF)
- Sparse matrix factorization
- Matrix completion

Building Block for Decomposition: Rank-One Tensors = Vector Outer Products

Matrix Version (2-way)

Given **two vectors**:

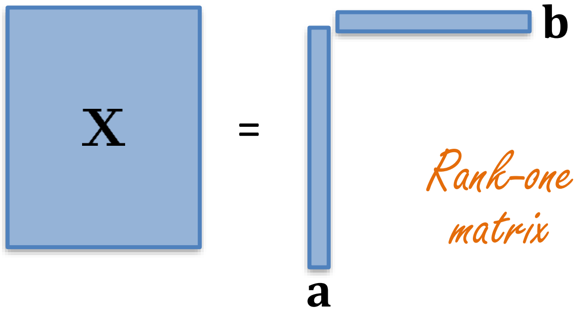
$$\mathbf{a} \in \mathbb{R}^m, \mathbf{b} \in \mathbb{R}^n$$

Their **outer product** is:

$$\mathbf{X} = \mathbf{a} \circ \mathbf{b} \quad \in \mathbb{R}^{m \times n}$$

Each entry is given by:

$$x(i,j) = a(i)\, b(j)$$



Rank-one matrix

Tensor Version (3-way)

Given **three vectors**:

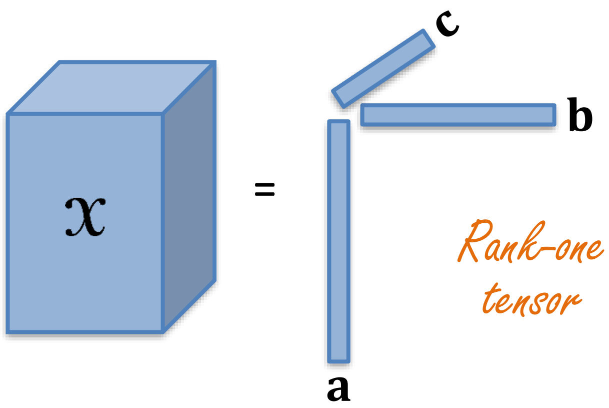
$$\mathbf{a} \in \mathbb{R}^m, \mathbf{b} \in \mathbb{R}^n, \mathbf{c} \in \mathbb{R}^p$$

Their **outer product** is:

$$\mathcal{X} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c} \quad \in \mathbb{R}^{m \times n \times p}$$

Each entry is given by:

$$x(i,j,k) = a(i)\, b(j)\, c(k)$$



Rank-one tensor

Tensor Version (d-way)

Given **d vectors**:

$$\mathbf{a}_k \in \mathbb{R}^{n_k}, \quad k = 1, \dots, d$$

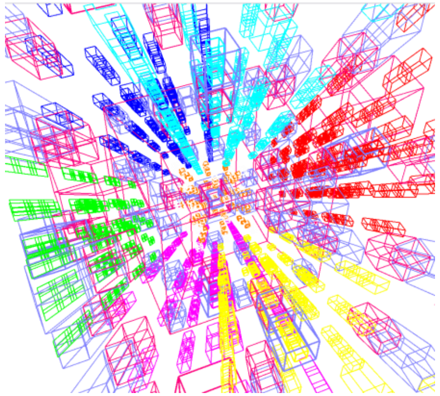
Their **outer product** is:

$$\mathcal{X} = \mathbf{a}_1 \circ \cdots \circ \mathbf{a}_d \quad \in \mathbb{R}^{n_1 \times \cdots \times n_d}$$

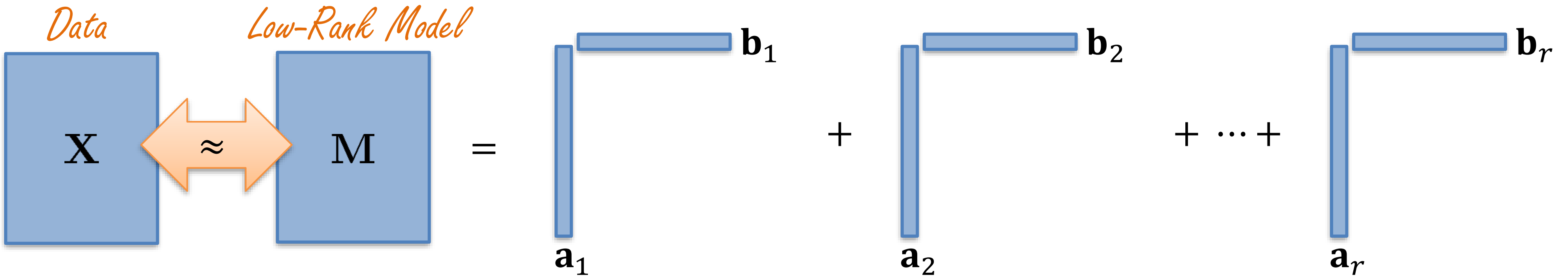
Each entry is given by:

$$x(i_1, \dots, i_d) = a_1(i_1) \cdots a_d(i_d)$$

Visualizing gets weird... But the math is still fine!



Matrix Decomposition: Detecting Low-Rank Structure



$$x(i,j) \approx m(i,j) = a(i,1)\,b(j,1) + a(i,2)\,b(j,2) + \cdots + a(i,r)\,b(j,r) = \sum_{\ell=1}^{r} a(i,\ell)\,b(j,\ell)$$

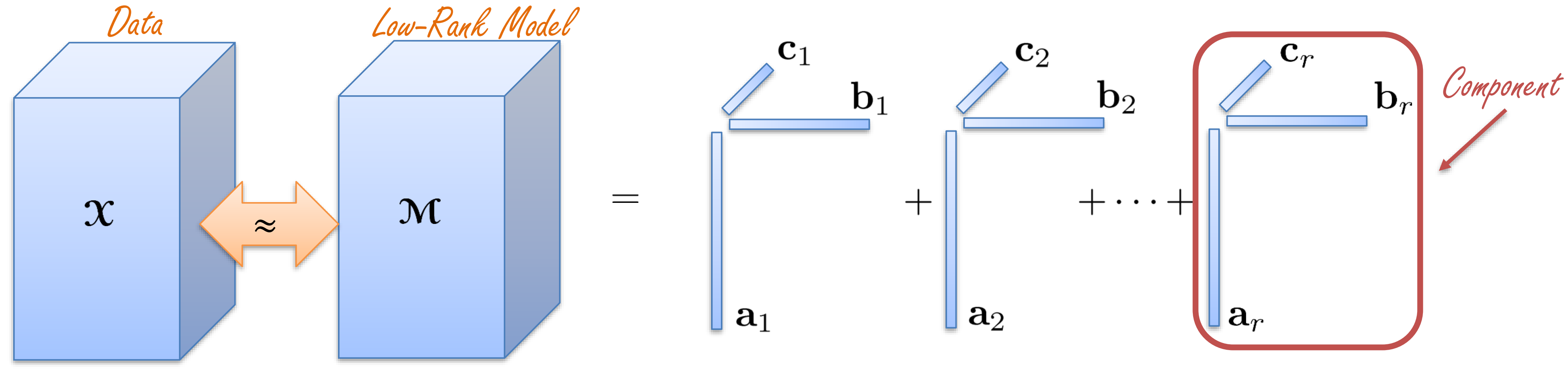
Matrix Notation $\Rightarrow \mathbf{X} \approx \mathbf{M} = \sum_{\ell=1}^{r} \mathbf{a}_\ell \circ \mathbf{b}_\ell = \mathbf{A}\mathbf{B}^T = [\![\mathbf{A},\mathbf{B}]\!]$

Sum of Squared Errors (SSE): $\sum_{ij} (\, x(i,j) - m(i,j)\,)^2 = \|\mathbf{X}-\mathbf{M}\|_F^2$

Eerily powerful tool for modeling data! Google search for “low-rank structure” turns up 5,590,000 results, and Google Scholar yields 127,000 papers!

CP Tensor Factorization (3-way): Detecting low-rank 3-way structure

Sandia National Laboratories



$$x(i,j,k) \quad \approx \quad m(i,j,k) = a(i,1)b(j,1)c(k,1) + a(i,2)b(j,2)c(k,2) + \cdots + a(i,r)b(j,r)c(k,r)$$

Tensor Notation $\Rightarrow$ $$\mathcal{X} \approx \mathcal{M} = \sum_{\ell=1}^{r} \mathbf{a}_\ell \circ \mathbf{b}_\ell \circ \mathbf{c}_\ell = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

Factor Matrices

Sum of Squared Errors (SSE): $$\sum_{ijk} (x(i,j,k) - m(i,j,k))^2 = \|\mathcal{X} - \mathcal{M}\|^2$$

Potentially an *even more* powerful tool for modeling data! But still new. Google search for “low-rank *tensor* structure” turns up only 550,000 results, and Google Scholar yields a mere 14,500 papers.

CP first invented in 1927

Frank Lauren Hitchcock
MIT Professor
(1875–1957)

THE EXPRESSION OF A TENSOR OR A POLYADIC AS A SUM OF PRODUCTS

By Frank L. Hitchcock

1. Addition and Multiplication.

Tensors are *added* by adding corresponding components. The *product* of a covariant tensor $A_{i_1 \cdot\cdot i_p}$ of order p *into* a covariant tensor $B_{i_{p+1} \cdot\cdot i_{p+q}}$ of order q is defined by writing

$$A_{i_1 \cdot\cdot i_p} B_{i_{p+1} \cdot\cdot i_{p+q}} = C_{i_1 \cdot\cdot i_{p+q}} \qquad (1)$$

where the product $C_{i_1 \cdot\cdot i_{p+q}}$ is a covariant tensor of order $p+q$. When no confusion results indices may be omitted giving

$$\mathbf{AB} = \mathbf{C} \qquad (1_a)$$

equivalent to the n^{p+q} equations (1). Boldface type is convenient for indicating that the letters do not denote merely numbers or scalars. Products of contravariant and of mixed tensors may be similarly defined.

A partial statement of the problem to be considered is as follows: to find under what conditions a given tensor can be expressed as a sum of products of assigned form. A more general statement of the problem will be given below.

2. Polyadic form of a tensor.

Any covariant tensor $A_{i_1 \cdot\cdot i_p}$ can be expressed as the sum of a finite number of tensors each of which is the product of p covariant vectors,

$$A_{i_1 \cdot\cdot i_p} = \sum_{j=1}^{j=h} a_{1j,\, i_1} a_{2j,\, i_2} \cdot\cdot a_{pj,\, i_p} \qquad (2)$$

where $a_{1j,\, i_1}$, etc., are a set of hp covariant vectors. When the indices $i_1 \cdot\cdot i_p$ can be omitted this may be written

$$\mathbf{A} = \sum_{j=1}^{j=h} \mathbf{a}_{1j} \mathbf{a}_{2j} \cdot\cdot \mathbf{a}_{pj}. \qquad (2_a)$$

The right member is now identical in appearance with a Gibbs

F. L. Hitchcock, ***The Expression of a Tensor or a Polyadic as a Sum of Products***, Journal of Mathematics and Physics, 1927

2. Polyadic form of a tensor.

Any covariant tensor $A_{i_1 \cdot\cdot i_p}$ can be expressed as the sum of a finite number of tensors each of which is the product of p covariant vectors,

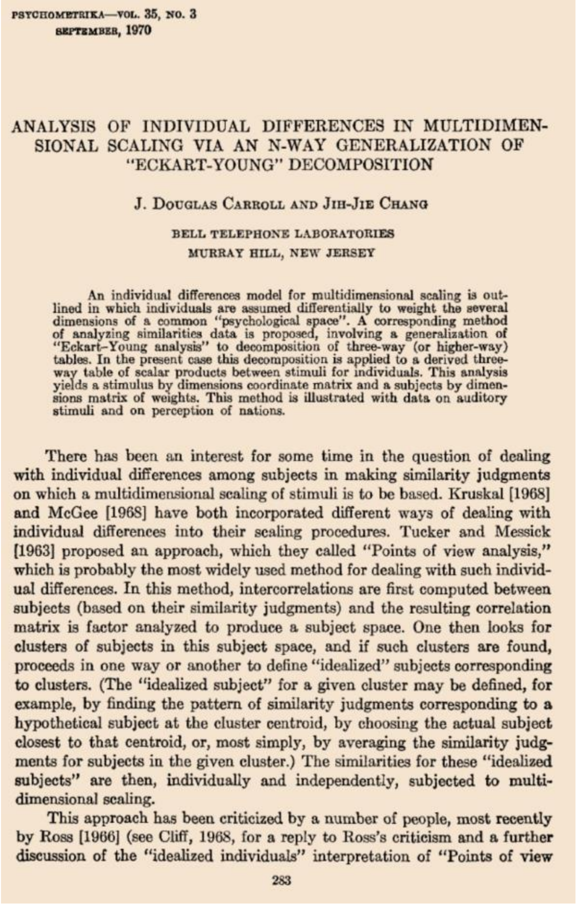
$$A_{i_1 \cdot\cdot i_p} = \sum_{j=1}^{j=h} a_{1j,\, i_1} a_{2j,\, i_2} \cdot\cdot a_{pj,\, i_p} \qquad (2)$$

where $a_{1j,\, i_1}$, etc., are a set of hp covariant vectors. When the indices $i_1 \cdot\cdot i_p$ can be omitted this may be written

$$\mathbf{A} = \sum_{j=1}^{j=h} \mathbf{a}_{1j} \mathbf{a}_{2j} \cdot\cdot \mathbf{a}_{pj}. \qquad (2_a)$$

CP Independently Reinvented (twice) in 1970

CANDECOMP: Canonical Decomposition


PSYCHOMETRIKA—VOL. 35, NO. 3
SEPTEMBER, 1970

ANALYSIS OF INDIVIDUAL DIFFERENCES IN MULTIDIMENSIONAL SCALING VIA AN N-WAY GENERALIZATION OF "ECKART-YOUNG" DECOMPOSITION

J. DOUGLAS CARROLL AND JIH-JIE CHANG

BELL TELEPHONE LABORATORIES
MURRAY HILL, NEW JERSEY

An individual differences model for multidimensional scaling is outlined in which individuals are assumed differentially to weight the several dimensions of a common "psychological space". A corresponding method of analyzing similarities data is proposed, involving a generalization of "Eckart–Young analysis" to decomposition of three-way (or higher-way) tables. In the present case this decomposition is applied to a derived three-way table of scalar products between stimuli for individuals. This analysis yields a stimulus by dimensions coordinate matrix and a subjects by dimensions matrix of weights. This method is illustrated with data on auditory stimuli and on perception of nations.

There has been an interest for some time in the question of dealing with individual differences among subjects in making similarity judgments on which a multidimensional scaling of stimuli is to be based. Kruskal [1968] and McGee [1968] have both incorporated different ways of dealing with individual differences into their scaling procedures. Tucker and Messick [1963] proposed an approach, which they called "Points of view analysis," which is probably the most widely used method for dealing with such individual differences. In this method, intercorrelations are first computed between subjects (based on their similarity judgments) and the resulting correlation matrix is factor analyzed to produce a subject space. One then looks for clusters of subjects in this subject space, and if such clusters are found, proceeds in one way or another to define "idealized" subjects corresponding to clusters. (The "idealized subject" for a given cluster may be defined, for example, by finding the pattern of similarity judgments corresponding to a hypothetical subject at the cluster centroid, by choosing the actual subject closest to that centroid, or, most simply, by averaging the similarity judgments for subjects in the given cluster.) The similarities for these "idealized subjects" are then, individually and independently, subjected to multidimensional scaling.

This approach has been criticized by a number of people, most recently by Ross [1966] (see Cliff, 1968, for a reply to Ross's criticism and a further discussion of the "idealized individuals" interpretation of "Points of view

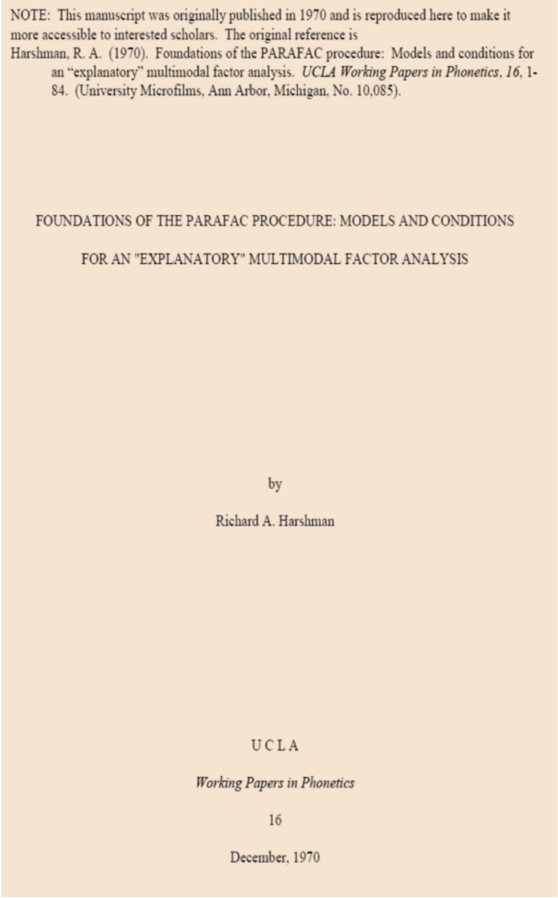
283

J. Douglas Carroll
Bell Labs
(1939-2011)

Jih-Jie Chang
Bell Labs
(1927-2007)

PARAFAC: Parallel Factors

Richard A. Harshman
Univ. Ontario
(1943-2008)


NOTE: This manuscript was originally published in 1970 and is reproduced here to make it more accessible to interested scholars. The original reference is
Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, *16*, 1-84. (University Microfilms, Ann Arbor, Michigan, No. 10,085).

FOUNDATIONS OF THE PARAFAC PROCEDURE: MODELS AND CONDITIONS FOR AN "EXPLANATORY" MULTIMODAL FACTOR ANALYSIS

by

Richard A. Harshman

UCLA

Working Papers in Phonetics

16

December, 1970

CP: CANDECOMP/PARAFAC — In 2000, Henk Kiers proposed this *compromise* name

CP: Canonical Polyadic — 2010: Pierre Comon, Lieven DeLathauwer, and others reverse-engineered CP, revising some of Hitchcock's terminology

Many thanks to the following persons for helping me learn about Jih-Jie Chang: Fan Chung, Ron Graham, Shen Lin (husband), May Chang (niece), Lili Bruer (daughter).

Example: CP for Mouse Neural Activity

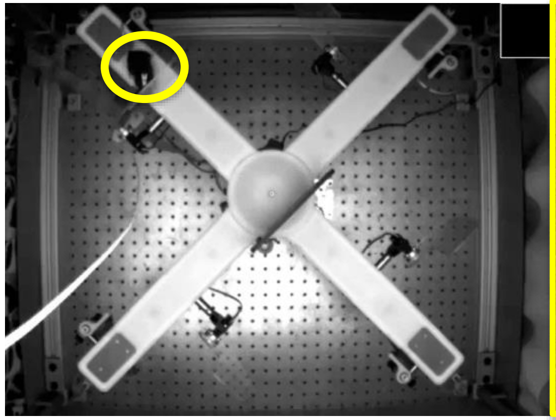
A. H. Williams, T. H. Kim, F. Wang, S. Vyas, S. I. Ryu, K. V. Shenoy, M. Schnitzer, T. G. Kolda, S. Ganguli. **Unsupervised Discovery of Demixed, Low-dimensional Neural Dynamics across Multiple Timescales through Tensor Components Analysis**. bioRxiv, 2017. https://doi.org/10.1101/211128

New Devices Enable Measuring Multiple Neurons Simultaneously

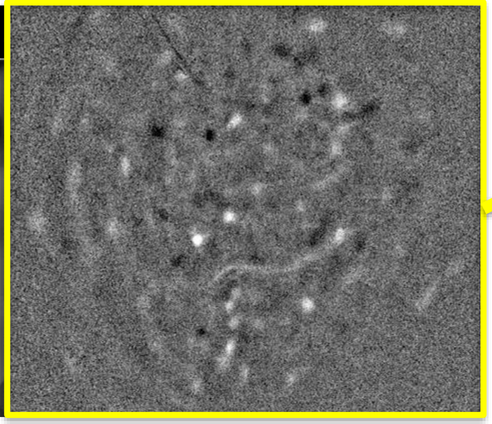
Thanks to Schnitzer Group @ Stanford
Mark Schnitzer, Fori Wang, Tony Kim

Microscope by Inscopix

mouse in “maze”



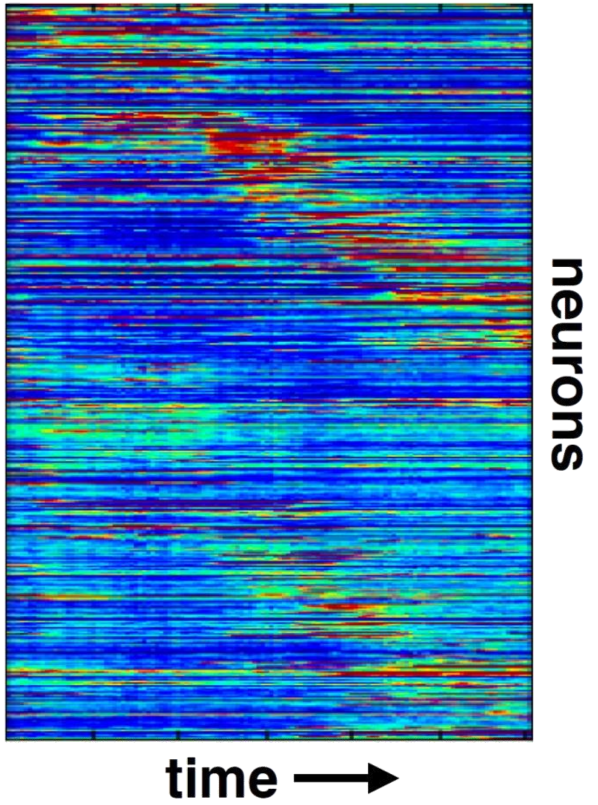
neural activity



One Column of Neuron x Time Matrix

One Trial

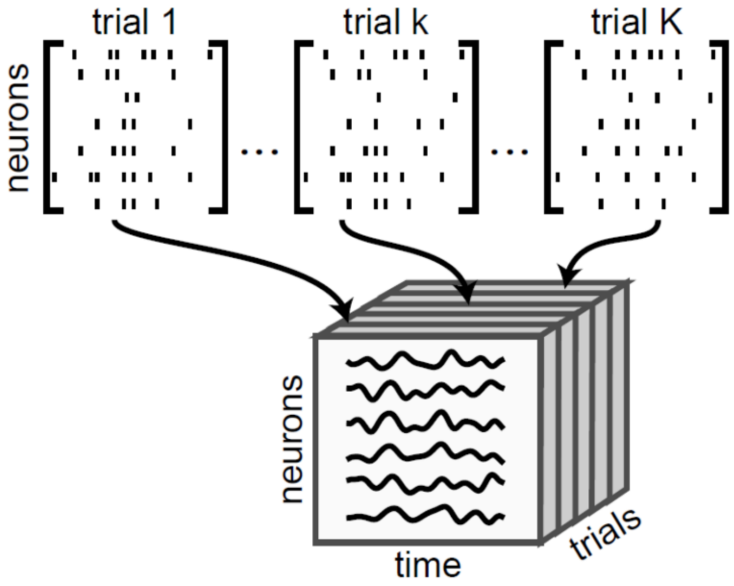
300 neurons × 120 time bins

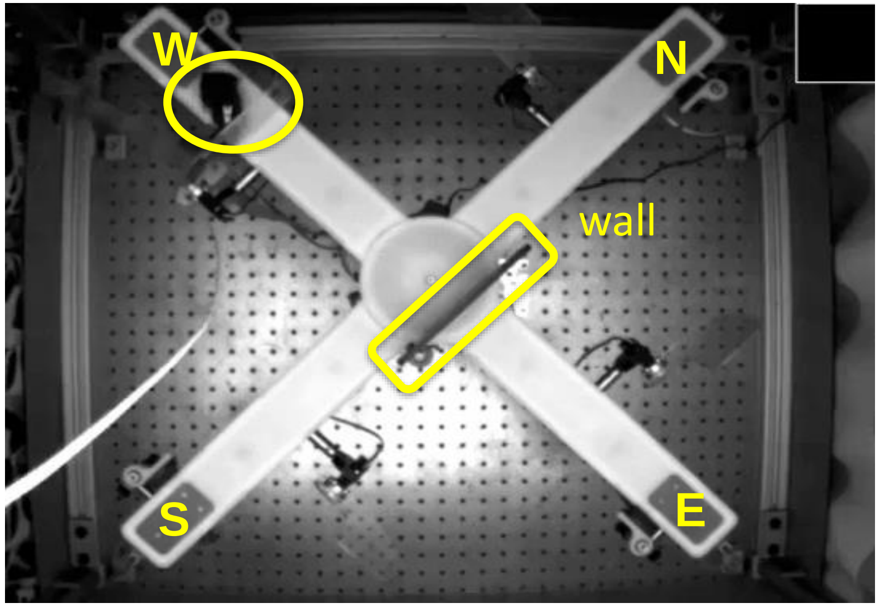


× 600 trials (over 5 days)

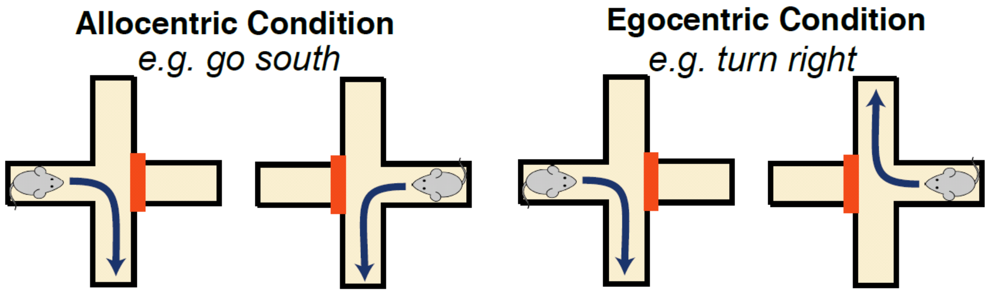
Williams et al., bioRxiv, 2017, DOI:10.1101/211128

Trials Vary Start Position and Strategies





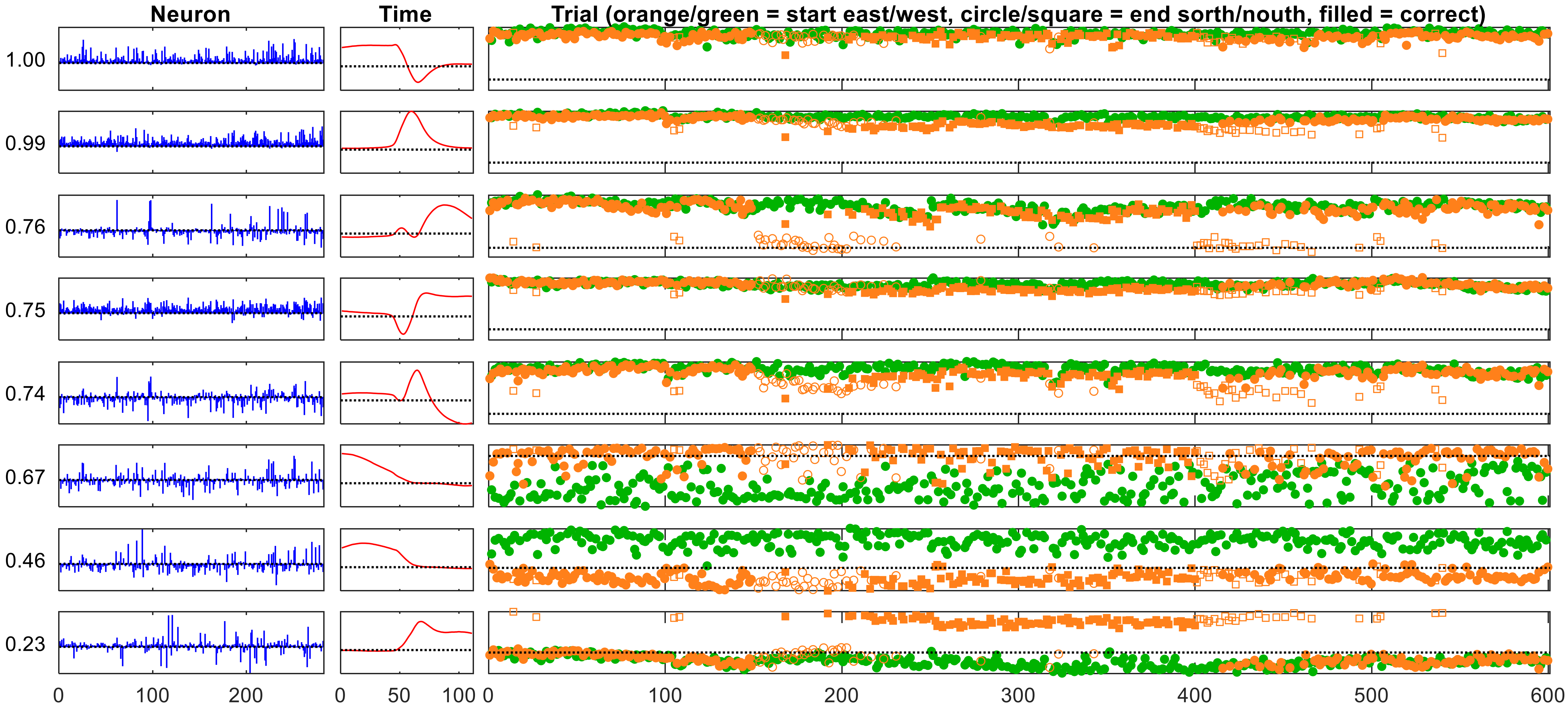
note different patterns on curtains



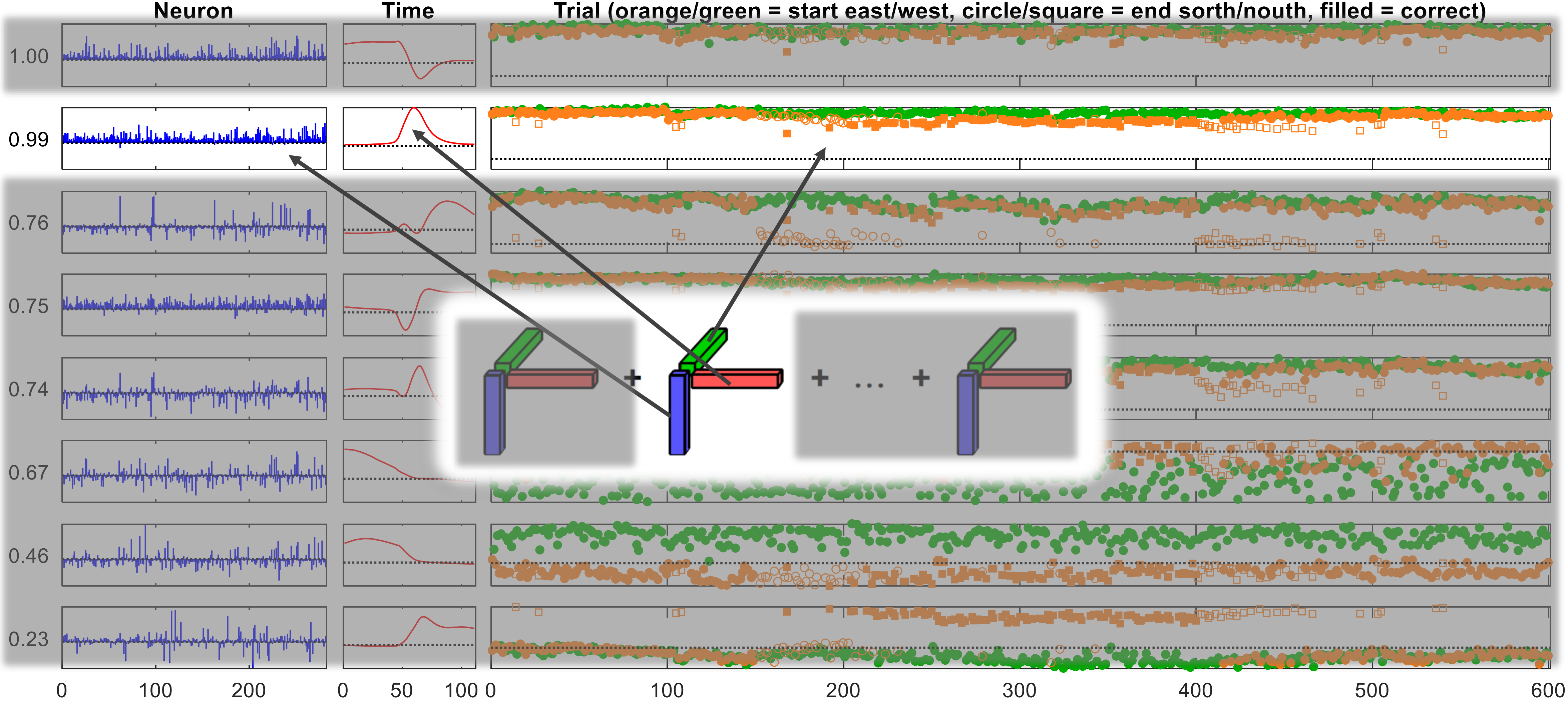
- 600 Trials over 5 Days
- Start West or East
- Conditions Swap Twice
  - ❖ Always Turn South
  - ❖ Always Turn Right
  - ❖ Always Turn South

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

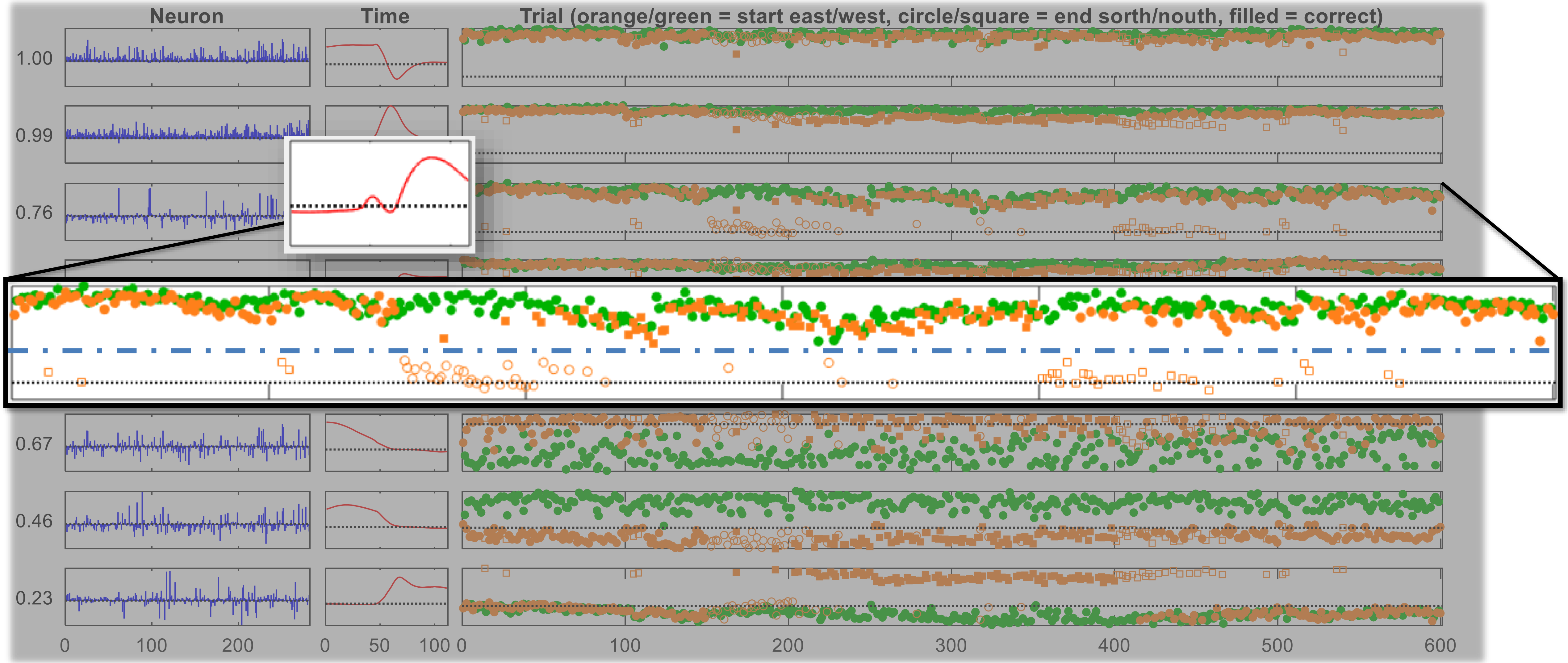
8-Component CP Decomposition of Mouse Neuron Data



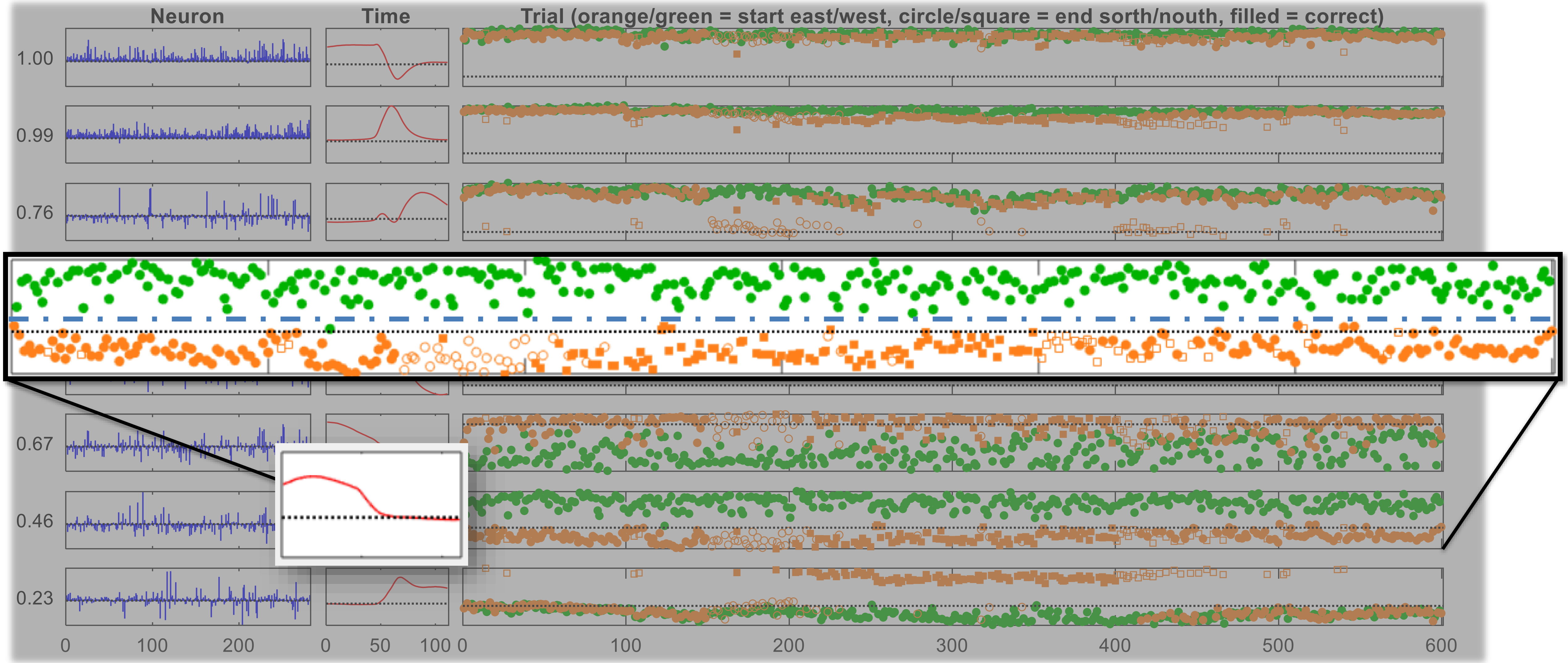
8-Component CP Decomposition of Mouse Neuron Data



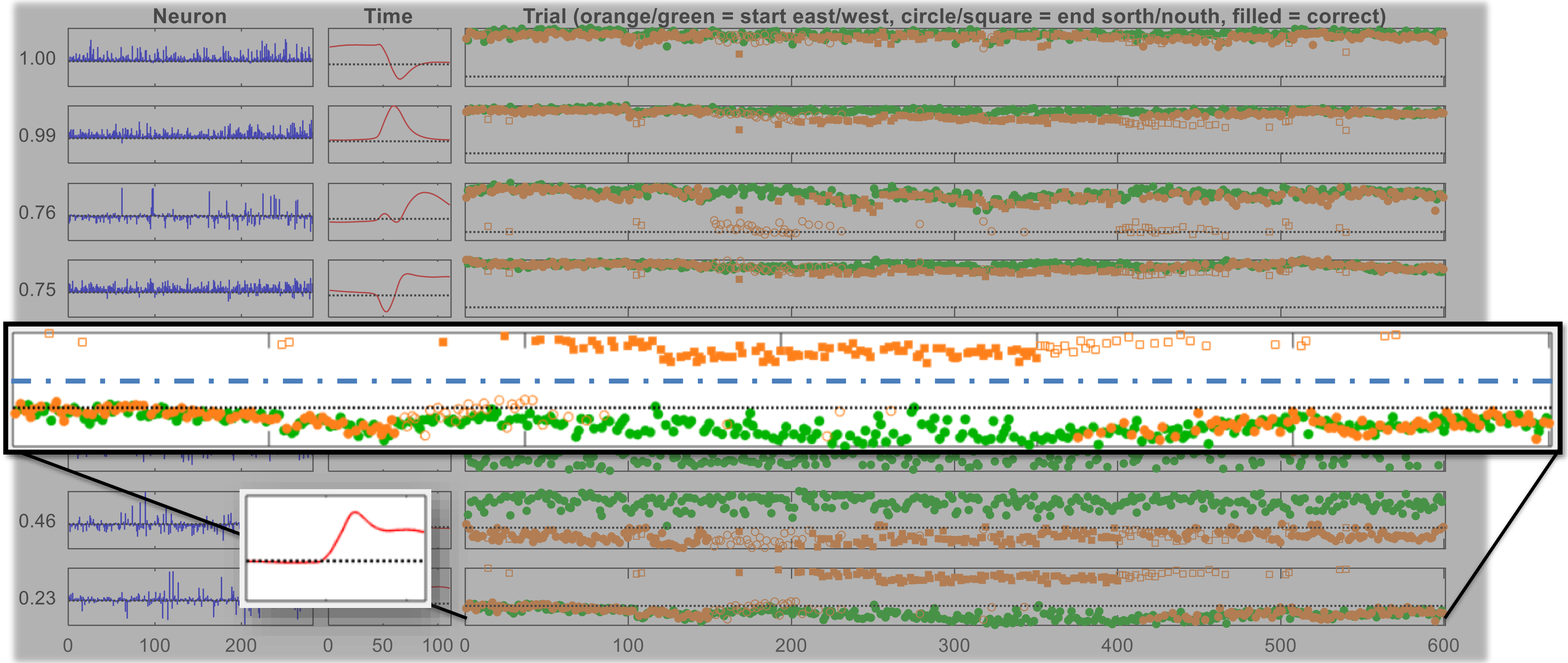
8-Component CP Decomposition of Mouse Neuron Data



8-Component CP Decomposition of Mouse Neuron Data



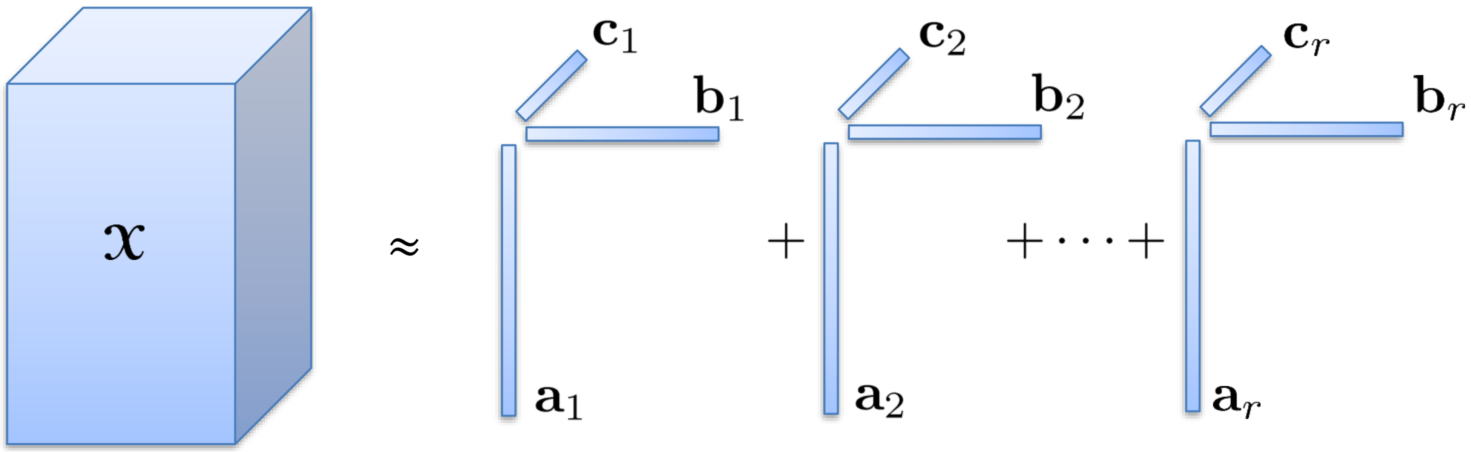
8-Component CP Decomposition of Mouse Neuron Data



Randomized Least Squares for CP Decomposition

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

Fitting CP

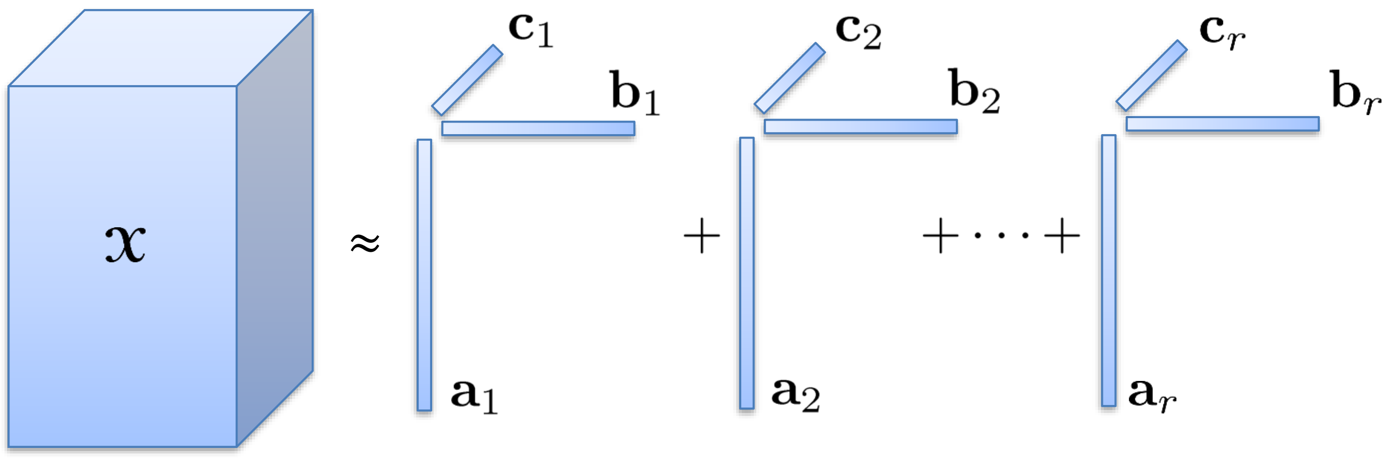


$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathcal{X} - \mathcal{M}\|^2 \text{ s.t. } \mathcal{M} = [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]$$

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{ijk} \left(x_{ijk} - \sum_{\ell} a_{i\ell} b_{j\ell} c_{k\ell} \right)^2$$

- **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)
- **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 (Kruskal 1977)

Fitting CP: Alternating Least Squares



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

$$\Downarrow$$

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{ijk} \left(x_{ijk} - \sum_{\ell} a_{i\ell} b_{j\ell} c_{k\ell} \right)^2$$

Repeat until convergence:

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

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

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

Nonconvex problem with convex subproblems.

Solving the Least Squares Problem

$$\min_{\mathbf{A}} \sum_{ijk} \Big(x_{ijk} - \sum_{\ell} a_{i\ell}\, b_{j\ell}\, c_{k\ell} \Big)^2 \quad \Rightarrow \quad \min_{\mathbf{A}} \| \mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})' \|_F^2$$

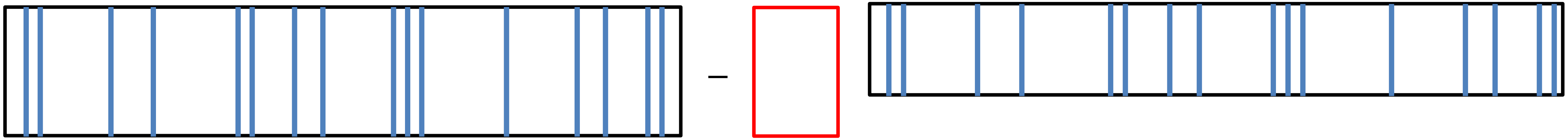
"right hand sides" — "matrix"

$$\mathbf{X}_{(1)} \quad - \quad \begin{bmatrix} \mathbf{a}_1 \cdots \mathbf{a}_r \end{bmatrix} \begin{bmatrix} (\mathbf{c}_1 \otimes \mathbf{b}_1)' \\ \vdots \\ (\mathbf{c}_r \otimes \mathbf{b}_r)' \end{bmatrix}$$

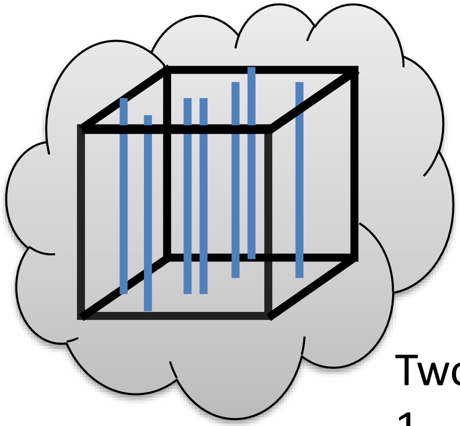
	Matrix Unfolding	$\mathbf{A}$	Khatri-Rao Product $(\mathbf{C} \odot \mathbf{B})'$
3-way case	$n \times n^2$	$n \times r$	$r \times n^2$
d-way case	$n \times n^{d-1}$	$n \times r$	$r \times n^{d-1}$

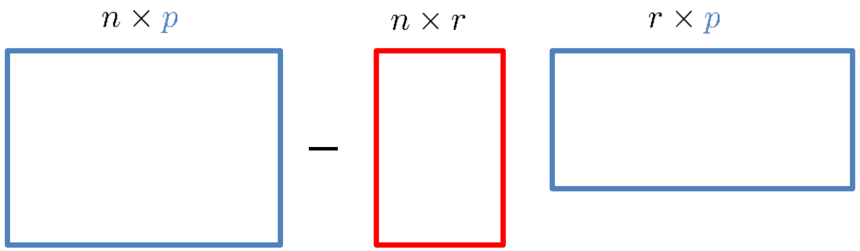
Short & Very Wide Matrix

CPRAND: Randomized Matrix Least Squares Subproblem



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





Each column in the sample is of the form: (C(ℓ,:) .* B(k,:))'

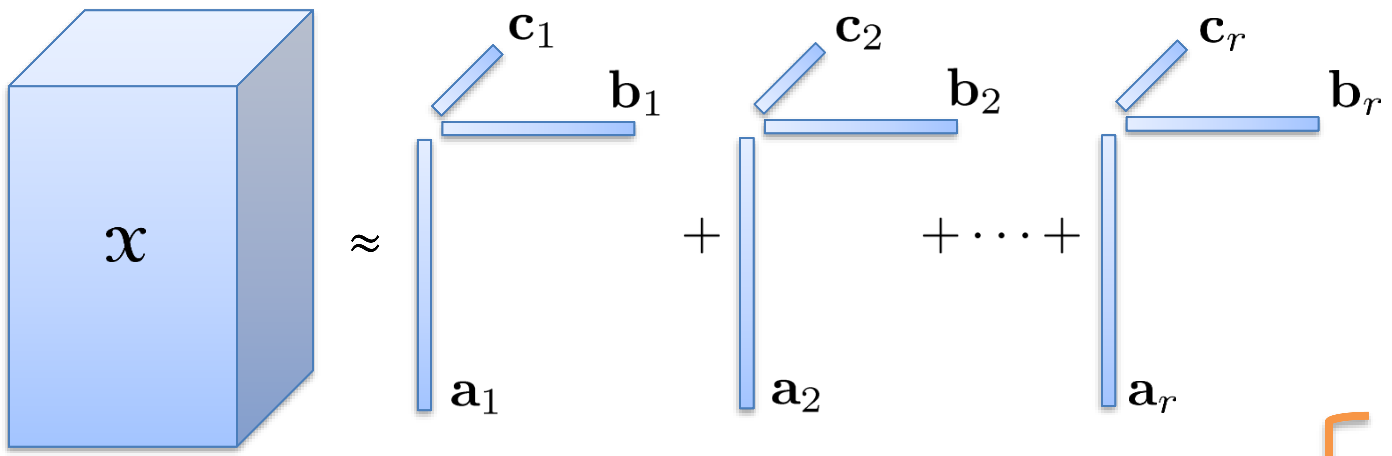
Two "tricks"

1. Never permute elements of tensor X intro $n \times n^2$ matrix form
2. Never form full Khatri-Rao product of size $r \times n^2$

CPRAND-MIX: Apply fast Johnson-Lindenstrauss Transform to mix the data in each direction to ensure "incoherence" – introduces some preprocessing cost

Battaglino, Ballard, Kolda, *A Practical Randomized CP Tensor Decomposition*, Jan 2017, arXiv:1701.06600

Convergence Check Become the Bottleneck!



Repeat until convergence:

Very fast with matrix sketching

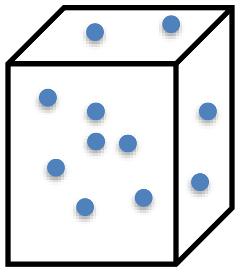
Step 1: $\min_{\mathbf{A}} \sum_{ijk} \left(x_{ijk} - \sum_{\ell} a_{i\ell}\, b_{j\ell}\, c_{k\ell} \right)^2$

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

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

Randomizing the Convergence Check

$$F(\mathbf{A},\mathbf{B},\mathbf{C}) = \sum_{ijk} \left(x_{ijk} - \sum_{\ell} a_{i\ell} b_{j\ell} c_{k\ell} \right)^2$$

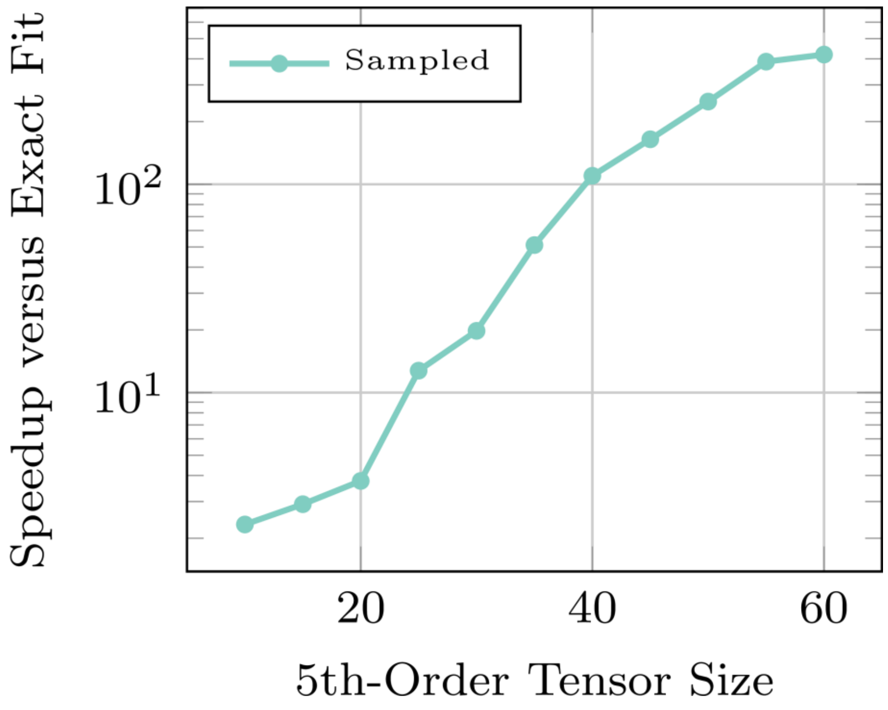


Estimate convergence of function values using small random subset of elements in function evaluation (use Chernoff-Hoeffding to bound accuracy)

$$\hat{F}(\mathbf{A},\mathbf{B},\mathbf{C}) = \omega \sum_{ijk \in \Omega} \left(x_{ijk} - \sum_{\ell} a_{i\ell} b_{j\ell} c_{k\ell} \right)^2$$

16000 samples < 1% of full data

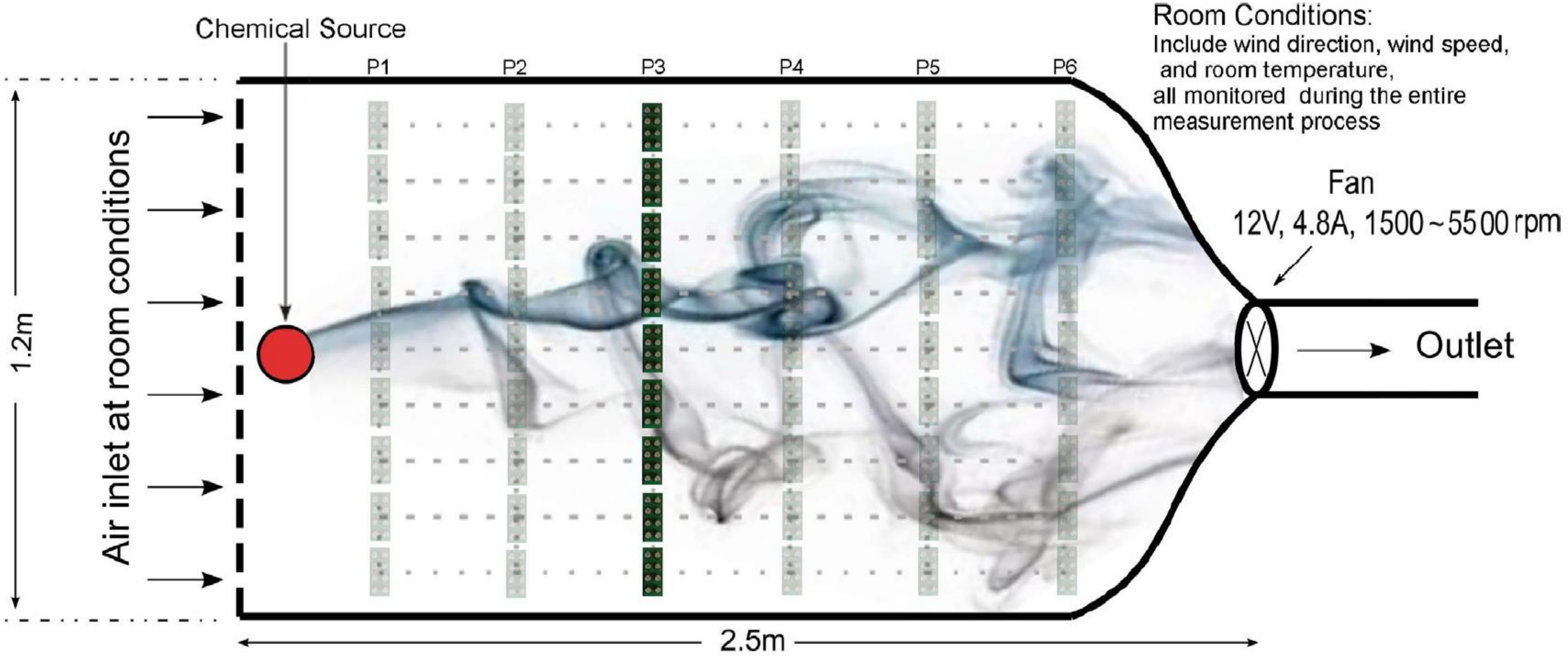
$$\frac{|F - \hat{F}|}{|F|} < 10^{-3}$$



Battaglino, Ballard, & Kolda 2017

Application to Hazardous Gas Dataset

71 Sensors × 5000 Timepoints × 5 Temperatures × 140 Experiments ≈ 2 GB



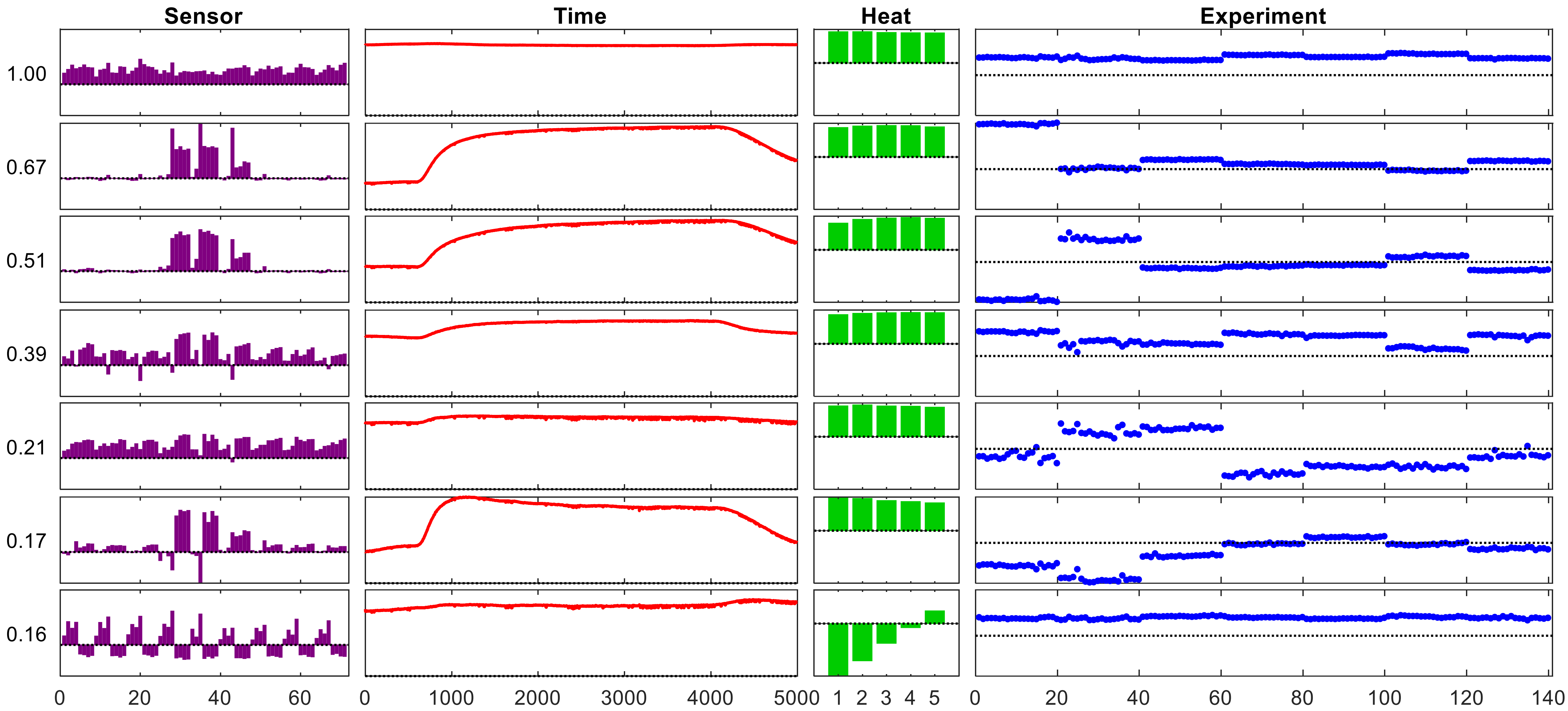
Position label	P 1	P 2	P 3	P 4	P 5	P 6
$x - axis\, distance\ (\underline{m})$	0.25	0.5	0.98	1.18	1.40	1.45

CP-ALS:
65 seconds

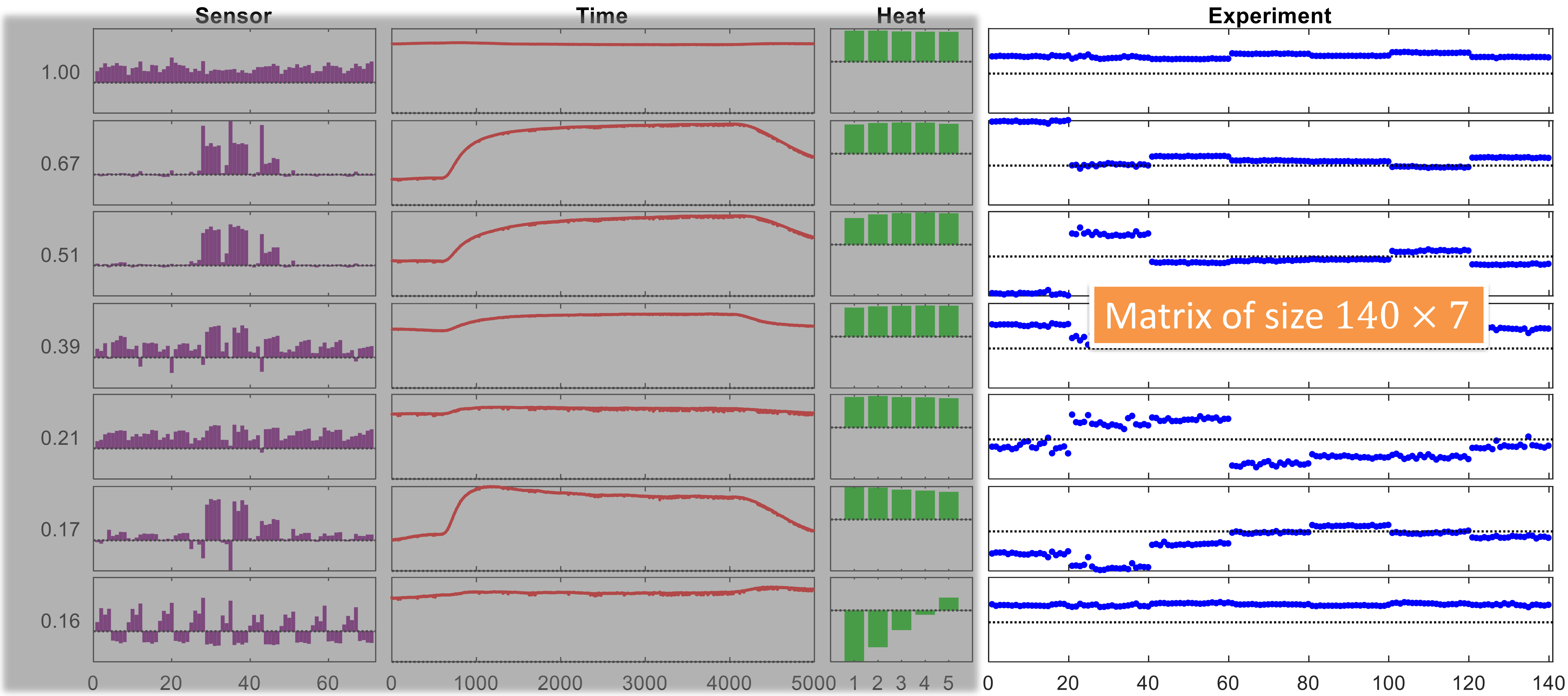
CP-ALS-RAND:
27 seconds

A. Vergara, J. Fonollosa, J. Mahiques, M. Trincavelli, N. Rulkov and R. Huerta, *On the performance of gas sensor arrays in open sampling systems using Inhibitory Support Vector Machines*, Sensors and Actuators B: Chemical, 2013, doi:10.1016/j.snb.2013.05.027

Factors from Gas Dataset

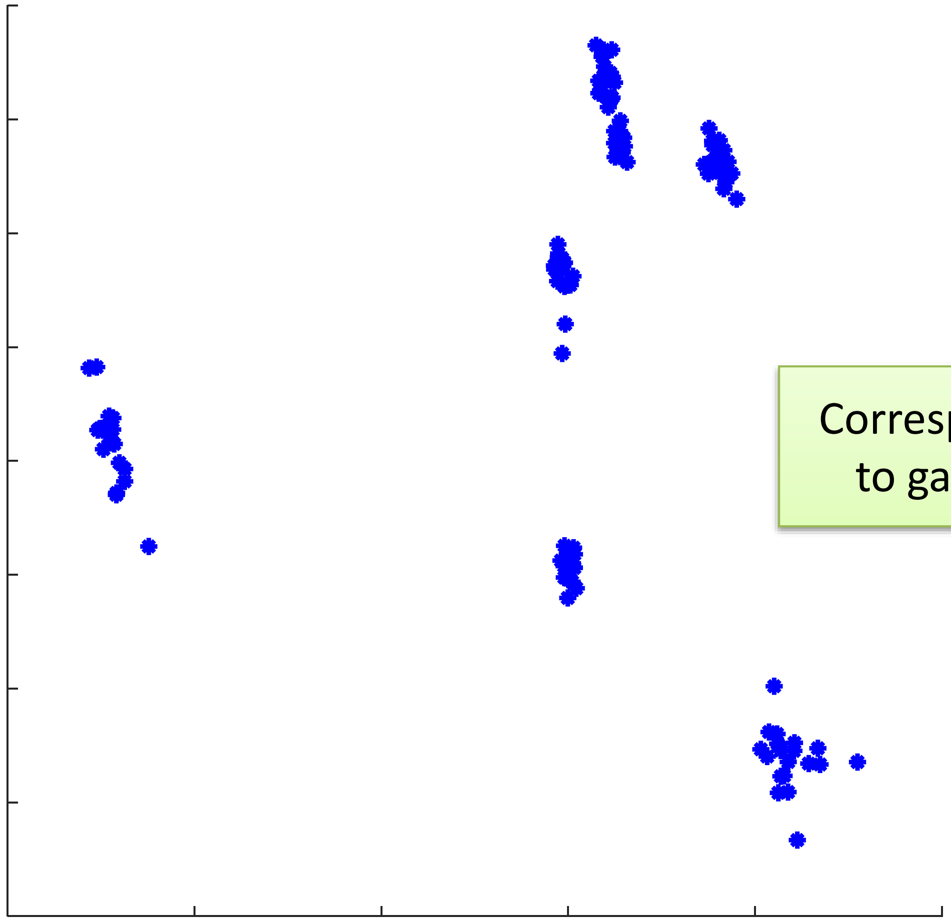


Factors from Gas Dataset

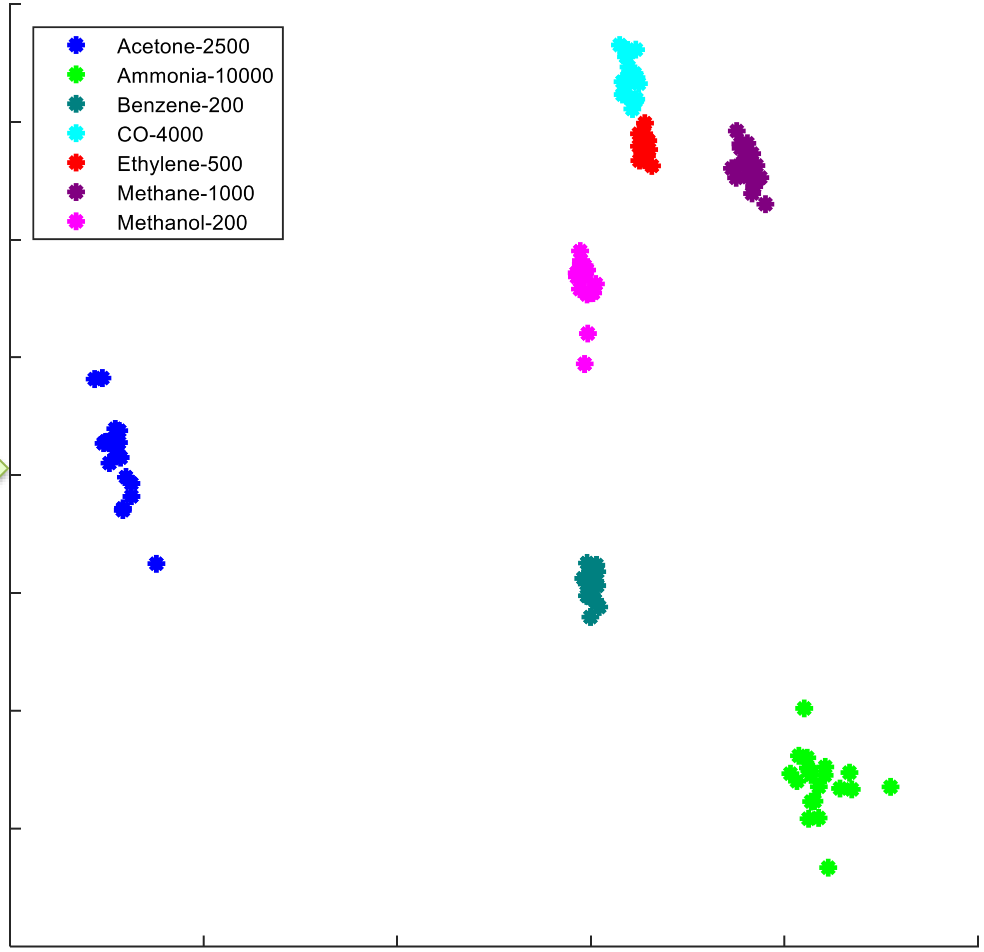


Viz of Experiment Factor Matrix Using PCA Projection

Experiments (140) projected onto 2D space using PCA



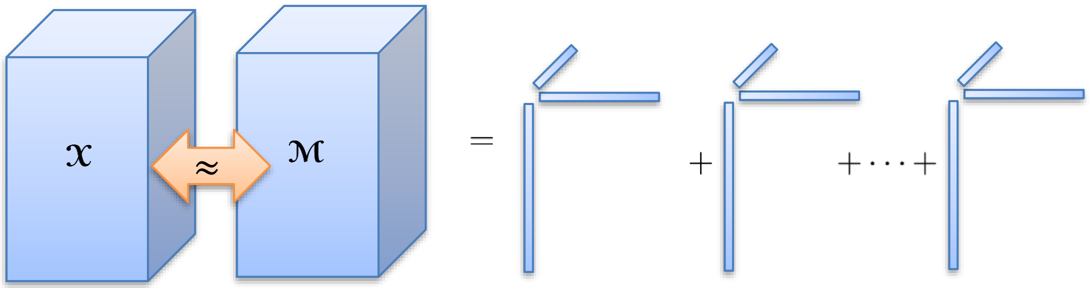
Correspondence to gas types!



Generalized CP Decomposition

Cliff Anderson-Bergman, J. Duersch, D. Hong, T. G. Kolda, **Generalized Canonical Polyadic Tensor Decomposition**, 2018 (coming soon)

Generalizing the Goodness-of-Fit Criteria



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

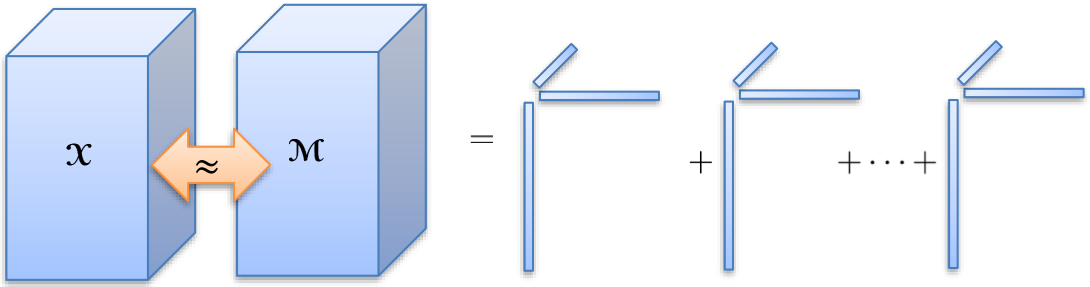
$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{ijk} (x_{ijk} - m_{ijk})^2 \text{ s.t. } \mathcal{M} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

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

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

Anderson-Bergman, Duersch, Hong, Kolda 2017

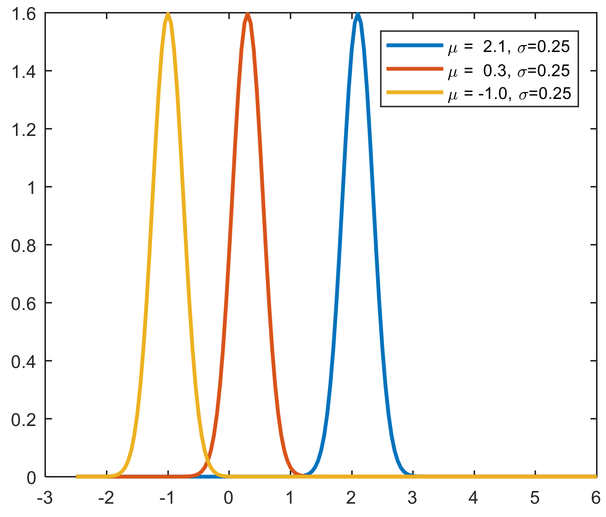
“Standard” CP via Maximum Likelihood



Gaussian Probability Density Function (PDF)

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

Probability Distribution Function: Normal-distributed with constant σ



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

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

Equivalently, Gaussian with mean m_{ijk}

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

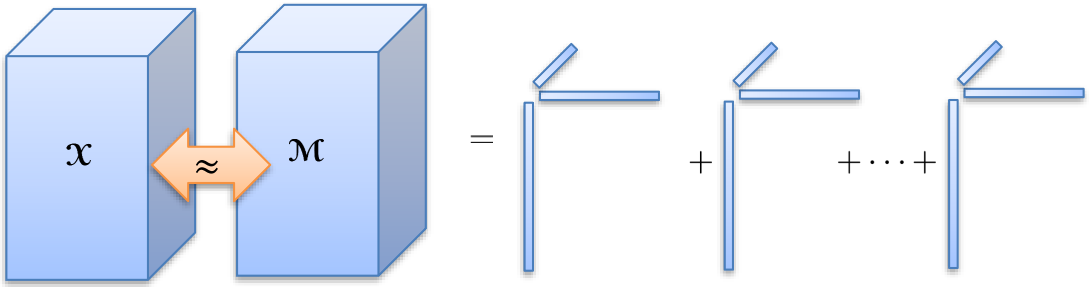
Minimize negative log likelihood with $\mu_{ijk} = m_{ijk}$ and σ constant for all entries:

$$-\log(\mathcal{L}(\mathcal{M})) = \sum_{ijk} \frac{(x_{ijk}-m_{ijk})^2}{\cancel{2\sigma^2}} + \cancel{{}^1\!/_2 \log 2\pi\sigma^2}$$

$$\min F(\mathcal{M}) = \sum_{ijk} (x_{ijk}-m_{ijk})^2$$

Anderson-Bergman, Duersch, Hong, Kolda 2017

“Rayleigh CP” with Linear Link



What if the data is nonnegative ($x_{ijk} \geq 0$)?

Assume data is Rayleigh-distributed.

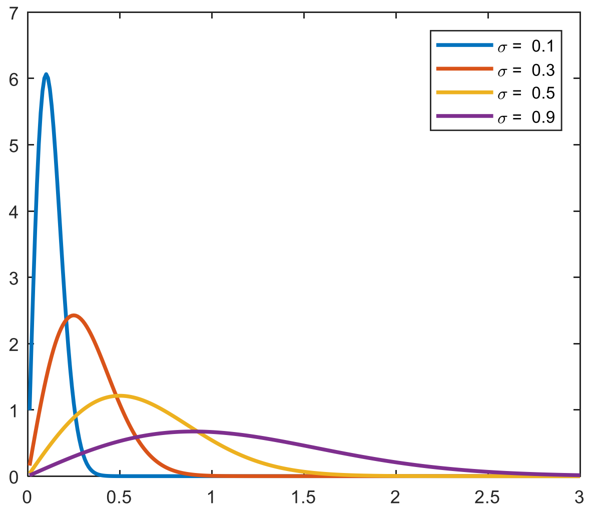
$$x_{ijk} \sim \text{Rayleigh}(m_{ijk})$$

Requires $m_{ijk} \geq 0$

Rayleigh Probability Density Function (PDF)

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

Probability Distribution Function: Rayleigh-distributed



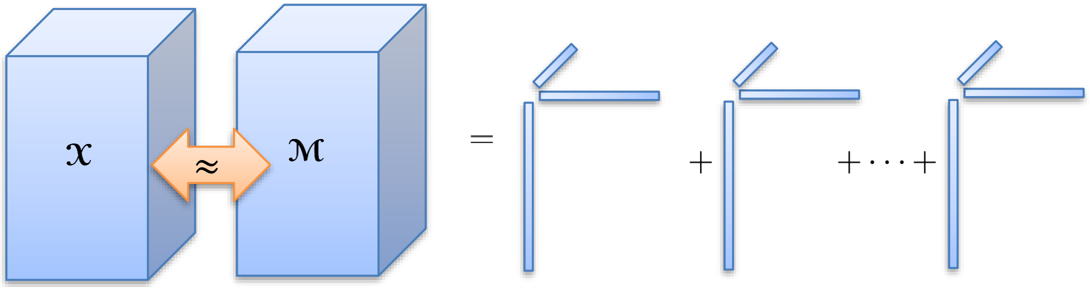
Minimize negative log likelihood with $\sigma_{ijk} = m_{ijk}$:

$$-\log(\mathcal{L}(\mathcal{M})) = \sum_{ijk} -\cancel{\log x_{ijk}} + 2\log m_{ijk} + \frac{x_{ijk}^2}{2m_{ijk}^2}$$

$$\min F(\mathcal{M}) = \sum_{ijk} 2\log m_{ijk} + \frac{x_{ijk}^2}{2m_{ijk}^2}$$

$$\mathbb{E}(x_{ijk}) = m_{ijk}\sqrt{\frac{\pi}{2}}$$

"Boolean CP" with Odds Link



What if data is binary ($x_{ijk} \in \{0,1\}$)?

m_{ijk}= odds ratio of $x_{ijk} = 1$.

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

$$\mathbb{E}(x_{ijk}) = \frac{m_{ijk}}{1+m_{ijk}}$$

Requires $m_{ijk} \geq 0$

Random Coin Flip: Probability versus Odds

$p \in [0,1]$: probability of 1

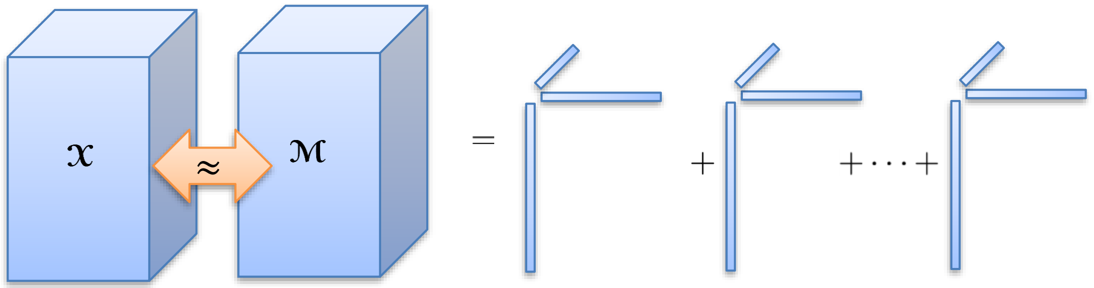
$r \geq 0$: odds ratio of 1

$$r = \frac{p}{1-p} \Leftrightarrow p = \frac{r}{1+r}$$

Probability Mass Distribution (PMF) $p^x(1-p)^{1-x} \Leftrightarrow \left(\frac{r}{1+r}\right)^x \left(\frac{1}{1+r}\right)^{1-x}$

$$\min F(\mathcal{M}) = \sum_{ijk} \log(m_{ijk}+1) - x_{ijk} \log m_{ijk}$$

Generalized CP



$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \sum_{ijk\in\Omega} f(x_{ijk}, m_{ijk}) \quad \text{s.t. } \mathcal{M} = [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]$$

Standard ($x, m \in \mathbb{R}$): $f(x,m) = (x-m)^2$

Rayleigh ($x, m \in \mathbb{R}_+$): $f(x,m) = 2\log(m) + x^2/(2m^2)$

Boolean-Odds ($x \in [0,1], m \in \mathbb{R}_+$): $f(x,m) = \log(m+1) - x\log(m)$

Poisson ($x \in \mathbb{N}, m \in \mathbb{R}_+$): $f(x,m) = m - x\log(m)$

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

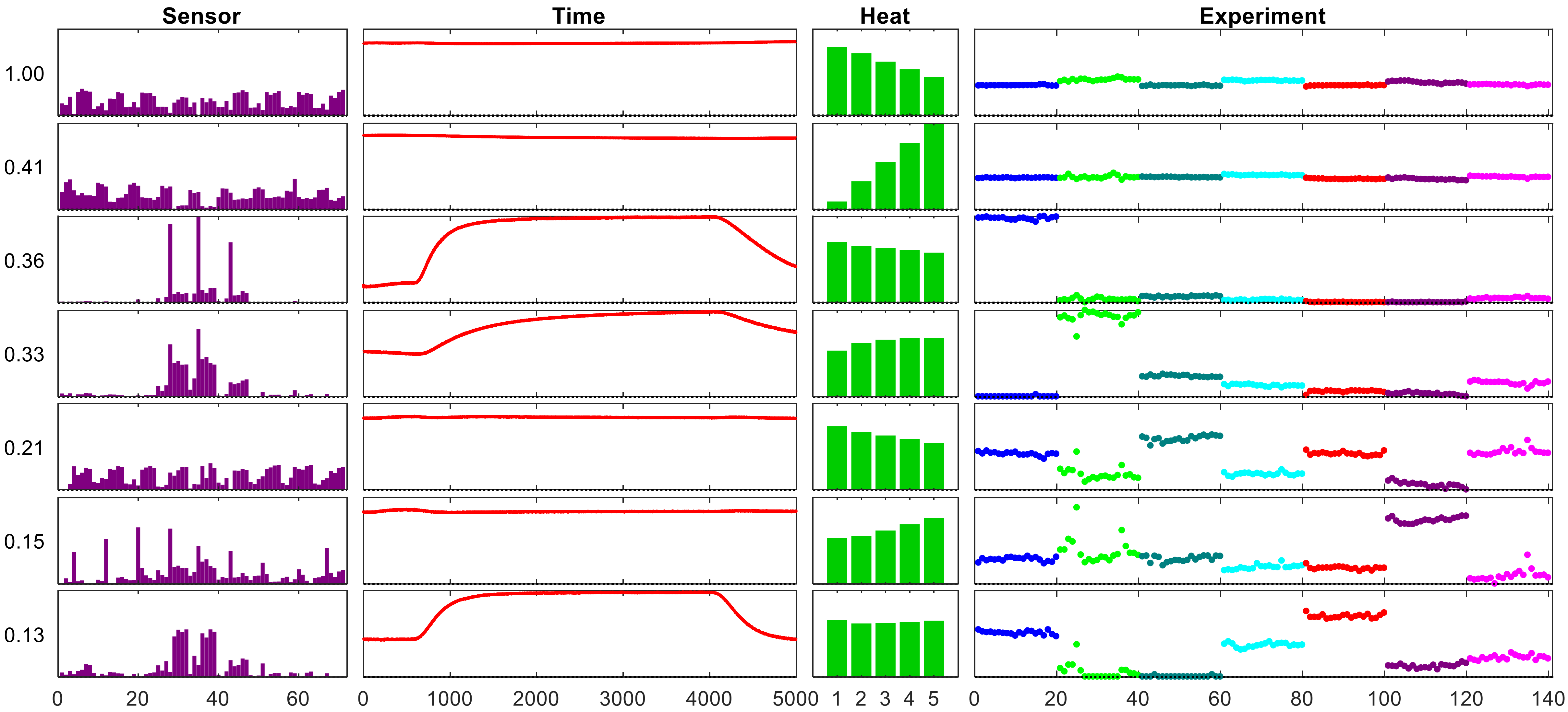
Algorithm Notes

- Can be solved via alternating or all-at-once optimization
  - ✓ Fewer knobs to tweak for all-at-once
  - ✓ Prefer all-at-once if any data is missing
- Gradient has an elegant form
  - ✓ Involves "MTTKRP"
- Missing data is handled by omitting from the sum in the objective function
  - ✓ Introduces sparsity into the gradient computation
- Large-scale problems requires stochastic approach
  - ✓ Stratification needed for sparse problems

Chi & Kolda 2012; Anderson-Bergman, Duersch, Hong, Kolda 2017

Mouse Data using Rayleigh (Nonneg)

Gas Data Using Rayleigh



A Sparse Binary Dataset

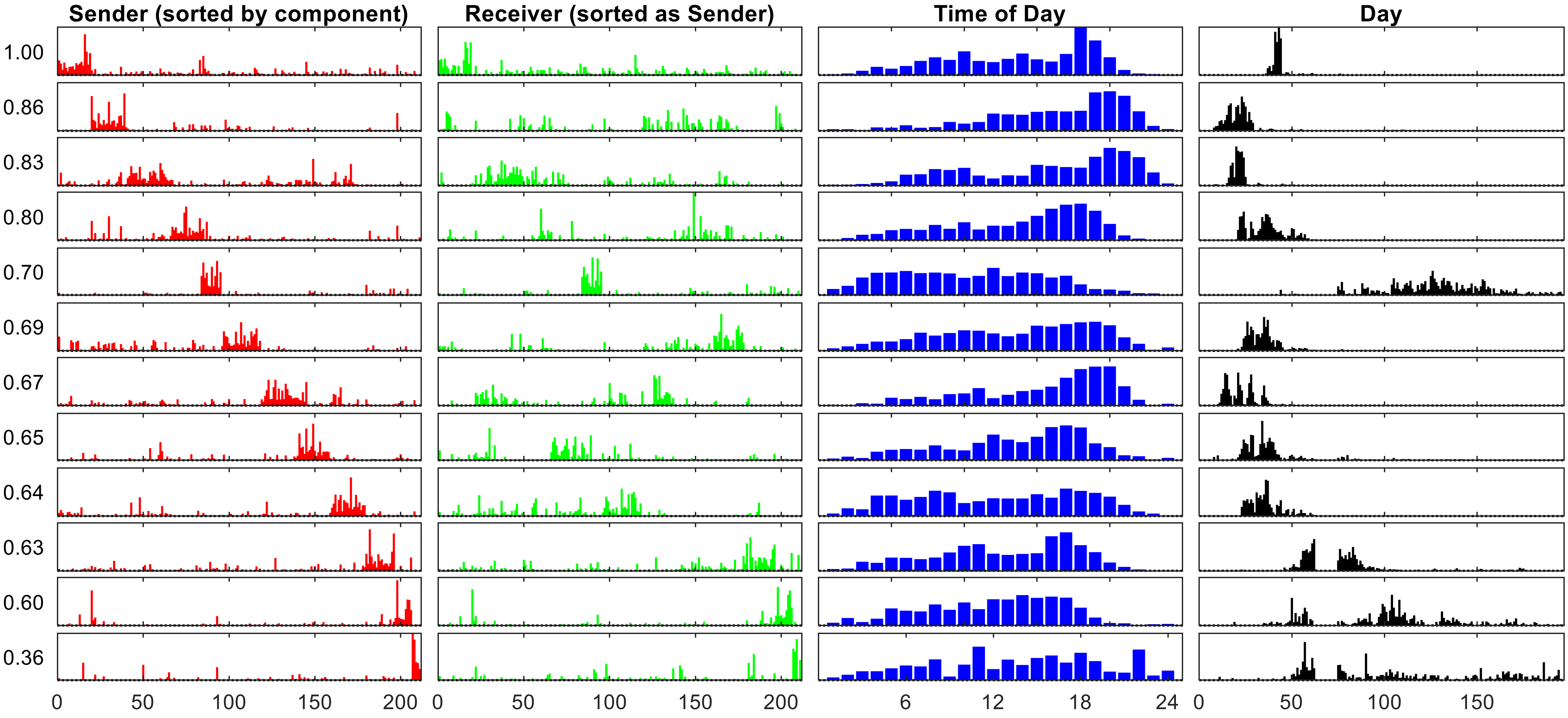
- UC Irvine Chat Network
  - 4-way binary tensor
    - Sender (211)
    - Receiver (211)
    - Hour of Day (24)
    - Day (196)
  - 14,849 nonzeros (very sparse)
- Goodness-of-fit (Boolean-odds):

$$f(x, m) = \log(m + 1) - x \log m$$

- Use GCP to compute 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



SIAM Journal on Mathematics of Data Science

- New journal, launching in Spring 2018

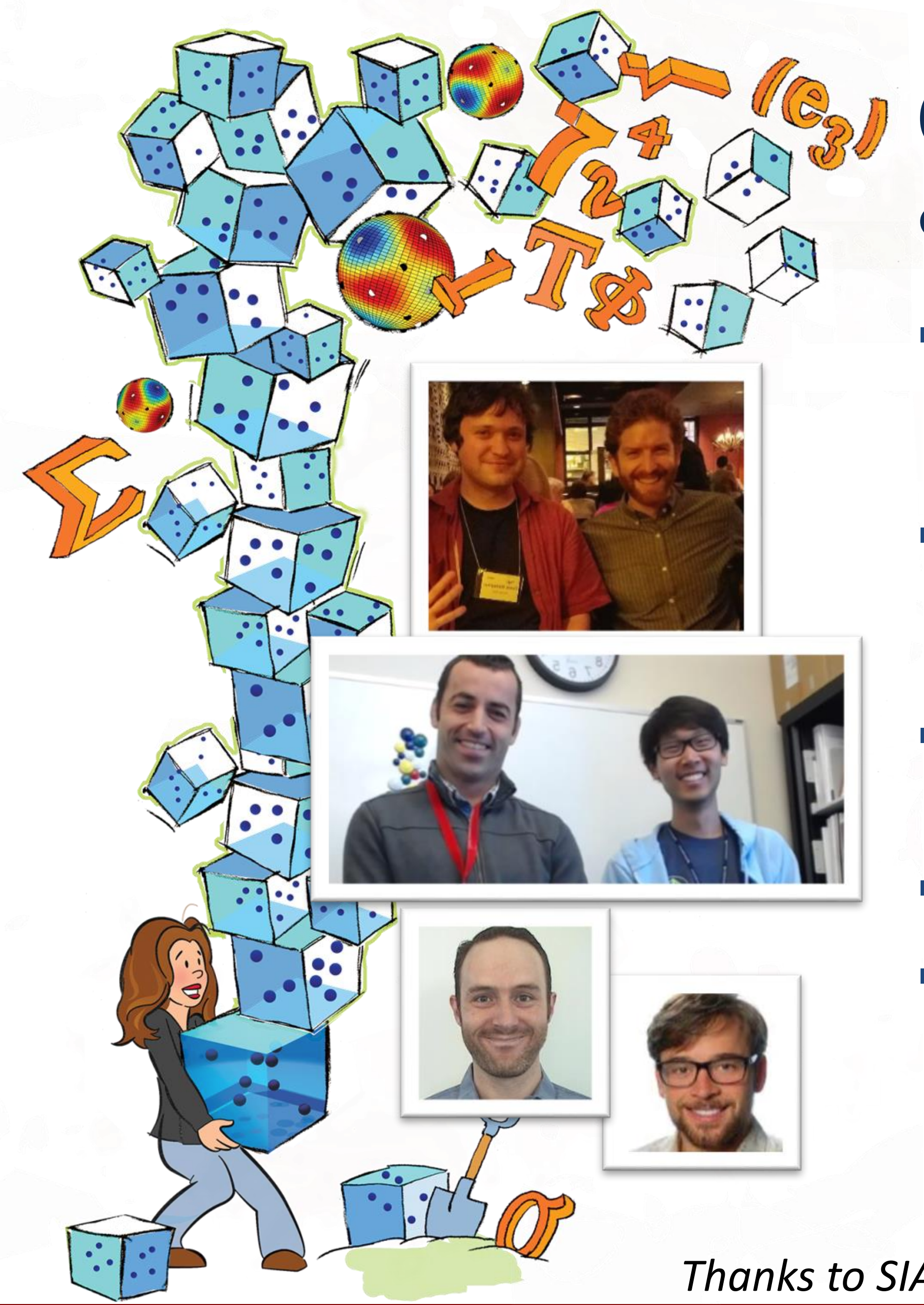
Focus

- Role of applied mathematics in data science, as complemented and intertwined with other key areas: statistics, computer science, network science, signal processing, etc.

Editor in chief: Tamara G. Kolda, Sandia

Section editors

- Alfred Hero, Michigan
- Michel Jordan, Berkeley
- Robert Nowak, Wisconsin
- Joel Tropp, CalTech



CP Tensor Decomposition & Data Analysis

- CP Tensor Decomposition is a key tool for data analysis
  - Latent factor analysis
  - Dimensionality reduction
- Randomized methods enable scaling
  - Initial evidence for increased robustness in global optimization
  - Many, many algorithm and implementation details
- Flexible data type via Generalized CP
  - Nonnegative, Boolean, Poisson data
- *Many open math problems remain!*
- Links
  - Tensor Toolbox for MATLAB: www.tensortoolbox.org
  - Parallel CP and GCP implementations: https://gitlab.com/tensors/genten
  - My web page: www.kolda.net

Thanks to SIAM for the invitation to speak and to YOU, the audience, for your attention!