Illustration by Chris Brigman

Sandia National Laboratories

Generalized Tensor Decompositions for Non-Normal Data

Tamara G. Kolda, Sandia Natl. Labs.

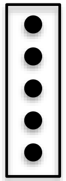
Joint work with David Hong (Michigan), Jed Duersch (SNL)

Supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, 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-NA-0003525.

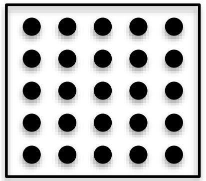
U.S. DEPARTMENT OF ENERGY

A Tensor is an Multi-Way Array

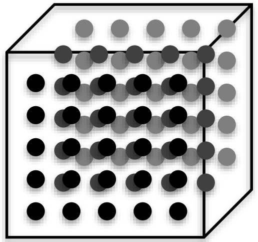
Vector
$d = 1$



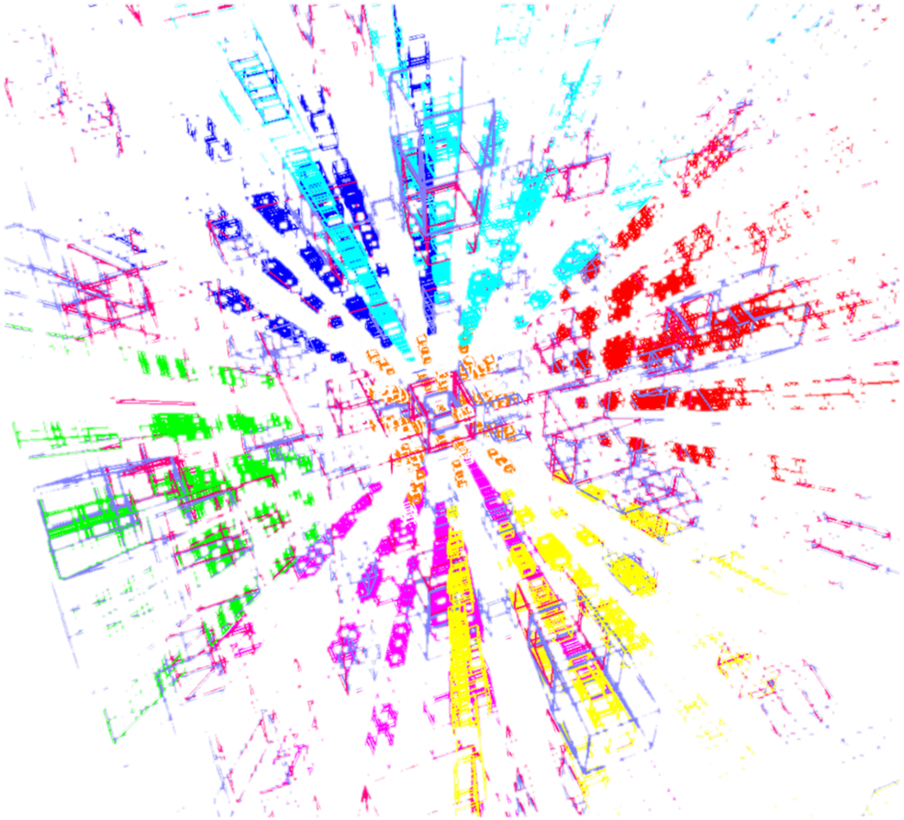
Matrix
$d = 2$



3^{rd}-order Tensor
$d = 3$



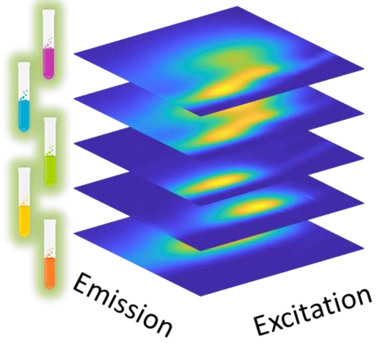
d^{th}-order Tensor
$d > 3$



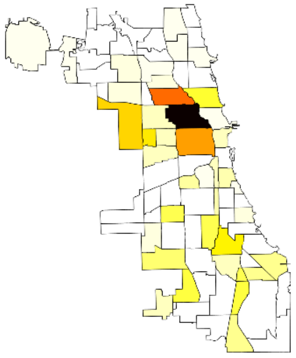
Tensors Come From Many Applications

- **Chemometrics:** Emission x Excitation x Samples (Fluorescence Spectroscopy)
- **Neuroscience:** Neuron x Time x Trial (Calcium Imaging)
- **Criminology:** Day x Hour x Location x Crime (Chicago Crime Reports)
- **Medicine:** Channel x Wavelength x Time (EEG measurements)
- **Sports:** Player x Statistic x Season
- **Cyber-Traffic:** IP x IP x Port x Time
- **Social Network:** Person x Person x Time x Interaction-Type

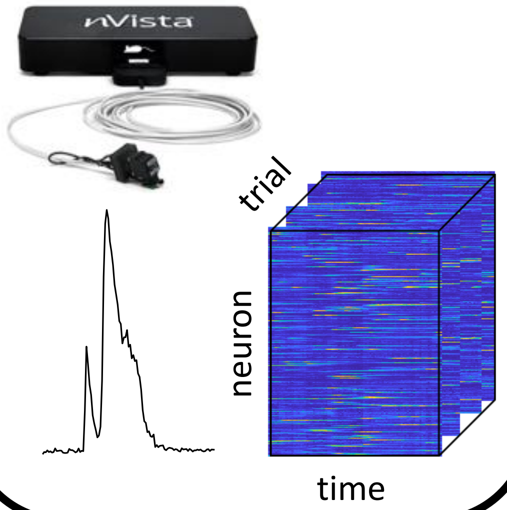
Chemometrics



Criminology

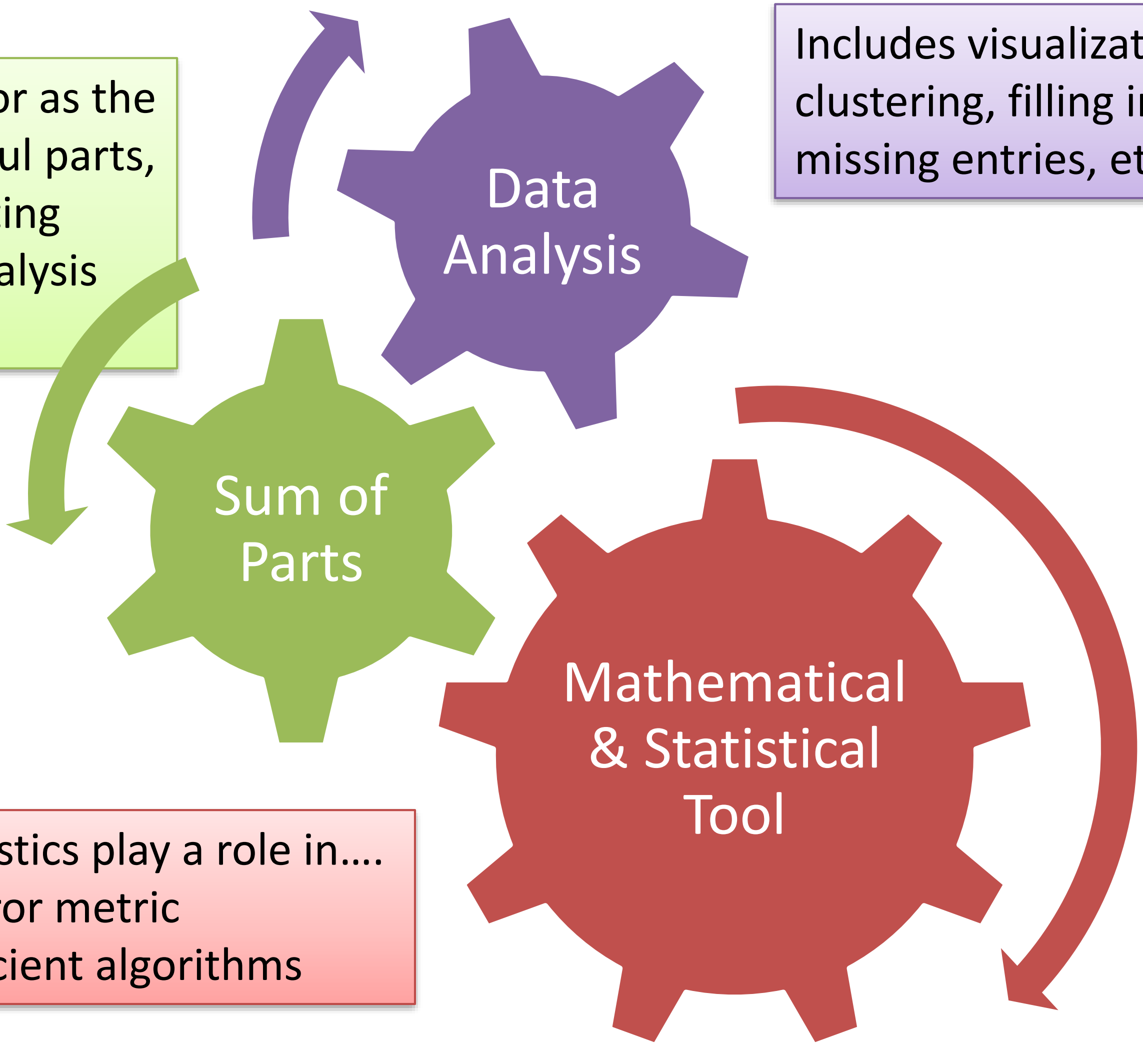


Neuroscience



Tensor Decomposition: A Mathematical & Statistical Tool for Analysis of Tensor Data

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.

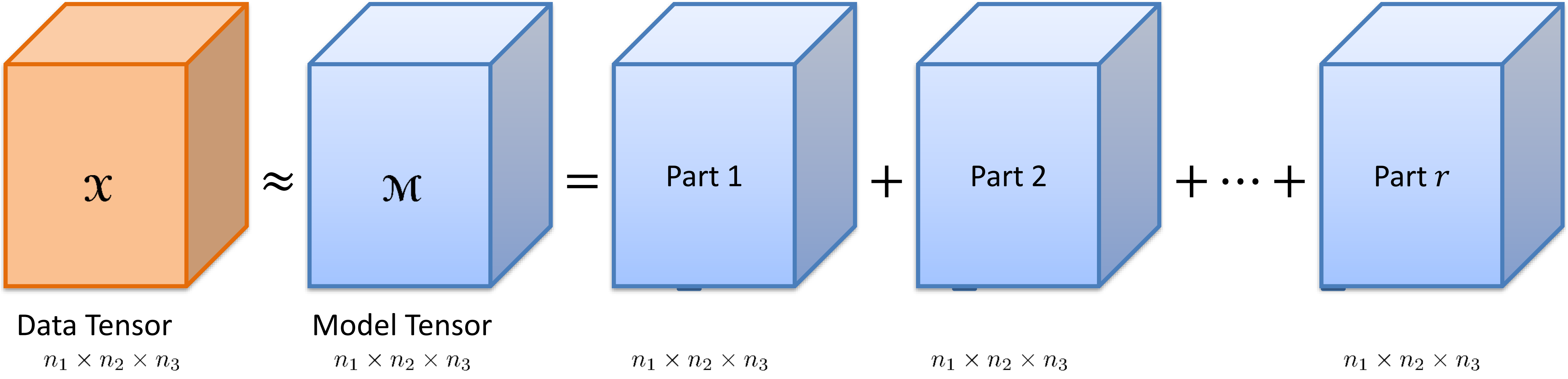
Mathematics/Statistics play a role in….

- Defining the error metric
- Developing efficient algorithms

Related Concepts for Matrices

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

Break Tensor into Understandable Parts…



Key: The parts have structure!

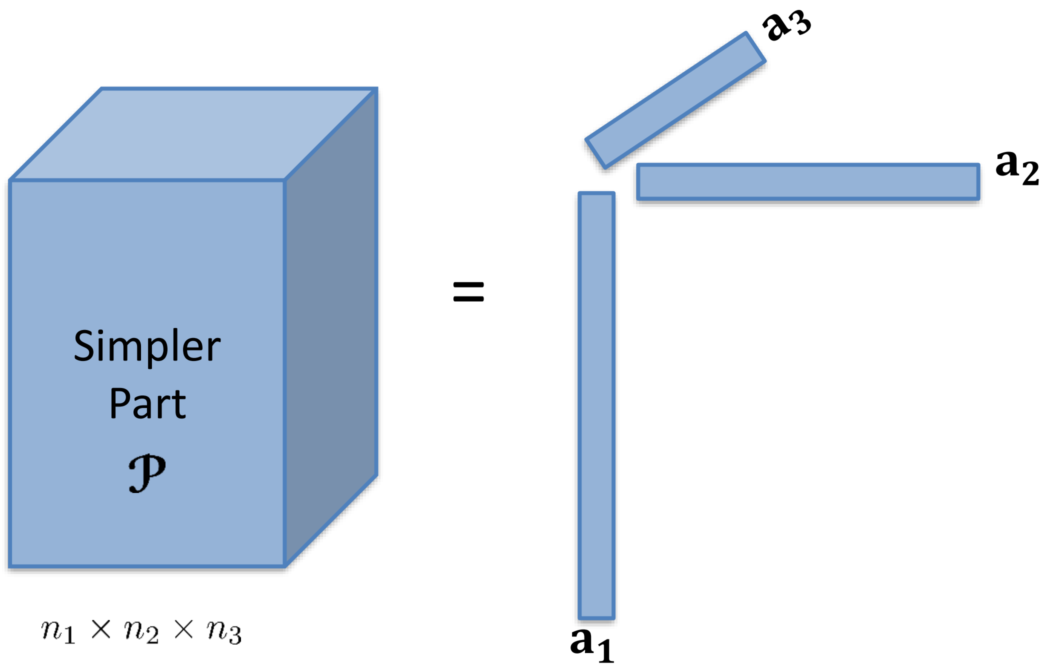
Rank-1 Tensors are the “Parts”

Given ***d*** **vectors**:

$$\mathbf{a}_k \in \mathbb{R}^{n_k} \text{ for } k = 1, \ldots, d$$

The **outer product** is

$$\mathcal{P} = \mathbf{a}_1 \circ \mathbf{a}_2 \cdots \circ \mathbf{a}_d \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$$

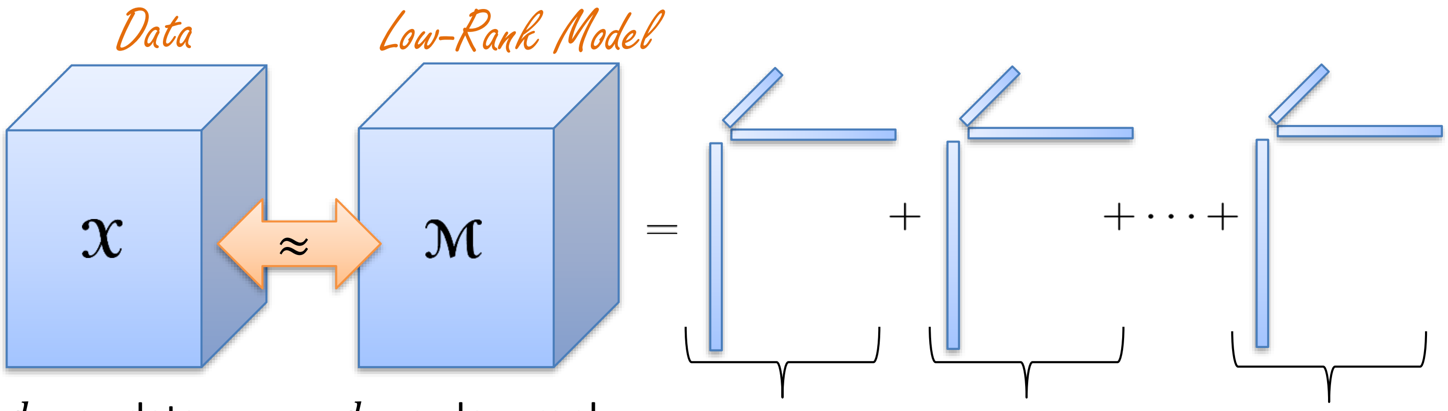


$$\mathcal{P}(i_1, i_2, i_3) = \mathbf{a}_1(i_1)\,\mathbf{a}_2(i_2)\,\mathbf{a}_3(i_3)$$

CANDECOMP/PARAFAC (CP) Tensor Factorization Uncovers the Rank-1 Parts

Images are three-way ($d = 3$), but assume all tensors are of size $n_1 \times n_2 \times \cdots \times n_d$

WLOG, $n = n_1 = \cdots = n_d$



d-way data tensor of size n^d

d-way low-rank model tensor of size n^d and rank r, $\approx nrd$ storage

rank-one component $j = 1$

rank-one component $j = 2$

rank-one component $j = r$

$$\mathcal{X} \approx \mathcal{M} \quad \text{where} \quad \mathcal{M} = \sum_{j=1}^{r} \mathbf{A}_1(:,j) \circ \mathbf{A}_2(:,j) \circ \cdots \circ \mathbf{A}_d(:,j)$$

Factor Matrices

Low-rank: $\operatorname{rank}(\mathcal{M}) \leq r \ll n^d$

Factor matrices: $\mathbf{A}_k \in \mathbb{R}^{n_k \times r}$ for $k \in \{1, \ldots, d\}$

Hitchcock, 1927; Carroll and Chang, 1970; Harshman, 1970

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 $A_{i_1 \cdots i_p}$ of order p *into* a covariant tensor $B_{i_{p+1} \cdots i_{p+q}}$ of order q is defined by writing

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

where the product $C_{i_1 \cdots i_{p+q}}$ 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 $A_{i_1 \cdots i_p}$ can be expressed as the sum of a finite number of tensors each of which is the product of p covariant vectors,

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

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

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

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

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

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

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

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)

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

PARAFAC: Parallel Factors

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

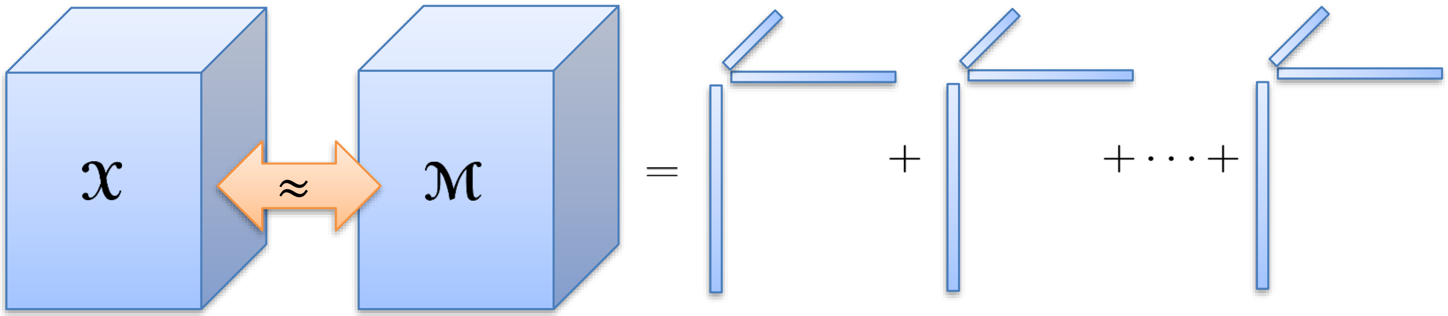
December, 1970

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).

Standard CP: Sum of Squares Error (SSE)



Standard CP

$$\min \;\; F(\mathcal{X},\mathcal{M}) \equiv \sum_{i\in\Omega} (x_i - m_i)^2$$

$$\text{s.t.} \;\; \operatorname{rank}(\mathcal{M}) \leq r$$

Shorthand for element of data tensor:

$$x_i \equiv x(i_1, i_2, \ldots, i_d)$$

Element of model low-rank tensor:

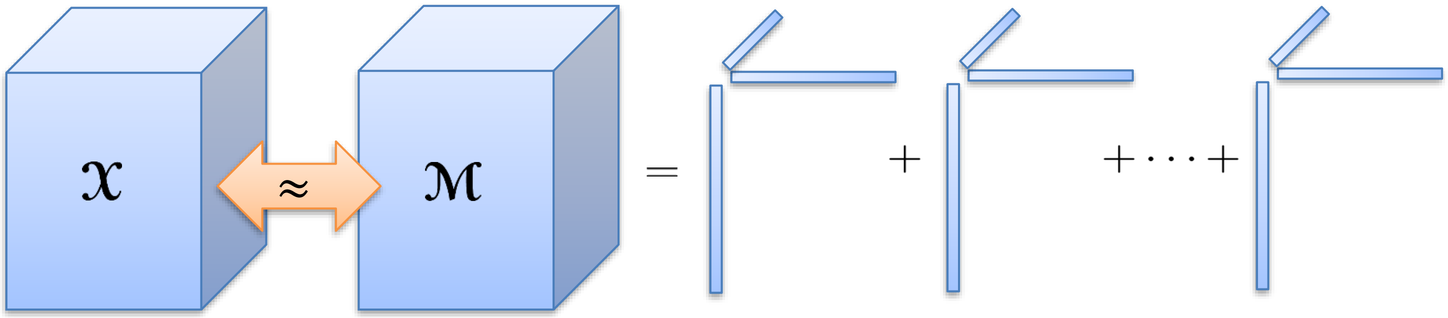
$$m_i \equiv \sum_{j=1}^{r} \prod_{k=1}^{d} \mathbf{A}_k(i_k, j)$$

(defined in terms of factor matrices)

Ω = set of all n^d elements in tensor

Hitchcock, 1927; Carroll and Chang, 1970; Harshman, 1970

Generalized CP (GCP)



GCP

$$\min \; F(\mathcal{X}, \mathcal{M}) \equiv \sum_{i \in \Omega} f(x_i, m_i)$$

$$\text{s.t.} \;\; \text{rank}(\mathcal{M}) \leq r$$

Why?

- SSE: maximum likelihood estimate (MLE) for Gaussian distribution

$$x_i = m_i + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma)$$

$$x_i \sim \mathcal{N}(m_i, \sigma)$$

- Different MLEs for different distributions
  - Poisson (counts)
  - Bernoulli (binary)

Hong, Kolda, Duersch, SIAM Review, 2019

Probability Distribution ⇒ Maximum Likelihood Estimator

Data Value

"Natural" Parameter

Model Value

$$x_i \sim p(x_i|\theta_i) \text{ where } \ell(\theta_i) = m_i$$

Link Function

Probability Distribution Function (PDF) or Probability Mass Function (PMF)

Maximize Likelihood of Data Tensor

$$\prod_{i \in \Omega} p(x_i, \theta_i)$$

Maximize Log-Likelihood

$$\sum_{i \in \Omega} \log p(x_i, \theta_i)$$

GCP

$$\min \; F(\mathfrak{X}, \mathfrak{M}) \equiv \sum_{i \in \Omega} f(x_i, m_i)$$

$$\text{s.t. } \operatorname{rank}(\mathfrak{M}) \leq r$$

Given PDF/PMF $p(x|\theta)$ and link function $\ell(\theta)$, GCP MLE by minimizing

$$f(x, m) = -\log p(x, \ell^{-1}(m))$$

Hong, Kolda, Duersch, SIAM Review, 2019

Gaussian MLE (Standard CP)

PDF for Normal Distribution

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

and

Link Function

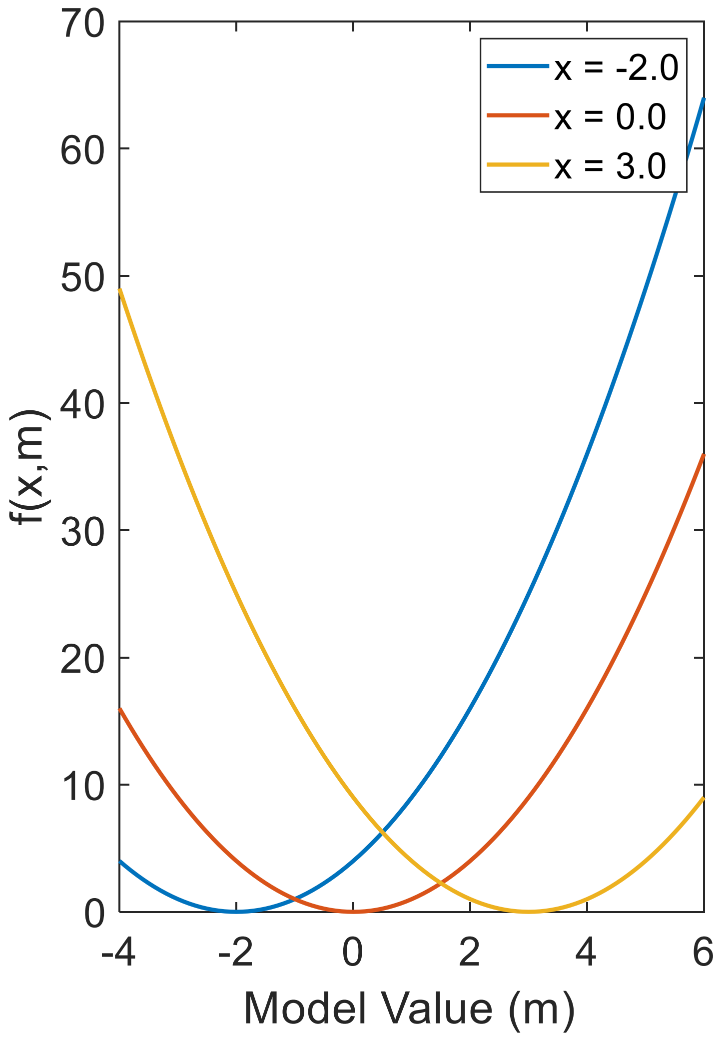
$$m = \mu$$

σ constant

Negative log-likelihood: $$-\log p(x|\mu,\sigma) = \frac{(x-u)^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$$

Eliminate natural parameter via link function: $$f(x,m) = \frac{(x-m)^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$$

Eliminate constants: $$f(x,m) = (x-m)^2$$



Hong, Kolda, Duersch, SIAM Review, 2019

Bernoulli MLE with Odds Link (Binary Data)

Bernoulli random variable

$x \in \{0,1\}$

ρ = probability of a 1

$$p(x \mid \rho) = \rho^x (1-\rho)^{(1-x)}, \quad x \in \{\,0,1\,\}$$

PMF for Bernoulli Distribution

$$p(x \mid \rho) = \rho^x (1-\rho)^{(1-x)}$$
$$x \in \{\,0,1\,\}$$

and

Link Function

$$m = \frac{\rho}{(1-\rho)}$$

Odds Link

$$\ell(\rho) = \rho \,/\, (1-\rho)$$
$$\ell^{-1}(m) = m \,/\, (1+m)$$

Odds (m)	Probability (ρ)
¼	20%
1	50%
4	80%
10	90%

Negative log-likelihood:

$$-\log p(x \mid \rho) = \log \frac{1}{1-\rho} - x \log \frac{\rho}{1-\rho}$$

Eliminate natural parameter via link function:

$$f(x,m) = \log(1+m) - x \log m \quad \text{for} \quad m > 0$$

Hong, Kolda, Duersch, SIAM Review, 2019

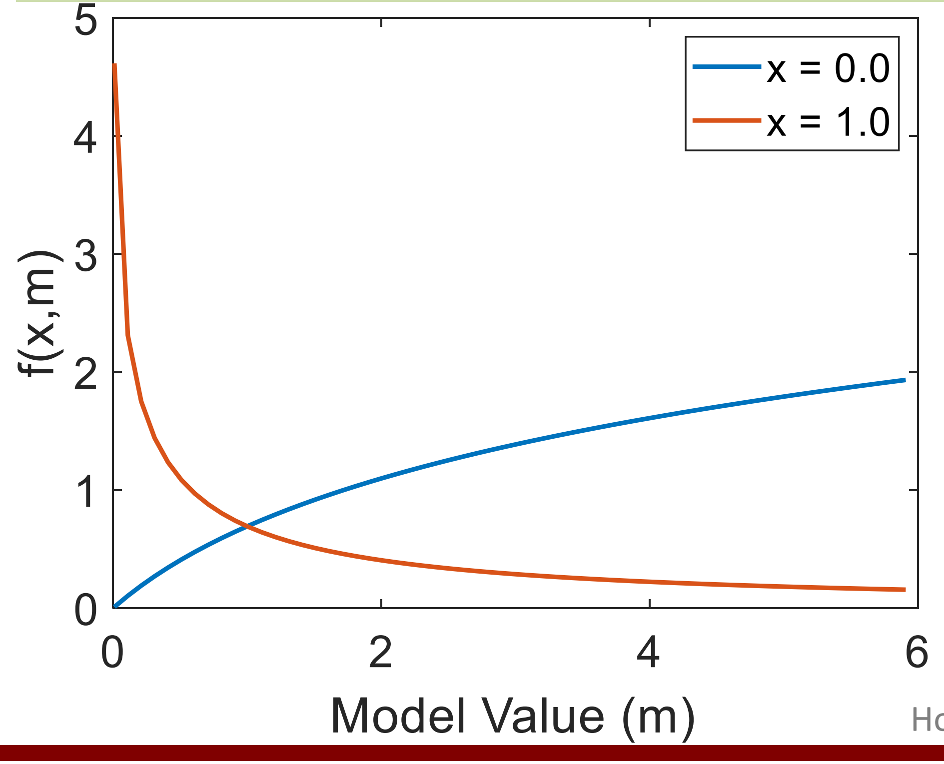
Bernoulli MLE with Odds Link (Binary Data)

Bernoulli random variable

$x \in \{0,1\}$

ρ = probability of a 1

$$p(x \mid \rho) = \rho^x (1-\rho)^{(1-x)}, \quad x \in \{0,1\}$$



PMF for Bernoulli Distribution

$$p(x \mid \rho) = \rho^x (1-\rho)^{(1-x)}$$
$$x \in \{0,1\}$$

and

Link Function

$$m = \frac{\rho}{(1-\rho)}$$

Negative log-likelihood:

$$-\log p(x \mid \rho) = \log \frac{1}{1-\rho} - x \log \frac{\rho}{1-\rho}$$

Eliminate natural parameter via link function:

$$f(x,m) = \log(1+m) - x \log m \quad \text{for} \quad m > 0$$

Hong, Kolda, Duersch, SIAM Review, 2019

Bernoulli MLE with Logit Link (Binary Data)

Bernoulli random variable

$x \in \{0,1\}$

ρ = probability of a 1

$$p(x \mid \rho) = \rho^x (1-\rho)^{(1-x)}, \quad x \in \{0,1\}$$

Logit (Log-Odds) Link

$$\ell(\rho) = \log\left(\rho \,/\, (1-\rho)\right)$$

$$\ell^{-1}(m) = e^m \,/\, (1+e^m)$$

Log-Odds(m)	Probability (ρ)
-1.39	20%
0	50%
1.39	80%
2.30	90%

PMF for Bernoulli Distribution

$$p(x \mid \rho) = \rho^x (1-\rho)^{(1-x)}$$

$$x \in \{0,1\}$$

and

Link Function

$$m = \log \frac{\rho}{(1-\rho)}$$

Negative log-likelihood:

$$-\log p(x \mid \rho) = \log \frac{1}{1-\rho} - x \log \frac{\rho}{1-\rho}$$

Eliminate natural parameter via link function:

$$f(x,m) = \log(1+e^m) - xm \quad \text{for} \quad m \in \mathbb{R}$$

Hong, Kolda, Duersch, SIAM Review, 2019

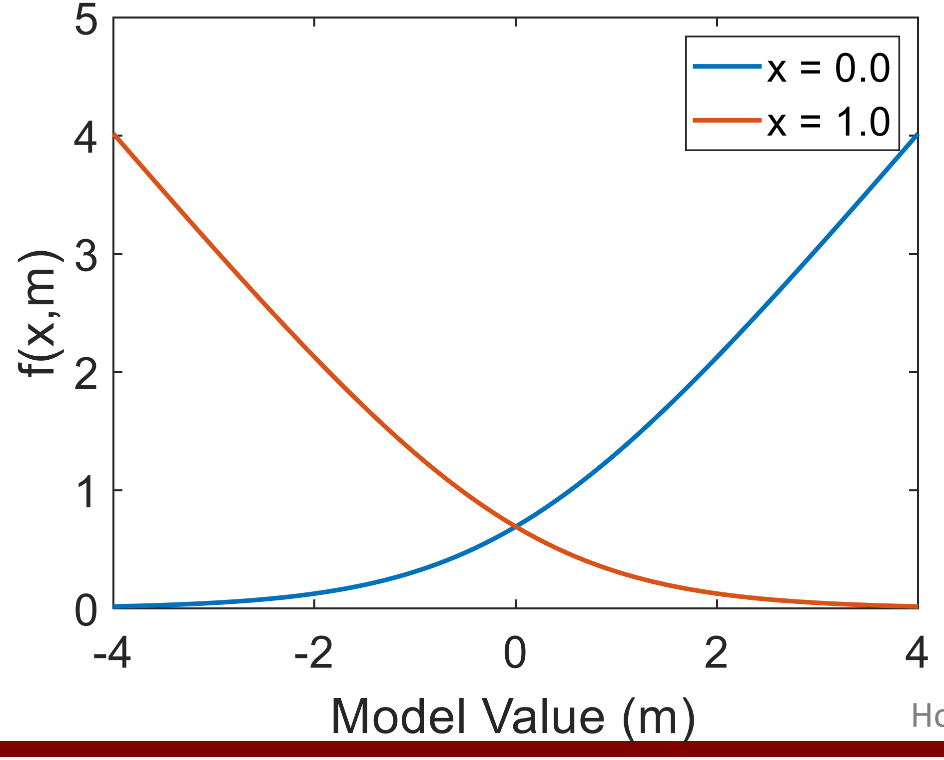
Bernoulli MLE with Logit Link (Binary Data)

Bernoulli random variable

$x \in \{0,1\}$

ρ = probability of a 1

$$p(x \mid \rho) = \rho^x (1-\rho)^{(1-x)}, \quad x \in \{0,1\}$$



PMF for Bernoulli Distribution

$$p(x \mid \rho) = \rho^x (1-\rho)^{(1-x)}$$
$$x \in \{0,1\}$$

and

Link Function

$$m = \log \frac{\rho}{(1-\rho)}$$

Negative log-likelihood:

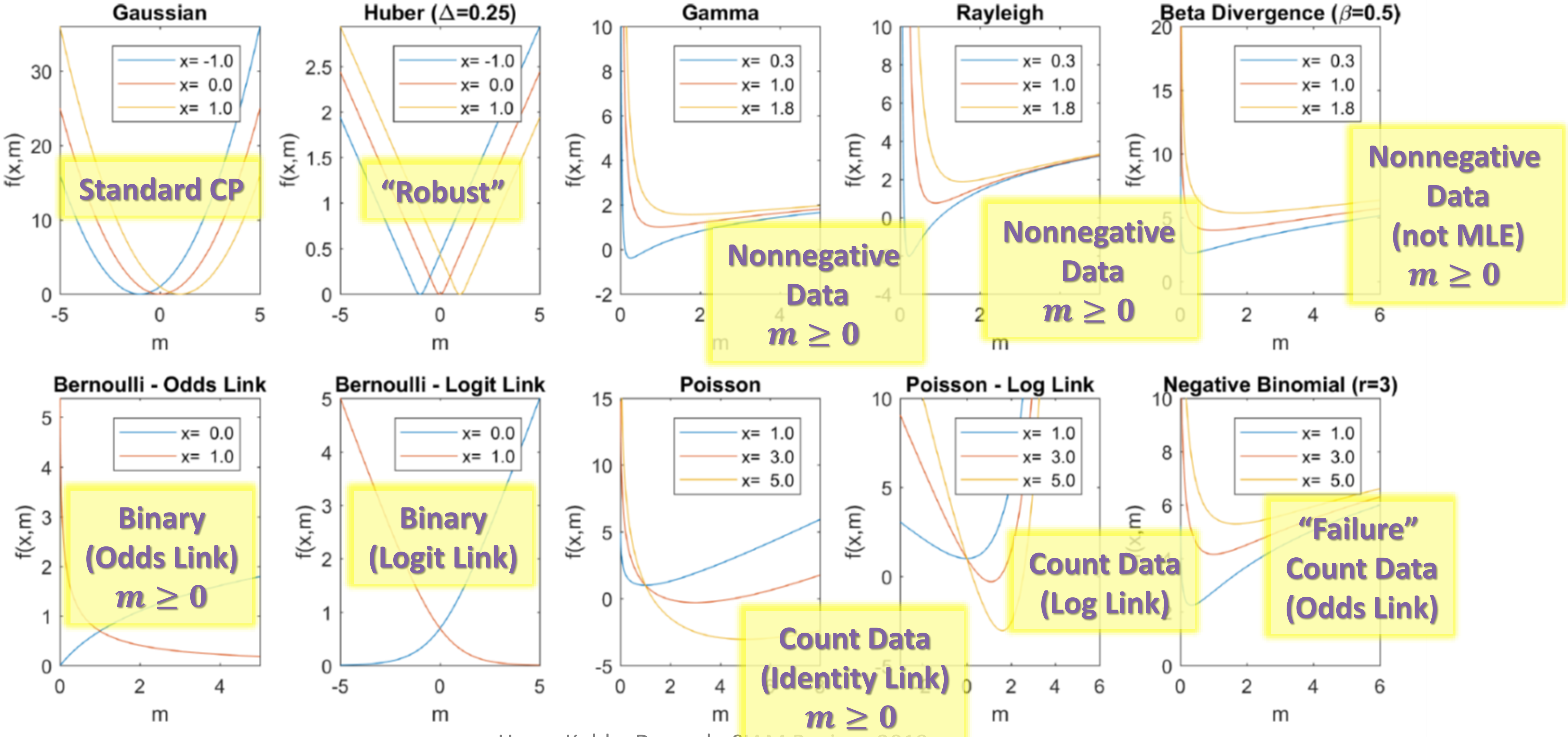
$$-\log p(x \mid \rho) = \log \frac{1}{1-\rho} - x \log \frac{\rho}{1-\rho}$$

Eliminate natural parameter via link function:

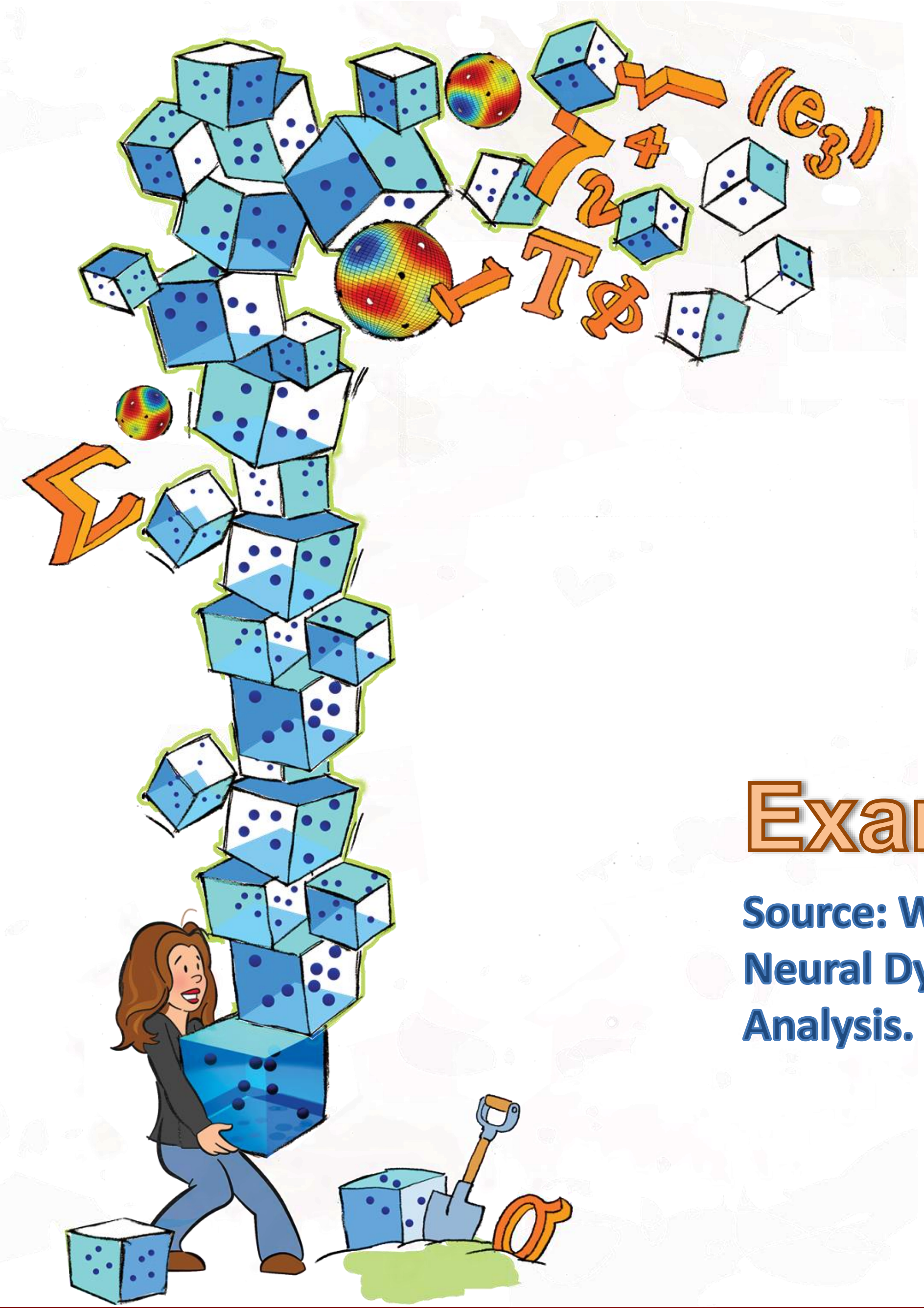
$$f(x,m) = \log(1+e^m) - xm \quad \text{for} \quad m \in \mathbb{R}$$

Hong, Kolda, Duersch, SIAM Review, 2019

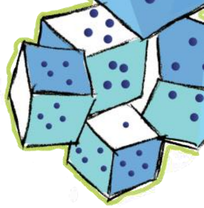
Sampling of Loss Functions



Hong, Kolda, Duersch, SIAM Review, 2019



Sandia National Laboratories



Example Tensor from Neuroscience

Source: Williams, et al. Unsupervised Discovery of Demixed, Low-dimensional Neural Dynamics across Multiple Timescales through Tensor Components Analysis. Neuron, 2018. https://doi.org/10.1016/j.neuron.2018.05.015

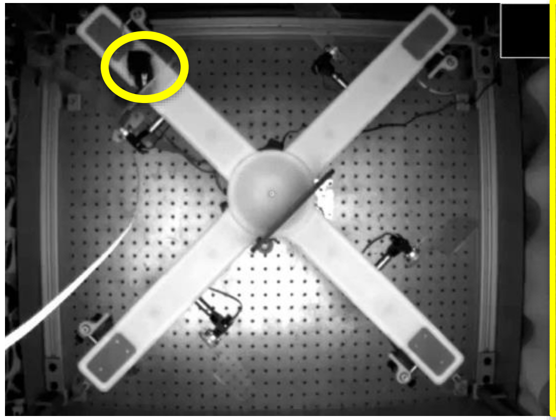
Activity of Single Neuron Measured Over Time Produces Vector Data

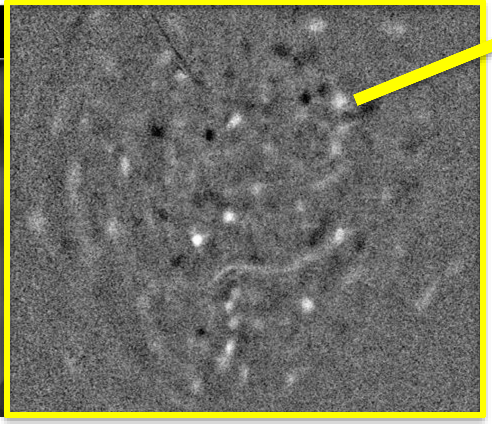
Thanks to Schnitzer Group @ Stanford
Mark Schnitzer, Fori Wang, Tony Kim

Microscope by Inscopix

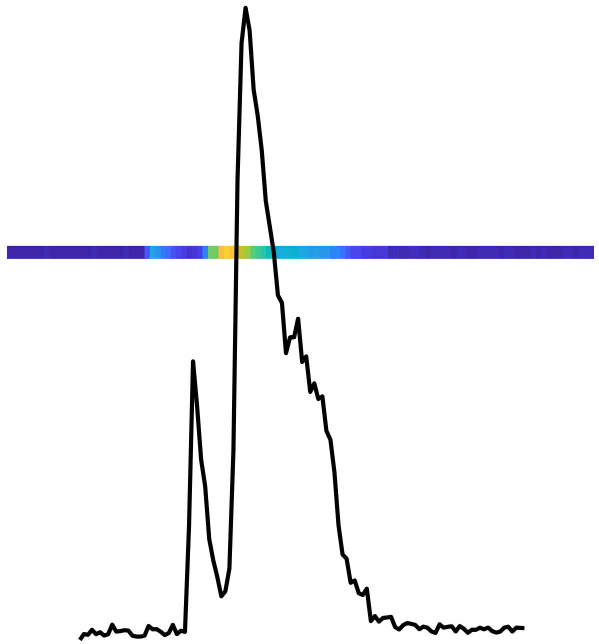
mouse in maze

neural activity via calcium imaging





111 time bins



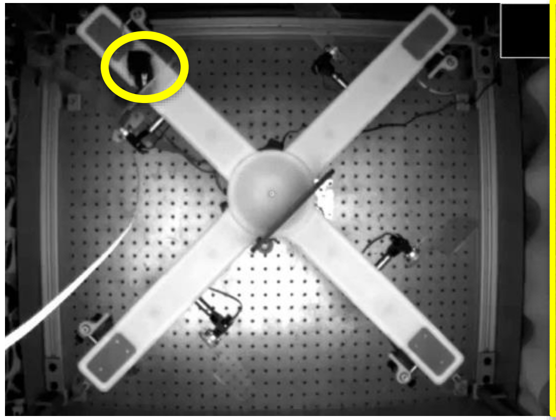
Williams et al., Neuron, 2018

Multiple Neurons Measured Over Time Produces Matrix

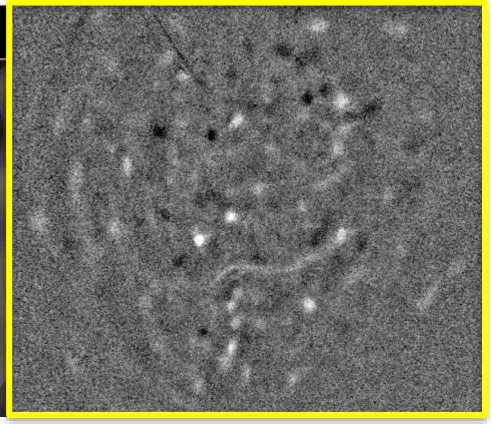
Thanks to Schnitzer Group @ Stanford
Mark Schnitzer, Fori Wang, Tony Kim

Microscope by Inscopix

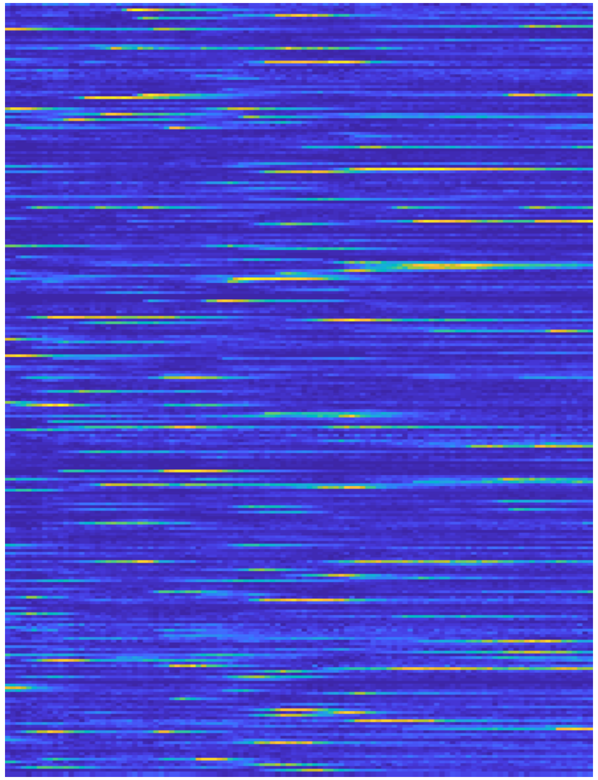
mouse
in “maze”



neural activity

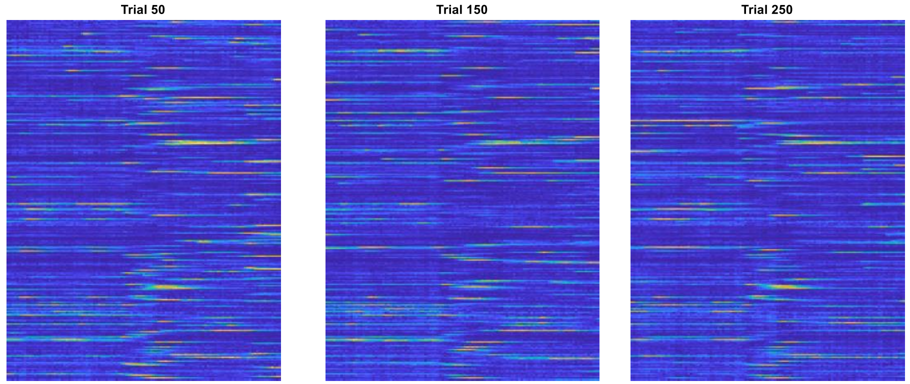


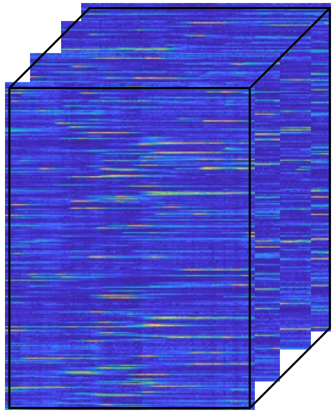
282 neurons × 111 time bins

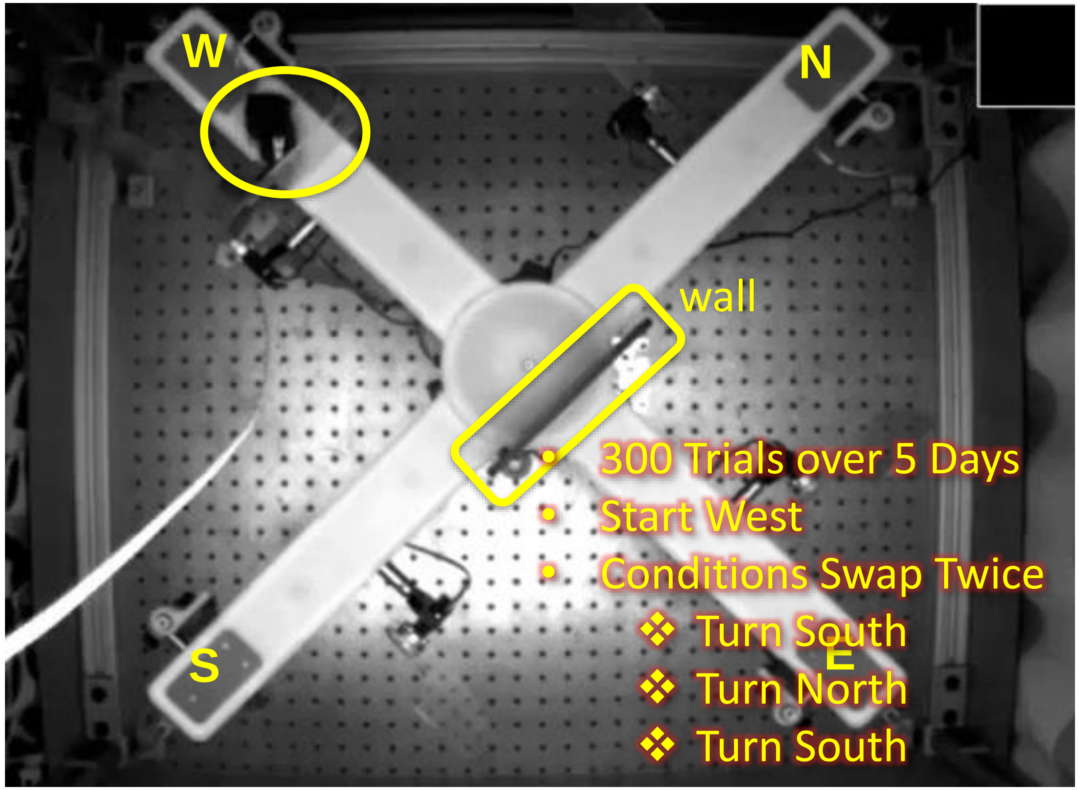


Williams et al., Neuron, 2018

Multiple Trials Produces 3-way Tensor







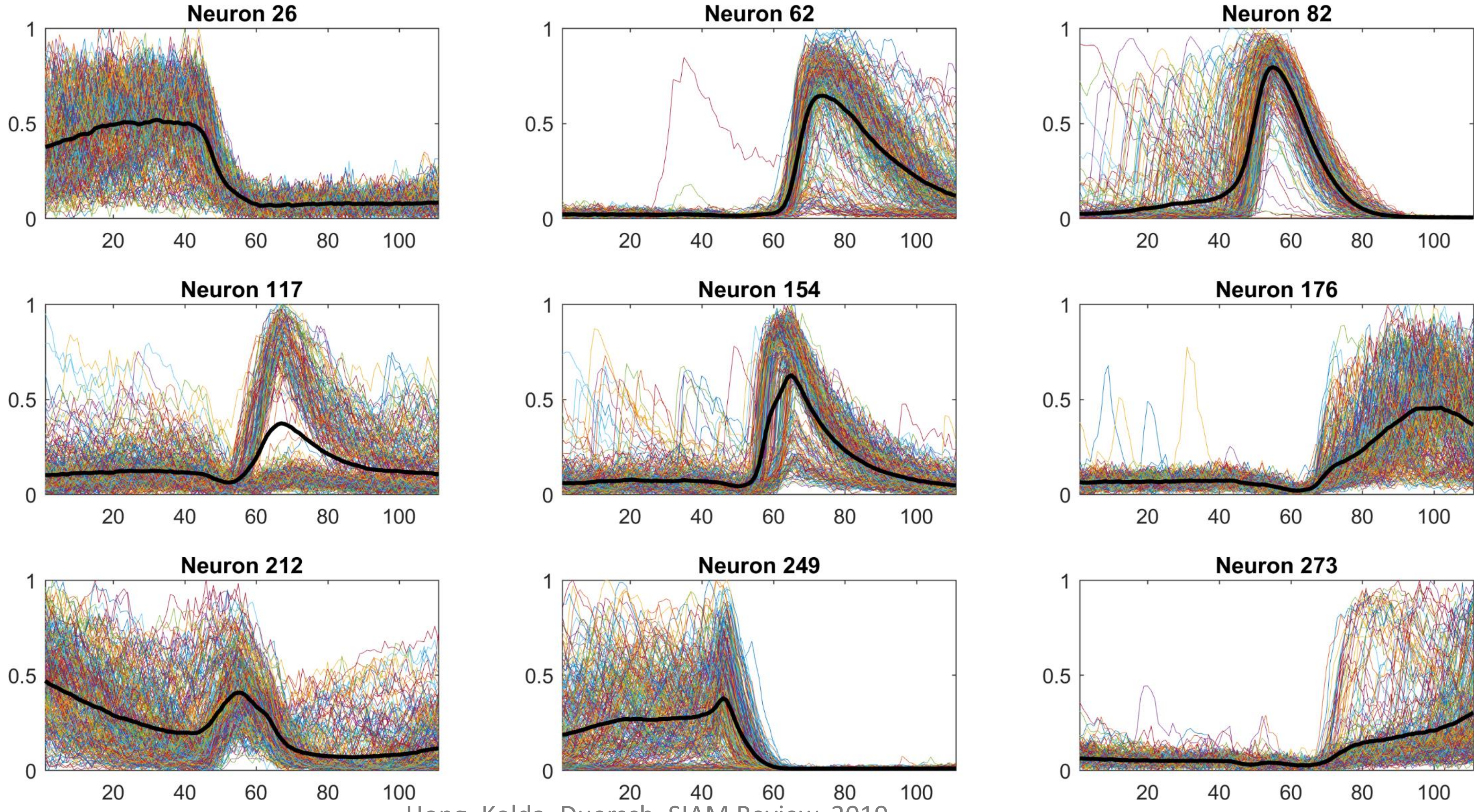
282 neurons × 111 time bins × 300 trials

Williams et al., Neuron, 2018

Example Neuron Activity

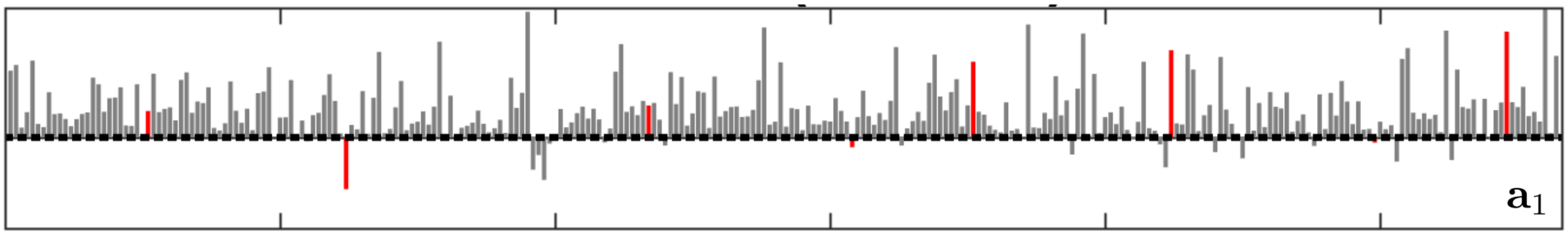
Thin lines show 300 individual trials

Thick line is average

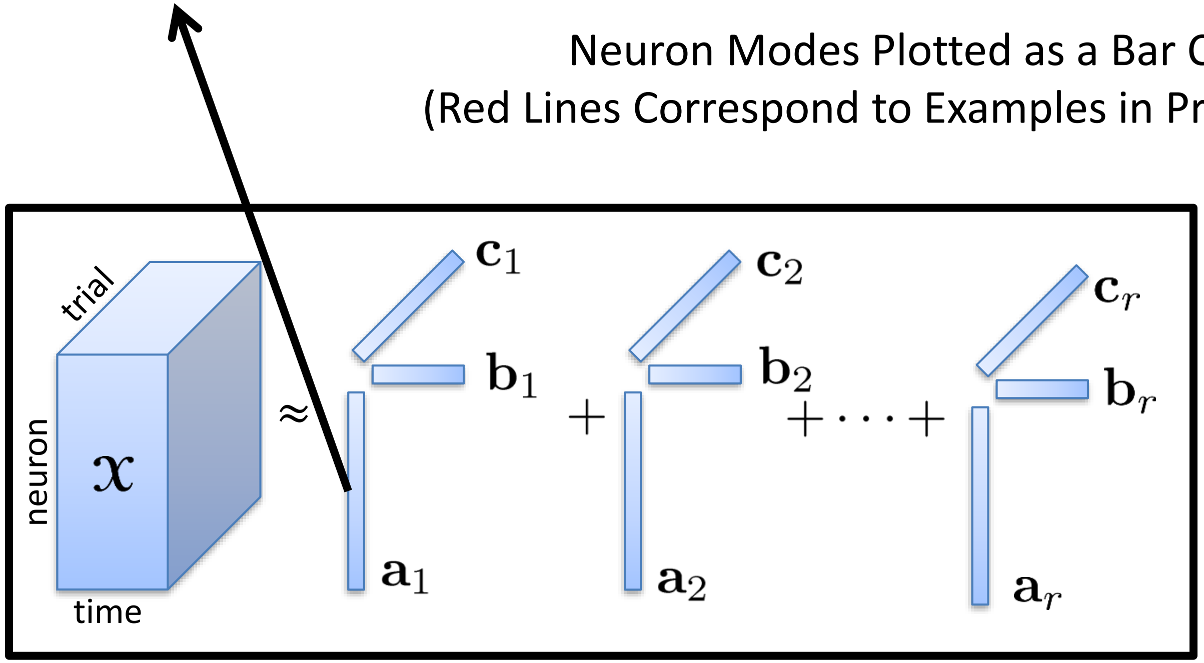


Hong, Kolda, Duersch, SIAM Review, 2019

Neuron Factor Vector Visualized as Bar Chart

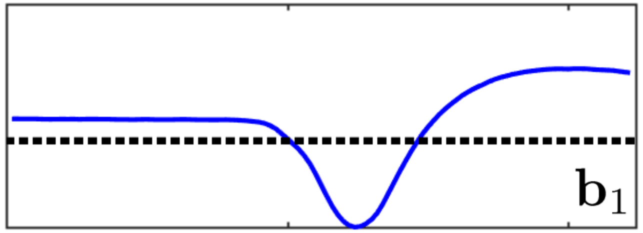


Neuron Modes Plotted as a Bar Chart
(Red Lines Correspond to Examples in Previous Slide)

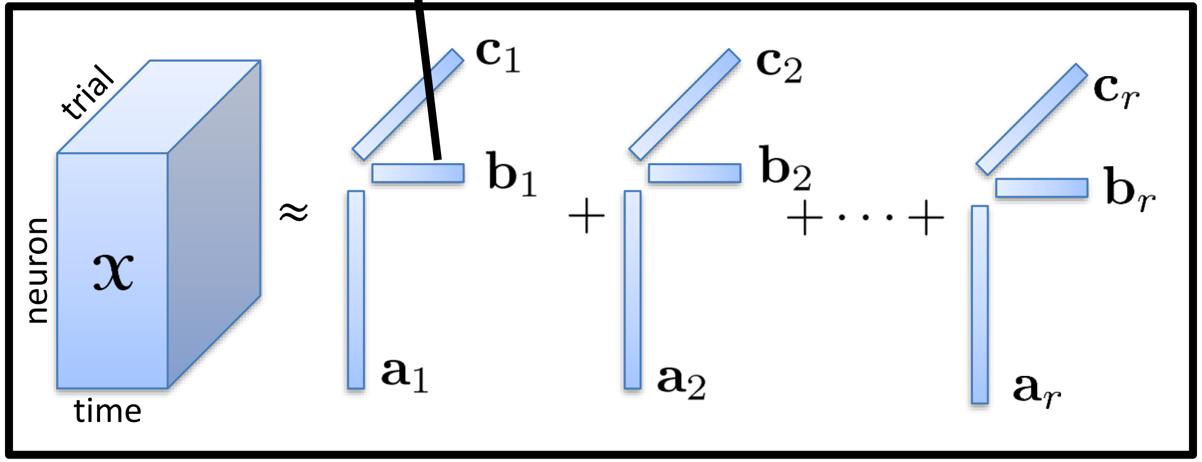


Hong, Kolda, Duersch, SIAM Review, 2019

Time Factor Vector Visualized as Line

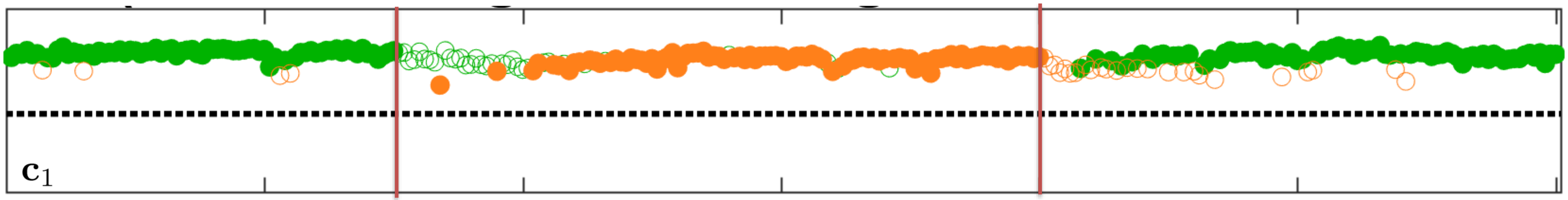


Time (within trial) Plotted as a Line
(Dashed Line is Zero)



Hong, Kolda, Duersch, SIAM Review, 2019

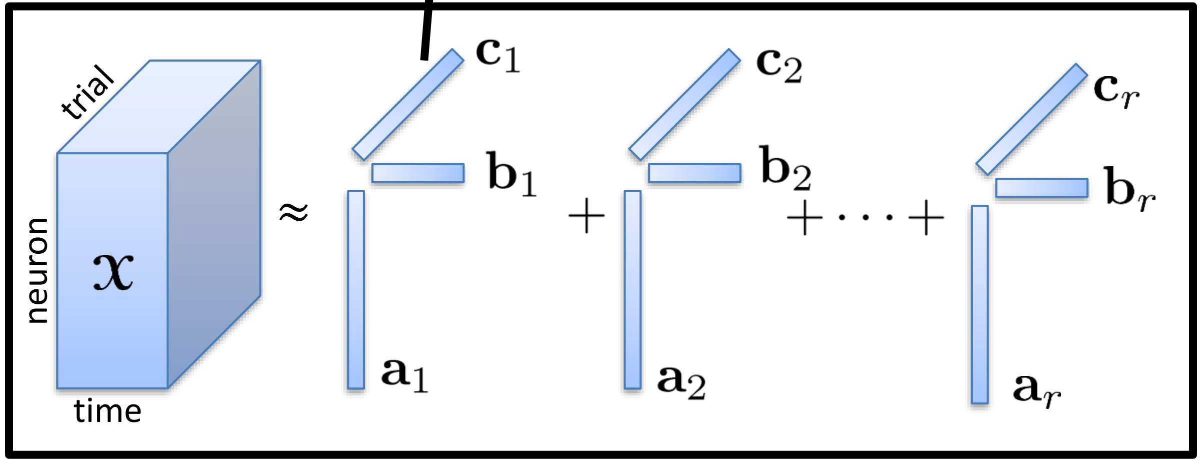
Trial Factor Vector Visualized as Color-Coded Scatter Plot



Rule Change

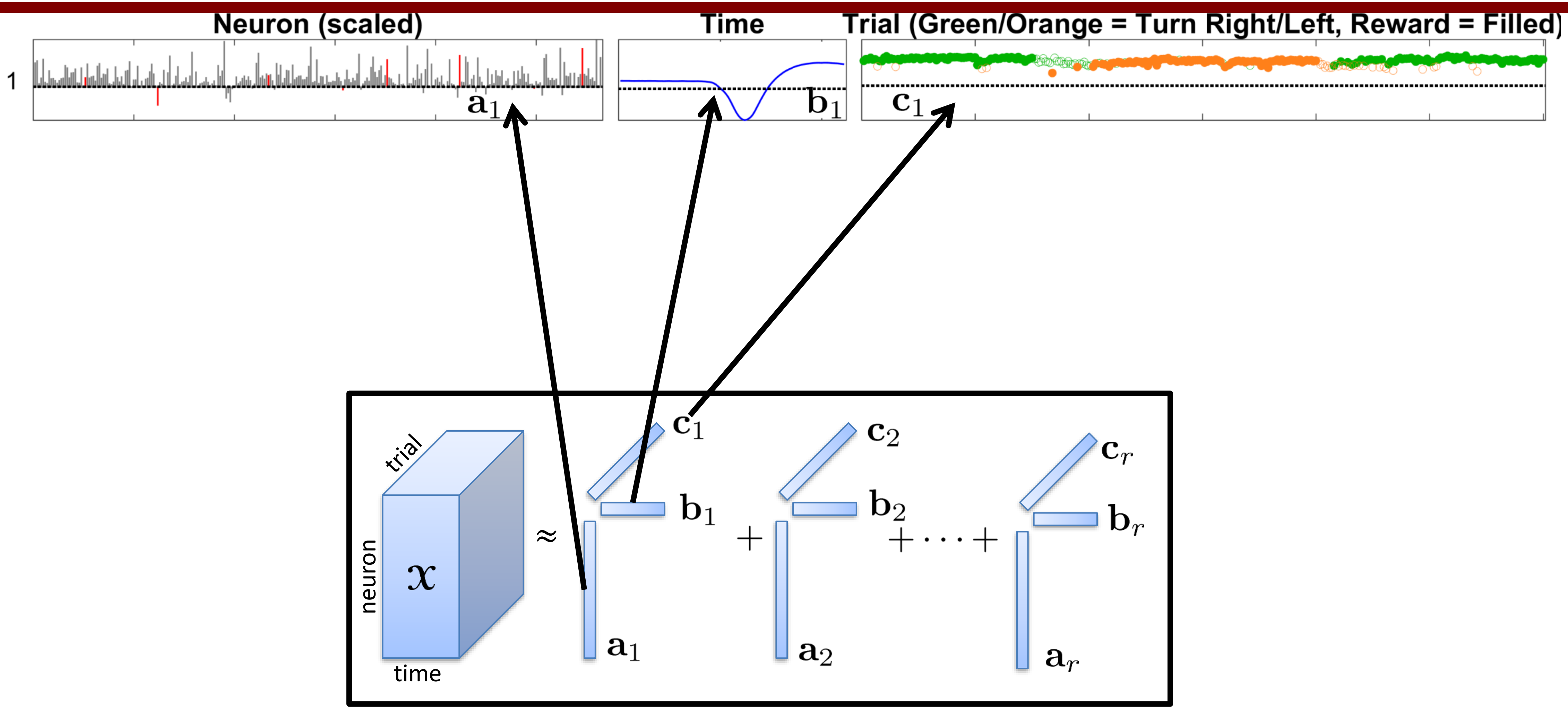
Trial Plotted as Scatter Graph
Right turn = Green
Left turn = Orange
Filled = Reward

Rule Change



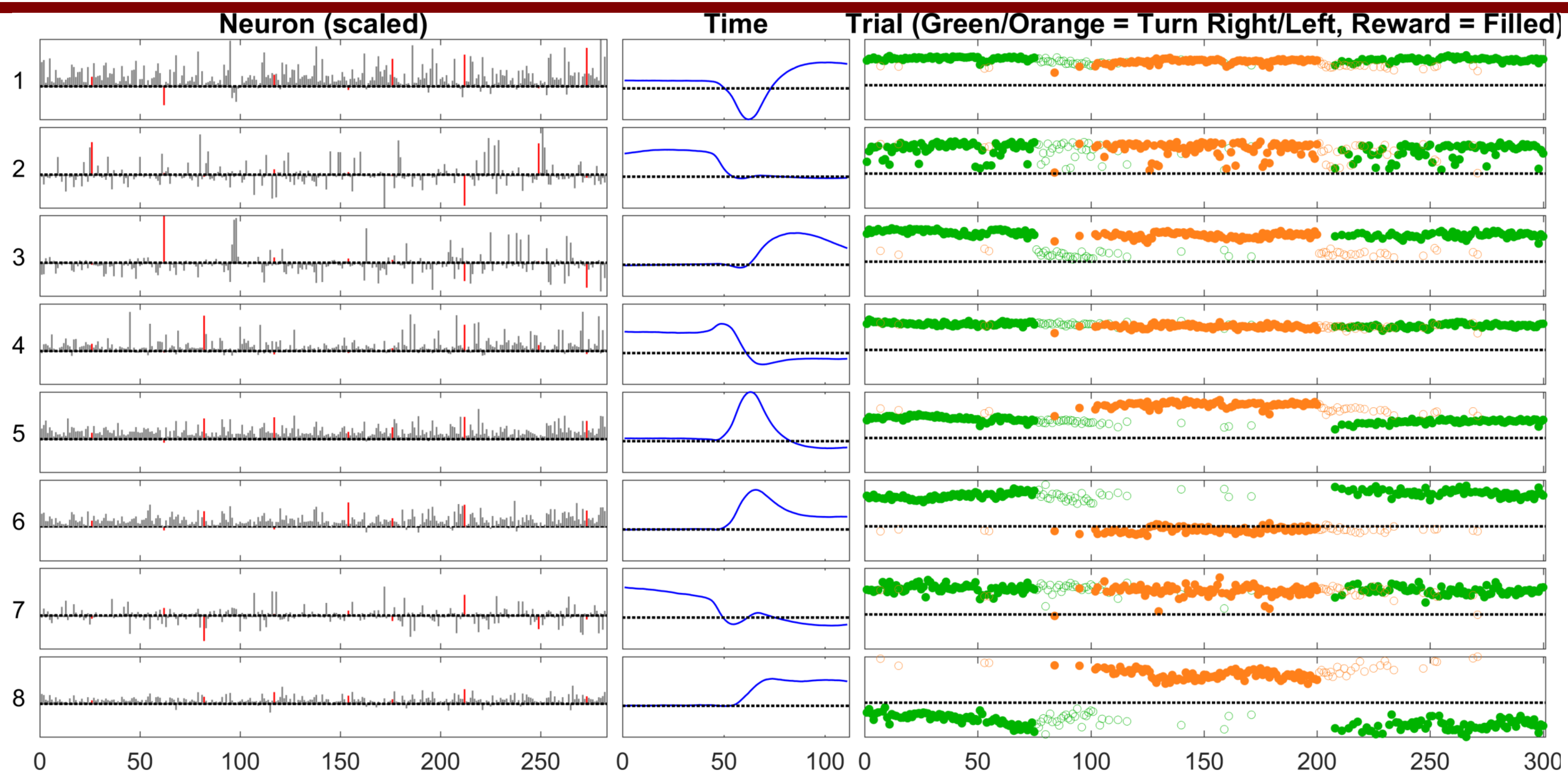
Hong, Kolda, Duersch, SIAM Review, 2019

Visualization of CP Tensor Decomposition Shows the Factors (Vectors)

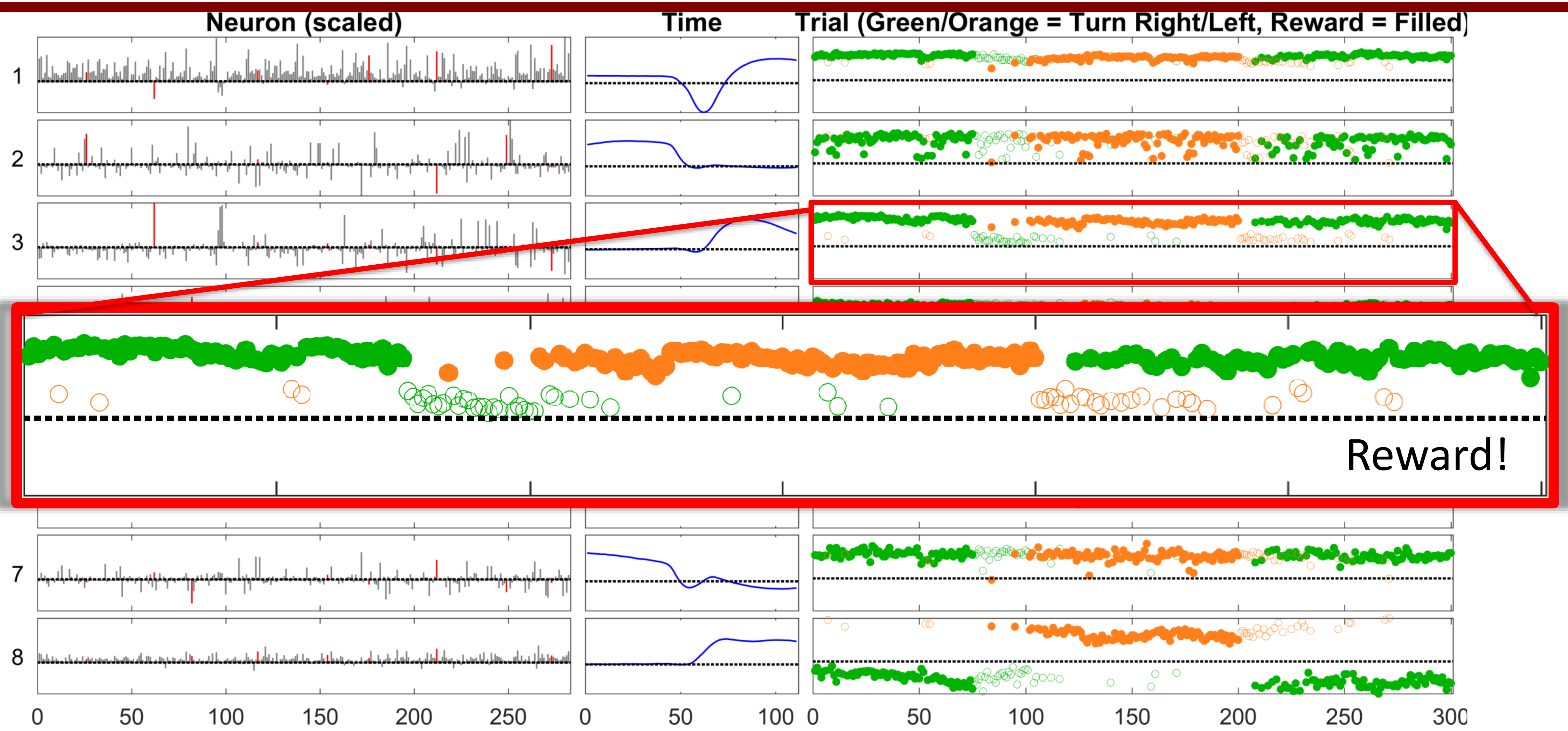


Hong, Kolda, Duersch, SIAM Review, 2019

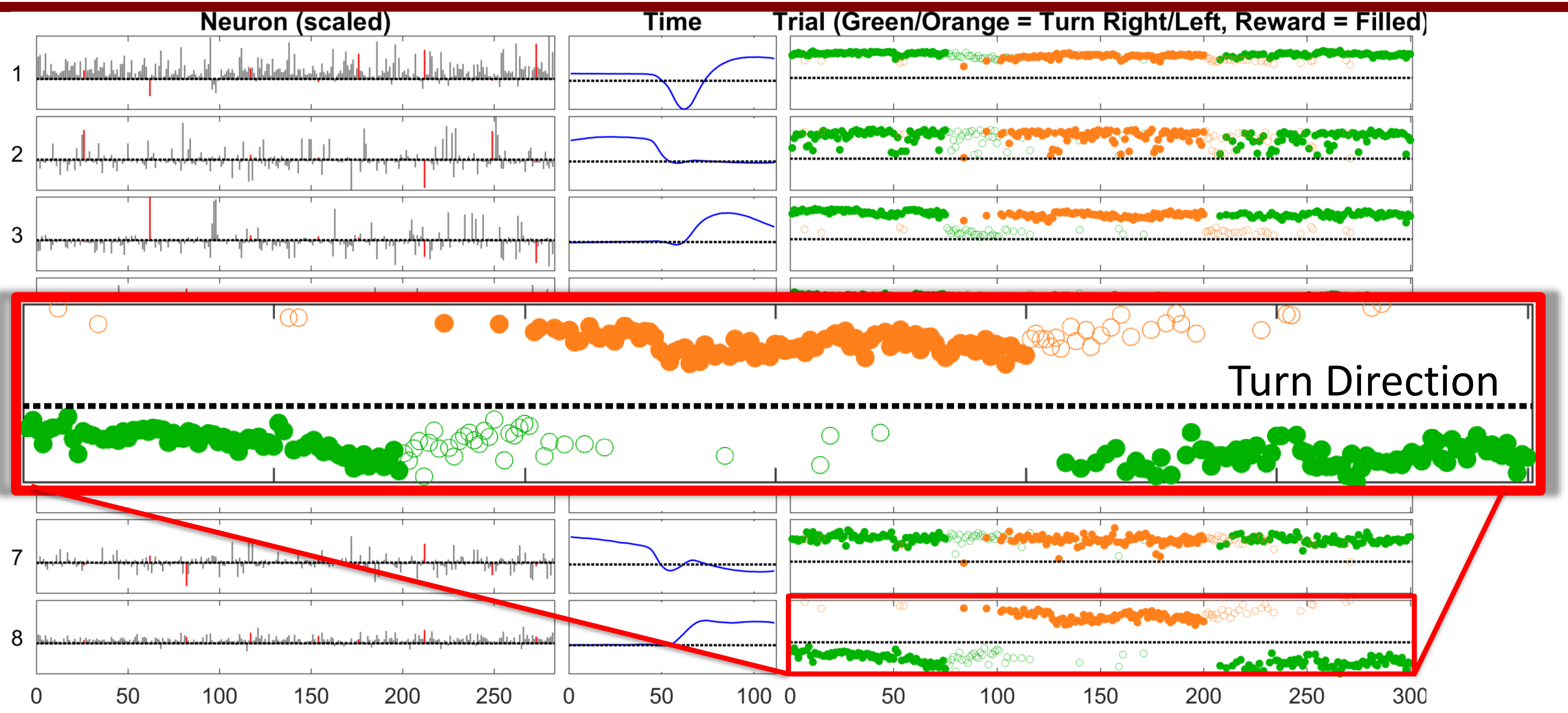
"Standard" CP Decomposition of Mouse Data, aka Gaussian ($f(x, m) = (x - m)^2$)



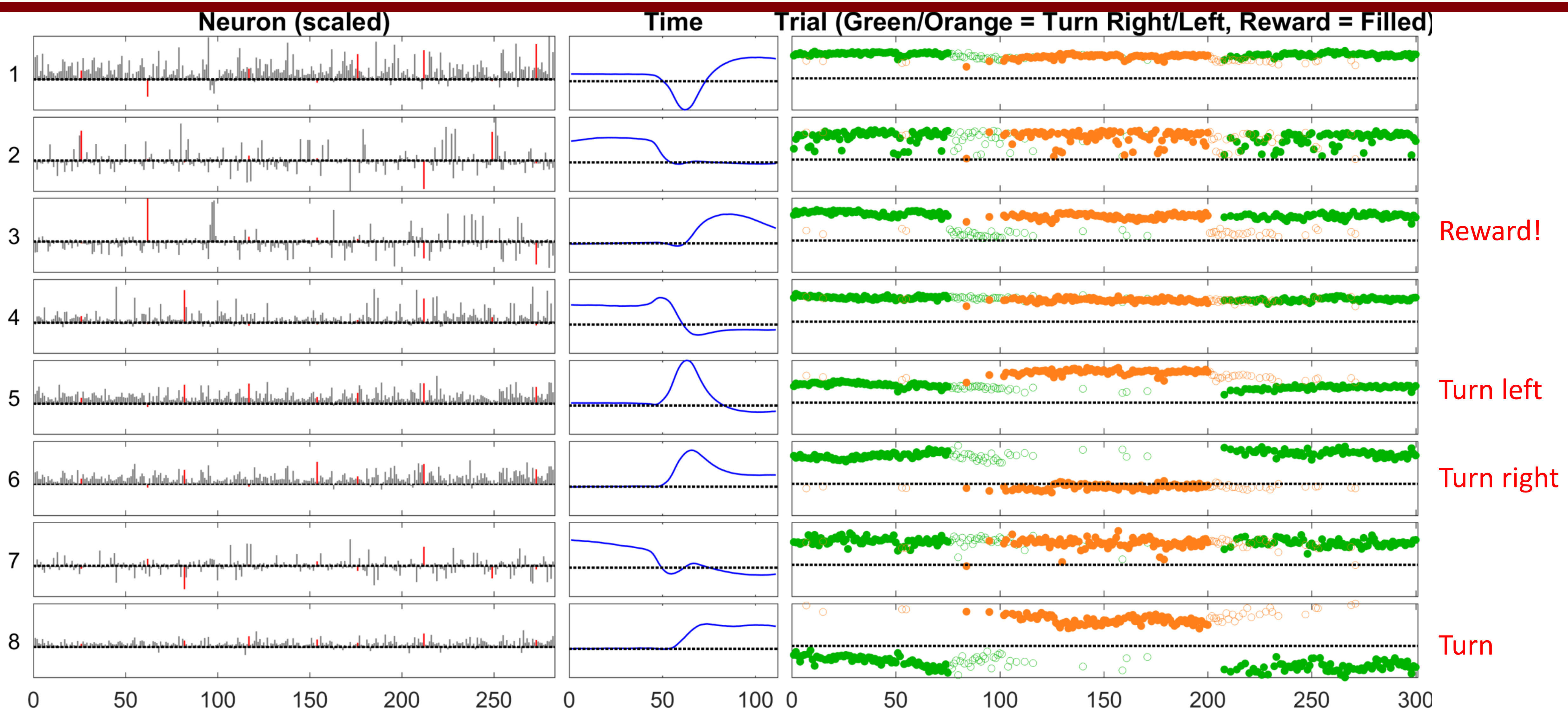
CP Tensor Decomposition "Sees" Reward



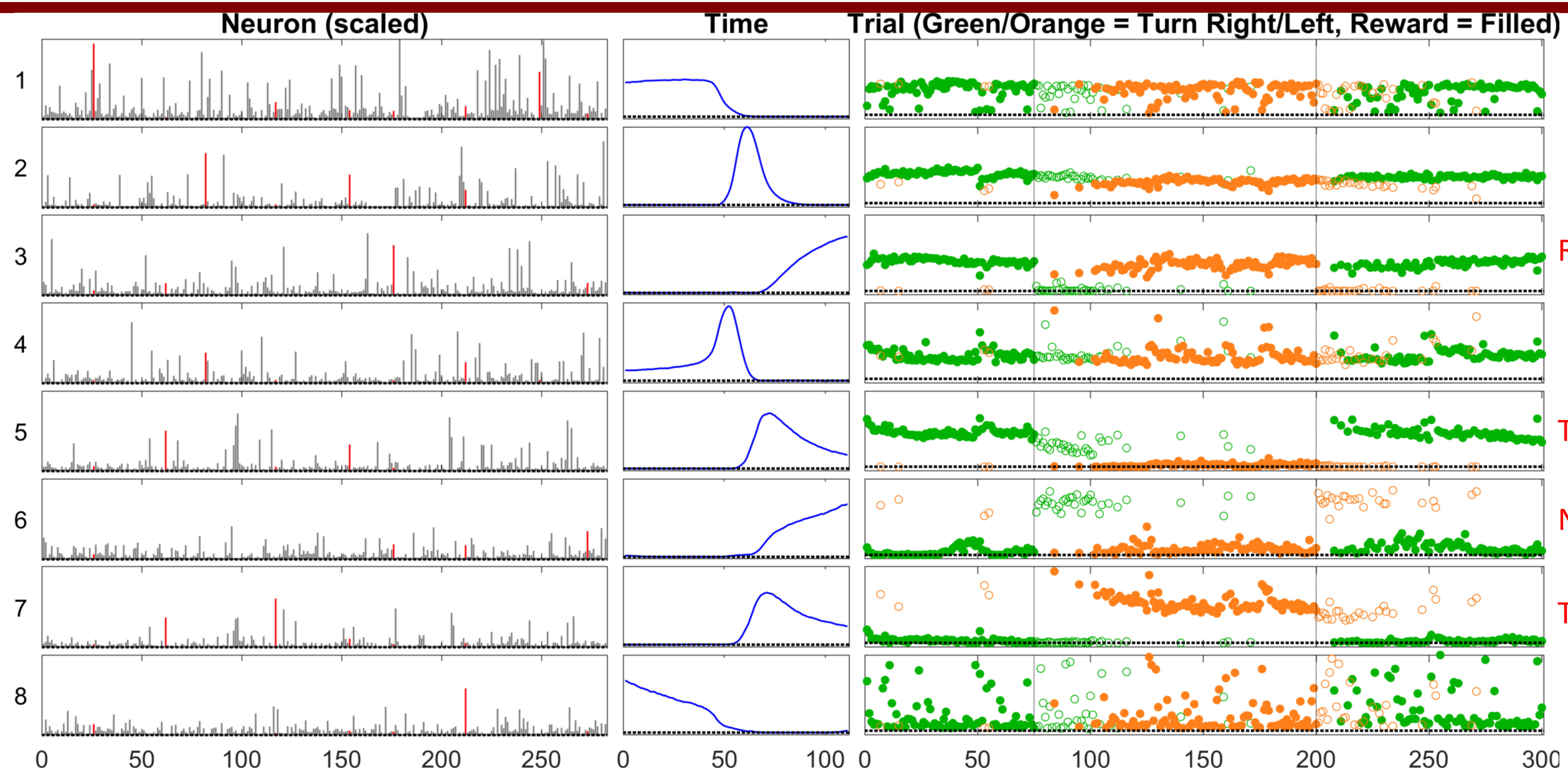
CP Tensor Decomposition "Sees" Turn Direction



CP Tensor Decomposition Can be Tough to Interpret due to Negative Entries



GCP Decomposition with Beta Divergence ($\beta = 0.5, f(x,m) = \sqrt{m} + x/\sqrt{m}$)



Reward!

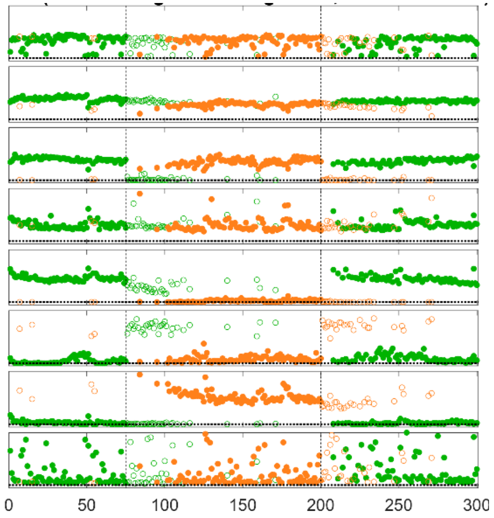
Turn right

No reward!

Turn left

Regression Using GCP Factors on Trial Mode

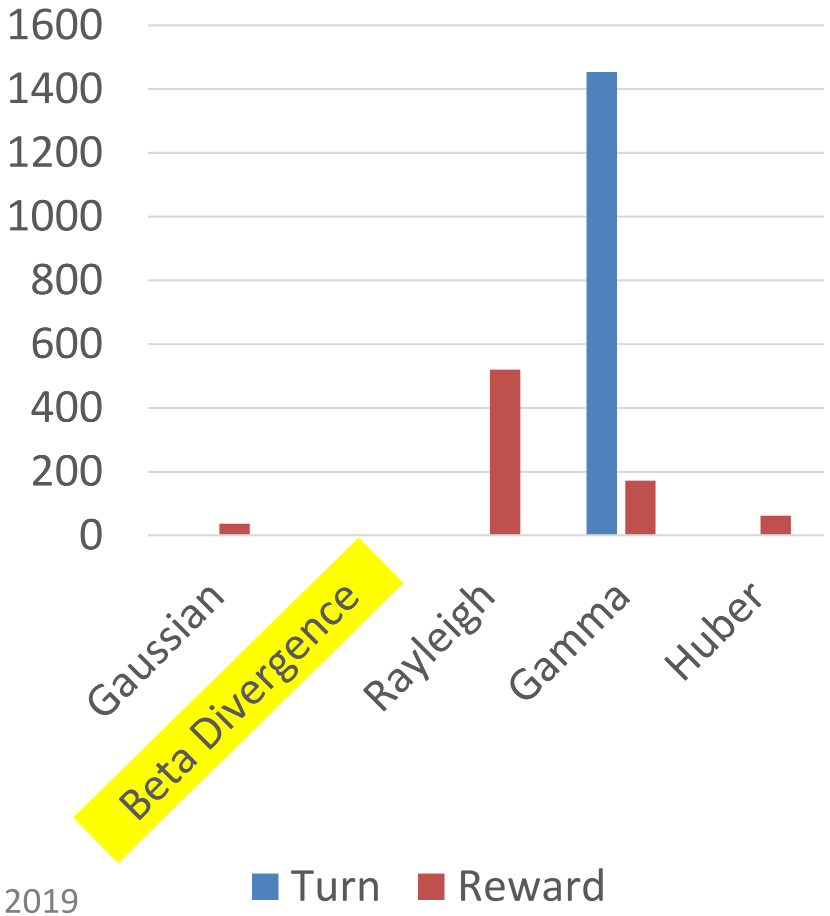
Trial Factor Matrix is 300×8



$$\min_{\boldsymbol{\beta}} \|\mathbf{A}_3^{\text{train}}\boldsymbol{\beta} - \mathbf{y}^{\text{train}}\|$$

$$\hat{\mathbf{y}}^{\text{test}} = [\mathbf{A}_3^{\text{test}}\boldsymbol{\beta} \geq 0.5]$$

Look at predicting turn and reward.
Split into two groups of 150 trials.
Train regression model with 1st group.
Test with 2nd group.
Repeat 100 times.



Hong, Kolda, Duersch, SIAM Review, 2019

Optimization Formulation for GCP Tensor Decomposition

GCP

$$\min \; F(\mathcal{X},\mathcal{M}) \equiv \sum_{i\in\Omega} f(x_i, m_i)$$

$$\text{s.t.} \;\; \text{rank}(\mathcal{M}) \leq r$$

i = multi-index
Ω = all indices

- Standard CP [Hitchcock, 1927; Carrol & Chang, 1970; Harshman, 1970]

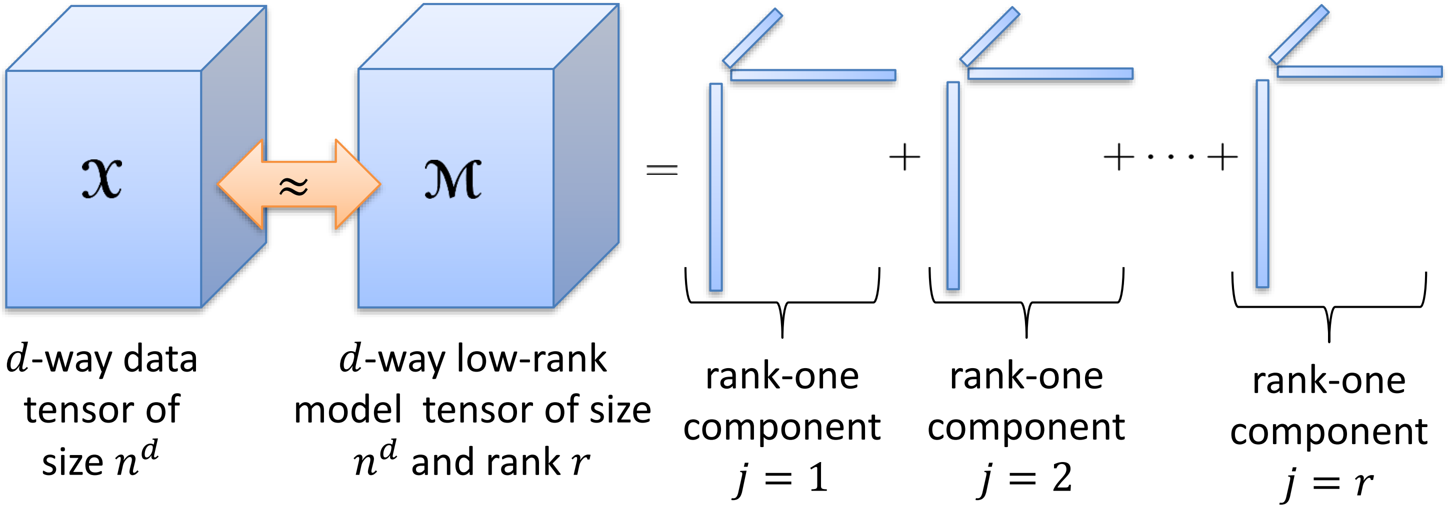
  $$f(x,m) = (x - m)^2$$

- Poisson CP (Identity Link) [Welling & Webber, 2001; Chi & Kolda, 2009]

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

- Logistic CP, etc. [Hong, Kolda, Duersch, 2018]

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



$$\mathcal{X} \approx \mathcal{M} \quad \text{where} \quad \mathcal{M} = \sum_{j=1}^{r} \mathbf{A}_1(:,j) \circ \mathbf{A}_2(:,j) \circ \cdots \circ \mathbf{A}_d(:,j)$$

Low-rank: $\text{rank}(\mathcal{M}) \leq r \ll n^d$

Factor matrices: $\mathbf{A}_k \in \mathbb{R}^{n_k \times r}$ for $k \in \{1,\ldots,d\}$

WLOG, $n = n_1 = \cdots = n_d$

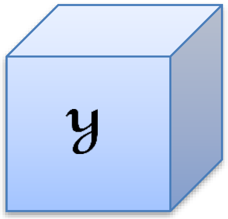
Gradient-based Optimization for Fitting the GCP Model

GCP

$$\min \; F(\mathcal{X}, \mathcal{M}) \equiv \sum_{i \in \Omega} f(x_i, m_i)$$

$$\text{s.t.} \;\; \text{rank}(\mathcal{M}) \leq r$$

<u>Define</u>: Elementwise partial gradient tensor, same size as data tensor = n^d



$$y_i = \frac{\partial f}{\partial m}(x_i, m_i)$$

<u>Define</u>: Khatri-Rao product in all modes but one of size $n^{d-1} \times r$

$$\mathbf{Z}_k = \mathbf{A}_d \odot \cdots \odot \mathbf{A}_{k+1} \odot \mathbf{A}_{k-1} \odot \cdots \odot \mathbf{A}_1$$

Gradients computed via a sequence of matricized-tensor times Khatri-Rao product (MTTKRPs):

$$\mathbf{G}_k \equiv \frac{\partial F}{\partial \mathbf{A}_k} = \mathbf{Y}_{(k)} \mathbf{Z}_k \text{ for } k = 1, \ldots, d$$

MTTKRP

gradient for mode k factor matrix of size $n \times r$

tensor unfolded in mode k into matrix of size $n \times n^{d-1}$

MTTKRPs can be computed efficiently…

- Bader & Kolda, SISC, 2007 – Dense and sparse
- Phan, Tichavsky, Cichocki, 2013 – Sequence
- Smith et al., IPDPS 2015 – Sparse
- Kaya & Ucar, SC 2015 – Sparse
- Li et al., IPDPS 2017 – Sparse
- Hayashi et al., 2017 – Dense
- Ballard, Knight, Rouse, 2017 – Dense

Stochastic Gradient Descent (SGD) for GCP

30-Second Tutorial on SGD

$$\min F(x)$$

Gradient Descent (GD)

α = learning rate

$$x^{(t+1)} = x^{(t)} - \alpha g^{(t)}$$

Stochastic Gradient Descent (SGD)

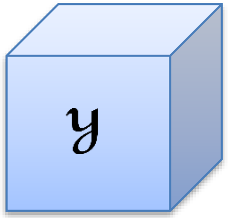
$$x^{(t+1)} = x^{(t)} - \alpha \tilde{g}^{(t)}$$

$$\mathbb{E}[\tilde{g}^{(t)}] = g^{(t)} \equiv \nabla F(x^{(t)})$$

Adam (Kingma & Ba, 2015)

Adaptive momentum SGD

Standard gradient

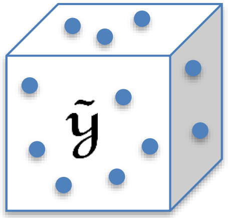


$$\mathbf{G}_k = \mathbf{Y}_{(k)} \mathbf{Z}_k$$

Cost: $O(rn^d)$ flops

$$y_i = \frac{\partial f}{\partial m}(x_i, m_i)$$

Stochastic gradient



$$\tilde{\mathbf{G}}_k = \tilde{\mathbf{Y}}_{(k)} \mathbf{Z}_k$$

Cost: $O(rs)$ flops

Choose stochastic *sparse* Y-tensor

$$\mathbb{E}[\tilde{\mathcal{Y}}] = \mathcal{Y}$$

such that

$$\text{nnz}(\tilde{\mathcal{Y}}) \leq s \ll n^d$$

By linearity of expectation: $\mathbb{E}[\tilde{\mathbf{G}}_k] = \mathbf{G}_k$

Uniform Sampling

Goal: Random ***sparse*** tensor of size n^d that equals the "Y-tensor" in expectation

Sample $s \ll n^d$ random tensor entries (with replacement)

$$\tilde{s}_i = \#\text{ times } i \text{ sampled}$$

$$\tilde{y}_i = \tilde{s}_i \cdot \frac{n^d}{s} \cdot y_i$$

$$\sum_{i \in \Omega} \tilde{s}_i = s$$

$$y_i = \frac{\partial f}{\partial m}(x_i, m_i)$$

Theory

Claim: $\mathbb{E}[\tilde{\mathcal{Y}}] = \mathcal{Y}$

Proof: $\mathbb{E}[\tilde{s}_i] = \dfrac{s}{n^d}$

$$\mathbb{E}[\tilde{y}_i] = \mathbb{E}[\tilde{s}_i] \cdot \frac{n^d}{s} \cdot y_i = y_i$$

Choosing s, the number of sampled elements…

- Choose $s = O(n)$
- Gradient = $O(rs) = O(rn)$ versus $O(rn^d)$

Downside…

- If data tensor is sparse, few entries corresponding to nonzeros will be chosen

Stratified 0/1 Sampling

Goal: Random ***sparse*** tensor of size n^d that equals the "Y-tensor" in expectation

Sample p nonzeros and q zeros.

$$\tilde{p}_i = \#\text{ times nonzero } i \text{ sampled} \qquad \eta = \#\text{ nonzeros}$$

$$\tilde{q}_i = \#\text{ times zero } i \text{ sampled} \qquad \zeta = \#\text{ zeros}$$

$$\tilde{y}_i = \left(\tilde{p}_i \cdot \frac{\eta}{p} + \tilde{q}_i \cdot \frac{\zeta}{q} \right) \cdot y_i \qquad y_i = \frac{\partial f}{\partial m}(x_i, m_i)$$

Theory

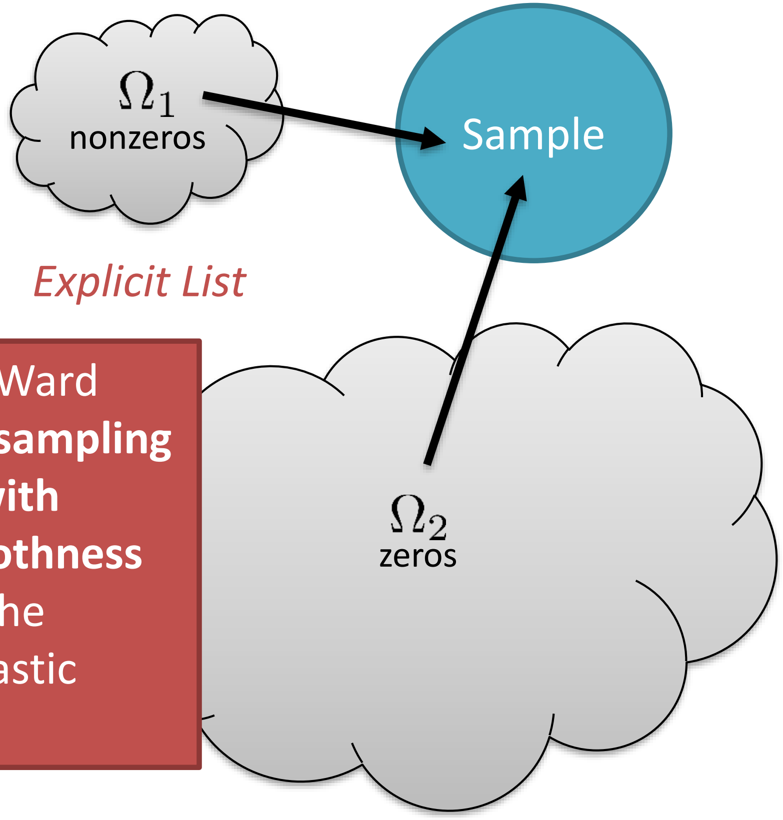
Claim: $\mathbb{E}[\tilde{\mathcal{Y}}] = \mathcal{Y}$

Proof: $\mathbb{E}[\tilde{p}_i] = \frac{p}{\eta}, \; \mathbb{E}[\tilde{q}_i] = \frac{q}{\zeta}$

$$x_i = 1 \Rightarrow \mathbb{E}[\tilde{y}_i] = \mathbb{E}[\tilde{p}_i] \cdot \frac{\eta}{p} \cdot y_i = y_i$$

$$x_i = 0 \Rightarrow \mathbb{E}[\tilde{y}_i] = \mathbb{E}[\tilde{q}_i] \cdot \frac{\zeta}{q} \cdot y_i = y_i$$

Needell, Srebro, and Ward (2013) justify **biased sampling toward functionals with higher Lipschitz smoothness** constants to reduce the variance in the stochastic gradient.



Implicit List (Requires Rejection Sampling)

Semi-Stratified 0/1 Sampling

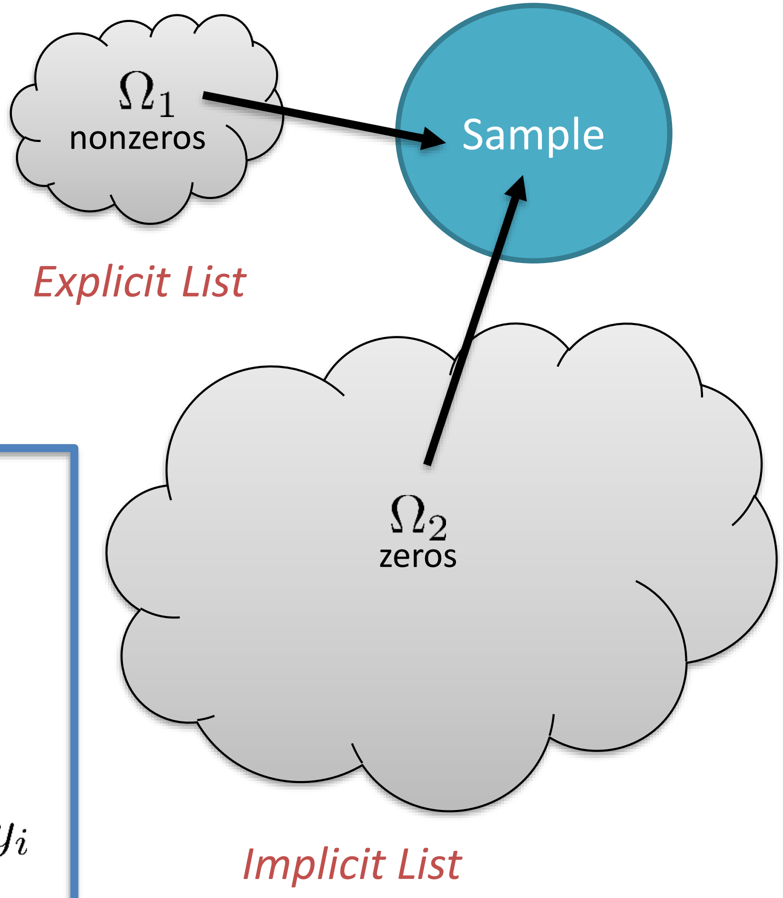
Goal: Random ***sparse*** tensor of size n^d that equals the "Y-tensor" in expectation

Sample p nonzeros and q assumed zeros.

$$\tilde{p}_i = \#\text{ times nonzero } i \text{ sampled} \qquad \eta = \#\text{ nonzeros}$$
$$\tilde{q}_i = \#\text{ times ``zero'' } i \text{ sampled} \qquad \zeta = \#\text{ zeros}$$

$$\tilde{y}_i = \tilde{p}_i \cdot \frac{\eta}{p} \cdot (y_i - c_i) + \tilde{q}_i \cdot \frac{(\eta+\zeta)}{q} \cdot c_i \text{ with } c_i \equiv \frac{\partial f}{\partial m}(0, m_i)$$

$$y_i = \frac{\partial f}{\partial m}(x_i, m_i)$$



Theory

Claim: $\mathbb{E}[\tilde{\mathfrak{Y}}] = \mathfrak{Y}$

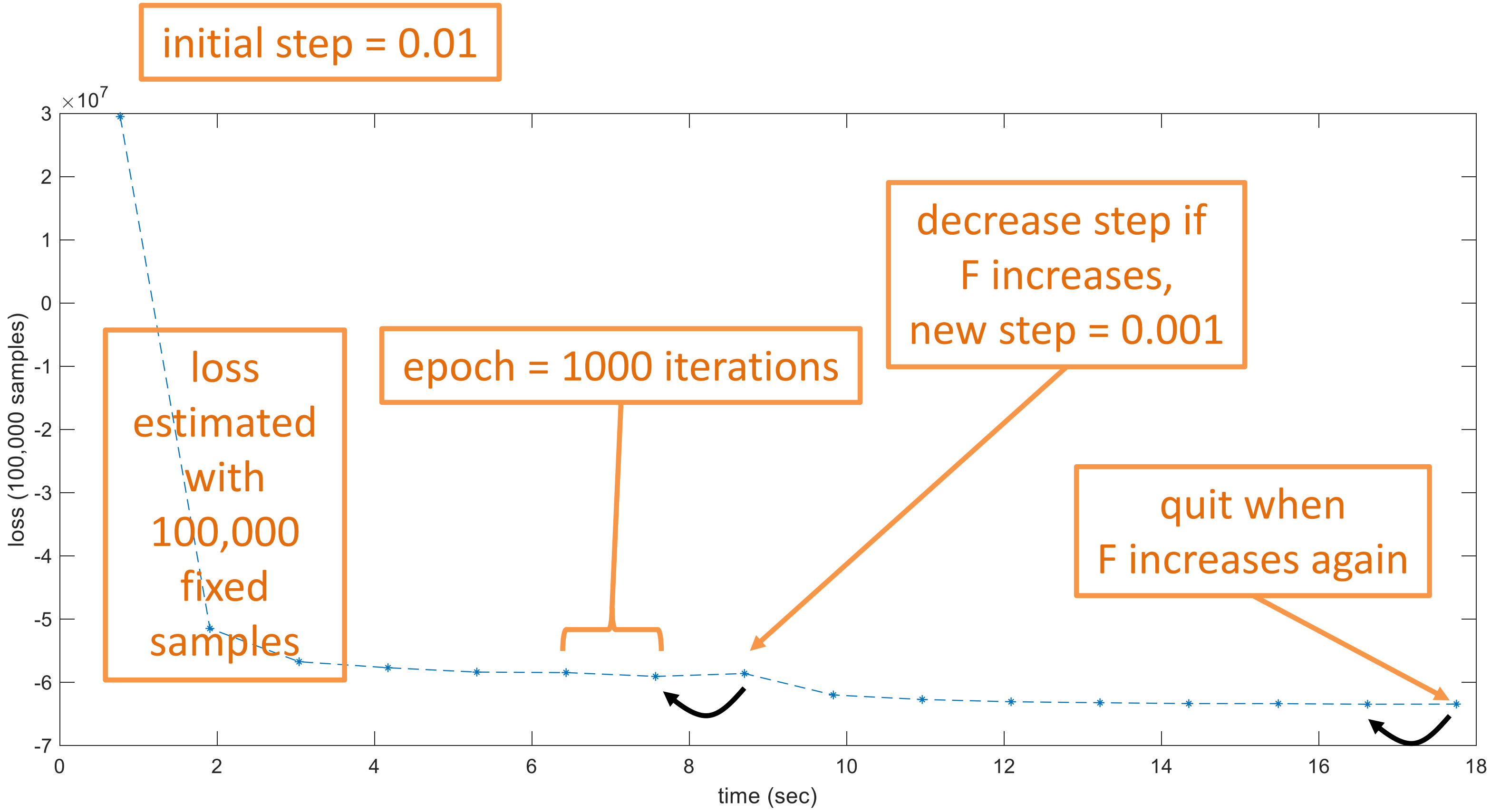
Proof: $\mathbb{E}[\tilde{p}_i] = \frac{p}{\eta}$, $\mathbb{E}[\tilde{q}_i] = \frac{q}{(\zeta+\eta)}$

$$x_i = 0 \Rightarrow \mathbb{E}[\tilde{y}_i] = \mathbb{E}[\tilde{q}_i] \cdot \frac{(\eta+\zeta)}{q} \cdot y_i = y_i$$

$$x_i = 1 \Rightarrow \mathbb{E}[\tilde{y}_i] = \mathbb{E}[\tilde{p}_i] \cdot \frac{\eta}{p} \cdot (y_i - c_i) + \mathbb{E}[\tilde{q}_i] \cdot \frac{\eta+\zeta}{q} \cdot c_i = y_i$$

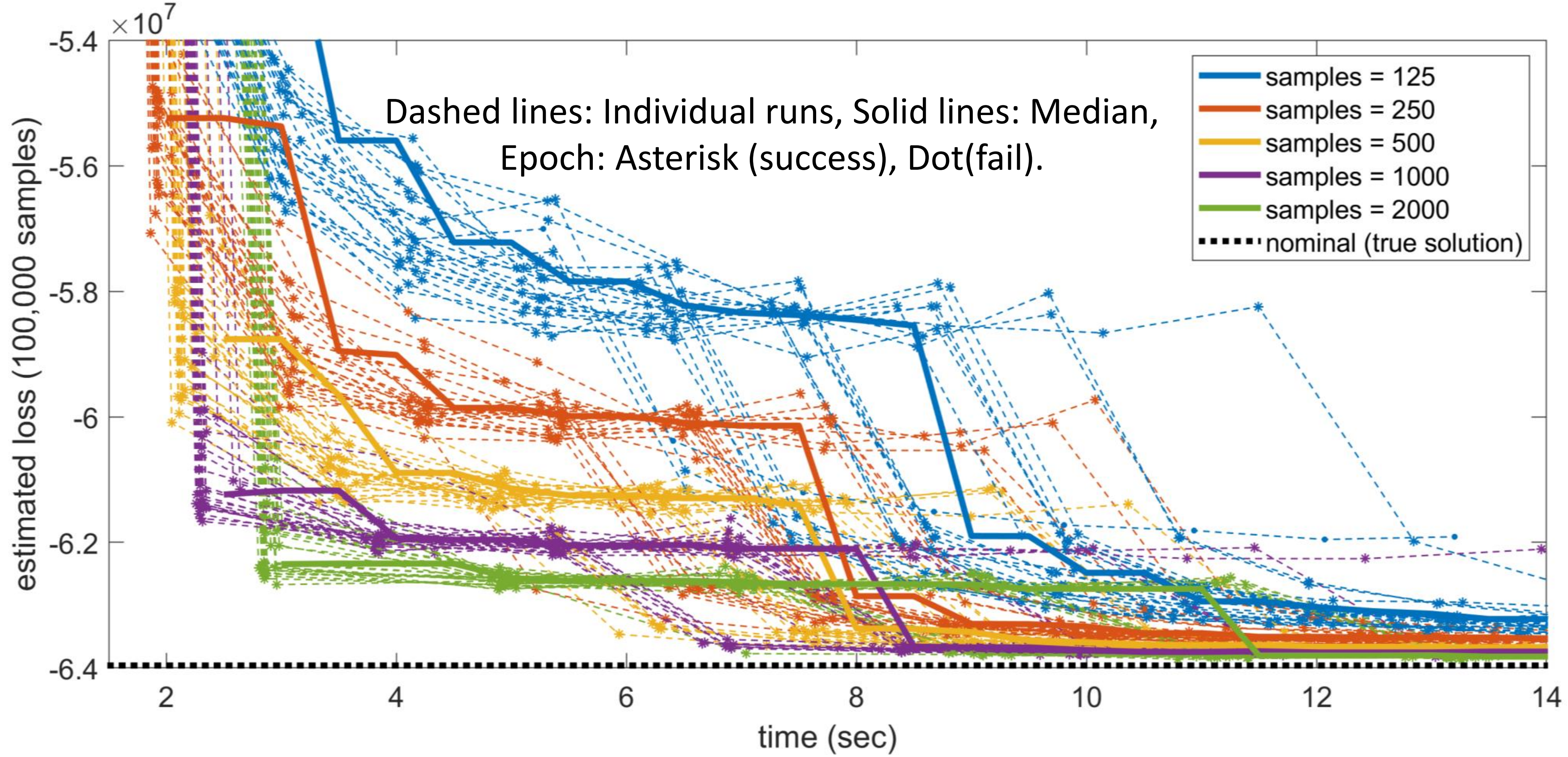
GCP with Stochastic Optimization

- Nonconvex problem
  - No guarantees of finding minimizer
- Using Adam (Kingma & Ba, 2015)
  - Default parameters
  - Some tweaks for checking convergence
- Past work on recommender systems uses SGD but ignores zeros
  - Gemulla, Nijkamp, Hass, Sismanis, KDD'11
  - Zhuang, Chin, Juan, and Lin, RecSys'13
- Past work on streaming uses SGD but data appears one slice at a time
  - Mardani, Mateos, Giannakis, IEEE TSP 2015
  - Maehara, Hayashi, Kawarabayashi,

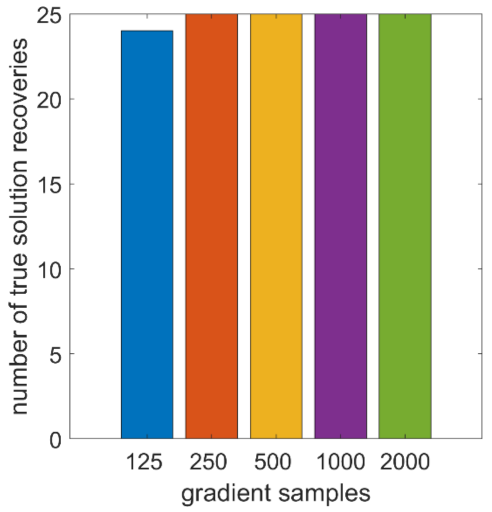


Example on Gamma-Distributed Data

$200 \times 150 \times 100 \times 50$ Tensor with low-rank ($r = 5$) structure based on Gamma distribution ($k = 1$, θ from model). Gamma loss: $f(x, m) = \frac{x}{m} + \log m$. Running stochastic GCP with 25 random starts and varying numbers of samples.



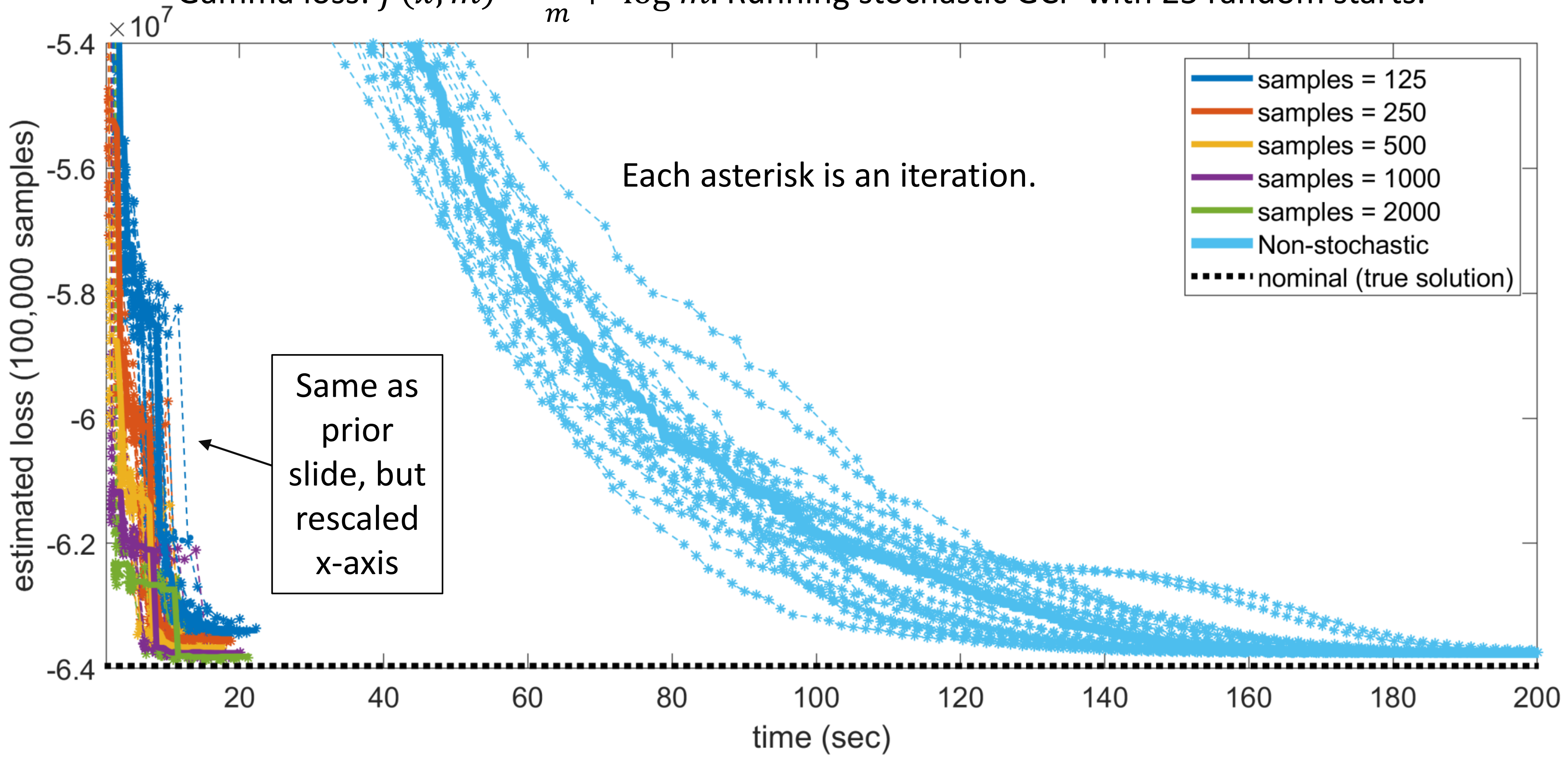
Success at Recovering Underlying Generative Factors



Stochastic vs. Non-Stochastic

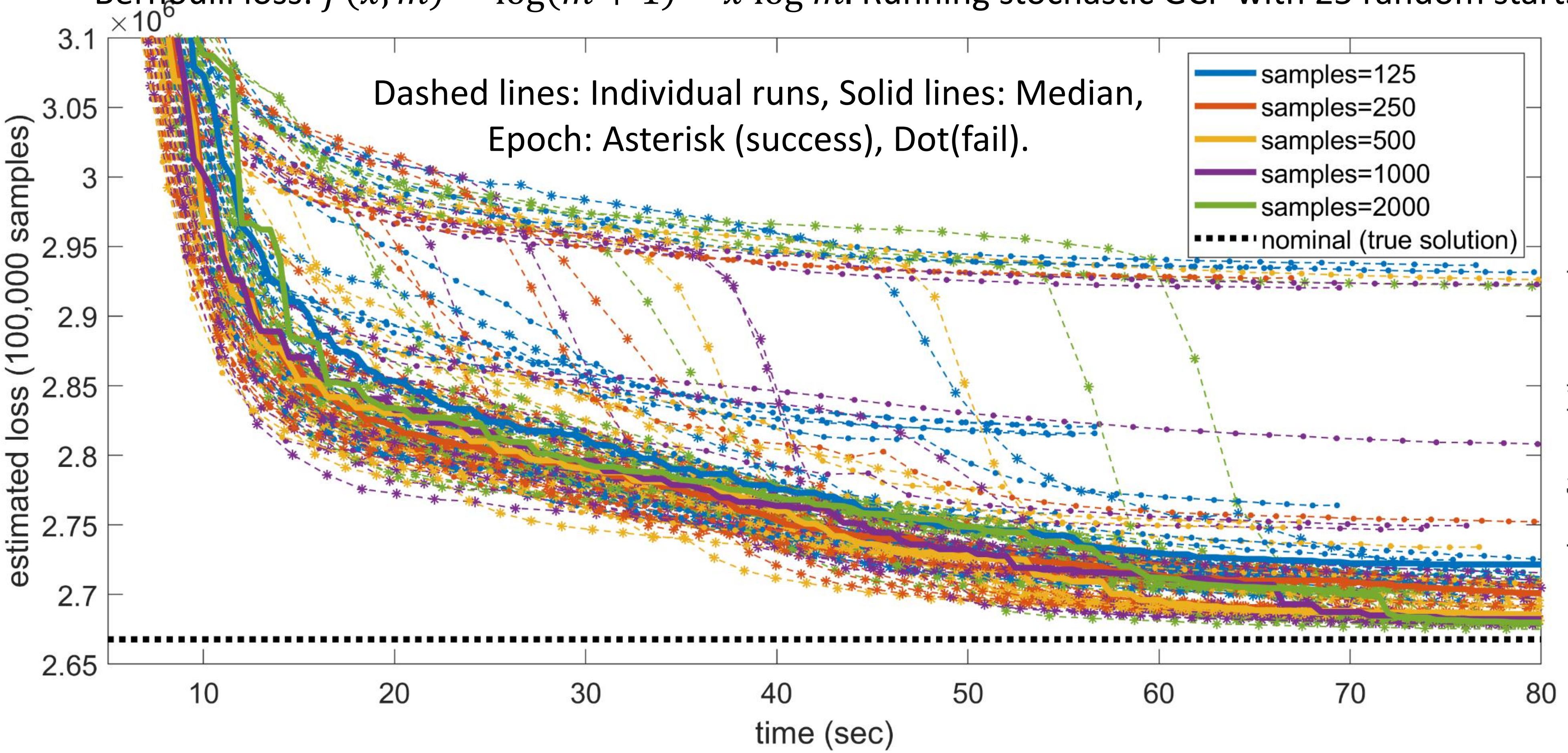
$200 \times 150 \times 100 \times 50$ Tensor with low-rank ($r = 5$) structure based on Gamma distribution ($k = 1$, θ from model).

Gamma loss: $f(x,m) = \frac{x}{m} + \log m$. Running stochastic GCP with 25 random starts.

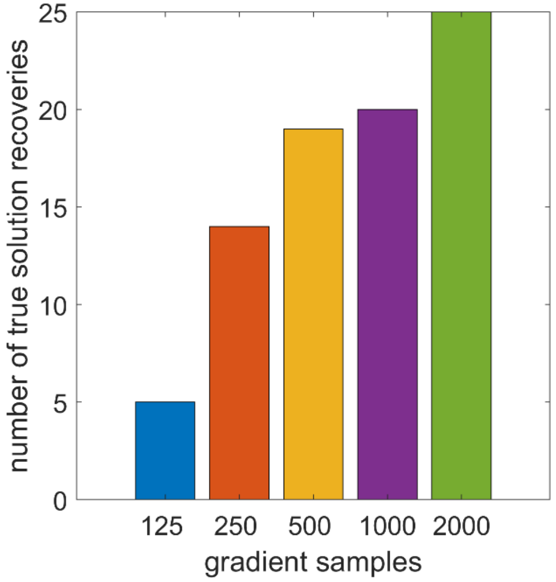


Example on Bernoulli-Distributed Data

$200 \times 150 \times 100 \times 50$ Tensor with low-rank ($r = 5$) structure based on Bernoulli distribution (odds from model).
Sparse tensor, less than 0.35% dense (~500K nonzeros).
Bernoulli loss: $f(x, m) = \log(m+1) - x \log m$. Running stochastic GCP with 25 random starts, varying # of samples.

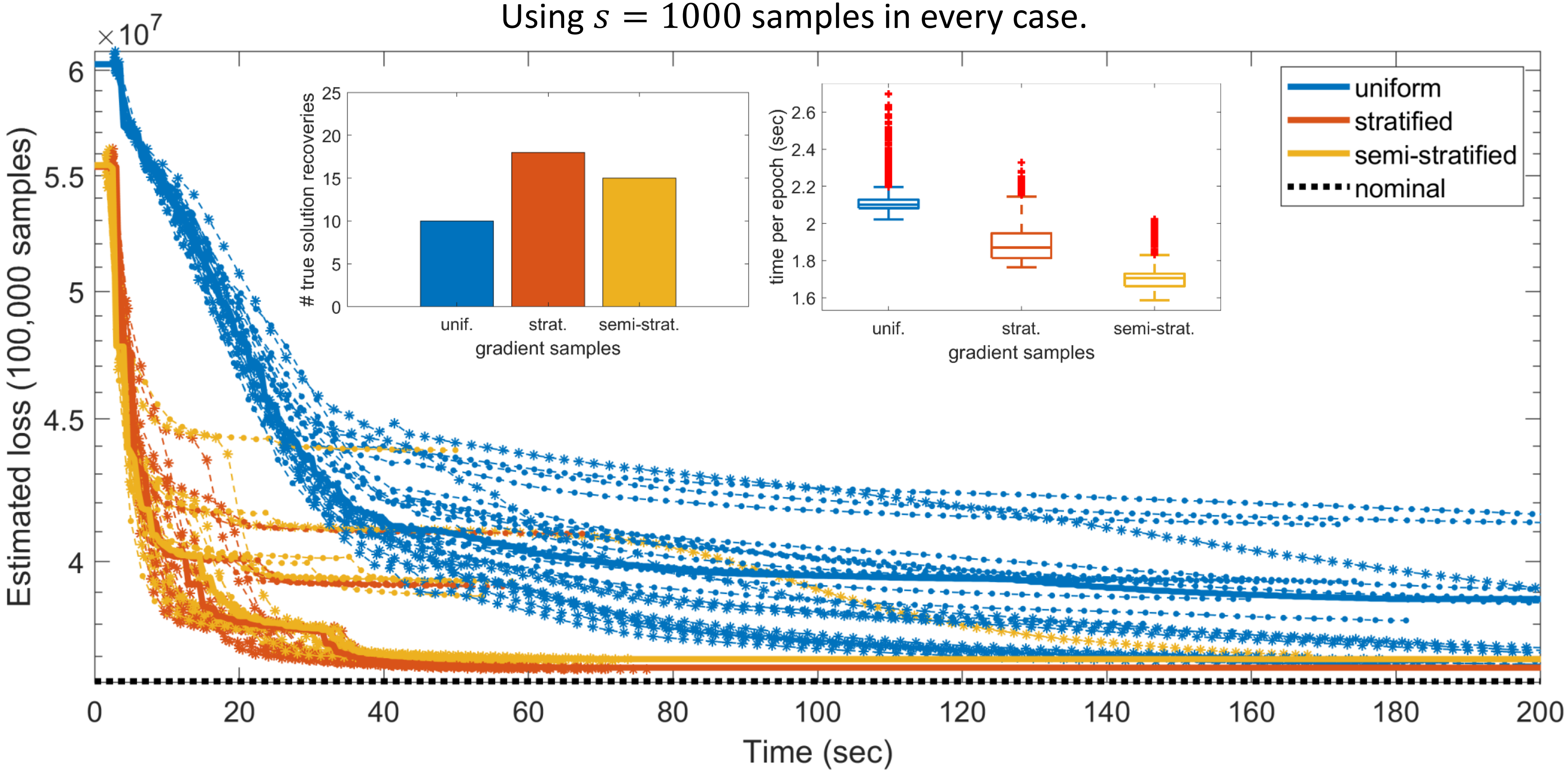


Success at Recovering Underlying Generative Factors



Uniform Sampling is Worse than Stratified for Sparse Tensors

Same set-up as binary experiments, but bigger tensor: $400 \times 300 \times 200 \times 100$, 0.38% dense (9M nonzeroes).
Using $s = 1000$ samples in every case.



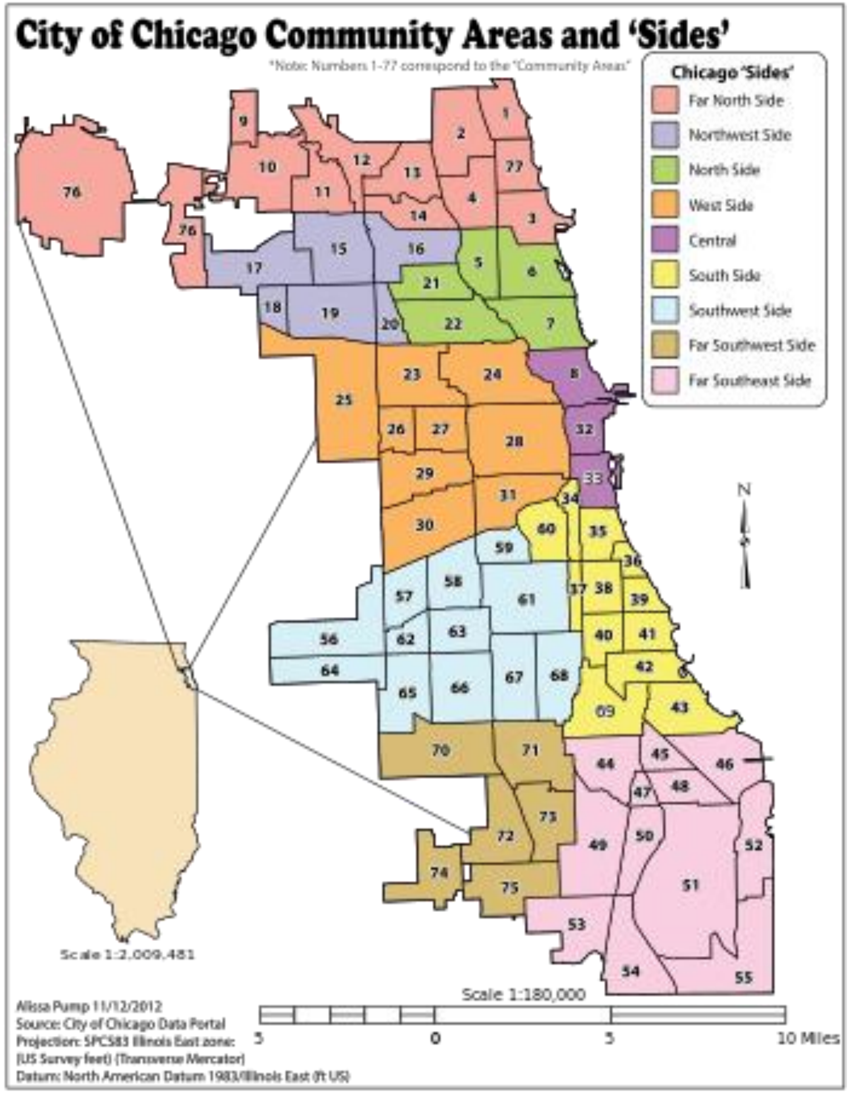
Chicago Crime Data

Sandia National Laboratories

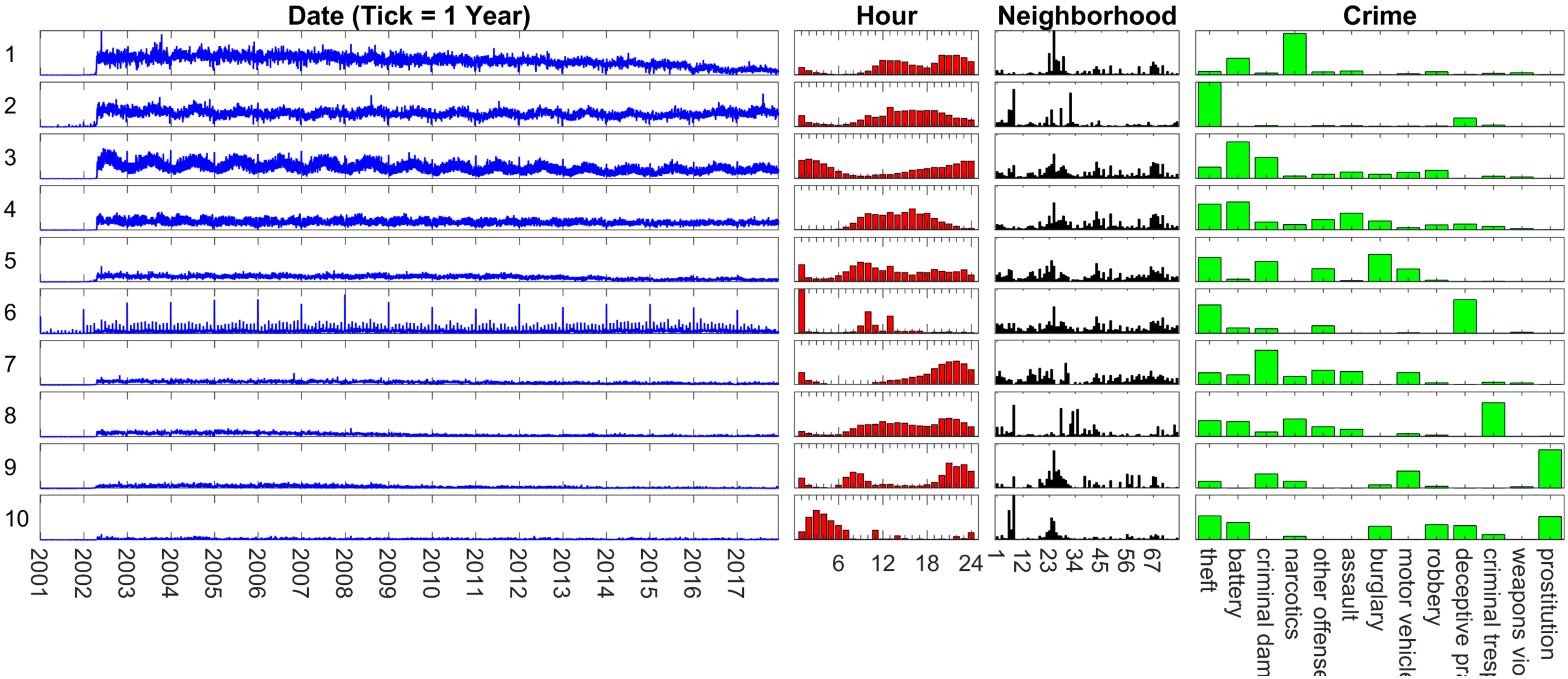
- 4-way count tensor
  - 6,186 Days
  - 24 Hours of the Day
  - 77 Community Areas
  - 32 Crime Types
- Non-zeros: 5,330,673
  - Storage: 0.21GB for sparse tensor
- Distribution of entries
  - 0: 98.54%
  - 1: 1.33%
  - ≥ 2: 0.12%
- Obtained from FROSTT (http://frostt.io/tensors/chicago-crime/)
- Data originally from Chicago Data Portal (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2)

GCP-Count
Rank = 10
$s = 6{,}319$

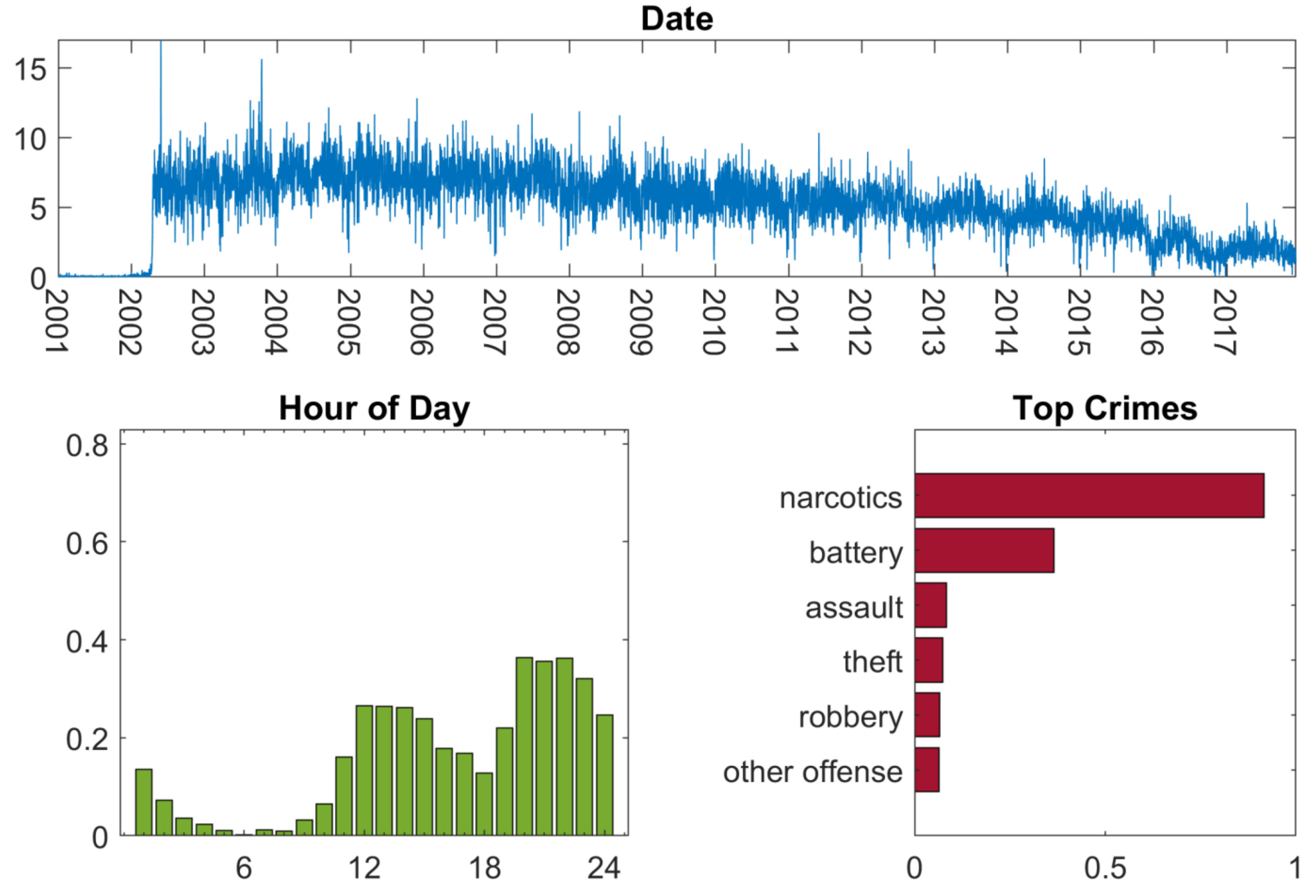
$$f(x,m) = m - x \log m$$



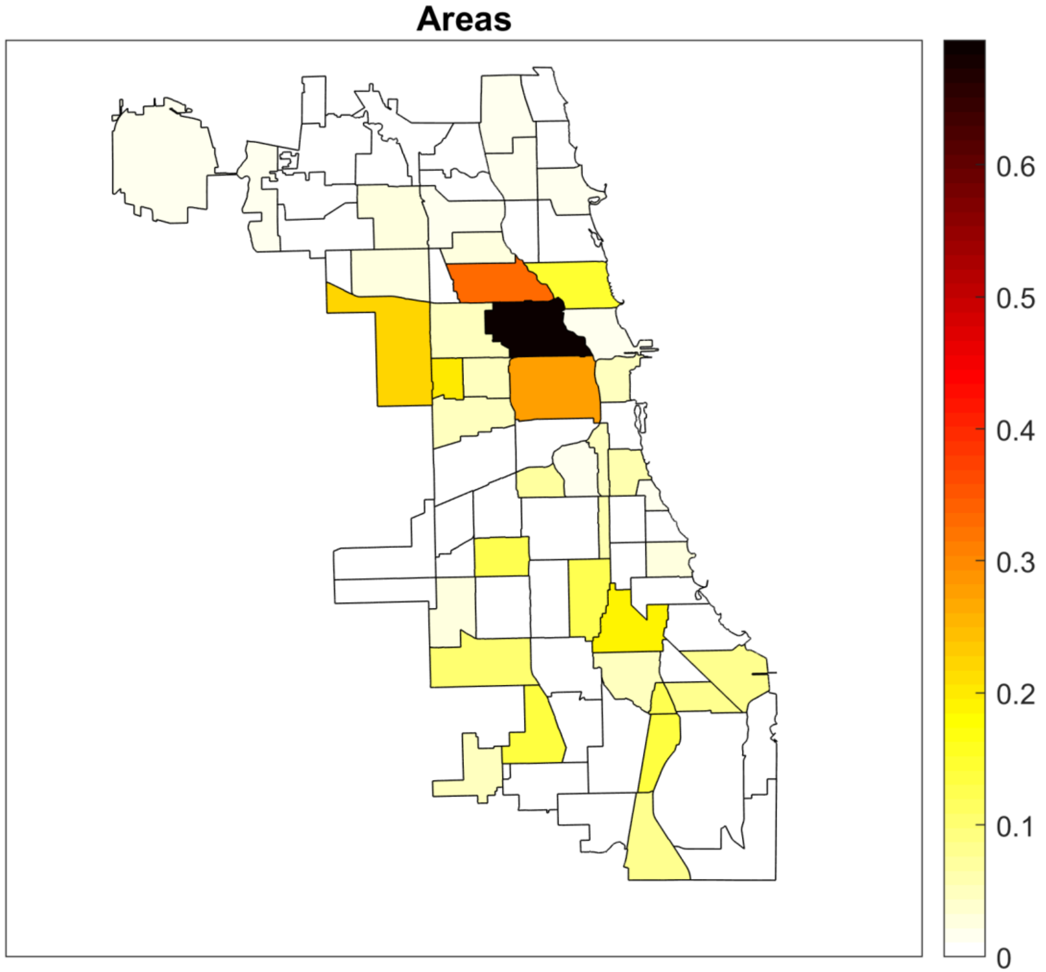
Application to Sparse Crime Binary Tensor (Semi-stratified results)



Component #1

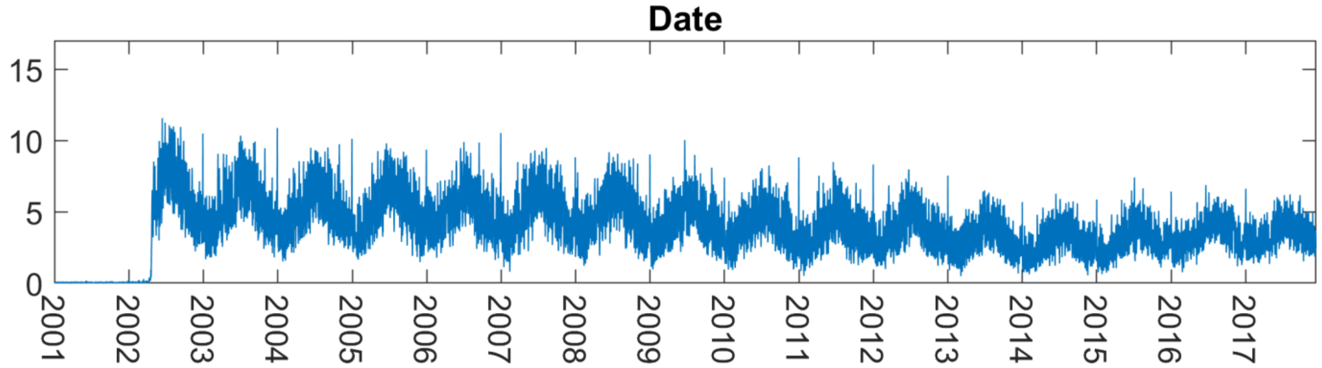


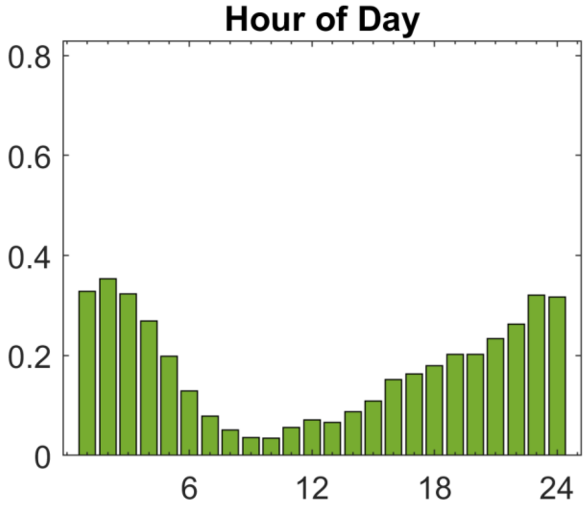
Date
15
10
5
0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Hour of Day
0.8
0.6
0.4
0.2
0
6
12
18
24
Top Crimes
narcotics
battery
assault
theft
robbery
other offense
0
0.5
1

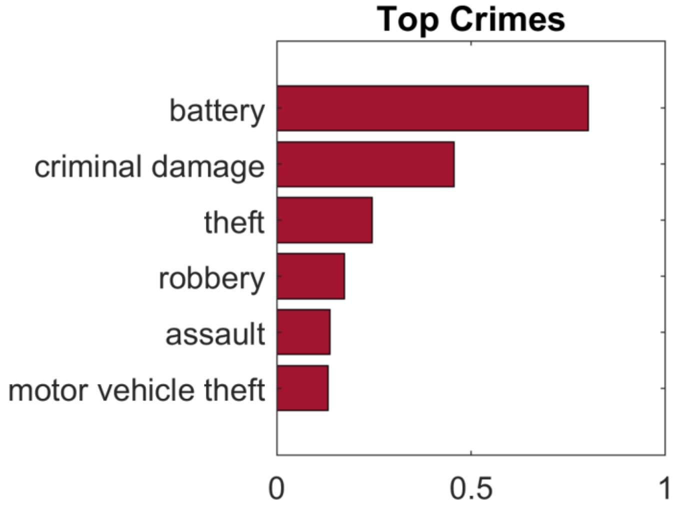


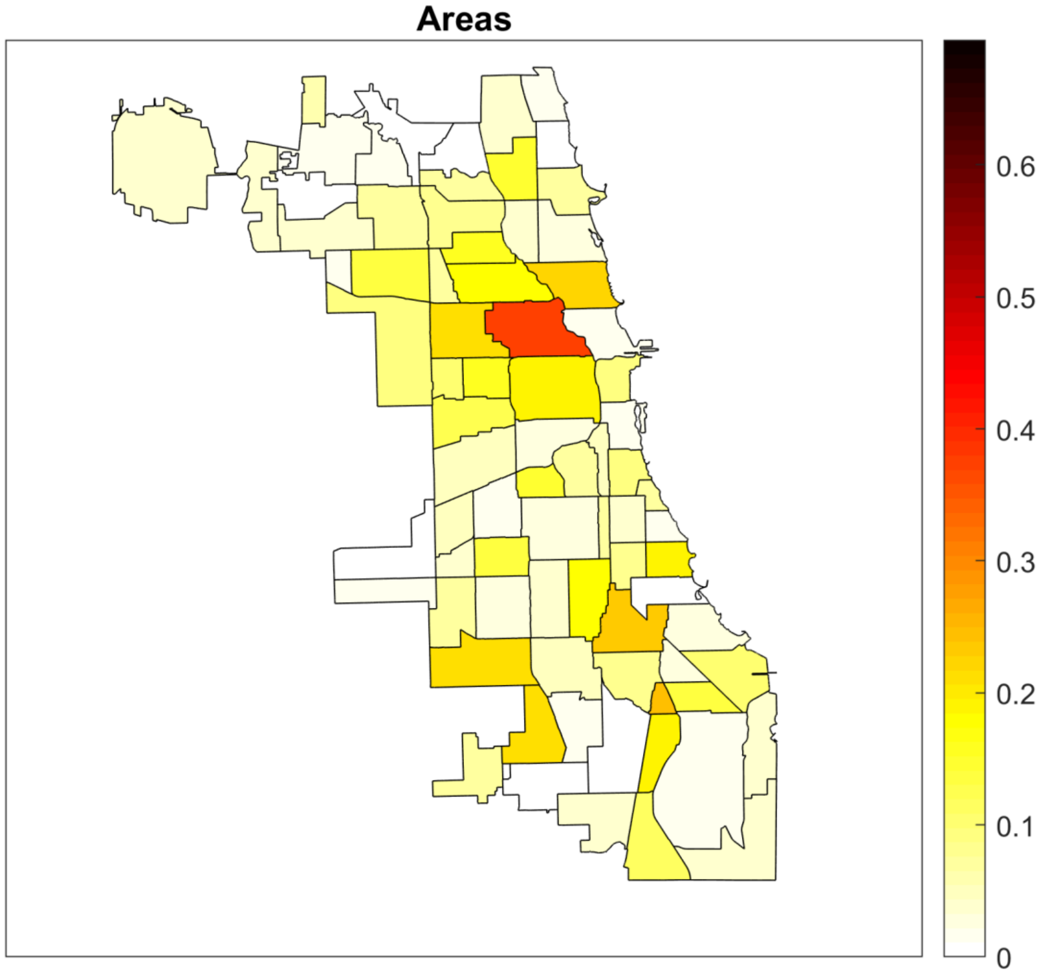
Areas
0.6
0.5
0.4
0.3
0.2
0.1
0

Component #3

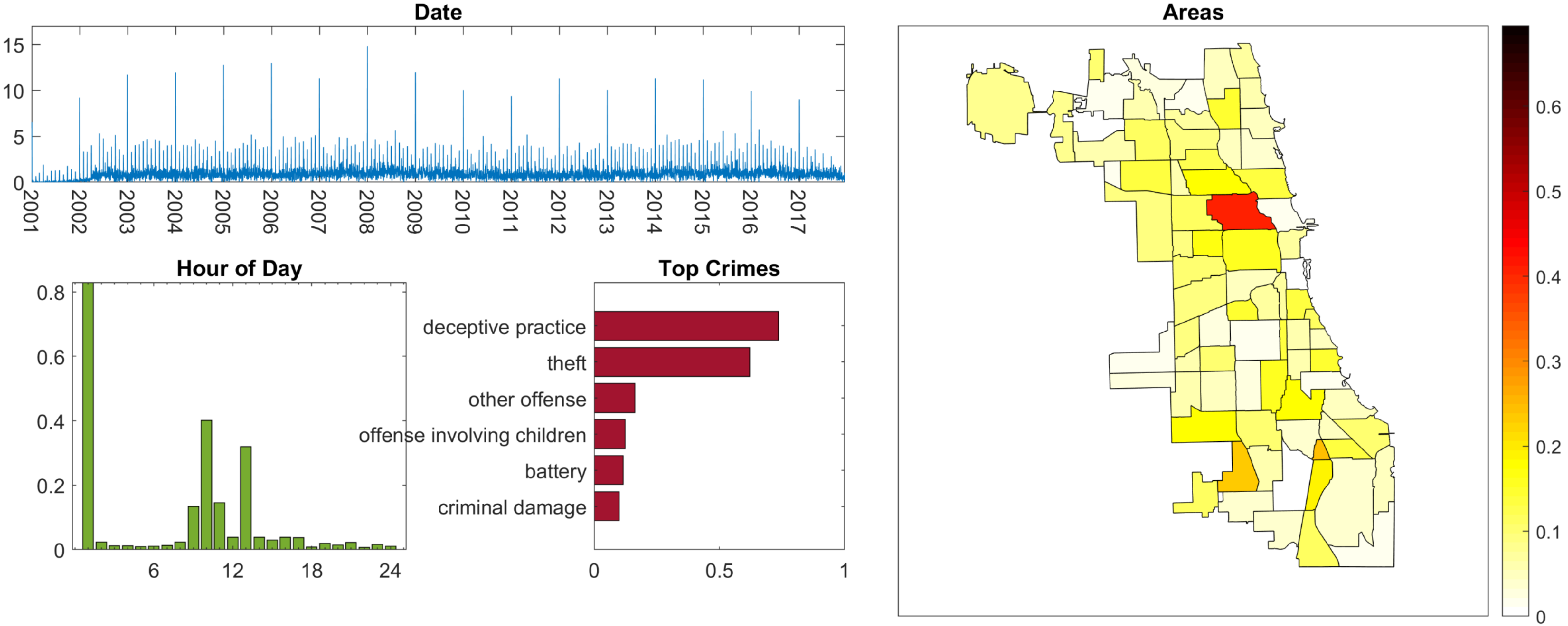

Date
15
10
5
0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017


Hour of Day
0.8
0.6
0.4
0.2
0
6
12
18
24


Top Crimes
battery
criminal damage
theft
robbery
assault
motor vehicle theft
0
0.5
1


Areas
0.6
0.5
0.4
0.3
0.2
0.1
0

Component #6


Date
15
10
5
0
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Hour of Day
0.8
0.6
0.4
0.2
0
6
12
18
24
Top Crimes
deceptive practice
theft
other offense
offense involving children
battery
criminal damage
0
0.5
1
Areas
0.6
0.5
0.4
0.3
0.2
0.1
0

Aside: Estimating Higher-Order Moments via Symmetric Tensor Factorization

Joint work with Sam Sherman, Notre Dame

Given a set of p observations: $\mathbf{a}_i \in \mathbb{R}^n, i = 1, 2, \ldots, p$

First-order moment (mean): $$\frac{1}{p}\sum_{i=1}^{p} \mathbf{a}_i$$

Second-order moment: $$\frac{1}{p}\sum_{i=1}^{p} \mathbf{a}_i \circ \mathbf{a}_i$$

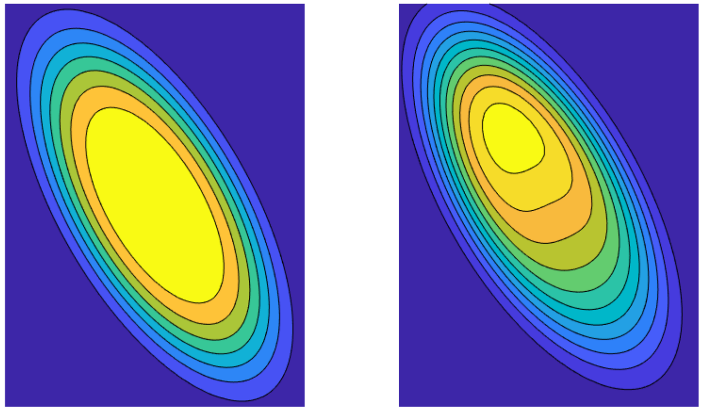
Third-order moment: $$\frac{1}{p}\sum_{i=1}^{p} \mathbf{a}_i \circ \mathbf{a}_i \circ \mathbf{a}_i$$

Fourth-order moment: $$\frac{1}{p}\sum_{i=1}^{p} \mathbf{a}_i \circ \mathbf{a}_i \circ \mathbf{a}_i \circ \mathbf{a}_i$$

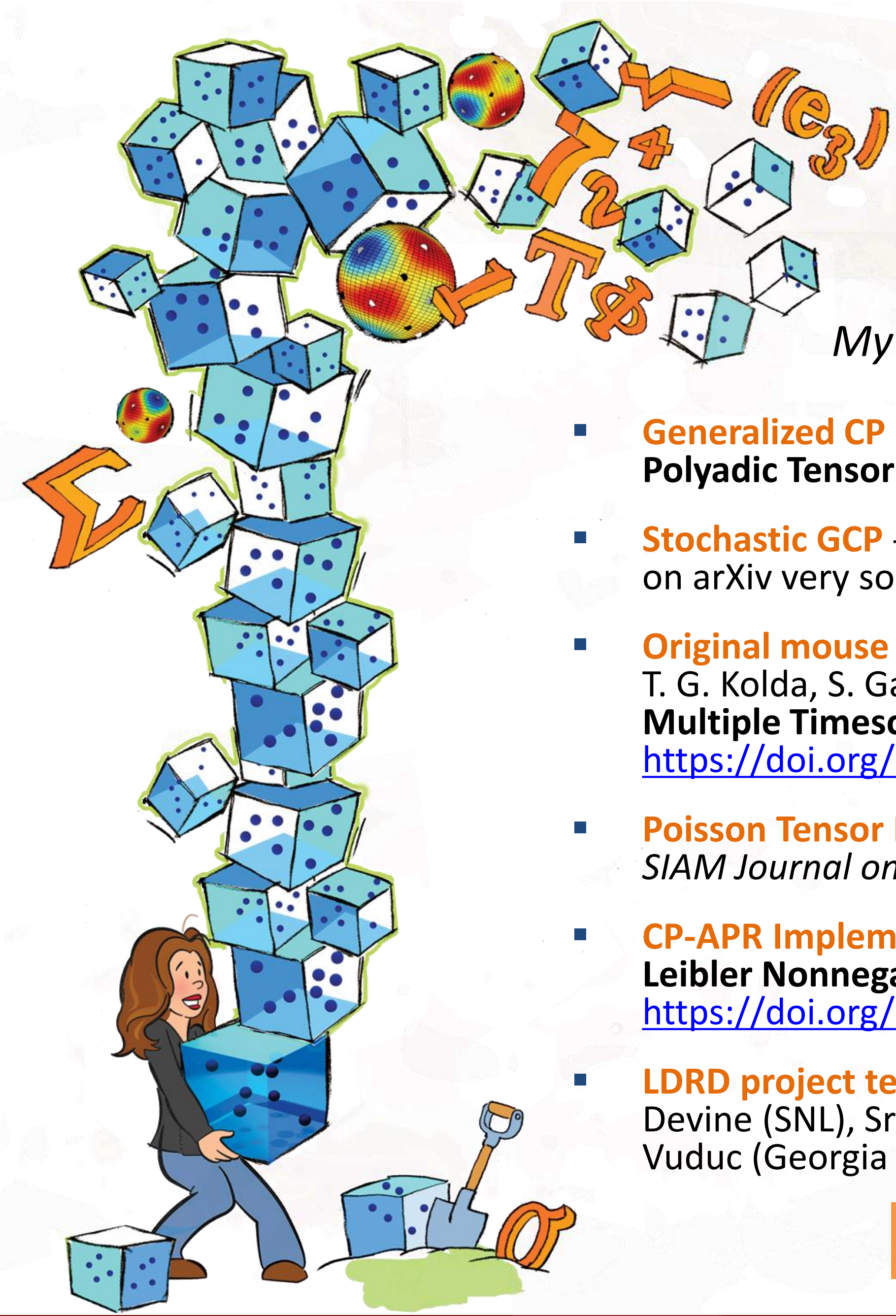
We can compute low-rank ($r \ll p$) symmetric tensor estimated to higher-order moments…

$$\frac{1}{r}\sum_{i=1}^{r} \mathbf{b}_i \circ \mathbf{b}_i \circ \mathbf{b}_i$$

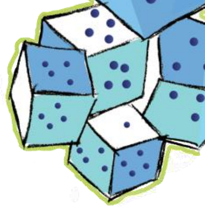
$$\frac{1}{r}\sum_{i=1}^{r} \mathbf{c}_i \circ \mathbf{c}_i \circ \mathbf{c}_i \circ \mathbf{c}_i$$



What are good applications, if any?



Sandia National Laboratories



References & Collaborators

My department is hiring statisticians! Talk to me to learn more.

- **Generalized CP (GCP) Tensor Decomposition** - D. Hong, T. G. Kolda, J. A. Duersch. **Generalized Canonical Polyadic Tensor Decomposition**, SIAM Review (to appear). http://arxiv.org/abs/1808.07452
- **Stochastic GCP** - D. Hong. T. G. Kolda. **Stochastic Gradients for Large-Scale Tensor Decomposition**, to appear on arXiv very soon!
- **Original mouse experiments** - 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**. *Neuron*, 98(6), 2018. https://doi.org/10.1016/j.neuron.2018.05.015
- **Poisson Tensor Factorization** - E. C. Chi, T. G. Kolda. **On Tensors, Sparsity, and Nonnegative Factorizations**. *SIAM Journal on Matrix Analysis and Applications*, 33(4), 2013. https://doi.org/10.1137/110859063
- **CP-APR Implementation** - S. Hansen, T. Plantenga, T. G. Kolda. **Newton-Based Optimization for Kullback-Leibler Nonnegative Tensor Factorizations**. *Optimization Methods and Software*, 30(5), 2015. https://doi.org/10.1080/10556788.2015.1009977
- **LDRD project team** - Cliff Anderson-Bergman (LLNL), Grey Ballard (Wake Forrest), Jed Duersch (SNL), Karen Devine (SNL), Srinivas Eswar (Georgia Tech), David Hong (Michigan), Jiajia Li (PNNL), Eric Phipps (SNL), Rich Vuduc (Georgia Tech), Jeff Young (Georgia Tech)

For more information and references: www.kolda.net