SAND2008-4971C

Multivariate Analysis of Hyperspectral Images: A Tutorial

Mark Van Benthem and Michael Keenan

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Overview

- Introduction and motivation of work
- Data types: Univariate and multivariate
  - From simple principles to the more complex
- Principal Component Analysis
  - The workhorse of multivariate analysis
- Multivariate curve resolution
  - Two-way data analysis
- Spatial Image Compression
  - Faster-better
- Trilinear Analysis

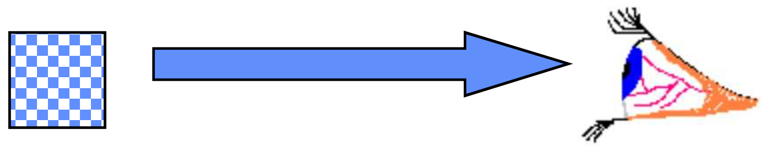
Sandia National Laboratories

Motivation

- Hyperspectral imaging is becoming commonplace in science and engineering
  - Chemical analysis
  - Materials characterizing - from nanoscale on up
  - Biology application with regard to cell function
  - Remote sensing
- Data sets are very large and can easily overwhelm researchers and users
- The rapid advance of hyperspectral imaging hardware has overwhelmed the available analysis software

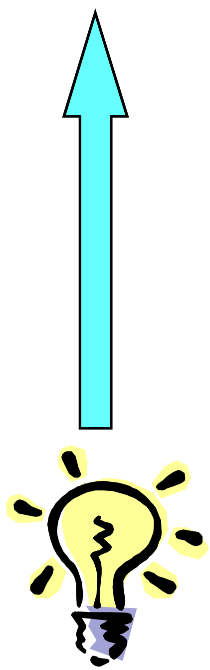
Sandia National Laboratories

Univariate Linear Model-Fluorescence



Model

$F = Cs$

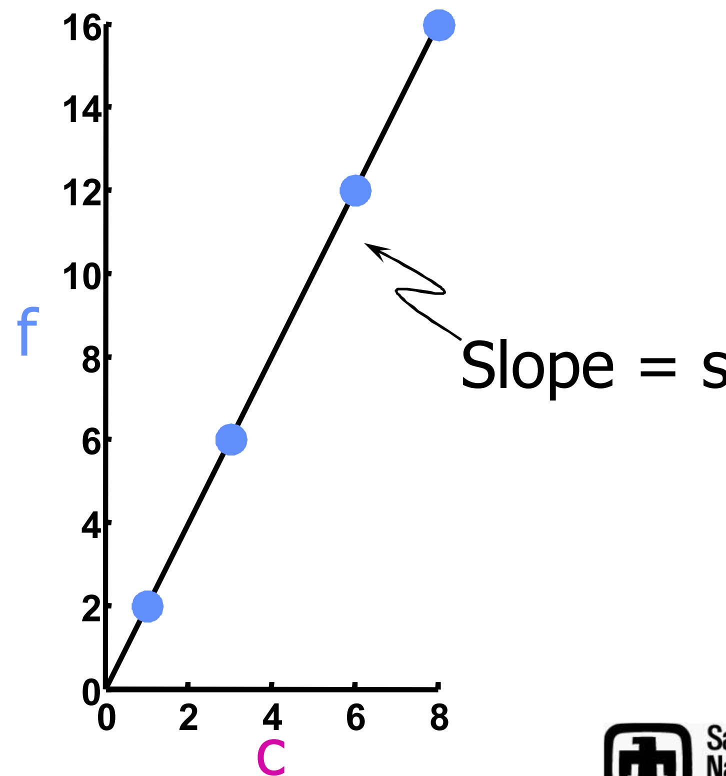


$f_1 = c_1 s$

$f_2 = c_2 s$

$f_3 = c_3 s$

$f_4 = c_4 s$



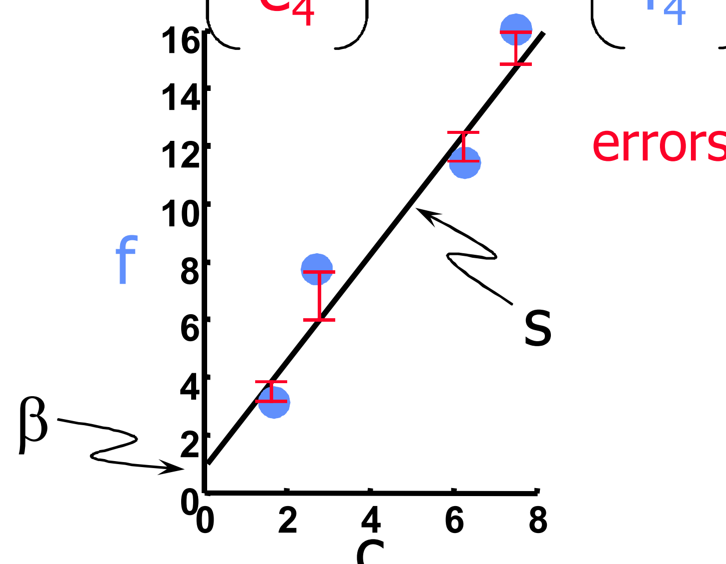
Sandia National Laboratories

Univariate Model

$$\mathbf{c} \times s + \beta + \mathbf{e} = \mathbf{F}$$

$$\begin{pmatrix} c_1 & 1 \\ c_2 & 1 \\ c_3 & 1 \\ c_4 & 1 \end{pmatrix} \times \begin{bmatrix} s \\ \beta \end{bmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix}$$

Add Intercept β



Sandia National Laboratories

Univariate Model and Least Squares

$$\mathbf{C}^T\mathbf{C} \quad \mathbf{S} = \mathbf{C}^T \mathbf{F}$$

$$\begin{bmatrix} c_1 & c_2 & c_3 & c_4 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 & 1 \\ c_2 & 1 \\ c_3 & 1 \\ c_4 & 1 \end{bmatrix} \begin{bmatrix} s \\ \beta \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & c_3 & c_4 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix}$$

$$\begin{bmatrix} \Sigma c_i^2 & \Sigma c_i \\ \Sigma c_i & n \end{bmatrix} \begin{bmatrix} s \\ \beta \end{bmatrix} = \begin{bmatrix} \Sigma c_i f_i \\ \Sigma f_i \end{bmatrix}$$

$$\begin{bmatrix} \hat{s} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} \Sigma c_i^2 & \Sigma c_i \\ \Sigma c_i & n \end{bmatrix}^{-1} \begin{bmatrix} \Sigma c_i f_i \\ \Sigma f_i \end{bmatrix}$$

$$(\mathbf{C}^T\mathbf{C})^{-1}(\mathbf{C}^T\mathbf{C})\, \hat{\mathbf{S}} = (\mathbf{C}^T\mathbf{C})^{-1} \quad \mathbf{C}^T\mathbf{F}$$

^ represents least squares estimate

Normal Equations

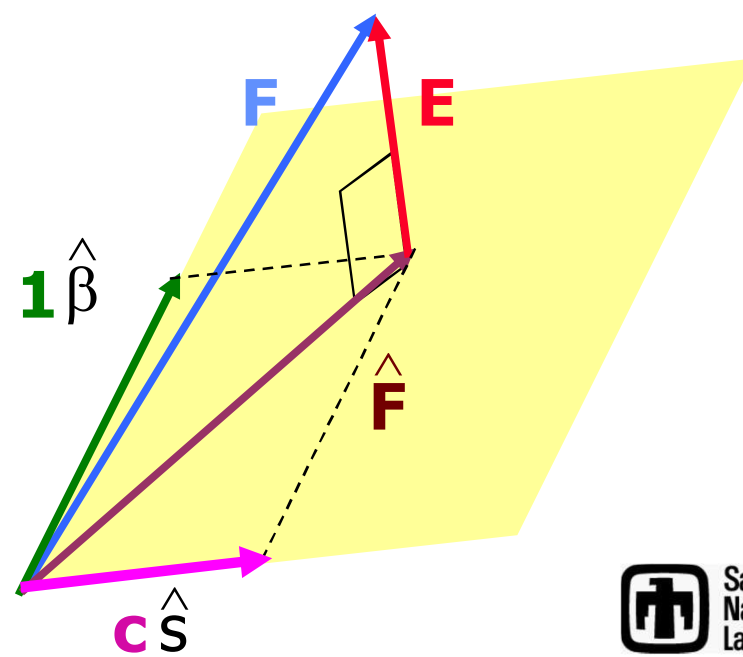
The model: $\mathbf{F} = \mathbf{C}\mathbf{S} + \mathbf{E}$

Given $\mathbf{C}$, estimate $\mathbf{S}$ $\quad \hat{\mathbf{S}} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{F} = \mathbf{C}^{\ddagger}\mathbf{F}$

And then $\mathbf{F}$ $\quad \hat{\mathbf{F}} = \mathbf{C}\hat{\mathbf{S}} = \underbrace{\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T}_{\text{“Hat” matrix}}\mathbf{F} = \mathbf{C}(\mathbf{C}^{\ddagger}\mathbf{F})$

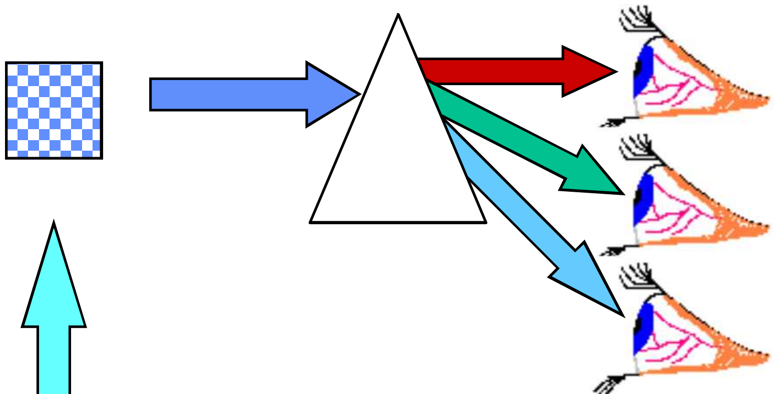
$$\mathbf{F} = \mathbf{C}\,\mathbf{S}$$

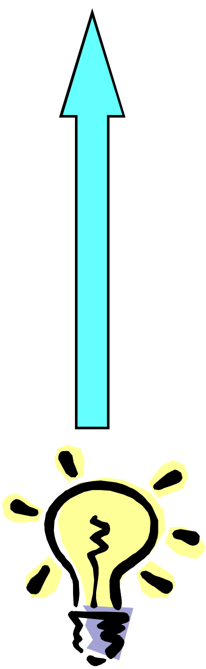
$$\begin{pmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \\ \hat{f}_4 \end{pmatrix} = \begin{pmatrix} c_1 & 1 \\ c_2 & 1 \\ c_3 & 1 \\ c_4 & 1 \end{pmatrix} \begin{pmatrix} \hat{s} \\ \hat{\beta} \end{pmatrix}$$



‡ represents matrix pseudoinverse

Multivariate Linear Model





Model

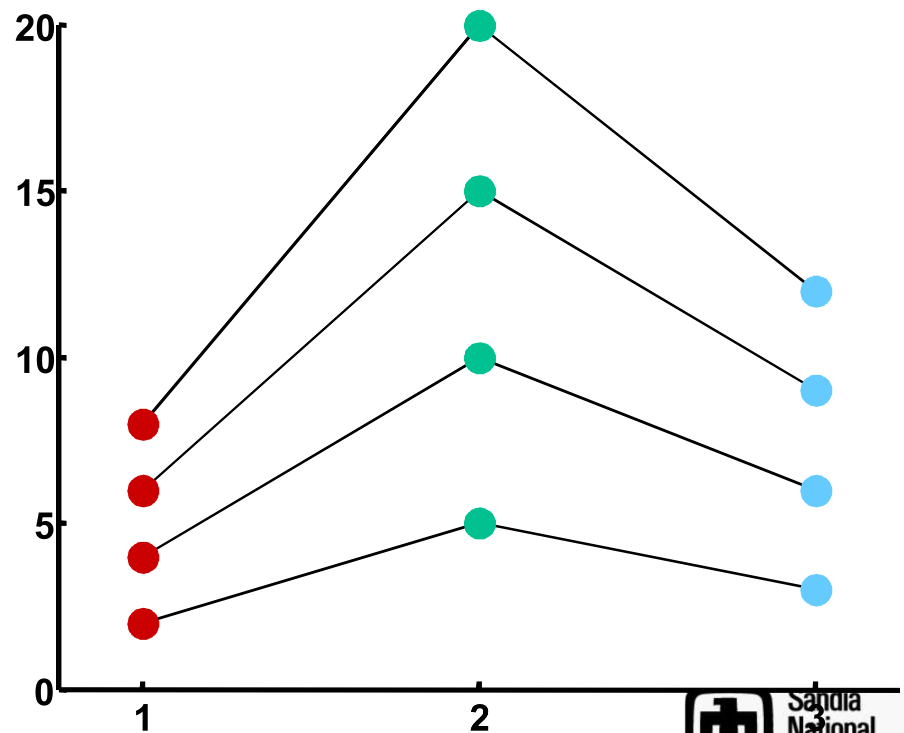
$$[f_{\bullet 1}\ f_{\bullet 2}\ f_{\bullet 3}] = c[s_1\ s_2\ s_3]$$

$$[f_{11}\ f_{12}\ f_{13}] = c_1[s_1\ s_2\ s_3]$$

$$[f_{21}\ f_{22}\ f_{23}] = c_2[s_1\ s_2\ s_3]$$

$$[f_{31}\ f_{32}\ f_{33}] = c_3[s_1\ s_2\ s_3]$$

$$[f_{41}\ f_{42}\ f_{43}] = c_4[s_1\ s_2\ s_3]$$



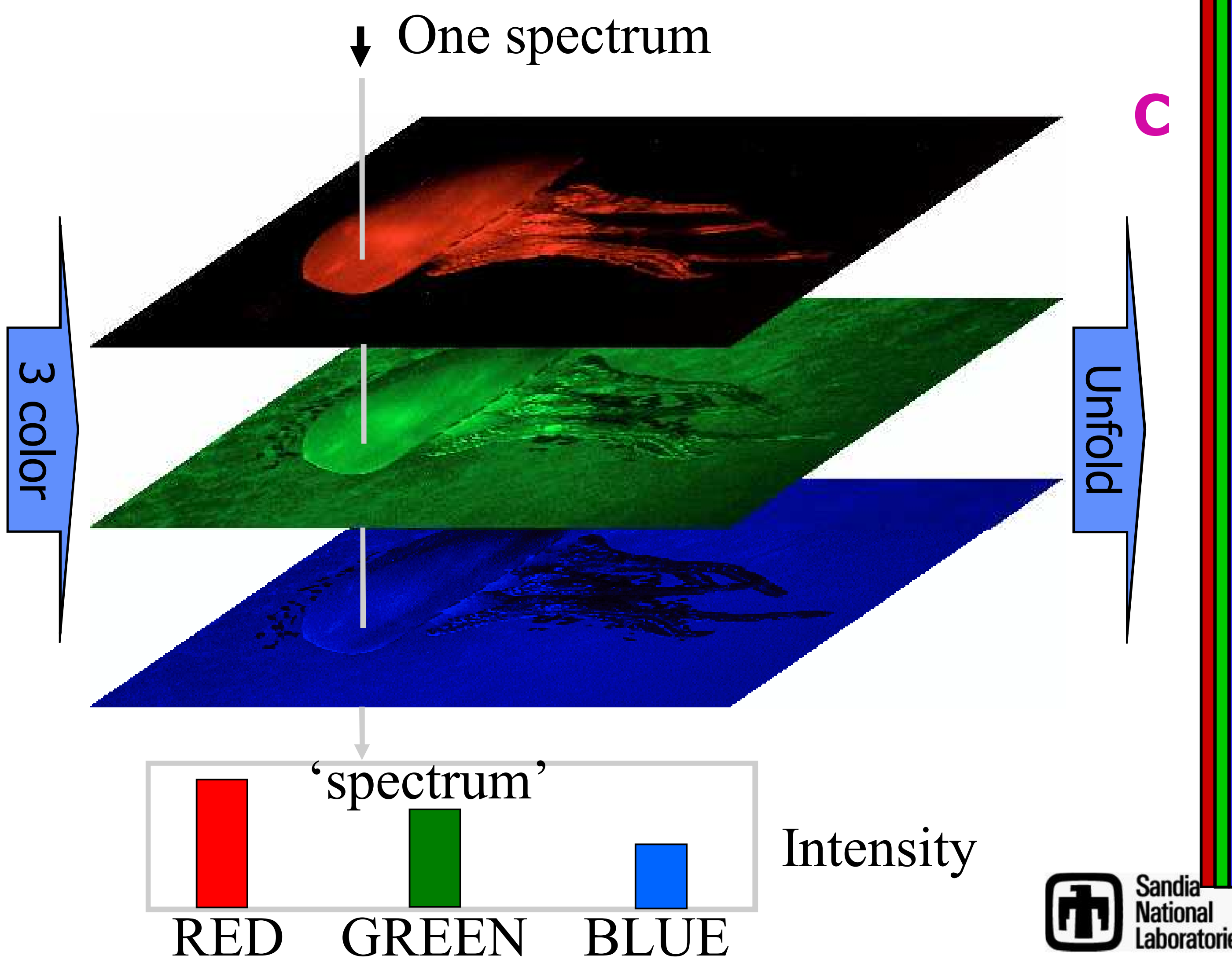
Multicomponent Multivariate Model

$$\mathbf{C} \times \mathbf{S} = \mathbf{F}$$

$$\begin{pmatrix} c_{11} & c_{12} & c_{13} & \alpha_1 \\ c_{21} & c_{22} & c_{23} & \alpha_2 \\ c_{31} & c_{32} & c_{33} & \alpha_3 \\ c_{41} & c_{42} & c_{43} & \alpha_4 \\ c_{51} & c_{52} & c_{53} & \alpha_5 \end{pmatrix} \times \begin{pmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ s_{31} & s_{32} & s_{33} & s_{34} \\ 1 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} & f_{13} & f_{14} \\ f_{21} & f_{22} & f_{23} & f_{24} \\ f_{31} & f_{32} & f_{33} & f_{34} \\ f_{41} & f_{42} & f_{43} & f_{44} \\ f_{51} & f_{52} & f_{53} & f_{54} \end{pmatrix}$$

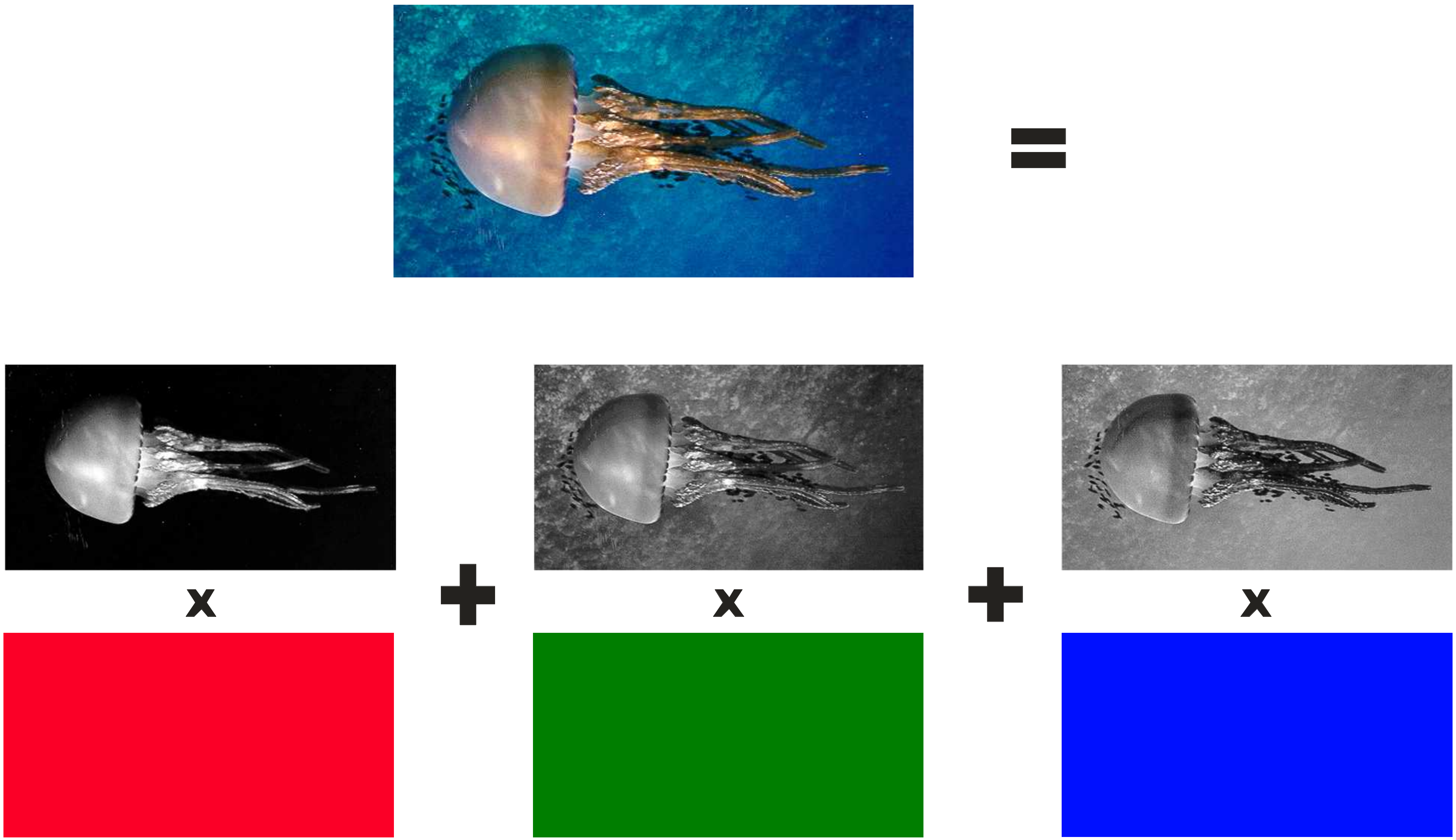
5 samples, 4 Wavelengths , 3 Components

Application to "Hyperspectral" RGB Image



Sandia National Laboratories

RGB Image = C · S

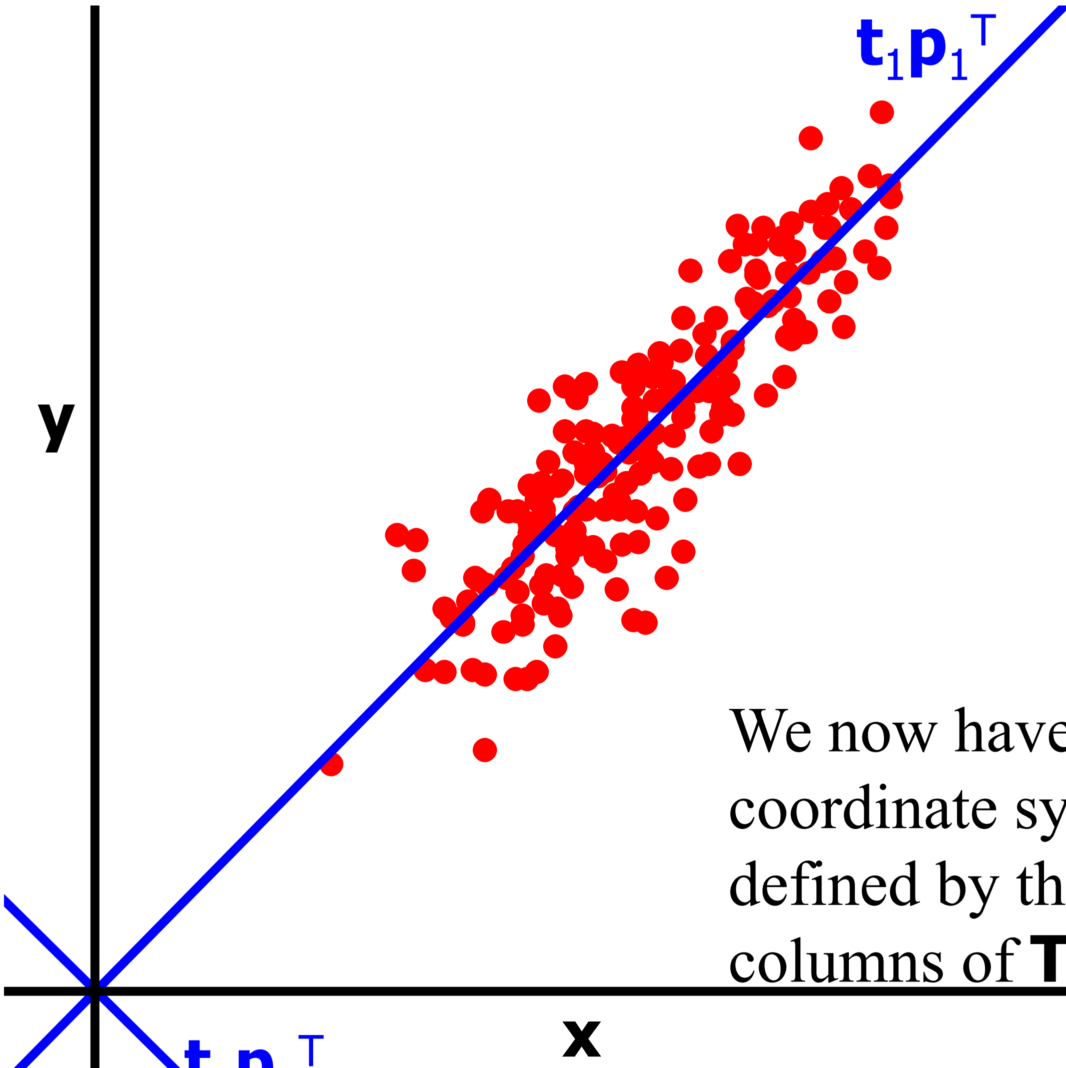


Note the linearity assumption

Principal Component Analysis (PCA)

- Data reduction technique
  - Reduces the dimensionality of or "factors" the data to the appropriate number of components
  - Creates an ordered set of orthogonal "scores," **T** and "loadings," **P**.
    - These are vectors representing the **C** and **S** modes ordered by decreasing variance contribution to the data set.
- Related to eigenanalysis, which is the usual method of computation
  - For a large image, compute the small cross-product matrix $\mathbf{Z} = \mathbf{F}^T\mathbf{F}$ and then solves the eigenvalue problem $\mathbf{ZP} = \mathbf{PD}$, where $\mathbf{D}$ is the diagonal matrix of eigenvalues
  - Then, sort the eigenvalues by size, order the vectors in **P** accordingly, select the appropriate number of vectors
  - Next solve for **T** using $\mathbf{T} = \mathbf{FP}$
  - And the result is the model $\mathbf{F} = \mathbf{TP}^T + \mathbf{E}$

PCA of Univariate Data

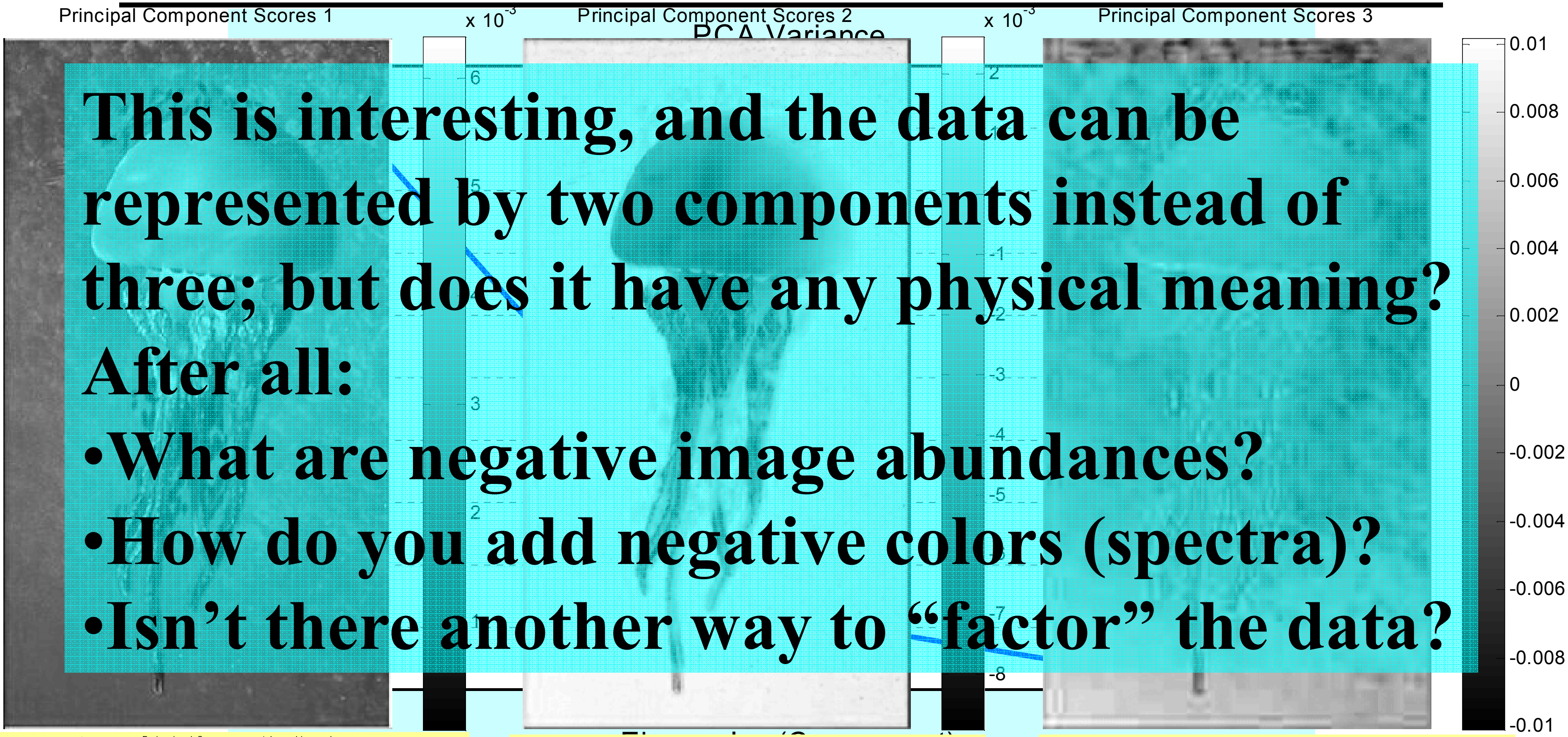


Each point is defined by an x and y coordinate. Let **F** = [**x y**], and **Z** = $\mathbf{F}^T\mathbf{F}$. Solve the eigenvalue problem **ZP** = **PD** and order them by size. Then project **F** into the **P**-space **T** = **FP**.

We now have a new coordinate system defined by the two columns of **T** and **P**

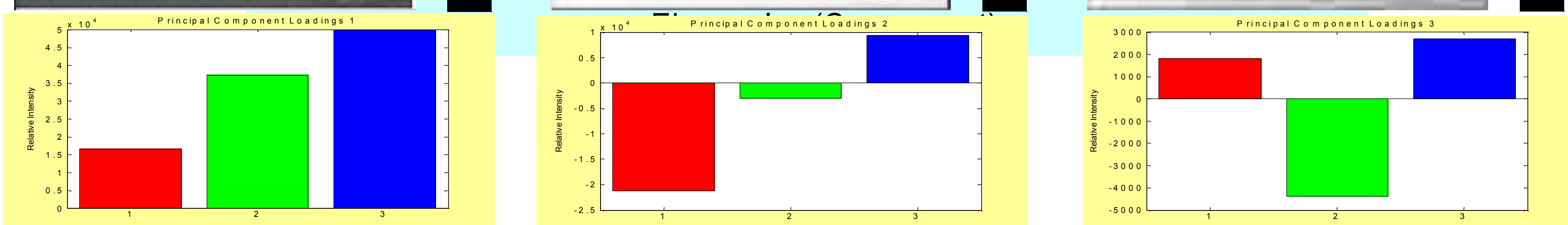
Sandia National Laboratories

PCA of the Jellyfish Image



This is interesting, and the data can be represented by two components instead of three; but does it have any physical meaning? After all:

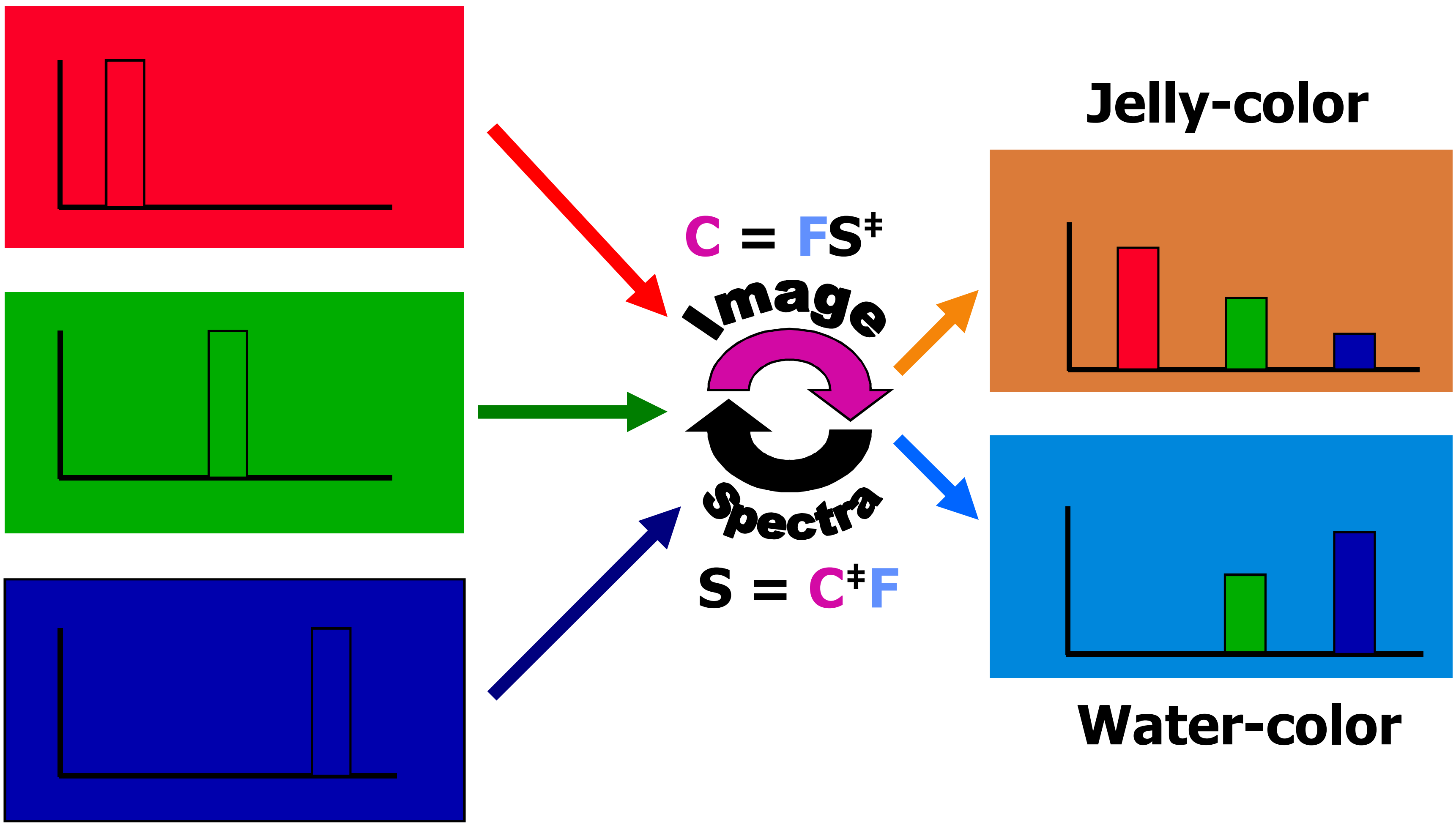
- What are negative image abundances?
- How do you add negative colors (spectra)?
- Isn’t there another way to “factor” the data?



Multivariate Curve Resolution

- Assumption: linear additive model: $\mathbf{F} = \mathbf{C}\,\mathbf{S}$
- Solve $\mathbf{F} = \mathbf{C}\,\mathbf{S}$ in least squares sense
  - $\hat{\mathbf{C}} = \mathbf{F}\,\mathbf{S}^{\ddagger}$ or use PCA-compressed $\mathbf{C} = \mathbf{T}(\mathbf{P}^{T}\mathbf{S}^{\ddagger})$
  - $\hat{\mathbf{S}} = \mathbf{C}^{\ddagger}\,\mathbf{F}$ or use PCA-compressed $\mathbf{S} = (\mathbf{C}^{\ddagger}\mathbf{T})\mathbf{P}^{T}$
  - More unknowns than equations
  - Have infinite number of solutions
- Apply constraints to narrow the solution space to those that are physically acceptable
  - nonnegativity of concentrations and/or spectra
  - Closure (sum-to-one constraint)
  - many others (e.g., unimodality, smoothness, etc.)
- Iterate until converged

The algorithm Transforms Red, Green and Blue to "Water" and "Jelly" colors

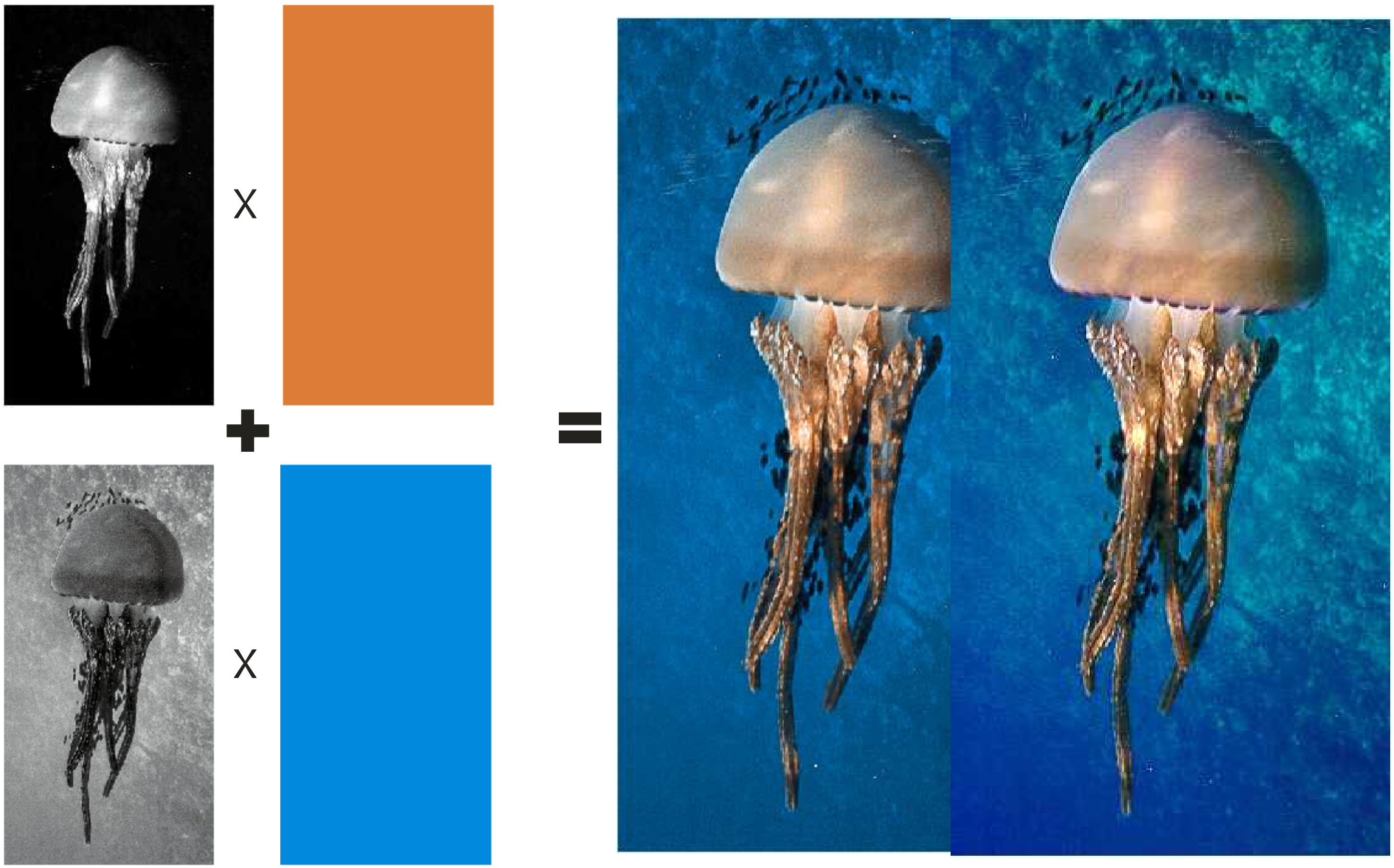


Certain proportions of R, G, and B correlate with each other as determined by the algorithms

Transformation from RGB to WJ Colorspace

- The 3-dimensional RGB color description may not be best if our interests are sea water and jellyfish.
- A 2-dimensional color model might lend more insight into the spectral signatures of sea water and jellyfish.
- Our algorithms can effect the dimensional reduction/transformation.
- Red, Green and Blue become Water-color and and Jelly-color in the new model.

"Water" and "Jelly" are the "pure components" in the WJ color space



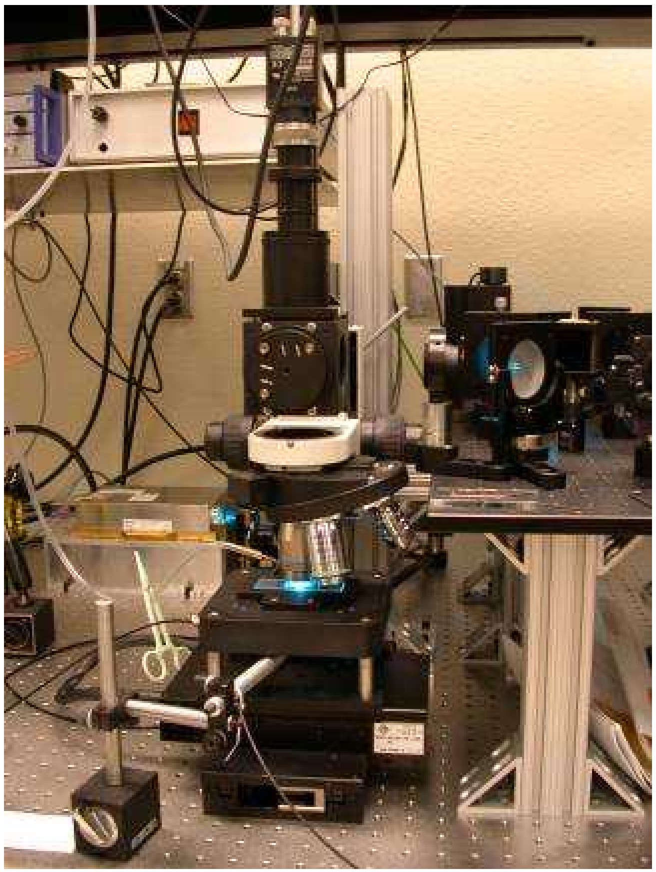
Two-color WJ image

Original RGB image

Note: "Water" and "Jelly" are not listed as Crayola® crayon colors.

Hyperspectral Imaging

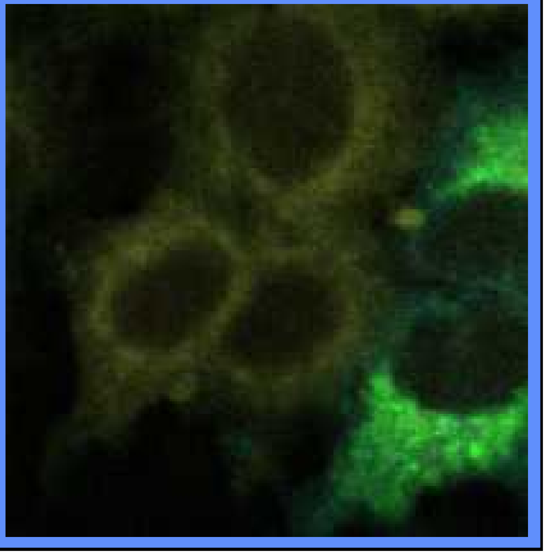
3D Hyperspectral Imaging Confocal Microscope (HSI-CM)



- HSI-CM Performance Specifications:
  - 488 nm laser excitation
  - 10x, 20x, 60x, 100x objectives
  - Lateral Resolution = 0.25 μm
  - Axial Resolution = 0.60 μm
  - Spectral range 490-800 nm
    - 512 channels
  - Spectral resolution = 1-3 nm
  - Acquisition rate = 8300 spectra-sec^{-1}
- Data Collection
  - 20x objective
  - Exposure time 0.24 msec
  - Image size 204x208 pixels
    - Step size .24 μm in lateral dimensions
  - Data size 512x204x208

Hyperspectral Cell Image

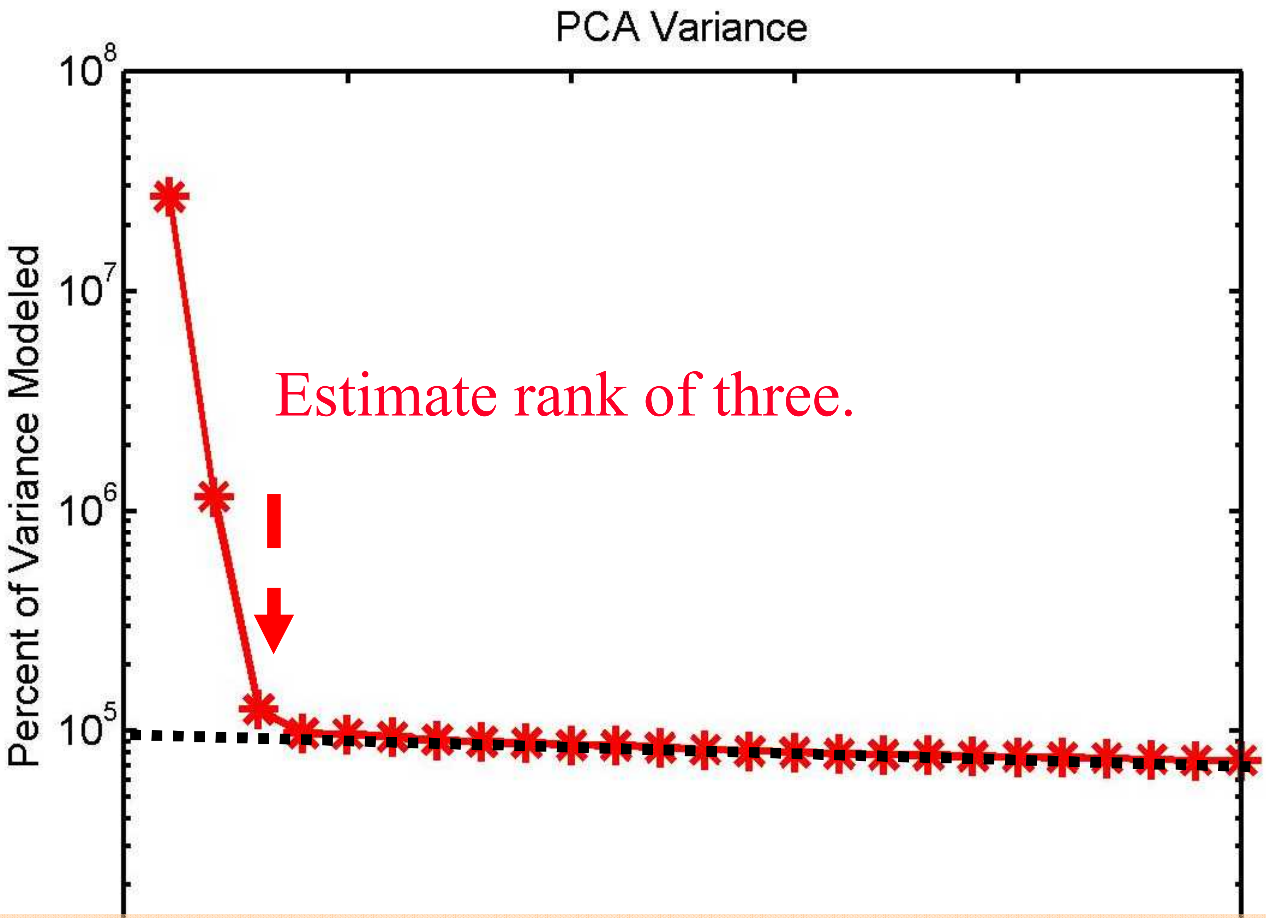
RGB image of a Human A549 pulmonary type II epithelial cell labeled with GFP and YFP collected with HSI-CM. Produced by integrating over three wavelength bands chosen for responses of GFP, YFP and cell autofluorescence.

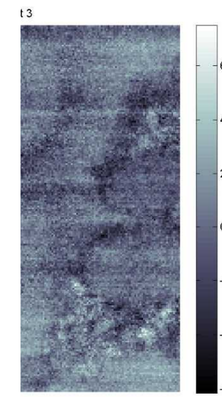


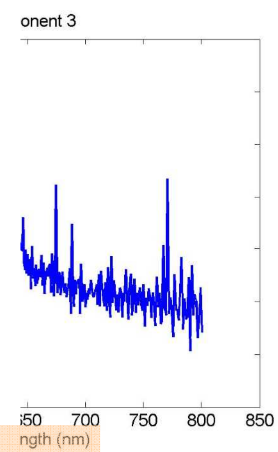
- Data Preprocessing-A critical process
  - Remove cosmic ray-induced "spikes"
  - Remove dark shape—rank one PCA component
    - Wavelength dependent spectral response of EMCCD
  - Trimmed first six wavelength-mode data points
    - Remove some baseline spikes
  - Poisson scale data in spectral mode, attempt to make noise iid normal
    - For Poisson distributed data $\sigma^2 = \mu$
    - Use inverse of spectral-mode mean times gain factor plus read noise
    - **H**: a diagonal spectral scaling matrix
    - $\tilde{F} = HF$

Sandia National Laboratories

PCA of Cell HSI Data

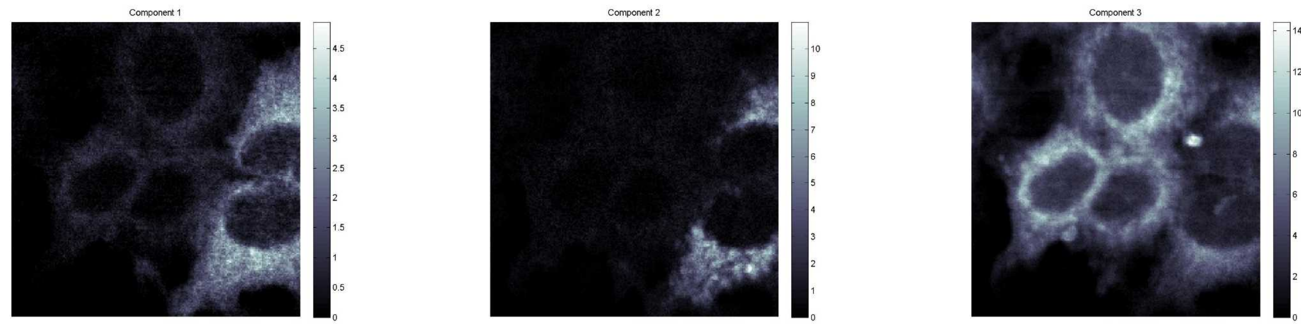


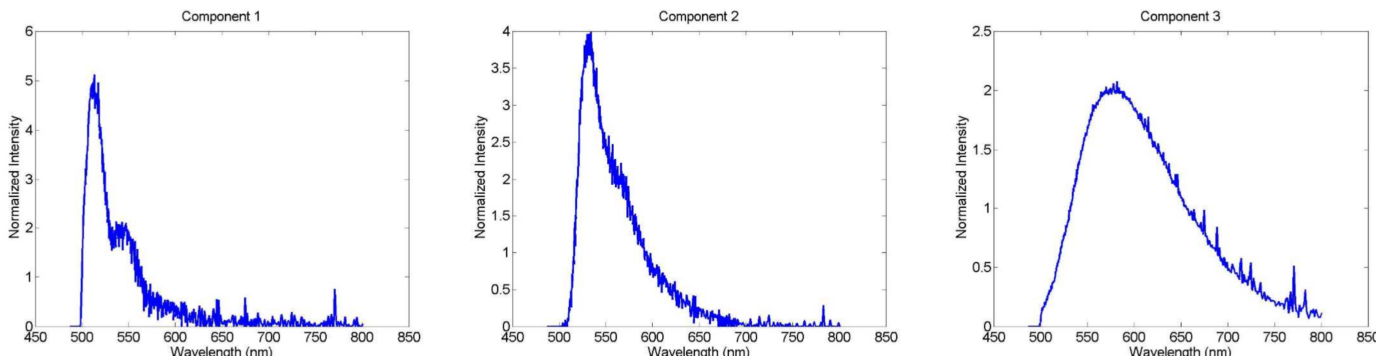


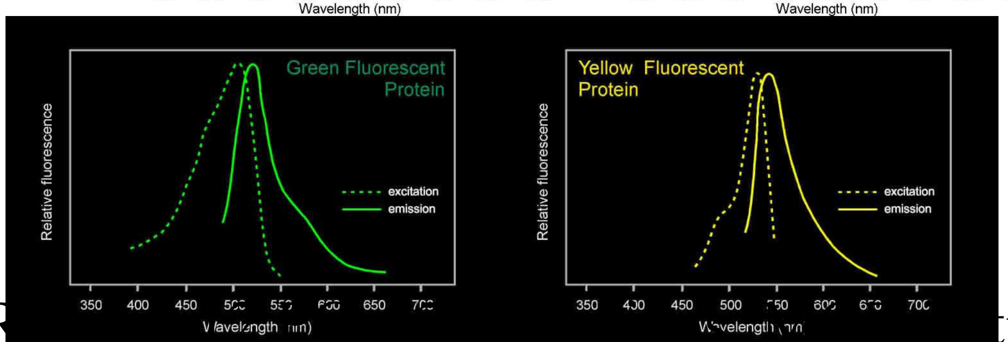


PCA gives us information about the data and a reduced data set with which to perform further analyses

MCR of Cell HSI Data





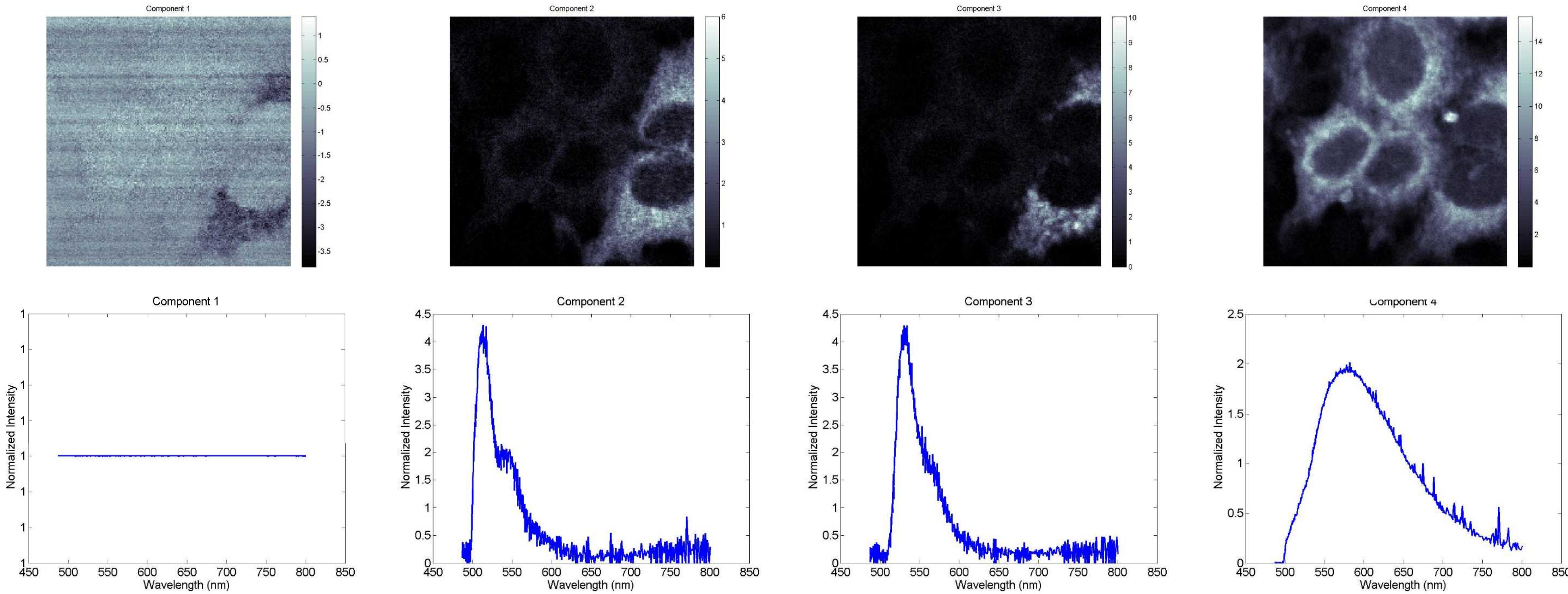


Spectra from http://home.ncifcrf.gov/ccr/flowcore/spectra.htm

MCR … tialized with previously resolved GFP, YFP and AF. Time to run 4.9 seconds.

Sandia National Laboratories

MCR of Same Cell HSI Data with Baseline



MCR using 12 PCs, 50 iterations, MCR initialized with previously resolved GFP, YFP and AF. Time to run 4.9 seconds.

These results are not as satisfactory as we would like. The YFG and GFP components are somewhat noisy and there is some “mixing” in factors. What else can we do?

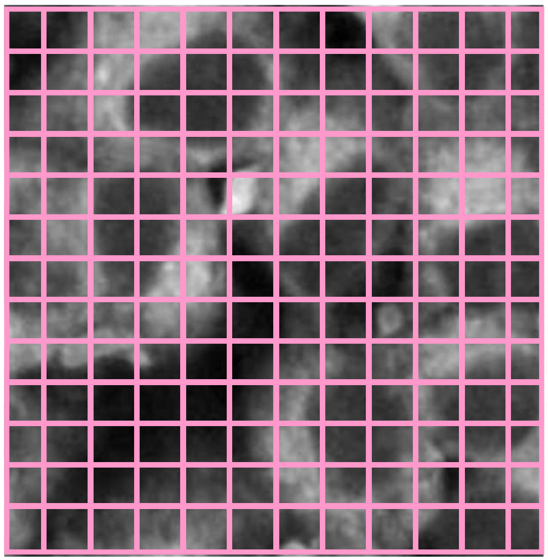
Sandia National Laboratories

Data Compression

- Can be simply and easy to apply (e.g., binning)
- Can achieve good analysis results with very high compression factors
- Can improve signal-to-noise characteristics of data
  - Random noise adds destructively
  - Signal adds constructively, enhancing signal
  - Binning preserves Poisson noise (Poisson adds as Poisson) and read noise (variances add)
- Subsequent analysis proceeds much faster!

Spatial Binning Compression of a Cell Image

Sum all of the pixels
in each square



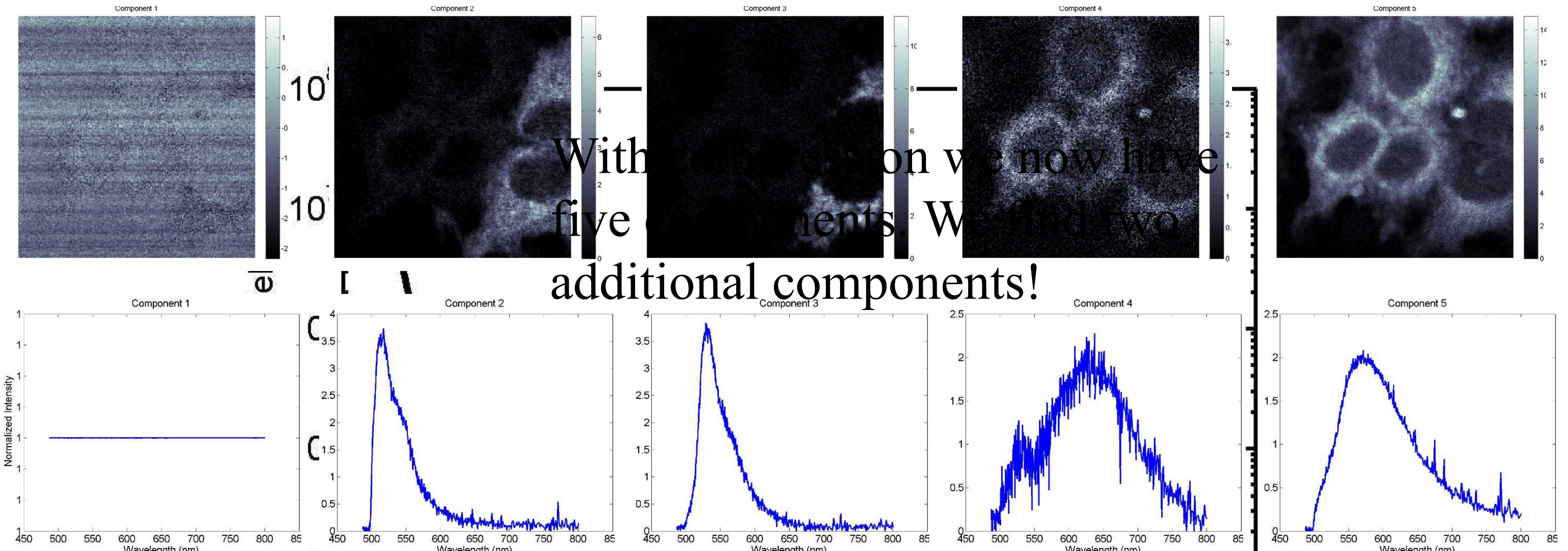
Original Image
208×204 pixels

Compression factors of
16 and 17 (total 272)
chosen for convenience

Compressed Image
13×12 pixels

Compression doesn’t have to preserve the image “quality.”
But, if you stand way back…

MCR of Cell Binned HSI Data



With compression we now have five components. We find two additional components!

MCR using 12 PCs, 100 iterations, MCR initialized with previously resolved GFP, YFP and AF. Time to run 0.9 seconds. Full resolution images obtained by projecting full data (loading and scores) into spectral-mode pure component space. (This calculation is include in the time above.)

Limitations of MCR and Two-Way Analyses

- **MCR is not magic!**
  - **Ensure you are employing appropriate constraints.**
  - **Other methods with different cost functions may be more reasonable for your data.**
  - **Heavily dependent on constraints to limit solution space.**
- **Plagued by the "Rotation Problem"**
  - **For the model $F = CS$, we minimize $Q = ||F\text{-}CS||^2$ in least squares**
  - **Can insert any rotation or transformation matrix, R, in the equation $F = (CR^{-1})(RS)$, with $I = R^{-1}R$ and retain the same Q**
  - **Never sure of unique decomposition**
- **Is there a way to get a unique decomposition?**

Sandia National Laboratories

Three-Way Data and Trilinear Methods

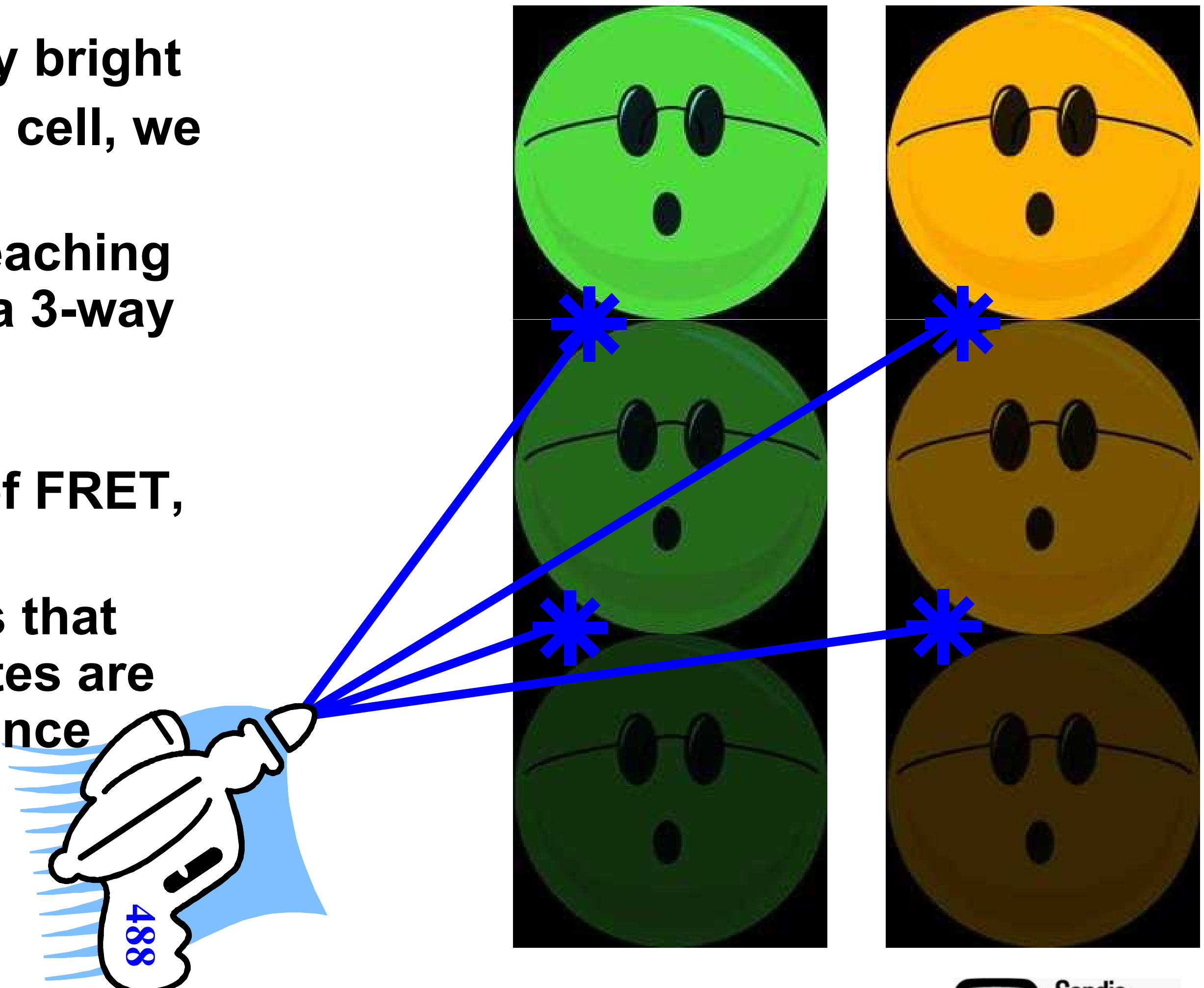
- **Three-way fluorescence data that follow a trilinear model**
  - **The fluorescence intensity is collected as a function of three independent parameters**
  - **Intensity varies linearly as a function of each parameter**
- **The trilinear model:**
  - $f_{ijk} = c_{ip} s_{jp} b_{cp} + e_{ijk}$ **for data with p-components**
  - $\mathbb{F} = \otimes(\mathbf{S}, \mathbf{C}, \mathbf{B})$, **which is a three-way array**
- **Provides rotationally unique decompositions**
  - **Given appropriate data rank structure**

$\otimes$ here is the triple-product operator.

Sandia National Laboratories

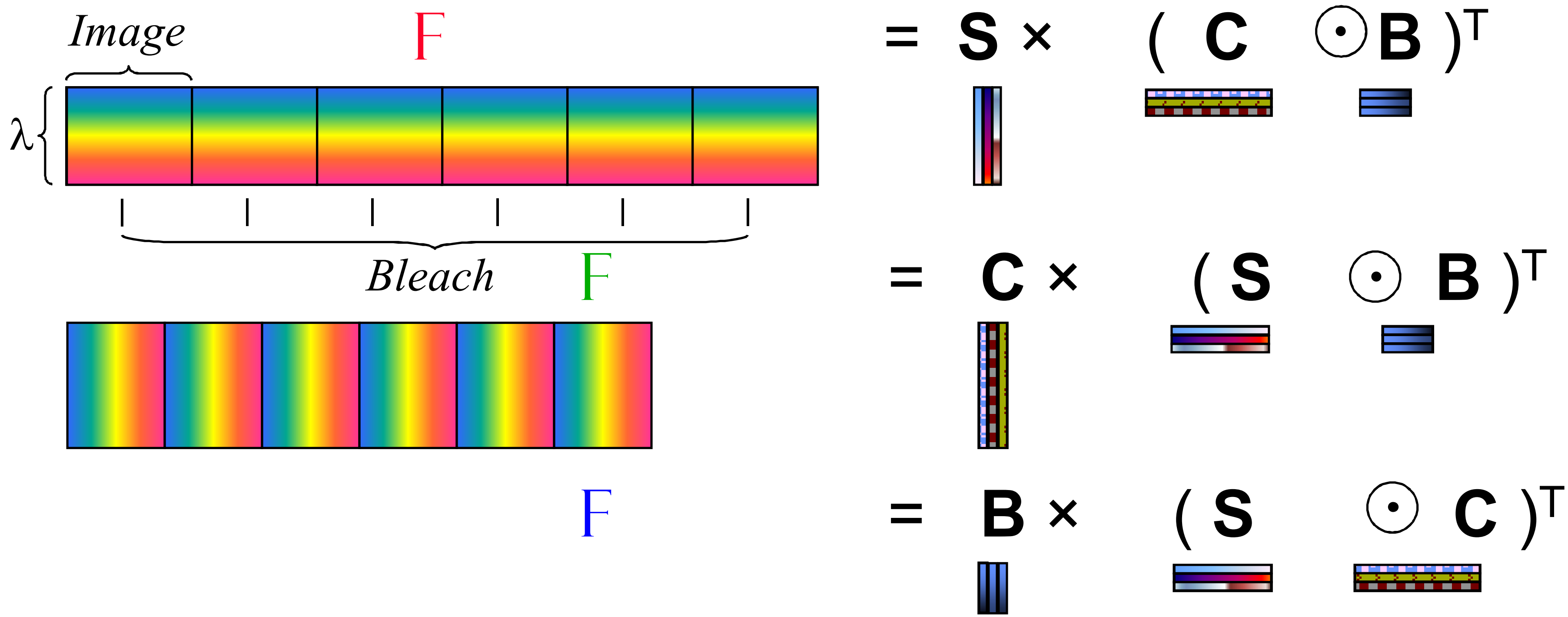
Higher Order Data and Analysis

- **Microscope laser is very bright**
- **Each time we image the cell, we coincidentally bleach it!**
- **At the same time we bleaching the cell, we're creating a 3-way array of data!**
- **If we are interested in determining presence of FRET, we can photobleach.**
- **Of critical importance is that GFP and YFP bleach rates are dependent on the presence or absence of FRET.**



FRET: fluorescence resonance energy transfer.

Three-Way Analysis: PARAFAC – PARallel FACtors Analysis (Conceptual)



$$\mathbf{S} \cong \mathbf{F} \times ((\mathbf{C} \odot \mathbf{B})^T)^{\dagger} \quad \mathbf{C} \cong \mathbf{F} \times ((\mathbf{S} \odot \mathbf{B})^T)^{\dagger} \quad \mathbf{B} \cong \mathbf{F} \times ((\mathbf{S} \odot \mathbf{C})^T)^{\dagger}$$

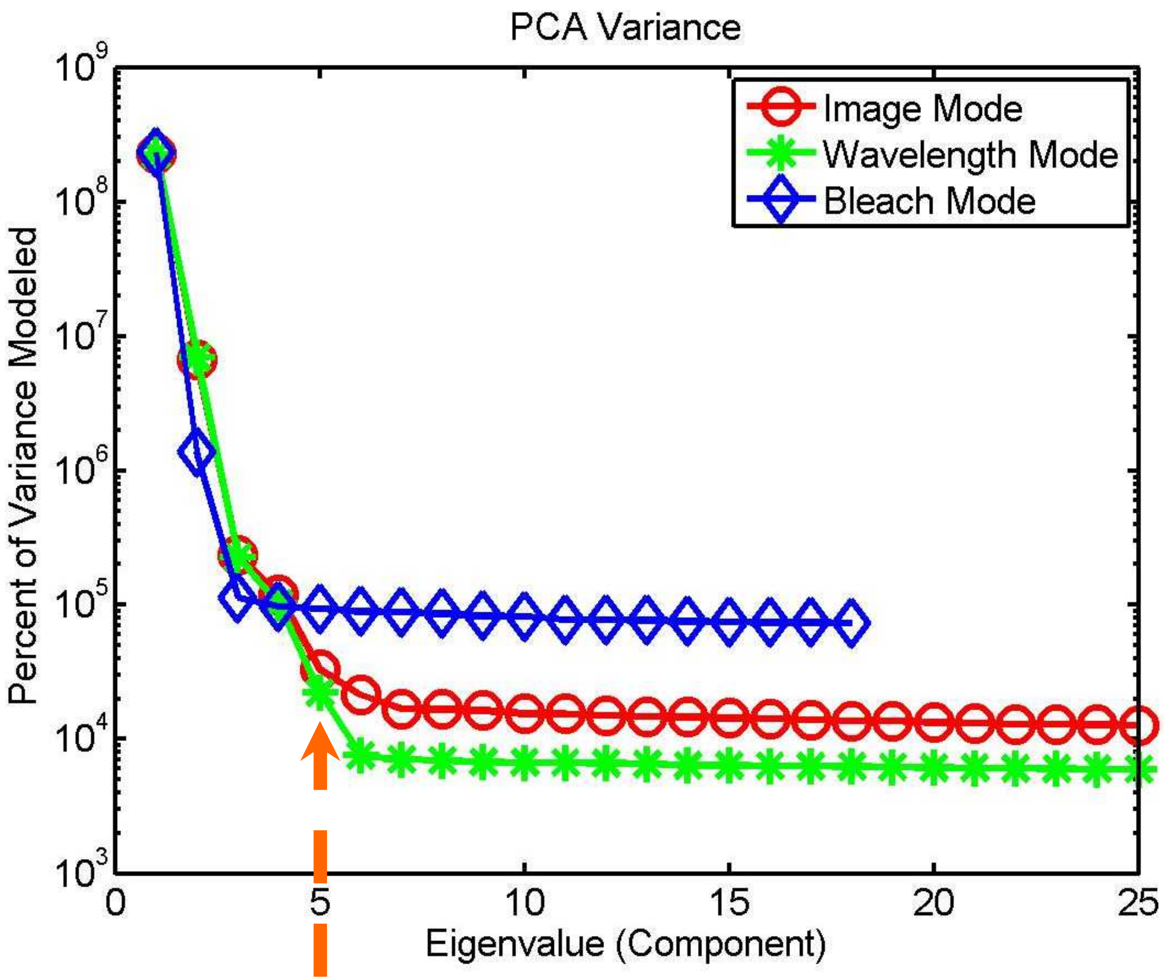
$\odot$ is known as the Khatri-Rao operator.

Application of nonnegative least squares (NNLS) optional.

Three-Way Cell Data

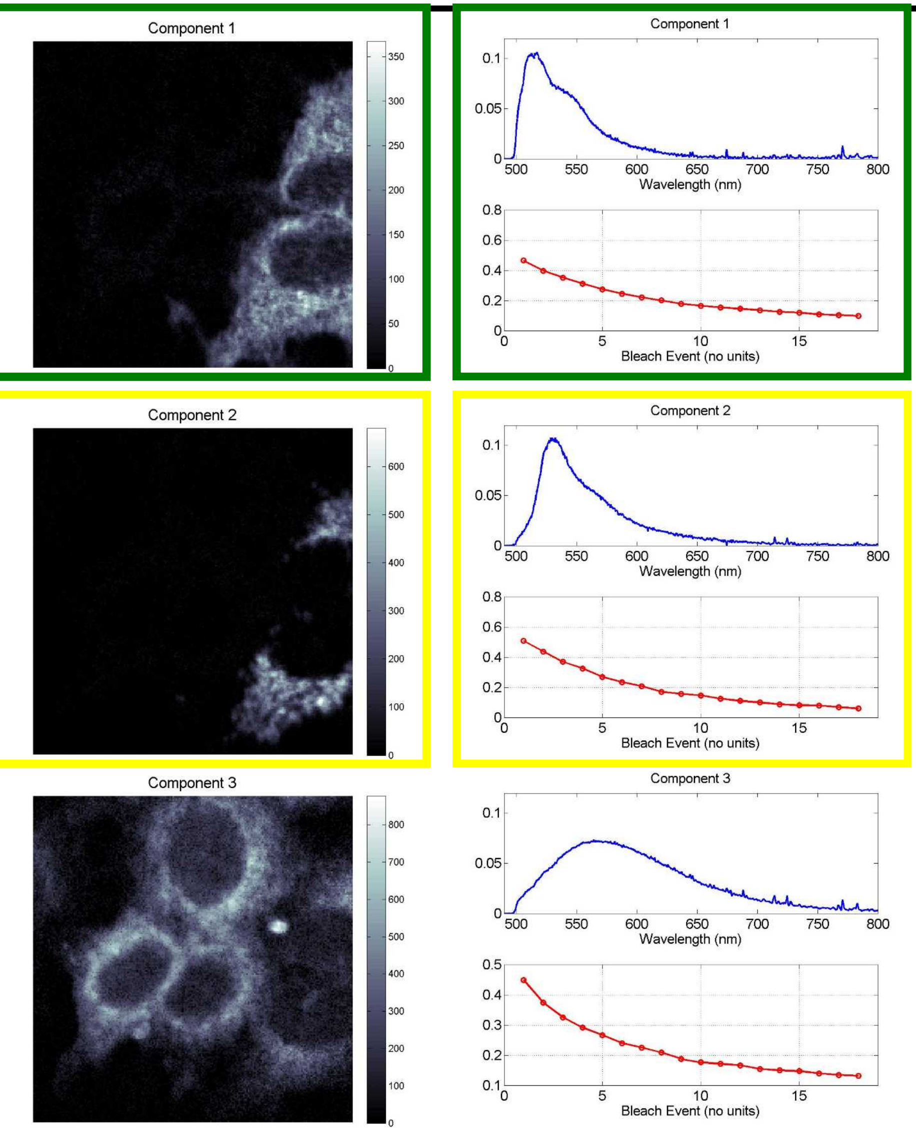
- Collected 18 consecutive photobleach images of same cells previously examined
- Preprocessing
  - Remove cosmic ray-induced "spikes"
  - Remove dark shape—rank one PCA component
    - Wavelength dependent spectral response of EMCCD
  - Trimmed first six wavelength-mode data points
  - Subtract offset for each image pixel
    - Baseline correct using known zero-signal elements
  - Binning compression in image-mode as previously described
  - Poisson scale in spectral mode
- Analyzed with a PARAFAC-ALS routine written in-house

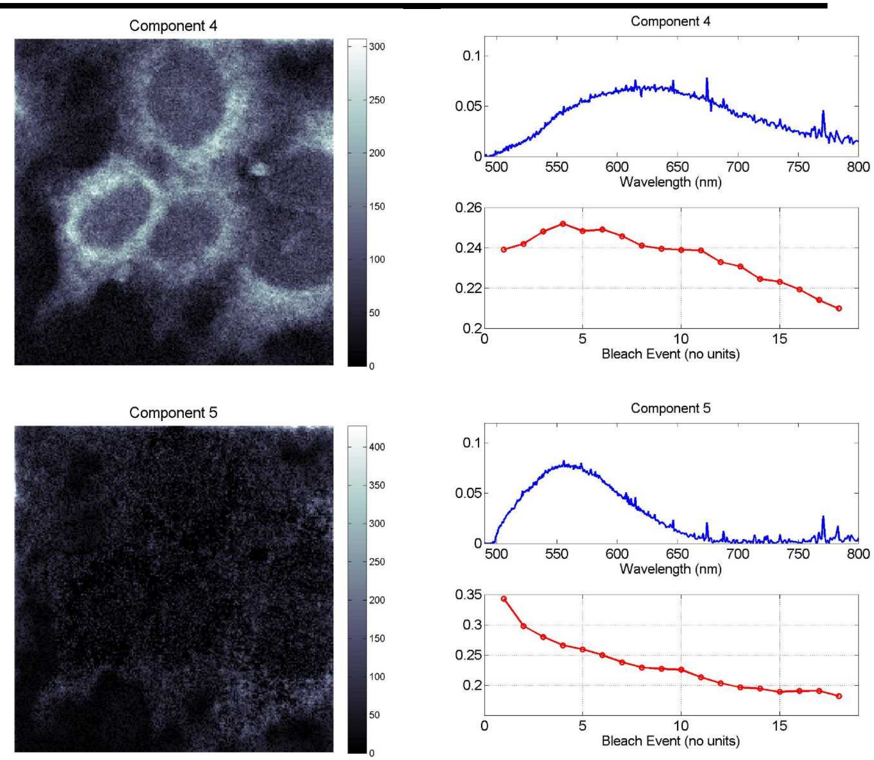
HSI Cells Three-Mode Rank Estimate



Largest rank in wavelength- and image-modes is five.

GFP-YFP-AF PARAFAC Model





PARAFAC easily finds GFP and YFP fluors and returns their photobleach curves and their cell locations.

Sandia National Laboratories

Summary

- Introduction and motivation of work
- Data types: Univariate and multivariate
  - From simple principles to the more complex
- Principal Component Analysis
  - The workhorse of multivariate analysis
- Multivariate curve resolution
  - Two-way data analysis
- Spatial Image Compression
  - Faster-better
- Trilinear Analysis

Recap and Summary

- Presented least squares methods for multivariate analysis of hyperspectral images
  - Least squares for linear models
- PCA the workhorse that is useful for data compression and rank estimation
- MCR: appropriate for many two-way models, but is not the answer for every problem.
  - You must choose the right tool for the job!
- Compression can take many forms
  - Binning images is simple and works well for linear additive data
- PARAFAC is a powerful tool for three-way analysis
  - Data must follow the trilinear model
- It’s all mathematics—not magic

Acknowledgements

- Sandia National Laboratories
  - Paul Kotula, Dave Haaland, Howland Jones, Jeri Timlin, and Tony Ohlhausen
- University of Texas Medical Branch
  - Allan Brasier, Ping Liu

Sandia National Laboratories