Argonne NATIONAL LABORATORY | Advanced Photon Source

NORTHWESTERN UNIVERSITY

Finding patterns in data: an overview

Chris Jacobsen

Advanced Photon Source, Argonne National Laboratory, USA

Department of Physics & Astronomy, Northwestern University, USA

Chris.J.Jacobsen@gmail.com

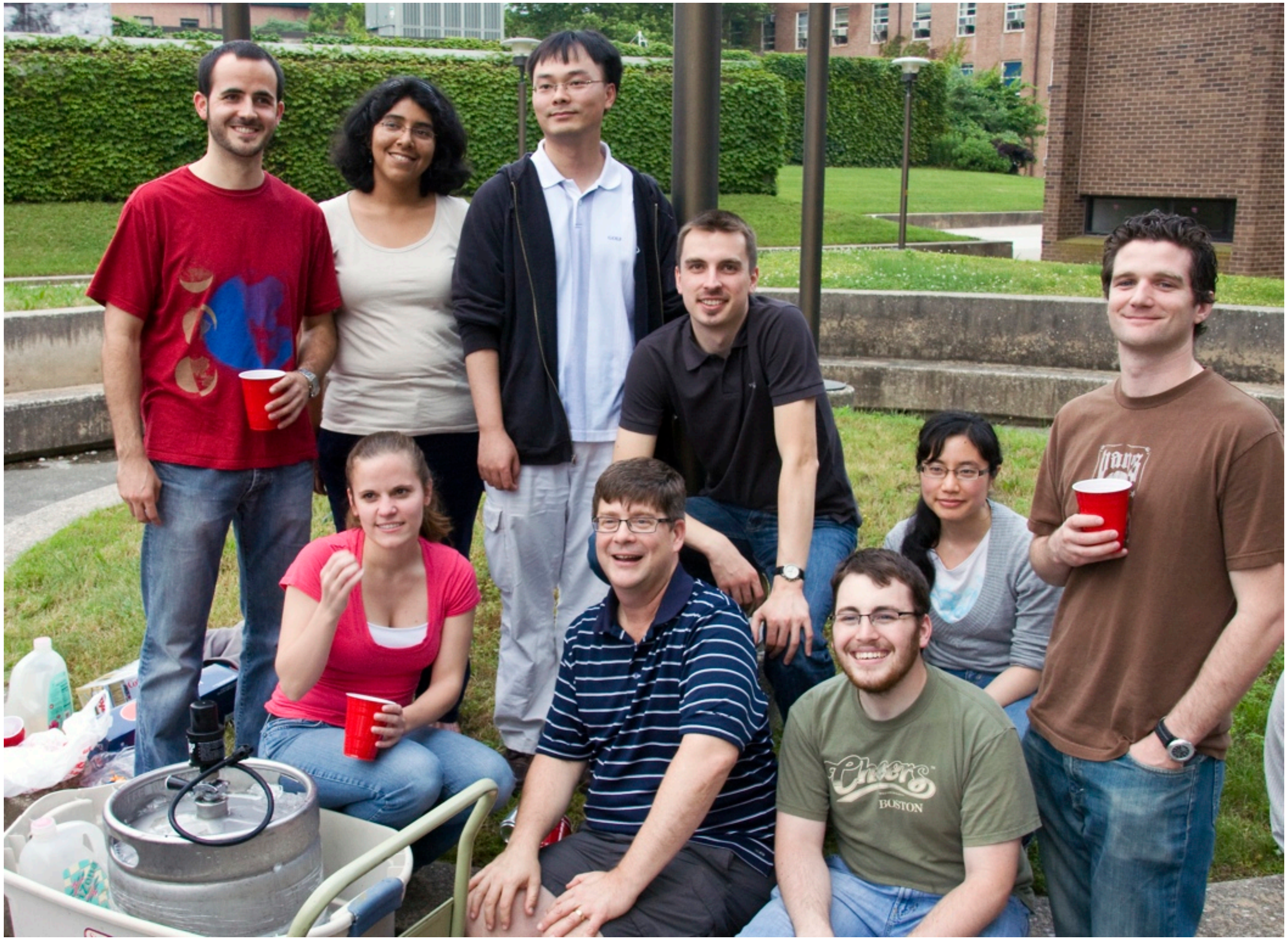
This workshop

- Thanks to Armando Solé and Andy Götz for organizing this workshop!
- Nitty gritty details of data storage are often swept under the rug at conferences and workshops, yet play a huge role in practice.
- How does one try other analysis programs? How does one share data with other researchers?

X-ray microscopy group at Stony Brook

Jan Steinbrener, Lisseth Gavilan, Xiaojing Huang, Christian Holzner, Rachel Mak, Josh Turner, Johanna Nelson, Chris Jacobsen, Robert Towers. Not shown: Sue Wirick, Chris Peltzer.

Phase contrast and fluorescence
Spectromicroscopy
XDM/CXDI



Summer at Stony Brook: groups take turns sponsoring the 4:30 pm Friday beer keg

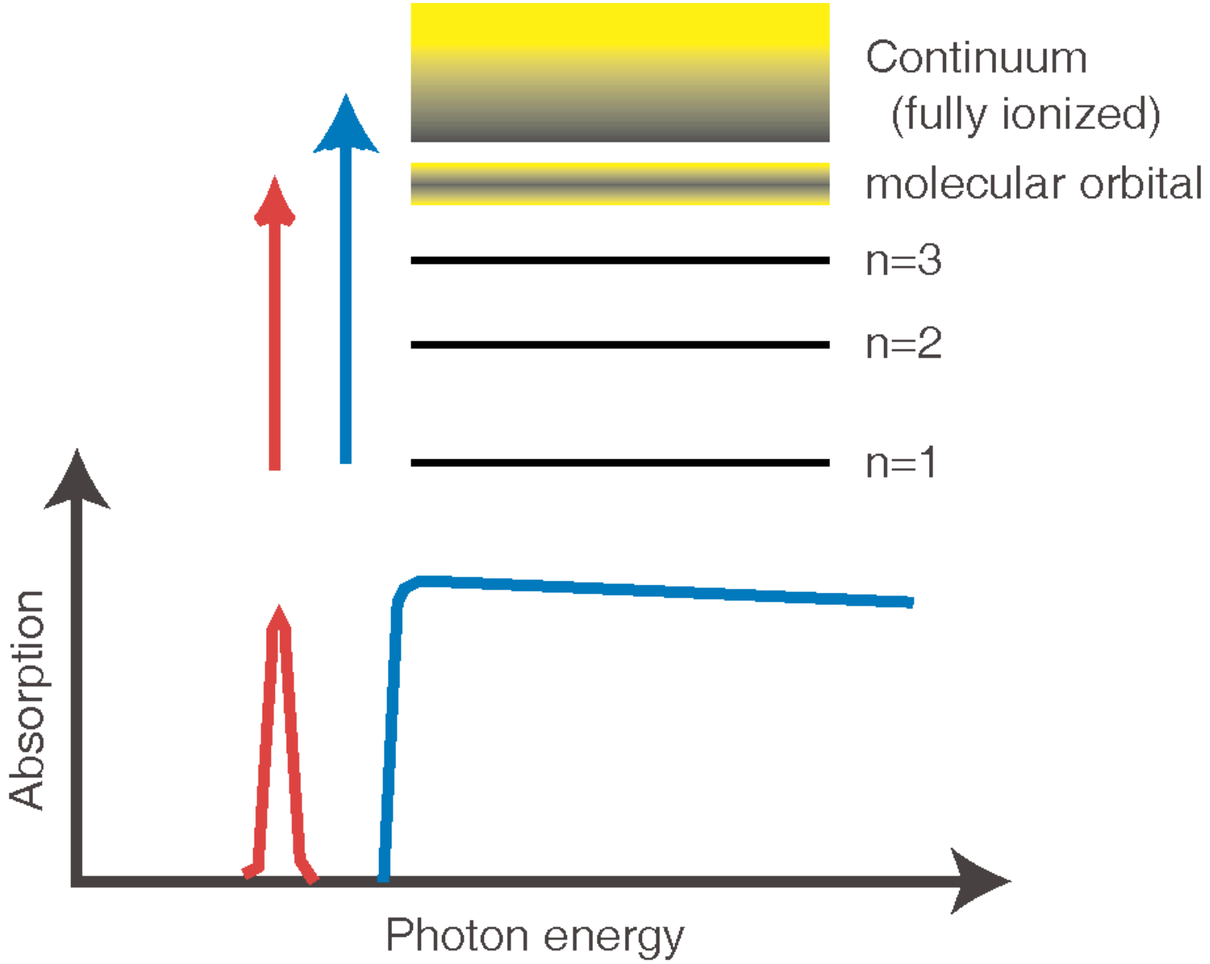
Moving to Northwestern University!

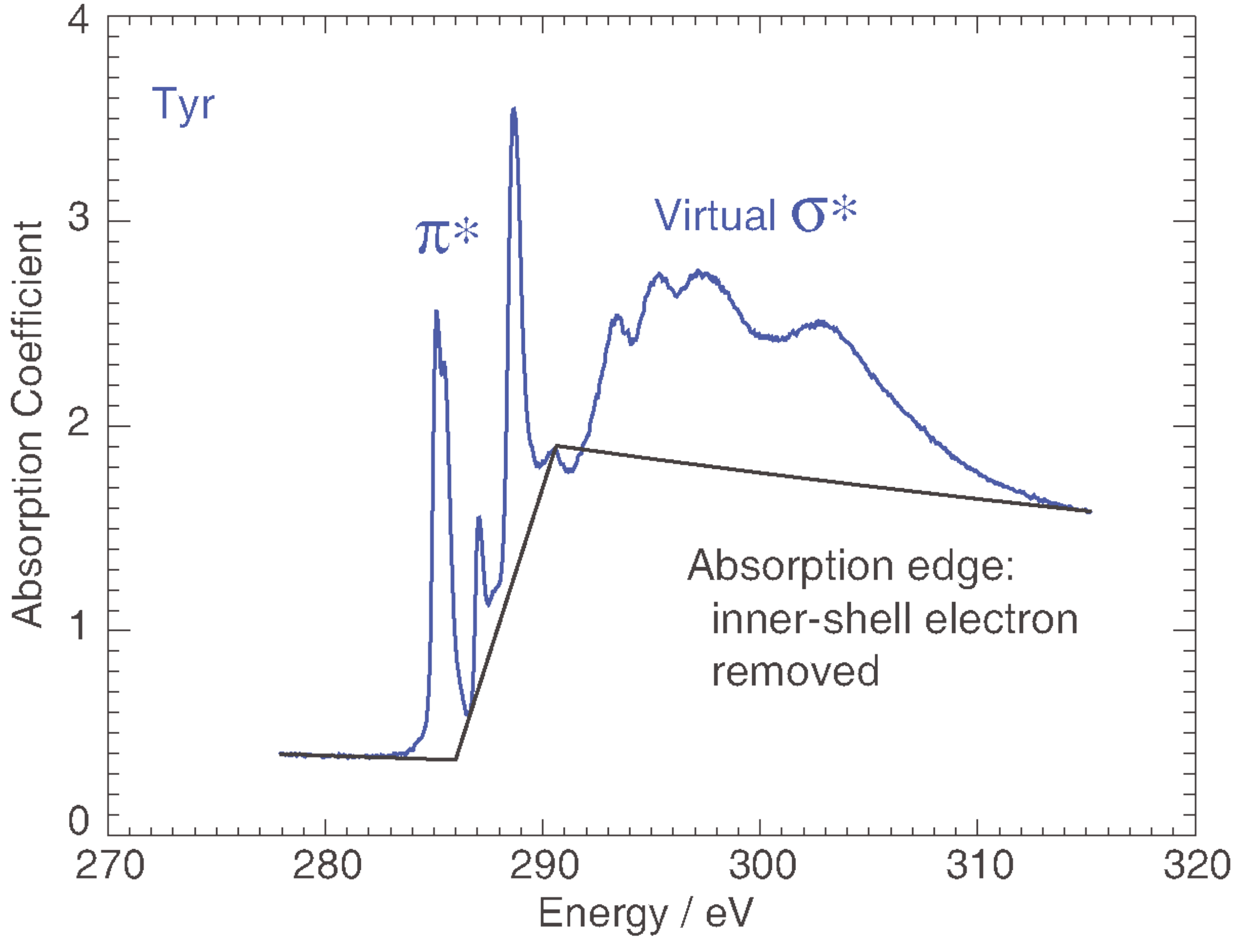
This talk

- Soft x-ray spectromicroscopy: what we do and how we process the data
  - Principal components, clusters, and non-negative matrices
- Connections with problems in other fields
  - X rays, electrons, satellites, shopping…
- Some thoughts on data formats

Near-edge absorption fine structure (NEXAFS) or X-ray absorption near-edge structure (XANES)

- Fine-tuning of the x-ray energy near an atom’s edge gives sensitivity to the chemical bonding state of atoms of that type
- First exploitation for chemical state transmission imaging: Ade, Zhang *et al.*, *Science* **258**, 972 (1992) – Stony Brook/X1A

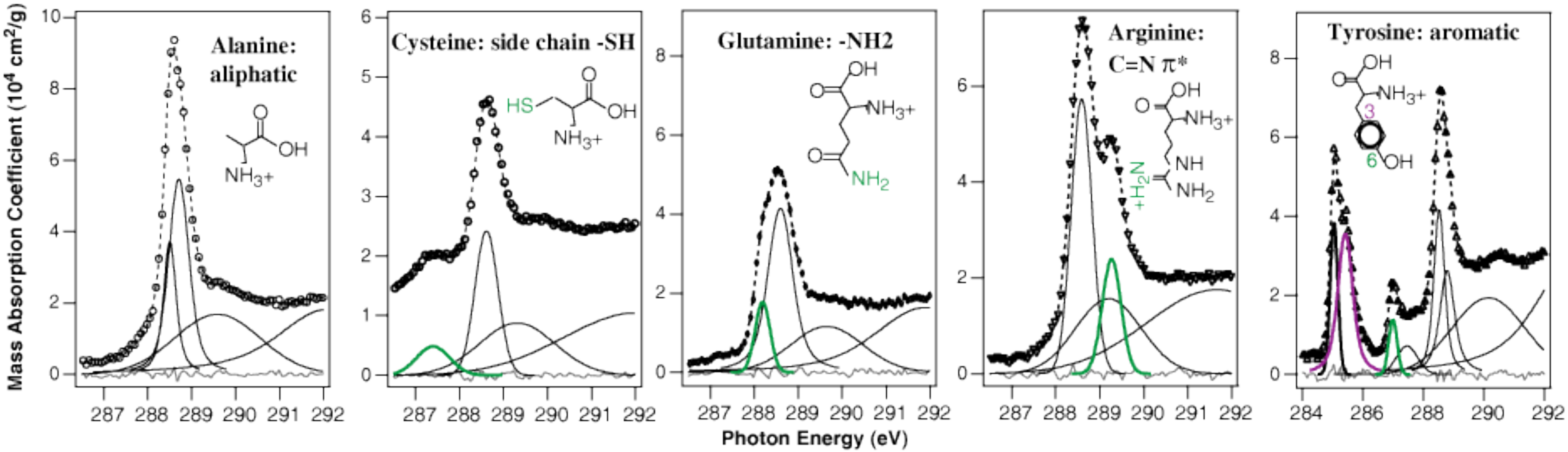




Compared with UV “tickling” of molecular orbitals, core-level electrons come from a single, well-defined state!

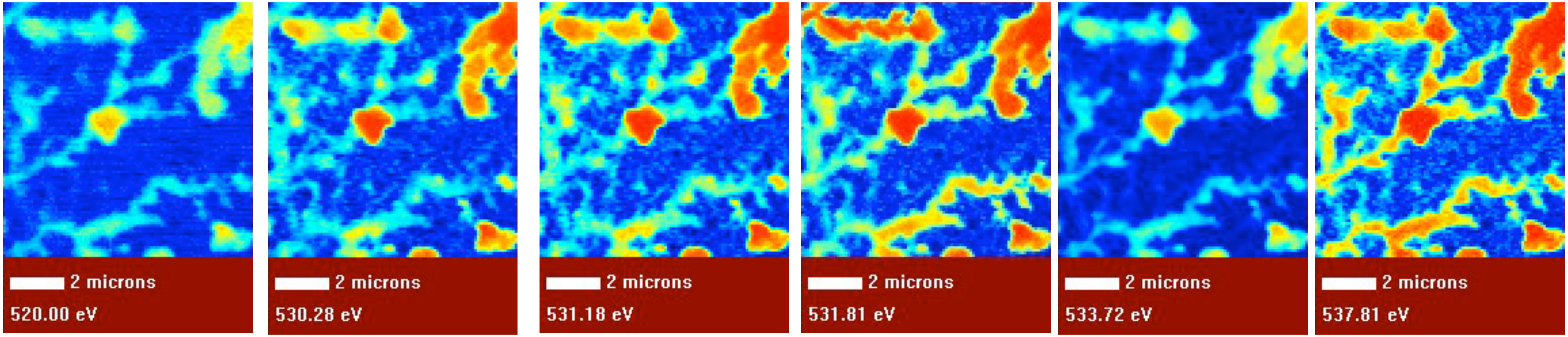
C-XANES of amino acids

- K. Kaznacheyev *et al.*, *J. Phys. Chem.* **A 106**, 3153 (2002)
- Experiment: K. Kaznacheyev *et al.*, Stony Brook (now CLS)
- Theory: O. Plashkevych, H. Ågren *et al.*, KTH Stockholm; A. Hitchcock, McMaster

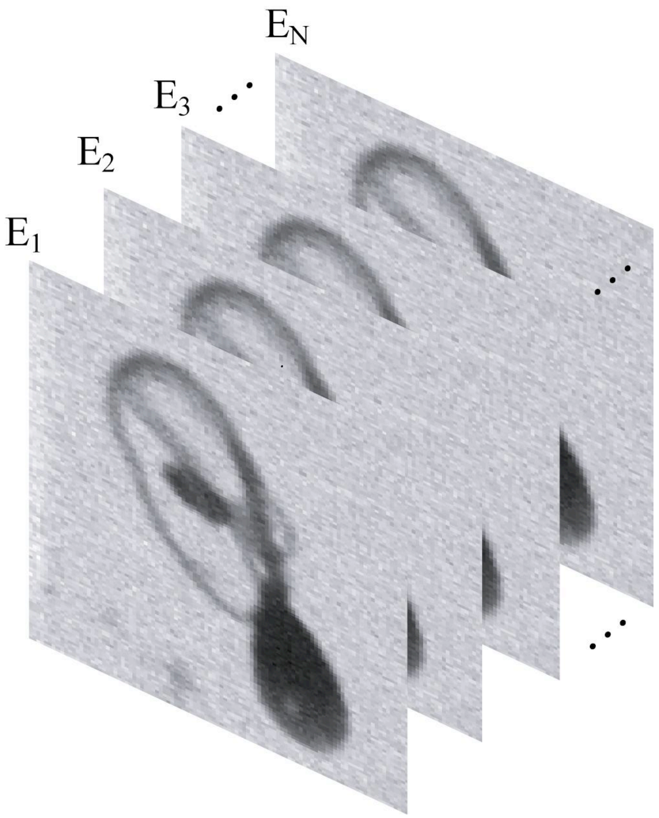


Polymers: see e.g., Dhez, Ade, and Urquhart, *JESRP* **128**, 85 (2003)

Spectromicroscopy: nanoscale heterogeneity



Lu in hematite (T. Schäfer)



Use of XANES for imaging chemical speciation: Ade, Zhang *et al.*, *Science* **258**, 972 (1992).

Aligned spectral image sequences: Jacobsen *et al.*, *J. Microscopy* **197**, 173 (2000)

Spectrum per pixel: spectromicroscopy, spectrum imaging, hyperspectral imaging...

Spectromicroscopy data

- Spectrum per pixel: cube in x, y, and E
- But we will treat pixels as independent, without worrying about spatial correlations: $p=i_x+i_y\cdot N_x$
- Thus we have data in energies $n=1\ldots N$ and pixels $p=1\ldots P$
- We measure a data matrix $D_{N\times P}$:

$$D_{N\times P} = \begin{bmatrix} D_{11} & \text{pixels} & D_{1P} \\ \text{spectra} & & \vdots \\ D_{N1} & \ldots & D_{NP} \end{bmatrix}$$

Spectromicroscopy analysis

We measure the optical density D=μt from I=I_0exp[-μt], which gives us a matrix over n=1..N energies and p=1..P pixels of the data:

$$D_{N\times P} = \begin{bmatrix} D_{11} & \text{pixels} & D_{1P} \\ \text{spectra} & & \vdots \\ D_{N1} & \dots & D_{NP} \end{bmatrix}$$

What we have

We wish we could interpret this in terms of a set of s=1..S components. We would then have a matrix of their spectra

$$\mu_{N\times S} = \begin{bmatrix} \mu_{11} & \text{components} & \mu_{1S} \\ \text{spectra} & & \vdots \\ \mu_{N1} & \dots & \mu_{NS} \end{bmatrix}$$

We would also have a matrix of their thicknesses

$$t_{S\times P} = \begin{bmatrix} t_{11} & \text{pixels} & t_{1P} \\ \text{components} & & \vdots \\ t_{S1} & \dots & t_{SP} \end{bmatrix}$$

What we wish we had

Analysis with known spectra

- Again, data are spectra times thicknesses:

$$\begin{bmatrix} D_{11} & \text{pixels} & D_{1P} \\ \text{spectra} & & \vdots \\ D_{N1} & \ldots & D_{NP} \end{bmatrix} = \begin{bmatrix} \mu_{11} & \text{components} & \mu_{1S} \\ \text{spectra} & & \vdots \\ \mu_{N1} & \ldots & \mu_{NS} \end{bmatrix} \cdot \begin{bmatrix} t_{11} & \text{pixels} & t_{1P} \\ \text{components} & & \vdots \\ t_{S1} & \ldots & t_{SP} \end{bmatrix}$$

or $$D_{N\times P} = \mu_{N\times S} \cdot t_{S\times P}$$

- Example: polymer blend. We may know that we have two or three polymers present, with no reactive phases.
  - Can measure spectra of all components from hand-selected regions
  - We therefore know $\mu_{N\times S}$
  - We can obtain thickness maps (images) by matrix inversion:

$$t_{S\times P} = \mu_{S\times N}^{-1} \cdot D_{N\times P}$$

  - Matrix $\mu_{N\times S}$ of all spectra can be inverted using singular matrix decomposition (SVD). See e.g., Zhang *et al.*, *J. Struct. Biol.* **116**, 335 (1996); Koprinarov *et al.*, *J. Phys. Chem. B* **106**, 5358 (2002).

What if we don’t know the components or their spectra $\mu_{N\times S}$?

- “Natural” specimens, such as in biology or environmental science
- Reactive phases, rather than simple mixing
- Complexity! 300x300 pixel image contains 10^5 spectra!
- Can we find the “organizer” from the data?

The “organizer”: components *S*

- If we know the *s=1…S* components (e.g., known pure compounds) and their spectra, we know $\mu_{N\times S}$ and thus $t_{S\times P} = \mu^{-1}_{S\times N} \cdot D_{N\times P}$

$$\begin{bmatrix} D_{11} & \dots & D_{1P} \\ \vdots & & \vdots \\ D_{N1} & \dots & D_{NP} \end{bmatrix} = \begin{bmatrix} \mu_{11} & \dots & \mu_{1S} \\ \vdots & & \vdots \\ \mu_{N1} & \dots & \mu_{NS} \end{bmatrix} \cdot \begin{bmatrix} t_{11} & \dots & t_{1P} \\ \vdots & & \vdots \\ t_{S1} & \dots & t_{SP} \end{bmatrix}$$

- And if not? Can we uncover an “organizer” for our data anyway? General problem: find *S*!

$$\begin{bmatrix} D_{11} & \dots & D_{1P} \\ \vdots & & \vdots \\ D_{N1} & \dots & D_{NP} \end{bmatrix} = \begin{bmatrix} C_{11} & \dots & C_{1S} \\ \vdots & & \vdots \\ C_{N1} & \dots & C_{NS} \end{bmatrix} \cdot \begin{bmatrix} R_{11} & \dots & R_{1P} \\ \vdots & & \vdots \\ R_{S1} & \dots & R_{SP} \end{bmatrix}$$

Finding *S*: eigenvalues through covariance

- We want to find the "hidden" dimension S
- Form covariance matrices:

$$Z_{N\times N} = D_{N\times P}\cdot D^T_{P\times N} \qquad \text{and} \qquad Z_{P\times P} = D^T_{P\times N}\cdot D_{N\times P}$$

Symmetric, and from same information. If $N = 10^2$ and $P = 10^4$, guess which is quicker to compute?

- Using $Z_{N\times N}$, find $s = 1\cdot N$ eigenvalues $\lambda(s)$:

$$Z_{N\times N}\cdot r(s)_{N\times 1} = \lambda(s)\, r(s)_{N\times 1}$$

- With eigenvalues $\lambda(s)$ over $s = 1\ldots N$, the orthogonal matrix C is formed from the eigenvectors $r(s)_N$:

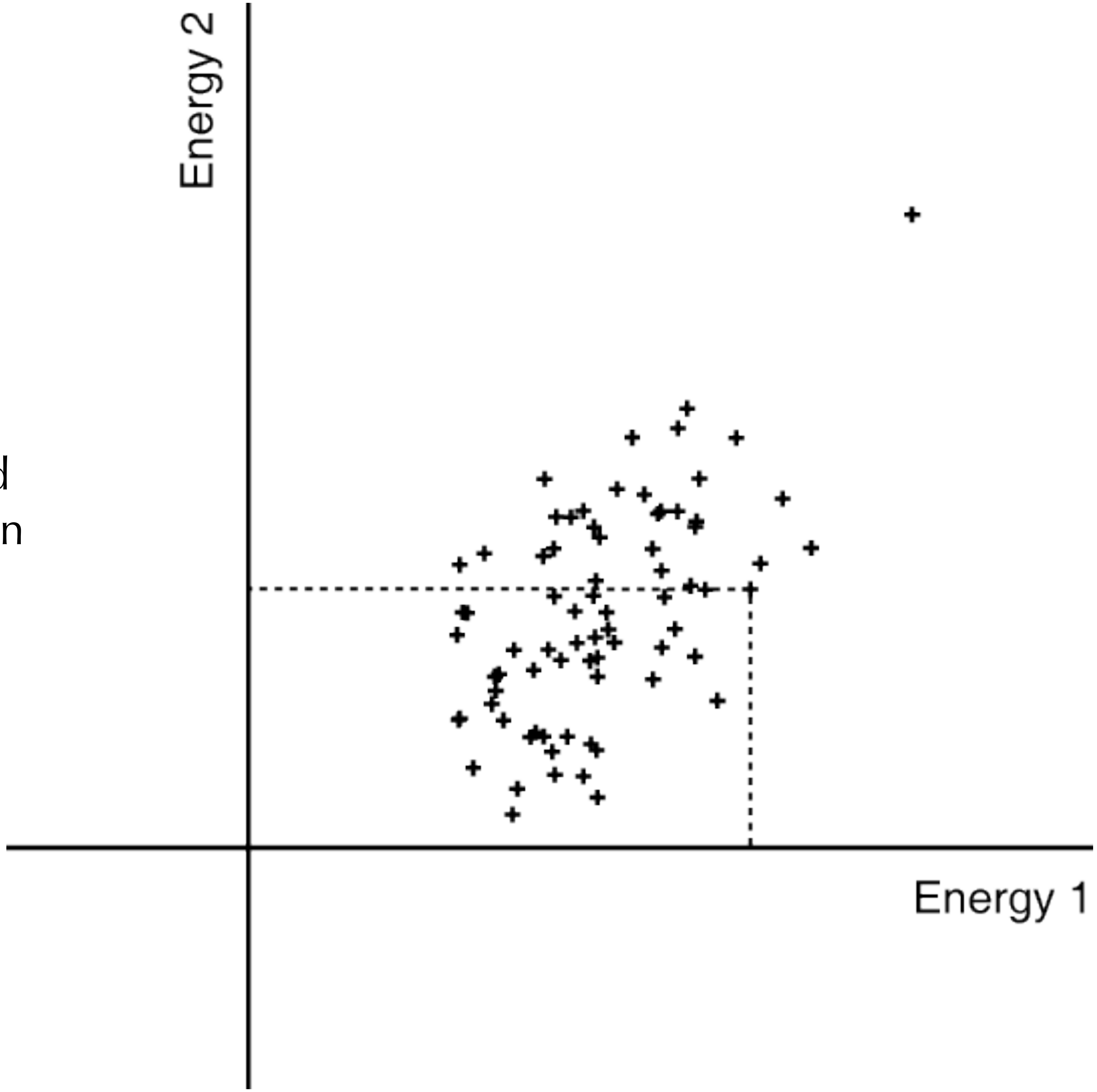
$$C_{N\times S} = \begin{bmatrix} r(1)_1 & \ldots & r(S)_1 \\ \vdots & & \vdots \\ r(1)_N & \ldots & r(S)_N \end{bmatrix}$$

Again, when $C_{N\times S}$ is known, one can calculate $R_{S\times P}$.

Principal component analysis (PCA)

Find set of components *S* that reflect intrinsic properties of the data

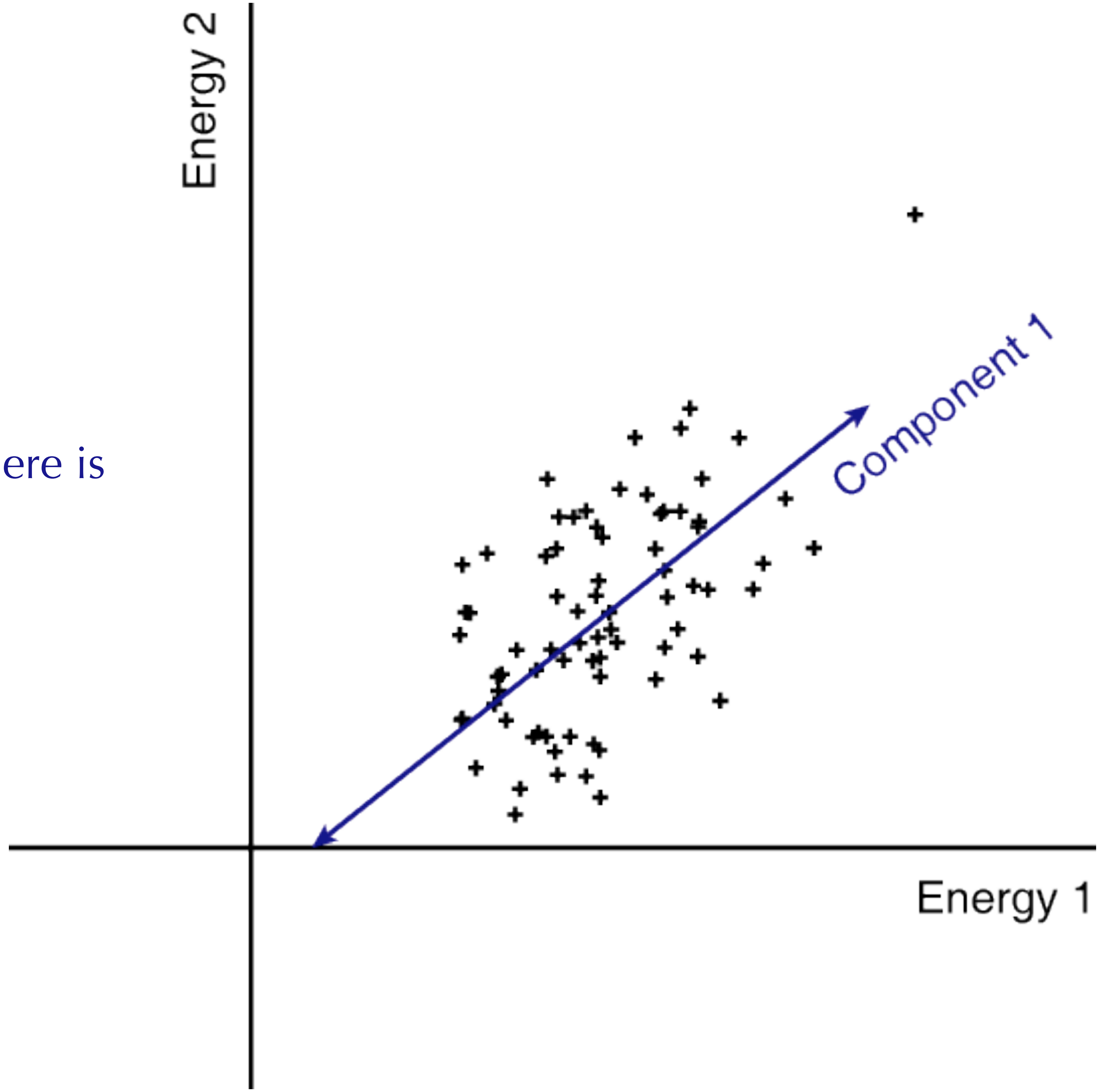
Scatterplot: pixels plotted based on signal at two different photon energies



Principal component analysis (PCA)

Find set of components *S* that reflect intrinsic properties of the data

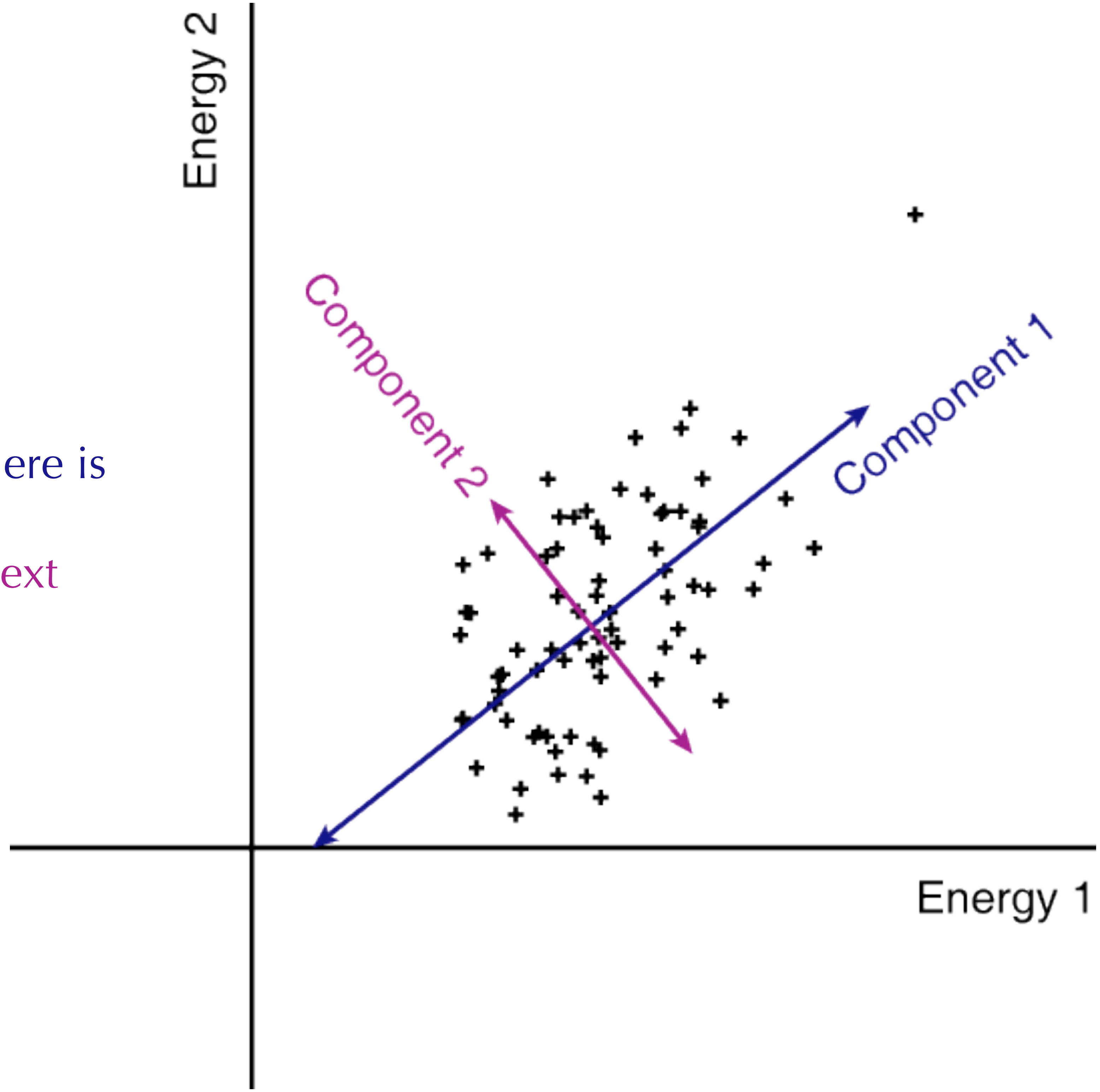
1. Find the axis along which there is the greatest variance



Principal component analysis (PCA)

Find set of components *S* that reflect intrinsic properties of the data

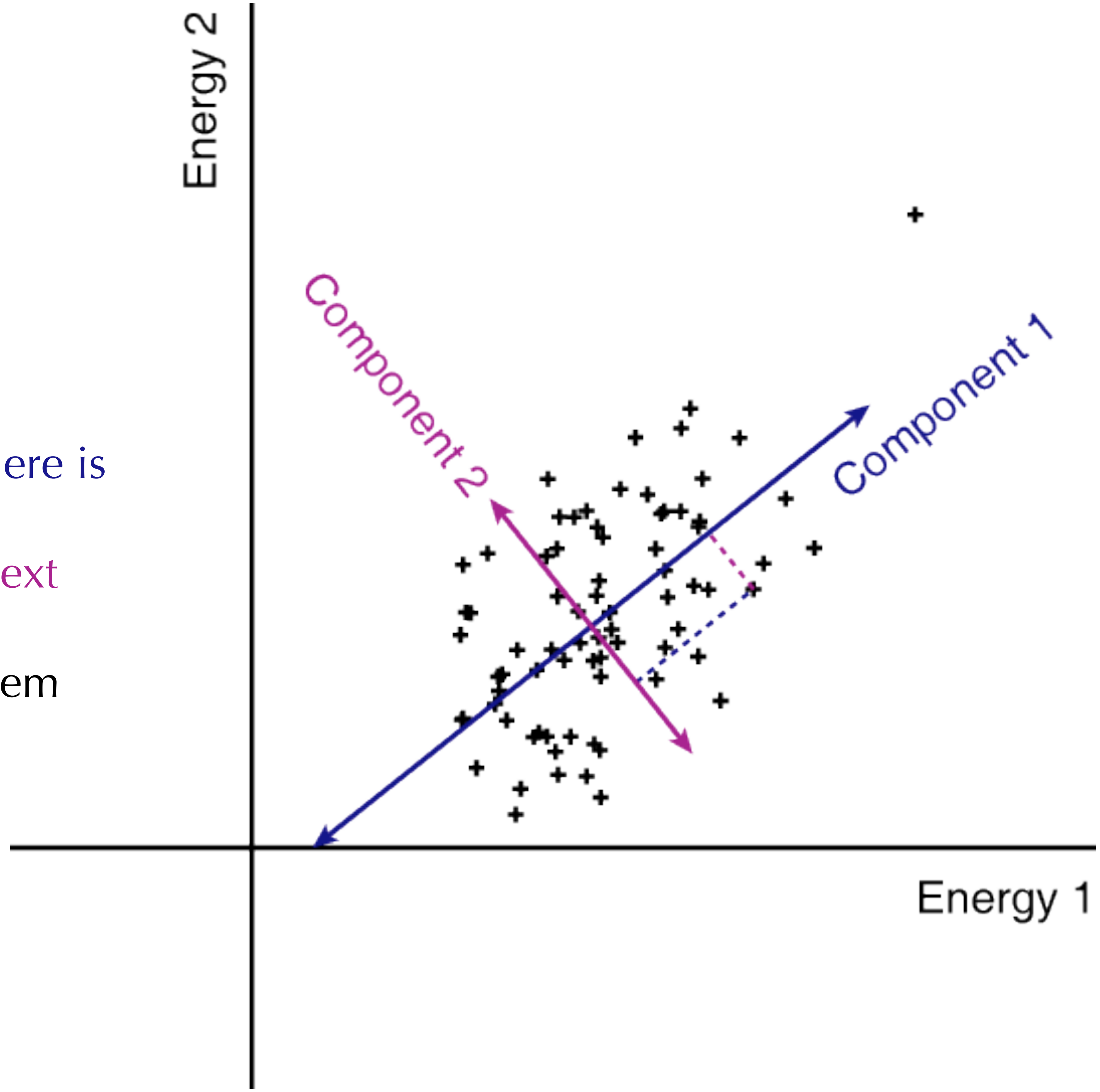
1. Find the axis along which there is the greatest variance
2. Find an orthogonal axis of next greatest variance



Principal component analysis (PCA)

Find set of components *S* that reflect intrinsic properties of the data

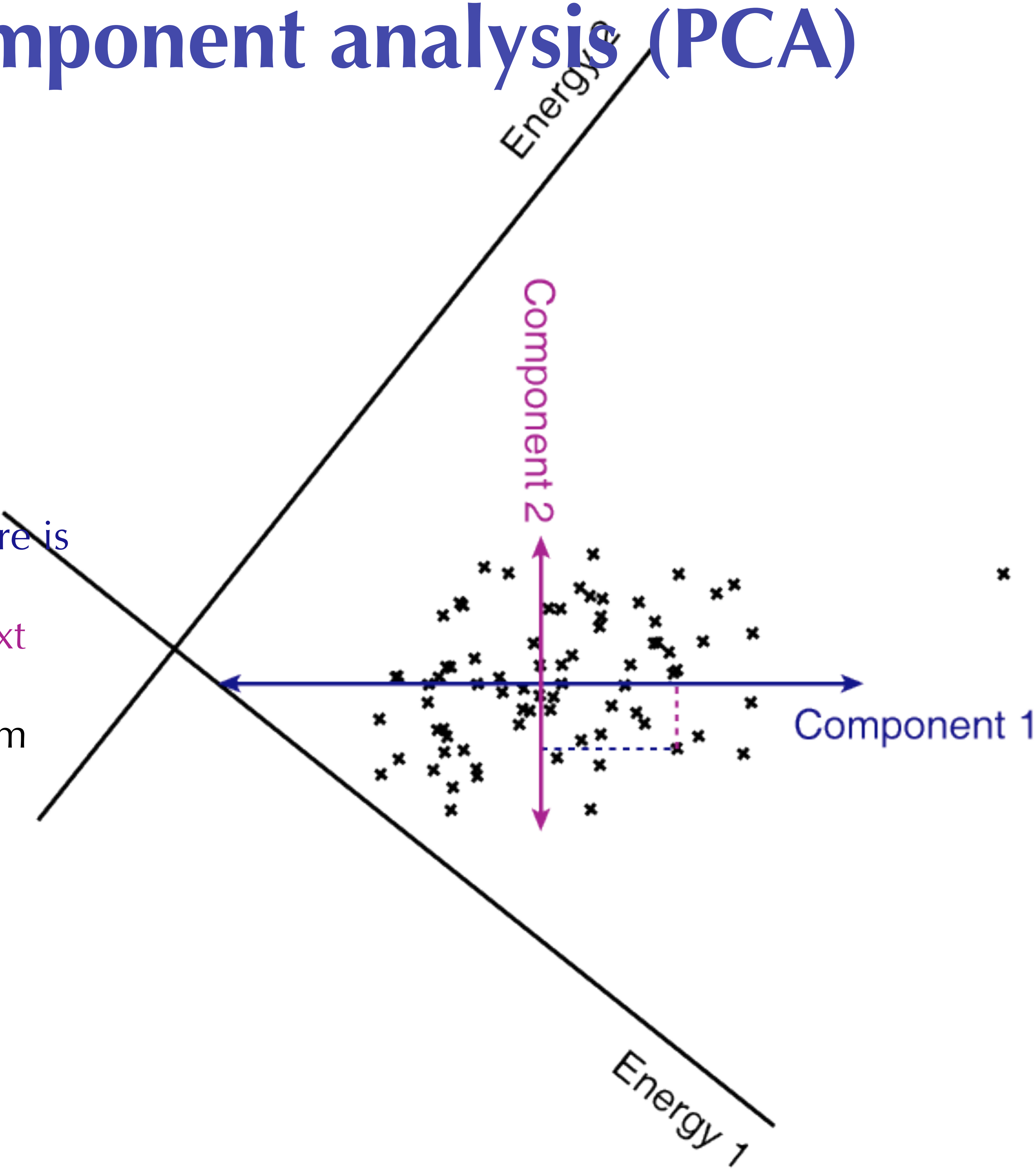
1. Find the axis along which there is the greatest variance
2. Find an orthogonal axis of next greatest variance
3. Gives a new coordinate system



Principal component analysis (PCA)

Find set of components *S* that reflect intrinsic properties of the data

1. Find the axis along which there is the greatest variance
2. Find an orthogonal axis of next greatest variance
3. Gives a new coordinate system
4. Rotate onto new, orthogonal coordinate system



Are eigenvalues *S* enough? Are we done?

- We can find eigenvalues which give us *one* way to find an "organizer" *S*.

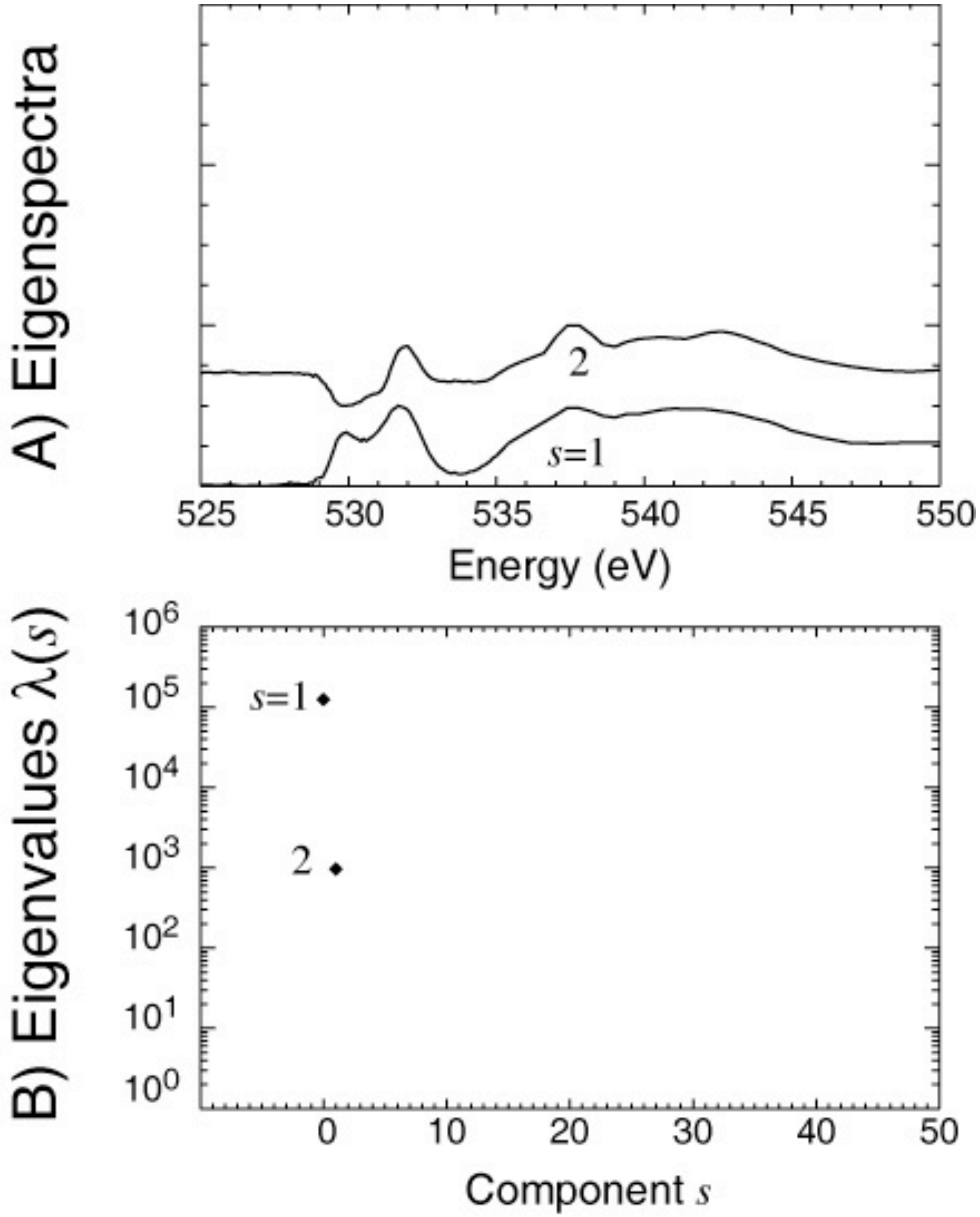
$$\begin{bmatrix} D_{11} & \dots & D_{1P} \\ \vdots & & \vdots \\ D_{N1} & \dots & D_{NP} \end{bmatrix} = \begin{bmatrix} C_{11} & \dots & C_{1S} \\ \vdots & & \vdots \\ C_{N1} & \dots & C_{NS} \end{bmatrix} \cdot \begin{bmatrix} R_{11} & \dots & R_{1P} \\ \vdots & & \vdots \\ R_{S1} & \dots & R_{SP} \end{bmatrix}$$

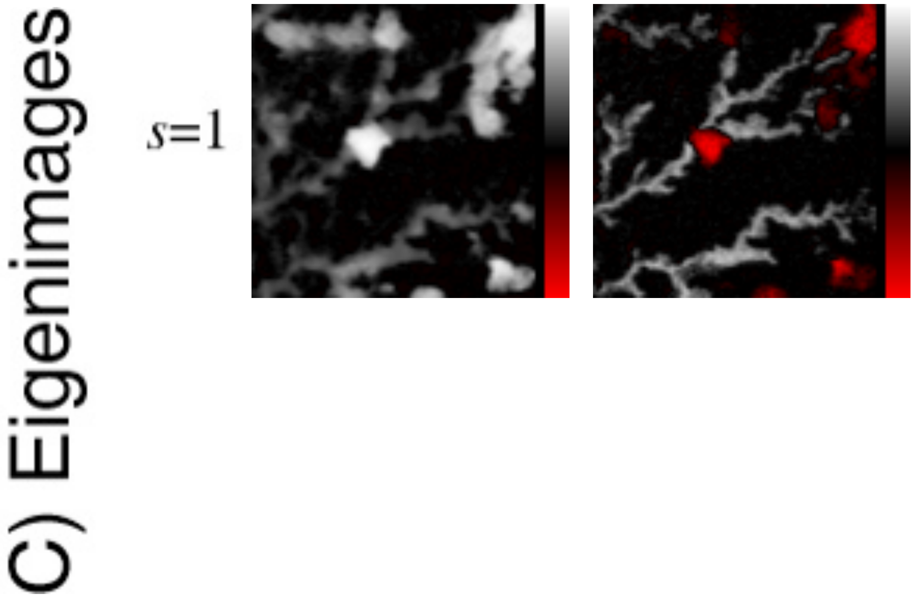
PCA in spectromicroscopy: King *et al.*, *J. Vac. Sci. Tech. A* **7**, 3301 (1989); A. Osanna & C. Jacobsen, XRM99 proceedings; Bonnet *et al.*, *Ultramicroscopy* **77**, 97 (1999).

- But is it the right "organizer" *S*?

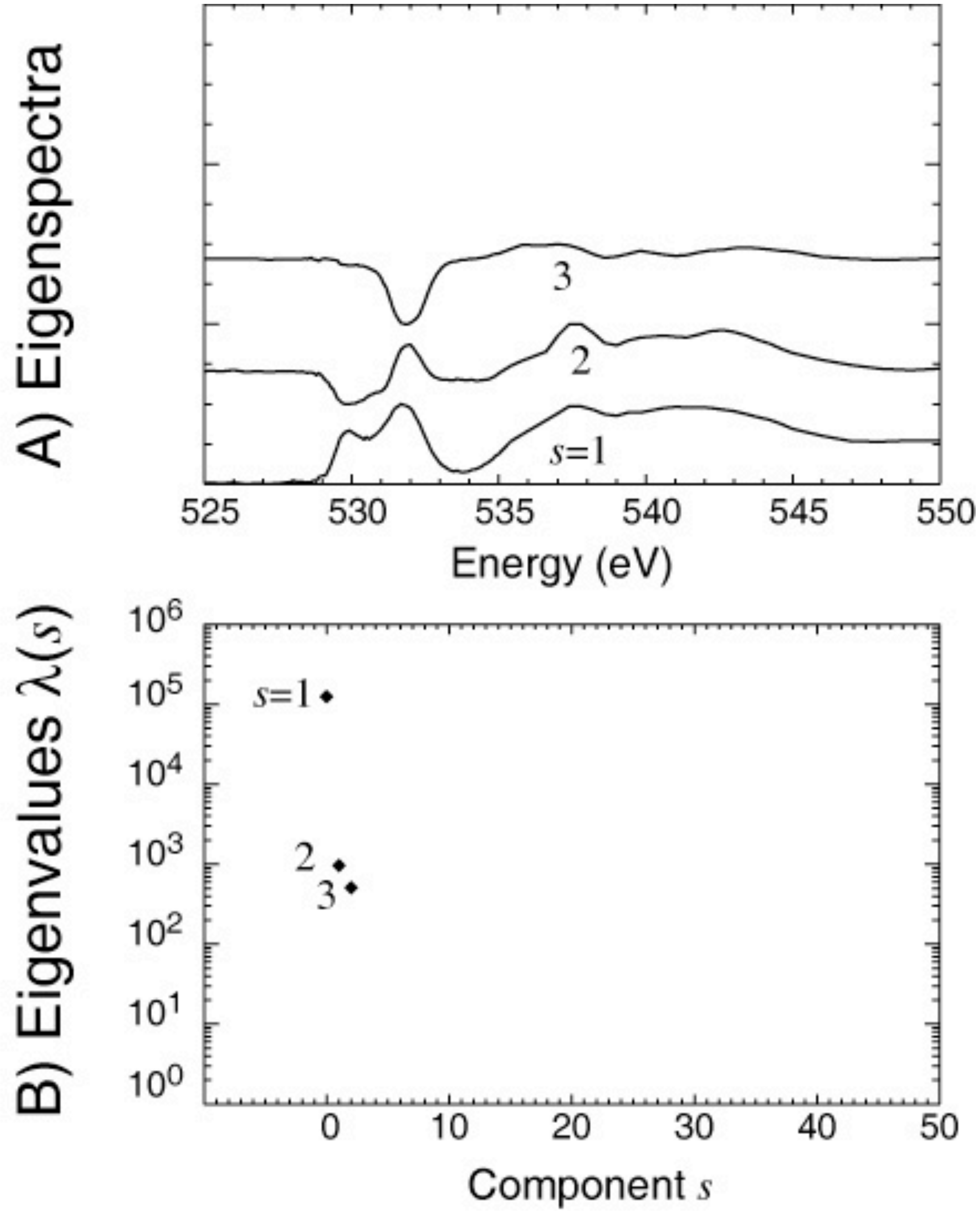
Eigenspectra and eigenimages

Find reduced number of **significant** components $\bar{S}_{abstract}$

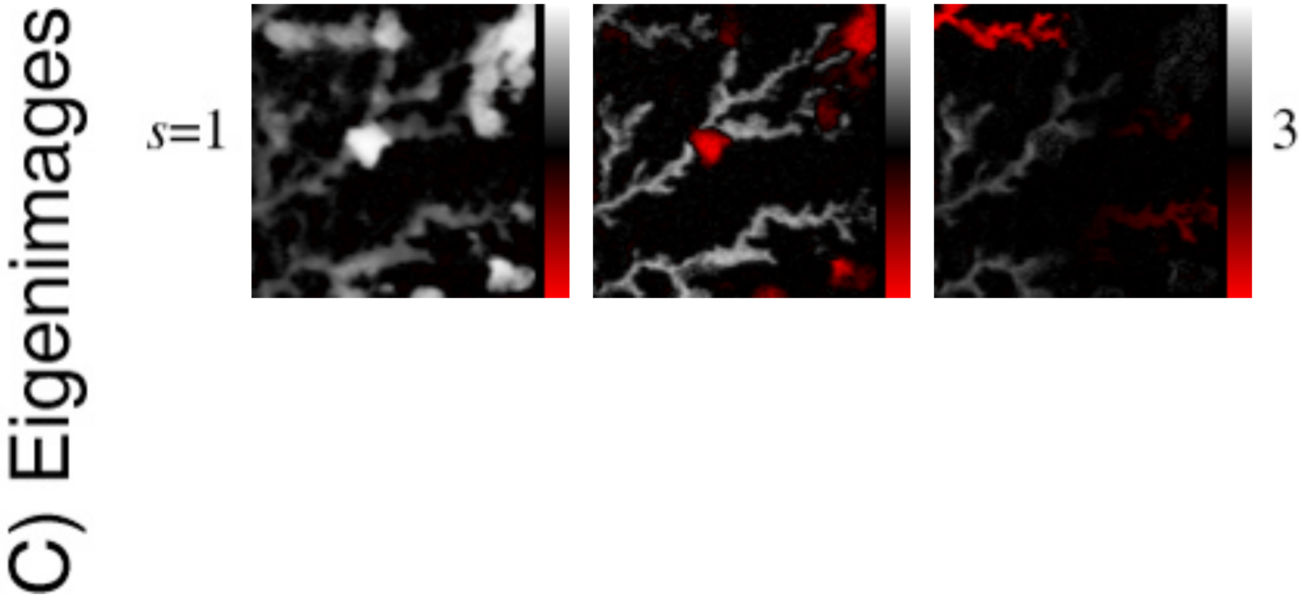




Eigenspectra and eigenimages

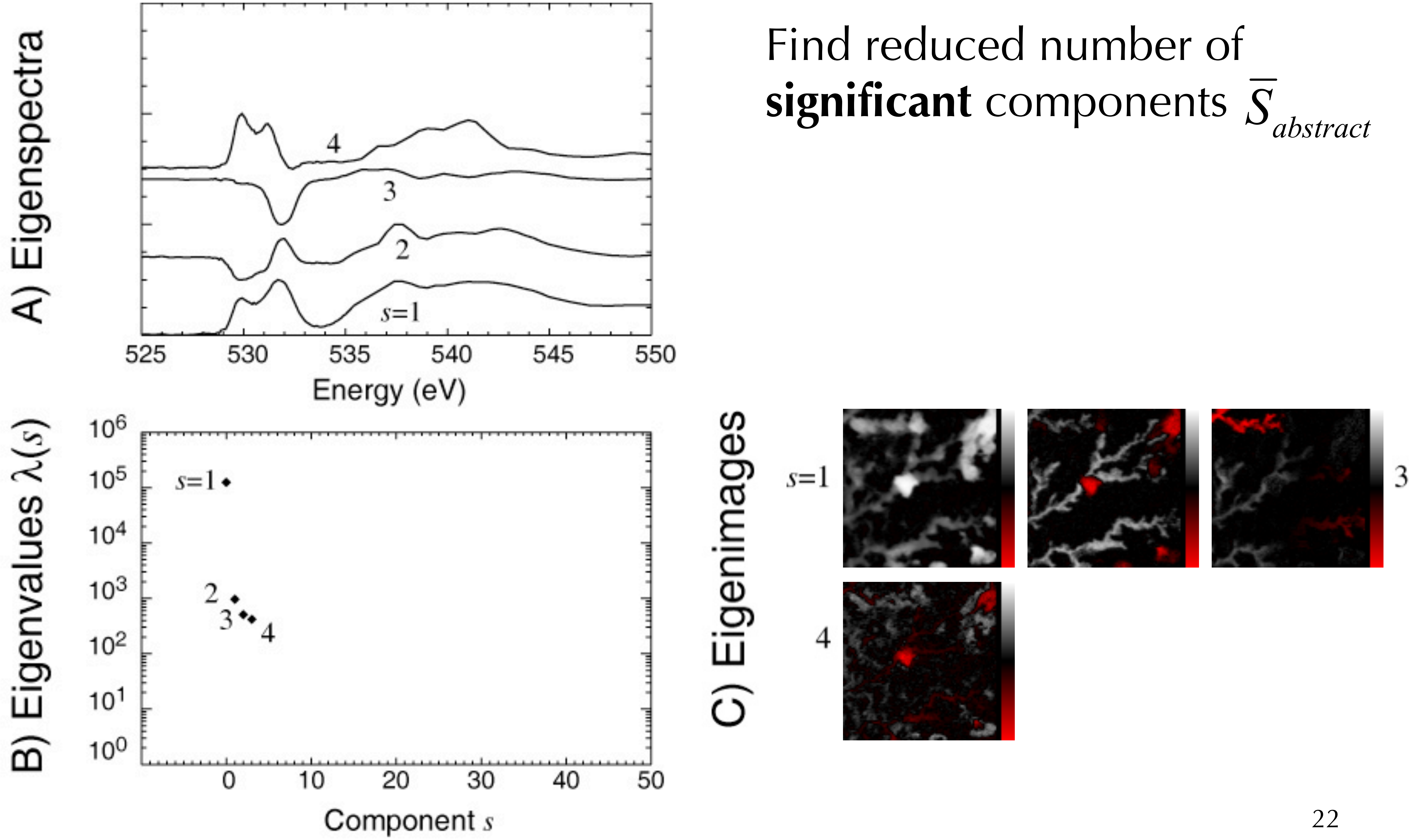


Find reduced number of **significant** components $\bar{S}_{abstract}$

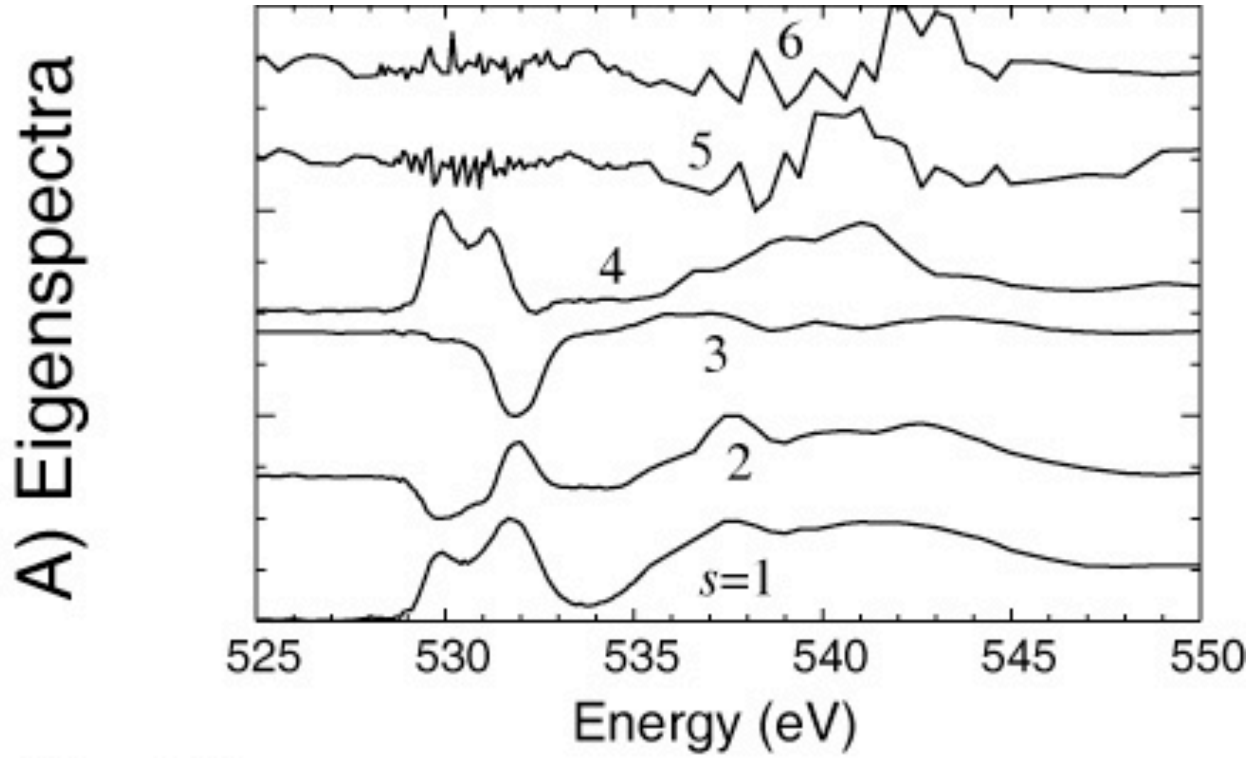


Eigenspectra and eigenimages

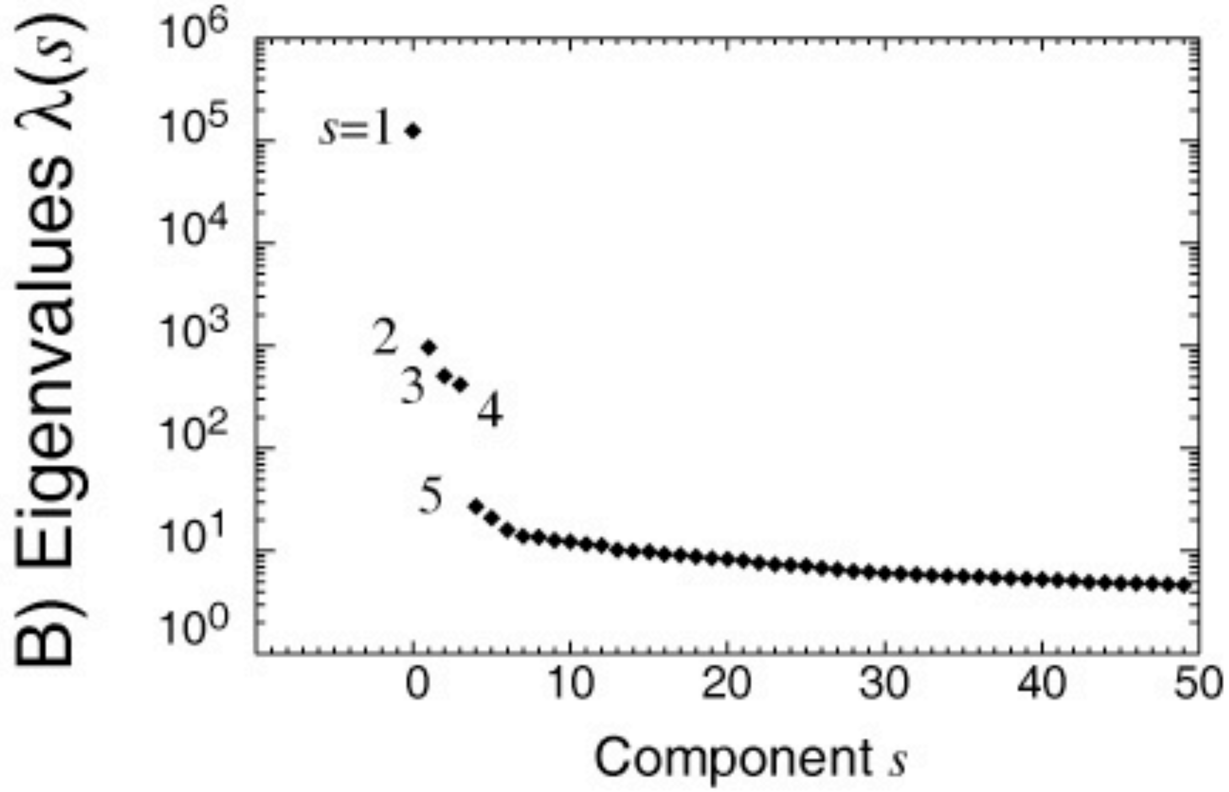
Find reduced number of **significant** components $\overline{S}_{abstract}$

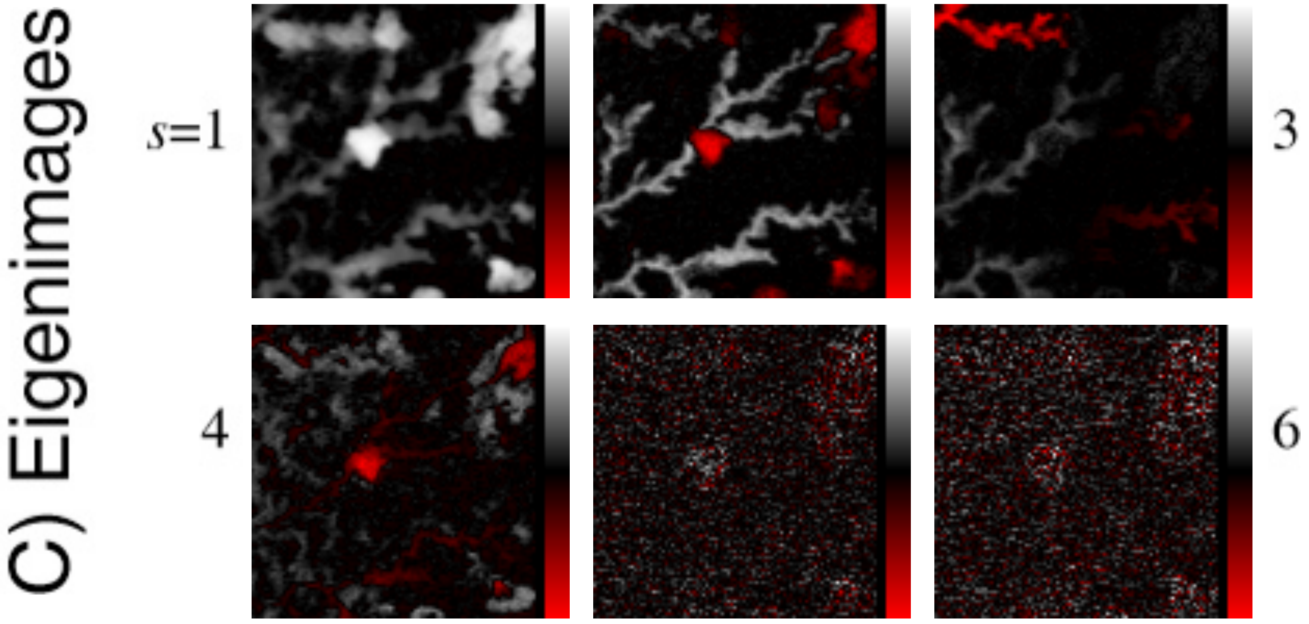


Eigenspectra and eigenimages



Find reduced number of **significant** components $\bar{S}_{abstract}$





Eigenspectra get us *something*…

- Principal component analysis lets you reduce and orthogonalize the data set!
- Reduction: filter out spectral variations that are poorly correlated throughout the dataset (smells like photon noise!). We went from N=140 energies to $\bar{S}_{\text{abstract}}$=4 components.
- Orthogonality might have nice consequences.

But we have a problem…

- Eigenspectra > 1 are abstract. They have negative optical densities, so they are not readily interpretable.

A well-known problem

We chose to follow an approach which is well known in the literature:

"You can't always get what you want; but if you try sometimes, well you just might find you get what you need"

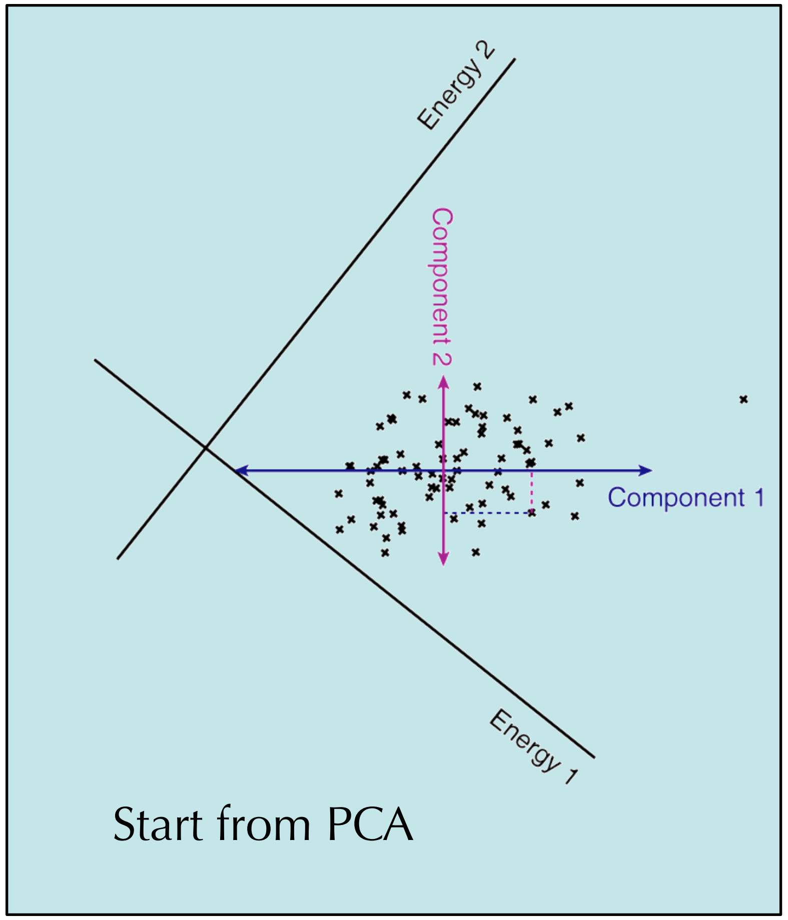
M. Jagger, K. Richards *et al.*, *Let it Bleed* **1**, 1 (1969)

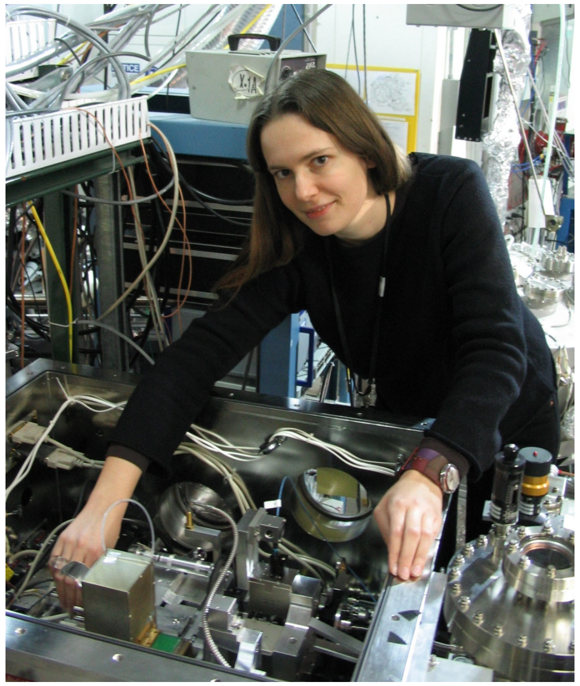
Finding *useful* organizers *S*

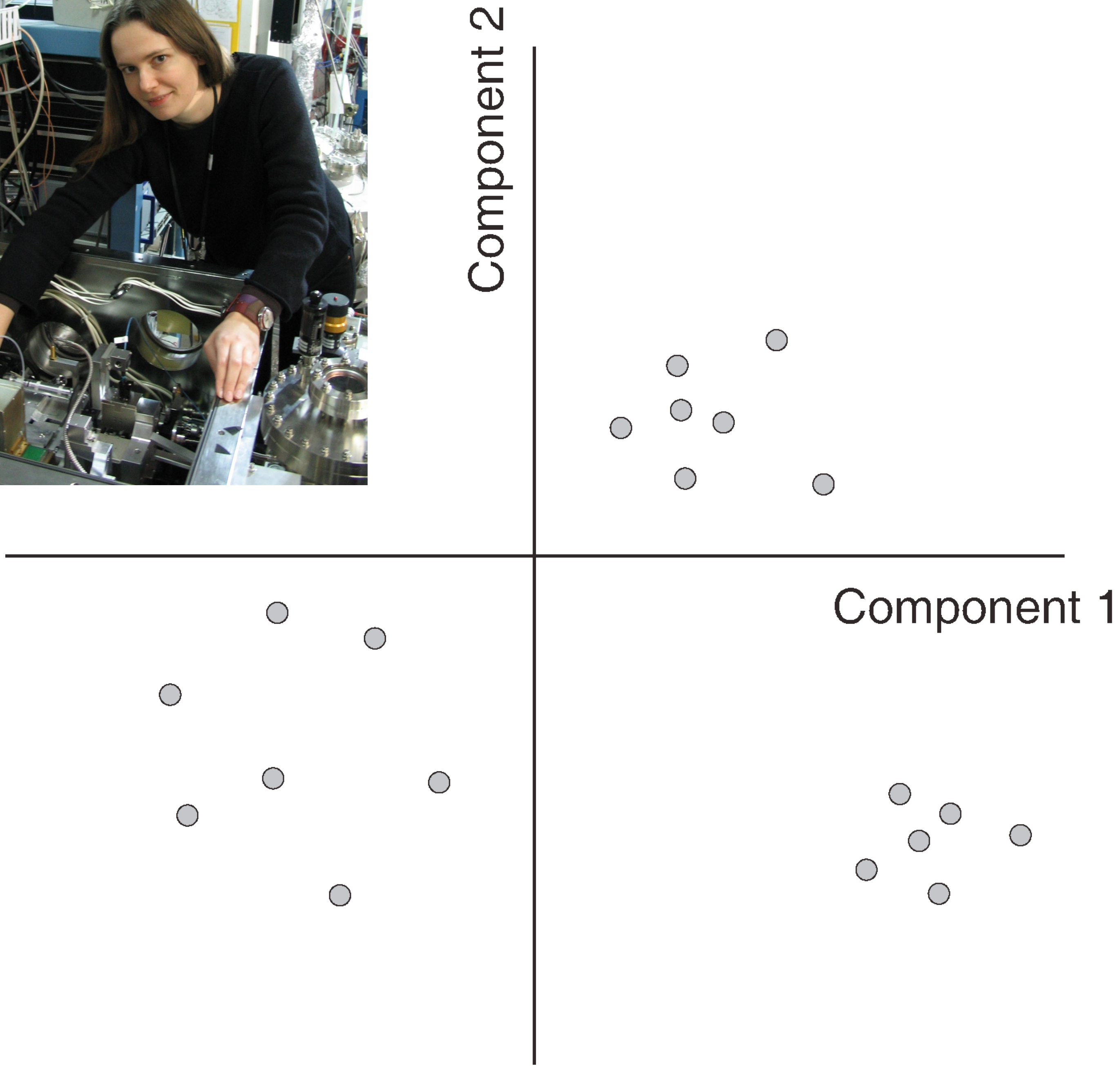
- Cluster analysis: pixels with common spectroscopic signatures yield $C_{N \times S}$
- Varimax: “rotate” $C_{N \times S}$ to make all spectra positive. Works well with discrete elemental signals in fluorescence, TOF-SIMS (P. Kotula, Sandia Labs).
- Non-negative matrix factorization: build $C_{N \times S}$ from noise, constraining for positivity and a minimum set of *S*

Cluster analysis: Euclidian distance learning algorithm

- Kohonen, *Proc. IEEE* **78**, 1464 (1990)
- Pixels are scattered according to weighting of each component

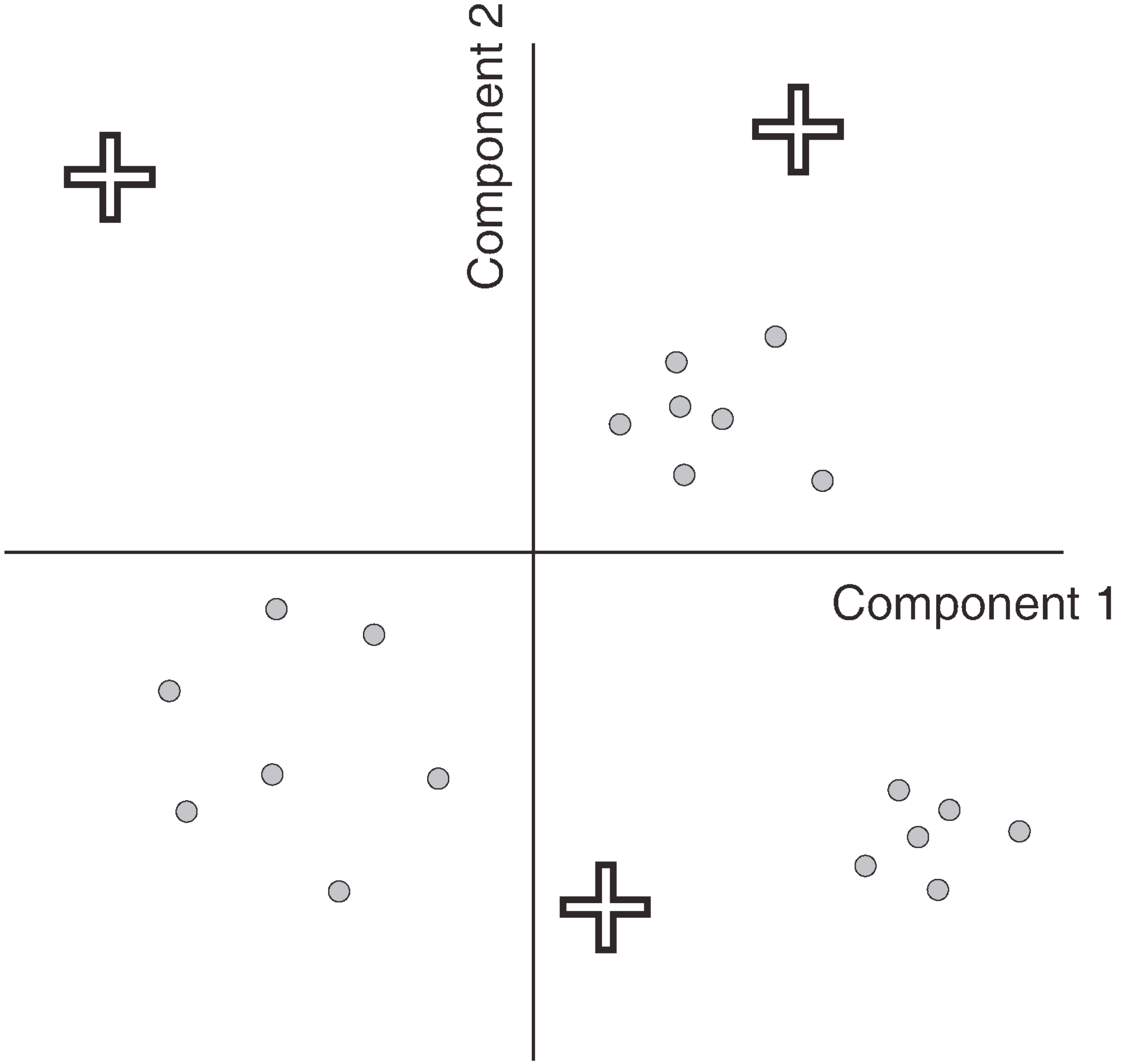






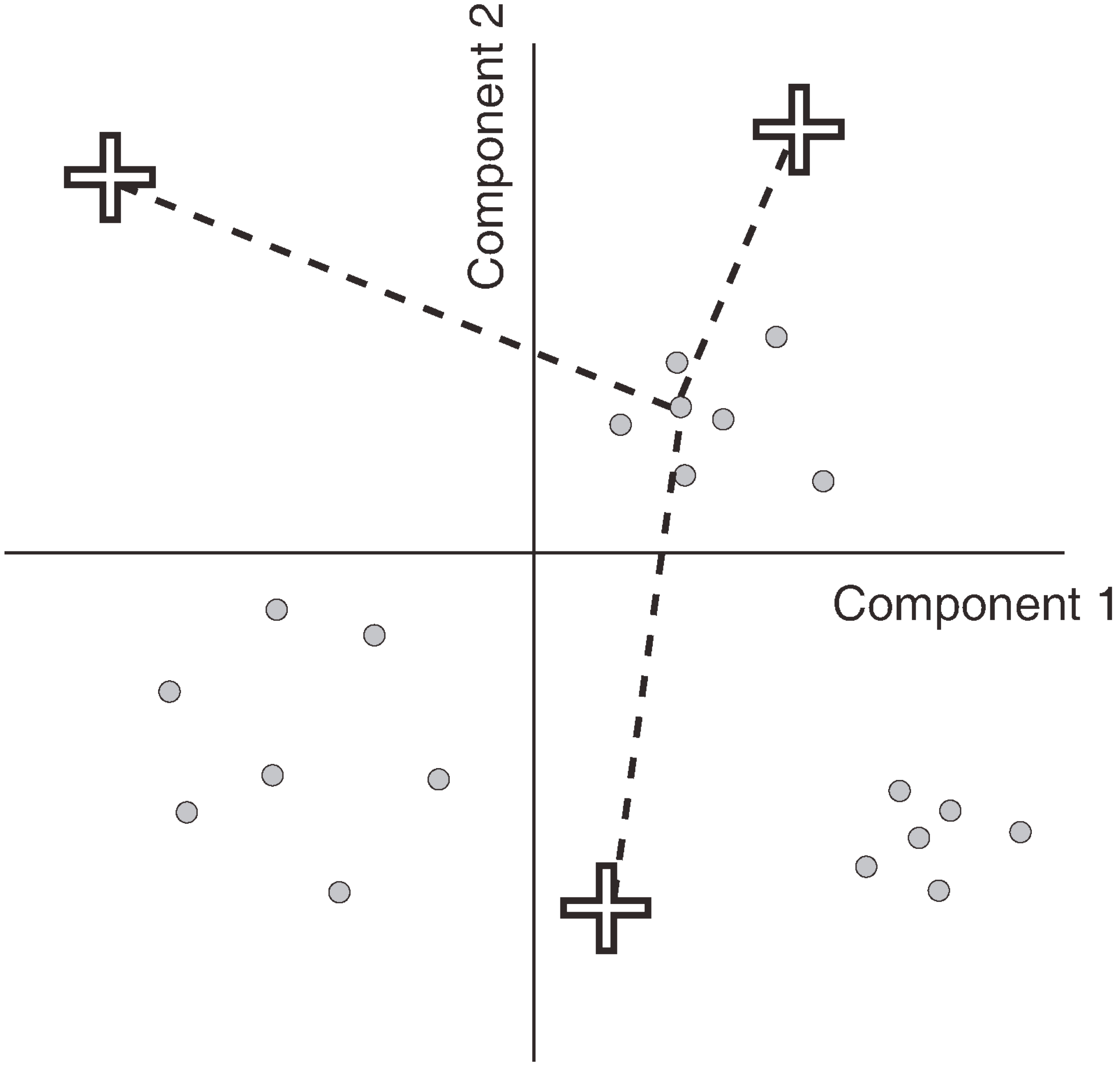
Cluster analysis: Euclidian distance learning algorithm

- Kohonen, *Proc. IEEE* **78**, 1464 (1990)
- Pixels are scattered according to weighting of each component
- Put down cluster centers at random positions.



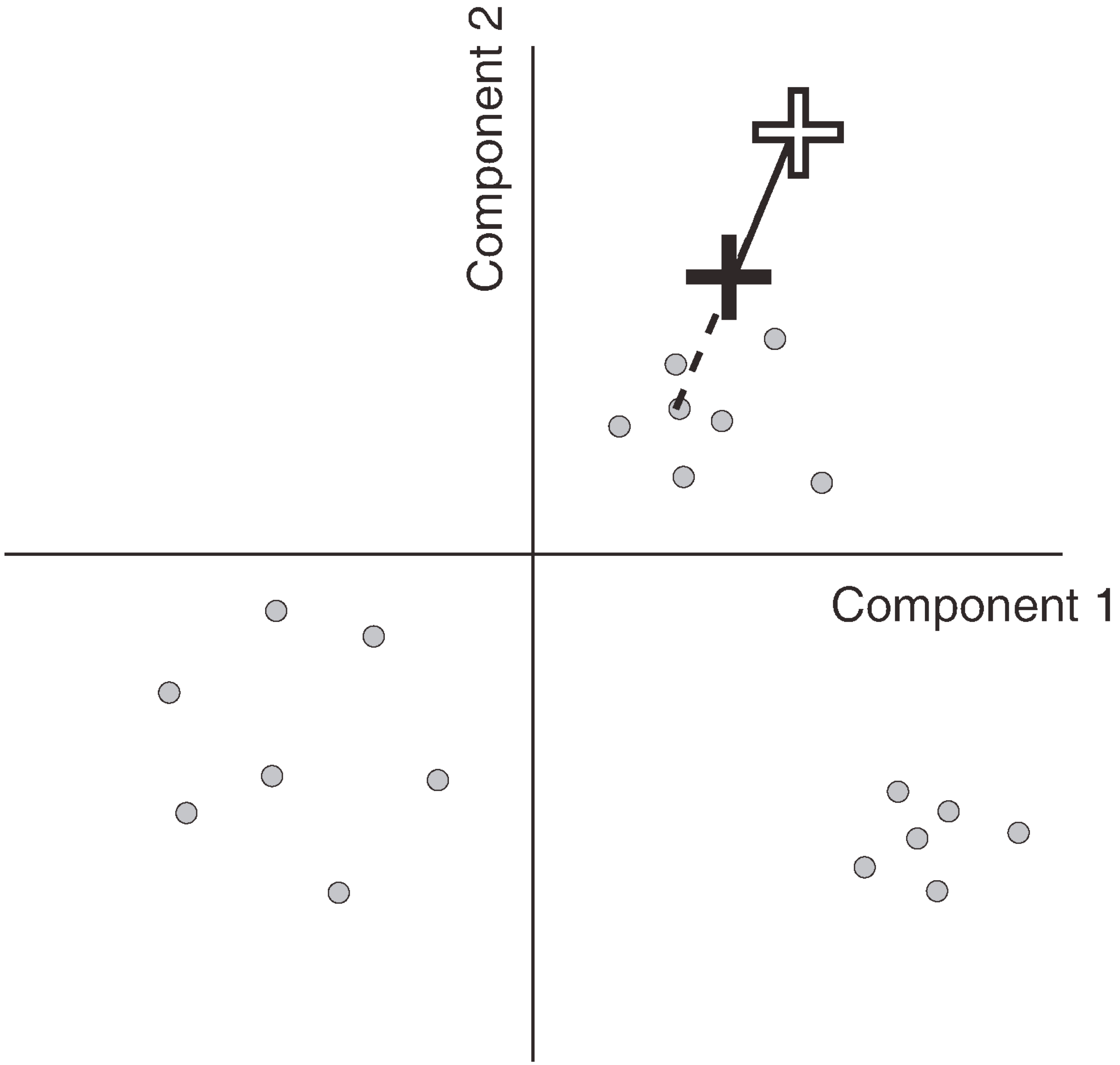
Cluster analysis:
Euclidian distance learning algorithm

- Kohonen, *Proc. IEEE* **78**, 1464 (1990)
- Pixels are scattered according to weighting of each component
- Put down cluster centers at random positions.
- Iterate through all pixels, several times:
  - Calculate distances from one pixel to all cluster centers.



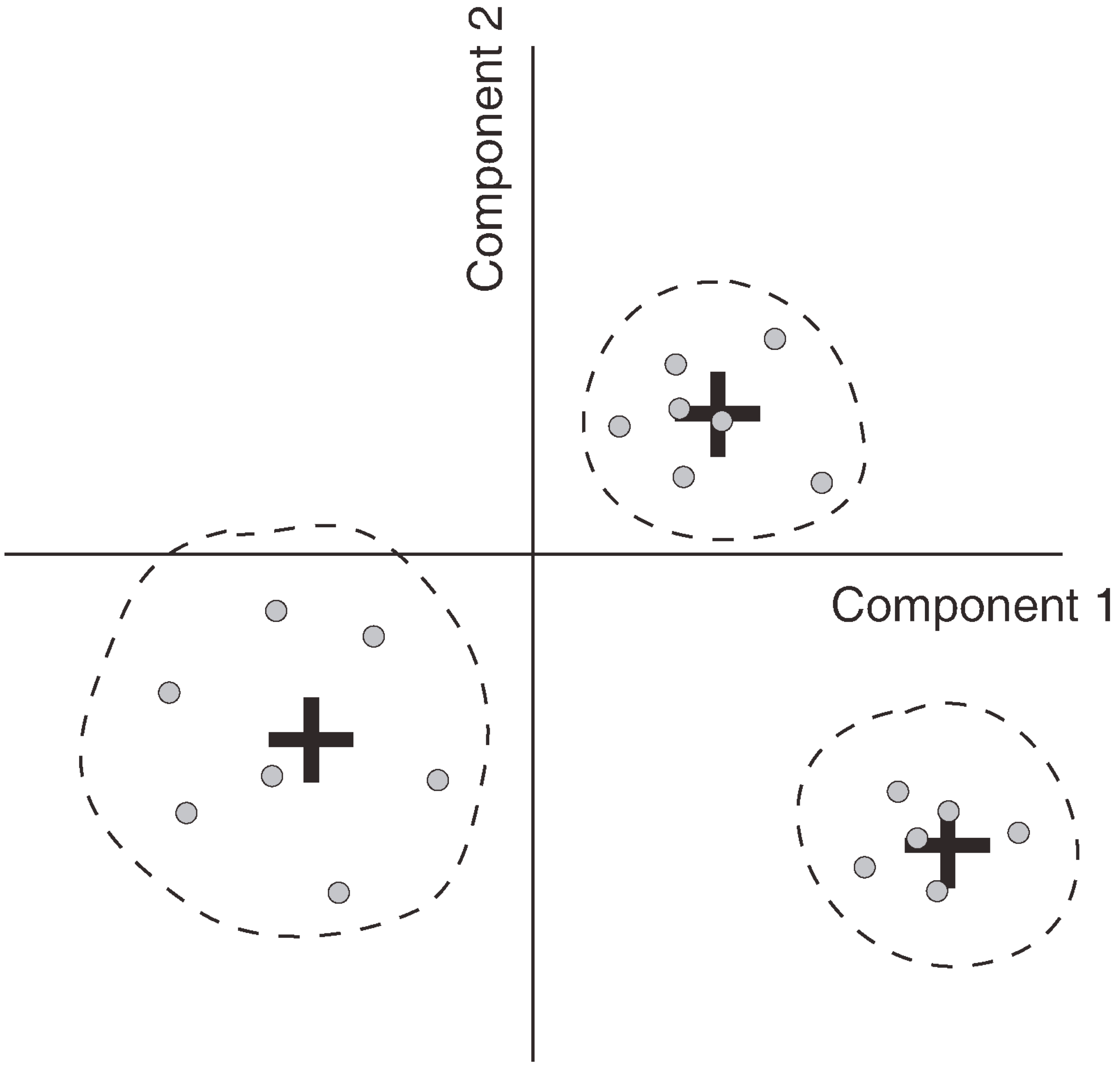
Cluster analysis: Euclidian distance learning algorithm

- Kohonen, *Proc. IEEE* **78**, 1464 (1990)
- Pixels are scattered according to weighting of each component
- Put down cluster centers at random positions.
- Iterate through all pixels, several times:
  - Calculate distances from one pixel to all cluster centers.
  - Pick shortest distance.
  - Move cluster center partway to pixel.



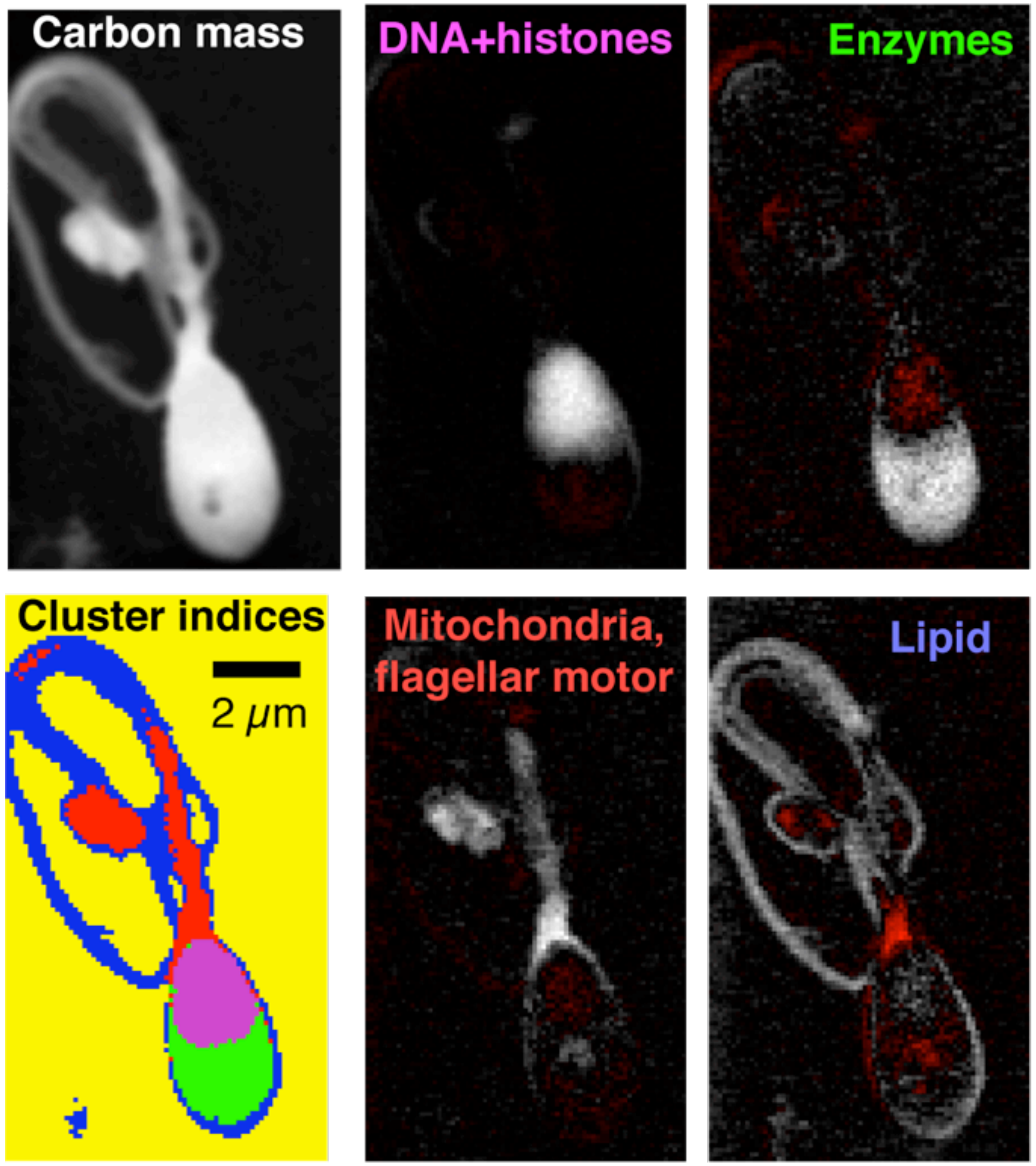
Cluster analysis: Euclidian distance learning algorithm

- Kohonen, *Proc. IEEE* **78**, 1464 (1990)
- Pixels are scattered according to weighting of each component
- Put down cluster centers at random positions.
- Iterate through all pixels, several times:
  - Calculate distances from one pixel to all cluster centers.
  - Pick shortest distance.
  - Move cluster center partway to pixel.
- Cluster pixels with their nearest cluster center

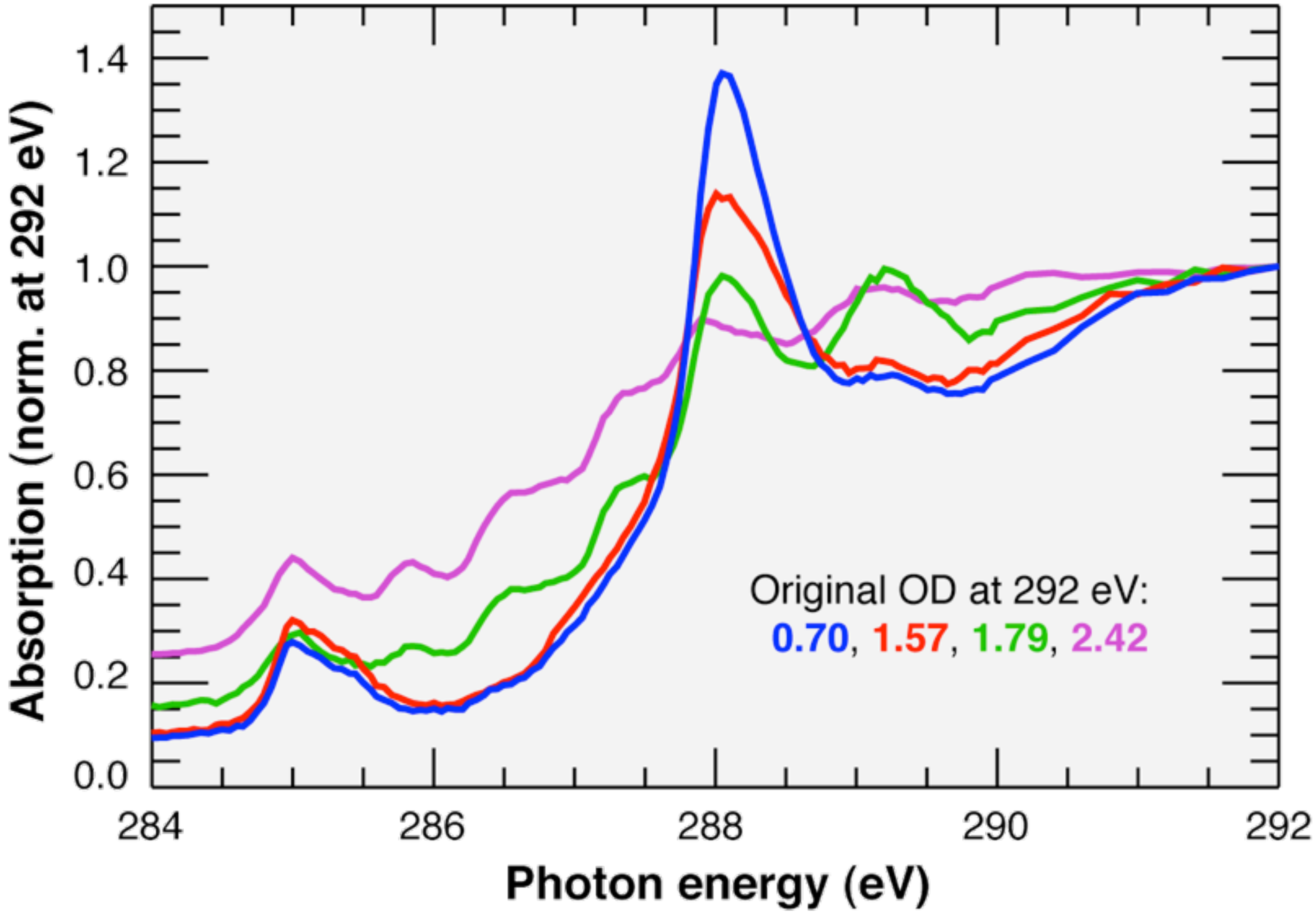


Cluster analysis: human sperm

Biochemical organization of sperm revealed directly from data: enzyme-rich region, DNA+histones, mitochondria and flagellar motor, lipid



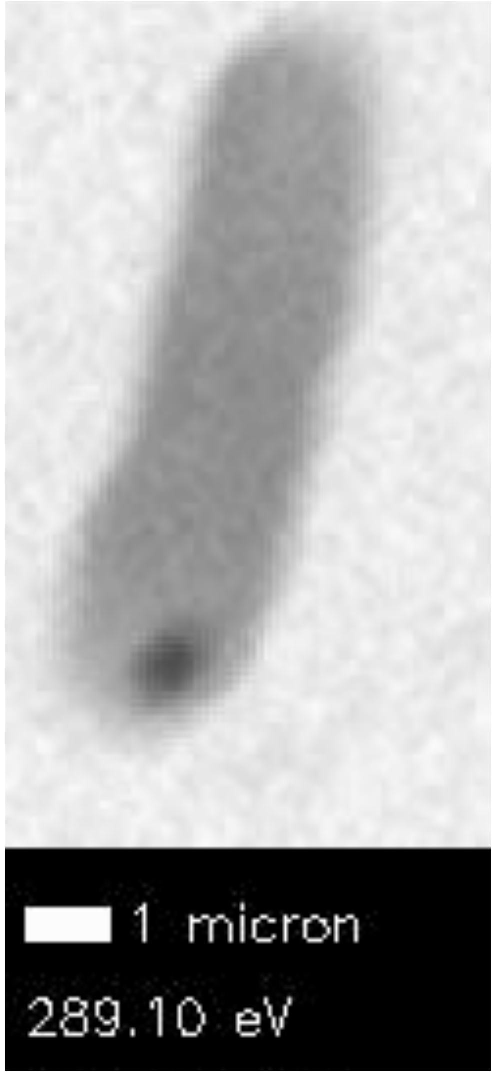
H. Fleckenstein, M. Lerotic, Y. Sheynkin *et al*., Stony Brook. Human sperm, air-dried.



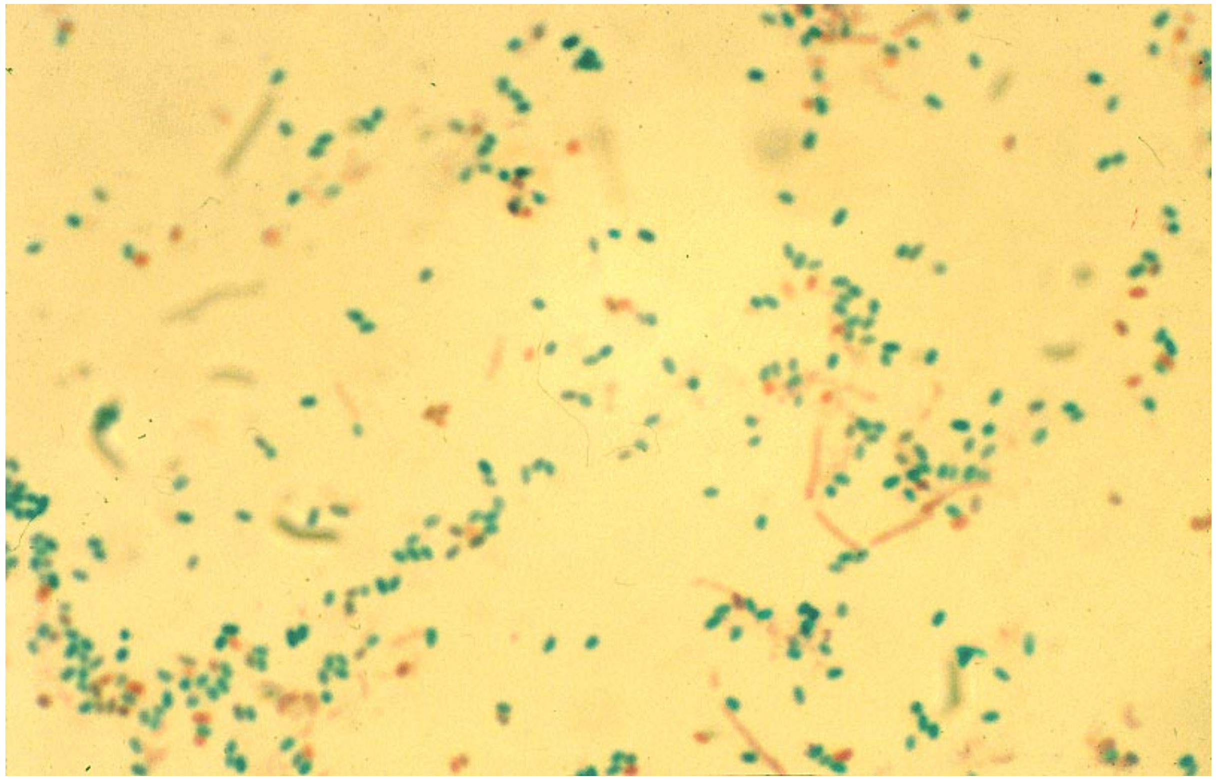
Red spots: negative values, because cluster spectra are not guaranteed to "span the set"

Bacterial Sub-Cellular Features

Studies of uranium reduction by Clostridium sp. (B. Larson, Stony Brook; J.B. Gillow, A.J. Francis, BNL Applied Science)



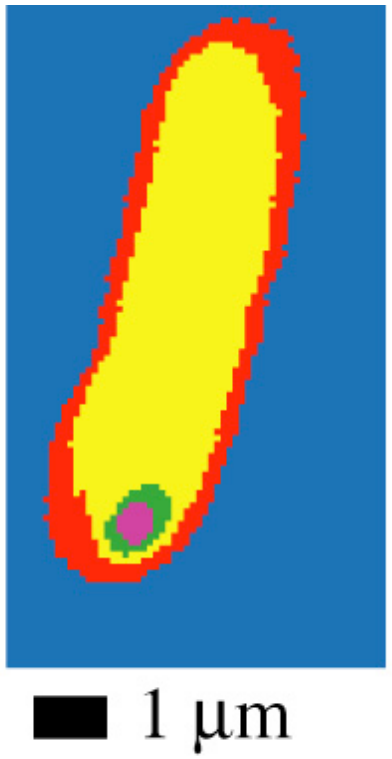
Clostridium sp.

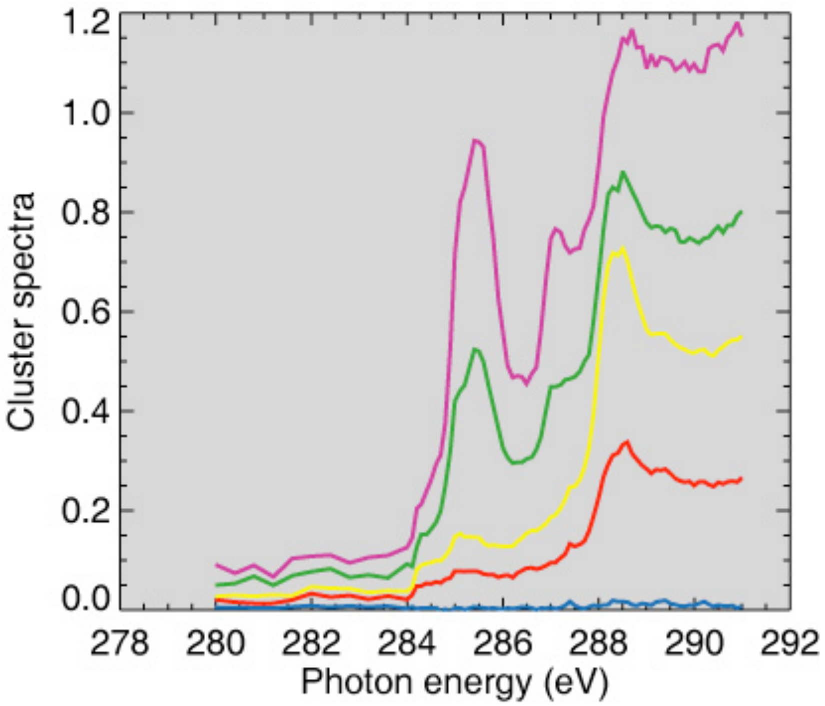


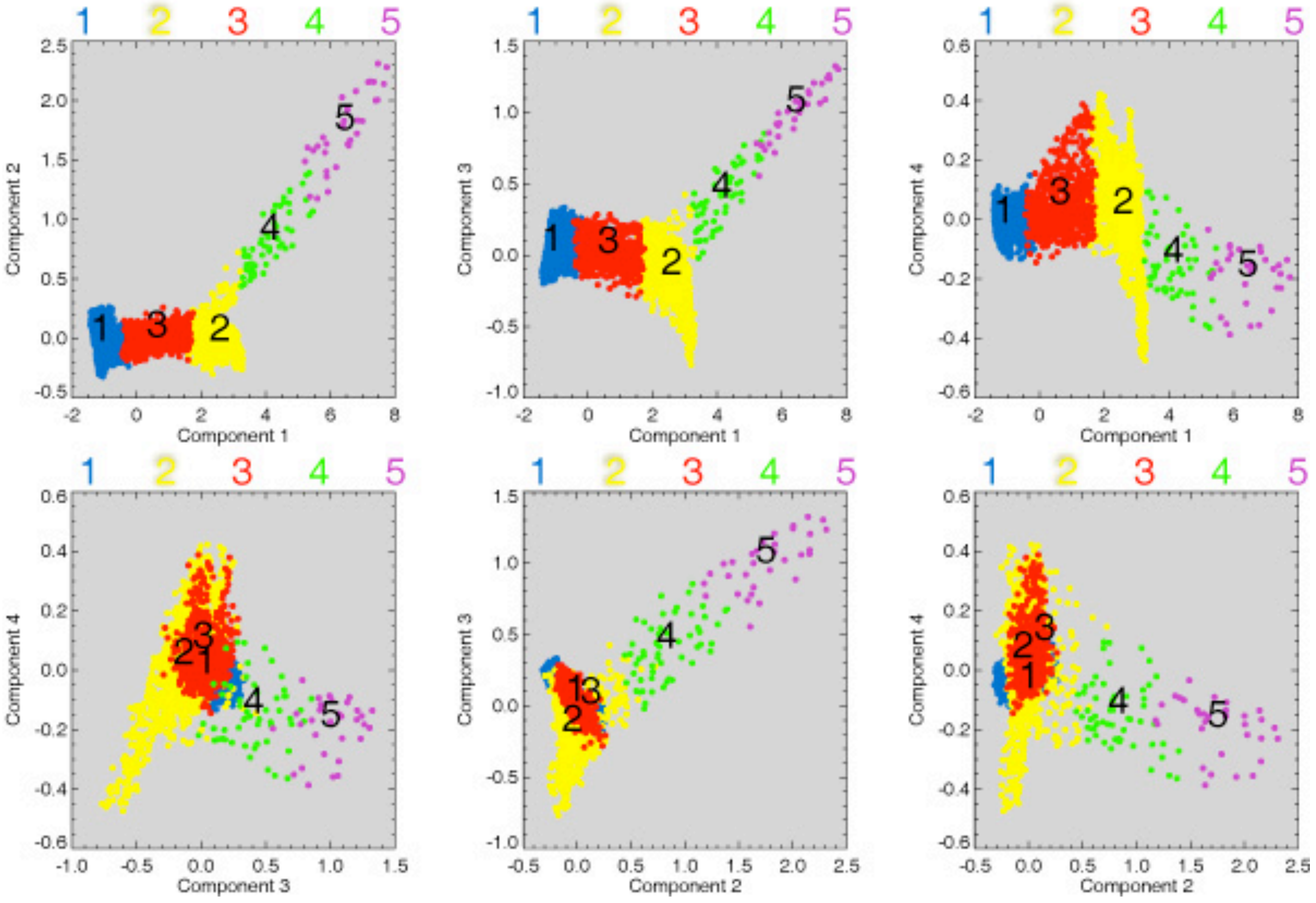
Sporulation

Great! But thickness versus chemistry?

- Scatterplots: each dot is a pixel. Can only show weighting in two components at a time in a 2D plot.
- Bacterium Clostridium sp. (J.B. Gillow, A.J. Francis)
- For some samples, clustering is dominated by thickness variations rather than spectral differences!

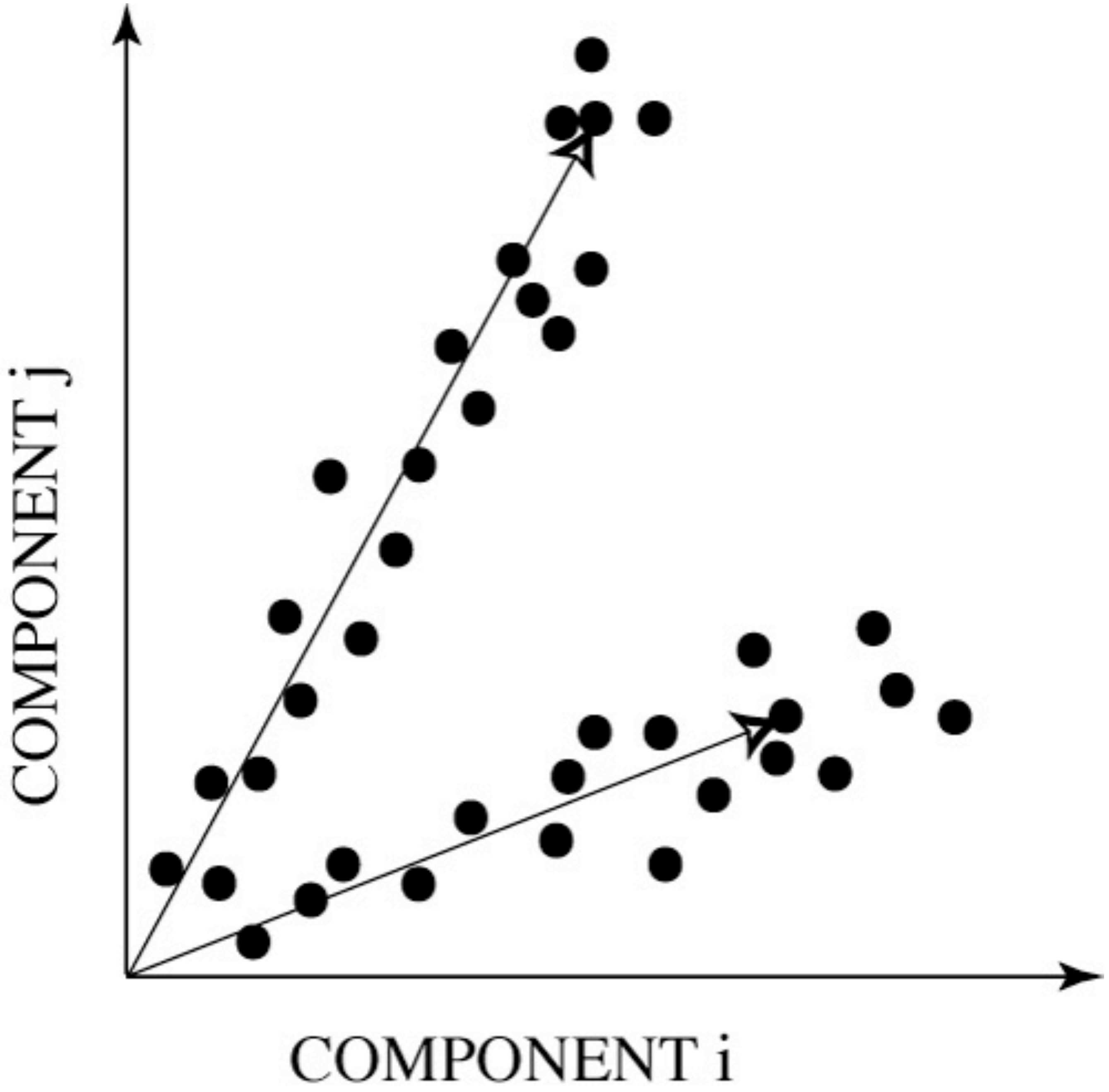






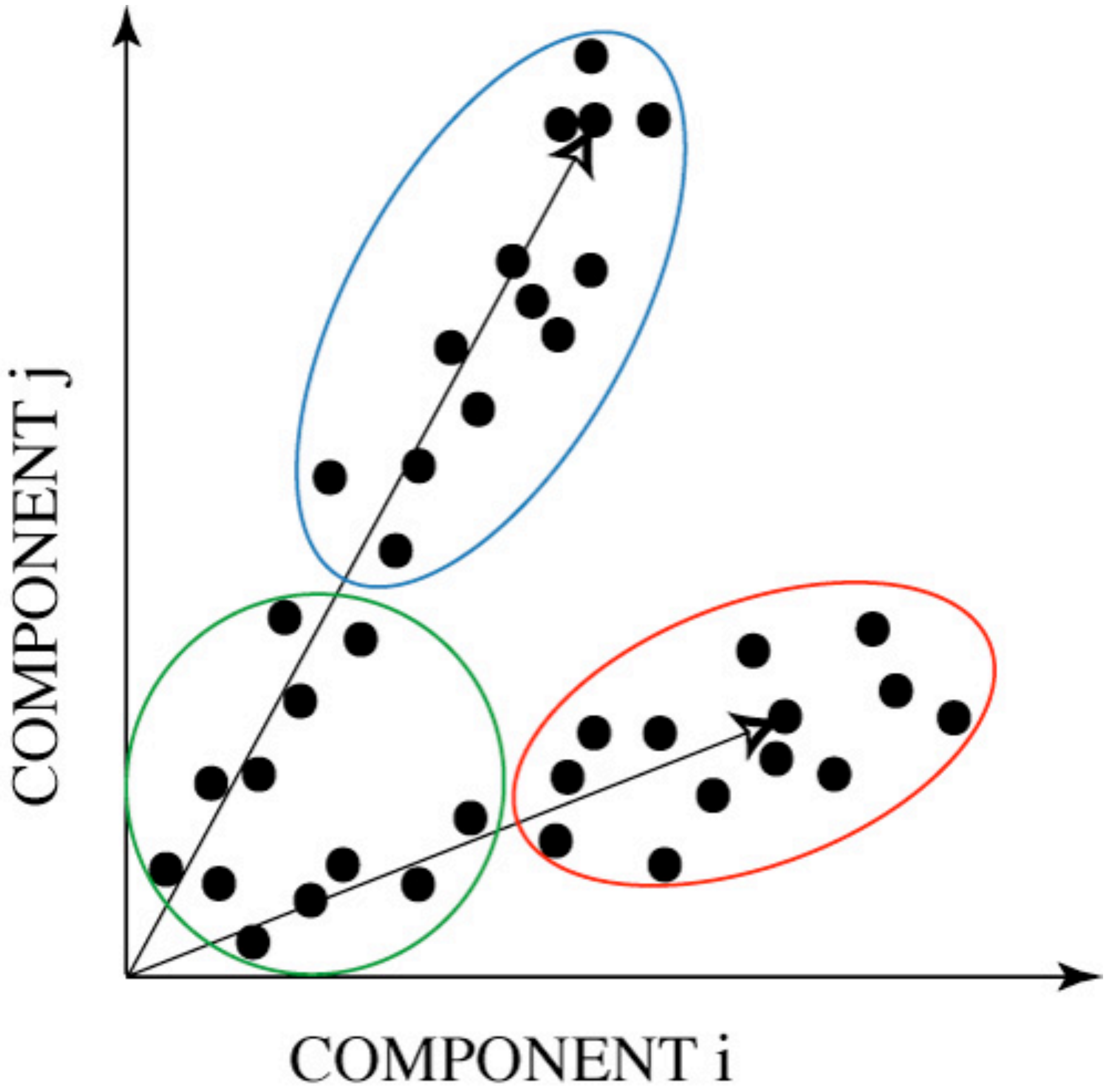
Different distance measure

- One compositional type should involve a constant **ratio** of components; radius from center is thickness



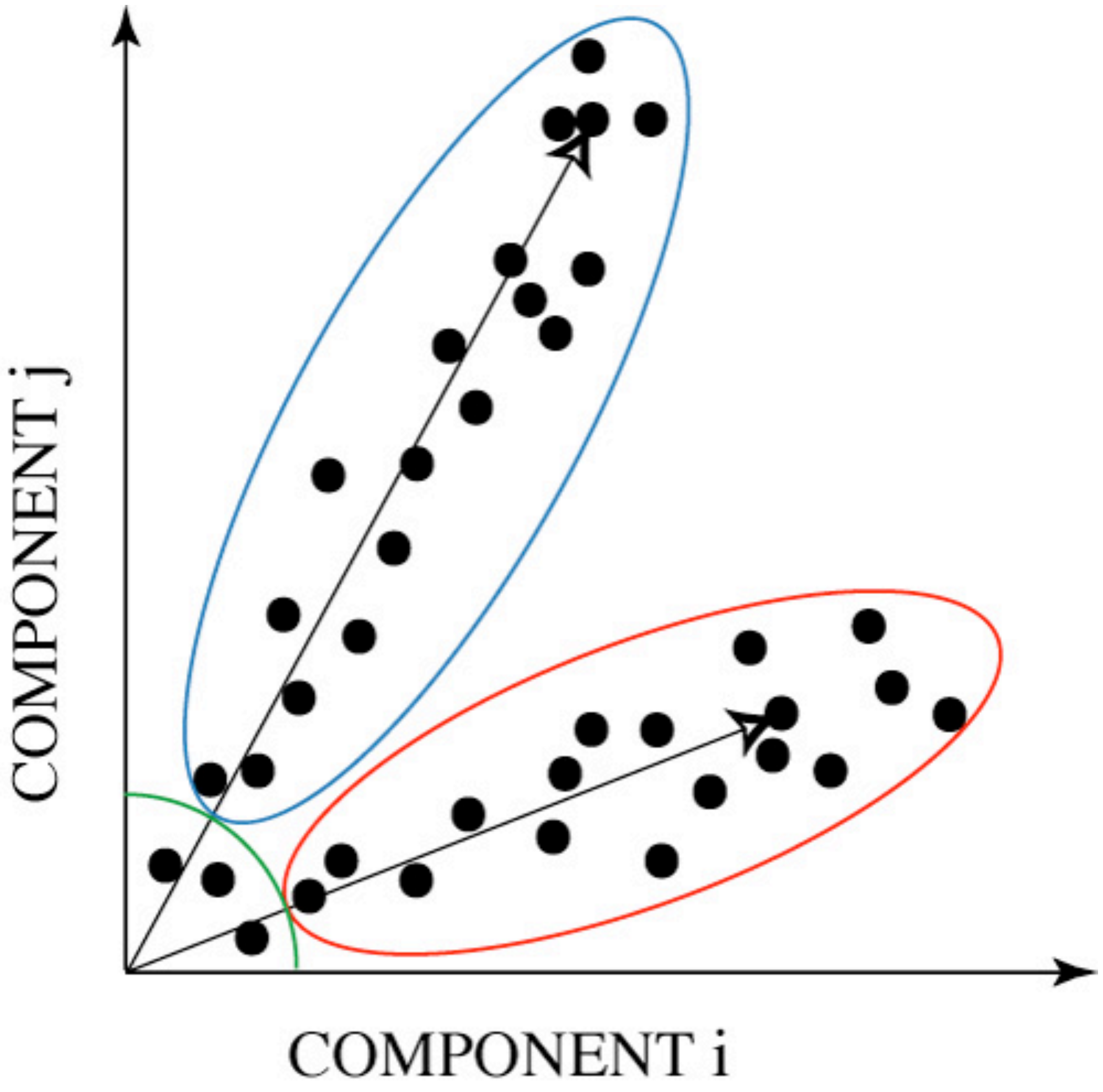
Different distance measure

- One compositional type should involve a constant **ratio** of components; radius from center is thickness
- When Euclidean distance is used, clustering algorithm finds spherical clusters



Different distance measure

- One compositional type should involve a constant **ratio** of components; radius from center is thickness
- When Euclidean distance is used, clustering algorithm finds spherical clusters
- To compensate for thickness, use cosine angle distance (θ) instead of Euclidean distance (d)

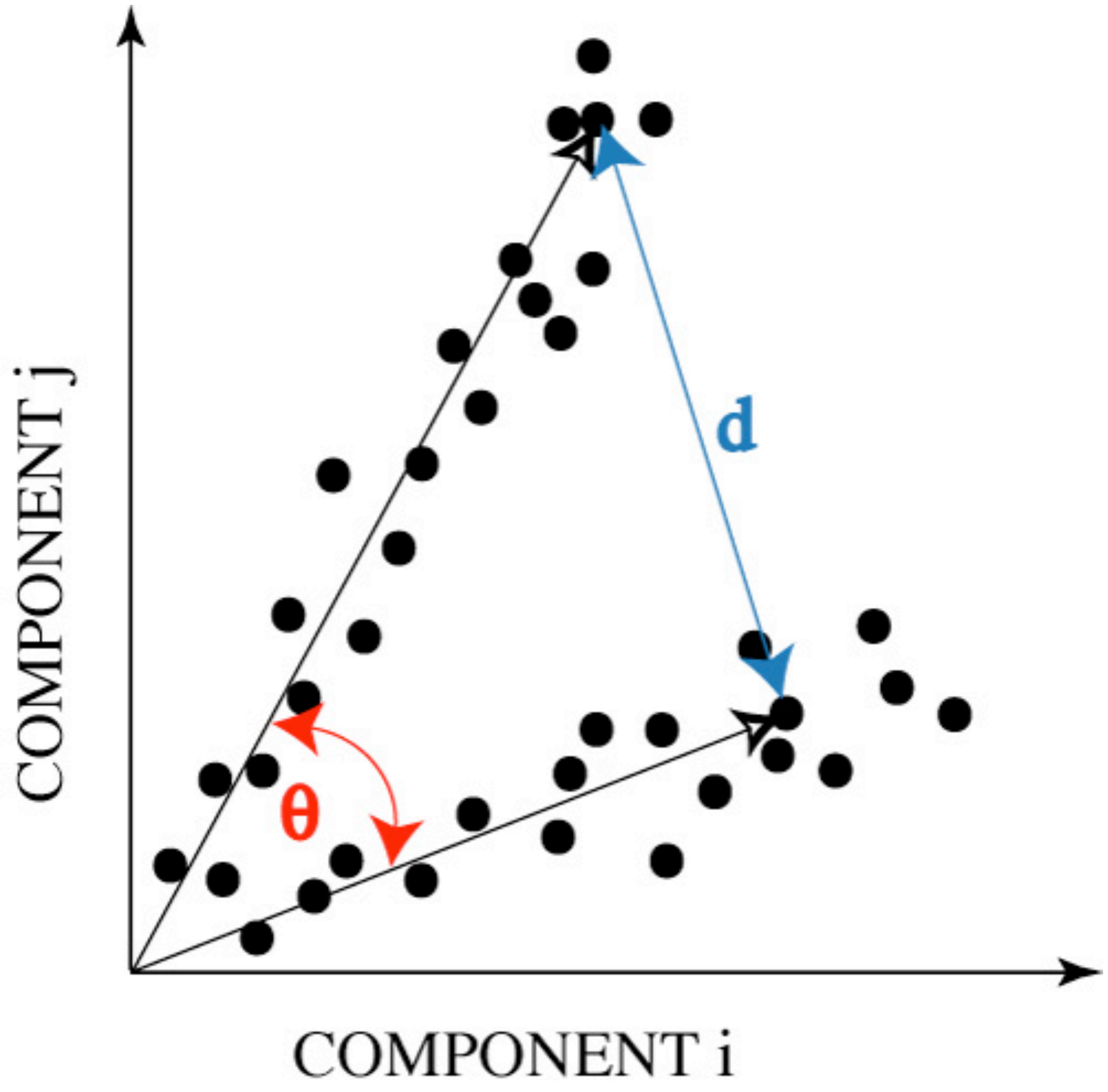


Different distance measure

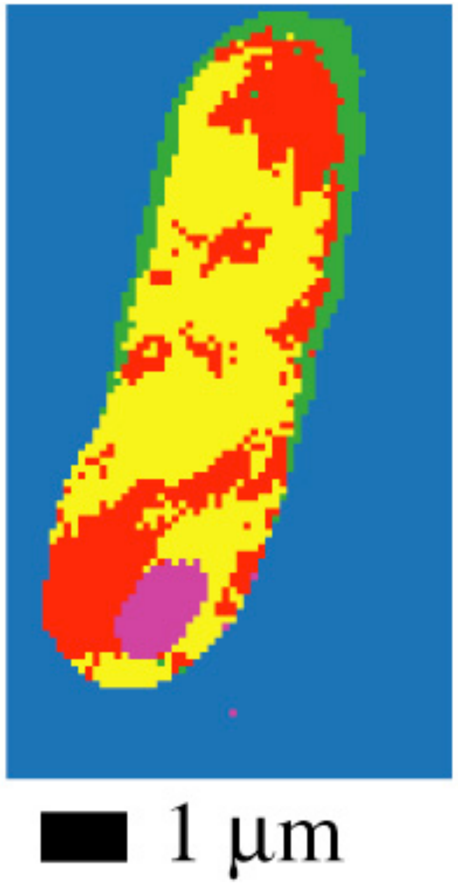
- One compositional type should involve a constant **ratio** of components; radius from center is thickness
- When Euclidean distance is used, clustering algorithm finds spherical clusters
- To compensate for thickness, use cosine angle distance (θ) instead of Euclidean distance (d)

$$\theta = \arccos\left(\vec{X}, \vec{Y}\right)$$

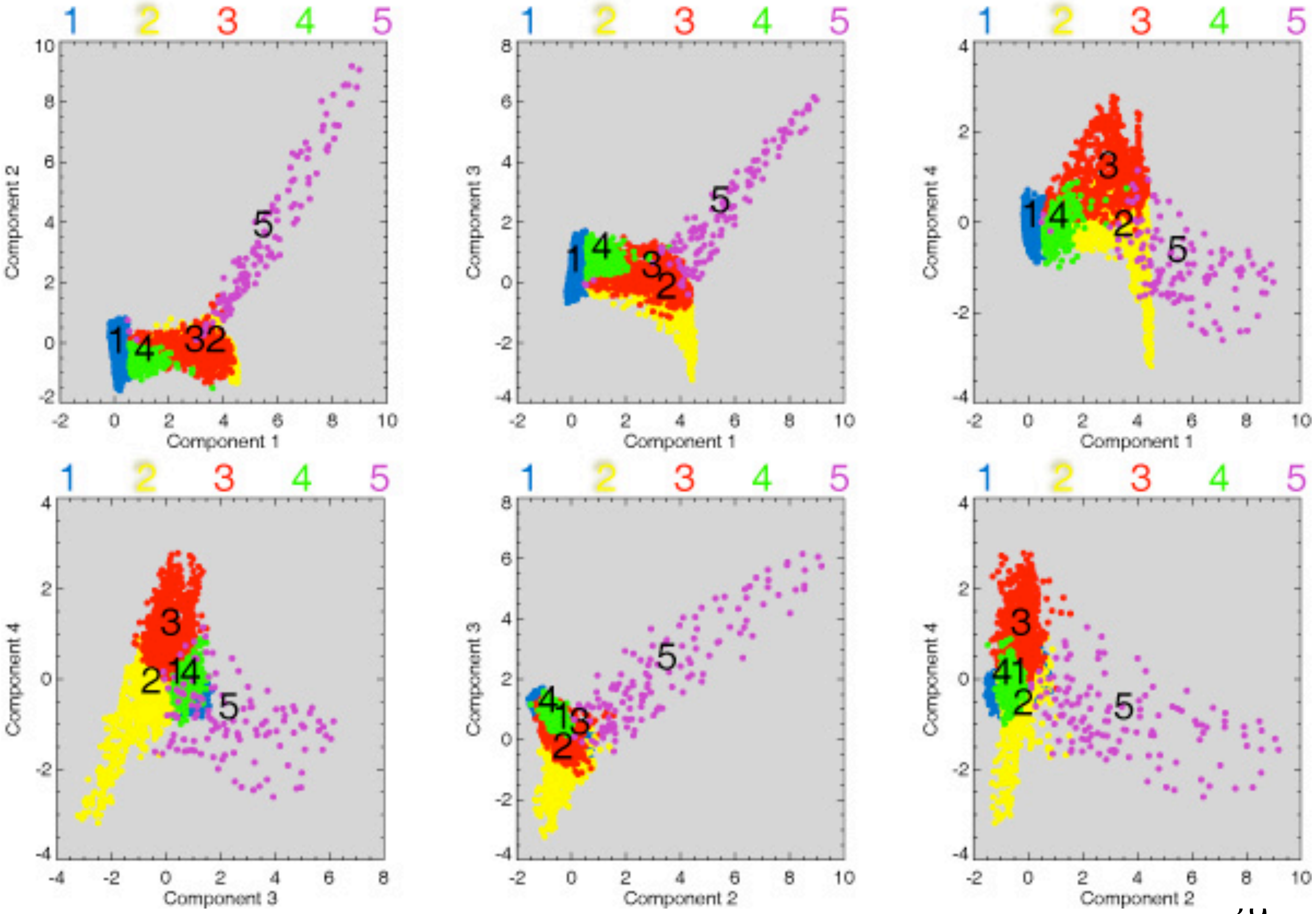
$$\theta = \arccos\left(\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \times \sum_i y_i^2}}\right)$$

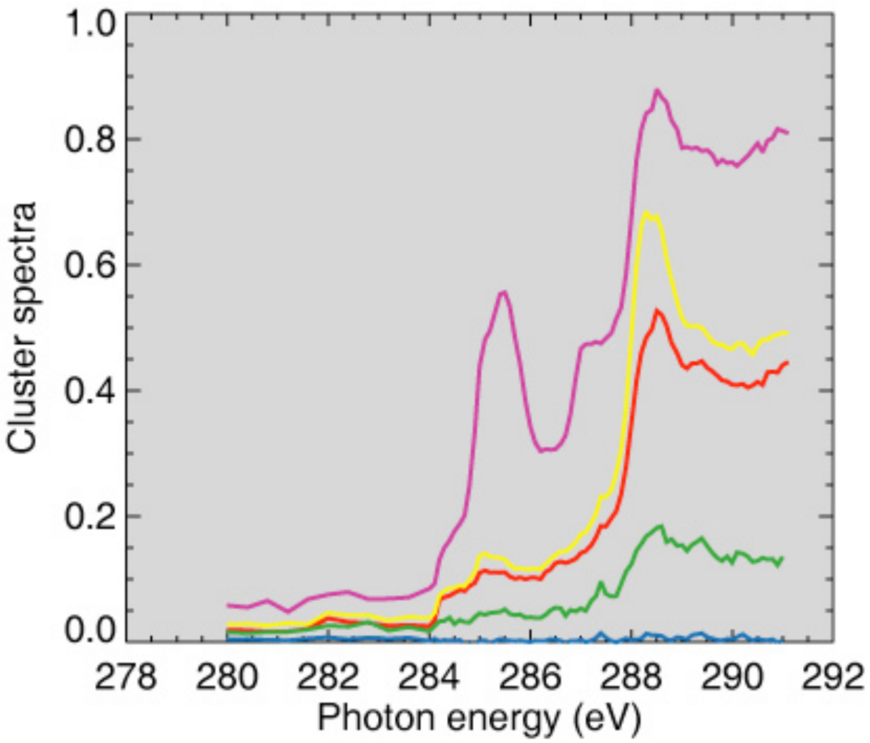


Clostridium sp. - reexamined



- Cluster analysis with cosine angle distance measure is much better!
- Now classifying by compositional (and not thickness) variations!



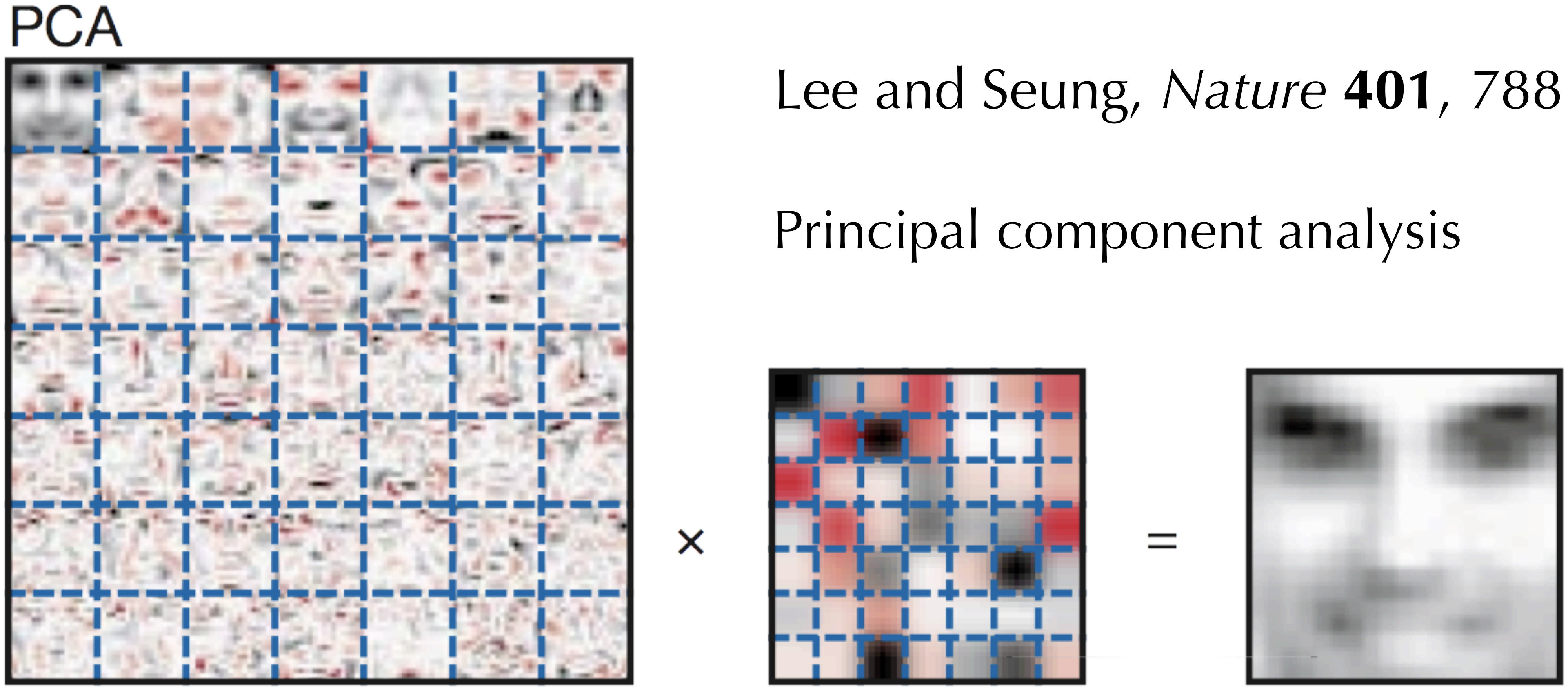


Finding *useful* organizers S

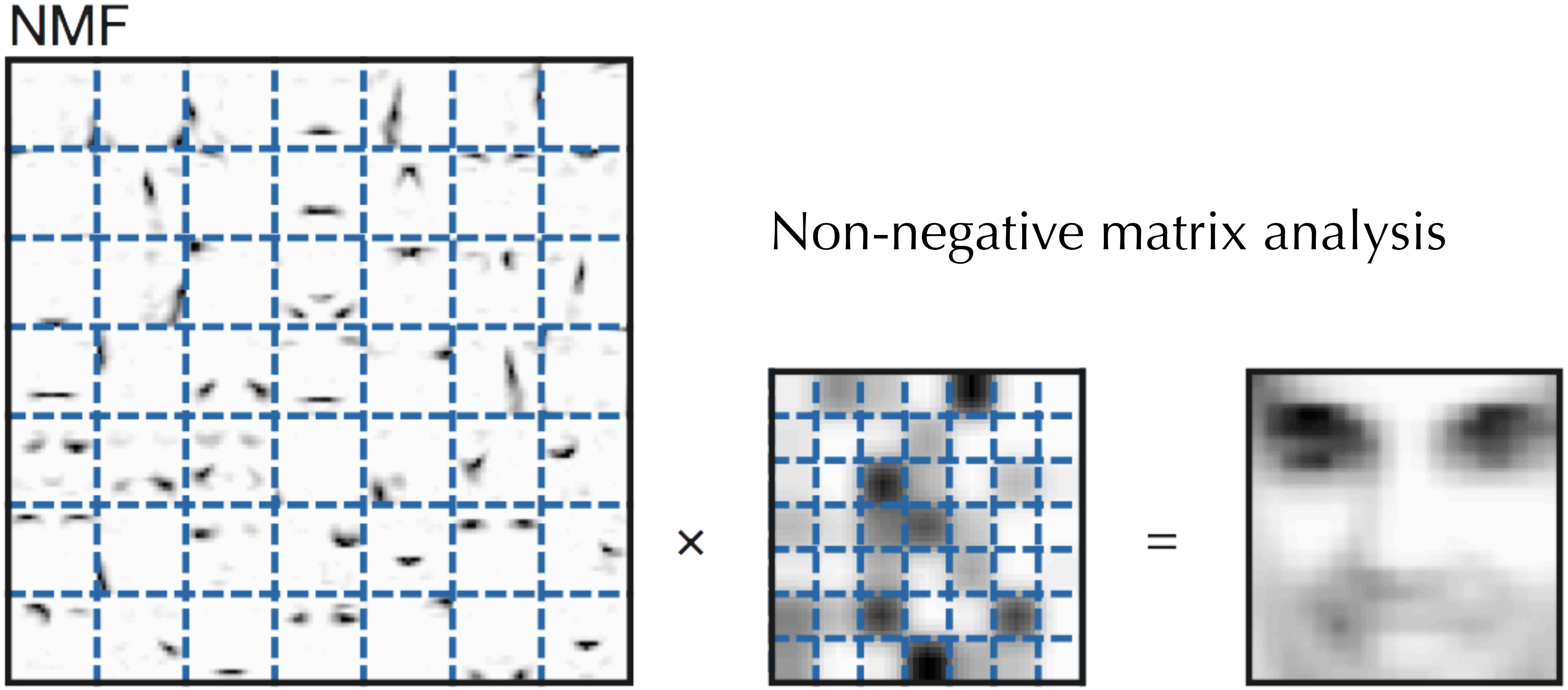
- Cluster analysis: pixels with common spectroscopic signatures yield $C_{N\times S}$
- Varimax: “rotate” $C_{N\times S}$ to make all spectra positive. Works well with discrete elemental signals in fluorescence, TOF-SIMS (P. Kotula, Sandia Labs).
- Non-negative matrix analysis: build $C_{N\times S}$ from noise, constraining for positivity and a minimum set of S

Avoiding negativity: non-negative matrix analysis

Lee and Seung, *Nature* **401**, 788 (1999)



Principal component analysis



Non-negative matrix analysis

NNMA: the goal

- Again we wish to find the organizer *S*:

$$D_{N\times P} \;=\; C_{N\times S}\cdot R_{s\times P}$$

$$\begin{bmatrix} D_{11} & \text{pixels} & D_{1P} \\ \text{spectra} & & \vdots \\ D_{N1} & \ldots & D_{NP} \end{bmatrix} = \begin{bmatrix} C_{11} & \text{pixels} & C_{1S} \\ \text{spectra} & & \vdots \\ C_{N1} & \ldots & C_{NS} \end{bmatrix} \cdot \begin{bmatrix} R_{11} & \text{pixels} & R_{1P} \\ \text{spectra} & & \vdots \\ R_{S1} & \ldots & R_{SP} \end{bmatrix}$$

- Our constraint: minimize $|D - C\cdot R|_F$ with $C, R, D \geq 0$

Algorithms for NNMA

Fall into 3 main categories:

- multiplicative update
  - prototype: Lee & Seung
  - requires more iterations to converge
  - not flexible -- once value hits zero, it stays zero for all subsequent iterations, even if result is not optimal
- gradient descent
  - take steps **ε** in direction of negative gradient
  - convergence depends on **ε**
- alternating least squares
  - can be very fast for *unconstrained* problems
  - **non-negatively constrained least squares** (NNLS) guaranteed to converge to local minimum
  - but NNLS more costly

  → Fast-combinatorial NNLS (FC-NNLS)?

Multiplicative update NNMA

- Iterative procedure:

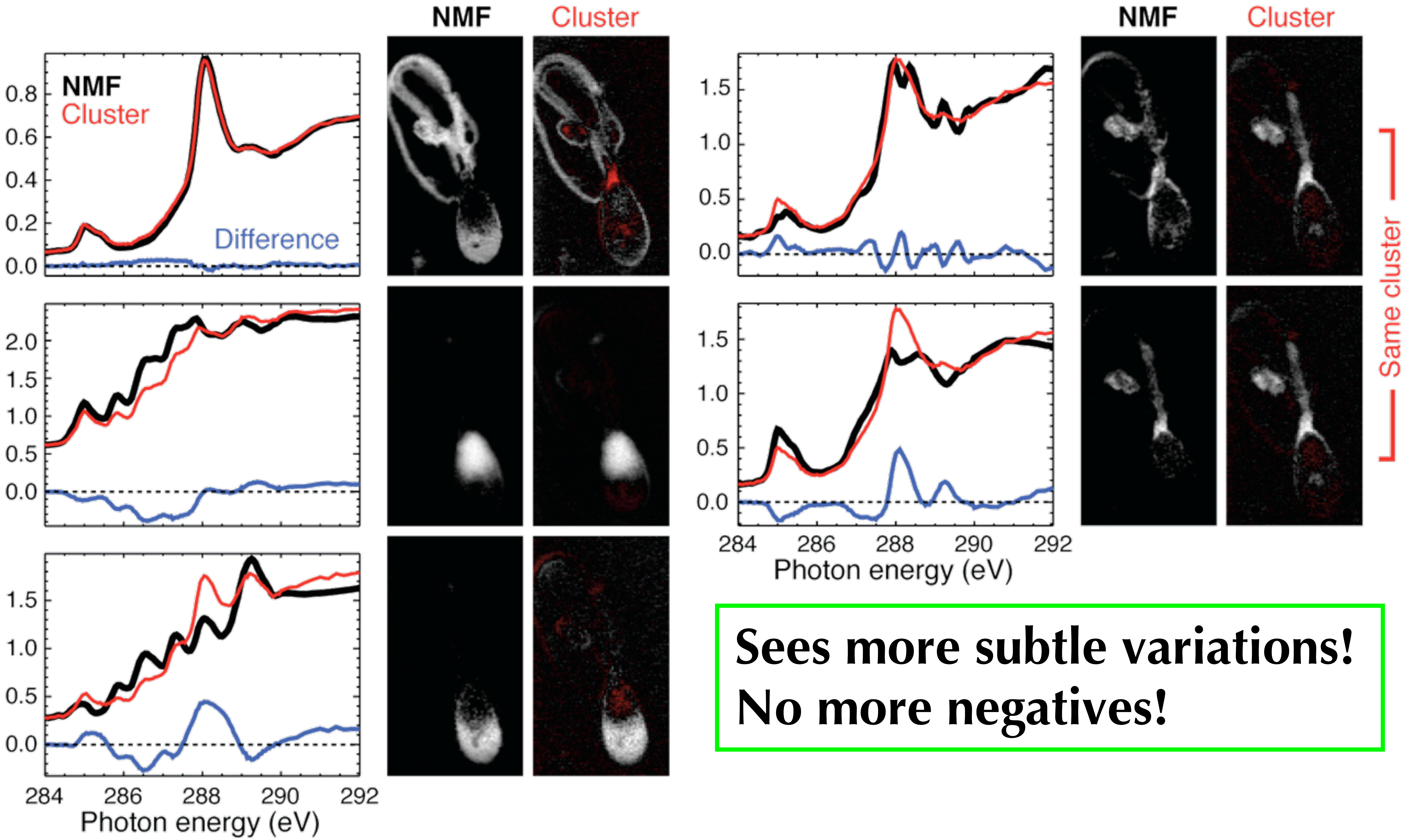
$$\mu_{N\times S} := \mu_{N\times S} \cdot \left(\frac{D_{N\times P}}{\mu_{N\times S} \times t_{S\times P}} \times t^T_{P\times S} \right)$$

For those pixels where $\mu_{N\times S} \times t_{S\times P}$ gives larger values than those present in the data $D_{N\times P}$, this estimate update will drive $\mu_{N\times S}$ towards smaller values and vice versa.

$$t_{S\times P} := t_{S\times P} \cdot \left(\mu^T_{S\times N} \times \frac{D_{N\times P}}{\mu_{N\times S} \times t_{S\times P}} \right)$$

- Fleckenstein and Jacobsen (unpublished), after Lee and Seung (1999)

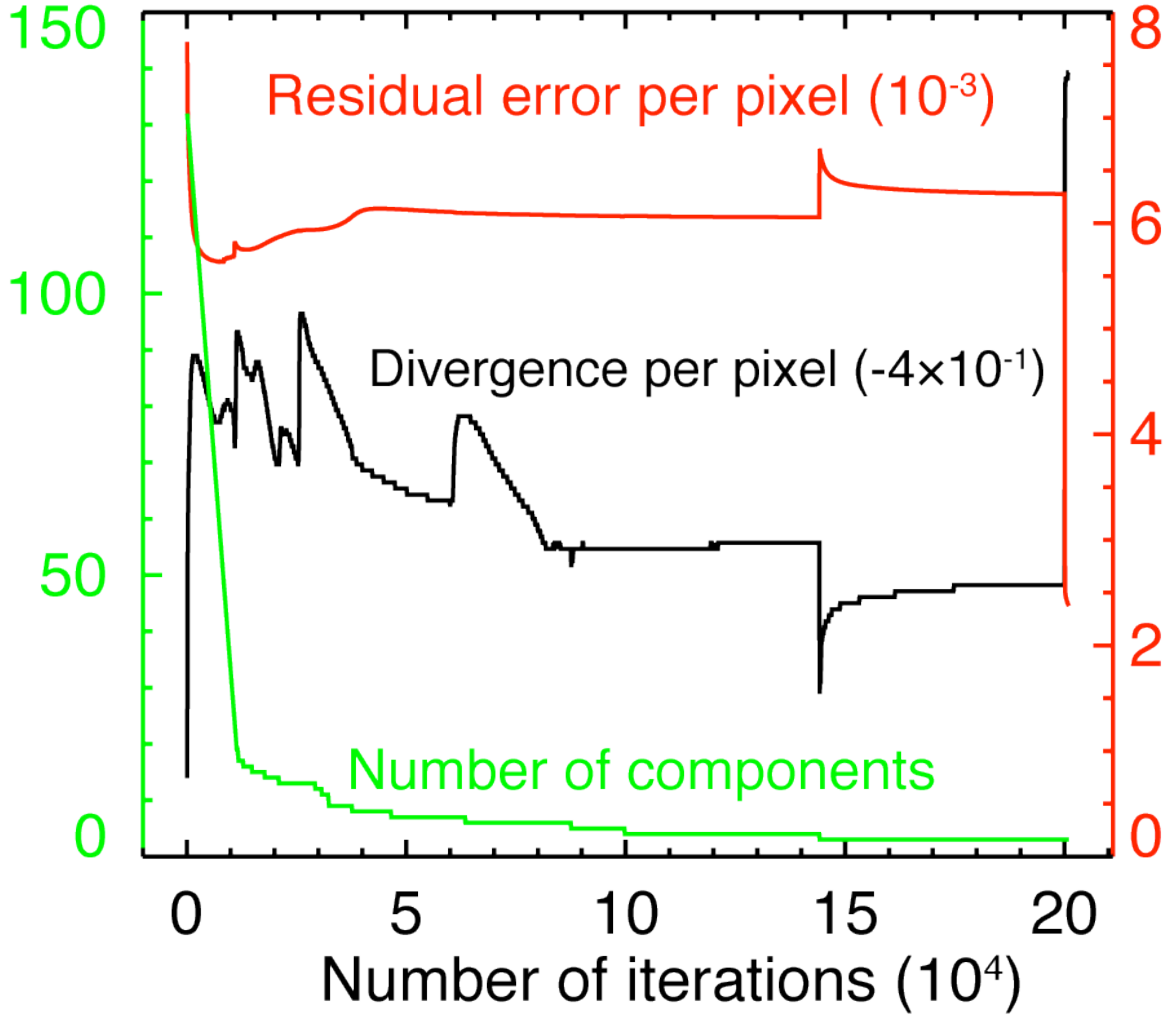
NNMA analysis of sperm



Sees more subtle variations!
No more negatives!

NNMA: many, many iterations

- On a single processor, Lee and Seung NMF algorithm can be slow!



Fast-Combinatorial NNLS

(Benthem & Keenan, 2005)

Problem: find R in

$$CR = D$$

for which $||CR - D||_F^2$ is minimized and subject to constraint

$$R \geq 0$$

- For **unconstrained** problems, the "pseudoinverse" $C^\dagger$ is the same for all columns of R:

  $$R = C^\dagger D$$

  so we only need to calculate $C^\dagger$ once.

- For **constrained** problems, this is not the case. Do we need to calculate *n* inverses for D with *n* columns?
- No...

Fast-Combinatorial NNLS (cont’d)

(Benthem & Keenan, 2005)

- At each iteration, first find columns of R sharing same zero positions:

  e.g.

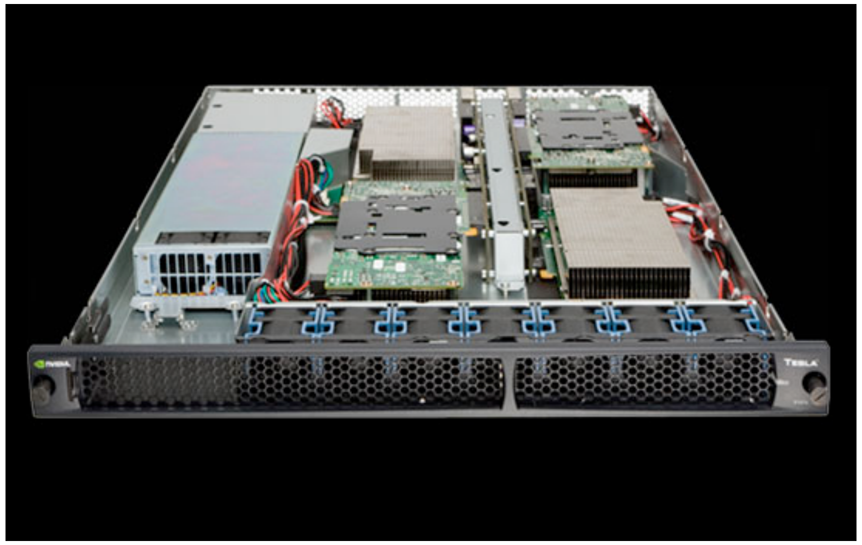
$$R = \begin{pmatrix} 0 & 5 & 0 \\ 3 & 0 & 8 \\ 7 & 4 & 1 \end{pmatrix}$$

Columns 1 and 3 both have their first elements equal zero.

- Pseudoinverse $C^{\dagger}$ is the same for these columns, thus we can compute $C^{\dagger}$ in a “**column-parallel**” way -- this is the “fast” part.
- Benthem also provides a method for identifying and grouping similar columns -- the “combinatorial” part.
- FC-NNLS looks like a promising algorithm for solving the NNMA problem -- but still work in progress!

Parallel computing

- Parallel nature of FC-NNLS algorithm may be well-suited to parallel computing architecture: NVidia CUDA, or OpenCL →



NVidia Tesla S1070

- Each inverse calculation for each "group" of columns is independent.
- Each inverse calculation can be computed by a "block" of "threads".
- Many blocks can execute in parallel, thus potentially speeding up the parallel algorithm even more...

This talk

- Soft x-ray spectromicroscopy: what we do and how we process the data
  - Principal components, clusters, and non-negative matrices
- Connections with problems in other fields
  - X rays, electrons, satellites, shopping...
- Some thoughts on data formats

Hyperspectral data in satellite imaging

- Image with multiple color filters; sort out vegetation types, or look for armored tanks in the desert
- Often time a small number of wavelengths: like ~5 to 128
- Large body of work, but not all of it is published!

Near-edge spectroscopy: ELNES and XANES

ELNES (electron **E**nergy **L**oss)

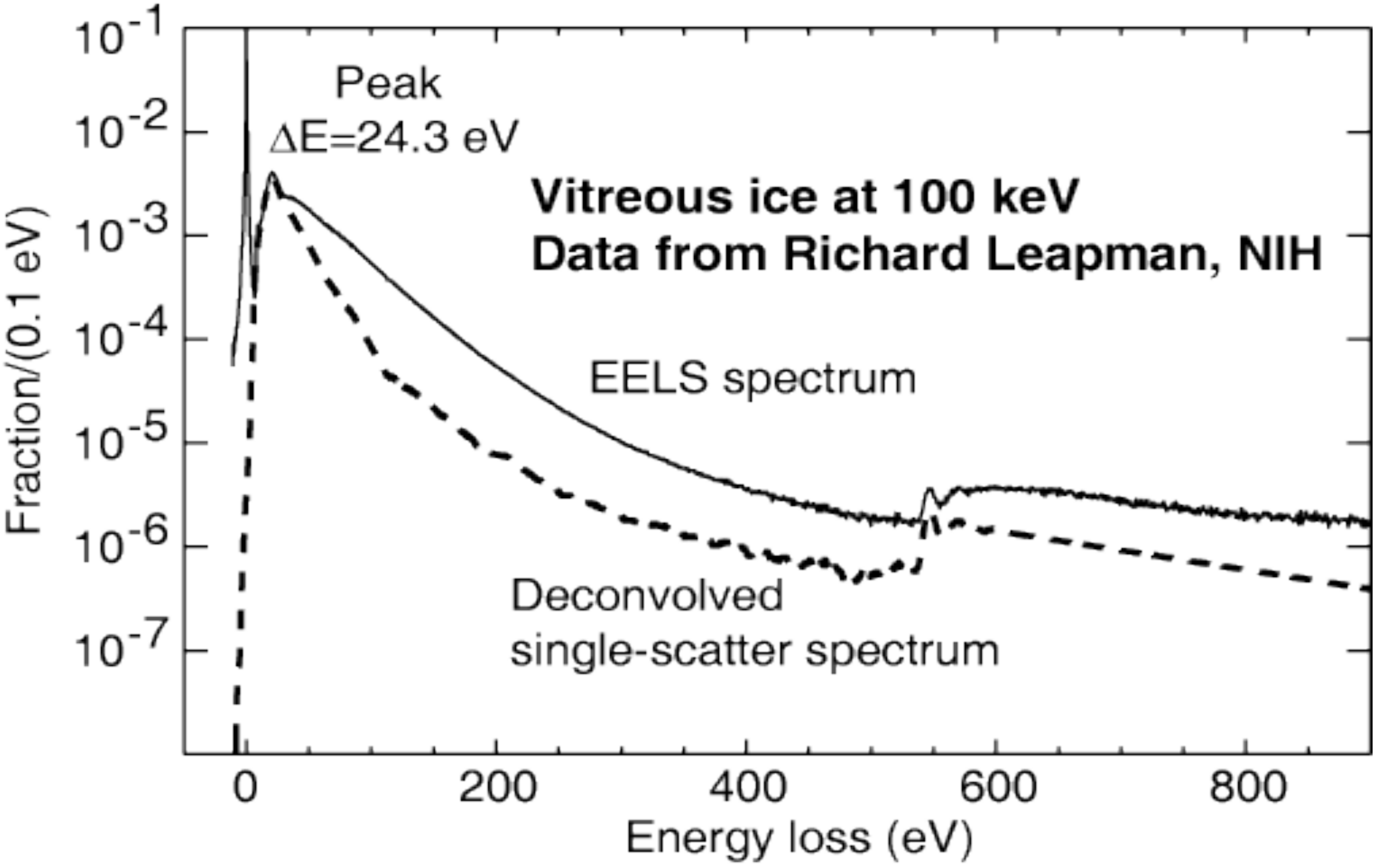
- Plural inelastic scattering
- Many elements at once - but plasmon modes are always excited (damage)
- Δ*E* was ~0.6 eV, but now 0.1 eV in some cases

XANES (**X**-ray **A**bsorption)

- No plural scattering
- One element at a time - slow but less damage
- Δ*E* of 0.05-0.1 eV is common

Electrons ~1000x more damaging:

- Isaacson and Utlaut, *Optik* **50**, 213 (1978)
- Rightor *et al.*, *J. Phys. Chem. B* **101**, 1950 (1997).



Spectrum imaging in EELS

- EELS: electron energy-loss spectroscopy
- "Spectrum-image: the next step in EELS digital acquisition and processing," Jeanguillaume and Colliex, *Ultramicroscopy* **28**, 252 (1989)
- "Electron energy-loss spectrum-imaging," Hunt and Williams, *Ultramicroscopy* **38**, 47 (1991)

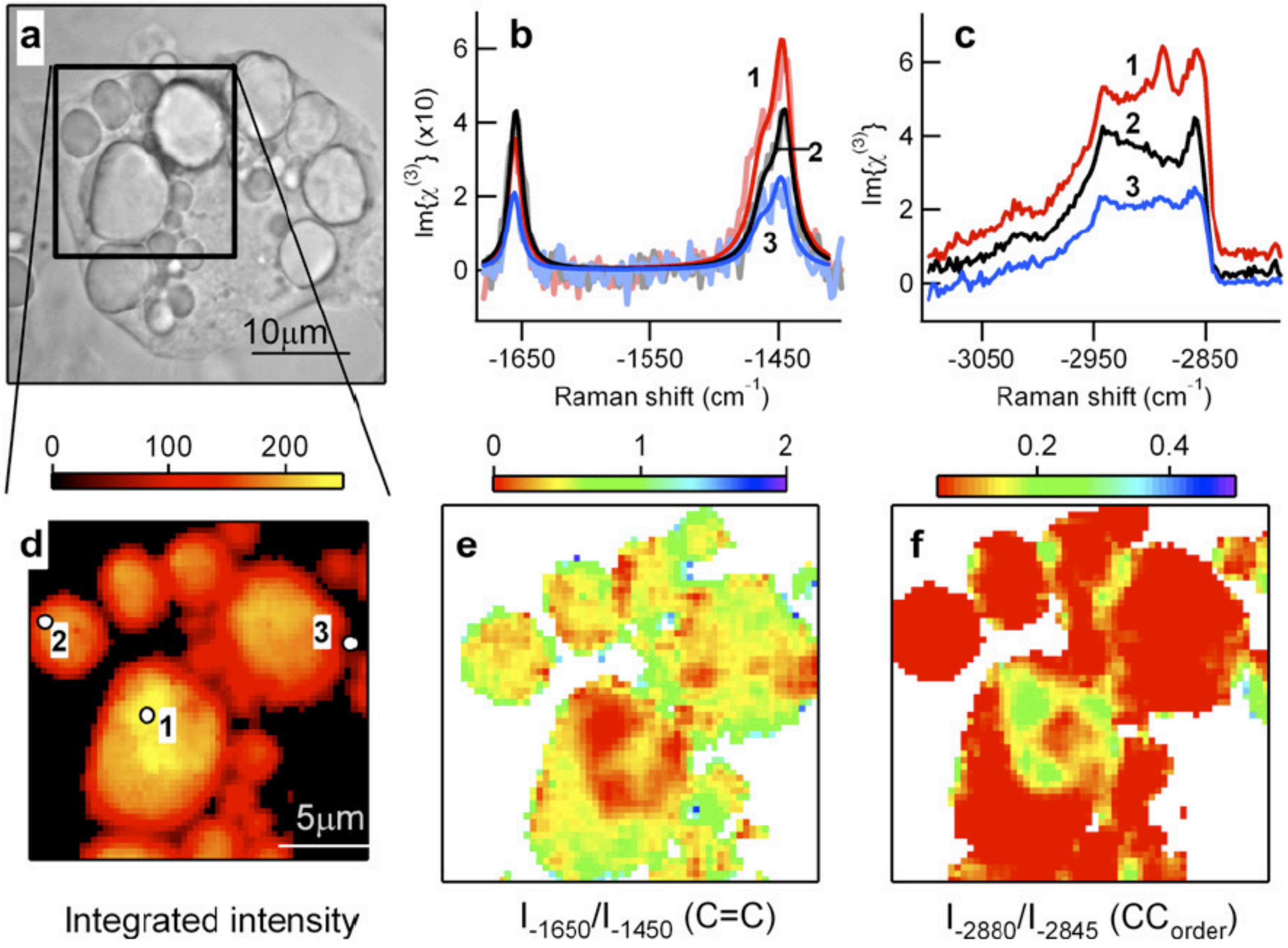
4908 Biophysical Journal Volume 95 November 2008 4908–4914

Quantitative Label-Free Imaging of Lipid Composition and Packing of Individual Cellular Lipid Droplets Using Multiplex CARS Microscopy

Hilde A. Rinia,* Koert N. J. Burger,† Mischa Bonn,‡ and Michiel Müller*

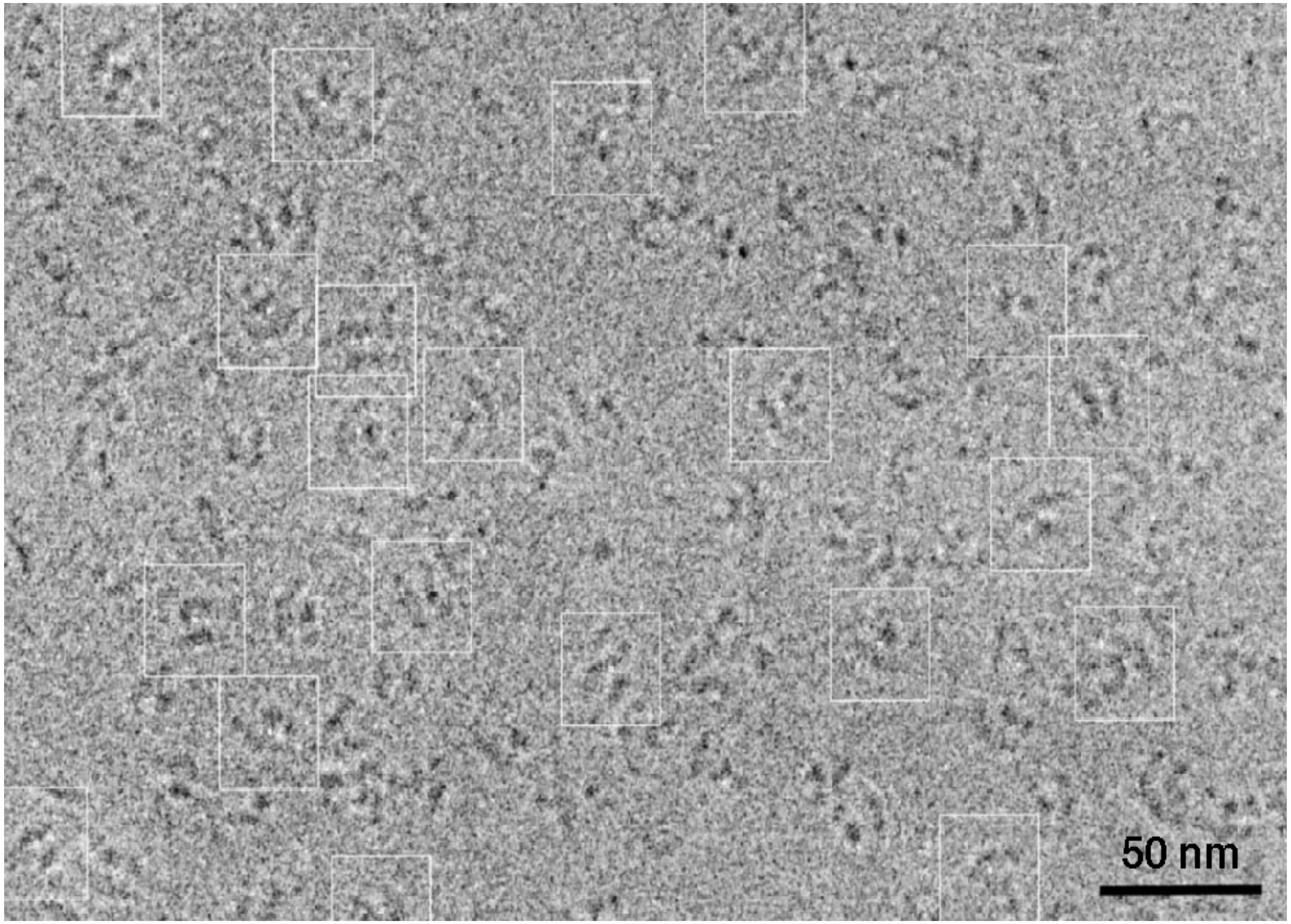
*Swammerdam Institute for Life Sciences, University of Amsterdam, 1098 SM Amsterdam, The Netherlands; †Division of Endocrinology and Metabolism, Department of Biology and Institute of Biomembranes, Utrecht University, 3584 CH Utrecht, The Netherlands; and ‡FOM Institute for Atomic and Molecular Physics (AMOLF), Kruislaan 407, 1098 SJ Amsterdam, The Netherlands

CARS: coherent anti-Stokes Raman (visible light probe of what are normally infrared signals)



Untangling complexity: single particle electron microscopy

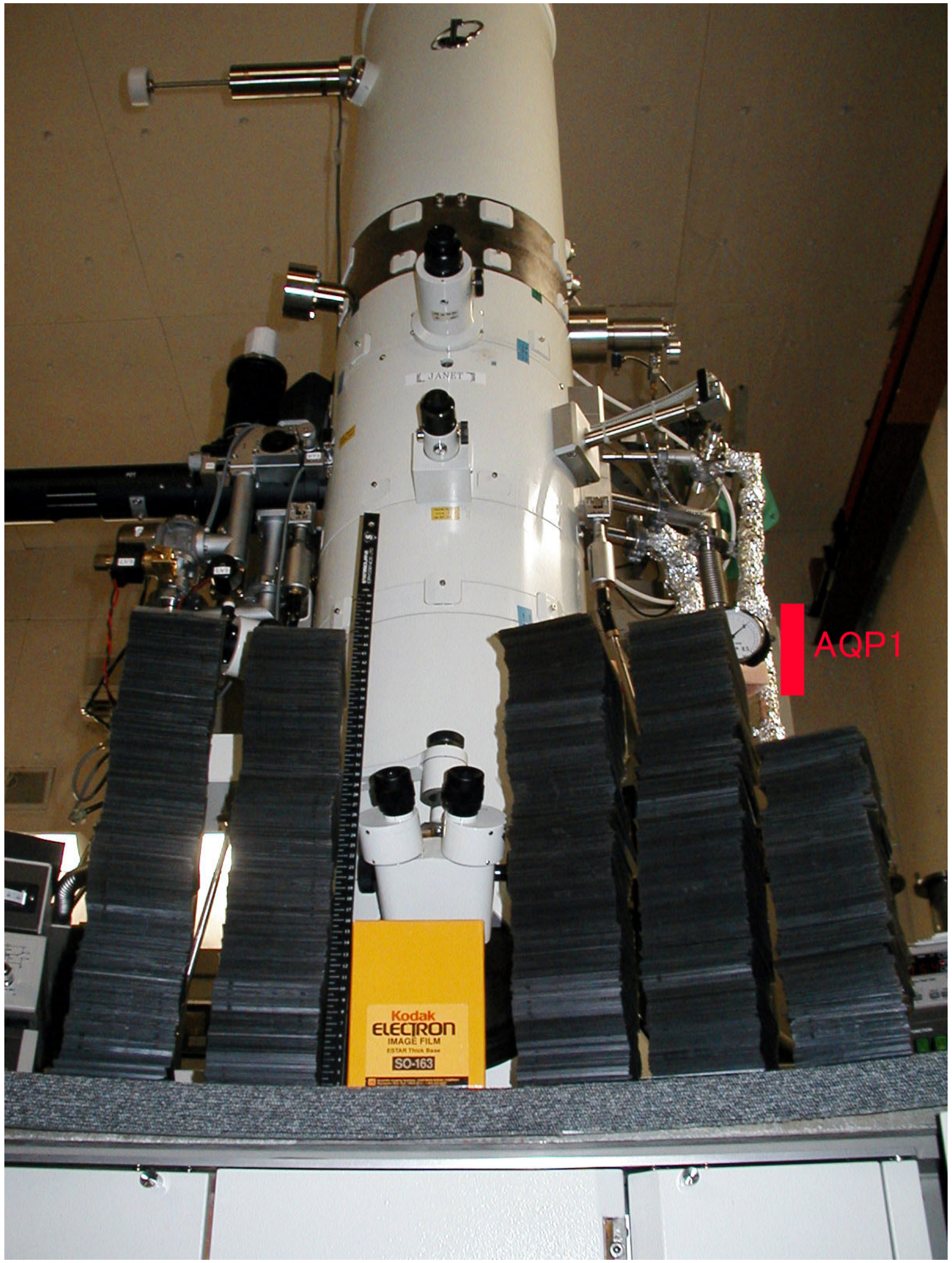
Goal: atomic-resolution EM. Pioneers: Franck (Albany) and others.



Brink *et al.*, *PNAS* **99**, 138 (2002): many molecules of fatty acid synthase, in thin ice, at random orientations.

Labor of love…

- Shown here: film plates of acetylcholine receptor. Miyazawa, Fujiyoshi, and Unwin, *Nature* **423**, 949 (2003).
- AQP1: aquaporin-1, Murata, Mitsuoka, Hirai, Walz, Agre, Heymann, Engel, and Fujiyoshi, *Nature* **407**, 599 (2000). (Agre: 2003 Nobel Prize in Chemistry)



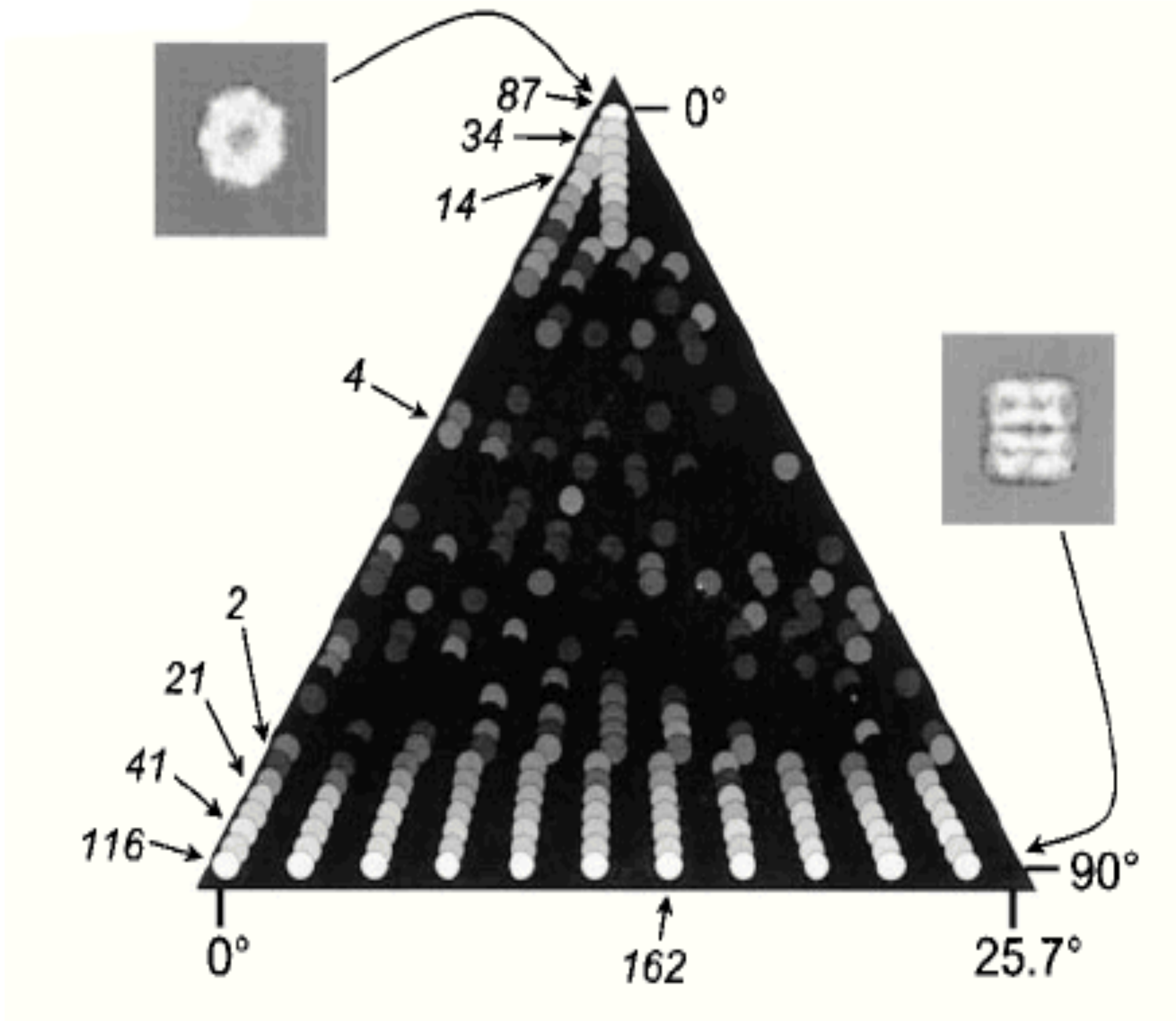
Multivariable statistical analysis

- We’ve seen that there are solutions for $D_{N\times P} = C_{N\times S} \cdot R_{S\times P}$
- Single particle imaging: we have N separate images of our sample, with P pixels per image, or $D_{N\times P}$
- We want to characterize the data in terms of θ distinct viewing directions (rather than S distinct chemical components)
- Which of N images map into which of θ viewing angles? $C_{N\times\theta}$
- What do representative images with pixels P look like from representative viewing angles θ? $R_{\theta\times P}$
- Single particle imaging: find $D_{N\times P} = C_{N\times\theta} \cdot R_{\theta\times P}$

Forming a tomographic dataset

Group similar projections, then iterate:

1. Correlate data to a view of a model
2. Tomographic reconstruction of data to obtain new model

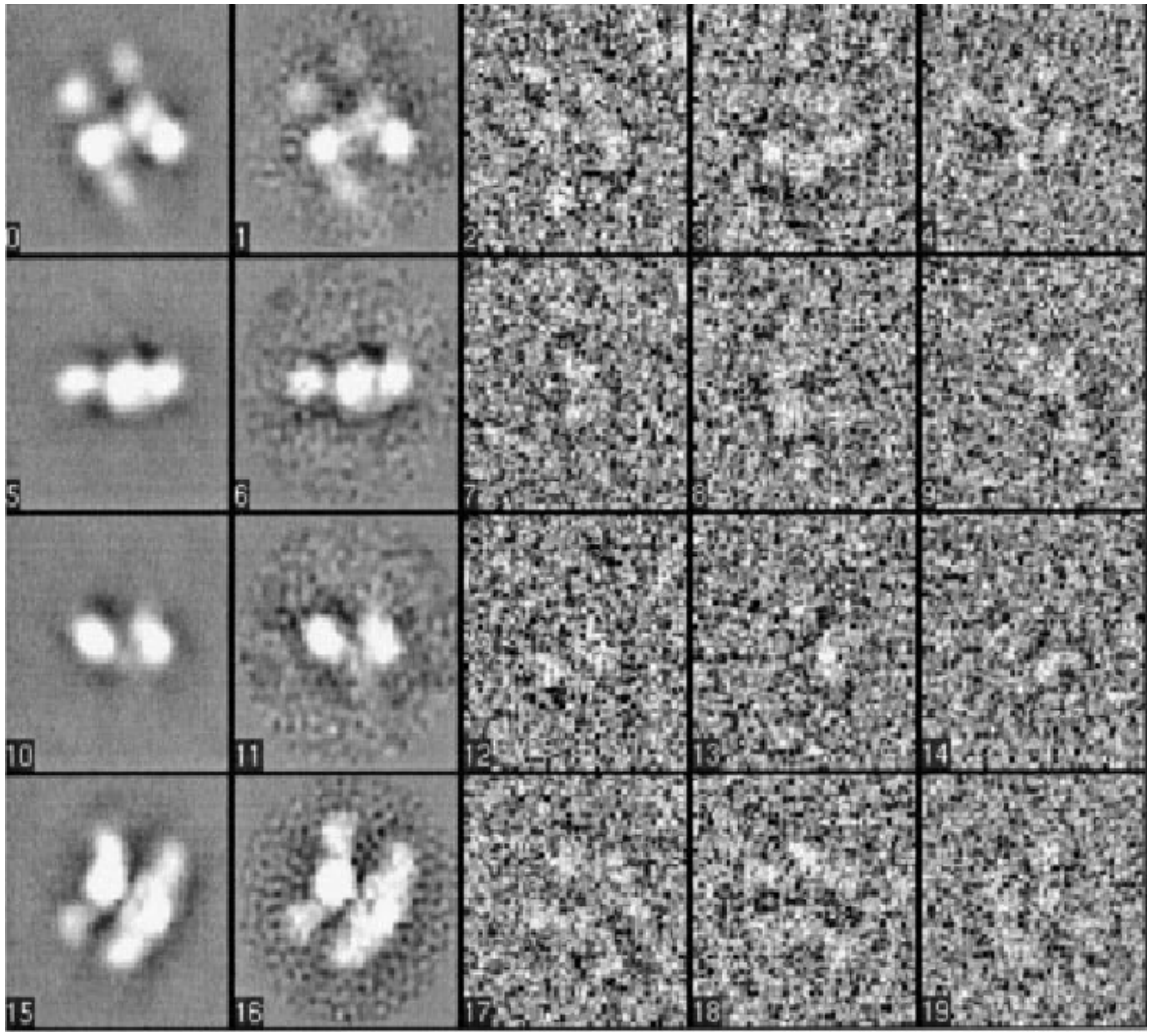


Ludtke et al., J. Mol. Bio. **314**, 253 (2001)

Model projection

Sum of projections

Individual images (in-plane rotation not corrected)



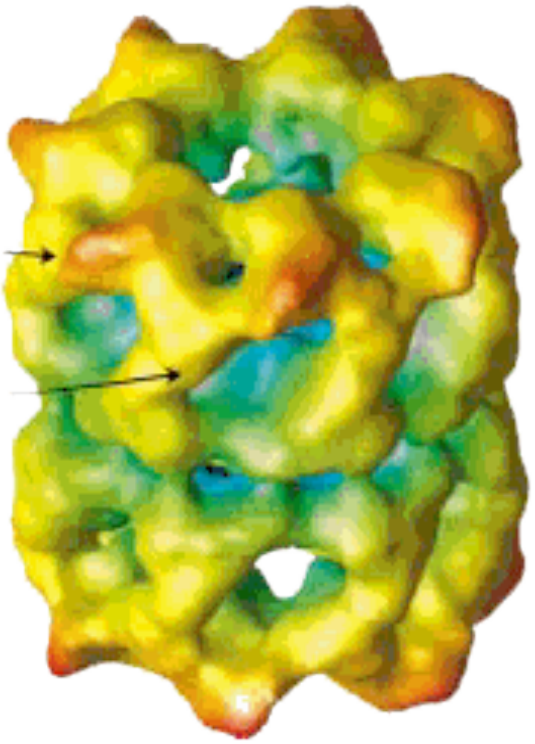
Brink et al., PNAS **99**, 138 (2002)

Images grouped by θ

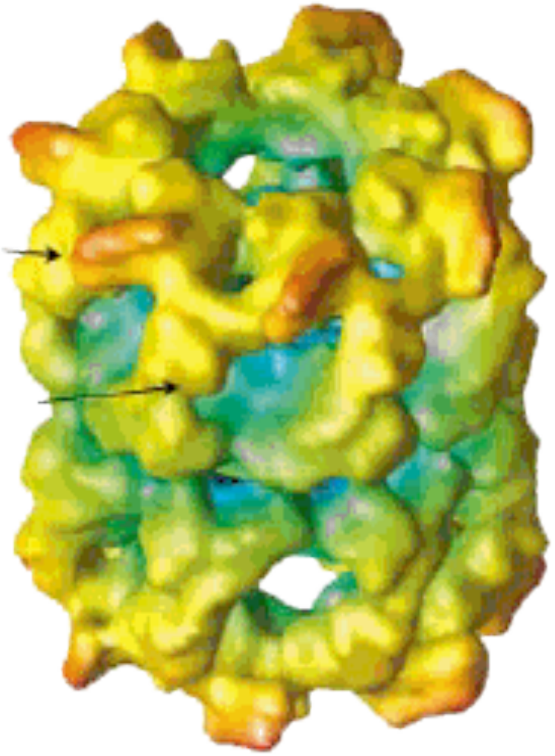
- Once images have been grouped by θ, we guess their viewing angle by comparing with projections from a first-guess model.
- Now do a tomographic reconstruction!
- With a better model, refine:
  - Re-do classification of images to θ
  - Guess again at projection angle
  - Do a new tomographic reconstruction

Single particle tomography example

- GroEL: a molecular chaperone to promote protein folding (essentially an inner sanctuary, hidden from chemical environment of a cell)
- Is there molecular-level variability in GroEL?
- Ludtke *et al.*, *J. Mol. Bio*. **314**, 253 (2001)



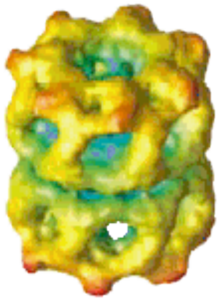
Cryo-EM



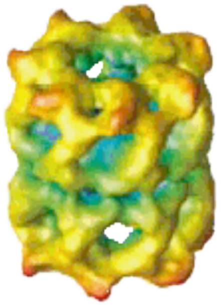
X-ray crystallography blurred to 1.2 nm

Iter 1

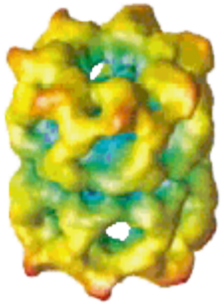
Iter 2



Iter 3



Iter 4

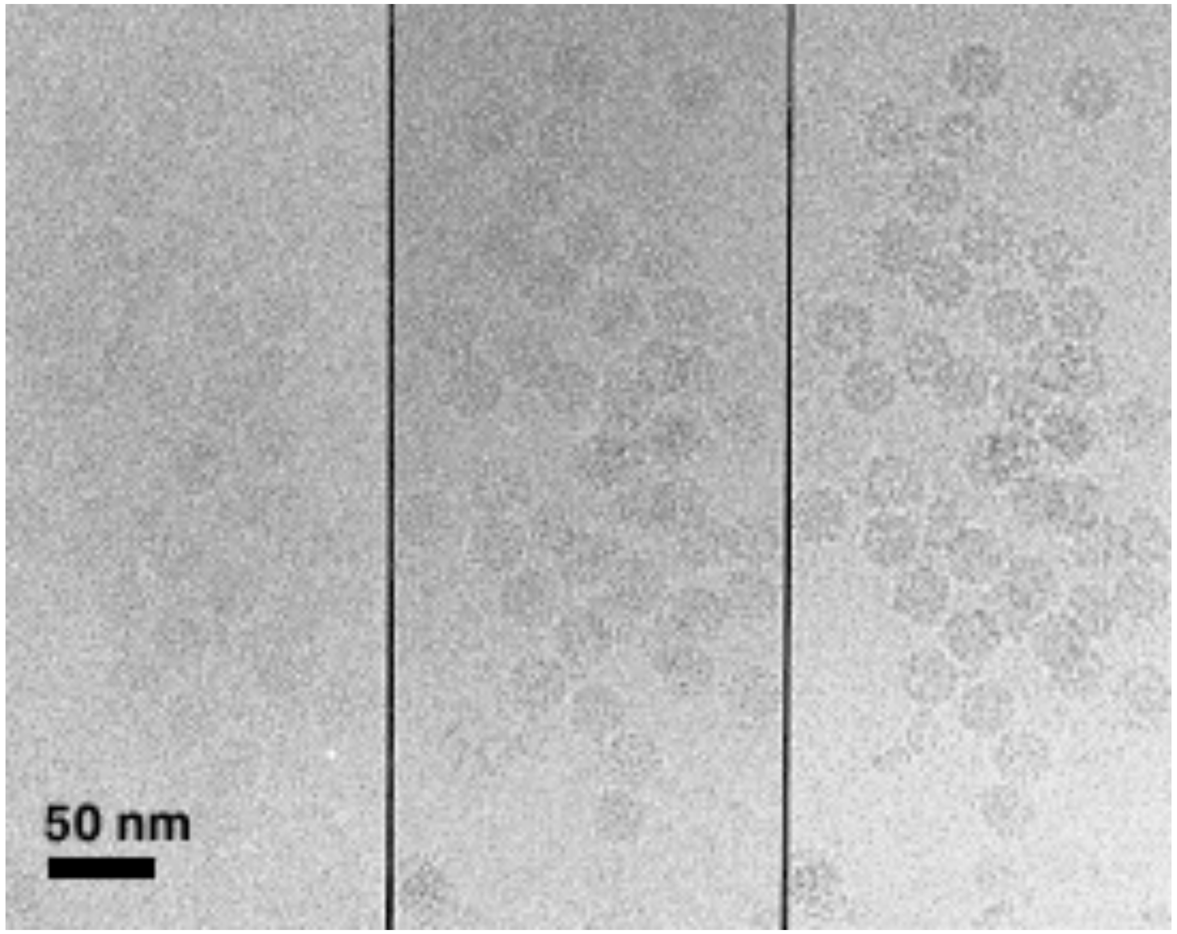


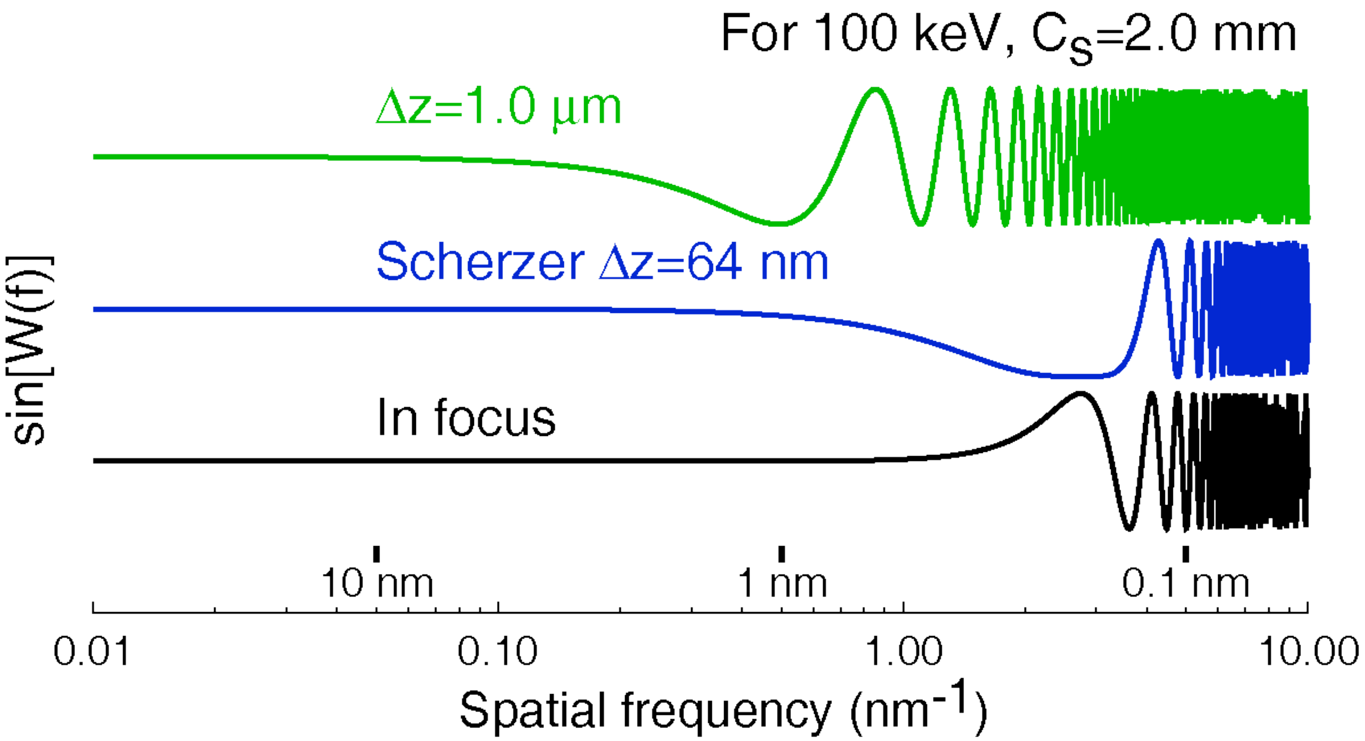
Iter 5

Of special note: Miyazawa, Fujiyoshi, Stowell, and Unwin, “Nicotinic acetylcholine recepter at 4.6 Å resolution: transverse tunnels in the channel wall,” J. Mol. Biol. **288**, 765 (1999)

Challenges in electron microscopy

- Radiation damage limits information (and thus alignment) in any single image
- Contrast transfer function: what you see depends strongly on the focus! Must measure and correct for (except for zeroes)

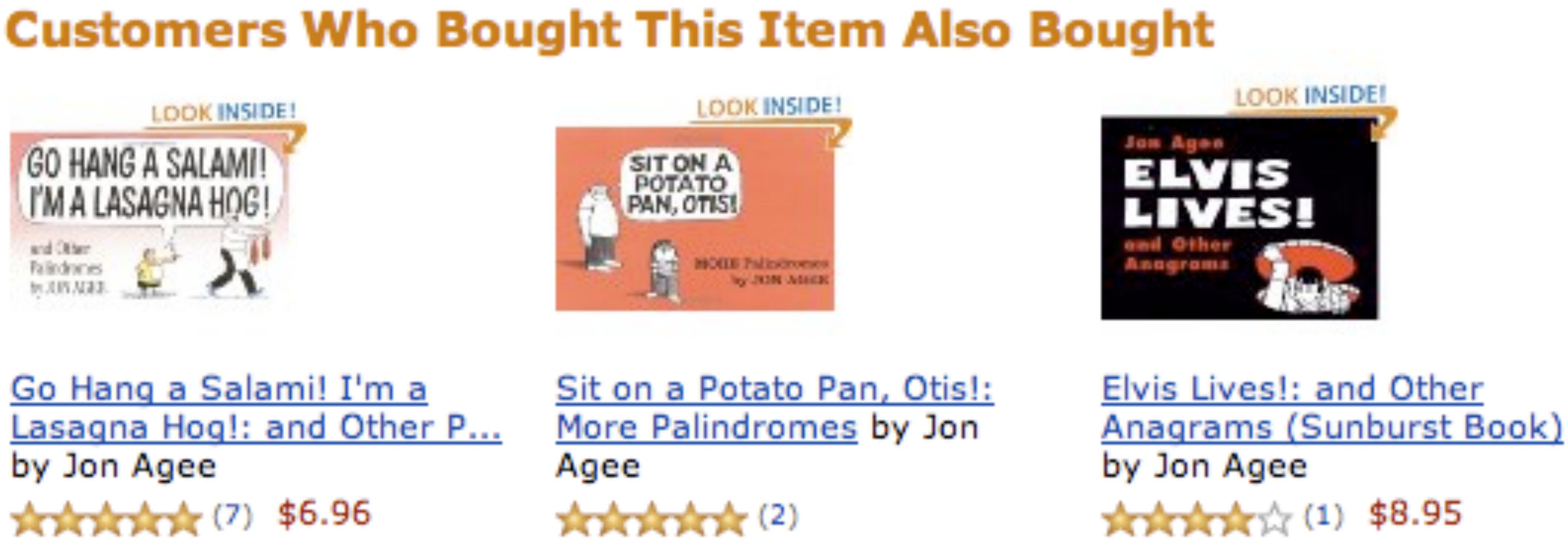




-1, 3, and 6 µm defocus images
Bowen et al., Tomato bushy stunt virus

amazon.com

- We have N shoppers, and P items for purchase
- We want to find S customer types: $D_{N\times P} = C_{N\times S} \cdot R_{S\times P}$
- You, as customer n, match customer types given in $C_{N\times S}$
- Customer types S like to purchase items P as given by $R_{S\times P}$



Customers Who Bought This Item Also Bought

LOOK INSIDE!
GO HANG A SALAMI! I'M A LASAGNA HOG!
and Other Palindromes by JON AGEE
LOOK INSIDE!
SIT ON A POTATO PAN, OTIS!
MORE Palindromes by JON AGEE
LOOK INSIDE!
Jon Agee
ELVIS LIVES!
and Other Anagrams

Go Hang a Salami! I'm a Lasagna Hog!: and Other P... by Jon Agee
★★★★★ (7) $6.96

Sit on a Potato Pan, Otis!: More Palindromes by Jon Agee
★★★★★ (2)

Elvis Lives!: and Other Anagrams (Sunburst Book) by Jon Agee
★★★★☆ (1) $8.95

The New York Times
Monday, September 21, 2009

Bits
Business ▪ Innovation ▪ Technology ▪ Society

Netflix Awards $1 Million Prize and Starts a New Contest

Jason Kempin/Getty Images

“Netflix, the movie rental company, has decided its million-dollar-prize competition was such a good investment that it is planning another one. The company’s challenge, begun in October 2006, was both geeky and formidable: come up with a recommendation software that could do a better job accurately predicting the movies customers would like than Netflix’s in-house software, Cinematch. To qualify for the prize, entries had to be at least 10 percent better than Cinematch.”

Making the future more simple

- With better synchrotron radiation instrumentation, we get richer data of more complex specimens!
- We can learn from other fields how to find the patterns in rich, complex data.
- We can learn from each other!
  - CCP4 in crystallography: shared data formats and data I/O routines, leading to mix-and-match analysis programs.

This talk

- Soft x-ray spectromicroscopy: what we do and how we process the data
  - Principal components, clusters, and non-negative matrices
- Connections with problems in other fields
  - X rays, electrons, satellites, shopping…
- Some thoughts on data sharing

Data storage

- First priority: data exchange, rather than original data storage
- Agreeing on HDF5 (or NeXus) is a great first step!
  - In our lab: spectromicroscopy
  - In our lab: coherent x-ray diffraction imaging/ diffraction microscopy
- But *how* is the HDF5 file organized?
  - Ideally we don't have to write a separate HDF5 file read/write routine for each dataset!

The tests were done on an array of $512 \times 512 \times 128 = 33\,554\,432$ floating point random numbers, either arranged as a single 1D array, or arranged as a 3D array of dimension $[512, 512, 128]$ to see if there are extra overhead costs associated with multidimensional data. Each test was repeated 10 times, leading to the averages shown below:

Time (seconds)	**Binary**	**XDR, 1D**	**XDR, 3D**	**HDF5, 1D**	**HDF5, 3D**
Read	2.72	2.75	2.81	4.12	5.07
Write	2.66	3.77	3.23	2.65	2.76

Here are some comments on the test results:

1. It's not shown in the above table, but using HDF5 added only 2080 bytes to the size of the file.

2. To do these tests properly, it is important to write a series of files before going back to read from the first file. Otherwise one can be fooled into thinking that the readback of the data happened very quickly due to buffering of the data in memory by either the operating system or the hard disk drive's firmware.

3. It might be that HDF5 writes the data in machine-native format and does byte-swapping[2] upon readback; this would explain why it writes data as quickly as binary streaming while imposing a 50–90% performance penalty on readback. With the XDR format, byte-swapping is presumably happening when the file is written, as there is a 20–40% performance penalty relative to a binary write, but little penalty on file readback.

Commonality versus local details

- For **portability**, we want a minimum set of agreed-upon names and attributes
- For **completeness**, we want all the details of the experiment
- For **history**, we want to record all the processing steps that have been applied to the data

Portability

- Agree upon a generic group name for multi-image data, such as `/images` or `/main_array`
- Assign an attribute to explain the type of data, such as `/images/type=transmission` or `type=fluorescence` or...
- Agree on simple minimal attribute names, such as `energies=` with attribute of “`eV`”
- Programs should skip fields they don’t understand, instead of crash

Simplicity=portability

- The less we are *required* to specify in the file, the better.
- The more generic we can make the definitions, the better.
- Any program could look for `/images` and read them in without knowing the “physics” of the measurement.
- One can still store all the detailed, non-generic information you want - just use additional non-generic HDF5 groups!

Non-portability

- If you call the primary data group `/maps` or `/hyperspectra` or... then one has to decide...
- If you use “`wavelengths`” or “`keV`” versus “`energies`” then one has to decide...
- If you use NeXuS conventions then you have problems sharing programs and data with electron microscopists, CARS microscopists, satellite imagers...
- Use HDF5 links to standardized names?

Completeness

- Create a local group like `/nsls_x1a` to store all local beamline parameters, like what detector was used and its settings, motor positions, and so on.
- Include a version number, so that as a beamline evolves you can select what parameters are to be read based on the version number.

History

- Create a group `/history`:
  - Name and date of originally recorded raw data file
  - Text lines that record processing steps carried out on the data
- This could be automatically added to by all analysis programs.
- Can one go to a graphical flow-chart representation of processing steps, such as is done in Amira for tomography data visualization?

Conclusions

- Multidimensional data are all over the place!
- Lots of people have developed algorithms for simplification, classification, and analysis
- There could be more cross-fertilization of methods and software between different scientific communities - even beyond synchrotrons and neutron facilities!
- To communicate, we need a common language: HDF5 with agreed-upon naming conventions.