3D Classification

Marin van Heel

Leiden University/Imperial College London/LNNano Campinas

Strasbourg 7 October 2016

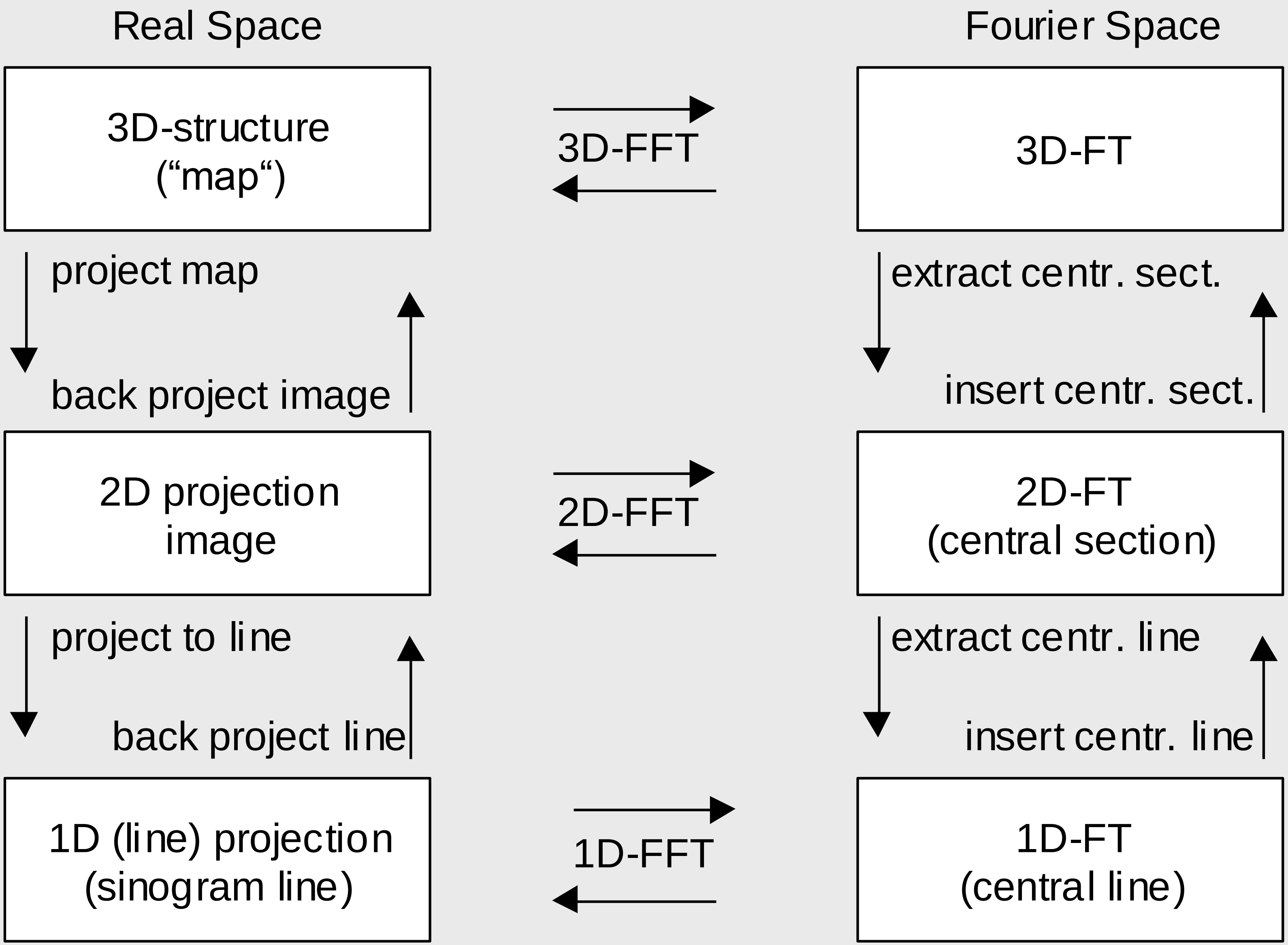
3D Classification

Marin van Heel

Leiden University/Imperial College London/LNNano Campinas

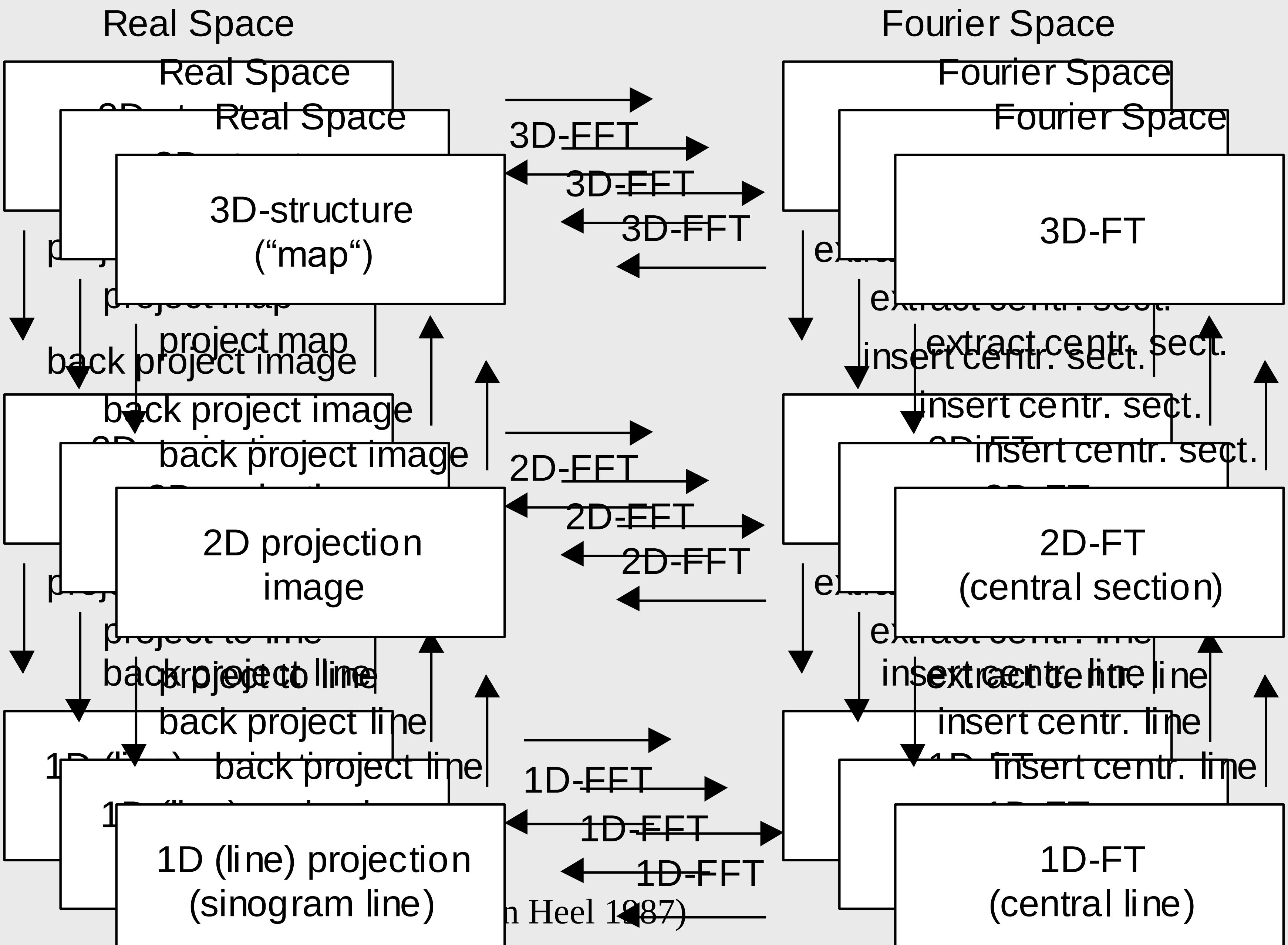
Strasbourg 7 October 2016

Fundamentals operations in reconstructions and projections



(van Heel 1987)

Fundamentals operations with heterogeneous data



3D Classification

Marin van Heel

Leiden University/Imperial College London/LNNano Campinas

Strasbourg 7 October 2016

Classification ...?

Supervised Classification

(looking for a specific thing, matched filtering, find my reference, reference bias)

Unsupervised Classification

(comparing everything to everything and let the data speak for itself, avoid bias)

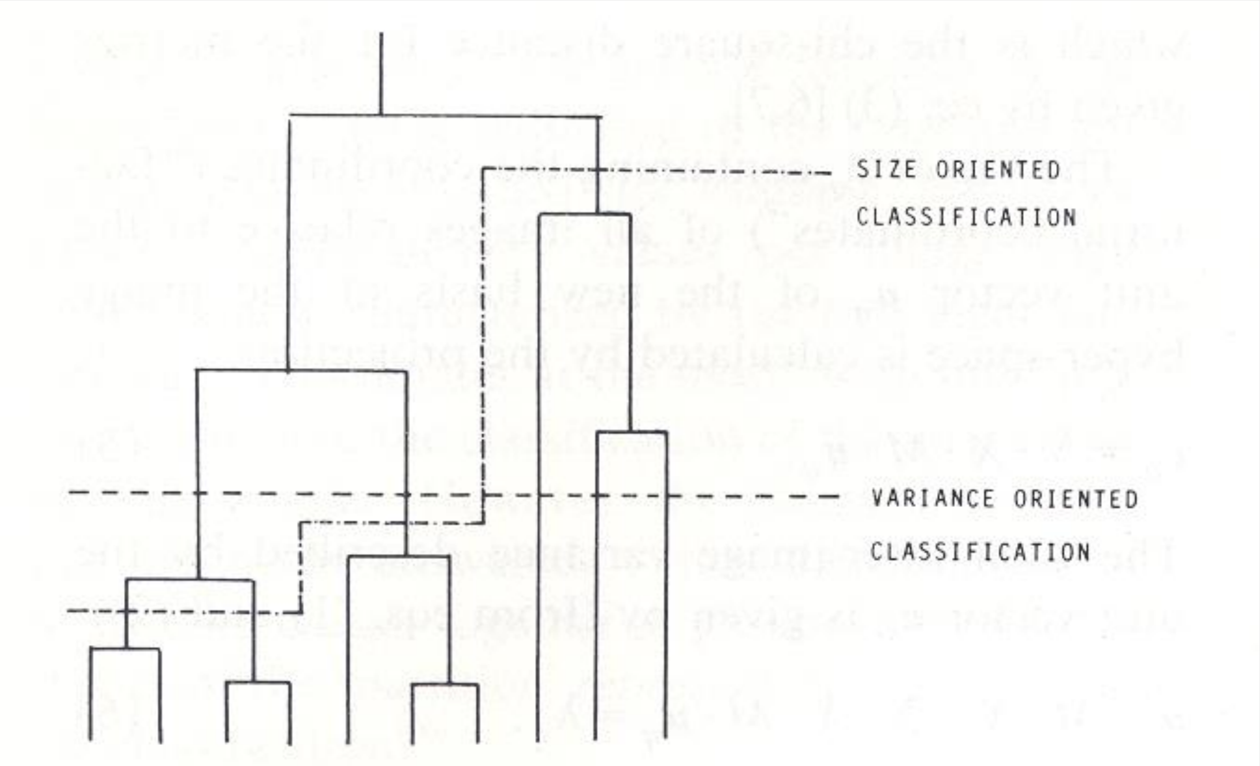
Classification ...?

Nomenclature has become a mess…

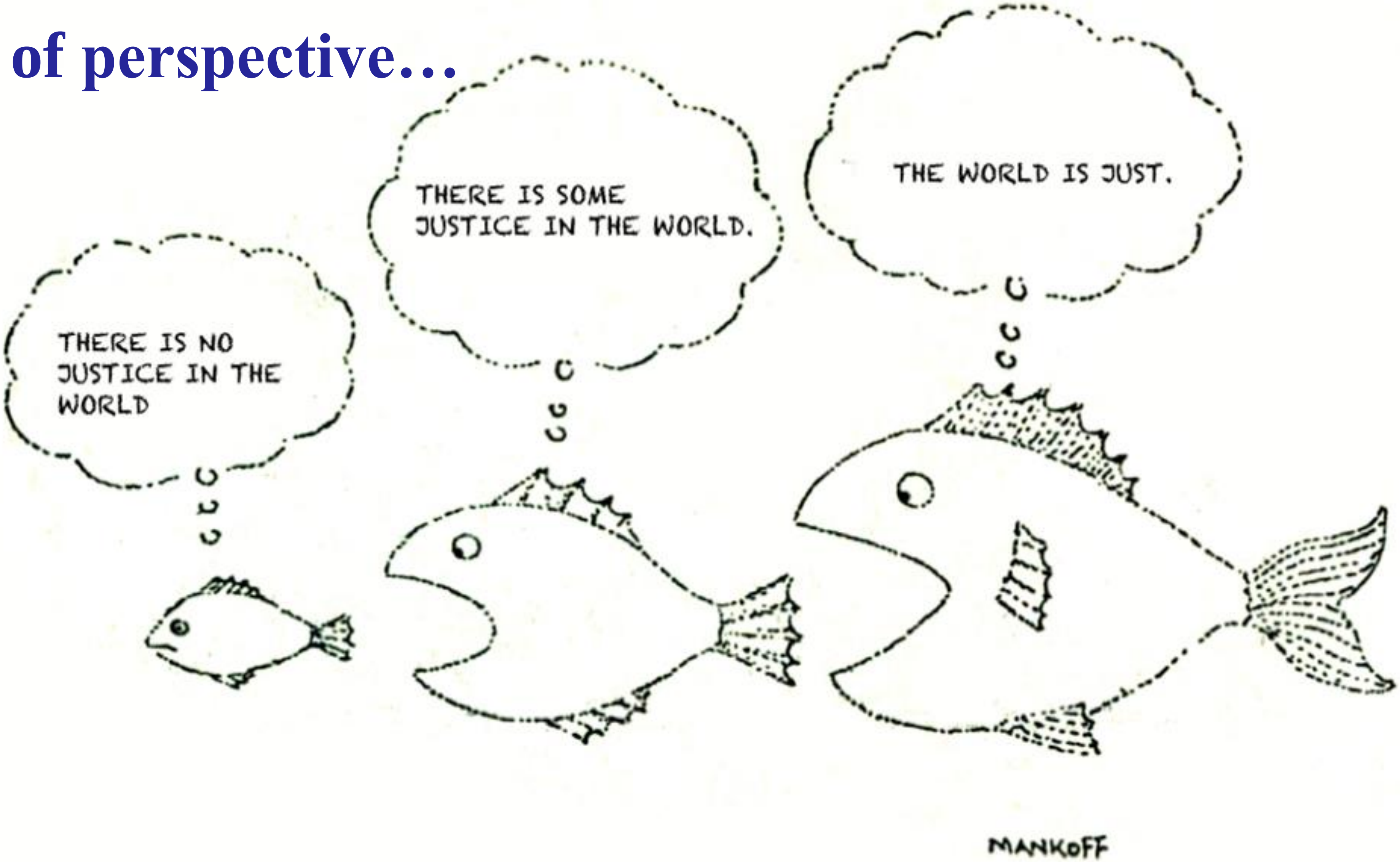
2D Classification ?
3D Classification ?

A priori frustration:

(you decide what makes sense, not the classification program)



Classification is a matter of perspective…



Biology is a mess...

(you need powerful multivariate statistical tools to make sense of it all)

Van Heel M, Portugal RV, Schatz M: **MSA of large datasets single particle electron microscopy**. *OJS* **6** (2016) 701-739. http://dx.doi.org/10.4236/ojs.2016.64059

Multivariate statistics:

is all about distances, correlations…

Correlation / Inner Product $= \sum_{a} F(a) \,.\, G(a)$

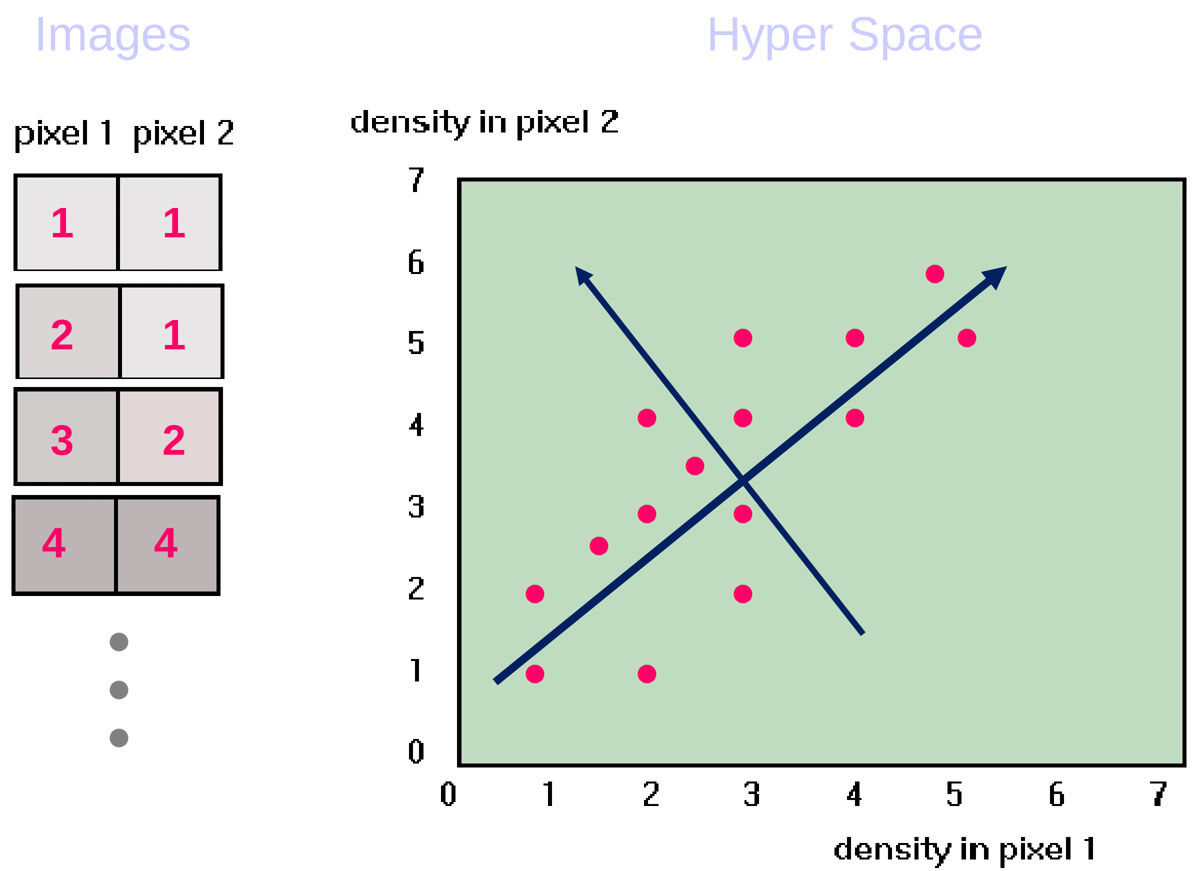
$$(\text{Euclidian Distance})^2 = \sum_{a} \big(F(a) - G(a)\big)^2$$

$$\text{Dist.}^2 = \sum_{a} \big(F^2(a) - 2\,\big(F(a).\,G(a)\big) + G^2(a) \big)$$

$$\text{Dist.}^2 = \sum_{a} F^2(a) \;-\; 2\,.\sum_{a} \big(F(a).\,G(a)\big) \;+\; \sum_{a} G^2(a)\,\big)$$

$$\text{Normalised Signal} = F(a) \;/\; \sqrt{\textstyle\sum_{a} F^2(a)} \;=\; F(a) / SD_F$$

Hyperspace Data Representation of single particle images

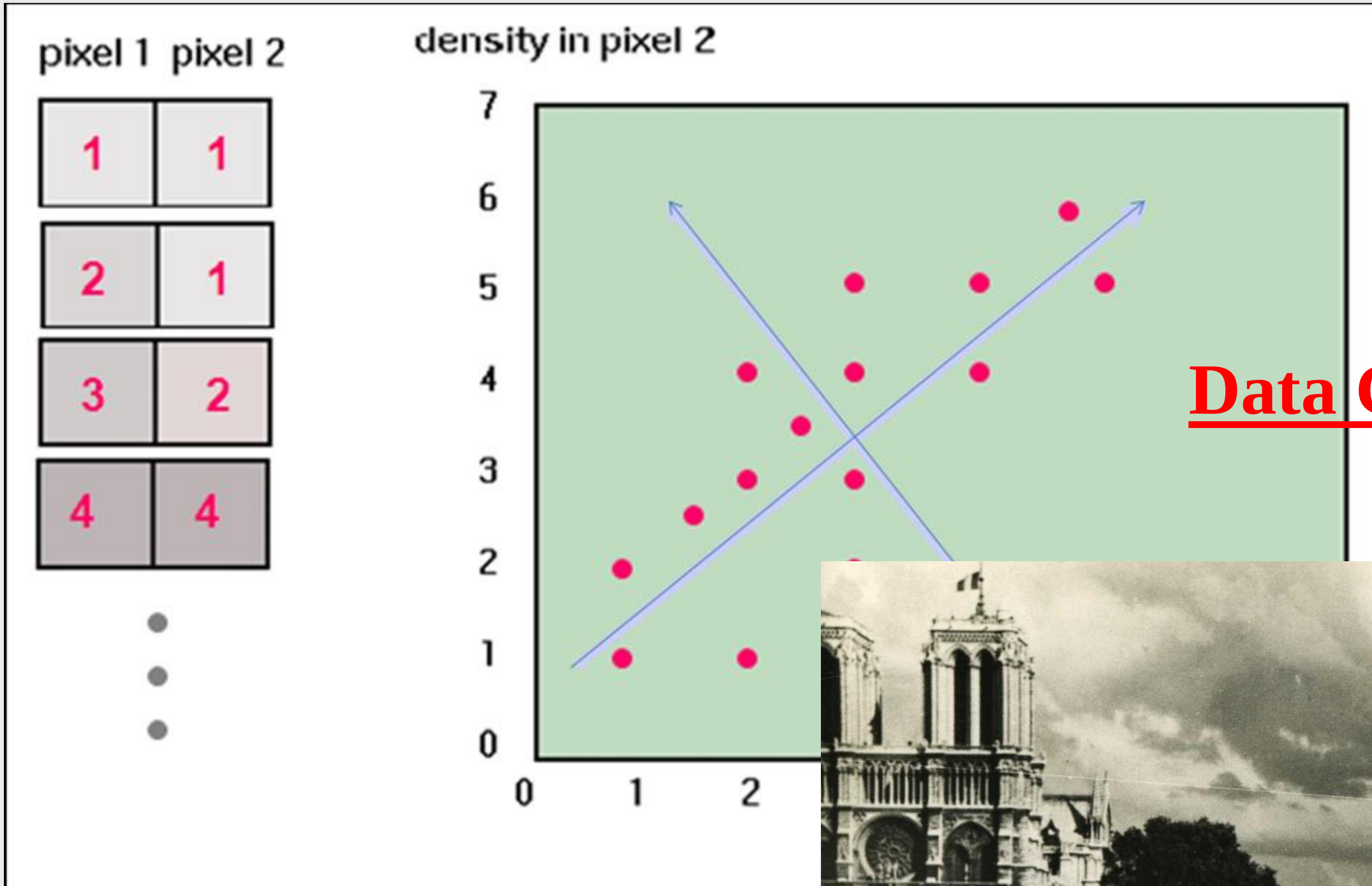


Data Compression!

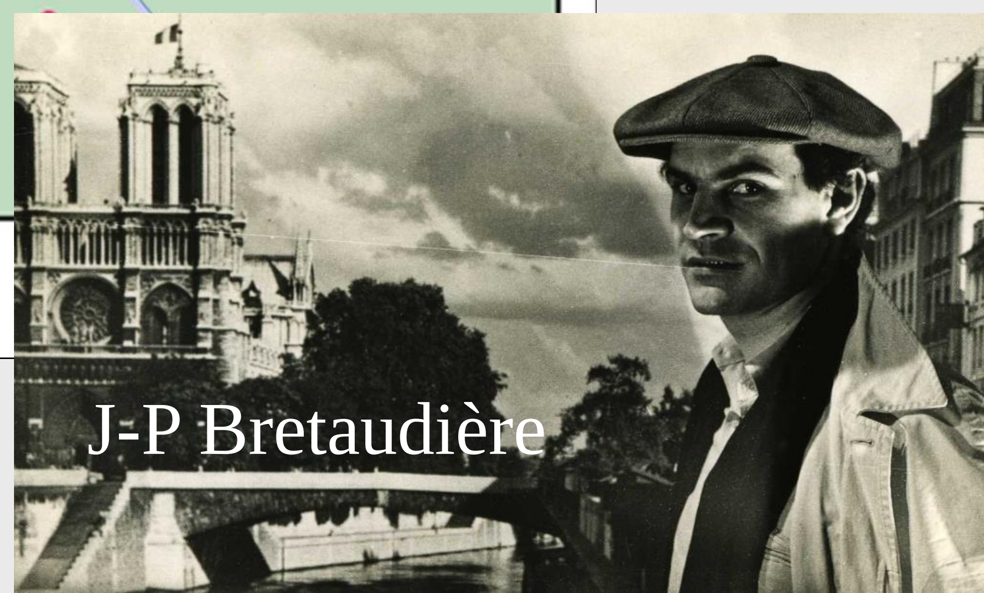
USE OF MULTIVARIATE STATISTICS IN ANALYSING THE IMAGES OF BIOLOGICAL MACROMOLECULES

Ultramicroscopy 6 (1981) 187–194
North-Holland Publishing Company

Marin VAN HEEL Joachim FRANK

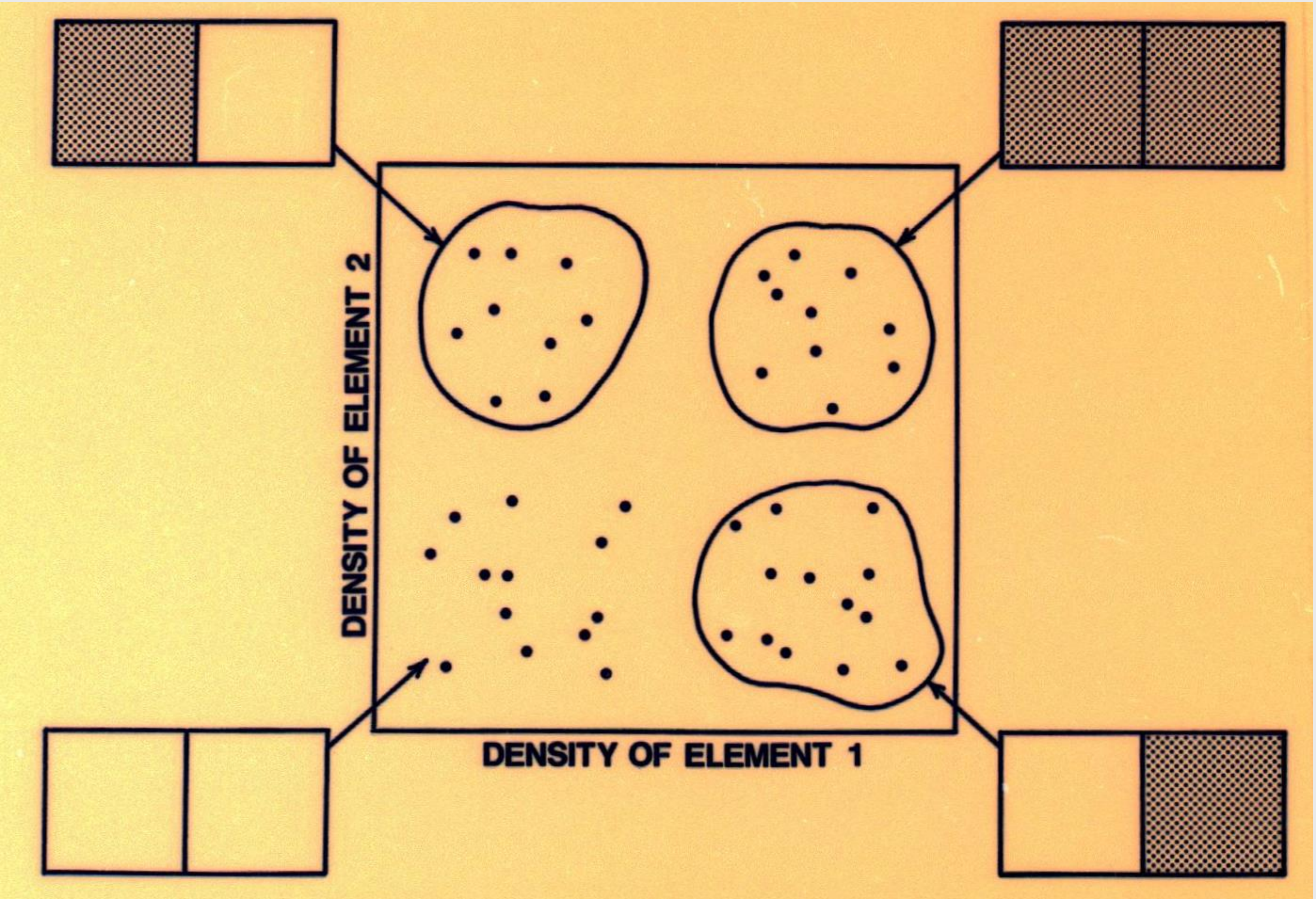


Data Compression!



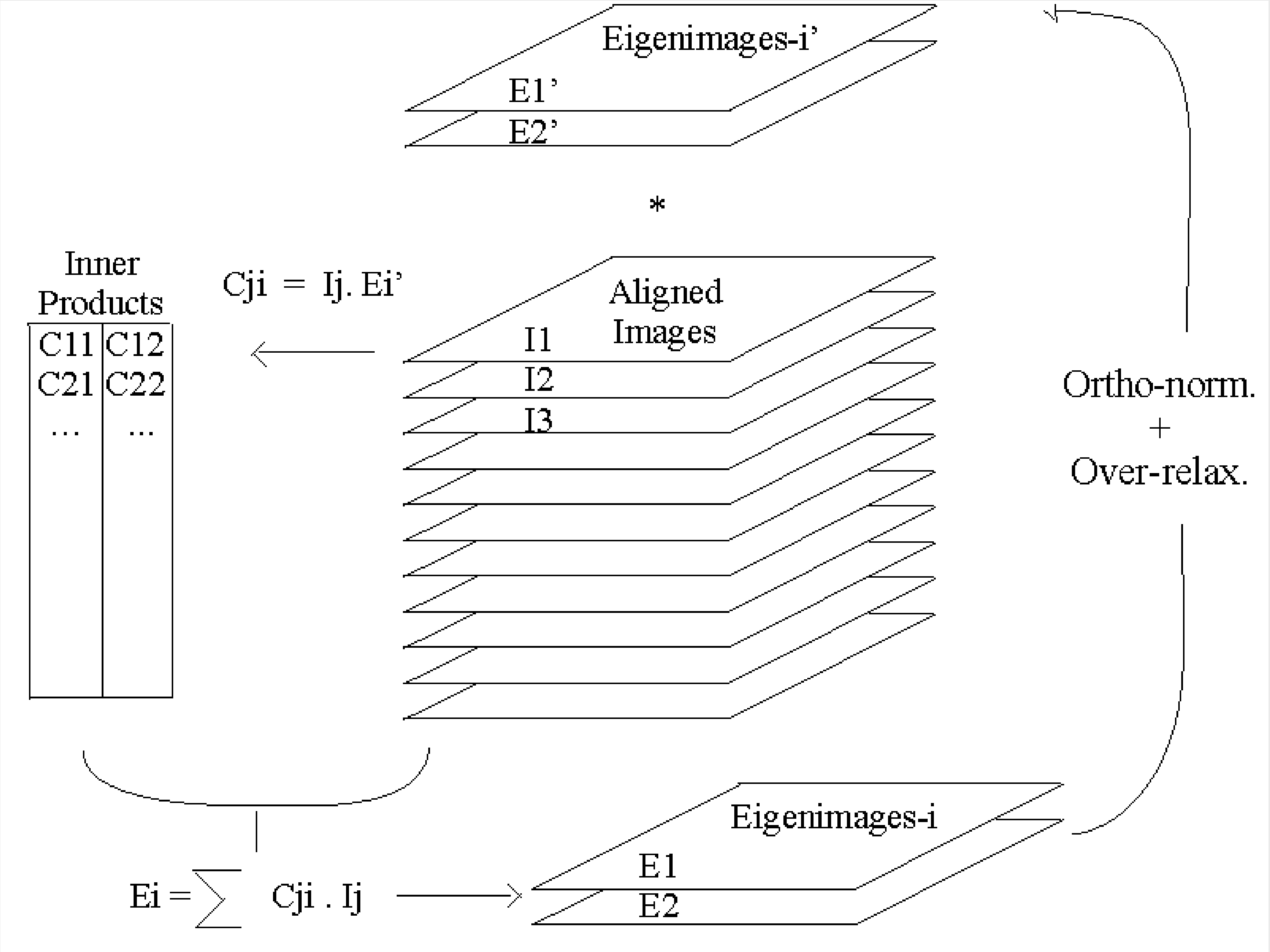
Hyperspace representation of single-particle data

Hyperspace representation and Classification

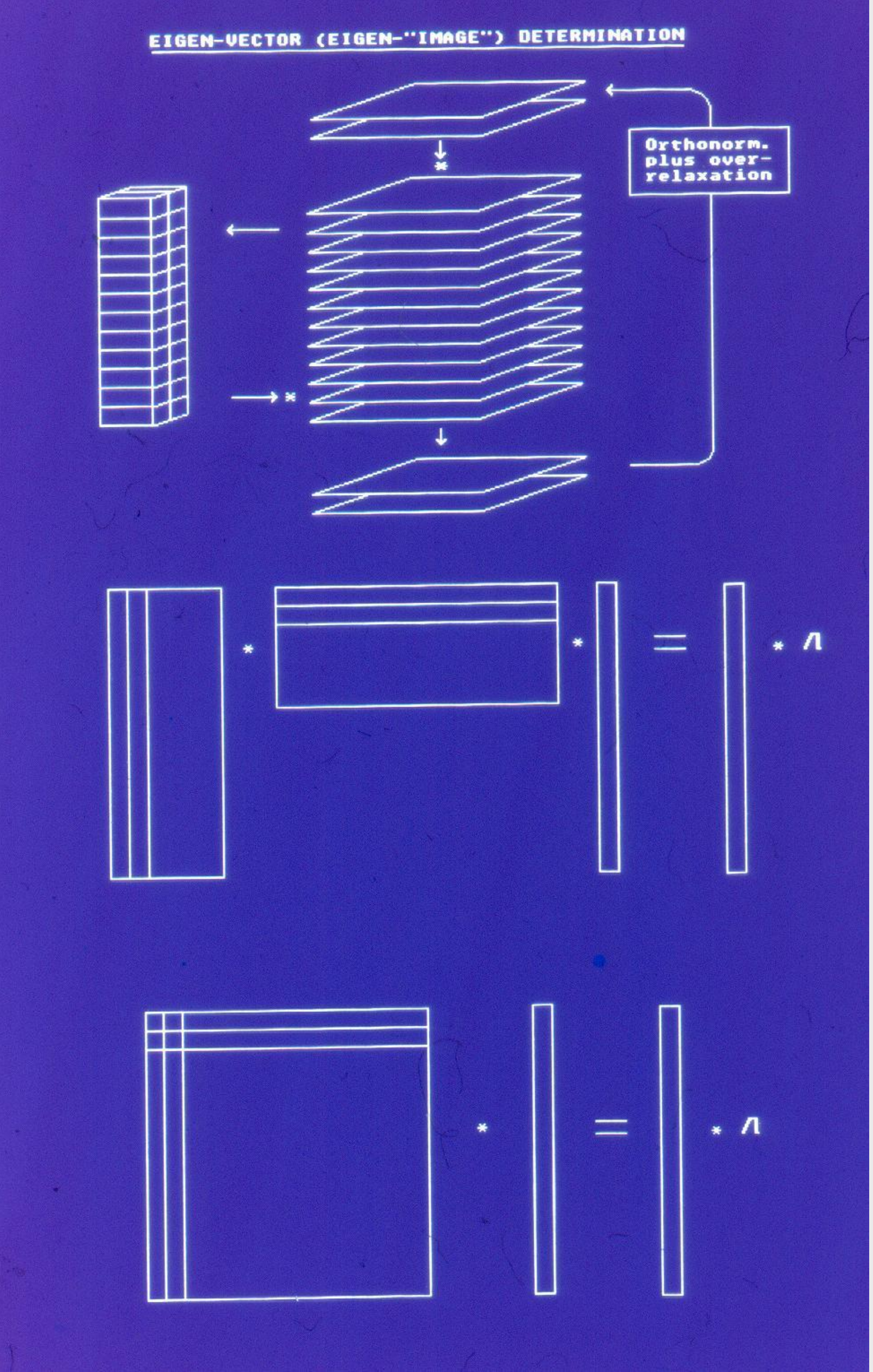


Data Matrix "X"

$$X = \begin{array}{cc} \begin{pmatrix} x_{1,p} & x_{1,2} & x_{1,3} & \cdots & \cdots & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & \cdots & \cdots & x_{2,p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & \cdots & \cdots & x_{n,p} \end{pmatrix} & \mathrm{n} \\ \mathrm{p} & \end{array}$$



(Ref.: Review 2000)



<u>From algorithm to mathematics</u>:

X'. X .U = U. Л

In detail eigenvector equation:

X' . N . X . M . U = U . Л

With orthonormalisation constraint:

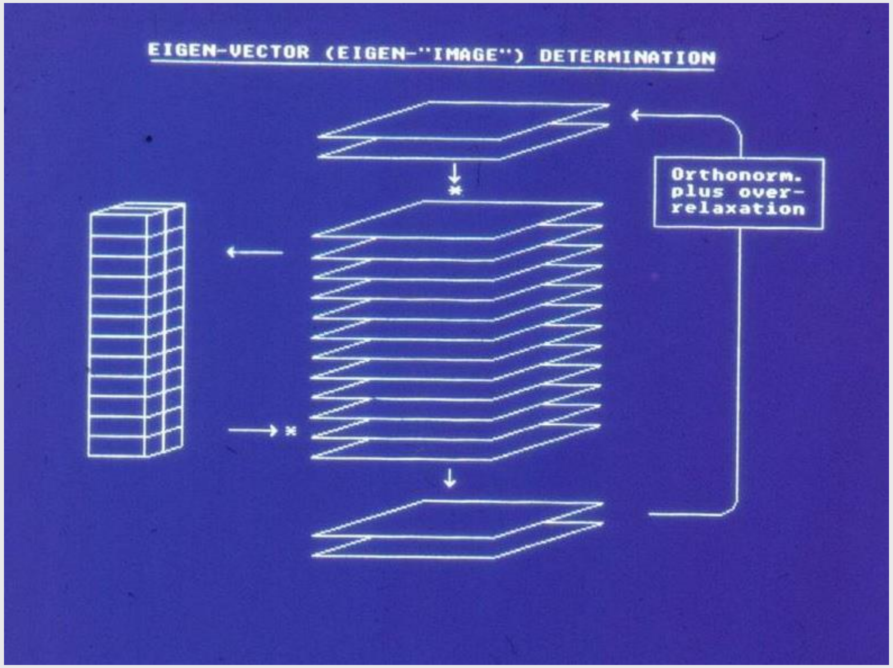
U' . M . U = Ip

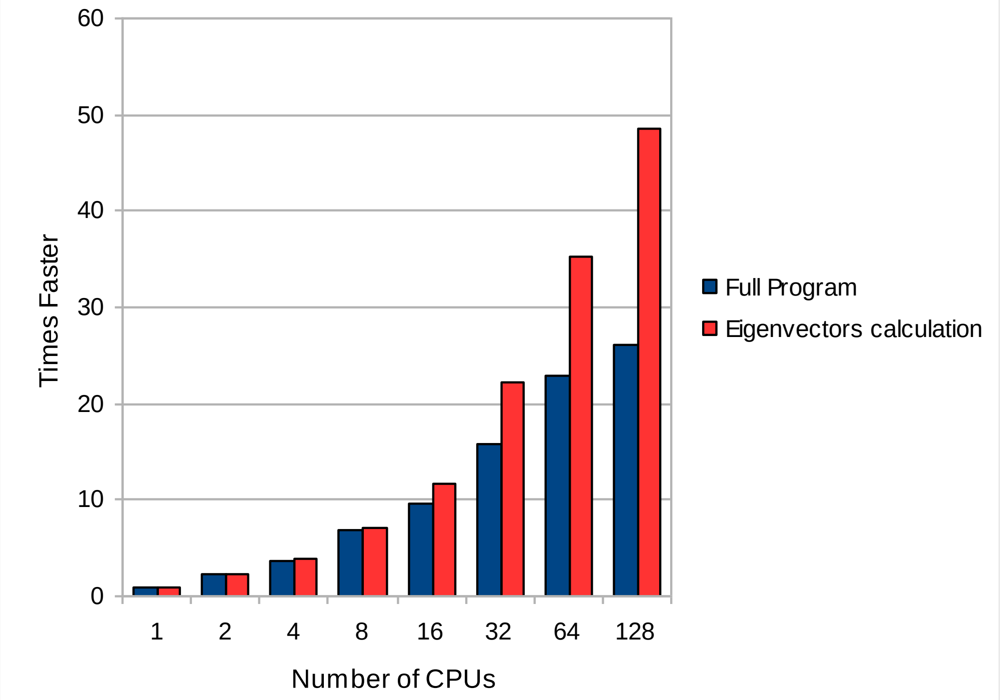
In conjugate space:

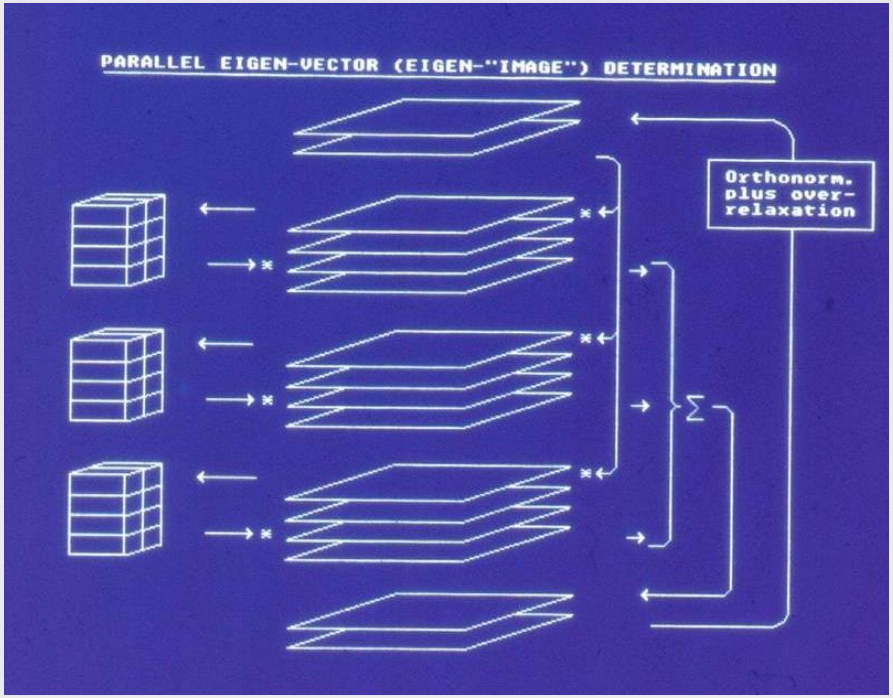
X . M . X' . N . V = V . Л

With orthonormalisation constraint:

U' . M . U = Ip



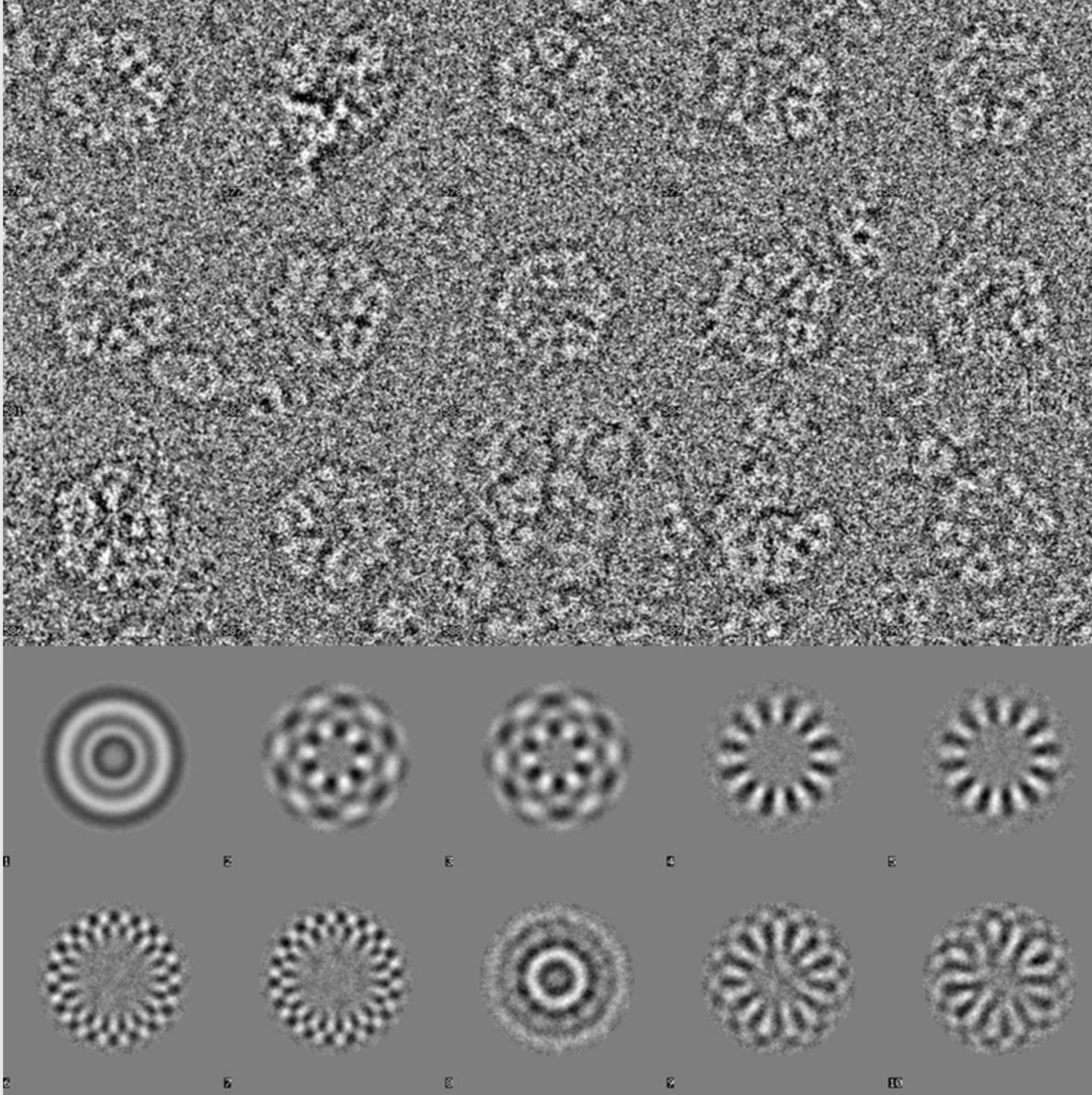


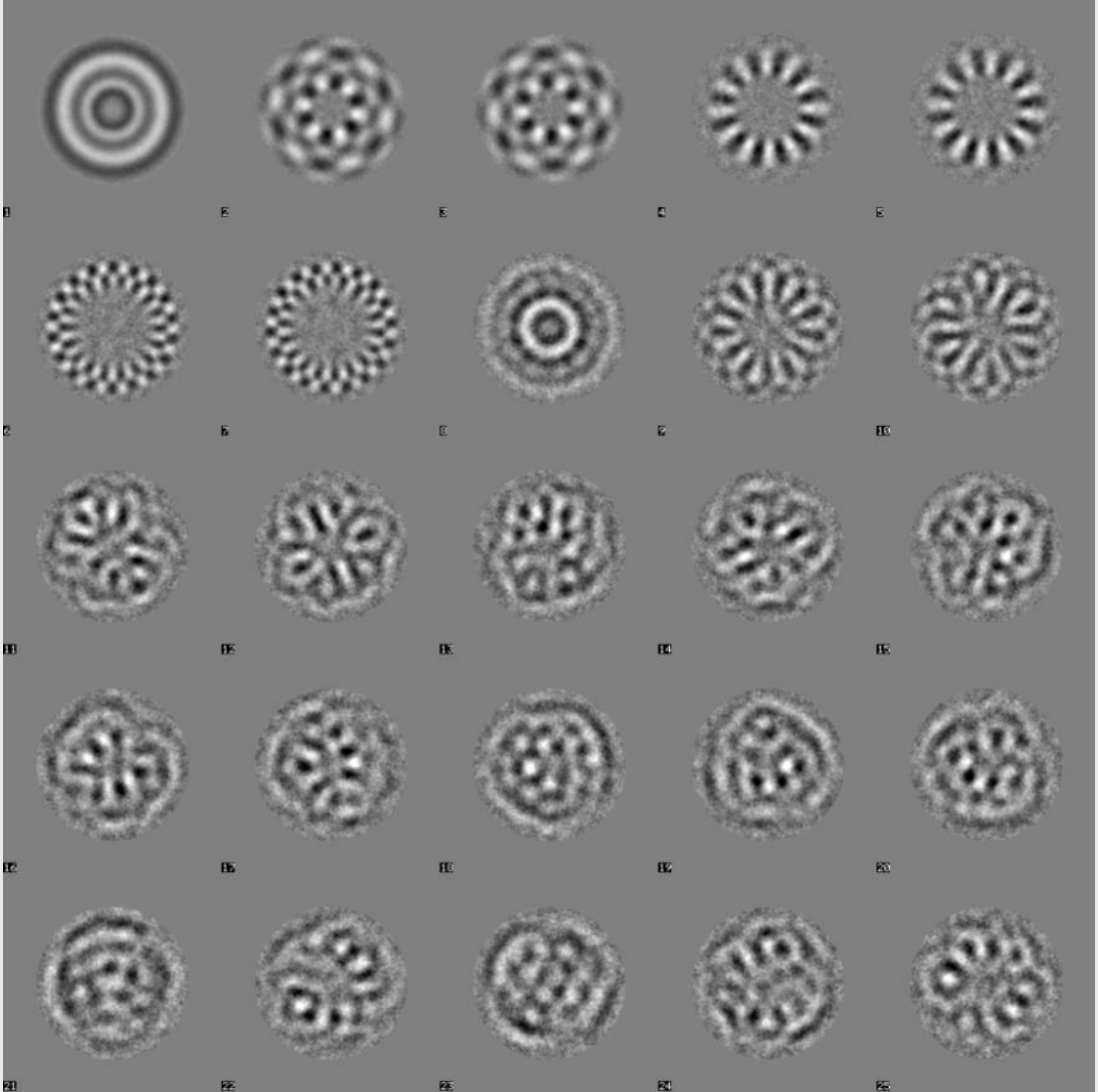


MSA Parallel Programming

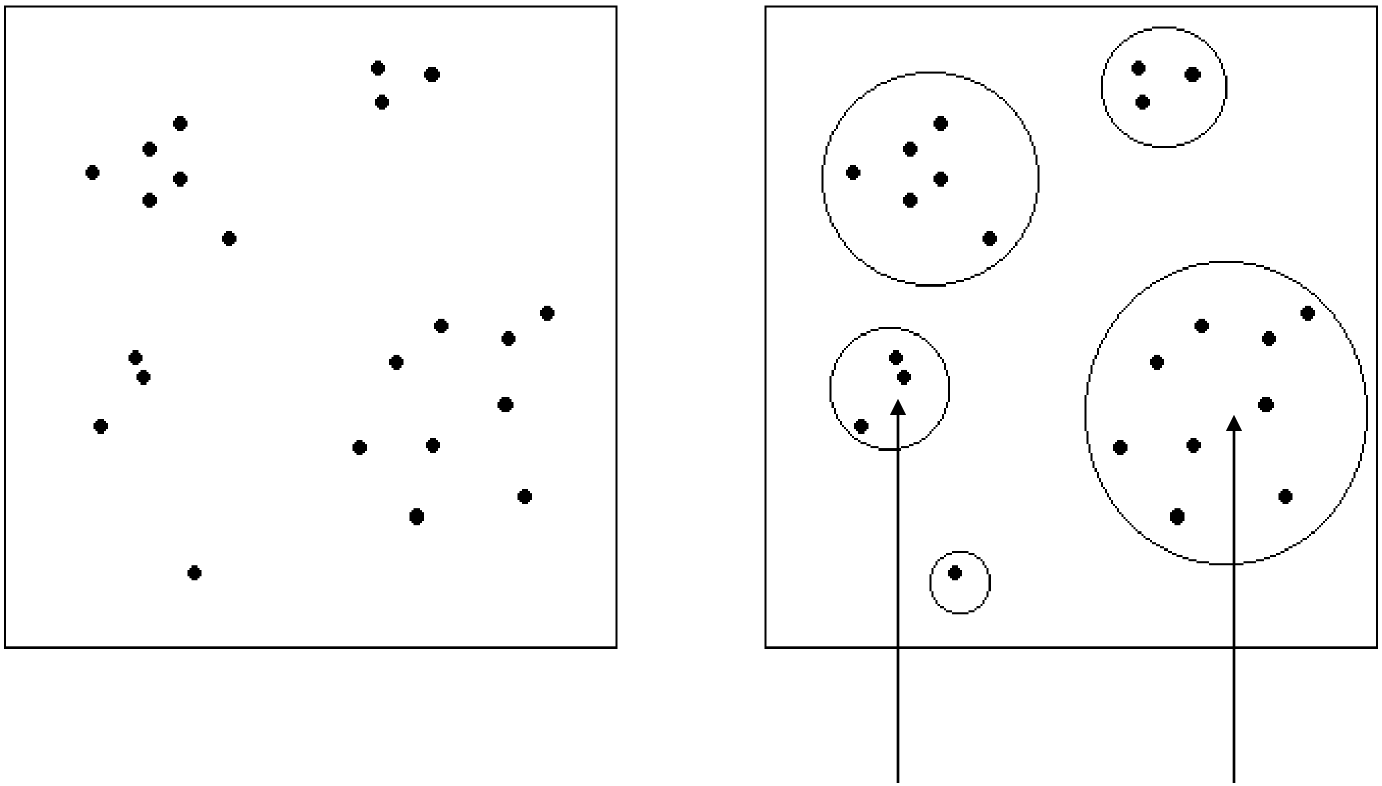
Parallel MSA and its scaling with the number of available CPUs (tested on 32 nodes with each 4 CPUs). The calculations necessary for standard MSA algorithm (top left) are distributed over the available CPUs (lower left). I/O is parallelized by copying the relevant part of the huge input data file to the local scratch file available on each node. Overall speed increase with the current version of the MSA (see text) and images of size 256x256 is around 27 times (Full Program).

MSA analysis of *Lumbricus Terrestris* hemoglobin





Automatic Classification



Average images in each class

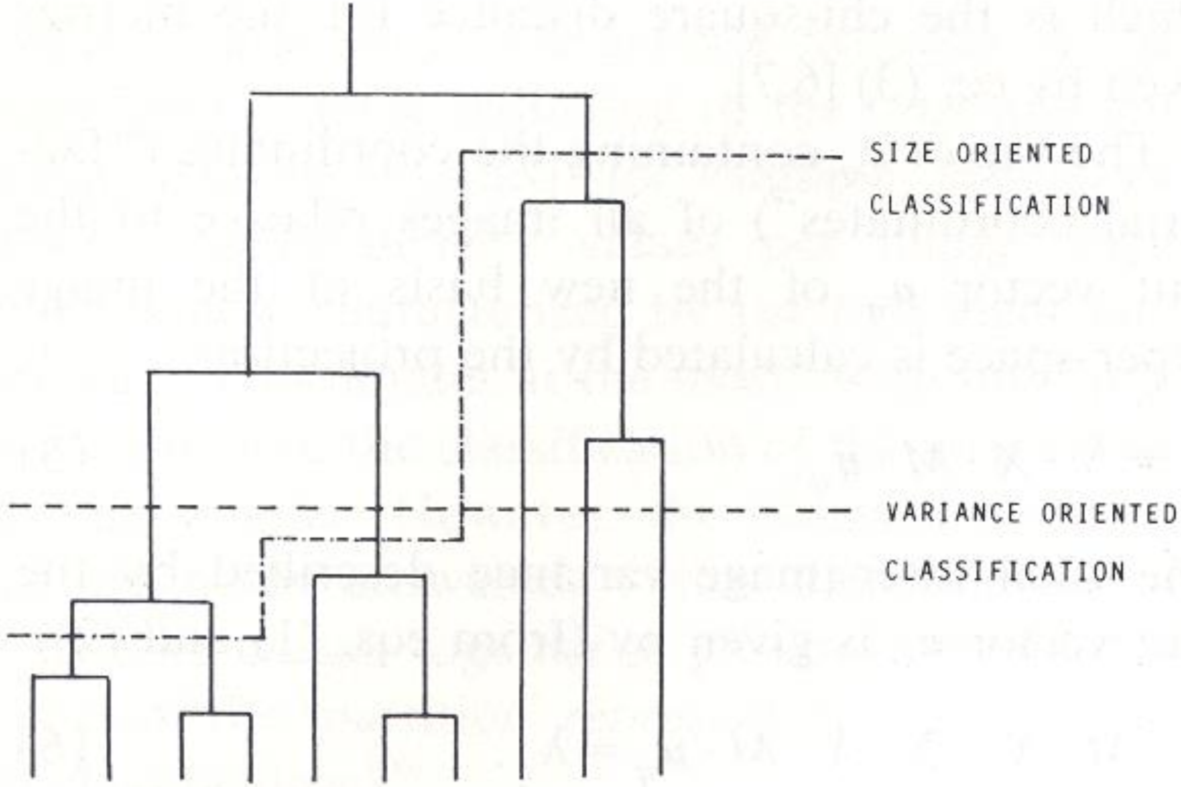


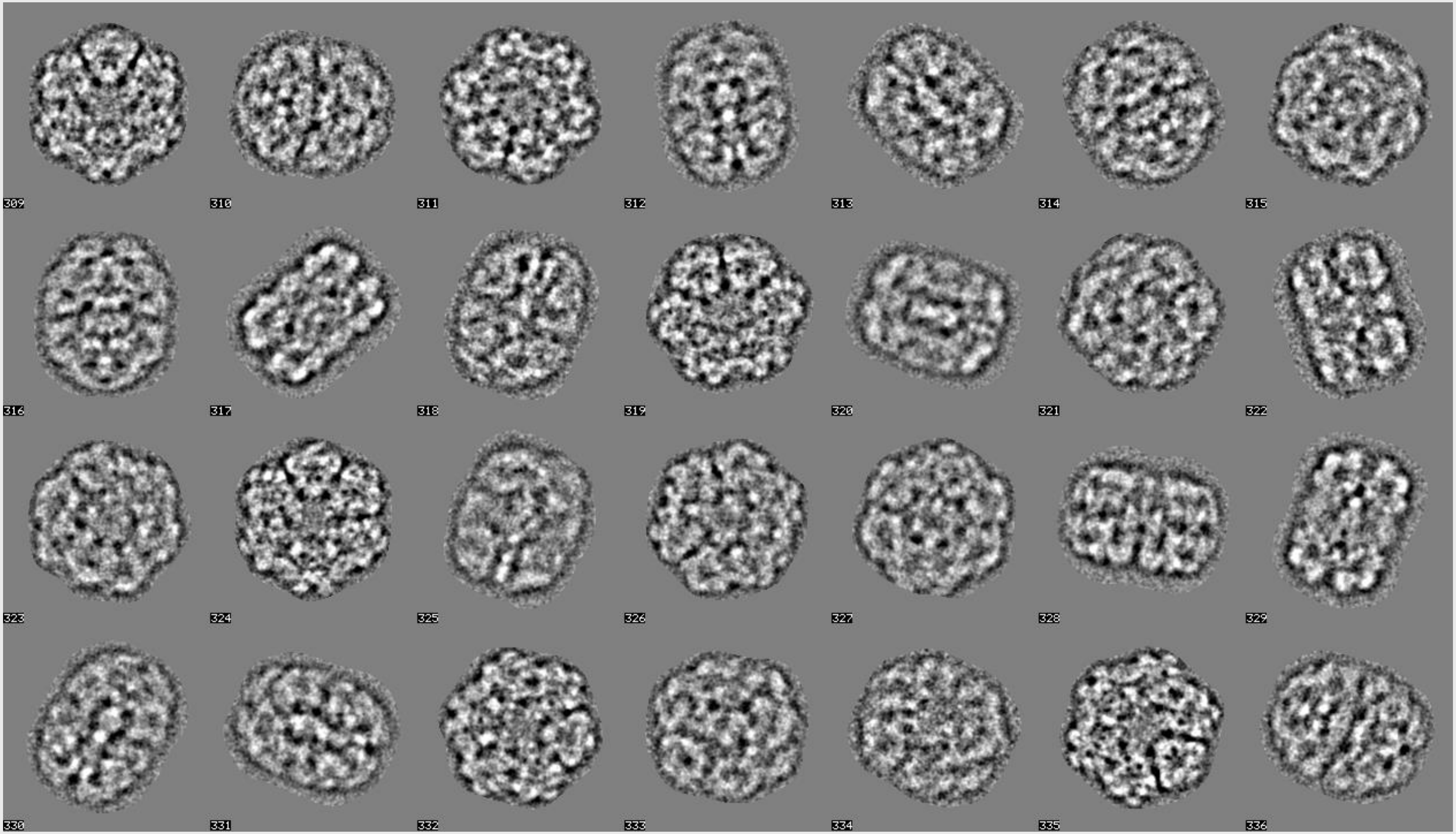
Fig. 1. An example of a hierarchical classification tree. In an hierarchical ascendant classification the procedure starts at the bottom of the figure with as many classes as there are images (10 in this example). The two classes that are closest together in terms of a classification criterion are merged into a larger class. The straight cut through the classification tree leads to a variance-oriented partitioning. A useful alternative is to follow the tree up and down, to obtain a class-size-oriented partition. For details see text.

$$\text{Add. Var.}_{i,i'} = \frac{w_i w_{i'}}{w_i + w_{i'}} d^2_{i,i'}$$

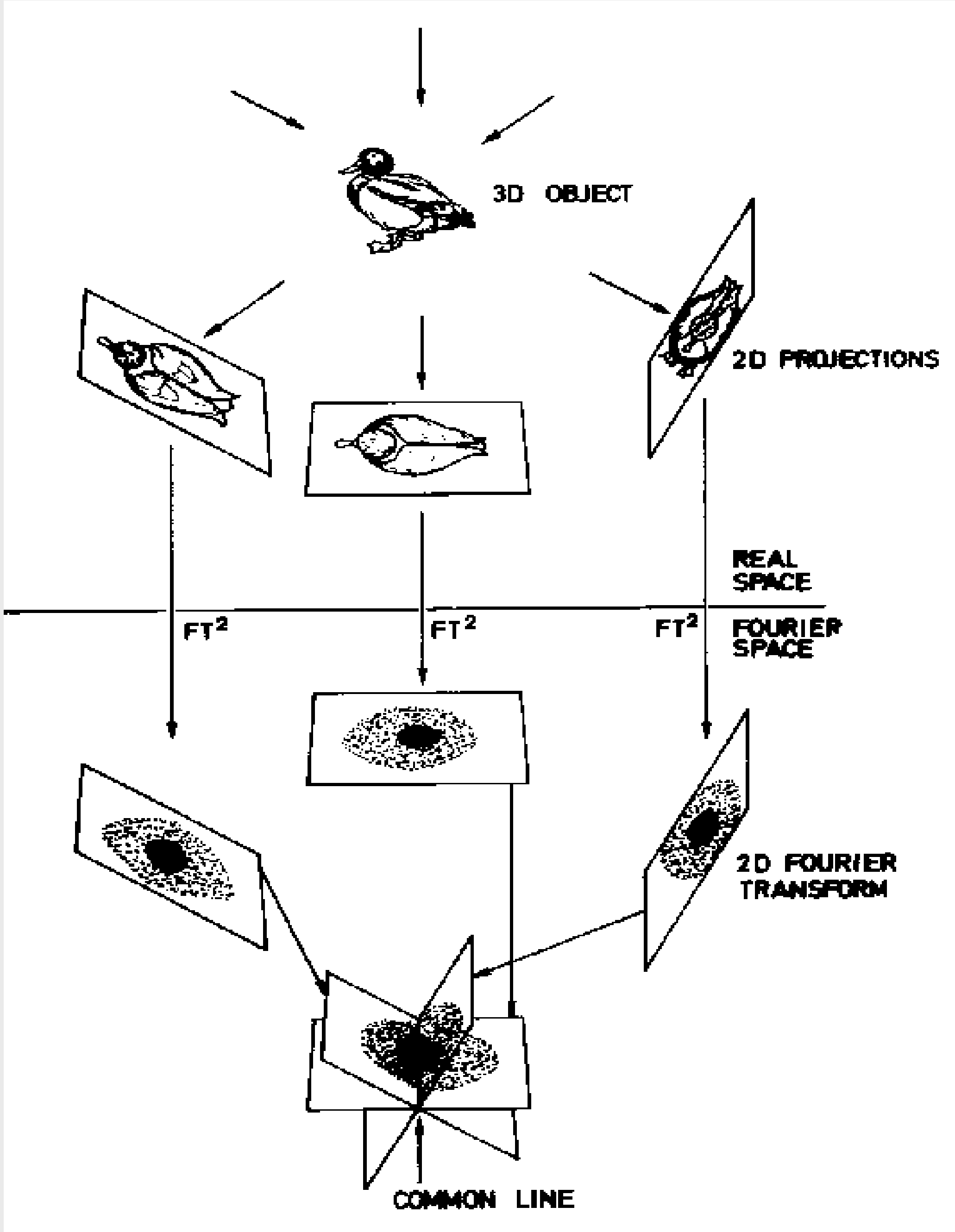
Ward Criterion

ABC-4D (Alignment by classification)

1000 class averages (144,000 particles)



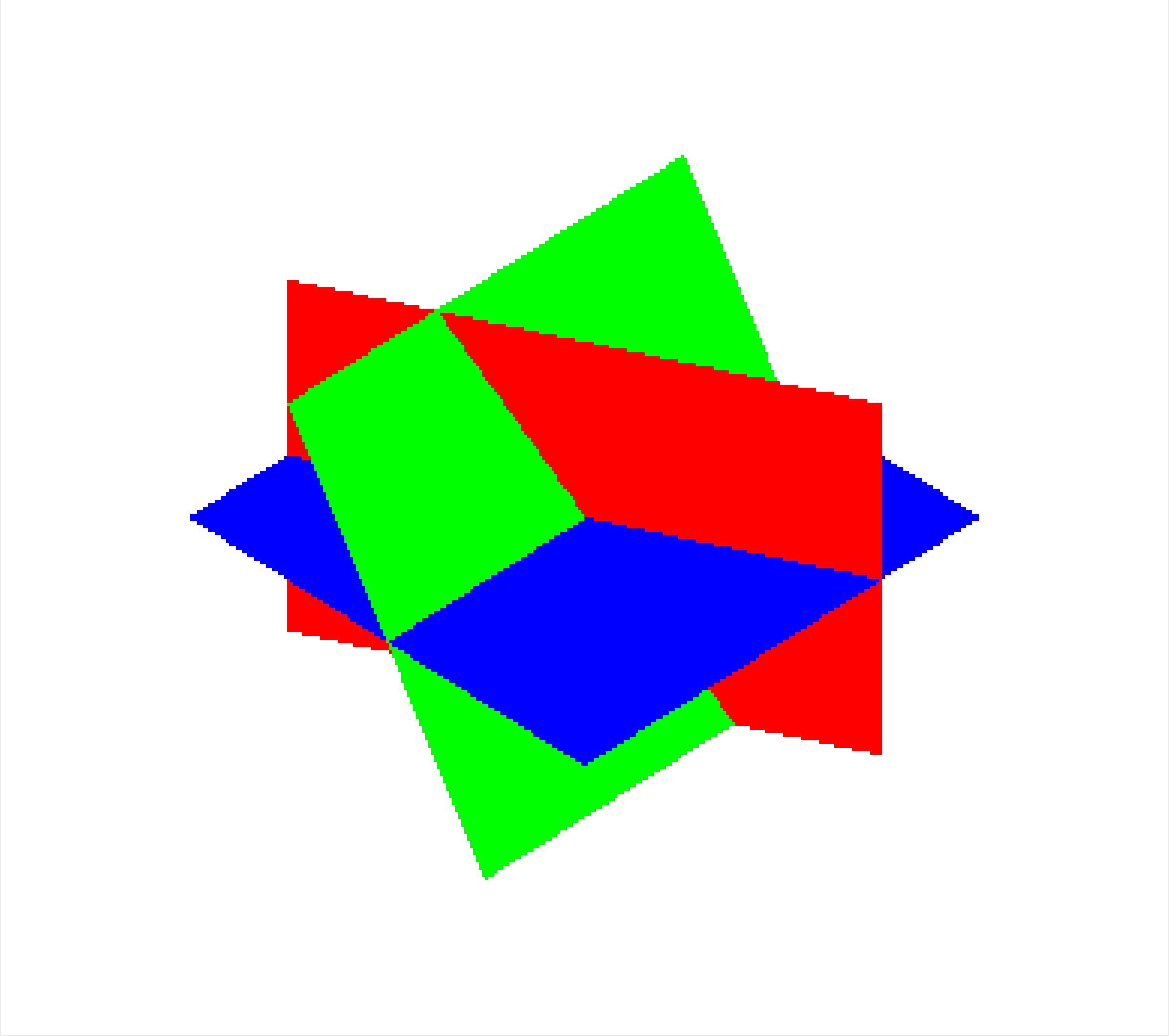
From 2D classes to 3D Structure(s)



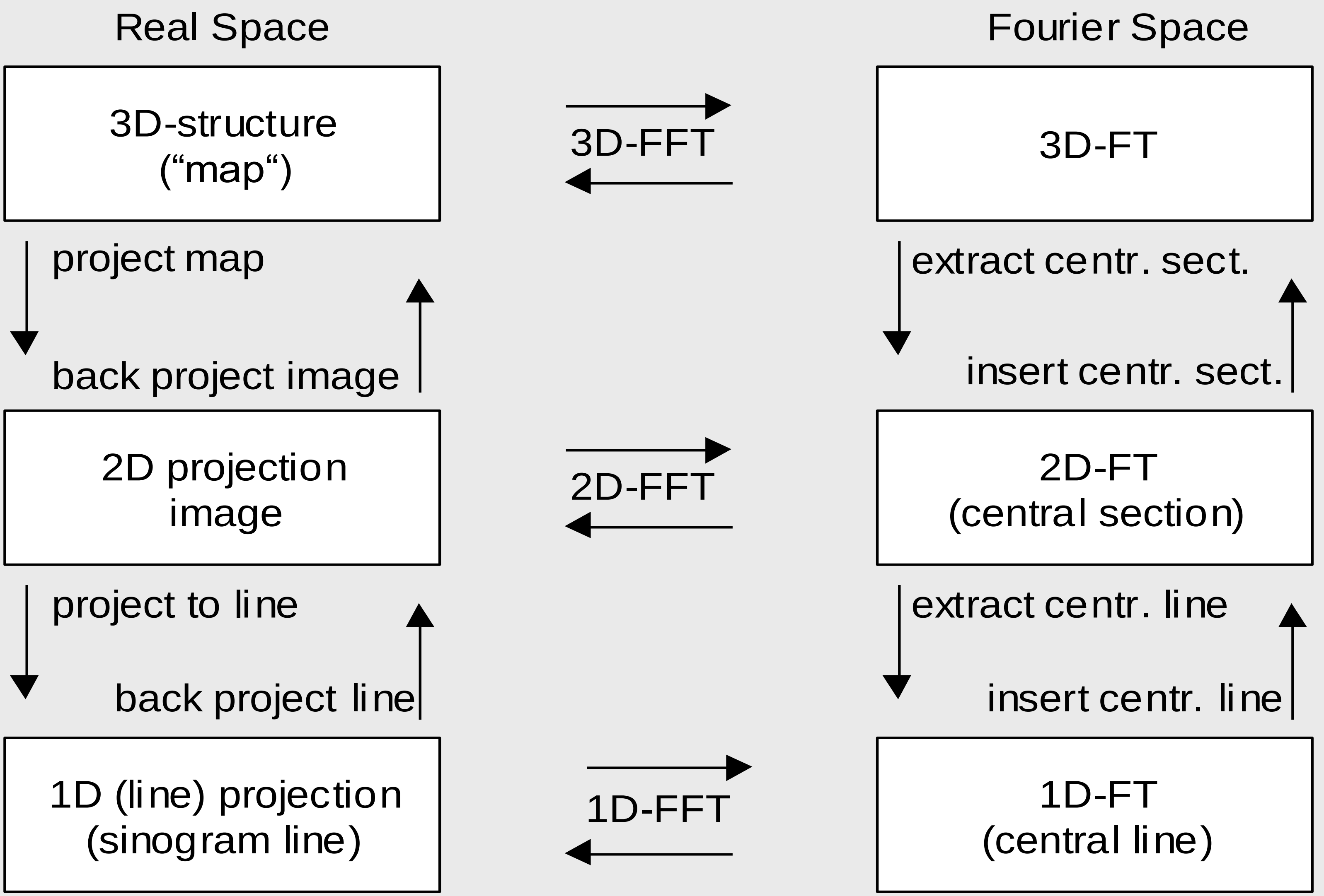
Three-Dimensional reconstruction from projections

We first need to find the relative angles to do a 3D reconstruction…

Intersecting Central Sections
(DeRosier Klug 1968; Crowther1971

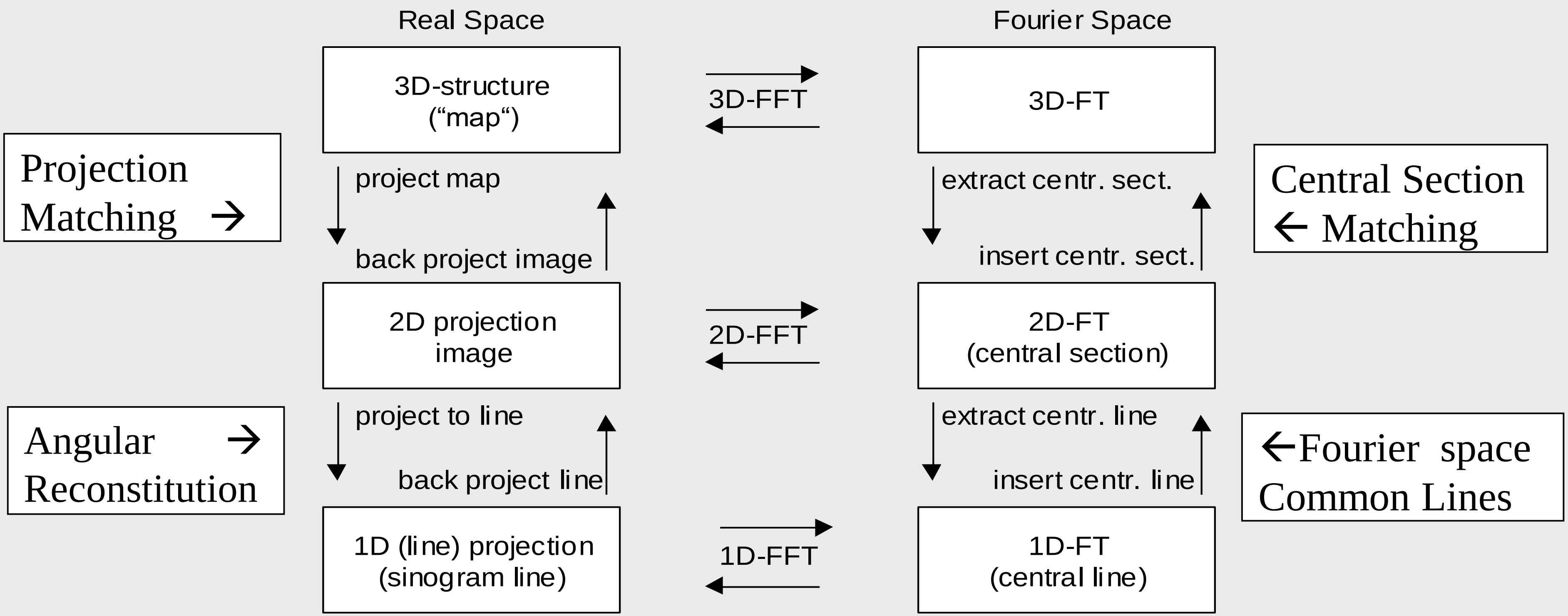


Fundamentals operations in 3D reconstructions and projections

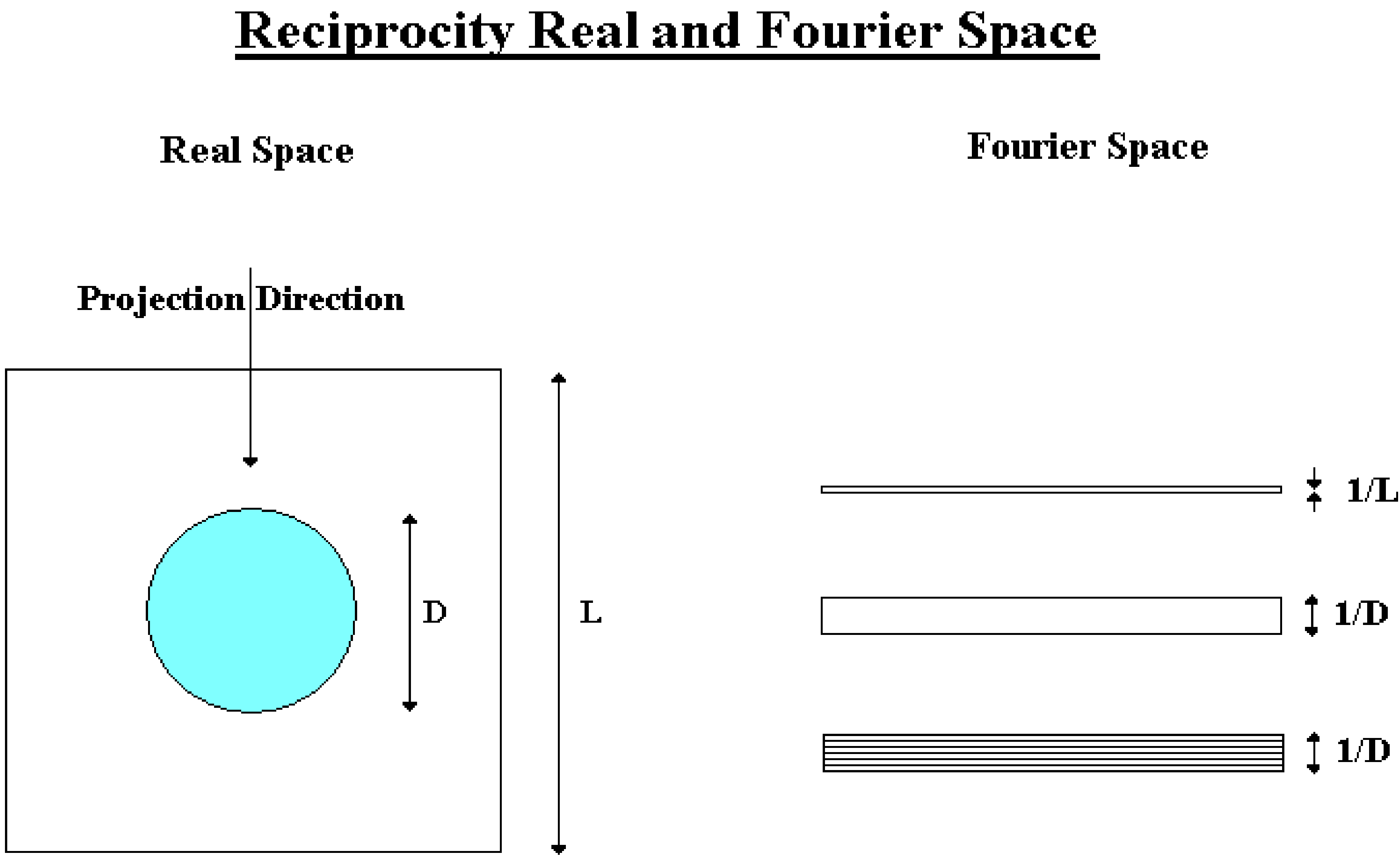


(van Heel *1987*)

Fundamentals operations in reconstructions and projections



Width of Central Section: Central Section Slab



Overlapping central sections in Fourier space

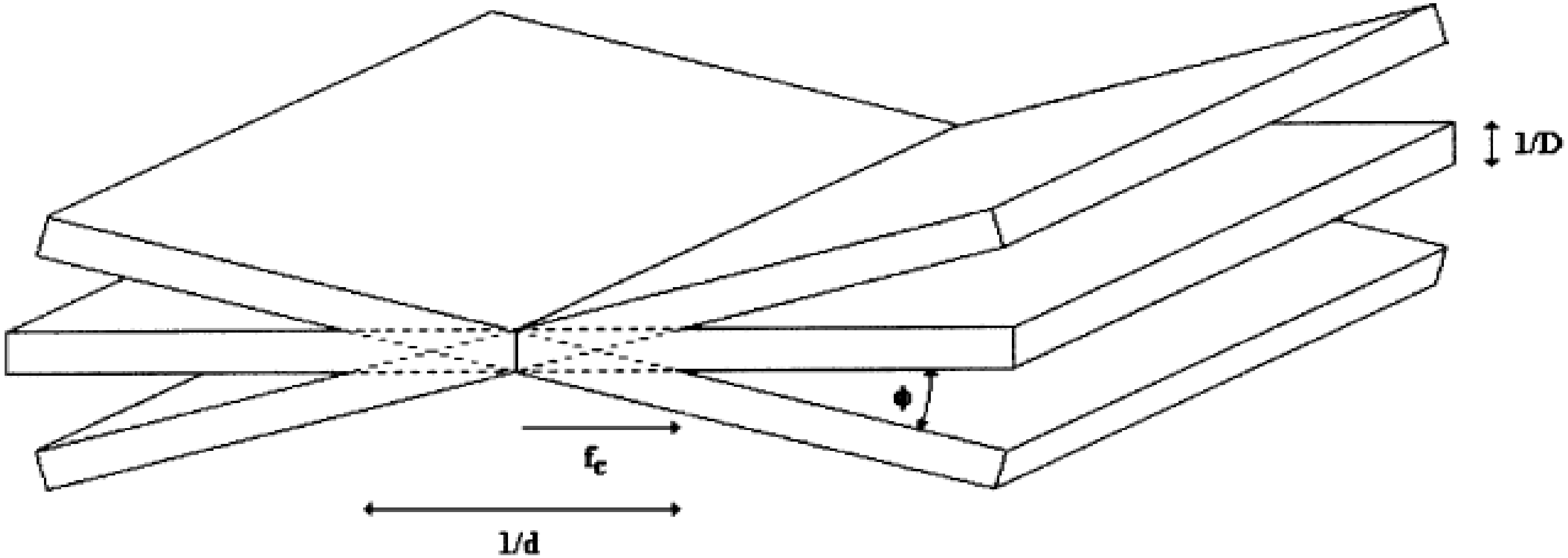


Fig. 14. Overlapping central sections. Fourier-space central sections, associated with 2D projections a 3D object of linear size 'D', have a width '$1/D$'. Central sections in Fourier space always overlap at very low frequencies, that is, close to the origin. Neighbouring central section, separated by an angle ϕ, cover largely the same information up to spatial frequency 'f_c'. The overlap of central sections is fundamental in both 3D reconstruction algorithms and in determining the highest isotropic resolution achievable for a given 3D reconstruction geometry. The areas of the 3D Fourier space volume not covered by the central section 'slabs' are not measured and are referred to as 'missing cone' or 'missing wedge' depending on the 3D reconstruction geometry.

Early **projection matching** experiments: Budapest 1984 Abstract

Figure 1. Continuous stereographic representation of "phantom" used to investigate properties of self-optimizing 3D reconstruction

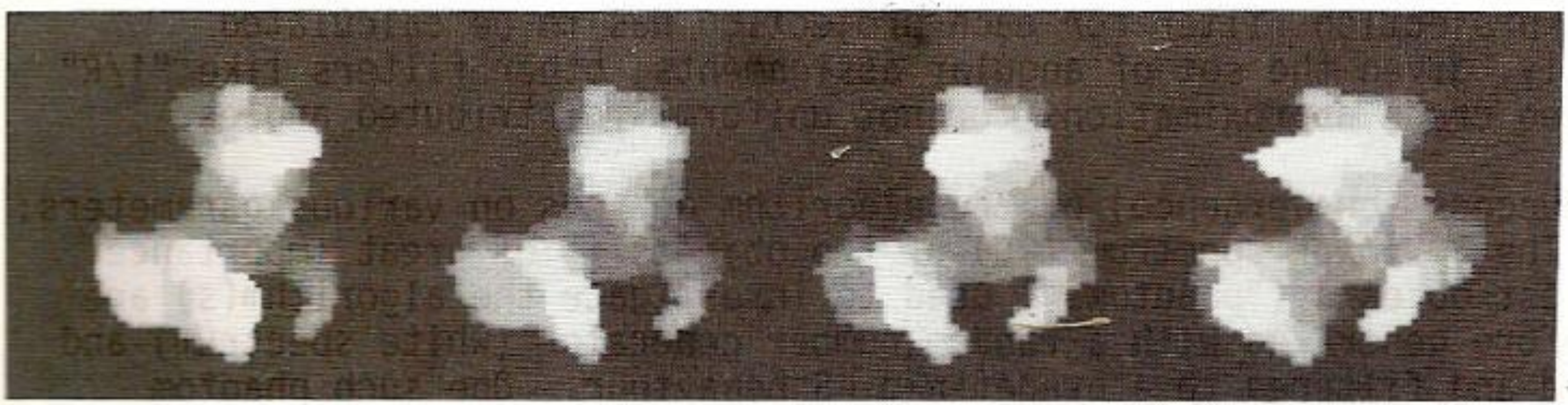
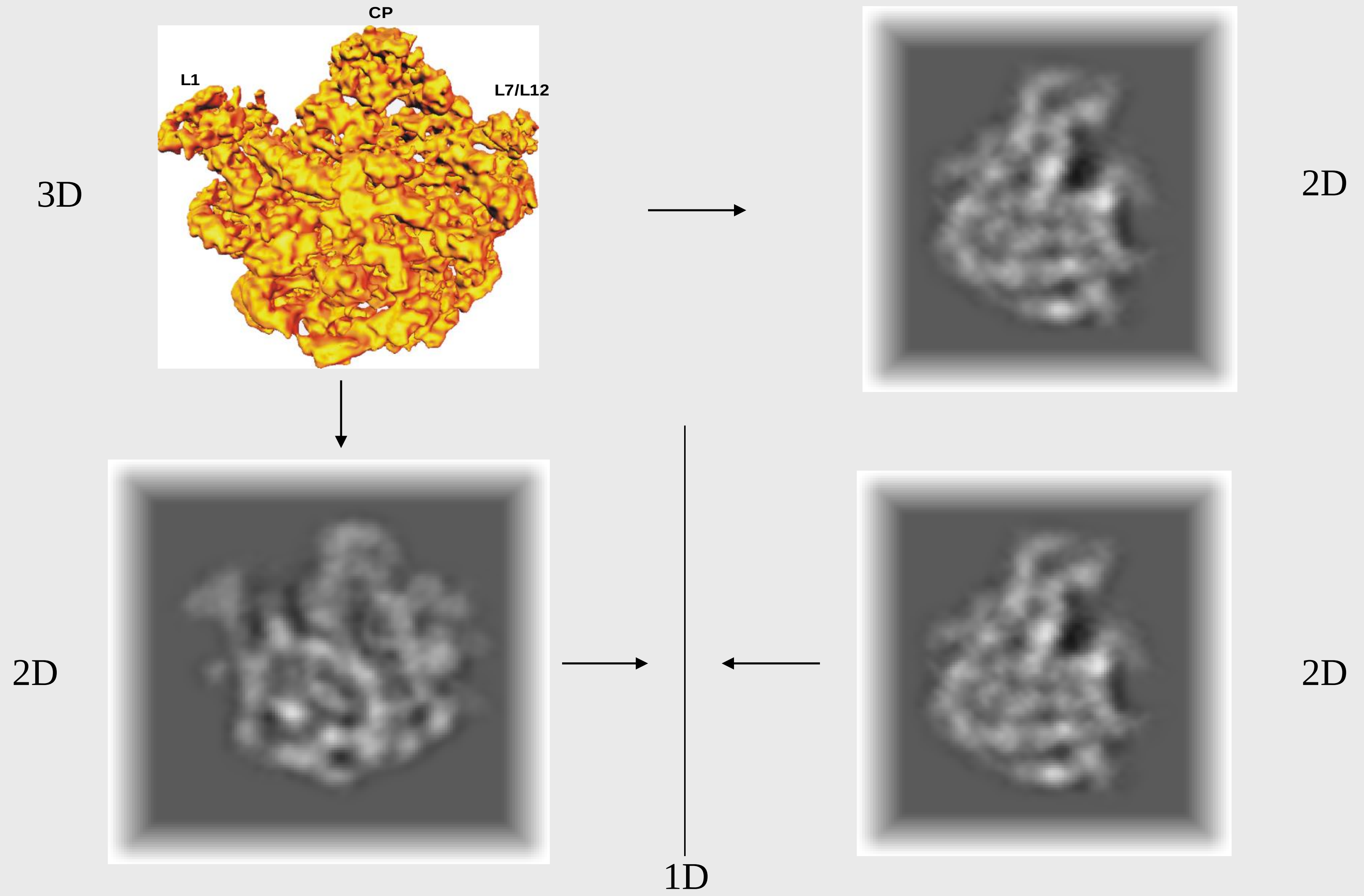
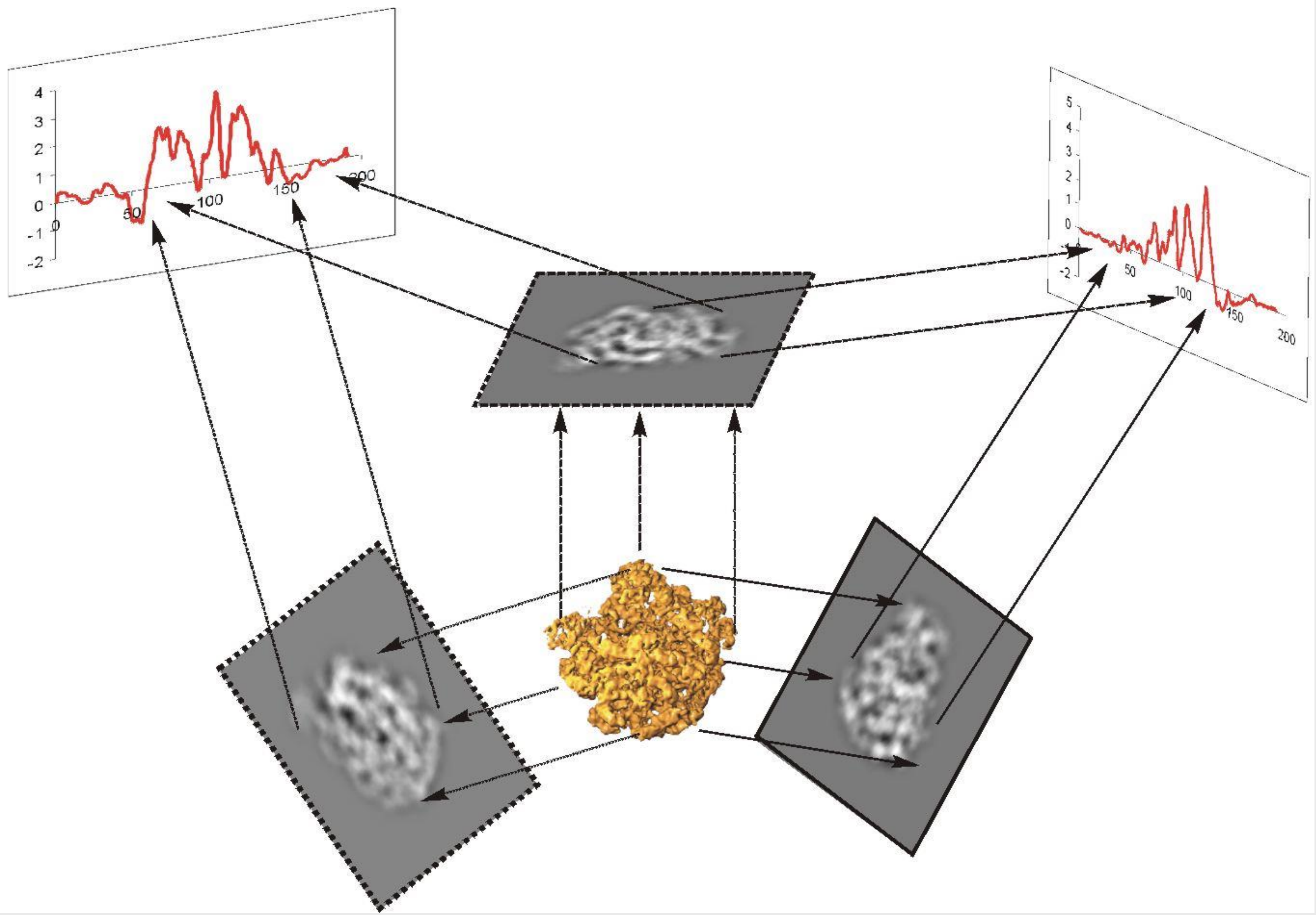


Figure 2. Stereo representation of "phantom" reconstructed using random starting angles; viewed from same directions as above.

Angular Reconstitution

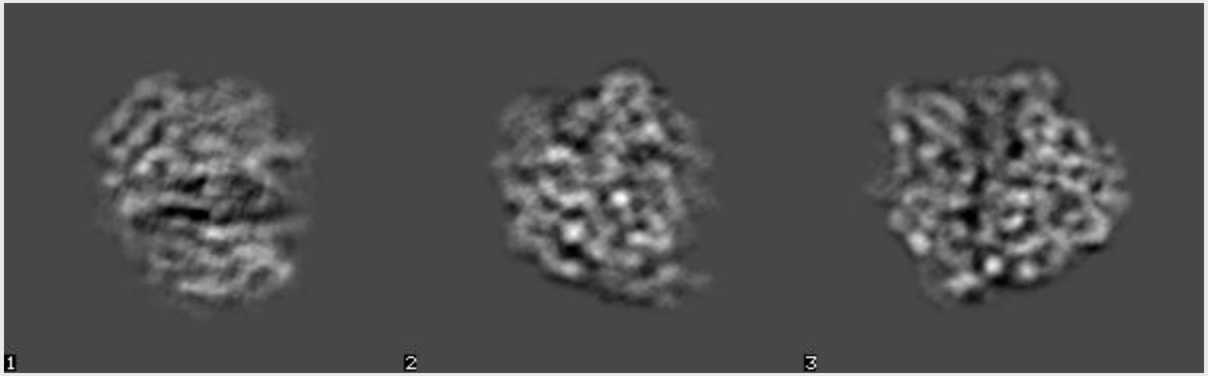


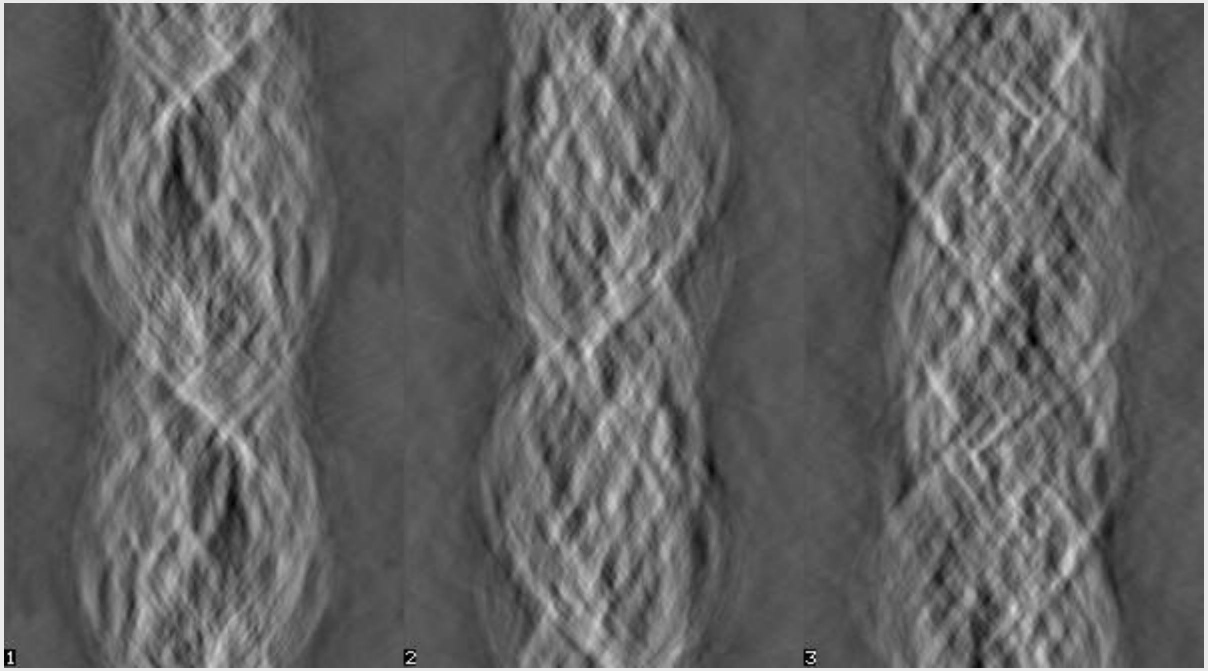
(MvH 1987)

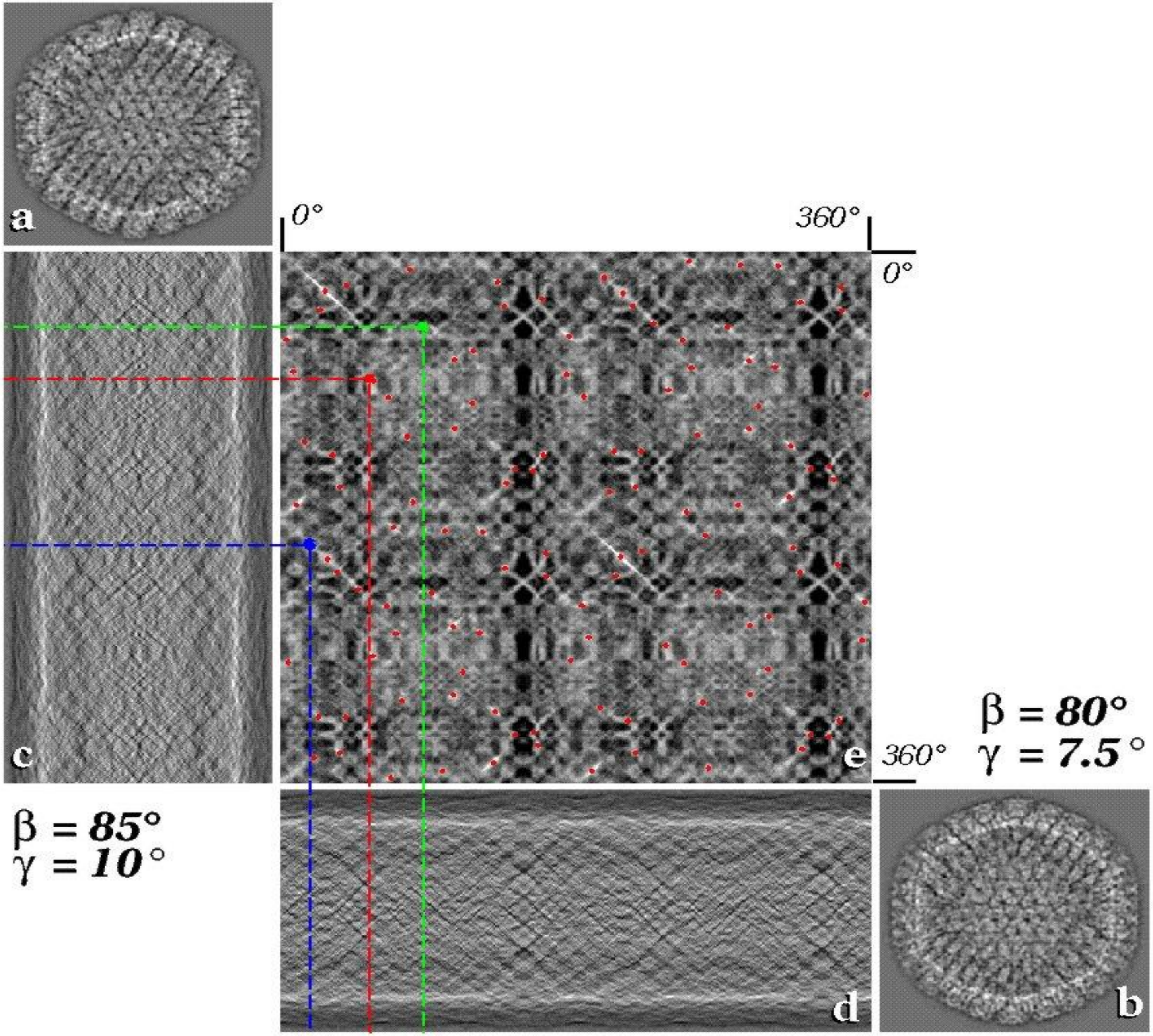


4
3
2
1
0
-1
-2
0
50
100
150
200
5
4
3
2
1
0
-2
50
100
150
200

70S *E. coli* projection plus sinograms

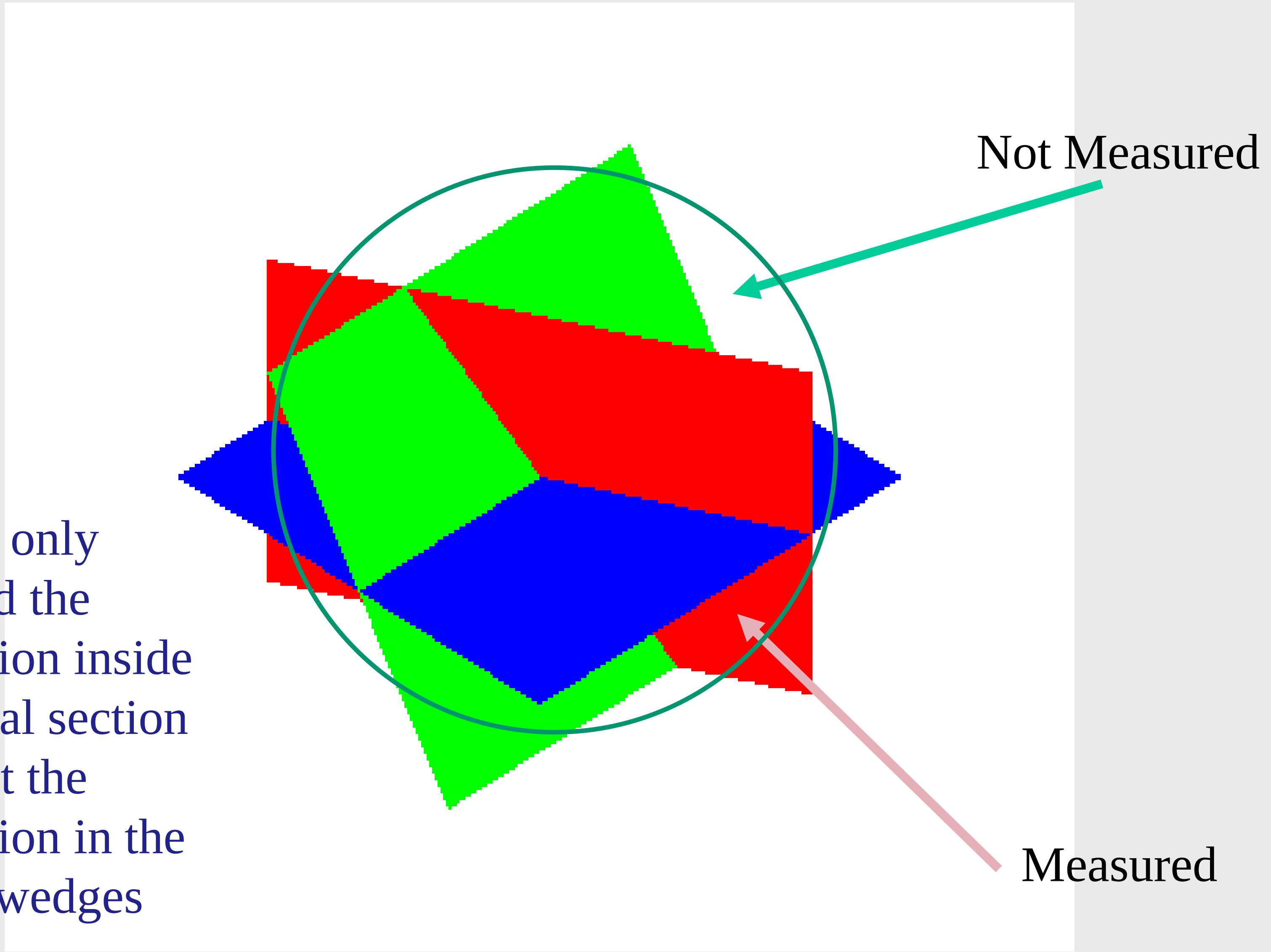






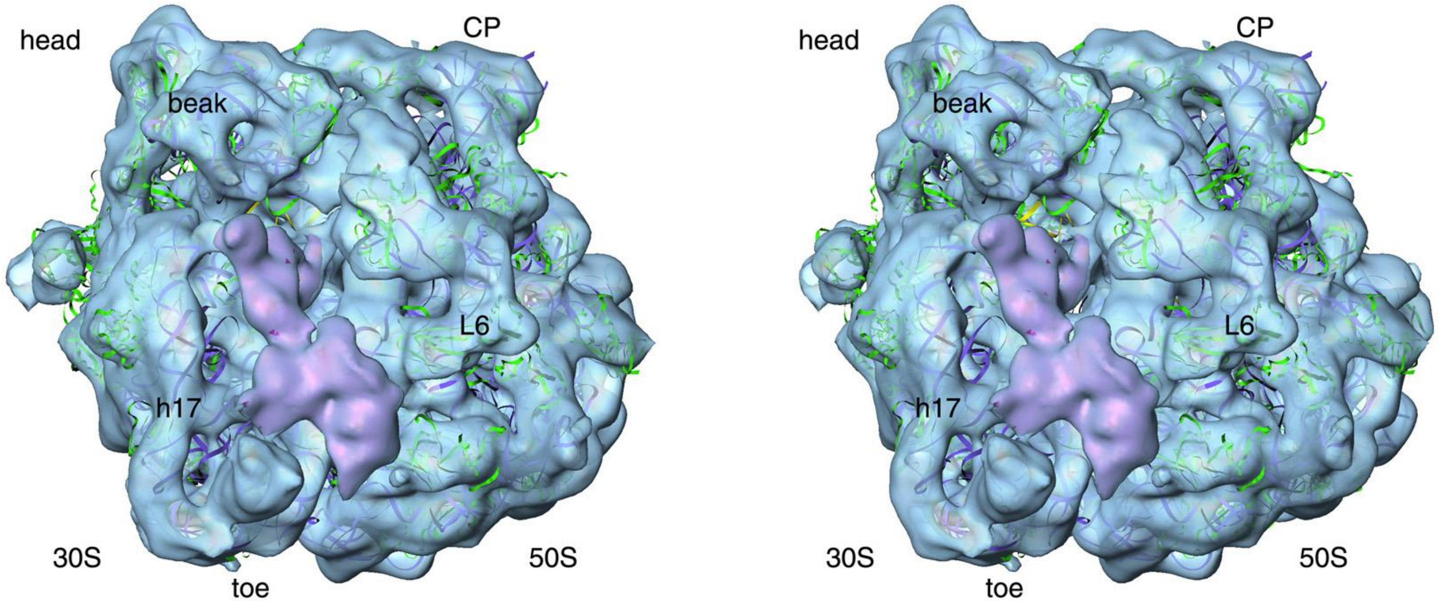
a
0°
360°
0°
c
e
360°
β = 80°
γ = 7.5 °
β = 85°
γ = 10 °
d
b

Intersecting Central Sections in 3D Fourier space

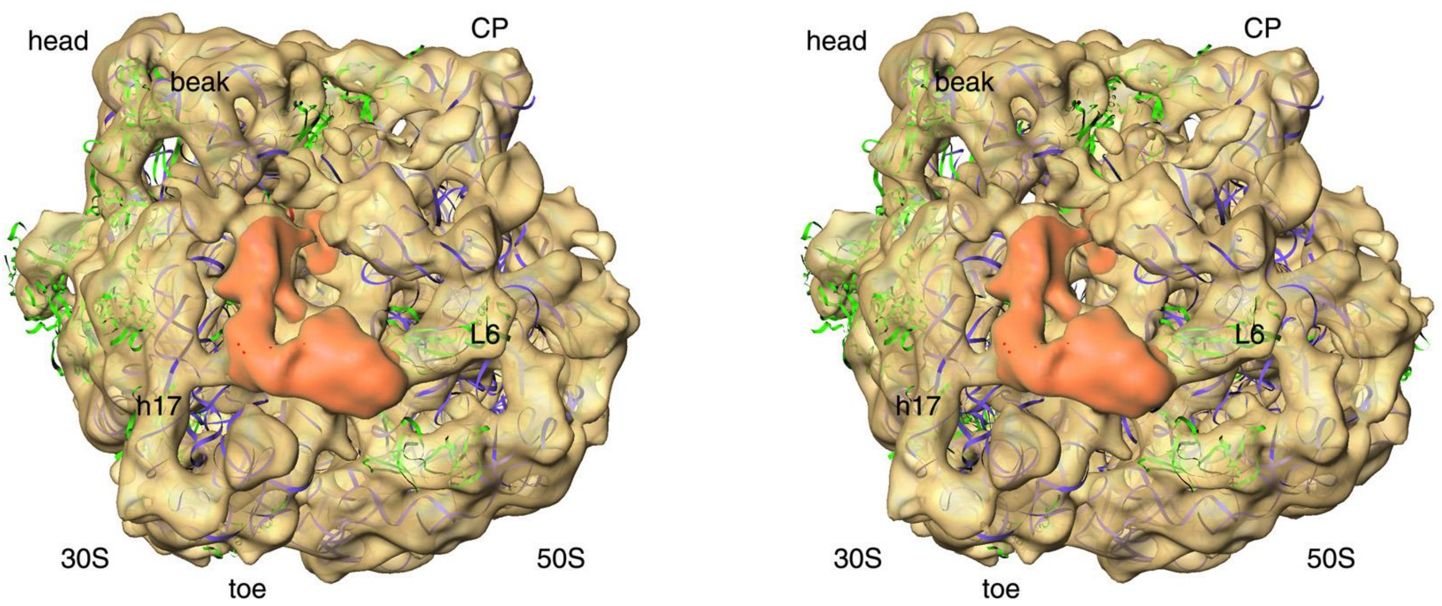


We have only measured the information inside the central section slabs, not the information in the missing wedges

4D Data Processing



4D cryo-EM!



RF3 Complex

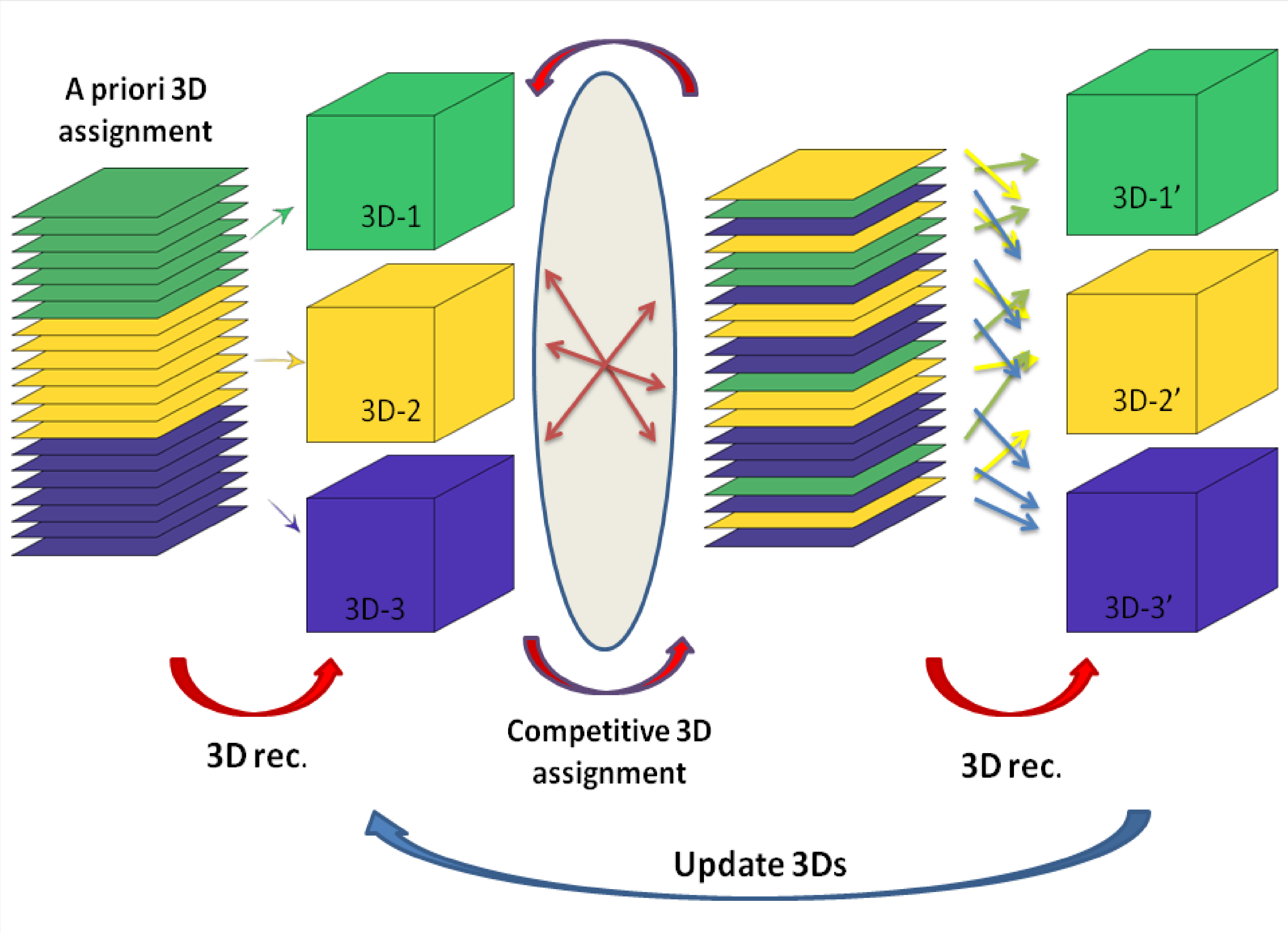
Type-1

vs.

Type-2

(Klaholz *et al*, Nature, 2004)

IMAGIC 4D:



A priori 3D assignment
3D-1
3D-2
3D-3
3D-1′
3D-2′
3D-3′
3D rec.
Competitive 3D assignment
3D rec.
Update 3Ds

A typical 4D refinement round)

```
Histogram of OLD 3D references

   0  0 |
 312  1 |************************************
 293  2 |**********************************
 188  3 |**********************
 587  4 |********************************************************************
   0  5 |

Meaning of columns: # in this bin, bin value


Histogram of NEW 3D references

   0  0 |
 278  1 |******************************
 304  2 |*********************************
 175  3 |*******************
 623  4 |********************************************************************
   0  5 |

Meaning of columns: # in this bin, bin value
```

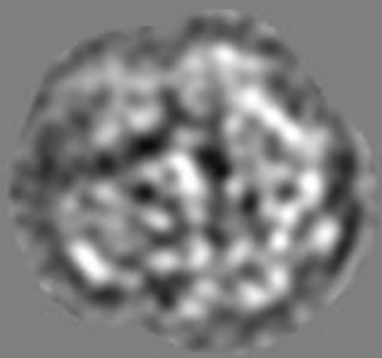
A typical 4D refinement round

```
Histogram of AR ERROR

1380    0    0   4.78E+01 |
1345   35   35   5.81E+01 |****
1020  360  325   6.33E+01 |**************************************
 641  739  379   6.85E+01 |********************************************
 436  944  205   7.37E+01 |************************
 341 1039   95   7.89E+01 |***********
 256 1124   85   8.41E+01 |**********
 206 1174   50   8.93E+01 |******
 163 1217   43   9.45E+01 |*****
 119 1261   44   9.96E+01 |*****
  89 1291   30   1.05E+02 |***
  70 1310   19   1.10E+02 |**
  59 1321   11   1.15E+02 |*
  42 1338   17   1.20E+02 |**
  32 1348   10   1.26E+02 |*
  25 1355    7   1.31E+02 |*
  19 1361    6   1.36E+02 |*
  17 1363    2   1.41E+02 |
  12 1368    5   1.46E+02 |*
   0 1380   12   1.57E+02 |*

Low values binned at  lower edge : 0
High values binned at higher edge: 12

Meaning of columns: # remaining, # accumulated, # in this bin, bin value
```



1/91

1/92

1/93

1/94

1/97

1/98

1/99

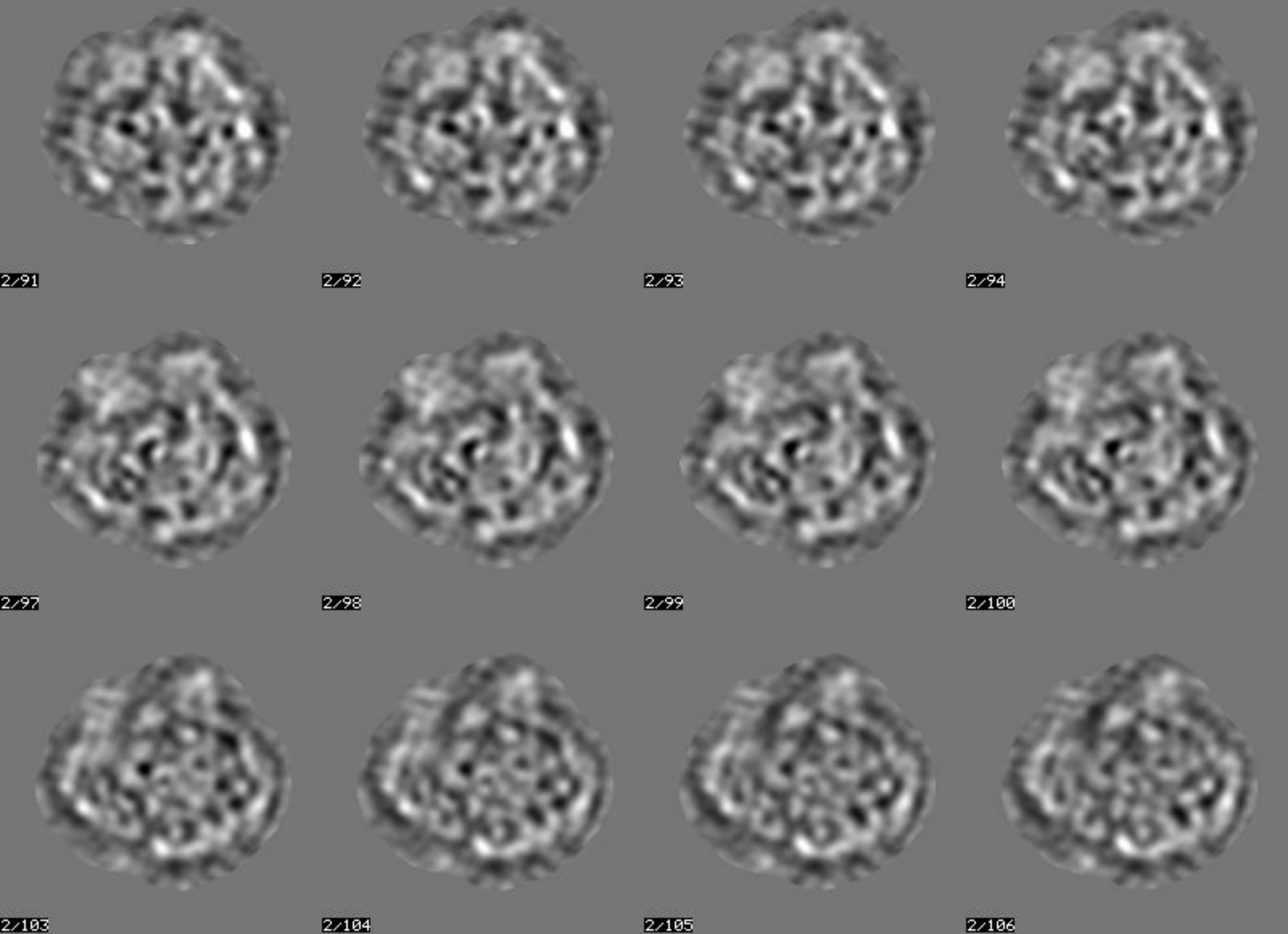
1/100

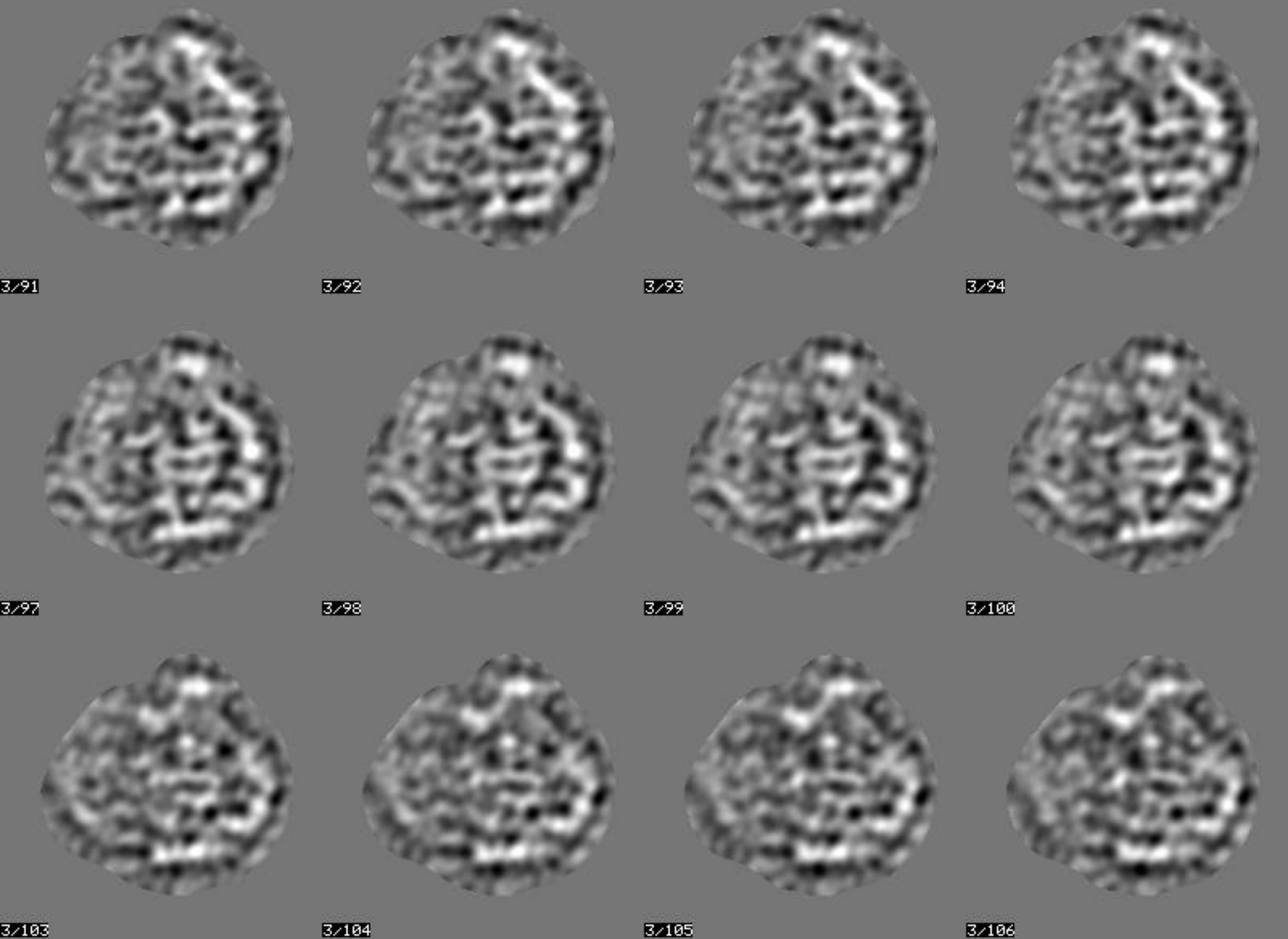
1/103

1/104

1/105

1/106


2/91
2/92
2/93
2/94
2/97
2/98
2/99
2/100
2/103
2/104
2/105
2/106


3/91
3/92
3/93
3/94
3/97
3/98
3/99
3/100
3/103
3/104
3/105
3/106

4/91

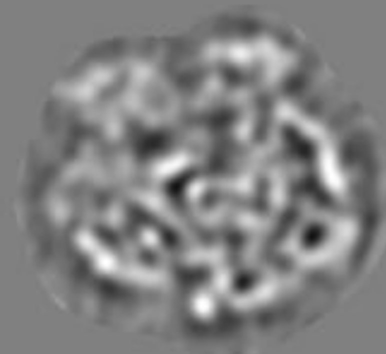
4/92

4/93

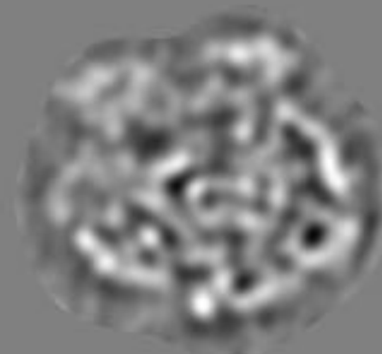
4/94

4/97

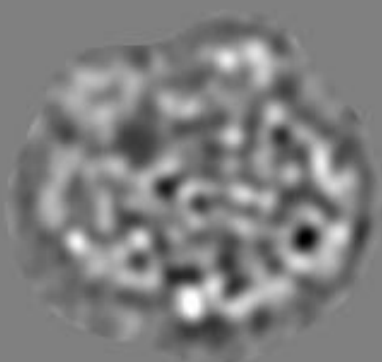
4/98



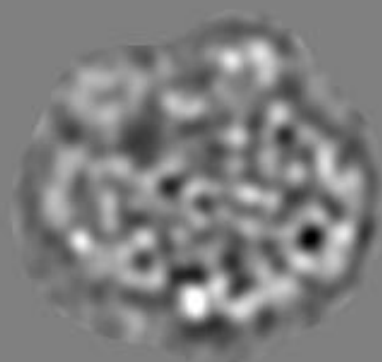
4/99



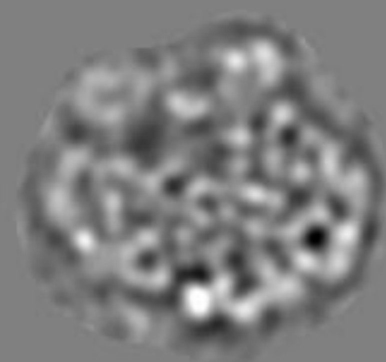
4/100



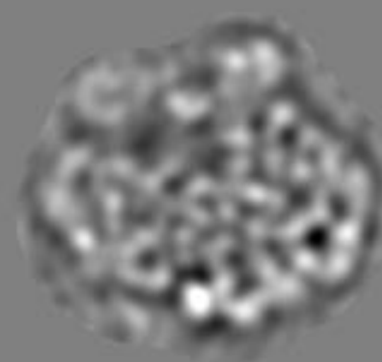
4/103



4/104

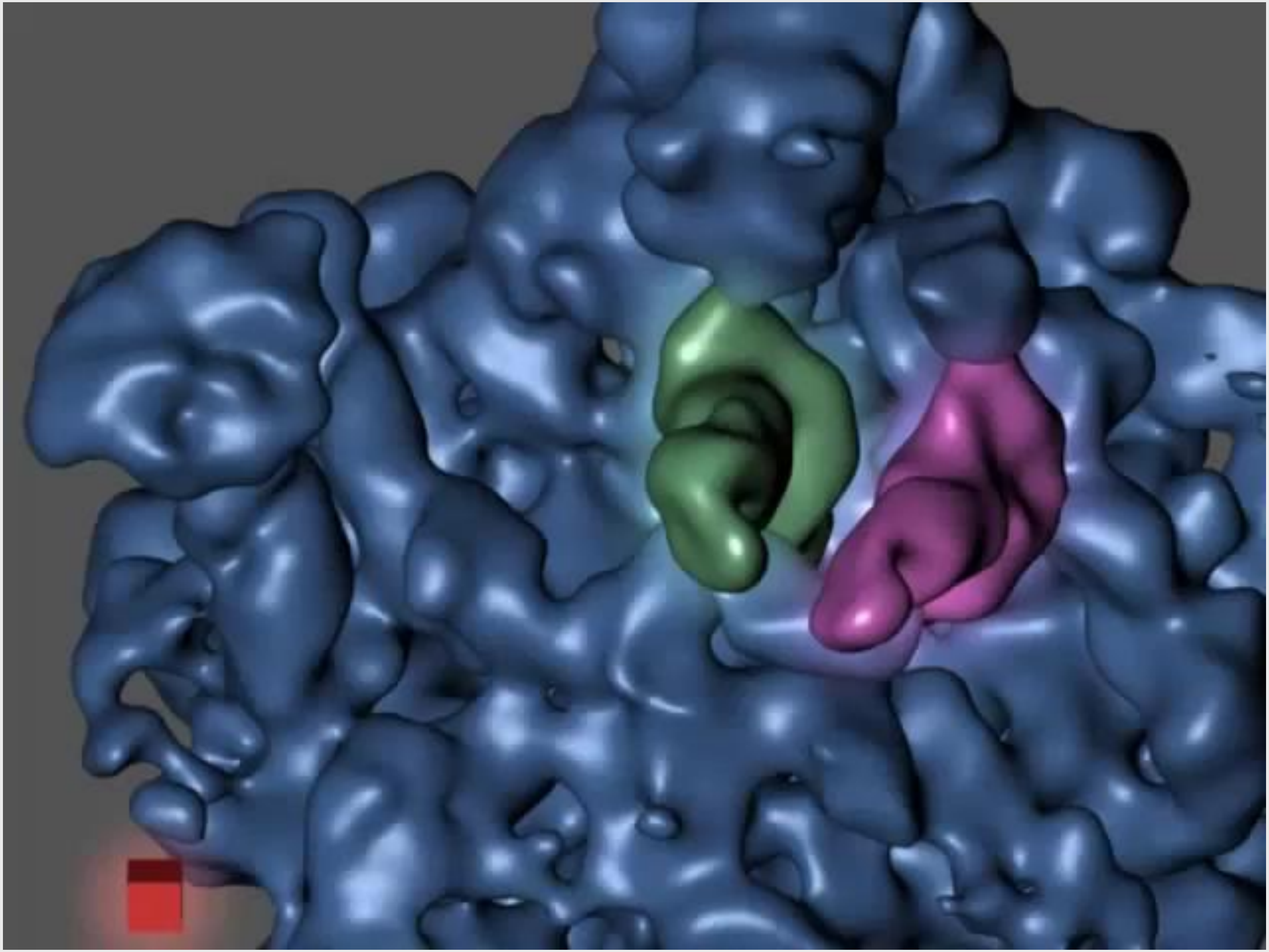


4/105

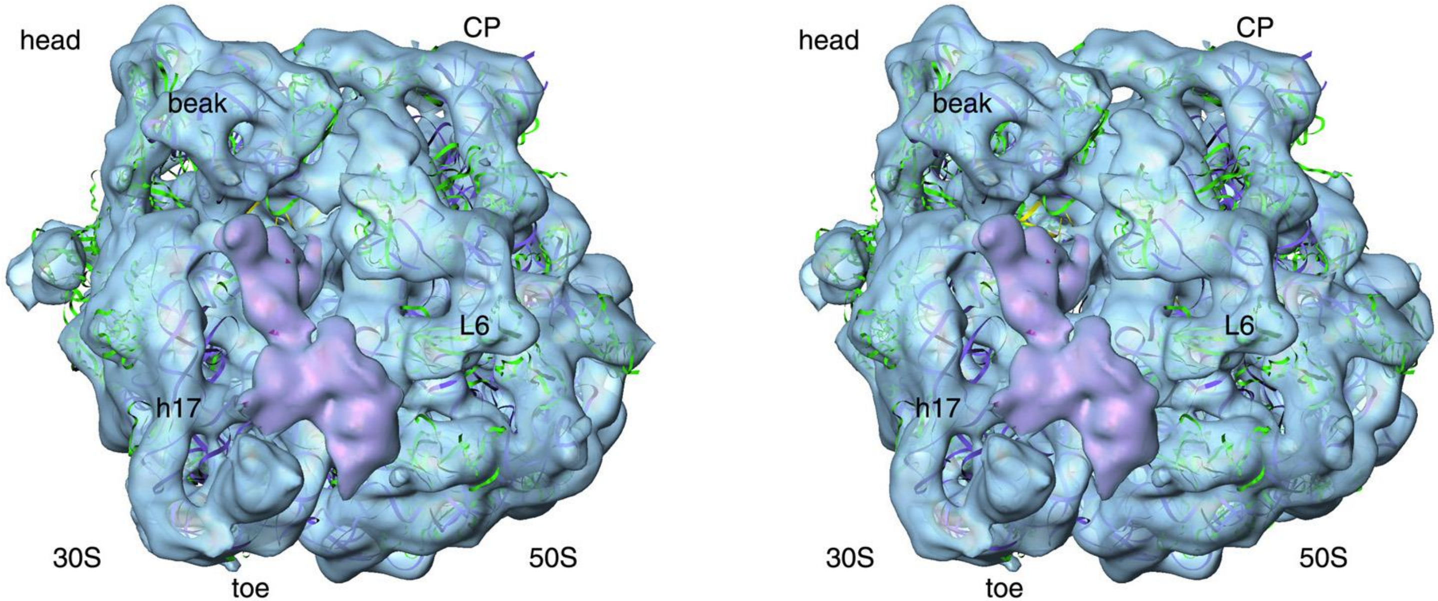


4/106

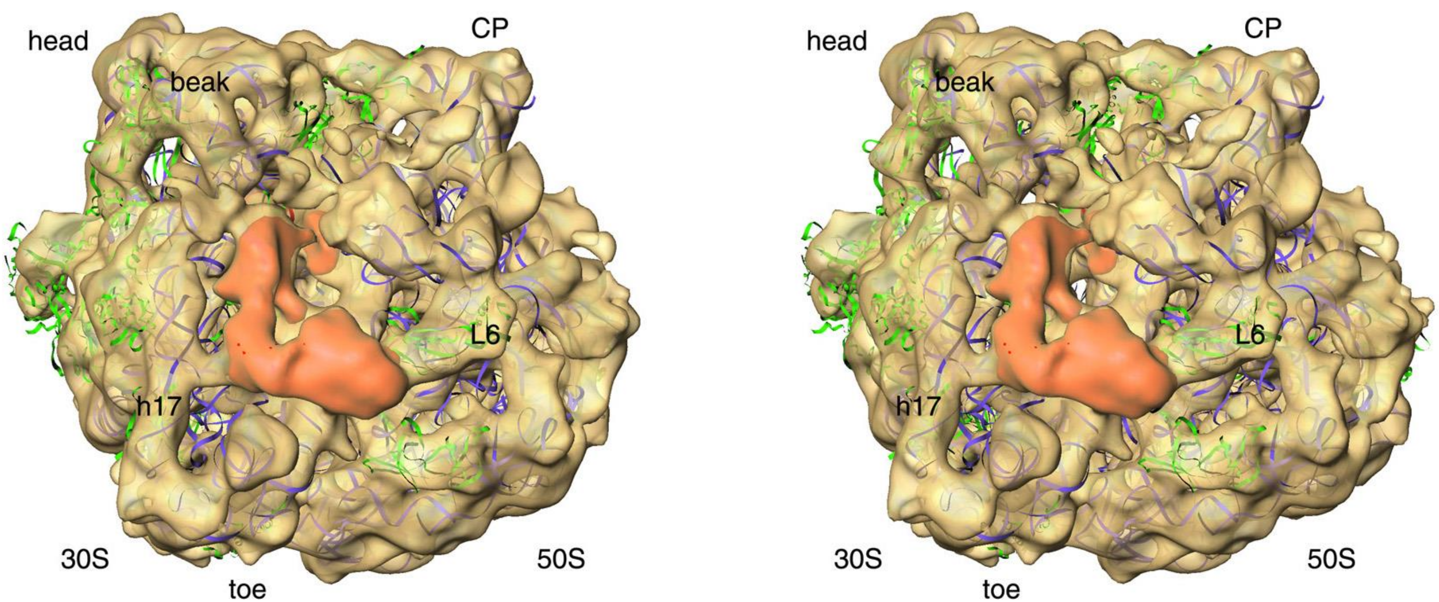
Stark *et al* Nature 2010



Focussed classification in 2D



4D cryo-EM!



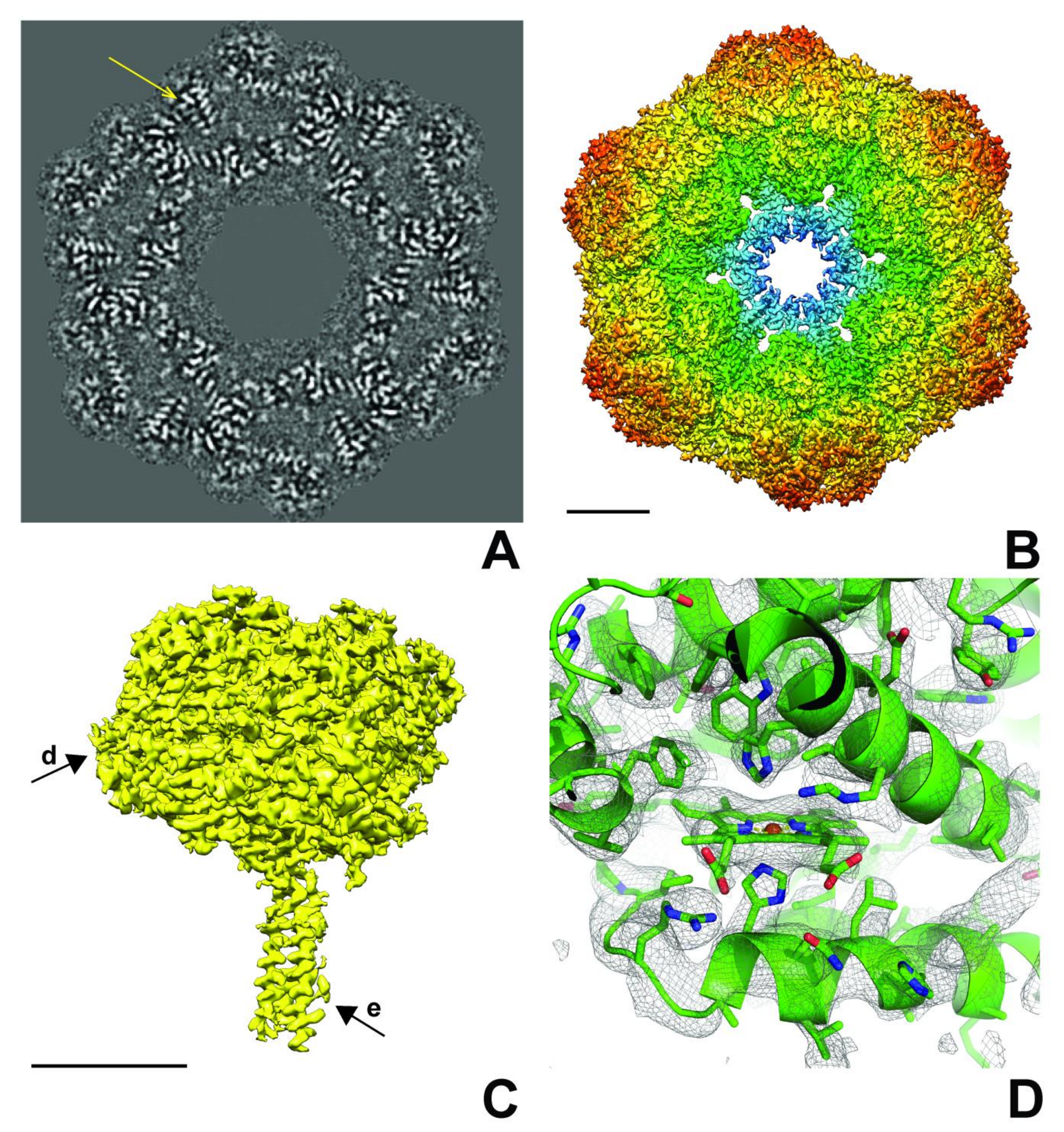
RF3 Complex

Type-1

vs.

Type-2

(Klaholz *et al*, Nature, 2004)



Focussed classification in 3D/4D: process each Globin fold independently


Pavel Afanasyev (Leiden/Maastricht)
Charlotte Linnemayr-Seer (Zurich)
Bart Alewijnse (NeCEN)
Michael Schatz (ImSc/Berlin)
Ralf Schmidt (ImSc/Berlin)
Sacha de Carlo (FEI/NeCEN)
Rishi Matadeen (NeCEN)
Rodrigo Portugal (LNNano)
and many others…