

2D Alignment as an *advanced* problem

Pawel A. Penczek

The University of Texas – Houston Medical School,
Department of Biochemistry



THE UNIVERSITY *of* TEXAS

HEALTH SCIENCE CENTER AT HOUSTON

MEDICAL SCHOOL

STABILITY AND REPRODUCIBILITY

AS STANDARDS FOR

OBJECTIVE 2D ALIGNMENT AND CLUSTERING

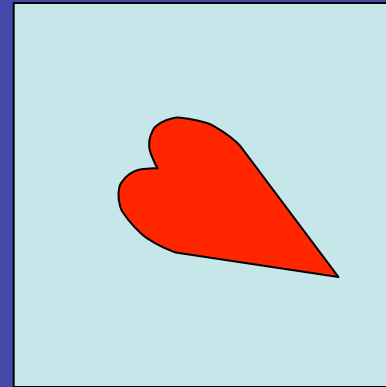
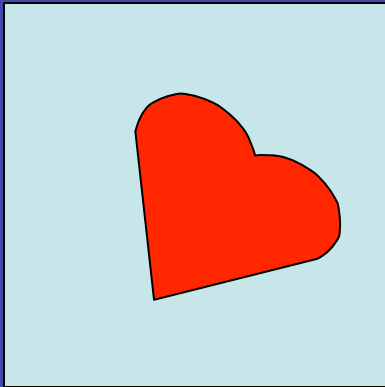
Pawel A. Penczek

**The University of Texas – Houston Medical School,
Department of Biochemistry**

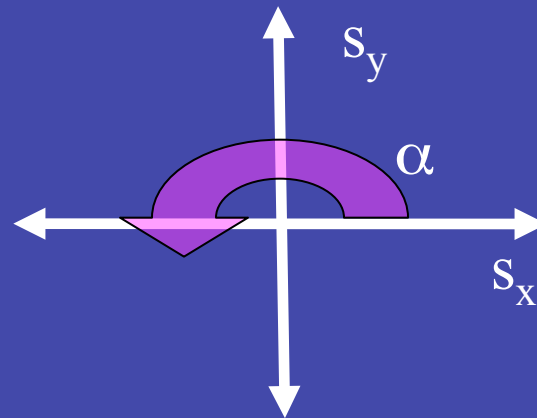


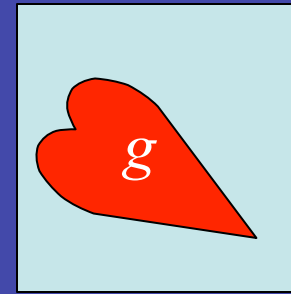
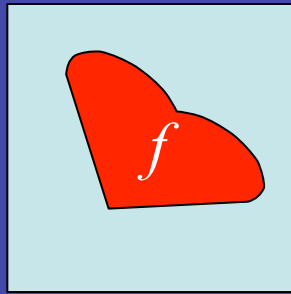
The alignment problem

Two 2D images:



Three degrees of freedom:





Two images are aligned if the least square discrepancy between them is minimized:

$$\int |f(\mathbf{x}; s_x, s_y, \alpha) - g(\mathbf{x})|^2 d\mathbf{x} \rightarrow \min$$

$$\int |f|^2 d\mathbf{x} + \int |g|^2 d\mathbf{x} - 2 \int f(\mathbf{x}; s_x, s_y, \alpha) g(\mathbf{x}) dx \rightarrow \min$$

$$\text{const} + \text{const} - c(\mathbf{x}; s_x, s_y, \alpha) \rightarrow \min$$

$$c(\mathbf{x}; s_x, s_y, \alpha) \rightarrow \max$$

Two images are aligned if the least square discrepancy between them is minimized:

$$c(s_x, s_y, \alpha) \rightarrow \max$$

Maximum of the cross-correlation function

Valid only if the noise is additive and white (its power spectrum is straight horizontal line)!

CRYO-EM IMAGE FORMATION MODEL

$$G = CTF \cdot F + N$$

✱ *N* - background noise

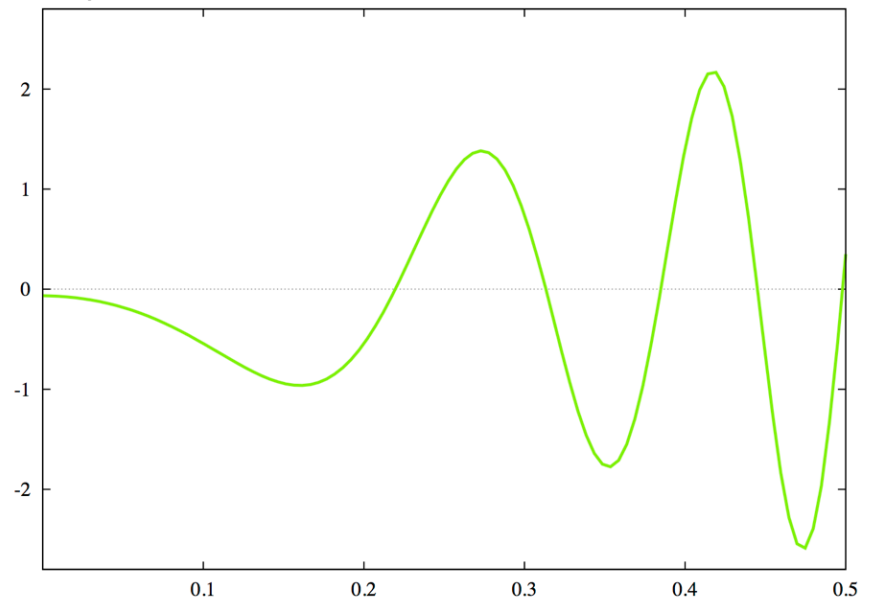
MATCHED FILTER

DETECTION OF A TEMPLATE IN A NOISY FIELD

$$G = CTF \cdot F + N_{background}$$

✱ *ccf* - cross-correlation function

$$ccf_{ice} \approx \frac{CTF}{P_N} (GF^*)$$



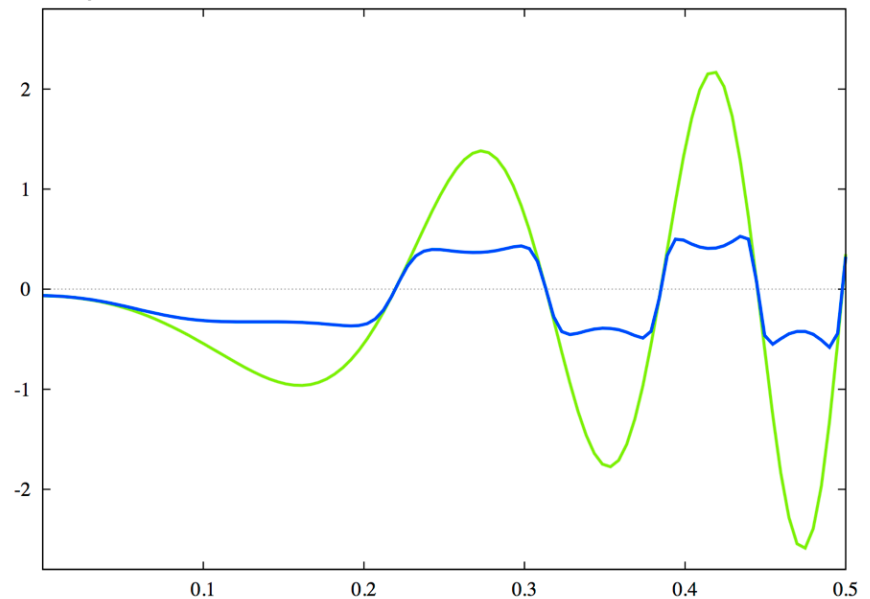
MATCHED FILTER

DETECTION OF A TEMPLATE IN A NOISY FIELD

$$G = CTF \cdot F + N_{background} + CTF \cdot M_{carbon}$$

✻ *ccf* - cross-correlation function

$$ccf_{carbon} \approx \frac{CTF}{P_N + P_M} (GF^*)$$



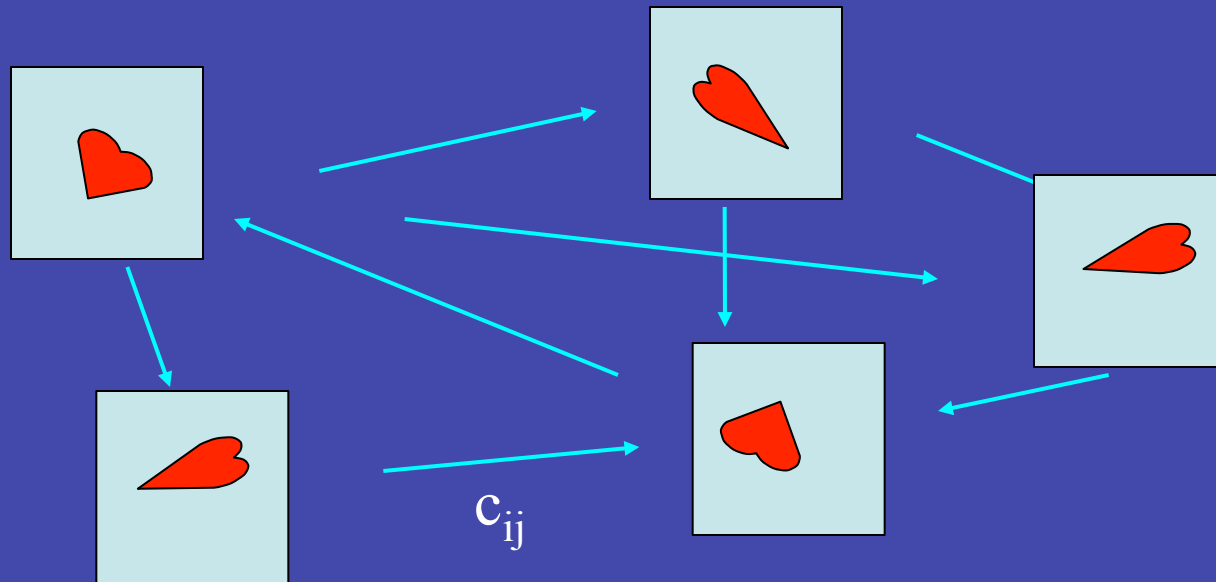
Two images are aligned if the least square discrepancy between them is minimized:

$$c(s_x, s_y, \alpha) \rightarrow \max$$

Maximum of the cross-correlation function

How to define the best alignment for n objects?

Alignment of n objects



The distances between all pairs of images have to be minimized simultaneously.

$$\sum_{k=1}^{n-1} \sum_{l=k+1}^n \int \left| f_k(\mathbf{x}; s_x^k, s_y^k, \alpha^k) - f_l(\mathbf{x}; s_x^l, s_y^l, \alpha^l) \right|^2 d\mathbf{x} \rightarrow \min$$

Sum of distances between each image and sums
(average) of all remaining images.

$$\sum_{k=1}^{n-1} \sum_{l=k+1}^n \int \left| f_k(\mathbf{x}; s_x^k, s_y^k, \alpha^k) - f_l(\mathbf{x}; s_x^l, s_y^l, \alpha^l) \right|^2 d\mathbf{x} \rightarrow \min$$

$$\int \left| \sum_{l=1}^n f_l(\mathbf{x}; s_x^l, s_y^l, \alpha^l) \right|^2 d\mathbf{x} \rightarrow \max$$

Squared norm of the sum of all images!

The three alignment criteria are equivalent:

1. Sum of all pairs-distances between images $\rightarrow \min$
2. Sum of distances between each image and sums (average) of all remaining images (variance) $\rightarrow \min$
3. Squared norm of the sum of all images $\rightarrow \max$

Sum of distances between all pairs of images.

$$\sum_{k=1}^n \int \left| f_k(\mathbf{x}; s_x^k, s_y^k, \alpha^k) - \langle f \rangle_k \right|^2 d\mathbf{x} \rightarrow \min,$$

where

$$\langle f \rangle_k = \frac{1}{n-1} \sum_{\substack{l=1 \\ l \neq k}}^n f_l(\mathbf{x}; s_x^l, s_y^l, \alpha^l)$$

Sum of distances between each image and sums (average) of all remaining images.

Suggests an alignment algorithm:

orientation of each image is refined against the current average of remaining images.

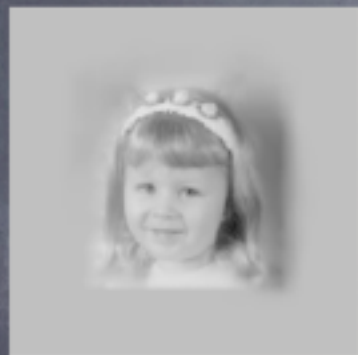
There is no algorithm that would guarantee location of the global minimum (best possible alignment of a set of n images).

Model Bias ?

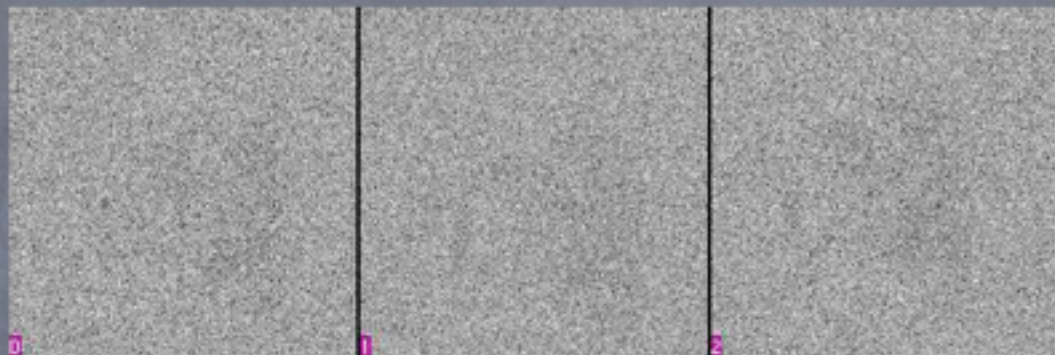
Next few slides are courtesy of Steve Ludtke

Model Bias

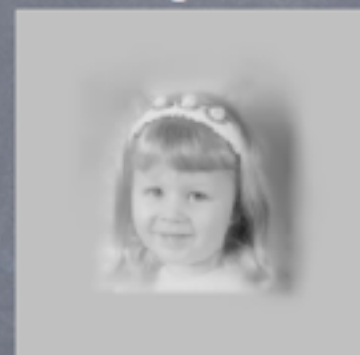
Base



Noisy (~10% contrast)



Align to



25

100

250

1000

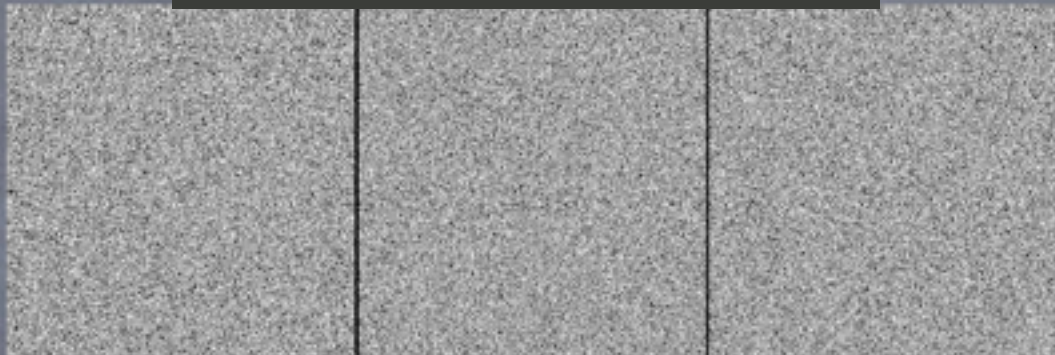
2000

Model Bias

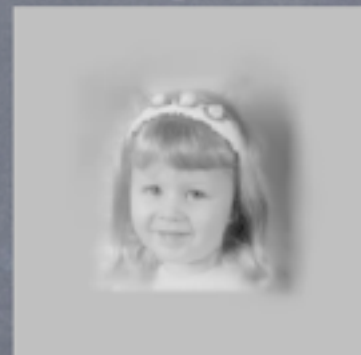
Base



Noisy versions of the base



Align to



25

100

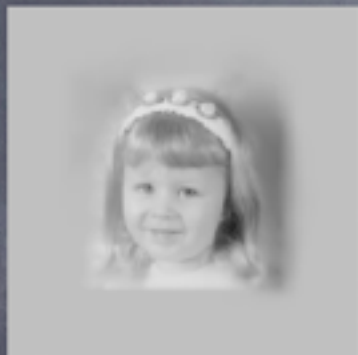
250

1000

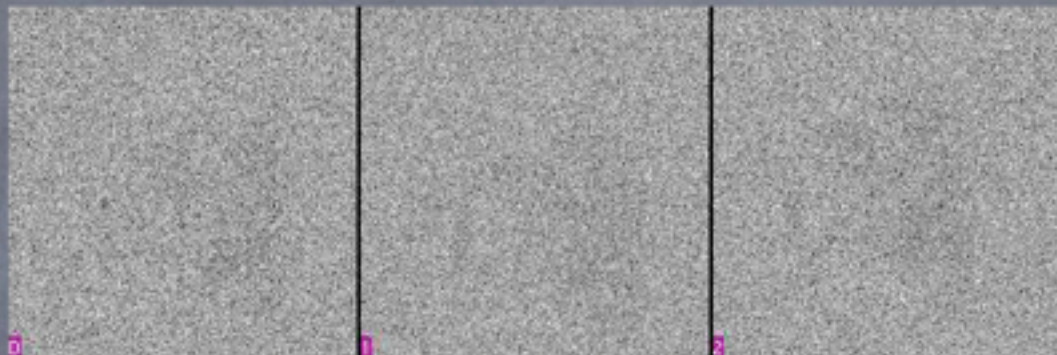
2000

Model Bias

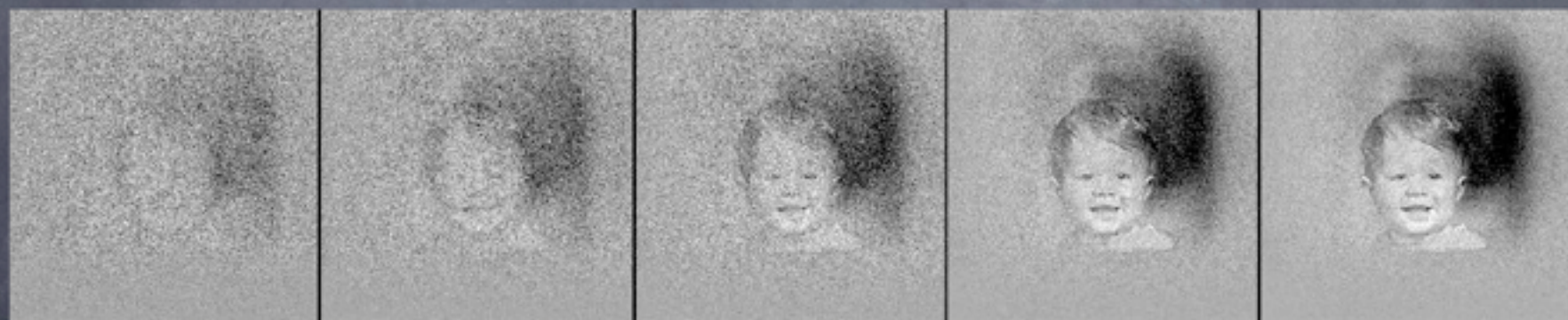
Base



Noisy (~10% contrast)



Align to



25

100

250

1000

2000

Reference bias in alignment of n images

42

P. Penczek, J. Frank / 3D reconstruction of single particles embedded in ice / 1992

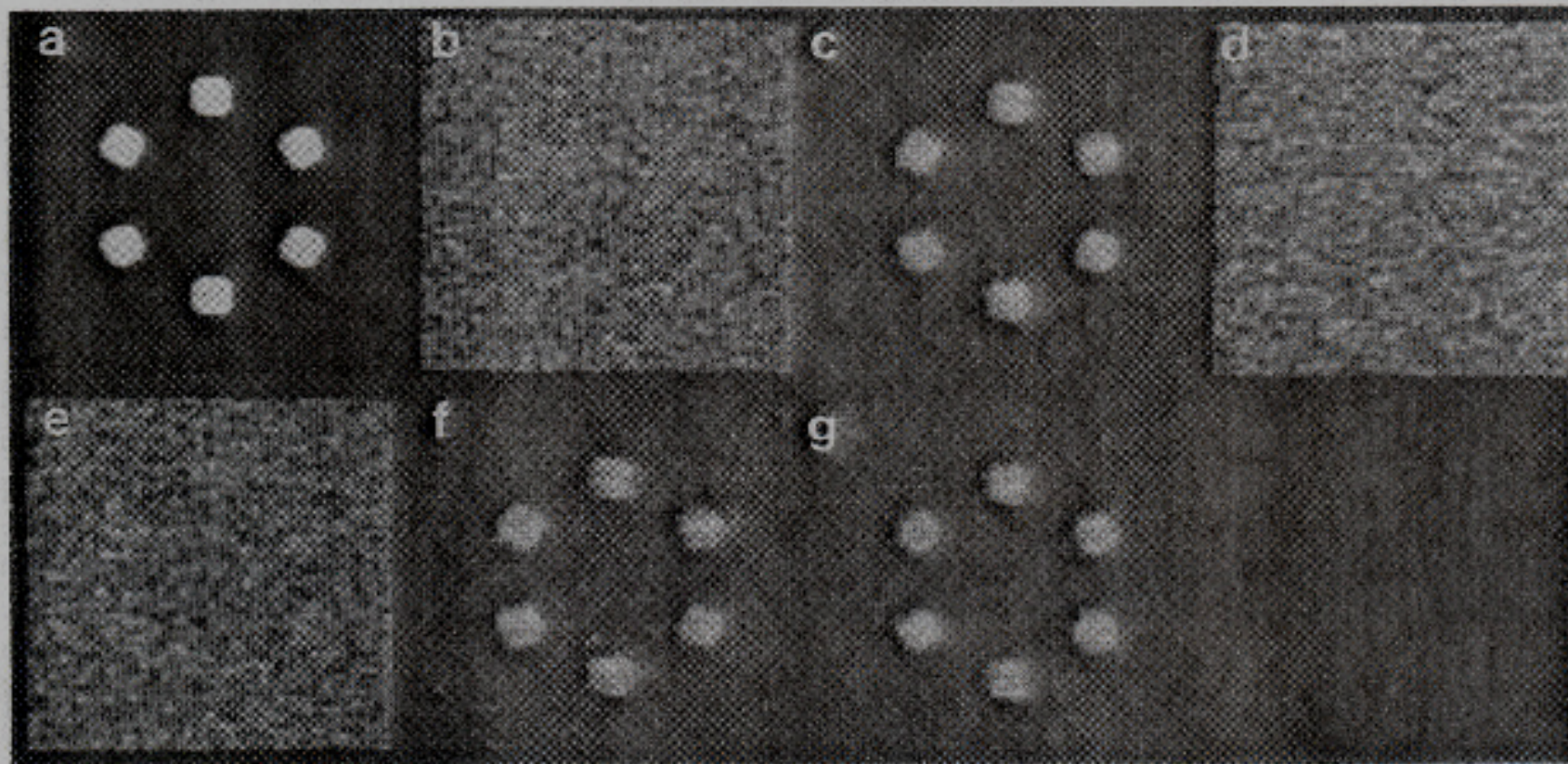
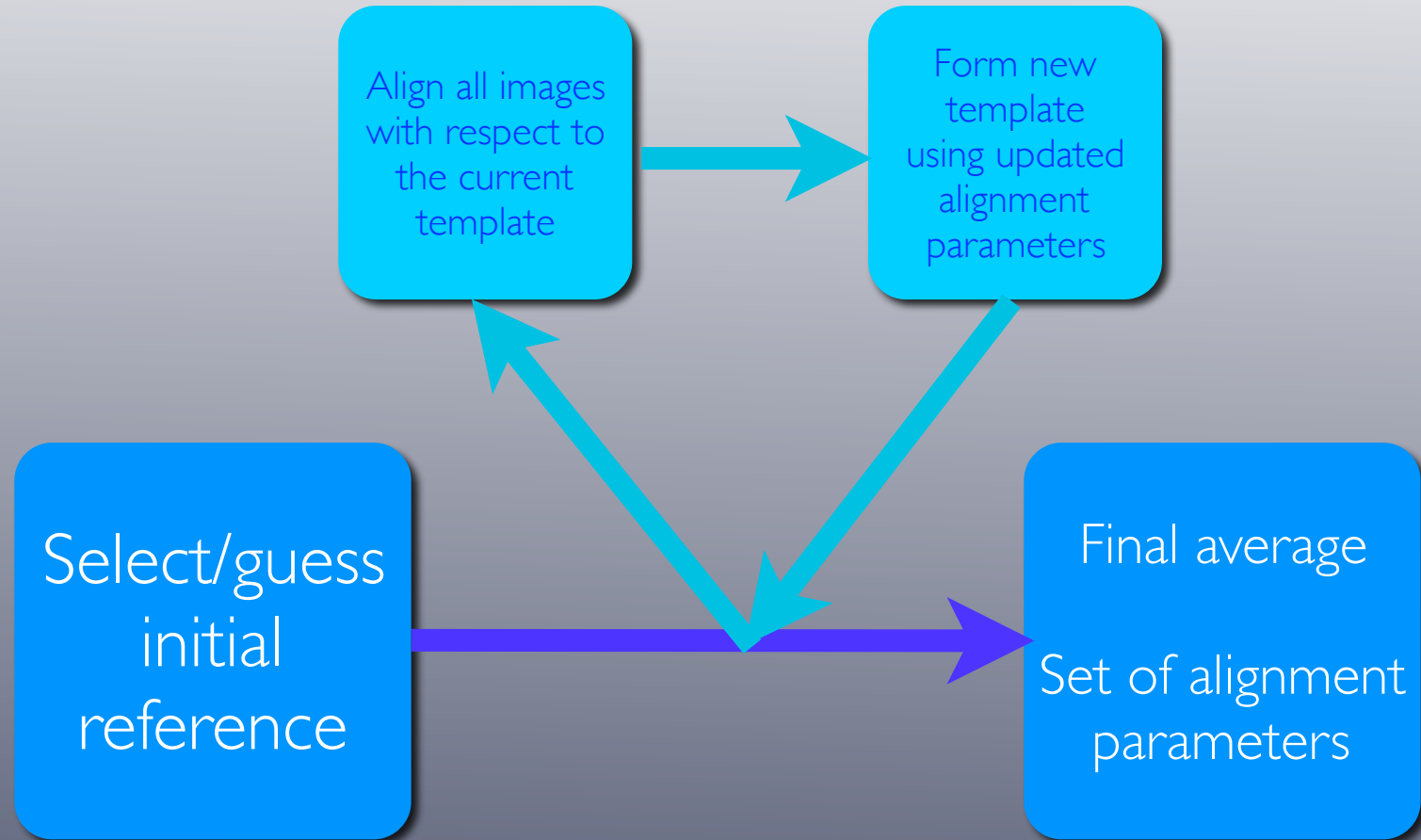


Fig. 2. Test of the reference-free alignment algorithm using model data. (a) Reference image used to test the standard alignment procedure. (b) Example of one of 128 64×64 test images containing Gaussian noise with the average equal to zero and variance equal to one. (c) Average of the series of 128 noise images aligned by the standard procedure. (d) Average of the series of 129 images (including the reference of the first test) after using the "reference-free" alignment program. (e) Modified version of the random image (b): the hexagonal reference image is added with a weight such that the resulting SNR is 0.75. (f) Average of the series of 128 modified noise images aligned by the standard procedure. (g) Average of the series of 129 modified images (including the reference of the first test) after using the "reference-free" alignment program.

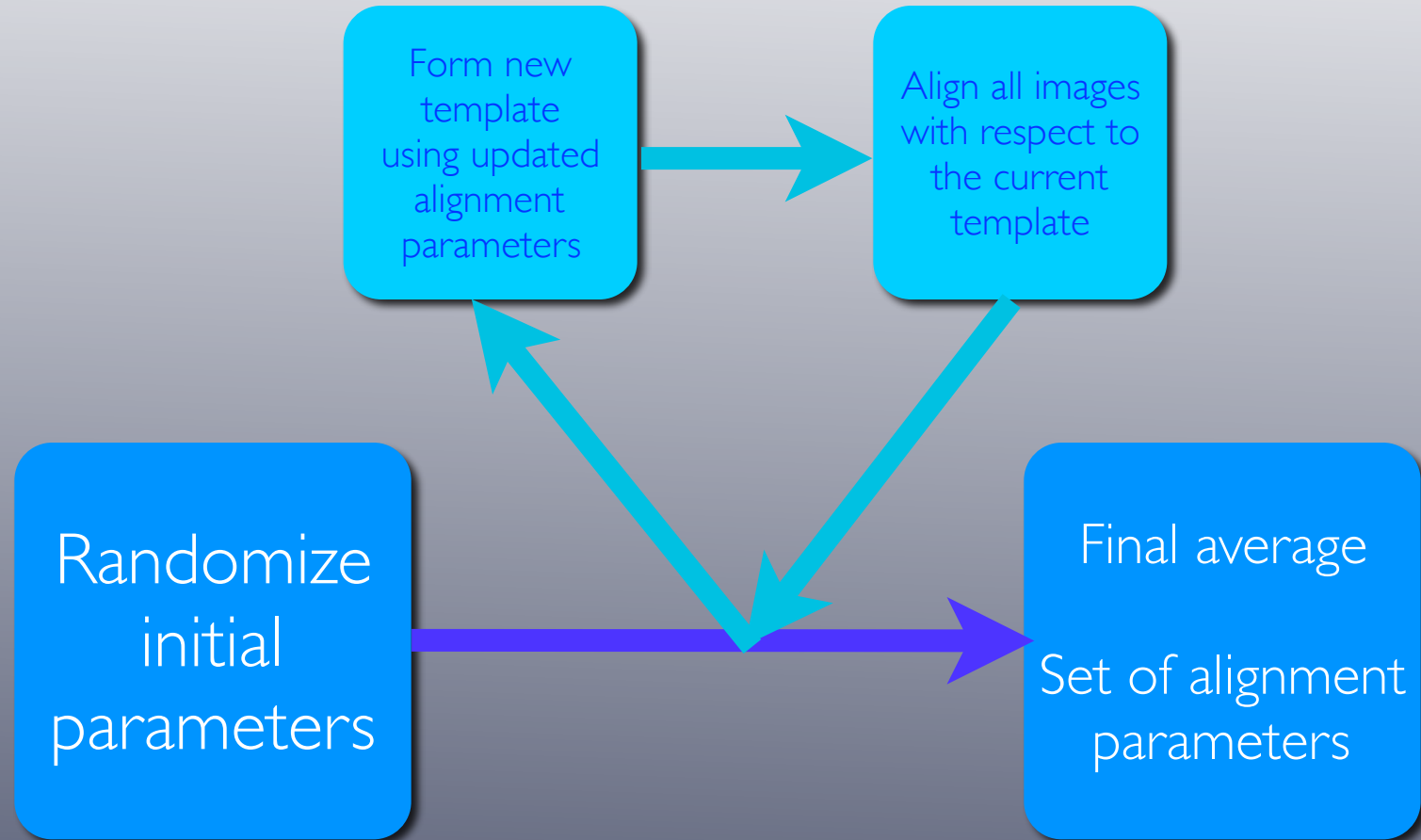
ALIGNMENT SCHEME

REFERENCE BASED



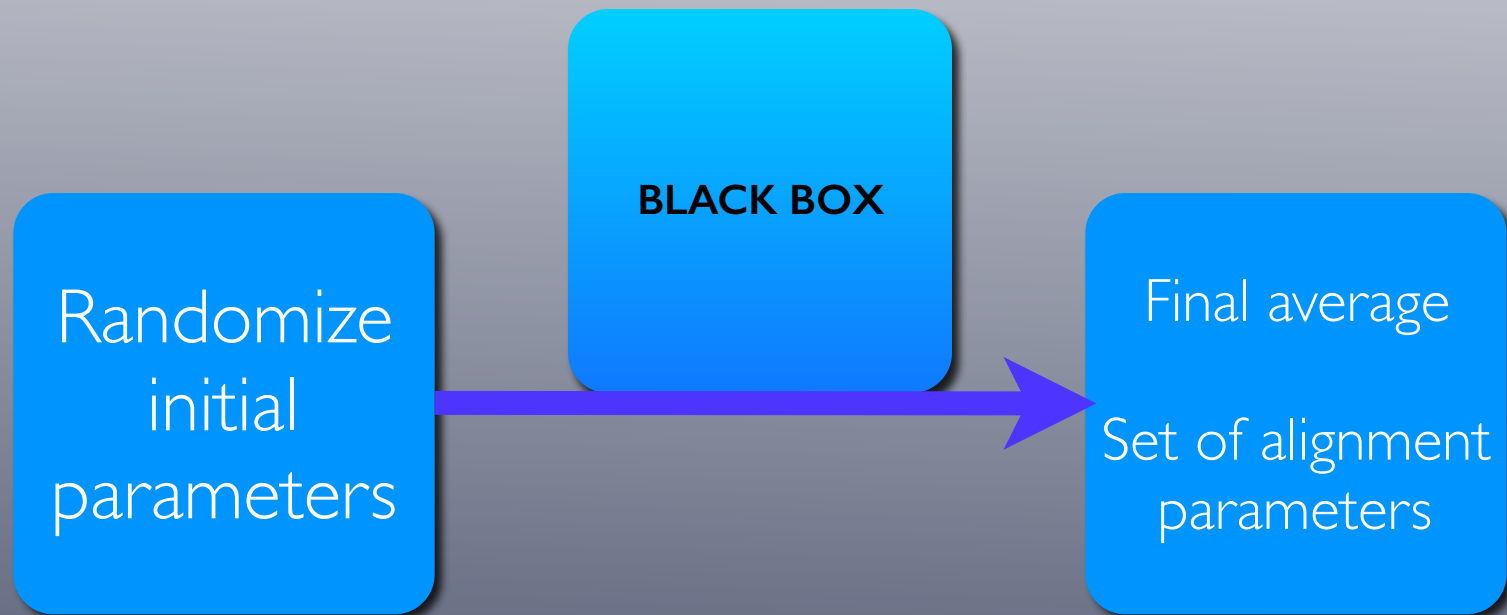
ALIGNMENT SCHEME

REFERENCE-FREE



ALIGNMENT SCHEME

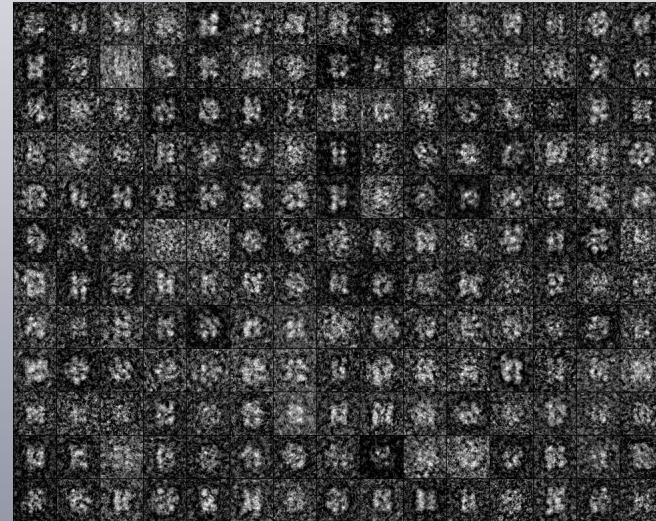
REFERENCE-FREE PROPER



2D MULTI-REFERENCE ALIGNMENT (MRA)

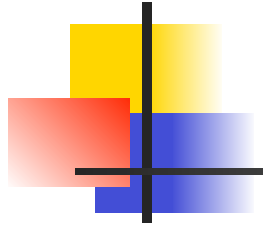
MRA is equivalent to K -means clustering, with the distance between images defined as a maximum similarity over the permissible range of image rotations and translations.

n images



K averages (clusters)

K-Means

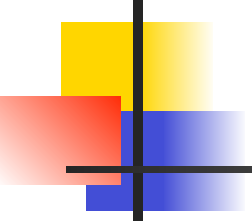


The K -means method is by far the most popular clustering algorithm used in scientific and industrial application.

K -means is both very simple and very fast, which makes it appealing in practice.

K -means begins with an arbitrary clustering based on K centers, and then repeatedly makes local improvements until the clustering stabilizes.

K-Means



Algorithm: K-means

Input: k number of clusters
 t number of iterations
 data the data
Output: C a set of k clusters

$cent$ = arbitrarily select k objects as initial centers

While(any d changed assignment) do

 for each d in $data$ do

 assign label x to d such that $\text{dist}(d, cent[x])$ is minimized;

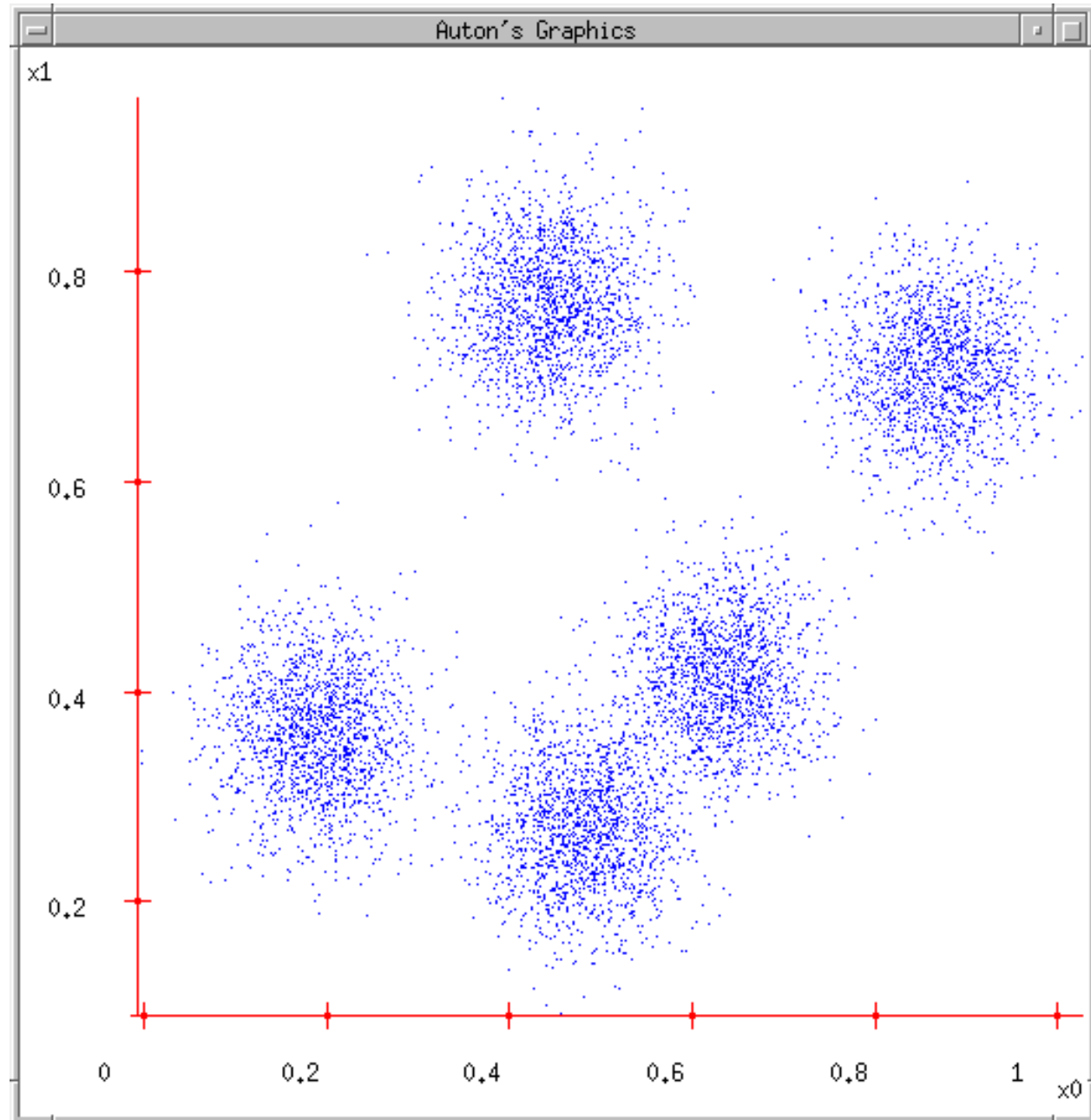
 for $x = 1$ to k do

$cent[x]$ = average value of all data with label x ;

Enddo

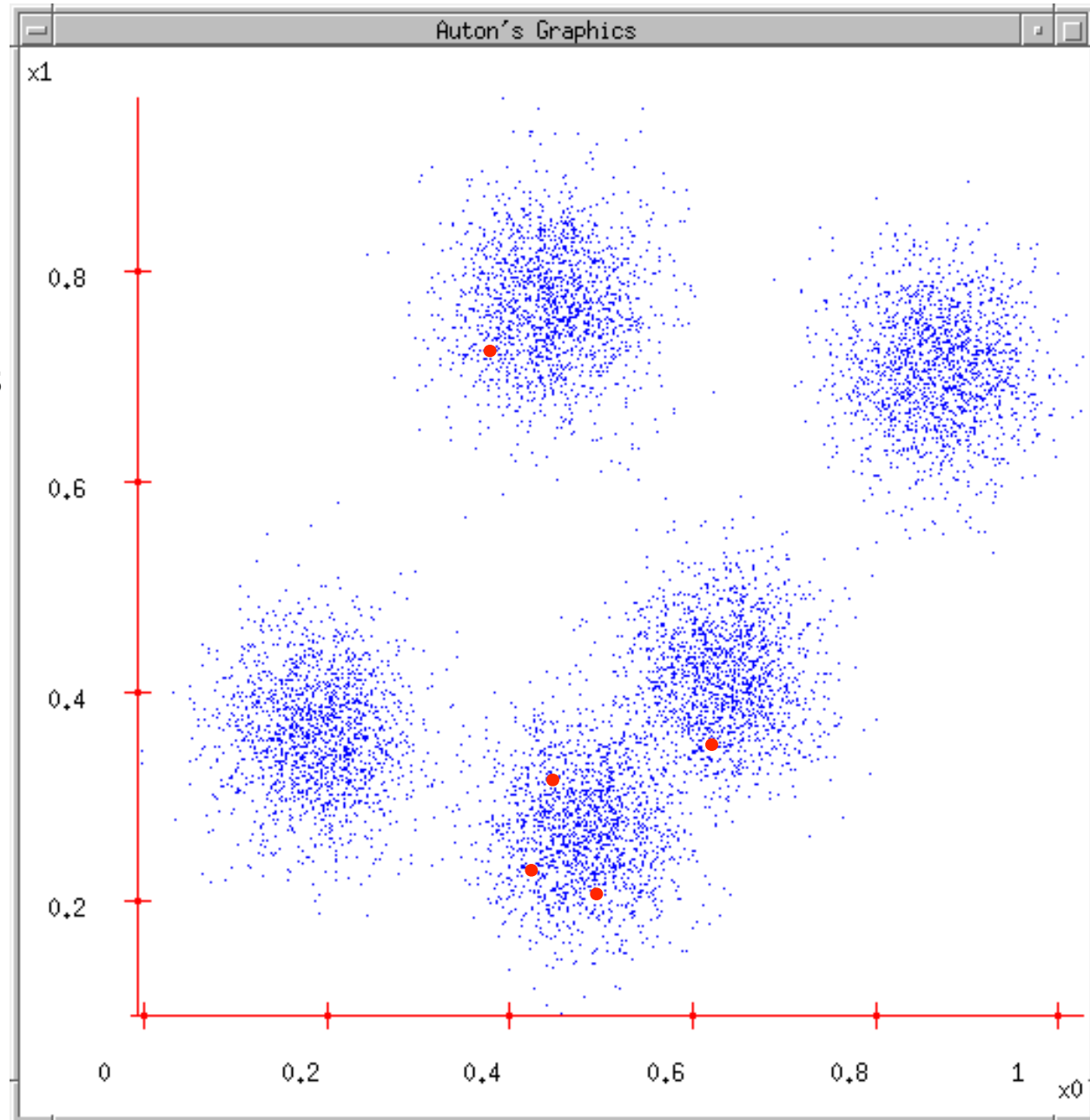
K-means

1. Ask user how many clusters they'd like.
(e.g. $K=5$)



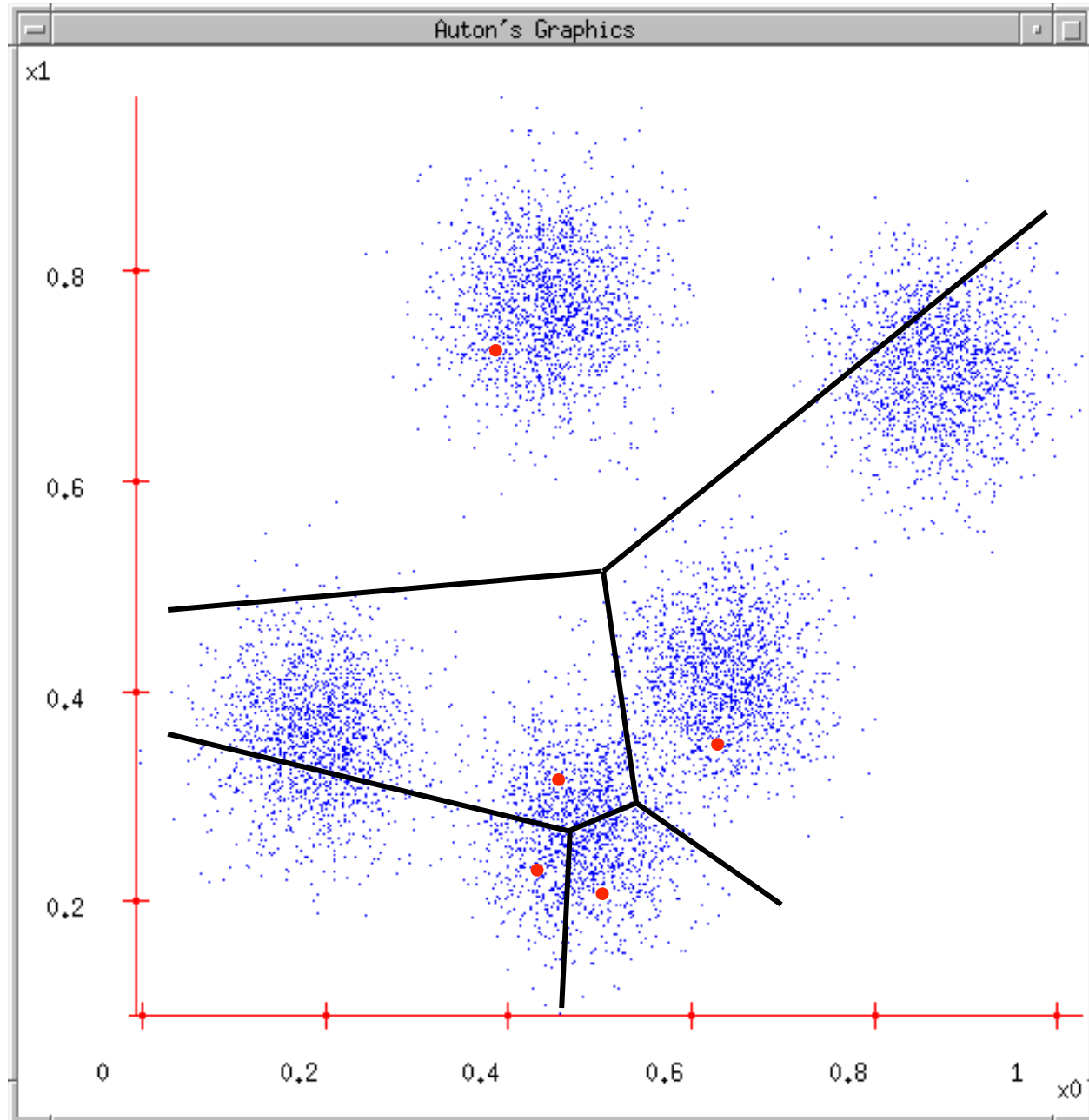
K-means

1. Ask user how many clusters they'd like.
(e.g. $K=5$)
2. Randomly guess K cluster Center locations



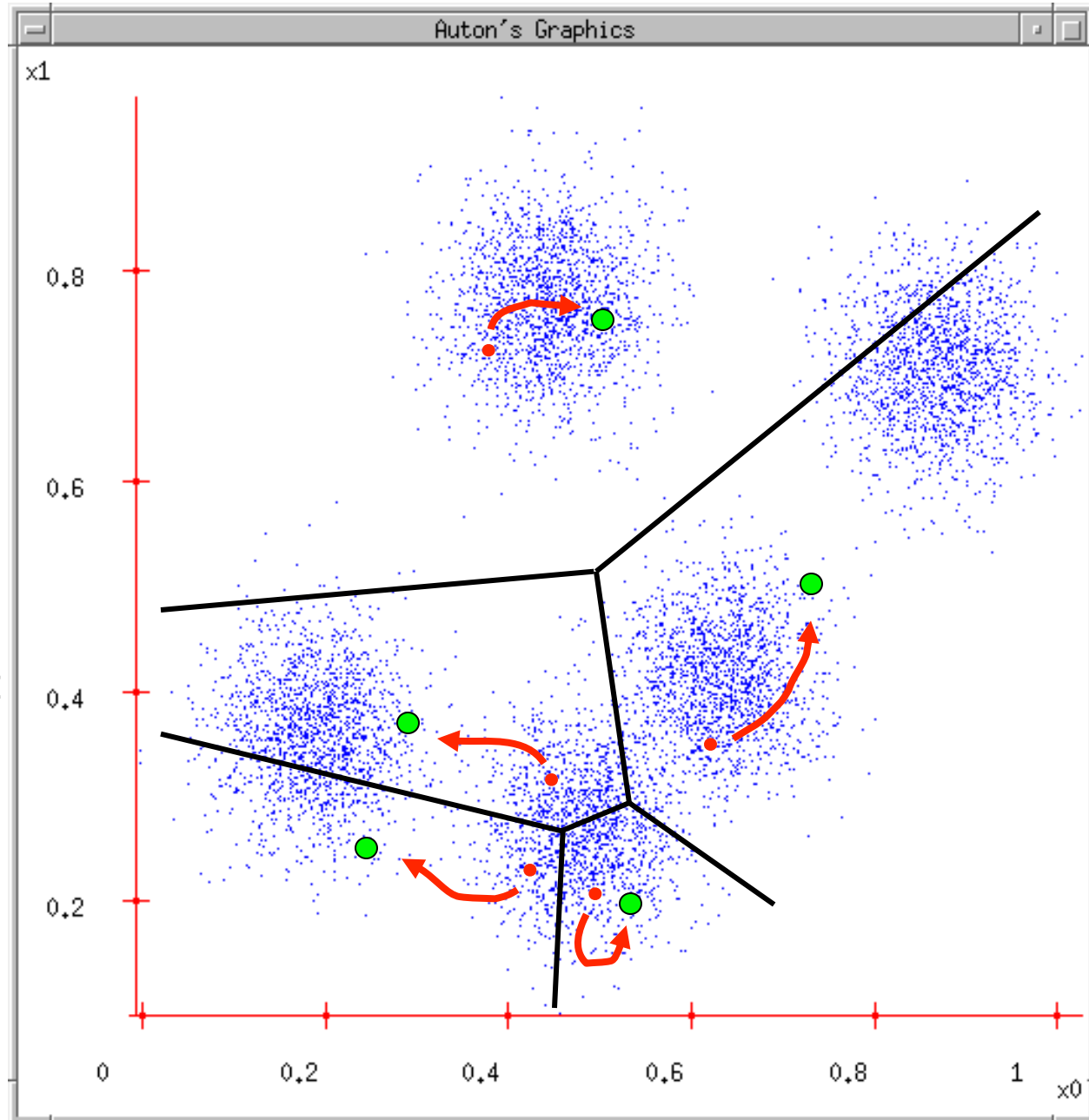
K-means

1. Ask user how many clusters they'd like.
(e.g. $K=5$)
2. Randomly guess K cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



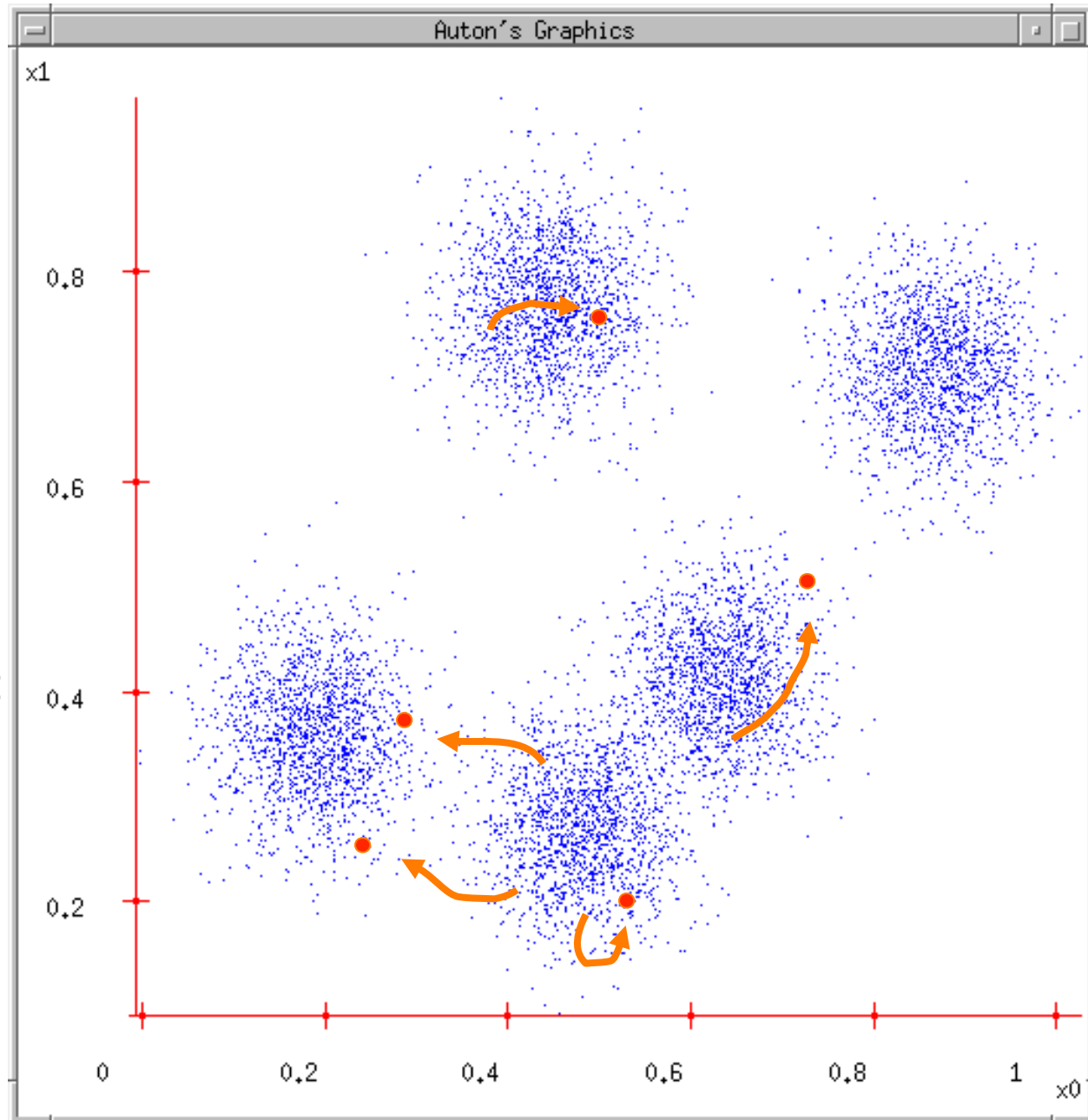
K-means

1. Ask user how many clusters they'd like.
(e.g. $K=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns

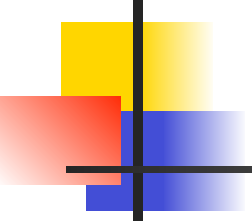


K-means

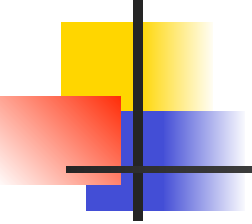
1. Ask user how many clusters they'd like.
(e.g. $K=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!



K-means properties

- 
-
- + Very simple algorithm
 - + Works very well if groups are well separated and number of groups K was guessed correctly
 - + $O(KNt)$ time complexity
 - + Guaranteed to converge in a finite number of steps
 - + In the SSE version, optimizes well-defined and intuitive notion of “natural grouping” (i.e., within-group variance)

K -means properties

- 
-
- Circular cluster shape only
 - Not guaranteed to converge to a global minimum
 - Finding global minimum not feasible in practice
 - Outliers can have very negative impact
 - If K not guessed correctly and/or groups are not well separated (i.e., almost always), the result dramatically depends on initialization.

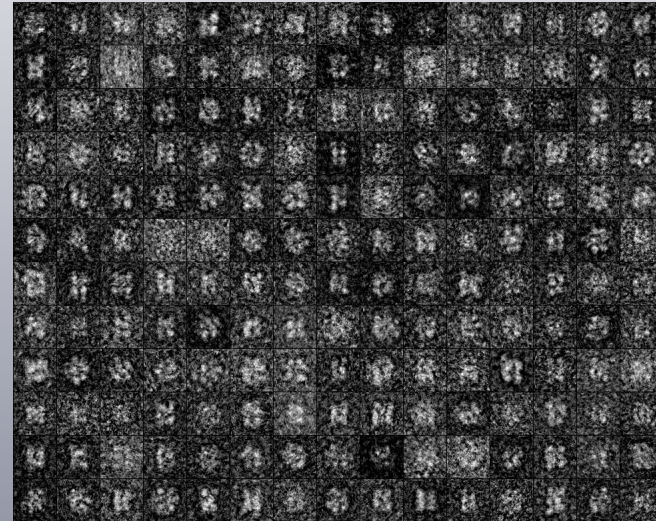
2D MULTI-REFERENCE ALIGNMENT (MRA)

MRA is equivalent to K -means clustering, with the distance between images defined as a maximum similarity over the permissible range of image rotations and translations.

K -means results depend on the solution to another nontrivial problem: the alignment of a set of 2D images.

Because neither of these two problems can be easily solved, the difficulty is compounded.

n images



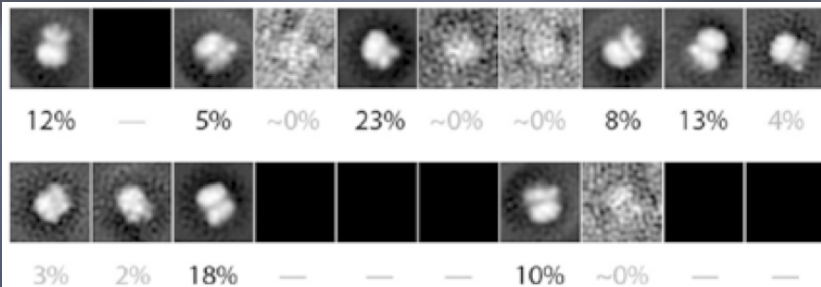
K averages (clusters)

K-MEANS CLUSTERING

KNOWN PROPERTIES:

- Very fast convergence guaranteed in a finite number of steps
- Converges only to a local minimum
- Unclear how to determine the appropriate number of classes (K)
- All images must be assigned to an average
- The solution (final averages) depends on the initial set of averages, and will change if clustering is repeated using different initial averages
- In EM, when alignment is added, classes tend to collapse

K-means group assignments
minimum distance to a template within a row












$EQK^{(EQUAL\ GROUP\ SIZE)}$ -MEANS CLUSTERING

Assign n images to K classes
such that each class contains

$$\frac{n}{K} \text{ images}$$

EQK -means group assignments
minimum distance to all templates, maximum number per group=3

			...	
	d_{11}^2	d_{12}^2	...	d_{1K}^2
	d_{21}^2	d_{22}^2	...	d_{2K}^2
	d_{31}^2	d_{32}^2	...	d_{3K}^2
	d_{41}^2	d_{42}^2	...	d_{4K}^2
	d_{51}^2	d_{52}^2	...	d_{5K}^2
⋮	⋮	⋮	⋮	⋮
	d_{n1}^2	d_{n2}^2	...	d_{nK}^2

2D ALIGNMENT AND STABILITY

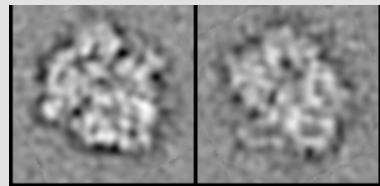
2D alignment is **stable** if perturbation of initial alignment parameters does not produce dramatically different results.

1. If a set of images is homogeneous, the result from reference-free alignment is stable even for very low SNR data.
2. The converse is true, i.e., if a set of images is stable, it must be homogeneous.

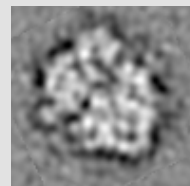
Assuming 1 and 2 are correct:
If we can find homogeneous subsets of images, we can solve the multi-reference alignment problem.

STABLE VS. UNSTABLE CLASSES: A TEST CASE

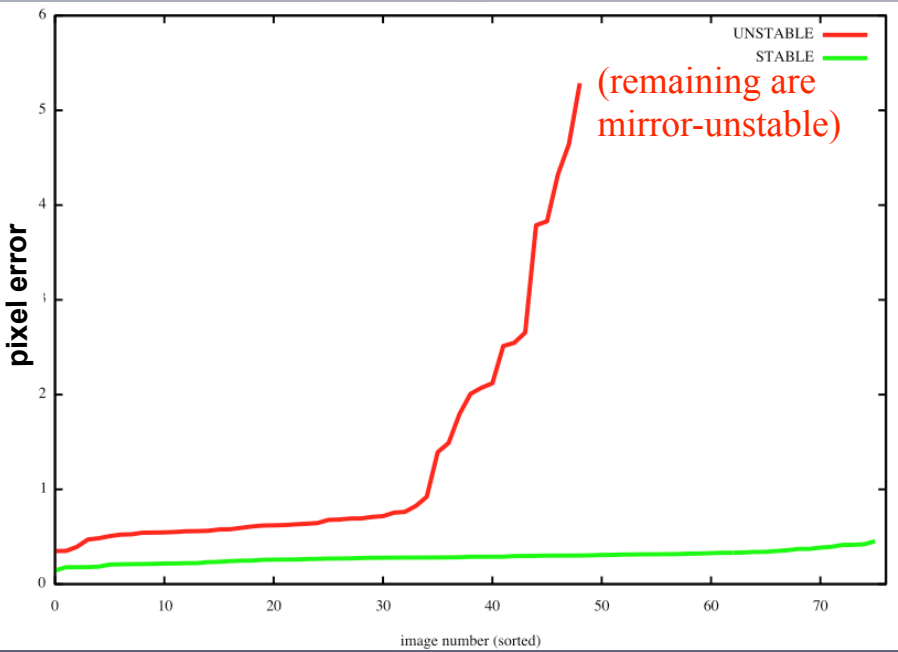
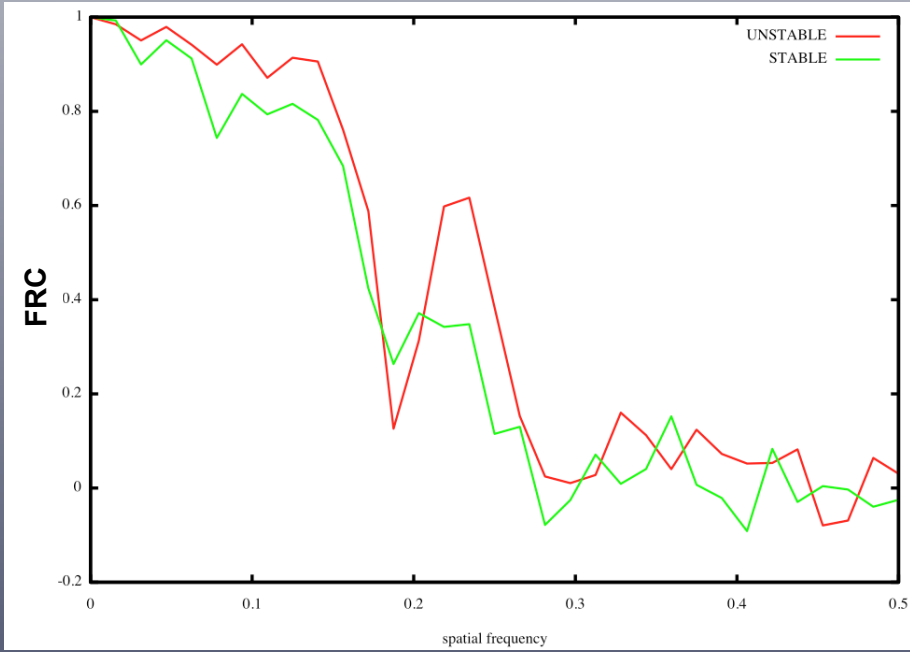
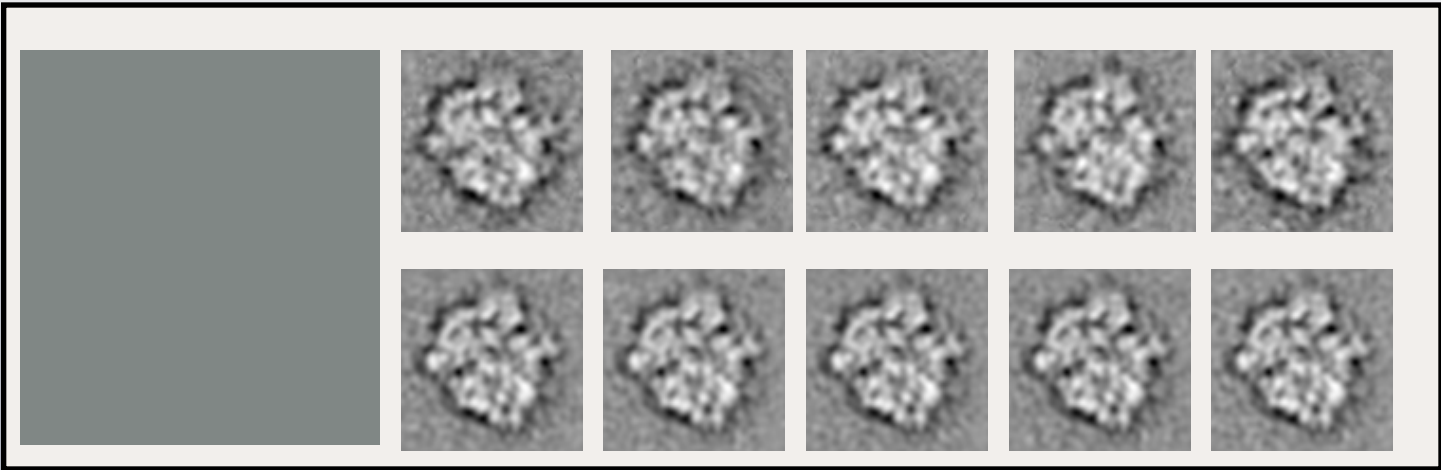
- Two groups were mixed 50-50, their respective averages are:



- The sum of these two averages:

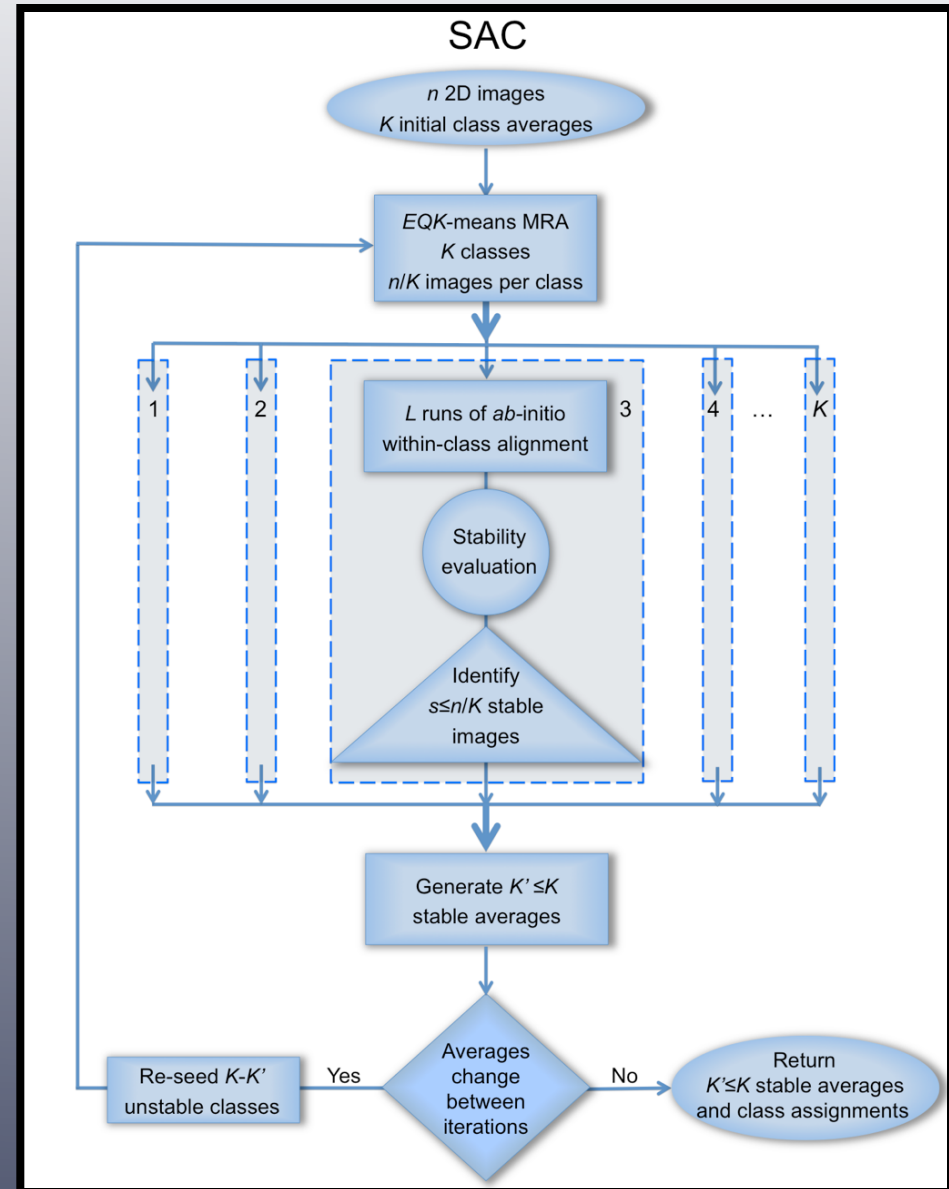


STABLE VS. UNSTABLE CLASSES: TEST RESULTS



A PROTOCOL FOR TESTING ALIGNMENT STABILITY

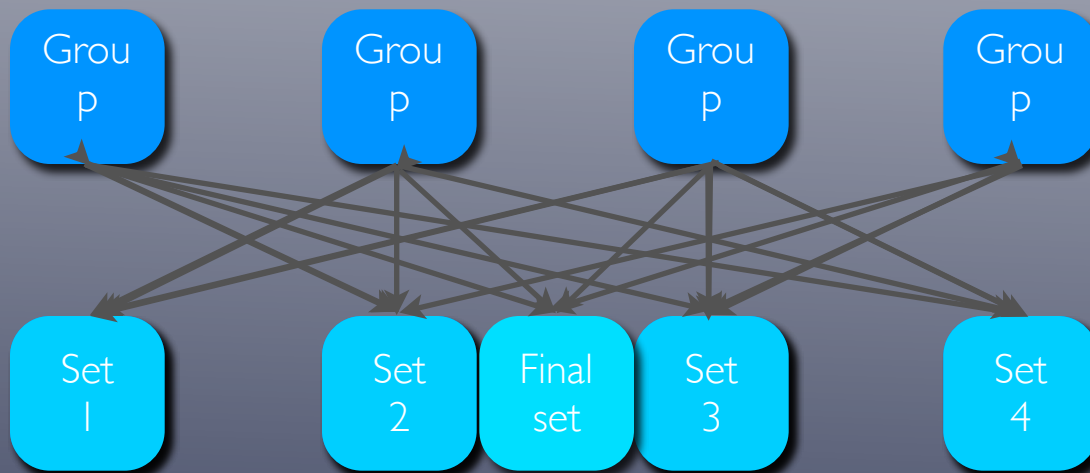
1. Run reference-free alignment L -times, using randomized initial orientation parameters
2. Bring all L sets of solutions into register by simultaneous minimization of the variance of orientation parameters (similar but not equivalent to alignment of resulting averages)
3. Compute pixel error for each image using orientation parameters for L positions it adopted
4. The set is called stable if the average of pixel errors for all images in L alignments is less than a predefined threshold (usually one pixel).



REPRODUCIBILITY

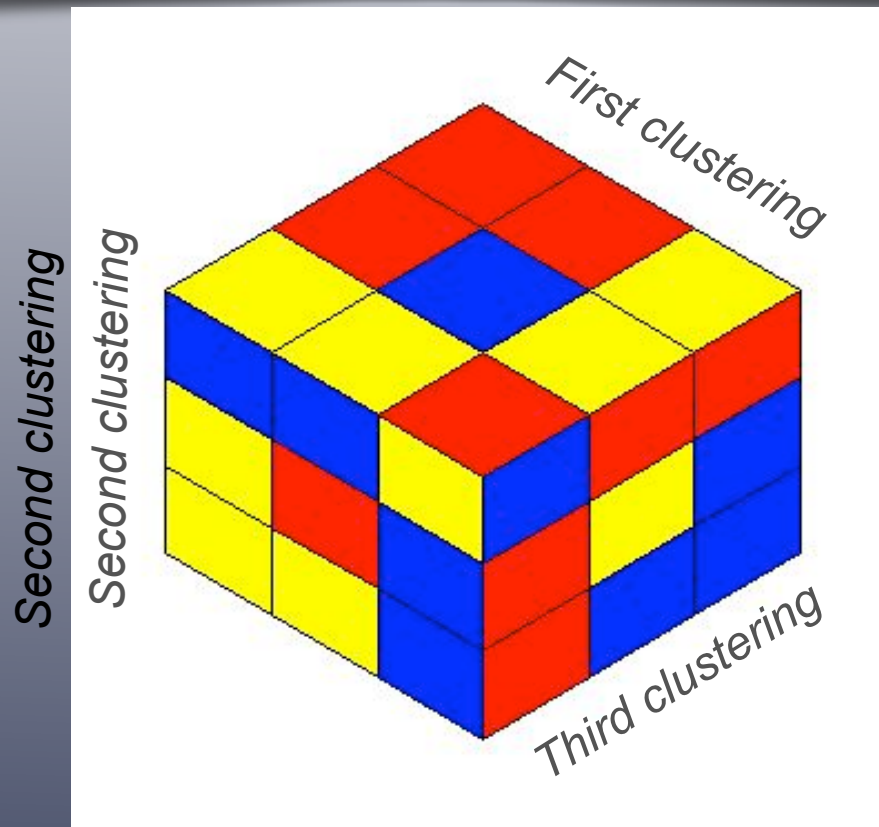
- Since *EQK*-means, even if combined with an alignment stability test, does not guarantee an optimum solution (global minimum) and stable groups can be fake, we require the solution to be reproducible over a number of quasi-independent runs.
- We have $m=4$ *EQK*-means runs analyzing the data in parallel. Once all runs produce their respective averages, we compare assignments of images to class averages and select as reproducible subsets shared among quasi-independent runs.

$m=4$



MULTIPLE ASSIGNMENT PROBLEM

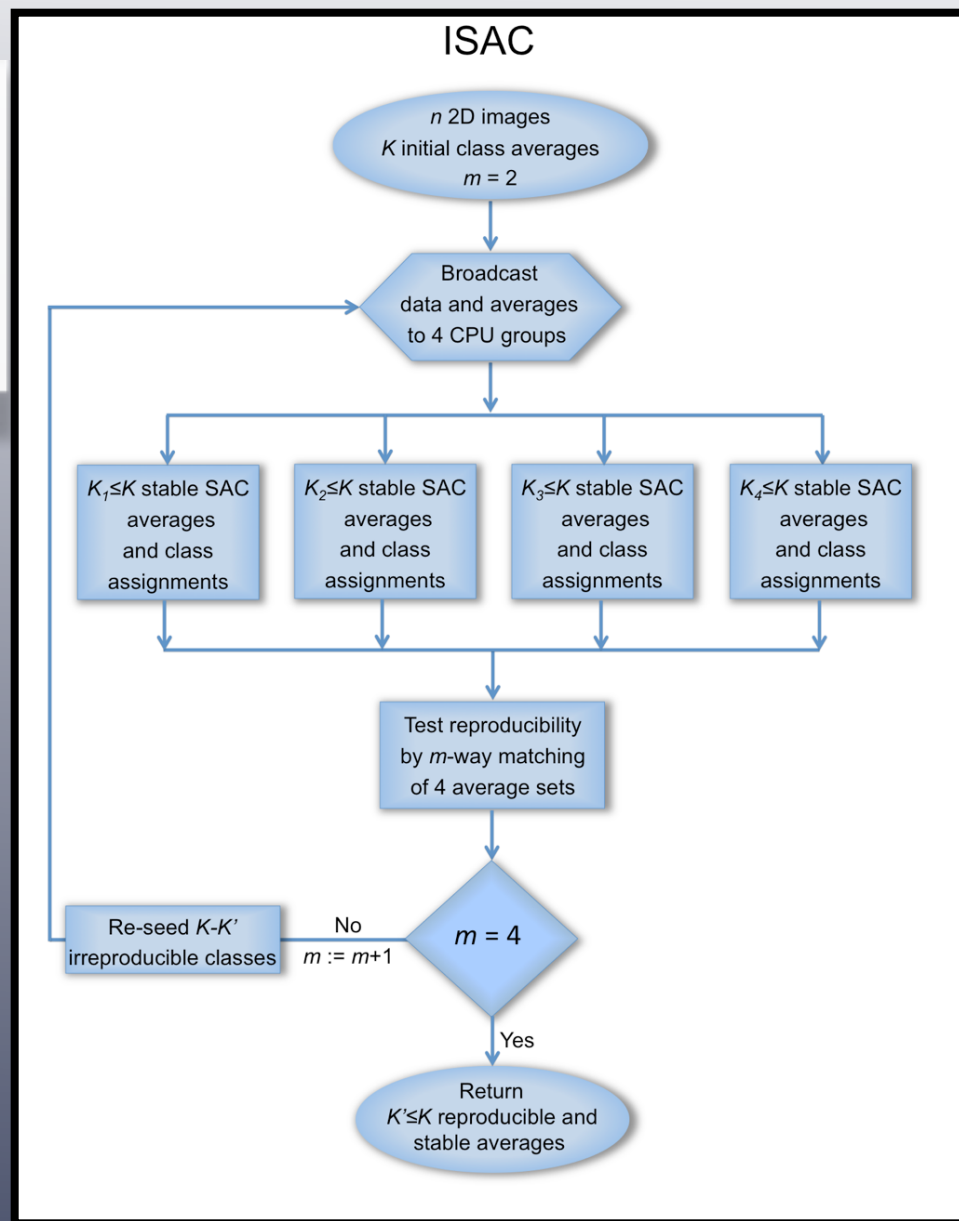
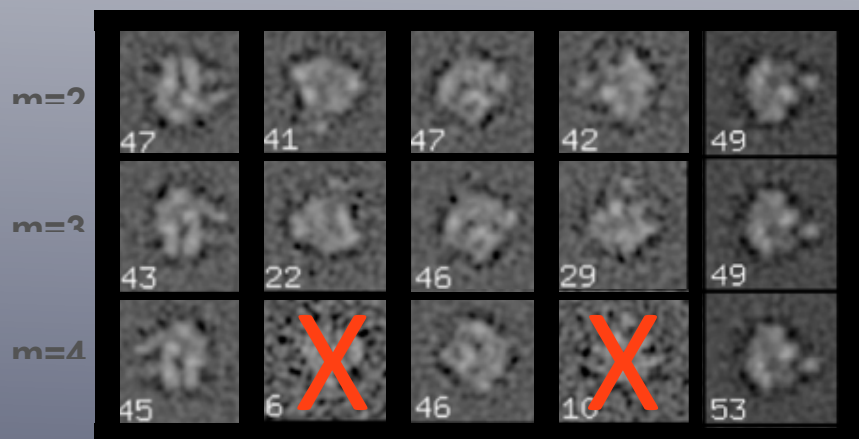
- For matching two sets of assignments, the solution is given in polynomial time by the Hungarian algorithm
- We developed a branching algorithm for matching m sets of assignments, that finds a nearly optimal solution in a reasonable time



ISAC: ITERATIVE STABLE ALIGNMENT AND CLUSTERING

We use 4 CPU groups to analyze the data set simultaneously

Irreproducible averages are eliminated

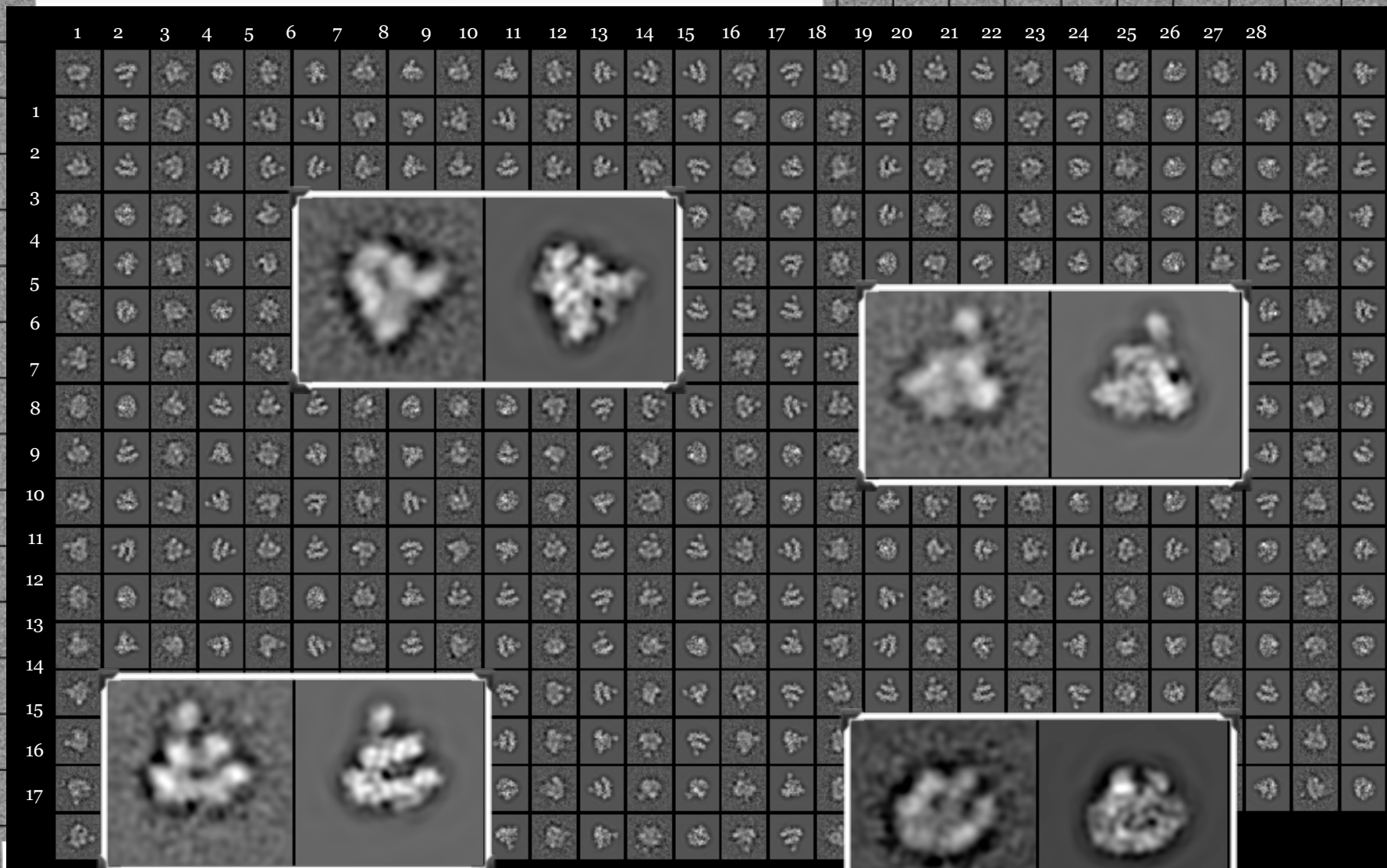


ISAC....

... is a very mysterious and powerful algorithm whose mystery is exceeded only by its power.



ISAC RESULTS for hRNAPII DATA SET



235 ISAC class averages matched to projections of a homologous X-ray structure

CONCLUSIONS

1. ISAC is a simple and intuitive approach (no equations!) based on concepts of stability and reproducibility.
2. ISAC operates exclusively on parameters and labels, not on image similarities as these are unreliable.
3. ISAC objectively generates reliable, validated 2D averages.
4. ISAC requires a minimal number of parameters:
 - Desirable number of images per group
 - Number of re-alignments L for the stability tests in SAC.
5. Reproducibility and stability test result in a relatively long computation time.

ACKNOWLEDGMENTS

Zhengfan Yang
Houston, TX



Francisco J. Asturias
Johnathan Chittulru
La Jolla, CA



Steve Ludtke
Baylor College, Houston

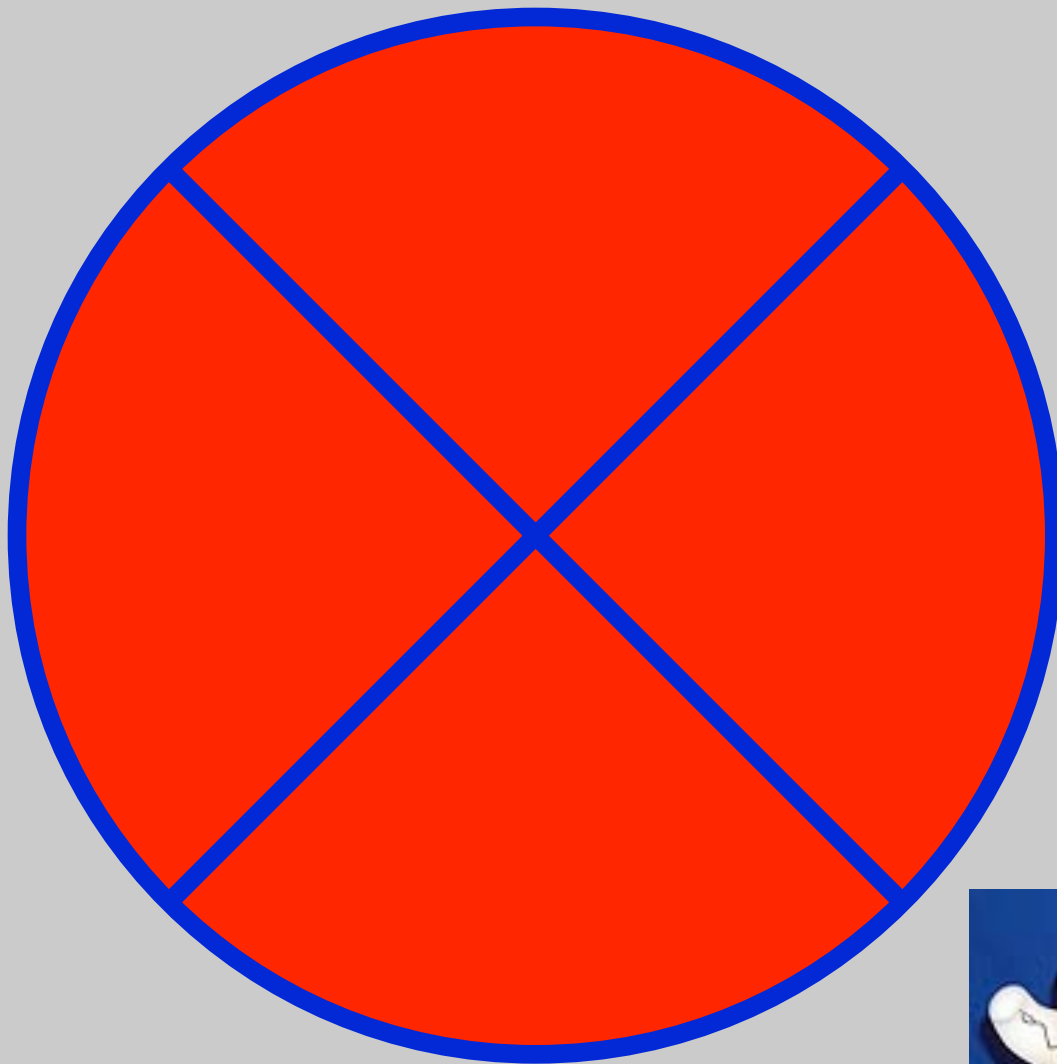


Christian M.T. Spahn
Charité, Berlin



NIH





Recommended reading

1. Penczek, P.A., Frank, J.: Resolution in Electron Tomography, in J. Frank (Ed.), Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell, 2 edn., Springer, Berlin, 307-330, 2006.
2. Vainshtein, B.K., Penczek, P.A.: Three-dimensional reconstruction, in U. Shmueli (Ed.), International Tables for Crystallography 3 edn., vol. B Reciprocal Space, 2008.
3. Penczek, P.A.: Single Particle Reconstruction, in U. Shmueli (Ed.), International Tables for Crystallography 3 edn., vol. B Reciprocal Space, 2008.
4. Penczek, P.A.: Fundamentals of three-dimensional reconstruction from projections. Methods Enzymol 2010, 482, 1-33.
5. Penczek, P.A.: Image restoration in cryo-electron microscopy. Methods Enzymol 2010, 482, 35-72.
6. Penczek, P.A.: Resolution measures in molecular electron microscopy. Methods Enzymol 2010, 482, 73-100.