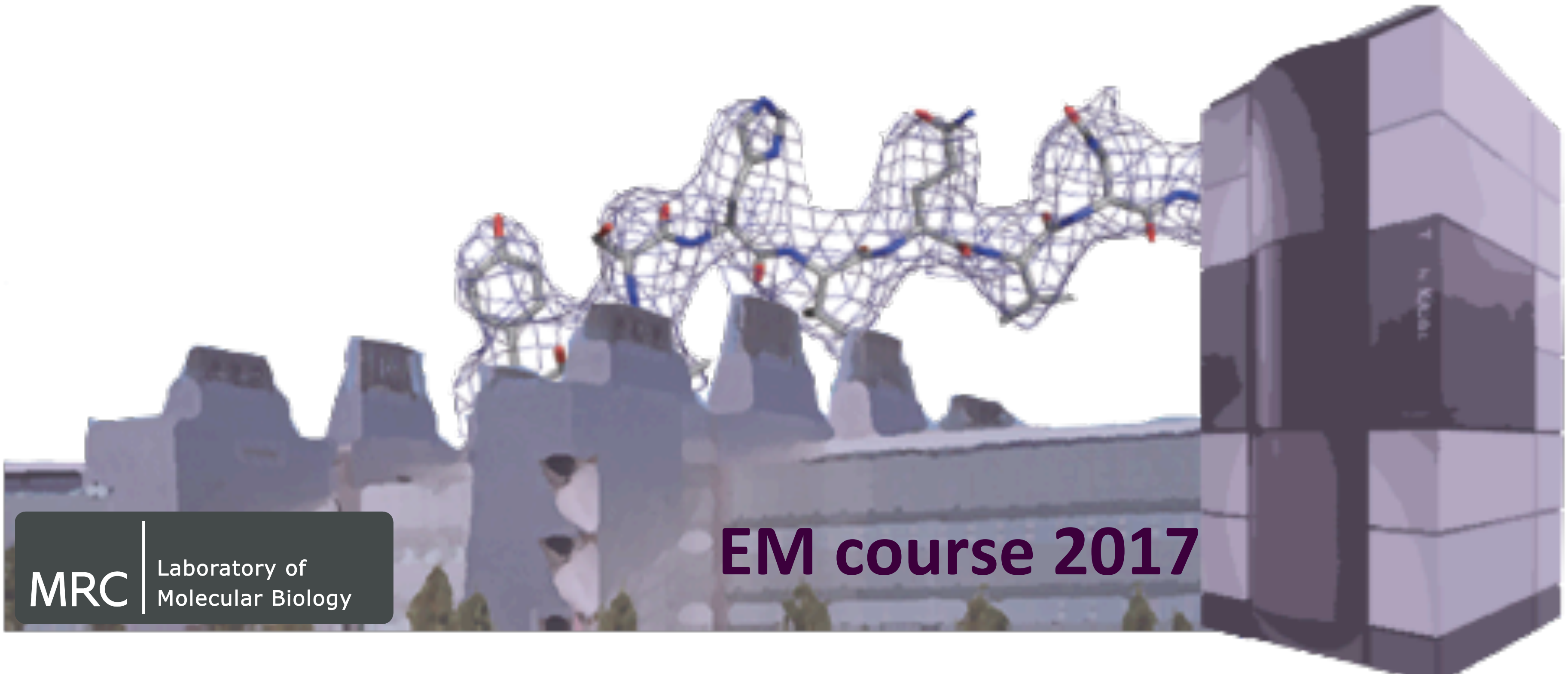
6. Image refinement

Sjors Scheres



Agenda

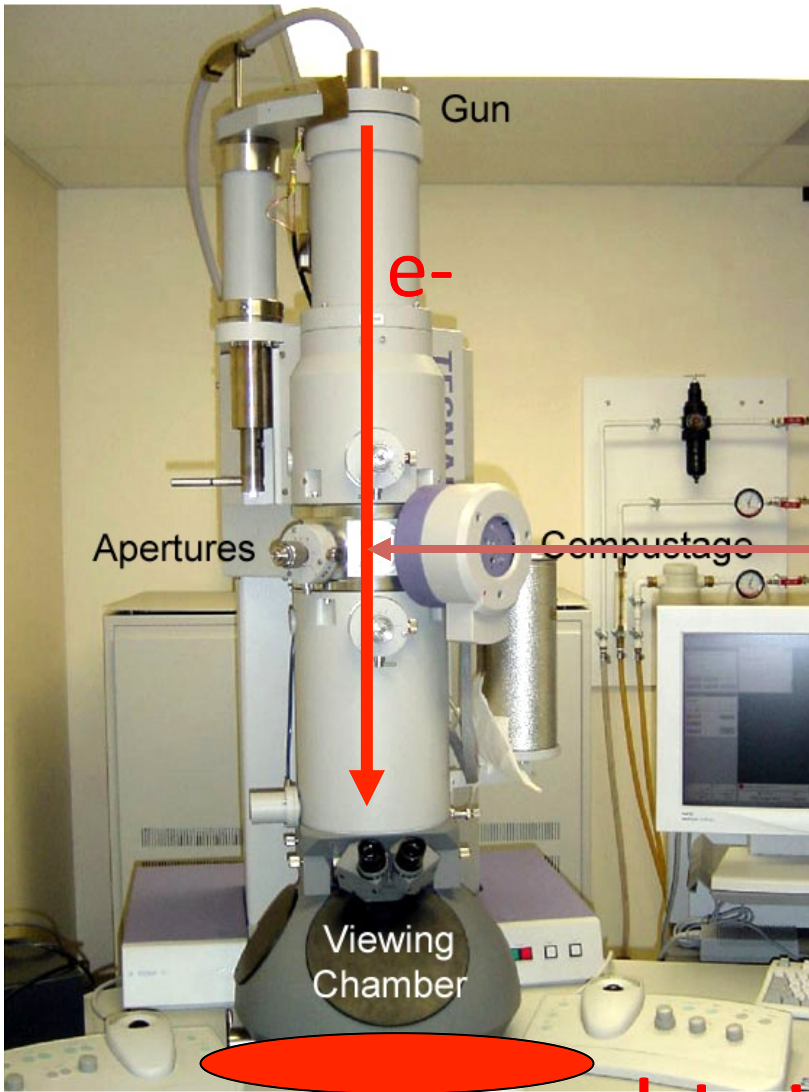
- An intuitive introduction
- Alignment
  - Dealing with the incomplete problem
  - maxCC vs ML (real-space)
- Classification
  - Multi-reference alignment in 2D
  - and in 3D
- Fourier-space formulation
  - Regularised likelihood optimisation
  - CTF correction

An intuitive introduction

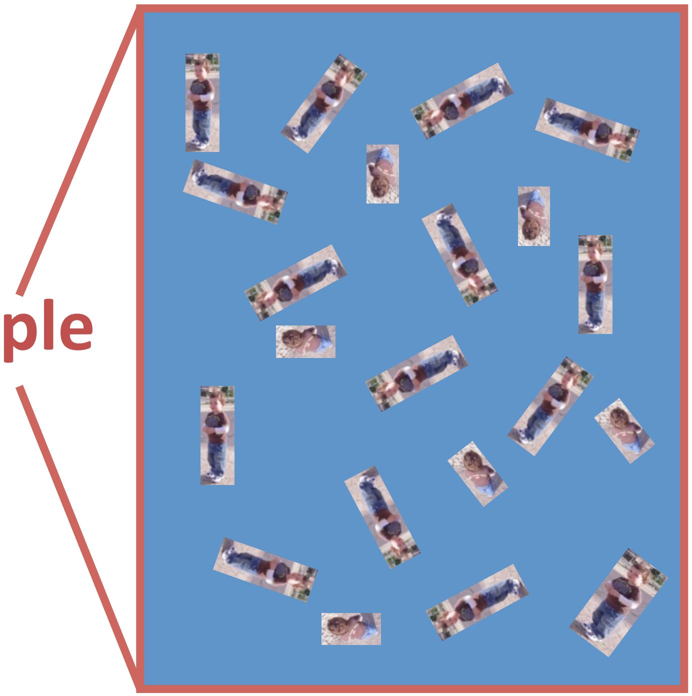
An example “protein”

Jan

Experimental setup



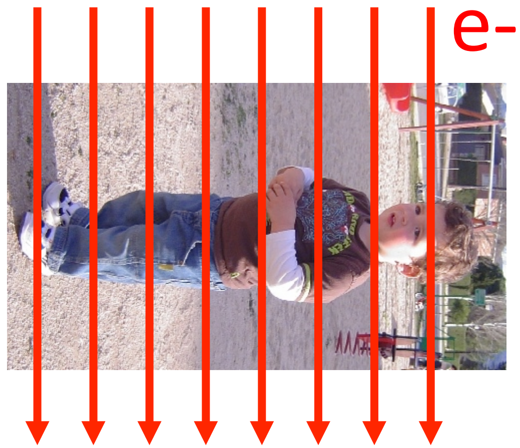
sample



detector

Electron microscopy imaging

3D object

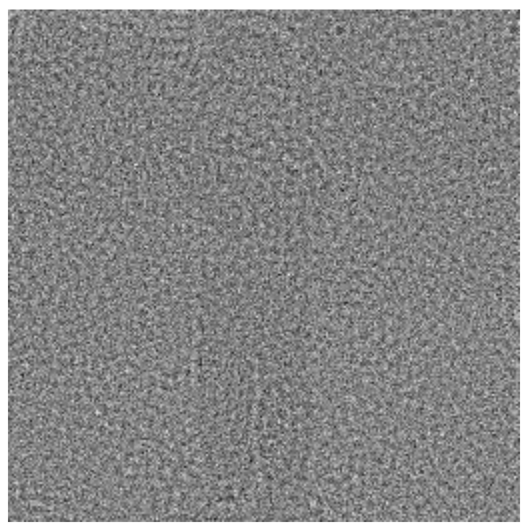


2D projection

We collect data in 2D,
but we want 3D info!

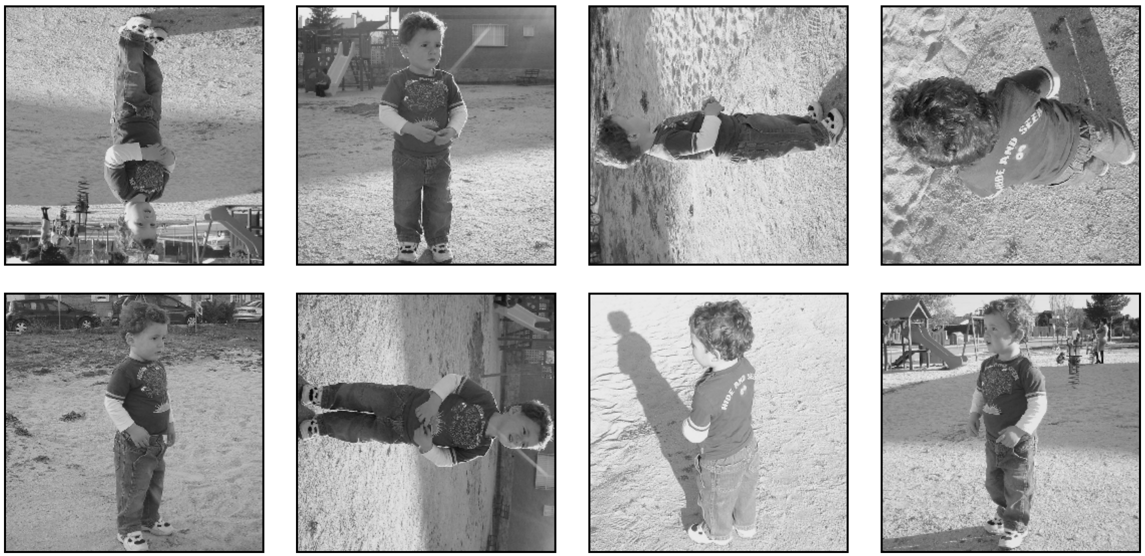
Further inconveniences

- Microscope imperfections introduce artefacts
  - Contrast Transfer Function (CTF)
- Large amounts of noise



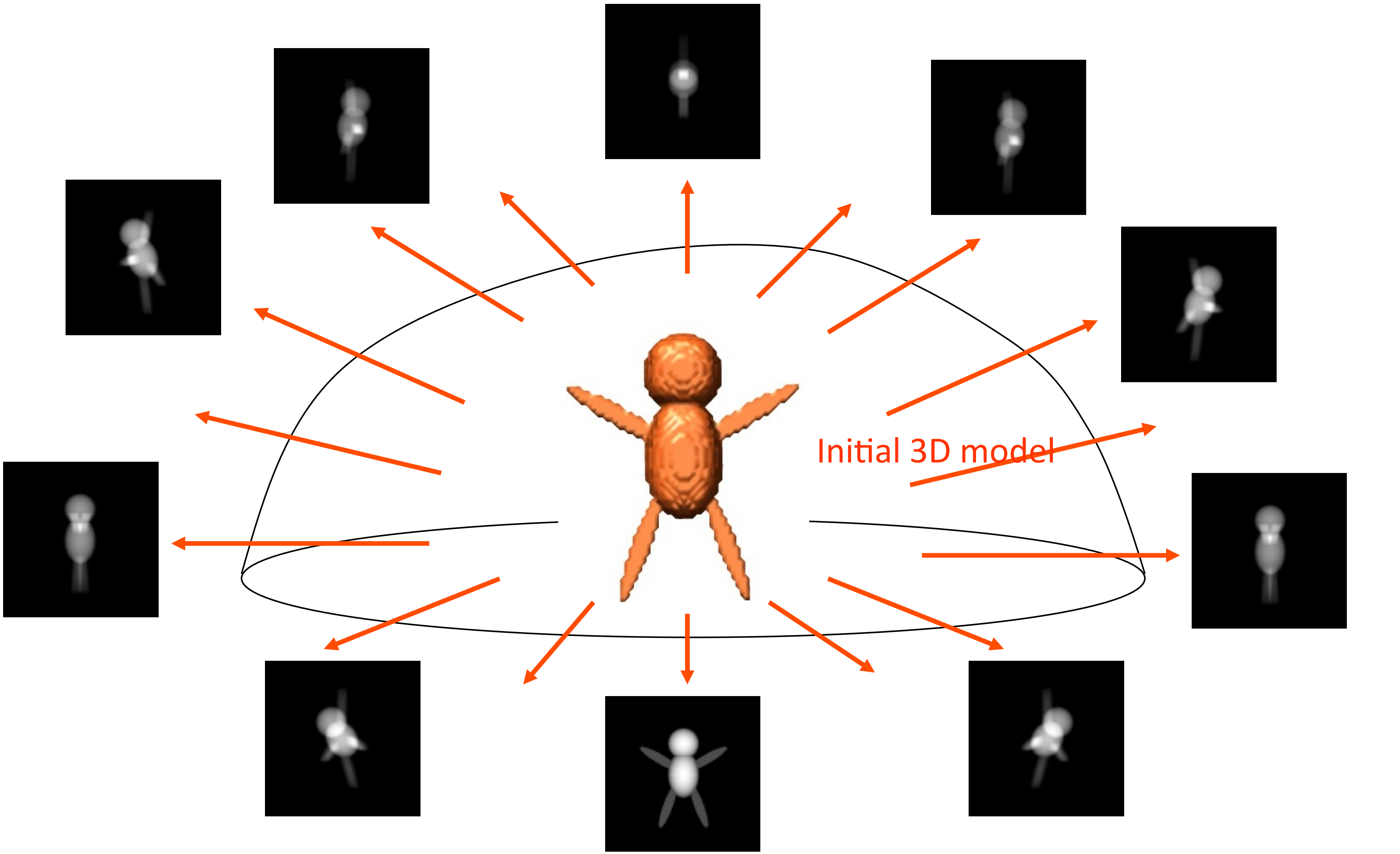
Single particle analysis

- Embedded in ice: many unknown orientations

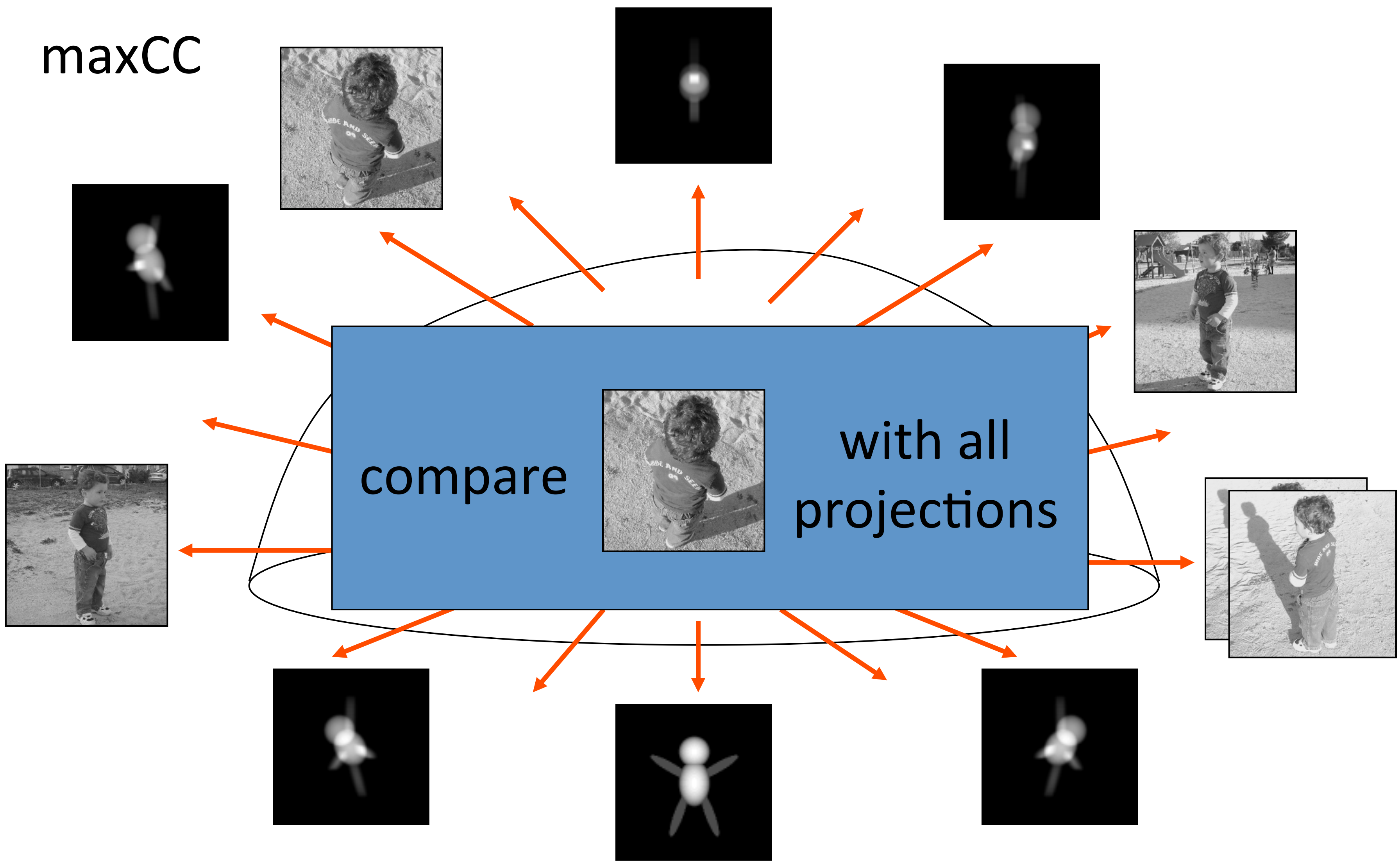


- Combine all 2D projections into a 3D reconstruction

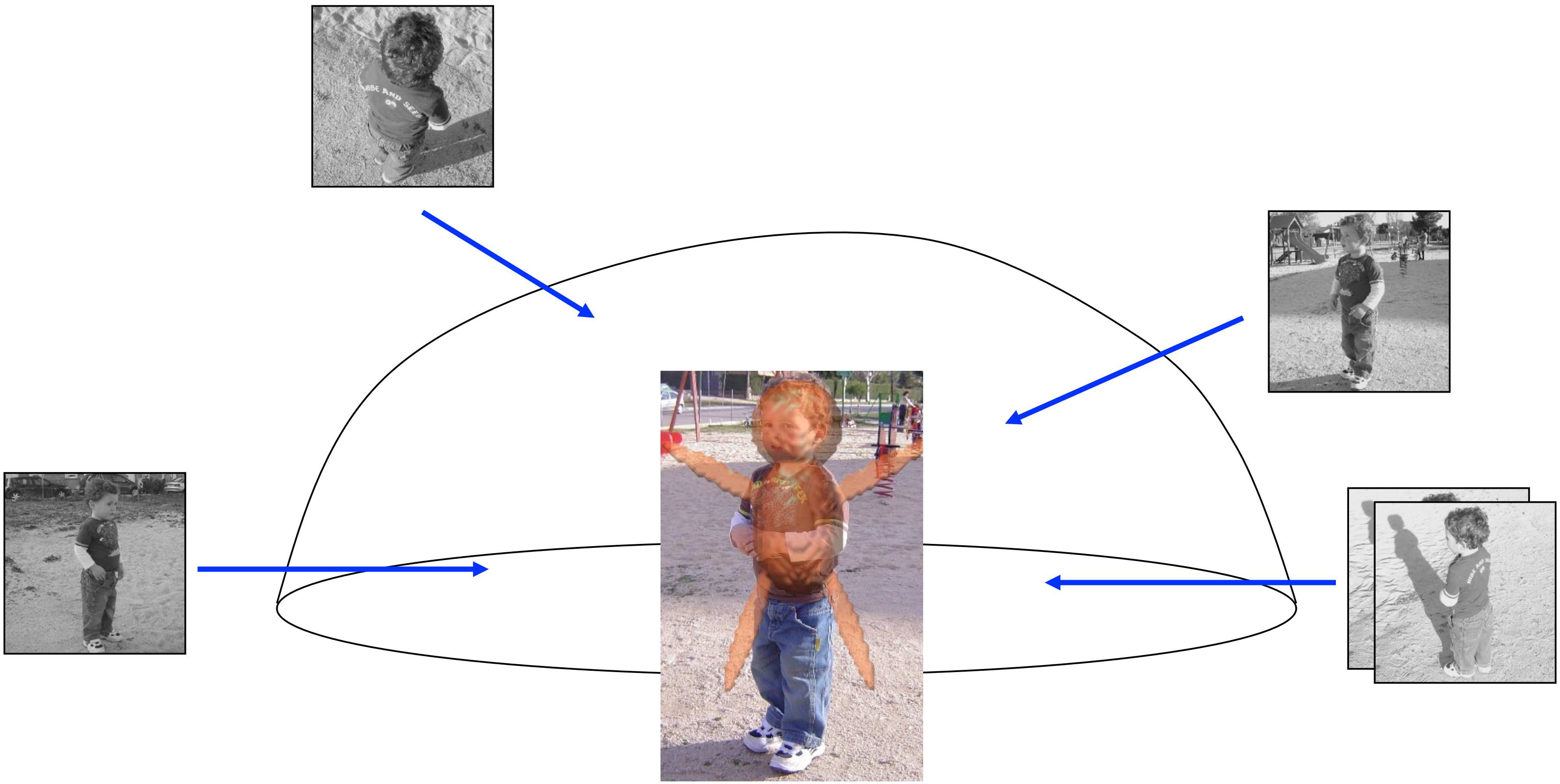
Projection matching



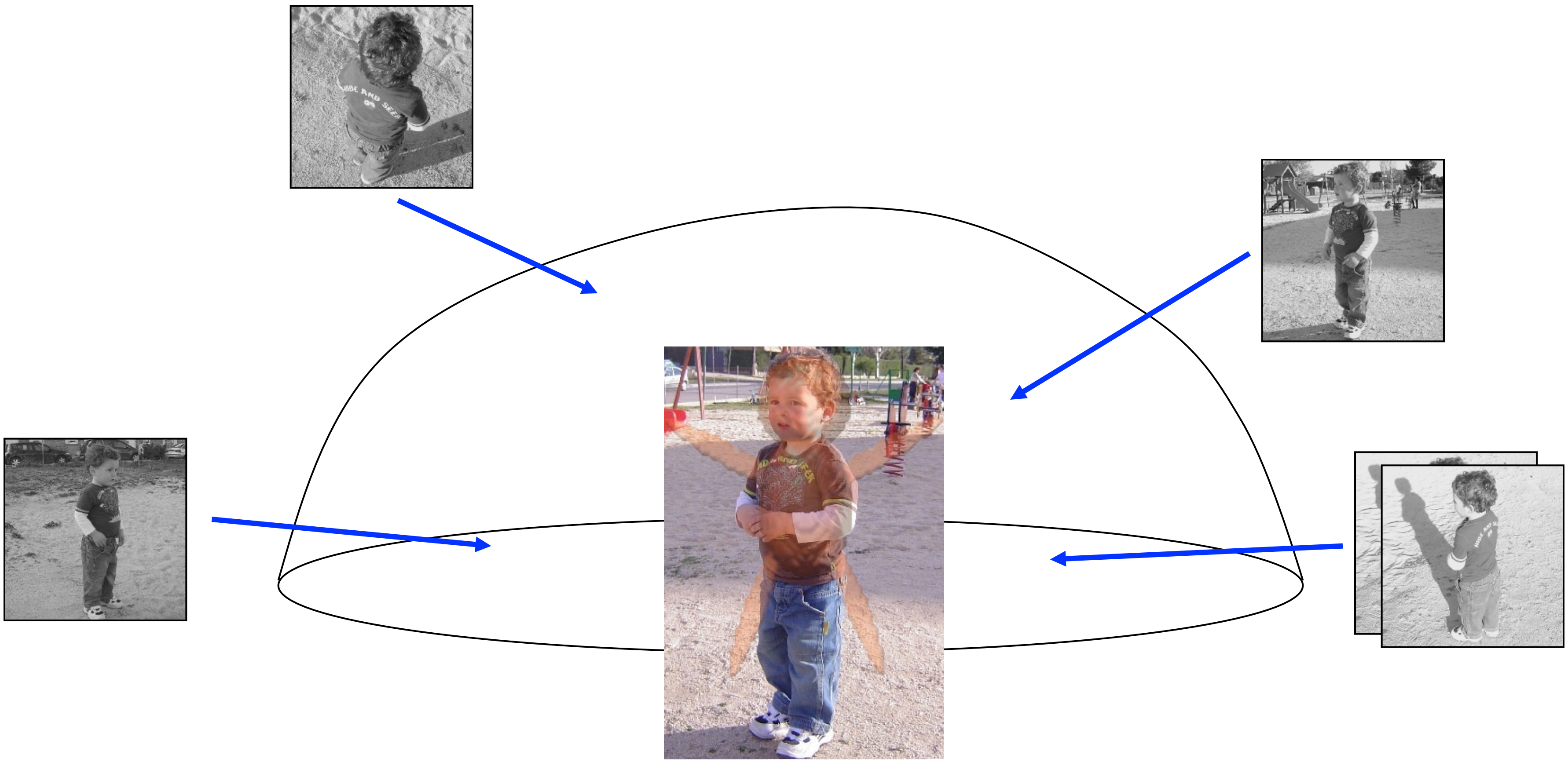
Projection matching



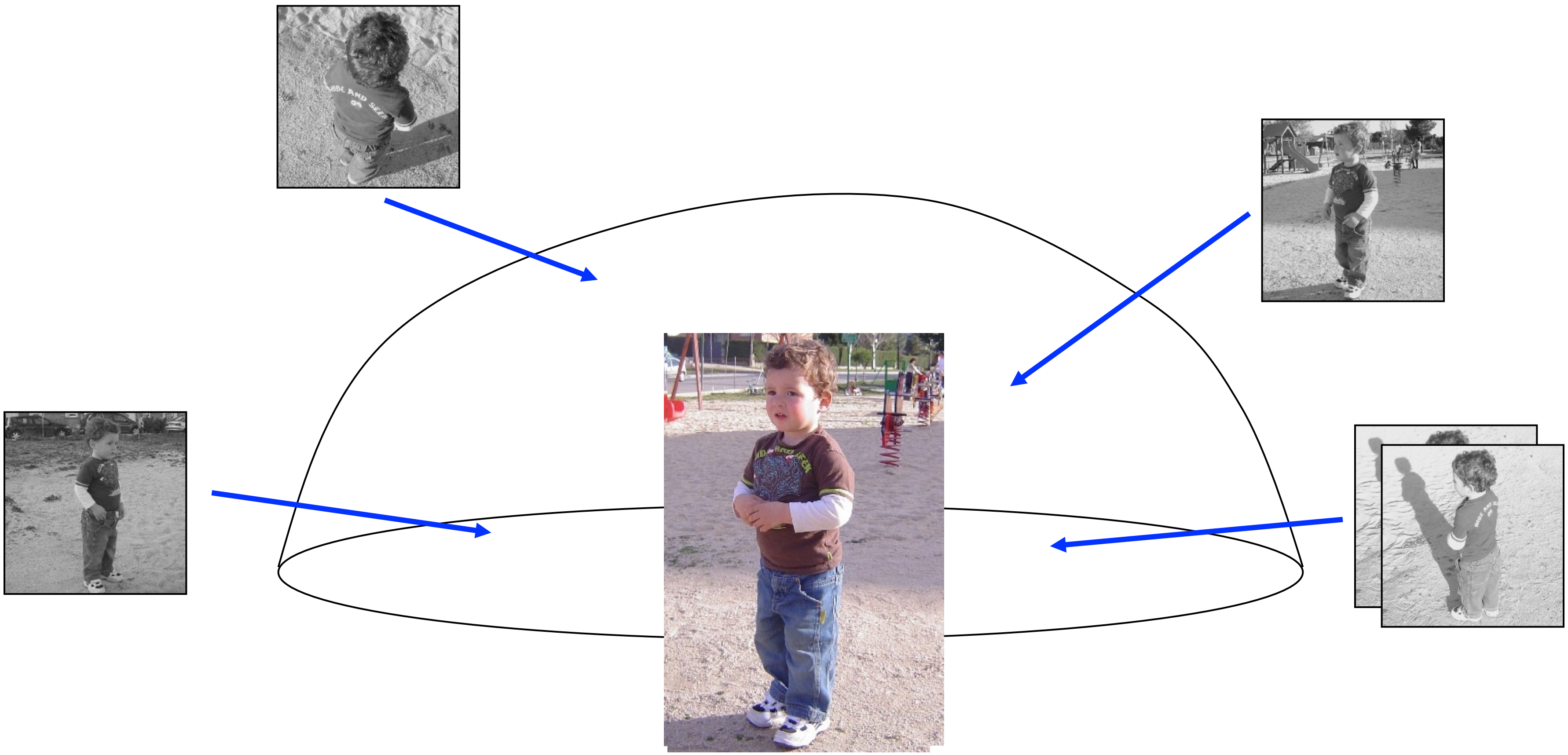
3D reconstruction



Iterative refinement



Iterative refinement



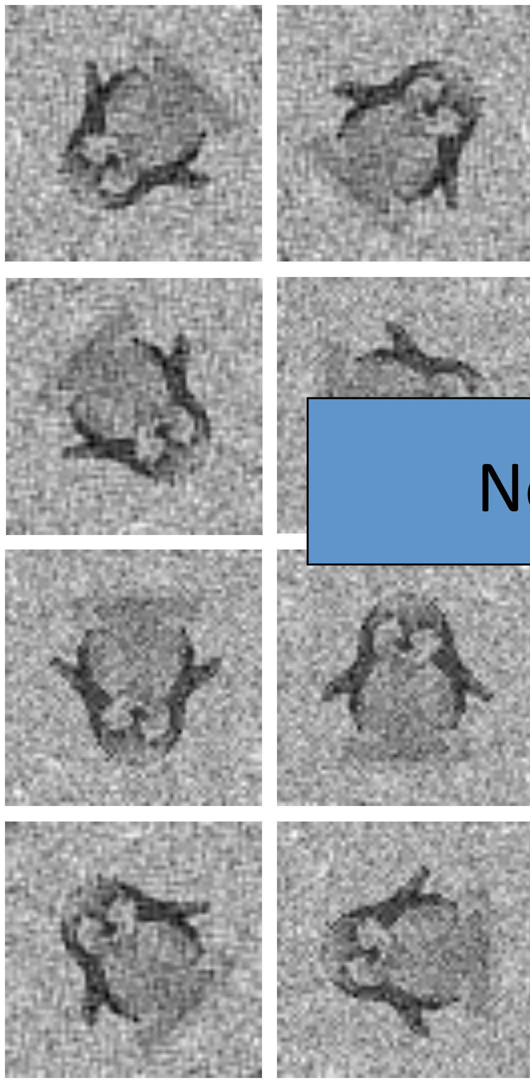
Alignment

Or how to ‘match’ projections

Incomplete data problems

- Part of the data was not observed experimentally
  - Orientations
  - Class assignments
- Difficult to solve!
  - Iterative methods?
- Complete data problem would be very easy to solve
- (Another famous one: the phase problem in XRD)

Incomplete data problems

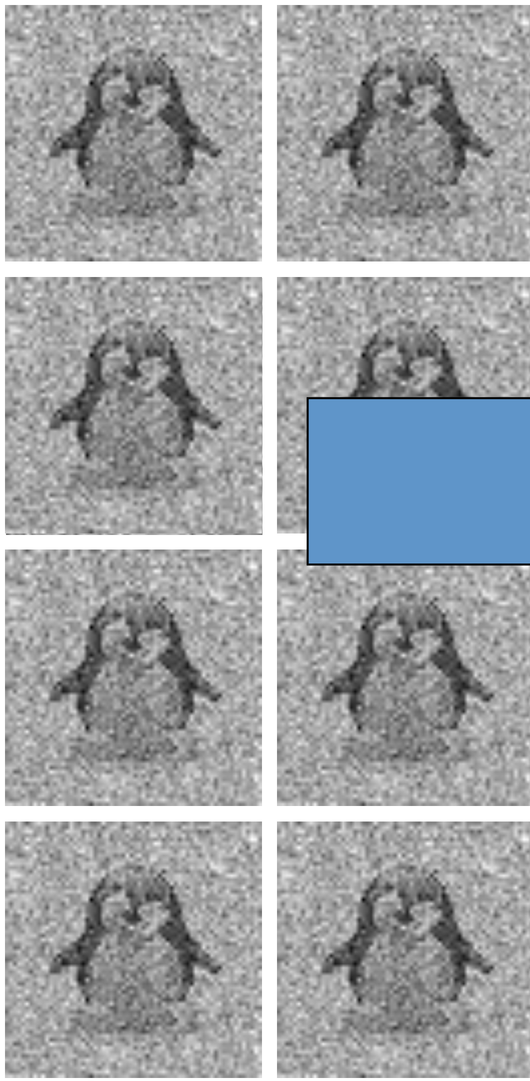


Not easy

Observed data (*X*): images

Missing data (Y): orientations

Complete data problems



white Gaussian noise

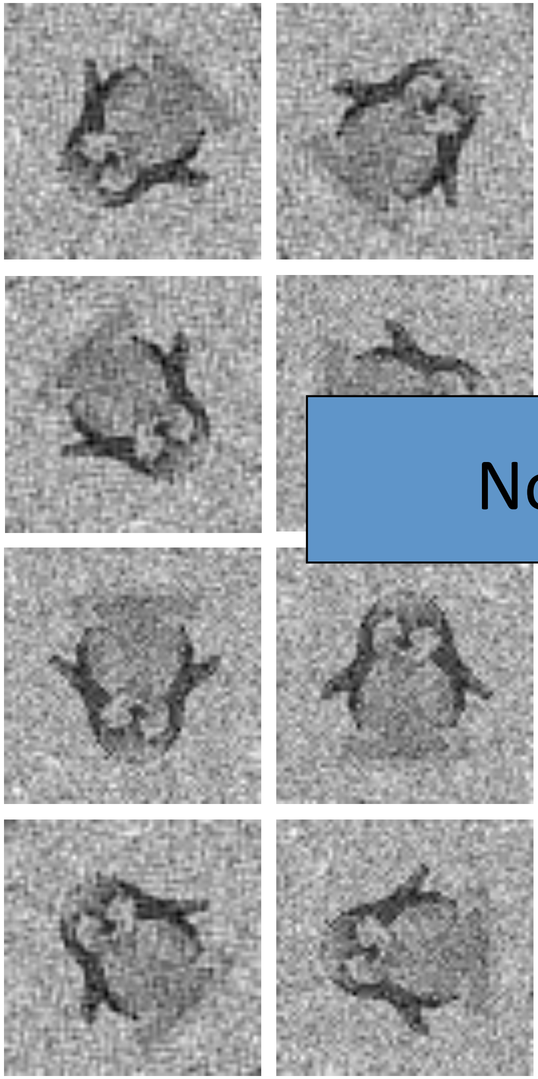
$$L(\Theta) = P(X \mid \Theta)$$

Easy!

$$A^{MLE} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Observed data (*X*): images

Incomplete data problems



Not easy

Observed data (*X*): images

Missing data (Y): orientations

Incomplete data problems

- Option 1: add *Y* to the model → Maximum cross-correlation / least-squares

$$L(Y,\Theta)=P(X\mid Y,\Theta)$$

- Option 2: marginalize over *Y* → Maximum Likelihood

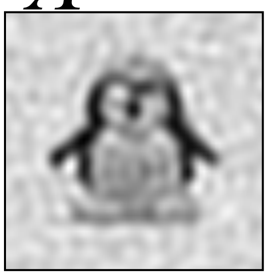
$$L(\Theta)=\underbrace{P(X\mid\Theta)}_{\text{Probability of X, regardless Y}}=\int_Y P(X\mid Y,\Theta)P(Y\mid\Theta)d\phi$$

The maxCC approach

Reference-based alignment

- Starts from some initial guess about the structure

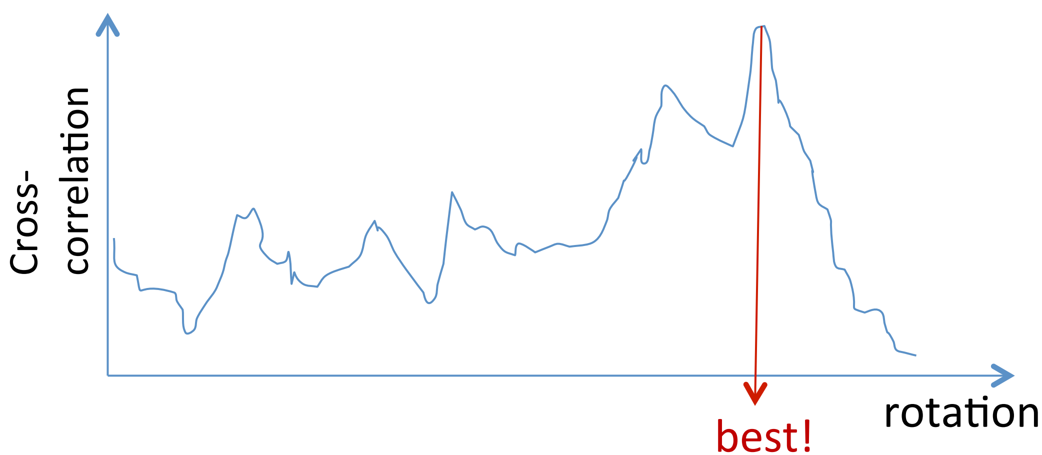
$A^{(n)}$



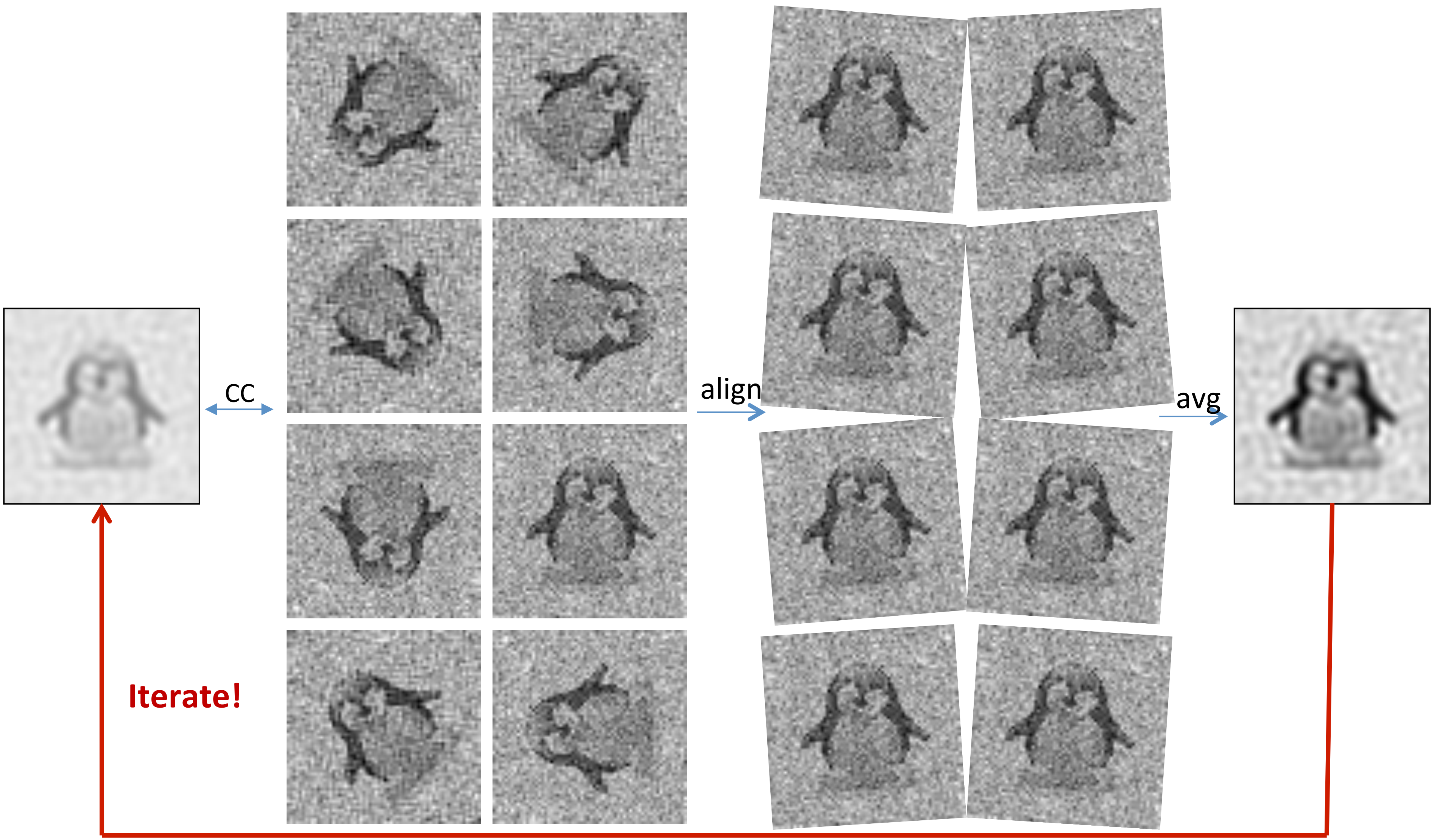
Cross-correlation

Compare initial guess with each experimental particle

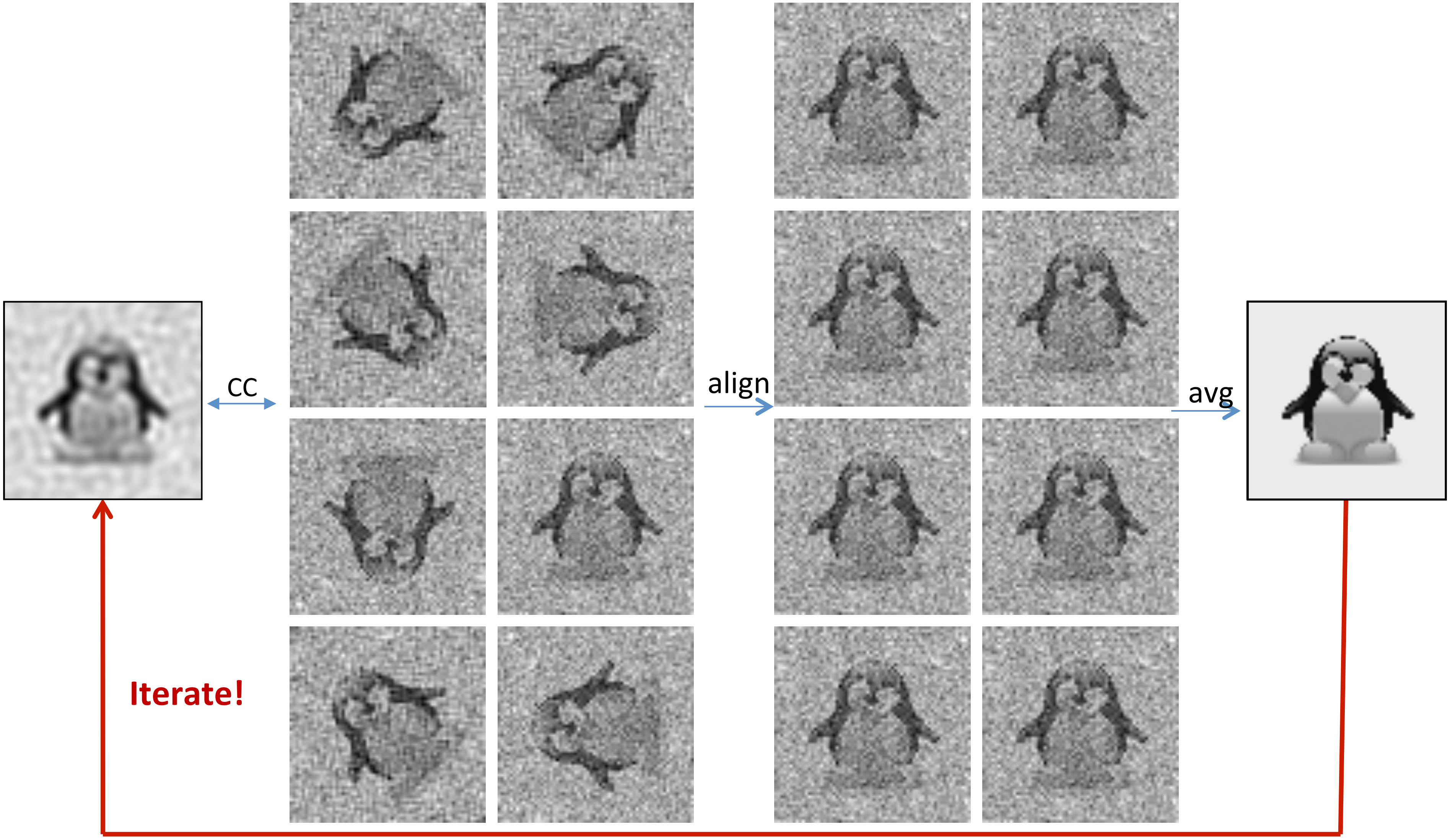
Illustrate CCF on the board



Align and average



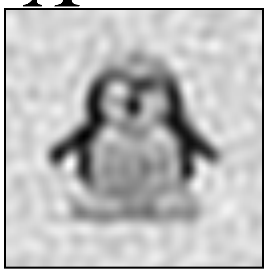
Align and average

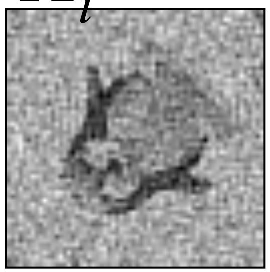


The ML approach

Maximum likelihood

$A^{(n)}$ X_i





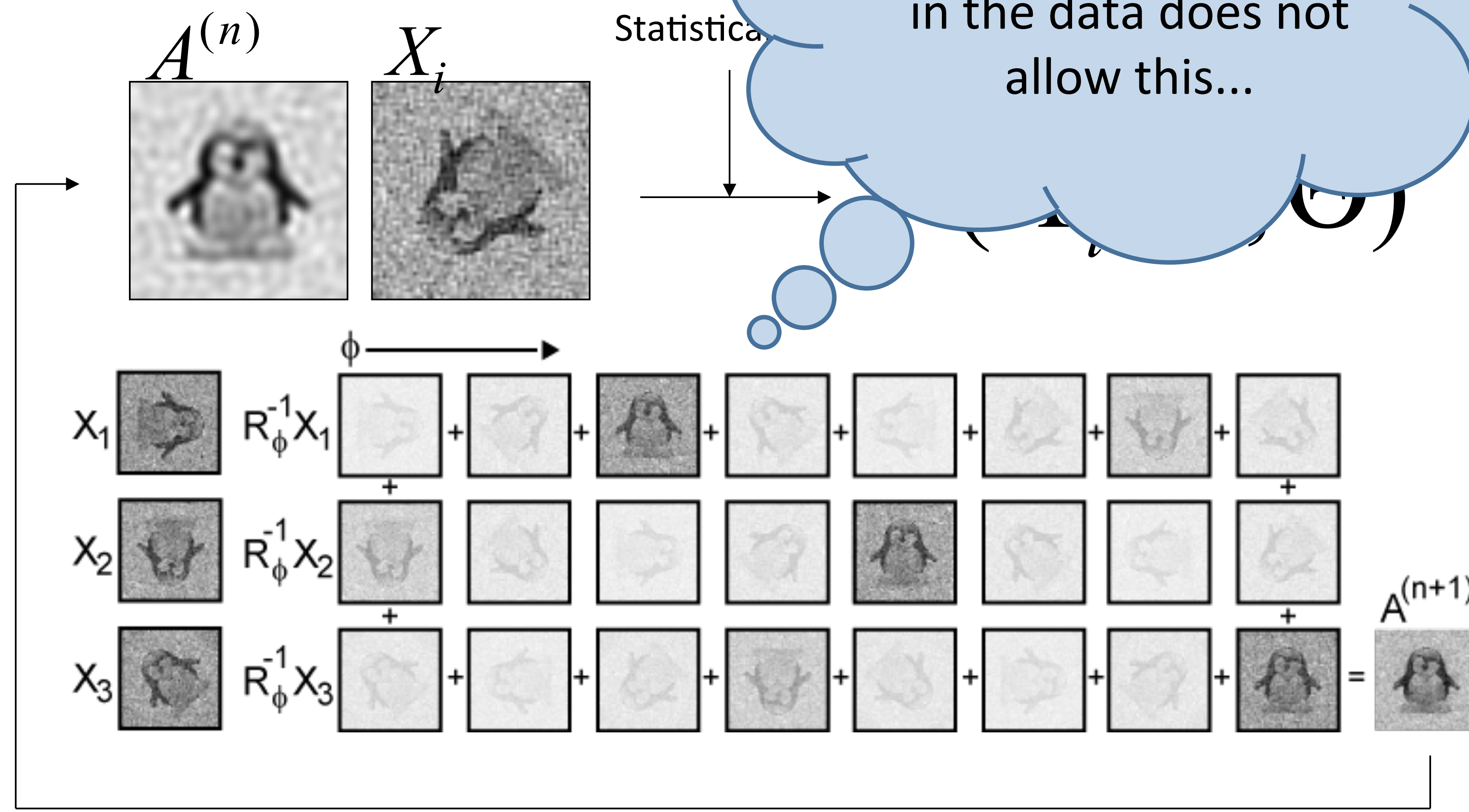
Statistical model

$$\longrightarrow P(X_i \mid \phi, \Theta)$$

Illustrate PDF on the board

Maximum li

Do not assign discrete orientations if the noise in the data does not allow this...



Sigworth, J. Struct. Biol., 1998

Incomplete data problems

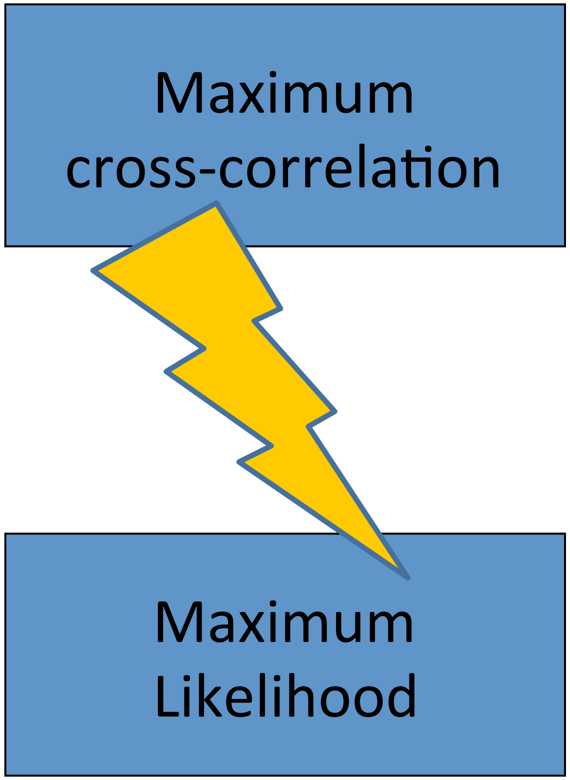
- Option 1: add *Y* to the model

$$L(Y,\Theta) = P(X \mid Y,\Theta)$$

- Option 2: marginalize over *Y*

$$L(\Theta) = P(X \mid \Theta) = \int_Y P(X \mid Y,\Theta) P(Y \mid \Theta) d\phi$$

Probability of X,
regardless Y



Incomplete data problems

Maximum cross-correlation

In the limit of **noiseless data** the Two techniques are equivalent!

Maximum Likelihood

Many software packages now use ML: cryoSPARC, SPARX/SPHIRE, FREALIGN, XMIPP, RELION

Read more? See *Methods in Enzymology*, **482** (2010)

Classification

The 2D multi-reference algorithm

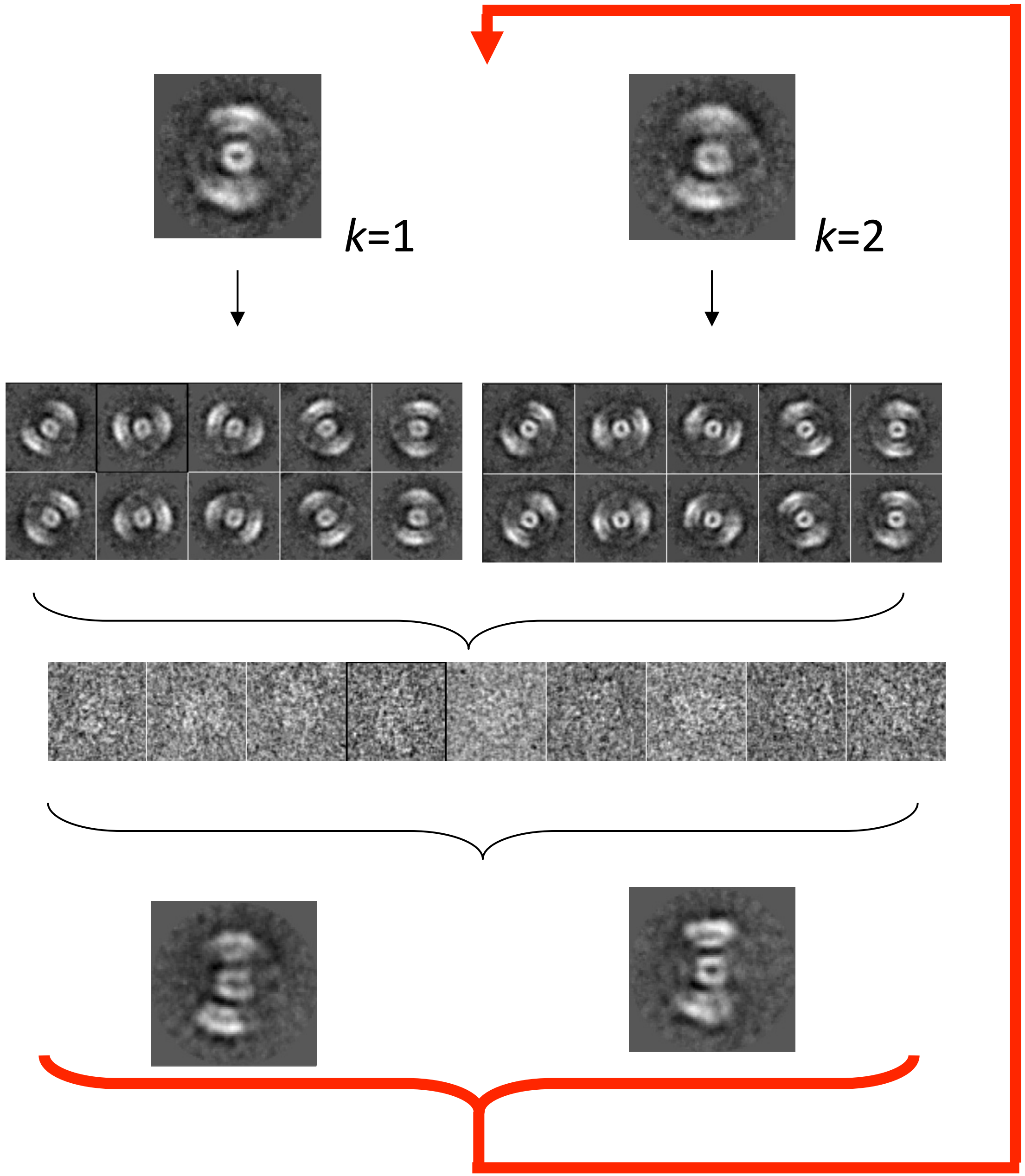
estimates for K
2D objects

sampled rotations 360°

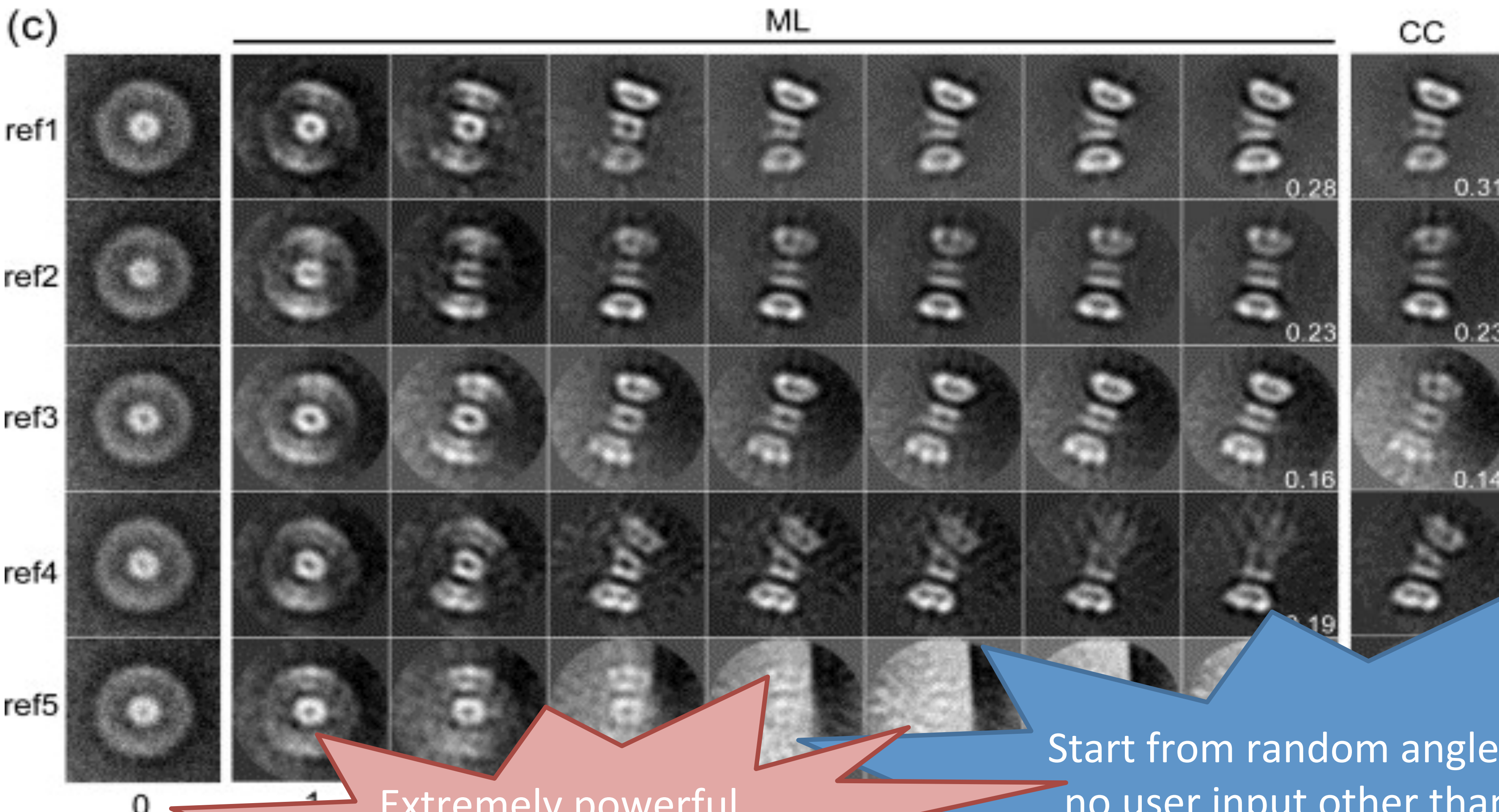
for each image, calculate all

$$P(\text{image}_i \mid k, \text{rot})$$

calculate new 2D average
as *probability weighted averages*



Reference-free 2D class averaging



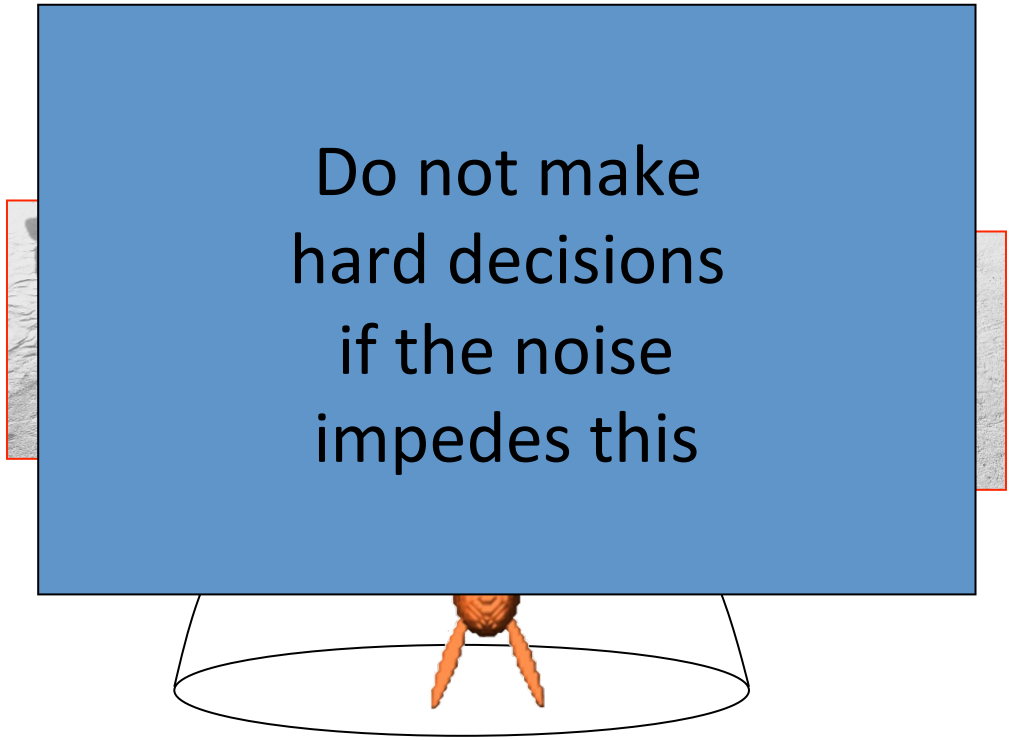
Extremely powerful to clean & assess your data

Start from random angles: no user input other than number of classes!!

Scheres et al (2005) J.Mol.Biol.

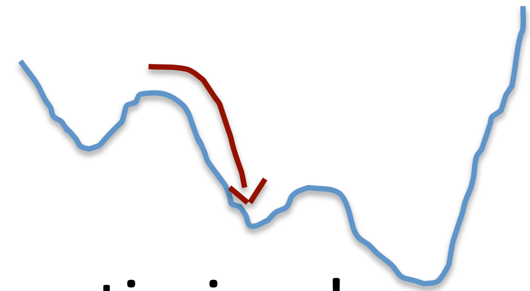
3D alignment & classification

3D ML refinement



“Probability-weighted angular assignment”

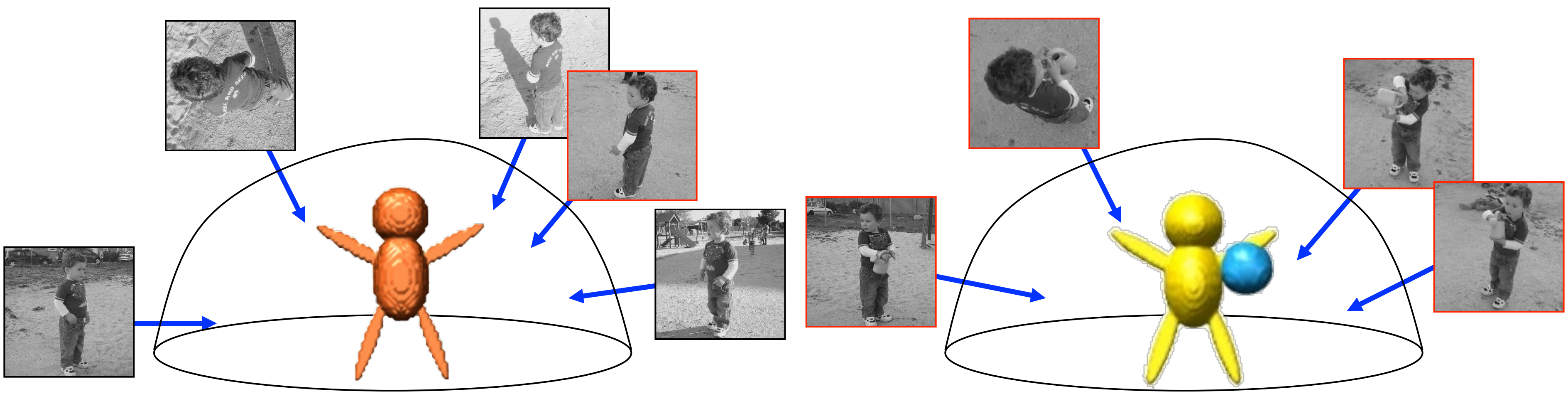
Initial model



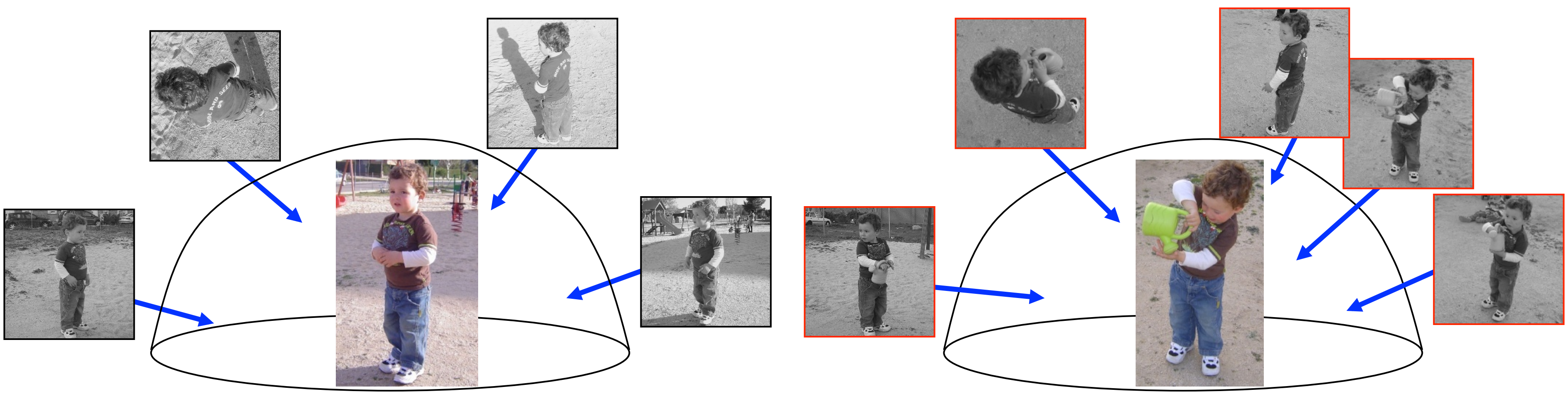
- Expectation-Maximisation is a local optimizer!
  - Gets stuck in nearest (local) minimum
- Bad model in -> bad model out!!!
  - Much less of a problem with high-resolution data
- Stochastic methods may reach global minimum
  - Stochastic Hill Climbing (SIMPLE)
  - Stochastic Gradient Descent (cryoSPARC & RELION)

Structural heterogeneity

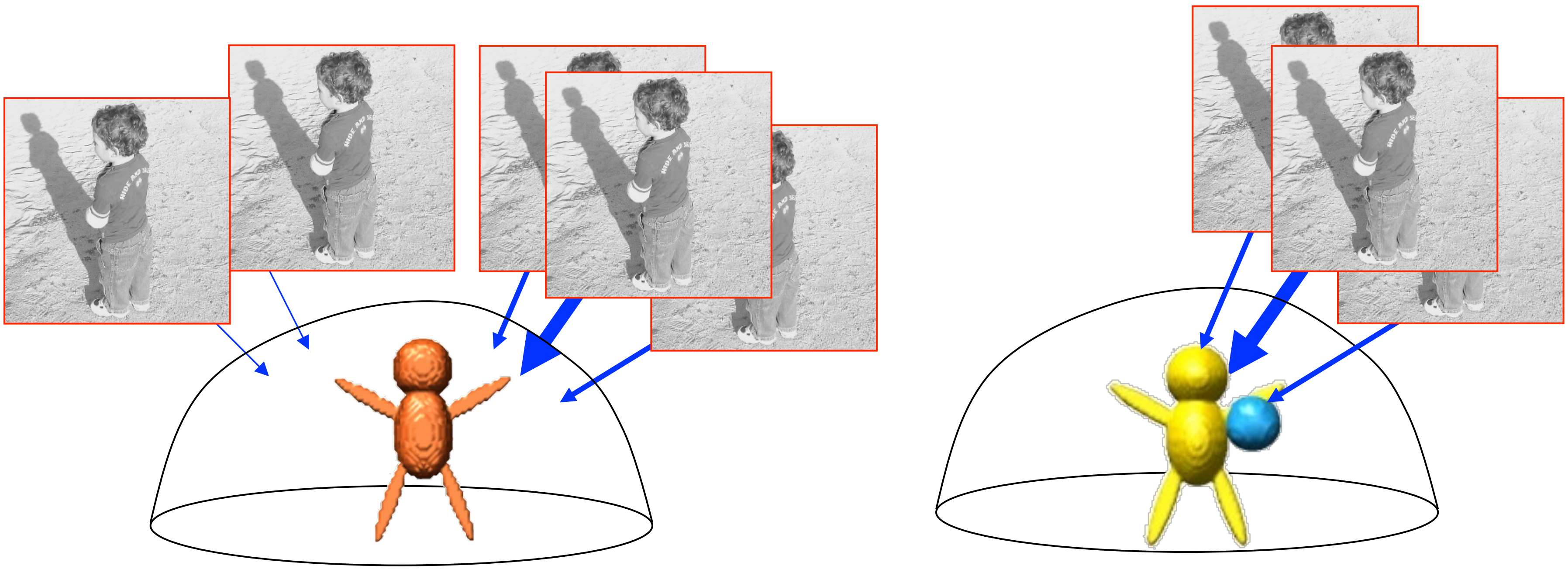
Multi-reference refinement



Multi-reference refinement



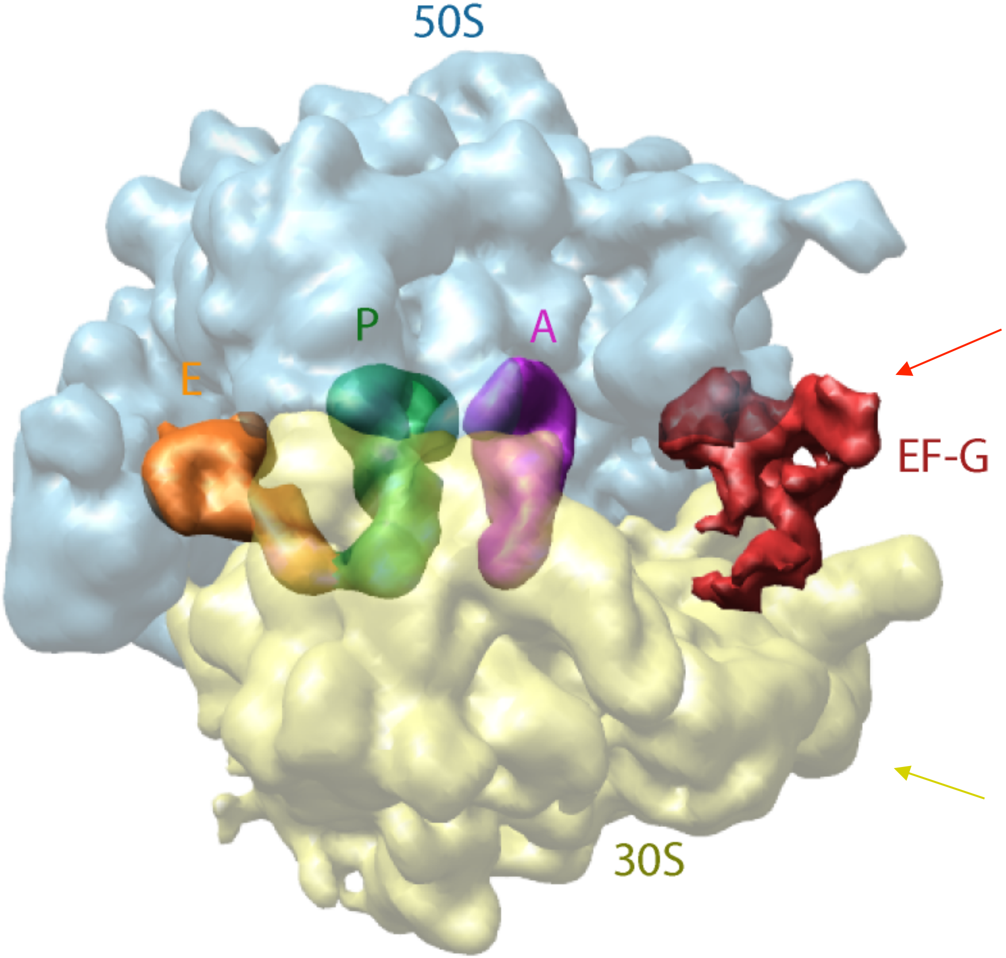
ML3D classification



"Probability-weighted angular assignment"

Prelim. ribosome reconstruction

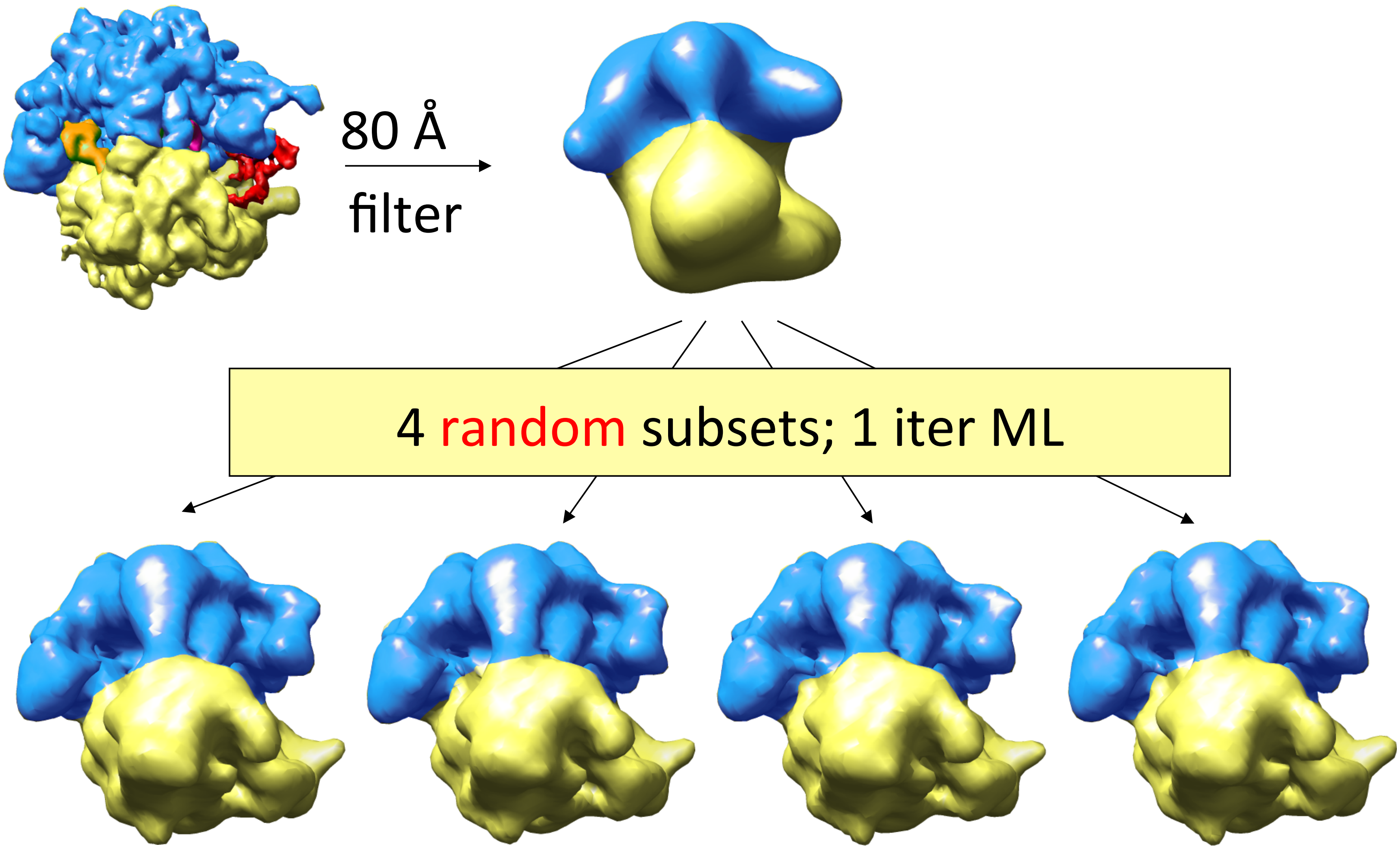
91,114 particles; 9.9 Å resolution



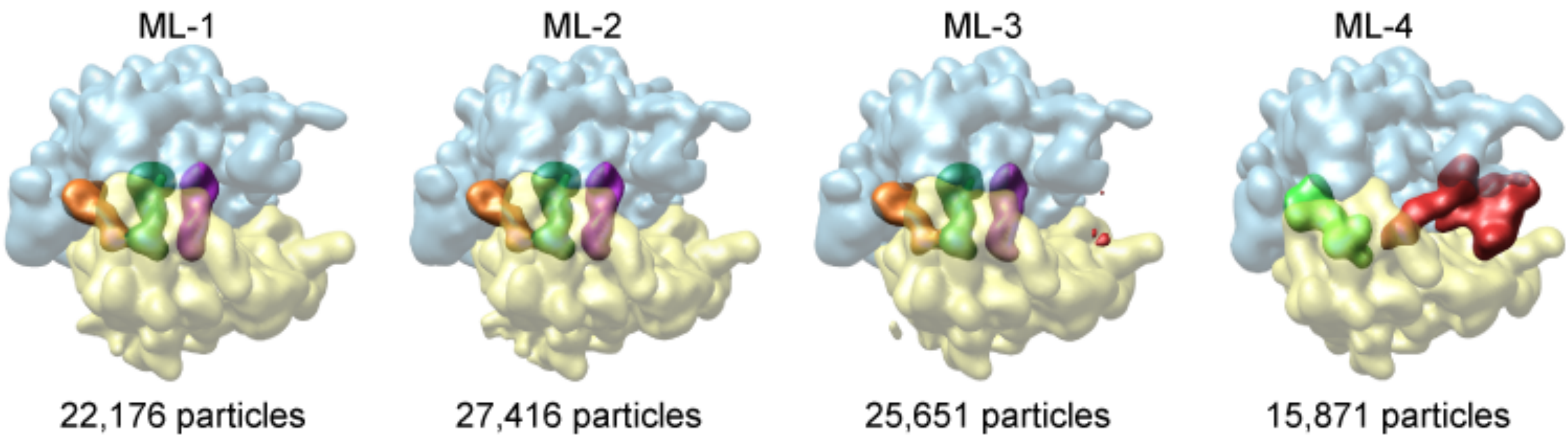
fragmented
(depicted at a
lower threshold)

blurred

Seed generation



ML-derived classes



no ratcheting; no EF-G; 3 tRNAs
differences: overall rotations

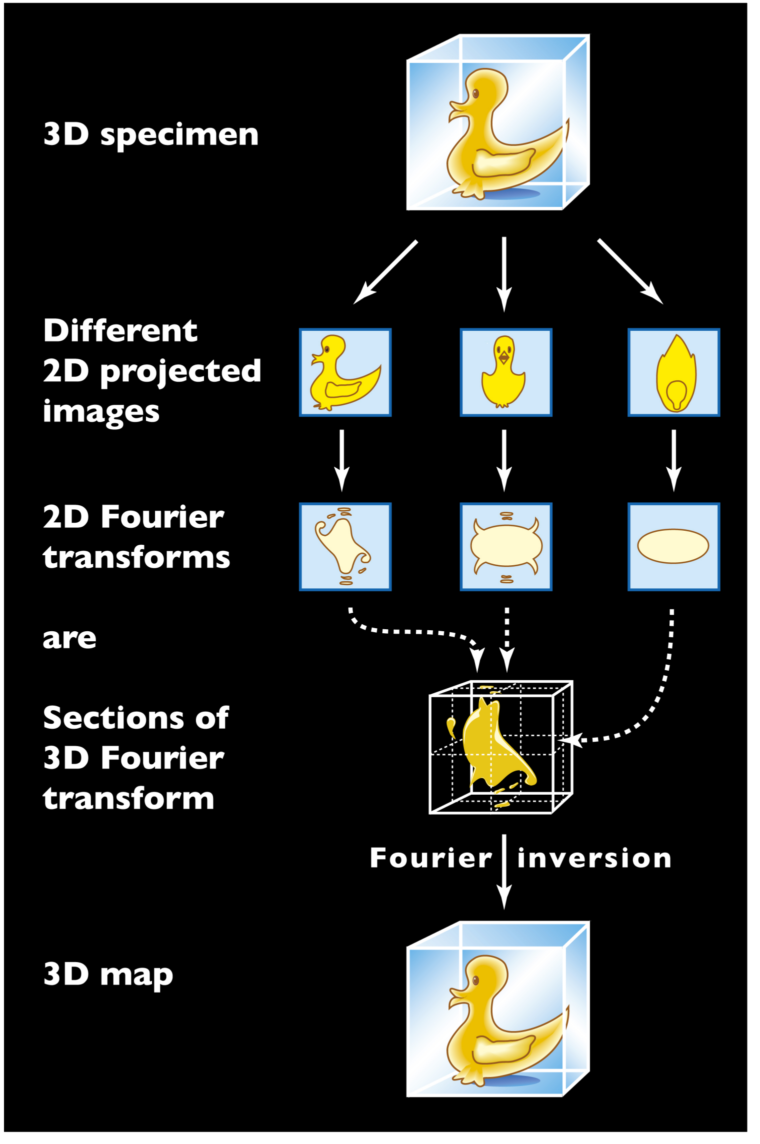
ratcheting,
EF-G, 1 tRNA

(Results coincided with a supervised classification)

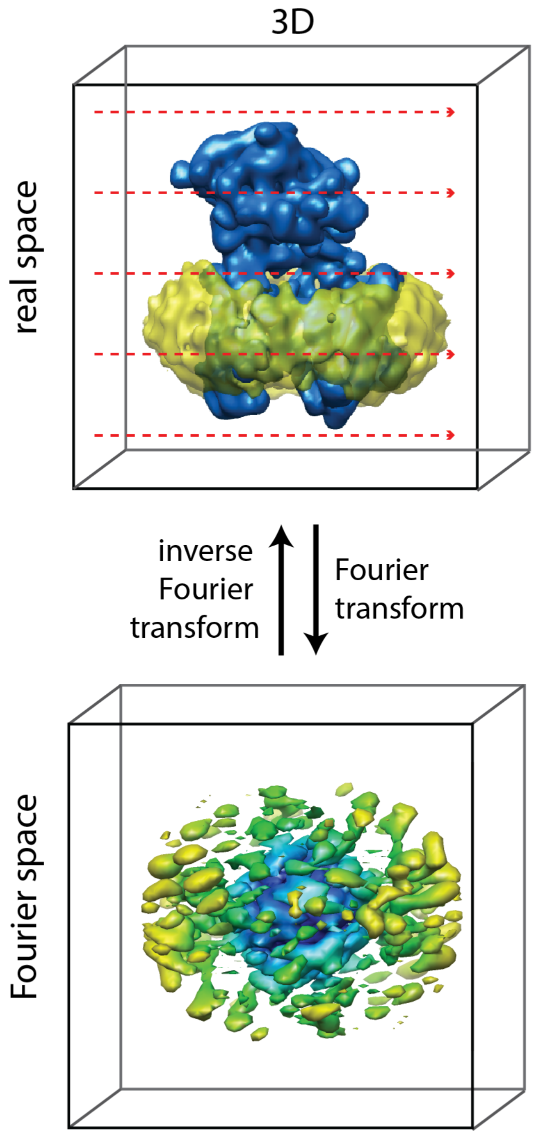
Scheres et al (2007) Nat. Meth.

Fourier-space formulation

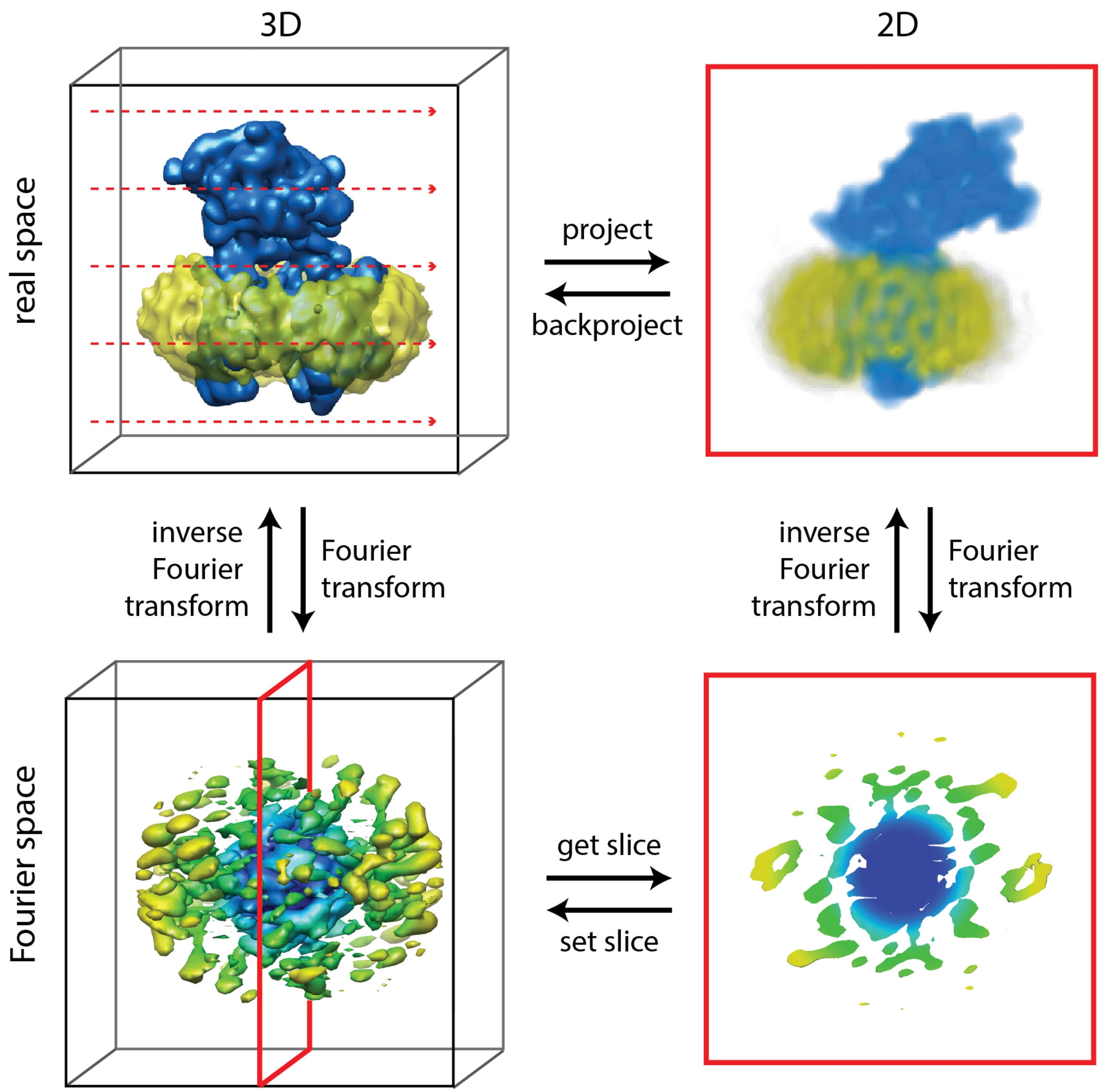
Projection-slice theorem



Projection-slice theorem



Projection slice theorem



Data model

- Real-space

$$X_i = \mathrm{CTF}_i \otimes \mathbf{P}_\varphi V_k + N_i$$

- Convolute w/ CTF
- $\mathbf{P}_\phi$ implements integrals
- N_i describes white noise

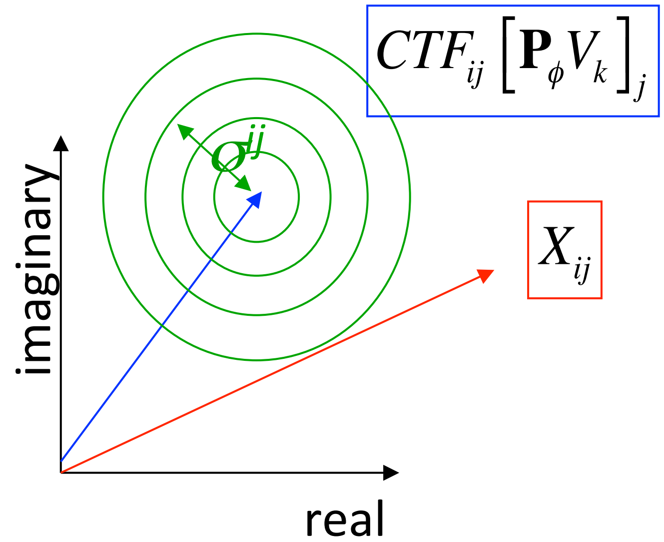
- Fourier space

$$X_i = \mathrm{CTF}_i \mathbf{P}_\varphi V_k + N_i$$

- Multiply w/ CTF
- $\mathbf{P}_\phi$ takes a slice
- N_i describes coloured noise

Coloured noise model

for each
Fourier pixel *h*
2D-Gaussian
in complex
plane



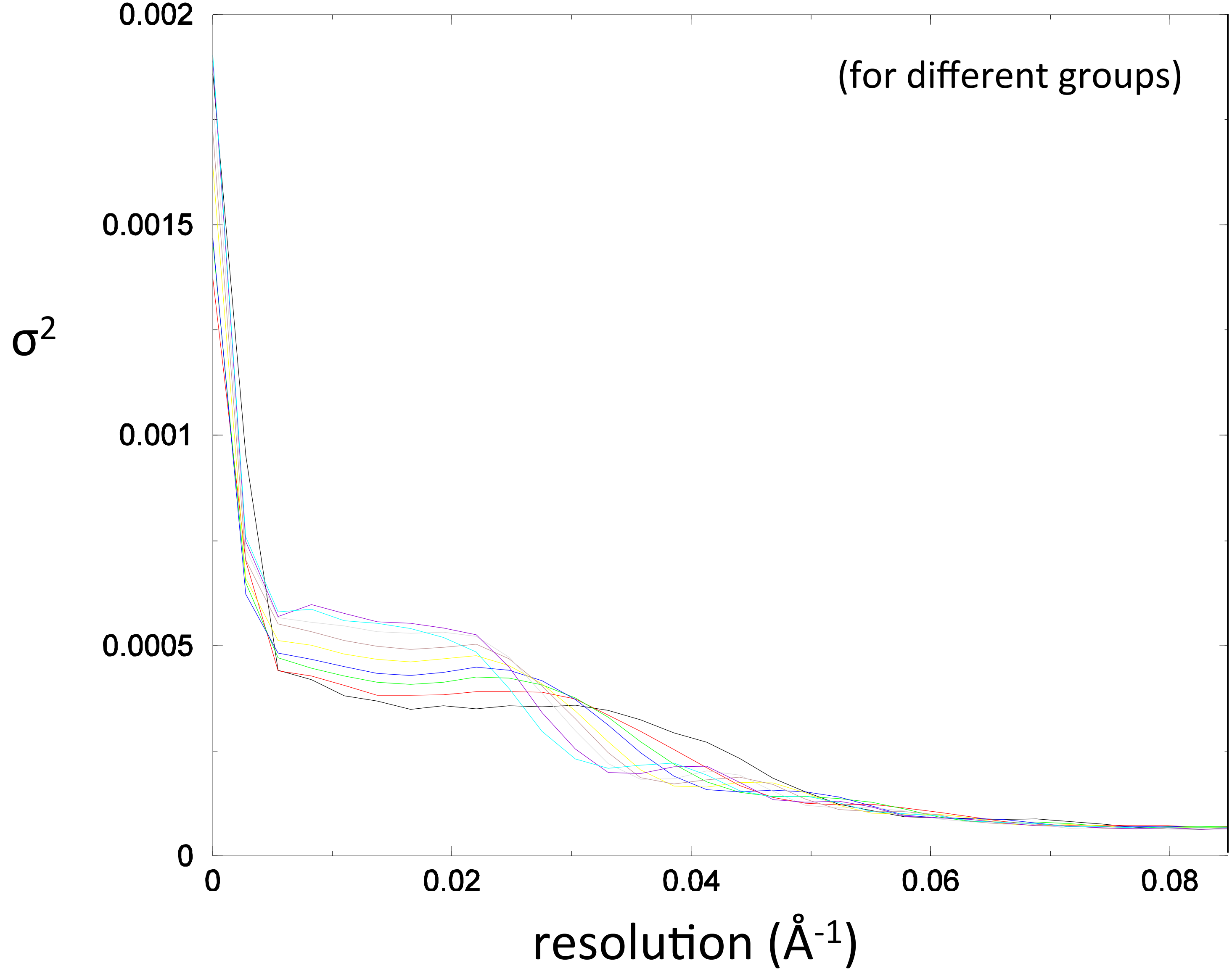
Assuming independence of noise between all Fourier terms:

$$P\left(X_i \mid k,\phi,\Theta\right)=\prod_{j=1}^{J}\frac{1}{2\pi\sigma_{ij}^2}\exp\left(\frac{\left|CTF_{ij}\left[\mathbf{P}_\phi V_k\right]_j - X_{ij}\right|^2}{-2\sigma_{ij}^2}\right)$$

resolution-dependent noise model!

Scheres *et al.* (2007) *Structure*

Coloured noise!!



‘Optimal’ filtering

- Paula: low, high and band-pass filters may be useful for denoising images for alignment

- By measuring power of noise and power of signal at every frequency, an ‘optimal’ filter is learnt automatically from the data
  - No user-expertise required to tune filters!

Regularised Likelihood

Maximum-likelihood estimators

- The best one can do…
- …in the limit of *infinitely large data sets*

- But my data set is limited in size, right?!
  - Even with Krios, K3 & EPU!

The bad news

- The experimental data alone is not enough to determine a unique solution!

- There are many noisy reconstructions that describe the data equally well…

- Danger of incorrect interpretation…

The good news

- By incorporating external information, a different problem may be solved for which a unique solution does exist!

- Regularisation

- Conventional regularisation approaches
  - Wiener filtering
  - Low-pass filtering

A Bayesian view on regularization

$$P(\Theta \mid X) = \frac{P(X \mid \Theta)P(\Theta)}{\cancel{P(X)}}$$

$$\text{Posterior} = \frac{\text{Likelihood * Prior}}{\text{Evidence}}$$

Regularised likelihood optimisation

Likelihood

- Assume noise is Gaussian and independent
  - in Fourier space
  - with spectral power $\sigma^2(\upsilon)$: *coloured noise*

$$P(X_i \mid k, \phi, \Theta) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}} \exp\left(\frac{\left\| X_{ij} - \mathrm{CTF}_{ij}(\mathbf{P}_\phi V_k)_j \right\|^2}{-2\sigma_{ij}^2} \right)$$

Prior

- Assume signal is Gaussian and independent
  - in Fourier space
  - Limited power $\tau^2(\upsilon)$: *smoothness in real space!*

$$P(\Theta) = \prod_l \frac{1}{2\pi\tau_{kl}} \exp\left\{ \frac{\left\| V_{kl} \right\|^2}{-2\tau_{kl}^2} \right\}$$

Expectation maximization

$$V^{(n+1)} = \frac{\sum_{i=1}^{N} \int_{\phi} \Gamma_{i\phi}^{(n)} \mathbf{P}_{\phi}^{\mathrm{T}} \frac{\mathrm{CTF}_i}{\sigma_i^{2(n)}} X_i d\phi}{\sum_{i=1}^{N} \int_{\phi} \Gamma_{i\phi}^{(n)} \mathbf{P}_{\phi}^{\mathrm{T}} \frac{\mathrm{CTF}_i^2}{\sigma_i^{2(n)}} d\phi + \frac{1}{\tau^{2(n)}}}$$

Wiener (optimal) filter for CTF-corrected 3D reconstruction / 2D class averages

$$\sigma_i^{2(n+1)} = \frac{1}{2} \int_{\phi} \Gamma_{i\phi}^{(n)} \left\| X_i - \mathrm{CTF}_i \mathbf{P}_{\phi} V^{(n)} \right\|^2 d\phi$$

Estimate resolution-dependent power of noise from the data

$$\tau^{2(n+1)} = \frac{1}{2} \left\| V^{(n)} \right\|^2$$

Estimate resolution-dependent power of signal from the data

$$\Gamma_{i\phi}^{(n)} = \frac{P(X_i \mid \phi, \Theta^{(n)}) P(\phi \mid \Theta^{(n)})}{\int_{\phi'} P(X_i \mid \phi', \Theta^{(n)}) P(\phi' \mid \Theta^{(n)}) d\phi'}$$

3D Wiener filter

$$V^{(n+1)} = \frac{\displaystyle\sum_{i=1}^{N} \int_{\phi} \Gamma_{i\phi}^{(n)} \mathbf{P}_{\phi}^{\mathrm{T}} \frac{\mathrm{CTF}_i}{\sigma_i^{2(n)}} X_i \, d\phi}{\displaystyle\sum_{i=1}^{N} \int_{\phi} \Gamma_{i\phi}^{(n)} \mathbf{P}_{\phi}^{\mathrm{T}} \frac{\mathrm{CTF}_i^2}{\sigma_i^{2(n)}} \, d\phi + \frac{1}{\tau^{2(n)}}}$$

- Calculates SSNR(υ) (as a 3D function)
- Handles uneven orientational distribution
- Handles astigmatic CTFs & CTF en
- Corrects CTF & low-pass
- *Optimal linear filter*

WITHOUT ARBITRARINESS!

Recapitulating

- Alignment & classification are incomplete problems
  - Best dealt with by marginalisation (ML)

- 2D and 3D problems are very similar

- Fourier-space is most convenient
  - CTF multiplication
  - Slices instead of line integral projections
  - Coloured noise-model
  - Regularised Likelihood function -> ‘optimal’ filters

Further Reading

- Penczek, Fundamentals of Three-Dimensional Reconstruction from Projections, *Methods in Enzymology*, , **482** (2010) p 1
- Penczek, Image restoration in cryo-electron microscopy, *Methods in Enzymology*, , **482** (2010) p 35
- Sigworth, Doerschuk, Carazo & Scheres, An Introduction to Maximum-Likelihood Methods in Cryo-EM, *Methods in Enzymology*, **482** (2010) p 263
- Scheres, Classification of Structural Heterogeneity by Maximum-Likelihood Methods, *Methods in Enzymology*, **482** (2010) p 295
- Scheres, Processing of Structurally Heterogeneous Cryo-EM Data in RELION, *Methods in Enzymology*, **579** (2016) p 125
- www2.mrc-lmb.cam.ac.uk/relion (tutorial & Wiki pages)