Bayesian methods; particle classification

Sjors H.W. Scheres

EMBO course 2019
Birkbeck College, London

MRC Laboratory of Molecular Biology

Agenda

- An intuitive introduction

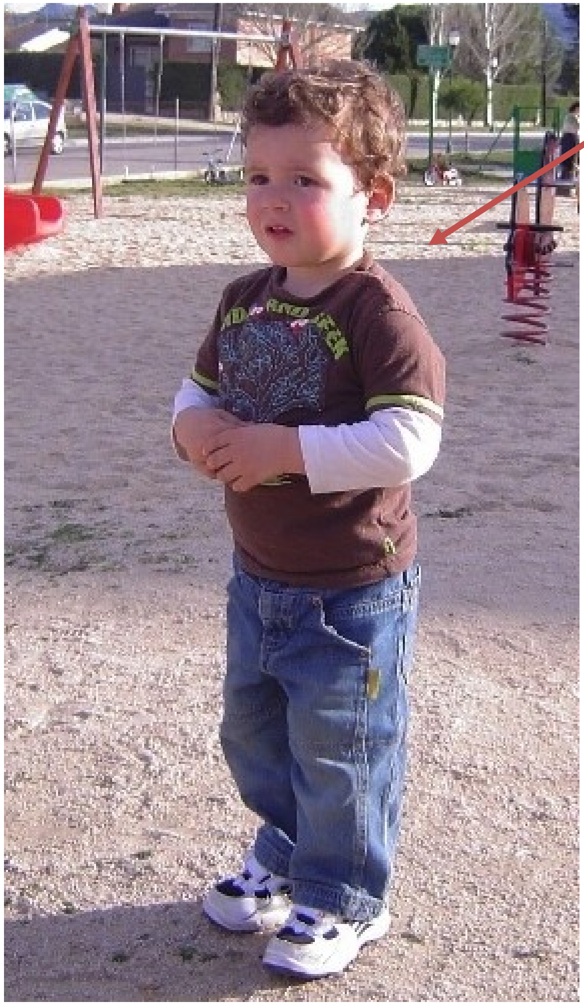
- Alignment
  - Dealing with the incomplete problem
  - maxCC vs ML (real-space)

- Classification
  - Multi-reference alignment in 2D
  - and in 3D

- Fourier-space formulation
  - Regularised likelihood optimisation (Bayesian approach)

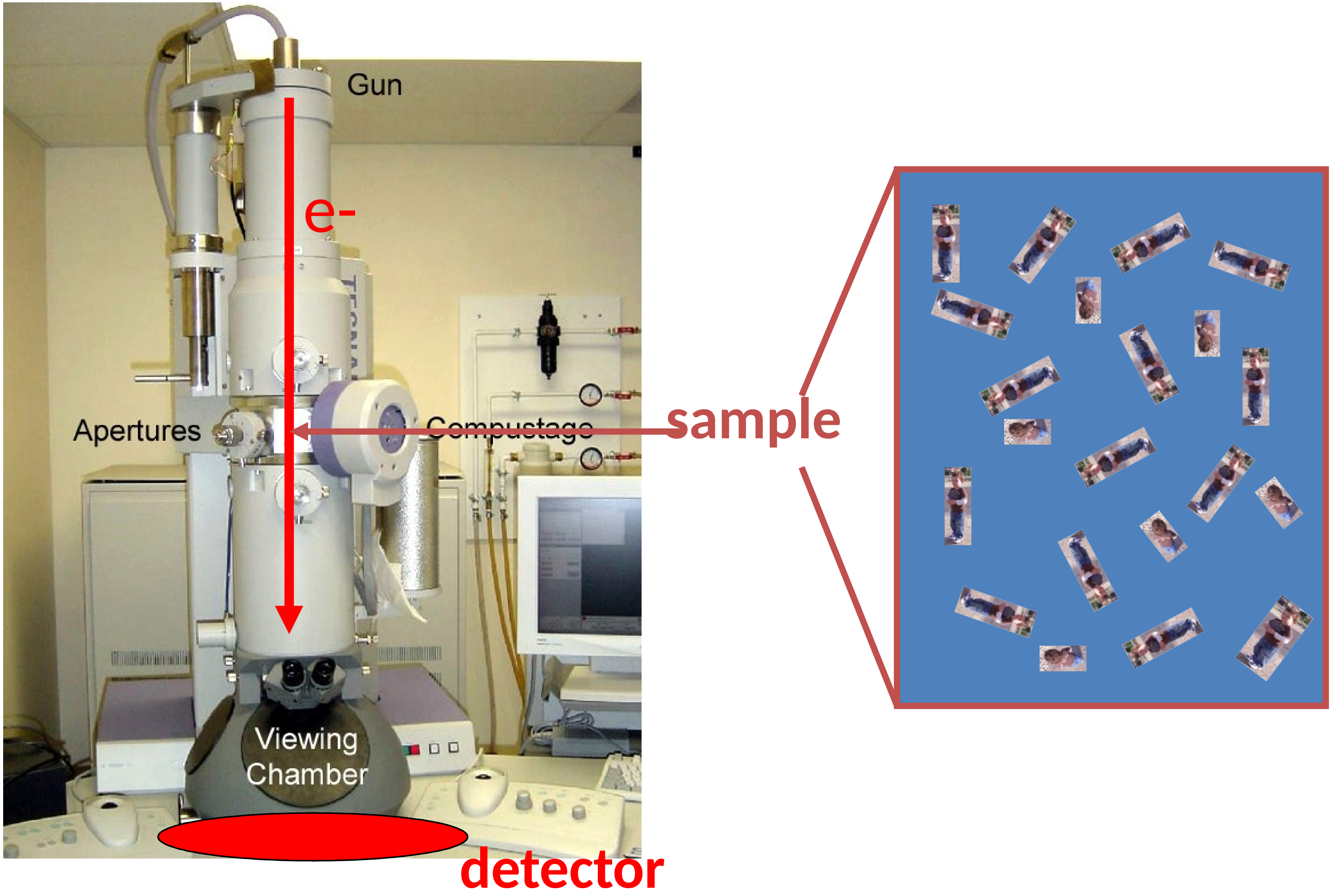
An intuitive introduction

An example “protein”

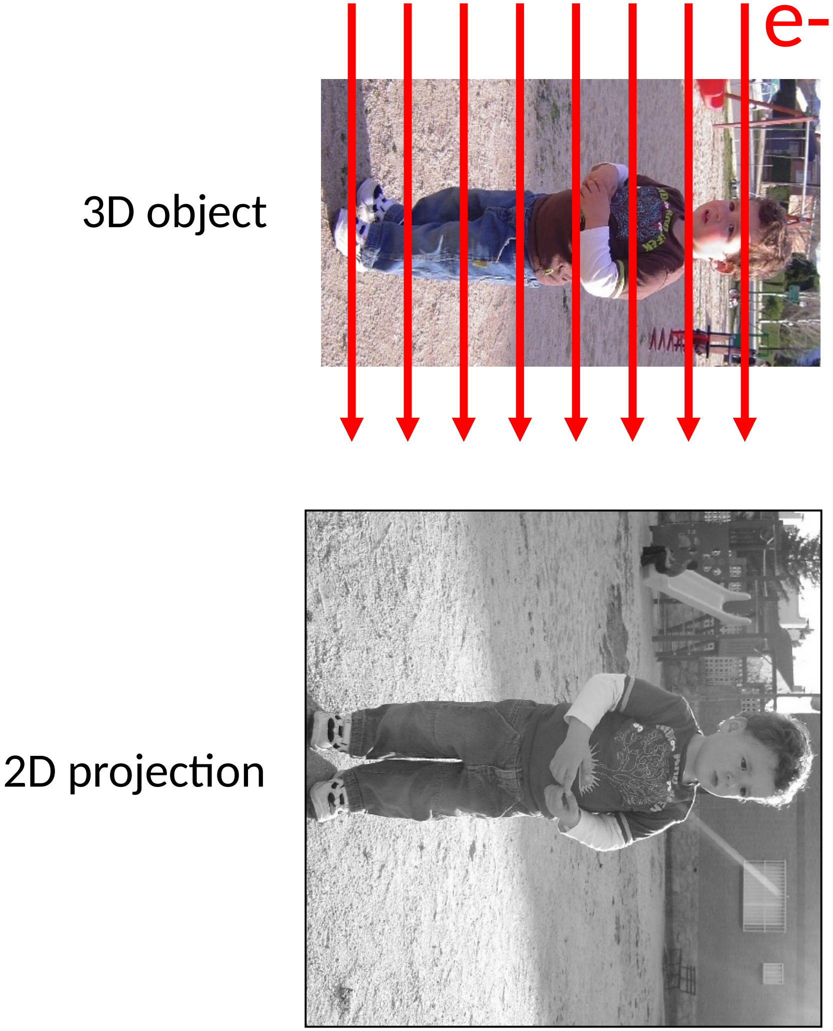


Jan

Experimental setup



Electron microscopy imaging

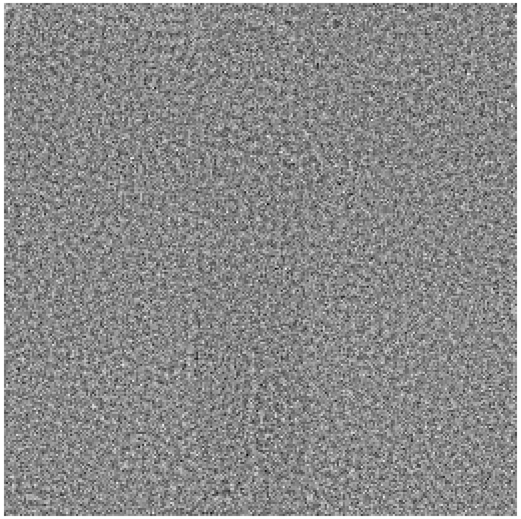


We collect data in 2D,

but we want 3D info!

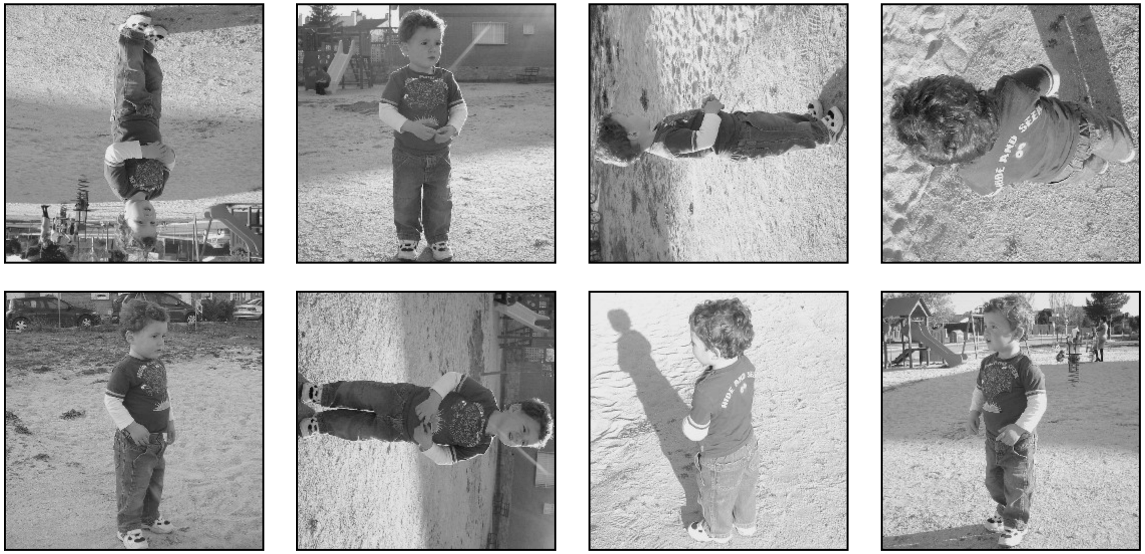
Further inconveniences

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



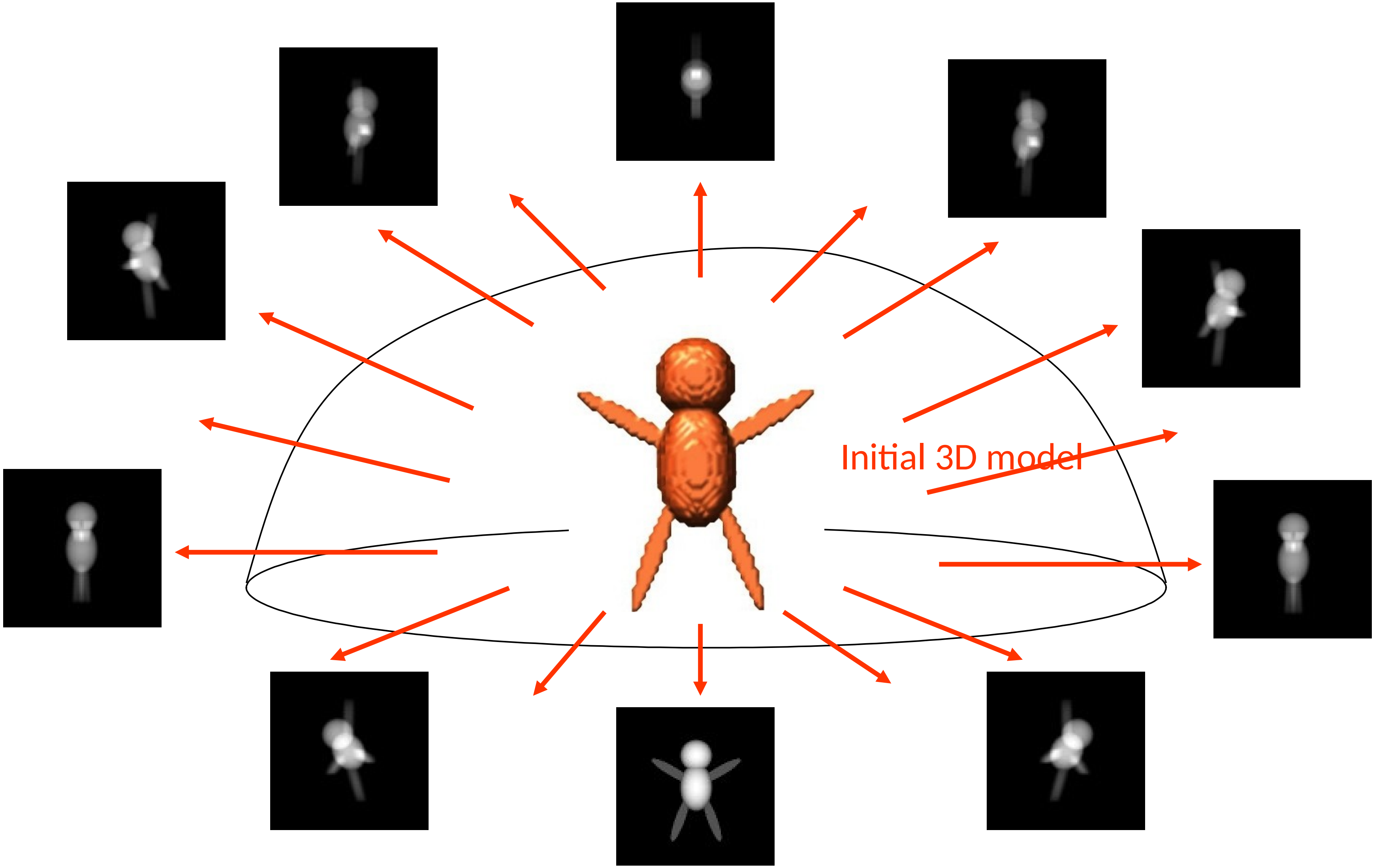
Single particle analysis

- Embedded in ice: many unknown orientations

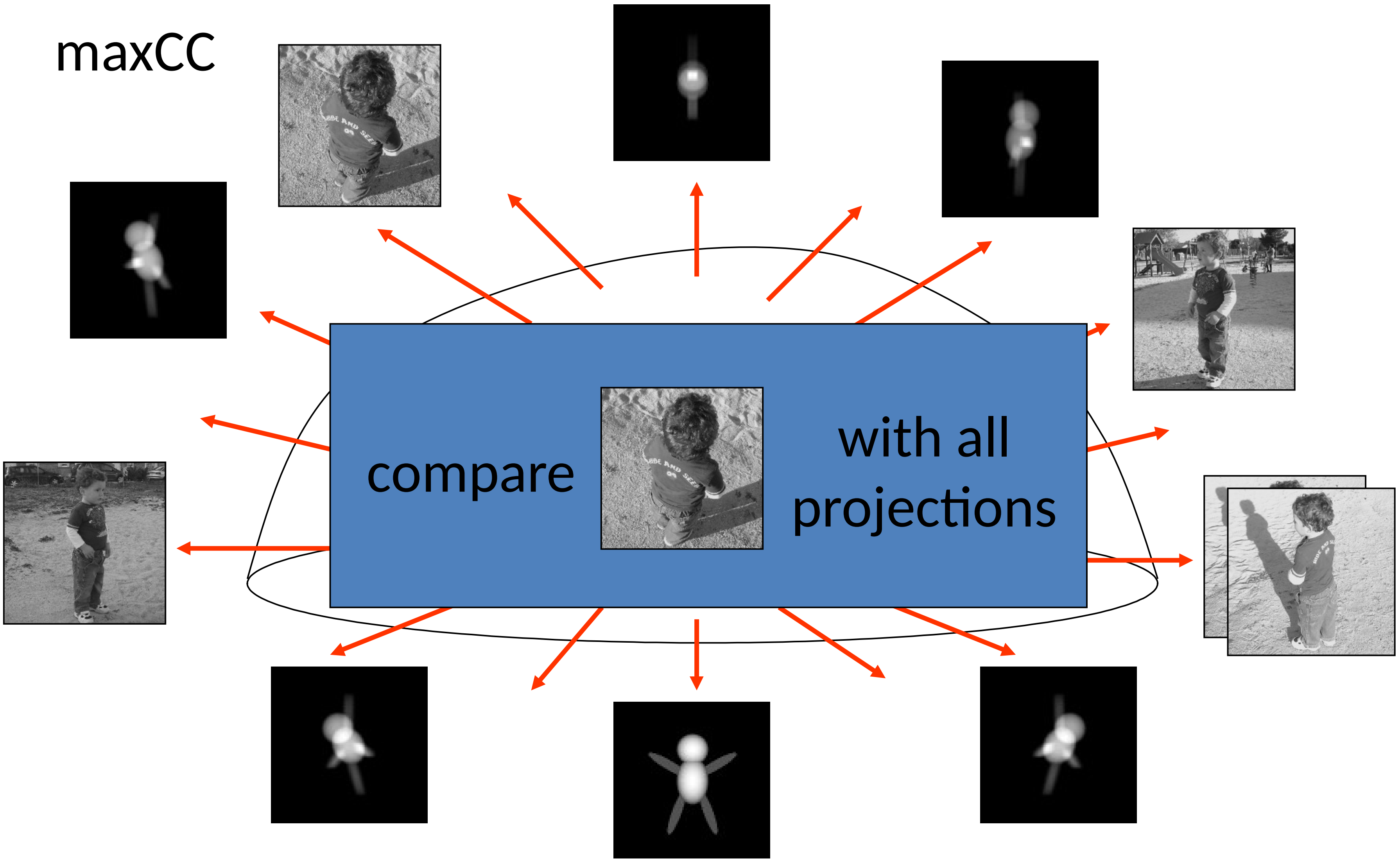


- Combine all 2D projections into a 3D reconstruction

Projection matching



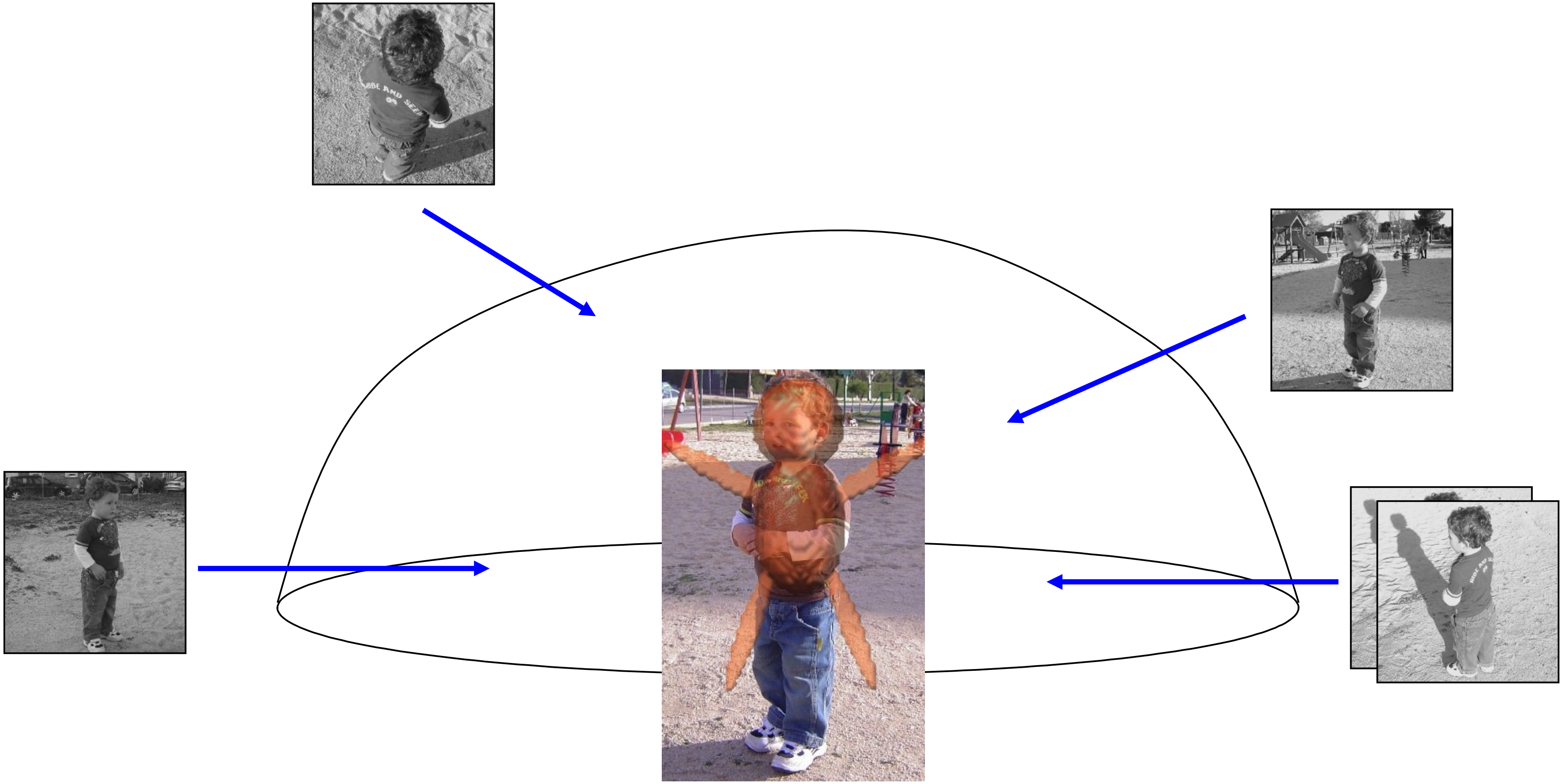
Projection matching



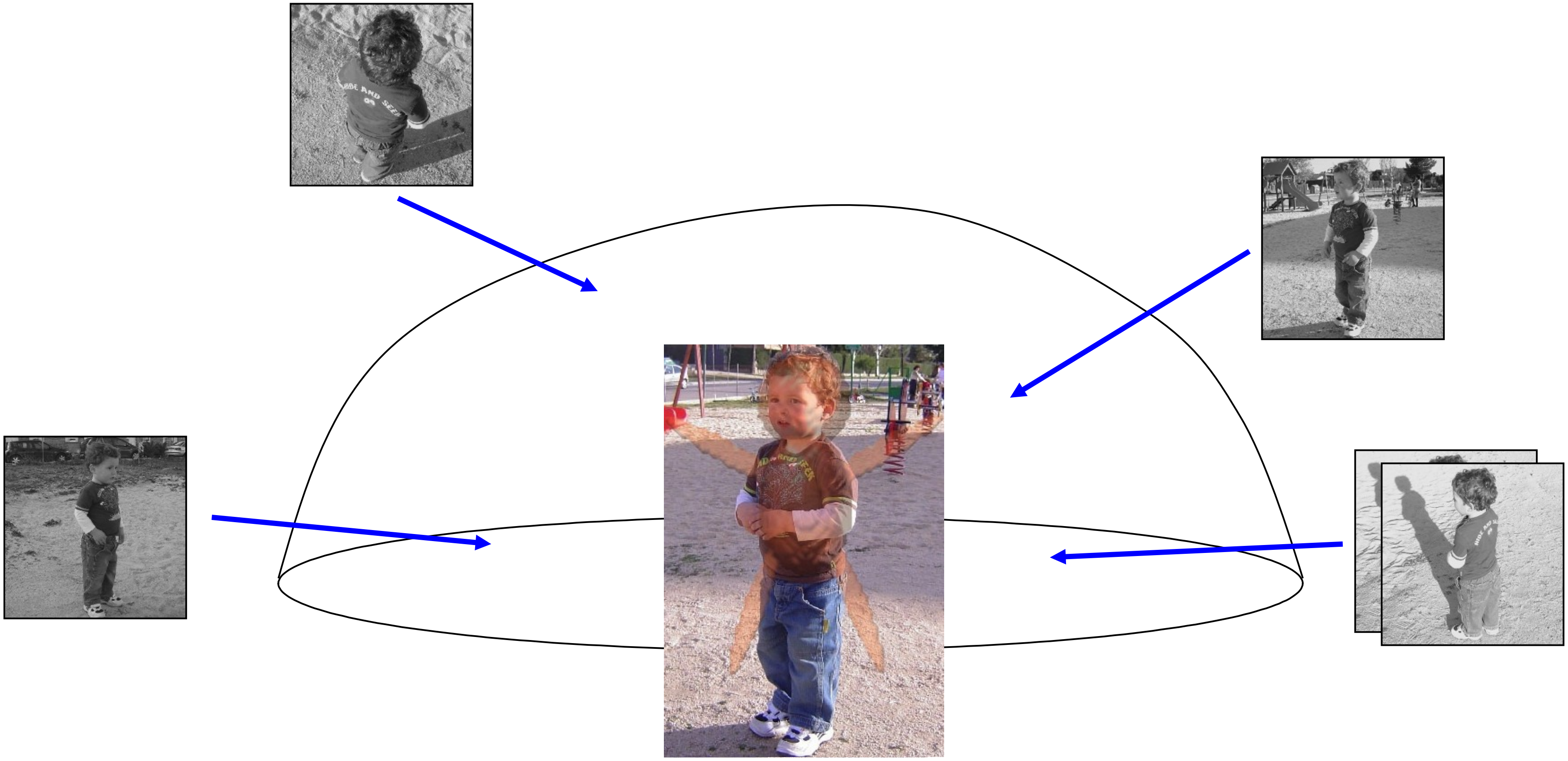
maxCC

compare with all projections

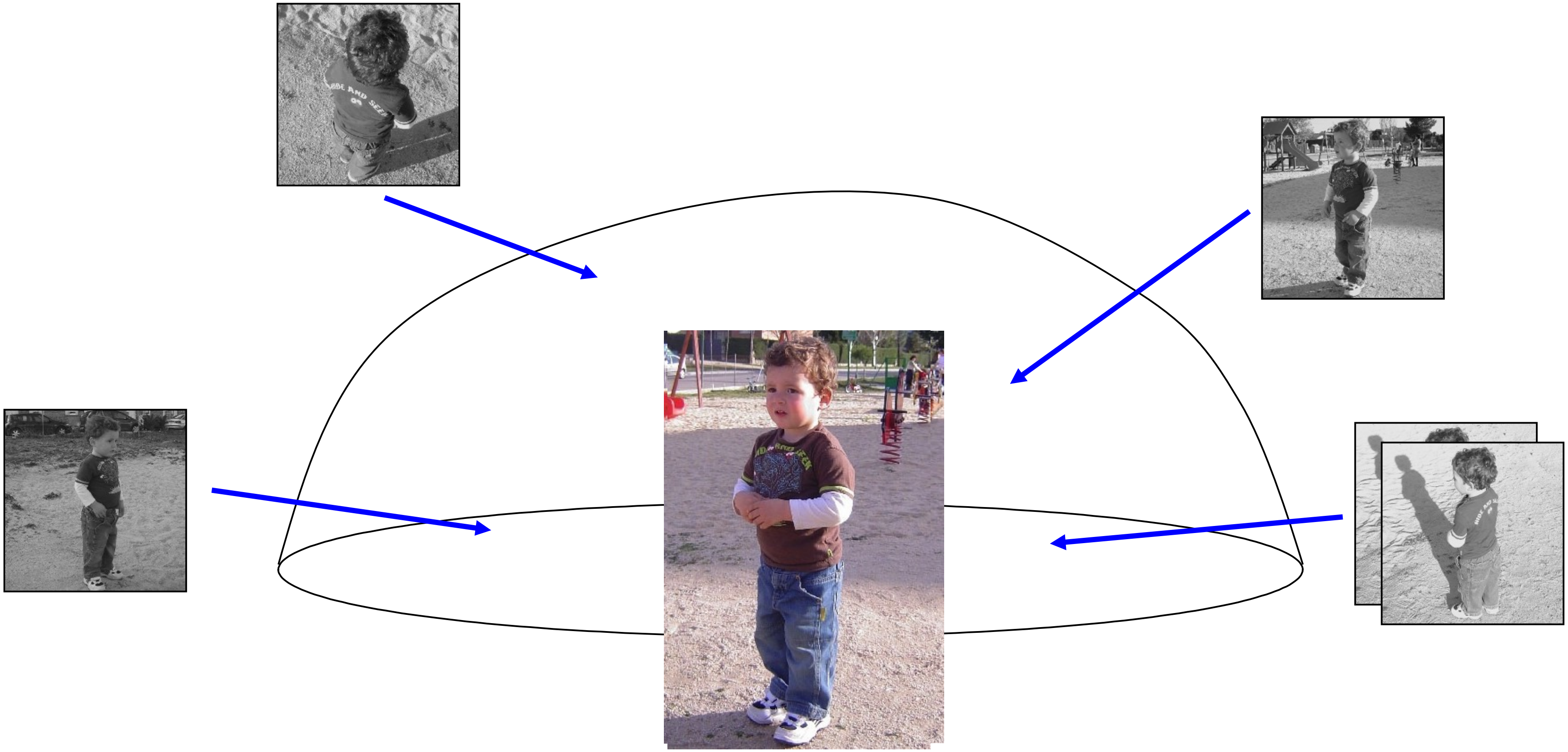
3D reconstruction



Iterative refinement



Iterative refinement



Alignment

Or how to ‘match’ projections

Incomplete data problems

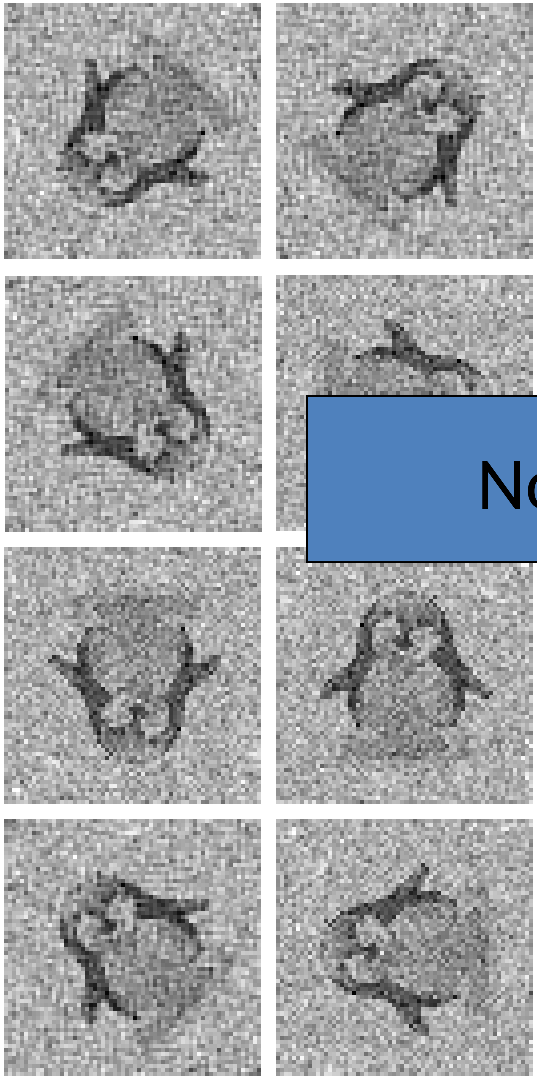
- Part of the data was not observed experimentally
  - Orientations
  - Class assignments

- 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

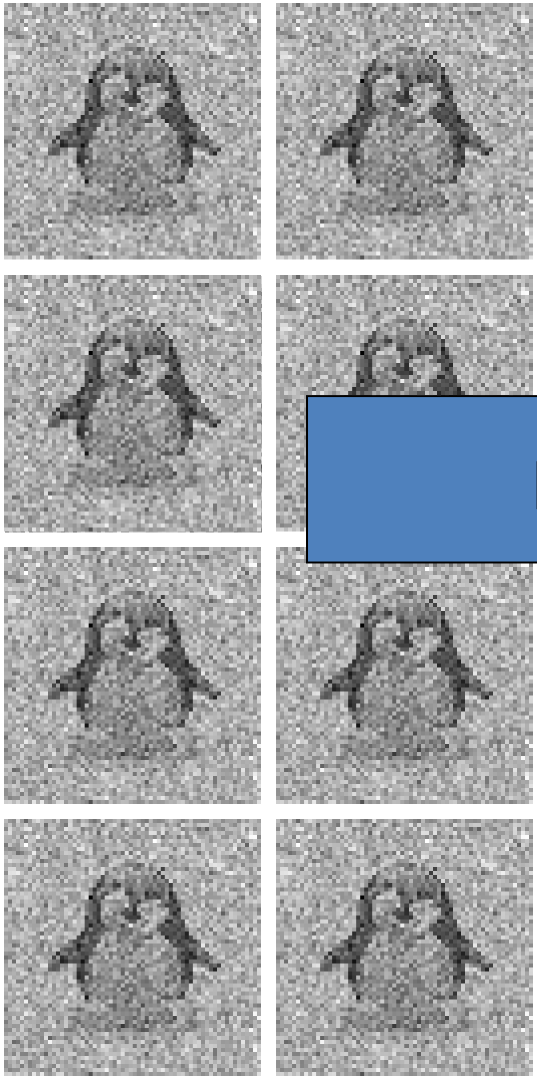


Not easy

Observed data (*X*): images

Missing data (Y): orientations

Complete data problems



Easy!

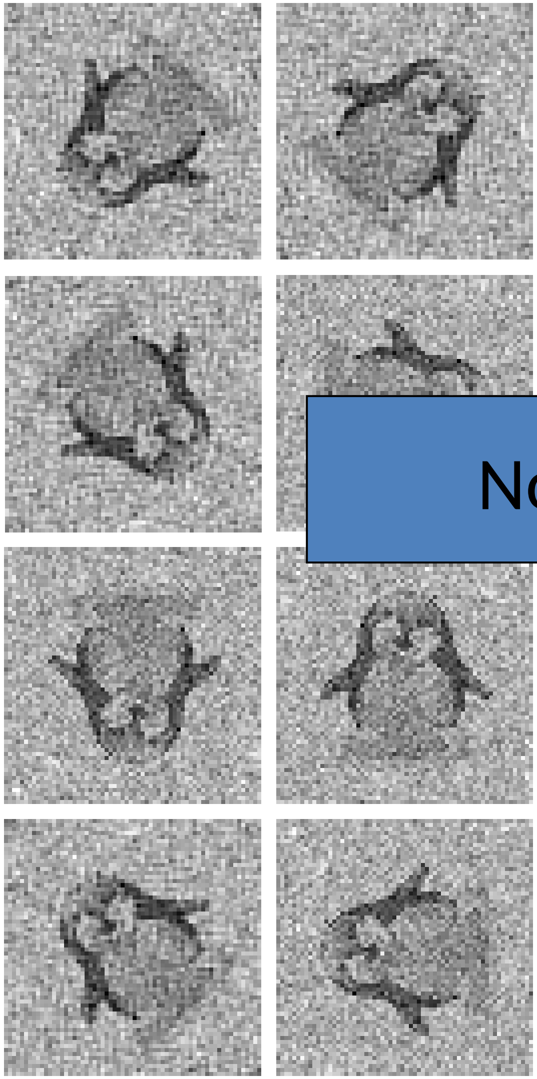
white Gaussian noise

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

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

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

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

- Option 2: marginalize over *Y* →

Maximum Likelihood

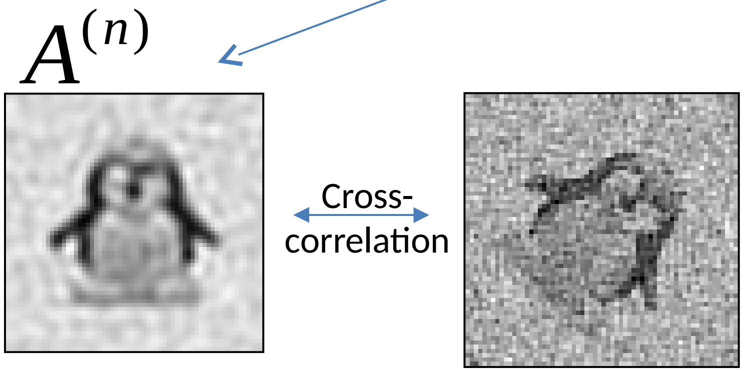
$$L(\Theta) = P(X \mid \Theta) = \int_Y P\left(X \mid Y,\Theta\right) P\left(Y \mid \Theta\right) d\varphi$$

Probability of X, regardless Y

The maxCC approach

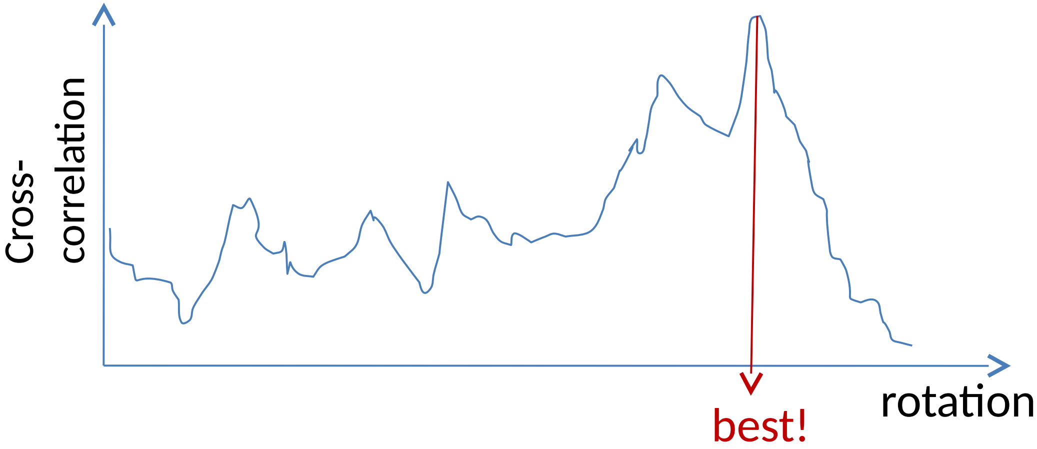
Reference-based alignment

- Starts from some initial guess about the structure

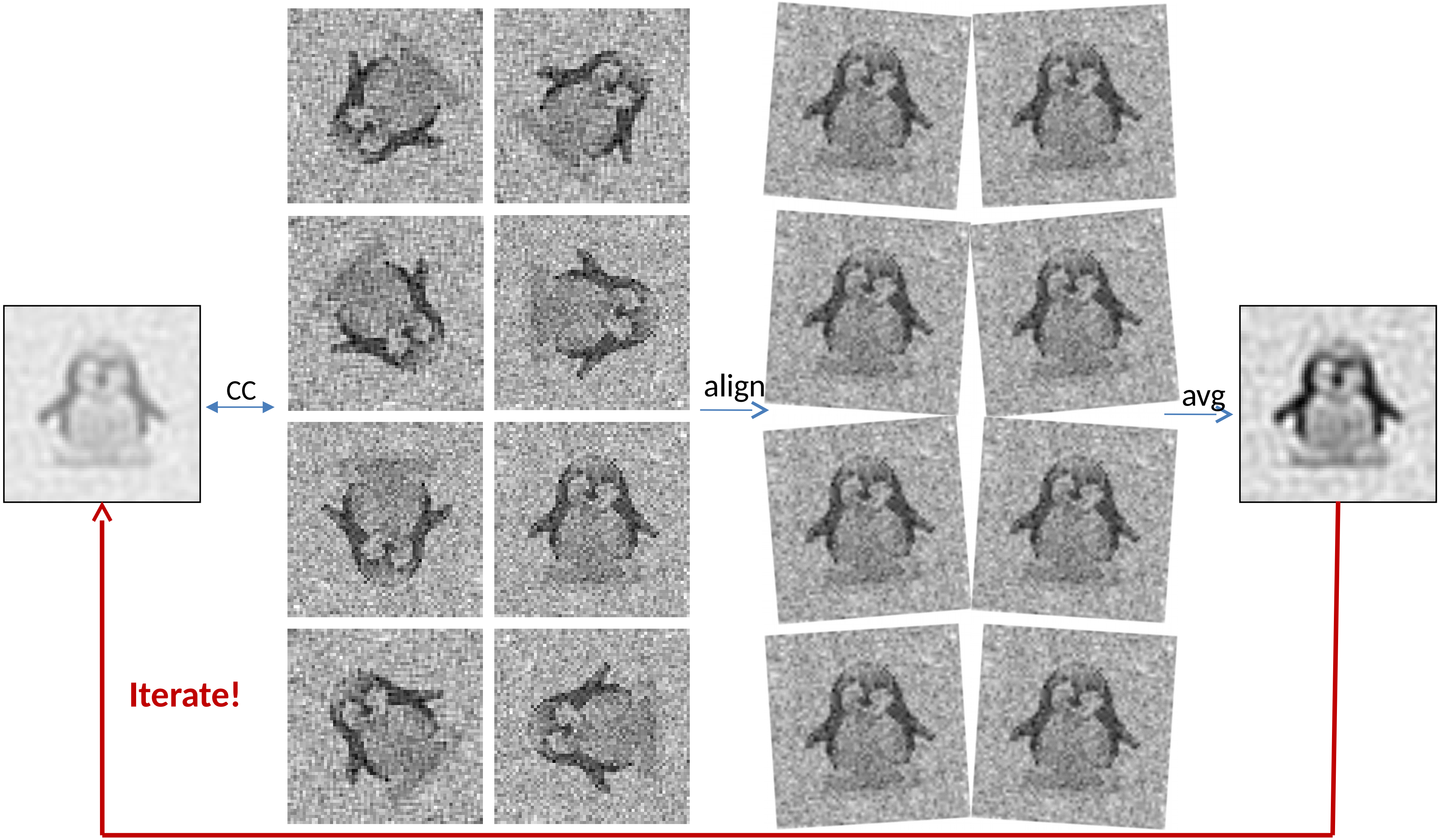


Compare initial guess with each experimental particle

Illustrate CCF on the board



Align and average



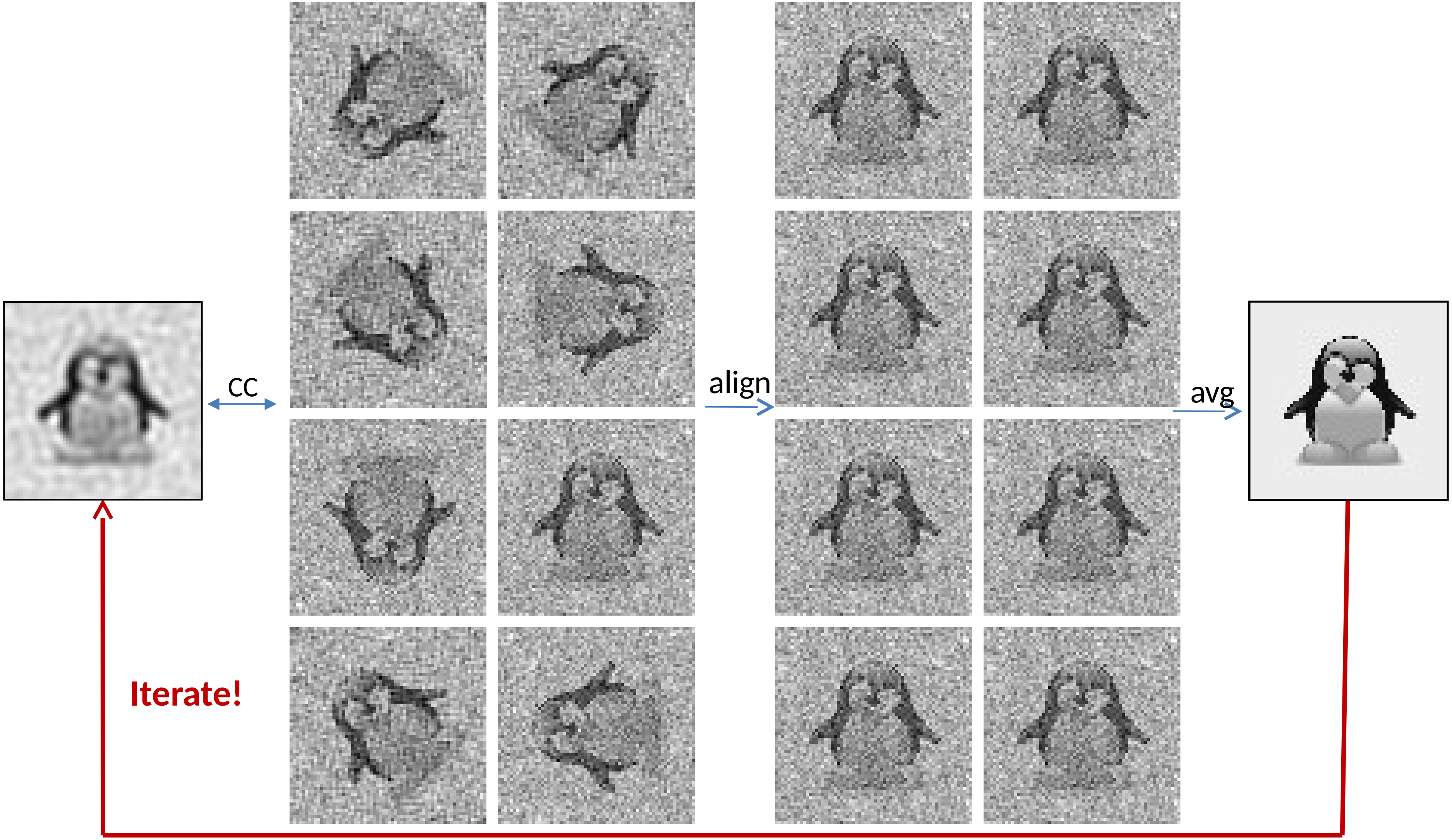
CC

align

avg

Iterate!

Align and average



cc

align

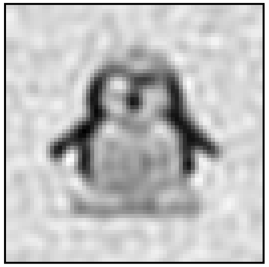
avg

Iterate!

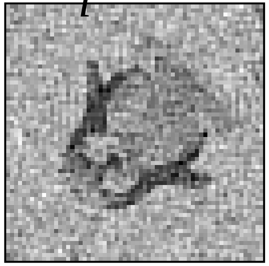
The ML approach

Maximum likelihood

$A^{(n)}$



X_i



Statistical model

$$P(X_i \mid \varphi, \Theta)$$

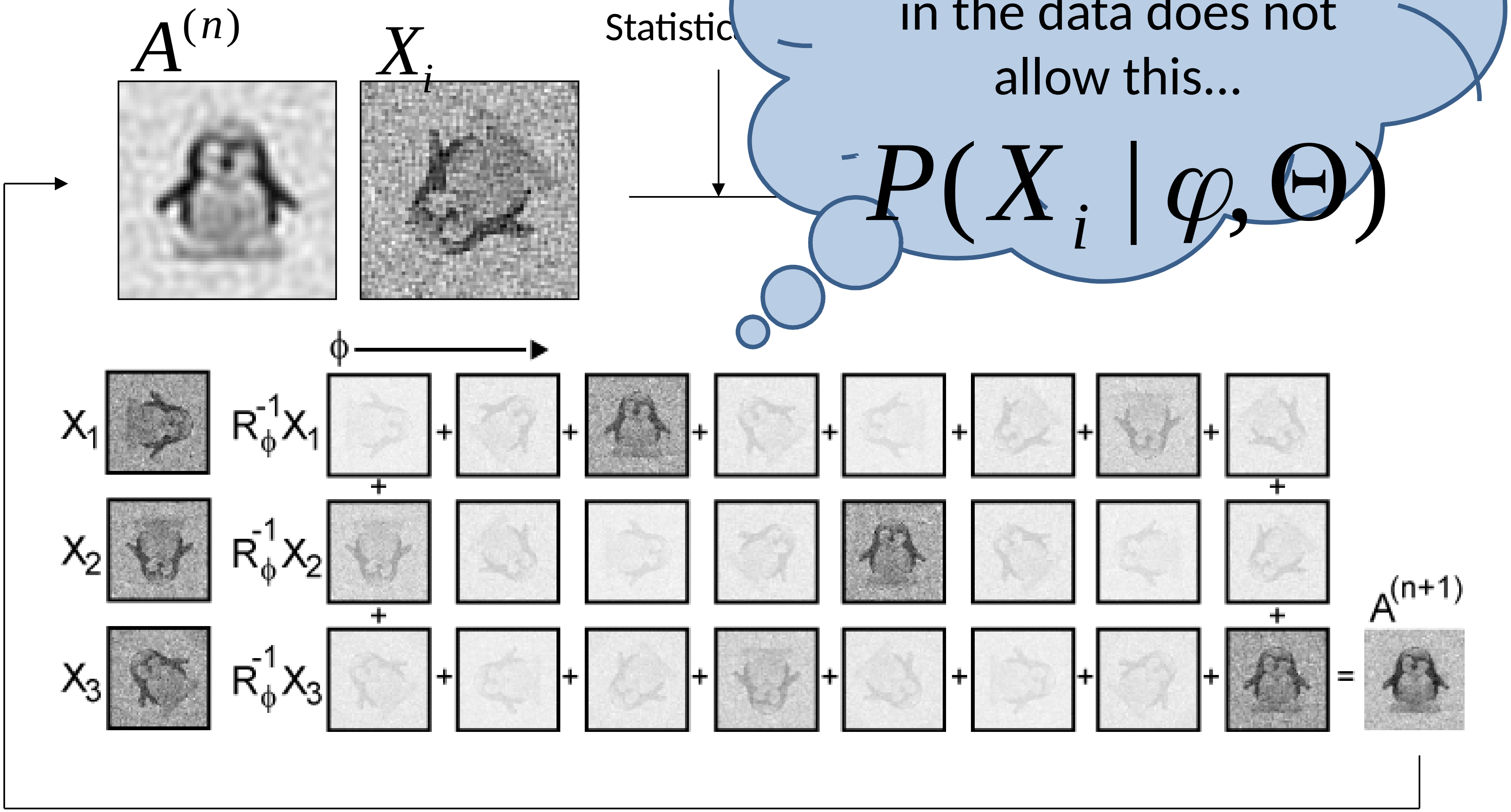
Based on Gaussian error model

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

Maximum li

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

$$P(X_i \mid \varphi, \Theta)$$



Sigworth, J. Struct. Biol., 1998

Incomplete data problems

- Option 1: add *Y* to the model

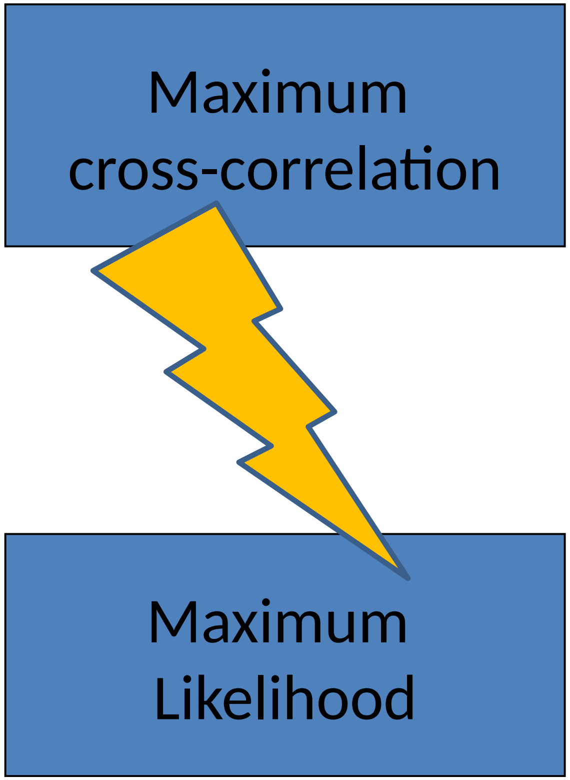
$$L(Y,\Theta) = P(X \mid Y,\Theta)$$

- Option 2: marginalize over *Y*

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

↓

Probability of X,
regardless Y



Maximum cross-correlation

Maximum Likelihood

Incomplete data problems

Maximum cross-correlation

Maximum Likelihood

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

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

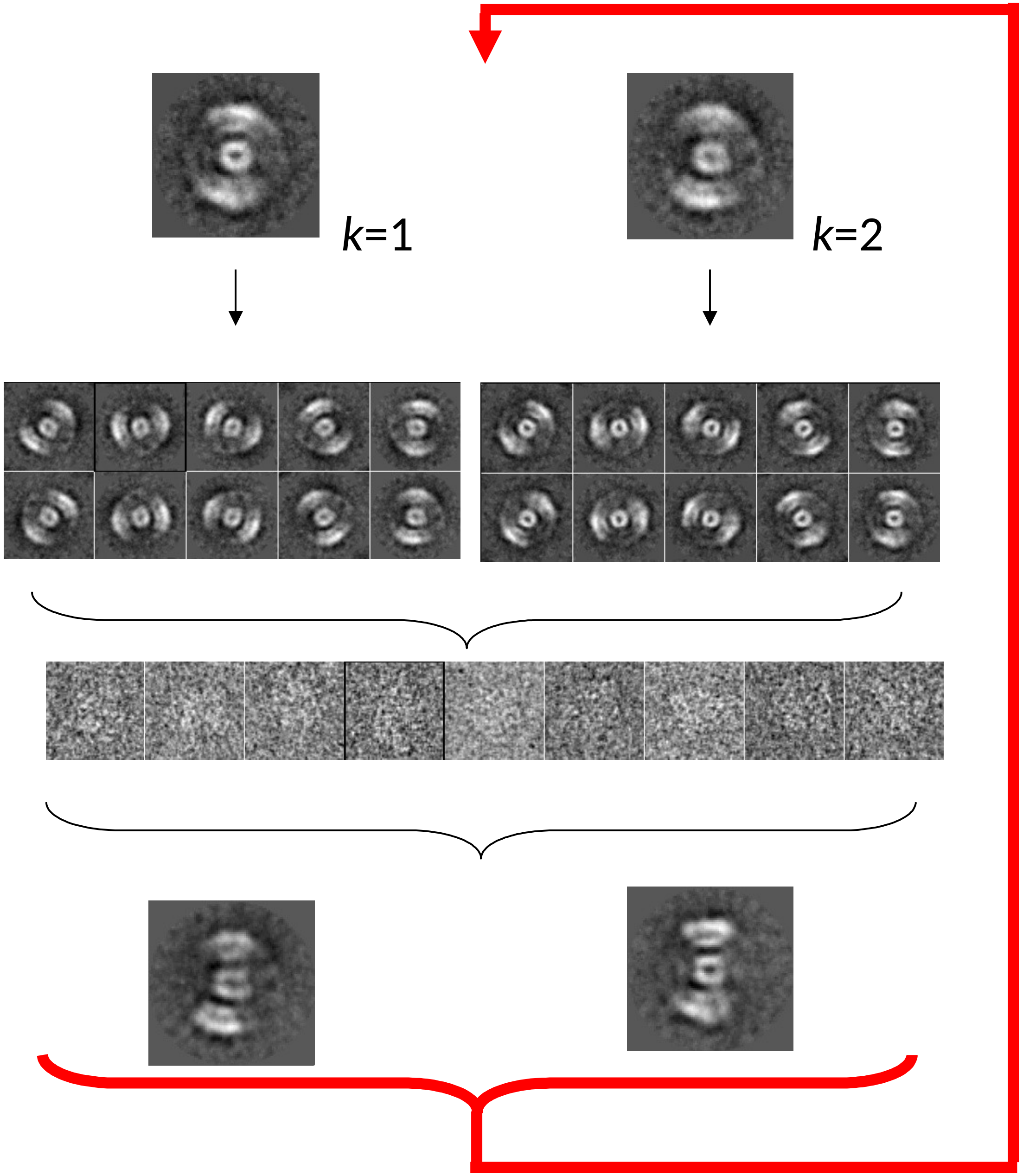
estimates for *K*
2D objects

sampled rotations 360°

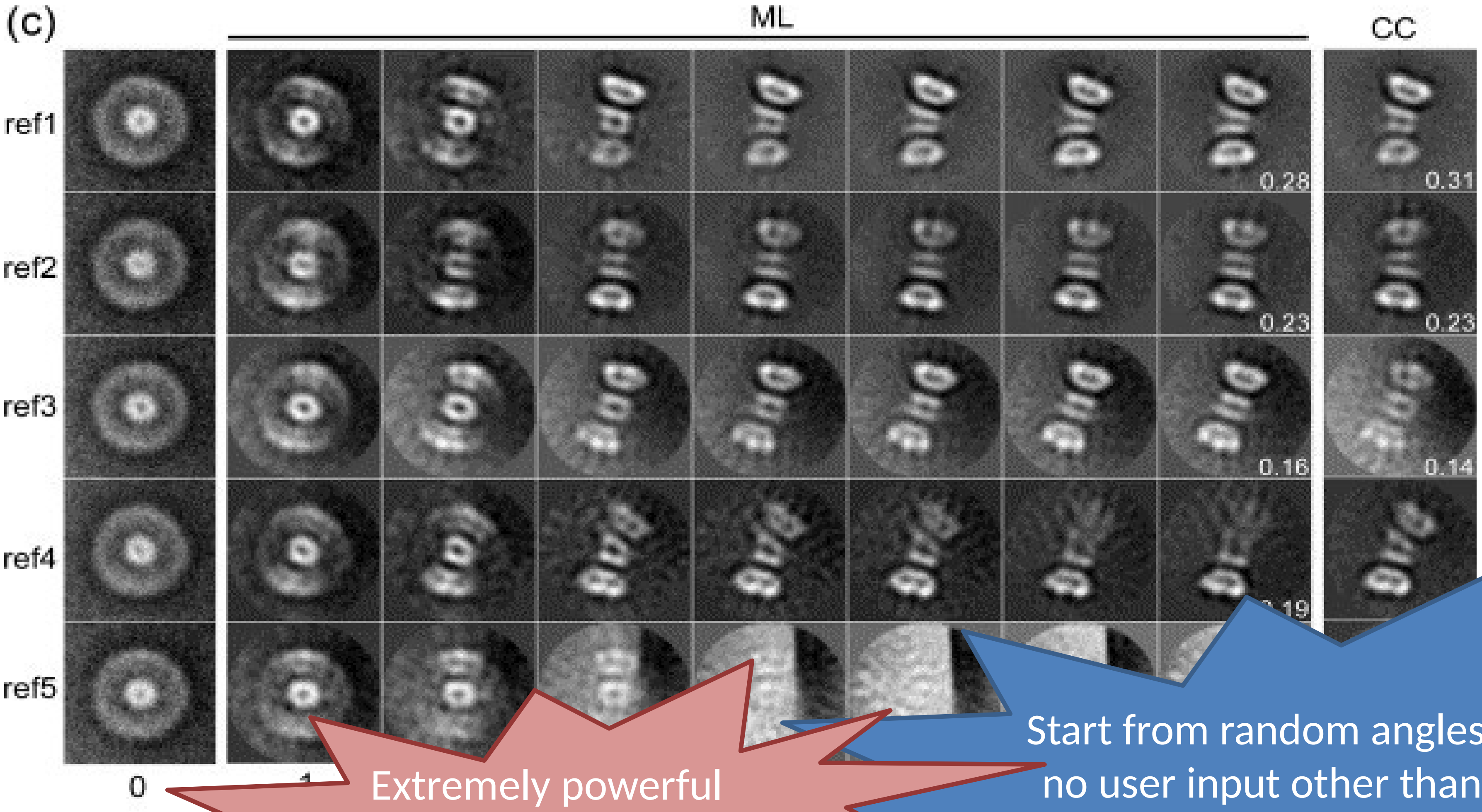
for each image, calculate all

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

calculate new 2D average
as *probability weighted*
averages



Reference-free 2D class averaging



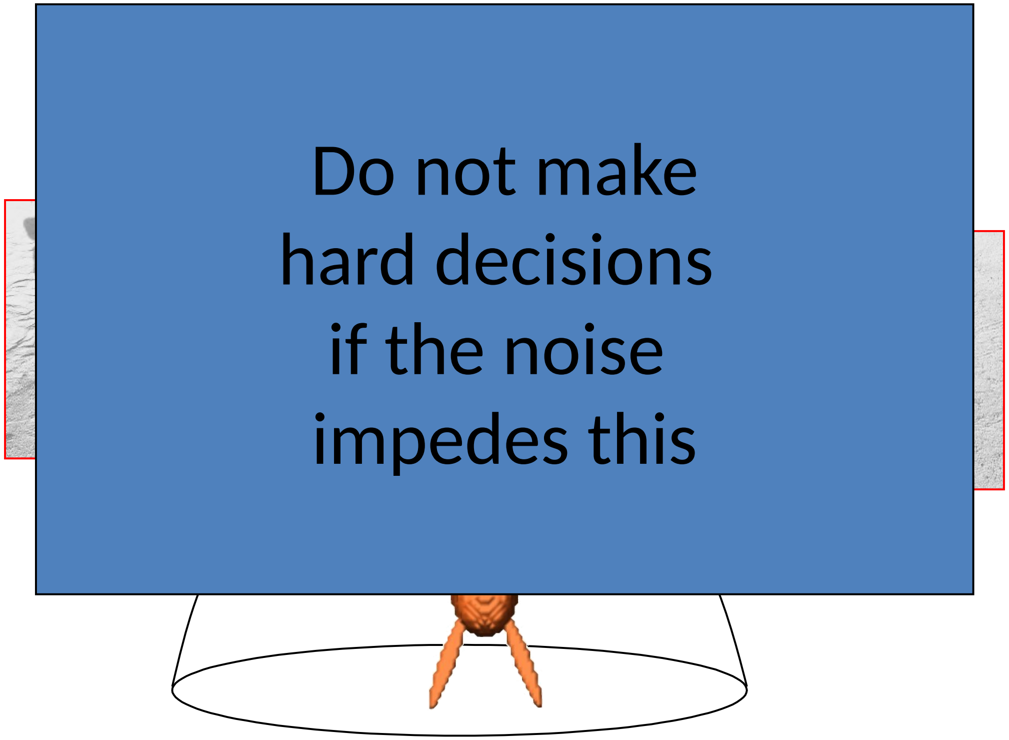
Extremely powerful to clean & assess your data

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

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

3D alignment & classification

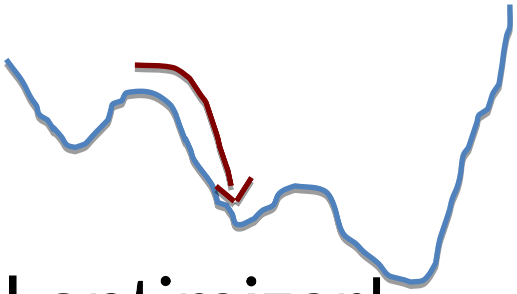
3D ML refinement



Do not make hard decisions if the noise impedes this

“Probability-weighted angular assignment”

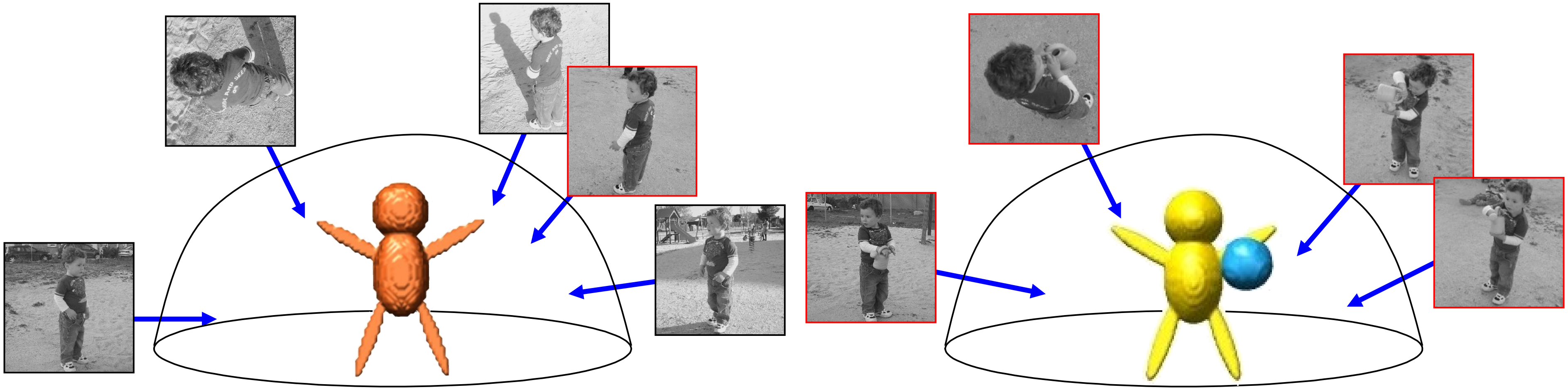
Initial model



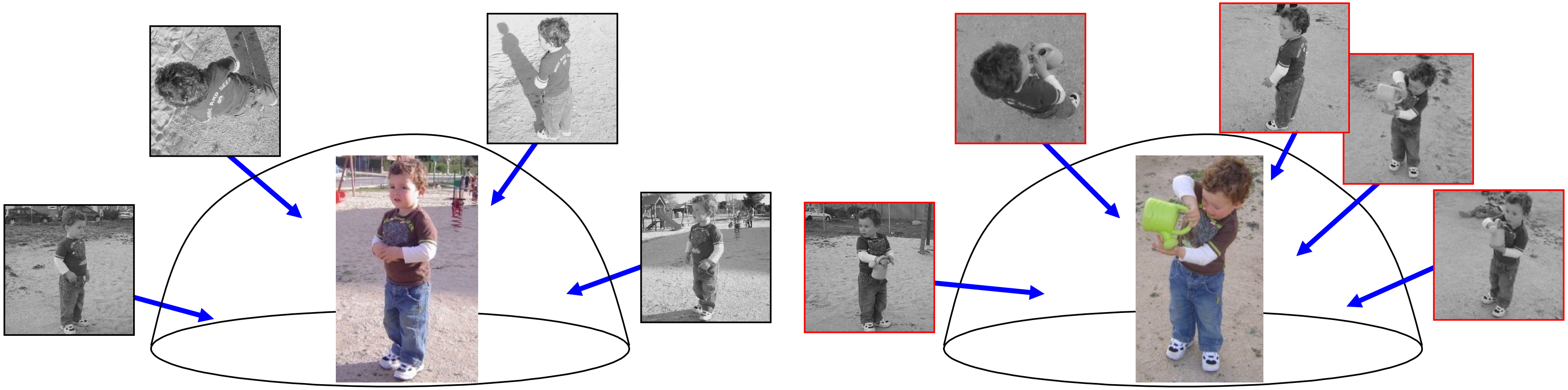
- Expectation-Maximisation is a local optimizer!
  - Gets stuck in nearest (local) minimum
- Bad model in -> bad model out!!!
  - Much less of a problem with high-resolution data
- Stochastic methods may reach global minimum
  - Stochastic Hill Climbing (SIMPLE)
  - Stochastic Gradient Descent (cryoSPARC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

Structural heterogeneity

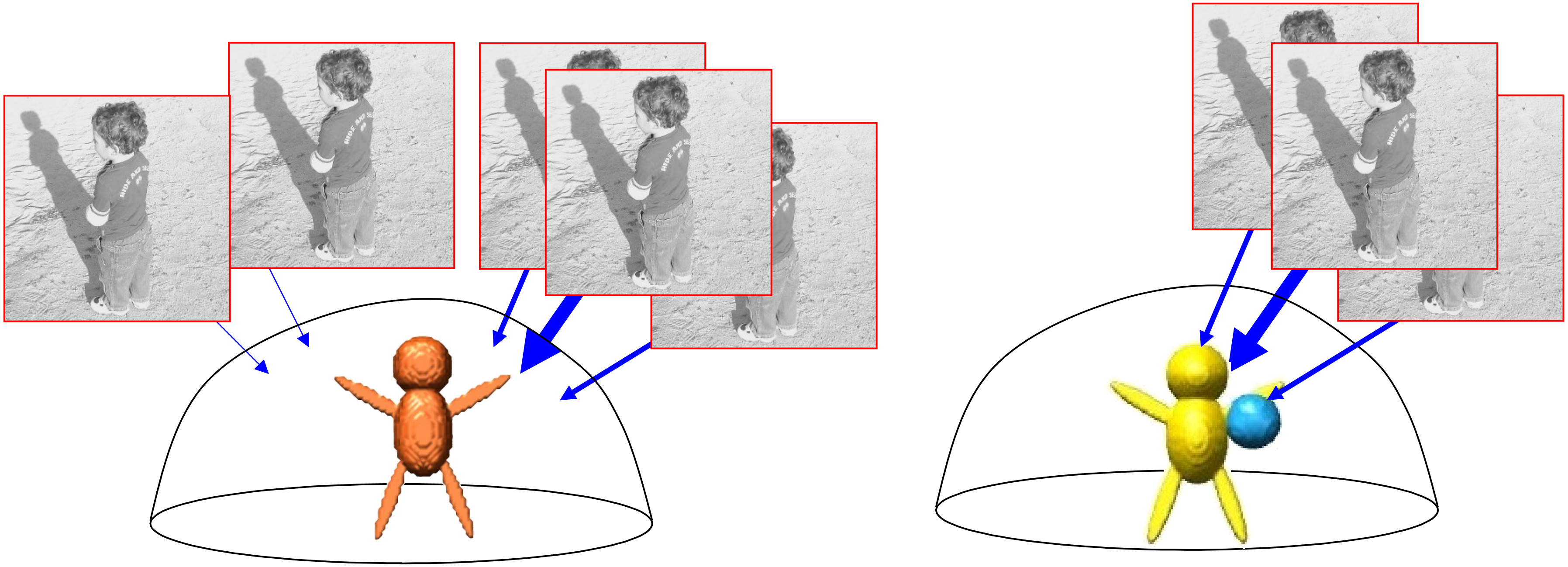
Multi-reference refinement



Multi-reference refinement



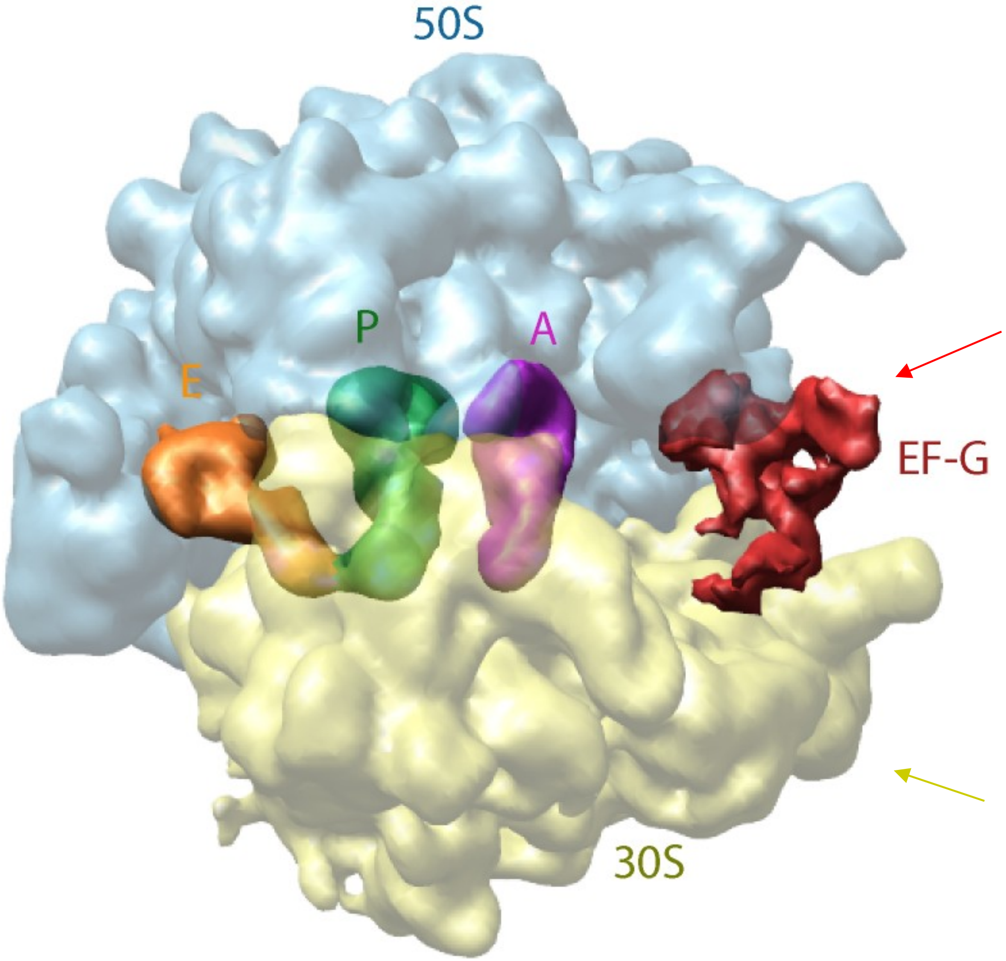
ML3D classification



“Probability-weighted angular assignment”

Prelim. ribosome reconstruction

91,114 particles; 9.9 Å resolution

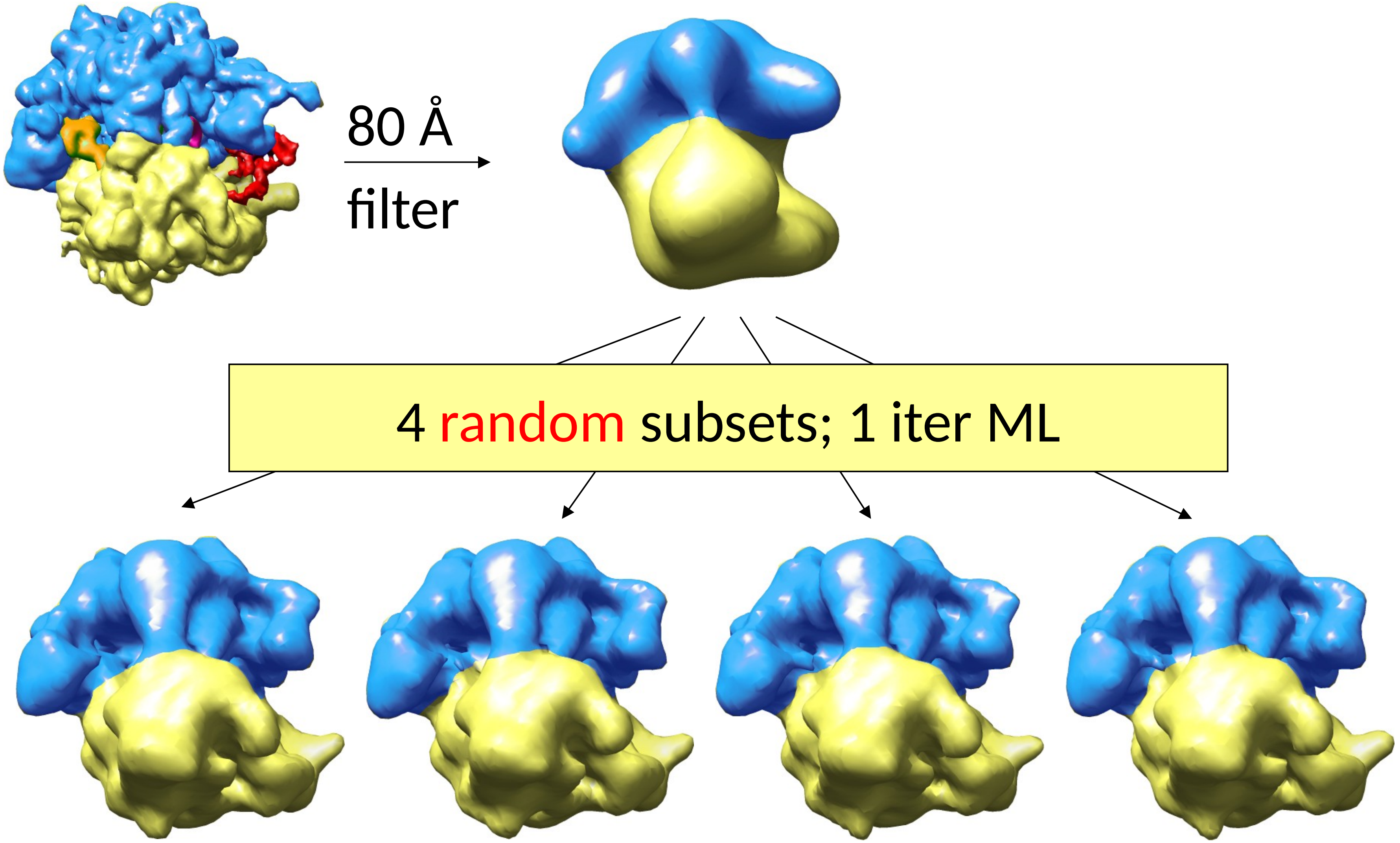


fragmented

(depicted at a lower threshold)

blurred

Seed generation

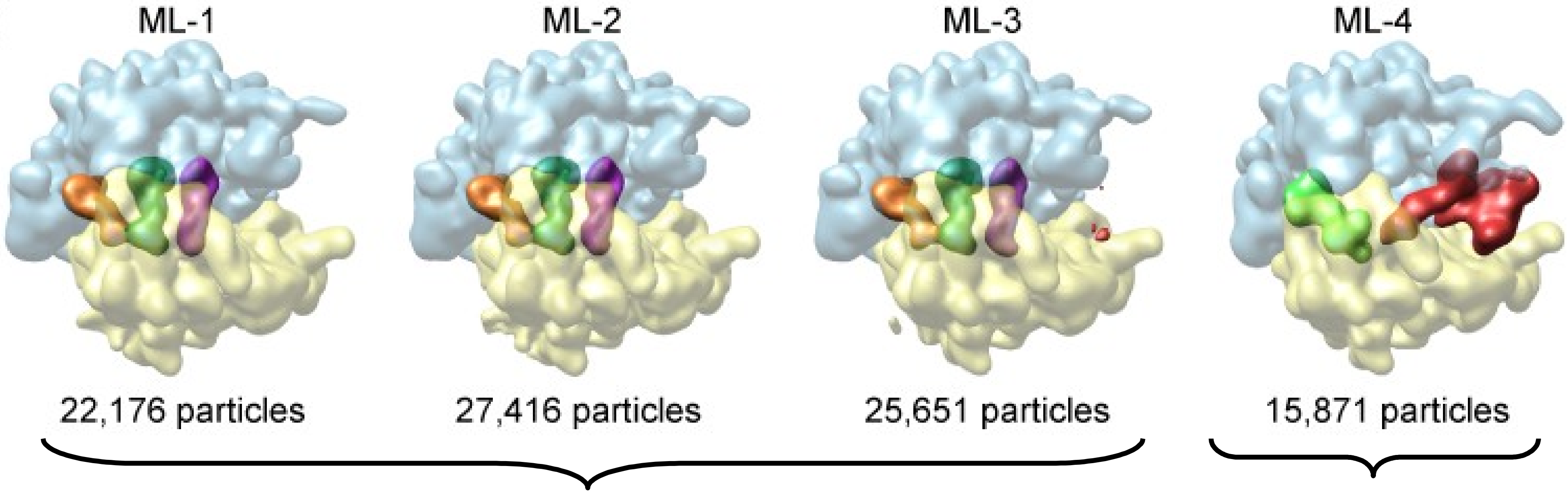


80 Å

filter

4 random subsets; 1 iter ML

ML-derived classes



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

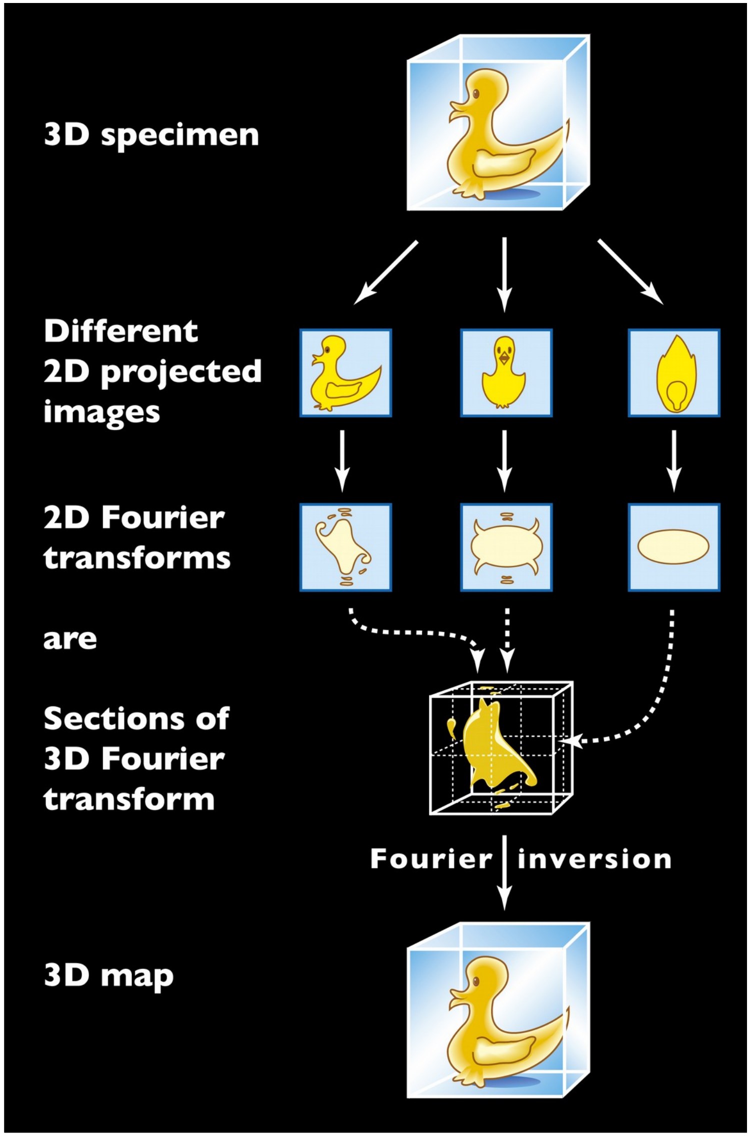
ratcheting,
EF-G, 1 tRNA

(Results coincided with a supervised classification)

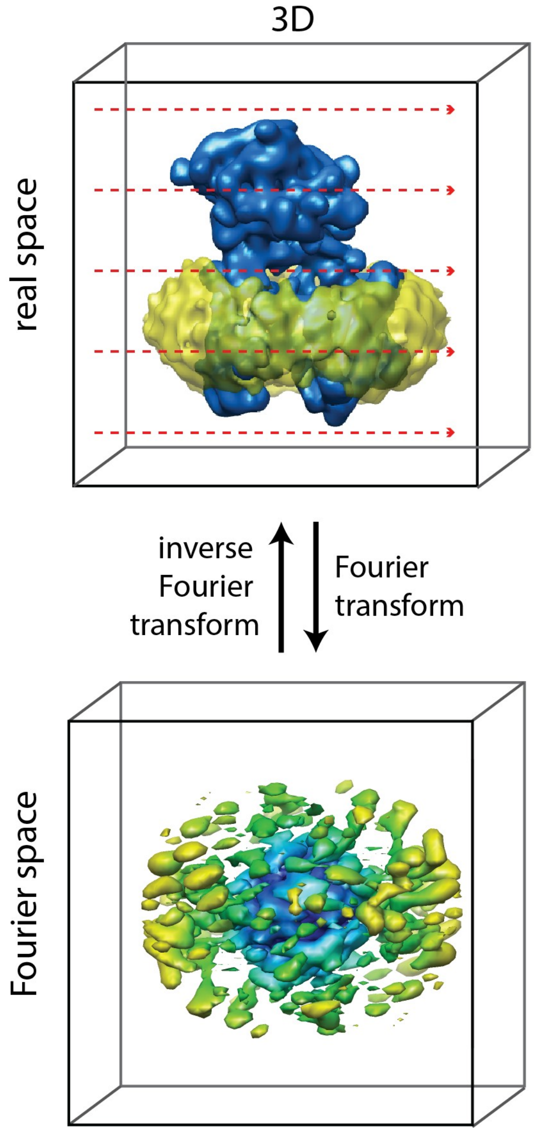
Scheres et al (2007) Nat. Meth.

Fourier-space formulation

Projection-slice theorem


3D specimen
Different 2D projected images
2D Fourier transforms
are
Sections of 3D Fourier transform
Fourier inversion
3D map

Projection-slice theorem



3D

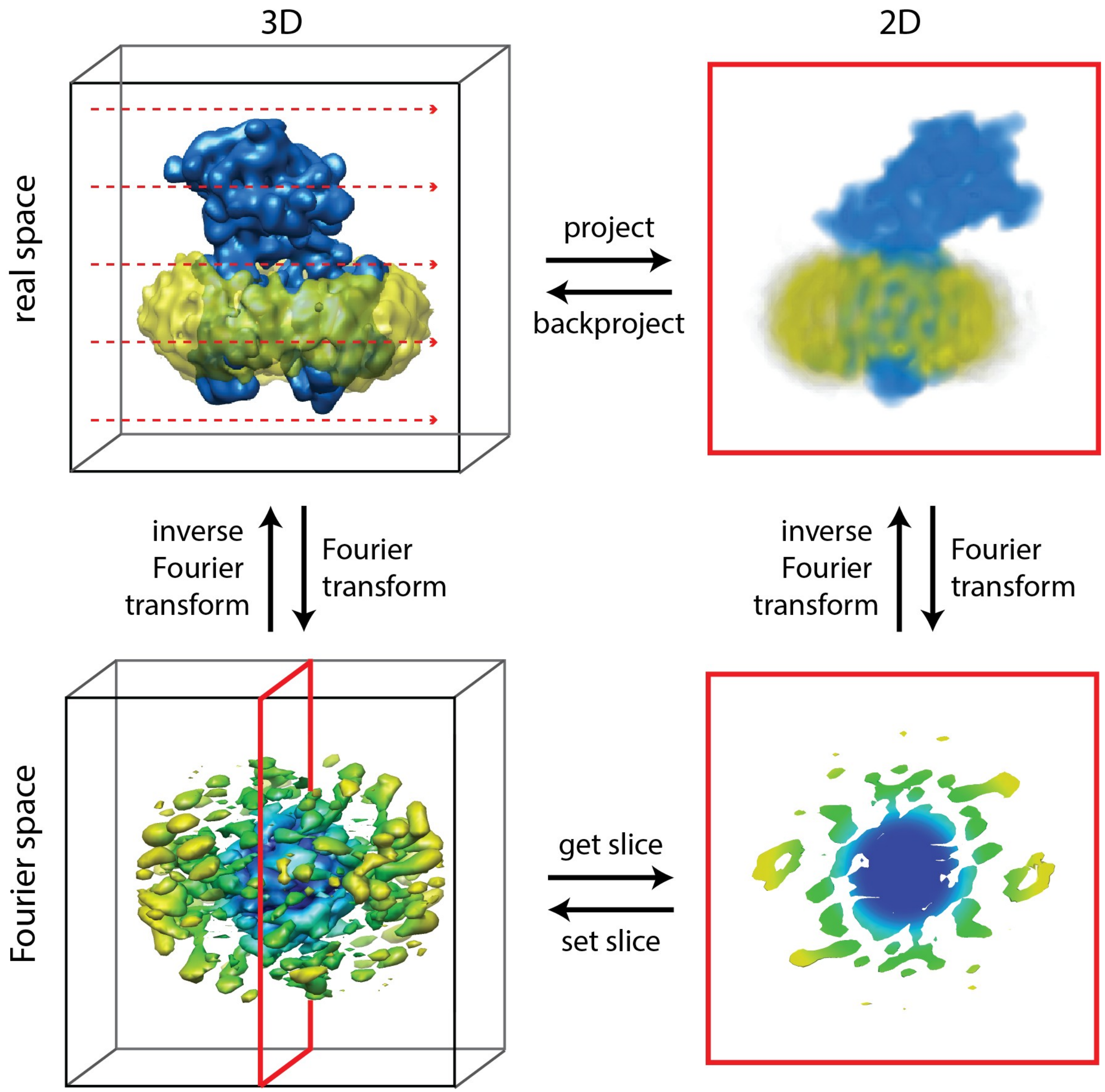
real space

inverse Fourier transform

Fourier transform

Fourier space

Projection slice theorem



3D

2D

real space

project

backproject

inverse Fourier transform

Fourier transform

inverse Fourier transform

Fourier transform

Fourier space

get slice

set slice

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

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

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

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

Regularised likelihood optimisation

Likelihood

- Assume noise is Gaussian and independent
  - in Fourier space
  - with spectral power σ²(υ): *coloured noise*

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

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

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

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

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

Estimate resolution-dependent power of noise from the data

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

Estimate resolution-dependent power of signal from the data

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

3D Wiener filter

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

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

WITHOUT ARBITRARINESS!

Recapitulating

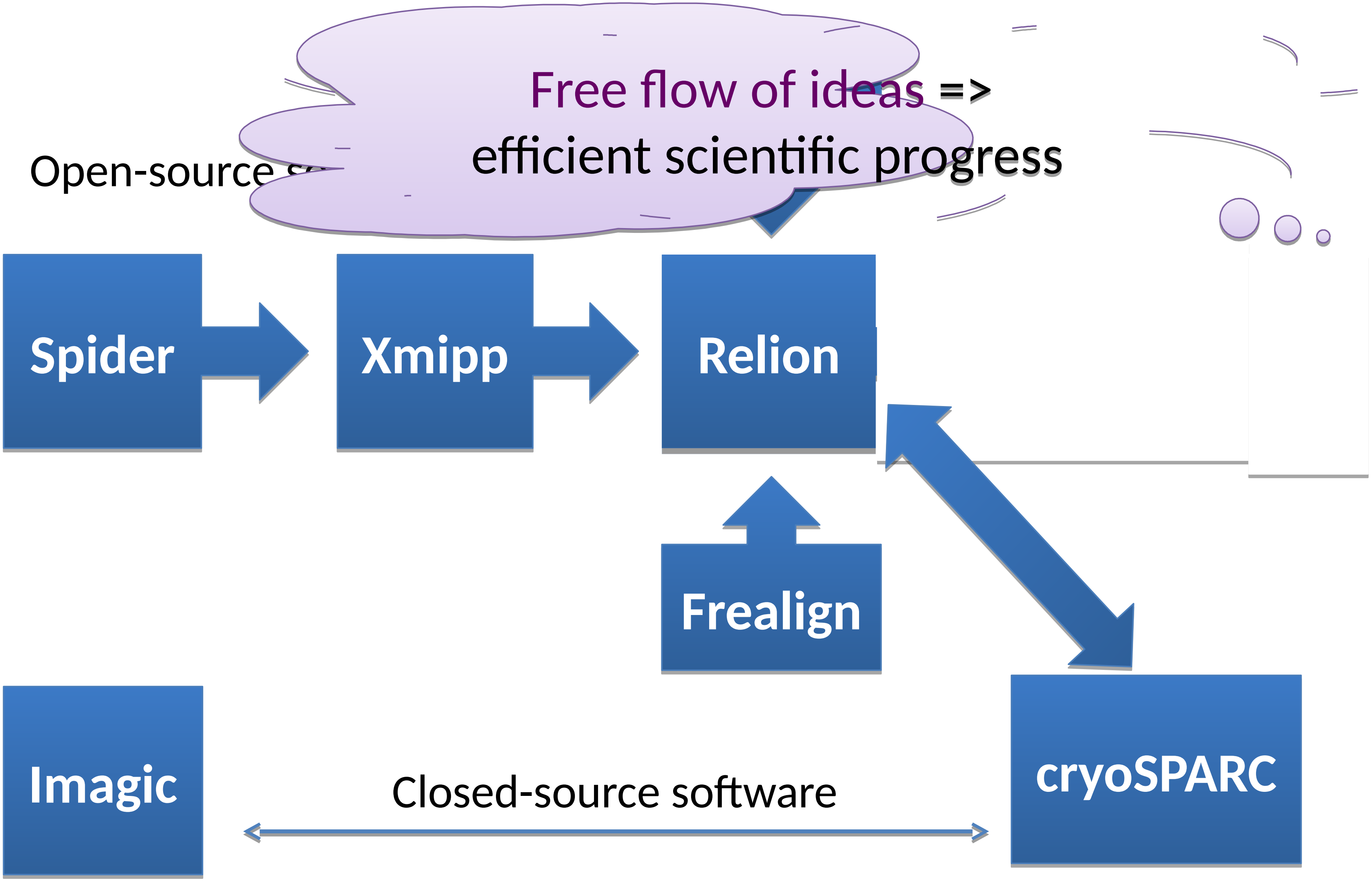
- Alignment & classification are incomplete problems
  - Best dealt with by marginalisation (ML)
- 2D and 3D problems are very similar
- Fourier-space is most convenient
  - CTF multiplication
  - Slices instead of line integral projections
  - Coloured noise-model
  - Regularised Likelihood function -> ‘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)

Some thoughts on cryo-EM software

Open software in a sharing community



Free flow of ideas =>

efficient scientific progress

Open-source s

Spider

Xmipp

Relion

Frealign

Imagic

Closed-source software

cryoSPARC

Recent trend of commercialisation

- Pharmaceutical interest -> commercial interest

- Protective measures
  - Restrictive licenses
  - Closed-source
  - Patents

Patents in cryo-EM software (I)

- We’re used to patents for hardware
- Not so for mathematical concepts
- Software development is much cheaper!
- Academics typically do software development themselves, but not hardware

Patents in cryo-EM software (II)

- Apply widely, rely on patent offices to restrict
  - Which patent officer will be expert on cryo-EM algorithms?
  - In US many things possible, EU is more restrictive
  - US-only patents still hard as companies are international

- Separation between academics/industry is extremely difficult
  - Collaborations, spin-offs, liability, etc.

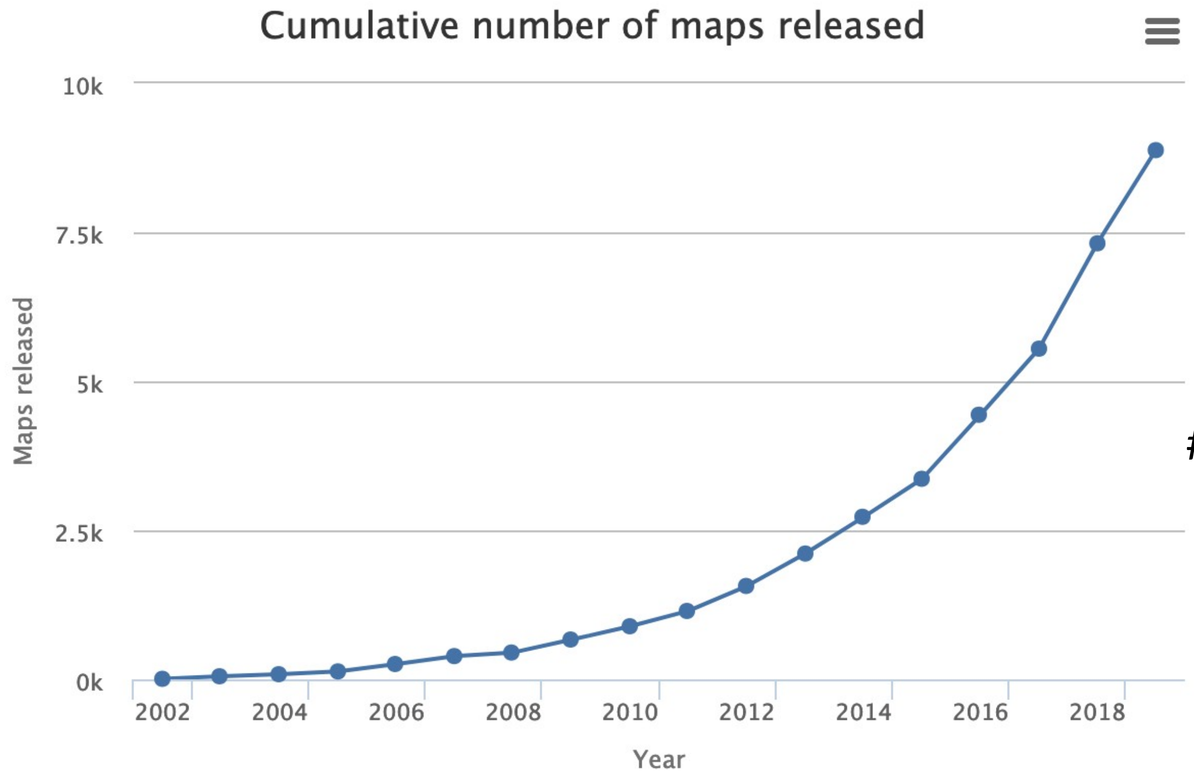
A warning from the past

- Commercial distribution rights to Xplor were owned by a small company
  - Good intentions; highly academic

- 15-20 years later, in hands of other company, these rights caused trouble
  - Xplor -> CNS -> CNX (now ~dead)
  - Academics had to restart from scratch: Phenix

Open software in a sharing community

Free flow of ideas =>
efficient scientific progress



@SjorsScheres

#OpenSoftwareAcceleratesScience

(EMStats: EMDB-Statistics)