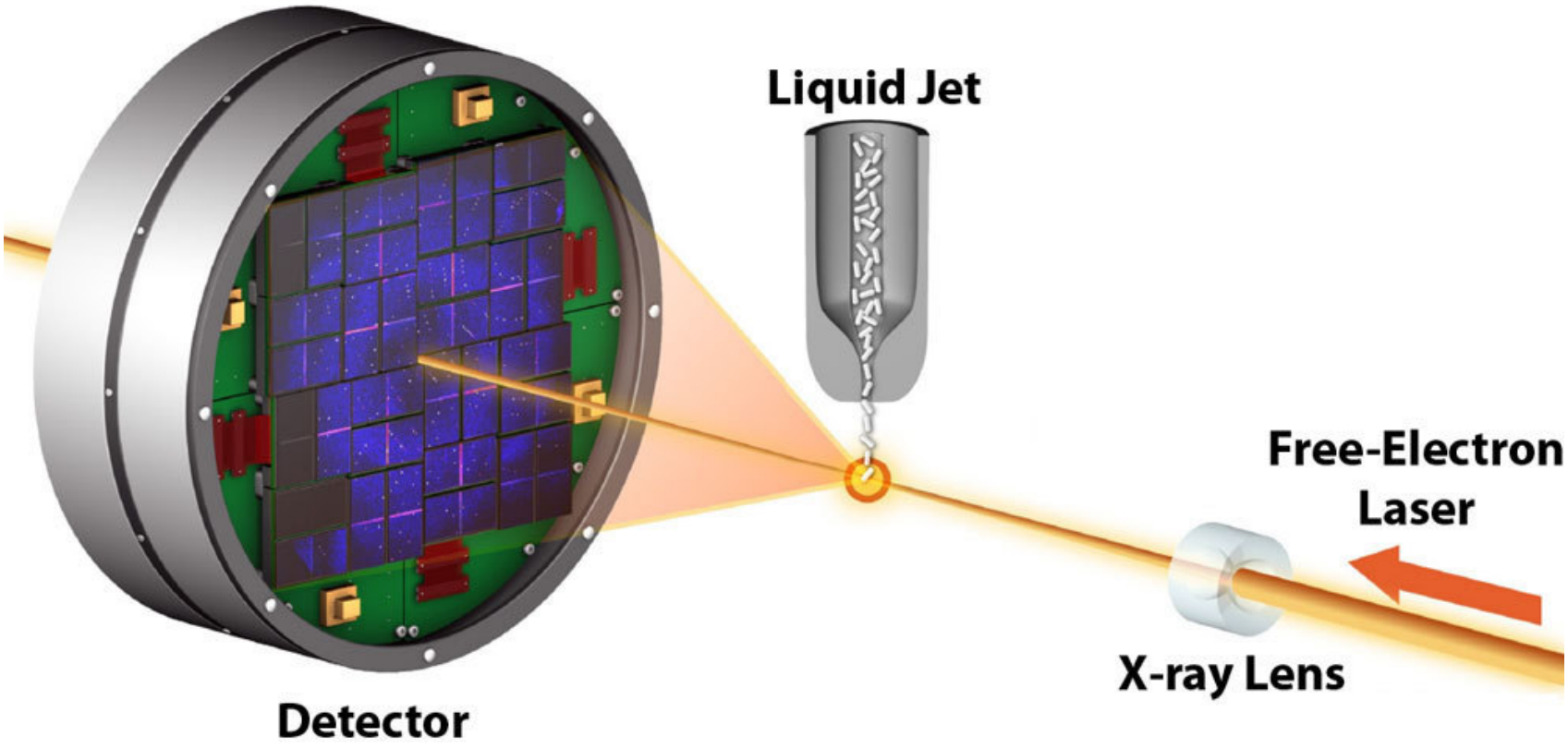
Clustering and main sources of variance of SFX data

Wolfgang Brehm[1]

[1] Faculty of Biology
University of Konstanz

2014-08-24

Free electron lasers

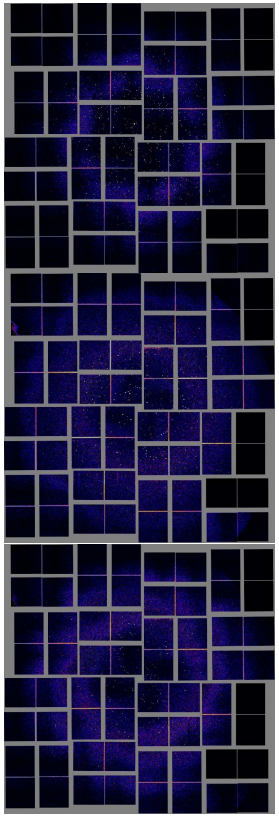


Max Plank Gesellschaft: http://www.mpg.de/5826245/free_electron_laser

The data

- many crystals
- many diffraction patterns
- indexed and integrated intensities for each pattern
- merged intensities

data in detail



- indexing
- integration

Reflections measured after indexing

h	k	l	I	sigma(I)
-43	20	-4	-206.69	239.90
-42	19	-4	-189.67	228.11
-42	19	-3	157.33	260.68
-42	19	-2	302.28	250.14
...				

Reflections measured after indexing

h	k	l	I	sigma(I)
-16	-17	7	74.39	1140.31
-16	-11	-11	-3044.06	2432.59
-16	-8	-13	480.81	817.36
-16	-5	17	0.00	3.13
...				

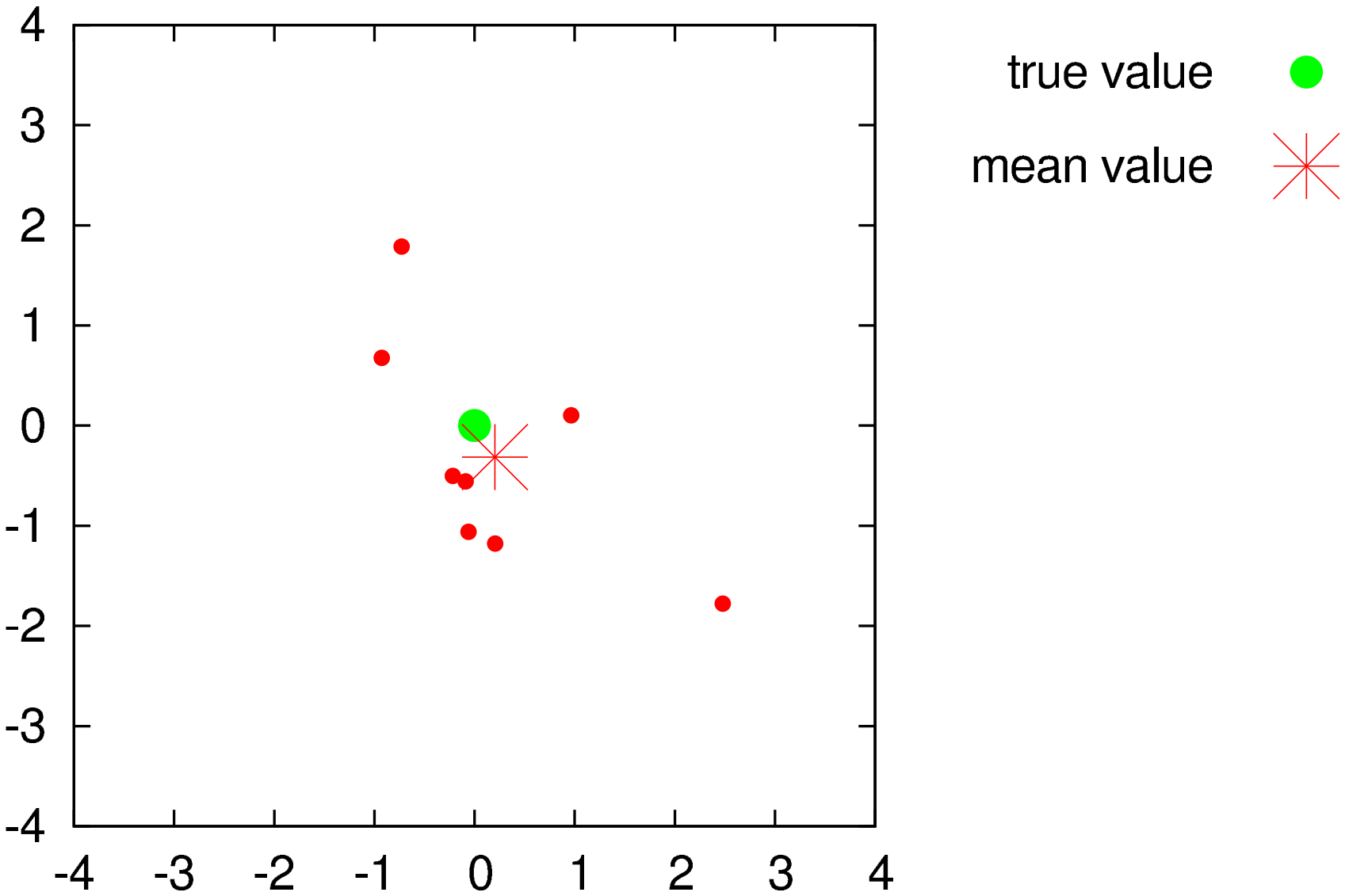
Reflections measured after indexing

h	k	l	I	sigma(I)
-16	-4	-15	-225.42	749.51
-16	-3	18	0.00	3.13
-16	-2	-16	5785.33	952.90
-16	2	20	0.00	3.13
...				

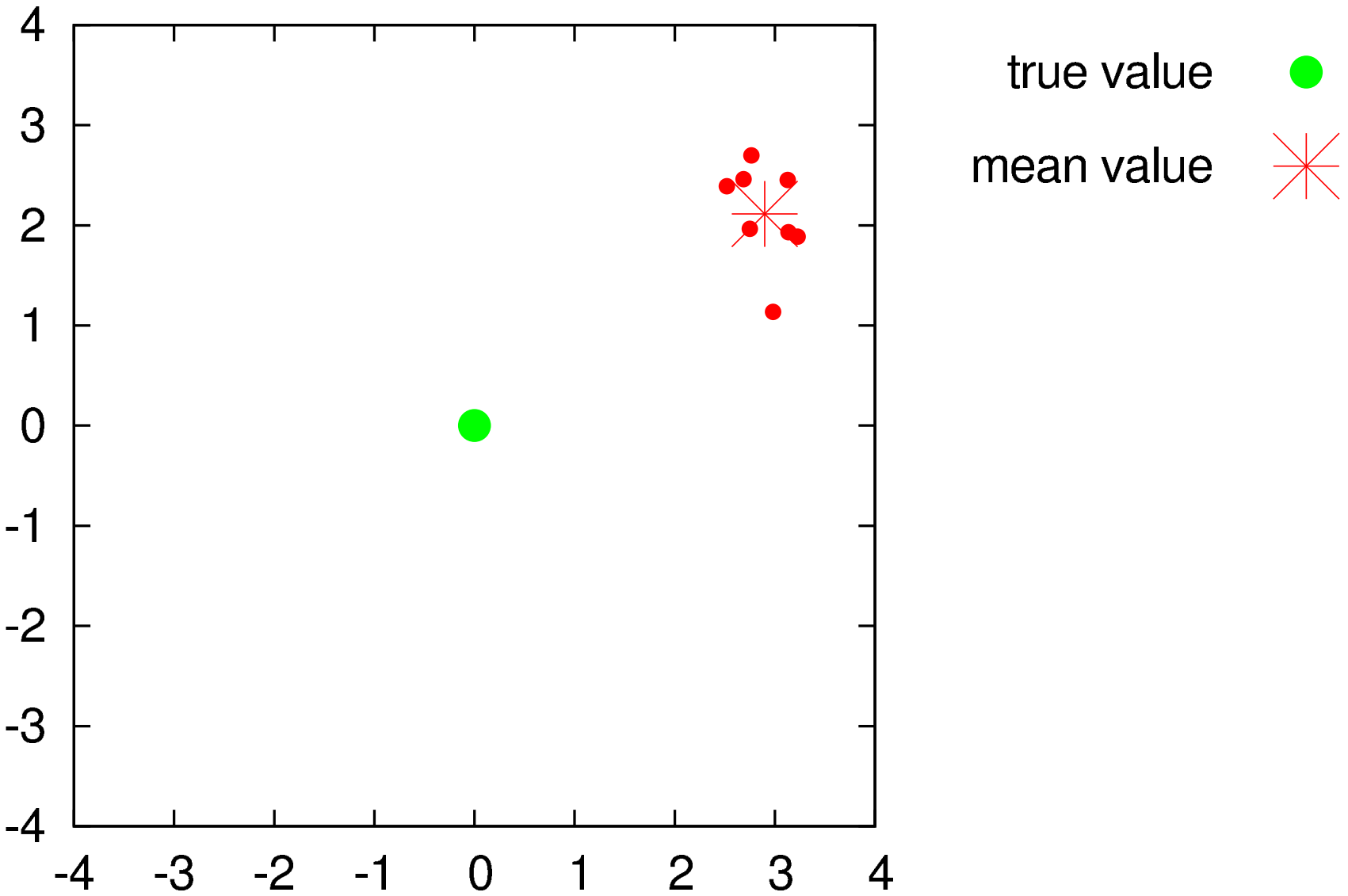
merging

h	k	l	I	sigma(I)
-43	20	-4	-206.69	239.90
-42	19	-4	-189.67	228.11
-42	19	-3	157.33	260.68
-42	19	-2	302.28	250.14
...				

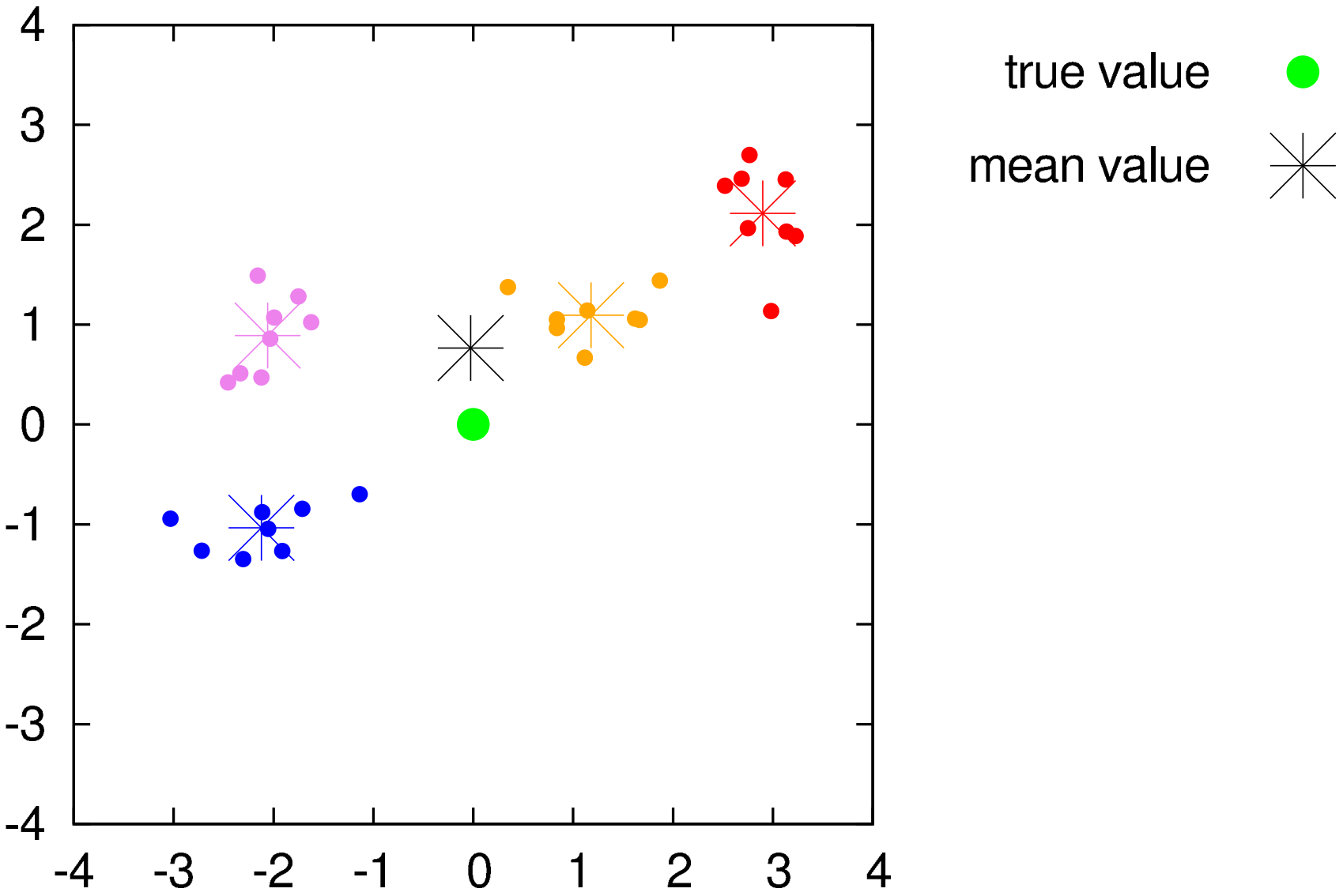
Systematic vs. random errors, Precision vs. Accuracy



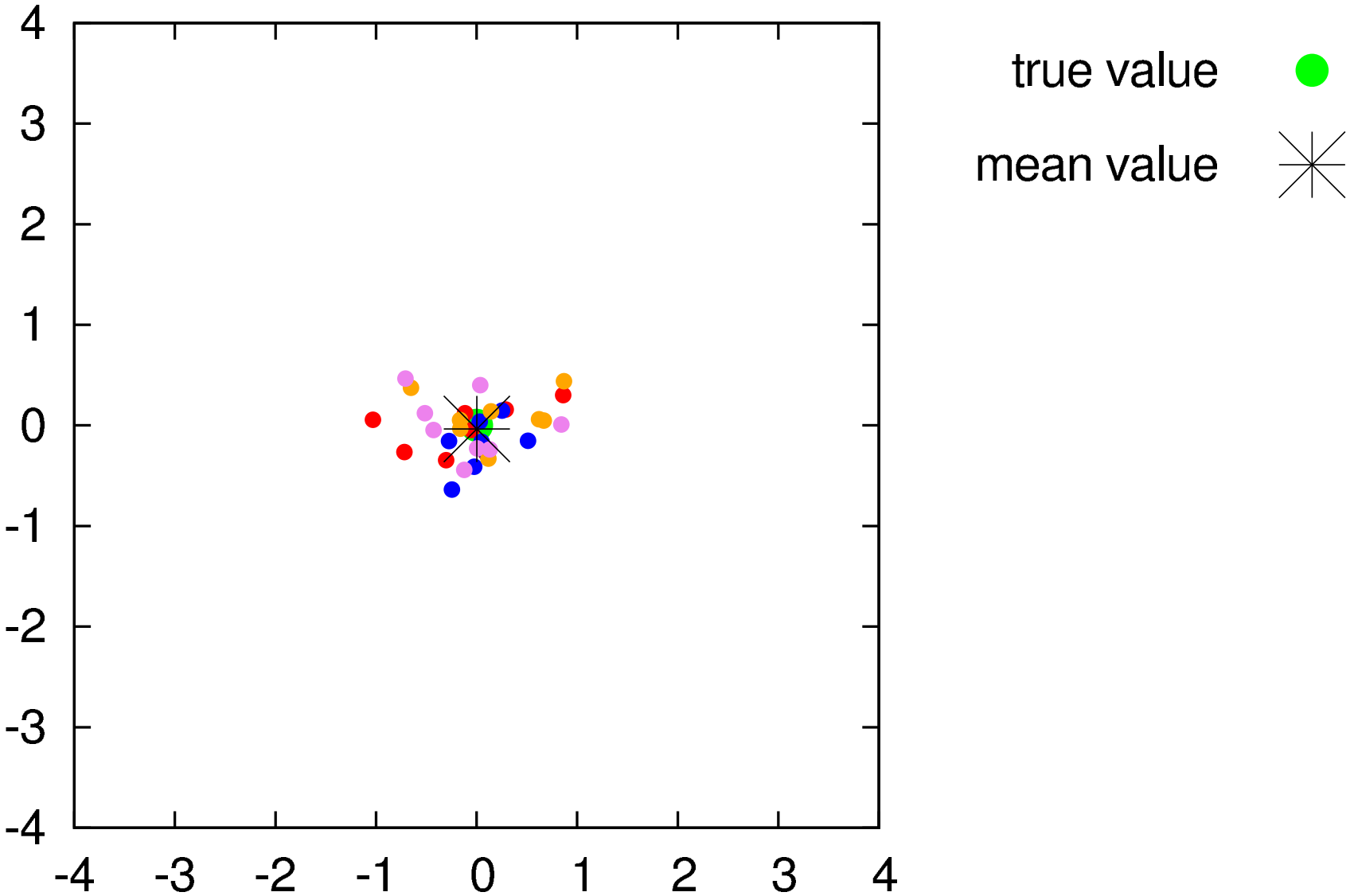
Systematic vs. random errors, Precision vs. Accuracy



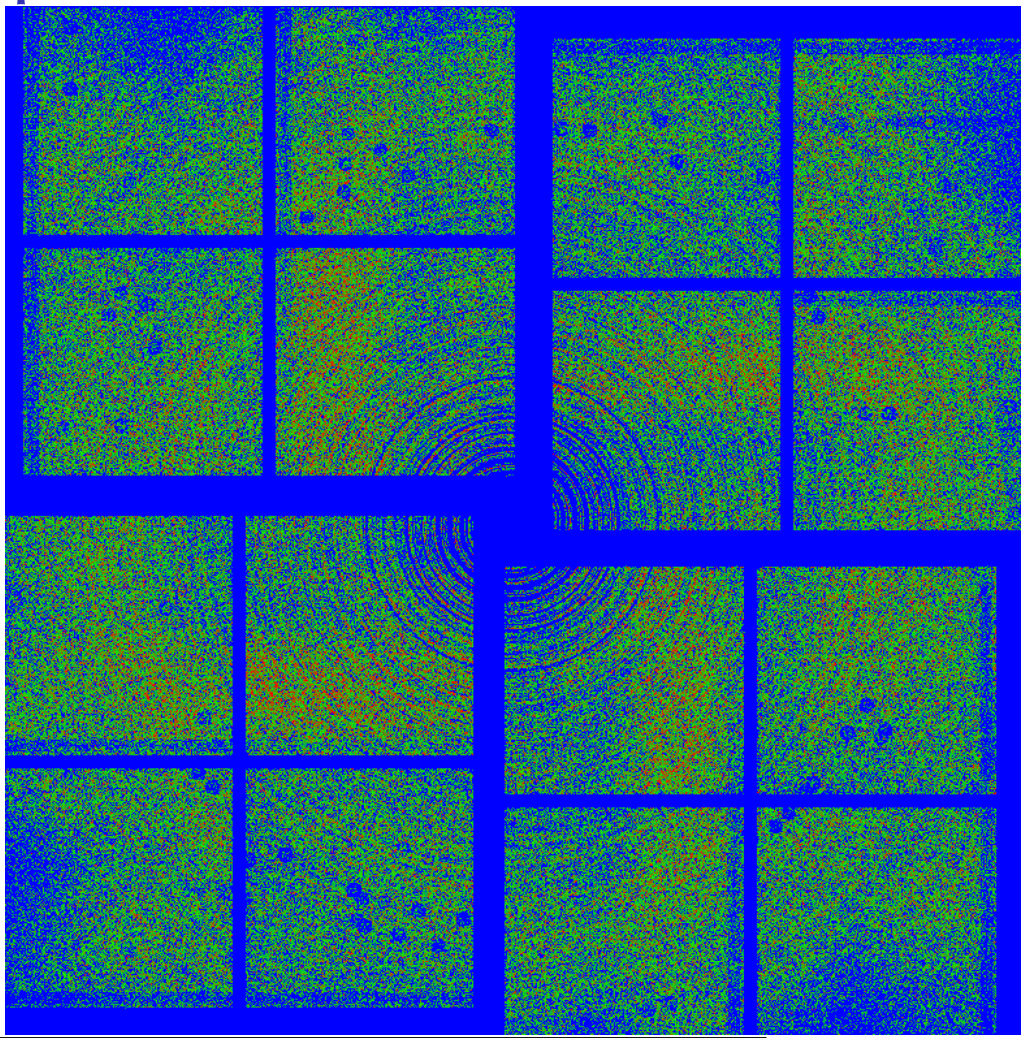
Many groups with high precision, overall low accuracy



Many groups with high precision, corrected for systematic errors



Example for systematic errors in CFEL: Bad parts of detector



- histogram of the position of integrated spots of the innermost modules.
- red = many, blue = none

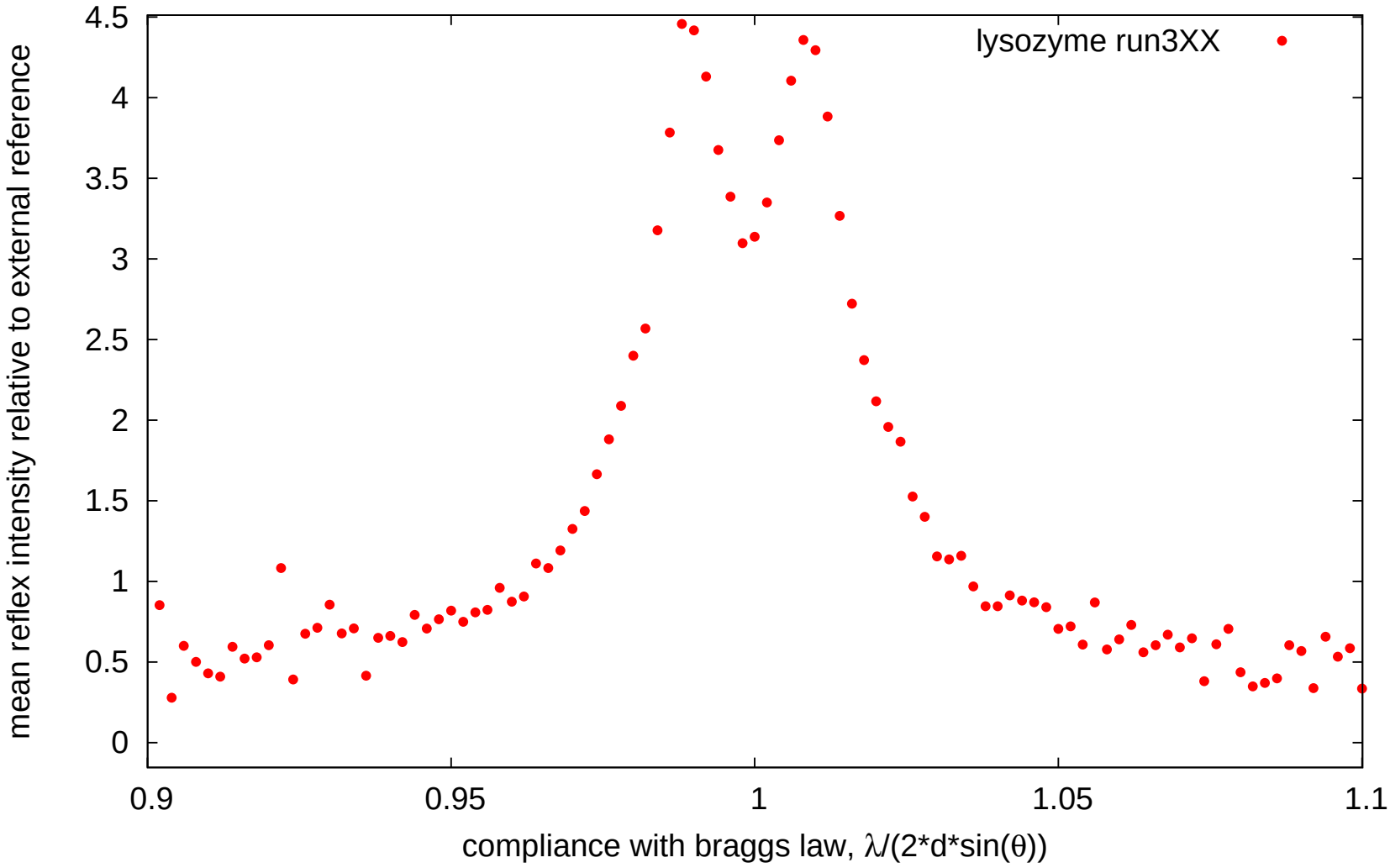
[1] Sébastien Boutet et al. "High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography". In: *Science* 337.6092 (2012), pp. 362–364. DOI: 10.1126/science.1217737

Example for systematic errors in CFEL: Partiality

Janet Smith Lab. “Complete X-ray Diffraction Dataset Collected From One Crystal”
Movie by Dr. Todd Geders

- reflections “come and go:” grow in intensity and disappear again
- crystal is rotated slowly
- Bragg’s law states the condition for which the reflex intensity is maximal

Example for systematic errors in CFEL: Partiality



Variance analysis

- To fix systematic errors we need to find the variables the errors are depending on

The XFEL data has some special properties which should be accounted for:

- different mean intensity per shot as big source of variance (scaling)
- sparse data (<1% of all possible reflexes are in one shot)
- different intensities are determined to a varying degree
- large errors

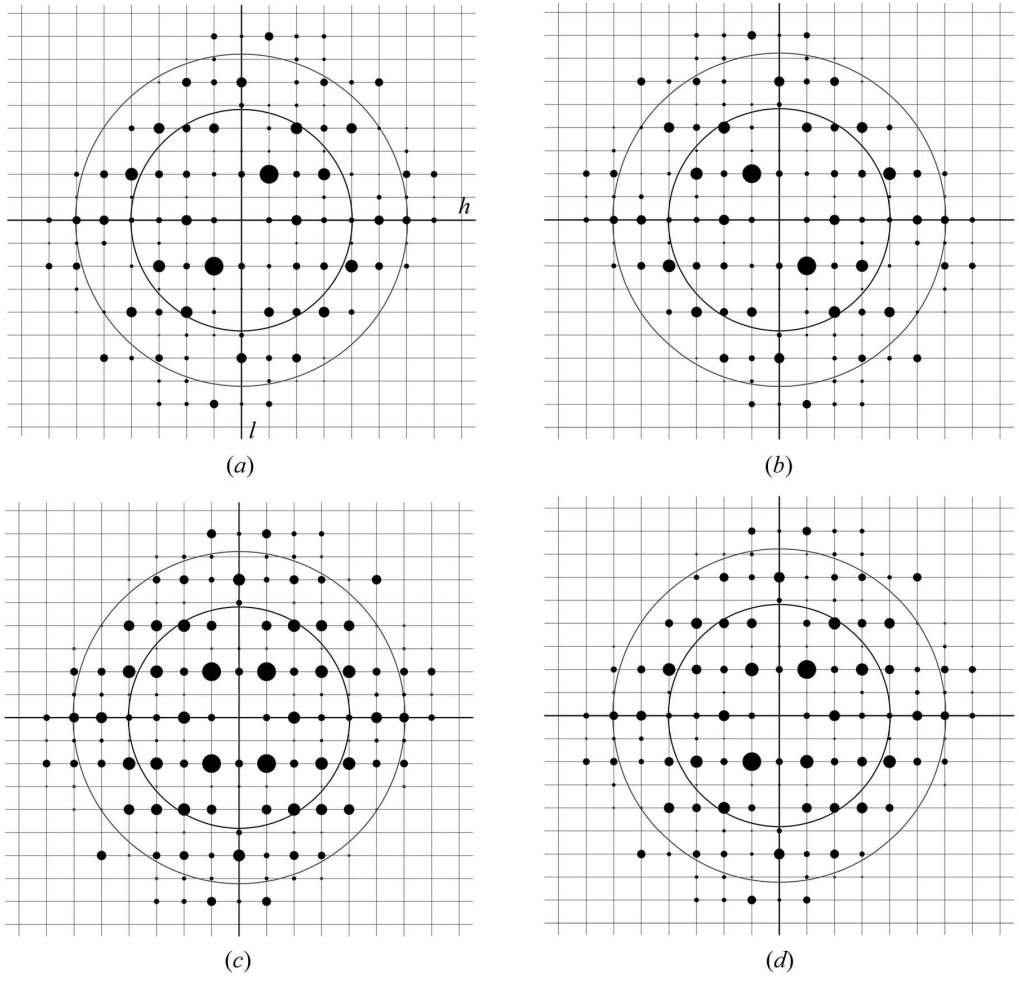
Correcting systematic errors

- find the physical law that explains the variance
  - textbook, physicist, mathematician
- revert the effect of the law on the data

In most cases the law depends on variables:

- if possible measure the variables
- else fit the variables to the data

Crystallographic twinning



(a) diffraction pattern of one crystal volume

(b) diffraction pattern of the same crystal rotated by 90° C e.g. from a twin

(c) Superposition of (a) and (b) with a twinning fraction of 0.5

(d) Superposition of (a) and (b) with a twinning fraction of 0.2

[7] Simon Parsons. "Introduction to twinning". In: *Acta Crystallographica Section D* 59.11 (2003), pp. 1995–2003. DOI: 10.1107/S0907444903017657

Systematic error when indexing: Computational twinning

- roughly 1/3 of all projects have crystals with a spacegroup which allowes for different indexing modes
- the indexing cannot be determined unambiguously from the positions of thte intensities
- every crystal is indexed in a random indexing mode, with many frames every indexing mode will have roughly the same probability
- merging will introduce an artificial symmetry
  - computational twinning

The idea

- the mean agreement between shots of the same indexing mode should be high
- the mean agreement between shots of differing indexing mode should be low

We are using the correlation coefficient [4] as a measure for agreement :

$$r_{ij} = \frac{\sum_n (I_i(h) - \bar{I}_i)(I_j(h) - \bar{(I_j)})}{\sum_n \sqrt{(I_i(h) - \bar{I}_i)^2} \times \sum_n (I_i(h) - \bar{I}_i)^2} \tag{1}$$

[4] Sir Ronald A. Fisher. *Statistical Methods for Research Workers.* Hafner Publishing Company Inc., 1954

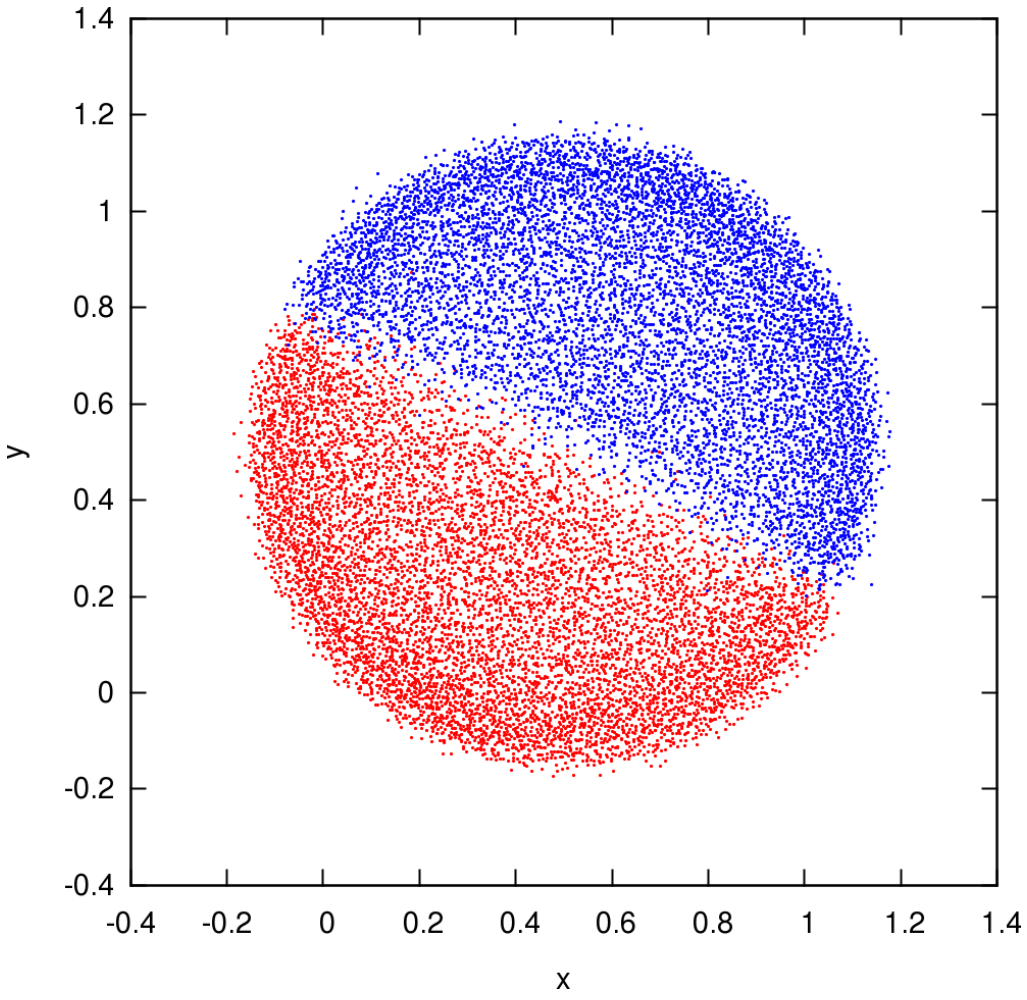
Proof of concept: Classical multidimensional scaling, Algorithm 1 [2]

- Each shot is represented by a point x in two dimensional space
- Interprete $1 - r_{ij}$ as a distance and minimize Ψ

$$\Psi = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[(1 - r_{ij}) - |x_i - x_j| \right]^2 \tag{2}$$

[2] Wolfgang Brehm and Kay Diederichs. "Breaking the indexing ambiguity in serial crystallography". In: *Acta Crystallographica Section D* 70.1 (Jan. 2014), pp. 101–109. DOI: 10.1107/S1399004713025431

Result of Algorithm 1



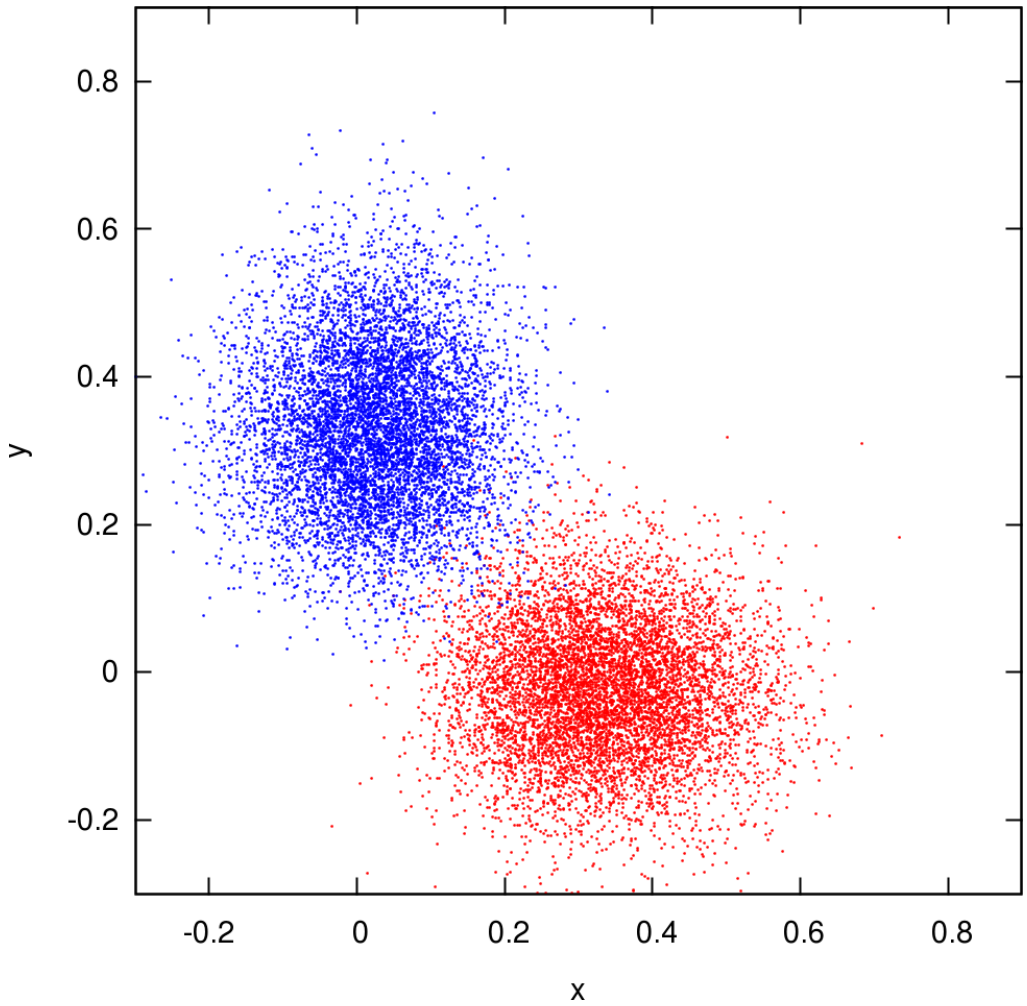
- synthetic data set with 15445 shots
- known indexing modes in red and blue
- x, y are abstract coordinates
- 5 min until convergence on my Laptop
- 1 % wrong assignments when separating based on this outcome

Next step: Dot product, Algorithm 2 [2]

- each shot is represented by a point x in a 2 dimensional space
- minimize Φ and the deviation between the dot product between all pairs of shots that can be compared and the correlation coefficient between both shots is minimal
- interprete the correlation coefficient as scalar product

$$\Phi = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [r_{ij} - x_i \cdot x_j]^2 \tag{3}$$

- better, easier, faster separation
- scalar product similar to the correlation coefficient without "scaling"
- visual result

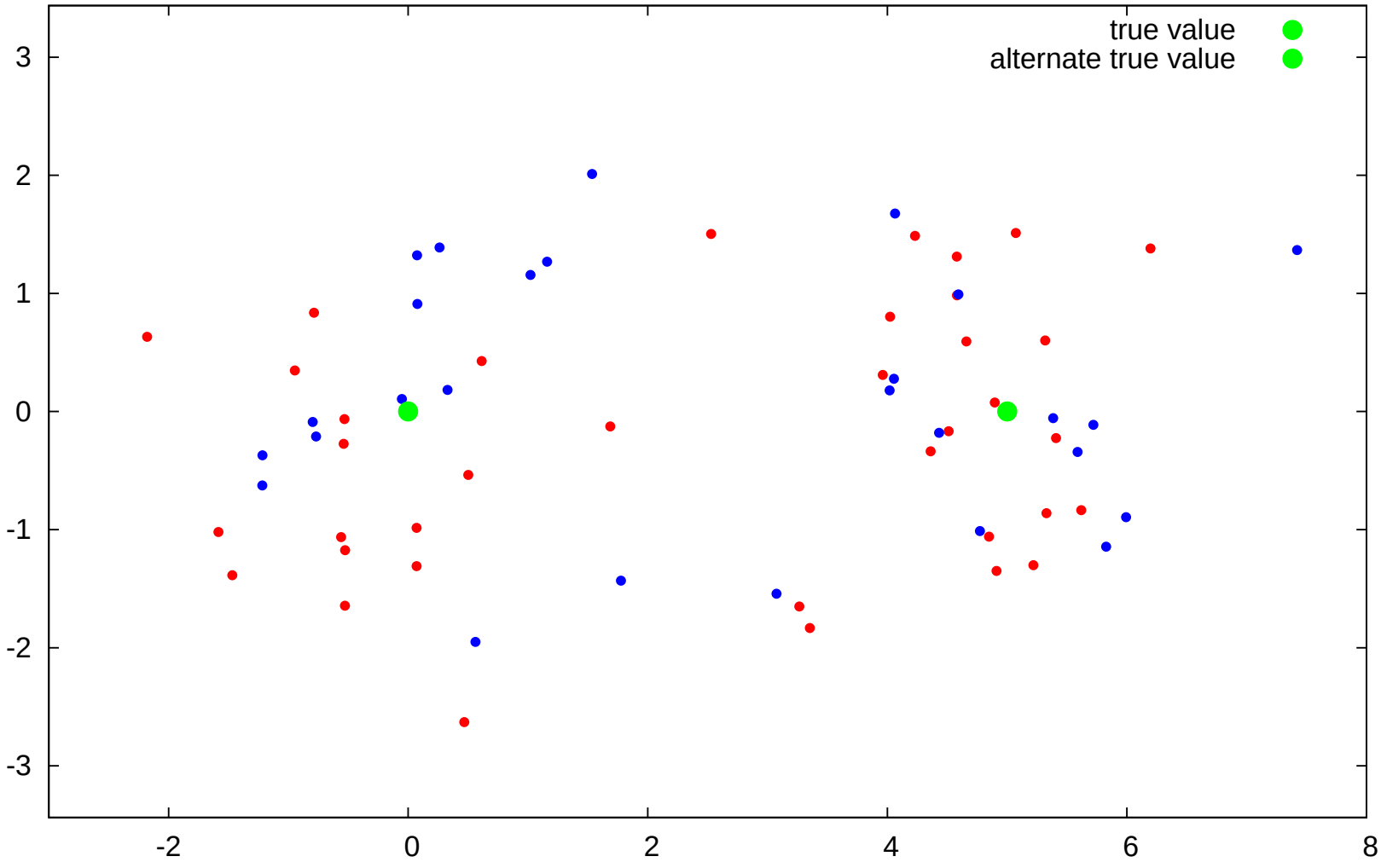


- synthetic data set with 15445 shots
- known indexing modes in red and blue
- x, y are abstract coordinates
- 50s until convergence on my Laptop
- 0.8 % wrong assignments when separating based on this outcome

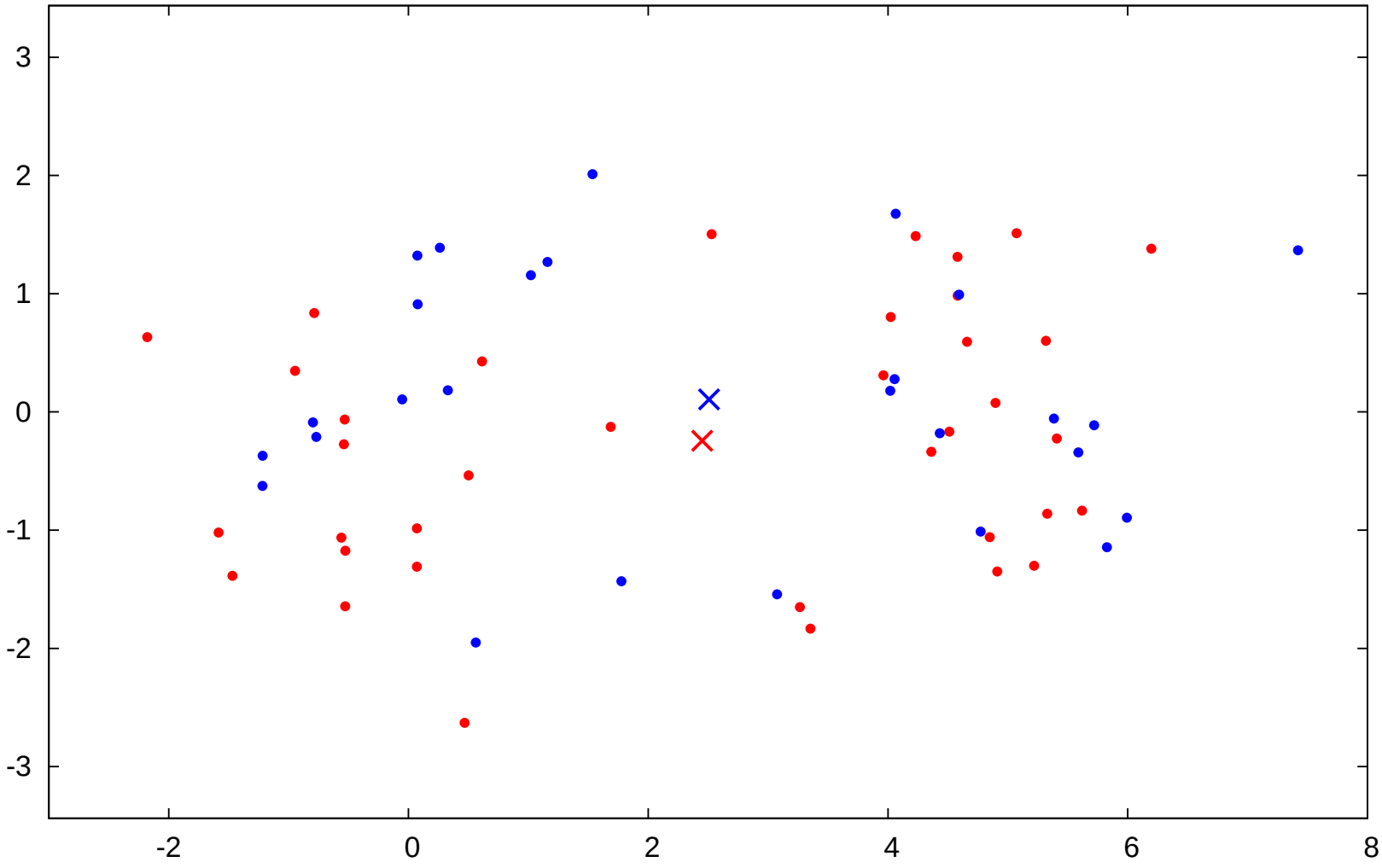
k-means clustering, Algorithm 3 (unpublished)

Find two groups that have a high agreement within and a low agreement between them.
-> k-means clustering

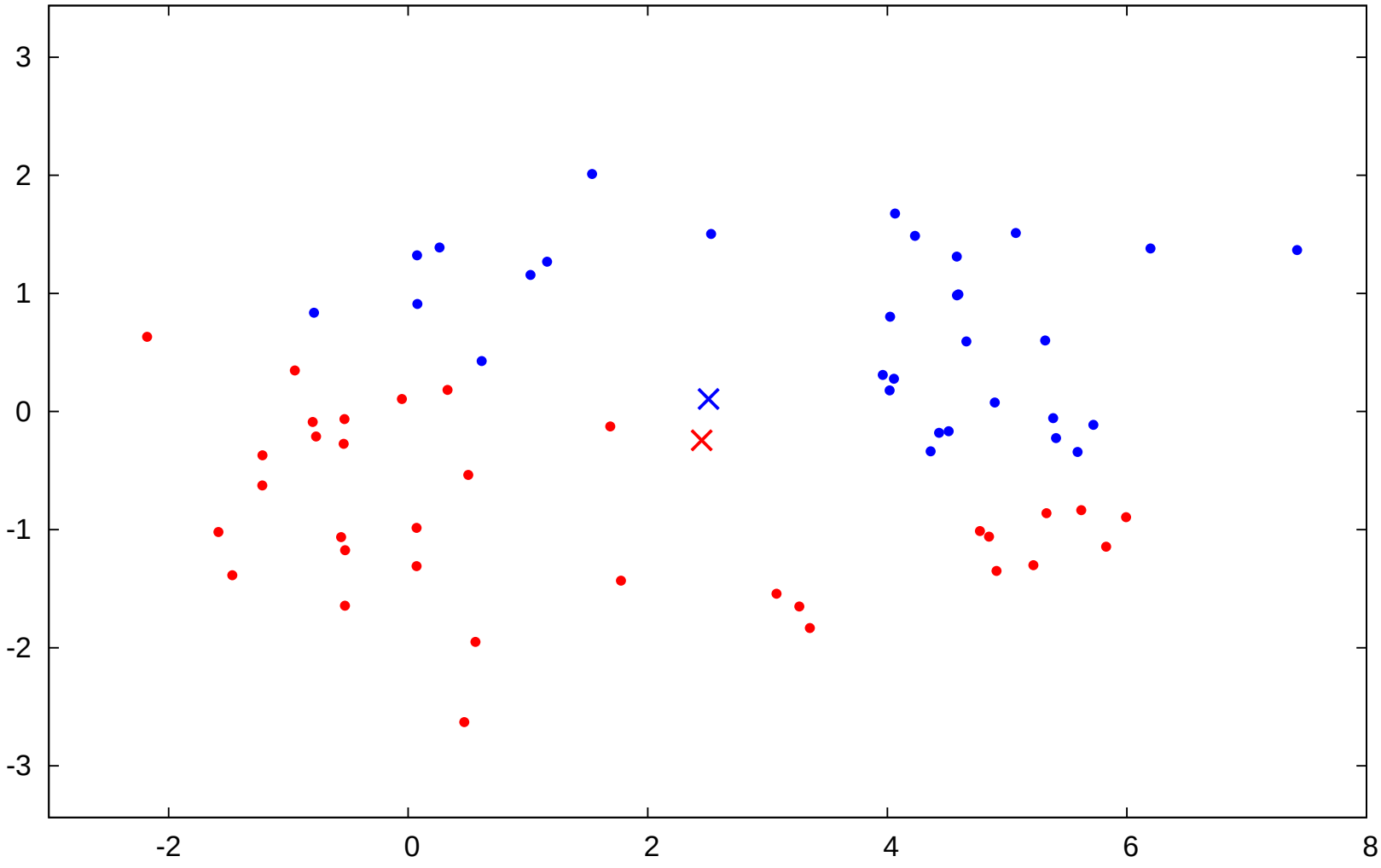
k-means clustering



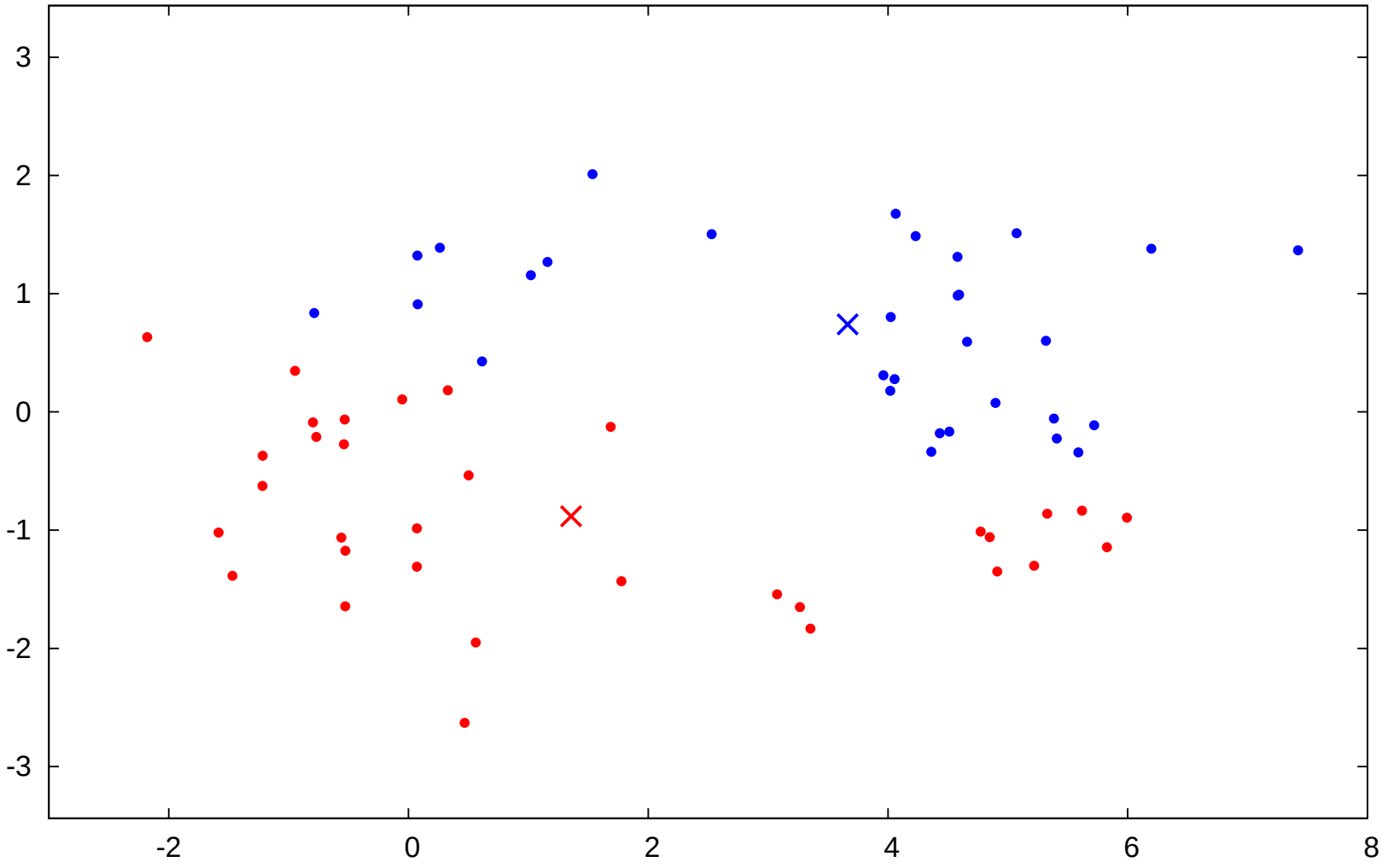
k-means clustering



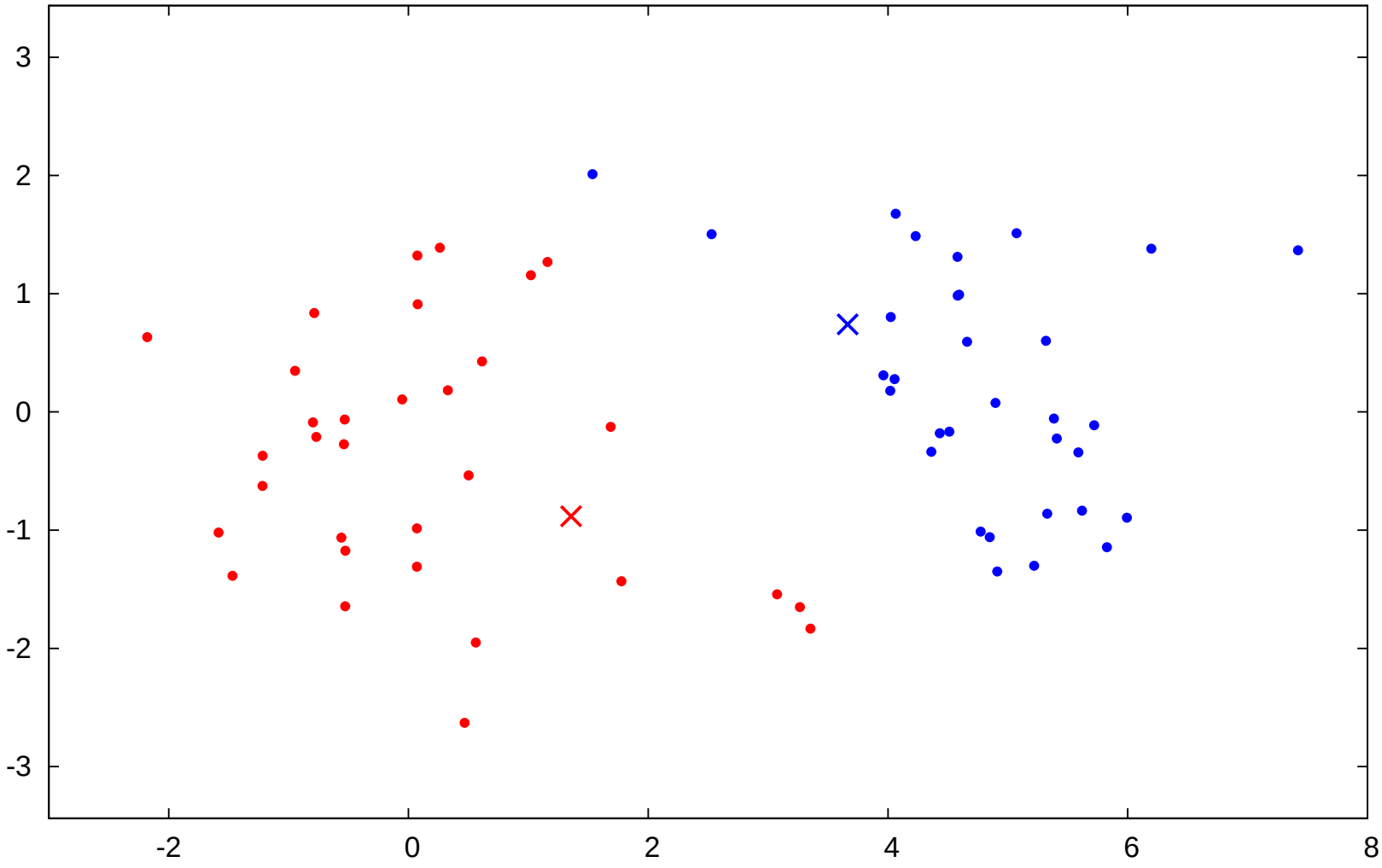
k-means clustering



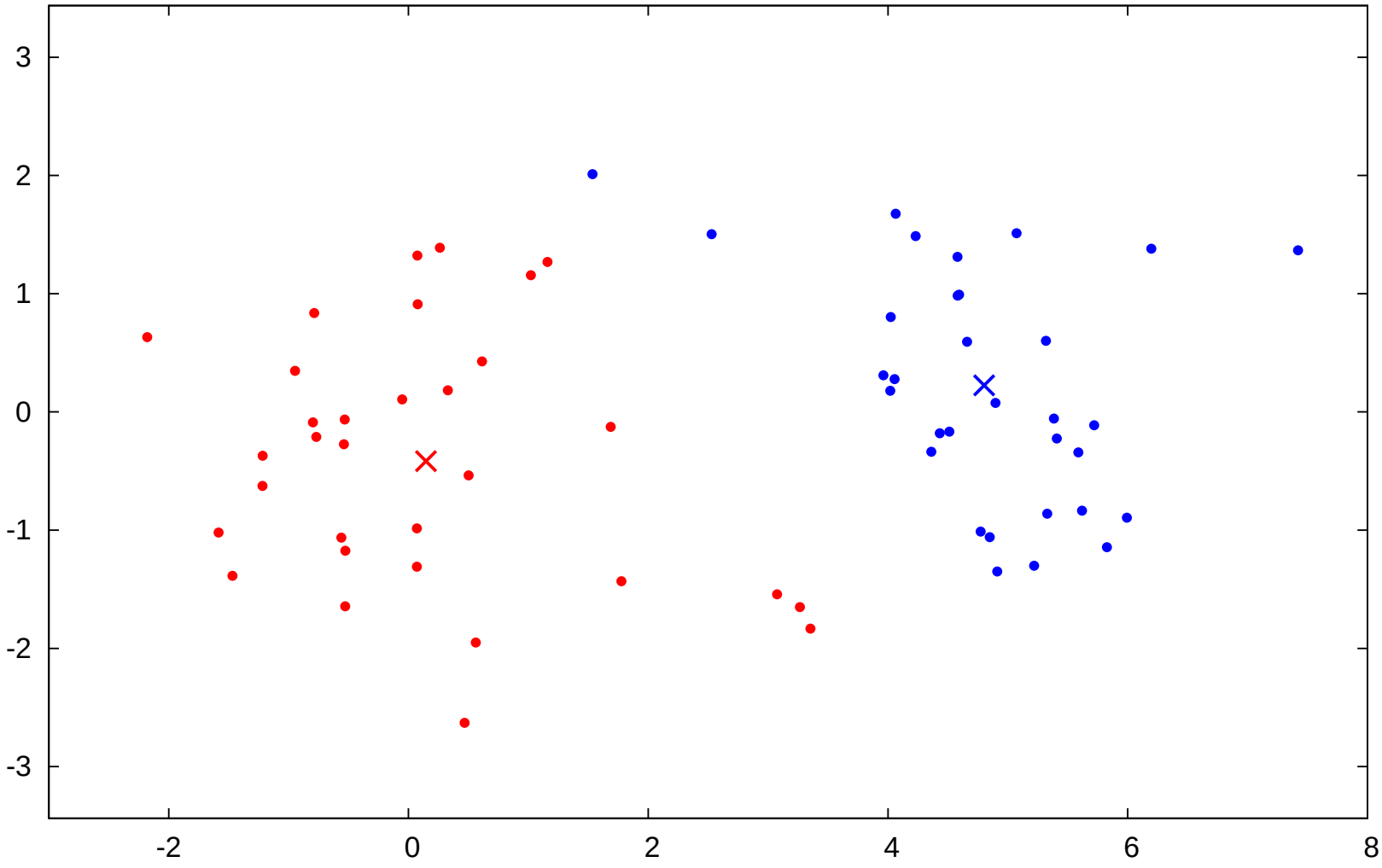
k-means clustering



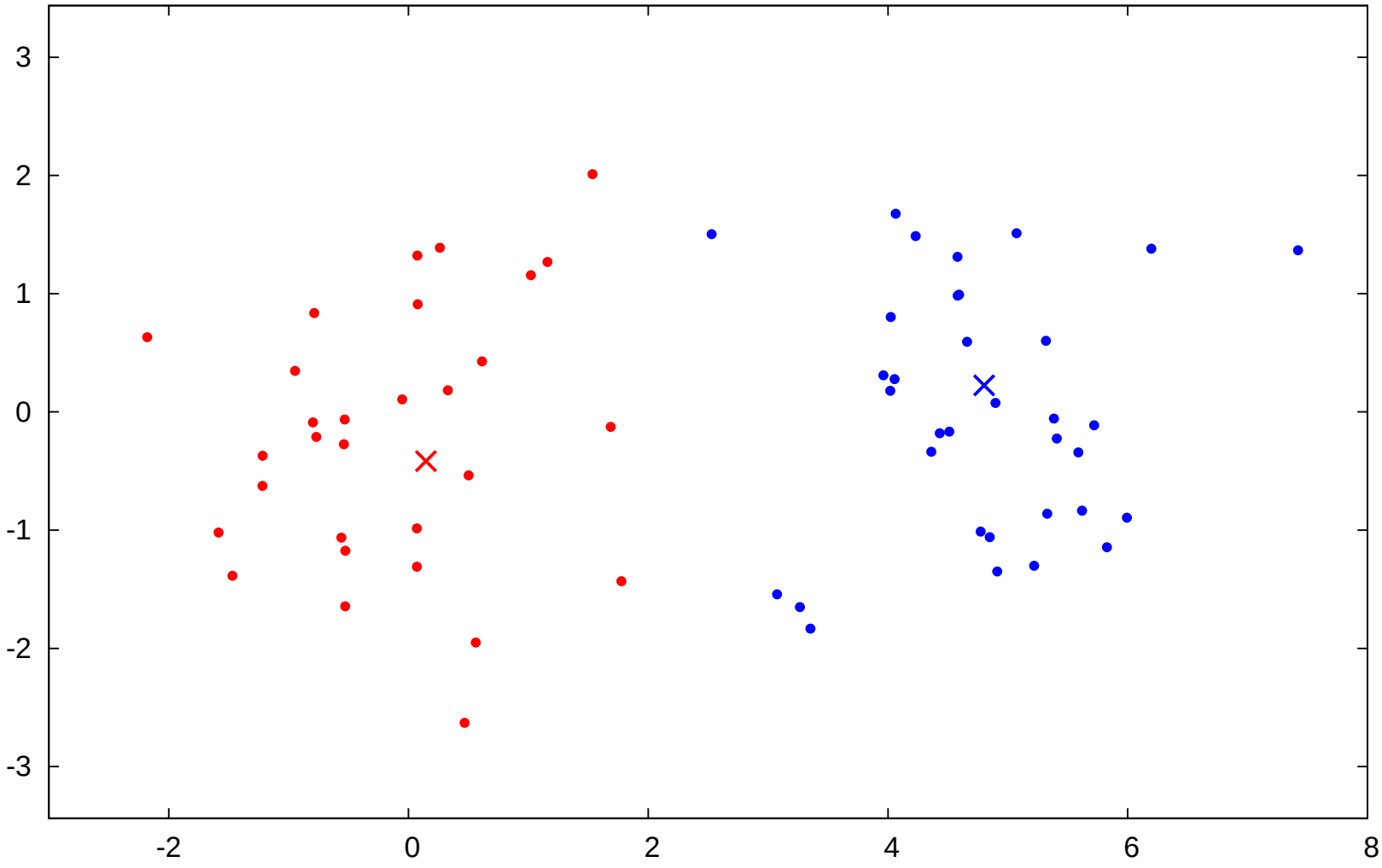
k-means clustering



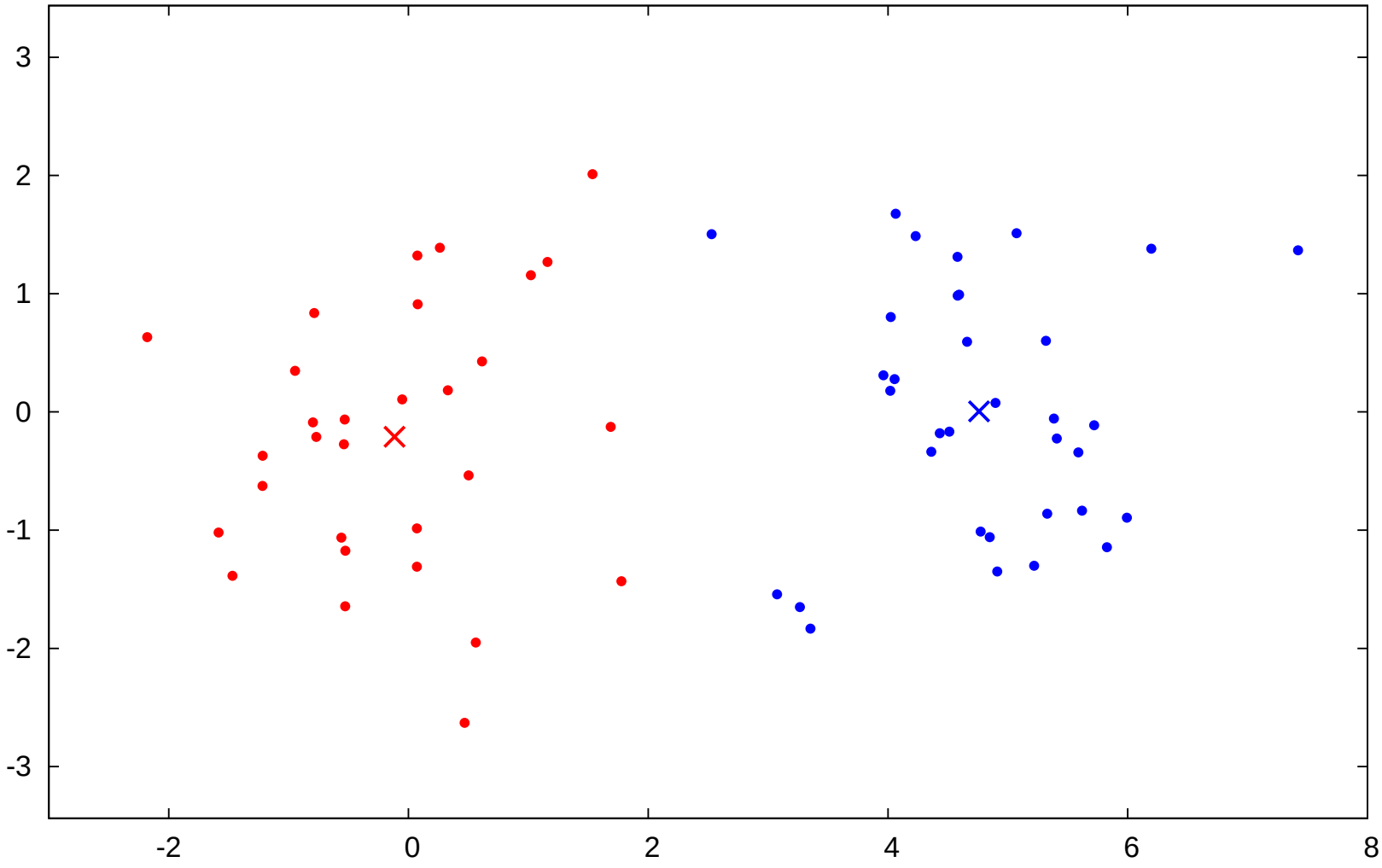
k-means clustering



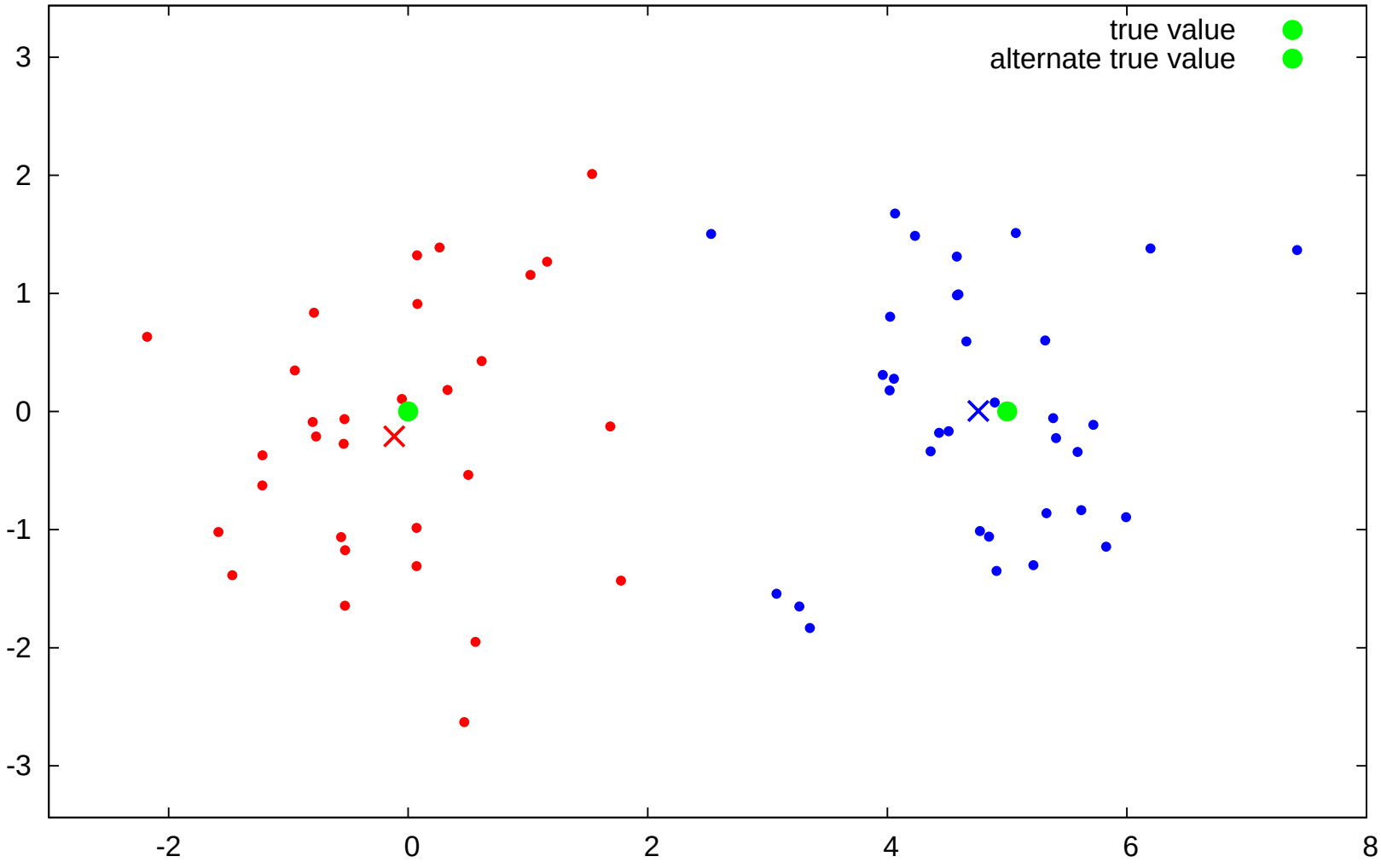
k-means clustering



k-means clustering



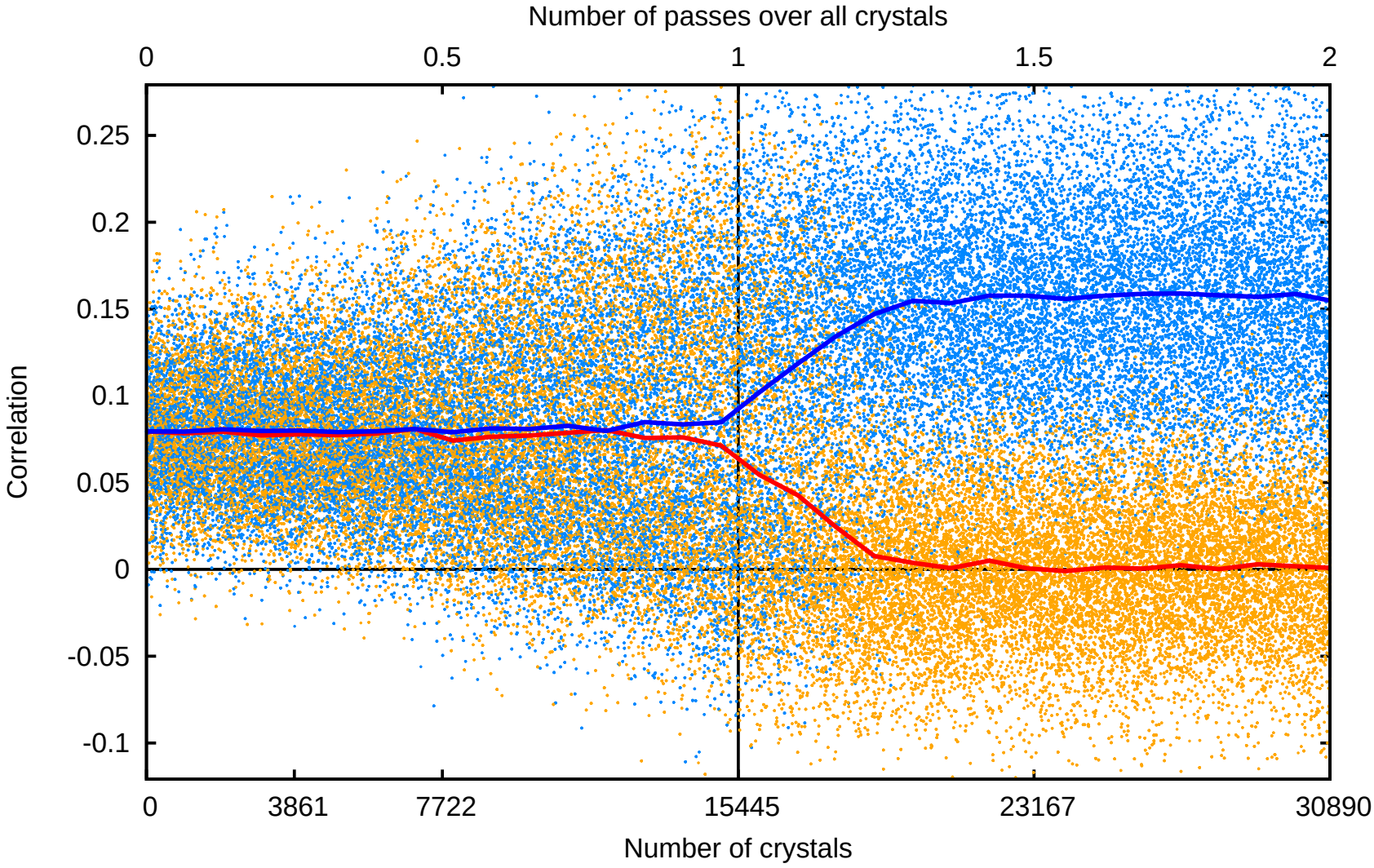
k-means clustering



Algorithm 3

1. starting from random assignments
2. reassign each shot to the group it agrees to more until convergence

- usually converges very fast, finished after reassigning every shot twice.
- basically a k-means clustering or expectation maximization algorithm



- 0.7 % wrong assignments
- 5 s until convergence on my laptop

General strategies to improve power of discrimination

- if known use twinning operators to effectively double the amount of shots
- use the point group with the highest symmetry that applies
  - merge symmetry related reflexes
  - increasing the number of reflex intensities that can be compared

Correcting the indexing ambiguity

- If twinning operator unknown: Test the possible twinning operators.
  Which one improves the agreement between two groups the most if applied to one of the two groups?
- If twinning operator is known: Apply to one of the groups to transform the shots to the same indexing mode

On the way to good data

Eliminate one systematic error after the other, the following list of systematic errors is in no way complete and ordered loosely by the size

- scaling/normalizing
- cell parameters and orientation
  - twinning - should it occur
  - partiality
  - indexing (e.g. multiple lattices)
- detector inhomogenity
- anisomorphy

Anisomorphy

The crystals differ in their physical properties.

- different conformations in which the protein can crystallize
- new kind of measurements that introduce another variable
  - time dependance
  - introduce conformational change of the protein in the crystal
  - et cetera

Acknowledgements

Kay Diederichs
Thomas White
Felix Baumann

Thank you!

Citations

- Sébastien Boutet et al. “High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography”. In: *Science* 337.6092 (2012), pp. 362–364. DOI: 10.1126/science.1217737.
- Wolfgang Brehm and Kay Diederichs. “Breaking the indexing ambiguity in serial crystallography”. In: *Acta Crystallographica Section D* 70.1 (Jan. 2014), pp. 101–109. DOI: 10.1107/S1399004713025431.
- Henry N. Chapman et al. “Femtosecond X-ray protein nanocrystallography”. In: *Nature* 470 (7332 Feb. 3, 2011), pp. 73–77. DOI: 10.1038/nature09750.
- Sir Ronald A. Fisher. *Statistical Methods for Research Workers*. Hafner Publishing Company Inc., 1954.
- Kunio Hirata et al. “Determination of damage-free crystal structure of an X-ray sensitive protein using an XFEL”. In: *Nature Methods* 11.7 (May 11, 2014), pp. 734–736. ISSN: 1548-7091. DOI: 10.1038/nmeth.2962.

Algorithm 3 in detail

n, g, k, i are element of the natural numbers
f is element of rational numbers
There are n shots that have to be grouped each group represents one combination of twinning operators

2. Every shot is assigned a random group
3. Repeat the following for every shot s until convergence
   - 3.1 Calculate the factor f that is considered the agreement from s to each group g: f=Sum over the correlation coefficients of s to the shots of each group reindexed to match the group g divided by the number of correlation coefficients in the sum
   - 3.2 Assign the shot s to the group that has the highest f
3. Reindex every shot according to its group

Answering questions