Quality of diffraction data

Kay Diederichs

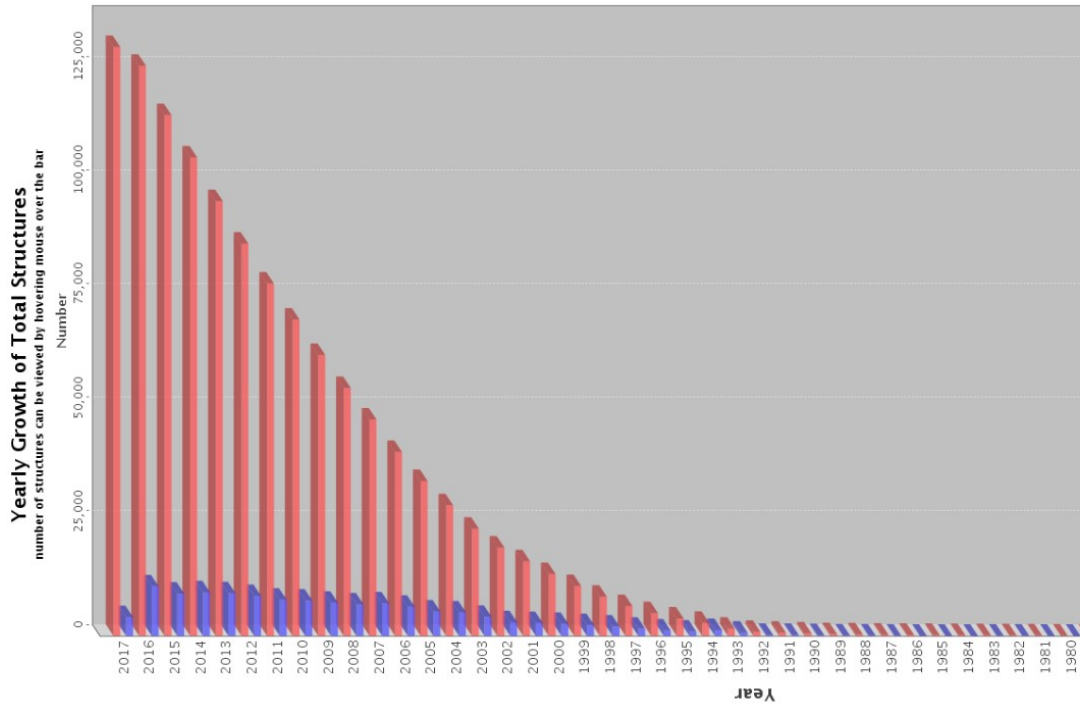
Protein Crystallography /
Molecular Bioinformatics
University of Konstanz, Germany

Outline

- Introduction and statement of problem
- 1st example: meaning of “quality”: accuracy *versus* precision
- 2nd example: measuring “quality”
- 3rd example: common misunderstandings
- Recent work: measuring non-isomorphism (i.e. systematic deviations of datasets)

Crystallography has been extremely successful

Protein Data Bank : ~133.000 entries



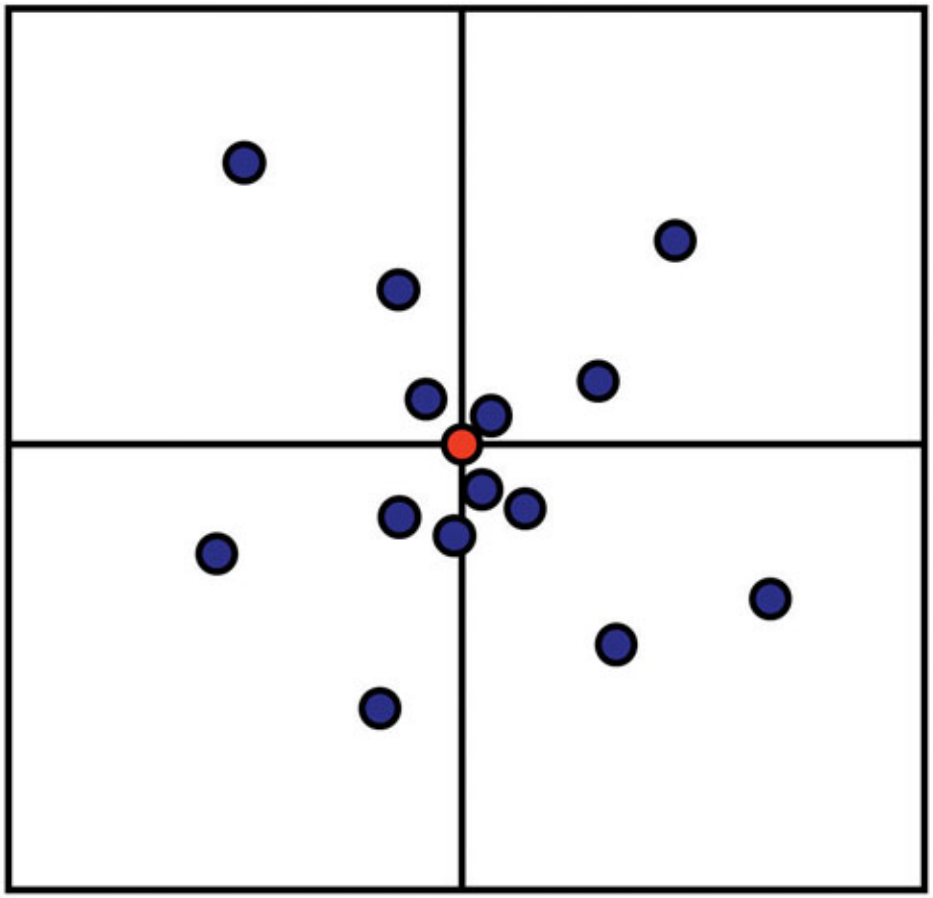
Could it be any better?

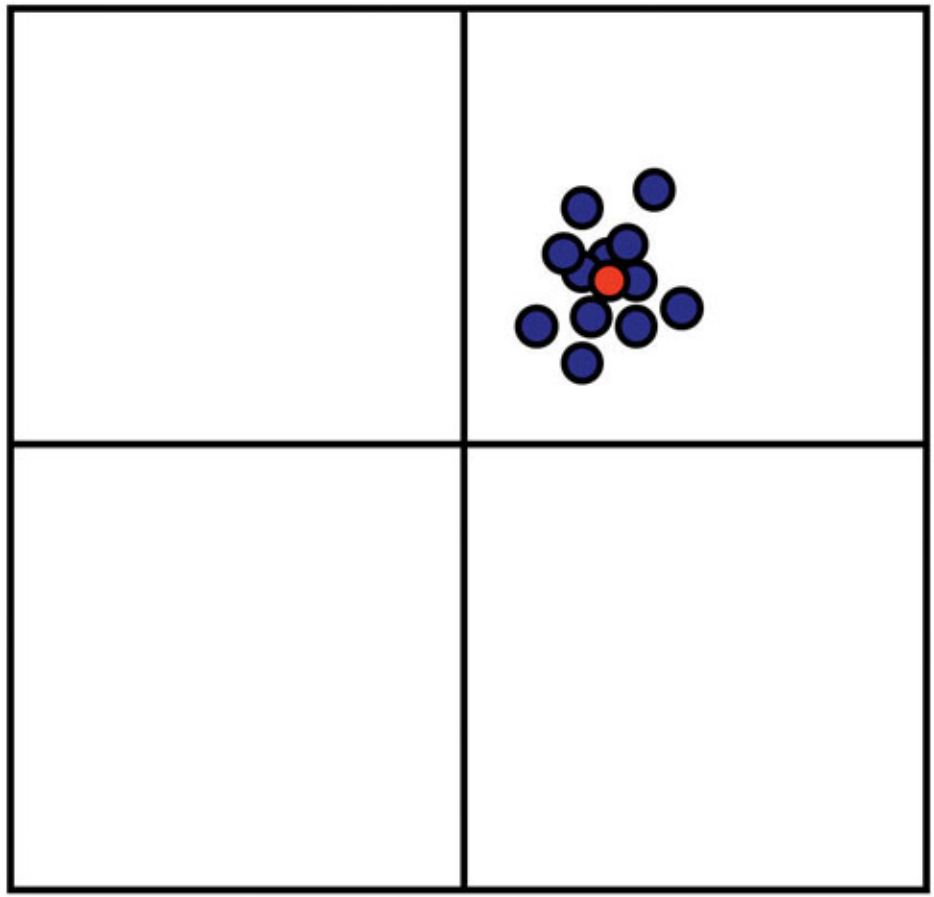
Three examples for

- *Rules* that may have been useful in the past under different circumstances, but are still commonly used today and result in wrong decisions
- *Concepts* resulting from first principles that would, if applied, deliver the information to reach the correct decision

1st example: Not understanding the difference between, and the relevance of **precision** and **accuracy**

"Quality"





B. Rupp, Biomolecular Crystallography

Accuracy – how different from the *true value*?

Precision – how different are *measurements*?

Numerical example

Repeatedly determine π=3.14... as 3.1, 3.2, 3.0 :
observations have medium precision, medium accuracy

Precision= mean relative absolute deviation from average value=
(0+0.1+0.1)/(3.1+3.2+3.0) = 2.2%

Accuracy= mean relative absolute deviation from true value:
=(|3.14-3.1| + |3.14-3.2| + |3.14-3.0|)/(3*3.14) = 2.5%

R_{merge} formula!

$$R_{merge} = \frac{\sum_{hkl}\sum_{i=1}^{n} |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl}\sum_{i=1}^{n} I_i(hkl)}$$

Repeatedly determine π=3.14... as 2.70, 2.71, 2.72 :
observations have high precision, low accuracy.

Precision= mean relative absolute deviation from average value=
(0.01+0+0.01)/(2.70+2.71+2.72) = 0.24%

Accuracy= mean relative absolute deviation from true value=
(|3.14-2.70| + |3.14-2.71| + |3.14-2.72|)/(3*3.14) = 13.7%

R_{merge} formula!

What is the “true value“?

- if only **random error** exists, accuracy = precision (on average)
- if unknown **systematic error** exists, true value cannot be found from the data themselves
- precision can easily be calculated, but not accuracy
- accuracy and precision differ by the unknown systematic error
- true values may be known from other approaches (e.g. F_{calc}^2 may be considered an estimate of the true value)

All data quality indicators estimate *precision* (only), but YOU (should) want to know *accuracy*!

➔ ***Rules***: "The data processing statistics tells me (and the reviewers!) how good my data are.
To satisfy reviewers, the indicators must be good."

- *Suboptimal result*: these rules encourage
  - overexposure of crystal to lower R_{merge}
  - data collection "strategy" with low multiplicity
  - data massaging: rejecting many "outliers", throwing away negative or weak data

➔ ***Concepts***:
  - Data processing output reports the *precision* of the data, *not* their accuracy.
  - averaging increases accuracy unless the data repeat systematic errors
  - rejecting too many data as outliers may *increase* the precision, but *decreases* accuracy!

2nd example: confusion by multitude and properties of crystallographic indicators

Confusion – what do these mean?

R_{merge}

$CC_{1/2}$

R_{sym}

Mn(I/sd)

CC_{anom}

I/σ

R_{meas}

R_{pim}

R_{cum}

Calculating the precision of unmerged (individual) observations

$\langle I_i/\sigma_i \rangle$ (σ_i from error propagation, i=individual)

$$R_{merge} = \frac{\sum_{hkl} \sum_{i=1}^{n} |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i(hkl)}$$

$$R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{i=1}^{n} |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i(hkl)}$$

$R_{meas} \sim 0.8 / \langle I_i/\sigma_i \rangle$

Calculating the precision of merged data

using the √n law of error propagation (Wikipedia "weighted arithmetic mean"):

<I/σ(I)>

$$R_{pim} = \frac{\sum_{hkl} \sqrt{\frac{1}{n-1}} \sum_{i=1}^{n} |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i(hkl)}$$

$R_{pim} \sim 0.8 / \langle I/\sigma \rangle$

by comparing averages of two randomly selected half-datasets X,Y:

H,K,L	I_i in order of measurement	Assignment to half-dataset	Average I of X	Y
1,2,3	100 110 120 90 80 100	X, X, Y, X, Y, Y	100	100
1,2,4	50 60 45 60	Y X Y X	60	47.5
1,2,5	1000 1050 1100 1200	X Y Y X	1100	1075
...				

(calculate the R-factor (D&K1997) or correlation coefficient $CC_{1/2}$ (K&D 2012) on X, Y)

Measuring the precision of merged data with a correlation coefficient

- Correlation coefficient has clear meaning and well-known statistical properties
- Significance of its value can be assessed by Student's t-test:
  e.g. CC>0.3 is significant at p=0.01 for n>100;
  CC>0.08 is significant at p=0.01 for n>1000
- Using “random half-datasets” of crystallographic intensity data: → $CC_{1/2}$
- From $CC_{1/2}$, we can analytically estimate **CC of the merged dataset against the true** (usually unmeasurable) **intensities** using

$$CC^* = \sqrt{\frac{2\,CC_{1/2}}{1 + CC_{1/2}}}$$

- (Karplus and Diederichs (2012) *Science* **336**, 1030)

- ***Rule***: "the quality of the data that I use for refinement can be assessed by R_{merge}/R_{meas} . Data with R_{merge}/R_{meas} > e.g. 60% are useless."

- Suboptimal result: Wrong indicator. Wrong high-resolution cutoff. Wrong data-collection strategy.

Concept: - use an indicator for the precision of the *merged* data if you are interested in the suitability of the data for MR, phasing and refinement.

- Use <I/σ> or <I>/<σ> (but how to calculate σ; and which cutoff??)

- Use $CC^* = \sqrt{\frac{2\,CC_{1/2}}{1+CC_{1/2}}}$ if you want to know how high (numerically) CC_{work} , CC_{free} in refinement can become (i.e. how *data quality limits model quality*):
CC_{work} larger than CC* implies overfitting, because in that case the model agrees better with the experimental data than the true signal does.

This does not work with R-values because data R-values and model R-values have different definitions!

3rd example: *improper* crystallographic reasoning

situation: data to 2.0 Å resolution

using all data: R_{work}=19%, R_{free}=24% (overall)

cut at 2.2 Å resolution: R_{work}=17%, R_{free}=23%

- ***Rule***: "The lower the R-value, the better."
„cutting at 2.2 Å is better because it gives lower R-values"

- (Potentially) suboptimal result: throwing away data.

- ***Concept***: indicators may only be compared if they refer to the *same* reflections.

Proper crystallographic reasoning

…. requires three concepts:

1. Better data allow to obtain a better model
2. A better model has a lower R_{free}, and a lower R_{free}-R_{work} gap
3. *Comparison* of model R-values is only *meaningful* when using the *same* data

Taking these together, this leads us to the *„paired refinement technique“*: compare models in terms of their R-values against the *same* data.

P.A. Karplus and K. Diederichs (2012) Linking Crystallographic Data with Model Quality. *Science* **336**, 1030-1033.

Recent work: Measuring non-isomorphism

Kinds of errors in (crystallographic/image) data -

- *Random*: mostly quantum effects (photon/electron emission/absorption)
- *Systematic*: macroscopic/experimental differences (nonlinearity, differences in absorption, conformation, composition, …)

Non-isomorphism denotes those systematic effects on measured signal that differ between individual datasets, or groups of datasets.

Crystallographic example: two forms of lysoyzme

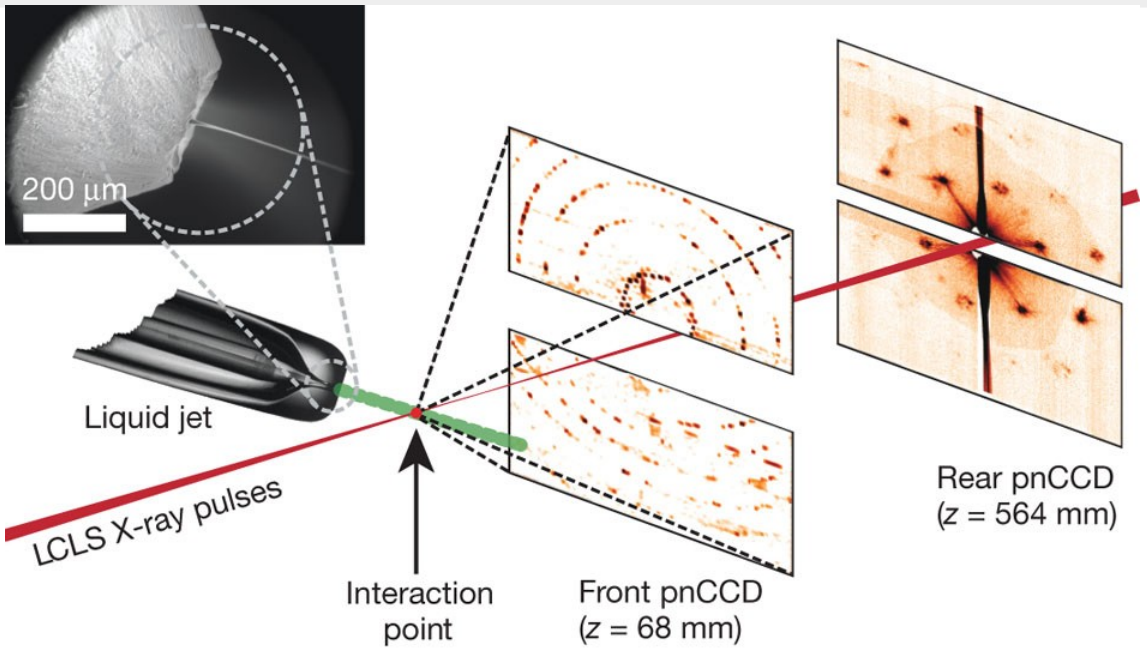
3aw6
3aw7

RH: 84.2%
vs 71.9%

RMSD = 0.18 Å Δcell = 0.7 % R_{iso} = 44.5%

Crystallography: multiple crystals/datasets

Femtosecond X-ray protein nanocrystallography
Chapman et al. (2011) *Nature* 470, 73-77
"... nanocrystals of photosystem I, one of the largest membrane protein complexes. More than 3,000,000 diffraction patterns were collected in this study, and a three-dimensional data set was assembled from individual photosystem I nanocrystals (~200 nm to 2 μm in size). …"
(15445 xtals used; Data collection at XFEL (LCLS, Stanford)



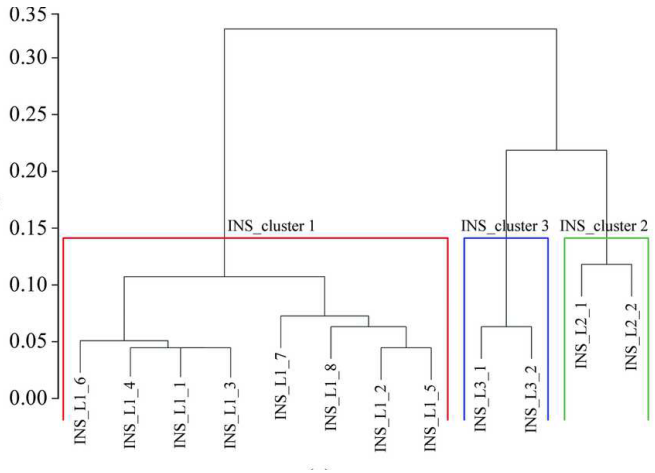
Separating random and systematic errors in data

- data-based (rather than cell-based) approach
- comparison of datasets based on pairwise correlation coefficients

$$cc_{ij} = \frac{\sum (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum (x_k - \bar{x})^2 \sum (y_k - \bar{y})^2}}$$

cc_{ij} = -1 .. 1

- hierarchical cluster analysis
  - allows no distinction between random and systematic error



Making sense of pairwise differences

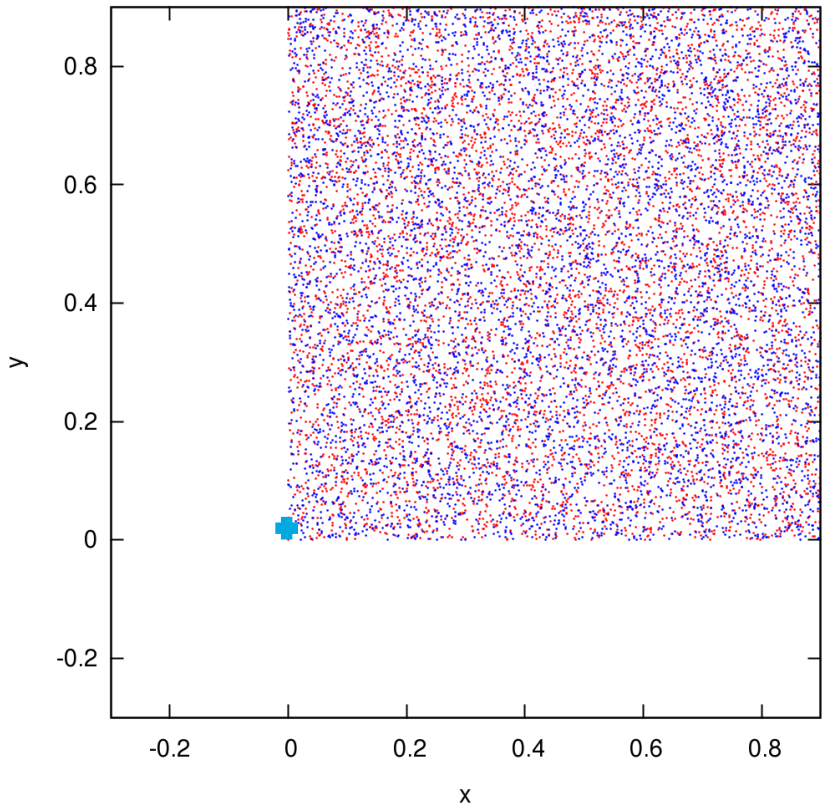
Need to separate the random error from the systematic error

- total error (difference of values that should be equal) is $\sqrt{random^2 + systematic^2}$
- pairwise CC has contributions from total error, i.e. from both sources of error
- **separation of random and systematic errors is not generally possible**

New way to analyze pairwise CCs: CC_ANALYSIS

- Brehm and Diederichs (2014) minimize $\phi(\{\vec{x}\}) = \sum_{i>j} (cc_{ij} - \vec{x}_i * \vec{x}_j)^2$ with $\{x\}=\{x_1,x_2,\ldots x_N\}$ where x_i and x_j are *N* vectors in *n*-dimensional space representing the datasets, and cc_{ij} is (Pearson's) correlation coefficient between intensities of datasets i and j
- with *n* = 2 or 4, this solves the **indexing ambiguity** (→ twinning) present in point groups 3, 4, 6, 312, 321 and 23, and additional cases with particular values of cell parameters.
- This type of analysis is called **Multidimensional Scaling**
- It turns XFEL data collection into a technique with general applicability

Least-squares iterations starting from random positions - each point represents one dataset with one of two indexing modes



Brehm, W. & Diederichs, K. (2014) Breaking the indexing ambiguity in serial crystallography. Acta Cryst. (2014). D70, 101-109

Which information can be extracted from the matrix of pairwise CCs?

The analysis (Diederichs 2017, Acta D73, 286-293) shows that …

- the least-squares solution of $\phi(\{\vec{x}\}) = \sum_{i>j} (cc_{ij} - \vec{x}_i * \vec{x}_j)^2$ exists and is "unique" if cc_{ij} known
- it can be obtained from the n Eigenvalues/Eigenvector of the cc_{ij} matrix
- the **x** vectors are arranged in a sphere with radius 1, in n-dimensional space
- vectors can be given as coordinates, or (better) length and spherical angles

Amount of signal

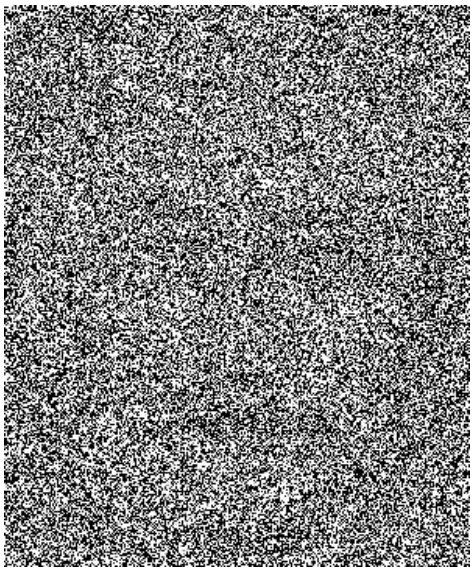
the length of a vector is CC*, the correlation with its prototype (“true”) dataset, and depends on the random error of the dataset

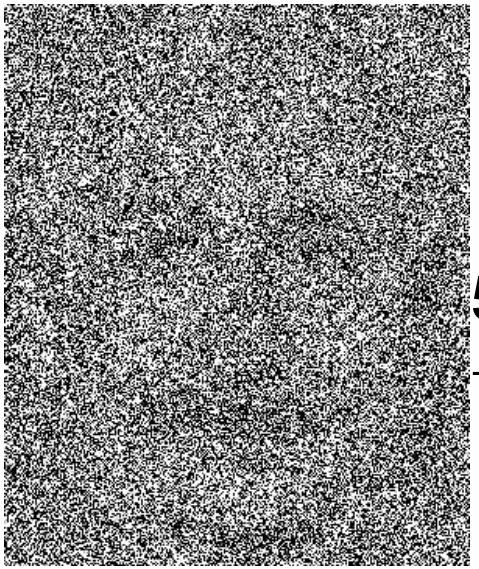
- CC* may be calculated from multiple observations in a dataset (crystallography)

Relation between datasets

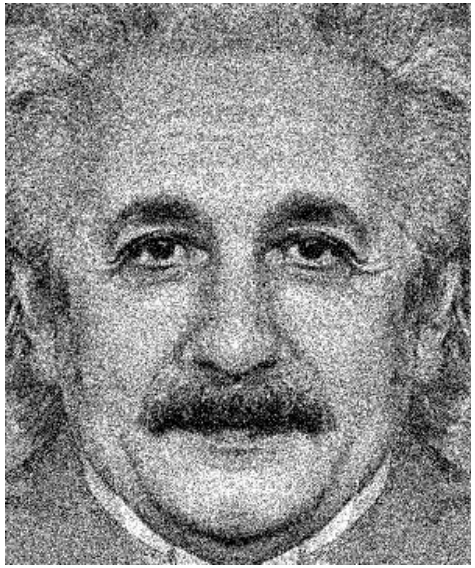
- angle between xi and x_j is proportional to the systematic difference between i and j
- $cc_{ij} = CC^*_i \cdot CC^*_j \cdot \cos(\text{angle}(x_i, x_j))$

Example: two kinds of noisy images





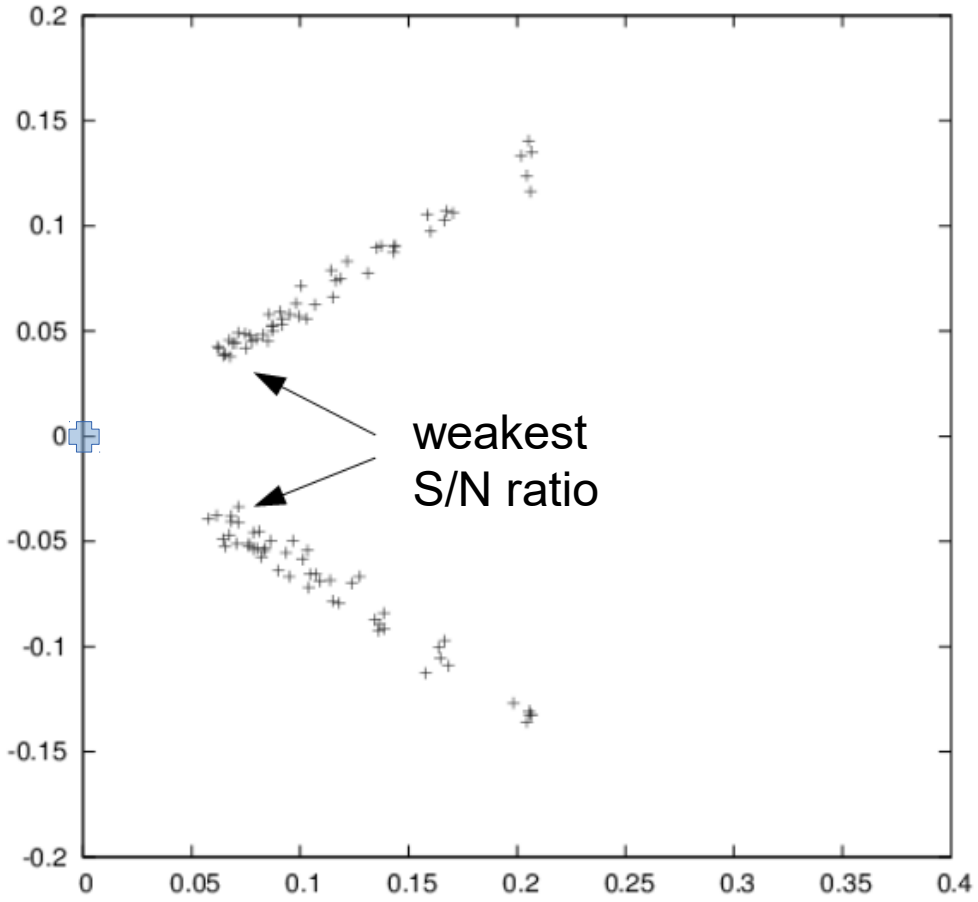
50+50
→



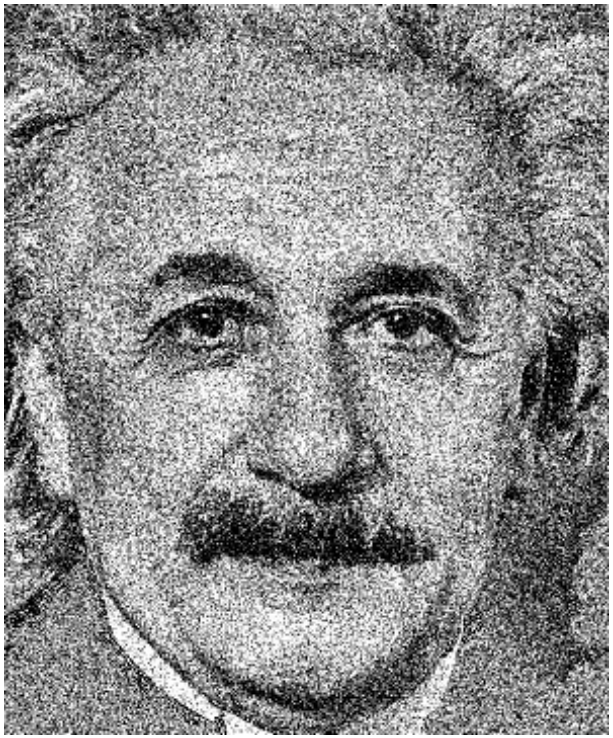
noisy images (SNR=1/13 and SNR=1/9) of original and mirror picture

Result of averaging without knowledge whether original image, or its mirror

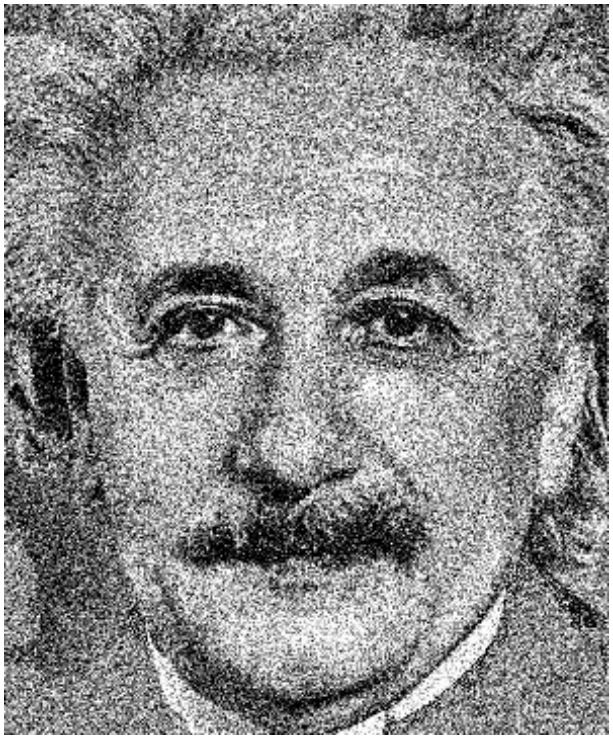
CC analysis with n=2



After clustering and *separate* averaging



original



mirror

Summary

- Crystallographic decisions are often based on *rules* of (if anything) only historical interest. These rules frequently lead to *improper shortcuts* being taken
- “make everything as simple as possible, but not simpler” (attributed to A. Einstein)
- Rules may be needed in expert systems; however, humans should rather learn, apply and further develop the underlying *concepts*
- Random and systematic differences of datasets (or images) can be separated within a simple and general framework. (Unpublished) implementations for classical and serial crystallography (XDS and CrystFEL) exist.

Thank you for your attention!

References:

Karplus, P.A. and Diederichs, K. (2015) Assessing and maximizing data quality in macromolecular crystallography. *Current Opinion in Struct.Biol.* **34**, 60-68.

Diederichs, K. (2015) Crystallographic data and model quality. in: Nucleic Acids Crystallography (Ed. E. Ennifar), Methods in Molecular Biology **1320**, 147-173.

Diederichs, K. (2017) Dissecting random and systematic differences between noisy composite data sets. *Acta Cryst*. **D***73*, 286-293.
(PDFs at https://www.biologie.uni-konstanz.de/diederichs/publications)