# *Data processing with XDS and associated programs*

Kay Diederichs


Protein Crystallography /
Molecular Bioinformatics
University of Konstanz, Germany

# Outline

I. *XDS* – the program package

II. *XDS* – how to obtain good data

III. Serial crystallography – some aspects

From tomorrow – tutorial using *your* data. I leave after my talk on tuesday so pls talk to me today or tomorrow!

throughout this talk: *program*, file

# The *XDS* program suite

Original author:
Wolfgang Kabsch
(Max-Planck-Institute
Heidelberg)

Since ~1986

I joined in 2007

# The *XDS* program suite

- *XDS*: the main program (indexing, integrating, scaling)
- *XSCALE*: scale several *XDS* intensity data sets together; zero-dose extrapolation; statistics
- *XDSCONV*: convert to other programs' formats

The following programs are independent of the *XDS* distribution:

- *XDS-Viewer* - inspect diagnostic images written by *XDS*, or (single) data frames (open source: sourceforge.net).Instead, *adxv* may be used
- *XDSSTAT* - additional statistics (not part of main distribution; download and use: see XDSwiki)
- *XDSGUI* – graphical user interface (open source: sourceforge.net) for *XDS* and *SHELX C/D/E* (since latest version)

Distribution for 64bit Linux & Mac: latest VERSION: Jan-2018 (w/ last error correction BUILT 8-Aug-2018); http://xds.mpimf-heidelberg.mpg.de/; see XDSwiki: Installation.

# interfaced to ...

- beamline software (generating XDS.INP)
- scripts: *xia2* (CCP4), *autoPROC* (Globalphasing), *xdsme* (Soleil), *autoxds* (SSRL), *autoprocess* (CMCF), ... *generate_XDS.INP* (XDSwiki), *fast_dp* (Diamond) - *please cite XDS!*
- CCP4: *pointless, xdsconv* (type CCP4_I+F, or CCP4, or CCP4_I, or CCP4_F)

# *XDSGUI*

- Simple GUI using Qt
- Adapted to the XDS philosophy
- User – extensible / modifiable commands
- Plots synchronously while processing
- Documentation and availability: XDSwiki

# Sources of information

- *XDS* main website: http://xds.mpimf-heidelberg.mpg.de - complete, accurate, up-to-date documentation; download

- XDSwiki: http://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/Main_Page

- CCP4 bulletin board

- “XDS webinar” (http://www.rigaku.com/downloads/webinars/kay-diederichs/)

- “X-ray tutorial” (Faust *et al.* JAC 2008, 2010)

- Email to kay.diederichs@uni-konstanz.de

# *XDS*wiki

- started Feb 2008; ~ 60 pages at
http://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/Main_Page
- e.g. „Optimization“; explanations of task output
- „Tips and Tricks“, „FAQ“
- „Quality Control“ with datasets and results, and links to the projects of the ACA2011 and ACA2014 „data processing“ workshop
- anybody can contribute!

(same holds for CCP4wiki: ~ 90 pages at
http://strucbio.biologie.uni-konstanz.de/ccp4wiki/index.php/Main_Page )

# *XDS* features

(just a short selection)

- 3D integration – partials/fullies implicit; fine Φ slicing
- 3D profiles of reflections are transformed into their own coordinate systems which makes them highly similar (Kabsch 1988 *J. Appl. Cryst*. **21**, 916-924)
- Smooth scaling
- Zero-dose extrapolation (*XSCALE*) can help a lot in sub-structure determination (Diederichs *et al.* 2003, *Acta Cryst.* **D59**, 903-909.)
- Fast - two levels of parallelization

# *XDS* non-features

- Old-fashioned: ASCII output to files, graphics, no mouse-over „help“ bubbles
- Not automatic, user is in full control
- Geometry not read from data frame headers
- Incomplete space-group determination – enough information to process data (see „Space group determination“ in XDSwiki)
- No support organization, XDS workshops, advertising, funding, income (Max Planck society)
- Source code not publicly available - but papers document features thoroughly

# *XDS* philosophy

(just a short selection)

- Do very little, but do it very well: gives good data
- Very robust – small molecule to ribosome
- Structure and information flow easily understandable; thoroughly documented

# Data processing steps

- set up geometric corrections: XYCORR
- find background and gain: INIT
- find reflection positions for indexing: COLSPOT
- index - establish cell and lattice): IDXREF
- [optional] strategy for data collection: XPLAN
- find shadowed areas on detector: DEFPIX
- evaluate intensities on all frames: INTEGRATE
- scale and calculate statistics: CORRECT

# Principle of *XDS* processing

- There is one JOB= line in XDS.INP which specifies a list of tasks:

JOB= XYCORR INIT COLSPOT IDXREF DEFPIX INTEGRATE CORRECT

- data reduction is divided into tasks in a **modular** way
- information storage/exchange/flow between tasks by data files which may be inspected/analyzed
- each task needs the result from the previous tasks
- fine-tuning of a task does *not* require previous tasks to be repeated
- each task writes its output file <TASK>.LP

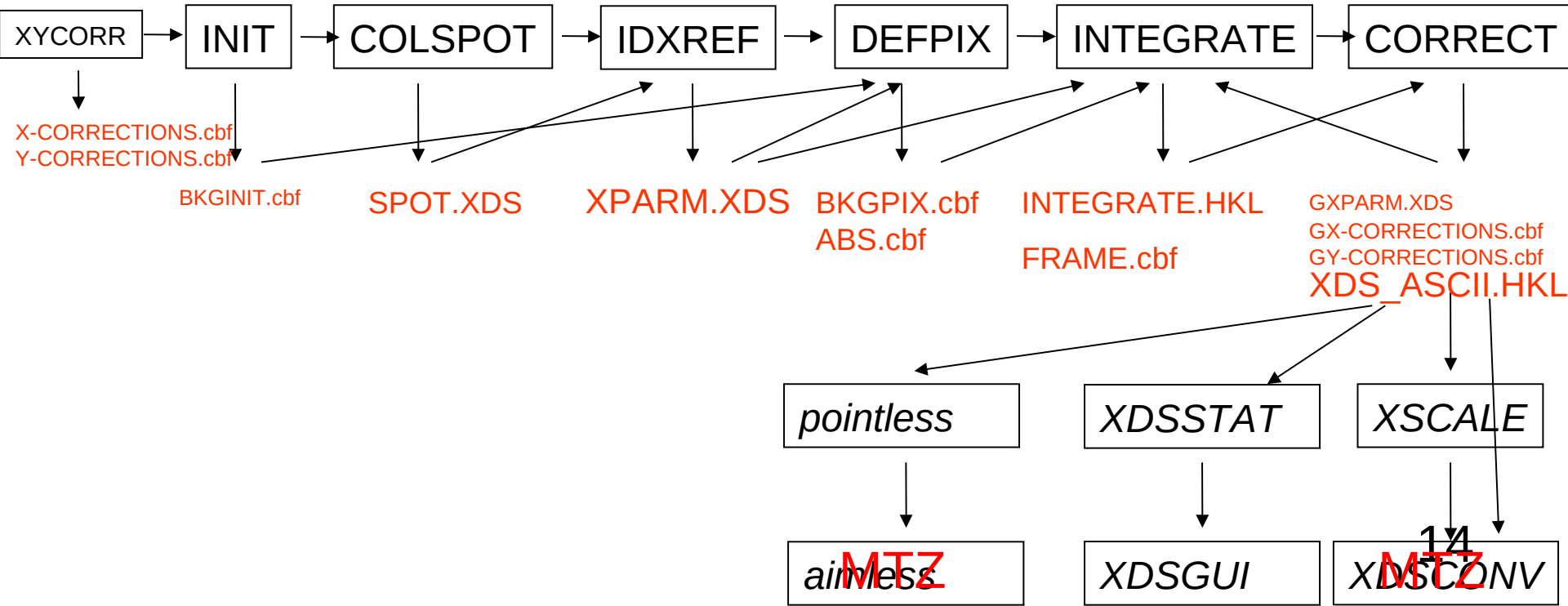
# Information flow

NAME_TEMPLATE_OF_DATA_FRAMES
DETECTOR

OSCILLATION_RANGE
SEPMIN
STRONG_PIXEL

ORGX
ORGY
DETECTOR_DISTANCE
X_RAY_WAVELENGTH
SPACE_GROUP_NUMBER

DATA_RANGE



```
!FORMAT=XDS_ASCII    MERGE=FALSE    FRIEDEL'S_LAW=TRUE
!OUTPUT_FILE=XDS_ASCII.HKL        DATE= 3-Oct-2006
!Generated by CORRECT   (XDS VERSION  August 18, 2006)
!PROFILE_FITTING= TRUE
!SPACE_GROUP_NUMBER=   92
!UNIT_CELL_CONSTANTS=    57.71    57.71   150.08  90.000  90.000  90.000
!NAME_TEMPLATE_OF_DATA_FRAMES= ../series_2_????.img
!DATA_RANGE=       1     399
!X-RAY_WAVELENGTH=  0.939010
!INCIDENT_BEAM_DIRECTION=  0.001872 -0.002230  1.064947
!FRACTION_OF_POLARIZATION=   0.980
!POLARIZATION_PLANE_NORMAL=  0.000000  1.000000  0.000000
!ROTATION_AXIS=  0.999995  0.002477 -0.001917
!OSCILLATION_RANGE=  0.500000
!STARTING_ANGLE=    30.000
!STARTING_FRAME=       1
!DETECTOR=ADSC
!DIRECTION_OF_DETECTOR_X-AXIS=   1.00000   0.00000   0.00000
!DIRECTION_OF_DETECTOR_Y-AXIS=   0.00000   1.00000   0.00000
!DETECTOR_DISTANCE=   189.286
!ORGX=   1541.25  ORGY=   1535.30
!NX=  3072  NY=  3072    QX=  0.102600  QY=  0.102600
!NUMBER_OF_ITEMS_IN_EACH_DATA_RECORD=12
!ITEM_H=1
!ITEM_K=2
!ITEM_L=3
!ITEM_IOBS=4
!ITEM_SIGMA(IOBS)=5
!ITEM_XD=6
!ITEM_YD=7
!ITEM_ZD=8
!ITEM_RLP=9
!ITEM_PEAK=10
!ITEM_CORR=11
!ITEM_PSI=12
!END_OF_HEADER
     0     0     4  4.287E-01  2.814E-01  1501.6  1514.4     99.4 0.00920 100  27   75.39
     0     0    -4  2.243E-01  2.386E-01  1587.4  1548.6     91.6 0.00920 100  30  -79.02
     0     0     5  5.976E-03  3.443E-01  1490.9  1510.2    100.4 0.01150 100  22   74.94
```

*XDS* output file:
XDS_ASCII.HKL

# II. Getting the best data

non-obvious

# How do random and systematic *error* depend on the *signal*?

random error obeys *Poisson statistics*
**error = square root of signal**

Systematic error is *proportional* to signal
**error = x * signal** (e.g. x=0.02 ... 0.10 )

(which is why James Holton calls it „fractional error“; there are exceptions)

non-obvious

# Systematic errors (noise)

- beam flicker (instability) in flux or direction
- shutter jitter
- vibration due to cryo stream
- split reflections, secondary lattice(s), ice
- absorption from crystal and loop
- radiation damage
- detector calibration and inhomogeneity; overload
- shadows on detector
- deadtime in shutterless mode
- imperfect assumptions about the experiment and its geometric parameters in the processing software
- ...

# The “error model”

Random error: $\sigma_r(I) \approx \sqrt{I}$

this is what INTEGRATE calculates

Systematic errors: $\sigma_s(I) \approx I$

this leads to deviations > $\sigma_r(I)$ between sym-related reflections

New $\sigma(I)$ estimate: $\sigma(I) = \sqrt{(a*(\sigma_r(I)^2+b*I^2))}$

with constants a,b fitted by CORRECT for the dataset

When random error vanishes (“asymptotically”), this results in $I/\sigma(I) = 1/\sqrt{(a*b)}$

# A *proxy* for good data

$(I/sigma)_{asymptotic}$=ISa (reported in CORRECT.LP) is a measure of systematic error arising from beamline, crystal, and data processing

For a given data set, ISa increases: if the geometric parameterization is improved; if the correct choice of "FRIEDEL'S_LAW=TRUE" versus "FALSE" is made; if BEAM_DIVERGENCE and REFLECTING_RANGE are correct. In short: when the experimental data are well processed

*ISa is (other than $R_{meas}$) independent of random error*

*Maximizing ISa* (good values are 30 and higher) *means minimizing systematic errors;*

*This usually also optimizes $CC_{1/2}$ at high resolution*

# III. Serial crystallography, and analysis of non-isomorphism

# Xtallography requires merging of data

Small crystals → low signal → small rotation range → multiple data sets

A) medium-sized crystals - conventional
multiple (~2-10) complete data sets (>90° rotation)

B) small crystals - serial synchrotron xtallography (SSX)
~20-1000 incomplete data sets (1-20° rotation)

C) tiny crystals - XFEL
~200-100.000 incomplete data sets (0° rotation: stills)

# When merging data, one has to worry about non-isomorphism!

- Data sets can differ in several ways
- A single value – like R-factor - does not express this
- It is useful to consider a data set as a point in a low-dimensional space

# Two types of systematic errors/differences

- Some kinds of systematic errors (e.g. beam intensity fluctuation, crystal vibration) “average out” with high multiplicity: let’s call these “isomorphous systematic errors”

- Other systematic errors (better: differences) between crystallographic data sets lead to “non-isomorphism” (*****)

  - Indexing ambiguity (occurs in almost every third project)
  - Cell parameters (which anyway are not determined precisely)
  - Radiation damage (within or between data sets)
  - Composition of crystal (ligands, HA, Se, …)
  - Conformation of molecules / loops / sidechains

***** Other fields call these systematic differences “heterogeneity”

(James Holton)

# Crystallographic example: two forms of lysoyzme

3aw6
3aw7

RH: 84.2%
vs 71.9%

RMSD = 0.18 Å

Δcell ~ 1%

$R_{iso}$ = 44.5%

# What does a correlation coefficient tell us?

$$cc_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \qquad ; \ -1 \leq cc_{xy} \leq 1$$

1) high (+ or -) correlation coefficients: data sets agree well (- : except for sign)

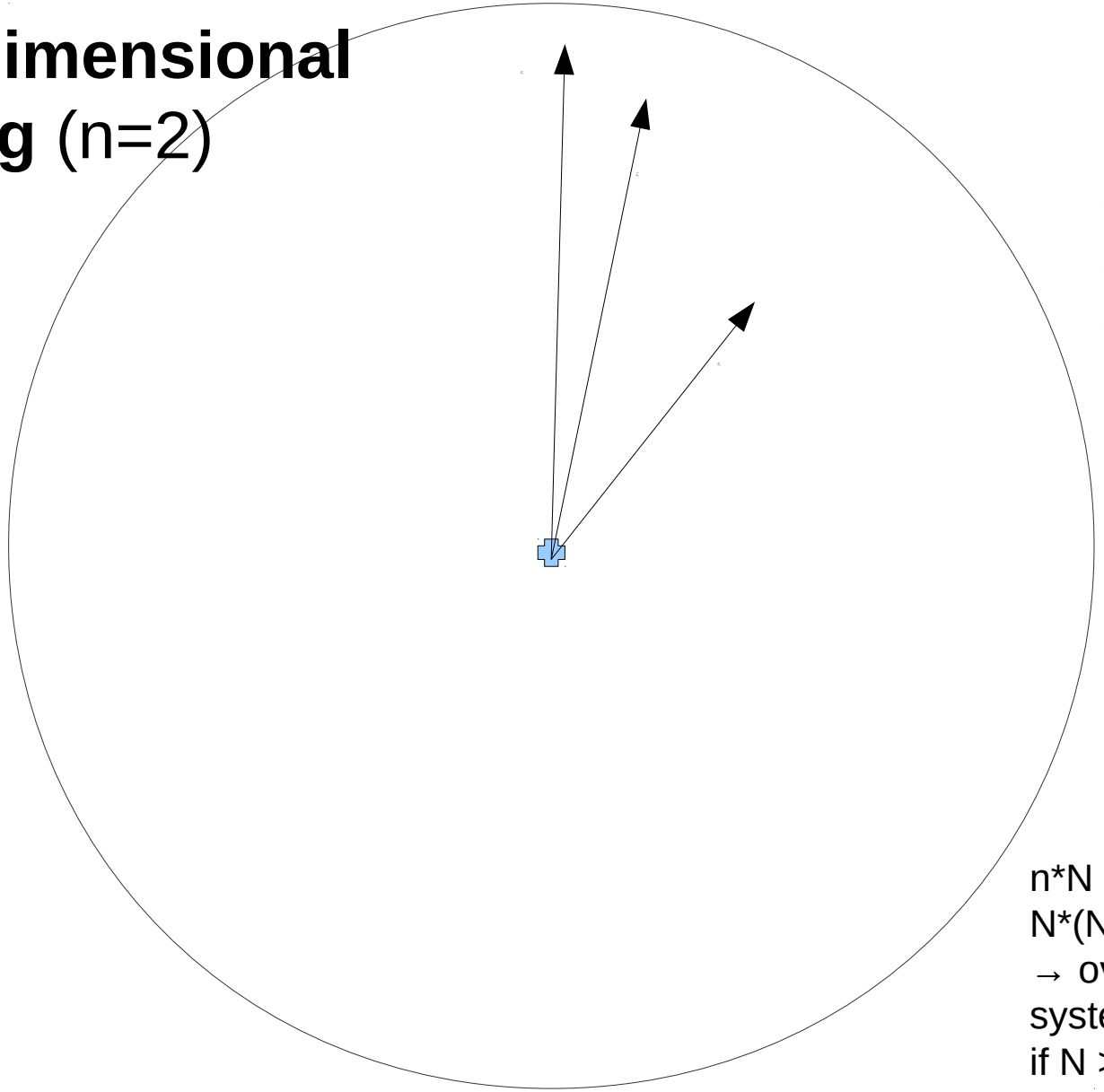
2) low correlation coefficient: data sets agree poorly

Possible reasons for 2):

- strong data, but high non-isomorphism
- weak data, and good isomorphism

*These two possibilities cannot be distinguished – a single correlation coefficient doesn't tell!*

# **Multidimensional scaling** (n=2)



$\vec{x_1} * \vec{x_2} = cc_{12}$

$\vec{x_1} * \vec{x_3} = cc_{13}$

$\vec{x_2} * \vec{x_3} = cc_{23}$

....
....
....

n*N unknowns
N*(N-1)/2 knowns
→ overdetermined system of equations if N > 2*n+1

# Analysing heterogeneity (non-isomorphism)

**It can be shown (Diederichs 2017, Acta D73, 286-293) 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/Eigenvectors of the $cc_{ij}$ matrix
- the **x** vectors are arranged in a sphere with radius 1, in n-dimensional space

**Amount of signal**

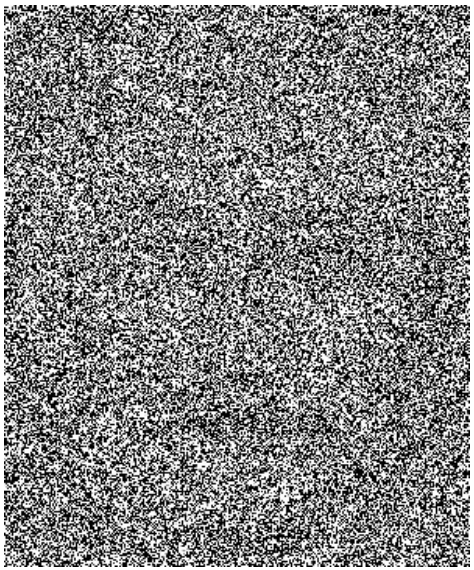
- the length of a vector is CC*, the correlation with its prototype ("true") dataset, and depends on the *random and isomorphous systematic error* of the data
- Data sets that are isomorphous have vectors with same direction
- CC* may also be calculated from $CC_{1/2}$, if multiple observations in a dataset are available

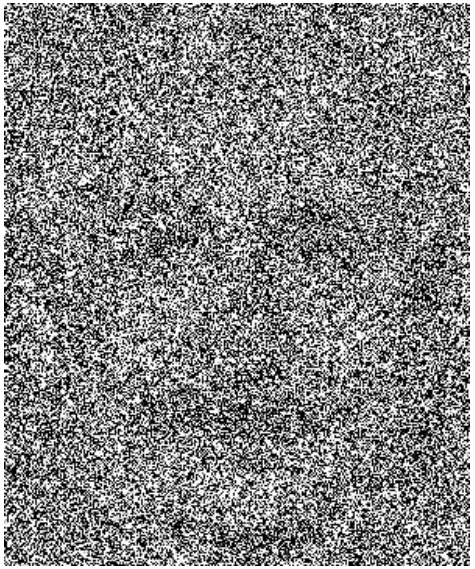
**Relation between datasets**

- angle between $x_i$ and $x_j$ : *non-isomorphous systematic difference* of i and j
- $cc_{ij} = CC^*_i \cdot CC^*_j \cdot \cos(\text{angle}(x_i, x_j))$ ("equation 5" in paper)

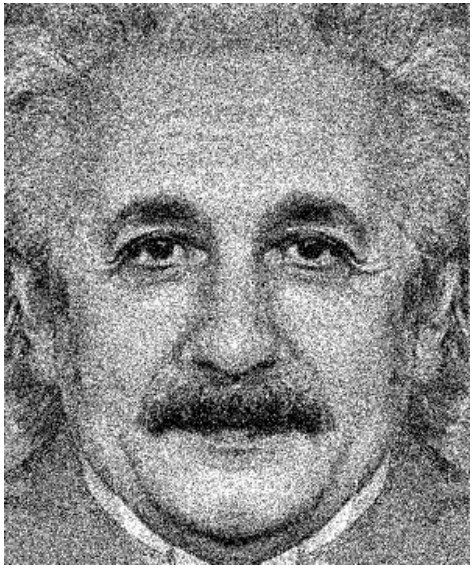
*The procedure separates the isomorphous from the non-isomorphous differences!*

# Example: two kinds of noisy images





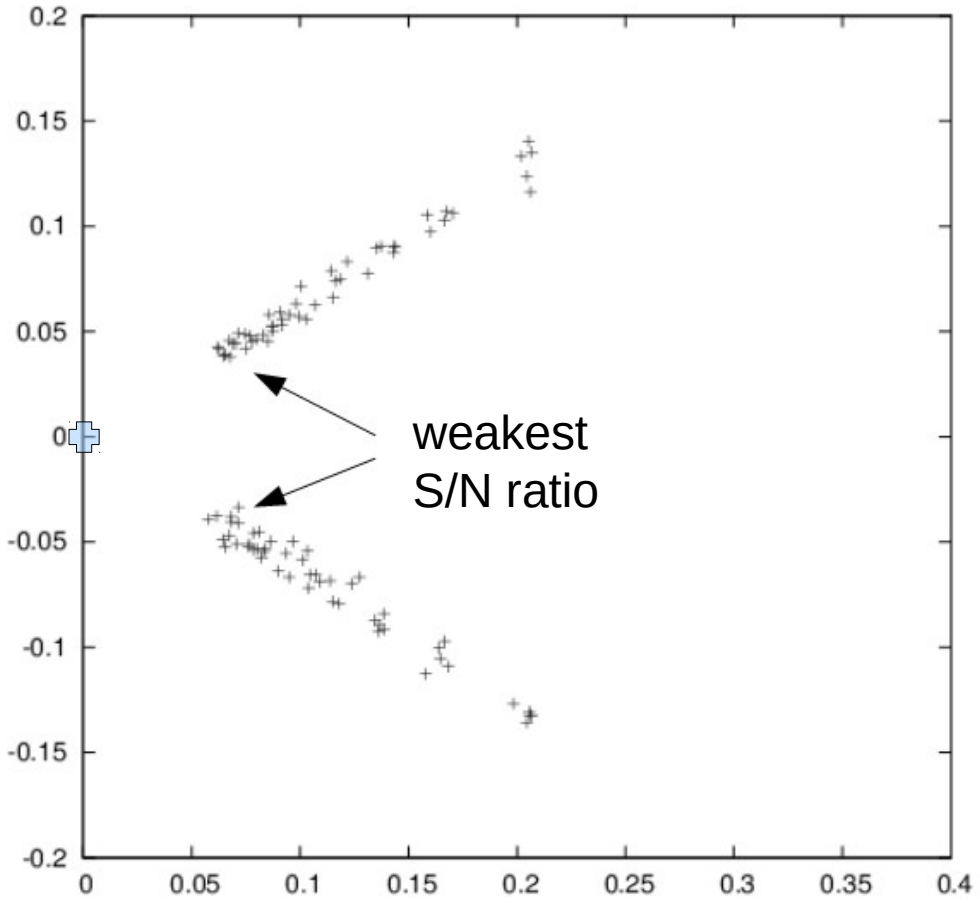
50+50
→



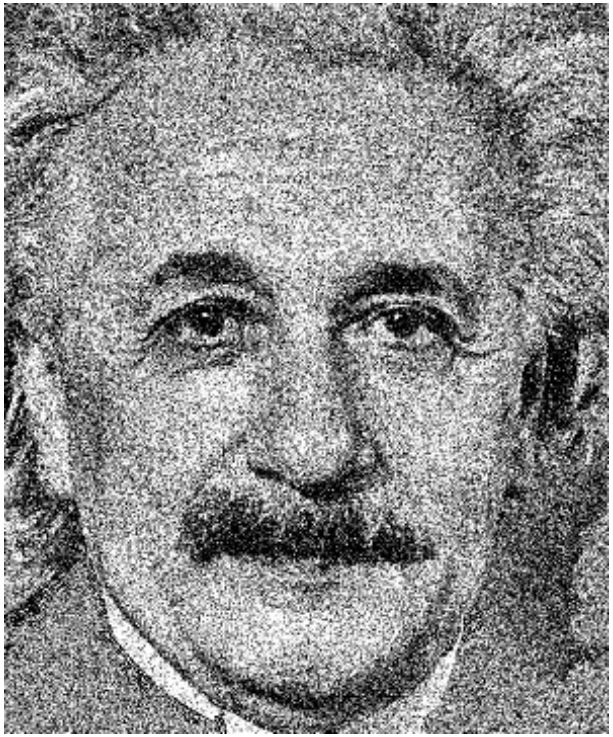
**noisy images (SNR=1/13 and SNR=1/9) of original and mirror picture**

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

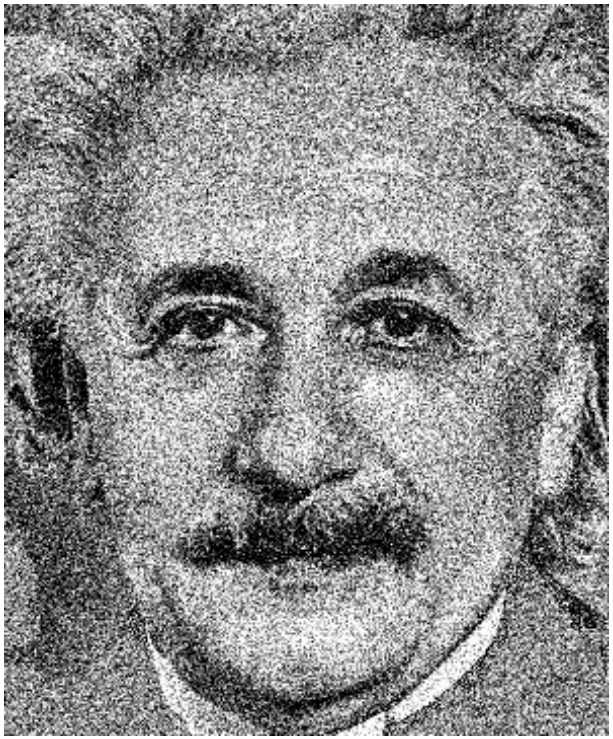
# CC analysis with n=2



# Removing the ambiguity



**Result of averaging: original**

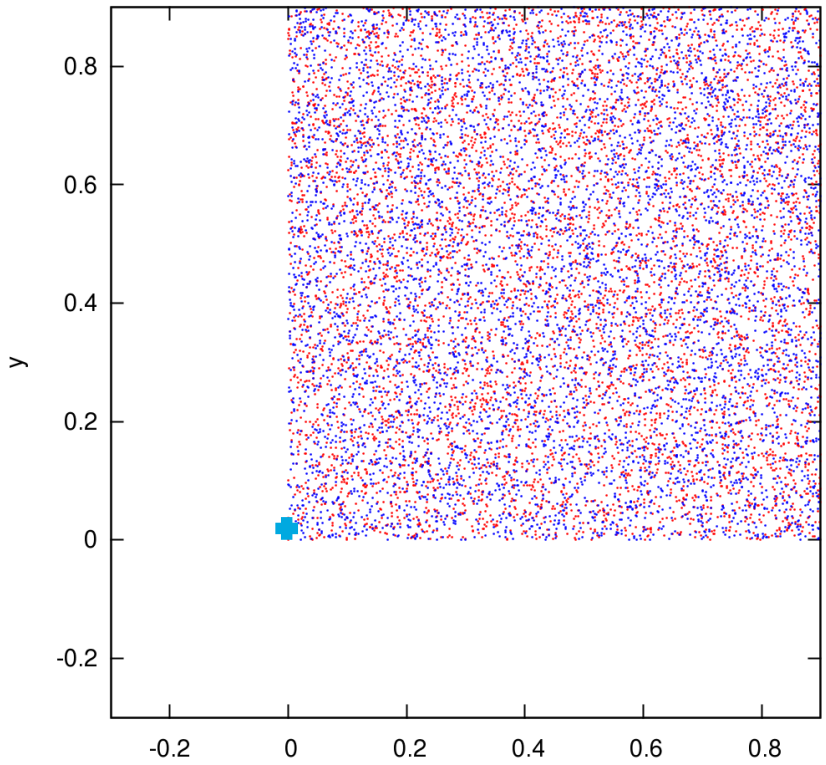


**Result of averaging: mirror**

# Programs

- **XDS** used for data processing of *n* partial data sets with common cell parameters & spacegroup
  - completeness: angular coverage / spacegroup
  - ISa: a measure of systematic error (need
  - $CC_{1/2}$: a measure of precision multiplicity!)
  - XDS_ASCII.HKL: reflection file
- **XSCALE** scales & merges all *n* XDS_ASCII.HKL files together
  - adjusts the error model to achieve consistency
  - statistics as XDS
- **XSCALE_ISOCLUSTER**: → iso.pdb; *n* atoms representing *n* data sets tentatively cluster; write XSCALE.#.INP files for clusters
- **XDSCC12**: analyze individual data sets by calculating their influence on $CC_{1/2}$ / $CC_{1/2}$ano. Optimize $CC_{1/2}$ by removing datasets that decrease $CC_{1/2}$ – this is called "delta-$CC_{1/2}$-method" (Assmann 2016)
- **XDS_NONISOMORPHISM**: analyze complete datasets using equation 2

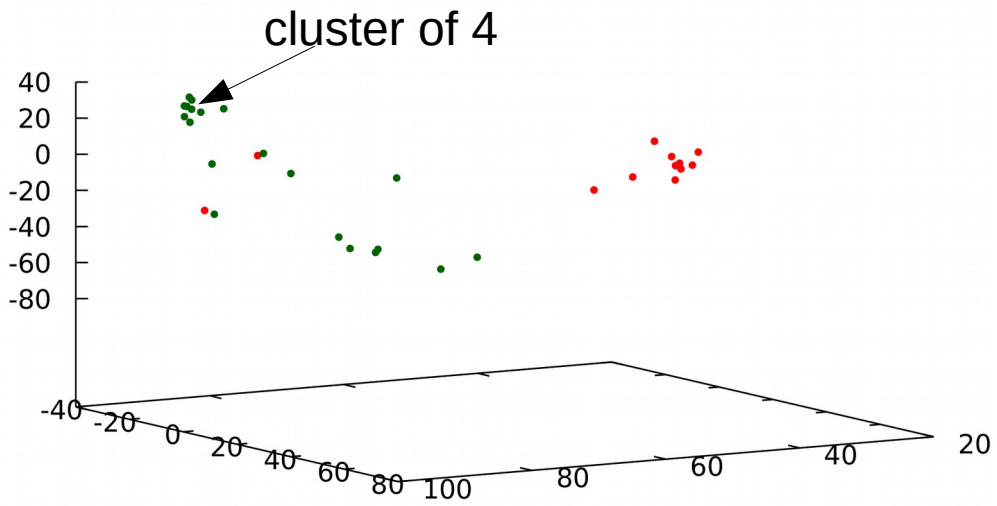
# PS-I (N=15445, n=2) [Chapman 2011]

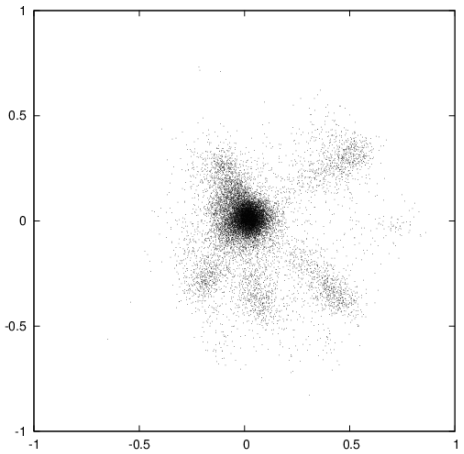



Least-squares iterations starting from random positions - each point represents a dataset with one of two indexing modes

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

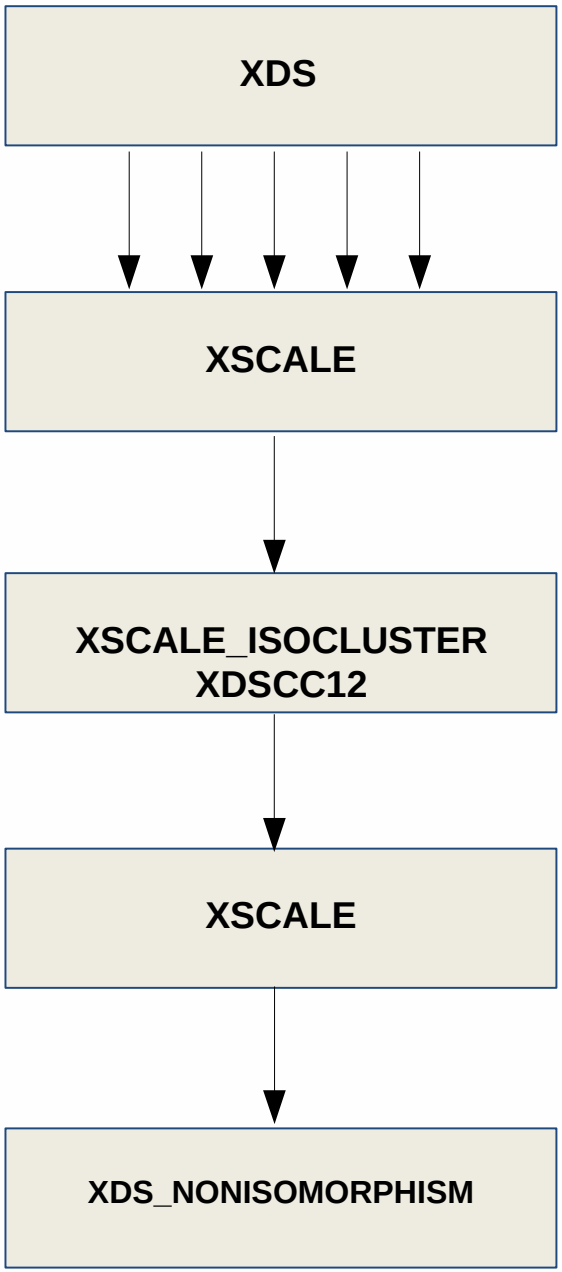
# Examples: conventional / XFEL




Data from Knott, East-Seletsky, Cofsky, Holton, Charles, O'Connell and Doundna (2017) Guide-bound structures of an RNA-targeting A-cleaving CRISPR–Cas13a enzyme. *Nature Structural & Molecular Biology* 24, 825–833

XFEL : PS-I (N=15445, n=3) *after* removing the indexing ambiguity




output:

XSCALE.HKL

XSCALE.#.INP

XSCALE.#.HKL

explanation:

first time: use all datasets
(MINIMUM_I/SIGMA=0 or 1)

separate into clusters & rank

specify cluster to work with: (e.g.)
`ln -s XSCALE.1.INP XSCALE.INP`
possibly iterate with
XSCALE_ISOCLUSTER

Use equation 2 to quantify coordinate differences between datasets

# Serial xtallography: Summary

- Analyse relations between data sets:  quantify isomorphous errors and non-isomorphism
- **Potential to identify clusters of data sets with biological meaning/relevance**
- A method that separates homogeneous (isomorphous) and heterogeneous (non-isomorphous) differences between data sets in (user-defined) low-dimensional space
- The axes of low-dimensional space are not "labelled" with specific properties
- Applications in Structural Biology: crystallographic datasets, or images of molecules
- (Analogously applicable in other Sciences: spectra, expression patterns, sequences, ...
  – wherever correlation coefficients between noisy data sets are available)

# References


Kabsch, W. (2010) *XDS*. and Integration, scaling, space-group assignment and post-refinement. *Acta Cryst.* **D66**, 124-132 and 133-144 (open access)

Diederichs K. (2009) Simulation of X-ray frames from macromolecular crystals using a ray-tracing approach. *Acta Cryst.* **D65**, 535-42

Diederichs K. (2010) Quantifying instrument errors in macromolecular X-ray datasets. *Acta Cryst.* **D66**, 733-740

Diederichs K., "Crystallographic data and model quality" in *Nucleic Acids Crystallography* (Ed. Ennifar), Methods in Molecular Biology (Springer 2015)

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

Diederichs, K., Wang, M. (2017) Serial Synchrotron X-Ray Crystallography (SSX). Methods in Molecular Biology **1607**, 239-272.

Assmann, G., Brehm, W. and Diederichs, K. (2016) Identification of rogue datasets in serial crystallography. *J. Appl. Cryst.* **49**, 1021-1028.

Huang, C.Y., *et al.* (2016) In meso in situ serial X-ray crystallography of soluble and membrane proteins at cryogenic temperatures. *Acta Cryst.* **D72**, 93-112.

Huang, C.Y., *et al.* (2015) In meso in situ serial X-ray crystallography of soluble and membrane proteins. **Acta Cryst. D71**, 1238-1256.

Diederichs, K. (2017) Dissecting random and systematic differences between noisy composite data sets. *Acta Cryst.* **D73**, 286-293.