Data processing with XDS and associated programs

Kay Diederichs

Protein Crystallography /
Molecular Bioinformatics
University of Konstanz, Germany

Outline

I. Overview of program package
II. *XDS* – process rotation data
III. *XSCALE* – scale data
IV. Error model
V. *XSCALE_ISOCLUSTER*, *XDSCC12* - enable serial crystallography; analyse non-isomorphism

Tutorial: *XDS*, *XSCALE*, *XDSGUI* ...

Processing of students' data

throughout this talk: *program*, file

The *XDS* program suite

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

Since ~1986

I joined in 2007

I. The *XDS*+ programs

- *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

Distribution for 64bit Linux & Mac: latest VERSION: Jan 31, 2020 (BUILT=20200131 small fixes and features); http://xds.mpimf-heidelberg.mpg.de/

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 and ARCIMBOLDO* (since latest version)
- *XSCALE_ISOCLUSTER*, *XDSCC12* – which data sets to re-index and merge?

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 if it is used!*
- CCP4: *pointless, xdsconv* (type CCP4_I+F, or CCP4, or CCP4_I, or CCP4_F)
- SHELX: *shelxc* reads XDS_ASCII.HKL

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 CCP4 wiki: ~ 90 pages at
http://strucbio.biologie.uni-konstanz.de/ccp4wiki/index.php/Main_Page)

XDS features

(just a short selection)

- Autoindexing (Kabsch 1988 *J. Appl. Cryst*. **21**, 67-72)
- 3D profiles of reflections are transformed into their own coordinate systems which makes them highly similar (Kabsch 1988 *J. Appl. Cryst*. **21**, 916-924)
- 3D integration – partials/fullies implicit; fine Φ slicing
- 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: no graphics, just (text) output to files
- Nothing automatic, user is in full control
- No frame header reading
- Incomplete space-group determination: translational component not automatic - use *POINTLESS* (also through *XDSGUI*) or learn the simple rules („Space group determination“ in XDSwiki)
- No support organization or XDS workshops - thanks to CCP4, I'm here!
- Code is not open source, but WK‘s papers document features thoroughly

XDS philosophy

(just a short selection)

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

II. Using *XDS*

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

The tasks are ...

- XYCORR : write positional correction files (X-CORRECTIONS.cbf, Y-CORRECTIONS.cbf)
- INIT : find background pixels (defaults usually OK)
- COLSPOT: find reflection positions
- IDXREF : "index" reflections; user may supply/choose spacegroup
- XPLAN [not required] : strategy for data collection
- DEFPIX : mask shadows on detector (use XDSGUI!)
- INTEGRATE : evaluates intensities on all frames, writes INTEGRATE.HKL and FRAME.cbf
- CORRECT : scales, rejects outliers, statistics, writes scaled, unmerged XDS_ASCII.HKL (and other files)

Example XDS.INP

```
JOB= XYCORR INIT COLSPOT IDXREF DEFPIX INTEGRATE CORRECT
ORGX=1546 ORGY=1552        !Detector origin (pixels); e.g. NX/2 NY/2
DETECTOR_DISTANCE=180      !(mm)
OSCILLATION_RANGE=0.50     !degrees (>0)
X-RAY_WAVELENGTH=0.980243 !Angstroem
NAME_TEMPLATE_OF_DATA_FRAMES=frms/wga2-27_1_???.img
DATA_RANGE=1 360           !Numbers of first and last data image collected
BACKGROUND_RANGE=1 10      !Numbers of first and last data image for background
SPACE_GROUP_NUMBER=  19    !0 for unknown crystals; cell constants are ignored.
UNIT_CELL_CONSTANTS= 44.4   86.4   104.5  90 90  90  ! not required if spgr=0
REFINE(IDXREF)=BEAM AXIS ORIENTATION CELL POSITION
REFINE(INTEGRATE)=DISTANCE BEAM ORIENTATION CELL ! AXIS
ROTATION_AXIS= 1.0 0.0 0.0
INCIDENT_BEAM_DIRECTION=0.0 0.0 1.0
FRACTION_OF_POLARIZATION=0.99                    ! SLS X06SA
POLARIZATION_PLANE_NORMAL= 0.0 1.0 0.0
DETECTOR=CCDCHESS        MINIMUM_VALID_PIXEL_VALUE=1     OVERLOAD=65000
DIRECTION_OF_DETECTOR_X-AXIS= 1.0 0.0 0.0
DIRECTION_OF_DETECTOR_Y-AXIS= 0.0 1.0 0.0
VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS= 7000 30000 !Used by DEFPIX
                                    !for excluding shaded parts of the detector.
INCLUDE_RESOLUTION_RANGE=50.0 1.3 !Angstroem; used by DEFPIX,INTEGRATE,CORRECT
```

Bold keyword/parameter pairs are required. Complete documentation at http://xds.mpimf-heidelberg.mpg.de/html_doc/xds_parameters.html

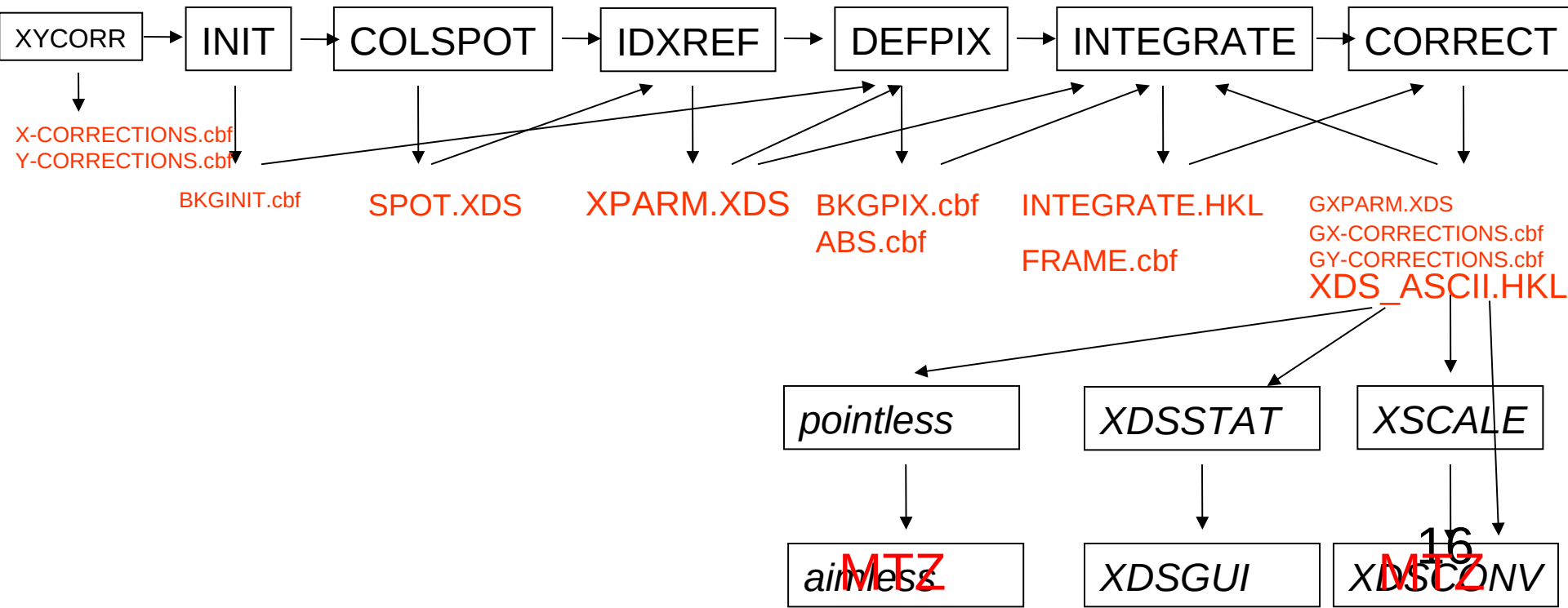
Information flow

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

III. Using *XSCALE*

- stand-alone scaling program (similar to *AIMLESS*, *SCALEPACK, DIALS.SCALE*)
- one (or more if several wavelengths) OUTPUT_FILE(s); optionally specify resolution shells and other option
- one or several (up to 1000s) INPUT_FILE(s) in XDS_ASCII.HKL format; optionally re-index, select resolution limits, ...

XSCALE.INP:

```
OUTPUT_FILE=dhdps.hkl
SAVE_CORRECTION_IMAGES=FALSE
! PRINT_CORRELATIONS=FALSE
INPUT_FILE=*../../../2018/DATA_FOR_KAY_FROM_SANTOSH/DHDPS_REN/c3_1/XDS_ASCII.HKL
INPUT_FILE= ../../../2018/DATA_FOR_KAY_FROM_SANTOSH/DHDPS_REN/c1_1/XDS_ASCII.HKL
REIDX_ISET=-1 0 0 0   0 -1 0  0  0 0 1 0
INPUT_FILE= ../../../2018/DATA_FOR_KAY_FROM_SANTOSH/DHDPS_REN/c1_2/XDS_ASCII.HKL
INPUT_FILE= ../../../2018/DATA_FOR_KAY_FROM_SANTOSH/DHDPS_REN/c1_3/XDS_ASCII.HKL
INPUT_FILE= ../../../2018/DATA_FOR_KAY_FROM_SANTOSH/DHDPS_REN/c2_1/XDS_ASCII.HKL
REIDX_ISET=-1 0 0 0   0 -1 0  0  0 0 1 0
INPUT_FILE= ../../../2018/DATA_FOR_KAY_FROM_SANTOSH/DHDPS_REN/c2_2/XDS_ASCII.HKL
REIDX_ISET=-1 0 0 0   0 -1 0  0  0 0 1 0
```

XSCALE.LP

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected	COMPARED	I/SIGMA	R-meas	CC(1/2)	Anomal Corr	SigAno	Nano
8.39	142620	1657	1660	99.8%	5.5%	6.8%	142620	106.33	5.6%	100.0*	94*	4.021	705
5.93	278547	3019	3019	100.0%	6.4%	7.4%	278547	96.87	6.4%	100.0*	79*	2.461	1384
4.84	357331	3879	3879	100.0%	7.1%	7.6%	357331	93.15	7.1%	100.0*	67*	2.034	1811
4.19	421292	4638	4638	100.0%	7.0%	7.3%	421292	98.61	7.0%	100.0*	52*	1.626	2191
3.75	471903	5224	5224	100.0%	7.8%	7.6%	471903	91.79	7.8%	100.0*	44*	1.458	2486
3.42	524016	5802	5802	100.0%	8.8%	8.2%	524016	85.27	8.9%	100.0*	37*	1.373	2770
3.17	571733	6296	6296	100.0%	10.3%	9.4%	571733	75.44	10.3%	100.0*	28*	1.258	3028
2.96	615612	6767	6767	100.0%	11.9%	11.6%	615612	63.61	11.9%	100.0*	19*	1.115	3249
2.80	657246	7227	7227	100.0%	14.7%	15.3%	657246	51.61	14.8%	99.9*	23*	1.064	3486
2.65	691472	7634	7634	100.0%	17.1%	18.9%	691472	44.60	17.2%	99.9*	15*	0.959	3689
2.53	720646	7991	7991	100.0%	20.5%	23.6%	720646	38.66	20.6%	99.9*	16*	0.936	3868
2.42	754494	8402	8403	100.0%	26.0%	30.8%	754494	32.09	26.1%	99.9*	10*	0.864	4072
2.33	774815	8675	8675	100.0%	31.0%	37.3%	774815	28.24	31.1%	99.8*	8	0.824	4213
2.24	775006	9075	9075	100.0%	37.3%	45.2%	775006	23.93	37.5%	99.8*	7	0.798	4407
2.17	659048	9486	9486	100.0%	34.6%	37.5%	659048	19.82	34.8%	99.6*	5	0.798	4615
2.10	652274	9700	9700	100.0%	41.0%	44.2%	652274	16.36	41.3%	99.4*	6	0.796	4723
2.03	671365	10023	10023	100.0%	49.7%	53.8%	671365	13.48	50.1%	99.2*	3	0.757	4882
1.98	670962	10327	10327	100.0%	68.0%	74.0%	670962	9.73	68.6%	98.4*	2	0.774	5038
1.92	472516	10617	10617	100.0%	83.2%	90.8%	472516	6.45	84.2%	96.1*	-1	0.739	5178
1.88	144587	9747	10941	89.1%	99.0%	107.5%	144392	2.96	102.5%	78.6*	2	0.717	4565
total	11027485	146186	147384	99.2%	13.3%	14.2%	11027290	38.31	13.4%	100.0*	15*	1.021	70360

IV. Error model – getting the best data

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)

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: $\sigma_s(I)$ is *proportional* to 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

V. Analysis of non-isomorphism (my own research; general i.e. not XDS-related)

Xtallography requires merging of data

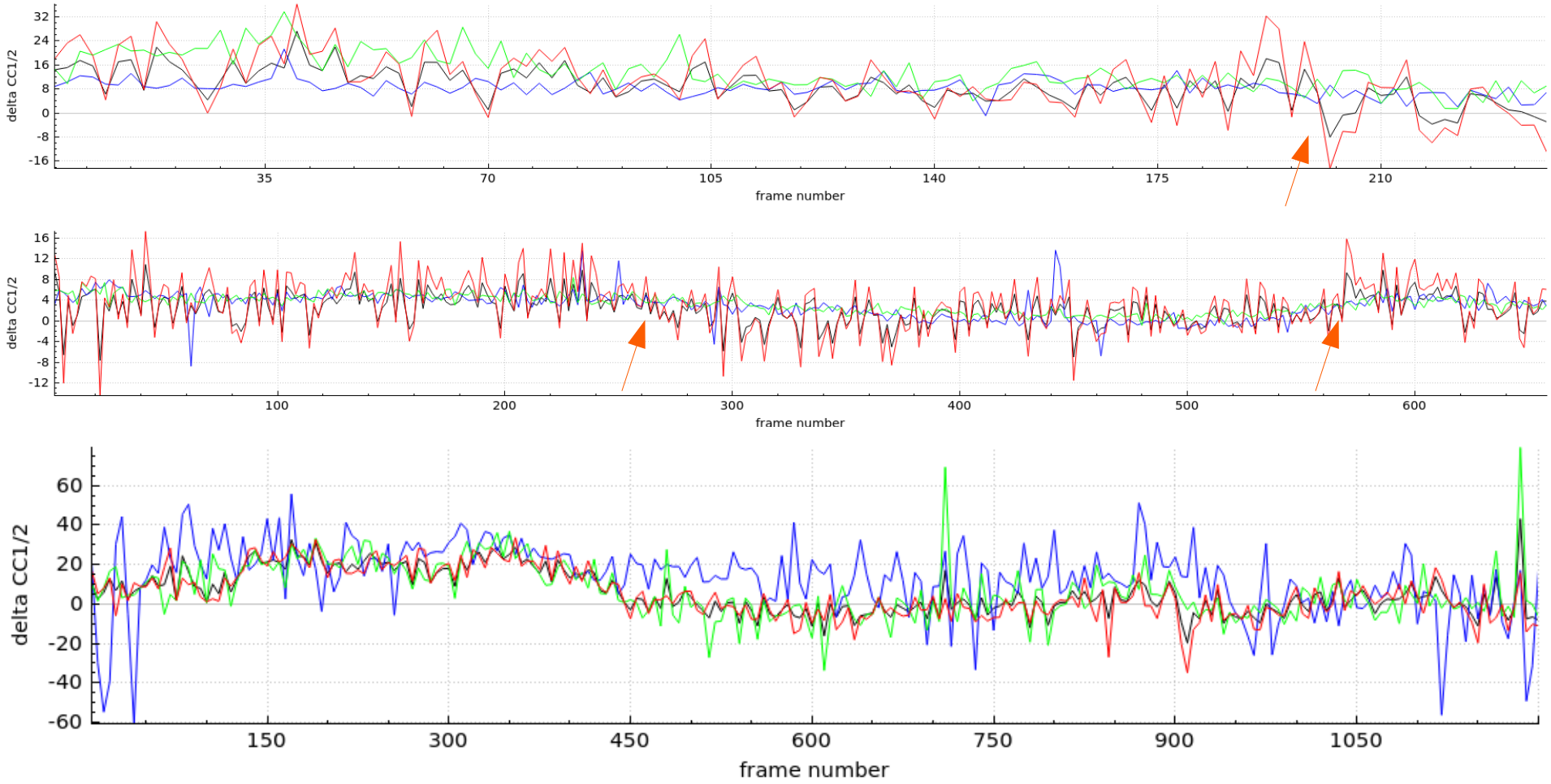
Small crystals → low signal → small rotation range → multiple data sets

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)

XDSCC12: calculates $\Delta CC_{1/2,i} = CC_{1/2,with\_i} - CC_{1/2,without\_i}$



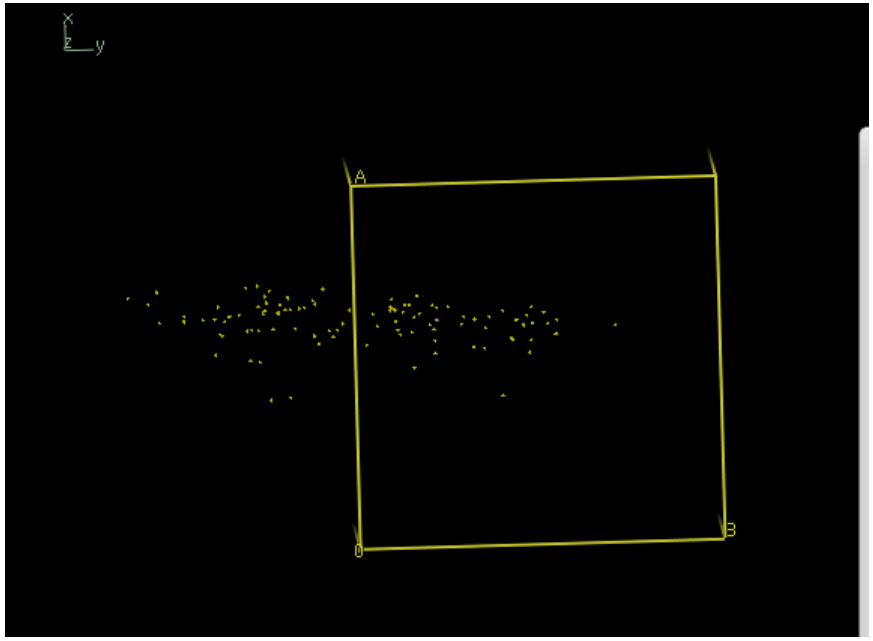
Examples of single data sets (plots from XDSGUI)

three resolution ranges (blue=low green=medium red=high) - i refers to batches of width 1°

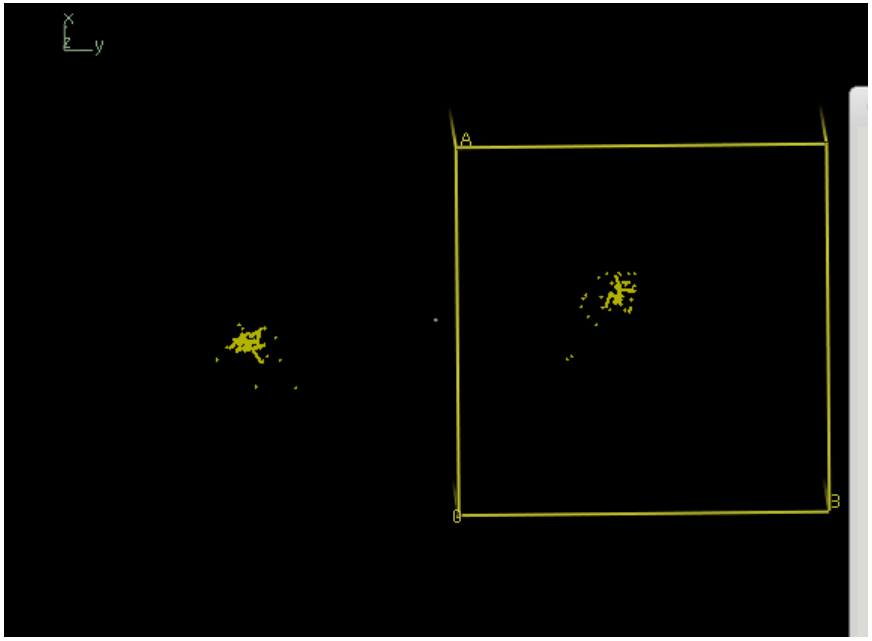
- find bad frame ranges
- radiation damage

Different talk: XDSCC12 for selection of data sets

Serial synchrotron crystallography data - James Holton's „Micro-focus Data Processing Challenge“



Original indexing suffering from over-merging (422): cc analysis shows large systematic differences



Two clusters after resolving the indexing ambiguity (222) using XSCALE_ISOCLUSTER (n=3): Structure can be solved by S-SAD

data are from http://bl831.als.lbl.gov/~jamesh/ challenge/microfocus/; complete processing pipeline is in SSX article of XDSwiki

Thank you!

(obtain PDF from kay.diederichs@uni-konstanz.de)