A Novel Statistic for Radiation Damage Analysis

ACA 2012

Graeme Winter

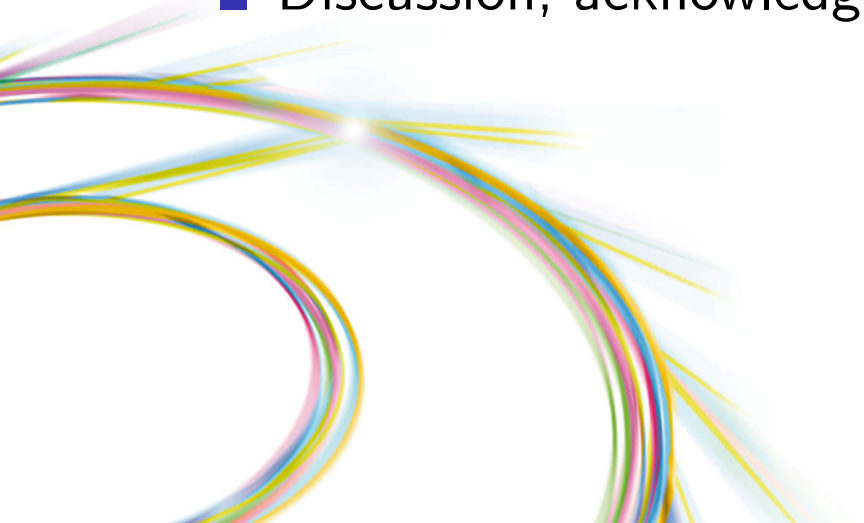
Diamond Light Source

July 2012

diamond

Overview

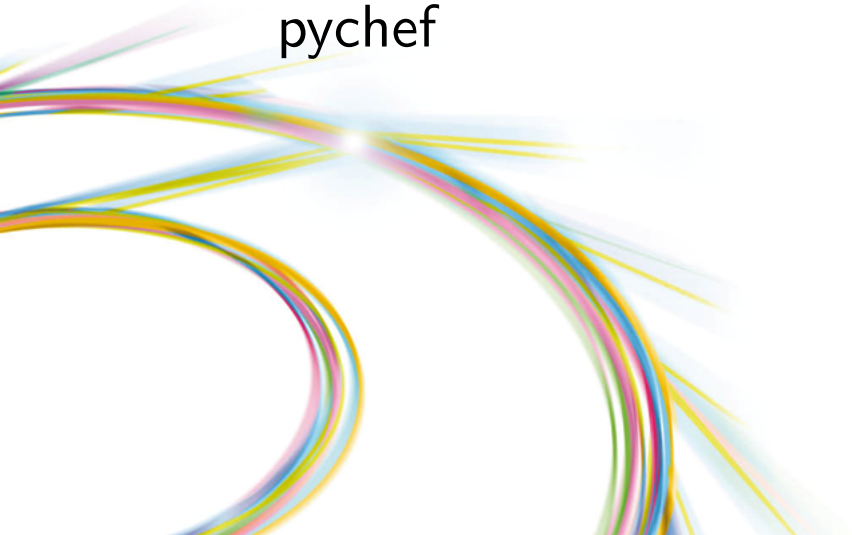
- Strategy, background, caveats
- Review of statistics
- Example cases
- New statistic - $R_{cp}(d)$
- More example cases
- Discussion, acknowledgements



diamond

Caveats

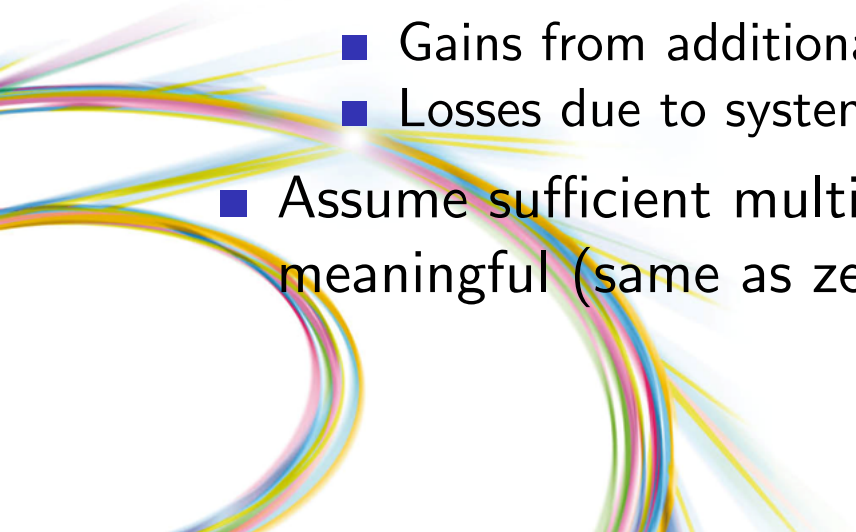
- All data described processed with xia2 / XDS i.e. using "standard methods"
- All this will show is a new tool and some instances where it could be useful
- All calculations independent of data analysis program used
- Program to perform these calculations included with xia2: pychef



diamond

Strategy

- Radiation damage gives rise to lots of changes
- Will only discuss changes to measured *intensities* i.e. not
  - Changes to unit cell
  - Sample B-factor
  - Vanishing diffraction
- Assume sufficient data for scaling, analyse after corrections applied, no assumptions about scaling *program*
- Looking to answer the question: would this data set have been more useful if I stopped collecting data earlier - balancing:
  - Gains from additional measurements i.e. multiplicity
  - Losses due to systematic changes reducing signal
- Assume sufficient multiplicity that above question is meaningful (same as zero-dose)



diamond

R_{merge}

$$R_{\mathrm{merge}}(i) = \frac{\sum_{\underline{h}} \sum_{j:i=\mathrm{image}} |I_{\underline{h}j} - \bar{I}_{\underline{h}}|}{\sum_{\underline{h}} \sum_{j:i=\mathrm{image}} I_{\underline{h}j}} \quad (1)$$

Measures: how well reflections on frame i agree with the average values of those reflections. This is reported by Scala.

diamond

R_{d}, Diederichs (2006)

$$R_d = \frac{\sum_{\underline{h}} \sum_{|b_j - b_i| = d} |I_{\underline{h}j} - I_{\underline{h}i}|}{\sum_{\underline{h}} \sum_{|b_j - b_i| = d} \frac{1}{2} |I_{\underline{h}j} + I_{\underline{h}i}|} \tag{2}$$

Measures: how well reflections separated by d frames agree. This is reported by XDSSTAT.

diamond

Application to Example Sets

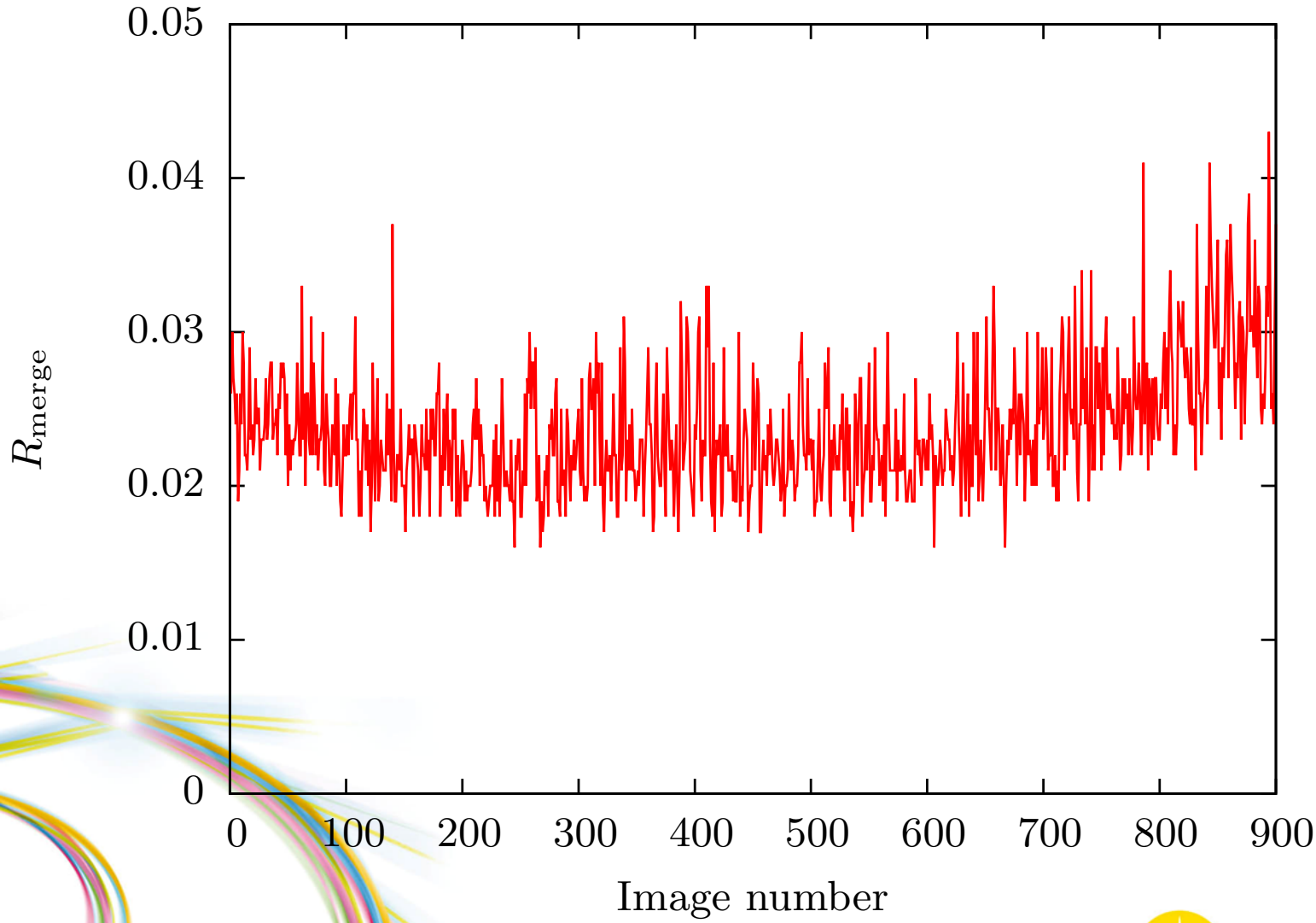
- Straightforward thaumatin example
- Radiation-damaged SAD
- Radiation-damaged MAD

diamond

Example 1: thaumatin

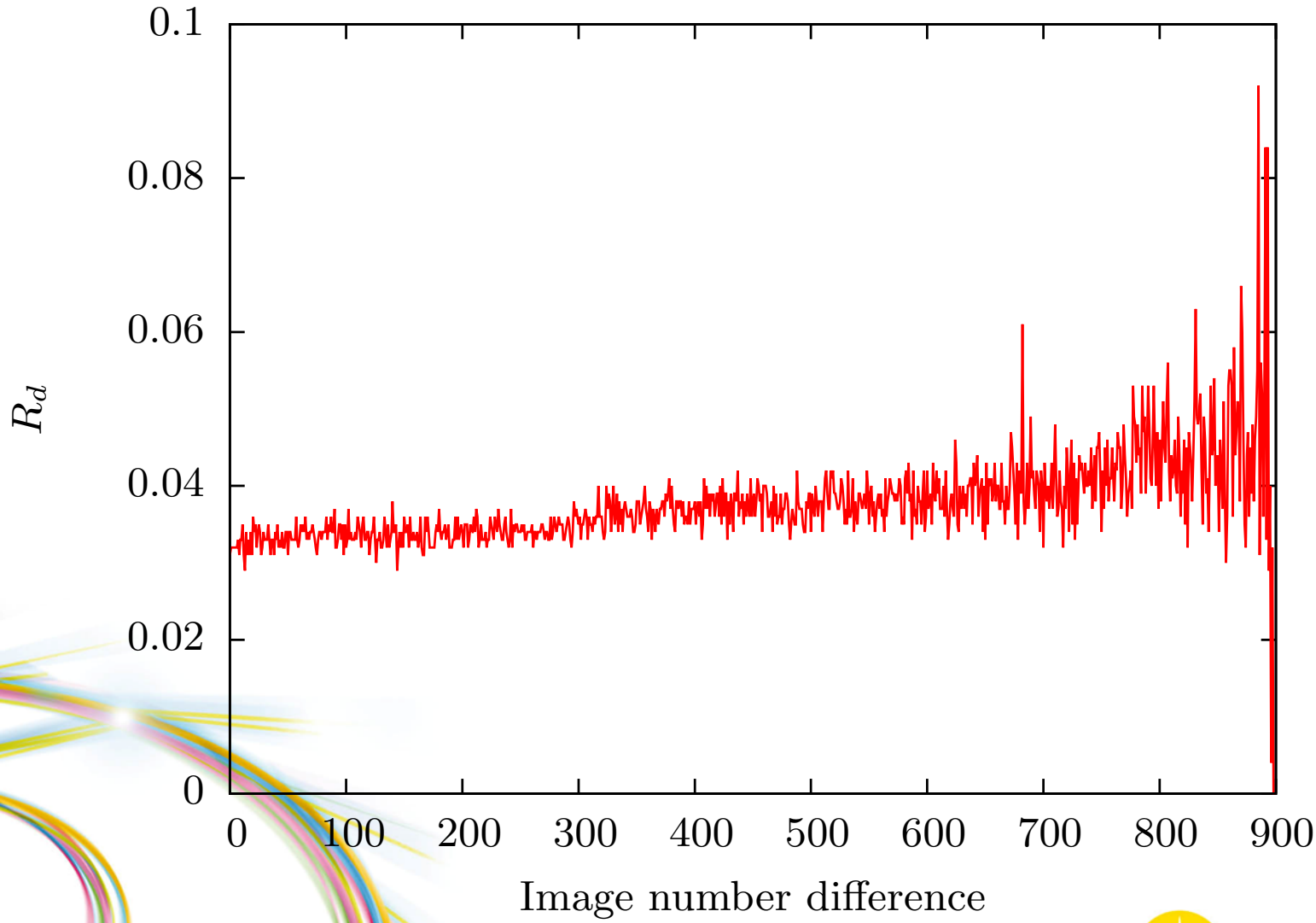
- A dull, native, good-quality sample
- Recorded during beamline setup at Diamond I03
- 900 frames $\times 0.1^\circ$
- Purpose: illustrate properties of existing residuals

Thaumatin: R_{merge}



diamond

Thaumatin: R_d

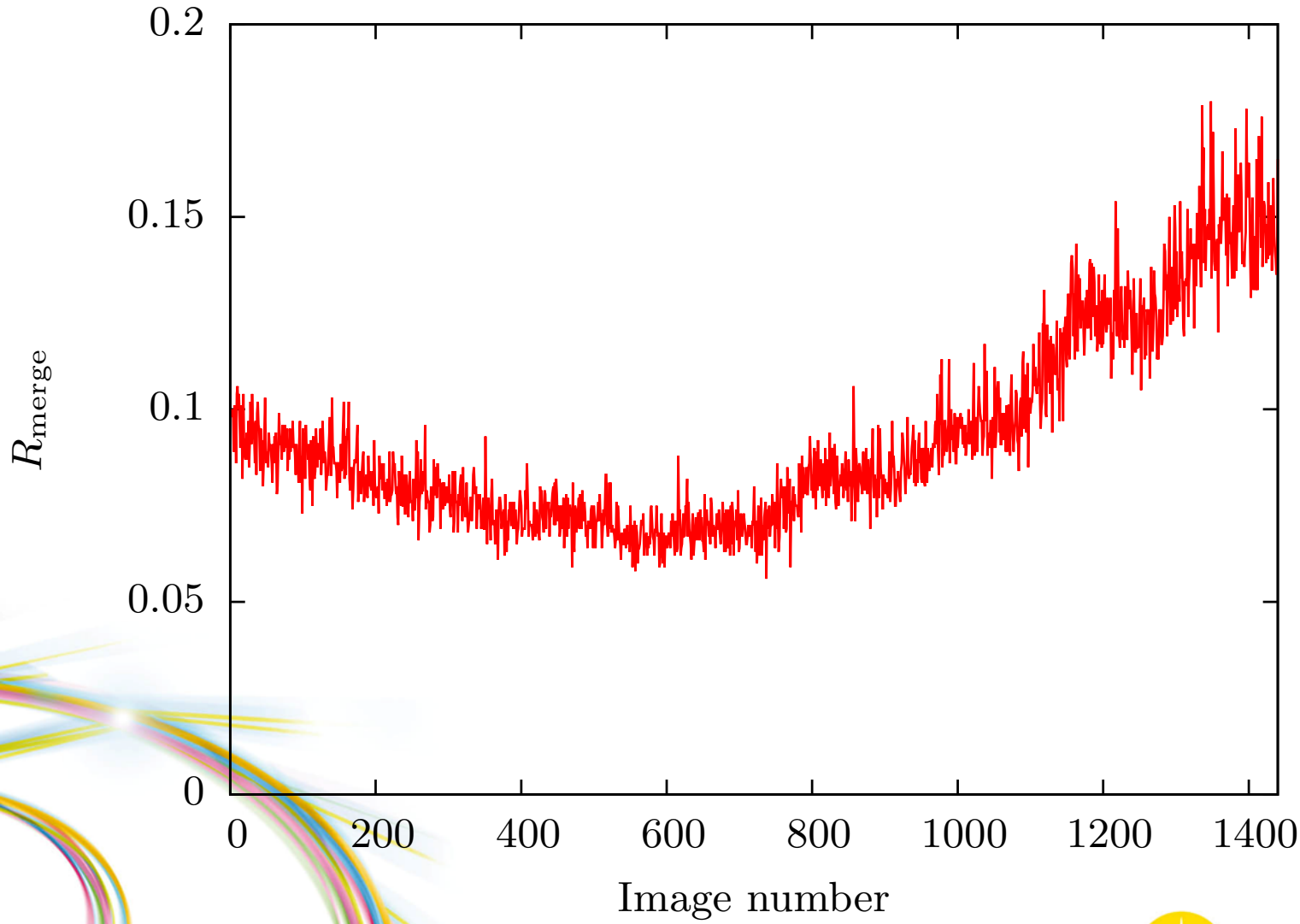


Example 2: radiation damaged SAD data

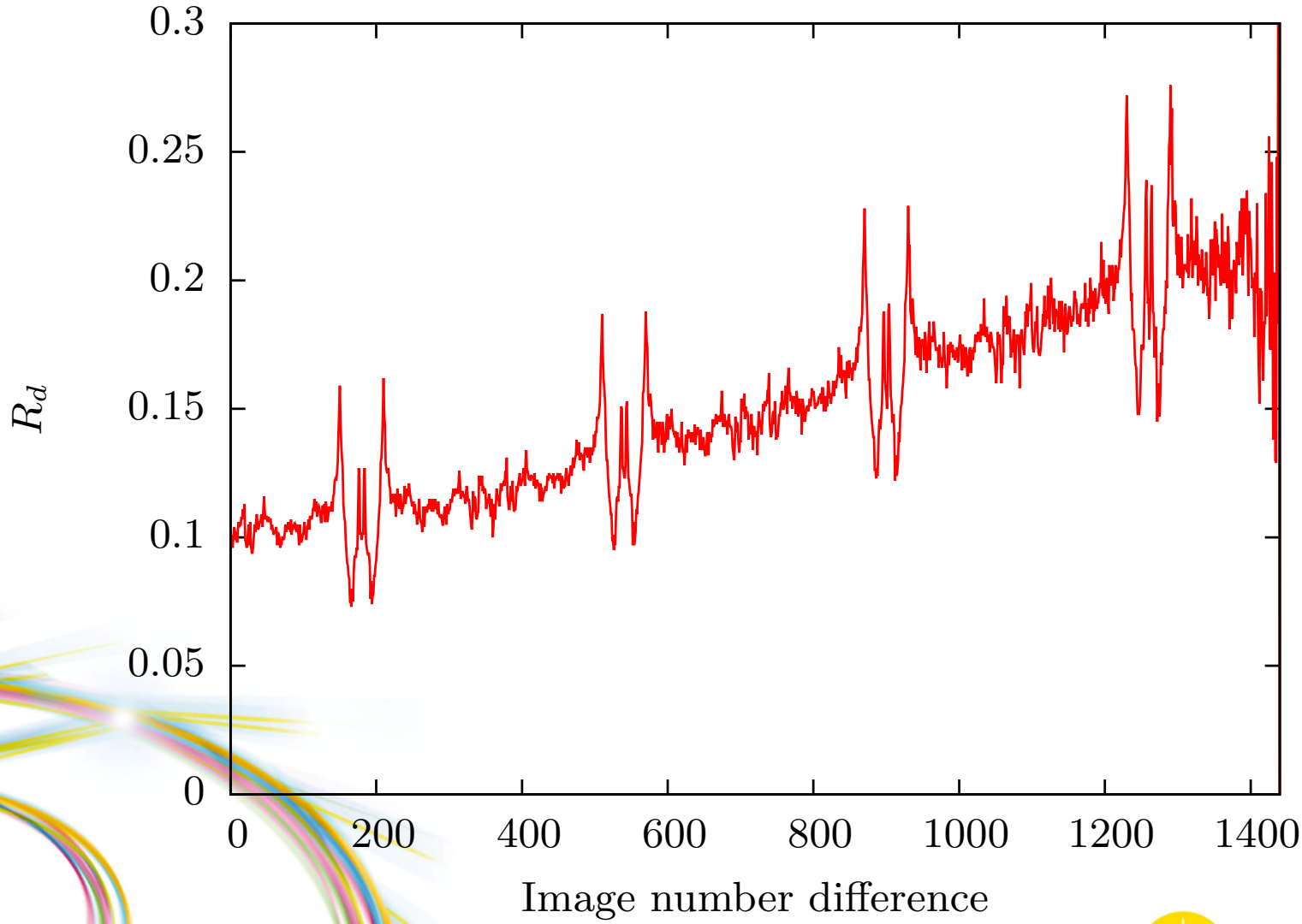
- High multiplicity SAD data set recorded from *Helix pomatia Agglutinin* at ESRF beamline ID14EH2 by Ed Mitchell as part of ongoing research
- 1440° of data, of which 720° were used in the structure solution due to radiation damage
- Symmetry is *H*32, so the data have $\sim$ 80-fold multiplicity

diamond

HPA: R_{merge}



HPA: R_d

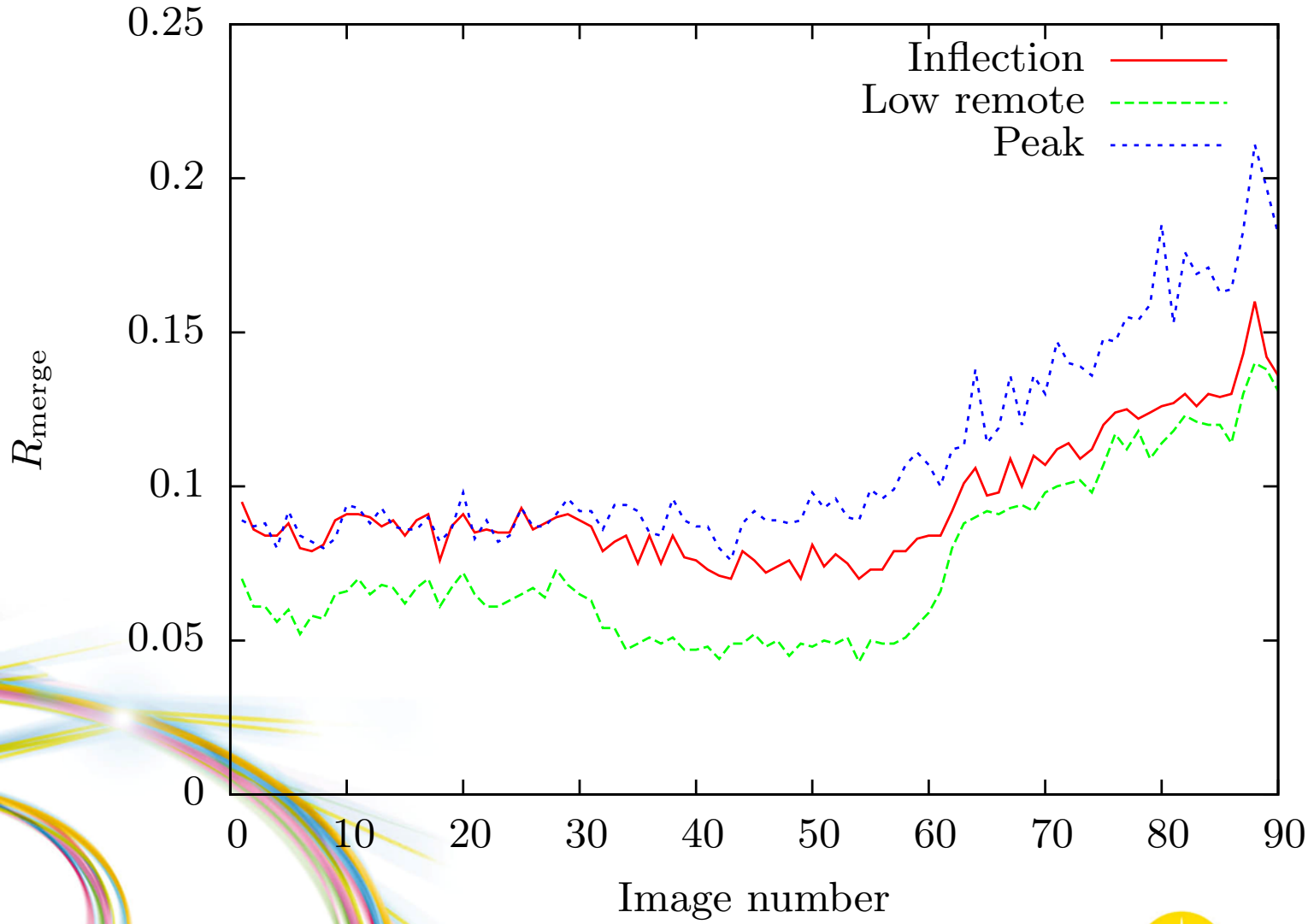

0.3
0.25
0.2
0.15
0.1
0.05
0
R_d
0
200
400
600
800
1000
1200
1400
Image number difference

Example 3: radiation damaged MAD data

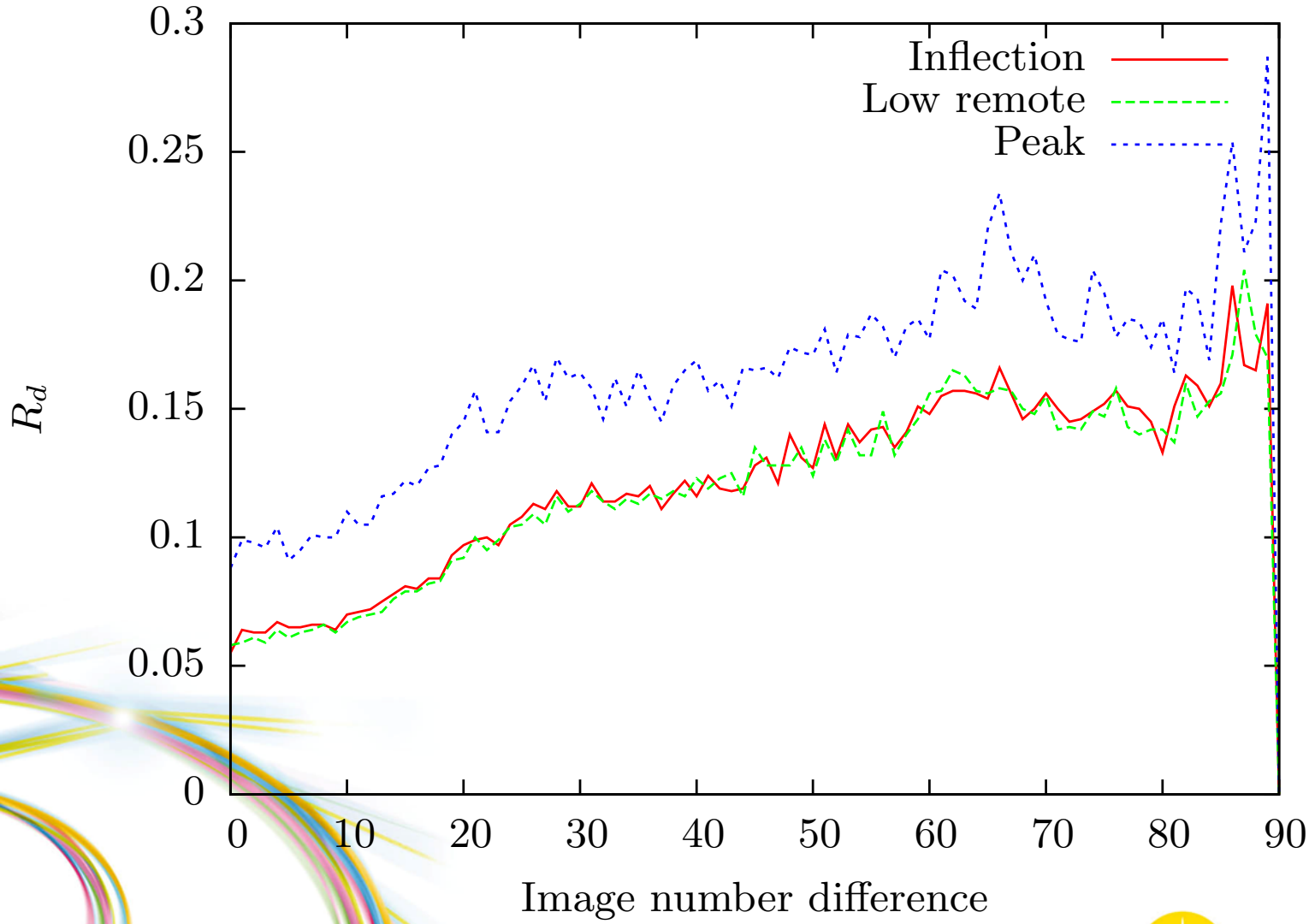
- JCSG target TB0541B
- 8 Se / 200 AA, 3 λ MAD with inverse beam on inflection and low energy remote
- Classic example of nasty radiation damage

diamond

2ISB: R_{merge}



2ISB: R_d


0.3
0.25
0.2
0.15
0.1
0.05
0
R_d
0
10
20
30
40
50
60
70
80
90
Image number difference
Inflection
Low remote
Peak

Review

- Stats behave as expected - presence or absence of radiation damage is clear
- No information provided so far about subset options i.e. when I clearly do have radiation damage, what to do?
- MAD data sets not considered very gracefully

diamond

Novel Statistic

- Question: could things have been better if I had stopped collecting data earlier? I.e. think about wall-clock *time*
- R_{merge} contains time information, R_d shows difference information, wouldn't it be great to have both?
- Including how *complete* the data are would also be important

diamond

Properties

- Computing pairwise differences avoids need for average value
- Integrating differences up to some dose d reproduces experiment
- Organising in terms of dose, integrating across wavelengths and sweeps allows evolution of sample to be described
- While performing analysis, compute completeness (anomalous and native) for each wavelength as function of d

diamond

$R_{cp}(d)$, *Cumulative-Pairwise* Residual

$$R_{cp}(d) = \frac{\sum_{\lambda} \sum_{\underline{h}} \sum_{i \neq j, i:d_i \leq d, j:d_j \leq d} |I_{\underline{h}i} - I_{\underline{h}j}|}{\sum_{\lambda} \sum_{\underline{h}} \sum_{i \neq j, i:d_i \leq d, j:d_j \leq d} \frac{1}{2}|I_{\underline{h}i} + I_{\underline{h}j}|} \tag{3}$$

Measures: including data to a point d, how internally consistent are the measurements?

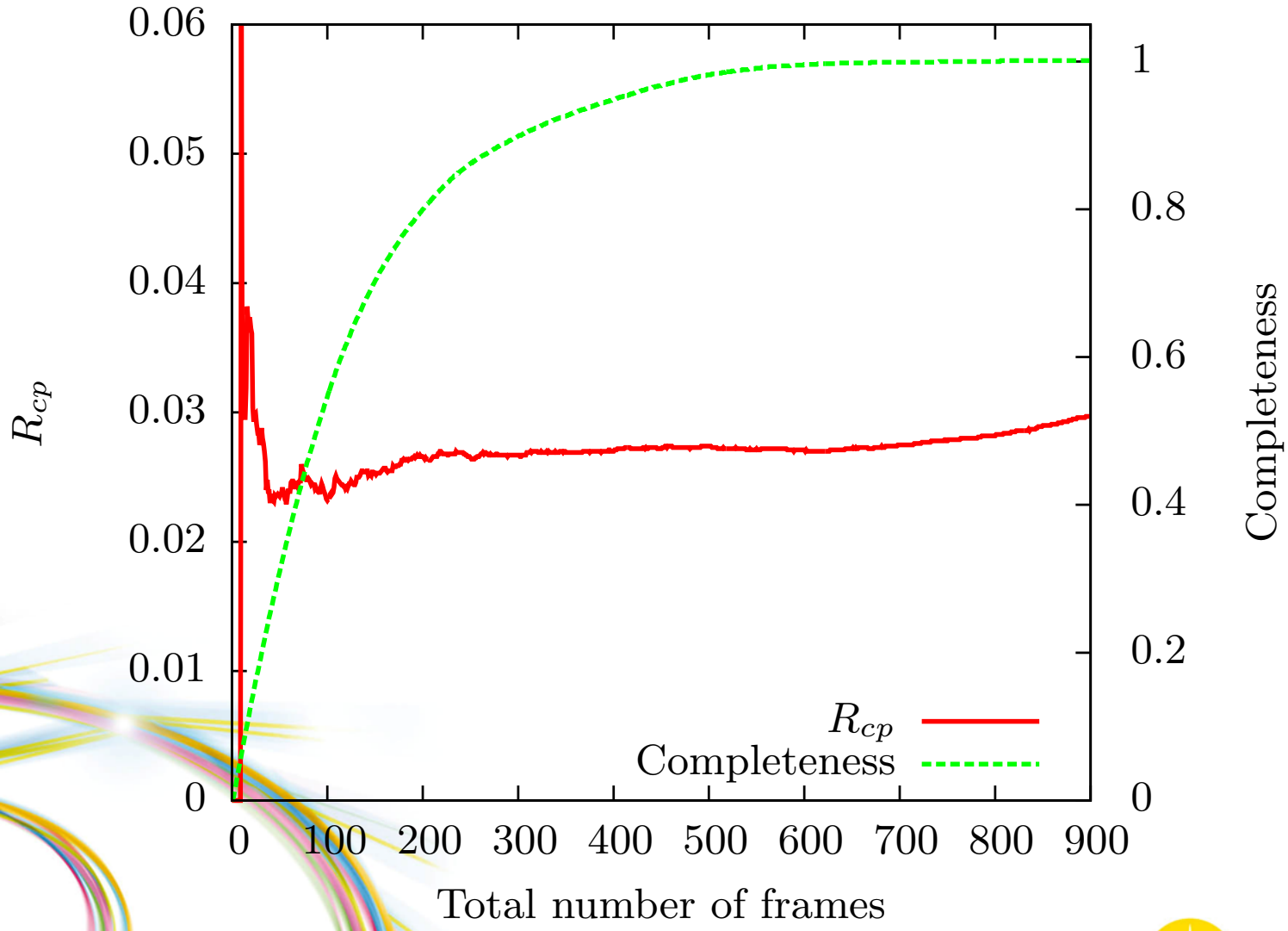
diamond

In English?

- For each wavelength, for each *HKL*, accumulate generally how different the intensities are up to some dose *d*
- Then plot this as a function of *d*
- And while you're there record how complete wavelengths are up to point *d* too...

diamond

Application to Thaumatin

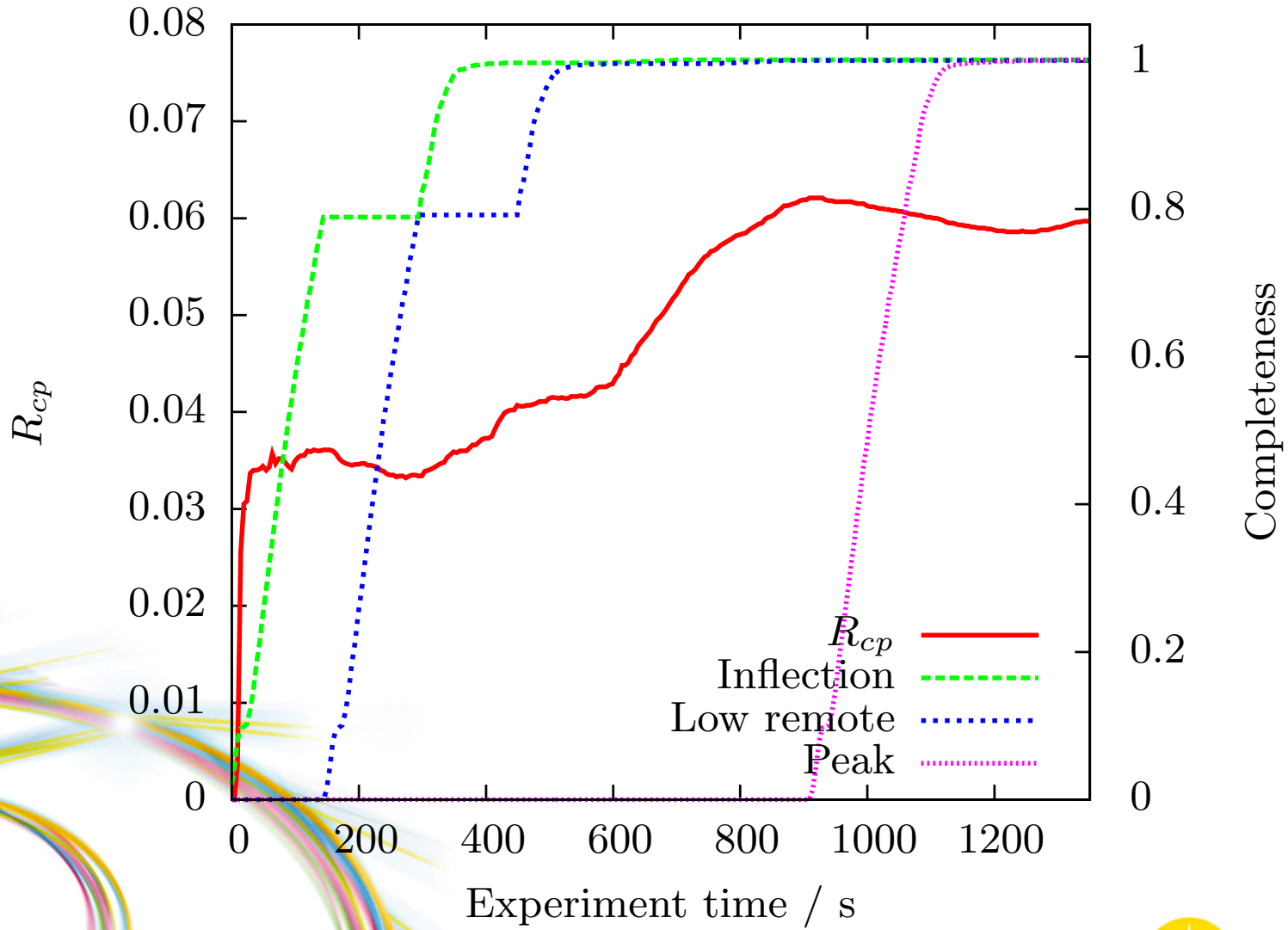


Comments and observations

- At the start always 'noisy' as there are few pairs - only trust once data are e.g. 50% complete - this is very similar noise to the end of an R_d plot
- Most interesting is what happens after the data are fairly complete - in this case, nothing
- Even more interesting is to use *dose* as the baseline rather than batch, though equivalent for SAD

diamond

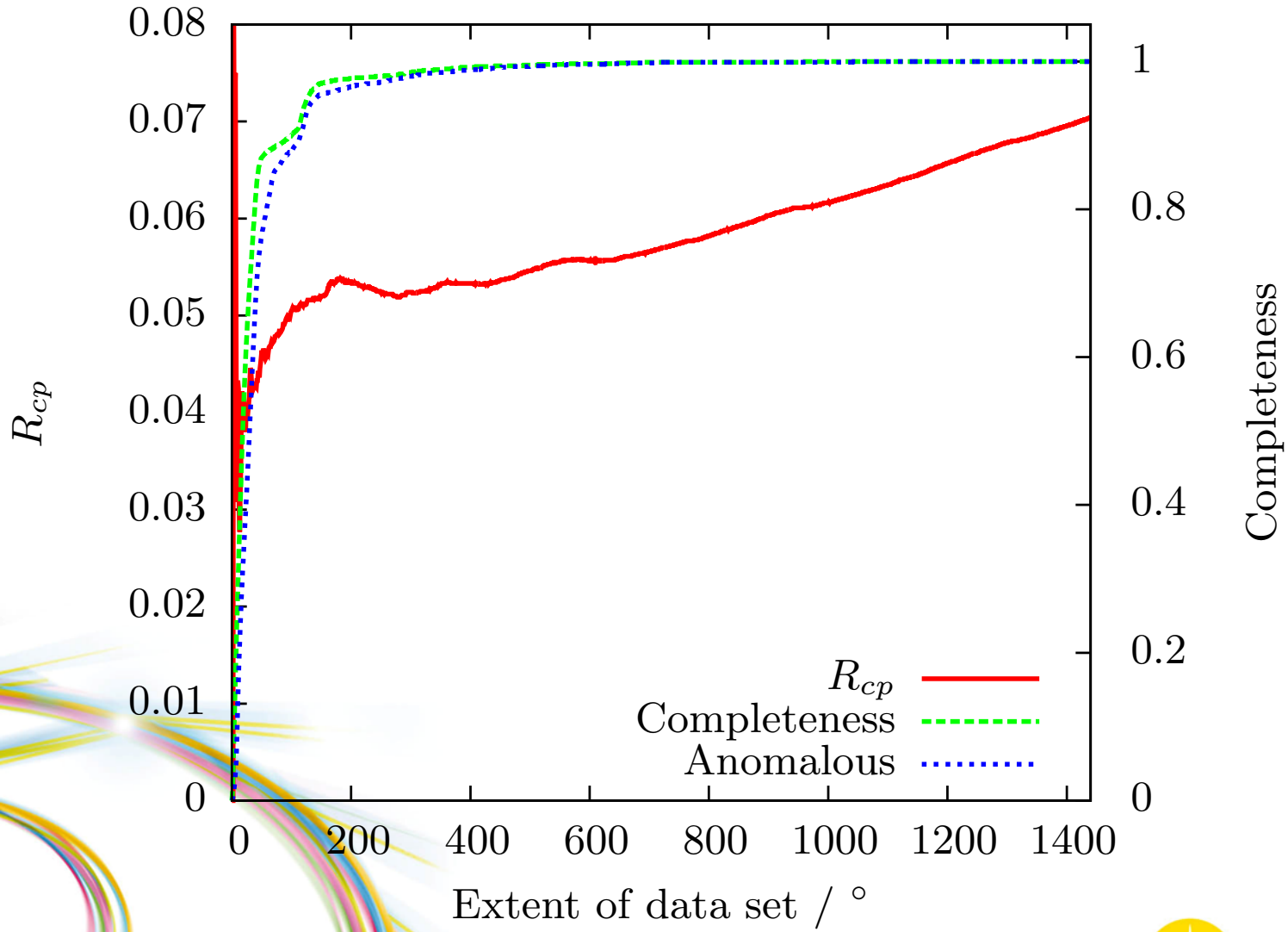
Application to MAD data



Interpretation

- Full three wavelength data has high residual (i.e. after 1350s)
- After 600s we have two complete wavelengths, much lower residual

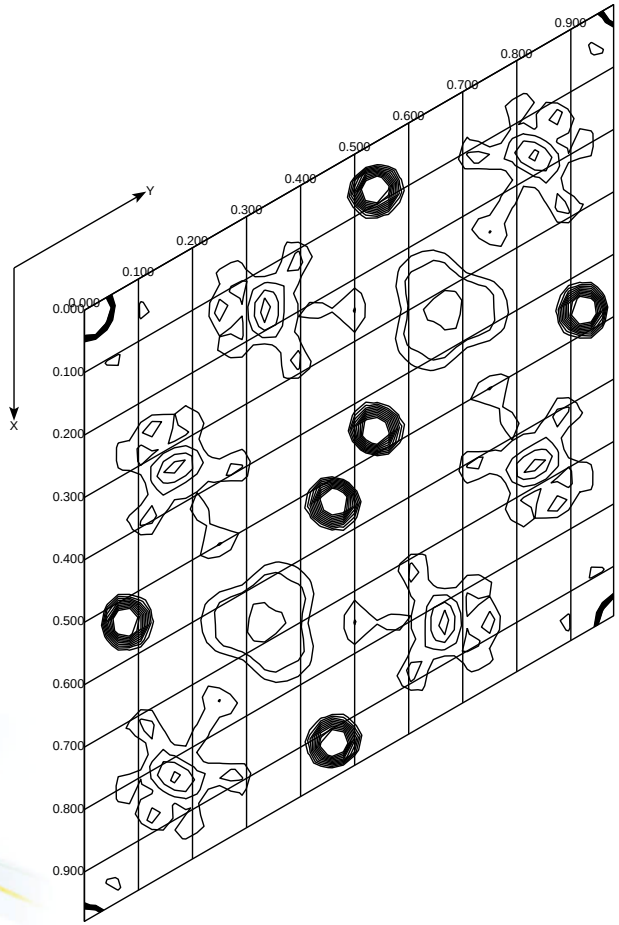
Application to HPA



Interpretation

- Time equivalent to number of frames
- Complete data (anomalous, native) pretty early on
- R_{cp} appears to increase systematically ~ 400 images - just over one full rotation
- One Zn site on special position - anomalous difference Patterson easy to interpret
- Processed data set 32 times - $0^\circ - 45^\circ$, $0^\circ - 90^\circ$, ..., $0^\circ - 1440^\circ$ to fixed resolution limit (1.65Å), used 30Å to 3Å to compute Patterson

HPA Anomalous Difference Patterson

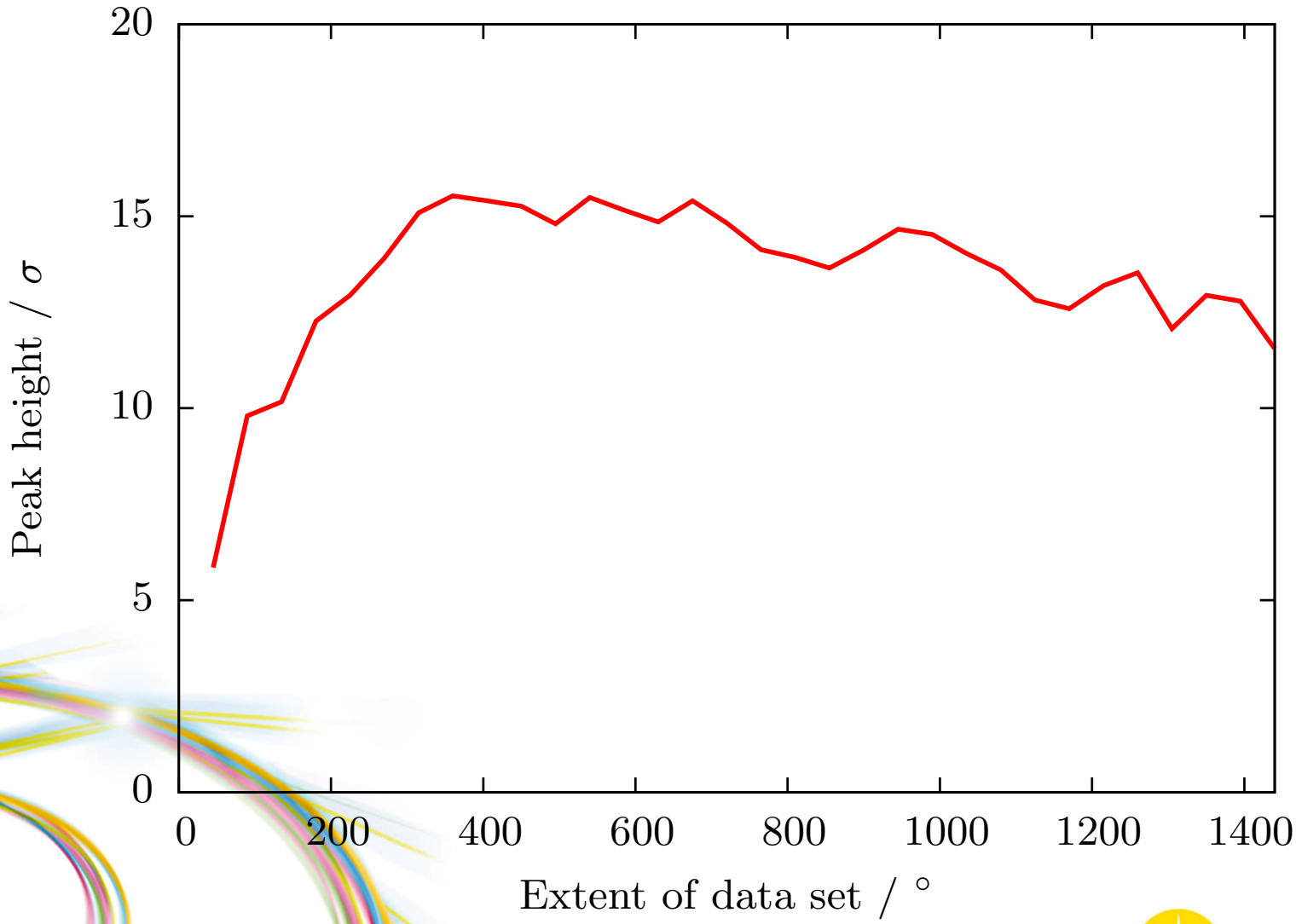


Scale = 2.500000 mm/A

Section 0 (0.000) on z axis

Map from fft

Peak Height / σ



Interpretation

- After $\sim 360^\circ$ radiation damage becomes measurable
- Complete subset around here gives "most internally consistent" representation of the sample
- Peak in anomalous difference Patterson using this much data
- N.B. phasing works with subset, full data set, just using SHELX C/D/E
- Zero-dose gave peak height of 16.07σ - slightly better than the 360° set but with $4\times$ the measurements used

Summary to date

- Looked at some examples
- Showed new statistic
- Illustrated statistic may be useful, perhaps more useful than R_{merge} *vs.* batch and complementary to R_d
- No cases shown where the insoluble becomes soluble - let's fix that

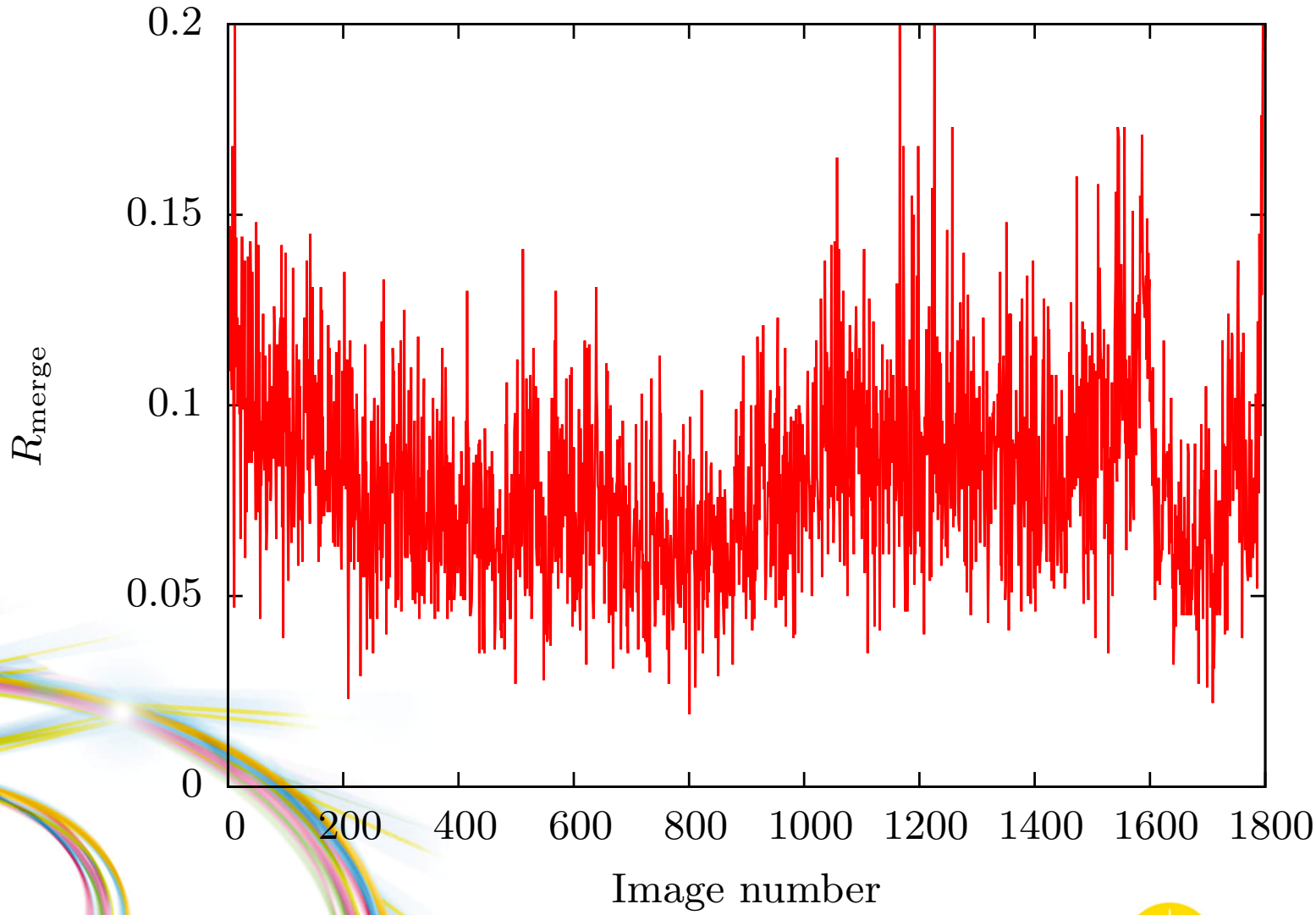
diamond

Unpublished Example - Au soak SAD

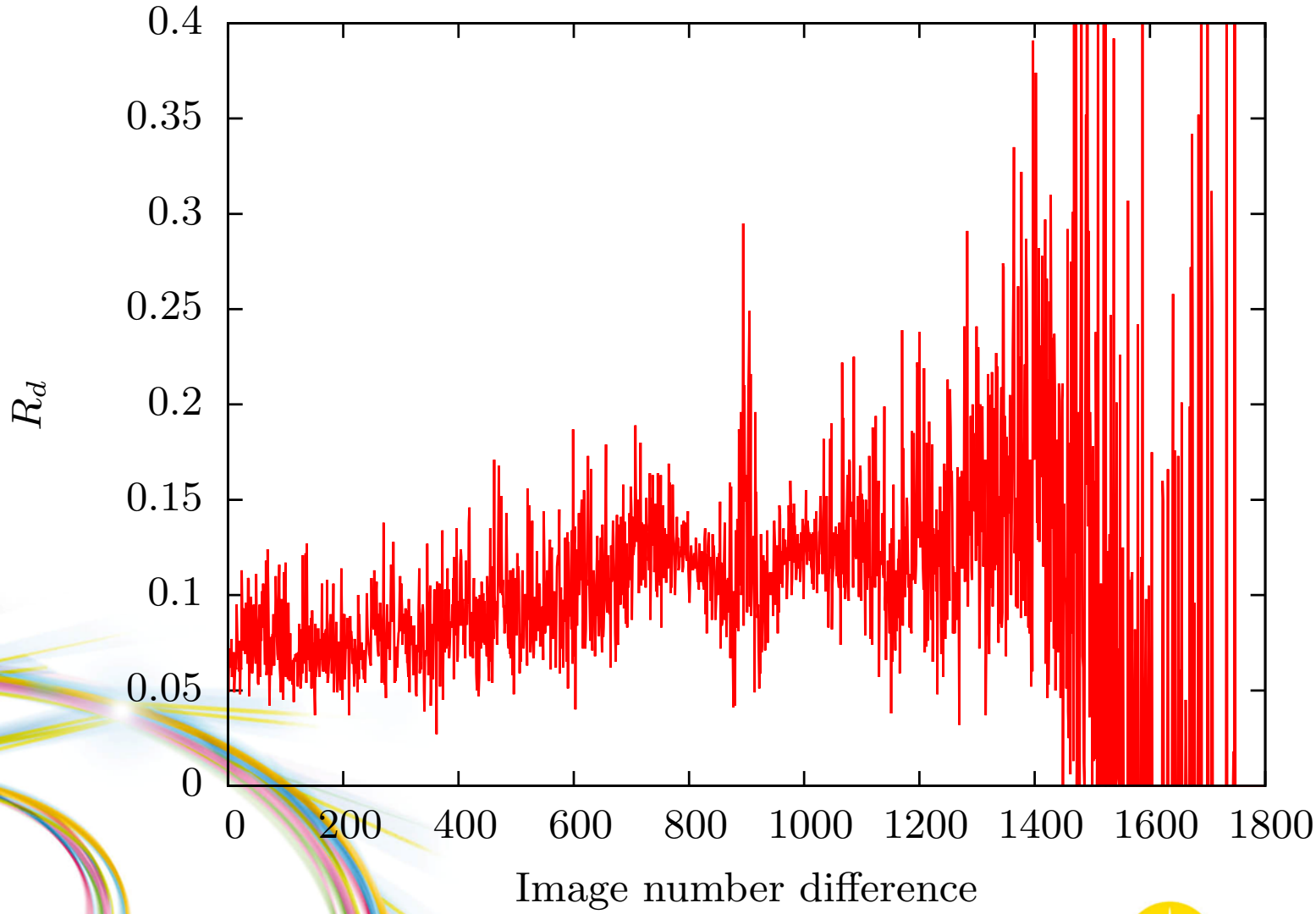
- Example from Christian Siebold at STRUBI
- Measured at Diamond Light Source I03 using Pilatus 6M
- Native data available, issue with this data set is substructure determination for phasing

High resolution limit	2.01	8.98	2.01
Low resolution limit	29.85	29.84	2.06
Completeness	71.2	97.3	13.5
Multiplicity	5.3	6.0	2.1
I/sigma	14.7	21.6	2.3
Rmerge	0.063	0.043	0.405
Rmeas(I)	0.086	0.067	0.707
Rmeas(I+/-)	0.076	0.052	0.568
Rpim(I)	0.034	0.027	0.470
Rpim(I+/-)	0.041	0.028	0.397
Wilson B factor	33.264	.	.
Partial bias	0.000	0.000	0.000
Anomalous completeness	65.1	97.6	10.6
Anomalous multiplicity	2.9	3.3	1.1
Anomalous correlation	0.388	0.613	0.690
Anomalous slope	1.589	0.000	0.000
Total observations	40714.	775.	228.
Total unique	7685.	130.	111.

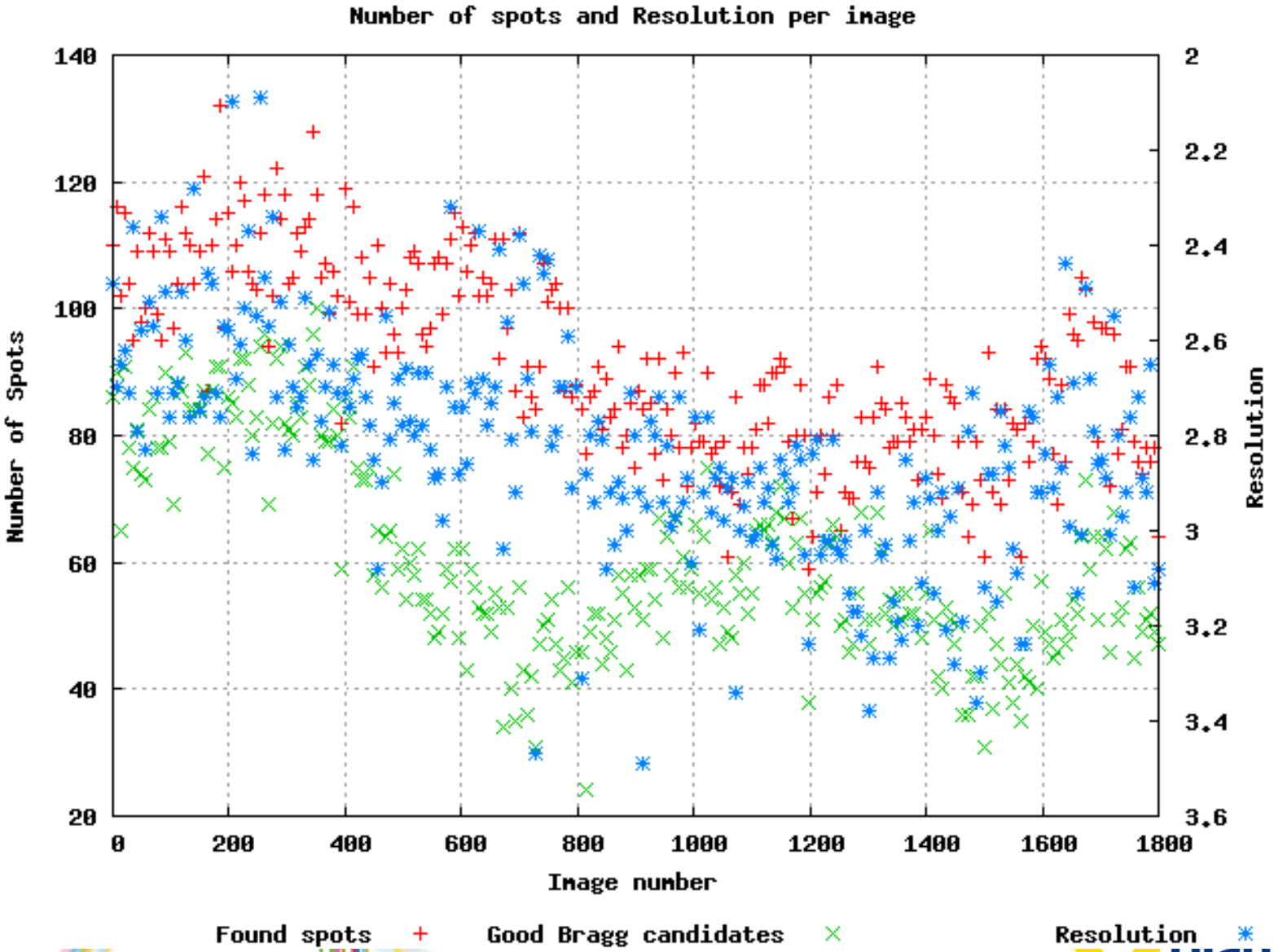
Au soak: R_{merge}



Au soak: R_d


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
R_d
0
200
400
600
800
1000
1200
1400
1600
1800
Image number difference

Per-image analysis (DISTL)

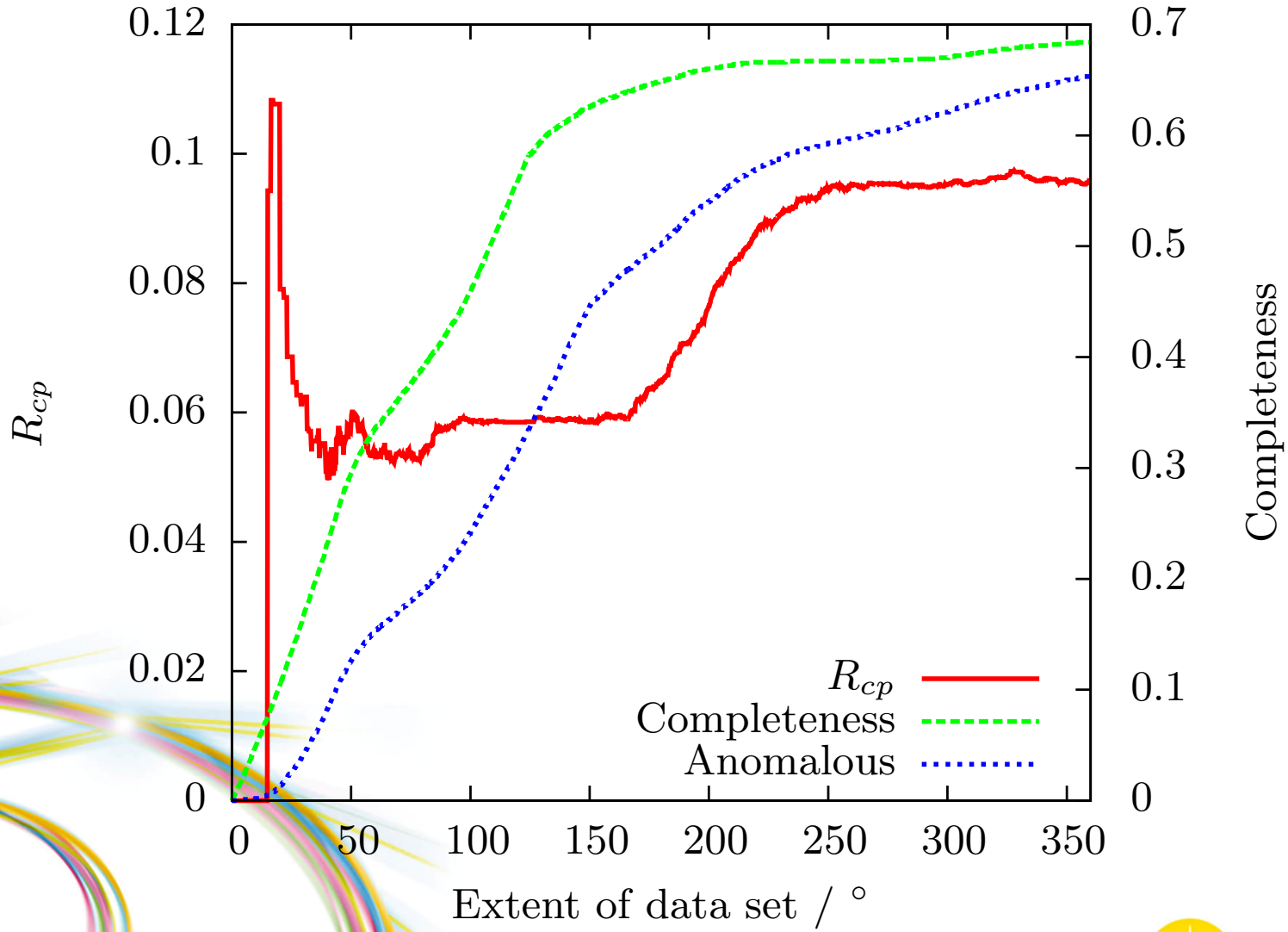


Interpretation

- Detector too far back - low completeness at high resolution
- Merging stats reasonable, clearly we have some radiation damage
- What does R_{cp} say?

diamond

Au soak: R_{cp}



Clear radiation damge

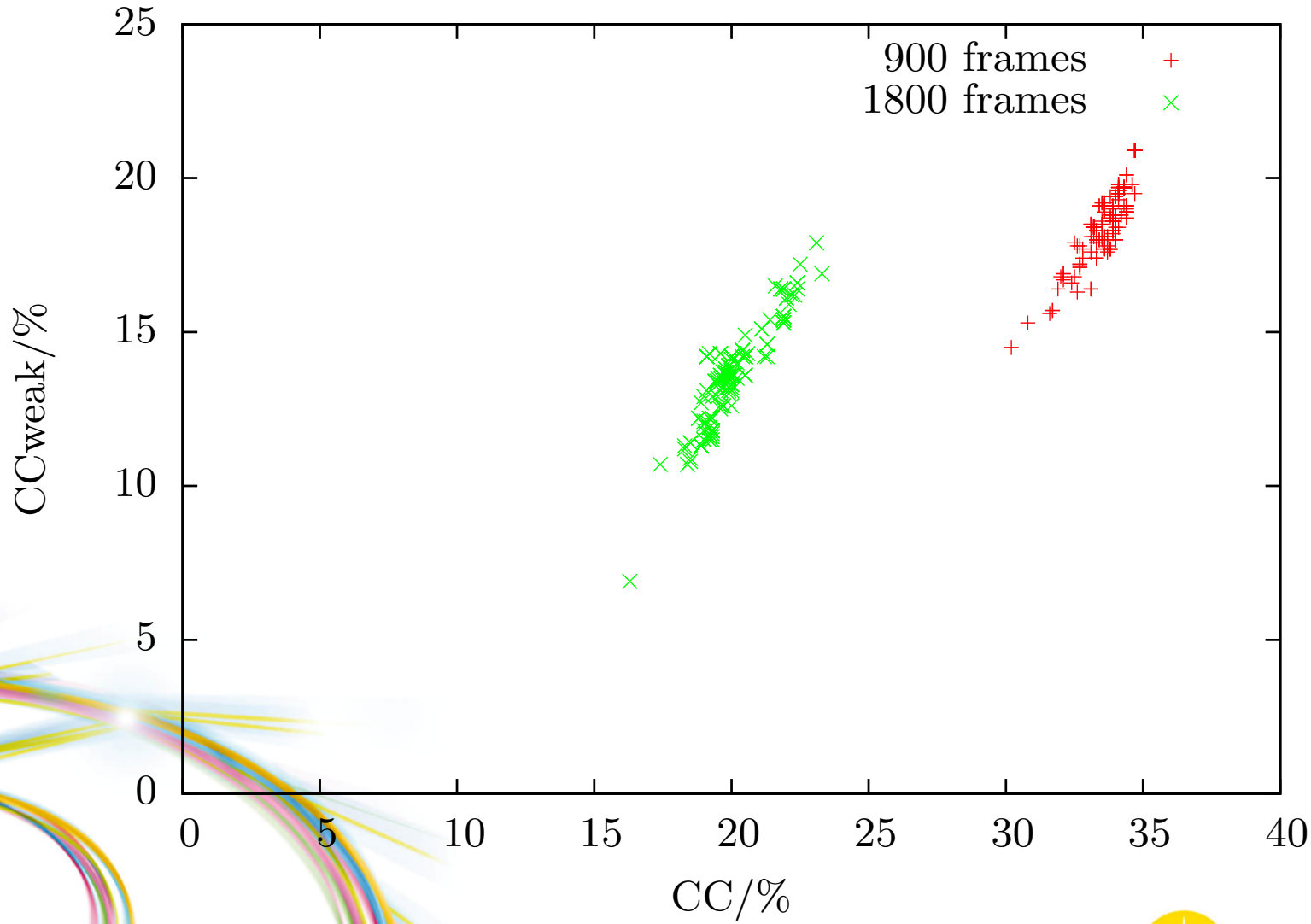
- Clear radiation damage according to this measure
- Data essentially as complete as they can be half way through
- Assert: using 900 frames / 180° of the data will work better than the full set
- Test: use SHELXD (via "Fast EP" at Diamond)

diamond

High resolution limit	2.01	8.98	2.01
Low resolution limit	29.83	29.83	2.06
Completeness	70.9	95.1	12.5
Multiplicity	2.9	3.1	1.2
I/sigma	16.0	31.5	1.4
Rmerge	0.036	0.022	0.348
Rmeas(I)	0.063	0.051	0.703
Rmeas(I+/-)	0.049	0.030	0.493
Rpim(I)	0.035	0.027	0.497
Rpim(I+/-)	0.034	0.021	0.348
Wilson B factor	32.688	.	.
Partial bias	0.000	0.000	0.000
Anomalous completeness	55.7	86.7	2.0
Anomalous multiplicity	1.6	1.8	1.0
Anomalous correlation	0.609	0.835	0.000
Anomalous slope	1.312	0.000	0.000
Total observations	21812.	385.	123.
Total unique	7646.	126.	101.

High resolution limit	2.01	8.98	2.01
Low resolution limit	29.85	29.84	2.06
Completeness	71.2	97.3	13.5
Multiplicity	5.3	6.0	2.1
I/sigma	14.7	21.6	2.3
Rmerge	0.063	0.043	0.405
Rmeas(I)	0.086	0.067	0.707
Rmeas(I+/-)	0.076	0.052	0.568
Rpim(I)	0.034	0.027	0.470
Rpim(I+/-)	0.041	0.028	0.397
Wilson B factor	33.264	.	.
Partial bias	0.000	0.000	0.000
Anomalous completeness	65.1	97.6	10.6
Anomalous multiplicity	2.9	3.3	1.1
Anomalous correlation	0.388	0.613	0.690
Anomalous slope	1.589	0.000	0.000
Total observations	40714.	775.	228.
Total unique	7685.	130.	111.

SHELXD CC, CCweak



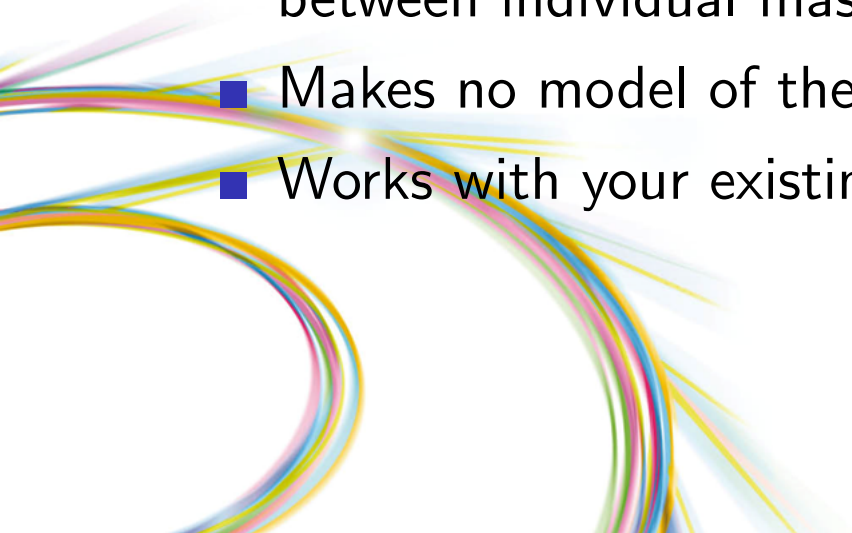
Discussion

- Subset of data currently being used with native for structure solution
- “Good” solutions from SHELXD were verified to be correct
- For reference: used zero-dose extrapolation in XSCALE, get 27.23 / 13.57 for CC / CCweak

diamond

Conclusions

- This tool possibly most useful with shutterless data collection
- Anisotropic diffraction: typically related to symmetry, symmetry used in calculation so not sure what effects are likely to be...
- Makes no changes to the intensity values - only asks the question "what happens if I stopped collecting earlier?"
- Does not depend on an average value, only on differences between individual masurements
- Makes no model of the radiation damage
- Works with your existing software in a graceful way



diamond

Next step

- Extending to multi-crystal data analysis - all samples start at dose 0, may decay at different rates, total completeness available

Acknowledgements

- Support and input from Diamond Light Source and Diamond Staff
- Miroslav Papiz, Steve Prince for converstion which got this started, inspriation for the name *chef*
- Kay Diederichs for R_d, as well as 0-dose with Sean McSweeney and Raimond Ravelli
- Dave Stuart, Phil Evans for useful discussions on this subject
- Ed Mitchell, I03 Staff, JCSG and Christian Siebold for the examples shown