Data collection strategy

Zbigniew Dauter

Natl. Cancer Inst. & Argonne Natl. Lal

Advance preparations

1. Crystals must be prepared

2. You must be prepared

Advance preparations

1. Crystals must be prepared
2. You must be prepared

because

Anything that can go wrong,
will go wrong
(Murphy, 2000 BC)

Work at the beam line

- Very hectic and/or very boring
- Requires quick responses (time is precious)
- Full of surprises
    - technical problems usually on Friday evening
    - your “best” crystals do not diffract
    - cryo stream develops an ice block
    - storm in Connecticut, no power
- Rarely fully satisfying, sometimes rewarding (lesson of optimism and perseverance)

Data collection process

- Involves lots of technical problems
- But it is science, not technicality
- Easy to screw-up in many ways
- Pays off to "engage your brain"
- Last truly experimental step

  later mostly computing (and writing-up

  which may be repeated many times
- good quality data make all subsequent

  steps much easier

Beam line selection

- Intensity and brilliance
  - 2^{nd} vs. 3^{rd} generation synchrotron
- Collimation, divergence and focusing
  - undulator or bending magnet
- Wavelength range
- Detector type and size
- Crystal characteristics
  - diffraction strength, cell dimensions
- Accessibility (APS vs. NSLS)
- User friendliness

No marvels

But synchrotron beam makes no miracles

bad crystal at home will be bad also
at the synchrotron

"This diffraction is so bad – how good
we did not bring our best crystals…"
(Hamburg, 1988)

Type of experiment

Always best to have diffraction data
complete, high resolution and accurate
but of particular importance are:

- Native data for refinement
  highest resolution (multiple passes)
- Molecular replacement
  medium resolution, no overloads
- Heavy atom derivative
  medium resolution, accurate
- Anomalous (MAD, SAD)
  modest resolution (radiation damage)
  very accurate and complete at low res

Detector and software

- In general all CCD or IP detectors and all data processing programs give equally good data (if working properly)
- Sometimes important is the size of detector front window (e.g. viruses)
- Some programs are better for particular applications (e.g. d*trek for fine slicing) or more automatic (user friendly)
- experimenter's experience may be more important than data processing program

Quality criteria

What means "good data" ?

Quantitatively and Qualitatively

Quantitatively	and	Qualitatively
Complete		Accurate
All reflections in the asymmetric or the anomalous unit have to be measured		Intensities have to be meaningful and have realistic error estimates (sigmas or uncertainties)

Very easy, but not good to collect indices without intensities (and their error estimations

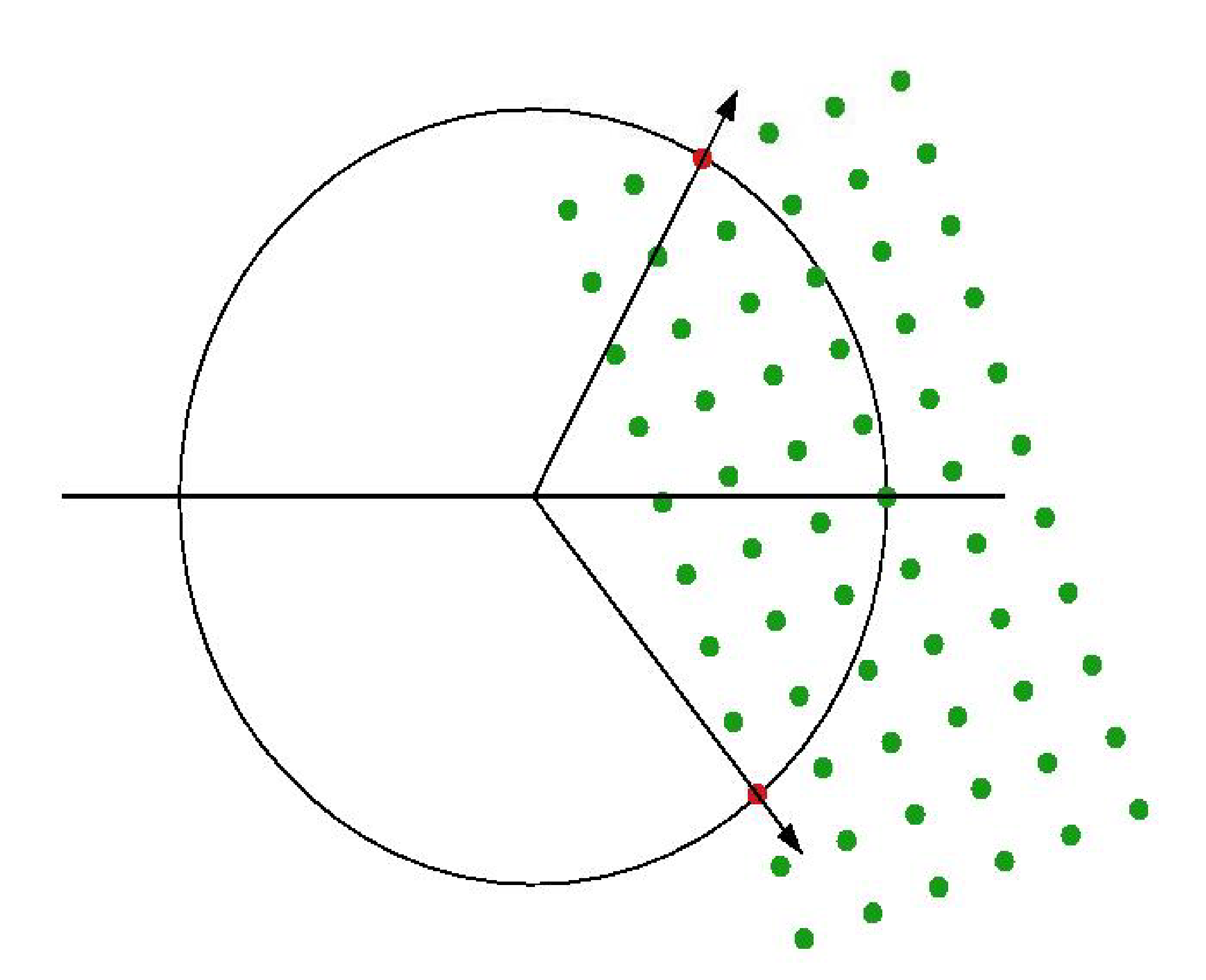
Quantitative completeness of indices

Depends entirely on the geometry
and mutual disposition of

Reciprocal lattice (crystal)
and
Ewald sphere (radiation)

Ewald construction

3-D illustration of Braggs law: $n \cdot \lambda = 2 \cdot d \cdot \sin$



Ewald sphere represents radiation

Reciprocal lattice represents crystal

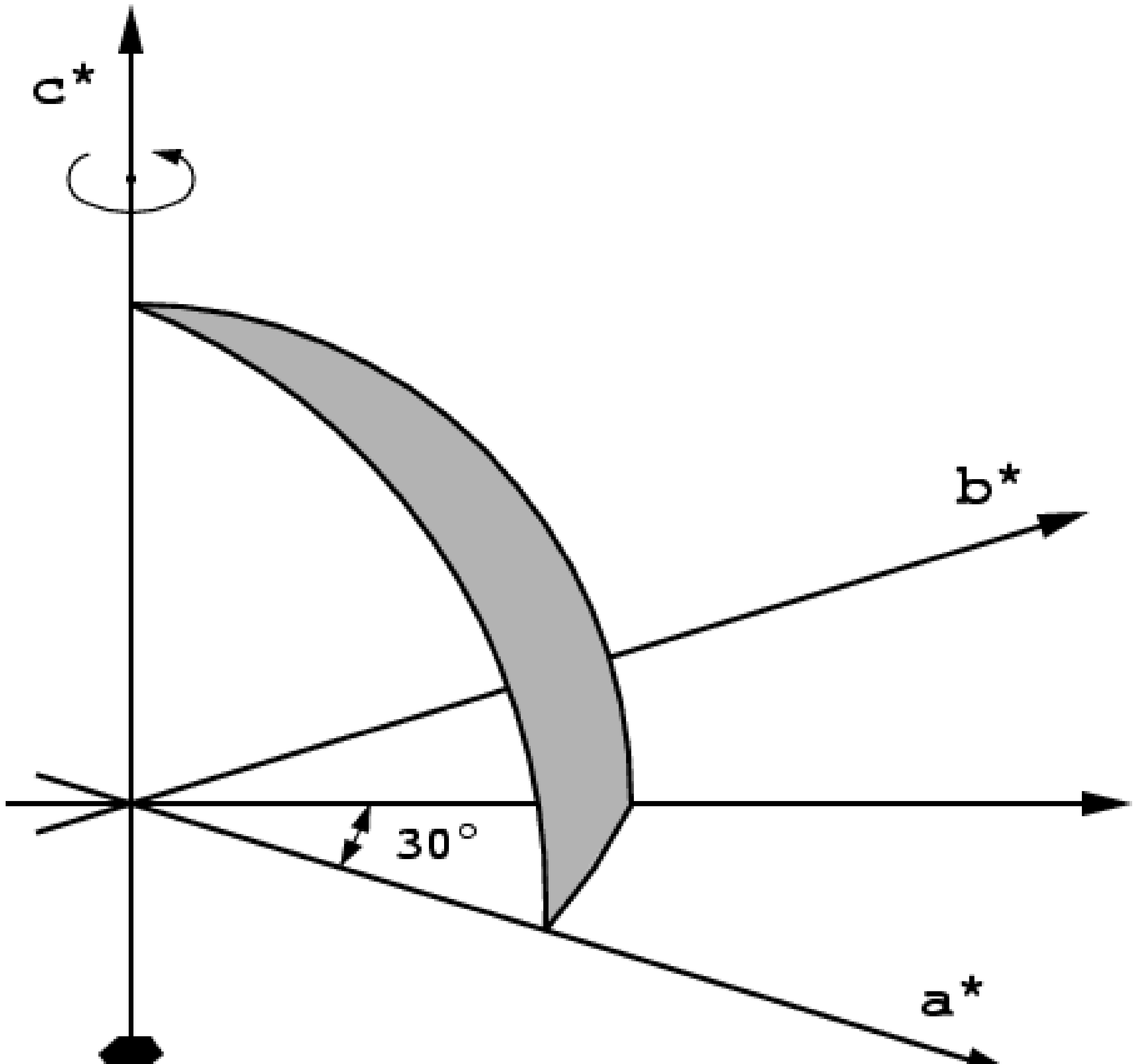
Asymmetric unit in reciprocal space

Asymmetric unit in reciprocal space is always a wedge bounded by rotation axes (or planes in Laue group):

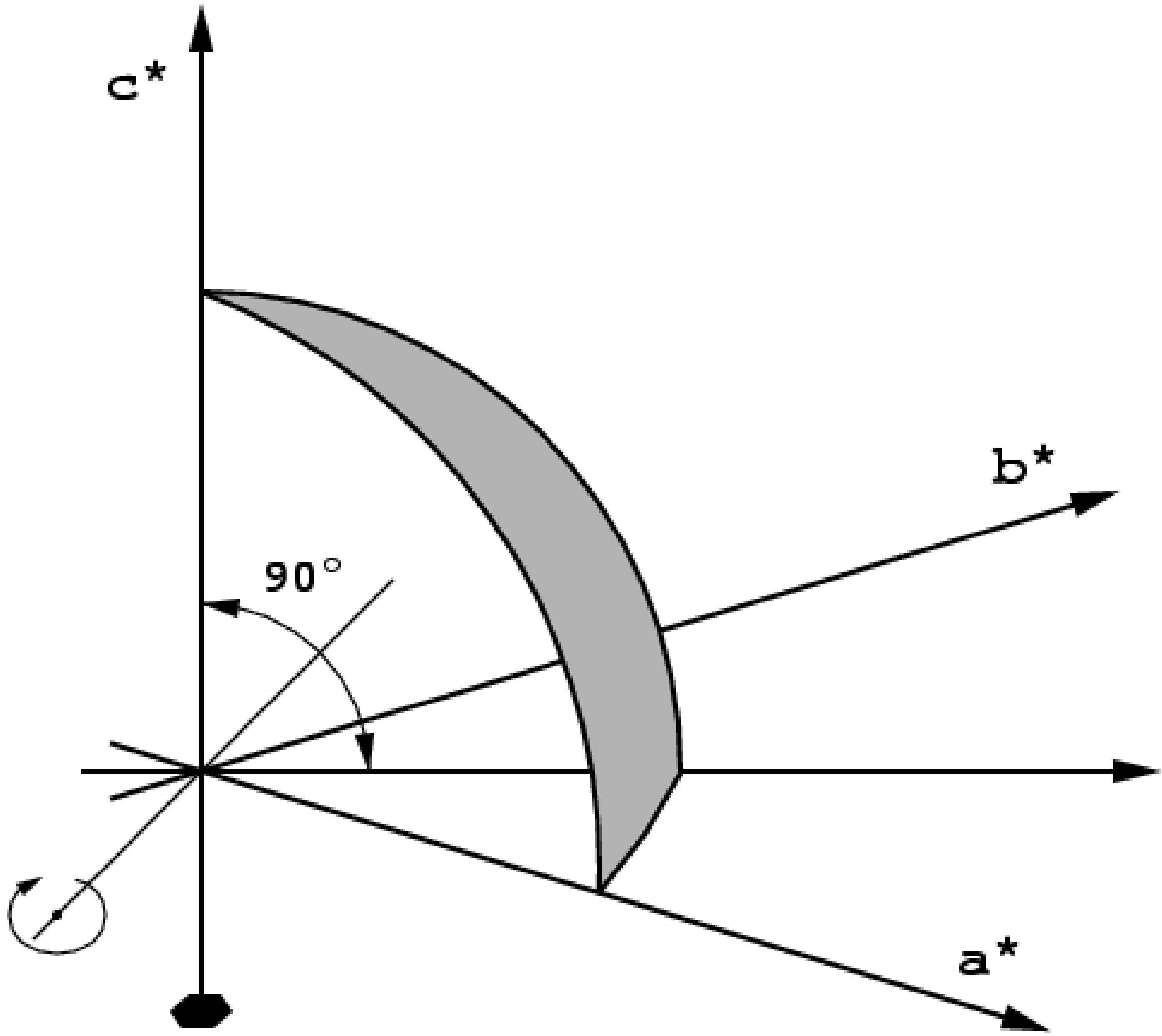
		native		anomalous
		native		anomalous
Triclinic	-	hemisphere	-	sphere
Orthorhombic	-	octant	-	quadrant
etc.				

It is important to know where to start and how much to rotate the crystal

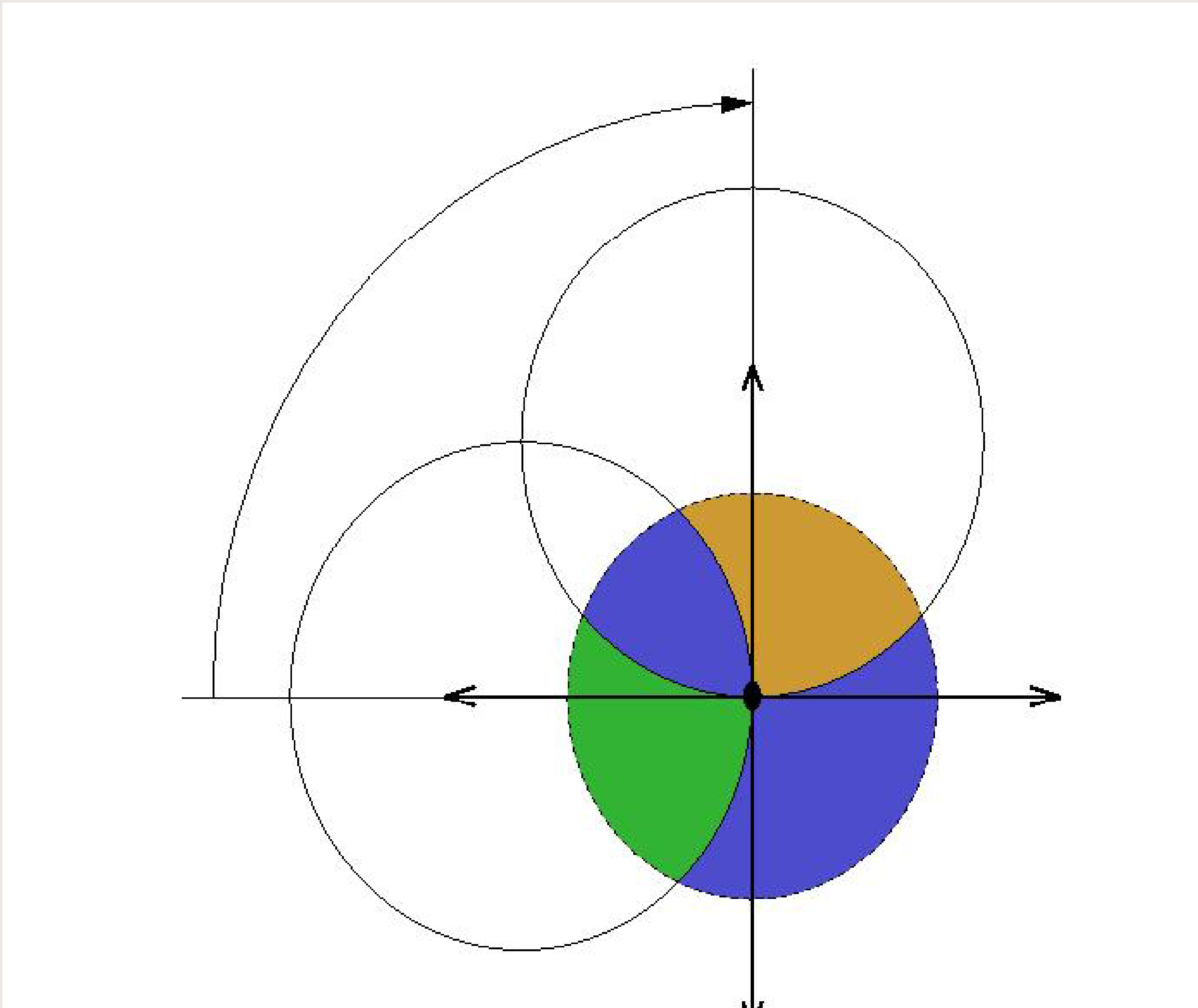
Asymmetric unit in 622 - c axis rotation



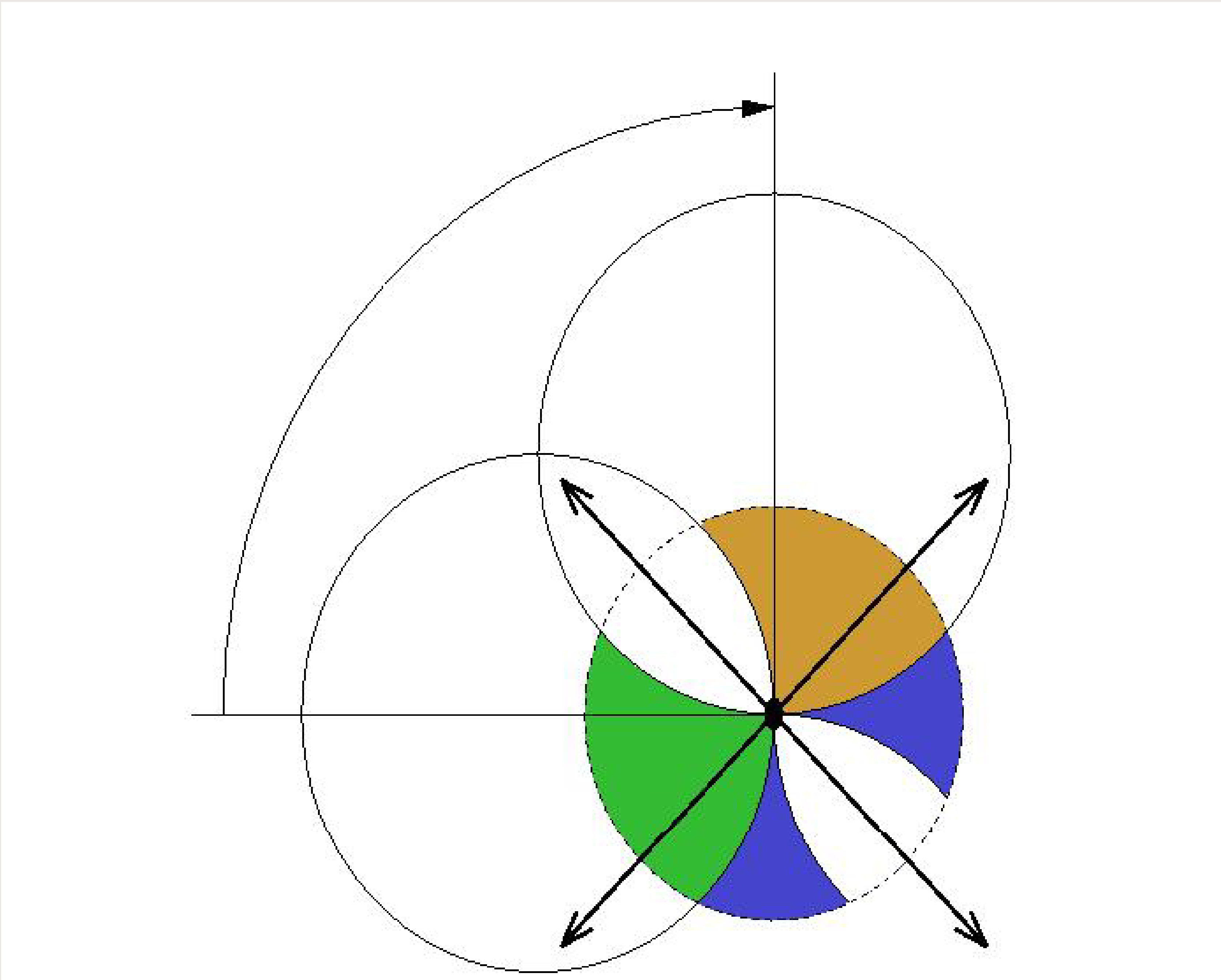
Asymmetric unit in 622 - a/b axis rotatio



Asymmetric unit in 222 - 90° axial



Asymmetric unit in 222 – 90° diagonal



Strategy programs

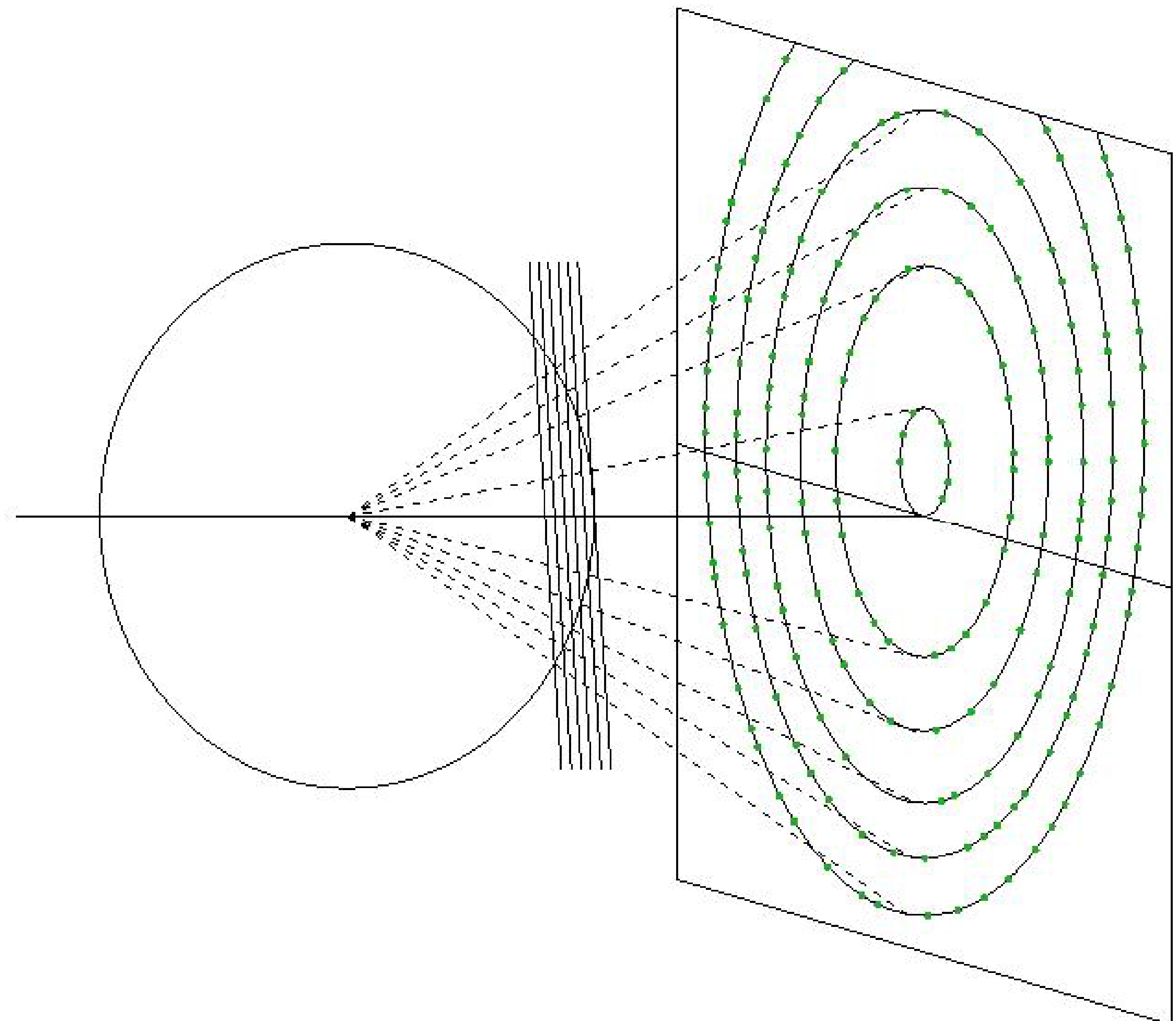
Such considerations are easy
with crystals in axial orientation,

but in arbitrary orientation it may be difficult

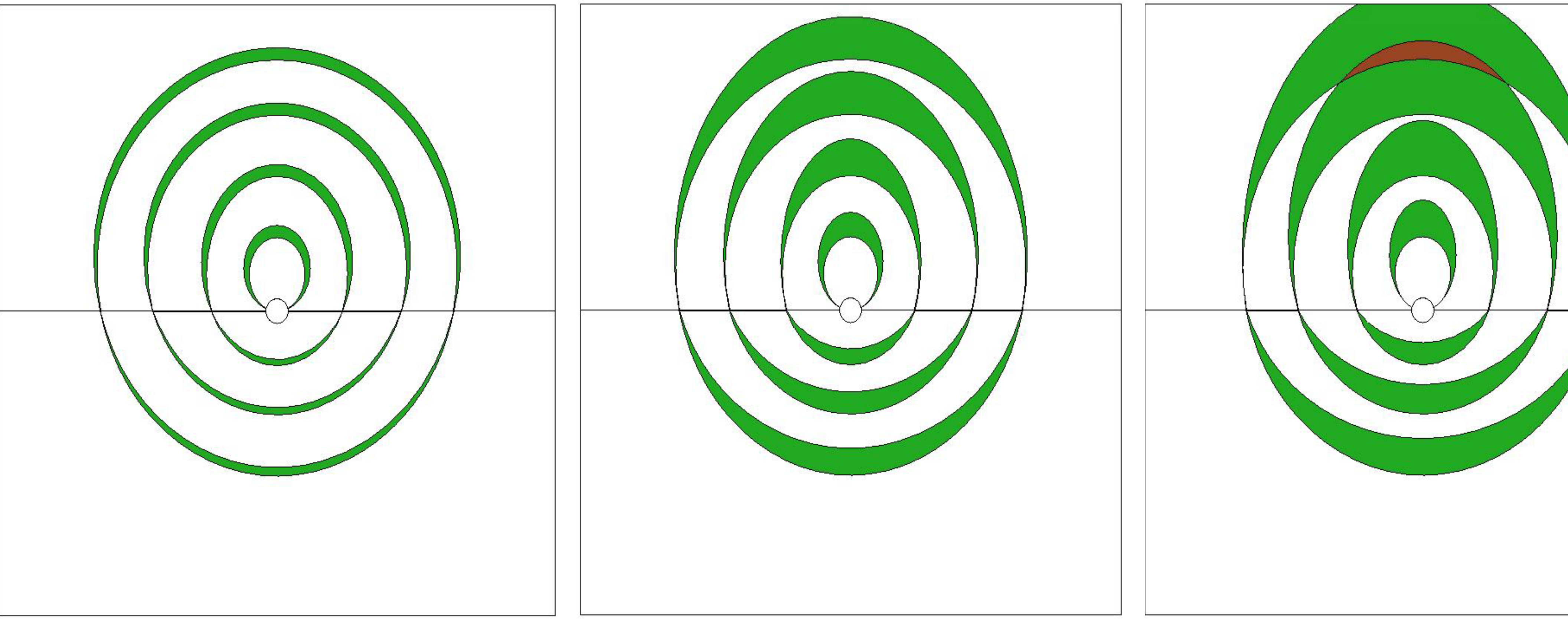
Good to rely on the strategy programs

(This is a minimalist approach,
360° will always give complete data set,
but beware of radiation damage !)

Still image

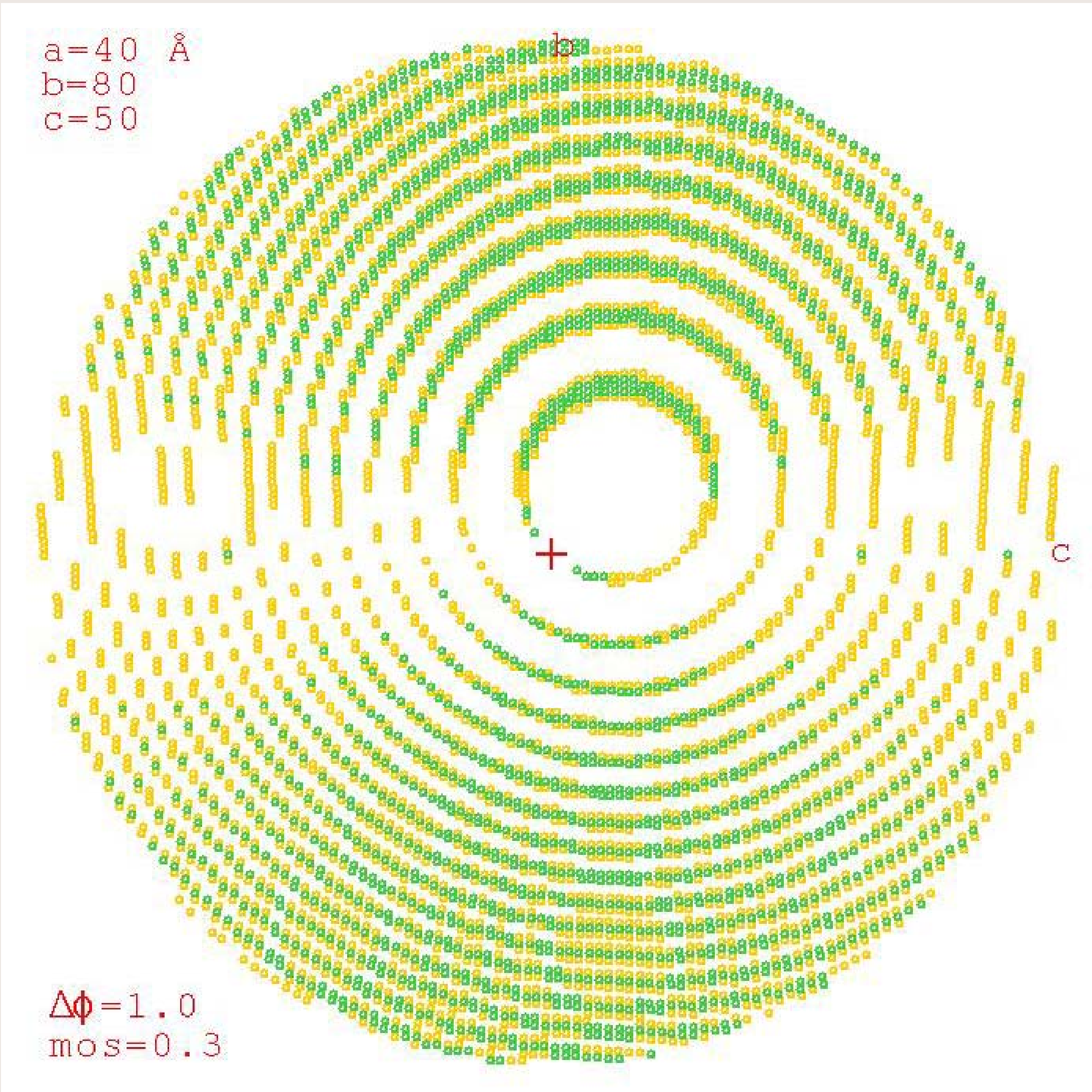


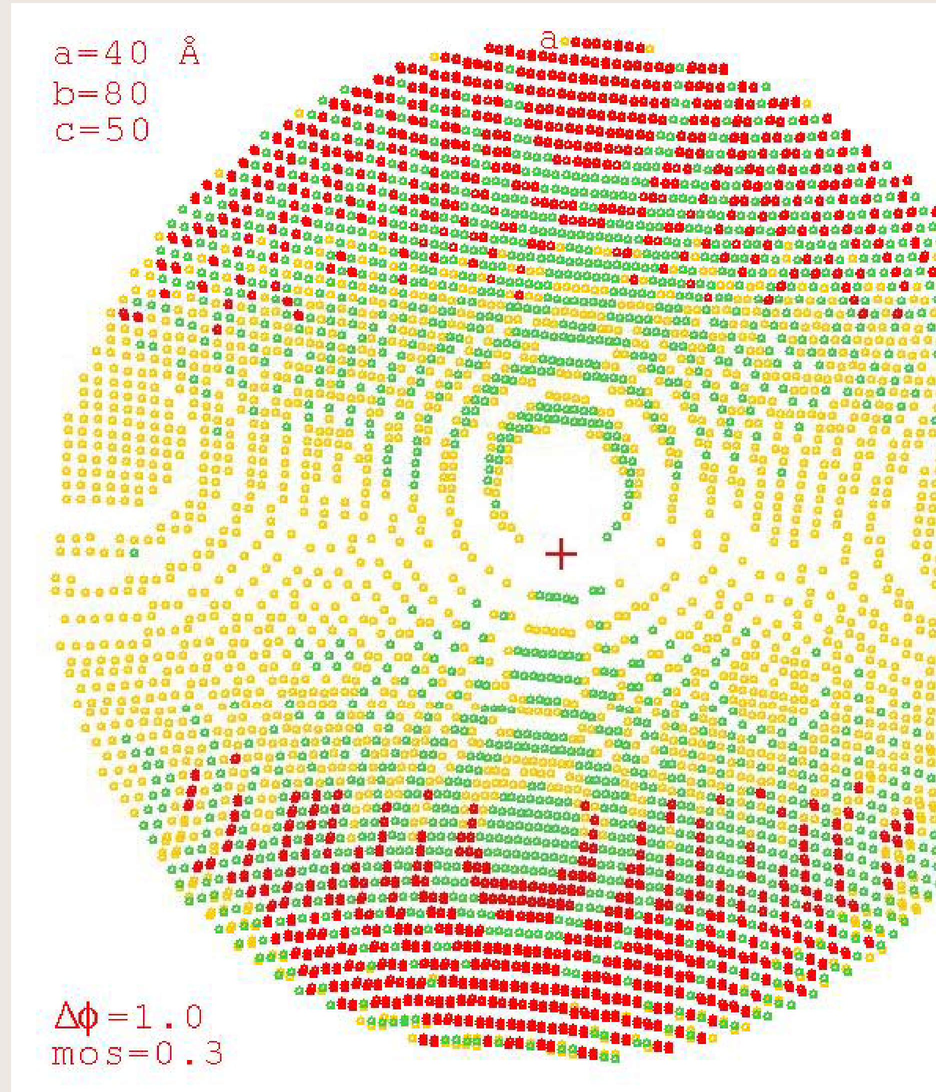
Increasing rotationMosaicity in reciproc space



$$\Delta\varphi_{max} = \frac{180 \cdot d}{\pi \cdot a} - \eta$$

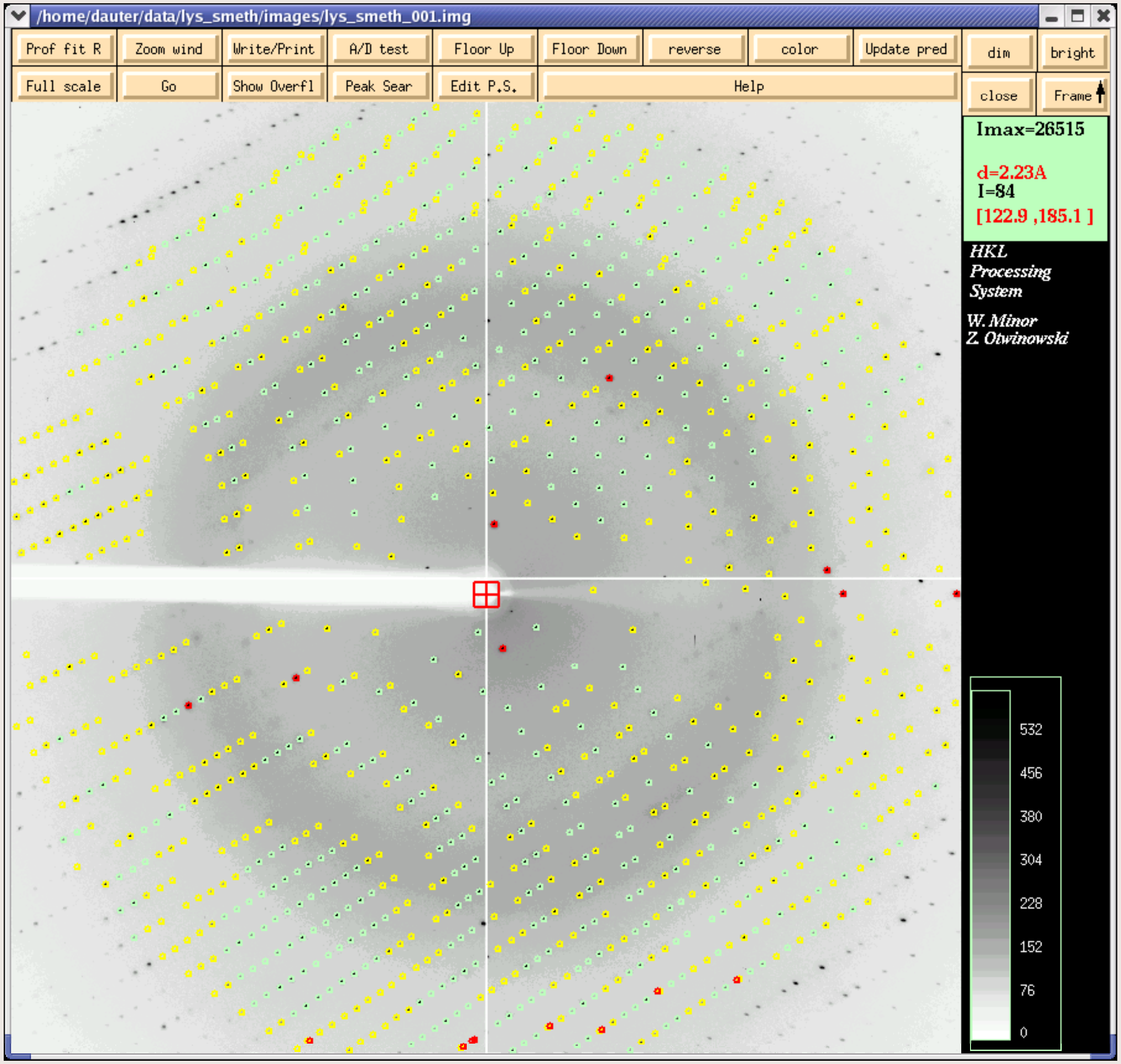
Cell length along the beam

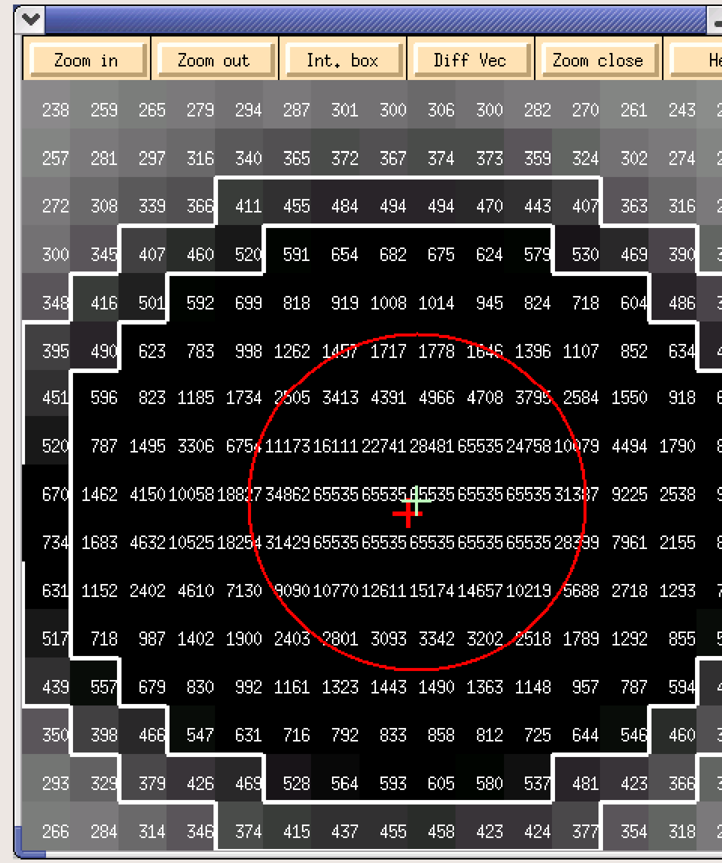




$$\Delta\varphi_{max} = \frac{180 \cdot d}{\pi \cdot a} - \eta$$

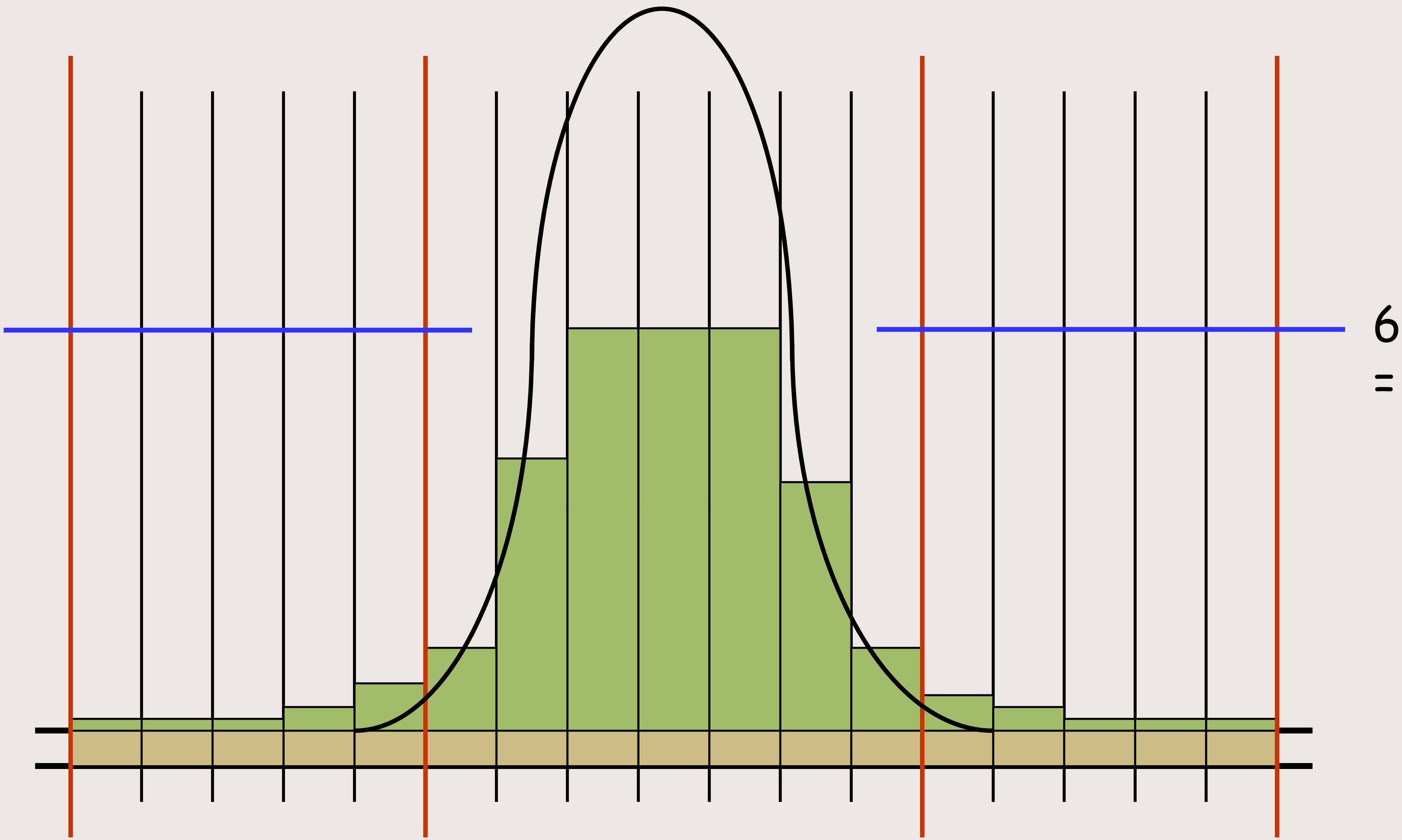
Overloaded profiles





Best, strongest reflections - very important fo

Overload extrapolated



standard profile fitted on shoulders

Missing reflections are never random

If missing reflections were spread randomly there would be no serious problem

However, for any of the above reasons they are always missing systematically

Kevin Cowtan's duck, back-transformed with missing segment of data:

Intensities (and their uncertainties)

Very easy to collect indices without intensities

however, intensities should be well measured
(accompanied by realistic error estimations)

Accuracy criteria of intensities

R_{merge} (R_{sym}, R_{int})

$$= \frac{\sum_{hkl} \sum_i |\langle I \rangle - I_i|}{\sum_{hkl} \sum_i I_i}$$

$I/\sigma(I)$ - generally > 2, or 50% of reflections > 3

Mulitiplicity – generally the higher the better

Improved versions of R_{merge}

R_{merge} - bad criterion from statistical point of vie
(depends on multiplicity)

Improved forms (unfortunately rarely used):

$$R_{meas} = \frac{\sum_{hkl} [n/(n-1)]^{1/2} \sum_i |\langle I \rangle - I_i|}{\sum_{hkl} \sum_i I_i}$$

(Diederichs & Karplu

$$R_{p.i.m.} = \frac{\sum_{hkl} [1/(n-1)]^{1/2} \sum_i |\langle I \rangle - I_i|}{\sum_{hkl} \sum_i I_i}$$

(Weiss & Hilgenfeld,

Criteria of anomalous data

Friedel mates can be treated as equivalent in scaling but kept separate on output

indicators of anomalous signal:

- higher χ^2 when Friedels merged
- higher R_{merge} when Friedels merged
- list of outliers

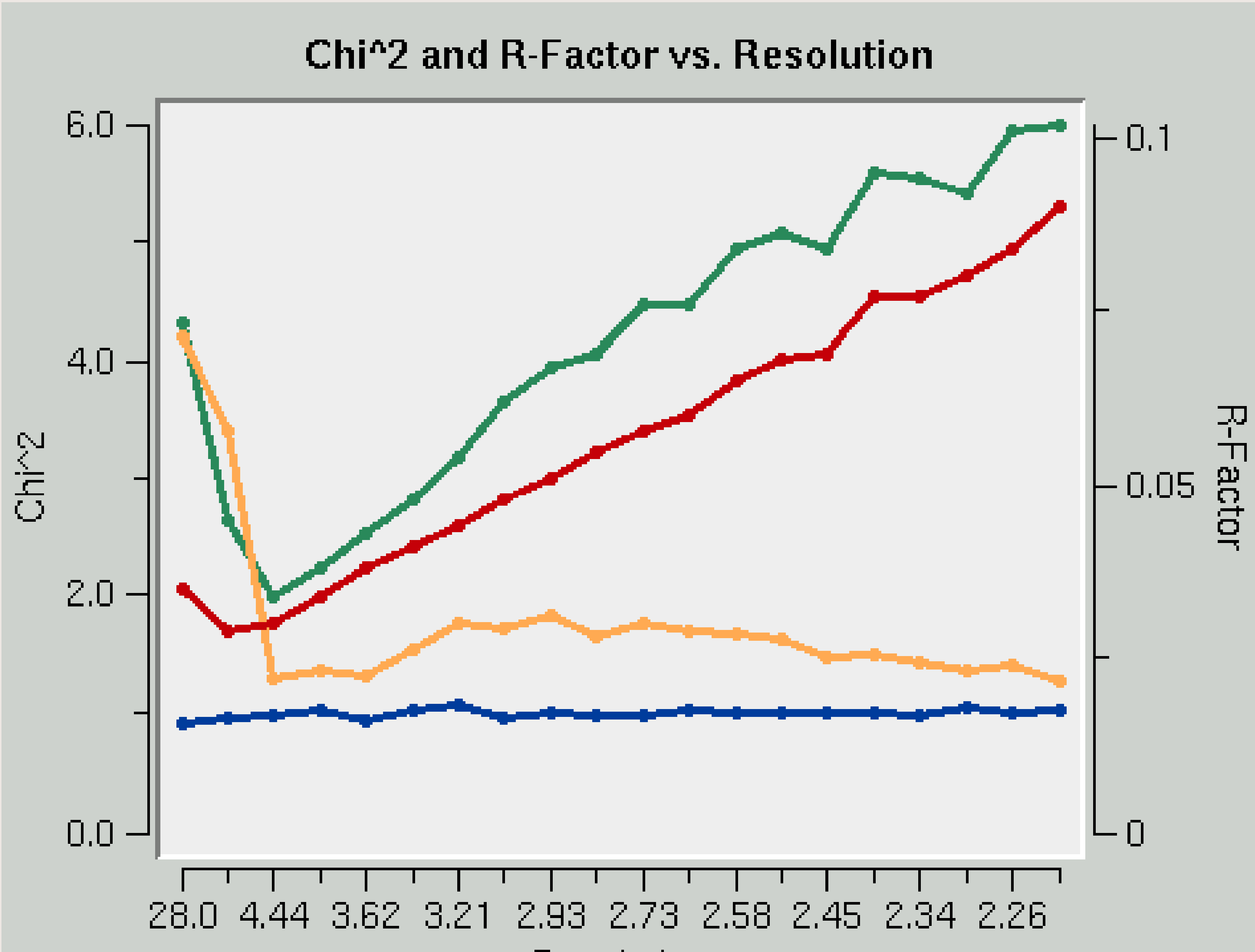
Numerical criteria of anomalous data

$$R_{anom} = \frac{\Sigma_{hkl} |I^+ - I^-|}{\Sigma_{hkl} |I^+ + I^-|/2}$$

$$\frac{\langle \Delta F \rangle}{\langle F \rangle} = \frac{\Sigma_{hkl} |F^+ - F^-|}{\Sigma_{hkl} |F^+ + F^-|/2}$$

$$\frac{\langle \Delta F \rangle}{\langle \sigma \Delta F \rangle} = \frac{\Sigma_{hkl} \Delta F}{\Sigma_{hkl} \sigma(\Delta F)}$$

R_{merge} and χ^2 (Ta_6Br_{12}-soaked cryst



Anomalous data - outliers

7	6	3		229.2	7456.0	7.00		
7	6	3	f+	418	-0.7	8669.6	0	416.1
-7	-6	3	a+	543	-0.5	8795.7	0	411.9
6	-7	3	f+	519	-0.5	8775.6	0	418.1
-6	7	3	a+	460	0.1	9060.5	0	430.4
6	7	-3	a+	553	0.0	9010.5	0	402.4
-6	-7	-3	a+	413	0.4	9189.9	0	429.8
-7	6	-3	f+	532	0.5	9231.9	0	413.8
7	-6	-3	a+	450	0.7	9341.0	0	421.3
-7	-6	-3	f-	404	-6.1	5935.6	100	399.1
7	6	-3	a-	567	-5.7	6154.0	100	400.1
6	-7	-3	a-	450	-6.4	5750.9	100	404.3
-6	-7	3	a-	532	-6.3	5850.3	100	401.6
6	7	3	f-	427	-6.6	5682.4	100	400.2
7	-6	3	a-	521	-6.5	5699.1	100	402.8
-7	6	3	a-	461	-6.6	5739.7	100	392.9

Standard uncertainties (σ's)

2-D detectors do not measure individual X-ray qua
but something proportional

therefore counting statistics is not valid

and σ's must be corrected for detector "gain"

$$\text{t-plot} = \frac{\langle I \rangle - I_i}{\sigma(I)}$$

average = 0.0, s.d. =

χ^2 criterion - agreement with expectations

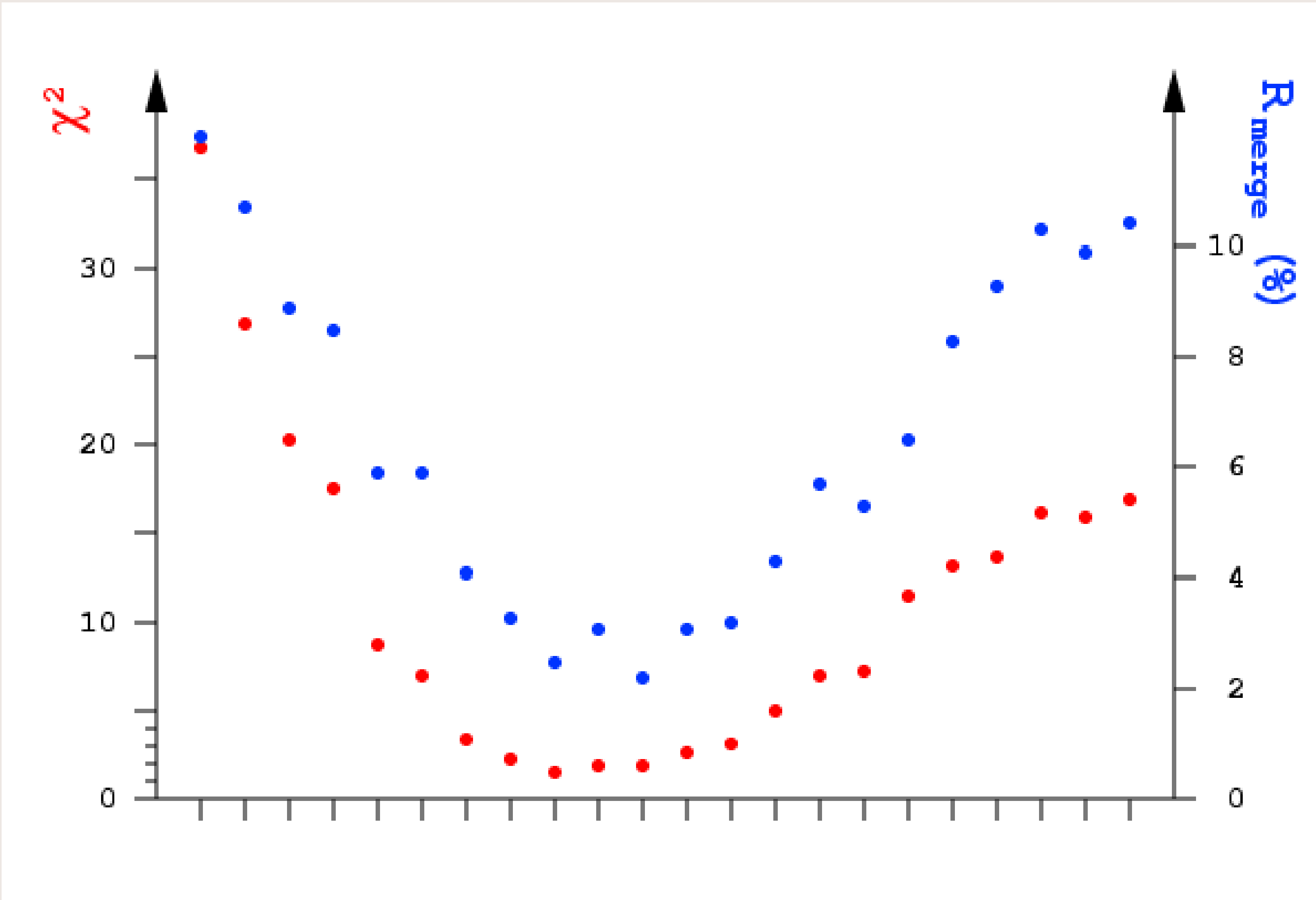
Multiplicity

More measurements of equivalent reflections lead to more accurate average and σ estimation

Also scaling and merging is more effective

But beware of radiation damage

Radiation damage



Typical syndrome of radiation damage –
first and last data do not agree with averages

Conclusions

X-ray data collection (with 2D detectors)

- scientific process, not technicality

- irreversible consequences (often)

- even more important due to progress in automation, phasing, refinement etc.

Always involves a compromise between time, redundancy, completeness etc.

- but it should be a wise compromise