Processing data collected with Pilatus/Eiger detectors

...

James Parkhurst

IUCR Computing School, Bangalore, August 2017

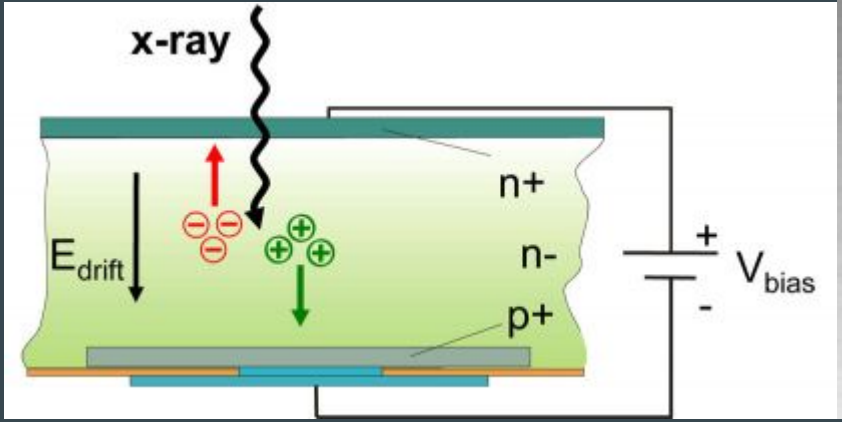
Introduction

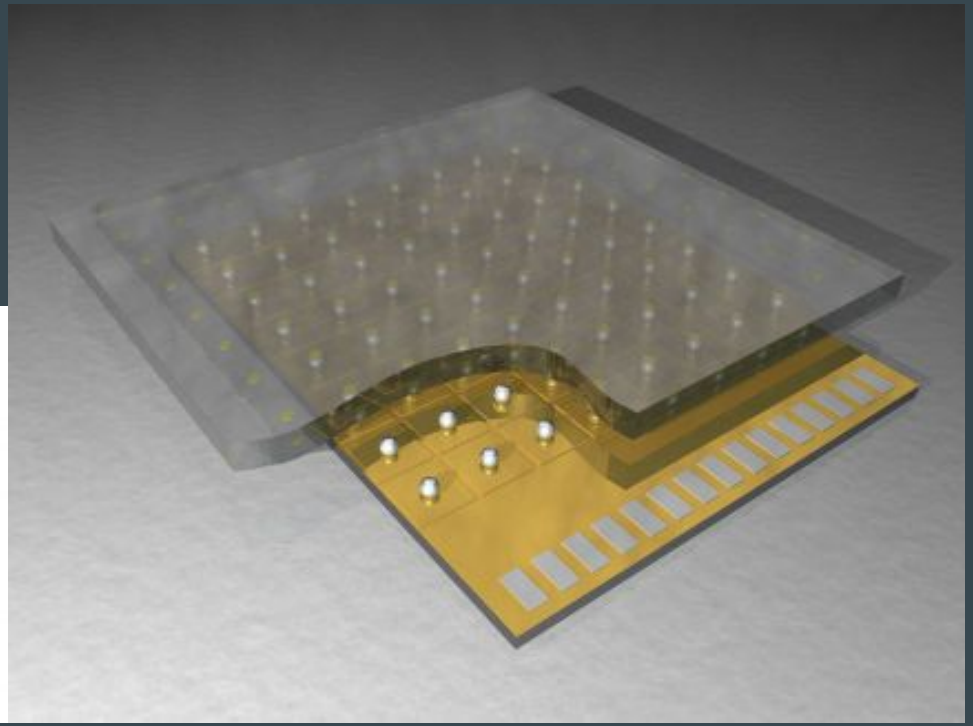
- Overview of Pilatus/Eiger detectors
- Overview of the DIALS integration program
- Data processing for Pilatus/Eiger detectors
  - Weak data
  - Spot finding
  - Background modelling
- Performance issues and parallelism

How does a Pilatus/Eiger detector work?

Sensor pixel: direct detection of X-ray photons -> one e-/hole pair per 3.6 eV.

Pixel electronics: counting of charge pulses.





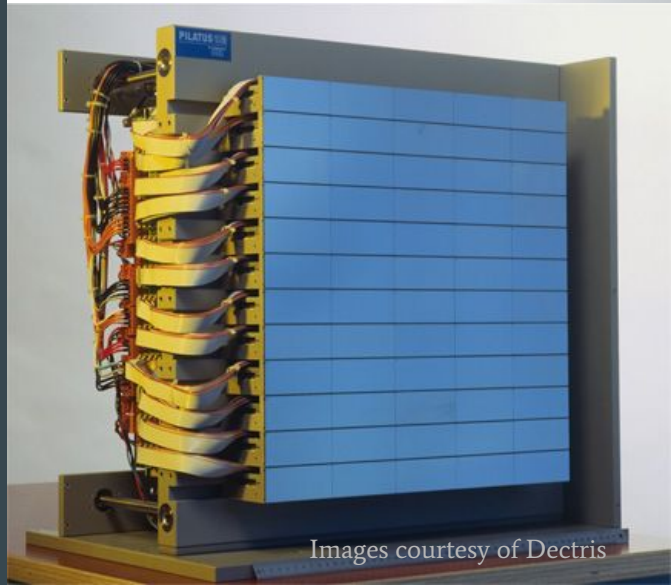
Images courtesy of Dectris

Modular detector

Pilatus/Eiger detectors are composed of modules: 8x2 array of CMOS ASICs

Each sensor module is a continuous 487 x 195 array of 94,965 pixels covering an active area of 83.8 mm x 33.5 mm

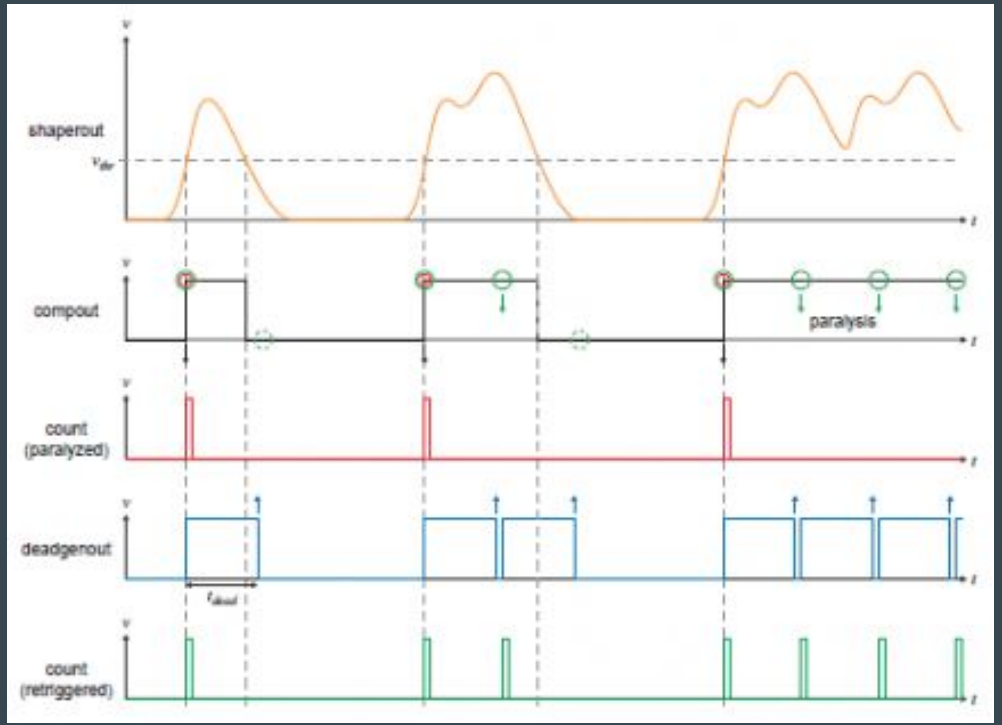
Modules are arranged to form larger detectors (Pilatus 6M contains 60 modules in a 5 x 12 grid)



Images courtesy of Dectris

High flux - retriggering

- Incident X-rays converted to electric charge
- Once charge is greater than a threshold, a count is registered.
- In “paralyzed” mode (Pilatus 2/Eiger), another count is only registered after the charge decreases below and increases above the threshold again.
- In “retriggering” mode (Pilatus 3), if the charge stays above the threshold, another count is registered after a certain time.



Count rate

- Due to the counting process there is a small dead time after each hit
- This becomes significant at high flux where some counts are lost
- The measured count rate is linear up to about 1Mcps

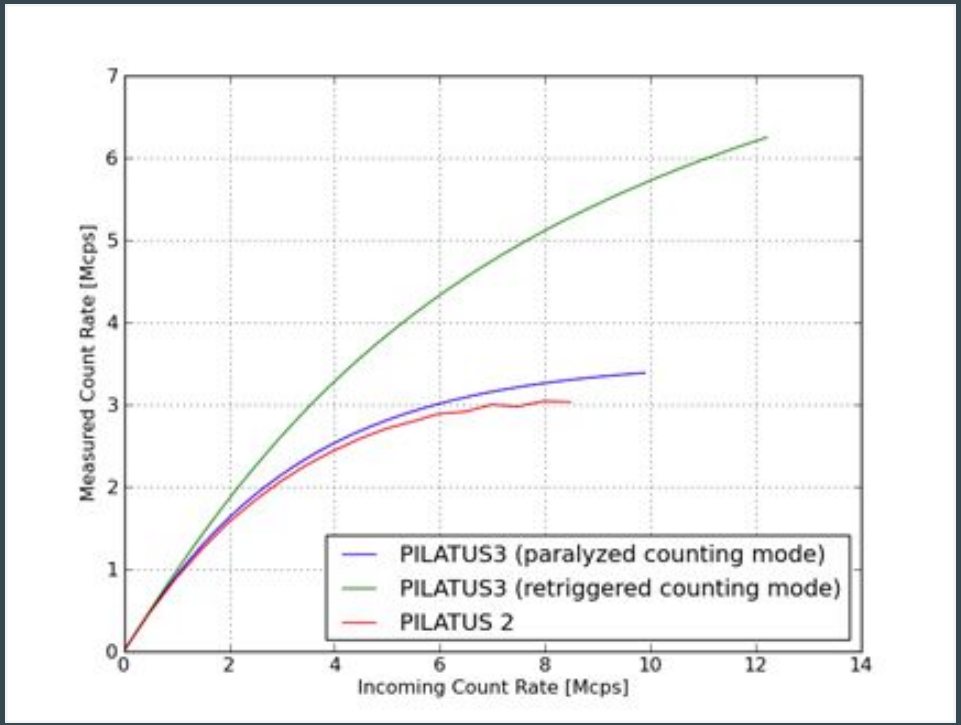


Image courtesy of Dectris

Virtual pixels

- Each module contains an 8x2 array of CMOS readout chips
- Chips have a small gap between them.
- This is spanned by 2 larger pixels (1.5x size of normal pixel)
- Counts are distributed into three “virtual pixels” after readout.
- The counts in the virtual pixels are therefore correlated.

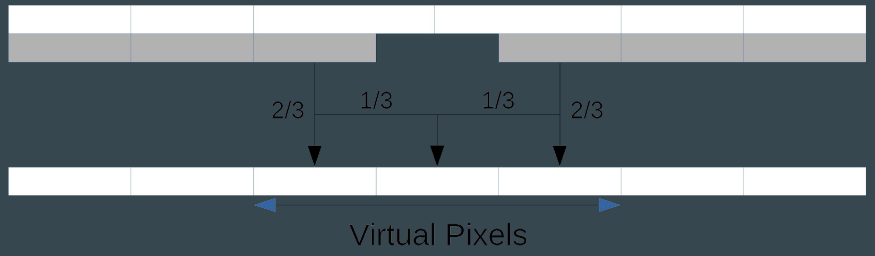


Image courtesy of Dectris

Pixel array detectors

- Direct detection of X-rays
- Single-photon counting
- Good signal-to-noise ratio and high dynamic range (zero dark signal, zero noise)
- Low-energy X-ray suppression (energy resolution by single energy threshold)
- Short readout time and high frame rates
- Modular detectors enabling multi-module detectors with large active area

DIALS

Acknowledgements

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

XDS

Wolfgang Kabsch

Max-Planck-Institut für Medizinische Forschung, Abteilung Biophysik, Jahnstrasse 29, 69120 Heidelberg, Germany

Correspondence e-mail: wolfgang.kabsch@mpimf-heidelberg.mpg.de

The usage and contr
package *XDS* for
described in the con
include automatic d
range and recogniti
Moreover, the limita
number of correction
pixel contents have b
been restructured fo
and completeness of
measurement.

1. Functional spec

The program packag
developed for the re
recorded on a plana
monochromatic X-ra
XDS accepts a
rotation images fron
and multiwire area
metrics and produce
of the reflections occ
way. The program a
positive amount of c
incident beam and cr
imposes no limitati
directions of the rot
oscillation range cov

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The finer things in

J. W. Pflugrath

Molecular Structure Corporation, 9009 New Trails Drive, The Woodlands, TX 77381, USA

Correspondence e-mail: jwp@msc.com

X-ray diffraction images from
sensitive detectors can be cha
depending on whether the rotati
is greater than or less than the c
The expectations and consequen
and thin images in terms of spa
X-ray background and $I/\sigma(I)$
software suite for processing
introduced, and results from *d*
those from another popular pacl

1. Introduction

Two-dimensional position-sensiti
for many years in X-ray diffract
cular, data from crystals of macr
oligonucleotides and their cor
acquired with an area detector
obsolete), a multi-wire system
recently commercialized char
coupled to a phosphor-coated fit
detectors, the crystal, centered in
oscillated around a single axis th
~2.0°, while counts from diffrac
for a specified time. At the en
detector is read out and the cou
two-dimensional array with each
to a distinct position on the d

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The integrati

Andrew G. W. Leslie

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England

Correspondence e-mail: andrew@mrc-lmb.cam.ac.uk

The objective of any
produce from a set of
with their associated
uncertainties), togeth
crystal unit-cell param
reliable, but should i
intervention. The pro
three stages. The first
parameters and the c
parameters may indic
The second step is to r
parameters and also t
known as post-refiner
images, which consists
reflections on each in
intensity of each refle
out while simultaneou
parameters. Basic fea
each of these three s
with reference to the

1. Introduction

The collection of mac
gone dramatic advan
advent of two-dimensi
and CCDs, crystal cry
monochromatic and

Centre National de la Recherche Scientifique
Université Paris-Sud

Laboratoire pour l'Utilisation du Rayonnement Electromagnétique

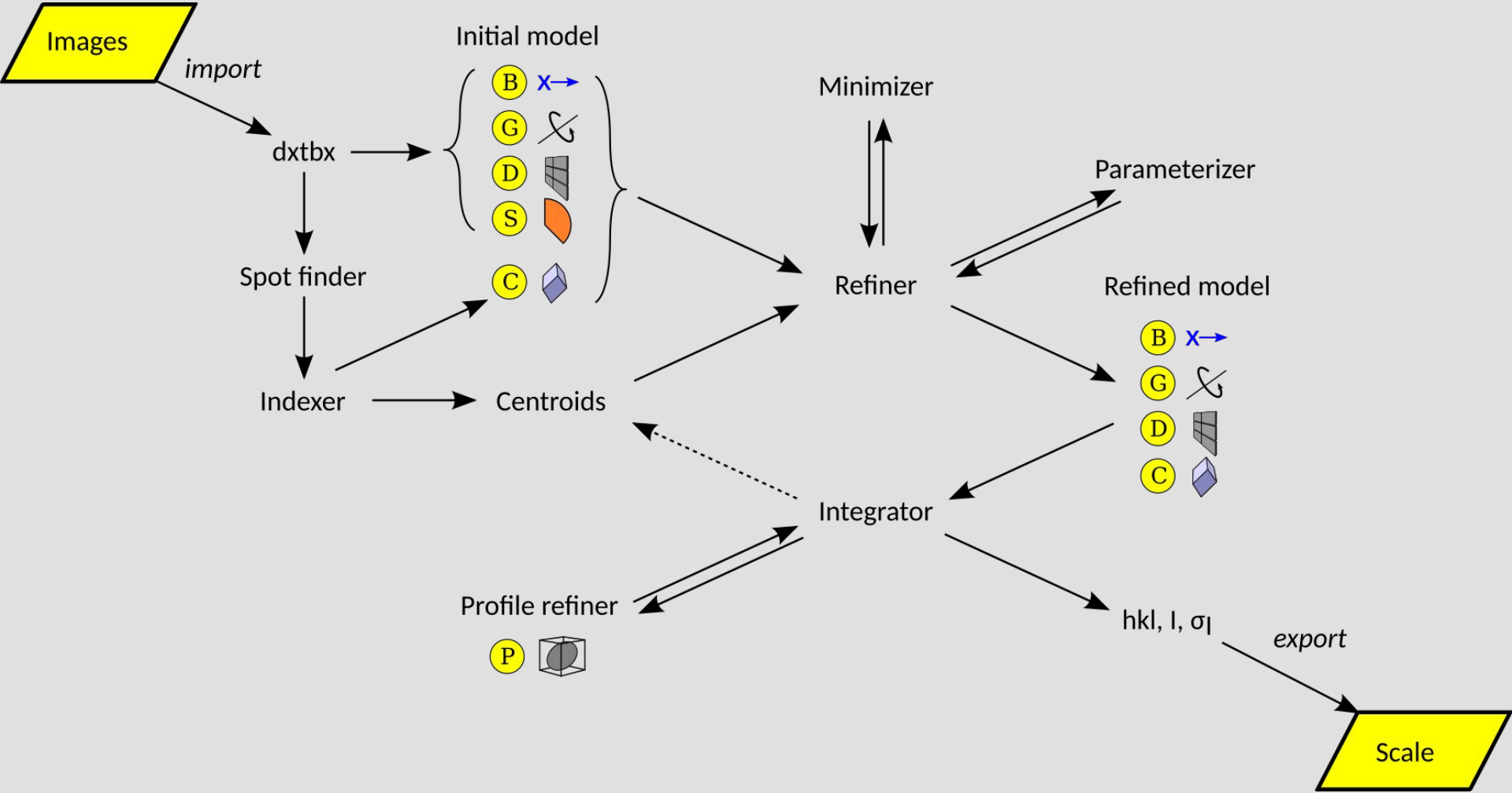
Proceedings

of the EEC Cooperative Workshop

on Position-Sensitive Detector Software

(Phases I & II)

held at L.U.R.E. from May 26 to June 7, 1986.


Images
import
dxtbx
Initial model
B
G
D
S
C
Spot finder
Indexer
Centroids
Minimizer
Refiner
Parameterizer
Refined model
B
G
D
C
Integrator
Profile refiner
P
hkl, I, σI
export
Scale

Transitions

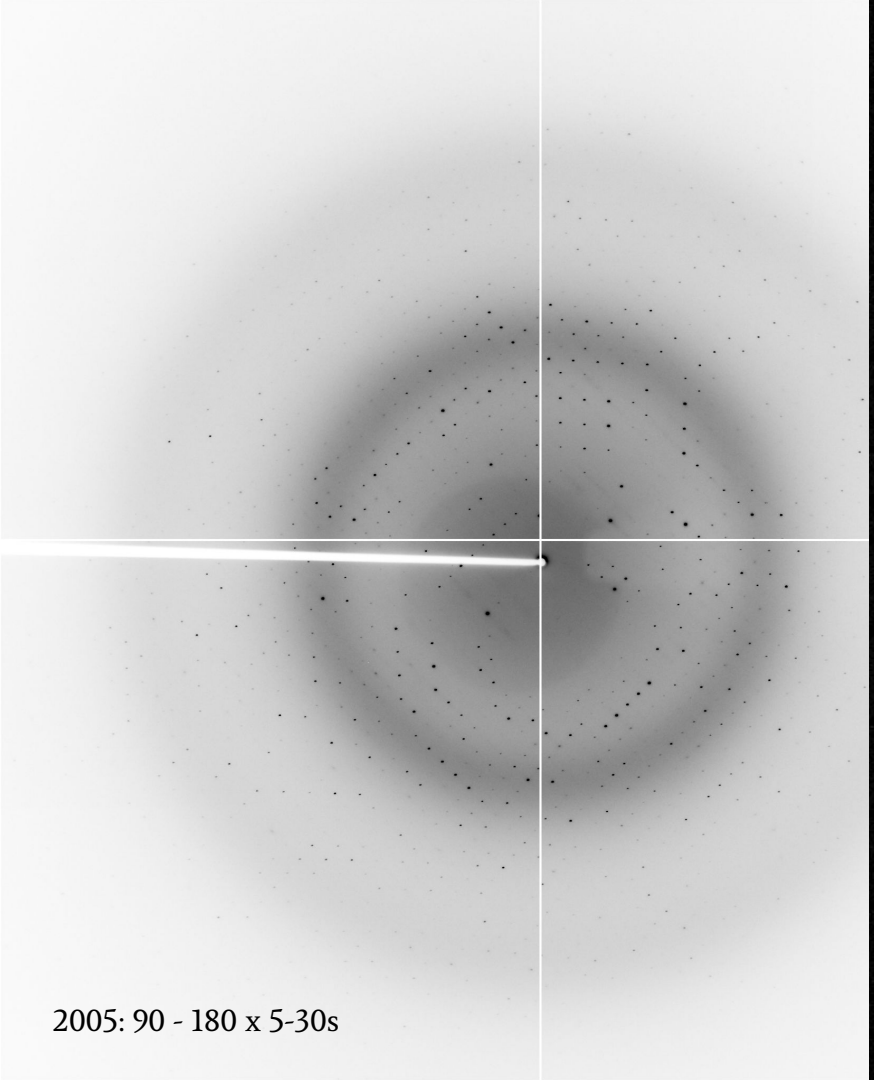
CCD → PAD: Pilatus → PAD: Eiger

PILATUS3 S and X product pages ...

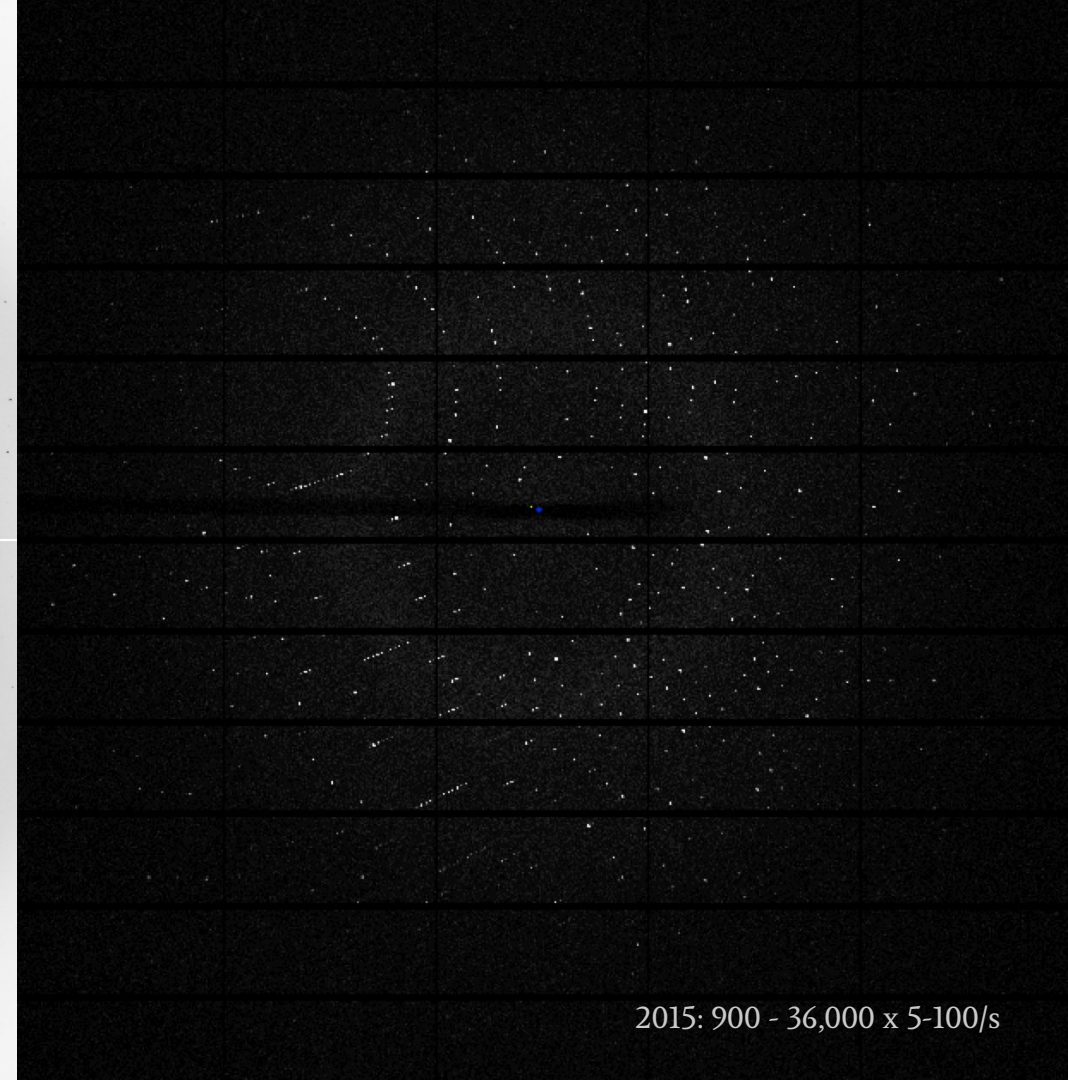
EIGER X product pages...

New Algorithms

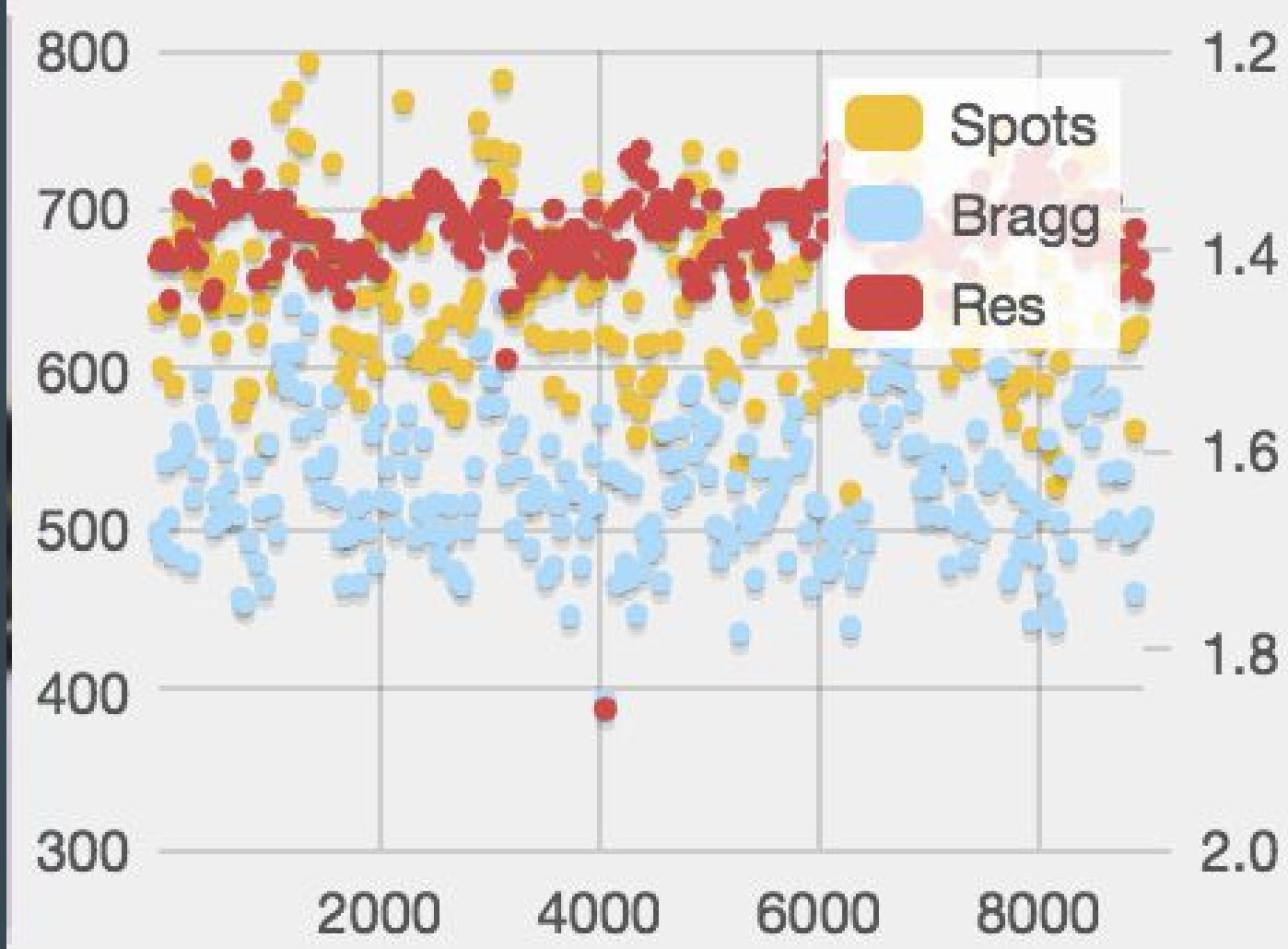
New infrastructure



2005: 90 - 180 x 5-30s



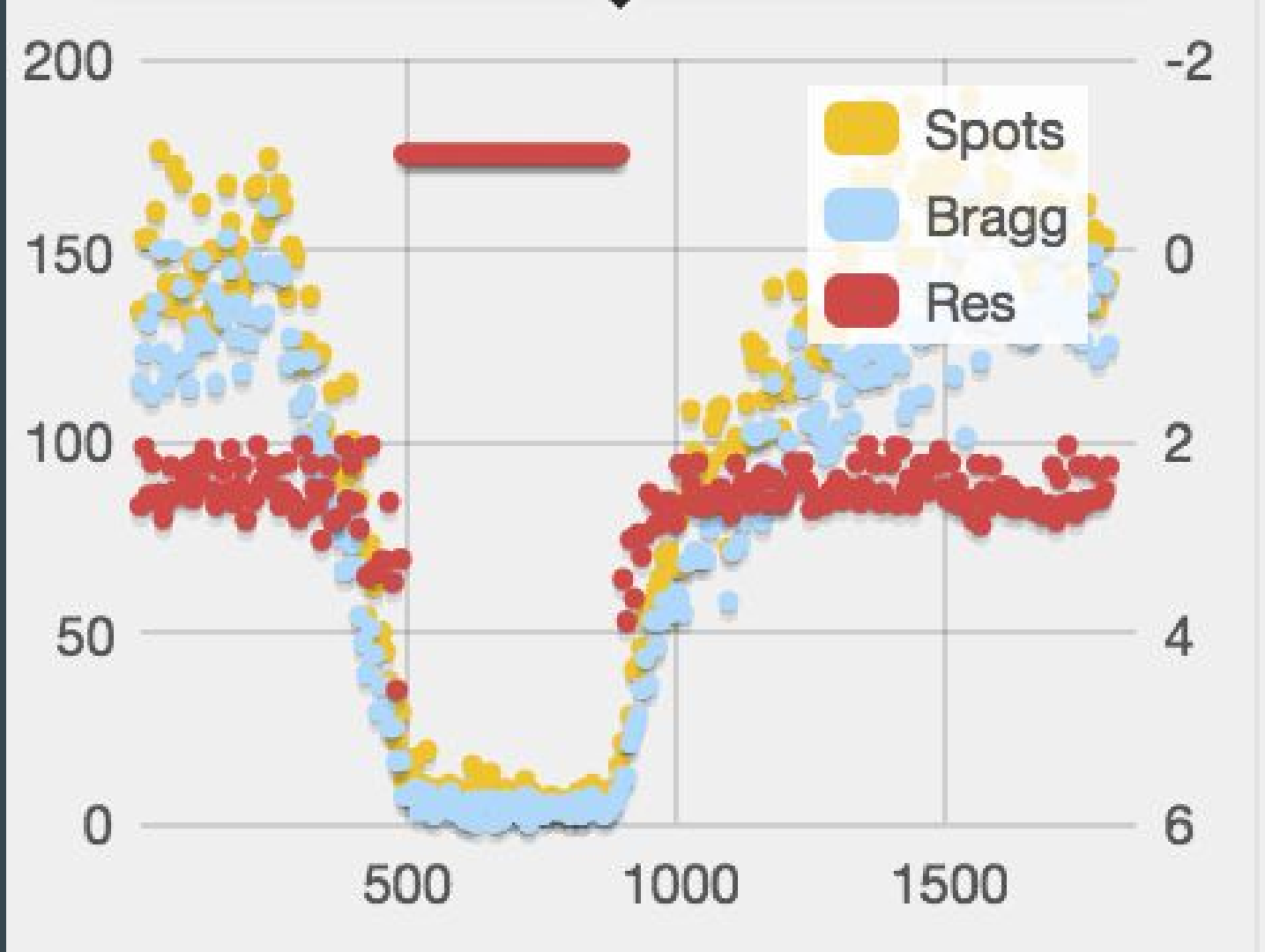
2015: 900 - 36,000 x 5-100/s



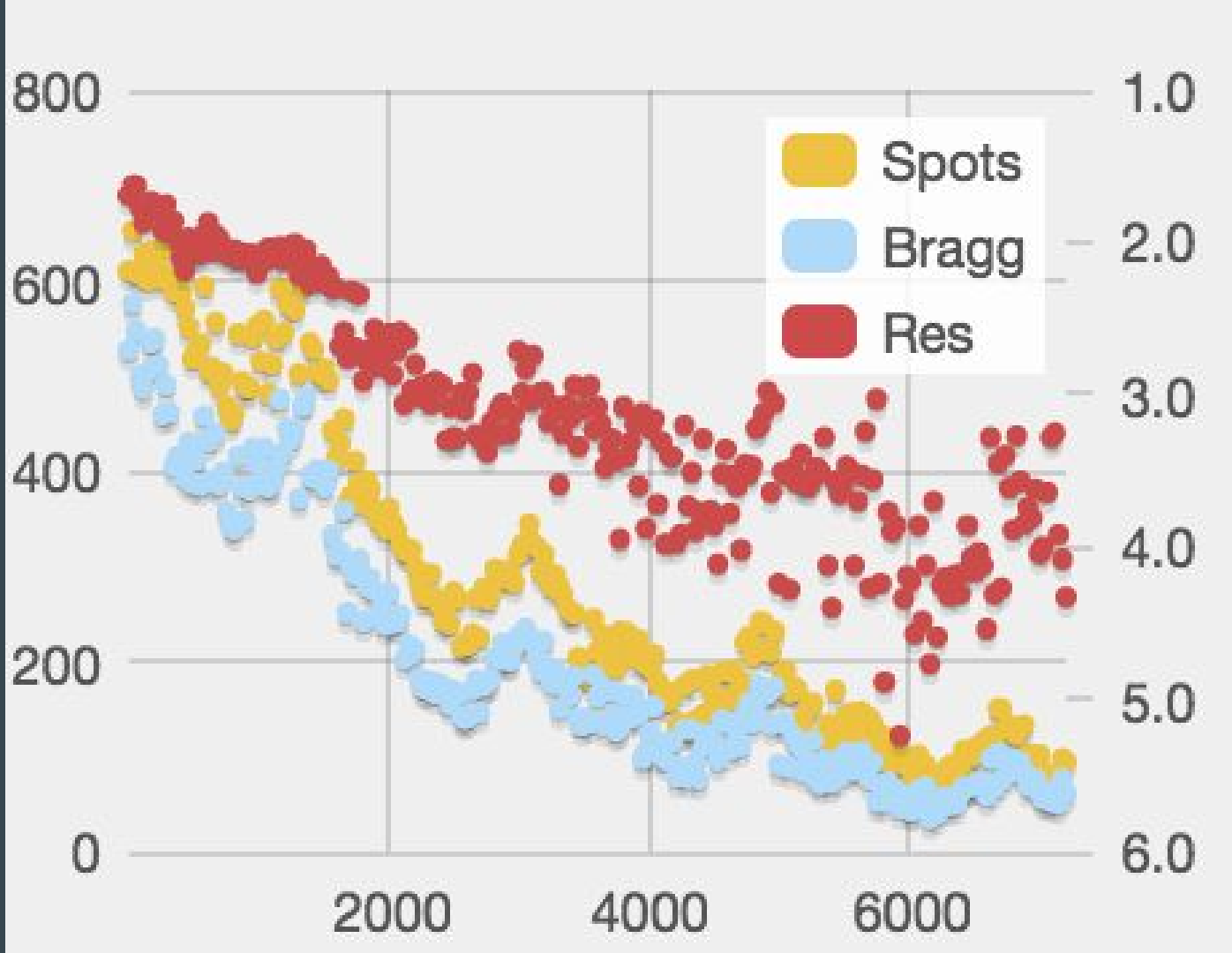
Good - lots of indexed spots

Bad - crystal leaving beam

Could treat as two sweeps



Bad - radiation damage



To what resolution do my spots extend?

Spot finder view of the image

Fine slicing: single image

Fine slicing: single image + 50 images summed

Fine sliced, weak data can obscure presence of ice rings

Complex detectors: DLS BL-I23

Traditionally, integration programs supported collection from a single flat panel detector.

Modern integration programs need to support multi-panel detectors with complex configurations such as the Pilatus DLS 12M @ Diamond beamline I23

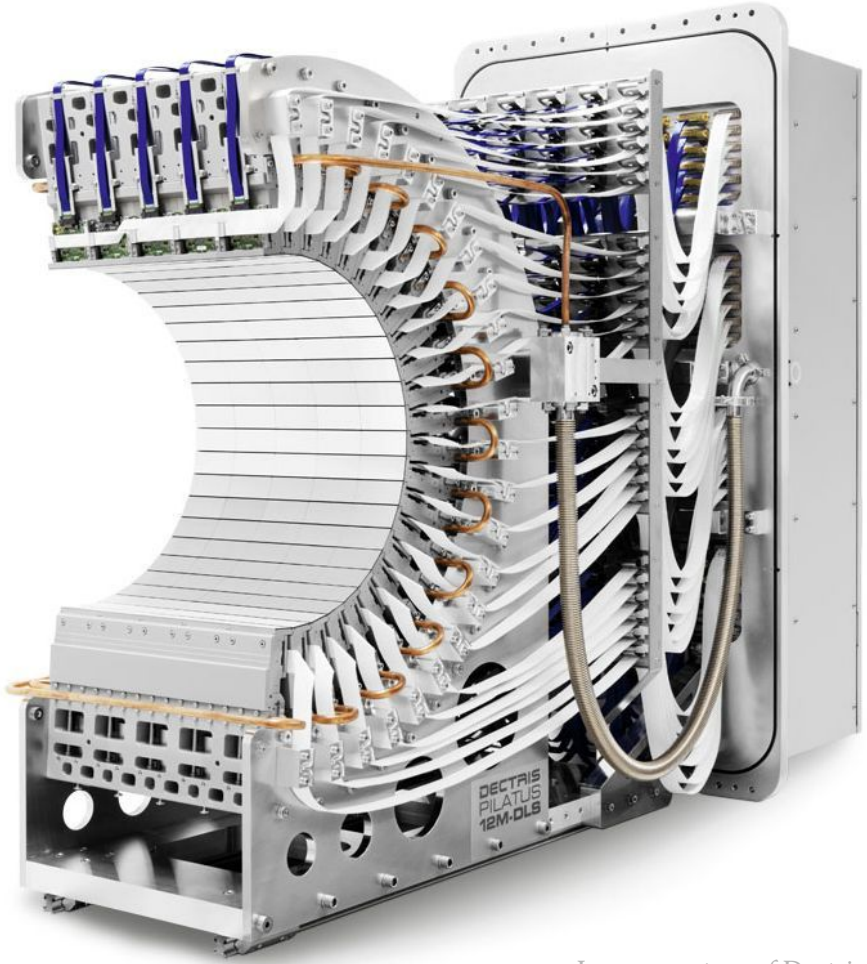
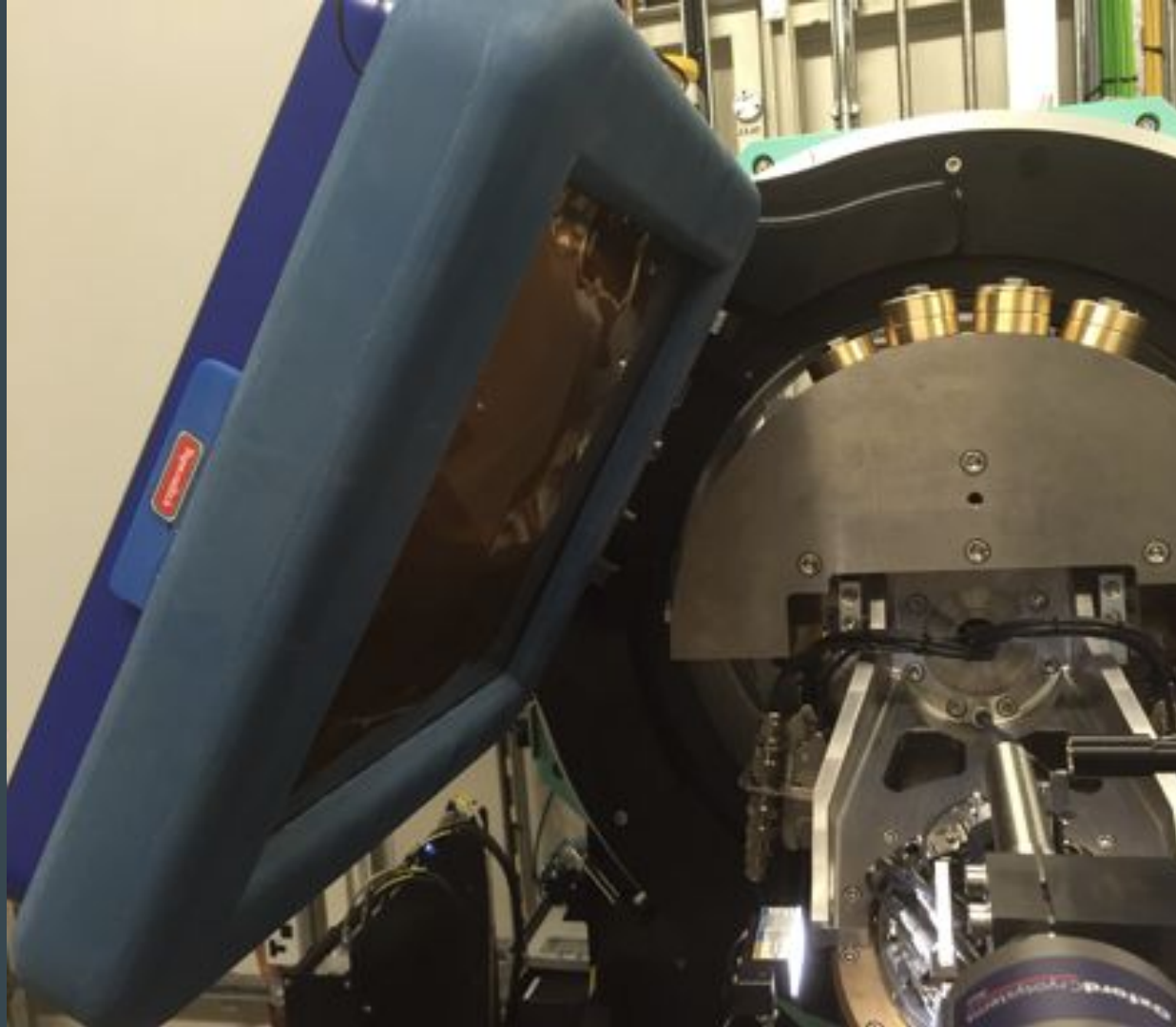


Image courtesy of Dectris

Complex detectors: DLS BL-I19

Detector is mounted on a goniometer so it can be positioned around the sample - including vertically above the sample (i.e. 90 degree to the incident beam).

The familiar concept of the “beam centre” is not really appropriate for this scenario.



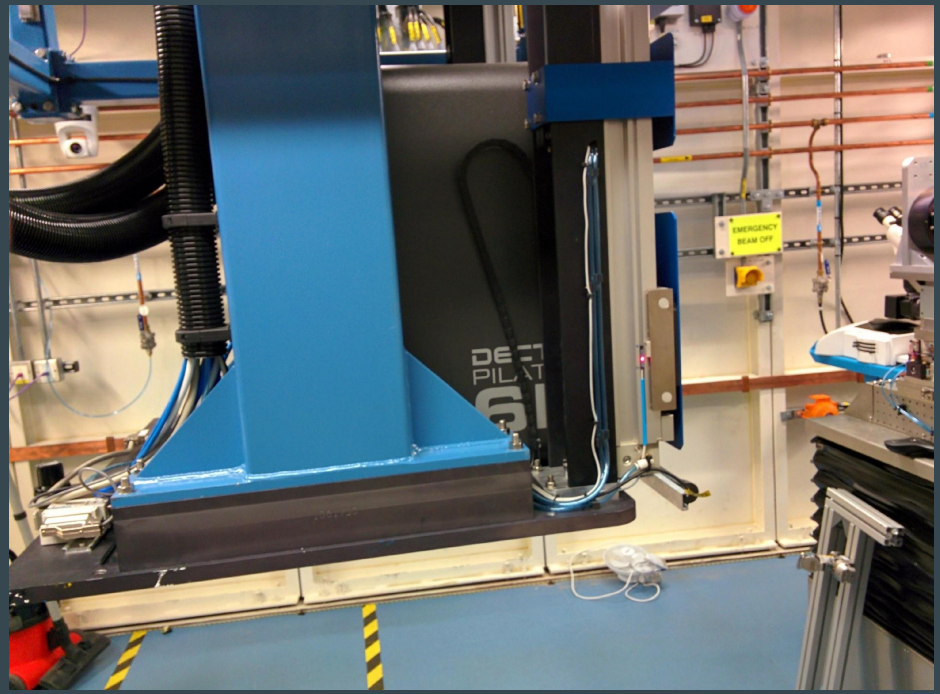
CCD -> PAD: weak data

Data collection with CCD:

- Need to balance signal to readout noise, dark current, etc.
- Therefore need to collect strong data to get good I / Sig(I).

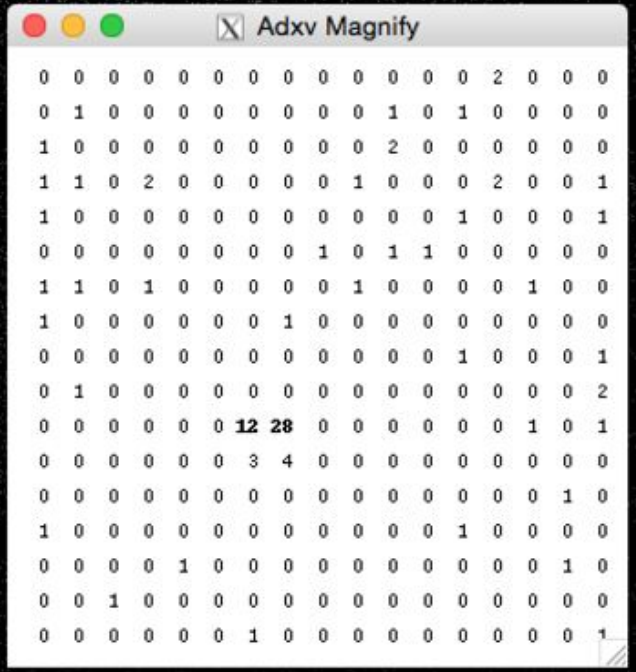
Data collection with PAD (Pilatus/Eiger):

- Very low readout noise in detection process means no compromise necessary
- Therefore dose / radiation damage can be spread around reciprocal space more uniformly.
- We can collect weak data with a very low background.



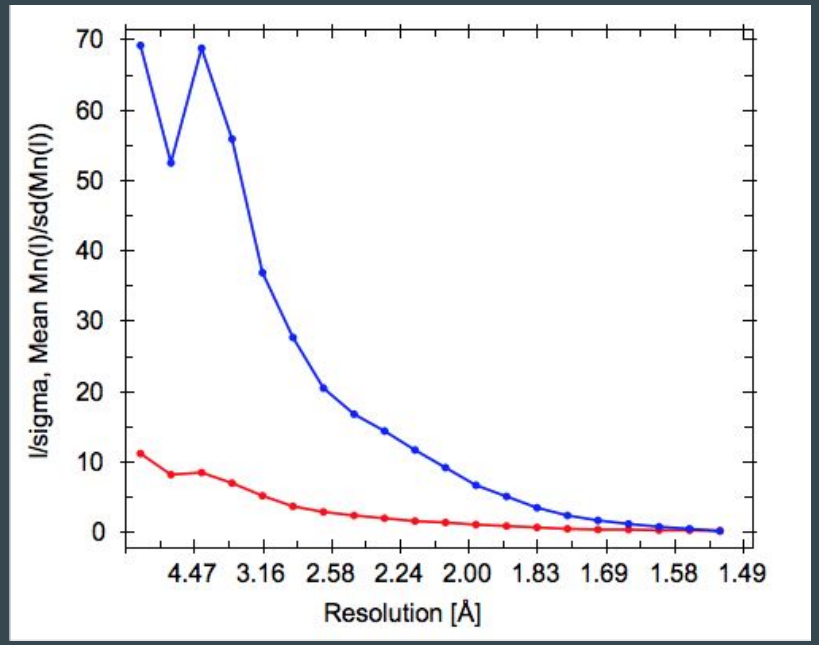
Strong low resolution spot

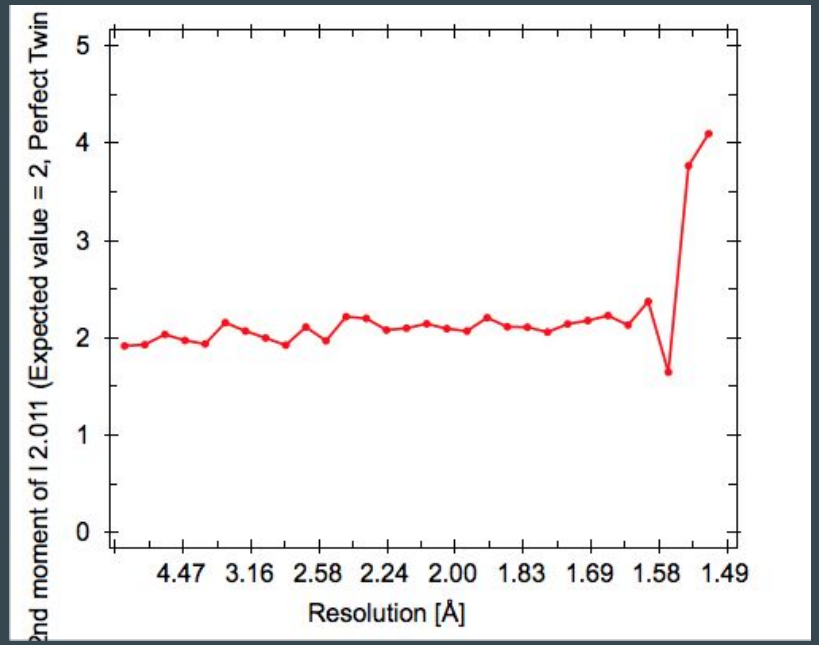
Example: Thaumatin recorded on I03, Pilatus2 6M with 0.05% transmission. Most pixels are zero.



Data processing

- I/sig(I) well behaved, tend to 0 at high resolution
- 4th moment of E well behaved to high resolution
- Rmerge in outer shell crazy, but data broadly good, Rpim overall < 1%



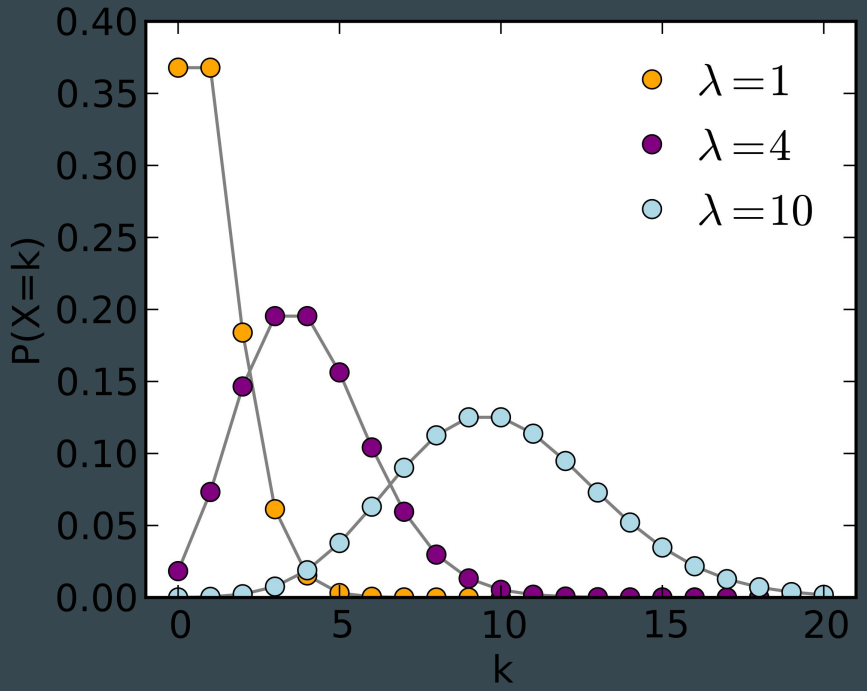


Poisson Distribution

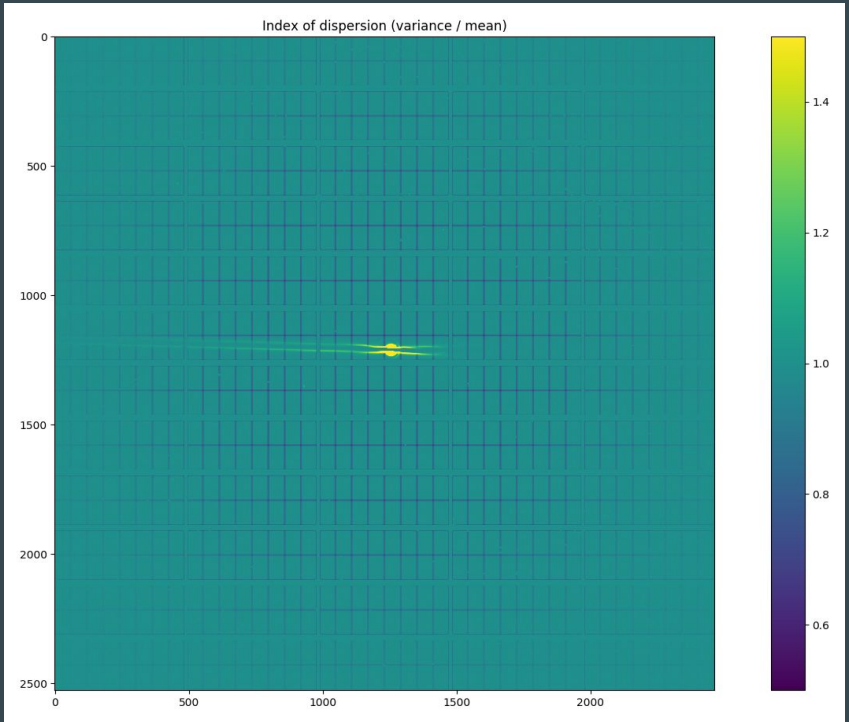
- Approximately normal for large lambda
- Not the case for small lambda

Useful properties

- variance = mean
- D = variance / mean = 1
- D is chi-squared distributed with n (n-1) degrees of freedom
- variance of D = 2 / (N-1)



Pixel array detectors: statistics



Analysed 9000 blank images and computed the index of dispersion (D = variance / mean) at each pixel.

For a Poisson distribution variance = mean, so we expect D = 1

Background data is Poisson distributed

Virtual pixels show under-dispersion due to correlations with neighbouring pixels

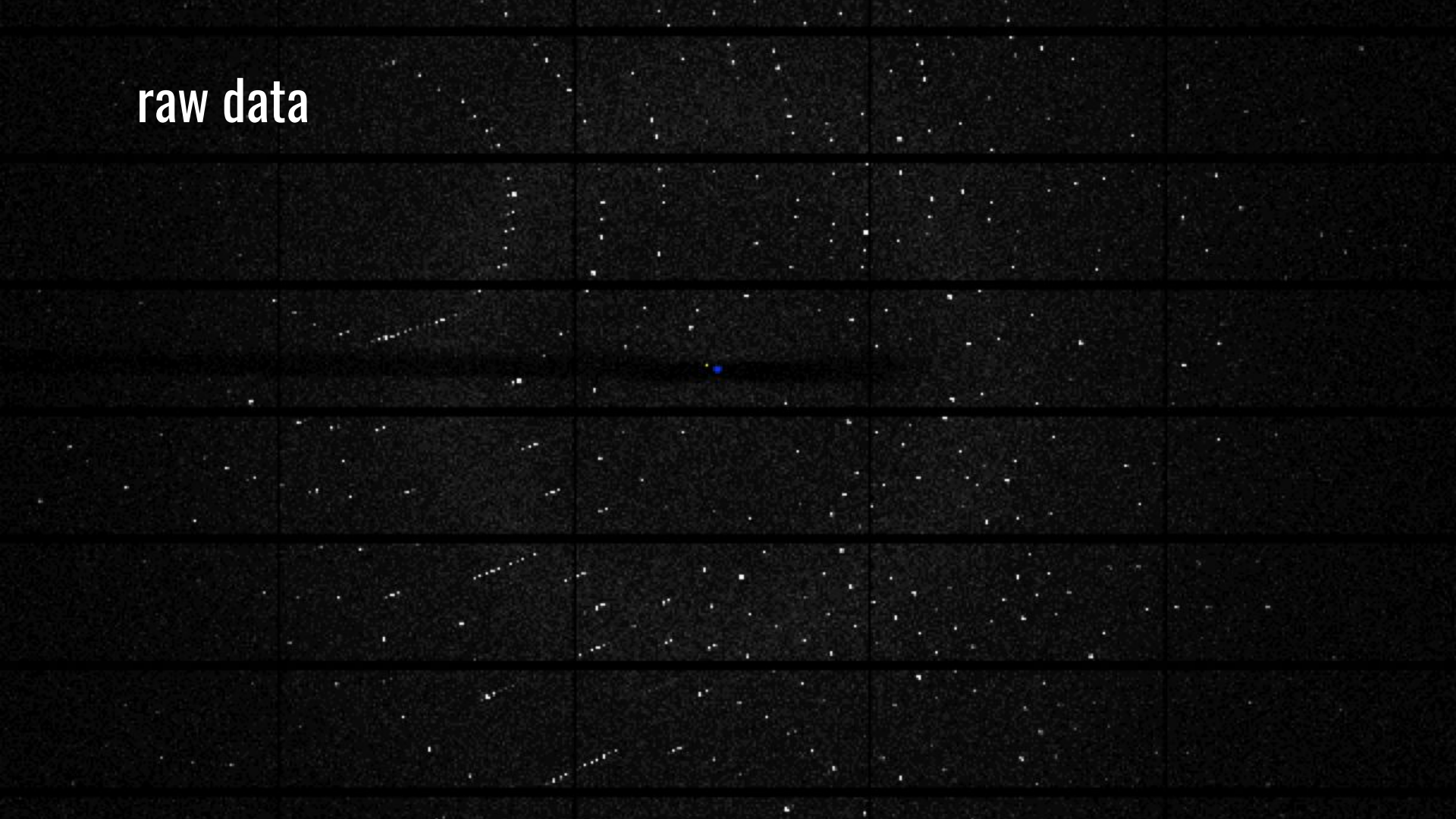
~7.2% of pixels are affected

Spot finding

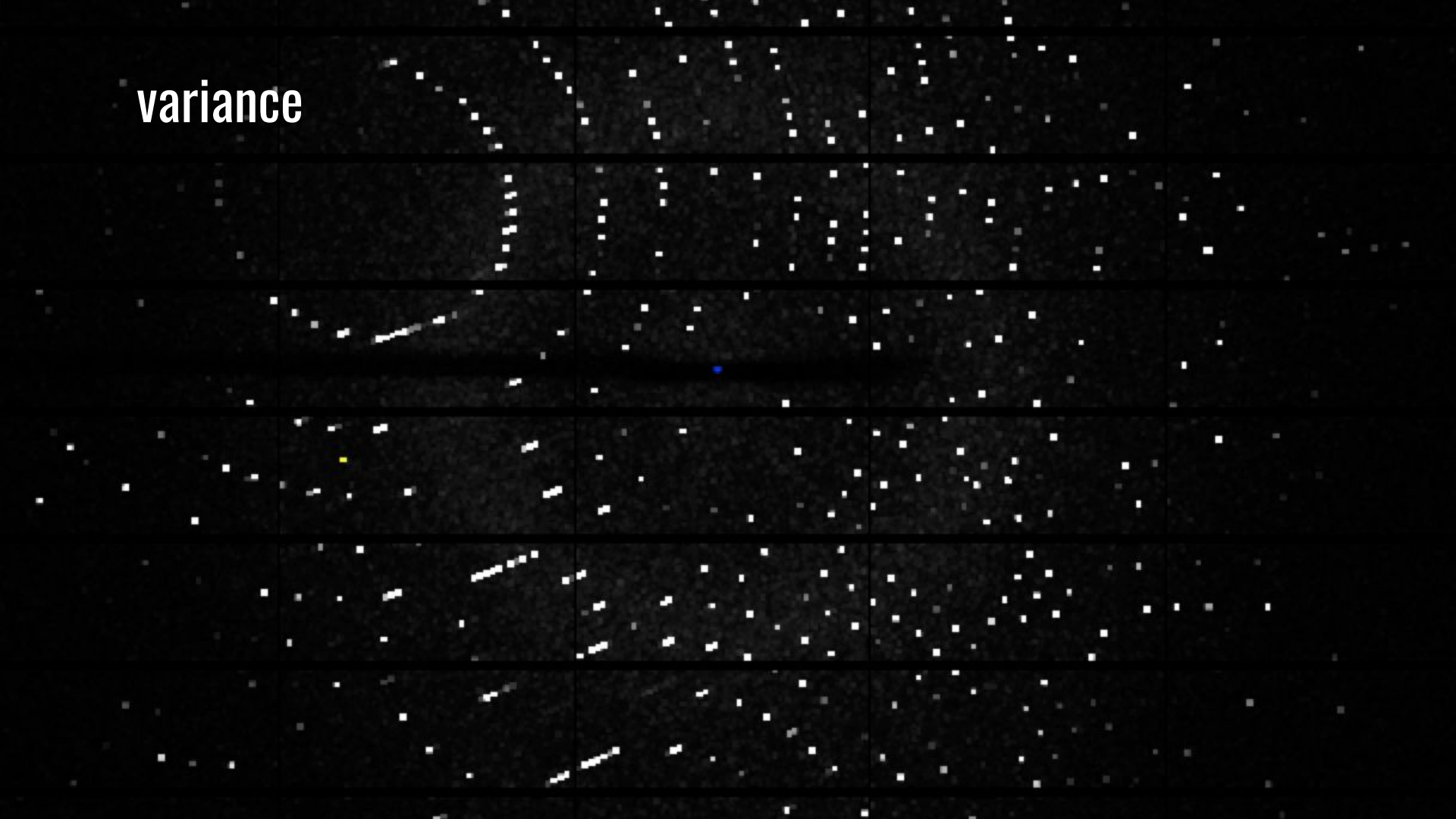
dials.find_spots

- Sequence of per-image filters to find strong pixels
- 3D analysis of strong pixels to identify strong spots
- Filter spots by
  - number of pixels
  - peak-centroid distance
  - resolution
  - ice rings
  - untrusted regions

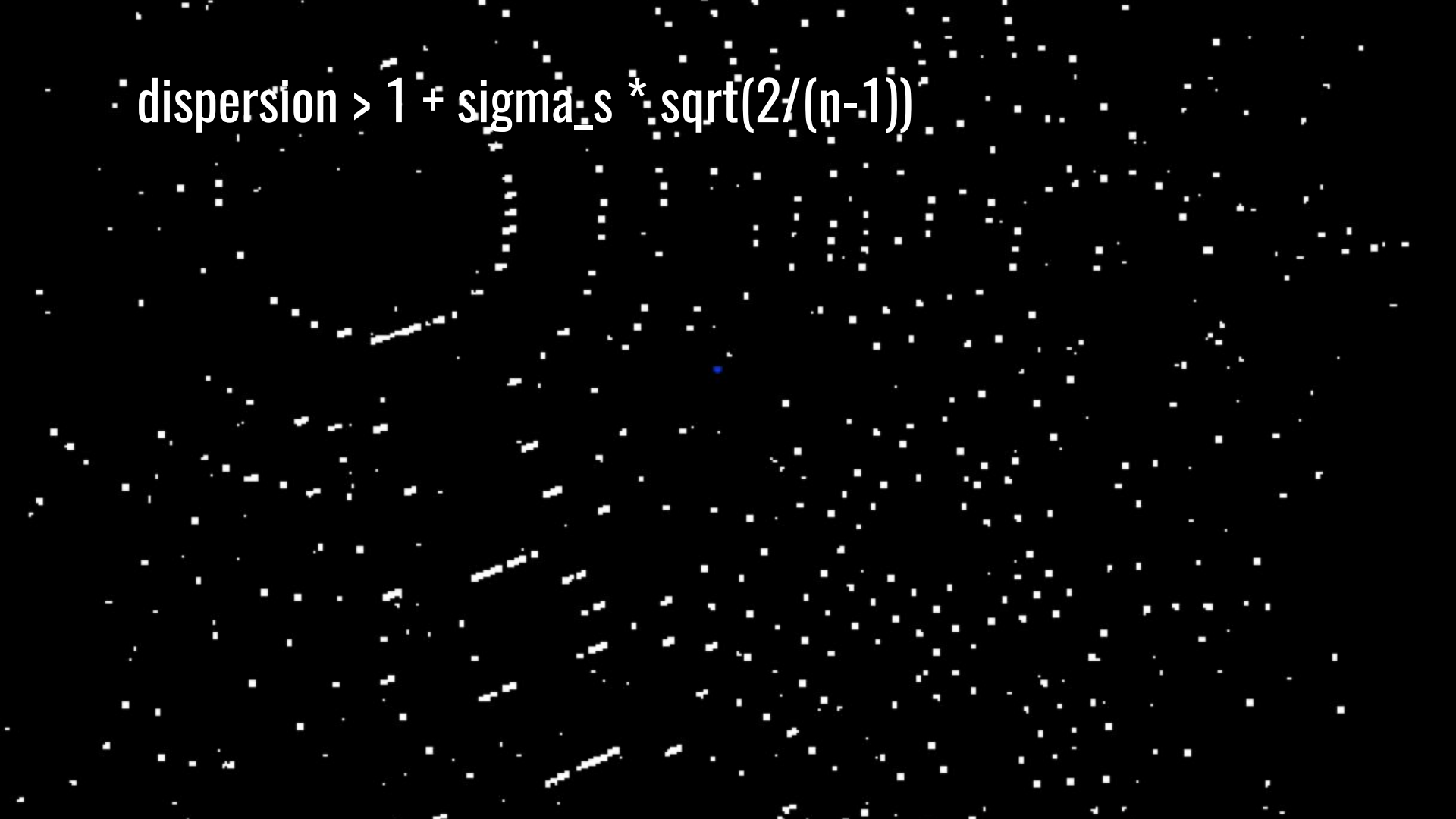
```
$ dials.find_spots datablock.json nproc=8

Setting spotfinder.filter.min_spot_size=3
Configuring spot finder from input parameters
-----------------------------------------------
Finding strong spots in imageset 0
-----------------------------------------------
Finding spots in image 1 to 540...
Extracting strong pixels from images (may take a while)
Extracted strong pixels from images
Merging 8 pixel lists
Merged 8 pixel lists with 922120 pixels
Extracting spots
Extracted 219125 spots
Calculating 219125 spot centroids
Calculated 219125 spot centroids
Calculating 219125 spot intensities
Calculated 219125 spot intensities
Found 1 possible hot spots
Found 1 possible hot pixel(s)
Filtering 219125 spots by number of pixels
Filtered 116321 spots by number of pixels
Filtering 116321 spots by peak-centroid distance
Filtered 116082 spots by peak-centroid distance

-----------------------------------------------
Saving 116082 reflections to strong.pickle
Saved 116082 reflections to strong.pickle
Time Taken: 31.768495
```

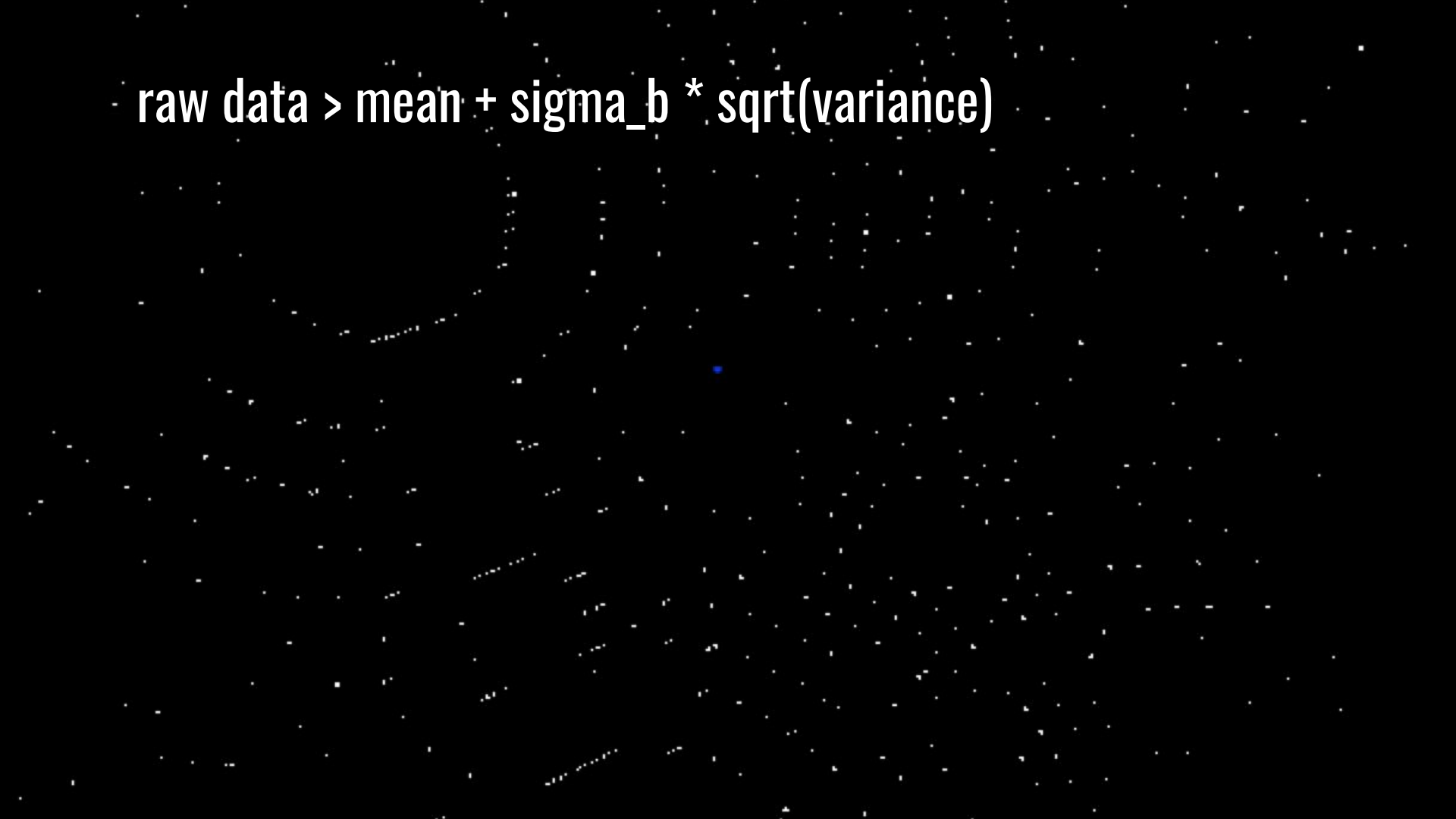

raw data

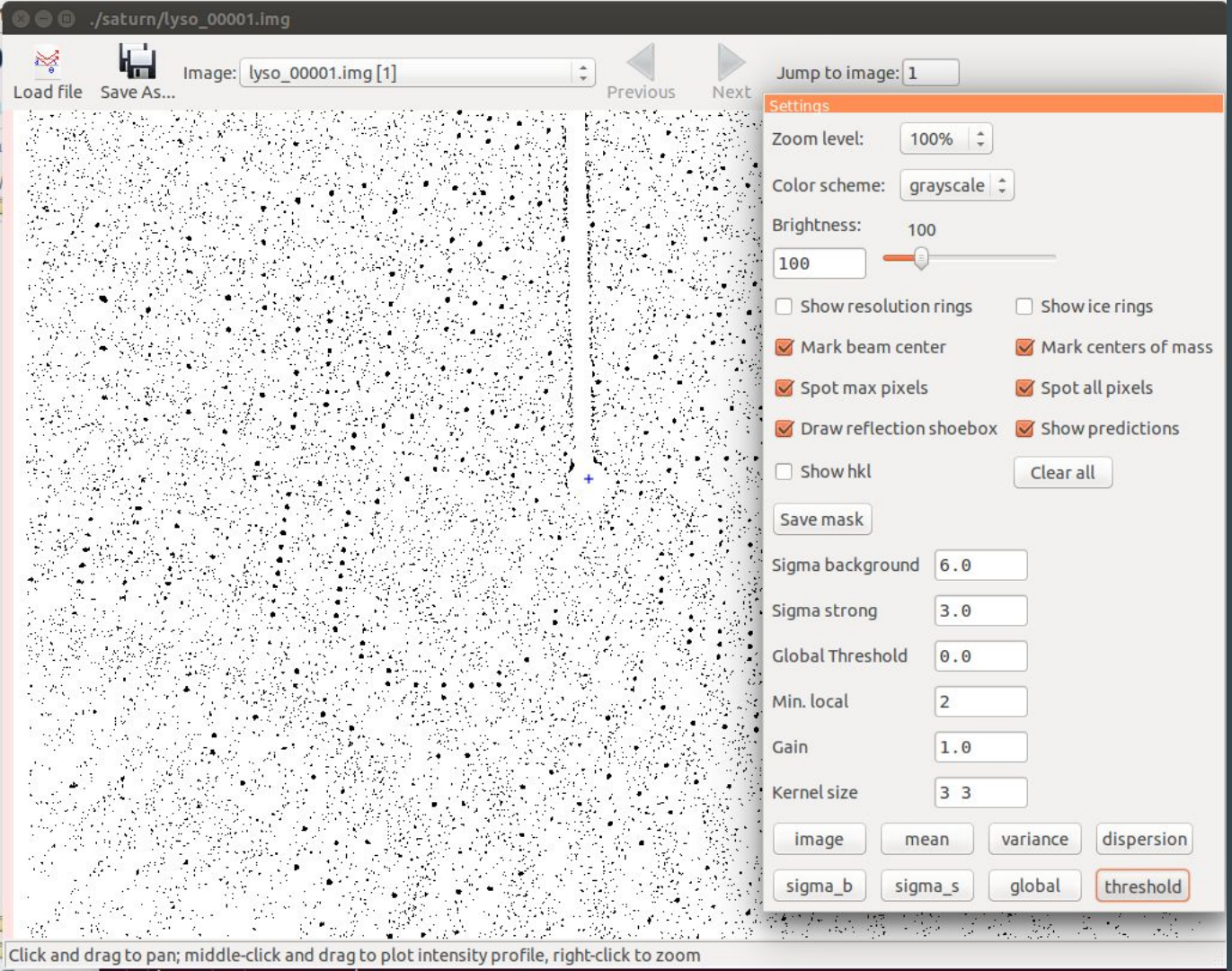
mean


variance

dispersion = variance / mean

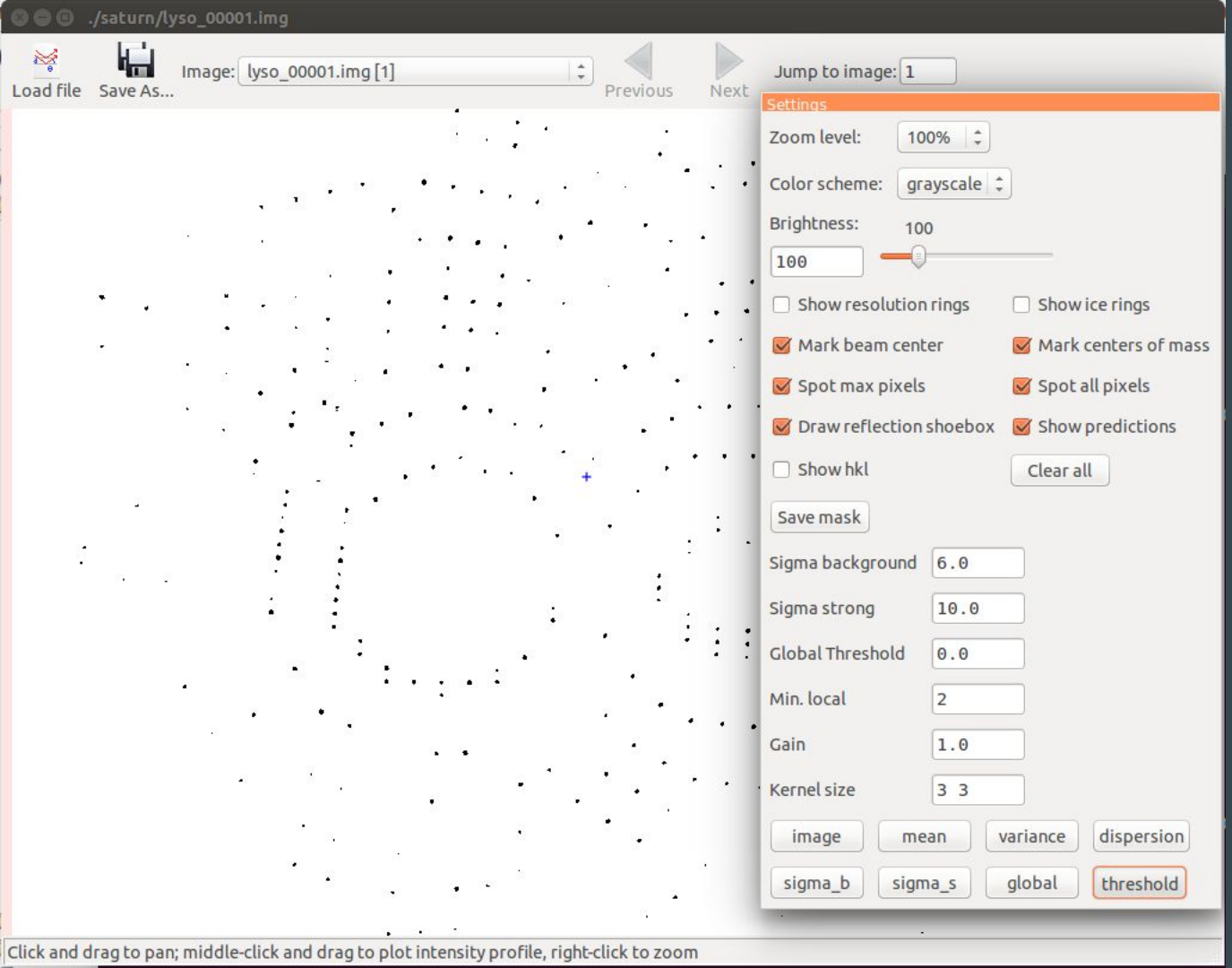

dispersion > 1 + sigma_s * sqrt(2/(n-1))


raw data > mean + sigma_b * sqrt(variance)



Default spot finding parameters are often not suitable for CCD images

Image is from Rigaku Saturn 92 detector



Default spot finding parameters are often not suitable for CCD images

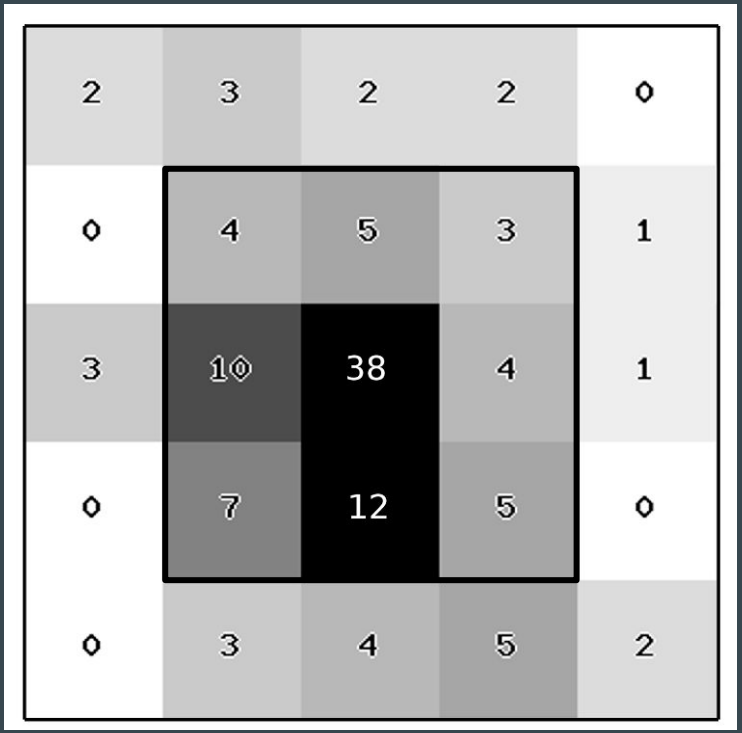
Image is from Rigaku Saturn 92 detector

Summary

- Pilatus and Eiger detectors are statistically well-behaved
- Pixels obey Poisson statistics
- Counts in virtual pixels are under-dispersed relative to a Poisson distribution
- Gain is equal to 1 across the detector, unlike CCDs which can have different per pixel gain values
- Spot finding works very well for Pilatus detectors, even when “strong” spots are “weak”.

Background modelling

Integration



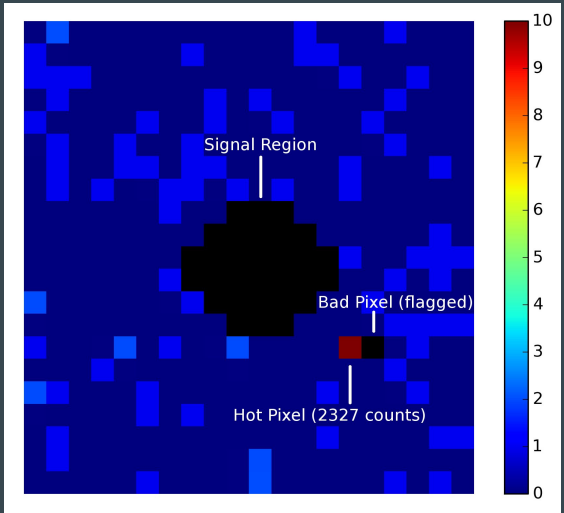
Summation integration: estimate the reflection intensity by summing the counts contributing to the reflection and subtracting the background

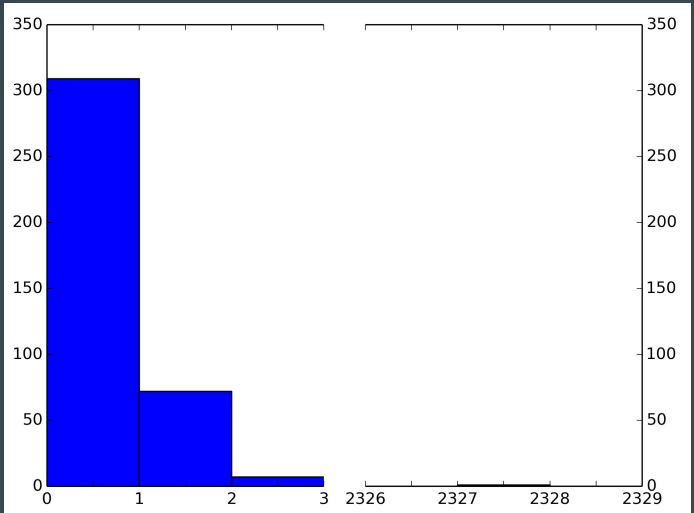
I = SUM(Counts - Background)

Profile fitting: fit a known profile shape to the reflection to estimate the intensity

Need to estimate background under reflection peak since it can't be measured directly

Background outlier pixels



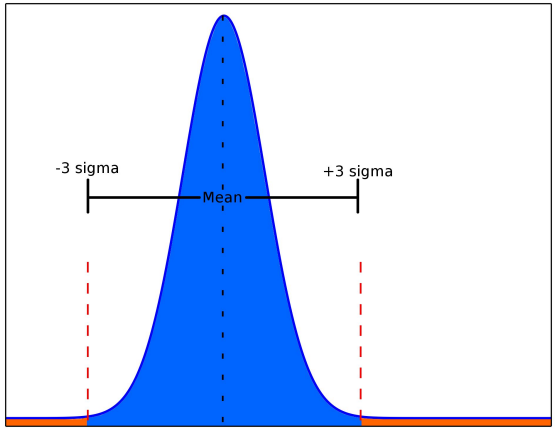


	With Hot Pixel	Without Hot Pixel
Mean	6.20	0.22
Variance/Mean	2237.90	0.926

~1 for Poisson distribution

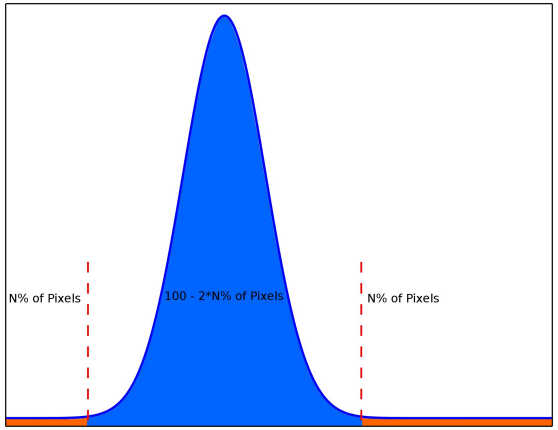
Outlier handling methods: simple

`outlier.algorithm=nsigma`



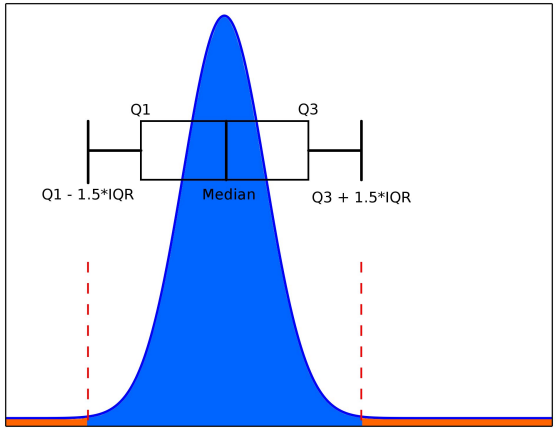
Reject pixels N sigma from the mean

`outlier.algorithm=truncated`



Reject N% of the highest and lowest valued pixels

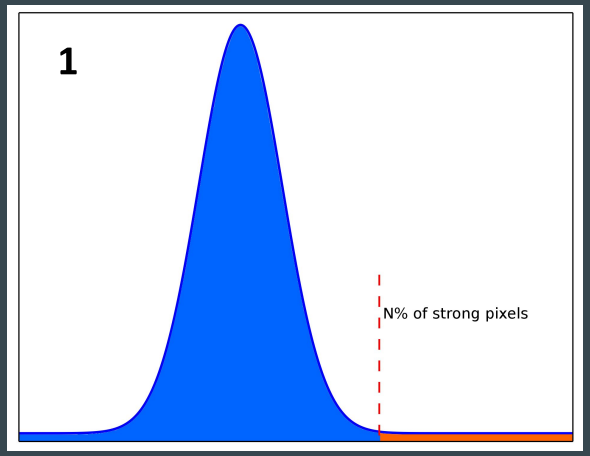
`outlier.algorithm=tukey`



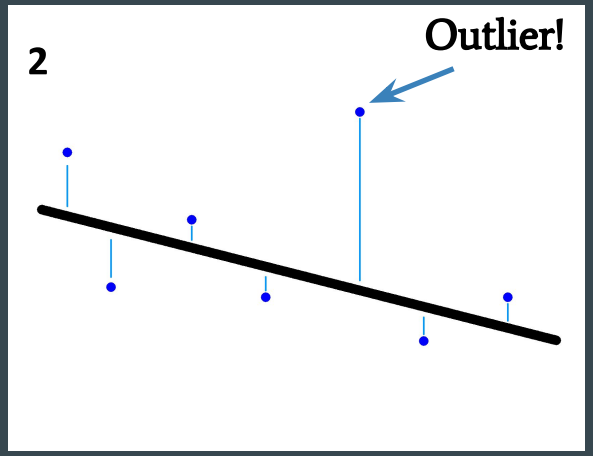
Reject pixels based on the interquartile range

Outlier handling methods: mosflm algorithm

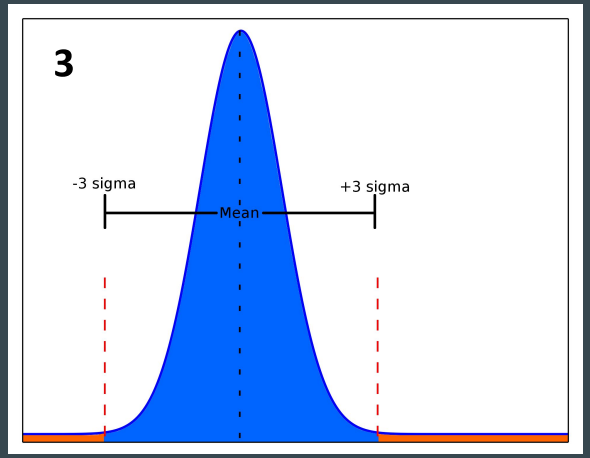
`outlier.algorithm=plane`



Remove N% of strongest pixels and compute the background plane



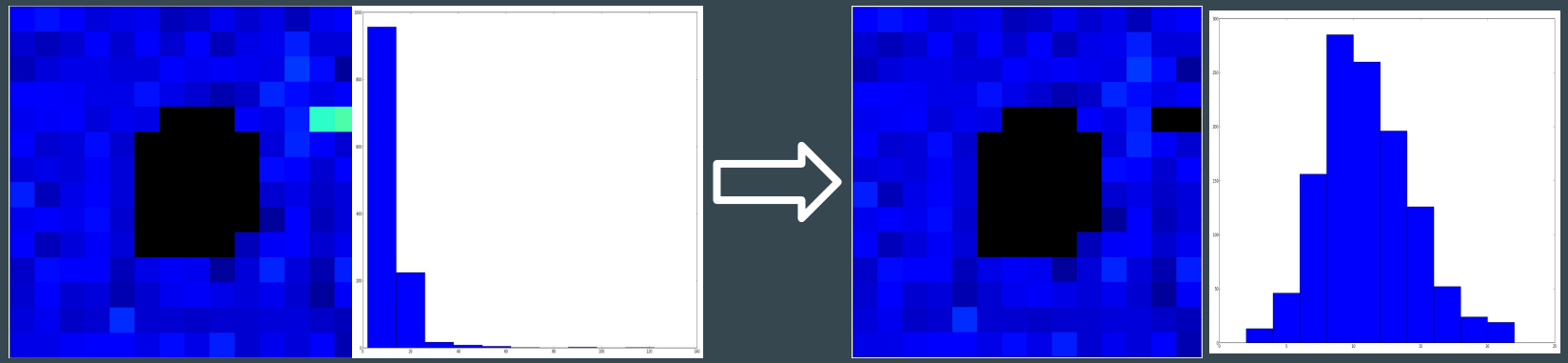
Compute the residuals of all background pixels to the plane



Remove pixels whose residuals are greater than N sigma from the plane

Outlier handling methods: xds algorithm*

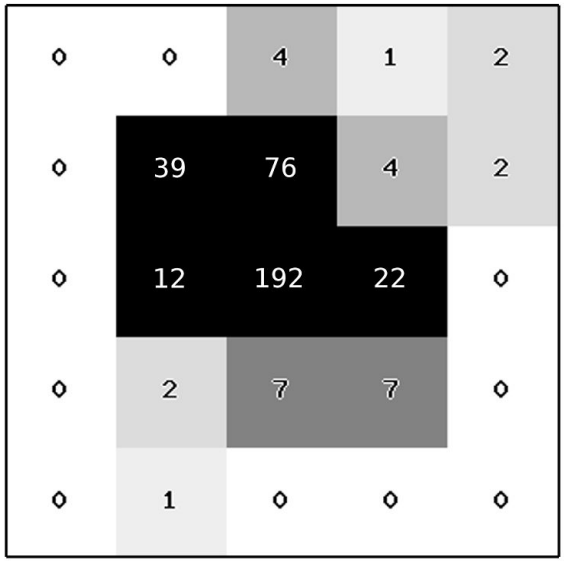
`outlier.algorithm=normal`



Iteratively remove high valued pixels until the distribution of pixel counts resembles a normal distribution

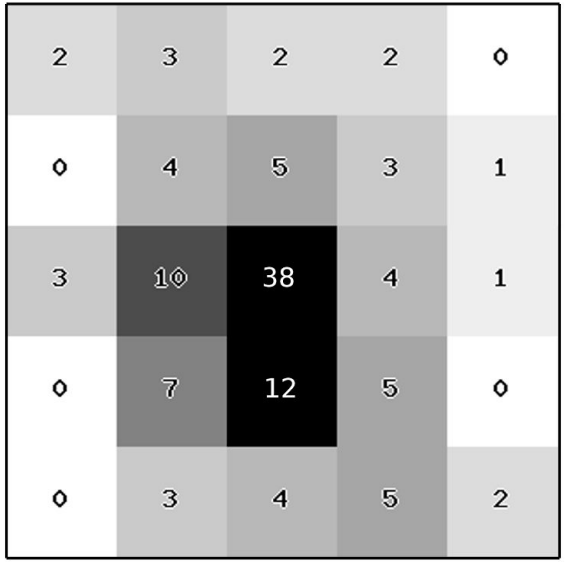
* As described in Kabsch (2010) 'Integration, scaling, space-group assignment and post-refinement', *Acta Cryst. D.* 66(2), 133–44.

Pixel array detectors: low background



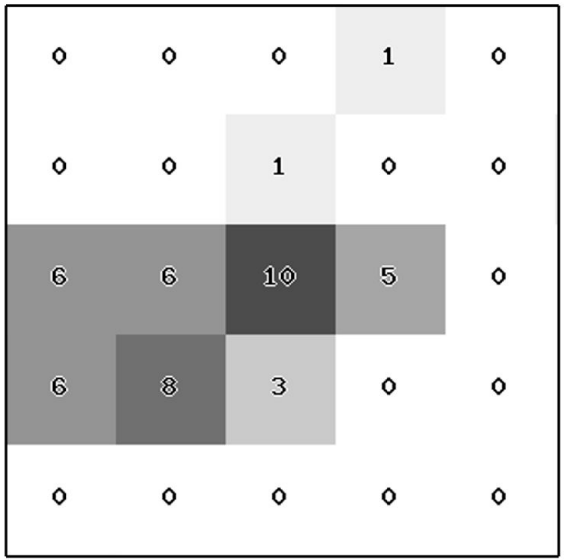
(a)

Thaumatin



(b)

DNA

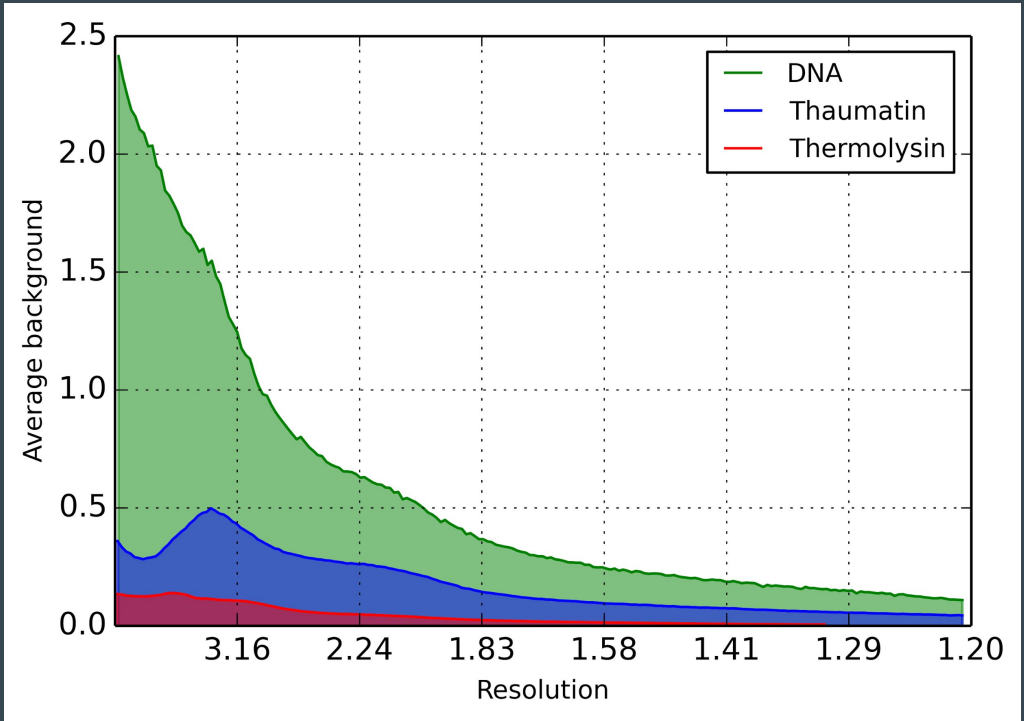


(c)

Thermolysin

Parkhurst et. al (2016) 'Robust background modelling in DIALS', *J. Appl. Cryst.* 49(6), 1912–1921.

Pixel array detectors: low background



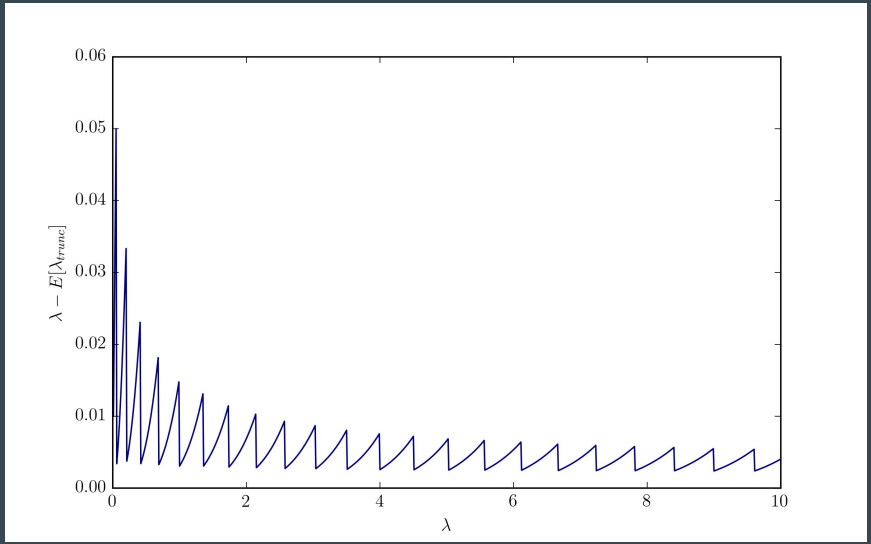
Each dataset has low background over entire resolution range.

Thaumatin and Thermolysin datasets have background less than 1 count per pixel over the whole resolution range

DNA dataset has background less than 1 count per pixel at high resolution

Bias in background determination

- Poisson distribution is asymmetric
- Truncation of the data results in bias in the background determination

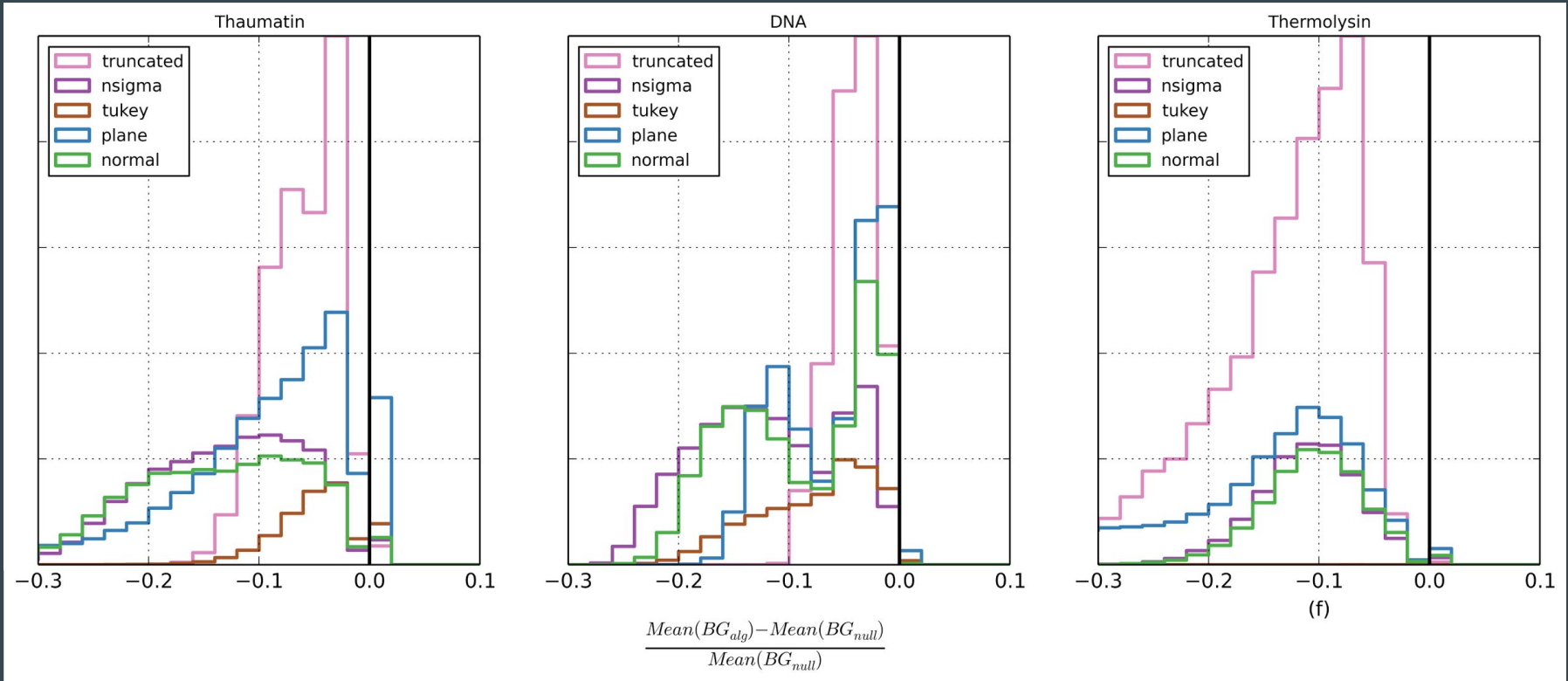


$$\lambda - E[\lambda_{trunc}] = \lambda \left[1 - \frac{Q(b, \lambda) - Q(a-1, \lambda)}{Q(b+1, \lambda) - Q(a, \lambda)} \right]$$

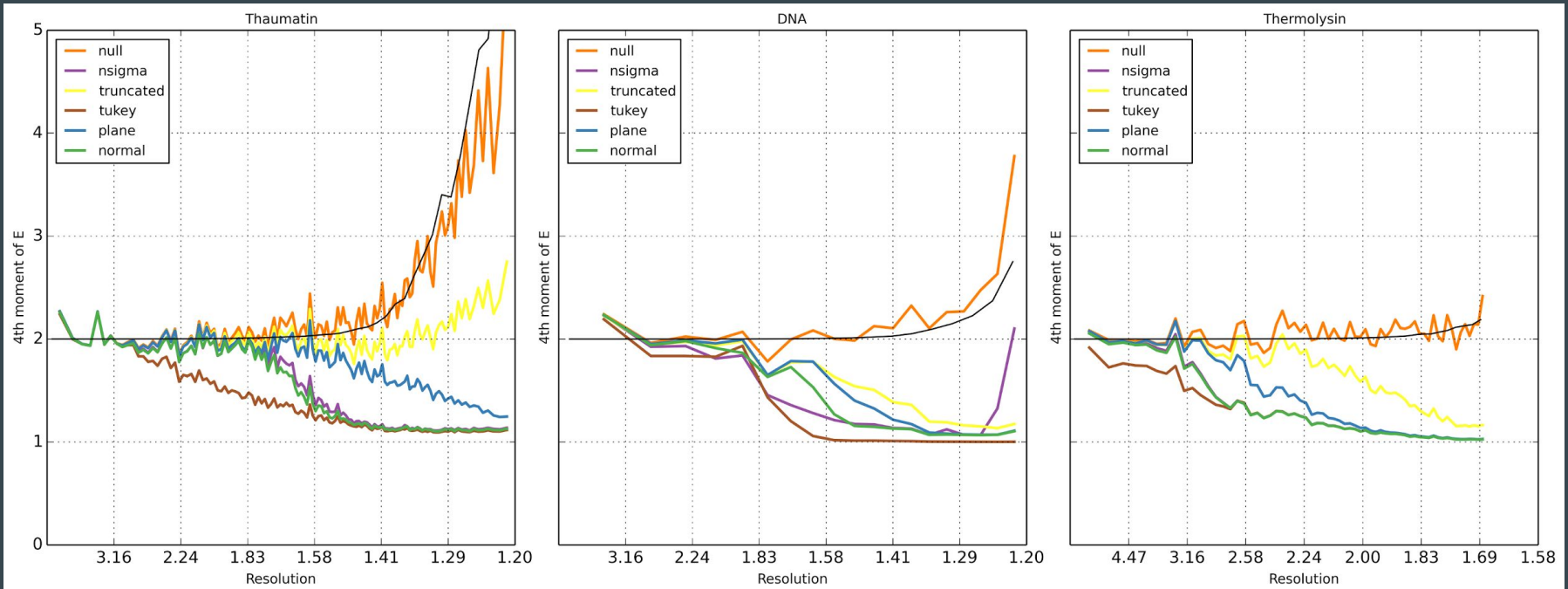
Q is the regularized gamma function

$$Q(a, x) = \frac{\Gamma(a, x)}{\Gamma(x)}$$

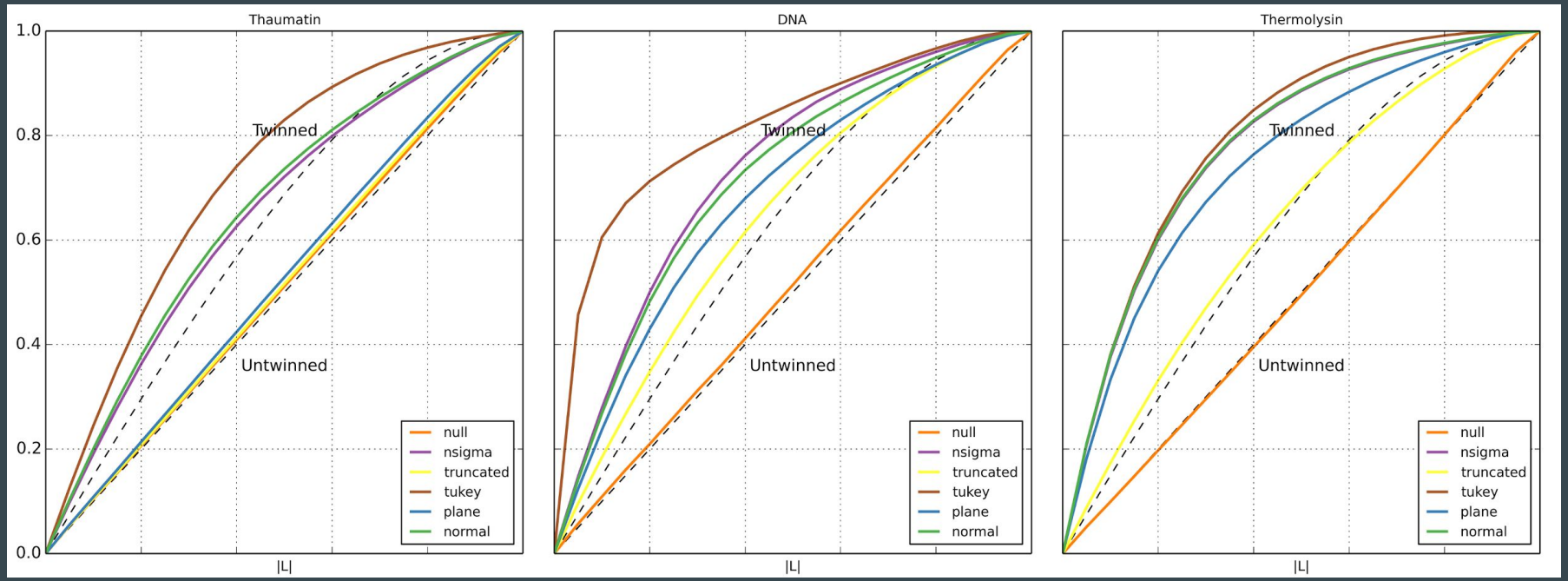
Bias in background determination



Bias in intensity statistics



Bias in intensity statistics



GLM background modelling

Eva Cantoni and Elvezio Ronchetti (2001), "*Robust Inference for Generalized Linear Models*", Journal of the American Statistical Association, Vol. 96, No. 455

Solve $$\sum_{i=1}^{n}\left[\psi_c(r_i)w(x_i)\frac{\mu_i}{\sqrt{V(\mu_i)}} - a(\beta)\right] = 0$$

$$r_i = \frac{y_i - \mu_i}{\sqrt{V(\mu_i)}}$$ Pearson residuals

$$V(\mu) = \mu$$ Variance function

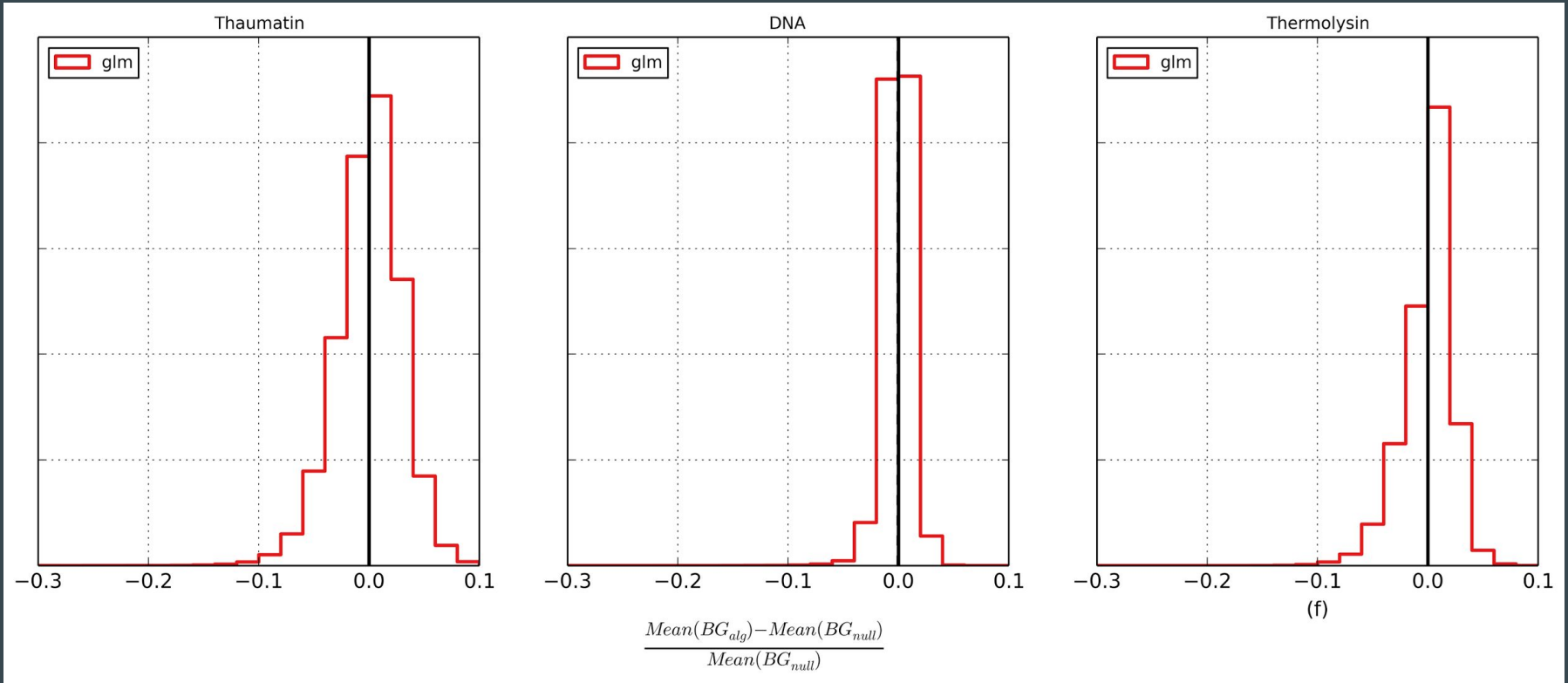
$$w(x) = 1$$ Weights for explanatory variables

$$\psi_c(r_i) = \begin{cases} r_i, & |r_i| \le c \\ c * \operatorname{sgn}(r_i), & |r_i| > c \end{cases}.$$ Weights for dependant variables

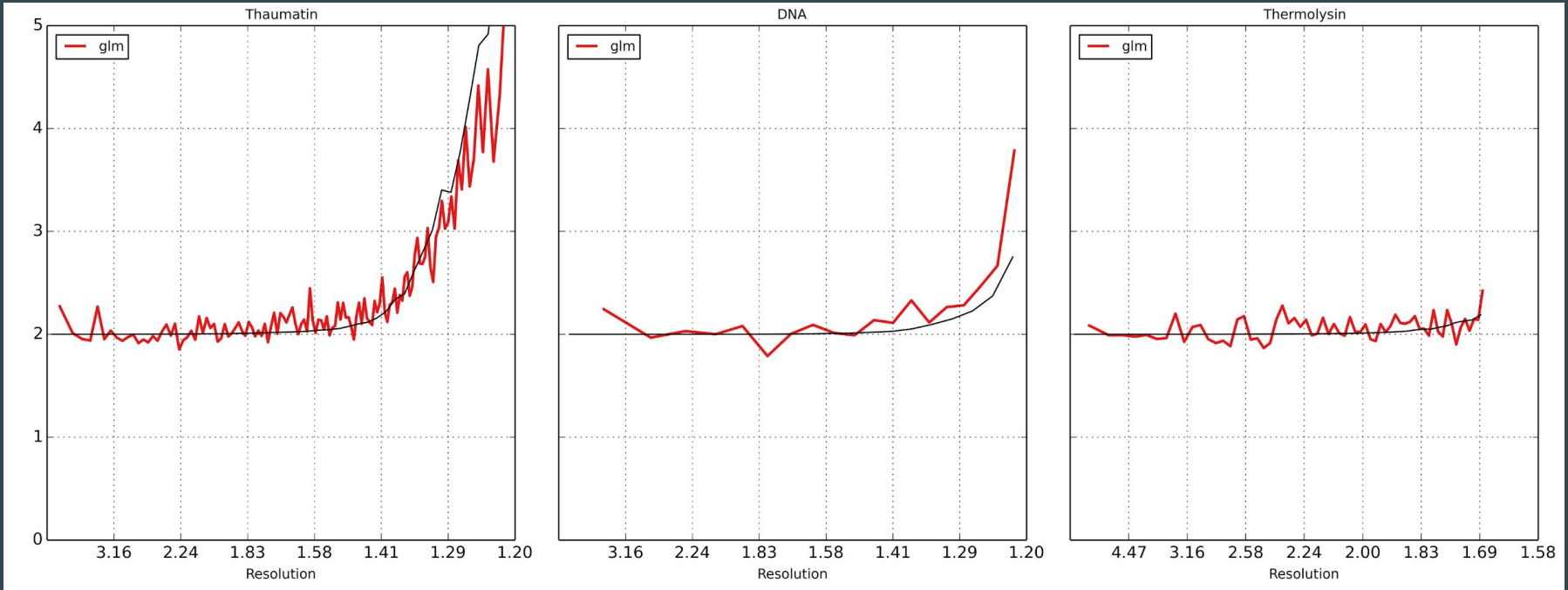
$$c = 1.345$$ Tuning constant

$$a(\beta) = \frac{1}{n}\sum_{i=1}^{n} E[\psi_c(r_i)]w(x_i)\frac{\mu_i}{\sqrt{V(\mu_i)}}$$ Consistency correction

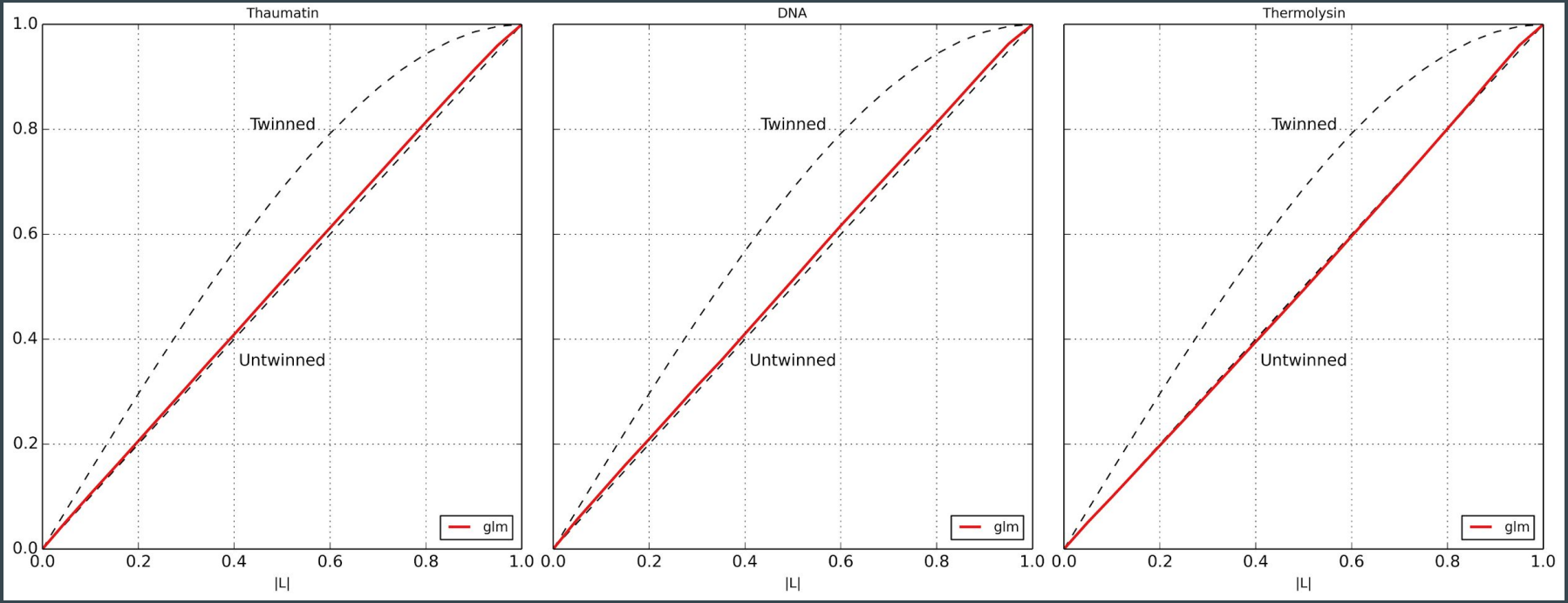
GLM method is unbiased



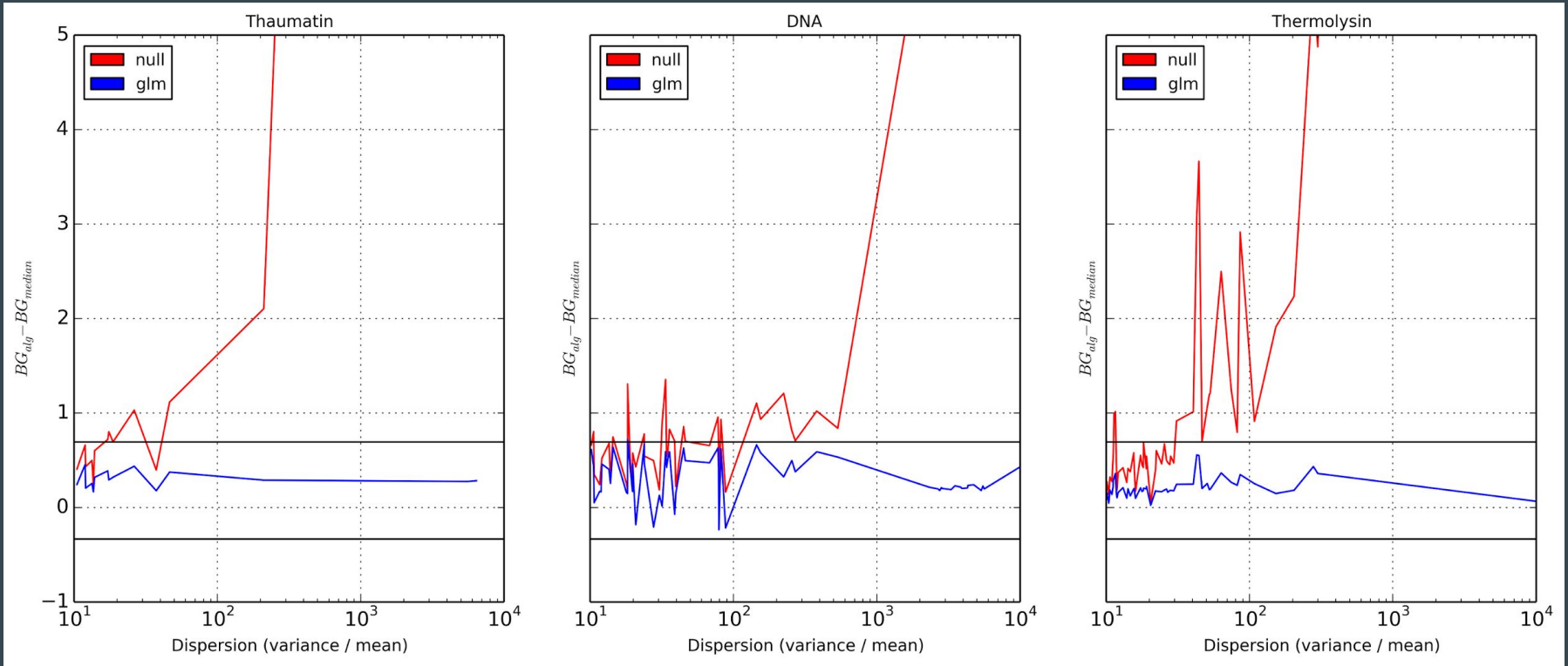
GLM method is unbiased



GLM method is unbiased



GLM method: handling pixel outliers



Twin test results

	Thaumatin		DNA		Thermolysin	
	L test	4th moment	L test	4th moment	L test	4th moment
truncated	0.04	0.00	0.50	0.28	0.50	0.23
nsigma	0.50	0.27	0.50	0.50	0.50	0.50
tukey	0.50	0.50	0.50	0.50	0.50	0.50
plane	0.06	0.01	0.50	0.42	0.50	0.50
normal	0.50	0.30	0.50	0.50	0.50	0.50
glm	0.03	0.00	0.04	0.00	0.03	0.00
null	0.03	0.00	0.05	0.00	0.03	0.00

Summary

- Traditional methods for handling pixel outliers systematically underestimate the background level
- Consequently they overestimate the reflection intensities even in the absence of any pixel outliers in the raw data.
- This can cause statistical tests to give the false impression that a crystal is twinned.
- The GLM method is robust against such effects.
    - When no outliers are present, the estimates given by the GLM algorithm are, on average, the same as those with no outlier handling;
    - When outliers are present, the method gives values within the expected bounds of the median.

Performance

Transitions

CCD → PAD: Pilatus → PAD: Eiger

PILATUS3 S and X product pages ...

EIGER X product pages...

New Algorithms

New infrastructure

Pilatus -> Eiger: algorithms and data

For DIALS:

- Detector behaviour is the same in both cases - identical mathematical problem which is well supported
- One file per image (Pilatus CBF), now one file per scan (Eiger HDF5) - easily handled via dxtbx
- Metadata stored in binary arrays in HDF5 - easily handled via dxtbx
- HDF5 external references just work
- Fine slicing works fine with 3D profile fitting - use the same algorithms.

HDF5 and Nexus

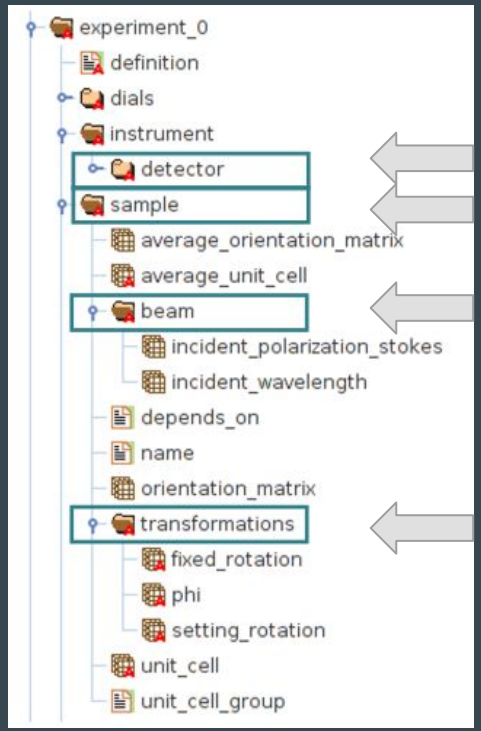
In the past, detectors typically wrote a file for each image. This is ok if the data rate is low and the number of files output is small.

This becomes difficult for the file system to handle when writing out huge numbers of files at a high rate.

The EIGER writes out 1 HDF5 file containing all the images from a single data collection.

EIGER HDF5 files use the Nexus data specification and can be read natively by DIALS.

Nexus HDF5 files



HDF5 is the file container. Image and metadata is stored in a hierarchical format.

Nexus provides the definition that allows programs to understand the HDF5 file.

Full NXmx specification available from:

http://download.nexusformat.org/sphinx/classes/applications/NXmx.html

CCD -> Pilatus -> Eiger: detector performance

- CCD detector - processing data during collection feasible
- Pilatus @ 10 Hz - data set around 3 minutes, fast processing OK, xia2 already “too slow” for interactive feedback
- Pilatus6M @ 10 Hz - may as well wait for data to be finished before processing
- Pilatus6M @ 100 Hz - data set in 18s - time to give up on processing in real time, fast processing now too slow for real-time
- Eiger9M @ 200 Hz - real-time effectively impossible

The problem

- Current detectors (DECTRIS Pilatus) run at up to 100 frames / s
- Next generation Eiger detectors run up to 750 frames / s (for 4M)
- This rate will probably not be routinely used for data collection
- This rate will be used for raster scanning i.e. to allow a large loop to be sampled with a fine beam in a short time (e.g. X-ray centering)
- For raster scanning the experiment has to wait for the results so this is time critical
- Therefore in first instance principle benchmarking problem is spot finding
- Need to make use of parallel processing

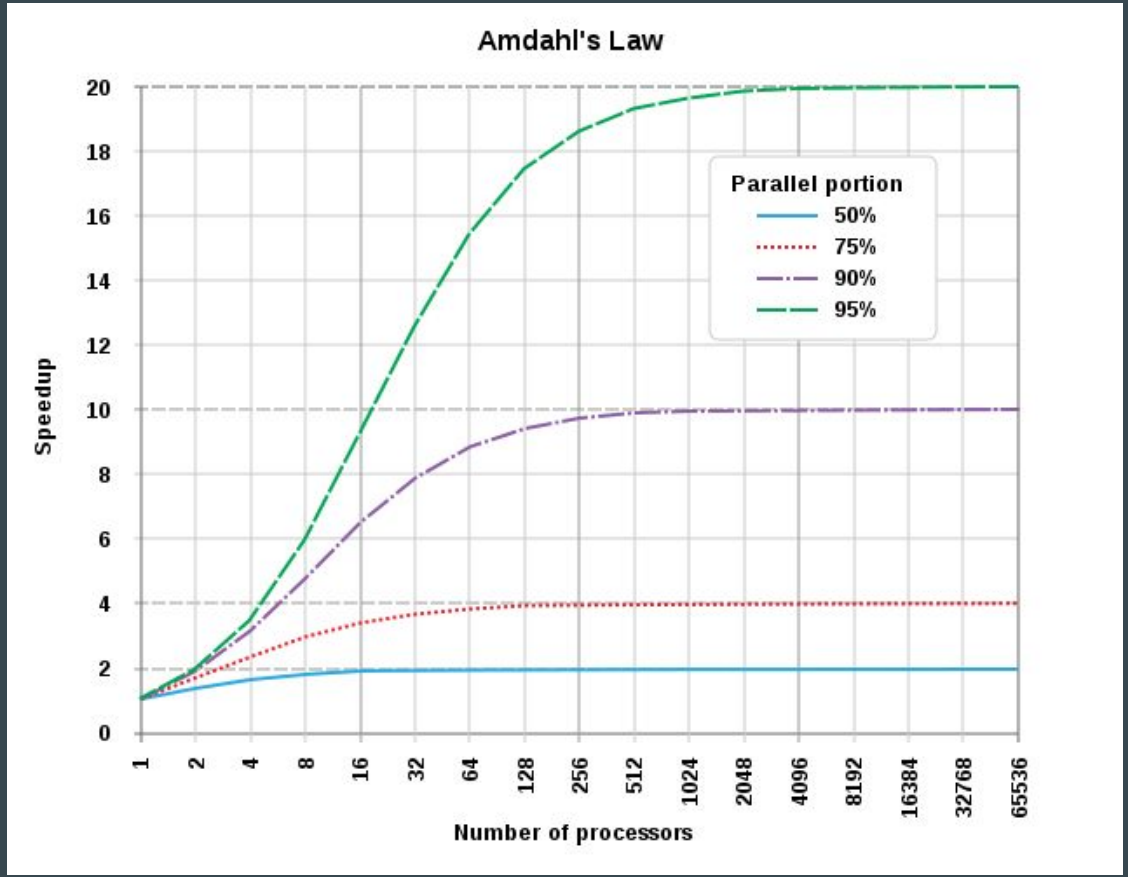
Amdahl's Law

Expected performance improvement from increased number of processors

$$S = \frac{1}{(1-p) + \frac{p}{s}}$$

p = parallel percentage

s = speed up (i.e. number of cores)

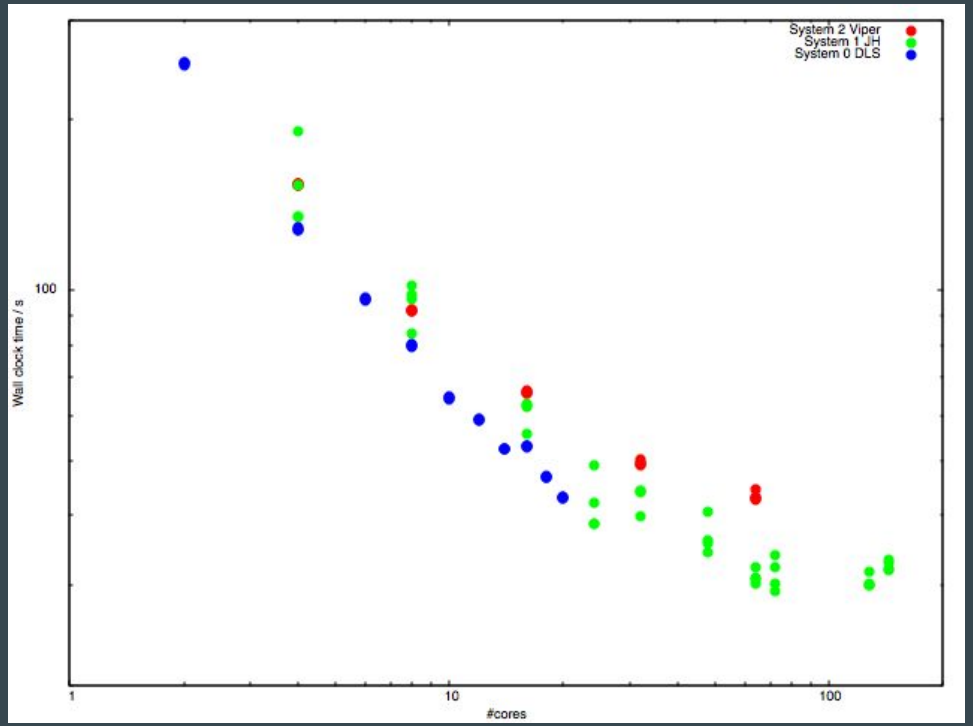


Benchmark

- Performed with dials 1.3.1 linux binaries (same binary set for all systems)
- dials.find_spots datablock.json nproc=${nproc} shoebox=false
- Principle consideration wall clock time i.e. from starting process to results becoming available
- Here nproc=4...# in system
- Data come from RAMDISK => file system performance not a consideration

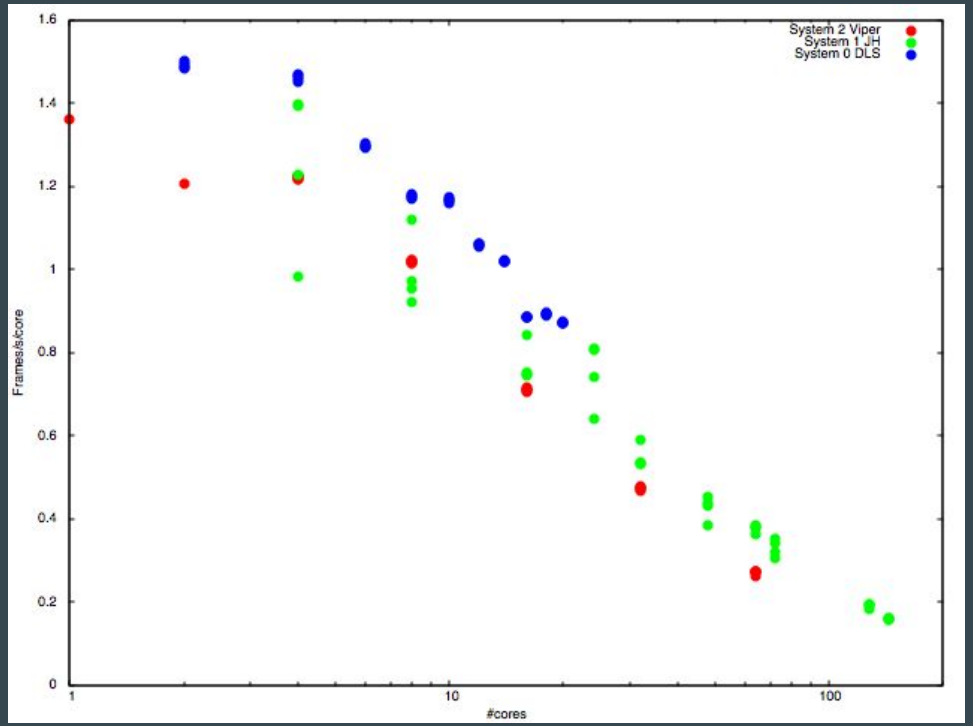
Wall clock time vs #cores

- Wall clock time decreases with increasing number of cores
- Decrease in wall clock time tails off at around 100 cores.



Frames/second/core (“efficiency”) vs #cores

“Efficiency” of spot finding drops with increasing number of cores.



Summary

- “Efficiency” drops off rather quickly with increasing #cores [1]
- Wall clock time flattens off - around 40 s for system 0 using 20 cores; ~ 30 s for system 1 using 144 cores
- For small #frames start up time (~ 4s) dominates
- For large #frames wall clock time ~ linear 0.08 s / frame (10 cores)
- We maybe need to put some effort into optimizing DIALS for many core architectures (e.g. system 1 above; Xeon phi; ...)
- Using small #cores but analysing each row of a grid scan on a separate node in a round-robin manner may be optimum for responsiveness

Acknowledgements

DIALS East

Gwyndaf Evans, Graeme Winter, David Waterman, James Parkhurst, Richard Gildea, Luis Fuentes-Montero, Markus Gerstel, Melanie Vollmar

DIALS West

Nick Sauter, Aaron Brewster, Iris Young

Lots of other people

Garib Murshudov, Andrew Leslie, Phil Evans, Harry Powell, Takanori Nakane, Andrea Thorn

DIALS East– Diamond / CCP4

diamond

BioStructX

Research Complex
at Harwell

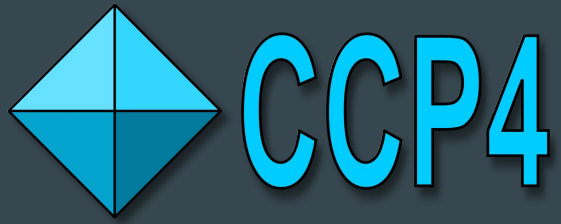
Science & Technology
Facilities Council

DIALS
Diffraction Integration for Advanced Light Sources

wellcome

BERKELEY LAB
Lawrence Berkeley National Laboratory

MRC
Laboratory of
Molecular Biology


CCP4

Thanks for listening!

•••

https://dials.diamond.ac.uk