DIALS: integration

James Parkhurst

DIALS-6 workshop May 2015

DIALS: integration

DATA FLOW

Internals: top-level

Tasks in dials.integate:

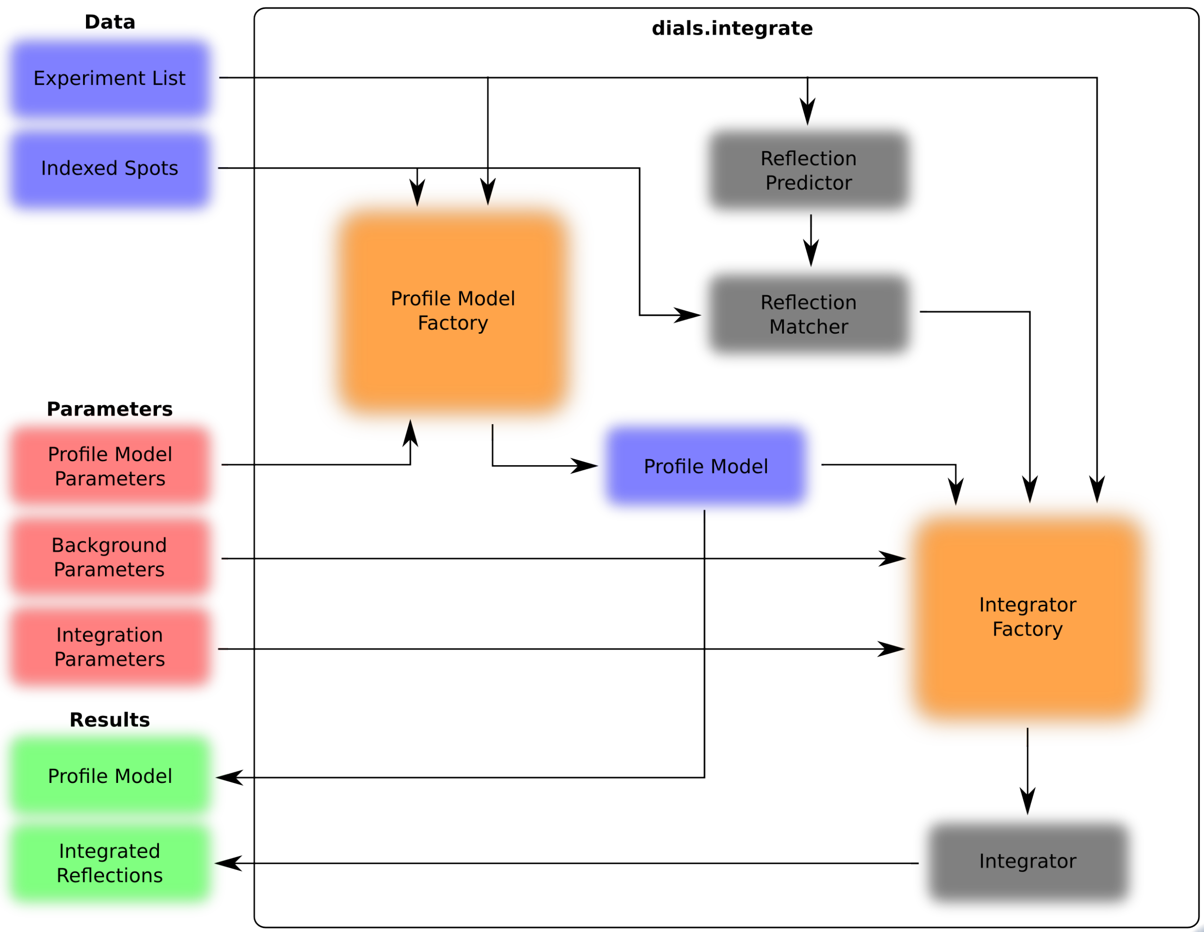
Calculate the bounding box parameters from strong reflections

Predict the positions of reflections on the images

Build reference profiles across all images

Integrate the reflections and save output

Internals: top-level



Internals: types of integrator

3D Integrator
Each integration job reads a block of images and extracts reflections into 3D shoeboxes for processing.

Flattened 3D Integrator
Each integration job reads a block of images and extracts 3D shoeboxes which are "flattened" for processing.

2D Integrator
Each integration job reads a block of images and extracts partial reflections into 2D shoeboxes for processing.

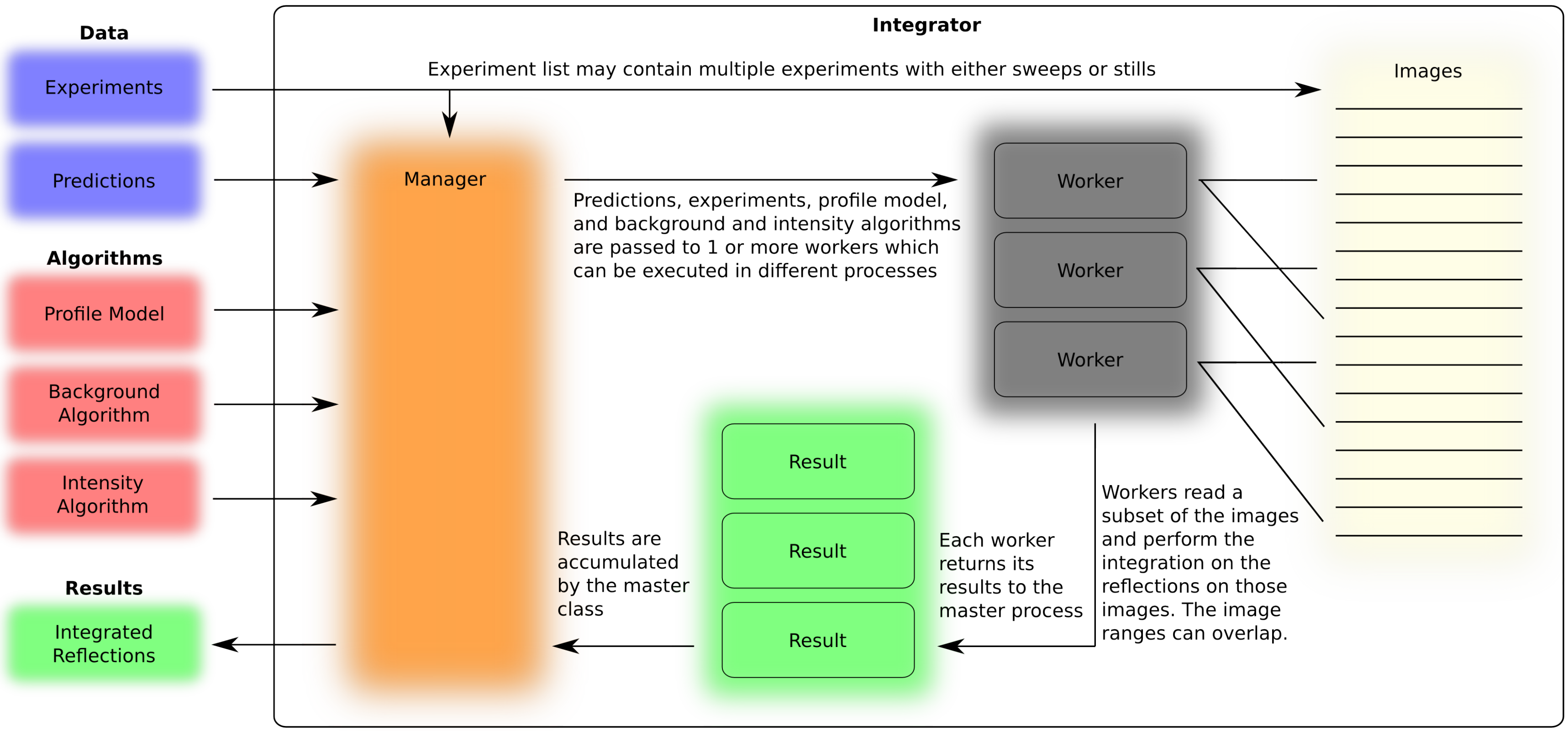
Single Frame 2D Integrator
Each integration job reads a single image and extracts partials reflections into 2D shoeboxes for processing.

Stills Integrator
Same as the single frame 2d integrator but specialized to accept still experiments rather than rotation experiments.

Integrator

DIALS
Diffraction Integration for Advanced Light Sources

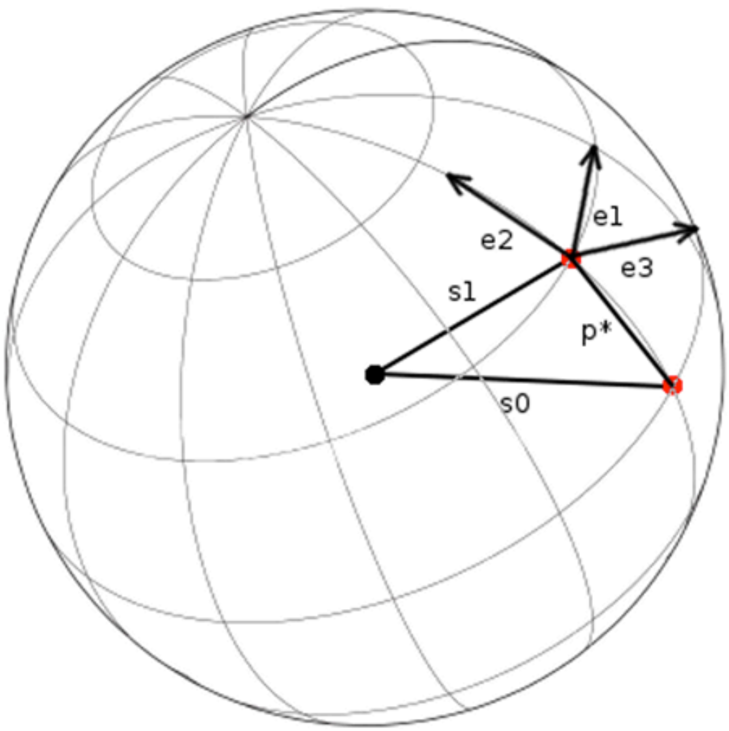
Internals: integrator



DIALS: integration

REFLECTION SHOEBOXES

Computing reflection shoeboxes



Profile coordinate system

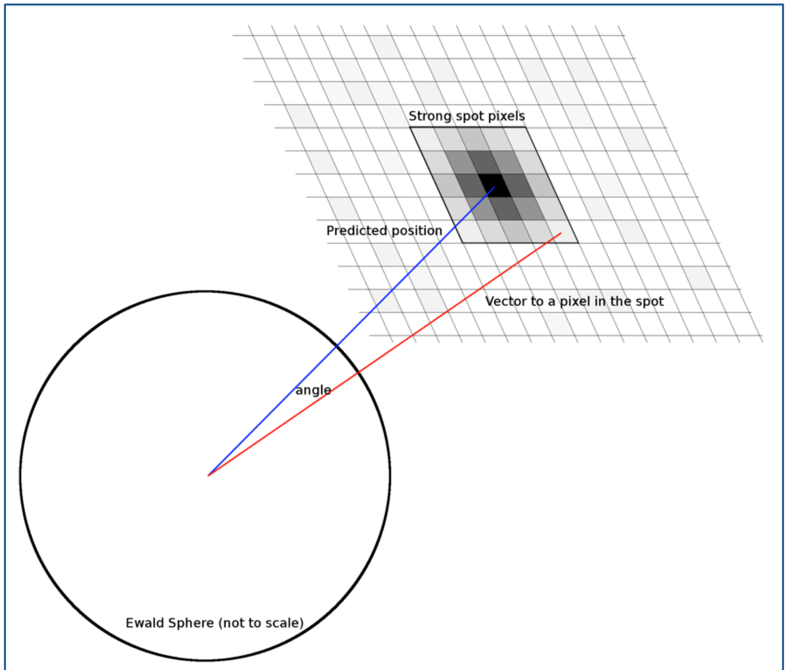
Use the kabsch model of a normal distribution on the surface of the Ewald sphere

$$\exp(-\epsilon_1^2/2\sigma_D^2)\ \exp(-\epsilon_2^2/2\sigma$$

$$\boldsymbol{e}_1 = \boldsymbol{S}_1 \times \boldsymbol{S}_0 / |\boldsymbol{S}_1 \times \boldsymbol{S}_0|$$
$$\boldsymbol{e}_2 = \boldsymbol{S}_1 \times \boldsymbol{e}_1 / |\boldsymbol{S}_1 \times \boldsymbol{e}_1|$$
$$\boldsymbol{e}_3 = (\boldsymbol{S}_1 + \boldsymbol{S}_0) / |\boldsymbol{S}_1 + \boldsymbol{S}_0|$$

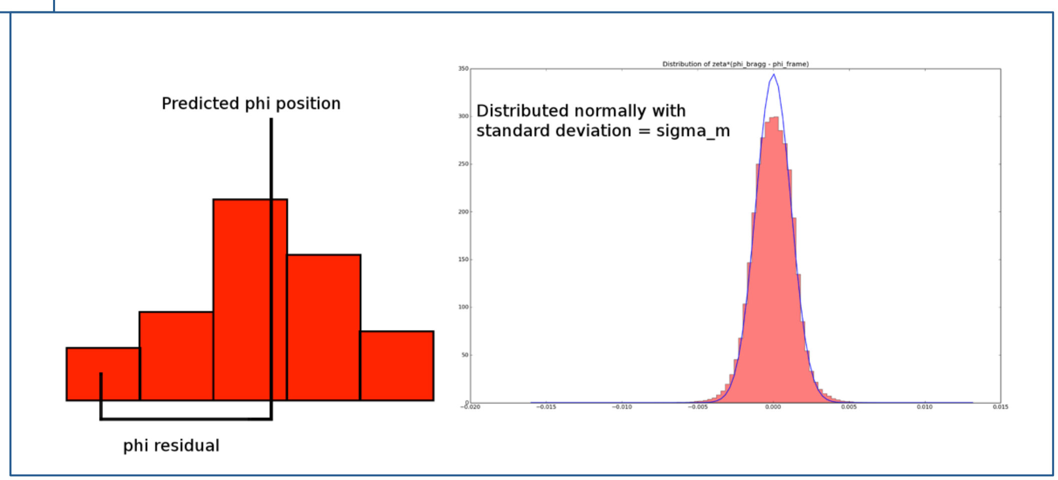
DIALS
Diffraction Integration for Advanced Light Sources

Computing reflection shoeboxes



σ_D is calculated from the spread of angles between the predicted diffracted beam vector and the vector for each strong pixel in the spot

σ_M is calculated by maximum likelihood method assuming a normal distribution of phi residuals for each strong pixel in the spot



DIALS: integration

BACKGROUND MODELLING

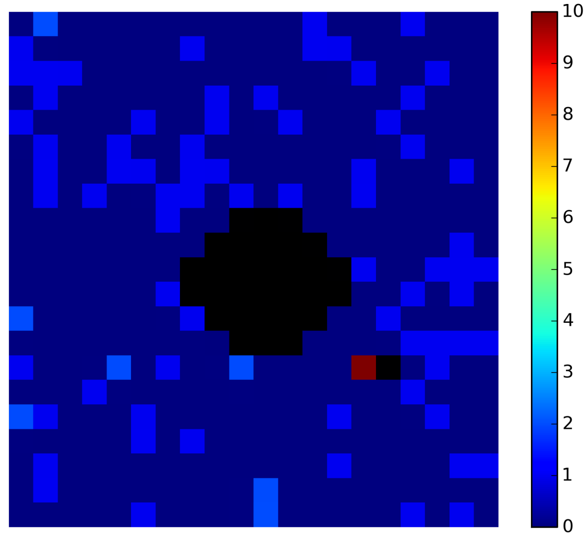
Models

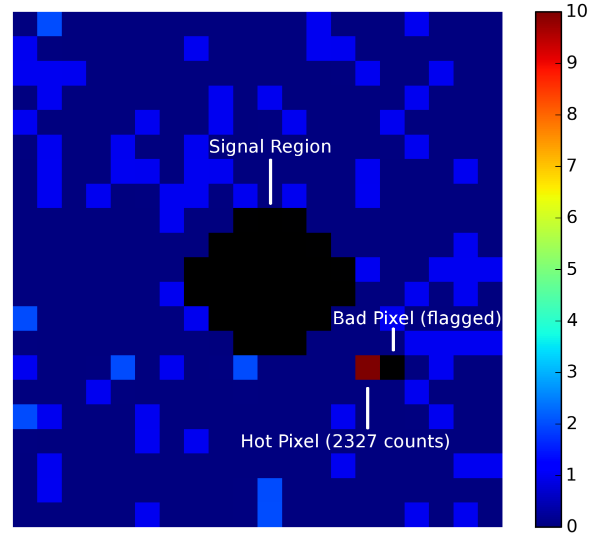
- Options to model the background under the peak as either
  - A constant across each image
  - A constant across all images
  - A plane across each image
  - A hyper-plane across all images
- Computed using simple linear least squares

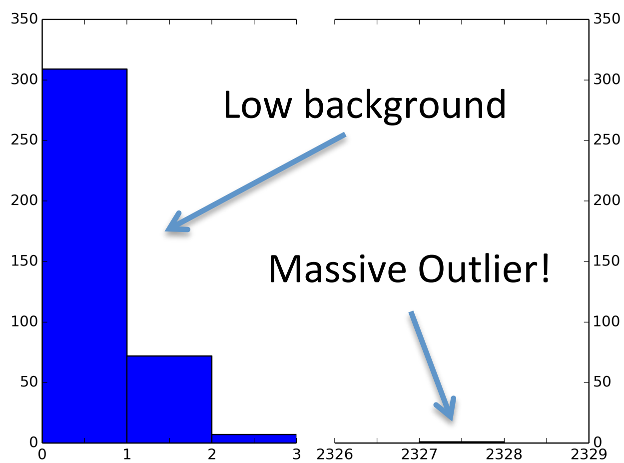
Outliers

- Large valued outliers can cause the background to be over-estimated
- This then causes the reflection intensity to be under-estimated
- Outliers in the background can come from:
  - Intensity from neighbouring spots
  - Hot pixels
  - Zingers
  - Unpredicted reflections
  - Ice rings
  - etc

Outliers







With Hot Pixel

Mean	6.20
Variance/Mean	2237.90

Without Hot Pixel

Mean	0.22
Variance/Mean	0.926

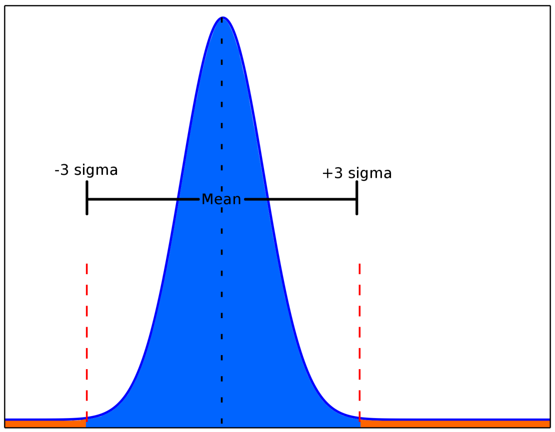
Catastropic over-estimation!

Should be ~1 for poisson distribution

DIALS
Diffraction Integration for Advanced Light Sources

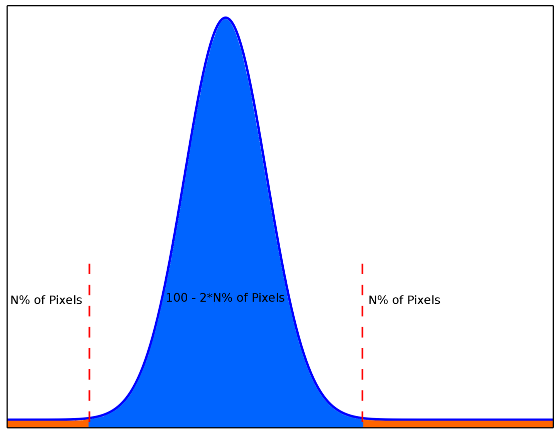
Simple outlier rejection

outlier.algorithm=nsigma



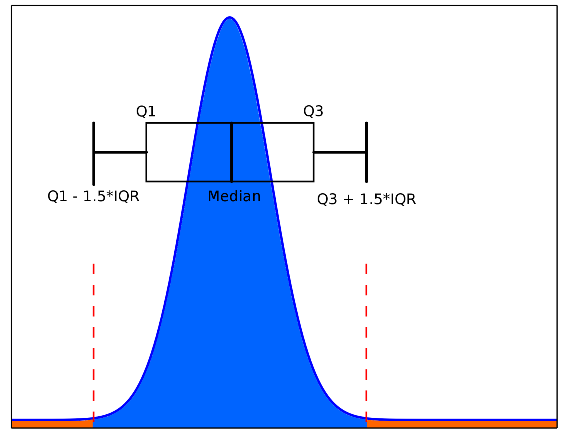
Reject pixels N sigma from the mean

outlier.algorithm=truncated



Reject N% of the highest and lowest valued pixels

outlier.algorithm=tukey

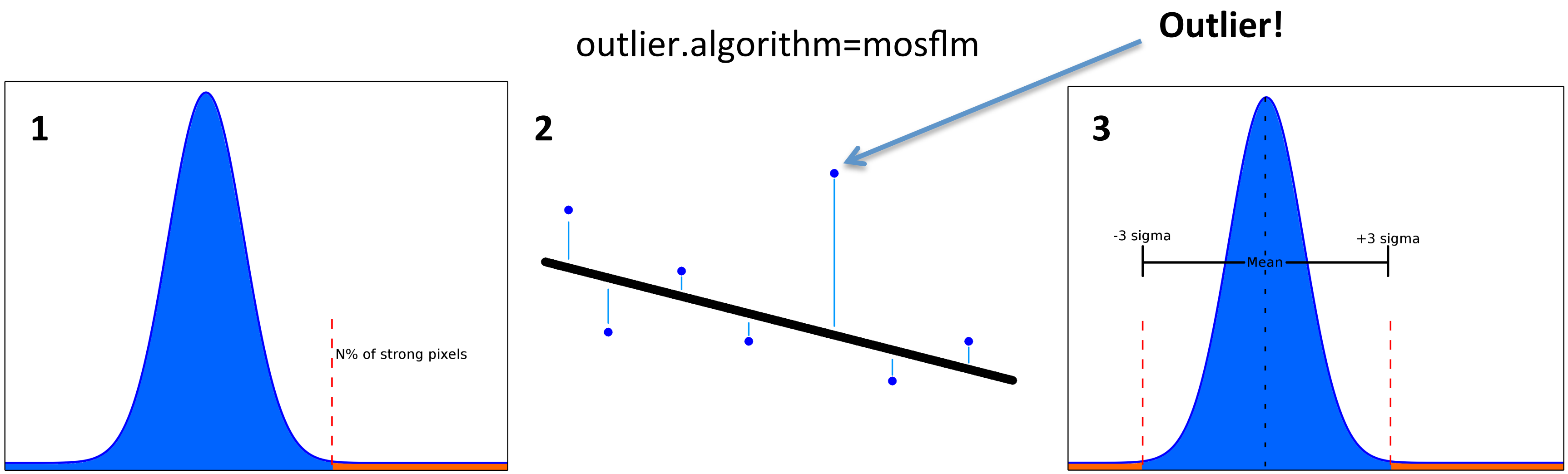


Reject pixels based on the interquartile range

DIALS
Diffraction Integration for Advanced Light Sources

Mosflm-style outlier rejection

outlier.algorithm=mosflm

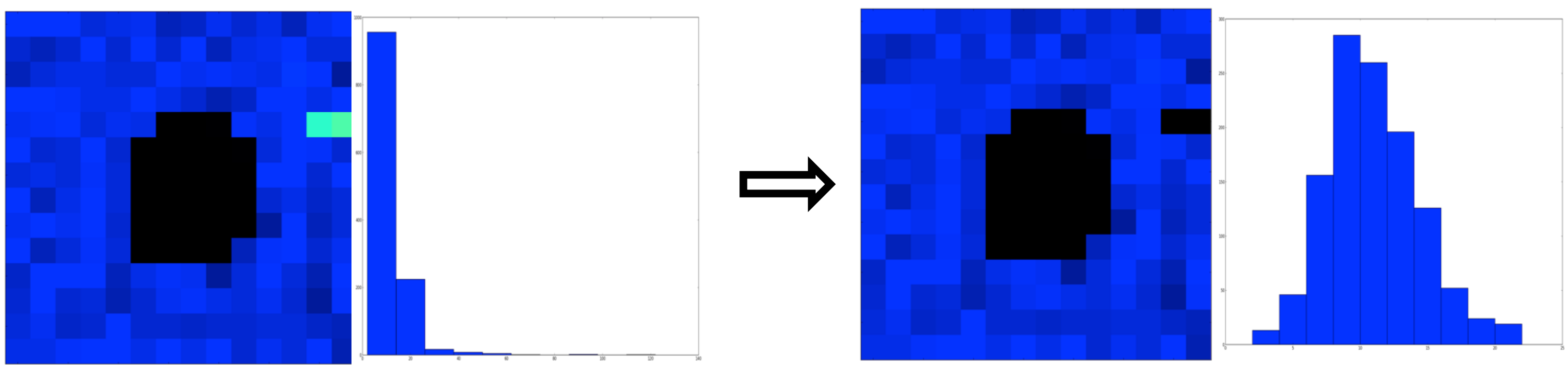


Remove N% of strongest pixels and compute the background plane

Compute the residuals of all background pixels to the plane

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

XDS-style outlier rejection



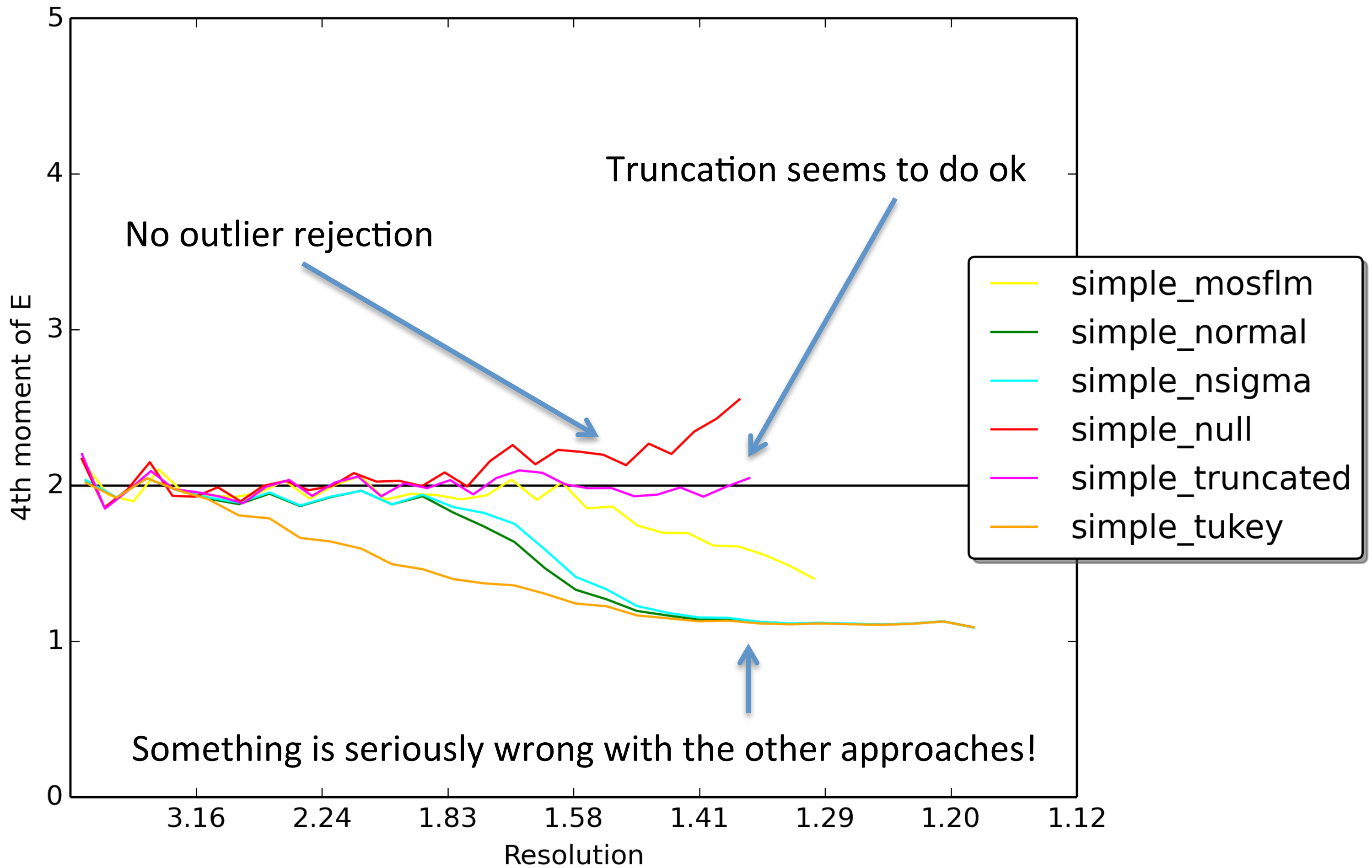
Iteratively remove high valued pixels until the distribution of pixel counts resembles a normal distribution

What effect does outlier rejection have

- Looked at two datasets
  - I04 Bag training. Good data with very few outliers.
  - PNAS data. Good data with some serious outliers.

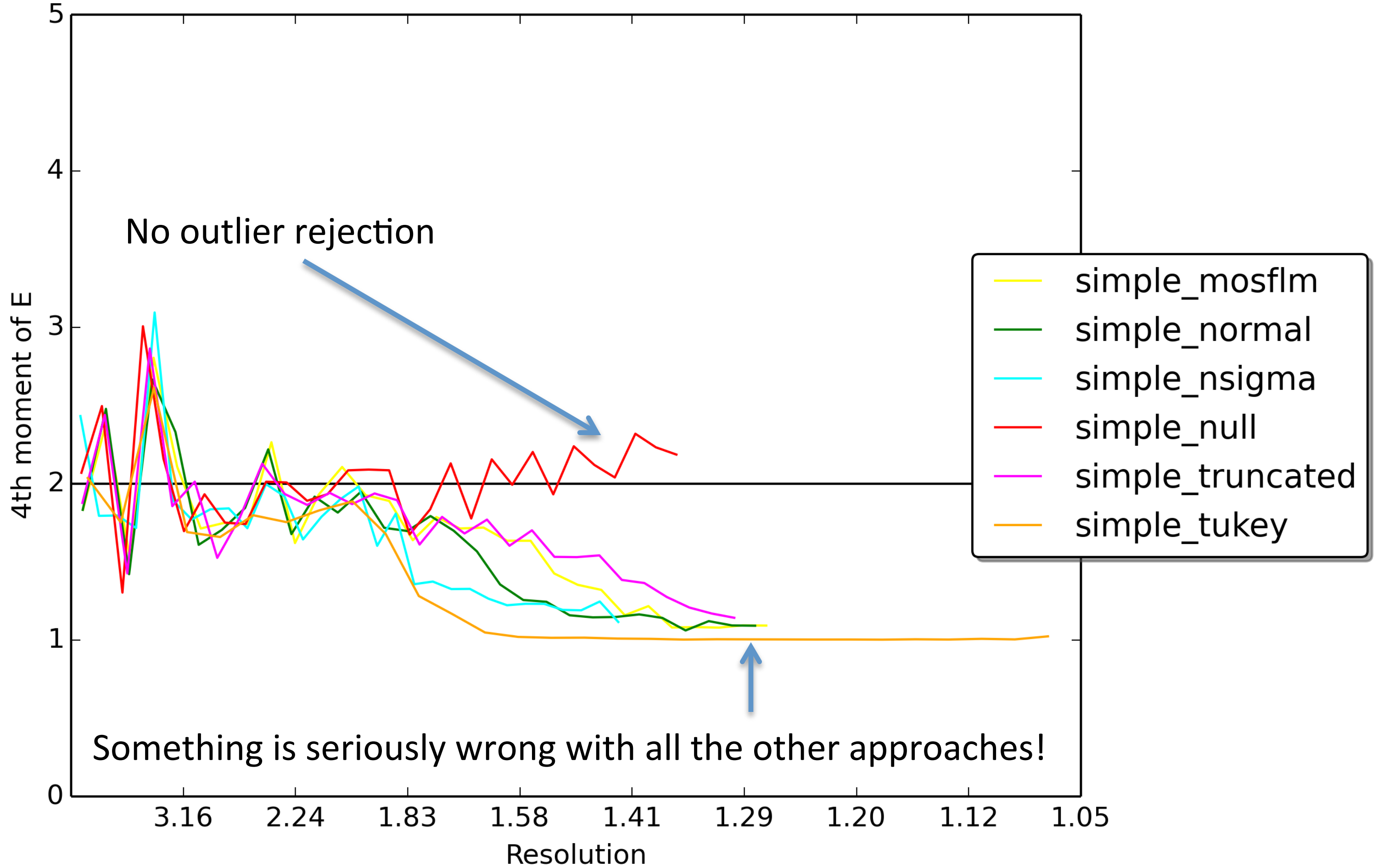
    These outliers caused pointless to find the wrong point group when the data was processed without outlier rejection (pointless has now been fixed so this error no longer occurs).

Handling outliers badly



I04 bag training data processed with xia2

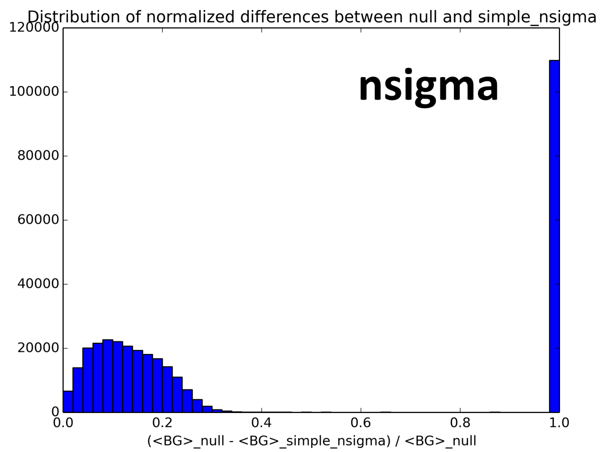
Handling outliers badly

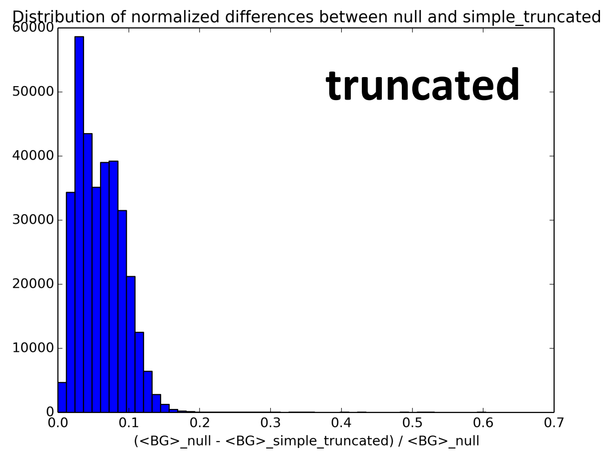


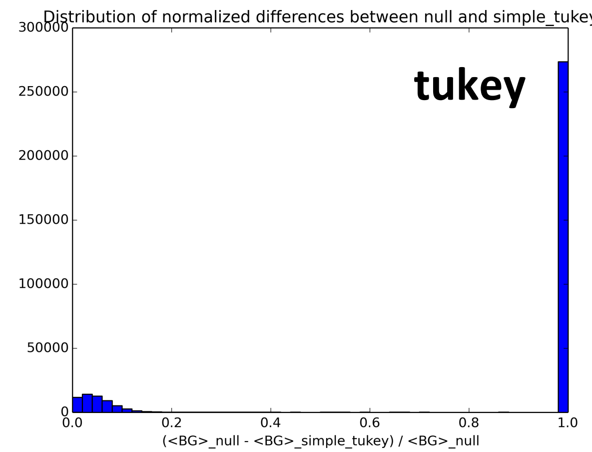
PNAS data processed with xia2

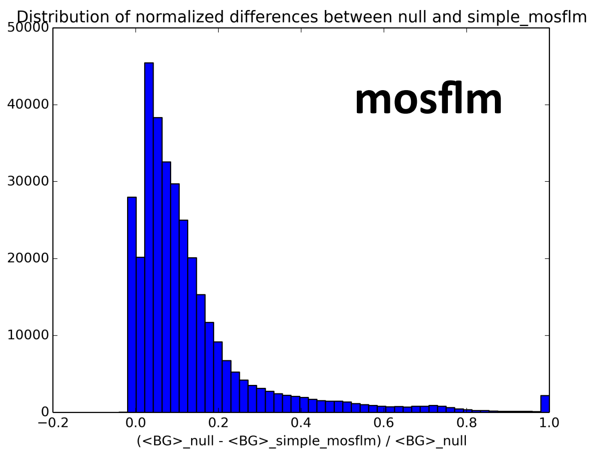
Handling outliers badly

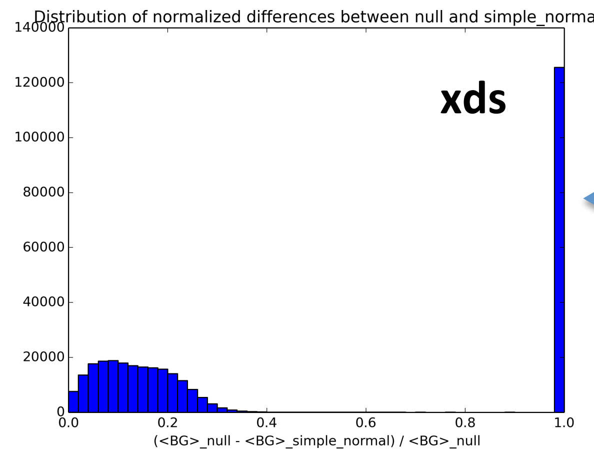
Histograms of background differences vs no outlier rejection for I04 bag training data











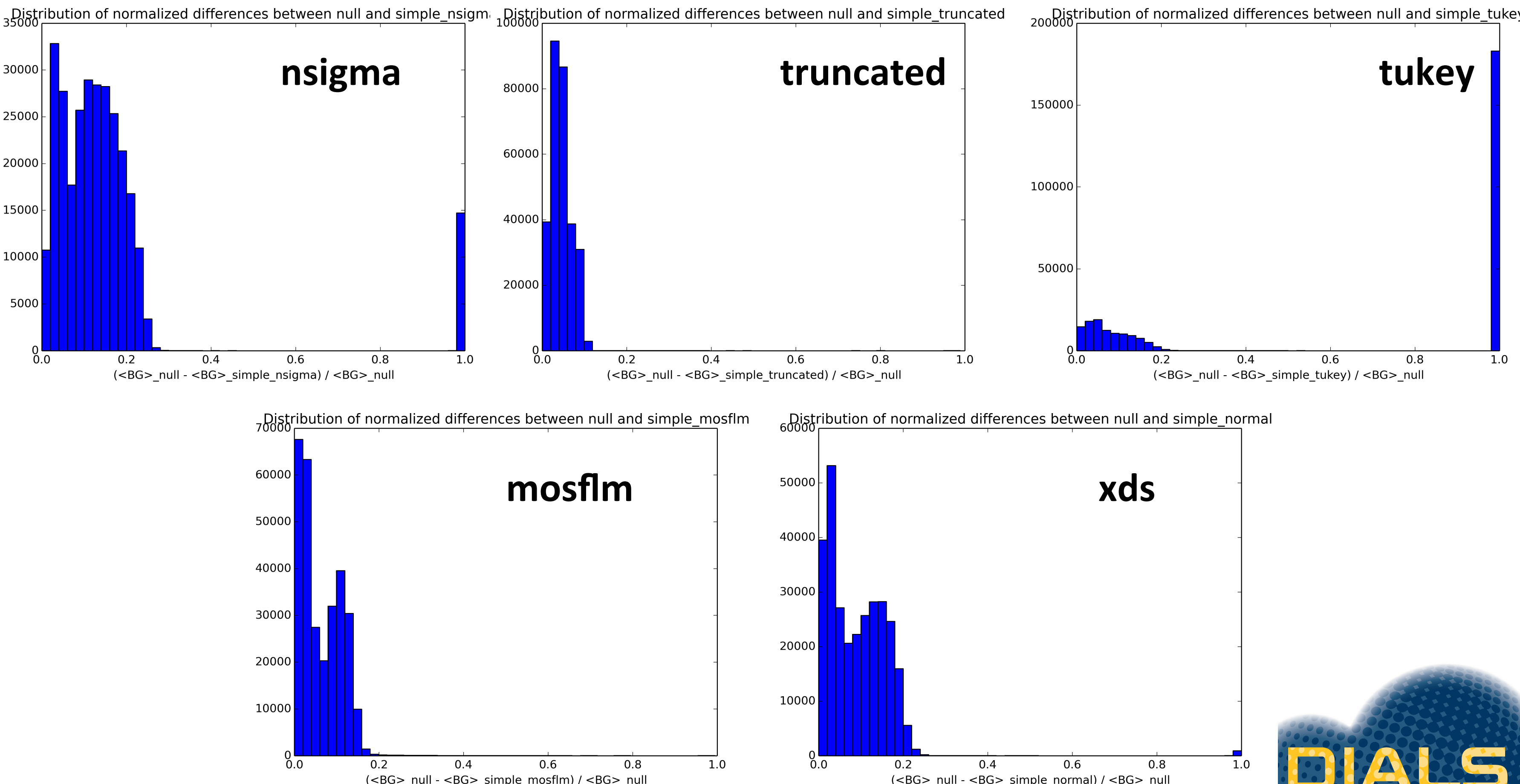
Sometimes removes all non-zero pixels as outliers!

Background is systematically under-estimated

DIALS
Diffraction Integration for Advanced Light Sources

Handling outliers badly

Histograms of background differences vs no outlier rejection for PNAS data



Background is systematically under-estimated

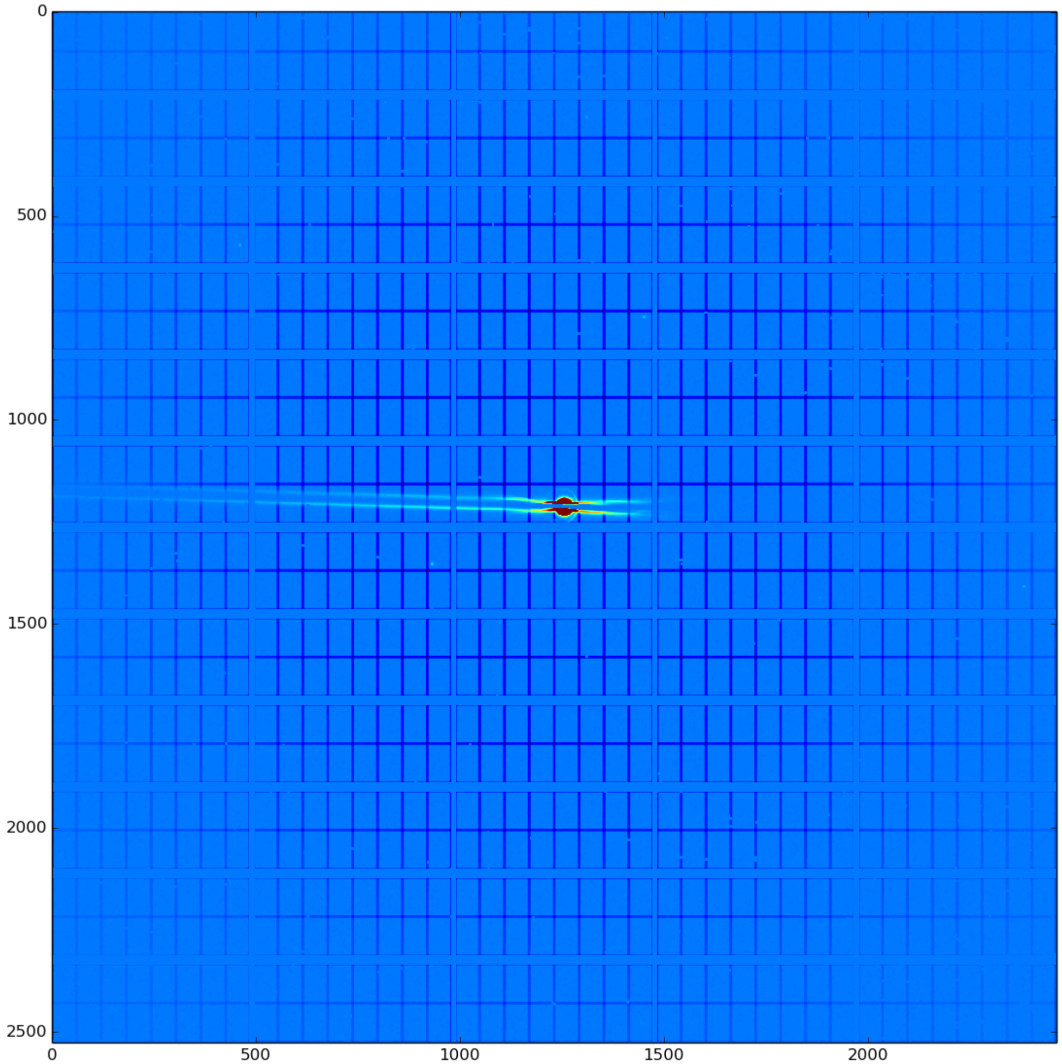
DIALS
Diffraction Integration for Advanced Light Sources

Handling outliers better

- Most of our methods assume a normal distribution of counts – not a good approximation for data with low background
- Instead of rejecting outliers could we use a robust estimation method?
- Most methods also focus on normally distributed data – leads to under-estimation
- Could we use a robust generalized linear model approach assuming a Poisson distribution?

DIALS
Diffraction Integration for Advanced Light Sources

Is the background Poisson distributed?



- Analysed 9000 blank images
- Local index of dispersion computed at each pixel
- Average index of dispersion computed for each pixel

Background data is Poisson distributed

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

~7.2% of pixels are affected

DIALS
Diffraction Integration for Advanced Light Sources

Robust GLM algorithm

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

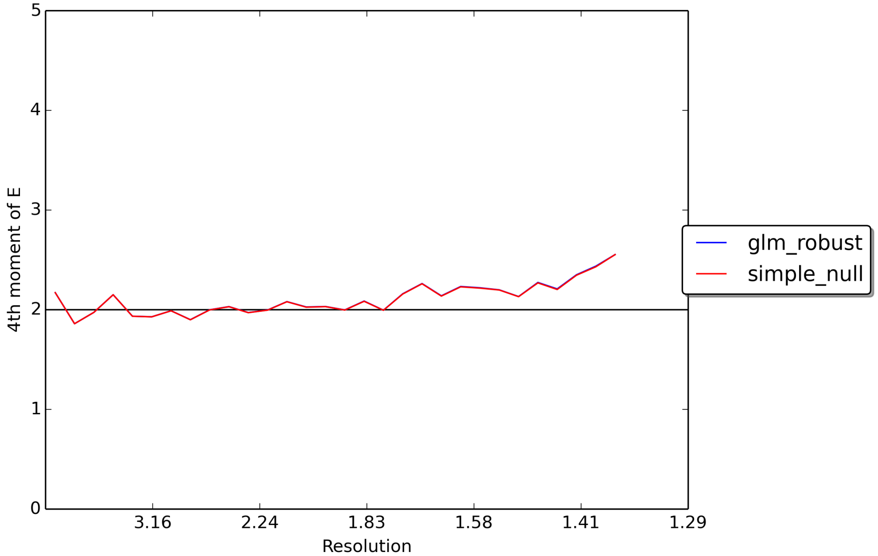
$\psi_c(r)=\begin{cases} r, & |r|\le c \\ c\,\mathrm{sign}(r), & |r|>c \end{cases}$ Weights for dependant variables

$c=1.345$ Tuning constant

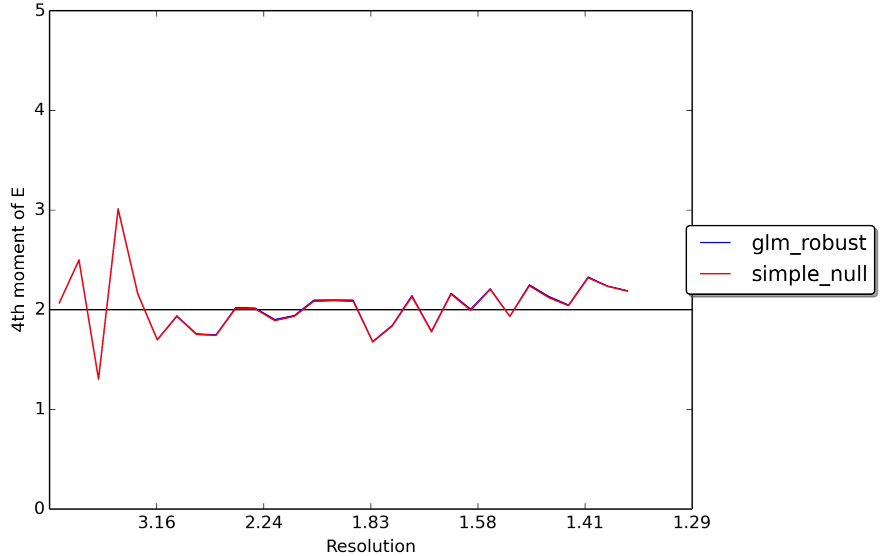
$a(\beta)=\frac{1}{n}\sum_{i=1}^{n}E[\psi_c(r_i)]\,w(x_i)\,\frac{\mu_i'}{\sqrt{V(\mu_i)}}$ Consistency correction

Handling outliers better (?)

In both cases robust GLM method gives sensible results for the 4th moment of E plots



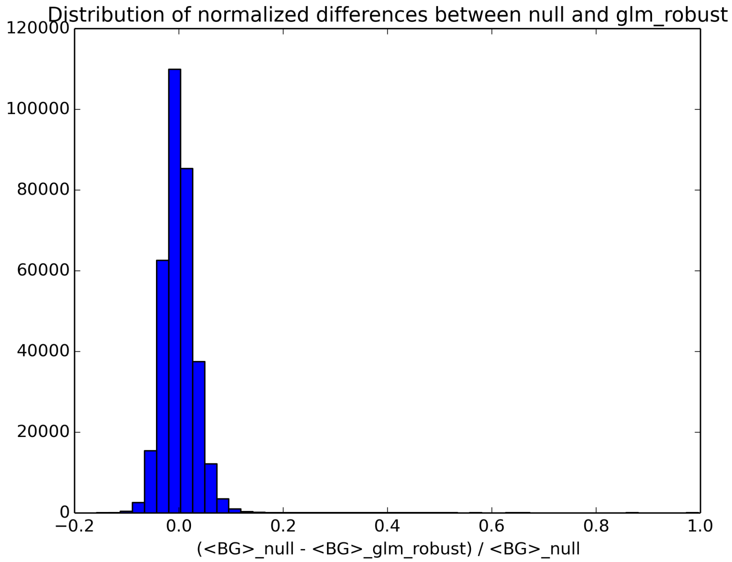
I04 bag training data



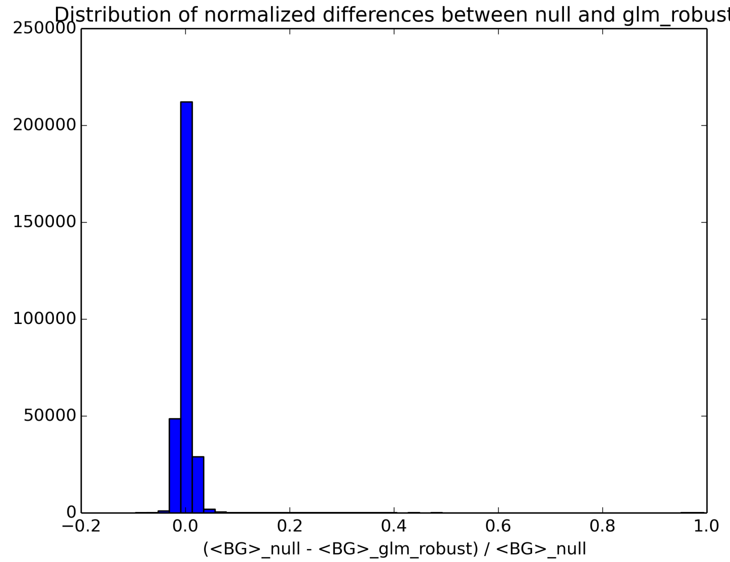
PNAS data

Handling outliers better (?)

Histograms of background differences vs no outlier rejection for I04 bag training data

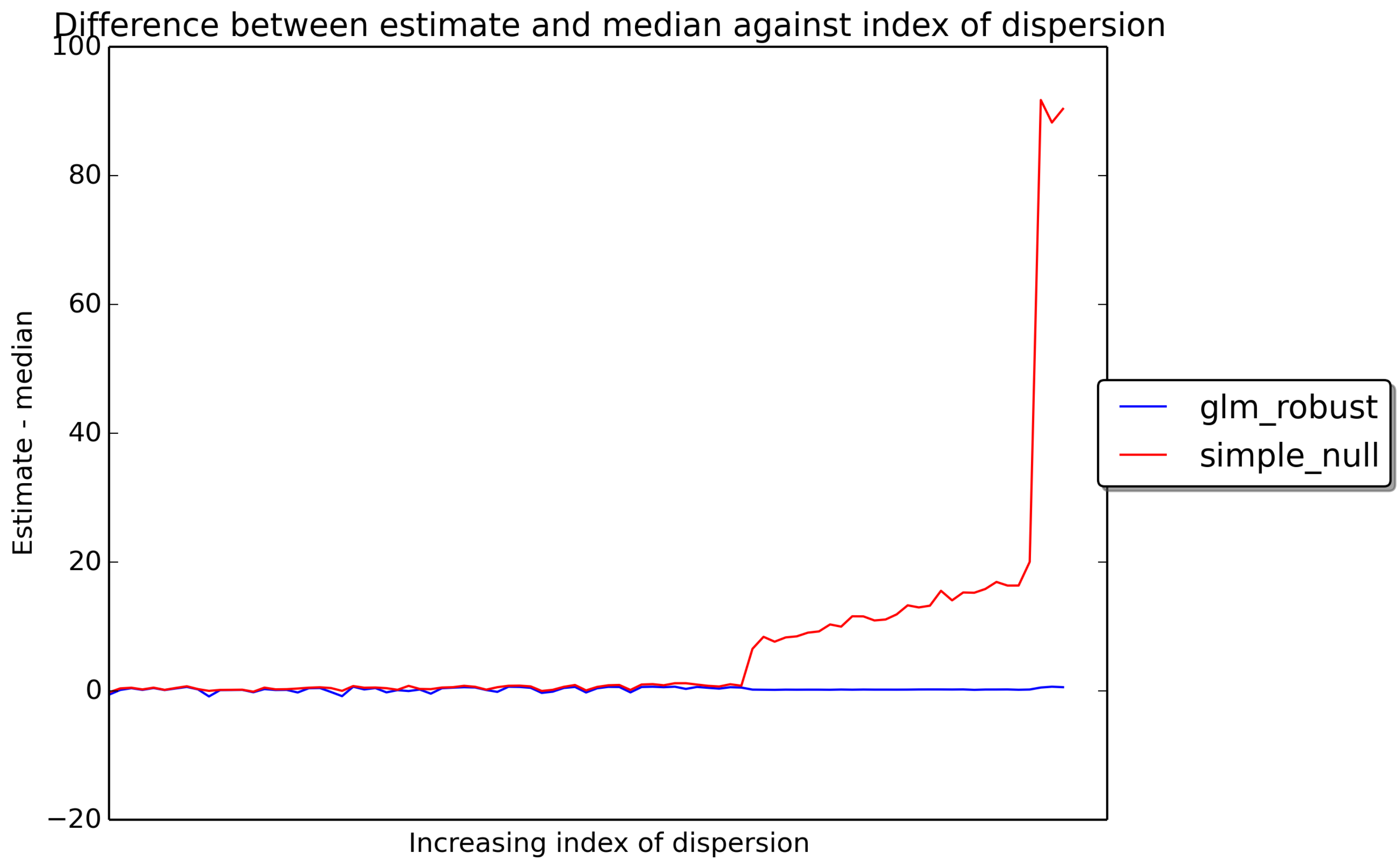


Histograms of background differences vs no outlier rejection for PNAS data



No systematic difference in the background!

Are we actually doing anything?



Looks like we're handling outliers ok

Robust algorithm

- Algorithm requires a good seed value or it won't converge – use the median
- Can't represent a straight line so currently using constant background – plan to look at more sophisticated background if needed
- Current results look promising

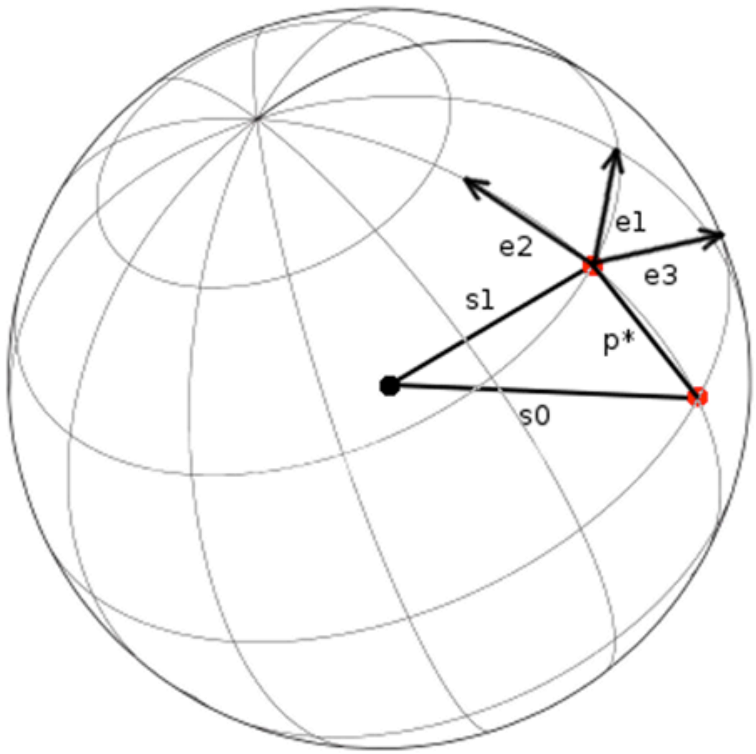
DIALS: integration

SIGNAL INTEGRATION

Integration

- Integration algorithm options:
  - Summation
  - 3D profile fitting (as in XDS)
  - 2D profile fitting (future)

3D profile fitting coordinate system



Profile coordinate system

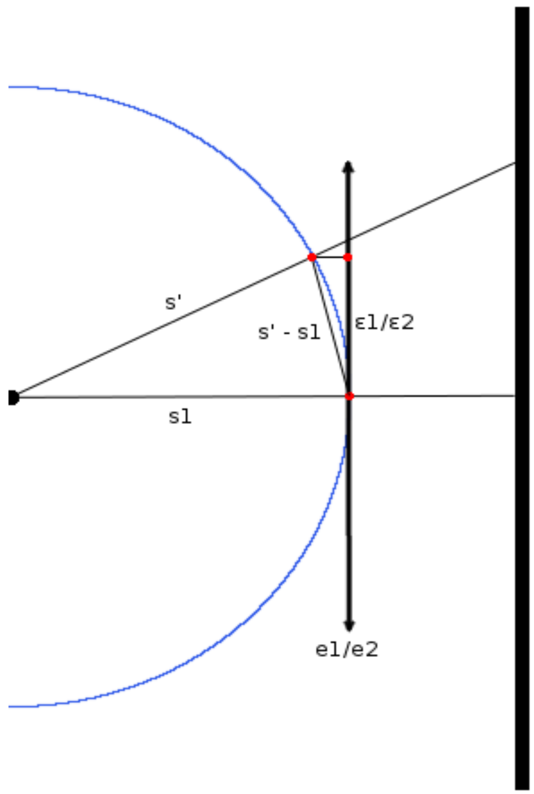
Use Kabsch coordinate system

- Corrects for geometrical distortions
- Makes spots appear to have taken shortest path through Ewald sphere
- Model assumes a Gaussian profile in Kabsch coordinate system

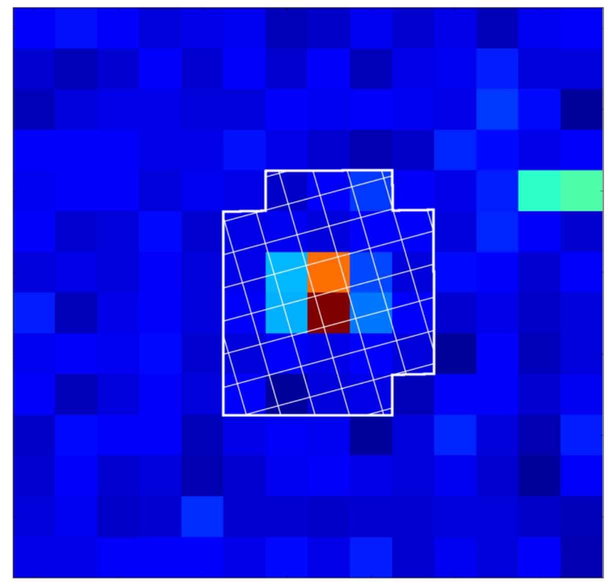
$$\boldsymbol{e}_1 = \boldsymbol{S}_1 \times \boldsymbol{S}_0 / |\boldsymbol{S}_1 \times \boldsymbol{S}_0|$$
$$\boldsymbol{e}_2 = \boldsymbol{S}_1 \times \boldsymbol{e}_1 / |\boldsymbol{S}_1 \times \boldsymbol{e}_1|$$
$$\boldsymbol{e}_3 = (\boldsymbol{S}_1 + \boldsymbol{S}_0) / |\boldsymbol{S}_1 + \boldsymbol{S}_0|$$

DIALS
Diffraction Integration for Advanced Light Sources

3D profile fitting pixel gridding

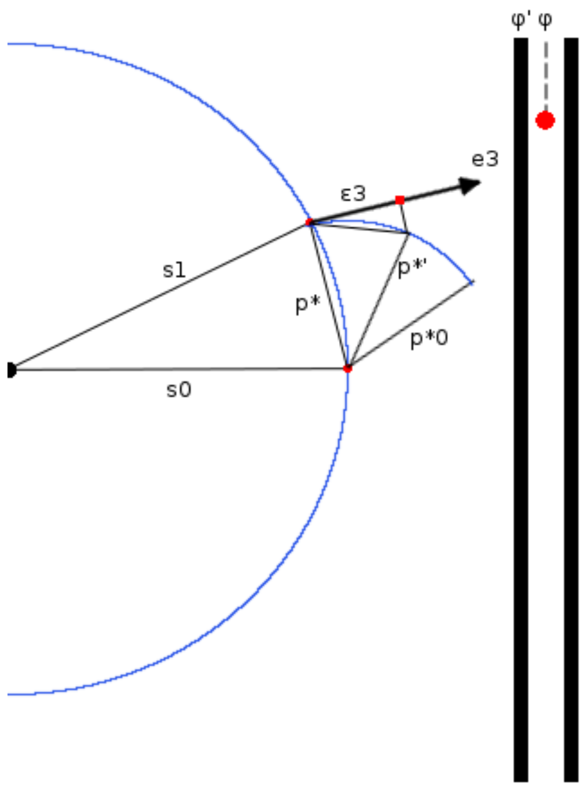


Pixels are mapped to the Ewald sphere

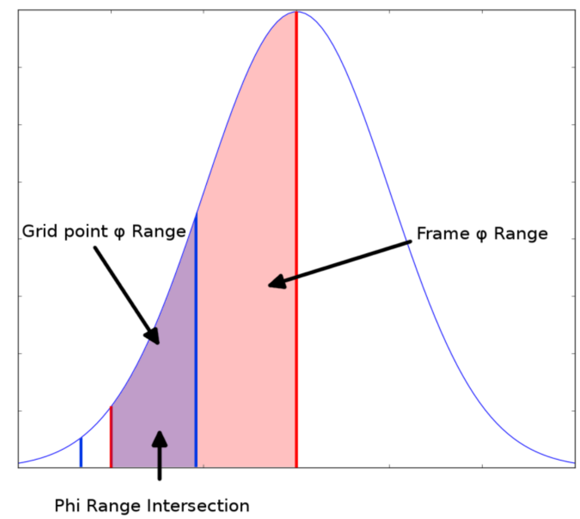


Counts are redistributed to Ewald sphere grid by computing fractional overlap of each pixel and Ewald sphere grid point

3D profile fitting phi gridding

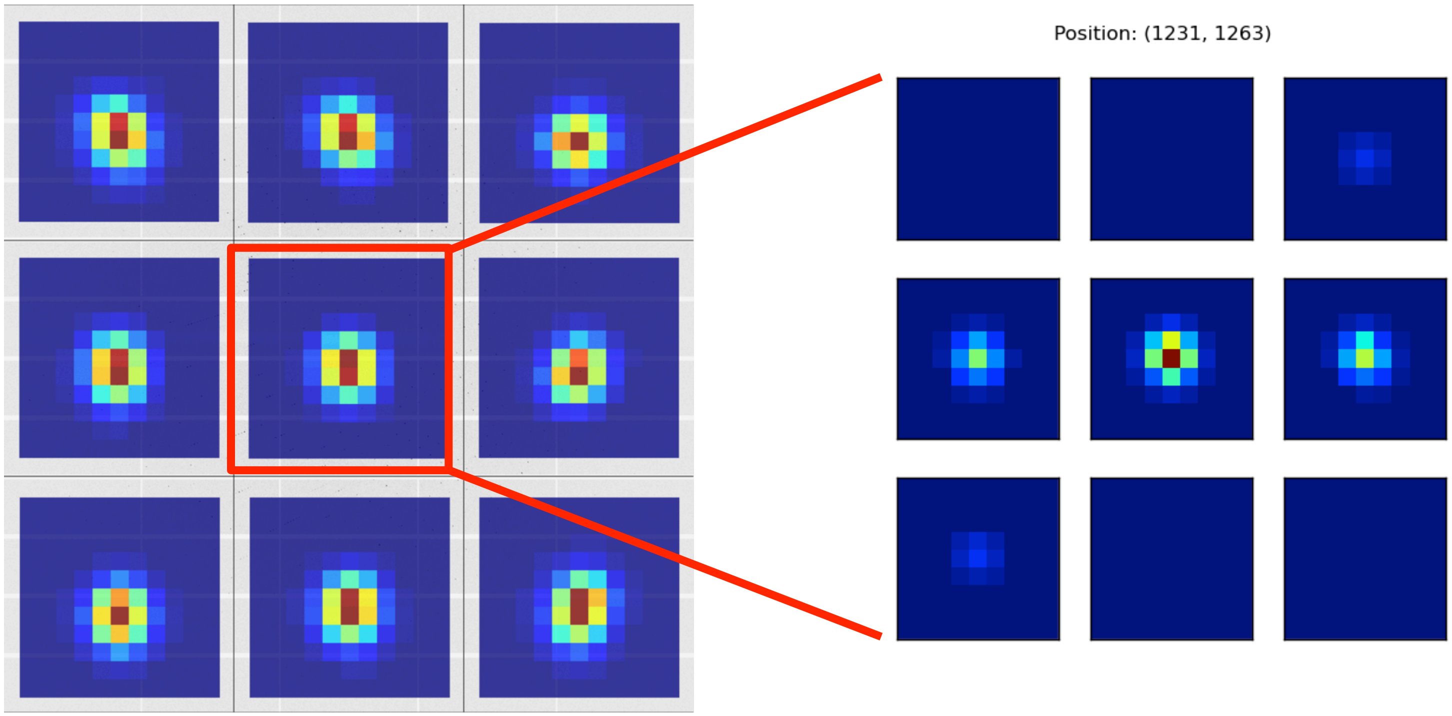


Frames are transformed to make reflection appear as if it took the shortest path through the Ewald sphere



Counts on each image are distributed by finding the angular overlap between each grid point and each image and integrating over the intersection

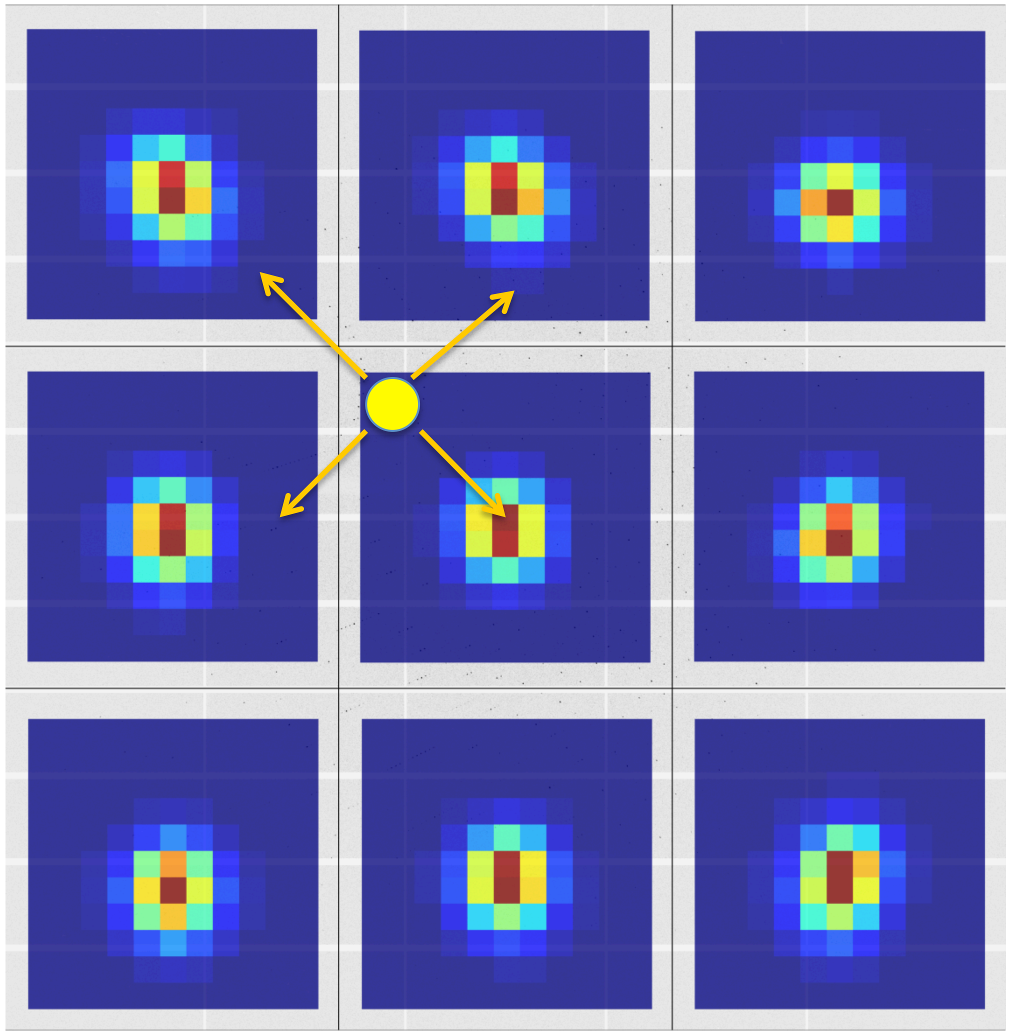
Building reference profiles



- Reference profiles are formed on a grid covering a given angular range
- Grid options include:
  - Rectangular grid (as in Mosflm)
  - Circular grid (as in XDS)
  - Single reflection (currently for multi-panel detectors)

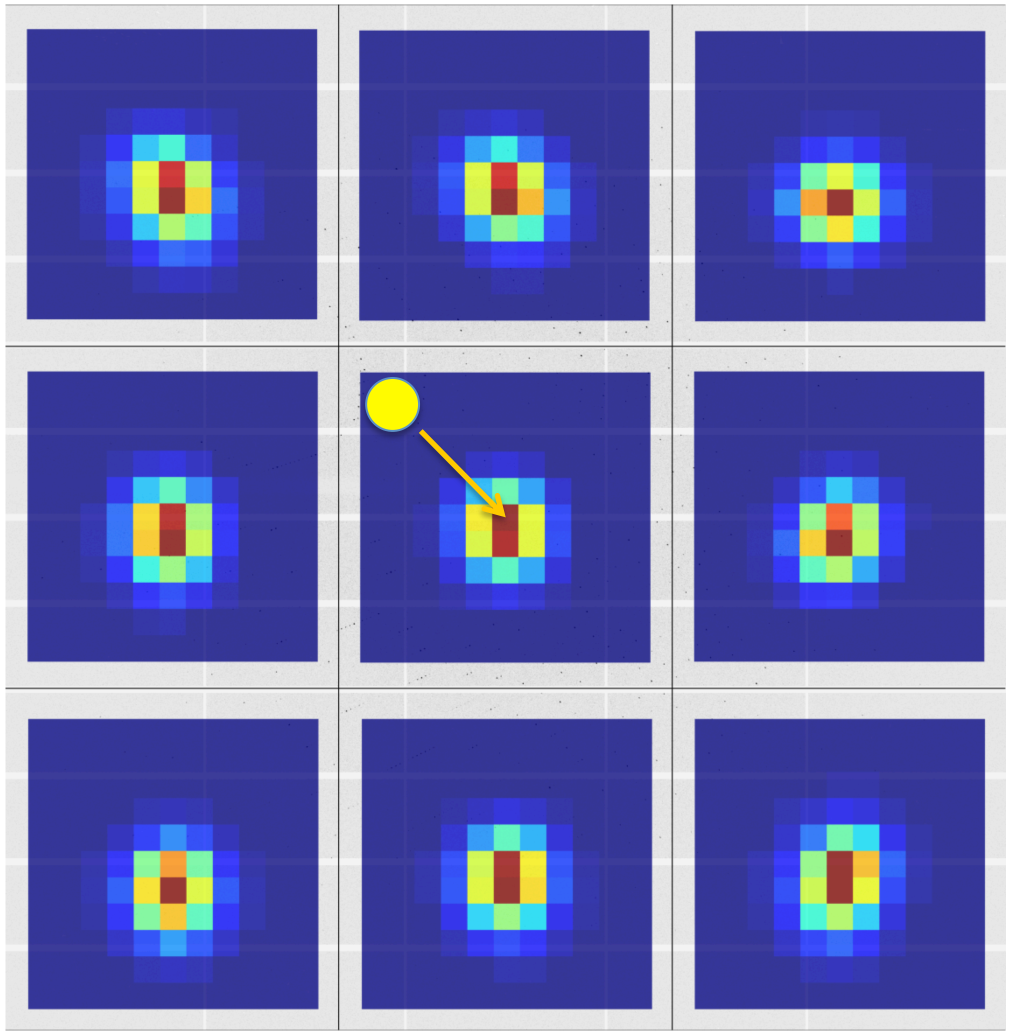
Building reference profiles

Each strong spot contributes to building the profile at adjacent grid points



Fitting reference profiles

Each reflection is fitted against its closest reference profile



Thanks!