U.S. DEPARTMENT OF
ENERGY

UNIVERSITY OF
CALIFORNIA

Practical considerations during processing of serial crystallographic XFEL data.

Aaron S. Brewster

BioXFEL workshop

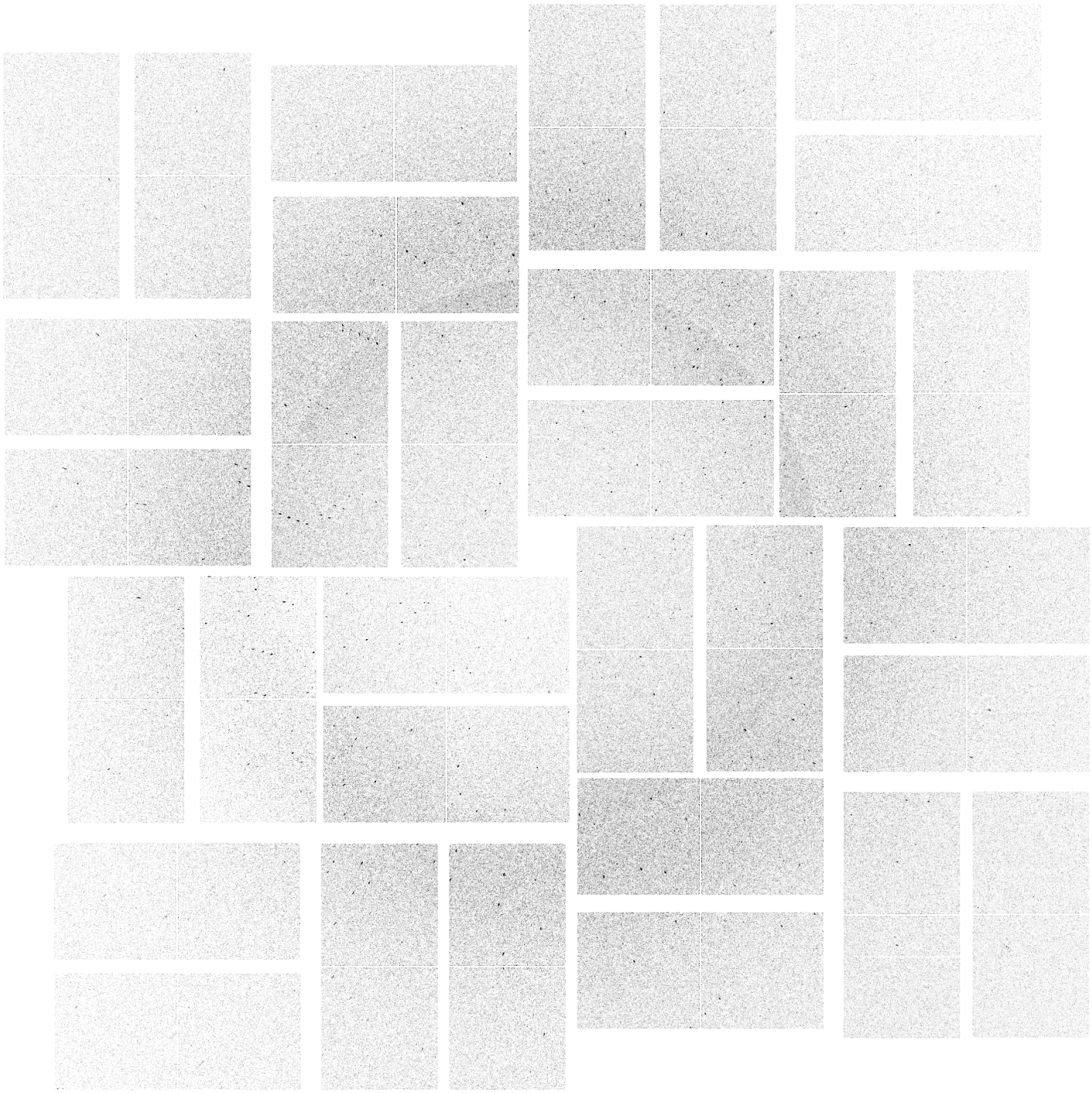
August 21st, 2014

Dataset

- Thermolysin dataset collected at CXI in March, 2014 (end of LCLS run 8)
- Focus on a single run, 2 minutes long, 14483 frames total
- Hitfinding: 3053 hits
  - At least 16 spots > 450 ADU per hit
- Indexing: 1923 images

All commands from this talk on the wiki:

http://cci.lbl.gov/xfel/index.php/LB67_Thermolysin

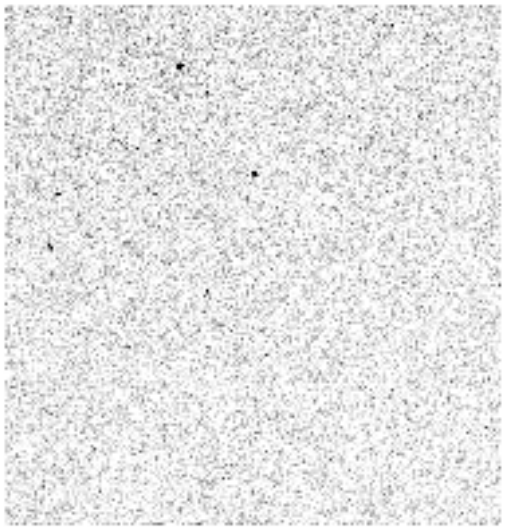
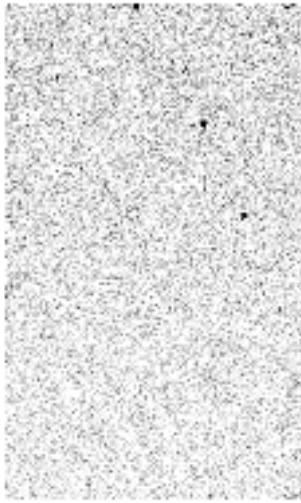


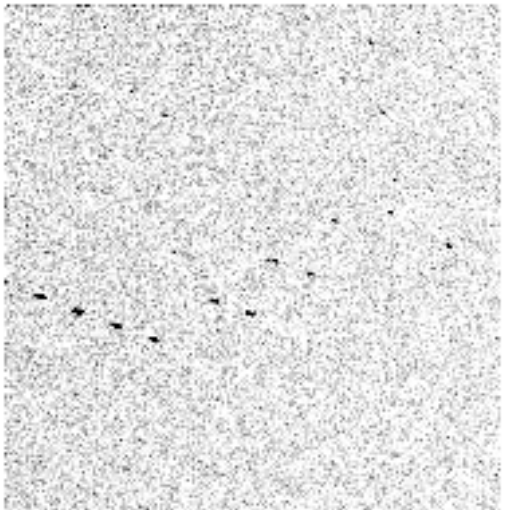
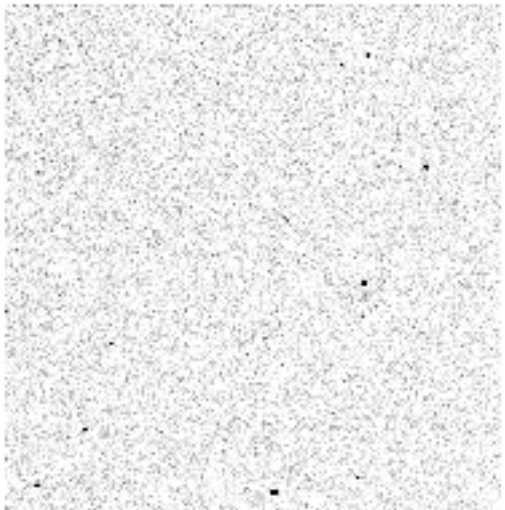
BERKELEY LAB



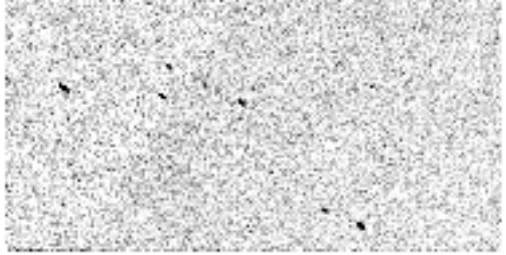


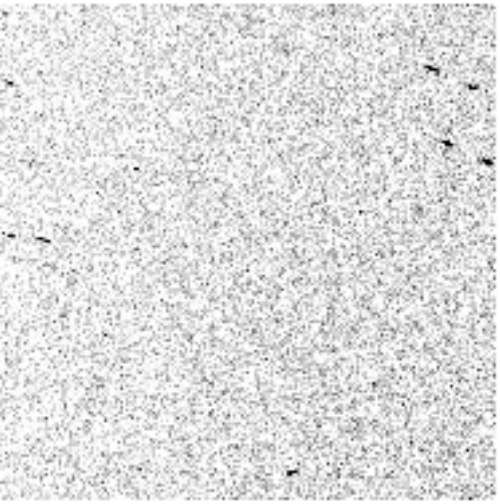


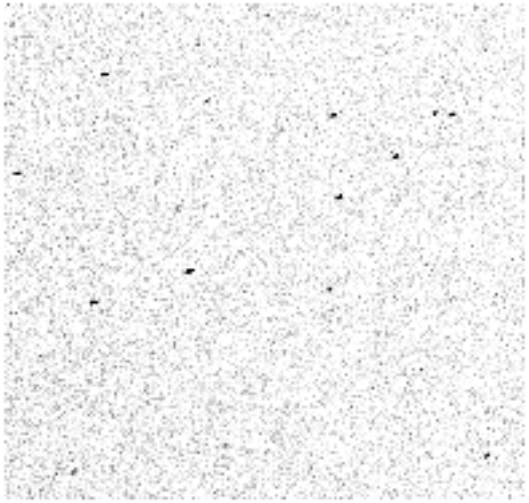
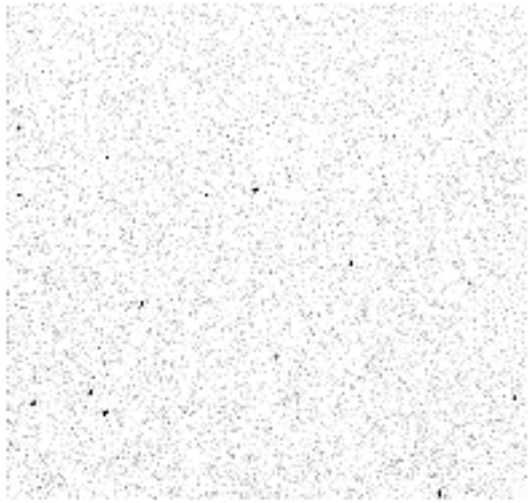




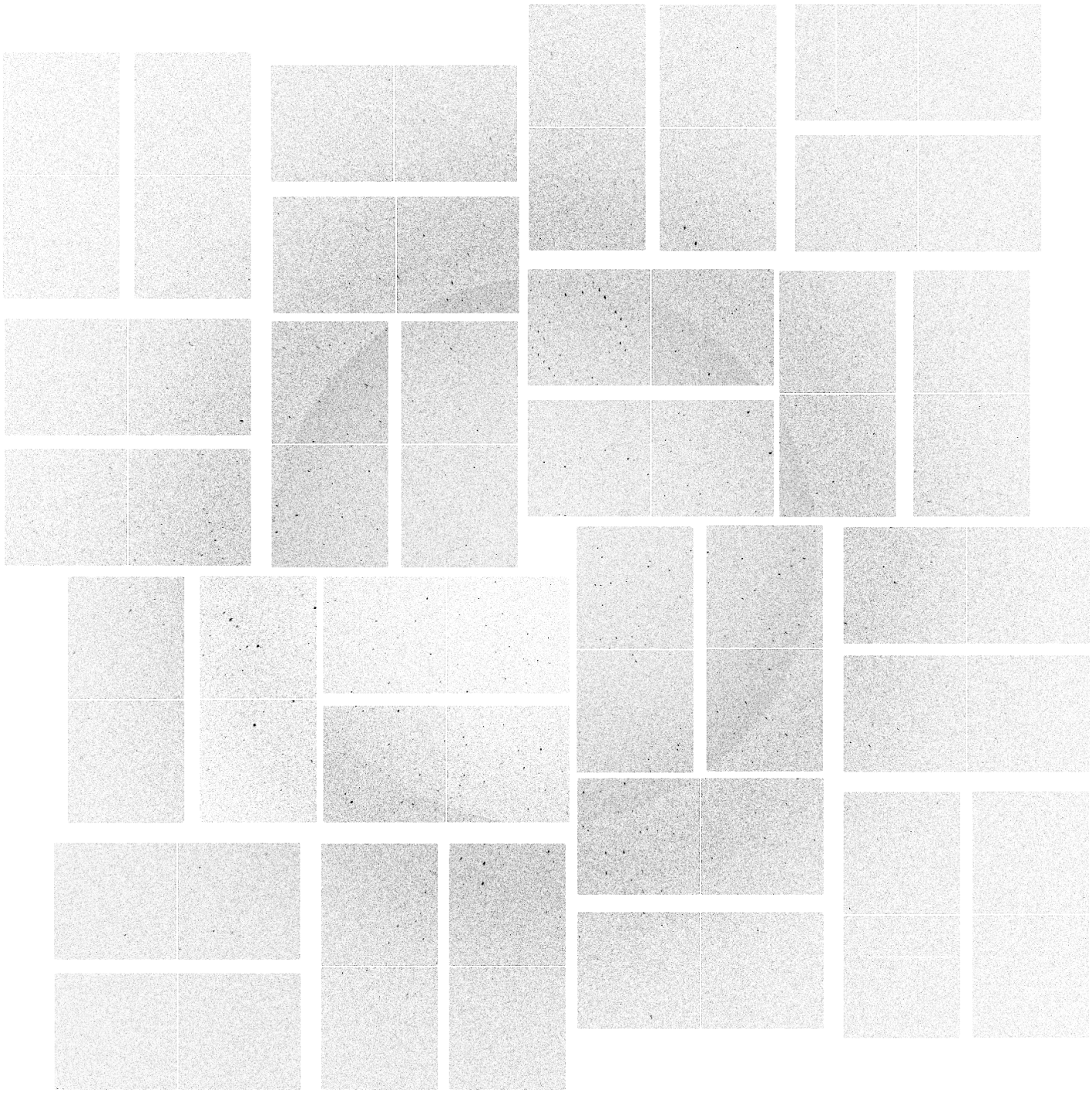









BERKELEY LAB

Three phases of an XFEL experiment

- Metrology
- Discovery
- Process

Metrology

"Where are my pixels in space?"

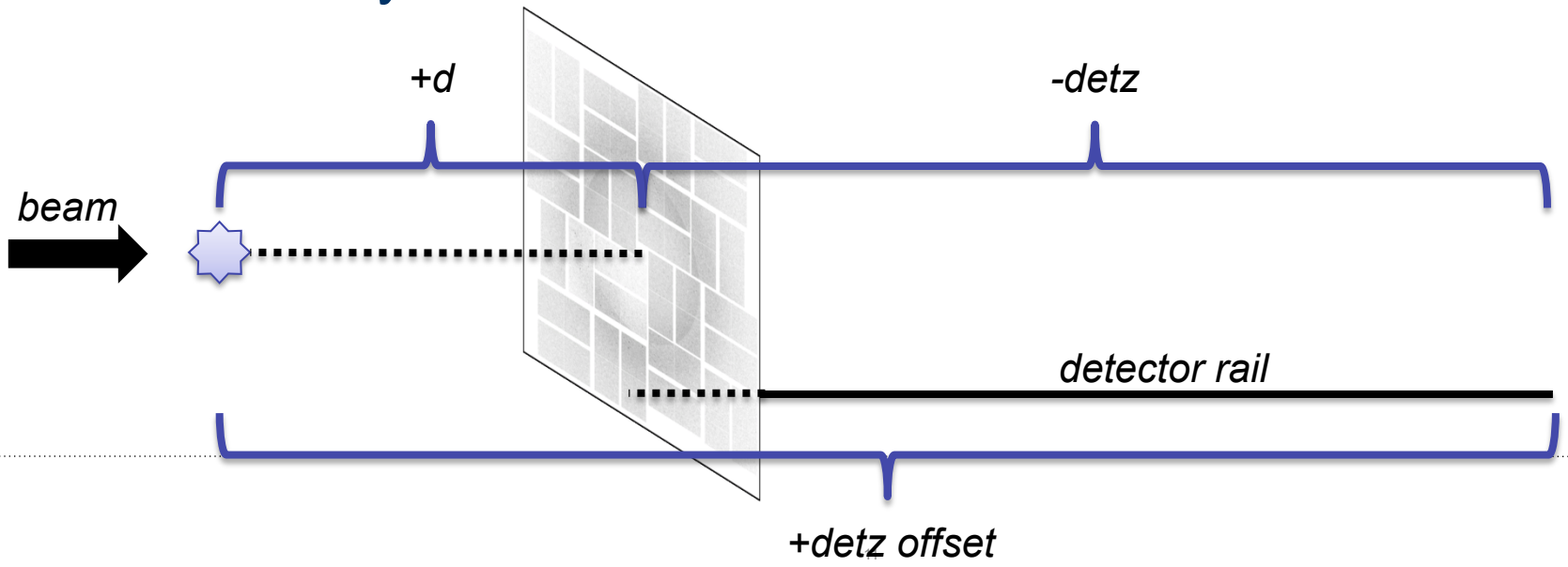
- Quadrants
  - Use averages to place quadrants aligning to rings
  - Account for beam center
- Tiles
  - cspad.metrology
- Detector distance/wavelength corrections

Metrology: beam center

- Beam center is defined as center of CSPAD
- Detector rail not parallel to beam
- Effect of 2 pixel shift in beam center on thermolysin dataset from tutorial
  - Normally: 1923 images indexed
  - With shift: 543 images indexed

Metrology: detector distance

- In the XTC stream: distance from back of rail to detector (detz)
- Desired: distance from detector to crystal (d)
- Optimize “detector z offset” (detz offset): distance from crystal to back of rail



Procedure: optimize detector distance

- Initial detz_offset: 572mm
- Write out new config files, changing detz_offset from 565 to 580:

```
for i in `seq 565 580`; do vi -c "%s/572/$i/g" -c "w
LB67-thermolysin_$i.cfg" -c q LB67-
thermolysin.cfg ; done
```

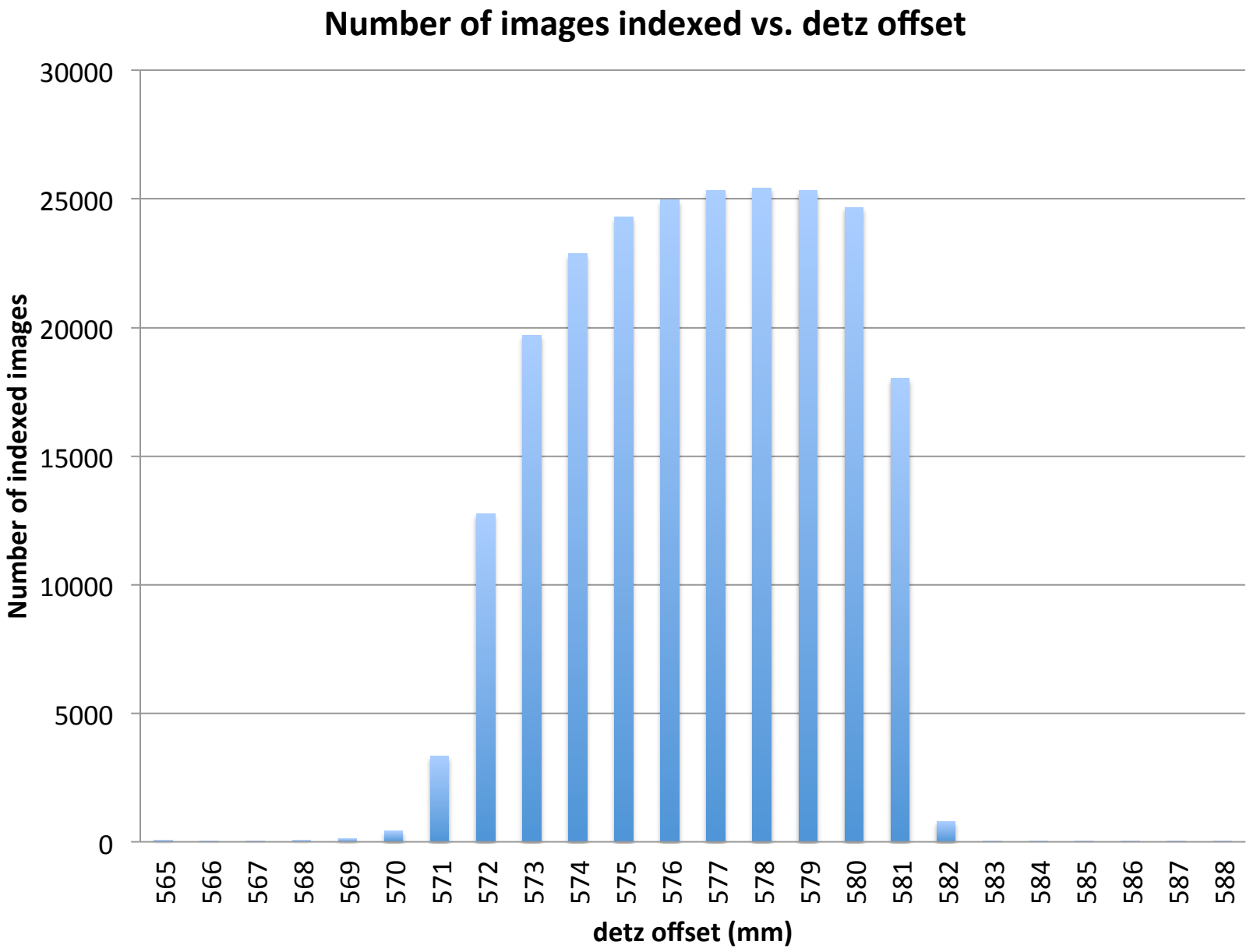
See wiki

- Submit batch jobs for these offsets, varying the trial number to match the offset:

```
for i in `seq 565 580`; do cxi.lsf -c cxib6714/
dist_trials/LB67-thermolysin_$i.cfg -o /reg/d/psdm/
cxi/cxib6714/ftc/brewster/dist_trials/ -x cxib6714 -r 30
-q psanacsq -p 8 -t $i; done
```

See wiki

BERKELEY LAB



Wavelength correction

- Sometime in 2013 the conversion between electron energy and photon energy drifted
- For data collected in runs 8 and 9, need to apply a correction, delta_k

$$\lambda = \frac{L}{2\gamma}\left(1 + \frac{(k + \Delta k)^2}{2}\right)$$

- Not applicable for last few experiments in run 9, onward

Margaritondoa and Ribica (2011)


Number of images indexed vs. detz offset
Number of indexed images
30000
25000
20000
15000
10000
5000
0
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
detz offset (mm)
Without Δk correction for λ
With Δk correction for λ

Metrology takeaways

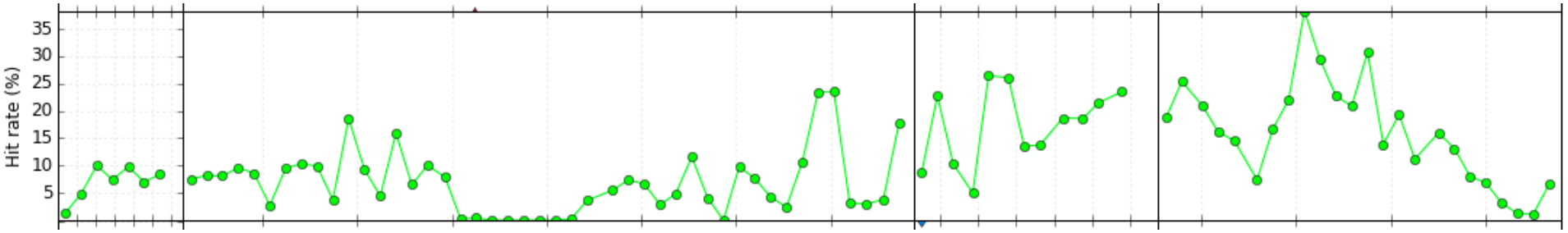
- Always look at an average image and verify the beam is centered
- Refine detz_offset
- Pay attention to unit cell parameters from indexing: they are the best indication that distance and wavelength are accurate

Discovery

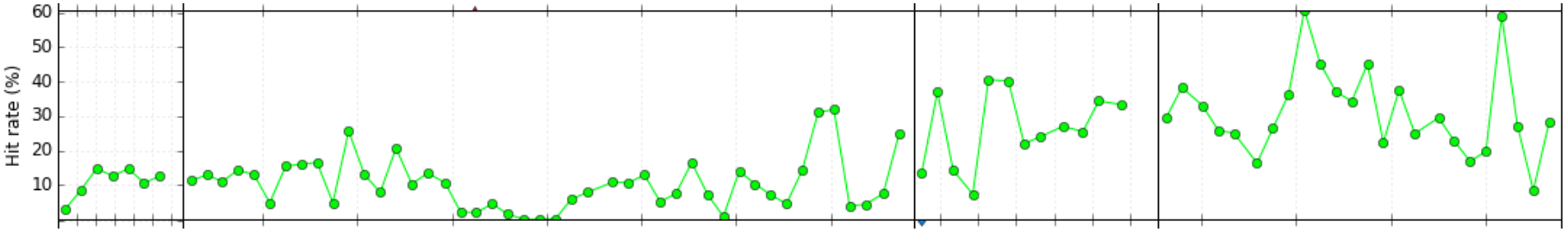
"What's the best way to process my data?"

- Get initial indexing solutions
  - If nothing is indexing, use hitfinder to find candidate images
  - Use distl.image_viewer to get first guesses at spotfinder parameters
  - Index with no target cell to discover unit cell dimensions
- Optimize parameters
  - Use lots of trials to find the best parameters
  - Consider grid searches

Hitfinder: parameterization effects



Hit: min number of spots = 16



Hit: min number of spots = 5

Spotfinding

- Which cctbx.xfel parameters should I focus on?
- cxi.parameters

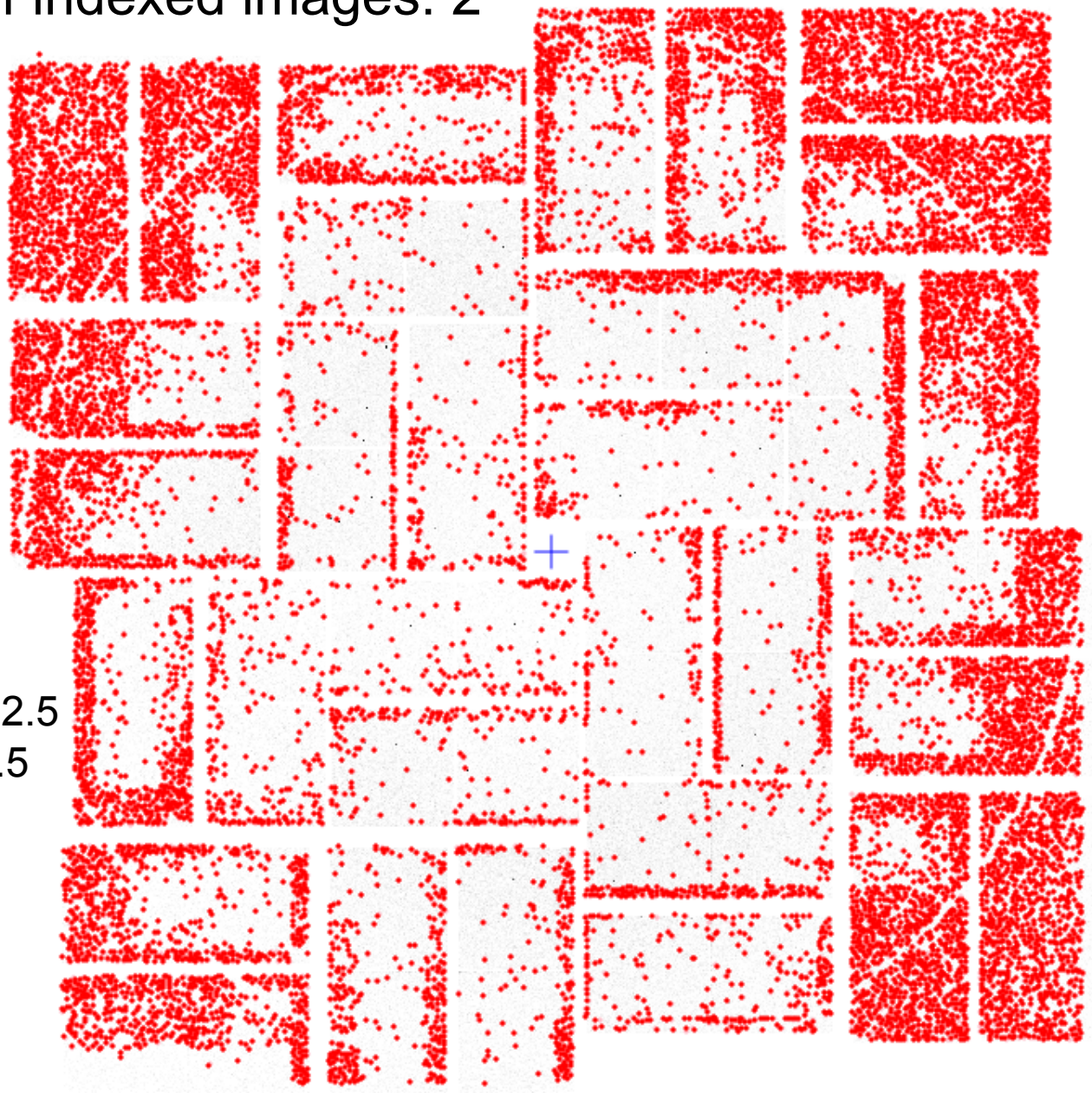
```
viewer {
  powder_arcs {
    show = False
    code = None
  }
  calibrate_silver = False
  calibrate_pdb {
    code = None
    d_min = 20
  }
  calibrate_unitcell {
    unitcell = None
    d_min = 20
    spacegroup = None
  }
}
target_cell = None
target_cell_centring_type = *P C I R F
known_symmetry = None
known_cell = None
distl {
  image = None
  res {
    outer = None
    inner = None
  }
  verbose = False
  dxtbx = False
  bins {
    verbose = True
    N = 20
    corner = True
  }
}
wedgelimit = 2
goniometer_rotation = ""
convention_override = None
spot_convention = None
override_pickled_spotfinders = True
spotfinder_header_tests = True
spotfinder_mode = "distl"
spotfinder_verbose = False
force_method2_resolution_limit = None
distl_lowres_limit = 50
distl_highres_limit = None
distl_binned_image_spot_size = 4
distl_maximum_number_spots_for_indexi
ng = 300
distl_minimum_number_spots_for_indexin
g = 40
distl_profile_bumpiness = 2
distl_report_overloads = True
distl_keep_Zdata = True
percent_overlap_forcing_detail = 30
overlapping_spot_criterion = 1.2
spots_pickle = "./DISTL_pickle"
distl_spotcenter_algorithm =
"center_of_mass"
distl_permit_binning = False
distl_force_binning = False
autoindex_override_beam = None
autoindex_override_distance = None
autoindex_override_wavelength = None
autoindex_override_twotheta = None
autoindex_override_deltaphi = None
image_specific_osc_start = None
codecamp {
  maxcell = None
  minimum_spot_count = None
}
pdf_output {
  file = ""
  box_selection = "all"
  enable_legend = False
  enable_legend_font_size = 10
  enable_legend_ink_color = "black"
  enable_legend_vertical_offset = 10
  box_linewidth = 0.04
  window_fraction = 0.666666
  window_offset_x = 0.16667
  window_offset_y = 0.16667
  markup_inliers = True
  render_all = False
  profile_shrink = 0
}
distl {
  minimum_spot_area = 3
  minimum_signal_height = None
  minimum_spot_height = None
  spot_area_maximum_factor = None
  peak_intensity_maximum_factor = 10000
  method2_cutoff_percentage = 20
  compactness_filter = True
  detector_tiling = None
  tile_translations = None
  tile_flags = None
  quad_translations = None
  detector_format_version = None
  scanbox_windows = 101 51 51
  peripheral_margin = 1
  pdf_output = None
  image_viewer = Fa
  port = 8125
  processors = 1
  nproc = 1
}
spotfinder = *distl speck
speckfinder {
  dark_stddev = ""
  dark_adu_scale = 100
}
indexing {
  data = None
  indexing_pickle = None
  completeness_pickle = None
  open_wx_viewer = False
  verbose_cv = False
  lattice_model_scoring_cutoff = 2
  devel_algorithm = None
  outlier_detection {
    allow = True
    switch = False
    verbose = False
    pdf = None
  }
  plot_search_scope = False
  mm_search_scope = 4
  improve_local_scope = *origin_offset
S0_vector
  }
  predictions_file = ""
parallel = 0
integration {
  file_template = None
  file_range = None
  rocking_curve = *none gh1982a
  mosaicity_deg = 0
  guard_width_sq = 11
  detector_gain = 0
  background_factor = 1
  model = *rossmann1979jac12-225
use_case_2_simulated_annealing \
                    ulated_annealing_7
                    ulated_annealing_9 \
      user_supplied
  use_subpixel_translations = None
  subpixel_joint_model {
    rotations = None
    translations = None
  }
  spot_shape_verbose = False
  signal_penetration = 0.5
  spotfinder_subset = *inlier_spots
goodspots spots_non-ice
  mask_pixel_value = None
  mosaic {
    refinement_target = *LSQ ML
    kludge1 = 1
    bugfix2_enable = True
    domain_size_lower_limit = 0
    enable_rotational_target_highsym =
True
    enable_rotational_target_triclinic = True
  }
  center_on_water_ring = False
  fft_sampling_granularity = 1
  beam_search_scope = 4
  autoindex_distance_grid_search = False
  beamplot_pdf_file = None
  mosaicity_spot_coverage = 0.8
  mosaicity_limit = 1.5
  autoindex_min_osc_spacing = 4
  ad_hoc_transformer = None
  model_refinement_minimum_N = 40
  refinement_partiality_weight = 0
  lepage_max_delta = 1.4
  heuristic_stdev_factor = 15
  rmsd_tolerance = 3.5
  subgroups_pickle = "./LABELIT_possible"
  index_only = False
  mosflm_integration_reslimit_override =
None
  best_support = False
  refinements_pickle = "./LABELIT_pickle"
  mosflm_rmsd_tolerance = 2.5
  difflimit_sigma_cutoff = 0.75
  difflimit_verbose = False
  difflimit_table = False
  rsymop_integration_permissible_resolutio
n = None
  rsymop_statistics_sigma_cutoff = 5
  exceptions_pickle = "./
  LABELIT_exceptions"
  override_pickled_exceptions = True
  mosflm_safety_algorithm = "corner"
  known_setting = None
  mosaicity = None
  image_brightness = 1
  reticular_allow = False
  reticular_command = "Supercell"
  reticular_pdf_file = ""
  sublattice_verbose = False
  sublattice_allow = True
  sublattice_maximum_modulus = 3
  sublattice_force_index = ""
  sublattice_significance_cutoff = 2
  sublattice_filter_next_layer = True
  sublattice_print_coset_signal = False
  sublattice_bin_count = None
  sublattice_bin_limit = None
  sublattice_bin_precision = *intermediate
  publication
  take_superlattice = False
  compatibility_allow = False
  compatibility_file = ""
  compatibility_column_label = IMEAN I F
  ad_hoc_transformation = "1.0"
  ad_hoc_resolution = 3
```

See wiki

Spotfinding

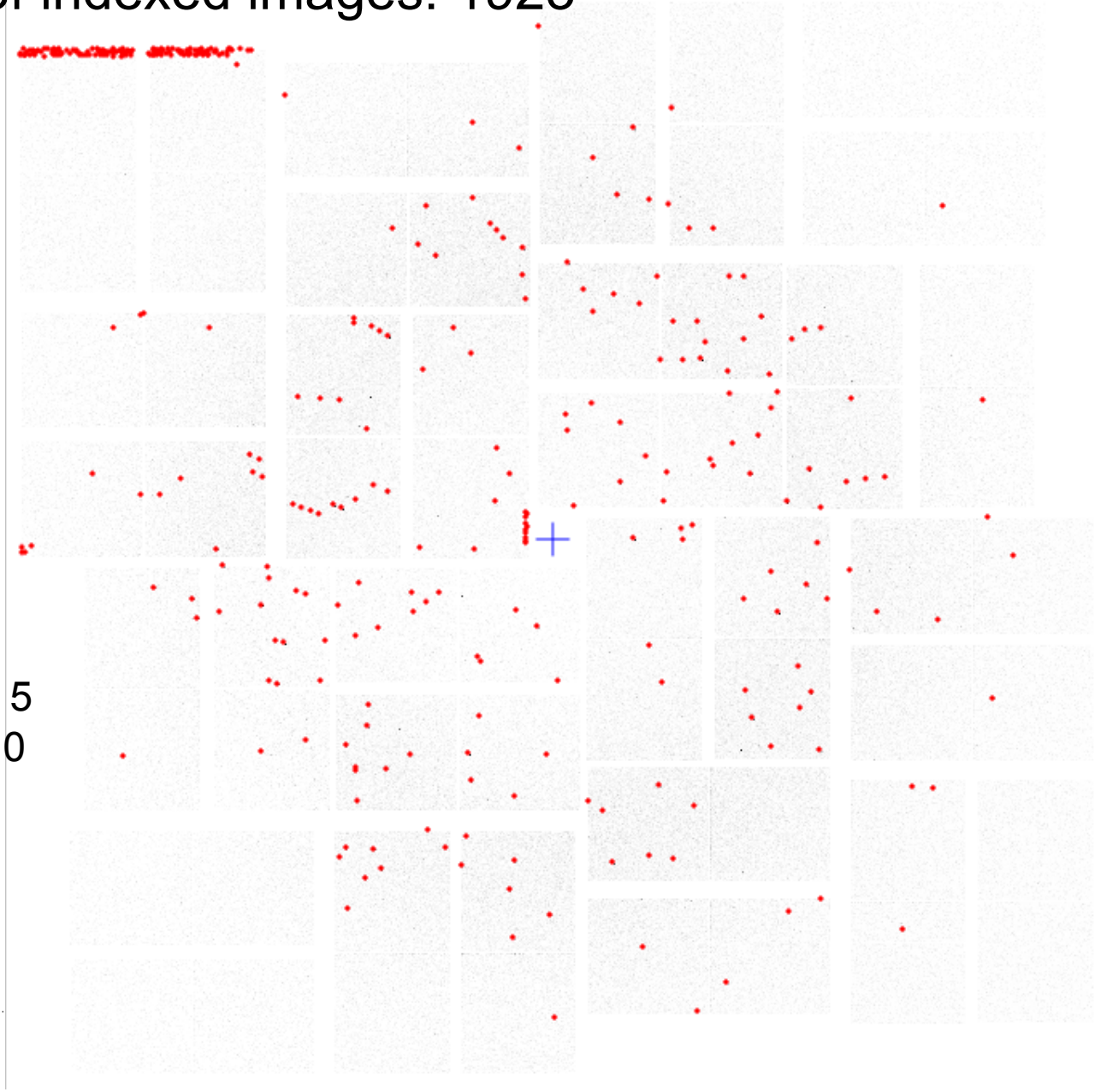
- Parameters with strong effects on indexing
  - distl.minimum_spot_area = 1
    - Only accept spots larger than this area in pixels
    - In LABELIT, defaults are larger, accounting for larger spots on CCDs
  - distl.minimum_signal_height = 5
    - Minimum number of sigmas above background for pixels to be signal
  - distl.minimum_spot_height = 10
    - If signal, number of sigmas above background required to be peak maximum
- What kinds of effects can these have?

Number of indexed images: 2



Spot area: 3
Signal height: 2.5
Spot height: 3.5

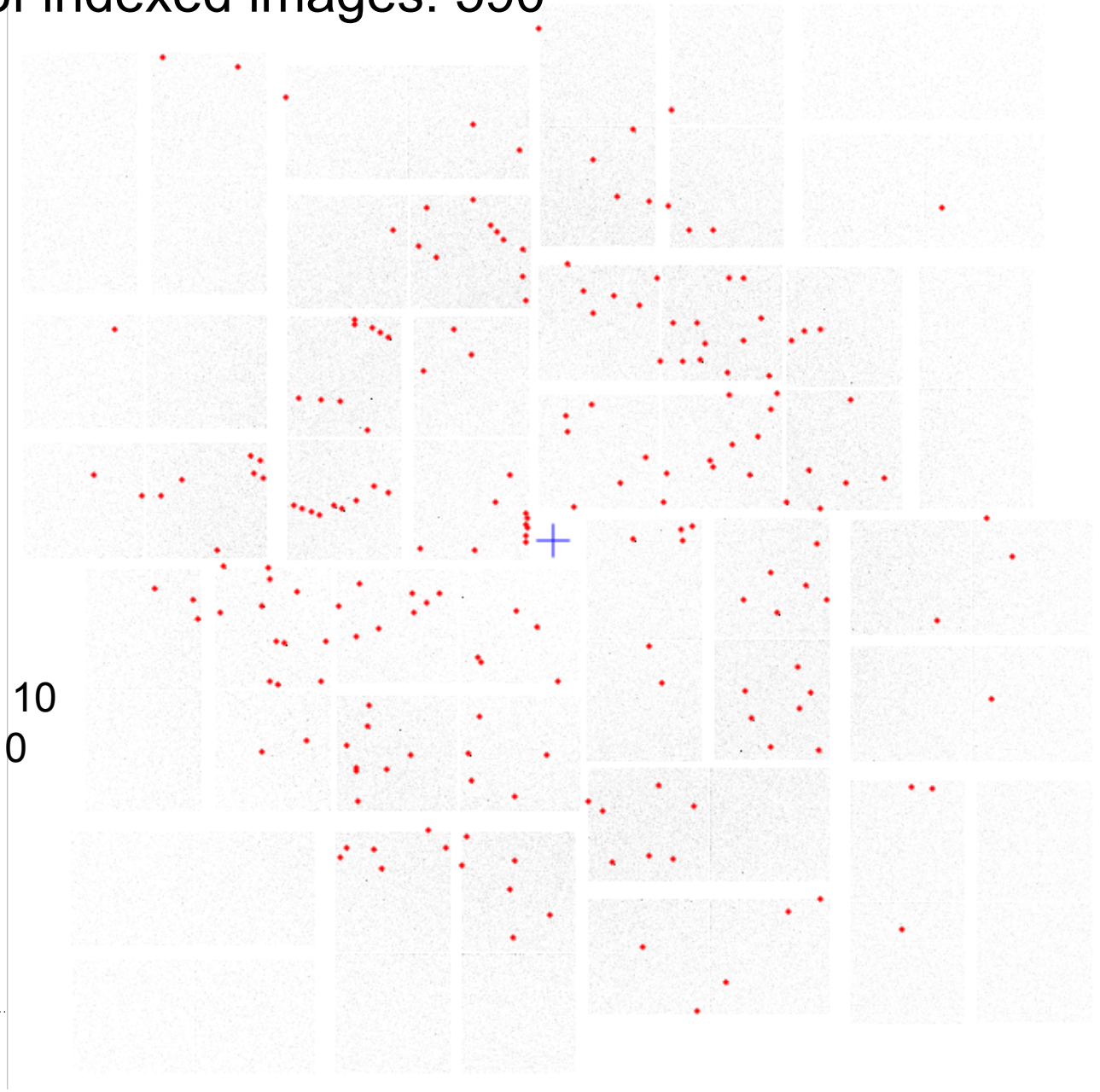
Number of indexed images: 1923



Spot area: 1
Signal height: 5
Spot height: 10

BERKELEY LAB

Number of indexed images: 590



Spot area: 1
Signal height: 10
Spot height: 10

Optimize spotfinder parameters

- Empirically: change parameters one at a time using distl.image_viewer and cxi.index to visualize results
- Systematically: grid search
  - Choose N parameters to test exhaustively
  - Time consuming:
    - Spot area: 1-5
    - Min signal height 1-10
    - Min spot height N-10
    - 275 combinations * 7 minutes each * 12 streams each / 48 nodes = 8 hours

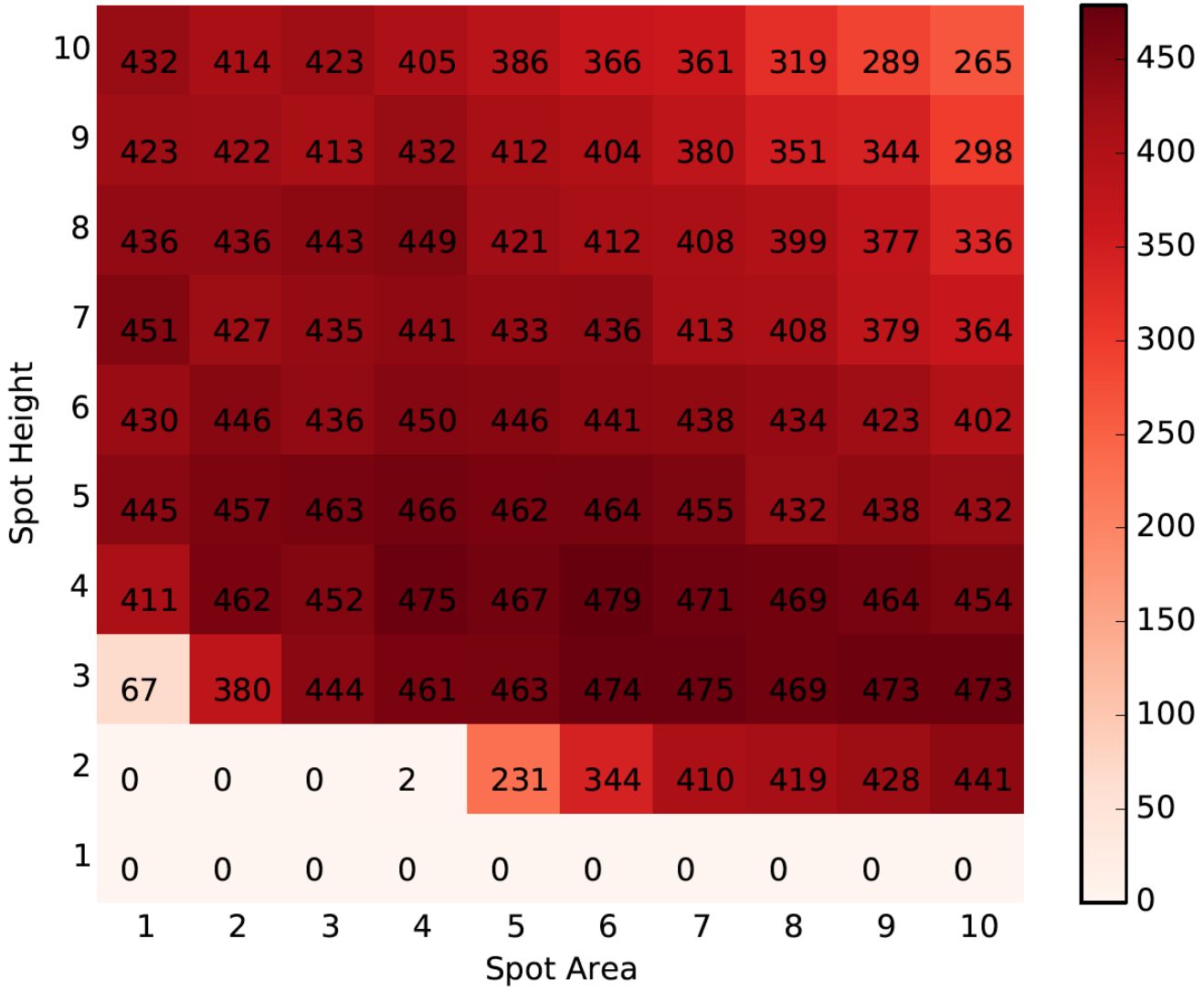
Example grids

s4	Signal/spot height								
Spot area	2	3	4	5	6	7	8	9	10
1	0	0	0	29	57	70	66	64	56
2	0	4	58	63	58	57	55	50	42
3	0	20	64	63	51	46	48	39	36
4	0	35	56	53	41	35	27	25	20
5	0	35	45	38	30	23	13	15	8

s6	Signal/spot height								
Spot area	2	3	4	5	6	7	8	9	10
1	0	0	29	120	128	112	97	93	79
2	0	35	124	110	92	81	71	57	45
3	0	75	104	98	69	55	47	38	35
4	2	83	73	50	35	19	17	12	3
5	2	79	50	34	17	9	6	1	1

s7	Signal/spot height								
Spot area	2	3	4	5	6	7	8	9	10
1	0	0	0	4	85	122	135	127	119
2	0	7	98	126	133	124	104	96	85
3	0	23	112	120	113	102	84	69	56
4	0	48	108	101	81	53	30	25	12
5	0	60	98	74	39	27	16	7	5

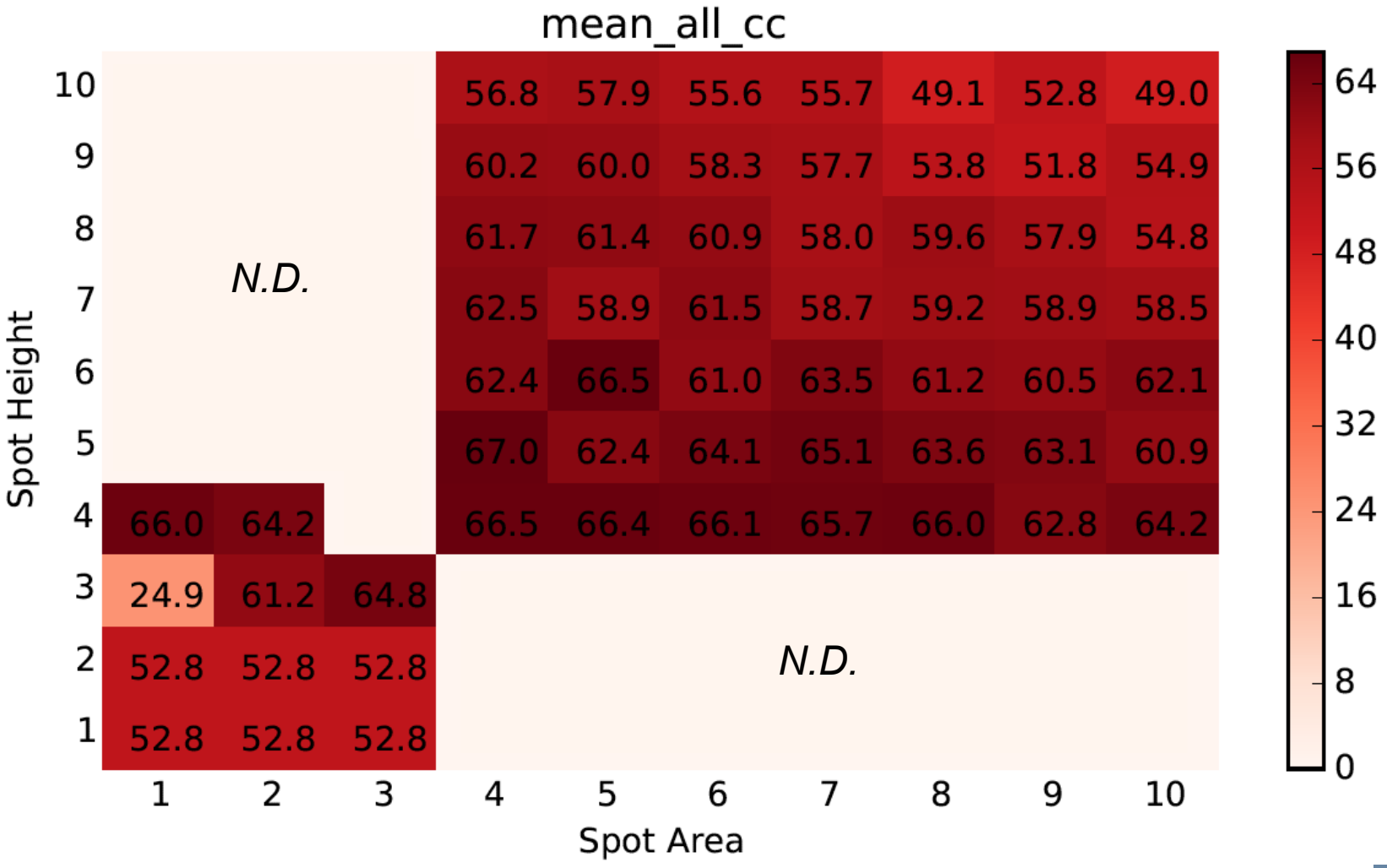
Example grids



Courtesy Oliver Zeldin, Stanford

BERKELEY LAB

Gridding: CC1/2



Courtesy Oliver Zeldin, Stanford

BERKELEY LAB

Discovery takeaways

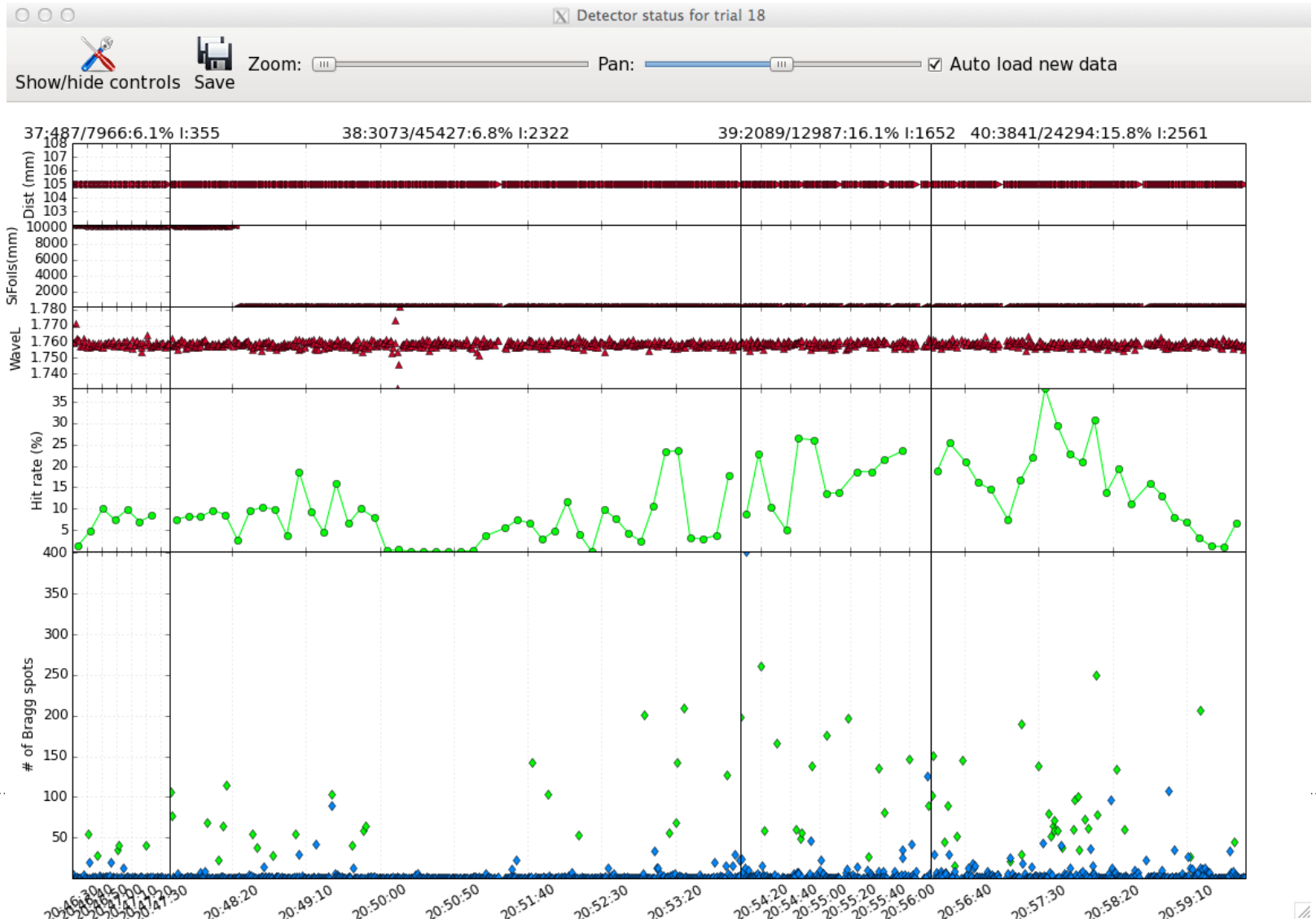
- Can’t index? Use hitfinder to generate images, find a good one, and try to index it with cxi.index
- Spot area, signal height, spot height are critical parameters
- Consider grid searches for good parameters

Process

"What's the fastest way to process my data?"

- Now that the parameters are known, process the entire experiment
- During the experiment
  - Real-time feedback on indexing rates
  - Real-time monitoring of completeness/resolution
- After the experiment
  - Submit large batches with new parameters
  - Remove hitfinder altogether

cxi.monitor_trials



cxi.trial_stats

DEMO

Processing takeaways

- Runtime goal: keep up with processing as close to real-time as possible in order to provide meaningful feedback for beamline operators and sample injection scientists that they can use to change and improve their operations on the fly, to the end of improving data quality and completeness.

Acknowledgements

Berkeley National Lab
Nicholas Sauter
Muhamed Amin
Tara Michels-Clark
Iris Young
Nat Echols
Paul Adams
Peter Zwart
Vittal Yachandra
Junko Yano
Jan Kern
James Holton

Janelia Farm
Johan Hattne

LCLS
Uwe Bergmann
…and many others

Diamond Light Source
David Stuart
Gwyndaf Evans
Graeme Winter
Jonathan Grimes
Richard Gildea
James Parkhurst
Luis Fuentes-Montero

CCP4
David Waterman

UCLA
David Eisenberg
Duilio Cascio
Michael Sawaya
Jose Rodriguez
Luki Goldschmidt

IBS
Jacques-Philippe Colletier

Stanford University
Axel Brunger
Mona Uervirojnangkoorn
Oliver Zeldin

SSRL
Mike Soltis
Ana Gonzalez
Ashley Deacon
Aina Cohen
Yingssu Tsai
Scott McPhillips

BNL
Allen Orville

NIH/NIGMS grants 1R01GM095887 and 1R01GM102520
DOE/Office of Science contract DE-AC02-05CH11231

BERKELEY LAB