NIH Data Science at NIH

HKL approach to handling Eiger images

Wladek Minor

HDMRX, BNL, May 2016

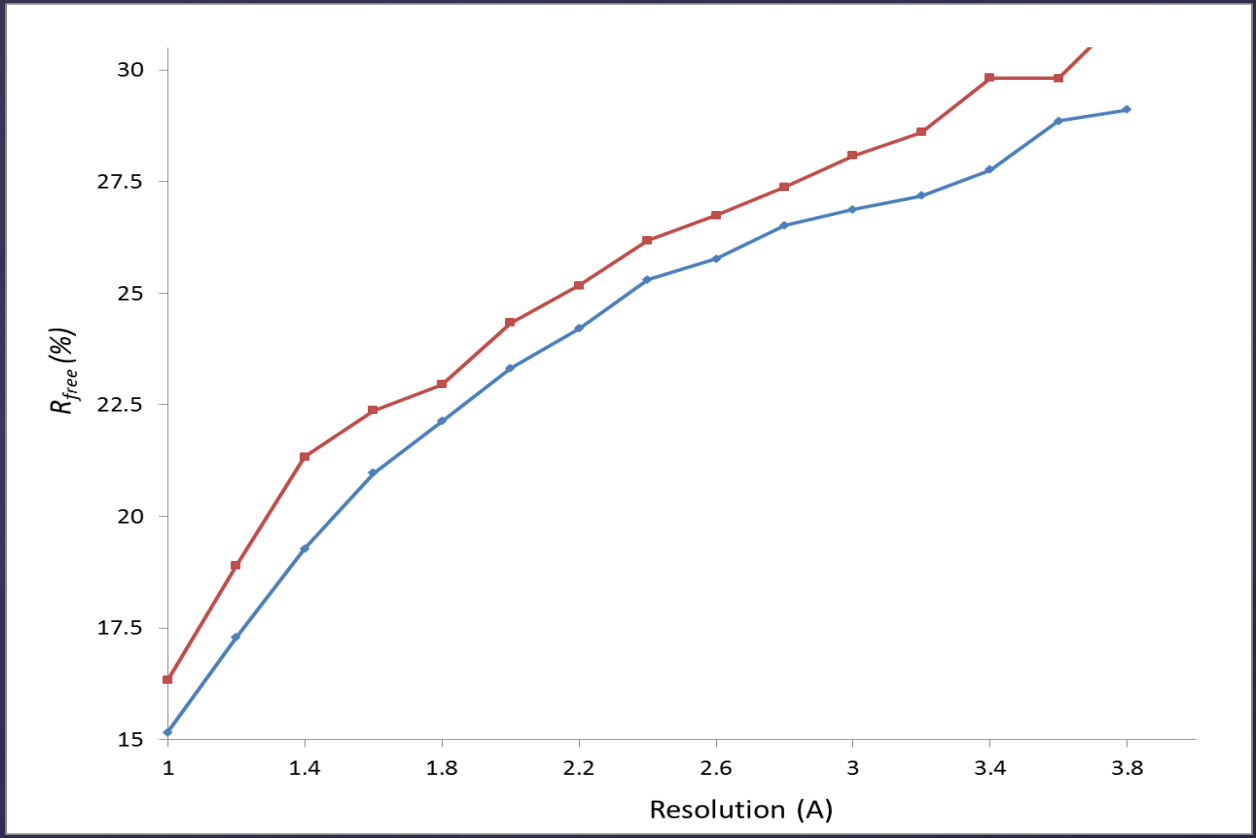
What experimenters know about data collection ?

```
REMARK   3  ESTIMATED OVERALL COORDINATE ERROR.
REMARK   3   ESU BASED ON R VALUE                            (A): NULL
REMARK   3   ESU BASED ON FREE R VALUE                       (A): NULL
REMARK   3   ESU BASED ON MAXIMUM LIKELIHOOD                 (A): NULL
REMARK   3   ESU FOR B VALUES BASED ON MAXIMUM LIKELIHOOD (A**2): NULL
REMARK   3
REMARK   3  RMS DEVIATIONS FROM IDEAL VALUES.
REMARK   3   DISTANCE RESTRAINTS.                      RMS    SIGMA
REMARK   3    BOND LENGTH                     (A) : NULL  ; NULL
REMARK   3    ANGLE DISTANCE                  (A) : NULL  ; NULL
REMARK   3    INTRAPLANAR 1-4 DISTANCE        (A) : NULL  ; NULL
REMARK   3    H-BOND OR METAL COORDINATION    (A) : NULL  ; NULL
REMARK   3
REMARK   3   PLANE RESTRAINT                  (A) : NULL  ; NULL
REMARK   3   CHIRAL-CENTER RESTRAINT       (A**3) : NULL  ; NULL
REMARK   3
REMARK   3   NON-BONDED CONTACT RESTRAINTS.
REMARK   3    SINGLE TORSION                  (A) : NULL  ; NULL
REMARK   3    MULTIPLE TORSION                (A) : NULL  ; NULL
REMARK   3    H-BOND (X...Y)                  (A) : NULL  ; NULL
REMARK   3    H-BOND (X-H...Y)                (A) : NULL  ; NULL
REMARK   3
REMARK   3   CONFORMATIONAL TORSION ANGLE RESTRAINTS.
REMARK   3    SPECIFIED                 (DEGREES) : NULL  ; NULL
REMARK   3    PLANAR                    (DEGREES) : NULL  ; NULL
REMARK   3    STAGGERED                 (DEGREES) : NULL  ; NULL
REMARK   3    TRANSVERSE                (DEGREES) : NULL  ; NULL
REMARK   3
REMARK   3  ISOTROPIC THERMAL FACTOR RESTRAINTS.     RMS    SIGMA
REMARK   3   MAIN-CHAIN BOND                (A**2) : NULL  ; NULL
REMARK   3   MAIN-CHAIN ANGLE               (A**2) : NULL  ; NULL
REMARK   3   SIDE-CHAIN BOND                (A**2) : NULL  ; NULL
```

What experimenters know about data collection ?

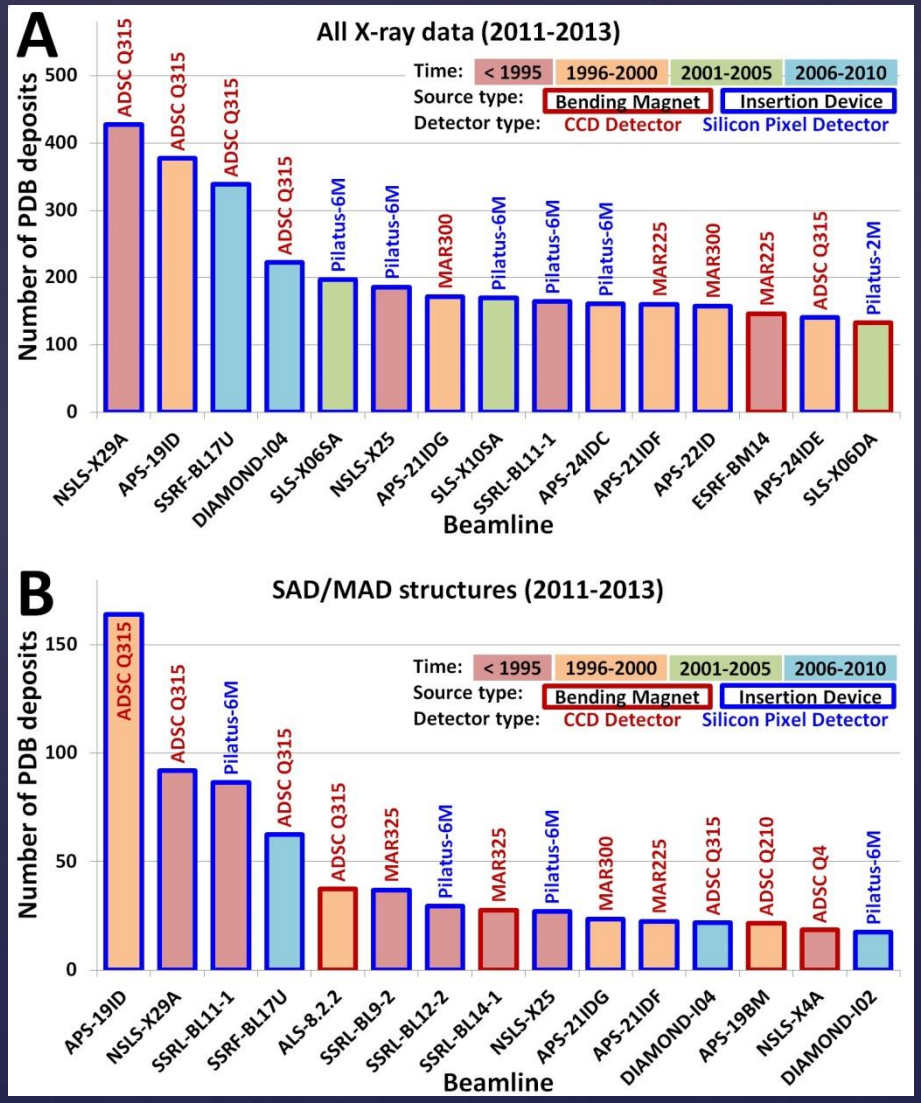
```
REMARK 200  DETECTOR TYPE                    : CCD
REMARK 200  DETECTOR MANUFACTURER            : ADSC QUANTUM 4
REMARK 200  INTENSITY-INTEGRATION SOFTWARE   : BLU-ICE
REMARK 200  DATA SCALING SOFTWARE            : MOSFLM, CCP4, SCALEPACK
REMARK 200
REMARK 200  NUMBER OF UNIQUE REFLECTIONS     : 17575
REMARK 200  RESOLUTION RANGE HIGH      (A)   : 2.950
REMARK 200  RESOLUTION RANGE LOW       (A)   : 47.870
REMARK 200  REJECTION CRITERIA  (SIGMA(I))   : NULL
REMARK 200
REMARK 200 OVERALL.
REMARK 200  COMPLETENESS FOR RANGE     (%)   : 94.3
REMARK 200  DATA REDUNDANCY                  : 3.200
REMARK 200  R MERGE                    (I)   : 0.07800
REMARK 200  R SYM                      (I)   : 0.08600
REMARK 200  <I/SIGMA(I)> FOR THE DATA SET    : NULL
REMARK 200
REMARK 200 IN THE HIGHEST RESOLUTION SHELL.
REMARK 200  HIGHEST RESOLUTION SHELL, RANGE HIGH (A) : 2.95
REMARK 200  HIGHEST RESOLUTION SHELL, RANGE LOW  (A) : 3.03
REMARK 200  COMPLETENESS FOR SHELL     (%)   : 92.4
REMARK 200  DATA REDUNDANCY IN SHELL         : 2.80
REMARK 200  R MERGE FOR SHELL          (I)   : NULL
REMARK 200  R SYM FOR SHELL            (I)   : NULL
REMARK 200  <I/SIGMA(I)> FOR SHELL           : NULL
REMARK 200
REMARK 200 DIFFRACTION PROTOCOL: SINGLE WAVELENGTH
REMARK 200 METHOD USED TO DETERMINE THE STRUCTURE: MAD SE-MET
REMARK 200 SOFTWARE USED: SNB, MLPHARE, CCP4, SOLVE, HKL, RESOLE
REMARK 200 STARTING MODEL: NULL
REMARK 200
REMARK 200 REMARK: NULL
REMARK 280
REMARK 280 CRYSTAL
REMARK 280 SOLVENT CONTENT, VS    (%): NULL
REMARK 280 MATTHEWS COEFFICIENT, VM (ANGSTROMS**3/DA): NULL
REMARK 280
```

Unexpected correlation?



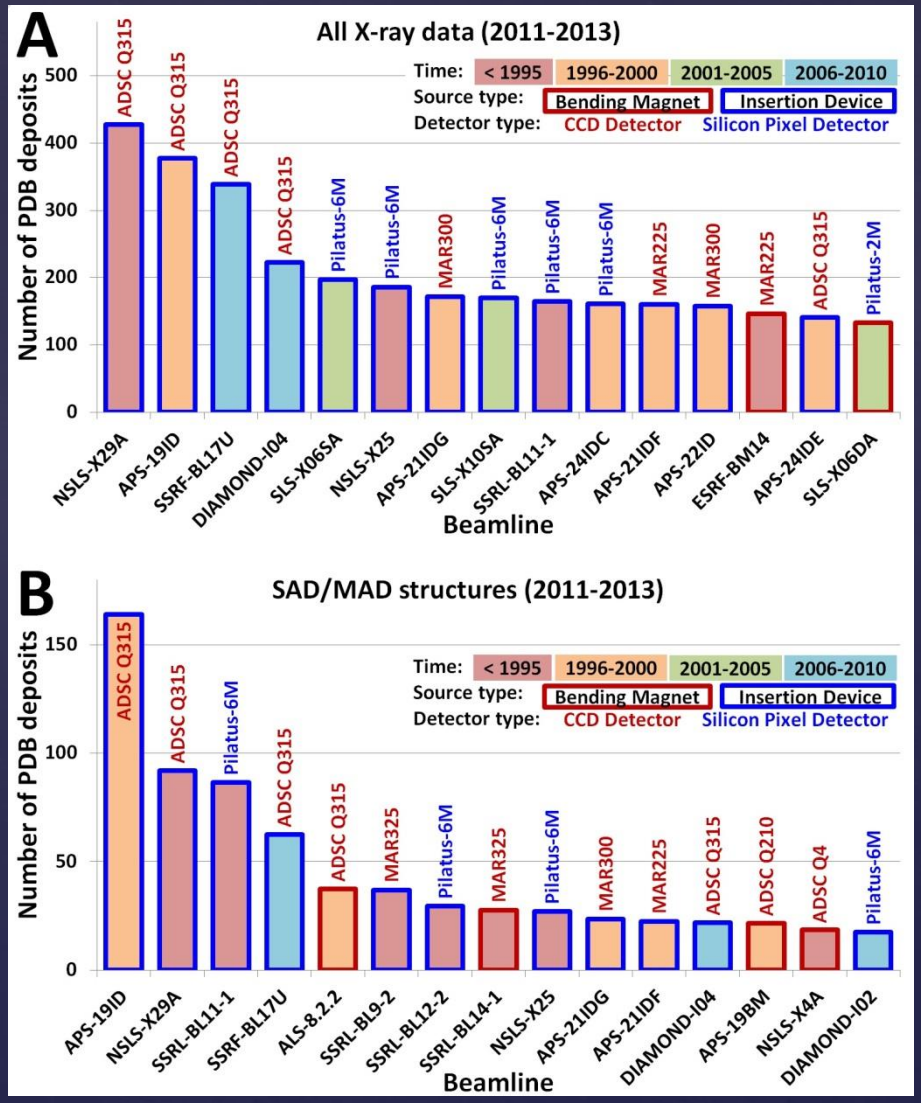
Average R_{free} by resolution bin (with a width of 0.2 Å for X-ray crystallography PDB structures deposited after January 1, 2001, divided into two groups by the number of missing data items ("NULLs") in the PDB file. The means for "high-completion" deposits (20 NULLs or less) are shown in blue, and the means for "low-completion" deposits (50 or more NULLs) are shown in red.

Where we should collect data ?



Zheng *et al*, *Expert Opinion on Drug Discovery(2014) 9: 125-37*

High Data-Rate?



Zheng *et al*, *Expert Opinion on Drug Discovery(2014) 9: 125-37*

Diffraction experiment - the last experiment before deposition to PDB

Dataset – 2minutes, sample change 2minutes -> 10minutes

6 datasets/hour -> 144 datasets/day

180 days -> 25920 datasets/day -> 2.5 PDB

125 synchrotron stations -> 324 PDB

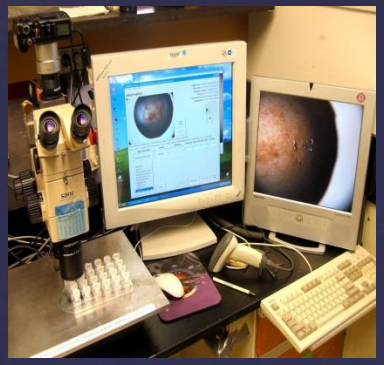
Efficiency -> 0.3%

HKL-3000 at SBC

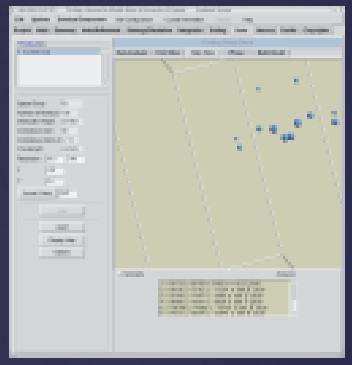
Database-controlled pipeline

lab e-book

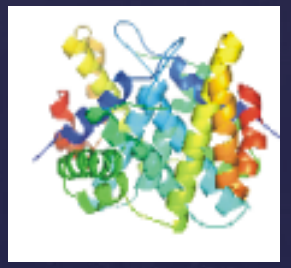
FPLC

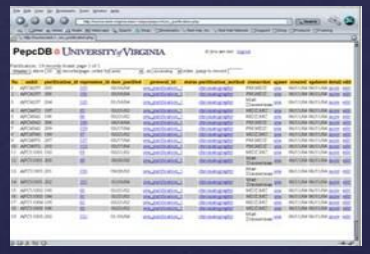


EQUIPMENT

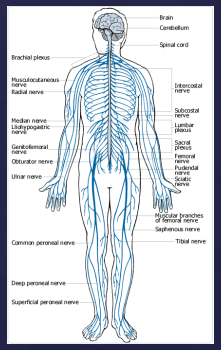


HKL-3000





DATABASES



Big brother?

Statistics / Progress in Minor Lab LIMS by researcher

Last week (17 Apr 2015 - 24 Apr 2015)

Person	Clones	Exprs	Purifs	Macro preps	Plates	Drops	Crystals	Datasets processed	Structure refs	Kinetic assays	Thermal shift assays
Cooper, David	0	0	0	0	0	0	23	18	0	0	0
Handing, Katarzyna	0	0	0	0	0	0	51	53	13	0	0
Hou, Jing	0	0	0	0	1	30	0	1	1	0	0
Kowiel, Marcin	0	0	0	0	0	0	1	8	3	0	0
Shabalin, Ivan	0	0	0	0	0	0	125	14	9	0	0
Shumilin, Igor	0	0	0	0	0	0	0	3	2	0	0
Szlachta, Karol	0	0	0	0	0	0	34	20	3	0	0

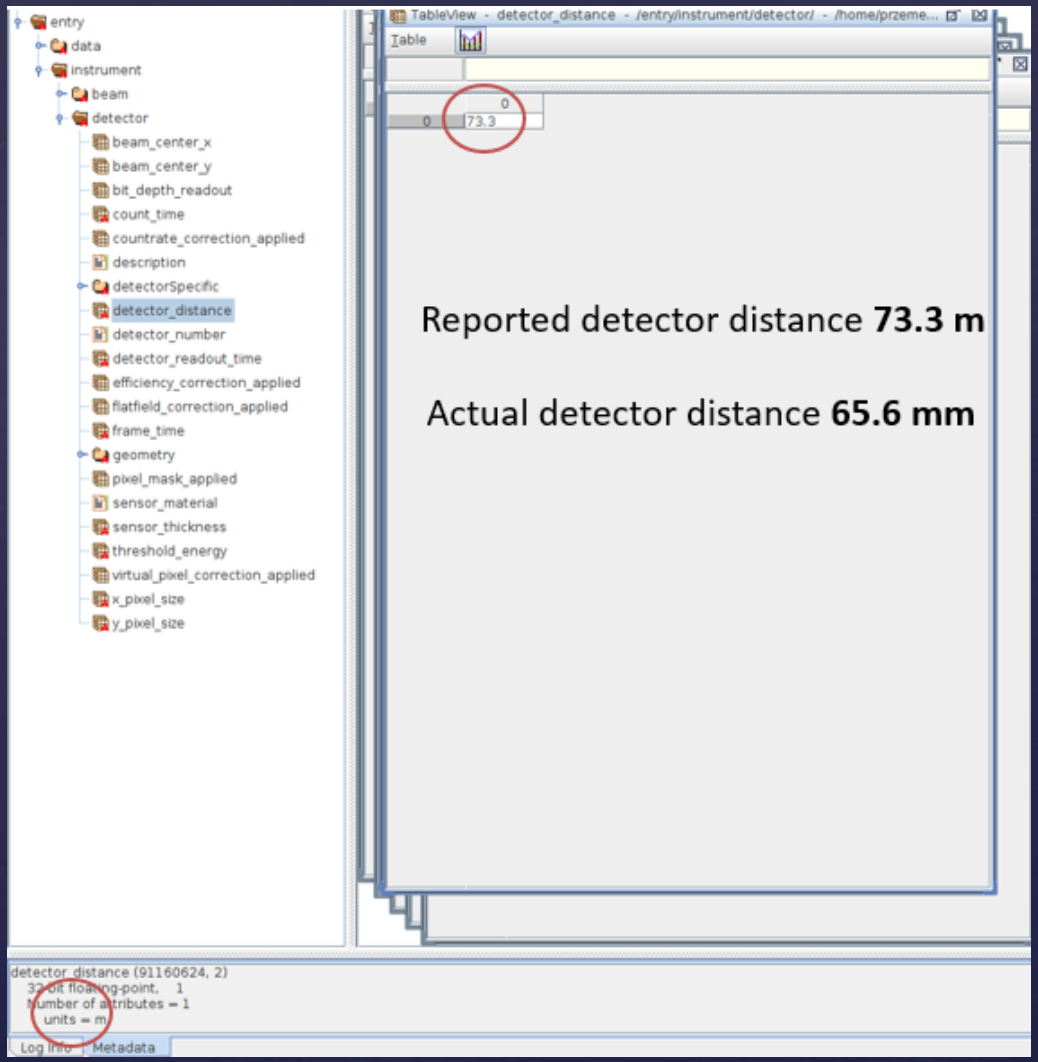
Last month (25 Mar 2015 - 24 Apr 2015)

HDF5 vs CBF
Structural Biologist Perspective

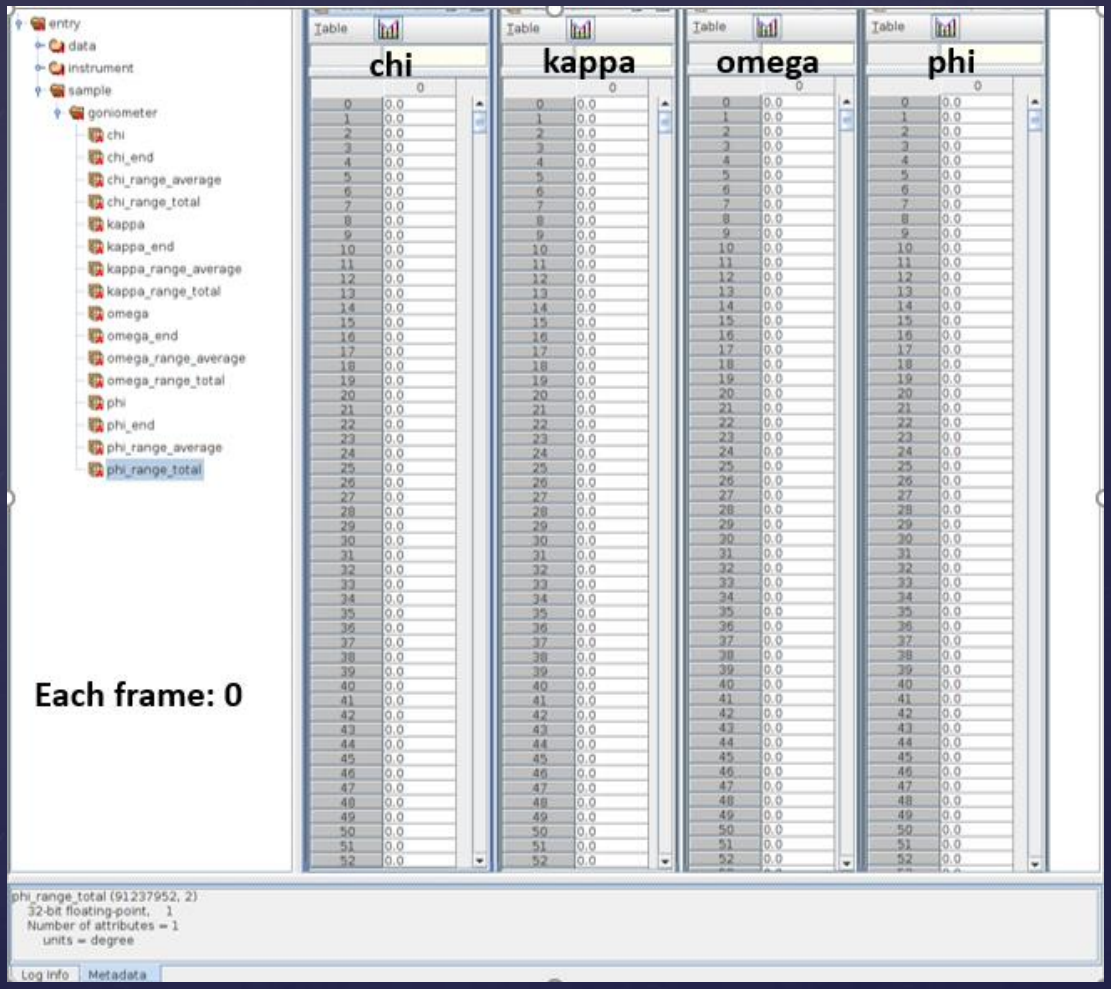
	Size	Processing time	Rmerge/Rmeas/Rpim
HDF5	3GB	21 min	0.054/0.057/0.018
CBF	3GB	21 min	0.054/0.057/0.017

Changing compiler suite **GCC** -> **Intel**
processing time **21 min** -> **4 min**

HDF5 vs CBF
Structural Biologist Perspective



HDF5 vs CBF
Structural Biologist Perspective



HDF5 *vs* CBF *vs* 321 other formats
Structural Biologist Perspective

New format new trivial inconveniences:

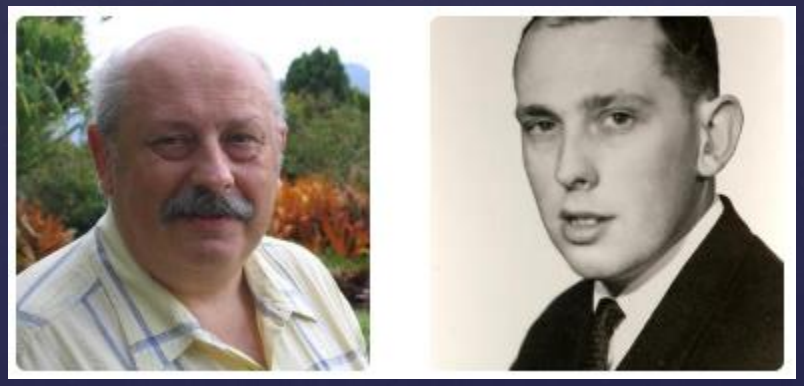
- File renaming
  - need to modify links in HDF5 master file
- New libraries and plugins
  - LZ4 compression plugin compatible only with newer versions of HDF5 library

What is special in Eiger and HD5 ?

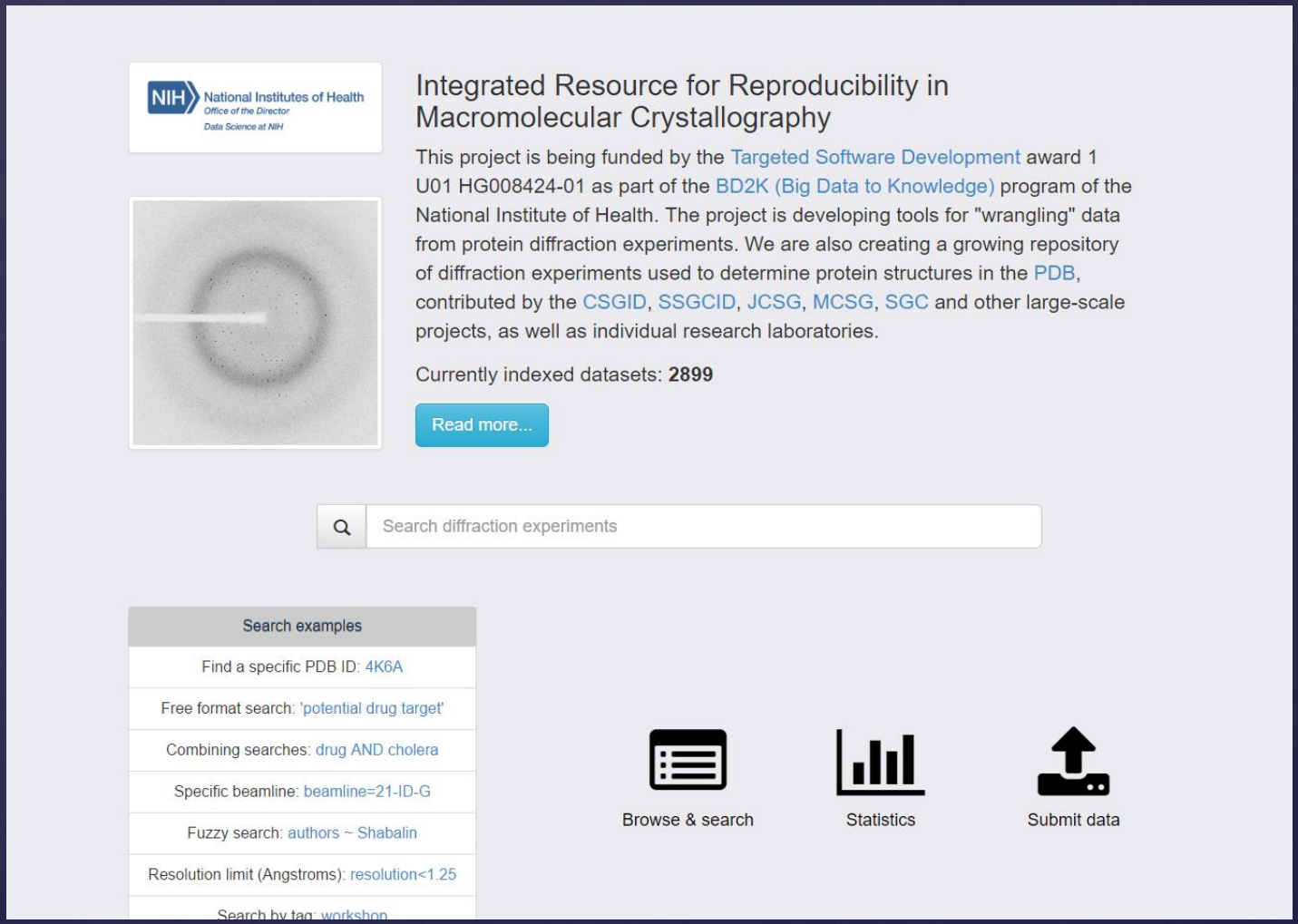
CBF – MAF
HDF5 – MAF

What is special in Eiger and HD5 ?

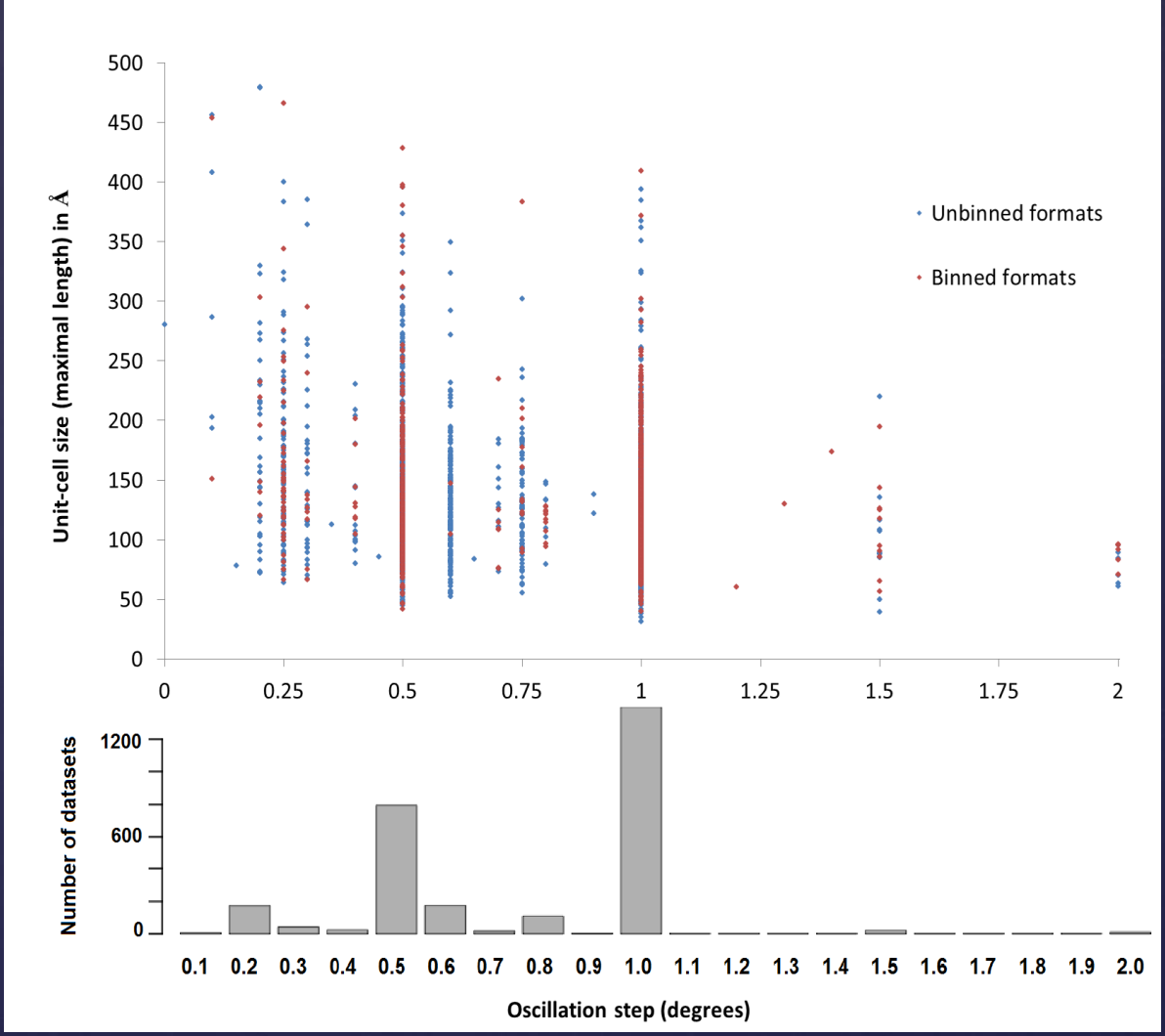
……. - MAH
CBF – MAF
HDF5 – MAF



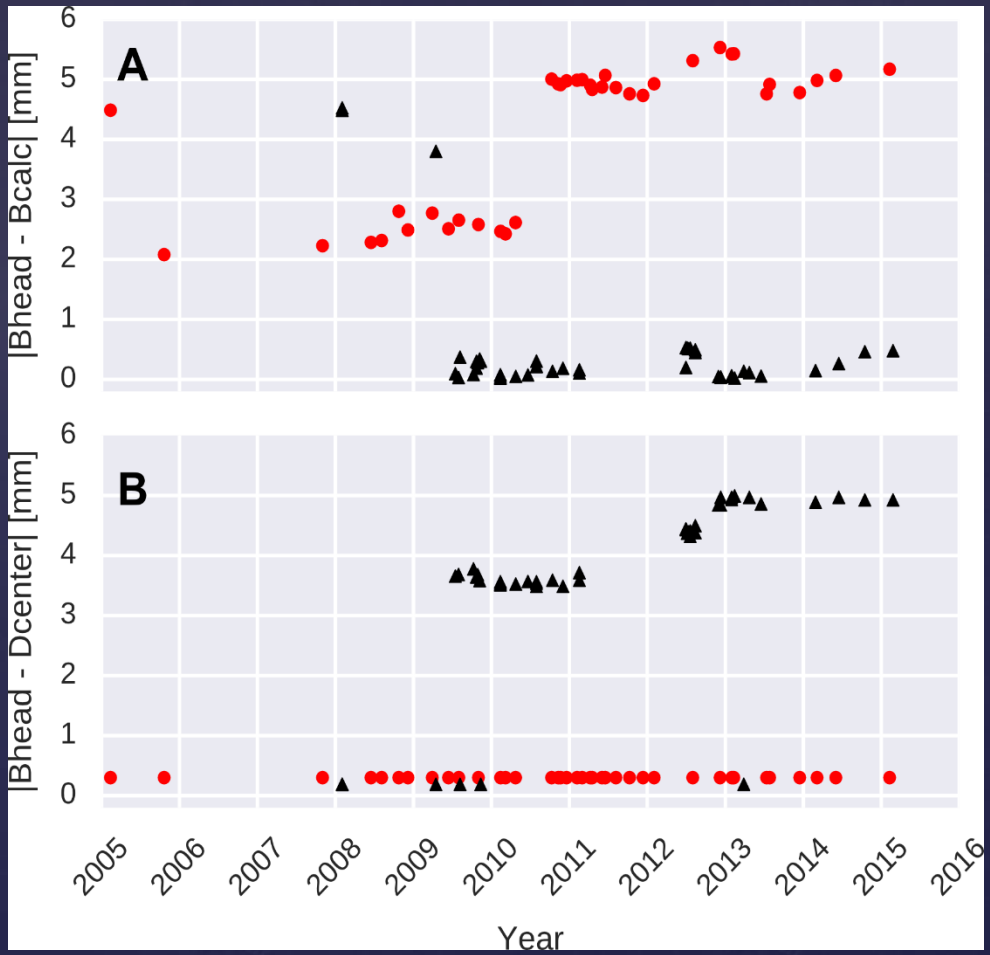
http://www.proteindiffraction.org



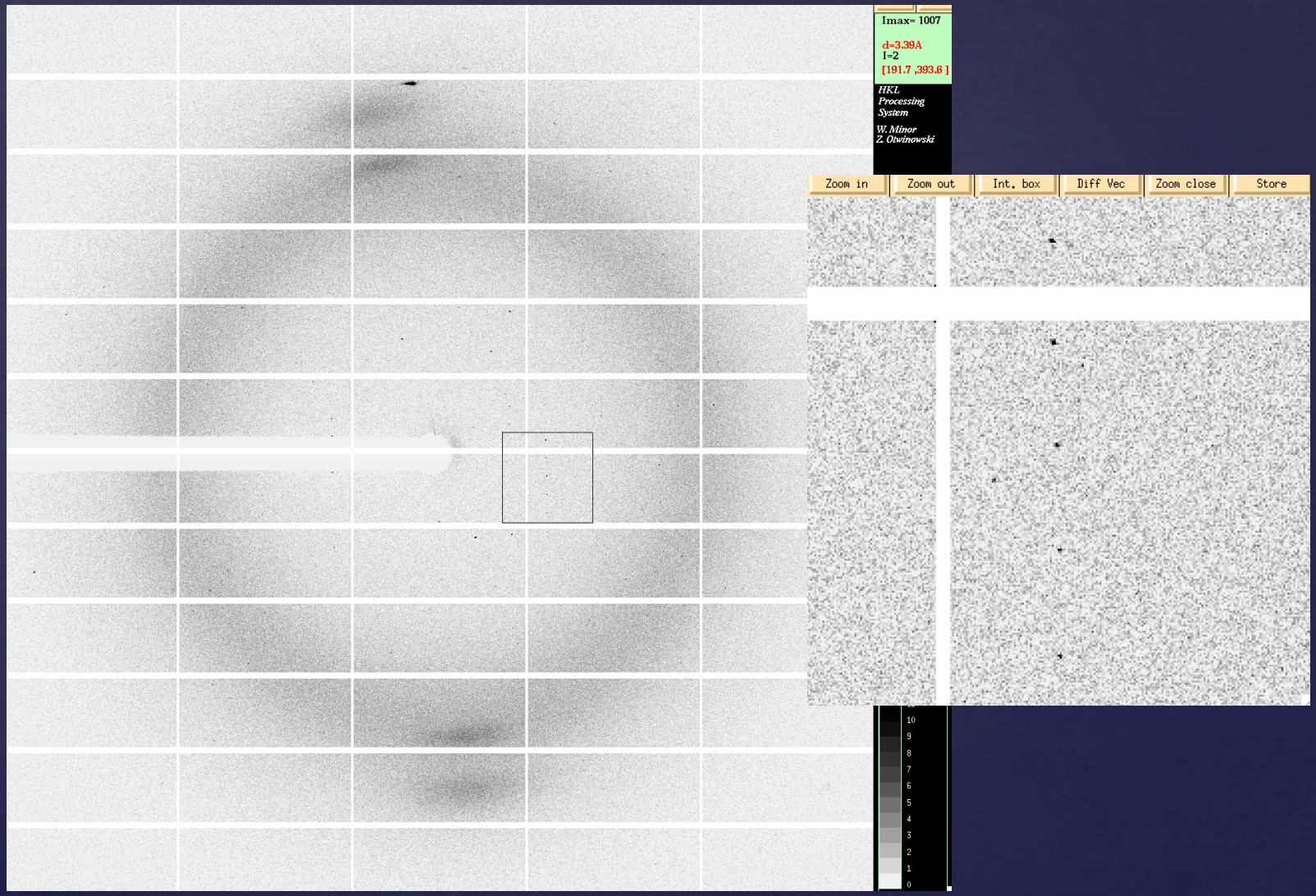
http://www.proteindiffraction.org



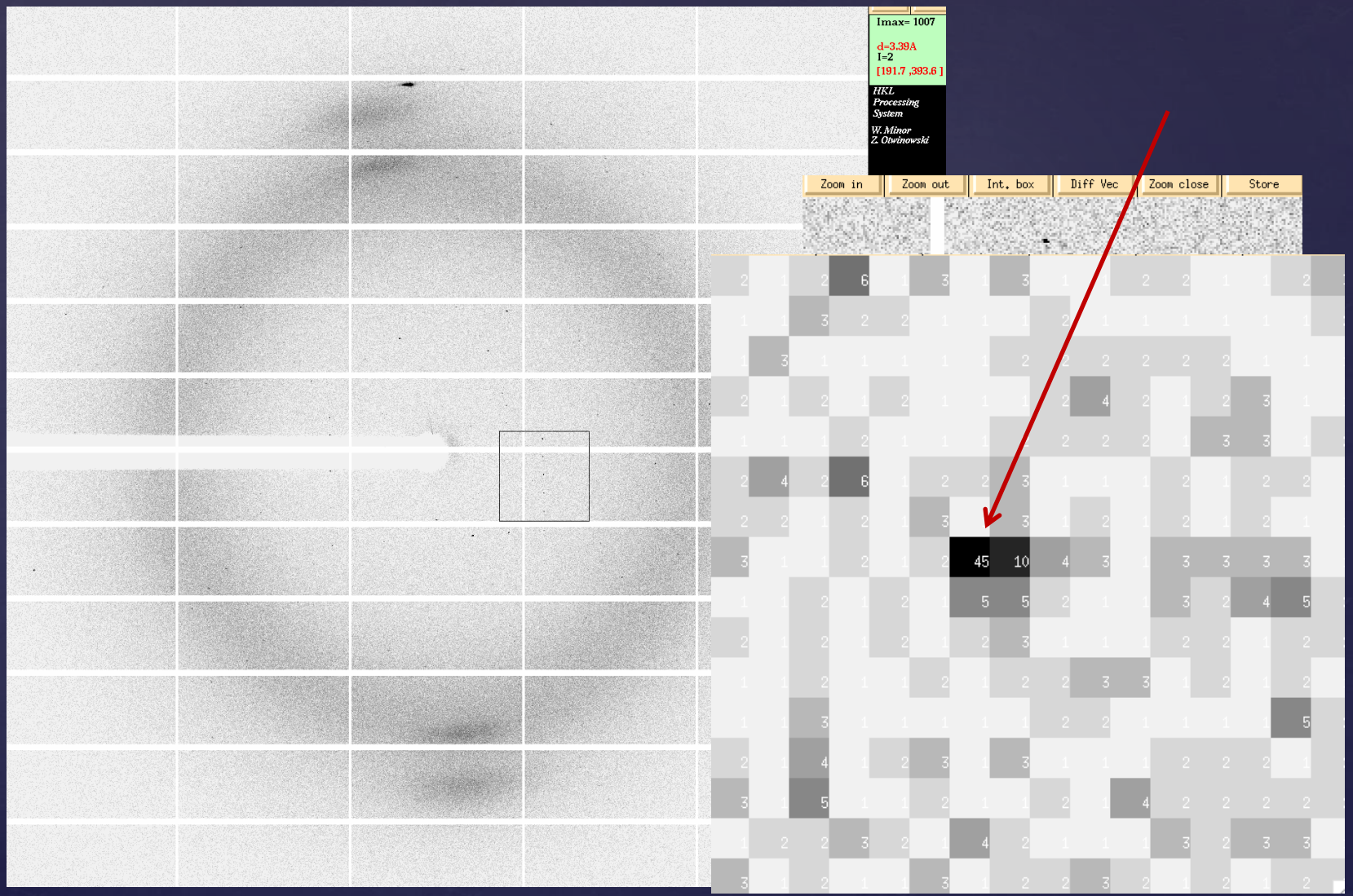
Beam position
Structural Biologist Perspective



Optimal data collection ?



Optimal data collection ?



Header – is CBF header a MAH ?

```
# 2015/May/06 10:30:40
# Pixel_size 172e-6 m x 172e-6 m
# Silicon sensor, thickness 0.001 m
# Oscillation_axis omega
# Excluded_pixels:  badpix_mask.tif
# Chi 0.0000 deg.
# Angle_increment 0.1000 deg.
# Polarization 0.99
# file_comments
# N_oscillations 2500
# Beam_xy (1223.03, 1256.56) pixels
# Exposure_time 0.020000 s
# Phi 0.0020 deg.
# Energy_range (0, 0) eV
# Start_angle 160.6000 deg.
# Detector_distance 0.617619 m
# Detector_Voffset 0.0000 m
# Alpha 0.0000 deg.
# Flat_field: (nil)
# Threshold_setting 7619 eV
# Exposure_period 0.020950 s
# N_excluded_pixels: = 321
# Kappa 0.0020 deg.
# Tau = 0 s
```

Do you like this image?

Do you like this image ?

How expensive is bright lens ?

See more choices

Canon EF 85mm f1.2L II USM Lens for Canon DSLR Cameras - Fixed
by Canon

$1,999.00 Prime
Get it by **Monday, Aug 24**

More Buying Choices
$1,999.00 new (22 offers)
$1,499.99 used (24 offers)

Trade-in eligible for an Amazon gift card

★★★★★ ▾ 159

See Style Options

Canon EF 85mm f/1.8 USM Medium Telephoto Lens for Canon SLR Cameras - Fixed
by Canon

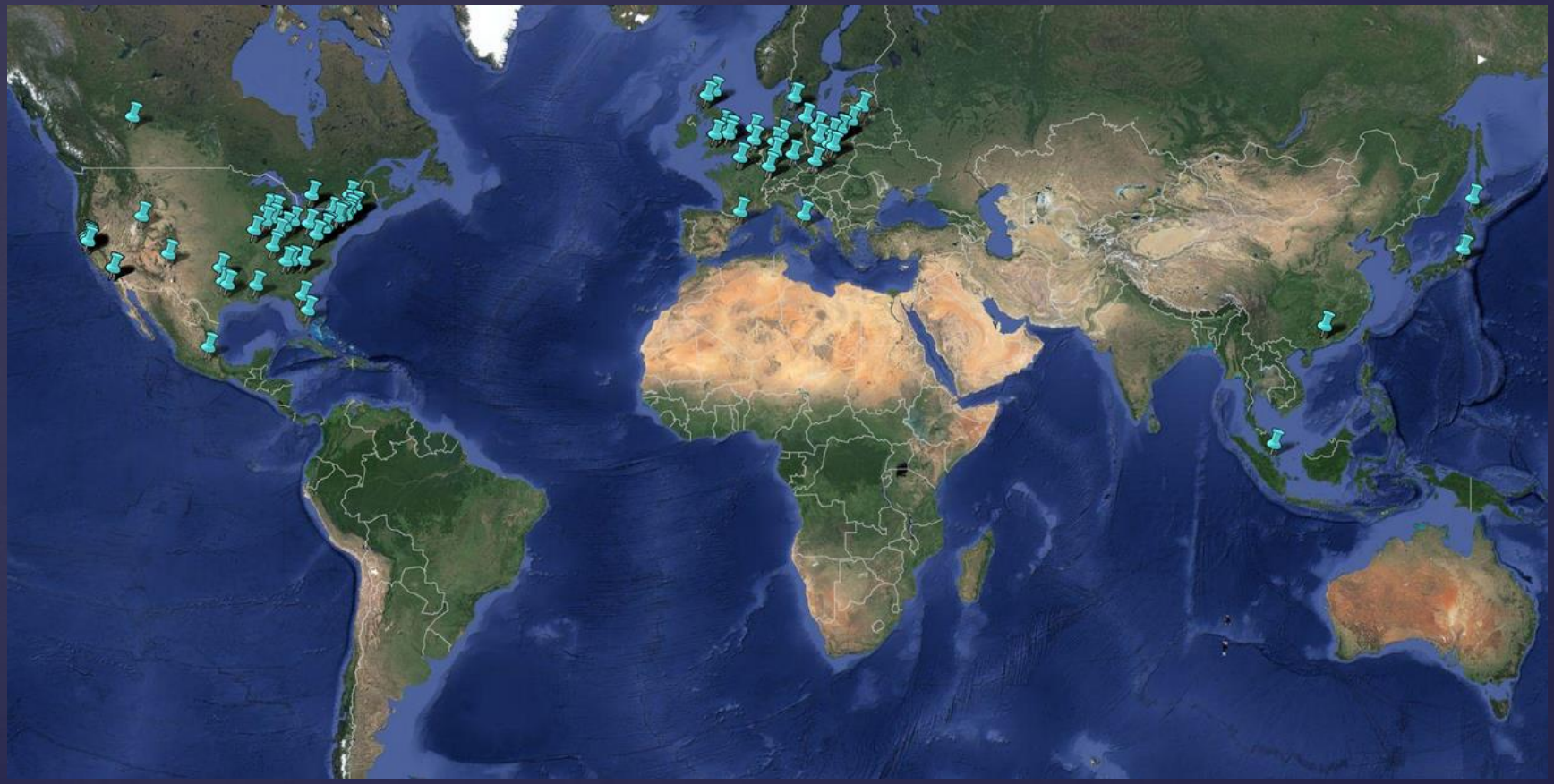
$369.00 Prime
Get it by **Monday, Aug 24**

More Buying Choices
$369.00 new (27 offers)
$298.00 used (26 offers)

Trade-in eligible for an Amazon gift card

★★★★★ ▾ 770

Collaborations documented by at least one paper



Acknowledgment

Acknowledgments

Wladek Minor
- Przemek Porebski
- Marcin Cymborowski
- Matt Zimmerman (CSIRC)
- Marek Grabowski
- Heping Zheng
- Karol Langner (Google)
- Piotr Sroka
- Ivan Shabalin
- Katherine Handing

Zbyszek Otwinowski
- Dominika Borek

Andrzej Joachimiak
MCSG and SBC staff

Wayne Anderson and CSGID staff

Steve Almo and NYSGRC Staff

Ian Wilson, Marc Elslinger and JCSG staff

Steven Burley, John Westbrook and PDB staff

Tom Terwilliger
Zbyszek Dauter

Grants:
U01-HG008424
NIH GM53163, GM62414, GM74942
GM093342, GM094585, GM094662
DOE, NCI
NIAID HHSN272200700058C
NIAID HHSN272201200026C
HKL Research. Inc.