SDS – Jonathan McDowell

Chandra Users' Committee, Apr 2006

Chandra Source Catalog

Repro 3

CIAO status

Testing

Data analysis issues

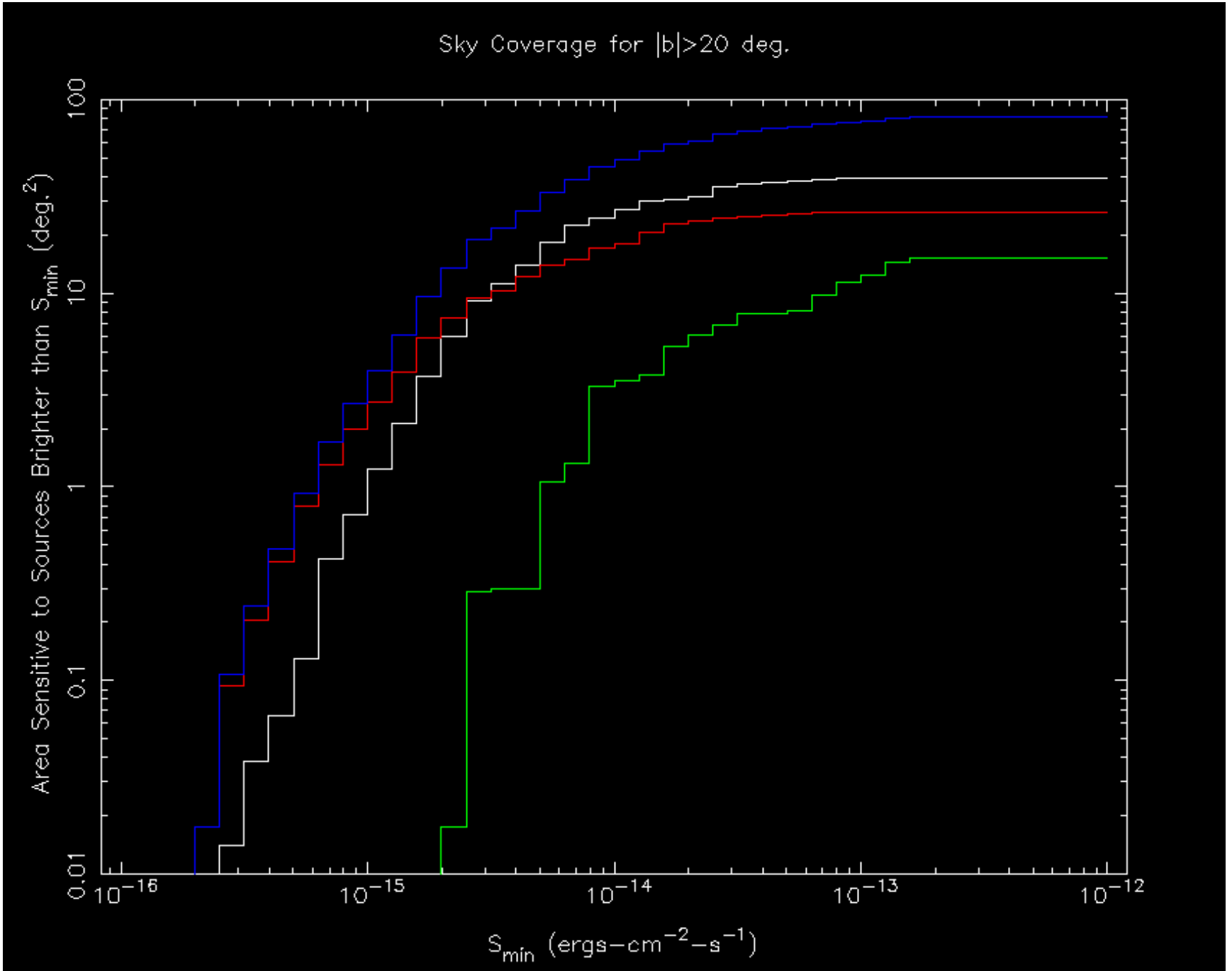
SDS

Chandra Source Catalog

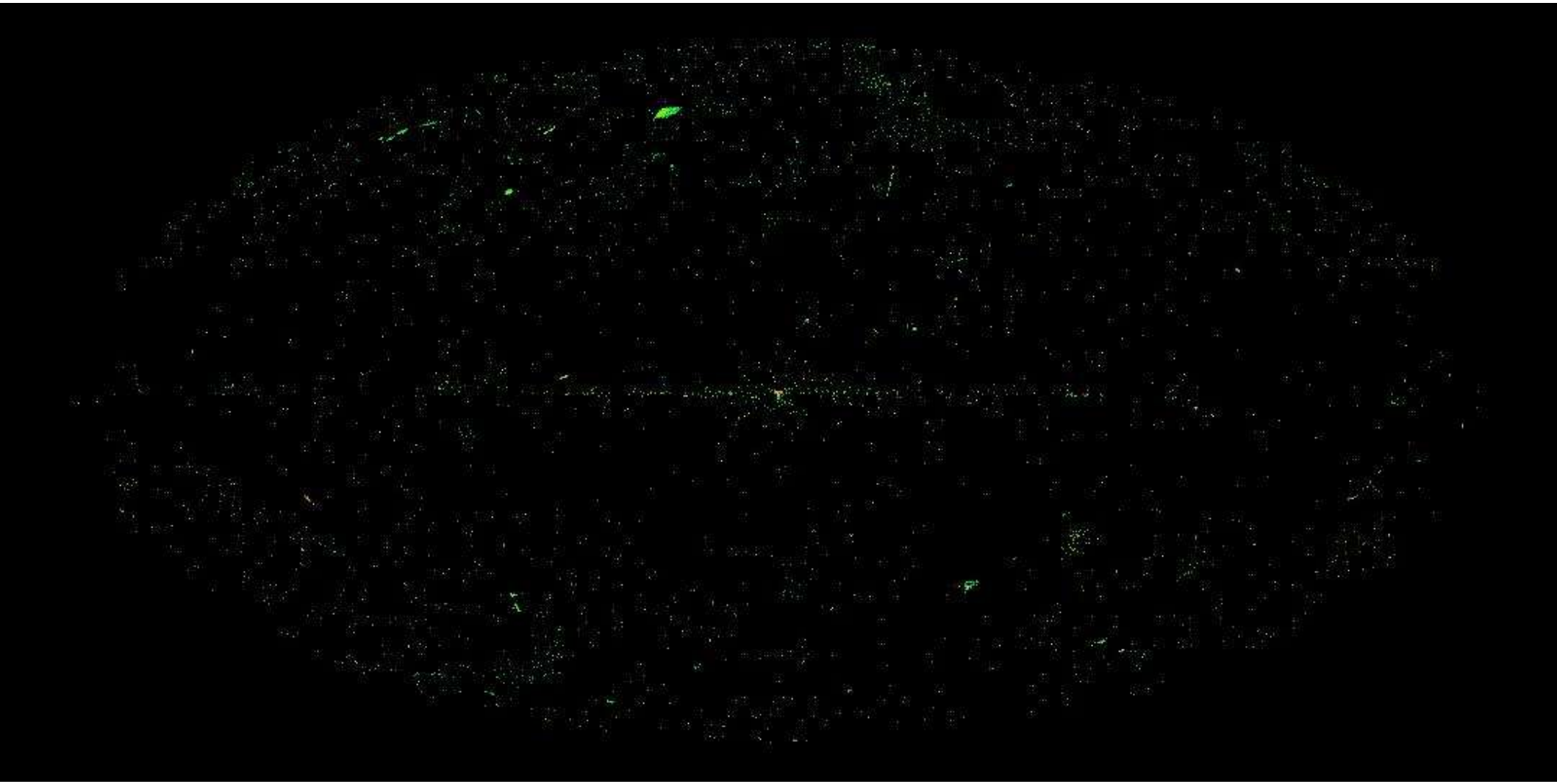
Chandra Source Catalog

- Goal: Catalog of Chandra sources for use in crossmatch with other catalogs, for analysis of X-ray source populations, and as all-sky X-ray astrometric catalog.
- Post-processing all Chandra imaging fields (ACIS and HRC); 160 sq deg by end 2005. Predict 400k sources by end of mission.
- Will handle mildly (1-30”) extended sources in first release
- Phased release, with later iterations doing a better job on extended sources.
- Science enabler for sample selection, prototype science studies, etc.

Sky Coverage: |b| > 20



Chandra fields 1999-2005 (galactic coordinates)



Chandra Source Catalog

- SDS and Data Systems working on project
- External review committee met Feb 8-10
- Endorsed goals of project but made significant recommendations for 'course correction'.
- New requirements document draft

Chandra Source Catalog

- Review committee gave a number of very positive comments
  - good general state of preparedness
  - important, exciting, timely project
  - blazing a path for other facilities
  - recognized key issues: content heterogeneity, reusing L1/L2 code, VO interoperability, phased delivery
- OK, that's great, but where do we need to improve? They gave us 11 key recommendations.
  - 6 URGENT ones, R1-R6, and 5 longer term ones, R7-R11

Review Panel Recommendations

- R1: A Requirements Document should be produced
  - First draft distributed
  - Not just a Reqs Doc, but a “project bible” describing what we are doing. Initial draft to capture current analysis only; update as we go. (Some sections still empty)
  - Current draft describes catalog contents and goals but is light on algorithms, which are currently documented elsewhere and will be incorporated in a later draft.

Review Panel Recommendations

- R2: Descope first release to support quick completion
  - Panel suggested possibilities: exclude some difficult kinds of regions, omit HRC, don't do fancy variability, crude UI
  - Our initial review makes it clear that the driving factor for a quick release is the scientist time needed for catalog characterization (and not coding time, processing time, etc.). We therefore expect to run the pipeline more or less as presently planned, but some outputs may not be included in catalog.
  - Basic outputs of source existence, position, flux considered critical – but getting them right implies getting a lot of other things right too (e.g. exposure) – requires interdependency analysis to see what can be tossed without impact.
  - Other things e.g. HRC are independent but don't need much characterization time so we may not gain much schedule by dropping them.

Review Panel Recommendations

- R3: Aim at multi-wavelength astronomer
  - Panel emphasized catalog should be targeted at general multi-wavelength astronomer as the most important customer rather than the X-ray expert astronomer
  - Implication: basic source catalog tables are the most important product (vs data objects)
  - Implication: worth doing energy flux (erg /cm2/s/keV) and not just number flux event though the latter is better constrained
  - Generated a new set of use cases focusing on general astronomer use of the catalog; deriving requirements on catalog and UI.

Review Panel Recommendations

- R4: Distinguish between database and catalog
- R5: Run pipeline to faint limit
- Panel directed us to add an extra stage of catalog definition, involving filtering, merging and quality assurance. There will then be a 'database', containing all the latest pipeline results, and a 'catalog', which is both a subset and a snapshot in time, containing a well characterised product. Both database and catalog consist of a master source table, per-observation source table, and data objects such as PHA files. The difference is that the catalog has controlled (versioned) releases and has a subset of the sources and a subset of the table columns and data products whose characterization we have a higher level of confidence in.
- This allows us to run the pipeline to a deep threshold (limited by per-source computing resources needed) while using a more conservative threshold for the catalog.

Review Panel Recommendations

- R6: Scope UI soon
- The panel felt that our UI plans were both vague and overambitious.
- The UI can drive some aspects of the back-end functionality.
- We should “complete a very simple first UI design as soon as possible”.

Outline UI requirements

- Minimum requirements
  - Web based interface (no download of application required)
  - Access to all fields in master and per-observation source catalogs
  - Support cone-search type (location crossmatch) interface
  - Support SQL-based interface implementing a subset of ADQL
  - Include ability to upload lists of target positions/errors to search
  - Interface will have links to L3 data objects
  - User able to select fields (columns) to be returned, and constrain number of rows to return
  - Return sorted sources with top N values of query
  - Return results in plain text, HTML

Baseline UI requirements

- Highly desirable:
  - Access to upper limit/sensitivity data
  - Link between sources and full field images
  - Name resolver in query interface
  - Support VOTABLE output
  - Virtual column definitions (query on functions of columns)

Baseline UI requirements

- Longer term requirements:
  - Full ADQL implementation
  - Integrate functionality with NED, SIMBAD, DataScope
  - Integrated link to Vizier and USNO-B (or successor)
  - Link to Chandra observation catalog for proposal info
  - Ability to query previous editions of catalog
  - Ability to query underlying database directly
  - Return flux in user-defined band (uses event or pha data)
  - ADQL equation scripting
  - User API (e.g. web service) access
  - Links to VOPLOT and other VO applications

Review Panel Recommendations

- R7: Investigating External Solutions
- Panel drew attention to ACIS Extract, 1XMM and XASSIST, and felt that we had not sufficiently described how we had looked at these solutions, and why we had or had not adopted their approaches.
- In fact the Panel's phrasing was stronger: “team were often not aware of, or seemed to have ignored, existing solutions...”. We believe this criticism is unfair, as we have indeed reviewed the three main approaches they cite, and they have influenced our design. It may be true that our rationale for not adopting some approaches needs to be revisited.

Review Panel Recommendations

- R8: Quality Assurance Plan
- The panel believed that fully automated quality assurance is not workable, and that we should plan manual spot checks.
- They also recommended we clearly separate a catalog production and quality assurance step from the pipeline production of the database, which could be rerun as needed independently of the database pipeline.
- We did not describe our plans in sufficient detail in the presentation, but we agree with this recommendation and it is essentially in line with our existing plans. The separation of the merge/filter/QA step has only a minor impact on our development.

Review Panel Recommendations

- R9: Extended Emission – panel agreed this can wait till later release, emphasized R&D work needed soon. We have an ongoing effort on this in SDS in the CIAO context.
- R10: Merging observations: For later releases, the panel emphasized the importance of running detect on merged observations of the same field, not just merging source lists from separate detect runs. Again, SDS needs to figure this out for normal CIAO users anyway. Panel also asked for full-field background-corrected smoothed images.
- R11: Avoiding low priority issues. For example, we shouldn't waste time worrying about pileup since it only affects a small fraction of sources. Point taken – although in many of these cases we're just taking for free the hard-won expertise from supporting general user data analysis.

Other Panel Recommendations

- Use cases too complicated and VO-oriented. We have begun working the new use case list.
- Choice of energy bands. We accept the suggestion to separate detect bands from color measurement bands. It was also suggested that the source finding bands be reviewed – our new simulations show that the detect results are not sensitive to the exact energy boundaries.
- Other recommendations: our responses are in the formal response document to be completed shortly.

Revised Schedule

- *Subject to project constraints, e.g. spacecraft support needs*
- 2006 Q1
  - Complete prototype per-observation pipeline definition (DONE)
  - First draft requirements doc (STARTED)
  - Review Committee (DONE)
  - Response to Review Committee Recs. (NEAR DONE)
  - Define use cases and begin flowing requirements (DONE)
  - Begin characterization plan (STARTED)
  - Begin UI definition/design (STARTED)
  - Complete prototype per-observation pipeline implementation (DONE)

Revised Schedule

- 2006 Q2
  - Complete data archive ingest/retrieve definition (MOSTLY COMPLETE)
  - Complete merge pipeline definition (STARTED)
  - Start per-observation science evaluation testing (STARTED)
  - Start baseline catalog characterization (NOT STARTED)
  - Complete pipeline/archive ingest/retrieve implementation (MOSTLY COMPLETE)
  - Complete catalog ingest/retrieve definition (STARTED)
  - Revise prototype pipeline based on Review Committee recommendations (NOT STARTED)
  - The above items (except the last) delayed from Q1 due to review committee preparations and response.

Revised Schedule

- 2006 Q2-Q3 (under revision)
  - Complete per-obs pipeline science eval testing, pipeline revision
  - Complete baseline catalog characterization
  - Complete baseline UI definition
- 2006 Q3-Q4
  - Merge/QA pipeline science evaluation testing and revision
- 2006 Q4-2007 Q1
  - Integration and test. initial production run
- 2007 Q1
  - Operational catalog characterization, initial UI release
  - Tweak production system; revised production run if needed
- 2007 Q2 - First catalog public release

SDS

CIAO STATUS

CIAO STATUS

- CIAO 3.3 release Nov 2005 – New user tools
  - reproject_aspect, reproject_image, reproject_image_grid
  - specextract as supplement for psextract
  - data cube support in DM tools, region area bug fixes
- CIAO 3.3.0.1 release Jan 2006 – new PIMMS file for proposal
- CALDB 3.2.0 release Nov 2005
  - Improved ACIS CTI, TGAIN, P2RESP; ACIS bad pixels; HRC-S gaps, HRC-I gain
- CALDB 3.2.1 release Dec 2005
  - New HRMA area, HRC-S QE, HETG efficiency

Download Statistics

- CIAO 3.3 released Nov 15
- 444 downloads
  - 317 Linux, 103 MacOSX, 24 Solaris
  - Includes 21 downloads marked as '10 or more users'

Forthcoming

- CIAO 4.0 now scheduled for late 2006 with extended testing phase
  - Sherpa 2 and ChIPS 2; currently testing initial code drops
  - New architecture includes internal use of Python; possibility of user interface in Python under evaluation
  - Working on support for S1/S3 CTI correction, dead area correction, better grating order separation files. May trigger CIAO 3.3.1 release

Forthcoming

- SAOSAC release plan
  - CXC Optics team continuing work on portable version; 40 out of 55 packages ported with testing on Sparc, Linux 32-bit, Linux 64-bit.
  - Behind schedule; differences in hardware floating point implementations require algorithm changes for better numerical stability; licensing issue identified
  - SDS prototyped Slang scripts to run SAOSAC and psf_project_ray to make images, radial profiles in prototype form; will add MARX.
- R&D: Merging observations; modelling ACIS background

SDS

Data analysis issues

Testing CIAO

- CIAO is a big system
- Tools, Sherpa, ChIPS/UI, DataModel, Configuration:
  - Algorithm development, spec, design, development, maintenance, test, portability, documentation, OTS integration
  - 10.5 FTEs in Data Systems and 6 FTEs in SDS not counting parts of pipeline not in CIAO or proposal support work
  - Unit testing by DS
  - Science unit testing and thread testing by SDS
- 860 k lines of code: mostly C, C++; some Fortran, Perl, Slang, and XML help files

CIAO Resources

- Detail of FTEs
  - Tools 4 DS, 1.5 SDS
  - Sherpa 3 DS, 1 SDS
  - ChIPS/UI 2 DS, 0.5 SDS
  - DM+Config 1.5 DS, 0.5 SDS
  - General doc and test – 2.5 SDS
  - SDS Test Lead: Margarita Karovska
  - SDS Doc Lead: Antonella Fruscione

Testing CIAO

- Stage 1: Unit Tests
  - Developers and scientists run unit tests on new and modified tools
  - Scientists run tools in science threads
  - Scientists report via test worksheets with pointers to example data; incorporated into automated test scripts used for portability and later regression testing
  - SDS test lead coordinates inputs

Testing CIAO

- Stage 2: Mini-Test
  - Specialized regression test
  - Selected key CIAO tools
  - All new or modified tools
  - Selected tools to test new library functionality
  - Test out high risk areas, ensure stability during preparations for a release

Testing CIAO

- Stage 3: Full test
  - Regression test for all tools in system
  - Add new tests for each release via input from worksheets
  - Run on all portability platforms
  - SDS test lead signs off on results
- Stage 4: Package testing
  - SDS verifies download tar files on each platform: 'smoke test' confirms that as-packaged system does run.
  - Validate web links, install instructions, tar files
  - SDS/DS go for release; ECR reviewed by CXC senior staff

Limitations

- There is always more to test!
- Example: 85 CIAO tools in CIAO 3.3 (plus Sherpa,ChIPS, scripts);
  - each tool has many parameters
  - obviously not practical to test all possible paths through the code
  - we do a suite of tests attempting to sample likely user cases and parameter values, but our resources don't allow us to test all the cases that will be encountered.
- Some areas have given us particular trouble: e.g. Sherpa, region areas; the region bugs are mostly fixed and Sherpa is getting a rewrite.

Helpdesk

- Since mid-October:
  - 161 help desk tickets
  - 78 were to do with CIAO
  - Most were usage issues resolved by email
  - 3 were identified as bugs (none new in CIAO3.3), now fixed
  - 3 requests for enhancement (e.g. ARFs in chip gaps)
  - 3 under investigation (e.g. dmgti on light curves: bug or doc workaround?)

SDS

Repro 3 status

Repro 3 status

- Repro 1: Last full reprocessing in 2001
- Repro 2: HRC-only, in 2002
- Repro 3 production started February 15
- Reprocessing 2005 first, then work backwards
- Done Jan-Oct 2005 so far
- Community alerted via email bulletin
- New 'how does this affect my processing' web page goes up this week
- Estimate completion in early 2007

How does this affect me?

- Post Repro3, downloaded evt2 files have the latest best processing (for now!).
- When you download archival data, it's usually a good idea to recalibrate it (via acis_process_events etc) with the latest version of CIAO/CALDB – not a big overhead
- Not mandatory – web pages give details on which cal changes matter for which data
- If you've done this to your data in the past 1-2 years, you probably don't need to worry
- If you are still working on data that hasn't been reprocessed since early in the mission, you should redo it

An improved archive - ACIS

- Less area set bad around node boundaries (Nov 2005)
- Time dependent gain for S0,S4,S5 (Jun 2005)
- CTI correction for S0,S4,S5 (Jun 2005)
- CTI-corrected gain for I0-3, S1-3 (Dec 2004)
- Improved destreak for chip S4 (Nov 2005)
- Improved geometry files (Feb 2005) (small HETG wavelength corrections)
- Improved ACIS-S fid lights (Dec 2003)

An improved archive - HRC

- Improved gap removal (Jan 2001, in Repro 2)
- Ghost image removal (Mar 2001, in Repro 2)
- Timing mode correction (2003)
- Tap ringing update (2004)
- Gain correction map (Nov 2005)
- Improved HRC-S degap (giving better LETG wavelengths) (Nov 2005)

SDS

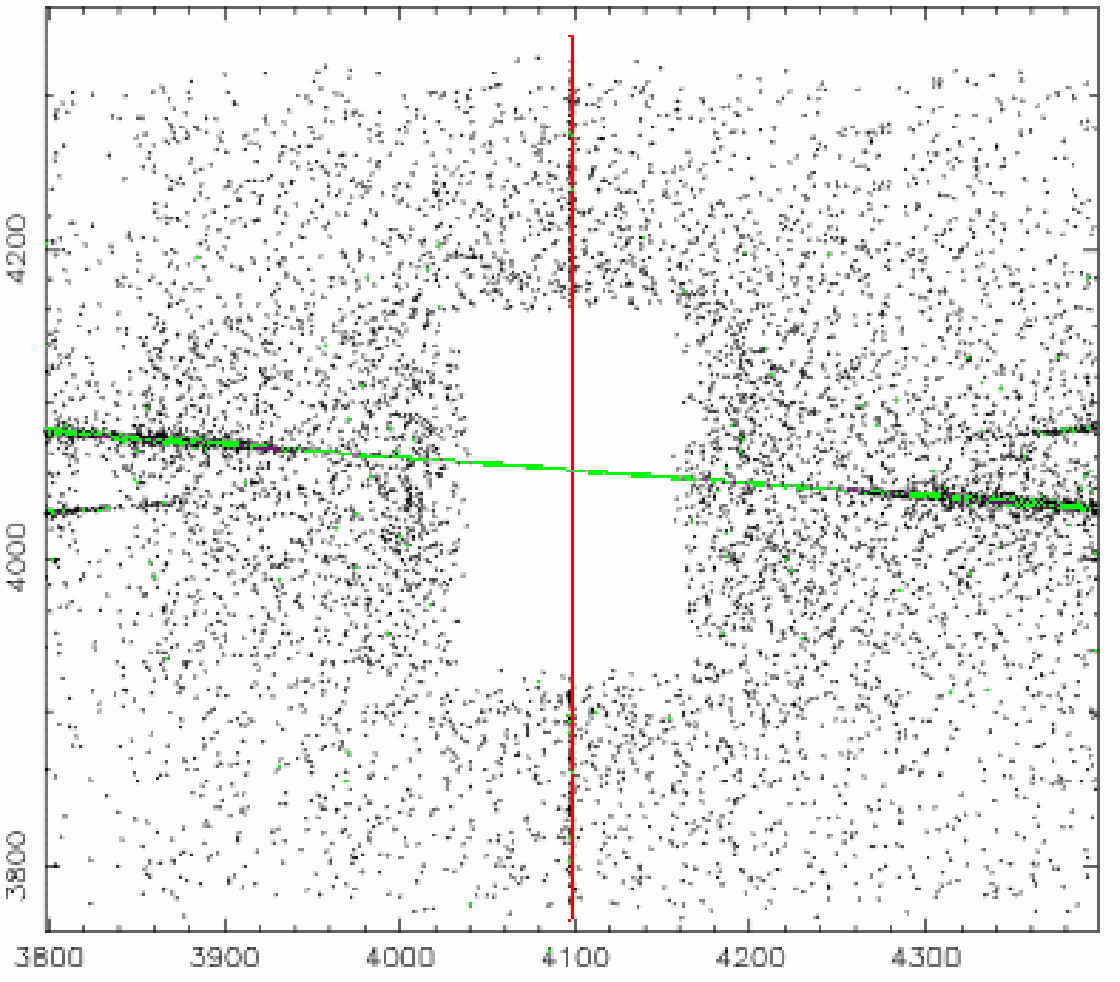
Data analysis issues

Specific Issues: Zero Order

- Grating data with ZO image piled or blocked
  - Pipeline doesn't find ZO location – wavelengths wrong
- Were users alerted to problem? Yes:
  - Analysis Guide for Chandra High Resolution Spectroscopy
    - http://space.mit.edu/ASC/analysis/AGfCHRS/AGfCHRS.html#nozo
    - "Cases requiring Customized Processing"
  - Thread "Correct Zero Order Source Position"
    - http://cxc.harvard.edu/ciao/threads/tgdetect
- But docs to deal with it were inadequate; new threads added
    - http://cxc.harvard.edu/ciao3.3/threads/tg_piled_zero/
    - http://cxc.harvard.edu/ciao3.3/threads/tg_blocked_zero
- New algorithm developed – prototype, will evaluate for pipeline

New ZO algorithm

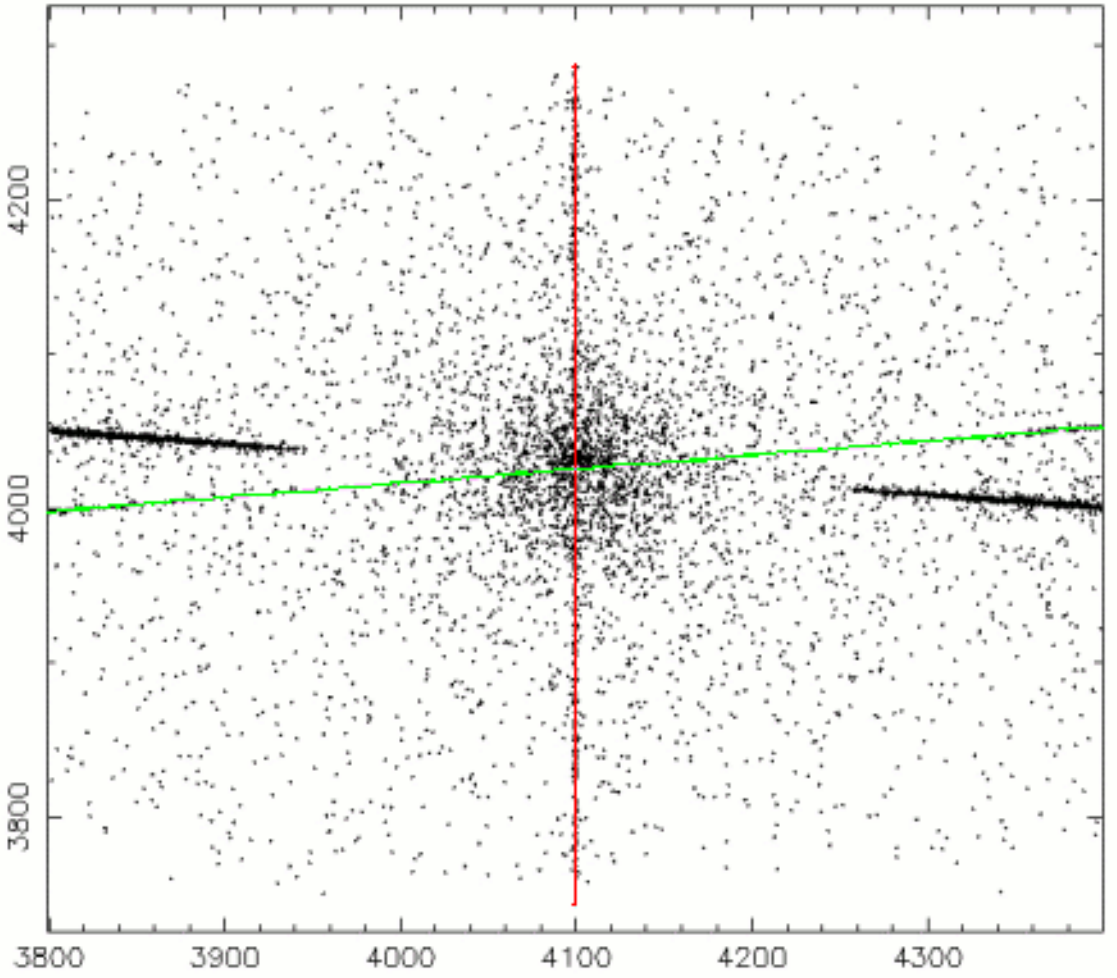
```
slsh> (x,y)=findzo("acisf00660_002N003_evt2.fits","m");
```



Handles zero order blocks, and …

New ZO algorithm

```
slsh> (x,y)=findzo("acisf00660_002N003_evt2.fits[EXPNO=1:100000]","m");
```



… and piled sources.

Specific Issues: Combining Spectra

- Data sliced into multiple observations (now common)
- How to combine the extracted PHA spectra and responses?
- The acisspec script only handles cases with very similar responses.
- Many users use FTOOLS addspec but don't read the dire warnings in its help file about how it can lead to wrong results. We plan to enhance specextract but must carefully address the ways in which you can get the wrong answer.