Science Considerations for Data Models

Some lessons from the Chandra Source Catalog

Dr. Ian N. Evans

Chandra X-ray Center Data System End-to-End Scientist
Chandra Source Catalog Project Scientist

Center for Astrophysics | Harvard & Smithsonian

IVOA Interop
2021 May 27

Scientifically complex datasets I

- The era of computationally intensive data analyses
  - Over the next decade data providers are going to have to build and serve *scientifically complex datasets* that incorporate increasingly sophisticated and robust scientific analyses

- Datasets that require algorithms that are too complex *or* too computationally expensive (*or* both) for individual researchers to easily perform bulk data analyses

- Such datasets MUST be scientifically rigorous and provide to the end user ALL the details needed to fully understand the data

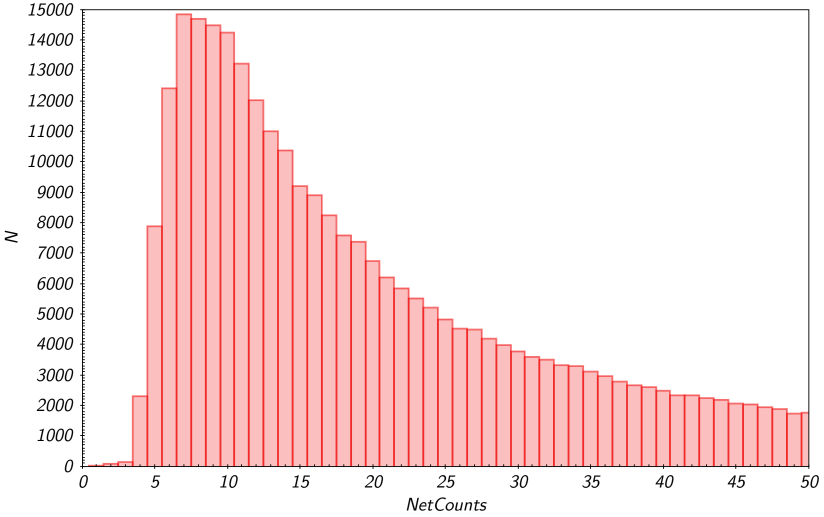
Scientifically complex datasets II

- X-ray astronomy is there now
  - *Ex.: Chandra Source Catalog* rel. 2.0 required ~600 CPU years to process
    - ~317K X-ray sources, ~928K detections (~1.42M w/photometric upper limits), 3 main + 6 ancillary tables totaling 1687 data columns, 38 types of FITS data products (~36TB total)
    - Precision astrometry, matching detections, source extent, multi-band photometry, spectral fits, temporal variability, …

A quick X-ray astronomy primer

- X-ray astronomy instruments typically detect individual X-ray photons
  - 4-dimensional very sparse data cube (x_{raw}, y_{raw}, t_{obs}, PHA) of photon *events*

Raw detected position	(x_{raw}, y_{raw})	→	Photon sky position	(α, δ)
Raw time of arrival	t_{obs}	→	Photon time of arrival	t_{TT}
Raw pulse height	PHA	→	Photon energy	E



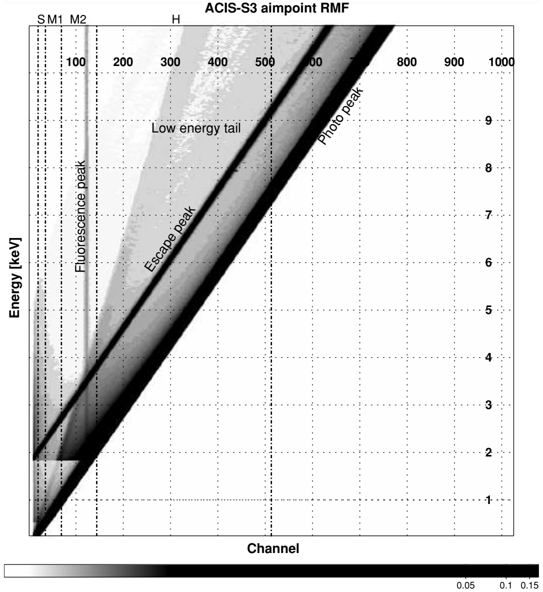
CSC 2.0 ACIS detections

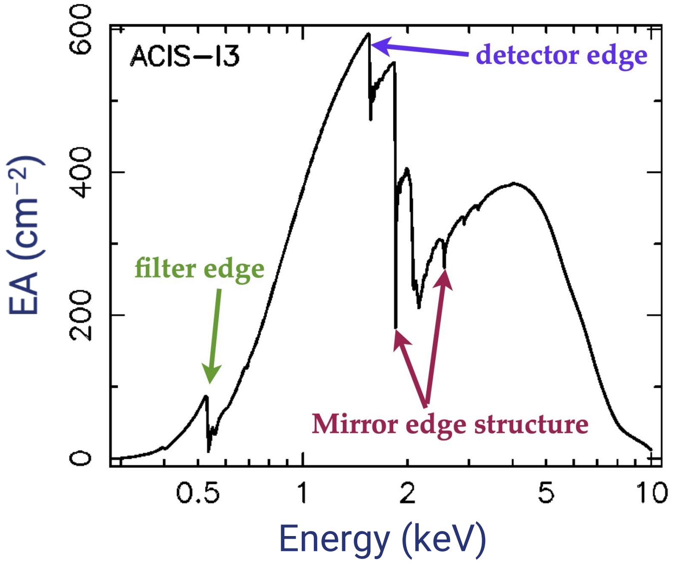
Sources with as few as ~4–5 net counts can be detected reliably with low-background instruments such as those present on *Chandra*

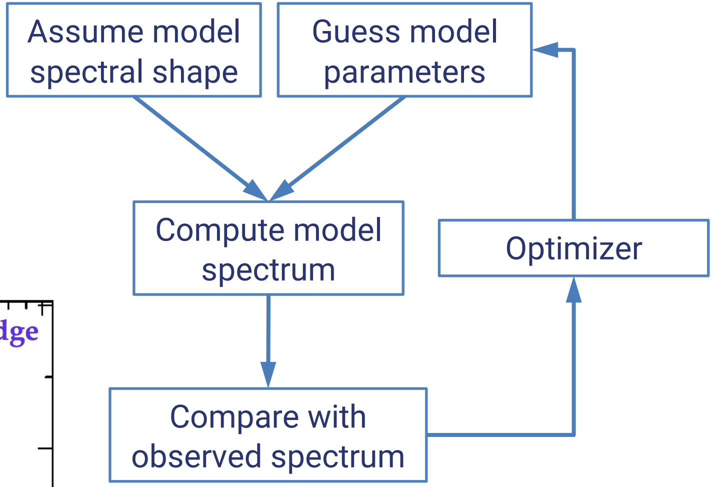
CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

X-ray photometry I

- The mapping from event PHA to photon energy is complex because it *depends on the source spectrum* which you are trying to determine





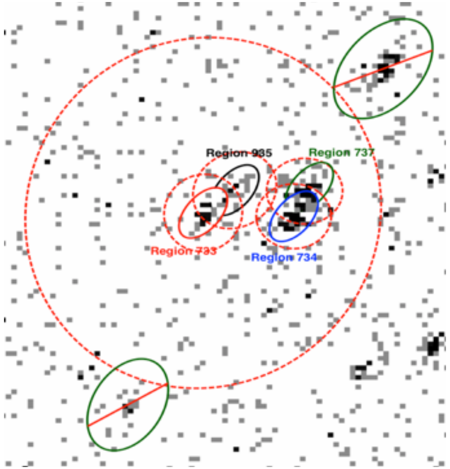


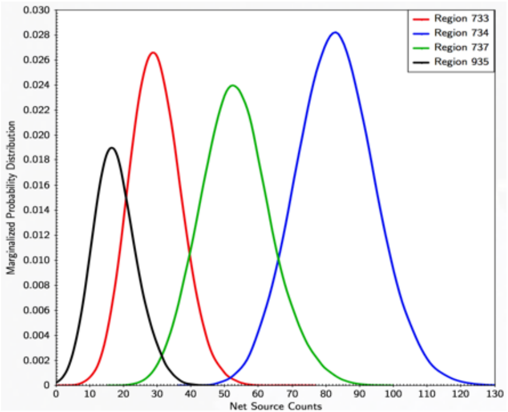
Forward Fitting

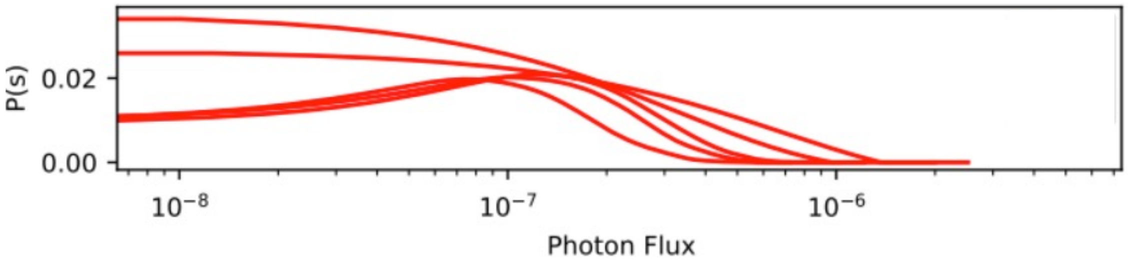
CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

X-ray photometry II

- CSC uses Bayesian X-ray aperture photometry approach (Primini & Kashyap 2014 ApJ 796, 24)
  - Multiple detections/overlapping apertures are solved for simultaneously
  - Joint posterior probability density functions (PDFs) for source and background fluxes in an *n*-source bundle are computed
  - Posteriors are optimized and sampled using MCMC







Courtesy Frank Primini (CfA)

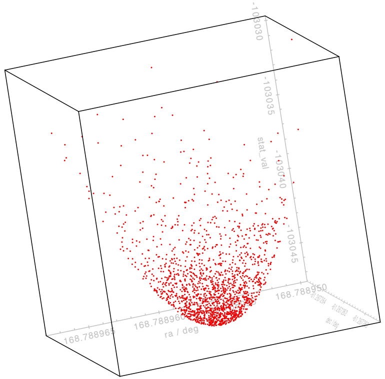
CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

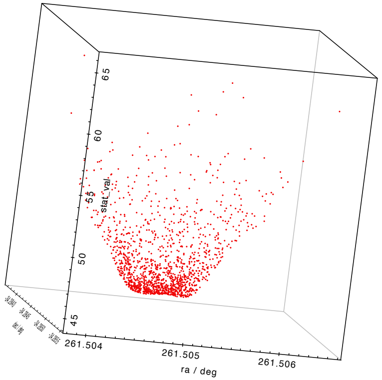
Confidence intervals I

- Measurements such as X-ray photometry fluxes are only as robust as the confidence intervals on the measurements
- Confidence intervals are typically estimated from
  - Raw data uncertainties (*e.g.*, photon statistics)
  - Calibration data uncertainties
  - Calibration systematic uncertainties
  - Model/fitting uncertainties
- Combining uncertainties means that confidence intervals are *rarely Gaussian*, *seldom symmetric*, and *often not analytic*

Confidence intervals II

- PDFs and MCMC draws often provide better representations of the true confidence distributions
- The CSC quotes independent lower and upper confidence limits but in some cases also provides PDFs and/or MCMC draws
  - The end user can calculate whatever confidence percentile they choose
  - MCMC draws can provide both measurement information and confidence intervals





RA/Dec draws for two separate detections in CSC 2.0

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Temporal variability

- Many X-ray sources are temporally variable (both within a single observation and from observation to observation)
  - Users want both individual epoch data (for temporal studies) and multi-epoch data
  - Combining data from multiple epochs often improves S/N
    - CSC does this via a Bayesian Blocks analysis to identify epochs to be combined for which the source is in a similar state
  - *Non*-detections can provide photometric upper limits
  - Users also want “canonical” properties for the source
    - Best estimate (“most useful”) values and global averages

Some additional considerations

- Significant ancillary data are required to characterize a measured source property
  - For the CSC, response matrices, auxiliary responses, spectral models, observation epochs, combined observations, … all go into defining a measurement
  - Some measurements have built in assumptions (such as the source spectral model) and there may be multiple alternatives
- There may be a many-to-many relationship between *detections* and *sources* that must be captured
  - For *Chandra* the PSF size varies by ~100× across the field of view

Data model notes I

- Data models must provide information about the *shape* of the confidence distribution (*especially* if it is not analytic)
  - Comparing confidence limits (e.g., 95% confidence) between different datasets is impossible if the distribution is not known
- Data models must support use of PDFs and MCMC draws representations of measurement confidence intervals
  - The use of Bayesian models to compute measurement PDFs via MCMC draws is rapidly becoming the norm in X-ray astronomy data analyses and will be assumed in the future

Data model notes II

- Data models must associate epoch information with measurements
  - Properties derived from single epochs and groups of epochs may be relevant and highlight different facets of the sources
    - *Ex.:* For a flaring source a global average and an average of epochs during the flares will provide vary different information
- Data models must associate ancillary data including assumptions with measurements
  - Measurements depend on those data and assumption
- Data models must support lower/upper limits on measurements

Conclusions: Data models and X-ray data

- Can IVOA data models represent X-ray and X-ray source data robustly?
  - Mark C.-D.'s use case examples suggest few (or no?) changes may be required for many basic models (measurement, coordinates, cube, …)
  - Some models may require more extensive revisions before they can robustly represent X-ray astronomy data
  - For X-ray source models, MANGO appears to provide a good framework on which to build the necessary robust representations in the future

THANK YOU!