From Chandra to Lynx : 2017 Aug 8-10, Cambridge MA

Lynx Data: Analysis Challenges

Vinay Kashyap (CHASC/CXC/CfA)
Pat Broos (Penn State), Peter Freeman (CMU), Andrew Ptak (GSFC), Aneta Siemiginowska (CfA), Alexey Vikhlinin (CfA), Andreas Zezas (Crete)

What do we want?

More Astronomy

The type of analysis you bring to bear on the data can have a significant impact on what inference is possible.

Example: Source Significance

- Back in the '90s, the best measure of the reality of a source was S/N. Now, we compute the probability of observing a background fluctuation of the same size as the observed data.
- Switching from

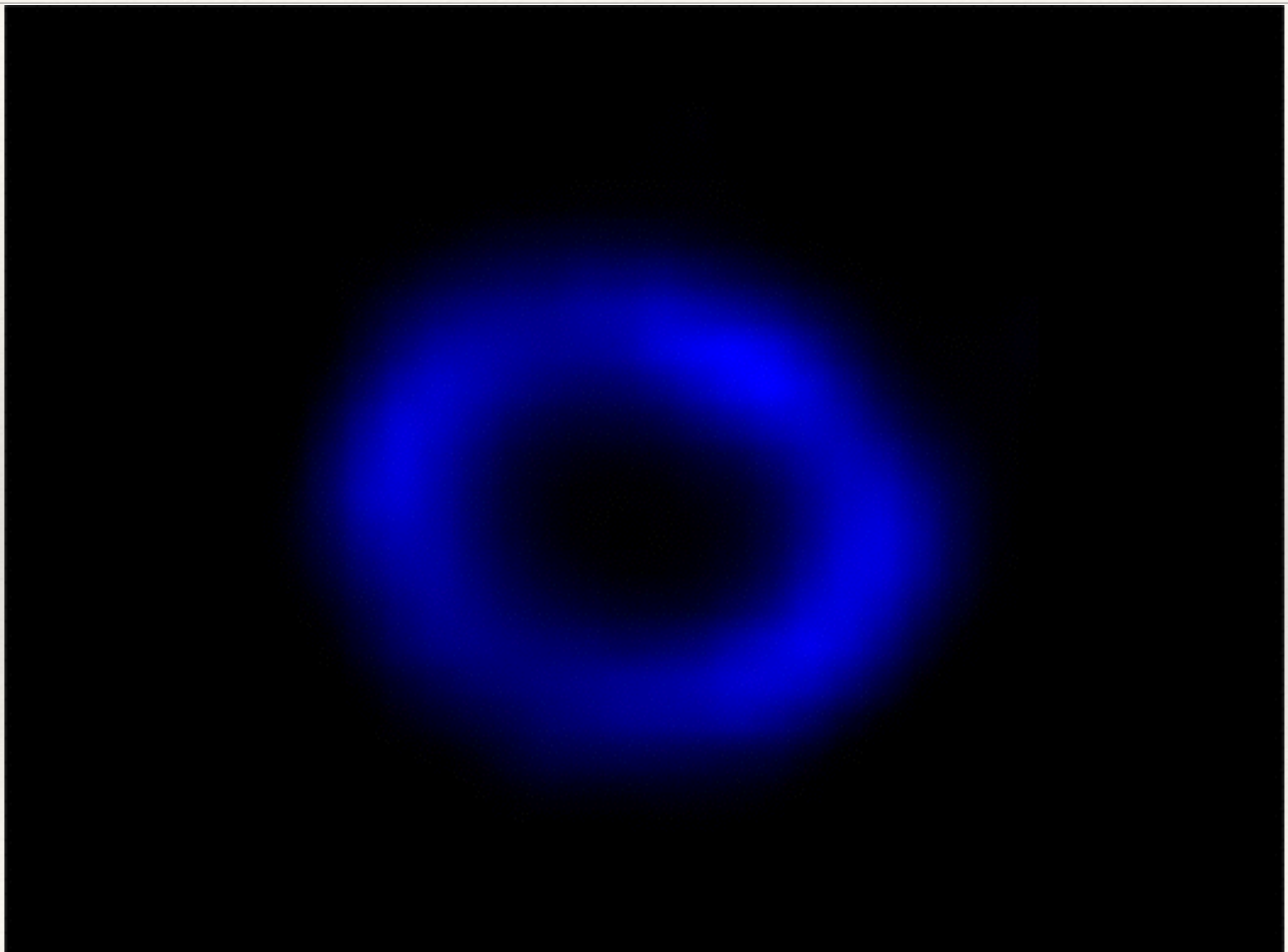
$$\frac{S}{N} = \frac{N_S - N_B/r_B}{\sqrt{N_S + N_B/r_B^2}}$$

to

$$Pr(k \geq N_S) = \sum_{k \geq N_S} \frac{\left(\frac{N_B}{r_B}\right)^k e^{-\frac{N_B}{r_B}}}{\Gamma(k+1)}$$

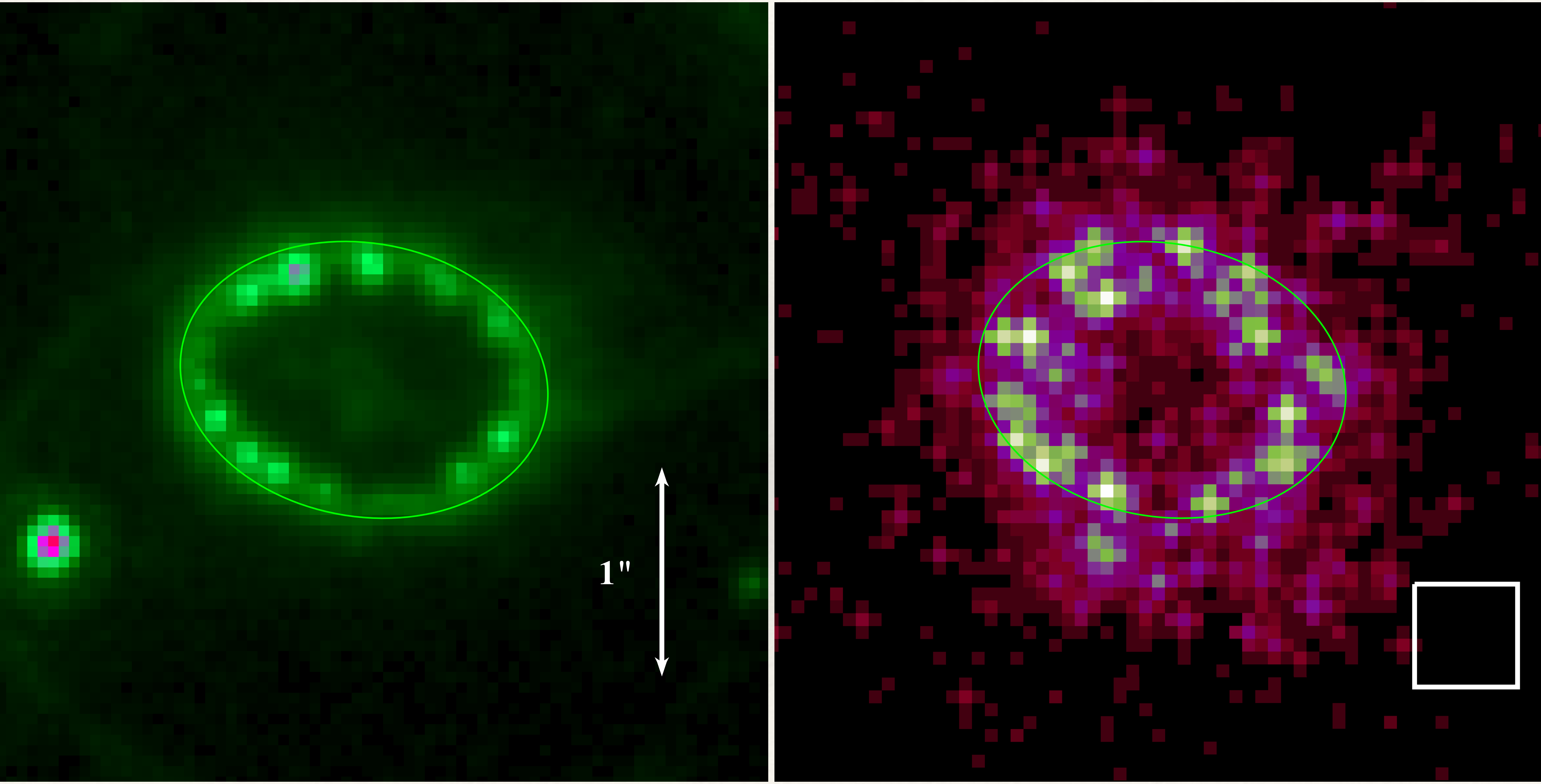
meant you went from needing 10 counts for a detection to needing 3

Example: SN 1987A



http://chandra.harvard.edu/photo/2017/sn1987a/

Example: SN 1987A



Contemporaneous *HST* (left) and *Chandra* (right) from 2001-dec

The lesson from AXAF

AXAF deliberately and explicitly invested in analysis technology.

The AXAF Beta Sites at Chicago and Hawaii produced `wavdetect`[1], and `vtpdetect`[2], and helped to plan the toolset for `CIAO`.

and from whence the statistical foundations of Sherpa were aquihired

Chandra supported the collaboration between high-energy astrophysicists and statisticians via CHASC[3],

which has given us `pyBLoCXS`[4], the MCMC tool in Sherpa, also used to handle calibration uncertainty[5,6], hardness ratio[7] and aperture photometry[8] tools in `CIAO` and CSC, and `LIRA`[9,10,11], among others.

[1]Freeman et al. 2002, [2]Ebeling & Wiedenman 1993, [3]Siemiginowska et al. 1997, [4]van Dyk et al. 2001, [5]Lee et al. 2011, [6]Xu et al 2014, [7]Park et al. 2006, [8]Primini & Kashyap 2014, [9]Esch et al. 2004, [10]Connors & van Dyk 2007, [11]McKeough et al. 2016

A Laundry List

A. Calibration issues

Analysis algorithms are often constrained by what is made possible by spacecraft design and what can be calibrated

B. New algorithms

Many new algorithms are currently being developed with *Chandra* data in mind, could make *Lynx* data more valuable

C. Advances in Statistics

New techniques are being developed by Statisticians, and will allow for better inferences to be drawn

(A) PSF

- *Lynx*'s PSF will have more degrees of freedom (more shells, mirror adjustability) than *Chandra*'s and will need a correspondingly greater effort to characterize and use
- Need high-fidelity models of the mirrors and the detectors, and tools to deal with variations in energy and across the FOV
- Photometry via PSF-fitting in the Poisson regime is still not bread-and-butter as in optical/IR
- Pileup could be a big problem because of high EA — mitigation via hardware (higher frame rates, oversampling) or software (modeling the pileup process, bootstrapping from the wings)

(A) Pointings

- *Chandra* has shown the value of mosaic observations. Analysis tools to deal with such re-aligned datasets are still kludgey
- Need to consider strategies to handle absolute alignments of multiple observations
- Need tools for source confusion analysis

(A) RMF

- Need to consider strategies to ameliorate and correct for long-term CTI and contamination
- Fitting global models to high-resolution calorimeter data is fraught with peril — we have had a taste with Chandra and XMM grating data, but Lynx data will push the boundaries in counts, resolution, and number
  - fitting algorithms must learn to guard against model misspecification[1], become more intelligent at discounting $\delta\chi$ where systematics are known to be large, find better ways to simultaneously fit spectra of different resolutions
- Improvements to atomic line databases (e.g., AtomDB, Chianti) must continue, and new algorithms are needed to propagate the highly non-linear error structure into analysis and inference

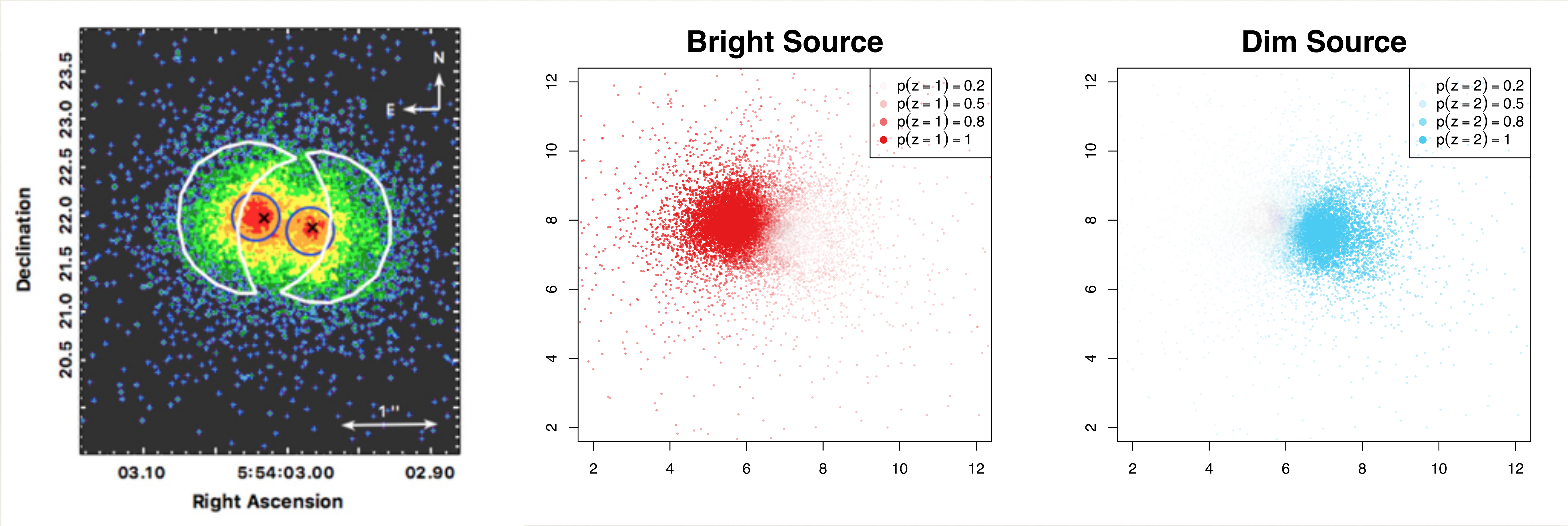
[1] *All models are wrong, but some are useful.* — George Box (British Statistician)

(B) Disambiguate Overlaps

- The goal is to sift the photons that belong to overlapping sources into separate piles probabilistically and carry out spectral and timing analyses on them
  - Use both spatial and rudimentary gross spectral information — Jones et al. 2015, ApJ 808, 137
  - Use spatial, gross spectral, *and* temporal information — Campos et al., in development
  - Use spatial and temporal information, and astrophysical spectral modeling information, hooked into Sherpa — Campos et al., contemplated

(B) Disambiguate Overlaps

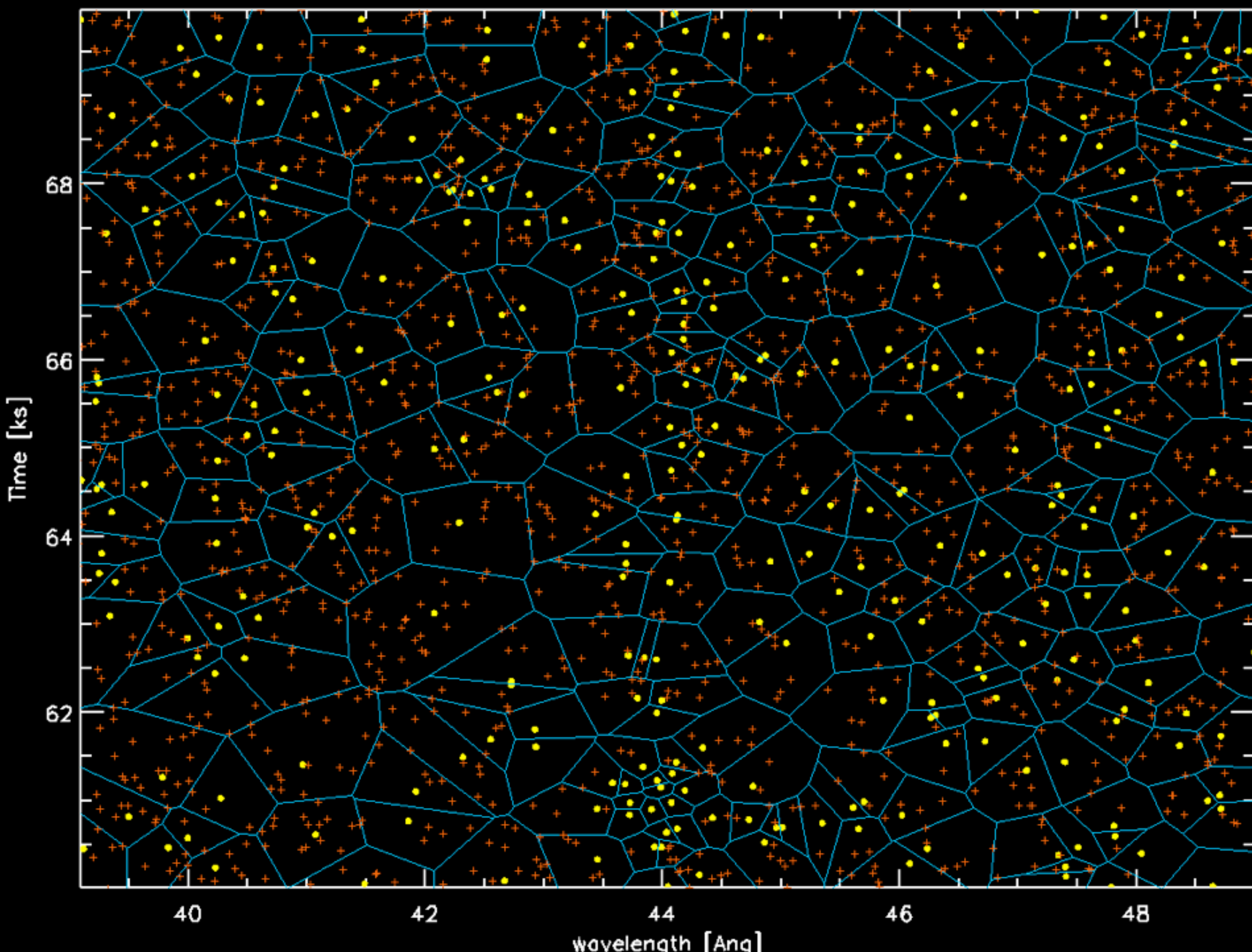
HBC 515 Aa+Ab weak-lined T-Tauri binary (Principe et al. 2016)



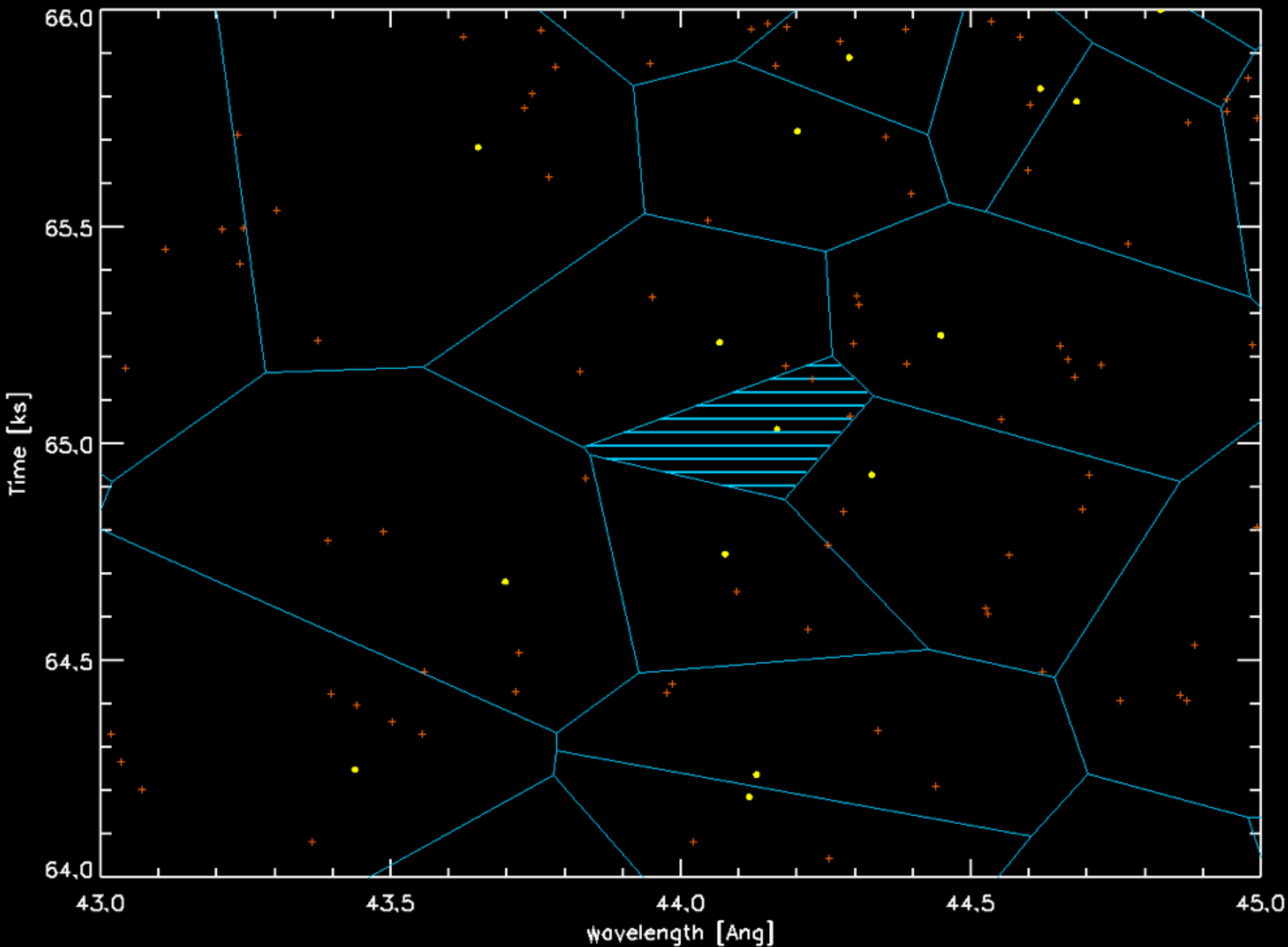
E-BASCS probability assignments based on spectral and temporal disambiguation

(B) Non-parametric Fluxes

- eff2evt: convert measured photon energies to flux using detector QE and telescope EA
  - Works fine when there are a lot of photons, but blows up when EA is small or events are sparse, and does not provide error bars
- New technique that accounts for possible range over which event can appear, and draws information from likely spectral model if available is in development



Time [ks]
68
66
64
62
40
42
44
46
48
wavelength [Ang]

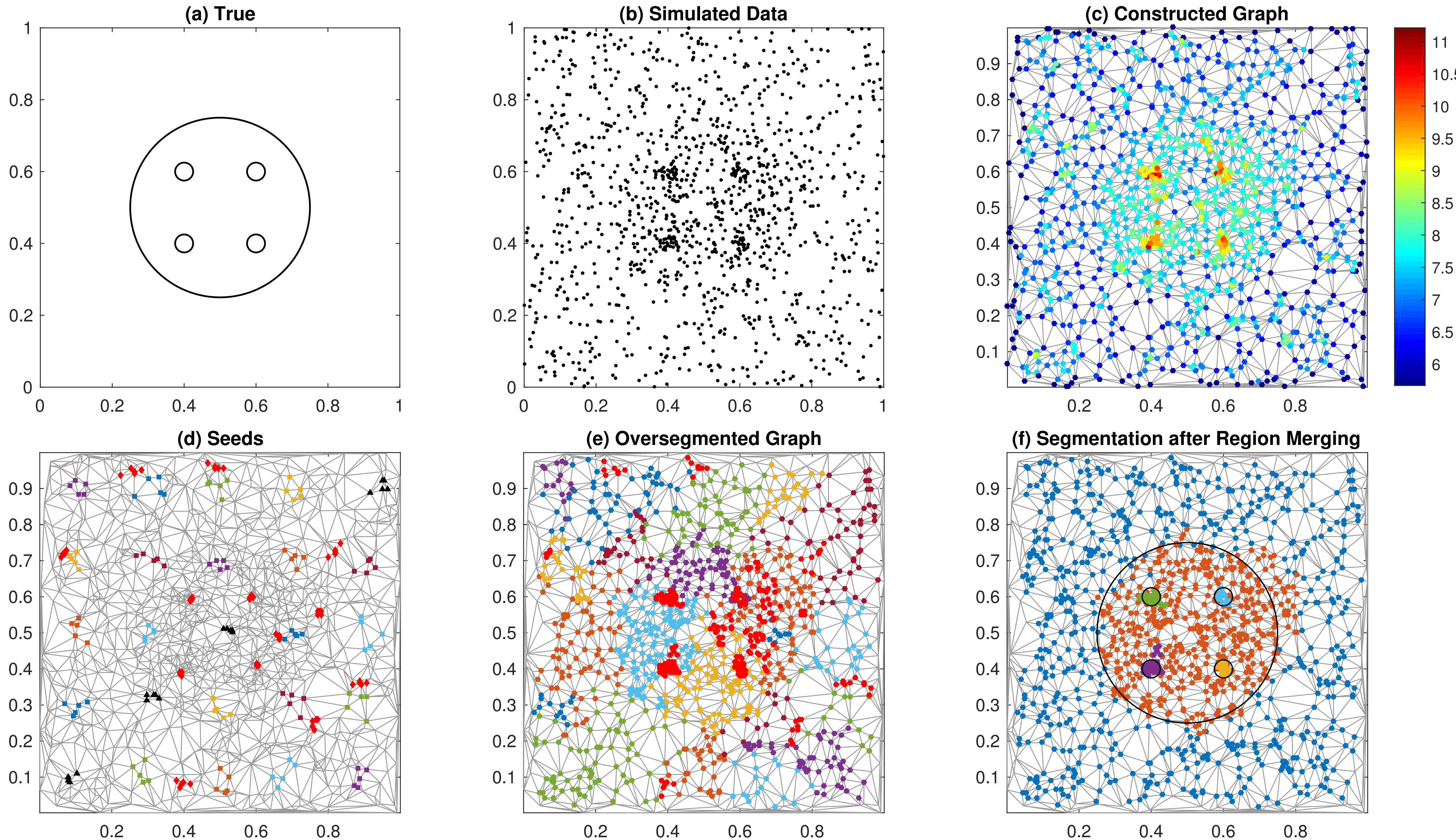


66.0
65.5
65.0
64.5
64.0
Time [ks]
43.0
43.5
44.0
44.5
45.0
wavelength [Ang]

(B) Adaptive Segmentation

- `csmooth`: adaptively smooth image by enforcing a S/N
  - Highly successful for displaying Chandra data, but difficult to do science with
- What if we could segment the events list based on some criterion for local similarity?
  - Graphed oversegmented seeded region growing, with subsequent merging using likelihood ratio type tests — Minjie Fan et al. 2017, in preparation

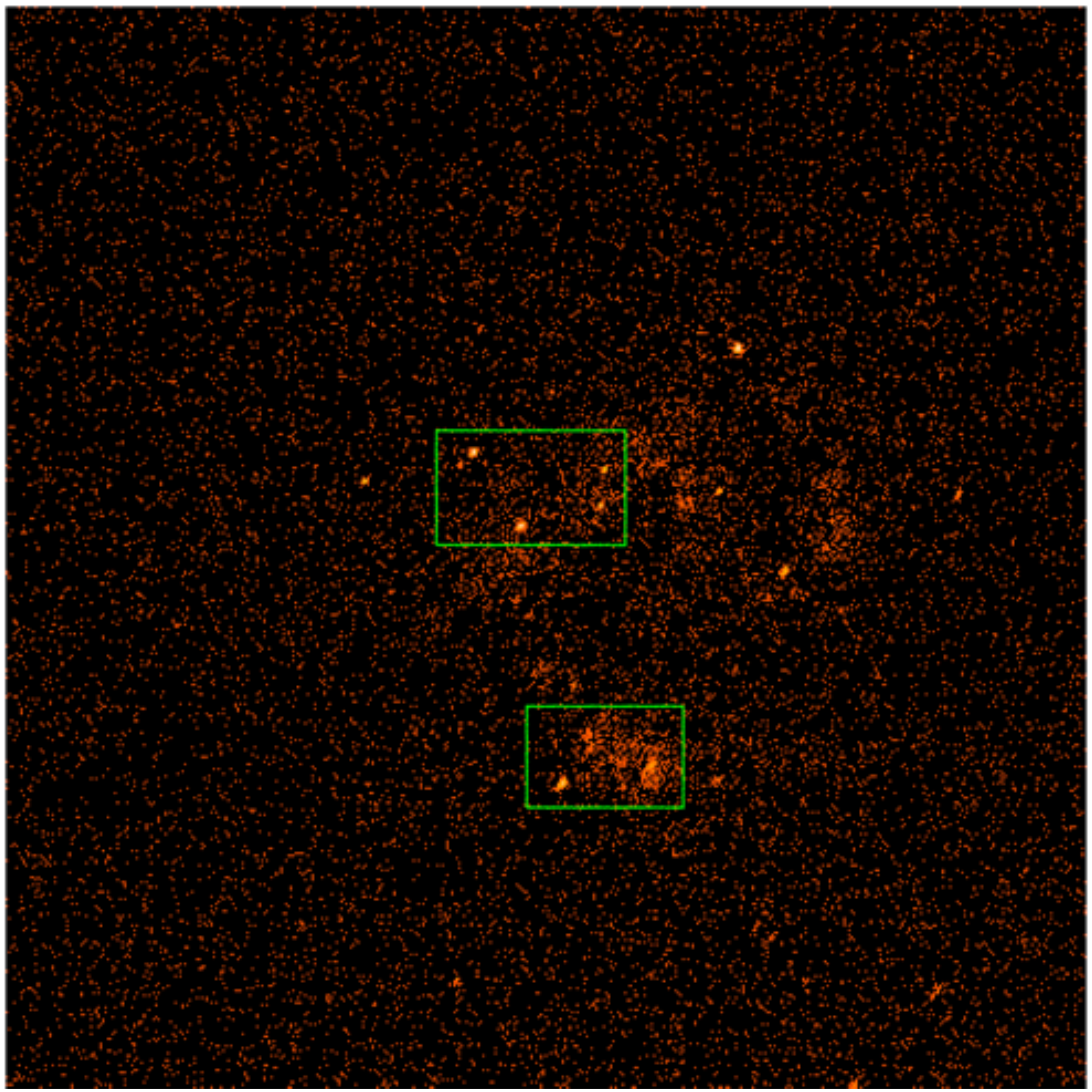
Seeded Region Growing in Poisson Regime

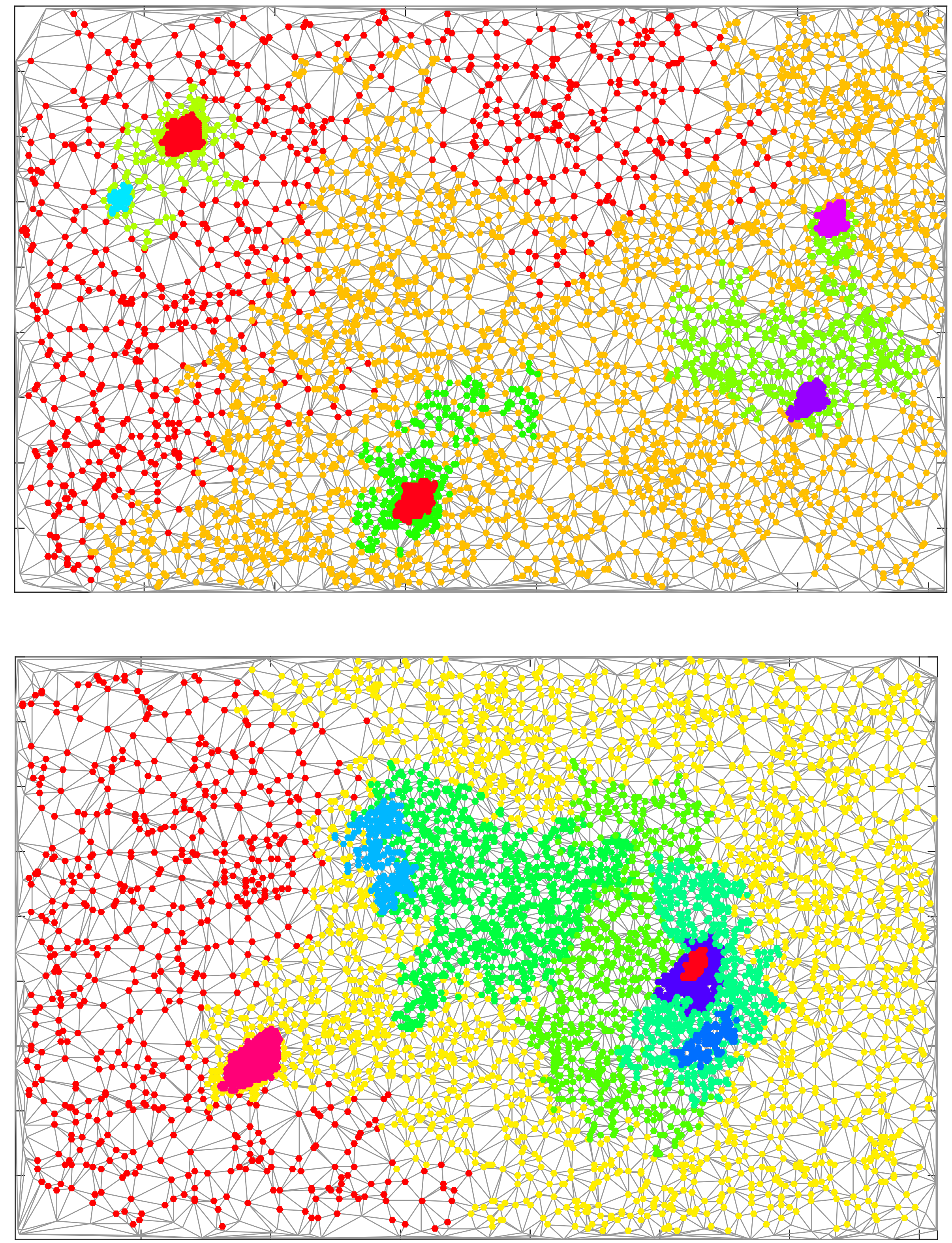


Fan, Lee et al.

[from Andreas Zezas]

Seeded Region Growing in Poisson Regime





Fan, Lee et al.

[from Andreas Zezas]

(B) Multi-band Deconvolution

- Deconvolution and/or reconstruction is currently limited to images. To derive spectral information requires making images in different bands and independently analyzing them
  - Not optimal, because fewer counts in each image means larger errors, and independent analyses imply loss of connecting information
- Work is in progress to upgrade LIRA to simultaneously reconstruct images in multiple passbands

(B) Robust Fitting

- There is a big problem with simultaneously fitting multiple datasets using a likelihood-based (χ^2, cstat) statistic, if the sizes of the datasets differ significantly.
  - You can't easily fit a high-resolution grating spectrum together with a low-resolution CCD spectrum, or an SED to spectroscopic and photometric data, or a small point source in the wing of a bright source
- Work is in progress to develop suitable weighting functions to loosen the tyranny of the bins

(B) cstat gof

- A long standing problem with fitting spectra in the Poisson regime has been the lack of a measure of the goodness of fit when using cstat.
- A new parameterization of goodness of fit using the mean and stddev of expected cstat has been derived recently by Kaastra 2017, arXiv:1707.09202
- This is an encouraging breakthrough, but more work is needed!

(C) New Stats

- We have got a lot of mileage out of χ^2 and Maximum Likelihood and MaxEnt and wavelets
- Markov Chain Monte Carlo is becoming widely used
- What could be next?

(C) New Stats

- **Hierarchical Bayes**
  - ability to build complex models for inference and classification and account for large amount of interrelationships among model parameters and instrument behavior
- **Gaussian Processes**
  - Continuous stochastic process that can be used to make extrapolations and distinguishing multiple trends from known or trained data
- **Fiducial Inference**
  - Compute probabilities and confidence bounds without having to set up prior probability distributions
- **Deep learning**
  - Applying multi-level, cascading non-linear transformations (aka artificial neural networks) to extract relevant features from a dataset (aka Machine Learning)

The Ω Group

- An informal ωG, just send one of us an email to "join". We will also be recruiting real statisticians to consult with.
  - Pat Broos
  - Peter Freeman
  - Vinay Kashyap
  - Andrew Ptak
  - Aneta Siemiginowska
  - Alexey Vikhlinin
  - Andreas Zezas