BIRS Astrostatistics : 2023 Nov 01

Modeling Multi-Dimensional Astronomical Data

VINAY KASHYAP

CHASC AstroStatistics Center
Center for Astrophysics | Harvard & Smithsonian

Highlights from CHASC

https://hea-www.harvard.edu/AstroStat/

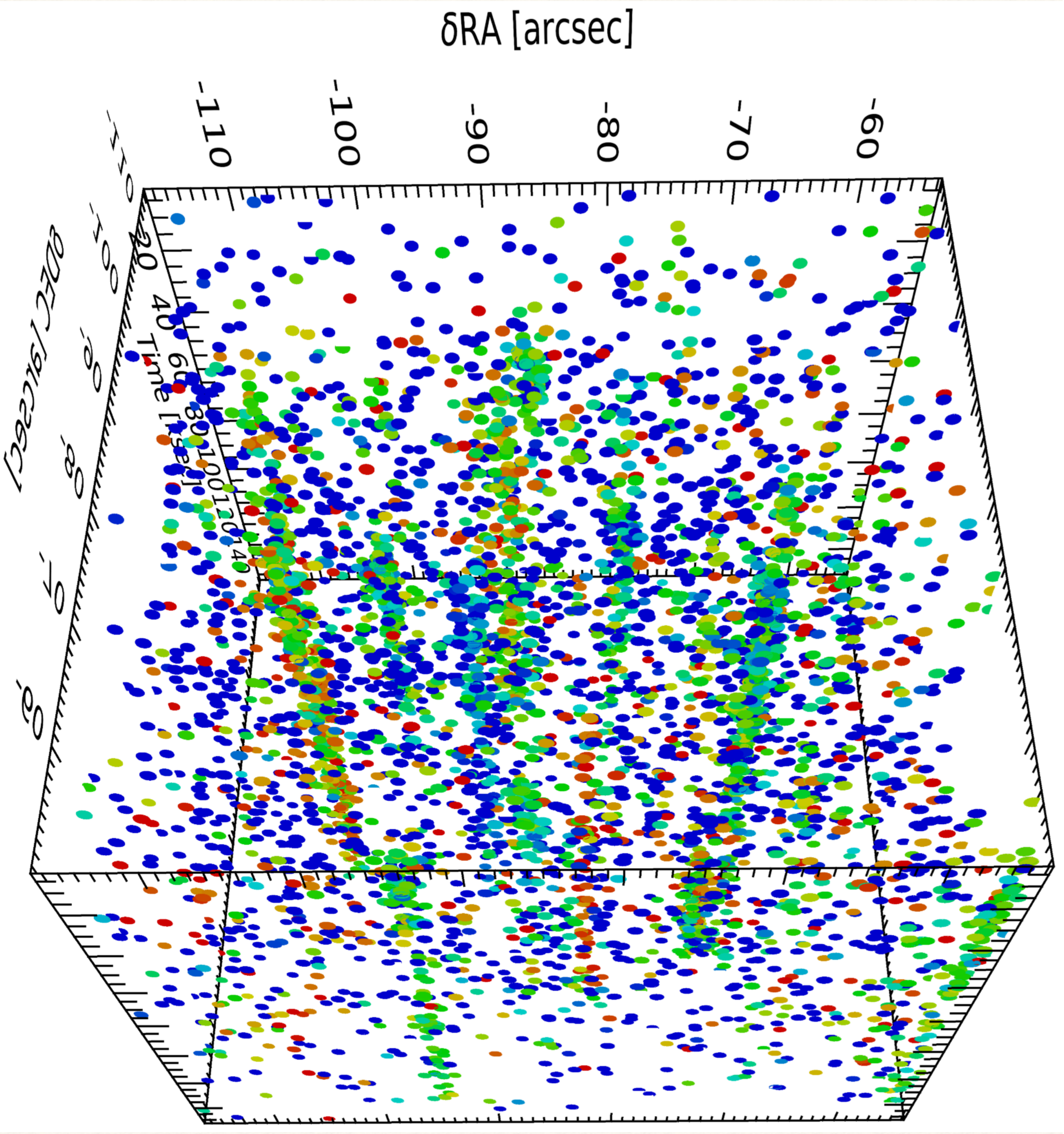
- The CHASC AstroStatistics Collaboration has been operating since c.1997
- Started as a collaboration between astrophysicists at the Center for Astrophysics and statisticians at Harvard to handle challenges of high-quality data anticipated from the Chandra X-ray Observatory
- Has now expanded to involve astrophysicists from CfA, MIT, Crete, Cambridge, IUCAA, GSFC, and statisticians from Harvard, Imperial, UC Davis, Michigan, Penn State, Simon Fraser, Columbia, Williams College
- Responsible for ≈40 PhD theses, ≳10 Masters theses

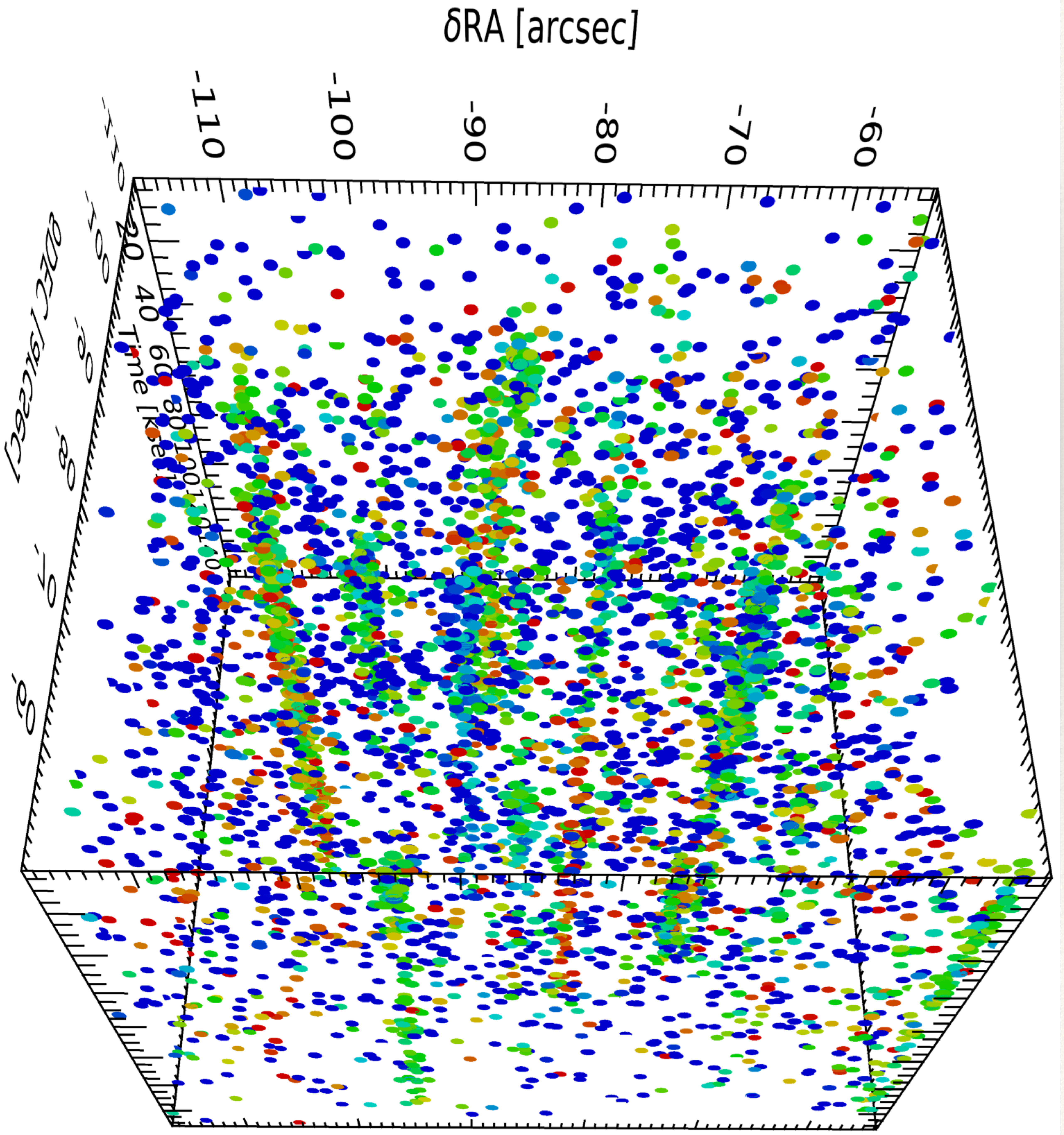
Astronomical Data are Multi-dimensional

- For several years now, we at CHASC have been developing algorithms to analyze multi-dimensional data focused on “lists of events” — photons, sunspots, flares, collections of fluxes, etc.
- I will give a broad overview of some of the highlights; ask David van Dyk, Aneta Siemiginowska, David Stenning, Yang Chen, or Max Autenrieth for details

High-Energy Astro: marked Poisson process

- High-energy data are 4-way tables of photons, with spatial, spectral, and temporal marks associated with each photon: $\{x, y, t, E\}$
- $\{x, y, t, E\}$ are projected onto a smaller set of axes, and 1-D or 2-D histograms are used to extract sources or variability events or spectral features
  - $\{x,y\} \rightarrow$ counts image I_{ij}
  - $\{t\} \rightarrow$ light curve l_{k}
  - $\{E\} \rightarrow$ energy or wavelength spectrum s_{p}
- combinations are also interesting
  - $\{x,y,t\} \rightarrow$ spatio–temporal variations
  - $\{x,y,E\} \rightarrow$ spatio–spectral variations
  - $\{t,E\} \rightarrow$ spectral variability
  - $\{x,y,t,E\} \rightarrow$ everything everywhere all at once


δRA [arcsec]
Time [ksec]
δDEC [arcsec]


δRA [arcsec]
Time [ksec]
δDEC [arcsec]

0-D and 1-D

Application	Analysis	Reference
Non detections/upper limits	balance of Type I and Type II, smooth tests and LRT	Kashyap et al. 2010 ApJ, Zhang et al. 2023 MNRAS
Spectral hardness (BEHR)	hierarchical Bayesian modeling	Park et al. 2006 ApJ
Modeling low-counts spectra, narrow lines in low-res spectra (BLoCXS)	MCMC with multimodal posteriors	van Dyk et al. 2001 ApJ, Park et al. 2008 ApJ
Collections (logN-logS, luminosity functions, power-law distributions, sunspot numbers, flare distributions, time delays)	data augmentation, Maximum Product of Spacings, multi-stage Bayesian, O-U/CARMA, Gaussian Processes	Yu et al. 2012 SolPhys, Tak et al. 2017 ApJ, Yan et al. 2021 RNAAS, Meyers et al. 2023; and in prep: Autenrieth et al., Wang et al., Yan et al., Ingram et al.

David van Dyk, Aneta Siemiginowska, David Stenning, Yang Chen, Max Autenrieth

1 1/2 D and 2-D

Application	Analysis	Reference
Incorporate calibration uncertainty in spectral modeling	Pragmatic and Fully Bayes, shrinkage estimation	Lee et al. 2011 ApJ, Xu et al. 2014 ApJ, Chen et al. 2018 AoAS, Yu et al. 2018 ApJ, Marshall et al. 2021 AJ, Yu et al., 2023 (submitted ApJ)
Image deconvolution with error bars (LIRA, jolideco)	Multiscale hierarchical Bayesian, p-value upper bounds, Ising, Genetic algorithms	Esch et al. 2004 ApJ, Stein et al. 2015 ApJ, McKeough et al. (in prep), Donath et al. (in prep)
(Spatial) segmentation and boundaries in event lists (SRGonG, BFD-SRGonG)	Graphed seeded region growing	Fan et al. 2023 ApJ, Wang et al. (in prep)
Spectro-temporal change points (Automark)	Minimum descriptor lengths	Wong et al. 2016 AoAS

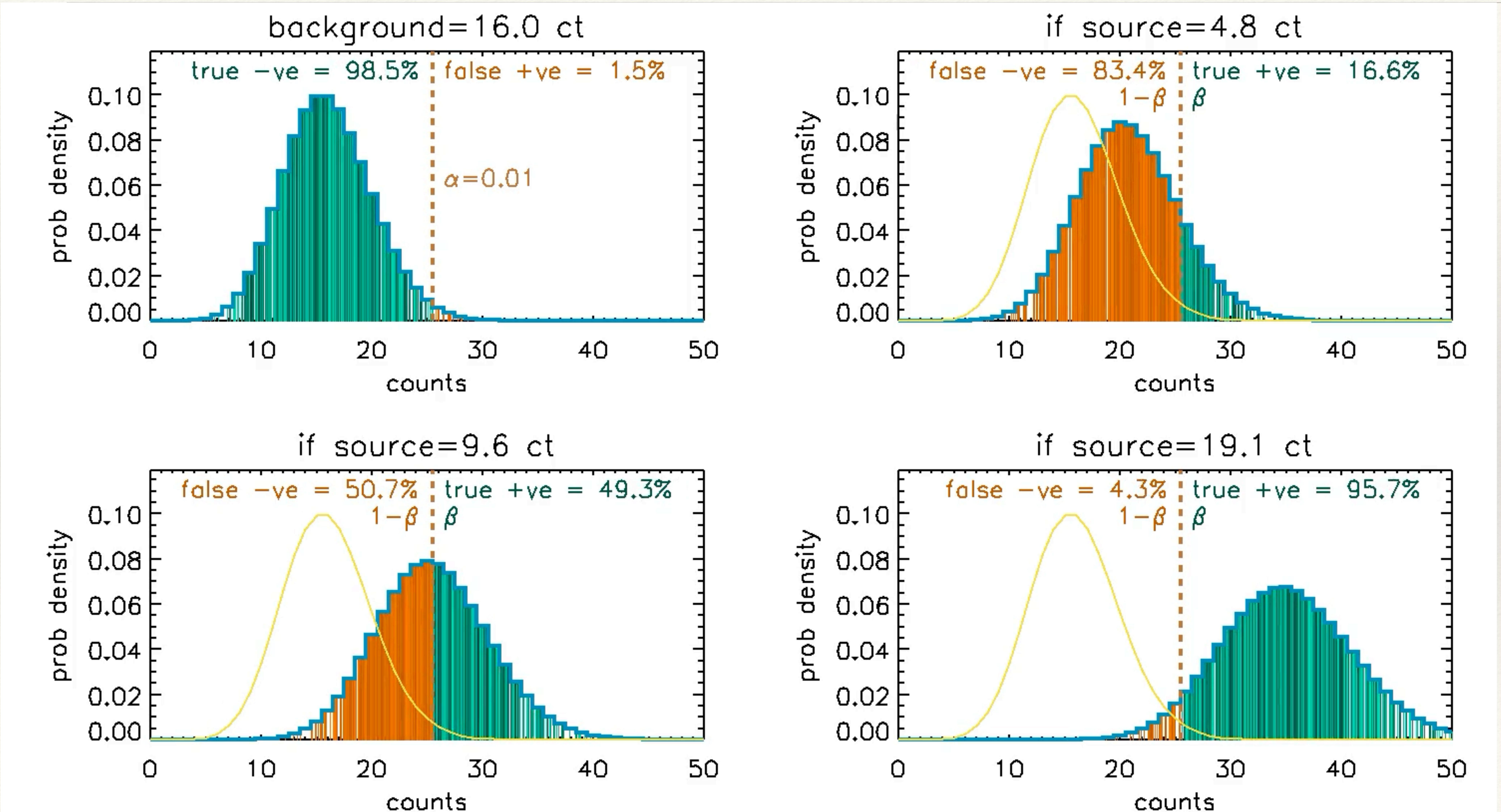
3-D and 4-D

Application	Analysis	Reference
spatio-spectral disambiguation of overlapping sources (BASCS)	Bayesian mixtures and Reversible Jump MCMC	Jones et al. 2015 ApJ
spatio-spectro-temporal disambiguation of overlapping sources (EBASCS)	Bayesian mixtures	Meyer et al. 2021 MNRAS
spatio-spectro-temporal change points in multi-filter data cubes (4D Automark)	Seeded region growing and minium descriptor lengths	Xu et al. 2022 AJ

Non detections/upper limits	balance of Type I and Type II, smooth tests
Spectral hardness (BEHR)	hierarchical Bayesian modeling
Narrow lines in low-res spectra (BLoCXS)	MCMC with multimodal posteriors
Collections (logN-logS, luminosity functions, power-law distributions, sunspot numbers, flare distributions, time delays)	data augmentation, Maximum Product of Spacings, multi-stage Bayesian, O-U/CARMA, Gaussian Processes
Incorporate calibration uncertainty in spectral modeling	Pragmatic and Fully Bayes, shrinkage estimation
Image deconvolution with error bars (LIRA, jolideco)	Multiscale hierarchical Bayesian, p-value upper bounds, Ising, Genetic algorithms
(Spatial) segmentation and boundaries in event lists (SRGonG, BFD-SRGonG)	Graphed seeded region growing
Spectro-temporal change points (Automark)	Minimum descriptor lengths
spatio-spectral disambiguation of overlapping sources (BASCS)	Bayesian mixtures and Reversible Jump MCMC
spatio-spectro-temporal disambiguation of overlapping sources (EBASCS)	Bayesian mixtures
spatio-spectro-temporal change points in multi-filter data cubes (4D Automark)	Seeded region growing and minium descriptor lengths

Kashyap et al. 2010, ApJ 719, 900

"0D": flux upper limits to undetected sources

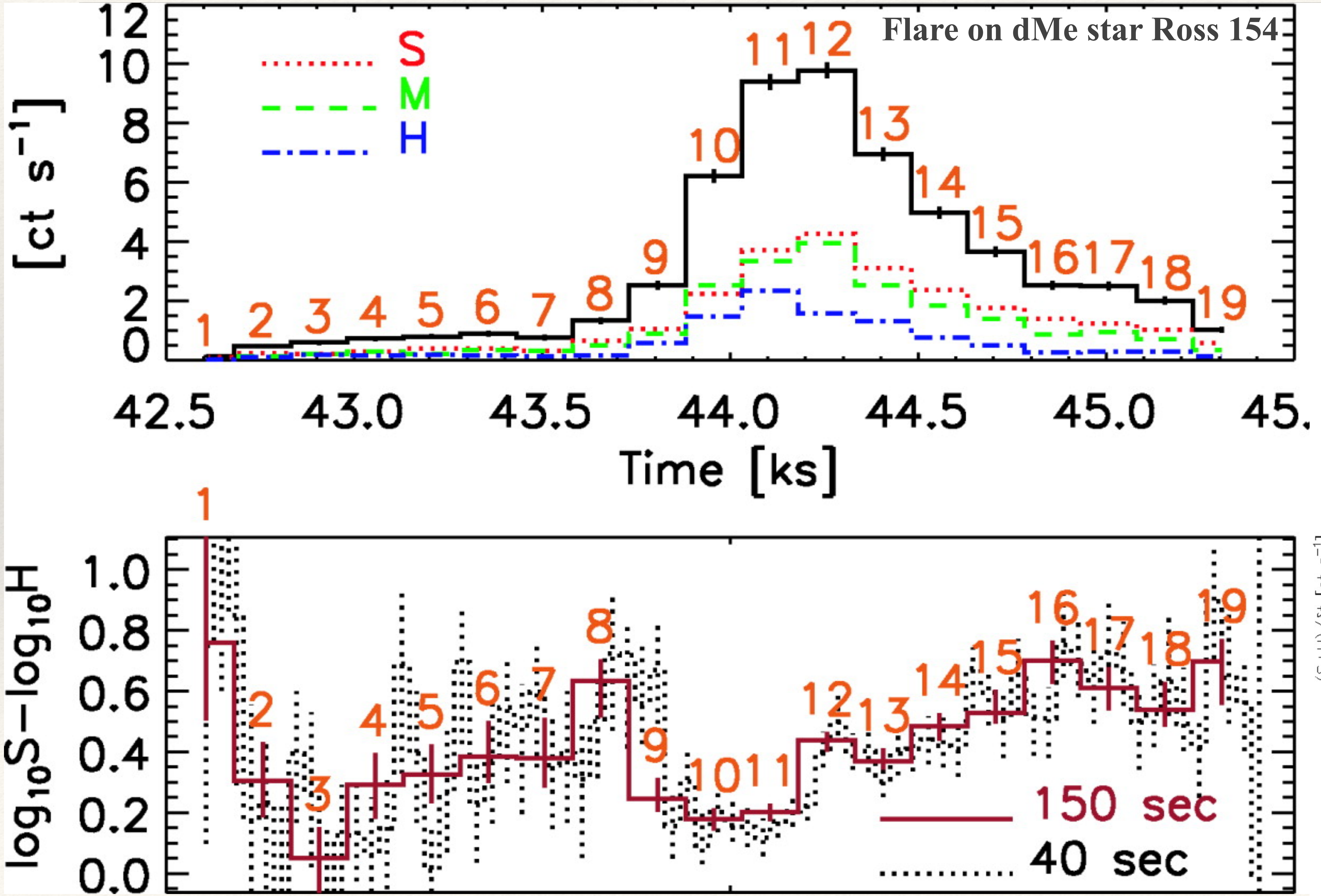


When no signal is detectable, it is useful to know what is the maximum brightness that a source could have at which point it would be detected.

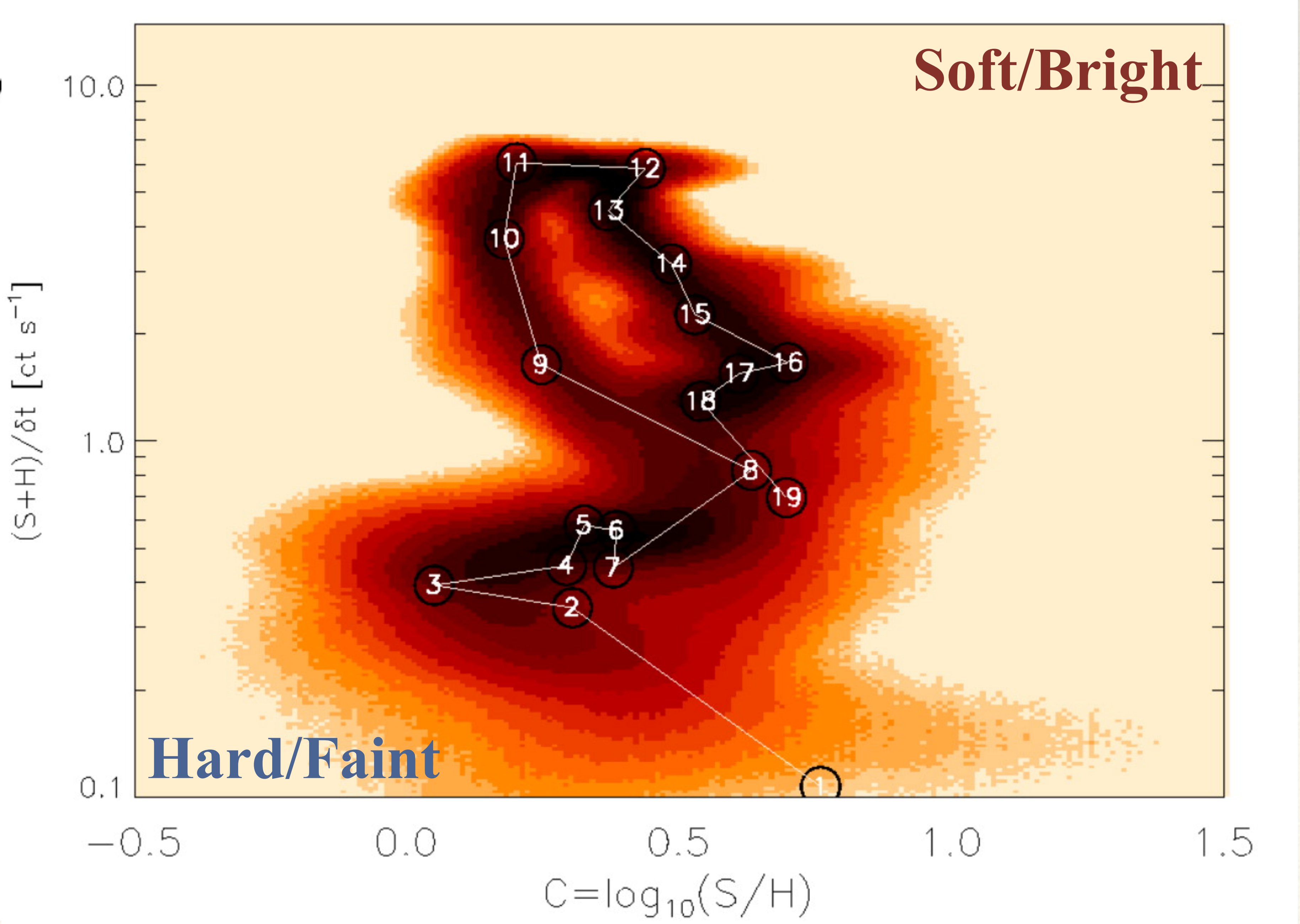
Compute upper limits based on probability of false –ves for a given acceptable false +ve threshold.

Park et al. 2006, ApJ 652, 610

1D: spectral shape in low counts regime

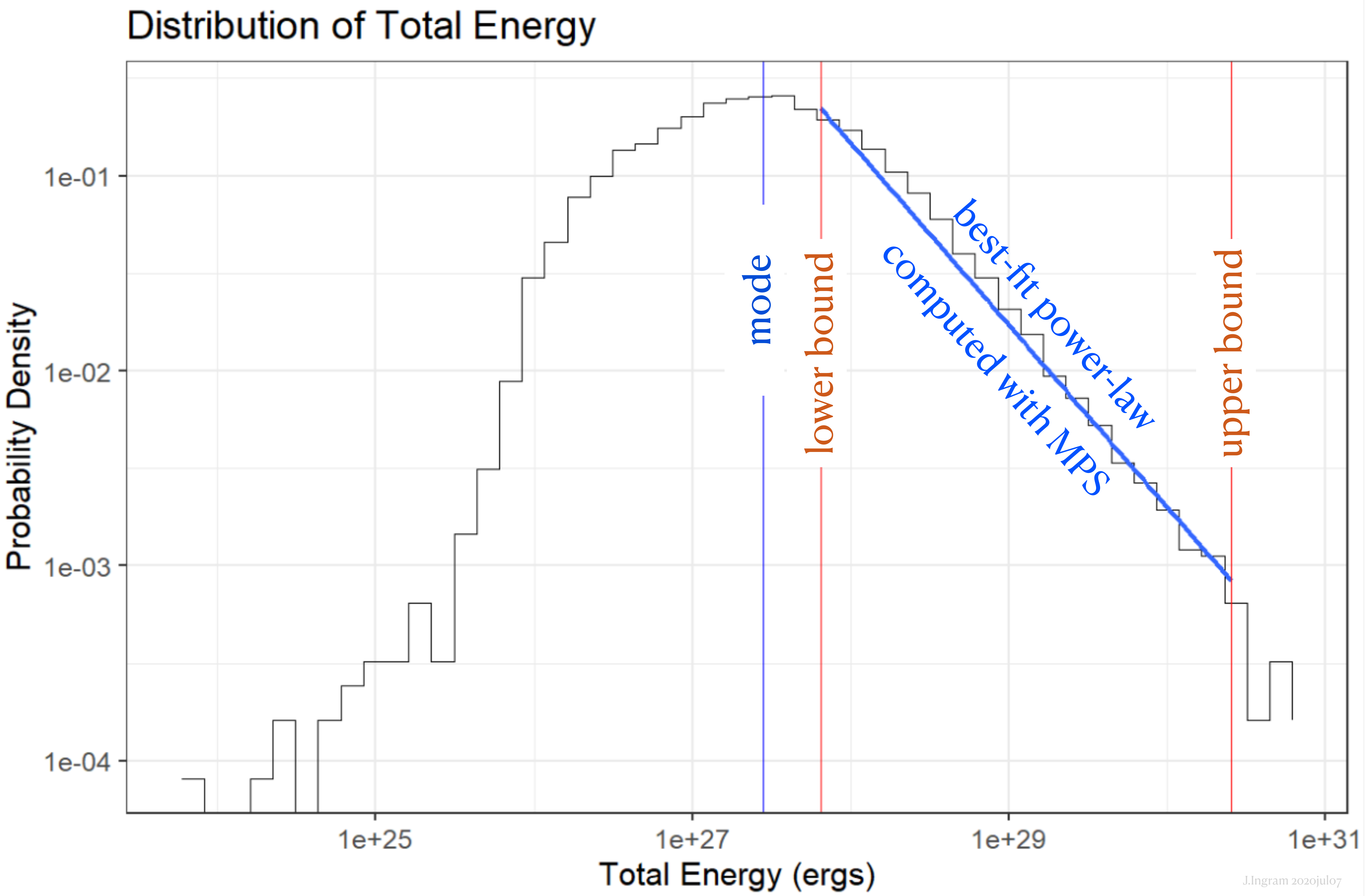


Even when there are too few counts to obtain a model fit to an energy spectrum, we can still get an estimate of the spectral shape via ratios of counts seen in broad bands via Hierarchical Bayesian modeling of Poisson count intensity, while accounting for background, instrument sensitivity, and exposure duration.



Wang, Meng, Ingram, Kashyap, Klingenberg (in prep)

1D: extent of a power-law distribution

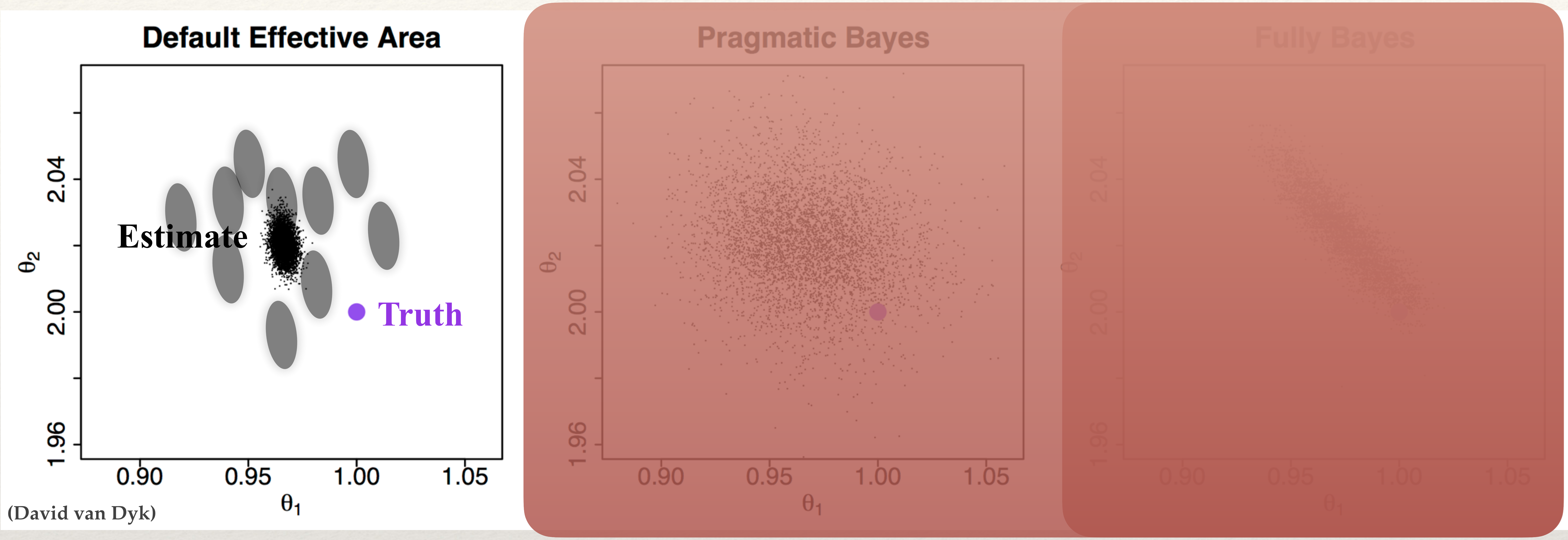


Solar flare energies appear to be distributed as a power-law, but the distribution turns over at both low and high energies. So a naive estimate of the power-law slope will give a biased estimate.

Maximum Product of Spacings is semi-parametric a technique that fits a power-law model over a small range but ignores the rest. So we can self-consistently estimate the upper and lower bounds of applicability of the power-law.

Lee et al. 2011 ApJ 731, 126; Xu et al. 2014 ApJ 794, 97; Yu et al. 2018, ApJ 866, 146; Yu et al. 2023, in prep

1D: incorporating systematic uncertainty



(David van Dyk)

$p(\theta|D,\varepsilon^{(m)})$

$p(\theta|D,\varepsilon^{(def)})$

$p(\varepsilon,\theta|D)$

$\rightarrow p(\theta|D,\varepsilon) \cdot p(\varepsilon)$

$p(\varepsilon,\theta|D)$

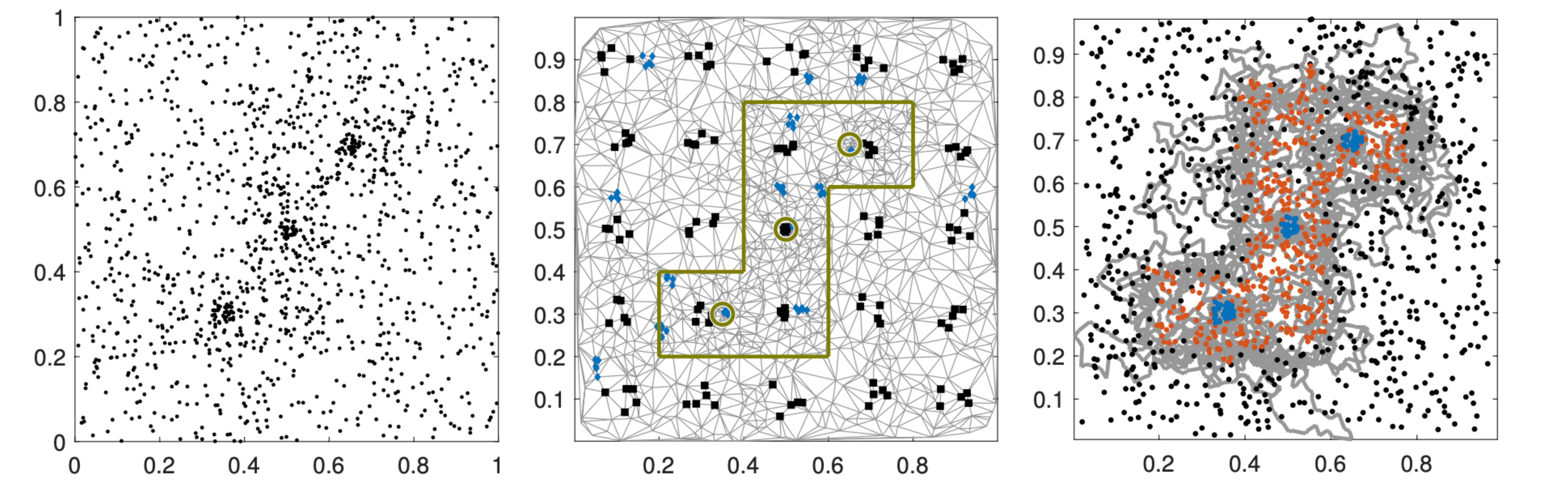
$\rightarrow p(\theta|D,\varepsilon) \cdot p(\varepsilon|D)$

Spectral analysis requires knowledge of instrument sensitivity, which is empirically measured on the ground prior to telescope launch. It is not known perfectly, and also evolves.

How to incorporate this uncertainty into the analysis?

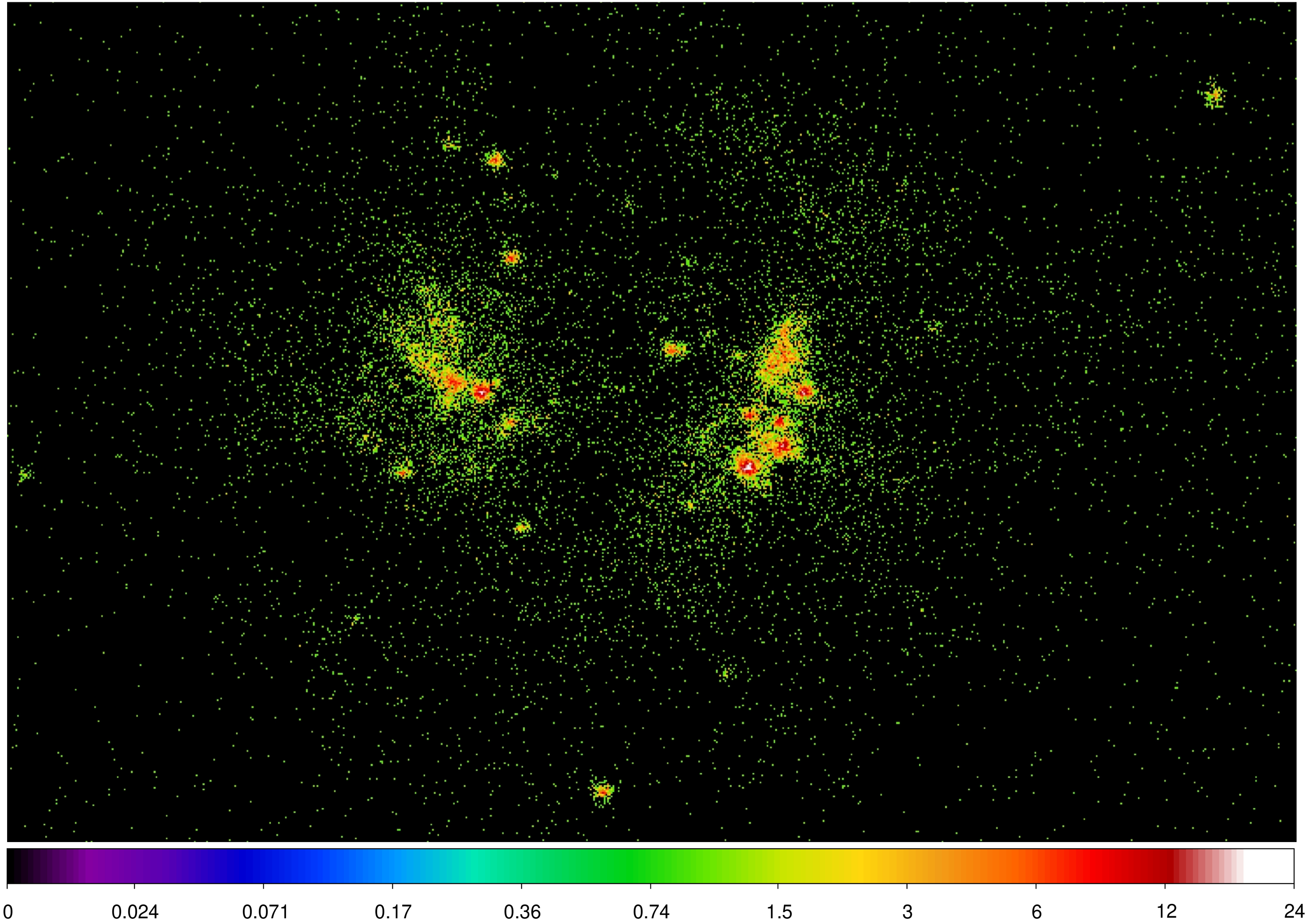
A *pragmatic Bayesian* way, where different choices of the sensitivity are sampled from a prior, and the *fully Bayesian* way where everything is estimated based on the data.

2D: segmentation of event lists



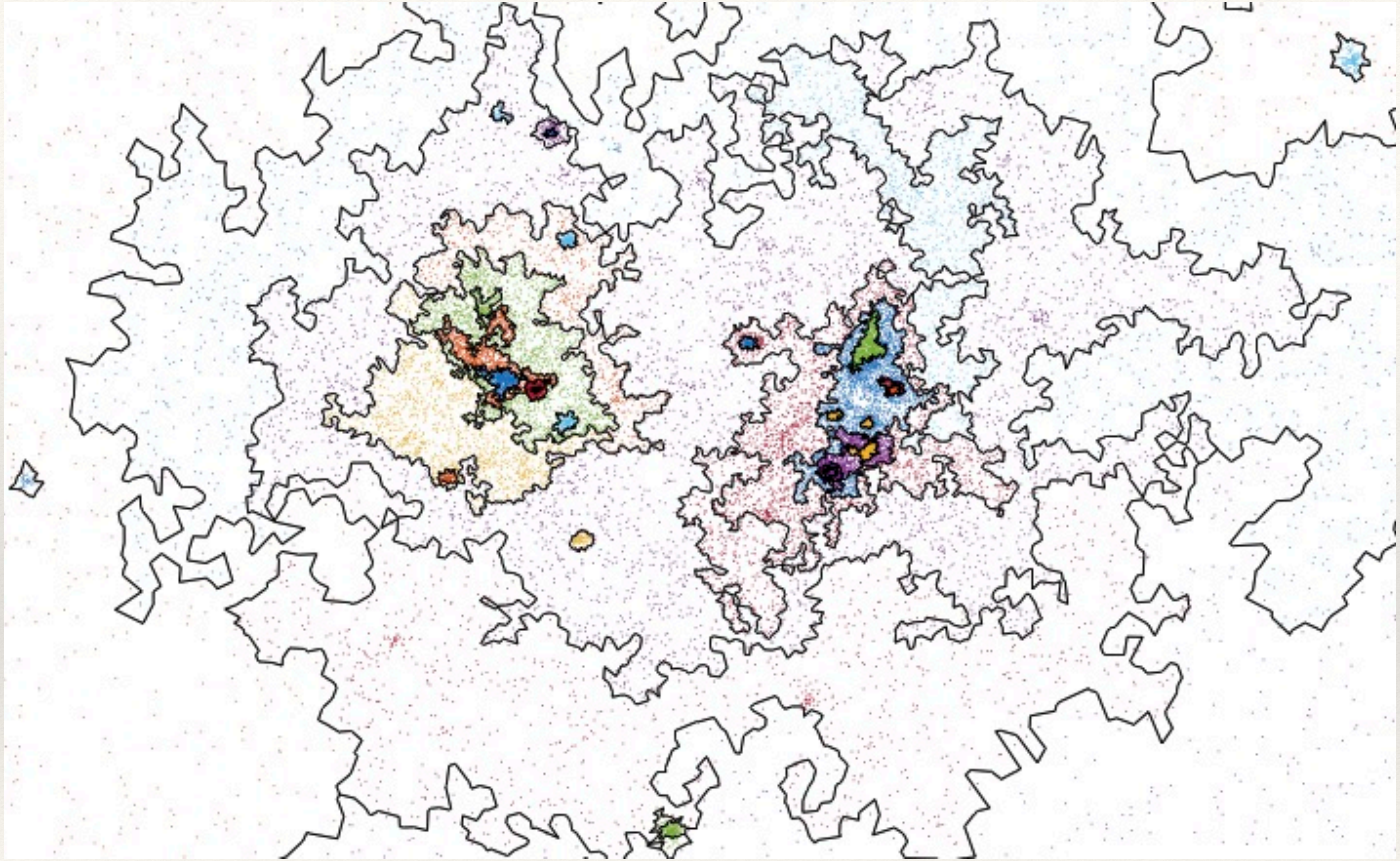
Using graphed Seeded Region Growing, we can define boundaries of diffuse regions and find segmentations without manual supervision.

Start with an oversampling of seeds, aggregate Voronoi cells into clusters based on similarity of surface brightness, and merge segments into an optimum number of ROIs via BIC



SRGonG

Chandra X-ray image of interacting starburst galaxies *Arp 299*



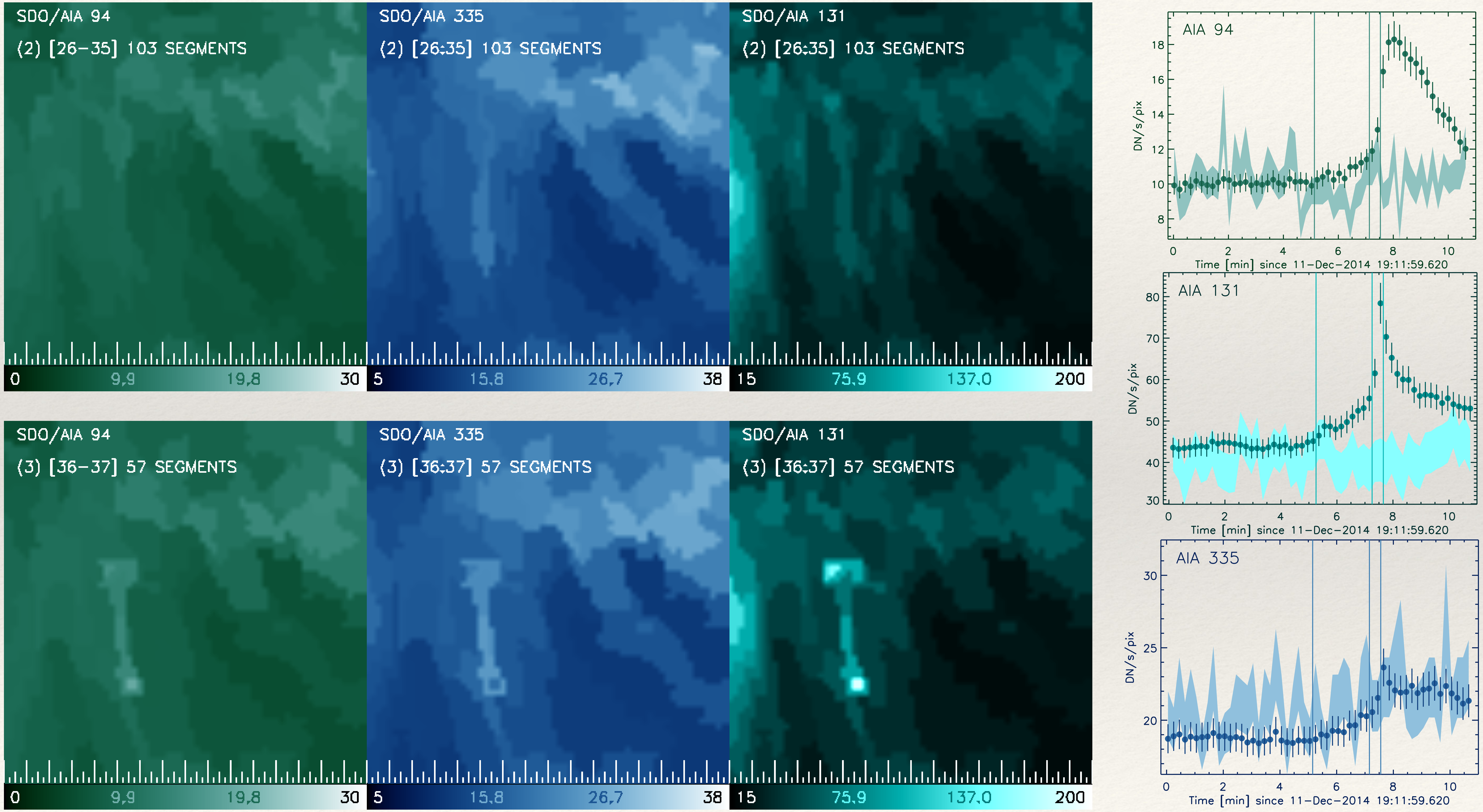
SRGonG

segmentation of *Arp 299* photons

Fan et al. 2023, AJ 165, 66; Wang et al. (in prep)

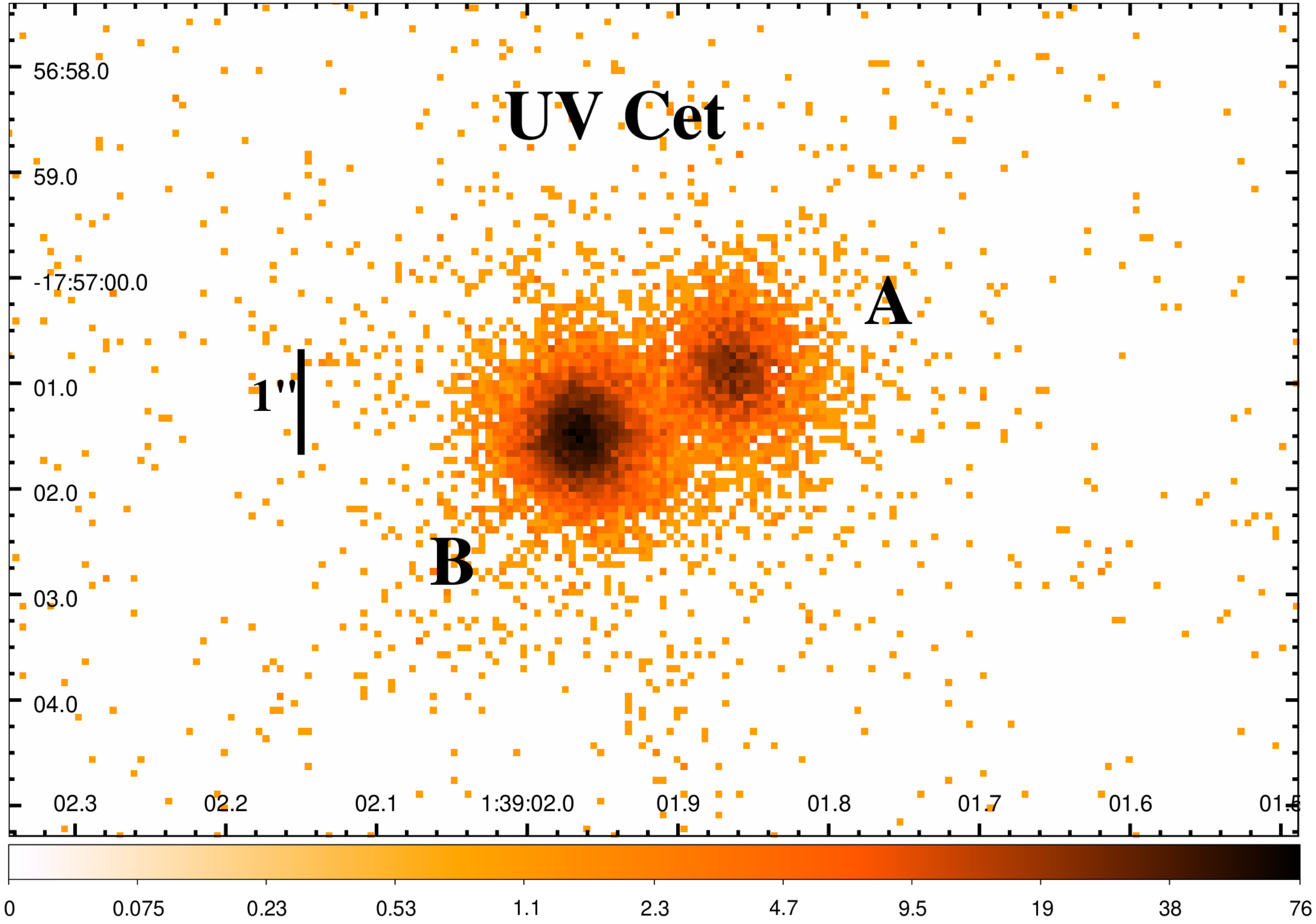
Wong et al. 2016, AoAS 10(2), 1107; Xu et al. 2021, AJ 161, 184

4D: change points in time across filter images



Example of detecting an evolving loop in the solar corona, where the loop is found in each of 6 filter images, and its growth and decay is identified over time. Uses MDL coupled with seeded region growing.

4D: disambiguation of overlapping photons



Probabilistically assign photons to one of several overlapping point sources by leveraging their spatial, spectral, and temporal patterns

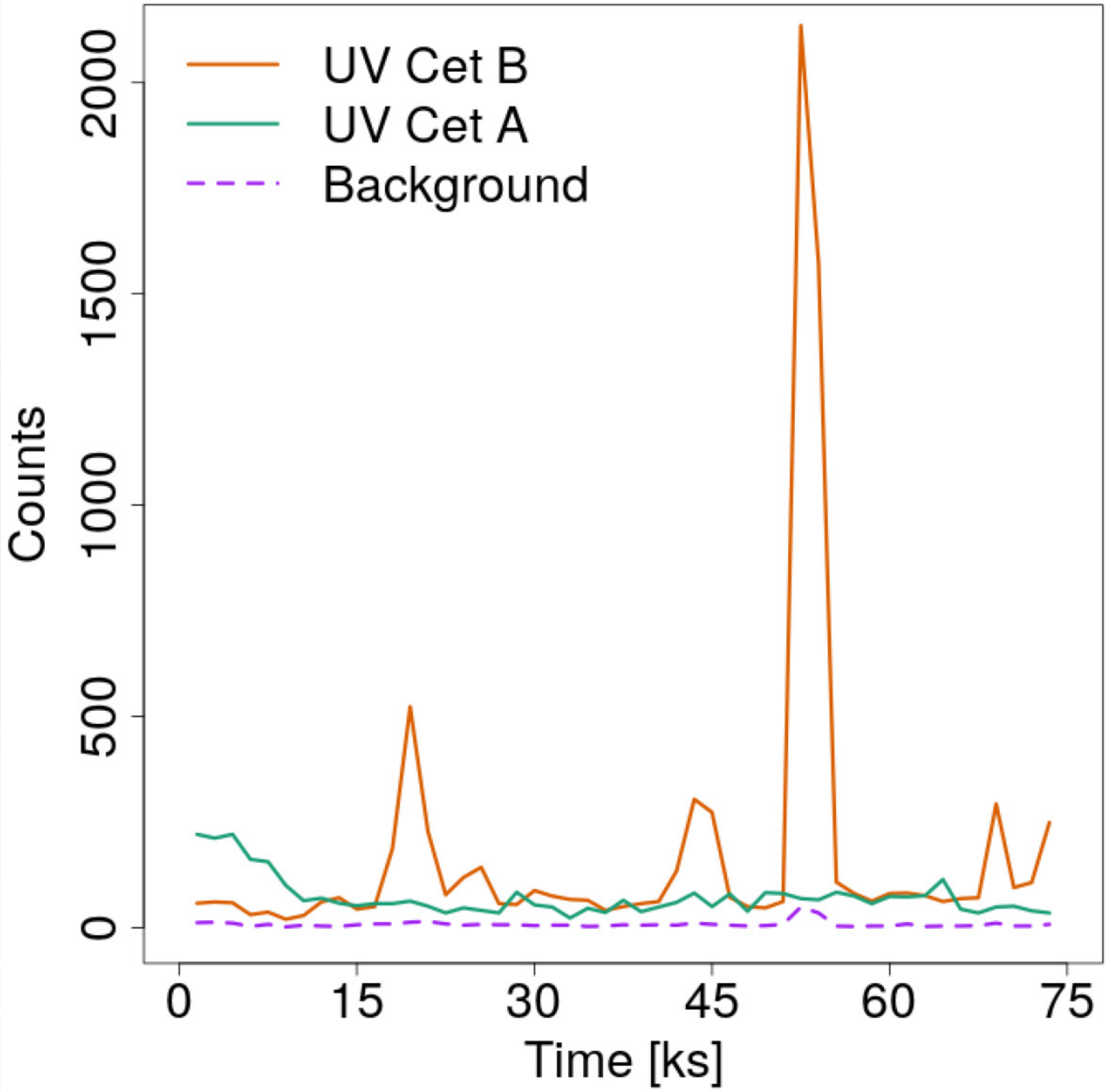
$\{x,y,E\}$ — BASCS (Jones et al. 2015)
$\{x,y,t,E\}$ — EBASCS (Meyer et al. 2021)

Finite Mixture model where each event is assumed to arise from one of several sources with the mixture weights representing proportion of photons from that source.

Each event is assigned a probability of belonging to each source and sifted, and the the sources are probabilistically separated.

Jones et al. 2015, ApJ 808, 137; Meyer et al. 2021, MNRAS 506, 6160

4D: disambiguation of overlapping photons



$L_{\text{Xpeak}} \approx 2{\cdot}10^{29}$ erg s^{-1} $\approx$ X1000
$L_{\text{Xmin}} \approx 10^{27}$ erg s^{-1} $\approx$200× increase
(Audard et al. 2003)