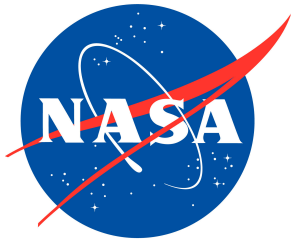
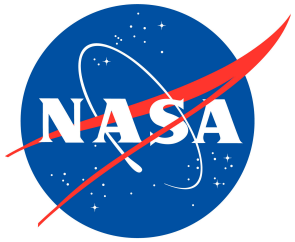


Bayesian Blocks for Particle Physics

Jeff Scargle
NASA Ames Research Center
Fermi Gamma Ray Space Telescope

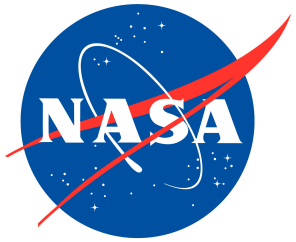
Special Thanks to Jim Chiang, Jay Norris, Brad Jackson
Banff Discovery Workshop, July 2010



Data Segmentation for Particle Physics

Jeff Scargle
NASA Ames Research Center
Fermi Gamma Ray Space Telescope

Special Thanks to Jim Chiang, Jay Norris, Brad Jackson
Banff Discovery Workshop, July 2010



Two Constructs:

- Segmentation of Data Spaces into Data Cells & Blocks
- Edelson and Krolik algorithm:
  correlation functions of unevenly sampled data

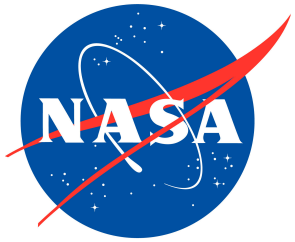
lead to many informative functions:

- Data adaptive histograms and time series representations
- Correlation functions
- Fourier Power spectra (amplitude and phase)
- Structure Functions
- Wavelet Power spectra (scalegrams)
- Time-Scale/Time Frequency Distributions

… in auto- and cross- modes

… for all data modes (events, counts in bins, measurements, etc.)

… for arbitrarily sampled data



Two Constructs:

- Segmentation of Data Spaces into Data Cells & Blocks
- Edelson and Krolik algorithm:
  correlation functions of unevenly sampled data

lead to many informative functions

- Data adaptive histograms and time series representations
- Correlation functions
- Fourier Power spectra (amplitude and phase)
- Structure Functions
- Wavelet Power spectra (scalegrams)
- Time-Scale/Time Frequency Distributions

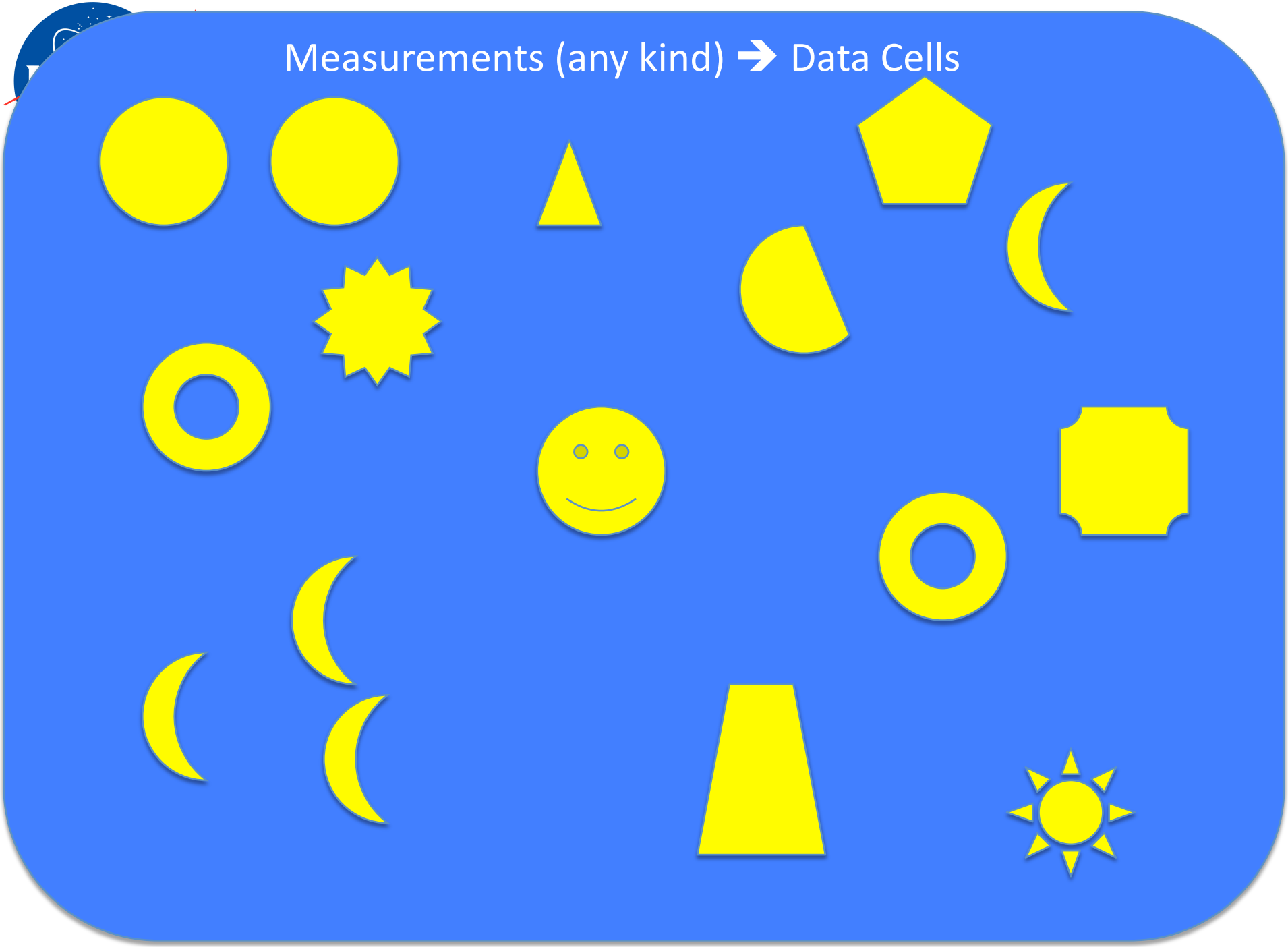
… in auto- and cross- modes

… for all data modes (events, counts in bins, measurements, etc.)

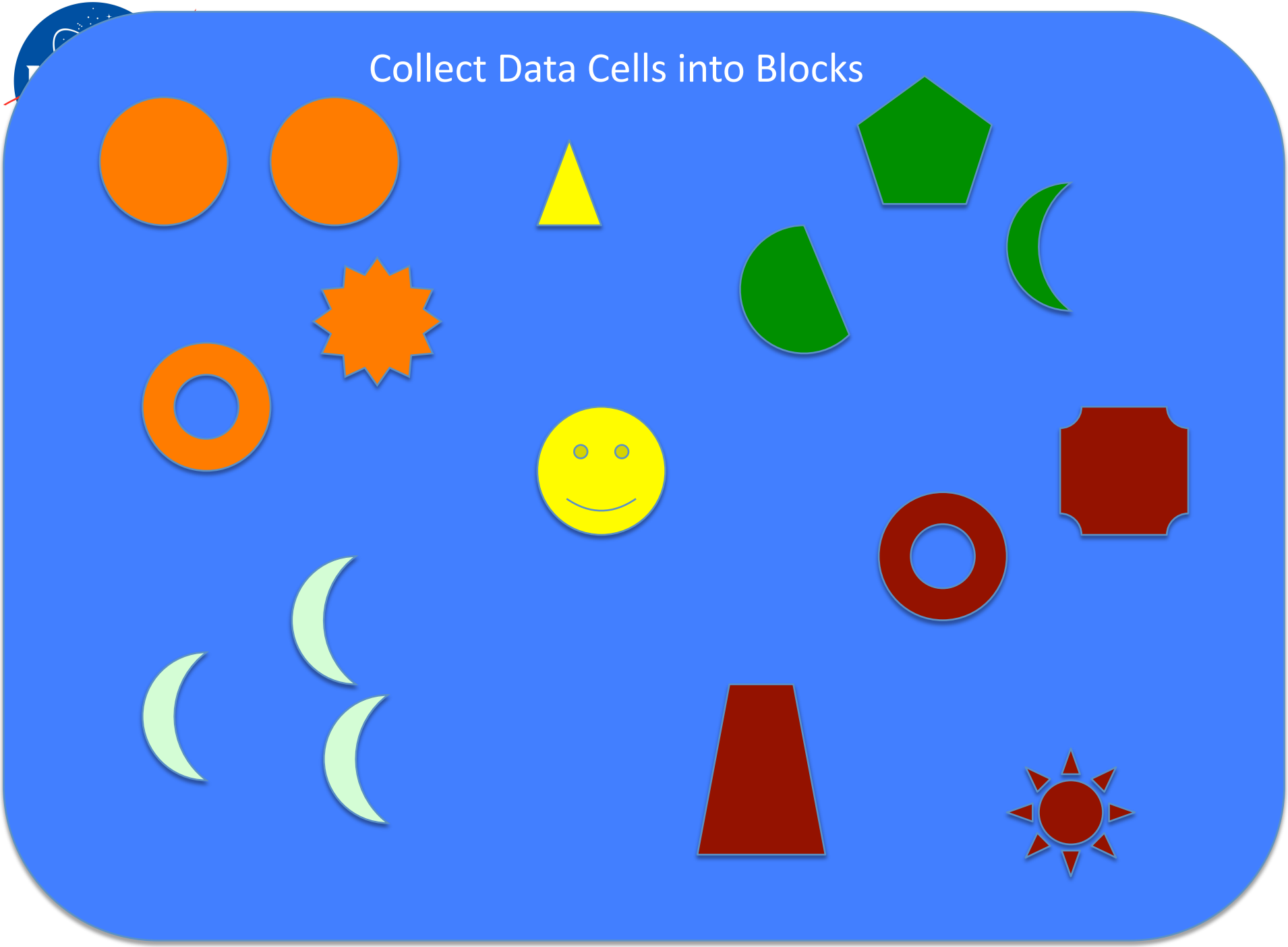
… for arbitrarily sampled data

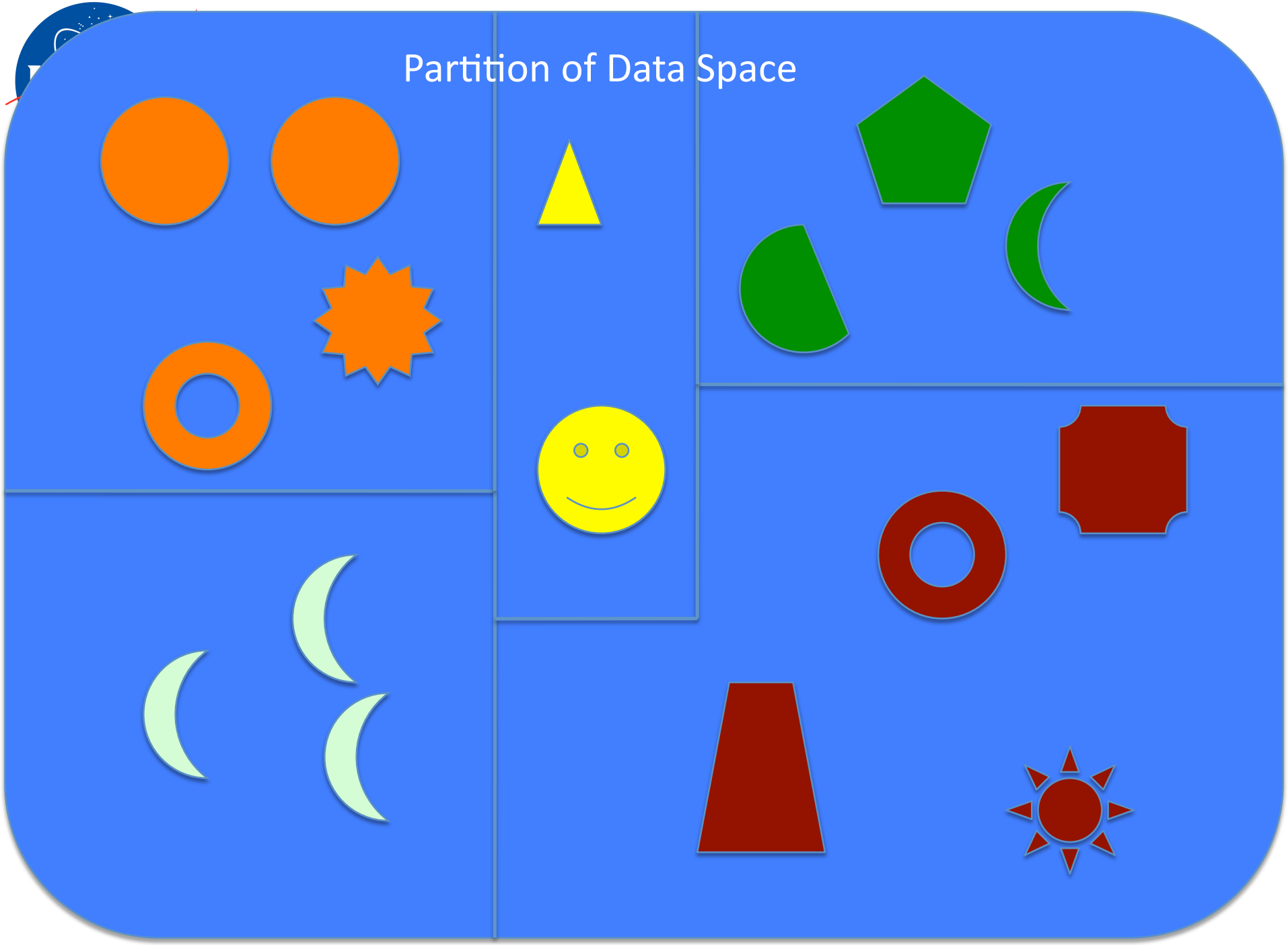
Data Space – Any Dimension

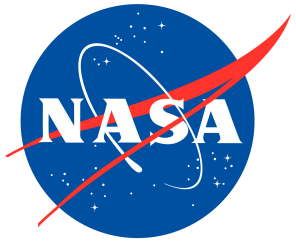
Measurements (any kind) ➔ Data Cells



Collect Data Cells into Blocks




Partition of Data Space

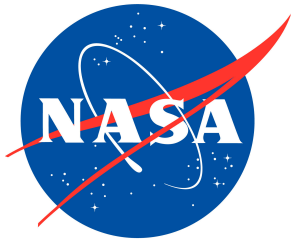


The Bin Myths

I. Point data must be binned in order to make sense out of them.

II. Bins must be equal in size.

III. The bins must be large enough so that each bin has a “statistically significant” sample.



Smoothing and Binning

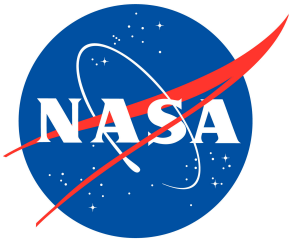
Old views: the best (only) way to reduce noise is to smooth the data
the best (only) way to deal with point data is to use bins

New philosophy: smoothing and binning should be avoided because they ...

- discard information
- degrade resolution
- introduce dependence on parameters:
  - degree of smoothing
  - bin size and location

Wavelet Denoising (Donoho, Johnstone) multiscale; no explicit smoothing
Adaptive Kernel Smoothing

Optimal Segmentation (*e.g.* Bayesian Blocks) Omni-scale -- uses neither explicit smoothing nor pre-defined binning



Bayesian Blocks

Piecewise-constant Model of Time Series Data

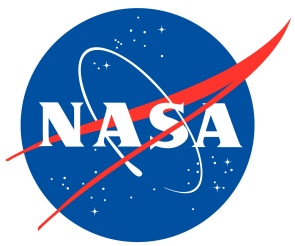
Optimum Partition of Interval, Maximizing Fitness Of Step Function Model

Segmentation of Interval into Blocks, Representing Data as Constant In the Blocks -- within Statistical Fluctuations

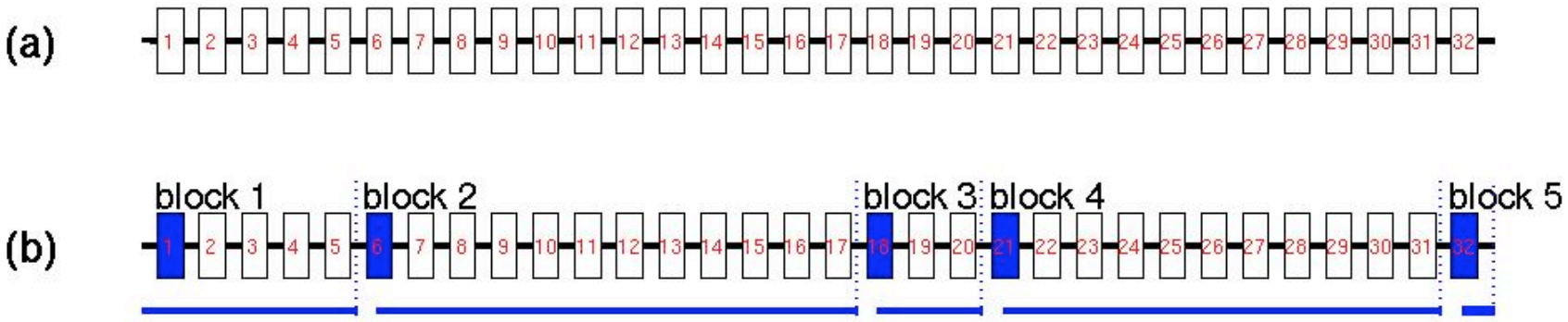
Histogram in Unequal Bins -- not Fixed A Priori but determined by Data

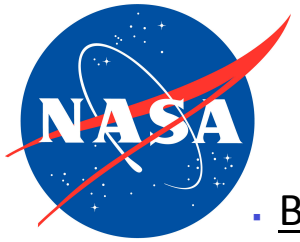
Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, a New Method to Analyze Structure in Photon Counting Data, Ap. J. 504 (1998) 405.

An Algorithm for the Optimal Partitioning of Data on an Interval," *IEEE Signal Processing Letters, 12 (2005) 105-108.*



Simple Example of 1D Data Cells and Blocks





Fitness Functions

- Block likelihood = product of likelihoods of its cells

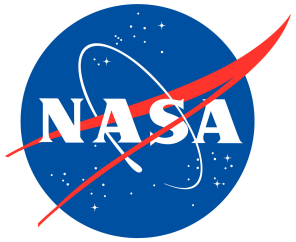
- Block Likelihood depends on
  - N = The Number of Events in the Block
  - M = The Size of the Block

- Model likelihood = product of likelihoods of its blocks

- Remove the dependence on the block event rates:
  - Marginalize, or
  - Maximize the Likelihood

- Adopt prior distribution for N_b, the number of blocks. (Parameter of this distribution acts like a smoothing parameter.)

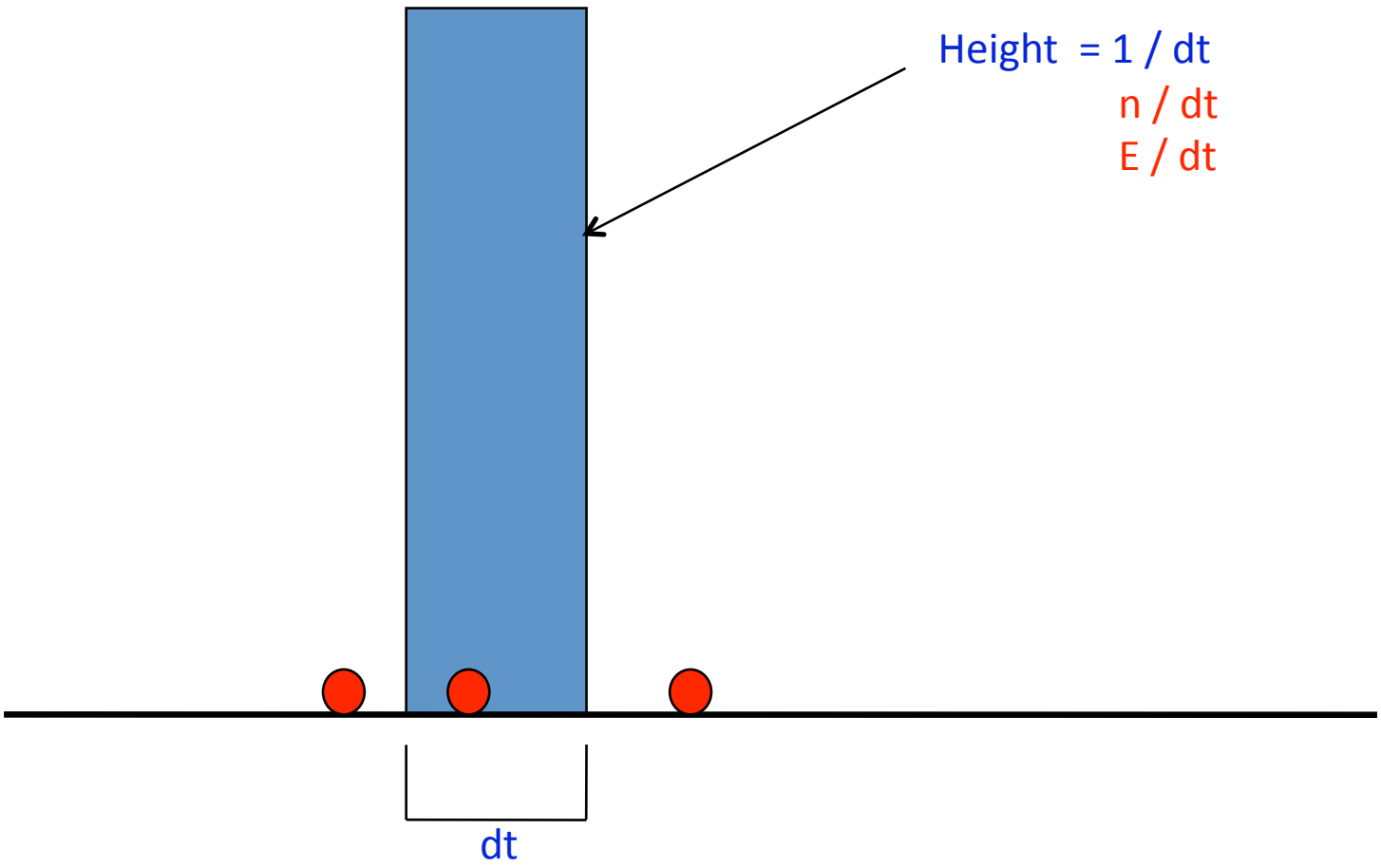
- Take log to yield an additive fitness function.

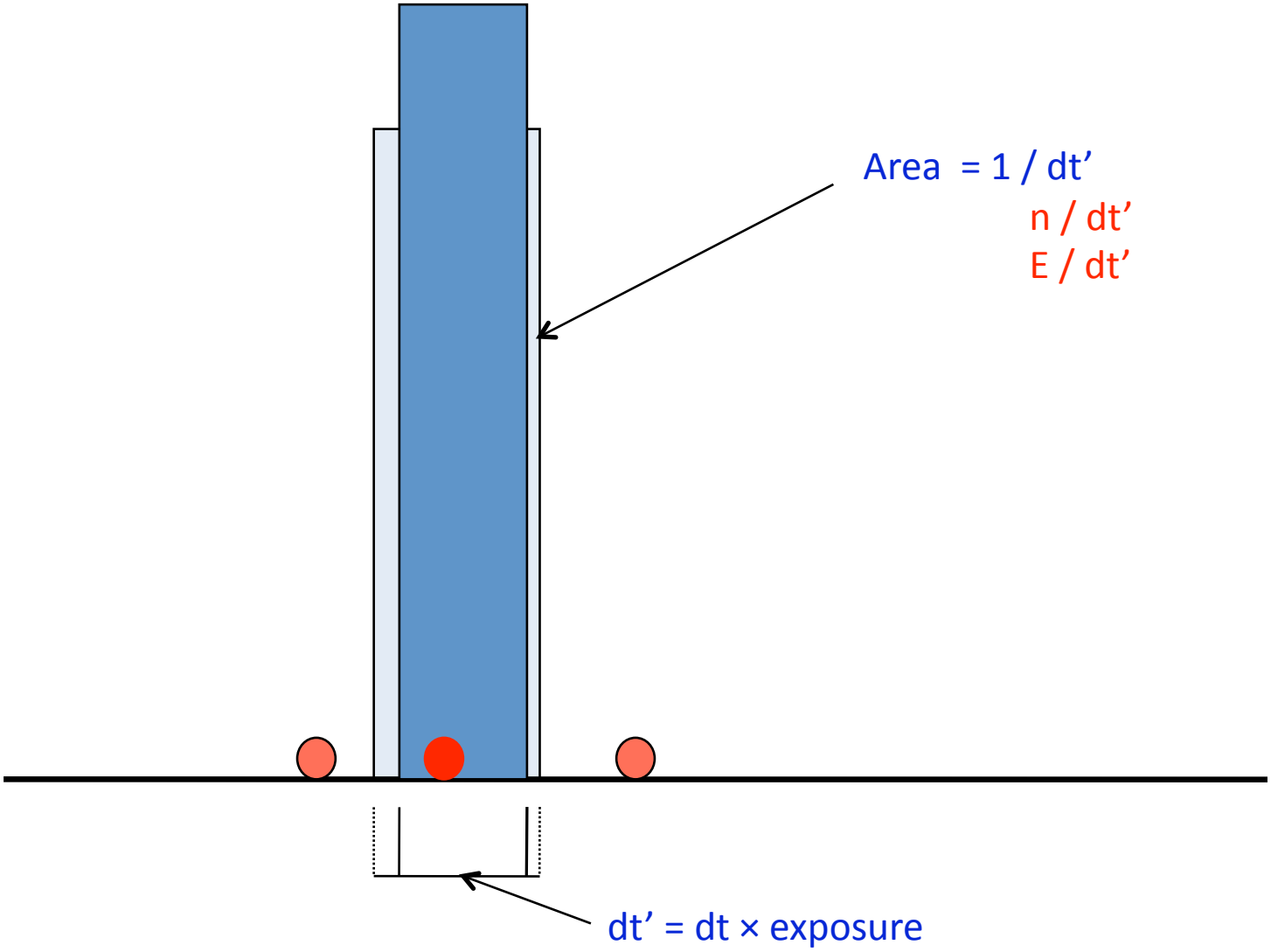


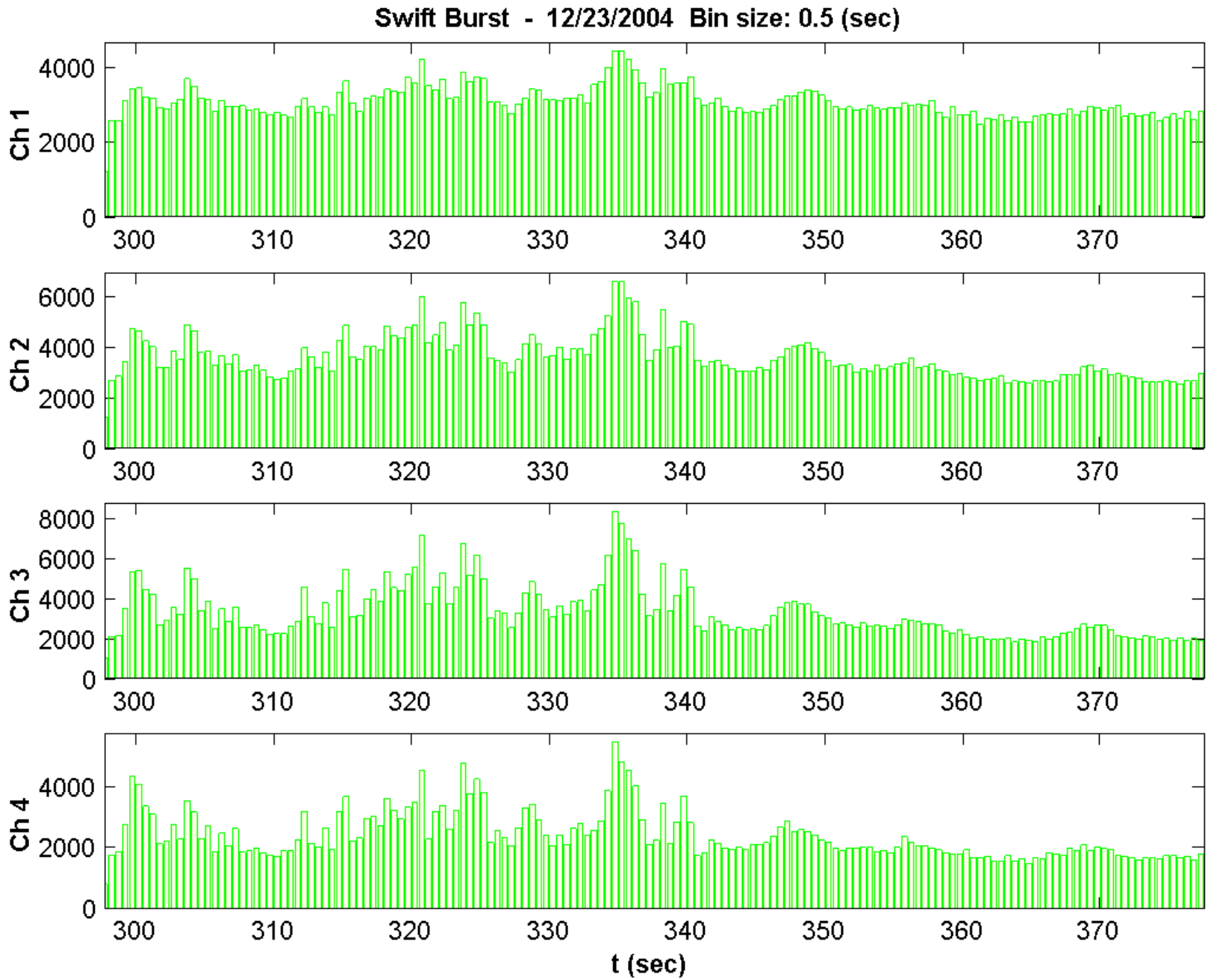
The Optimiser

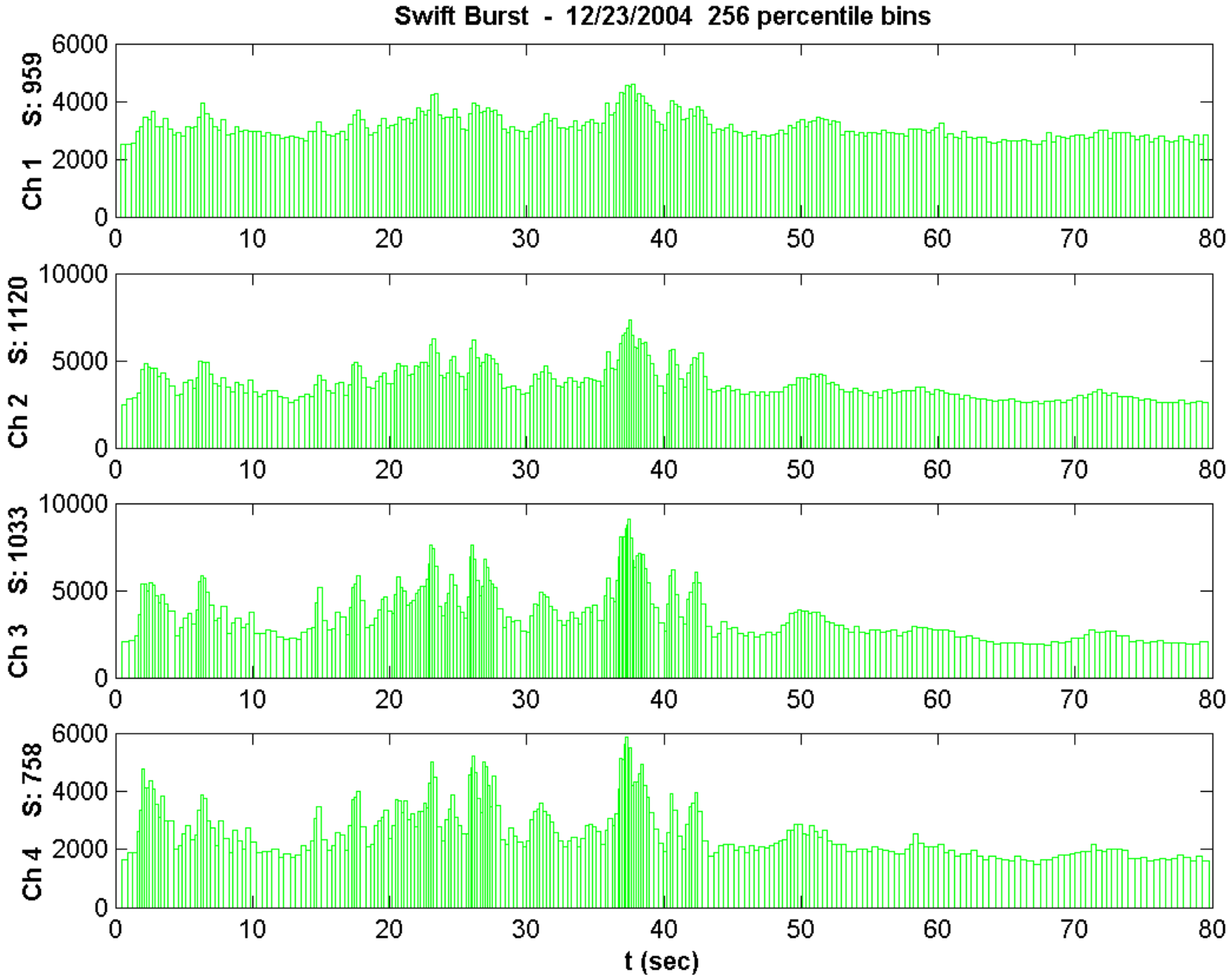
```
best = []; last = [];
for R = 1:num_cells
  [ best(R), last(R) ] = max( [0 best] + fitness( cumsum( data_cells(1:R, :) ) );

    if first > 0 & last(R) > first    % Option: trigger on first significant block
        changepoints = last(R); return
    end

end

% Now locate all the changepoints
index = last( num_cells );
changepoints = [];

while index > 1
  changepoints = [ index changepoints ];
   index = last( index - 1 );
end
```

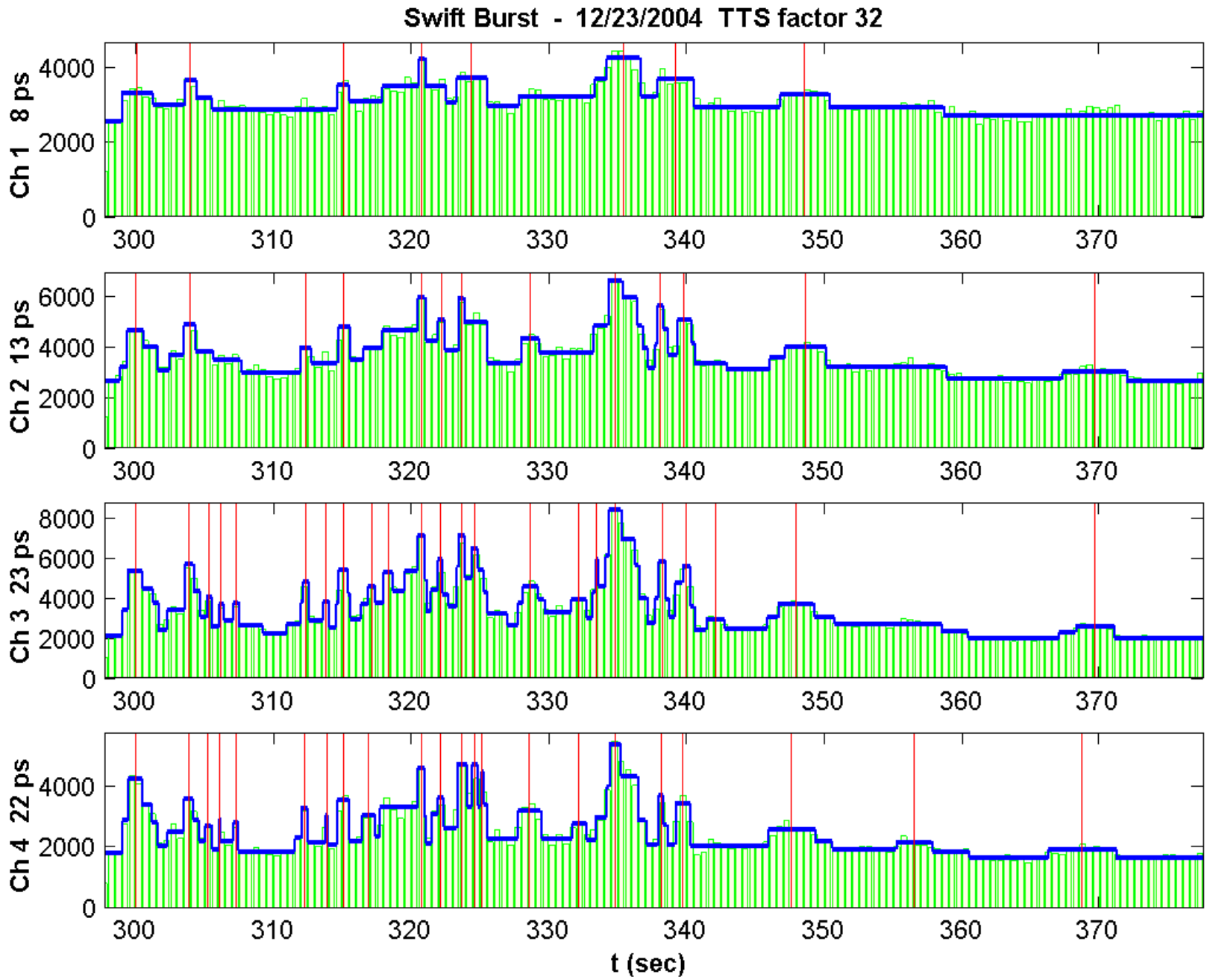
Do not use at home: a few details omitted!

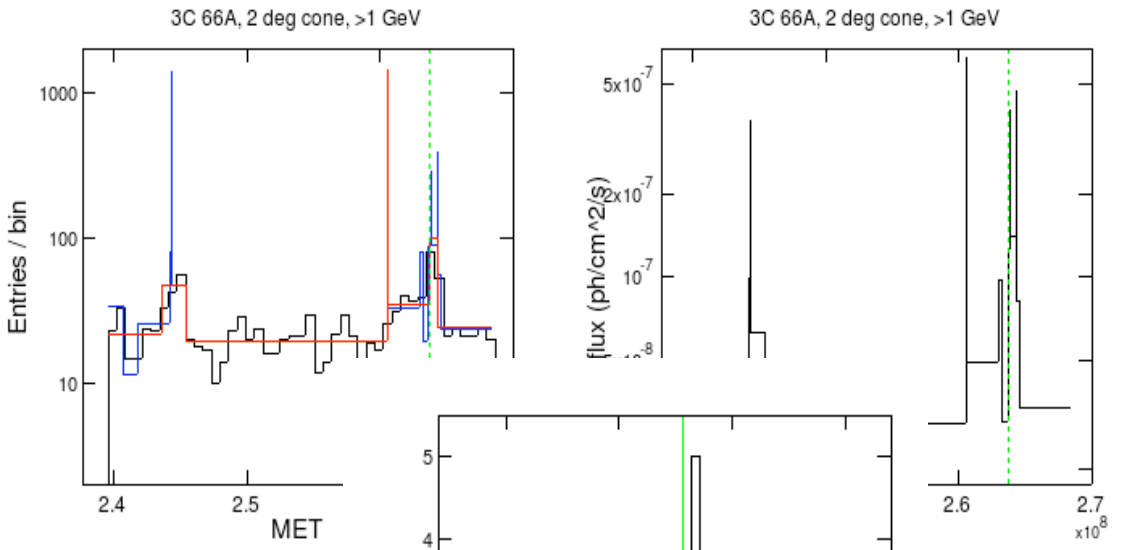

Height = 1 / dt
n / dt
E / dt
dt


Area = 1 / dt'
n / dt'
E / dt'
dt' = dt × exposure

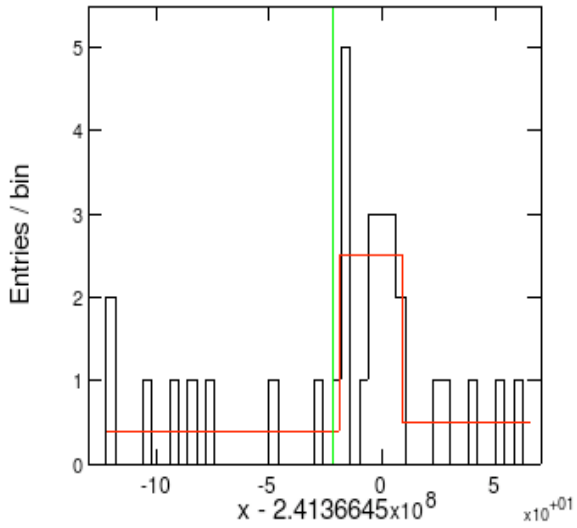

Swift Burst - 12/23/2004 Bin size: 0.5 (sec)
Ch 1
Ch 2
Ch 3
Ch 4
0
2000
4000
6000
8000
300
310
320
330
340
350
360
370
t (sec)

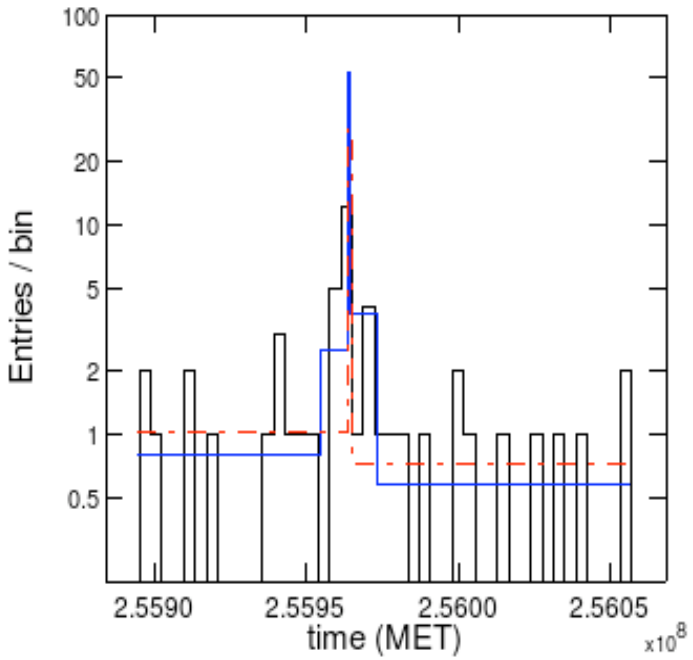

Swift Burst - 12/23/2004 256 percentile bins
Ch 1 S: 959
Ch 2 S: 1120
Ch 3 S: 1033
Ch 4 S: 758
0
2000
4000
6000
5000
10000
10
20
30
40
50
60
70
80
t (sec)

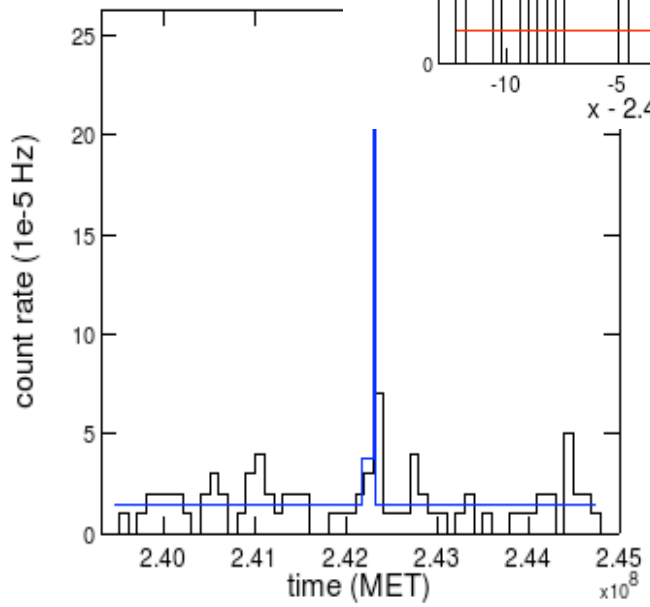

Swift Burst - 12/23/2004 TTS factor 32
Ch 1 8 ps
Ch 2 13 ps
Ch 3 23 ps
Ch 4 22 ps
0
2000
4000
6000
8000
300
310
320
330
340
350
360
370
t (sec)

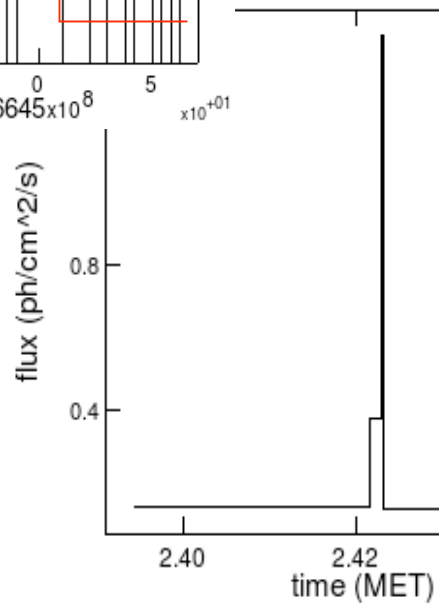

3C 66A, 2 deg cone, >1 GeV
Entries / bin
MET

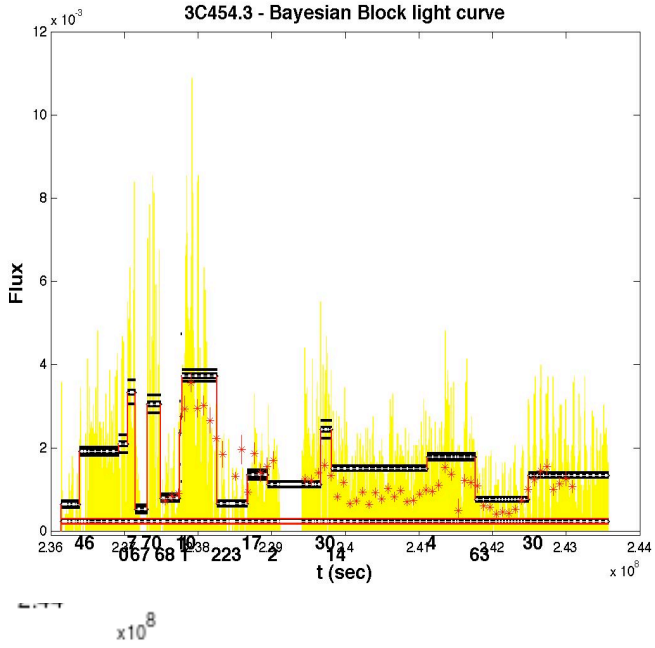
3C 66A, 2 deg cone, >1 GeV
flux (ph/cm^2/s)

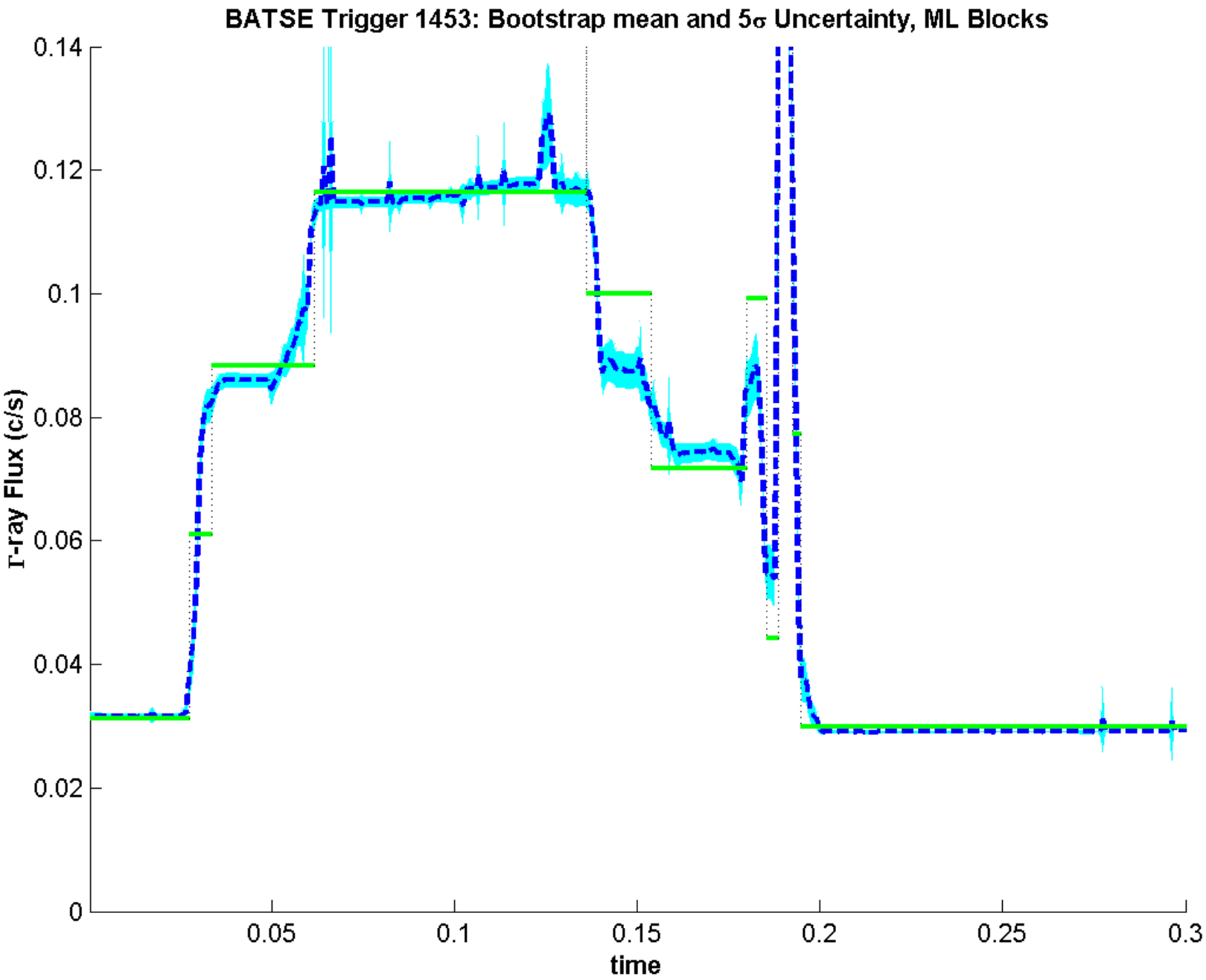

Entries / bin
x - 2.4136645x10^8

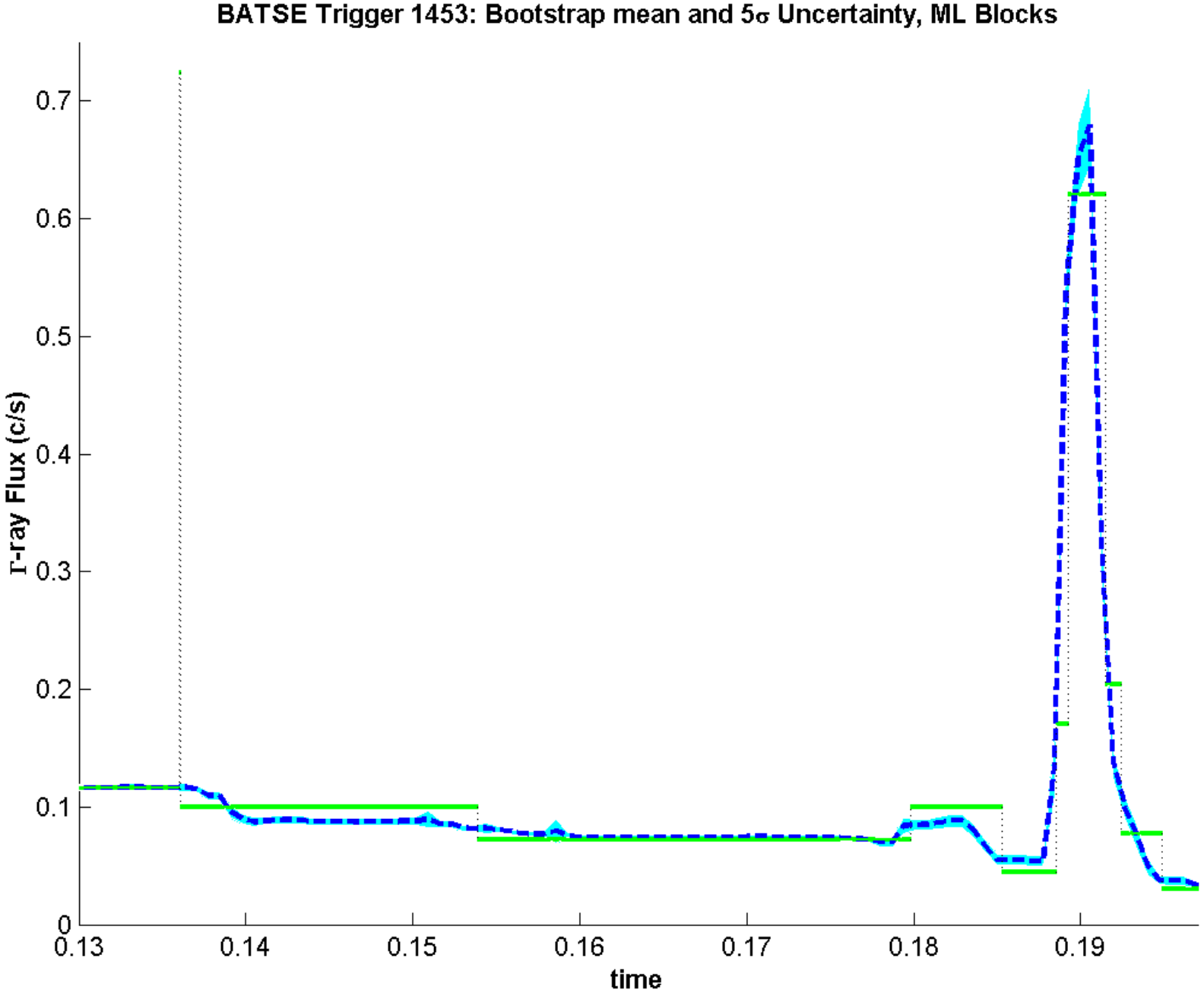

Entries / bin
time (MET)

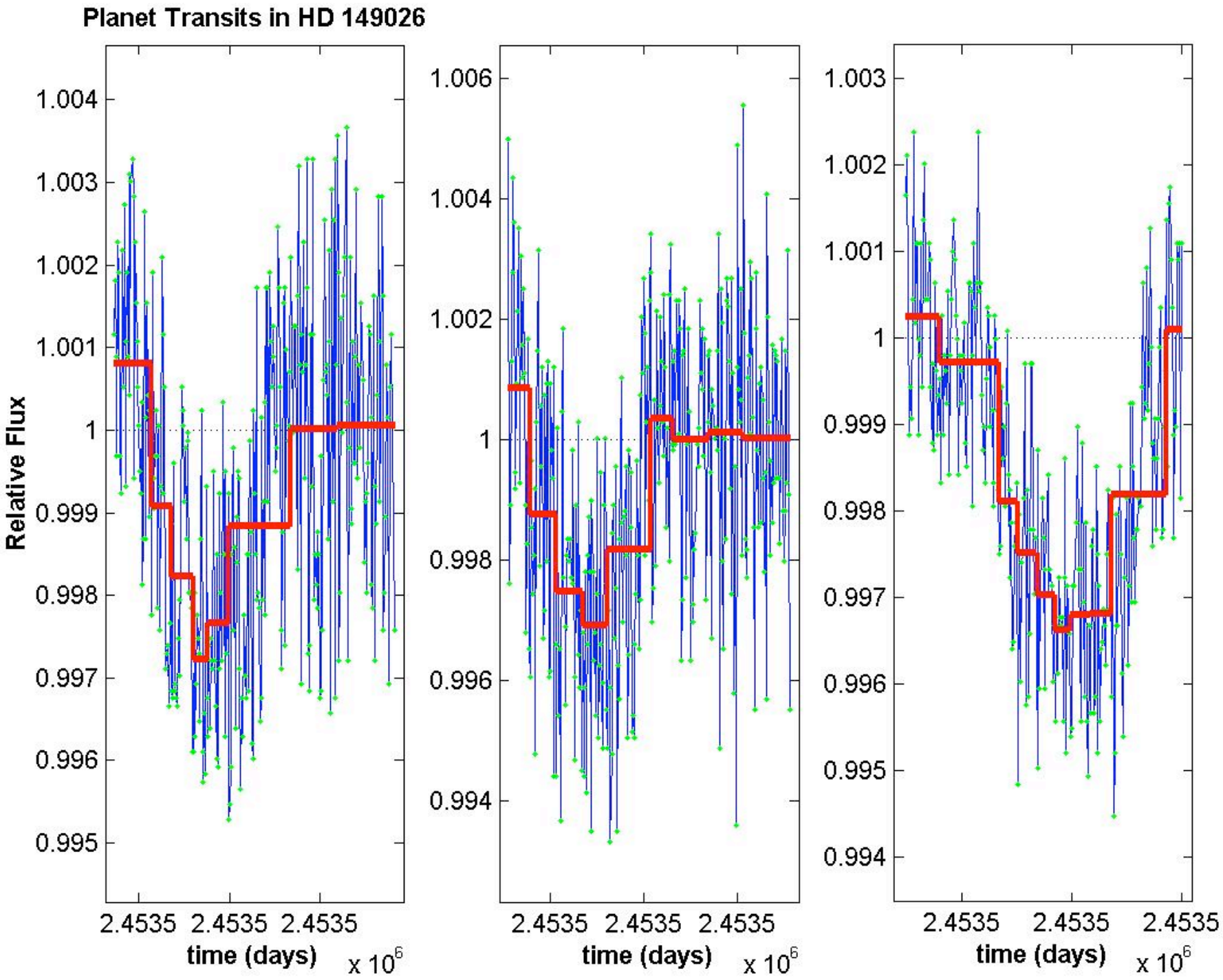

count rate (1e-5 Hz)
time (MET)


flux (ph/cm^2/s)
time (MET)

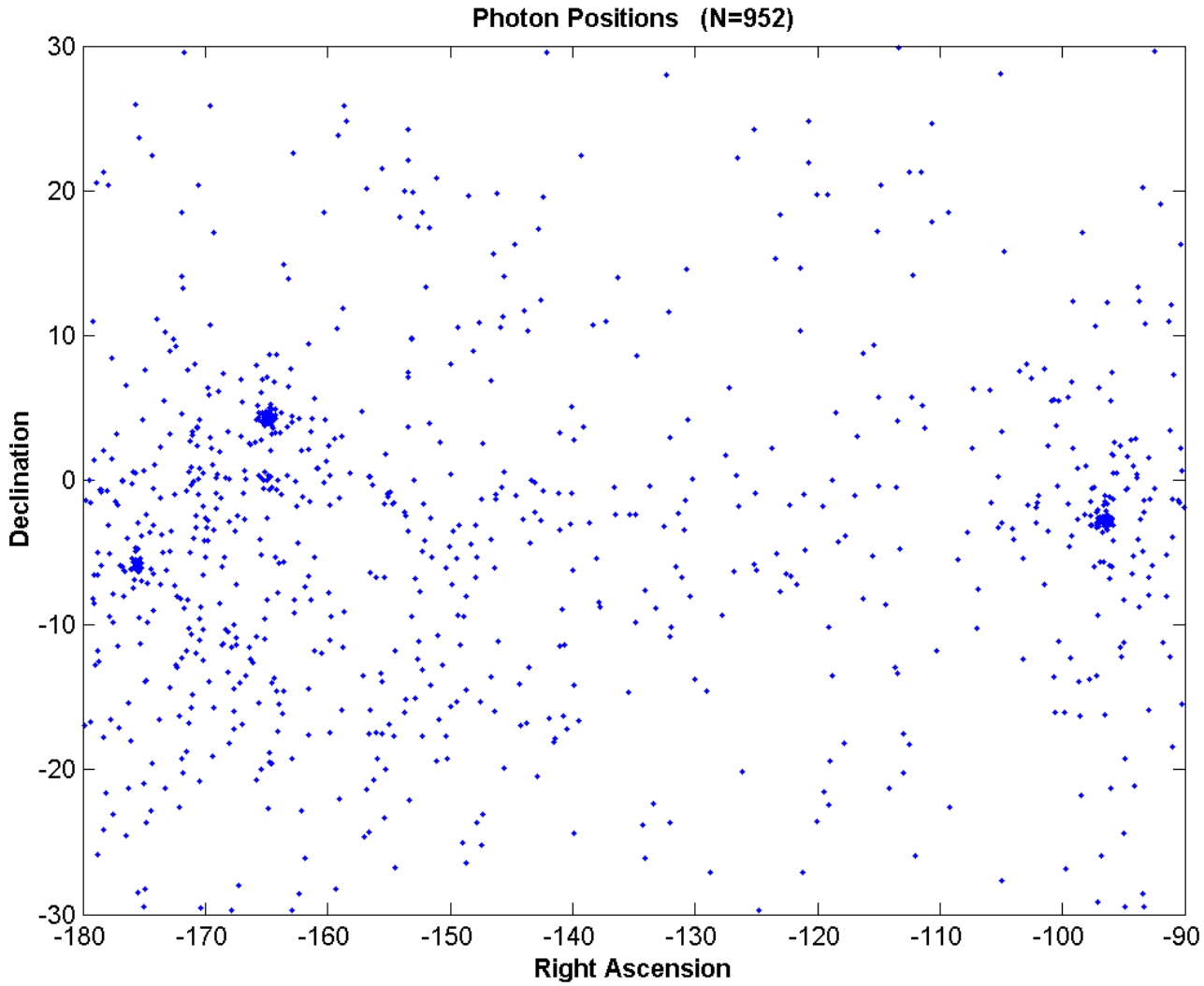

3C454.3 - Bayesian Block light curve
Flux
t (sec)

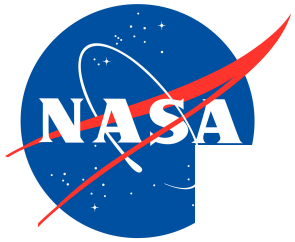

BATSE Trigger 1453: Bootstrap mean and 5σ Uncertainty, ML Blocks
Γ-ray Flux (c/s)
time
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0.05
0.1
0.15
0.2
0.25
0.3


BATSE Trigger 1453: Bootstrap mean and 5σ Uncertainty, ML Blocks
Γ-ray Flux (c/s)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.13
0.14
0.15
0.16
0.17
0.18
0.19
time

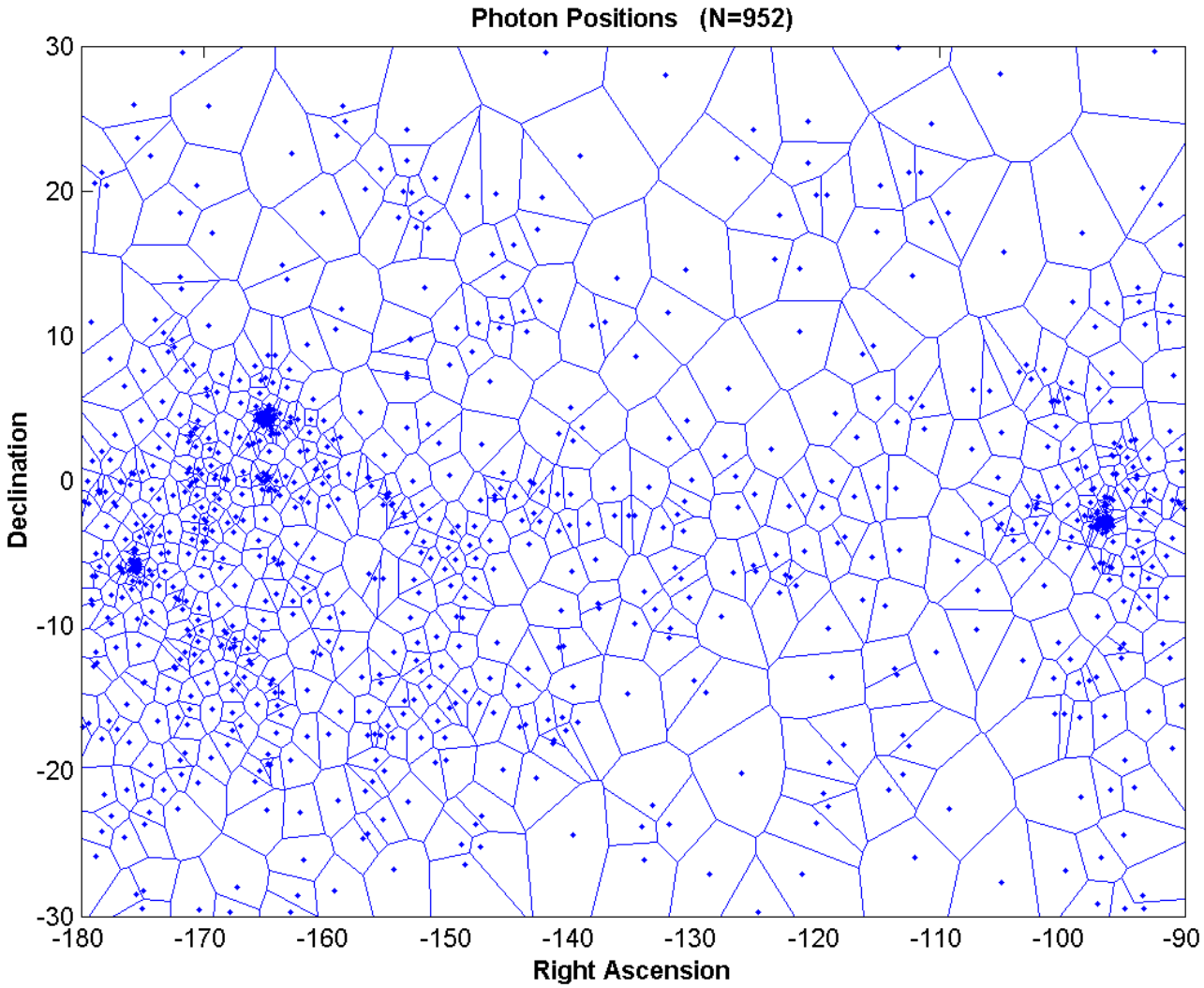

Planet Transits in HD 149026
Relative Flux
1.004
1.003
1.002
1.001
1
0.999
0.998
0.997
0.996
0.995
2.4535 2.4535 2.4535
time (days)
x 10^6
1.006
1.004
1.002
1
0.998
0.996
0.994
2.4535 2.4535 2.4535
time (days)
x 10^6
1.003
1.002
1.001
1
0.999
0.998
0.997
0.996
0.995
0.994
2.4535 2.4535 2.4535
time (days)
x 10^6

Voronoi Tessellation of data in any dimension

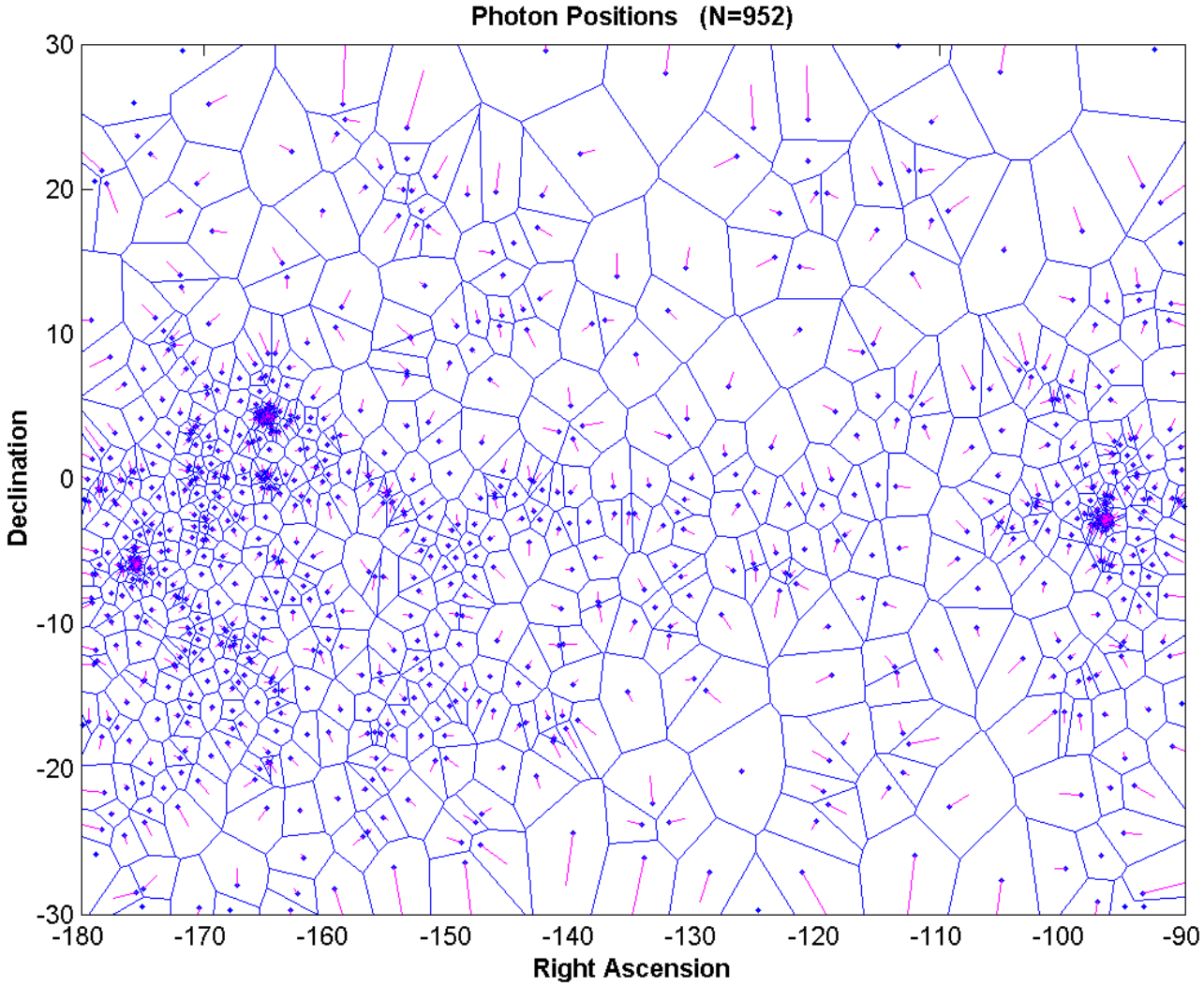


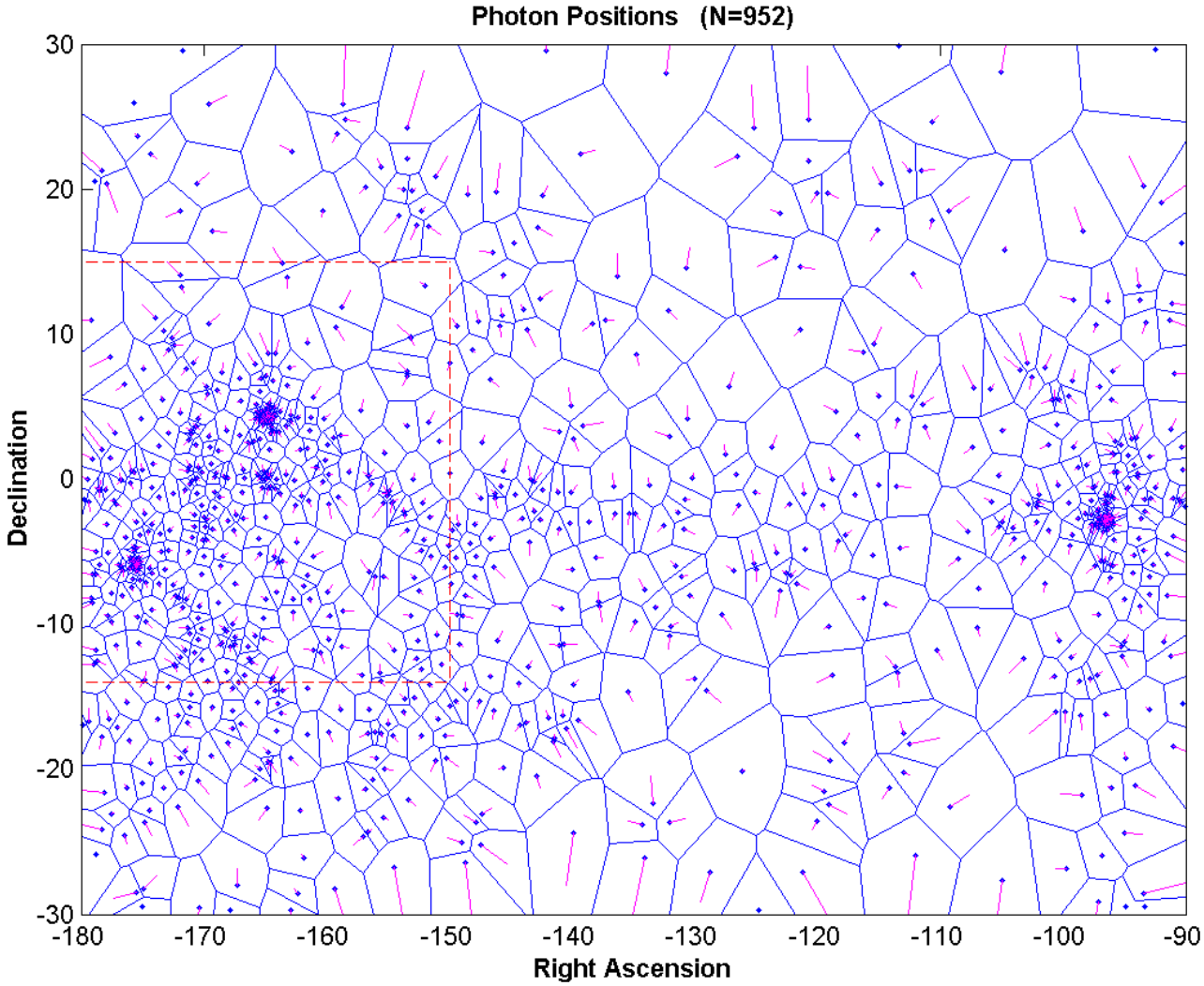


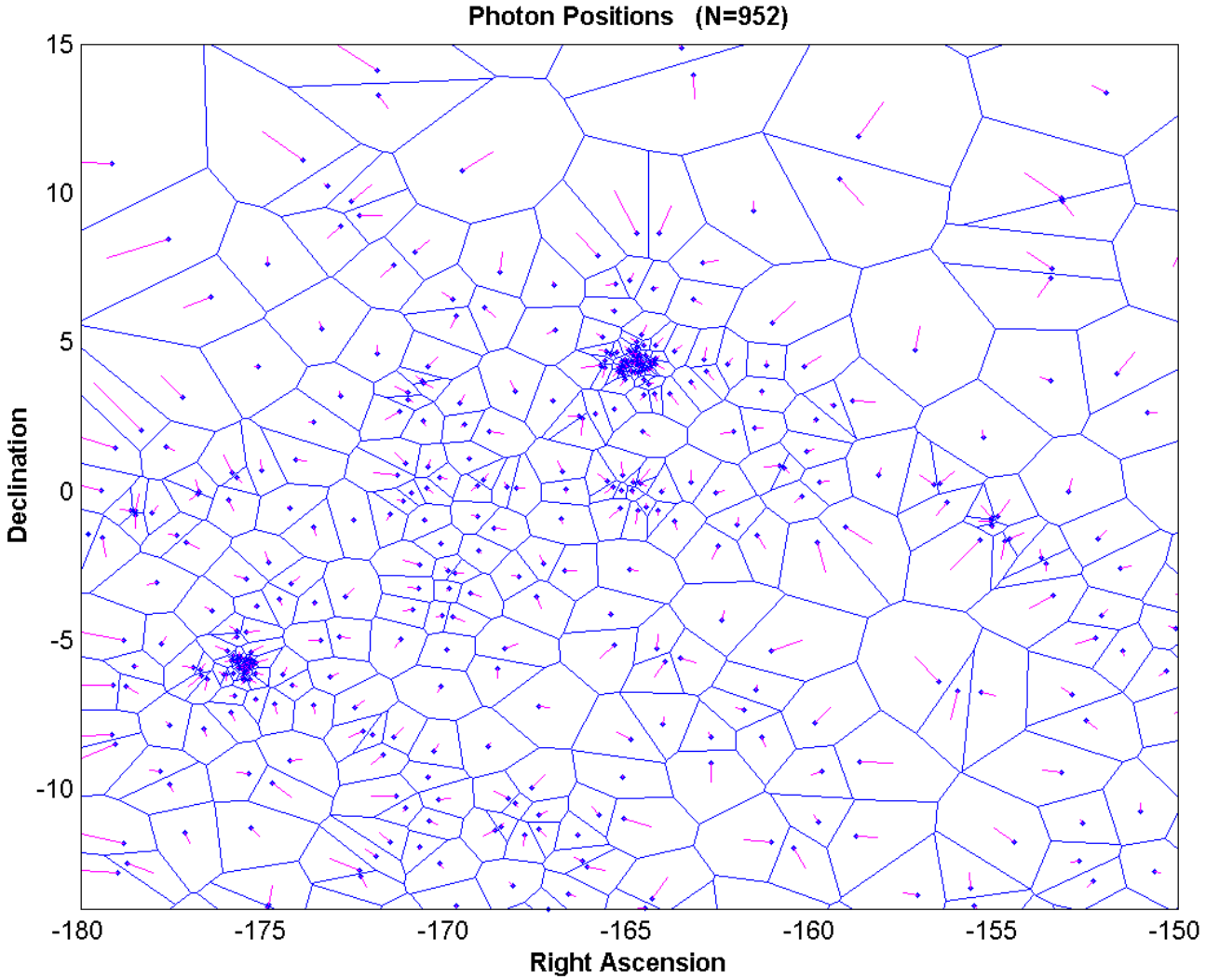
Construct Voronoi cells to represent local photon density

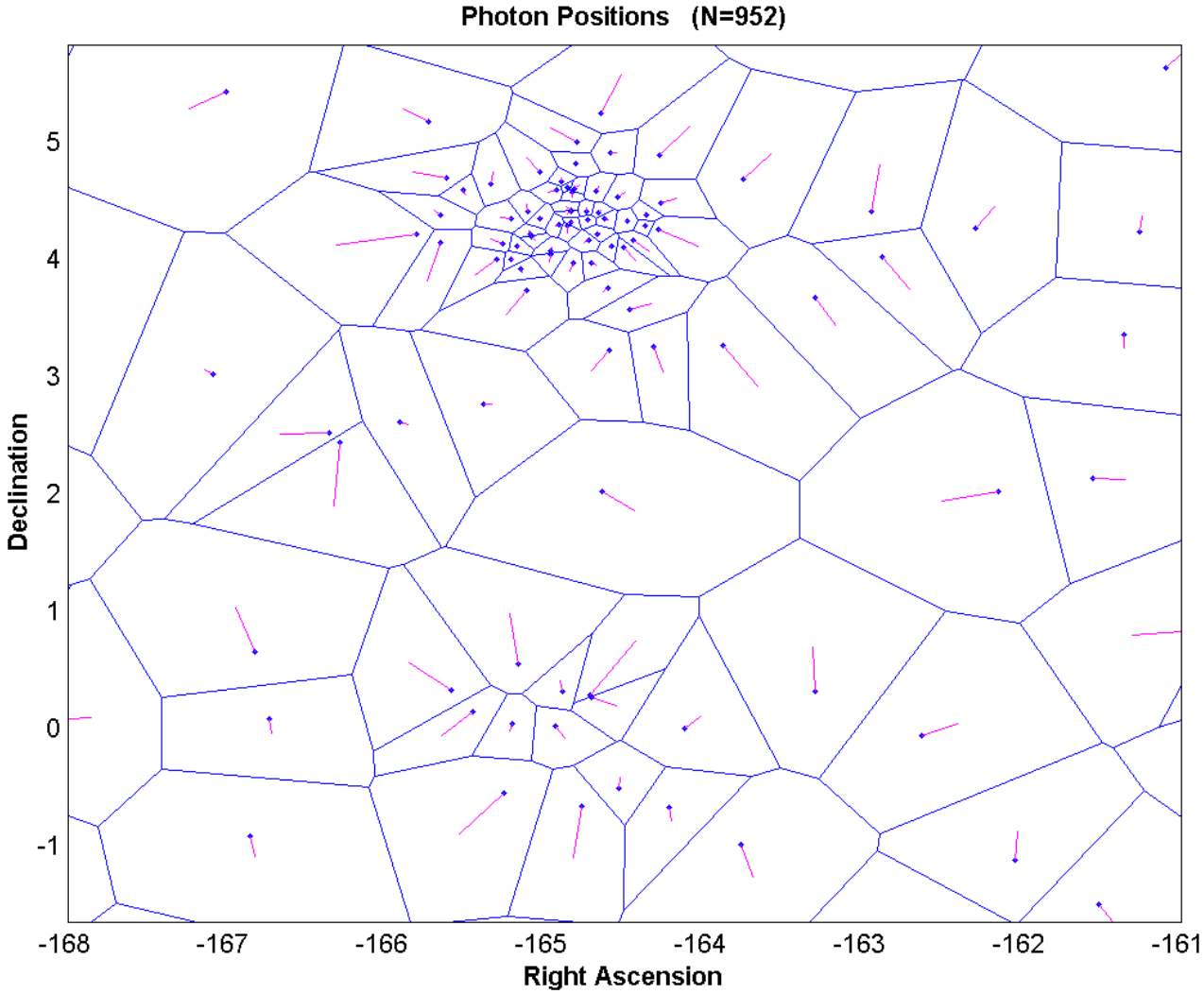


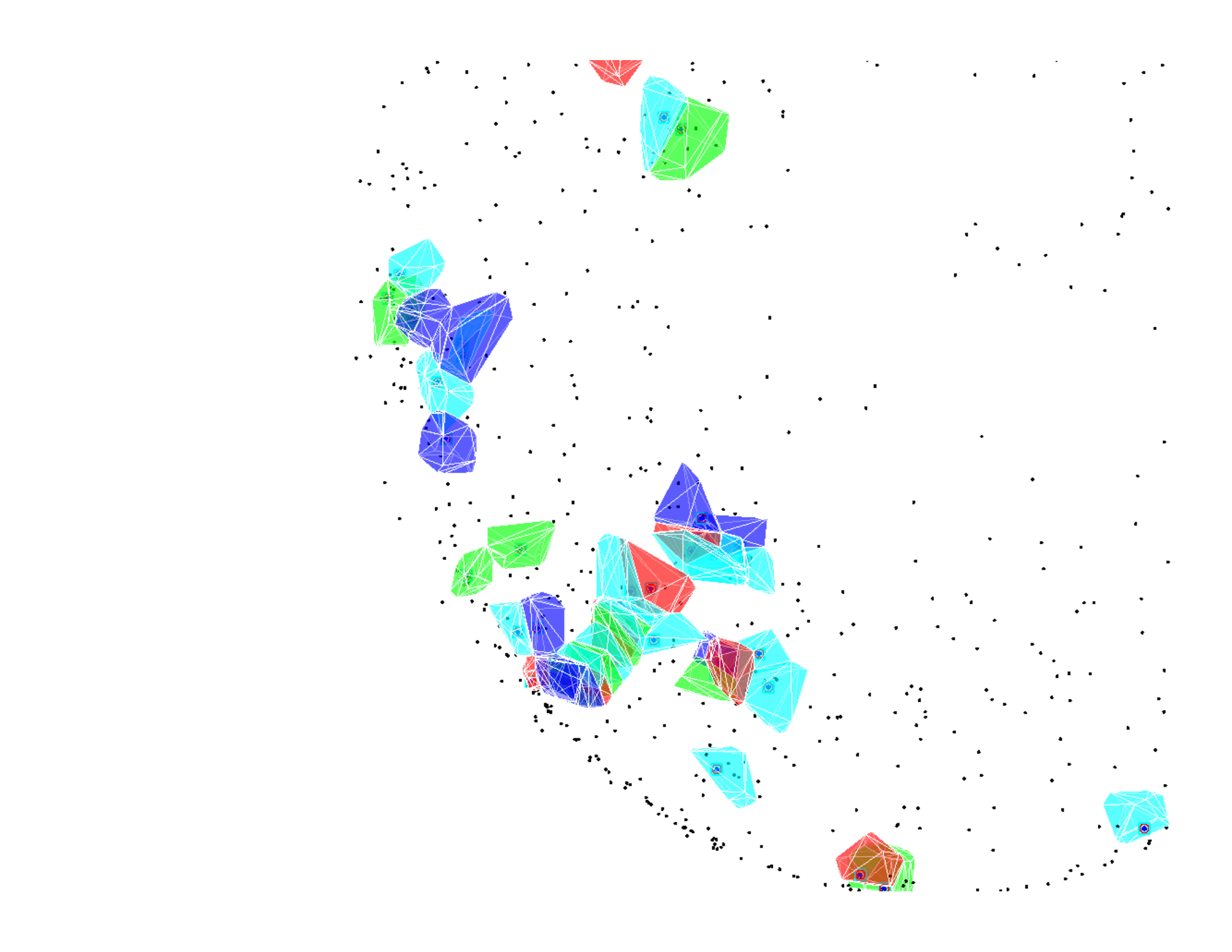
density ~ 1 / cell area

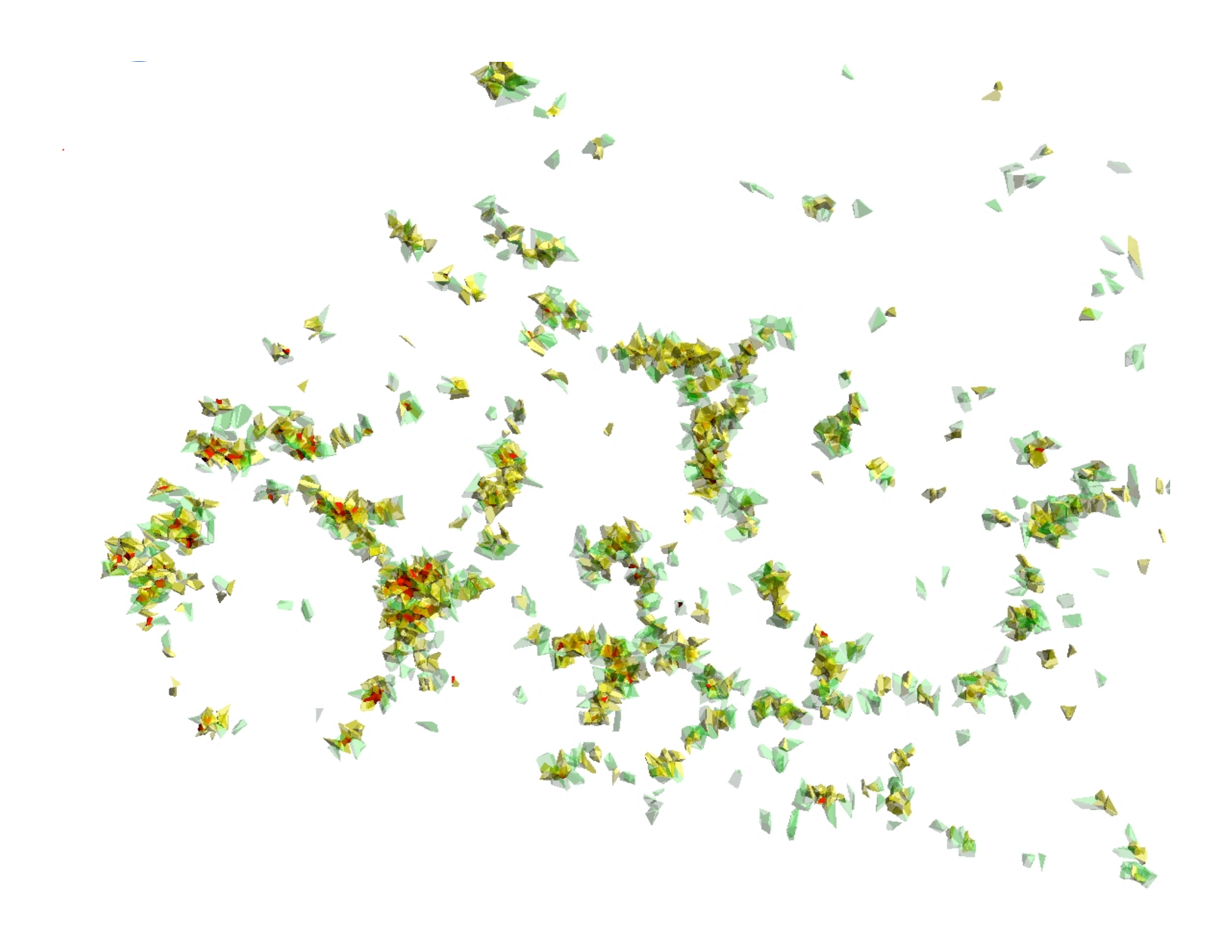

Photon Positions (N=952)
Declination
Right Ascension
30
20
10
0
-10
-20
-30
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90

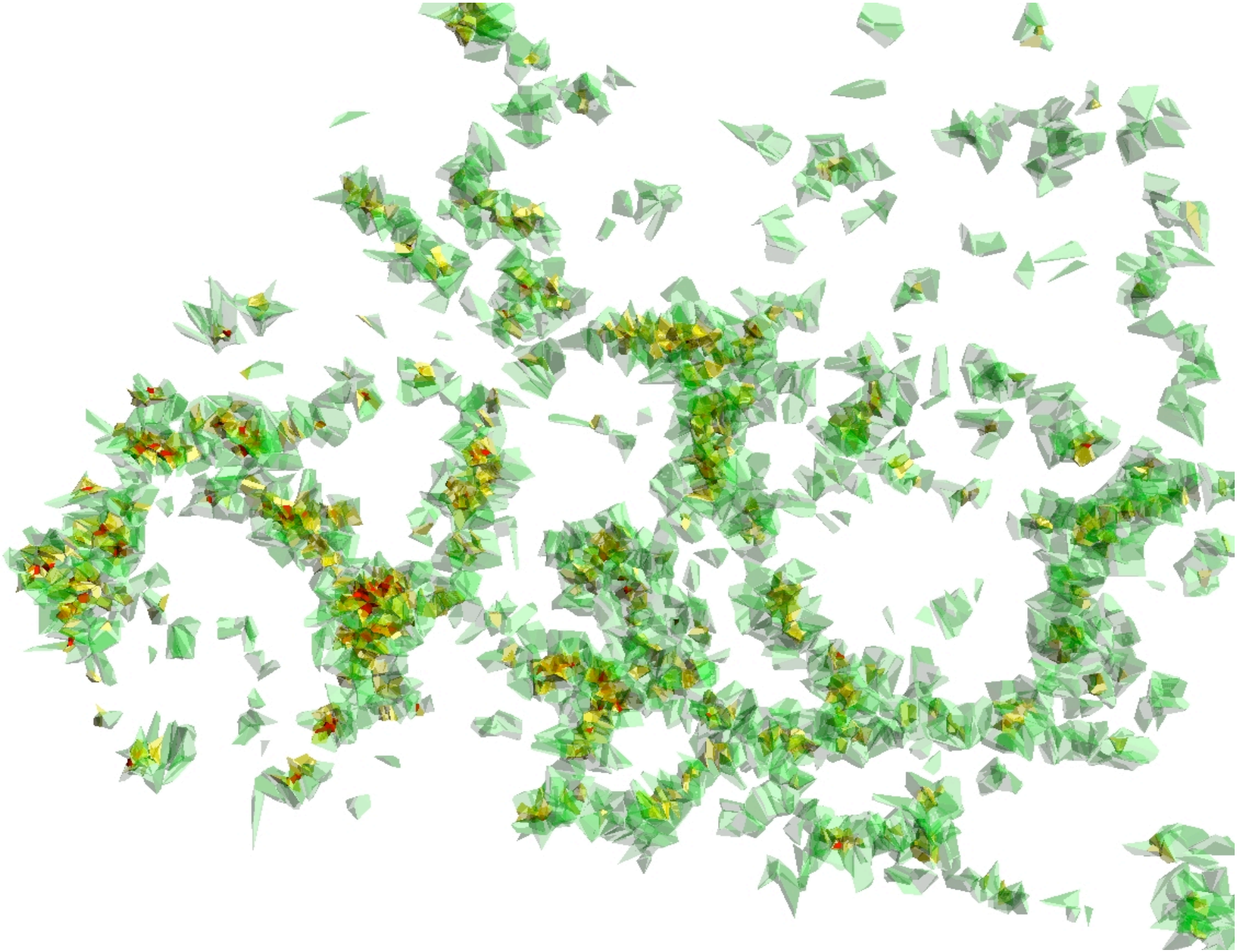

Photon Positions (N=952)
30
20
10
0
-10
-20
-30
Declination
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
Right Ascension

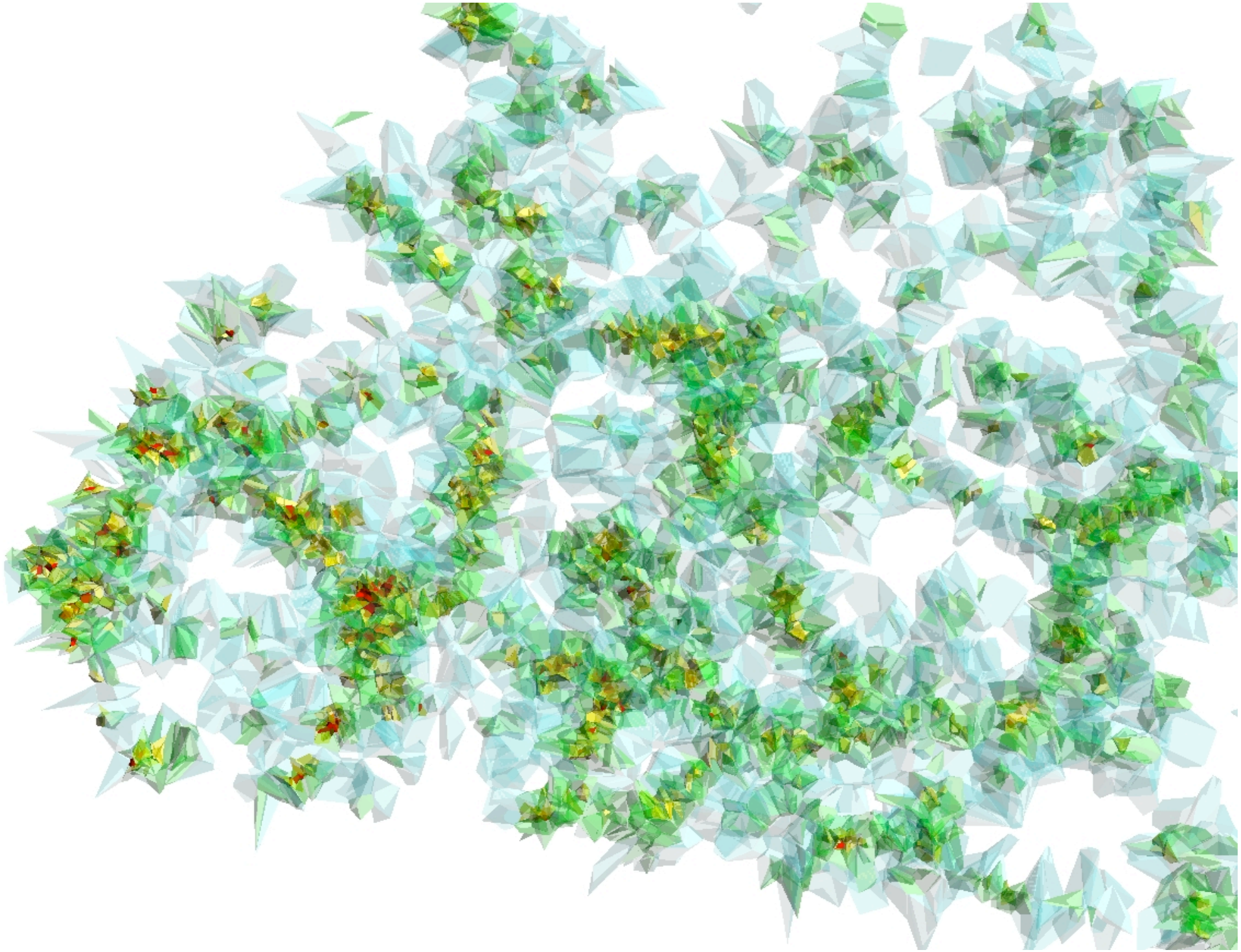

Photon Positions (N=952)
Declination
Right Ascension
15
10
5
0
-5
-10
-180
-175
-170
-165
-160
-155
-150


Photon Positions (N=952)
Declination
5
4
3
2
1
0
-1
-168
-167
-166
-165
-164
-163
-162
-161
Right Ascension

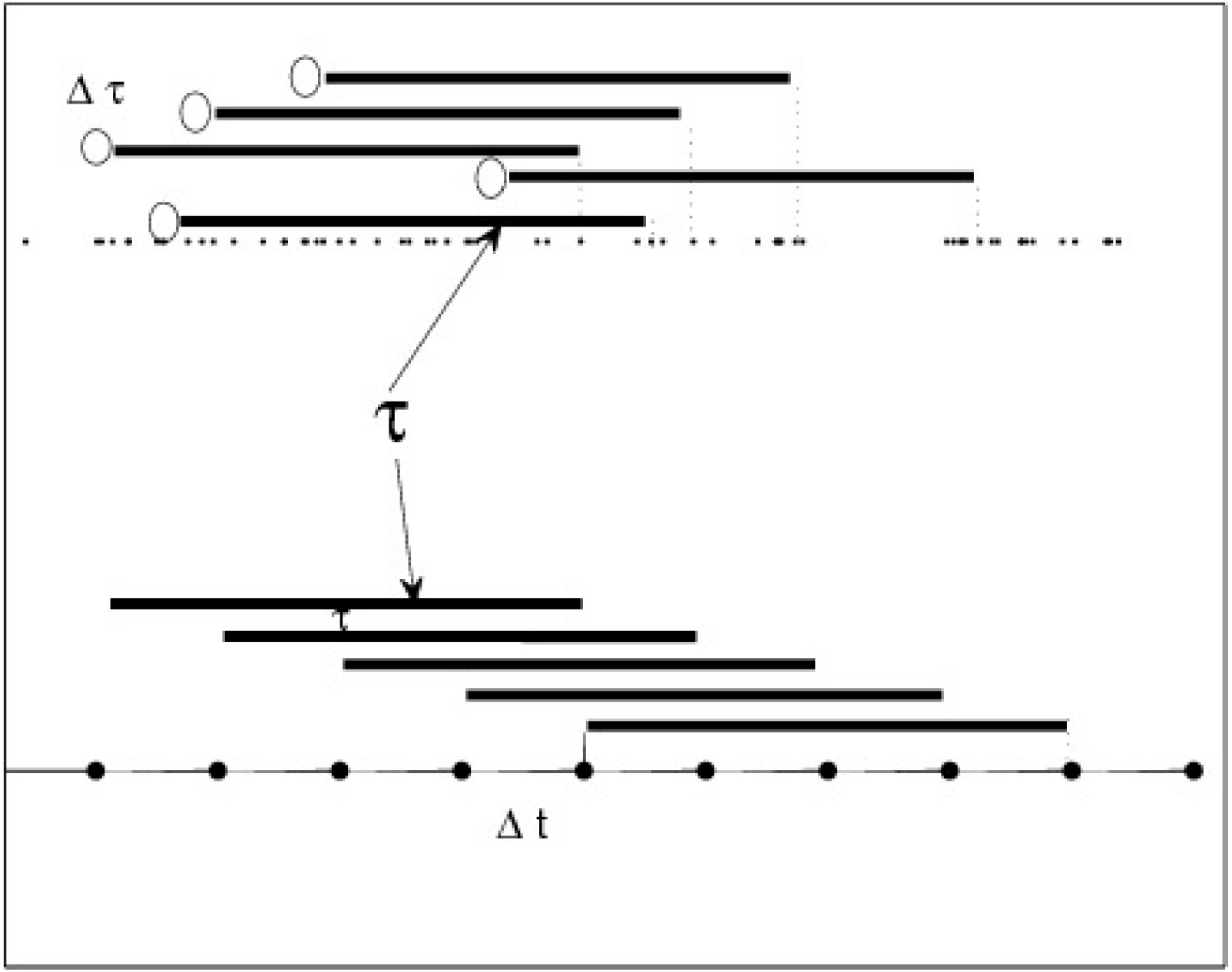


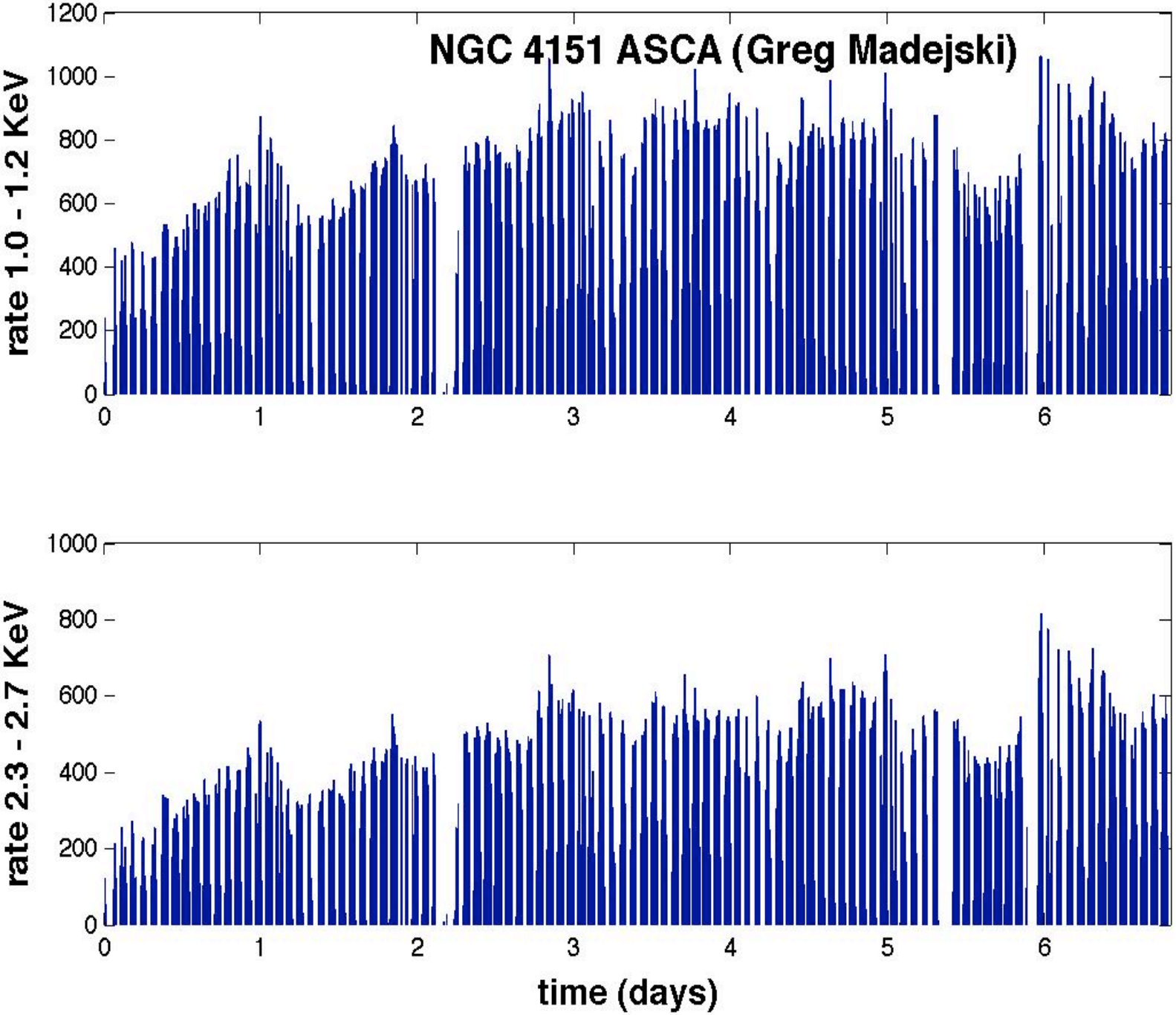


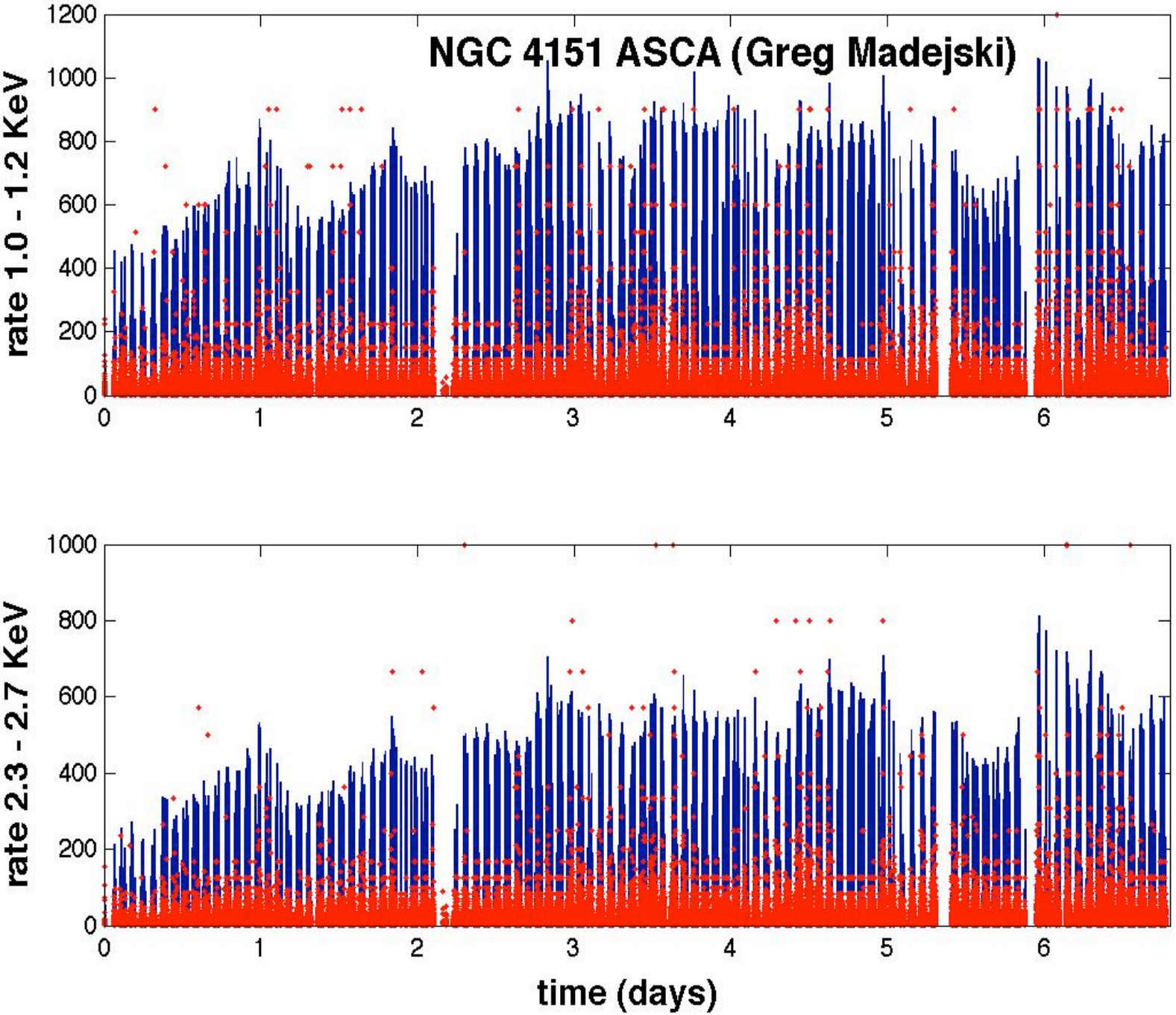


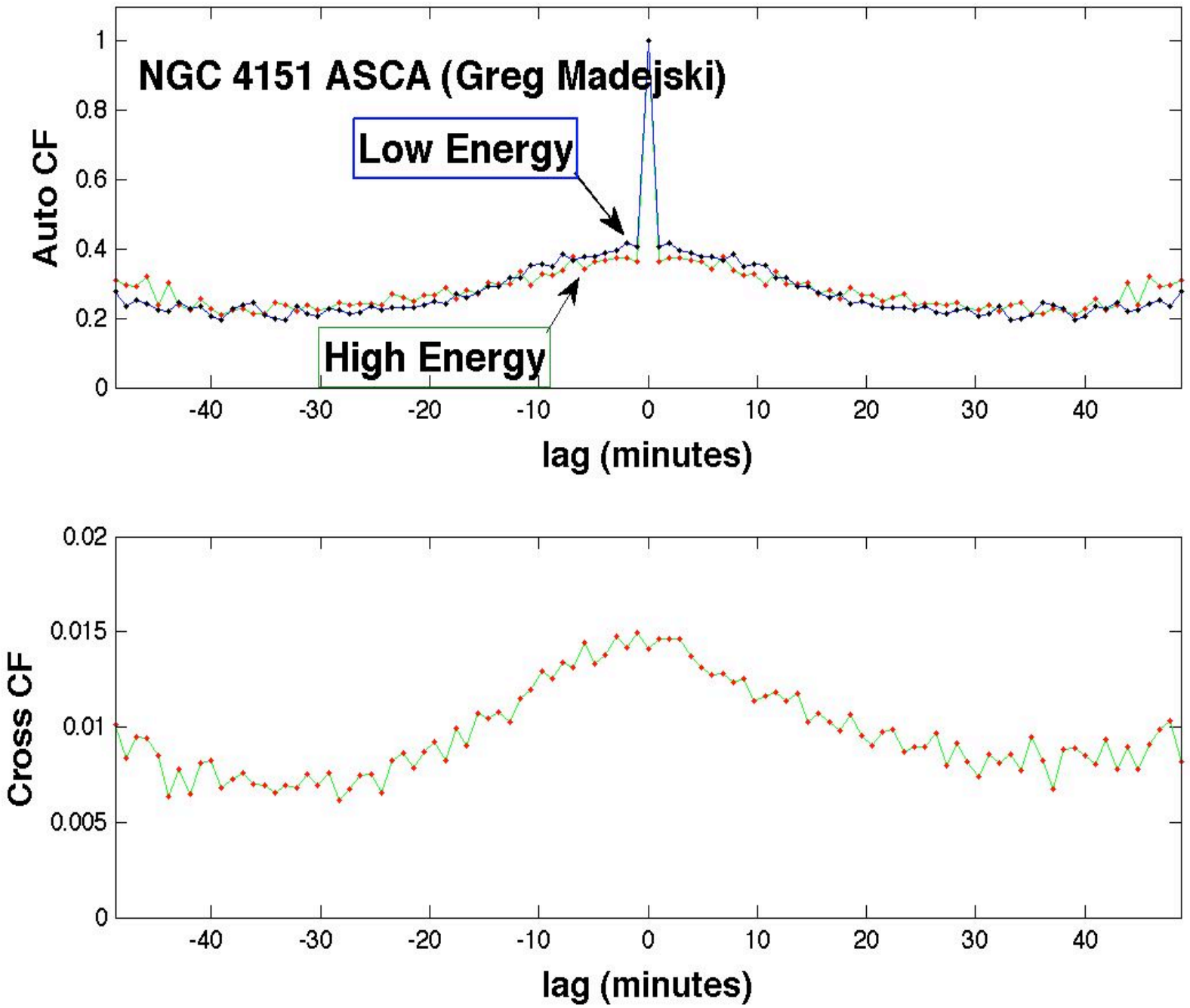


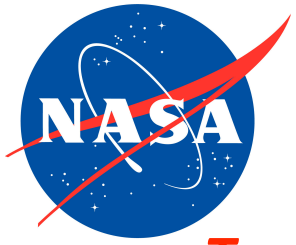
Edelson and Krolik: The Discrete Correlation Function: a New Method for Analyzing Unevenly Sampled Variability Data, Ap. J. 333, 1988, 646- starting point for all else!




NGC 4151 ASCA (Greg Madejski)
rate 1.0 - 1.2 KeV
rate 2.3 - 2.7 KeV
time (days)
0
200
400
600
800
1000
1200
0
1
2
3
4
5
6


NGC 4151 ASCA (Greg Madejski)
rate 1.0 - 1.2 KeV
rate 2.3 - 2.7 KeV
time (days)
0
200
400
600
800
1000
1200
1
2
3
4
5
6


NGC 4151 ASCA (Greg Madejski)
Low Energy
High Energy
Auto CF
lag (minutes)
Cross CF
lag (minutes)

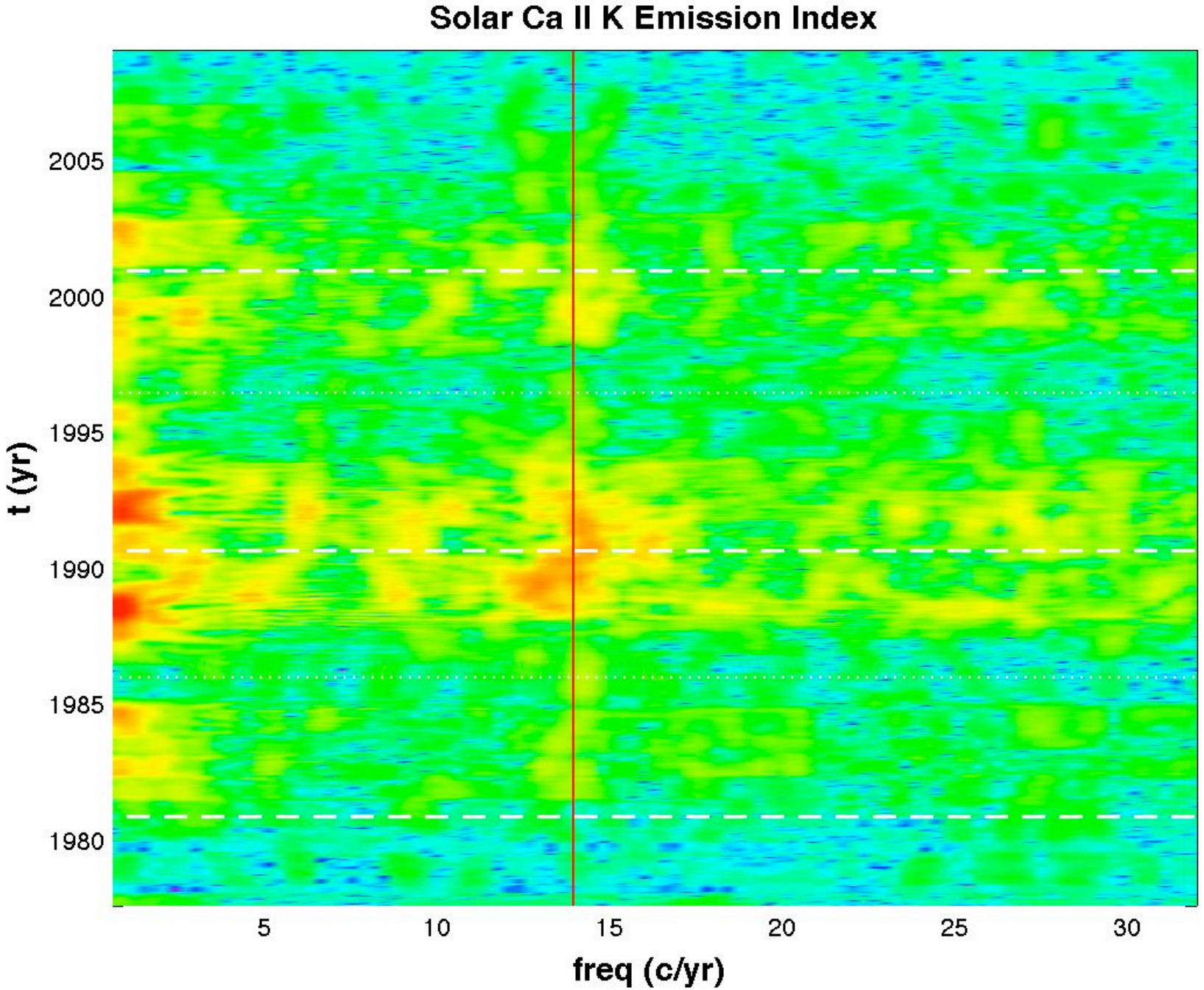


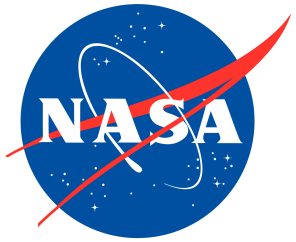
Time-Frequency/Time-Scale Analysis

Transform to a new view of the time series information.

- A Reality in joint time & frequency (or scale) representation
- Atomic decomposition
  - Time-frequency atoms
  - Over-complete representations
  - Optimal Basis Pursuit (Mallat), etc.
- Uncertainty Principle: T-F resolution tradeoff
- Non-stationary processes
  - Flares
  - Trends & Modulations
  - Statistical change-points
- Instantaneous Frequency
- Local vs. Global structure
- Interference (cross-terms in bi-linear representation)

Time-Frequency/Time-Scale Analysis (Temps-Fréquence) Patrick Flandrin http://perso.ens-lyon.fr/patrick.flandrin/publis.html; A Wavelet tour of Signal Processing (Une Exploration des Signaux en Ondelettes) Stéphane Mallat

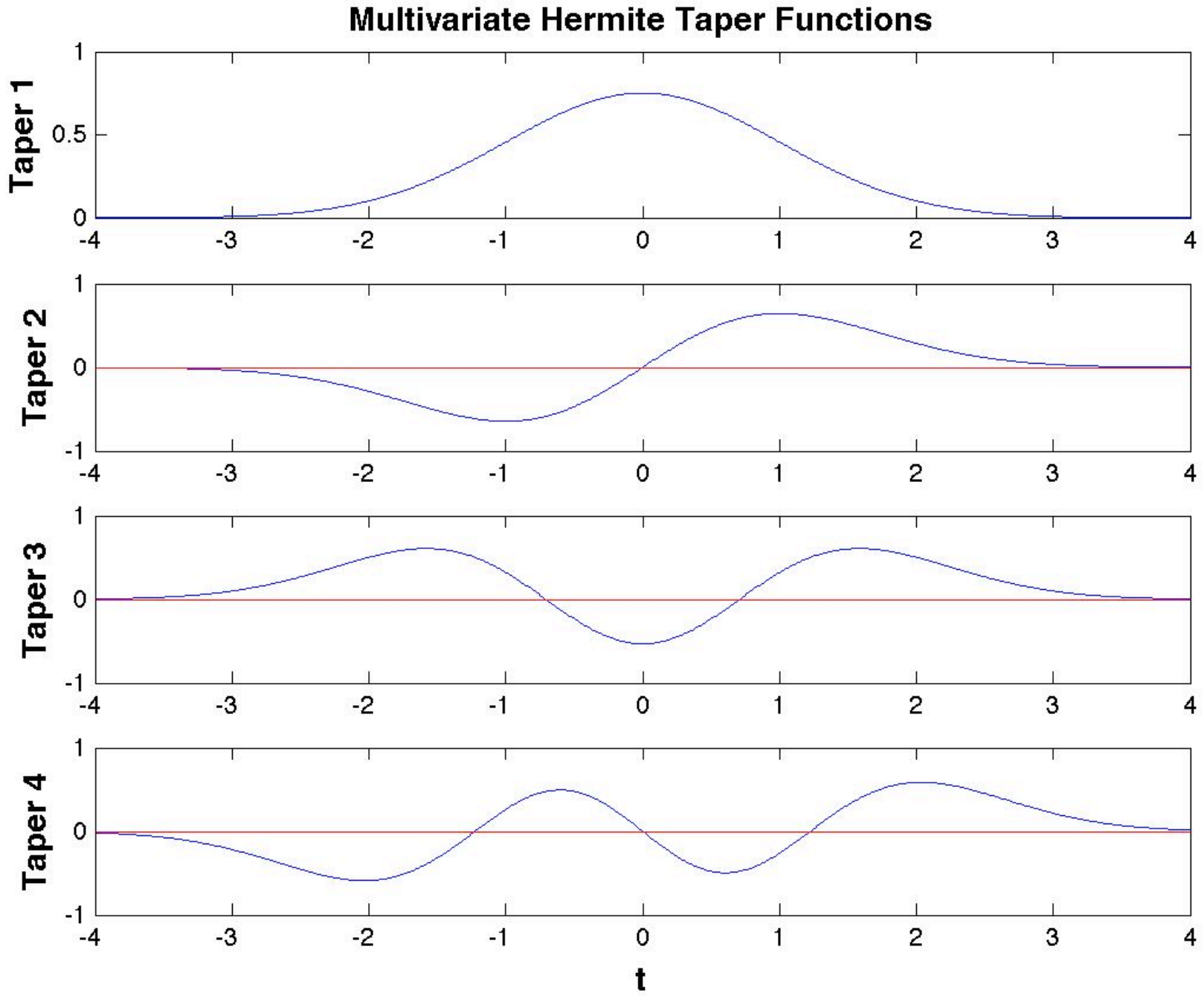

Solar Ca II K Emission Index
2005
2000
1995
1990
1985
1980
t (yr)
5
10
15
20
25
30
freq (c/yr)

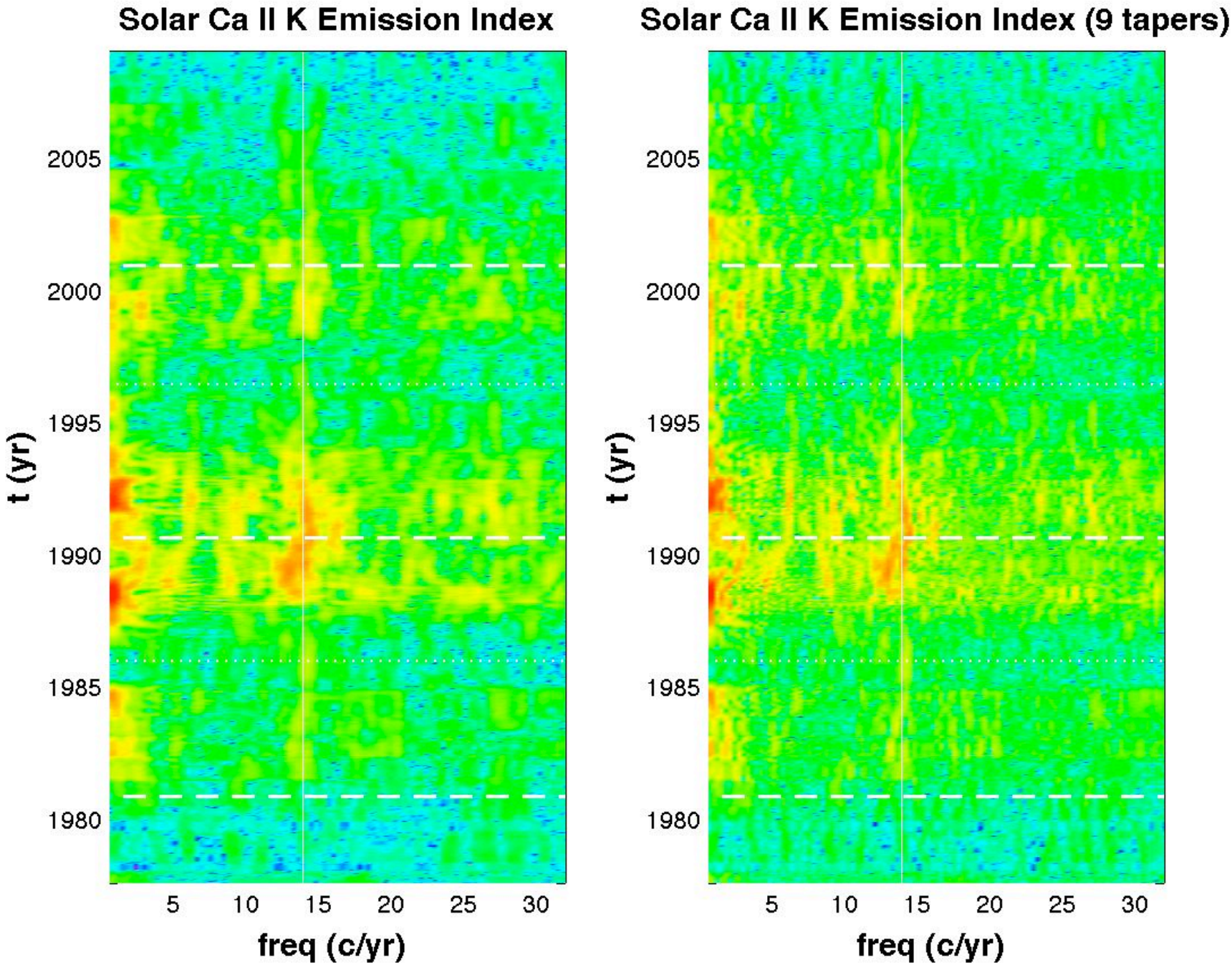


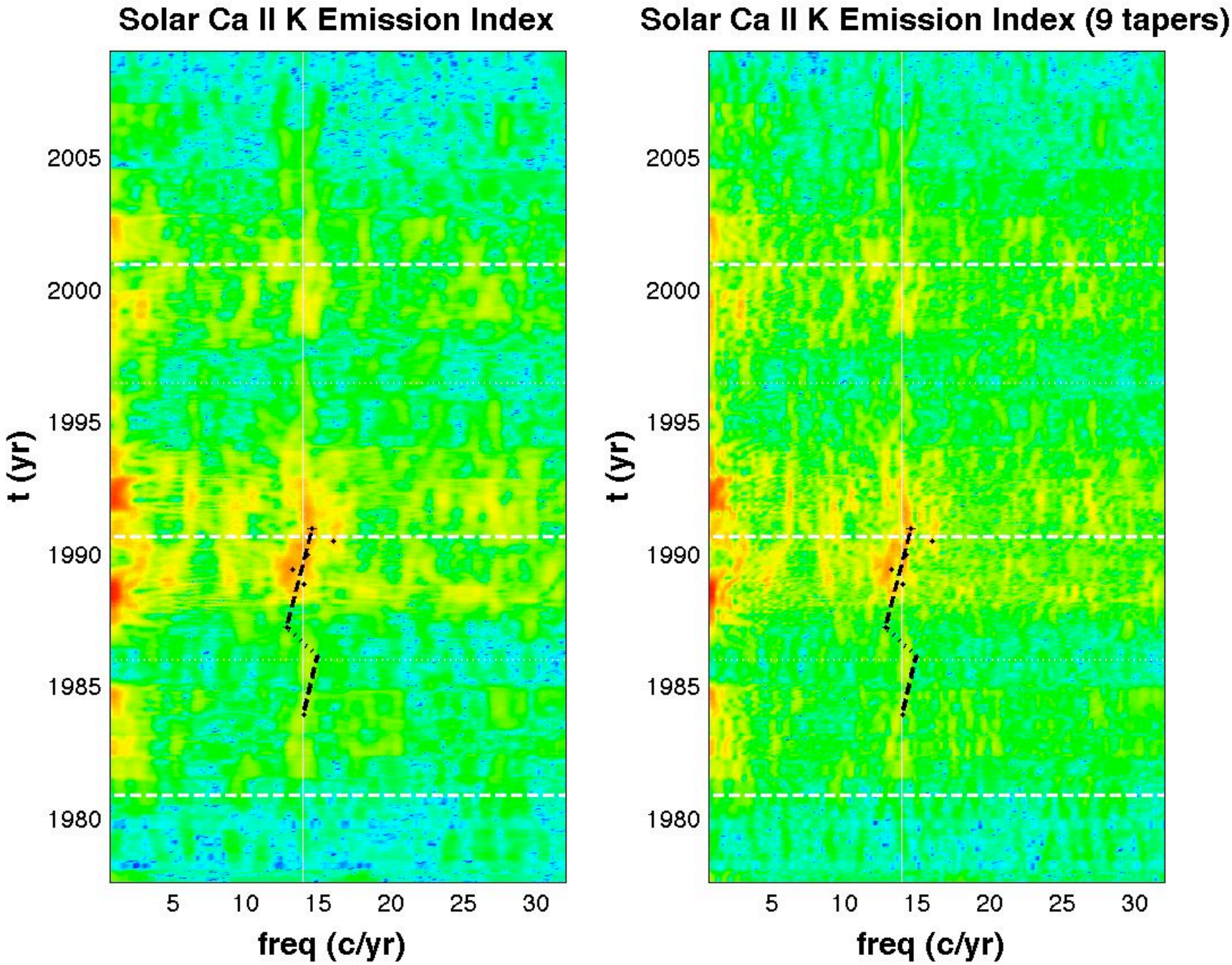
Multi-taper Analysis (Thomson 1982)

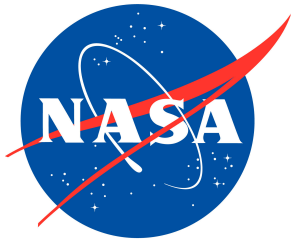
- Tapers (windows) reduce sidelobe leakage = bias
- Incomplete use of data → loss of information
- Multitapers recover this information
- Leakage minimization = eigenvalue problem
  - Eigenfunctions: efficient window functions
  - Eigenvalues
    - measure effectiveness
    - determine how many terms to include

Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques, Don Percival and Andrew Walden (1993)


Multivariate Hermite Taper Functions
Taper 1
Taper 2
Taper 3
Taper 4
t
1
0.5
0
-1
-4
-3
-2
-1
0
1
2
3
4


Solar Ca II K Emission Index
Solar Ca II K Emission Index (9 tapers)
t (yr)
2005
2000
1995
1990
1985
1980
5
10
15
20
25
30
freq (c/yr)


Solar Ca II K Emission Index
Solar Ca II K Emission Index (9 tapers)
t (yr)
2005
2000
1995
1990
1985
1980
5
10
15
20
25
30
freq (c/yr)



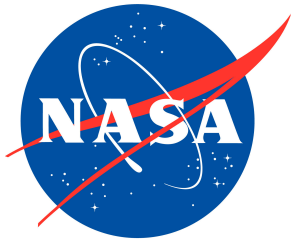
Machine Learning and Data Mining in Astronomy

Editors: Kamal Ali, Jeff Scargle, Ashok Srivastava, Mike Way
Chapman and Hall, Data Mining and Knowledge Discovery series
http://astrophysics.arc.nasa.gov/~mway/book/DMKD.pdf

Handbook of Statistical Analysis of Event Data

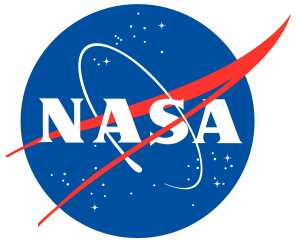
Jeff Scargle

Contributions welcome!



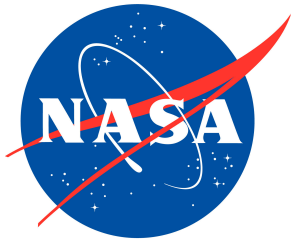
Statistical Interlude

- Clinical studies usually small and expensive
- "Meta-analysis" – Increase significance by combining statistical summaries of published studies (not re-analysis of original data)
- Role of publication bias (PB)
- Assess potential for PB with Rosenthal formula



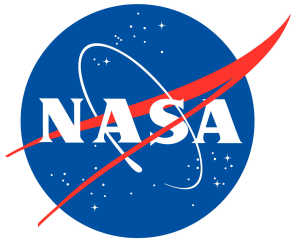
Statistical Interlude

- Publication bias is large!
- Editorial policy: Do not publish a study unless it achieves a 3-sigma positive result
- Rosenthal formula:
  - ✓ Completely wrong!
  - ✓ Used to justify hundreds of "meta-analytic" results in medicine, and psychology (real and para-)
  - ✓ Not a single applied scientist questioned the validity of the formula
- Many medical studies, especially those relevant to decisions about safety of drugs to be released to the market, are based on this statistical blunder.



Statistical Interlude

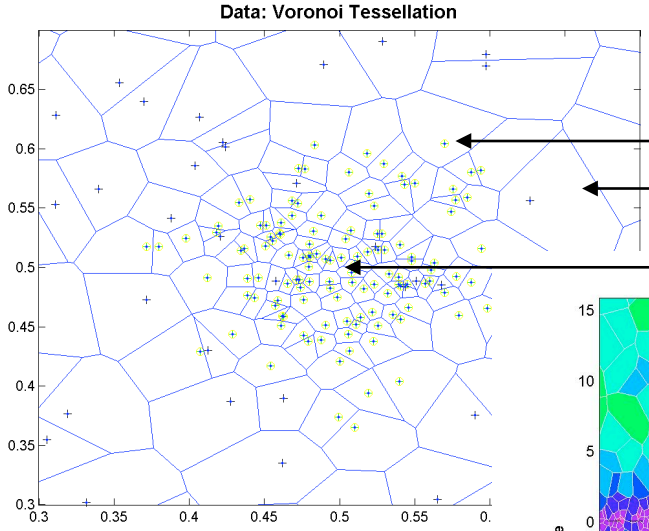
- Rosenthal, R. (1979) The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.
- Publication Bias: The “File-Drawer” Problem in Scientific Inference, J. D. Scargle. Journal of Scientific Exploration, Vol. 14, No. 1, pp. 91–106, 2000.
- A Generalized Publication Bias Model, P. H. Schonemann and J. D. Scargle, Chinese Journal of Psychology, 2008, Vol. 50, 1, 21-29.



Statistical Interlude

- Pre-election radio interview with the president of a major political polling organization ("Dr. Z").
- Caller: "I hang up on polling phone calls – intrusion of my privacy."
- Discussion of this as a potential bias.
- Dr. Z: "I don't worry about such biases. We just get a larger sample."
- JS calls the radio show and tries to verify Dr. Z's belief that increased sample size can fix a bias.
- Dr. Z does not understand; responds by puffing up the reliability of his polling organization.

Voronoi Tessellations on 3+ Scales



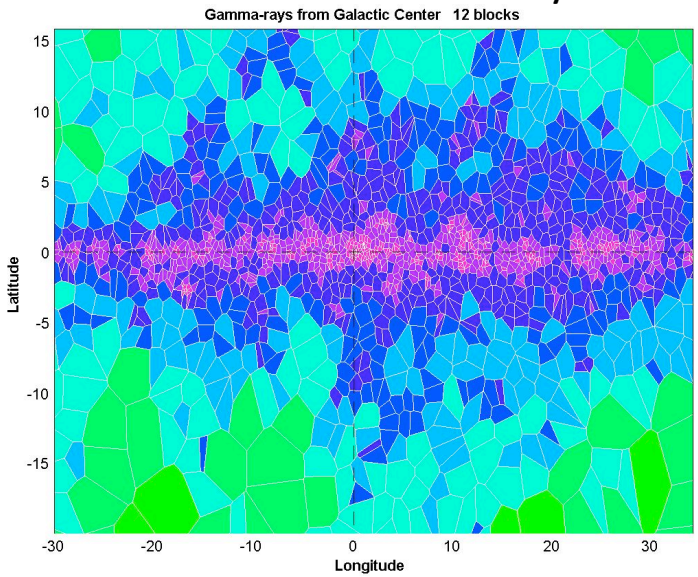
10^{-35} meters

Random space-time lattice (T. D. Lee)

Points: micro-partons?

Cells: Planck length cells

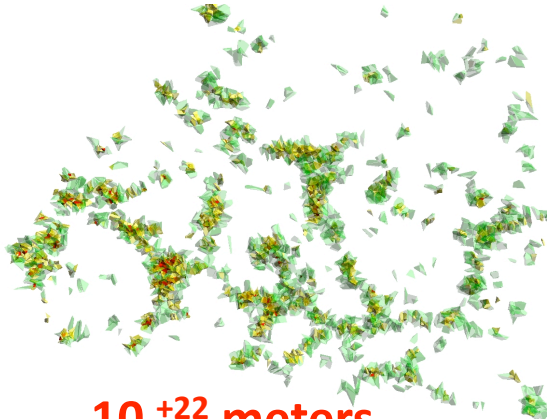
Blocks: Elementary Particles



GLAST Source Detection Algorithm

Points: Photons

Blocks: Point sources



10^{+22} meters

Cluster detection algorithm:

Points: Galaxies

Cells: Galaxy Neighborhoods

Blocks: Clusters, filaments, …

Large Scale Structure

Points: Galaxies

Cells: Voids