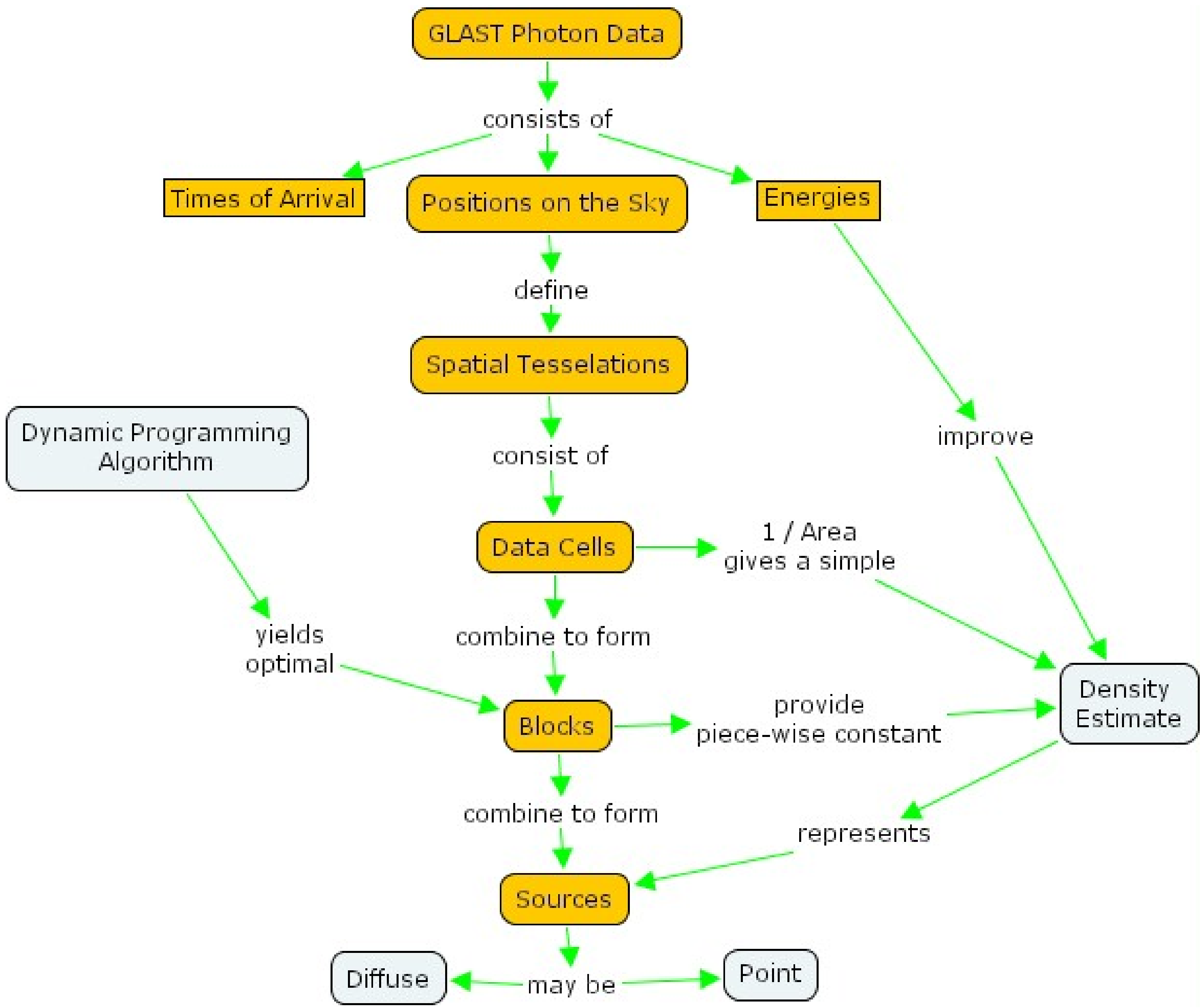
Bootstrap Segmentation Analysis and Expectation Maximization to Detect and Characterize Sources

Jeffrey.D.Scargle@nasa.gov

Space Science Division
NASA Ames Research Center

Diffuse Emission & LAT Source Catalog

SLAC Workshop: May 23, 2005


GLAST Photon Data
consists of
Times of Arrival
Positions on the Sky
Energies
define
Spatial Tesselations
improve
Dynamic Programming Algorithm
consist of
Data Cells
1 / Area gives a simple
yields optimal
combine to form
Density Estimate
Blocks
provide piece-wise constant
combine to form
represents
Sources
Diffuse
may be
Point

Problems (Solutions)

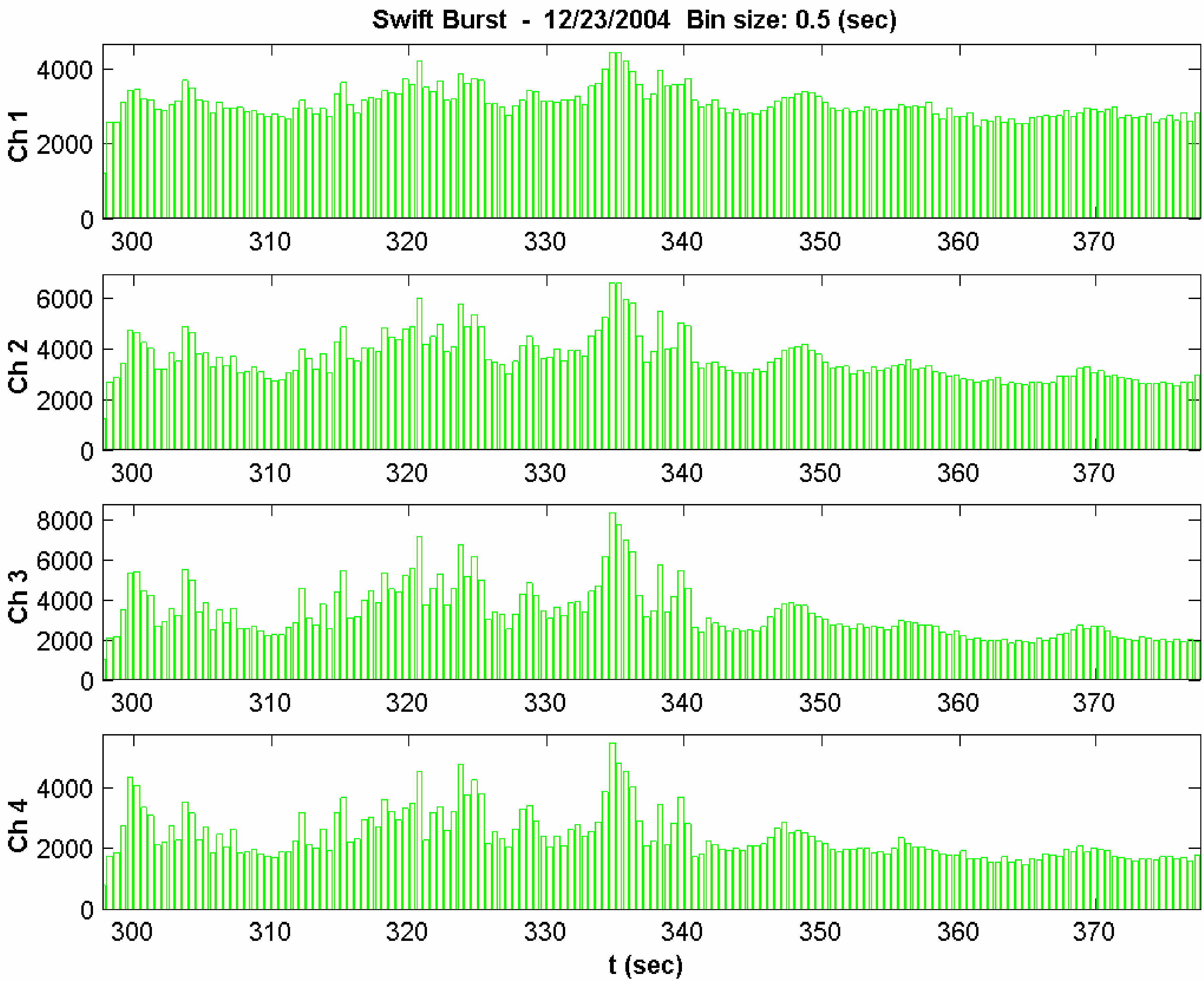
- Algorithm provides no Error Estimate
  - Bootstrap Errors
  - Block Posterior Probabilities

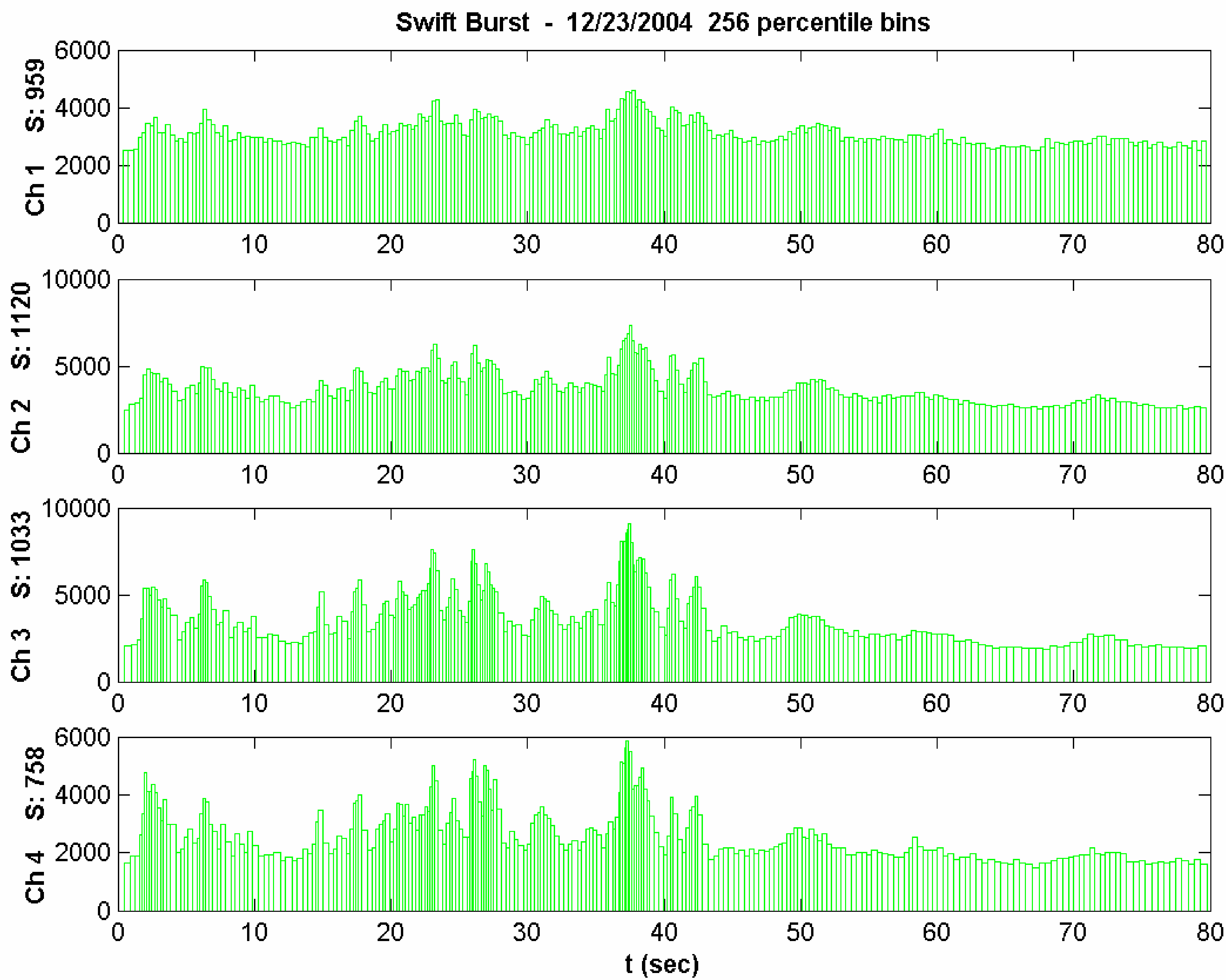
- Point Spread: Overlapping Sources
  - Expectation Maximization (EM)

- Point Spread: Function of Energy
  - Maximum Likelihood Models?

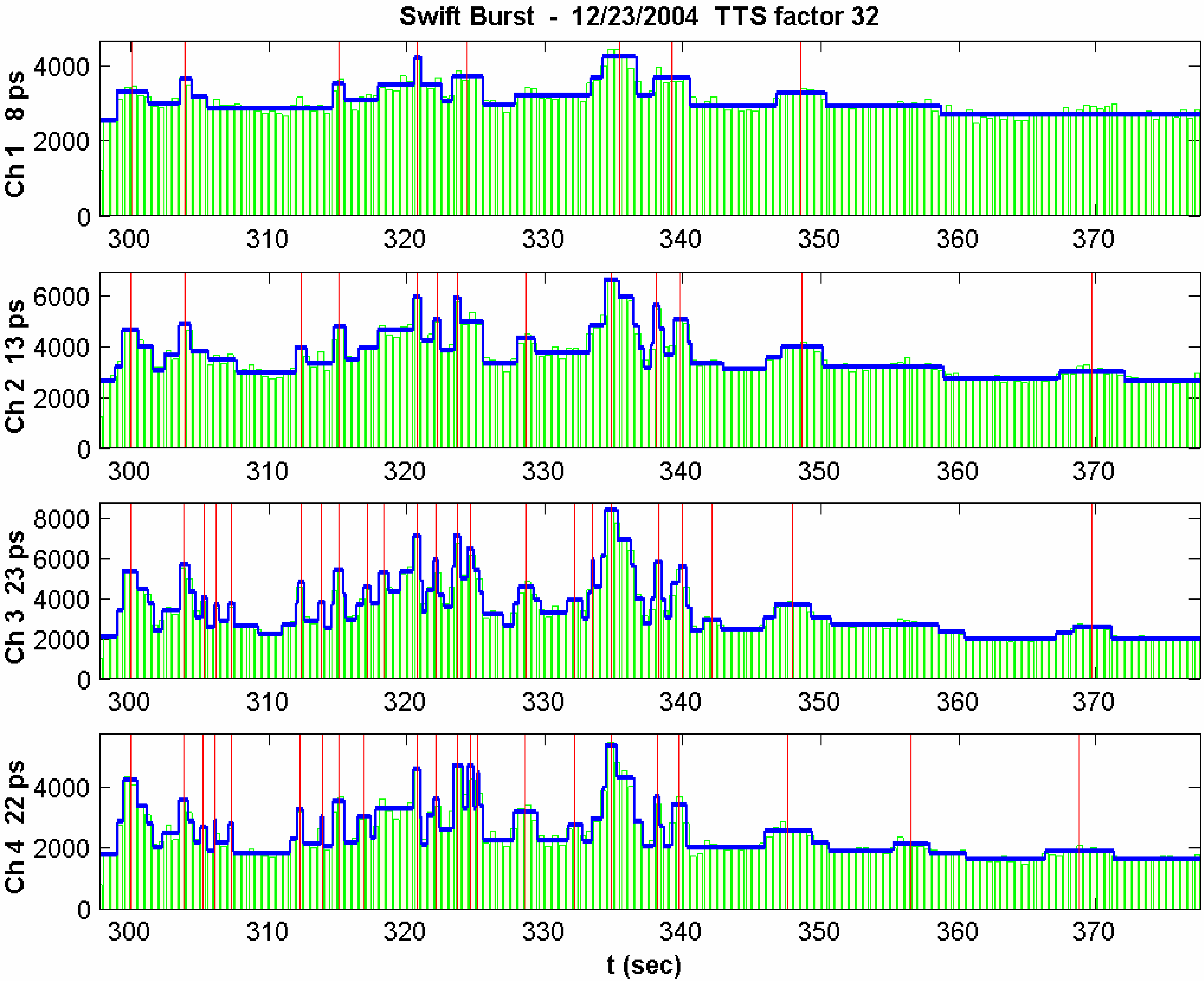
One Dimensional[1] Example: Swift GRB Data

1. But everything applies to 2D and higher!


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



Swift Burst - 12/23/2004 256 percentile bins
Ch 1 S: 959
Ch 2 S: 1120
Ch 3 S: 1033
Ch 4 S: 758
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)

Expectation Maximization (EM)

Initialize: Find a good guess at the mixture model:

- How Many Sources?
- Locations of the Sources
- Source Parameters (size, spectra, …)

Iterate:

1. From the model: Divide the data into pieces that are relevant to each Source separately

2. Re-determine the Source Locations and Parameters by fitting to the data pieces in 1

3. Repeat as needed

The “Maximization” Step with Point Data

Maximize the Unbinned log-Likelihood:

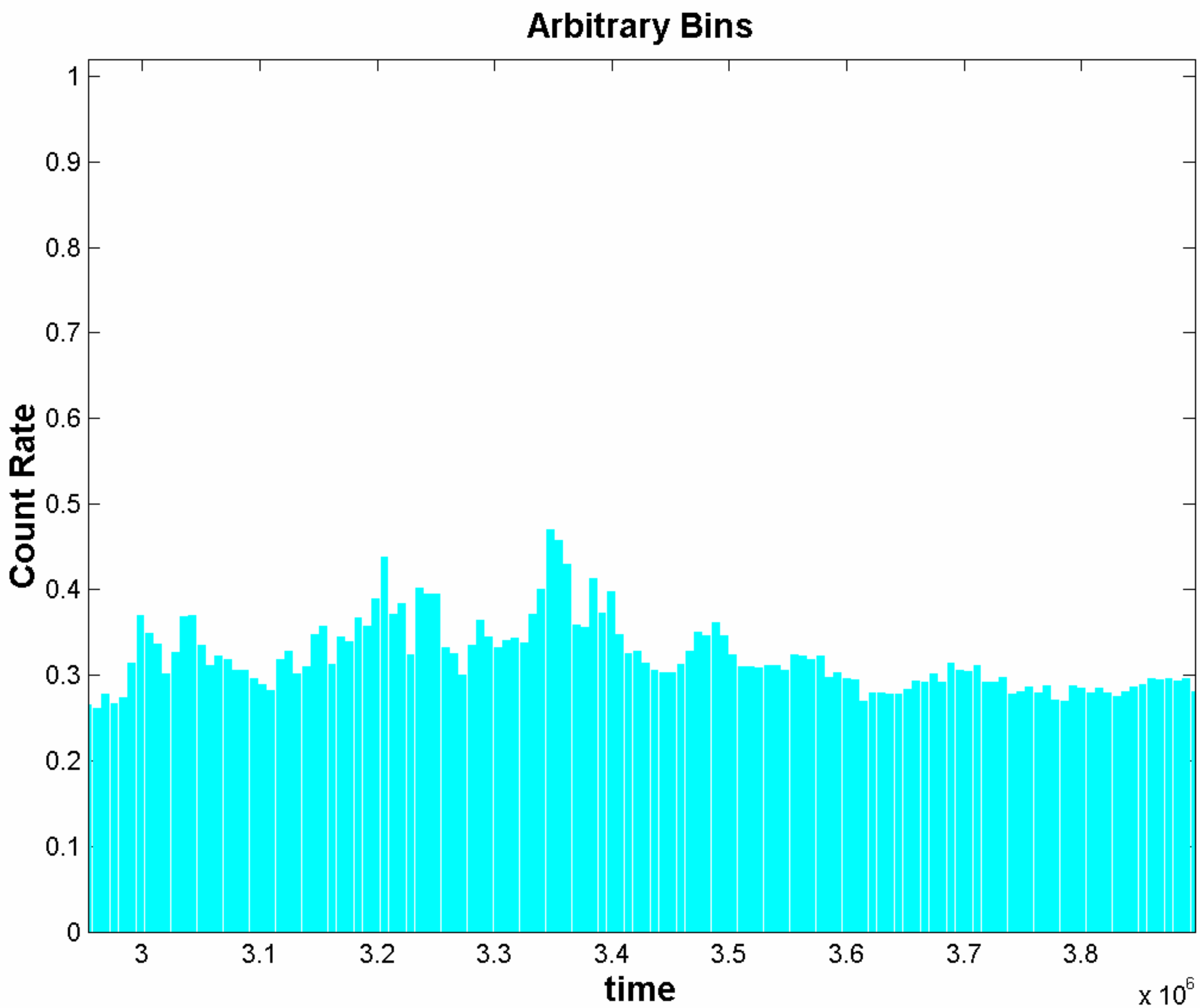
$$\mathrm{Log}(L) = \blacklozenge_{i} \log[\, \nearrow_a(t_i) + b \,] \; - \Downarrow w(t)\, [\, \nearrow_a(t) + b \,]$$

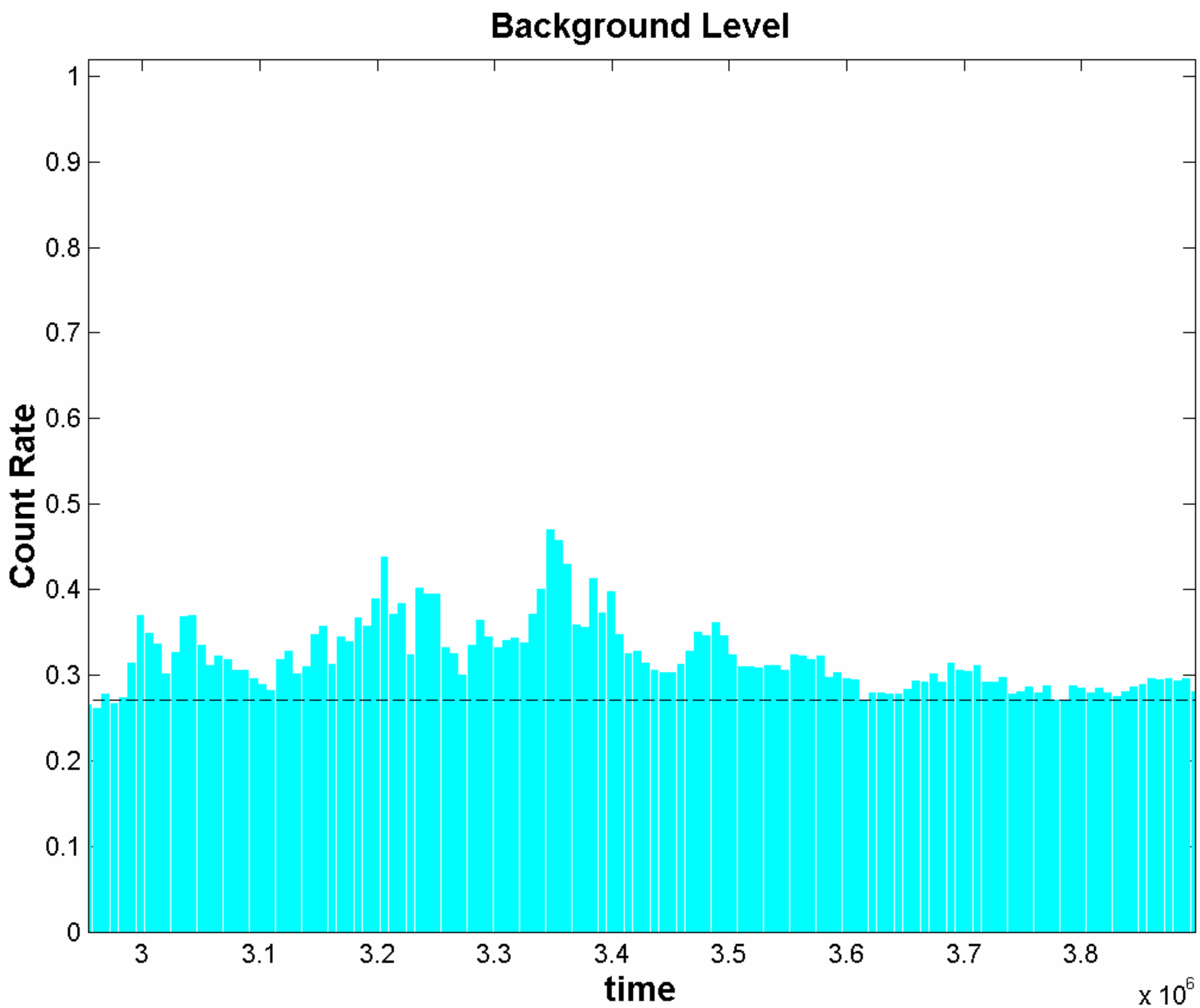
Where the model for each source (pulse) is:

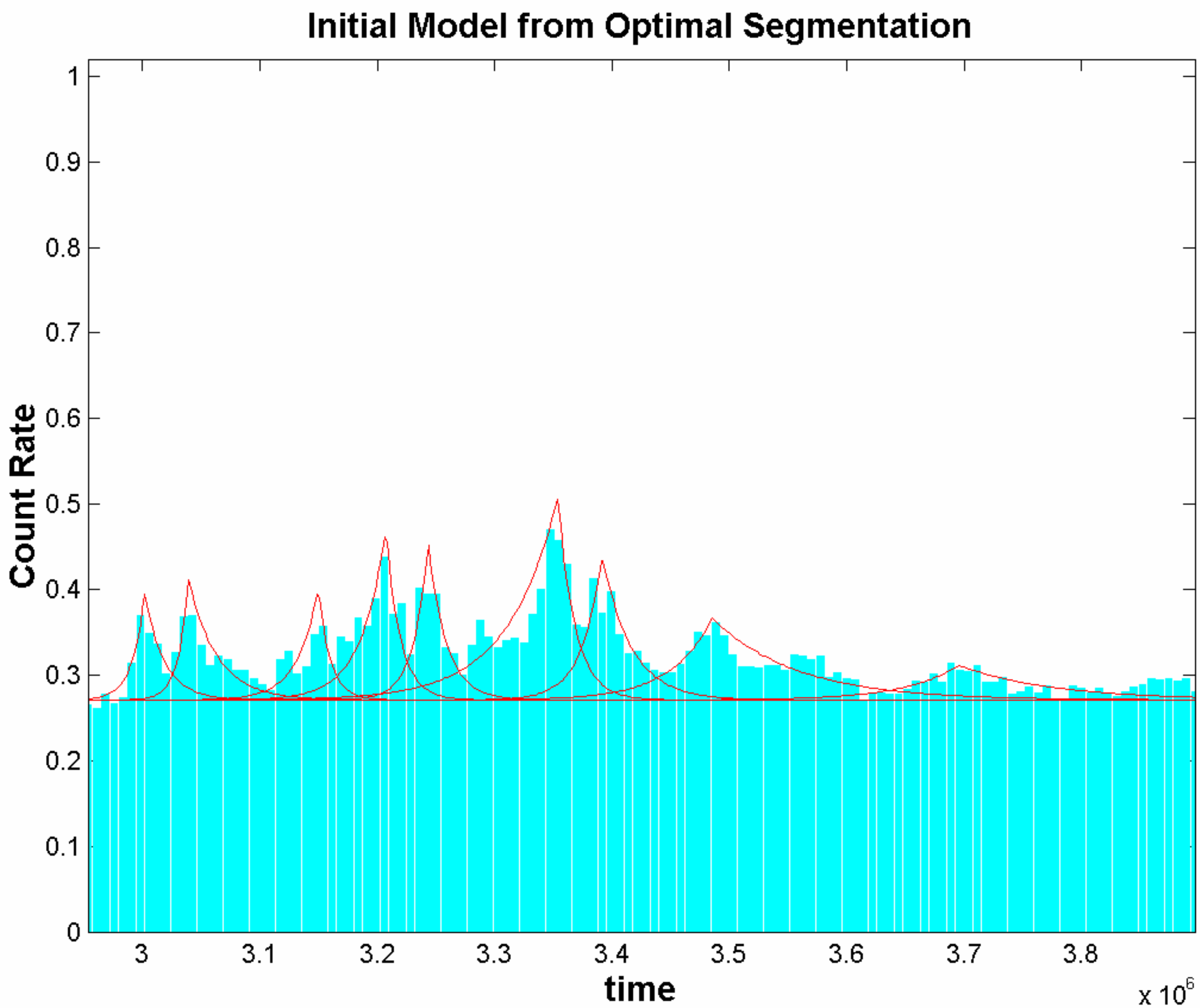
$$X(t) = [\, \nearrow_a(t) + b \,]$$

a = source parameters (location, size, …)
b = local background constant
w(t) is the EM partitioning, or weighting, function

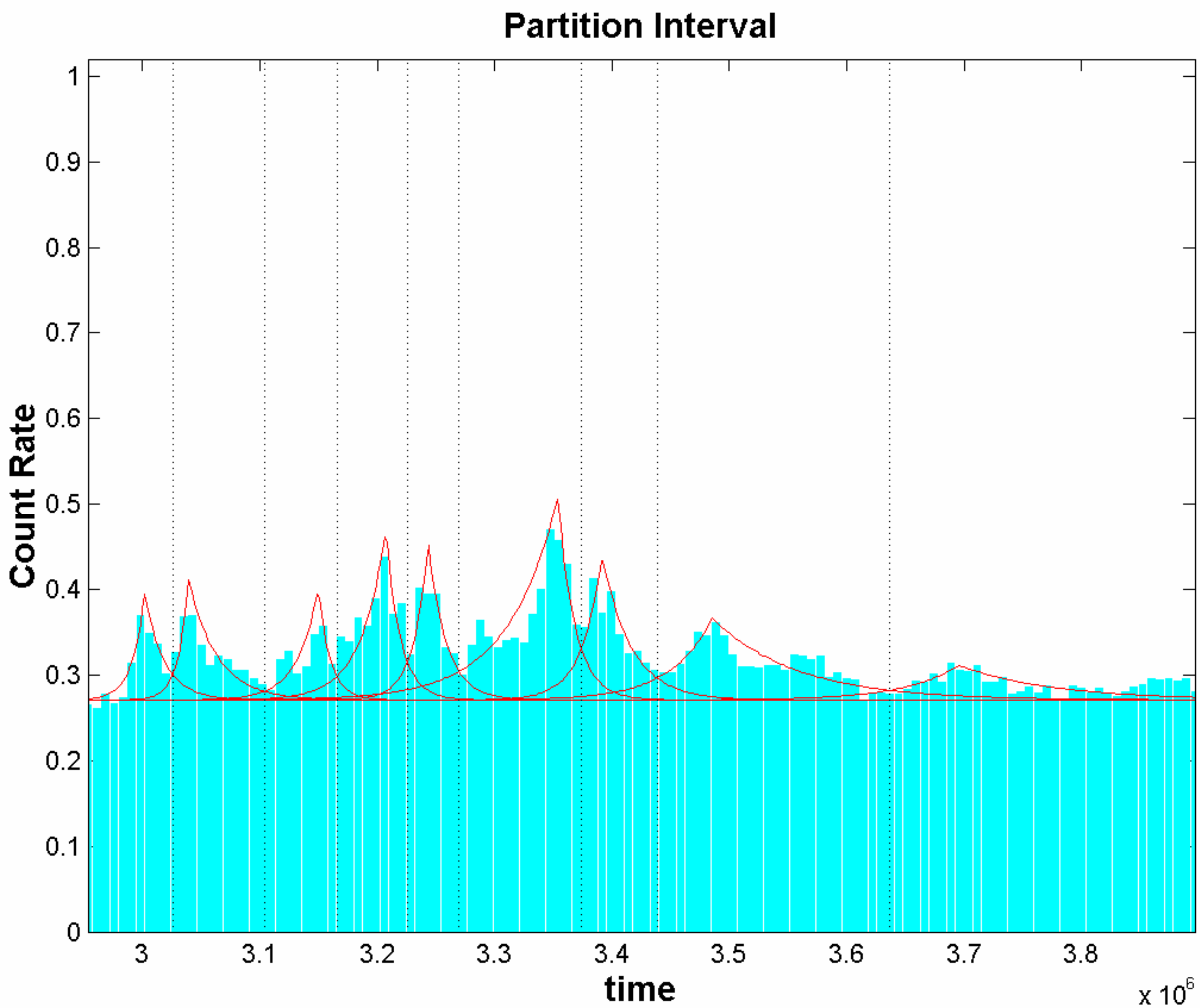
NB: the term $\blacklozenge_{i} \log[\, w(t_i) \,]$ in the full log-likelihood is a constant,
irrelevant for model fitting.

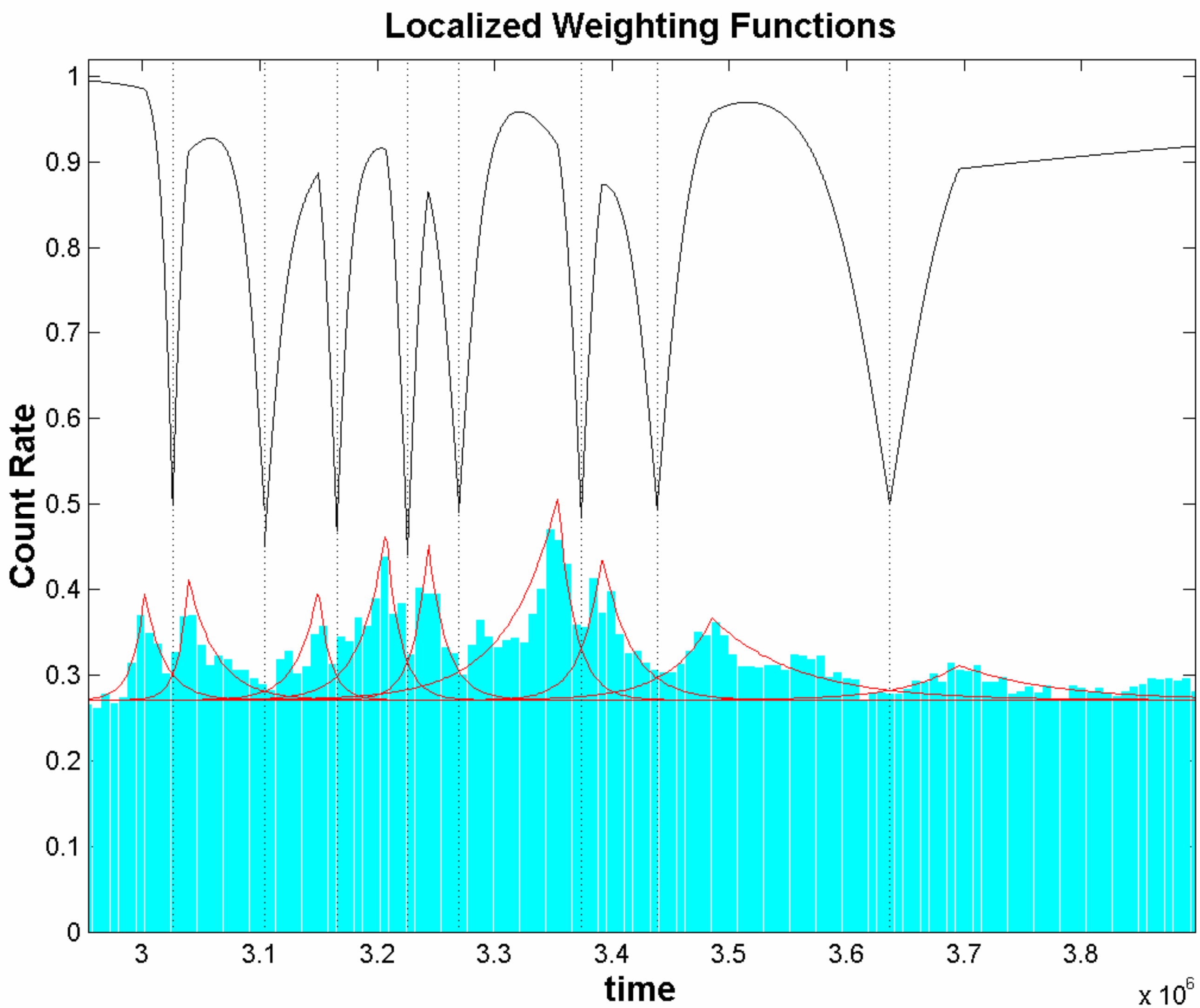

Arbitrary Bins
Count Rate
time
x 10^6
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
3
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8


Background Level
Count Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
3
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
time
x 10^6

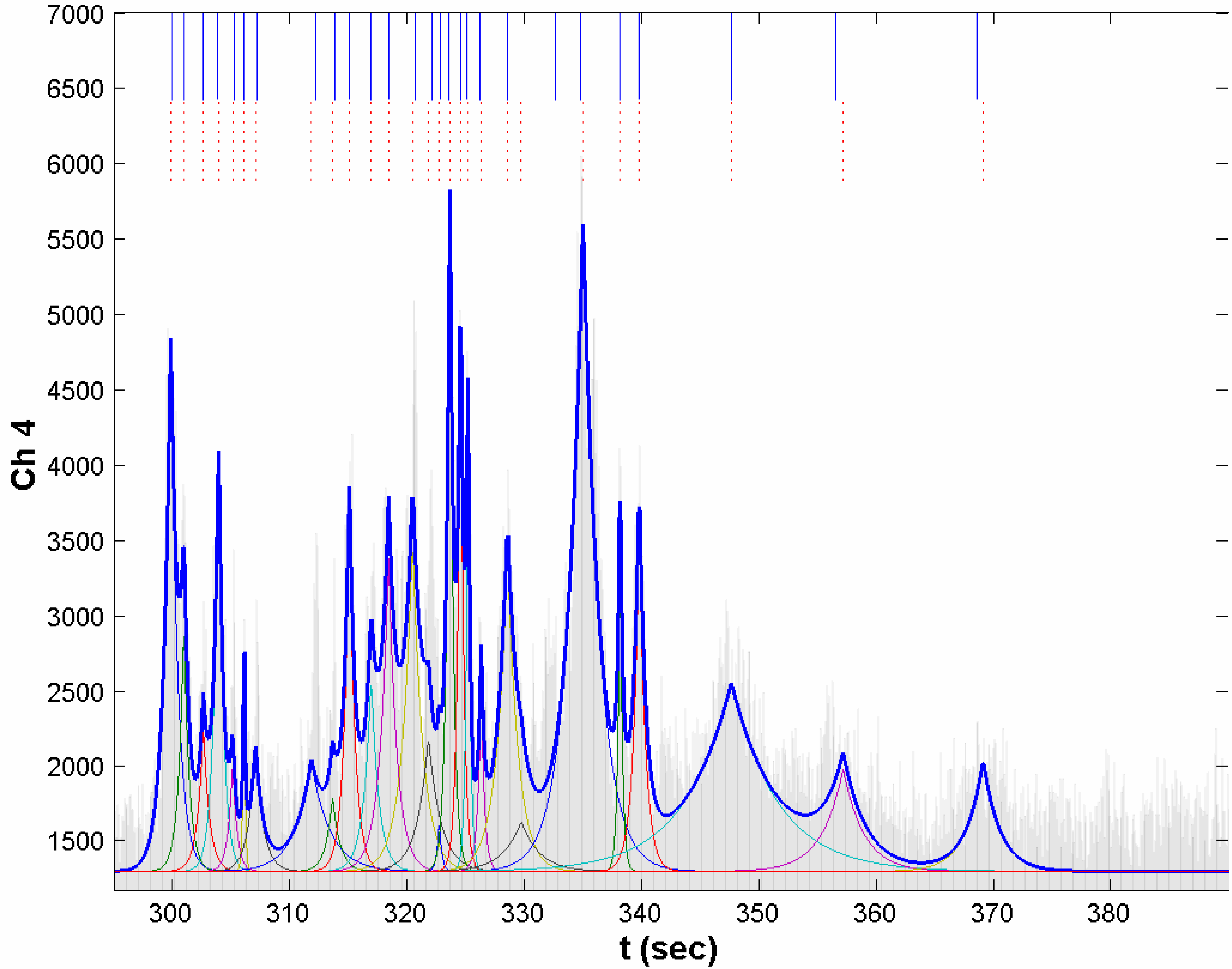


Initial Model from Optimal Segmentation
Count Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
3
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
time
x 10^6

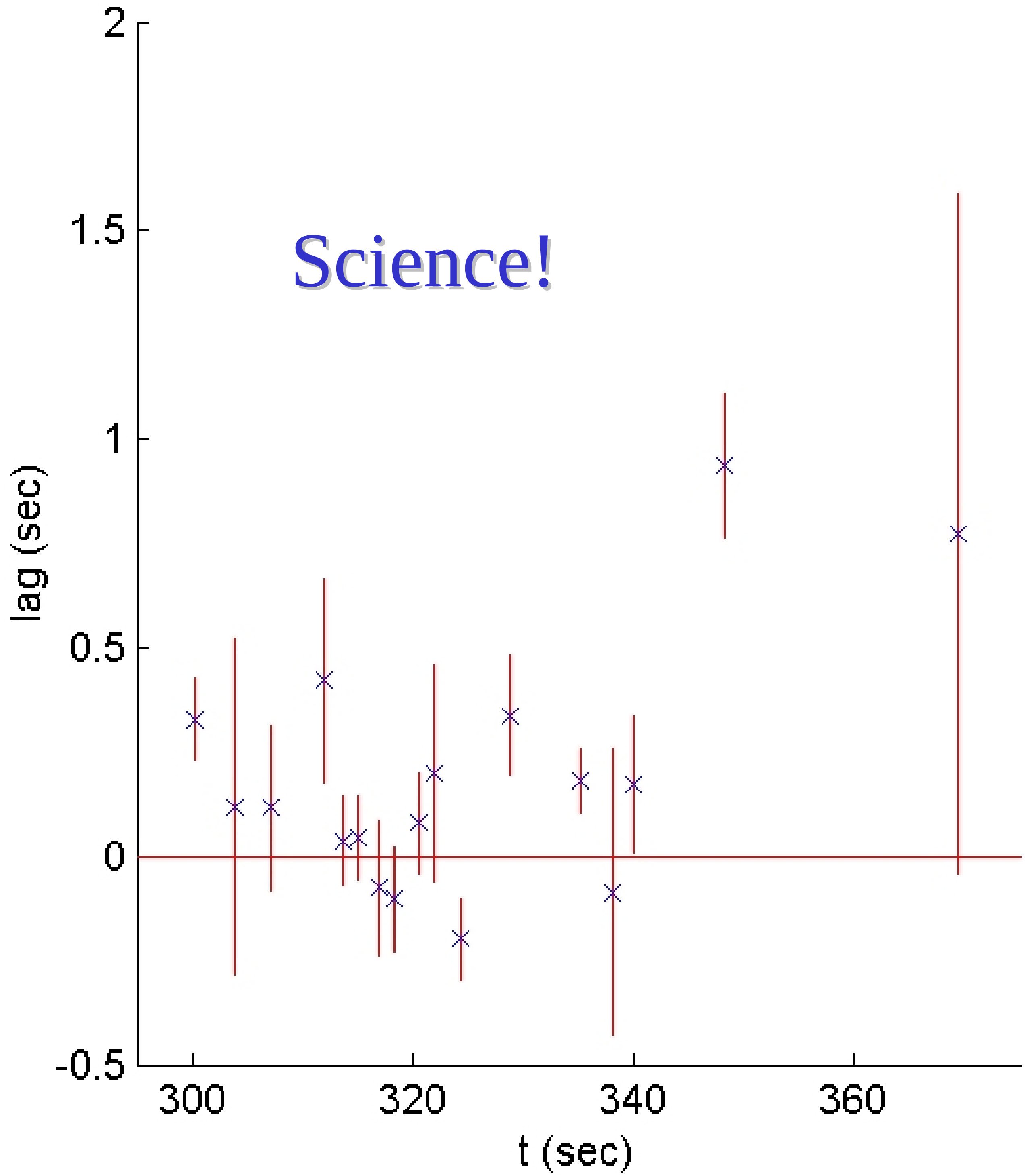

Partition Interval
Count Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
3
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
time
x 10^6



Localized Weighting Functions
Count Rate
time
x 10^6
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
3
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8



7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
Ch 4
300
310
320
330
340
350
360
370
380
t (sec)


Science!
lag (sec)
t (sec)
2
1.5
1
0.5
0
-0.5
300
320
340
360

Bootstrap Method:
Time Series of N Discrete Events

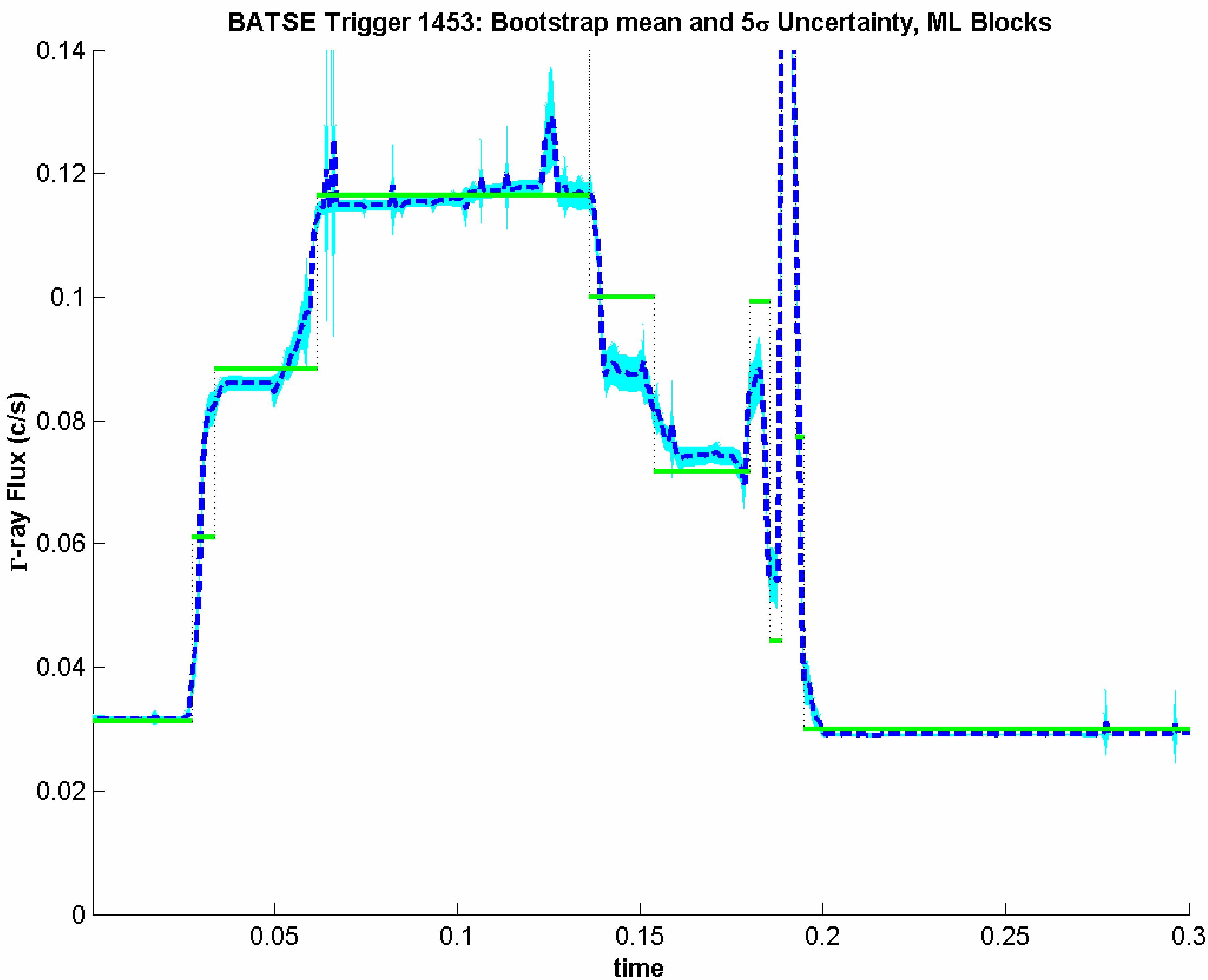
For many iterations:

- Randomly select N of the observed events *with replacement*
- Analyze this sample just as if it were real data

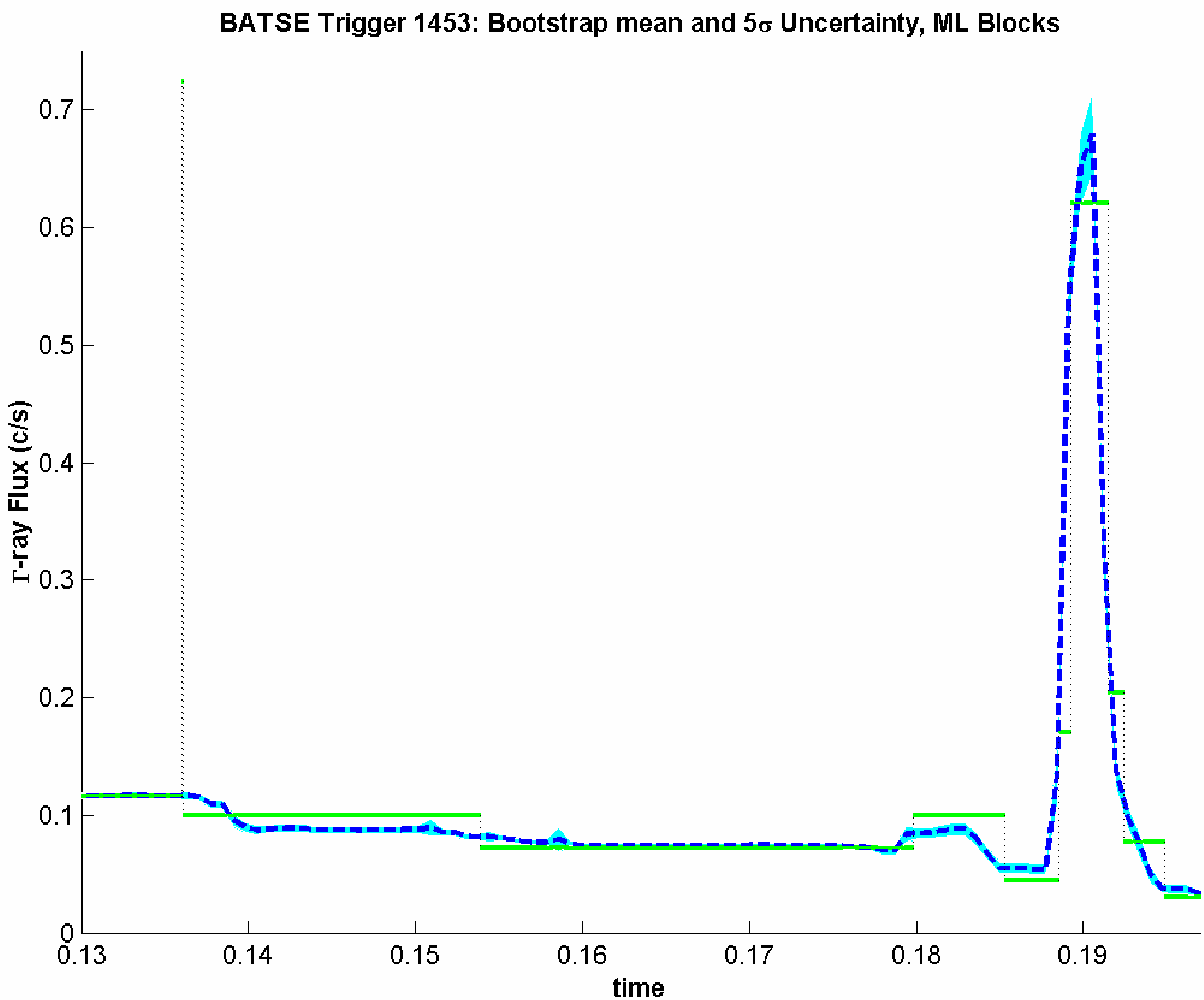
Compute mean and variance of the bootstrap samples

Bias = result for real data – bootstrap mean
RMS error derived from bootstrap variance

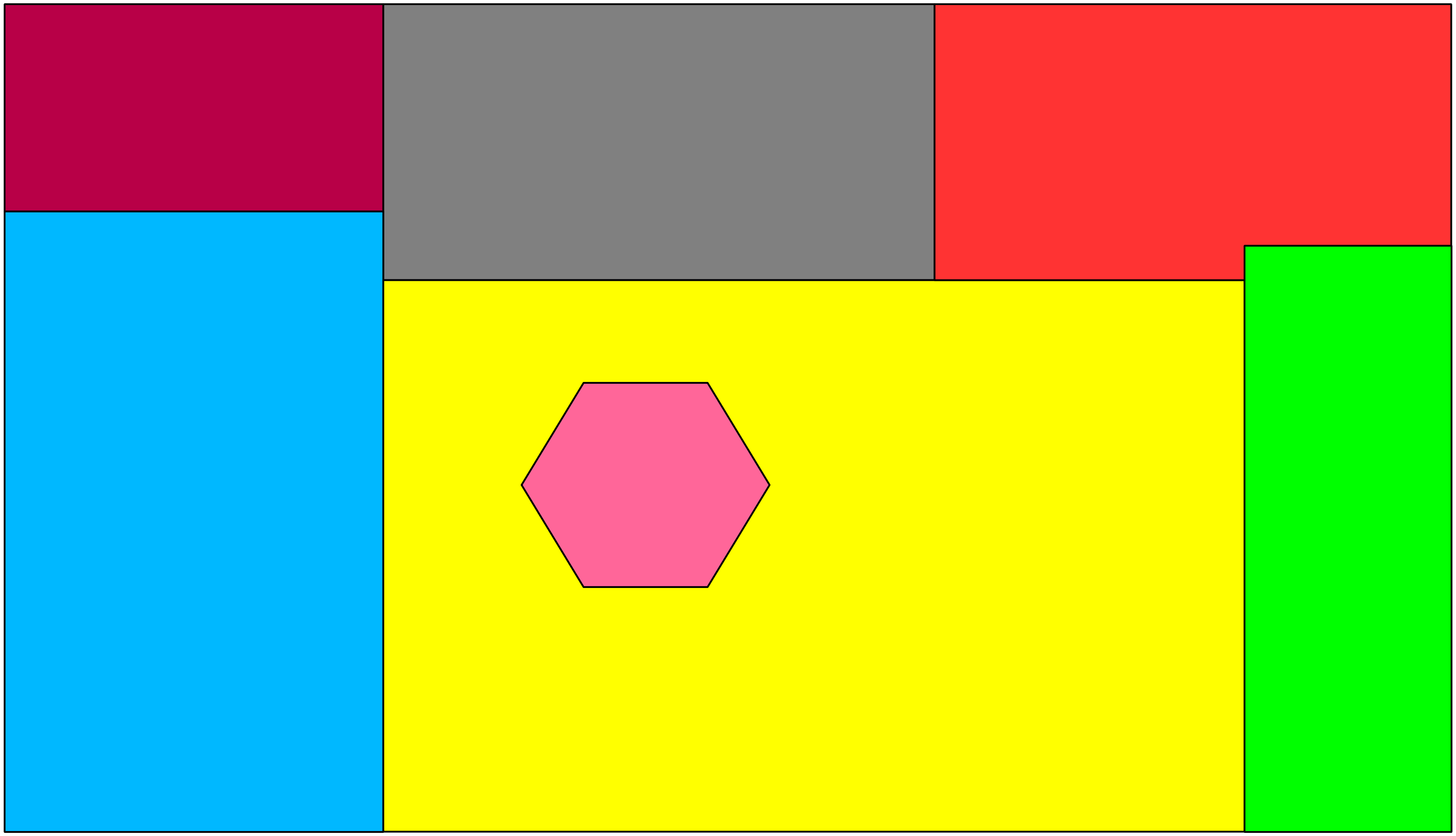
Caveat: The real data does not have the repeated events in bootstrap samples. I am not sure what effect this has.



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

Piecewise Constant Model
(partitions the data space)



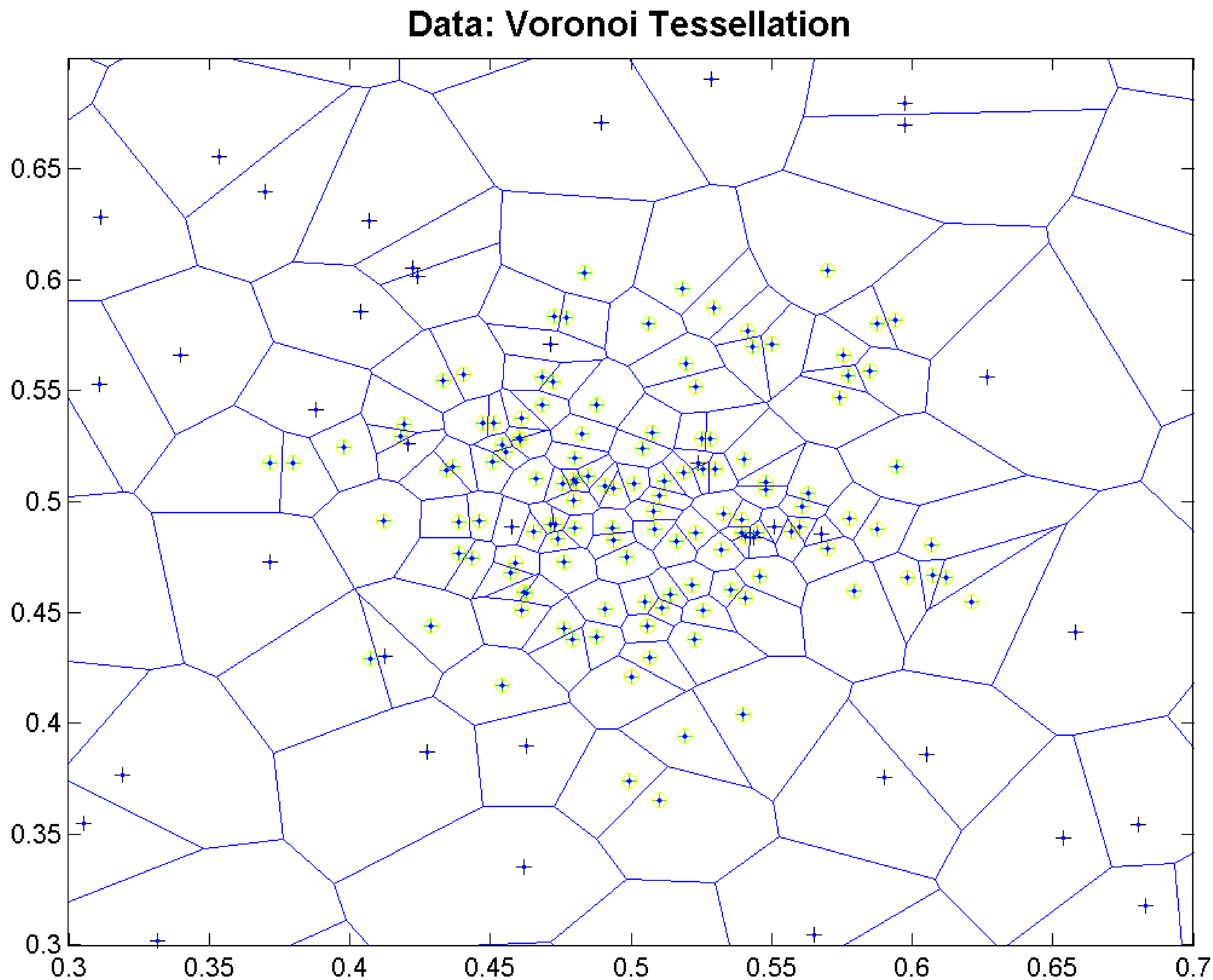
Signal modeled as constant over each partition element (block).

Optimum Partitions in Higher Dimensions

- Blocks are collections of Voronoi cells (1D,2D,...)
- Relax condition that blocks be connected
- Cell location now irrelevant
- Order cells by volume

Theorem: Optimum partition consists of blocks that are connected in this ordering

- Now can use the 1D algorithm, $O(N^2)$
- Postprocessing: identify connected block fragments



Data: Voronoi Tessellation
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7

Blocks

<u>Block: a set of data cells</u>

Two cases:

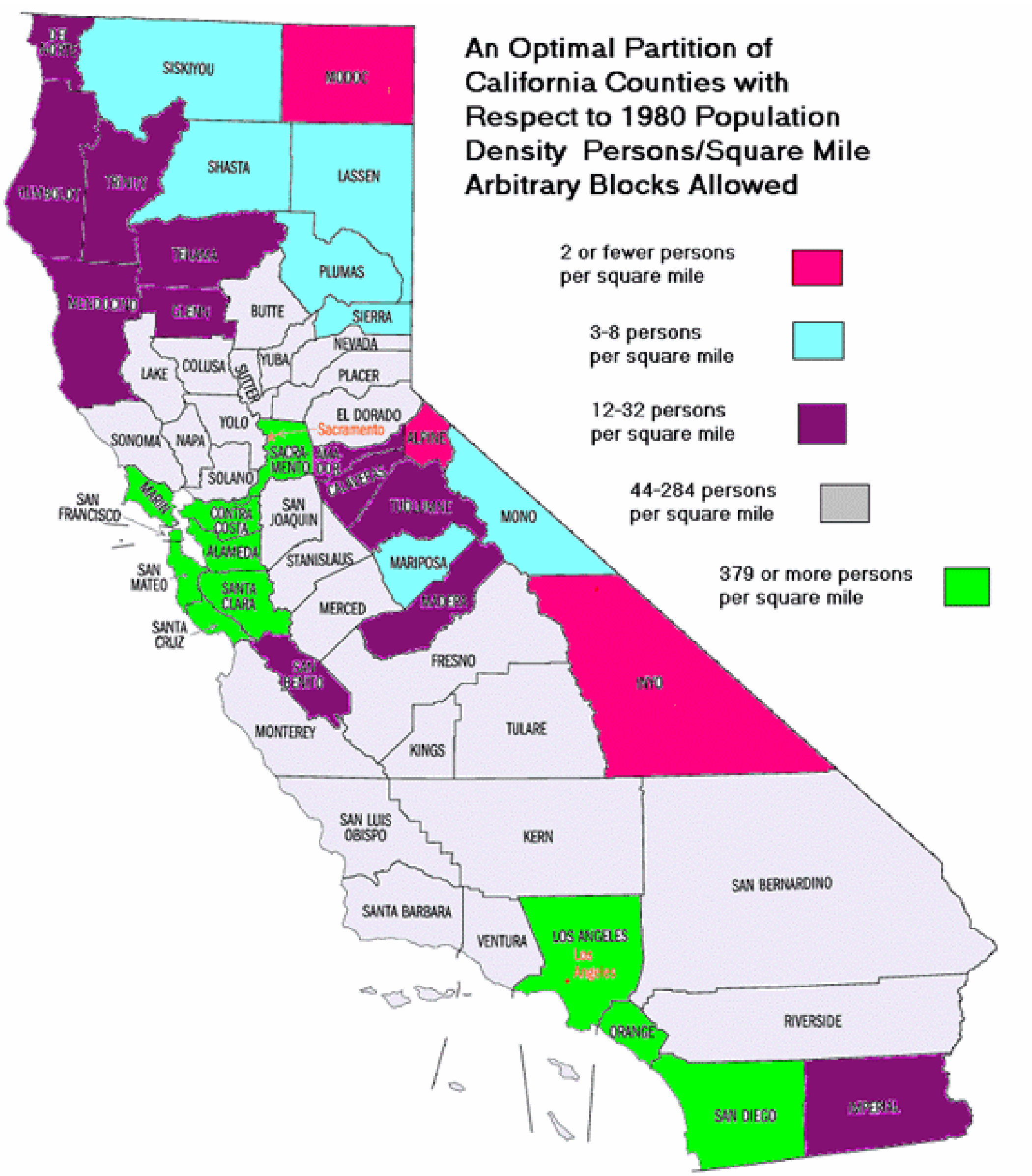
- Connected (can't break into distinct parts)
- Not constrained to be connected

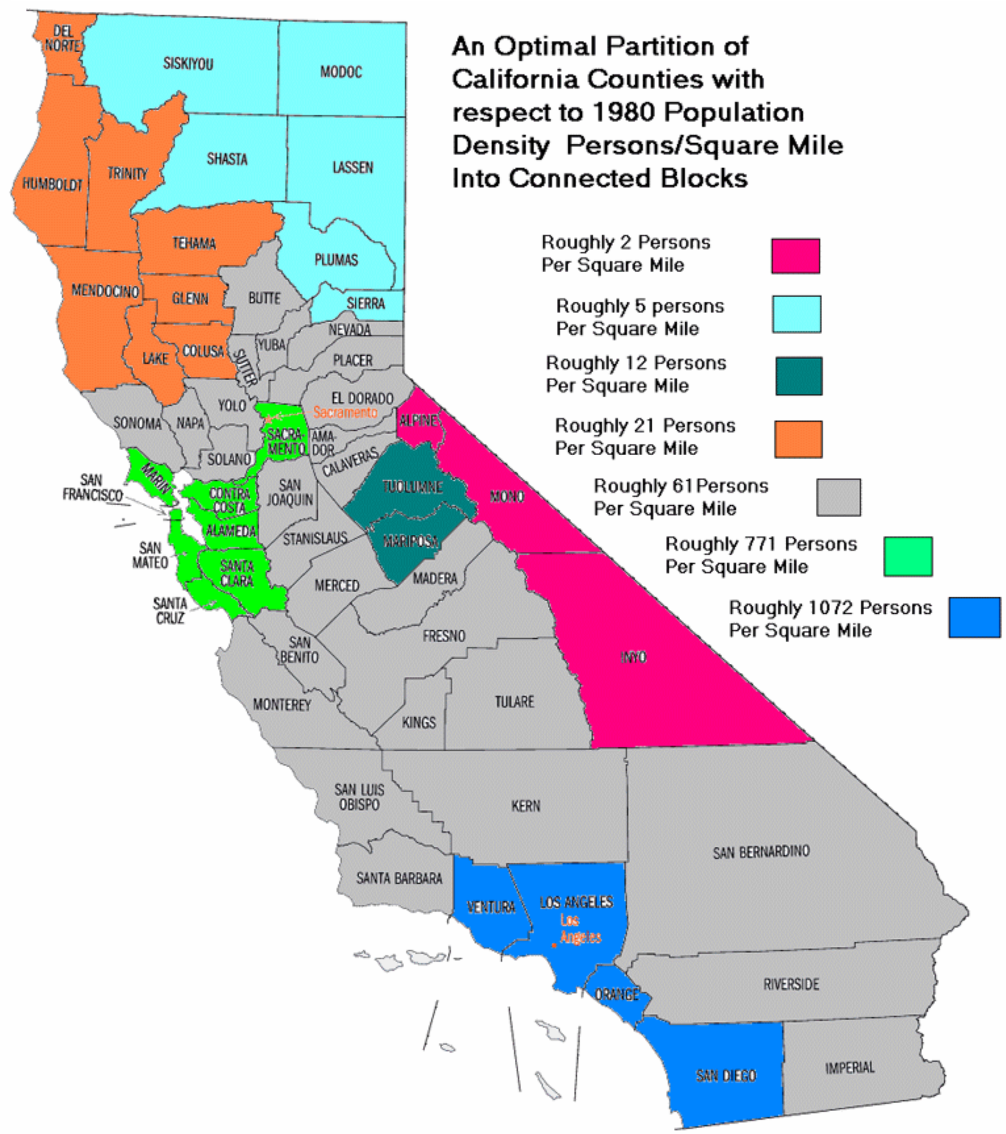
Model = set of blocks

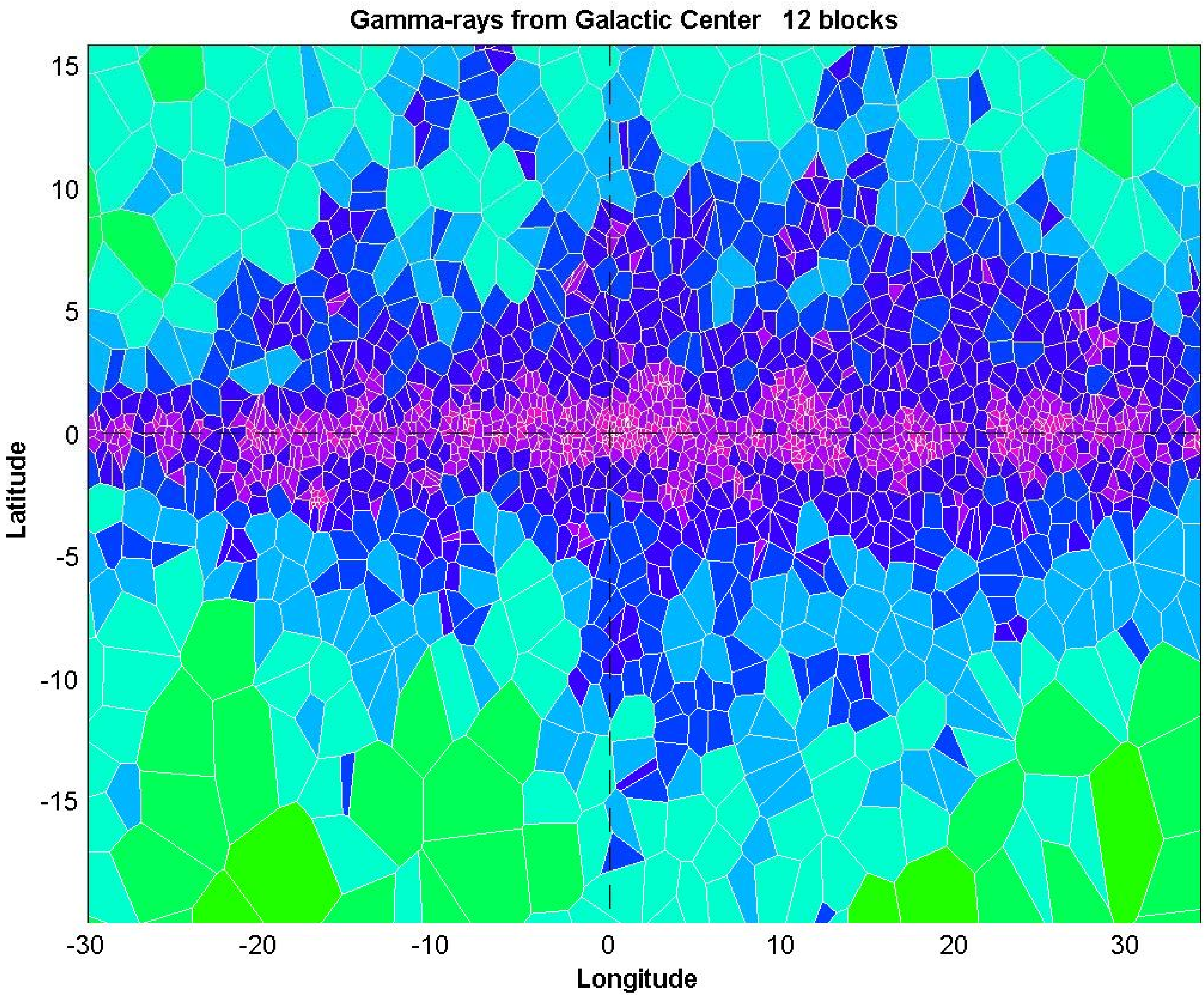
<u>Fitness function:</u>

F(Model) = sum over blocks F(Block)

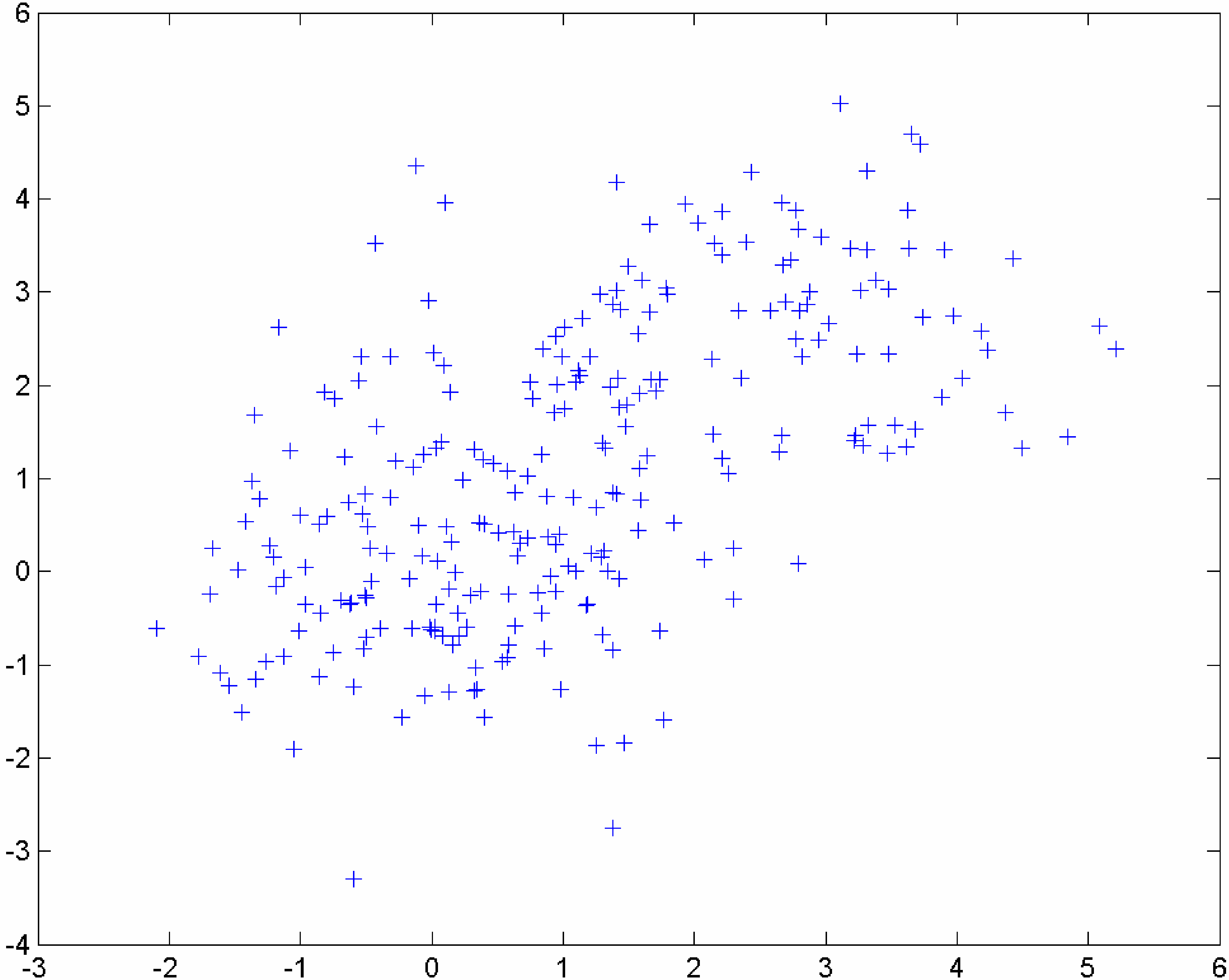
Connected vs. Arbitrary Blocks

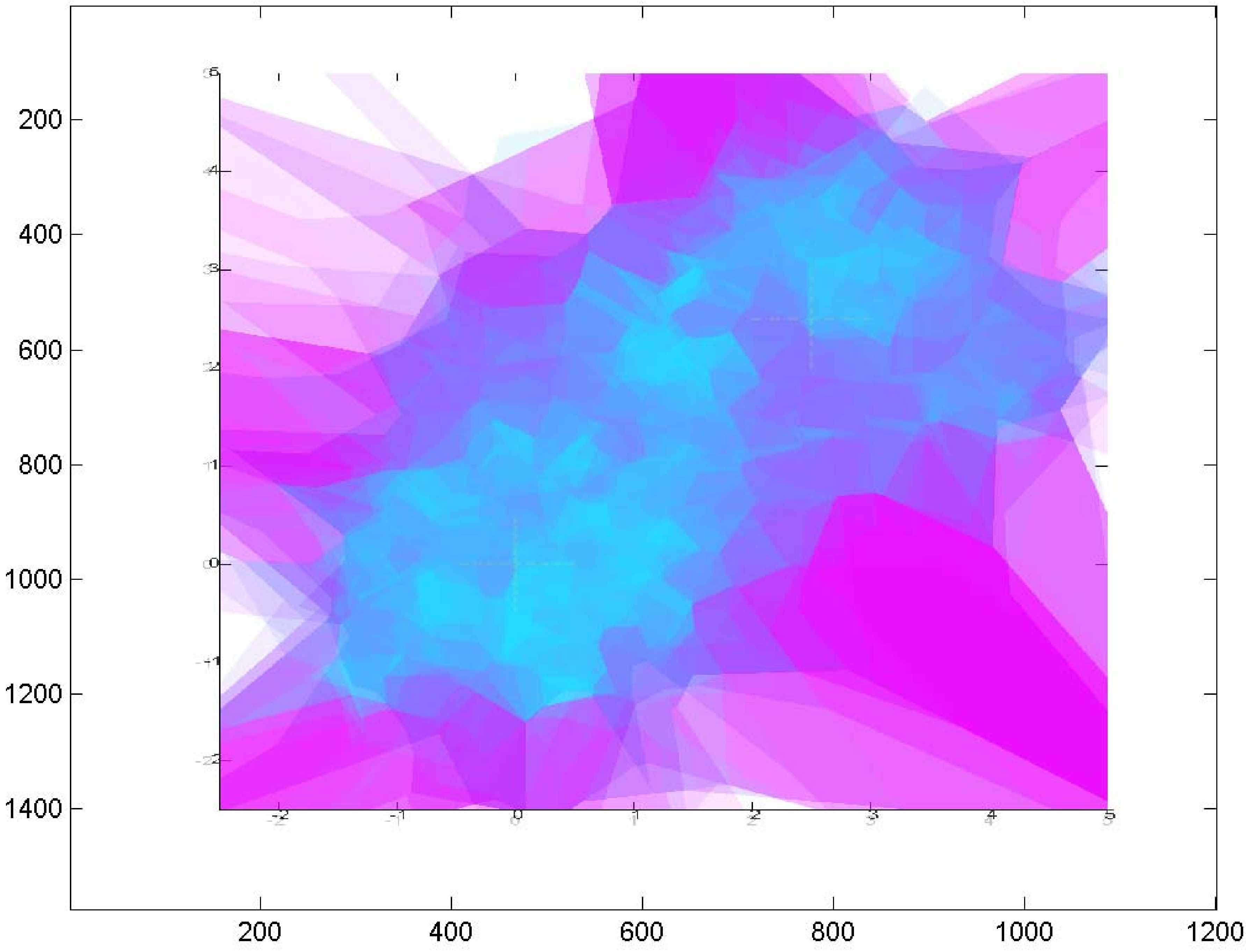




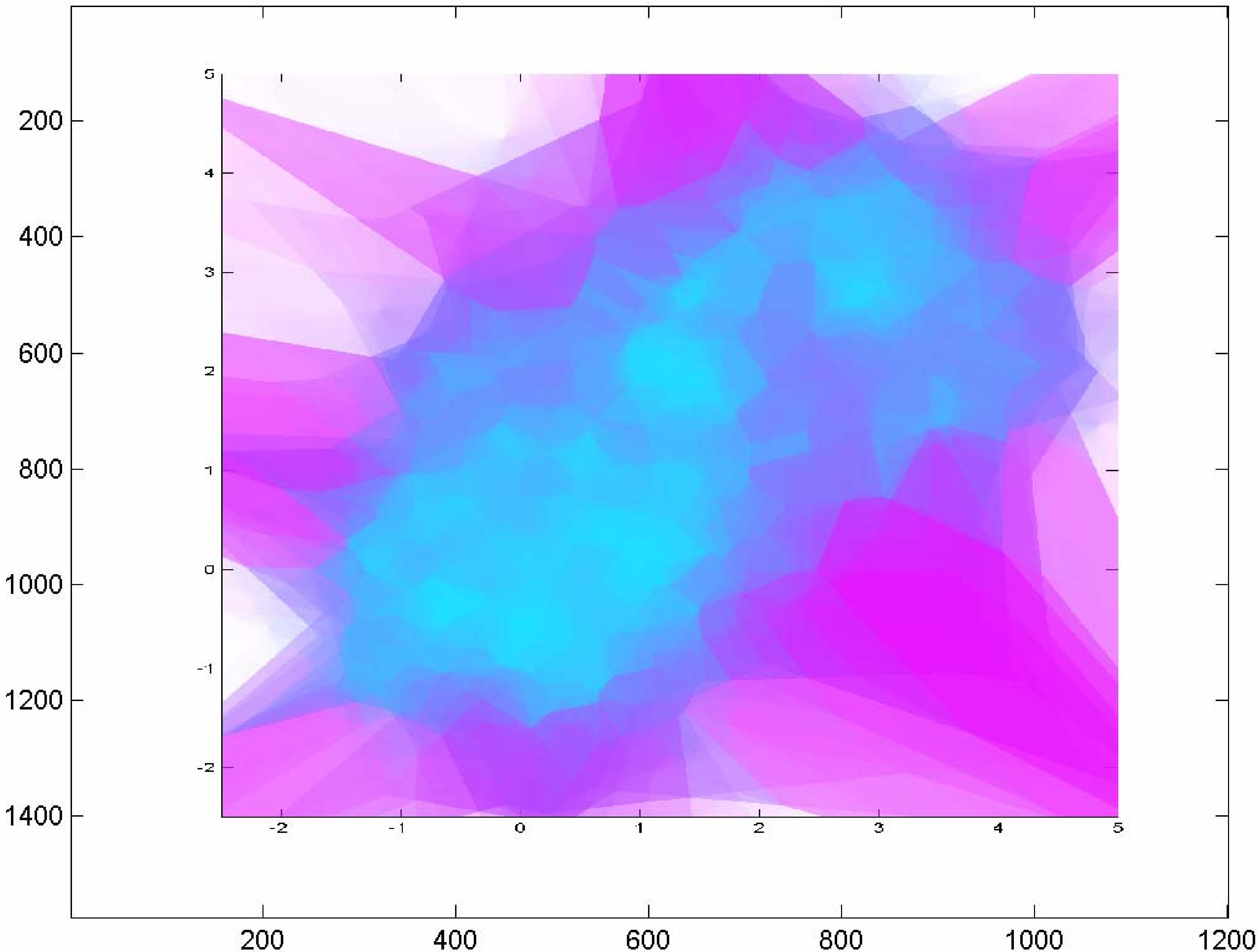

Gamma-rays from Galactic Center 12 blocks
Latitude
15
10
5
0
-5
-10
-15
-30
-20
-10
0
10
20
30
Longitude

2D Synthetic Bootstrap Example: Raw Data

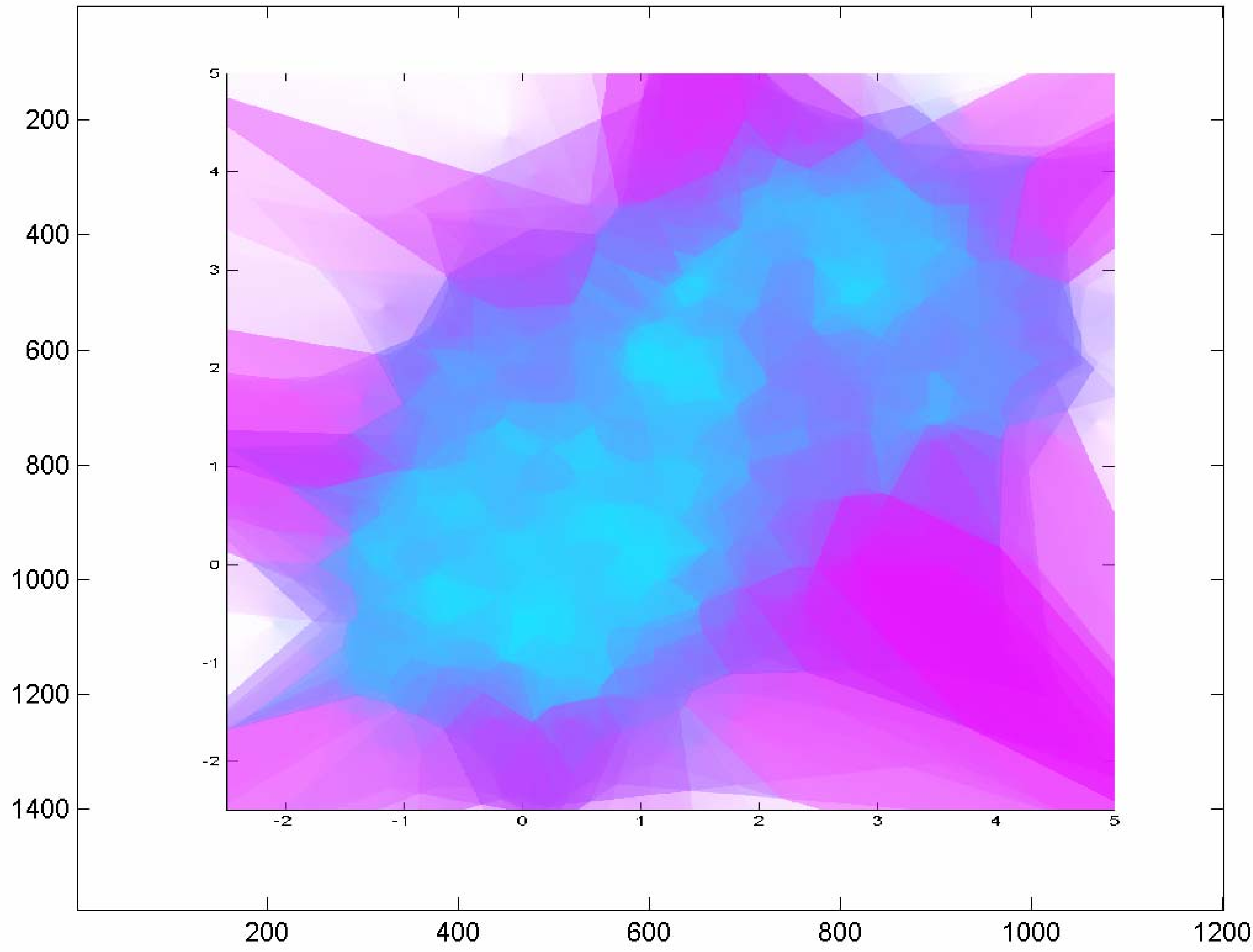




200
400
600
800
1000
1200
1400
200
400
600
800
1000
1200
5
4
3
2
1
0
-1
-2
-2
-1
0
1
2
3
4
5

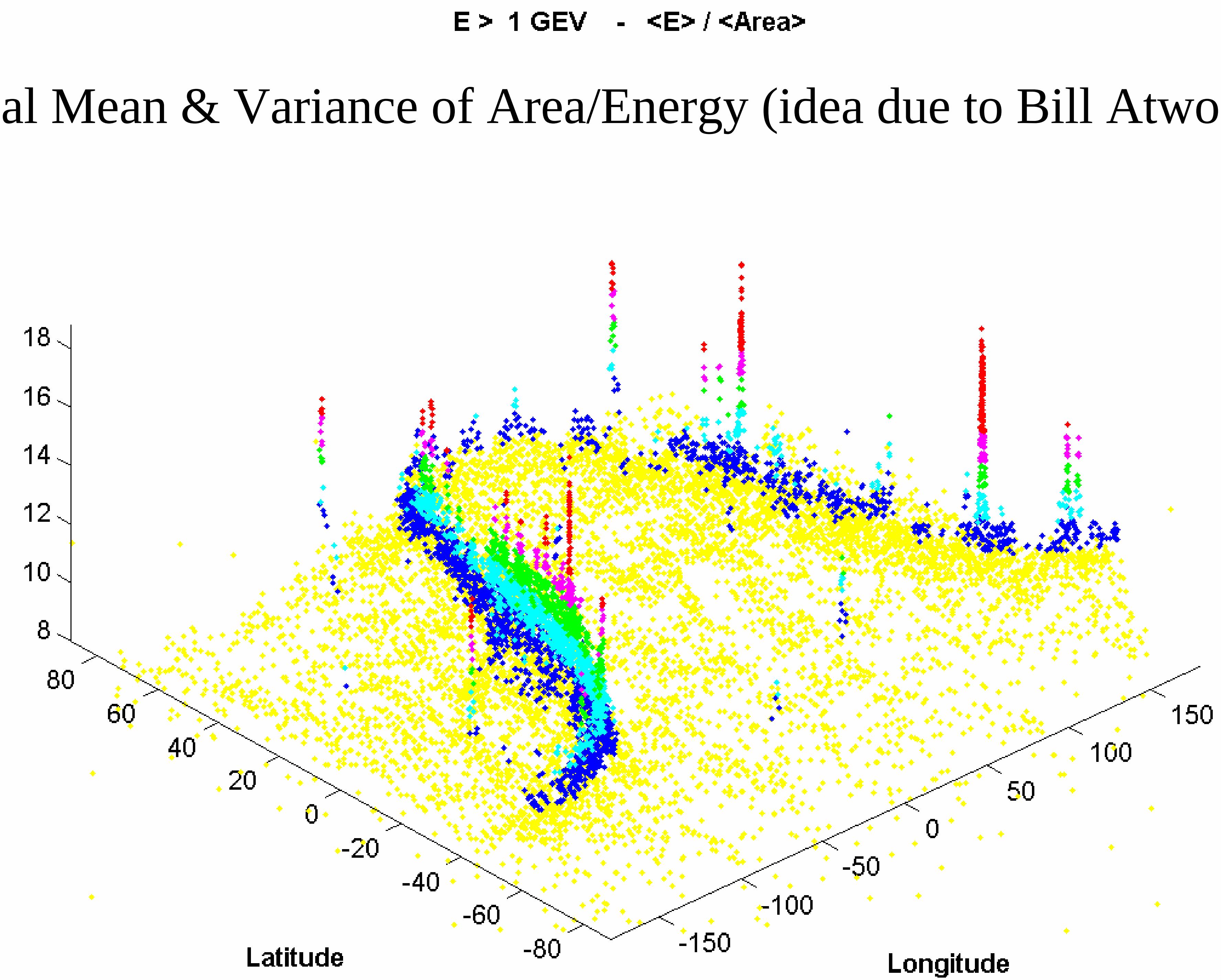


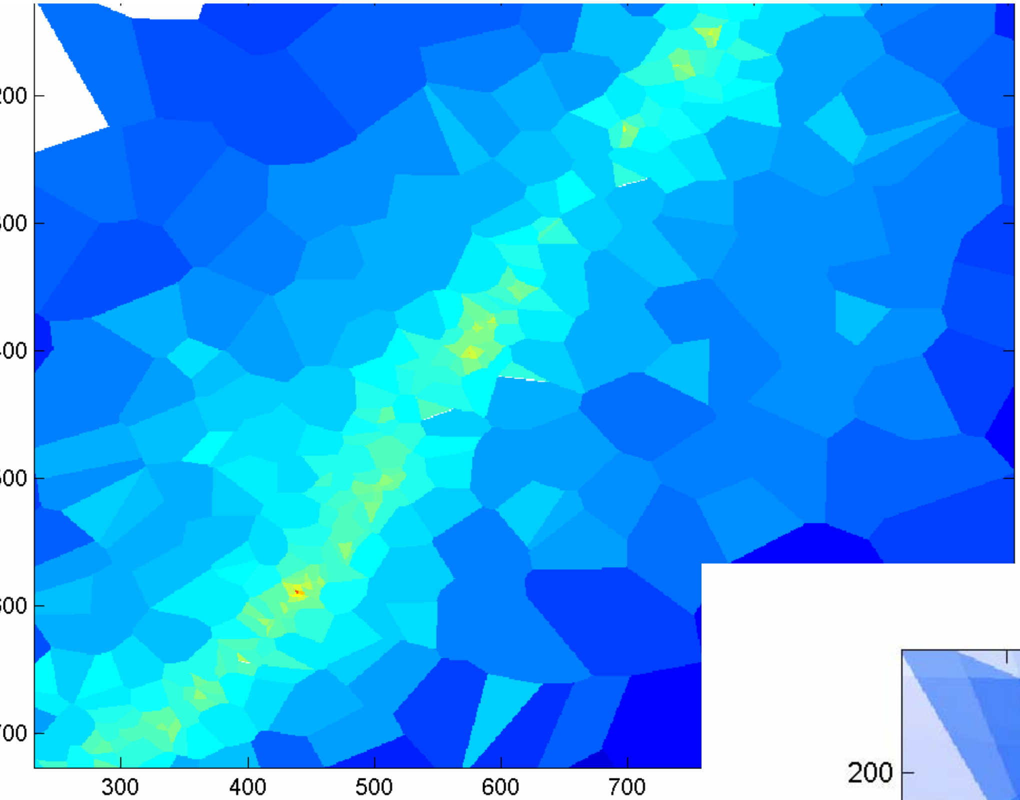
200
400
600
800
1000
1200
1400
200
400
600
800
1000
1200
5
4
3
2
1
0
-1
-2
-2
-1
0
1
2
3
4
5



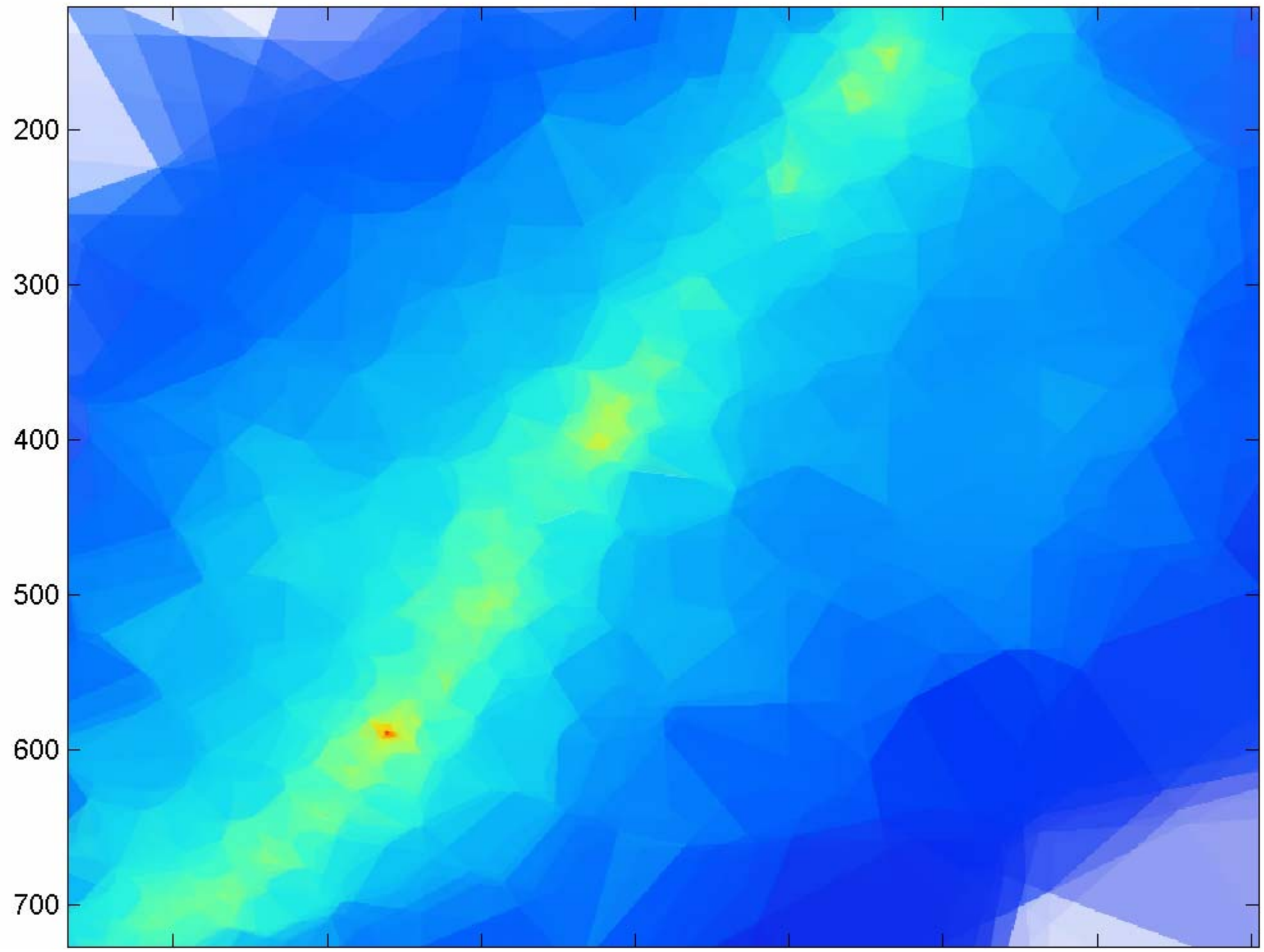
200
400
600
800
1000
1200
1400
200
400
600
800
1000
1200
5
4
3
2
1
0
-1
-2
-2
-1
0
1
2
3
4
5

Local Mean & Variance of Area/Energy (idea due to Bill Atwood)

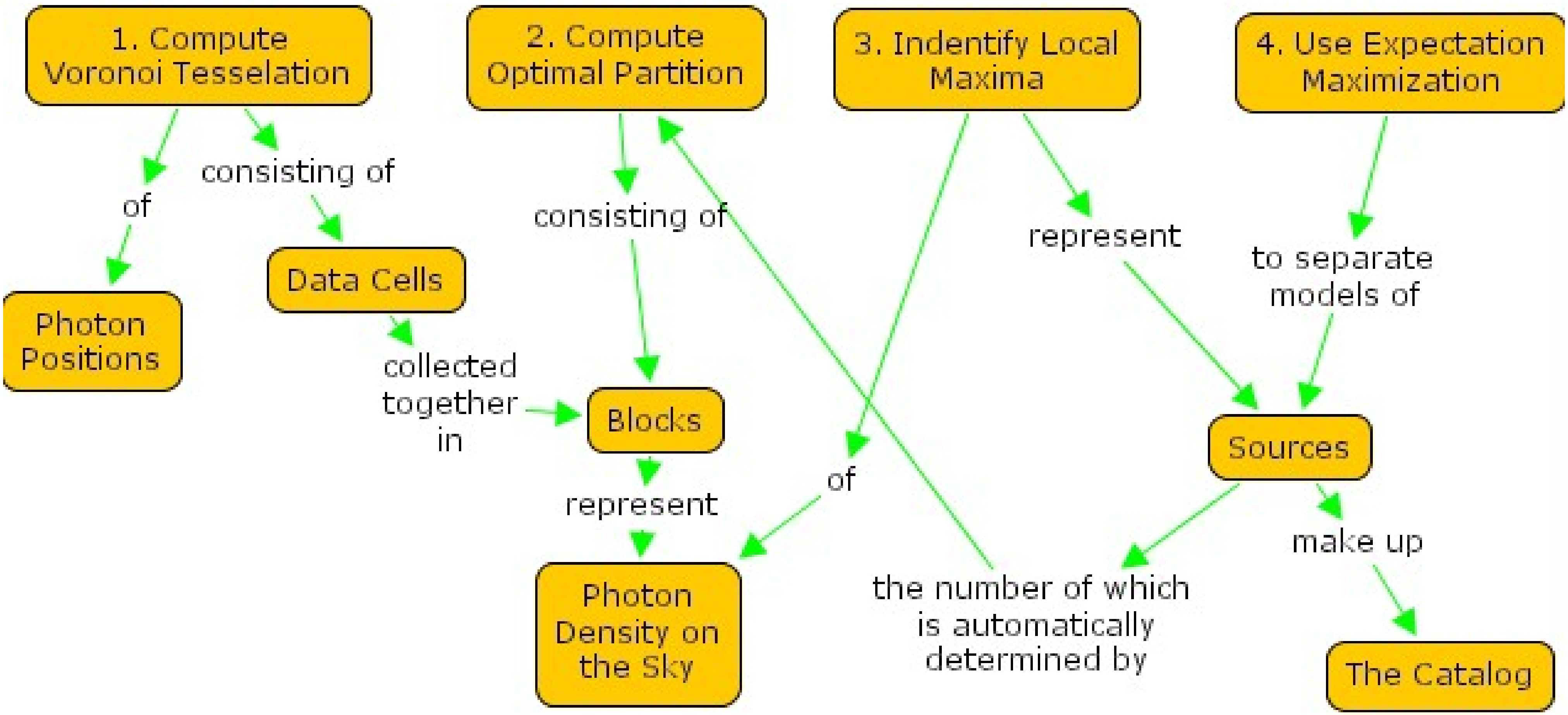




200
300
400
500
600
700
300
400
500
600
700



200
300
400
500
600
700


1. Compute Voronoi Tesselation
2. Compute Optimal Partition
3. Indentify Local Maxima
4. Use Expectation Maximization
of
consisting of
Photon Positions
Data Cells
consisting of
collected together in
Blocks
represent
Photon Density on the Sky
of
represent
to separate models of
Sources
make up
the number of which is automatically determined by
The Catalog