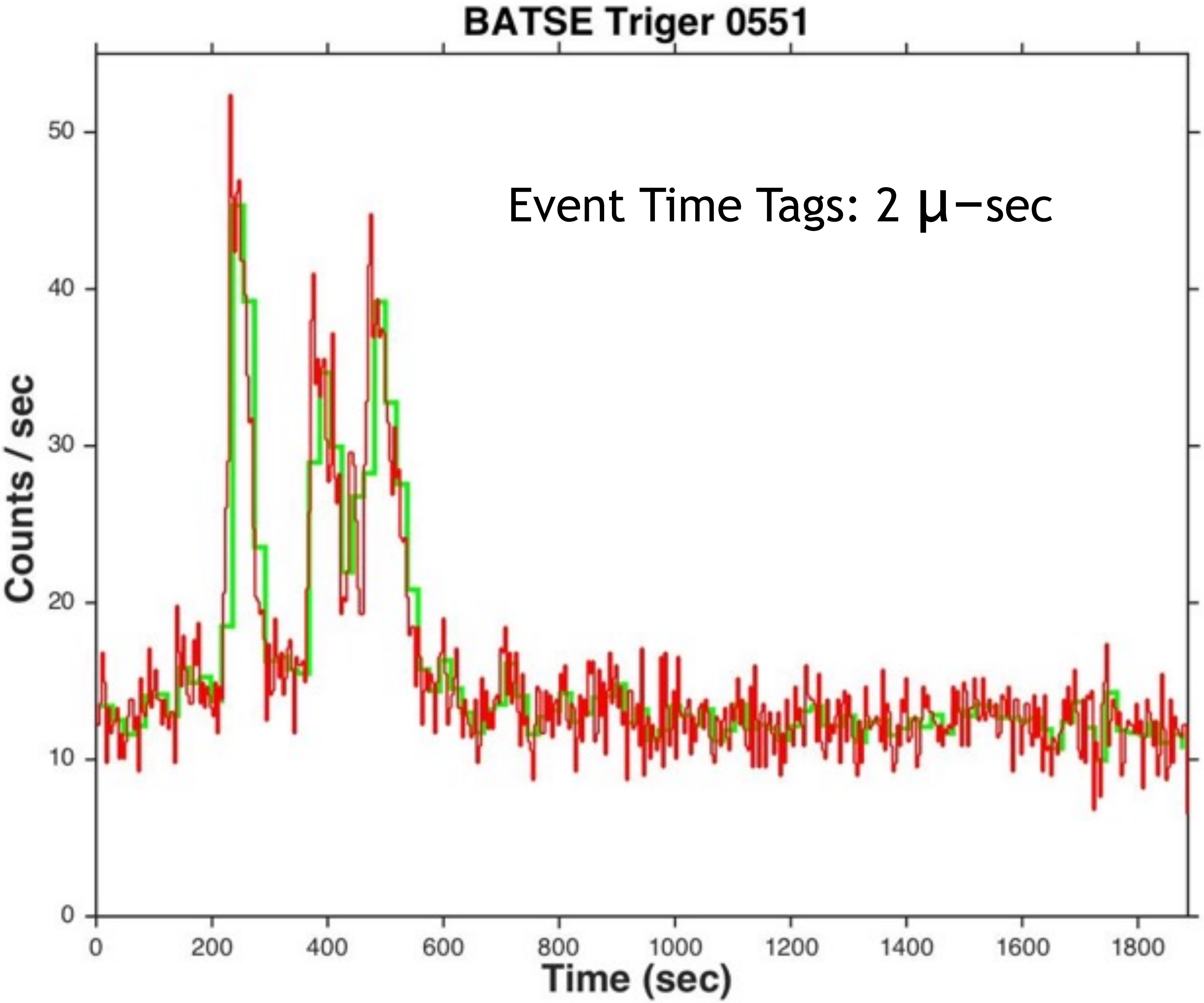
Optimal Segmentation with Pruned Dynamic Programming

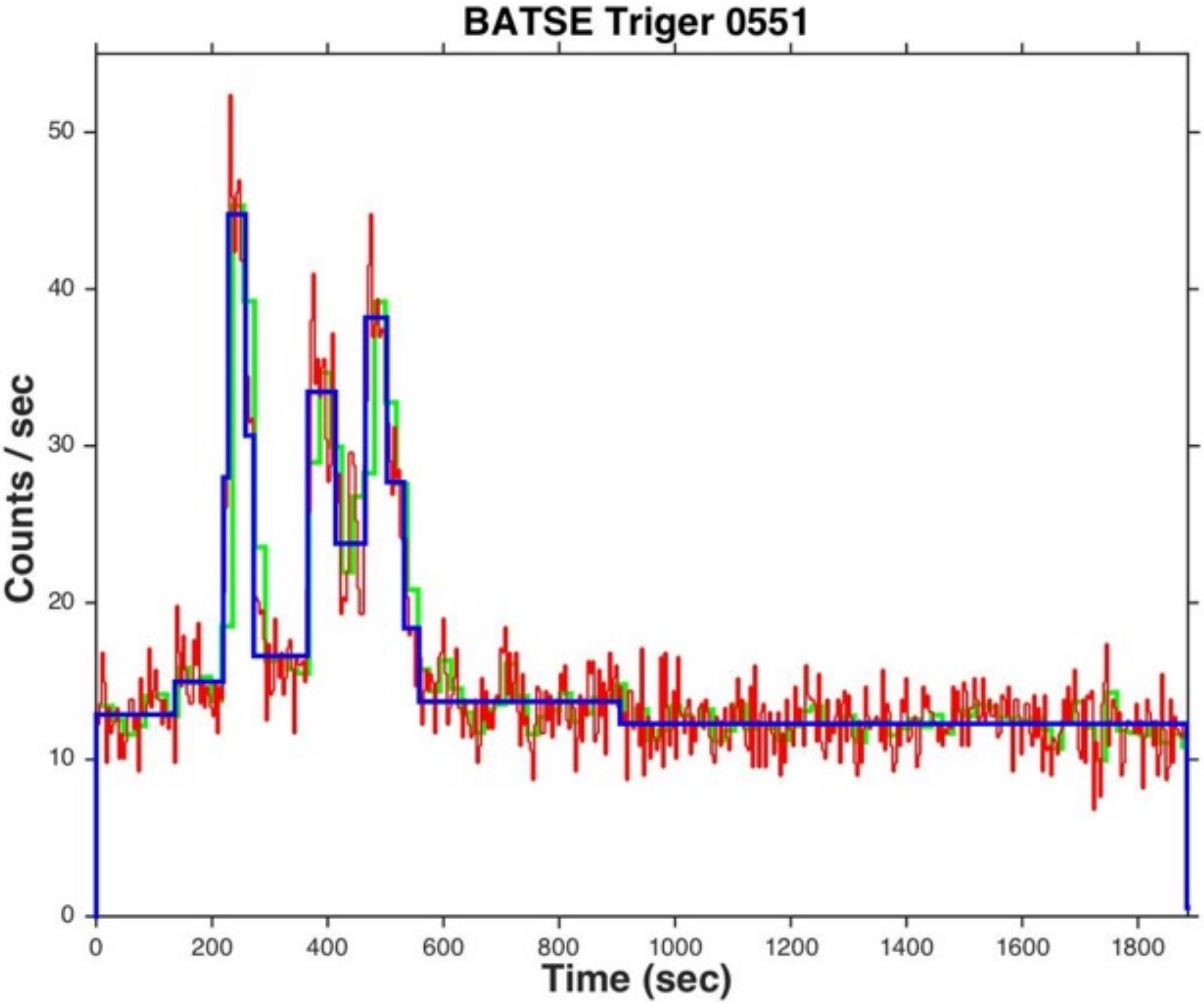
Jeff Scargle
NASA Ames Research Center
Jeffrey.D.Scargle@nasa.gov

Thanks: Brad Jackson + Students, Jay Norris, Jim Chiang

AI @ SLAC September 26, 2017



Event Time Tags: 2 μ−sec


BATSE Triger 0551
Counts / sec
Time (sec)
0
10
20
30
40
50
0
200
400
600
800
1000
1200
1400
1600
1800

What is Optimal Segmentation?

DATA:

- any signal measurements
- sequential in time, space, energy, …

MODEL:

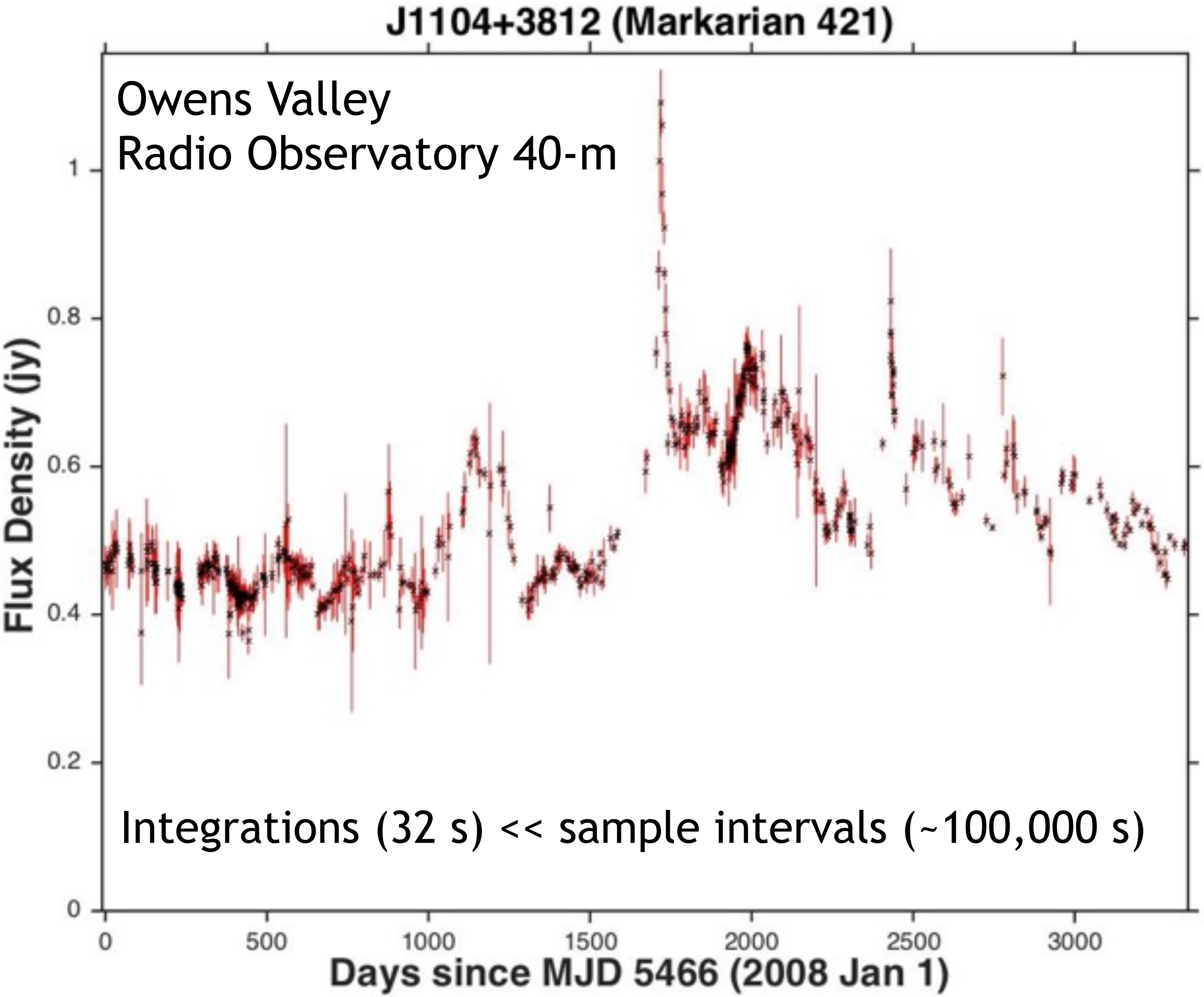
- partition data interval into blocks
- block = set of consecutive data points
- model of signal in each block
- block cost (aka fitness, risk, etc.)

OPTIMIZATION:

- total model cost = sum of block costs
- optimize over all possible partitions ($\sim 2^N$)

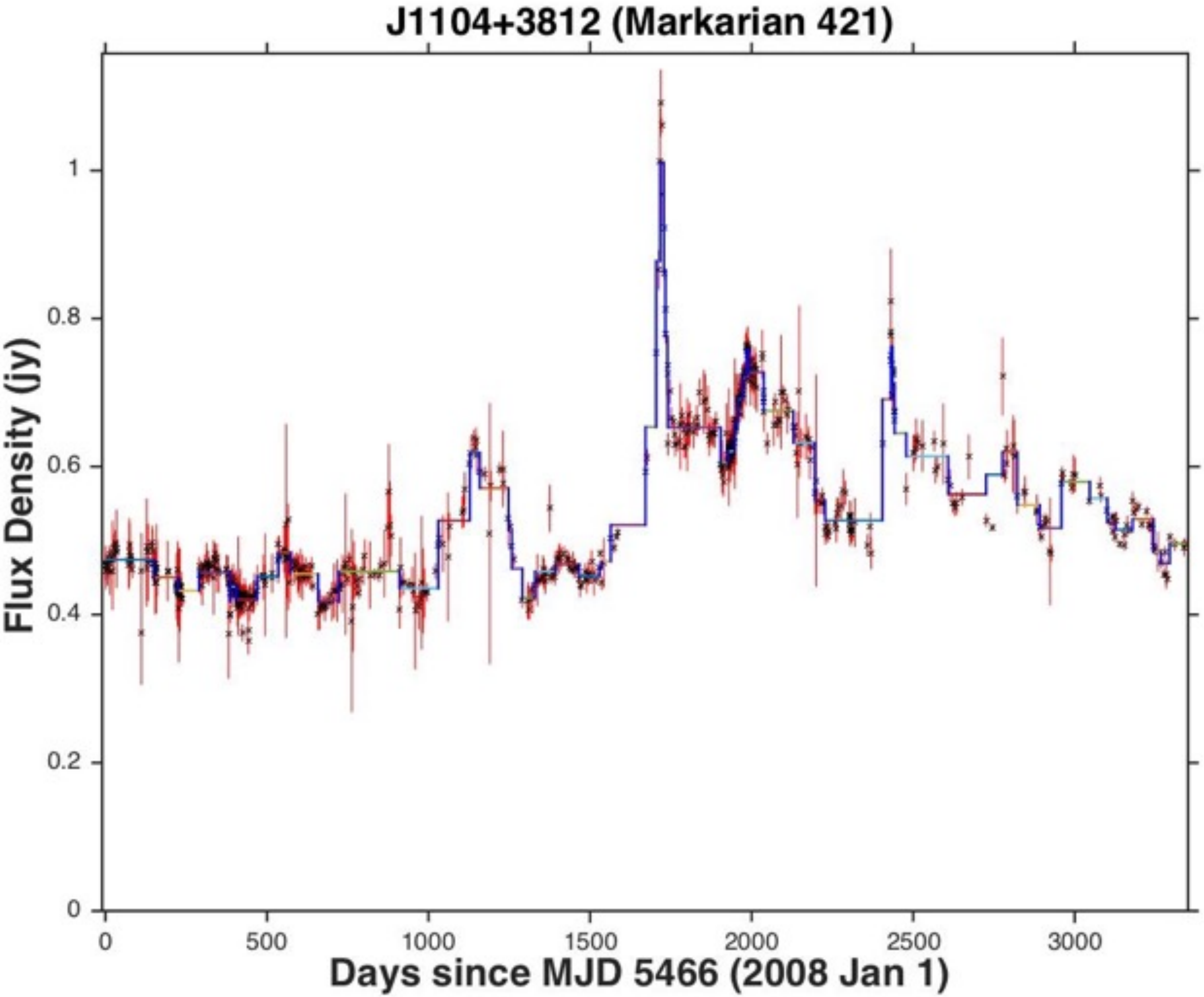
What good are Segmented Time Series Representations? (1)

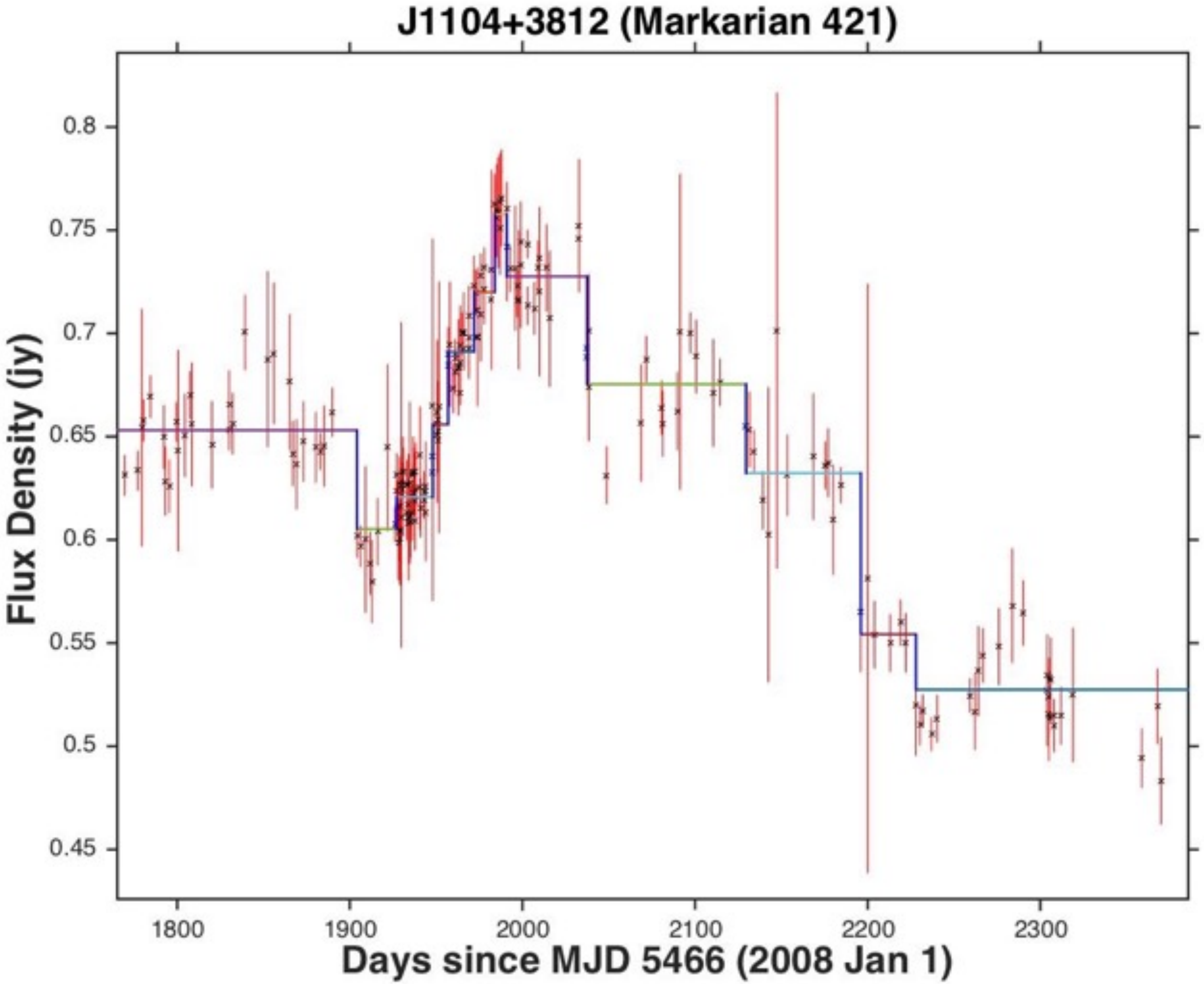
Detect and characterize any and all statistically signficiant variability supported by the data.



Owens Valley
Radio Observatory 40-m

Integrations (32 s) << sample intervals (~100,000 s)


J1104+3812 (Markarian 421)
Flux Density (jy)
0
0.2
0.4
0.6
0.8
1
0
500
1000
1500
2000
2500
3000
Days since MJD 5466 (2008 Jan 1)


J1104+3812 (Markarian 421)
Flux Density (jy)
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
1800
1900
2000
2100
2200
2300
Days since MJD 5466 (2008 Jan 1)

Bayesian Block Generalizations

Any Data Mode:

- *point measurements*
- *time-tagged events*
- *live-time intervals*
- *categorical*
- *gaps, uneven sampling, exposure variation, real-time*
- *… any data mode*

Any Block Shape:

- *constant*
- *linear*
- *exponential*
- *FRED (fast rise/exponential decay) /DERF*
- *… any block shape*

Improved speed:

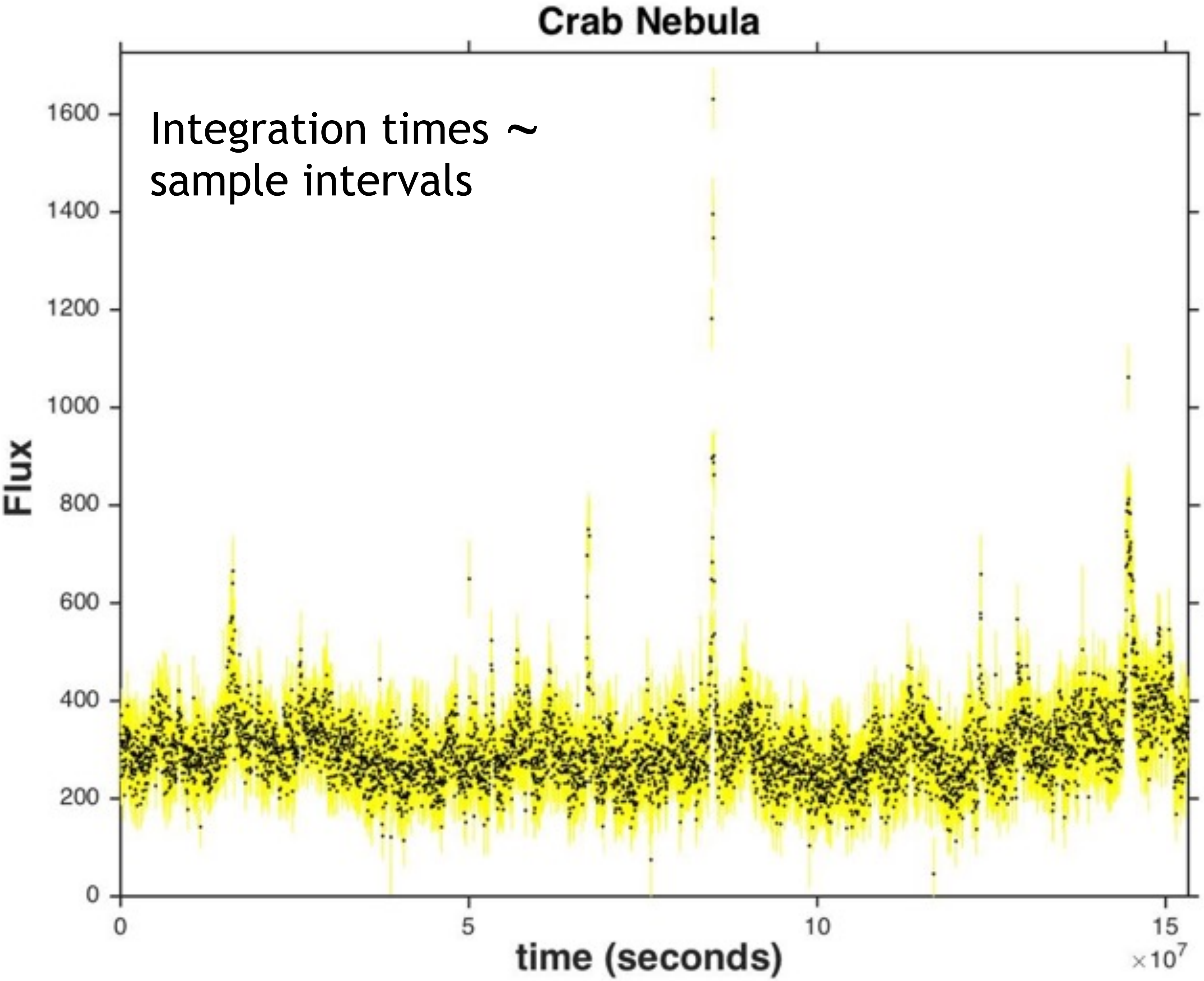
$O(N^2) \rightarrow \sim N$

Data Points

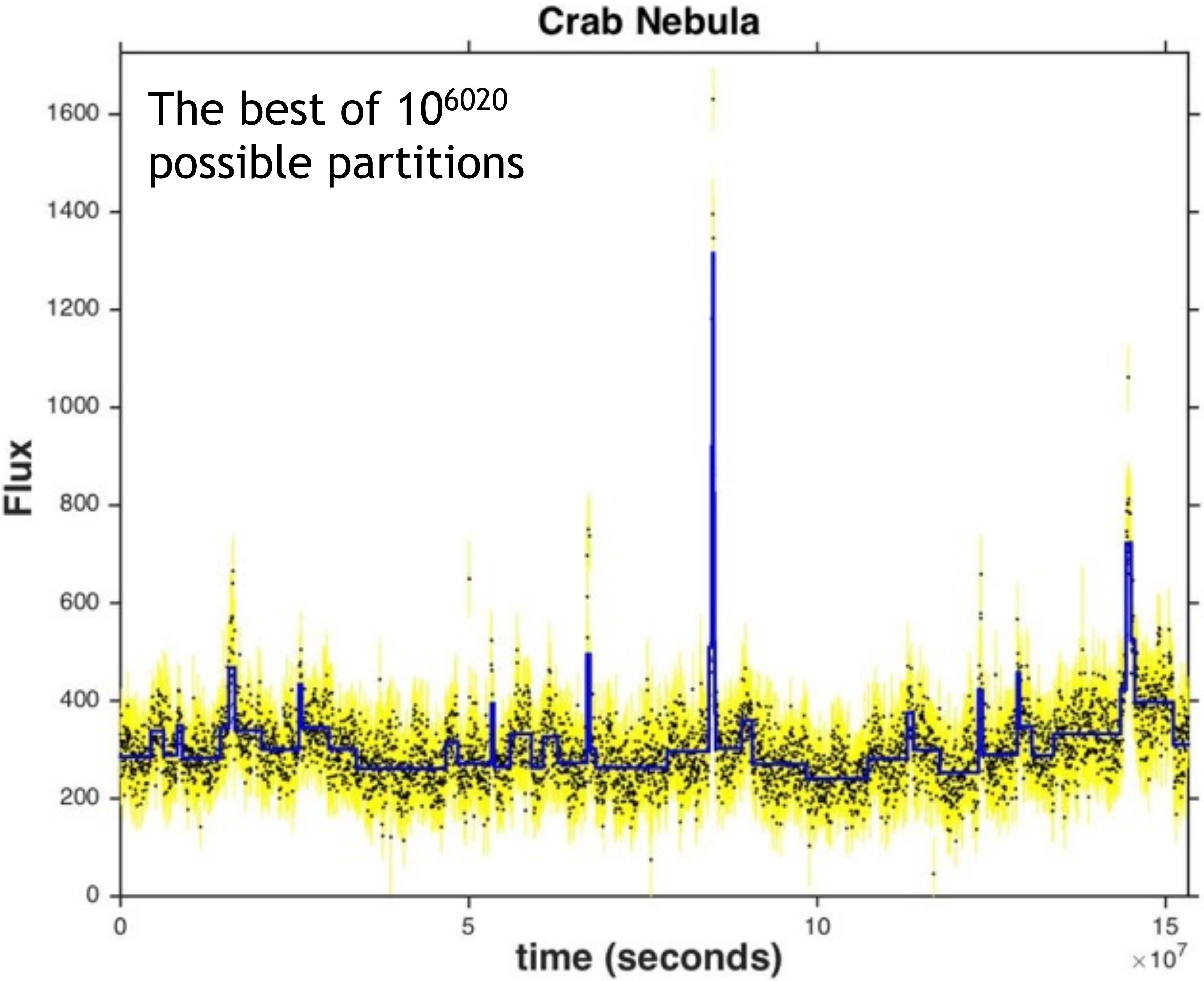
… must include whatever information is necessary to compute the *objective function* of a data block (Typically: time + intensity + error)

- measurements at a point in time - $X(t_i)$, $\sigma (t_i)$
- measurements averaged over a time interval
- event times (TTE)
- inter-event intervals
- live-time intervals
- auxiliary information (e.g. color)
- categorical (e.g. 0-1, ACGT, …)

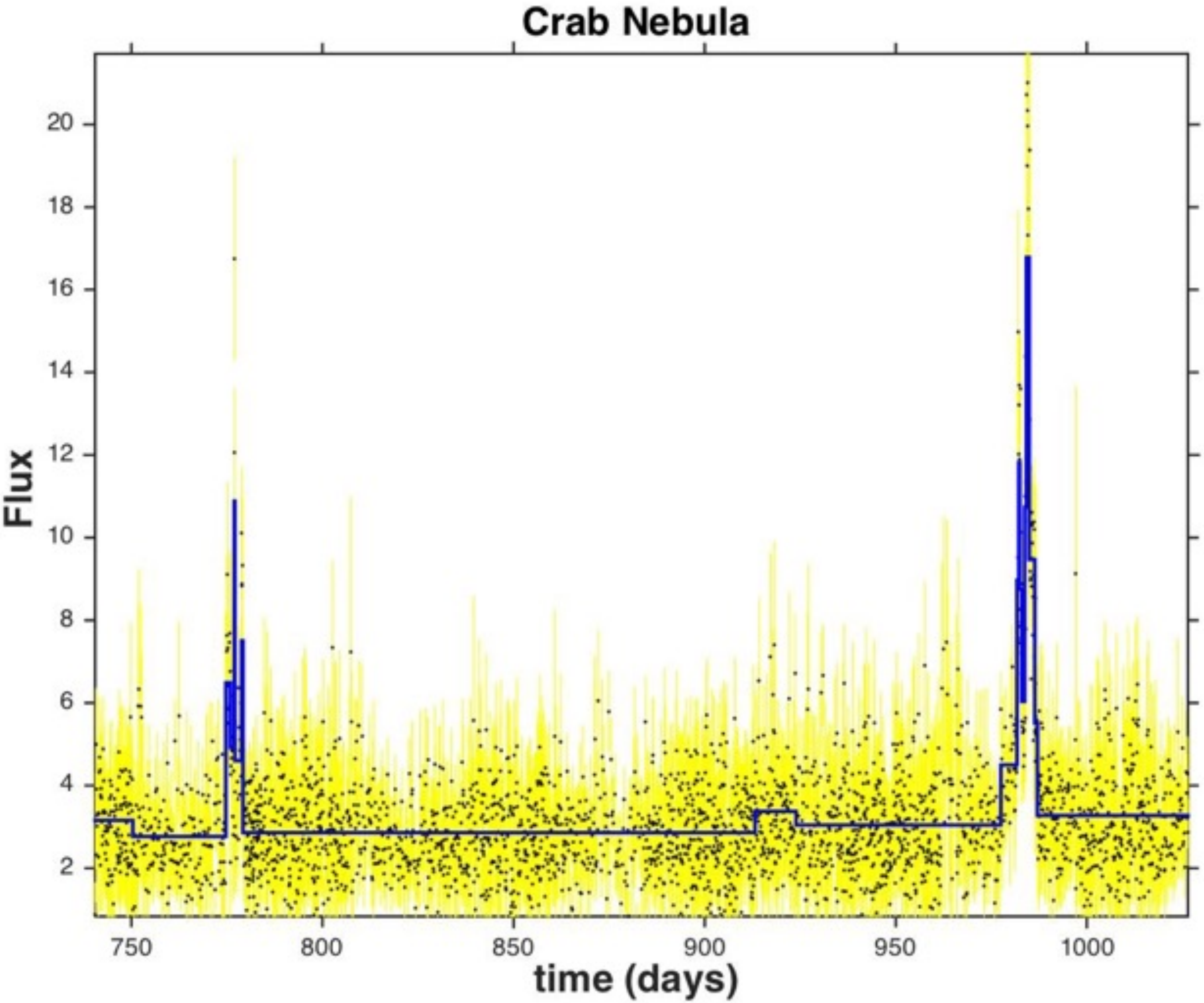
- multivariate (multi-wavelength, mulit-messenger)
- mixtures of any of the above
- data on the circle (e.g. angles)
- higher dimension (e.g. 3D galaxy positions)
- …

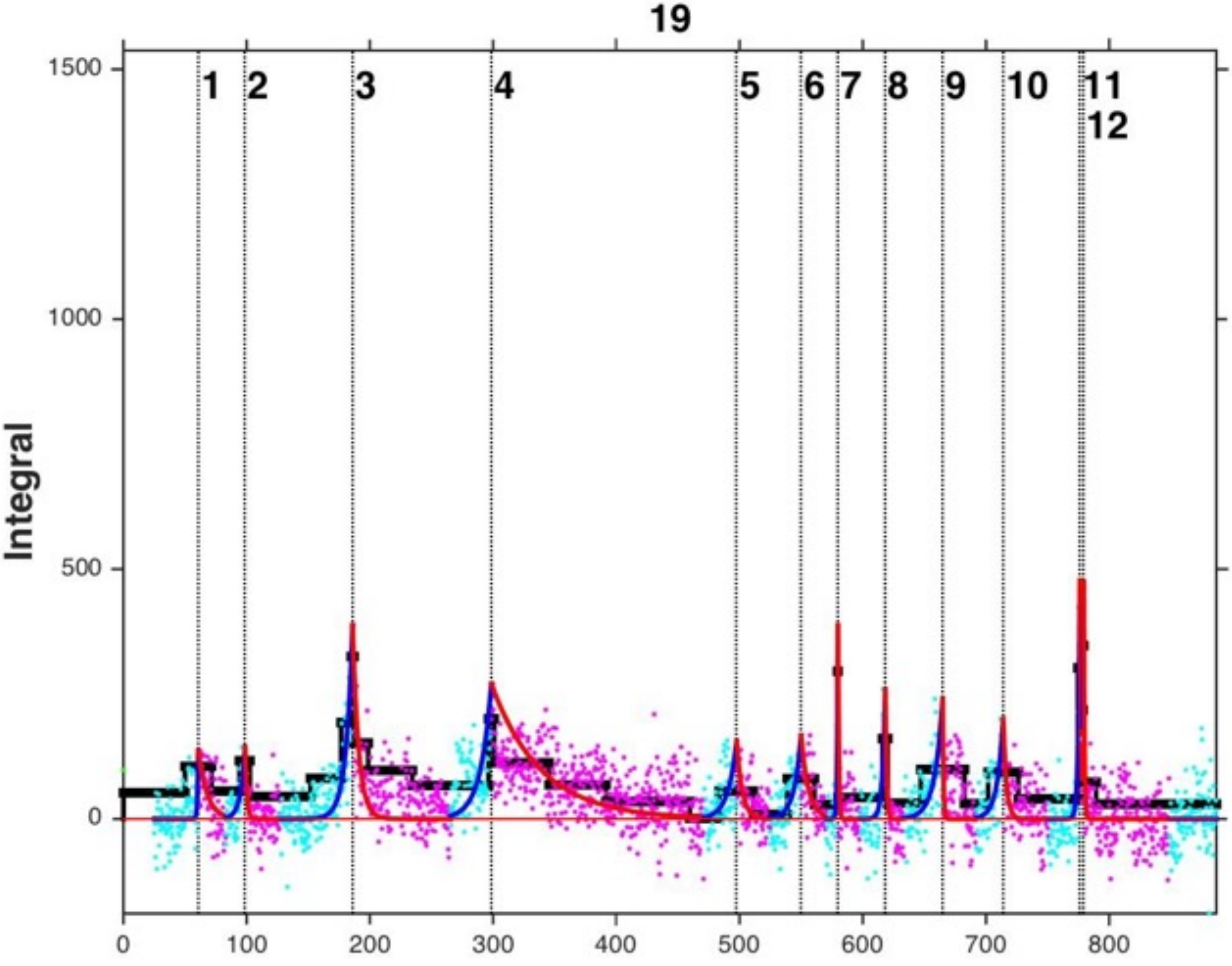


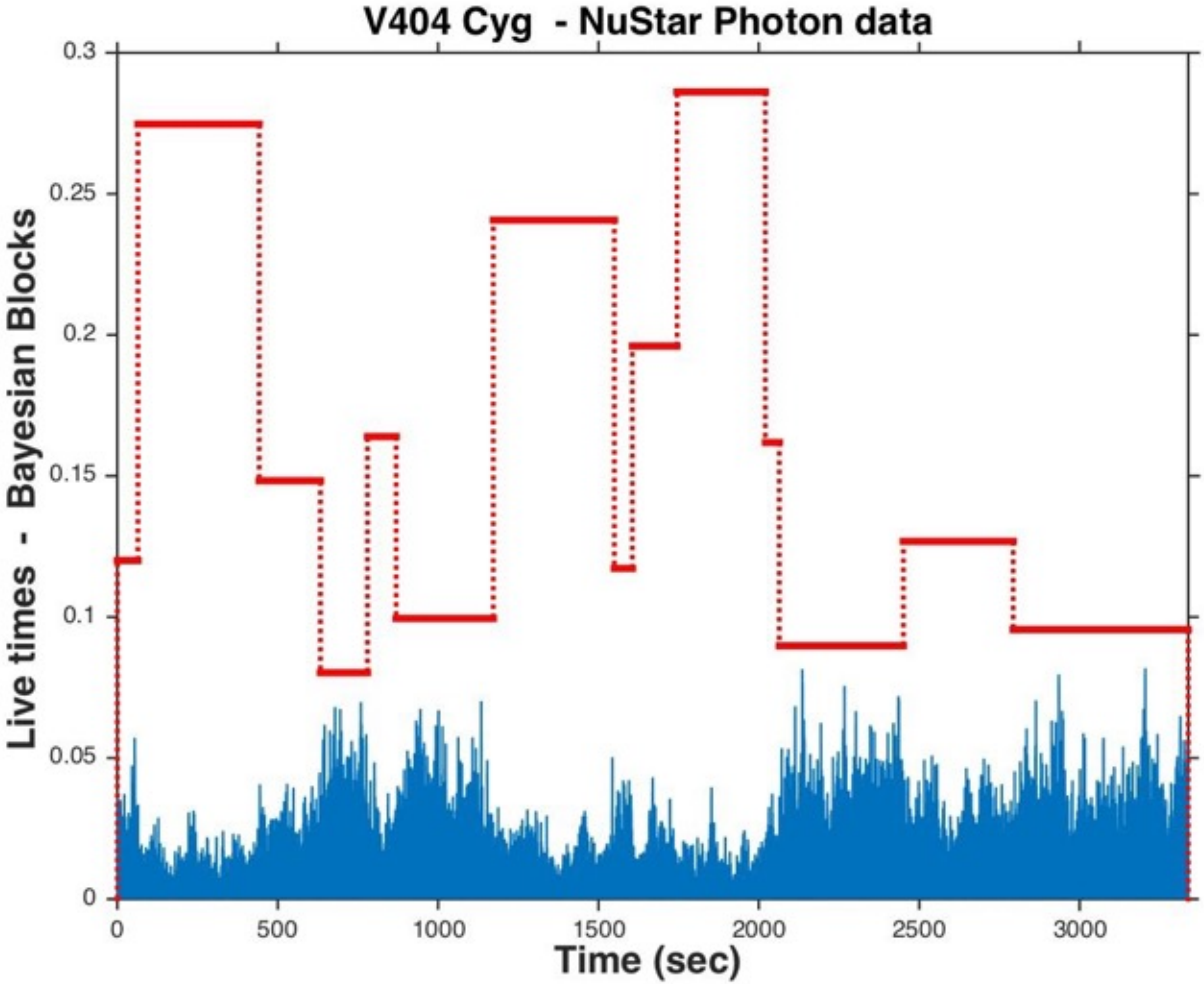
Integration times ~ sample intervals



The best of 10^{6020} possible partitions


Crab Nebula
Flux
20
18
16
14
12
10
8
6
4
2
750
800
850
900
950
1000
time (days)


19
1
2
3
4
5
6
7
8
9
10
11
12
Integral
1500
1000
500
0
0
100
200
300
400
500
600
700
800

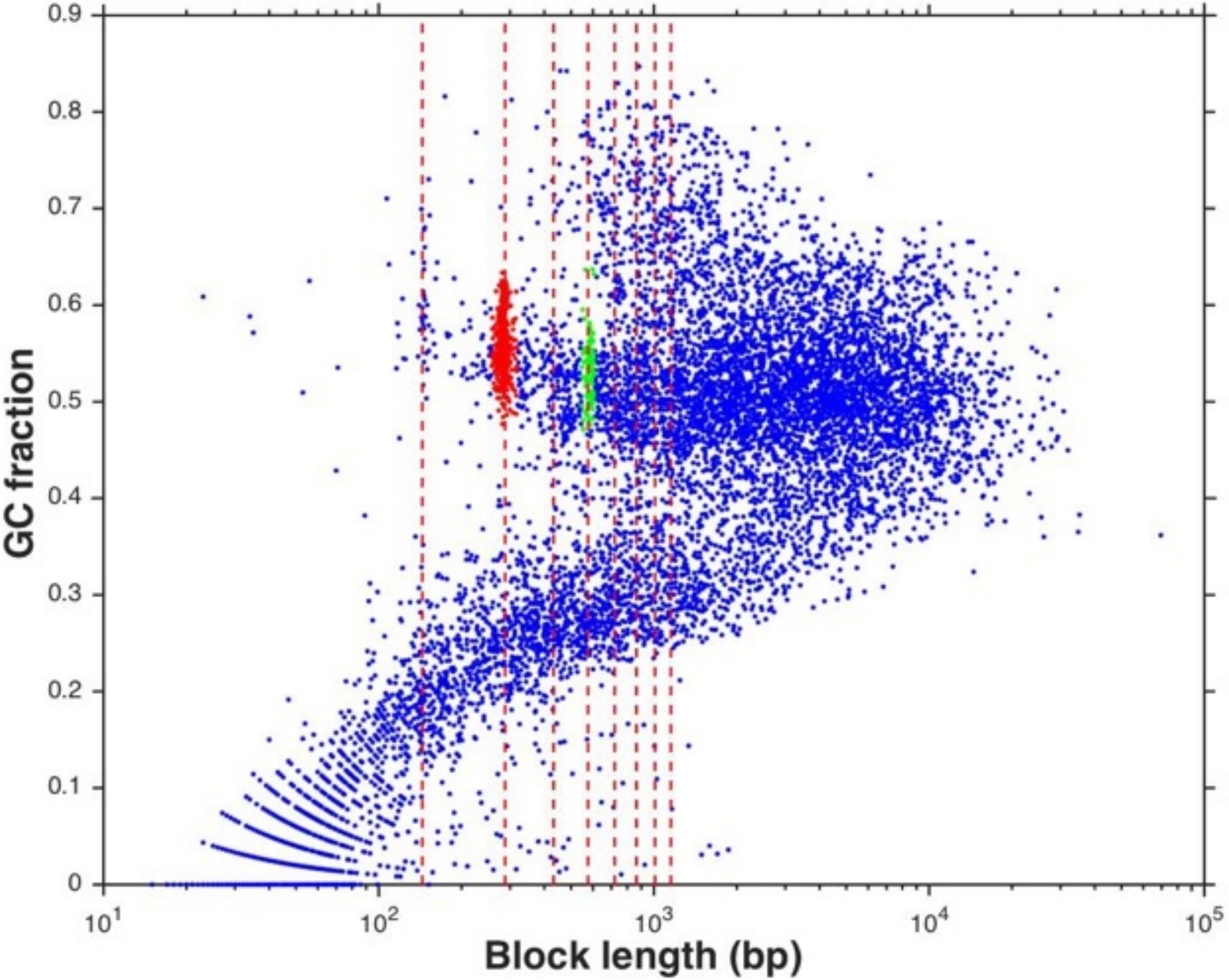

V404 Cyg - NuStar Photon data
Live times - Bayesian Blocks
Time (sec)
0
0.05
0.1
0.15
0.2
0.25
0.3
0
500
1000
1500
2000
2500
3000

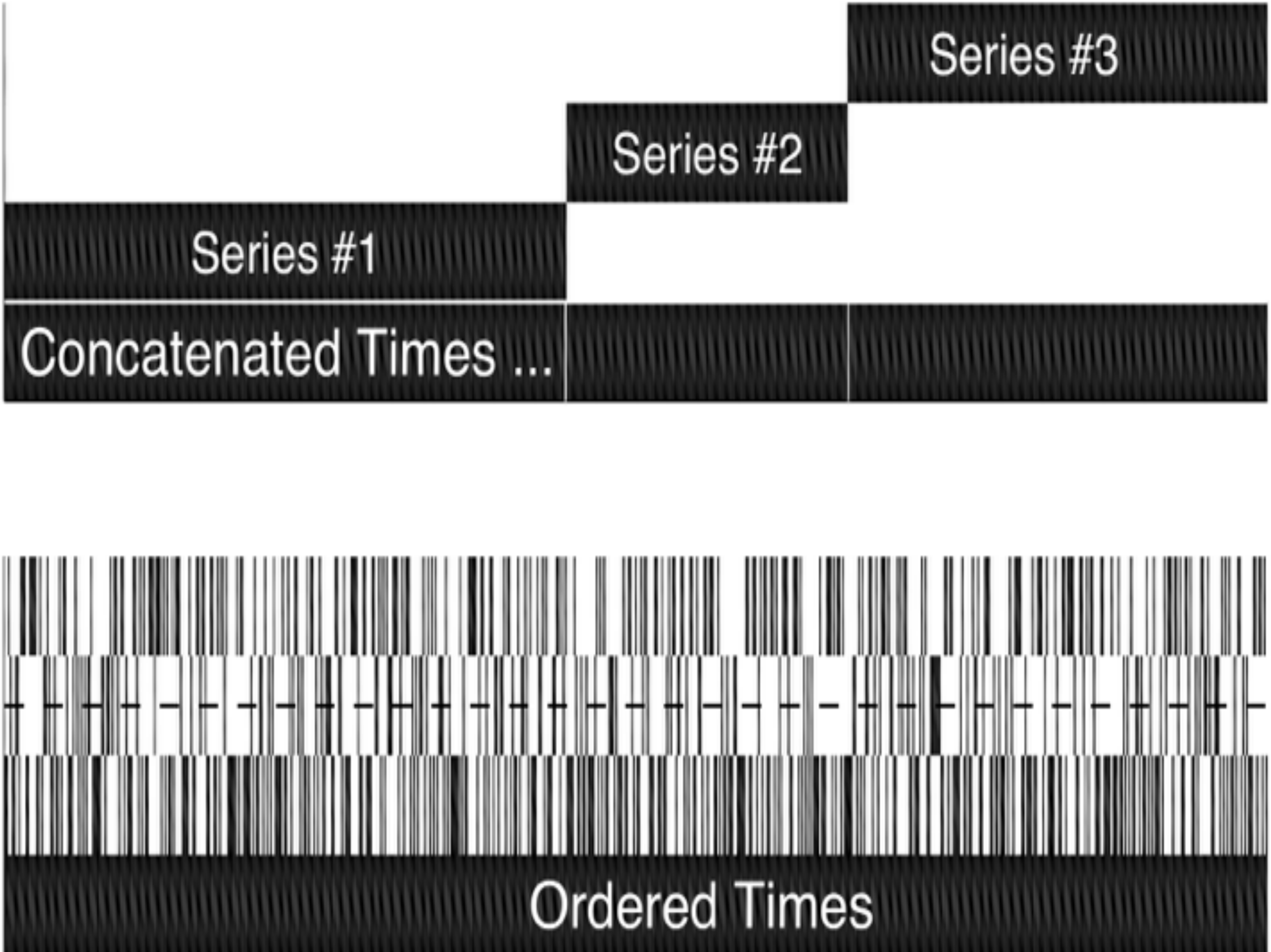
The Human Genome

ATCAGAGACTCGTACCCGTGCCTGAGCAGTGC
TGAAGAGGCACTCGTTTGGAAAGGGGGCCCAT
CCCCCGGGACTTCGGACACTCCTGGCTGAAGC
ATAACGTGTAGGCGCTCTAGAGGGCTCGCCTA
CCCGAGCCCACATACCAGCGATGATGAGATCG
ACCCGCACCGACGCATGGAGCGCAGGCGTGCT
GCTAGTCGAGGGCACGGTCGCCCCAACAAACG
ACCGCCCA … N ~100,000,000 bp

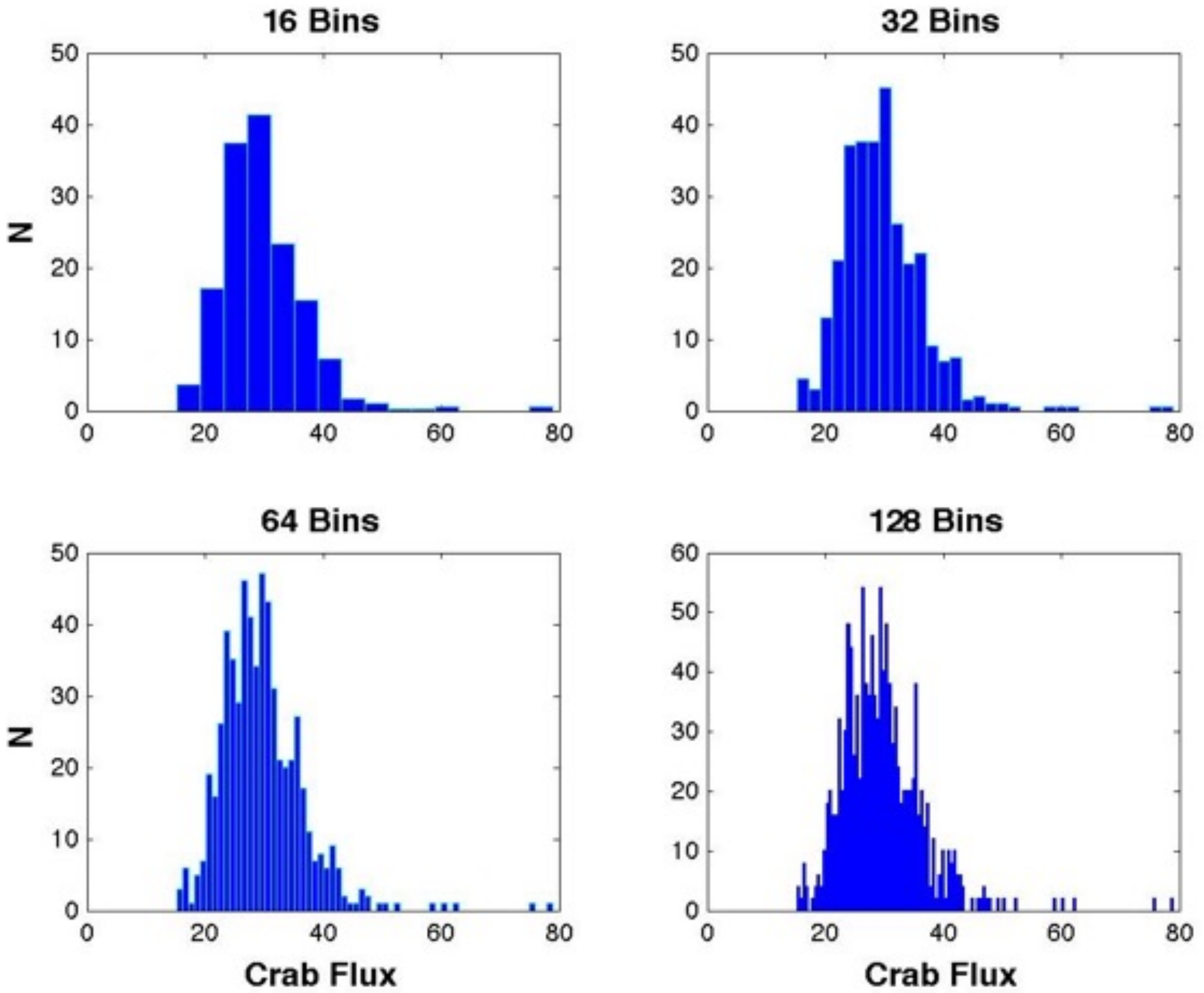
“GC islands”: G or C —> 1, A or T —> 0; two data modes:

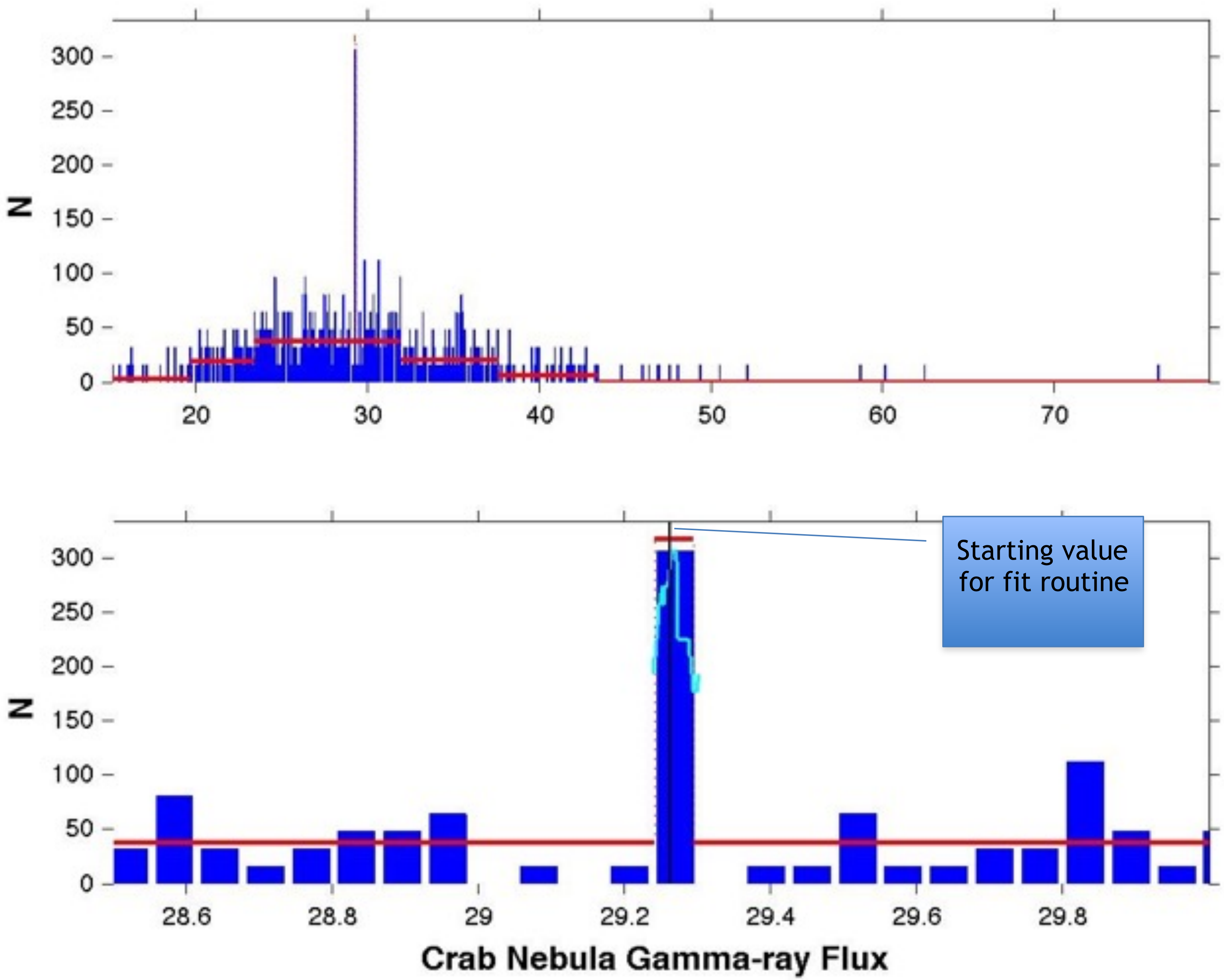
- 0-1 on fixed grid
- intervals between 1’s


GC fraction
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10^1
10^2
10^3
10^4
10^5
Block length (bp)

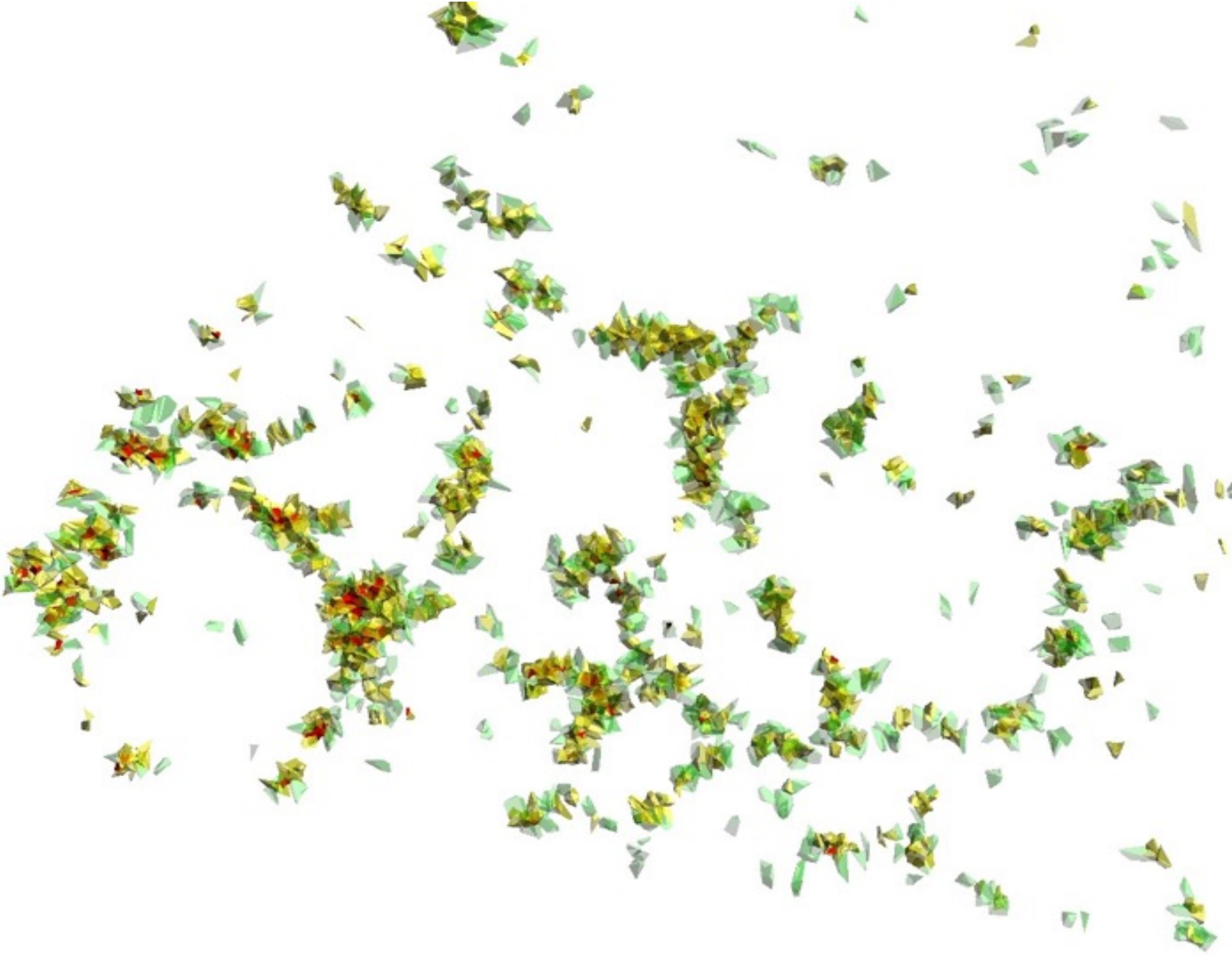

Series #3
Series #2
Series #1
Concatenated Times ...
Ordered Times

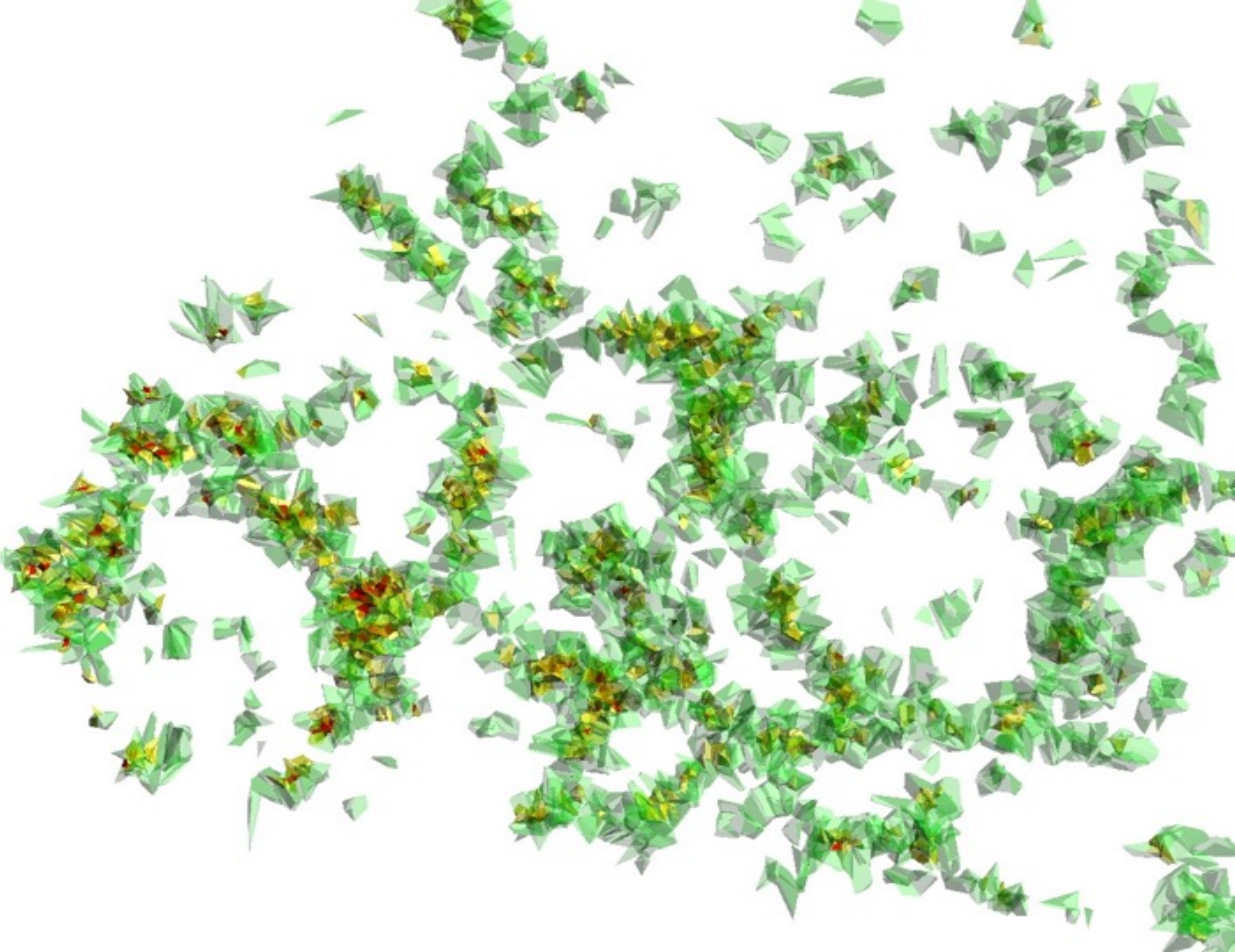
Bayesian Block analysis of event time series is mathematically the same problems as constructing (optimal) histograms!

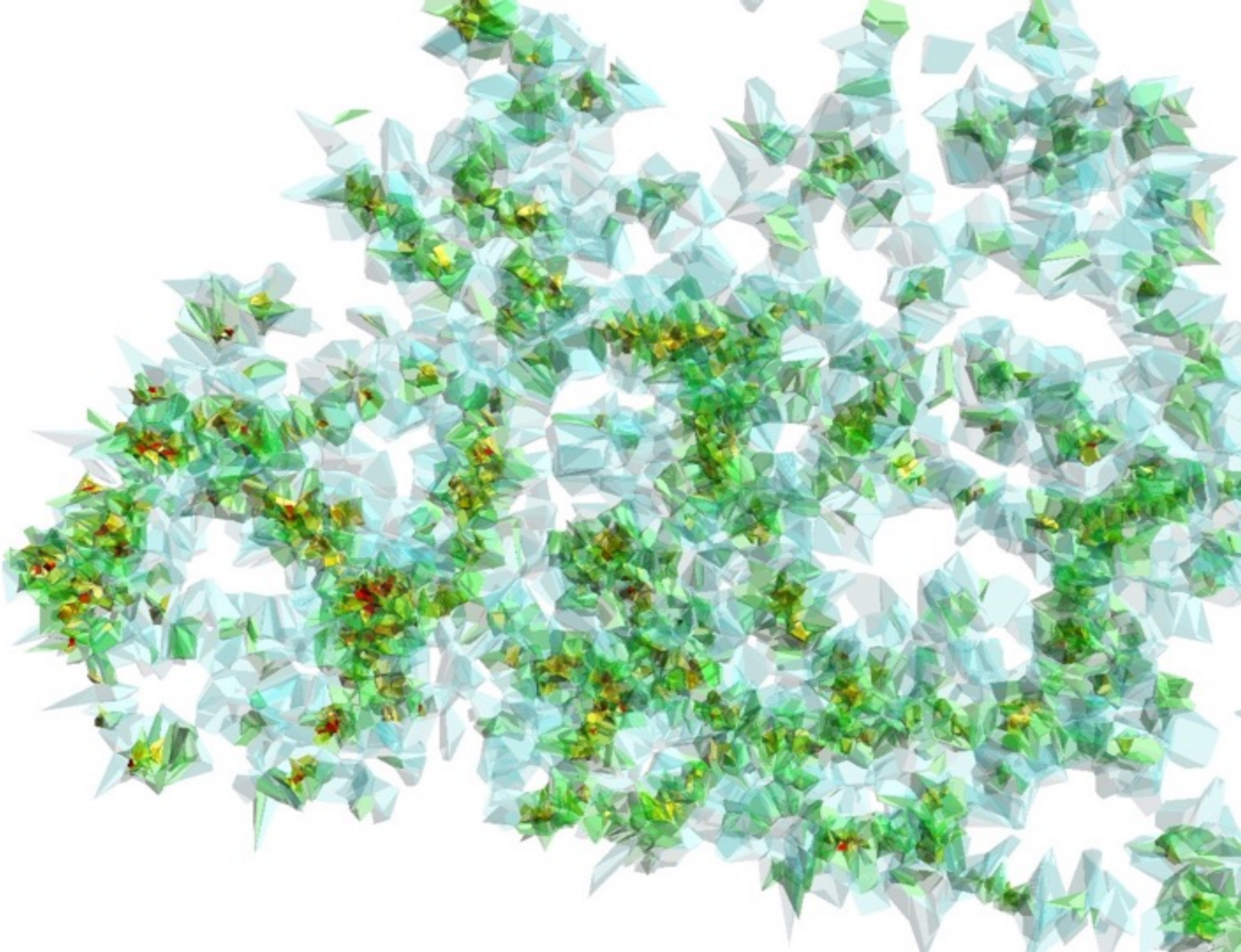




Starting value for fit routine







Maximum (log) Likelihood Block Costs

Event Data: $C_n = N_n \log(N_n / T_n)$

N_n = number of events in block n

T_n = length of block n

Point Measurements:

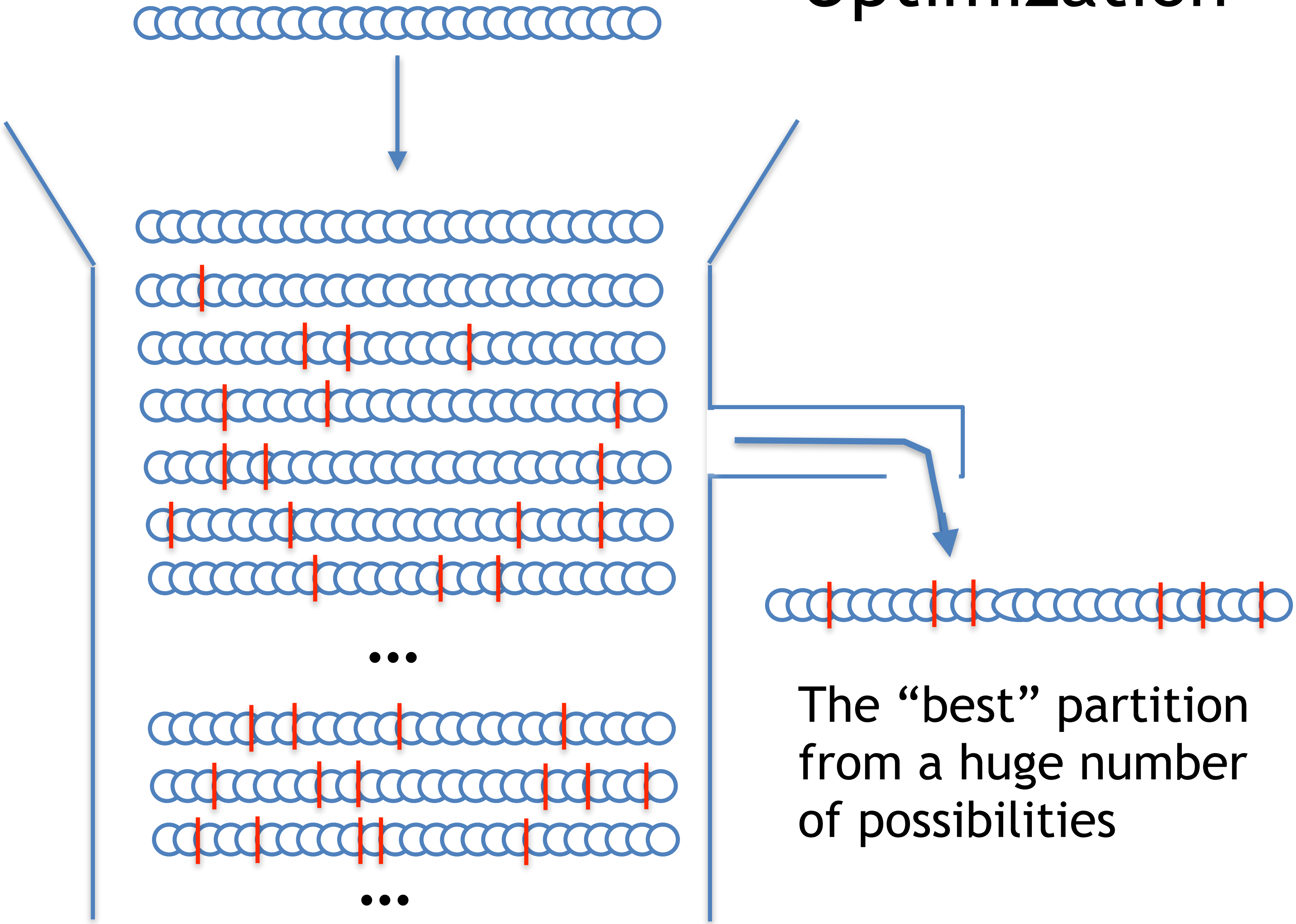
$C_n = (\sum w_n x_n)^2 / (\sum w_n)$ variance!

x_n = measured value

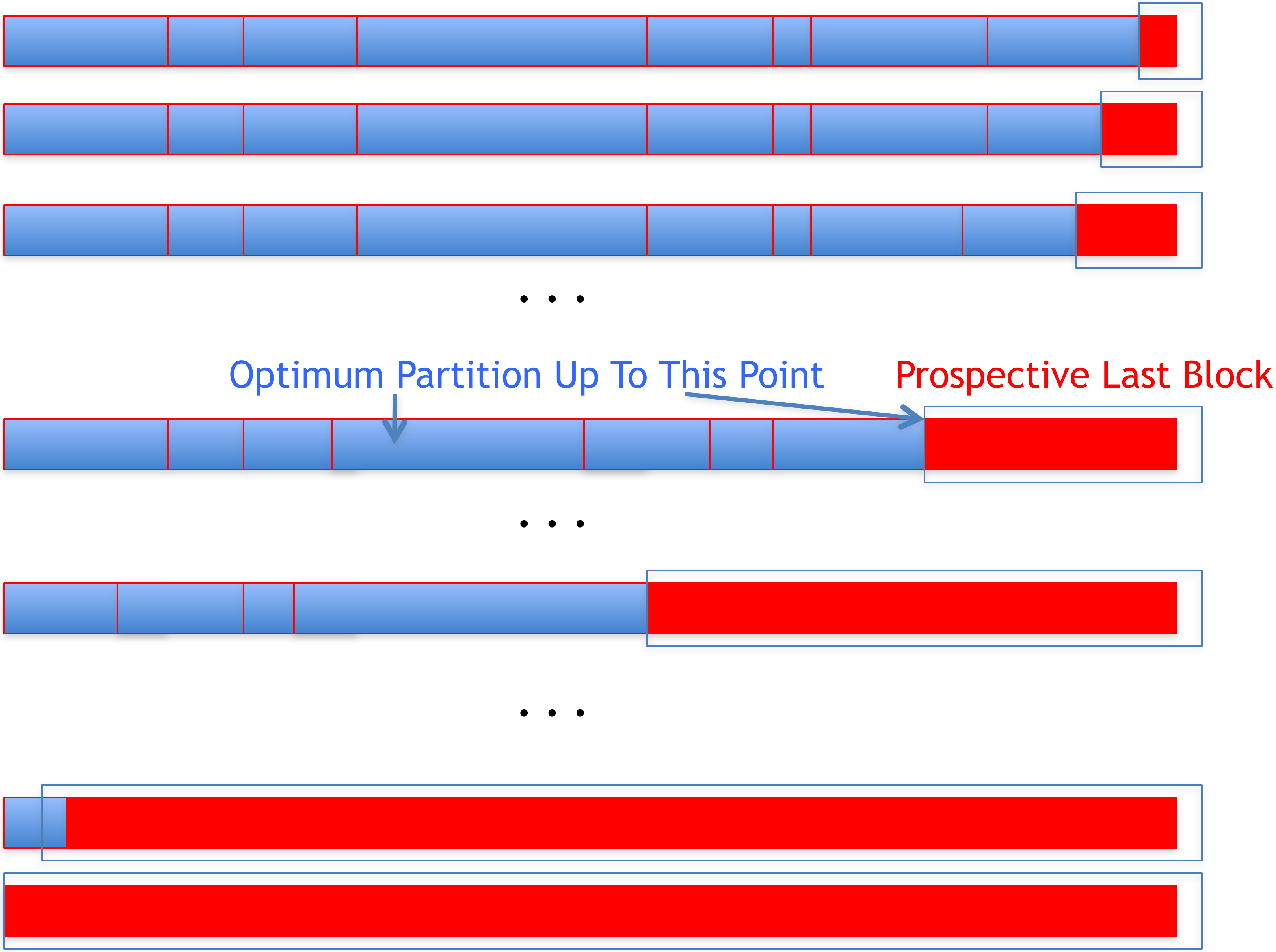
w_n = weight = 1 / σ_n^2

Total model cost: $\sum C_n$

Optimization



The "best" partition from a huge number of possibilities



Optimum Partition Up To This Point
Prospective Last Block
. . .
. . .
. . .

Dynamic Programming Algorithm

```
for ii = 1:num_points



   sum_num             = sum_num             + num_vec( ii );%count
   sum_len             = sum_len             + len_vec( ii );%length

   % Cost of last block:
   cost_last             = cost_function(sum_num            , ...
                                         sum_len            );

   % Compute and maximize total model cost:
   cost_total             = opt             + cost_last            ;
   [ opt(ii+1) last ] = max( cost_total             );

   last_change( ii ) =           last  ; % Store last change-point





end
```

Pruned Dynamic Programming Algorithm

```
for ii = 1:num_points

   unpruned( ii ) = ii; % New point unpruned until found otherwise

   sum_num( unpruned ) = sum_num( unpruned ) + num_vec( ii );%count
   sum_len( unpruned ) = sum_len( unpruned ) + len_vec( ii );%length

   % Cost of last block:
   cost_last( unpruned ) = cost_function(sum_num( unpruned ), ...
                                          sum_len( unpruned ));

   % Compute and maximize total model cost:
   cost_total( unpruned ) = opt( unpruned ) + cost_last( unpruned );
   [ opt(ii+1) last ] = max( cost_total( unpruned ) );

   last_change( ii ) = unpruned( last ); % Store last change-point

   % update pruning
   id_unpruned = find([ cost_total(unpruned) > opt(ii+1) - ncp_prior ]);
   unpruned = unpruned( id_unpruned );

end
```

What good are Segmented Time Series Representations? (2)

Detect and characterize,
without bins or smoothing:

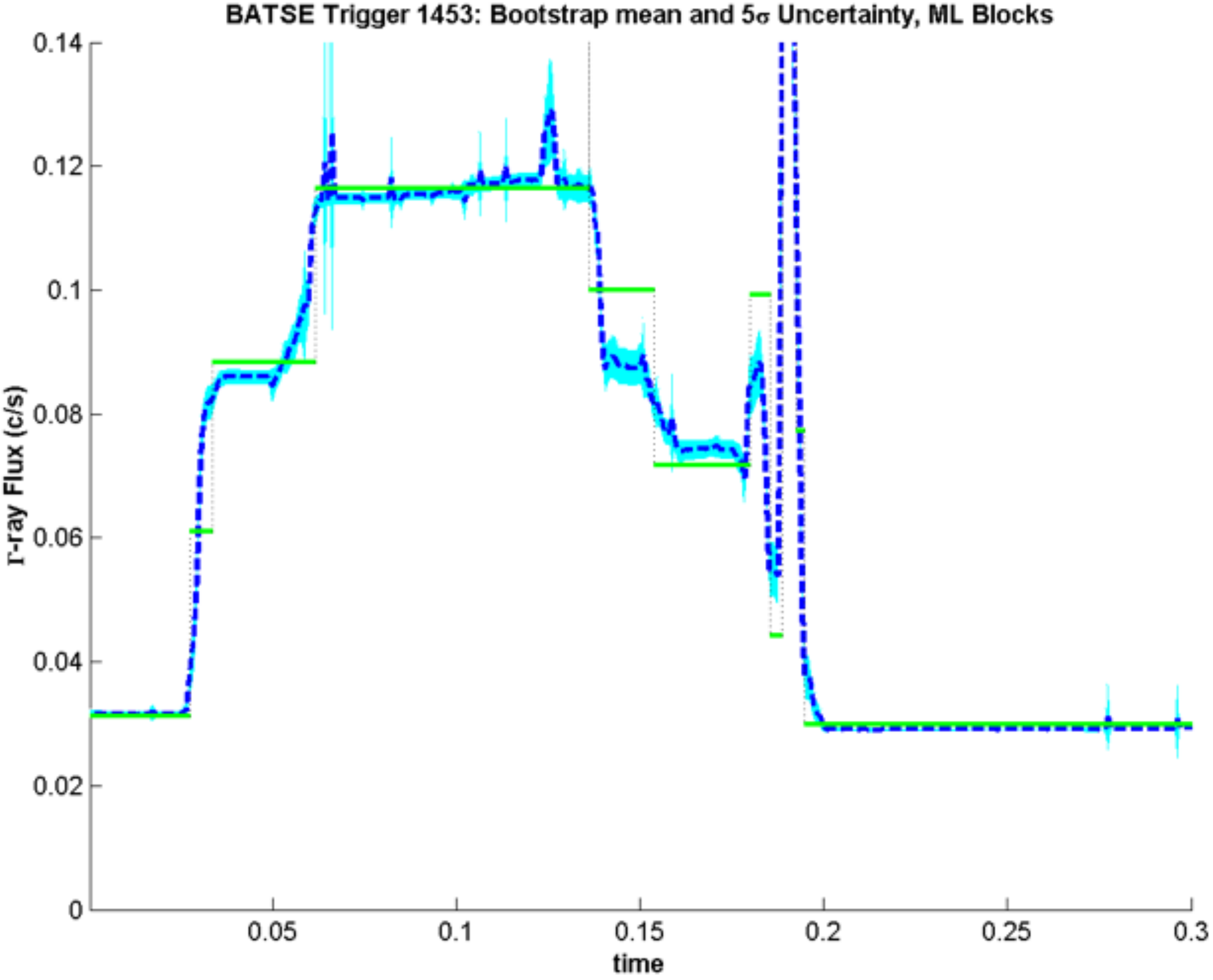
- *Pulses (aka “flares”)*
- *Pulse shapes (including the Arrow of Time)*
- *Variability index*
- *Variability time scales (min, max, dist, …)*
- *Transient event triggers (real-time mode)*

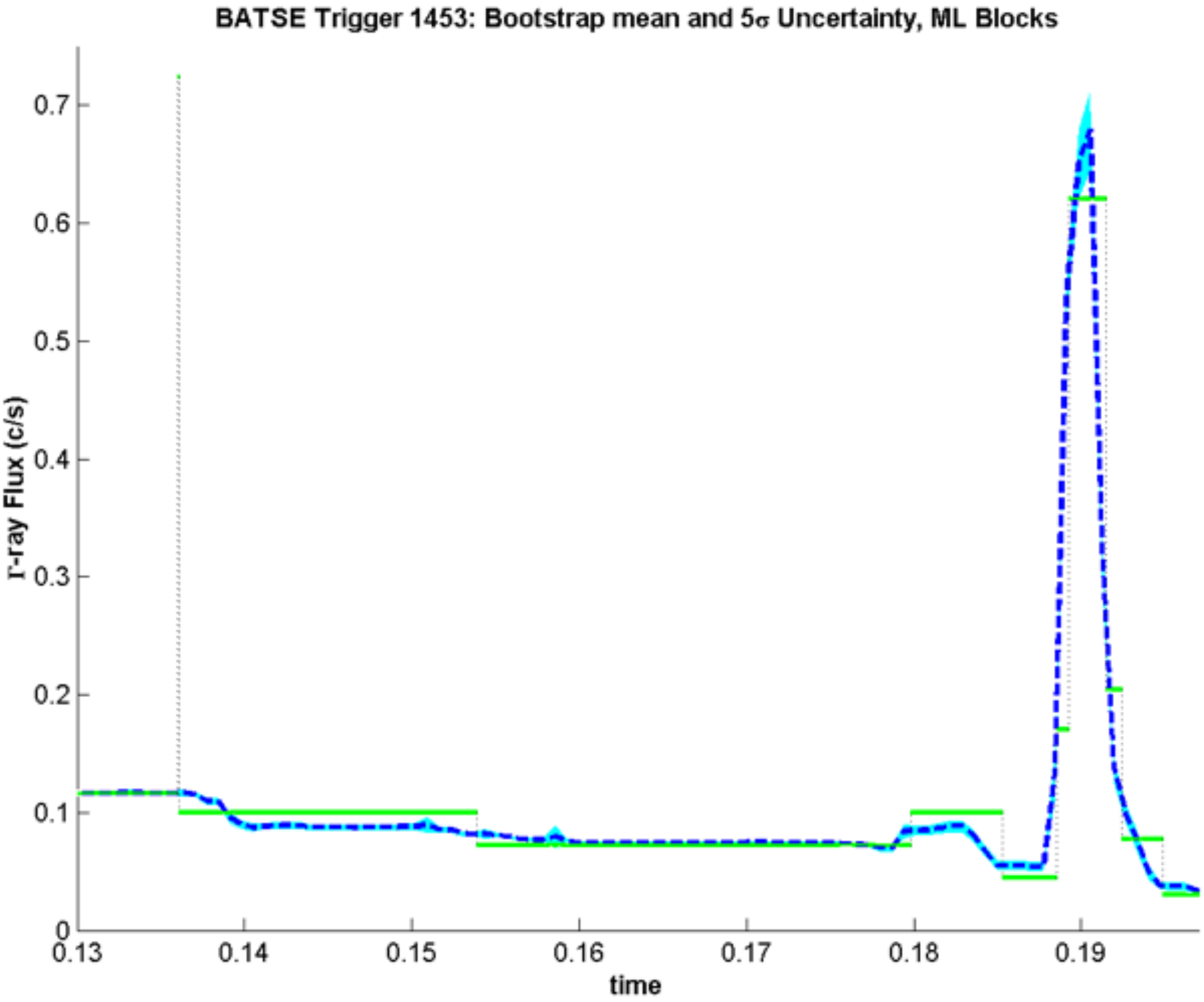
What good are Segmented Time Series Representations? (3)

Implement:

- *Exploratory Data Analysis*
- *Time series classification*
- *Noise suppression*
- *Visual displays*
- *Data compression*
- *Data adaptive histograms*

Thanks


BATSE Trigger 1453: Bootstrap mean and 5σ Uncertainty, ML Blocks
Γ-ray Flux (c/s)
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
time


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