Data Intensive Science

A prologue

Dipankar Bhattacharya
IUCAA, Pune

Workshop on Data Intensive Science, IUCAA, 13 Feb 2017

DATA INTENSIVE SCIENCE

- Data-intensive is a qualitative description, the benchmark continuously shifts as our data gathering and processing ability improves
- Natural sciences have always been data intensive. Major discoveries are made at the edge of the contemporary abilities

BIG DATA SCIENCE

- Normally involves compiling data collected by multiple sources and/or over a length of time
- Requires ingestion, analysis, reduction, visualisation, modelling and interpretation
- The larger the dataset, the more involved the techniques need to be

USE OF LARGE DATASETS

- Reveal signal hidden in noise
- Determine physical parameters of study subjects - model fitting
- Study populations - find outliers
- Measure correlations - find patterns

Extensive use of Statistics is necessary - both for drawing inference and for estimation of significance

ASTRONOMY AS LARGE DATA SCIENCE

Large datasets are in common use in astronomy

- Objects are distant, faint and numerous
- Real-life multivariate systems
- Study evolution through populations
- Cause and effect through correlations

Methods and techniques applicable to other branches of data intensive science

Data intensive Astronomy

17th century

Johannes Kepler

wikipedia

Tycho Brahe

wikipedia

Kepler's laws of planetary motion (1609)

- Required long study of planetary orbits
- Used precision measurements conducted over 24 years (1580-1604) by Tycho Brahe

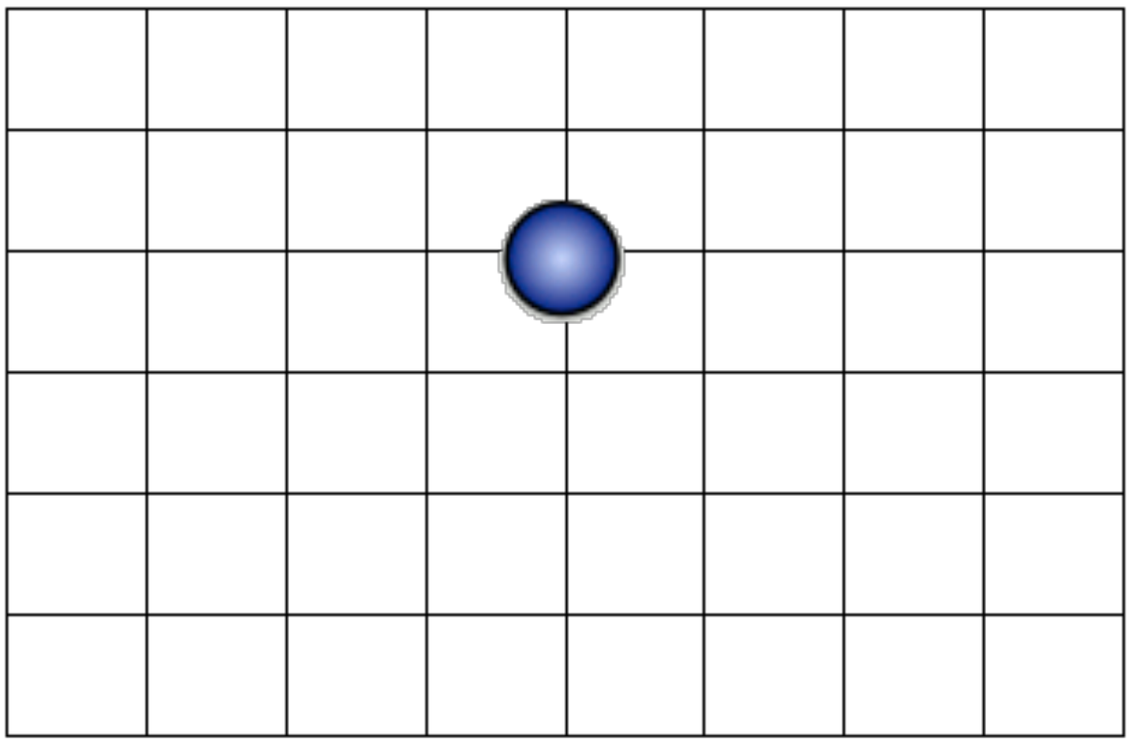
Present day Astronomical Data are Digital

Data holdings can be very large due to multi-dimensional nature

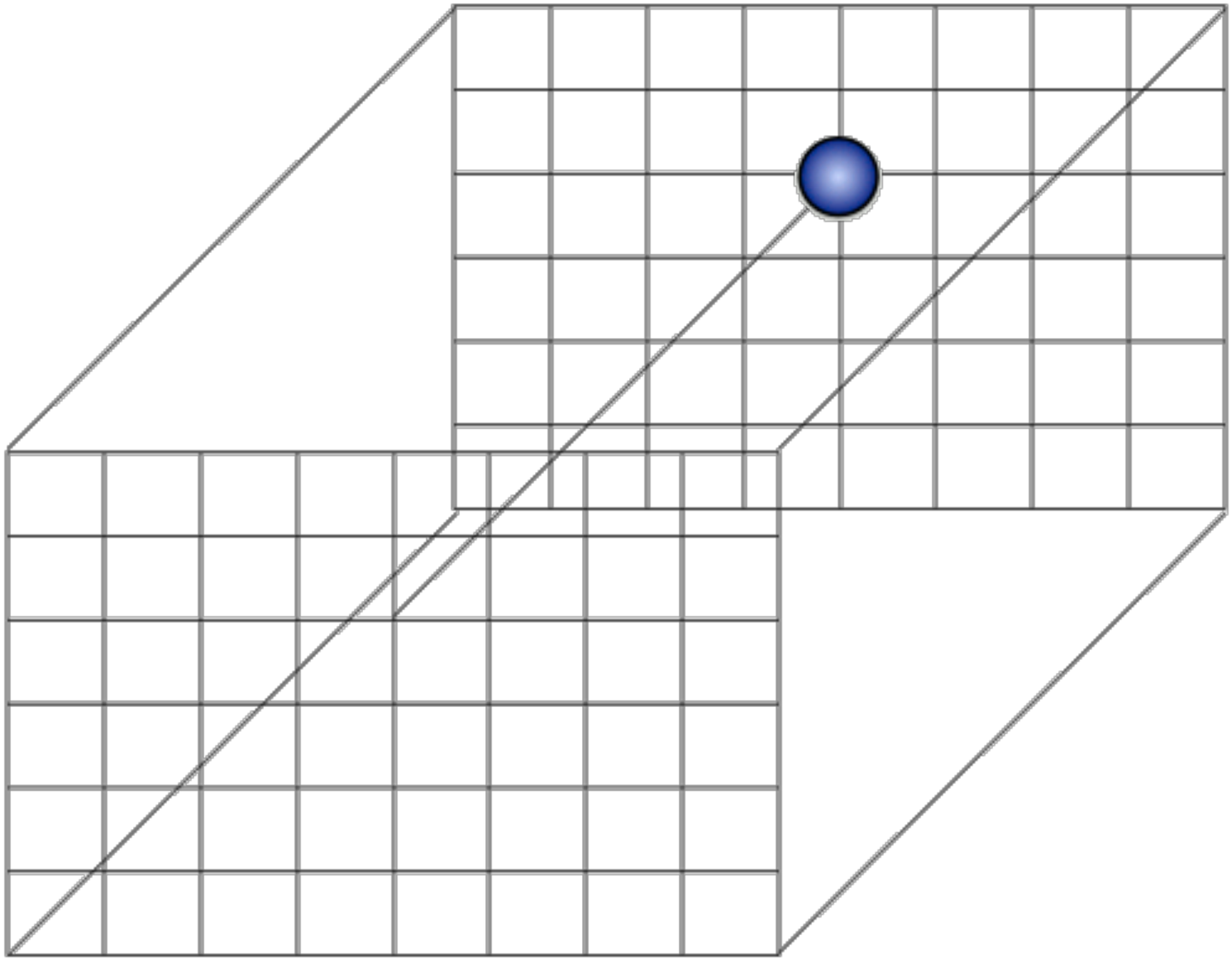
Poised to grow by many orders of magnitude

Multidimensionality of Astronomical Data

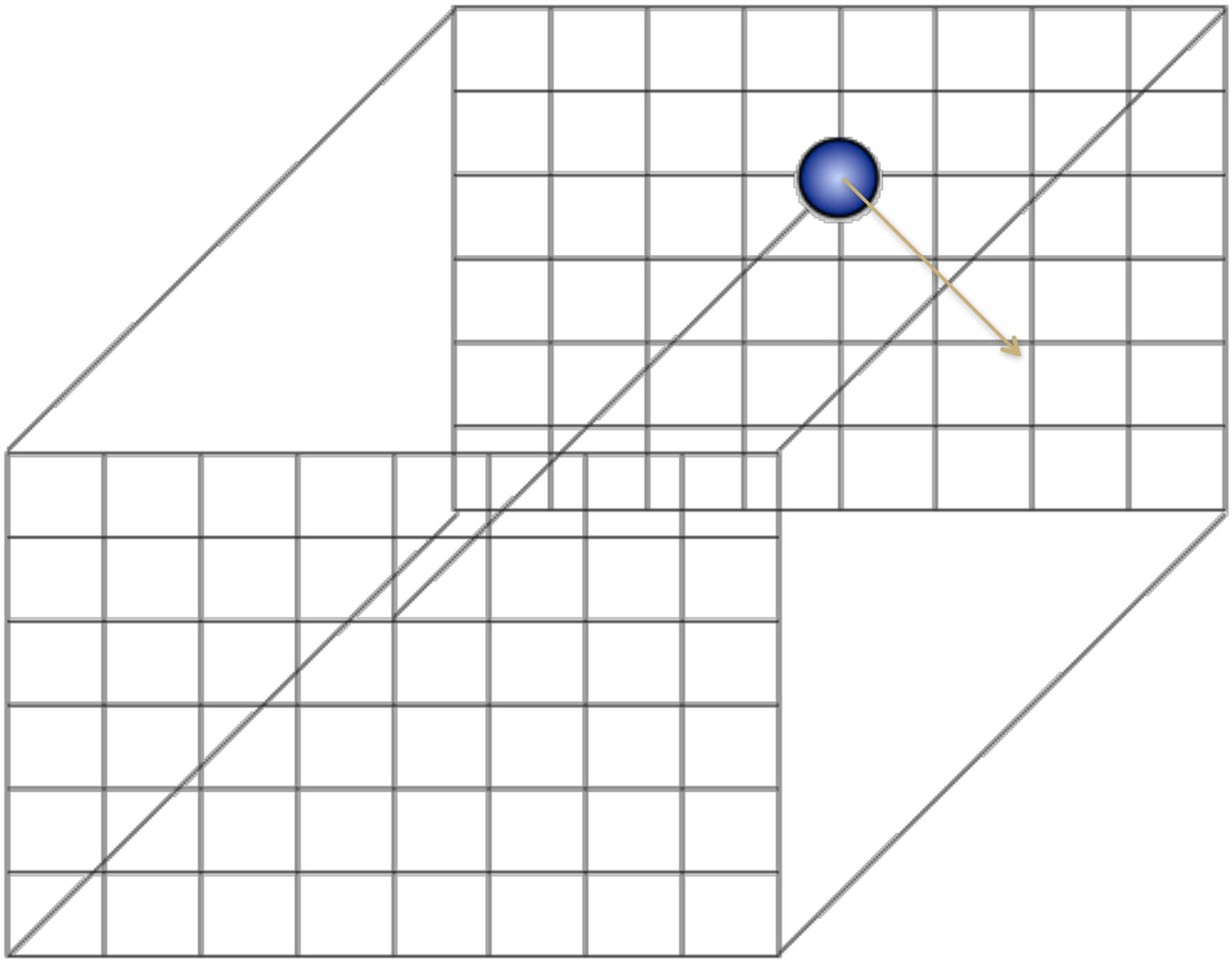
Position



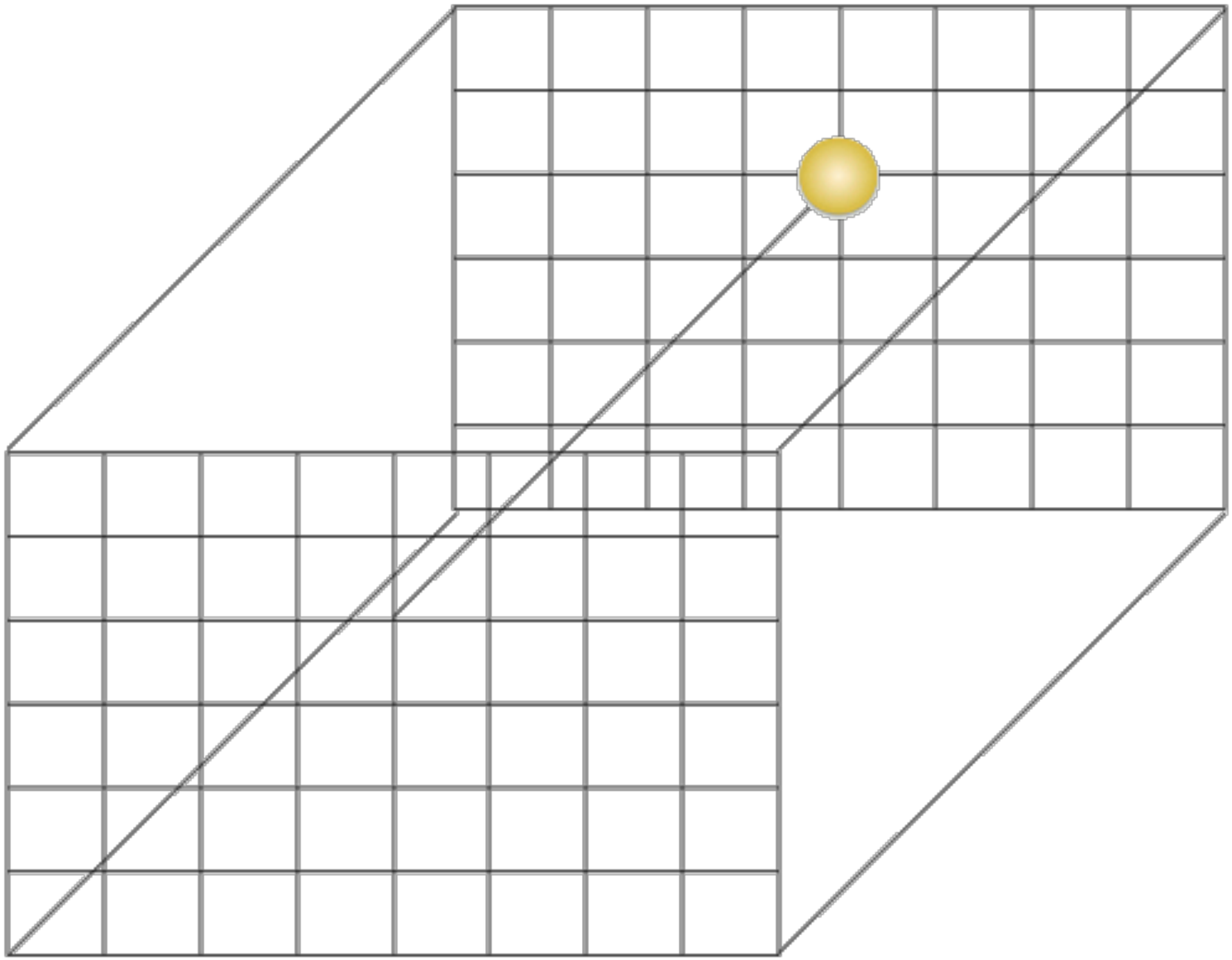
Distance



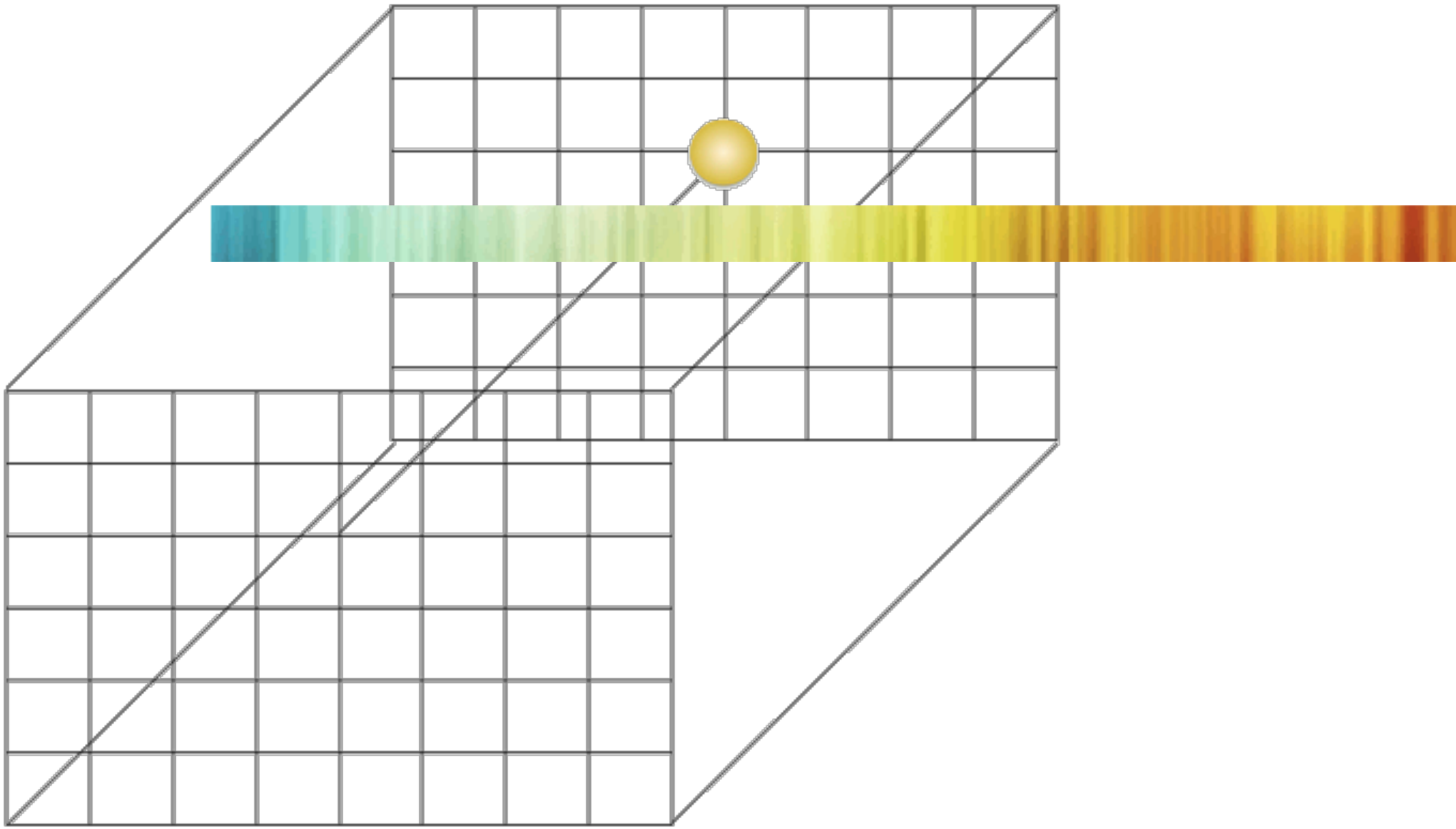
Velocity



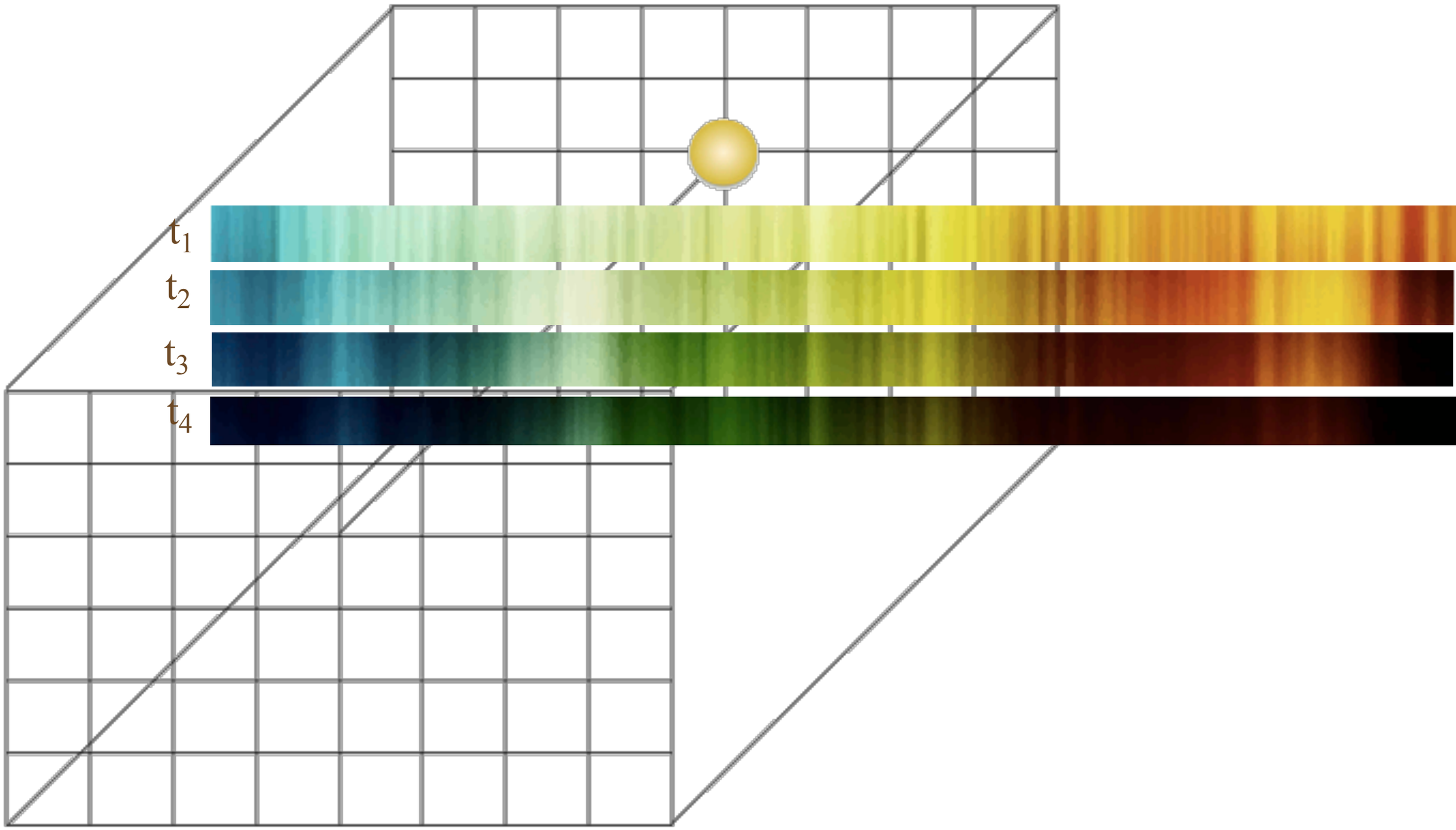
Brightness



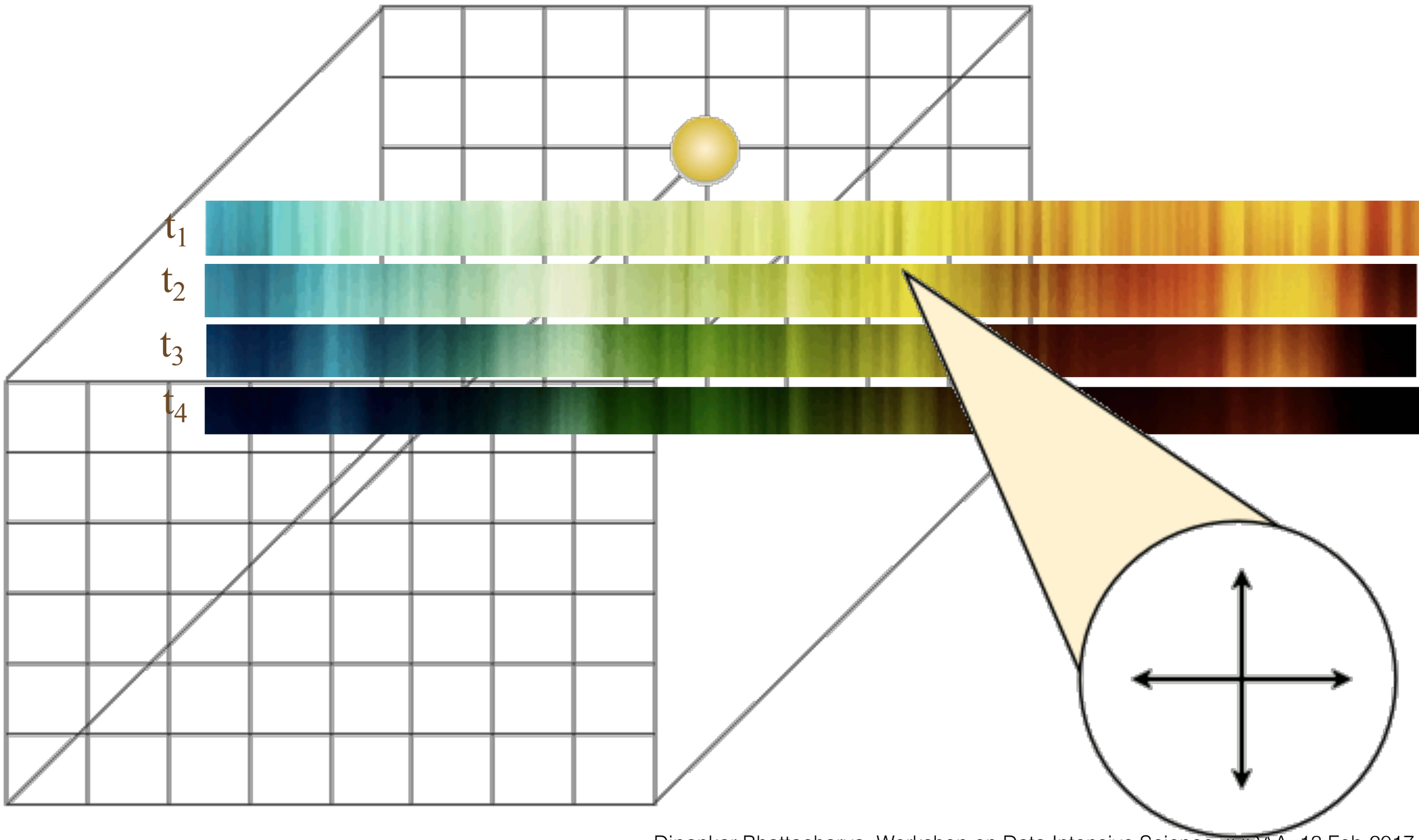
Spectrum: many wavelengths



Time Domain



Polarization



Multidimensionality of Astronomical Data

- Spatial location : 3D, angular size : 2D, velocity : 3D
- Brightness
- Spectroscopic : many wavelengths
- Time domain : many time samples at each wavelength
- Polarization : 4 stokes parameters at each time
- Morphology
- Classification

.....for every object.

In full detail, data volume can easily exceed 1GB/object

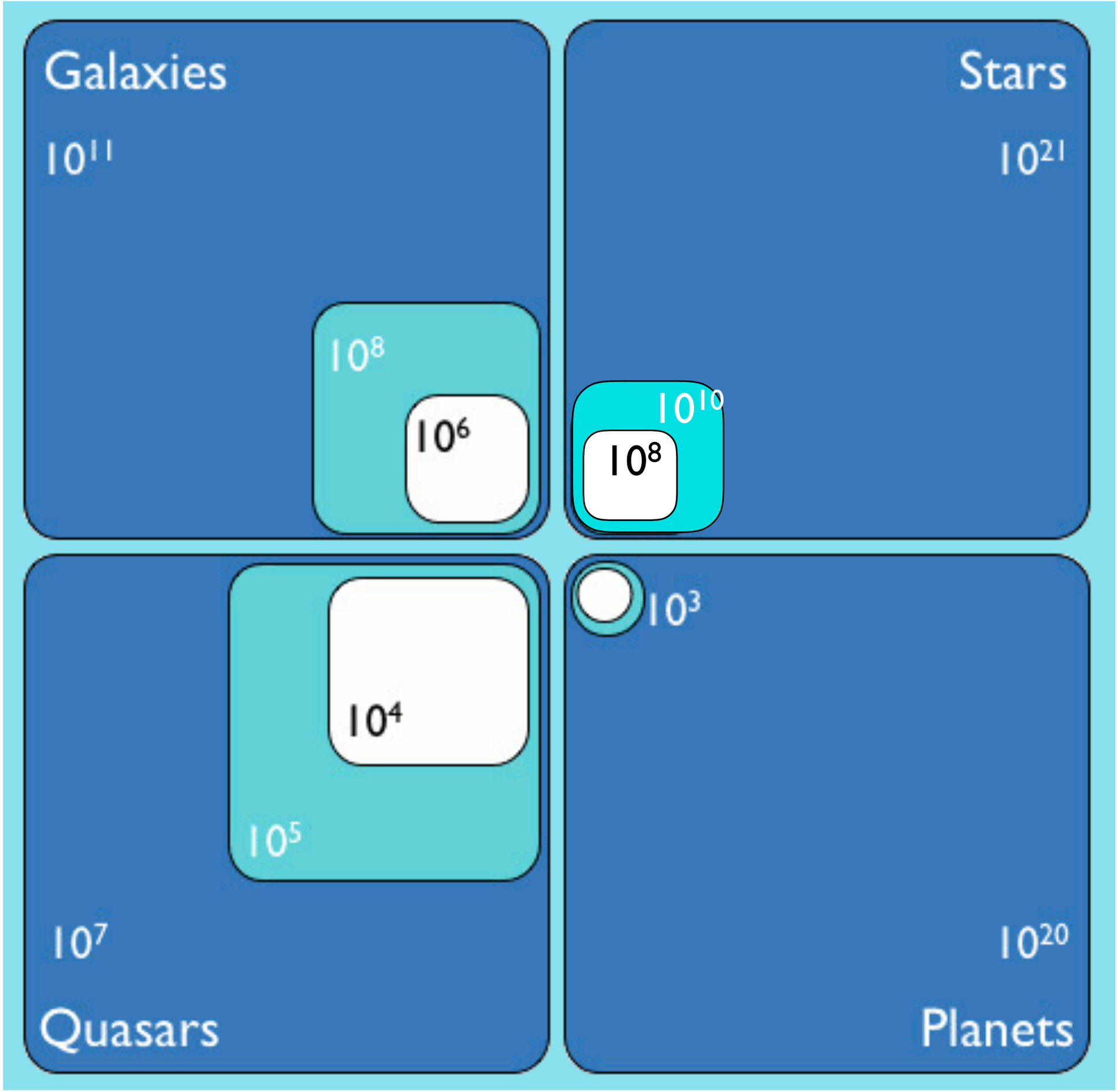
Total number of catalogued objects ~ 10^{10}

Current Data Holding in Astronomy

- Over 10 PetaBytes in Public Archives
- About 100x in Raw Data
- Doubling every year

A very small fraction of the objects in the universe has been catalogued,

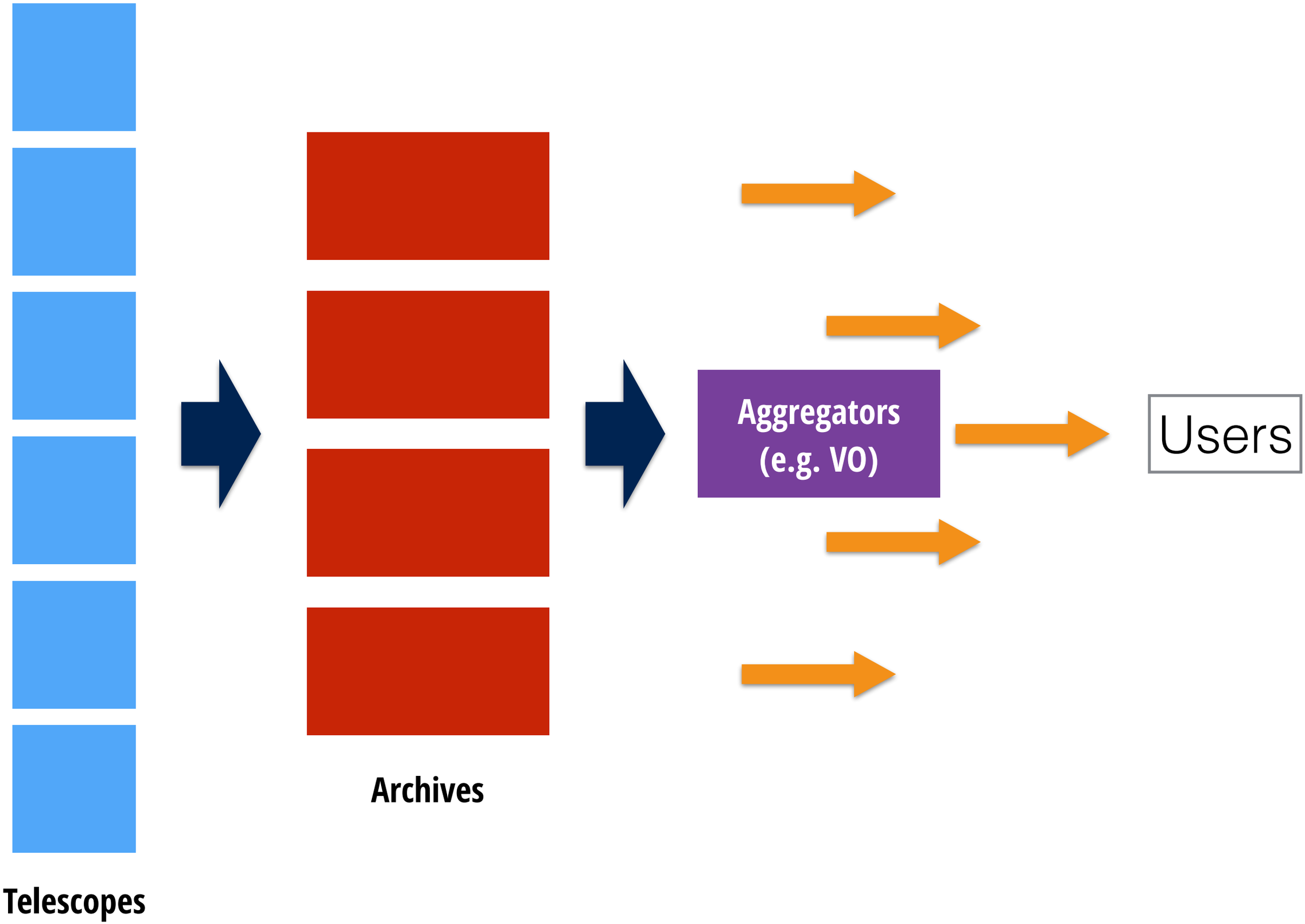
far fewer have been studied in depth



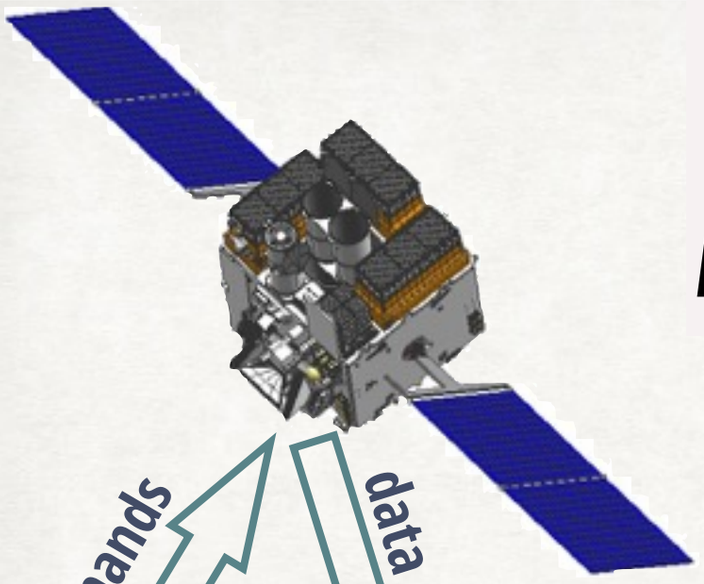
We are witnessing an explosive growth of Astronomical Data

- Telescopes continue to become more powerful, reaching deeper into the universe
- Multiple wavelengths – covering the entire electromagnetic band from radio to γ-rays
- Adding non-electromagnetic messengers, e.g. gravitational waves
- Entering a new era of time domain astronomy
- Huge quantity of data is entering public astronomical archives

Public Data Distribution in Astronomy



Example:
AstroSat data flow



commands

data

ISTRAC, PEENYA

ISSDC, BYLALU

raw data

science products

POCs

LAXPC: TIFR, RRI

UVIT: IIA

SXT: TIFR

CZTI: IUCAA

SSM: ISAC, IUCAA

DATA ARCHIVE

USER

PROPOSALS

ASTROSAT SCIENCE SUPPORT CELL

REVIEW, TIME ALLOCATION, SCHEDULING

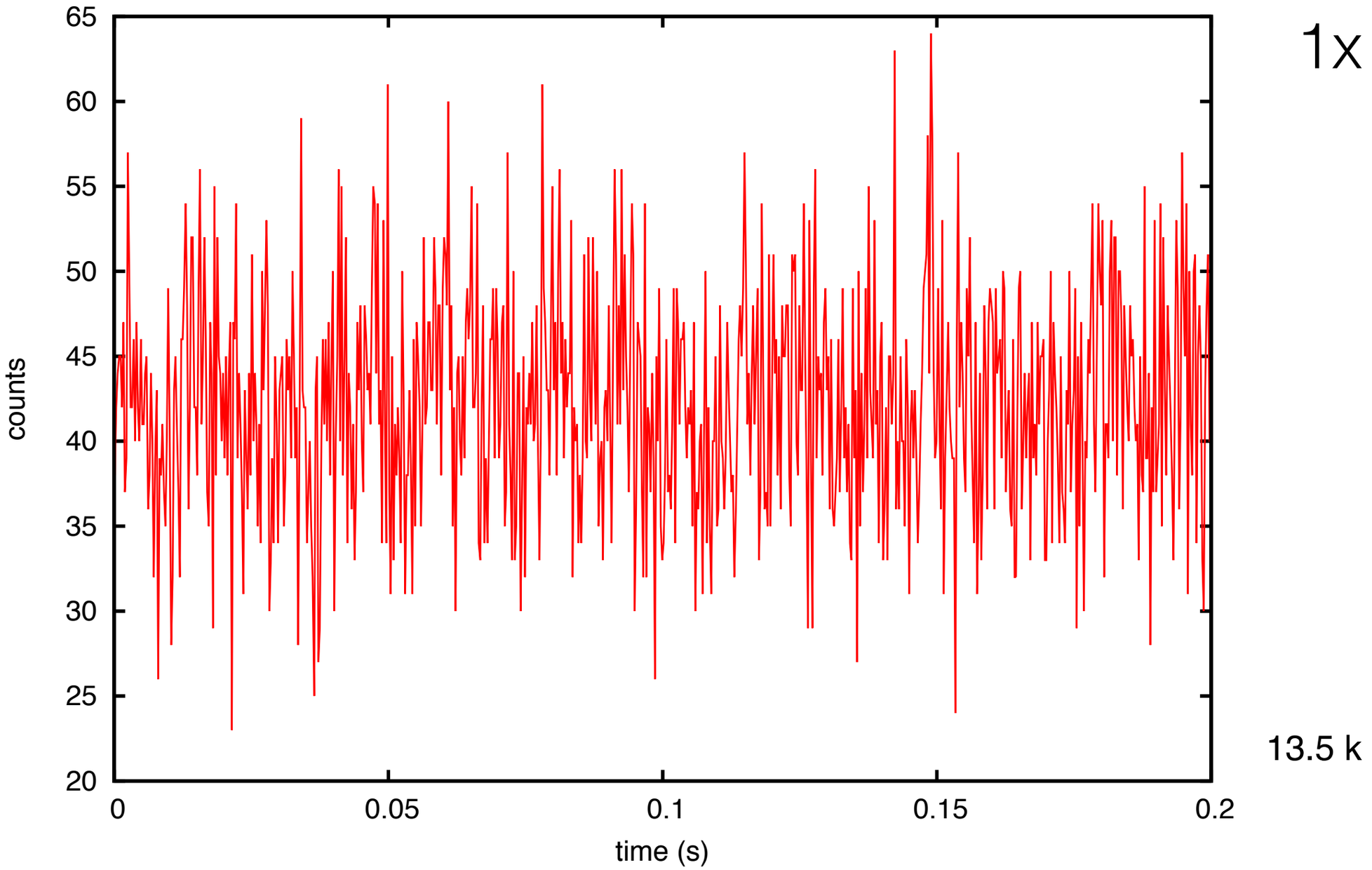
Data products being generated at a rate > 100 TB / year

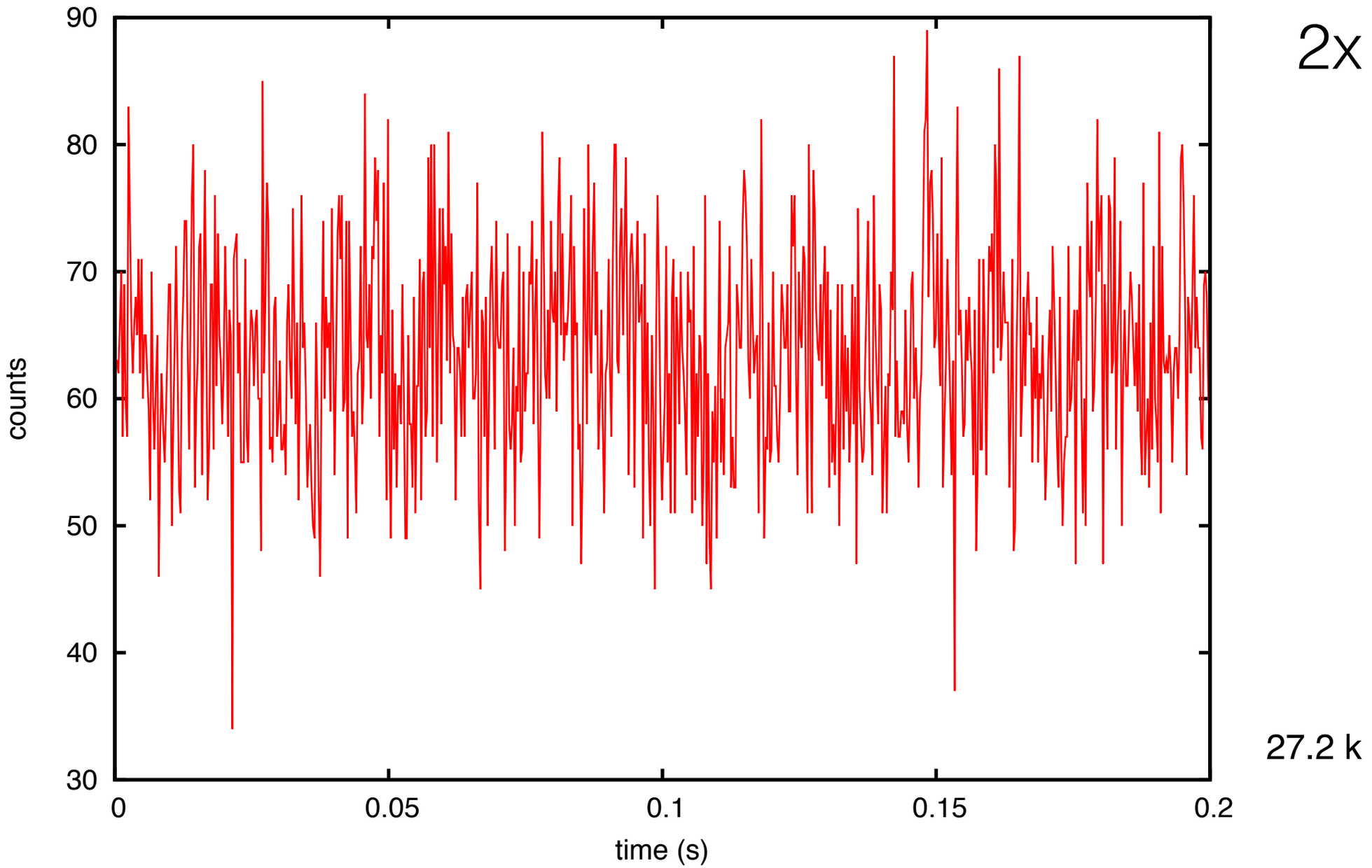
From Data to Inference

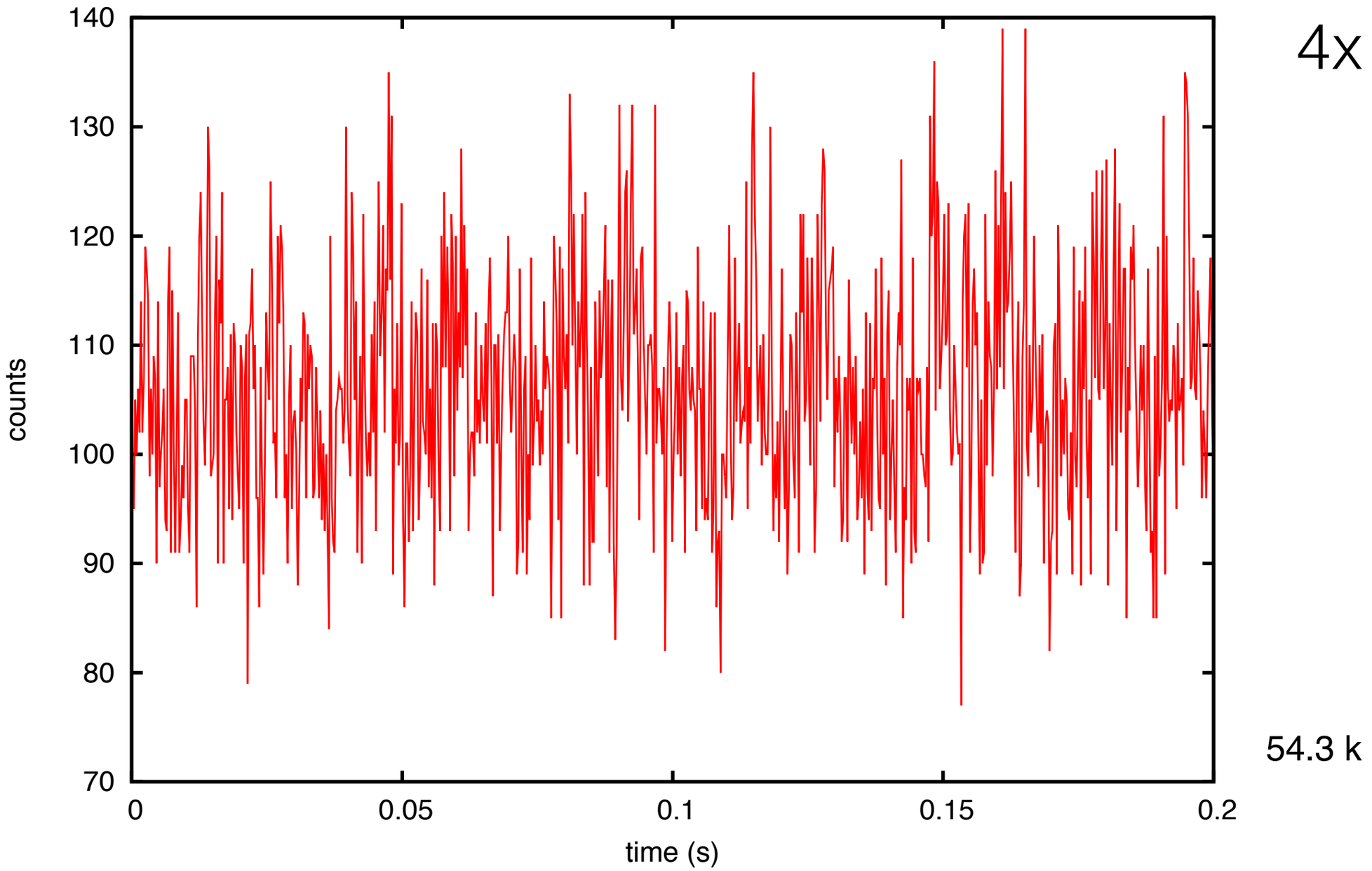
Finding faint signal in presence of noise

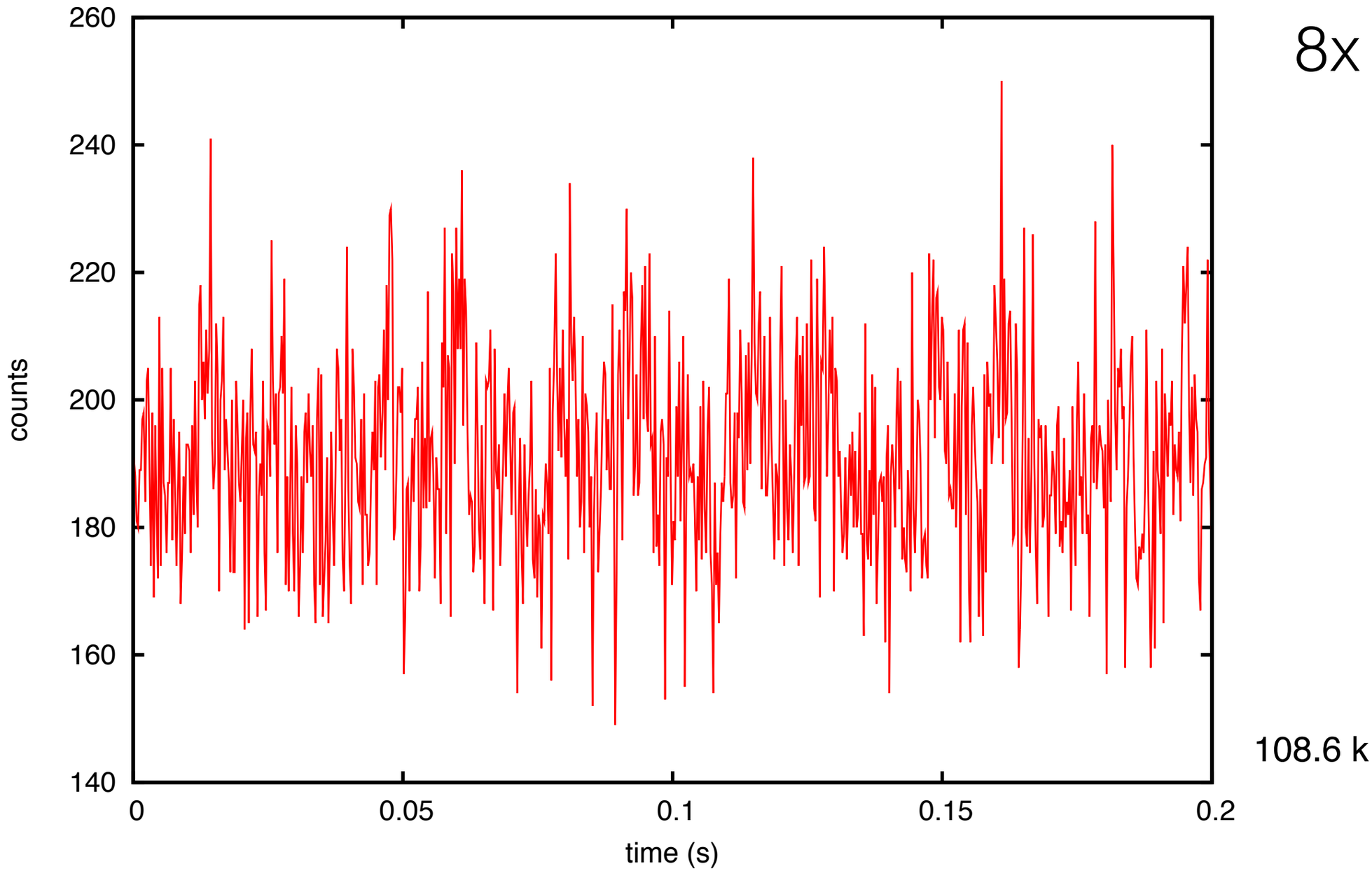
Accumulate large data

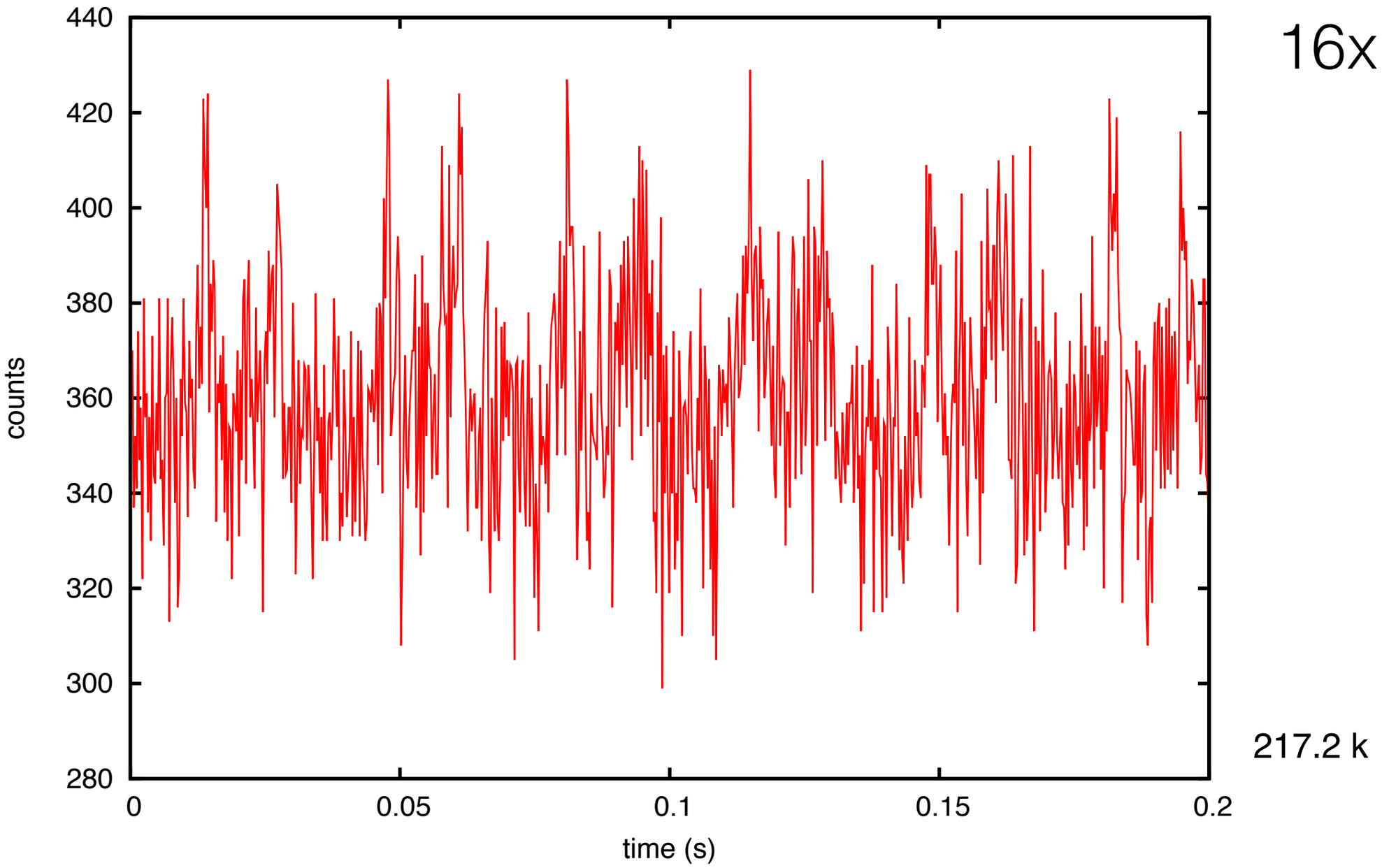
Average to improve signal-to-noise ratio

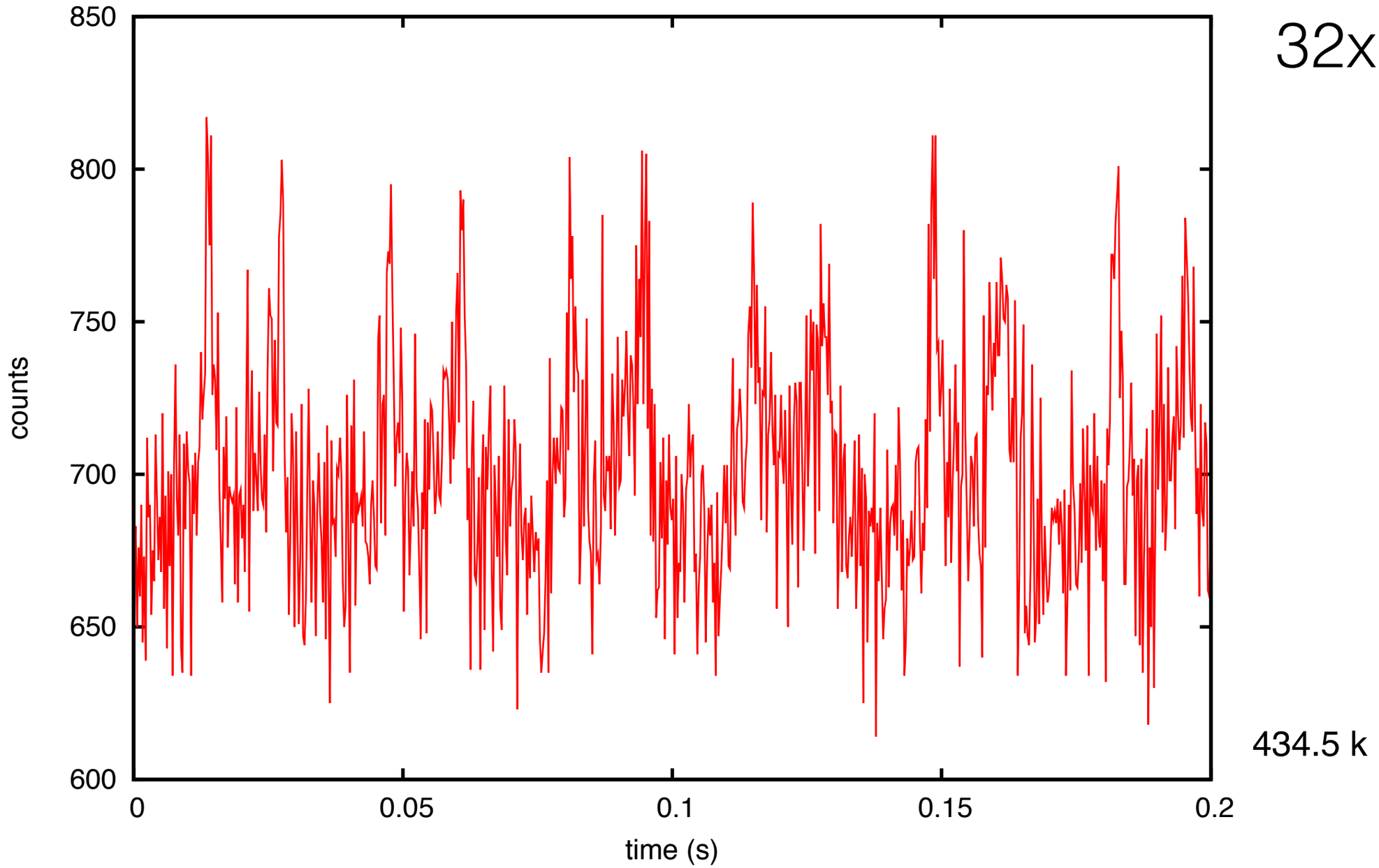

65
60
55
50
45
40
35
30
25
20
counts
0
0.05
0.1
0.15
0.2
time (s)
1x
13.5 k

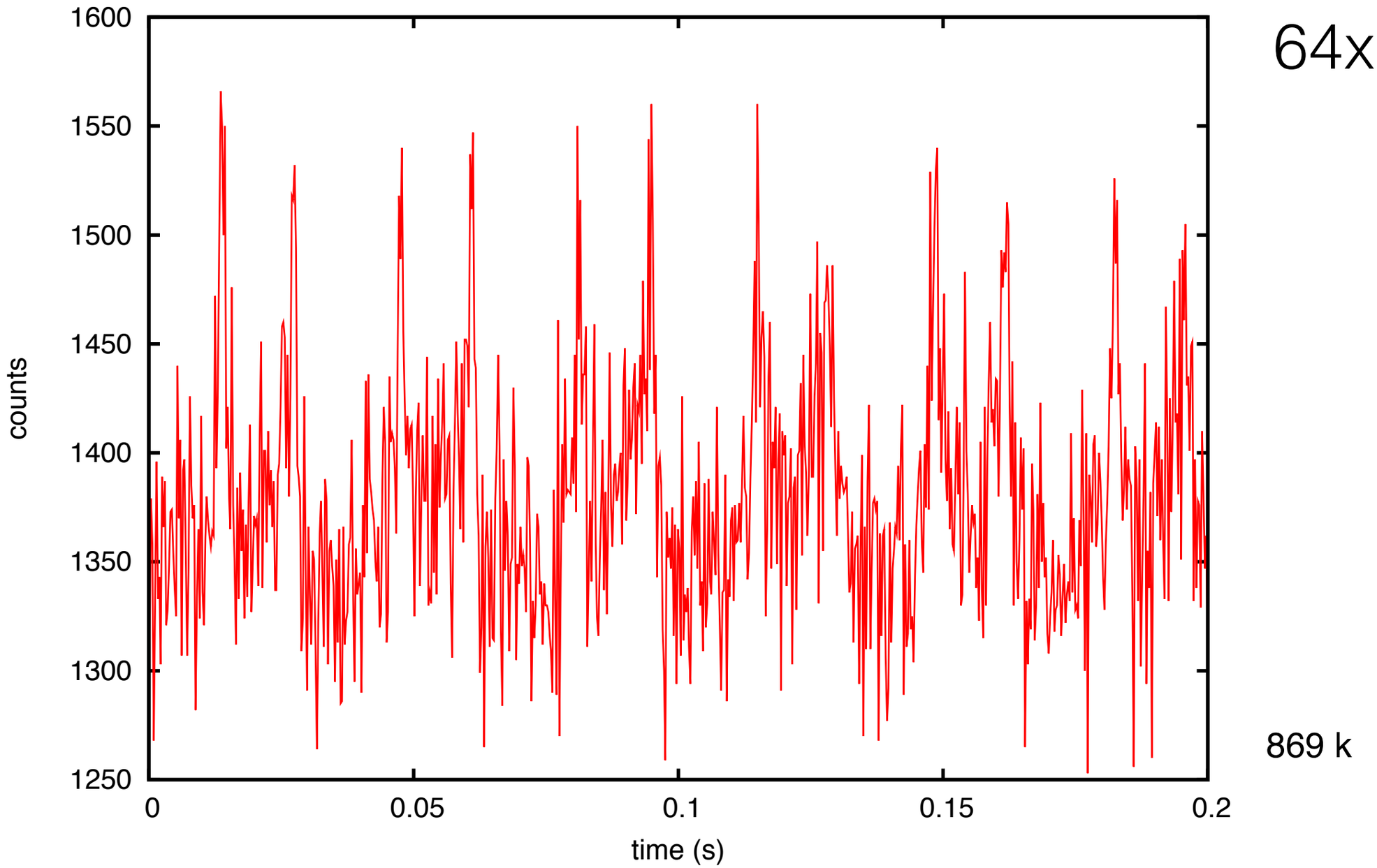

90
80
70
60
50
40
30
counts
0
0.05
0.1
0.15
0.2
time (s)
2x
27.2 k

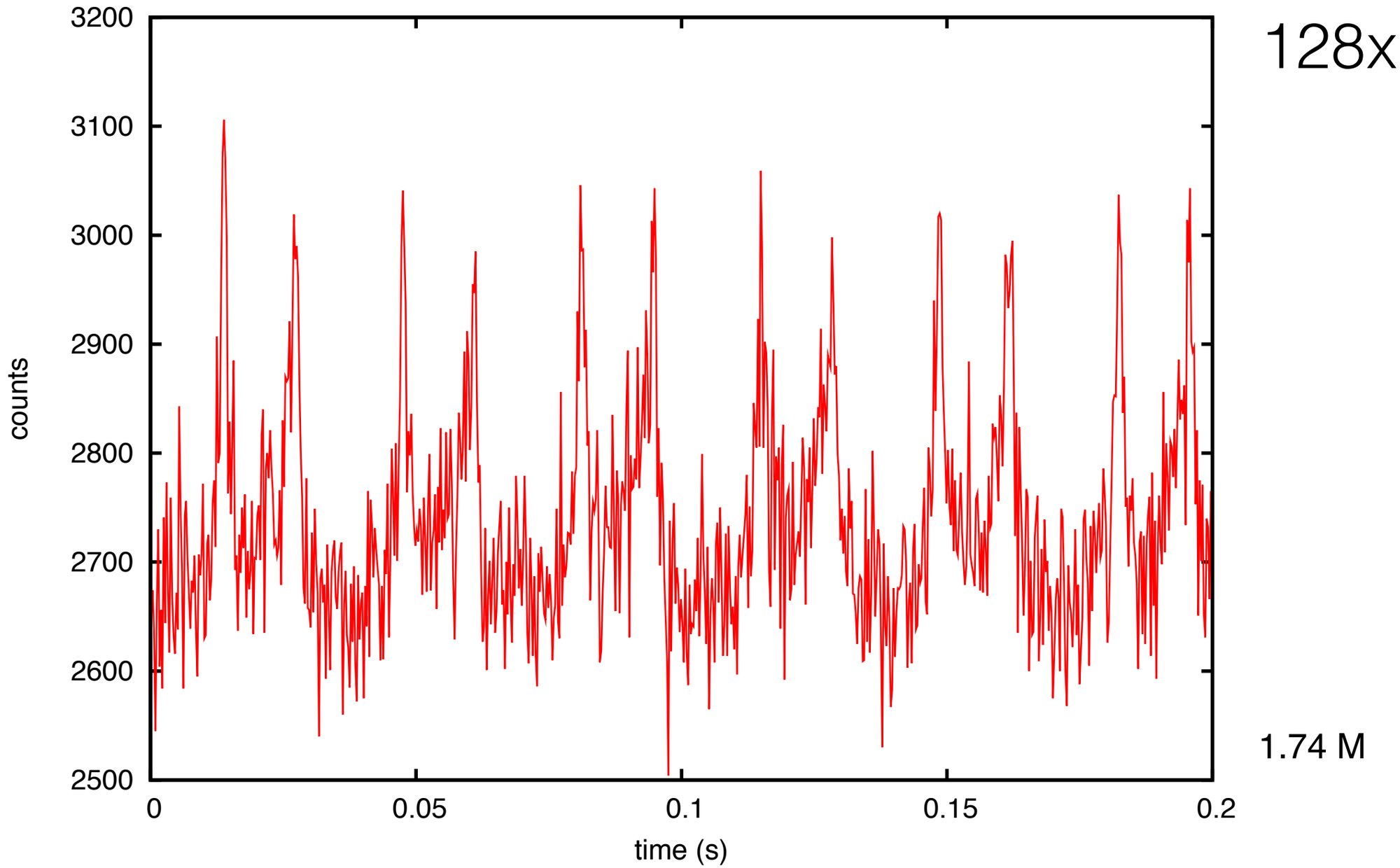

4x
140
130
120
110
100
90
80
70
counts
0
0.05
0.1
0.15
0.2
time (s)
54.3 k

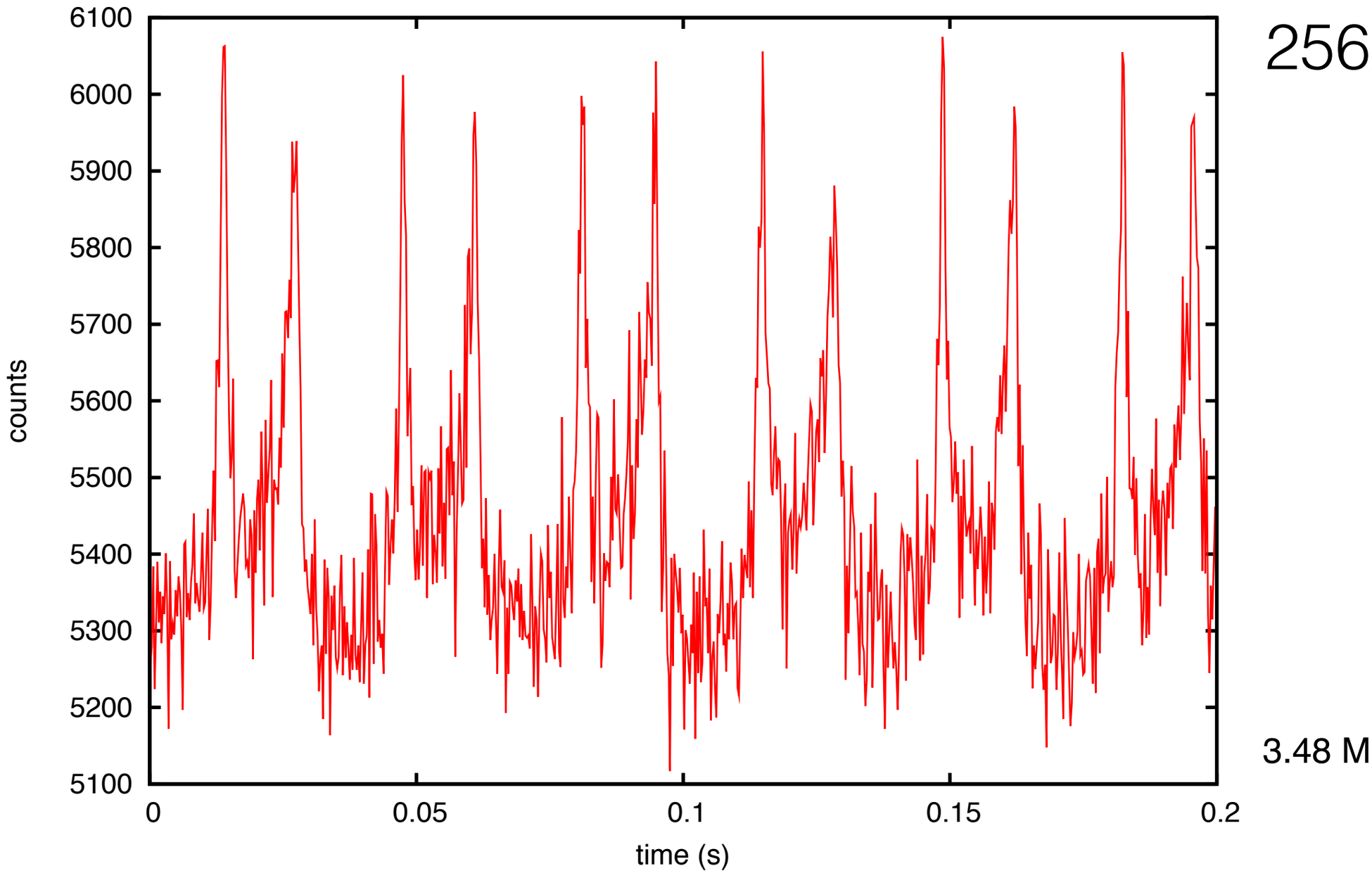

260
240
220
200
180
160
140
counts
0
0.05
0.1
0.15
0.2
time (s)
8x
108.6 k

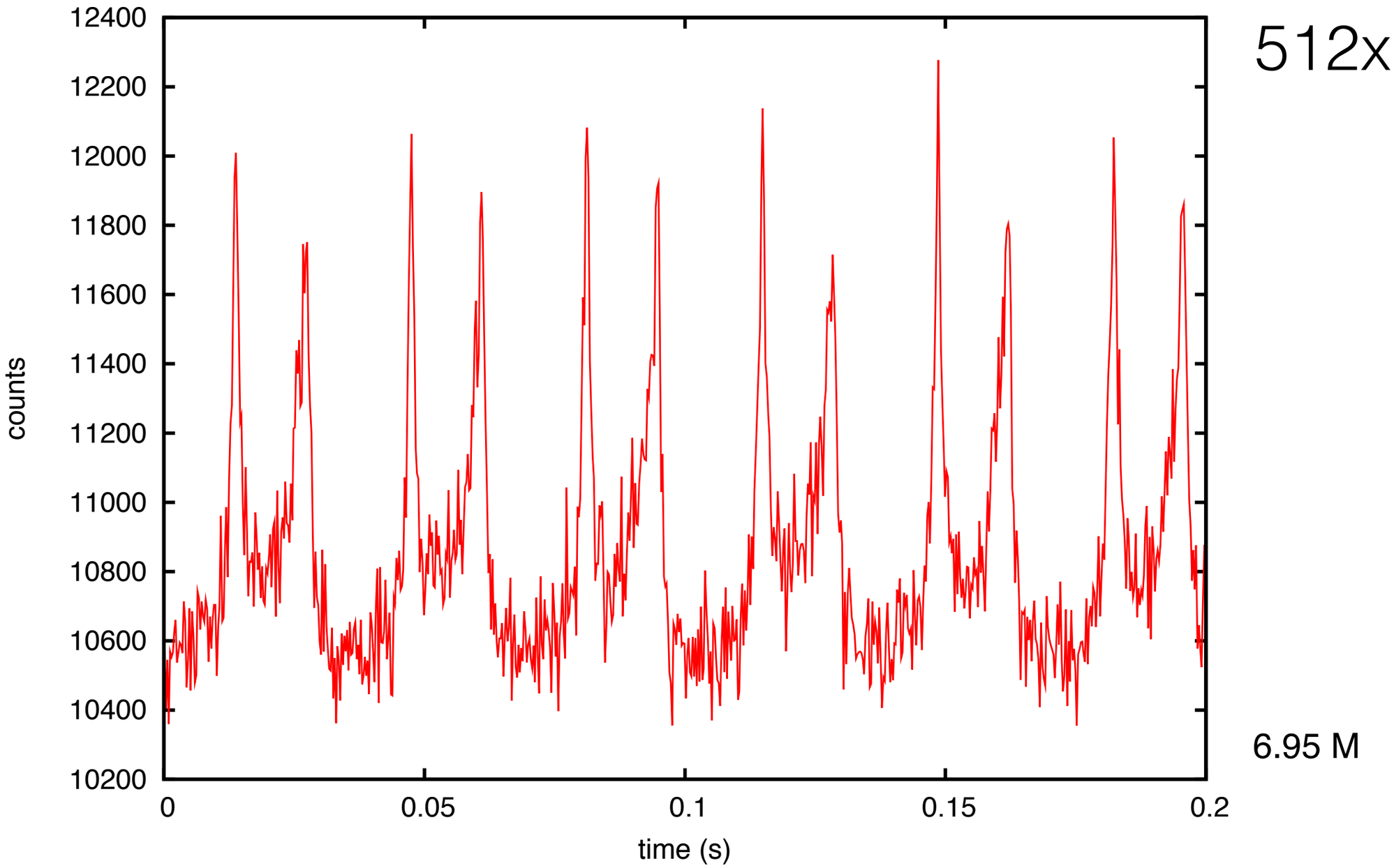

16x
440
420
400
380
360
340
320
300
280
counts
0
0.05
0.1
0.15
0.2
time (s)
217.2 k


32x
850
800
750
700
650
600
counts
0
0.05
0.1
0.15
0.2
time (s)
434.5 k


64x
1600
1550
1500
1450
1400
1350
1300
1250
counts
0
0.05
0.1
0.15
0.2
time (s)
869 k

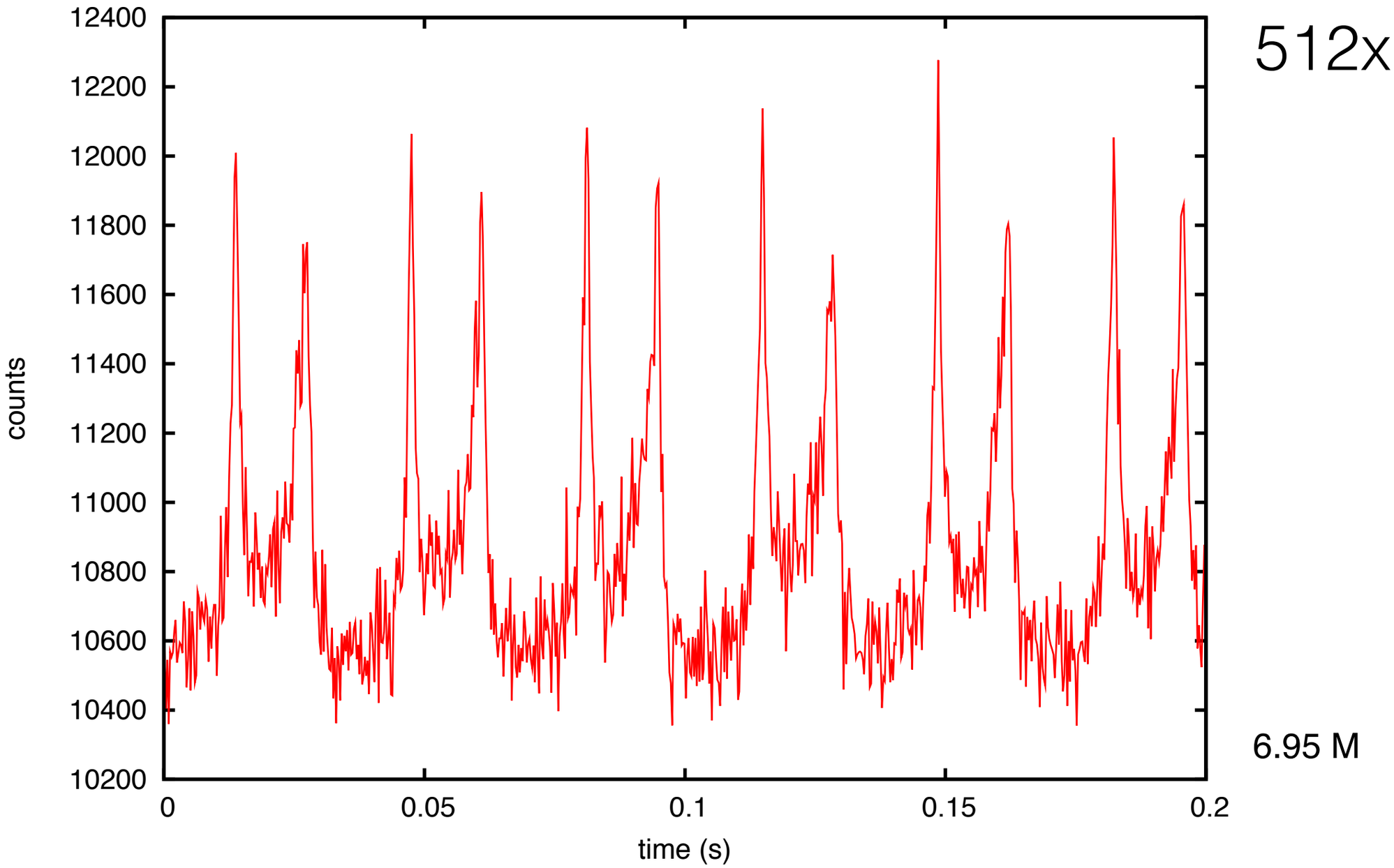

128x
1.74 M
counts
time (s)
3200
3100
3000
2900
2800
2700
2600
2500
0
0.05
0.1
0.15
0.2


256x
6100
6000
5900
5800
5700
5600
5500
5400
5300
5200
5100
counts
0
0.05
0.1
0.15
0.2
time (s)
3.48 M


12400
12200
12000
11800
11600
11400
11200
11000
10800
10600
10400
10200
counts
0
0.05
0.1
0.15
0.2
time (s)
512x
6.95 M

X-ray pulses from the Crab Pulsar

observed with the AstroSat mission, India



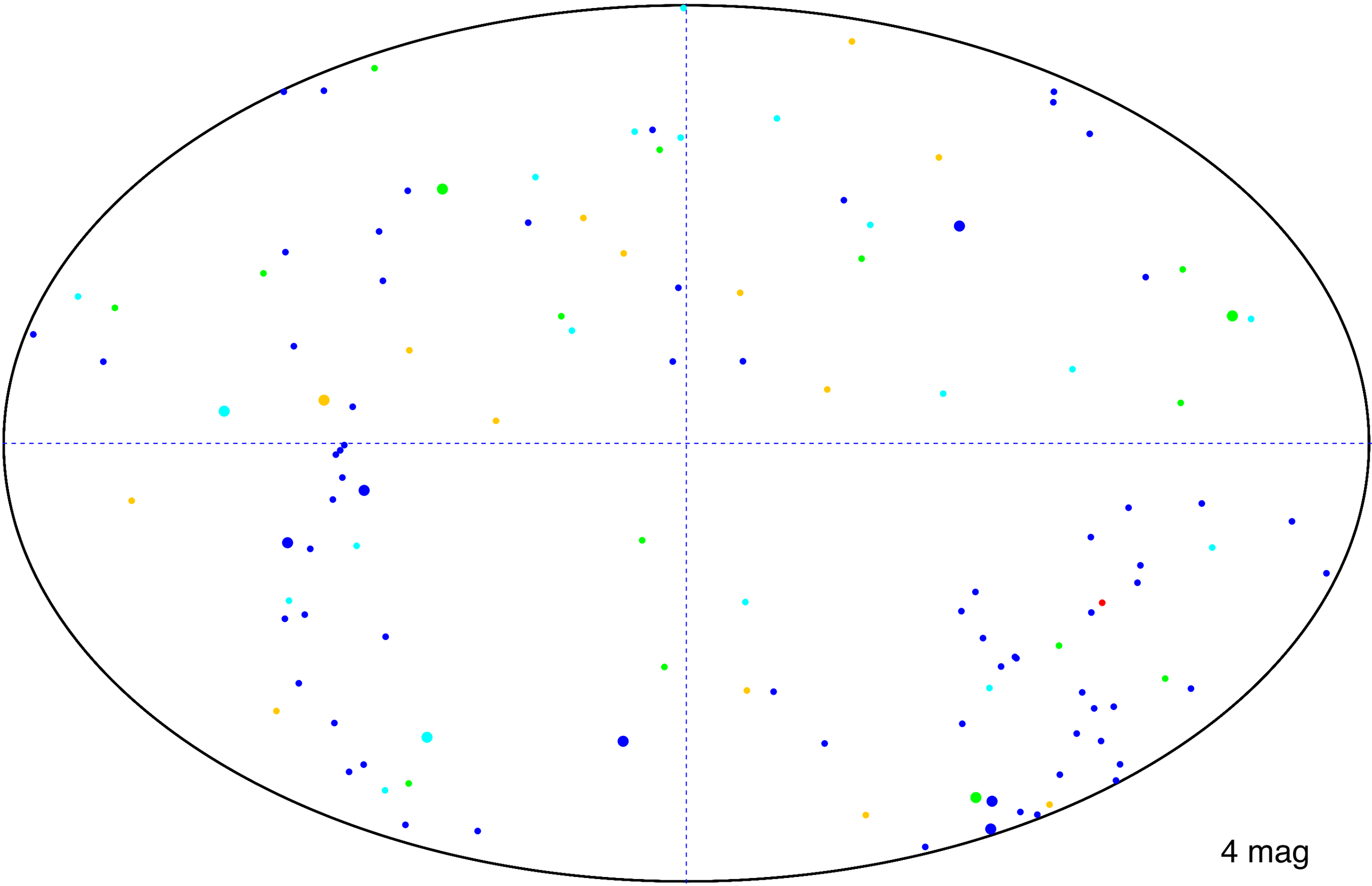
Unravelling patterns in distribution

Study large population

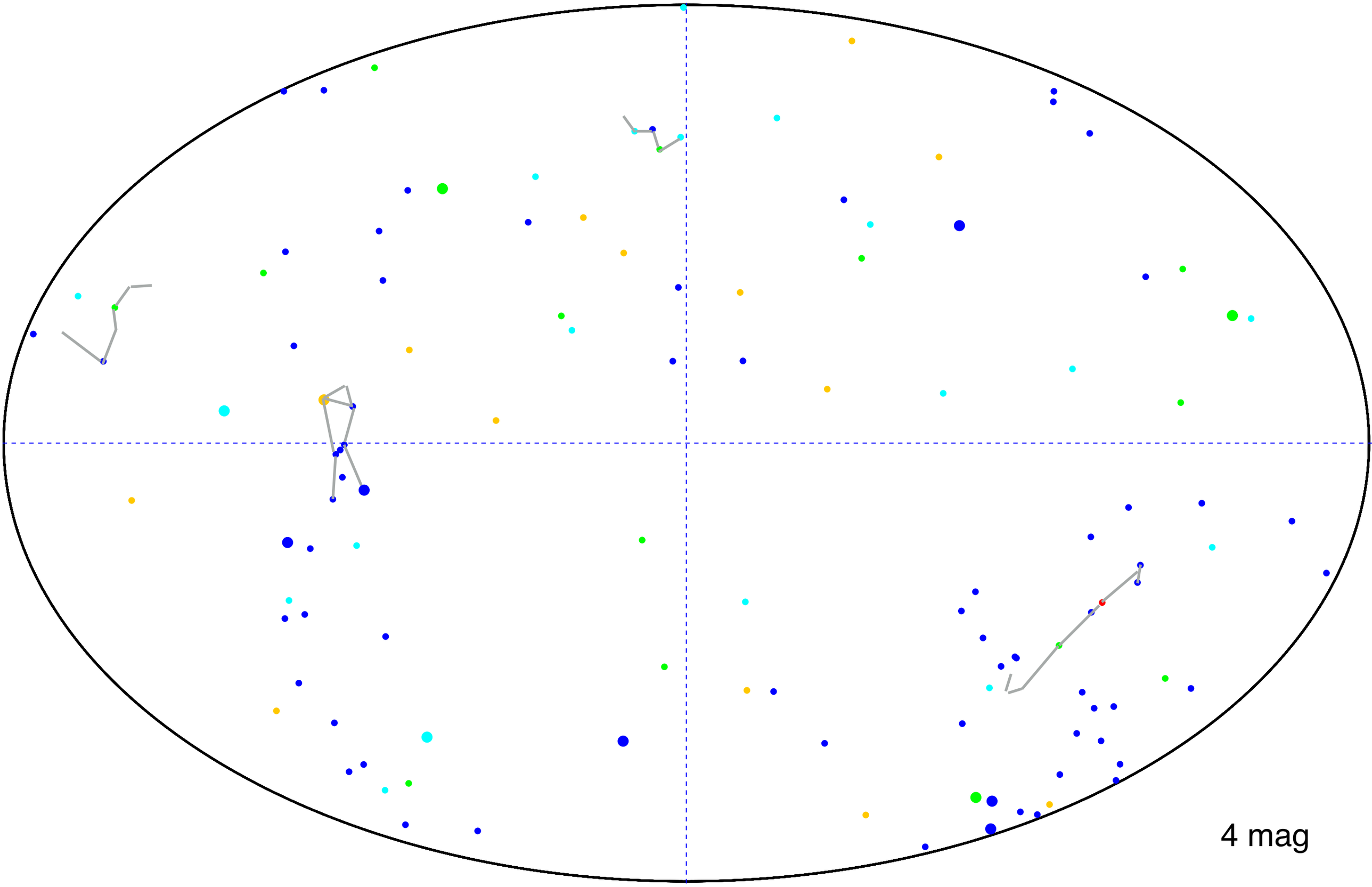
Arrange to reveal pattern

Measure statistical significance

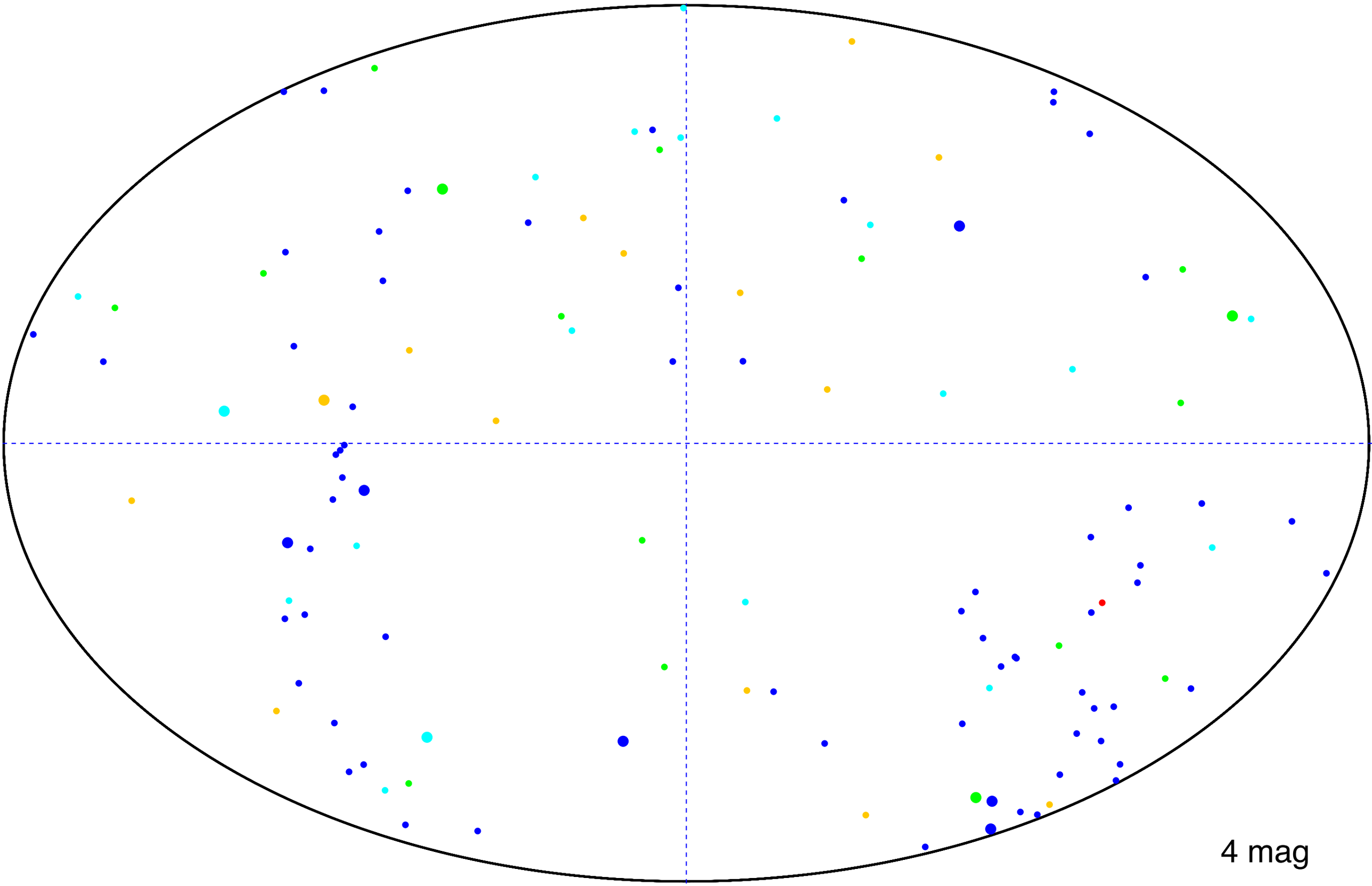
Hipparcos Sky



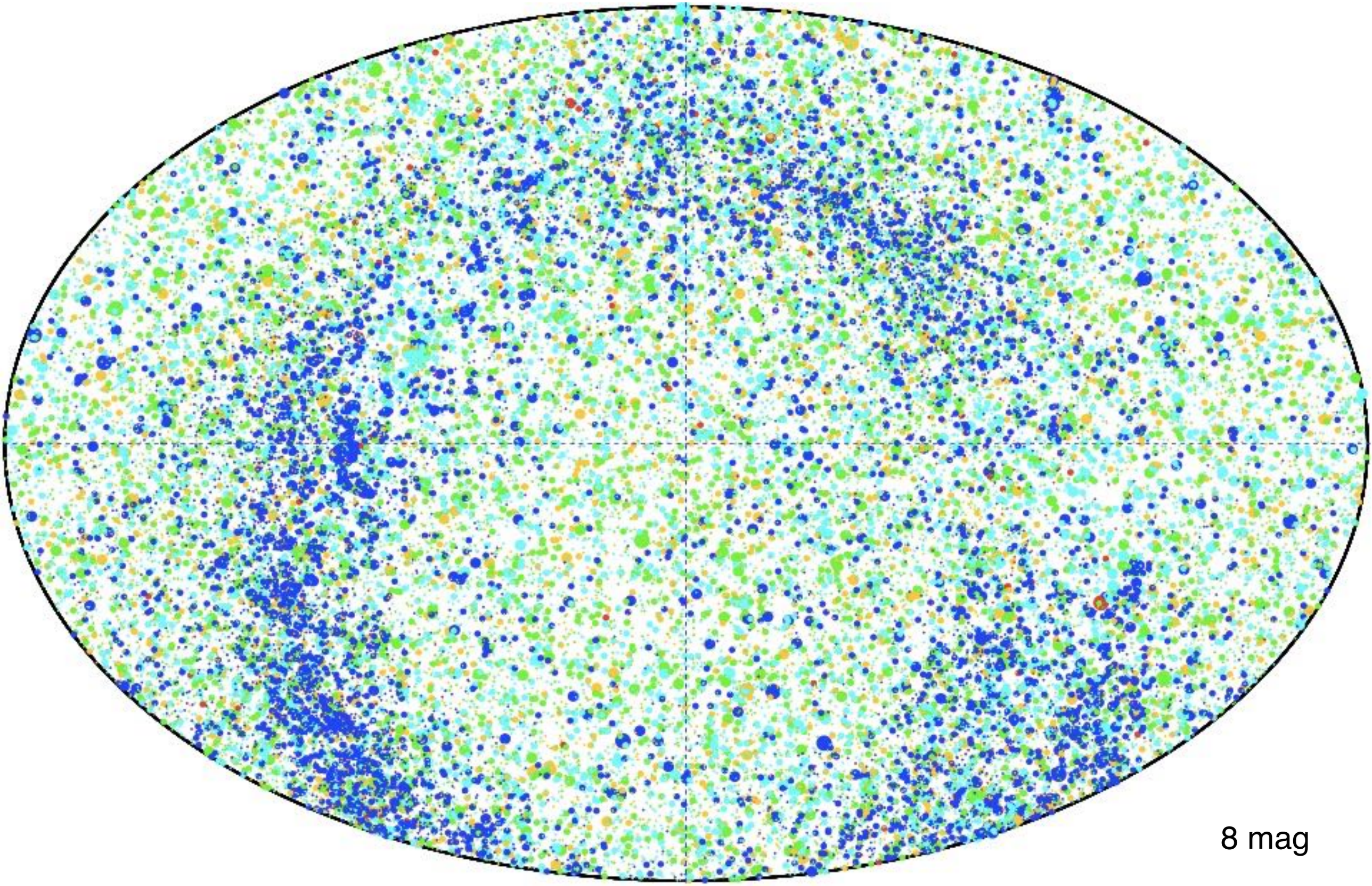
Hipparcos Sky



Hipparcos Sky

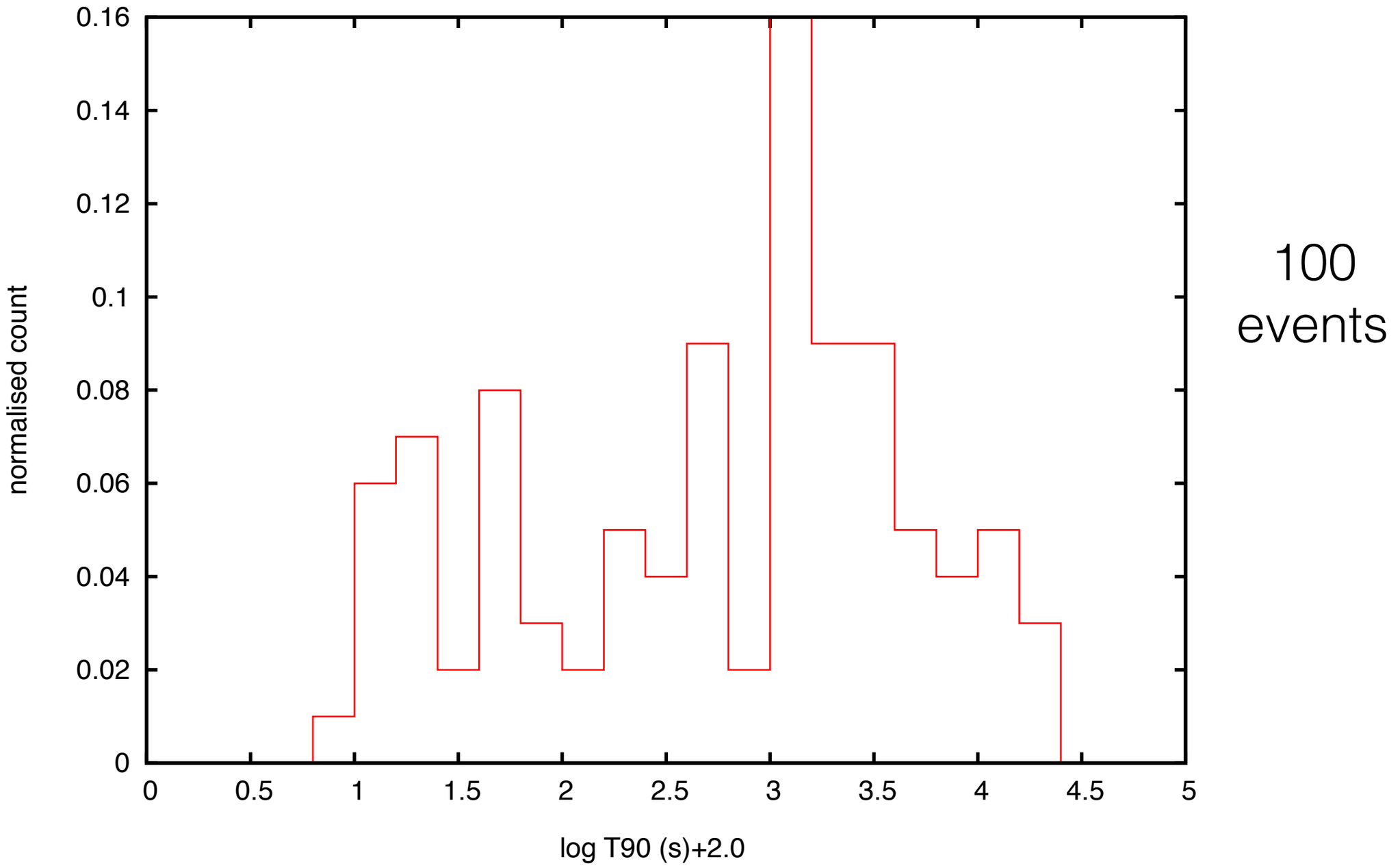


Hipparcos Sky

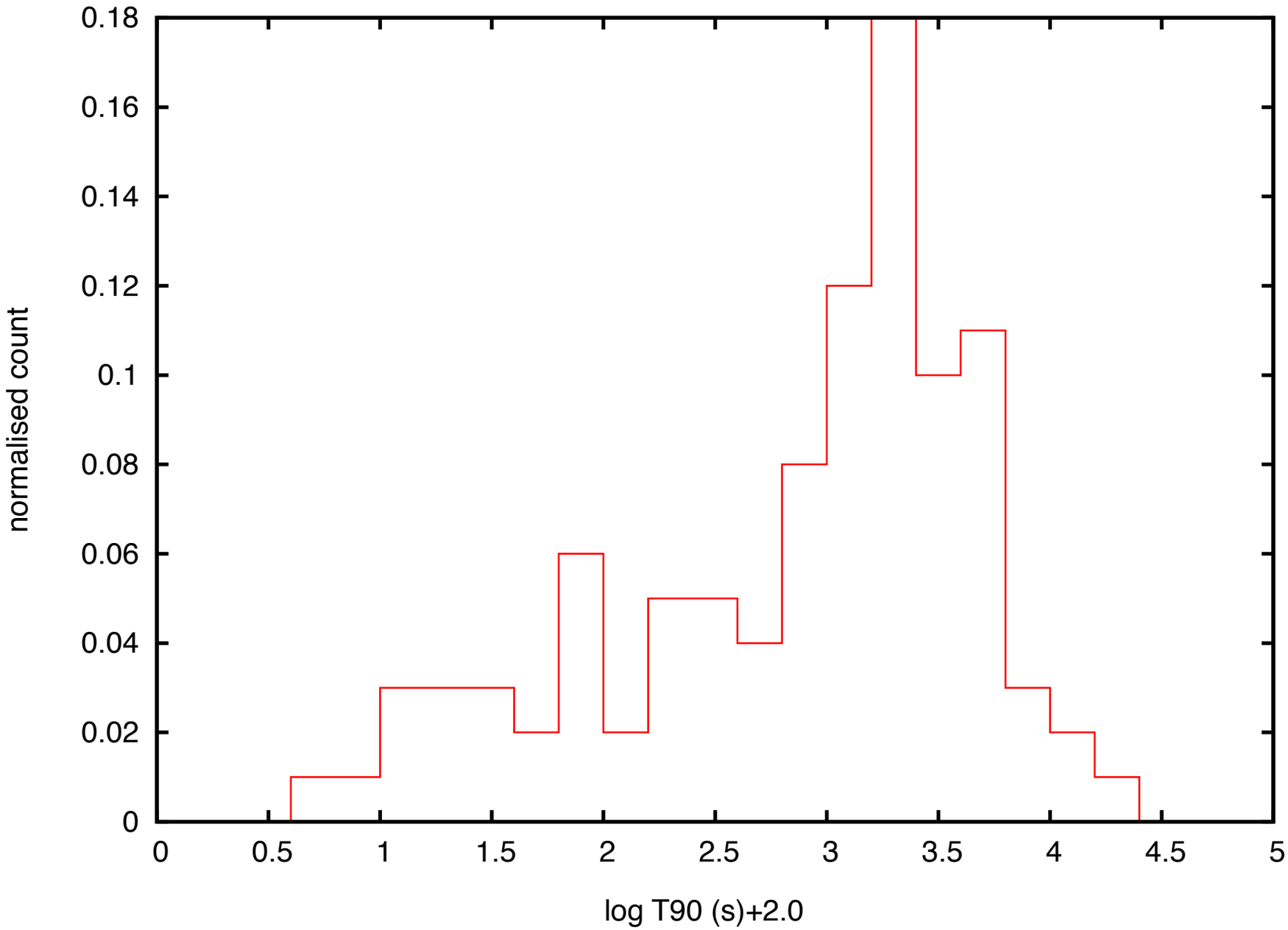


Distributions: Populations and sub-populations

Gamma Ray Bursts: Duration

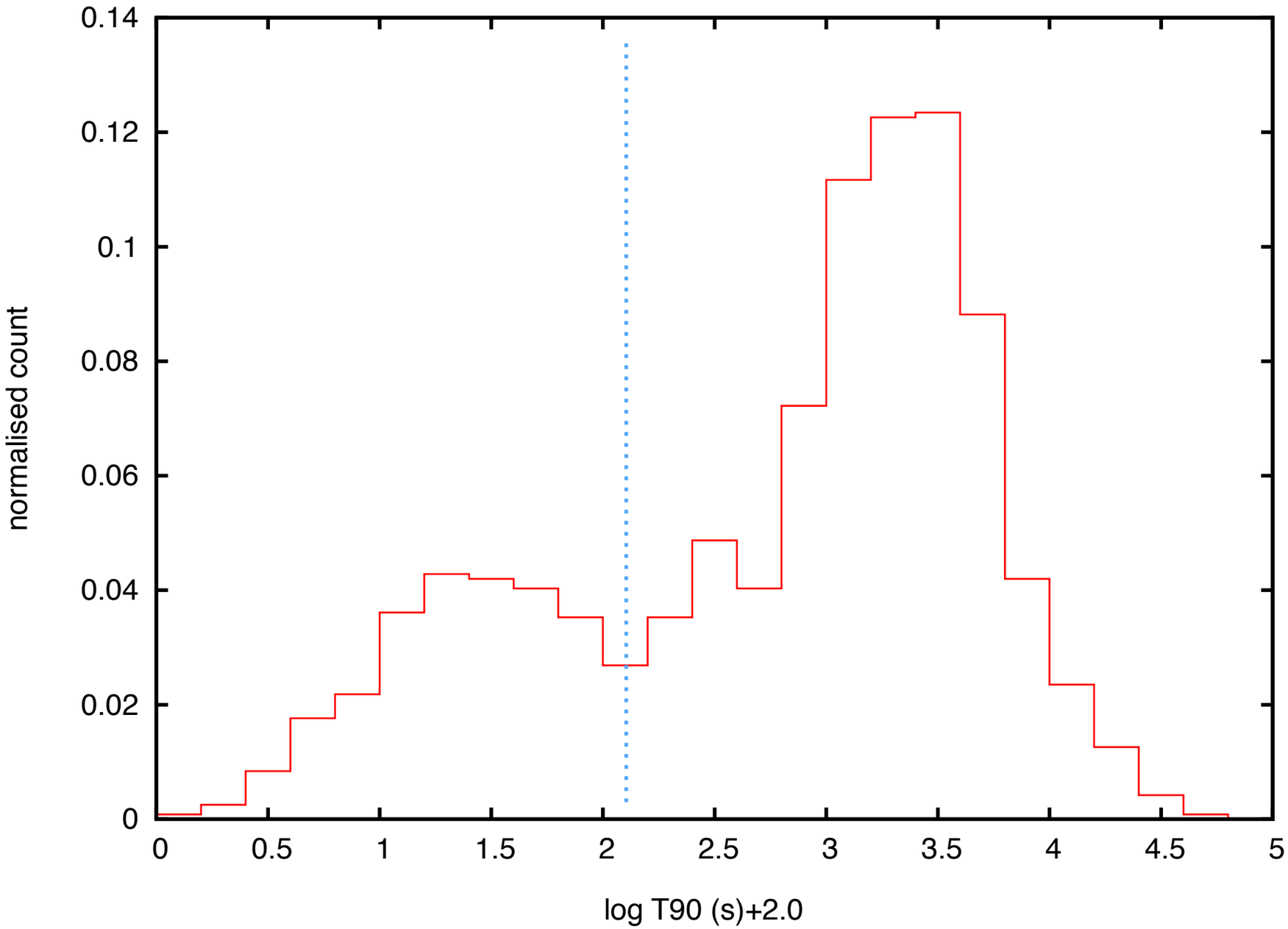


Gamma Ray Bursts: Duration



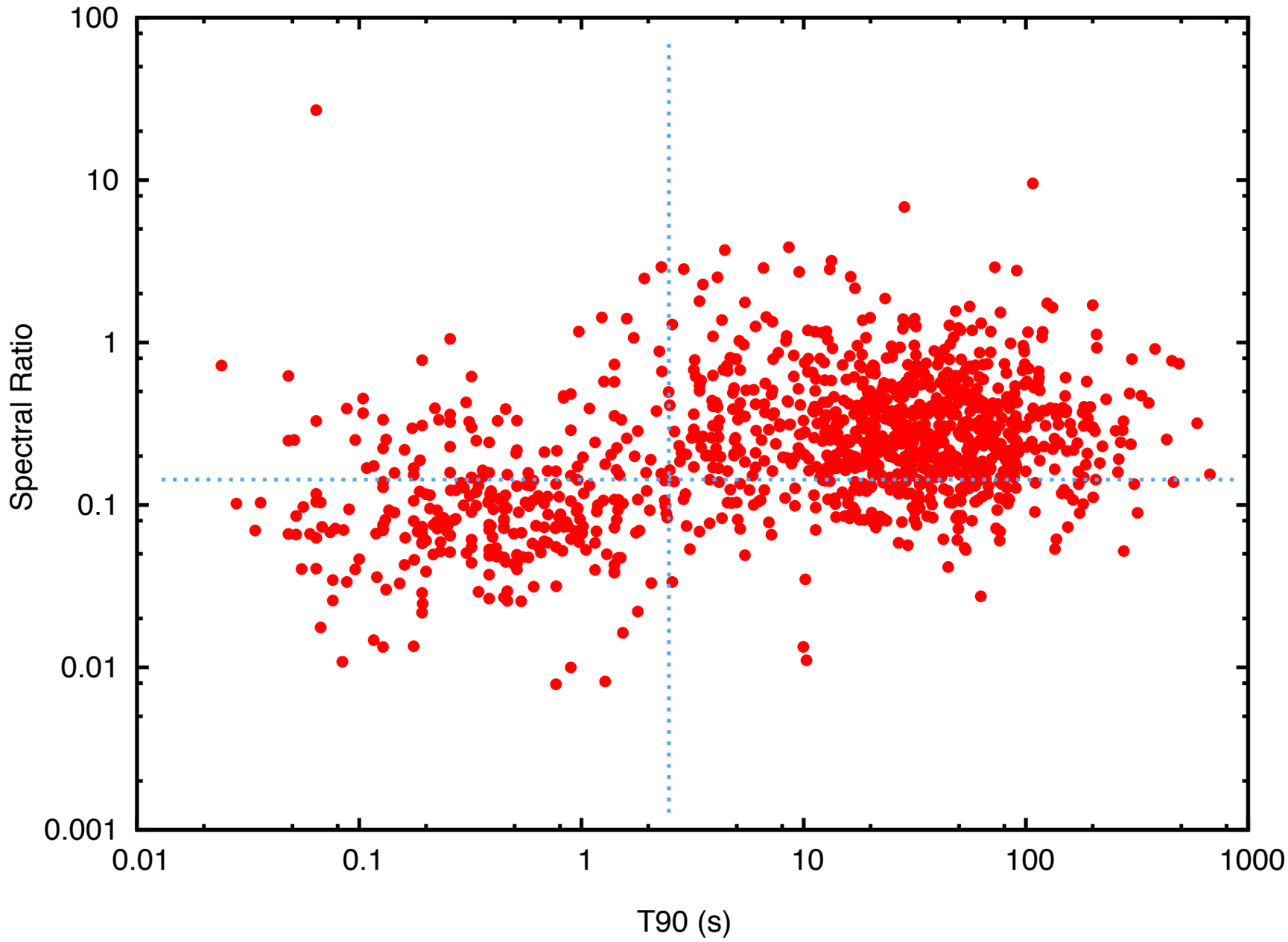
200 events

Gamma Ray Bursts: Duration



1200 events

Gamma Ray Bursts: Classification



Classification: a key step

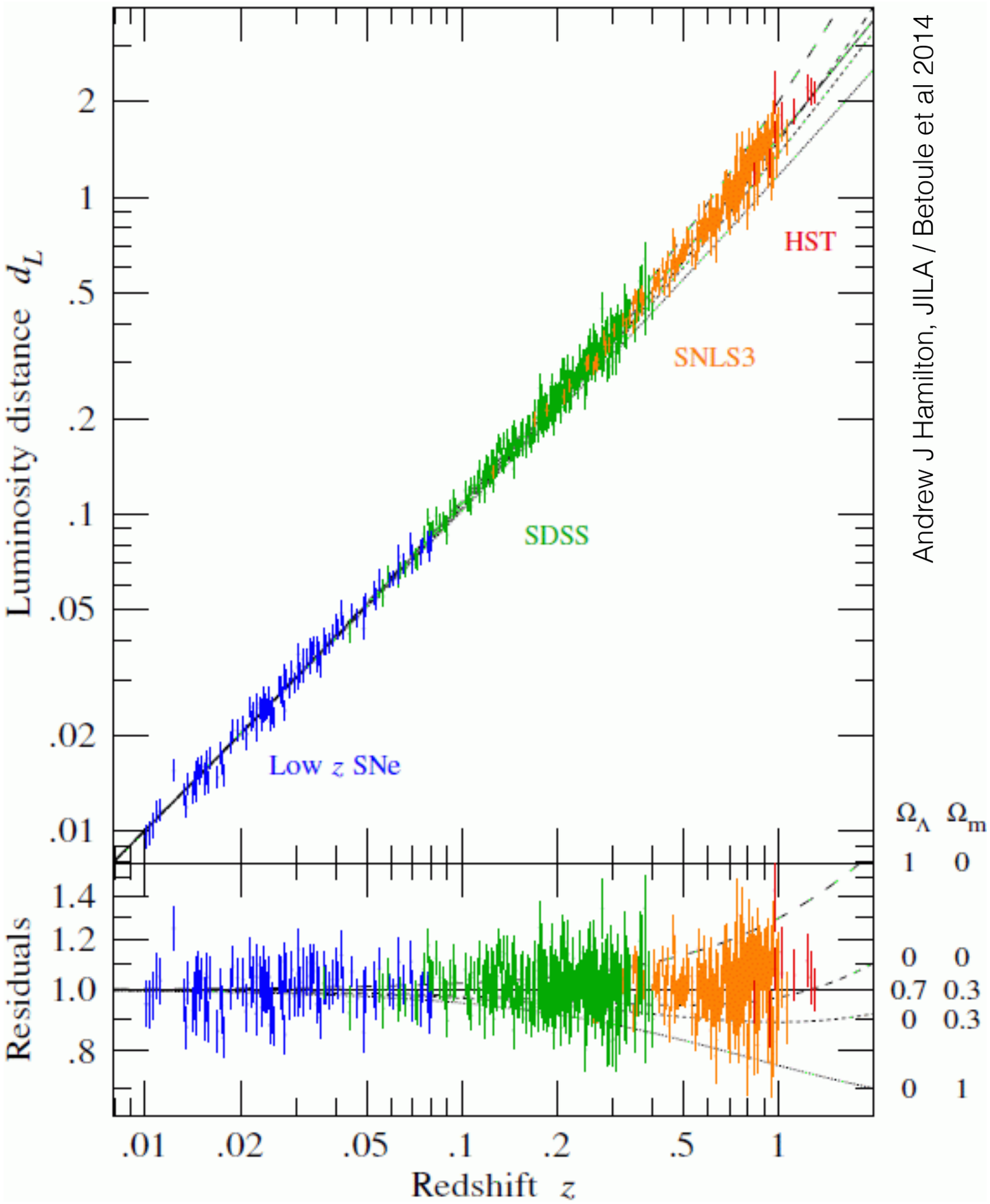
- Different phenomena can share some common properties, classification essential to separate them
- In case of transient events, classification essential to plan follow-up action
- Classification often needs multiple parameters
- Needs to be automated for large datasets and fast transients
- Various methods of Machine Learning are being employed

Correlations: Dynamics and Evolution

Supernova Cosmology Project

Measuring the history of expansion of the Universe

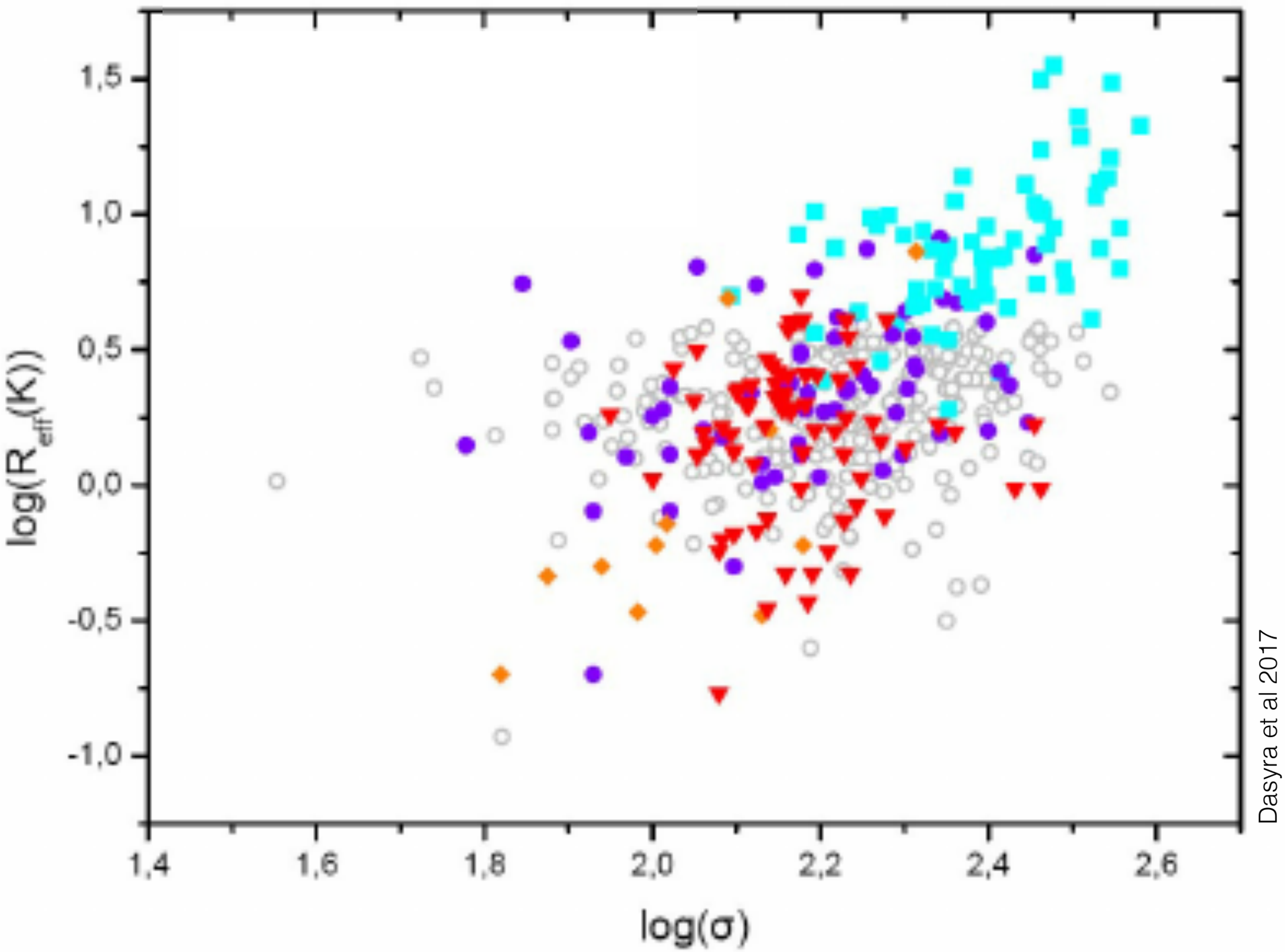
Larger and deeper sample led to the discovery of accelerated expansion: evidence of ***Dark Energy***



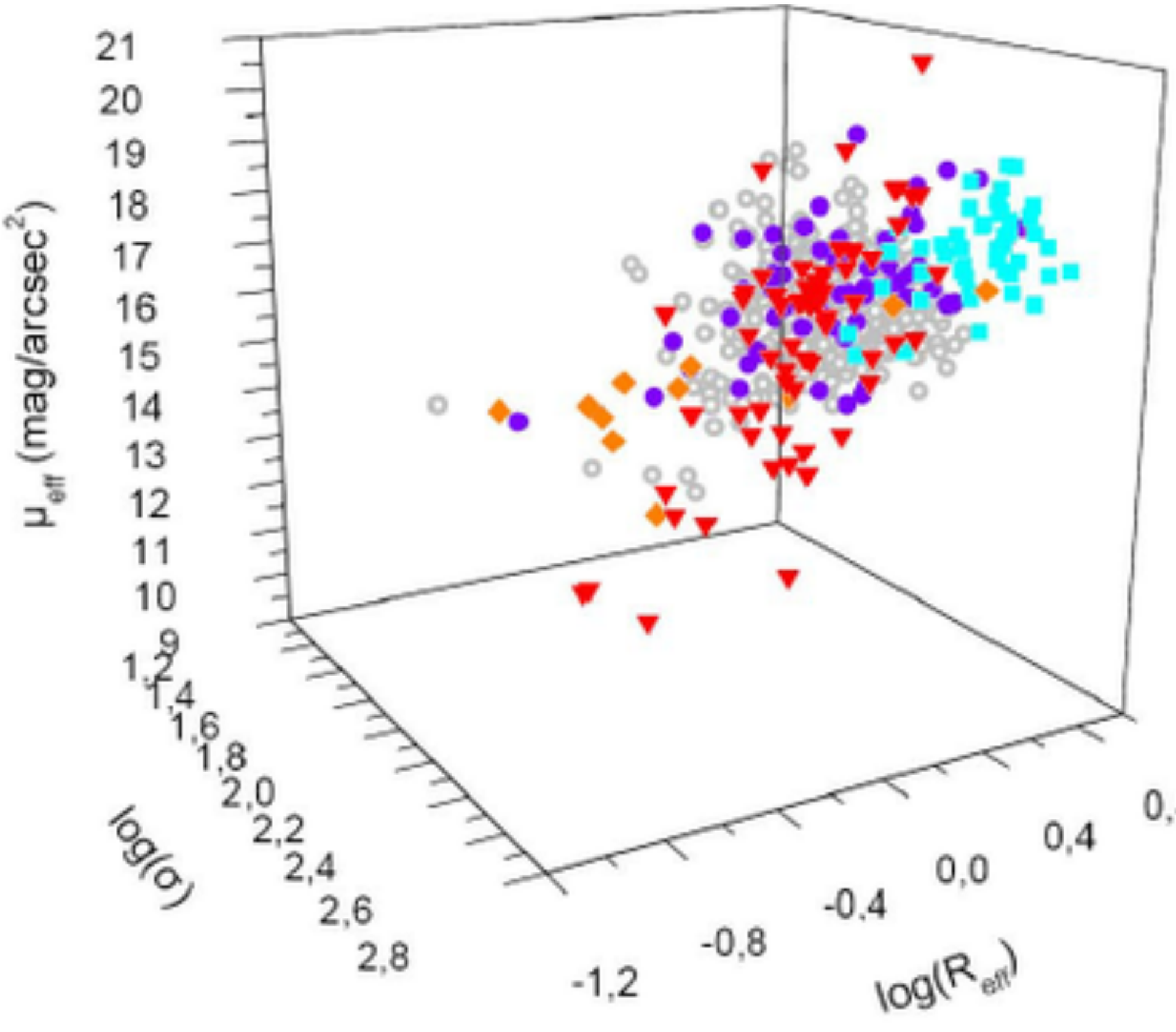
Andrew J Hamilton, JILA / Betoule et al 2014

Multivariate Correlations: Indication of underlying physical processes

Bivariate Correlation



Multivariate Fundamental Plane



Important manipulations on the data

- Search, Sort, Selection
- Visualisation
- Statistical Characterisation
  - moments of distribution
  - covariance, principal components
  - regression
- Hypothesis testing
  - maximum likelihood
  - Bayesian inference

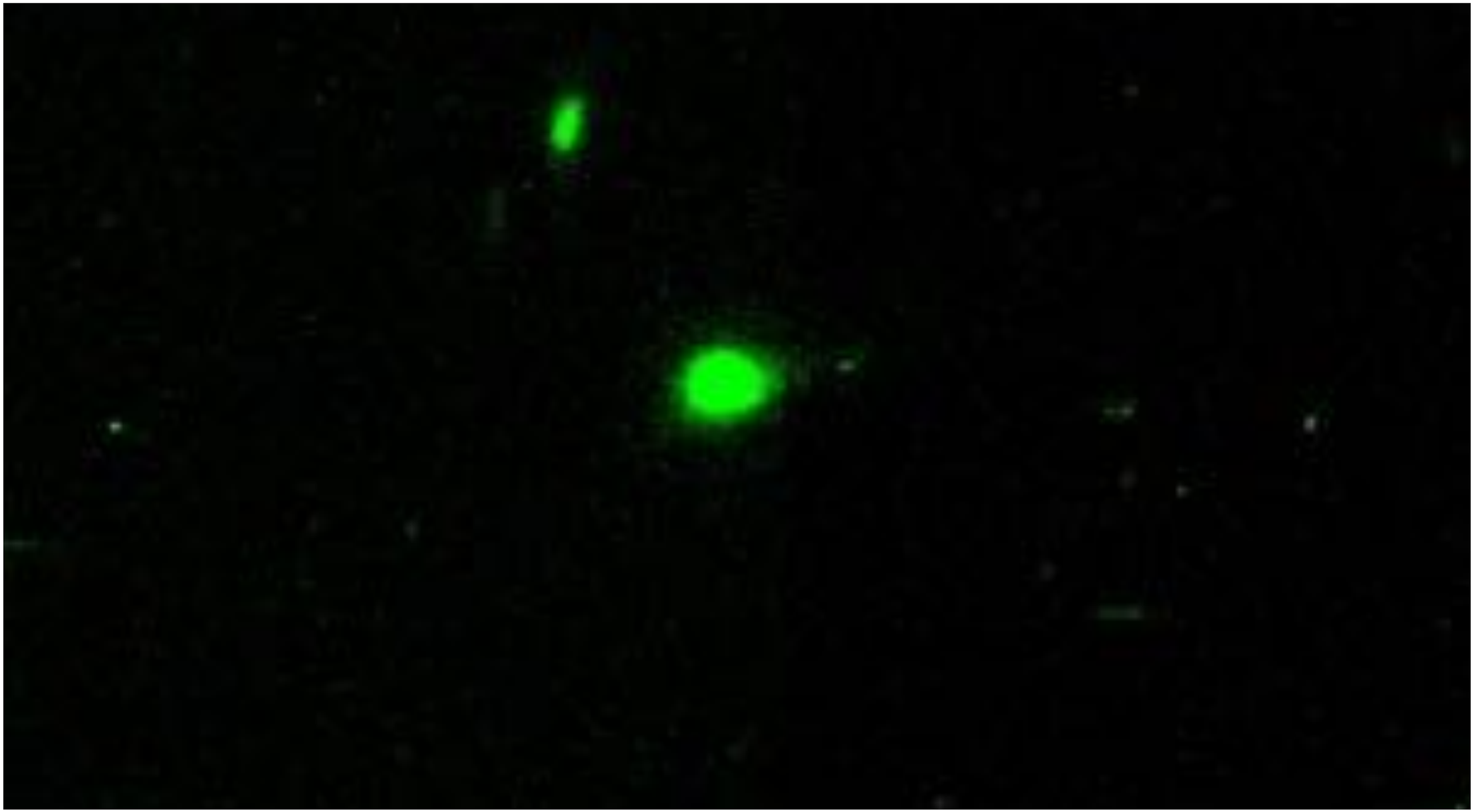
Large data size poses a challenge to all of these.
Special techniques are needed

Pattern recognition and classification using the human brain

Citizen Science:
The Galaxy Zoo Project

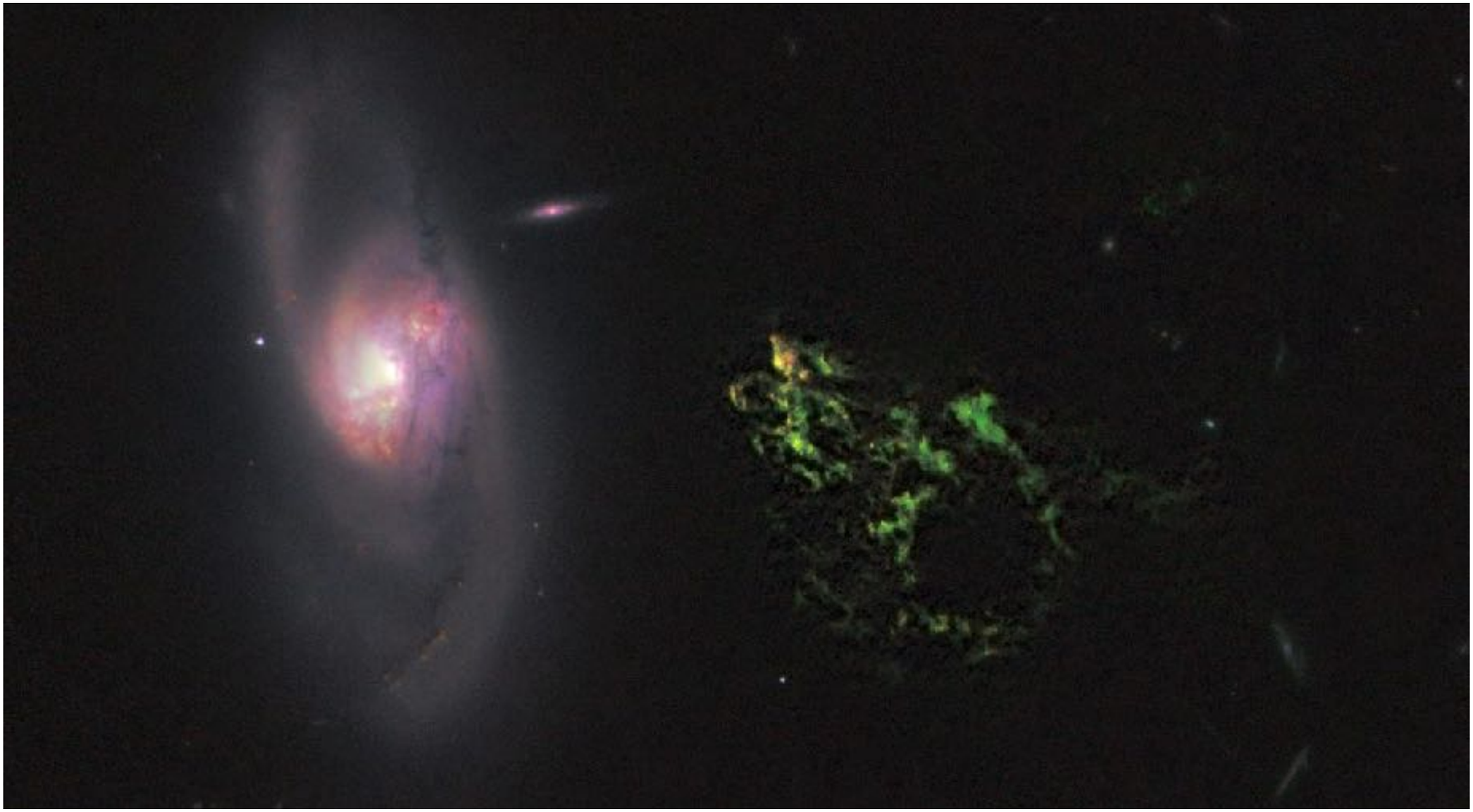
Citizen Science Discovery

“Green Peas” galaxies

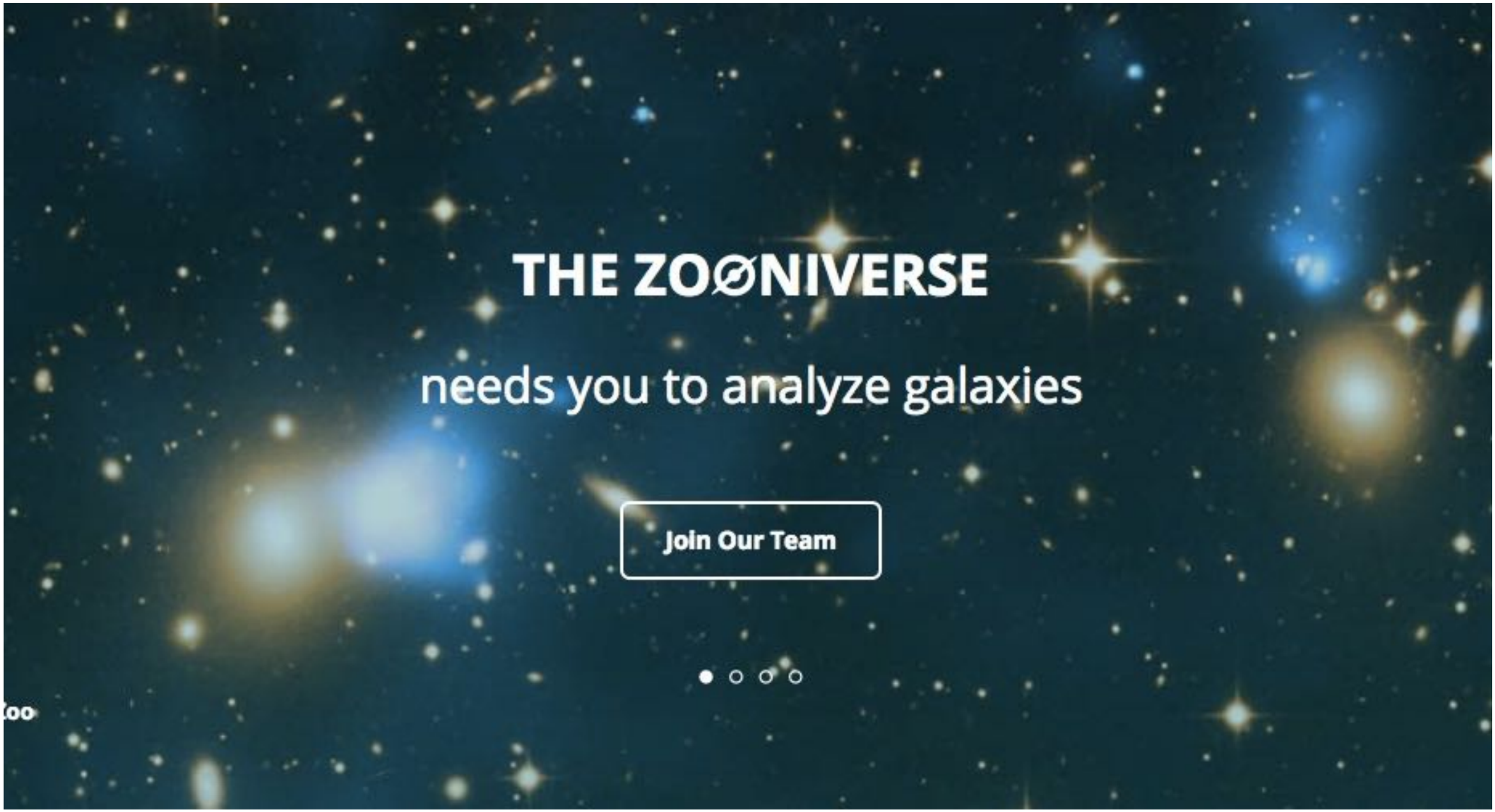


Citizen Science Discovery

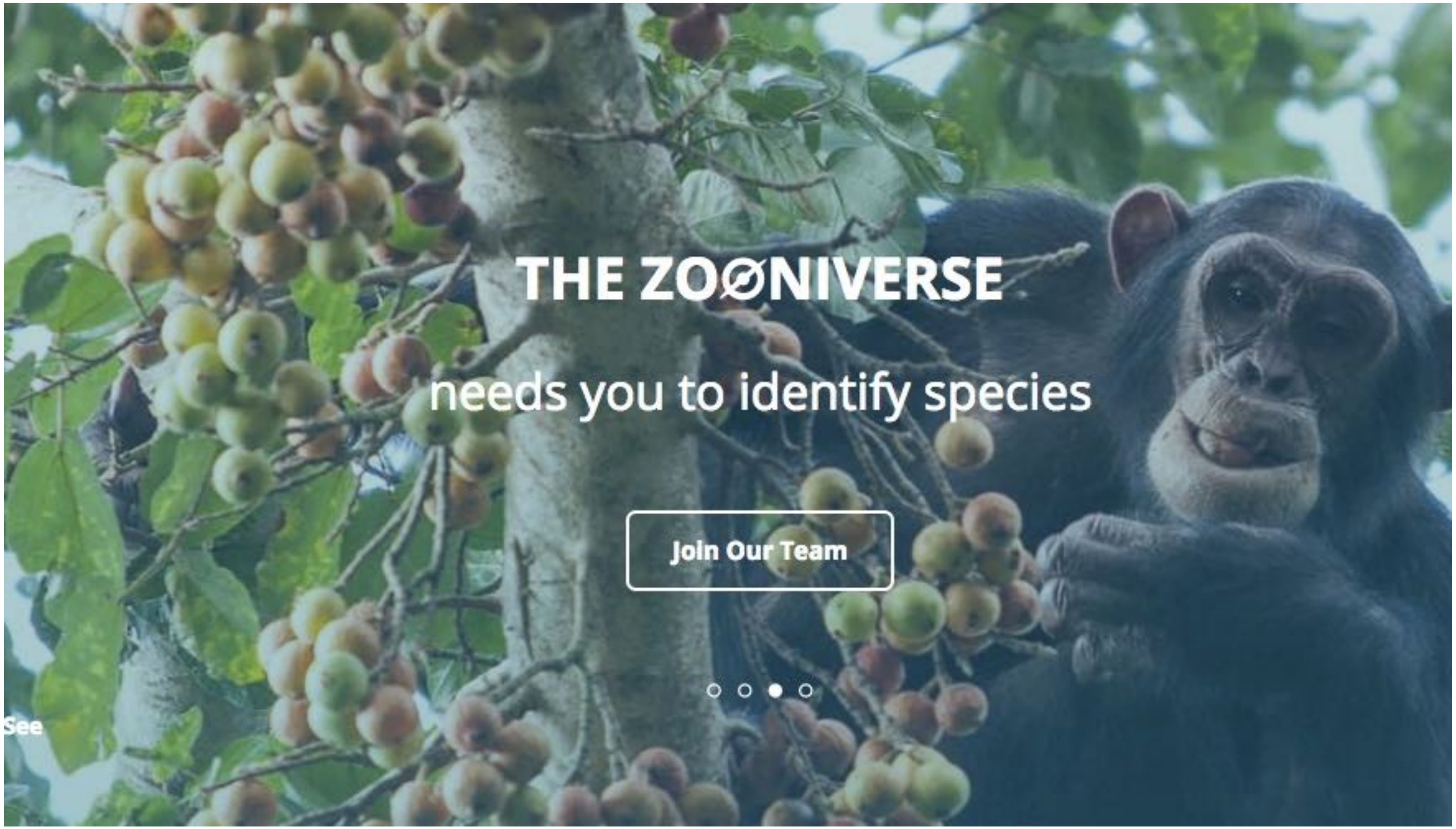
“Hanny’s Voorwerp”



www.zooniverse.org



www.zooniverse.org



www.zooniverse.org

Summary

- Contemporary science is being driven by rapidly increasing data volume
- Transportation, Analysis, Classification and Interpretation of large data sets pose new challenges
- Diverse new technologies, methods and algorithms are being developed to meet these challenges
- Visualisation a very important area, necessary for prototyping of automated methods - human ability of pattern recognition still wins in complex situations
- This workshop will familiarise you with some of the modern techniques of handling and using large data