

Computational AstroStatistics

Synergy between
statistics, computer
science and astronomy

Symbiotic Relationship e.g. PICA

PiCA Algorithms

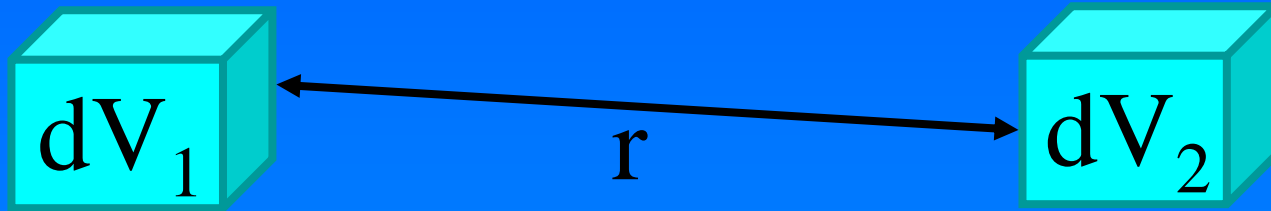
- Correlation functions (*Kayo et al. 2004; Scranton et al. 2004; Wake et al. 2004*)
- KDE codes (*Balogh et al. 2004*)
- Naïve Bayesian Classifier (*Richards et al. 2004*)
- Mixture models (*Connolly et al. 2000*)
- Anomaly Detection
- K-means clustering
- Kth nearest neighbors (*Balogh et al. 2004*)

All built for massive data sources

N-point correlation functions

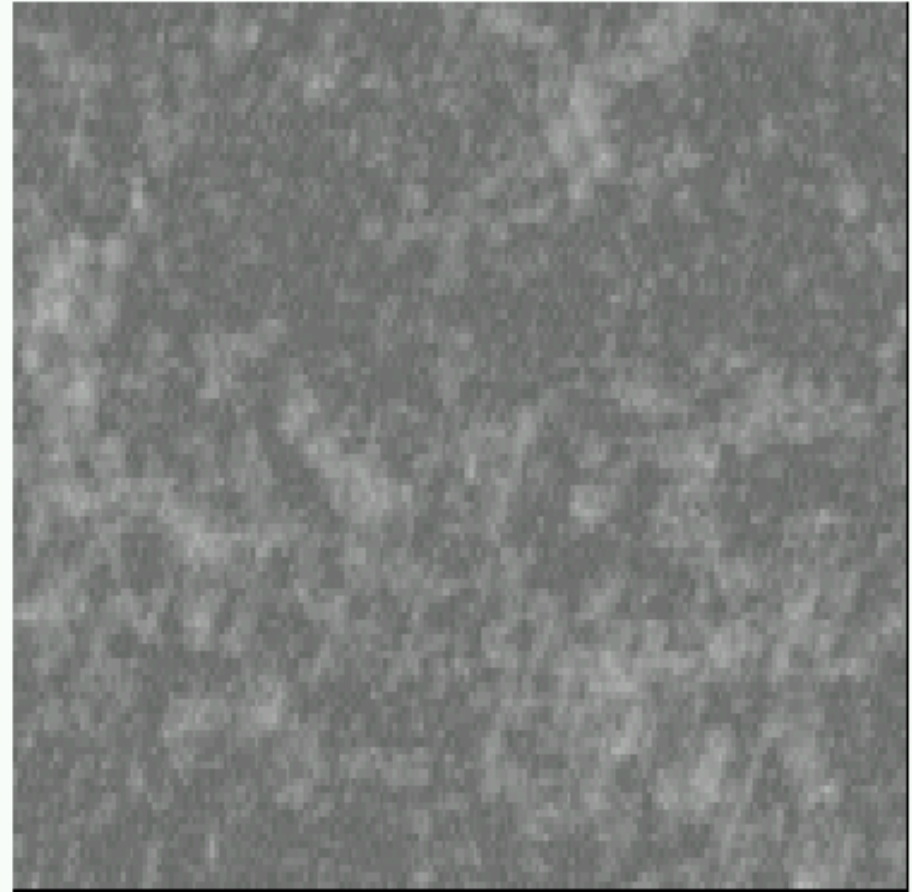
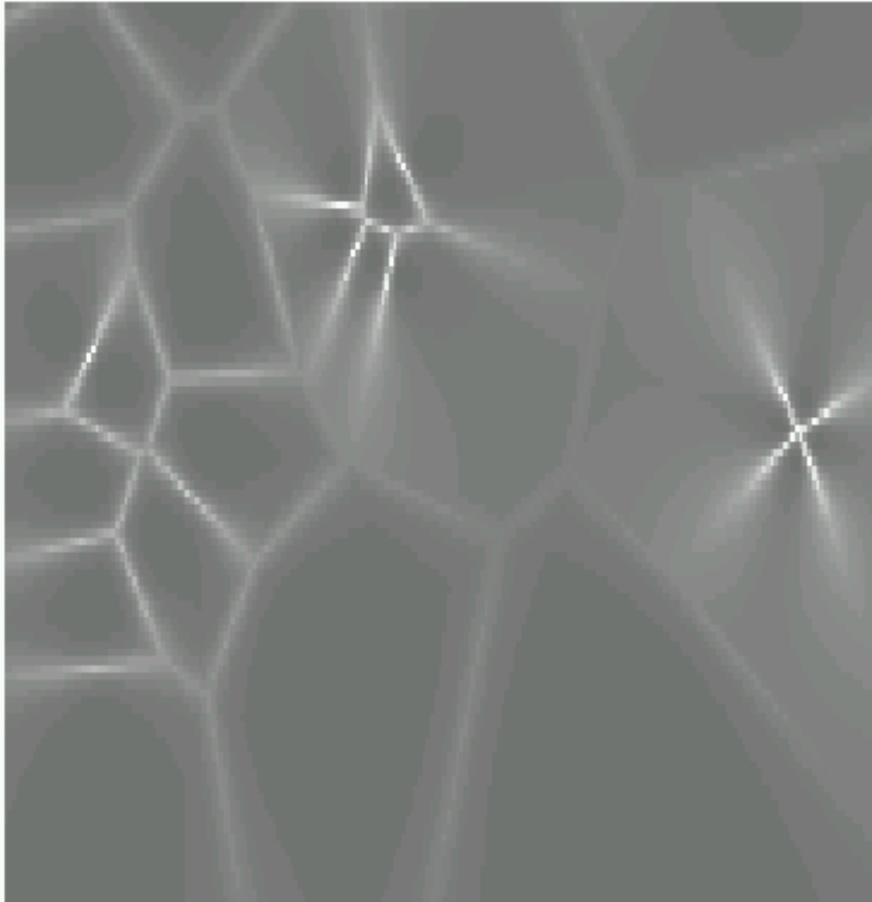
The 2-point function ($\xi(r)$) has a long history in cosmology (Peebles 1980). It is the excess joint probability (dP_{12}) of a pair of points over that expected from a Poisson process.

$$dP_{12} = n^2 dV_1 dV_2 [1 + \xi(r)]$$



$$dP_{123} = n^3 dV_1 dV_2 dV_3 [1 + \xi_{23}(r) + \xi_{13}(r) + \xi_{12}(r) + \xi_{123}(r)]$$

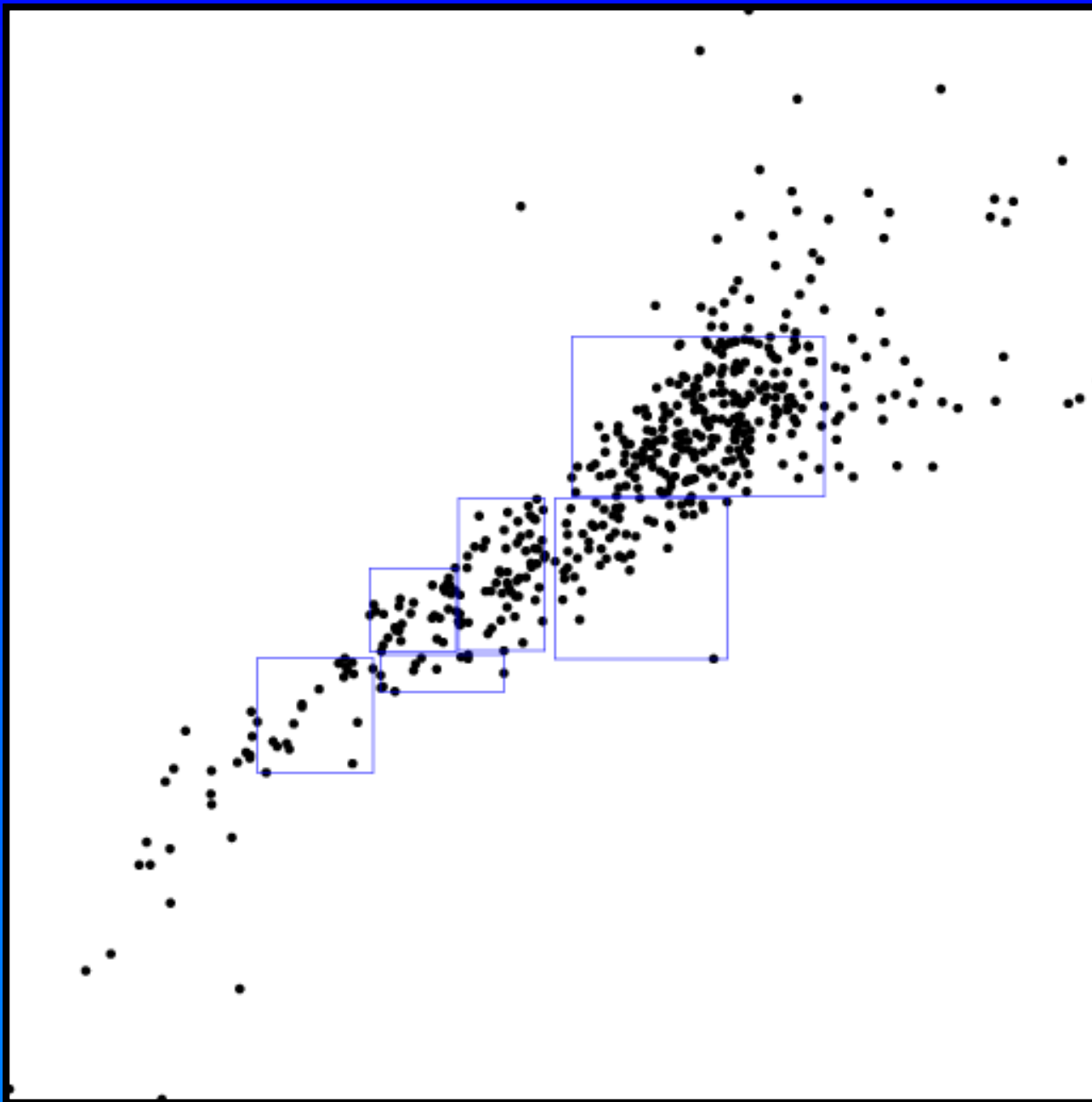
Motivation for the N-point functions: Measure of the topology of the large-scale structure in universe



Same 2pt, very different 3pt

Multi-resolutional KD-trees

- ❑ Scale to n-dimensions (although for very high dimensions use new tree structures)
- ❑ Use Cached Representation (store at each node summary sufficient statistics). Compute counts from these statistics
- ❑ Prune the tree which is stored in memory! (Moore et al. 2001 astro-ph/0012333)
 - ❑ Exact answers as it is all-pairs
 - ❑ Many applications; suite of algorithms!



Top Level

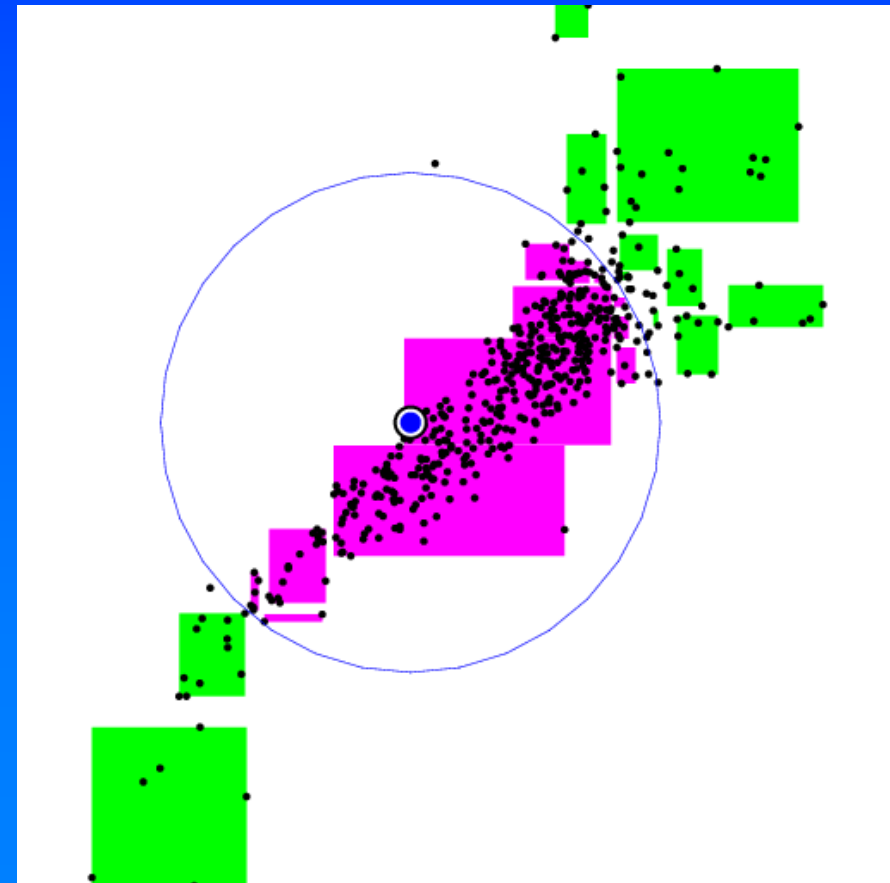
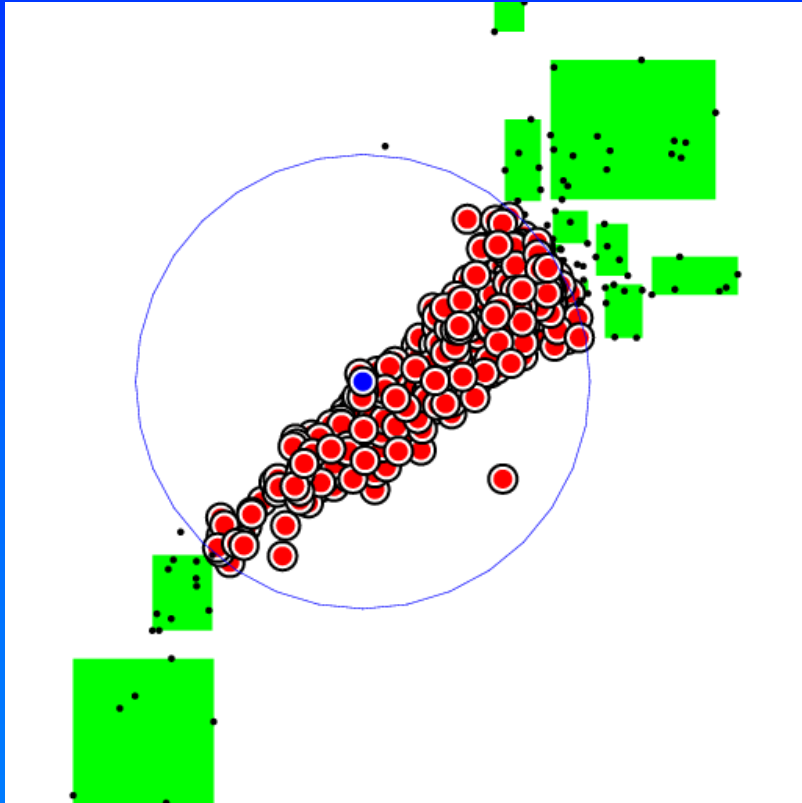
1st Level

2nd Level

5th Level

Just a set of range searches

Also Prune cells inside!
Greater saving in time



Prune cells outside range

Dual Tree Algorithm

Usually binned into annuli

$$r_{\min} < r < r_{\max}$$

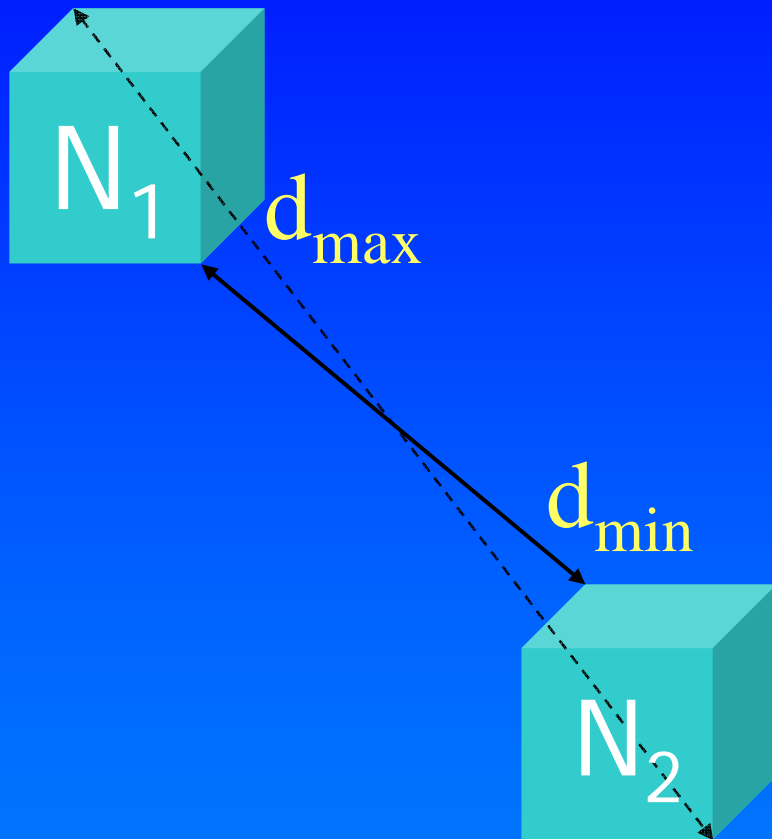
Thus, for each r transverse both trees and prune pairs of nodes

No count

$$d_{\min} < r_{\max} \quad \text{or} \quad d_{\max} < r_{\min}$$

$$N_1 \times N_2$$

$$r_{\min} > d_{\min} \quad \text{and} \quad r_{\max} < d_{\max}$$



Therefore, only need to calculate pairs cutting the boundaries.

Scales to n-point functions also do all r values at once

Faster!

How does one compute the 4pt function for a billion galaxies?

Need to accept regime of approximate answers. The tree provides a new form of stratification for the monte carlo variance-reduction techniques.

Build conditional probability functions for the counts and return these probabilities as an approximate answer rather than the true count
(Alex Gray 2003)

Also explore distributed data structures on distributed computing

Summary

- Techniques and codes now available to do massive computation on present data sets. Need to disseminate these via VO infrastructure
- Need to explore approximate answers and distributed computations for next generation of data sets.
- Synergy of visualization and data-mining is vital to efficiently guiding data-mining and observing results