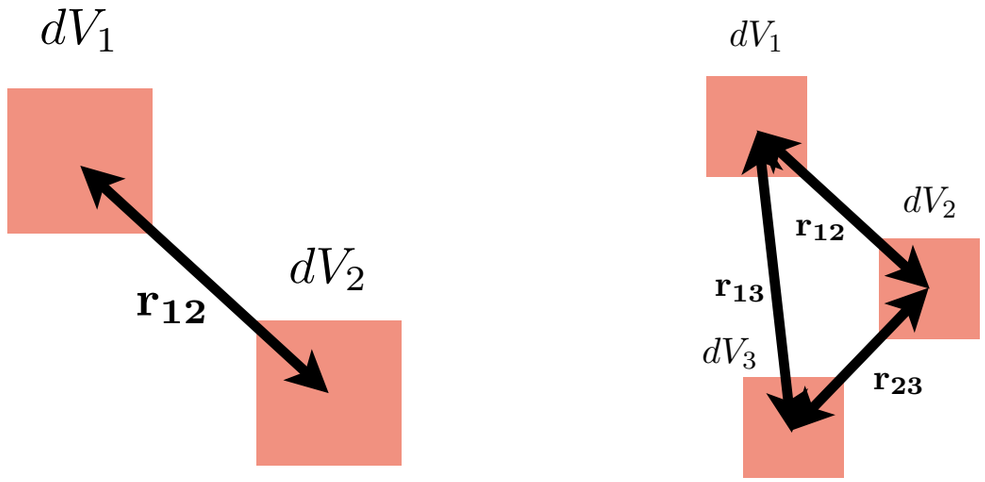
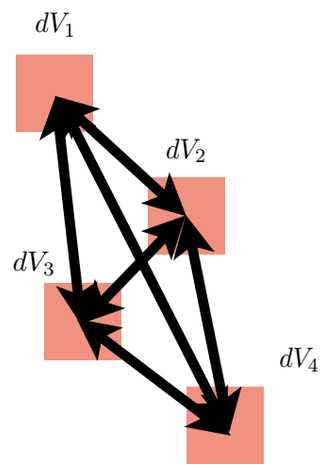
Efficient Estimation of N-point Spatial Statistics

- *n-point correlation functions* give the probability of points occurring in a given configuration



- A *general, powerful* spatial statistic, *capable of fully characterizing any distribution*

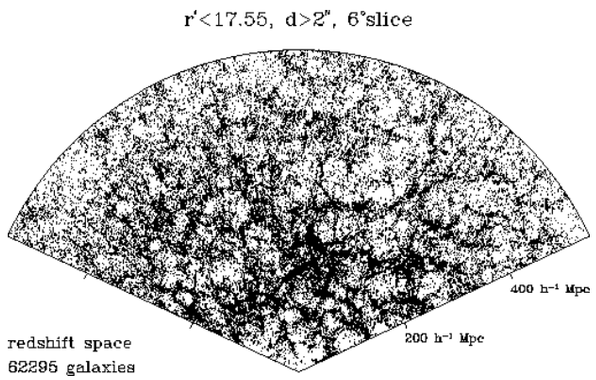
- Previously used to understand:
  - *Hierarchical structure formation*
  - *Gaussianity of the early universe*
  - *Models of galaxy mass bias*



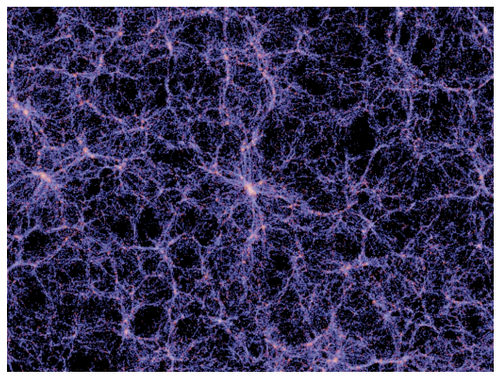
Computational Task

- Estimate n-point functions by counting *n*-tuples of points satisfying some distance constraints - $O(N^n)$ directly, *per set of constraints*
- Need many sets of constraints - repeat computation *M* times
- Need to estimate variance - repeat the computation for *J* subsamples
- Need large *n* (at least 3) to accurately distinguish distributions

SDSS (millions of points)



Virgo Sim. (billions of points)



Overall complexity: $O(J \cdot M \cdot N^n)$

Efficient Computation

- Build *kd*-trees on the data
- Compare *n* nodes, prune if distance bounds allow

New

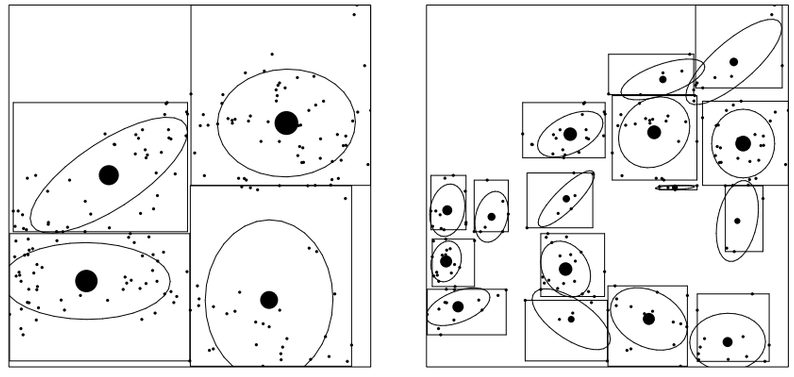
Share information among different matchers

- overcome dependence on M

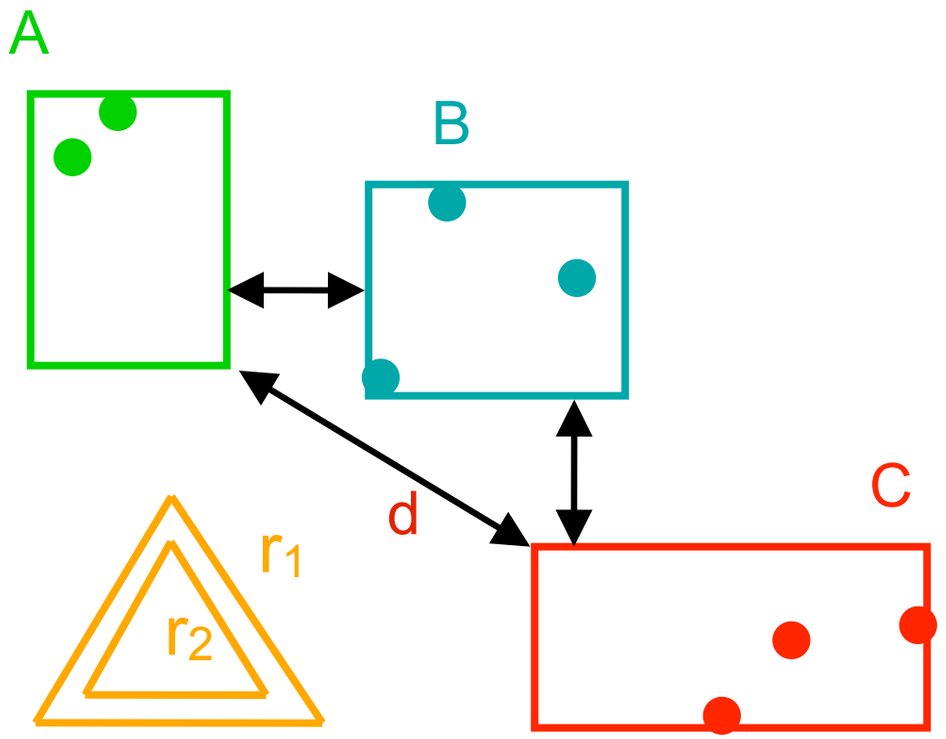
New

Incorporate jackknife resampling directly

- overcome dependence on J



kd-tree Level *2* *kd*-tree Level *4*



prune if $d > r_1$

Preliminary Results & Ongoing Work

	Multi-bandwidth *new*	Single bandwidth [Moore, et al, 2001]	Naive - $O(N^n)$ (estimated)
2 point cor. 100 matchers	4.96 s	352.8 s	2.0×10^7 s
3 point cor. 243 matchers	13.58 s	891.6 s	1.1×10^{11} s
4 point cor. 216 matchers	503.6 s	14530 s	2.3×10^{14} s

10^6 mock galaxies

- Heterogeneous Architectures: perform leaf-leaf computations very efficiently on GPU
- Massively Parallel tree code: scales to thousands of processors