Recent Geometric Multiscale Algorithms for n-point Correlations

Alexander Gray

Carnegie Mellon University

School of Computer Science

Joint work with

Andrew Moore, Andrew Connolly, Robert Nichol, Alexander Szalay, Istvan Szapudi, David Wake, Gauri Kulkarni, Christopher Genovese, Larry Wasserman

References

- [Gray-Moore, NIPS 2000]
- [Moore et al., Mining the Sky 2000]
- [Gray et al., ADASS 2004]
- [Gray et al., Computational Physics 2004]
- [Gray et al., to be submitted 2005]

Outline:

1. **Some cosmological questions**
2. **Spatial correlations**
3. **Divide-and-conquer in real-space**
4. **Multi-tree algorithm, exact**
5. **Multi-tree Monte Carlo**

Outline:

1. **Some cosmological questions**

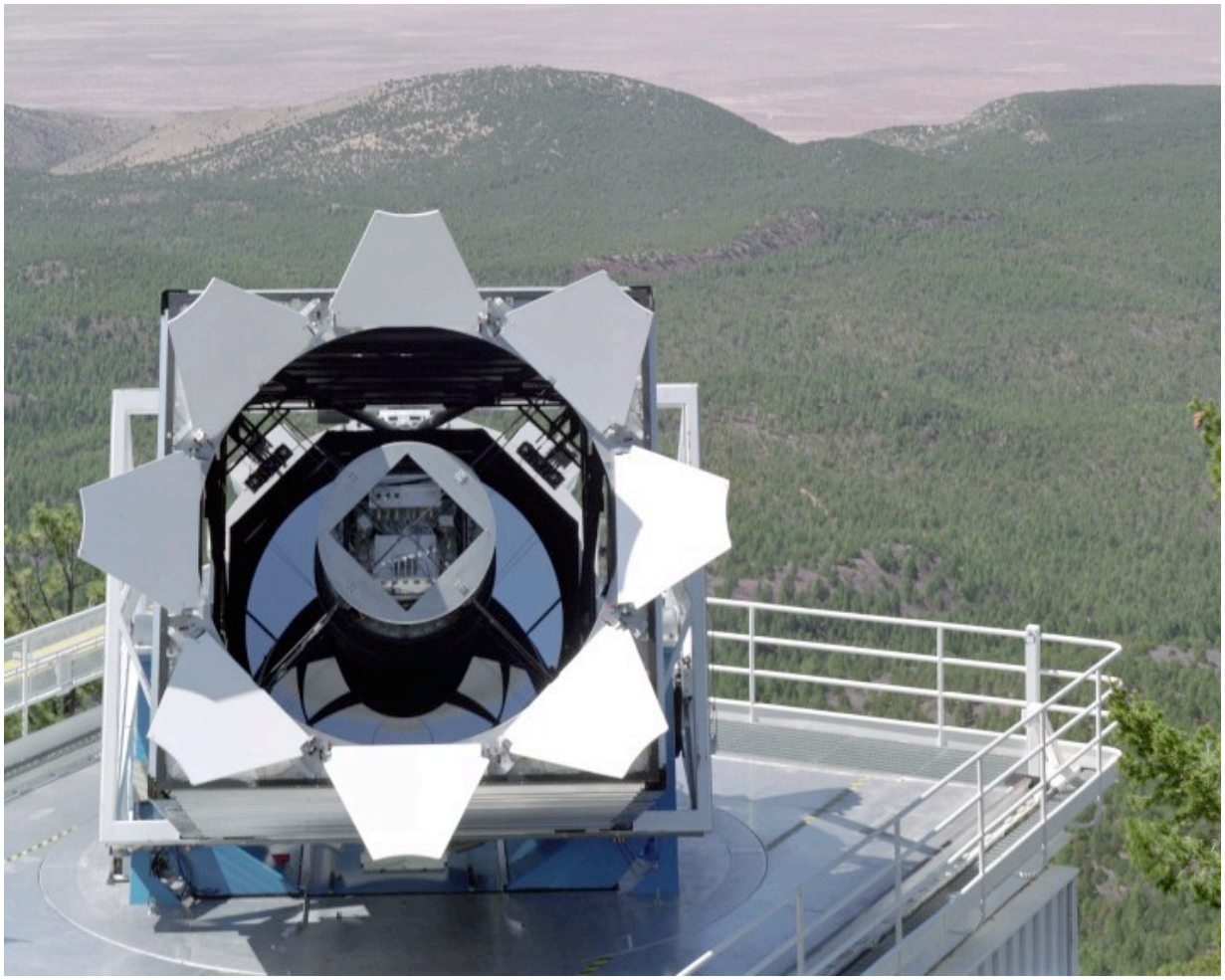
2. **Spatial correlations**

3. **Divide-and-conquer in real-space**

4. **Multi-tree algorithm, exact**

5. **Multi-tree Monte Carlo**

Optical telescopes have been getting bigger and faster…

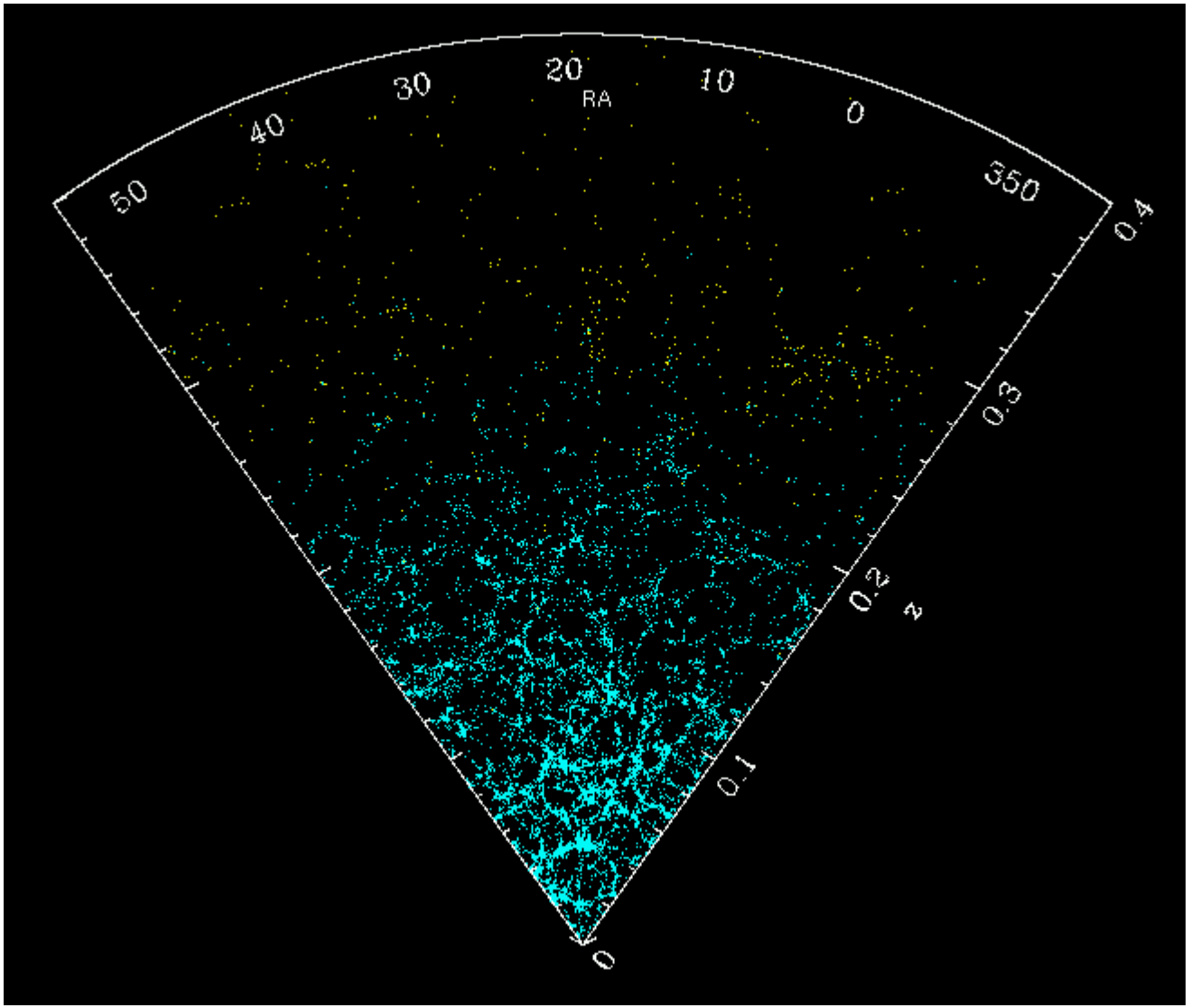


…resulting recently in the **Sloan Digital Sky Survey (SDSS)**:

~ 1 billion objects, 144 dimensions

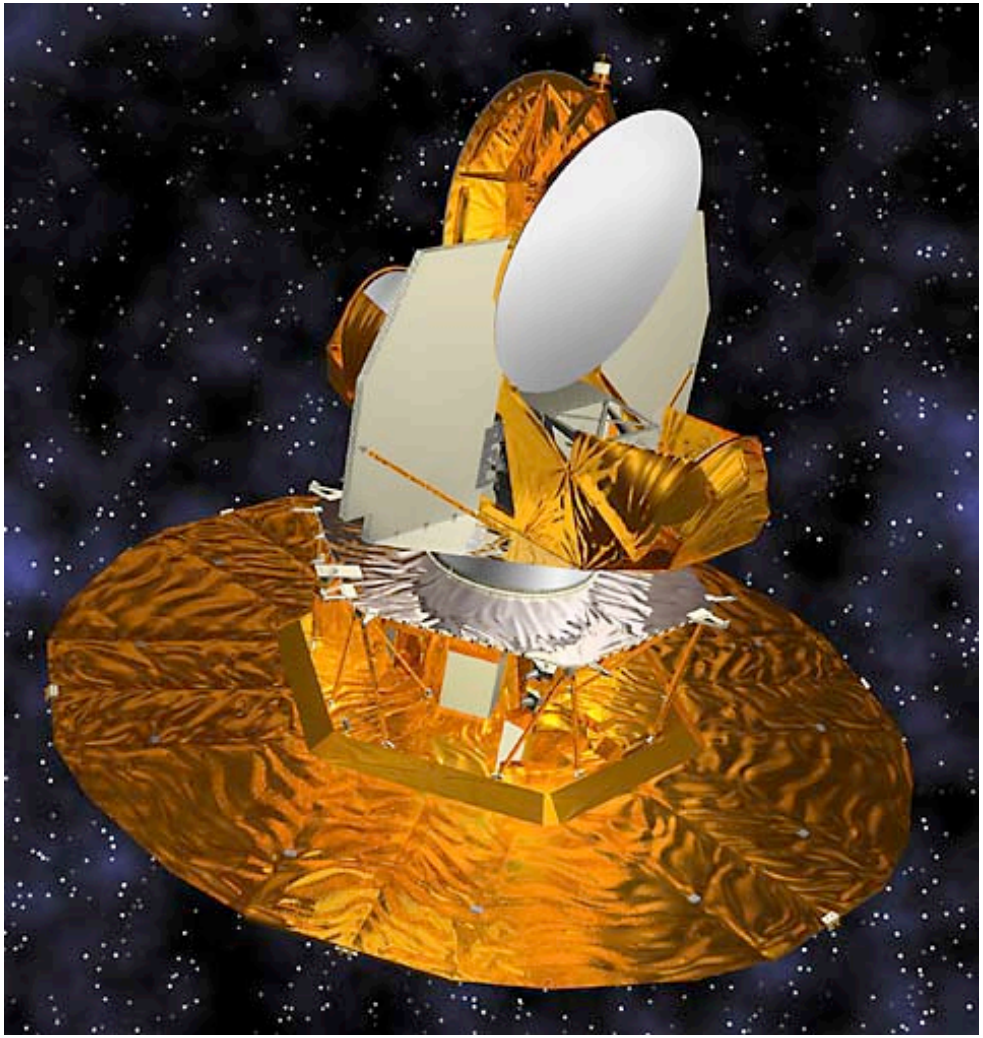
(250M now in 5 colors, 500K 2000-D spectra)

What we see:



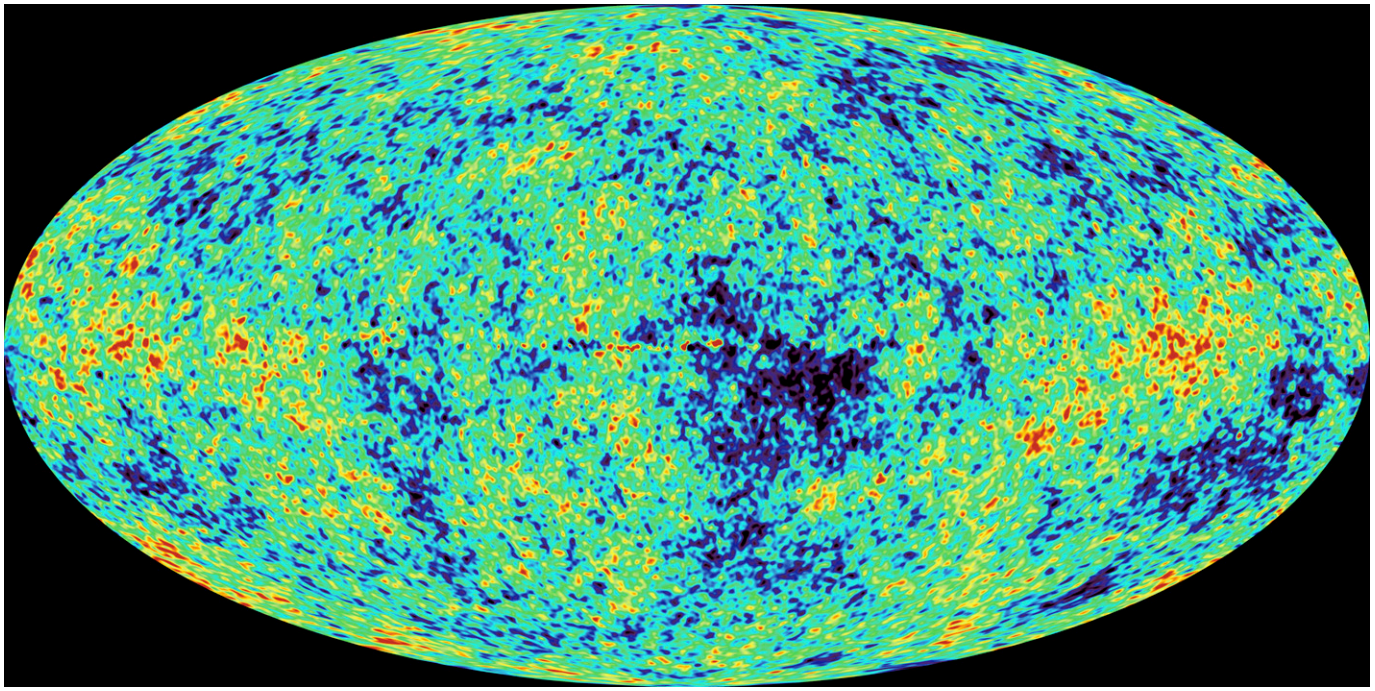
…So what gave rise to this **large-scale structure**?

We've been building fancier instruments too:



WMAP satellite, NASA

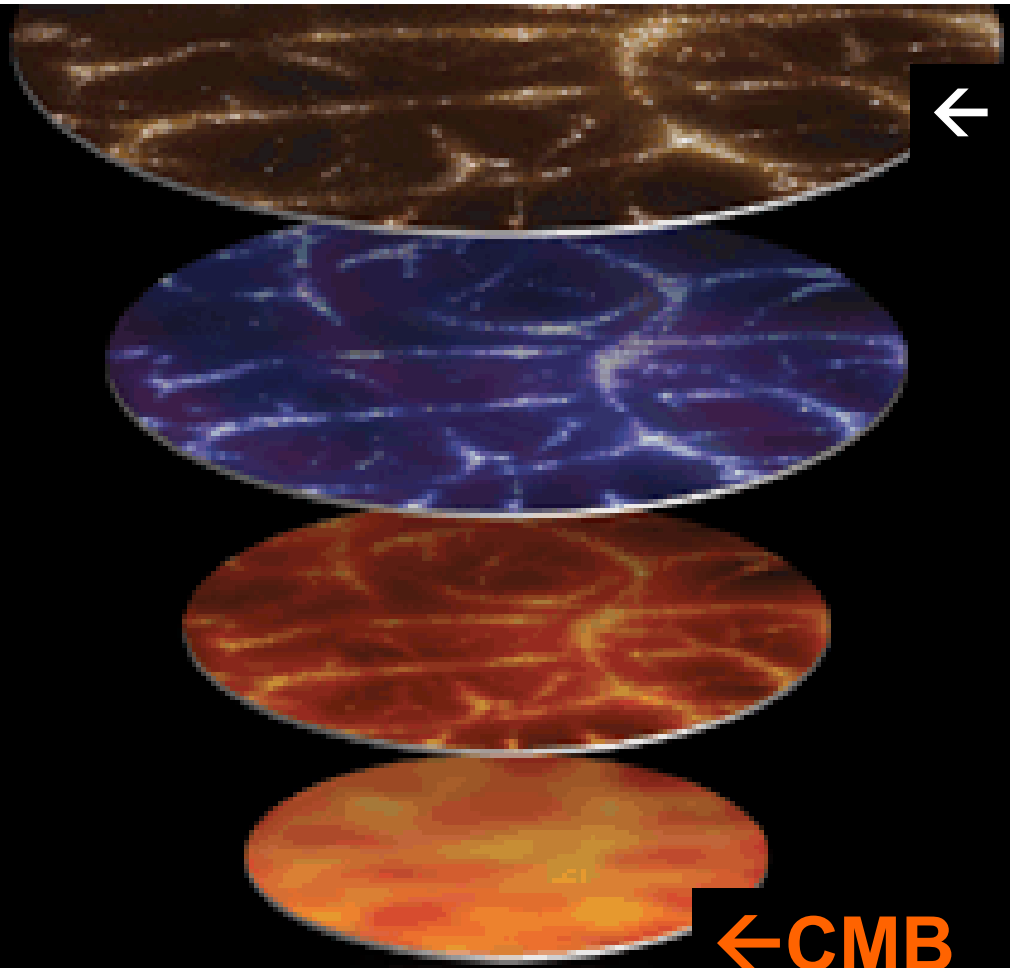
As a result, we also see this:



Cosmic Microwave Background (CMB)

~ 380,000 yrs after the Big Bang

(now = 13B)



← today's structure

←CMB

Breakthrough of th

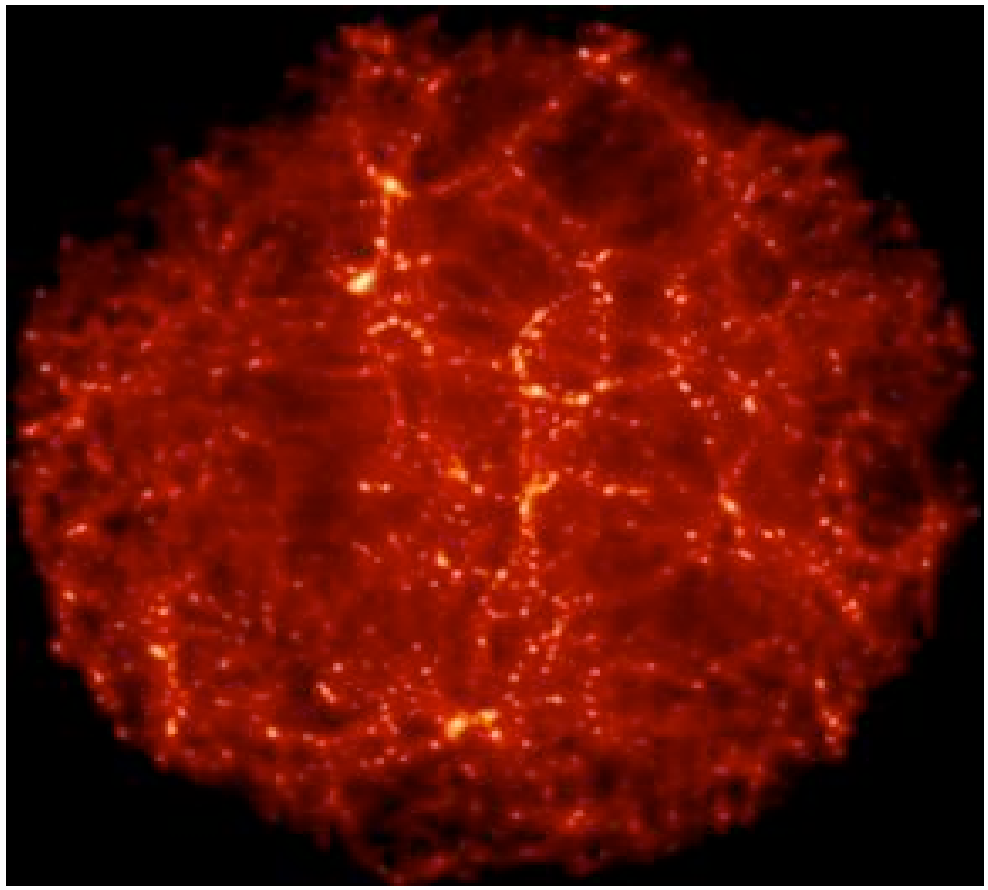
The Big Bang Model =

Theory of gravitation (General Relativity)

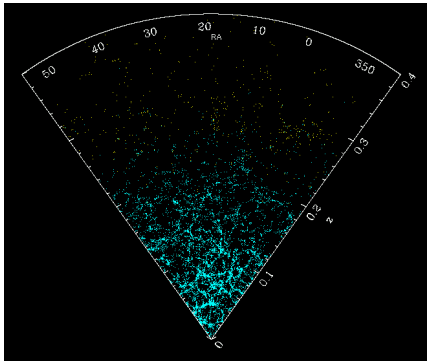
+

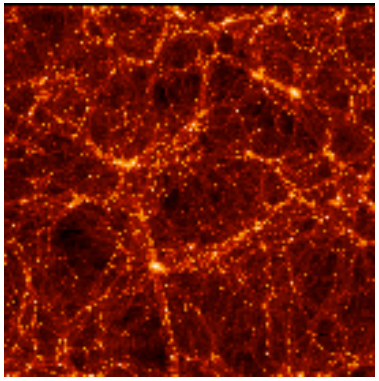
Cosmological Principle (uniformity)

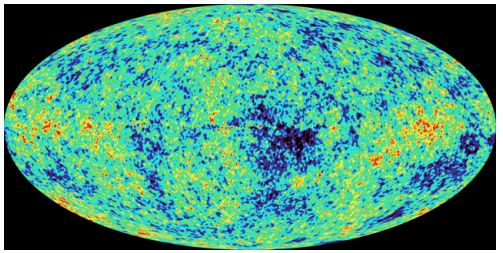
Model $(\Omega, \Lambda, \ldots)$



VIRGO N-body simulation



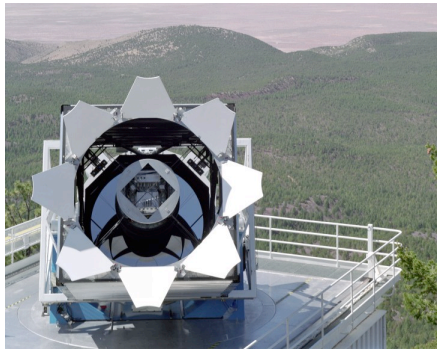


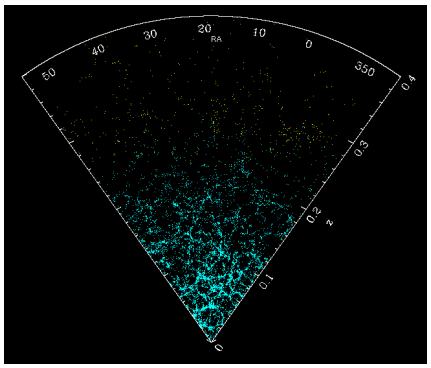


- Clusters?
- Filaments?
- Voids?
- Homogeneity?
- Isotropy?

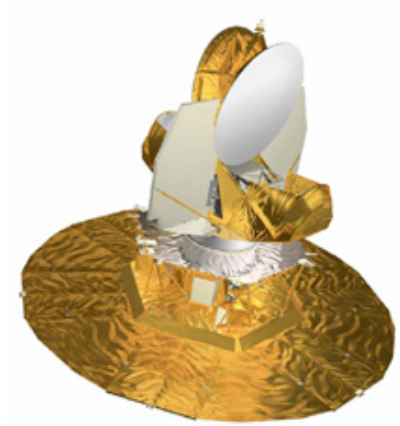
Q: Is the universe Gaussian?

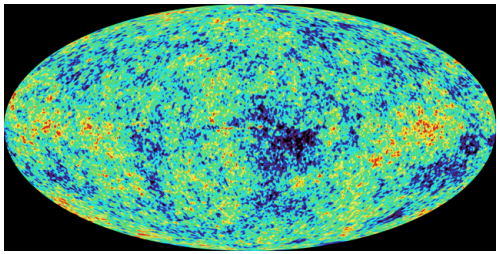
With R.Nichol, D. Wake, G. Kulkarni (CMU Physics)





Gaussianity?

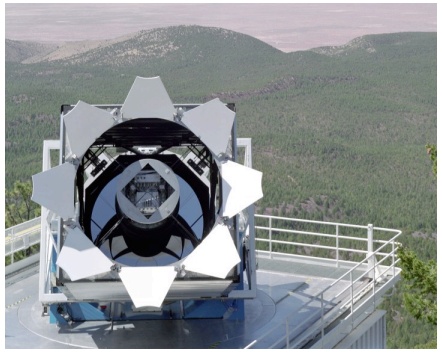


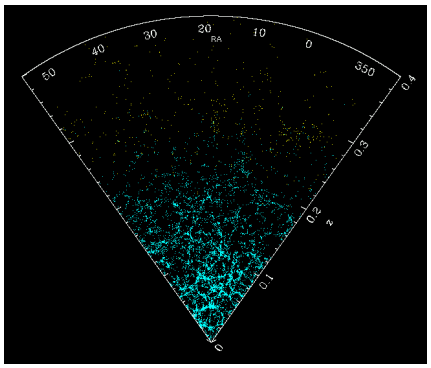


Gaussianity?

Q: Does the Model fit the data?

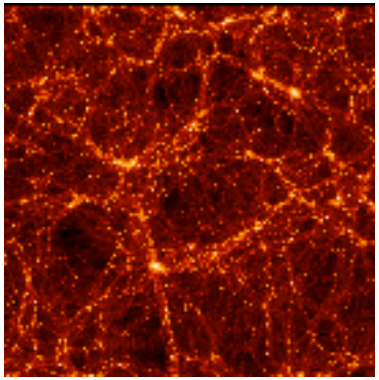
With R.Nichol (CMU Physics), A.Szalay (JHU Physics), I.Szapudi (Hawaii Physics)





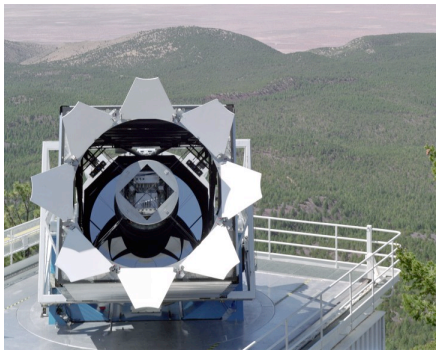
Same distribution?

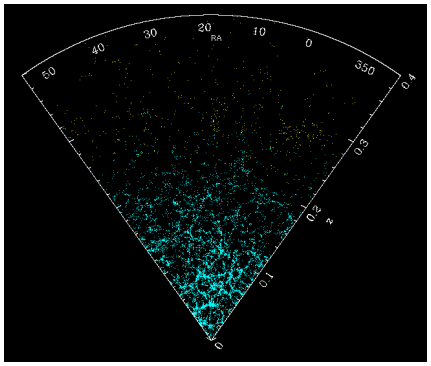
Model $(\Omega, \Lambda, \ldots)$

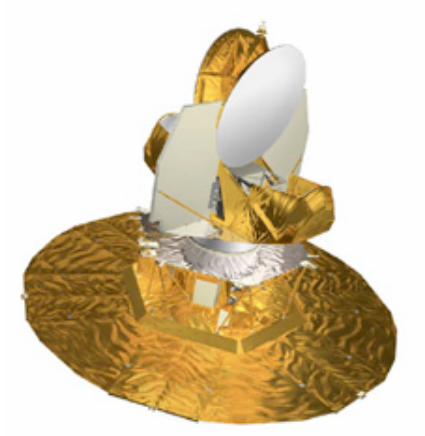


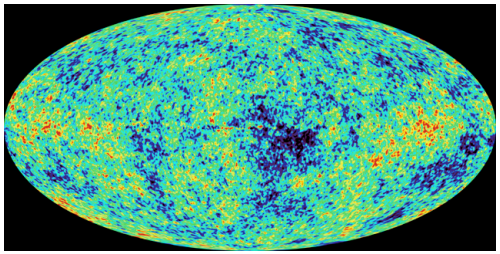
Q: Does dark energy exist?

R. Scranton (Pitt. Physics), A.Connolly (Pitt. Physics), R.Nichol (Physics)









Do we see the ISW Effect?

Outline:

1. Some cosmological questions

2. **Spatial correlations**

3. Divide-and-conquer in real-space

4. Multi-tree algorithm, exact

5. Multi-tree Monte Carlo

What are **pair correlations**, ***n*-point correlations**, etc.?

Synonyms: "Pair correlation function" $g(r)$ =
"Radial distribution function" $g(r)$ =
"2nd-order intensity function" $\lambda_2(r)$ =
"2-point correlation function" $\xi(r)$

Near-synonyms: *K*-function, fractal dimension

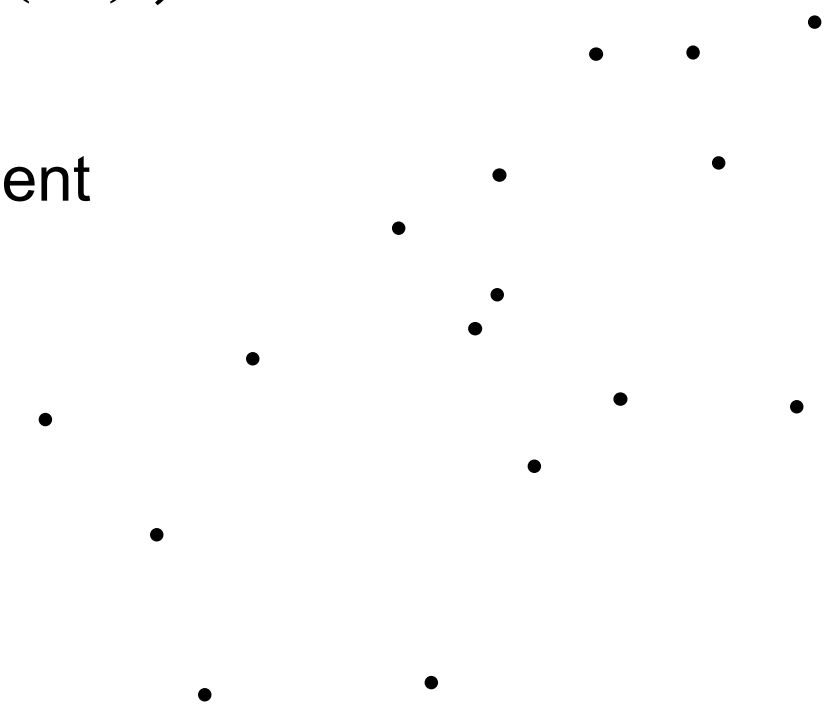
In general: "*n*th-order correlation function" =
"*n*-point correlation function" (*n*pcf)

Poisson point process

$$N(A) \sim Poisson(\lambda V(A))$$

$N(A_1), N(A_2)$ independent

(homogeneous, isotropic)



Moments: $E[N(A)]$

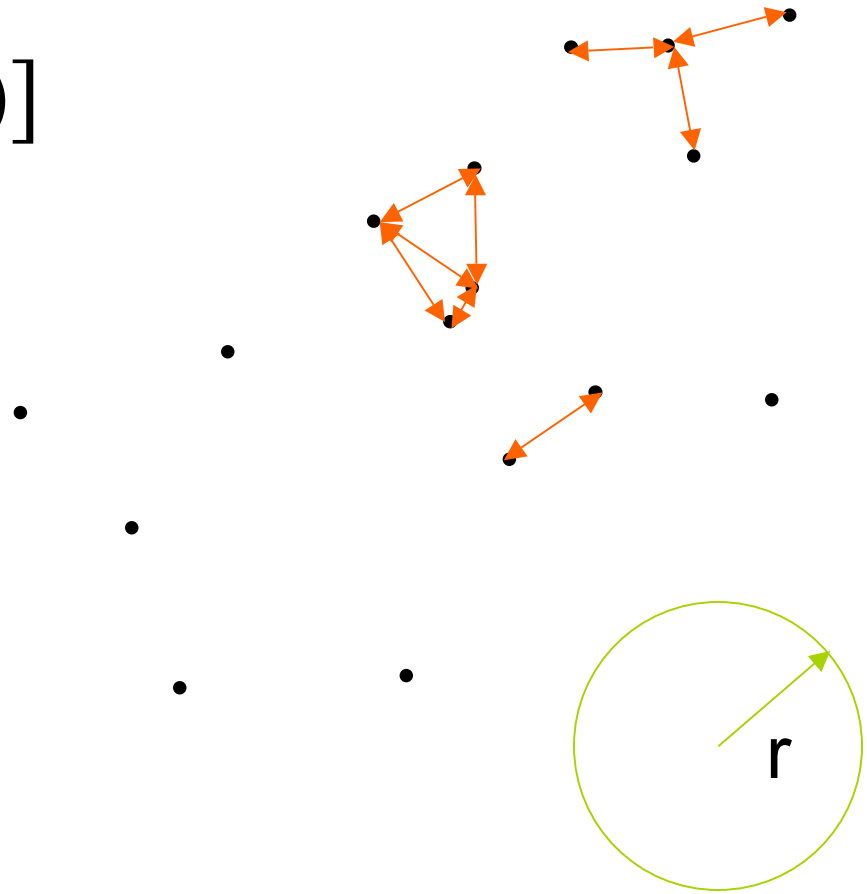
$E[N(A_1)N(A_2)]$

...

2-point correlation

Definition:

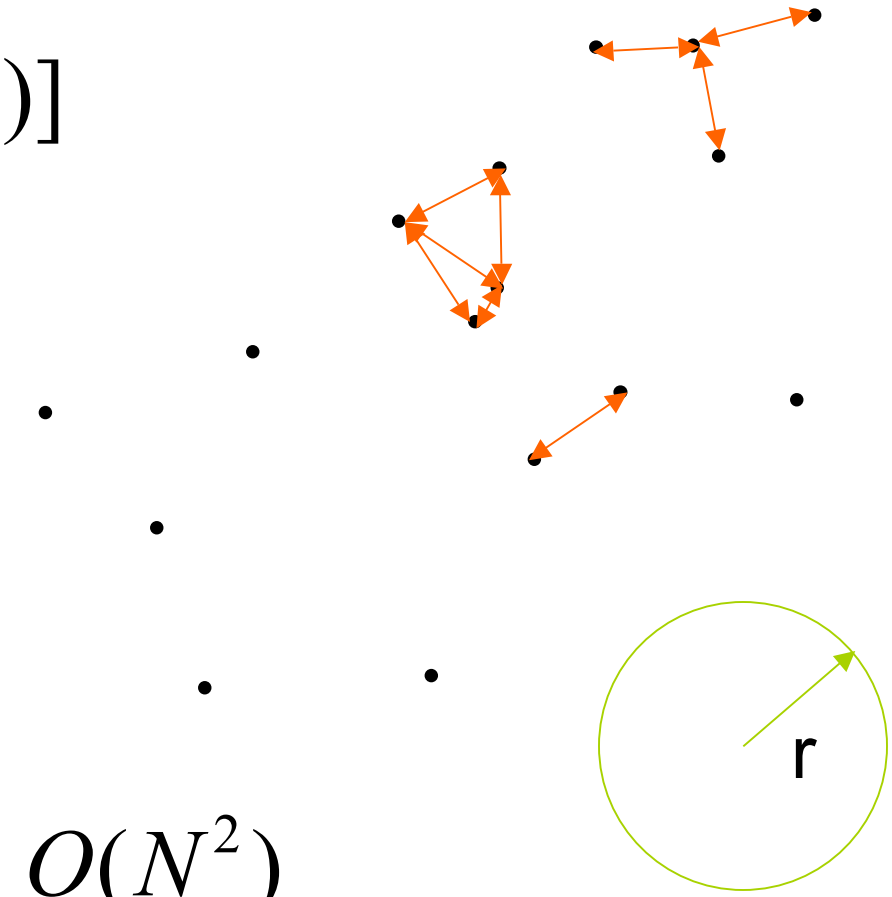
$$dP = \lambda^2 dV_1 dV_2 [1 + \xi(r_{12})]$$



2-point correlation

Definition:

$$dP = \lambda^2 dV_1 dV_2 [1 + \xi(r_{12})]$$

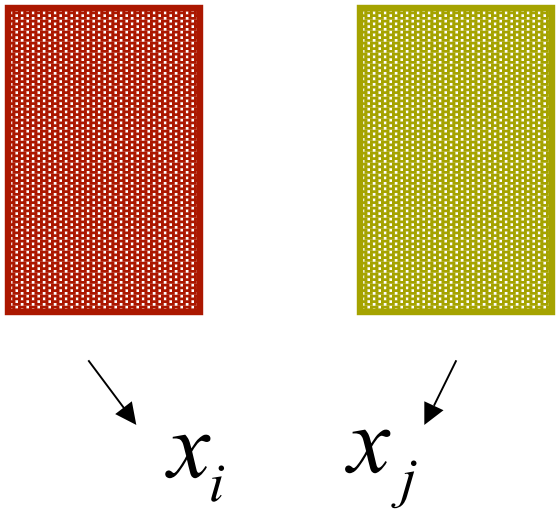


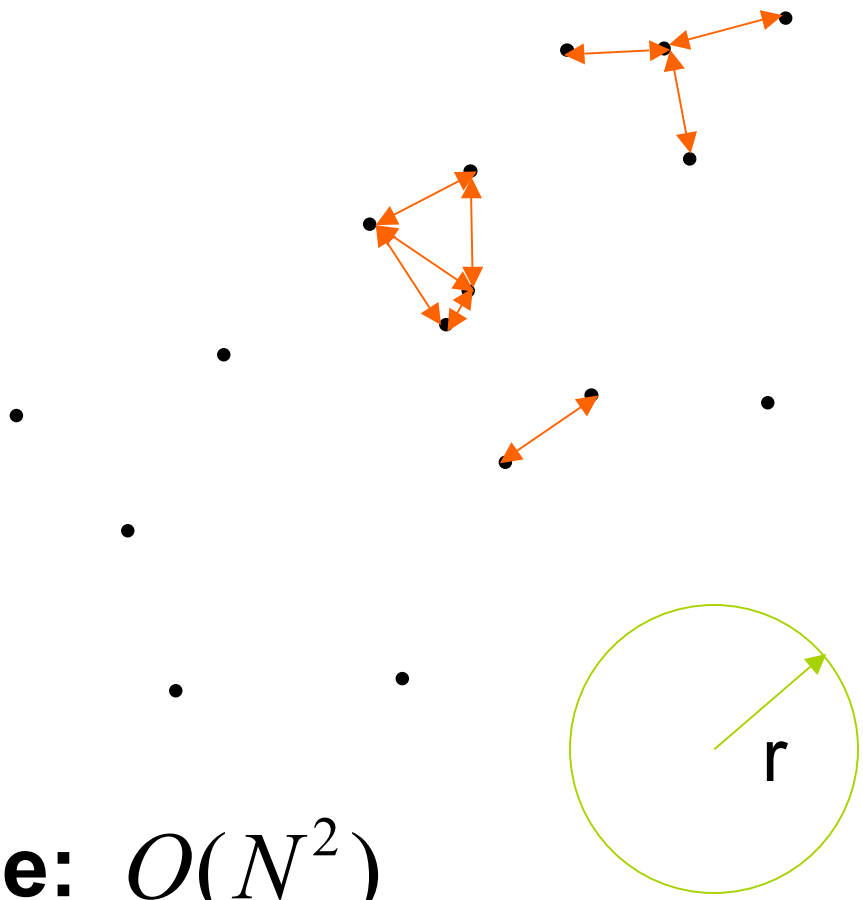
Must compute:

$$\sum_i^N \sum_{j \neq i}^N I(\|x_i - x_j\| < r)$$

➡ **Naïve cost:** $O(N^2)$

2-point cross-correlation



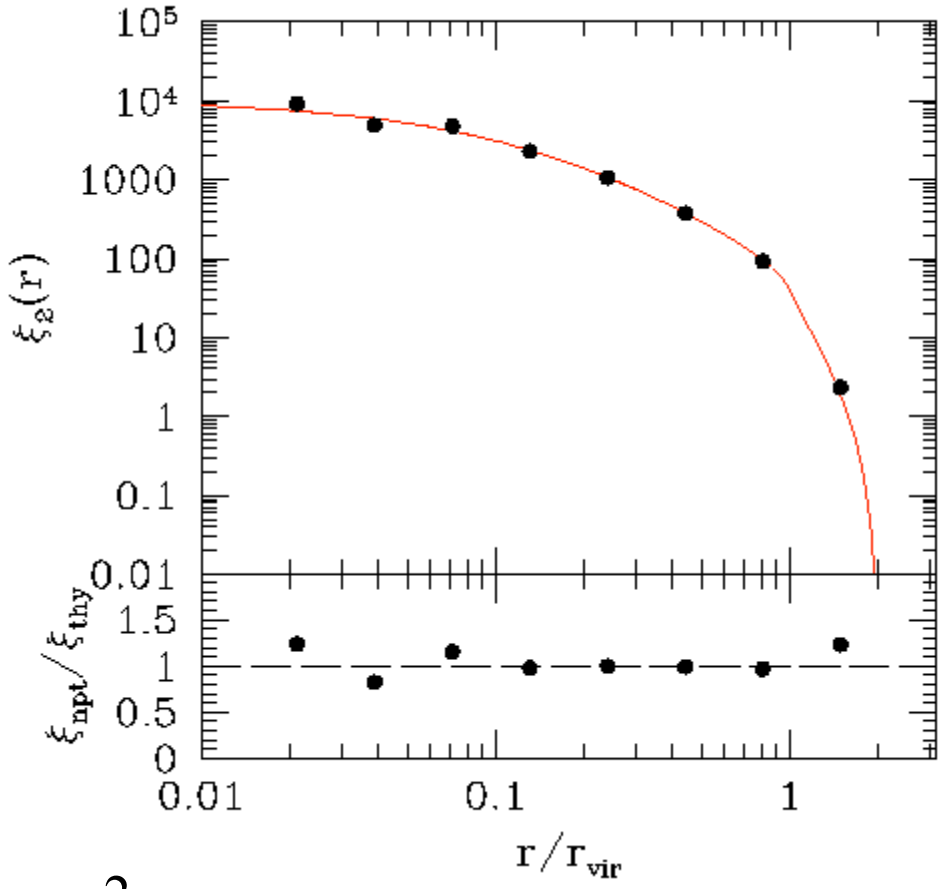


$$\sum_{i}^{N}\sum_{j\neq i}^{N} I(\|x_i - x_j\| < r)$$

➡ **Main obstacle:** $O(N^2)$

2-point correlation **function**

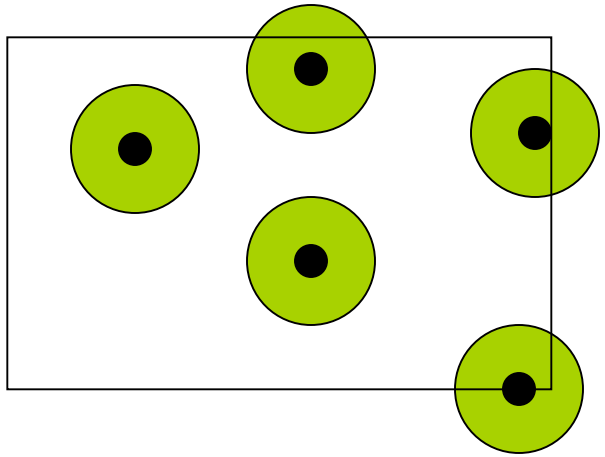
Must compute the 2-point correlation for *B* different values of *r*



➡ **Naïve cost:** $O(BN^2)$

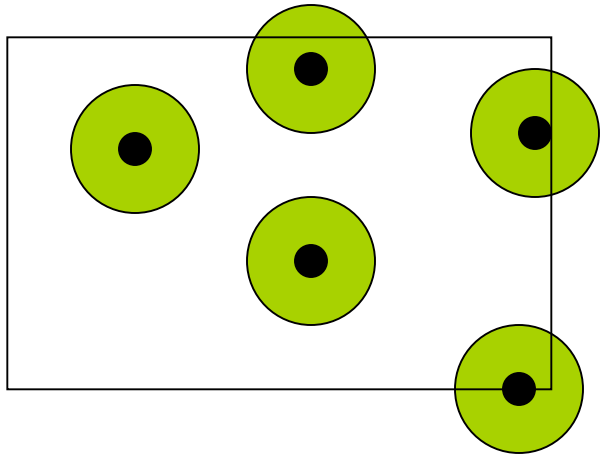
2-point correlation **estimator**

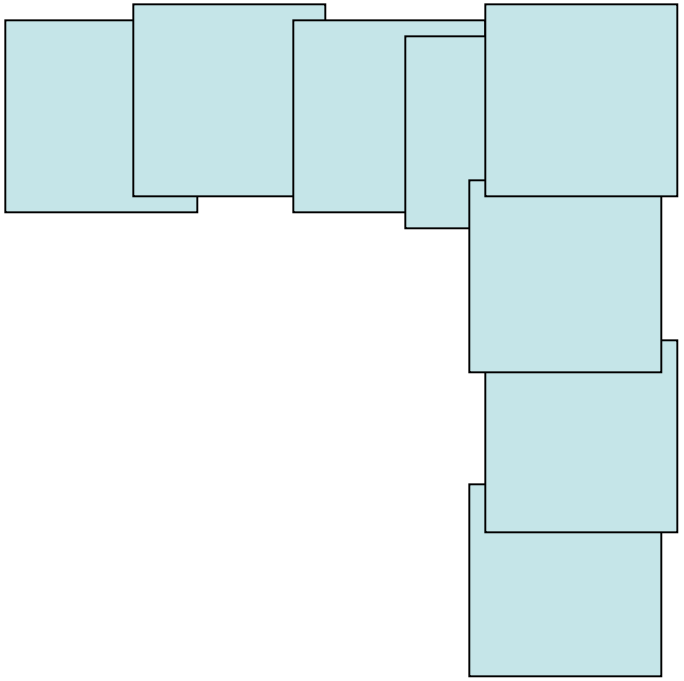
Major issue: *Must handle **edge effects.***



2-point correlation **estimator**

Major issue: *Must handle **edge effects.***





2-point correlation **estimator**

Major issue: *Must handle* ***edge effects.***

Landy-Szalay 2pcf estimator (1993):

$$\hat{\xi} = \frac{DD - 2DR + RR}{RR}$$

Cross-correlate with huge ‘random’ point set

2-point correlation **estimator**

Major issue: *Must handle* ***edge effects.***

Landy-Szalay 2pcf estimator (1993):

$$\hat{\xi} = \frac{DD - 2DR + RR}{RR}$$

1. Random set should be big
2. Do this many times!

2-point correlation **estimator**

Major issue: *Must handle* ***edge effects.***

Landy-Szalay 2pcf estimator (1993):

$$\hat{\xi} = \frac{DD - 2DR + RR}{RR}$$

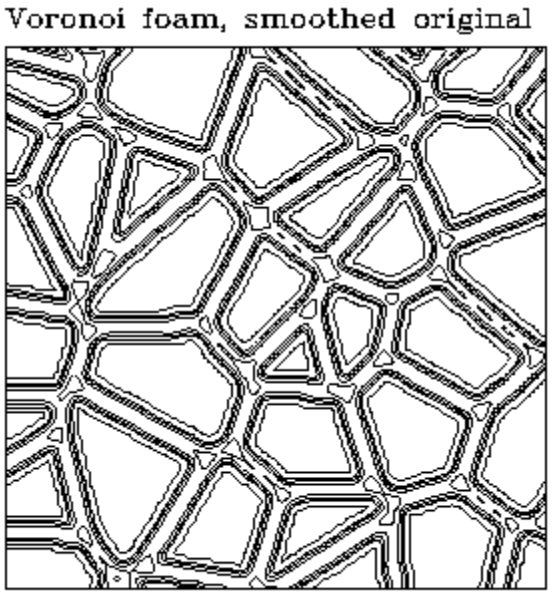
1. Random set should be big
2. Do this many times!

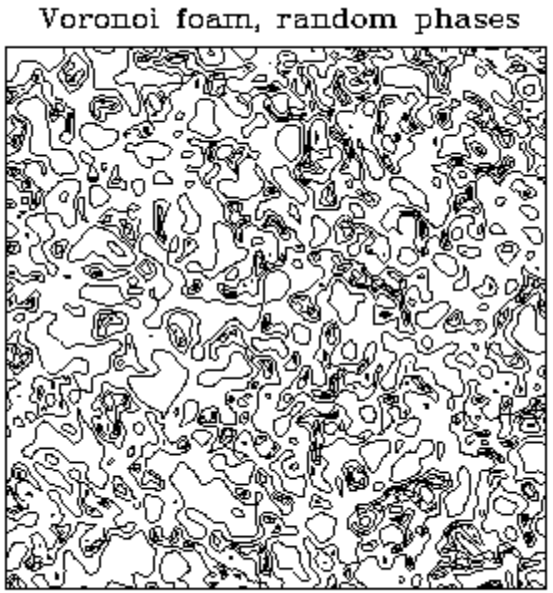
Variance estimate: Bootstrap.

n=2 is not enough

- Whole distribution: need n=1 to infinity
- [White 1979]: Distribution depends significantly on the higher-order terms
- Variance of lower-order terms is given by higher-order terms
- *Important:* Standard model: n>2 terms are 0

n=2 is not enough



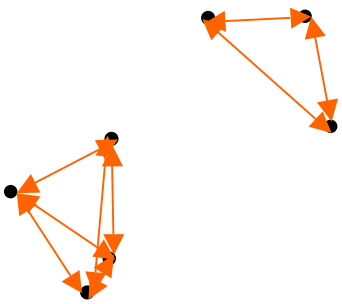


We need at least the 3-point to distinguish these…

3-point correlation function

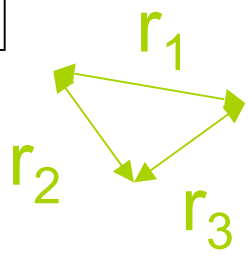
Def: $dP = \lambda^3 dV_1 dV_2 dV_3 \cdot$

$$[1 + \xi(r_{12}) + \xi(r_{23}) + \xi(r_{13}) + \zeta(r_{12}, r_{23}, r_{13})]$$



Must compute:

$$\sum_{i}^{N} \sum_{j \neq i}^{N} \sum_{k \neq j \neq i}^{N} I(\delta_{ij} < r_1) I(\delta_{jk} < r_2) I(\delta_{ki} < r_3)$$



➡ **Main obstacle:** $O(N^3)$

n-point correlation **estimator**

Szapudi-Szalay *n*pcf estimator (1997):

$$\hat{\xi}_n = \frac{(D_1 - R_1)(D_2 - R_2)\ldots(D_n - R_n)}{R_1 R_2 \ldots R_n}$$

What are **pair correlations**, ***n*-point correlations**, etc.?

Spatial statistics for characterizing distributions of points in space, e.g.:

- clusters of points [White 1979]
- nearest-neighbor distances, counts in cells [Peebles 1980]
- filaments [Fry 1986]
- voids [BF 1996]

Points can be atoms, molecules, galaxies…
Point sets can be from real or simulated data…

→ used throughout physics.

Foundation for theory of point processes [Daley,Vere-Jones 1972], [Ripley 1976].

Challenges

- Arbitrary point distributions
- Huge N
- Handle edges correctly
- General n
- B different scales
- Accuracy
  - exact if possible
  - if not, give error bounds

Main approach: Fourier space

Power spectrum: Fourier transform of 2pcf

[Peebles & Groth 1976]

Discrete Fourier transform:

- Exact if points lie on a grid (e.g. image) → approximate otherwise ('window function')

Multidimensional FFT:

- Replace N with M^D → $O(M^D \log M^D)$
- Error bounds (?)

Main approach: Fourier space

Higher moments?

- n=3? *Bispectrum* → O(N(NlogN))

  e.g. [Pen et al. 2003]
- n>3? *Trispectrum*, etc. → good luck…

Tractability/naturalness?

- For Gaussian model, good fit
- Hard to modify/extend in general

Edges? Fourier representation problematic.

What this talk is about

Let's go back in time…

Divide-and-conquer in Fourier space.

[Cooley-Tukey 1968]

→ [Peebles-Groth 1976] *2pcf feasible* (sort of)

Divide-and-conquer in real space.

[Bentley 1975], [Greengard-Rokhlin 1987]

→ 2004… *npcf feasible*??

Outline:

1. Some cosmological questions

2. Spatial correlations

3. **Divide-and-conquer in real-space**

4. Multi-tree algorithm, exact

5. Multi-tree Monte Carlo

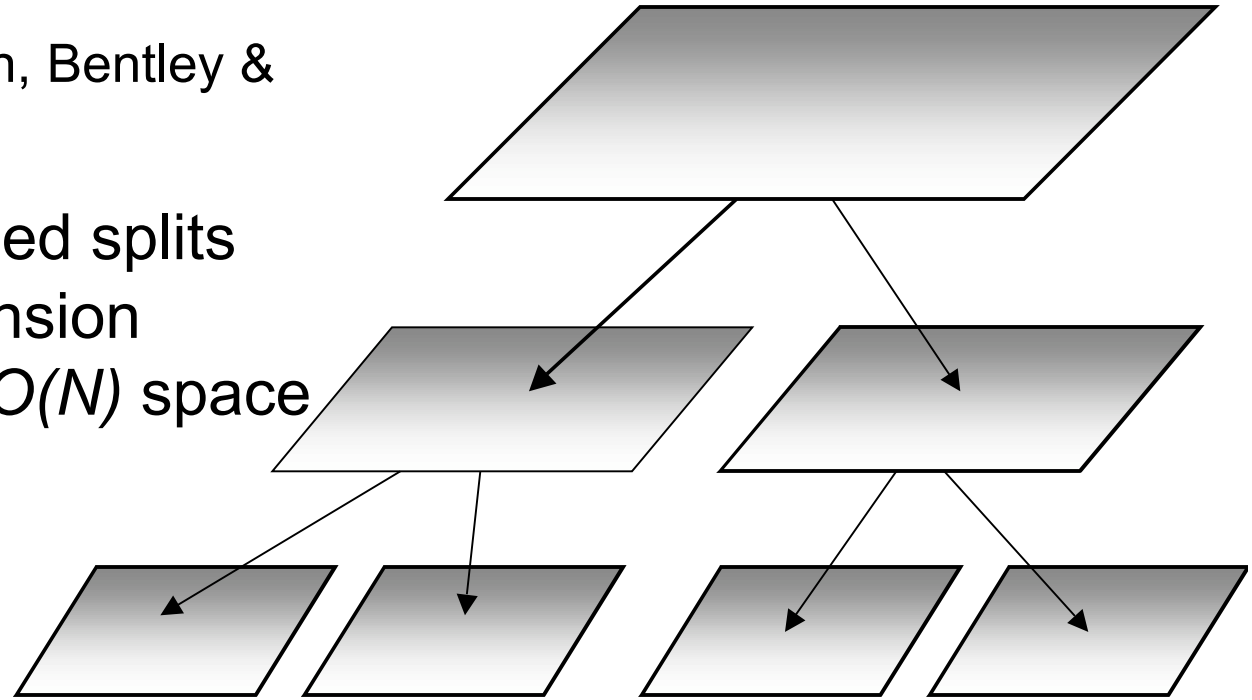
*k*d-trees:

most widely-used space-partitioning tree

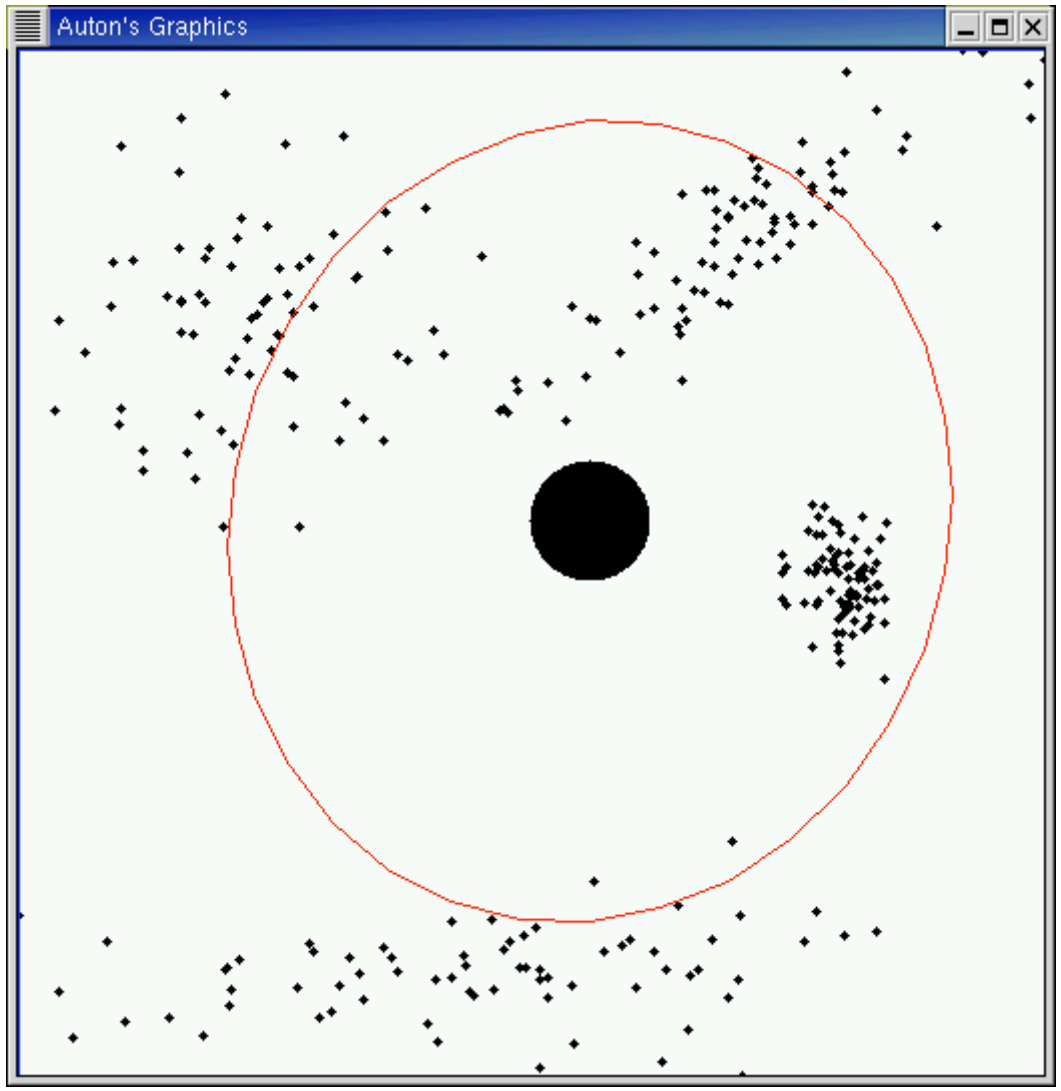
(computational geometry)

[Bentley 1975, Friedman, Bentley & Finkel 1977]

- Univariate axis-aligned splits
- Split on widest dimension
- *O(N log N)* to build, *O(N)* space



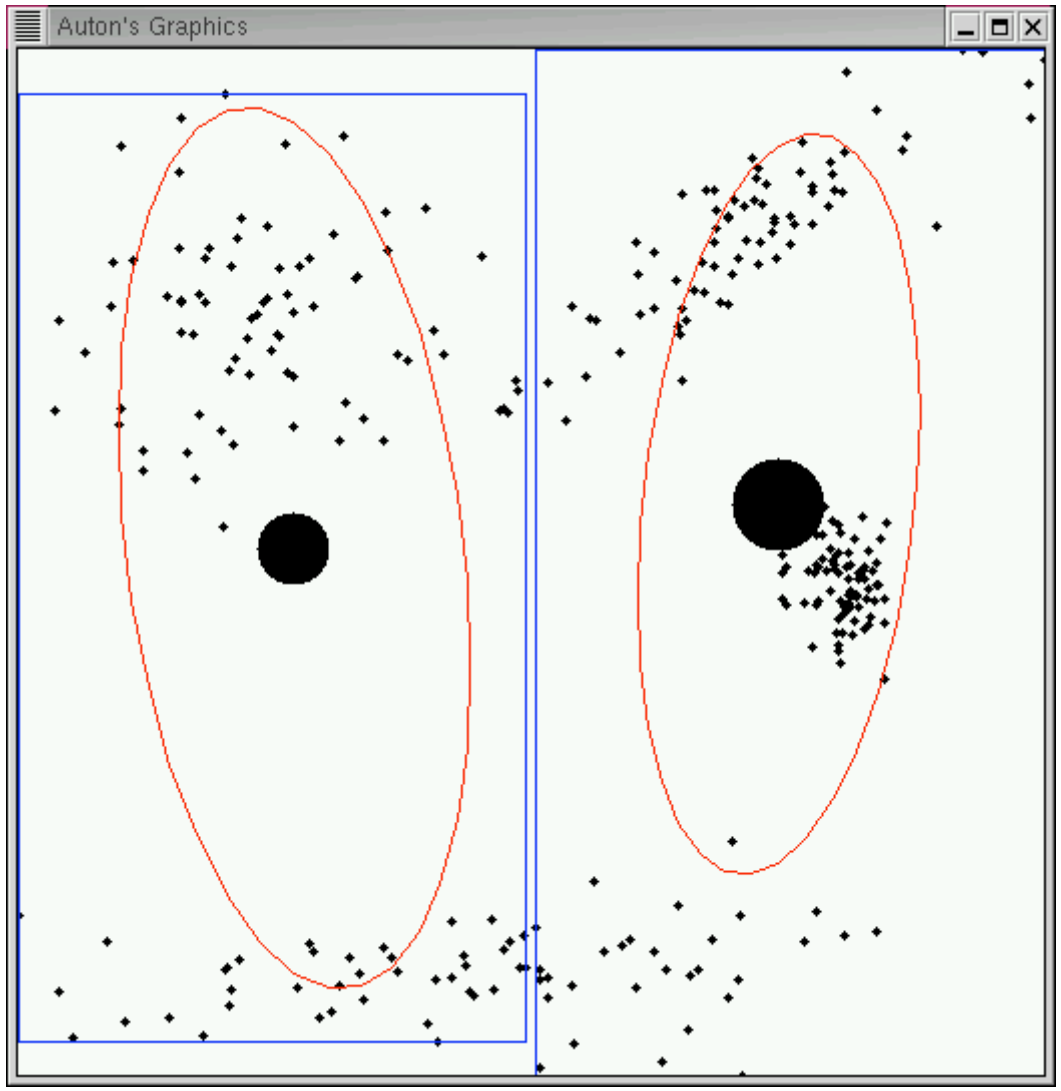
A *k*d-tree: level 1



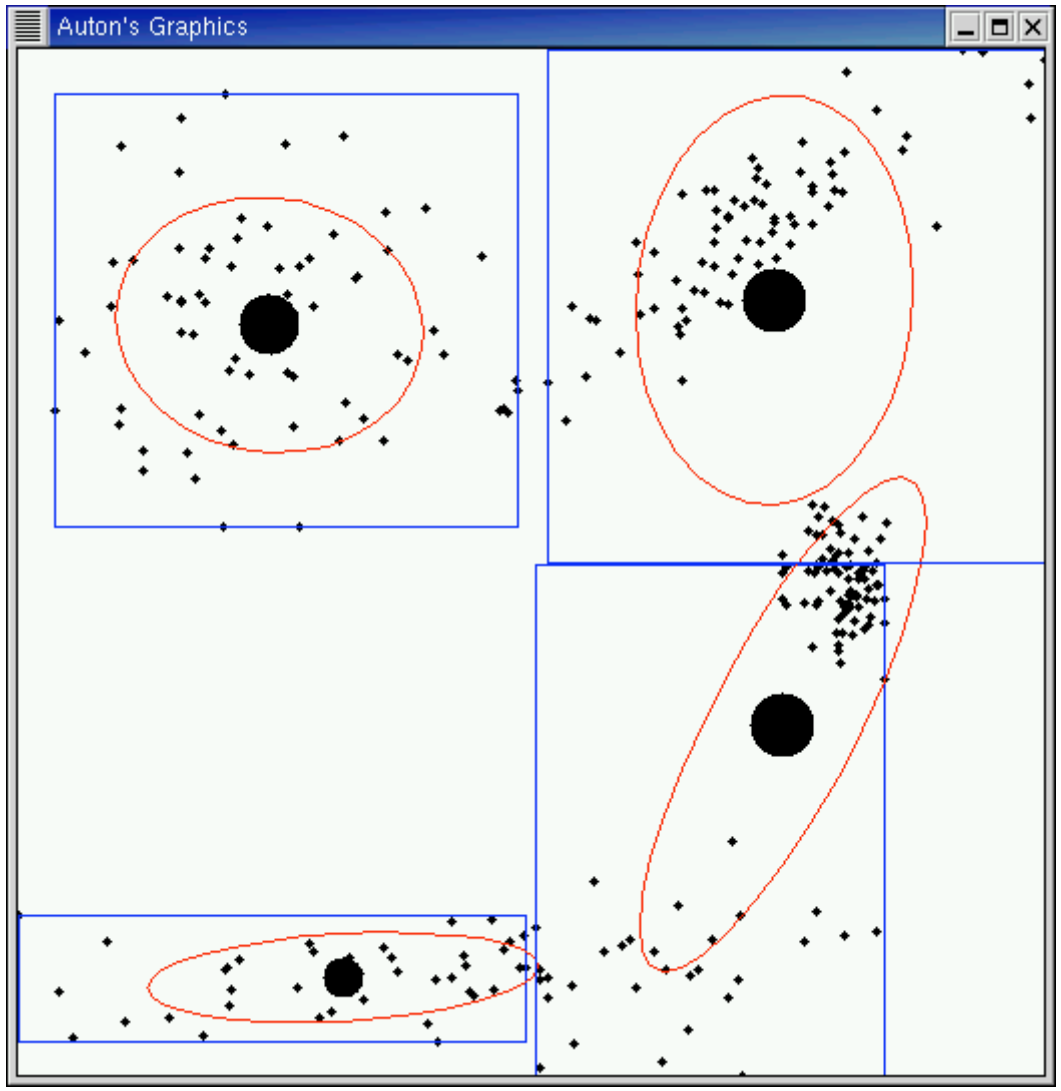
Add: **‘Cached sufficient statistics’**

[Moore and Lee, 1995]

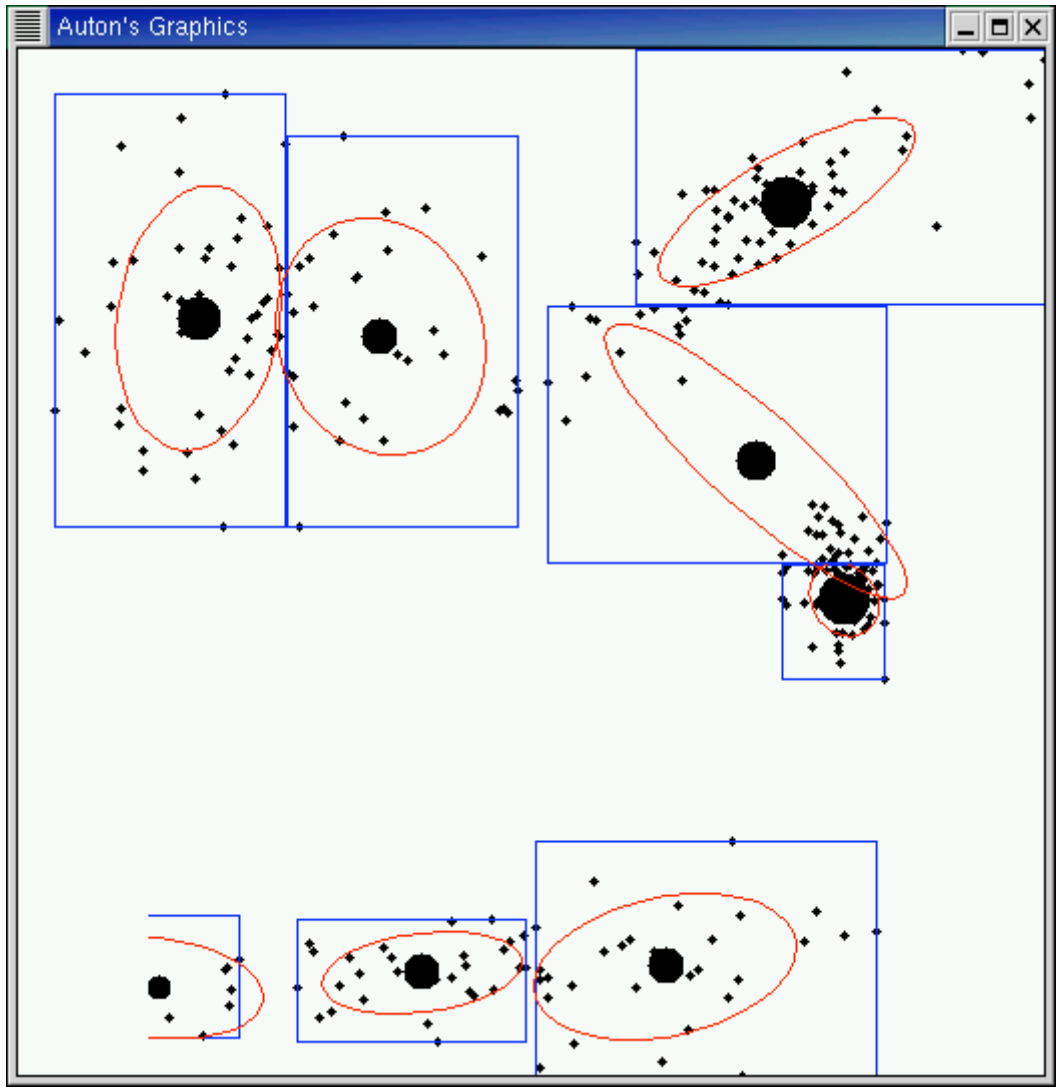
A *k*d-tree: level 2



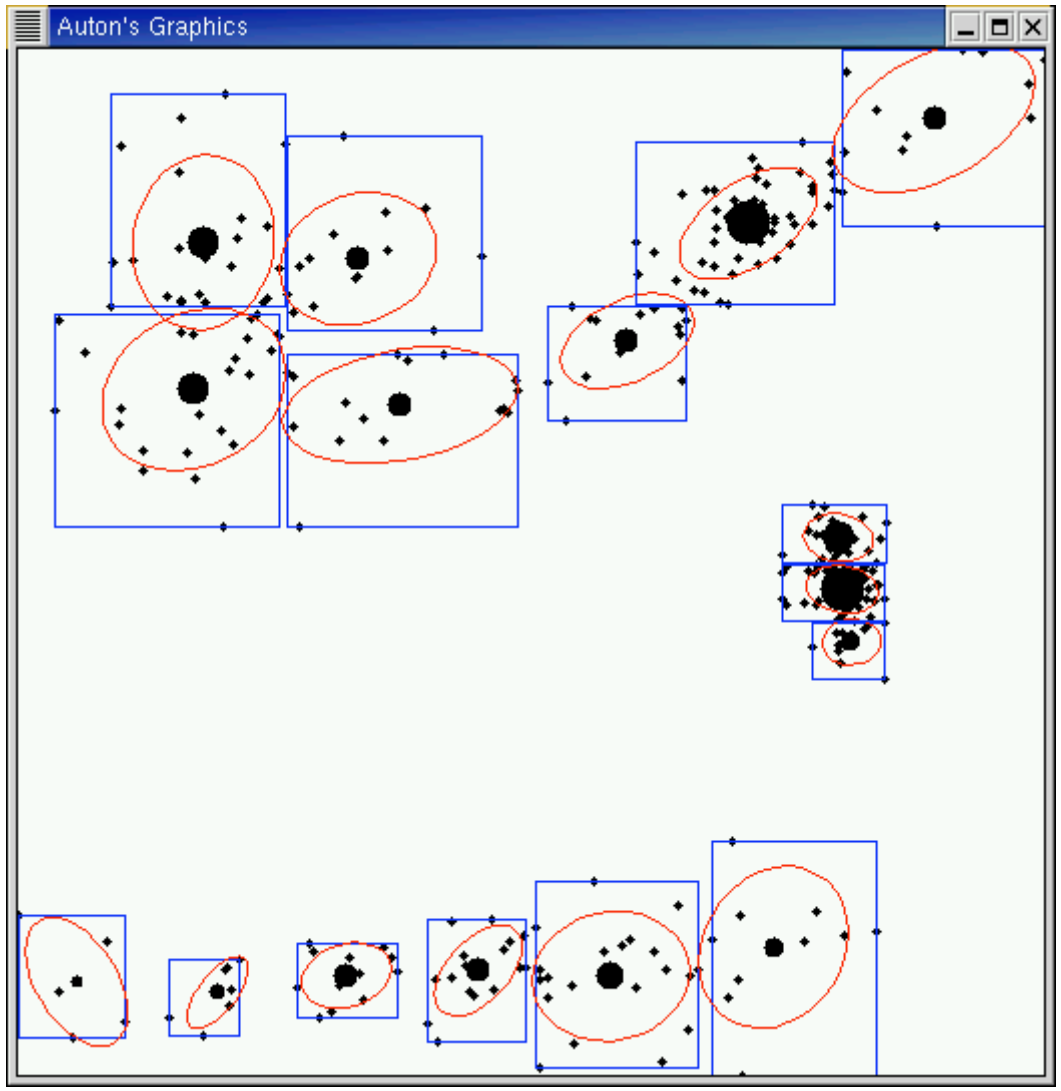
A *k*d-tree: level 3



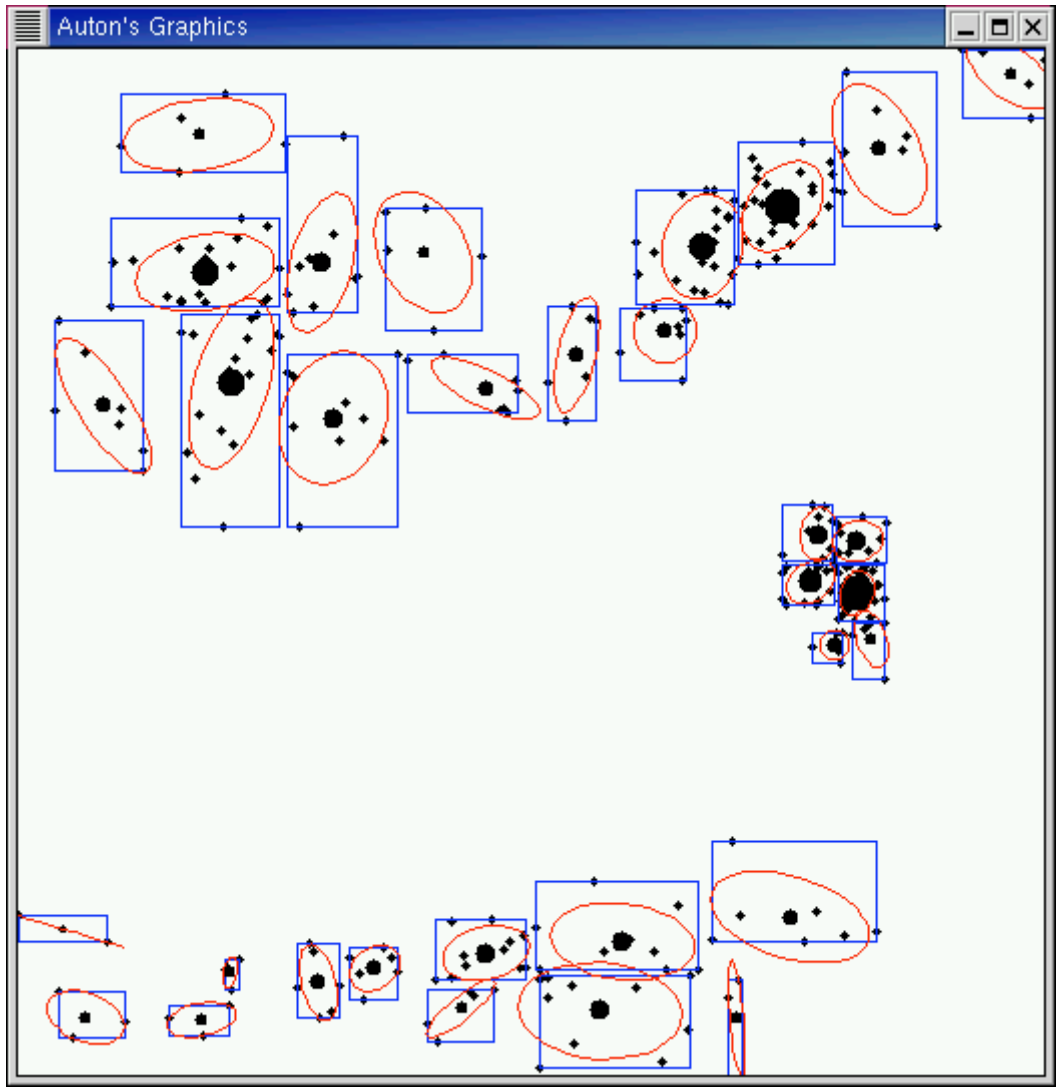
A *k*d-tree: level 4



A *k*d-tree: level 5



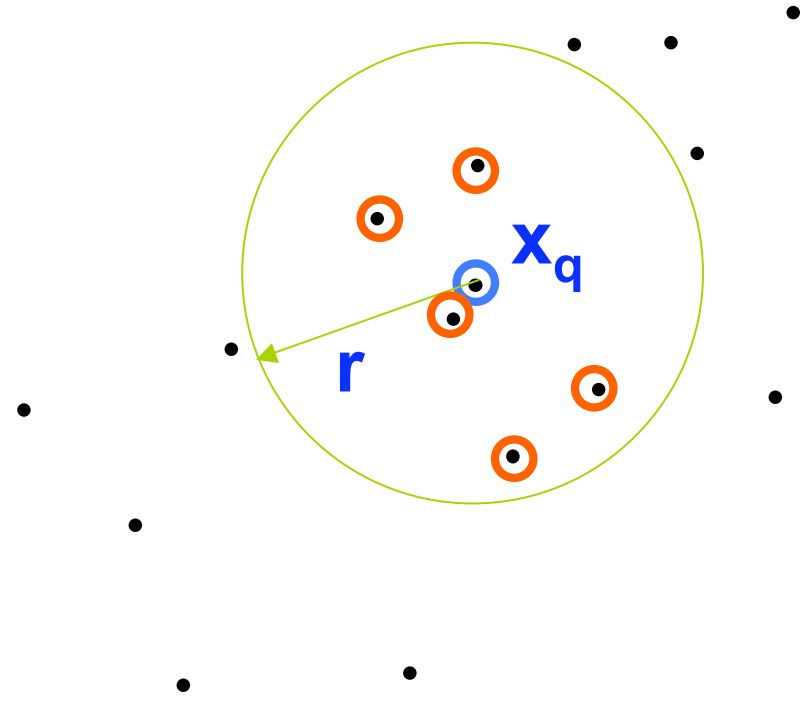
A *k*d-tree: level 6



Effectively, an adaptive hierarchical grid

1-pcf (using late-70's ideas + 1 trick)

aka Range-counting, or spherical-kernel KDE

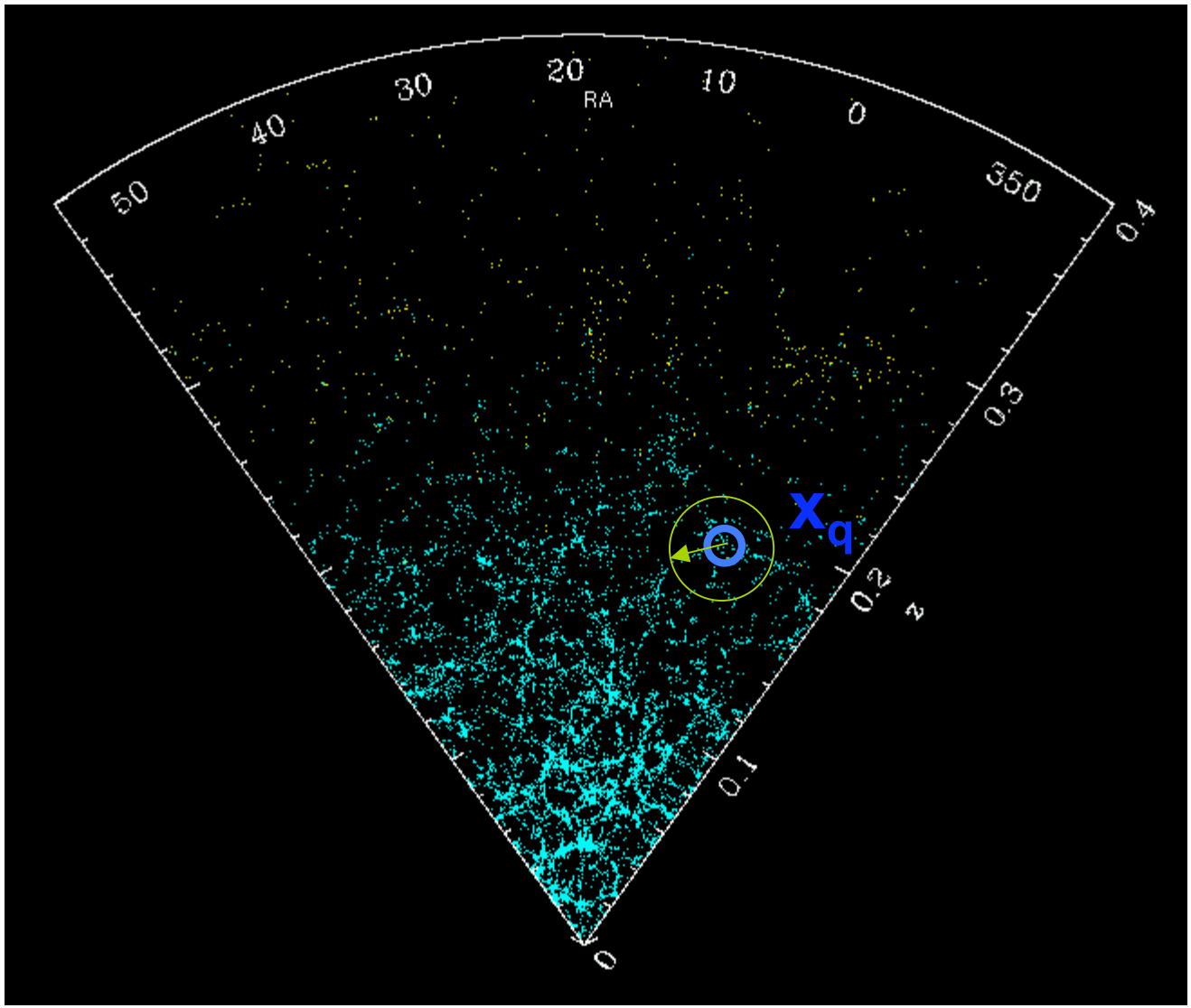


answer:

$c(x_q) = 5$

$$c(x_q) = \sum_{i \neq q}^{N} I(\|x_q - x_i\| < r)$$

Remember, the real dataset is big:

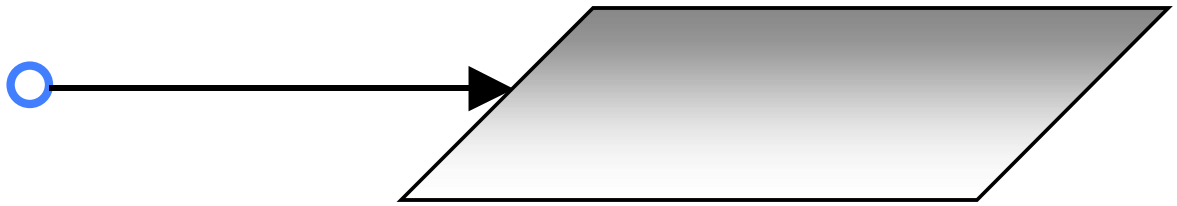


$$O(N) \longrightarrow ?$$

Exclusion and inclusion

using point-node *kd*-tree bounds

O(D) bounds on distance minima/maxima:



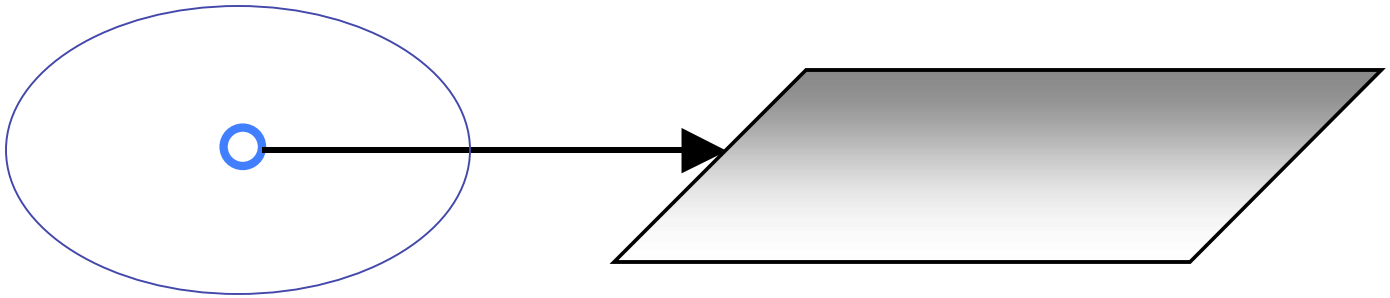
$$\min_i \|x - x_i\| \geq \sum_d^D \left[\max\left\{ (l_d - x_d)^2, 0 \right\} + \max\left\{ (x_d - u_d)^2, 0 \right\} \right]$$

$$\max_i \|x - x_i\| \leq \sum_d^D \max\left\{ (u_d - x_d)^2, (x_d - l_d)^2 \right\}$$

Exclusion and inclusion

using point-node *kd*-tree bounds

O(D) bounds on distance minima/maxima:



$$\min_i \|x - x_i\| \geq \sum_d^D \left[\max\left\{ (l_d - x_d)^2, 0 \right\} + \max\left\{ (x_d - u_d)^2, 0 \right\} \right]$$

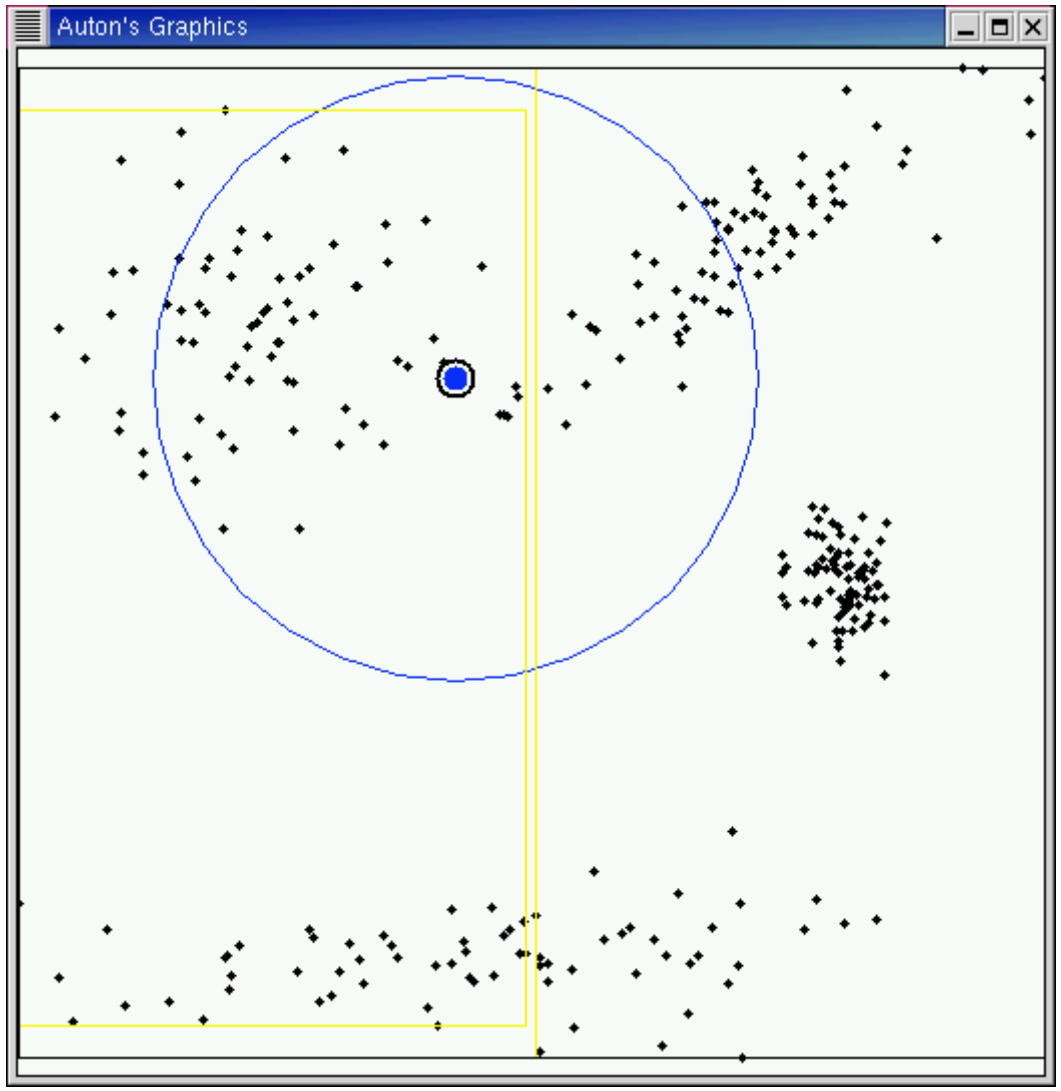
$$\max_i \|x - x_i\| \leq \sum_d^D \max\left\{ (u_d - x_d)^2, (x_d - l_d)^2 \right\}$$

Divide-and-conquer #0

‘Single-tree’ algorithm

[Bentley et al., 70’s], many others

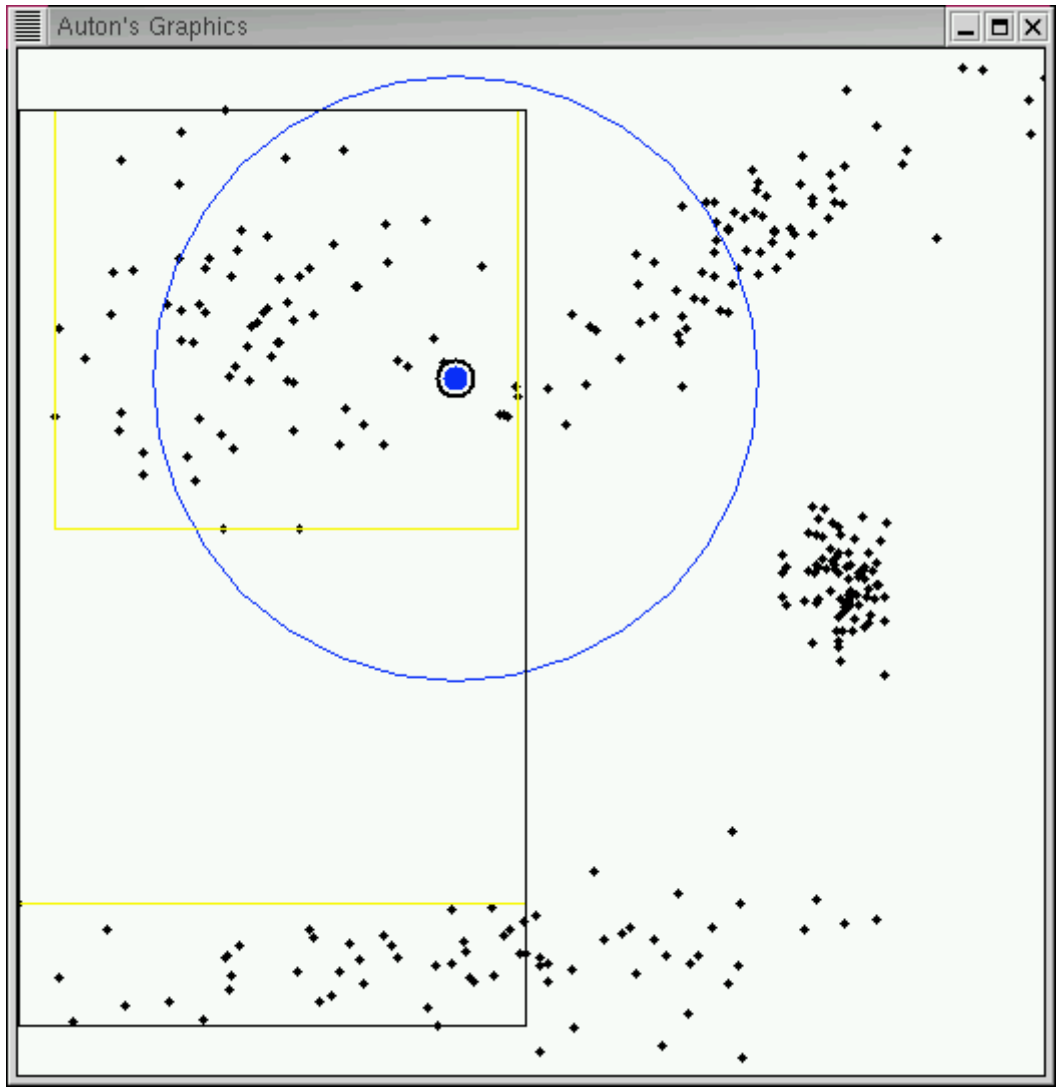
Range-count example



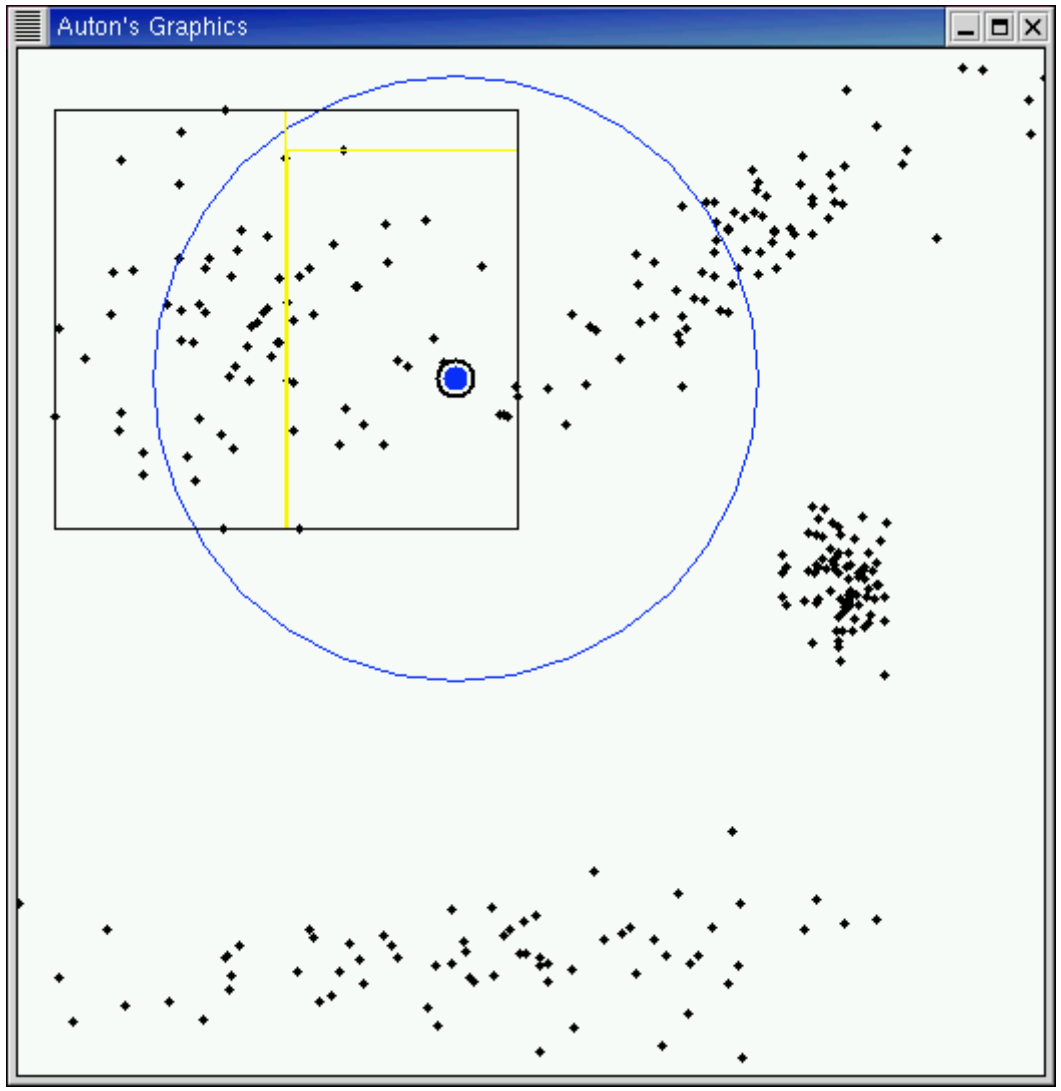
Recursive range-count algorithm

[folk algorithm]

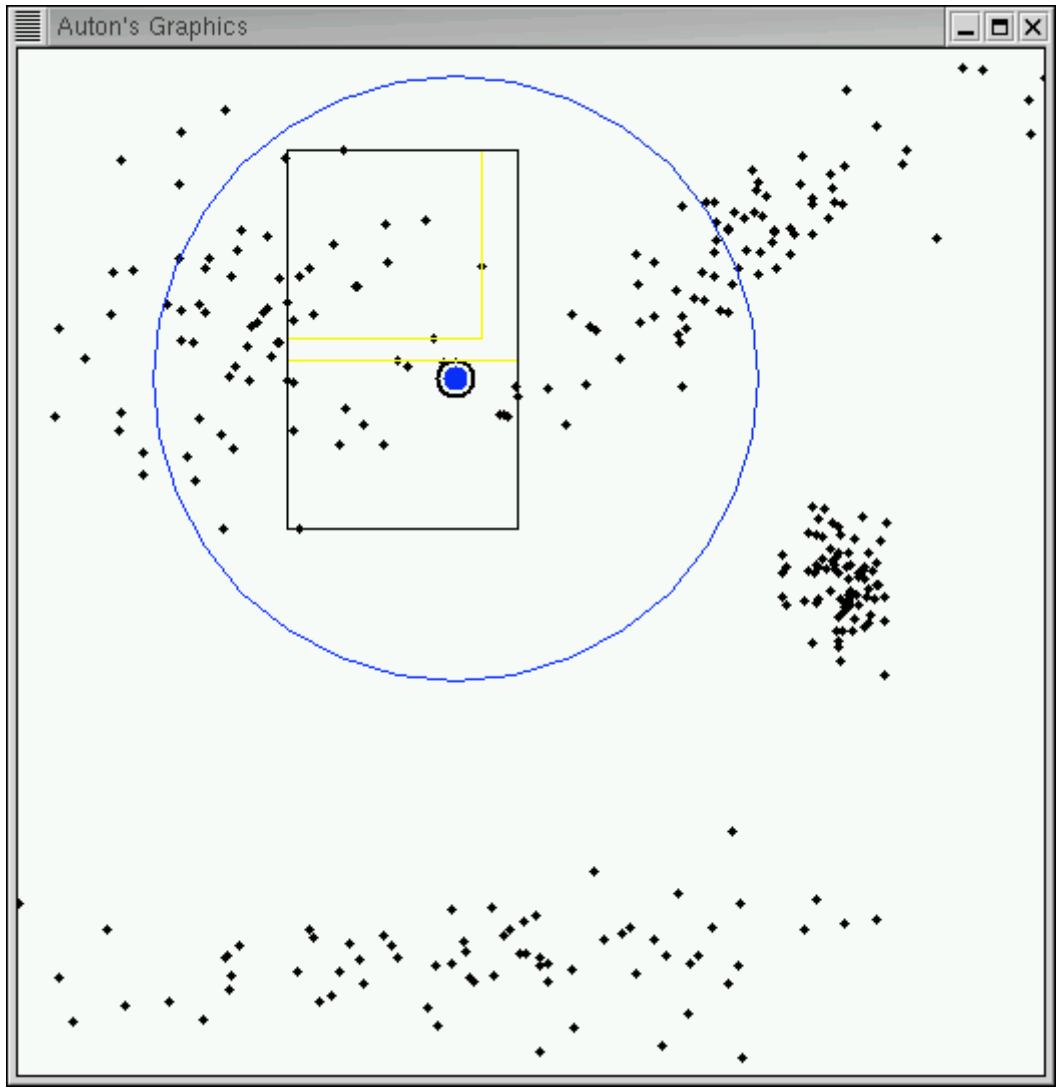
Range-count example



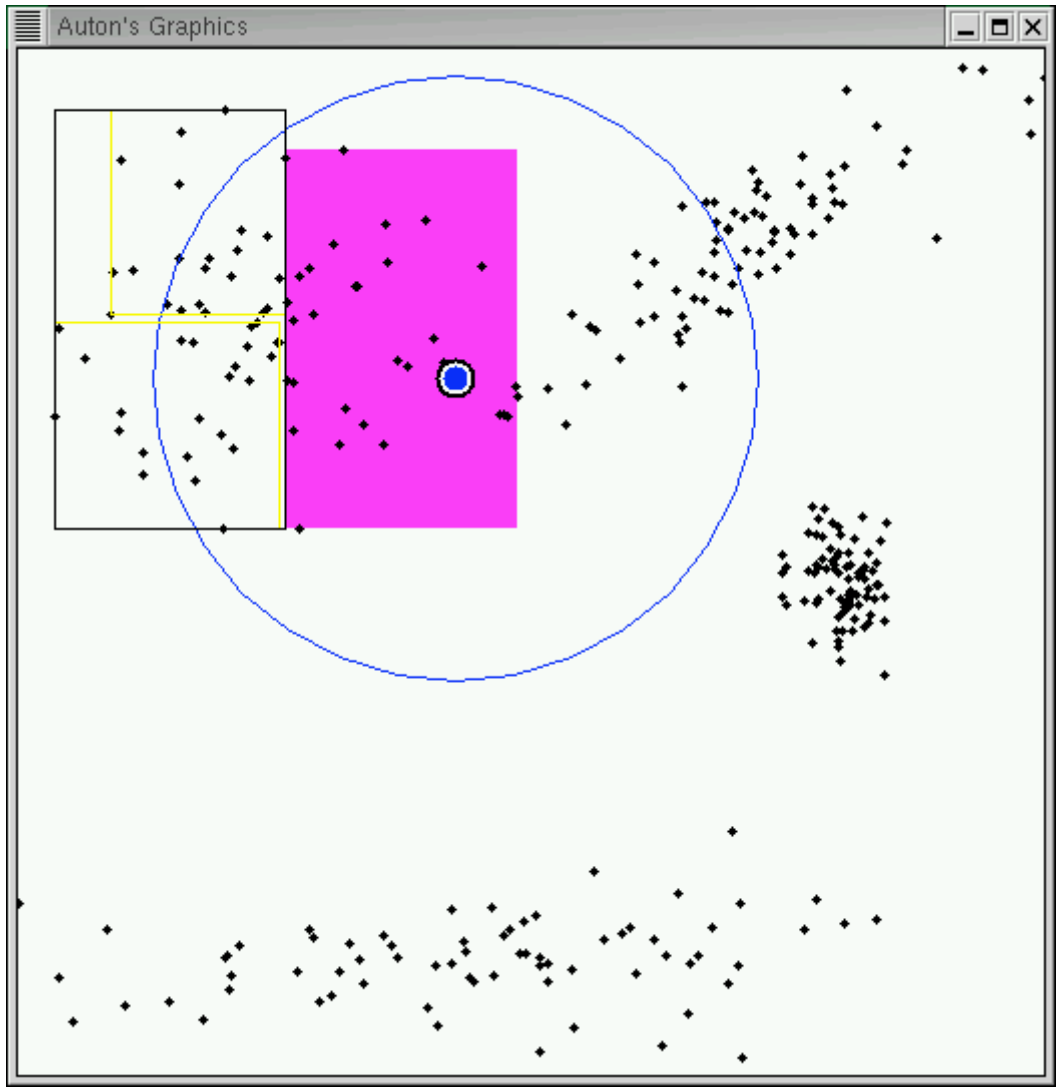
Range-count example



Range-count example

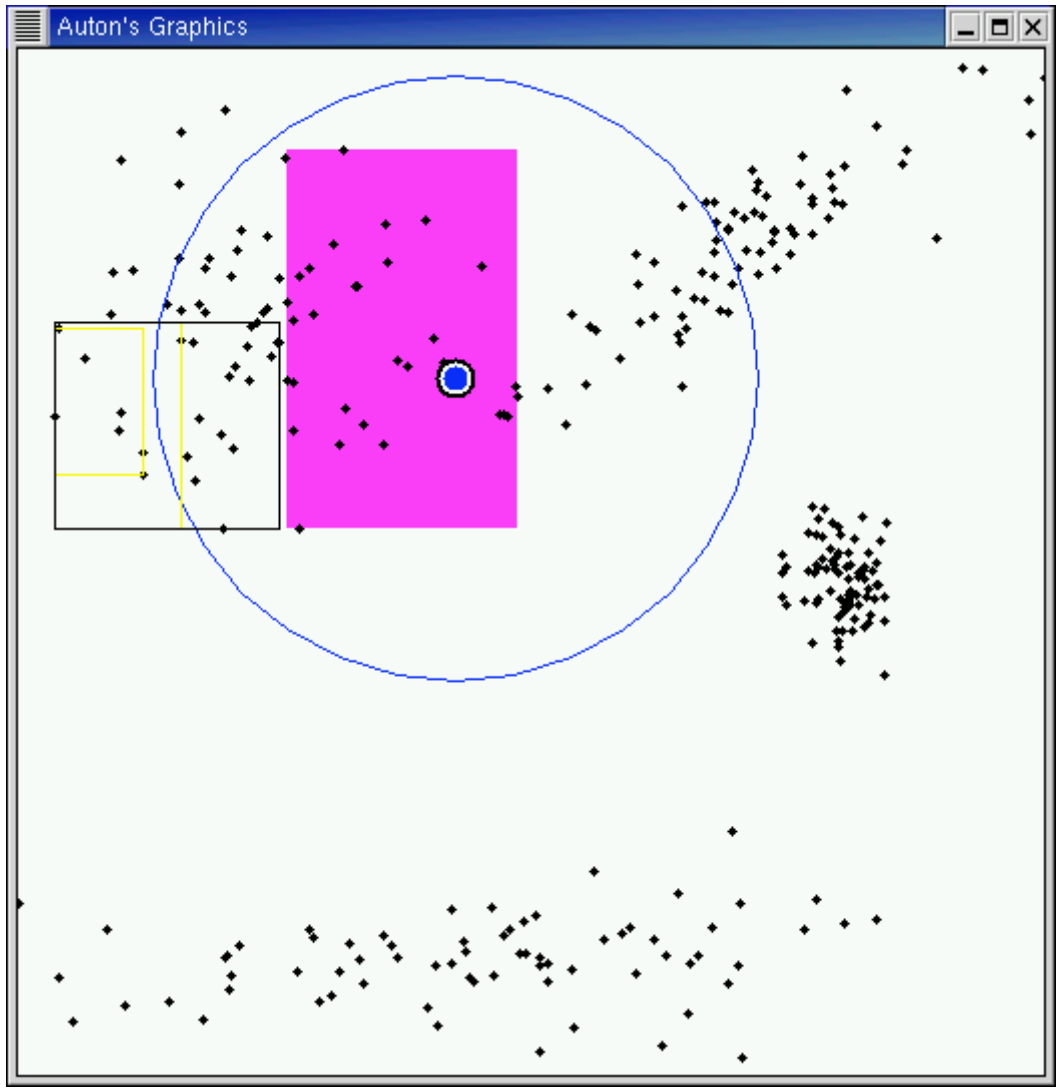


Range-count example

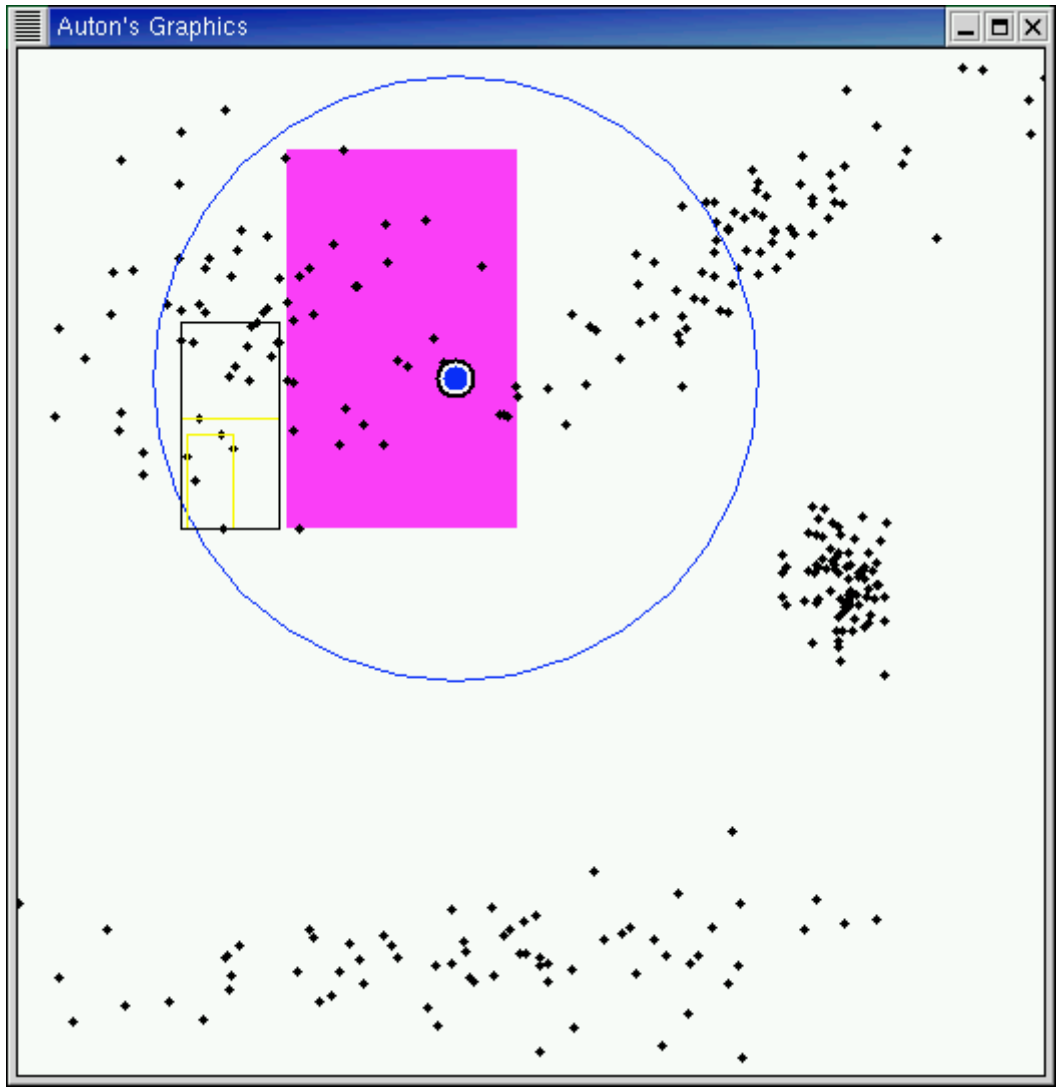


Pruned!
(inclusion)

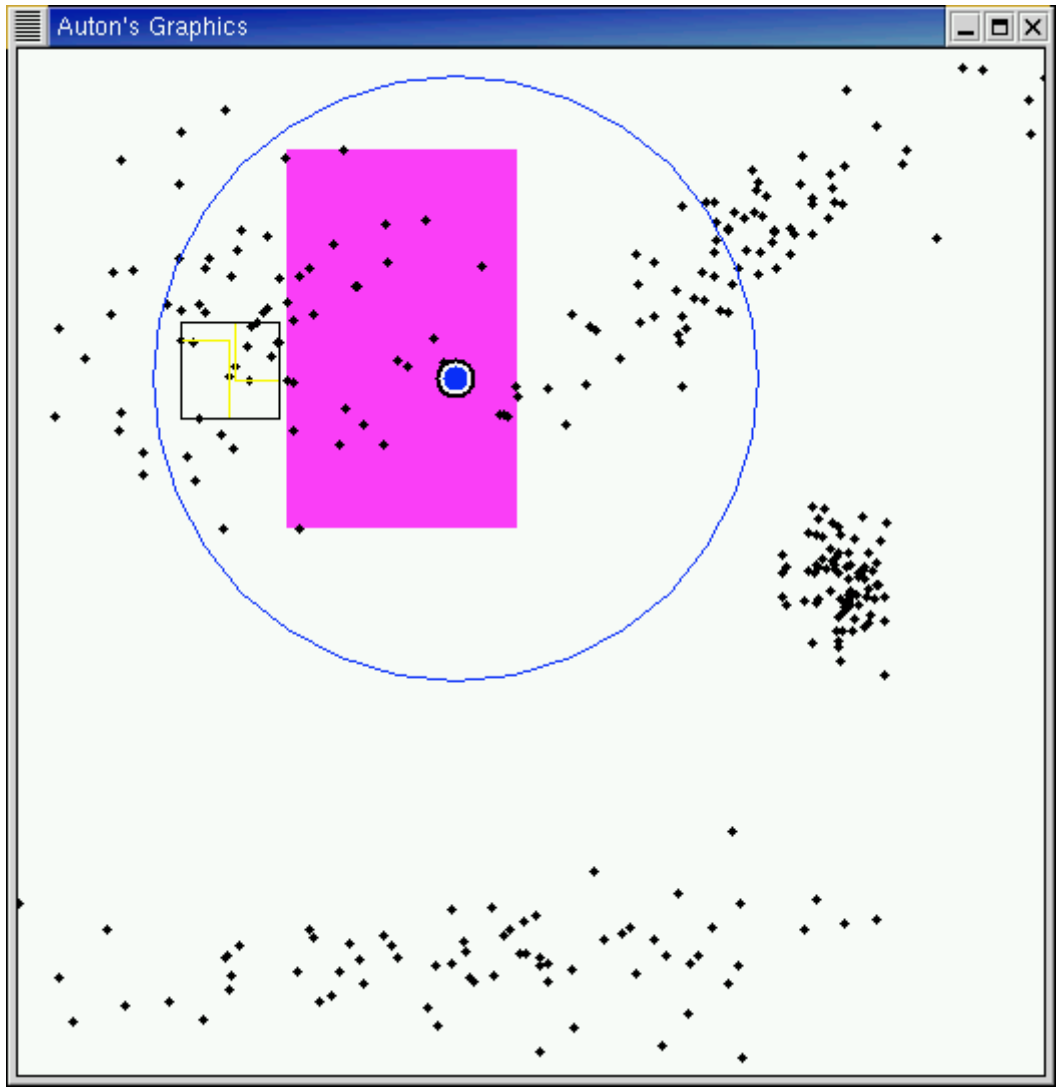
Range-count example



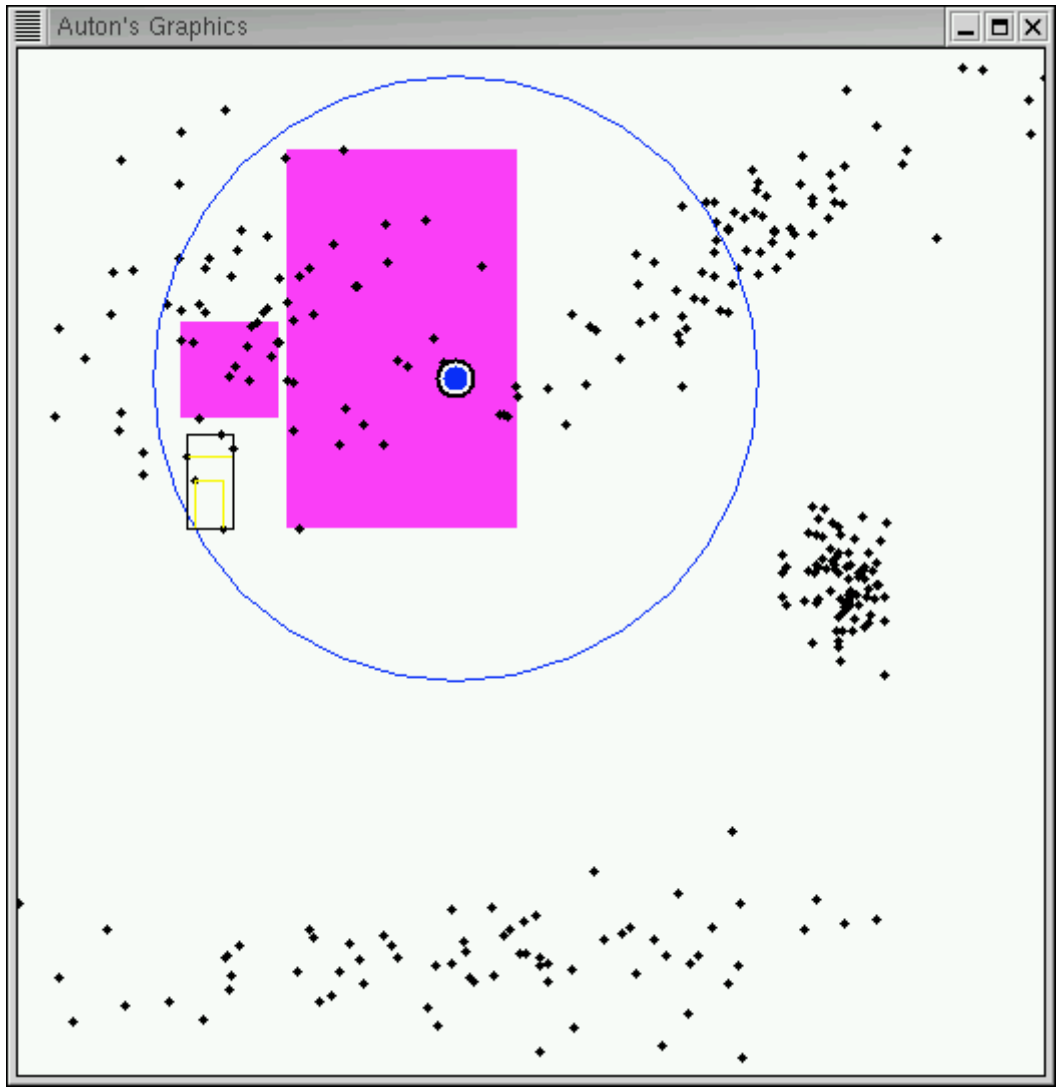
Range-count example



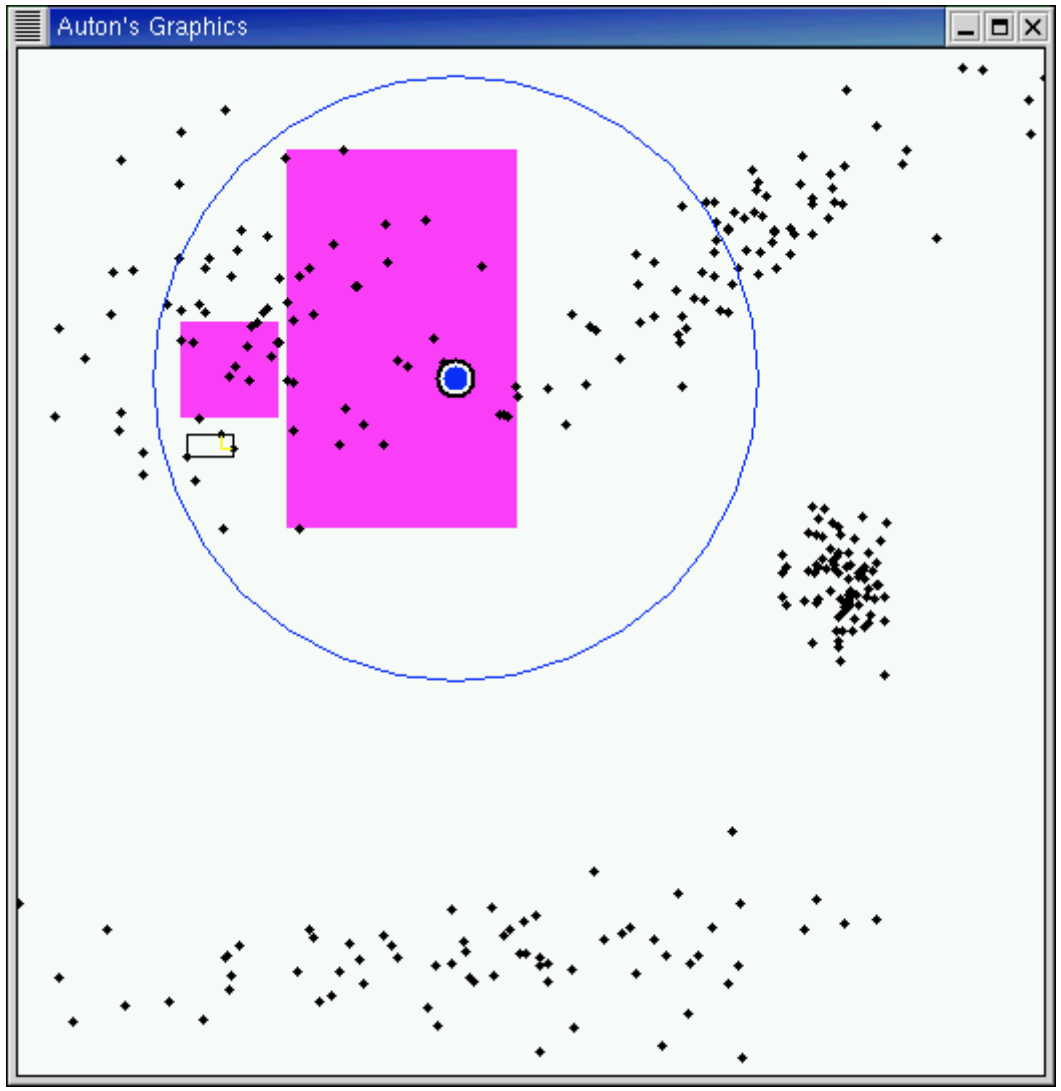
Range-count example



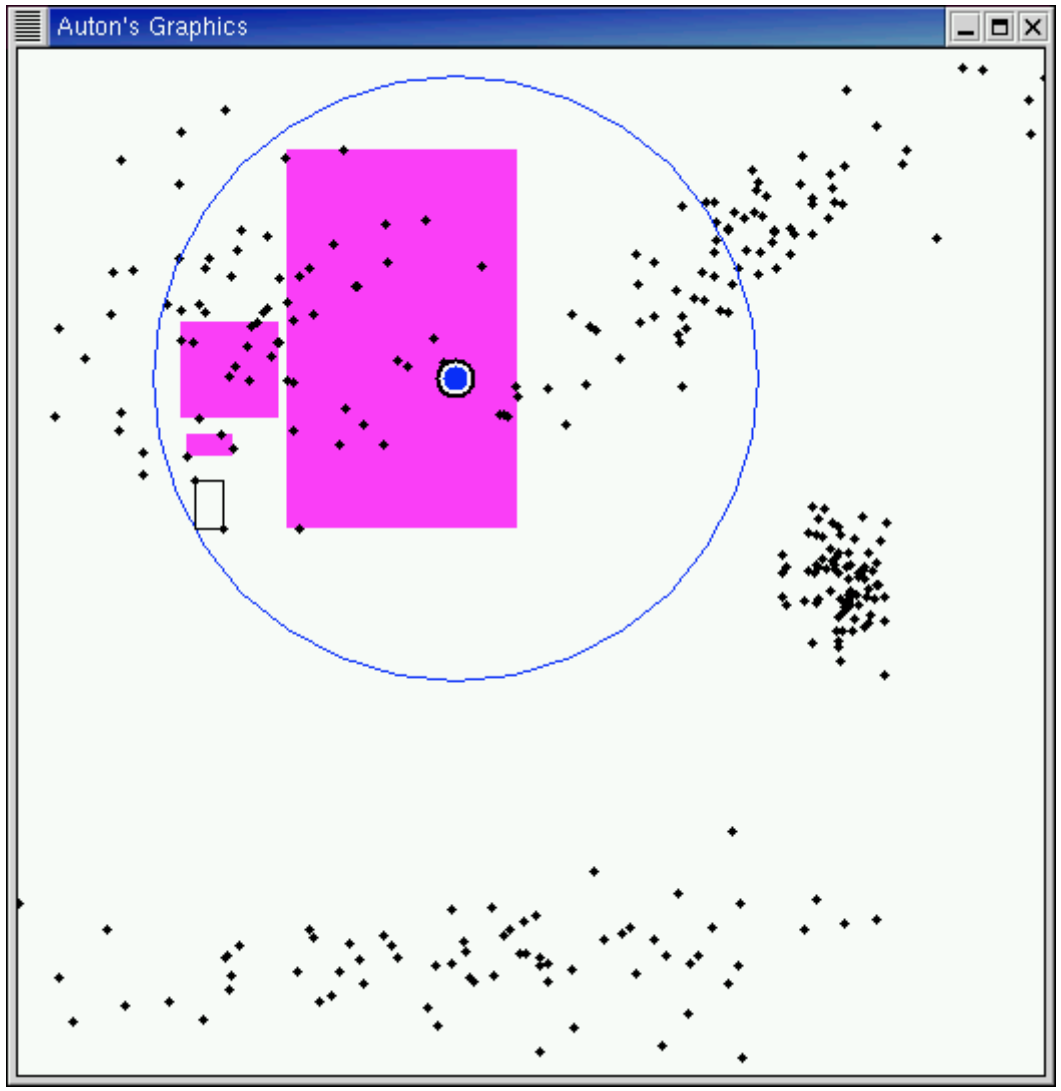
Range-count example



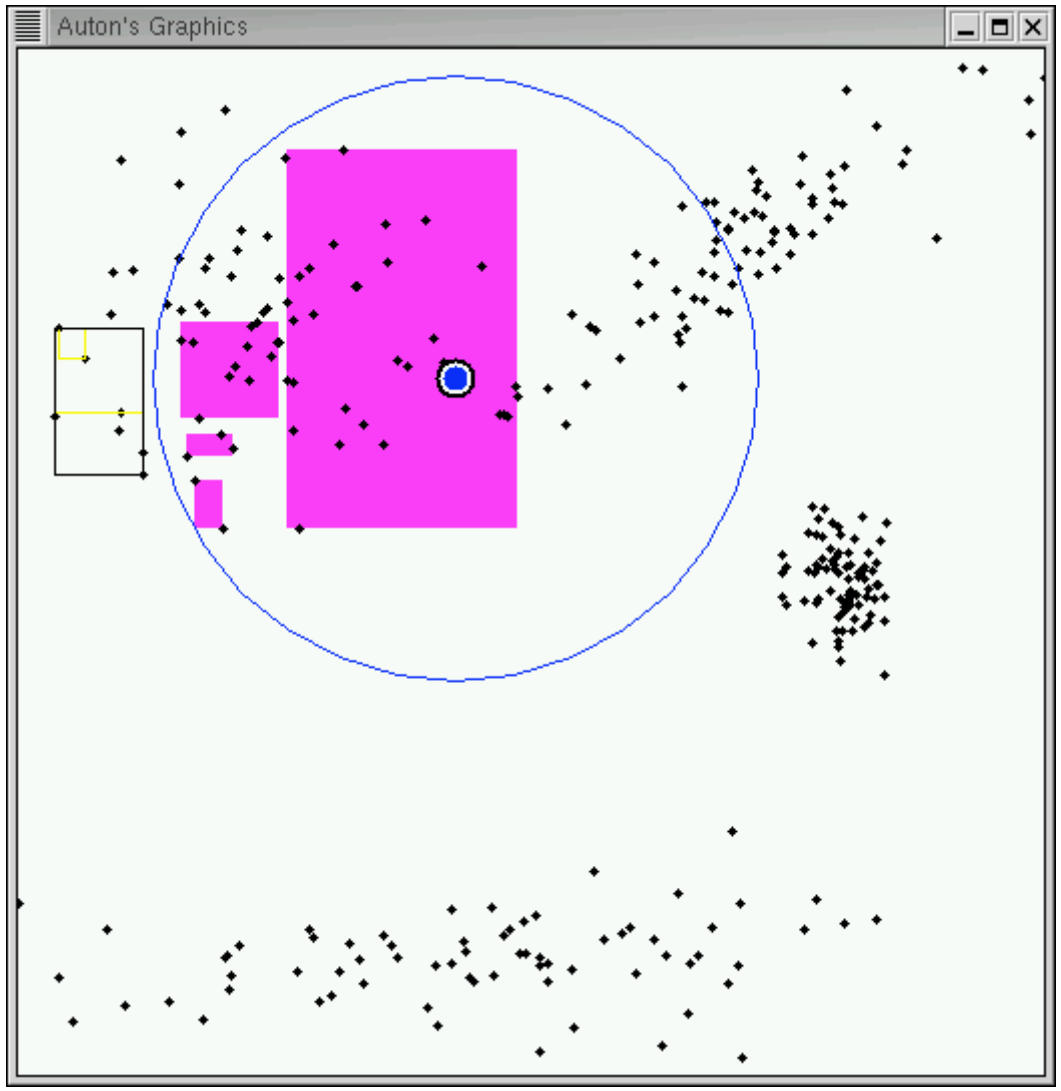
Range-count example



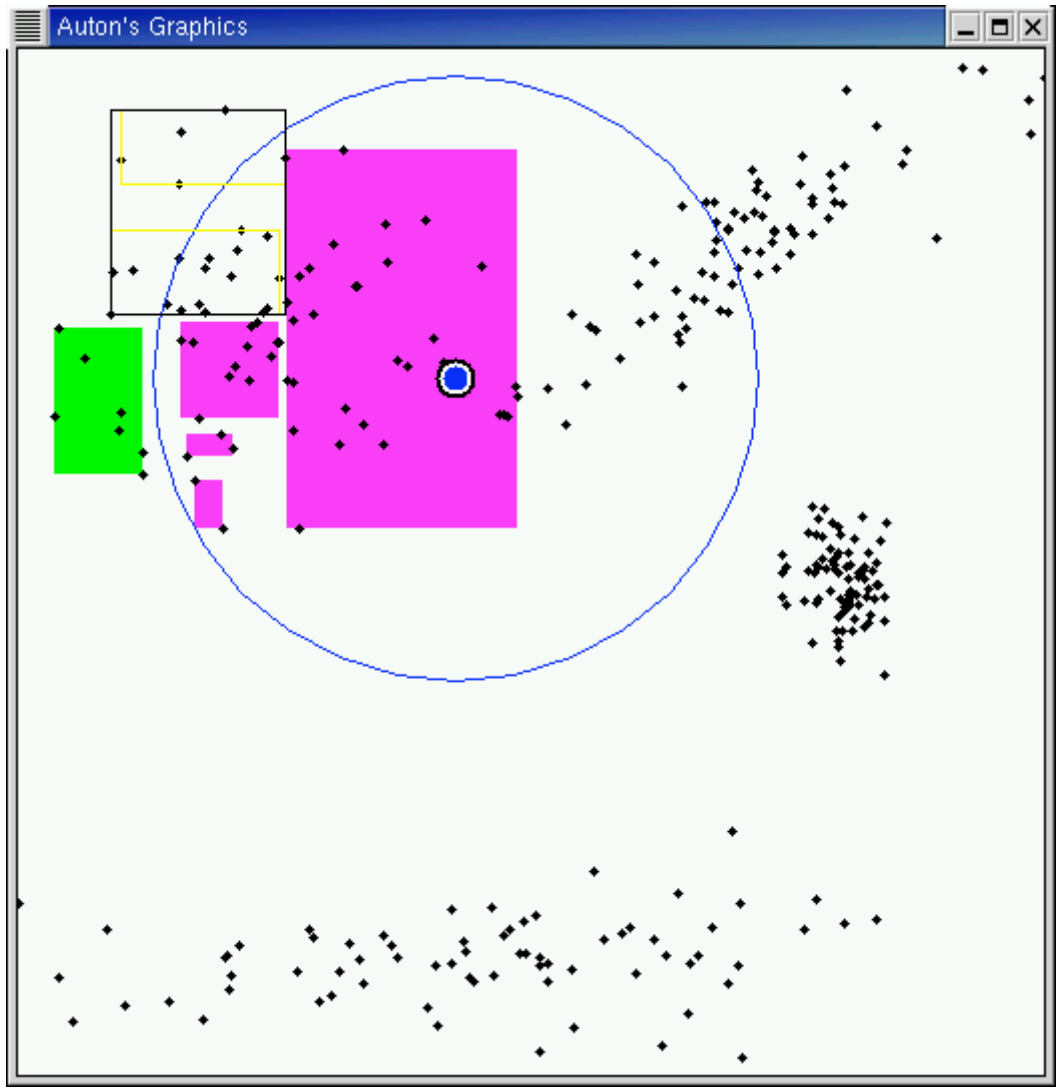
Range-count example



Range-count example

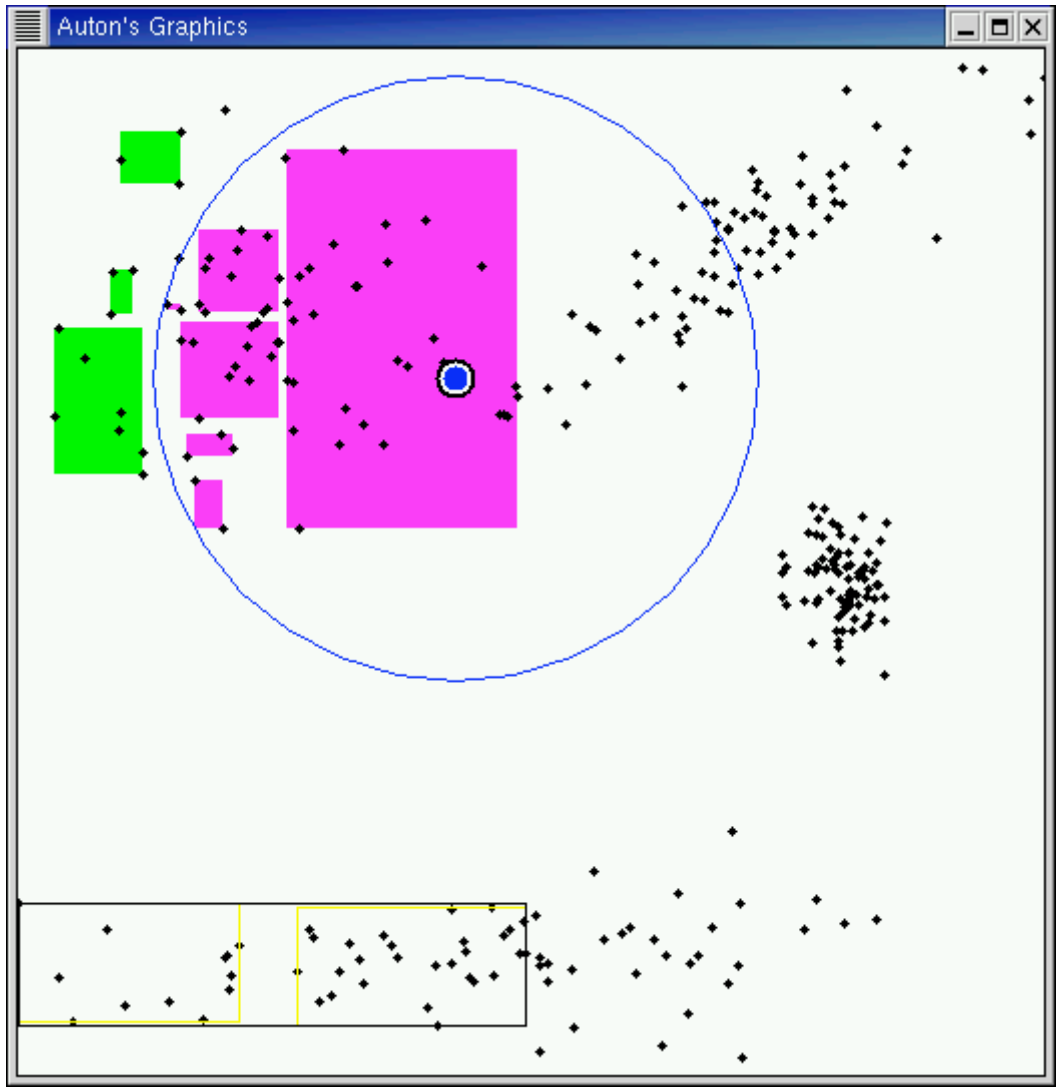


Range-count example

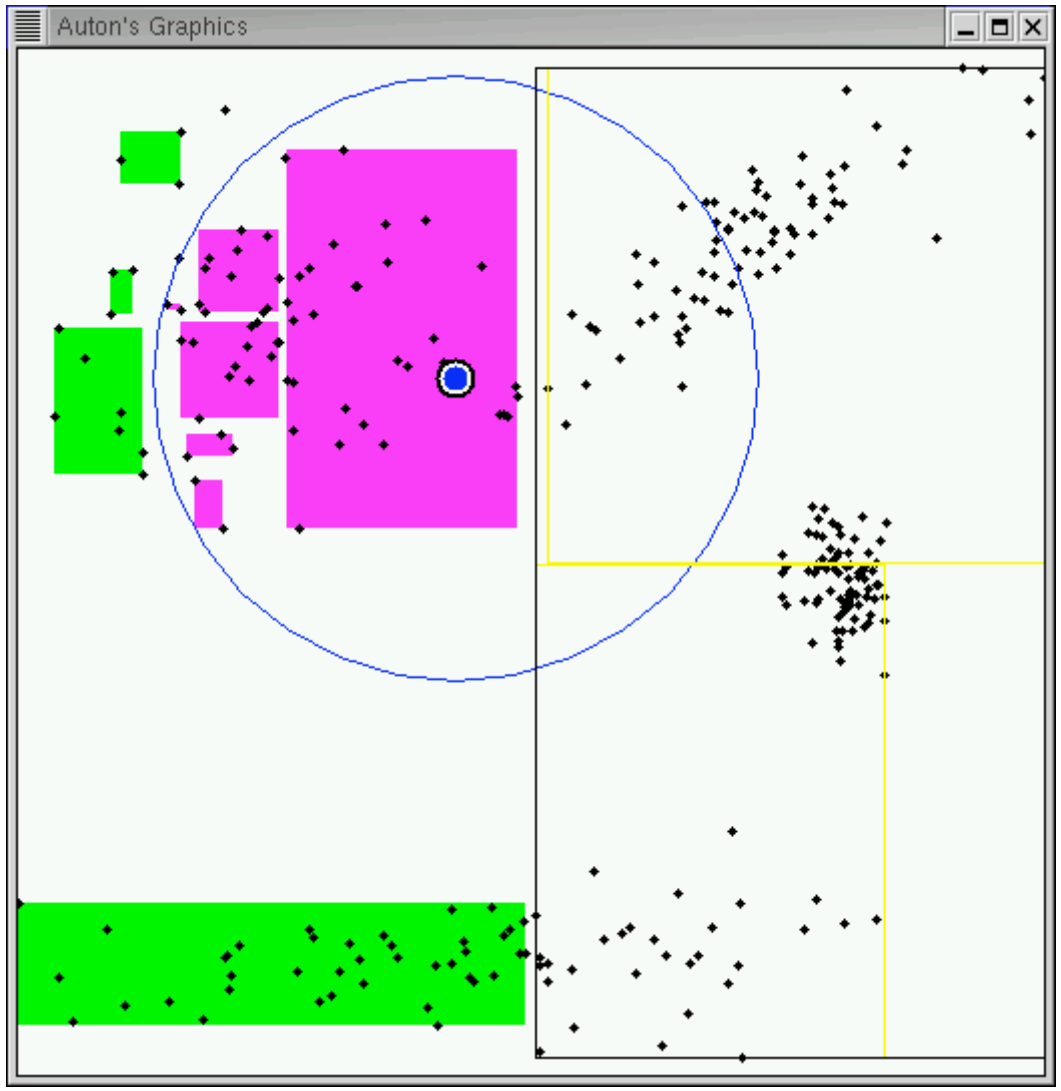


Pruned!
(exclusion)

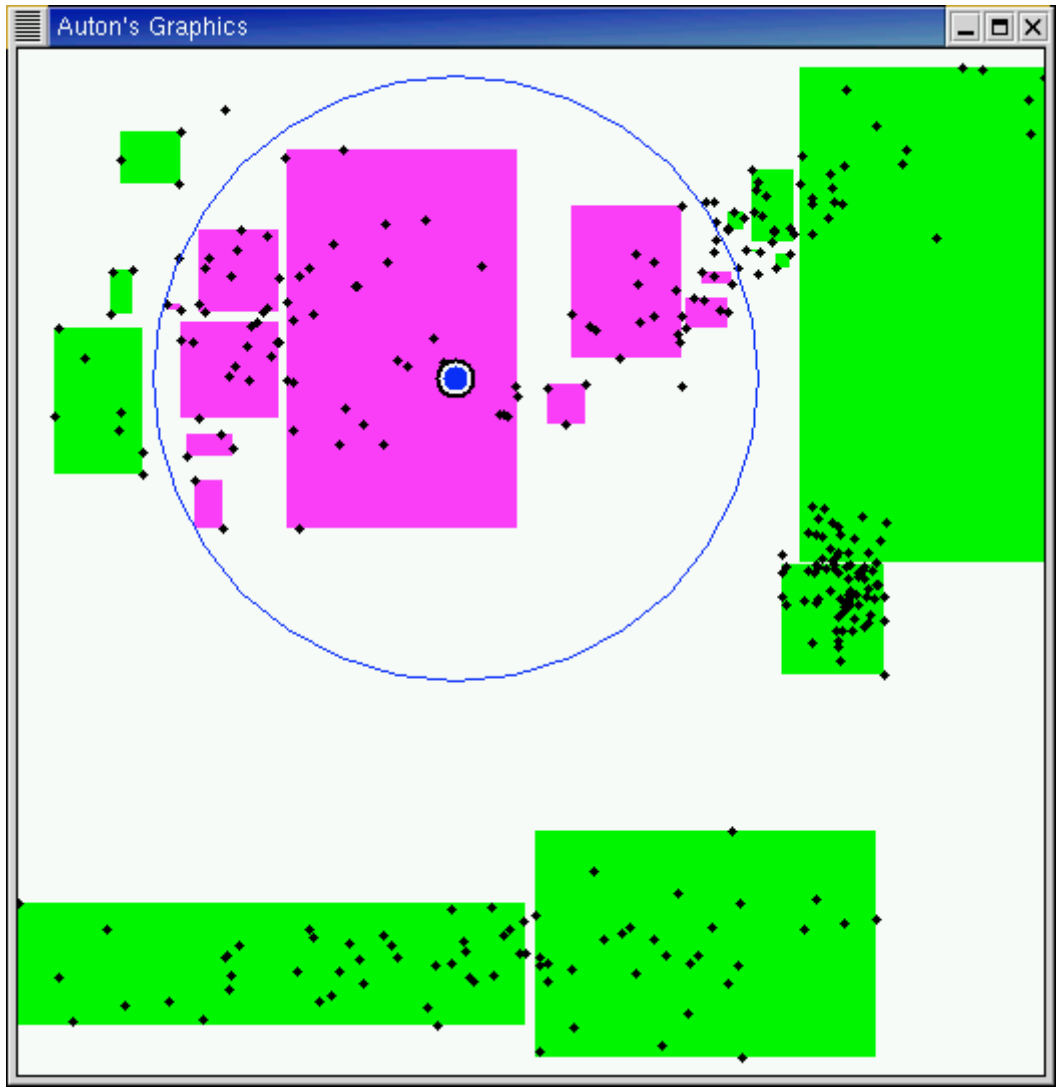
Range-count example



Range-count example



Range-count example



‘Single-tree’ solution

Do range-counting with inclusion,
with kd-trees.
(Then divide by 2 to get pair count.)

$$N \cdot O(N) \rightarrow N \cdot O(\log N)$$

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

Not bad, but we can do better….

Outline:

1. Some cosmological questions

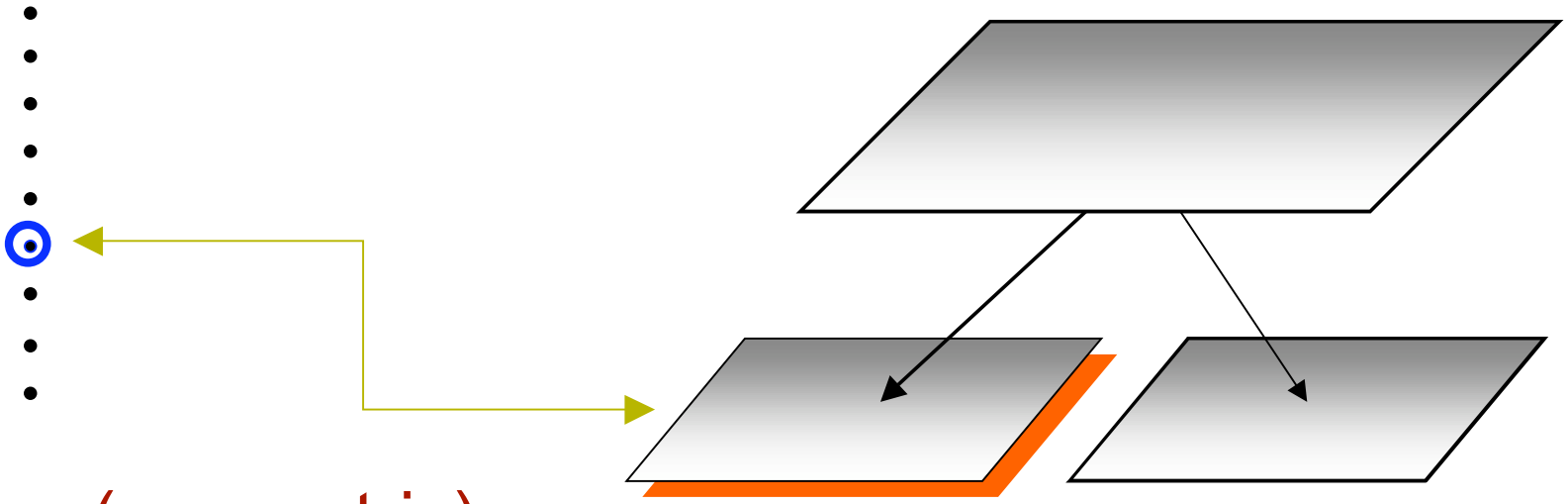
2. Spatial correlations

3. Divide-and-conquer in real-space

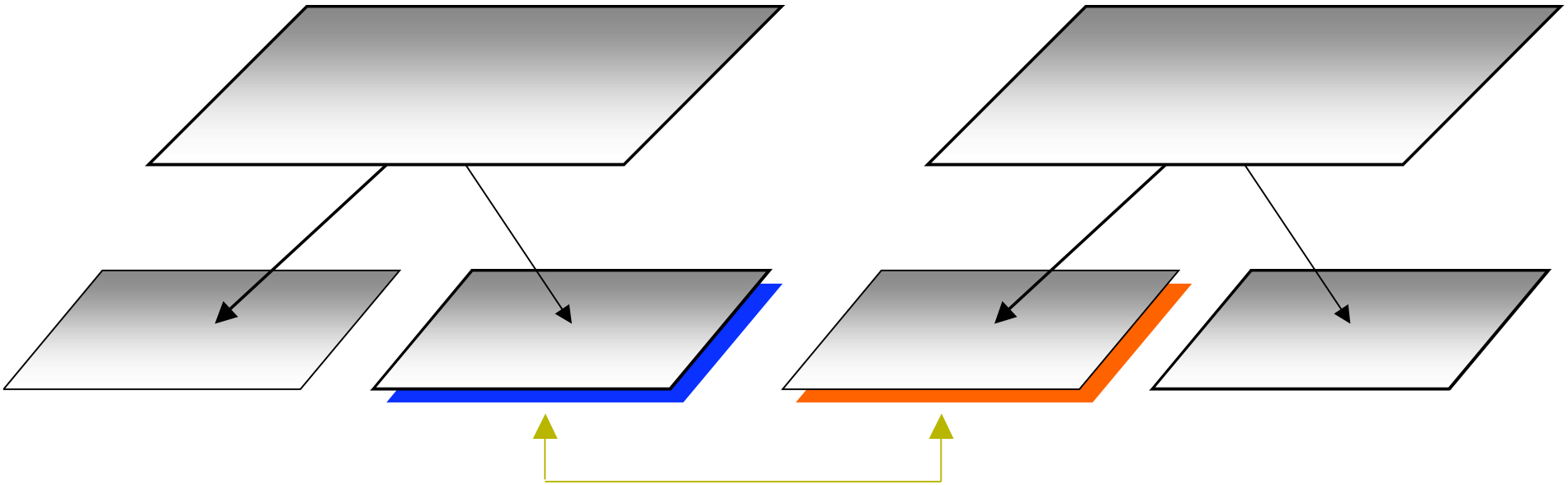
4. **Multi-tree algorithm, exact**

5. Multi-tree Monte Carlo

Single-tree:



Dual-tree (symmetric):



Divide-and-conquer #1

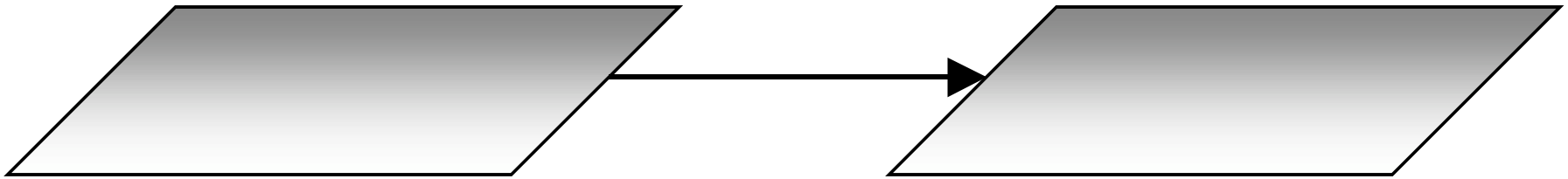
Divide-and-conquer over multiple sets simultaneously

[Appel 1983], [Callahan-Kosaraju 1993], [Gray-Moore 2000, Gray 2003]

Exclusion and inclusion

using *k*d-tree **node-node** bounds

O(D) bounds on distance minima/maxima:



(Analogous to point-node bounds.)

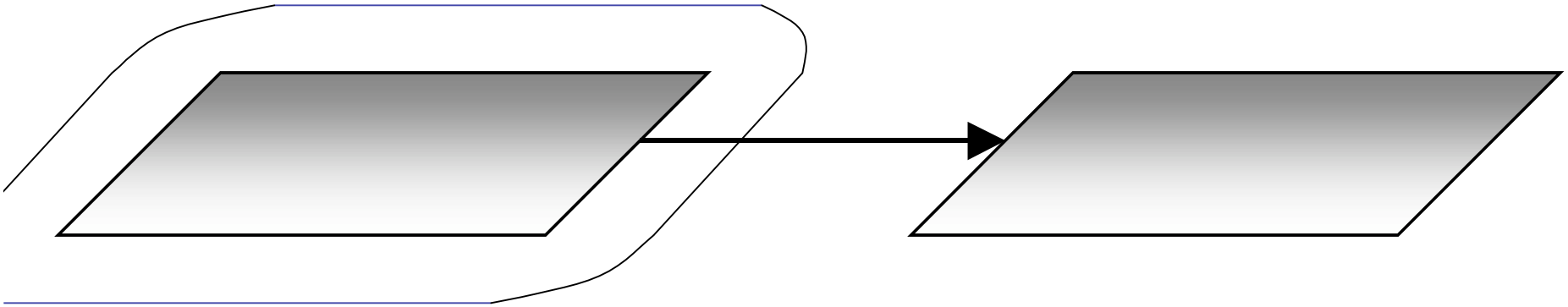
Also needed:

Nodewise bounds.

Exclusion and inclusion

using *k*d-tree **node-node** bounds

O(D) bounds on distance minima/maxima:



(Analogous to point-node bounds.)

Also needed:

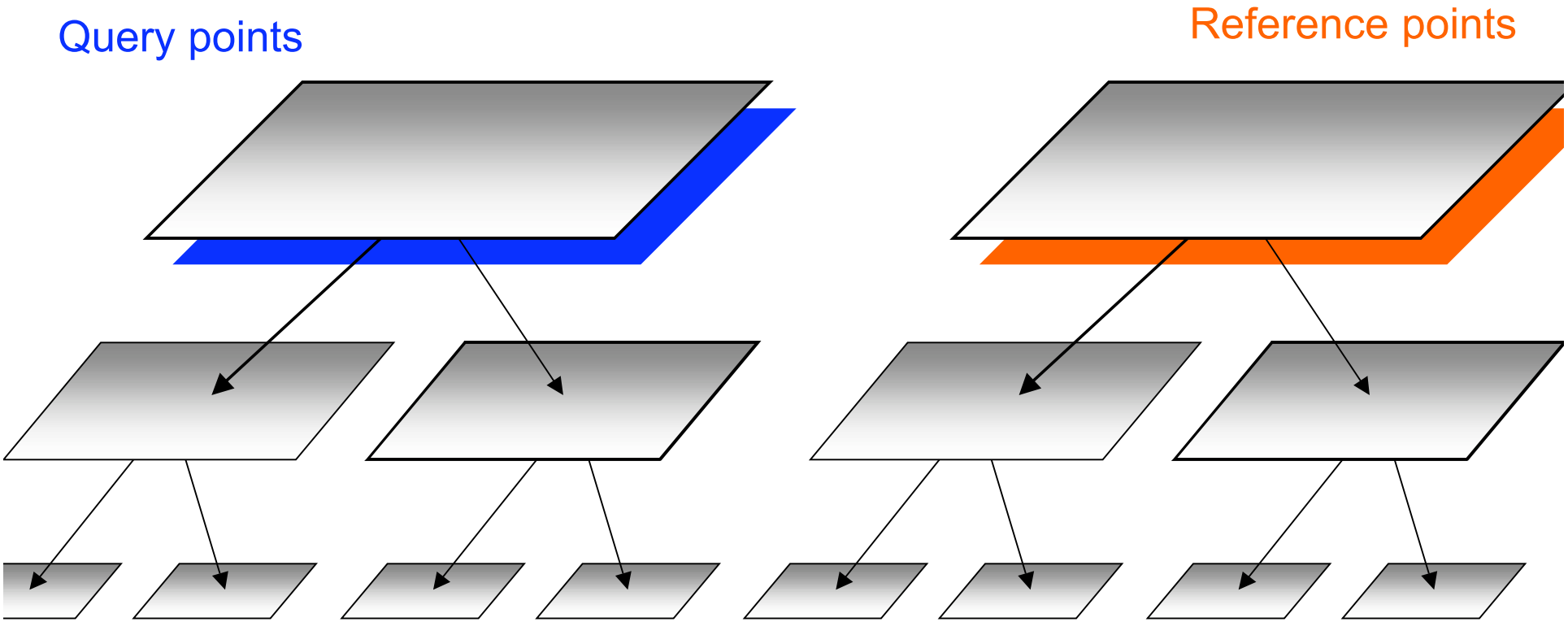
Nodewise bounds.

Simple recursive algorithm

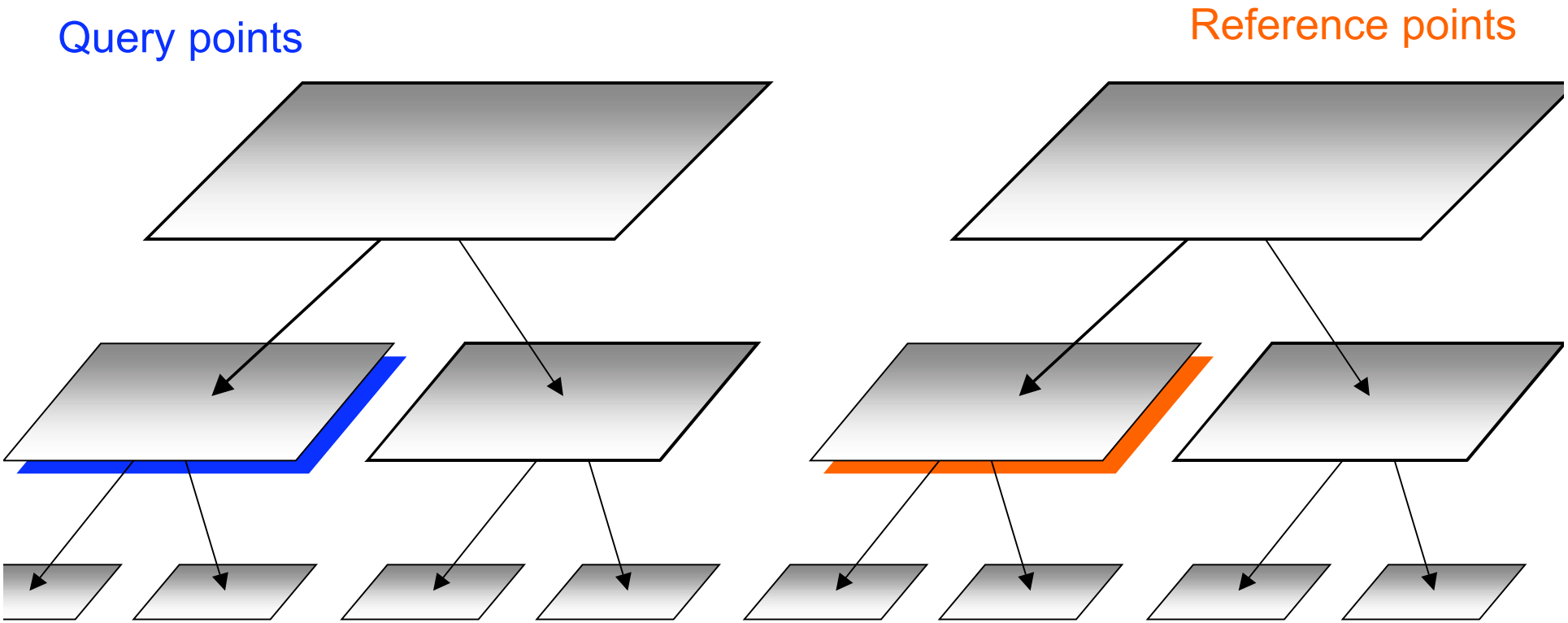
```
2pt(Q,R)
{
   if Q.id > R.id, return 0.
   if exclude(Q,R), return 0.
   if include(Q,R), return Q.count x R.count.

   if leaf(Q) and leaf(R), 2ptBase(Q,R).
   else,
      return 2pt(Q.left,R.left) + 2pt(Q.left,R.right) +
             2pt(Q.right,R.left) + 2pt(Q.right,R.right).
}
```

Dual-tree traversal

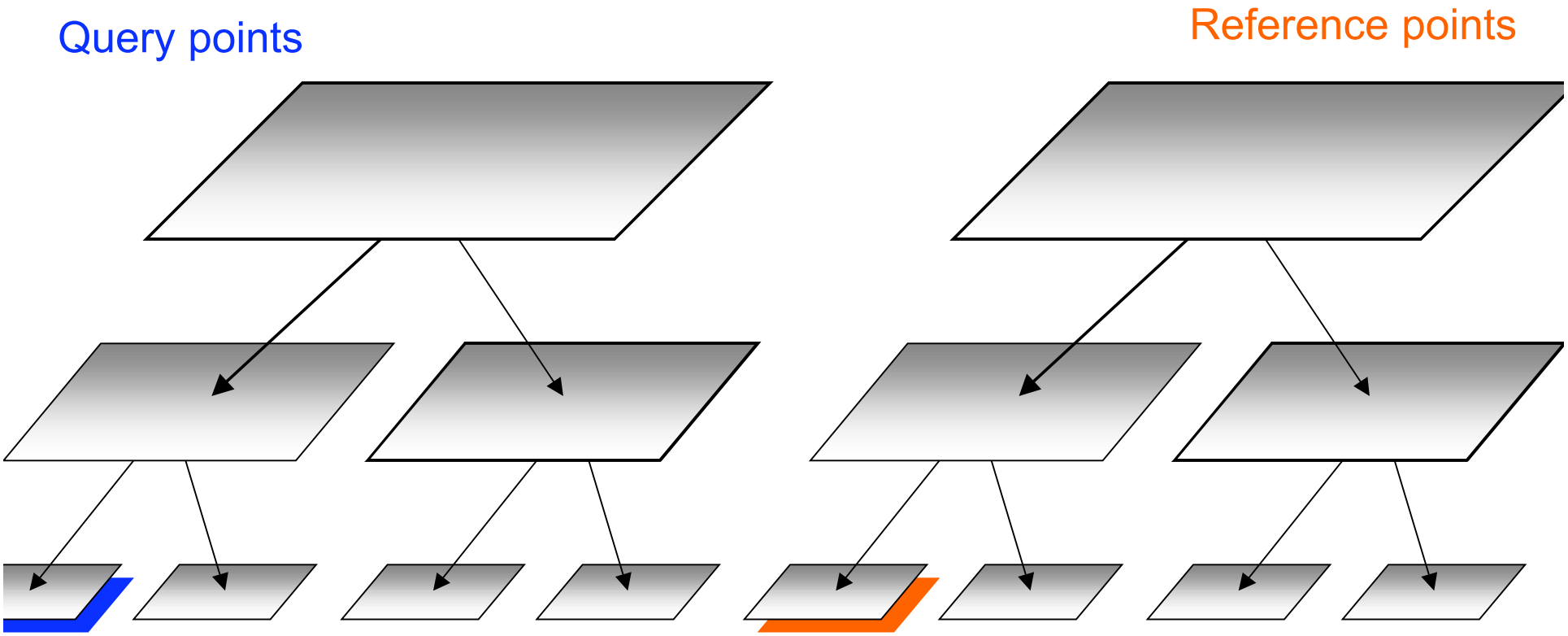
(depth-first)



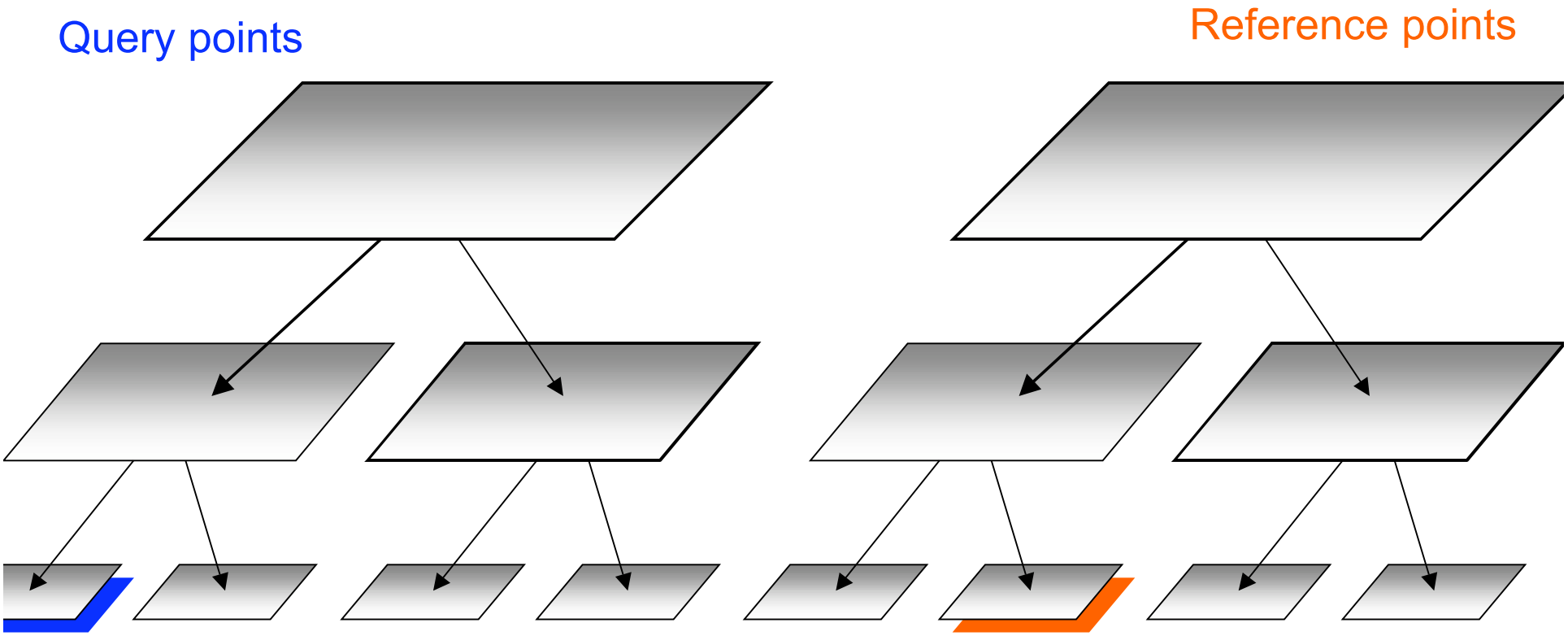
Dual-tree traversal



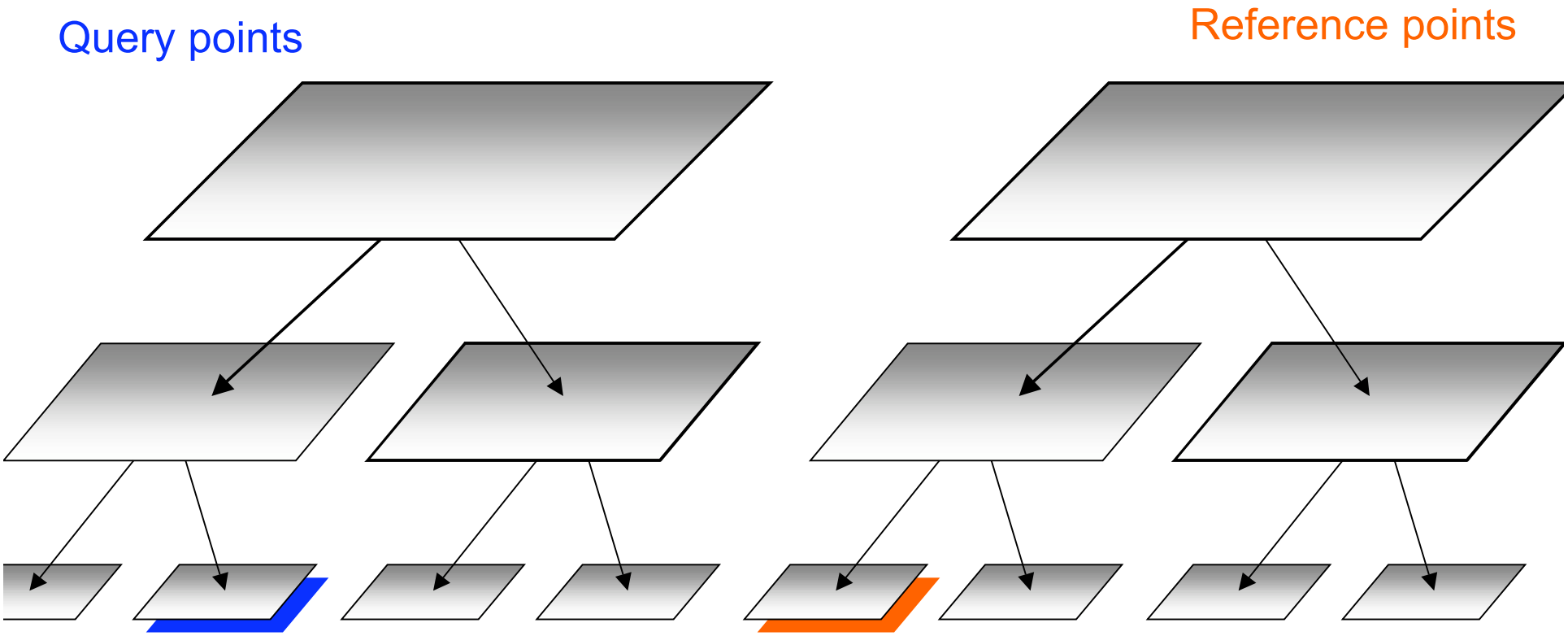
Dual-tree traversal



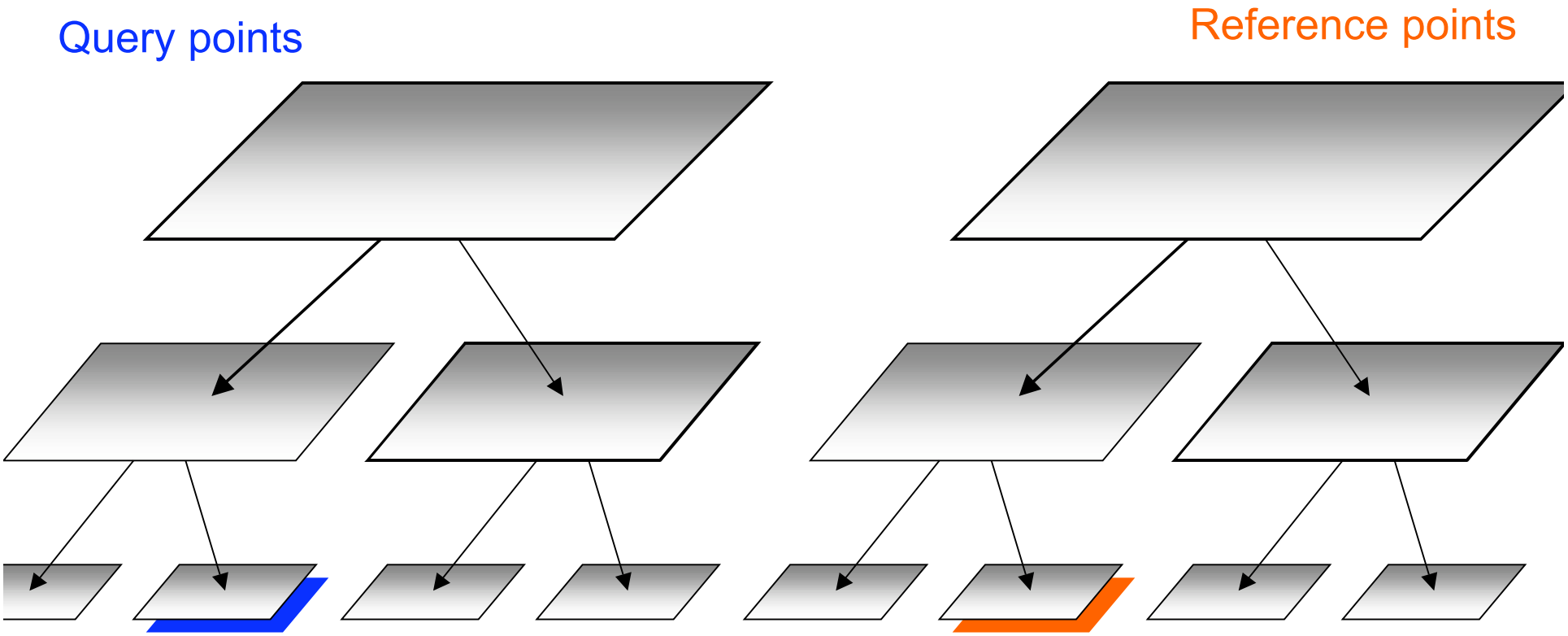
Dual-tree traversal



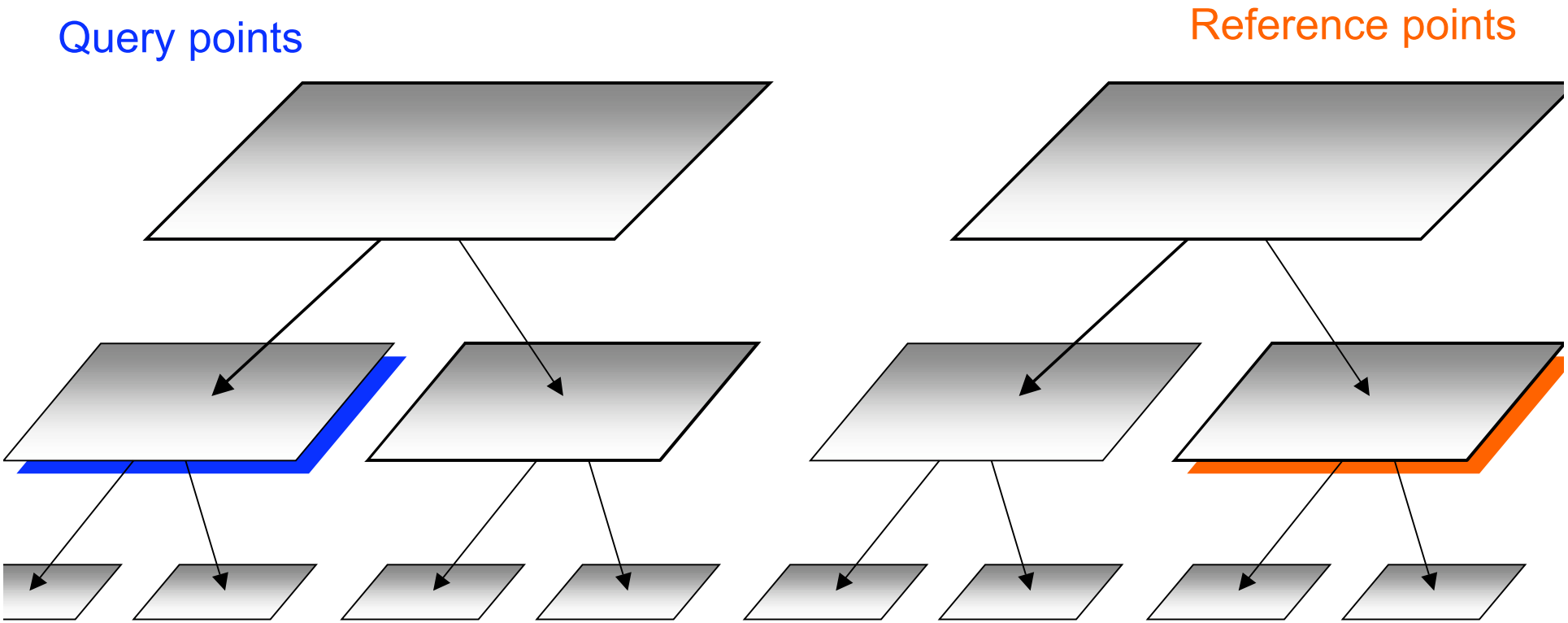
Dual-tree traversal



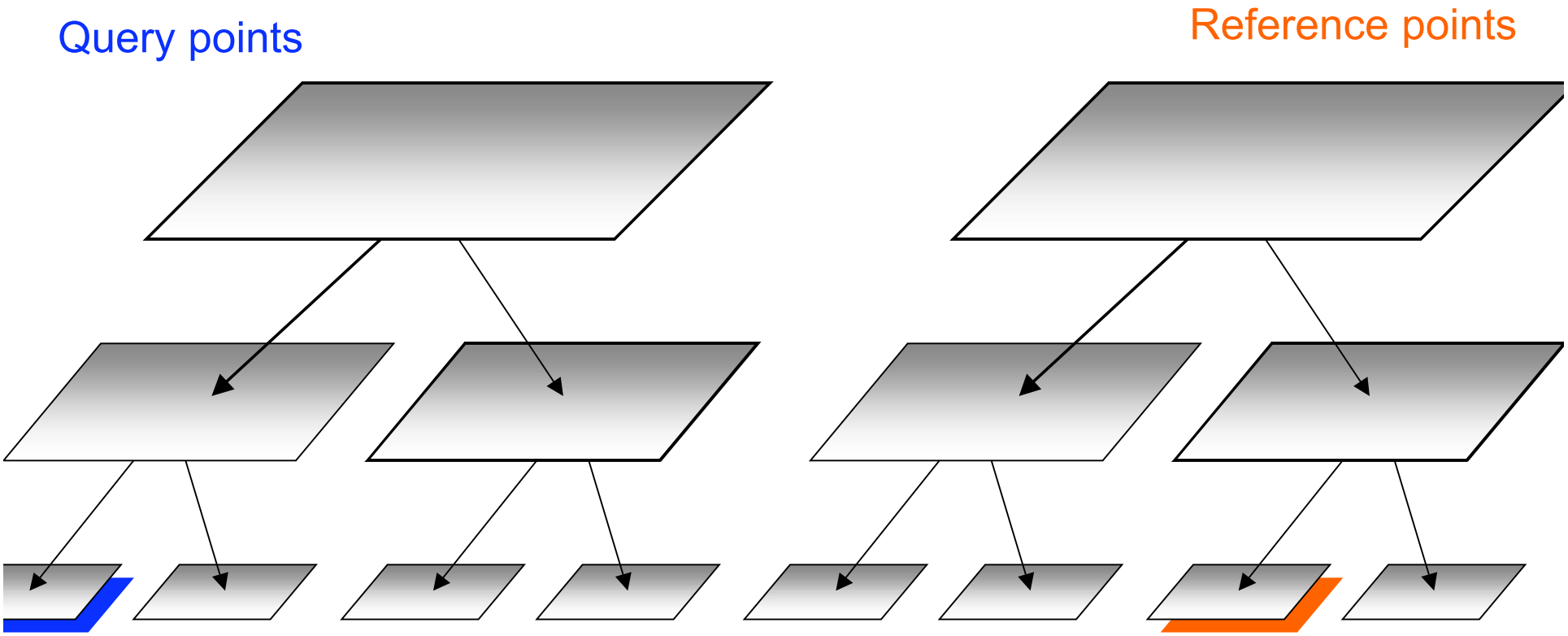
Dual-tree traversal



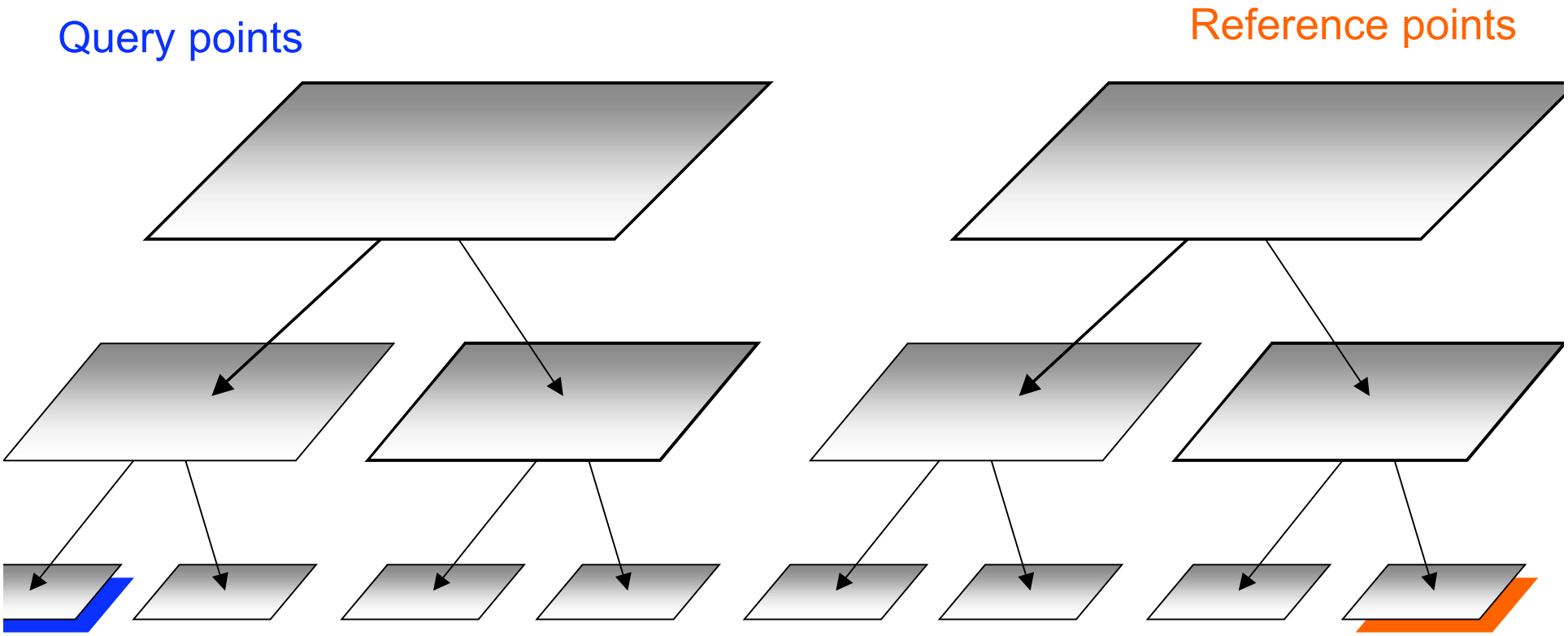
Dual-tree traversal



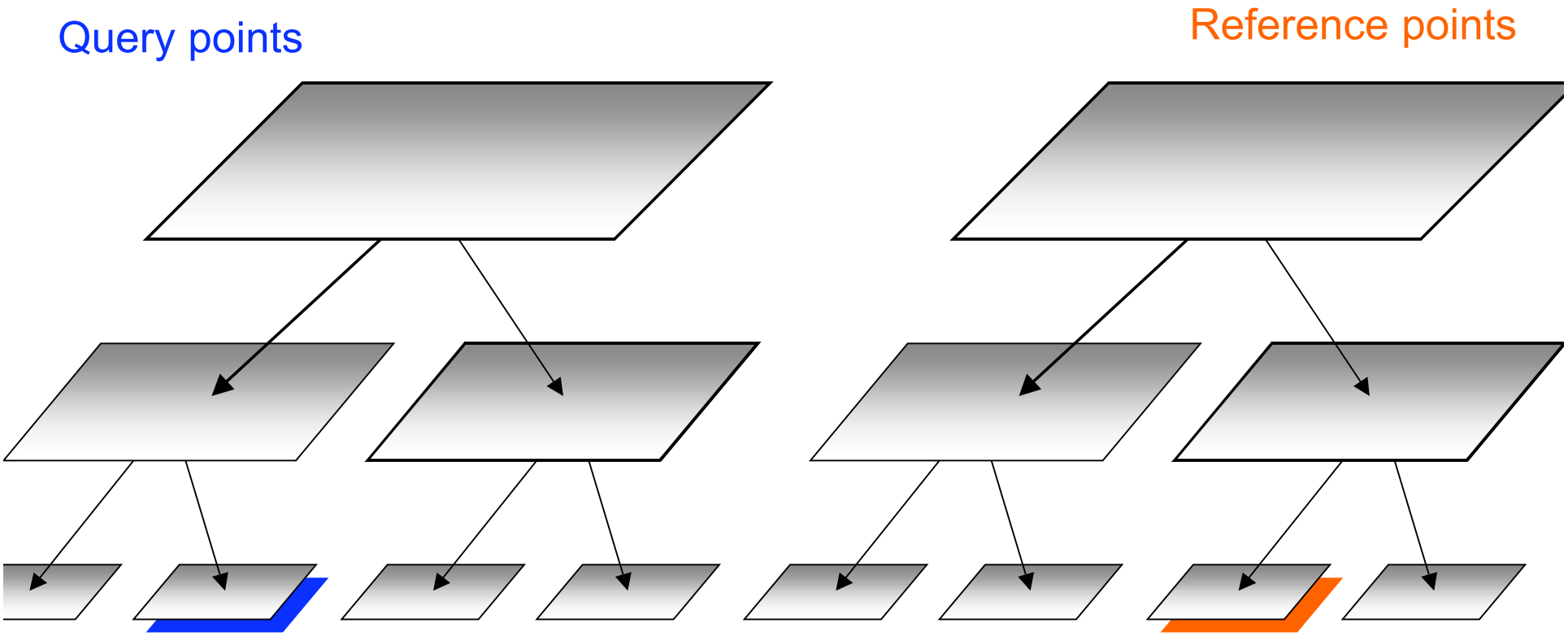
Dual-tree traversal



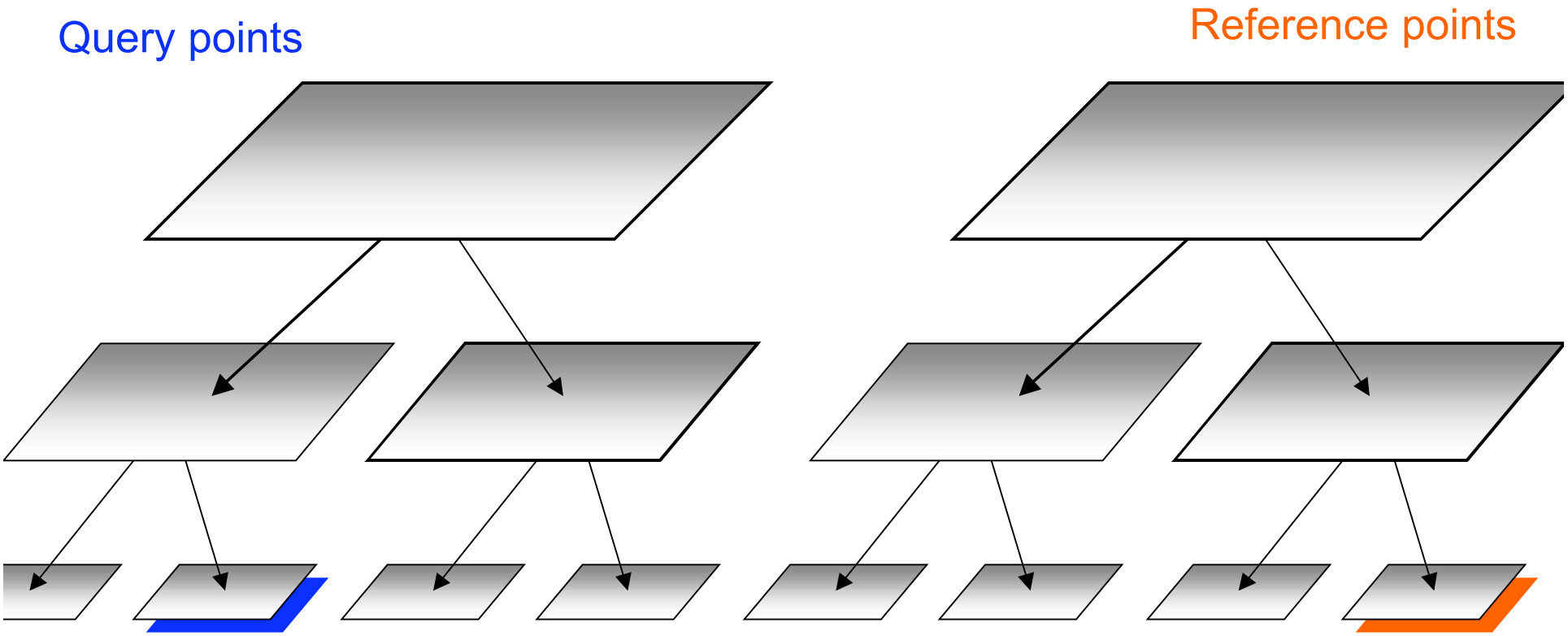
Dual-tree traversal



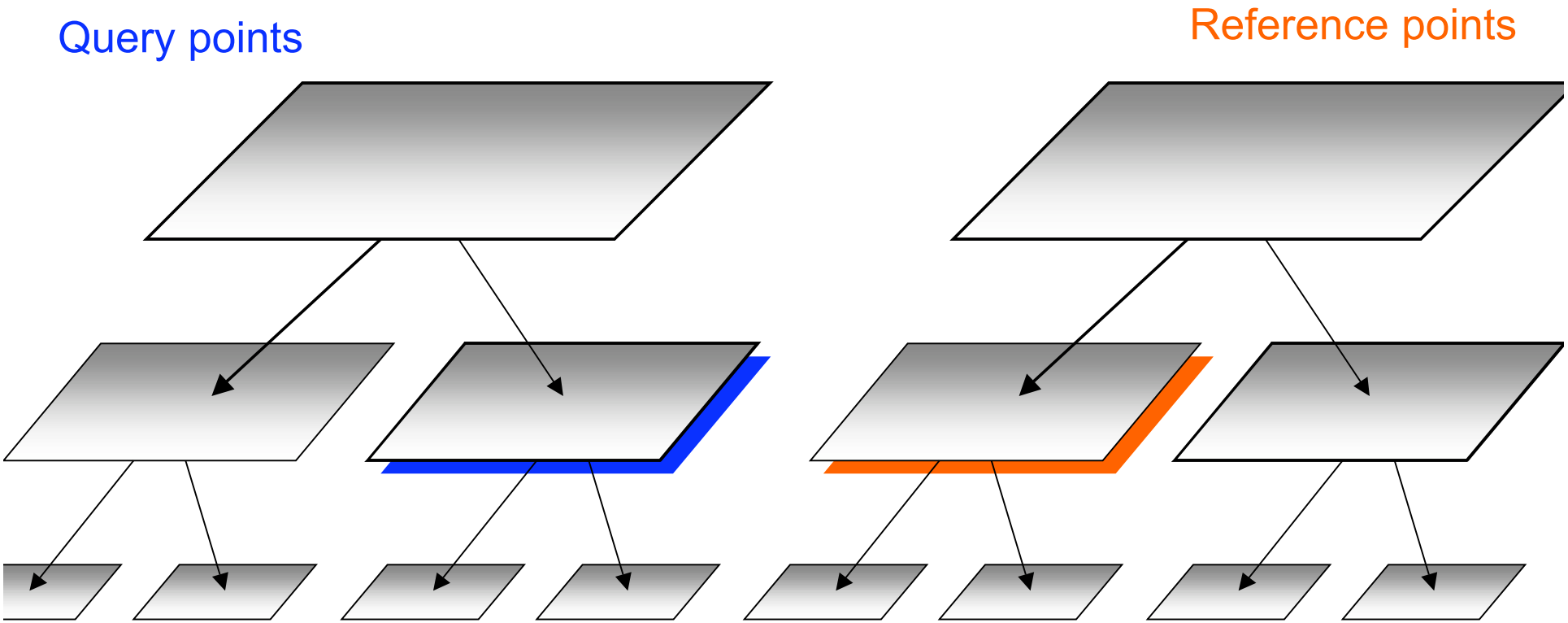
Dual-tree traversal



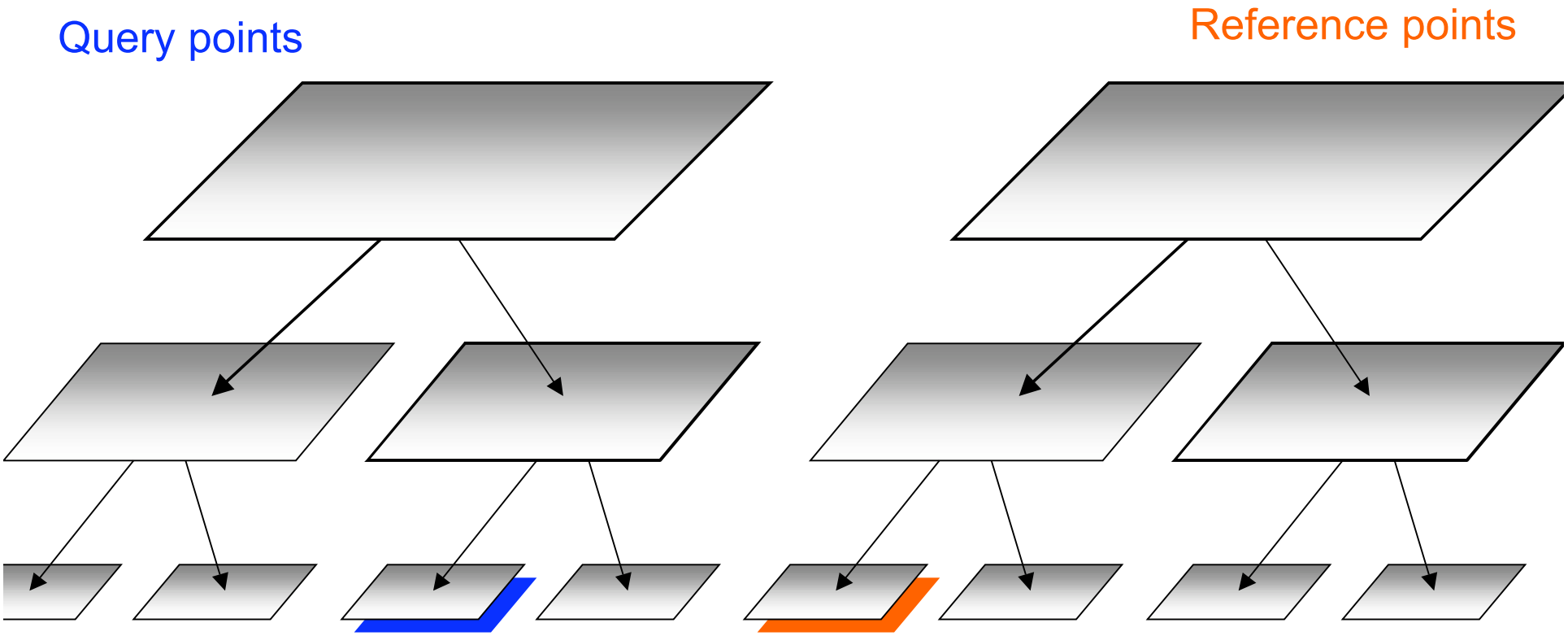
Dual-tree traversal



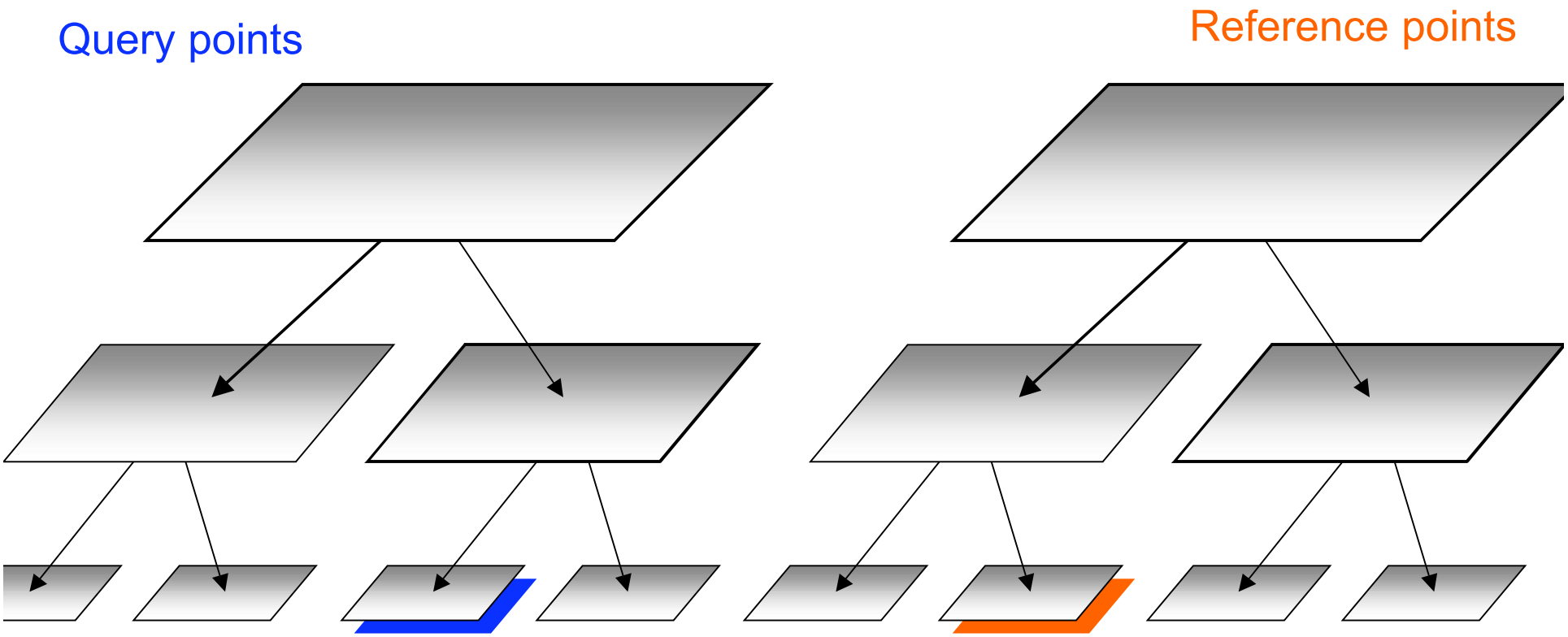
Dual-tree traversal



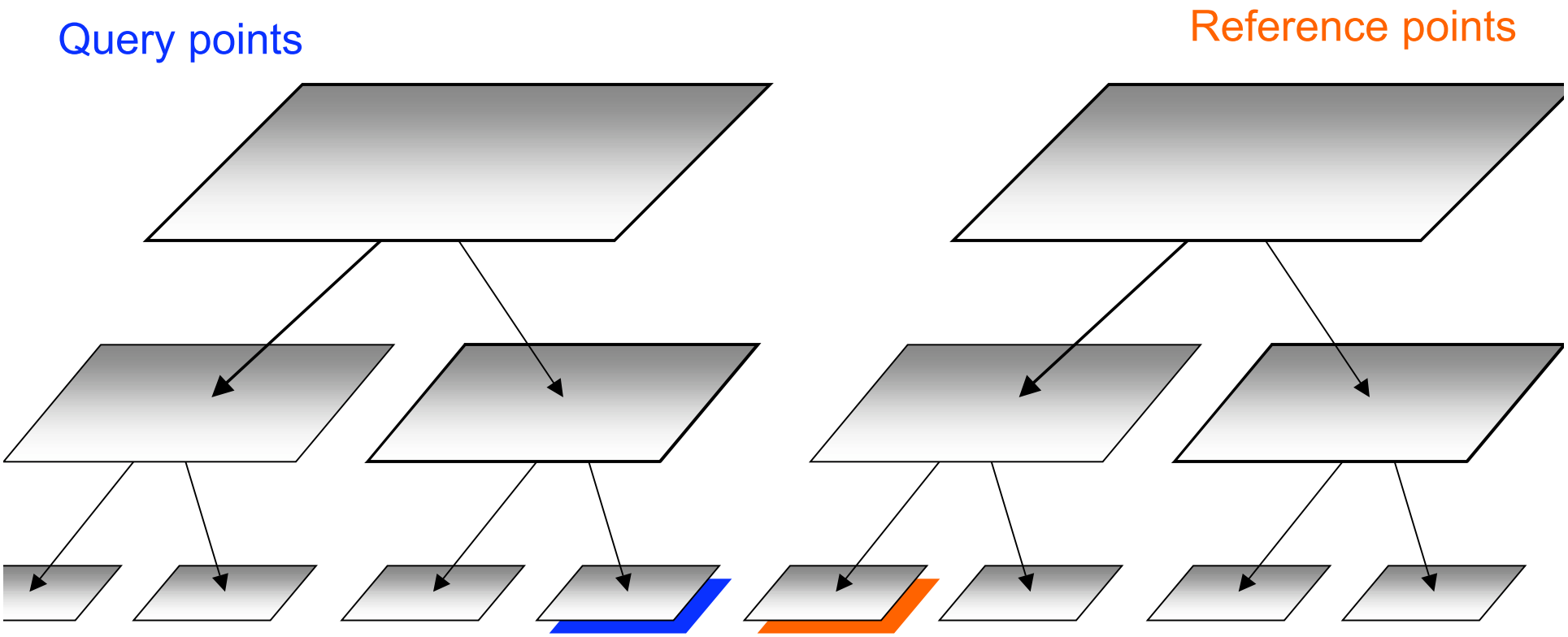
Dual-tree traversal



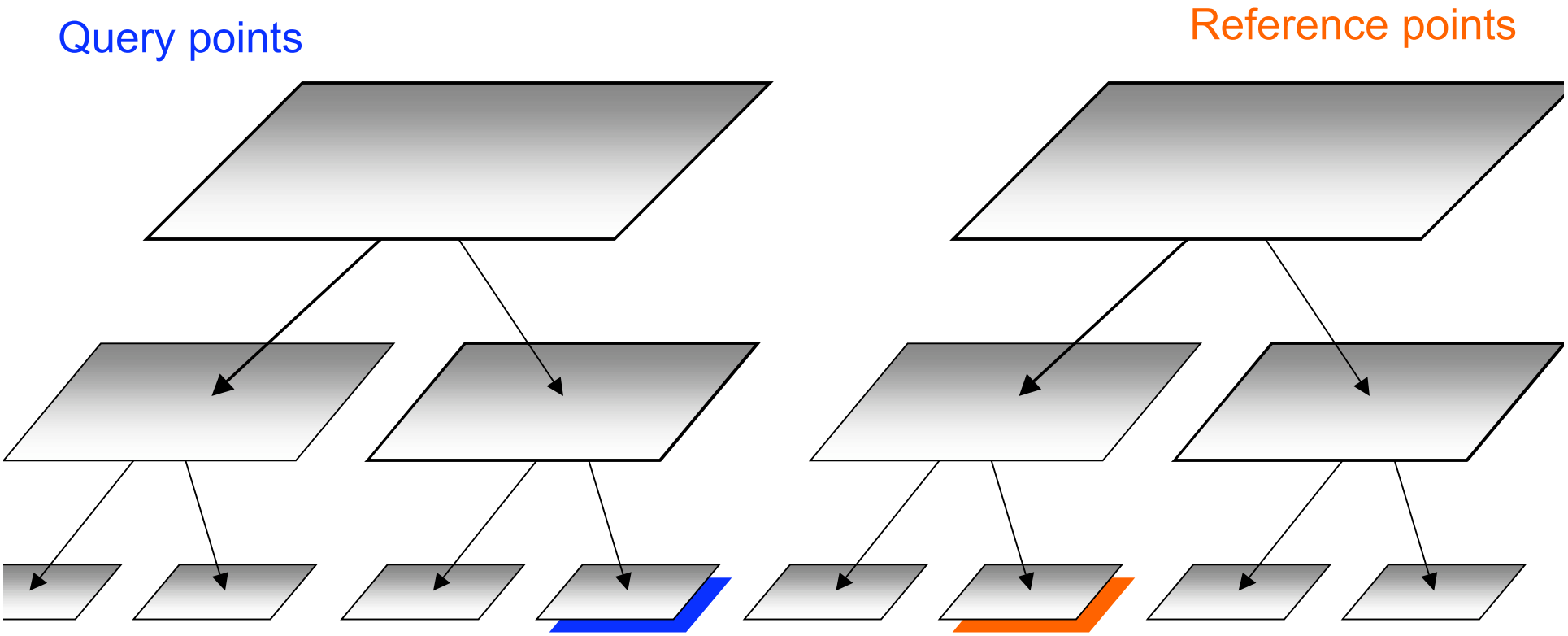
Dual-tree traversal



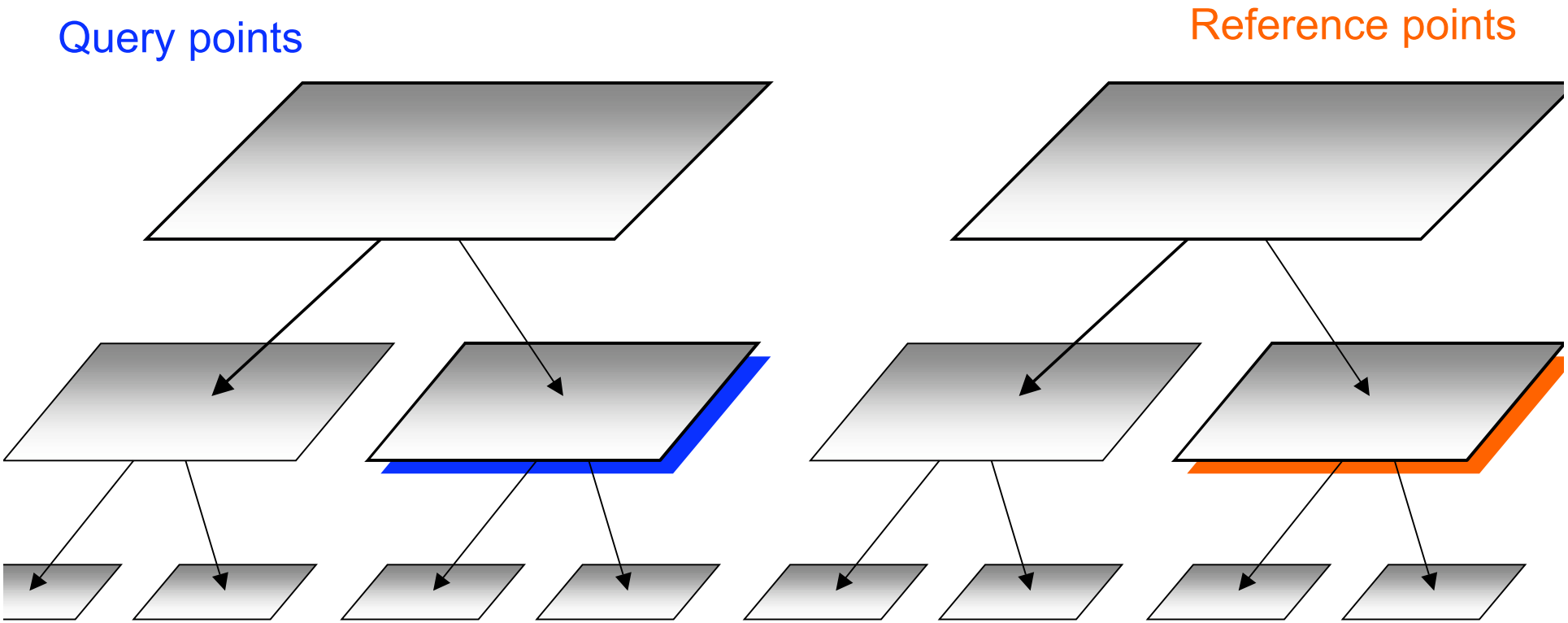
Dual-tree traversal



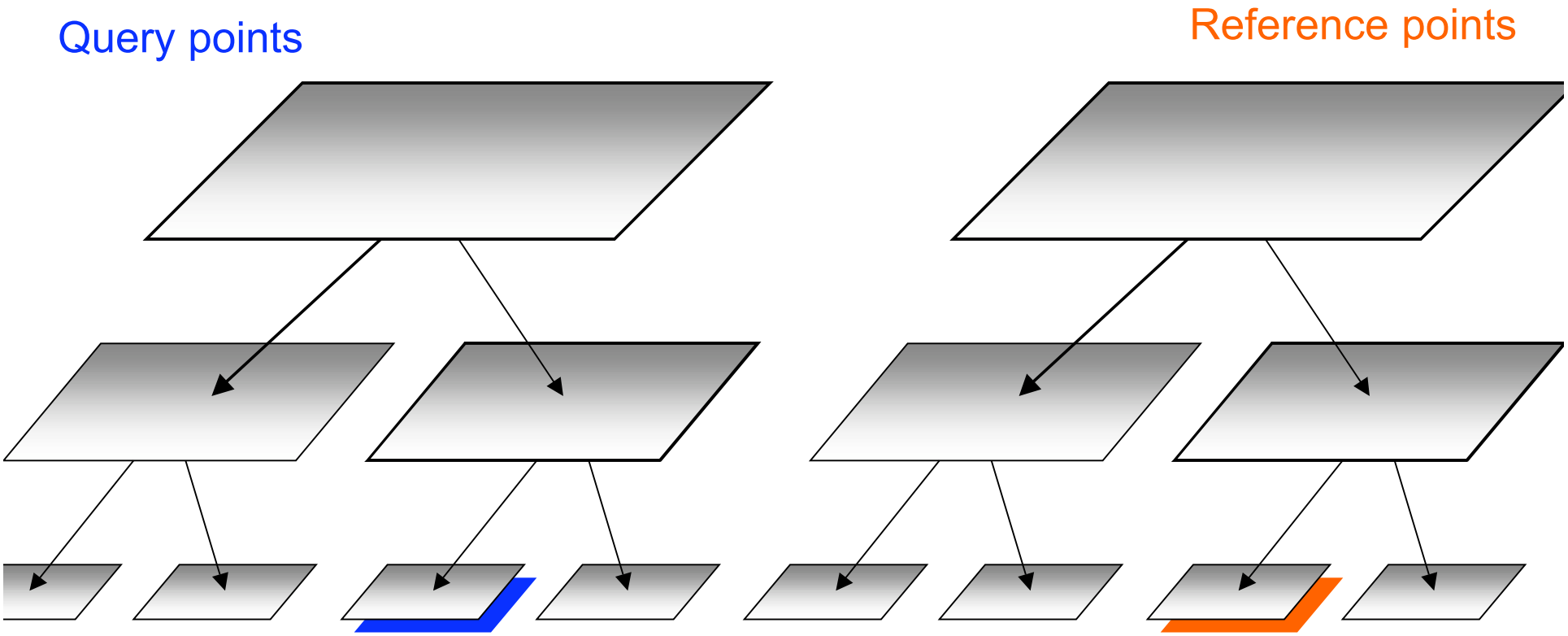
Dual-tree traversal



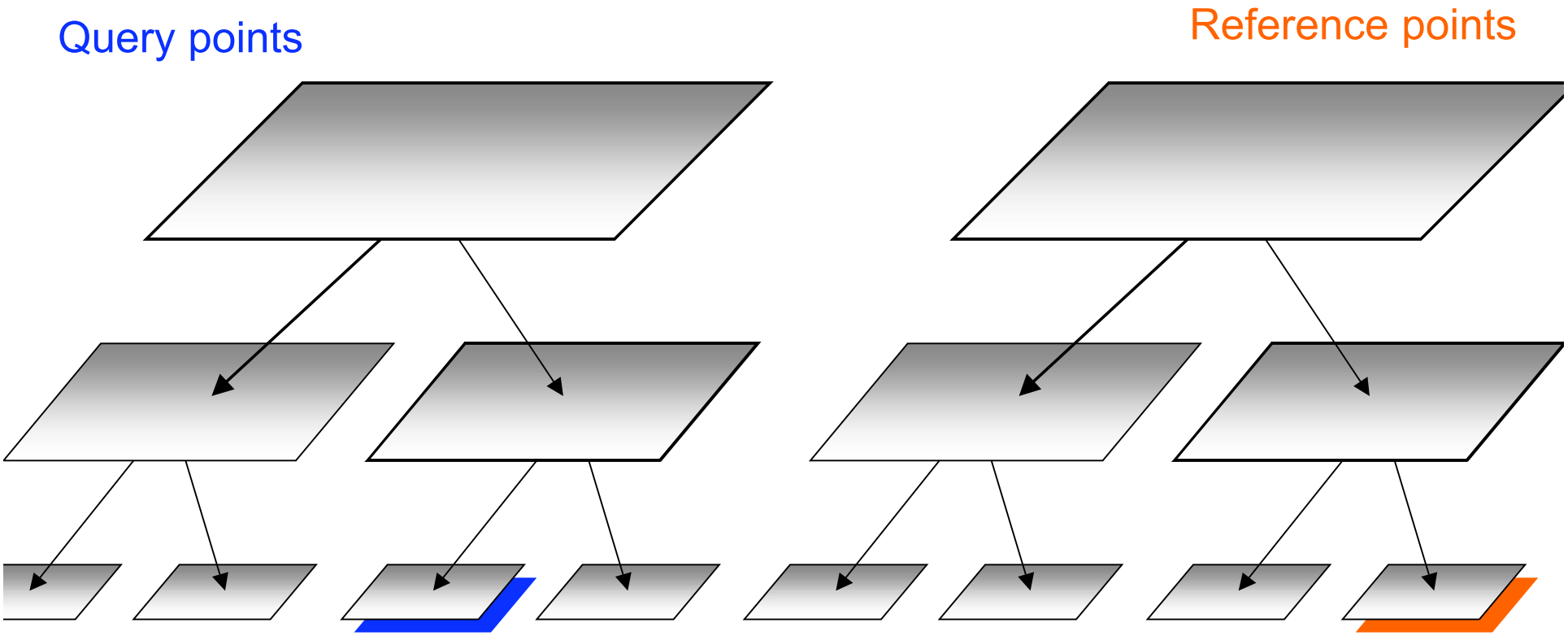
Dual-tree traversal



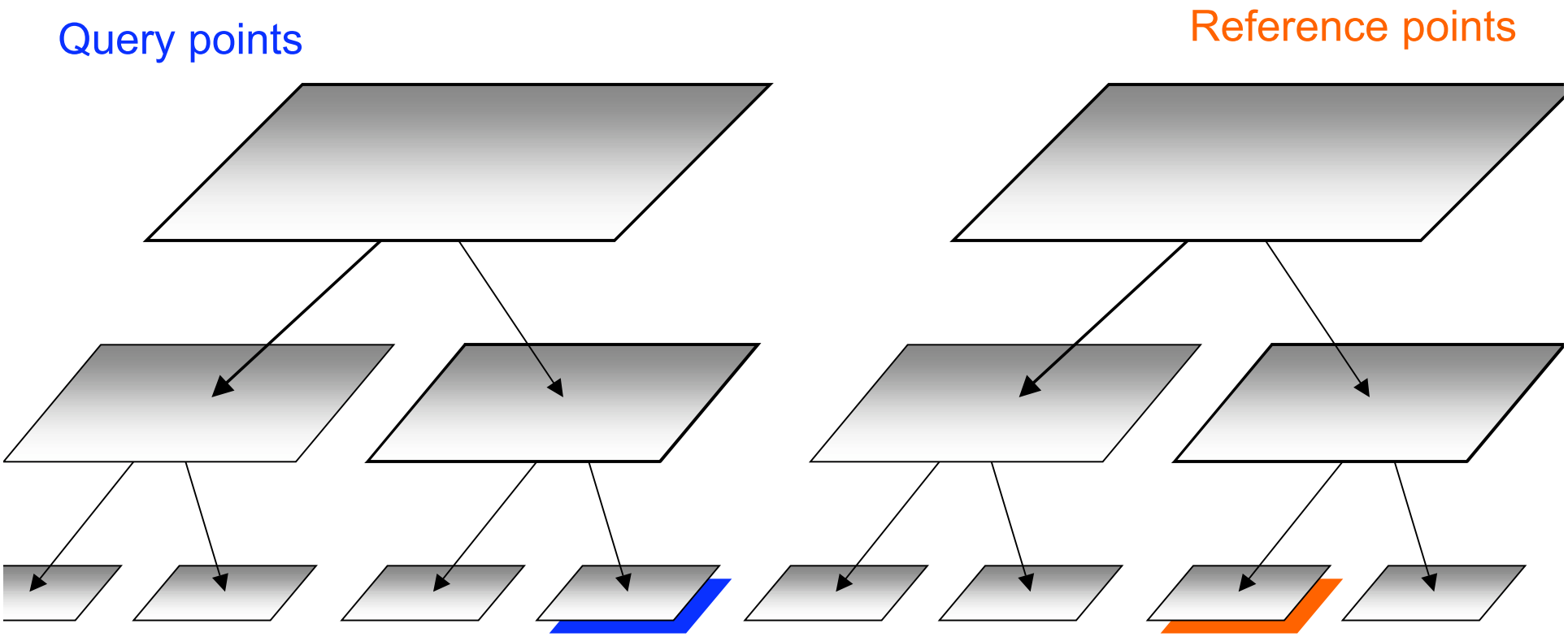
Dual-tree traversal



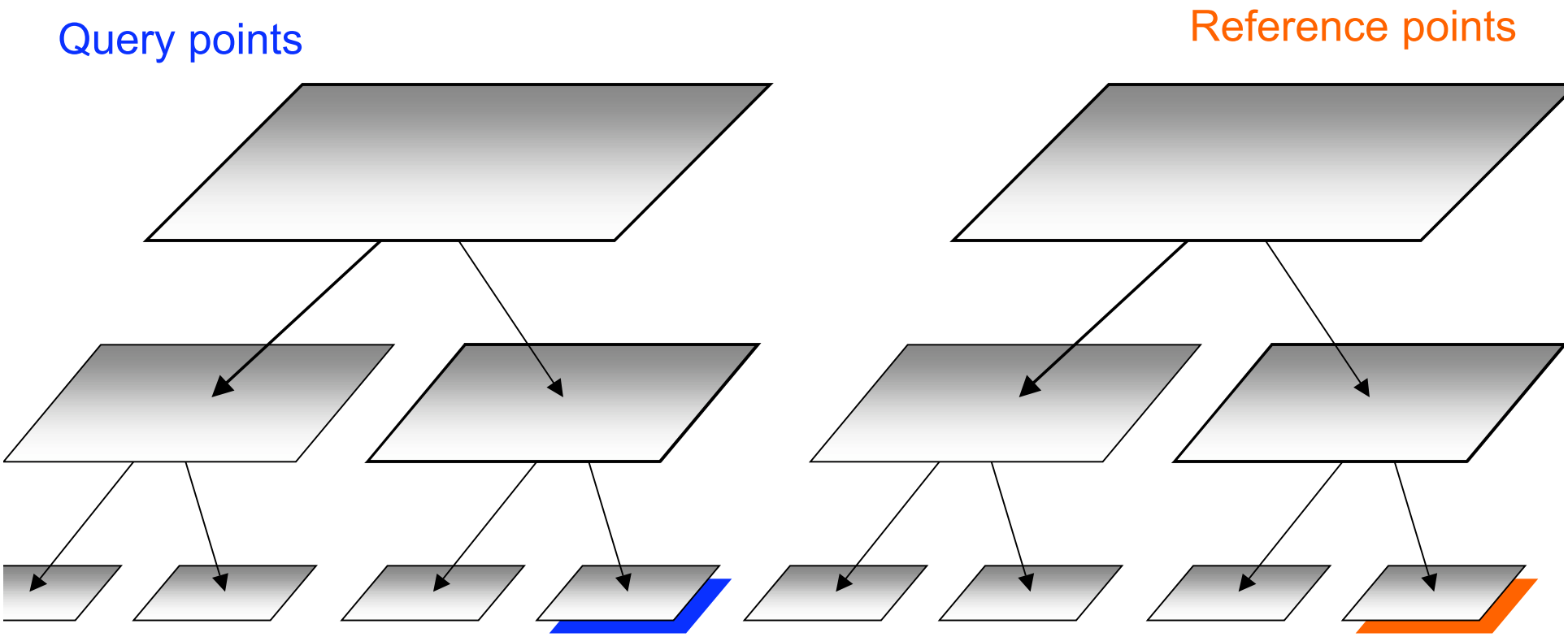
Dual-tree traversal



Dual-tree traversal



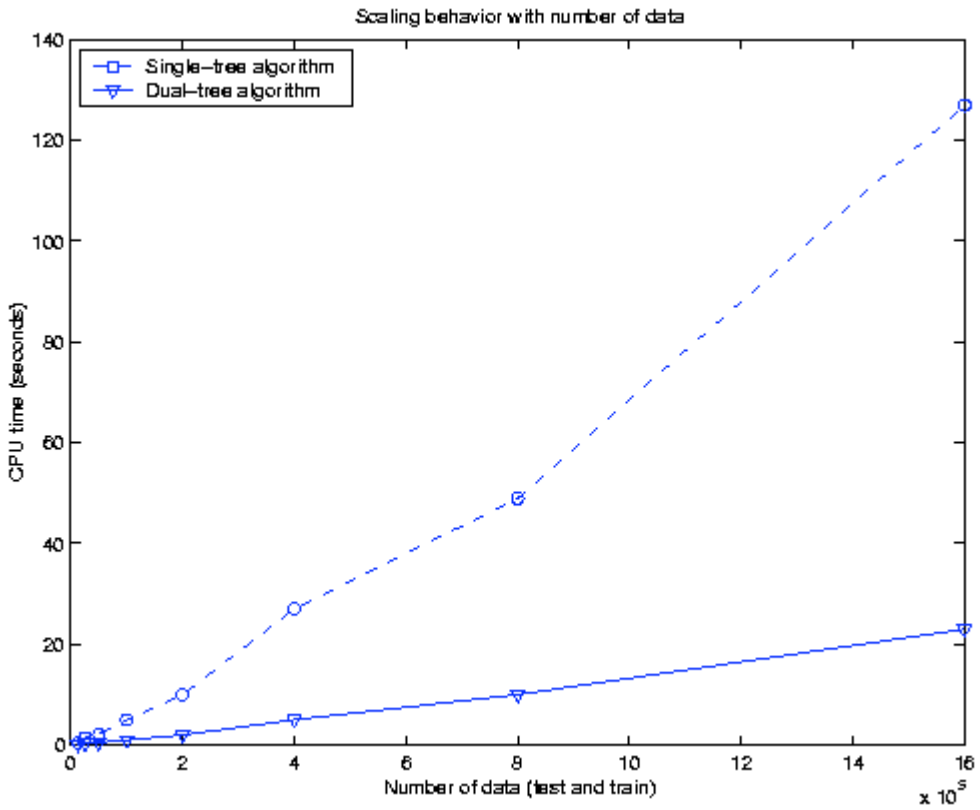
Dual-tree traversal



Speedup Results

N	naïve	dual-tree
12.5K	7	.12
25K	31	.31
50K	123	.46
100K	494	1.0
200K	1976*	2
400K	7904*	5
800K	31616*	10
1.6M	35 hrs	23

5500x



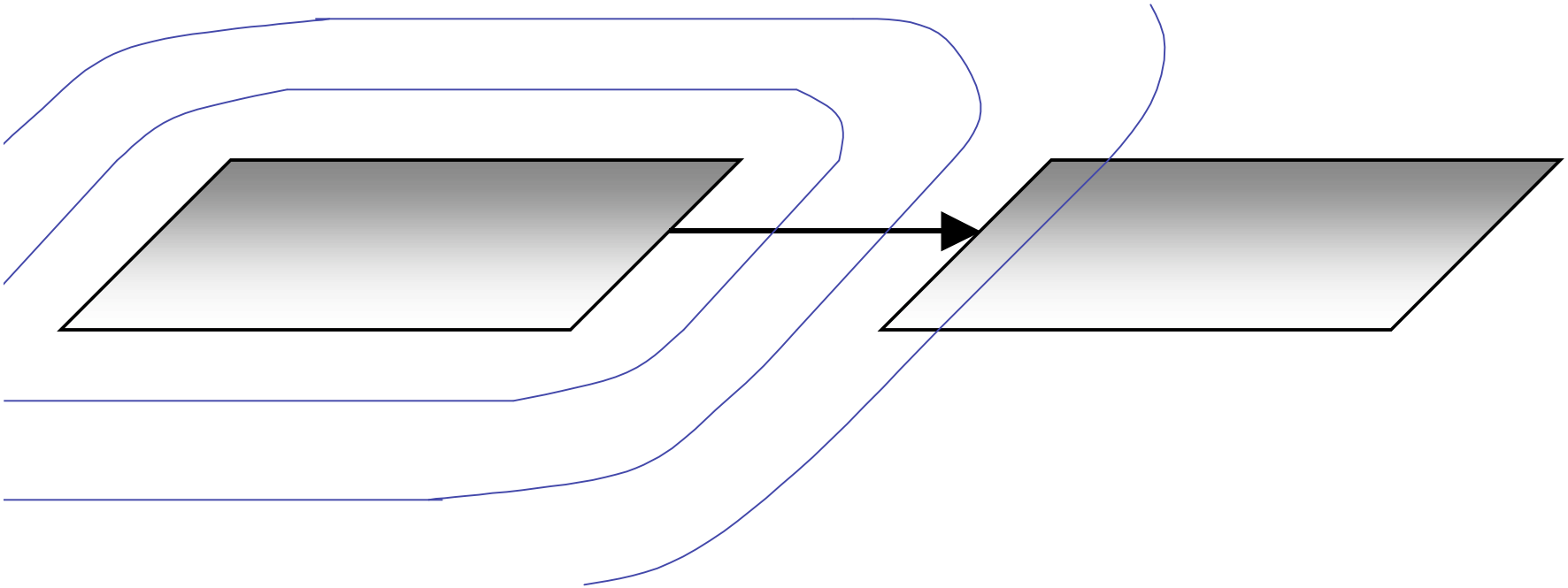
One order-of-magnitude speedup over single-tree at ~2M points

Divide-and-conquer #2

Divide-and-conquer over radii

[Gray-Moore, AI & Statistics 2003]

Exclusion and inclusion,
on multiple radii simultaneously.



$$\min||x-x_i|| < r_1 \Rightarrow \min||x-x_i|| < r_2$$

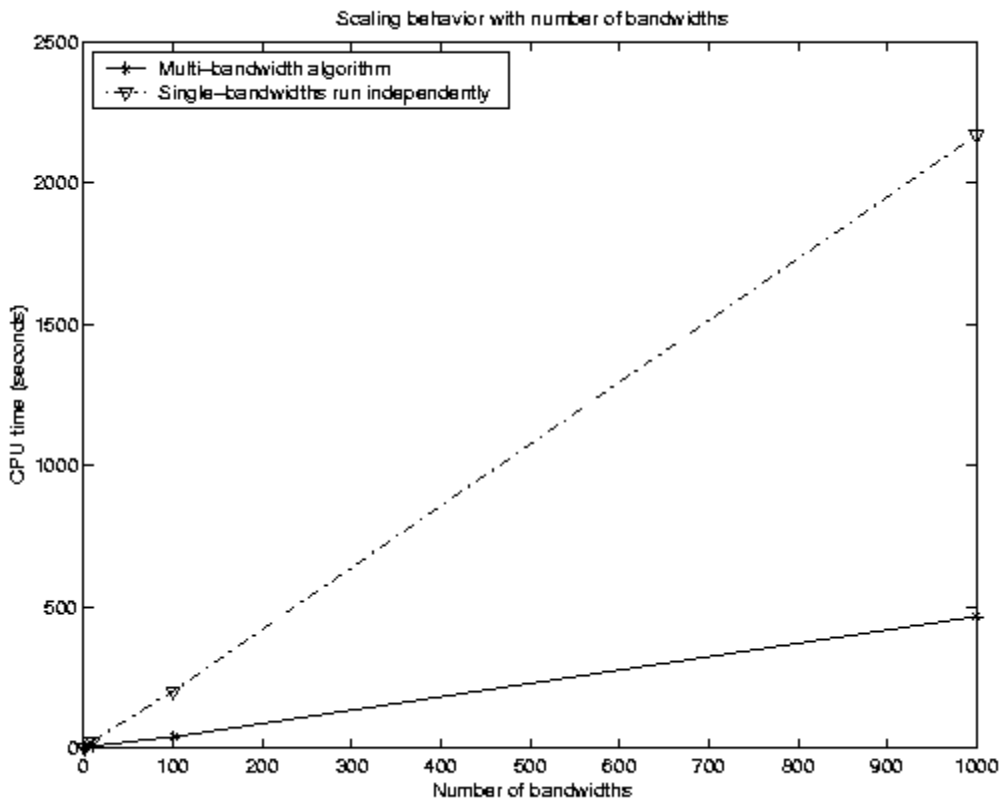
Use binary search to locate critical radius: O(logB)

Doubly recursive algorithm

2pt(*Q,R,{r}*)
{
  r_l=lowerradius*({r},Q,R).*
  r_u=upperradius*({r},Q,R).*
  {r} → *{r}'.*
  update bounds for all radii.

  if leaf(*Q*) and leaf(*R*), **2ptBase**(*Q,R,{r}'*).
  else,
    2pt(*Q*.left,*R*.left,*{r}'*), **2pt**(*Q*.left,*R*.right,*{r}'*).
    2pt(*Q*.right,*R*.left,*{r}'*), **2pt**(*Q*.right,*R*.right,*{r}'*).
}

Speedup Results



One order-of-magnitude speedup over single-radius at ~10,000 radii

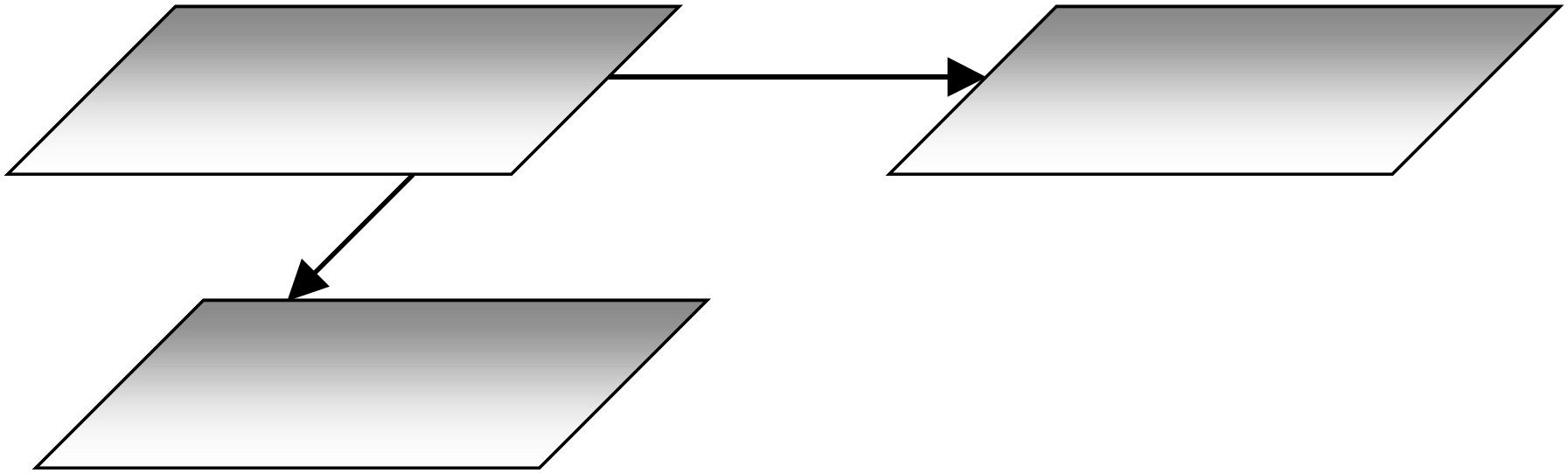
Divide-and-conquer #2

Divide-and-conquer over n>2 sets simultaneously

[Gray-Moore 2000, Gray 2003]

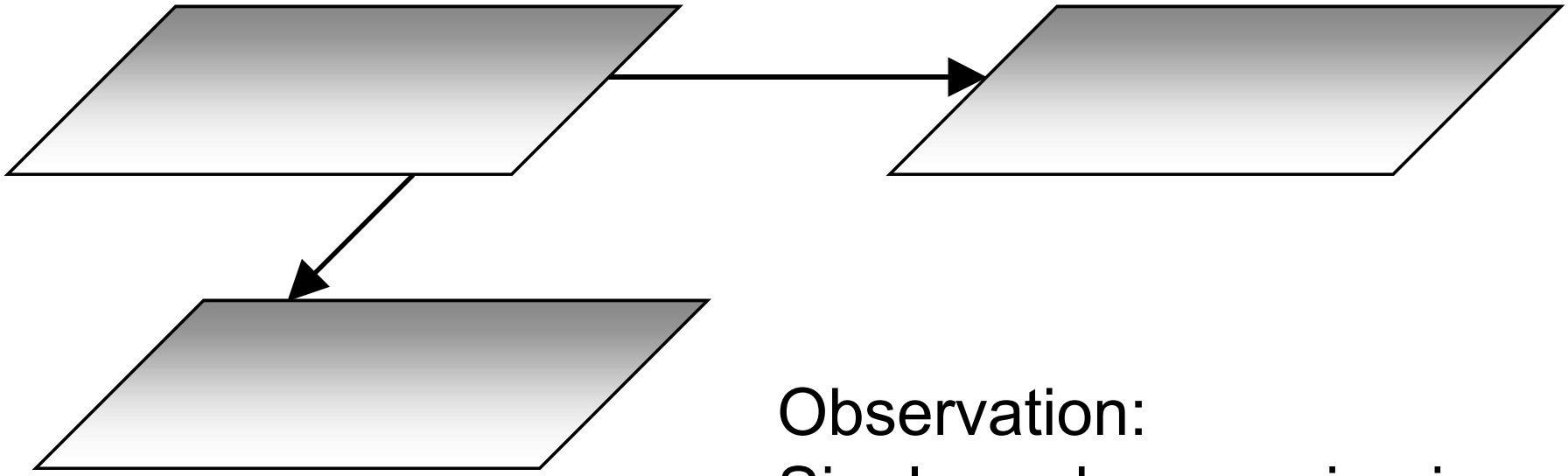
How do to *n*>2?

Use *n*-node recursion. *Check constraints on n-tuple of nodes*.



How do to *n*>2?

Use *n*-node recursion. *Check constraints on n-tuple of nodes*.



Observation:
Single-node recursion is equivalent

npt$(Q_1, Q_2, \ldots, Q_n)$
{
  C_{max}= maxtuples$(Q_1, Q_2, \ldots Q_n)$. if C_{max}=0, return.
  for i=1…n, j=i+1…n,
    S_{ij}=testpair(Q_i, Q_j).

  if S contains an exclusion, C_{hi}=C_{hi}-C_{max}.
  if S is all inclusions, C_{lo}=C_{lo}+C_{max}.
  if S is all leaves,
    nptBase$(Q_1, Q_2, \ldots, Q_n)$.
  else
    Q^*=choosenode$(Q_1, Q_2, \ldots, Q_n)$.
    npt$(Q_1, \ldots, Q^*.left, \ldots, Q_n)$.
    npt$(Q_1, \ldots, Q^*.right, \ldots, Q_n)$.
}

Anytime bounds

Notice:

We maintain strict lower and upper bounds on the count throughout the computation.

Divide-and-conquer #3

Combinatorial constraints

Don’t count permutations of same points more than once

- Number nodes in depth-first order
- Enforce depth-first ordering
  - in maxtuples()
  - in **nptBase**()
- These become recursive

Theoretical analysis

Theorem:
X ~ homogeneous isotropic Poisson, *r* s.t. $E[c_r(x)]=m$

➔ $E[T(N)] = O(N^{\log n})$

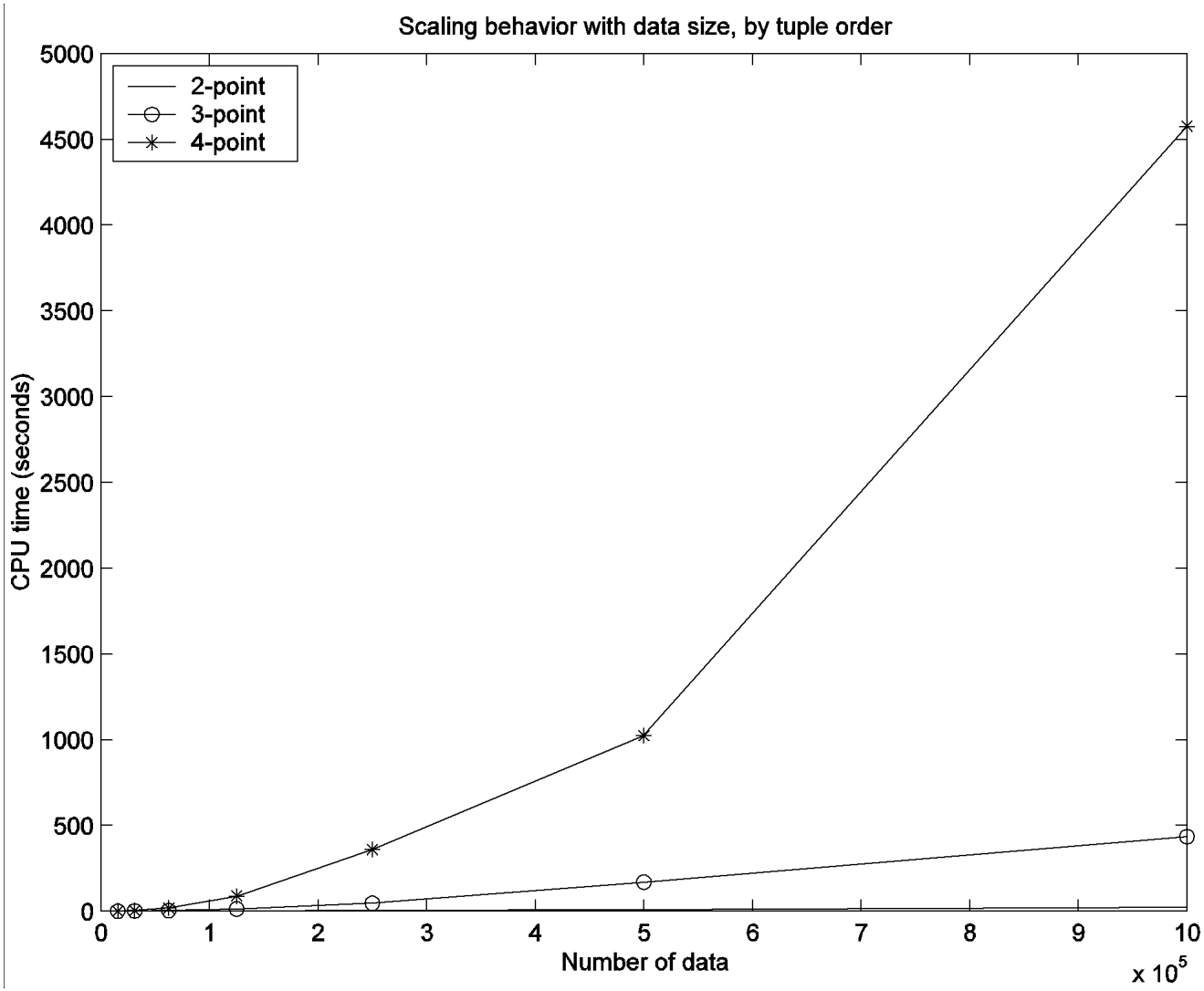
Proof sketch:
Recurrence analysis and a packing bound
$T(N)=nT(N/2) + O(1), T(1)=O(1).$

Validation of theoretical analysis

n=2: ***O(N)***

n=3: ***O(N^{log3})***

n=4: ***O(N^2)***



3-point performance

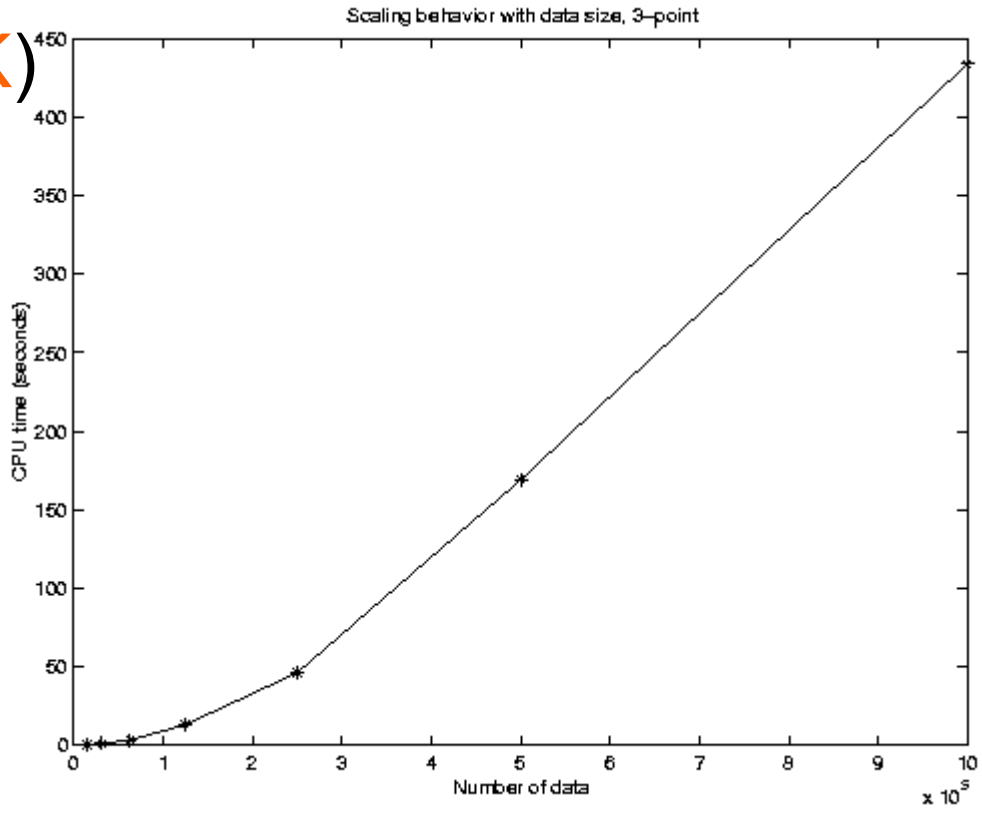
(biggest previous: 20K)

VIRGO
simulation data

N = 75,000,000

naïve: 5×10^9

multi-tree: **55 sec.**



But…

Depends on r^{D-1}.
Slow for large radii.

Example:
VIRGO data
N = 75,000,000

naïve: 5×10^9
multi-tree:
large h: 24 hrs

Let's develop a method for large radii.

Outline:

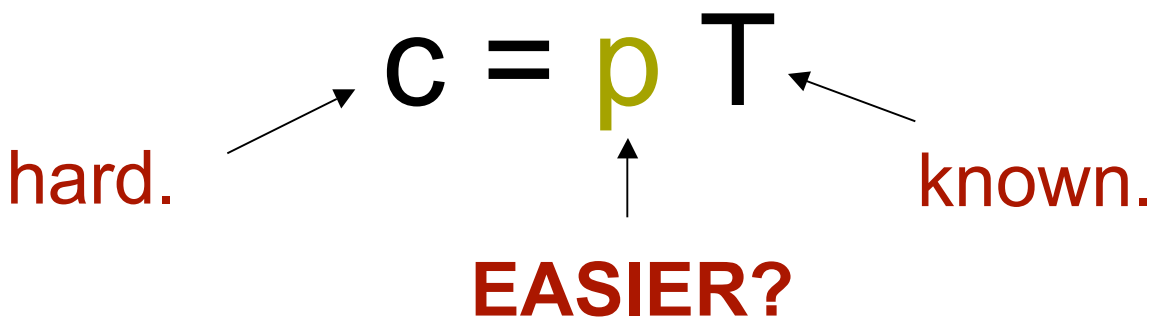
1. Some cosmological questions

2. Spatial correlations

3. Divide-and-conquer in real-space

4. Multi-tree algorithm, exact

5. **Multi-tree Monte Carlo**



→ Monte Carlo?

$$\hat{p} \pm z_{\varepsilon/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{S}}$$

No dependence on N!

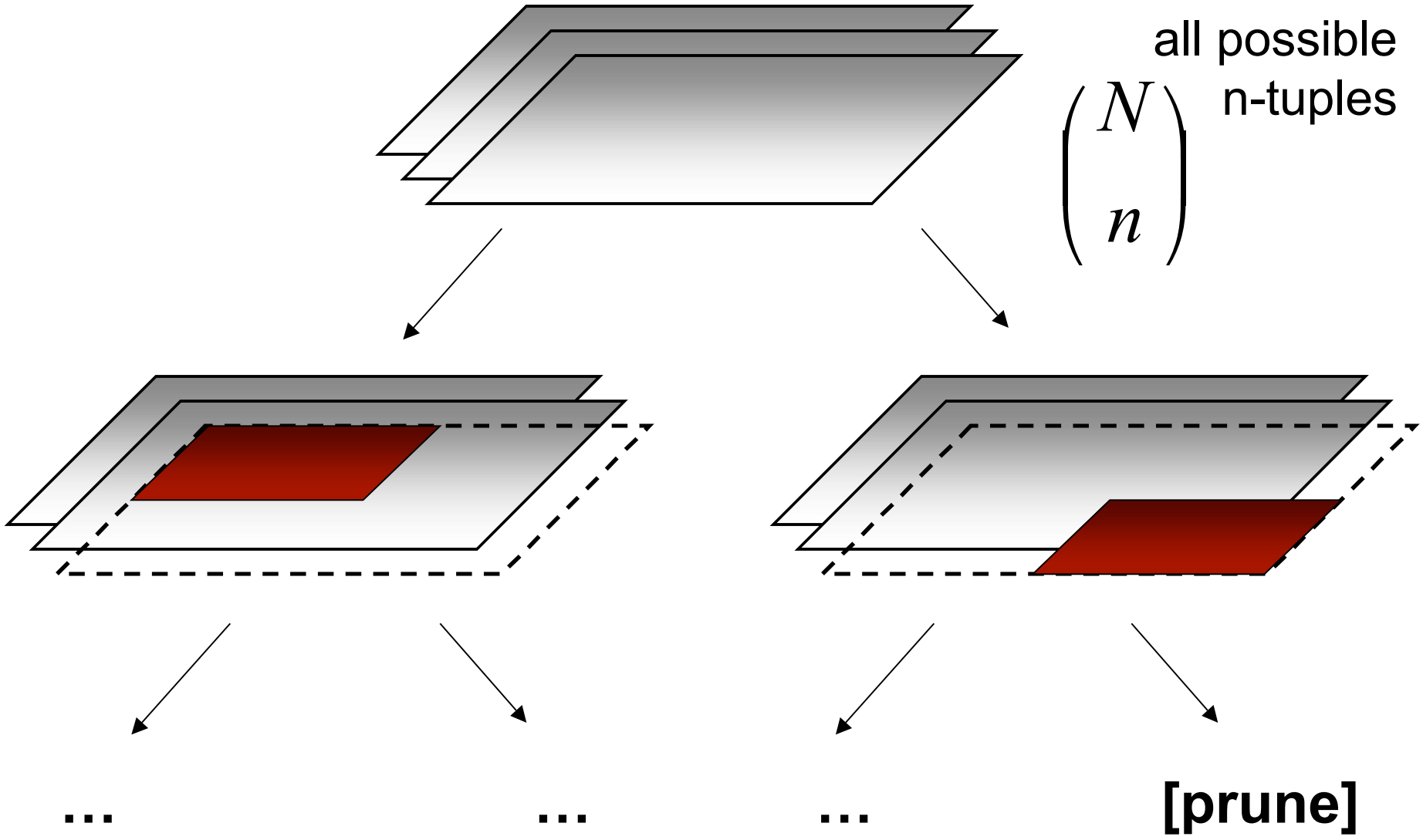
… but depends on p.

Basic idea:

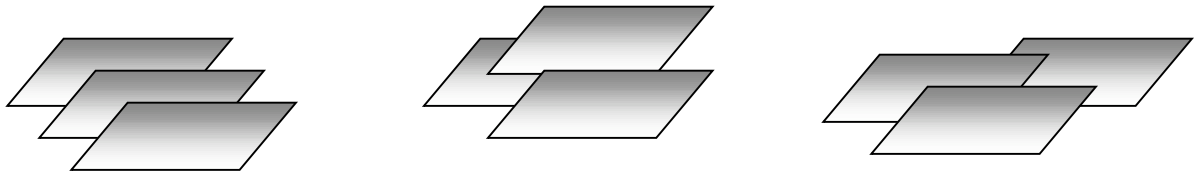
1. **Remove some junk**
(Run exact algorithm for a while)
→ **make p larger**

2. **Sample from the rest**

We get disjoint sets from the recursion tree



Now do stratified sampling



$$T_1 + T_2 + T_3 = T$$

$$\frac{T_1}{T}\hat{p}_1 + \frac{T_2}{T}\hat{p}_2 + \frac{T_3}{T}\hat{p}_3 = \hat{p}$$

$$\left(\frac{T_1}{T}\right)^2 \hat{\sigma}_1^{\,2} + \left(\frac{T_2}{T}\right)^2 \hat{\sigma}_2^{\,2} + \left(\frac{T_3}{T}\right)^2 \hat{\sigma}_3^{\,2} = \hat{\sigma}^2$$

Note 1: Wald interval not perfect

Poor coverage for small p

Agresti-Coull not usable here; use min. p

Note 2: Adaptive Neyman sampling

$$S_k^{opt} = \frac{\left(\frac{T_k}{T}\right)^2 \sigma_k^2}{\sigma}$$

→ Why not use $\hat{\sigma}_k^2$?

→ Update allocations periodically

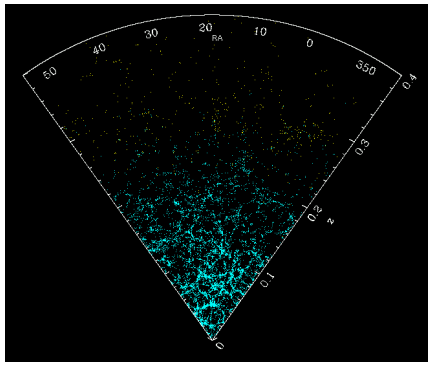
Speedup Results

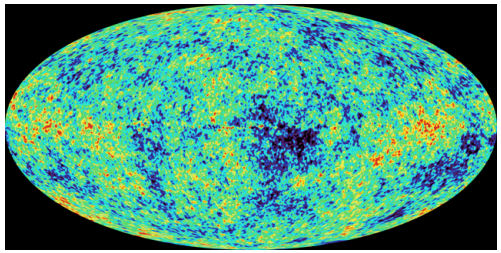
Example:
VIRGO data
N = 75,000,000

naïve: 5×10^9
multi-tree:
large h: 24 hrs

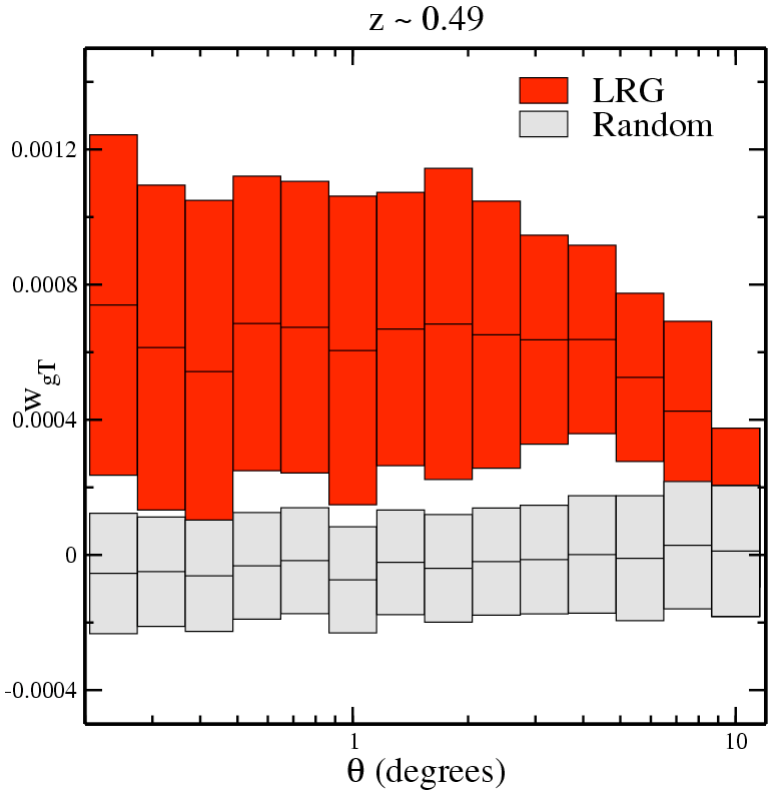
multi-tree monte carlo:
99% confidence:
96 sec

Question: Does dark energy exist?





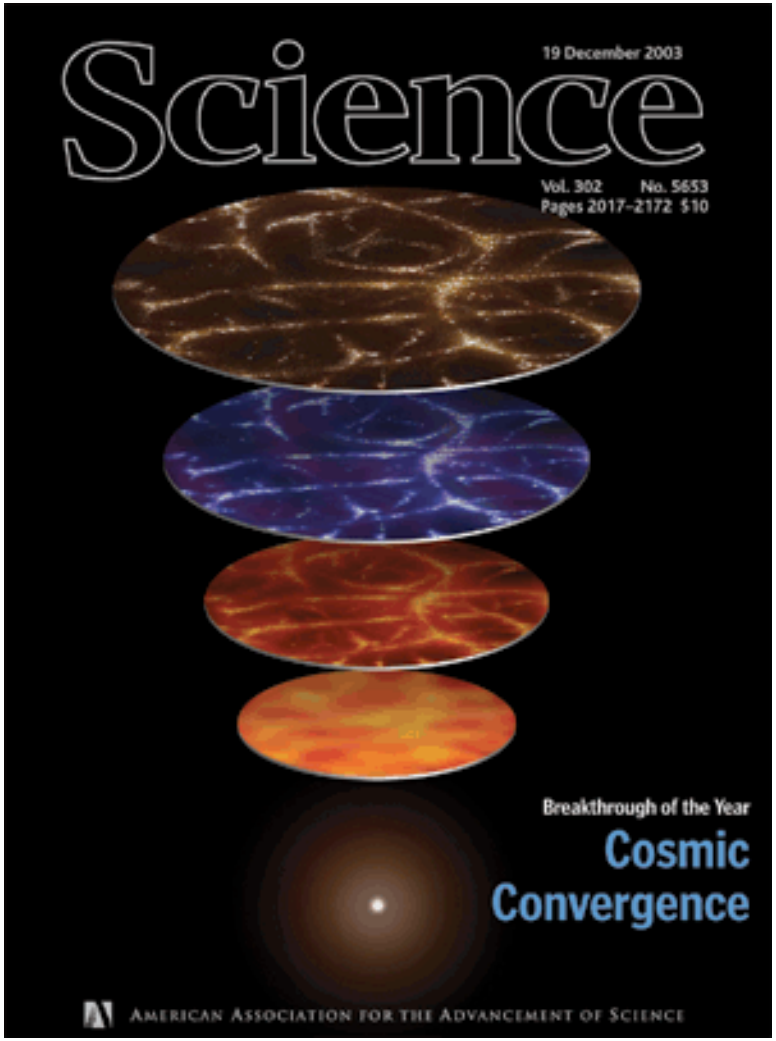
Do we see the ISW Effect?



2-point on 2,000,000 galaxies and WMAP pixels

First (second) physical evidence of dark energy.

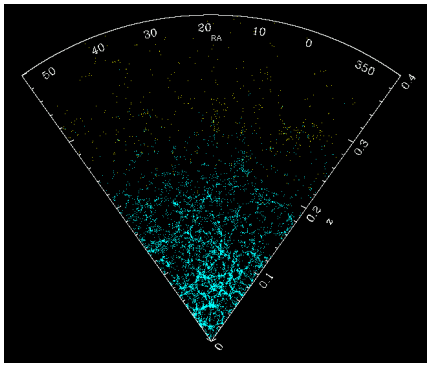
[Scranton et al., PRL 2003 submitted]

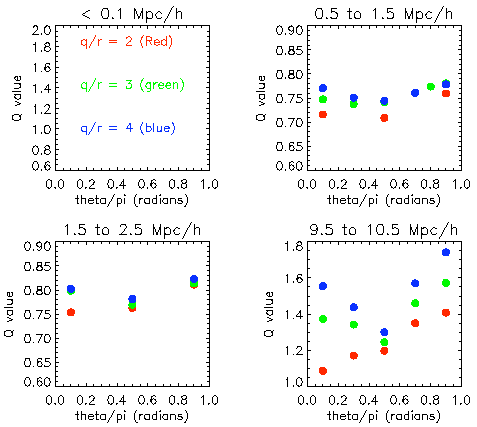


Science
#1 Breakthrough of 2003

Bob Nichol on David Letterman show
July 2003

Question: Is the universe Gaussian?



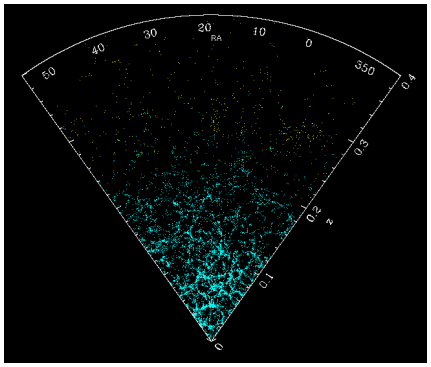


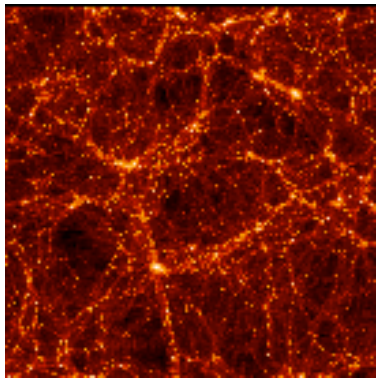
3-point on 130,000 galaxies, 1.3M random

Most comprehensive third-order statistics on universe to date.

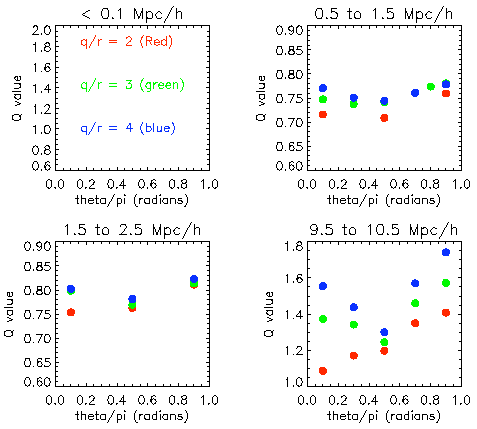
[Nichol et al., ApJL 2004 submitted]

Question: Does the model fit the data?





Same?



3-point on 130,000 galaxies, 1.3M random

Coming: 3-point on VIRGO

→ constrain cosmological parameters

Extensions

- Weighted points (**done**).
- Interval constraints (**done**).
- Projected-space correlations (**done**).
- Marked correlations (when someone needs it).

These are examples of…

Generalized N-body problems

All-NN: $\{\forall, \arg\min, \delta, \cdot\}$

2-point: $\{\Sigma, \Sigma, I_r(\delta), w\}$

3-point: $\{\Sigma, \Sigma, \Sigma, I_R(\delta), w\}$

KDE: $\{\forall, \Sigma, K_r(\delta), \cdot; \{r\}\}$

SPH: $\{\forall, \Sigma, K_r(\delta), w; t\}$

General theory and toolkit for designing algorithms for such problems

Next?

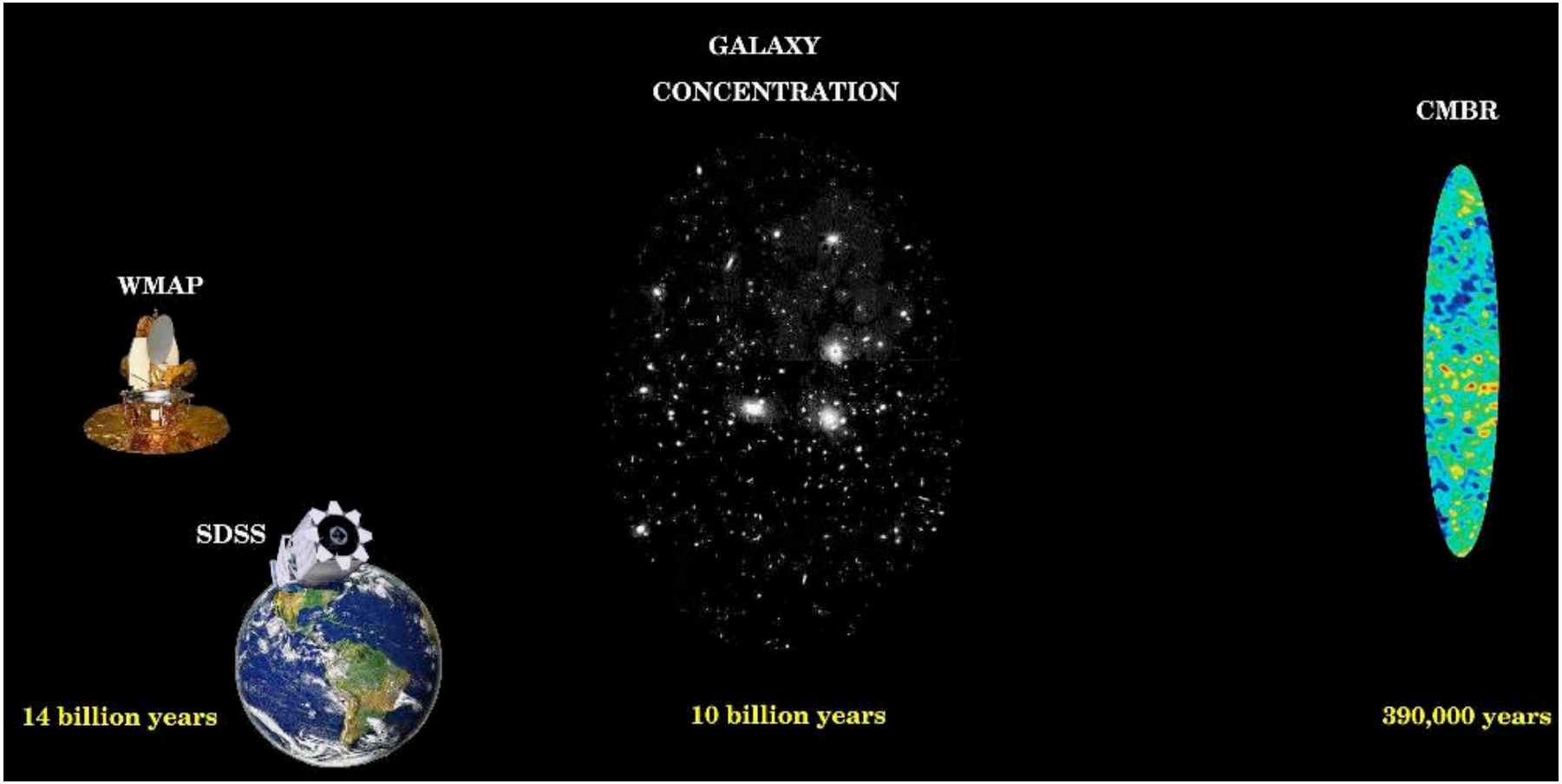
- Intrinsic dimension estimation: edge-corrected, fast.
- BBGKY hierarchy?
- Tell me!

For more info, or the code:

- email me: agray@cs.cmu.edu
- http://www.cs.cmu.edu/~agray

Please tell me your problems!

Experimental Setup



What Are We Measuring?