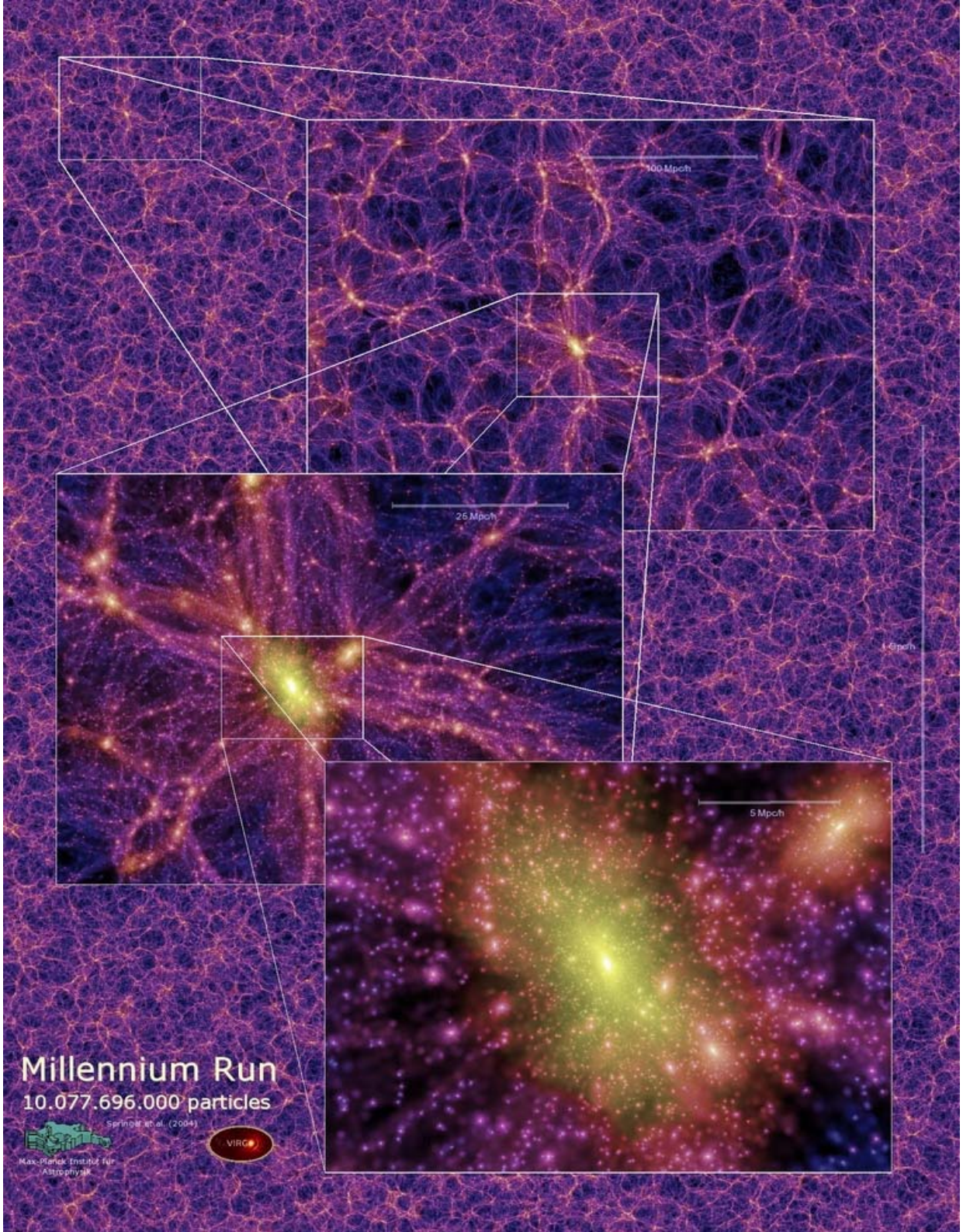
Millennium database

Using a relational database for analysing simulations and SAM results

With contributions from

- JHU : Alex Szalay
- MPA:
  - –Jeremy Blaizot,
  - –Ben Panter
  - –Guo Qi
  - –Volker Springel
  - –Simon White
  - –Vivienne Wild



Agenda

- Millennium simulation products
- Motivation for RDB approach
- Merger trees in RDB
- GAVO online query tool demos
- Plans
- Synthetic spectra: query-by-example using PCA

Millennium simulation

- 10 billion dark matter particles
- 500 Mpc box
- 1yr WMAP parameters
- 300000 CPU hours
- 25Tb stored output (63 snapshots)
- Density field binned in 256^3 grid cells
- 750000000 (sub)halos
- 20 Million halo merger trees
- SAM galaxy models: 1 billion galaxies

Science questions:

1.Return the galaxies residing in halos of mass between 10^13 and 10^14 solar masses.
2.Return the galaxy content at z=3 of the progenitors of a halo identified at z=0
3.Return all the galaxies within a sphere of radius 3Mpc around a particular halo
4.Return the complete halo merger tree for a halo identified at z=0
5.Find positions and velocities for all galaxies at redshift zero with B-luminosity, colour and bulge-to-disk ratio within given intervals.
6.Find properties of all galaxies in haloes of mass 1014 at redshift 1 which have had a major merger (mass-ratio < 4:1) since redshift 1.5.**
7.Find all the z=3 progenitors of z=0 red ellipticals (i.e. B-V>0.8 B/T > 0.5)
8.Find the descendents at z=1 of all LBG's (i.e. galaxies with SFR>10 Msun/yr) at z=3
9.Make a list of all haloes at z=3 which contain a galaxy of mass >109 Msun which is a progenitor of BCG's in z=0 cluster of mass >10**14.5**
10.Find all z=3 galaxies which have NO z=0 descendant.
11.Return the complete galaxy merging history for a given z=0 galaxy.
12.Find all the z=2 galaxies which were within 1Mpc of a LBG (i.e. SFR>10Msun/yr) at some previous redshift.
13.Find the multiplicity function of halos depending on their environment (overdensity of density field smoothed on certain scale)
14.Find the dependency of halo formation times on environment (“Gao-effect”)

Why use RDBM ?

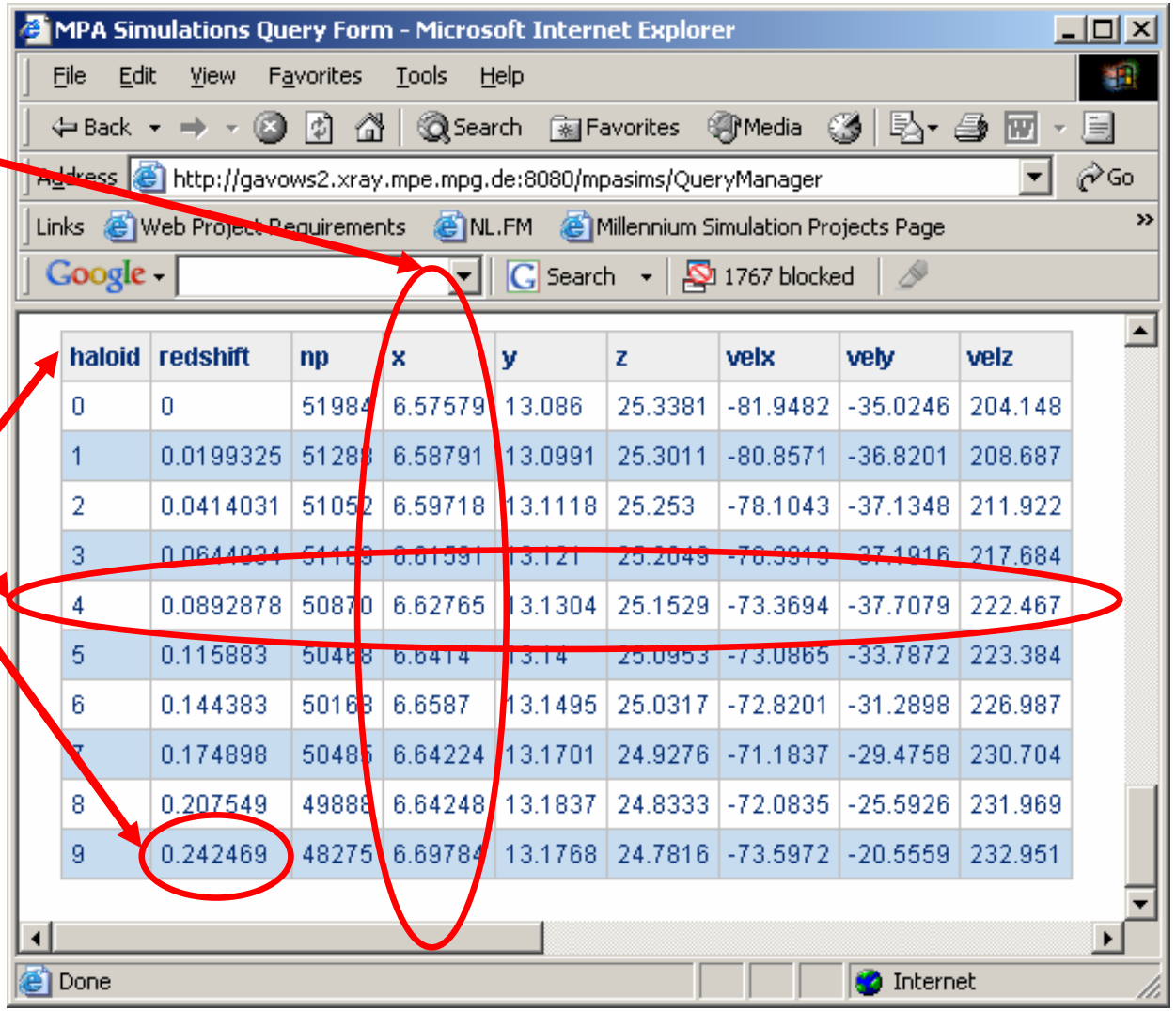
- encapsulation of data in terms of logical structure, no need to know about internals of data storage
- standard query language for finding information
- advanced query optimizers (indexes, clustering)
- transparent internal parallelization

- transactionally safe remote access to multiple users at same time
- security mechanisms
- standardized, transactional support for inserts/updates/deletes
- maintenance (backup, mirroring, etc)

- **forces one to think carefully about data structure**
- **speeds up path from science question to answer**
- **facilitates communication**

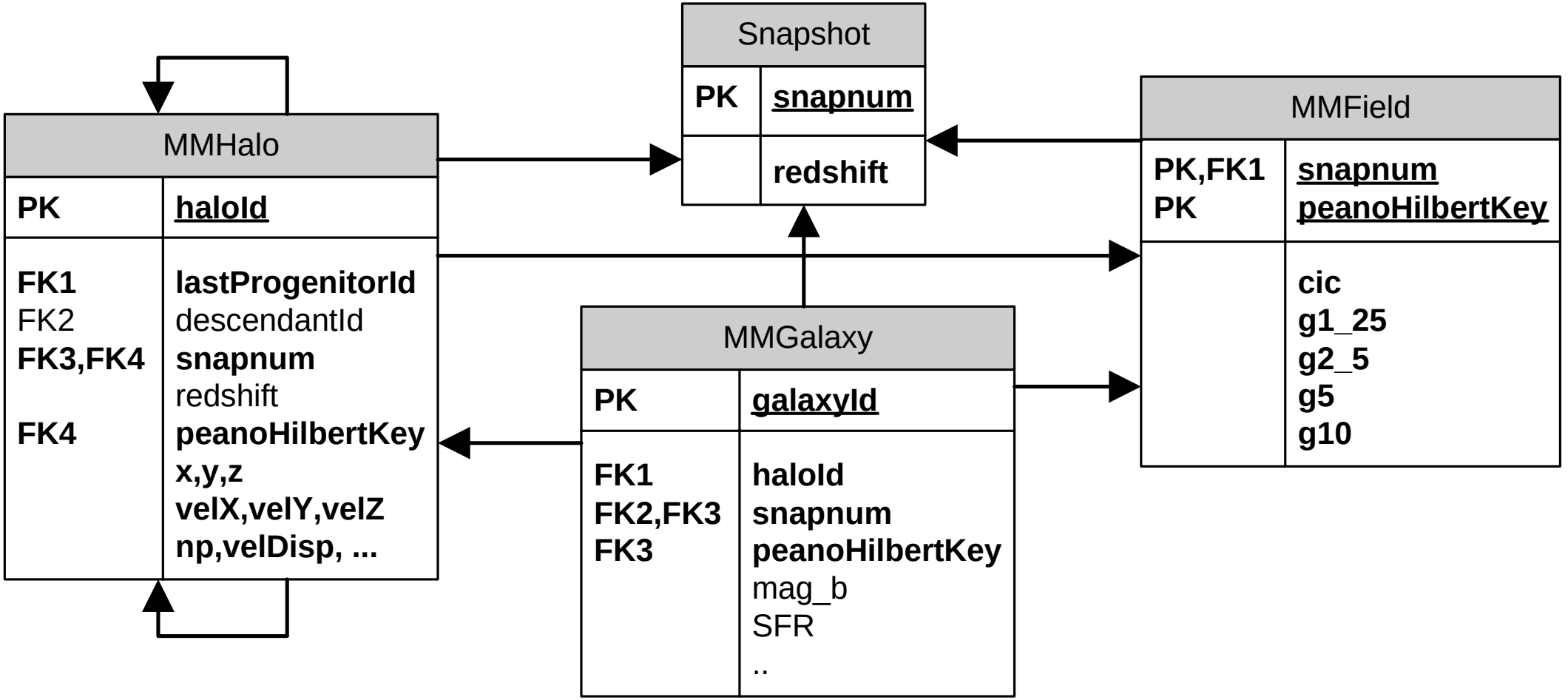
RDBM stores data in Tables

- Column: name, simple datatype
- Row: an instance of a relation
- Value: a cell in a row
- Primary key: unique identifier for rows, built up from >=1 columns

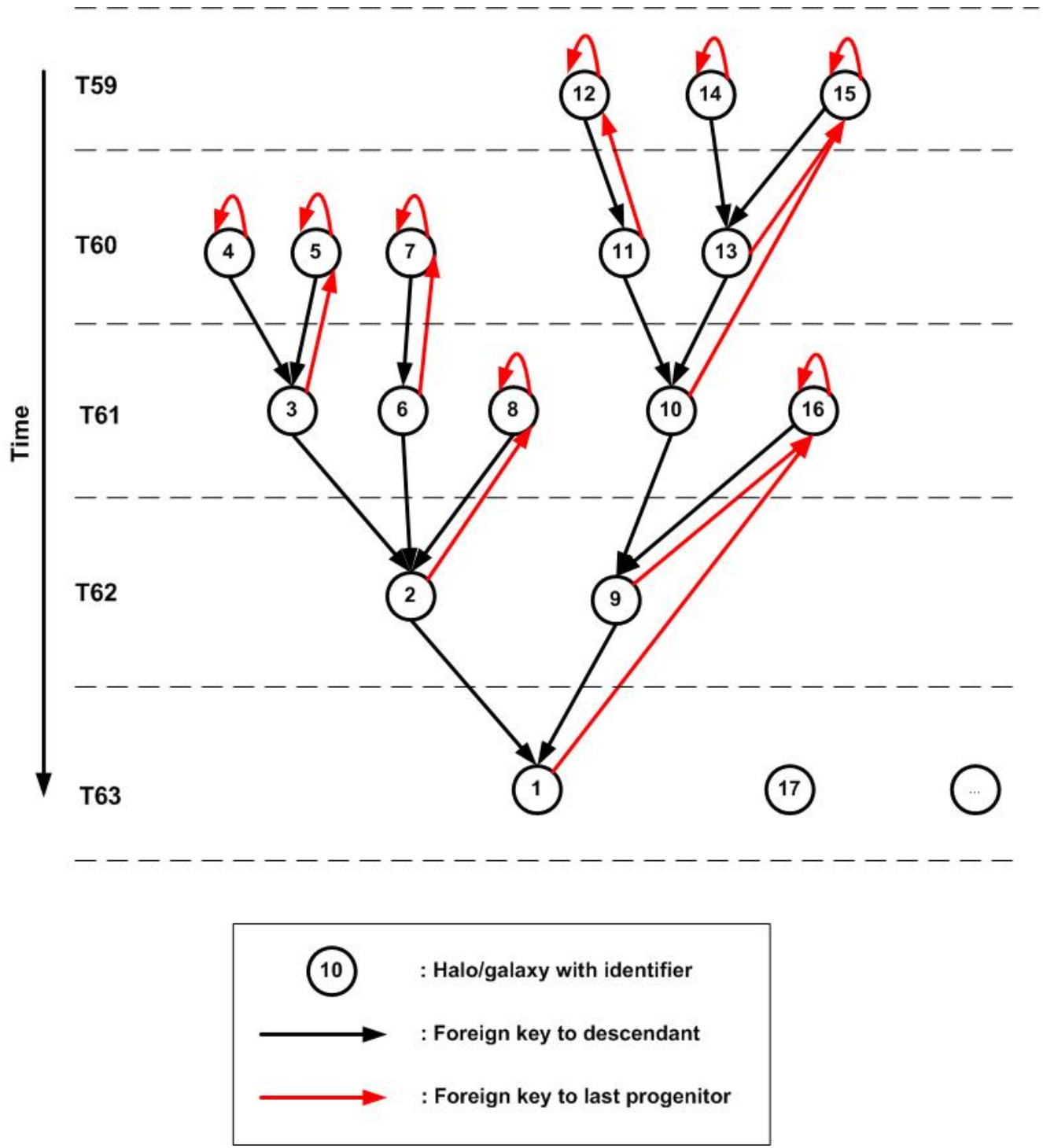


haloid	redshift	np	x	y	z	velx	vely	velz
0	0	51984	6.57579	13.086	25.3381	-81.9482	-35.0246	204.148
1	0.0199325	51283	6.58791	13.0991	25.3011	-80.8571	-36.8201	208.687
2	0.0414031	51052	6.59718	13.1118	25.253	-78.1043	-37.1348	211.922
3	0.0644934	51163	6.61591	13.121	25.2049	-76.3919	-37.1916	217.684
4	0.0892878	50870	6.62765	13.1304	25.1529	-73.3694	-37.7079	222.467
5	0.115883	50468	6.6414	13.14	25.0953	-73.0865	-33.7872	223.384
6	0.144383	50163	6.6587	13.1495	25.0317	-72.8201	-31.2898	226.987
7	0.174898	50485	6.64224	13.1701	24.9276	-71.1837	-29.4758	230.704
8	0.207549	49888	6.64248	13.1837	24.8333	-72.0835	-25.5926	231.969
9	0.242469	48275	6.69784	13.1768	24.7816	-73.5972	-20.5559	232.951

Millennium schema (schematic)



Merger trees



Millennium databases

- SQLServer2000 on 4 processor Opteron with 10Tb RAID
- milli-Millennium
  - Galaxies added: merger trees, links to their parent halos
  - Density field at various smoothings
  - Updated web site (demo)
- Millennium subset
  - Subset (~2%, 10x milli-Mil) of halo and galaxy trees
  - Z=0 density field
- Millennium
  - Halo trees in database (750000000, proprietary)
  - SAM galaxies in progress (10^9 galaxies)
  - Density fields at all Z will be added: 1056964608 rows
- Durham
  - milli_Millennium mirror (Postgres)
  - Durham halo tree and galaxy catalogues

Tools

- GAVO online query tool
  - Query results temporarily buffered on server: memory
- Streaming queries: faster, less limited (only timeout)
- IDL (with Ben Panter)
  - wget –http-user=*** --http-password=*** -O localfile.csv http://www.g-vo.org/sdssdr3/DBQueryStream?SQL=select * from moped..agebin
  - GUI asking for username/password
  - Interprets CSV stream, turned into IDL components
- TOPCAT with GAVO SQL plugin

Demo 1

- Return merger tree for halo identified at z=0

```
select prog.*
  from halo des,
       halo prog
 where prog.haloId
       between des.haloId and des.lastProgenitorId
   and des.haloId = 5000063000000
```

Demo 2

- Return B-band luminosity function of galaxies residing in halos of mass between 10^13 and 10^14 solar masses

```
select .2*round(5*g.mag_b) as magB,
       count(*) as num
  from MMGalaxy g, MMHalo h
 where g.haloId = h.haloId
   and h.mTopHat between 1000 and 10000
   and h.redshift=0
group by magB
```

Demo 3

- Return the formation time of halos, defined as the maximum time at which it still has a progenitor of greater than half its mass, as function of the matter density in its environment, defined by the matter density smoothed on scale of 10Mpc

```
select zForm, avg(g10) as g10
from MMField f,
 ( select des.haloId, des.phkey,
          max(PROG.redshift) as zForm
     from MMHalo PROG,
          MMHalo DES
    where DES.redshift = 0
      and PROG.haloId between DES.haloId and DES.lastProgenitorId
      and prog.np >= des.np/2
      and des.np between 100 and 200
   group by des.haloId, des.phkey ) t
 where t.phkey = f.phkey
   and f.snapnum=63
group by zForm
```

3. Return the formation time of halos, defined as the maximum time at which it still has a progenitor of greater than half its mass, as function of the matter density in its environment, defined by the matter density smoothed on scale of 10Mpc

```
select zForm, avg(g10) as g10
from MMField f,
 ( select des.haloId, des.phkey,
          max(PROG.redshift) as zForm
     from MMHalo PROG,
          MMHalo DES
    where DES.redshift = 0
      and PROG.haloId between DES.haloId and DES.lastProgenitorId
      and prog.np >= des.np/2
      and des.np between 100 and 200
   group by des.haloId, des.phkey ) t
 where t.phkey = f.phkey
   and f.snapnum=63
group by zForm
```

Plans

- Comparison to standard, file based approach
- Load full Millennium in RDB
- Support “MyDB” for SAGF producers
  SAM-online, working on results of queries
- More products:
  - Mock SDSS
  - Light cones online
  - Spectra + query-by-example
- Use for science

Theory VO: spectra

- Combine theory and observations
- Example: query-by-example on theory spectra
- Find similar spectra, from these the actual galaxy formation history
- Chi-squared on all stored spectra ? Slow, requires storing all of them
- Idea (not original): use PCA to compress data

PCA

- Need training sample of theory spectra to create eigenspectra
- Project all spectra
- Store PCA amplitudes in DB
- Provide web service:
  - Upload (observational) spectrum (IVOA SSA/SED)
  - Project onto theory eigenspectra
  - Use amplitudes as parameters in query for “nearby” amplitudes
  - Return corresponding theory spectra
  - Return corresponding galaxy formation histories, or their halos, or their environment …

Issues

- Dealing with errors, gaps: “gappy PCA” (Connolly & Szalay)
- Normalization:
  - incoming spectrum in general from very different dataset, needs common normalization
  - Incoming set will have gaps, errors
  - Ad hoc normalization possible (and works quite good)
- Indexing of complex multi-dimensional point set for quick nearest k neigbours search (Voronoi ? See Laszlo‘s work)

Normalized gappy PCA

- Fit normalization factor at same time as PCA amplitudes. Model:

$$\boldsymbol{F}_\lambda = N(\boldsymbol{c}_\lambda + \sum_{i=1}^{n_{pc}} a_i \boldsymbol{e}_{i,\lambda})$$

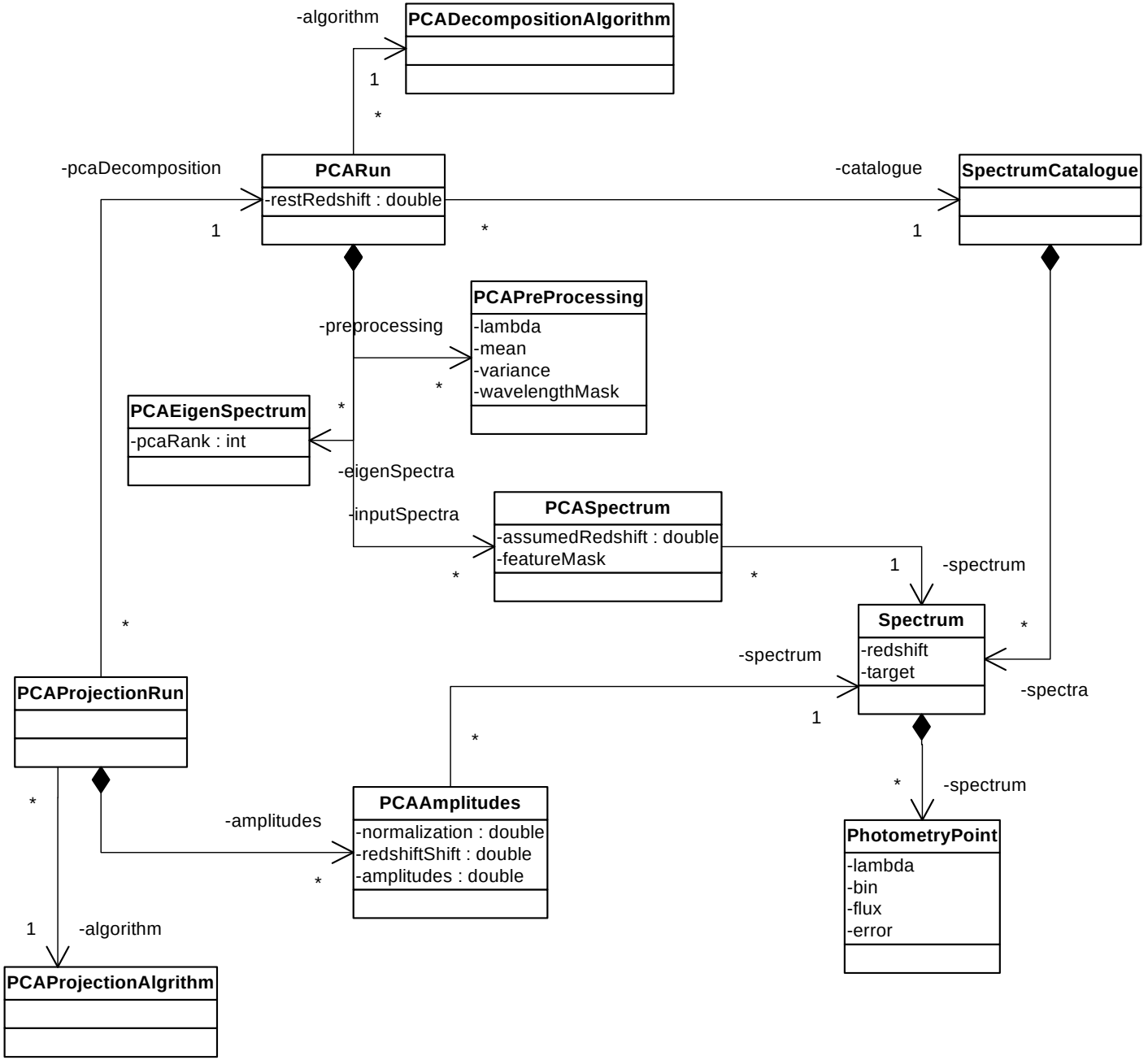
- Minimize (over a_i *and* N) :

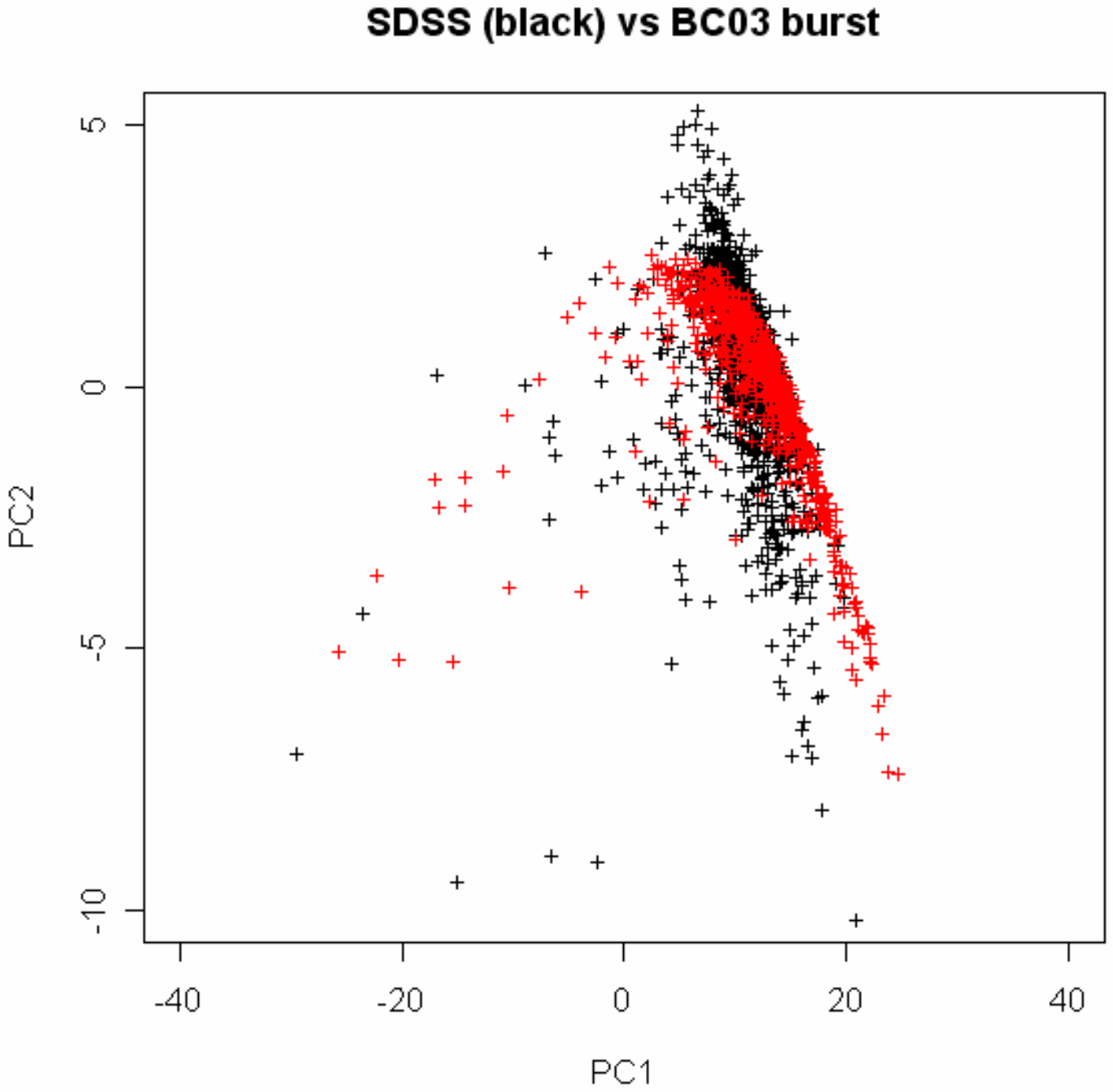
$$\chi^2 = \sum_\lambda \boldsymbol{w}_\lambda (\boldsymbol{F}_\lambda - N(\boldsymbol{c}_\lambda + \sum_i a_i \boldsymbol{e}_{i,\lambda}))^2$$

So far

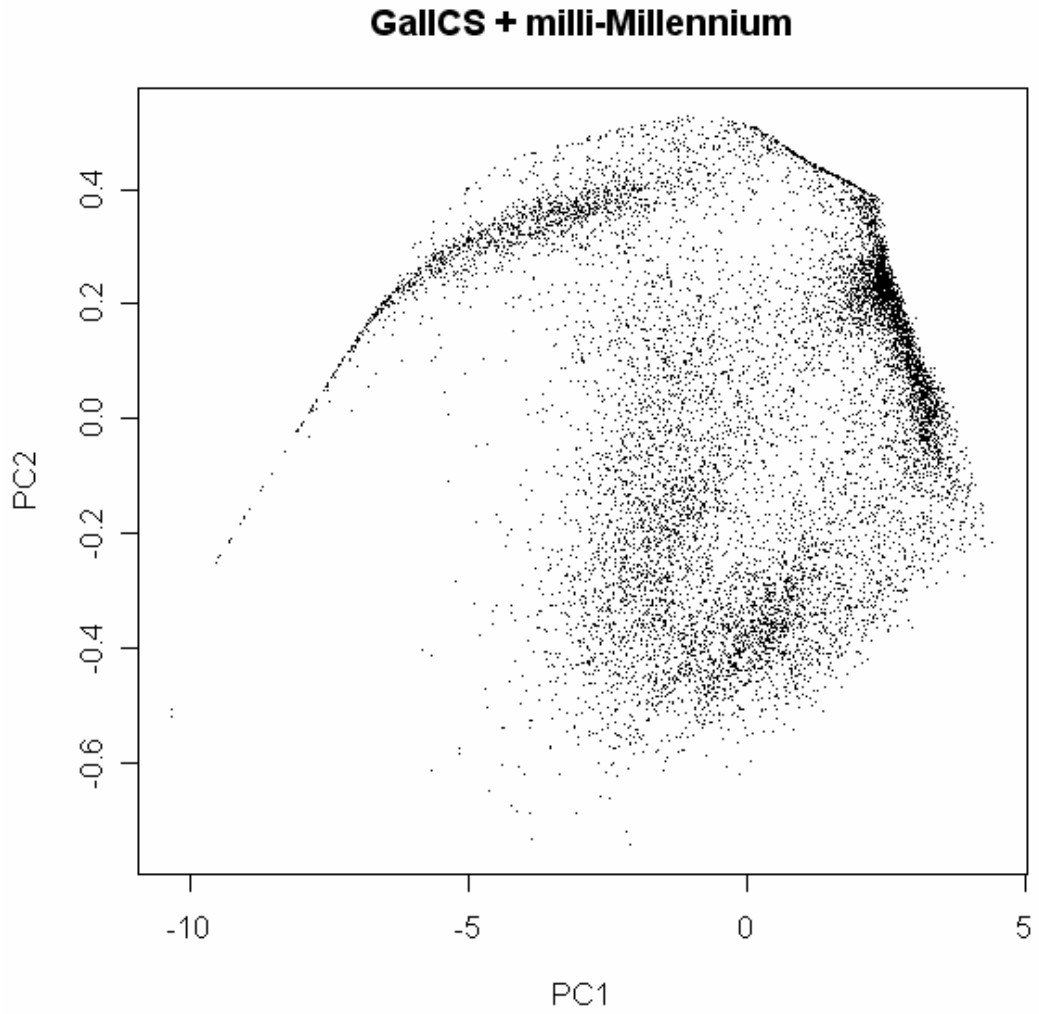
- Ran PCA on BC03 stochastic bursts (Vivienne Wild)
- On first GalICS+milli-Millennium spectra (Jeremy Blaizot)
- Projected SDSS spectra on both
- Defined a PCA data model/schema
- Stored PCAs in database
- TOPCAT access

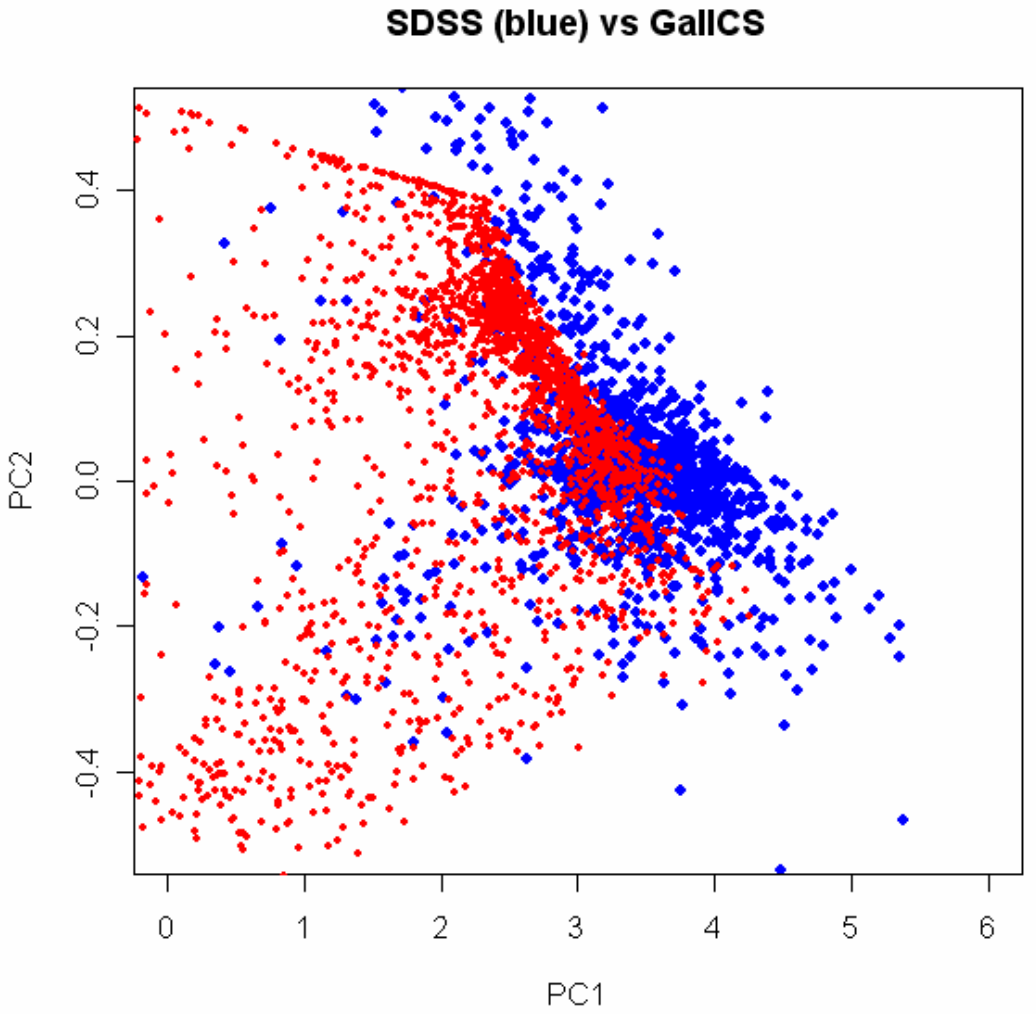
PCA data model (RDB schema available)



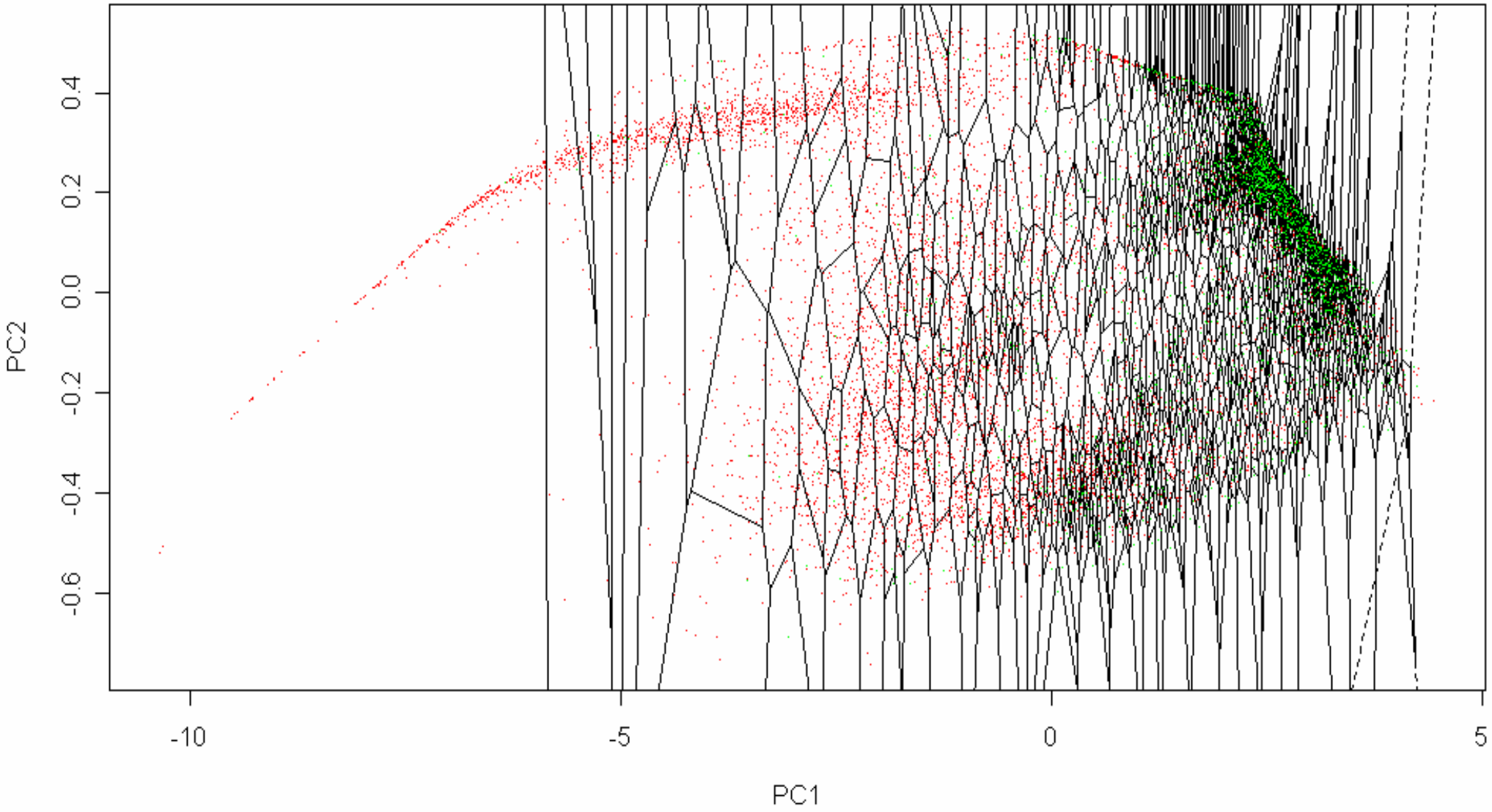


2006-04-05 IAP


GalICS + milli-Millennium
PC2
PC1
0.4
0.2
0.0
-0.2
-0.4
-0.6
-10
-5
0
5


SDSS (blue) vs GalICS
PC2
PC1
0.4
0.2
0.0
-0.2
-0.4
0
1
2
3
4
5
6

milliMil-GalICS
PC1 vs PC2 Voronoi tesselation (HVO/JHU)



2006-04-05 IAP