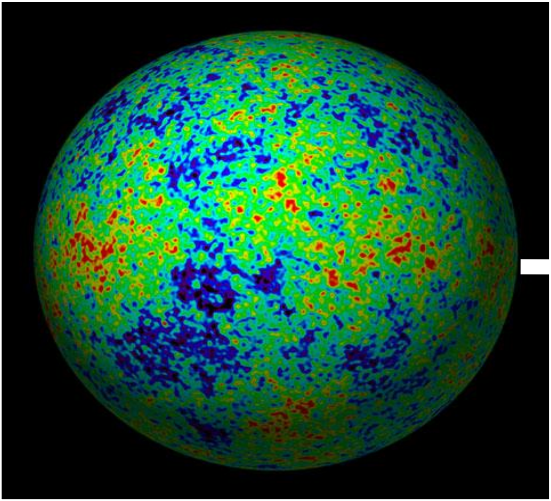
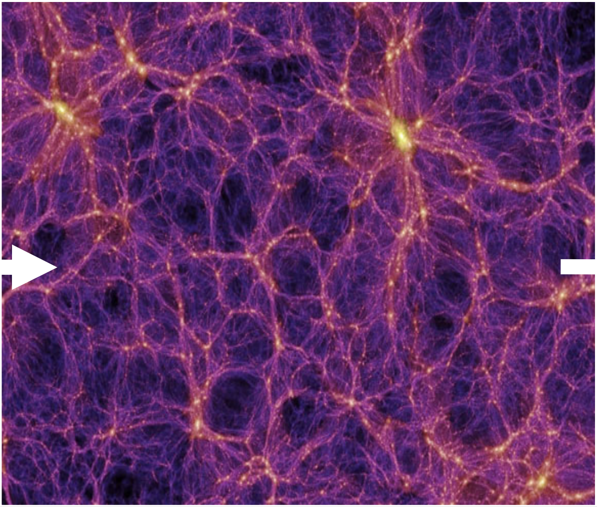
BIG DATA

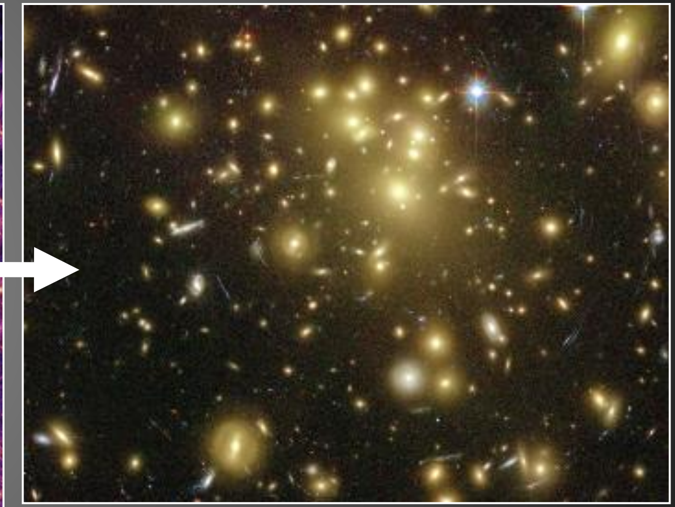
Processing, Analysing and Visualizing the SDSS Database



Cosmology



Large Scale Structure



Galaxies

Dr. Asa F. L. Bluck

University of Victoria

Overview of Session

- 9:30 – 10:15 Lecture on using the SDSS database for research
- 10:15 – 11:15 Workshop on using TopCat to visualize big data
- 11:15 – 11:30 Break
- 11:30 – 12:20 Group discussion on
  - – Woo et al. 2013 (halo mass quenching)
  - – Bluck et al. 2014 (bulge mass quenching)

Lecture:
What's in the SDSS Database?

- Information on over a million galaxies
  - Photometry for 1.1 million galaxies
  - Spectroscopy for 0.7 million galaxies
- Close to 1000 properties measured (for each galaxy)
- So, that means the database is:
  ~ a billion numbers!!
- How we get to grips with such a vast resource is the topic of today's session
- Key themes:
  1) Develop new tools to explore the database
  2) Visualize complex relationships
  3) Understand some common analysis techniques

Simple Database Queries

- use sdss; (selects database)
- show tables; (shows what tables are available)
- describe table_name; (lists column names)
- select count(*) from table_name;

  (returns column length)
- Select var1, var2, var3 … from table_name where ‘Boolean Opperators [e.g. var2 is not NULL and var6 > var5 …]’ limit N;

  (where N = number of required outputs, if absent the database will return all available values!)

Advanced Database Queries

The general form of a multi-table, multi-variable query is:

```
select a.column_name1, a.column_name2,
b.column_name1, c.column_name1 …

from table1 a, table2 b, table3 c …

where a.objID = b.objID and b.objID = c.objID …

and ‘Boolean Operators [e.g. a.column_name6 >
b.column_name9 …]’ ;
```

Master Table: “dr7_uberuber”

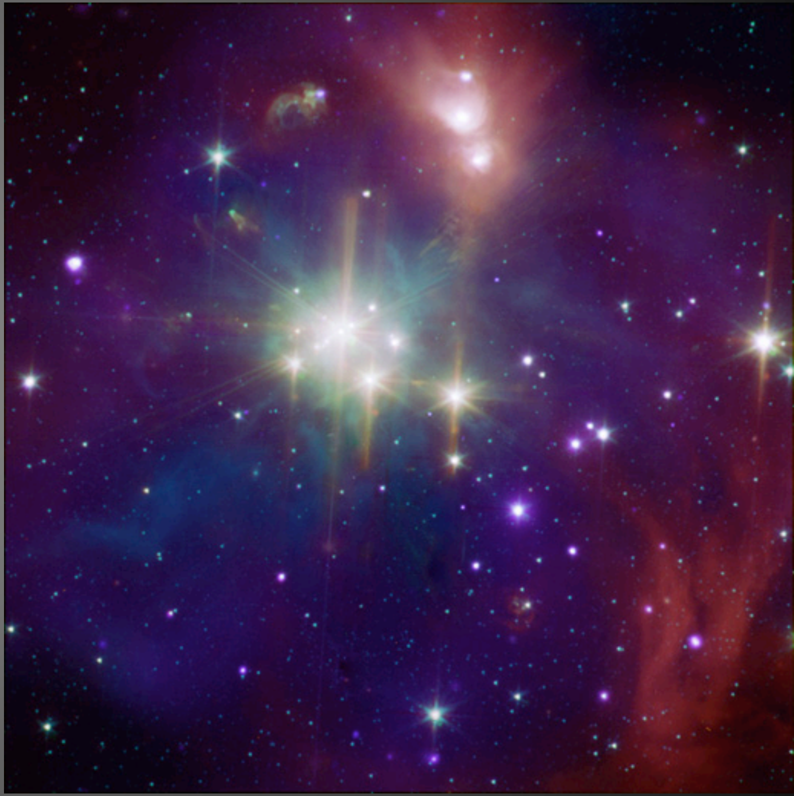
- ra, decl, z_spec
- total_SFR_med, fiber_SFR_med
- total_mass_med, fiber_mass_med (OLD values)
- delta_SFR (distance from SF Main Sequence)
- K01_flag, K03_flag, S06_flag (BPT flags)
  - AGN, SF, NULL (s/n > 5)
- new_K01_flag, new_K03_flag, new_S06_flag, agn_sn
  - AGN, SF, NULL (s/n > 1)
- Note: d“prop” -> error (e.g. dz_spec = err(z_spec)

Emission Lines: "dr7_emmlines"

- Halpha_flux, Hbeta_flux, NII_flux, OII_flux, OIII_5007_flux, OIII_4959_flux, …
  - – Line strengths no dust correction
- Corr_Halpha_flux_mw, corr_Hbeta_flux_mw, corr_NII_flux_mw, corr_OIII_5007_flux_mw, …
  - – Line strengths corrected by Milky Way extinction law, intended to remove dust effects.

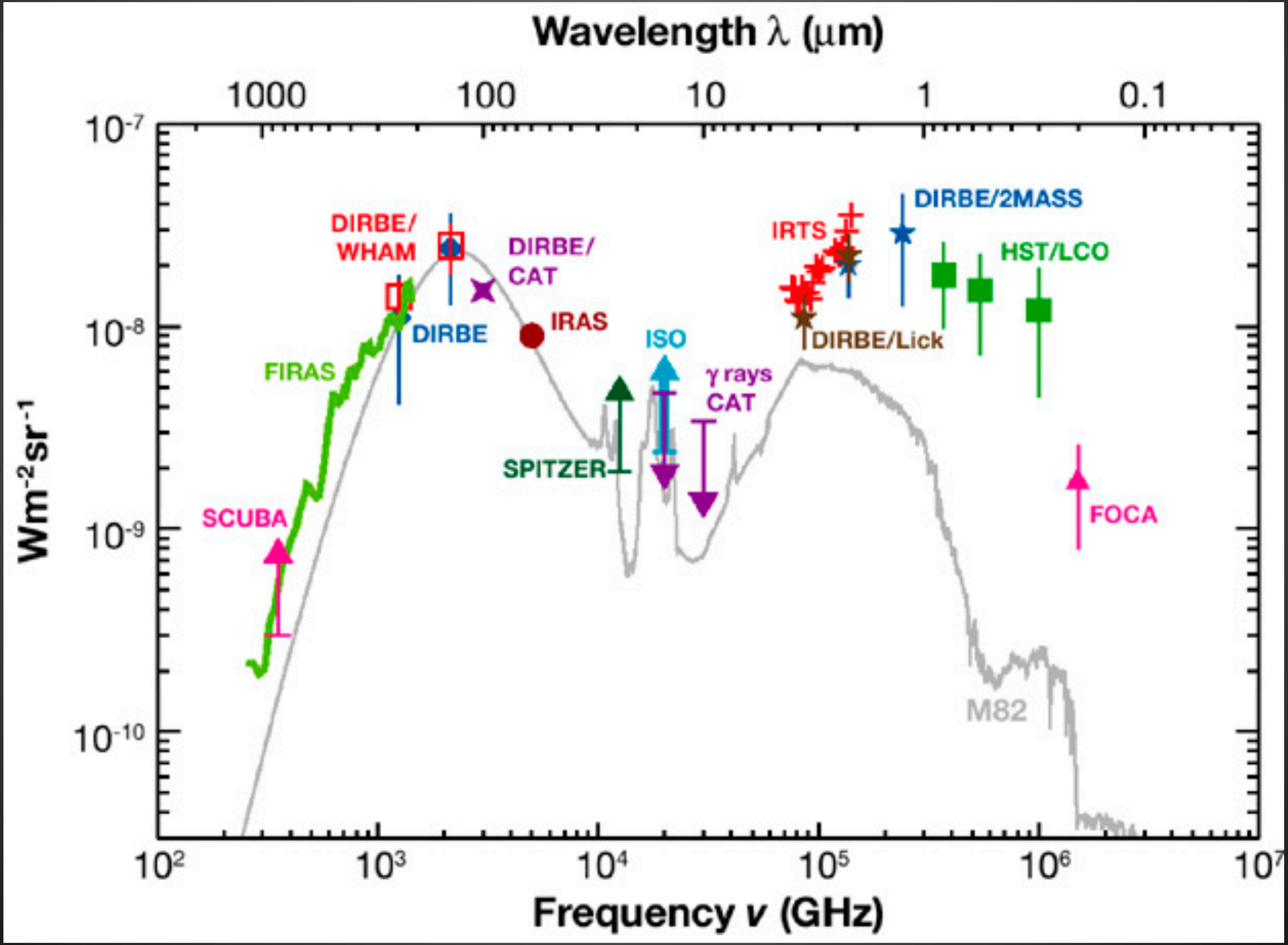
How to Measure Star Formation Rates?
1) UV + IR Fluxes

- The lifetime of a star varies as a strong function of mass: lower mass stars live longer
- More massive stars are brighter and bluer than less massive stars
- So, noting a population of very bright blue stars indicates recent star formation:
- high UV fluxes -> high SFR
- BUT UV photons are easily absorbed by dust particles
- So, we need to add the absorbed light too (which is re-emitted in IR):
- SFR_tot = SFR_UV + SFR_IR



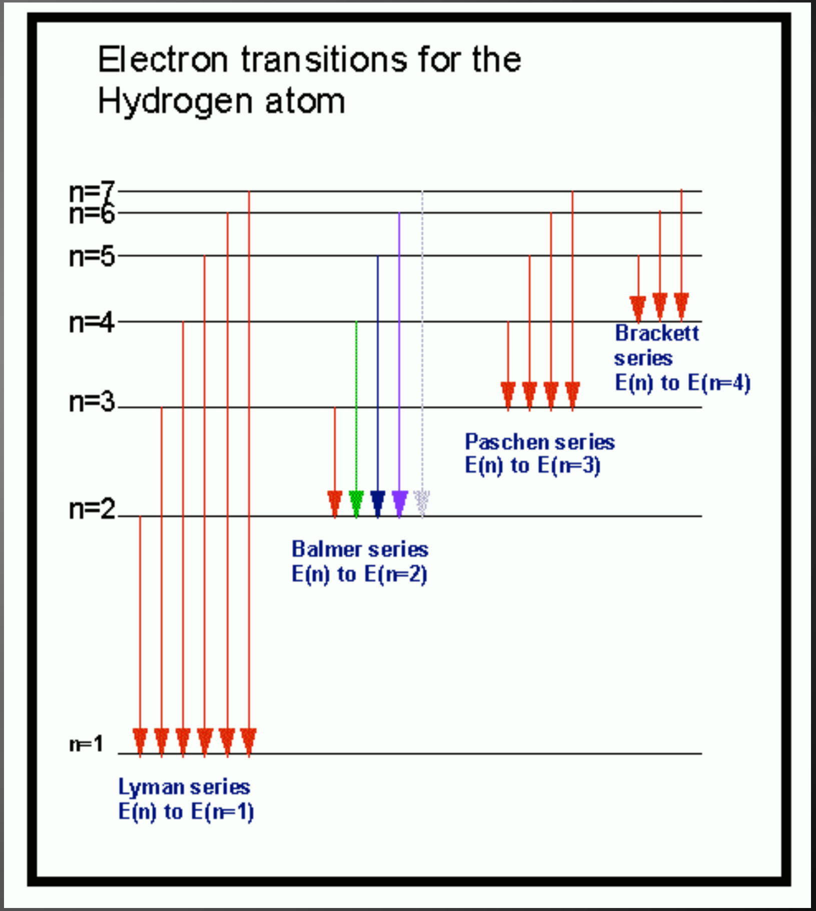
Coronet nebular
(star forming region in our Galaxy)

How to Measure Star Formation Rates?
1) UV + IR Fluxes

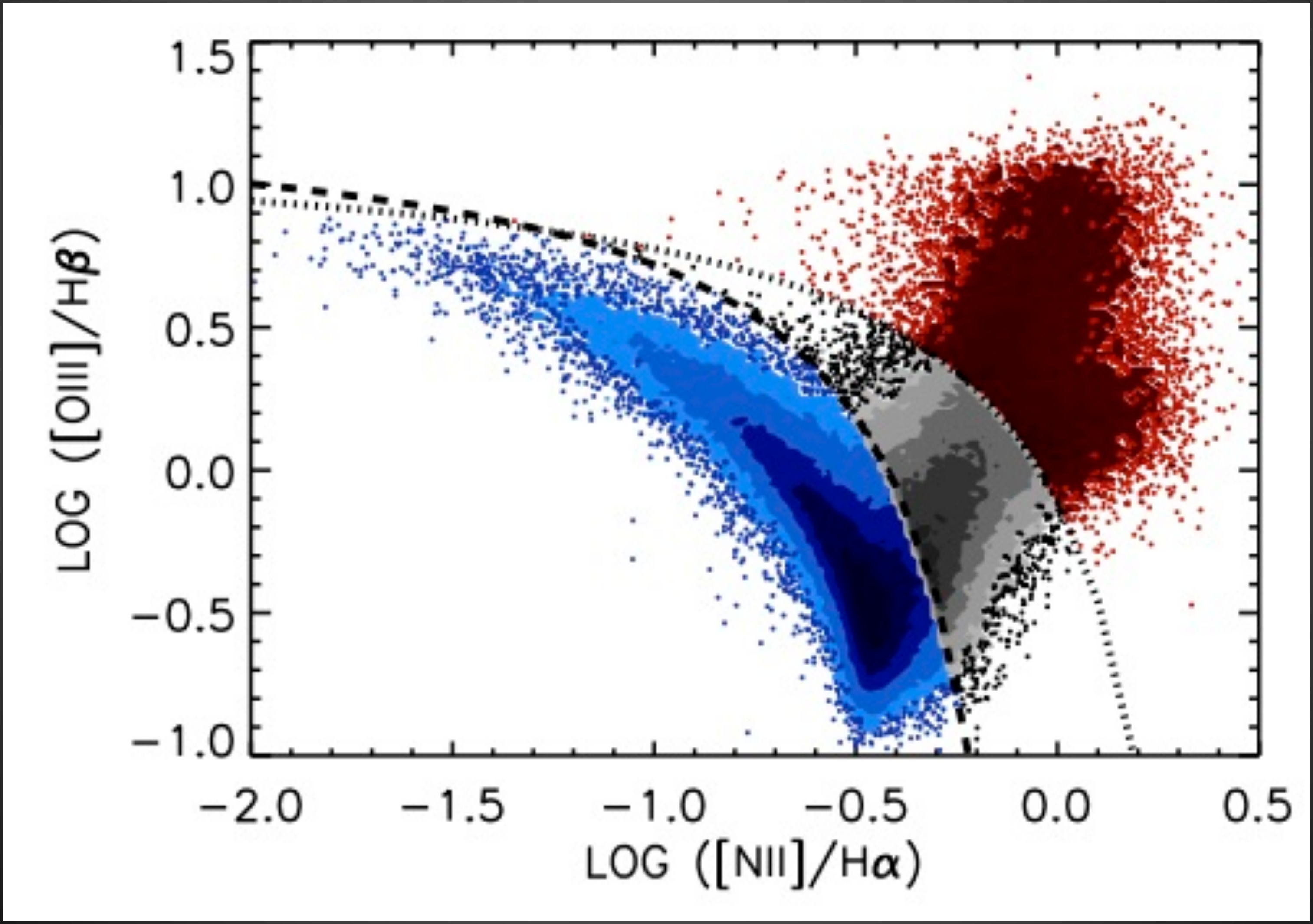


How to Measure Star Formation Rates?
2) Optical Emission Lines

- Massive stars give out a lot of highly ionizing radiation
- This can remove electrons from interstellar hydrogen gas, and also elevate energy levels in atoms
- When the radiation flux reduces, the electrons and protons re-combine and the electrons settle down in their energy levels
- This leads to Ly_alpha, H_alpha, H_beta etc. emission lines
- Strong emission lines -> high SFR
- However, AGN can also excite atomic Hydrogen, so there can be some confusion

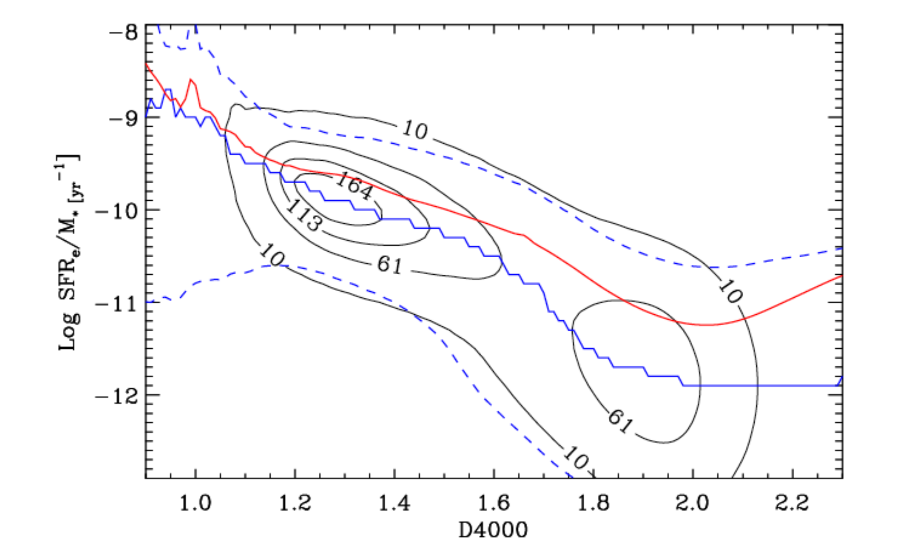

Electron transitions for the Hydrogen atom
n=7
n=6
n=5
n=4
n=3
n=2
n=1
Brackett series E(n) to E(n=4)
Paschen series E(n) to E(n=3)
Balmer series E(n) to E(n=2)
Lyman series E(n) to E(n=1)

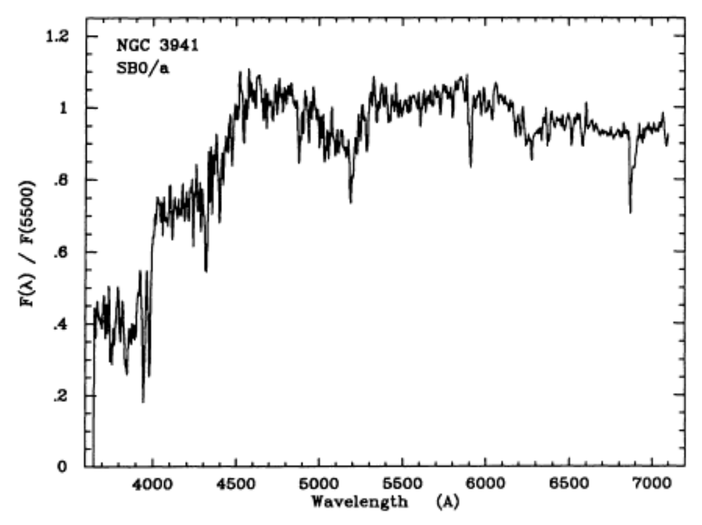
How to Measure Star Formation Rates?
3) Telling SF from AGN



How to Measure Star Formation Rates? 4) Indirect Spectroscopic Means

- For AGN and non-emission line galaxies it is VERY difficult to estimate reliable SFRs via optical astronomy alone
- However, Brinchmann et al. (2004) use an observed correlation between sSFR (=SFR/M*) and Dn4000 (the strength of the 4000 A break feature in a galaxy's sprectrum) to estimate a *rough* value
- This "break" is a result of metal and H absorption at wavelengths shorter that ~4000A





How to Measure Star Formation Rates?
5) Photometric Estimation

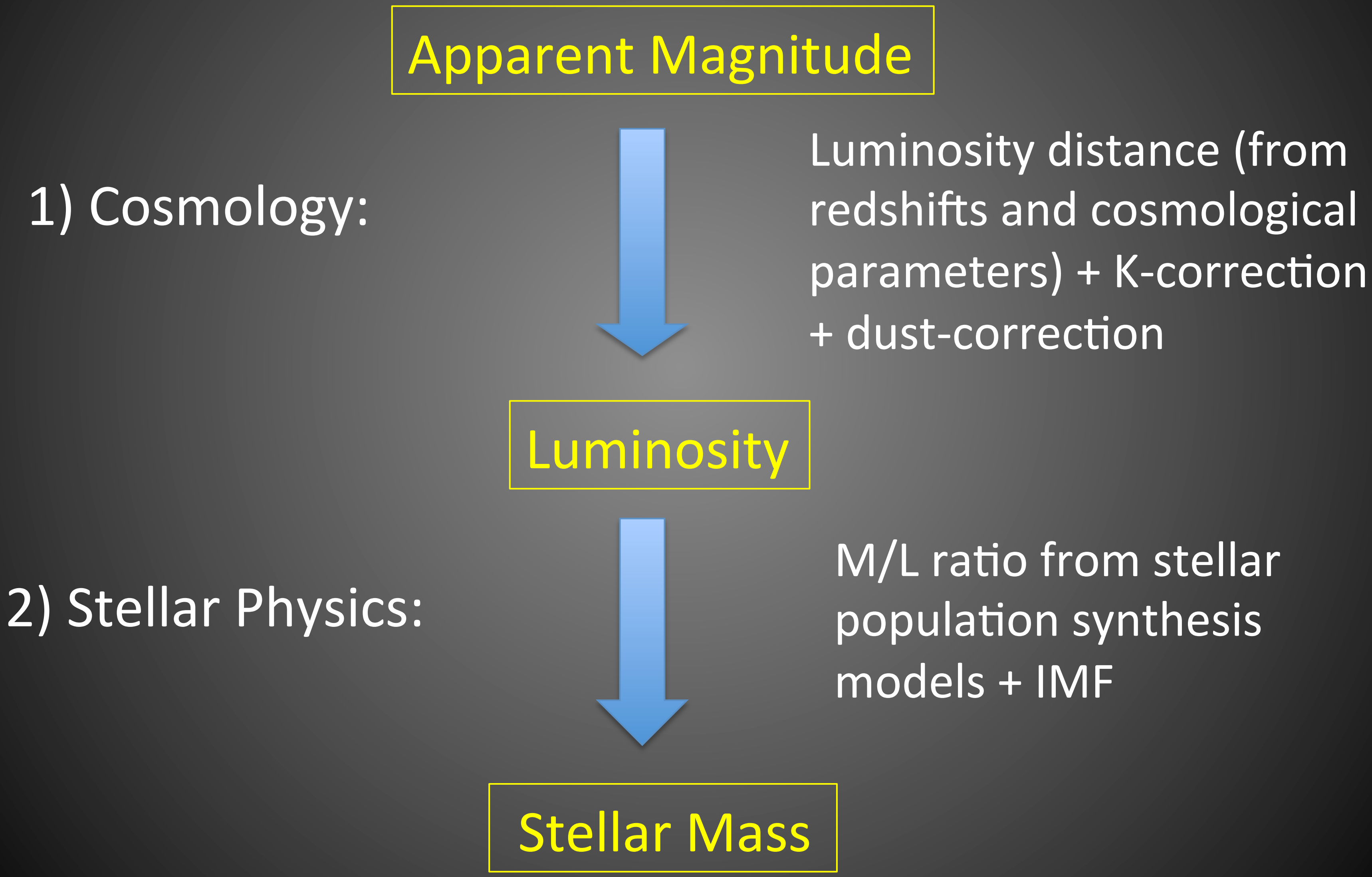
- The colour and magnitude of galaxies in varying (ugriz) wavebands can be used to constrain their full spectrum, or SED.
- By matching to large samples of model galaxy spectra, the fraction of young-to-old stars in the integrated stellar population can be estimated
- From this the SFR is determined
- Generally speaking, bluer galaxies have higher SFRs, but watch out for dust which reddens the colours!
- Dust is estimated by various extinction laws based on multi-wavelength data and empirical correlations

Component Masses: “dr7_nsed_spec”

- total_mass (NEW, B+D together)
- bulge_mass, disk_mass
- V_max (Volume over which galaxy is visible)
- z_min, z_max (min and max visible redshifts)
- P_pS (If < 0.32 -> composite; If > 0.32 -> single)
- contam_flag (set = 0 for “good” fit)
- delta_bd (set < 1 for “reliable” total masses)
- Note: ddisk_mass_m is –ve error on disk_mass

  ddisk_mass_p is +ve error on disk_mass

How to Measure Stellar Masses?



Apparent Magnitude

1) Cosmology:

Luminosity distance (from redshifts and cosmological parameters) + K-correction + dust-correction

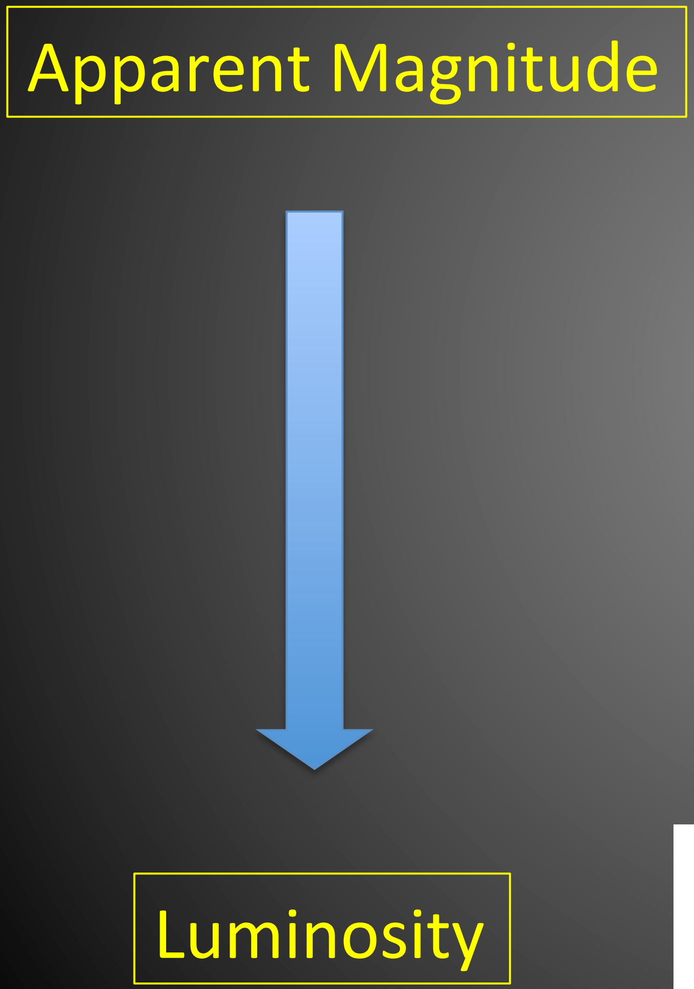
Luminosity

2) Stellar Physics:

M/L ratio from stellar population synthesis models + IMF

Stellar Mass

How to Measure Stellar Masses?
1) Cosmology



Apparent Magnitude

Luminosity

Definition:

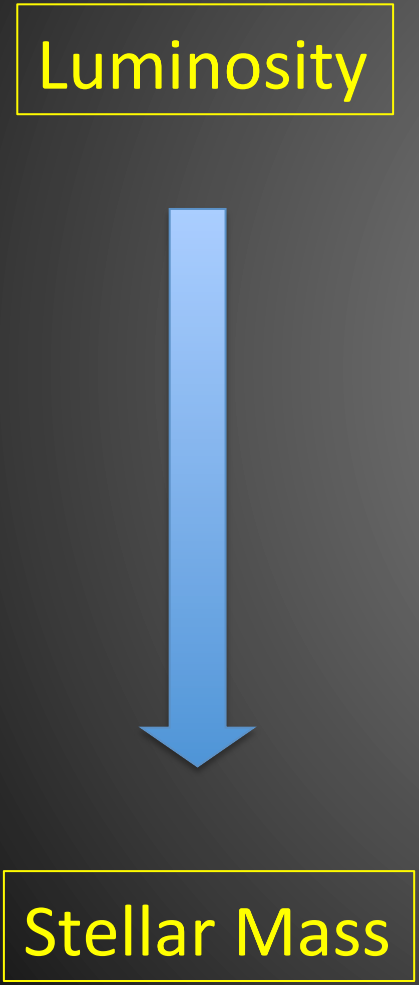
$$D_{\mathrm{L}} \equiv \sqrt{\frac{L}{4\pi\, S}}$$

$$D_{\mathrm{L}} = (1+z)\, D_{\mathrm{M}} = (1+z)^2\, D_{\mathrm{A}}$$

$$D_{\mathrm{C}} = D_{\mathrm{H}} \int_0^z \frac{dz'}{E(z')}$$

$$E(z) \equiv \sqrt{\Omega_{\mathrm{M}}\,(1+z)^3 + \Omega_k\,(1+z)^2 + \Omega_\Lambda}$$

How to Measure Stellar Masses?
2) Stellar Physics



Luminosity

Stellar Mass

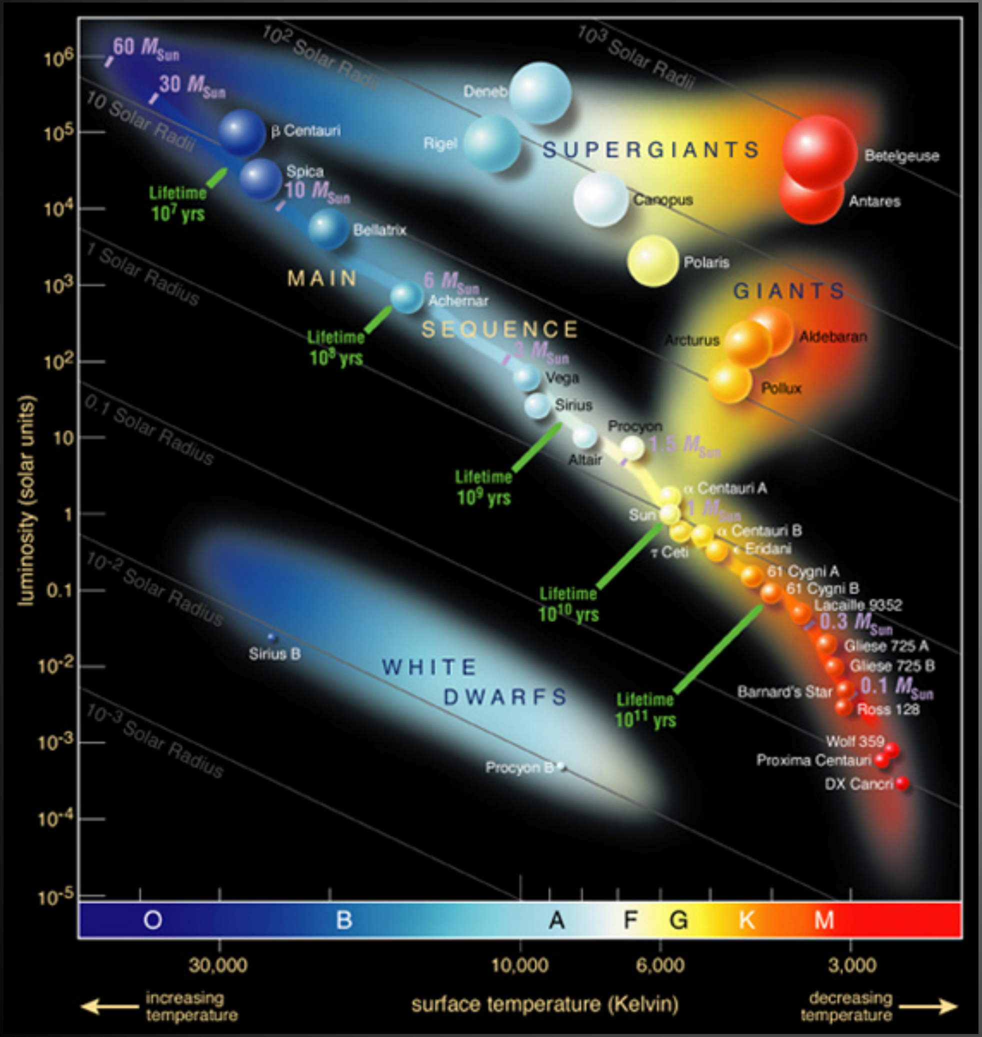
An (Overly) Simplistic Approach:

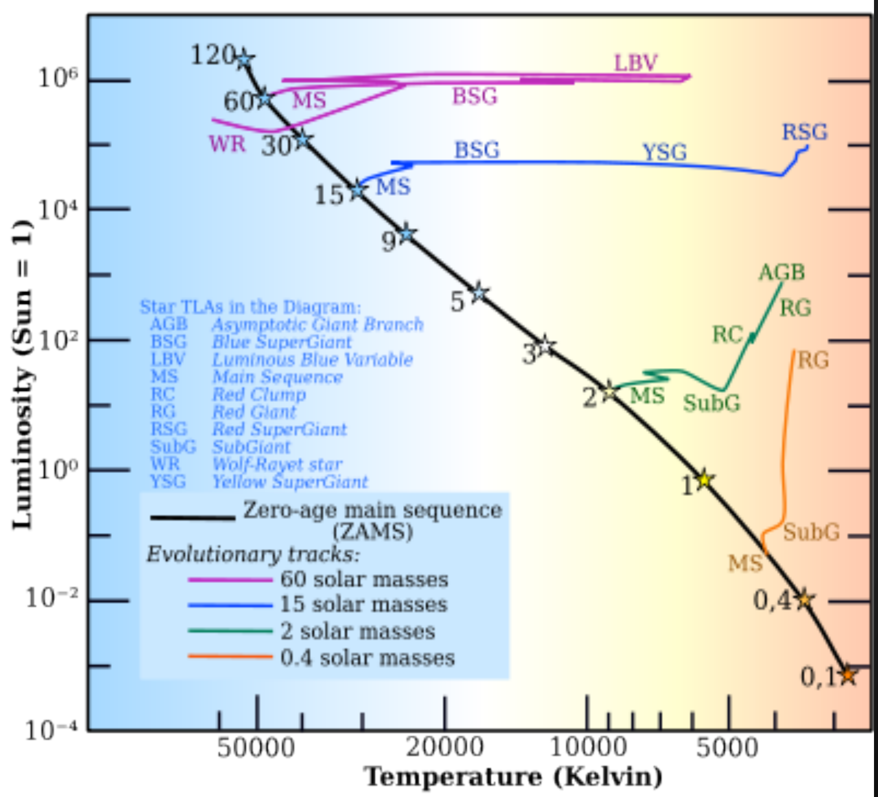
Galaxy Mass =
(Luminosity / Solar Luminosity) x Mass of Sun

BUT – Not all stars have solar M/L ratio. Roughly a factor of 10 - 100 out for Ellipticals, and a factor of 100 – 1000 out for star forming spiral galaxies!

So, we need to know what stars are in each galaxy, their M/L ratios, and in what abundances they exist at any given epoch.

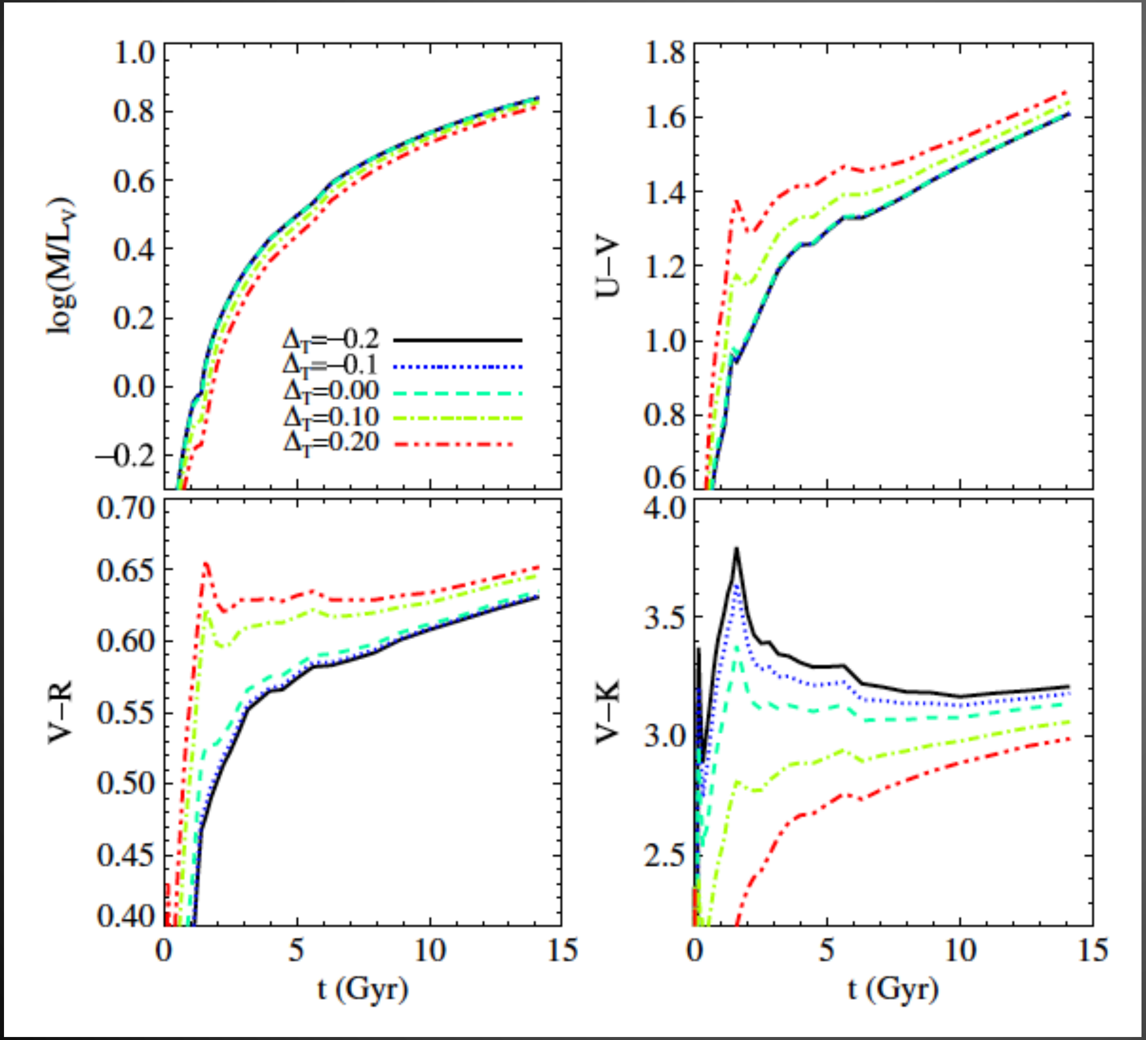
How to Measure Stellar Masses?
2) Stellar Physics – HR Diagrams





Stellar Mass Evolutionary Tracts

How to Measure Stellar Masses?
2) Stellar Physics – M/L (λ,t)



M/L varies with time
But so does colour

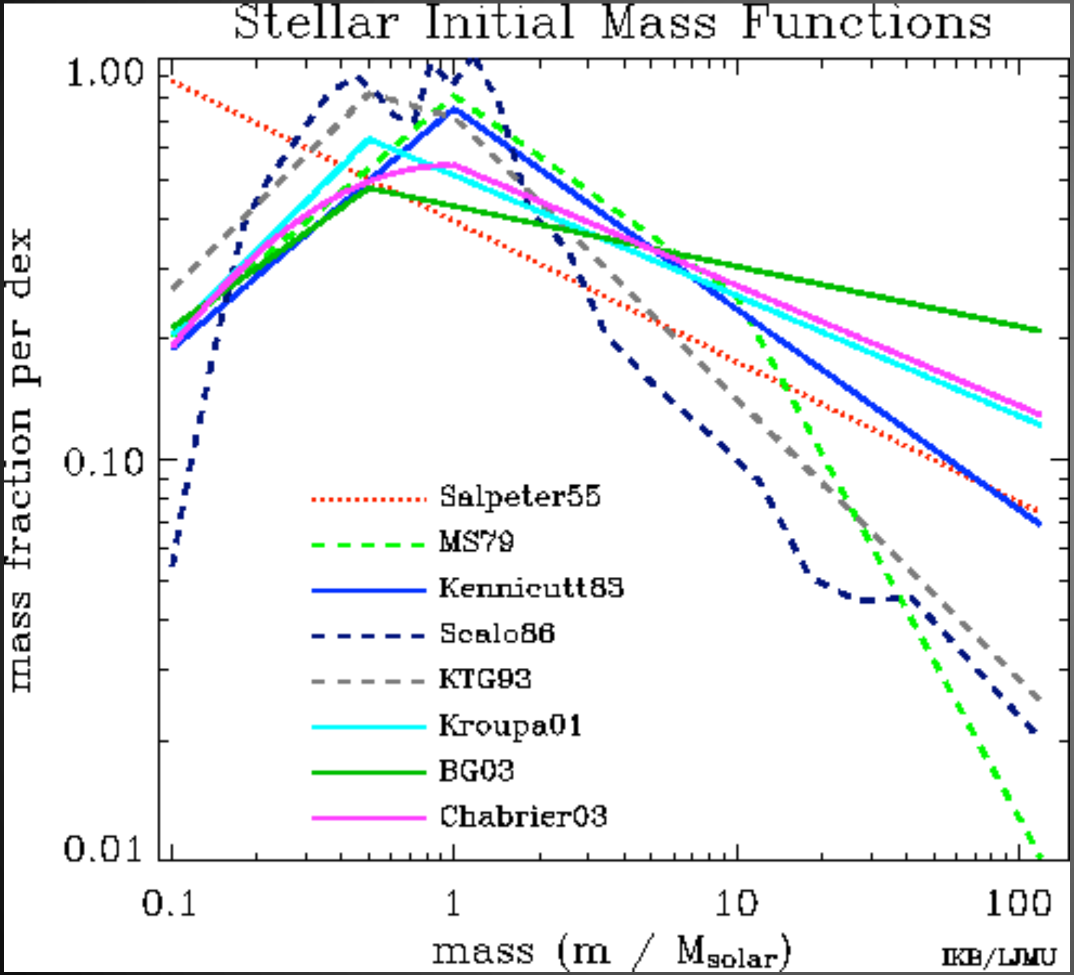
So we use colour to infer M/L

But we need to make assumptions about the initial mass function (IMF) of stars.
And we have to be careful about dust extinction again.

How to Measure Stellar Masses?
2) Stellar Physics – General Method

- Compare spectral energy distributions (SEDs) to a library of model versions (e.g. Bruzual & Charlot 2003; Conroy, Gunn & White 2009)
- Simultaneously fit for age of stellar population, metalicity, dust extinction, e-folding time, …
- Assume an IMF
- The stellar mass of the model galaxy which best fits the photometric data is assigned to the galaxy
- Errors are computed from the width of the probability distribution function across the library of model galaxies: typically 0.2 – 0.3 dex

The Initial Mass Function



Definition:

$$\Phi \equiv \frac{\mathrm{dn}}{d\ln M} =$$

Parameterization:

$$\begin{cases} A_l(0.5n_cm_c)^{-x}\exp\left[\frac{-(\log M - \log m_c)^2}{2\sigma^2}\right] \\ \qquad\qquad\qquad (M \leqslant n_cm_c) \\ A_hM^{-x} \\ \qquad\qquad\qquad (M > n_cm_c), \end{cases}$$

B/D Photometric Morphologies:

“sdss_dr7_morph_mybkg_mydeblend_Xr” (where X = {u,g,I,z})

- Mr_galaxy, MX_galaxy (absolute r, X band magnitude)
- Mr_bulge, Mr_disk, MX_bulge, MX_disk
- btr, btX (bulge-to-total *light* ratio in r, X band)
- rhalf_X (half light radius of galaxy in arcsecs)
- rhalf_kpc_X (half light radius of galaxy in kpc)
- re, re_kpc (half light radius of BULGE – arcsec, kpc)
- rd, rd_kpc (half light radius of DISK – arcsec, kpc)
- P_pS (If < 0.32 -> composite; If > 0.32 -> single)
- n (Sersic index of BULGE == 4)

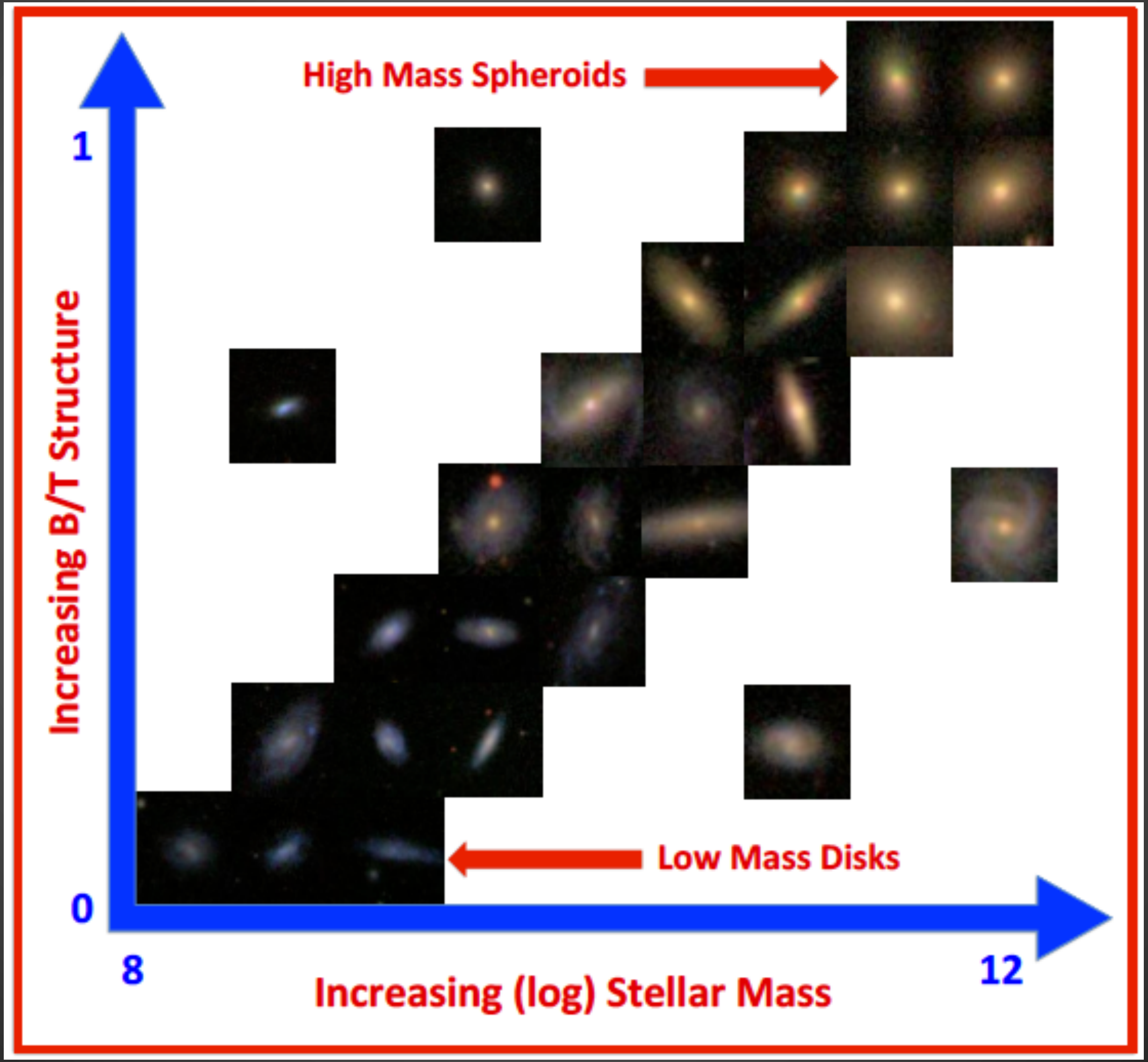
Single Sersic Morphologies:

“sdss_dr7_morph_mybkg_mydeblend_sersic_Xr”

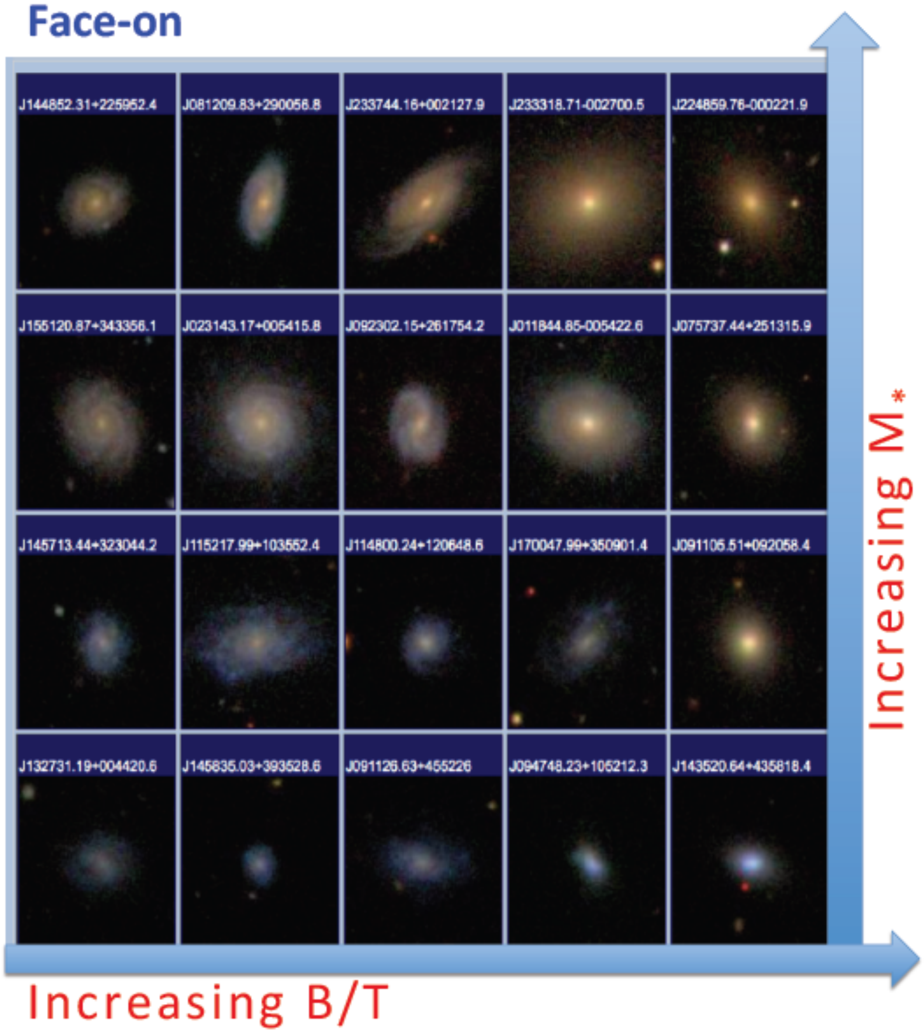
(where X = {u,g,i,z})

- n (Sersic index of galaxy in X, variable 0.5 – 8)
- rhalf_r, rhalf_X (half light radii in arcsecs)
- rhalf_kpc_r, rhalf_kpc_X (half light radii in kpc)

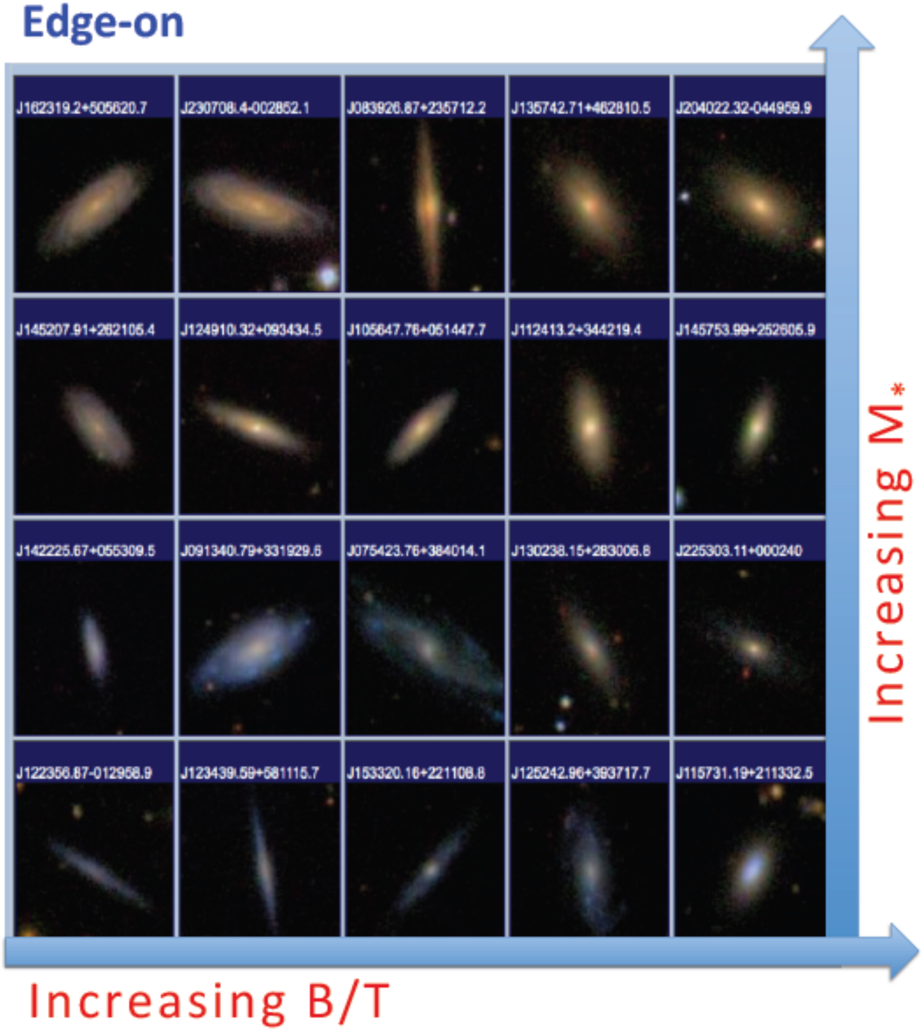
Schematic Rendering of B/T – M* Relationship



Colour – Mass & B/T Variations



Face-on



Edge-on

Environmental Metrics: “dr7_density”

- d2p, d3p, d5p, d10p (d“n”p, surface density of galaxies evaluated at the nth nearest projected neighbour)
- d“n”p_norm (where n = 2, 3, 5, 10. Same as above but normalized by average value at the redshift in question)

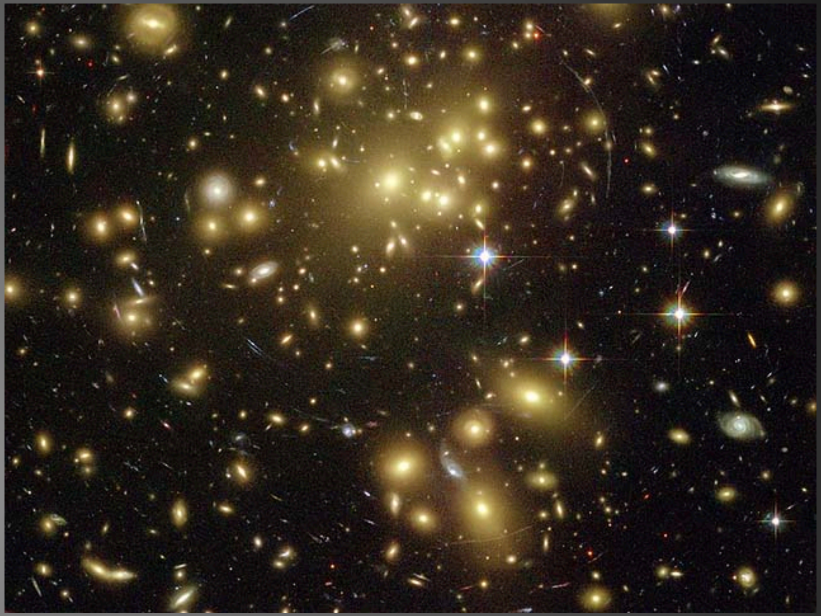
How to Measure Local Densities?

Galaxy Surface density to nth nearest neighbour (with dv < 1000 km/s):

$$\Sigma_n = \frac{n}{\pi r_{p,n}^2}$$

Normalized by average value At redshift of measurement:

$$\delta_n = \frac{\Sigma_n}{\langle \Sigma_n(z \pm \delta_z) \rangle}$$



High Density Galaxy Cluster – Abell 1689

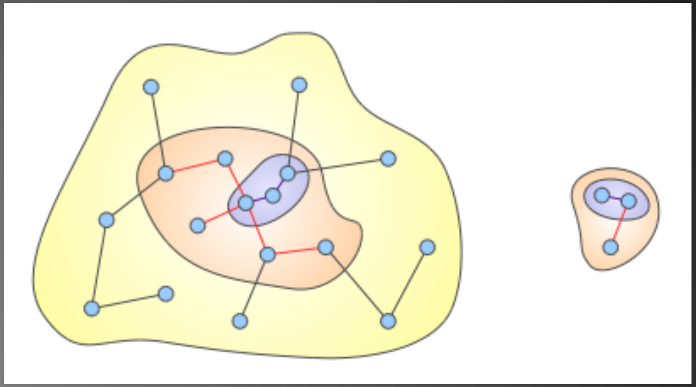
Low Density Galaxy Field

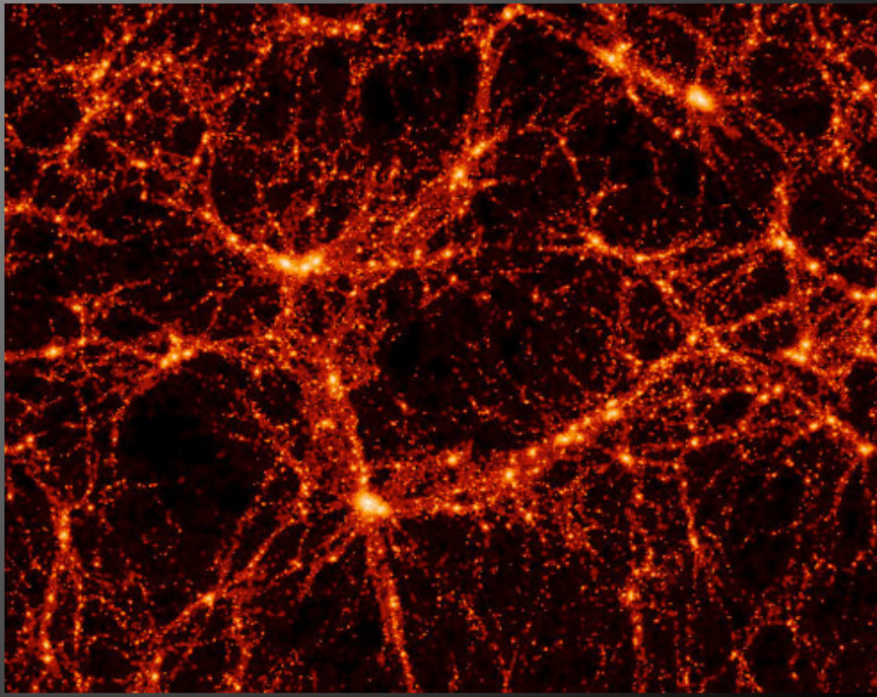
Environmental Metrics: “dr7_groups_yang”

- M_halo_mass (log solar masses)
- is_MMGG (is most massive galaxy in group? 1 = central, 0 = satellite)
- rp_mass (distance in kpc from projected centre of mass of group)
- r200_mass (approximate virial radius of group given in kpc)

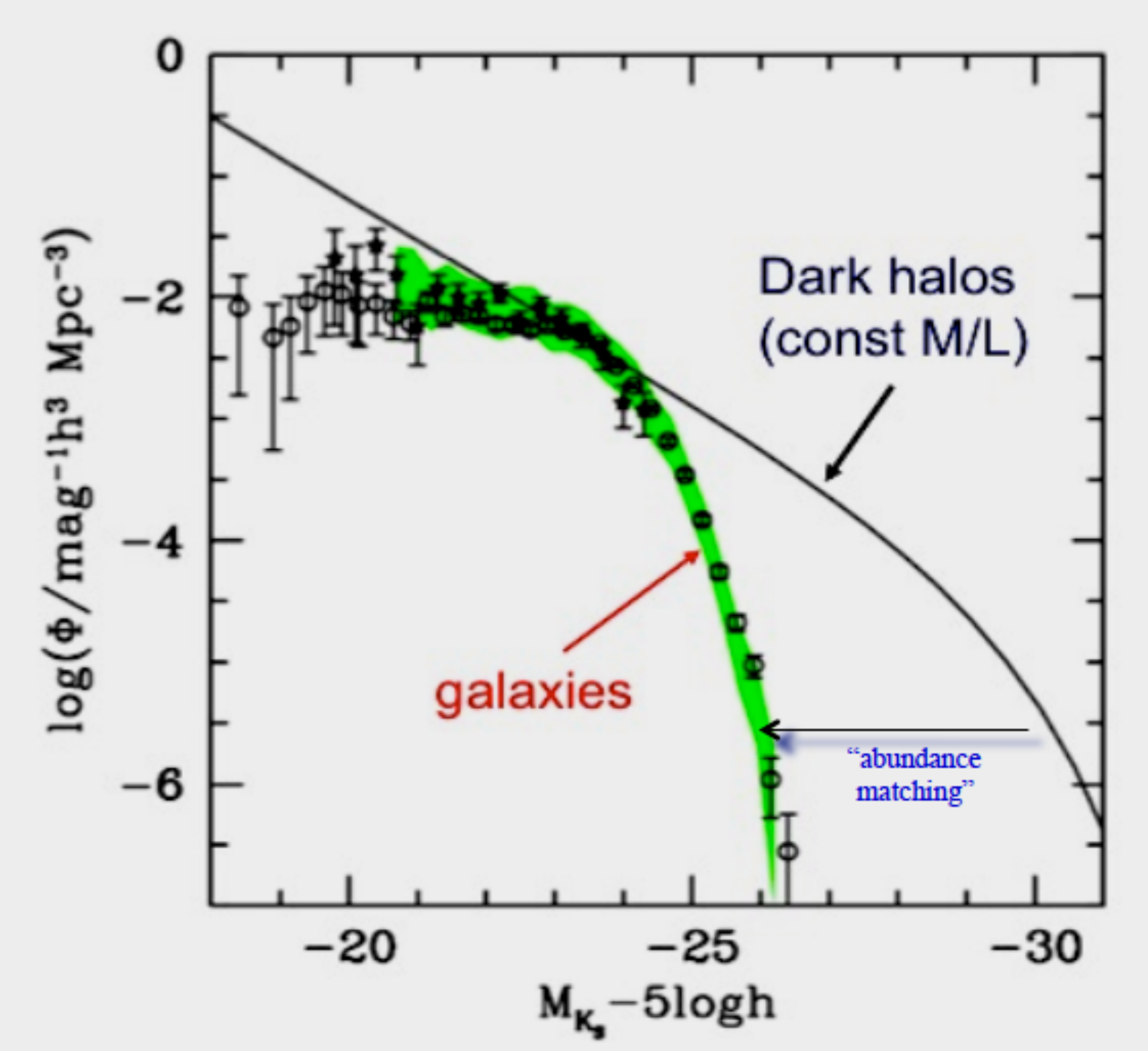
How to Find Galaxy Groups? First Steps – Linking Algorithm

1. Choose (‘guess’) a *linking length*
2. Count all galaxies within that distance from any neighbour as a member of the group
3. Sum their stellar masses together
4. Use “Abundance Matching” to estimate the halo mass



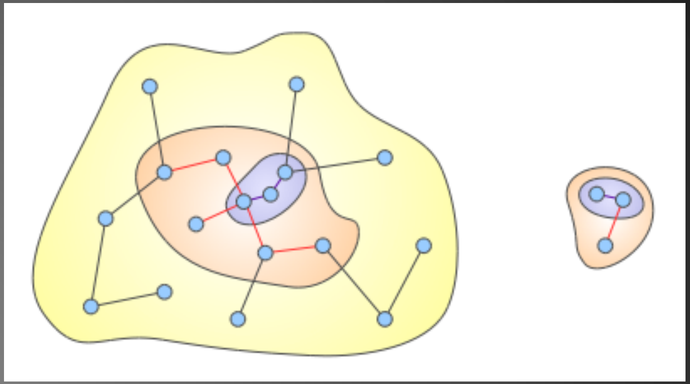


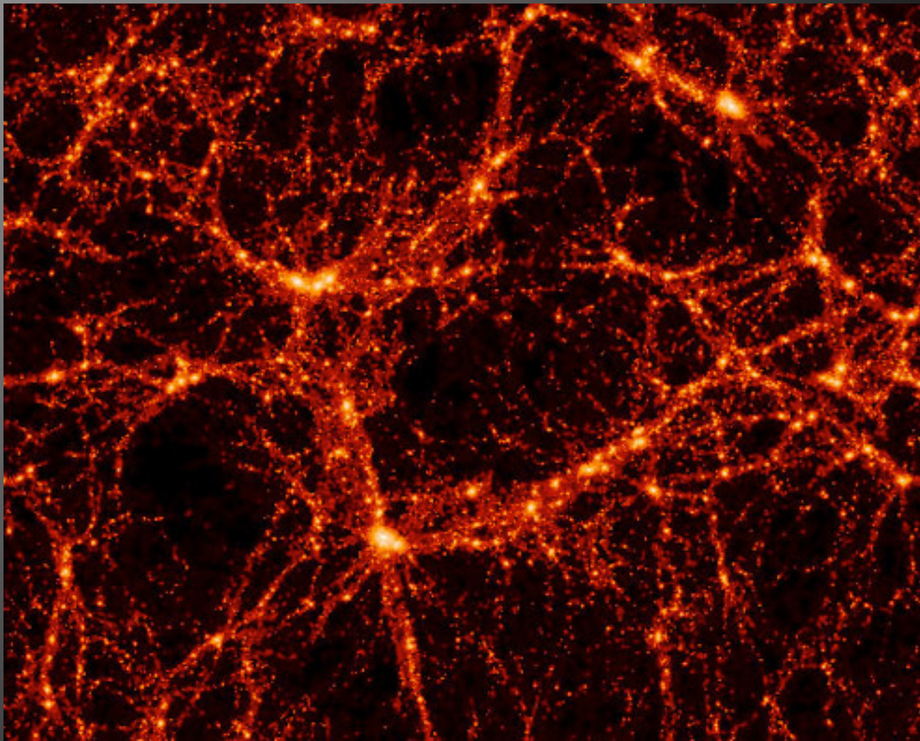
How to Measure Halo Masses?
First Steps - Abundance Matching



How to Find Galaxy Groups?
Next Steps - Iterate

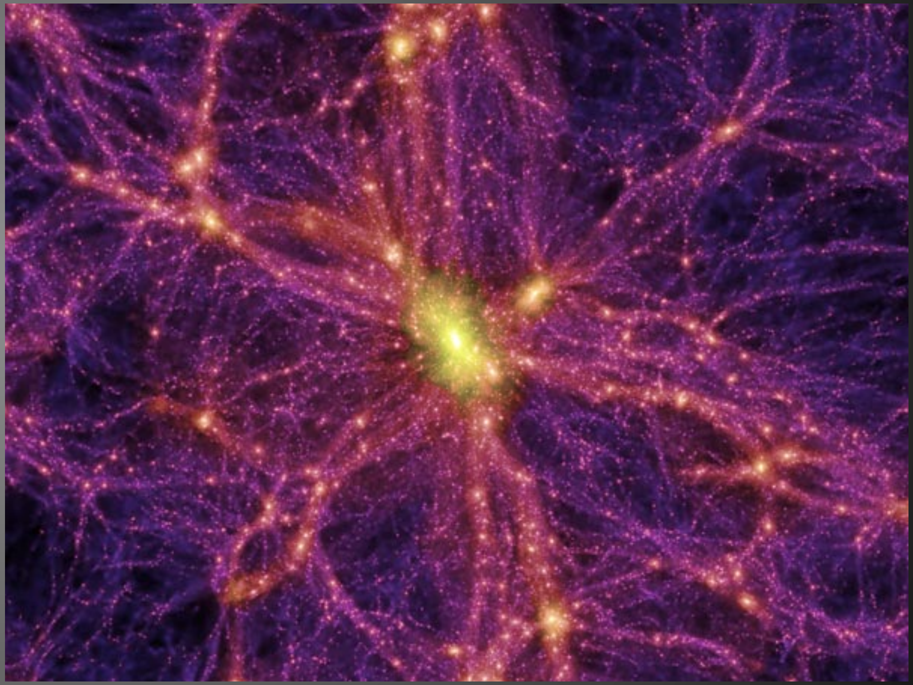
1. Calculate Rvir from Mhalo
2. Compute how many galaxies in “group” are beyond this distance
3. Discard those galaxies, and adjust the linking length to minimize their inclusion
4. Re-run group finder
5. Recompute Mhalo, Rvir
6. Iterate until you reach convergence





How to Find Galaxy Groups? Final Steps – Test Method

- Compare to Semi Analytic Model (SAM) run on Millennium dark matter simulation
- How many galaxies are successfully assigned to their correct DM halo group or cluster?
- 90% at log(M_halo) > 12
- Starts to break down at log(M_halo) < 12, with a very low success rate at log(M_halo) < 11



Millennium Simulation

Galaxy Pair Catalogues: Tables

- "cp_pairs_mendelz"
  - rp < 80 kpc, dv < 10,000 km/s, 0.1 < M1/M2 < 10
  - DR7 spectroscopic parent sample. Uses most up-to-date mass estimates available.
- "cp_nonpair_mendelz"
  - Everything not in cp_pairs_mendelz
- "cp_control_mendelz"
  - Matched in redshift and total stellar mass

Galaxy Pair Catalogues: Tables

- “cp_scudder_pairs”
  - Made from cp_pairs, dv < 300 kpc, 0.02 < z < 0.2, has metalicity measured, SF by K03 cut (i.e. star forming emission line galaxies only)
  - Uses slightly older masses (from g-r colours)
- “cp_scudder_control”
  - Made from cp_nonpair
  - Must have metalicity and be star forming
  - Matched by mass, redshift, and local density
  - 10 controls per galaxy in pair sample

Galaxy Pair Catalogues: Parameters

- rp (projected separation between pairs in kpc)
- delv (velocity difference between galaxy pairs in km/s)
- mratio (mass ratio of pair – M1 / M2)
- pair_objID (ObjID of pair to link control with)
- sfr_offset_tot (Total: SFR – med(SFR_controls))
- sfr_offset_fib (Fiber: SFR – med(SFR_controls))
- oh_offset (Total: [O/H] – med([O/H]_controls))

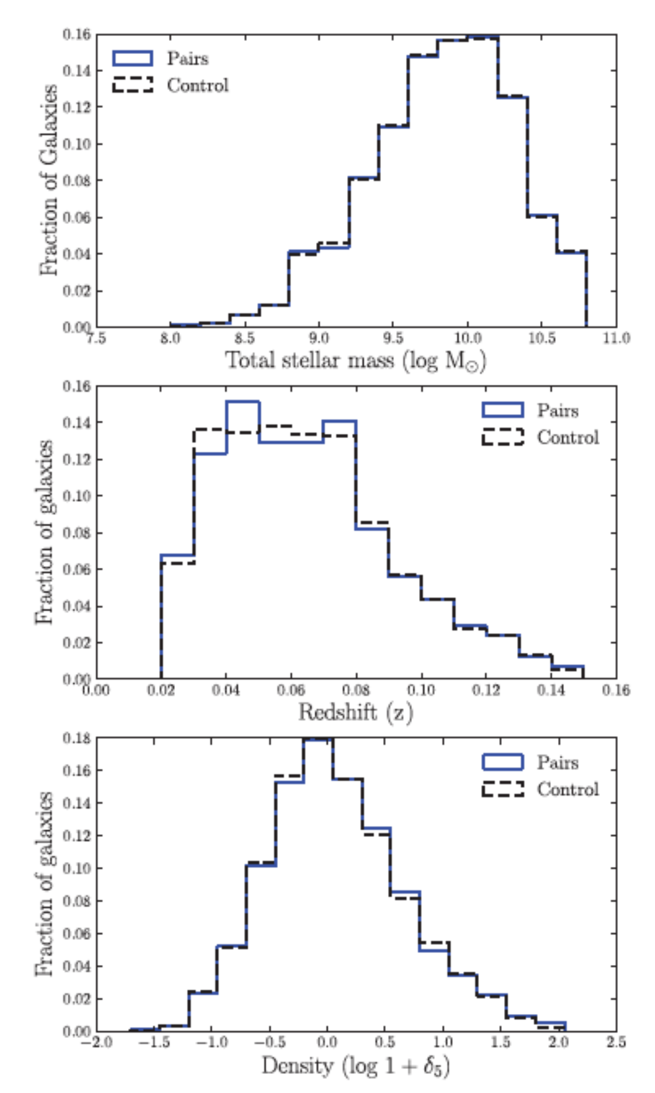
Why do we need Controls?

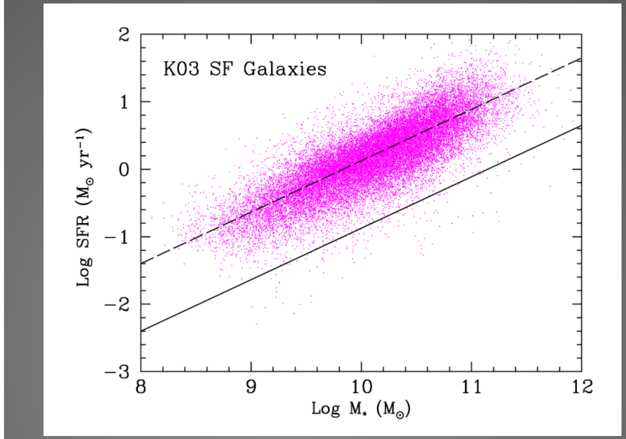
- Galaxies are highly complex objects!
- There are so many inter-relationships between galaxy properties that it is not always clear what the root cause is. Thus:
- Correlation does not imply causation!!
- The simplest way to test whether a given relationship is truly fundamental, is to see if it is dependent on any third variable.
- For example, the dependence of SFR on galaxy morphology is still present even at a *fixed stellar mass.*

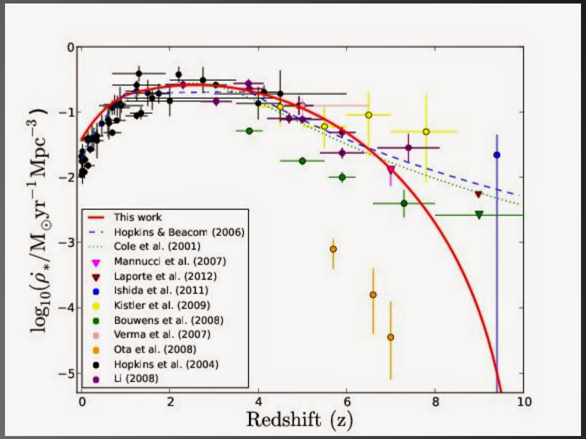
What exactly are Controls?

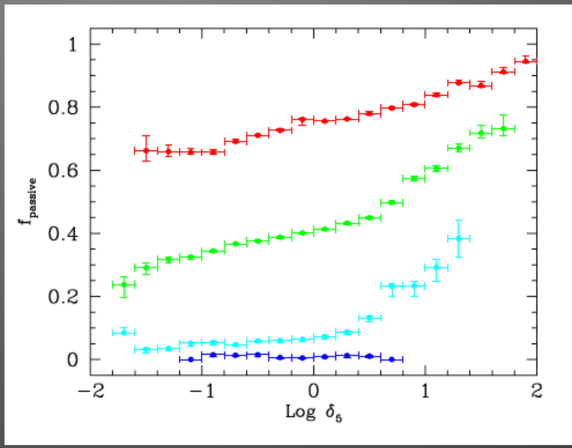
- A control group is a sample of objects intended for scientific study which do not have the property under investigation, or else have that property fixed.
- For example, if we want to investigate the effect of galaxy interactions on, say, SFR - the control sample will be galaxies not in pairs.
- It is important that the controls are as similar as possible to the science sample, except in the aspect of the variable under consideration.
- In practice it is very difficult to control for all relevant variables, but the main (known) factors at least should be thought about and matched on.

A Control Analysis Example: SFR Enhancement in Close Pairs





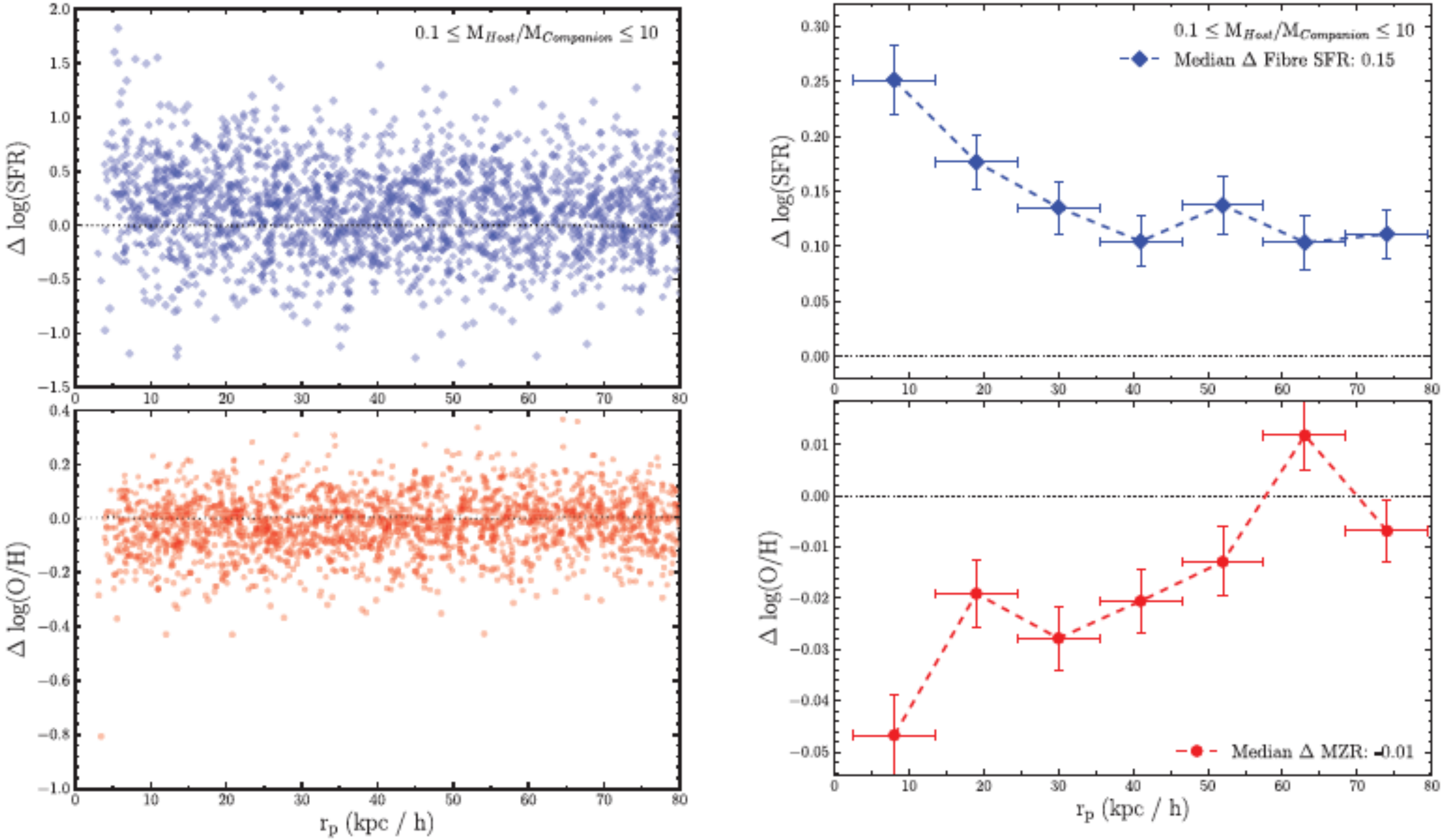




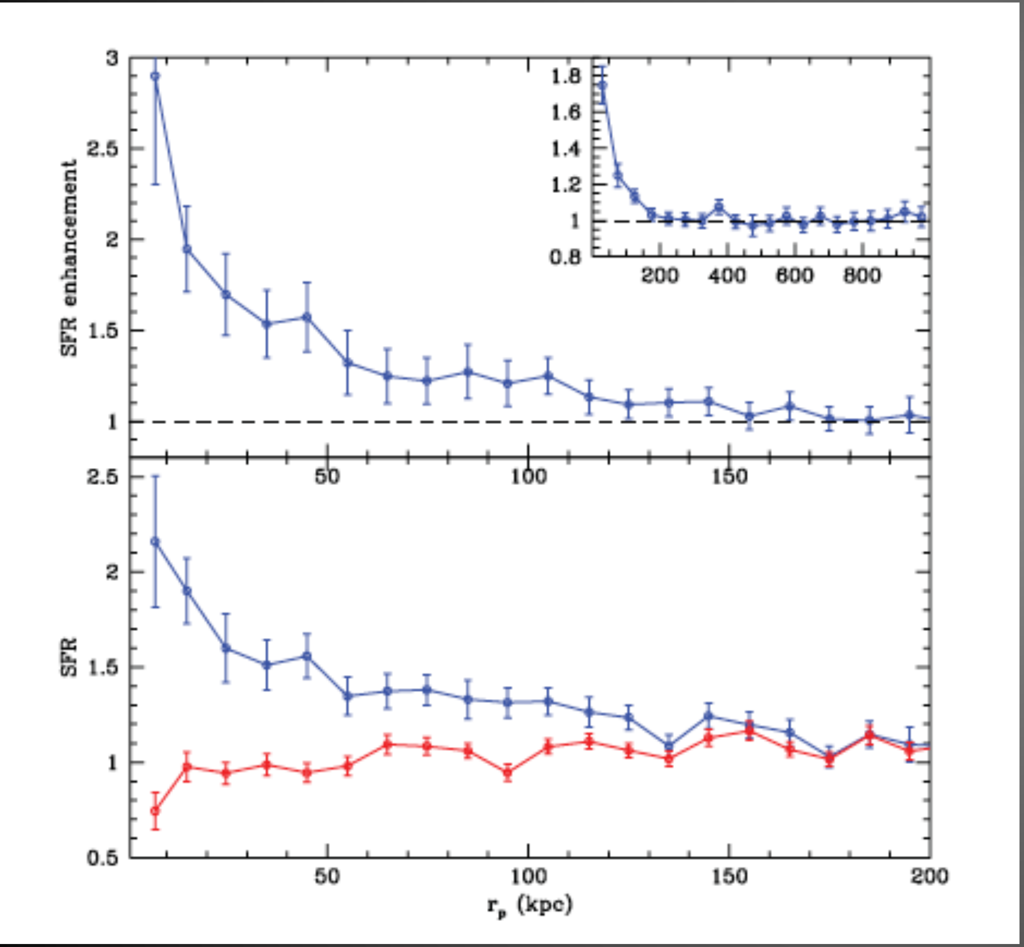
High Mass
Intermediate Mass
Low Mass
Very Low Mass

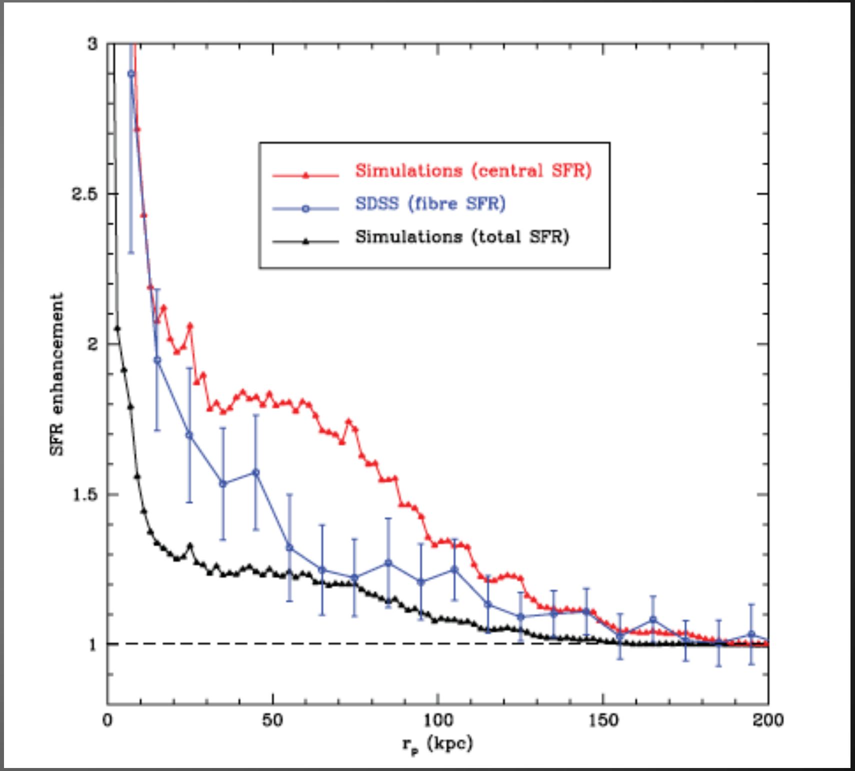
Scudder et al. 2012

A Control Analysis Example:
SFR Enhancement in Close Pairs



A Control Analysis Example: SFR Enhancement in Wide Pairs





Patton et al. 2013

Rest of Session

- 9:30 – 10:15 Lecture on using the SDSS database for research
- 10:15 – 11:15 Workshop on using TopCat to visualize big data
- 11:15 – 11:30 Break
- 11:30 – 12:20 Group discussion on
  - Woo et al. 2013 (halo mass quenching)
  - Bluck et al. 2014 (bulge mass quenching)

Opening TopCat with More Memory

- From the command line:
- **java -Xmx500M -jar topcat-full.jar &**
- **You can replace 500M with any value lower than your machine's RAM. But I find 500 is usually sufficient.**