From cosmological simulations to real observations: the ins and outs of the mock world… Part 1: the real universe

Peder Norberg (Durham University)

Mexican Numerical Simulations School
October 3rd - 6th 2016

Lectures content

- Aim for the next few days:
  - to introduce basic observational concept
  - to provide a path on how to create fake universes
  - to connect real observations with simulations by a clear understanding of observational measurements
  - to provide a background on how to compare simulations with observational data
  - …

- A possible take home message from these lecture could be:

  Do not forget to discuss observational data and measurements with those who gathers them to properly understand them

Programme for lectures 1 & 2

- Aim:
  - (brief & biased) overview of observational cosmology
    - survey types
    - survey characteristics (incl. statistical properties)
  - (brief & unbiased?) overview of statistical descriptors for large scale structure analysis:
    - 1-point statistics:
      - dN/dm (number counts)
      - dN/dz (redshift distributions)
      - LFs, SMFs, … (luminosity, stellar mass, … functions)
      - colour distributions (observed and intrinsic)…
      - …
    - multi-point statistics:
      - 2-point clustering statistics
      - …

Observational Cosmology… and what I will consider

Observational cosmology exist in many forms:

(a) large scale structure studies via photometric and spectroscopic galaxy catalogues

(b) cosmological footprint on the CMB

(c) cosmology with standard candles (e.g. SN1a)

(d) …

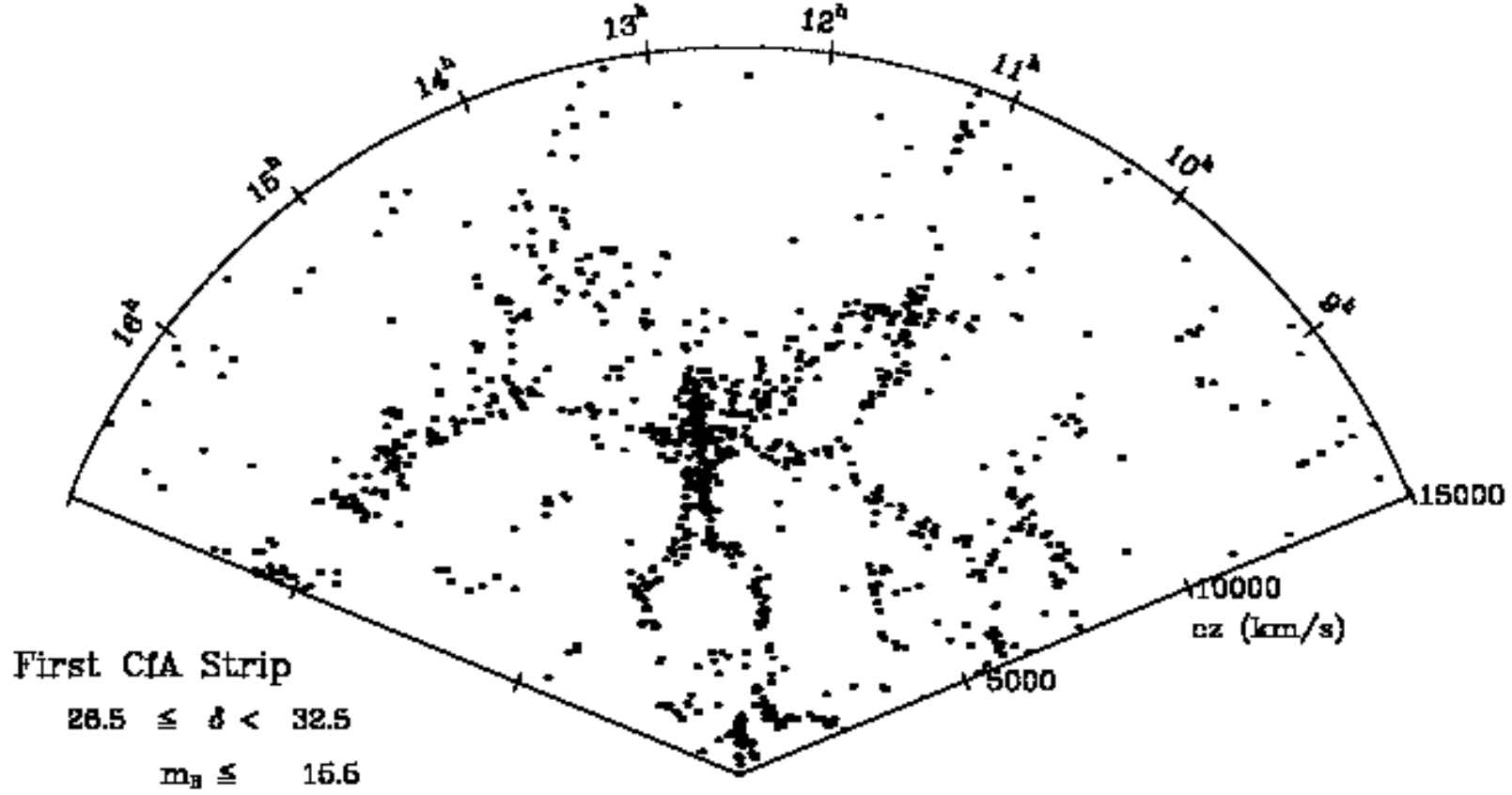
In this lecture I will only consider (a), even though all probes contribute to the definition of our cosmological model.

Aim of Large Scale Structure surveys:

- Mapping the universe using galaxies as tracers
- Cosmology and galaxy formation
- Cosmology often the main driver, even though early on main progress in galaxy formation studies
- Since Baryonic Acoustic Oscillation (BAO) measurements, design of dedicated cosmological surveys for BAO studies and redshift space distortions

CfA surveys… where it all started

- CfA survey started in 1977 by Davis, Huchra, Latham and Tonry.
- First CfA Survey (CfA), completed in 1982 with ~2k galaxies
- 2nd CfA survey (CfA2) started by Huchra and Geller in 1984 and carried on until 1995 resulting in 18k redshifts.

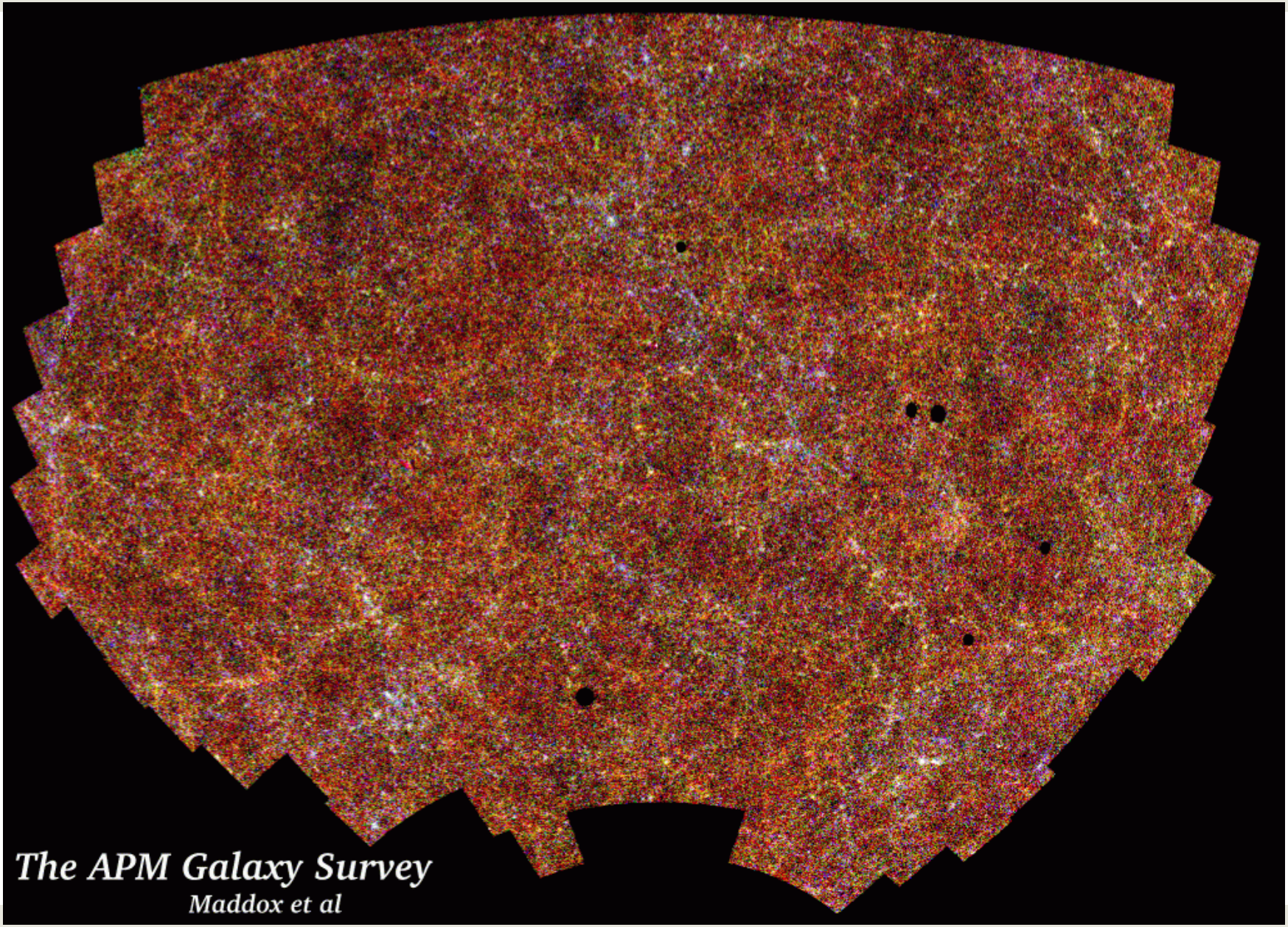


(Huchra et al . 1983)

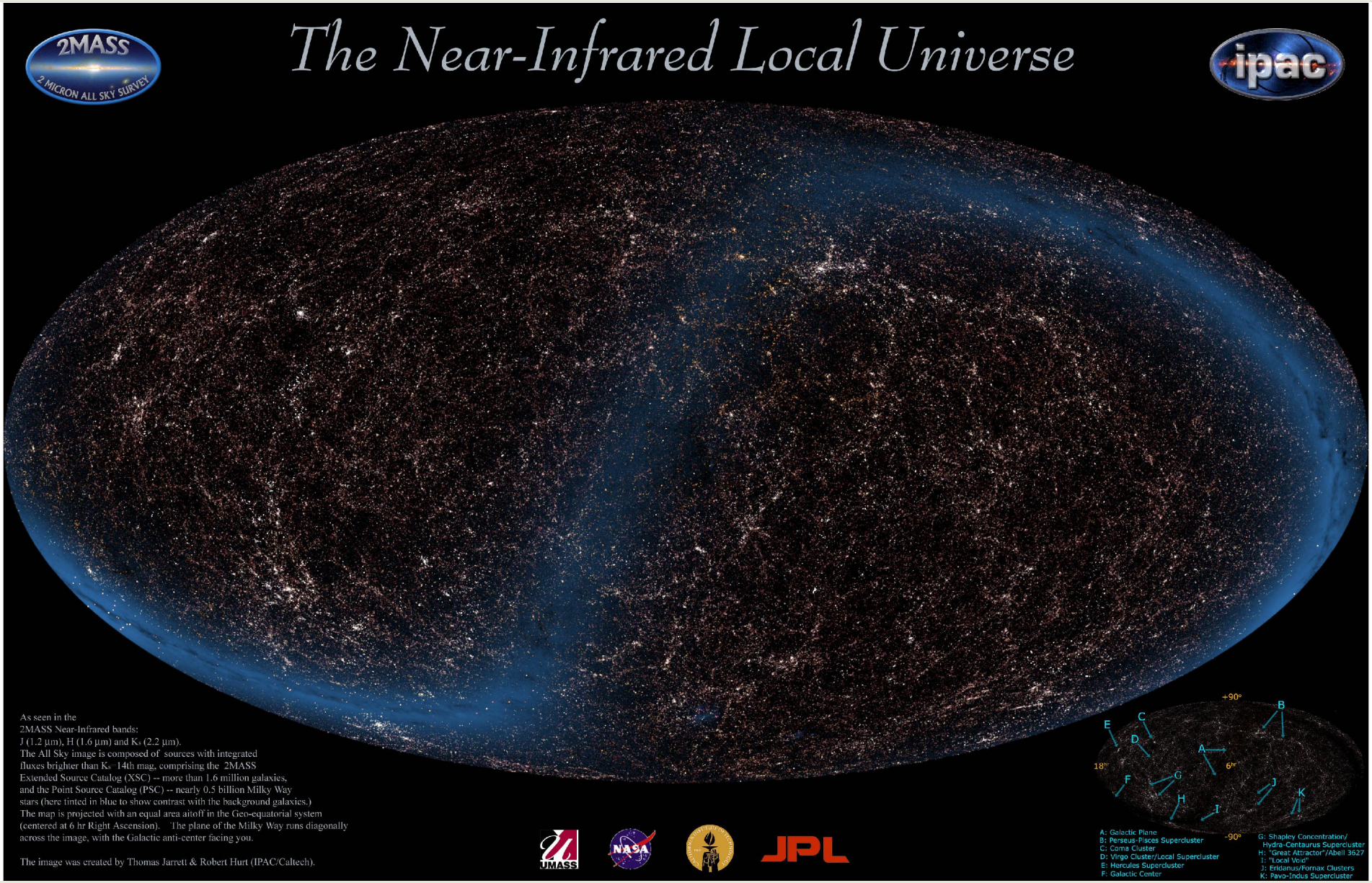
Galaxy Surveys: state-of-the-art in the 1990s

- Imaging surveys:
  - APM: scanned photographic plates of the southern sky
  - IRAS: shallow infra-red satellite survey of the whole sky
  - 2MASS: shallow near-IR survey of the whole sky with twin telescopes
  - …
- Spectroscopy:
  - CfA2 (mid-1990s): 18k redshifts
  - LCRSz (mid-1990s): 20k redshifts ($z<0.2$) over ~700 deg^2.
  - PSCz (late-1990s): 15k redshifts over 85% of the sky
  - CFRS (mid-1990s): first exploration of the z~1 universe with ~600(!) redshifts
  - …

Imaging survey: APM photographic galaxy survey



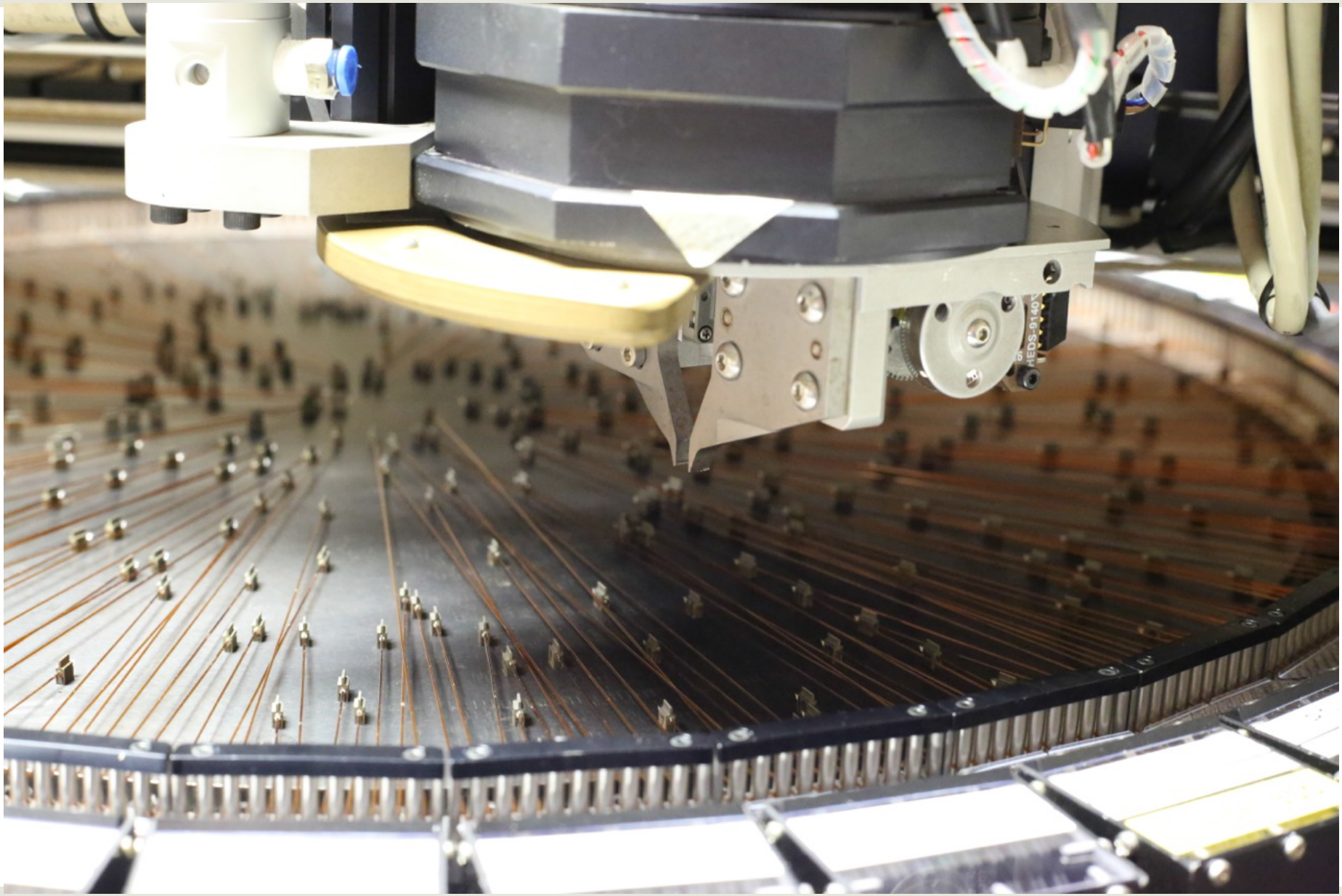
Imaging survey: 2MASS galaxy survey



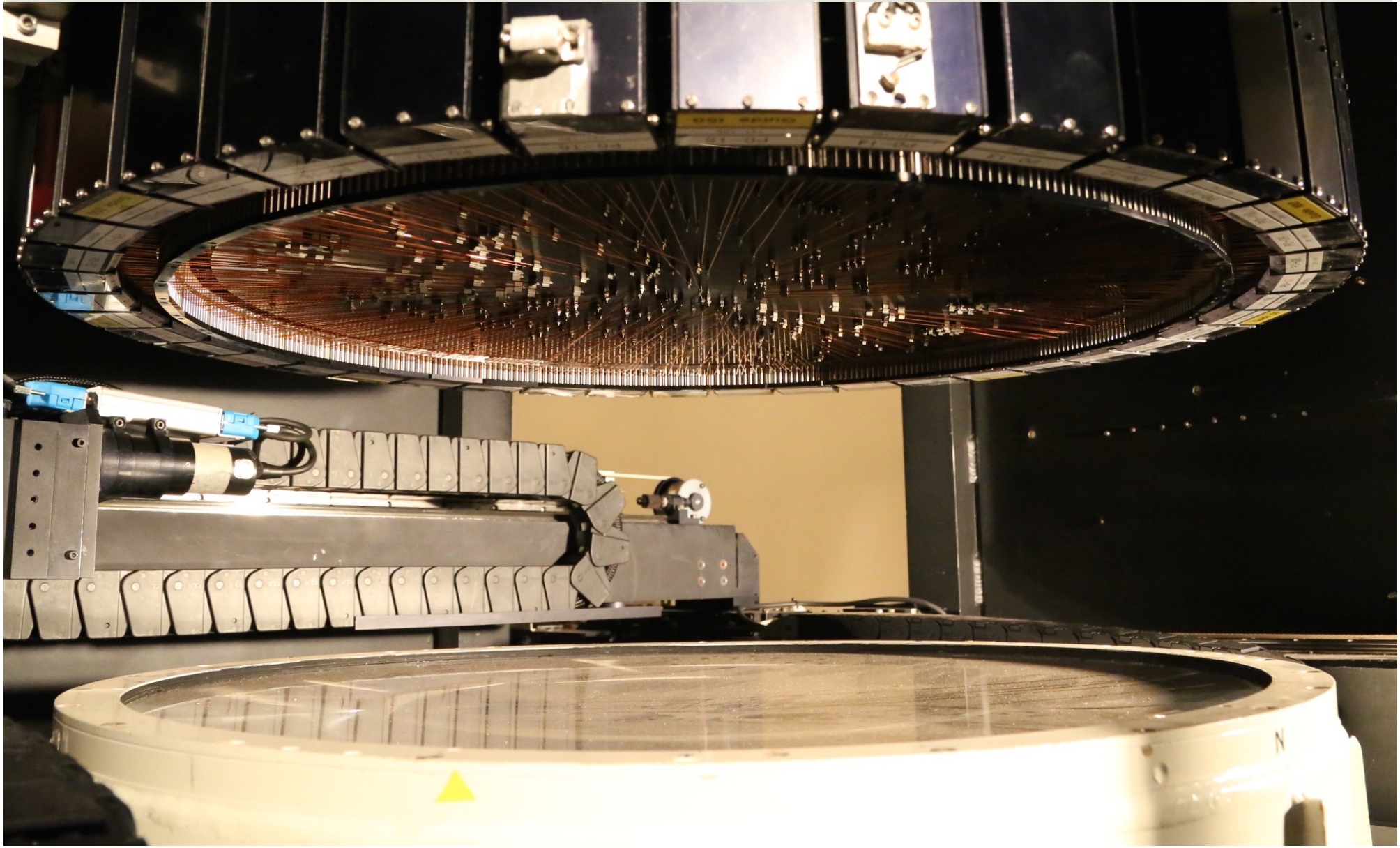
Peder Norberg, ICC & CEA, Durham University

Instrumentation: advent of multi-object spectrographs like 2dF on the AAT

Instrumentation: advent of multi-object spectrographs like 2dF on the AAT



Instrumentation: advent of multi-object spectrographs like 2dF on AAT

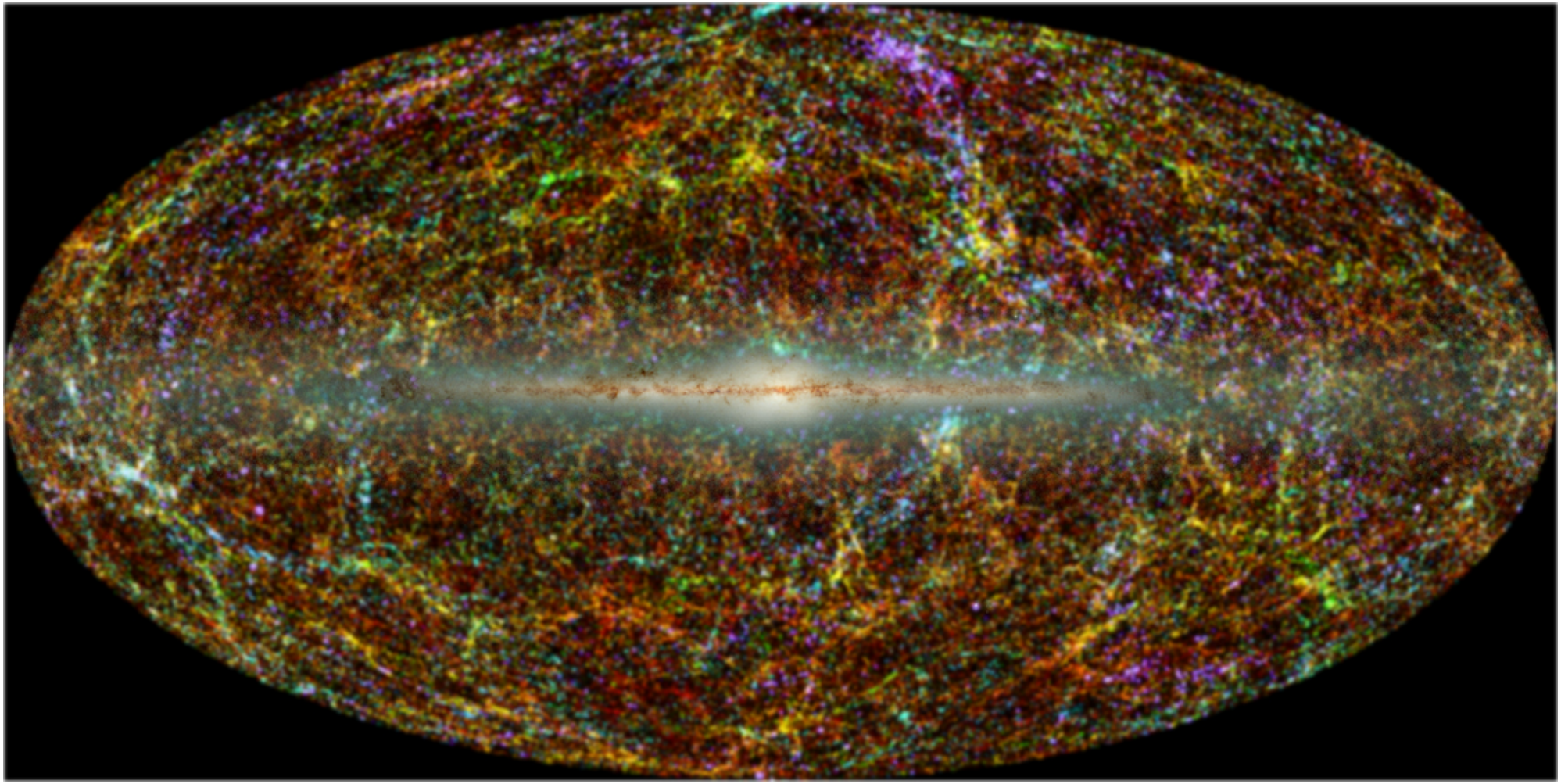


Galaxy Surveys: state-of-the art in mid-2000s

- Low redshift universe ($z<0.3$):
  - 2dFGRS: 250k redshifts over 1.5k deg^2 (by 2002)
  - 2MRS: 45k redshift over 42k deg^2
  - SDSS (I & II): ~1M redshifts (by 2008) over ~10k deg^2
- Baryonic Acoustic Oscillations (BAO) in SDSS and 2dFGRS
  - Planning of future large BAO surveys (BAO – "standard" ruler)
  - Focus on Redshift Space Distortion (RSD) probes (growth rate)
- "High" redshift universe ($z<1.4$):
  - zCOSMOS:
  - DEEP2 ($0.6<z<1.4$)
- The "high-z" survey lead to a better understanding of galaxy evolution, and provide key testbed for future cosmology surveys.

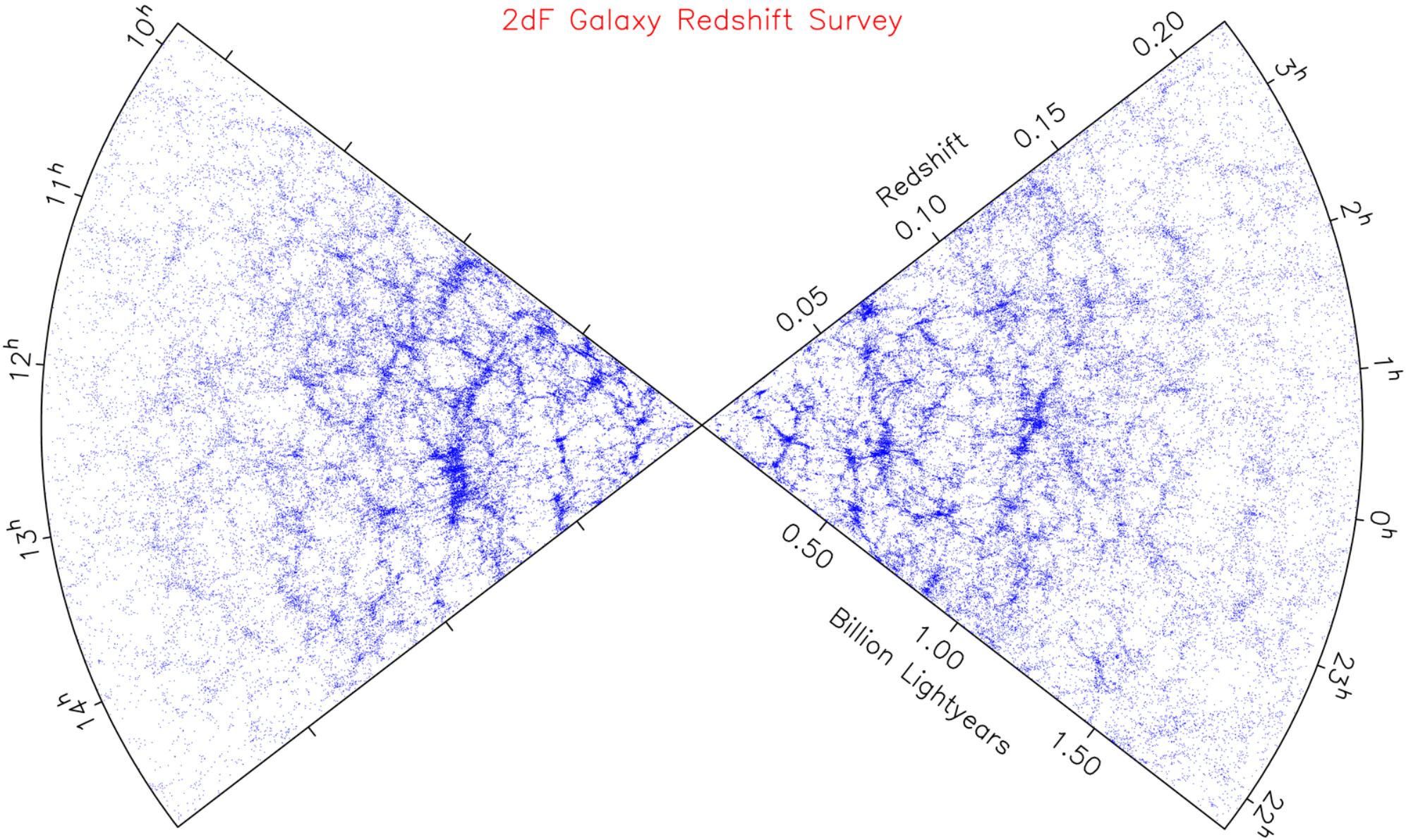
Redshift survey: 2MASS Redshift Survey (2MRS)

- 2Micron All Sky Survey: all sky imaging survey in the near-IR
- 2MRS: galaxies are colour coded according to distance away from us



(Huchra – 2011)

2dFGRS galaxy distribution

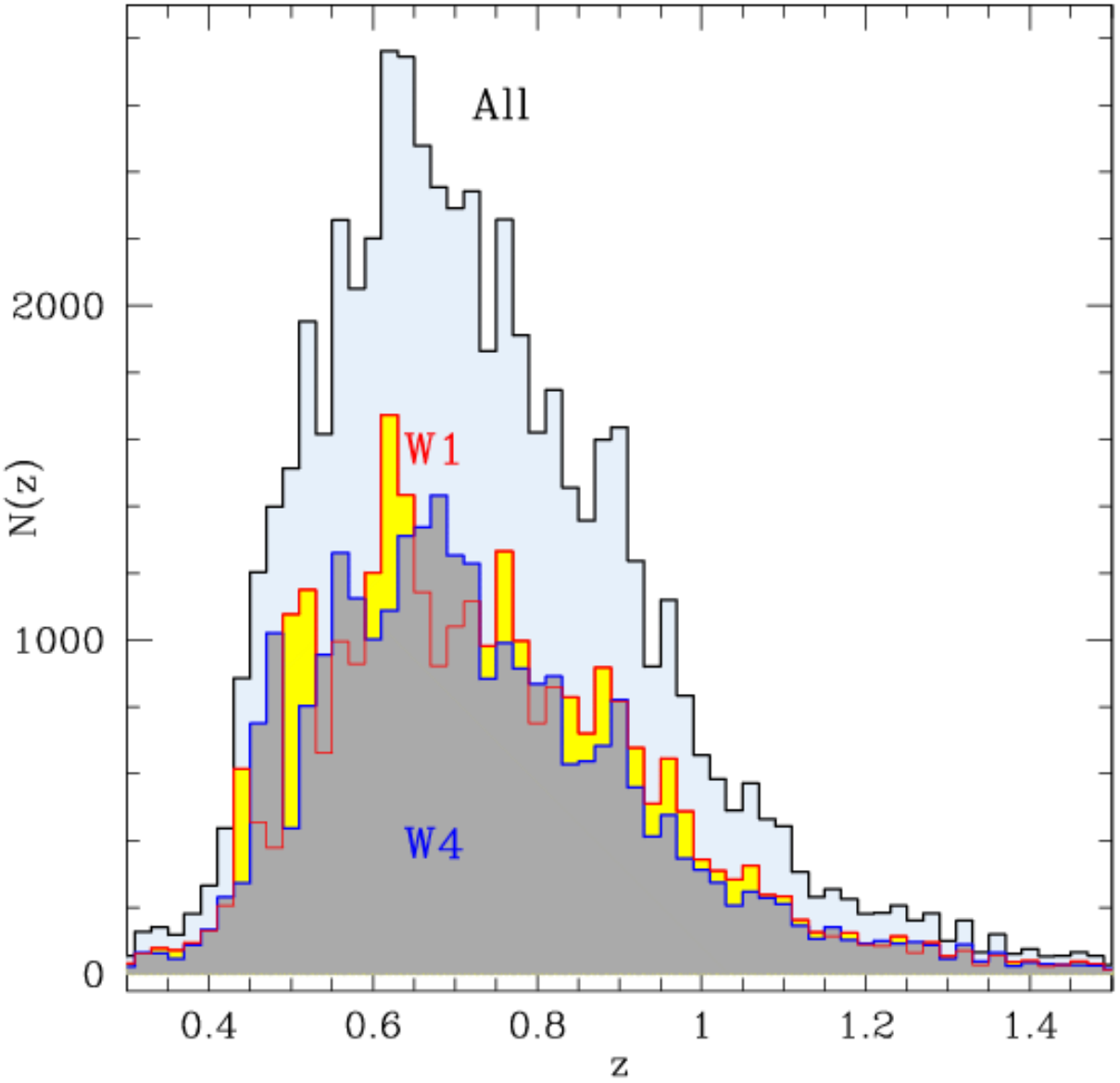


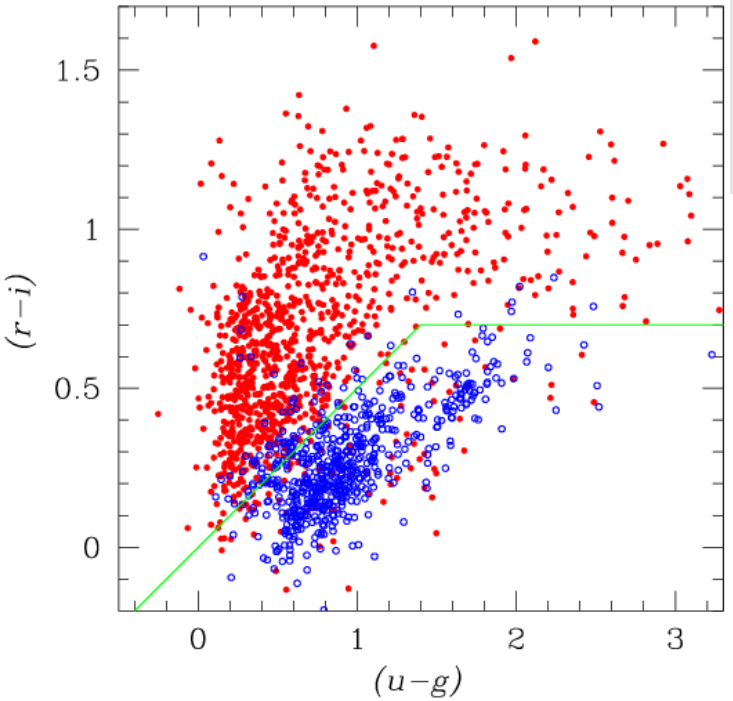
Peder Norberg, ICC & CEA, Durham University

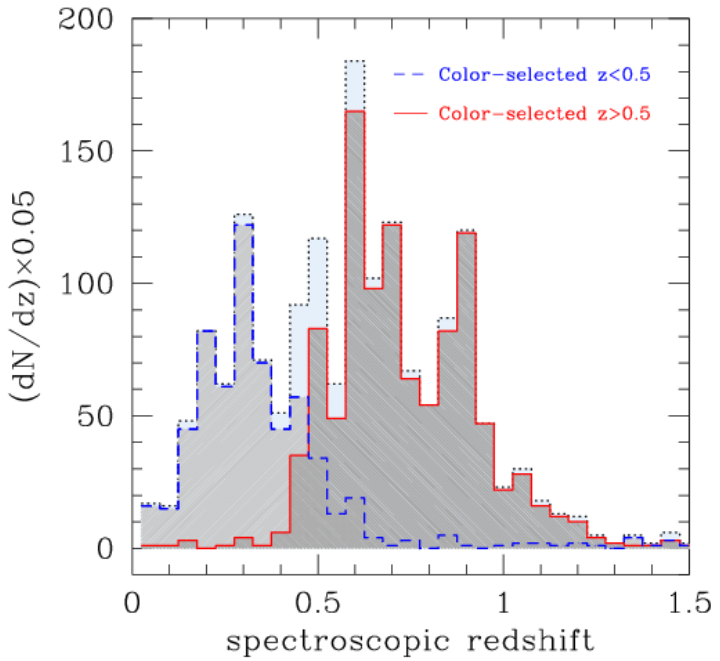
Survey Cosmology: state-of-the-art today

- Cosmological surveys:
  - Spectroscopic:
    - BOSS, eBOSS (SDSS-III, IV):
      - BAO & RSD with LRGs, ELGs and QSOs
    - VIPERS ($0.5<z<1.2$):
      - RSD with dense galaxy sampling
  - Multi-band imaging:
    - DES (6k deg^2): lensing
    - KiDS (~2k deg^2): lensing
  - …
- Galaxy formation surveys:
  - GAMA ($z<0.5$)
  - PAUS: narrow band imaging survey
  - …

Vipers survey





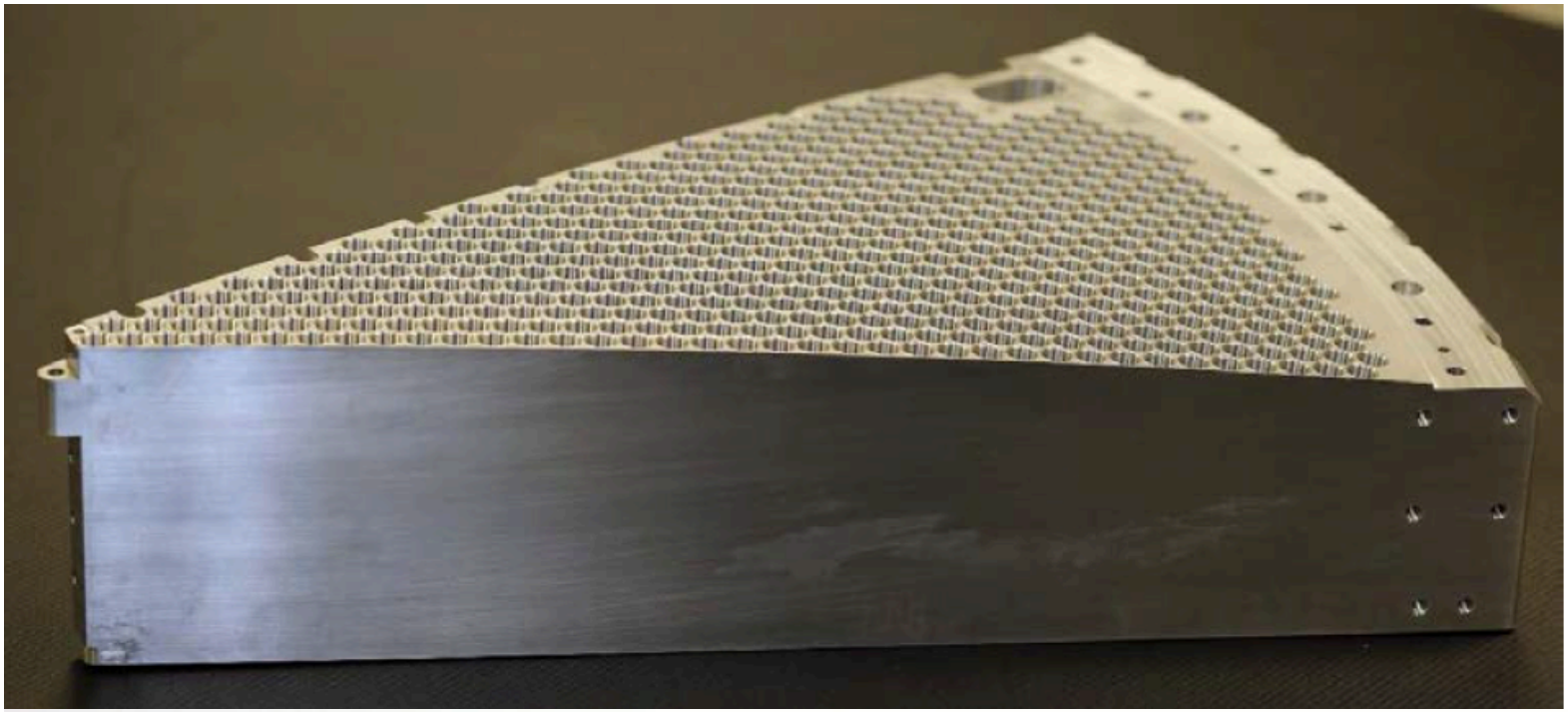


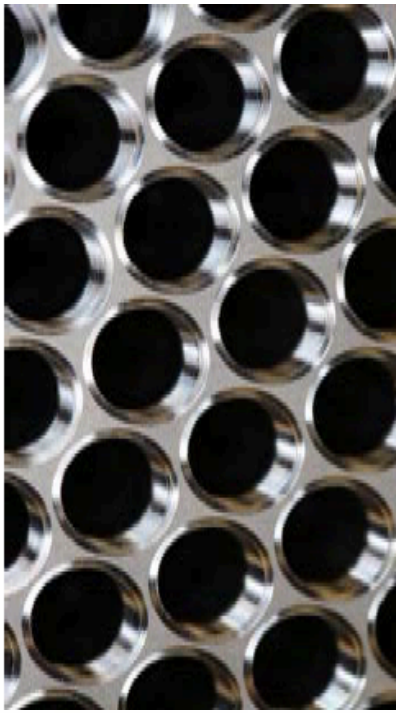
(Guzzo et al. 2014)

Advent of 2nd generation multi-object spectrographs

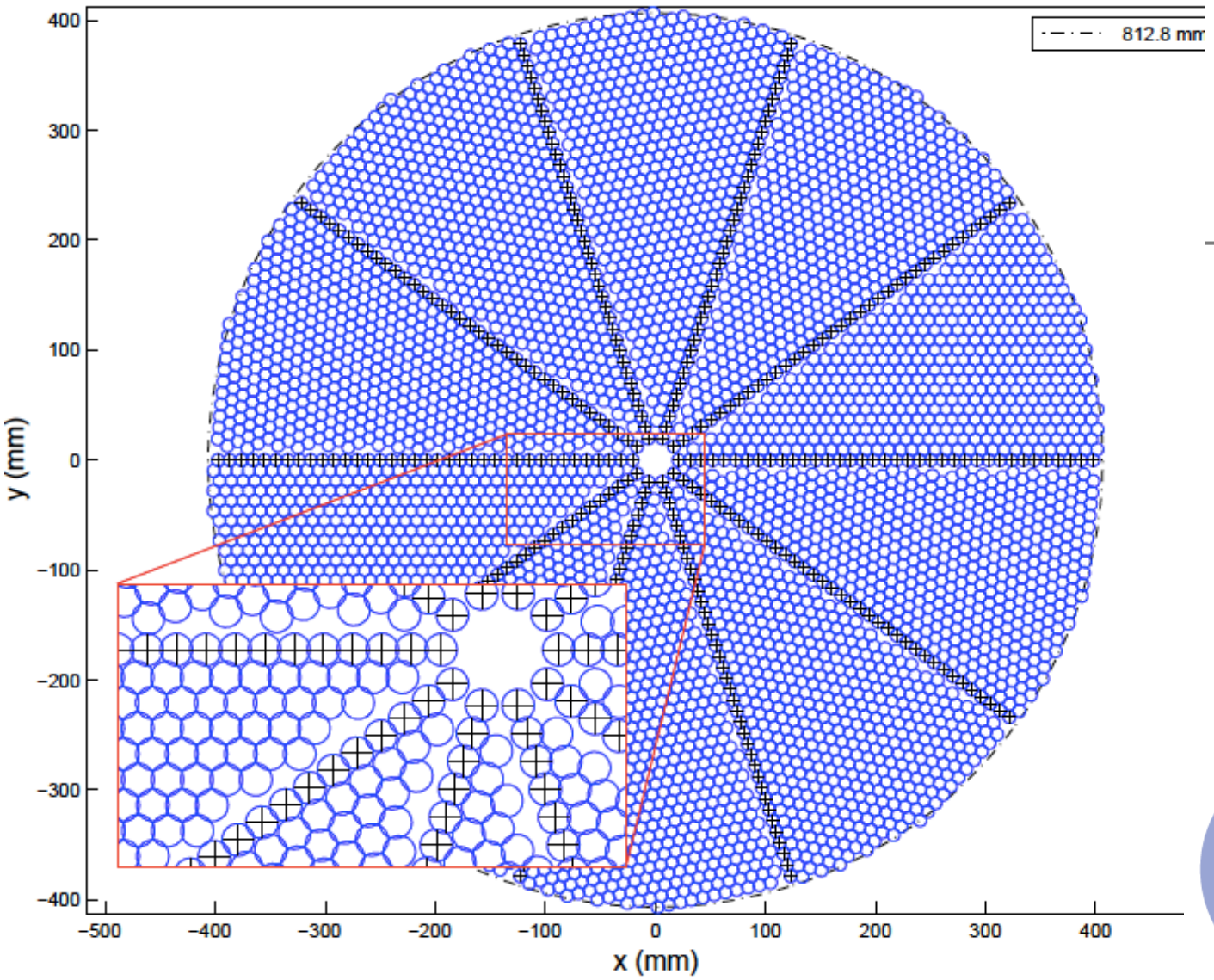
DESI petal: a 10th of the full focal plane with holes for fibre positioners:

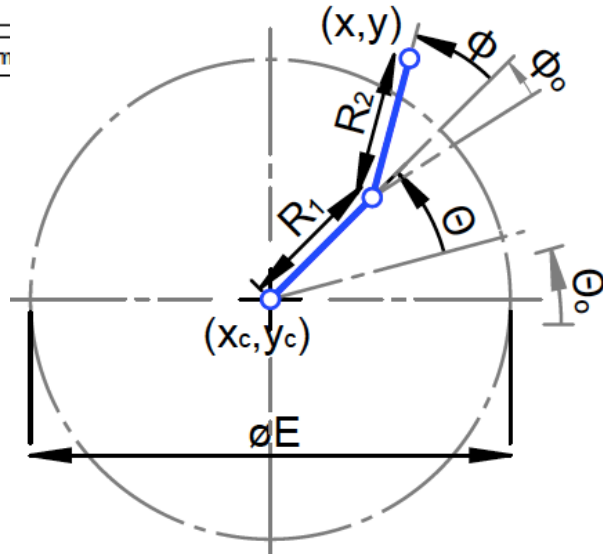
- 5000 fibres (DESI) vs 400/1000 fibres (2dF/SDSS)
- 2’ configuration (DESI) vs 50’ (2dF) (SDSS: configured “manually”)



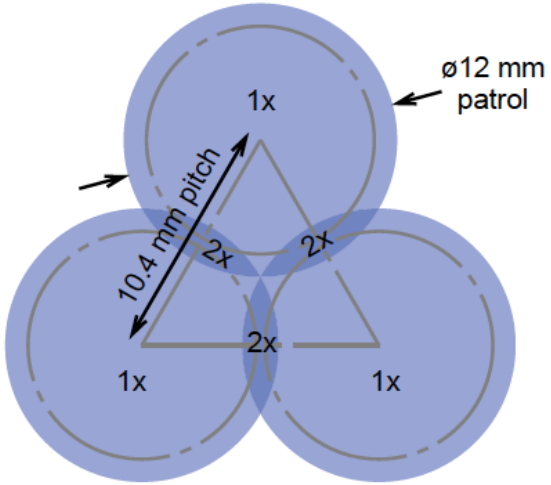


Advent of 2nd generation multi-object spectrographs



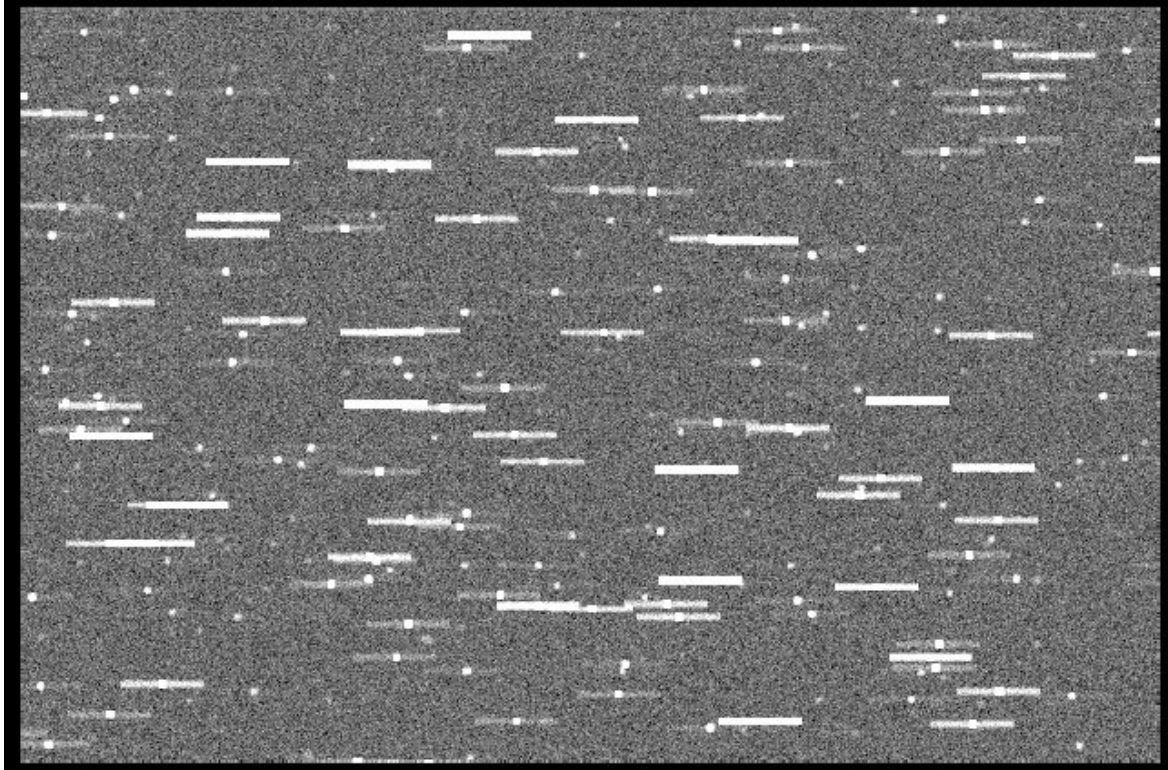


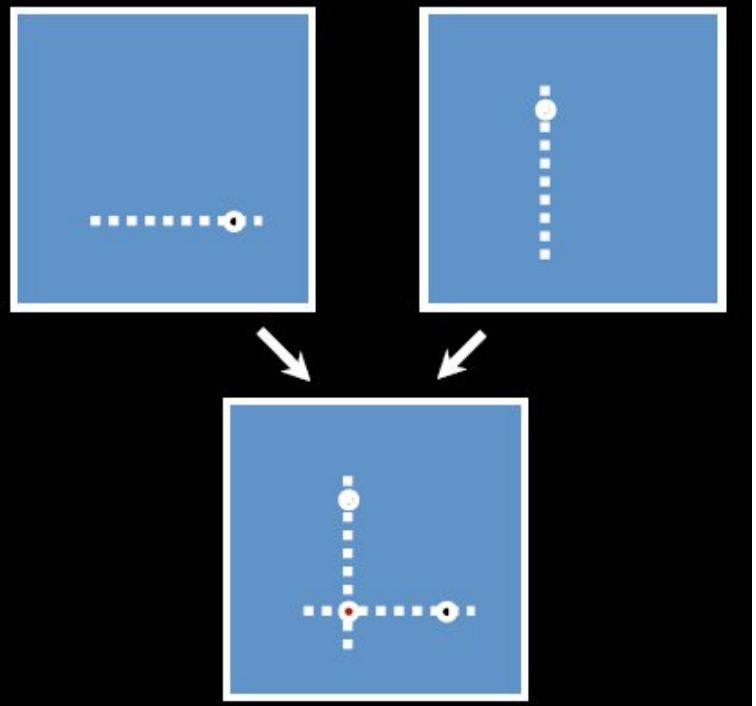
KINEMATICS / COORDINATES



Peder Norberg, ICC & CEA, Durham University

New challenges: slitless spectroscopy for Euclid



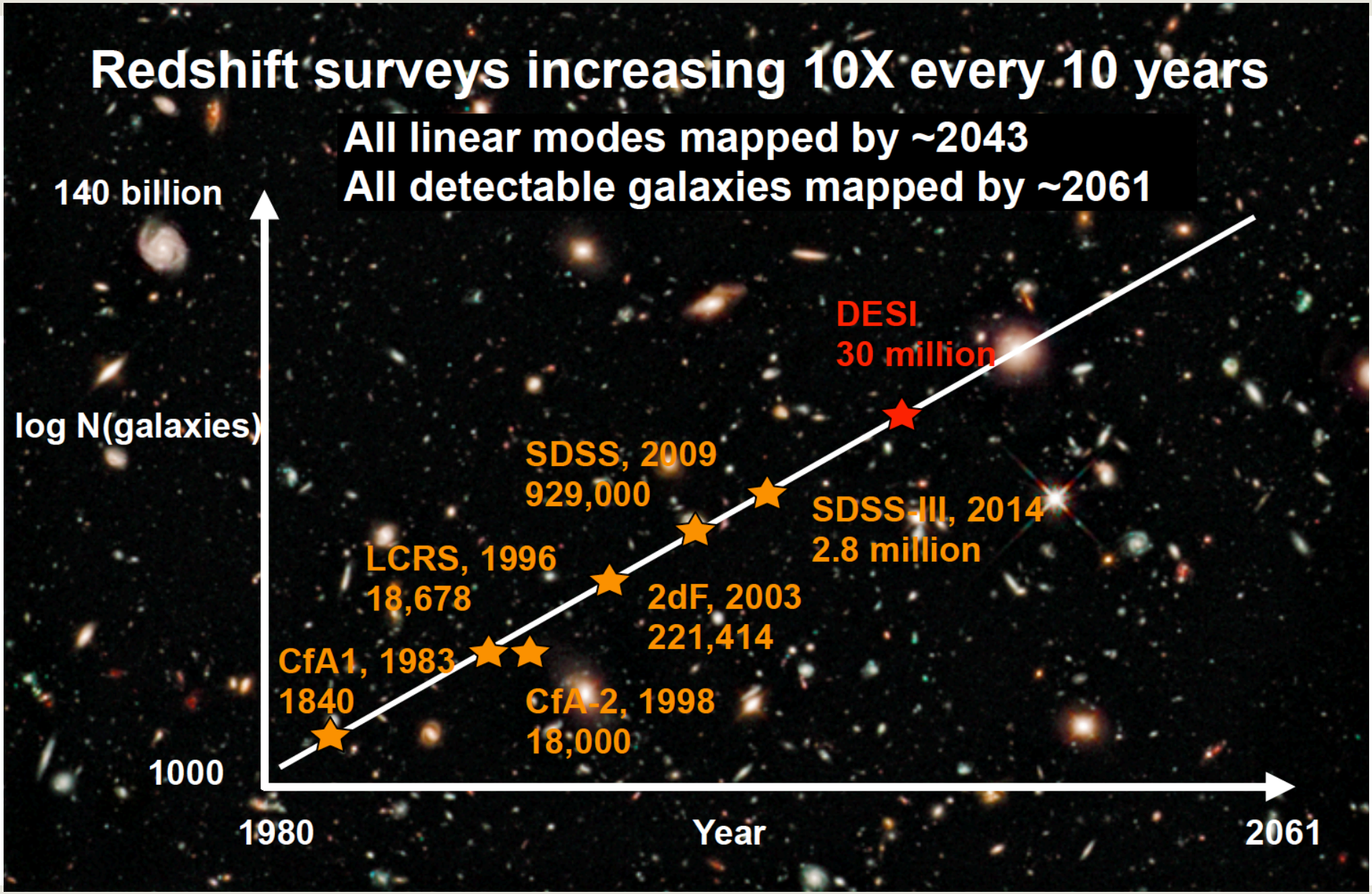


- A spectrum for every source in the field of view.
- Not restricted

Survey Cosmology: state-of-the art in ~10 years

- Cosmological spectroscopic surveys: (bright tracers)
  - PFS (~1.5k deg^2):
    - ~4M redshifts (0.6<z<2.4)
  - DESI (14k deg^2):
    - ~35M redshifts (0<z<3.5)
  - Euclid (15k deg^2):
    - ~50M redshifts (0.9<z<2.0)
    - ~1500M galaxies with shape and photo-z
  - …
- Galaxy formation surveys: (faint tracers)
  - MOONS (~50 deg^2):
    - ~0.1M spectra (0.8<z<3-4) (SDSS main at z~1.5)
  - WAVES:
    - WAVES Wide (~1k deg^2): ~1M redshifts (z<0.25)
    - WAVES Deep (~0.1k deg^2): ~1M redshifts (z<1)

Spectroscopic redshift acquisition with time



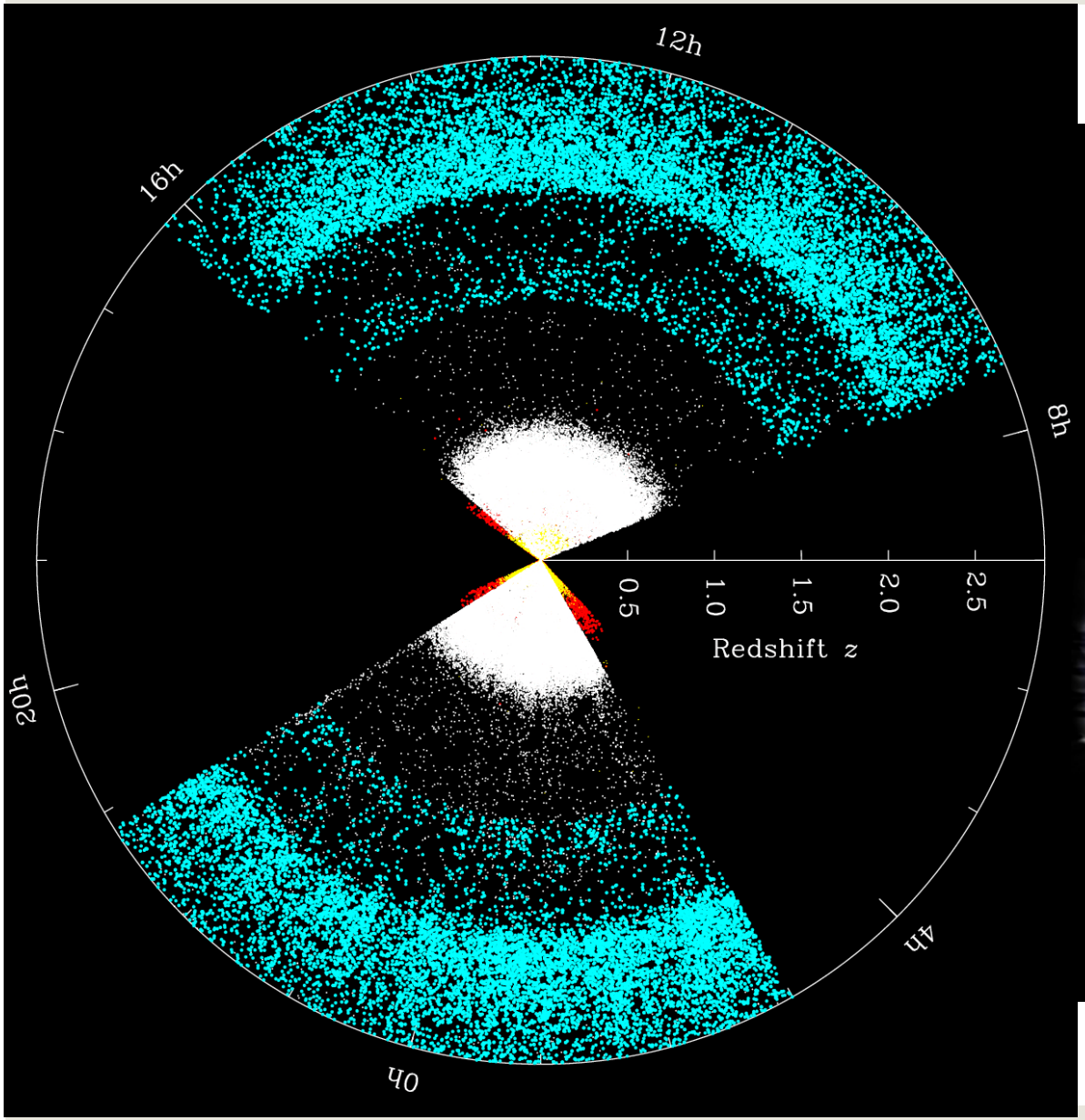
(Schlegel – 2016)

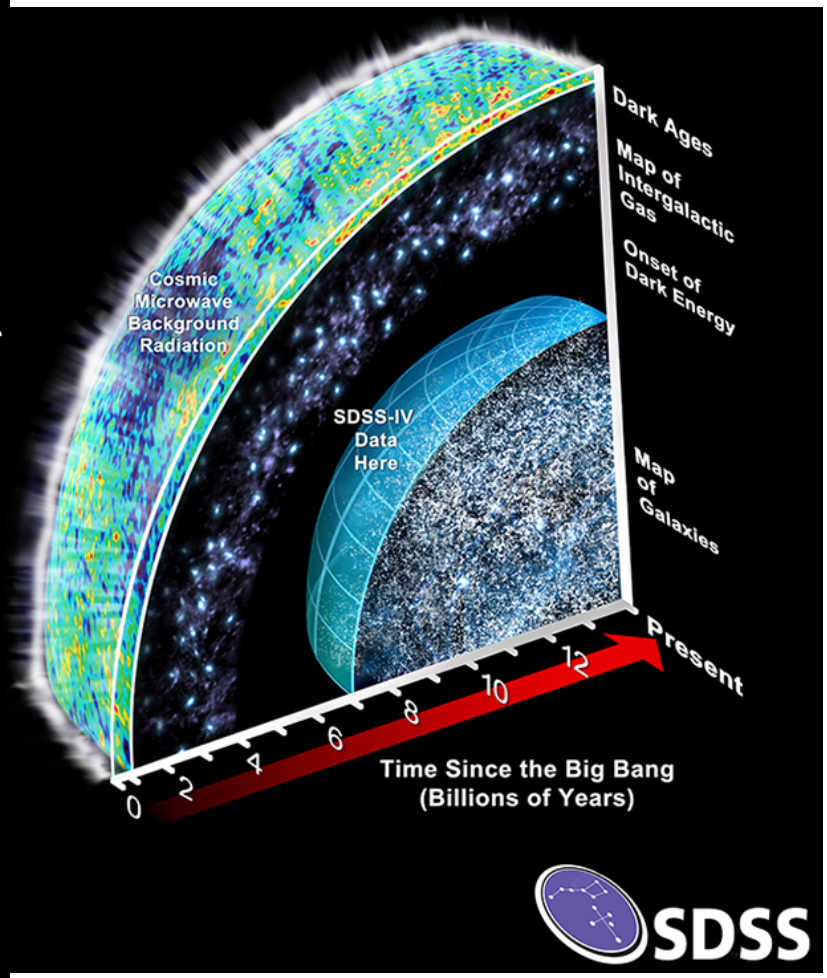
Three types of galaxy surveys (I)

- Cosmological spectroscopic surveys:
  - Large to extremely large volumes
  - Bright galaxies as tracers, often strongly biased (but not always!)
  - Strong emphasis on targeting algorithm and sampling rate
  - Statistical probe: BAO and/or redshift space distortions (RSD)
  - Key issue:
    - modeling the survey selection function to sub-% accuracy
    - …

- Examples:
  - SDSS-I/II LRG
  - VIPERS
  - SDSS-III/IV BOSS/eBOSS
  - PFS
  - DESI
  - Euclid spectroscopic

SDSS-III BOSS





(SDSS III – 2015)

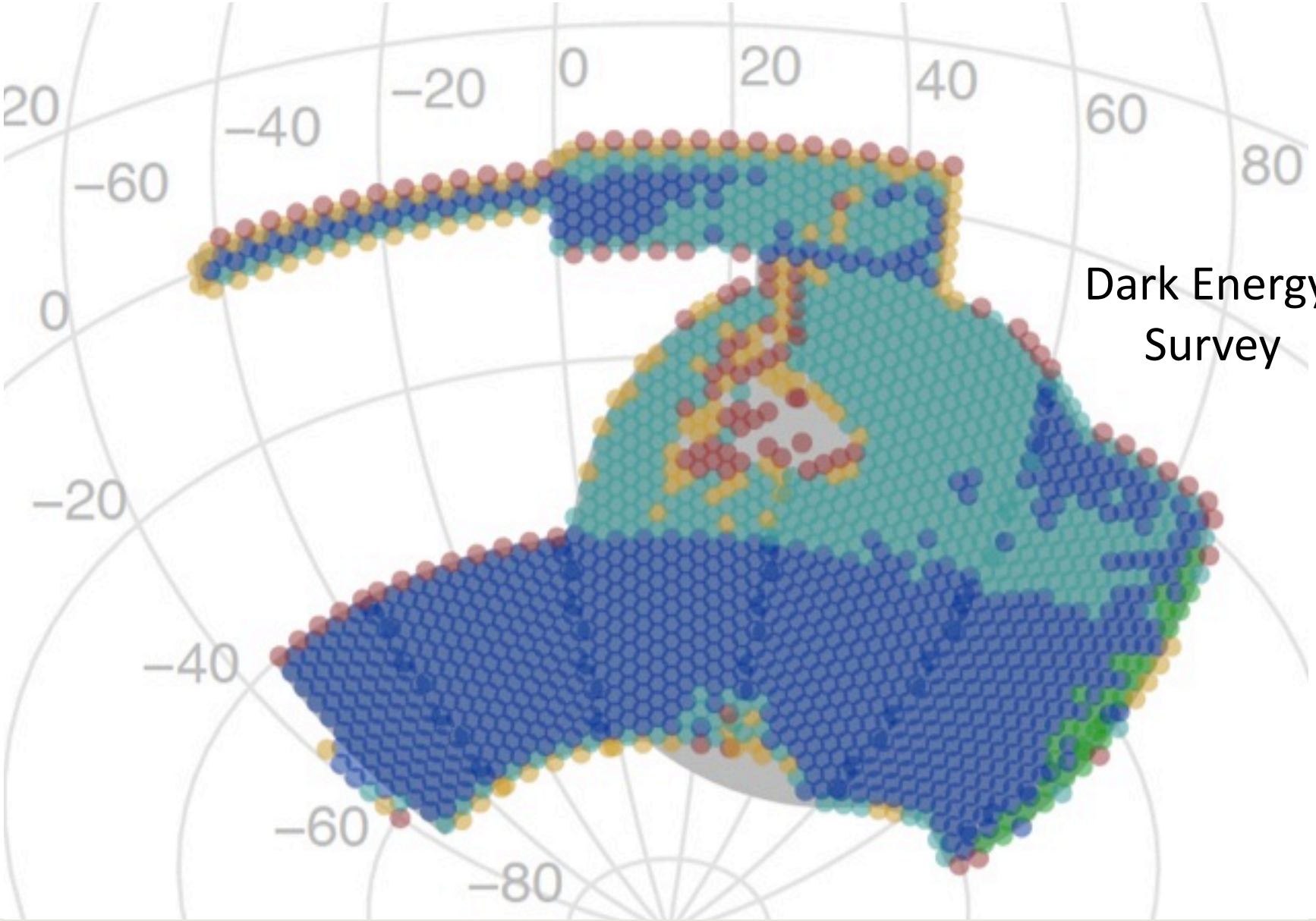
Peder Norberg, ICC & CEA, Durham University

Three types of galaxy surveys (II)

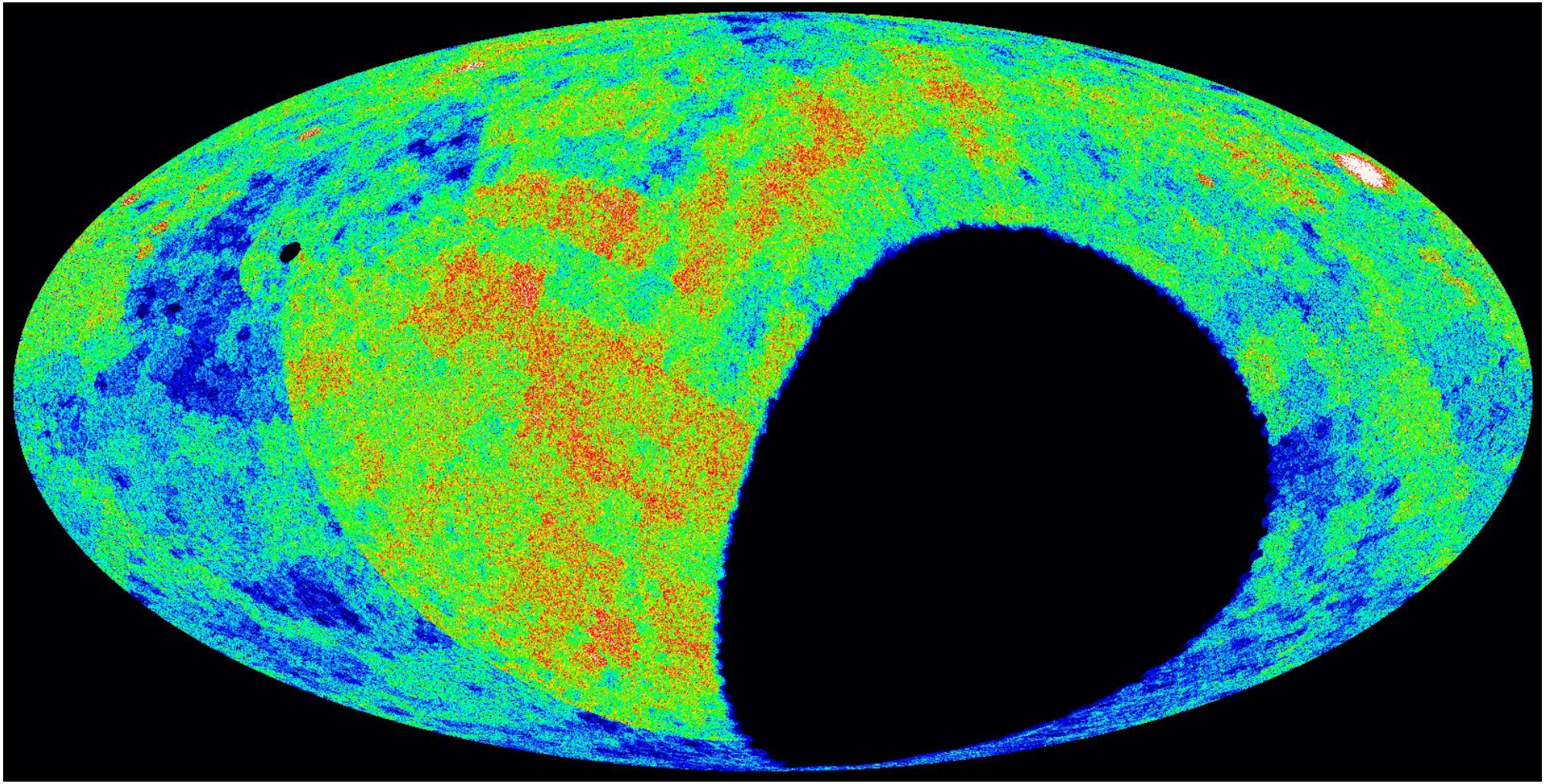
- Cosmological imaging surveys:
  - Large to extremely large volumes
  - All galaxies to some flux limit in some band
  - Strong emphasis on:
    - multi-wavelength (for reliable photo-z)
    - homogeneity of the imaging dataset (spatial and depth)
    - prime observing conditions (key for lensing studies)

- Examples:
  - 2MASS / IRAS / APM
  - SDSS: u, g, r, i, z over ~10k deg^2
  - DES: g, r, i, z, y over ~6k deg^2
  - DECaLS/BASS/MzLS: g, r, z over ~10k deg^2
  - KiDS / HSC: u, g, r, i / g, r, i, z, y over ~1-2k deg^2
  - Euclid imaging: riz , Y, J, H over 15k deg^2
  - LSST: u, g, r, i, z, y over the southern sky over 10 years

Dark Energy Survey: sky coverage (tiling)



Imaging surveys: PanSTARRS

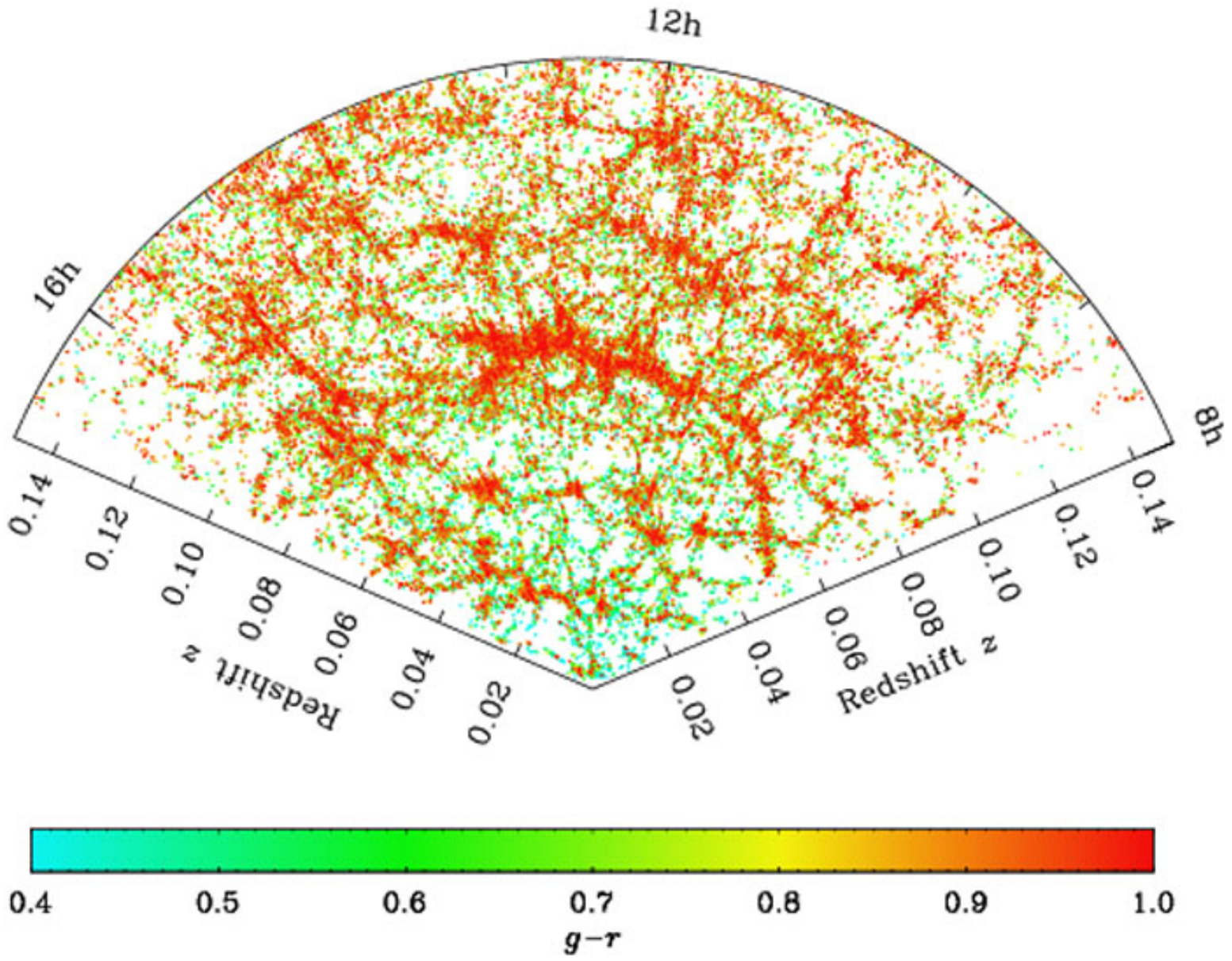


Three types of galaxy surveys (III)

- Galaxy formation surveys:
  - High S/N spectroscopy, including spectral flux calibration
  - Multi-wavelength coverage:
    - different physical processes probed with different wavelengths
  - Large enough volumes (to sample “all” environments “well”)
  - All galaxies down to some flux limit:
    - L* and fainter (at least)
  - High uniformity and spatial completeness on all scales:
    - Key for environment / group studies
  - Analysis: often focus on volume limited samples

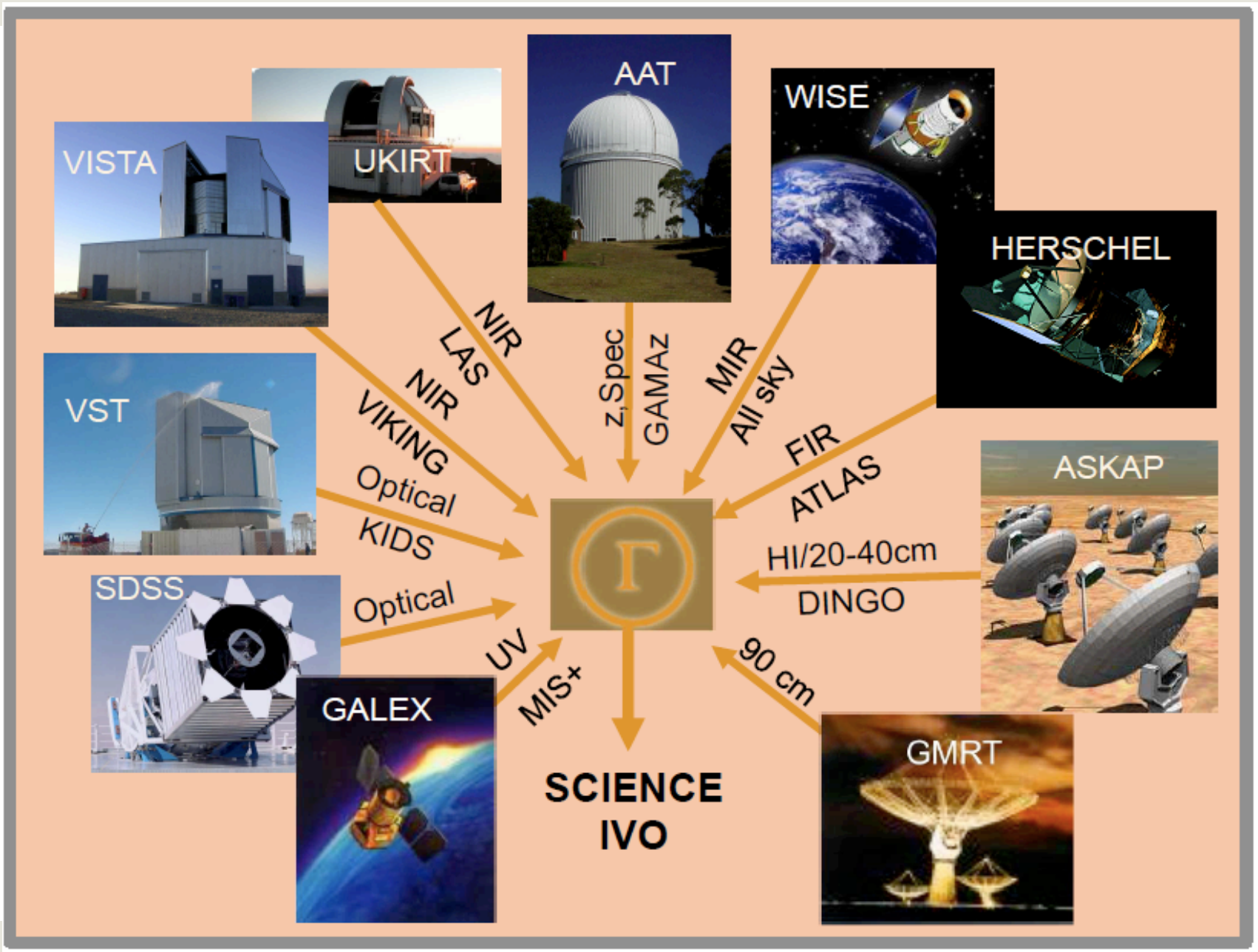
- Examples:
  - SDSS main (r<17.7); Issue: close pair spatial completeness
  - GAMA main (r<19.8); Issue: spectral flux calibration
  - Most “smaller” surveys, like COSMOS, DEEP2, VVDS, …
  - By 2020 / 2022, MOONS / WAVES

Large scale structure in SDSS main survey

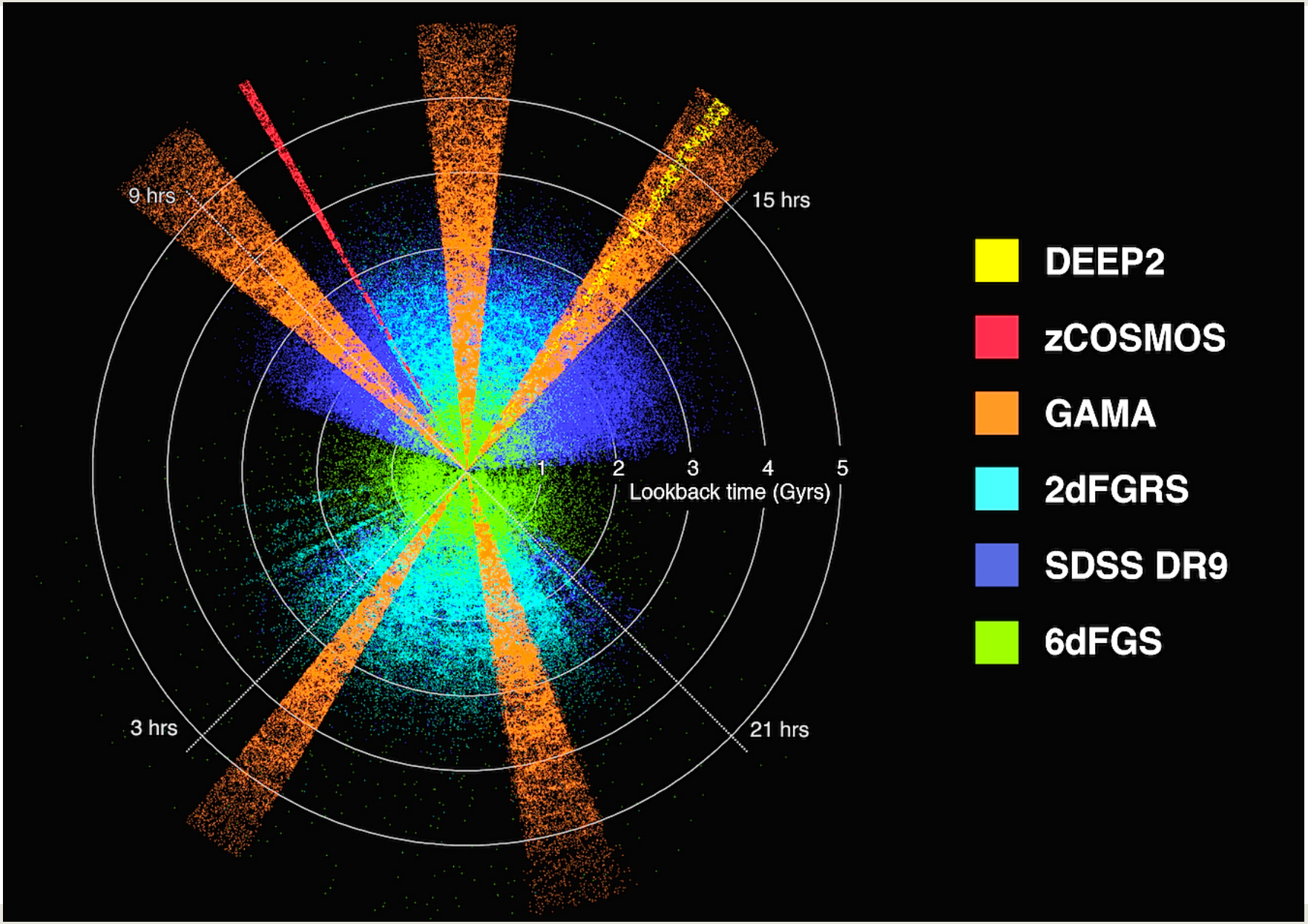


(Zehavi et al. 2011)

Galaxy formation survey: GAMA

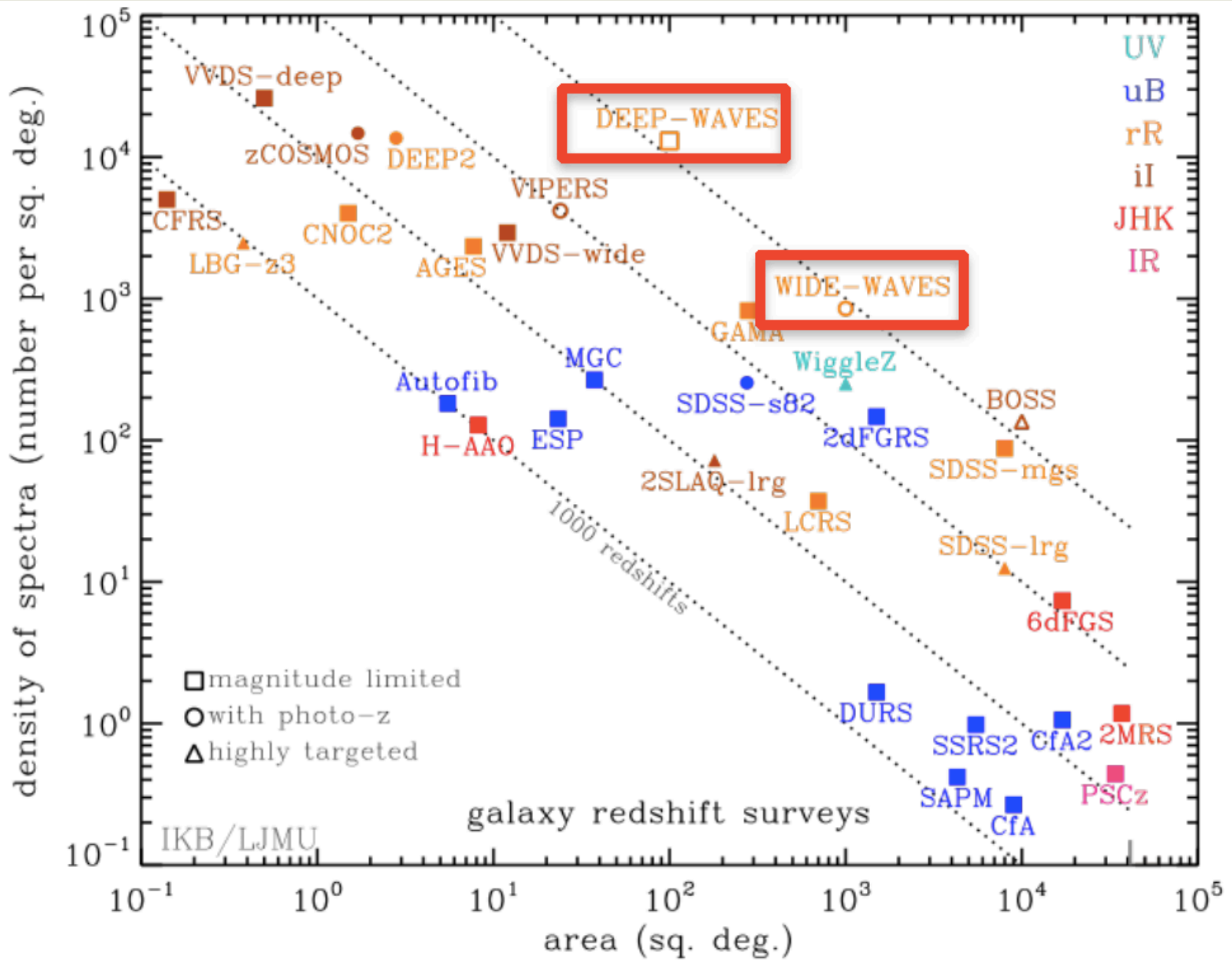


Survey comparisons: “coneplot”



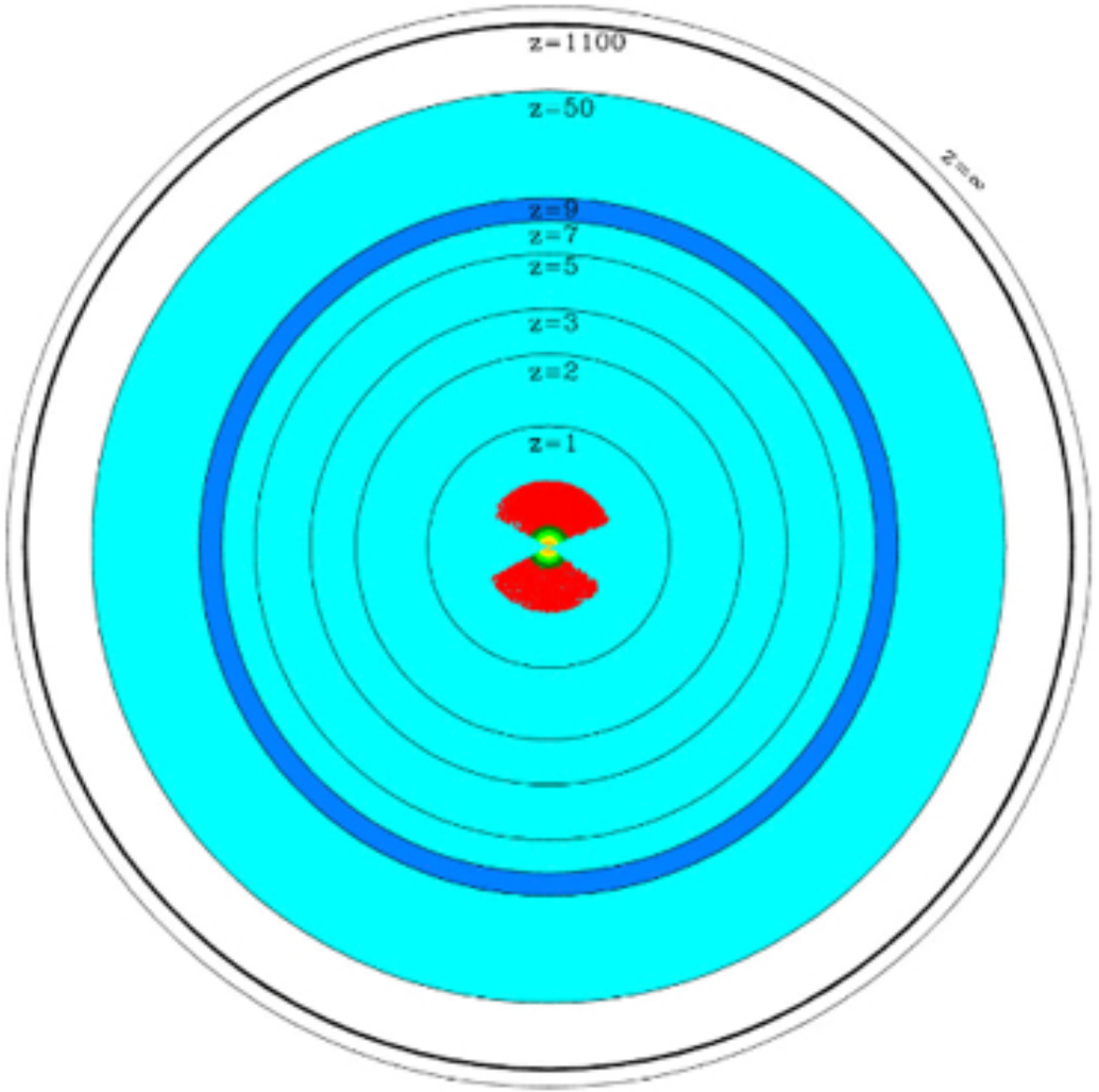
Peder Norberg, ICC & CEA, Durham University

Survey comparisons: density of spectra vs area



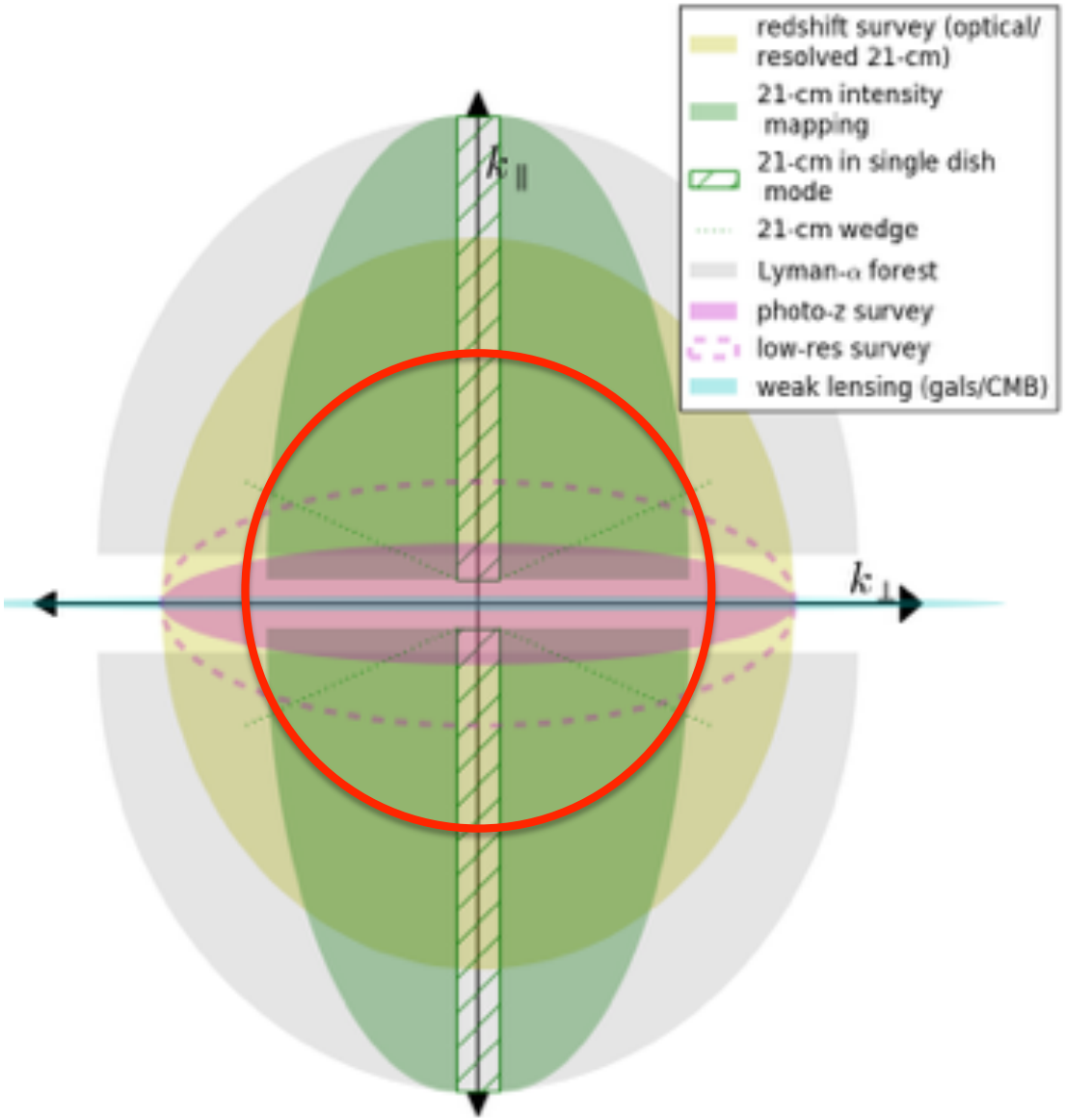
(Baldry – 2010+)

But this is the volume available and what has been sampled so far…



(Schlegel & Slosar – 2016)

And in terms of modes this is what can be sampled



(Schlegel & Slosar – 2016)

Survey characteristics

- Basic survey characteristics:
  - number of galaxies
  - area and/or effective area
  - flux limit / target selection
  - redshift range ($z_{min} < z < z_{max}$) and galaxy number density (nbar)
  - …
- Advanced characteristics:
  - selection function (angular and radial)
  - underlying galaxy population:
    - Luminosity function
    - Stellar mass function
    - …
  - clustering properties:
    - Angular: $w(\theta)$
    - 3-D: $\xi(s)$, $\xi(r_p, \pi)$, $w_p(r_p)$, …
    - …

Why do survey characteristics matter?

- All surveys have inherent selection effects:
  - flux limit (i.e. not volume limited)
  - completeness (i.e. not everything that could be seen is included)
  - …
- All numerical simulations have inherent limitations:
  - spatial and temporal resolution
  - numerical accuracy
  - physics implemented
  - …
- Spelling out limitations and selection effects explicitly is essential for any simulation – observation comparison.

Which survey characteristic matter?

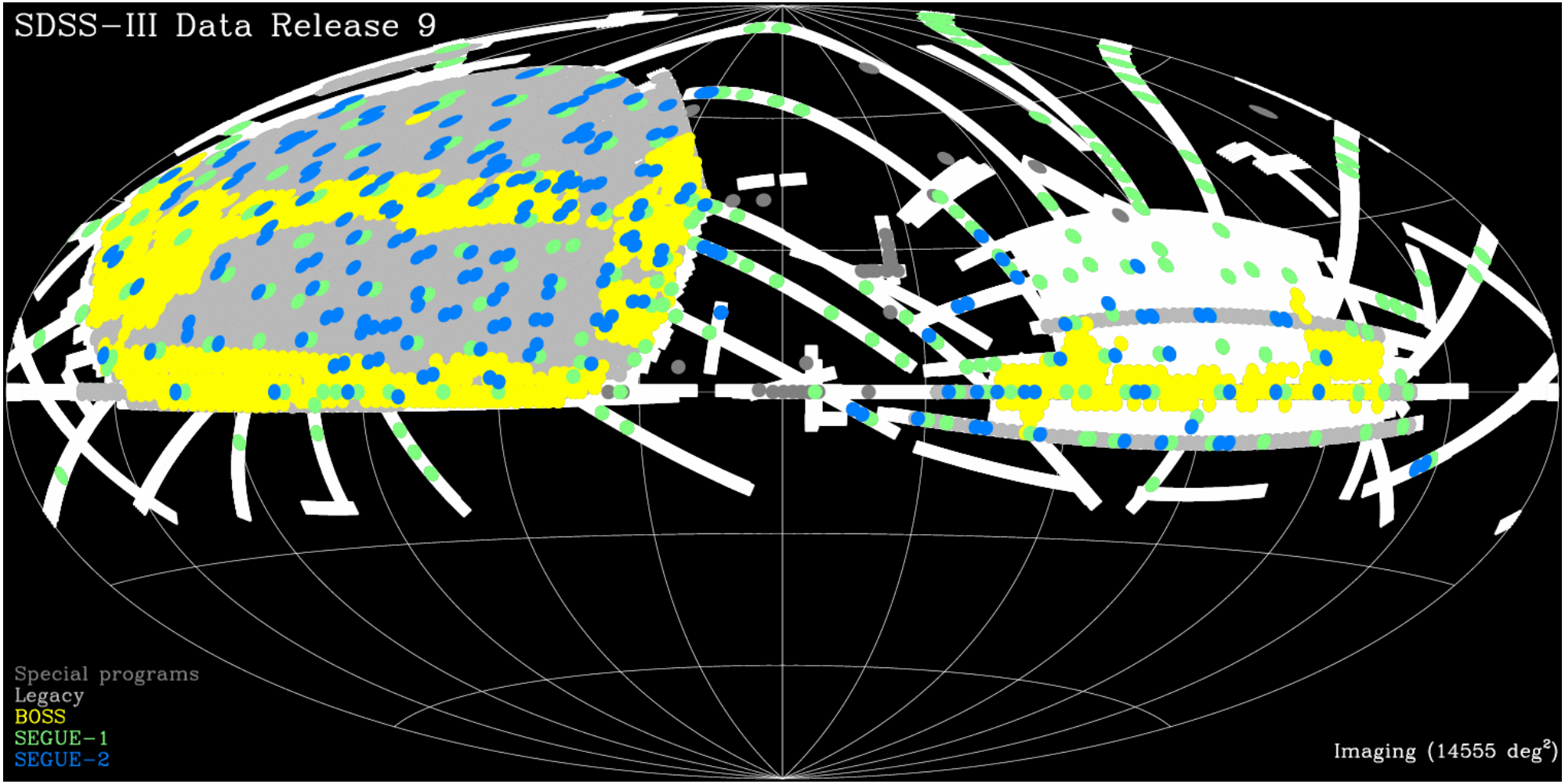
- It all depends on:
  - specific science goals
  - legacy value of survey

- Example:
  - SDSS main survey has a ~99% redshift completeness overall, but only 60-70% complete for galaxies with neighbours within 50'' of each other:
    - 1st statistic is great and sufficient for most sciences goals
    - 2nd impacts strongly on the analyses requiring small scale information (e.g. galaxy groups or 1-halo terms in clustering statistics).

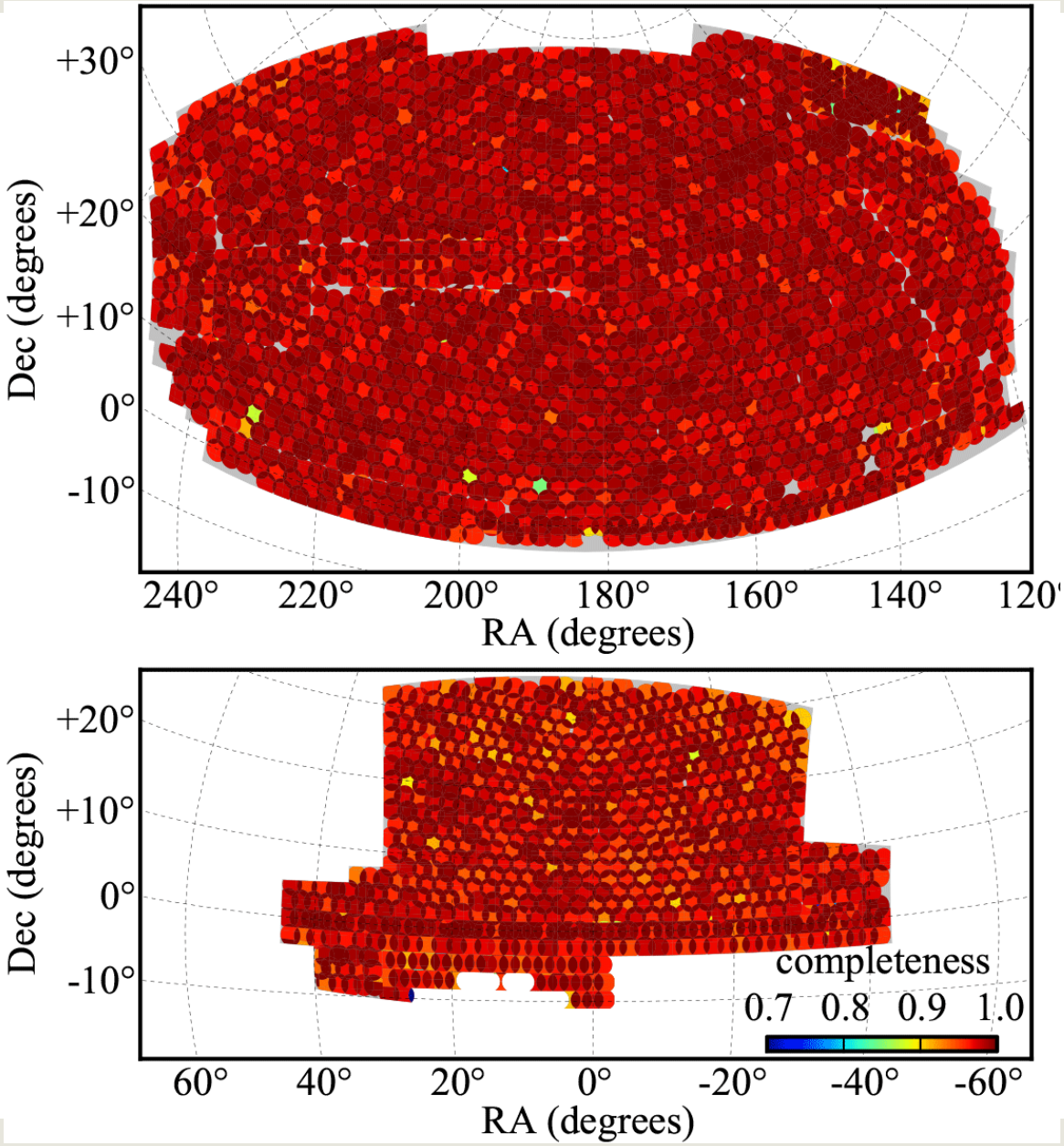
    GAMA addresses that 2nd issue, but only over ~200 deg^2.
  - SDSS imaging is generally very complete, except for low-surface brightness galaxies: not an issue for cosmological surveys with SDSS, but needs to be accounted for in galaxy formation studies.

Sky coverage: many options until…reality kicks in!

- SDSS-III (DR9) survey footprint: all surveys combined



SDSS-III: BOSS DR12 spatial coverage



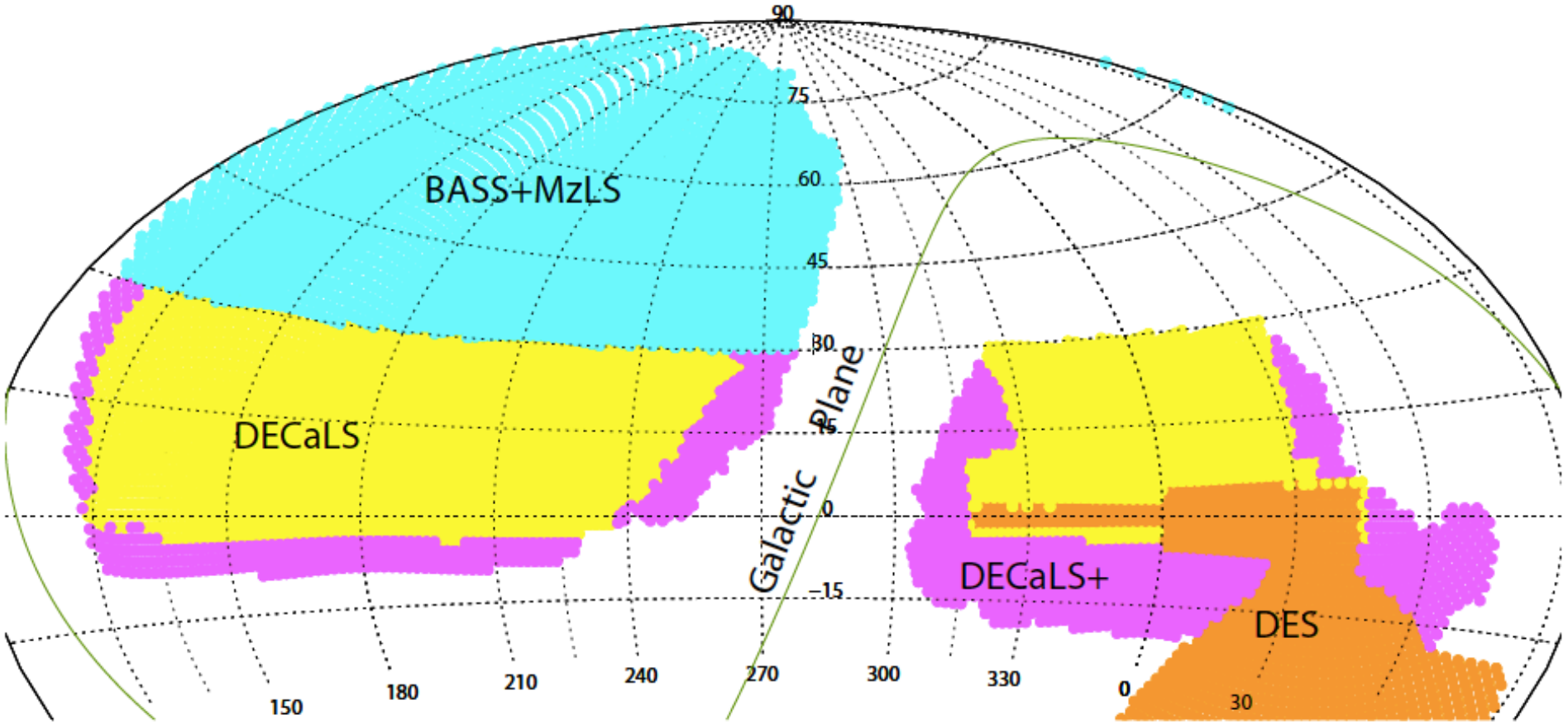
SDSS DR12 BOSS

Northern

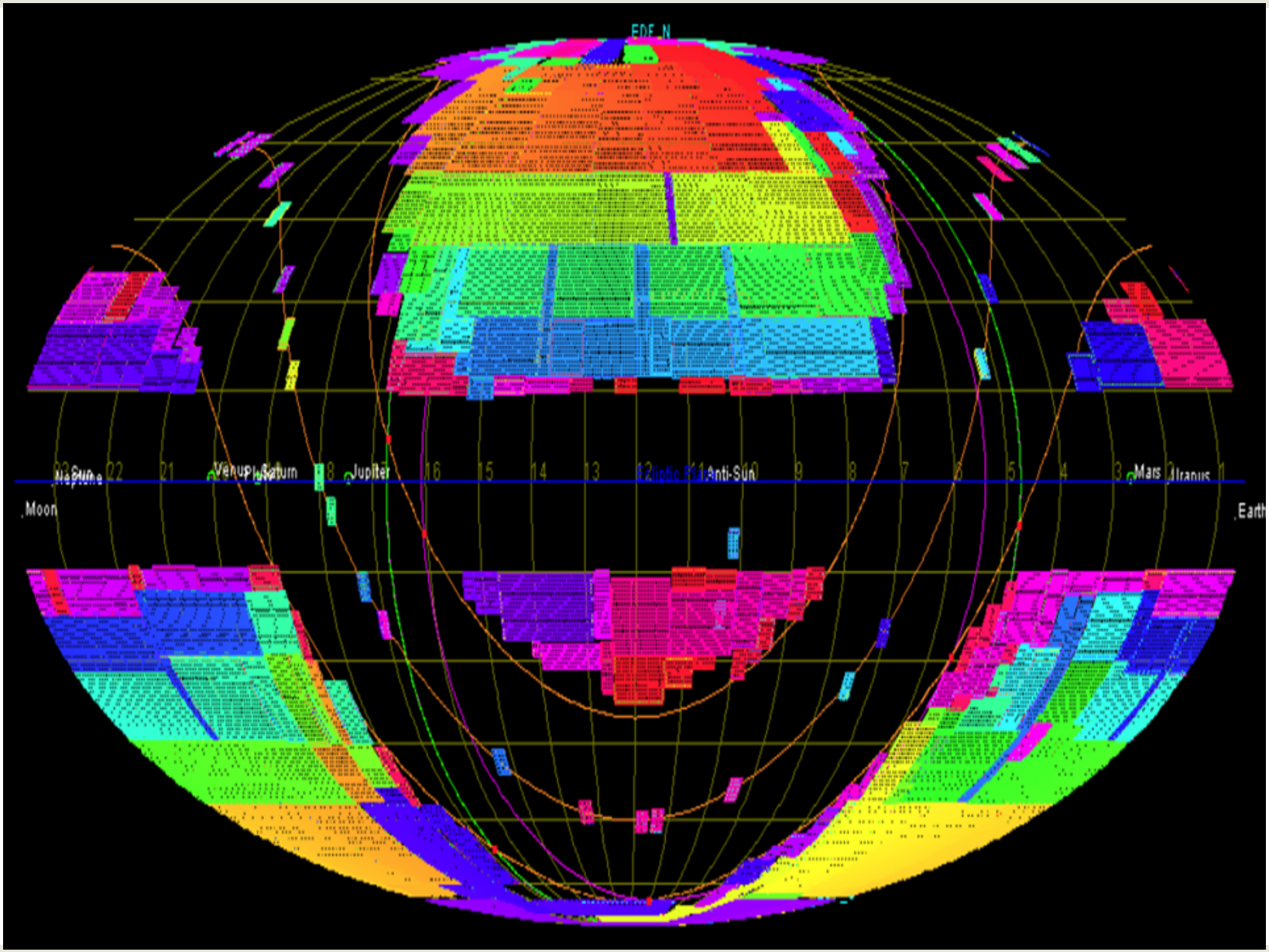
Southern

Sky coverage: simple in principle until…

DESI: Dark Energy Spectroscopic Instrument



Sky coverage: the Euclid options…



Peder Norberg, ICC & CEA, Durham University

Reasons behind those “crazy” survey footprints: basic observational astronomy concepts

- The Earth’s rotation axis is inclined w.r.t. its orbit around the sun:
  - Not all sky positions are available at any time
  - Observing “seasons” are influenced by:
    - Weather conditions (…)
    - Length of nights
    - Location of telescope

- The moon rotates around the earth:
  - The brightness of the sky varies with time (lunations) and affects bluer wavelengths more.

- We live in the Milky-Way:
  - Not all parts of the sky are observationally statistically equivalent for extra-galactic studies (e.g. galactic plane)

- Time allocation constraints (incl. politics and additional sciences):
  - Not all sky positions have equal observing pressure on them
  - Location of telescopes/instruments…

- For space missions, like Euclid, other constraints like the ecliptic…

Observing: some limitations

- A survey conducted over a period of time (usually years) will inherit:
  - Variability in the observing conditions:
    - with time (hourly, nightly, weekly, monthly, yearly changes)
    - seeing, depth, and filter response sensitivity to observing conditions
    - manual interventions…
    - …
  - Variability in the instrument performance:
    - telescope orientation (flexion, alignments,…)
    - wavelength (e.g. red arm more sensitive than blue)
    - upgrades/maintenance:
      - fibre replacements
      - new CCDs
      - Mirror recoating
      - …
  - Variability in the input catalogues (for spectroscopic surveys):
    - imaging homogeneity
    - calibration
    - star/galaxy separation
    - …

Identification of key effects for the analysis

- Not all limitations can (and should not need to) be accounted for:
  - Essential to understand the analysis pipeline
  - Essential to understand the statistical uncertainties
  - Essential to understand the systematic uncertainties

Classic observable: flux/magnitude definitions

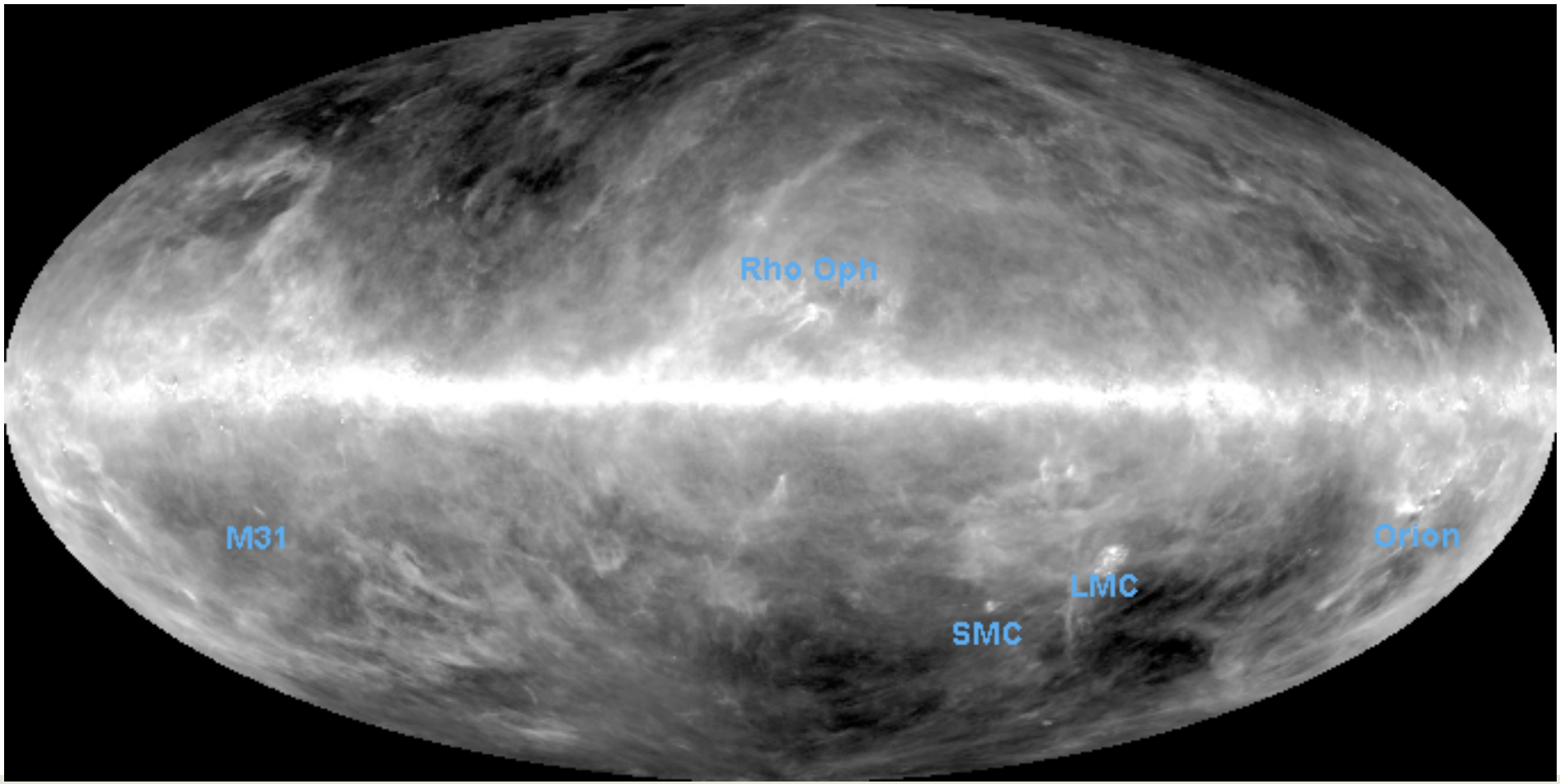
- Observationally there is a large number of way to measure the flux (or magnitude, i.e. $m = -2.5 \log_{10} f + c$) of an object:
  - Bolometric: flux accounting for all photons emitted by the object
  - Total: flux (in a band/filter) accounting for all the photons received from the object in a given band/filter

  But neither are realistically possible….

- In reality for extended objects (like galaxies) one resorts to:
  - Aperture: flux contained within a fixed circular aperture
  - Fibre: flux contained within the aperture of a spectroscopic fibre
  - PSF: flux contained within a Gaussian PSF model
  - Petrosian: flux within a circular aperture whose radius is defined by the shape of the azimuthally averaged light profile.
  - Model: flux within a fitted de Vaucouleurs or exponential light profile
  - Sersic: generalized model magnitude
  - ….
- All come with their pros and cons: science goal (or data availability or data reduction feasibility) will decide what is likely to be best.

Dust extinction: another observational effect

- The sky is not uniformly transparent: dust extinction
- Dust extinction is wavelength dependent
- Schlegel et al. (1998) dust map using IRAS and COBE/DIRBE data

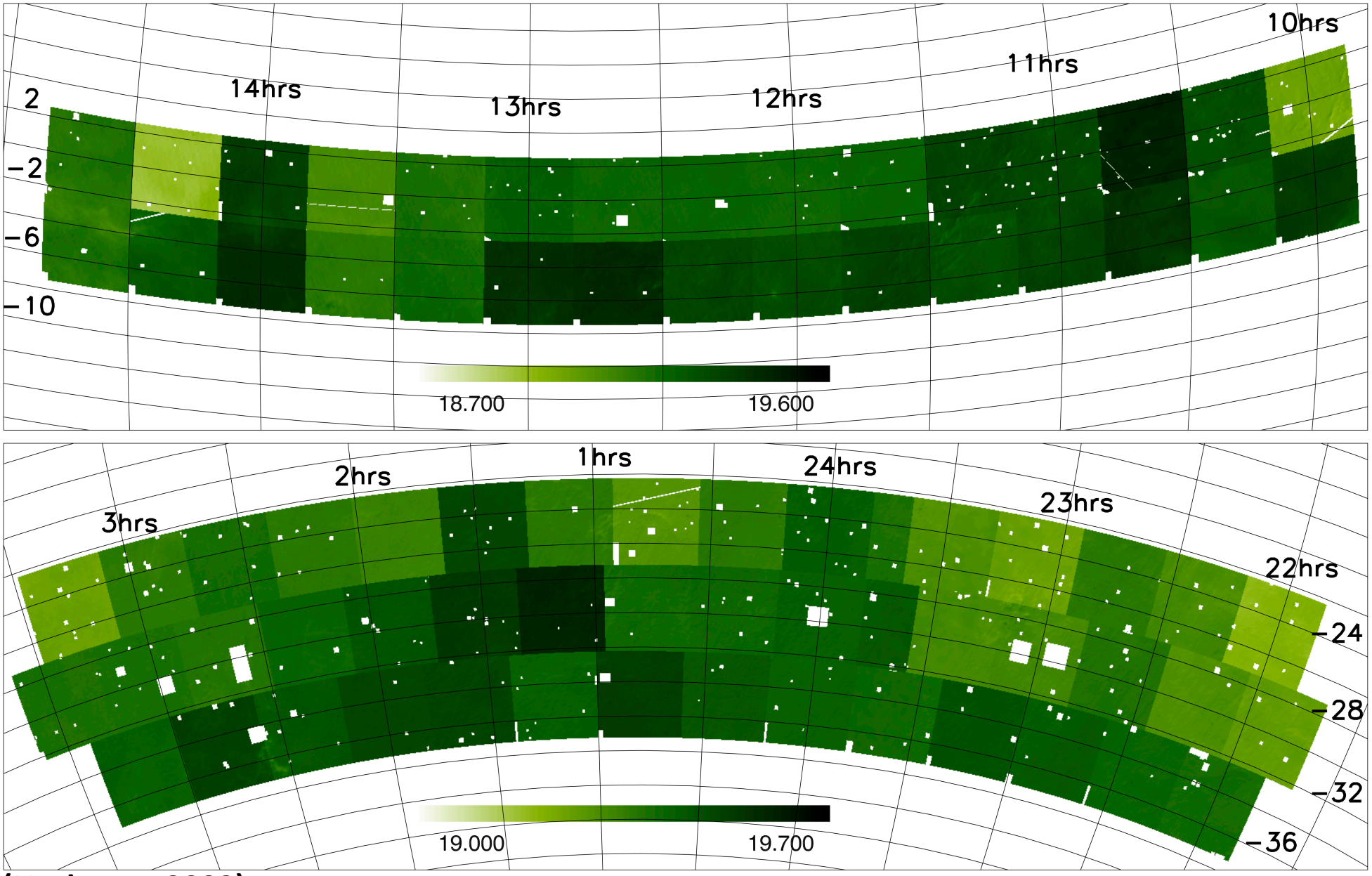


(Schlegel et al. 1998)

Survey selection function(s)

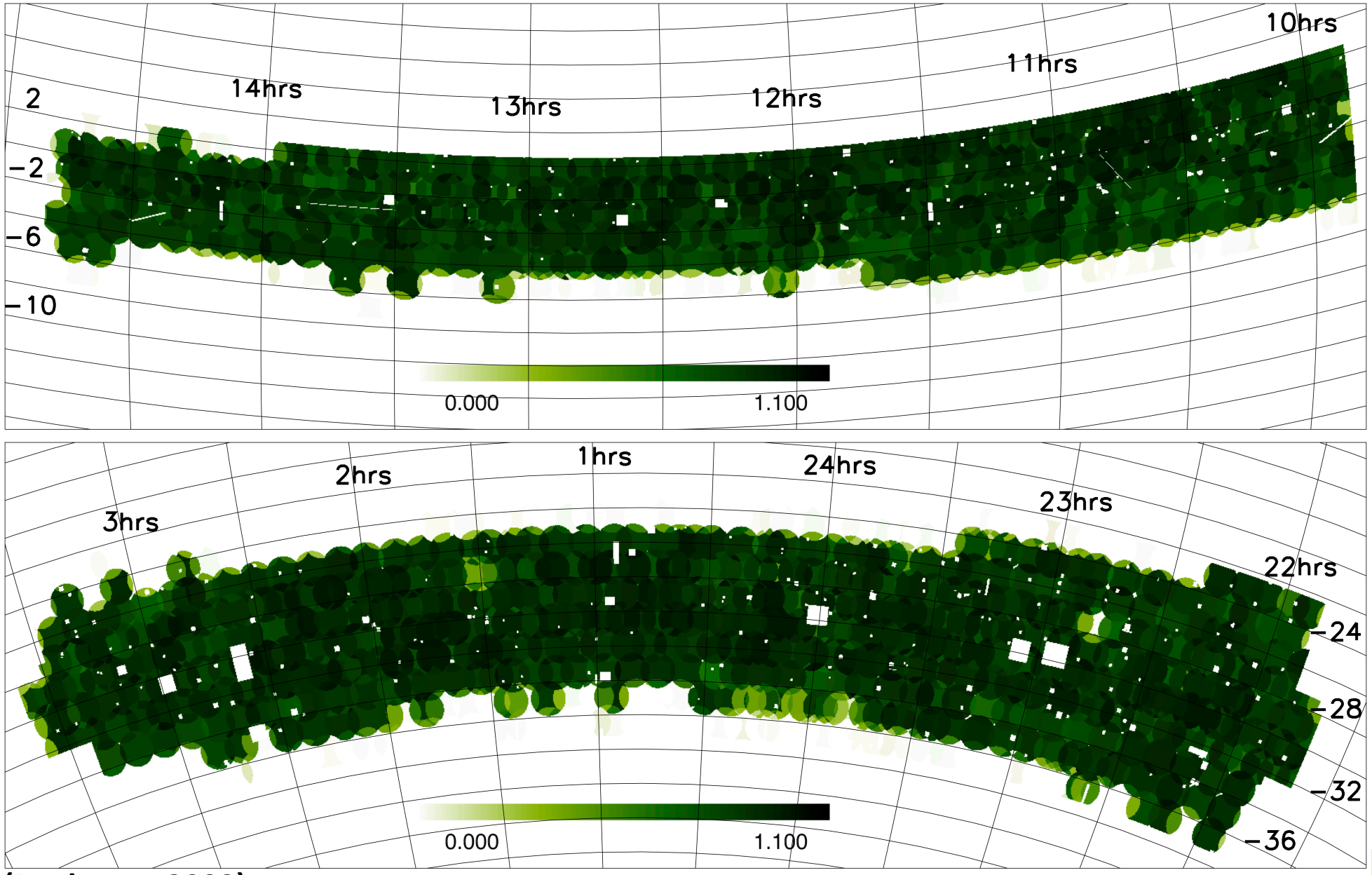
- The selection function is a survey characteristic which encapsulates the information of the probability that an object can be detected.
- The selection function, like the survey, is defined by:
  - the survey area
  - a set of magnitude/flux choices
  - a set of magnitude/flux limits
  - various maps (which may or not be needed / combined):
    - imaging completeness maps, incl. “bright” object mask
    - imaging depth maps
    - spectroscopic completeness maps:
      - angular map and radial distribution separately (most common situation)
      - 3D completeness distribution in angular and radial bins (e.g. Euclid for sure, but possibly also true for DESI)
    - …

2dFGRS maps: imaging depth



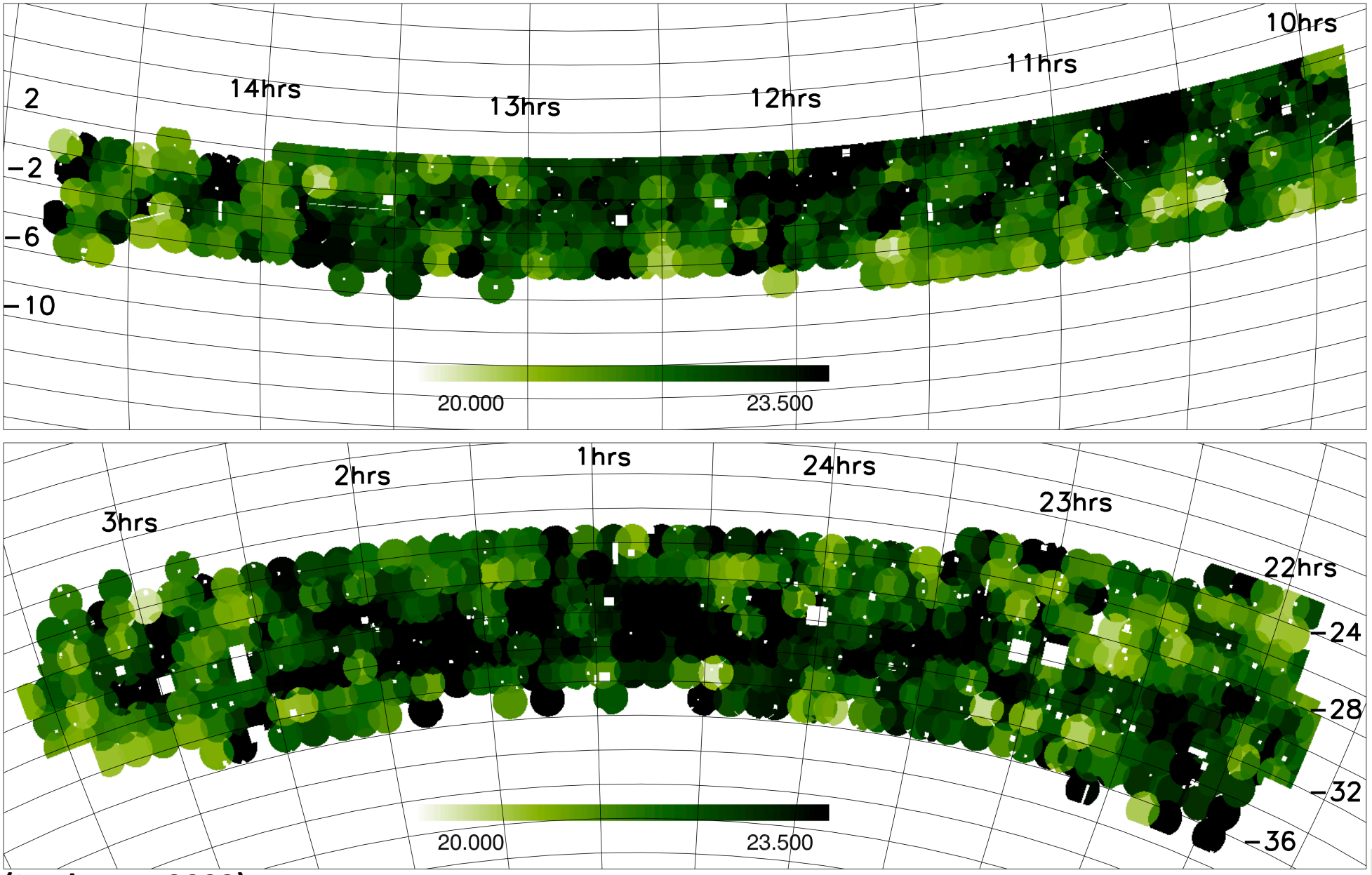
(Norberg – 2002)

2dFGRS maps: spectroscopic completeness



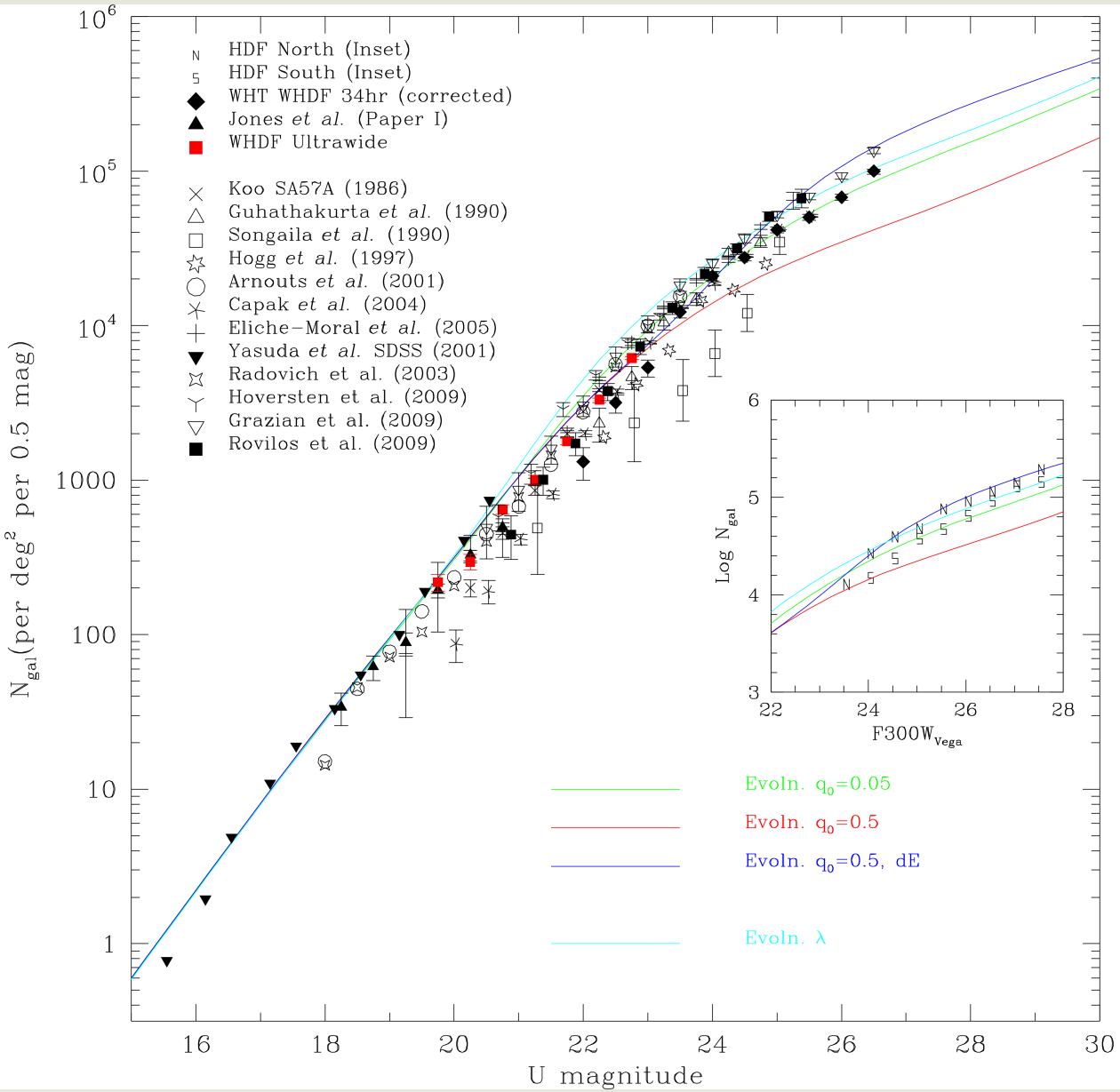
(Norberg – 2002)

2dFGRS maps: magnitude dependent spectroscopic completeness



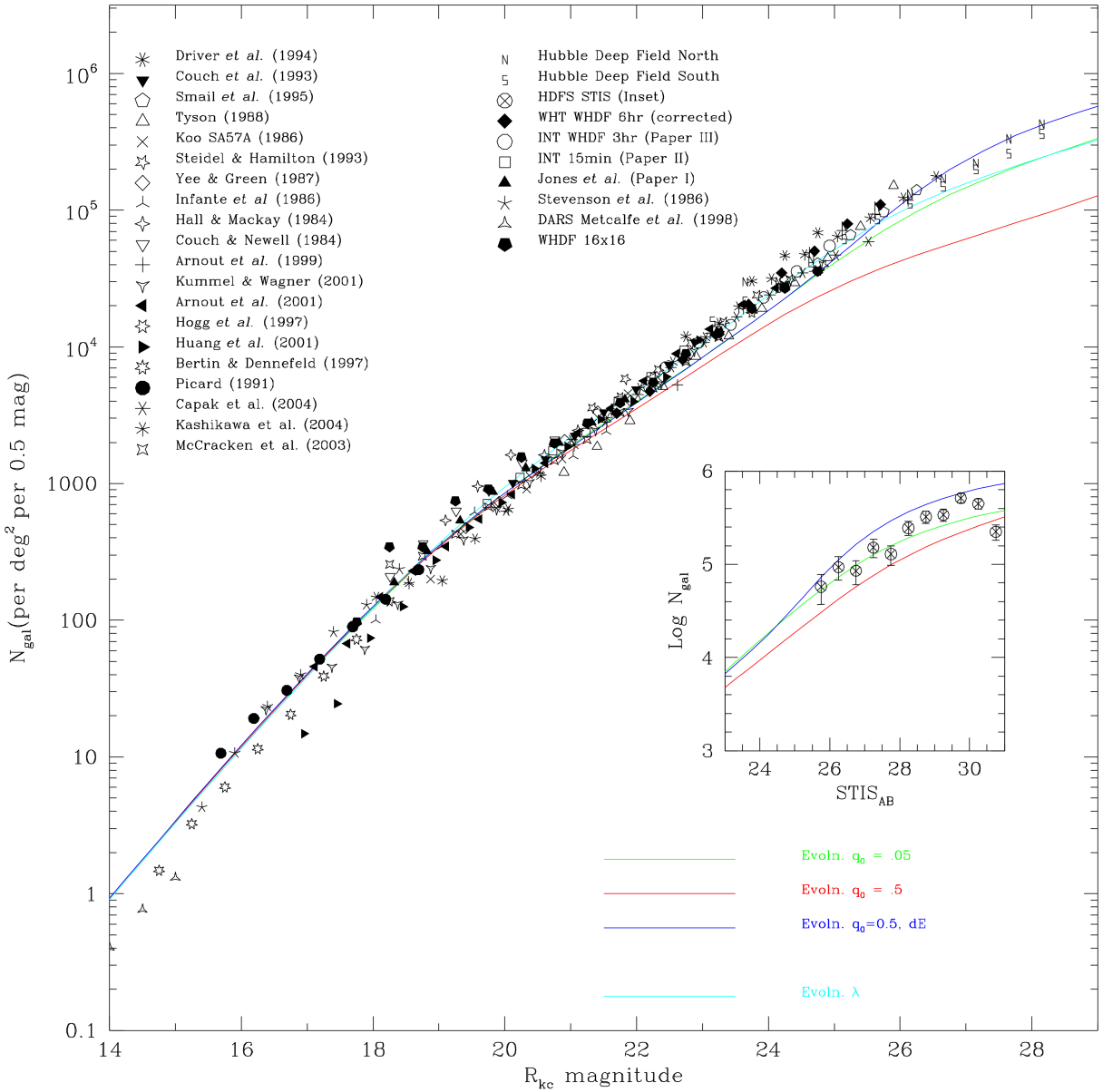
(Norberg – 2002)

Flux limited surveys: u-band number counts



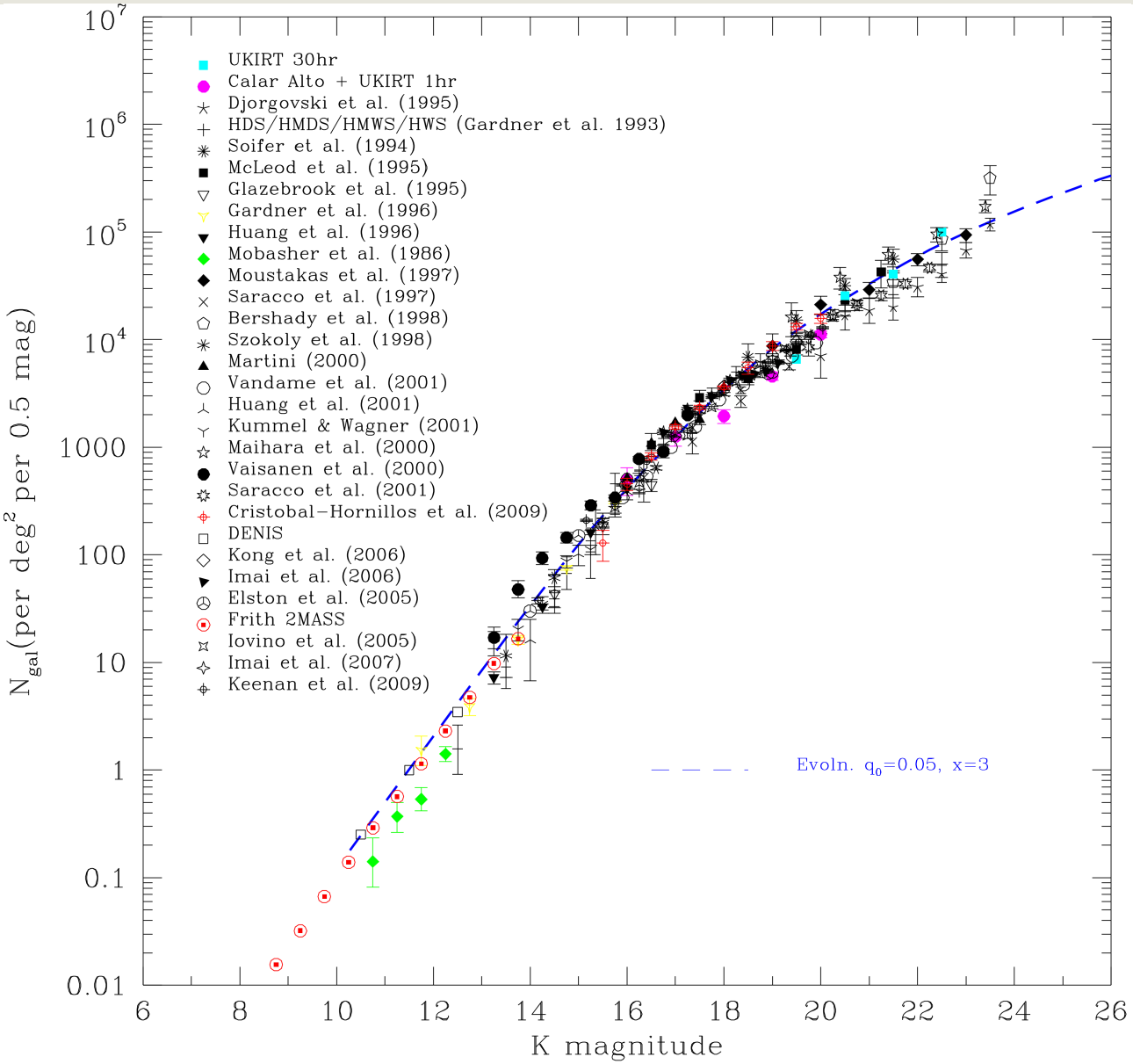
(Metcalfe – 2011)

Flux limited surveys: r-band number counts



(Metcalfe – 2011)

Flux limited surveys: k-band number counts



(Metcalfe – 2011)

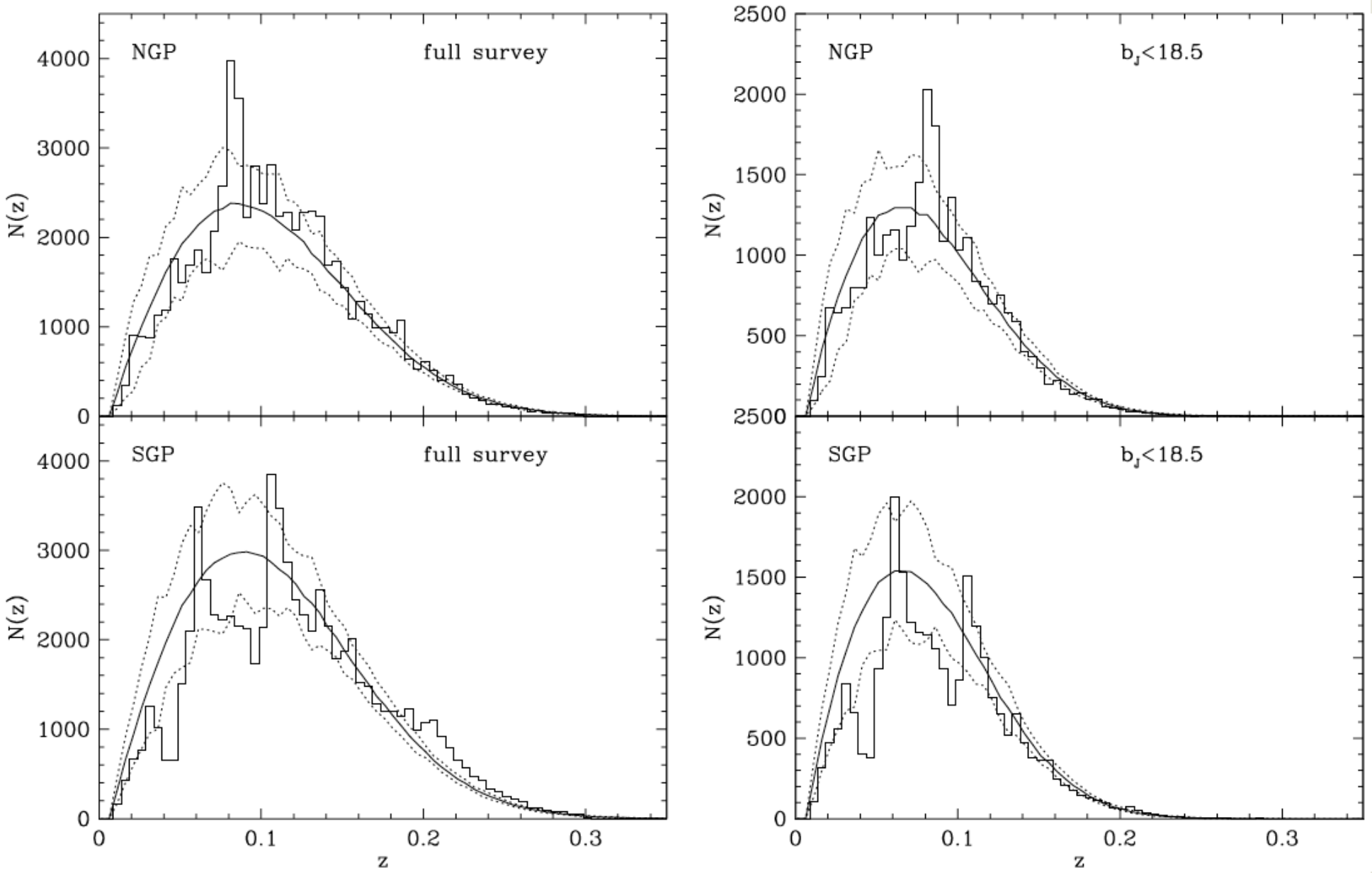
Star galaxy separation

- Star galaxy separation: easy in principle, but hard in reality as it all depends on the information at hands and the purpose of it.

- Useful information:
  - Multi-band photometry: stars and galaxies do not always reside in the same colour-colour or colour-magnitude space
  - Sample of high quality data (e.g. HST) with which to train against
  - High quality imaging:
    - Enables morphological discriminators

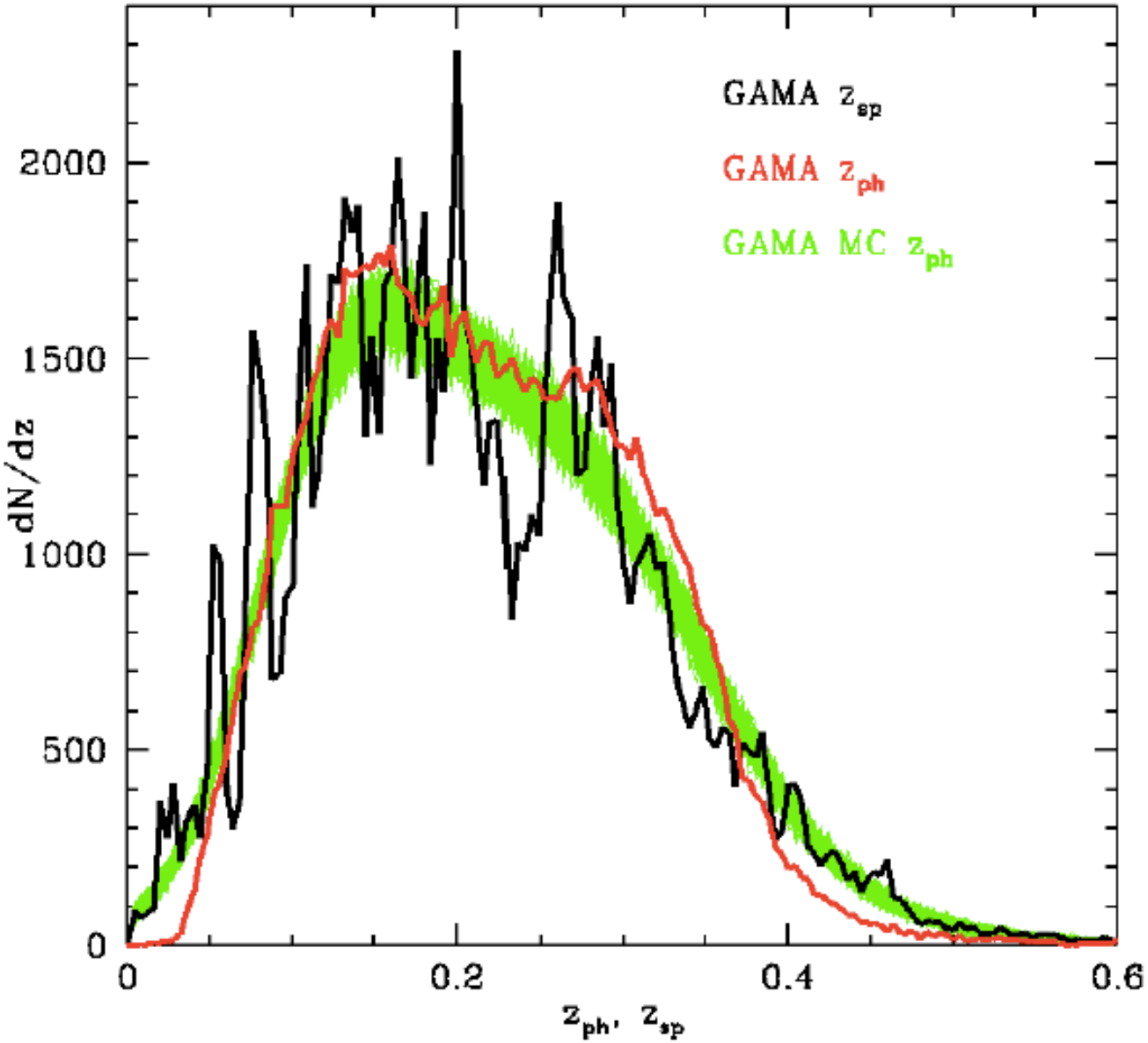
- Key points:
  - Define what fraction of interlopers is acceptable (tend to vary a lot from science goal / survey design)
  - Understand the quality of the S/G separation as function of magnitude, but also position within the survey (stellar density not uniform…)

Flux limited surveys: redshift distribution



(Norberg et al. 2002)

Flux limited surveys: spec-z & photo-z distribution



(Driver et al. 2010)

Modeling the radial selection function: two options

- Empirical modeling:
  - Fit a smooth curve to the observed dN/dz
  - Convolve the observed dN/dz with a smooth filter
- Intrinsic modeling:
  - Estimate the galaxy LF and infer a smooth dN/dz from it:
  - Estimate the joint density LF distribution and marginalize over the LF distribution (e.g. Cole 2011)
- Pros and Cons:
  - Empirical modeling cannot account for fluctuations on scales larger than the survey
  - Intrinsic modeling might require the modeling of the evolution of the galaxy population (likely uncertain to some extent).
  - Empirical modeling is very fast (and easy to implement), but requires to be redone/retuned for each sub-sample unlike intrinsic modeling.
  - Intrinsic modeling requires all assumptions to be spelt out, leading to a better understanding of the survey limitations.
  - Ideally both empirical and intrinsic approaches should be adopted!

From cosmological simulations to real observations: the ins and outs of the mock world… Part 1b: the real universe

Peder Norberg (Durham University)

Mexican Numerical Simulations School
October 3rd - 6th 2016

Intrinsic survey characteristics

- For some analyses the intrinsic characteristics are more important to understand/reproduce, but they require all a proper modeling of the selection function 1st.
- 1-point statistics:
  - Galaxy luminosity function:
    - Number of objects per unit volume per unit luminosity (in a given band)
  - Stellar mass function:
    - Number of galaxies per unit volume per unit stellar mass…
  - …
- 2-point statistics:
  - Galaxy clustering:
    - 2-point correlation function statistically describes the probability, given one object, that there is another object at a separation **x** from the former.
    - …
  - …

Luminosity function estimators

- Vmax (Schmidt 1968; Felten 1976):

$$\phi(\mathrm{L})\mathrm{dL} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \omega_{\mathrm{i}} \frac{\mathrm{W}(\mathrm{L}-\mathrm{L}_{\mathrm{i}})}{\mathrm{V}_{\mathrm{max}}(\mathrm{L}_{\mathrm{i}})}$$

$$\mathrm{W}(\mathrm{L}-\mathrm{L}_{\mathrm{i}}) = \Theta(\mathrm{L}_{\mathrm{i}}-\mathrm{L}+\mathrm{dL}/2) - \Theta(\mathrm{L}+\mathrm{dL}/2-\mathrm{L}_{\mathrm{i}})$$

- Standard maximum likelihood estimators:
  - STY (Sandage et al. 1979) – assumes a functional form
  - SMWL (Efstathiou et al. 1988) – assumes a "binned" LF

    The probability of a galaxy of having a luminosity L in a volume element dx centred on x is:

$$P(L,\mathbf{x})\,dL\,d^3\mathbf{x} = \phi(L)\rho(\mathbf{x})\,dL\,d^3\mathbf{x}$$

    The conditional probability of a galaxy α at redshift z_α will have luminosity L_α is:

$$p_\alpha = \frac{\phi(L_\alpha)}{\int_{L^{\mathrm{min}}(z_\alpha)}^{\infty} \phi(L)\,dL}$$

$$\mathcal{L} = \Pi_\alpha p_\alpha$$

Luminosity function estimators

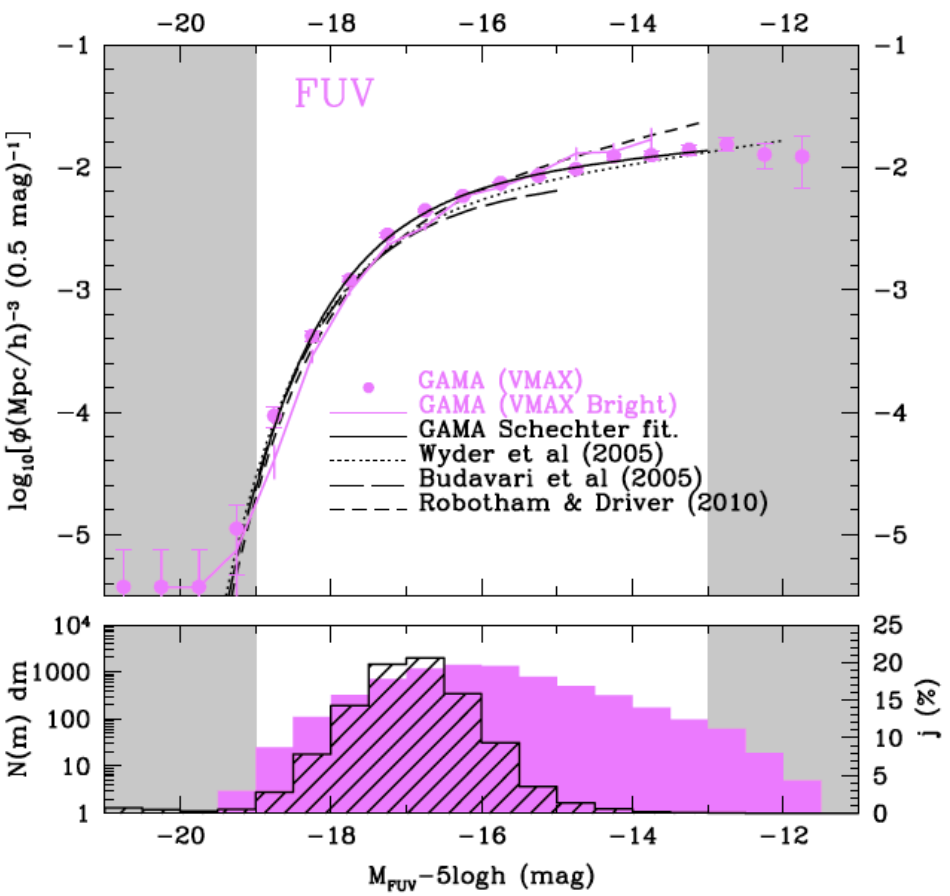
- Maximum likelihood estimators with joint density and luminosity estimates (Cole 2011; Loveday et al. 2015) gives for galaxy α:

$$p_\alpha = \frac{\Delta(z_\alpha)\frac{dV(z_\alpha)}{dz}\phi(L_\alpha)}{\int \Delta(z)\frac{dV}{dz}\int_{L^{\min}(z)}^{\infty}\phi(L)\,dL\,dz} \qquad \mathcal{L} = \Pi_\alpha p_\alpha$$

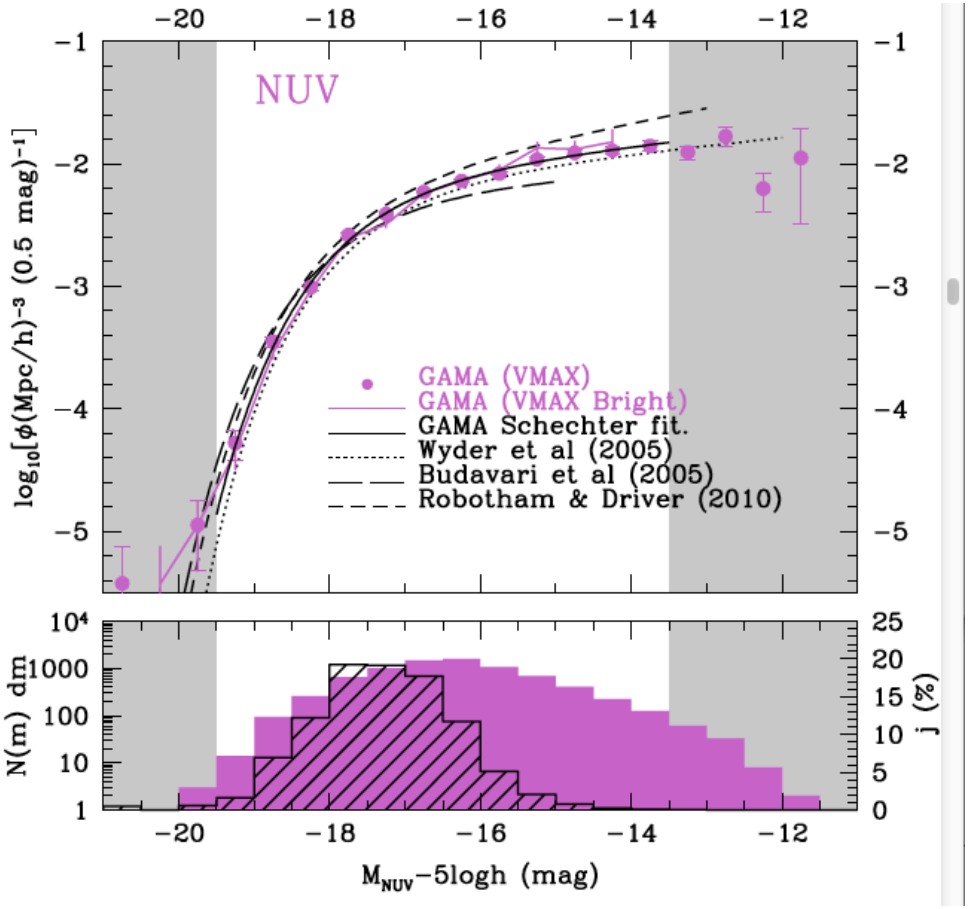
where $\Delta(z_\alpha)$ is the galaxy overdensity at z_α (assuming no evolution of the LF).

- Key ingredients for all estimators:
  - Cosmology
  - k-corrections: SED dependent correction to account for band shifting with redshift.
  - e-corrections: evolutionary corrections to account for the fact that galaxies evolve over the redshift range probed. Can be critical to include in some cases.

Galaxy Luminosity functions: FUV and NUV



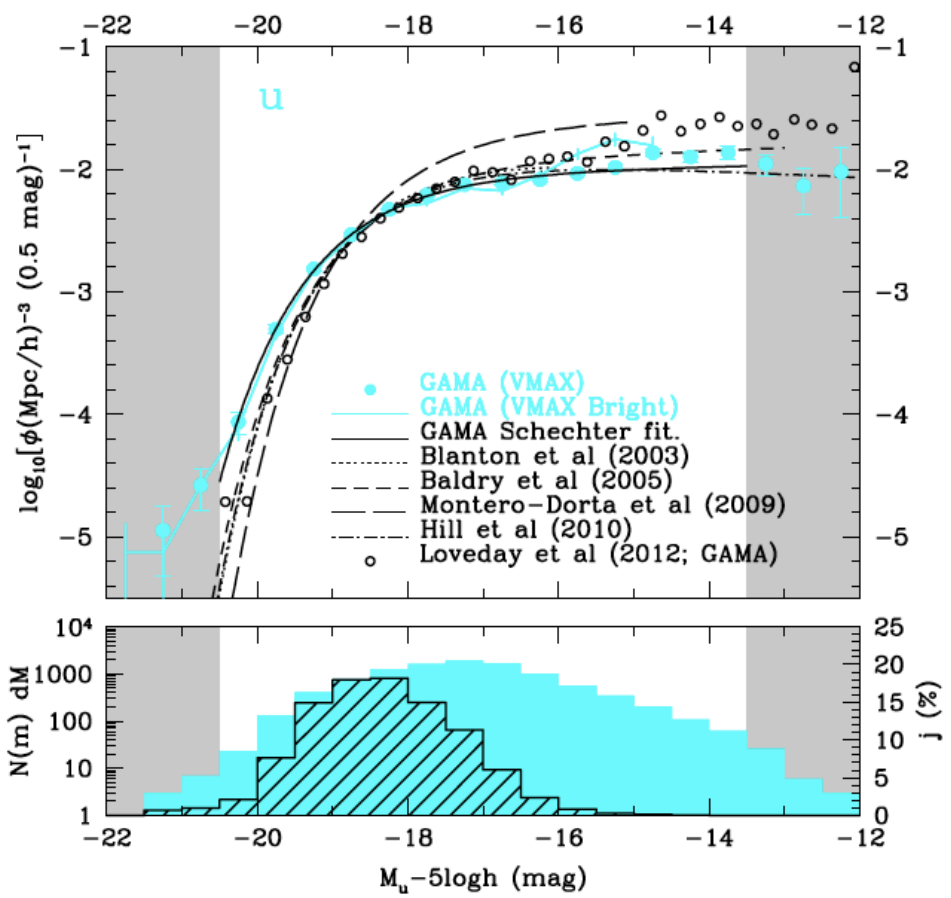
Bright Faint

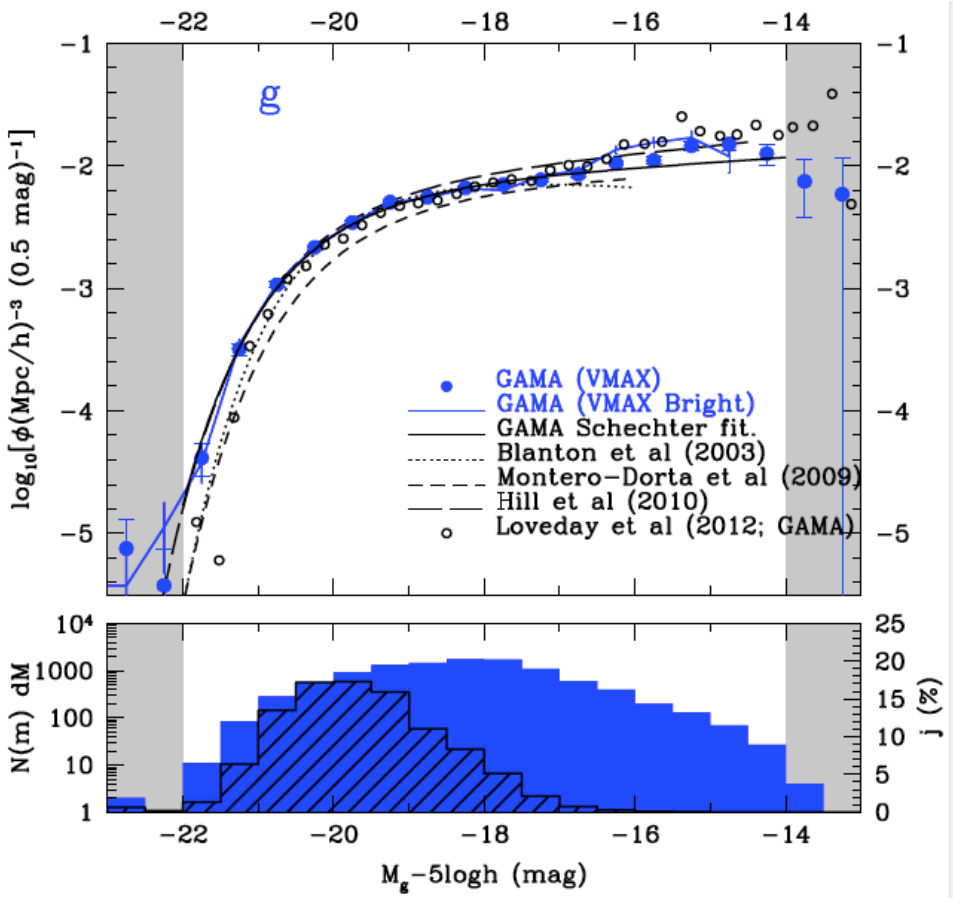


Bright Faint

(Driver et al. 2012)

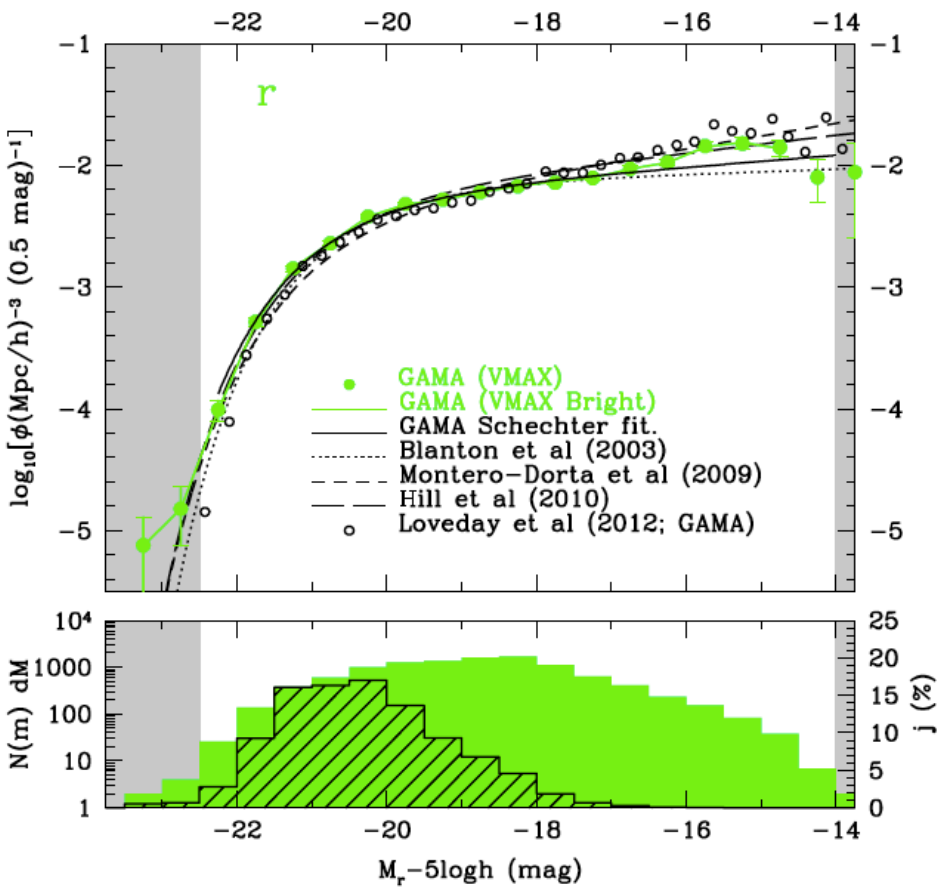
Galaxy Luminosity functions: optical blue (u & g)

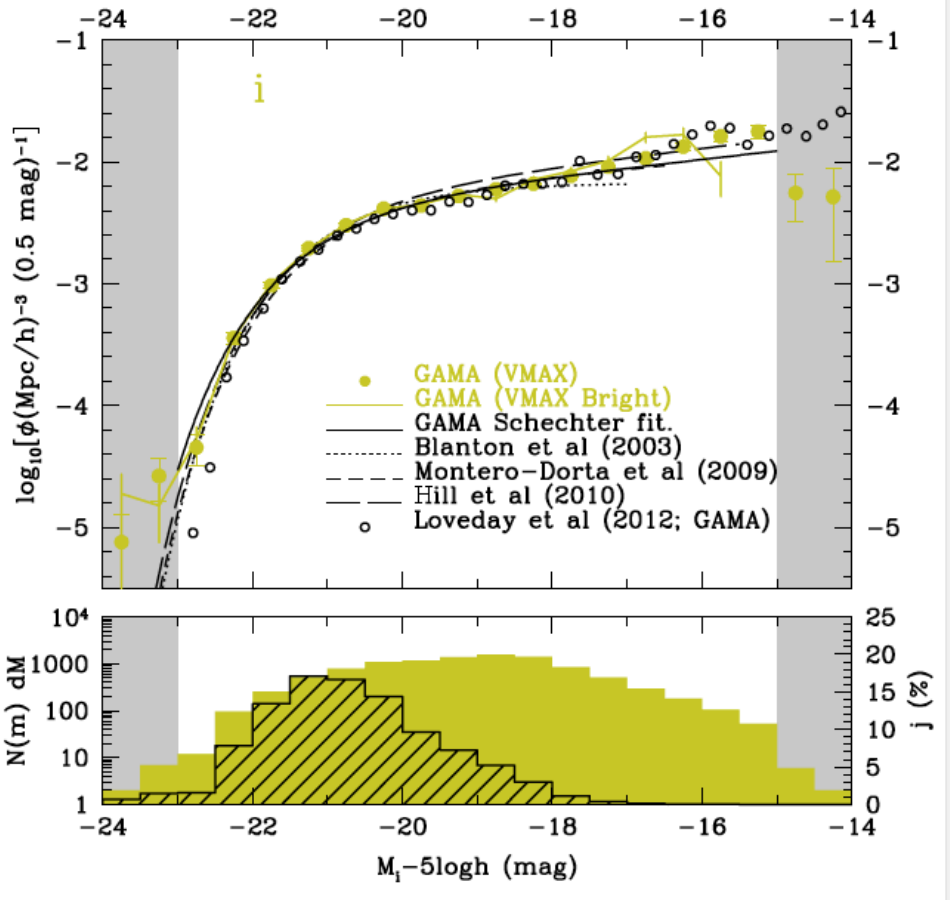




(Driver et al. 2012)

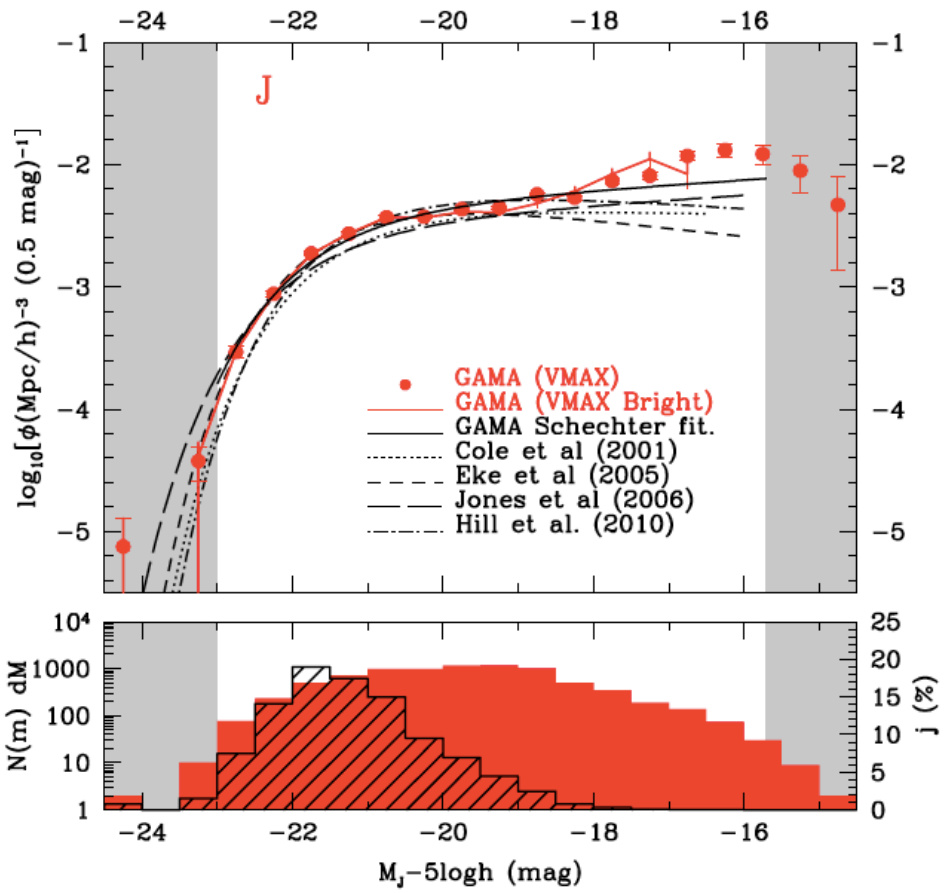
Galaxy Luminosity functions: optical red (r & i)

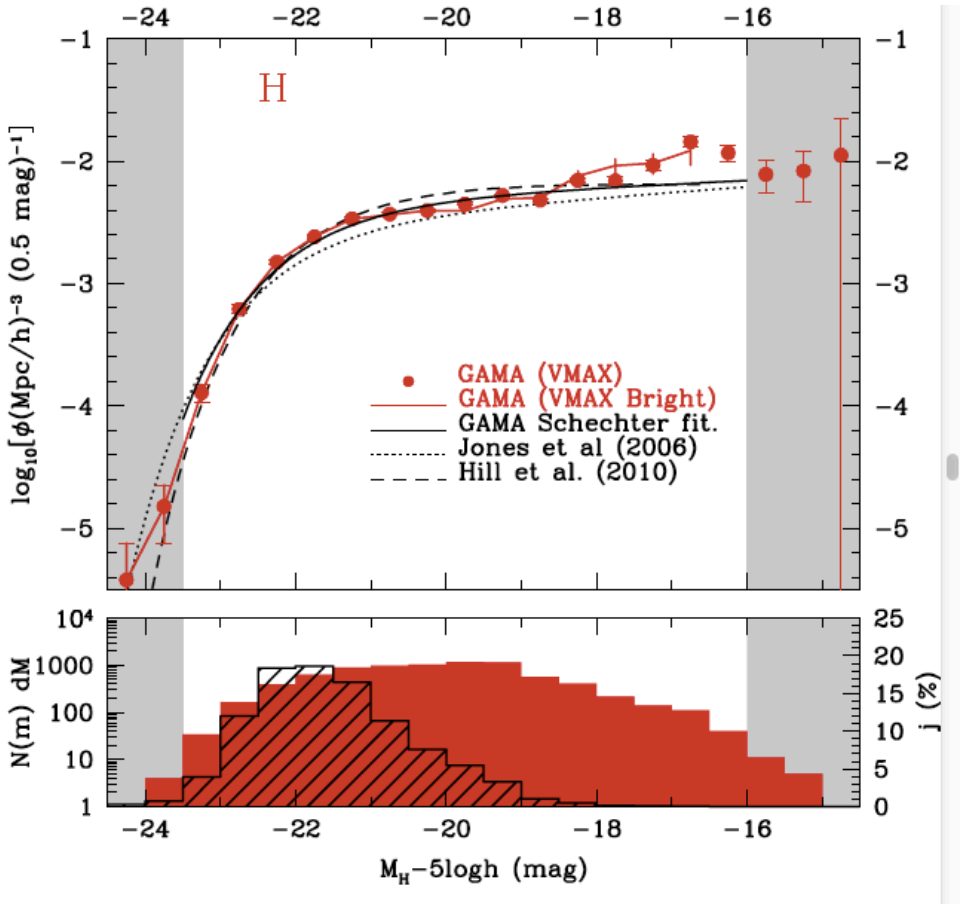




(Driver et al. 2012)

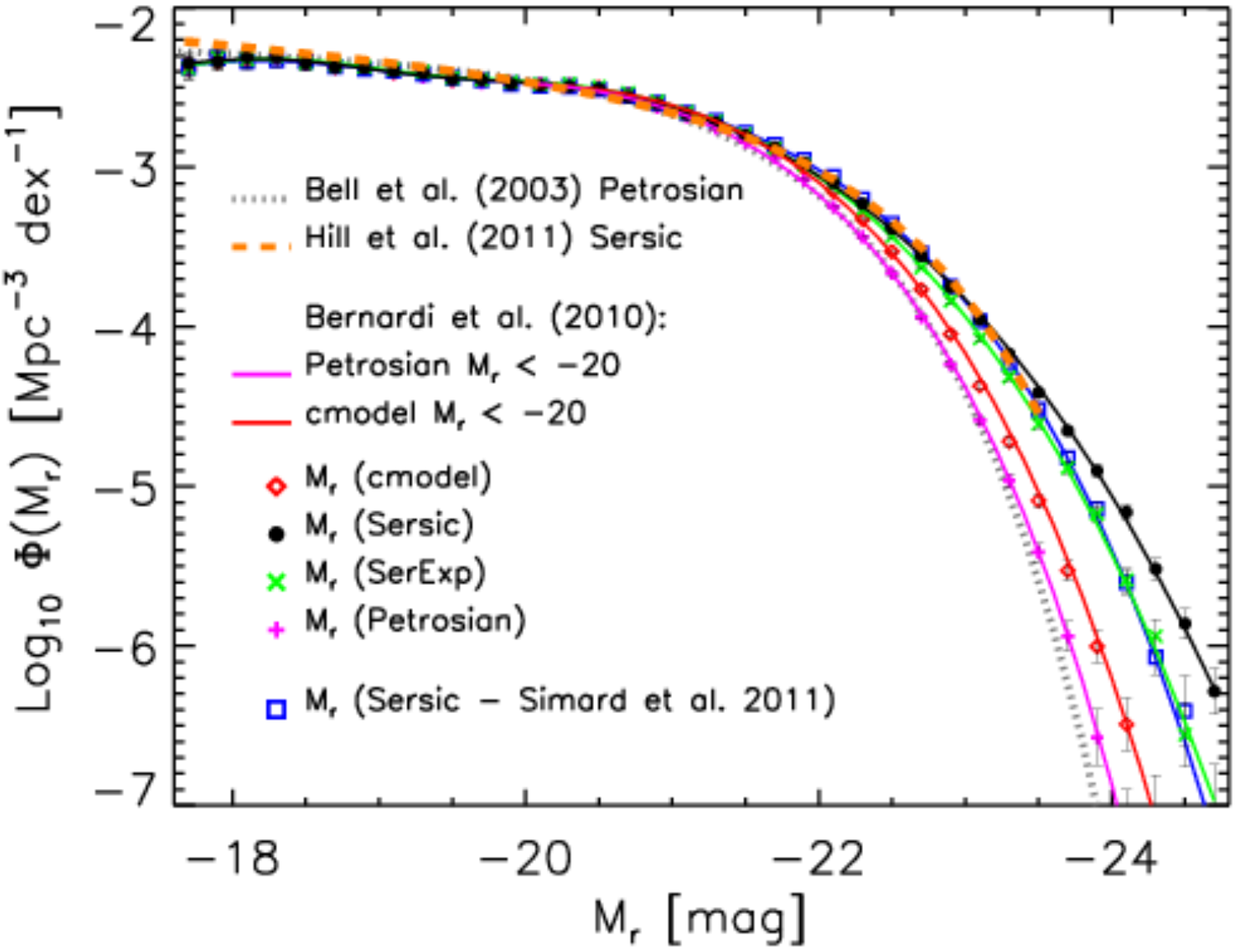
Galaxy Luminosity functions: near-IR (J & H)





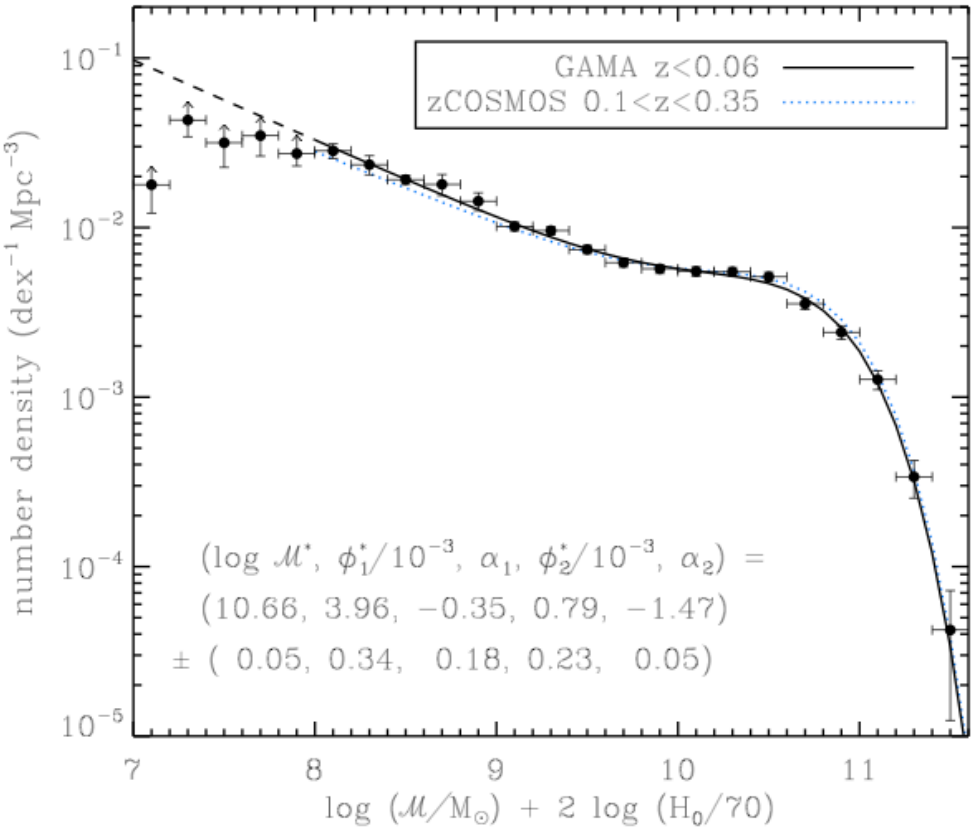
(Driver et al. 2012)

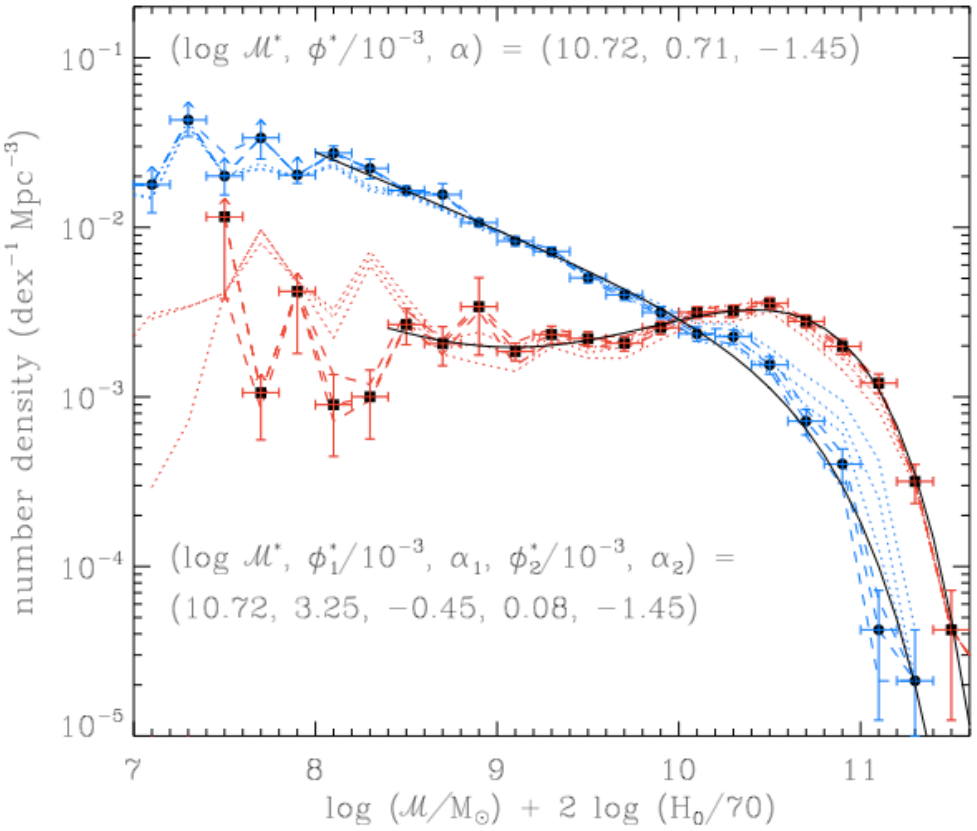
Luminosity functions: magnitude definitions matter



(Bernardi et al. 2014)

Stellar mass function: low-z with GAMA





(Baldry et al. 2012)

Stellar mass function: observed or inferred?

- For a theoretical viewpoint, the stellar mass function is the most fundamental characteristic of a galaxy sample. In simulations it is "easy" to derive, just need to add up all the stars in a galaxy.

- For an observational standpoint, the stellar mass function is not an observable (one collects photons, not stellar mass). Assumptions include:
  - initial mass function (e.g. Kennicut, Chabrier, …)
  - stellar population synthesis model (e.g. Bruzual & Charlot, Maraston, …)
  - stellar mass-to-light ratio (in some band, defined in some way)
  - dust modeling and metallicity…
  - …

- Could this possibly lead to misunderstandings on its accuracy? See e.g. Conroy et al. (2009) and Mitchell et al. (2013)

Clustering statistics: 2-point correlation function

- Following Peebles (1980), the 2-point correlation function, $\xi(\mathbf{x})$, is given by the excess probability of finding a galaxy within a volume dV at position **x** from another galaxy:

$$dP = n_G^2 \, (1+\xi(\mathbf{x})) \, dV$$

where n_G is the average galaxy number density.

- Cosmological Principle implies that ξ only depends separation (and not actual positions in 3D space)

- Common correlation function estimators (for galaxy surveys):

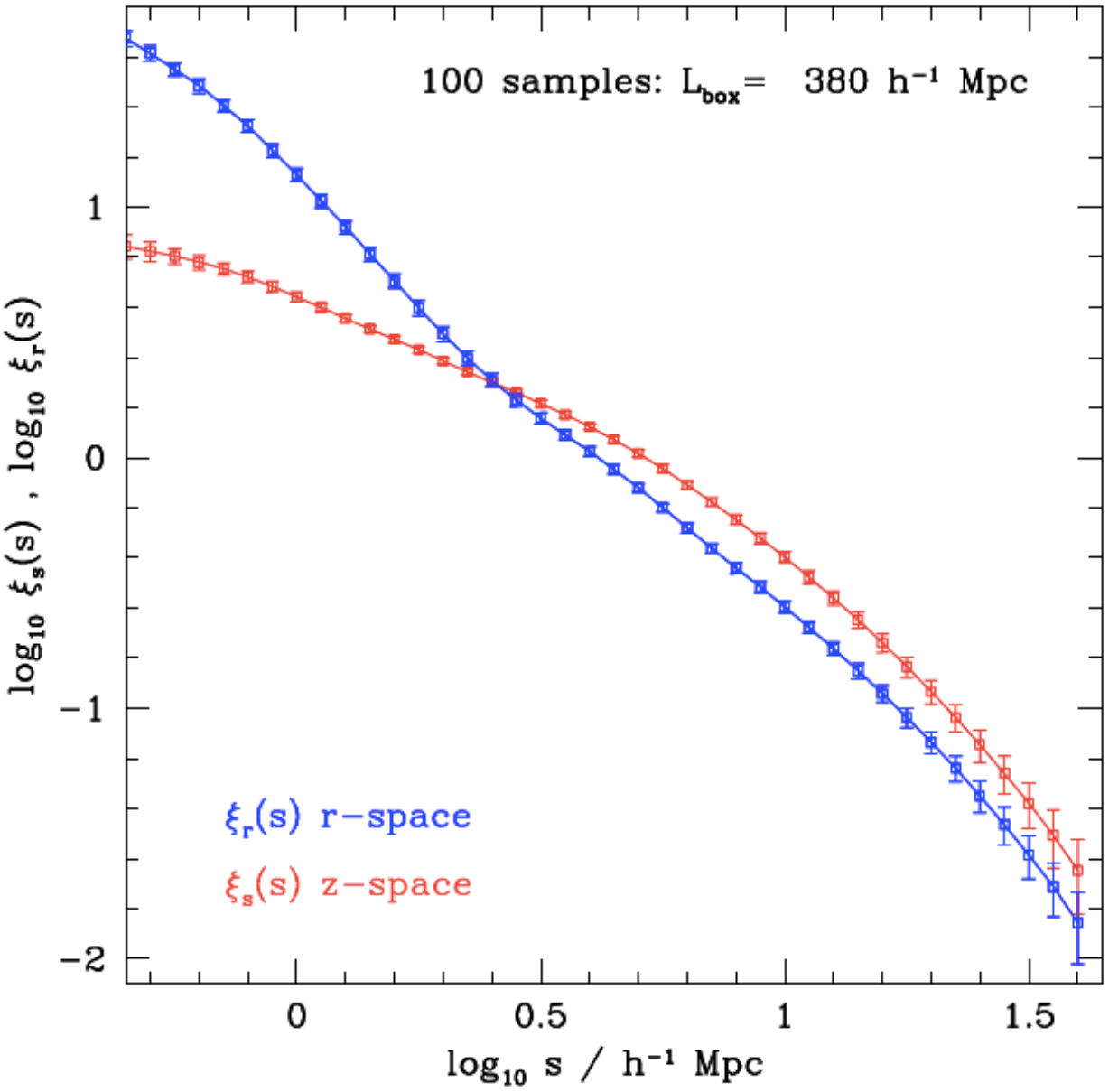
Hamilton (1993):

$$1 + \xi_{\rm H} = \frac{DD\,RR}{DR^2}$$

Landy & Szalay (1993):

$$\xi_{\rm LS} = \frac{DD - 2DR + RR}{RR}$$

Dark Matter correlation function: ξ(r) & ξ(s)

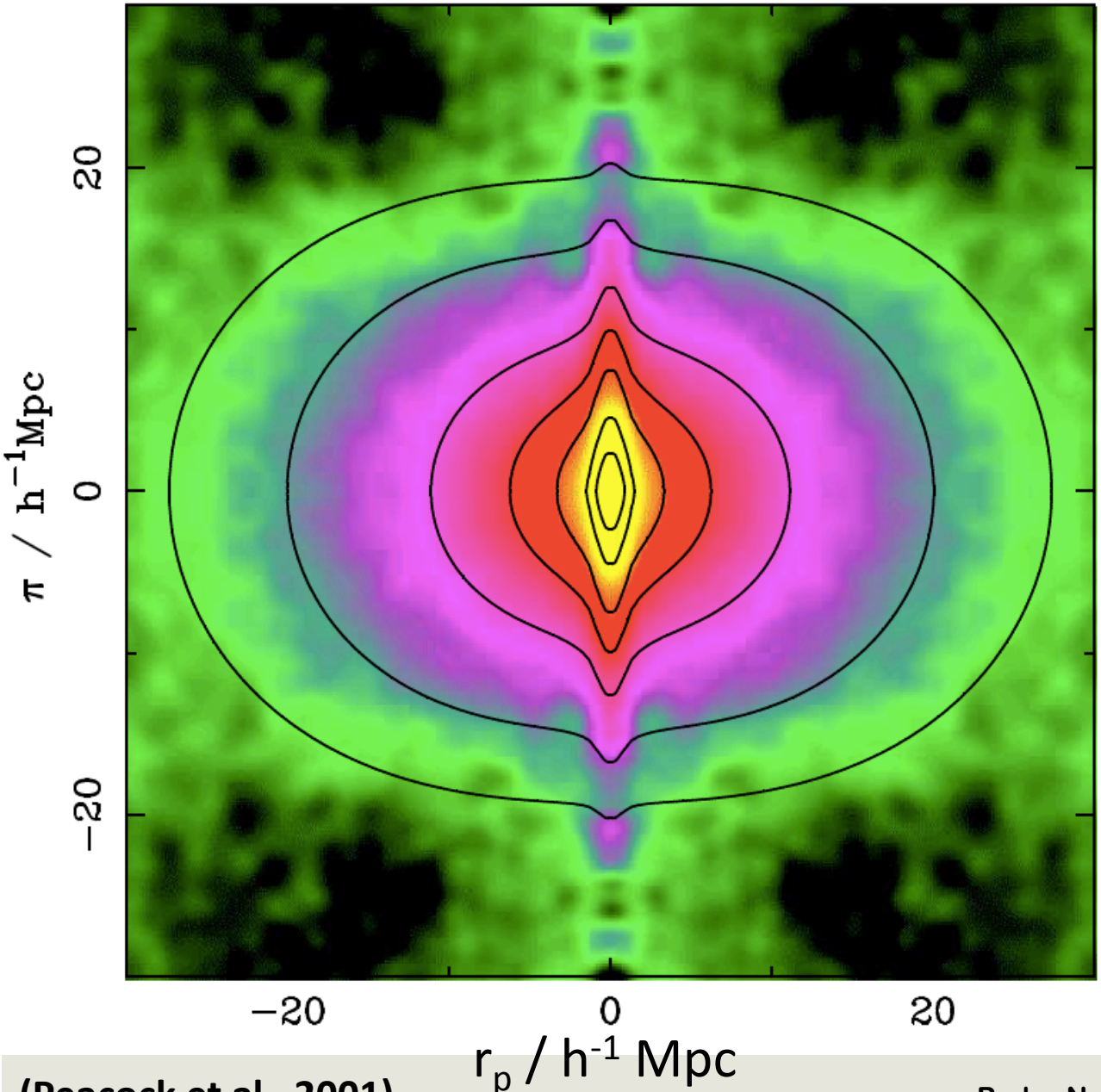


To go from real space to redshift space, one needs to account for the velocity along the line of sight.

Observationally it is:

$$z_{\rm tot} \quad = \quad z_{\rm hub} + \mathbf{v} \cdot \hat{\mathbf{r}}/c$$

2d Correlation function: $\xi(r_p,\pi)$ & RSD



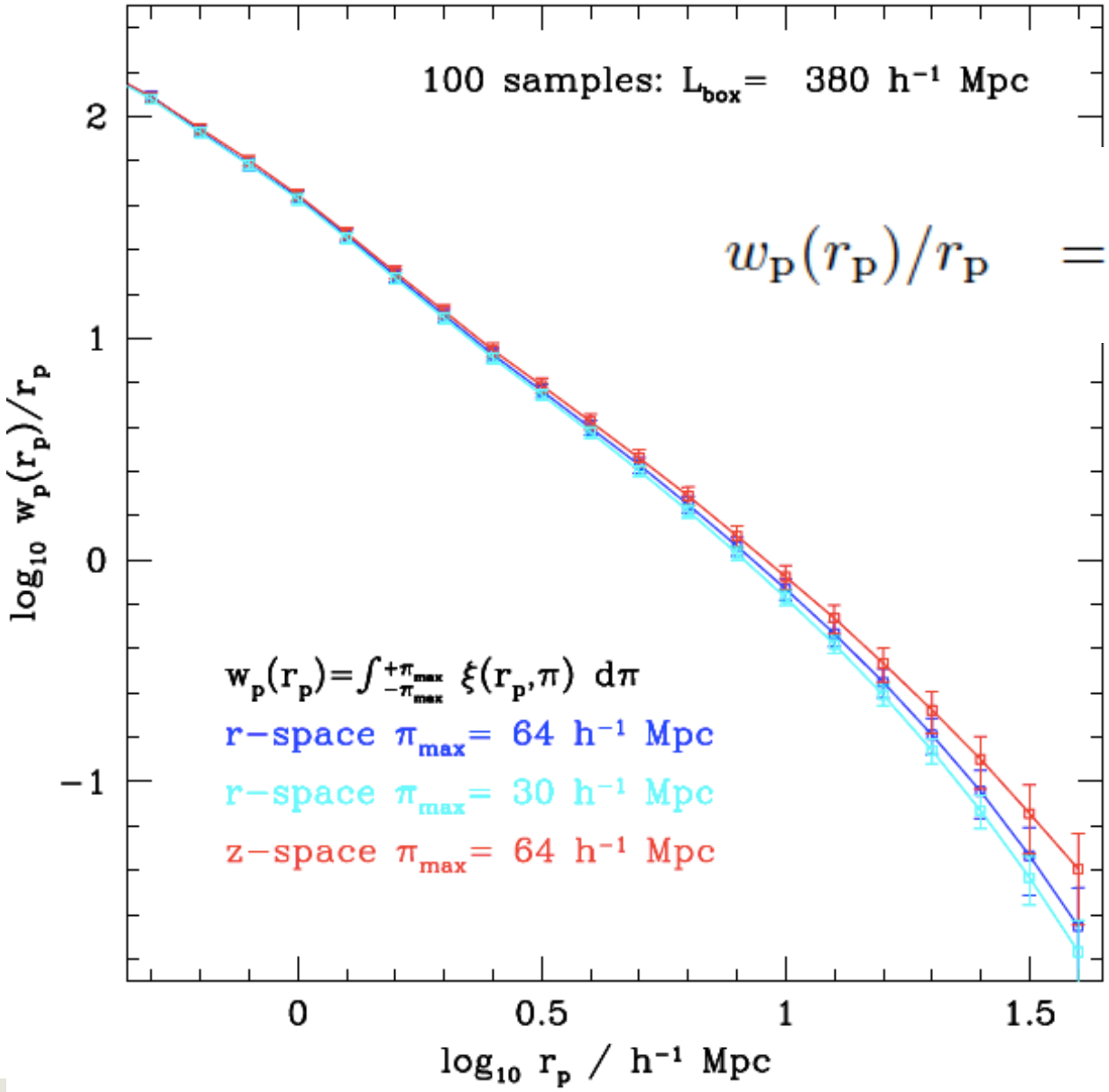
2dFGRS $\xi(r_p,\pi)$ with log colour scale.

Model: Kaiser + FoG

r_p: projected separation
π: line-of-sight separation

(Peacock et al. 2001)

Dark Matter correlation function: $w_p(r_p)$



$$w_{\rm p}(r_{\rm p})/r_{\rm p} = \frac{2}{r_{\rm p}} \int_0^{\pi_{\rm max}} \xi_{\rm X}(r_{\rm p}, \pi)\, d\pi$$

with:

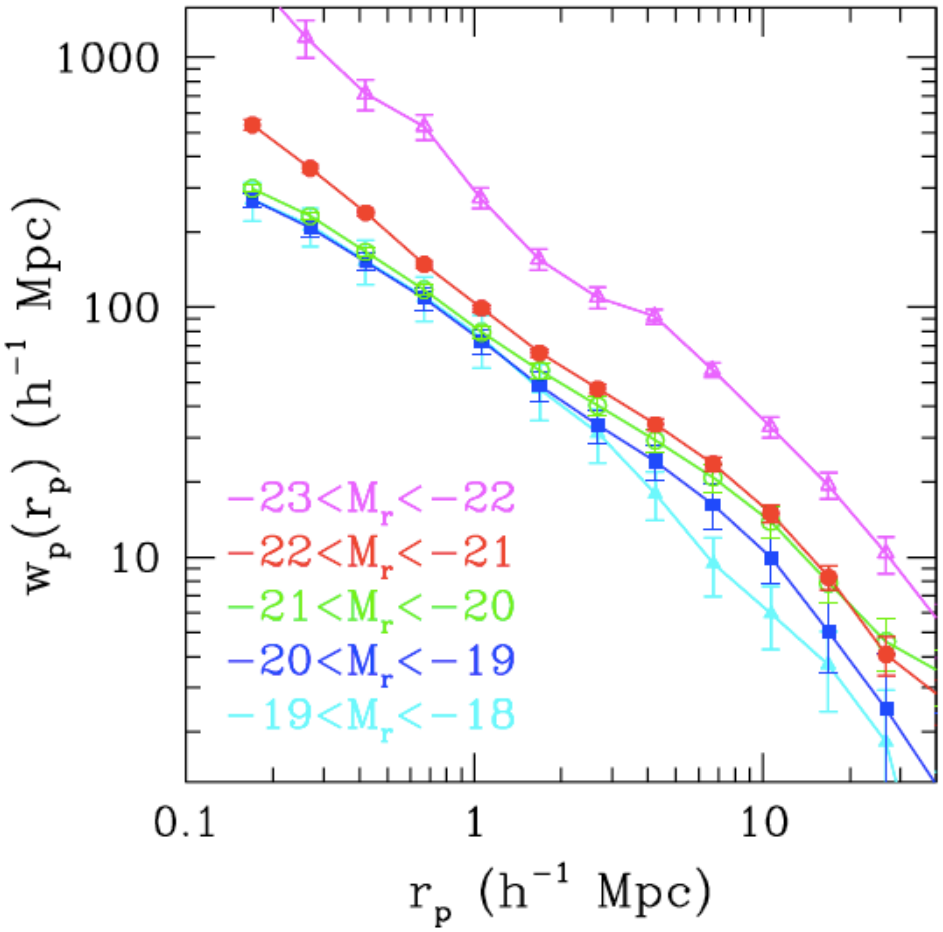
r_p: projected separation

π: line-of-sight separation

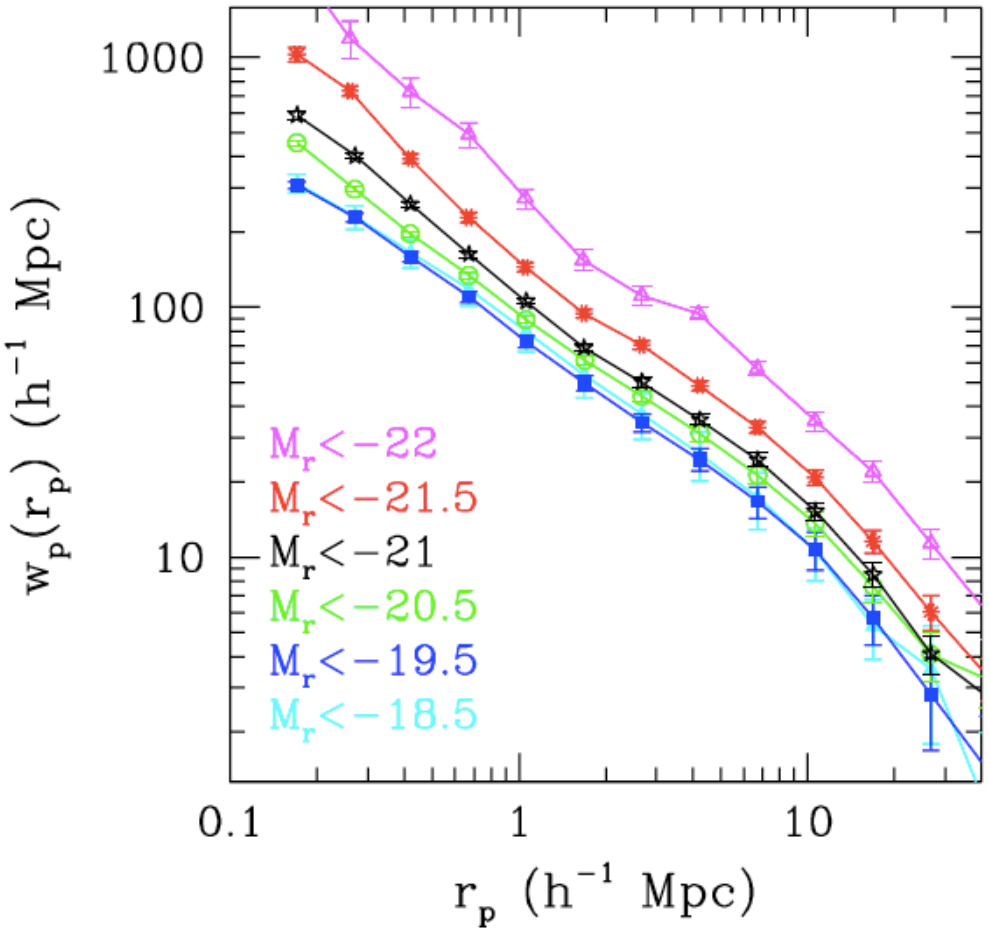
The projected correlation function, $w_p(r_p)$, is a statistic that provides *in principle* a RSD independent result…

Galaxy clustering: state-of-the-art with SDSS

Volume limited samples



Threshold samples



(Zehavi et al. 2011)

Error on clustering statistics: internal (i.e. data driven)

- Split survey into N sufficiently large sub-volumes
  - Bootstrap: Bootstrap (with replacement) the sub-volumes
    - Issue: the number of replacements not defined, but often assumed to be the same as the number of sub-volumes

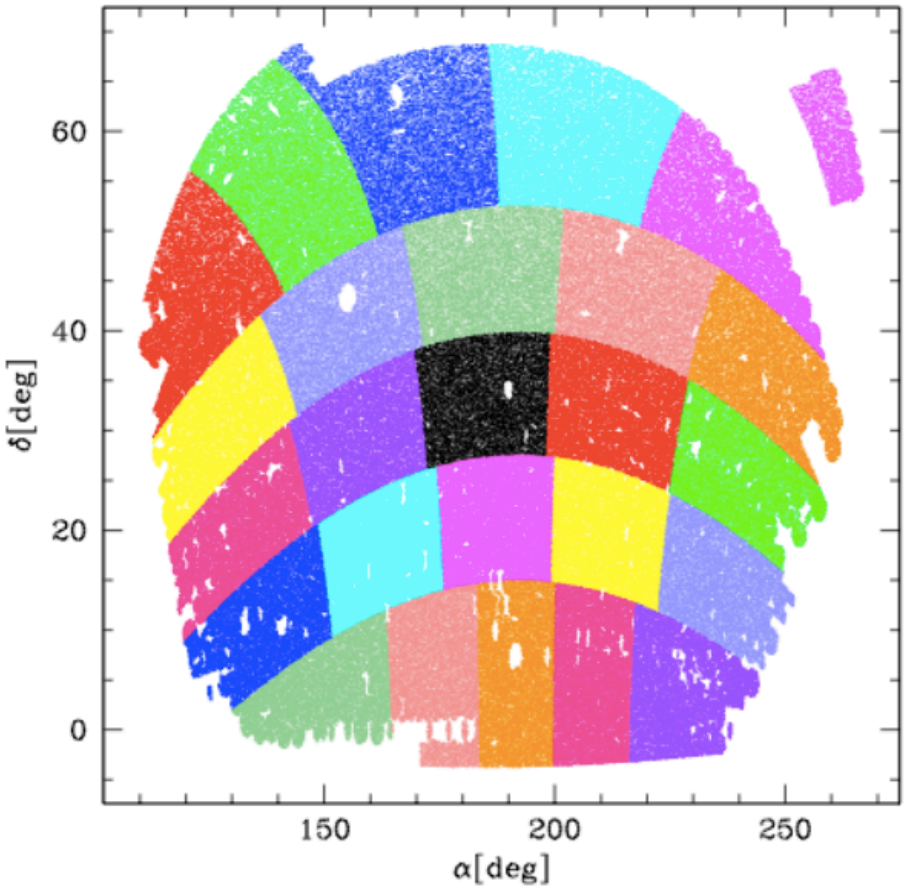
$$C_{\rm boot}(x_i, x_j) = \frac{1}{N-1}\sum_{k=1}^{N}(x_i^k - \bar{x_i})(x_j^k - \bar{x_j})$$

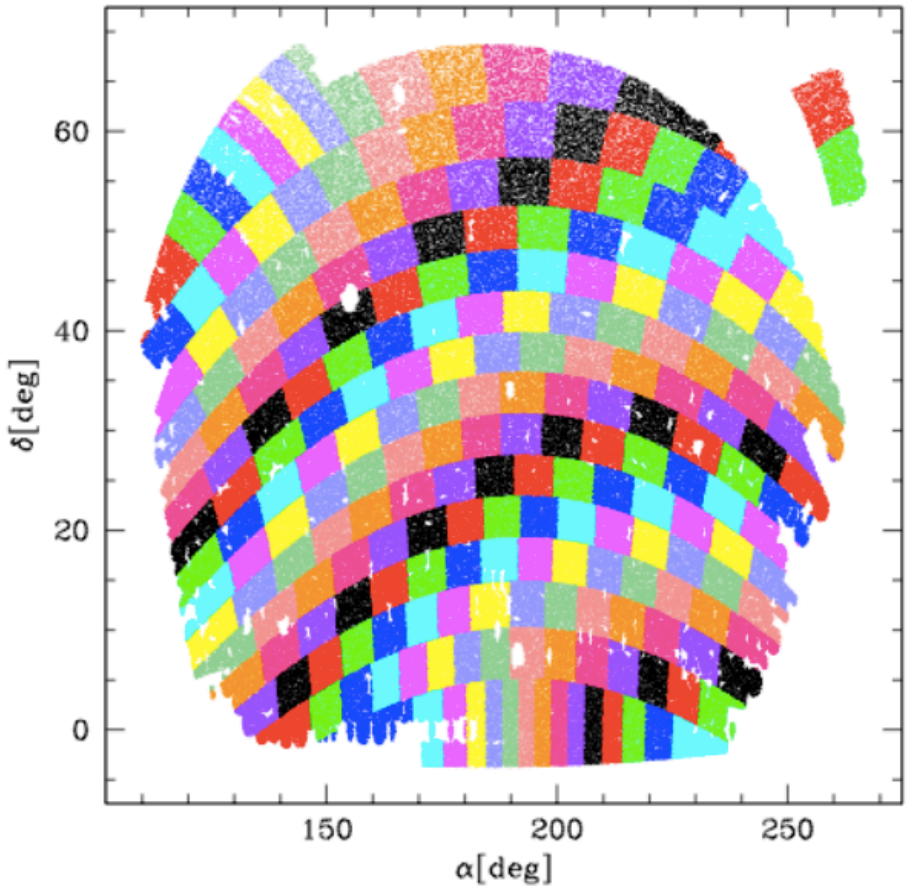
  - Jackknife: use N samples of N-1 sub-volumes, leaving each one out in turn
    - Issue: at most N realisations

$$C_{\rm jk}(x_i, x_j) = \frac{(N-1)}{N}\sum_{k=1}^{N}(x_i^k - \bar{x_i})(x_j^k - \bar{x_j})$$

- General points:
  - Number of sub-volumes not defined ⟷ minimum sub-volume size

Possible Jackknife/Bootstrap regions on SDSS DR7





(Norberg et al. 2012)

Internal vs external error estimators

	Internal	External
Visible statistical data properties (e.g. luminosity and colour)	Yes	Possible, but hard
Hidden statistical data properties (e.g. higher order statistics)	Yes	Hard or impossible
Construction/Stablity of covariance matrix (e.g. noise in the recovered covariance matrix)	Unclear	Yes
Inclusion of large scale samples variance (e.g. presence of large coherent superstructures)	Yes, but not correctly	Yes, but not always
Intrinsic limitations due to survey size	No	Yes

Internal: data driven error estimates like Bootstrap / Jackknife

External: mocks, analytic work

Summary (part 1)

- Many survey characteristics covered:
  - Empirical:
    - Redshift distributions
    - Number counts
  - Intrinsic:
    - Galaxy luminosity functions
    - Galaxy correlation functions

- Different type of surveys:
  - Galaxy formation vs cosmology focused

- We live in an era of galaxy surveys… and it will carry on at least to the mid-2030's, and plans are already starting to talk about the next ones…

The future is still bright!

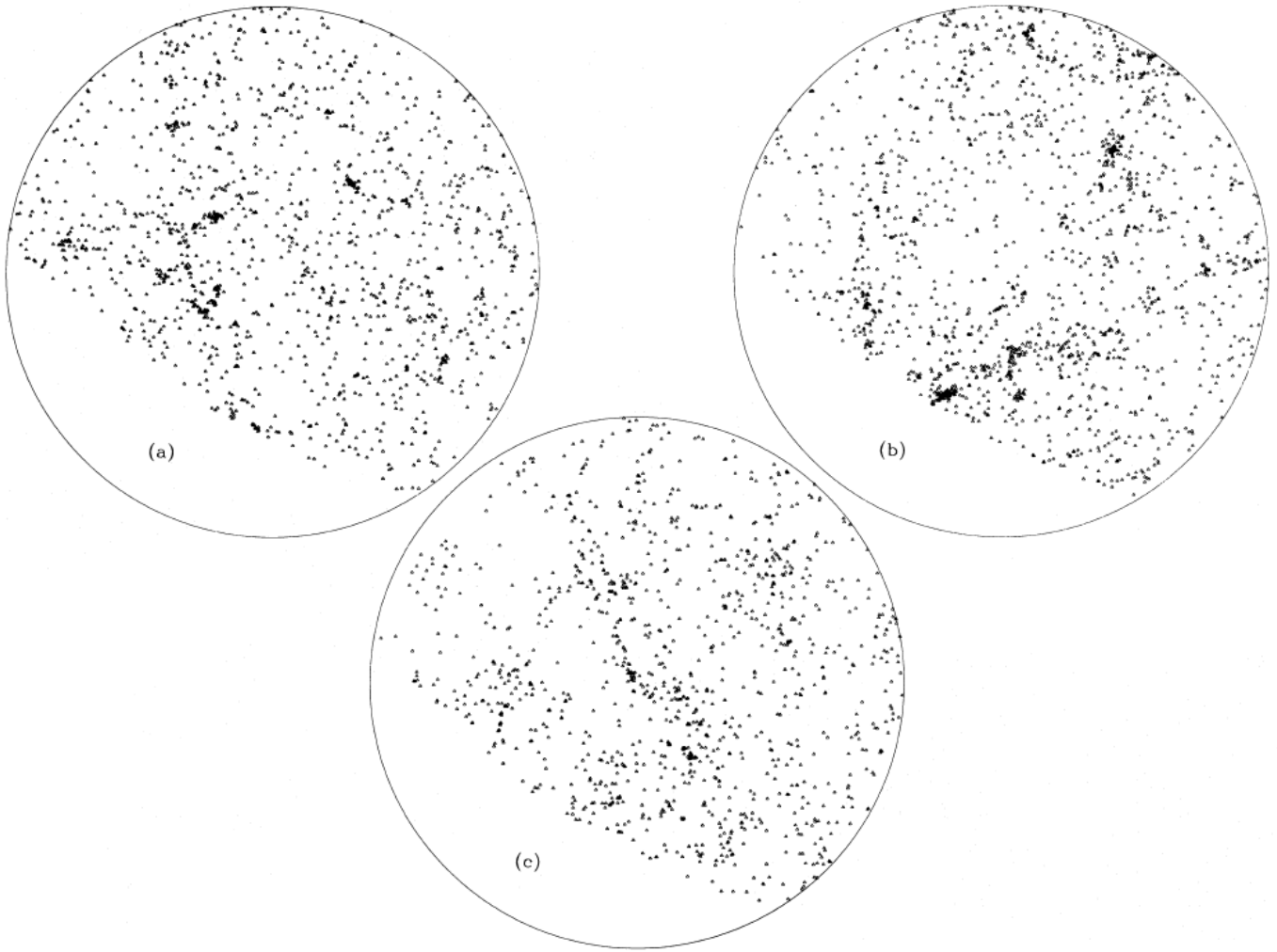
From cosmological simulations to real observations: the ins and outs of the mock world… Part 2: the fake universe

Peder Norberg (Durham University)

Mexican Numerical Simulations School
October 3rd - 6th 2016

Mocks: some historical background

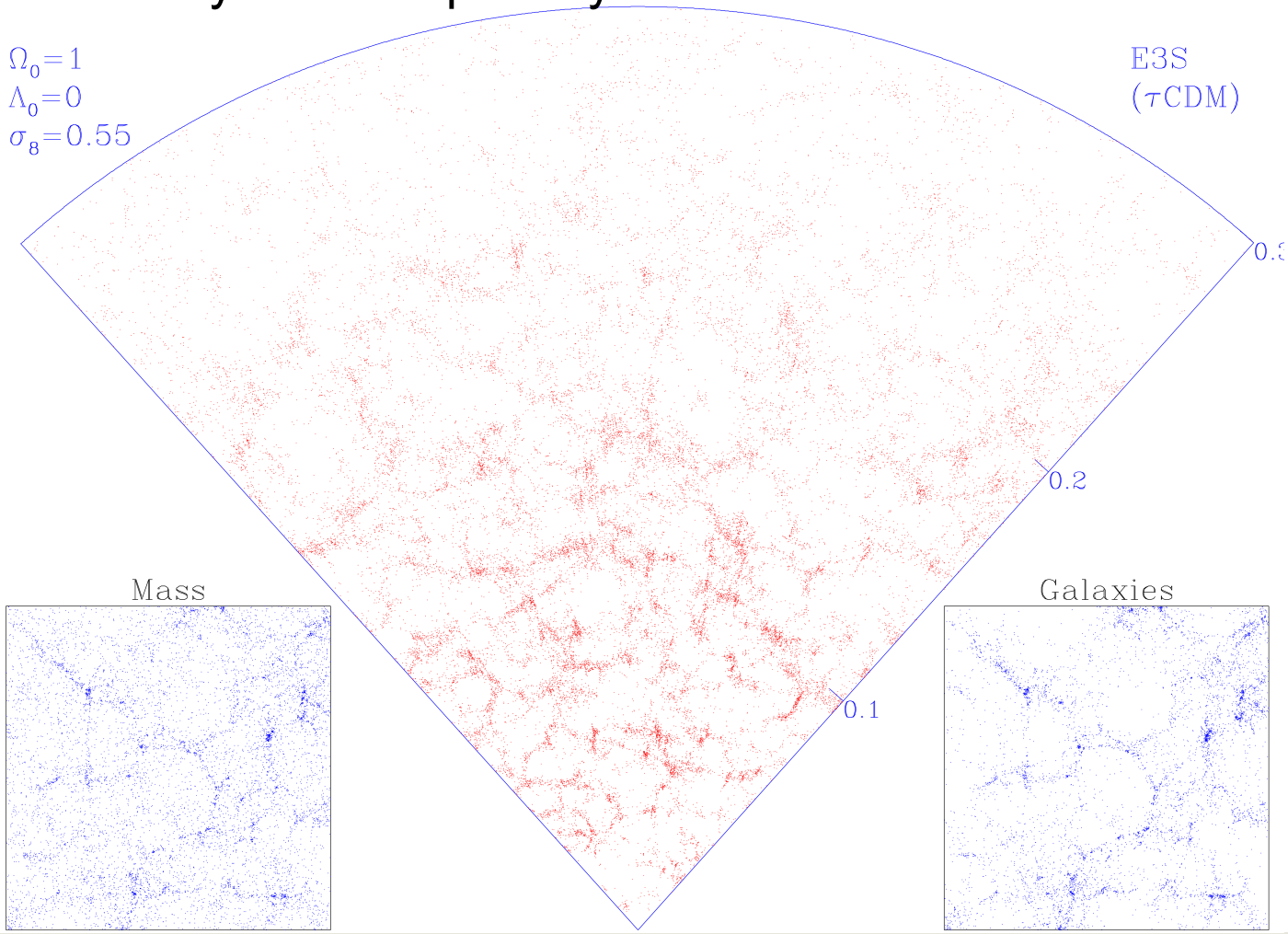
- Simulations have been used in earnest since the 1980's to rule out models or more precisely define the favoured ones.
- One of the 1st mocks (Davis et al. 1984) for the CfA survey:



(Davis et al. 1984)

Large scale breakthrough of mocks: 2dFGRS

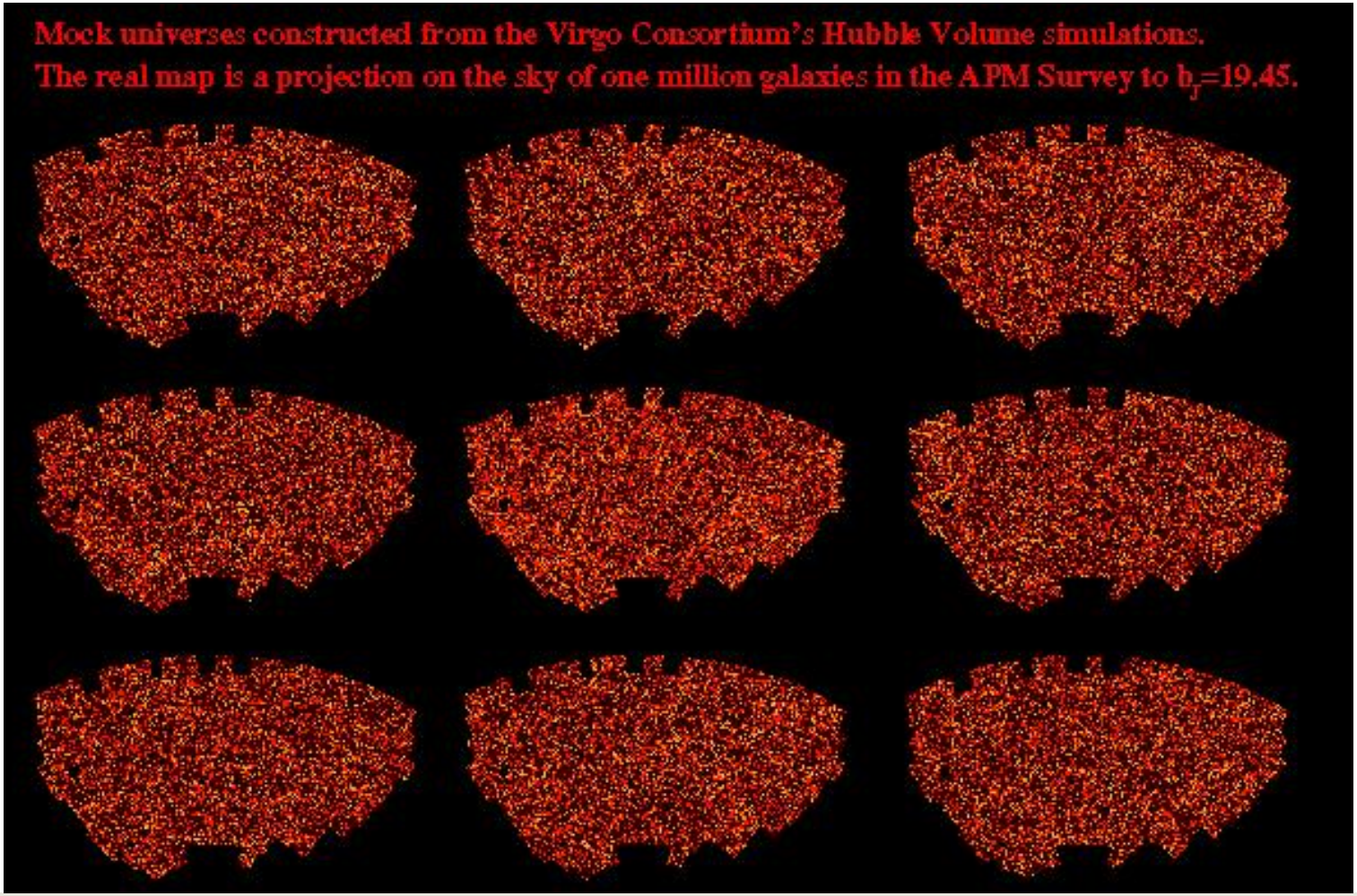
- Even though SDSS was clearly a superior data set compared to 2dFGRS, mocks were a key ingredient in the 2dFGRS analysis pipeline and only later adopted by SDSS.



(Cole et al. 1998)

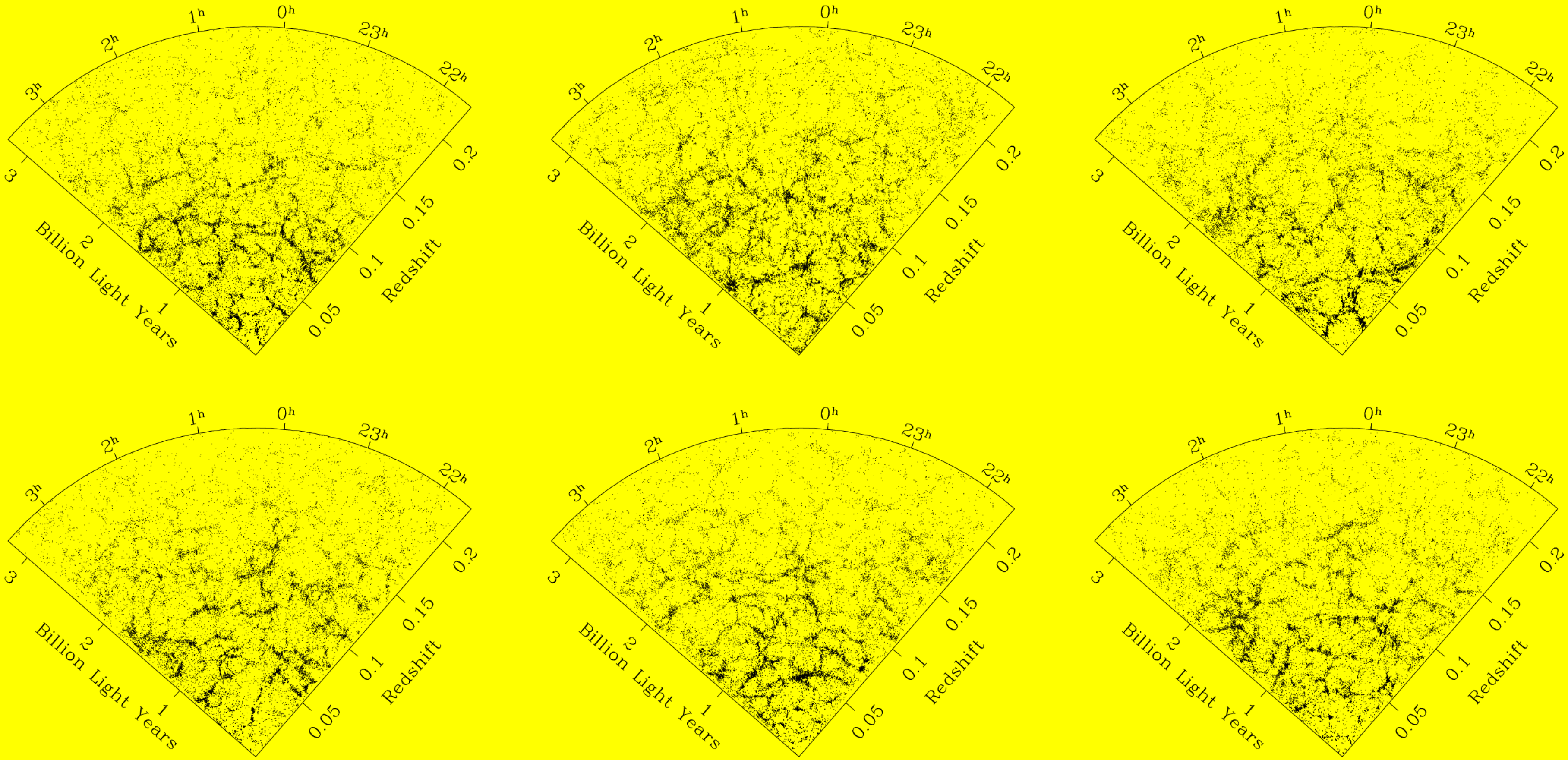
Large scale breakthrough of mocks: 2dFGRS

- Even though SDSS was clearly a superior data set compared to 2dFGRS, mocks were a key ingredient in the 2dFGRS analysis pipeline and only later adopted by SDSS.



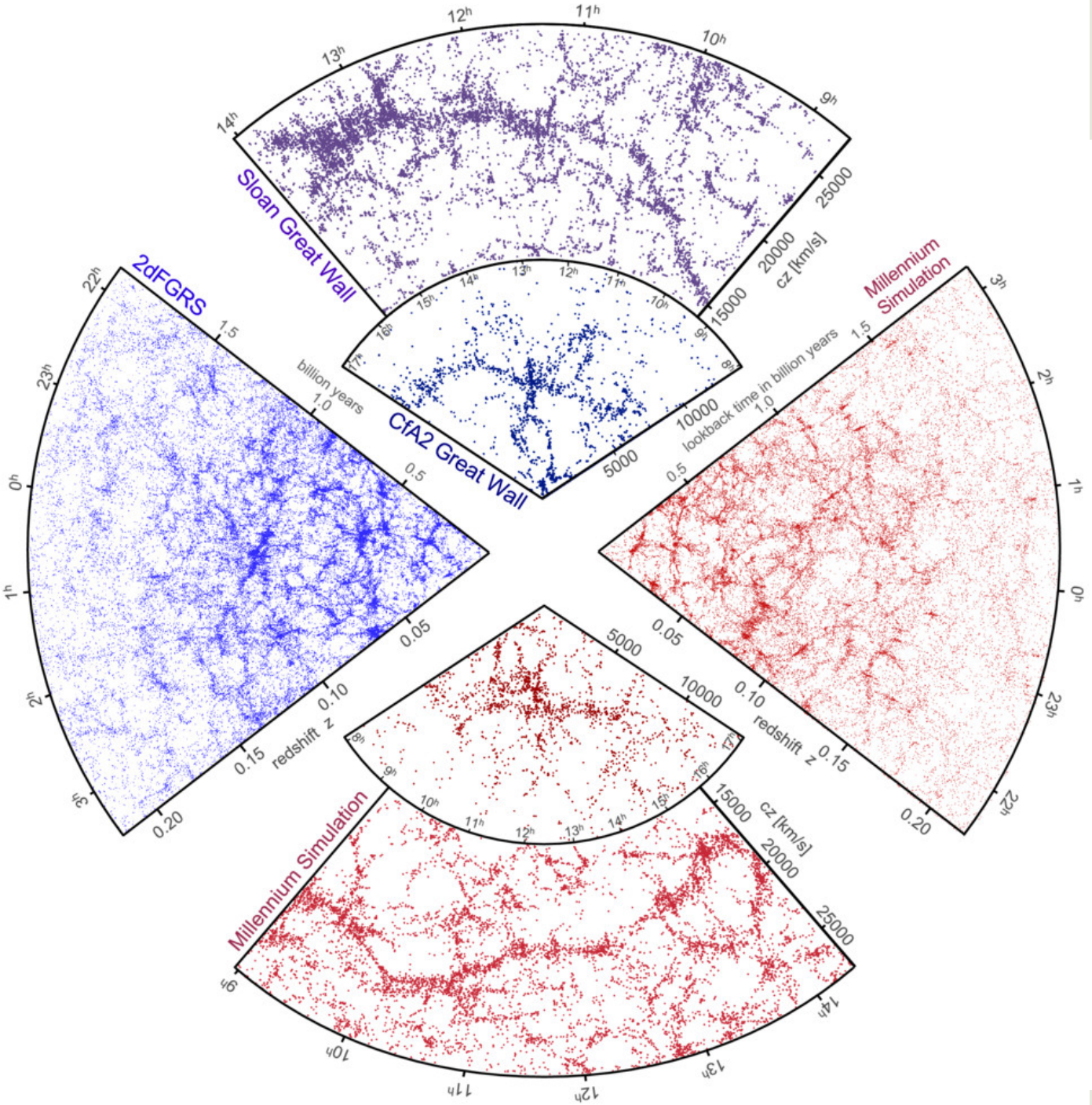
VIRTUAL vs REAL UNIVERSES II

Mock universes constructed from the Virgo Consortium's Hubble Volume simulations.
Each plot shows a projection of a 3 degree slice. The real map is from the Anglo-Australian 2dF Galaxy Redshift Survey. This is the biggest map of the local galaxy distribution to date (May 2000).
When completed, the 2dFGRS will consist of 250,000 galaxy redshifts selected from the APM Survey.



Peder Norberg, ICC & CEA, Durham University

Millennium



(Springel et al. 2005)

What are the basic steps?

- Most common and least exact:
  - Simulation in box
  - Use box (and/or its replications) to the survey volume
  - Define observer/orientation and select galaxies to match selection function
  - Key limitation: no evolution of the underlying density field

- Rather common (details in next lecture)
  - Simulation with outputs at different cosmic times (snapshots)
  - Define observer/orientation and select galaxies when they enter the lightcone
  - Key limitation: require interpolation between snapshots

- Most exact:
  - Simulation information is output on lightcone directly
  - Key limitation: cannot change observer position.

Many types of mocks:

- N-body simulations (in order of increasing mass resolution requirements):
  - Dark Matter particles: galaxies are biased tracers of the DM density field, with the biasing scheme more or less complex.
  - Halo catalogues: galaxies reside in halos, which are populated according to an assumed halo occupation distribution (HOD).
  - Sub-halo catalogues: galaxies reside in sub-halos (halos within halos), which are populated according to some pre-defined mechanism (e.g. SHAM,…)
  - Merger trees (w/o sub-halos): galaxies reside in (sub)-halos, with an evolutionary connection.
  - …
- Hydro-dynamical simulation:
  - very hard: largest box size available is ~100 Mpc (i.e. 70 Mpc/h)
- Approximate simulations:
  - Zeldovich (or 2LPT)
  - Lognormal
  - COmoving Lagrangian Acceleration (COLA) (Tassev et al. 2013)
  - …

Biased DM field mocks (e.g. 2dFGRS mocks)

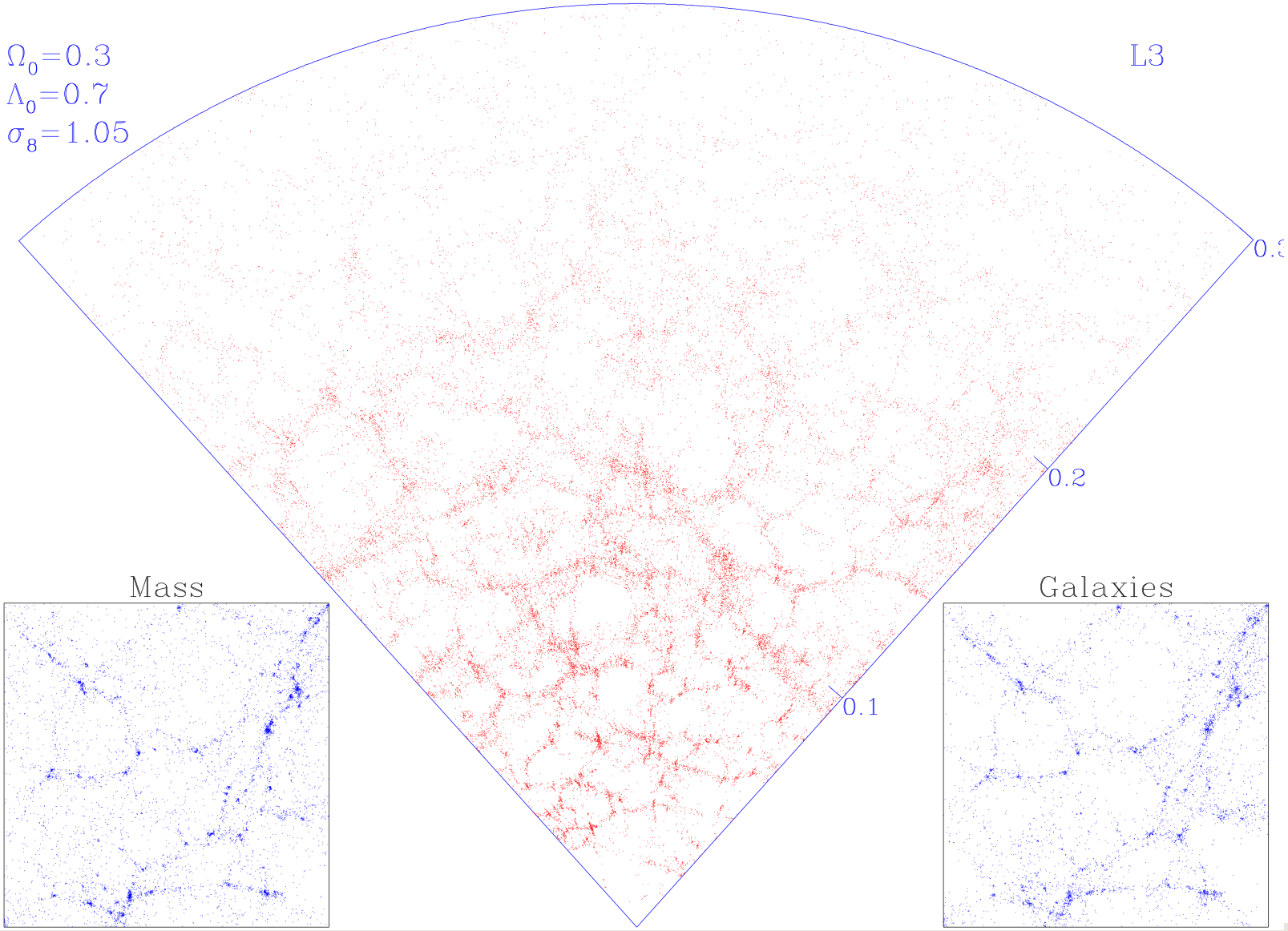
- Galaxies are a biased tracer of the dark matter density field
- Local biasing schemes as in Cole et al. (1998), i.e. the probability of a mass particle being selected is a function only of the neighbouring density field (leading to scale independent bias on large scales)
- Broadly two type of biasing schemes:
  - Lagrangian, based on the initial density field
  - Eulerian, based on the final density field
- Most successful:
  - Lagrangian with selection probability defined by:

$$P(\nu) \propto \begin{cases} \exp(\alpha\nu + \beta\nu^{3/2}) & \text{if } \nu \geq 0 \\ \exp(\alpha\nu) & \text{if } \nu \leq 0 \end{cases} \qquad \begin{aligned} \delta_S(\mathbf{r}) &= (\rho_S(\mathbf{r}) - \bar{\rho}) \,/\, \bar{\rho} \\ \nu(\mathbf{r}) &= \delta_S(\mathbf{r})/\sigma_S \end{aligned} \qquad \sigma_S^2 = \langle |\delta_S|^2 \rangle$$

- Many cosmologies: open, flat, COBE or Cluster normalised, τCDM,…

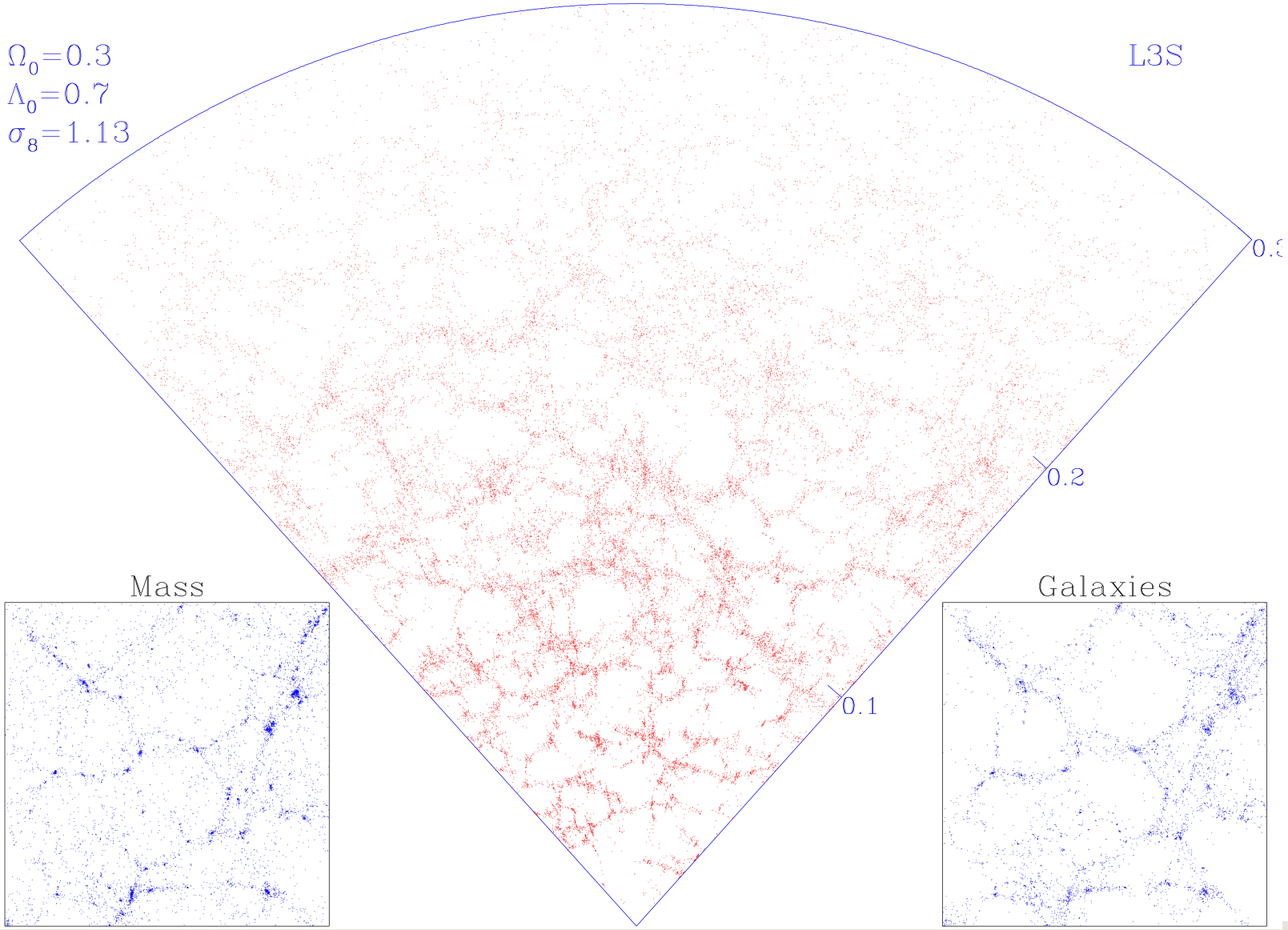
(Cole et al. 1998)

Biasing schemes comparison



(Cole et al. 1998)

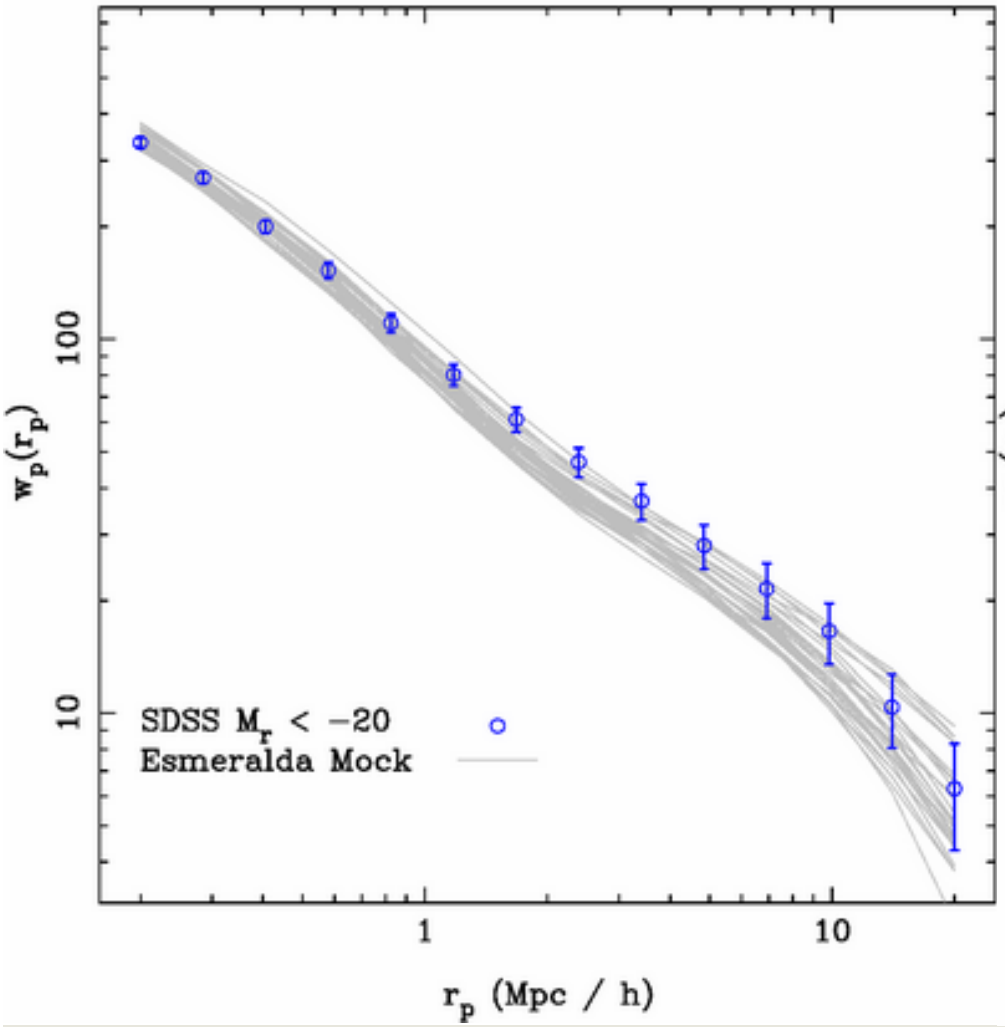
Biasing schemes comparison

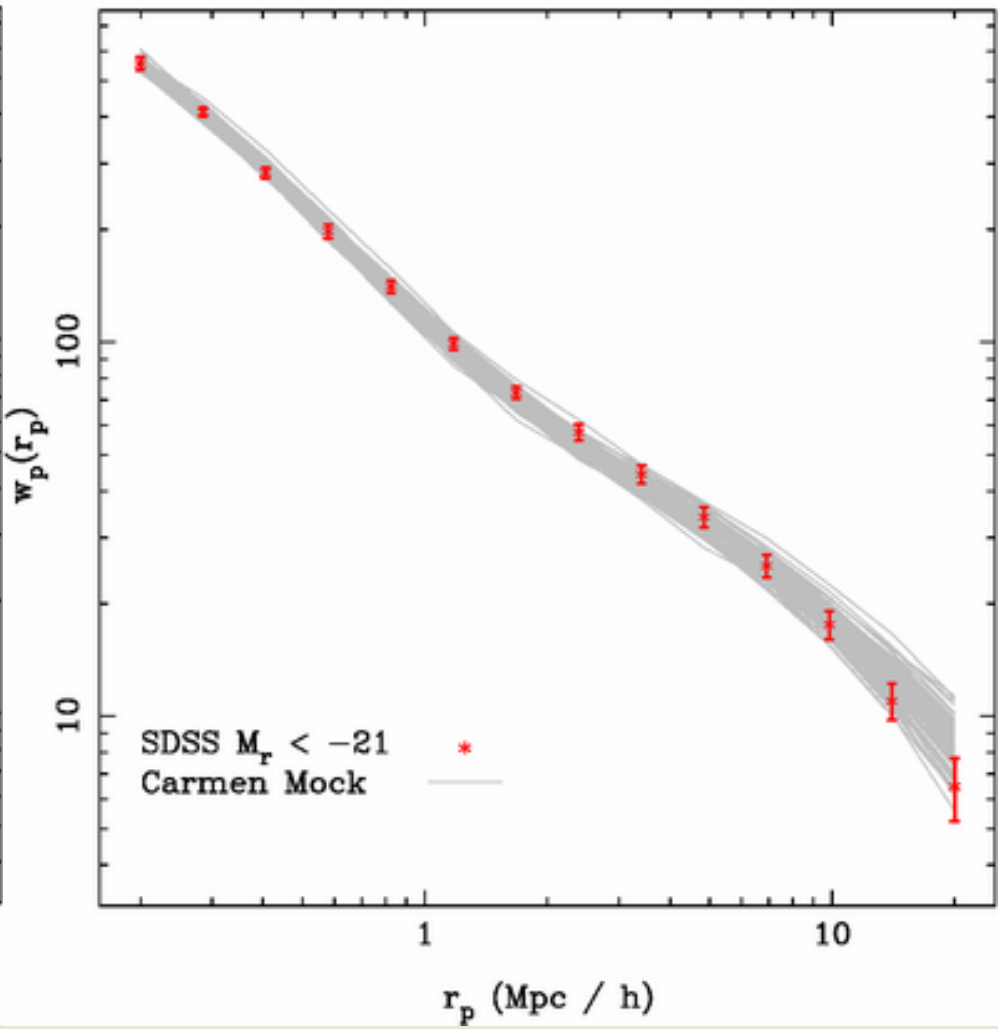


(Cole et al. 1998)

Halo catalogue mocks (e.g. SDSS LasDamas mocks)

- Halo model (see Aldo's talk)
- HOD evolution: see e.g. Contreras et al. (arXiv:1607.06154)





(Berlind et al. -- 2009 Las Damas Collaboration)

HOD mocks: some practical comments

- Definitions matter especially in HOD analyses:
  - Masses:
    - w.r.t. critical density: $M_{200,c}$ or $M_{vir,c}$
    - w.r.t. mean density: $M_{200,m}$ or $M_{vir,m}$

    These mass definitions have different redshift dependences…
  - Halo definition:
    - FoF
    - Subfind, Rockstar,…
    - …
  - Halo mass function:
    - Which fitting formula (e.g. Tinker et al., Jenkins et al. , Warren et al.,…) or the intrinsic one of the simulation?
  - Bias function:
    - Is the intrinsic one of the simulation in perfect agreement with the one used for the analysis?
  - Radial density profile (part of the HOD model)

Sub-halo catalogues (SHAM & empirical models)

- Sub-halo abundance matching & empirical models (see Aldo's talk)

- Key points:
  - Basic SHAM (i.e. with no scatter) has "no" free parameter, but depends on:
    - matching variable and its definition (V_{peak}, V_{max}, M_{infall},…)
    - numerical resolution of simulation
    - halo and merger tree (e.g. Rockstar, Subfind,…)
    - …
  - Unlike HOD mocks, there is a formalism for evolution (as they build on merger tree information)
  - Unlike HOD mocks, they have potential to provide predictions, but not much explored (yet).
  - Like HOD mocks, they will *by construction* mimic very well a range of observable data to which they have been statistically tuned.

Merger-Tree mocks (e.g. Galform / Millennium)

- Various semi-analytic galaxy formation models:
  - Galform: Durham Model and its derivatives, like Galacticus
  - L-Galaxies: Munich model and its derivatives, like SAGE
  - …

- By construction such models are suited for the creation of mocks

- Key limitations:
  - High level models with many physical ingredients: requires significant expertise in using/running them
  - accuracy with which the observables can be matched despite models being tuned

- Key strength:
  - predictive power well in advance of data being available
  - consistent set of observables (but do not need to match data!)

N-body Mocks: pros & cons

	Biased DM	HOD	SHAM	SAM
Min. mass resolution (m_p in Msol/h) Desired	$\sim 10^{12}$ $\sim 10^{11}$	$\sim 10^{10}$ $\sim 10^{9}$	$\sim 10^{9}$ $< 10^{8}$	$\sim 10^{9}$ $< 10^{8}$
Merger-Tree	No	No	Yes	Yes
Lightcone output only	Yes	Yes	No	No
Lightcone + Merger-Tree	Not needed	Not needed	Yes	Yes
Cosmological Volumes (snapshots)	Yes	Yes	Not yet	Not yet
Cosmological Volumes (lightcone)	Yes	In progress	TBD	TBD

Lightcones: from Hubble to Euclid via MICE

- Lightcone simulations are in principle simple:
  - Run the standard N-body simulation
  - Create output of a sub-set of the data at much higher time resolution than standard snapshots (for an observer, an orientation and an opening angle). Some interpolation might be required.
  - That's it!

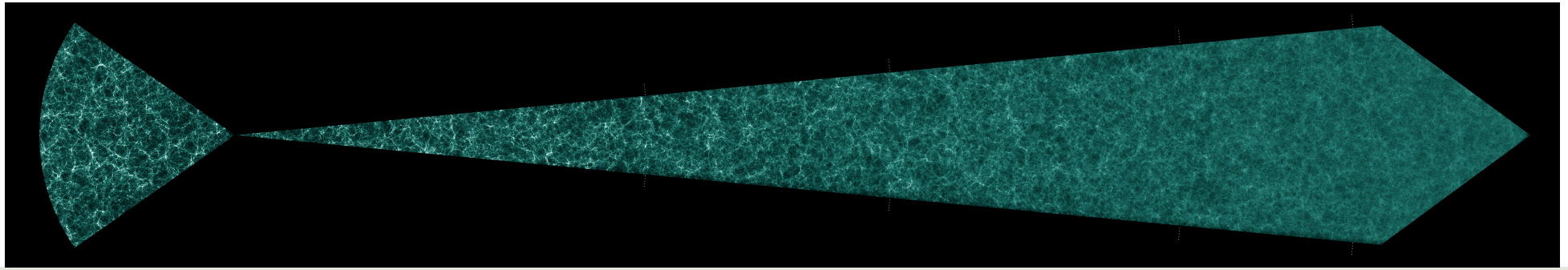
- But if it is so simple, why is it not regularly done?
  - typical simulator do not see much use of the lightcone data
  - N-body codes are not set up to provide lightcone output (snapshot information often sufficient, but not always)

- Lightcone output is cheap in terms of space:
  - typically ~50% of the particles information of one single snapshot (for an octant), and much less for any deeper option.
  - Usually only one lightcone per observer (etc) per run (unless different set up from start)

Hubble volume lightcone

- Hubble volume (1997-1999):
  - 1 billion particles in a 3 Gpc/h box
  - m_p ~ 2.2 10^{12} Msol/h
  - WMAP-1 cosmology (Ω_m=0.25, Ω_Λ=0.75, h=0.73,σ_8=0.9)
  - Evrard et al. (2002)

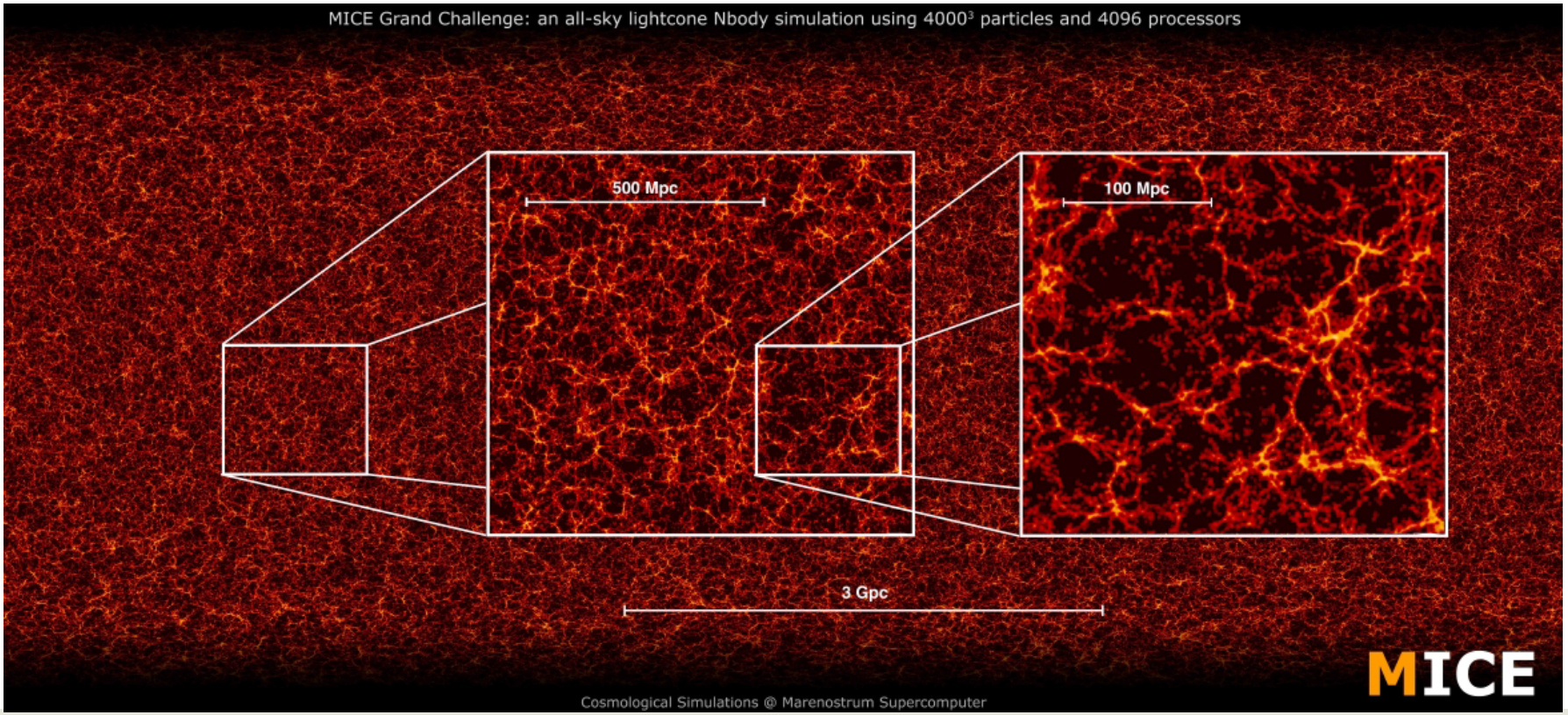
Name	center	solid angle	$z_{\max}$ (Λ)	$z_{\max}$ (τ)
MS	$(L/2, L/2, L/2)$	4π	0.57	0.42
VS	$(0, 0, 0)$	4π	0.57	0.42
PO	$(0, 0, 0)^{a}$	$\pi/2$	1.46	1.25
NO[b]	$(L, L, L)^{a}$	$\pi/2$	1.46	1.25
DW	$(0, 0, 0)^{a}$	$10° \times 10°$	4.4	4.6
XW	$(0, 0, 0)$	$16° \times 76°$	6.8	–



(Evrard et al. 2002)

MICE lightcone

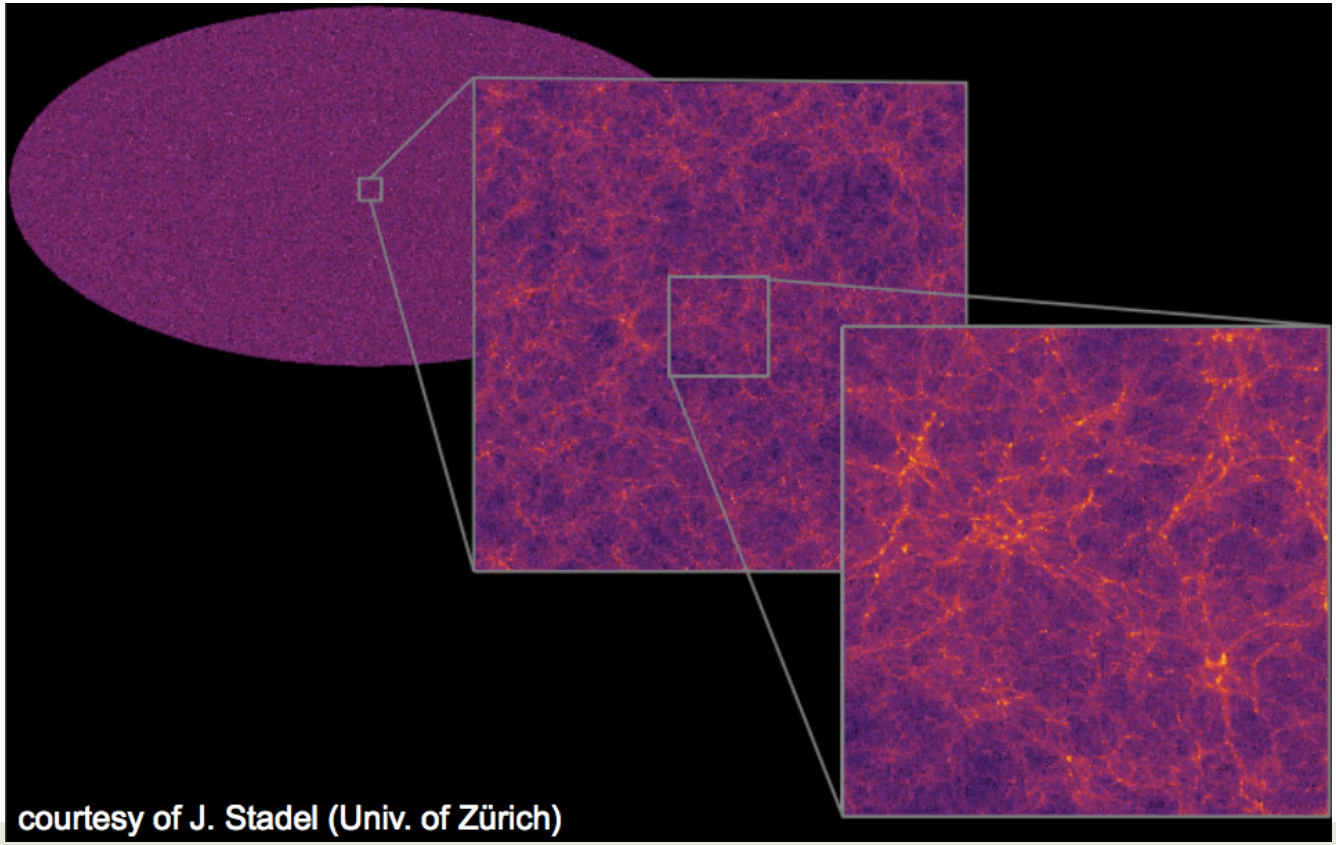
- MICE Grand Challenge (2013):
  - 70 billion (4096^3) particles in a 3.072 Gpc/h box
  - $m_p \sim 2.9\ 10^{10}$ Msol/h
  - WMAP-5 cosmology (Ω_m=0.25, Ω_Λ=0.75, h=0.7, σ_8=0.8)
  - Fosalba et al. (2015)



(Fosalba et al. 2015)

Euclid Flagship Simulation

- Euclid Flagship simulation (2016):
  - 2 trillion (12600^3) particles in a 3.8 Gpc/h box
  - $m_p \sim 2\ 10^9$ Msol/h
  - Euclid reference cosmology (Ω_m=0.??, Ω_Λ=0.??, h=0.??, σ_8=0.?)
  - Stadel/Teyssier et al. (in prep)



Summary (part 2)

- Mocks are integral part of analyses today
- Mock design is closely related to the science goals
- Unfortunately specific mock development happens most often when needed. Various reasons:
  - Limited use on their own right
  - Require a significant amount of work to develop