Simulations for Large-Scale Structure Surveys

Risa Wechsler
(Stanford/SLAC/KIPAC)

+ collaborators in DES + DESI & LSST

Large Scale Structure in Garching
July 23, 2015

- Goals
  - What do we need to simulate for large cosmological surveys?
- Methods
  - for both high resolution and low resolution dark matter simulations
- Uses
  - examples from the Dark Energy Survey

Dark Energy Survey

- 5000 sq. degree imaging survey, 2013-2018
- 300 million galaxies, grizY to ~ 24
- weak lensing, galaxy clustering, galaxy clusters, supernovae
- results this week on weak lensing, photo-z's, systematics, red galaxy selection, galaxy-CMB cross correlation all used simulations

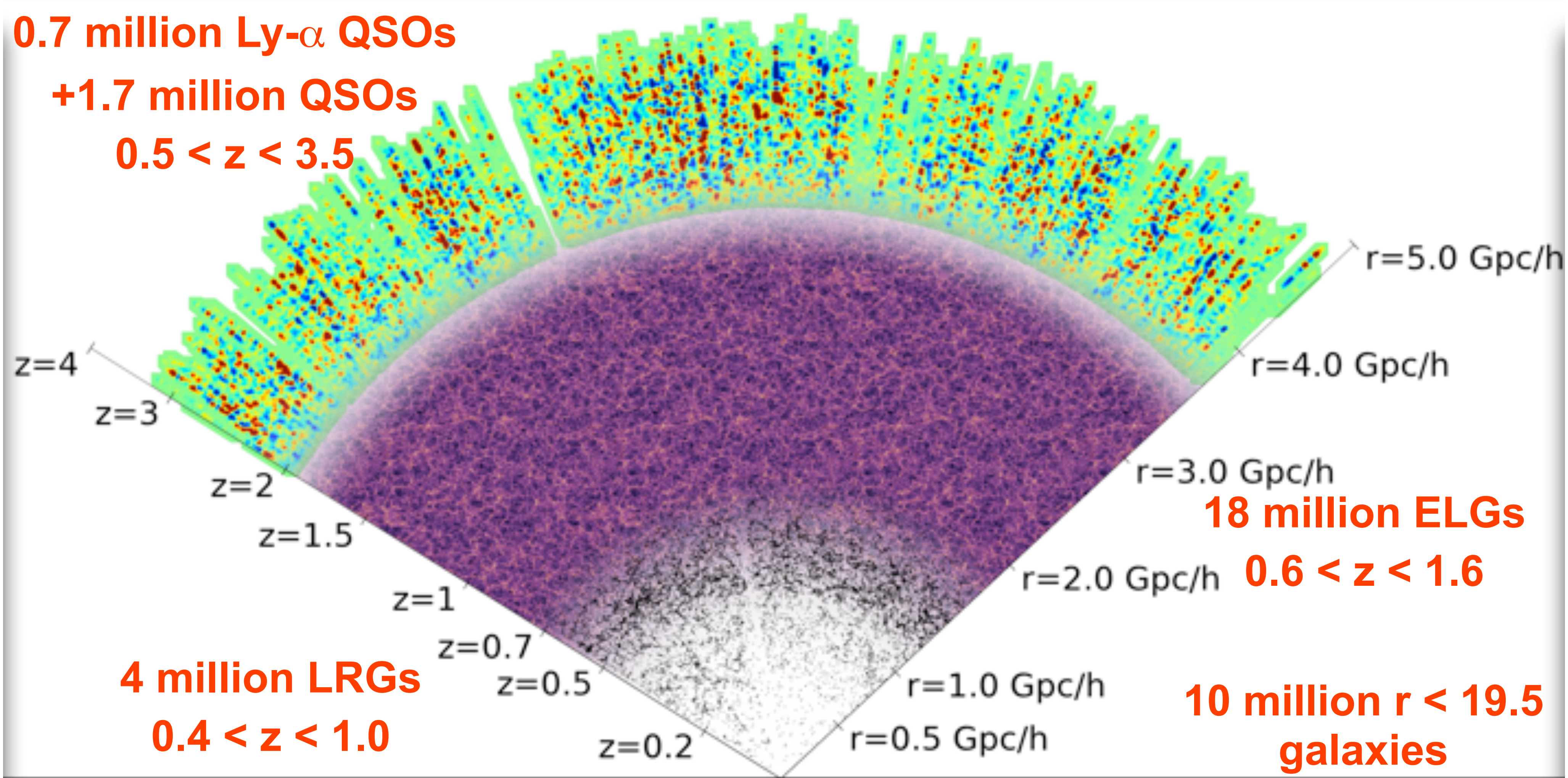
DESI:

The Dark Energy Spectroscopic Instrument

- new instrument: 5000 fibers, 8 deg^2 FOV, new spectrographs with R=2000-5500
- 14000 sq. degree spectroscopic survey, 2019+
- 35 million galaxy & quasar spectra
- more than an order of magnitude increase in # of spectra and volume probed compared to current state-of-the-art
- although main driver is BAO, significant power in joint probes, small scales

Overview of the DESI Survey

- Four target classes in dark time spanning redshifts z=0.4 ➔ 3.5
  - these are roughly the easiest 25 million spectra to measure.
  - includes nearly all the massive black holes in the Universe (LRGs + QSOs)
- Additional Bright Galaxy Survey will target all ~10 M galaxies with r < 19.5 (z=0-0.4)
- Milky Way Survey will map ~10 million stars



What are simulations good for, anyways?

- Accurate predictions of the mean signal, as a function of all cosmological models to be tested
  - e.g., matter power spectrum
  - galaxy statistics (need detailed understanding of how galaxies trace the dark matter)
  - more complex statistics: e.g. cross-correlation of galaxies with CMB, galaxies with clusters, galaxies with troughs, lensing x CMB lensing etc...
- Understanding covariances between observables
- Astrophysical systematics
  - e.g. impact of galaxy formation on matter power spectrum, intrinsic alignments, etc.
- Data systematics
  - e.g. photometric redshifts, calibration, dust, star-galaxy separation, deblending, etc etc -- simulations can provide realistic correlations with underlying structure
- Testing analysis pipelines
  - especially important for complex cosmological analyses that involve multipule probes and complex systematics

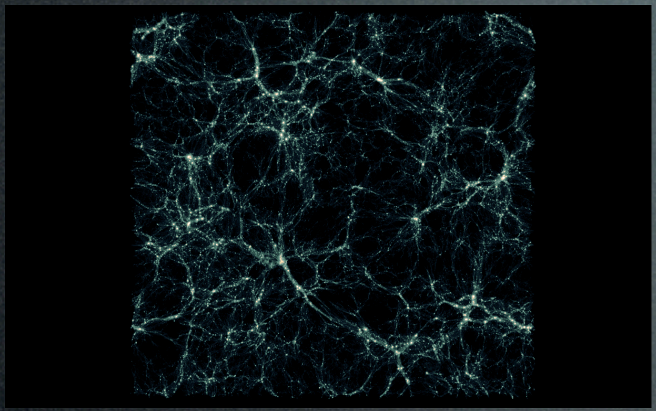
From Precision to Accuracy

- Data is already very precise, and rapidly getting more so. A full understanding of our systematic errors is essential for accuracy.
- Various ways to address this:
  - (Blind) simulation analysis
    - do a calculation as realistic as possible where you know the truth, and check that you got the right answer
    - do this again where you don't know the truth, but can check afterwards
  - Null tests
    - calculate signals that should not exist if you have controlled for systematics
  - Blind data analysis
    - shift your data vector by an unknown amount, analyze real and fake data together, etc
- Should do all of these!

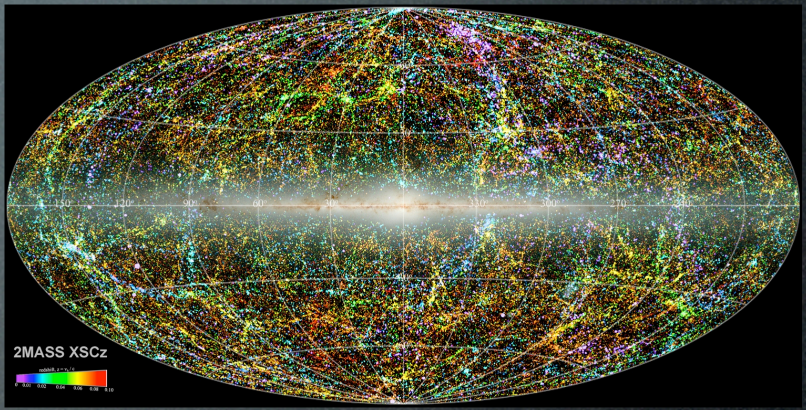
Simulated Sky Surveys

- Fundamental limit of how much cosmological information we can extract from surveys will be our ability to model and understand systematics over the full survey area and on small scales
- Need accurate simulations that model realistic galaxy populations as well as survey details
- Our understanding of galaxy formation will be an increasingly limiting factor as we push to smaller scales
- Ideally want to model all of the major cosmological probes for a survey in one simulation
    - observed properties of galaxies
    - large-scale structure of galaxies
    - realistic impact of lensing shear on galaxies
    - as many relevant observational systematics as possible
- Want to produce many full area and depth sky surveys; need lightweight simulations
    - many cosmological models
    - a variety of galaxy models for a given cosmology
    - multiple skies for covariance

Modeling galaxies



→



- ideal: predict the galaxy population for a given cosmological model e.g. P(k | L, SFR, color, etc) from first principles

  note that the previous speakers just told us this is not actually (yet) possible!

- practical: describe the galaxy population for a given cosmological model with a flexible parameterization; marginalize over this parameterization for the possible galaxy population when constraining cosmology

- + lots of possibilities in between...

Approaches to and considerations in galaxy modeling

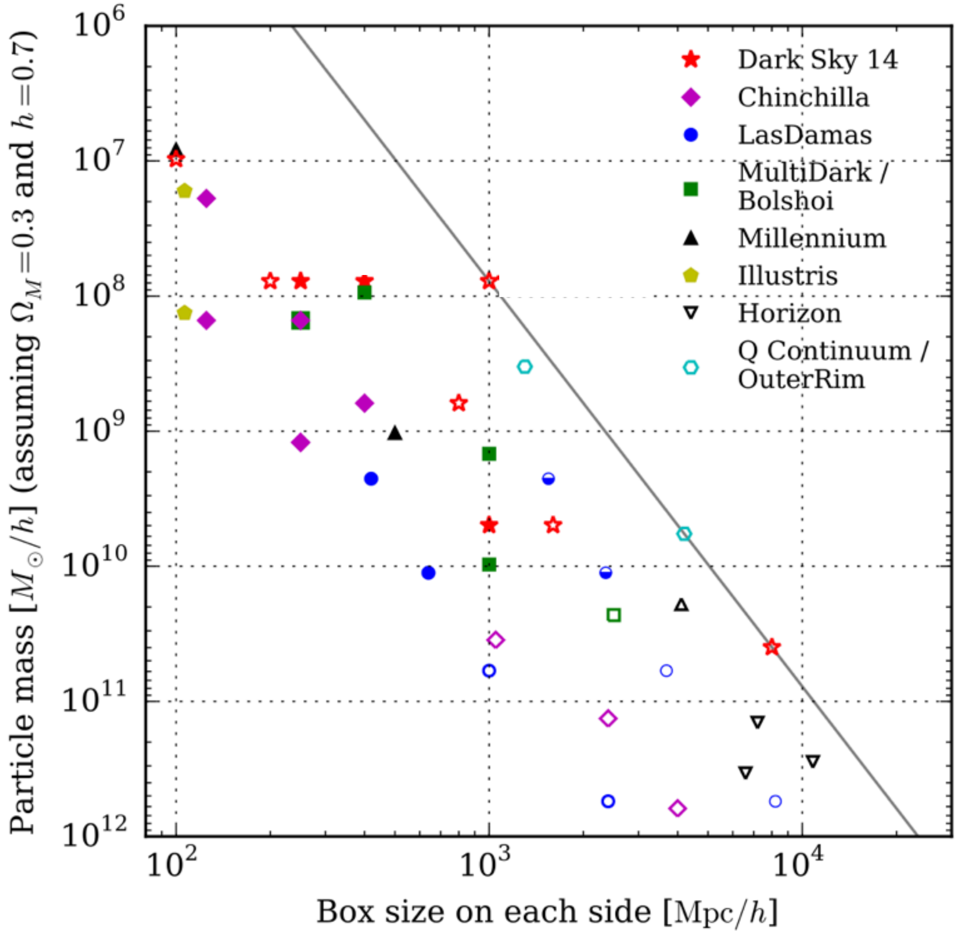
- hydro sims
- semi-analytic models
- empirical models
  - lots of flavors
  - HOD/CLF/SHAM/CAM/ADDGALS
- key distinctions:
  - are galaxy properties determined by direct simulation, by calculation from simplified physical recipes, or from empirical techniques?
  - are all galaxies in resolved host halos?
  - are all galaxies in resolved subhalos?
  - do galaxy properties depend on merger history?

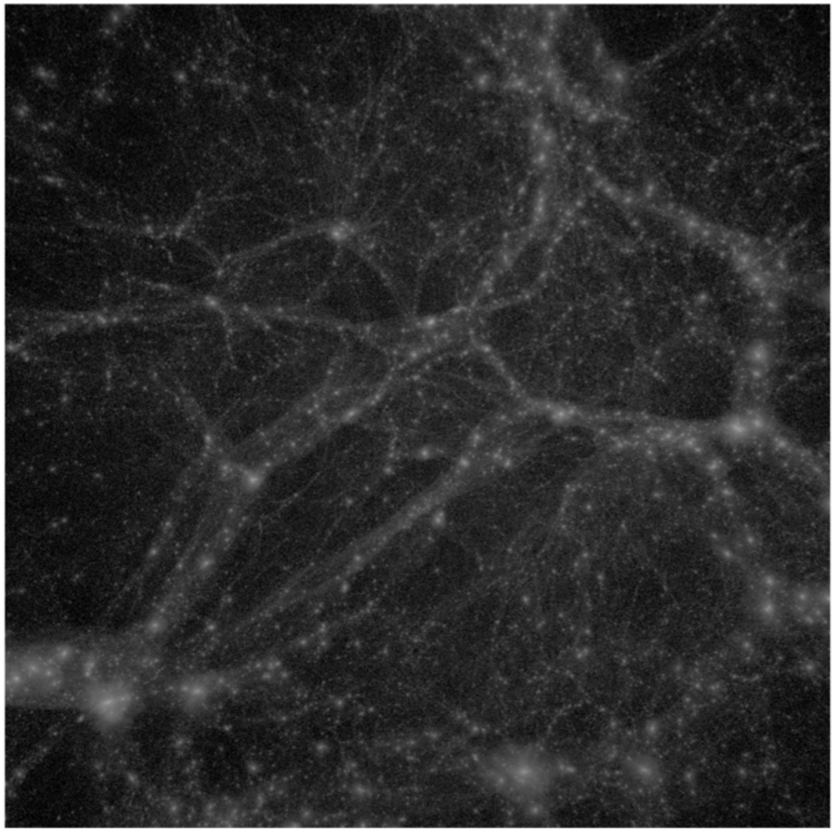
- how much volume?
  - surveys now probing several tests of Gpc3, ideally would like to simulate survey volumes many times, many cosmologies
- how faint?
  - typically cosmology probes use >0.1L* galaxies, but fainter galaxies matter e.g. for lensing source population
- what galaxy properties are important?
- what correlations are important?
  - e.g. correlations between galaxy properties and assembly history and large-scale structure

some challenges

- large dynamic range -- difficult to get volume and accuracy at the same time
- most models are computationally expensive, so it's hard to explore parameter space
- models are not general enough -- even when exploring parameter space, they don't match the data to the precision with which it is measured
- models do not capture all of the relevant physical processes, and thus may be missing some of the relevant correlations

N-body simulations





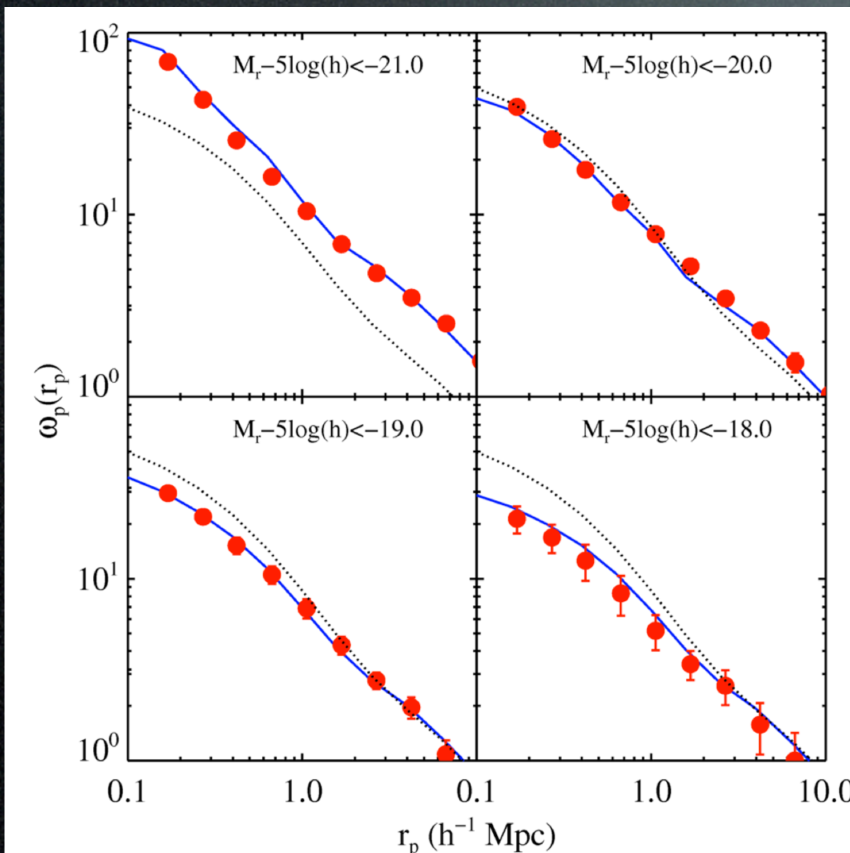
trade-offs between volume, resolution, cosmological model space, included physics.

my approach to these challenges

- large dynamic range, difficult to get volume and accuracy
  - tune models to high resolution simulations, learn method to apply to low resolution simulations
- the models are computationally expensive, so it is hard to explore parameter space fully
  - use inexpensive approaches -- empirical models win
- the models are not general enough -- even when exploring parameter space they don't match the data to the accuracy with which it is measured
  - use empirical models. start simple and generalize as much as necessary to match the data.
- the models do not capture all of the physical processes, and thus may be missing relevant correlations
  - this one is hard to prove until galaxy formation is solved. but indicates one should include correlations from halo properties (including history, substructure) where possible. can then ask: are models in which galaxy properties depend on halo properties/history/substructure sufficient?

- In practice, different approaches are very complementary -- as a community we should almost certainly be pursuing multiple approaches
- Starting with empirical models does not necessarily limit your ability to learn about physics later -- lots of great examples in the past few years of the synergy between empirical models and the most expensive and most physical models

What determines galaxy clustering?

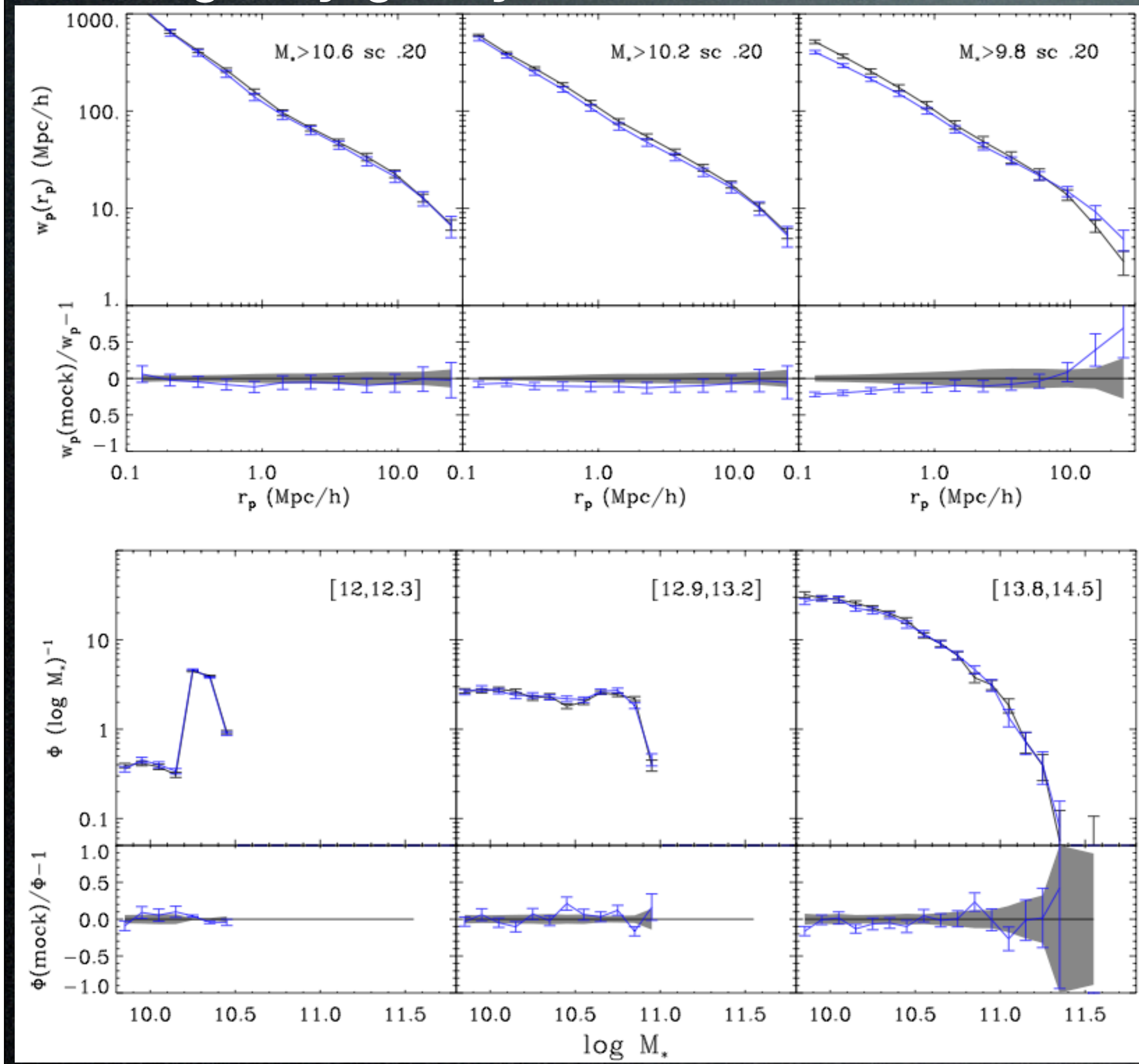


Conroy, RW, Kravtsov 2006
following Kravtsov, Berlind, RW et al 2004
+lots of later work!

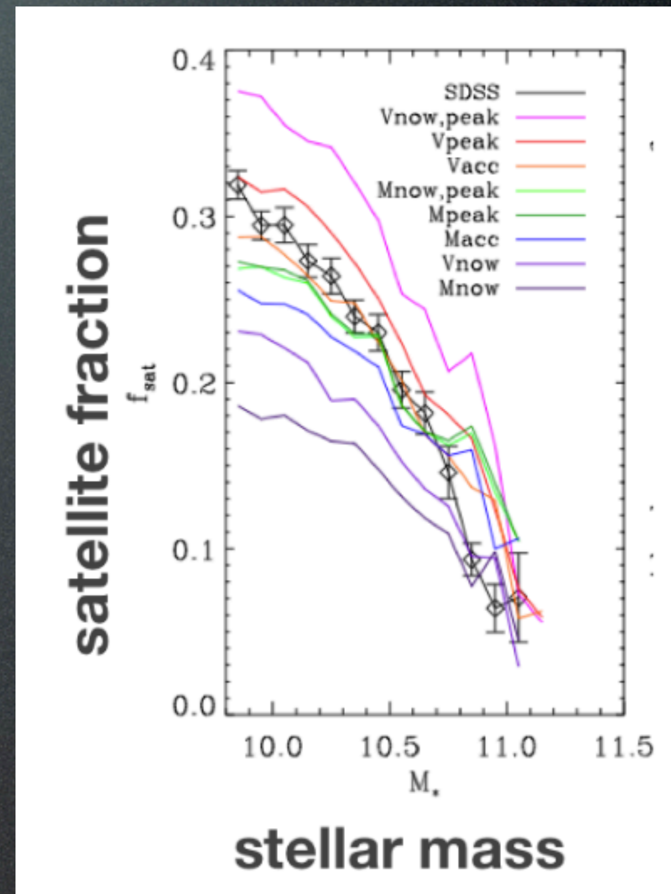
- galaxy properties seem to be tightly correlated to the properties of their halos -- (sub)halo abundance matching
- one galaxy per halo and subhalo, able to predict many galaxy spatial statistics extremely well
- but details matter:
  - need high resolution simulations to resolve subhalos and get merger trees
  - which halo property is most important in setting the galaxy properties?
  - how much scatter is there (and is it random, or correlated with other halo properties?)
  - how to treat satellites; are they they same as centrals at accretion, and what happens after accretion
  - scatter, proxy, reln between centrals and satellites can be constrained with data

Best fit model to galaxy clustering & conditional mass function

galaxy-galaxy correlation function



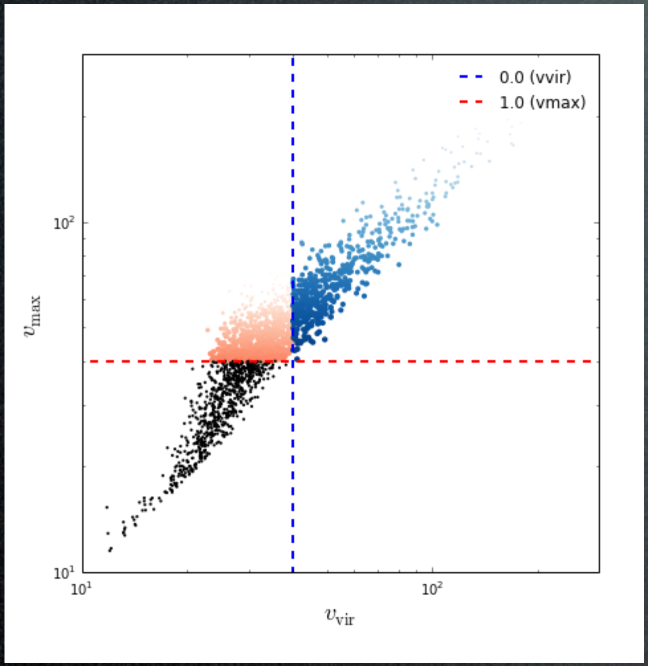
Conditional Stellar Mass Function

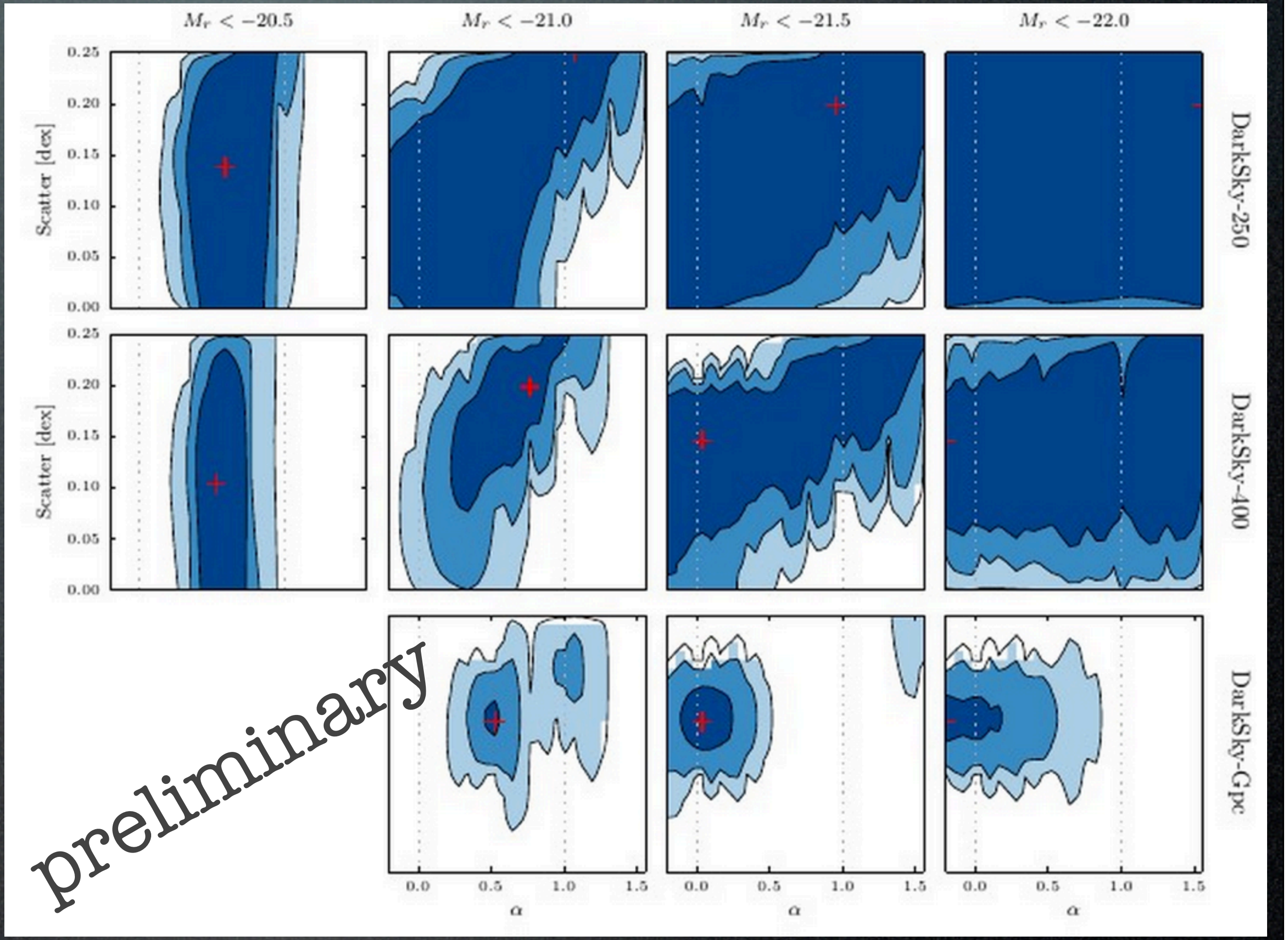


Reddick, RW, Tinker & Behroozi 2012
black: SDSS measurements
blue: best fit model

how much impact of assembly history?

Lehmann, Mao, Becker, RW in prep





- generalized abundance matching proxy: from v_{vir} to v_{max} at the peak of the mass accretion history *

* technical note for experts: v_{max} @ m_{peak} is used to avoid issues with mergers see in v_{peak}, very similar to v_{relax}

what about star formation?

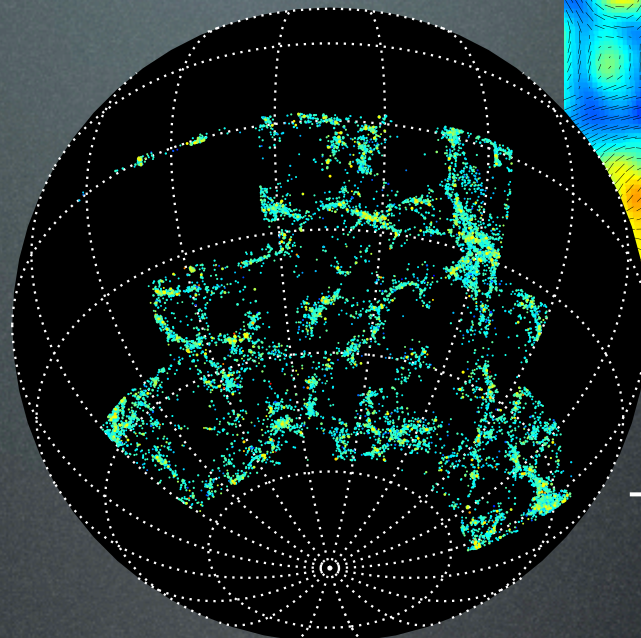
- these models work extremely well, but that doesn't mean they are right in detail! e.g. basic assumption in most abundance matching models -- nothing happens to galaxies after they stop accreting dark matter
- see e.g. Van Daalen et al 2014; Chaves-Montero et al 2015: detailed tests of SHAM assumptions vs. EAGLE simulation -- most hold, but see important impact of star formation after accretion as well as stripping of stars, both of which are ignored in most SHAM models
- Can be accounted in models that trace star formation in halos over time
  - e.g Conroy & Wechsler 2009; Behroozi, RW, Conroy 2013 -- abundance matching per snapshot, star formation connected to halo accretion history to connect them
  - Becker 2015 (1507.03605): Stellar Mass Assembly Distribution P(dM*/dt|X,a) assumes star formation history set by halo mass accretion history -- traces merger history like a SAM, but more empirical: does not invoke specific physics besides halo growth (see also e.g. Behroozi+ in prep, Wang et al 2007, Mutch et al 2013, Lu et al 2014)

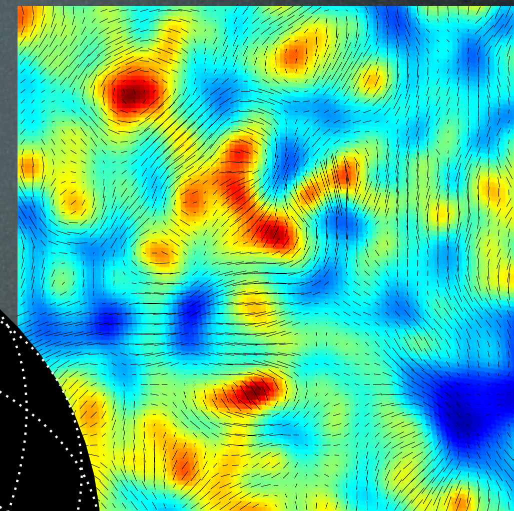
But what about the surveys?

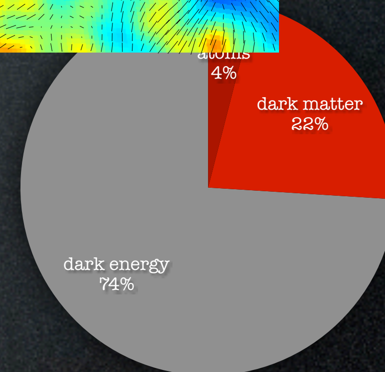
- These models (SHAM, CAM, SAM, SMAD) need very high resolution to get robust merger trees and resolve substructure
  - depends on the mass of the galaxy population being modeled, but generally need M < 1e8 particles, not practical for large survey volumes
- Need methods which can be applied to large volumes for surveys
- Basic idea: use high resolution simulations to tune models for low resolution simulations -- use only information from smooth particle distribution or high mass halos to add low mass halos or galaxies with luminosities and colors

Simulating full surveys

DES Mock Pipeline / R. Wechsler







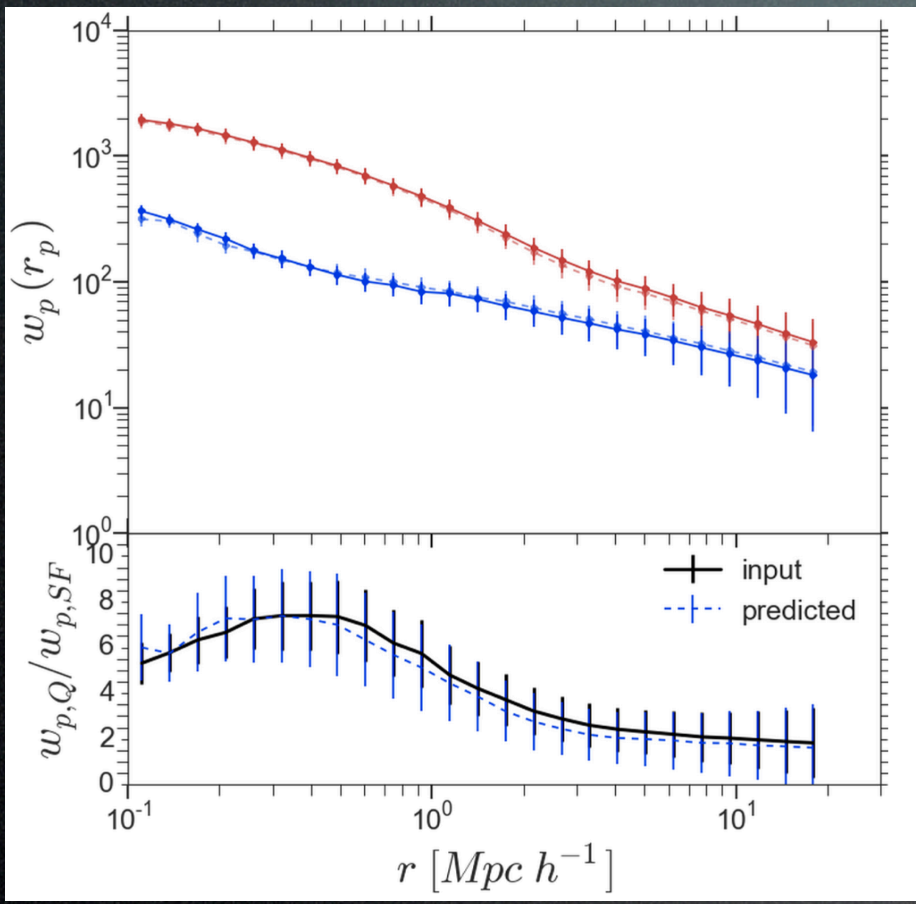
Produces simulated catalogs of ~1 billion galaxies (i ~ 26) over 1/4 of the sky, on a lightcone out to z=2

Includes galaxy photometry in many bands, galaxy magnitudes and shapes fully lensed along the lightcone

Extensively tested with SDSS, DES, etc.

Wechsler, Busha, Becker
with DES simulation working group

Learning color from high resolution



Su, DeRose, Becker, RW in prep

- train a model on a small high resolution box, apply the model to larger box
- model uses only the distance to and mass of massive halos
- similar approach (ADDGALS) used to put low mass halos in below the resolution limit (Wechsler/Busha/DeRose)

example simulation effort for DES

Cosmological simulations + galaxy populations + lensing (Busha, Becker, DeRose, RW)

UFIG imaging simulations (Chang et al)

Systematics maps (Leistedt et al)

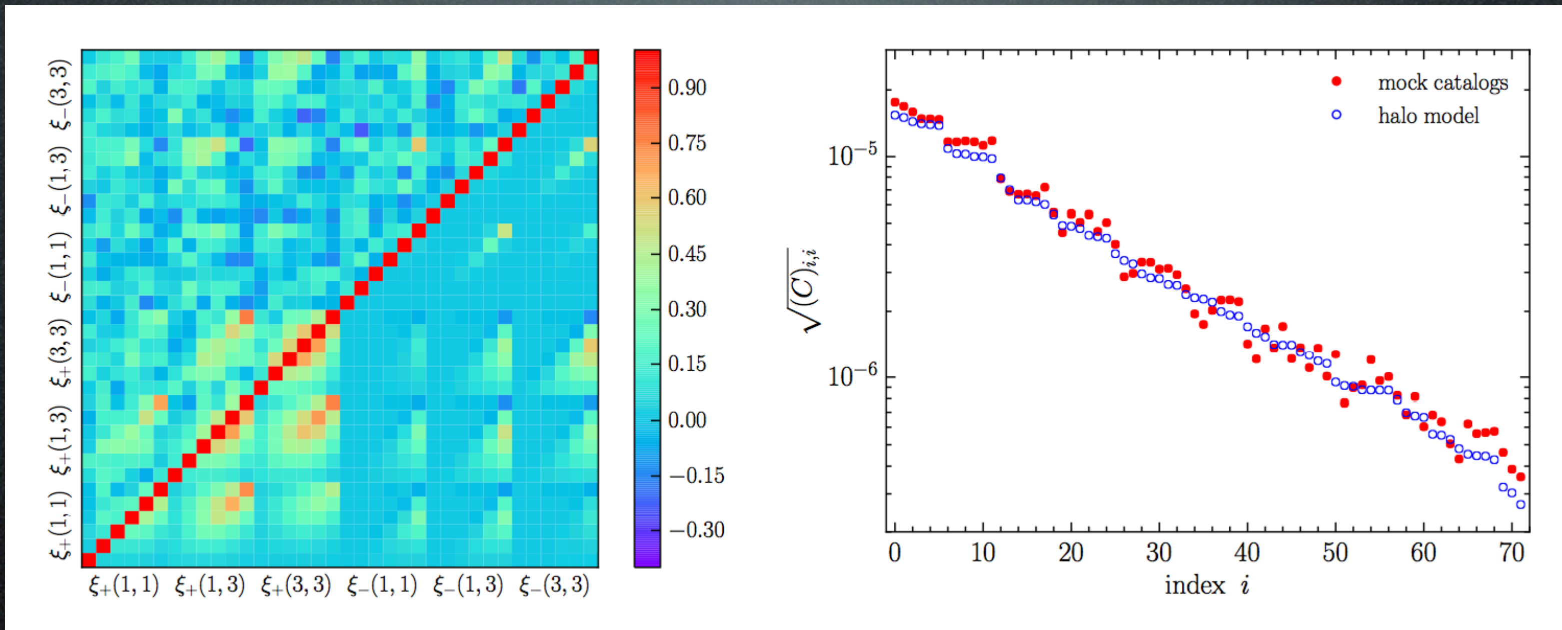
Photometric Redshifts for Lensing (Bonnett et al)

Shear 2pt Covariances (Becker et al)

Cosmology Constraints from Shear 2-pt (DES collaboration)

covariances for shear

simulation covariances
(Becker)



halo model covariances
(Krause & Eifler)

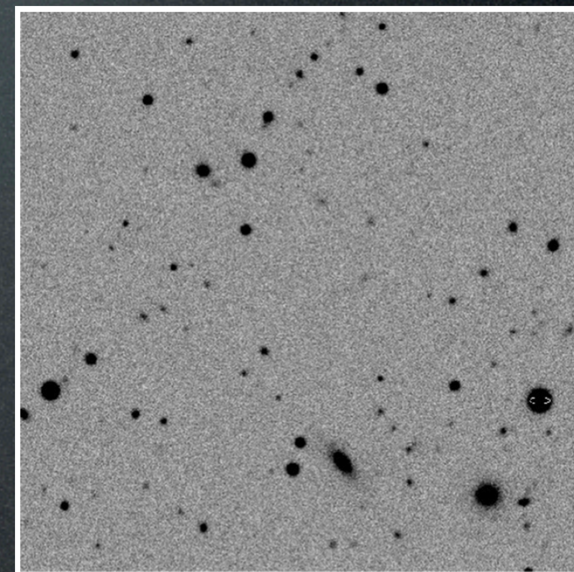
Becker + DES Collaboration 2015

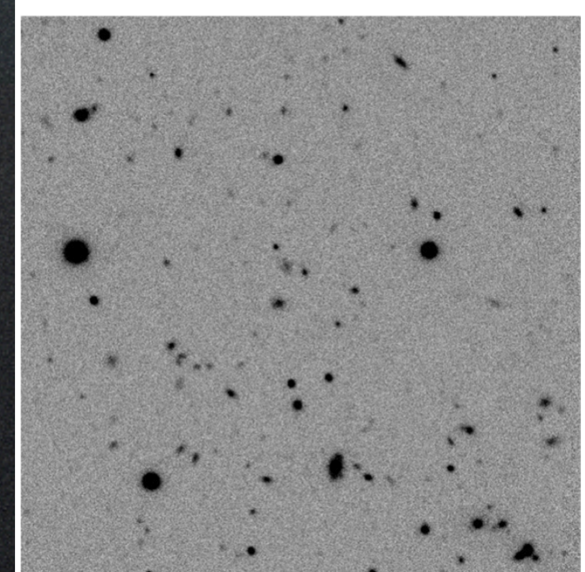
- used 126 realizations of the DES-SV footprint, populated with simple galaxy model
- compare with halo model estimator
- also compared jackknife on sims to jackknife on data, to look for extra sources of noise in the data

from galaxy catalogs to images

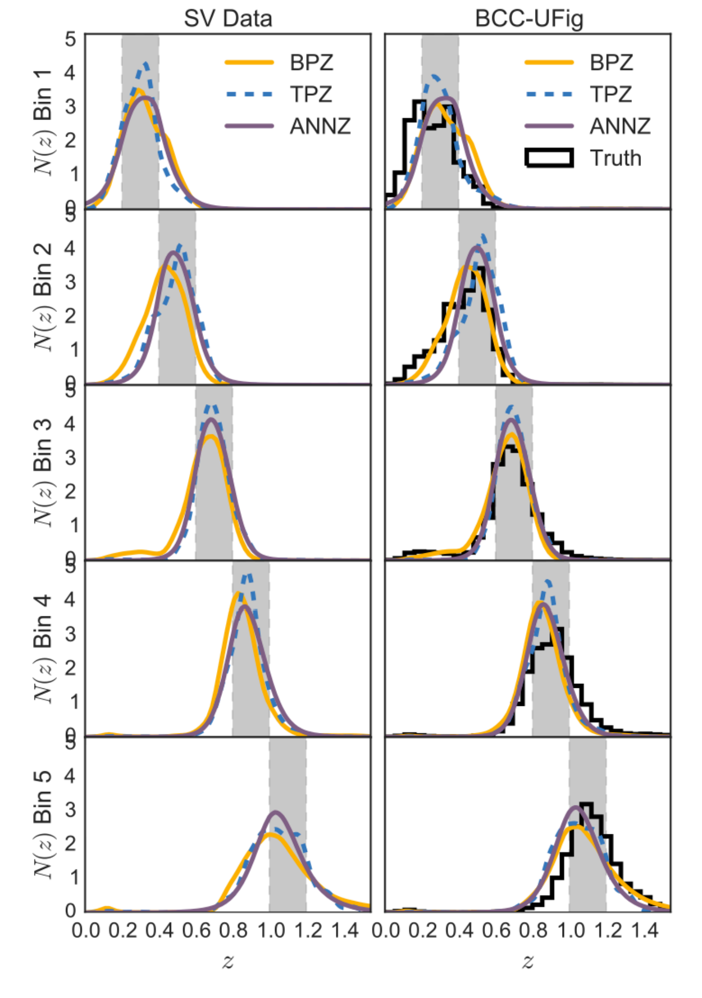
ETHZurich team: Amara, Akaret, Bruderer, Chang, Refregier
with Busha, Becker & RW

Chang et al 2015,
ApJ 801, 73





- Input galaxy catalogs into UFIG (Ultra Fast Image Generation)
- Create images, process similarly to DESDM pipeline
- Combine with DES systematics maps (Leistedt 2015 on arxiv yesterday), run analysis tools
- Allows us to understand how to connect observed properties (including complex systematics) with true properties

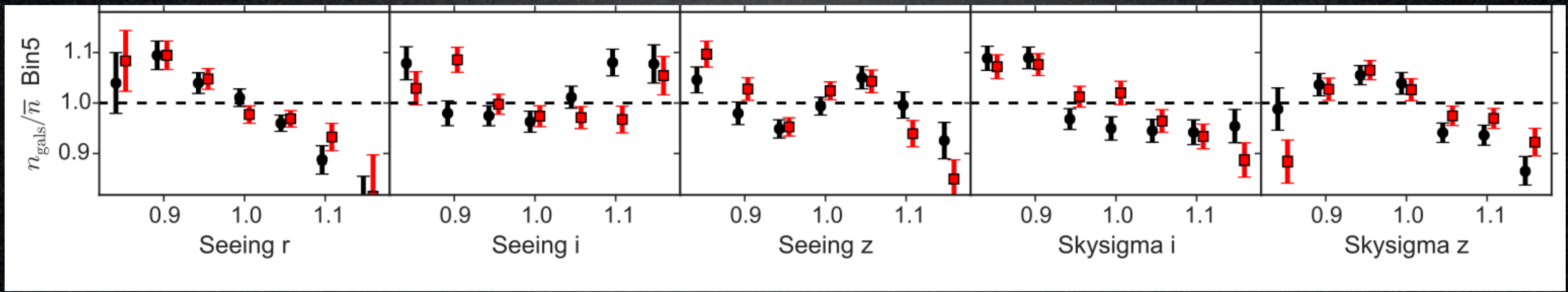


Photometric redshifts and systematics

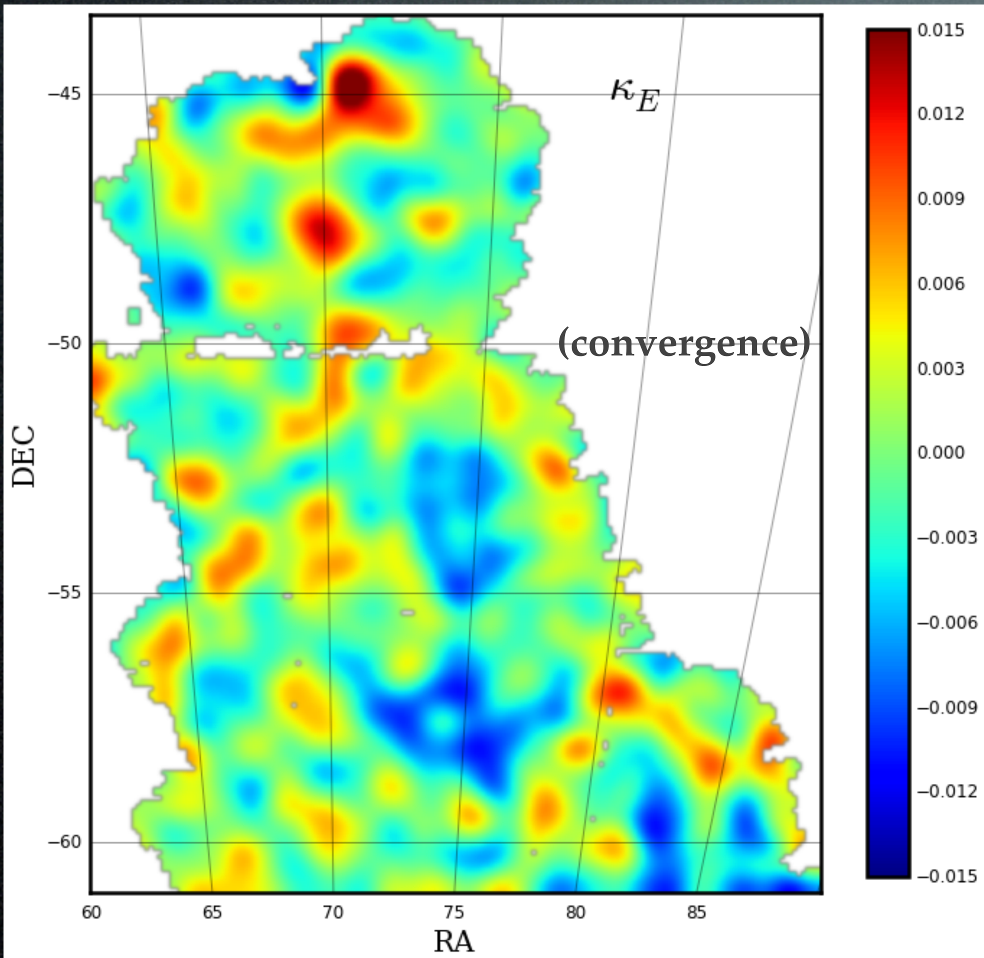
Leistedt et al 2015, see also Bonnett et al 2015

- Testing photo-z codes on realistic simulations (processed through images, back to catalogs, through systematics maps), where we know the underlying redshift distribution

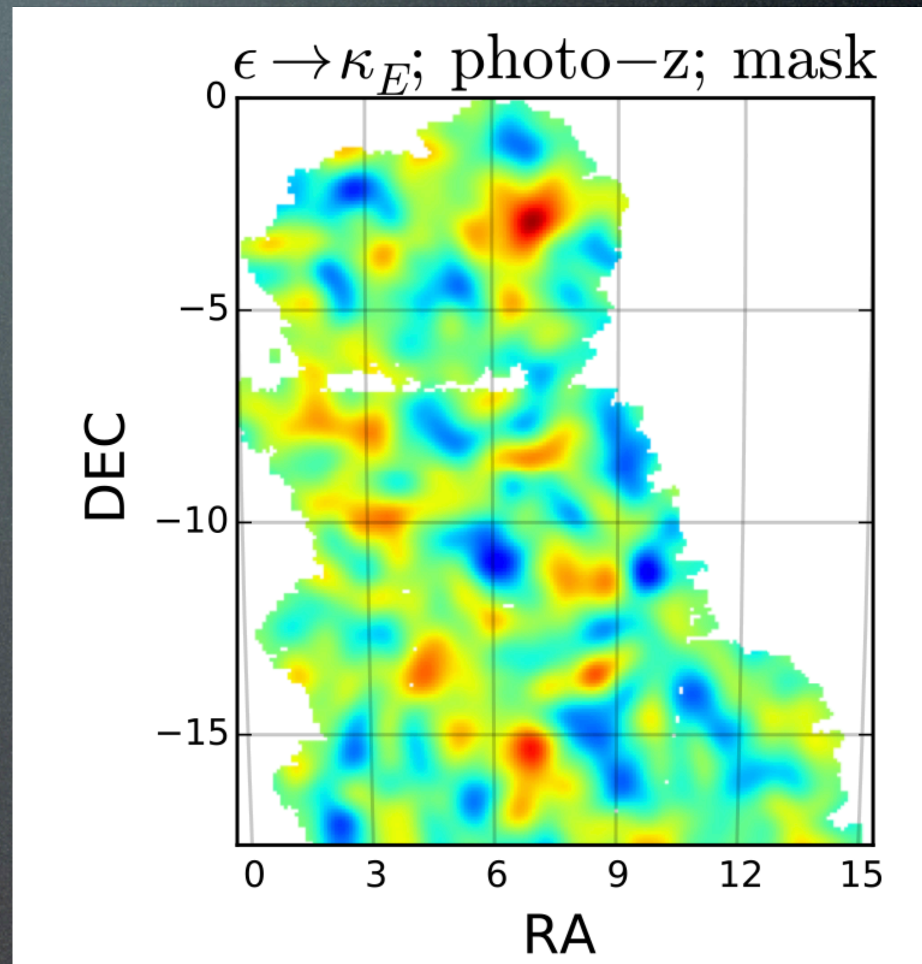
- Can do this for other systematics as well -- seeing, sky background, etc...



Mapping the Mass



Map mass in the Universe at z < 0.5 over140 sq. degrees from weak lensing shear



"Observed" mass in the DES BCC Simulations

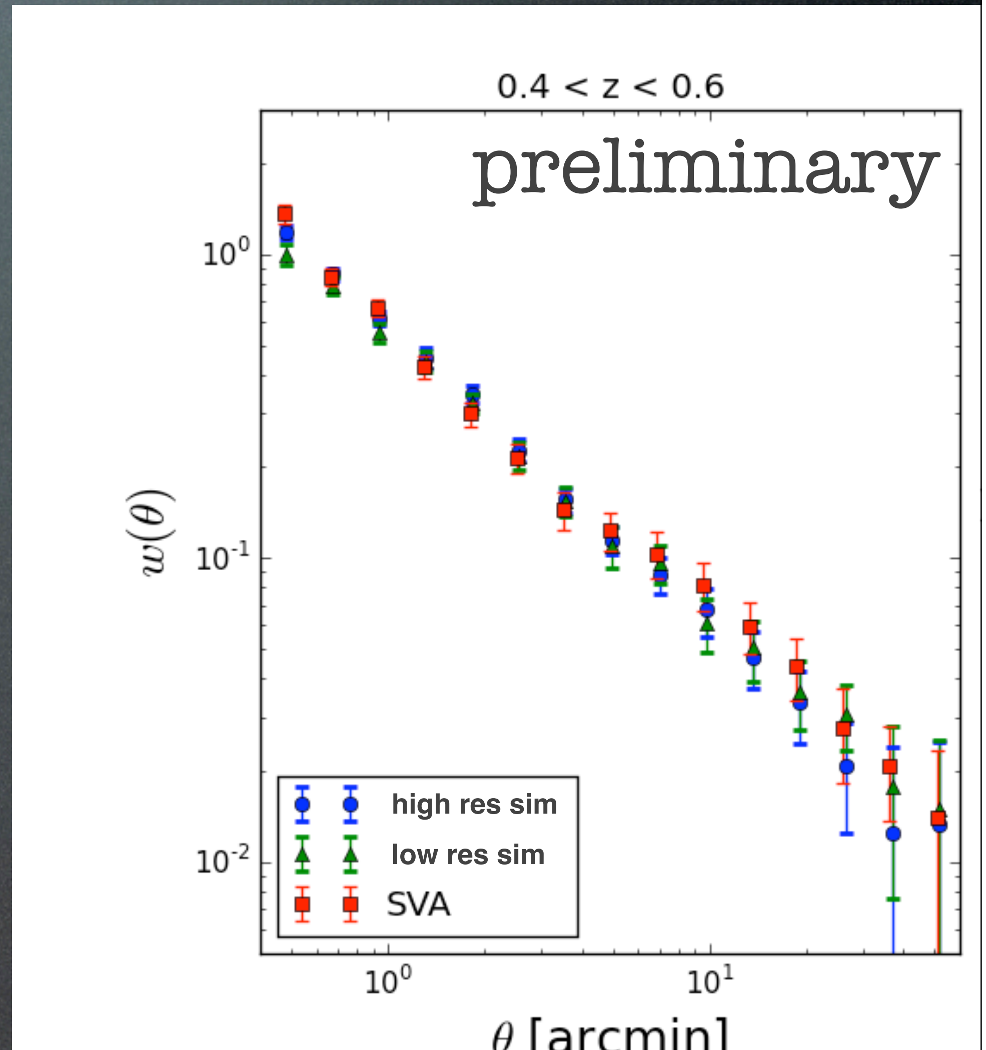
Vikram, Chang & the DES Collaboration 1504.03002

Clustering of red galaxies

Simulations allow for testing of algorithms and
detailed tests of modeling approaches

Redmapper galaxy sample (Rozo, Rykoff et al arxiv/ 1507.05460, see also Rozo talk tomorrow)

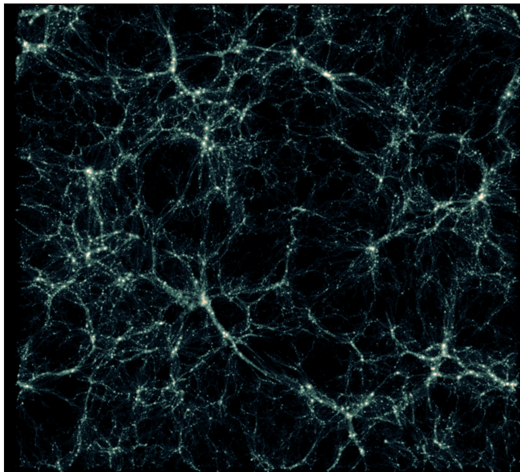
Simulations used to test photo-z's, redshift distribution, understand outliers (allowed us to detect an issue with star galaxy separation!),
test details of halo occupation models



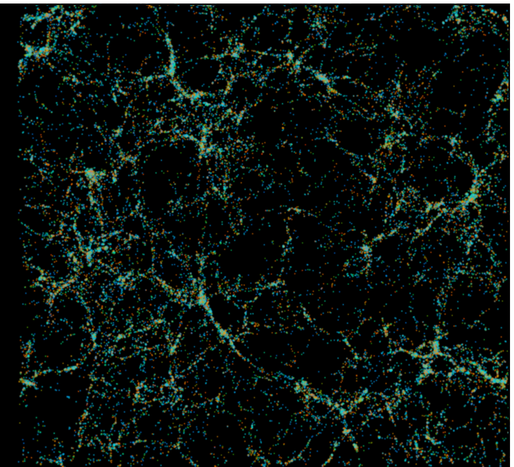
Krause, DeRose, RW
+ DES collaboration in prep

DES Blind Cosmology Challenge

- Goal: Run key cosmology analyses on simulations. Recover true cosmology and understand biases, before running on the data.
- Essential for testing analysis pipelines. Also provides very useful feedback on the simulations!
- Where possible don't unblind analysis on data until one has shown correct recovery in simulations
- Staged challenge; run on one with known cosmology, then try another unknown cosmology
- Six full skies in different cosmologies in hand
- Working towards simulating 300 DES Y1 skies to use for data analysis and covariances



Summary



- extracting accurate parameters of interest out of cosmological surveys increasingly requires simulations in many aspects of the analysis
    - predictions of the mean signal
    - covariances between observables
    - understanding systematics in data and analysis
- data are already good enough to start distinguishing between complex details of galaxy-halo connection
- need a hybrid of approaches to make accurate models that are large enough for surveys
- theorists have a lot of work to do to keep up with the data!