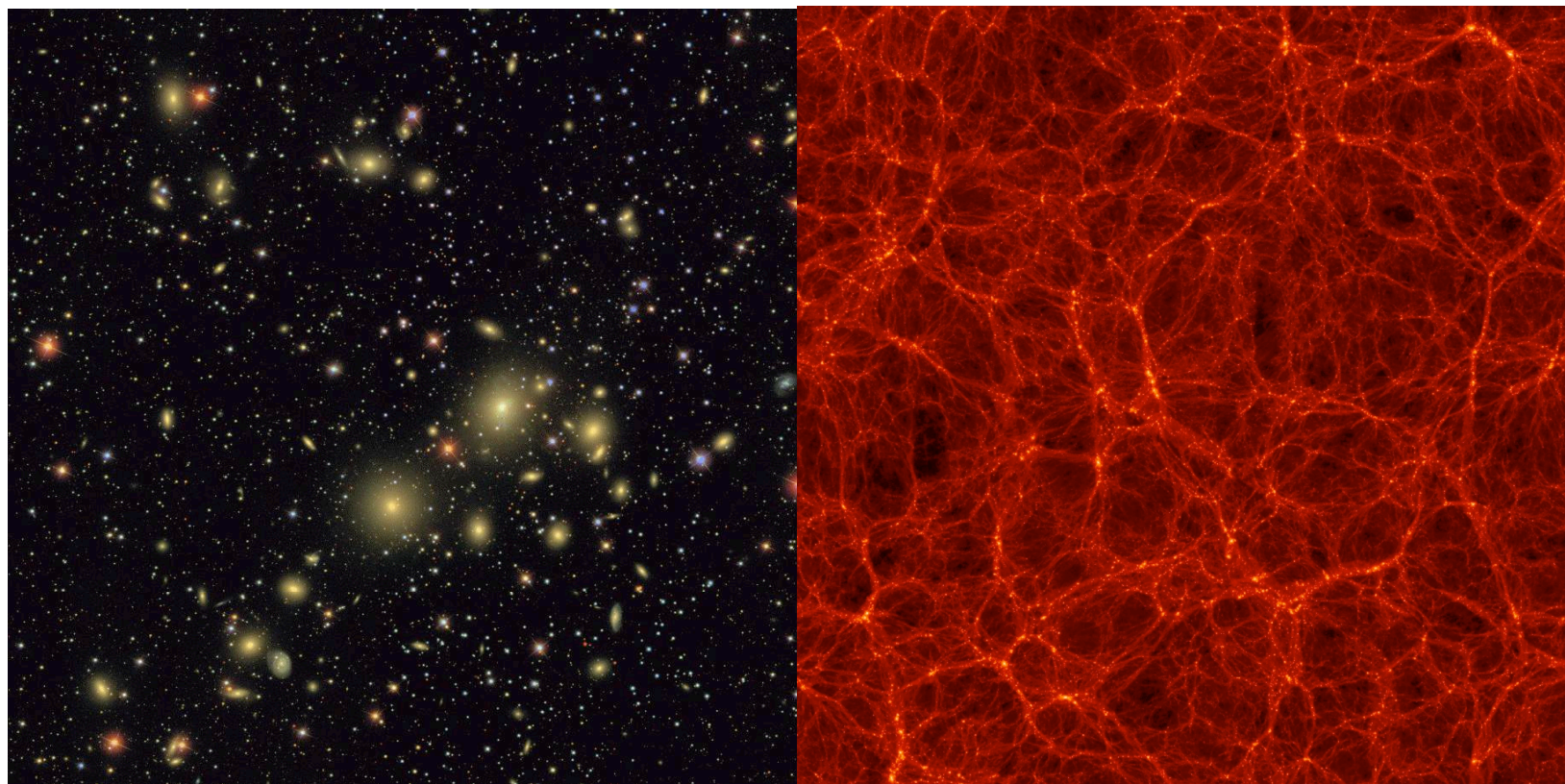


Simulation Challenges

In the Era of Surveys of 10 Billion Galaxies

Risa Wechsler



Futures of Astrophysical
Computing
December 16, 2010

Scale of the problem: data

■ Largest survey to date

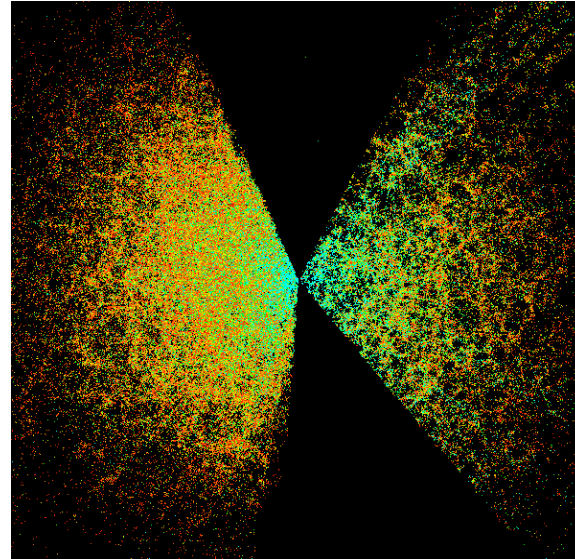
- Sloan Digital Sky Survey
- 200 million objects over 1/4 of the sky

■ Deepest survey to date

- Hubble Ultra Deep Field: 11 days on HST; 8 magnitudes deeper
- 10,000 galaxies over 1/13 millionth of the sky
- implies ~100 billion galaxies to this depth

■ The next decade

- few sq. deg -- deep, multi-wavelength data e.g. CANDELS
- Dark Energy Survey (2011-2016): 1/8 of the sky, 2.5 magnitudes deeper than SDSS. ~ 300 million galaxies
- PanStarrs; WFIRST; Euclid
- LSST (2017-2027): 1/2 the sky, 5 magnitudes deeper than SDSS. ~ 10 billion galaxies. Image each piece of sky every 3 nights; 30 TB/night, ~100 PB over 10 years.



side note: info in the human genome
unreduced: 10 GB per person
reduced (differences from reference
genome): 20 MB
--> 150 PB for every person on earth

■ Cosmological probes:

■ CMB

■ SN

■ Structure formation:

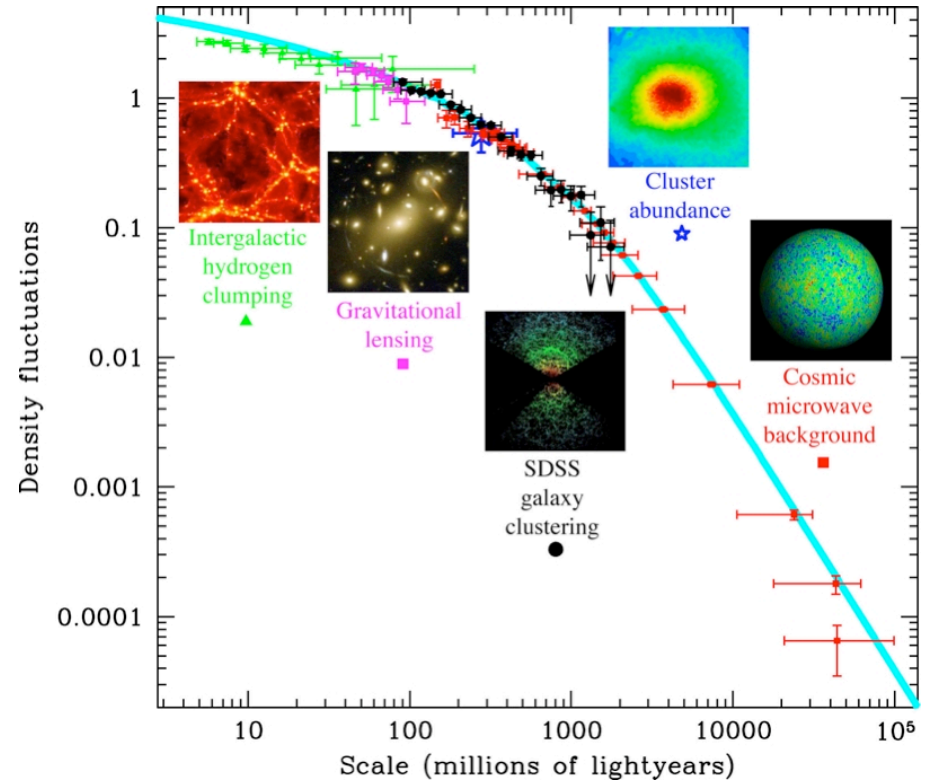
■ gravitational lensing

■ galaxy clusters

■ galaxy clustering

■ Ly-alpha forest

■ 21-cm



these are all based on structure formation simulations

What are simulations for in interpreting these data?

■ Cosmological parameters & Dark Energy

■ main cosmological probes already are or soon will be in the systematics dominated regime

- theory systematics: need to get from ~ 7 parameters specifying the cosmological model to better than 1% predictions for structure formation

- observational systematics: e.g. star galaxy separation, photometry, cluster miscentering

■ several related issues

- precise predictions for a variety of structure formation probes

- development and verification of science ready codes to work on large volumes

- understanding the instrument

- understanding observational systematics

- covariance matrices to determine error bars (e.g. Schneider), needed not just for one measurement, but for many (e.g.: lensing, galaxy clustering, galaxy clusters)

- impact of galaxy formation & galaxy selection (type dependent bias)

Simulation issues for the dark side

■ Dark Energy

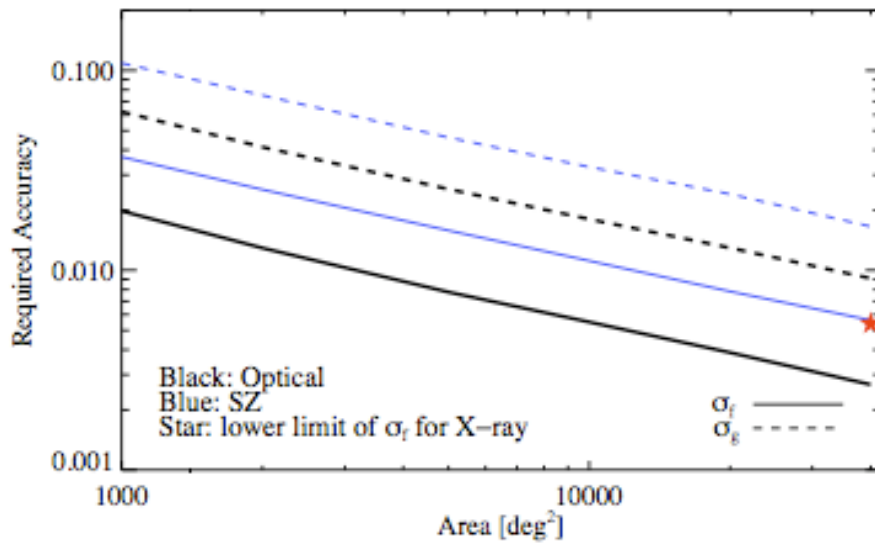
- BAO: 100 Mpc scales; need $\sim 0.5\%$ precision on position of the peaks (impacted by scale-dependent bias)
- Weak Lensing: need to understand the impact of baryonic physics on the DM power spectrum
- Clusters: need $\sim 0.5\%$ prediction for mass function given cosmological parameters; detailed understanding of the mass-observable relation for various observables
- Galaxy Clustering: need detailed understanding of scale-dependent, type-dependent bias on all scales

■ Dark Matter

- indirect detection: need detailed understanding of the inner regions of our Galaxy to distinguish astrophysical signals
- direct detection: need detailed understanding of the full phase space distribution of dark matter in our Galaxy to interpret any detections / limits as particle properties
- measuring neutrino mass; distinguishing CDM from warm DM: precise predictions for galaxy & matter PS as above, precise predictions for substructure

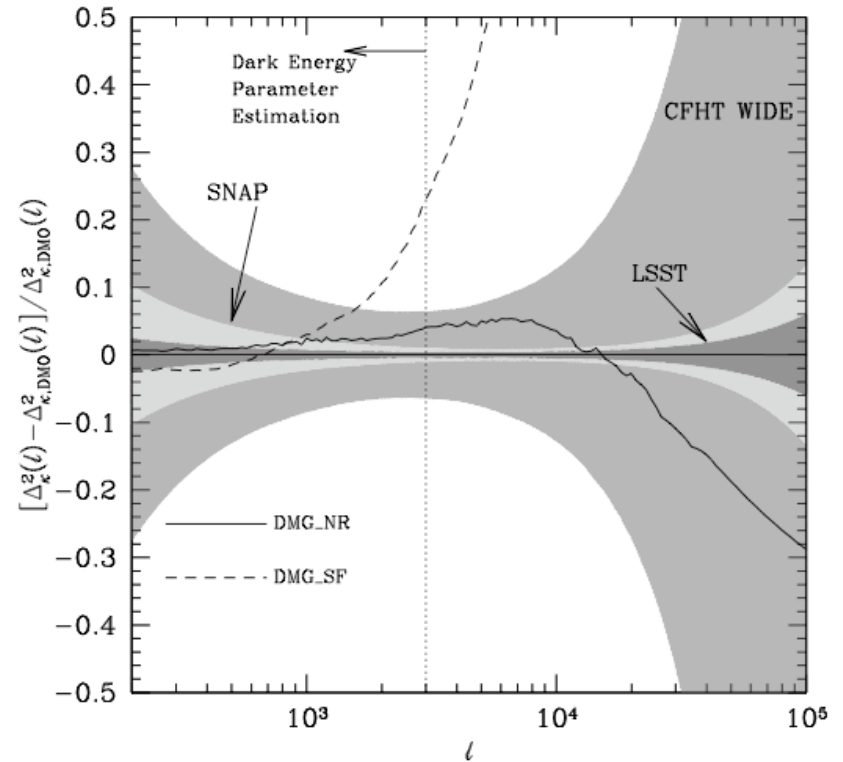
Few Examples

Wu, Zentner & Wechsler 2010



need to know the number of dark matter halos to better than 1%

Rudd, Zentner & Kravtsov 2008

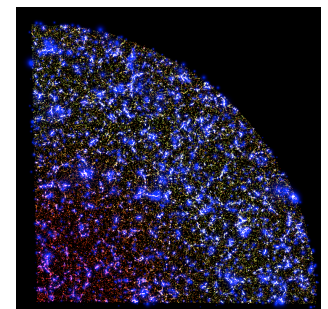
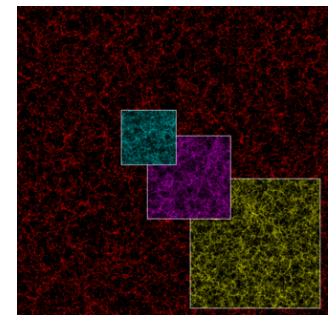
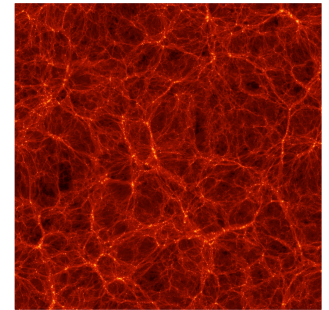


uncertain impact of baryons in the matter power spectrum dominates error budget

Simulation needs to understand SDSS

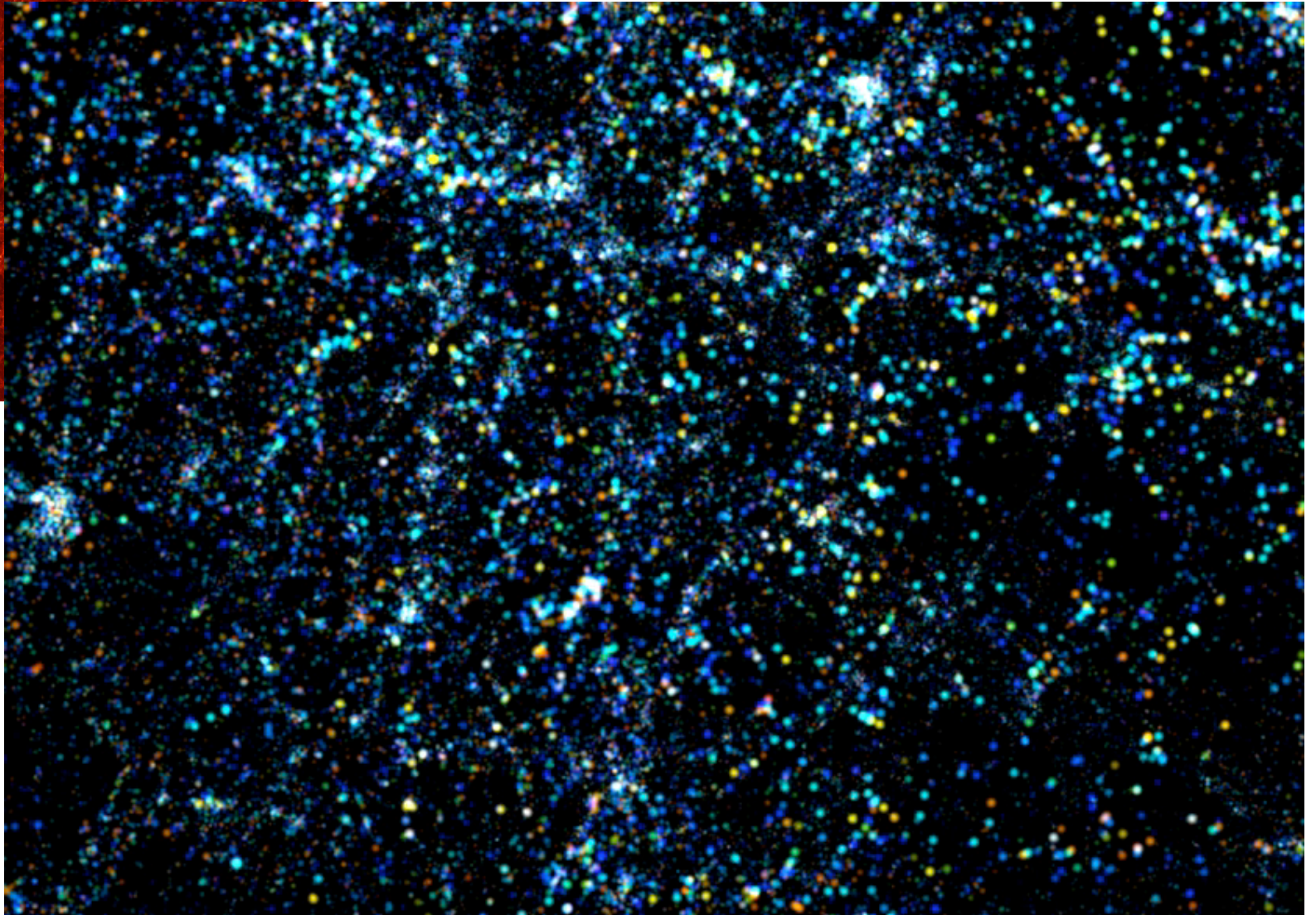
depends on science goals / galaxy assignment scheme

- if you want to predict clustering of satellites, connect every galaxy to a dark matter substructure, track merging history
- subhalo abundance matching (SHAM)
- Bolshoi: allows us to model down to SMC masses, histories for MWs
1 kpc force res, $1e8$ Msun mass res.
- if you want to associate every galaxy with a dark matter host halo (HOD modeling; assume you want to resolve every halo that hosts any galaxy with 100 particles)
- LASDAMAS project / 100 SDSS surveys
- if you want to really push your resolution and just match the clustering, can connect galaxies to dm density field
- ADDGALS: adding density determined galaxies to lightcone simulations (Wechsler et al 2004, 2011, Busha & Wechsler et al 2011)

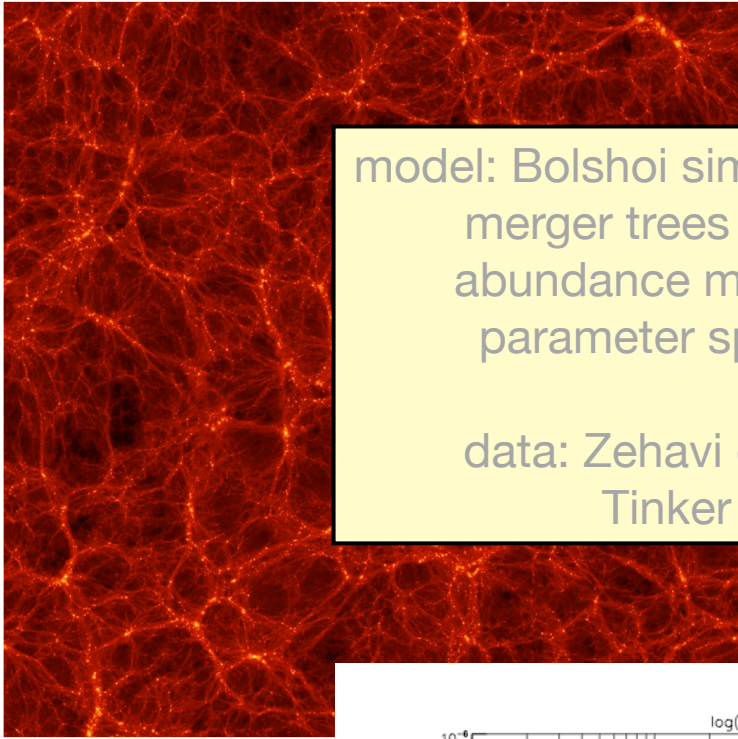


360 Mpc

model: Bolshoi simulation (Klypin et al 2010)
~ 7 million CPU hours; 1 kpc force res; $M_p = 1.3e8 M_{\text{sun}}$
merger trees (Behroozi et al 2011)
abundance matching to assign galaxies to halos

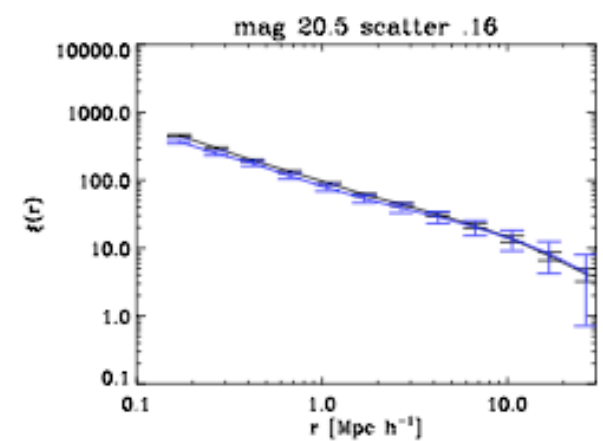
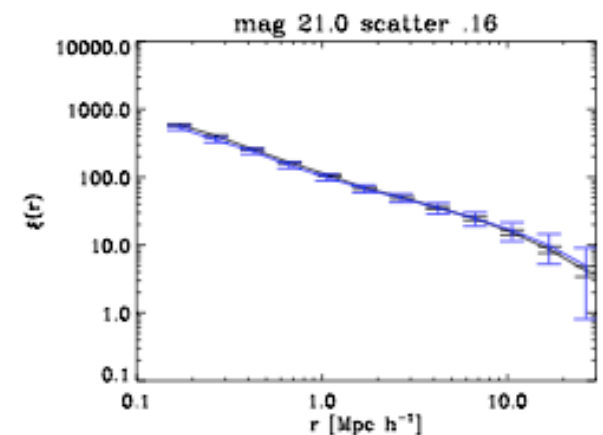
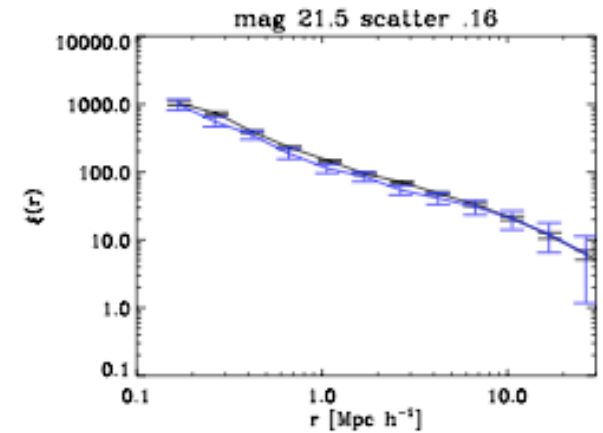
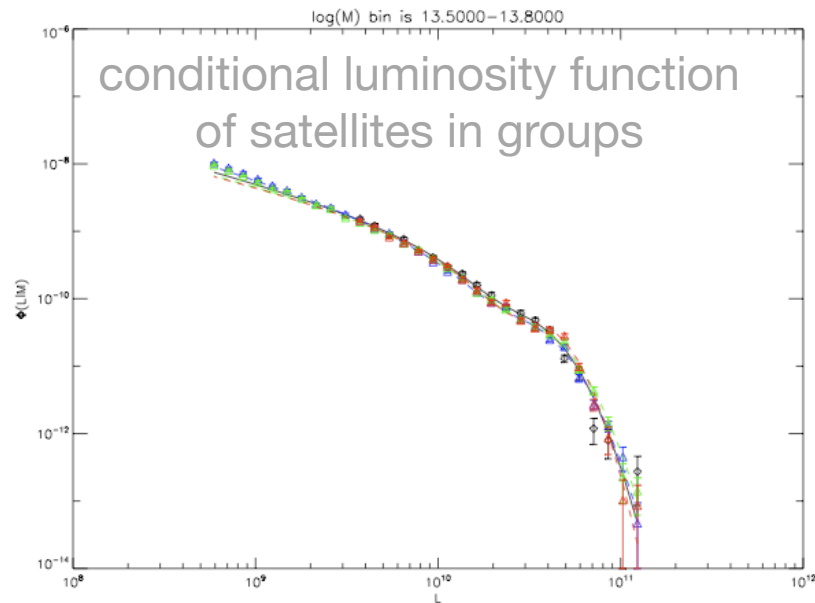


vis by Ralf Kaehler

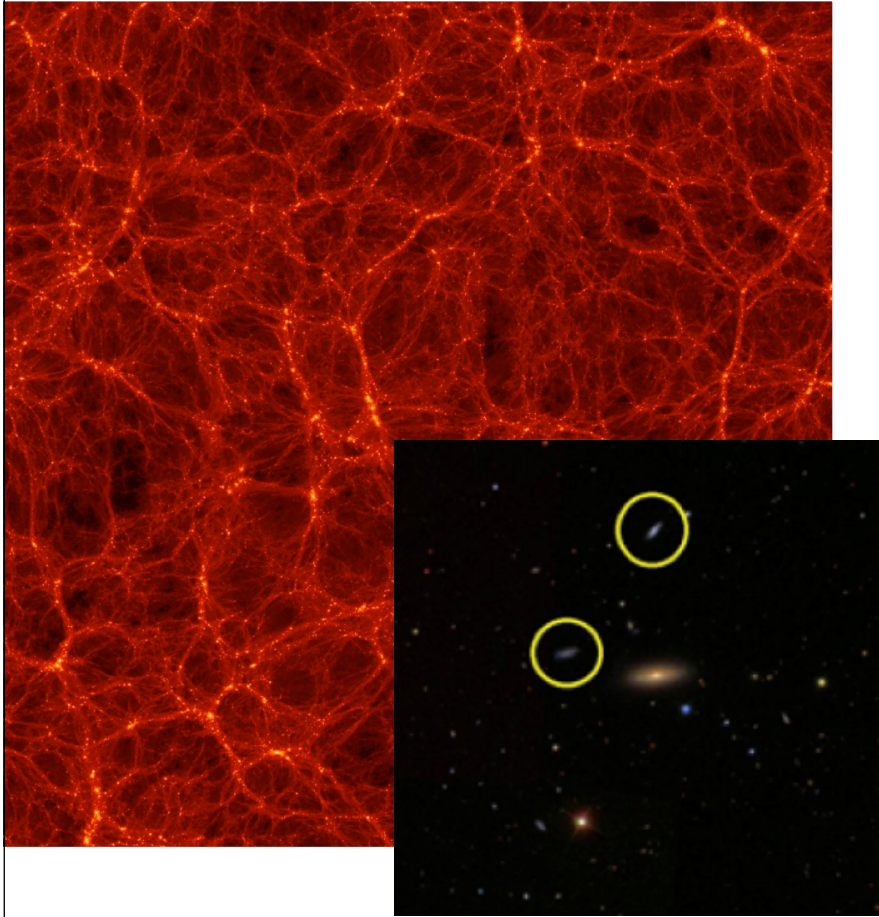


model: Bolshoi simulation (Klypin et al 2010)
merger trees (Behroozi et al 2011)
abundance matching with one free
parameter specifying the scatter

data: Zehavi et al 2010 clustering;
Tinker group catalog



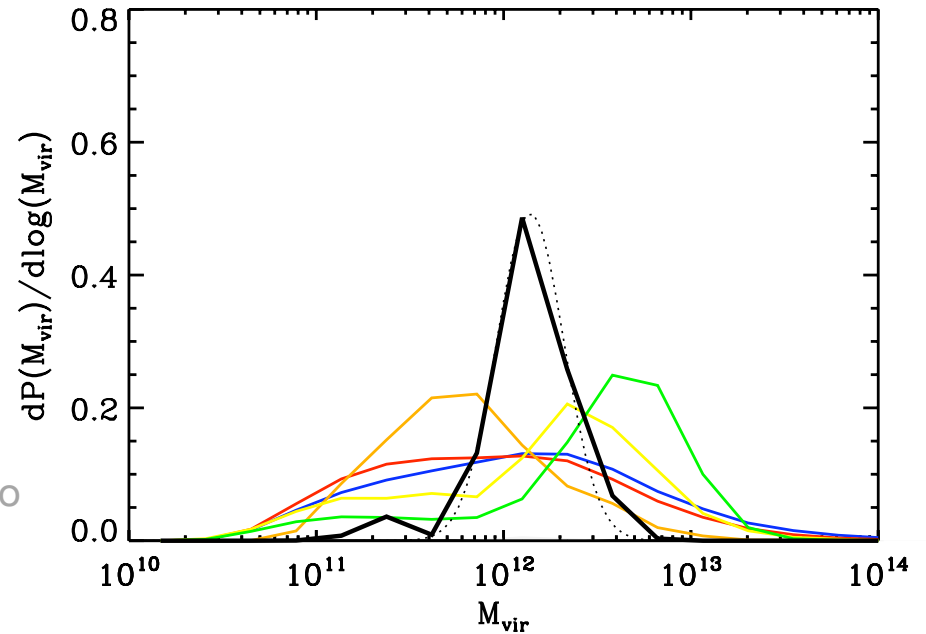
Understanding the Cosmological Context of the MW



How to understand the detailed properties (mass distribution, formation history) of the one system where we have the most detailed measurements?

If you have a simulation which is large enough to be cosmologically representative, you can treat the set of all halos as a prior for the properties of the system

Here: used dynamical properties of the LMC and SMC to determine the MW mass & assembly history (weight each halo by likelihood it has an LMC & SMC)
Future: want to use more properties, including fainter satellites / demanding resolution and volume challenge



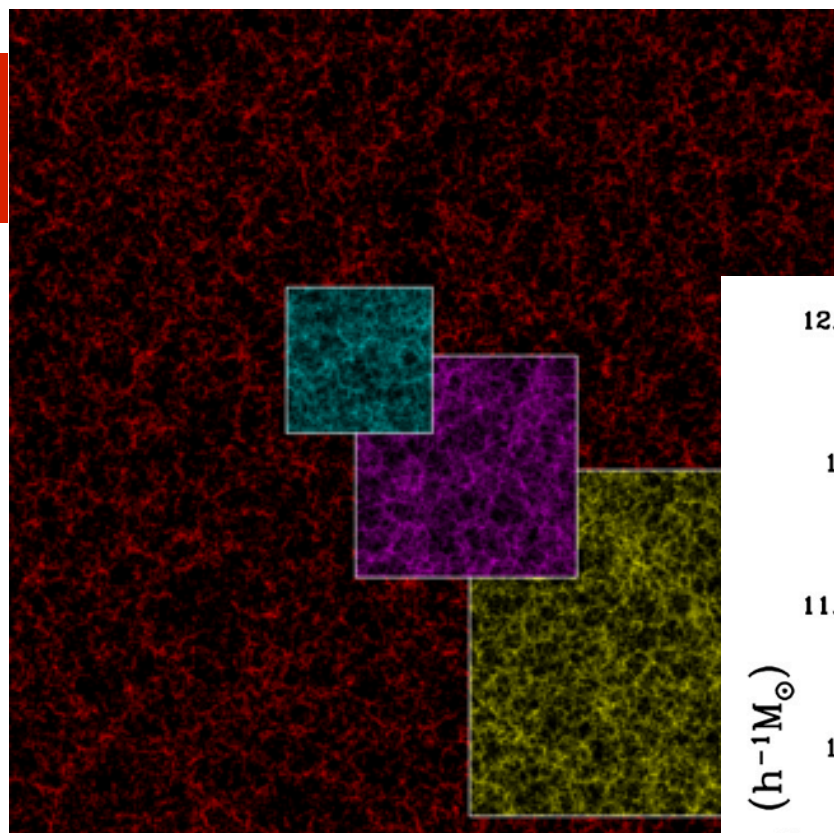
Busha et al 2011 (arXiv:1011.2203)

merger tree viz: Peter Behroozi

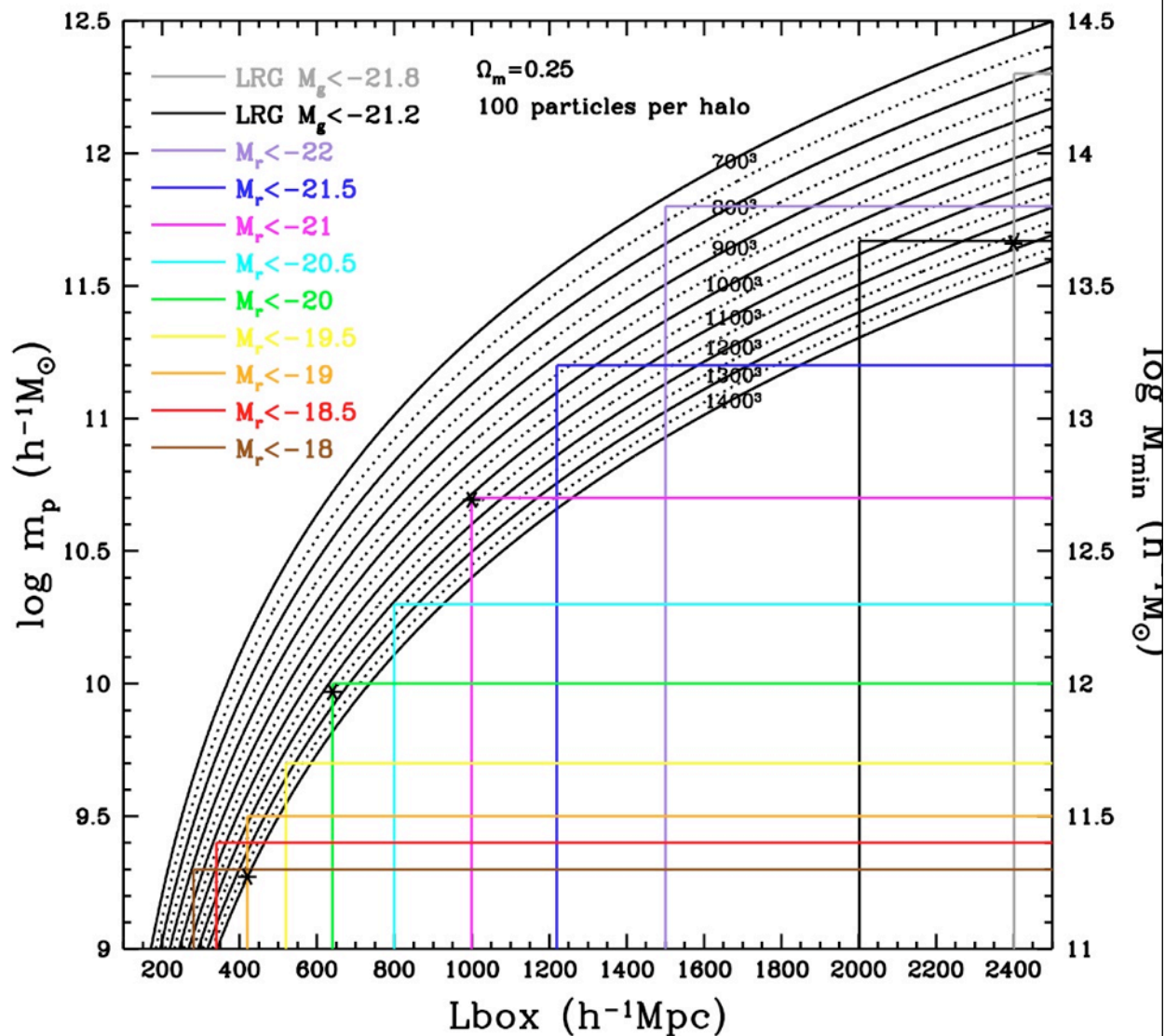


merger trees (~ 250 GB) and various galaxy catalogs now available
need database tools to make them public & useable by wide community

We only have one universe to observe... need many realizations to get error bars right



also need many additional realizations of different cosmological models



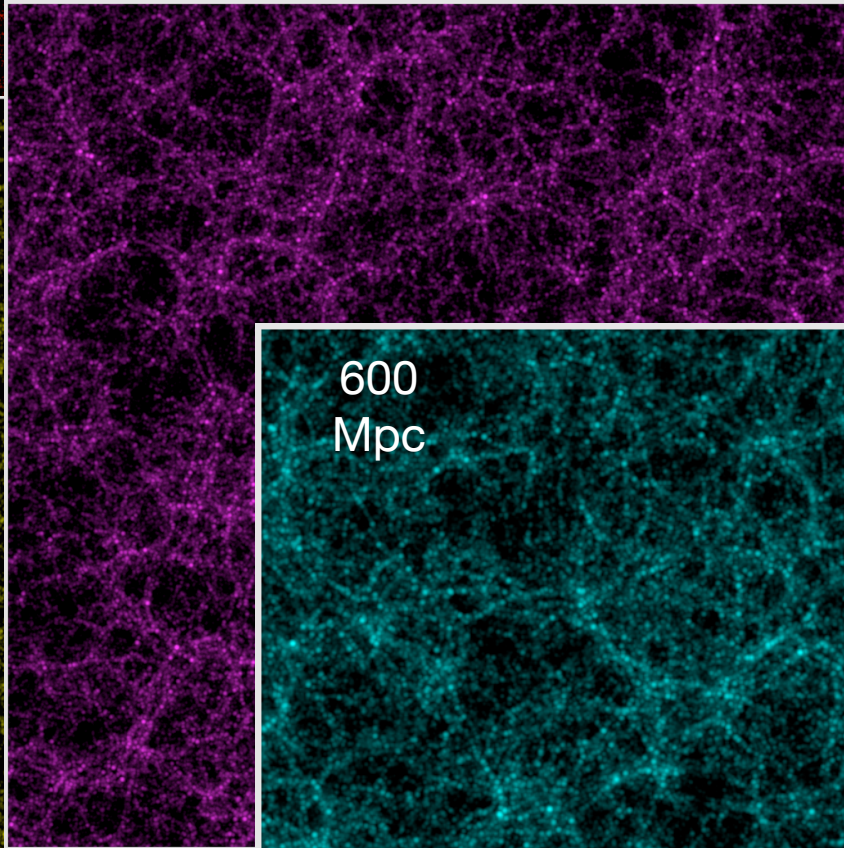
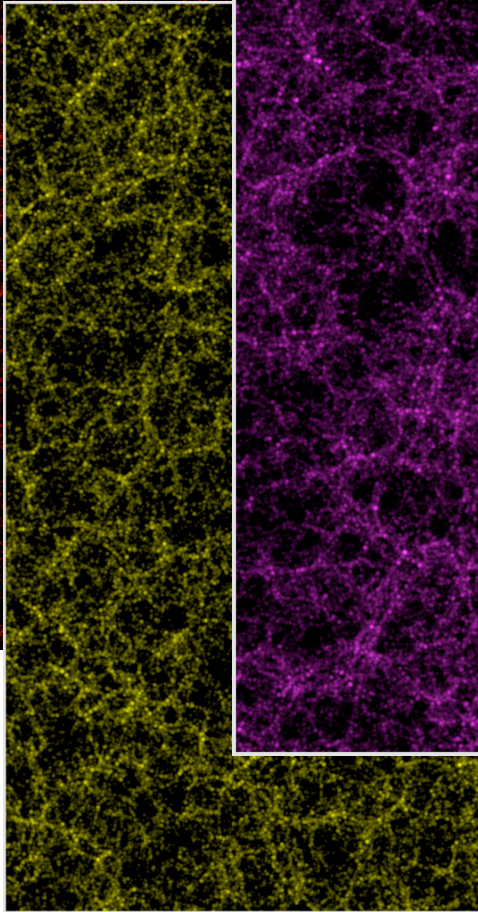
LASDAMAS: LargeSuite of DArk MATter Simulations

McBride & Berlind (Vanderbilt)
Busha & Wechsler (Stanford)
Scoccimarro & Manera (NYU)
van den Bosch (Yale)

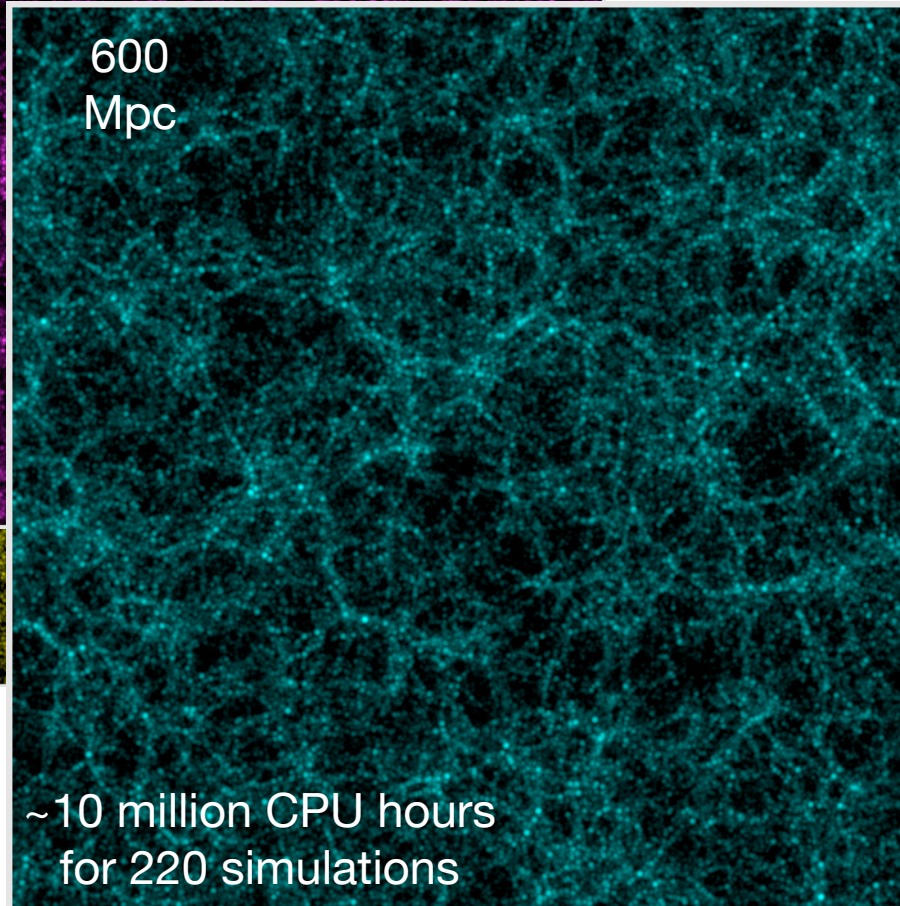


mocks publicly available

2.8
Gpc



600
Mpc



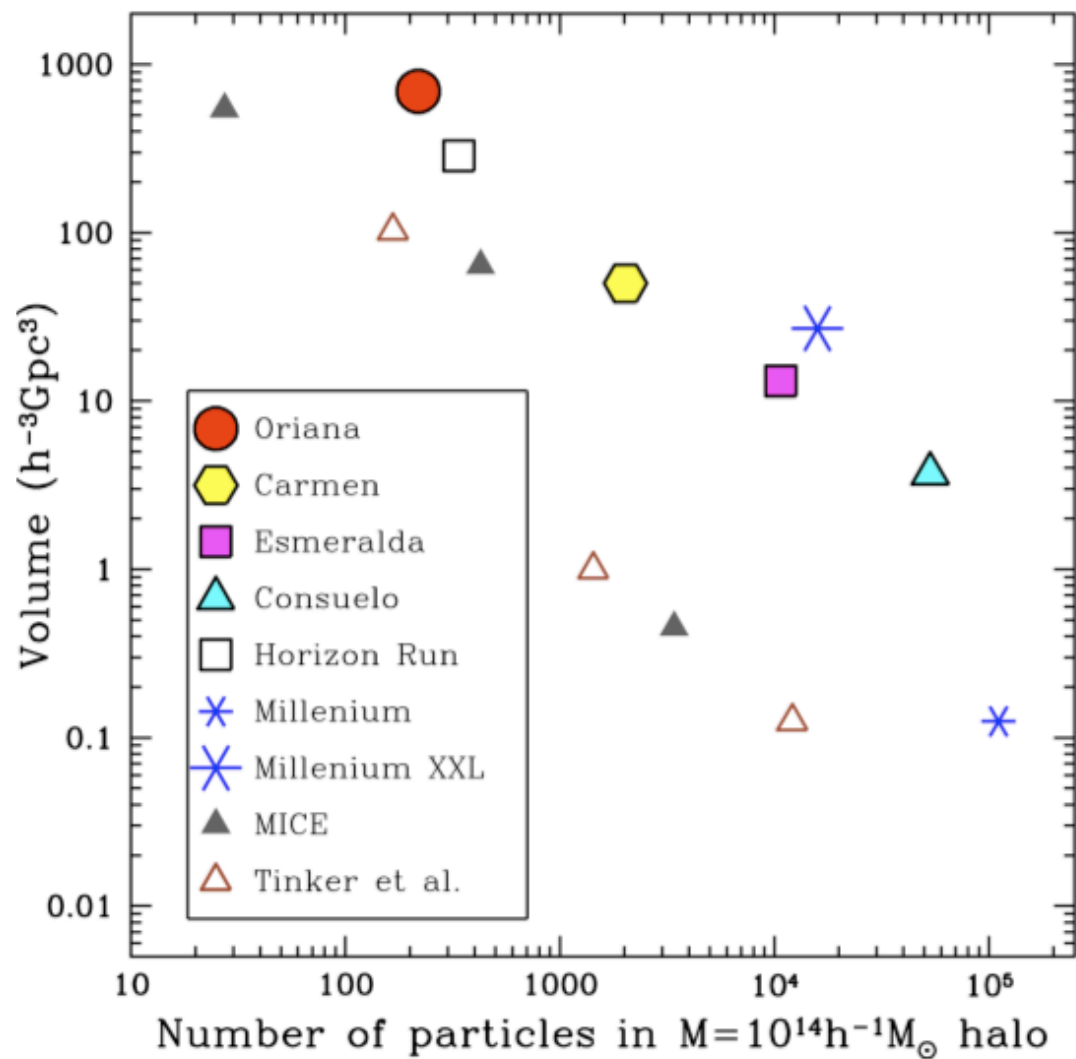
~10 million CPU hours
for 220 simulations

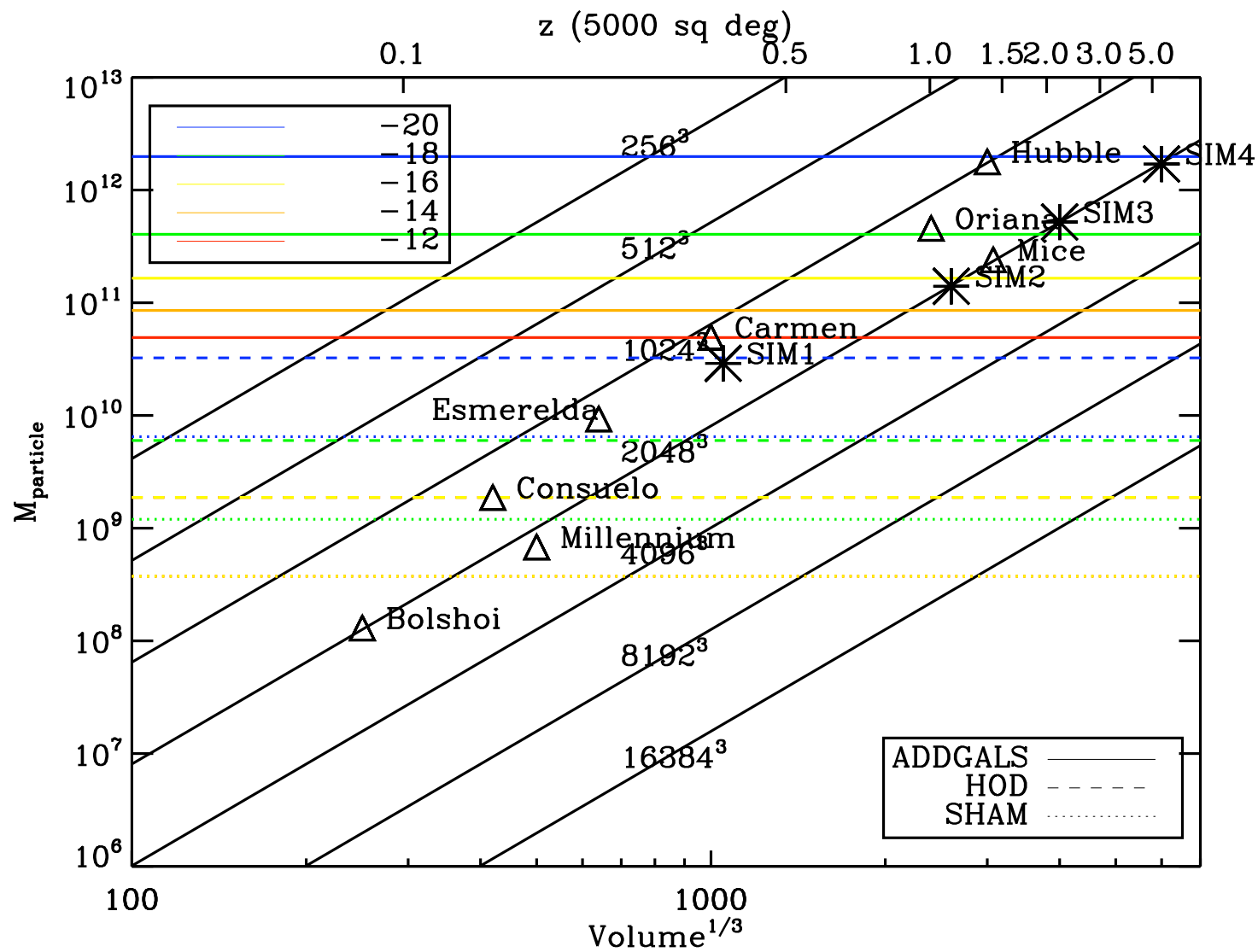
upcoming papers

McBride et al 2011 / LASDAMAS mocks
McBride et al 2011 / Mass function

total volume probed is ~ 1000 Gpc³

Current largest DM simulations

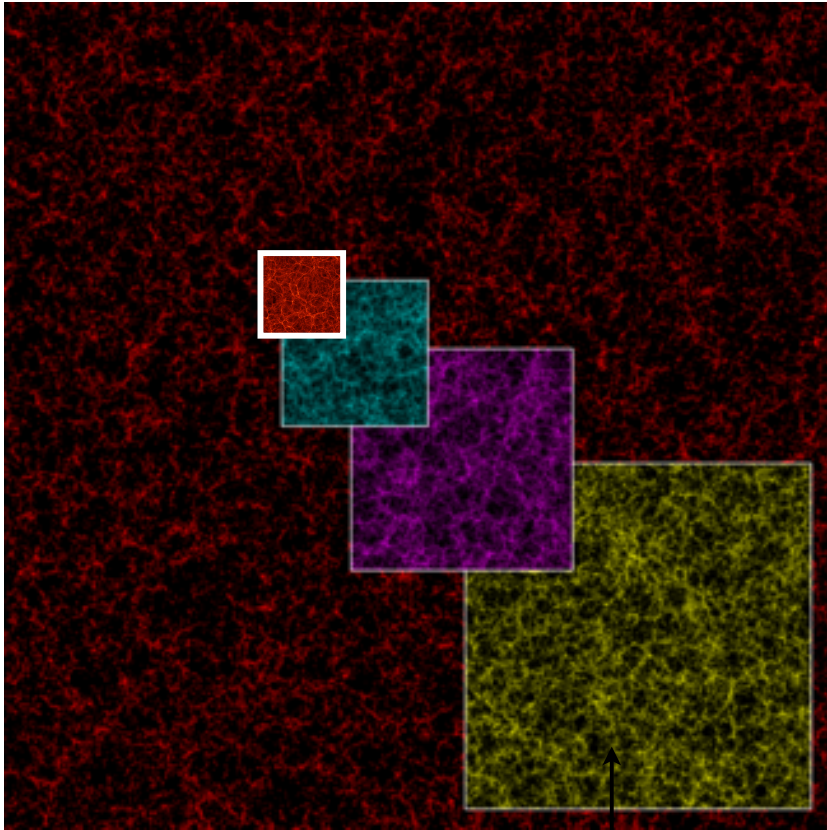




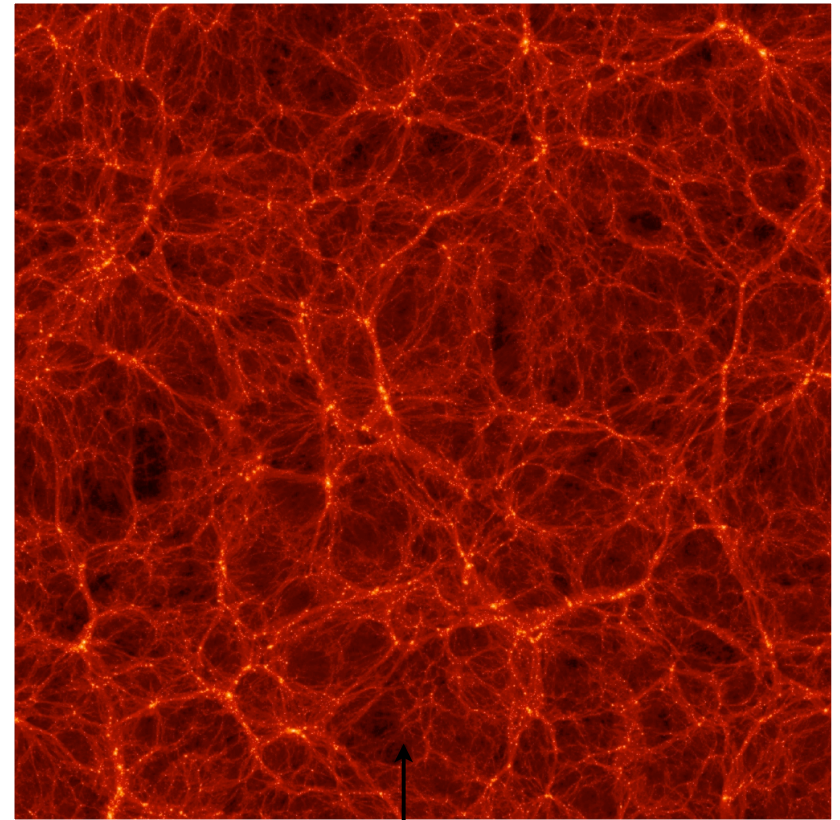
Data is already harder than flops

- In many cases the post-processing is harder than the simulation
 - (more ambiguities, more data challenges)
- need to find halos, see how they build up over time, understand how they are connected to galaxies...
- examples:
 - Bolshoi: 8 billion particle simulation x 200 snapshots
 - ~ 300 TB for one box
 - 8 million halos
 - hard to manipulate all particles in halos, so we often don't use all the information
 - Consuelo: 7TB for [100 snapshots of] one simulation
 - 300 TB for current runs if we saved everything so we preprocess (find halos) and then delete.
- compression seems reasonable and is necessary but:
 - halo finding is actually not a solved problem
 - additional statistics that require the dark matter particles (e.g. lensing) and we don't always know the details ahead of time.

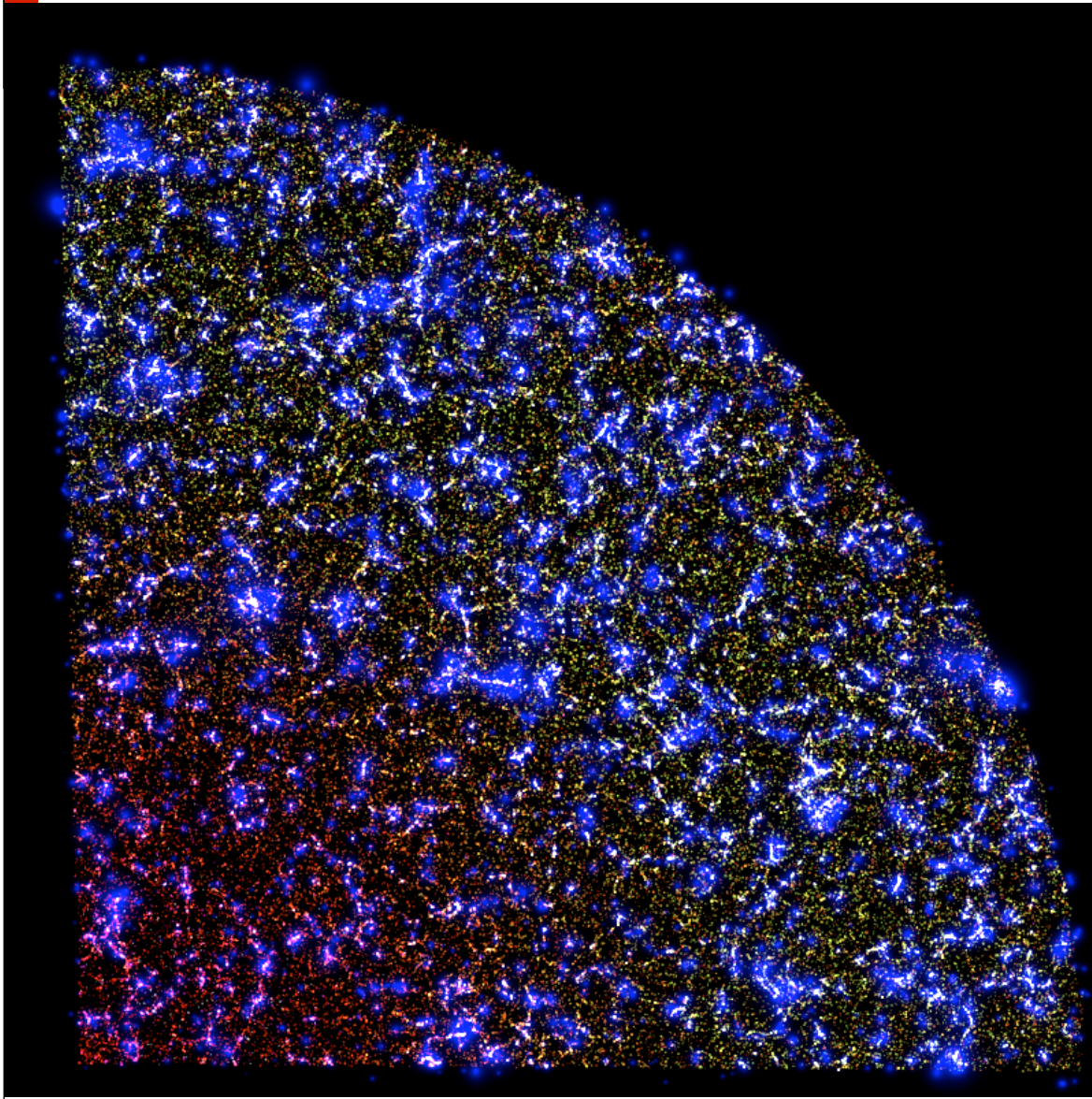
What do we need in the next decade just for one realization of the observed universe?



this one gets you 1/2 % of sky out to $z=1.3$
(or 1/8 of sky out to $z=0.3$)



this is ~ the resolution you want



with upcoming surveys,
both wide and deep:

simulate the whole thing first, before
you have the data

we are doing this now for both wide
surveys (DES)
and deep surveys (CANDELS)
+ wide/deep (LSST)

go through whole pipeline:
dm, galaxies, lensing, images,
galaxy catalogs, inferred parameters

especially for wide surveys, helps
build expertise with relevant data
volumes



DARK ENERGY
SURVEY

A Blind Cosmology Challenge for the DES

with lots of collaborators including

Michael Busha (Stanford --> Zurich)

Gus Evrard

Andrey Kravtsov

Brandon Erickson

Molly Swanson

Matt Becker

Joerg Deitrich

Huan Lin

Basilio Santiago

Nacho Sevilla

Eduardo Rozo

+ many, many folks who will do analysis!



DARK ENERGY
SURVEY

BCC: the basic idea

- Develop ability to run our codes, particularly core cosmology analyses, on 5000 sq. degrees of data that looks as much like the DES data as possible, before we are inundated with data.
- Convince ourselves and the community that we can recover key cosmological parameters from such data, and understand at what precision this can be done including the full range of observational systematics.

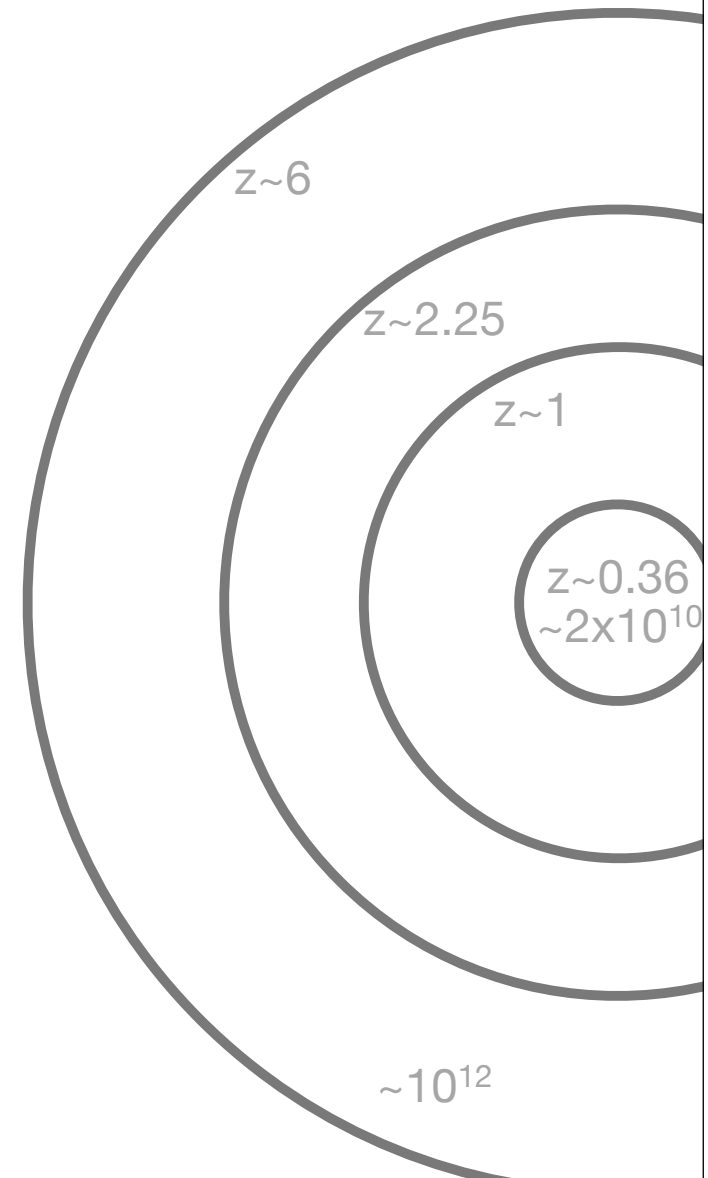
- ★ **Test ability of codes to run on full data**
- ★ **Test robustness and accuracy of cosmology codes**
- ★ **Assess realistic systematic errors for cosmology analyses**
- ★ **Assess spectroscopic followup plans**
- ★ **Accurately assess computational needs for analysis**
- ★ **Test out new ideas for analysis**
- ★ **Do fun stuff before we have data!**
- ★ **Convince community that what we know what we are doing**



DARK ENERGY
SURVEY

BCC simulation basics

- simulate a series of cosmologies, unknown to SWGs
- 5000 sq. degrees
- N-body lightcones to $z \sim 6$, constructed using 4 simulation boxes of varying resolution (2048^3 particles, few 10^{10} to 10^{12})
- galaxies with multiband photometry to full DES depth (including all sources with 10 sigma in any band)
- apply observational transfer function to including many observational effects without running through image simulator
- explore a range of dark energy models and other alternative cosmologies
- first cosmology done, on NSF Teragrid (Kravtsov, Evrard, Wechsler, Busha)
- total resources $\sim 650\text{K}$ CPU hours; 2-40 TB per run.





DARK ENERGY
SURVEY

BCC simulation pipeline

1. **Decide on set of cosmological models** (Busha, Wechsler, Kravtsov, Evrard, Lahav)
2. Initial conditions, run simulation, output light cone, run halo finder, validate (Busha, Erickson)
3. Add galaxies (Busha, Wechsler)
4. Run validation tests (Hansen, Busha, Wechsler, others)
5. Calculate shear at all galaxy positions (Becker)
6. Add shapes, lens (magnify & distort) galaxies (Dietrich)
7. **Add stars** (Santiago)
8. **Determine mask** (Swanson), including varying photometric depth & seeing, foreground stars
9. Determine photometric errors (Lin, Busha), **incorporating mask information**
10. **Misclassify stars and galaxies** (Sevilla, Hansen, Santiago)
11. **Blend galaxies** (Hansen)
12. Determine photometric redshifts (Busha, Cunha, Gerdes, etc)
13. Provide a lensed galaxy catalog in the Brazil portal with:
ra, dec, mags, magerrors, photoz's, $p(z)$, size, ellipticity, **star/galaxy probability, seeing**

★ **grey steps already implemented and in use (over 220 sq. degrees)**

★ **Science working groups do analysis!**

A few thoughts...

■ Data data data

- sometimes the data (both real & simulated) can be compressed, but only if you already know all of the questions you want to ask.
- we are still learning. even on the theory side compressed data is challenging.
- want to publish simulations and make data accessible to wide range of users
- new kinds of problems and inference

■ Astrophysics is not one problem (need more than one kind of computer)



- N-body simulations // simulation analysis (implications for moving data around)
- detailed simulations of star formation & galaxy formation
- simulations relativistic jets, SN, etc.
- exploring cosmological model parameter space (e.g. MCMC)

■ A good fraction of computing is not done in the efficiency limit

- hardware is often cheaper *and easier to get funding for* than developers
- structural problem: generally this work is done by people (students/postdocs) who need to find a new job in 1-2 yrs

■ Already limited by systematic uncertainties in many regimes

- no ab initio models can explain basic statistical properties of galaxies
- both theory systematics (e.g. mass function, impact of baryons) & observational systematics (which can sometimes only be solved by simulations)

- 
- 
- Computational challenge for getting the science out of next generation surveys is large.
 - Need for a more coordinated effort that integrates hardware, software development, data curation and dissemination training of developers and users.
 - Need for more collaboration within the field and with other experts in computational science.