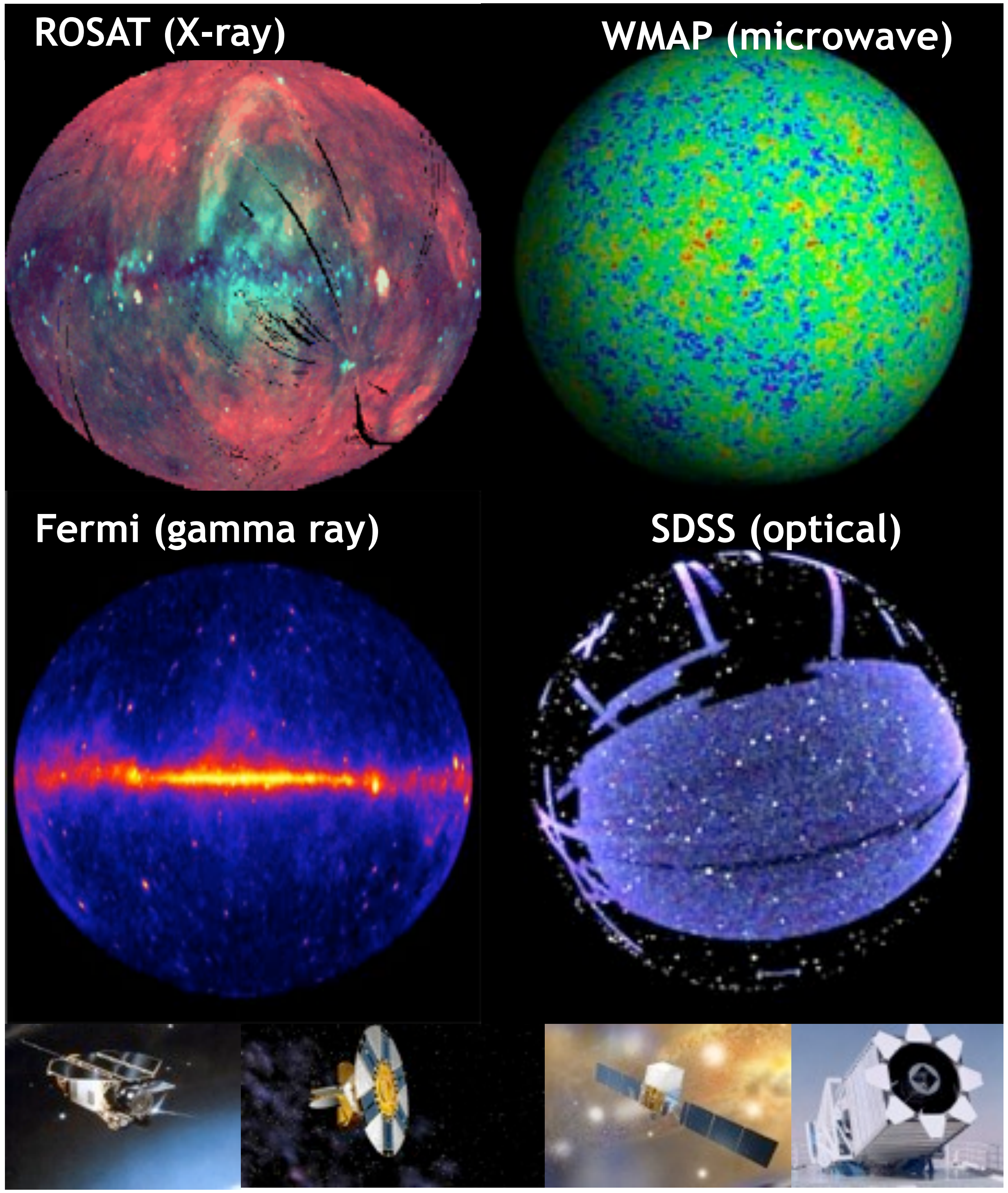
Exploring the Dark Universe: Statistical and Data Challenges

Katrin Heitmann
High Energy Physics Division
Mathematics & Computer Science Division
Argonne National Laboratory

Collaborators:
J. Ahrens, U. Alam, D. Daniel, P. Fasel, H. Finkel,
N. Frontiere, S. Habib, D. Higdon, T. Holsclaw, H. Lee,
E. Lawrence, Z. Lukic, C. Nakhleh, A. Pope, B. Sanso,
C. Wagner, M. White, B. Williams, J. Woodring,
and the ANL visualization team

Modern Cosmology and Sky Maps

- Modern cosmology is the story of mapping the sky in multiple wavebands
- Maps cover measurements of objects (stars, galaxies) and fields (temperature)
- Maps can be large (Sloan Digital Sky Survey has~200 million galaxies, many billions for planned surveys)
- Statistical analysis of sky maps
- All precision cosmological analyses constitute a statistical inverse problem: from sky maps to scientific inference
- Therefore: No cosmology without (large-scale) computing



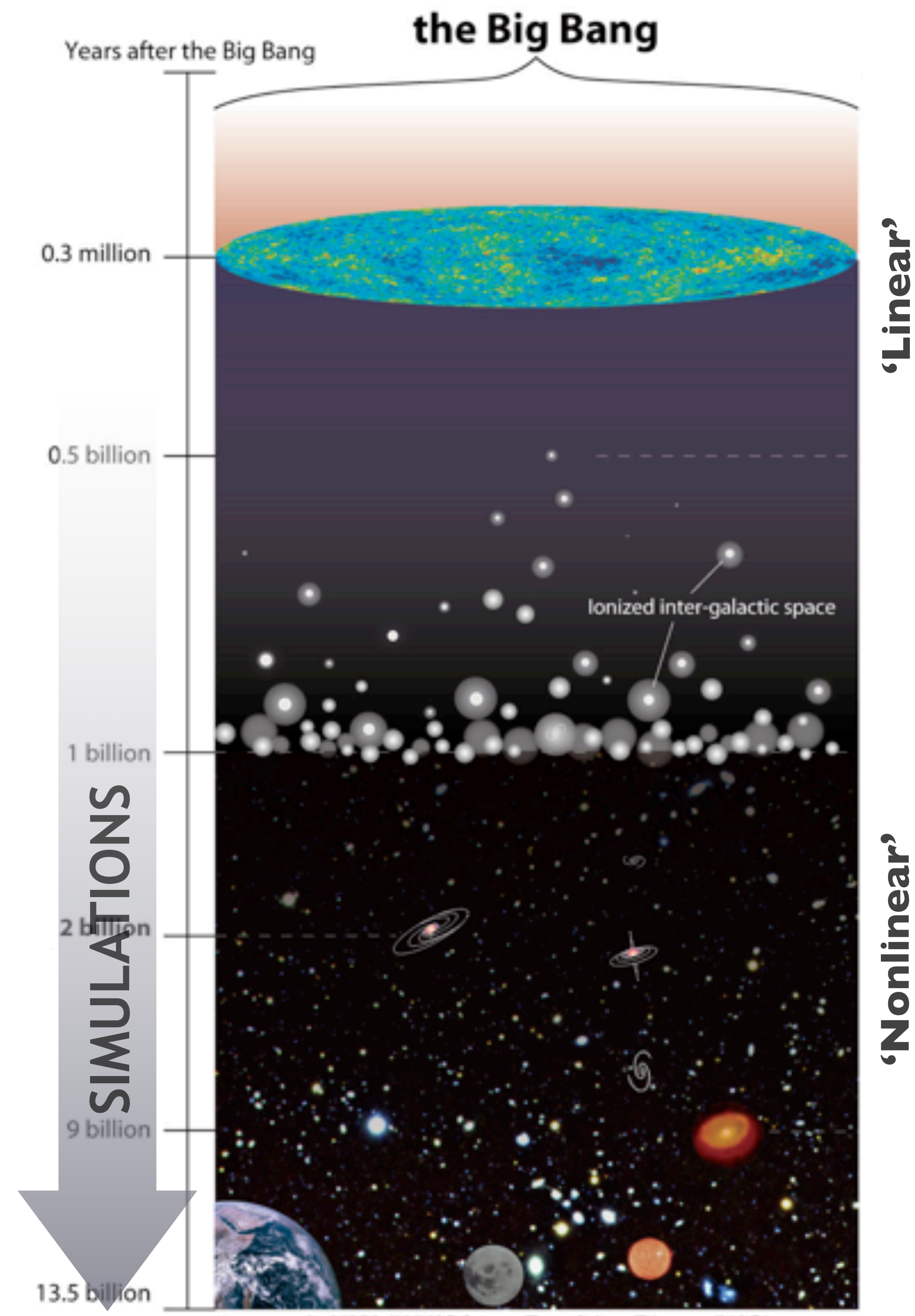
The Dark Universe

- Dark Energy: Multiple observations show that the expansion of the Universe is accelerating (first in 1998, Nobel prize 2011)
- Imagine you throw a ball in the air and instead of coming down it flies upwards faster and faster!
- Questions: What is it? Why is it important now? Being totally ignorant, currently our main task is to characterize it better and exclude some of the possible explanations
- Dark Matter: Observations show that ~27% of the matter in the Universe is “dark”, i.e. does not emit or absorb light
- So far: indirect detection, aims: characterize nature of dark matter and detect the actual dark matter particle

~95% of the Universe is “dark” -- we do not understand the nature and origin of dark energy and dark matter.

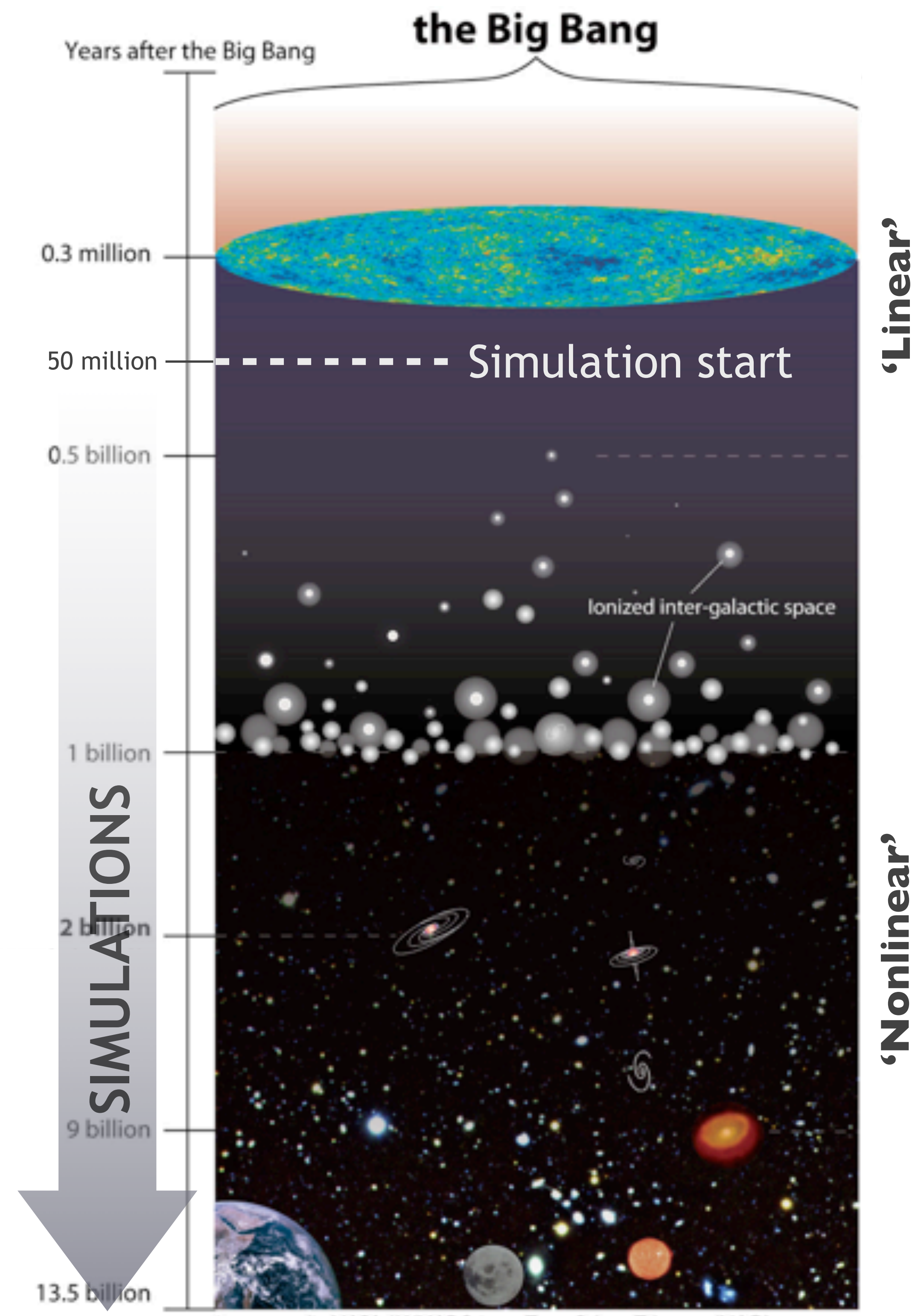
Structure Formation: The Basic Paradigm

- **Solid understanding of structure formation; success underpins most cosmic discovery**
  - Initial conditions determined by primordial fluctuations
  - Initial perturbations amplified by gravitational instability in a dark matter-dominated Universe
  - Relevant theory is gravity, field theory, and atomic physics ('first principles')
- **Early Universe:** Linear perturbation theory very successful (CMB)
- **Latter half of the history of the Universe:** Nonlinear domain of structure formation, impossible to treat without large-scale computing



Structure Formation: The Basic Paradigm

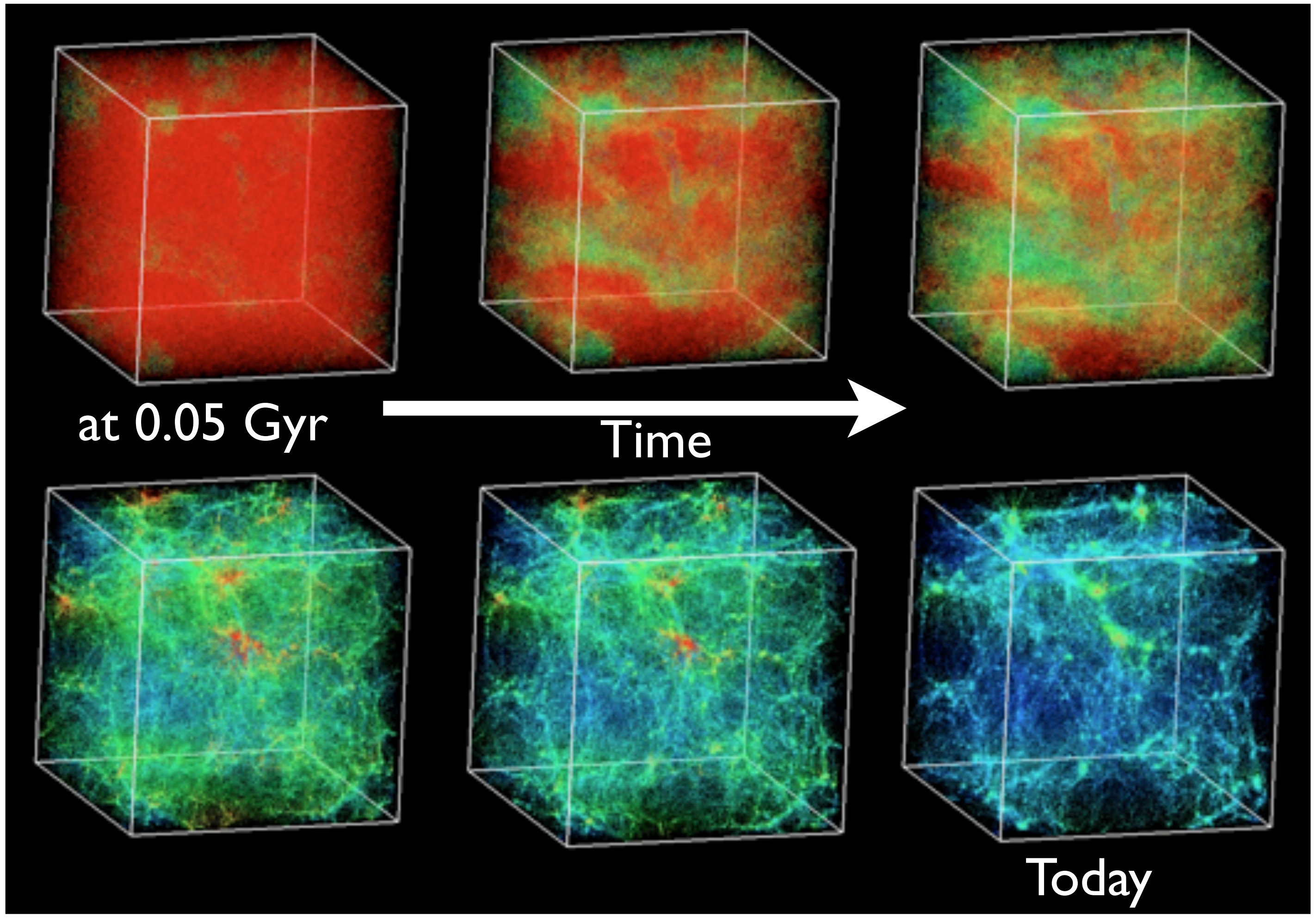
- **Solid understanding of structure formation; success underpins most cosmic discovery**
  - Initial conditions determined by primordial fluctuations
  - Initial perturbations amplified by gravitational instability in a dark matter-dominated Universe
  - Relevant theory is gravity, field theory, and atomic physics (‘first principles’)
- **Early Universe:** Linear perturbation theory very successful (CMB)
- **Latter half of the history of the Universe:** Nonlinear domain of structure formation, impossible to treat without large-scale computing



Computing the Universe

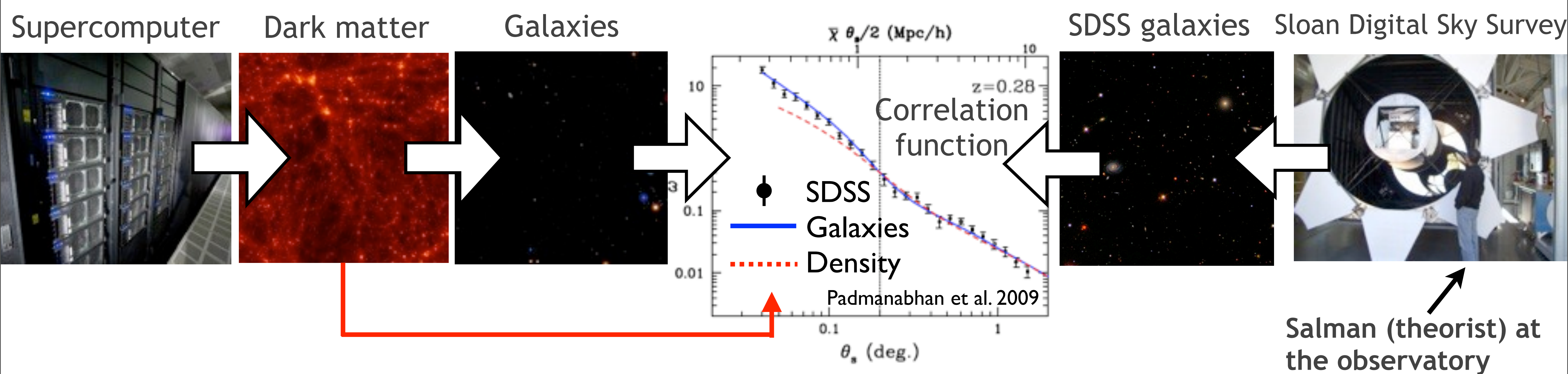
- Gravity dominates at large scales, key task: solve the Vlasov-Poison equation (VPE)
- VPE is 6-D and cannot be solved as PDE, therefore N-body methods
- Cosmological VPE: a “wrong-sign” electrostatic plasma with a time-dependent particle “charge”
- Particles are tracers of the dark matter in the Universe, mass typically at least ~10^9 $M_\odot$
- At smaller scales, add gas physics, feedback etc., sub-grid modeling inevitable

$$
\begin{aligned}
\frac{\partial f_i}{\partial t} + \dot{\mathbf{x}}\frac{\partial f_i}{\partial \mathbf{x}} - \nabla\phi\frac{\partial f_i}{\partial \mathbf{p}} = 0, \qquad \mathbf{p} = a^2\dot{\mathbf{x}}, \\
\nabla^2\phi &= 4\pi G a^2(\rho(\mathbf{x},t) - \langle\rho_{\rm dm}(t)\rangle) = 4\pi G a^2\Omega_{\rm dm}\delta_{\rm dm}\rho_{\rm cr}, \\
\delta_{\rm dm}(\mathbf{x},t) &= (\rho_{\rm dm} - \langle\rho_{\rm dm}\rangle)/\langle\rho_{\rm dm}\rangle, \\
\rho_{\rm dm}(\mathbf{x},t) &= a^{-3}\sum_i m_i \int d^3\mathbf{p}\, f_i(\mathbf{x},\dot{\mathbf{x}},t).
\end{aligned}
$$



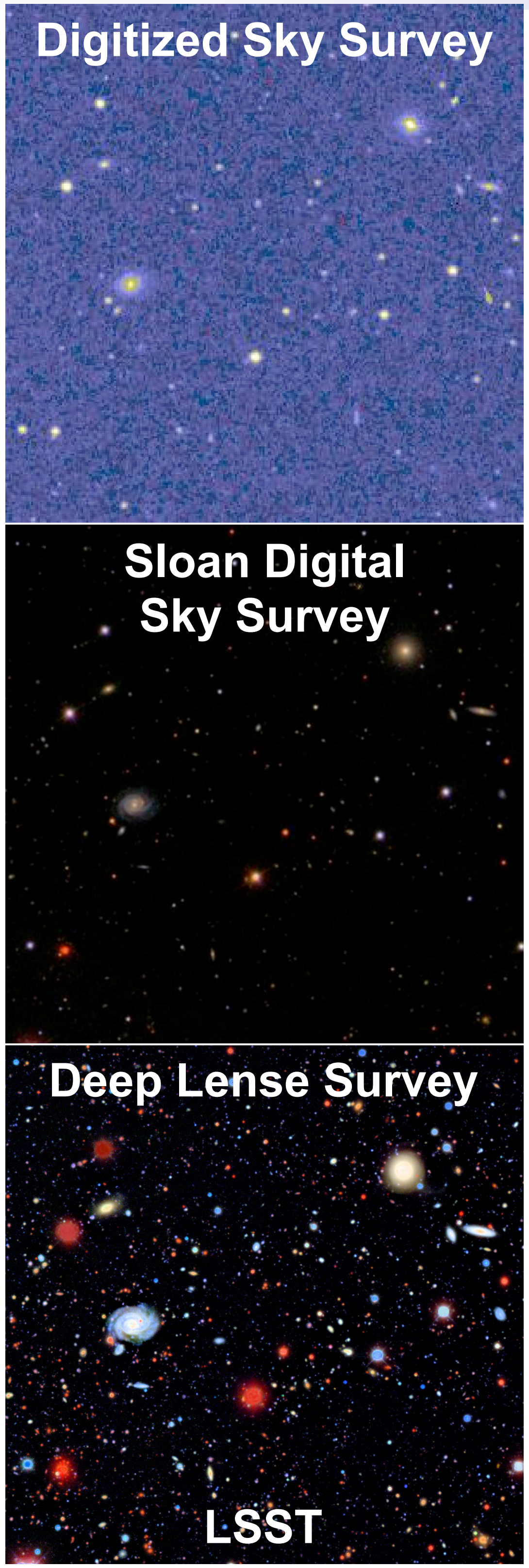
> “The Universe is far too complicated a structure to be studied deductively, starting from initial conditions and solving the equations of motion.”
>
> *Robert Dicke (Jayne Lectures, 1969)*

Connecting Theory and Observations

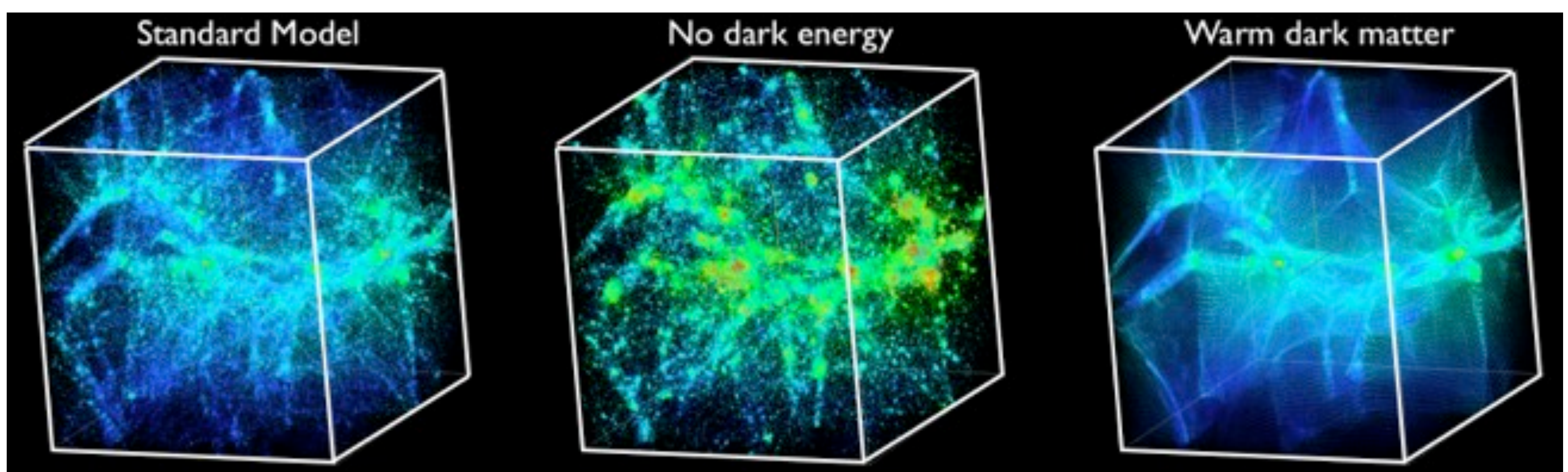


- Simulate the formation of the large scale structure of the Universe via dark matter tracer particles
- Take dark energy into account in the expansion history
- Measure the high-density peaks (dark matter halos) in the mass distribution
- “Light traces mass” to first approximation, therefore populate the halos with galaxies, number of galaxies depends on mass of halo (constraints from observations)
- Galaxy population prescription (hopefully) independent of cosmological model

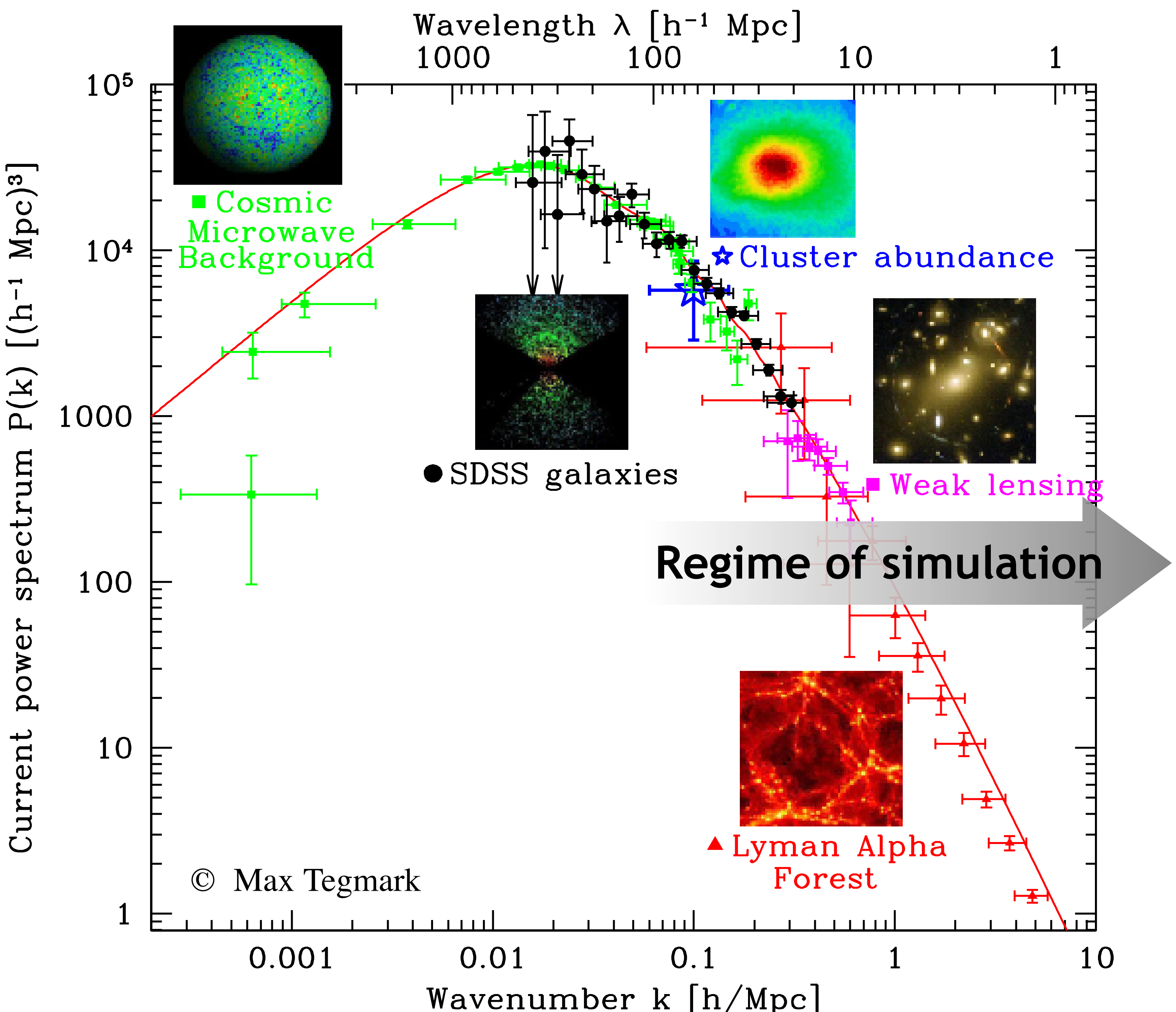
Challenges Ahead



- **Data Challenge:** Next generation cosmological observatories aim to understand the nature of the dark universe by going “deeper, faster, wider” (Large Synoptic Survey Telescope, LSST) -- pushing current boundaries **by orders of magnitude**
  - 30 terabytes of data *per night*; billions of galaxies
- **Modeling Challenge:** Scales that are resolved by future surveys become smaller and smaller, demanding (i) ever larger simulations with increased mass and force resolution; (ii) more details in the physics
  - Simulations are very costly, we need a large number
- **Analysis Challenge:** We have only one sky and cannot do controlled experiments, “inverting” the 3-D sky



The Matter Power Spectrum



2-point correlation function:

$$\xi(\vec{x}) = \int \frac{d^3\vec{y}}{V}\delta(\vec{y}-\vec{x})\delta(\vec{y}) = \int \frac{d^3\vec{k}}{(2\pi)^3 V}\underbrace{|\delta_k|^2}_{\text{power spectrum}} e^{i\vec{k}\cdot\vec{x}}$$

- 2-point correlation function: excess probability of finding an object pair separated by a distance r_{12} compared to that of a random distribution
- P(k): power spectrum, Fourier transform of correlation function

$$\Delta^2(k) = \frac{k^3 P(k)}{2\pi^2}$$

- Power spectrum very sensitive to physics of interest: amount and properties of dark matter, dark energy, neutrino mass, ...
- Many different probes for measuring P(k)

Length scale of interest:
1 parsec (pc)=3.26 light years~3·10^{13}km, separation of stars in a galaxies
Mpc=10^{6}pc: ~ separation of bright galaxies

The Advent of Precision Cosmology

- Cosmology has entered the era of precision science, from order of magnitude estimates to 10% accuracy measurements of mass content, geometry of the Universe, spectral index of primordial fluctuations and their normalization, dark energy EOS, --
- Next step: observations at the 1% accuracy limit; theory and predictions have to keep up!
- Why do we need higher accuracy?

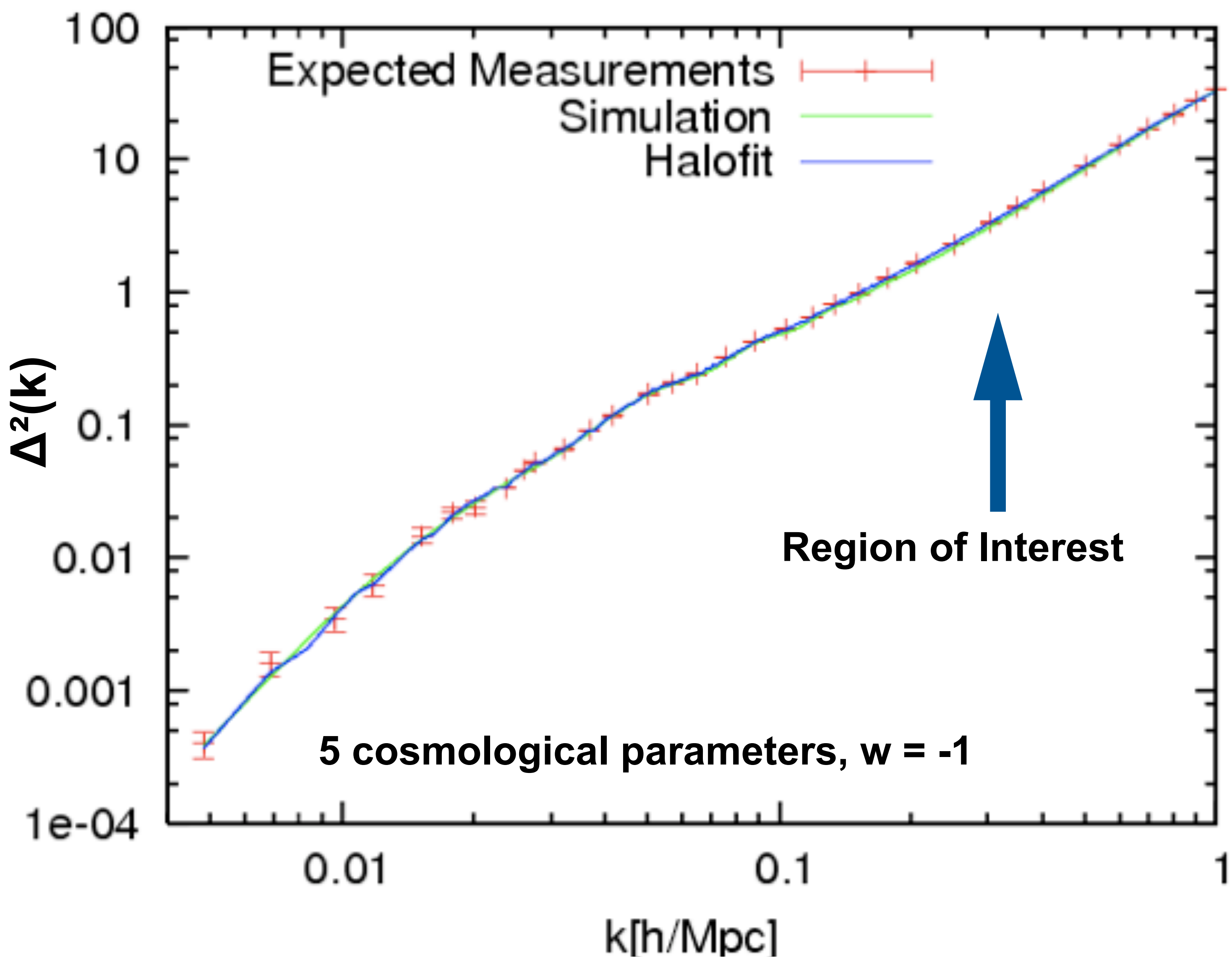
The Advent of Precision Cosmology

- Cosmology has entered the era of precision science, from order of magnitude estimates to 10% accuracy measurements of mass content, geometry of the Universe, spectral index of primordial fluctuations and their normalization, dark energy EOS, --
- Next step: observations at the 1% acc… have to keep up!
- Why do we need higher accuracy?

**Douglas Scott, UBC
at the Santa Fe Cosmology
Workshop in 2005**

The One Percent Challenge and its Importance

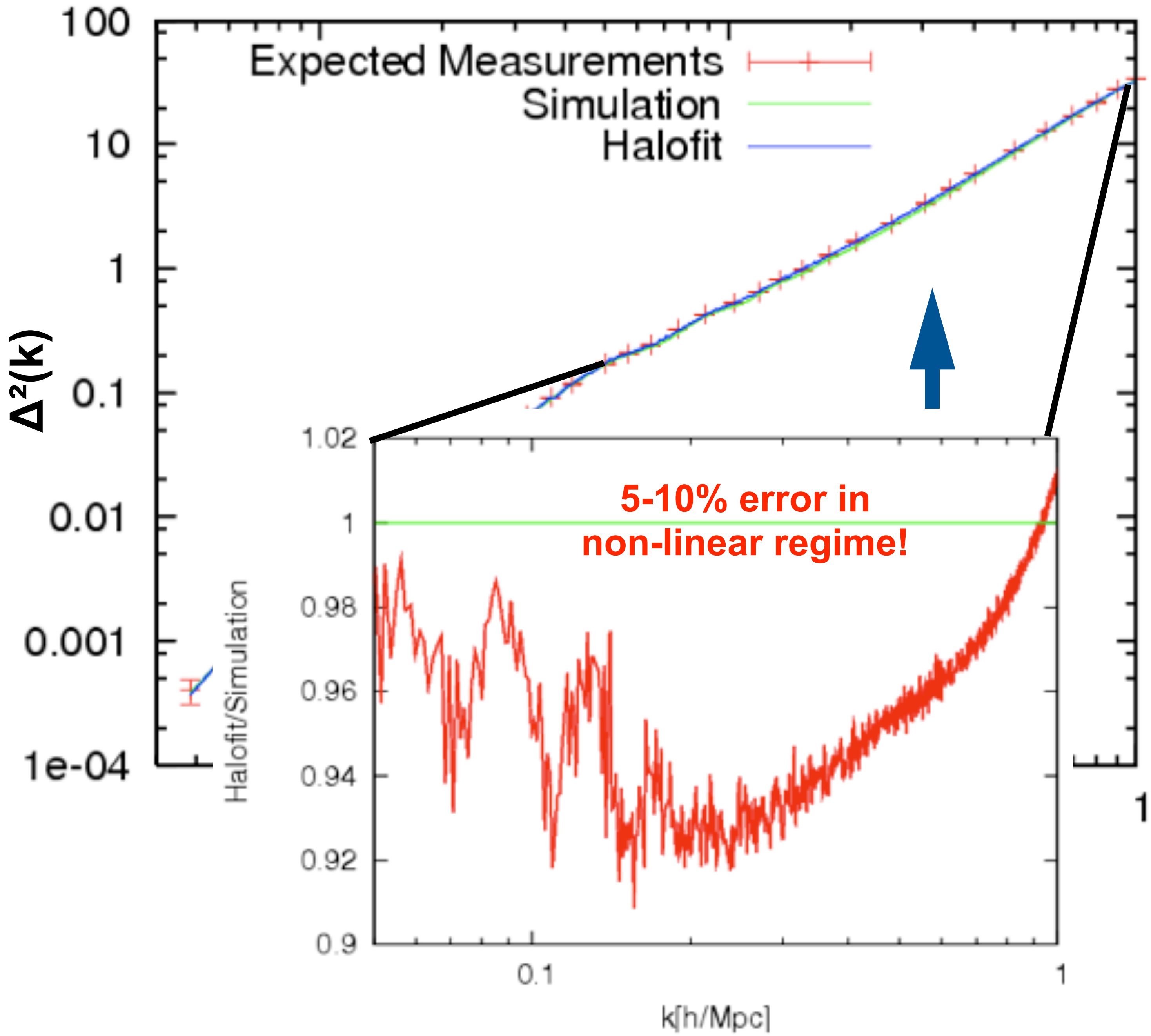
- Why do we need higher accuracy in our theoretical predictions?
- Example here: matter power spectrum
- Question: how badly will our constraints on dark energy be biased if we *do not* reach the same accuracy in our modeling as we might have in our data?
- Generate mock data set with the expected 1% error
- Analyze data with current method using HaloFit to model the matter power spectrum
    - HaloFit (Smith et al. 2003): semi-analytic fit for the power spectrum, based on modeling approach and tuned to simulations, accurate at the 5-10% level



$$\Delta^2(k) = \frac{k^3 P(k)}{2\pi^2};\ P(\vec{k}) = \langle \delta^2(\vec{k}) \rangle$$

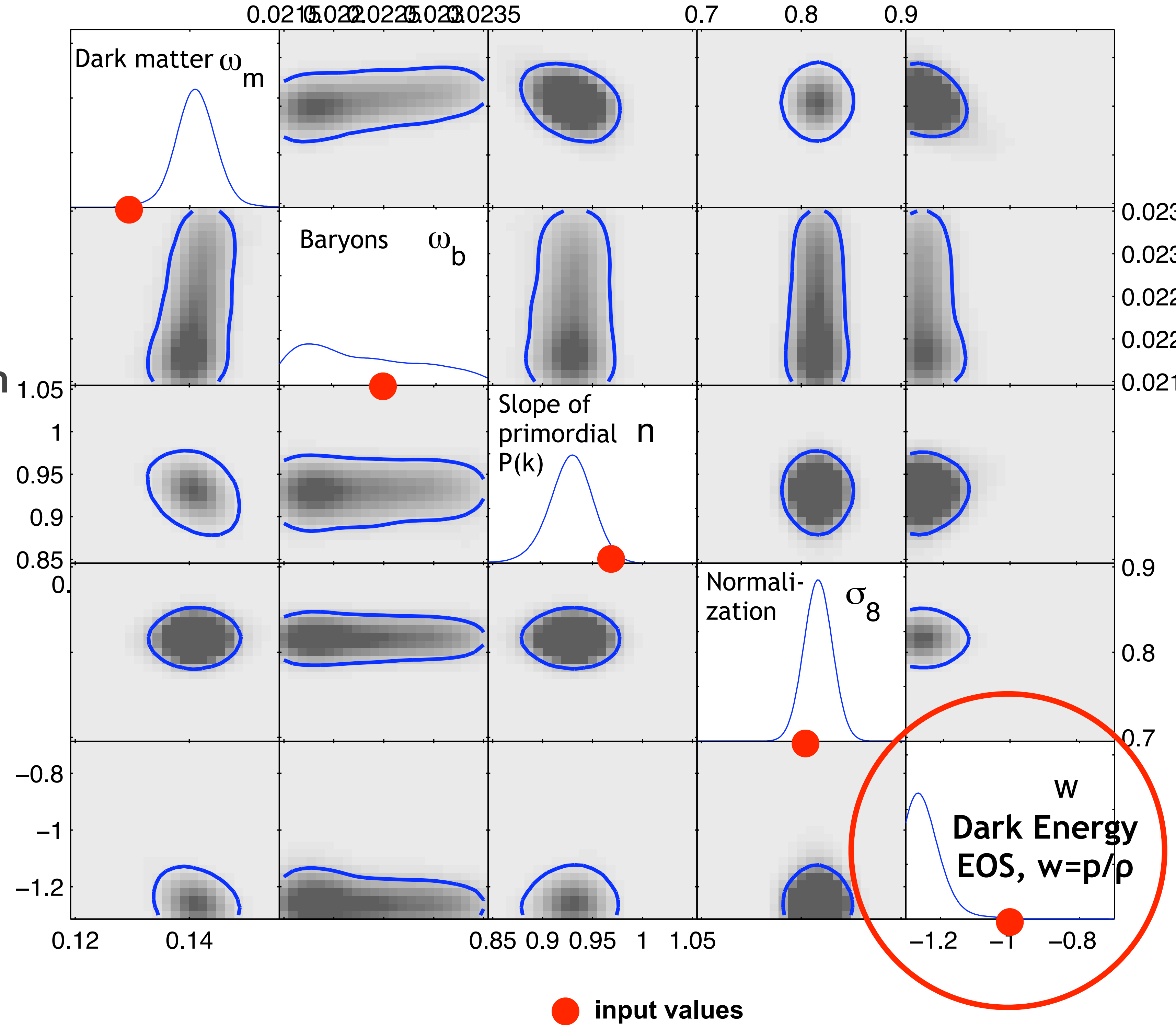
The One Percent Challenge and its Importance

- Why do we need higher accuracy in our theoretical predictions?
- Example here: matter power spectrum
- Question: how badly will our constraints on dark energy be biased if we *do not* reach the same accuracy in our modeling as we might have in our data?
- Generate mock data set with the expected 1% error
- Analyze data with current method using HaloFit to model the matter power spectrum
  - HaloFit (Smith et al. 2003): semi-analytic fit for the power spectrum, based on modeling approach and tuned to simulations, accurate at the 5-10% level



Analysis of the "True data"

- Generate mock data from high-resolution simulation
- Use Halofit for analysis; remember, halofit ~5-10% inaccurate on scales of interest
- Parameters are up to 20% wrong! (We checked that with more accurate predictions the answer is correct)
- Only solution: precision simulations
- Analysis takes at least 10,000 input power spectra for MCMC, each simulation takes ~20,000 CPU hours
- With a 2000 node cluster running 24/7, our analysis will take ~30 years, hmmm...

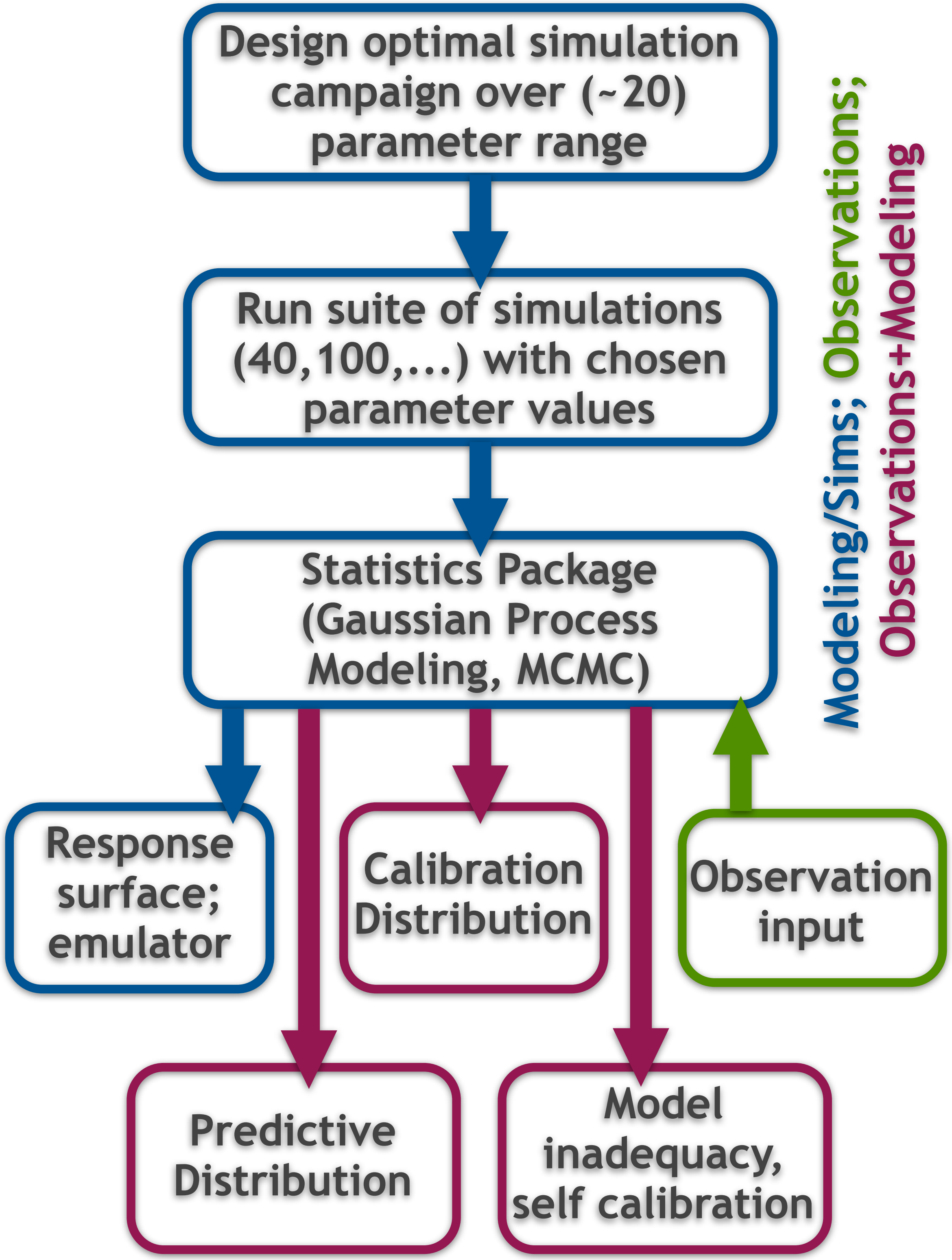


LSSFast: Sub-Percent Precision Prediction for P(k) in sub-seconds

- **Aim: predict P(k) out to scales of k~1 h/Mpc at 1% accuracy between z=0 and z=1**
  - Regime of interest for current weak lensing surveys
  - Baryonic physics at these scales is sub-dominant, so physics is "easy"
  - Dynamic range for simulations manageable
- **Step 1: Show that simulations can be run at the required accuracy (Heitmann et al. ApJ 2005; Heitmann et al., ApJ 2010)**
  - Code comparison
  - Initial conditions, force and mass resolution, ...
  - Minimal requirement: 1 billion particles, 1.3 Gpc volume, 50 kpc force resolution, ~ 20,000 CPU hours, few days on 250 processors + wait time in queue ~ 1 week per simulation on "Coyote", LANL cluster
- **Step 2: Cosmic Calibration Framework (Heitmann et al. ApJL 2006, Heitmann et al., ApJ 2009)**
  - With a small number of high-precision simulations, build a prediction scheme ("emulator") that provides the power spectrum for any cosmology within a given parameter space prior
  - ~ 40 cosmological models sufficient
- **Step 3: Cosmic Emulator (Lawrence et al., ApJ 2010)**
  - Carry out large number of simulations (~1,000) at varying resolution for 38 cosmologies, one high-resolution run per cosmology, emulator is effectively a "look-up" table
  - Emulator available at: www.lanl.gov/projects/cosmology/CosmicEmu

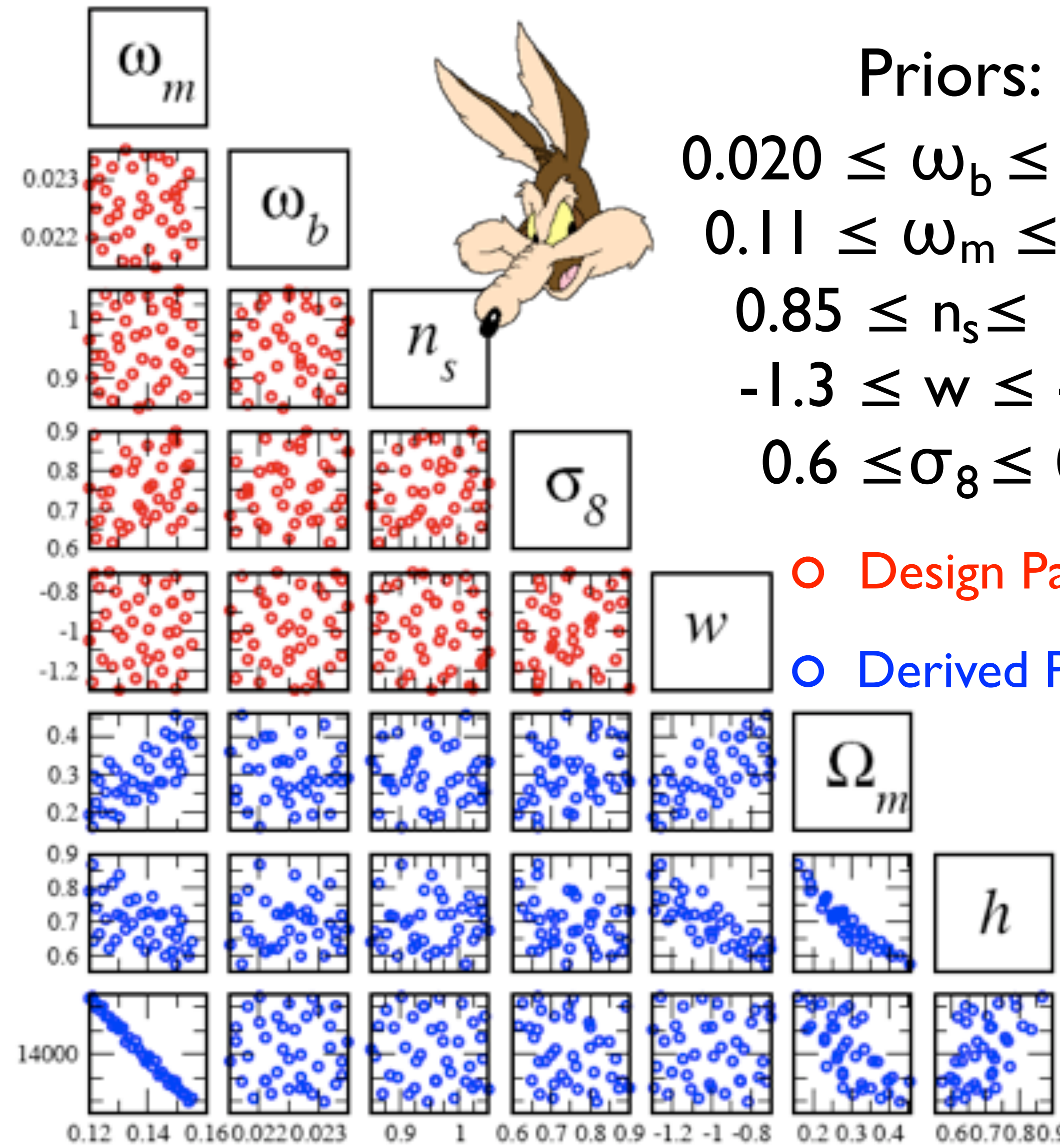
Cosmic Calibration Framework

- Step 1: Design simulation campaign, rule of thumb: O(10) models for each parameter
- Step 2: Carry out simulation campaign and extract quantity of interest, in our case, power spectrum
- Step 3: Choose suitable interpolation scheme to interpolate between models, here Gaussian Processes
- Step 4: Build emulator
- Step 5: Use emulator to analyze data, determine model inadequacy, refine simulation and modeling strategy...



The Simulation Design

- “Simulation design”: for a given set of parameters to be varied and a fixed number of runs, at what settings should the simulations be performed?
- In our case: five cosmological parameters, tens of high-resolution runs are affordable
- First idea: grid
  - Assume 5 parameters and each parameter should be sampled 3 times: 3^5=243 runs, not a small number, coverage of parameter space poor, only allows for estimating quadratic models ☹
- Second idea: random sampling
  - Good if we can perform many runs -- if not, most likely insufficient sampling of some of the parameter space due to clustering
- Our approach: orthogonal-array Latin hypercubes (OA-LH) design
  - Good coverage of parameter space
  - Good coverage in projected dimensions



Priors:

$0.020 \leq \omega_b \leq 0.025$

$0.11 \leq \omega_m \leq 0.15$

$0.85 \leq n_s \leq 1.05$

$-1.3 \leq w \leq -0.7$

$0.6 \leq \sigma_8 \leq 0.9$

o Design Parameters

o Derived Parameters

Priors are informed by current cosmological constraints, the tighter the priors, the easier to build a prediction tool. Restriction in number of parameters also helps!

The Coyote Universe

Priors:

$0.020 \leq \omega_b \leq 0.025$

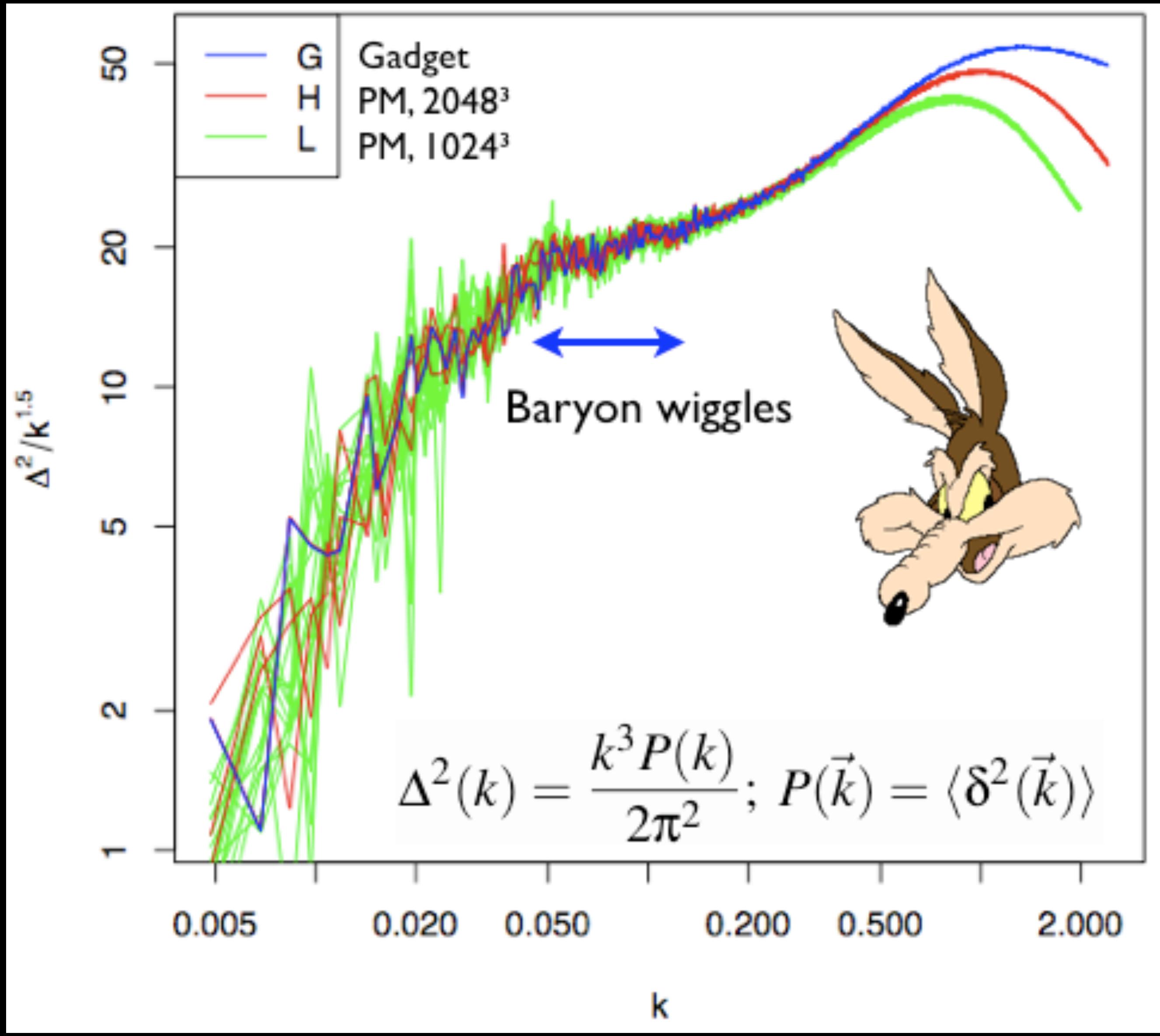
$0.11 \leq \omega_m \leq 0.15$

$0.85 \leq n_s \leq 1.05$

$-1.3 \leq w \leq -0.7$

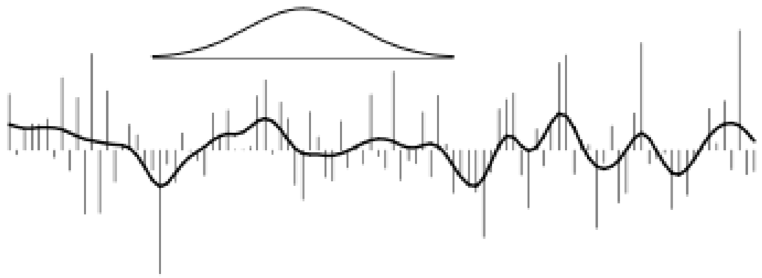
$0.6 \leq \sigma_8 \leq 0.9$

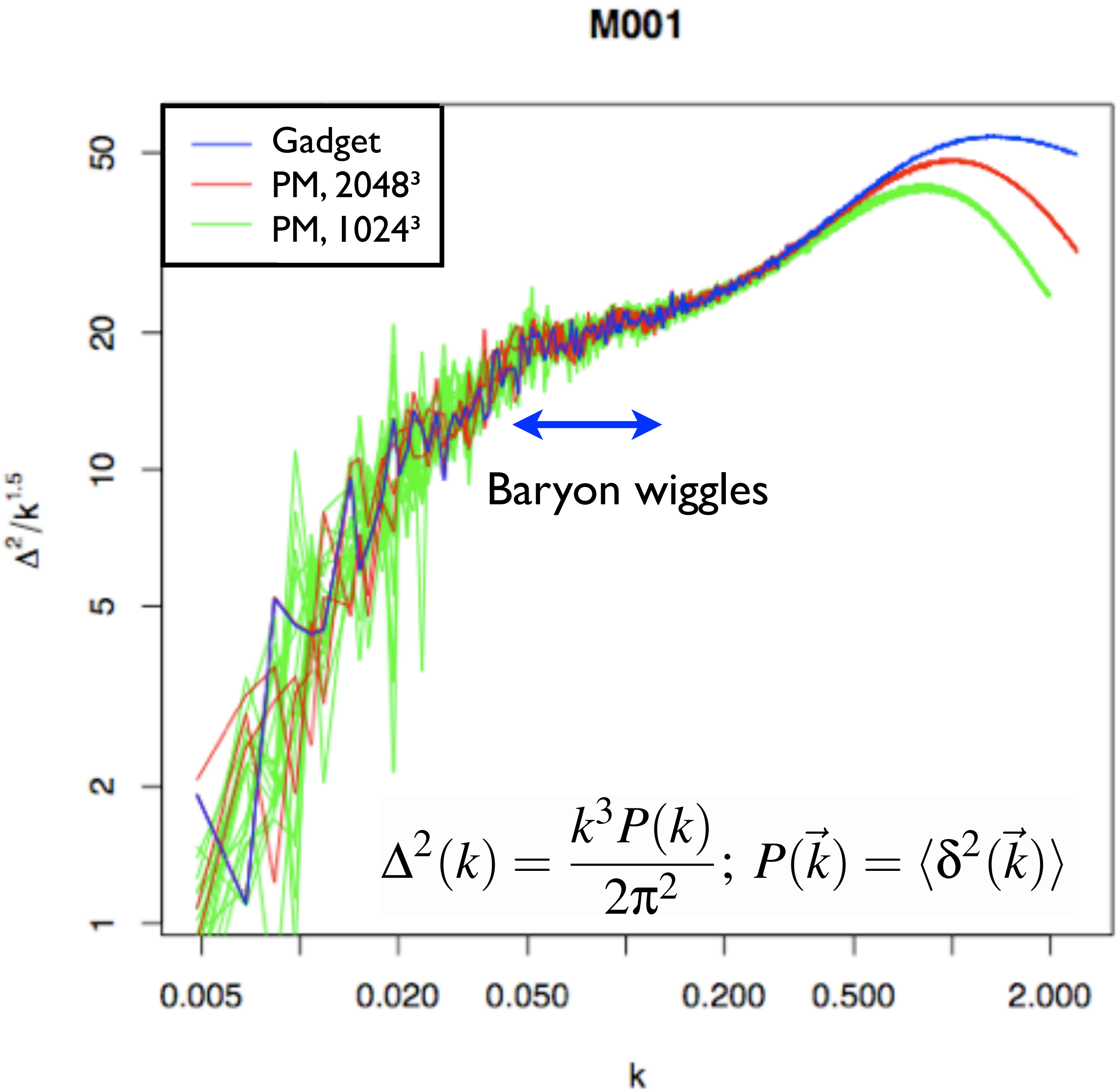
- 37 model runs + ΛCDM
  - 16 low resolution realizations (green)
  - 4 medium resolution realizations (red)
  - 1 high resolution realization (blue)
  - 11 outputs per run between z = 0 - 3
- Restricted priors to minimize necessary number of runs
- 1.3 Gpc boxes, m_P ~$10^{11} M_\odot$
- ~1000 simulations, 60TB



Next step: Smooth Power Spectrum

- Each simulation represents one possible realization of the Universe in a finite volume
- Need smooth prediction for building the emulator for each model
- Major challenge: Make sure that baryon features are not washed out or enhanced due to realization scatter
- Construct smooth power spectra using a process convolution model (Higdon 2002)
- Basic idea: calculate moving average using a kernel whose width is allowed to change to account for nonstationarity

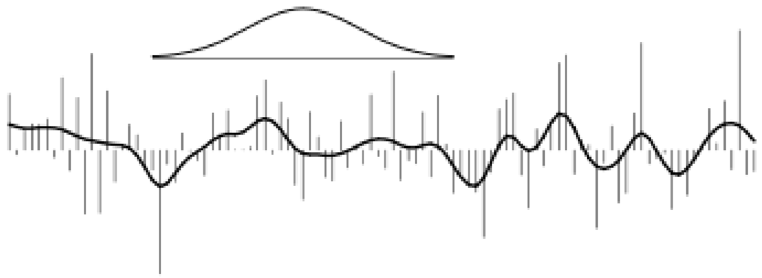


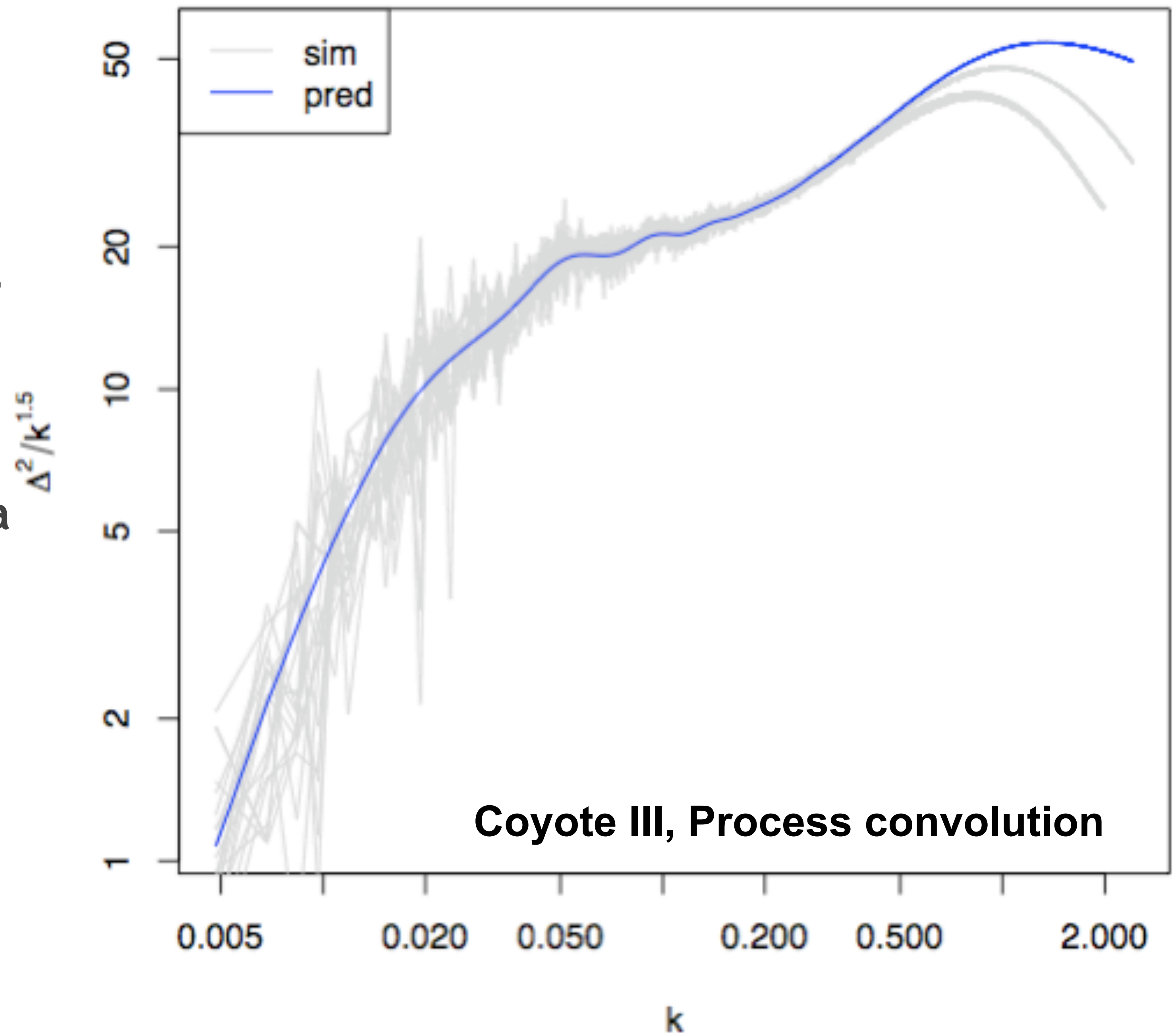


$$\Delta^2(k) = \frac{k^3 P(k)}{2\pi^2};\ P(\vec{k}) = \langle \delta^2(\vec{k}) \rangle$$

Next step: Smooth Power Spectrum

- Each simulation represents one possible realization of the Universe in a finite volume
- Need smooth prediction for building the emulator for each model
- Major challenge: Make sure that baryon features are not washed out or enhanced due to realization scatter
- Construct smooth power spectra using a process convolution model (Higdon 2002)
- Basic idea: calculate moving average using a kernel whose width is allowed to change to account for nonstationarity





The Interpolation Scheme: Gaussian Processing

- **After simulation design specification: Build interpolation scheme that yields predictions for any cosmology within the priors**
- **Model simulation outputs using a p_η - dimensional basis representation**
  - Find suitable set of orthogonal basis vectors $\phi_i(k,z)$, here: Principal Component Analysis
  - 5 PC bases needed, fifth PC basis pretty flat
  - Next step: modeling the weights
  - Here: Gaussian Process modeling (non-parametric regression approach, local interpolator; specified by mean function and covariance function)

Cosmological parameters

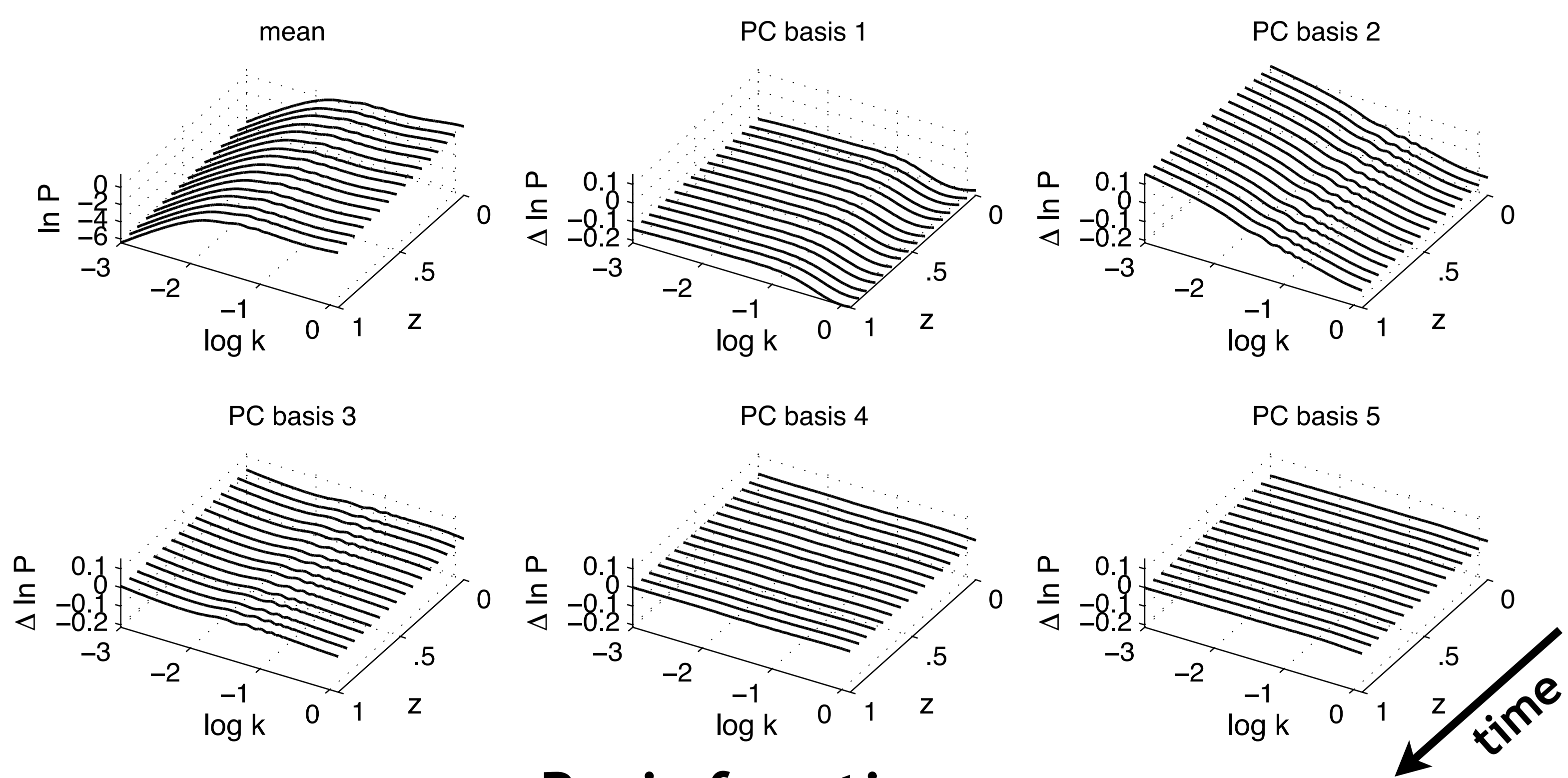
Number of parameters, 5

Number of basis functions, here: 5

$$\ln\left\{\frac{\Delta^2(k,z)}{2\pi k^{3/2}}\right\} = \sum_{i=1}^{p_\eta} \phi_i(k,z) w_i(\theta) + \epsilon \qquad \theta \in [0,1]^{p_\theta}$$

Basis functions, here: PC basis

Weights, here: GP model

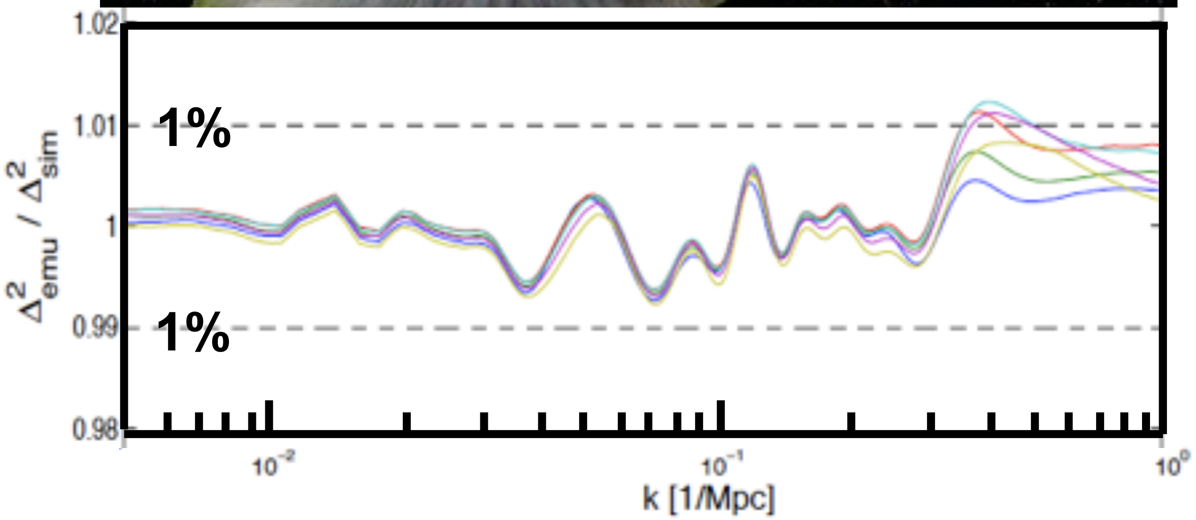


Basis functions

The Cosmic Emu(lator)

- Prediction tool for matter power spectrum has been constructed
- Accuracy within specified priors between z=0 and z=1 out to k=1 h/Mpc at the 1% level achieved
- Emulator has been publicly released, C code, Fortran wrapper available
- Next steps
  - Extend k-range ✓
  - Include more physics, e.g. neutrinos
  - Other statistics, e.g. shear spectrum ✓

Emulator performance: Comparison of prediction and simulation output for a model not used to build emulator at 6 redshifts.

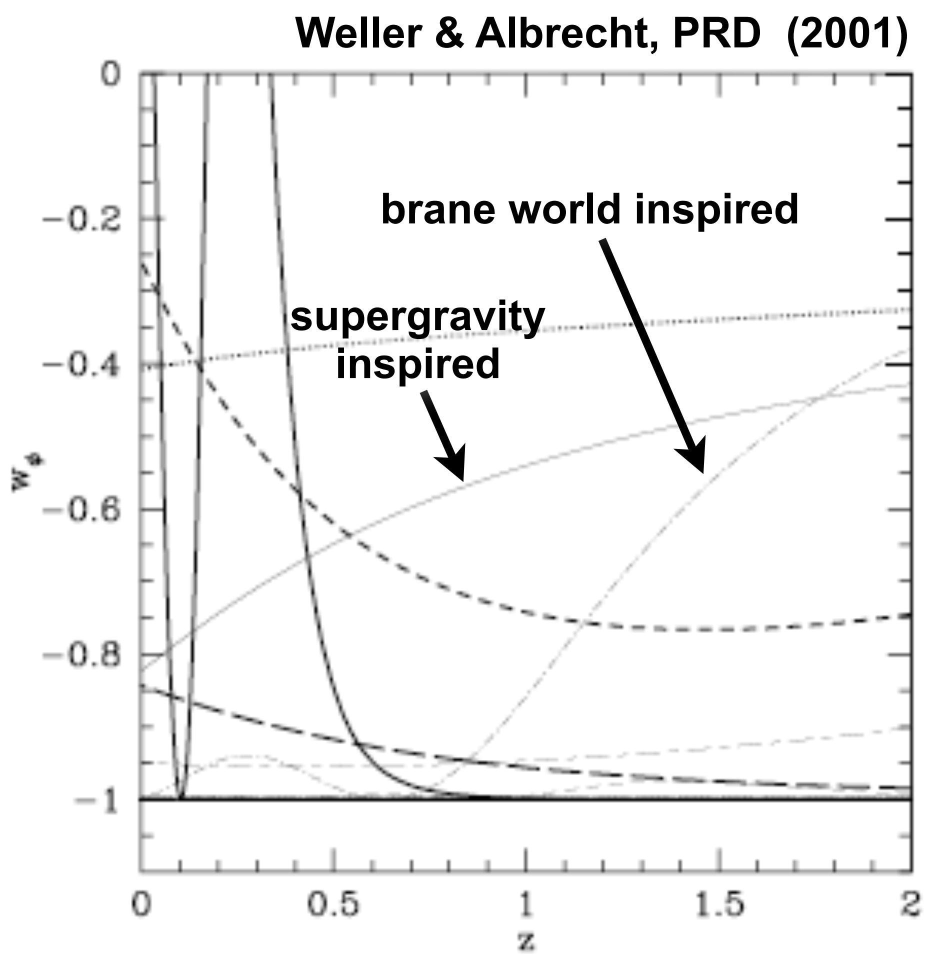


Cosmic Emulator in Action: LSSFast

- Instantaneous ‘oracle’ for nonlinear power spectrum, reduces compute time from weeks to negligible, accurate at 1% out to k~1/Mpc for wCDM cosmologies
- Enables direct MCMC with results from full simulations for the first time

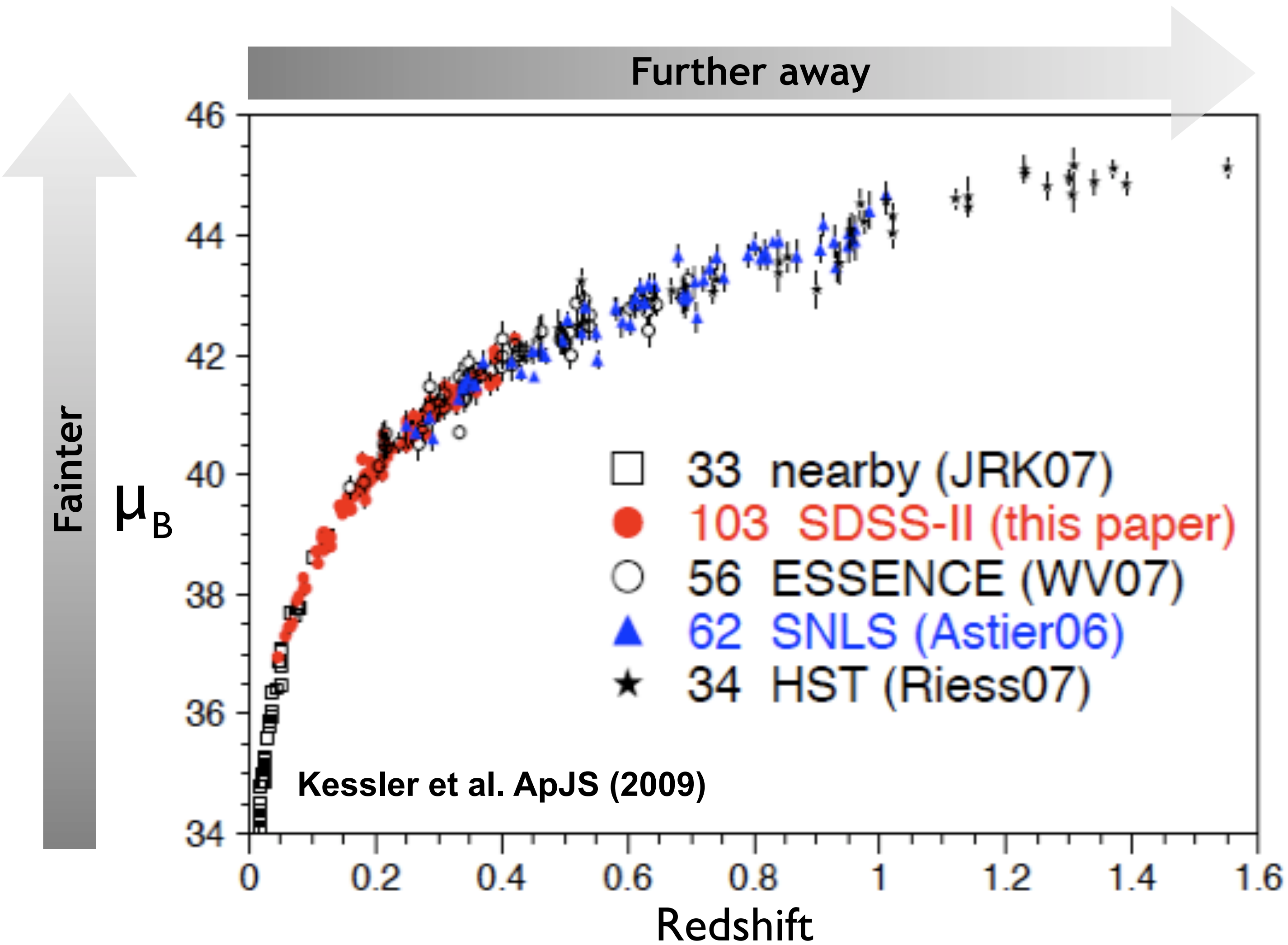
Analysis Challenge: The Nature of Dark Energy

- Problem: total ignorance about the origin and nature of dark energy
- So far in this talk: Assume the dark energy equation of state w=const.
- Key: we (the theorists) predict that for a “physically well motivated model”, EOS should be time varying
- More or less endless possibilities to invent models, theorists can calculate...
- Observers have something to look for... but we cannot test each and every model separately
- Aim: develop non-parametric reconstruction scheme



Equation of state parameter w as a function of redshift for various models

Reconstruction Task



- **Measurements of supernova magnitudes and w(z) connected via double-integral**
- **Some reconstruction approaches:**
  - Naive: fit μ and take two derivatives, bad approach for noisy data
  - Assume parametrized form for w, estimate associated parameters (e.g. Linder 2003)
  - Pick local basis representation for w(z) (bins, wavelets) and estimate associated coefficients (effectively piecewise constant description of w(z)) (e.g. Huterer & Cooray 2005)
- **Here: new, nonparametric reconstruction approach based on Gaussian Process models (Holsclaw et al. Phys. Rev. Lett 2010, Phys. Rev. D. 2010)**

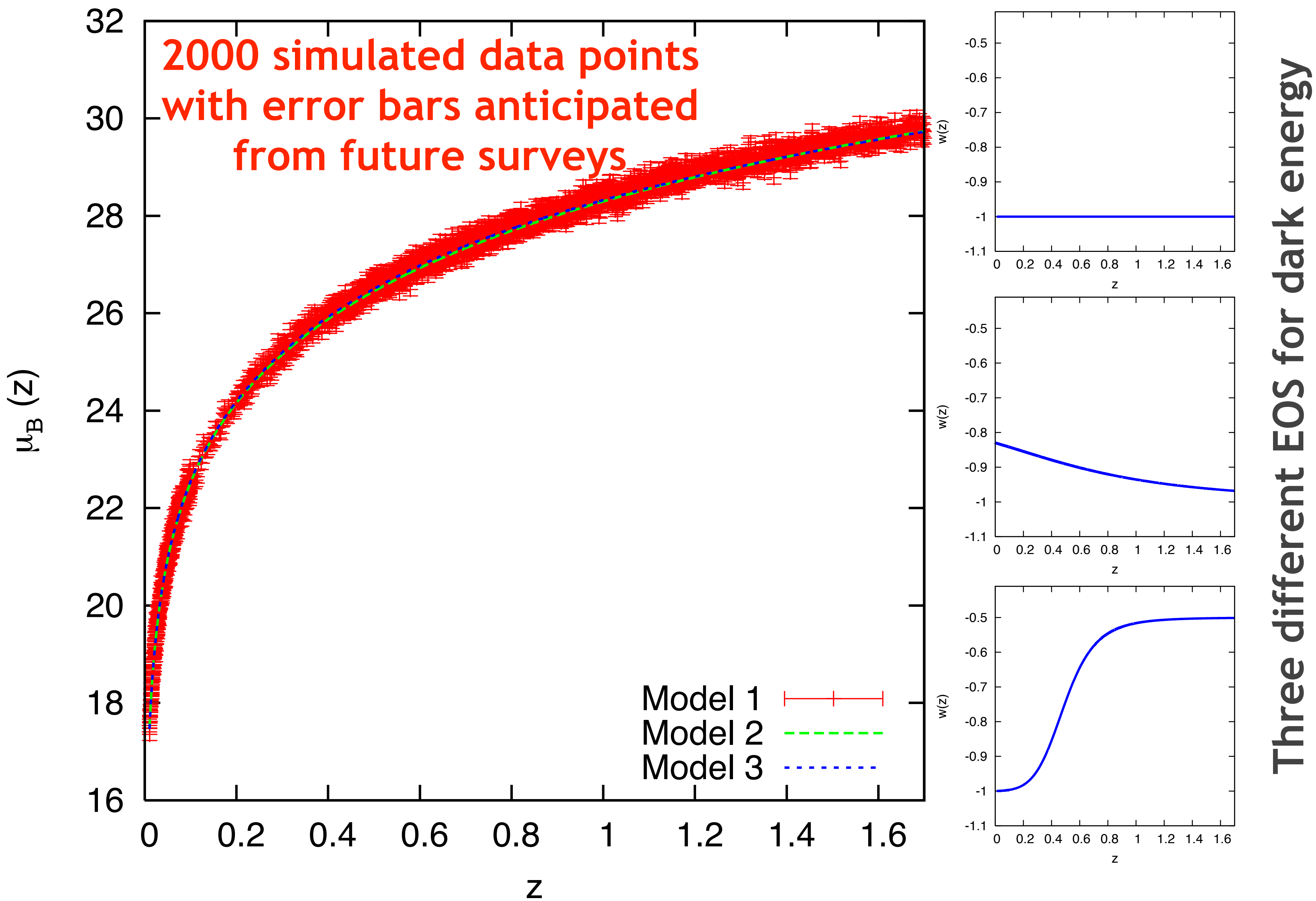
Observer

$$\mu_B(z) = m_B - M_B = 5\log_{10}\left(\frac{d_L(z)}{1\mathrm{Mpc}}\right) + 25$$

$$d_L(z) = (1+z)\frac{c}{H_0}\int_0^z ds\left[\Omega_m(1+s)^3 + (1-\Omega_m)(1+s)^3\exp\left(3\int_0^s \frac{w(u)}{1+u}du\right)\right]^{-\frac{1}{2}}$$

Theorist

The Challenge



- **Differences in the distance module μ are very small for different dynamical dark energy models**
- **To test our new method and compare with other methods we set up datasets for three different dark energy models with data quality of a future survey**

Reconstructing w(z) with GP Modeling

- Assume a GP for dark energy equation of state parameter

$$w(u) \sim \mathrm{GP}(-1, K(u,u')), \quad K(u,u') = \kappa^2 \rho^{|u-u'|^\alpha}$$

- Need to integrate over this in the expression for the distance modulus, where

$$y(s) = \int_0^s \frac{w(u)}{1+u} du$$

- Use the fact that the integral over a GP is another GP and specify covariance

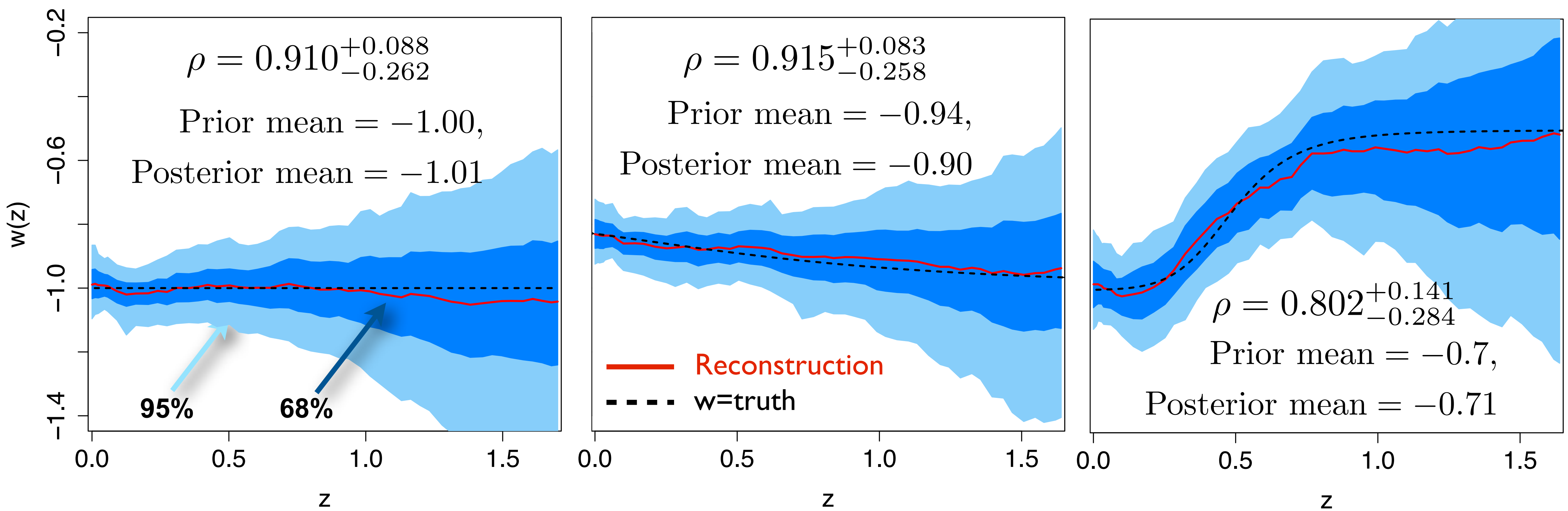
$$y(s) \sim \mathrm{GP}\left(-\ln(1+s), \kappa^2 \int_0^s \int_0^{s'} \frac{\rho^{|u-u'|^\alpha} du du'}{(1+u)(1+u')} \right)$$

- A joint GP for the two variables can be constructed

$$\begin{bmatrix} y(s) \\ w(u) \end{bmatrix} \sim \mathrm{GP}\left[\begin{bmatrix} -\ln(1+s) \\ -1 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right]$$

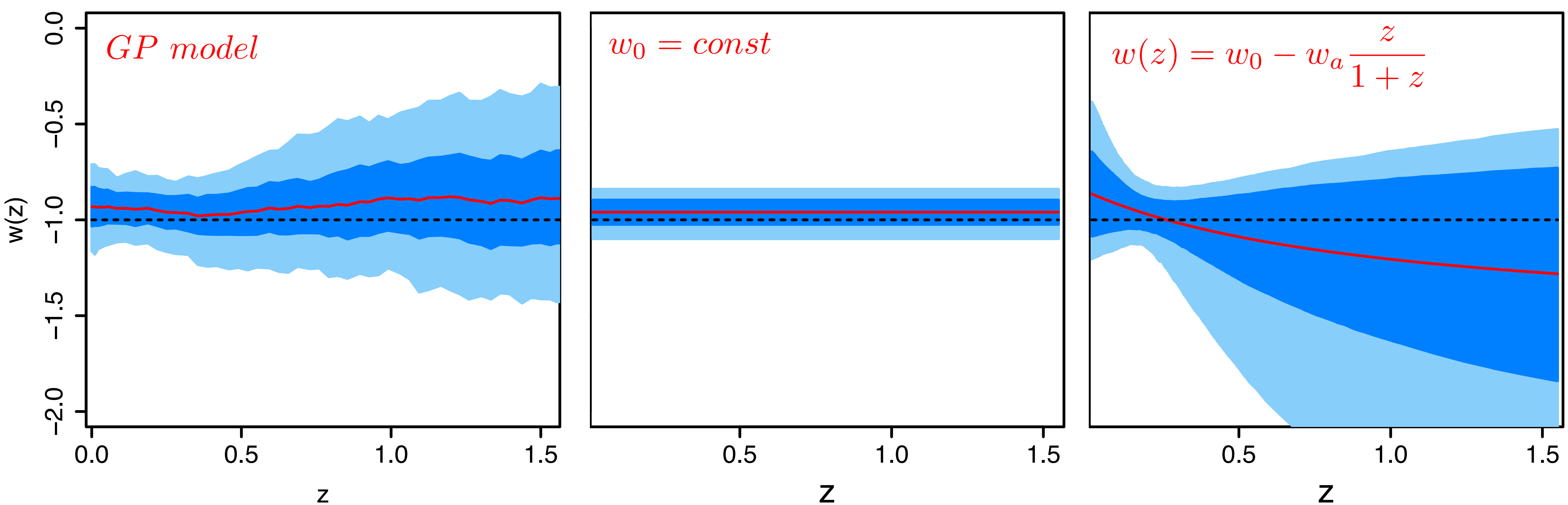
Results

- First: simplify task by fixing $\Omega_m = 0.27$ and $\Delta_\mu = 0$
- GP model: $w(u) \sim \mathrm{GP}(-1, K(u,u'))$ with $K(z,z') = \kappa^2 \rho^{|z-z'|}$
- Determine GP hyperparameters $\kappa,\ \rho$ from data
- Start with mean = -1, adjust after initial burn-in time
- Excellent results!



Results from Recent Data

- Combined data analysis of supernova data (Hicken et al.), cosmic microwave background data (WMAP), and data from the Sloan Digital Sky Survey (BAO)
- GP model and parametrization results (Holsclaw et al. Phys. Rev. D 2011)
- All are in agreement with a cosmological constant within error bars

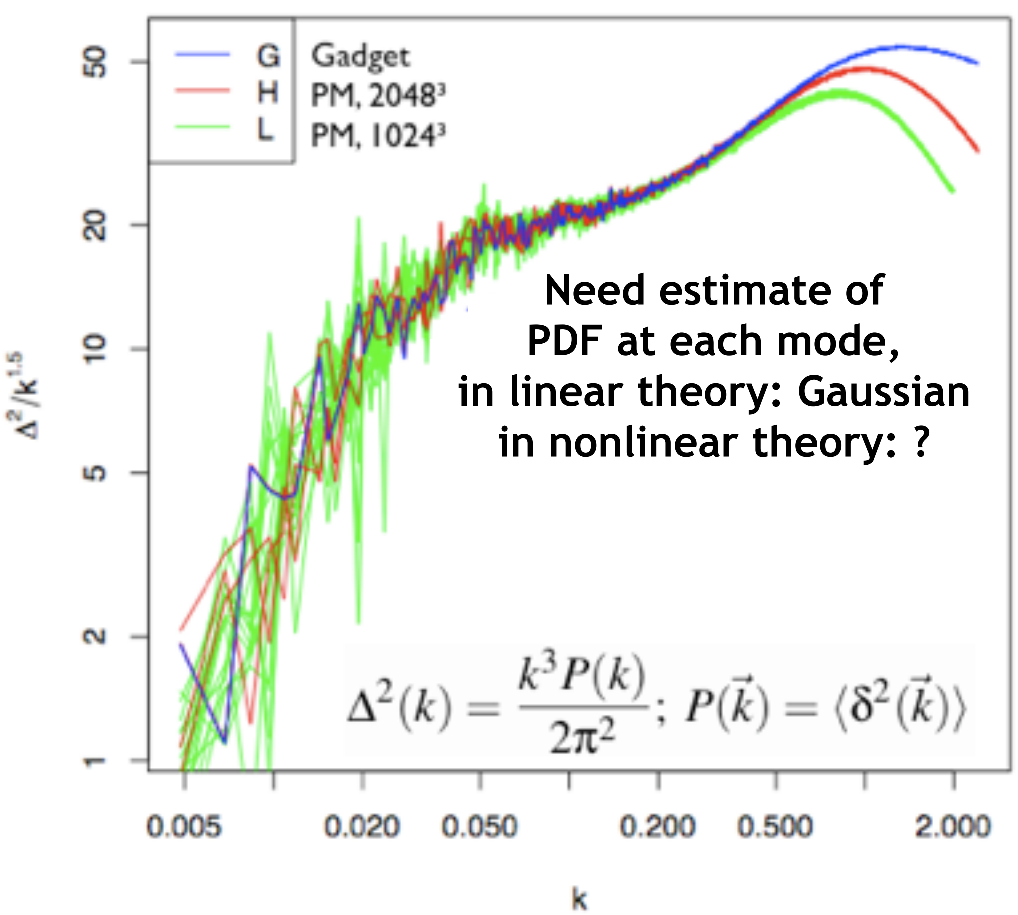


More Challenges Ahead, Some Examples

- LSST will gather equivalent of SDSS data within a couple of nights; equivalent of DES data within a couple of months
- We will not be any longer statistics limited but systematics limited, both observational and theoretical

Sloan Digital Sky Survey
~10 years of data taking

Example I: Covariances



- We only observe a finite part of the Universe, due to nonlinear coupling, modes are correlated
- Emulator provides diagonal part of covariance matrix, but we need full matrix for error estimate, Cov(k,k')
- We do not know the exact initial conditions, so we need many realizations to estimate the PDF at each mode and build up covariance matrix
- Thousands of simulations for each cosmology?

Dark Energy Survey
5 years, start 2012

Large Synoptic
Survey Telescope
10 years, start 2018

More Challenges Ahead, Some Examples

Example II: Combining Probes

- From the same survey, different cosmological probes are extracted
- E.g.: clustering statistics of galaxies, abundance of clusters of galaxies (bound, heavy objects)
- All measured from the same galaxies, will have same systematics
- Cross correlation between different probes
- Covariances?
- "Brute force": simulate the full survey with galaxy population thousands of times, measure correlations
- Difficulty: have to cover large range of scales

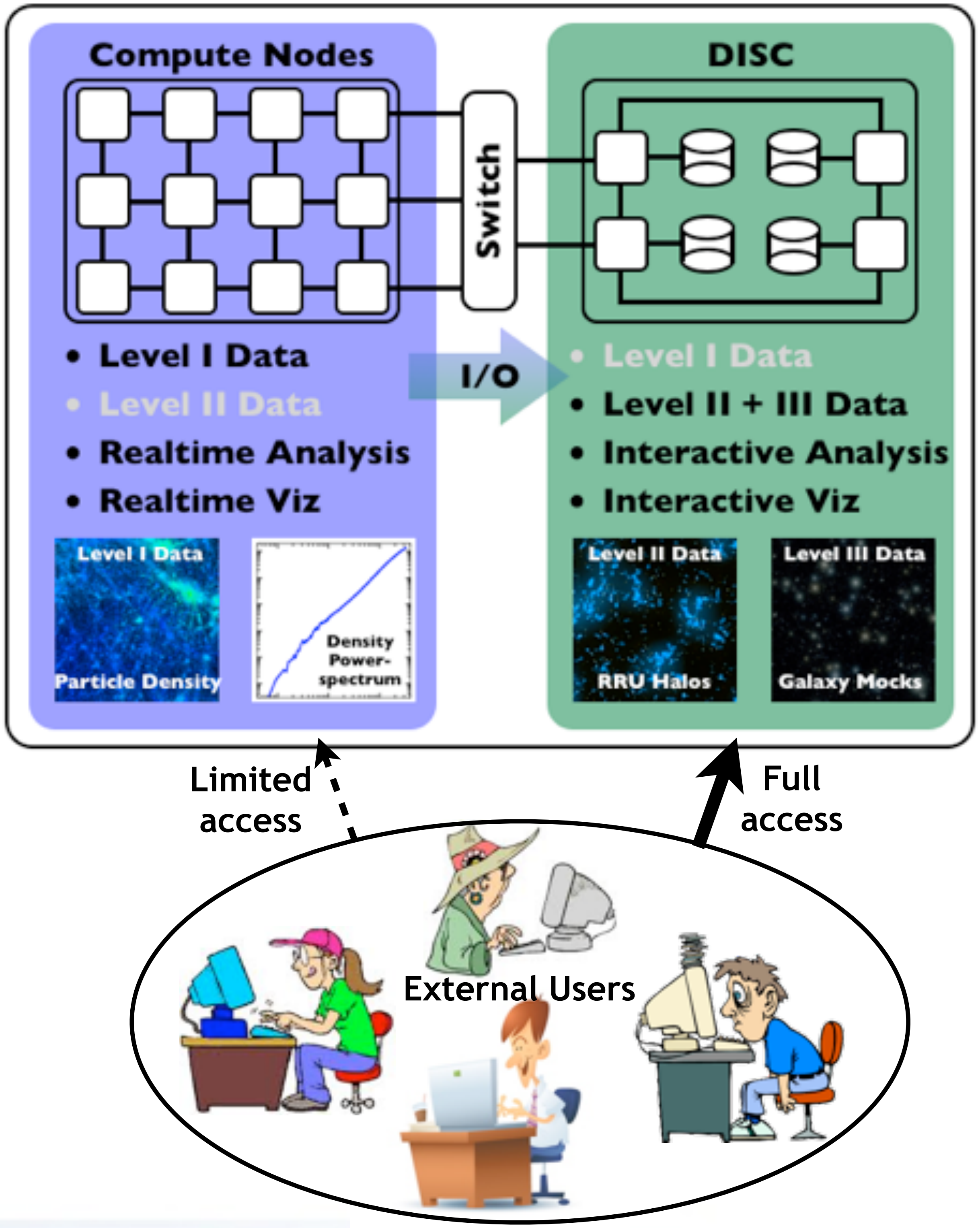
Example III: Modeling

- On large scales: gravity dominates
- On small scales: baryons become important, gas physics, feedback effects, not possible to do simulations from first principles
- Many modeling options, different groups find different results, if one observable is matched, another one will be off
- Simulations at least an order of magnitude more expensive than gravity only, many modeling parameters to be varied
- How do we incorporate our ignorance about the baryonic physics into our error budget and still get good constraints?

More Challenges Ahead, Some Examples

Example IV: The Data Challenge from a Simulator's Perspective

- **Simulation datasets: Currently simulation data generation is constrained only by storage and I/O bandwidth, ~PB datasets will be available in the near future**
  - In situ analysis: Large-scale analysis tasks on the compute platform; data compression
  - Post-processing: Post-run analyses on host system or associated 'active storage'
- **How can we efficiently share data?**
  - Simulation campaigns are carried out at very few places (supercomputer centers)
  - Outputs are very science rich, many people can contribute to the analysis
  - Moving raw data is impractical (at some point impossible), analysis often takes a lot of computing power
  - Need for making data *and* analysis opportunity available to the community



Thanks to all collaborators:

Our outstanding external collaborators (some pictures missing)

Our new group at Argonne National Laboratory