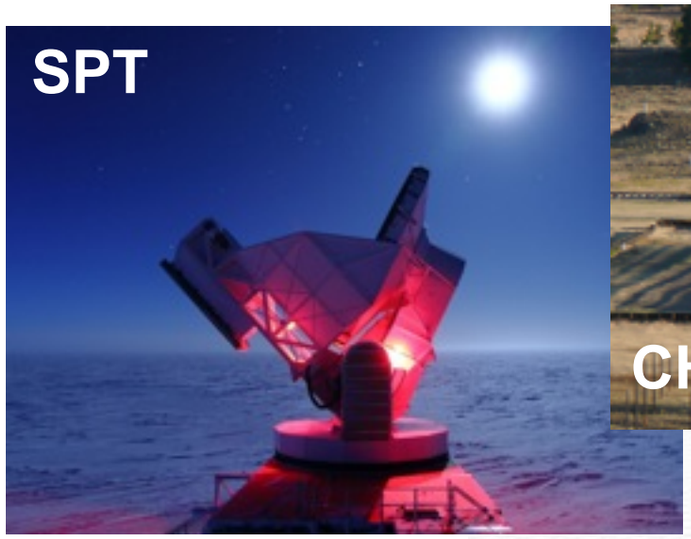
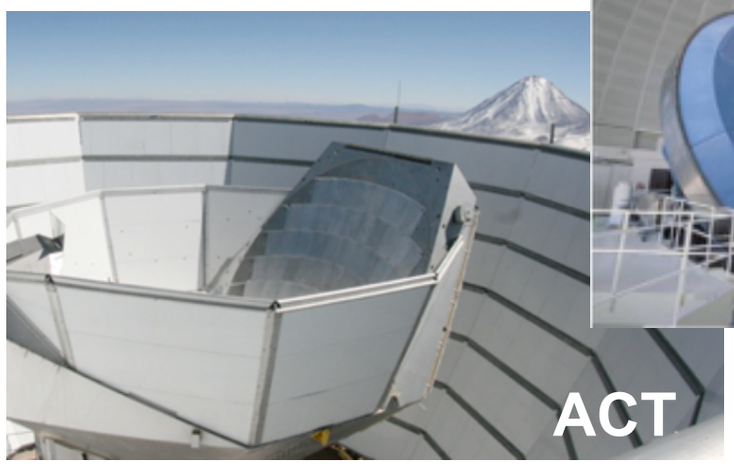
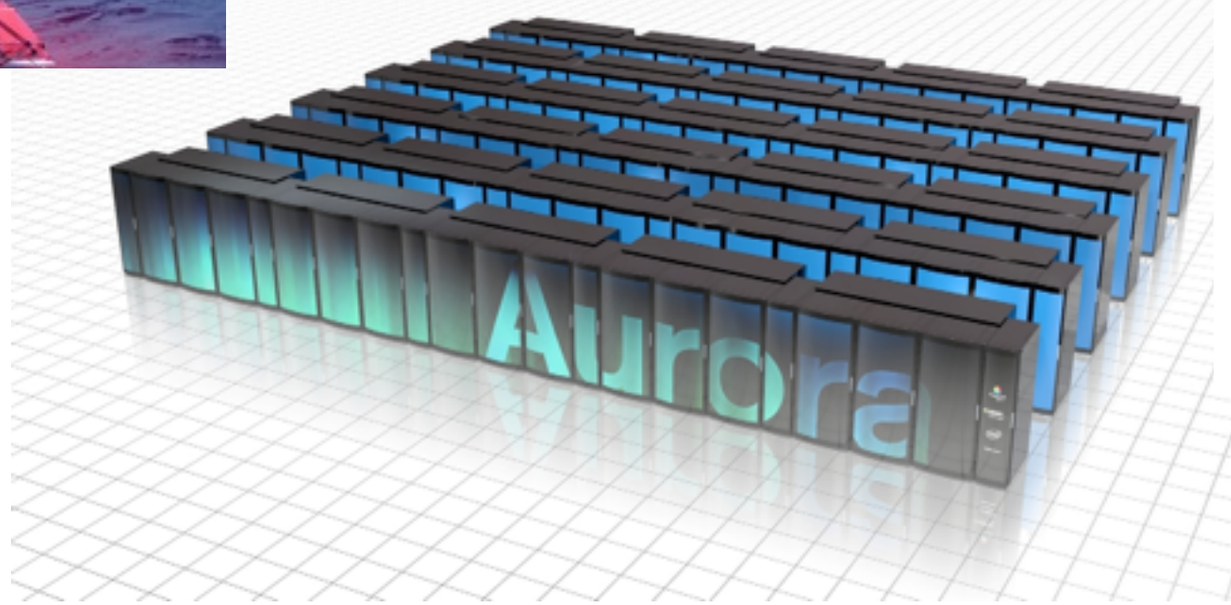
Cosmological Simulations: Under the Hood

Salman Habib
Argonne National Laboratory
Kavli Institute for Cosmological Physics





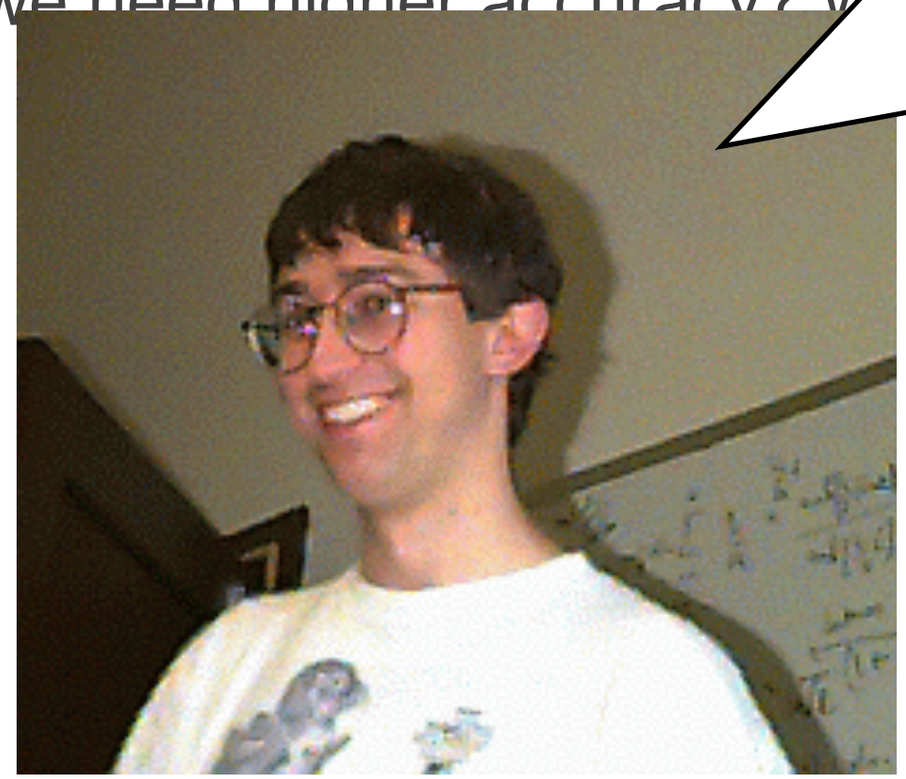


Mexican Numerical Simulations School
Lecture 4, October 6, 2016

Precision Cosmology —

- Cosmology has entered the era of precision science, from order of magnitude estimates to ~few % accuracy mea… of mass content, geometry of the Universe … fluctuations and their normaliza…
- Next step: observations at th… predictions have to keep up!
- Do we need higher accuracy? …

It's the f…… Universe, guys!
It deserves at least two decimal places!

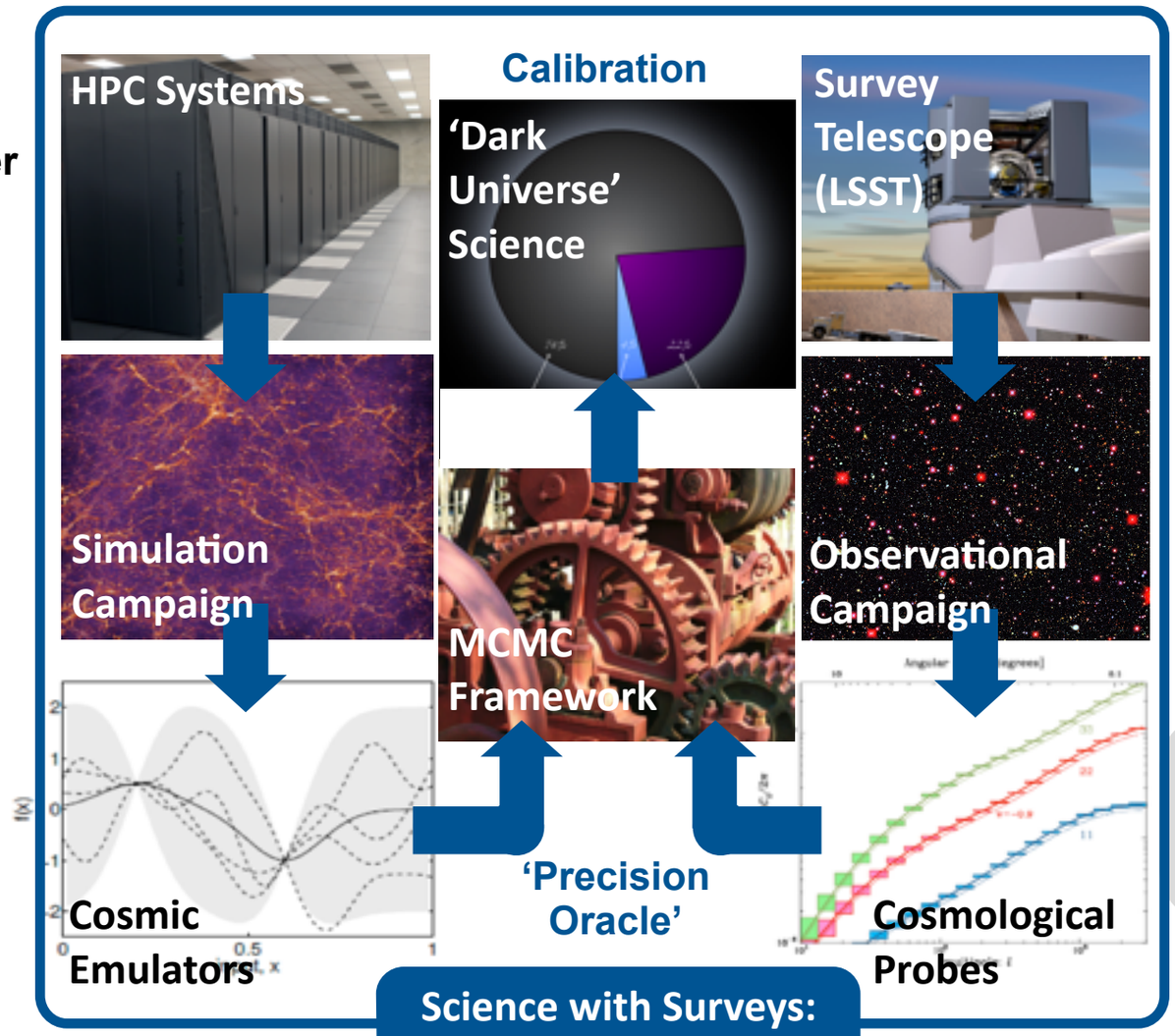


Douglas Scott, UBC
at the Santa Fe Cosmology Workshop (2005)

Precision Cosmology: Theory at Industrial Scales

Supercomputer Simulation Campaign

Simulations + CCF



Mapping the Sky with Survey Instruments

Observations: Statistical error bars will 'disappear' soon!

Measurement Error

Theoretical Prediction

Data Vector

CCF= Cosmic Calibration Framework

$$y = \eta(\theta) + \epsilon$$

$$\pi(\theta|y) \propto L(y|\eta(\theta)) \cdot \pi(\theta)$$

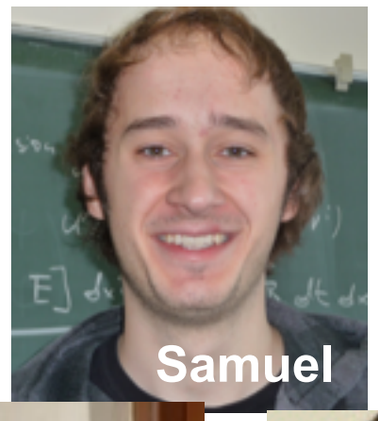
Posterior Distribution

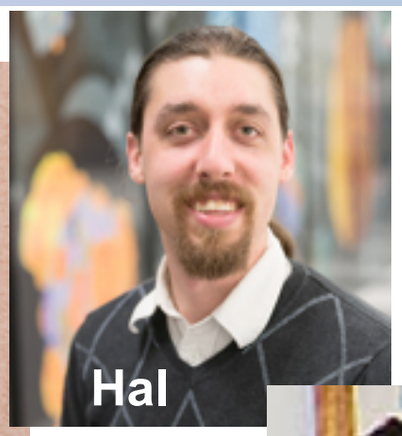
Likelihood Function (not always known!)

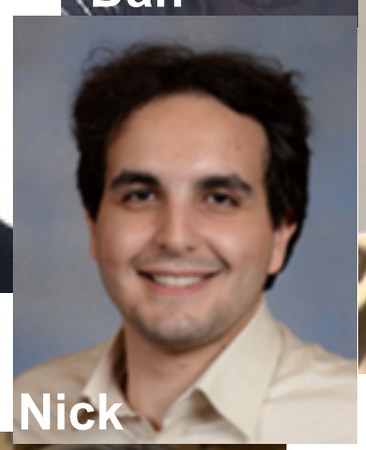
Parameter Prior Distbn.

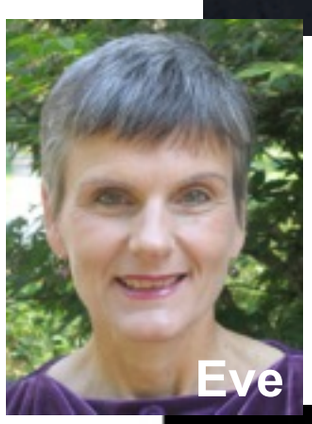
Bayesian Inverse Problem Solution

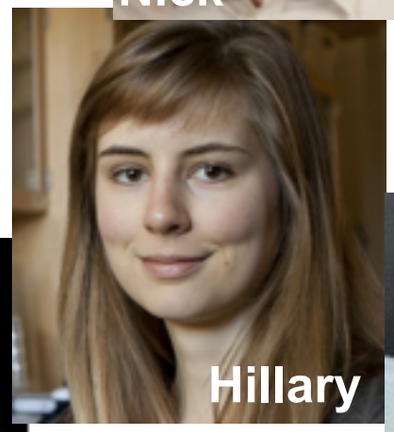
Argonne Cosmology Theory and Computing: People

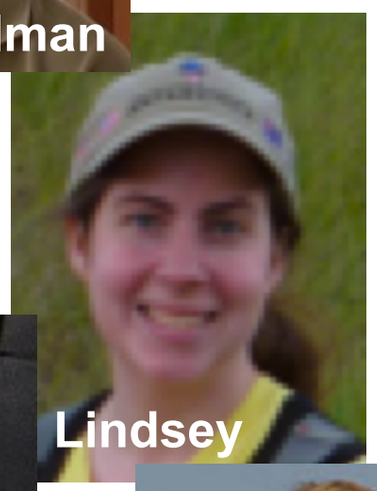


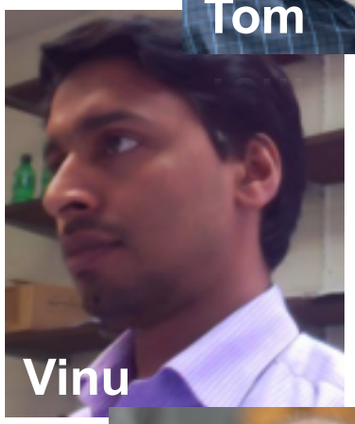


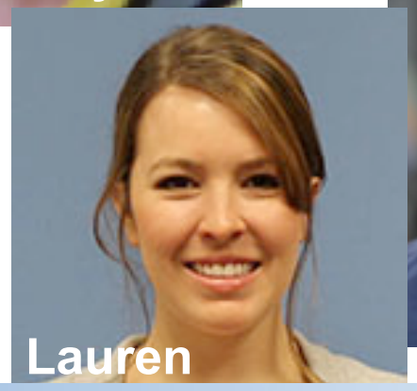












Cosmic Emulation Collaboration

- **The Beginnings — Proof of Concept**

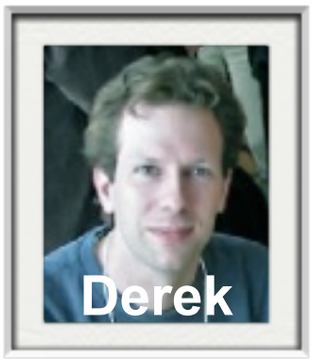
(Heitmann et al. 2006, Habib et al. 2007)

- **The Coyote Universe + Extension**

(Heitmann et al. 2009, 2010, 2013, Lawrence et al. 2010)

- **Emulators beyond P(k)**

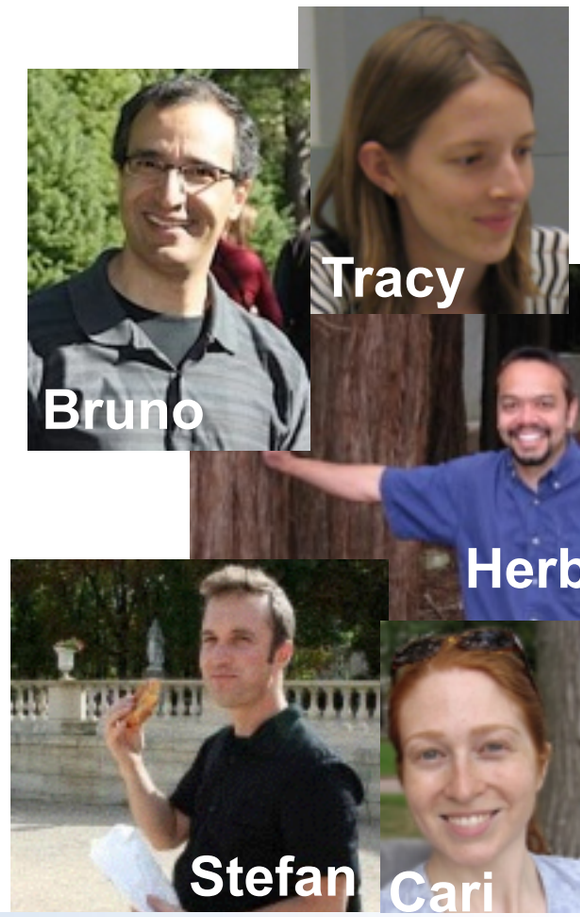
(Kwan et al. 2013a,b)



+ HACC team

- **The Mira-Titan Universe**

(Heitmann et al. 2015)



Key References

- **General Method:**
  - K. Heitmann, D. Higdon, C. Nakhleh, and S. Habib, ApJ Lett **646**, 1 (2006) [short]
  - S. Habib, K. Heitmann, D. Higdon, C. Nakhleh, and B. Williams, Phys. Rev. D **76**, 083503 (2007) [technical]
  - D. Higdon, K. Heitmann, C. Nakhleh, and S. Habib, in the *Oxford Handbook of Applied Bayesian Analysis* edited by O' Hagan and West (Oxford, 2010) [review with a worked out inverse problem]
  - K. Heitmann, D. Bingham, E. Lawrence, S. Bergner, S. Habib, D. Higdon, A. Pope, R. Biswas, H. Finkel, N. Frontiere, and S. Bhattacharya, ApJ **820**, 108 (2016) [nested sampling, strong convergence]

- **Power Spectra:**
  - K. Heitmann, M. White, C. Wagner, S. Habib, and D. Higdon, ApJ **715**, 104 (2010) [Coyote Universe I]; K. Heitmann, D. Higdon, M. White, S. Habib, B.J. Williams, and C. Wagner, ApJ **705**, 156 (2009) [Coyote Universe II]; E. Lawrence, K. Heitmann, M. White, D. Higdon, C. Wagner, S. Habib, and B. Williams, ApJ **713**, 1322 (2010) [Coyote Universe III]; K. Heitmann, E. Lawrence, J. Kwan, S. Habib, and D. Higdon, ApJ **780**, 111 (2014) [Coyote Universe IV]
  - J. Kwan, K. Heitmann, S. Habib, N. Padmanabhan, E. Lawrence, H. Finkel, N. Frontiere, and A. Pope, Phys. Rev. D **810**, 35 (2015) [galaxy power spectrum]

- **Other Examples:**
  - J. Kwan, S. Bhattacharya, K. Heitmann, and S. Habib, ApJ **768**, 123 (2013) [Halo shape emulation]
  - T. Holsclaw, U. Alam, B. Sanso, H. Lee, K. Heitmann, S. Habib, and D. Higdon, Phys. Rev. Lett. **105**, 241302 (2010) [Sn constraints on w(z) using GPs]
  - T. Holsclaw, U. Alam, B. Sanso, H. Lee, K. Heitmann, S. Habib, and D. Higdon, Phys. Rev. D **84**, 083501 (2011) [combining data sets]

Precision Cosmology: Statistics + Machine Learning

Old School Stats: Limited Data/Computing
('benign' black box,
controlled environment)

Find the (simple)
'data model'

Machine Learning: 'Big Data'
('vicious' black box,
potentially uncontrolled environment)

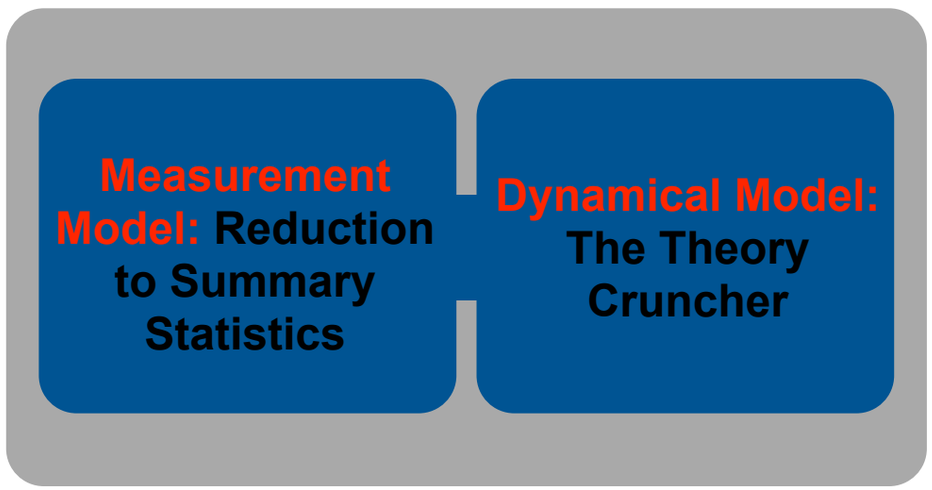
Find the (arbitrary)
'dynamical map'

Both (cartoon) approaches useful but by no means exhaustive in terms of scientific value--
with a good theory (forward model) and sufficient computing the box is not black!
We are not 'model-free' --

Statistics meets Machine Learning in the Supercomputing/Big Data Environment

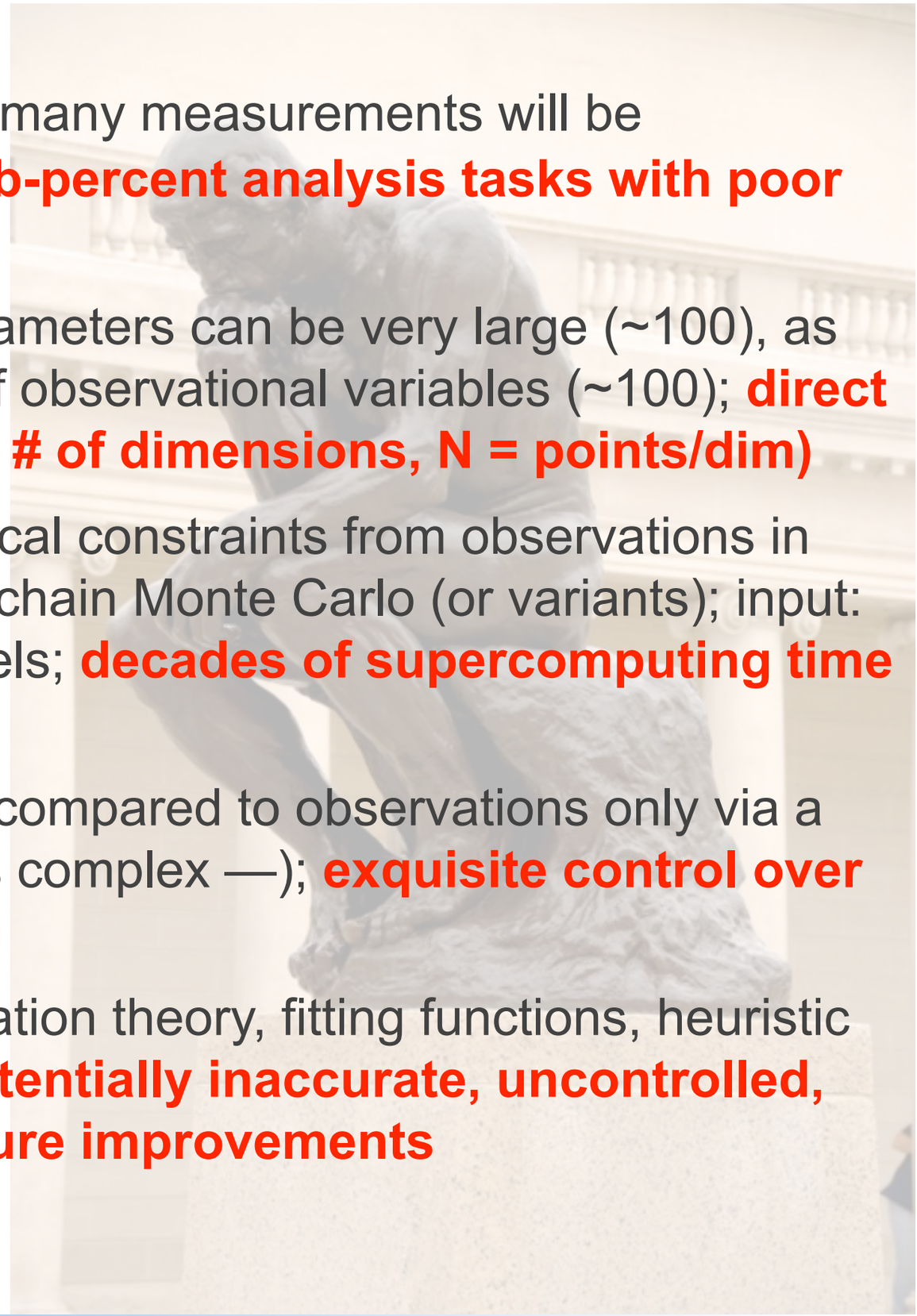
Modeling cycle: Predict, (Cross) Validate, Optimize

Summary Statistics
(restricted to what can be measured)



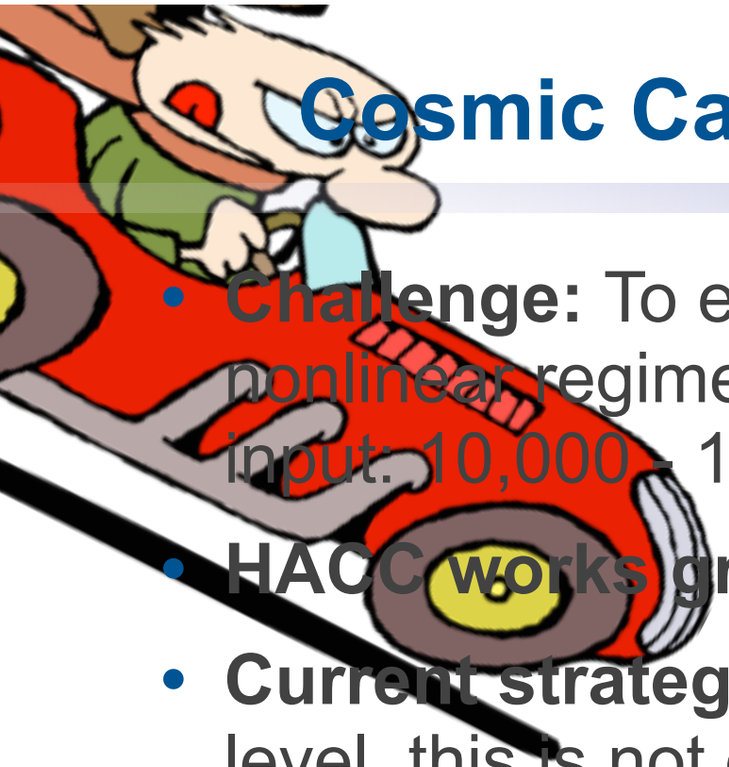
Theory and Modeling Inputs
(prior ranges)

Cosmic Calibration: The Challenge



- **Subtle Signal:** ΛCDM already very good, many measurements will be systematics limited; **need to carry out sub-percent analysis tasks with poor prognosis for (near future) validation**
- **Curse of Dimensionality:** Number of parameters can be very large (~100), as well as measured sampling of a number of observational variables (~100); **direct representation hopeless (N^d, where d = # of dimensions, N = points/dim)**
- **Complex Modeling:** To extract cosmological constraints from observations in the nonlinear regime, need to run Markov chain Monte Carlo (or variants); input: 100,000 - 1000,000 different forward models; **decades of supercomputing time needed to solve via brute force**
- **Discrepancy Analysis:** Modeling can be compared to observations only via a phenomenological filter (— astrophysics is complex —); **exquisite control over systematic errors crucial (very difficult)**
- **Criticism of Current Strategies:** Perturbation theory, fitting functions, heuristic models, computational short-cuts, etc.; **potentially inaccurate, uncontrolled, subject to bias, limited potential for future improvements**

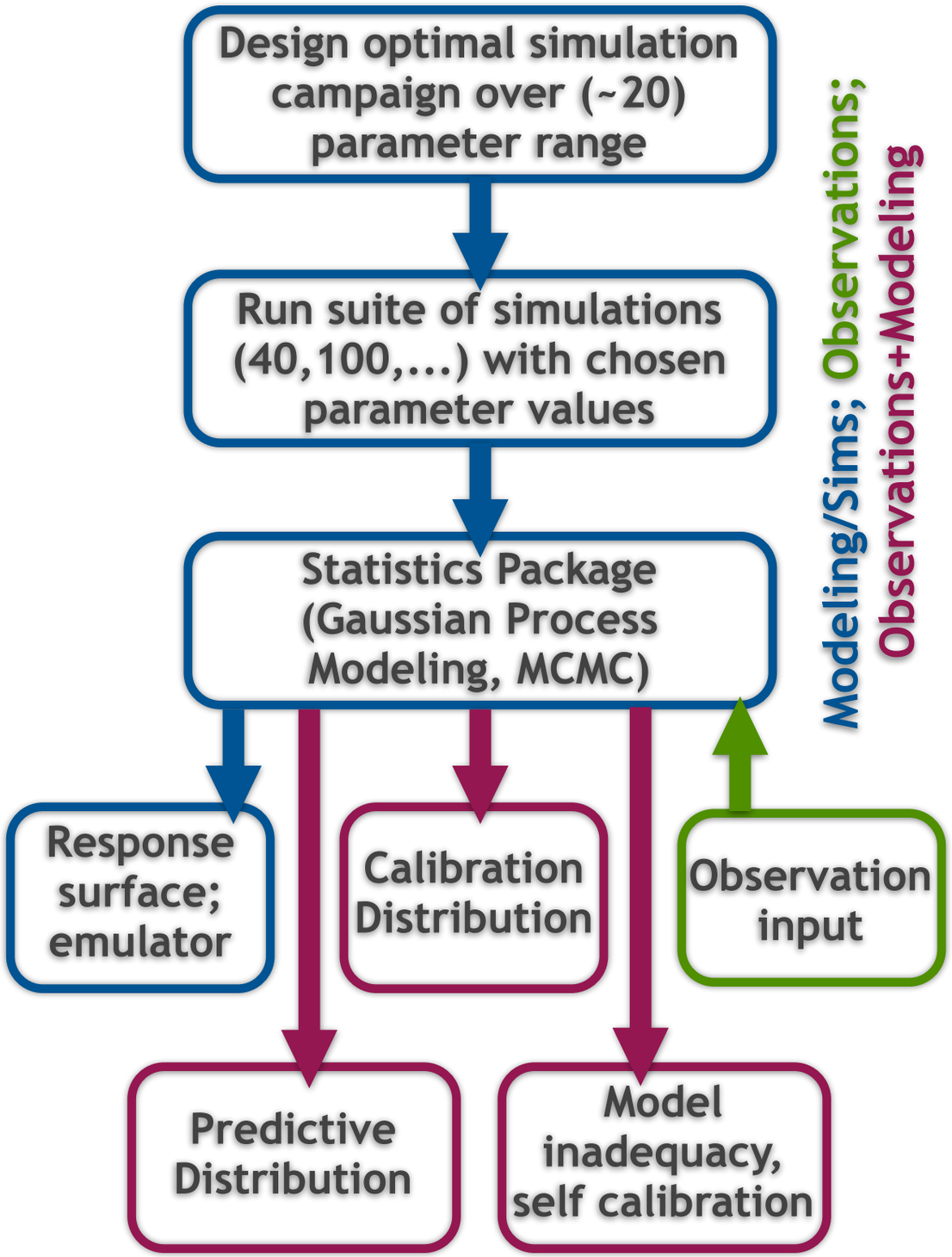
Cosmic Calibration: Solving the Inverse Problem



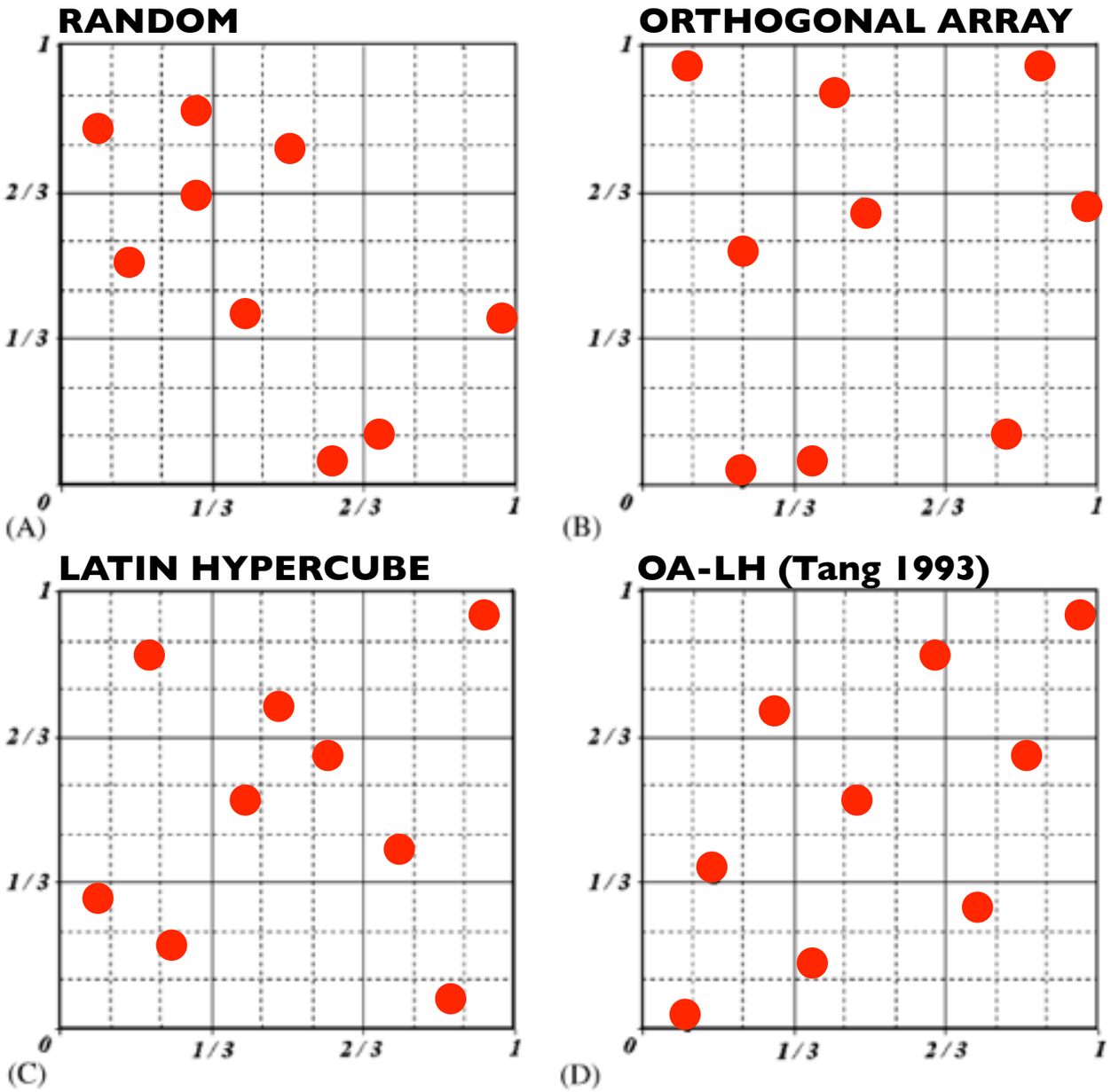
- **Challenge:** To extract cosmological constraints from observations in nonlinear regime, need to run Markov chain Monte Carlo (MCMC) code; input: 10,000 - 100,000 different models (or even more)
- **HACC works great!** But still cannot generate simulations in seconds ...
- **Current strategy:** Fitting functions for e.g. P(k), accurate at the 10% level, this is not good enough!
- **Our alternative:** Emulators, fast and accurate prediction schemes built from a manageably finite set of simulations
- "**Ingredients": 1)** Optimal sampling methods to decide which models to simulate, **2)** efficient representation of simulation outcome, **3)** powerful interpolation scheme
- **Example here: Power spectrum emulator**
  - Step 1: Show simulations have required accuracy (discussed already)
  - Step 2: Minimum number of simulations? Interpolation scheme that provides the power spectrum for any cosmology within a given parameter space prior
  - Step 3: Carry out simulation and build final emulator

Cosmic Calibration Framework

- **Step 1:** Design simulation campaign, rule of thumb: O(10) models for each parameter ('Higdon's Rule')
- **Step 2:** Carry out simulation campaign and extract quantity of interest, in our case, the power spectrum
- **Step 3:** Data transformation (e.g., smoothing) and reduction (here via an ***empirical Principal Components*** basis expansion)
- **Step 4:** Choose suitable interpolation scheme to interpolate between models, here ***Gaussian Processes***
- **Step 5:** Build emulator (unusual application of Bayesian ideas)
- **Step 6:** Use emulator to analyze data, determine model inadequacy, refine simulation and modeling strategy...

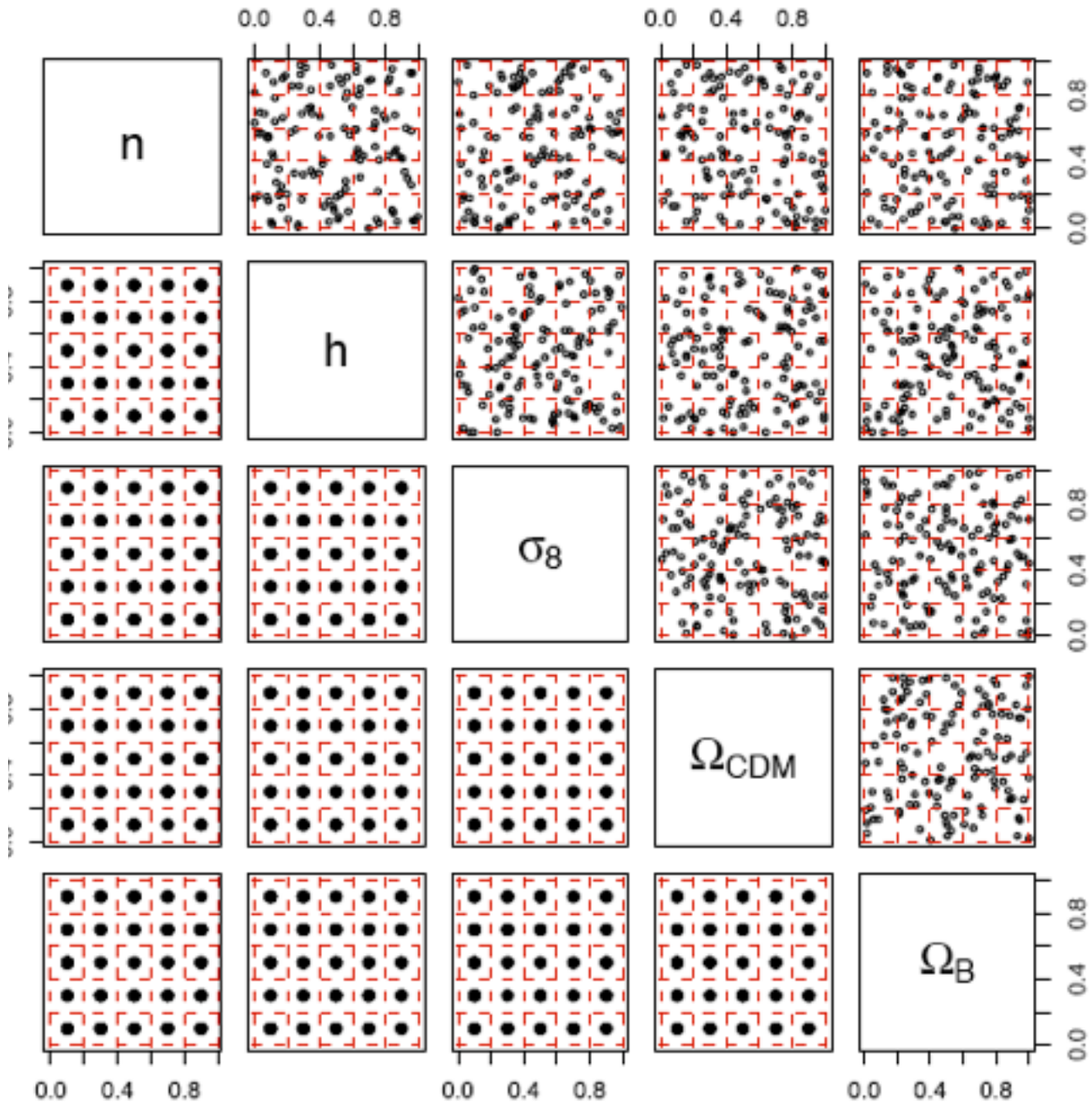


Sampling Parameter Space



Sandor & Andras 2003

Strive for “equidistribution” property over the sampling space, best approach when ignorant of functional variation, well-suited to GPMs

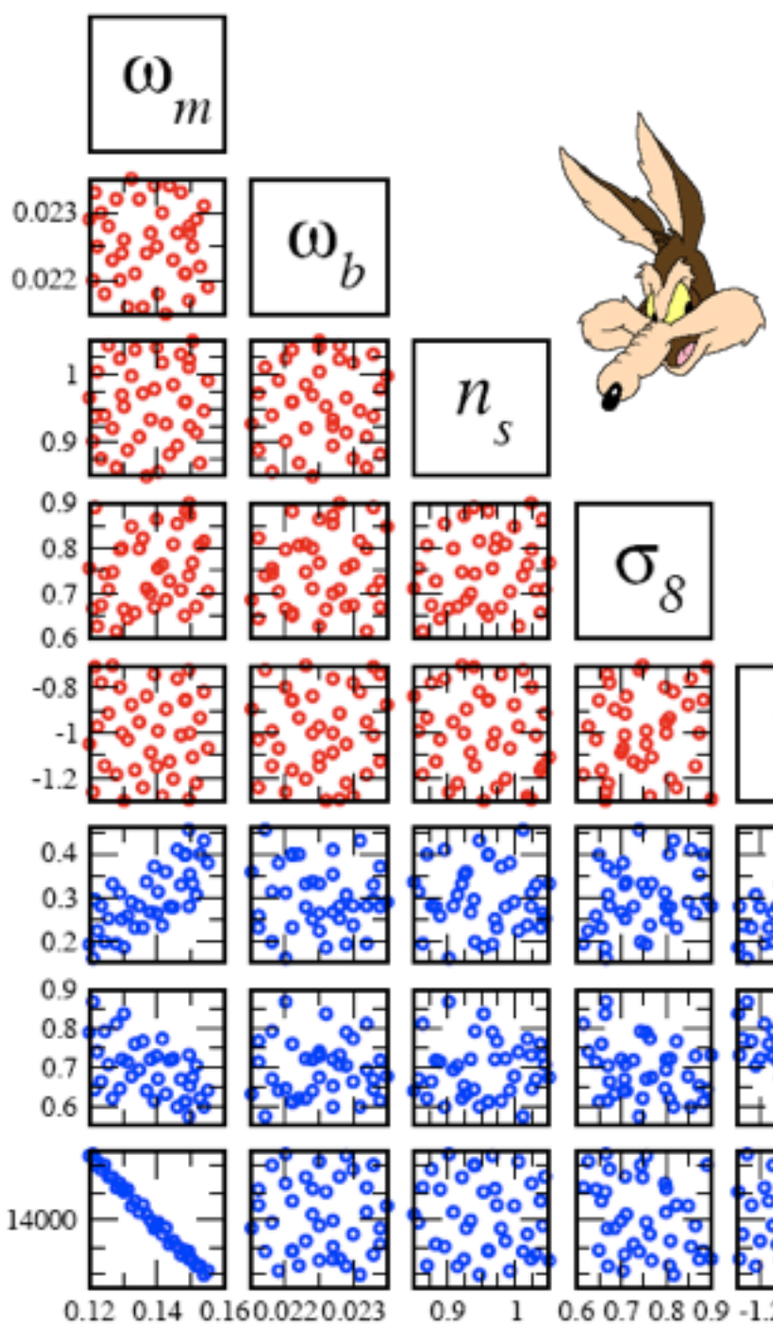


Practical 128 point, 5-level, strength-2-based design [level=#variable slices, strength=(lower) dimension to be sampled, #columns=#variables, #rows=#trials, **significant computation in its own right**]

The Coyote Simulation Design for wCDM Cosmologies

- **Observational considerations**
  - Planck provides very accurate measurements of “vanilla parameters”
  - In particular, ω_b , ω_m, n_s known at the 2-3% level
  - w, σ_8 less well known
  - Determine best-fit value for *h* for each model from distance to surface of large scattering, known at 0.3% (in later emulator release, *h* is a free parameter)
- **For good emulator performance from very small number of runs**
  - Not too broad priors
  - Not too many parameters

The (original) Coyote Universe



Priors:

$0.020 \leq \omega_b \leq 0.025$

$0.11 \leq \omega_m \leq 0.15$

$0.85 \leq n_s \leq 1.05$

$-1.3 \leq w \leq -0.7$

$0.6 \leq \sigma_8 \leq 0.9$

Design Parameters

Derived Parameters

The Coyote Universe

Priors:

$0.020 \leq \omega_b \leq 0.025$

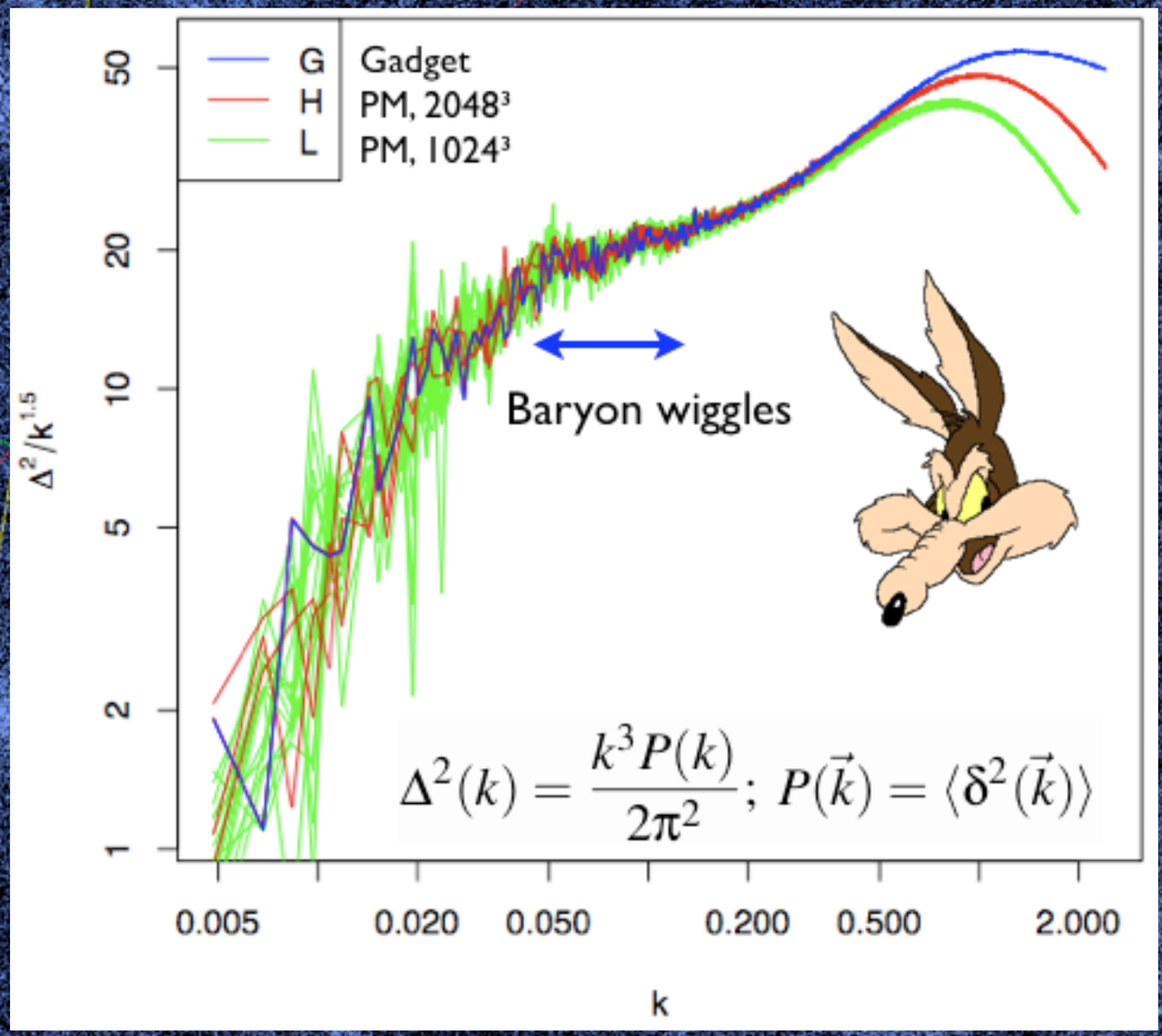
$0.11 \leq \omega_m \leq 0.15$

$0.85 \leq n_s \leq 1.05$

$-1.3 \leq w \leq -0.7$

$0.6 \leq \sigma_8 \leq 0.9$

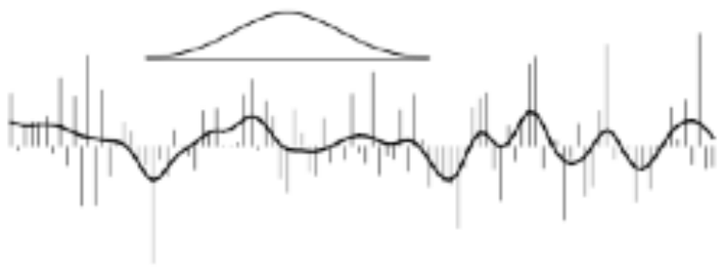
- **37 model runs + ΛCDM**
  - 16 low resolution realizations (green)
  - 4 medium resolution realizations (red)
  - 1 high resolution realization (blue)
  - 11 outputs per run between z = 0 - 3
- **Restricted priors to minimize necessary number of runs**
- **1.3 Gpc boxes, mp ~$10^{11} M_\odot$**
-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



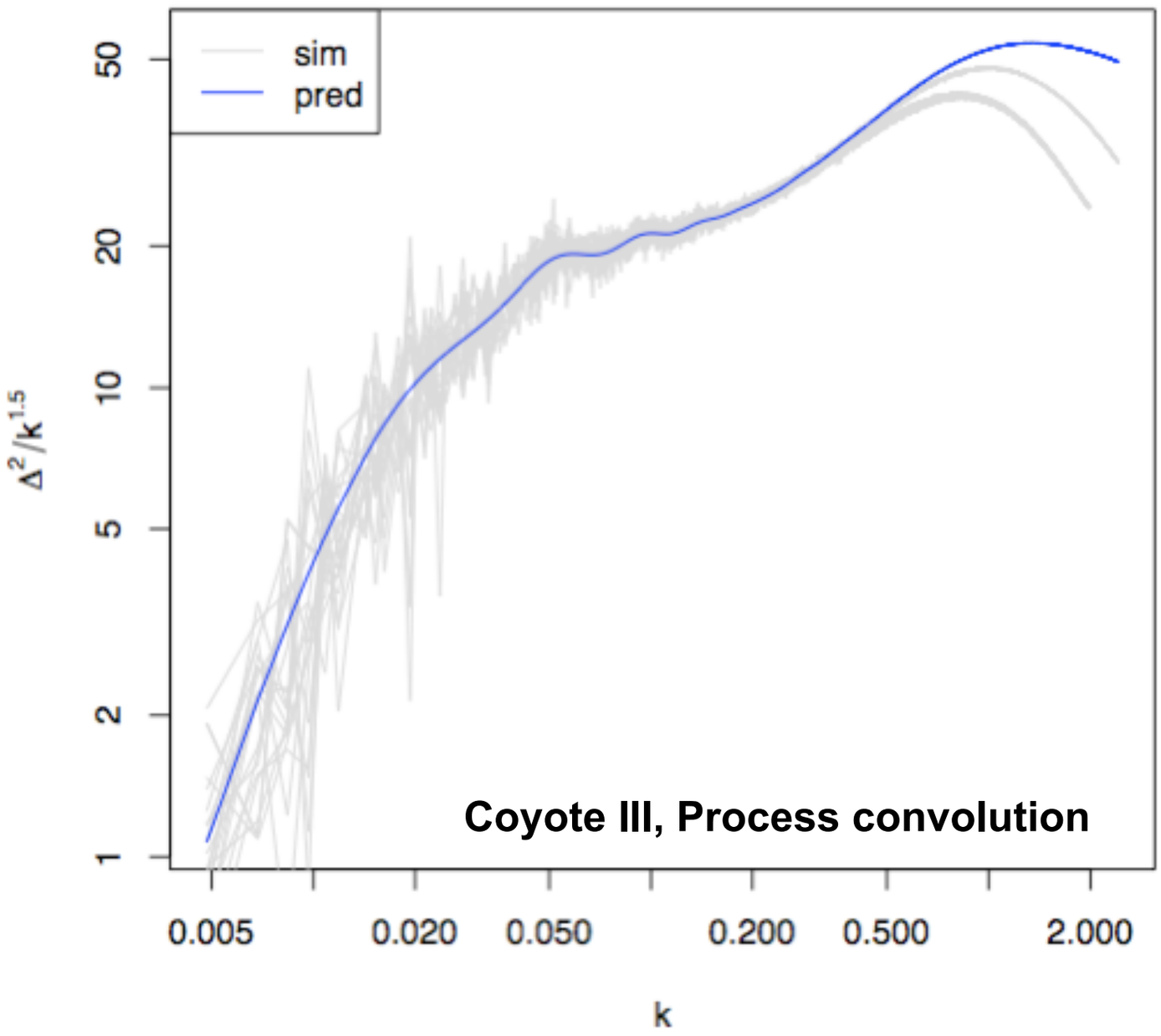
Background Visualization with ParaView by J. Woodring

Next step: Smoothing the Power Spectrum

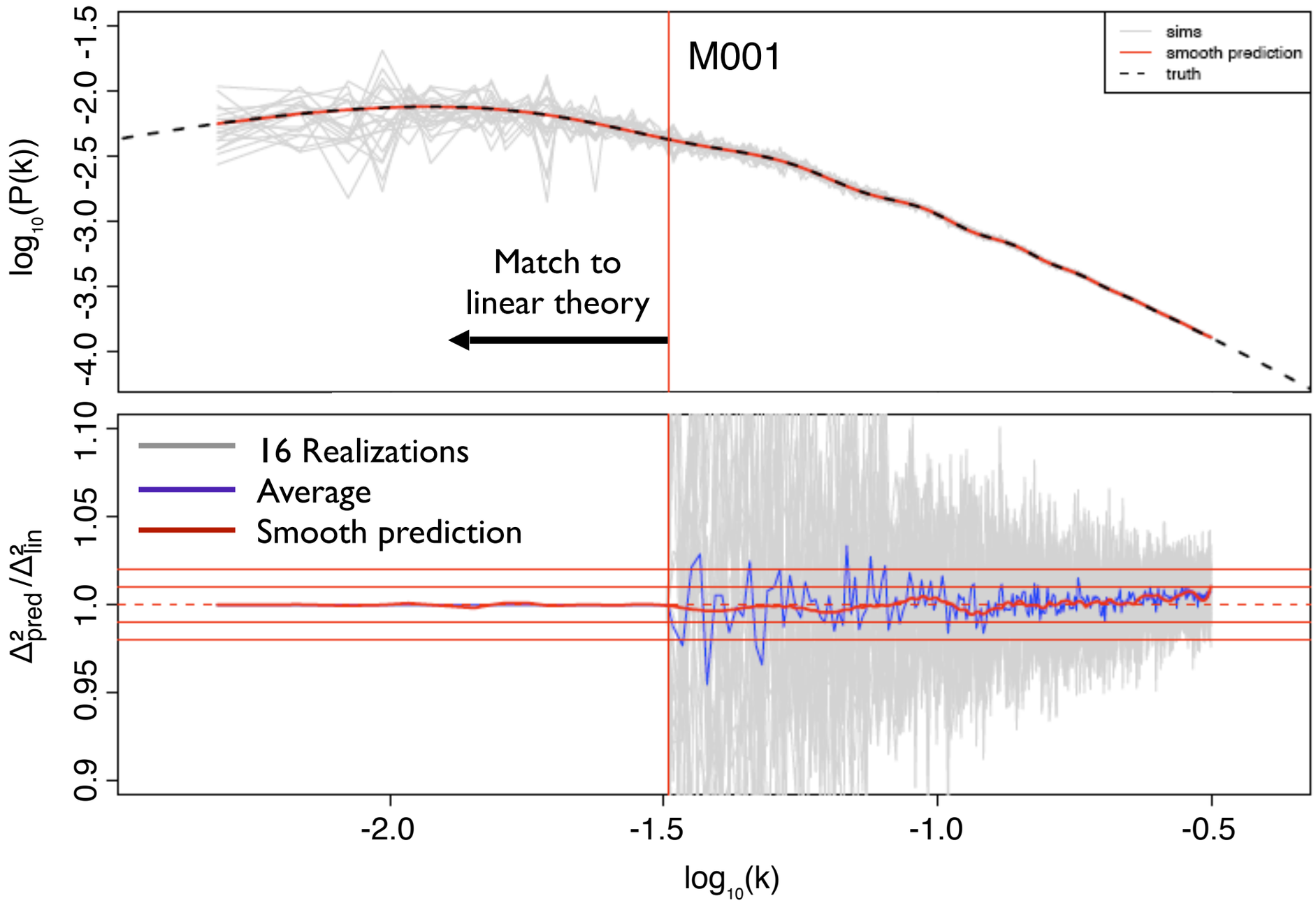
- Each simulation represents one possible realization of the Universe in a finite volume
- Need smooth prediction for building the emulator for each model
- Major challenge: Make sure that baryon features are not washed out or enhanced due to realization scatter
  - Construct smooth power spectra using a process convolution model (Higdon 2002)



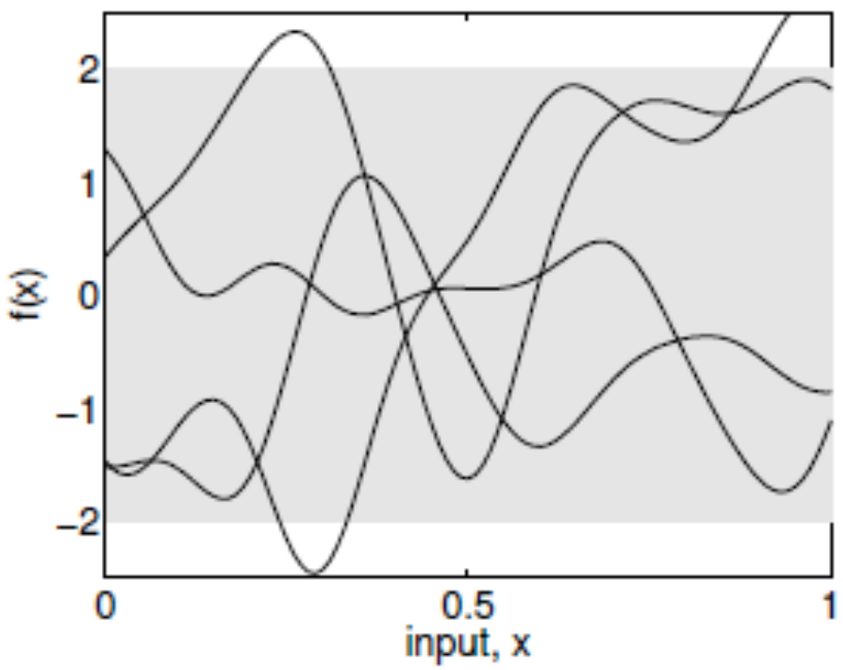
  - Basic idea: calculate moving average using a kernel whose width is allowed to change to account for nonstationarity



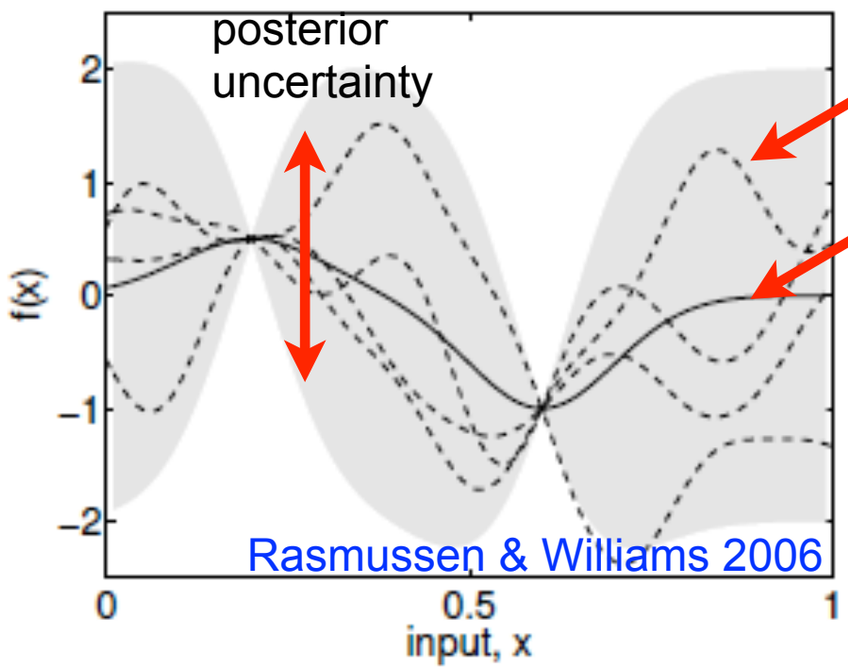
Test on the Linear Power Spectrum



Gaussian Process Modeling



Prior distribution over random functions: global mean zero (although individual choices clearly are not mean-zero), variance assumed to be independent of x, 2-SD band in gray



Posterior distribution conditioned on exact information at two x points, consider only those functions from the prior distribution that pass through these points

In actual practice, use data compression via PCA and avoid overfitting by using priors on hyperparameters and by controlling the (MCMC-based) learning process

Can solve both regression and reconstruction (statistical inverse) problems

GPs are nonparametric -- no need to worry if the functions can fit the data (e.g., linear functions against nonlinear data), even with many observations still have plenty of candidate functions

With GP models, the choice of prior distribution over random functions is essentially a statement of the properties of the initial covariance function, these properties can be specified in terms of a set of hyperparameters, using data to determine these then defines the learning problem for the GP approach

Gaussian Processes: Basics I

GPs are straightforward generalizations of Gaussian distributions over vectors to function spaces, and are specified by a ***mean function*** and a ***covariance function***

$$\mathbf{f} = (f_1, \ldots, f_n)^T \sim \mathcal{N}(\mu, \Sigma)$$

$$cov(f(\mathbf{x}), f(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}')$$

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

They have several convenient properties, of which the two most significant are

- ***Marginalization*** yields a Gaussian distribution

$$p(\mathbf{y_a}) = \int p(\mathbf{y_a}, \mathbf{y_b}) d\mathbf{y_b}$$

$$p(\mathbf{y_a}, \mathbf{y_b}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}\right) \Longrightarrow p(\mathbf{y_a}) = \mathcal{N}(\mathbf{a}, \mathbf{A})$$

Gaussian Processes: Basics II

- ***Conditioning*** yields a new Gaussian distribution

$$p(\mathbf{y_a}|\mathbf{y_b}) = \frac{p(\mathbf{y_a}, \mathbf{y_b})}{p(\mathbf{y_b})}$$

$$p(\mathbf{y_a}, \mathbf{y_b}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}\right)$$

$$\Longrightarrow p(\mathbf{y_a}|\mathbf{y_b}) = \mathcal{N}(\mathbf{a} + \mathbf{B}\mathbf{C}^{-1}(\mathbf{y_b} - \mathbf{b}), \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T)$$

The result also holds for conditioning with Gaussian errors. This property is important because it means that conditioning can be carried out “analytically”, without a brute force rejection algorithm being employed.

Note, however, that a ***matrix inversion*** is required for this step. This is one aspect of the “curse of dimensionality” in regression/inverse problems. Ideas on how to deal with this issue are at the cutting edge of current research.

Covariance Function I

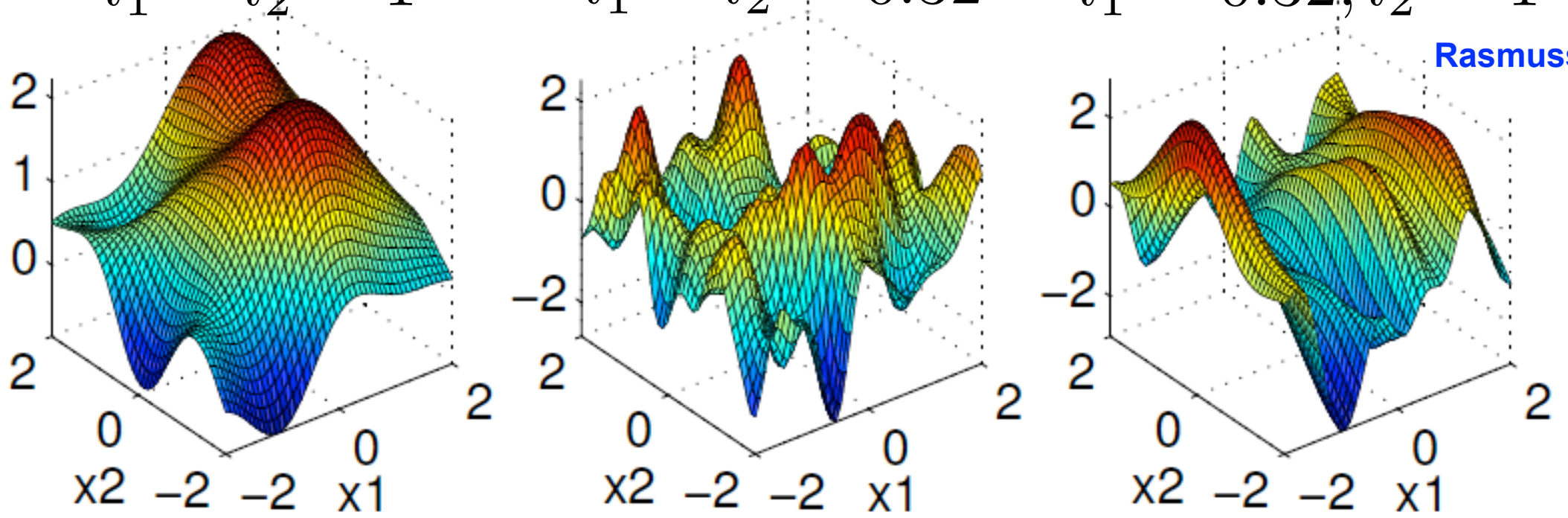
The (***symmetric, positive semi-definite***) covariance function is the key ingredient in GP modeling. Depending on the application, various choices of the covariance function are possible, both in terms of the ***form*** and the underlying ***parameters***

The ***squared exponential*** form is very common:

$$k_{SE}(r) = \exp\left(-\frac{r^2}{2l^2}\right)$$

here l defines a characteristic length scale; the realizations are infinitely differentiable (possibly unrealistic?)

$l_1 = l_2 = 1$ $\quad$ $l_1 = l_2 = 0.32$ $\quad$ $l_1 = 0.32, l_2 = 1$



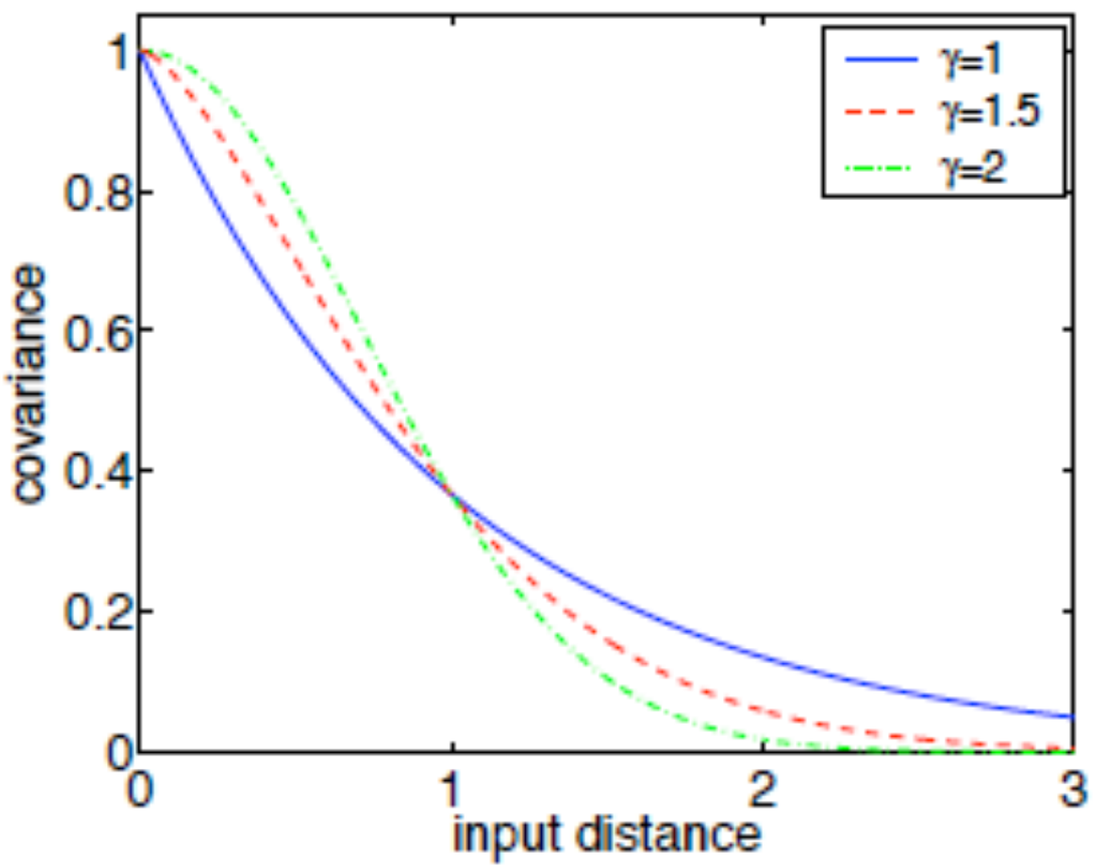
Rasmussen 2006

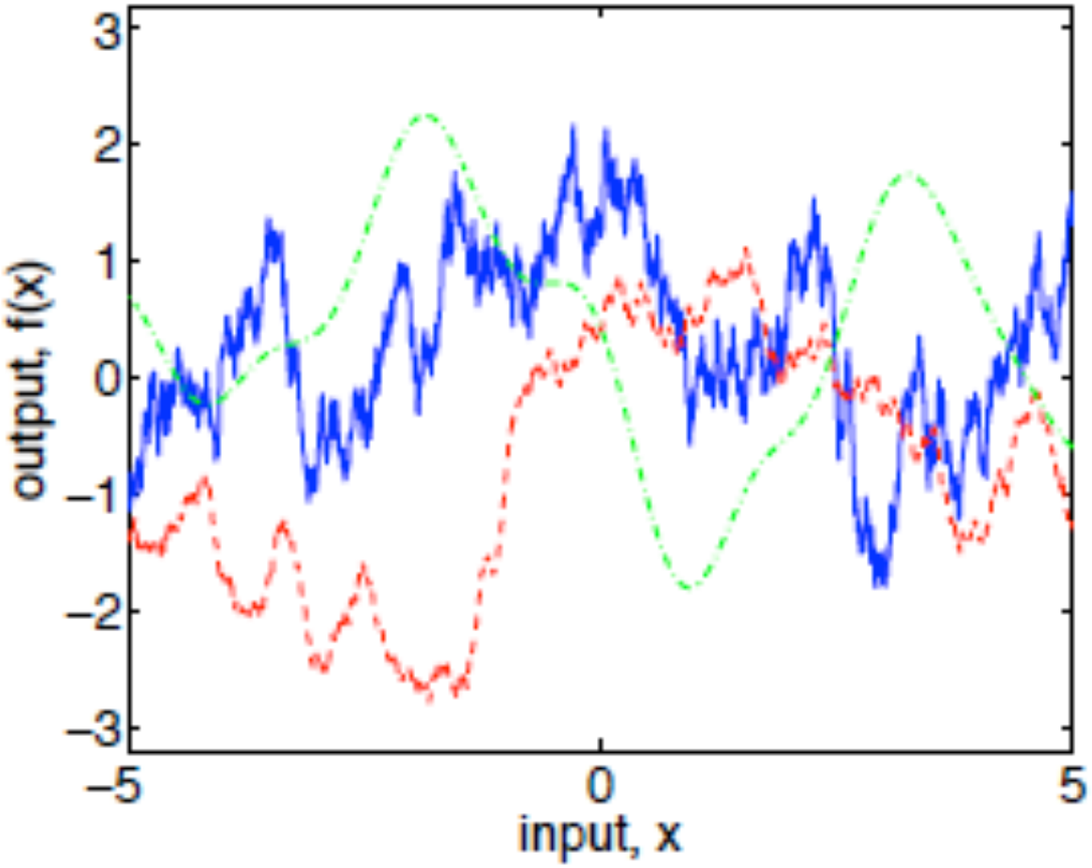
Covariance Function II

The ***Gamma-exponential*** form

$$k_{GE}(r) = \exp\left(-(r/l)^{\gamma}\right), \ 0 < \gamma \leq 2$$

corresponds to an Ornstein-Uhlenbeck process in one dimension for unit exponent, where it yields continuous but non-(MS) differentiable functions, except in the squared-exponential limit (Matern cov. fn. is smoother)

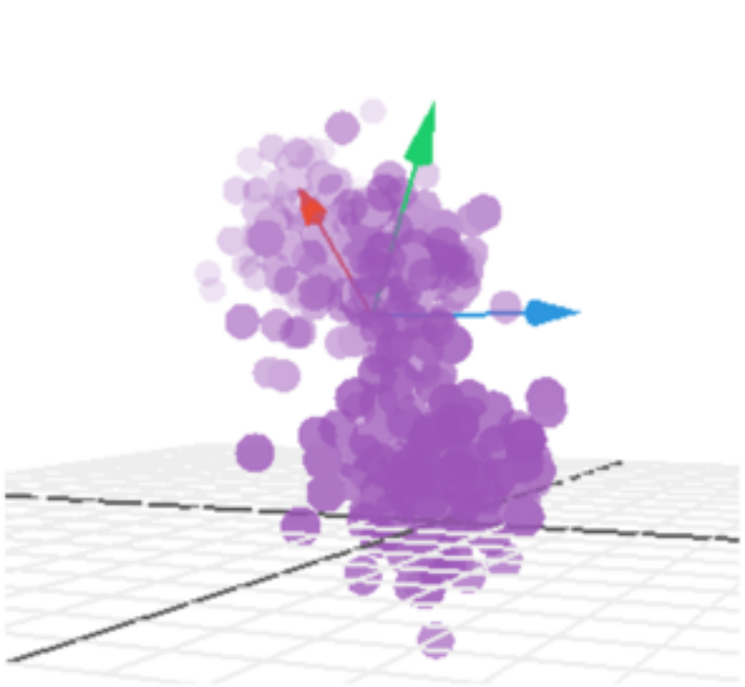


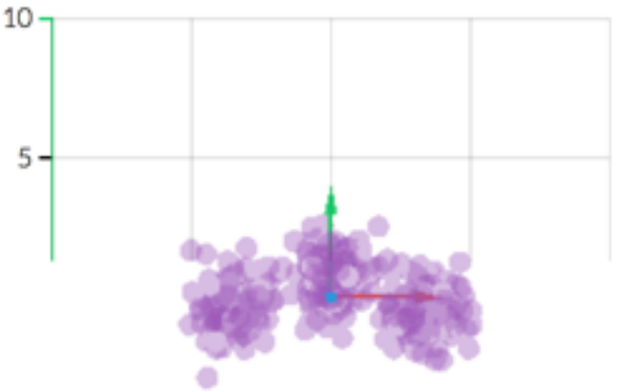


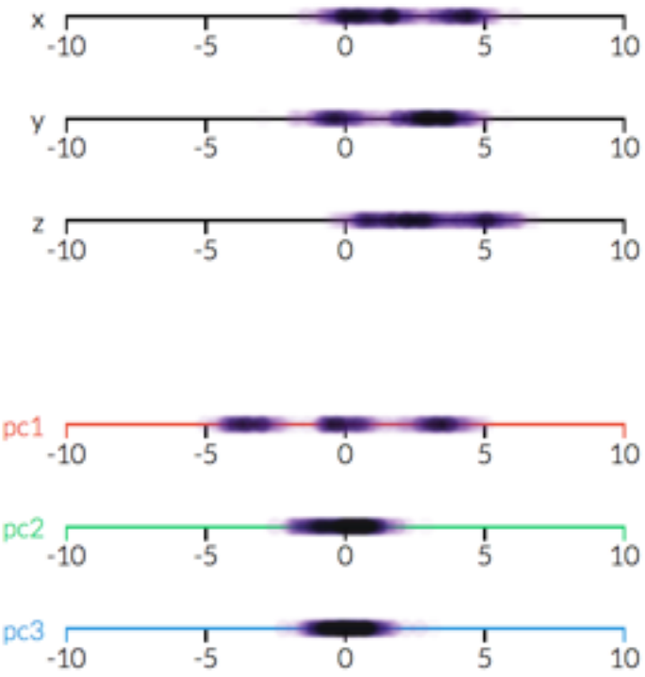
Rasmussen & Williams 2006

Principal Components I

- Principal Component Analysis (PCA) is a technique to emphasize variations and bring out strong patterns in a dataset (we use it for data reduction)
- In example below, find “camera” angle to project the data in such a way that the first principal component (PC1) captures most of the variation and every sub-sequent PC has less and less variation







- In our case: create a $n_{kz} \times m$ matrix, carry out a singular value decomposition …

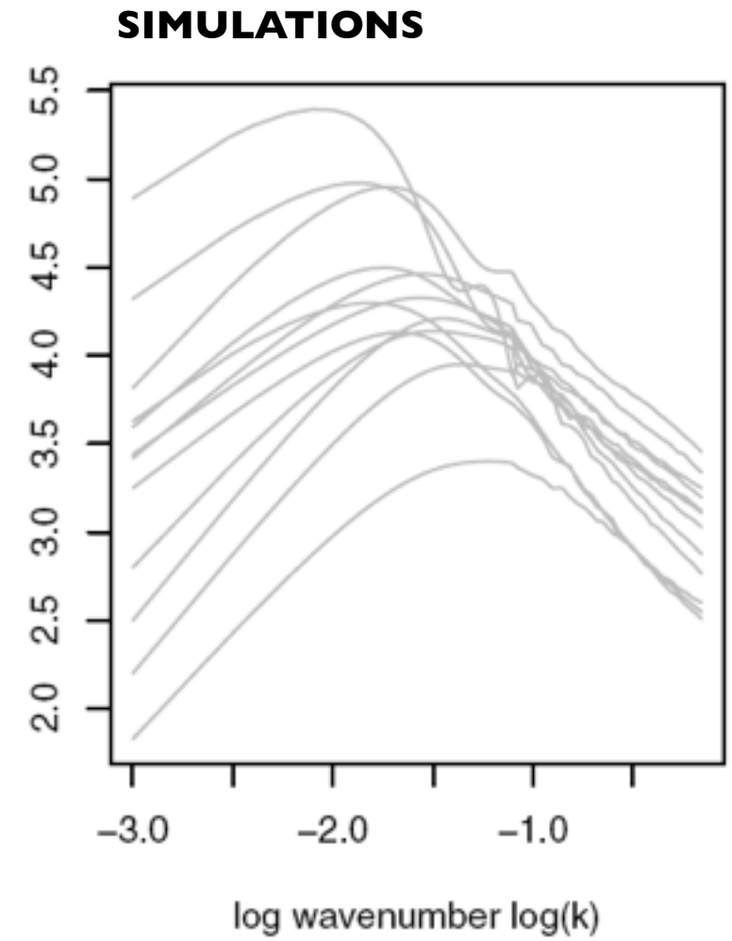
power spectra of *k*-*z* grid

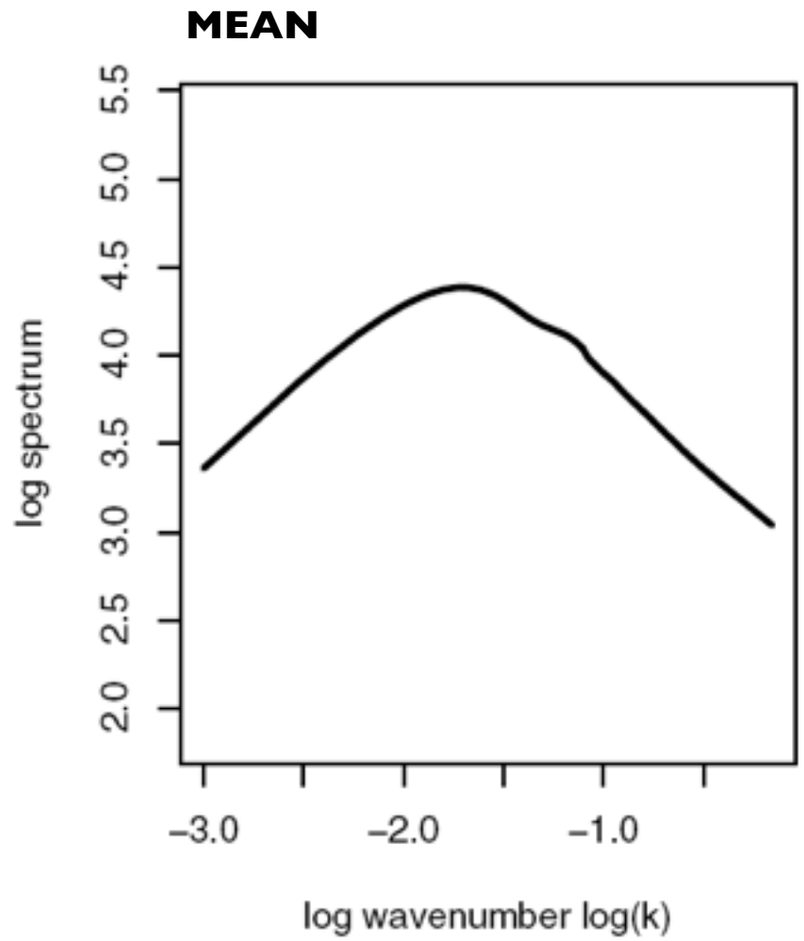
number of models

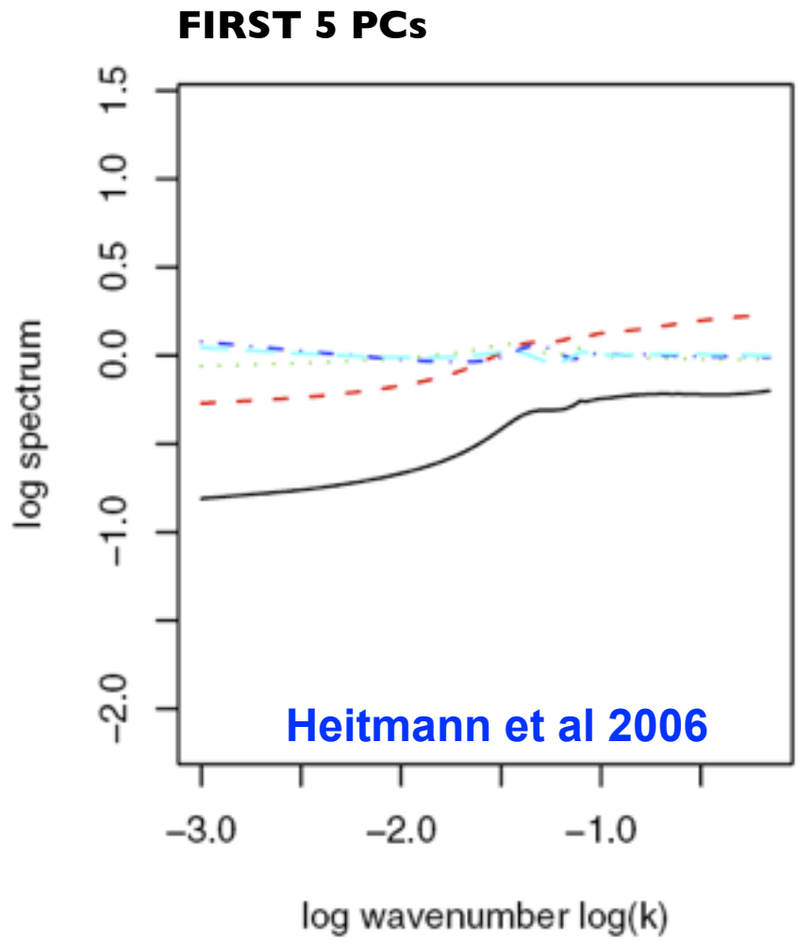
from http://setosa.io/ev/principal-component-analysis/

Principal Components II

P(k) example







Mean-adjusted Principal Component Representation

$$\eta(k;\theta) = \sum_{i=1}^{p_\eta} \phi_i(k) w_i(\theta) + \epsilon, \qquad \theta \in [0,1]^{p_\theta}$$

COSMOLOGICAL/MODELING PARAMETERS

PC BASIS FUNCTIONS

GP WEIGHTS

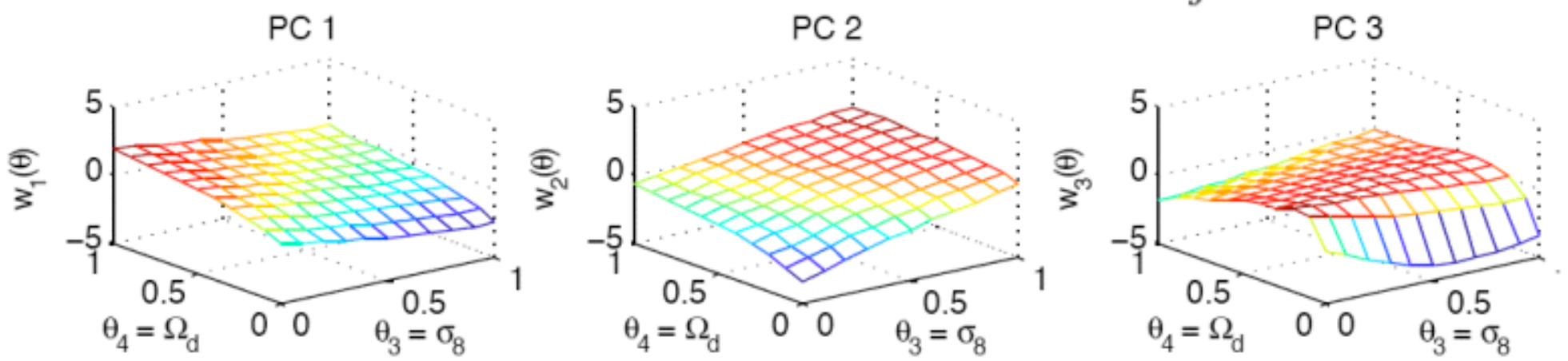
STANDARDIZED PARAMETER DOMAIN

Emulator Predictions (Includes Sensitivity Analysis)

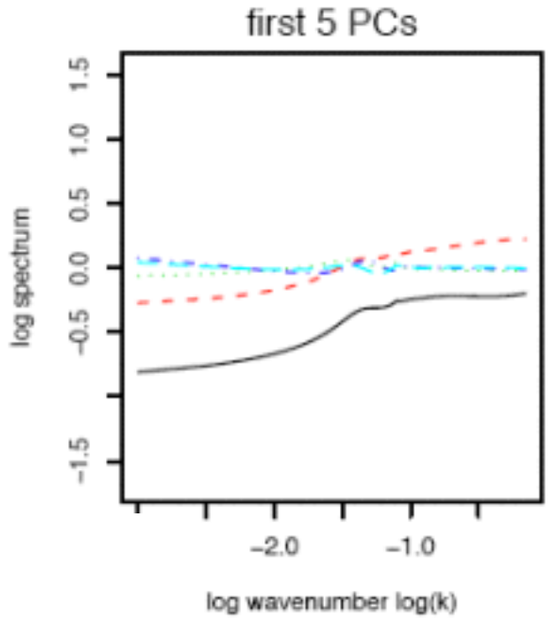
Gaussian process (GP) models are used to estimate the weights $w_j(\theta)$ at untried settings

$$\widehat{\eta}(\theta;k) = \sum_{j=1}^{p_\eta} w_j(\theta)\phi_j(k)$$

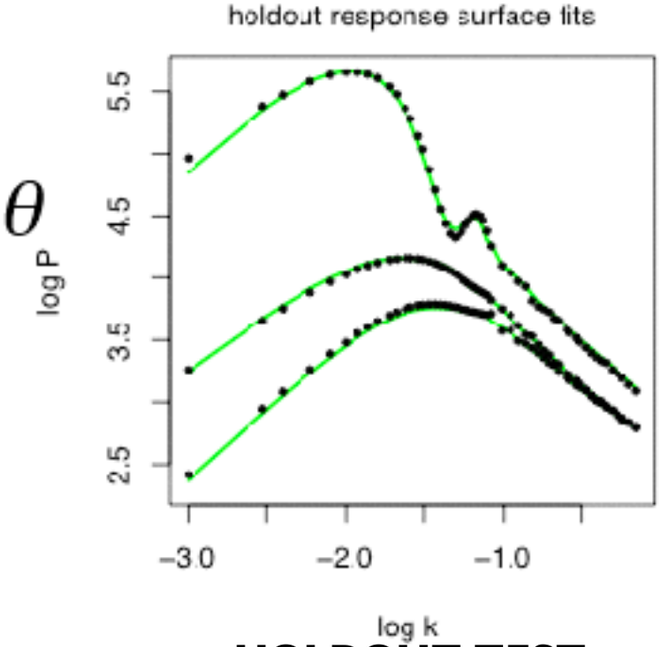
GP models $w_j(\theta)$



Bases $\phi_j(\theta)$

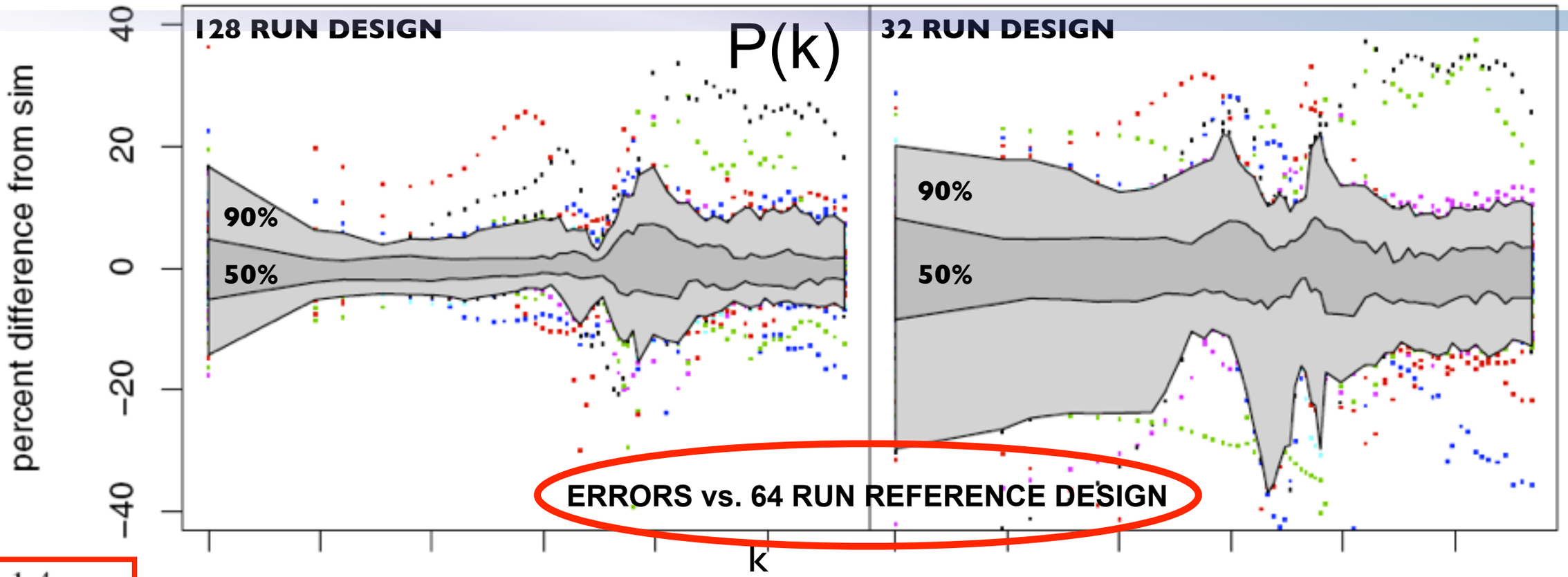


Prediction at new θ



HOLDOUT TEST

Convergence



$0.8 \leq n \leq 1.4,$

$0.5 \leq h \leq 1.1,$

$0.6 \leq \sigma_8 \leq 1.6,$

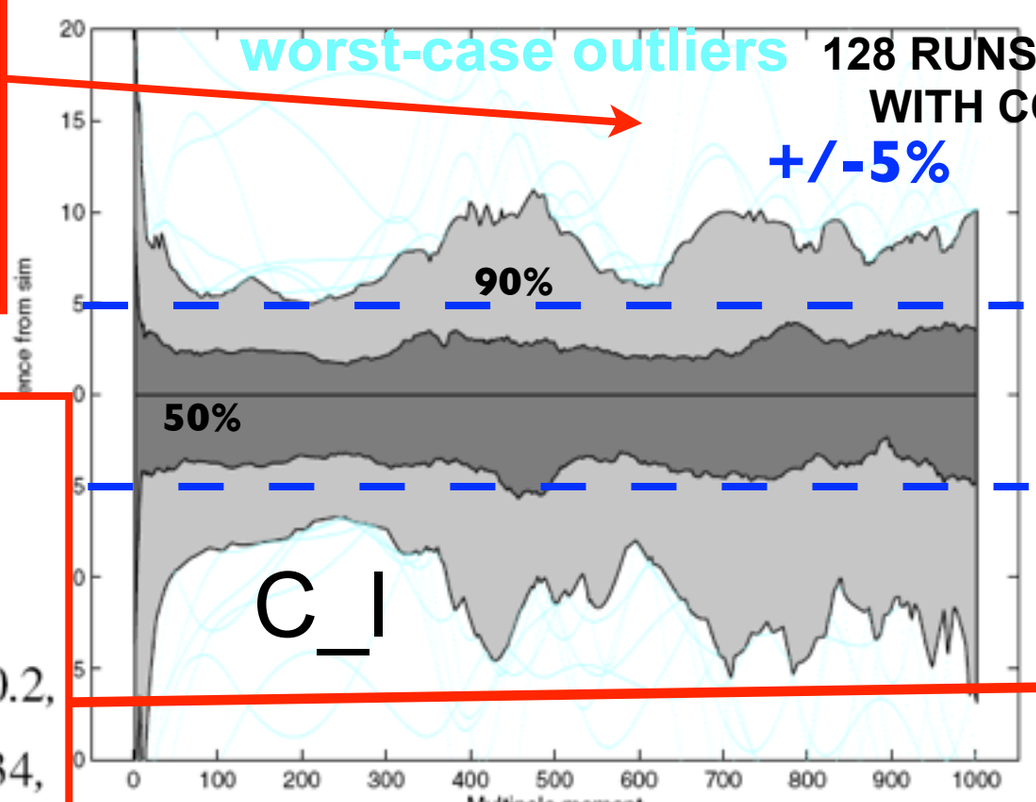
$0.05 \leq \Omega_{\rm CDM} \leq 0.6,$

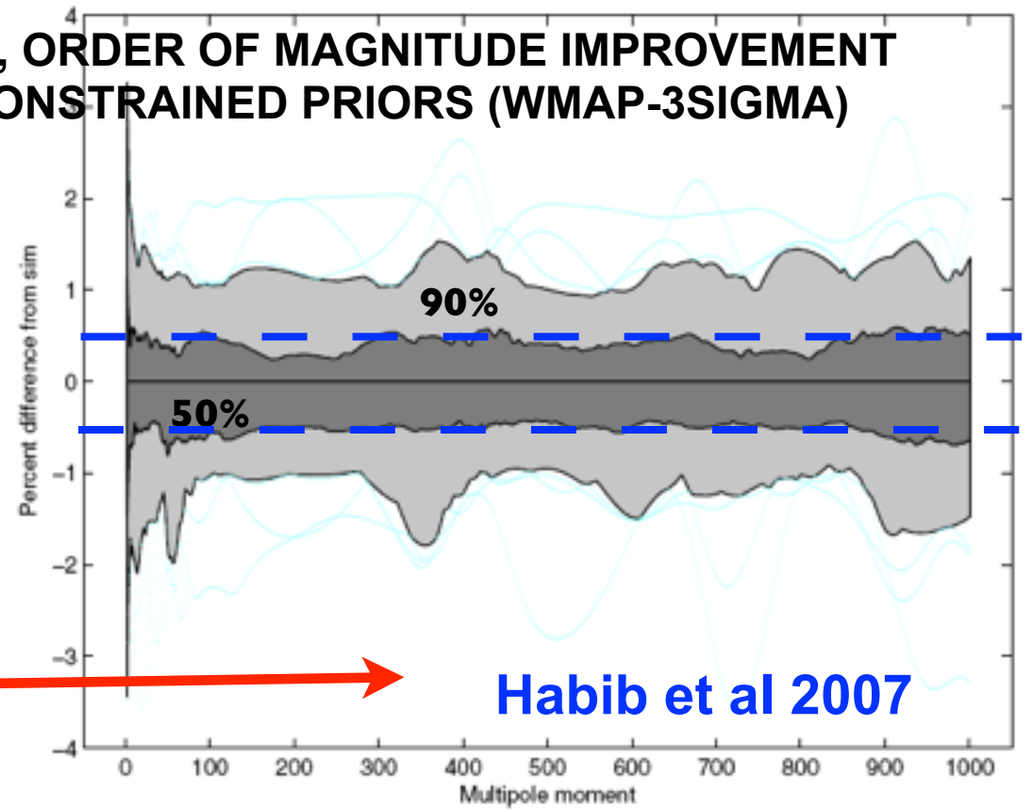
$0.02 \leq \Omega_{\rm b} \leq 0.12.$

worst-case outliers

128 RUNS, ORDER OF MAGNITUDE IMPROVEMENT WITH CONSTRAINED PRIORS (WMAP-3SIGMA)

+/-5%





$0.85 \leq n \leq 1.25,$

$0.6 \leq h \leq 0.9,$

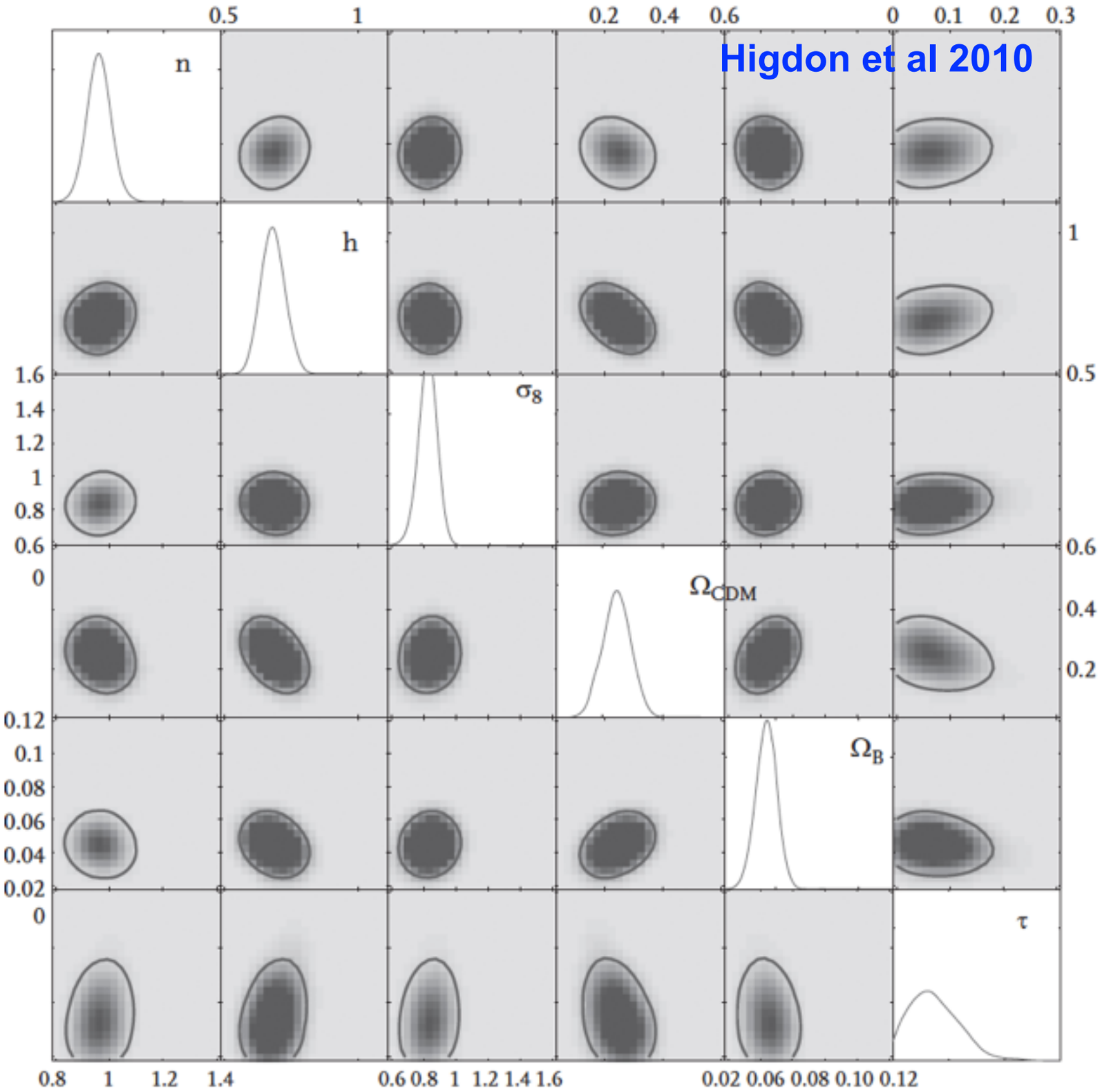
$0.6 \leq \sigma_8 \leq 1.2,$

$0.06 \leq \Omega_{\rm CDM} h^2 \leq 0.2,$

$0.018 \leq \Omega_{\rm b} h^2 \leq 0.034,$

Habib et al 2007

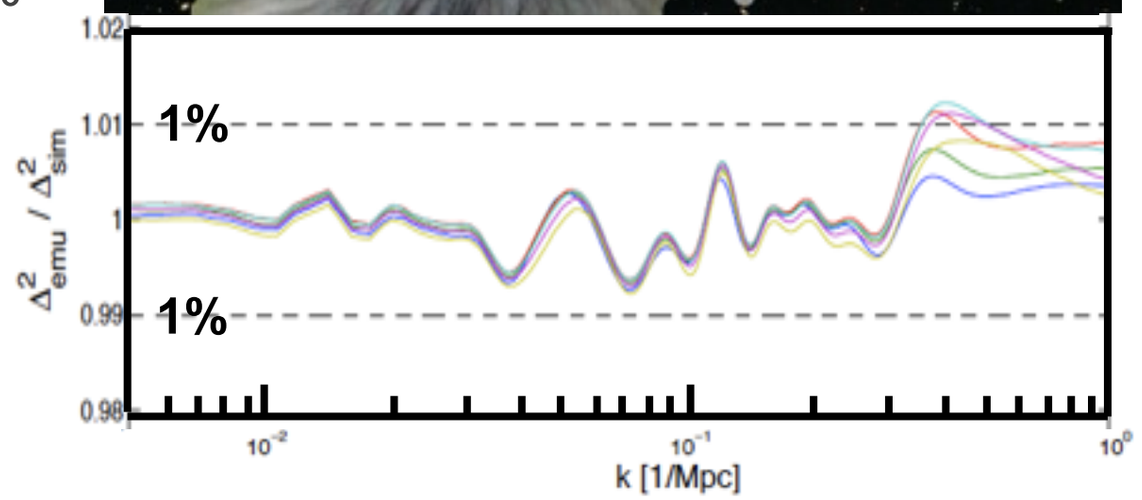
Inverse Problem Solution with SDSS and WMAP Data



The Cosmic Emu(lator)

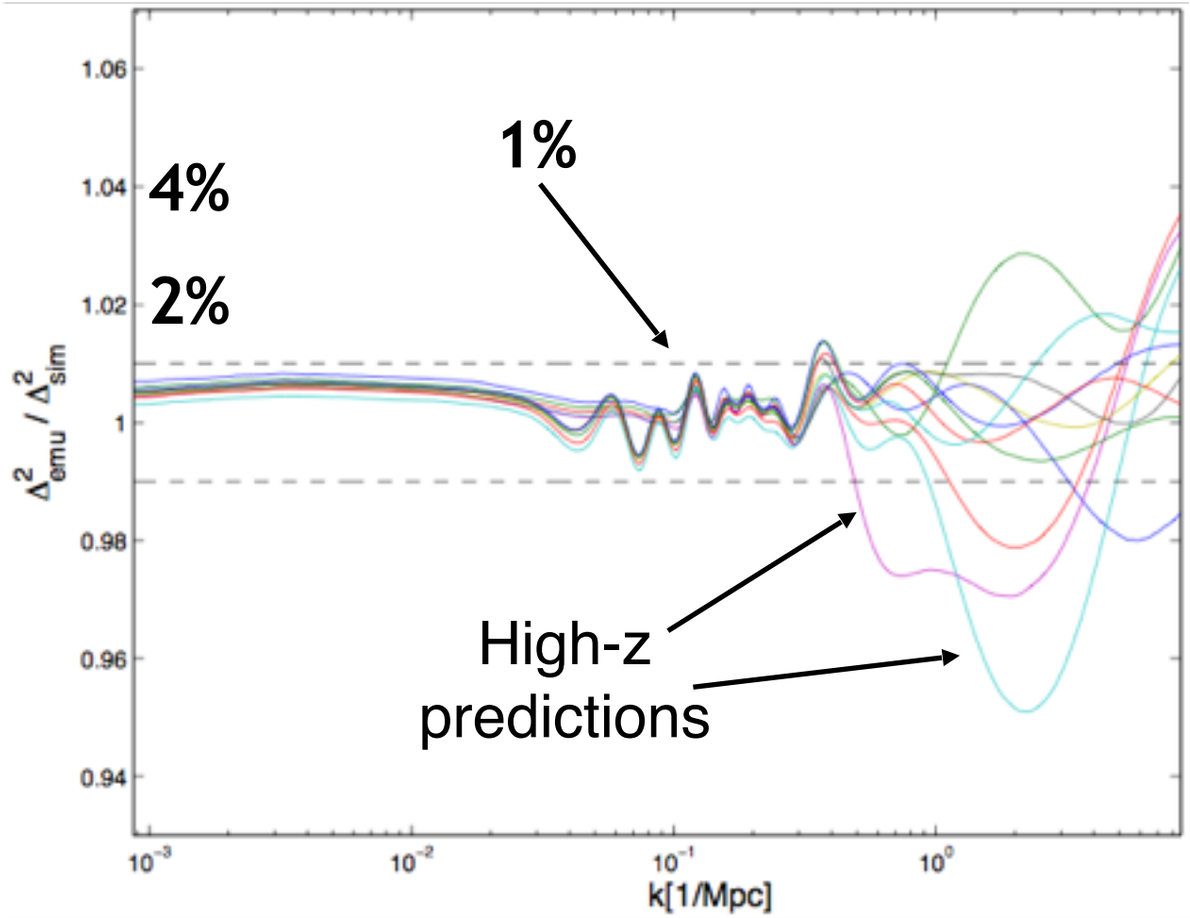
- Prediction tool for matter power spectrum has been constructed
- Accuracy within specified priors between z=0 and z=1 out to k=1 *h*/Mpc at the 1% level achieved
- Emulator has been publicly released, C code
- Extension: Include *h* as sixth parameter, out to k=10 *h*/Mpc and z=4 (many more simulations)
  - Nested simulations to cover large k-range
  - Approach degrades accuracy to ~3%

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

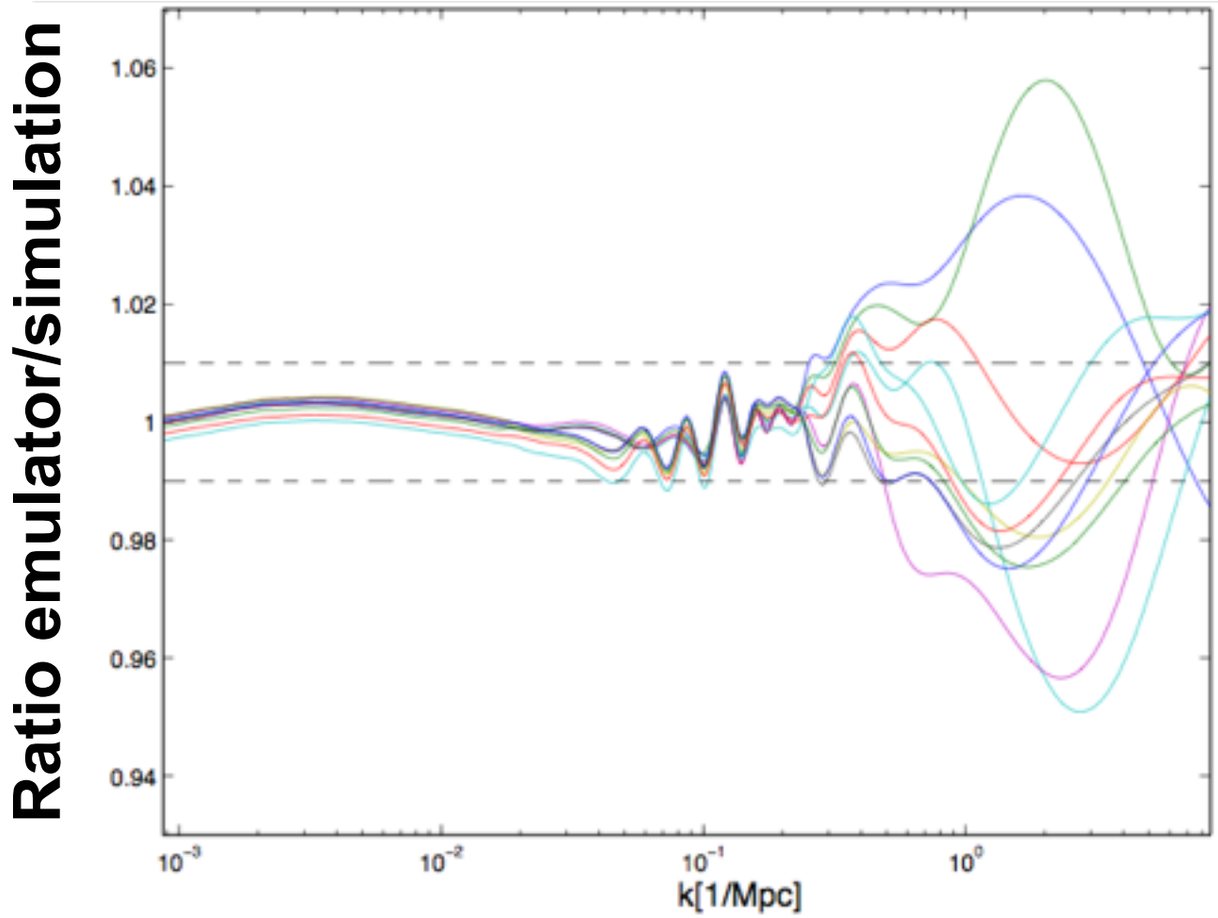


FrankenEmu Results

Extension in k- and z-range



Hubble treated as free parameter



- **Examples of applications to parameter estimation:**
- Weak lensing — CFHTLenS [Fu et al. (2014)]
- Weak lensing — SDSS [Huff et al. (2011)]
- LSS + CMB — SDSS + WMAP [Higdon et al. (2010)]

The Next Step: The Mira-Titan Universe

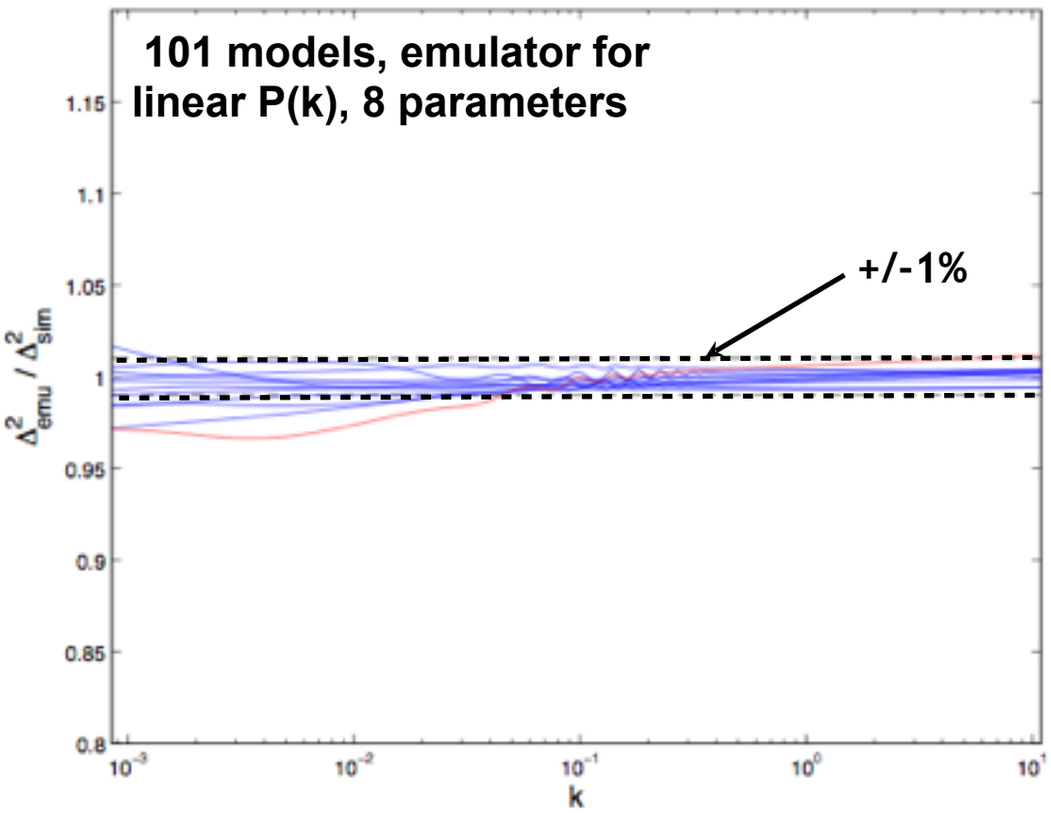
- Extend parameter space to include varying w(z) and massive neutrinos
- Build ***"nested designs"***: enable to build emulator from first set of 26 models, improve with additional 29 models, final precision with 101 models overall (more than halfway done)
- Various emulators for P(k), mass function, c-M relation, RSD predictions, derived quantities...
- LCDM done, finalized set-up based on this run

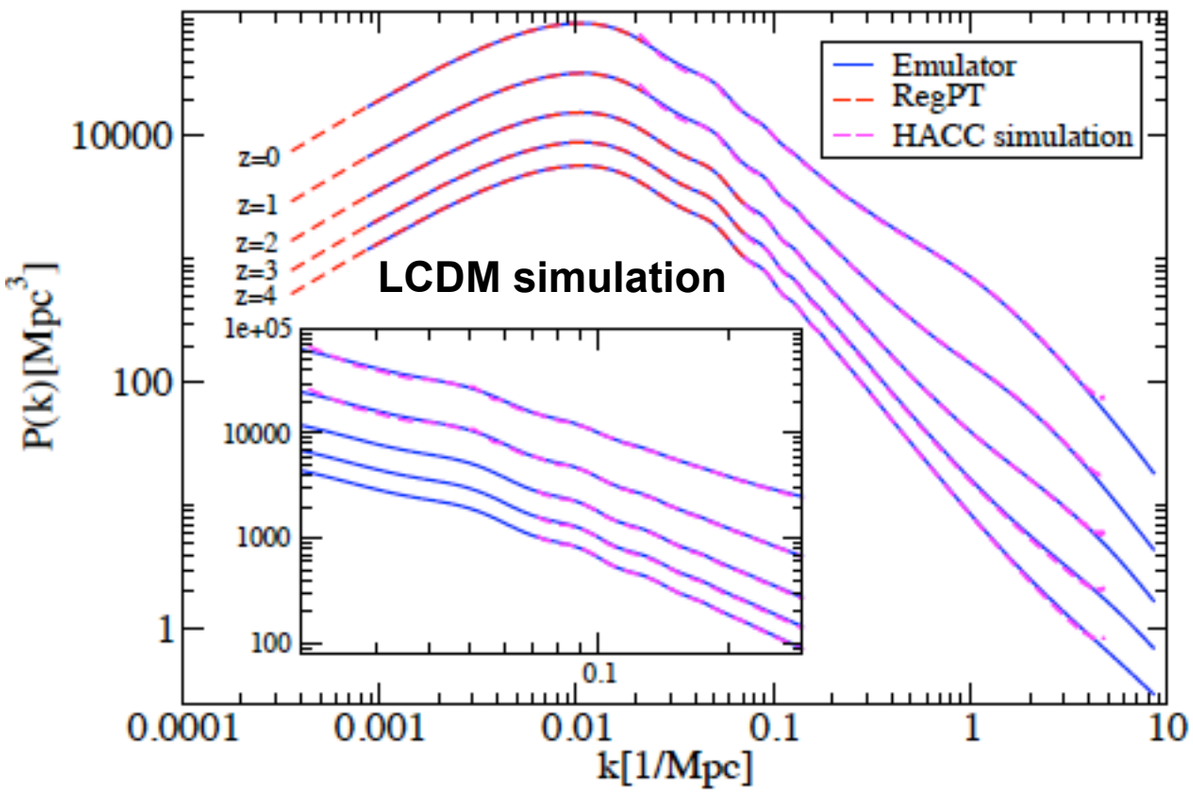
Key advance

Heitmann et al. 2015

Parameters

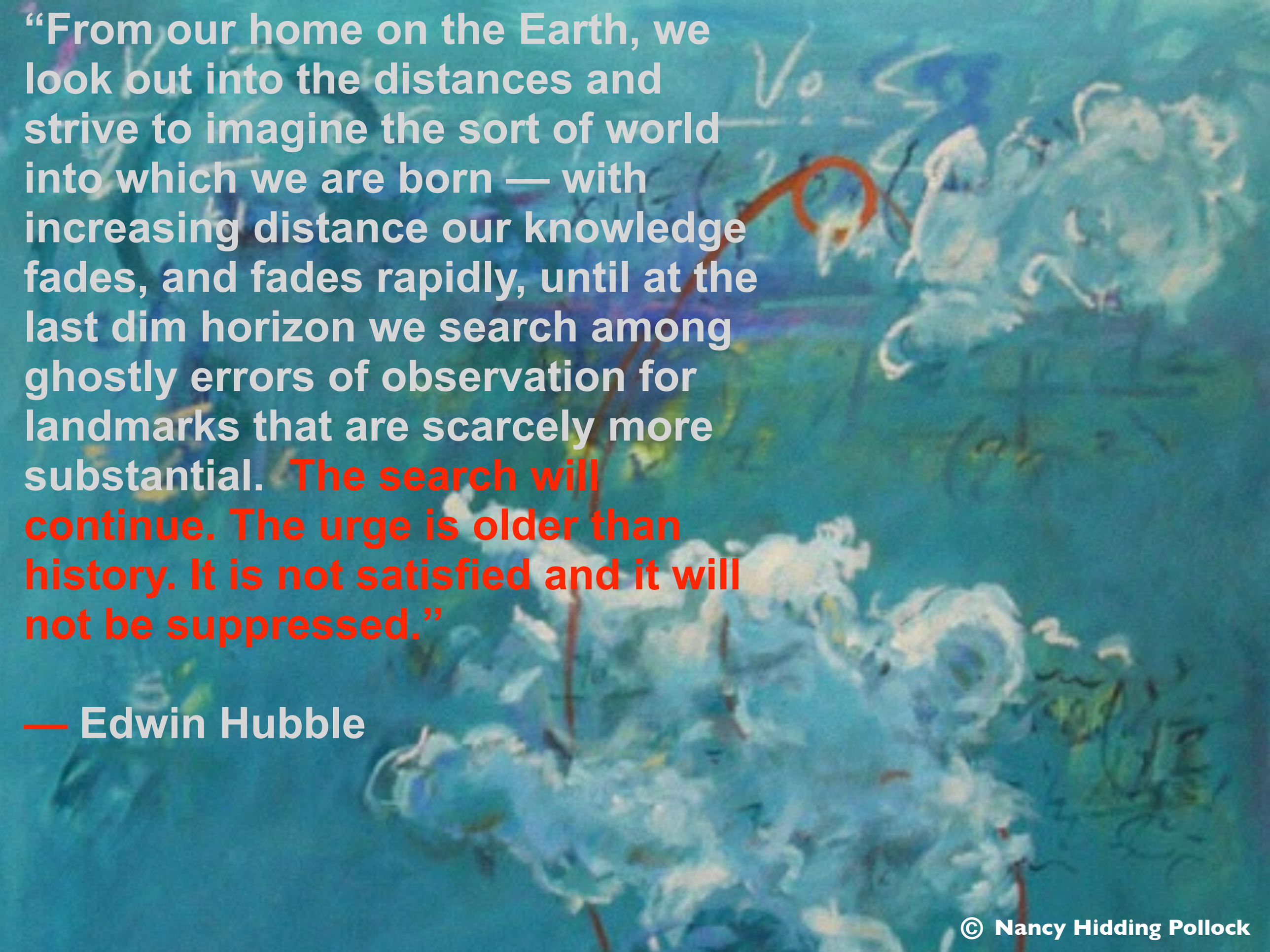
$$0.12 \leq \omega_m \leq 0.155$$
$$0.0215 \leq \omega_b \leq 0.0235$$
$$0.7 \leq \sigma_8 \leq 0.9$$
$$0.55 \leq h \leq 0.85$$
$$0.85 \leq n_s \leq 1.05$$
$$-1.3 \leq w_0 \leq -0.7$$
$$-1.5 \leq w_a \leq 1.15$$
$$0.0 \leq \omega_\nu \leq 0.01.$$





Some Lessons Learned

- Idea for Coyote emulator was born in a Sept. 2006 workshop (motivation was weak lensing), first power spectrum emulator released in Dec. 2009
- January 10, 2008, after a year of testing we think we understand the simulation errors at the 1% level and start the first runs for emulation construction; [February 2, email from Martin White: "*Sanity checks. I've been thinking about these runs some more and worrying about lots of different things. It's so hard to know we're doing things right to 1%!*"]
- First end-to-end calculation of "simplest" but nontrivial problem to provide precision prediction tool
- Collaboration of three different communities: cosmology, computer science, statistics
  - Many tools already exist, do not want to reinvent them! Communication with other communities important, but maybe slow at the start. Don't give up ...
- Simulation infrastructure: Running/analyzing 1000 simulations is not easy...
  - Need for integrated analysis tools
  - Automatization of running the code and checking outputs (working on "Smaash" tool for this)
- Serving the data
  - Data can be used for many projects (working on "PDACS" to provide access to data and analysis tools)



“From our home on the Earth, we look out into the distances and strive to imagine the sort of world into which we are born — with increasing distance our knowledge fades, and fades rapidly, until at the last dim horizon we search among ghostly errors of observation for landmarks that are scarcely more substantial. The search will continue. The urge is older than history. It is not satisfied and it will not be suppressed.”

— Edwin Hubble

More on Simulations —

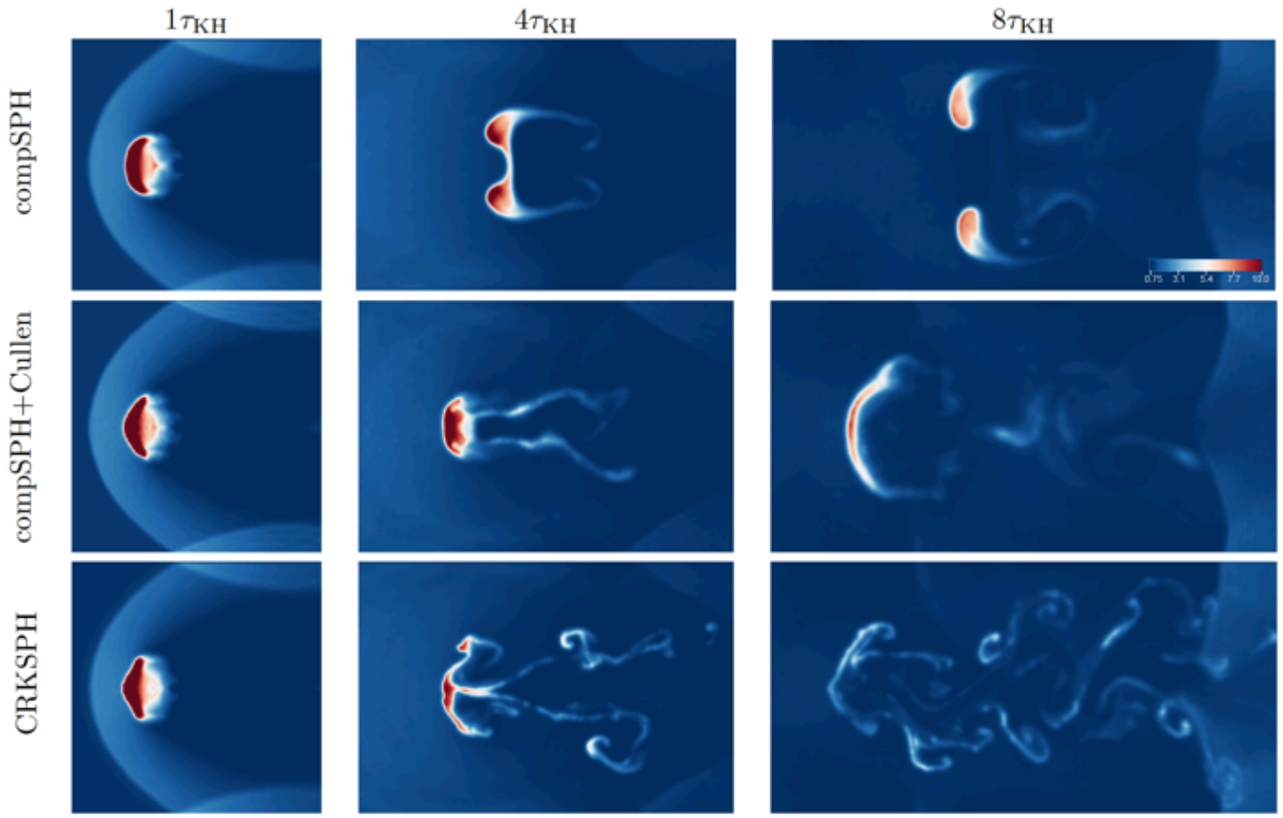
More on Simulations —

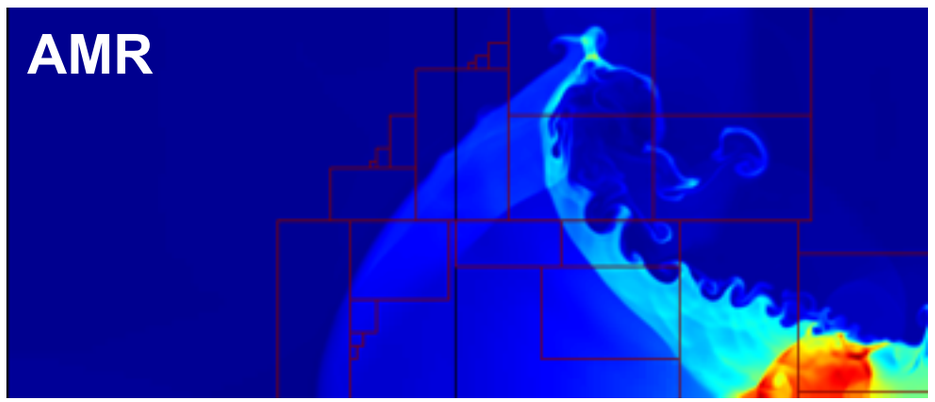
- **Initial conditions**
  - Why can't we run the Universe backwards?
  - How do we know what the Universe is doing at early times?
  - How can we guard against initial condition artifacts?
- **Time-Stepping**
  - What is a symplectic integrator and why do we use it?
  - Hierarchical time-stepping
  - Time-stepping errors and shadowing
- **Considerations for running simulations**
  - Effect of the box size, initial redshift, force, and mass resolution, etc. etc.
- **Analysing simulations**
  - What are the kinds of analyses we need to carry out? (halos, subhalos, merger trees, density estimation, etc.)
  - Approaching the "virtual universe"

Towards a “Virtual” Universe

- **Aim of our cosmological simulation effort**
  - Better understanding of physics in the simulations
  - Go from simulations directly to observations
  - Requires major effort on next-generation systems
  - Adding hydro — CRKSPH (Conservative Reproducing Kernel SPH)
  - Now supported by DOE ECP

EXASCALE COMPUTING PROJECT





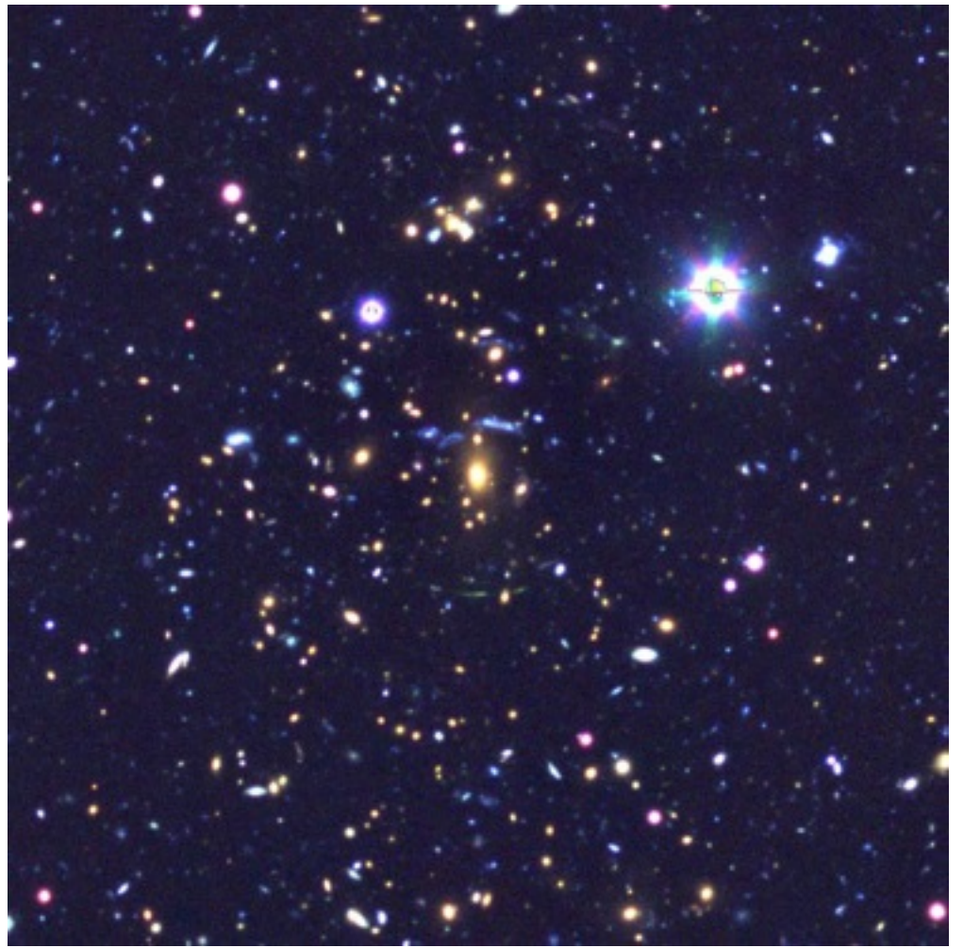
PICS Example: Generating Strong Lensing “Observations”

PICS: Pipeline for Images of Cosmological Strong lensing

Simulated strong lens image (left) to match SPT cluster observations taken with the MegaCAM camera on Magellan (right), work by L. Bleem, M. Florian, S. Habib, K. Heitmann, M. Gladders, N. Li, S. Rangel

N. Li et al., arXiv:1511.03673, ApJ (in press)

Lecture 2: Summary



Resources

- **Texts**
  - Dodelson, *Modern Cosmology*
  - Durrer, *The Cosmic Microwave Background*
  - Gorbunov and Rubakov, *Introduction to the Theory of the Early Universe*
  - Kaiser, *Elements of Astrophysics (*http://www.ifa.hawaii.edu/~kaiser/lectures/*)*
  - Kitchin, *Astrophysical Techniques*
  - Lyth and Liddle, *The Primordial Density Perturbation*
  - Mo, van den Bosch, and White, *Galaxy Formation and Evolution*
  - Martinez and Saar, *Statistics of the Galaxy Distribution*
  - Mukhanov, *Physical Foundations of Cosmology*
  - Peacock, *Cosmological Physics*
  - Ryden, *Introduction to Cosmology*
  - Weinberg, *Cosmology*

Cosmological ‘Standard Model’

- **ΛCDM**
  - General relativity describes gravity at all relevant scales; at large scales, the Universe is homogeneous and isotropic (FRW metric)
  - The cosmological constant is nonzero
  - Initial density fluctuations are given by a Gaussian random field (almost scale invariant; consistent with inflation)
  - Constituents include baryonic matter, neutrinos, and radiation
  - Characterized by a handful of parameters
  - Currently consistent with all known measurements (where systematics are controllable)

$$\rho_c = \frac{3H^2}{8\pi G} \text{ (critical density)}$$

$\Omega_{\rm i} \equiv \frac{\rho_{\rm i}}{\rho_{\rm c}}$ (density parameter)

$\rm H_0$ (Hubble constant)

σ_8 (Power spectrum normalization)

$\rm n_s$ (scalar spectral index)

r (tensor to scalar ratio)

$\rm m_\nu$ (neutrino mass sum)

$\rm N_\nu$ (effective number of neutrino species)

w (dark energy equation of state parameter)

Some Cosmological Standard Model Parameters

Puzzles: Cosmology is a Work in Progress

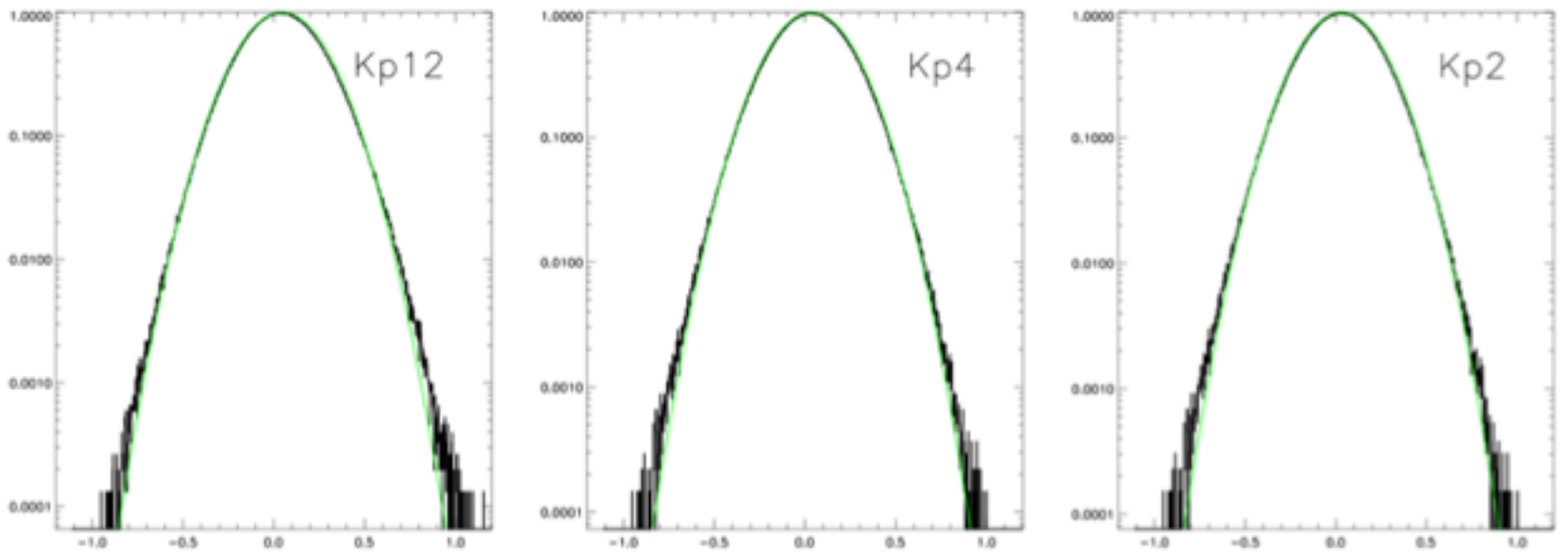
- **Specialness, etc.**
  - Why is the Universe so old?
  - Why is it spatially flat?
  - Why is the CMB temperature so isotropic (‘horizon problem’)?
  - Where do the primordial fluctuations come from, and why are they Gaussian?
- Why is the cosmological constant (‘vacuum energy’) so small?
- How sensitive are future states to initial conditions?
- What happened to topological relics?
- Baryogenesis?

$$l_P = \sqrt{\frac{\hbar G}{c^3}} = 1.616 \times 10^{-35} m$$

$$t_P = \frac{l_P}{c} = \sqrt{\frac{\hbar G}{c^5}} = 5.39 \times 10^{-44} s$$

$$m_p = \sqrt{\frac{\hbar c}{8\pi G}} = 4.34 \times 10^{-9} kg = 2.44 \times 10^{18} GeV/c^2$$

$$\frac{\Delta T}{T} \sim 10^{-5}$$

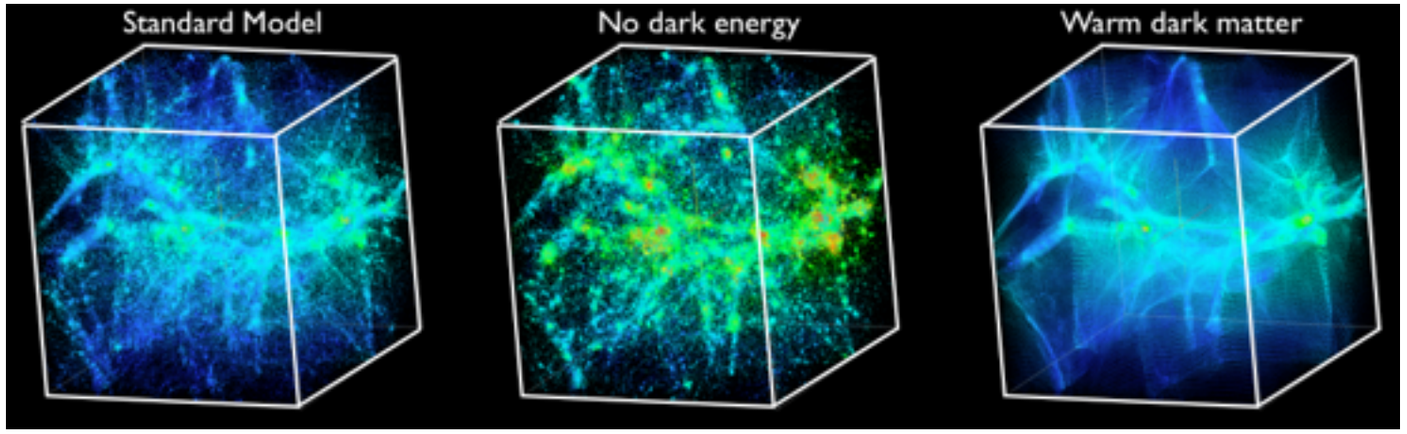


One-point PDFs of masked WMAP temperature maps (Andrew Jaffe)

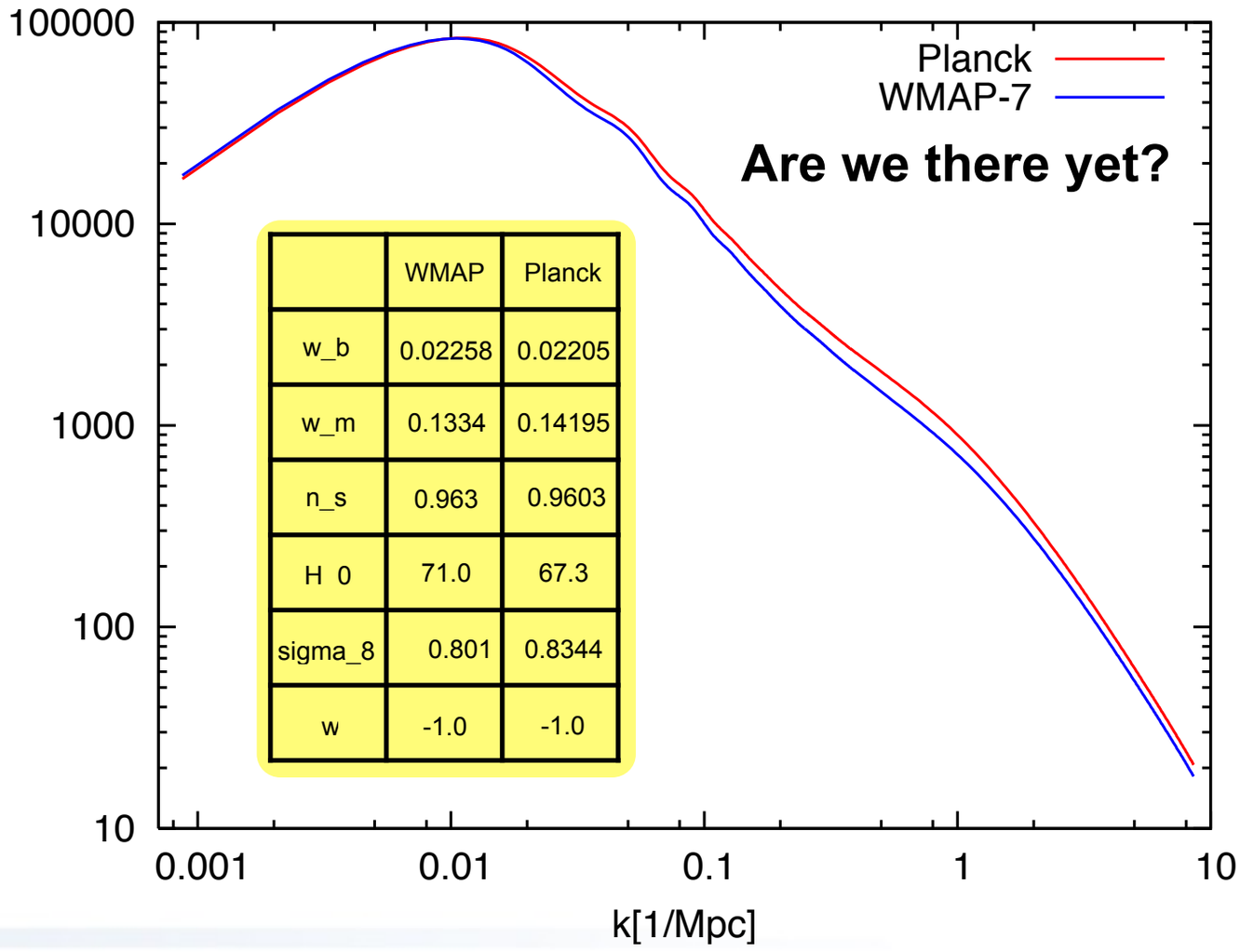
Precision Cosmology: Distinguishing Features

- **Precision Probes**
  - Complete understanding of measurement technique
  - Well-defined prediction from cosmological Standard Model; possible contamination from astrophysical uncertainties either understood or capable of being modeled out
  - Small measurement errors that are directly connected to small measurement errors in cosmological parameters
  - Examples: prediction/ measurement of geometry and growth of structure

focus of these lectures —



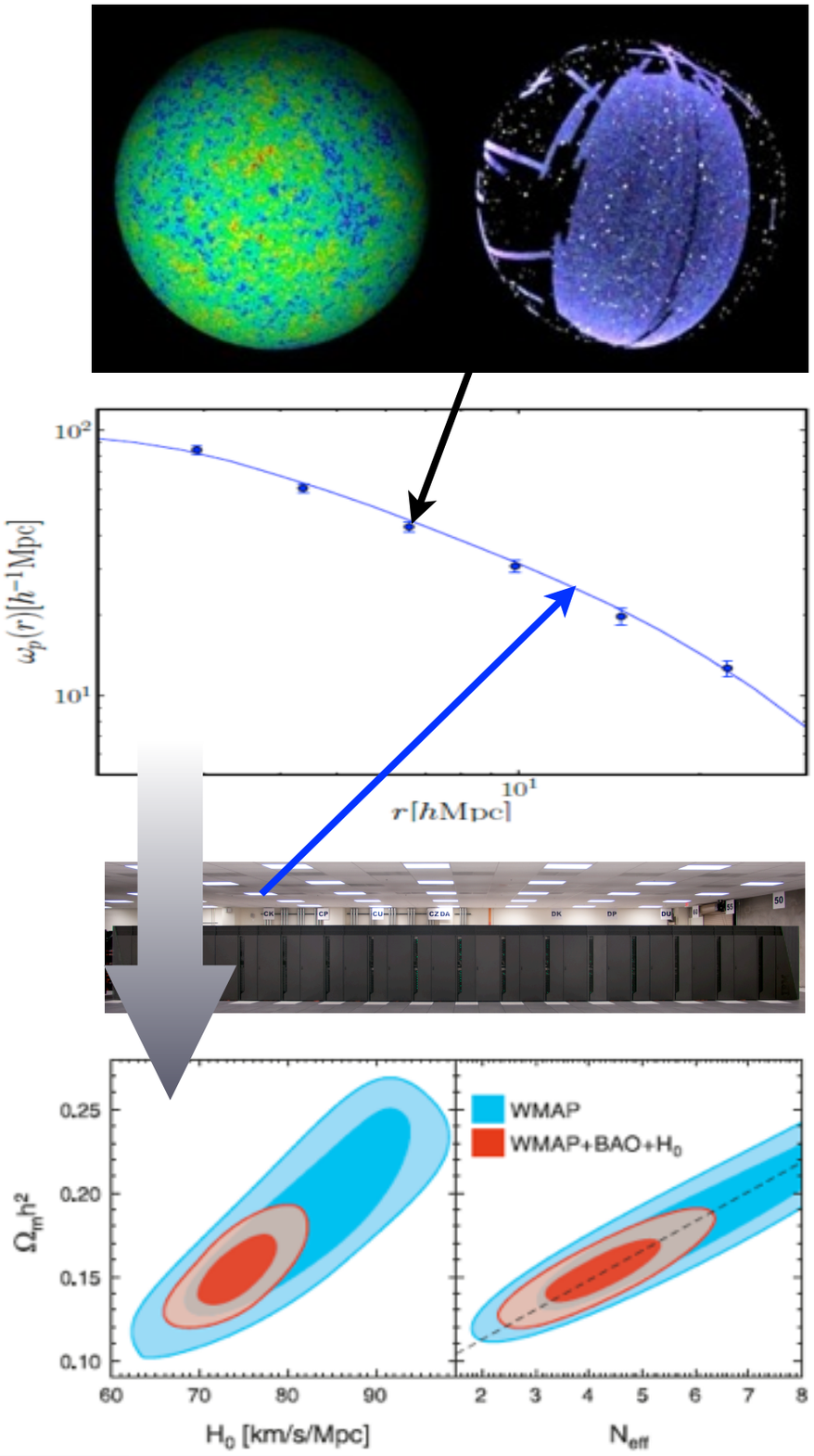
Precision cosmology with large-scale structure



	WMAP	Planck
w_b	0.02258	0.02205
w_m	0.1334	0.14195
n_s	0.963	0.9603
H_0	71.0	67.3
sigma_8	0.801	0.8344
w	-1.0	-1.0

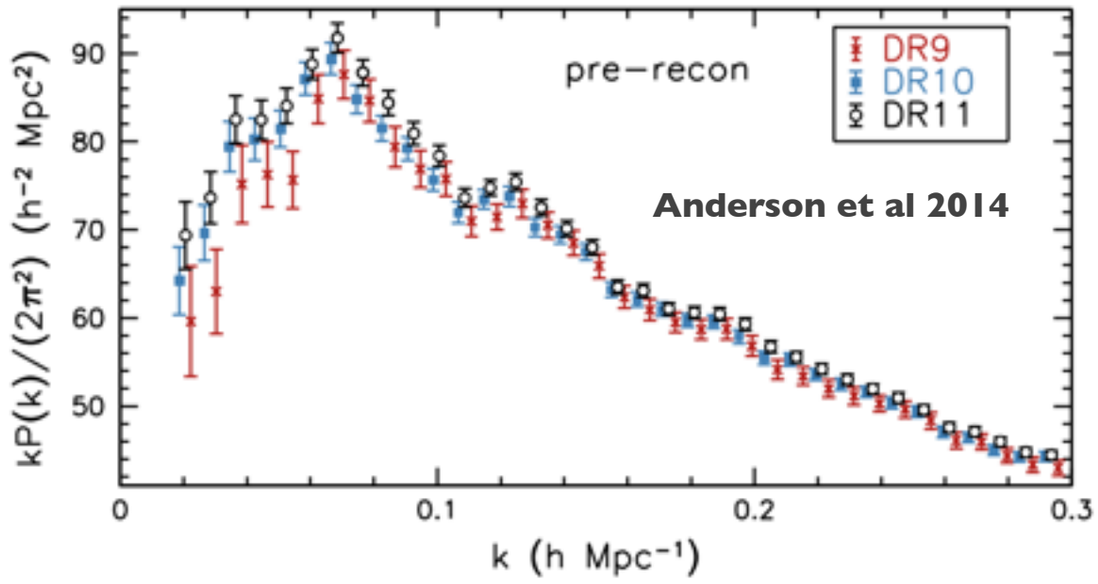
Cosmology as an Inverse Problem

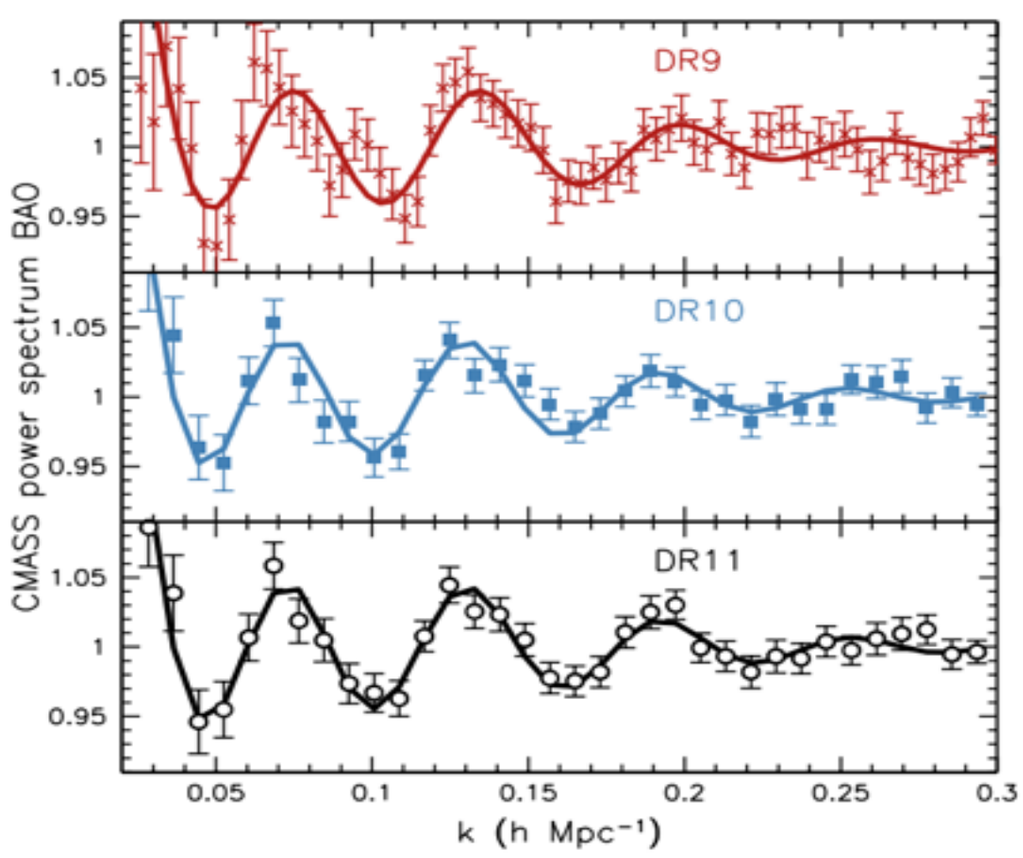
- **Measurements:** Raw objects are sky maps (objects/fields) at various degrees of coverage and completeness
- **Summary Statistics:** Distill information from maps via ‘summary statistics’ (e.g., correlation and cross-correlations)
- **Predictions:** Theory provides ‘forward model’ for summary statistics
- **Inverse Problem:** Explore posterior distribution of cosmological parameters via Bayes Theorem using MCMC (or analogous techniques)
- **Cosmic Calibration Framework:** Four-step process: (i) Sample parameter space; (ii) Build ‘interpolator’ for statistics of interest; (iii) Run MCMC engine for posterior exploration and parameter estimation; (iv) Make predictions for new observational probes/measurements



Two-Point Statistics

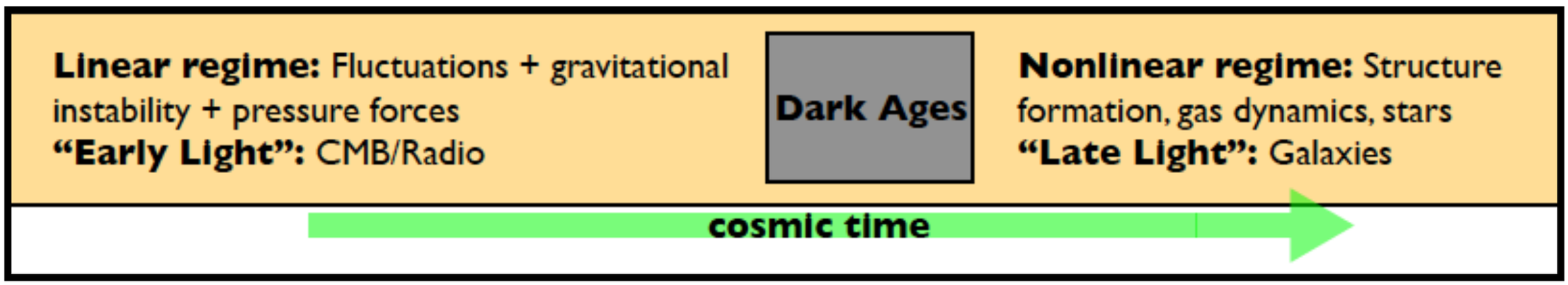
- All structure formation probes in some way study density fluctuations
- Hence desire **robust** ways to characterize clustering statistics of the underlying mass field and its tracers (e.g., galaxies)
- The (2-point) **correlation function** is the excess probability of finding an object pair separated by a distance r_{12} compared to that for a random distribution:
  $dP=n^2(1+\xi(r_{12}))dV_1dV_2$
  where n is the mean density; the **power spectrum** $P(k)$ is the Fourier transform of the correlation function
- The primordial fluctuations, as best known currently, are Gaussian, and completely specified by 2-point statistics
- Nonlinear structure formation induces non-zero higher point correlation functions



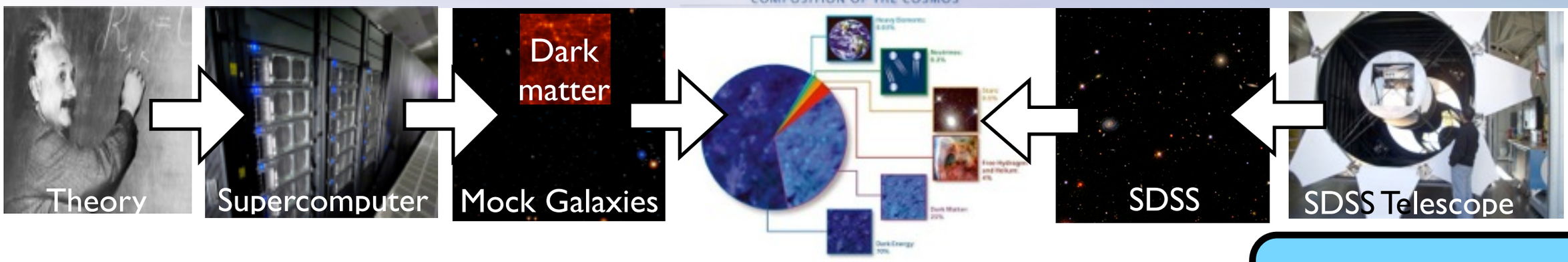


The Story of Primordial Fluctuations

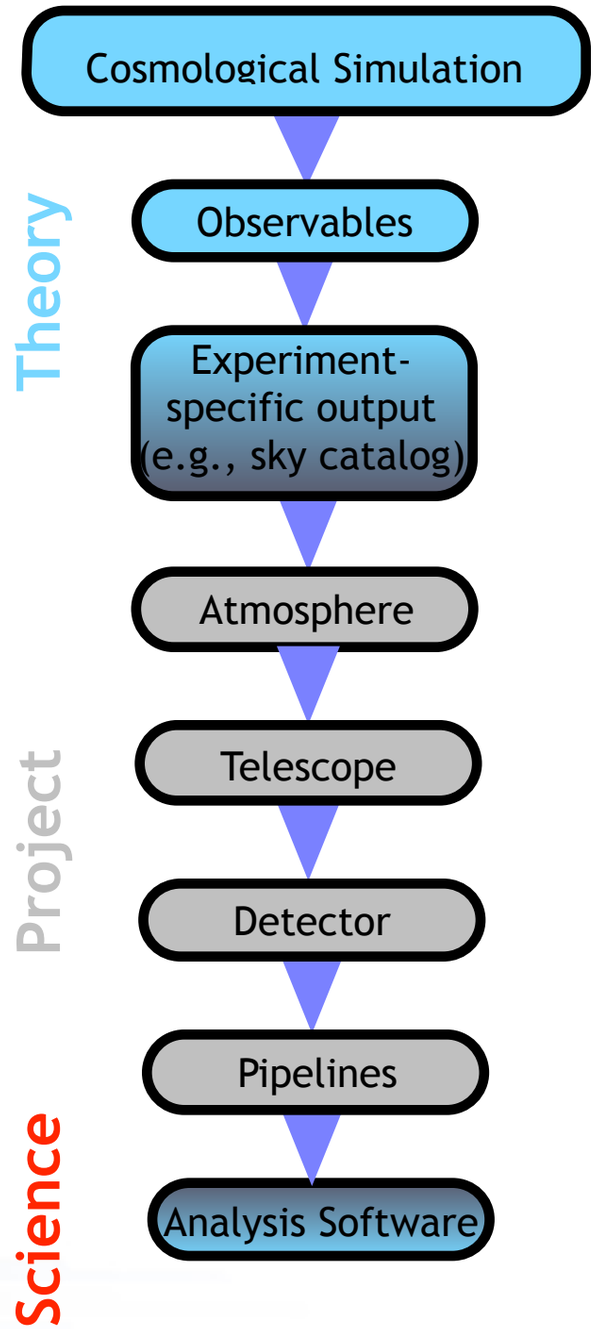
- **Fluctuation Timeline**
  - Imprinted by early universe processes such as quantum fluctuations or phase transitions
  - Initial density fluctuations same for all species
  - Growth of perturbations suppressed as long as universe as radiation-dominated
  - After radiation-matter equality, perturbations in the dark matter component begin to grow
  - Baryons and photons, essentially locked together follow the CDM potential
  - Photon pressure causes baryons to see a repulsive force leading to acoustic oscillations
  - Gravitational instability eventually leads to nonlinear mode mixing
  - Strong gravity and gas dynamics/feedback processes form present universe



Role of Computational Cosmology



- **Three Roles of Cosmological Simulations**
  - Basic theory of cosmological probes
  - Production of high-fidelity 'mock skys' for end-to-end tests of the observation/analysis chain
  - Essential component of analysis toolkits
- **Extreme Simulation and Analysis Challenges**
  - Large dynamic range simulations; control of subgrid modeling and feedback mechanisms
  - Design and implementation of complex analyses on large datasets; new fast (approximate) algorithms
  - Solution of large statistical inverse problems of scientific inference (many parameters, ~10-100) at the ~1% level

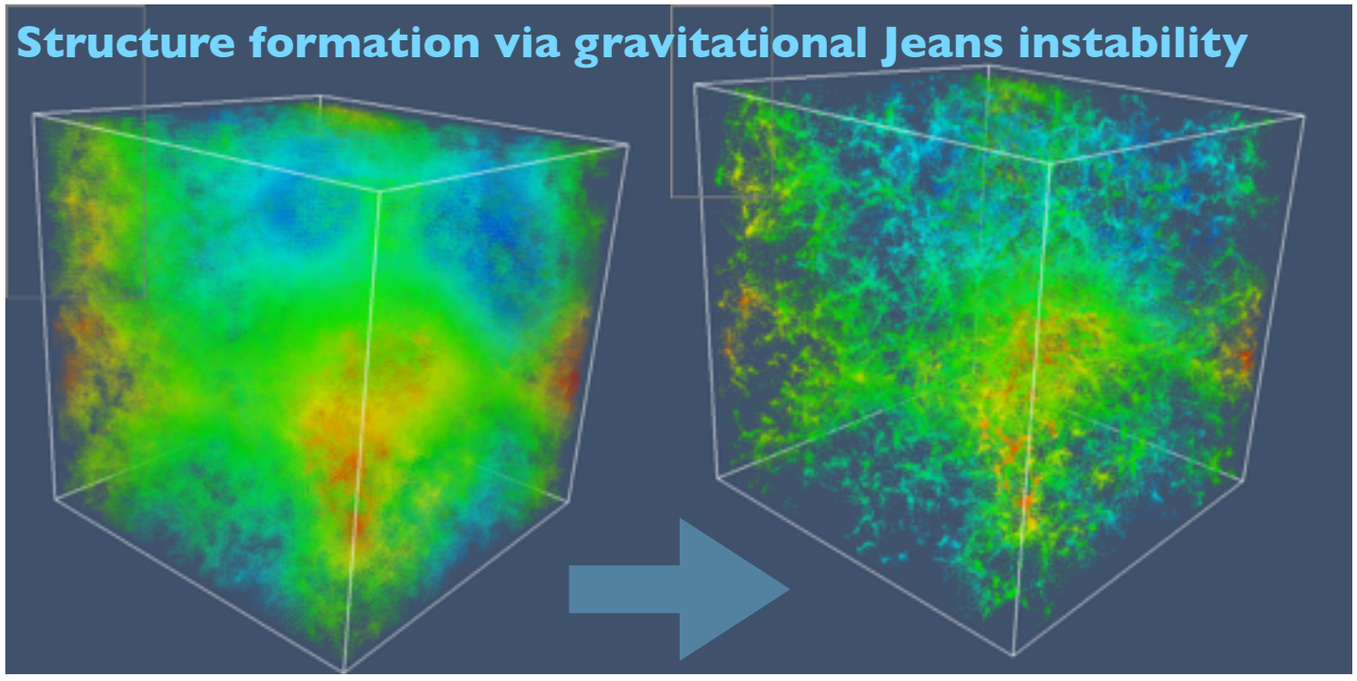


Simulating the Universe

- **Key Role of Gravity**
  - Gravity dominates at large scales: solve the Vlasov-Poisson equation (VPE)
  - VPE is 6-D and cannot be solved as a PDE
- **N-Body Methods**
  - No shielding in gravity (essentially long range interactions)
  - Technique is naturally Lagrangian
  - Are errors controllable?
- **More Physics**
  - Smaller scale ‘gastrophysics’ effects added via subgrid modeling or post-pocessing
- **Phenomenology**
  - Calibrate simulations against observations

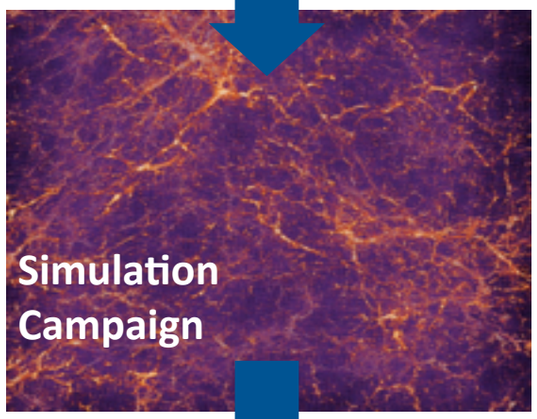
$$
\begin{aligned}
\frac{\partial f_i}{\partial t} + \dot{\mathbf{x}}\frac{\partial f_i}{\partial \mathbf{x}} - \nabla\phi\frac{\partial f_i}{\partial \mathbf{p}} = 0, \qquad \mathbf{p} = a^2\dot{\mathbf{x}}, \\
\nabla^2\phi = 4\pi G a^2(\rho(\mathbf{x},t) - \langle\rho_{\rm dm}(t)\rangle) = 4\pi G a^2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

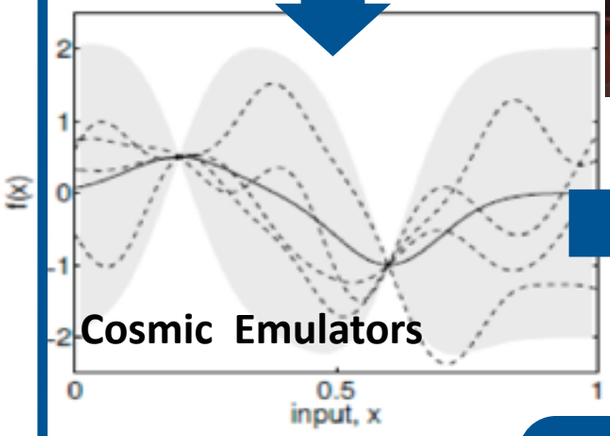
Cosmological Vlasov-Poisson Equation: A ‘wrong-sign’ electrostatic plasma with time-dependent particle ‘charge’, Newtonian limit of the Vlasov-Einstein equations



Precision Cosmology: 'Big Data' Meets Supercomputing

Supercomputer Simulation Campaign

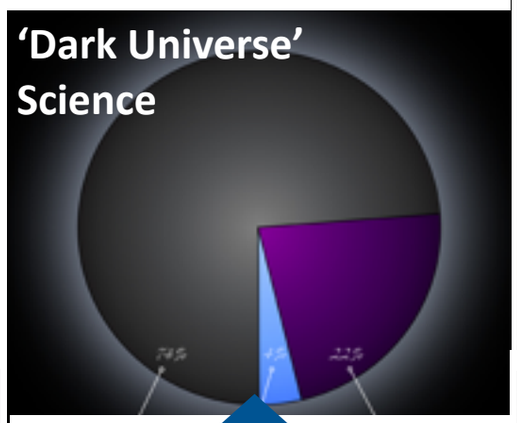




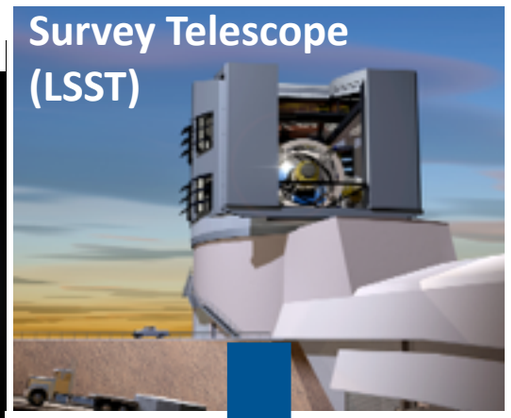
Simulations + CCF

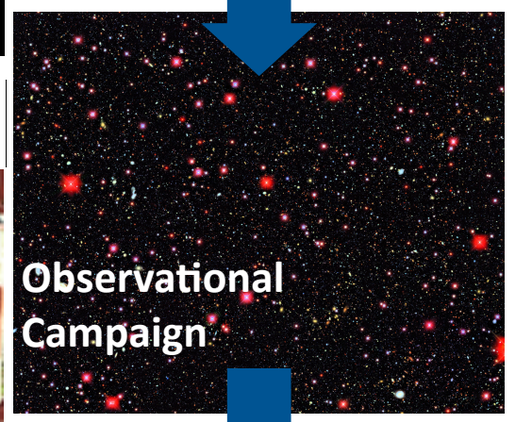
Emulator based on Gaussian Process Interpolation in High-Dimensional Spaces

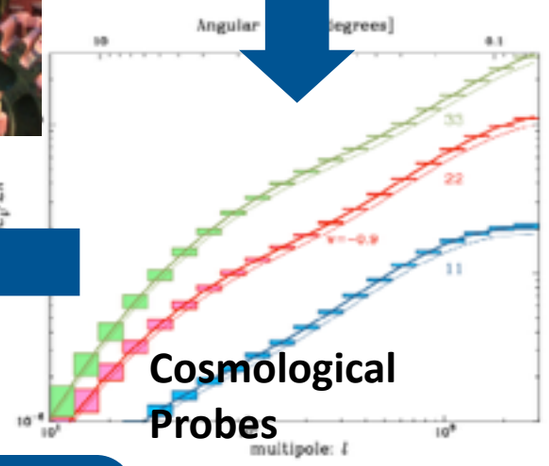
Calibration



'Precision Oracle'







Mapping the Sky with Survey Instruments

Observations: Statistical error bars will 'disappear' soon!

Science with Surveys: HPC meets Big Data

CCF= Cosmic Calibration Framework

The ‘Bleeding Edge’: Two Questions

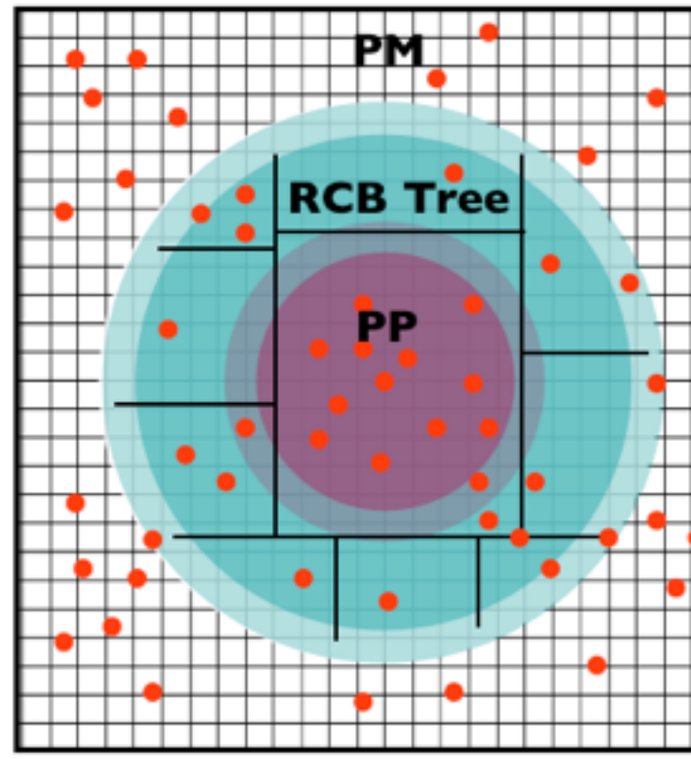
- **Recall: Cosmology = Physics + Statistics**
  - Mapping the sky with large-area surveys
  - LSST: ~4 billion galaxies total; ~200,000 galaxies per sq. deg. or ~40K galaxies over a sky patch the size of the moon
  - To ‘understand’ a single dataset this large (~100 PB), we need to model the distribution of matter down to the scales of the individual galaxies, and over the size of the entire survey: ~trillion particle simulations
- **Resolution:**
  - Force dynamic range greater than a million to one
  - Local overdensity variation is ~million to one
- **Computing ‘Boundary Conditions’:**
  - Total memory in the PB+ class
  - Performance in the 10 PFlops+ class
  - Wall-clock of ~days/week, in situ analysis

Can the entire observable Universe be ‘stuffed’ inside a supercomputer?

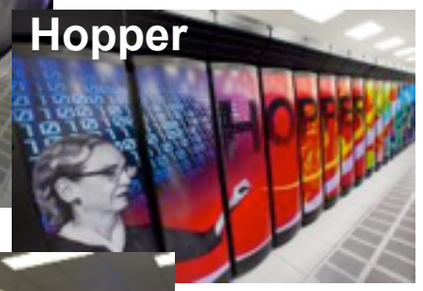
Can the Universe be run as a short computational ‘experiment’?

Opening the HACC ‘Black Box’: Design Principles

- **Optimize Next-Generation Code ‘Ecology’:** Numerical methods, algorithms, mixed precision, data locality, scalability, I/O, in situ analysis -- life-cycle significantly longer than architecture timescales
- **Framework design:** Support a ‘universal’ top layer + ‘plug-in’ optimized node-level components; minimize data structure complexity and data motion -- support multiple programming models
- **Performance:** Optimization stresses scalability, low memory overhead, and platform flexibility; assume ‘on your own’ for software support, but hook into tools as available (e.g., ESSL FFT)
- **Optimal Splitting of Gravitational Forces:** Spectral Particle-Mesh melded with direct and RCB tree force solvers, short hand-over scale (dynamic range splitting ~ 10,000 X 100)
- **Compute to Communication balance:** Particle Overloading
- **Time-Stepping:** Symplectic, sub-cycled
- **Force Kernel:** Highly optimized force kernel takes up large fraction of compute time, no look-ups due to short hand-over scale
- **Production Readiness:** runs on all supercomputer architectures; exascale ready!

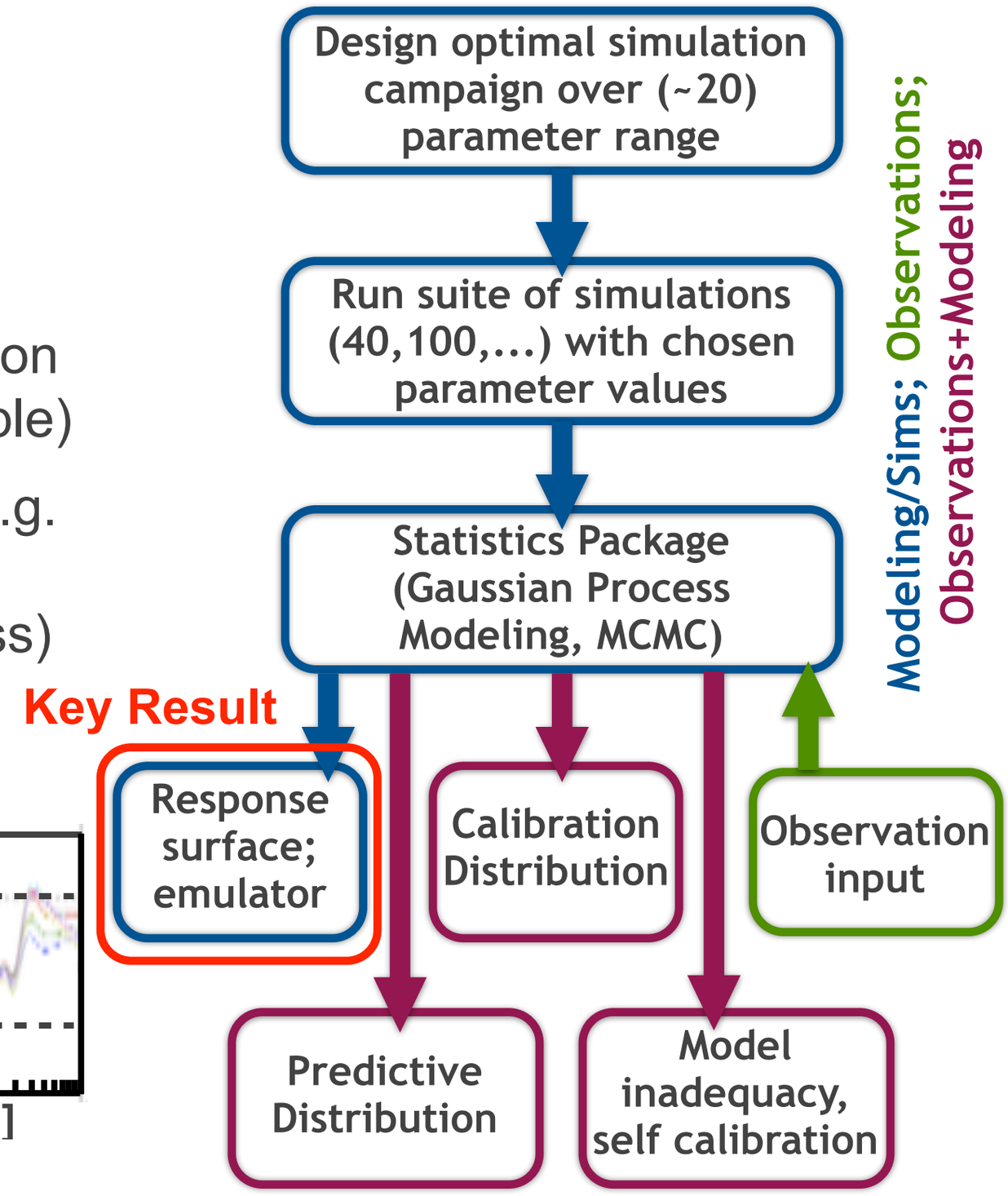


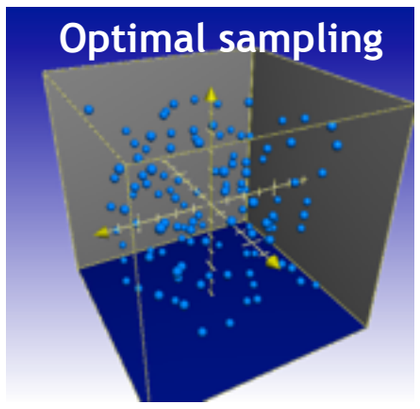
HACC force hierarchy (PPTreePM)

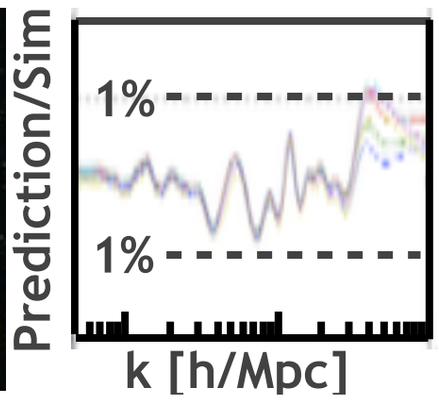


Cosmic Calibration: Solving the Cosmic Inverse Problem

- **Challenge:** To extract cosmological constraints from observations in the nonlinear regime, need to run Markov Chain Monte Carlo; input: 10,000 - 100,000 different models
- **Brute Force:** Simulations, ~30 years on 2000 processor cluster (simple example)
- **Current Strategy:** Fitting functions, e.g. for P(k), accurate at ~10% level, not good enough! (Pertbn. theory hopeless)
- **Our Solution:** Precision emulators







Heitmann et al. 2006, Habib et al. 2007, --