Theories of everything

David W. Hogg
(Center for Cosmology and Particle Physics, New York University)

2012 December 16
what I am going to say

- Uniquely in astronomy, it is conceivable to have a justified probabilistic model of every observable source that explains every observed pixel of every image ever taken.

- We are working towards a few-$10^9$ parameter model of the $10^{13}$ pixels of large survey data we have in hand.

- **We would like to have a hierarchical model with a structure that mirrors the causal structure implicit in the cosmological model.**

- Collaborators:
  - Jo Bovy (IAS)
  - Rob Fergus (NYU)
  - Dan Foreman-Mackey (NYU)
  - Dustin Lang (Princeton)
  - Phil Marshall (Oxford)
  - Sam Roweis (deceased)
  - NASA, NSF (AST and IIS), Humboldt Foundation
astronomy: scale

- in 2011: total imaging data $10^{13}$ to $10^{14}$ few-byte pixels
  - only hundreds of terabytes
  - I spin it in my own machine room
- $10^9$ to $10^{10}$ useful stars in the Milky Way
- $10^{10}$ useful galaxies in the observable Universe
- by 2020: $10^{16}$ pixels if we are lucky
  - note “optical” and “imaging” biases
astronomy is a sweet inference problem
astronomy: promise

- the Universe is driven by unseen constituents
  - dark matter
  - dark energy
- precise theory of gravity, the cosmological world model, and the growth of structure
- excellent physical models of stars
- working models for galaxies
- at the lowest level, the model is exceedingly sparse
  - each star touches only $10^{-11}$ of the full pixel database
  - each pixel touches only $\sim 1$ star or galaxy
- CCDs, IR arrays, receivers all well-understood in sensitivity and noise properties
Image modeling (Lang): data
Image modeling (Lang): model
what is a model?

- probability of the data given assumptions and parameters (a likelihood)
- priors over parameters
- priors over qualitative choices
- produces only probabilistic information
- inherently hierarchical
astronomers love catalogs

- list of stars and galaxies in the imaging you have
- SDSS Catalog has nearly $10^9$ sources
- objects detected using statistical significance tests
- object properties measured using “algorithms”
- no probabilistic interpretation
  - the likelihood function doesn’t permit freedom in “algorithm choice”
  - no sampling over parameters
  - no sampling over qualitatively different explanations
blind deconvolution

► have an image $b$ (unwrapped into one huge vector)
► model is

$$b = A \cdot x + e$$

where $A$ is a convolution matrix, $x$ is an unconvolved (true) scene, $e$ is noise

► $x$ has as many elements as $b$ and $A$ has far, far more (in general).
► regulation, restriction, or priors are required
► the likelihood function is just the noise model $p(e)$
► This is a standard idea underlying computer vision (see, e.g., Hirsch, Schölkopf).
Image modeling (Lang): tuning
Faint-source proper motions (0808.4004): brown dwarf
Faint-source proper motions (0808.4004): $z \sim 6$ quasar
Faint-source proper motions (0808.4004): faint galaxy
Faint-source proper motions (0808.4004): defect
“okay, you can model the stars, but that’s the easy part!”

- stars are laid down by a process set by the Milky Way
- the Milky Way is created by a process set by the cosmological model
Fig. 1.— Smoothed, summed weight image of the SDSS field after subtraction of a low-order polynomial surface fit. Darker areas indicate higher surface densities. The weight image has been smoothed with a Gaussian kernel with $\sigma = 0.2^\circ$. The white areas are either missing data, or clusters, or bright stars which have been masked out prior to analysis.

Stream-finding

going up the hierarchy

- stars are on orbits around the Milky Way
- the Milky Way is generated by gravitational collapse from initial conditions (and a whole lot of gastrophysics)
- the initial conditions and growth of structure are set by the cosmological parameters
- the cosmological parameters are set by...
- we don’t know how to perform this inference, even approximately!
  - models of no fixed complexity
  - likelihood “calls” involve enormous dynamical computations
  - numbers of parameters in the billions
  - butterfly effects abound
and we’re back:

**Key:** Telescope / Atmosphere / Detector / Star / Galaxy
baryon acoustic feature
two-point functions

- much of cosmology depends on two-point functions
  - correlation functions
  - power spectra
  - variances as a function of scale
- two point functions are invariably point-estimated
  - count pairs as a function of separation
  - evaluate harmonic basis functions at galaxies
  - FFT a pixel map
  - insane “window functions”
- principled inference happens after that lossy point estimation
- can we do better?
  - Gaussian processes?
  - infer the density field?
exoplanet direct imaging (Fergus, in prep)
what I said

- Uniquely in astronomy, it is conceivable to have a justified probabilistic model of every observable source that explains every observed pixel of every image ever taken.
- We are working towards a few-$10^9$ parameter model of the $10^{13}$ pixels of large survey data we have in hand.
- **We would like to have a hierarchical model with a structure that mirrors the causal structure implicit in the cosmological model.**
- Collaborators:
  - Jo Bovy (IAS)
  - Rob Fergus (NYU)
  - Dan Foreman-Mackey (NYU)
  - Dustin Lang (Princeton)
  - Phil Marshall (Oxford)
  - Sam Roweis (deceased)
  - NASA, NSF (AST and IIS), Humboldt Foundation