How is cosmology like exoplanets?

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

2012 June 7

punchlines

- Probabilistic inference with a generative model beats any point estimate for accuracy and precision.
- When you don't know how to model your data, use the data to build the model; think *hierarchically*.
- You usually need to spend even more time modeling the things you *don't care about*—the noise—than the things you do—the signal.
- "Images $\rightarrow$ coadd $\rightarrow$ catalog $\rightarrow$ best-fit model $\rightarrow$ high-level conclusions" just won't work in many circumstances.
  - warnings for *LSST* and *PanSTARRS* and *Gaia* and ...

principal collaborators

- **Jo Bovy** (IAS)
- *Rob Fergus (NYU)*
- Dan Foreman-Mackey (NYU)
- *Jonathan Goodman (NYU)*
- Joe Hennawi (MPIA)
- Rory Holmes (MPIA)
- Sergei Koposov (Cambridge)
- **Dustin Lang** (Princeton $\rightarrow$ CMU)
- Hans-Walter Rix (MPIA)
- *Sam Roweis (deceased)*
- David Schiminovich (Columbia)
- Vivi Tsalmantza (MPIA)

Hogg's decadal survey

- Money spent on *inference with real data* is much more productive, per dollar, than money spent on hardware or theory. . .

Hogg's decadal survey

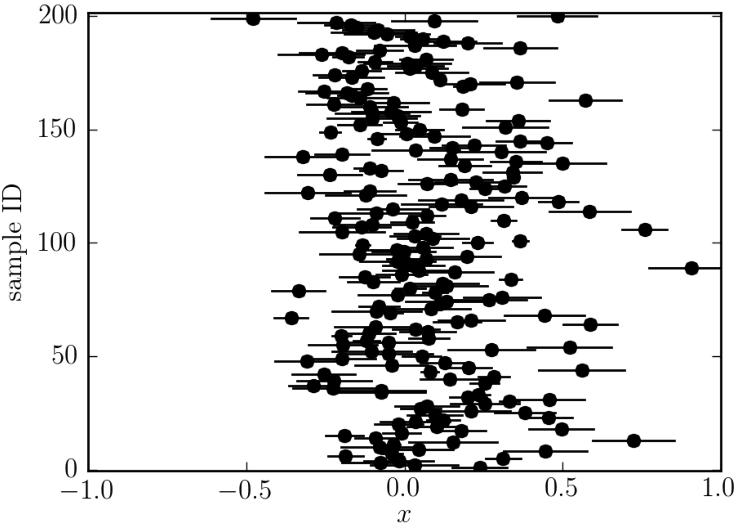
- Money spent on *inference with real data* is much more productive, per dollar, than money spent on hardware or theory. . .
  - . . . *and will help us survive the collapse of NASA and NSF!*

what is inference?

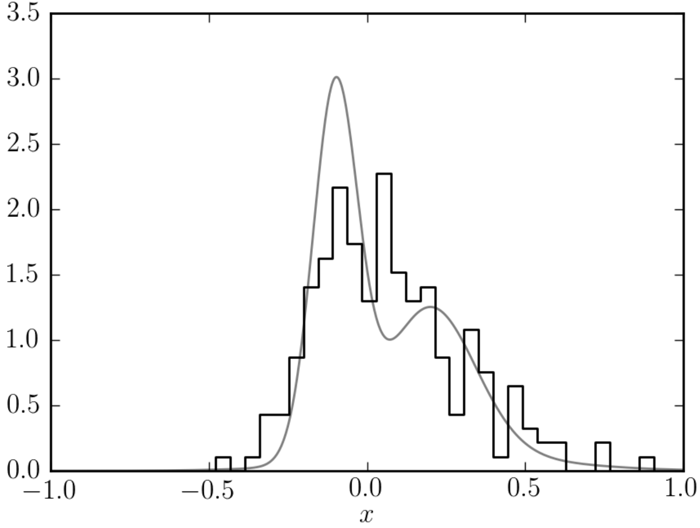
- I have some data $\mathbf{D}$, I need to measure x.
- theoretically inspired arithmetic operations on the data?
- maximum-likelihood estimator?
- *No: full likelihood function* $p(\mathbf{D}|x, \boldsymbol{\alpha})$
- *And* marginalize $p(\mathbf{D}|x) = \int p(\mathbf{D}|x, \boldsymbol{\alpha})\, p(\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\alpha}$
  - like a rotation and projection of the data into the x space
  - as lossless as possible (there are theorems)
  - likelihoods can be combined with other likelihoods to correctly combine multiple data sets relevant to x.

1. Data-driven models

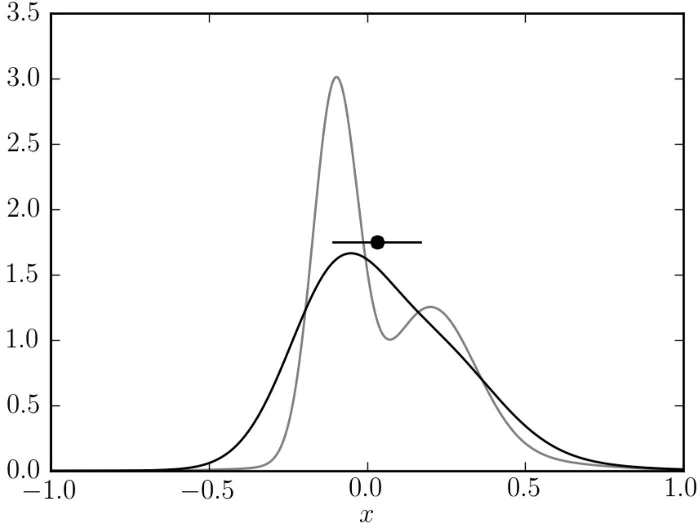
extreme deconvolution (Bovy, Hogg, Roweis 0905.2979): demo



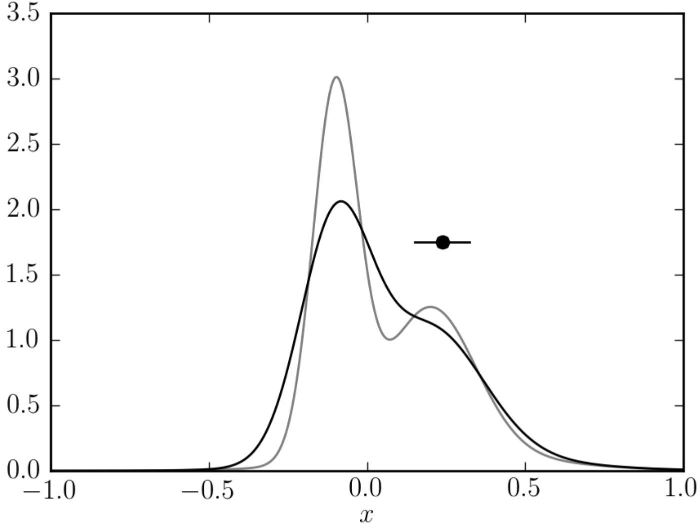
extreme deconvolution (Bovy, Hogg, Roweis 0905.2979): demo



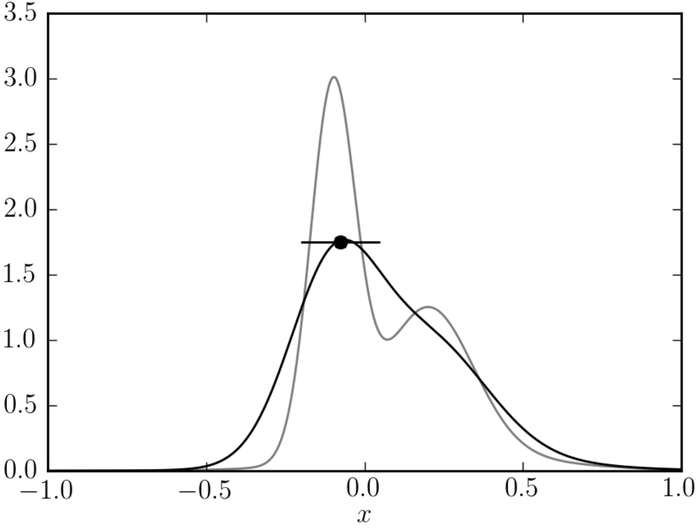
extreme deconvolution (Bovy, Hogg, Roweis 0905.2979): demo



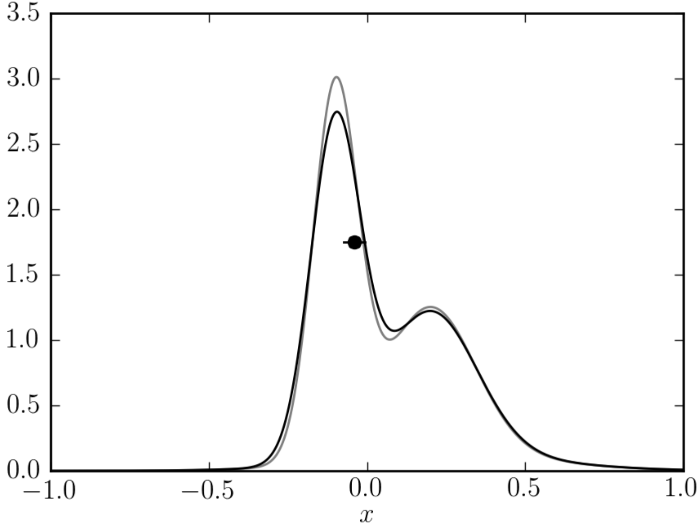
extreme deconvolution (Bovy, Hogg, Roweis 0905.2979): demo



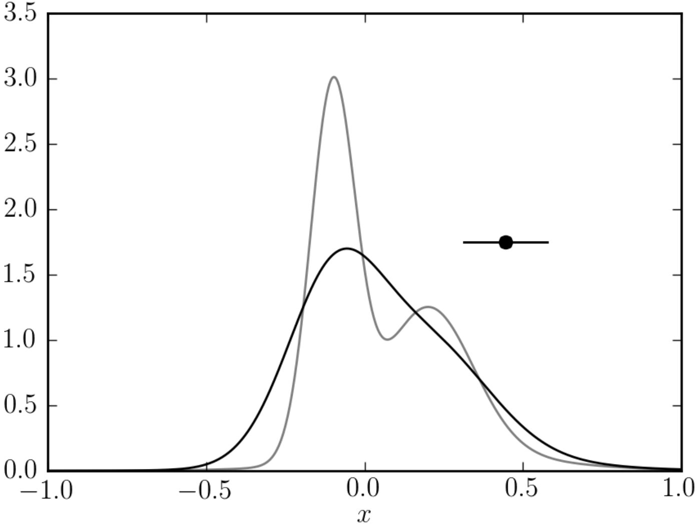
extreme deconvolution (Bovy, Hogg, Roweis 0905.2979): demo



extreme deconvolution (Bovy, Hogg, Roweis 0905.2979): demo



extreme deconvolution (Bovy, Hogg, Roweis 0905.2979): demo



extreme deconvolution (Bovy, Hogg, Roweis 0905.2979): idea

- Each datum x_n has its own error σ_n, therefore
- each datum x_n is drawn from it's own, individual pdf $p(x_n \,|\, \sigma_n, \theta)$.
- Parameterize the true (zero-error) PDF with "hyperparameters" θ and
- find the hyperparameters that optimize the combined likelihood of *all the data*.

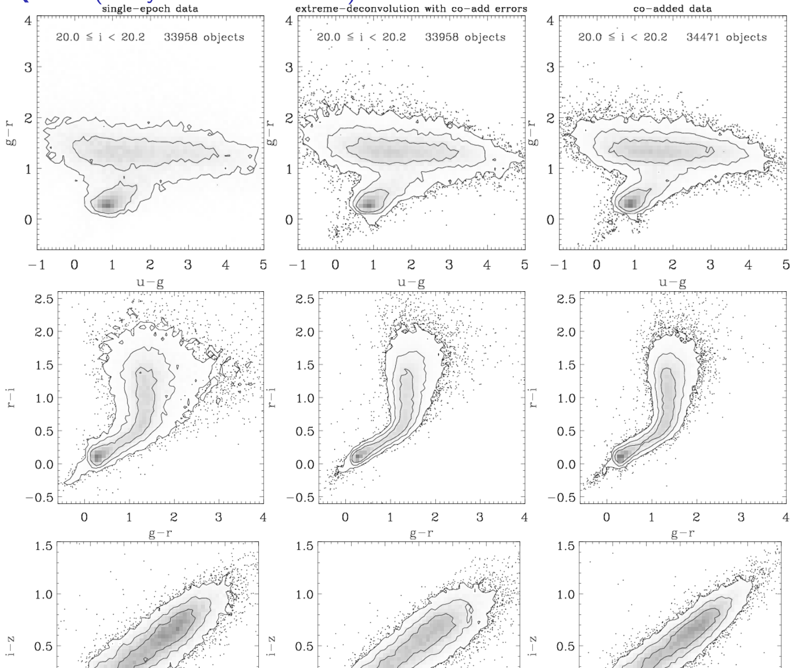
$$p(\{x_n\} \,|\, \theta) \;=\; \prod_n p(x_n \,|\, \sigma_n, \theta) \tag{1}$$

- This is a form of *hierarchical inference*.
- Generalize to D dimensions.

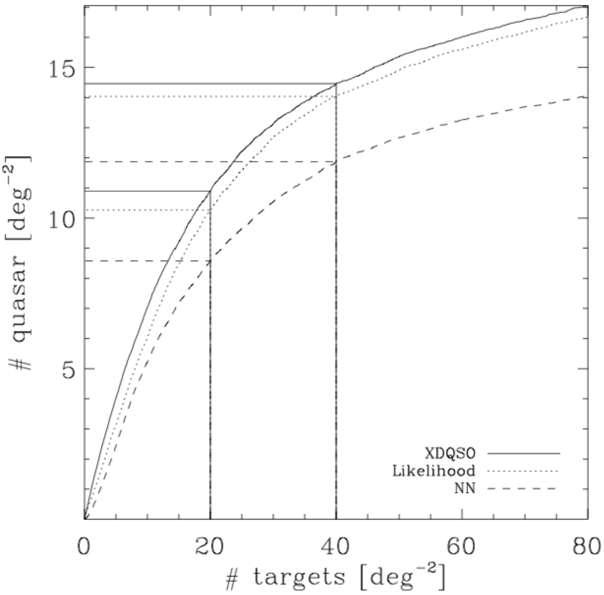
XDQSO (Bovy *et al.* 1011.6392): setup

- $2.2 < z < 3.5$ quasars can be used to measure the baryon acoustic oscillation in the Lyman alpha forest
- *SDSS-III BOSS*
- quasars in this range *look like stars* in *ugriz*
- This is a hard supervised classification problem.

XDQSO (Bovy *et al.* 1011.6392): results



XDQSO (Bovy *et al.* 1011.6392): results



XDQSO (Bovy *et al.* 1011.6392): why do we win?

- We are data-driven.
- We use the errors correctly and account properly for missing data; we have a *generative model*.
- That is true for both the training data and the test data.
  - We can predict high S/N data using *only* low S/N data!
- We are extensible to new prior information or other data.
  - *GALEX*
  - *UKIDSS*
  - variability
- *extreme-deconvolution*
  - Bovy, Hogg, & Roweis (0905.2979)
  - it Just Works (tm)
  - C code with Python and IDL wrappers / interface
  - can handle large data sets with large numbers of dimensions
- *SDSS-III BOSS* core target selection

polemic: What's wrong with typical classification algorithms?

- neural networks, boltzmann machines, support vector machines, boosting
- these are all *awesome*
- they require that *test data* have the same statistical and error properties as *training data*
- they require that all features be measured for all data points

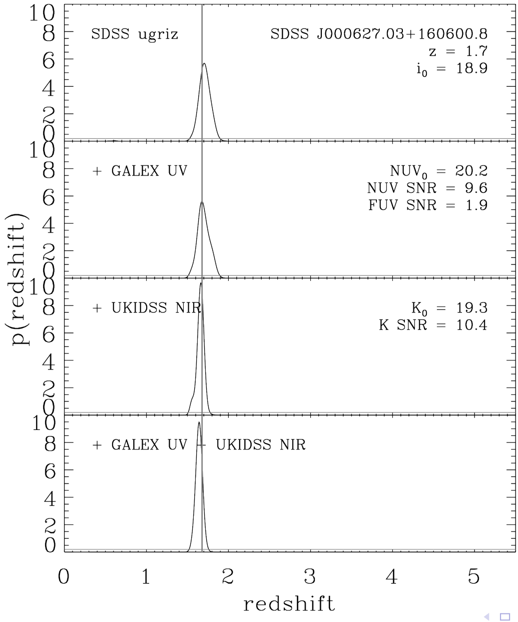
polemic: What's wrong with typical classification algorithms?

- neural networks, boltzmann machines, support vector machines, boosting
- these are all *awesome*
- they require that *test data* have the same statistical and error properties as *training data*
  - *never true!*
- they require that all features be measured for all data points
  - *never true!*

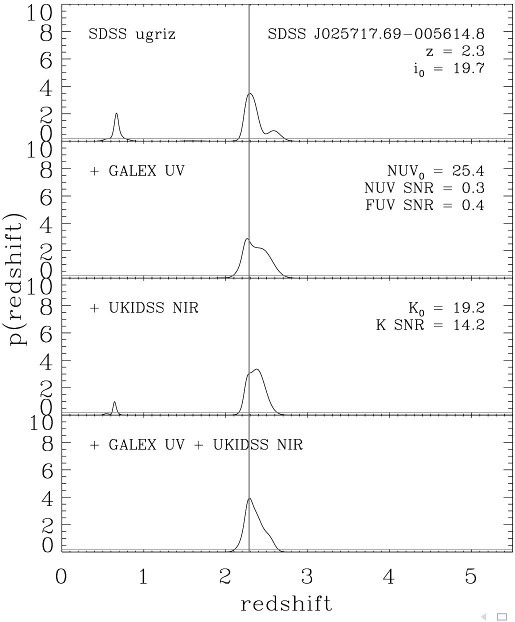
polemic: What's wrong with typical classification algorithms?

- neural networks, boltzmann machines, support vector machines, boosting
- these are all *awesome*
- they require that *test data* have the same statistical and error properties as *training data*
  - *never true!*
- they require that all features be measured for all data points
  - *never true!*
  - (If you know enough about your data to fix this problem, then *just write down a likelihood!*)

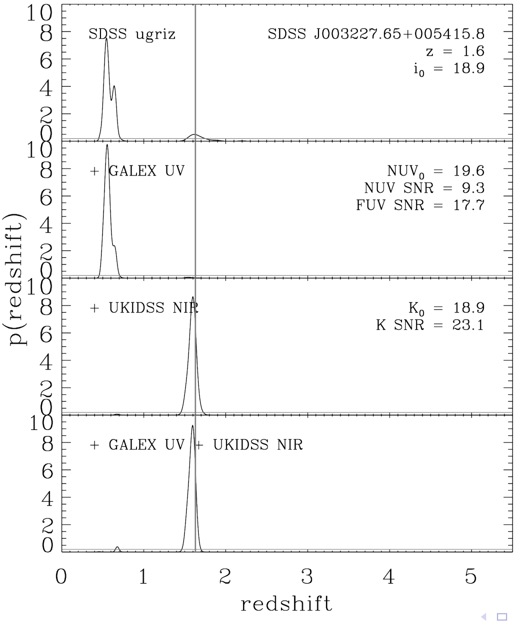
XDQSOz redshift prediction (Bovy *et al.* 1105.3975): example



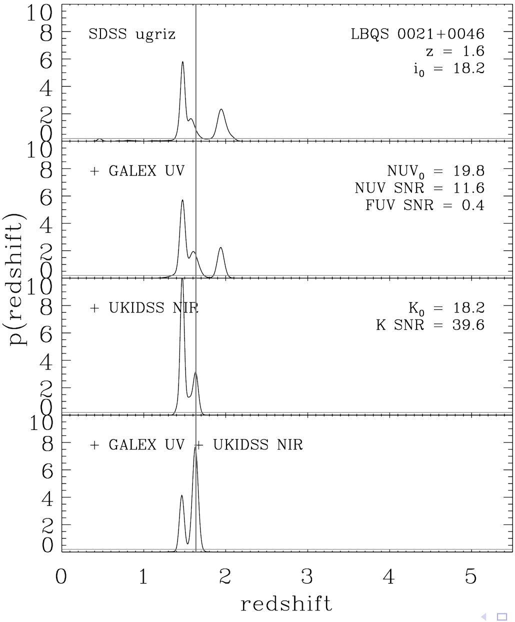
XDQSOz redshift prediction (Bovy *et al.* 1105.3975): example



XDQSOz redshift prediction (Bovy *et al.* 1105.3975): example



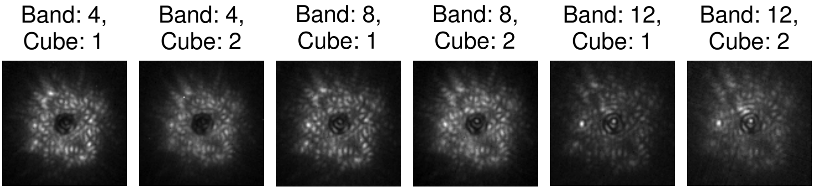
XDQSOz redshift prediction (Bovy *et al.* 1105.3975): example



XDQSOz redshift prediction (Bovy *et al.* 1105.3975): lessons

- When you have a probabilistic generative model, generating the raw data, even *extremely low signal-to-noise data can be decisive*.
- Catalogs are useless in this regime.

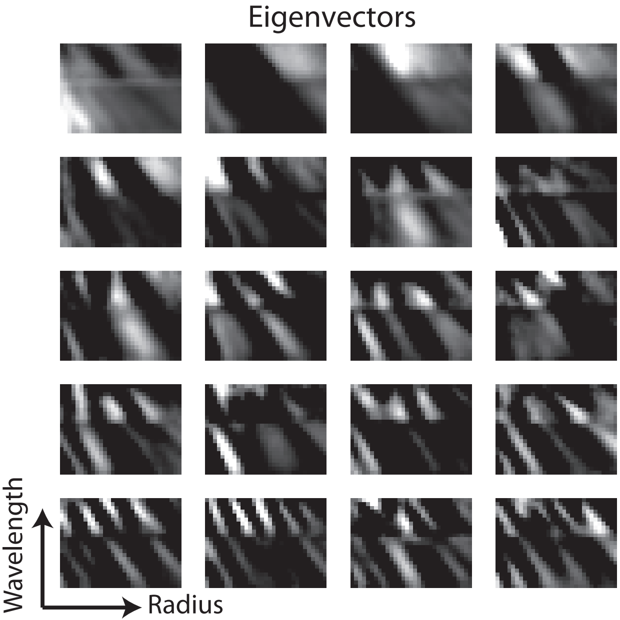
high contrast imaging (Fergus *et al.*): examples



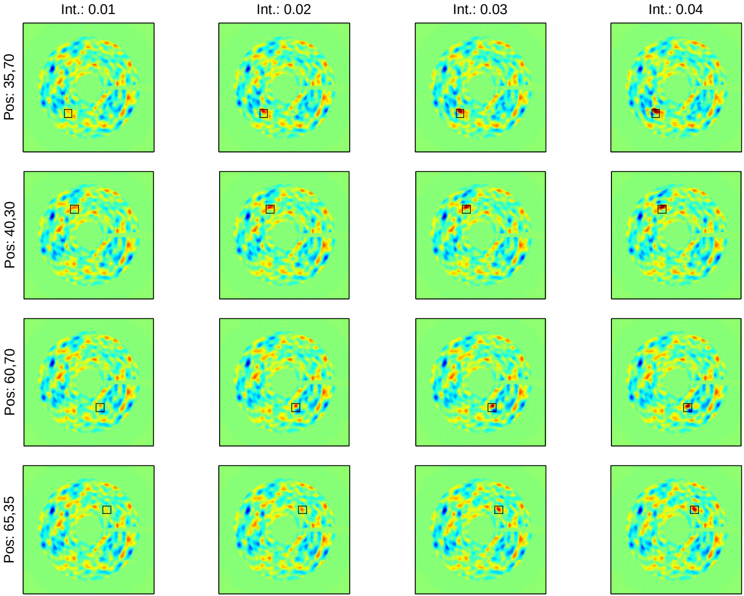
data from the *P1640* spectroscopic imaging coronograph (Oppenheimer *et al.*)

- Data are four dimensional: x, y, λ, n_{exp}.
- Expect strong structure in the radius–wavelength plane.
- We have made the instrument *an order of magnitude* more sensitive, by software alone.

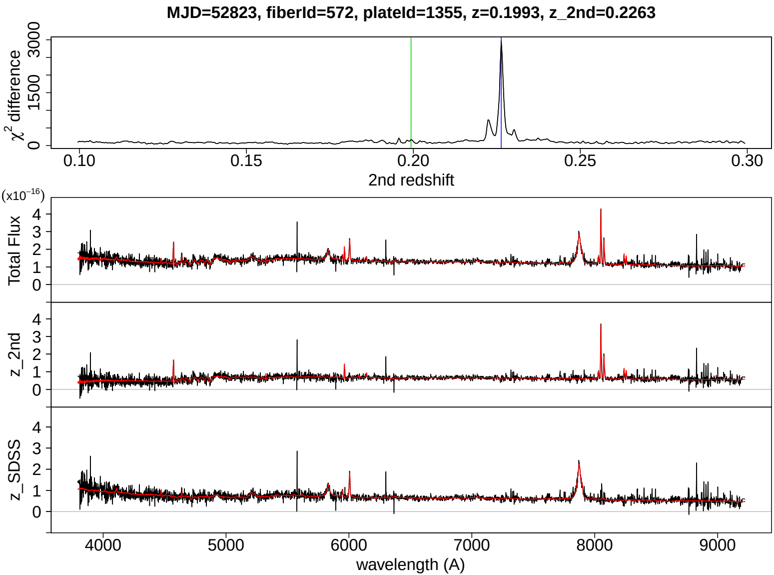
high contrast imaging (Fergus *et al.*): eigenvectors



high contrast imaging (Fergus *et al.*): sensitivity



binary quasars (Tsalmantza *et al.* 1106.1180): example



punchlines

- Probabilistic inference with a generative model beats any point estimate for accuracy and precision.
- When you don't know how to model your data, use the data to build the model; think *hierarchically*.
- You usually need to spend even more time modeling the things you *don't care about*—the noise—than the things you do—the signal.
- "Images $\rightarrow$ coadd $\rightarrow$ catalog $\rightarrow$ best-fit model $\rightarrow$ high-level conclusions" just won't work in many circumstances.
  - warnings for *LSST* and *PanSTARRS* and *Gaia* and ...

2. Foreground-background modeling

GD-1 stream (Grillmair & Dionatos 2006 *ApJL* **643** L17–L20.)

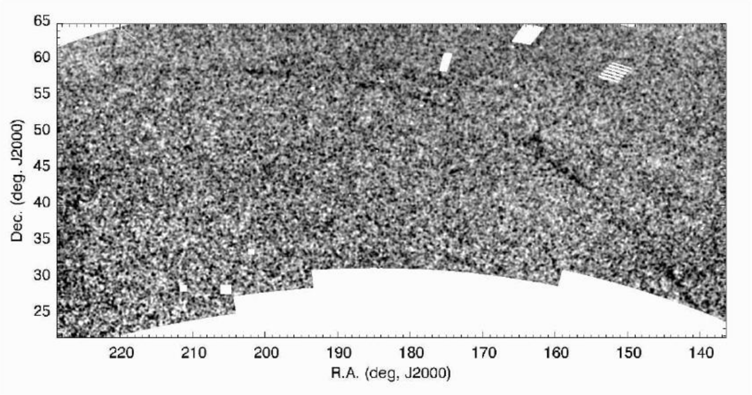


Fig. 1.— Smoothed, summed weight image of the SDSS field after subtraction of a low-order polynomial surface fit. Darker areas indicate higher surface densities. The weight image has been smoothed with a Gaussian kernel with $\sigma = 0.2°$. The white areas are either missing data, or clusters, or bright stars which have been masked out prior to analysis.

GD-1 stream (Koposov *et al.* 0907.1085): setup

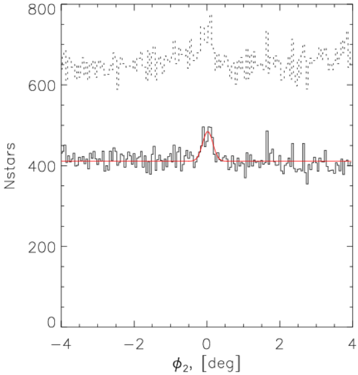


FIG. 2.— Profile in stars with $0.15 < g-r < 0.41$ $18.1 < r < 19.85$ across the $\phi_2 = 0$ axis. The dotted line shows the profile of stars of all stars with -70< ϕ_1¡10. The solid line shows the weighted profile of stars -70< ϕ_1 <10 with weights depending on ϕ_1. The Gaussian fit with 640 stars and sigma=9′is shown in red.

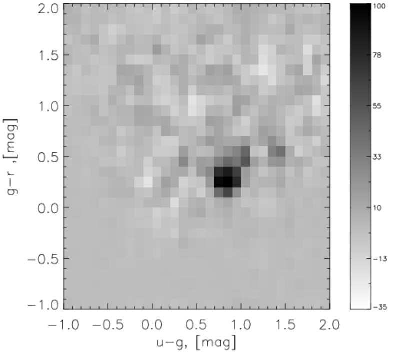


FIG. 3.— Color-color diagram of the stream. The metallicity according to the the Equation 4 from Ivezić et al. (2008) is [Fe/H]=-1.9±0.1

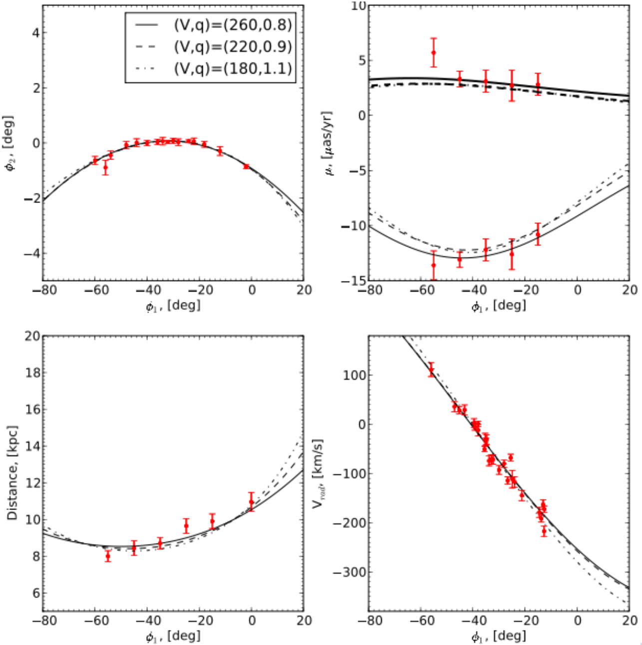
halo stars, and therefore the Ivezić et al. (2008) calibration is correct). We derive that $[Fe/H]_{phot}$=-1.9± 0.1.

To derive the metallicity age and distance in a more

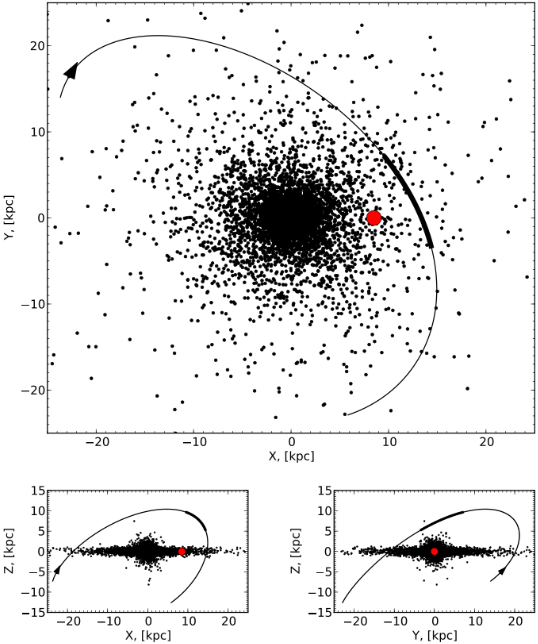
mixture models

[on the board]

GD-1 stream (Koposov *et al.* 0907.1085): results



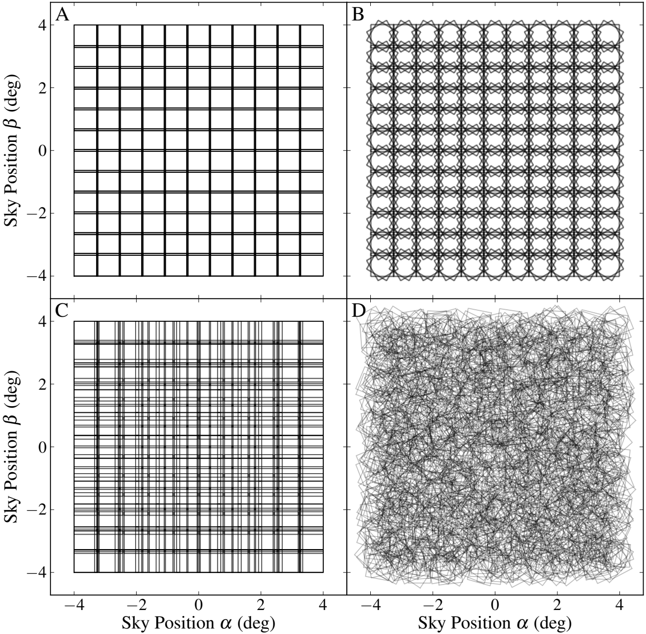
GD-1 stream (Koposov *et al.* 0907.1085): results



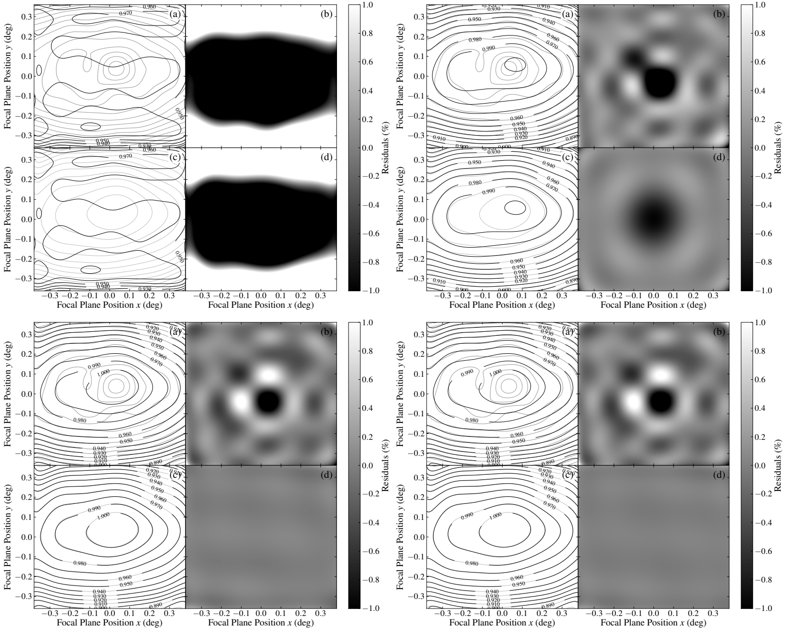
GD-1 stream (Koposov *et al.* 0907.1085): lessons

- We got the first-ever six-dimensional map of an orbit in the Milky Way.
- If we had required hard classification of every star, we would have *failed*.
- We had to put more parameters into our *background model* than the stream!

self-calibration of imaging (Holmes, Rix, Hogg)



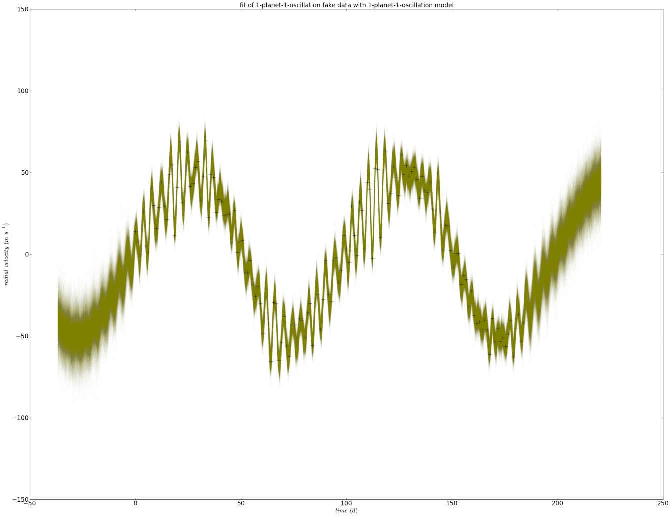
self-calibration of imaging (Holmes, Rix, Hogg)



self-calibration of imaging

- A good survey (Holmes *et al.*):
  - every star appears in many images
  - in different images, the star is in different places
  - every image contains many stars
- A good model (Foreman-Mackey & Hogg):
  - every star has some probability of being variable
    (actually every star is variable with unknown amplitude)
  - every datapoint has some probability of being corrupted
  - calibrate without hard classification
  - mixture model is a marginalization over good–bad decisions
  - *can recover many discarded SDSS-II Stripe 82 imaging runs*

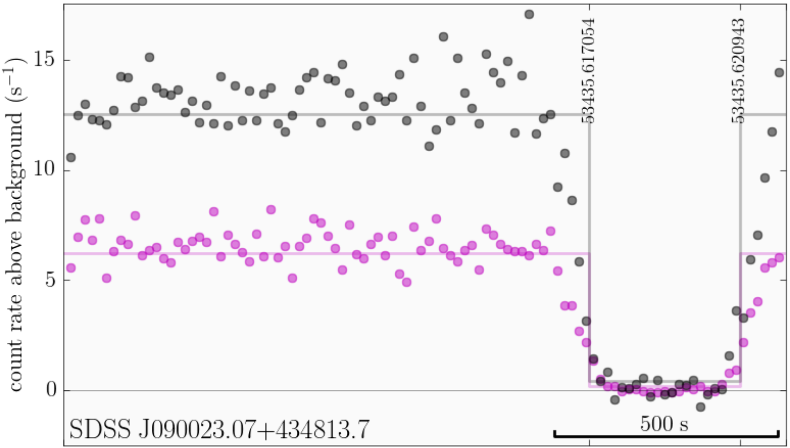
exoplanets around red giants (Hou, Goodman, Hogg)



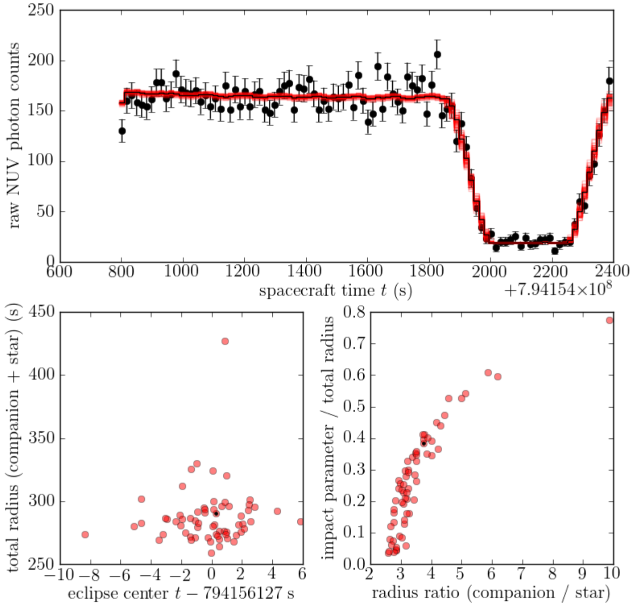
exoplanets around red giants (Hou, Goodman, Hogg)

- stars (especially giants) have surface oscillations
- radial-velocity signal is a superposition of exoplanet and oscillations
- need methods to model stochastically driven, damped oscillators
  - This is what *Gaussian Processes* are designed to do!
  - Also very appropriate for transits in the presence of *stochastic intensity variations*.
  - Inference is expensive. Suck it up.

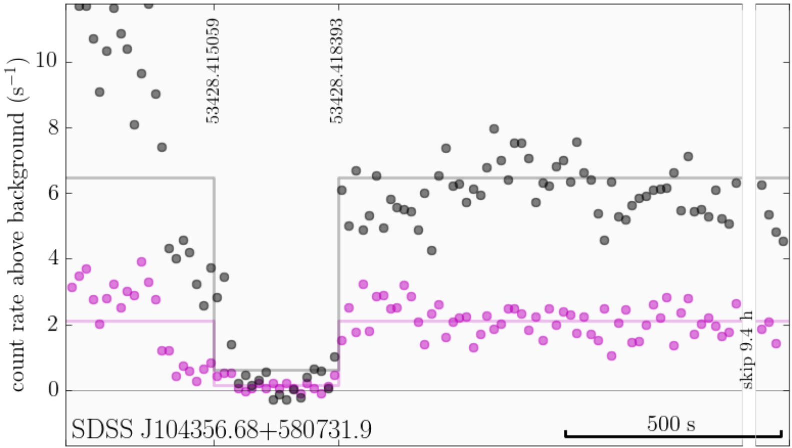
exoplanets around white dwarfs (Schiminovich, Lang, Hogg)



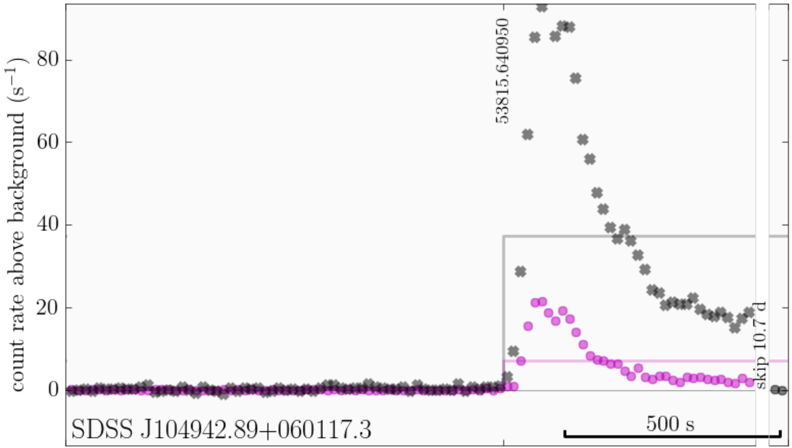
exoplanets around white dwarfs (Schiminovich, Lang, Hogg)



exoplanets around white dwarfs (Schiminovich, Lang, Hogg)



exoplanets around white dwarfs (Schiminovich, Lang, Hogg)

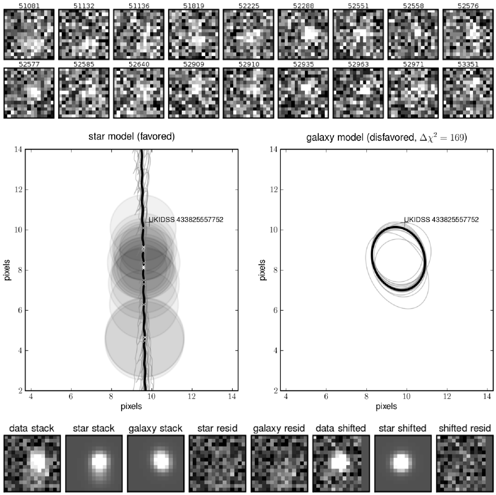


punchlines

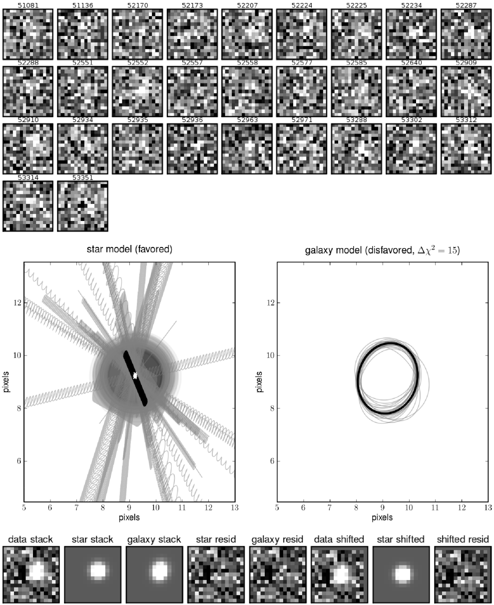
- Probabilistic inference with a generative model beats any point estimate for accuracy and precision.
- When you don't know how to model your data, use the data to build the model; think *hierarchically*.
- You usually need to spend even more time modeling the things you *don't care about*—the noise—than the things you do—the signal.
- "Images $\rightarrow$ coadd $\rightarrow$ catalog $\rightarrow$ best-fit model $\rightarrow$ high-level conclusions" just won't work in many circumstances.
  - warnings for *LSST* and *PanSTARRS* and *Gaia* and ...

3. Catalogs are bad; unstacked images are good

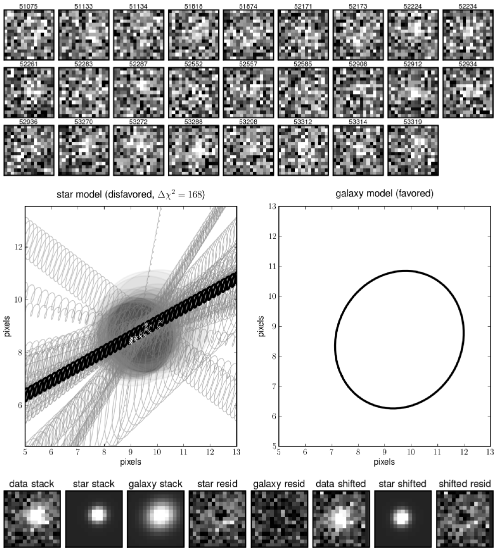
faint proper motions (Lang *et al.* 0808.4004): brown dwarf



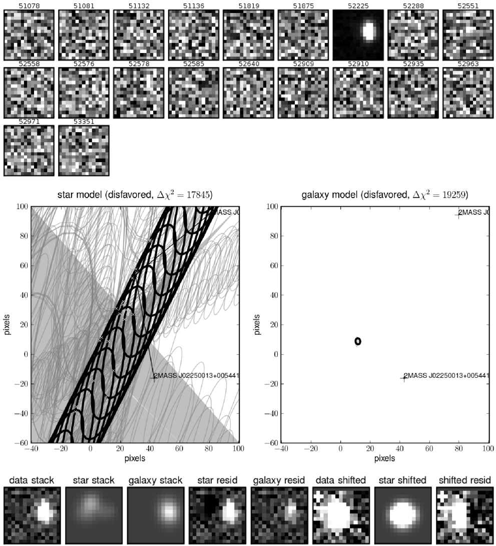
faint proper motions (Lang *et al.* 0808.4004): $z > 6$ QSO



faint proper motions (Lang *et al.* 0808.4004): faint galaxy



faint proper motions (Lang *et al.* 0808.4004): defect



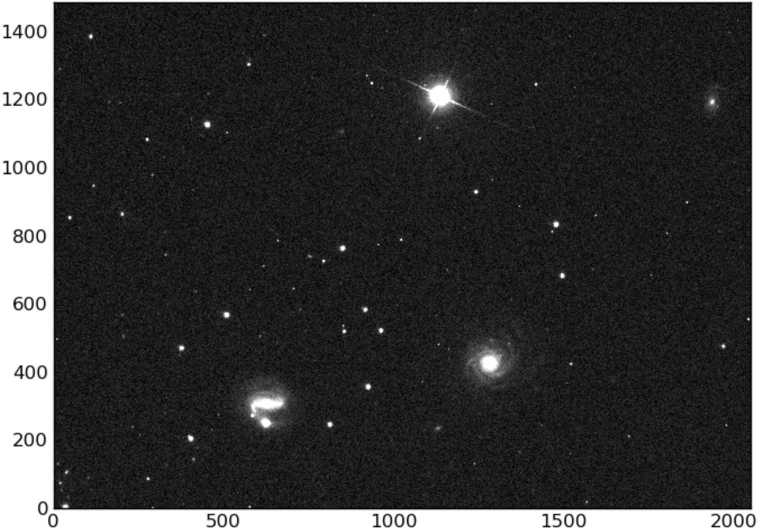
faint proper motions (Lang *et al.* 0808.4004): lessons

- If we had only a catalog, we would have *failed*.
- If we had only a coadd, we would have *failed*.

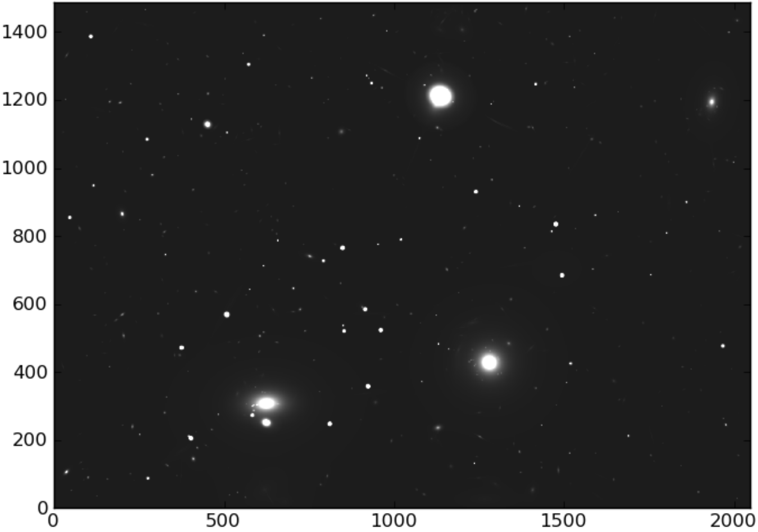
what's wrong with *LSST* and *PanSTARRS*?

- reducing data with point estimates
- building catalogs from "co-adds" with point estimates
- catalog matching
- *All of these throw away information. Does it matter?*
    - Lang and I are betting it does: *theTractor.org*

The *Tractor* (Lang *et al.*): data



The *Tractor* (Lang *et al.*): model



punchlines

- Probabilistic inference with a generative model beats any point estimate for accuracy and precision.
- When you don't know how to model your data, use the data to build the model; think *hierarchically*.
- You usually need to spend even more time modeling the things you *don't care about*—the noise—than the things you do—the signal.
- "Images $\rightarrow$ coadd $\rightarrow$ catalog $\rightarrow$ best-fit model $\rightarrow$ high-level conclusions" just won't work in many circumstances.
  - warnings for *LSST* and *PanSTARRS* and *Gaia* and ...

[polemical backup slides]

polemic: Weak lensing

- work very hard to make sensitive morphological measurements (think "ellipticities") on millions of galaxies
- then simply *average* in bins on the sky to make a shear map!
- Hierarchical inference beats averaging, provably.

$$p(\mathrm{data} \mid \mathrm{map}, \alpha) = \prod_n \int p(d_n \mid \mathrm{map}, \alpha) \tag{2}$$

$$p(d_n \mid \mathrm{map}, \alpha) = \int p(d_n \mid s_n, \mathrm{map})\, p(s_n \mid \alpha)\, \mathrm{d}s_n \tag{3}$$

$$p(\mathrm{data} \mid \alpha) = \int p(\mathrm{data} \mid \mathrm{map}, \alpha)\, p(\mathrm{map})\, \mathrm{d\,map} \tag{4}$$

polemic: The baryon acoustic feature

- build galaxy catalog from noisy imaging and spectroscopy data
- build two-point function from catalog
- fit baryon acoustic feature to two-point function
- *Can we write down the likelihood instead?*
  - model the density field
  - model how galaxies populate that field
  - enormous marginalization
  - impossible? (*e.g.*, Dodelson *et al.* 9712074)
- If we fail, will S/N rise with survey size?
  - certainly not guaranteed

polemic: Missing data

- Most machine-learning methods hate missing data.
- Interpolation or data censoring (both very, very bad) are required.
- Any model that properly accounts for *uncertainty* also properly accounts for *missing data*.
  - Missing data is (extreme) uncertainty; uncertainty is (mild) missing data.
- If you have a justified generative model $p(\mathbf{D}_n|\omega_n)$, you automatically deal with missing data.

polemic: Don't convolve your data, convolve your model!

- If you are uncertain about something (a redshift, a classification) so that you don't know which bin to put it in:
- *Don't* put a bit of your *data point* into each bin!
  - That re-convolves your noisy result with the noise again.
- *Do* put a bit of your *distribution model* into each bin.
  - That is, convolve your *model* for the object with the uncertainty.
  - Obvious, but frequently done wrong.

polemic: Catalogs are dangerous (Hogg & Lang 1008.0738)

- No objects are detected or classified with perfect confidence.
- Different investigators have different objectives and priors.
- As new data become available, the balance will shift for many objects.
- *Catalogs become wrong, likelihood functions are forever.*
  - and I mean *functions*, not optima