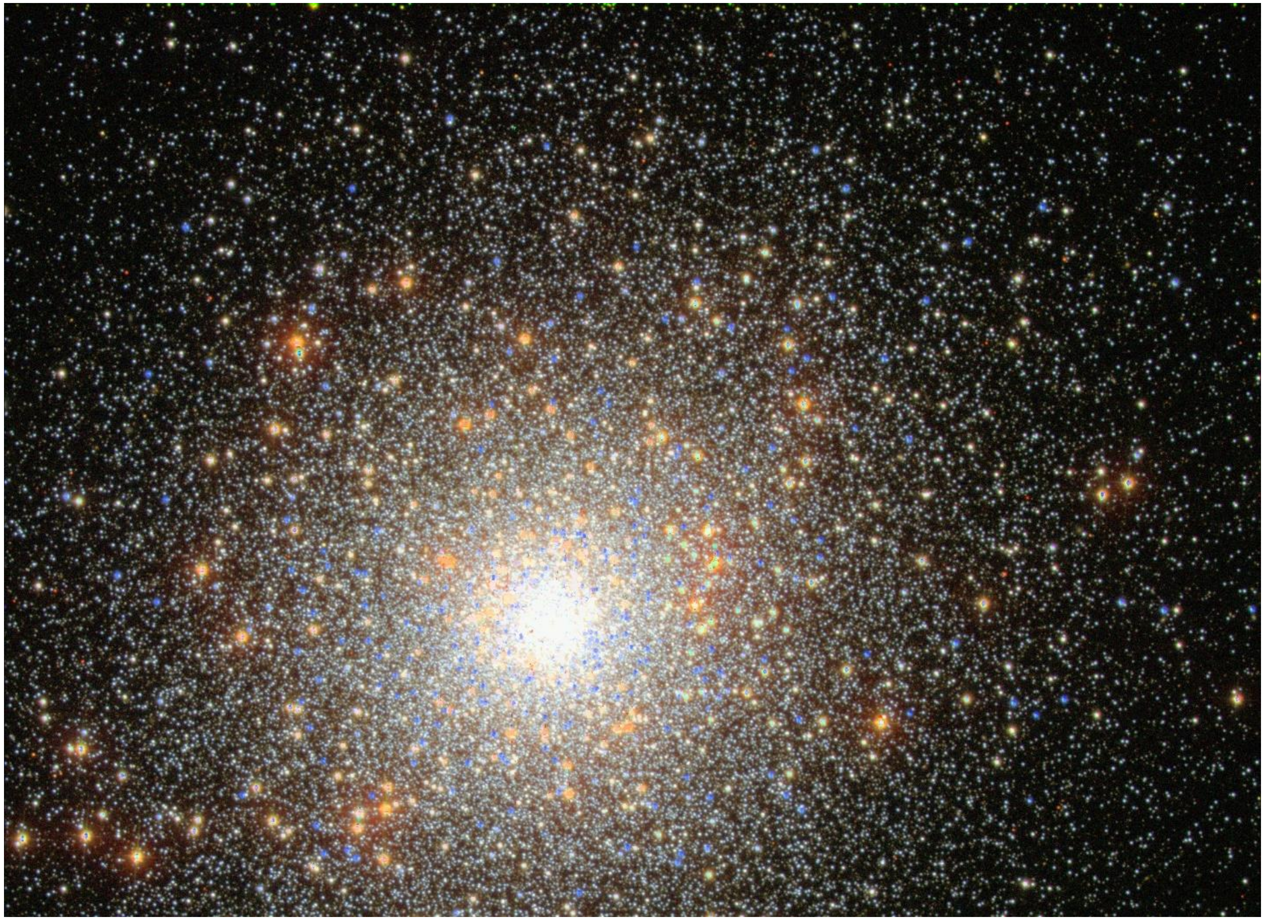
Probabilistic Cataloguing

Stephen K N Portillo
with Benjamin Lee, Tansu Daylan and
Douglas P Finkbeiner
25 October 2016
Harvard Astrostatistics Group

NSERC
CRSNG

VE RI TAS

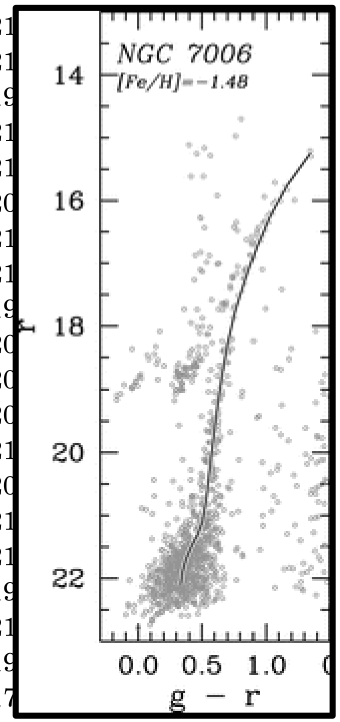
Telescopes don't make catalogues!



Slide title stolen from D.W. Hogg and D. Lang, EAS Publication Series 45, 351 (2011)
Image: Sloan Digital Sky Survey, DR12

People make catalogues

RA (J2000)	DEC (J2000)	g	r
229.4351	2.010923	19.344	19.114
229.4353	1.990166	23.07	21
229.4358	2.033374	21.809	21
229.4361	2.070269	20.107	19
229.4362	1.997957	22.894	21
229.4364	2.048578	22.386	21
229.4366	2.053515	20.853	20
229.4369	2.103516	21.827	21
229.4369	2.043476	23.067	21
229.437	2.051732	19.96	19
229.4371	2.102266	20.813	20
229.4373	2.052342	20.785	20
229.4374	1.996688	21.161	20
229.4376	2.13321	22.476	21
229.4378	2.039289	20.883	20
229.438	2.077996	22.682	21
229.438	2.043483	22.884	21
229.4381	2.045585	20.111	19
229.4382	2.011463	22.069	21
229.4382	2.029807	19.625	19
229.4382	2.030182	17.835	17
229.4385	2.157053	22.193	21.877
229.4385	2.147021	22.492	21.546
…	…	…	…



An et al. ApJS 179, 328 (2008)

(Deterministic) Catalogues

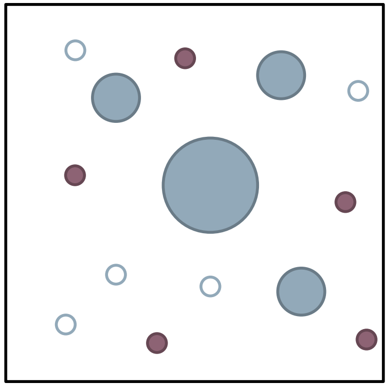
- A (deterministic) catalogue is a list of point source candidates above some inclusion threshold TS_{incl}

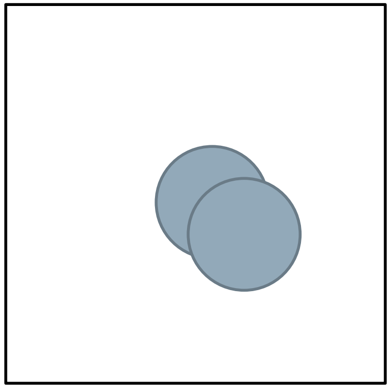
  $$Data, TS_{incl} \rightarrow \{\ell_i \pm \sigma_{\ell_i}, b_i \pm \sigma_{b_i}, F_i \pm \sigma_{F_i}\}_{i=1}^{N}$$

- **Inclusion threshold = detection threshold:**
  Almost all catalogue sources are true sources
  But faint true sources are not in the catalogue

- **Inclusion threshold < detection threshold:**
  More faint true sources are included in the catalogue
  But many catalogue sources are not true sources
  The data is overfitted

- Overlapping point sources may not be deblended



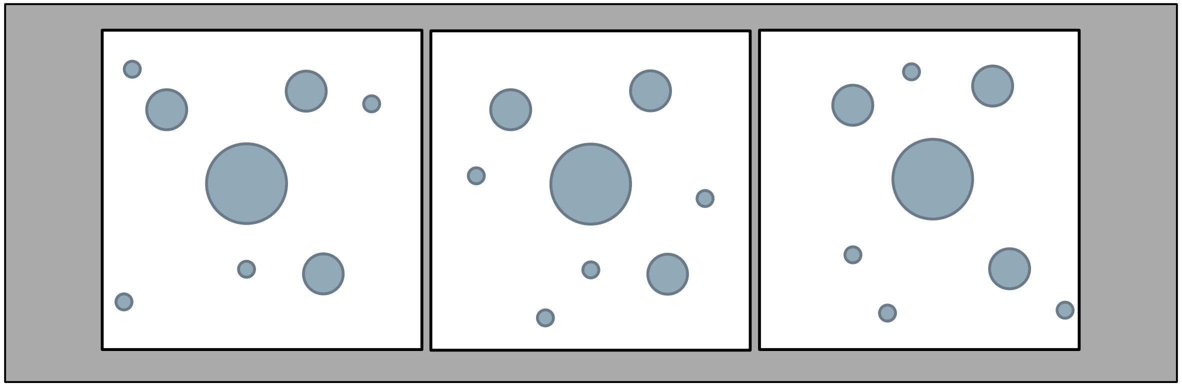


Probabilistic Catalogues

- A probabilistic catalogue is a posterior probability distribution over the space of lists of point source candidates

$$P\left(\{\ell_i, b_i, F_i\}_{i=1}^{N} \middle| Data\right)$$
$$= \pi\left(\{\ell_i, b_i, F_i\}_{i=1}^{N}\right)\mathcal{L}(Data|\{\ell_i, b_i, F_i\}_{i=1}^{N})$$

- Sampling the probabilistic catalogue provides an ***ensemble of catalogues*** inferred from the data



D. W. Hogg and D. Lang, arXiv:1008.0738v1 (2010)
B. J. Brewer, D. Foreman-Mackey, and D. W. Hogg, ApJ 146, 7 (2013)

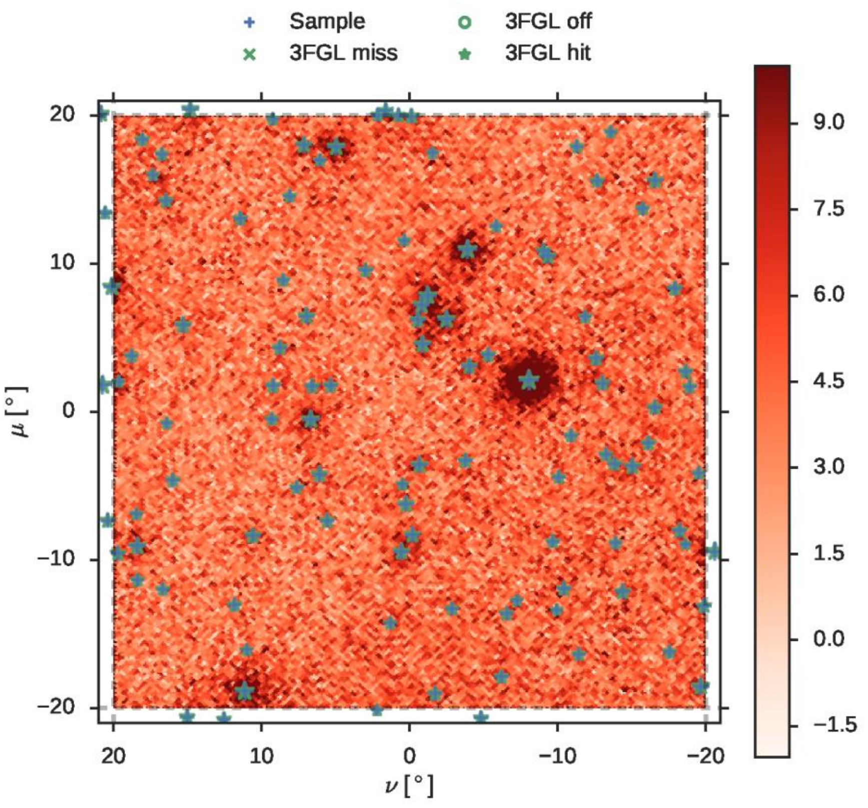
Why Probabilistic Catalogues?

- The reality of a single faint point source candidate will be very uncertain, but the properties of a faint population are constrained
- The uncertainty in deblending sources with overlapping PSFs can be captured
- Provides a framework to marginalize over uncertainties (modelling, instrumental, calibration, etc.)
- Probabilistic cataloguing more fully captures the information contained in the data and the ***inherent degeneracies*** of point source identification

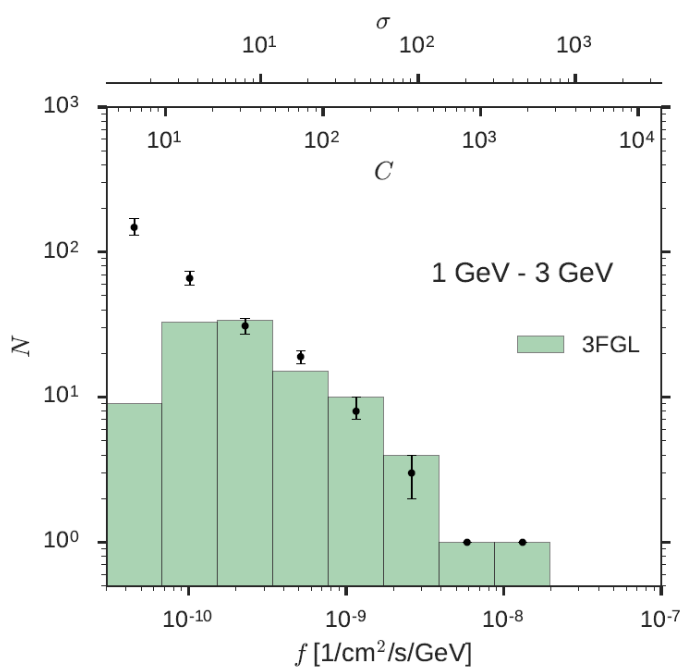
Application I: *Fermi* High Latitude

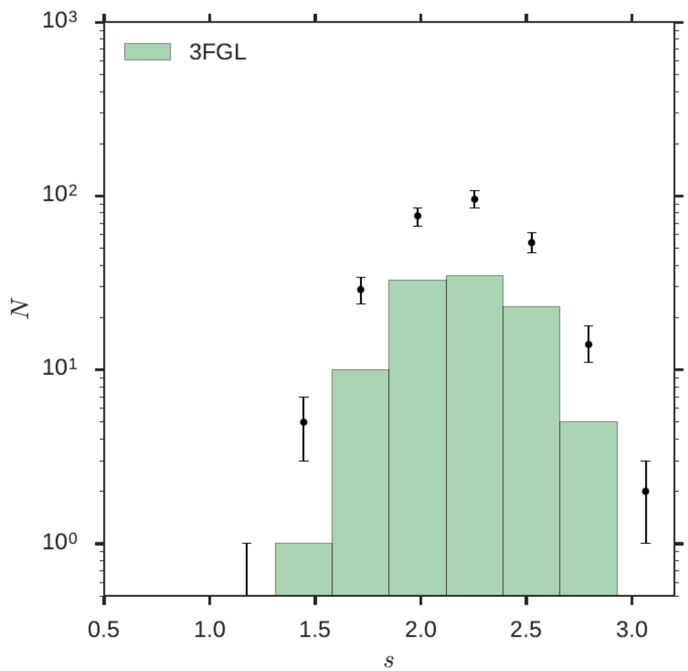
- North Galactic Pole $20^\circ \times 20^\circ$ ($N_{pix} = 29\,880$)
- 3 energy bins: 0.3-1 GeV, 1-3 GeV, 3-10 GeV
- Region includes 108 3FGL sources
- Run with ~250 CPU-hours
- Diffuse sources:
  - Galactic diffuse emission
  - Isotropic emission
- Point source population:
  - Mostly distant active galaxies
  - Assumed to be isotropically distributed
  - Unknown flux distribution parameterized as power law

Catalogue Samples



Flux and Colour Distributions

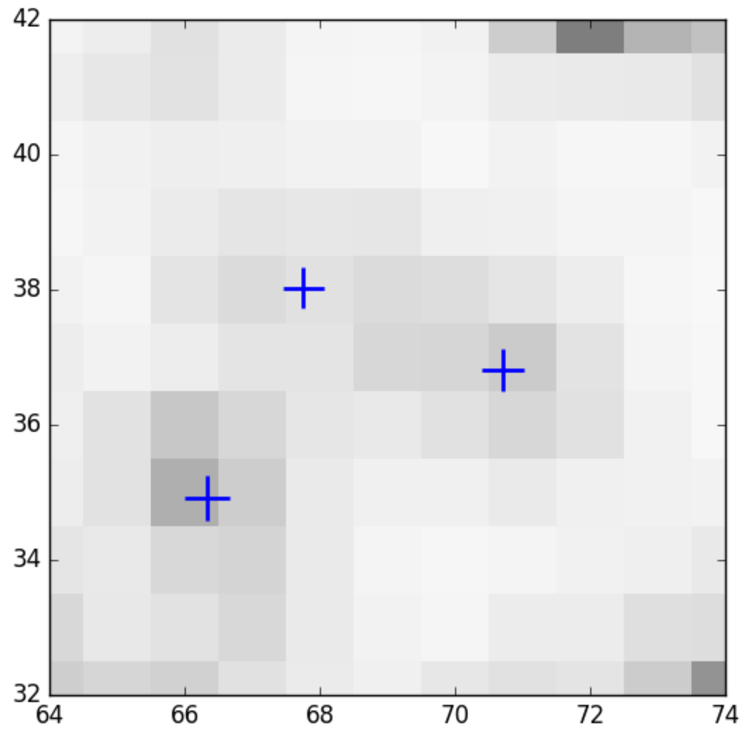




Application II: SDSS Globular Cluster

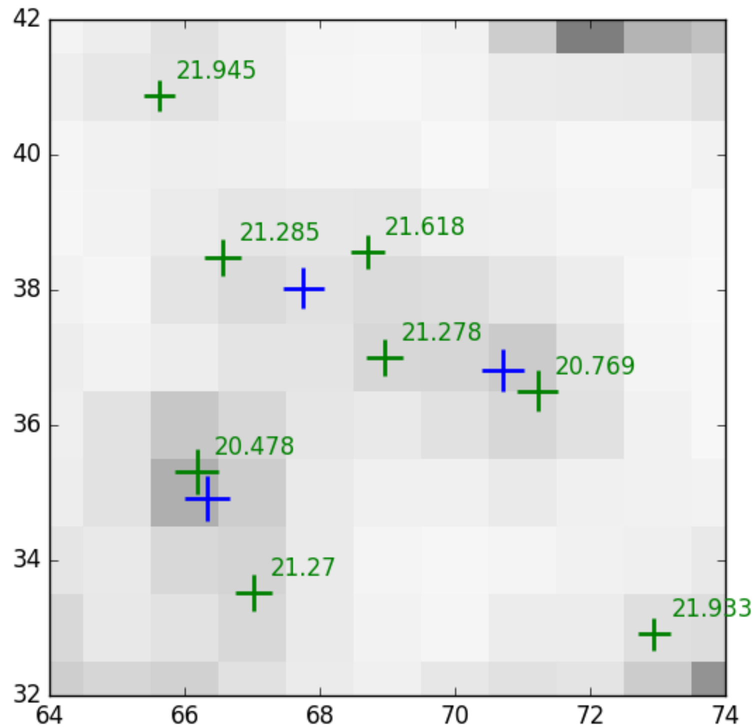
- Messier 2 $40'' \times 40''$ ($N_{pix} = 10\,000$)
- Region includes 337 DAOPhot sources
- Run with ~250 CPU-hours
- Region has also been observed with HST, which has better angular resolution, identifying 1 000 sources

Deterministic Catalogue of SDSS Data



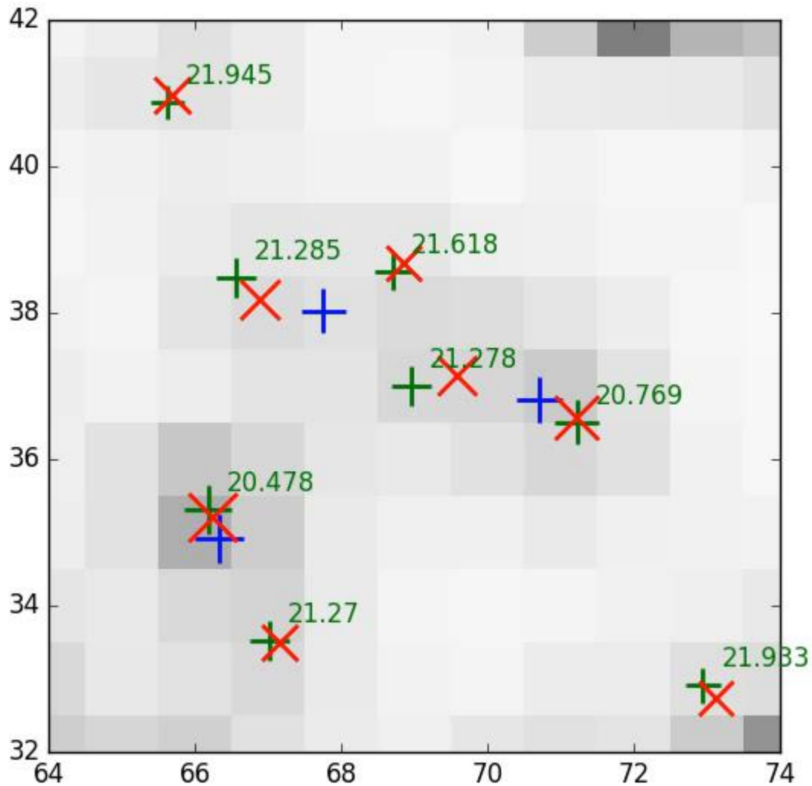
An, D. et al. (2008) ApJS, 179, 2

Deterministic Catalogue of HST Data



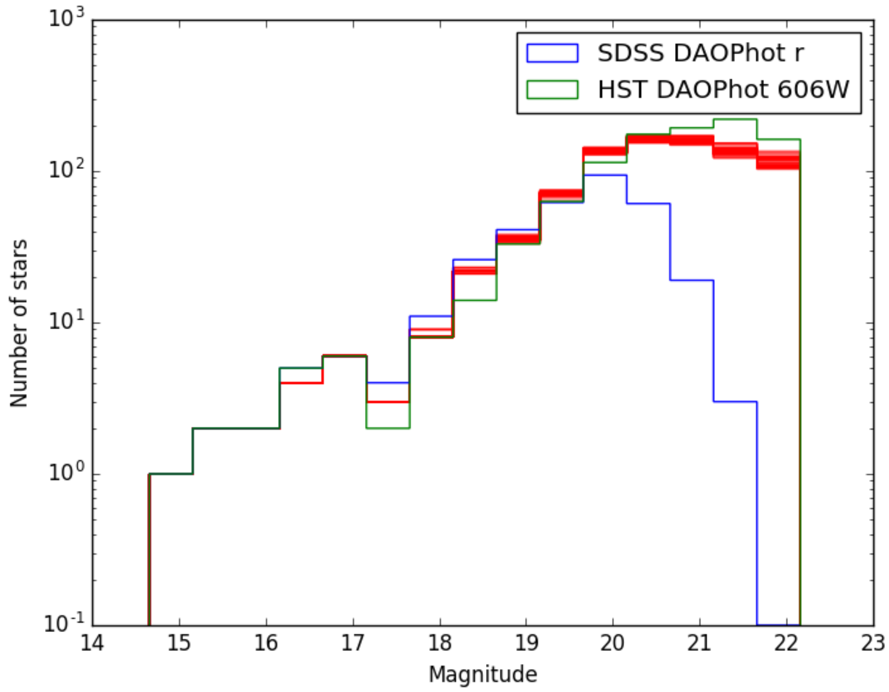
Sarajedini, A. et al. (2007) AJ, 133, 1658

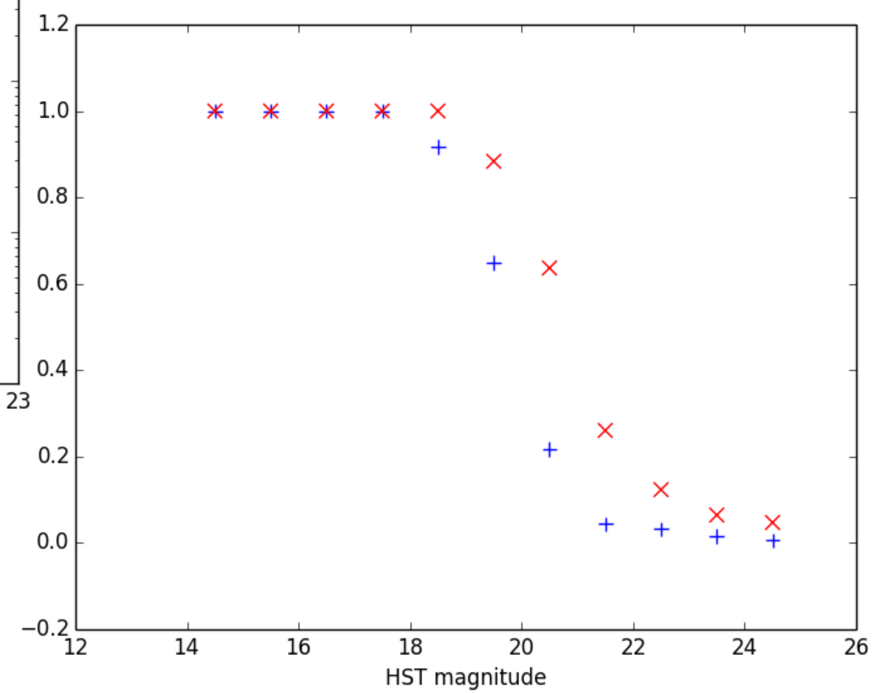
Probabilistic Catalogue of SDSS Data



SDSS DAOPhot
HST DAOPhot
SDSS PCat-Dnest

Completeness





Reversible Jump MCMC

- Allows proposals to change dimensionality of model
  - Move m takes x and generates auxillary u to propose x'
  - Move m' takes x' and generates auxillary u' to propose x
  - $\dim x + \dim u = \dim x' + \dim u'$ and $(x, u) \leftrightarrow (x', u')$ one-to-one

$$\alpha(x \to x') = \min\left(1, \frac{\pi(x')}{\pi(x)} \frac{\mathcal{L}(x'|D)}{\mathcal{L}(x|D)} \frac{j_{m'}(x')}{j_m(x)} \frac{g(u')}{g(u)} \left|\frac{\partial(x', u')}{\partial(x, u)}\right|\right)$$

- For example, birth/death between $x = \{x_1, \ldots, x_N\}$ and $x' = \{x_1, \ldots, x_{N+1}\}$ has $u = x_{N+1}$ and $u' = \emptyset$
  - If birth and death equally likely, sources independent in prior and new source x_2 generated from prior

$$\alpha(x \to x') = \min\left(1, \frac{\pi(N+1)}{\pi(N)} \frac{\mathcal{L}(x'|D)}{\mathcal{L}(x|D)}\right)$$

P.J. Green, Biometrika 82, 711 (1995)

Catalogue Priors

- Prior that sources are independent and described by population parameters β:

$$\pi\left(\{\ell_i, b_i, F_i\}_{i=1}^{N}, \beta\right) = \pi(\beta)\pi(N|\beta) \prod_{i=1}^{N} \pi(\ell_i, b_i, F_i|\beta)$$

- β can describe both spatial and flux distributions
- What should the prior on the number of sources look like? What do we mean by "the number of sources"?

How many sources are there with a flux above F_{min}?

- Prior on N through putting a log uniform prior on expected number of sources $\langle N \rangle$?

$$\log \frac{\pi(N+1)}{\pi(N)} = \log N - \log(N+1) \approx -\frac{1}{N}$$

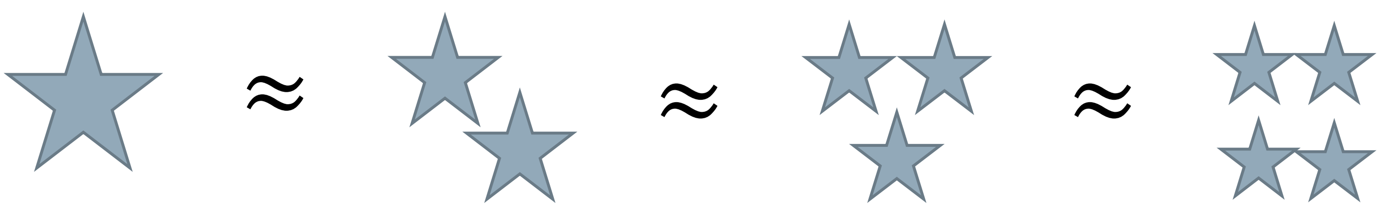
Source Number Prior

- But is this prior enough to counteract the fact that models with more sources will fit better?

- What about a prior that penalizes the $(N+1)^{\text{th}}$ source based on the expected improvement in χ^2 under the null hypothesis that there are N sources?

$$\log \frac{\pi(N+1)}{\pi(N)} = -\frac{3}{2}$$

How many sources meaningfully affect the current data?



What is the most compact representation of the data?

Conclusion

- Probabilistic catalogue samples are an ensemble of catalogues inferred from the data
- A point source population can be distinguished from a diffuse source, even if the individual sources are below the detection threshold
- Overlapping point sources can be better deblended
- This ensemble of catalogues captures the inherent degeneracies of point source identification