Probabilistic decision-making, data analysis, and discovery in astronomy

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

2012 April 20

Machine learning in astronomy

- classification (stars vs quasars, stars vs galaxies)
- anomaly detection
- data structures, search, retrieval
- modeling with large (or infinite) numbers of parameters
- decision making, resource allocation, experimental design
- *this workshop will not be a survey of the field*

Openness

- Everything I do is on the web all the time.
- Most astronomers lie on the "open" side of the openness spectrum.
- *Almost all astronomical data are free on the web for anyone to use.*

principal collaborators

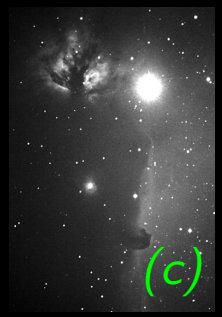
- **Jo Bovy** (IAS)
- **Dustin Lang** (Princeton $\rightarrow$ CMU)
- Sam Roweis (deceased)

Modeling noisy data

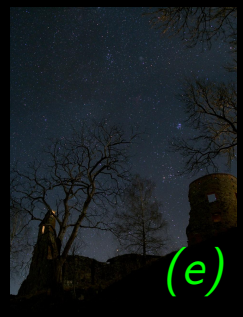
search “Comet Holmes” on *Yahoo!*

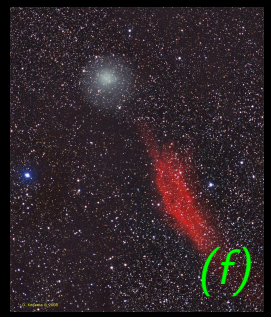

*(a)

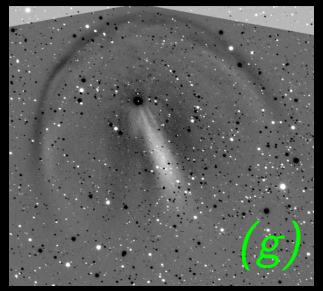

(b)


(c)


(d)

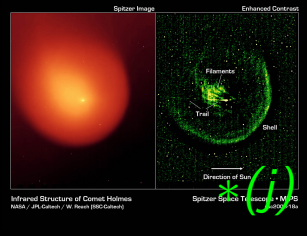

(e)


(f)


(g)


(h)


*(i)



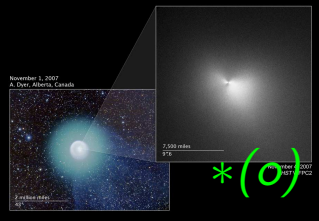
*(j)


*(k)


*(l)

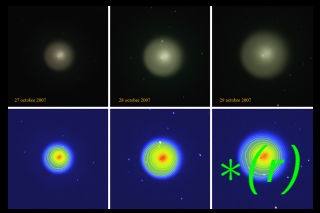

(m)


(n)

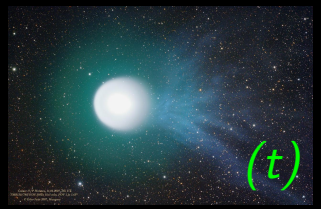

*(o)


(p)


(q)

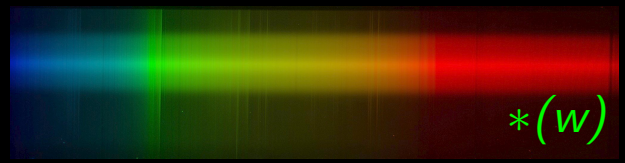

*(r)


*(s)

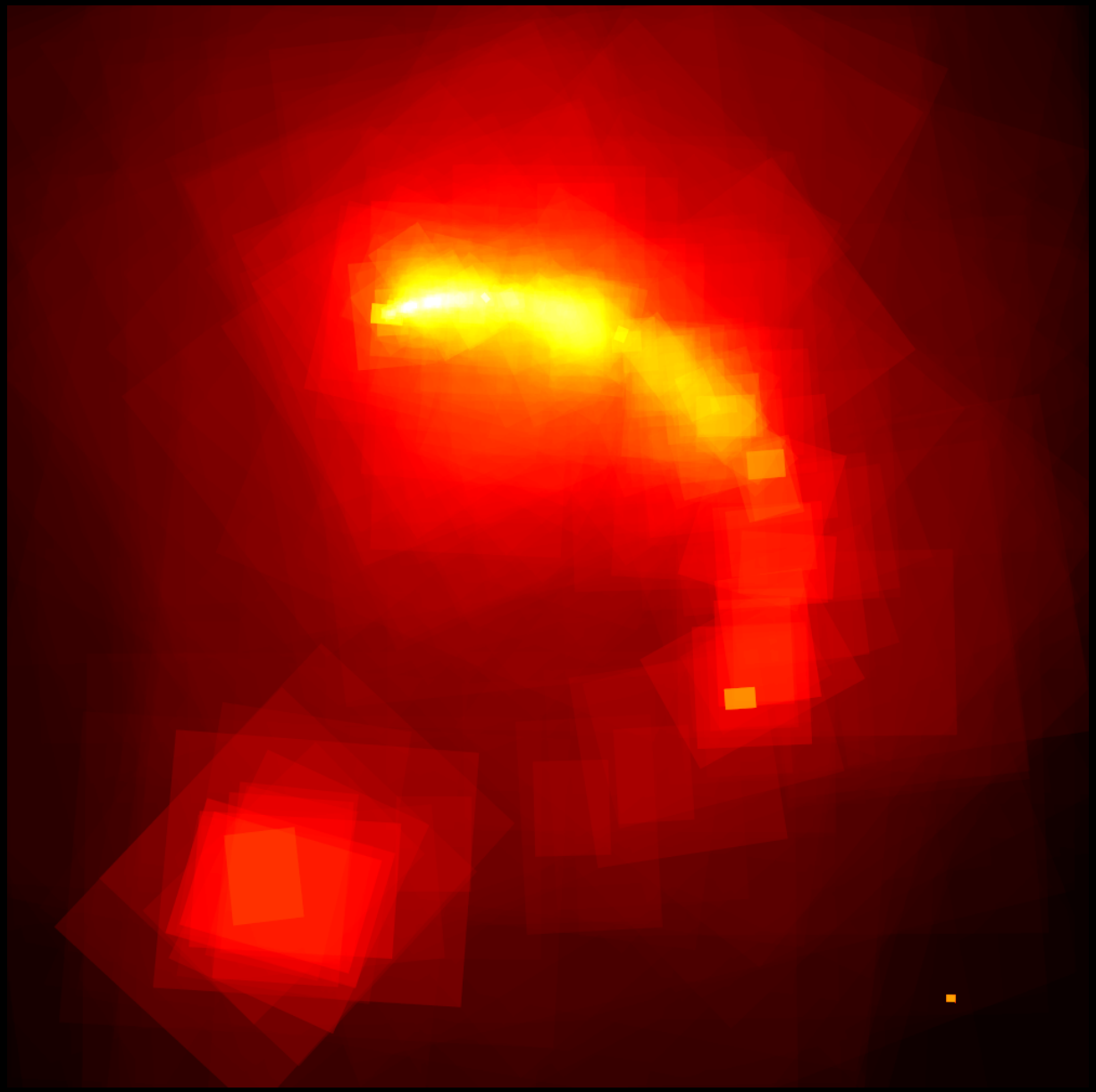

(t)


(u)

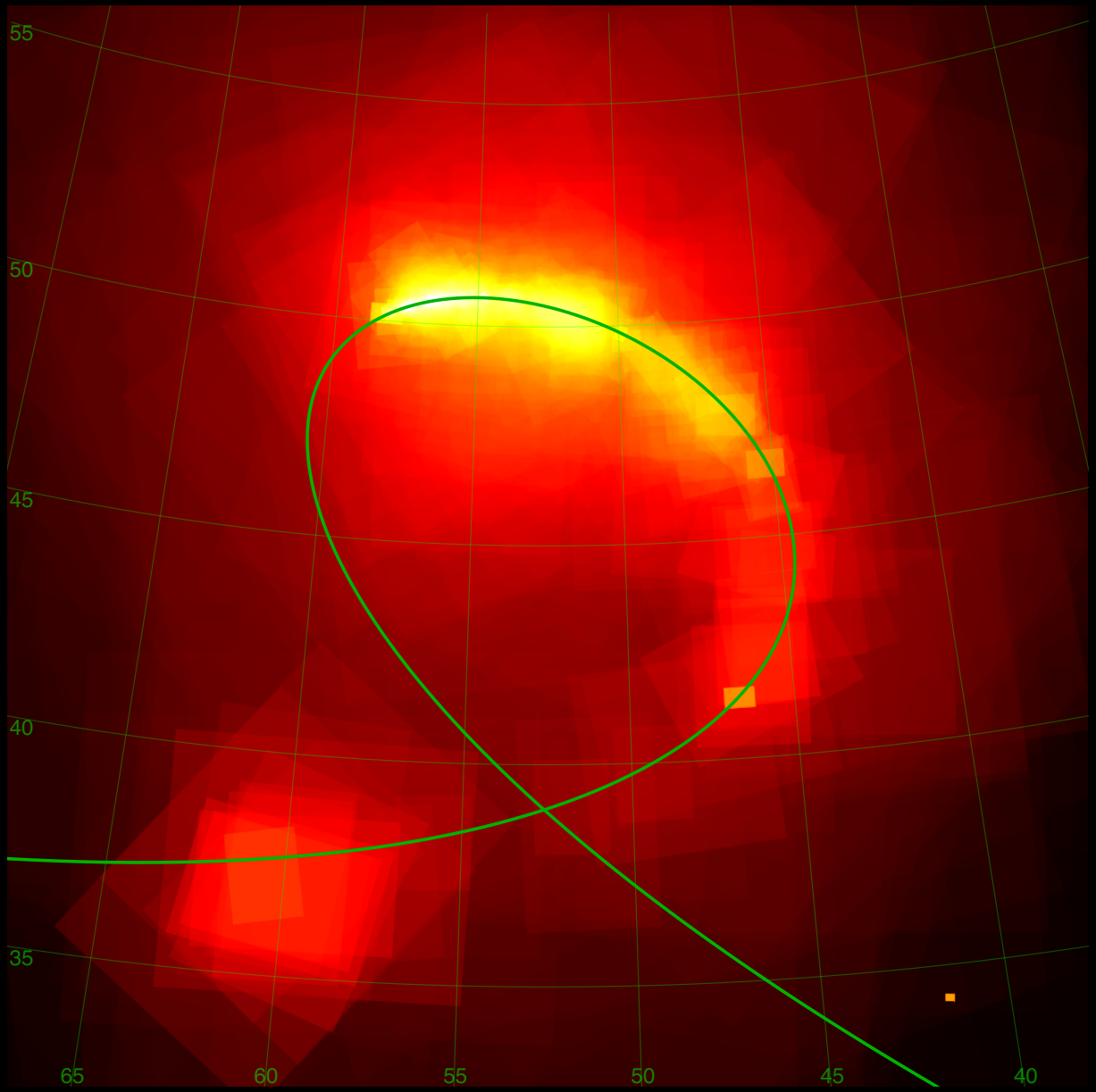
*(v)


*(w)

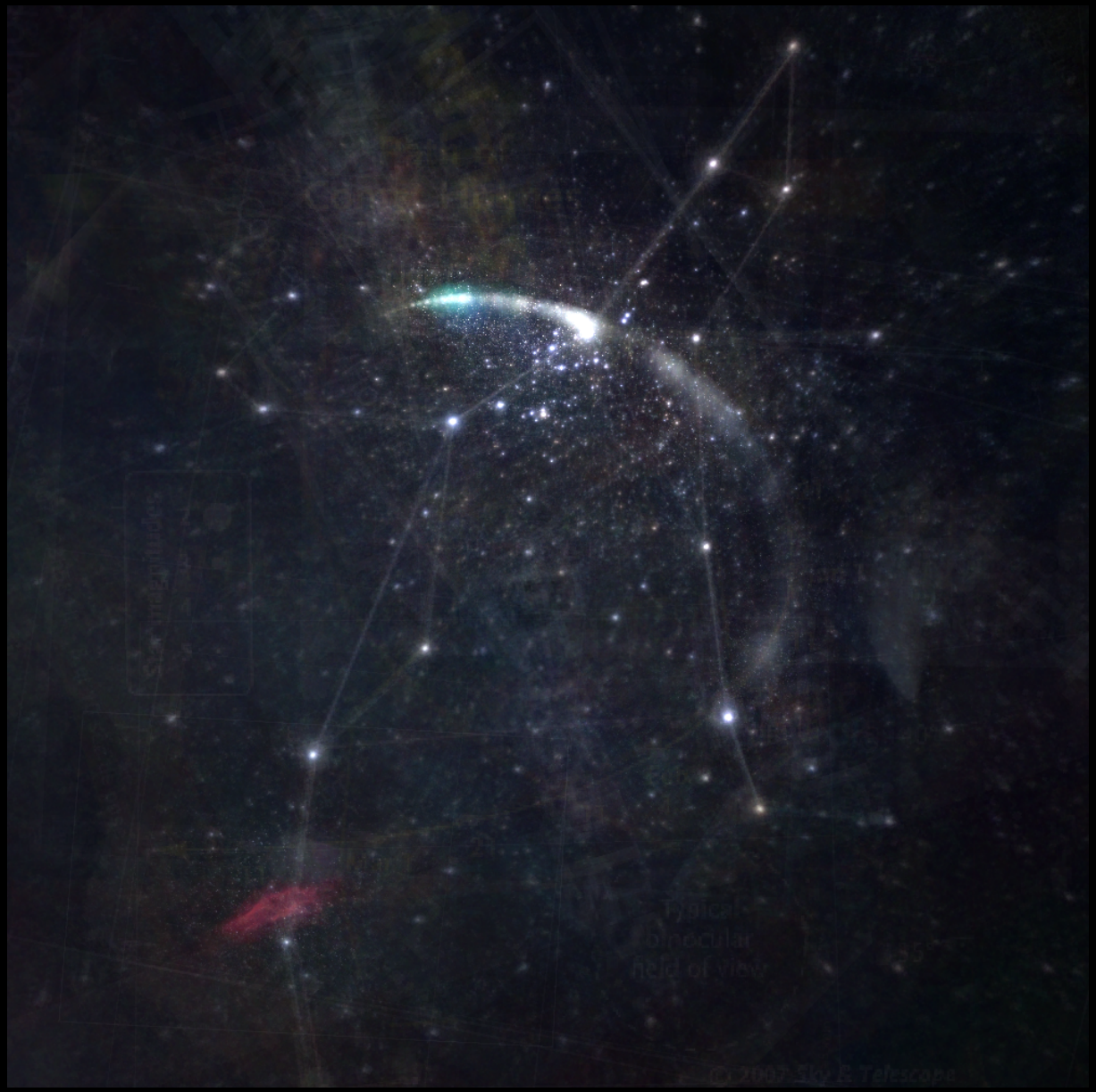
Comet Holmes



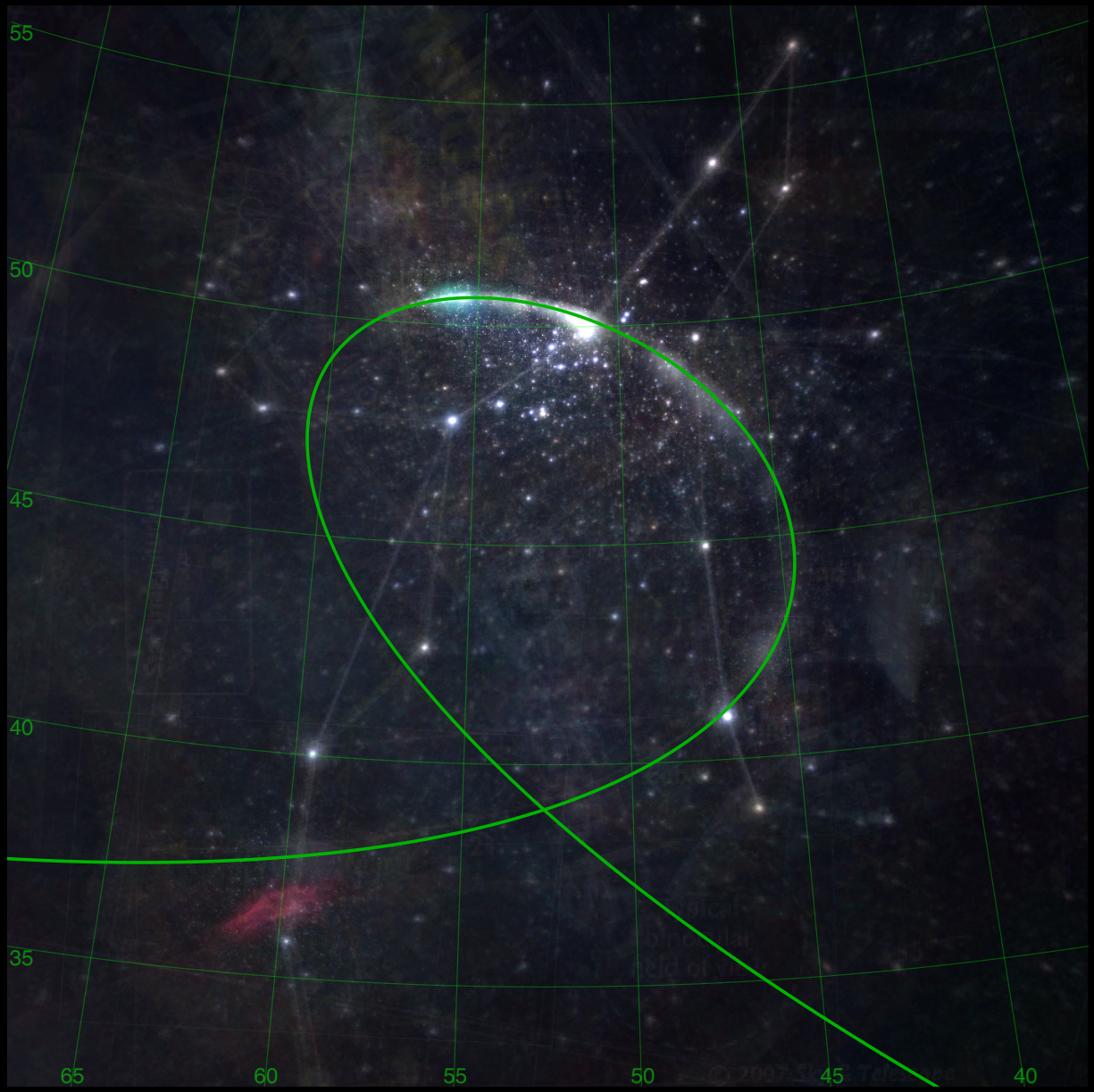
Comet Holmes



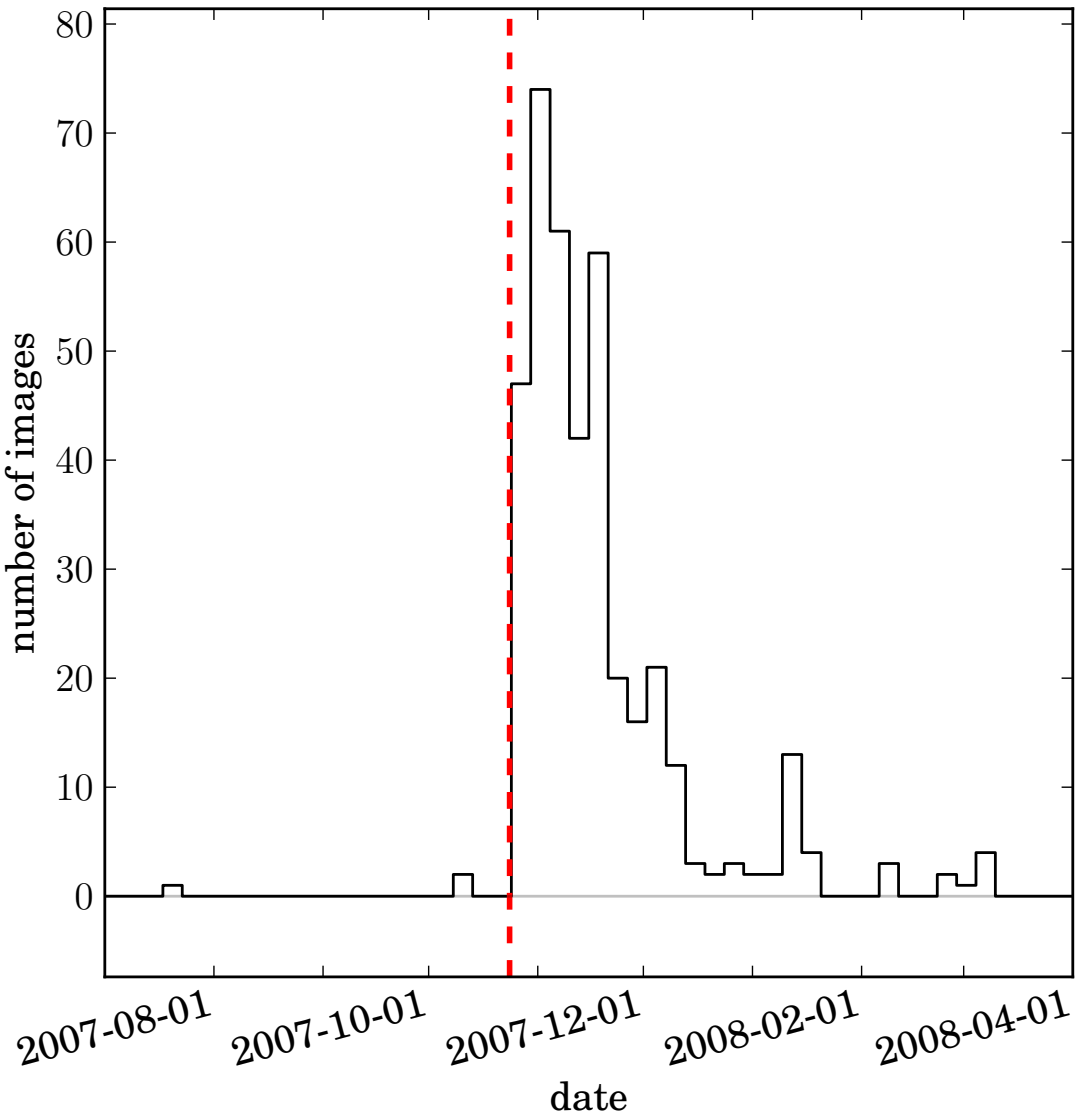
Comet Holmes



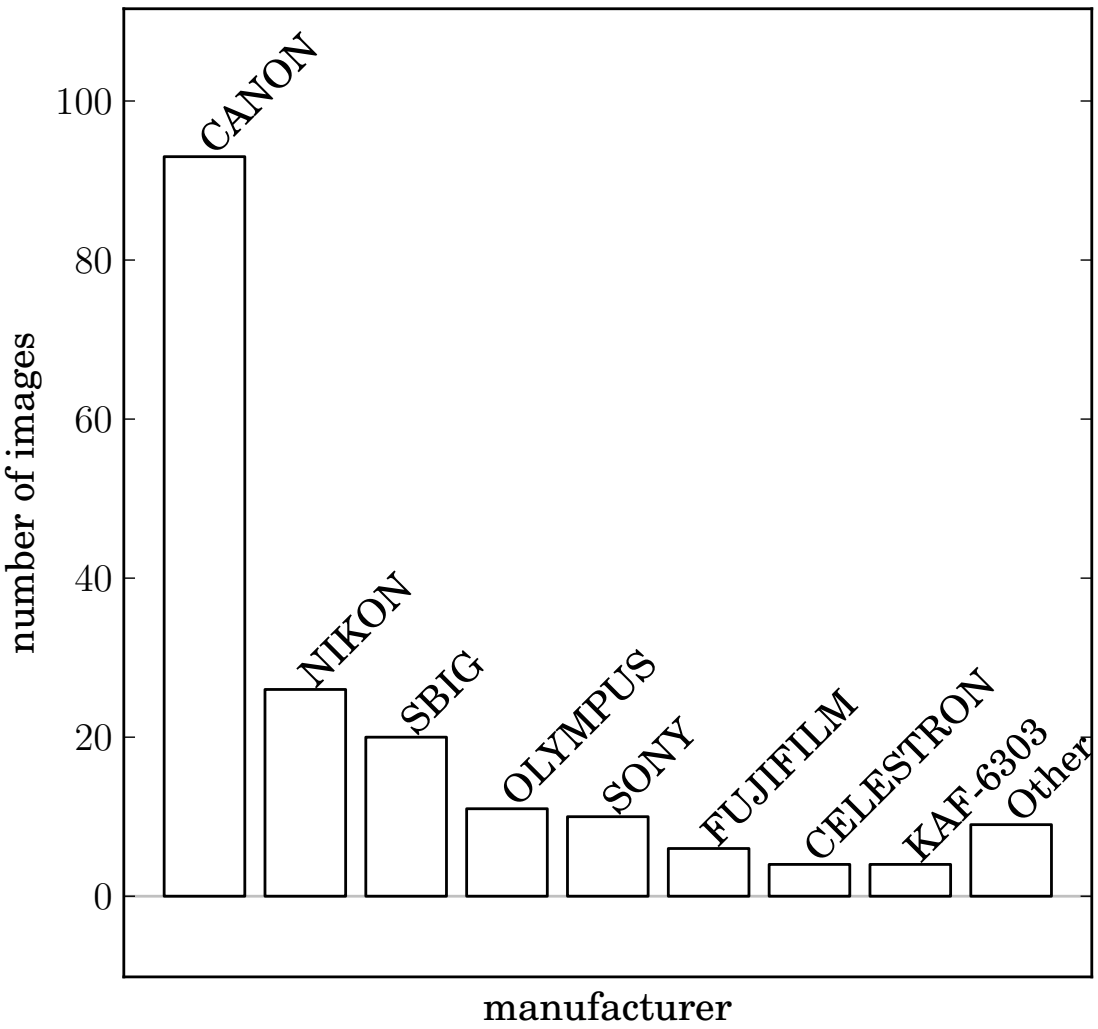
Comet Holmes



Comet Holmes



Comet Holmes



Comet Holmes: the model

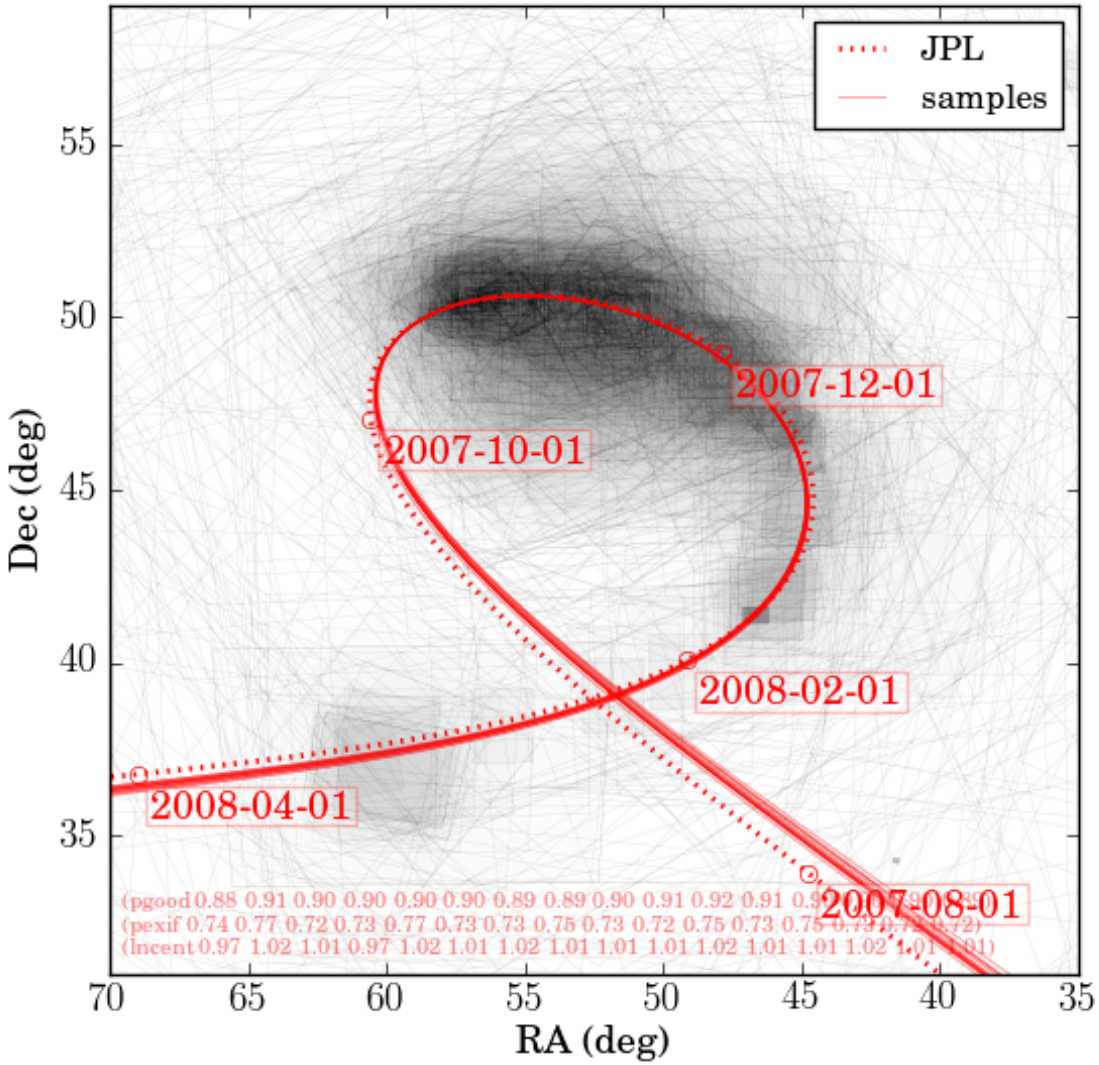
$$
\begin{aligned}
p(\boldsymbol{\alpha}_i|\boldsymbol{\Omega}_i,\boldsymbol{\omega},\boldsymbol{\theta}) &= \int p(\boldsymbol{\alpha}_i|t_i,\boldsymbol{\Omega}_i,\boldsymbol{\omega},\boldsymbol{\theta})\, p(t_i|\boldsymbol{\Omega}_i,\boldsymbol{\theta})\, dt_i \\
p(\boldsymbol{\alpha}_i|t_i,\boldsymbol{\Omega}_i,\boldsymbol{\omega},\boldsymbol{\theta}) &= p_{\rm good}\, p_{\rm fg}(\boldsymbol{\alpha}_i|t_i,\boldsymbol{\Omega}_i,\boldsymbol{\omega},\boldsymbol{\theta}) + [1-p_{\rm good}]\, p_{\rm bg}(\boldsymbol{\alpha}_i) \\
p_{\rm fg}(\boldsymbol{\alpha}_i|t_i,\boldsymbol{\Omega}_i,\boldsymbol{\omega},\boldsymbol{\theta}) &= \begin{cases} [\eta\,\Omega_i]^{-1} & \text{comet in } \eta \text{ sub-image} \\ 0 & \text{comet not in } \eta \text{ sub-image} \end{cases} \\
p_{\rm bg}(\boldsymbol{\alpha}_i) &= [4\pi]^{-1} \quad ; \qquad (1)
\end{aligned}
$$

$$
p(t_i|\boldsymbol{\Omega}_i,\boldsymbol{\theta}) = \begin{cases} p_{\rm emp}(t_i) & \text{if no } t_{\rm EXIF} \\ p_{\rm EXIF}\, p(t_i|t_{\rm EXIF}) + [1-p_{\rm EXIF}]\, p_{\rm emp}(t_i) & \text{if } t_{\rm EXIF} \text{ in } \end{cases}
$$

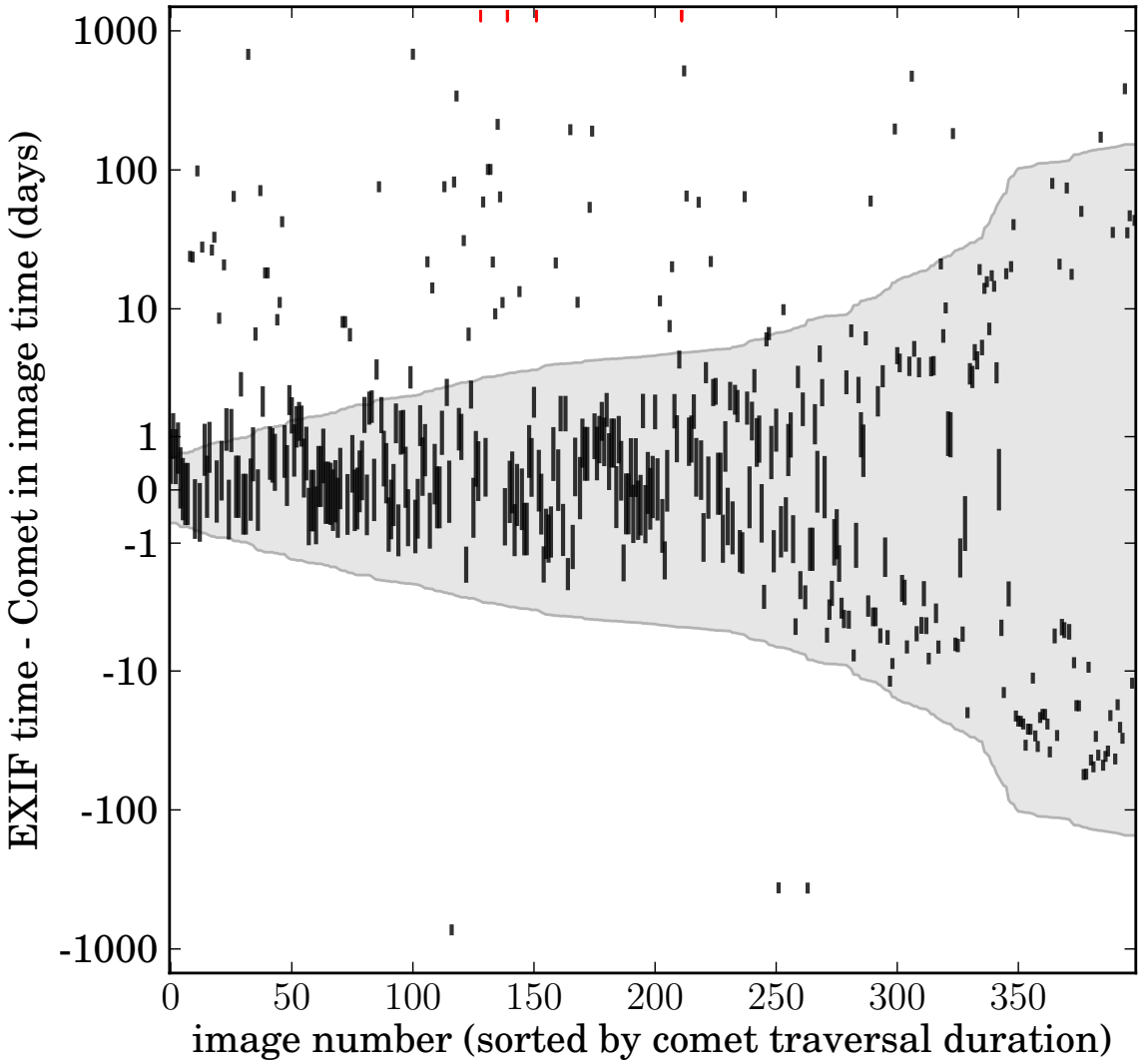
Comet Holmes: the model

- It is a model of the way *people* point their cameras.
  - We don't trust the meta-data.
  - Meta-data reconstruction often requires a model of the meta-data provider.
  - See also *GalaxyZoo*.
- It requires *informative priors*.
  - That doesn't mean we have to make strong assumptions.
- This is Citizen Science with unwitting participants.

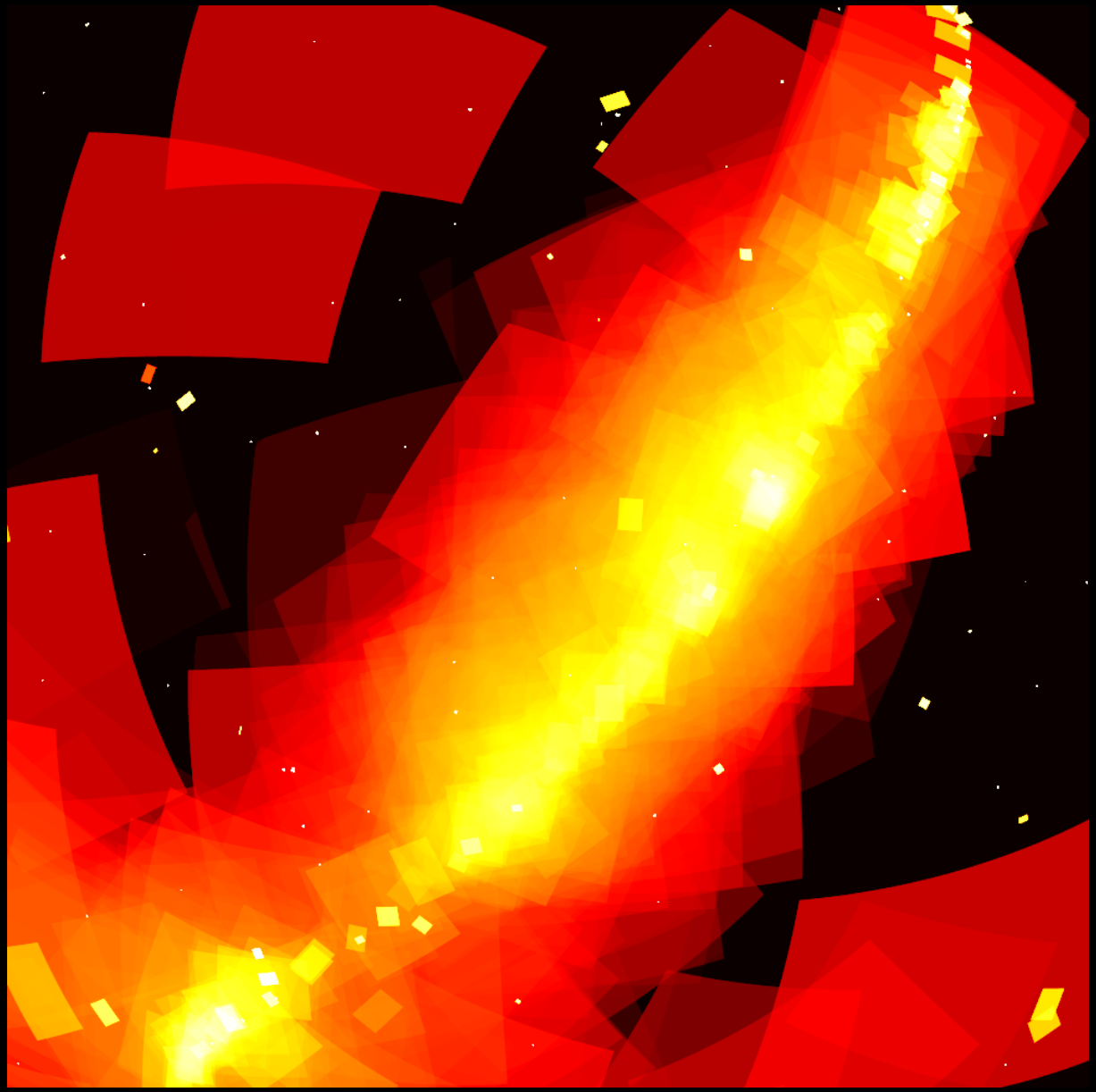
Comet Holmes: results



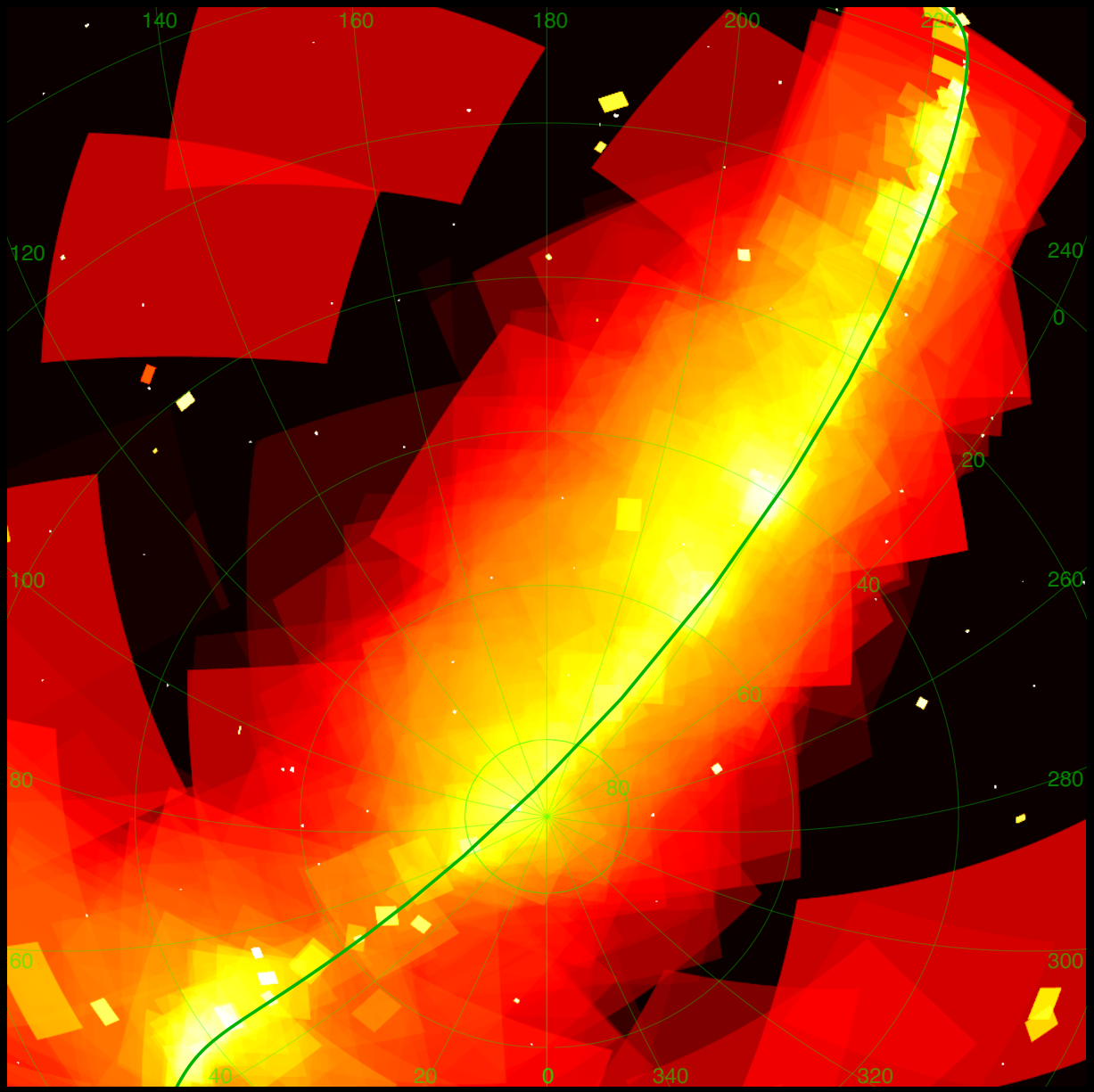
Comet Holmes: results



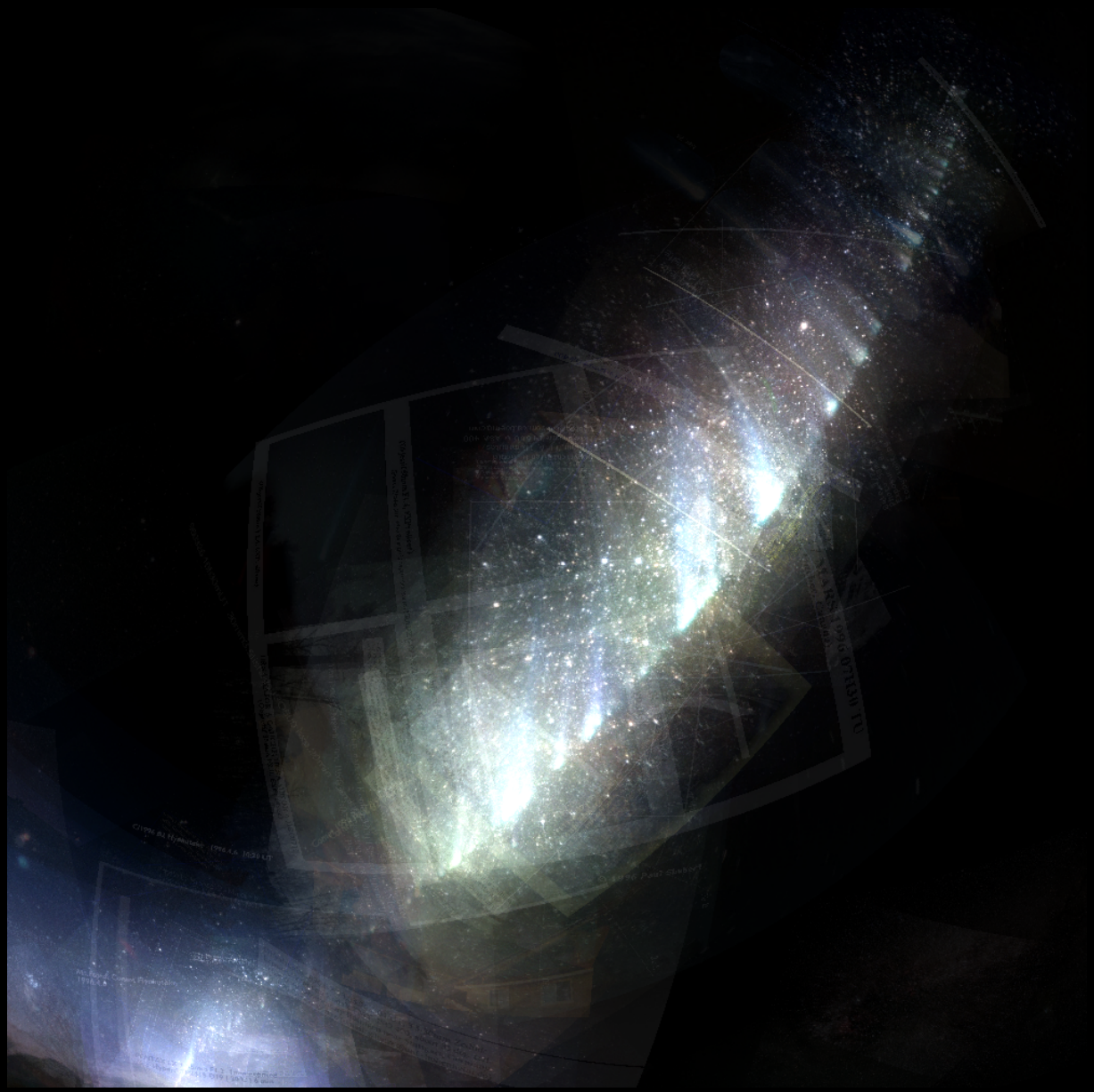
Comet Hyakutake



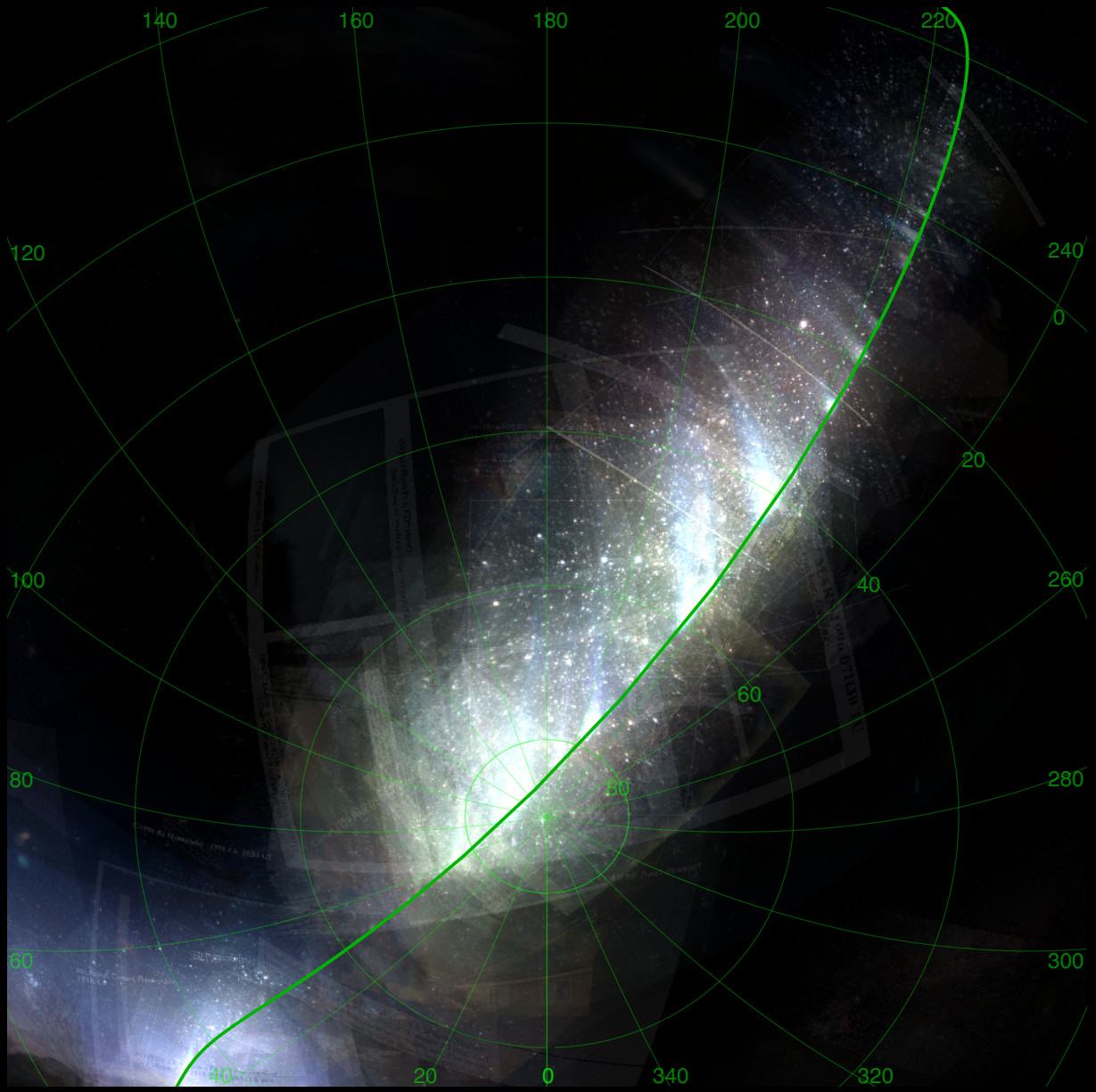
Comet Hyakutake



Comet Hyakutake



Comet Hyakutake



Quasar target selection

Quasar target selection: 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.
  - we knew in advance $> 10^4$ quasars
  - *but* those quasars are *brighter* and thus *higher signal-to-noise*

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

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!*

Quasar target selection (1011.6392): method

- classification by density modeling
- extreme deconvolution (TM):
  - each data point samples the true density (in color space), *convolved* with that data point's own unique uncertainty profile
  - model all this with mixtures of Gaussians for performance
- likelihood ratios (star *vs.* galaxy) become density ratios in the convolved model

What's wrong with typical density estimation methods?

- They estimate the density of the *observed* data.
- If two data sets are taken by different instruments or with different noise properties, density estimated by one will not be relevant to the other.
- In astronomy, we go to great lengths to understand our noise models and uncertainty variances; a good method will use them.

extreme deconvolution: before

- standard E-M density estimation using a mixture of Gaussians:
  - find a_k, m_k, V_k to maximize probability of data

$$p(D|\{a_k, m_k, V_k\}) = \prod_i p(D_i|\{a_k, m_k, V_k\}) \quad (3)$$

$$p(D_i|\{a_k, m_k, V_k\}) = \sum_k a_k \, N(D_i|m_k, V_k) \quad (4)$$

$$1 = \sum_k a_k \quad (5)$$

extreme deconvolution: after

- extreme deconvolution:
  - find a_k, m_k, V_k to maximize probability of data

$$p(D|\{a_k, m_k, V_k\}) = \prod_i p(D_i|\{a_k, m_k, V_k\}) \quad (6)$$

$$p(D_i|\{a_k, m_k, V_k\}) = \sum_k a_k \, N(D_i|m_k, V_k + S_i) \quad (7)$$

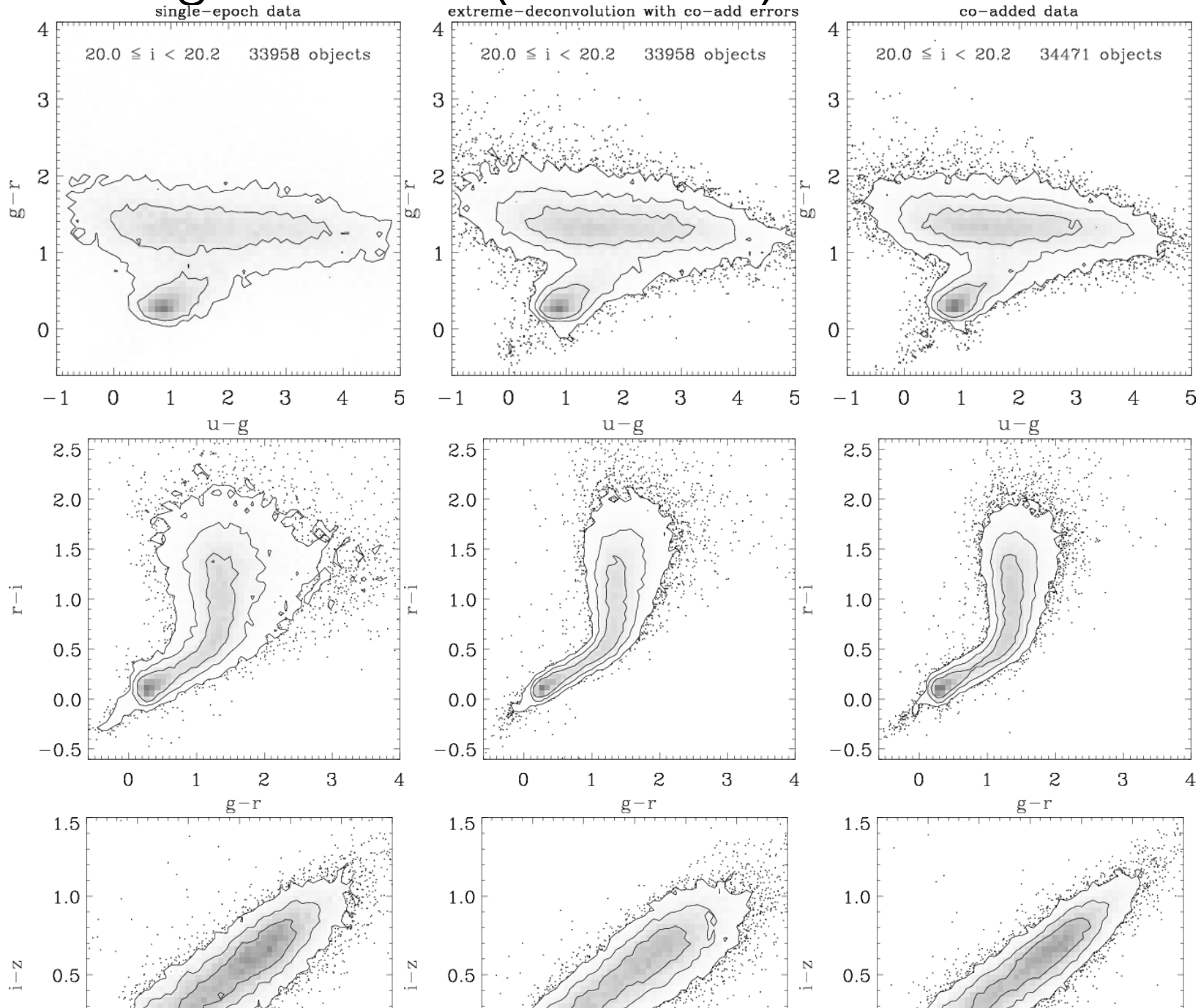
$$1 = \sum_k a_k \quad (8)$$

  - note Gaussian noise model—this can be generalized
  - also can deal with arbitrary missing data
  - *but* optimization is much *slower* and *more complicated*
  - Bovy et al, http://arxiv.org/abs/0905.2979 and *AOAS*

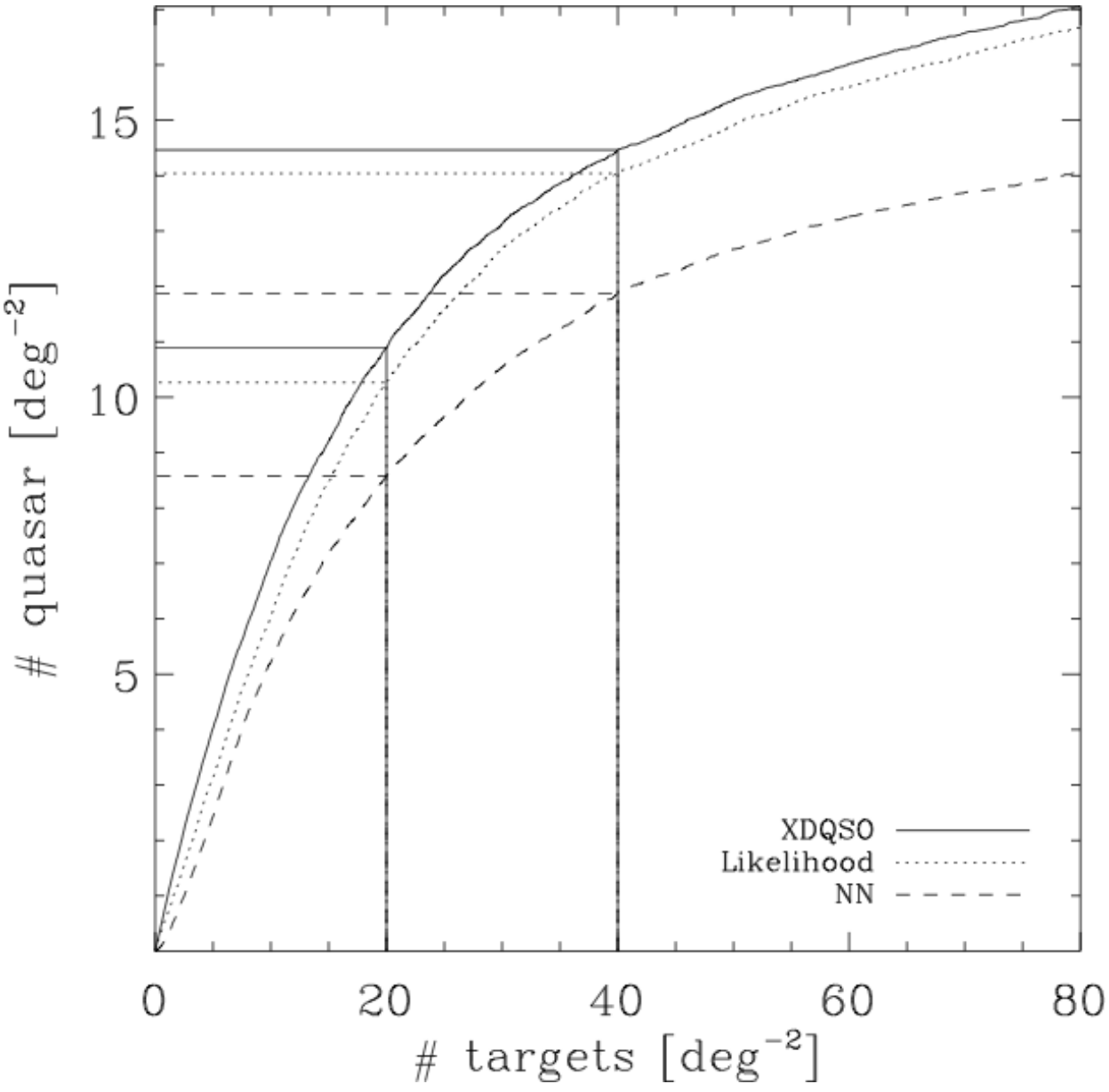
Polemic: missing data

- Many 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 pryobabilistic model $p(D|M, I)$, you automatically deal with missing data.

Quasar target selection (1011.6392): results



Quasar target selection (1011.6392): results



Quasar target selection (1011.6392): why do we perform well?

- 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 are extensible to new prior information or other data.
  - *GALEX*
  - *UKIDSS*
  - variability
- Bovy
- extreme-deconvolution (at code.google.com)
- *SDSS-III BOSS* core target selection

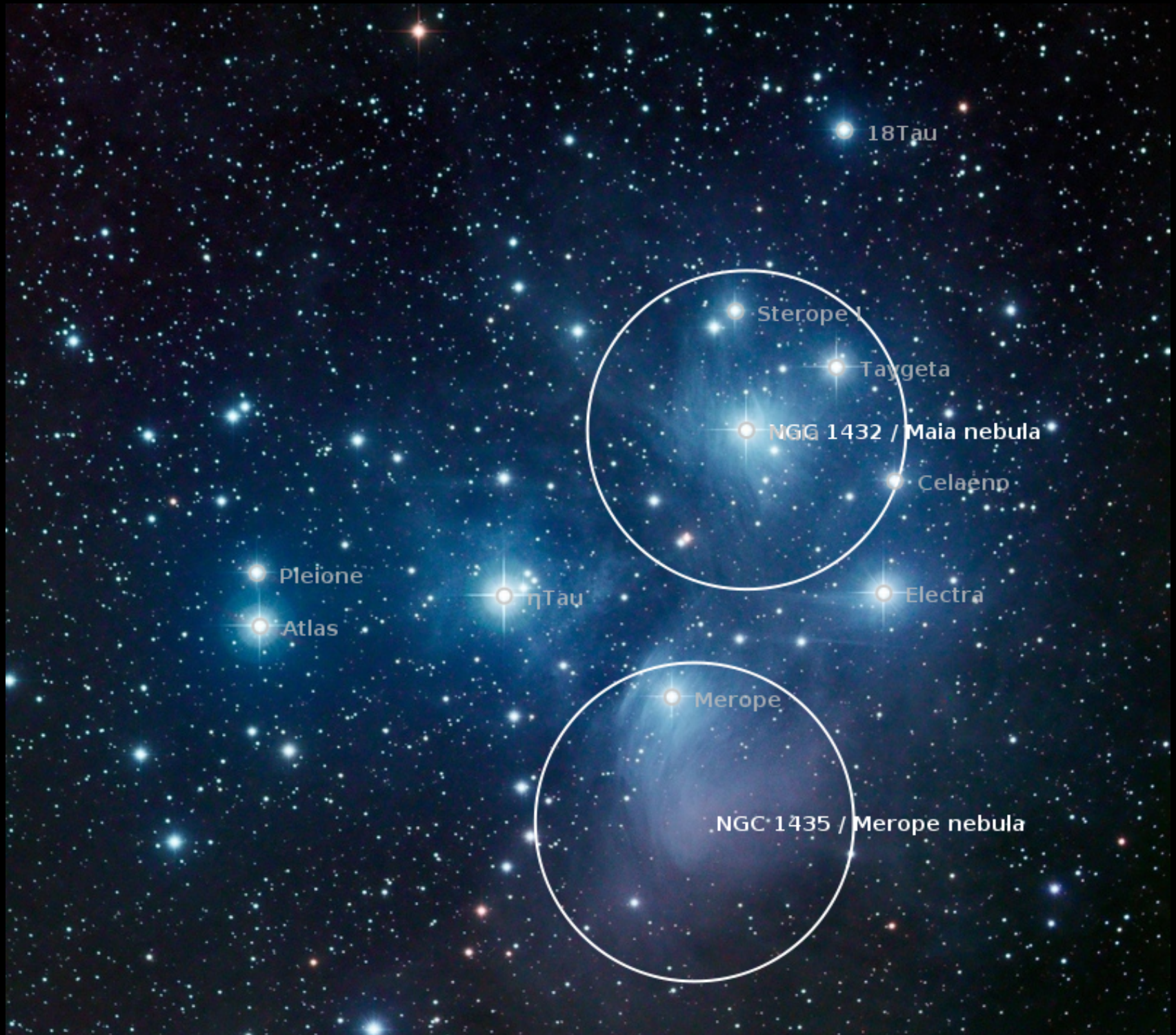
Automated astronomical image recognition

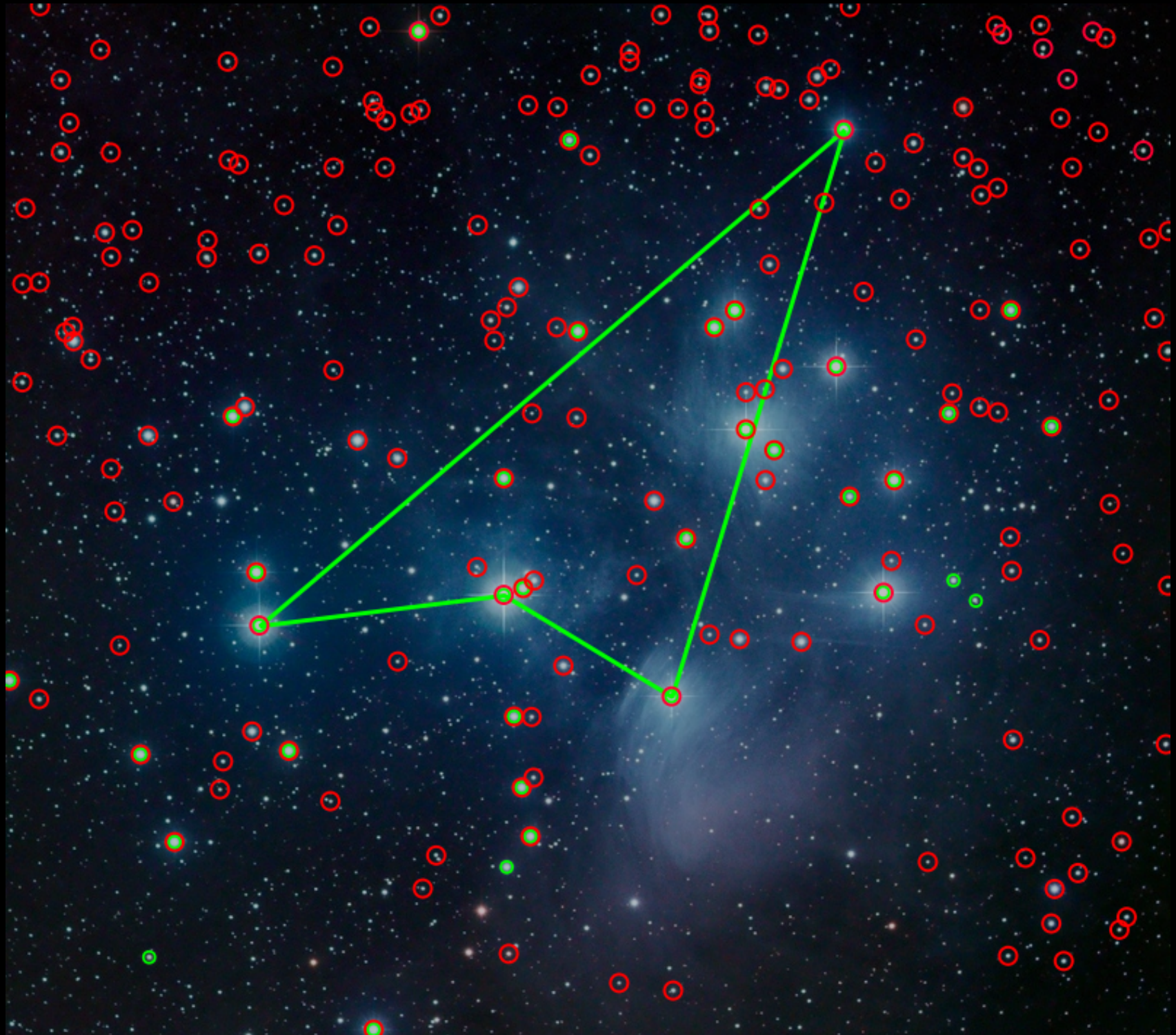
Astrometry.net (Lang *et al.* 0910.2233)

- Non-text search:
  - Here is an image, what is this an image of?
  - In the process of answering this, we also vet and calibrate it.
- Calibration:
  - Produce standards-compliant world-coordinate systems for images of unknown provenance.
  - Repair damaged or wrong image headers.
  - Provide astrotagging services.

Astrometry.net (Lang *et al.* 0910.2233)

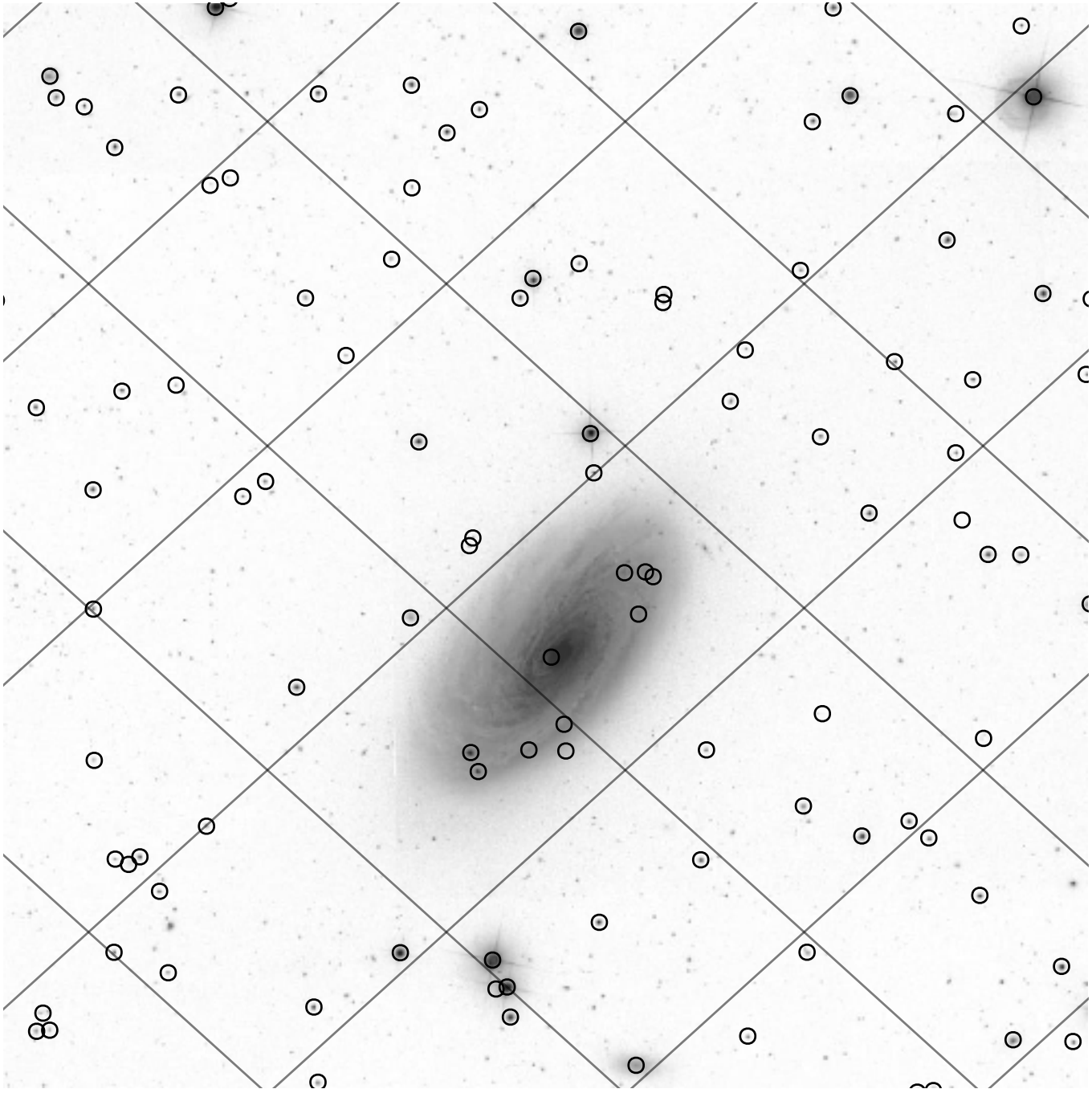
- In *flickr*: few 10^4 submissions.
- On the web: tens of thousands; in projects: millions.
- Source detection, geometric hashing, Bayesian decision theory.
- A probabilistic model of how detected stars are distributed within images!
  - mixture model or foreground–background model
- Make decisions that opimize our *long-term discounted free cash flow*.
  - requires utility specification
  - requires customer model

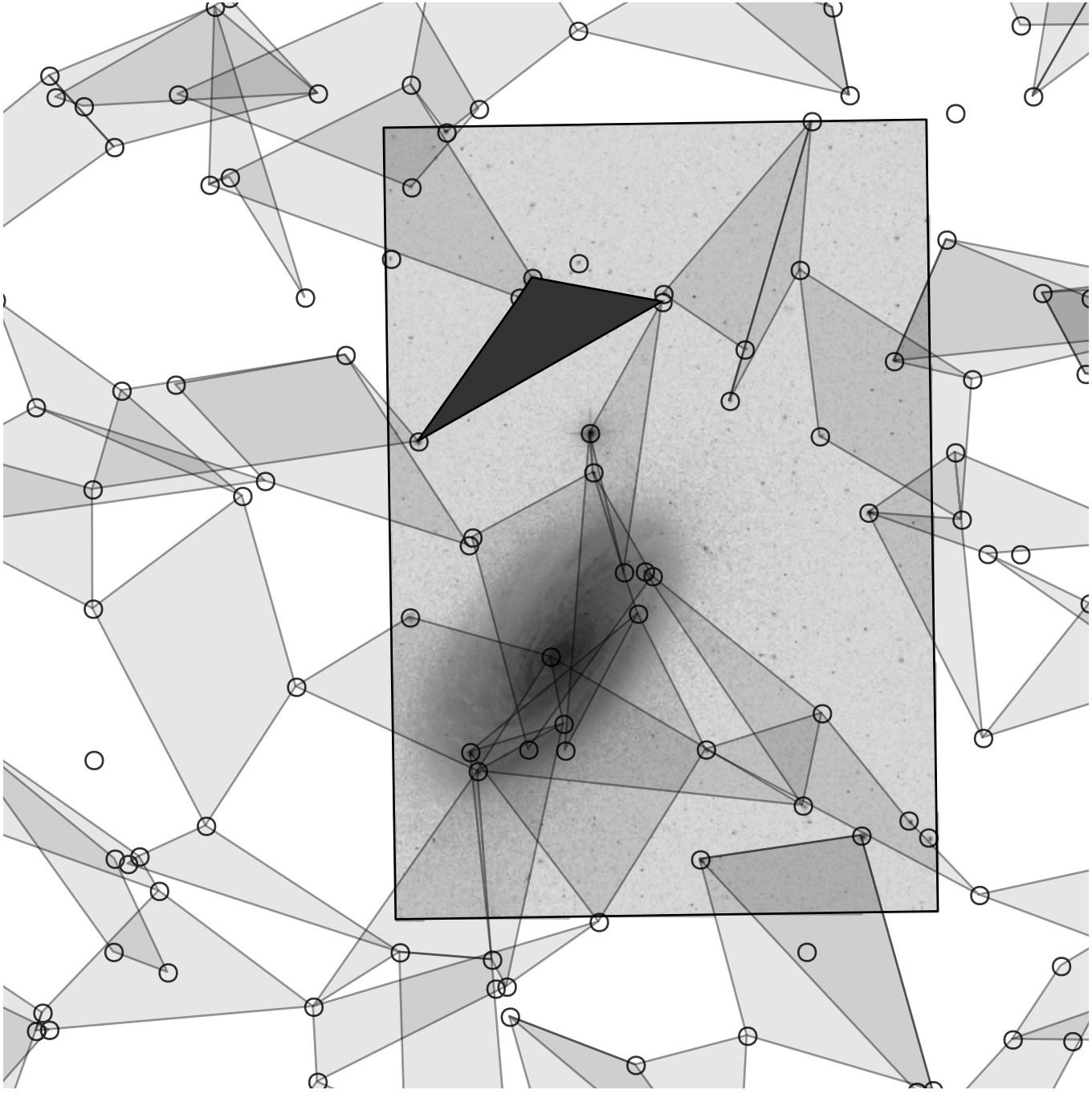

18Tau
Sterope I
Taygeta
NGC 1432 / Maia nebula
Celaeno
Pleione
ηTau
Electra
Atlas
Merope
NGC 1435 / Merope nebula

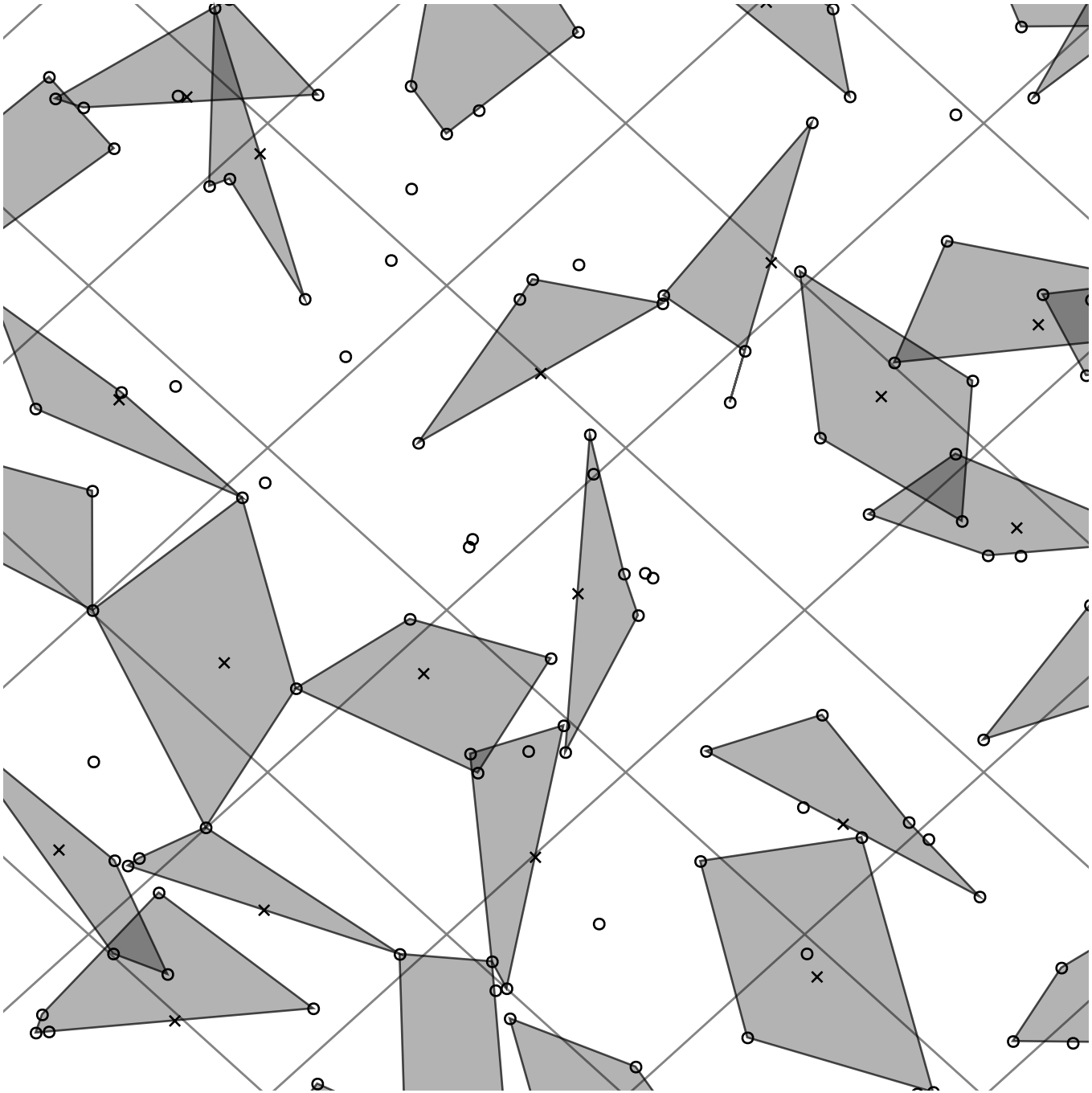


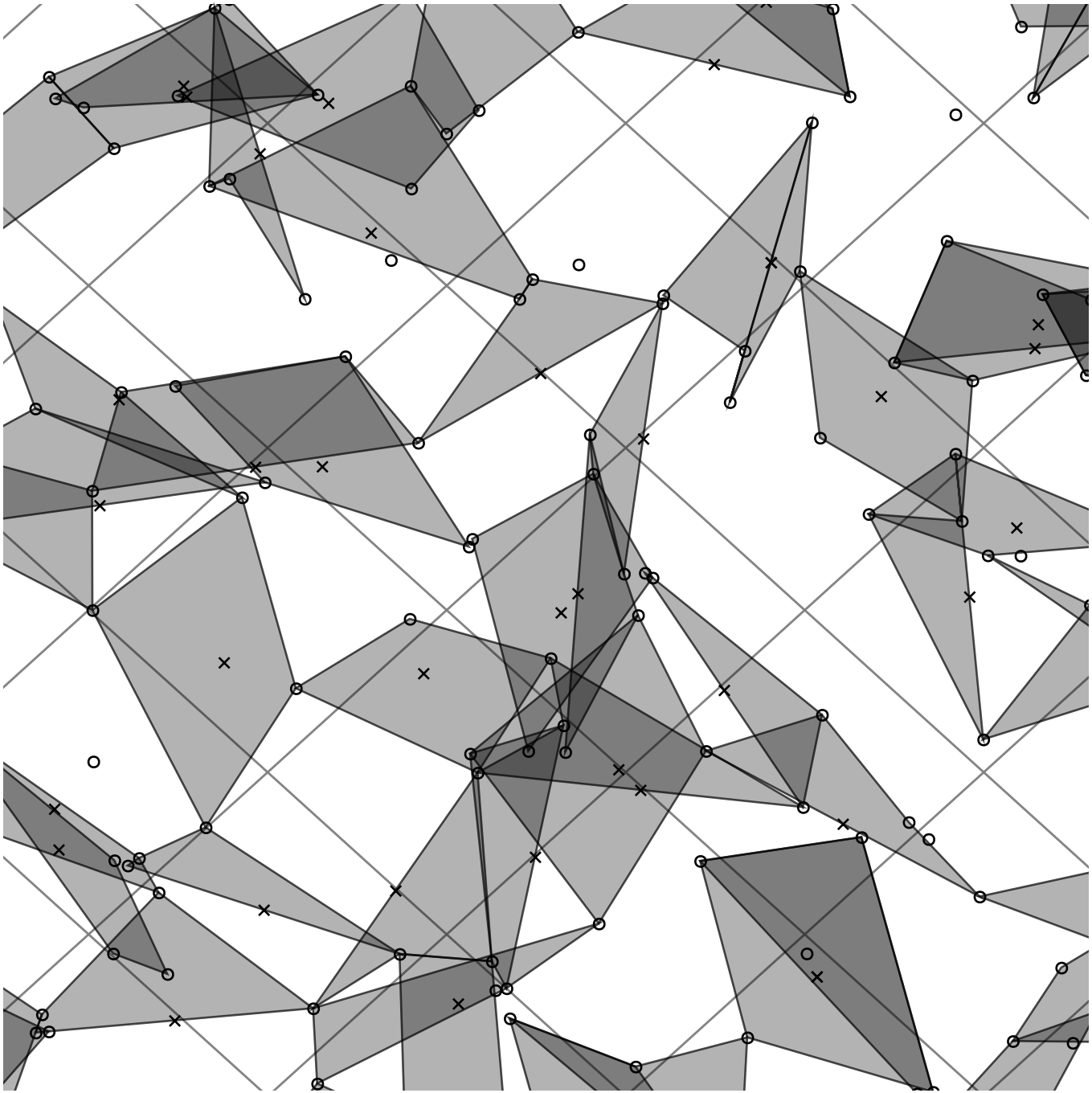
Astrometry.net (Lang *et al.* 0910.2233): how it works

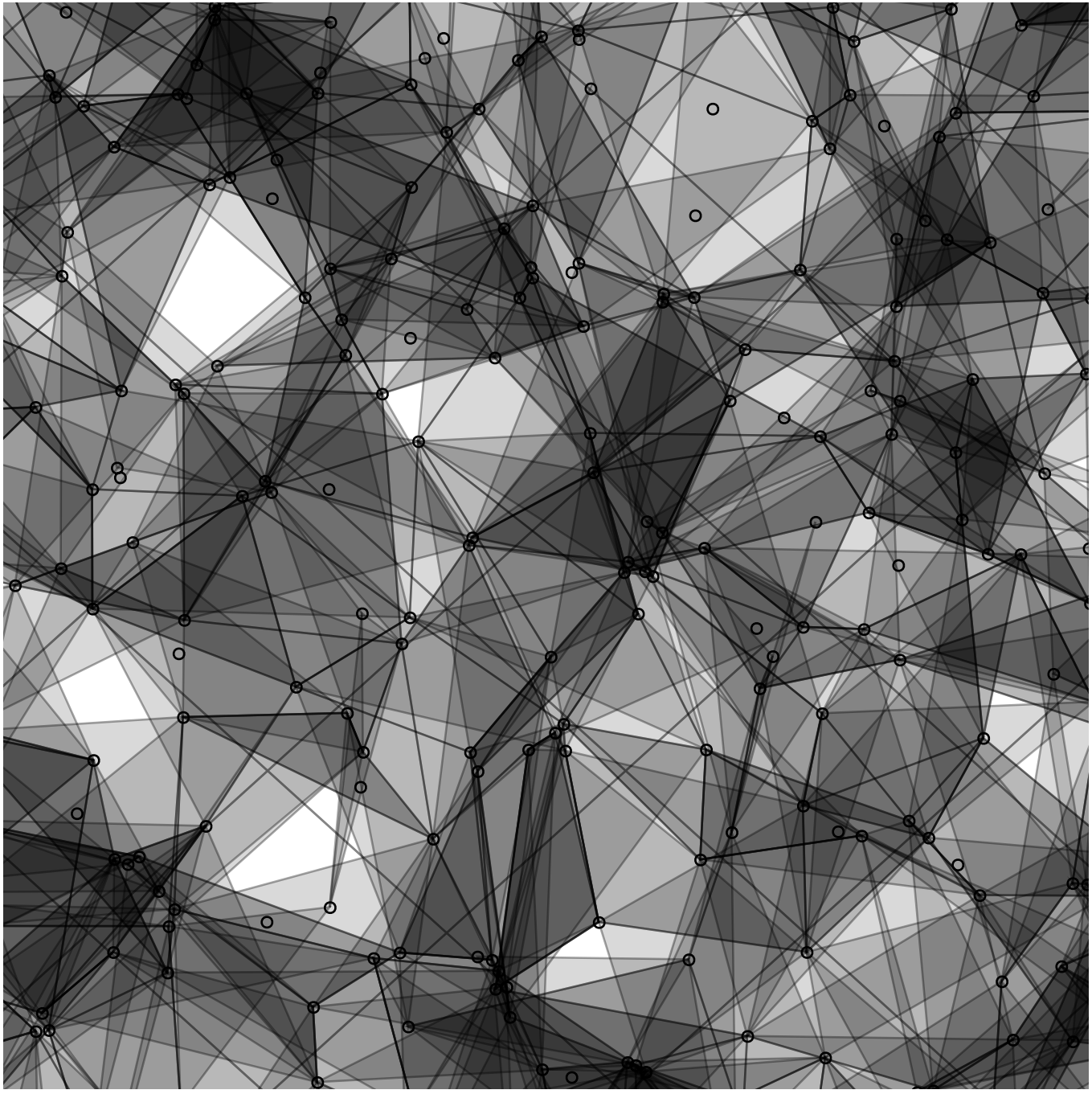
- identify stars in images (badly)
- match indexed 4-star or 5-star figures
- that implies a pointing, orientation, and scale
- how well do we predict the *other* stars in the image?
  - foreground-background model (because star identification is bad)
  - posterior probability that this match hit by chance
- utility model to decide whether to "accept" the match

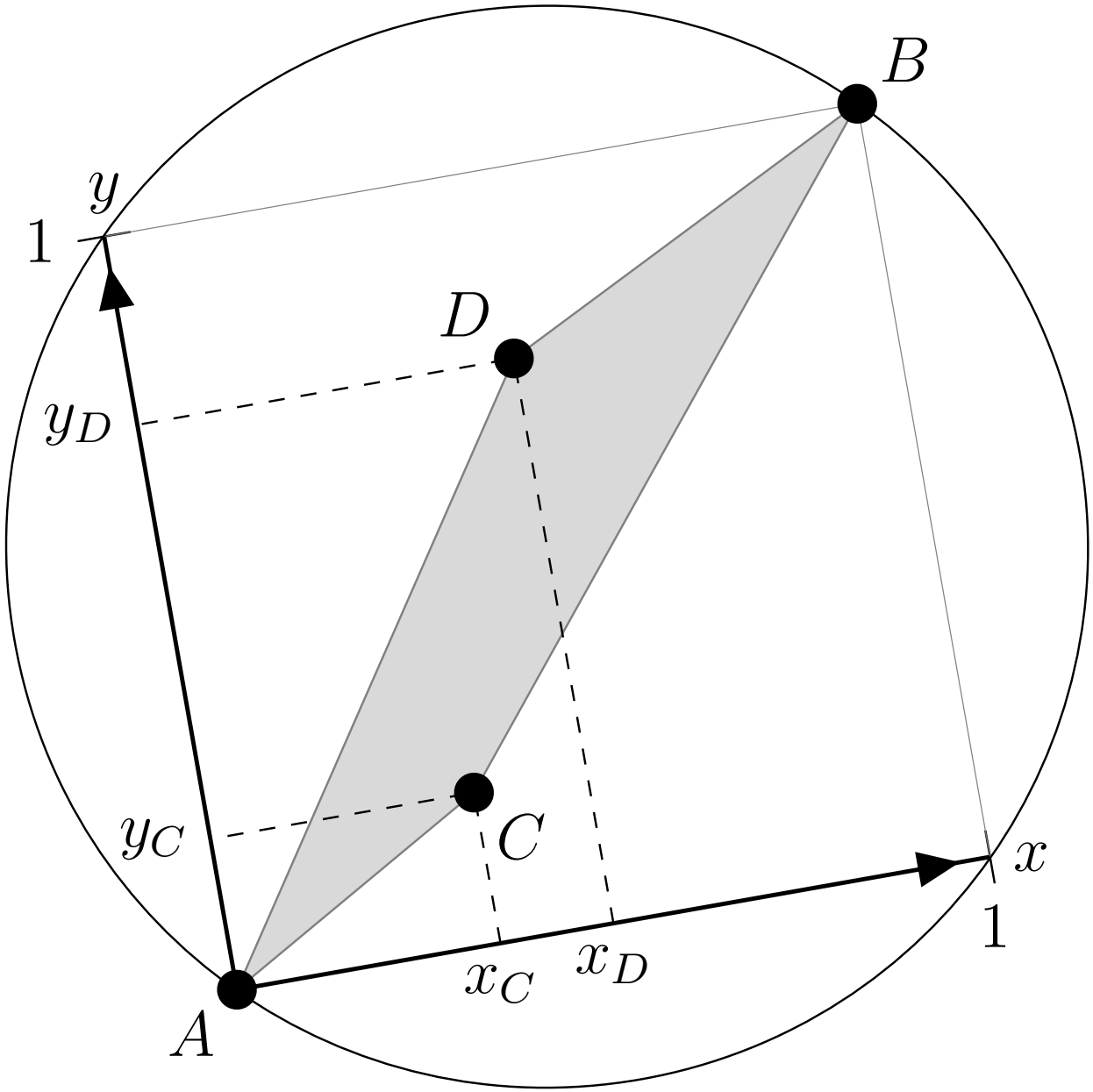








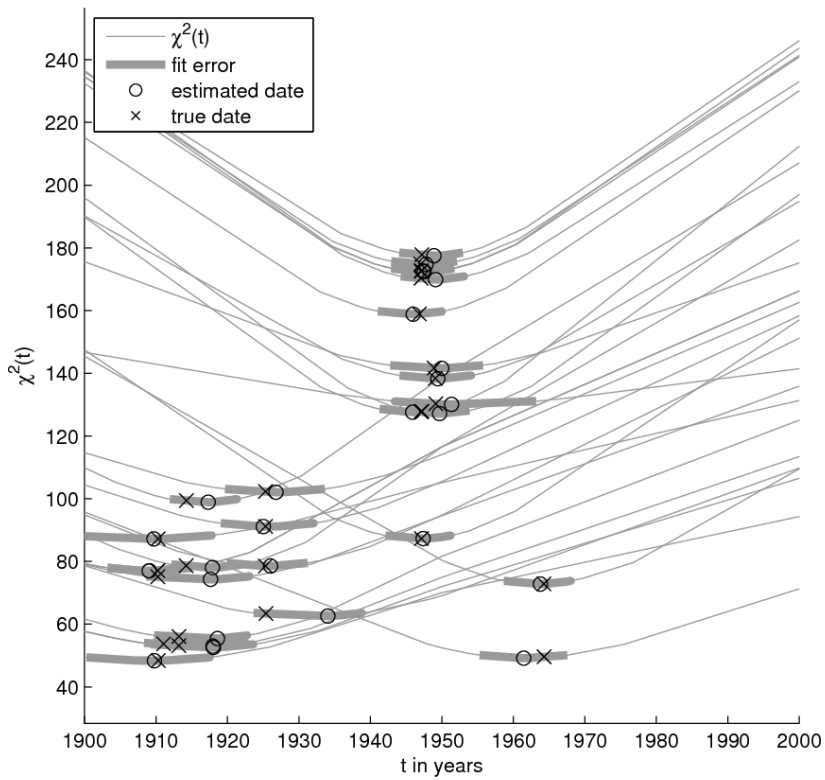


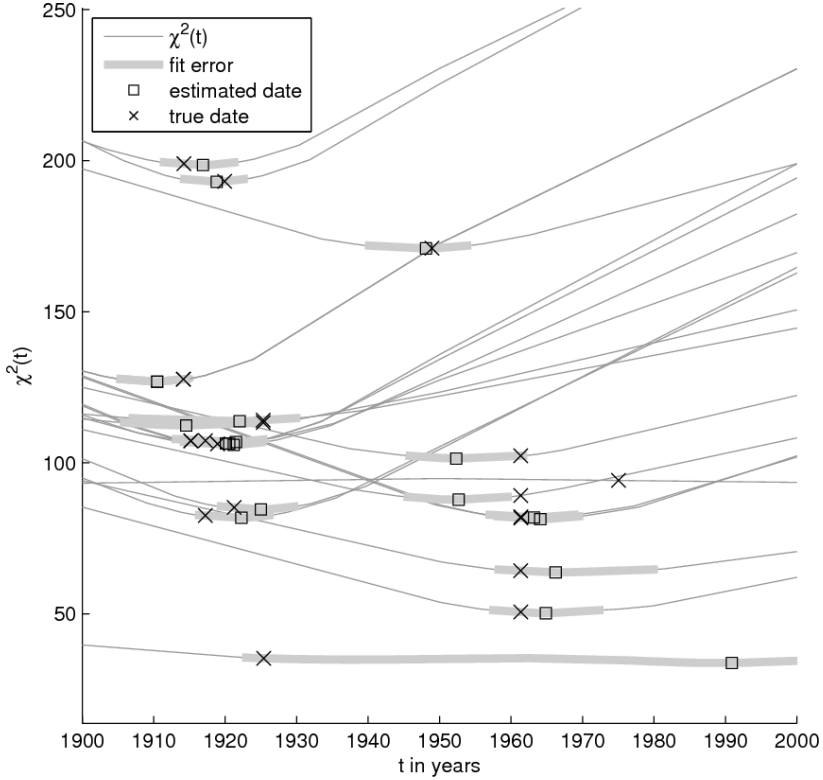

B
y
1
D
y_D
y_C
C
x
1
x_C
x_D
A

Astrometry.net (Lang *et al.* 0910.2233): performance

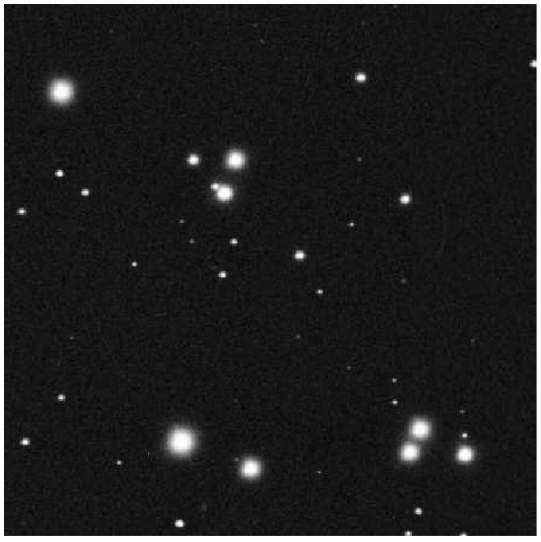
- essentially no false positives (exceptions insane)
- $>$ 99.9 percent success rate on *SDSS* and *GALEX* imaging
- large numbers of users (amateurs, professionals, educators, robots)
- can also do other kinds of calibration:
  - wavelength bandpass
  - photometric sensitivity
  - point-spread function
  - date (to within years 2008 AJ 136 1490)
- don't believe? Try it yourself at http://nova.astrometry.net/

Science with untrusted data: Blind Date

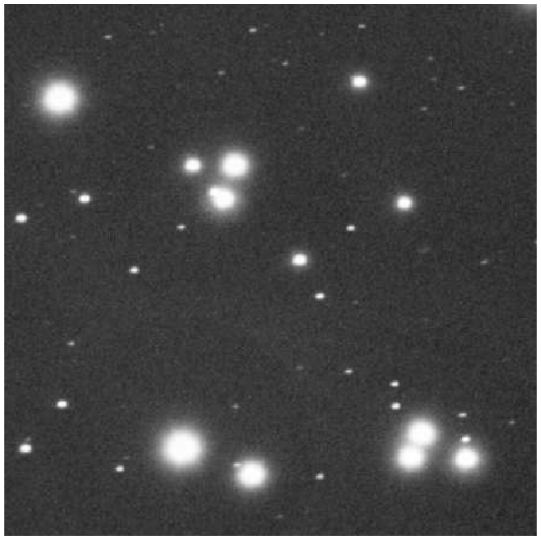




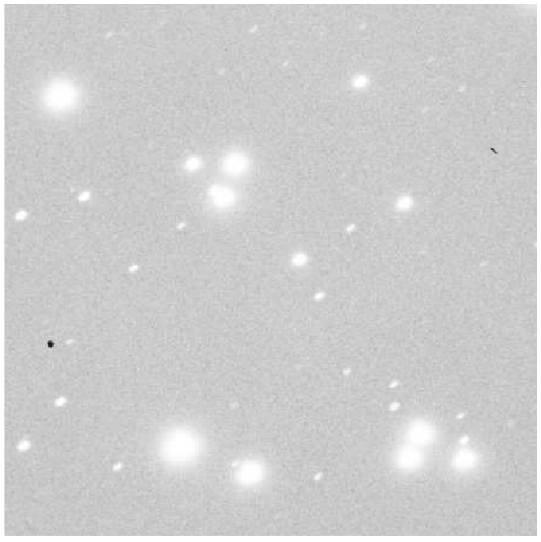
Science with untrusted data: Blind Date



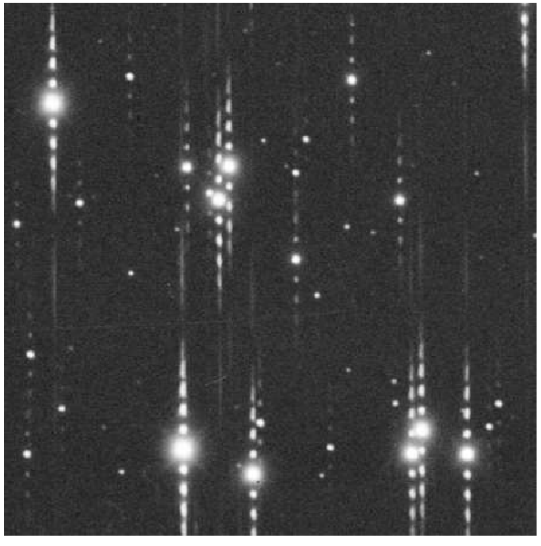
(a) Mar-1914 / Nov-1917



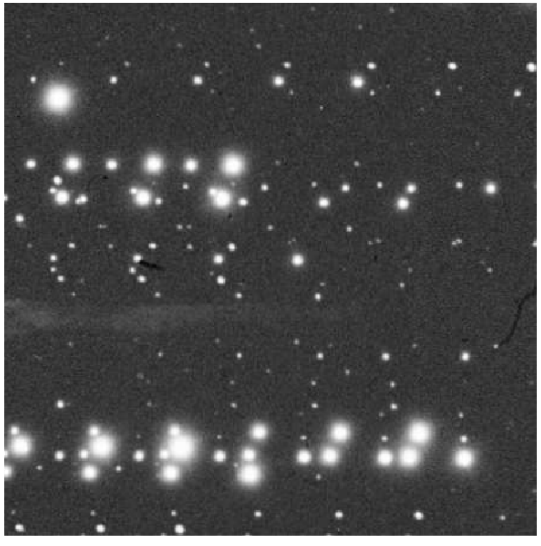
(b) Mar-1947 / Nov-1945



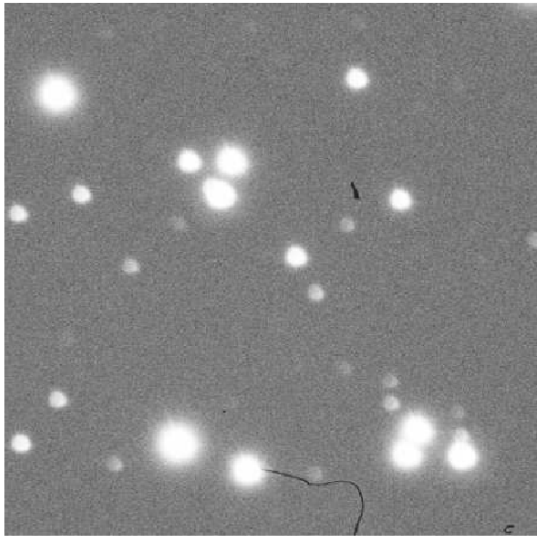
(c) Feb-1949 / May-1951



(d) Feb-1914 / Jun-1910

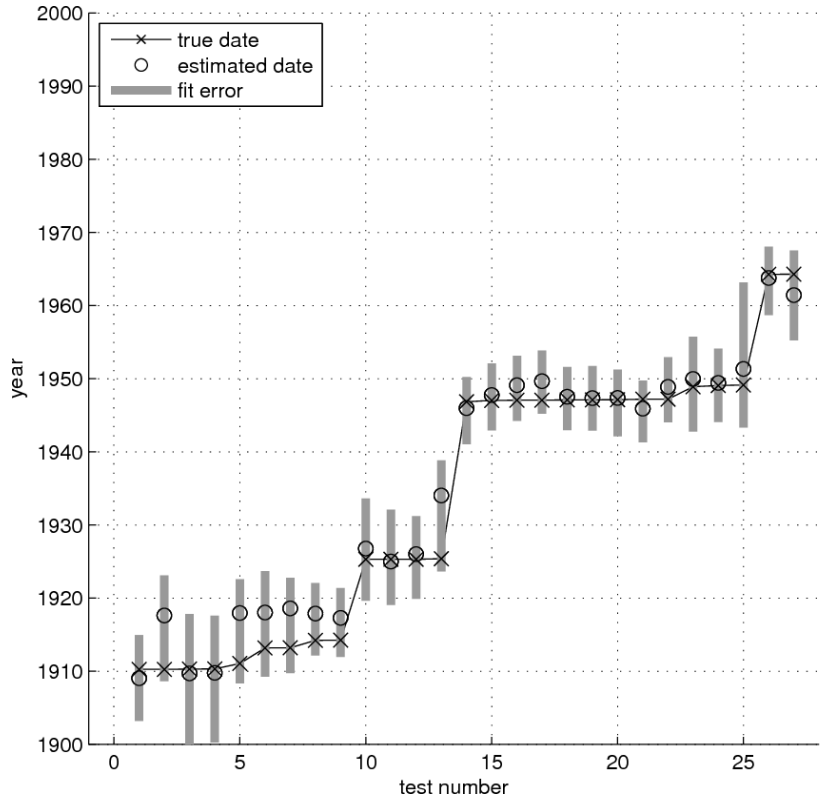


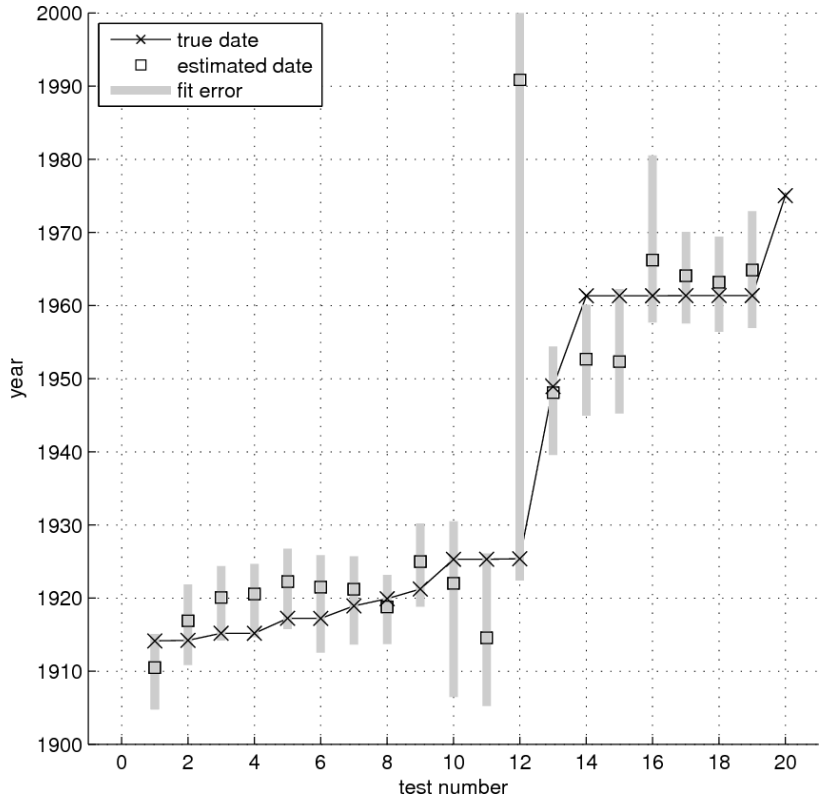
(e) Mar-1914 / Nov-1916



(f) Jan-1975 / Jan-1800

Science with untrusted data: Blind Date





Barron, Hogg, Lang, Roweis, 2008, *AJ* **136** 1490

trusting data

- data that have not been used to *do science* are *wrong data*
- a practical point, not a theoretical point (see *SDSS*)
- science is the *ultimate functional testing environment*
- thinking about a "trust model" for astronomy
  - we work with data from observatories, amateurs, and archives
  - data often have unknown provenance, wrong clocks, *etc.*
  - we need to calibrate, vet, verify—automatically

the theory of everything

- simultaneous modeling of all astronomical imaging (Hogg & Lang, 0810.3851)
- run in real time as an update system
- model parameters:
  - position and brightness of every star
  - pointing, orientation, bandpass, PSF, calibration of every image
  - camera parameters for every telescope + camera
- report "novel information content" about incoming imaging
  - could even direct new observations automatically
- basis of the *Open-Source Sky Survey* (TM but vapor-ware)
- calibration and vetting is produced naturally

Orion astrometry

Orion astrometry

Orion astrometry

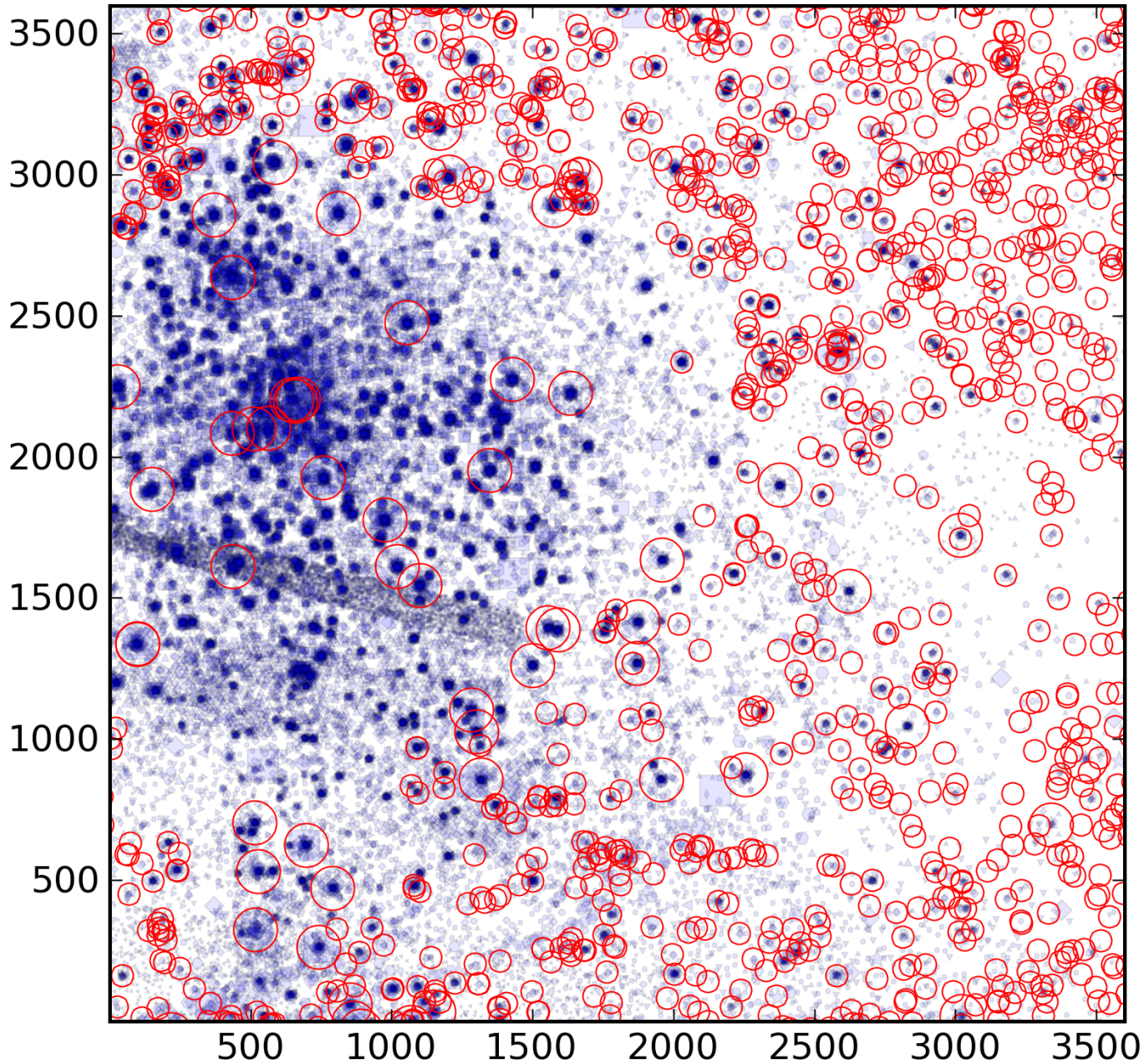
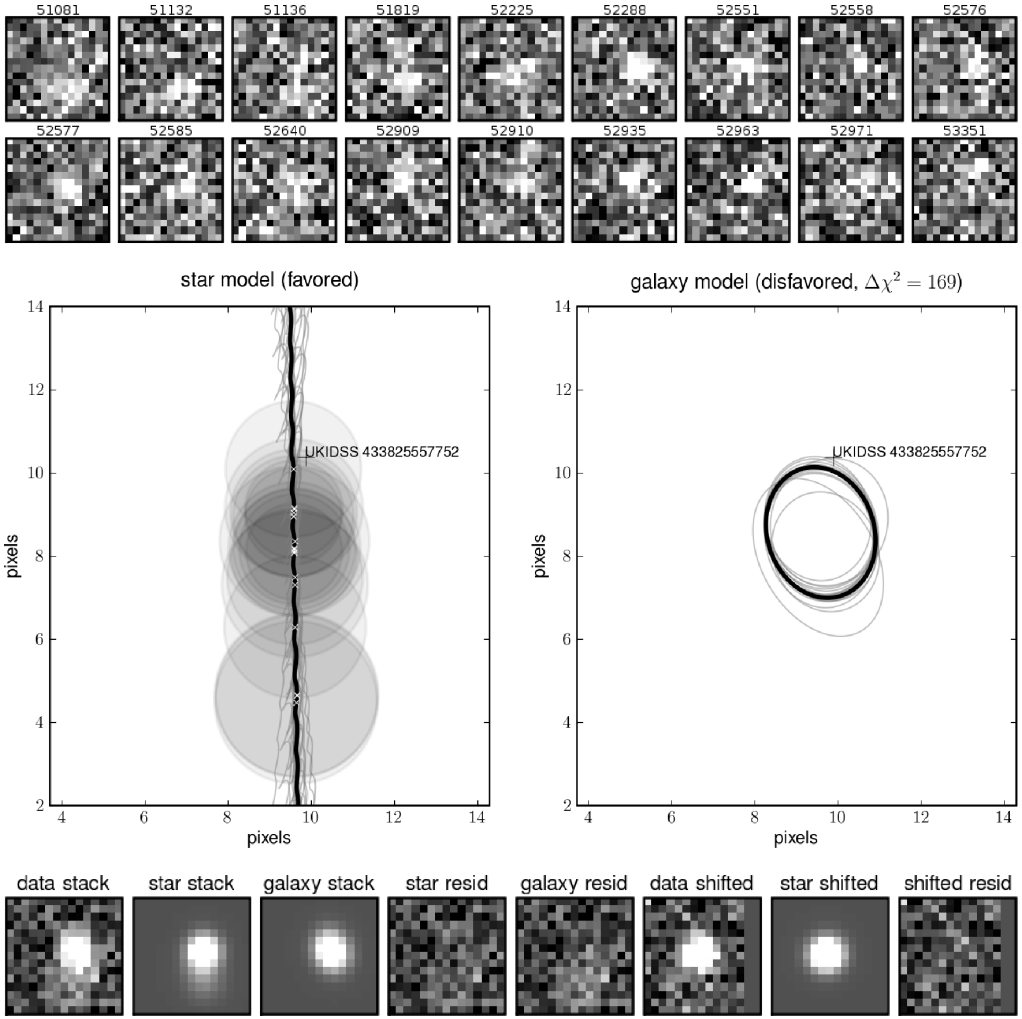


Image modeling

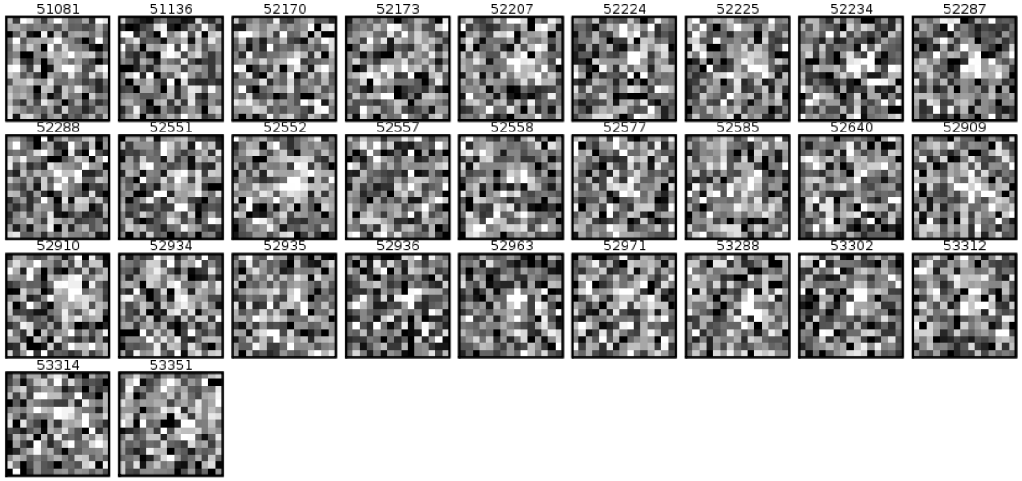
Polemic: Catalogs are dangerous (Hogg & Lang 1008.0738)

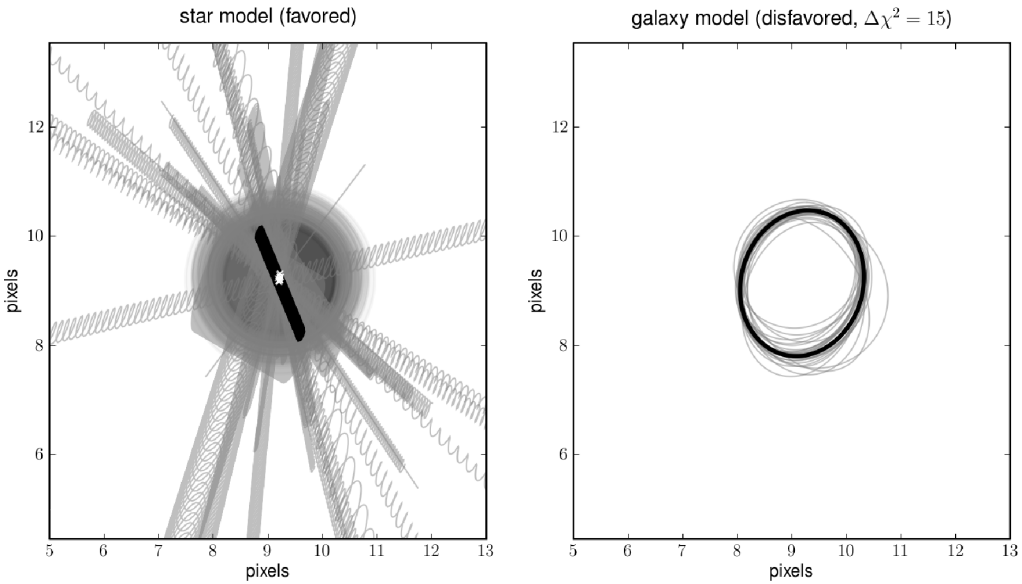
- Astronomers love catalogs. But:
- No objects are detected or classified with perfect confidence.
- Different investigators have different objectives and priors.
- As new data become available, the balance will shift for many objects.
- *Catalogs become wrong, likelihood functions are forever.*
  -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

Faint-source proper motions (Lang *et al.* 0808.4004): brown dwarf

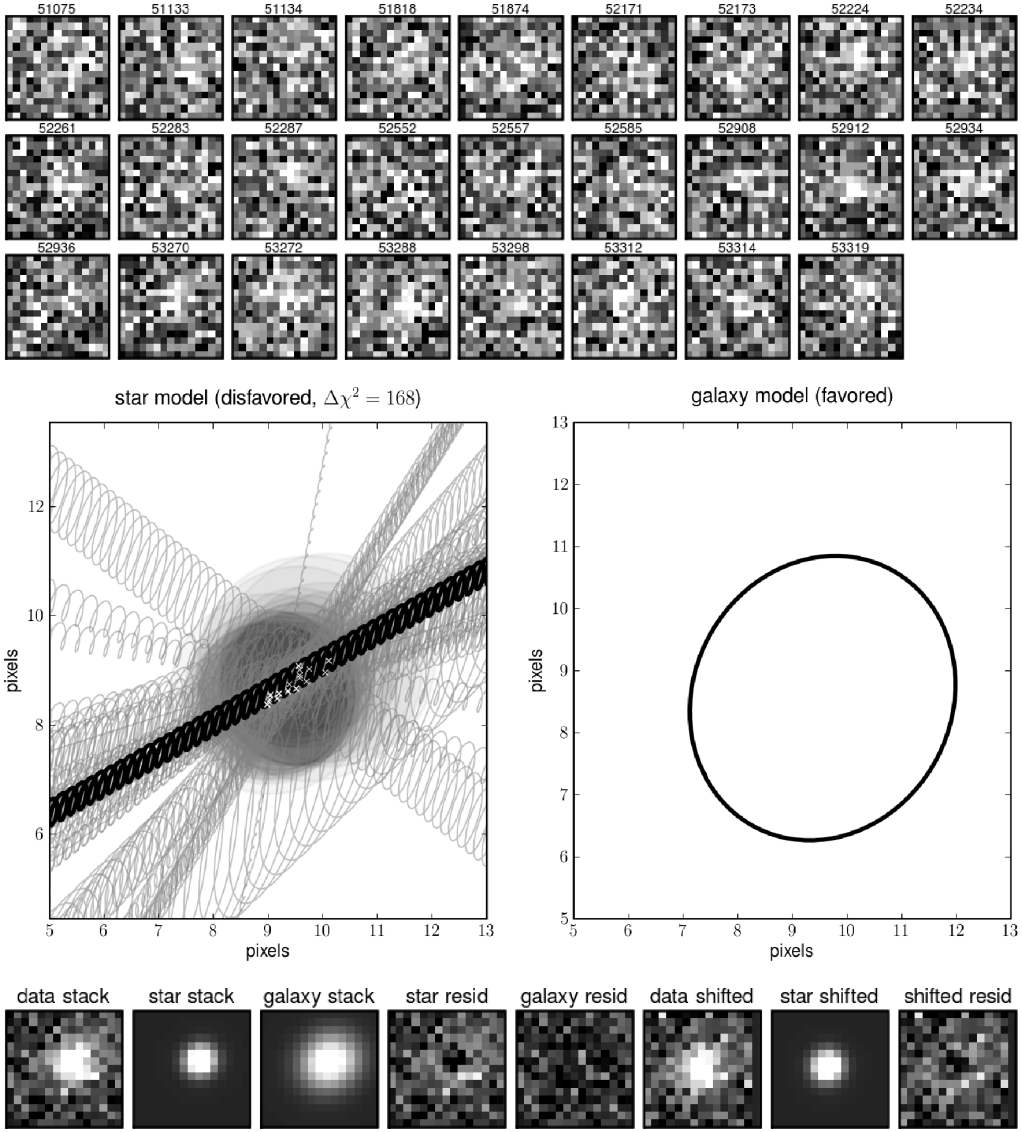


Faint-source proper motions (Lang *et al.* 0808.4004): $z \sim 6$ quasar

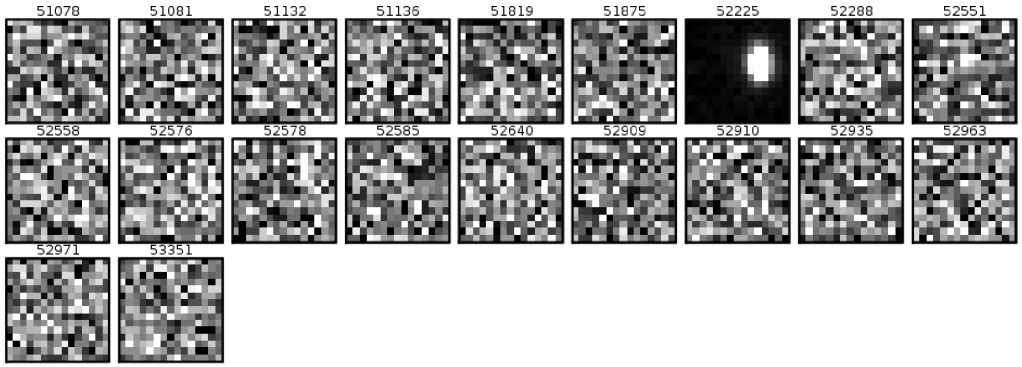




Faint-source proper motions (Lang *et al.* 0808.4004): faint galaxy



Faint-source proper motions (Lang *et al.* 0808.4004): defect



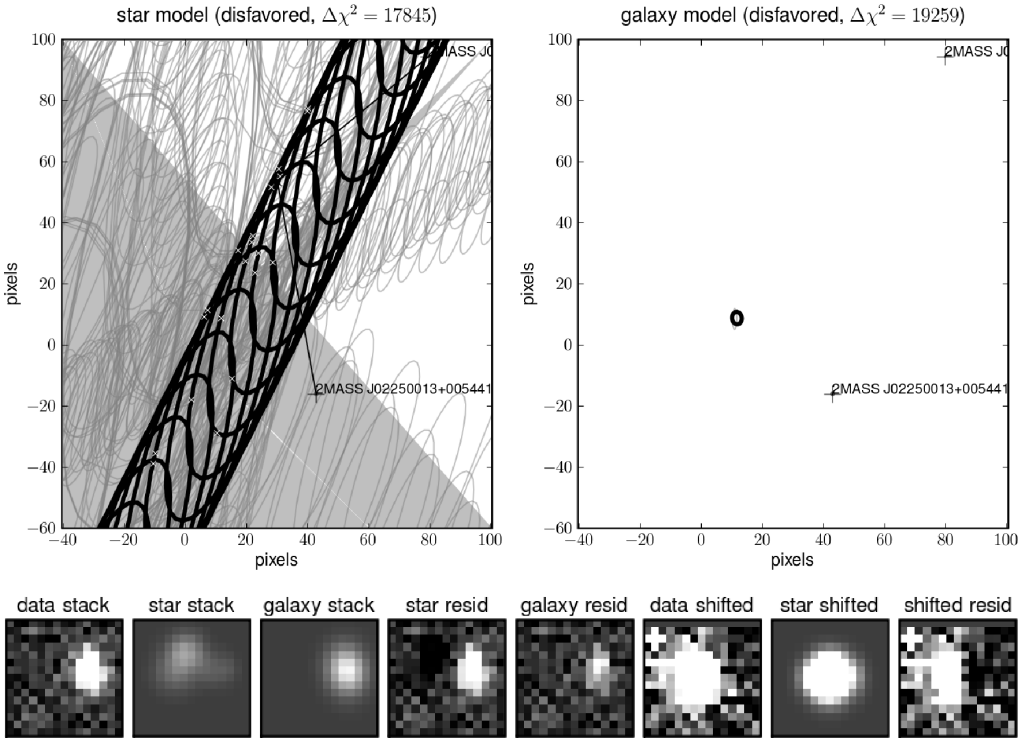


Image likelihoods (Lang): data

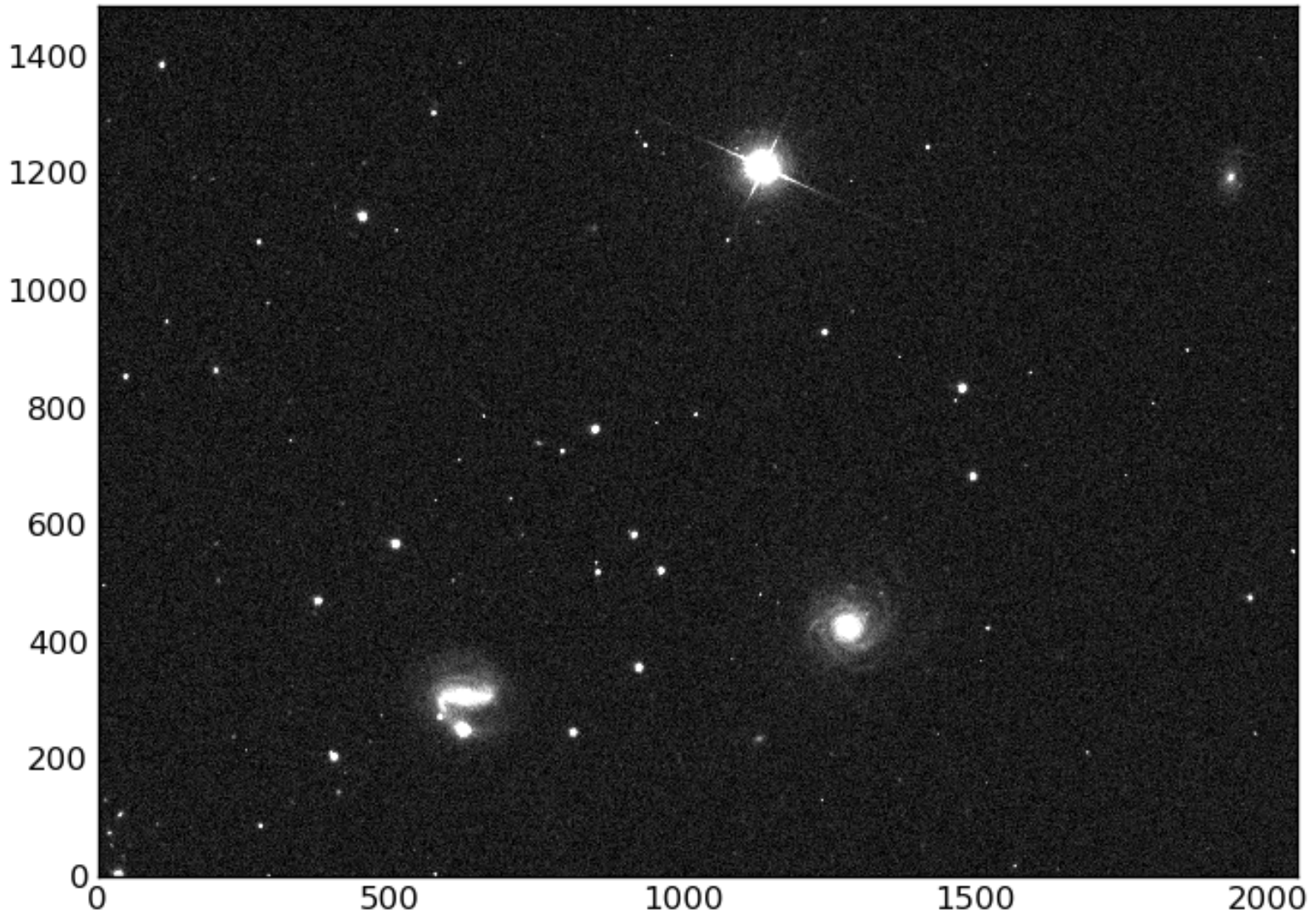
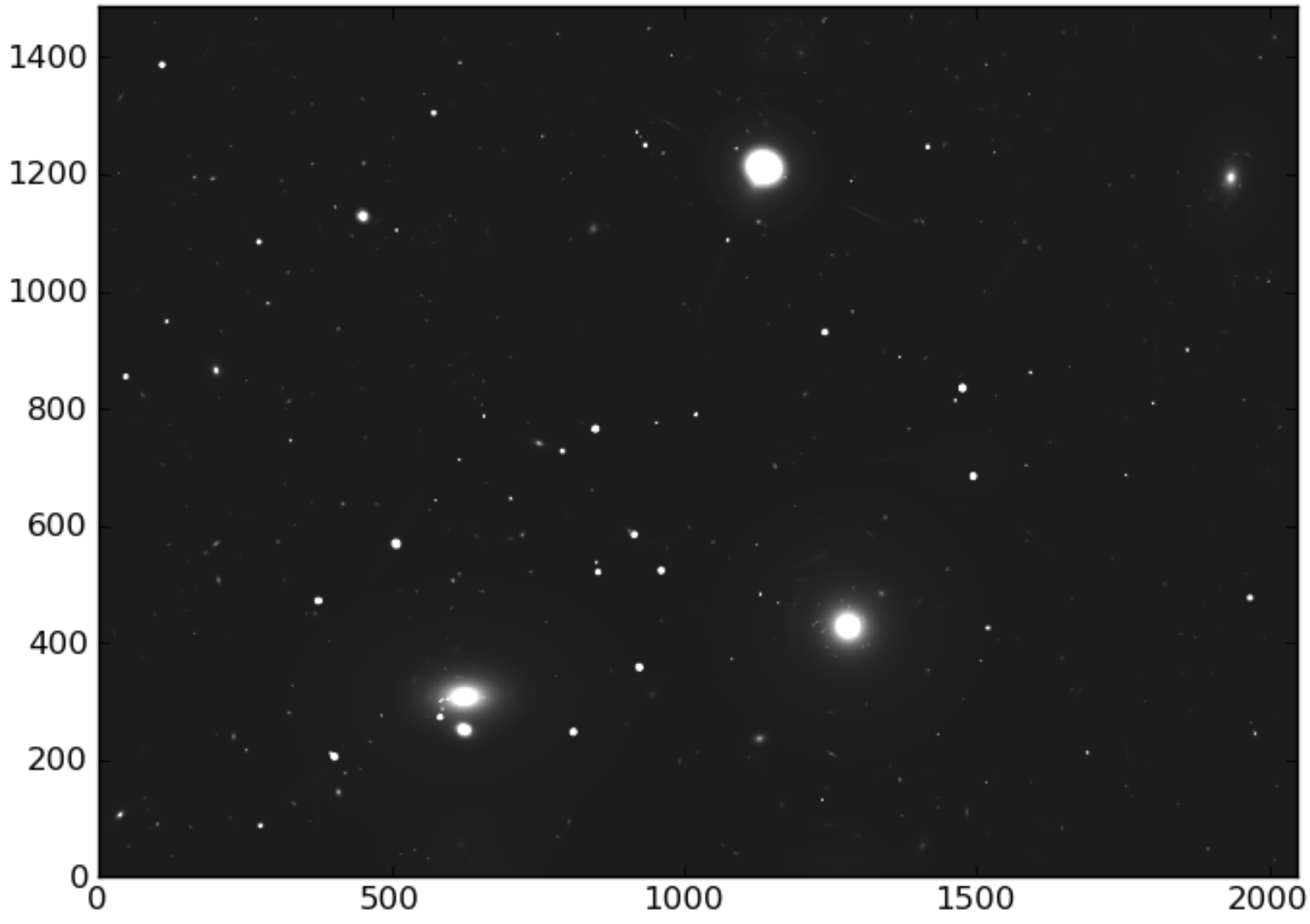


Image likelihoods (Lang): model



exoplanet direct imaging (Fergus & Hogg, in prep)