Semantic astronomy is doomed

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

2010 June 17

2. summary

- In it's naïve form, semantic astronomy is doomed to fail.
  - all meta-data are wrong
  - most interesting meta-data need to be *probabilistic*
- But we can save the *VO*.
  - third-party robotic endorsements
  - automated data analyses
  - standards for transmitting likelihood functions

3. A typical day at the *VO*

- fire up *DataScope*
- search for $(\mathrm{RA}, \mathrm{Dec}) = (177.322, 32.55)$ deg
- click an ungodly number of checkboxes
- get 193 resources, 37 sources of imaging, 133 images
- *most* of the images are either useless or duplicates!
- (in 2020, this query will return 10^4 images)

4. Do good data rise to the top?

- "In the VO architecture, there is nobody deciding what is good data and what is bad data, (although individual registries may impose such criteria if they wish). Instead, we expect that good data will rise to prominence organically, as it does on the World Wide Web. We note that while the web has no publishing restrictions, it is still an enormously useful resource; and we hope the same paradigm will make the VO registries useful."
  - *VO* Architecture Overview (Williams *et al.* 2004)
- Question: Was the rise to the top of great Web pages *organic*?

5. Hella not organic

- The reason when you search "dog" you get good pages is not that those pages rose to the top organically!
- The reason is that *Google* did an enormous amount of *computation* and *data analysis*.
- *Google* is the opposite of organic:
  - centralized authority; highly automated; objective
- *Google*'s power comes not from *containing* tons of pages; it comes from *understanding* tons of pages.

6. Why you need *Astrometry.net*

- It can create *ab initio*—or vet existing—meta-data.
- If it can run on an image, then that image is comprehensible.

7. *proposal:* Cryptographic robot taggers

- *Astrometry.net*, *PSF.net*, *Zeropoint.net*, *Bandpass.net*, *Noisemodel.net*, *etc.*
- could *rank* images on well-defined scientific criterion, *e.g.*:
  - sensitivity to quasars good for HeII reionization studies
  - ability to constrain proper motion of the SMC
  - capable of distinguishing brown dwarfs from high-redshift galaxies
- could operate autonomously in a scraping mode
- could generate meta-data and *sign* or tag versions
- could use simple cryptography for high levels of trust
- *VO*2020 simply *doesn't work* without this

8. A typical day at *Camp Hogg*

- (We just did this:)
- I have 10^8 point sources from *SDSS*; what are their *GALEX* fluxes?
- Catalog-match? No way:
  - resolutions differ by a factor of 30 (in solid angle)
  - signals-to-noise differ by factor of few to tens
  - blended sources are a mess
  - non-detects are not very informative
- The catalog entries are *not data*; they are *meta data*!
- We (Schiminovich & Hogg) had to photometer everything in the original pixels.

9. Catalog-level is dead

- Catalog matching is exponential-time:
- object i in catalog A and objects j, k in catalog B
- i could be j or i could be k
- i could be a blend of the two
- i could be a blend of j and something unseen
- i could be completely unseen
- j and k could both be blends of i and other sources
- and so on...
  - Budavari & Szalay, 2008, *ApJ* **679** 301
  - (though they imply it is linear in the catalog sizes)
- Catalogs are meta-data!
  - probabilistic
  - wrong in detail

10. What can be *known* about an image?

- observatory, telescope, instrument: YES
- observer, proposal title, abstract: YES
- date and time of observation, instrument mode: YES
- raw pixel values read out by the detector electronics: YES
- conditions: NO
- astrometric calibration or WCS: NO
- zeropoint, sky, and noise level: NO
- stars and galaxies inside the image: NO
- fluxes of stars inside the image: NO
- point-spread function: NO
- (why not?)

11. The problems with probabilistic

- Anything a calibration procedure returns is returned at finite precision!
- Anything said of a source is subject to change without notice.
- It is *literally impossible* for the meta-data to be correct.
  - Even a seemingly well-posed question like "how many stars brighter than $V = 20$ mag are inside this image?" has only a posterior probability distribution over answers.
  - calibration uncertainties
  - star–galaxy separation
  - PSF model improvements
- Only the *raw-data* electronics readout is stable.
- Precise experiments require *marginalization* over probabilistic meta-data.

12. *proposal:* World-scale image modeling

- imagine a detailed model of the entire sky
  - position and angular motion of every source
  - every SED and variability
- fit to every image ever taken
  - astrometric WCS
  - bandpass, zeropoint, PSF, noise model
- and then a *sampling* over the posterior PDF
  - sampling of information about every source
  - different samples might be qualitatively different
  - importantly, sampling over all calibration information
- You only get the probabilistic information right by *doing science*.

13. *proposal needed:* Meta-data likelihood functions

- How to distribute calibration meta-data with detailed enough uncertainty information that subsequent users can marginalize?
- K-element samplings are possible.
- Far future: All calibration outputs are callable functions?
  - thinking about this for catalogs too
  - Hogg & Lang, *Theory of Everything*, arXiv:0810.3851
  - Hogg & Lang, *SDSS* reprocessing?

14. summary

- In it's naïve form, semantic astronomy is doomed to fail.
  - all meta-data are wrong
  - most interesting meta-data need to be *probabilistic*
- But we can save the *VO*.
  - third-party robotic endorsements
  - automated data analyses
  - standards for transmitting likelihood functions