Designing Astronomical Projects for Real-Time Discovery, Operational Efficiency, and Statistical Re-Use of Legacy Data

David W. Hogg *(NYU) (MPIA) (Flatiron)*

Thank you

- Megan Bedell (Flatiron)
- Josh Bloom (Berkeley)
- Doug Finkbeiner (Harvard)
- Dan Foreman-Mackey (Flatiron)
- **Dustin Lang** (Toronto)
- Sam Roweis (deceased)
- Bernhard Schoelkopf (MPI-IS)
- **Dun Wang** (NYU)
- ...and many more

David W. Hogg *(NYU) (MPIA) (Flatiron)*

(this talk)

- This is intended to be a conversation starter about scientific projects.
    - (and my visuals suck)

Exoplanet discovery

- There are six methods for finding **planets around other stars**.
    - Transit, radial-velocity, timing, direct imaging, microlensing, astrometry.
- Transit (thousands) and radial-velocity (hundreds) are the market leaders.
    - But not for long!

What is a transit?

- Because of insane good luck, a planet passes between us and its host star.
- It blots out a fraction of the star’s light that is proportional to the area ratio.
  - Rp^2 / Rs^2
  - (modified by lots of details)

What is the radial-velocity method?

- Planet and star orbit a common center of mass.
- Look for star’s consequently variable Doppler shift.
- Jupiter-like planets induce few-m/s signals.
- Earth-like planets induce 10-ish-cm/s signals.

NASA *Kepler* Mission

- 42-CCD camera pointing at 100 square degrees continuously for 4.1 years.
- Designed with the sole purpose of finding Earth-like exoplanets.
    - All trades made for simplicity and stability.
    - Data designed to require minimal calibration or spacecraft knowledge.
    - Inexpensive.
- Delivered 2300 to 4000 planets (depending on definitions).
    - (far exceeding all expectations)
- Also did amazing science with stars
    - (and there is a whole awesome *K2* story.)
    - (and an awesome #openscience #otherpeoplesdata story)

A transiting exoplanet

- **A transit is a blunt tool**.
- Learn period, a radius ratio (planet to star) and a stellar density (yes, density).
- Almost no information about orbital eccentricity or planet mass.
- Also, the signal is exceedingly **sparse**.
    - Earth transits the Sun (for some exo-astronomers) for 13 hours every 365.26 days.
    - Exo-astronomers have to be both **lucky and persistent** to observe it.

Extreme-precision radial-velocity experiments

- *Typically:*
- Have a few to many nights per year on an expensive telescope.
- Measure Doppler shifts for a dozen or so stars per night.
  - These days: With a precision of 1 m/s! Soon: 10 cm/s!!
  - (precision and accuracy very different here!)
- Don’t want to “waste time” on stars that won’t produce planets.
- Harsh decisions are made.
  - *This star’s past data look more promising than this star’s*.
  - There are even exoplanet savants!
  - Telescope time-allocation committees demand efficient programs.
- Literally **no-one has ever fully automated these decisions**.

Humans

- *I can't say this enough:*
- We are swinging around literally hundreds of millions of dollars of equipment on the unrecorded **whims of humans**, talking to each other in windowless conference rooms.
    - and for what?

Exoplanet populations

- A (or the?) key goal of all this is to figure out the properties of the full population of planets.
- We want to account for (very very complex) selection effects.
    - Shorter periods easier to find.
    - Larger planets easier to find.
    - “Quieter” stars easier to search.
    - Plus non-trivial dynamical issues, like resonances and interactions.
- Radial-velocity experiments deliver more information per system, and yet:
- All populations inferences have been performed with the **transit data alone**.
    - Why?

Exoplanet populations

- There are more planets than stars!
- Super-Earths and mini-Neptunes are (by far) the most common planet types.
- The Solar System is not really “typical”.

Humans *vs* machines

- Humans are ~~fucking~~ great at efficiently finding planets!
    - Even NASA *Kepler* had human-vetting in the loop until its final data release.
- But humans are ~~fucking~~ hard to model quantitatively in any statistical model of planet populations!
    - And you will simply get wrong answers if you don’t model the human decisions.
    - And humans really, really don’t want to “waste time” on control samples!
- If the RV communities had accepted some efficiency hit for algorithmic operations and control samples, their data would have been useful for something other than just discovery.
    - But, to be fair, no-one would have given them telescope time.

Exoplanets

- I’m being negative.
- But seriously, the breakthroughs in exoplanet science over the last decade, from joint analysis of many data streams, have been absolutely incredible.

switch gears

Sloan Digital Sky Survey

- First really big university-based astronomy research project.
- Took imaging of ¼ of the sky in five optical passbands.
- Took spectra of 1,000,000 galaxies and quasars to obtain **redshifts**.
- Measured the inhomogeneous structure in the Universe and it's growth with cosmic time.
- **All parts of the survey were operationally algorithmic and repeatable**.
  - It was literally designed with long-term statistical legacy value in mind from day one.
  - (there is also a great #openscience and #otherpeoplesdata story here too)

Sloan Digital Sky Survey

- Project was enormously over-designed:
- Imaging was far deeper than needed to target spectroscopy.
  - (factor of hundreds in observing-time equivalent)
- Spectroscopy was far higher in signal-to-noise than needed to obtain redshifts.
  - (factor of tens in observing-time equivalent)
- Project went enormously over budget.
  - (it was completed by the University partners)
- Project was **enormously productive**.
  - There are almost ten thousand papers from the original survey, and thousands more from the subsequent surveys *SDSS-II*, *SDSS-III*, and *SDSS-IV*

What to conclude?

- *SDSS* didn’t get the balances all exactly right, but it was successful for two reasons:
- It was operationally algorithmic and statistically reliable.
- The over-design meant that there was lots of **additional, unplanned discovery space** for astronomers and cosmologists across all domains.
- Part of the *Survey*’s success was a result of its **inefficiency**.

Present-day cosmology experiments

- Cosmology is a mature field with extremely mature questions.
- We are required (by funding partners) to make new projects highly efficient for measuring particular parameters.
- This could have disastrous consequences.

(off topic?)

- What does this have to do with *Real-Time Decision Making*?
  - Nothing! But it does have a lot to do with *Decision Making*.

Valuation and discovery

- Fundamentally the issue is that we can put quantitative measures over improvements in parameter estimation.
    - Cramer–Rao bounds or Fisher information.
    - Information or entropies.
- We don’t know how to put quantitative measures over unplanned discovery space.
    - The situation is a bit like testing in schools: It leads you to value what you **can test**, rather than what you **want to test**.

switch gears

Astrometry.net

- Here’s an image of the night sky. What’s the pointing, rotation, and scale?
- The **first-ever reliable image-recognition system** in any domain.
  - *Make it rain!*
- That said, it is in a domain that has **no commercial value** whatsoever.
  - *(whoops)*

Astrometry.net

- Automated detection of stars and background in arbitrary imaging.
- Lookup of geometric hashes in immense database.
- Fastest kd-tree (of its type) in the world (at the time).
- Bayesian inference conditioned on stars in the image.
- Explicit decision-theory implementation for user response.
- Automatic visualization of results.
- Open data and code.
- Dustin Lang *et al* (2010) *The Astronomical Journal* **139** 1782–1800

Bayesian decision theory

- *Astrometry.net* obtains probabilistic information about the pointing, rotation, and scale of the astronomical image.
- It has to *decide* what to say to the user.
- We explicitly cast this as a decision theory problem:
  - What are our long-term cash-flow implications if we give an answer and it’s right?
  - What if we give an answer and it’s wrong?
  - What if we don’t give an answer but we could have?
- We make the decision that maximizes our expected long-term **cash flow**.

What?

- We had to make a judgement about the astronomical community’s long-term view of *Astrometry.net*:
  - Would they trust a system that gives wrong answers?
  - How would reputation propagate in the astronomical community?
  - How much do we value different kinds of users with different attitudes towards risk?
- This is hard!
  - In the end we just made up reasonable numbers.
- This is hard, and it’s in a domain of **absolutely no consequence whatsoever**.
  - If we want to get serious about quantitative decision making in real contexts (like, say, **self-driving cars**), it’s going to hurt.

Long-term future discounted free-cash flow

- **Long-term future**: All parameters and context are variable.
  - (so there is no confident computation of anything)
  - (the specific time frame for the long term is very context-dependent)
- **Discounted**: You prefer cash now to cash later.
  - (discount rate is different for investigators in different positions)
- **Free-cash flow**: Current revenue less current expenses.
  - (cash that could be paid out to investors without impact to the present-day scale of the business)
  - (doesn’t take account of expenditures to make capital improvements for growth; we do what we do in order to get those opportunities)

Scientific LTFDFCF

- The fact that the objective is explicitly long-term means that it can’t be precisely estimated.
- The discount rate is a strong function of career stage or project scale.
- The costs and benefits of individual scientific papers are measured in the **hundreds of thousands of dollars**.
- It is pointless to compute the LTFDFCF in any units other than real currency units (*eg*, USD or EUR or BTC).
    - You need to make trades between software, hardware, personnel, travel, and all other budget categories.

Implications for RTDM

- We need to build utility models for trades.
- Our utility models need to explicitly look towards **end goals** of discoveries:
    - publications,
    - future grant funding success,
    - junior-scientist careers,
- We need to think about outcomes in terms of probabilities and utilities.

Implications for model selection

- *On a tiny, tiny scale:*
- Any time you say “I have discovered X” or “I have ruled out Y” you are explicitly making (and announcing) a **decision**.
- Most hard-core Bayesians think you do this by marginalizing likelihoods.
  - But that’s wrong!
  - And it is **super-duper espensivo** to do these marginalization integrals precisely.
- You really ought to do this by integrating **utilities** over posterior beliefs about outcomes.
  - And if you don’t know the utilities precisely (see previous slides),
  - there is no point in doing these integrals precisely!
  - It’s not just the Bitcoin miners that are wasting CPU cycles.

switch gears

Star bucks?

- Imagine you had a heterogeneous network of astronomical resources.
  - And note: The astronomical community, viewed as a whole, does!
- How would you objectively and optimally apportion these resources?
- You would run a live auction (or something like it).
  - *cf*: Endless late-night conversations with Josh Bloom.

Global telescope networks

- There are now many global telescope networks, some operated by professionals and some operated by amateurs.
- There is much focus on making sure that the nodes in the network are as **similar as possible**.
- That choice is **pessimal**.
    - **Q:** Who, when hiring employees, tries to hire people who are as similar as possible?
    - **A:** Only extremely bad employers.
- But that choice (homogeneity) is made because there is no theory of how to make trades across heterogeneous assets.

Managing heterogeneous networks

- Our vision for RTDM in the *LSST* and *SKA* era is autonomous trades among assets in a heterogeneous network.
- Each asset will have a utility model.
  - A function of what it observes, and what it is trying to measure or discover.
- Each asset earns and spends cash (real if possible) making trades with other assets in a live market.
  - Notice that some assets could be human-operated; this encourages human–machine collaborations.
- **This is not a near-term goal**.

Heterogeneity in general

- Most astronomers abhor heterogeneity:
- Data are easier to use if they are flat, identical, evenly sampled.
- But in every context, there is more information if the data are heterogeneous.
  - This is not just a theoretical point:
  - Consider sampling theorems in evenly vs randomly sampled time series.
  - Consider time-scale sensitivity as exposure times are varied.
  - Consider spectral coverage as wavelength bandpasses are moved.

Sloan Digital Sky Survey, once again

- Original plan for imaging was to do a single pass over the sky. One shot. Calibrate externally, and rely on system stability for calibration stability.
- We moved the survey to a mode in which it did some overlapping imaging, so the same star would see a few different detector locations.
- This operational heterogeneity permitted us to self-calibrate the survey with **no use of any external calibration data whatsoever**.
    - The self-calibration ended up far more precise than the external calibration.
    - It was adopted at DR8 and for all subsequent data releases.
- Even this **tiny bit of heterogeneity** made the *Survey* far more capable.

NASA *Kepler*, once again

- Kepler kept its pointing as stable as possible, so that each star is always nailed to the same focal-plane position as precisely as possible.
- This led to very stable, homogeneous data, but prevented us from learning anything about the spacecraft calibration.
    - This limited our final precision, and precluded certain kinds of investigations.
- We could have learned more (but at substantial operational cost) with some small dithering of the spacecraft.
    - We show this empirically—in some sense—in Wang et al, *arXiv*:1508.01853

Operations and ethics

- Everything I have said here would substantially **complexify operations**.
  - That is, the hardware would become cheaper, and we could do more science, but not nearly as simply.
  - There will be **hardware–software trades** to consider.
- Astronomy is paid-for by the public; we have an obligation to maximize scientific return on those dollars.
  - (even astronomy funded by the *Simons Foundation* is ultimately paid-for by the public)
- **Full Disclosure:** My group is a computational data-analysis group. We benefit directly from complexity in operations!
  - So long as it doesn't involve modeling humans.

Radial-velocity experiments, once again

- If we can make progress towards specifying utility,
    - (even very, very poor approximations)
- then we can create algorithmic, adaptive programs for radial-velocity exoplanet discovery that are explicitly optimized to maximize utility.
    - Related to active learning.
- If the utility is sophisticated enough, the program will generate all needed statistical controls along with high efficiency at finding new systems.
    - We are working on these utility proxies now (though we are not close).
    - They are based on **discounted information about exoplanet populations**.
- And then we can use (or combine) RV data for population inferences.

Discussion triggers

- Exoplanet science and the problem of having **humans in the discovery loop**.
- Decisions are different from inferences! They involve **utility**.
- **Your utility is your LTFDFCF**.
- Astronomical projects need to be **designed for discovery** goals.
- Almost any scientific or engineering goal benefits from **heterogeneity** of hardware and observing.
- As heterogeneity increases, **software inevitably becomes more complex**.