Joint analyses of data: The future

David W. Hogg (NYU) (Flatiron) (representing no-one but myself)
Now, CMB missions need LSS surveys and *vice versa*.  
- lensing, ISW, contamination

NASA Lambda Archive was created (in part) to share **likelihood functions**  
- (will return to this point later)

Two (at least) kinds of joint analyses:  
- combine high-level constraints on cosmological parameters  
- combine low-level information about individual objects or sources or pixels

The cosmology community is extremely sophisticated here, and we can learn from them.
context: the Large Hadron Collider

- Data are exceedingly complex.
  - triggers, jet identification, missing transverse momentum, and so on
- Simulations are very big and slow.
  - full model of the standard model and the machine
- Hardware has enormous numbers of calibration parameters.
  - you don't just “see” the events.
- Data releases have been limited and are very hard to use.
- There are projects underway to build intermediate products for re-use
  - RECAST, for example
- There is no trivial solution to these problems.
context: Astronomical data growing in complexity

- CMB stage-4 goals are extremely foreground-sensitive.
- 21-cm and other line intensity mapping experiments even more so.
- SKA and ALMA producing interferometric visibilities; and very corrupted data.
- Exoplanet radial-velocity and transit missions looking for part-in-100,000 variability on top of part-in-100 systematics.
- In general: as scientific goals get more mature, projects produce data that is harder to naively process.
Growing issues with reproducibility and failures-to-reproduce.
  ○ many examples in the social sciences, but there are physical-science equivalents

Blinding and hypothesis pre-registration are key tools for the future.
  ○ again, cosmology leads here

Every scientific result in astrophysics suggests a hypothesis pre-registration for future data sets.
issue: Experimenter knowledge

- As data become more complex, the knowledge of the system builders becomes more valuable.
  - compare HST and Planck raw-data streams.
  - or SDSS and the new 21-cm experiments.
- The data are responsibly used by the experimental team for their goals.
- The team knowledge is encoded in the data-analysis procedures applied to the data.
  - team knowledge is folklore or implicit knowledge
  - rarely are scientific papers reproducible in all the relevant senses
  - by construction, (almost) everything the team knows is encoded in data-analysis software
issue: Likelihood functions

- If you want to combine data from different experiments, you want to **multiply the likelihood functions**!
  - (not multiply the posteriors)
- Team-built likelihood functions contain the team’s implicit knowledge about the data.
- This is true whether we are combining at high level (cosmological parameters) or low level (individual object or pixel properties).
  - the NASA Lambda Archive is aimed at the former.
  - among other things, it is a likelihood archive
  - there are currently **no standards for propagating likelihood information** at the pixel or object level
issue: Expecting the unexpected

- Open data support **investigations not imagined** by the experimental team.
- That is, we need to give tools and data products that are useful for all scientific investigators.
- This is an ill-posed problem!
Recommendations

● Since team knowledge is encoded in the software, **data releases must also be software releases.**
  ○ (and probably *vice versa*)

● **Level 1:** Data releases should be **accompanied by likelihood-function releases**, as software or as executable APIs.

● **Level 2:** All **data-analysis software should be released open-source** for re-use along with the raw-data inputs for that software.
  ○ such that published results are fully reproducible by (say) an advanced graduate student

● **Level 3:** Plus **full documentation** such that truly *ab initio* and qualitatively different data analyses are possible.
  ○ again, I suggest the standard of an advanced graduate student
Take-home

- Data releases only make sense with appropriate, rich, associated software releases.
  - This permits arbitrary future joint analyses and new discoveries.
  - This provides tools for pre-registration and reproducibility.