Detecting the Unexpected 2017

---

shared information, minutes, and running notes

Rules for this document: Wikipedia Rules! Edit. Be bold. Don't imagine that someone else will type in what you wanted to see here.

Meeting information

Dates: Feb 27 - Mar 2 2017
Location: STScI

Day 1

Morning talks

- *Help fill in if you have notes!* <Speaker name>
    - Notes here...

Unconference session summaries

- Jupyter notebooks as data products
    - Many use pipelines to create data products (results) that are served on the web.
    - These are static products. In some cases, that is enough, but reanalysis is useful some fraction of the time.
    - Jupyter notebooks can serve this role by encapsulating pipeline results, being portable, and supporting reanalysis outside of the initial computational environment.
    - Questions/concerns over fact that code/data that supports notebook must be built/downloaded separately. Actual environment (variables, state) is not portable with the notebook.
    - Still likely some fun in converting a few pipeline scripts into notebooks.
- Improving the integration of humans and machines in classification/discovery
    - Firstly: we all agreed that we'll always (at least within our lifetimes) need both humans and machines. There may be some use cases where humans are only required to physically interpret results, but we are not yet at the stage where we can even envision what a machine to do ~perfect discovery *and* classification would look like.

    - The question then becomes: when is the optimal time to insert humans into the process? Depends on specific goals:
        - Training
        - classification
        - discovery
        - regression
        - physical interpretation
    - There was a general feeling that much of what we're doing should be moving to active learning, perhaps with a human community running along with it to access collective intelligence
    - A note that no matter how awesome we make these machines, their use will be limited if we can't work out how to apply them efficiently, even to old data. Older catalogs still have plenty of unexplored discovery potential, but even if you do some work on an old catalog there's often no way to share what you've found or not found (if it's not worth publishing a full paper for).
    - The more complex our machines become, the more challenging it becomes to physically interpret their outputs. Many of the advancements made recently in ML tend to turn algorithms/networks/etc into a black box. This is probably the opposite direction we should try to go if we want to better integrate machines and humans. How can we get machines to output results with explanations that are understandable by humans?
- *Help fill in if you have notes!* <session title>
    - What did you talk about?

Afternoon talks

- Arjun Dey
    - Data sources are changing from targeted, planned observations to large coherent projects that (may) take data in survey mode and release publicly
    - So how do we discover new things in canned surveys?
    - DESI
        - Starting ~2019A
        - 40,000 spectra per night, R ~ 5000, 3800-10000 Angstroms
        - Primarily a redshift survey -> cosmological parameters
    - Pre-imaging surveys (e.g., DECaLS) data released regularly, no proprietary period
    - Finding interesting targets harder with spectroscopy -- need massive template libraries
- Dalya Baron: Anomaly detection in galaxy spectra
    - What is an outlier?
        - "Bad" objects (cosmic rays, artifacts, etc.)
        - Misclassified objects

        - Tail of a distribution
        - Unknown unknowns - completely new things, not predicted
    - Searching for outliers in galaxy spectra from SDSS (~2 million)
    - Assign a distance between all pairs of objects
        - Distance assignment done with a random forest
        - ??
        - Find objects that have largest distance from all pairs
    - Of the 400 weirdest galaxies:
        - Most (⅓) are BPT outliers - some AGN activity?
        - Small fraction are bad spectra
        - ~⅛ are stars
        - Find many “extremely red” galaxies -- huge column densities
    - Q: Are the stars are in there because they are also stellar outliers? That suggests you should be running on stars too!
    - Now working on APOGEE and MANGA spectra
    - https://github.com/dalya/WeirdestGalaxies
- Stephanie Juneau: Black Hole-Host Galaxy connection with large datasets
    - AGN identification is more contentious than one might think.
    - Biases abound and are still being revealed.
    - Also: http://datalab.noao.edu/
    - Data sets in datalab include DESI imaging, SMASH, ...see tomorrow’s bazaar
- Branimir Sesar: Enhancing the PS1 3pi Survey Catalog with Machine Learning: Outlier Detection and Star-Galaxy Classification
    - Pan-STARRS1 data is multi-epoch and multi-band, but 70 obs over 4 years and 5 (non-simultaneous) bands.
    - Combine that with occasional bad data and inference becomes very hard.
    - How to identify outliers? 50k G/K dwarfs (not variable) are used as input to supervised learning algorithm.
    - Tree-based algorithms are informative because trees are interpretable. E.g., now know that PS1 photometric flags are very *un*informative.
- Kiri Wagstaff: Discovery via Eigenbasis Modeling of Uninteresting Data
    - “There is no such thing as an objective anomaly”: Anomaly detection requires definition of normalcy.
    - Incorporate what user already knows/has seen
    - Process is to select a point as an anomaly, then add it to the normal set. Find next most anomalous, repeat…
    - This is unsupervised, but could imagine serving anomalies out for expert opinion.
    - https://github.com/wkiri/DEMUD
- Armin Rest: Pan-STARRS1 Data Archive
    - The good: relative/absolute photometry (ubercal/supercal), astrometry (after gaia recalibration)
    - The bad: orthogonal transfer ccds never worked well enough to use. Bad pixels.
    - The ugly: photometry sensitive to details (stacks, forced, etc.)

- Rick White: The Hubble Source Catalog
    - Heterogenous archive. Various instruments, irregular spatial coverage
    - Irregular astrometric quality is particularly tricky.
    - 10^8 objects have a single measurement (filter). 1 has ~180 measurements.

Day 2

Morning talks

- Chris Lintott: Citizen Science in the era of Big Data
    - 2 Arguments
        - We're very clever, but there's a place in the future for the public
        - Doing that is complicated, difficult, and interesting
    - Astronomy is full of zorillas, and LSST is not big data
        - I think the zorillas in LSST won't be big data (even if we get thousands of instances of each zorilla, that's not "big") while LSST is big data of normal stuff (Zebras, Lions, etc.)
    - Combining human & machine classification is valuable but should be done with care (blindly using human classification as a training set is not optimal, because it will primarily teach the machine about things that humans find easy to classify)
        - Classification in parallel, by humans and machines independently, looks like a good way to find unusual objects
        - Even "normal stuff" in astronomy is not always universally recognisable to non-astronomers (contrast lion/zebra with elliptical/spiral galaxy)
- Brooke Simmons: Building Citizen Science Projects with the Zooniverse
    - Jargon
        - Subjects = observations to be analysed
        - Workflow = research question defined as series of tasks for volunteers
        - Talk
            - Public involvement
        - Classification (Aggregation)
    - http://www.zooniverse.org/lab
- Steven Silverberg: Disk Detective: Searching WISE for New Debris Disks with ~30,000 New Colleagues
    -
- Julie Banfield: Radio Galaxy Zoo: citizen science and machine learning
    -
- Michael Zevin: The Future of Citizen Science: Coupling Crowdsourcing and Machine Learning
    - Gravity Spy
    -
- David Hogg: Tracing the Milky Way's assembly with data-driven spectroscopy

- Graziano Ucci: Inferring Physical Properties of Galaxies from their emission-line spectra - a Machine Learning Approach

Unconference session summaries

- How to train students for research methods?
  - What level of students are we talking about?
    - Recommendations depend on level of the student
  - On the topic of having students write a blog
    - Meant for marking progress, writing down intermediate progress, advisor and other group members can keep track
    - Some pushback: much more flexibility with pen and paper. Notebooks let you sketch, feels like a lower bar than writing something formally on a blog
      - Phil writes everything in little notebooks / journals
  - Specific topics:
    - Notetaking
    - Software
    - Workflows
    - Statistics?
    - Encapsulate ideas in writing
    - Visualization
  - For students, you want to reward the process not the result
    - Self-assessment can help - set goals for a time period
  - Slack
    - Lowers the bar for communication, inter-group and group-advisor
  - How much time do we let students explore, discover?
    - Important to show that projects fail and that's ok
  - "The hardest part of grad school is crippling self-doubt"
  - Having a novice teach a novice is far more useful
  - For group meetings, have a rotating leadership role between grad students and postdocs?

Afternoon talks

- Thomas Robitaille, Freelance: Interactive visualization and exploration of big, heterogeneous, and highly-dimensional data
  - Horse People - Zorillas of google earth
    - Billions of eyes on one big data set finds things that are impossible to predict
  - http://www.glueviz.org/en/stable/

- Amr Hassan: Real-time Data Analysis and Visualisation on Commercial Cloud Infrastructures
    - 3D big data visualization
- Tom Donaldson: Accessing data through MAST