

Evaluating Data Science Contributions in Teaching and Research

Or...

How can I get promoted/tenure as a Data Scientist?

Lance A. Waller

Department of Biostatistics and Bioinformatics

Rollins School of Public Health

Emory University

Outline

- Me, you, and your promotion/tenure process
- The evidence: Your dossier (Research, Teaching, and Service)
- The standards of evidence in your dossier (Exhibits 1, 2, 3)
- Data Science Research: What's different and how do you document it?
- Data Science Teaching: What's different and how do you document it?
- Bringing it all together

Me: I wanted to be cool...



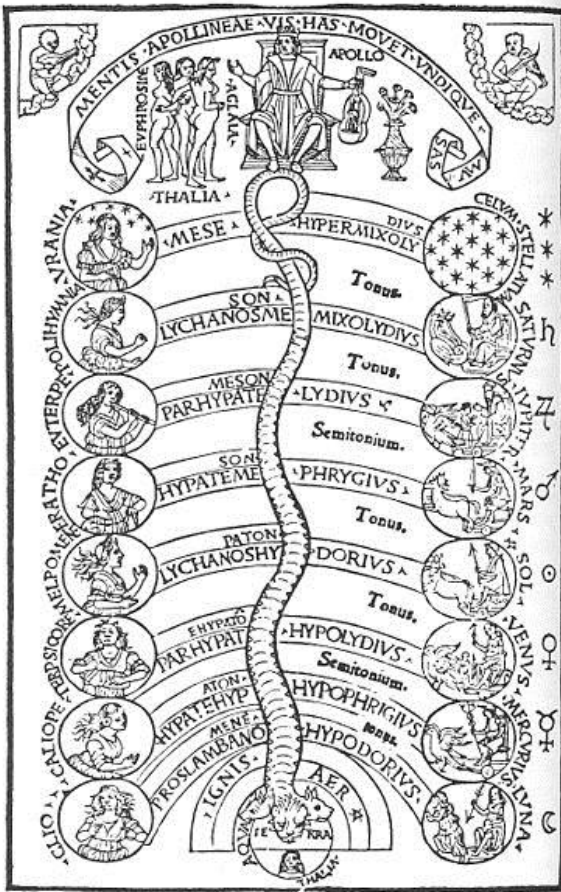
What actually happened...

- Biostatistics Professor
- Department Chair
- Faculty Hiring and Mentoring
- Promotion and Tenure (one step of the process)
- My goal: A slam-dunk promotion from Biostatistics and Bioinformatics, every time.
- What about Data Science?
 - JSM 2014: Presentation on Data Science to Chairs' Workshop by Jeff Leek.
 - Conversations with Roger Peng, Brian Caffo, others over past 12 months.

You:

- Cool.
- Doing Data Science, generating scholarly productivity in new areas of:
 - Research
 - Teaching
 - Service
- Your goal: A slam-dunk promotion.

Steps in the Pre-Copernican (academic) promotion process



- Trustees/Regents/etc.
- President
- Advisory Committee(s)
- Dean
- P&T Committee
- Chair
- Department
- You

Key steps in the promotion process

- Know the rules, know the rules, know the rules.
- Who will be evaluating you?
 - What fields do they represent?
 - What represents quality to them and their field?
- What documentation is required? Expected? Allowed?
- Discuss the process with your chair, your Promotion and Tenure Committee representative, your colleague who just went through the process, your Dean.

Boyer's *Scholarship Reconsidered*

- An important resource: In 1990, Earnest Boyer of the Carnegie Foundation authored a report entitled *Scholarship Reconsidered: Priorities for the Professorate*
 - Available on Amazon
- Widely cited, well known to the upper spheres of influence.
- What's so important about it for this conversation?

Boyer's *Scholarship Reconsidered*

- Boyer defined four types of academic scholarship
 - The **scholarship of discovery**
 - The **scholarship of integration**
 - The **scholarship of application** ([scholarship of engagement](#)); and
 - The **scholarship of teaching and learning**
 - https://en.wikipedia.org/wiki/Boyer%27s_model_of_scholarship
- **Provides a broader context for scholarship.**
- Deans, Provosts, and Presidents talk about this.

Evidence: The Dossier

- Key components:
 - Exhibit 1: Personal statement
 - Exhibit 2: CV highlighting accomplishments in Research, Teaching, and Service
 - Exhibit 3: External letters from experts in the field

Exhibit 1: Personal Statement

- Tell your story, put work in context.
- **Highlight past accomplishments.**
- Highlight focus and recognition.
- Highlight unique features and motivators.
- Highlight goals and **establish future trajectory.**

Exhibit 2: CV (typical measures of success)

- Research success
 - Peer-reviewed publications
 - Competitive grant funding
 - Invitations to speak
- Teaching success
 - Courses (with evaluations by students, peers)
 - New ideas? Did they work?
- Service success
 - Completed projects (publications again)
 - Strong collaborations

Exhibit 3: Letters of support

- “Arm’s length”
- Comment on accomplishments, unique features, and likelihood of continued success
- Three components to each letter:
 - Letter **content** (must be clear to all levels)
 - Letter **writer** (matters more at first few levels, recognized expert in the field?)
 - Letter**head** (matters more at higher levels...“peer institution”?)
- Need good writers, peer institutions, insightful comments.
- Chair and Dean will summarize for higher levels.

Out of the box...

- Challenge: How to package “out of the box” success so that those both inside AND outside of the box appreciate the accomplishments.
- Discuss with your chair.
- Discuss with your mentor(s).
- Discuss with Promotions and Tenure committee members.

Evidence of Research Success

- Peer-reviewed, citable publications!
- Authorship, order matter.
- Journal quality matters.

What about...

- Blogs?
- Social media?
- Key question: Are you having an impact? Can you show it?
 - Reposts? Media?
 - Will letter writers notice? Will they comment?
 - Can you impress reviewers?
- Still evolving...

Data Science Twist

- Duncan Temple Lang notes data preparation often takes 80% of a data scientist's time. (NRC Report 2015, *Training Students to Extract Value from Big Data*)
- How to document this effort?
- Key research scholarly products expanded to include:
 - Software
 - Data

Review committees

- Recognize peer review publications
 - Some variation between disciplines
 - Computer science: Conference papers great!
 - Statistics: Conference papers? Peer review journals!
 - Biology: Journals? High impact factor journals!
- Software?
- Data?

Downloads vs. citations

- Parallel to journal publications.
- Downloads = how many people read it (or intended to read it)?
- Citations = how many people used it?

Citation is key, but evolving

- Need to present productivity in forms familiar to reviewers (letter writers and review committees).
- A first step: Link software and data to motivating peer review publication.
 - In **publication list, add note regarding related software** (and download/citation statistics) along with motivating publication, if allowed.
 - **Mention your contribution** to data development (personal statement **and** near citation, if allowed).
 - **Separate section of CV** (“Software”). Discuss metrics of interest with review committee members early (and prepare for changes!). The weakest of the three...
- Other developments...

Software as a Publication

- Some peer-review journals (e.g., *Journal of Statistical Software*).
- The software works and people are using it. Do I *have* to write a paper?
- GitHub as publication?
- Downloads as citations?
- Software is dynamic, but for reproducibility, we need citable versions of software.
- Moving target but some recent developments of note...

2015 NSF Workshop

- *NSF Workshop on Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution*
 - Meeting: January 28, 2015, Final Report: April 20, 2015
 - <https://softwaredatacitation.org/Pages/home.aspx>
- One of three action items: “...the research community develop a primary consistent data and software citation record format (e.g., analogous to BibTex or RIS bibliography formats used in journal publishing) to support D/S citation. Journals and professional societies need to take a more active role in curating citation style files.”

Original Software Publications

- July 2015: In collaboration with GitHub, Elsevier announced a new academic content class: Original Software Publications
 - <http://www.journals.elsevier.com/science-of-computer-programming/call-for-software/a-new-software-track-on-original-software-publications-scico/>
- **“All software and code published is, and will remain, fully owned by their developers.”**
- “All software and code submitted for review and evaluation must be released under a number of pre-approved licenses” (e.g., GPL, Apache-2.0, MIT, etc.)

Data as a Publication

- Data dissemination plan required for most major research grants.
- Post to your or a lab's website?
- Post to public repository (e.g. genetics, imaging)?
- Details in Supplementary Materials?
- Also evolving rapidly...

Citing Data

- GenBank and others.
- DataCite: <https://www.datacite.org/>
- Research Data Alliance: Data Citation Working Group
 - <https://rd-alliance.org/groups/data-citation-wg.html>
- American Geophysical Union
 - <https://agu.confex.com/agu/fm14/meetingapp.cgi/Paper/19292>
- Joint Declaration of Data Citation Principles (2014)
- Statistics?

Joint Declaration 2014

- Joint Declaration of Data Citation Principles (2014)
 - 1. Importance
 - 2. Evidence
 - 3. Unique Identification
 - 4. Access
 - 5. Persistence
 - 6. Specificity and Verifiability
 - 7. Interoperability and Flexibility

- When citing this document please use: Data Citation Synthesis Group: **Joint Declaration of Data Citation Principles**. Martone M. (ed.) San Diego CA: FORCE11; 2014 [<https://www.force11.org/group/joint-declaration-data-citation-principles-final>].

Data as a Publication: Two recent examples

- Dryad (www.datadryad.org/)
 - Abstract, ReadMe.txt, Data in .zip
- *Scientific Data* (www.nature.com/sdata/)
 - Online, open-access, peer-reviewed publication from Nature Publishing Group for descriptions of scientifically valuable datasets.
 - Peer-reviewed content on how the dataset was constructed.
 - Narrative and data.
- Both provide DOIs for data sets.
- Developing citation protocol:
 - Cite original paper (peer review journal).
 - Cite data.

Example: Dryad Citation

- When using this data, please cite the original publication:
 - Yoshimi K, Kumada S, Weitemier A, Jo T, Inoue M (2015) Reward-induced phasic dopamine release in the monkey ventral striatum and putamen. PLOS ONE 10(6): e0130443. <http://dx.doi.org/10.1371/journal.pone.0130443>

- Additionally, please cite the Dryad data package:
 - Yoshimi K, Kumada S, Weitemier A, Jo T, Inoue M (2015) Data from: Reward-induced phasic dopamine release in the monkey ventral striatum and putamen. Dryad Digital Repository. <http://dx.doi.org/10.5061/dryad.r14bv>

Example: *Scientific Data* citation

- *Scientific Data* citation:
 - Roelfsema, C. M. et al. Field data sets for seagrass biophysical properties for the Eastern Banks, Moreton Bay, Australia, 2004–2014. *Sci. Data*. 2:150040 doi: 10.1038/sdata.2015.40 (2015).
 - Author Contributions: Chris Roelfsema, design (70%), methods (70%), field data collection (60%), writing (50%). Eva M. Kovacs, design (10%), methods (10%), field data collection (40%), writing (30%). Stuart R. Phinn, design (20%), methods (20%), field data collection (10%), writing (20%).

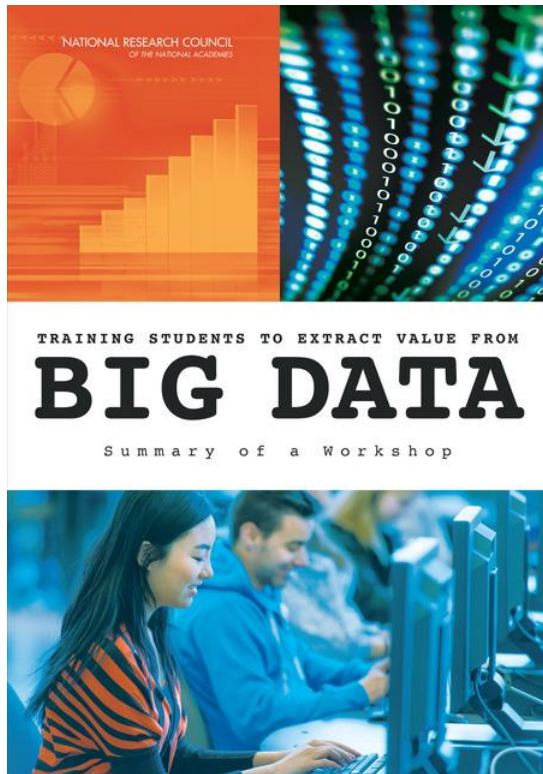
- Data Citation:
 - Roelfsema, C. M., Kovacs, E. M., Lyons, M. & Phinn, S. PANGAEA <http://doi.pangaea.de/10.1594/PANGAEA.846147> (2015).

- **These are in addition to the motivating article.**

Teaching Data Science

- Who are the students?
- What do they want to know?
- What do they need to know?
- Skills vs. core concepts.

Teaching



- National Research Council, Committee on Applied and Theoretical Statistics (CATS)
- *Training Students to Extract Value from Big Data: Summary of a Workshop*
- What do data science jobs require?
- Where/how do we teach it?

Training Opportunities

- www.mastersindatascience.org
 - (Currently) 23 great schools with Masters in Data Science
 - List of skills: Hadoop, Python, R, SQL, Tableau.
 - List of careers: Business Analyst, Data Analyst, Data Architect, Data Engineer, Marketing Analyst, Quantitative Analyst, Statistician.
- New courses in traditional format.
 - Good: Review committees know what to do with this.
 - Challenge: Not the only nor necessarily the most popular approach with instructors and trainees.

Novel teaching modalities

- MOOCs
 - Lots written, some strong opinions, Hopkins program.
- Boot camps
 - Short term, coding principles, set baseline for training.
- Hackathons
 - Weekend “analytic challenge”.
 - Pre-internship, teamwork, focus, short-term results.
 - Long-term impact?
- YouTube tutorials.

Challenge: Documentation

- Enrollees vs. participants vs. completers.
 - Downloads vs. citation all over again.
- Lots of analytics available. Which are compelling and to whom?
- Can/will letter writers comment?
- Be aware of and pre-empt preconceptions of voting faculty, review committee members, higher administrators.

Bringing It All Together: General Principles

- Informative personal statement.
 - Highlight accomplishments.
 - Highlight unique features and define as strengths.
 - Establish goals and clearly identify trajectory.
- Letter writers, Letter content, Letterhead.
- Frame accomplishments as evidence.
 - Novel elements as extensions of standards of evidence.
- Link all together. Chair makes your case.

Key ideas

- Know the rules.
- Know what counts as evidence, and by whom.
- Recognize your own research and teaching productivity.
- Provide context for your scholarly accomplishments (*Scholarship Reconsidered*).
- Think citations. DOI is your friend.
- Discuss with your Chair, early and often.
- Discuss with faculty, early and often.

Questions?

Schema huius præmissæ diuisionis Sphærarum .



Data Scientists vs. Statisticians

- From www.mastersindatascience.org/careers/statistician/ (emphasis added)
- “...a great debate about **whether data science is just statistics, sexed up.**”
- “Those who argue against the “sexing up” theory note that:
 - **Statisticians and Data Analysts are primarily concerned with set tasks.** ... They are given parameters and do their best to collect and analyze information from conventional sources...
 - **Data Scientists think outside the structured box.** They create their own questions/projects and use a **much wider range of tools – only some of which are statistical** – in order to establish unique connections between big data.”
- “Of course, experienced statisticians have been thinking outside the box since the dawn of the field. However, thanks to the surge of technology, **those who wish to call themselves data scientists must now have formidable software engineering, machine learning and predictive analytics skills.**”