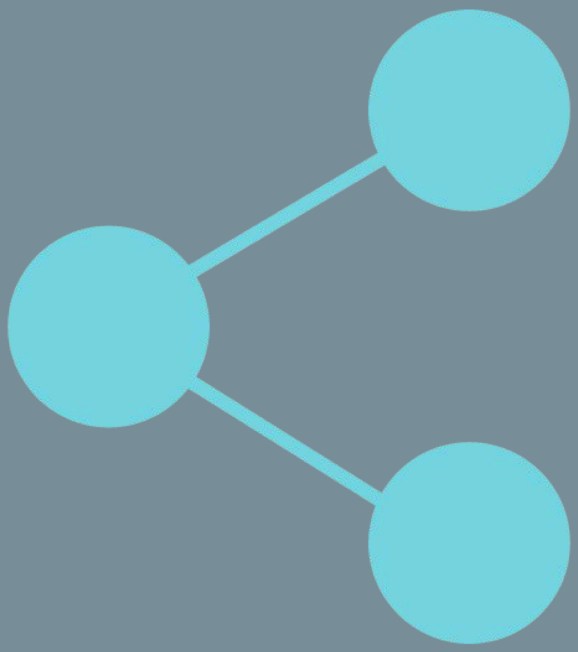
How to be a modern scientist



@jtleek

For

How to be a modern scientist

Jeffrey Leek

This book is for sale at http://leanpub.com/modernscientist

This version was published on 2016-04-08

Also By Jeffrey Leek

The Elements of Data Analytic Style

Contents

Introduction

This book is an opinionated guide to being a scientist using modern internet-enabled research, teaching, publishing, and communication tools. The primary audience for this book is students, postdocs, and faculty members in academic science roles. There are a now a large number of tools that are available to academic scientists that can speed and improve all aspects of a scientific career. As with many things, while the future is already here, it is not evenly distributed. I hope this book will smooth that out a little.

In this book I will cover a range of tools that can be used to build a modern scientific career. Many of these tools overlap with the open-science, reproducible research, and scientific communication communities. My goal with this book was not to cater to any one of these specific communities, but to the practicing scientist who wants to take advantage of modern technologies but whose research program may not focus on any specific scientific revolution.

The traditional academic scientist worked on research, submitted papers to peer reviewed journals, reviewed for those same journals, and was judged solely on their research productivity. Research productivity was usually defined as the number and quality of peer reviewed papers. Few other academic outputs carried much weight. That is slowly changing. Increasingly, scientific communication, good software, reproducible research, open science, and other non-traditional outputs are being recognized. As this happens, some of the tools and approaches discussed in this book will move from interesting and tangentially useful, to standard and required. But almost none have made that leap so far.

The tools that enable this book to be written have all arisen over the last two decades, with many of the tools becoming available only in the last several years. The nice thing about this rapid technological revolution in academic science is that there are still relatively few codified guidelines for how to use them. I'm going to explain some ways I have used the tools, but keep in mind that in every area I will discuss both the technologies and their use is rapidly evolving.

The modern academic scientist develops code in the open, publishes data and code open source, posts preprints of their academic work, still submits to traditional journals, and reviews for those journals, but may also write blog posts or use social media to critique published work in post-publication review fora. These activities can dramatically increase the profile of scientists, particularly junior scientists, if done well. But their value for important career milestones such as promotion and tenure or getting grants, is still often muted or fuzzy.

In this book I will cover a range of topics from writing, to publishing, social media to teaching. In each chapter there are three sections:

1. What should I do and why?
2. What tools should I use?

3. Further tips and issues

The point is to point you toward tools and ideas that will help you keep up with the modern scientific world without getting in your way. I will devote some energy to discussing the potential tradeoffs for junior scientists and ways to leverage increasingly limited time for non-traditional activities into maximum benefit.

There is a quote[1] is from an article in the Chronicle Review. I highly recommend reading the article, particularly check out the section on the author's "Uncreative writing" class at UPenn. The article is about how there is a trend in literature toward combining/using other people's words to create new content.

> The prominent literary critic Marjorie Perloff has recently begun using the term "unoriginal genius" to describe this tendency emerging in literature. Her idea is that, because of changes brought on by technology and the Internet, our notion of the genius—a romantic, isolated figure—is outdated. An updated notion of genius would have to center around one's mastery of information and its dissemination. Perloff has coined another term, "moving information," to signify both the act of pushing language around as well as the act of being emotionally moved by that process. She posits that today's writer resembles more a programmer than a tortured genius, brilliantly conceptualizing, constructing, executing, and maintaining a writing machine.

It is fascinating to see this happening in the world of literature; a similar trend seems to be happening in science. A ton of exciting and interesting work is done by people combining known ideas and tools and applying them to new problems. The modern scientist is "creative" in a whole new way - opening the doors to a scientific career that few could have envisioned only a few years ago. I hope this book will be a start for you to begin exploring modern science.

Note: Part of this chapter appeared in the Simply Statistics blog post "Unoriginal genius"[2].

[1] http://chronicle.com/article/Uncreative-Writing/128908/

[2] http://simplystatistics.org/2011/09/26/unoriginal-genius/

Paper writing

Writing - what should I do and why?

Write using collaborative software to avoid version control issues.

On almost all modern scientific papers you will have co-authors. The traditional way of handling this was to create a single working document and pass it around. Unfortunately this system always results in a long collection of versions of a manuscript, which are often hard to distinguish and definitely hard to synthesize.

An alternative approach is to use formal version control systems like Git[3] and Github[4]. However, the overhead for using these systems can be pretty heavy for paper authoring. They also require all parties participating in the writing of the paper to be familiar with version control and the command line. Alternative paper authoring tools are now available that provide some of the advantages of version control without the major overhead involved in using base version control systems.

The usual result of file naming by a group (image via https://xkcd.com/1459/)

Make figures the focus of your writing

[3] https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

[4] https://github.com/

Once you have a set of results and are ready to start writing up the paper the first thing is *not to write*. The first thing you should do is create a set of 1-10 publication-quality plots with 3-4 as the central focus (see Chapter 10 here[5] for more information on how to make plots). Show these to someone you trust to make sure they "get" your story before proceeding. Your writing should then be focused around explaining the story of those plots to your audience. Many people, when reading papers, read the title, the abstract, and then usually jump to the figures. If your figures tell the whole story you will dramatically increase your audience. It also helps you to clarify what you are writing about.

Write clearly and simply even though it may make your papers harder to publish.

Learn how to write papers in a very clear and simple style. Whenever you can write in plain English and make the approach you are using understandable and clear. This can (sometimes) make it harder to get your papers into journals. Referees are trained to find things to criticize and by simplifying your argument they will assume that what you have done is easy or just like what has been done before. This can be extremely frustrating during the peer review process. But the peer review process isn't the end goal of publishing! The point of publishing is to communicate your results to your community and beyond so they can use them. Simple, clear language leads to much higher use/reading/citation/impact of your work.

Include links to code, data, and software in your writing

Not everyone recognizes the value of re-analysis, scientific software, or data and code sharing. But it is the fundamental cornerstone of the modern scientific process to make all of your materials easily accessible, re-usable and checkable. Include links to data, code, and software prominently in your abstract, introduction and methods and you will dramatically increase the use and impact of your work.

Give credit to others

In academics the main currency we use is credit for publication. In general assigning authorship and getting credit can be a very tricky component of the publication process. It is almost always better to err on the side of offering credit. A very useful test that my advisor John Storey[6] taught me is if you are embarrassed to explain the authorship credit to anyone who was on the paper or not on the paper, then you probably haven't shared enough credit.

Writing - what tools should I use?

WYSIWYG software: Google Docs and Paperpile.

This system uses Google Docs[7] for writing and Paperpile[8] for reference management. If you have a Google account you can easily create documents and share them with your collaborators either

[5] http://leanpub.com/datastyle
[6] http://www.genomine.org/
[7] https://www.google.com/docs/about/
[8] https://paperpile.com/app

privately or publicly. Paperpile allows you to search for academic articles and insert references into the text using a system that will be familiar if you have previously used Endnote[9] and Microsoft Word[10].

This system has the advantage of being a what you see is what you get system - anyone with basic text processing skills should be immediately able to contribute. Google Docs also automatically saves versions of your work so that you can flip back to older versions if someone makes a mistake. You can also easily see which part of the document was written by which person and add comments.

Getting started

1. Set up accounts with Google[11] and with Paperpile[12]. If you are an academic the Paperpile account will cost $2.99 a month, but there is a 30 day free trial.
2. Go to Google Docs[13] and create a new document.
3. Set up the Paperpile add-on for Google Docs[14]
4. In the upper right hand corner of the document, click on the *Share* link and share the document with your collaborators
5. Start editing
6. When you want to include a reference, place the cursor where you want the reference to go, then using the *Paperpile* menu, choose insert citation. This should give you a search box where you can search by Pubmed ID or on the web for the reference you want.
7. Once you have added some references use the *Citation style* option under *Paperpile* to pick the citation style for the journal you care about.
8. Then use the *Format citations* option under *Paperpile* to create the bibliography at the end of the document

The nice thing about using this system is that everyone can easily directly edit the document simultaneously - which reduces conflict and difficulty of use. A disadvantage is getting the formatting just right for most journals is nearly impossible, so you will be sending in a version of your paper that is somewhat generic. For most journals this isn't a problem, but a few journals are sticklers about this.

Typesetting software: Overleaf or ShareLatex

An alternative approach is to use typesetting software like Latex. This requires a little bit more technical expertise since you need to understand the Latex typesetting language. But it allows for more precise control over what the document will look like. Using Latex on its own you will have

[9] http://endnote.com/
[10] https://products.office.com/en-us/word
[11] https://accounts.google.com/SignUp
[12] https://paperpile.com/
[13] https://docs.google.com/document/u/0/
[14] https://paperpile.com/blog/free-google-docs-add-on/

many of the same issues with version control that you would have for a word document. Fortunately there are now "Google Docs like" solutions for editing latex code that are readily available. Two of the most popular are Overleaf[15] and ShareLatex[16].

In either system you can create a document, share it with collaborators, and edit it in a similar manner to a Google Doc, with simultaneous editing. Under both systems you can save versions of your document easily as you move along so you can quickly return to older versions if mistakes are made.

I have used both kinds of software, but now primarily use Overleaf because they have a killer feature. Once you have finished writing your paper you can directly submit it to some preprint servers like arXiv[17] or biorXiv[18] and even some journals like Peerj[19] or f1000research[20].

Getting started

1. Create an Overleaf account[21]. There is a free version of the software. Paying $8/month will give you easy saving to Dropbox.
2. Click on *New Project* to create a new document and select from the available templates
3. Open your document and start editing
4. Share with colleagues by clicking on the *Share* button within the project. You can share either a read only version or a read and edit version.

Once you have finished writing your document you can click on the *Publish* button to automatically submit your paper to the available preprint servers and journals. Or you can download a pdf version of your document and submit it to any other journal.

Writing - further tips and issues

When to write your first paper

As soon as possible! The purpose of graduate school is (in some order):

- Freedom
- Time to discover new knowledge
- Time to dive deep
- Opportunity for leadership

[15] https://www.overleaf.com/
[16] https://www.sharelatex.com/
[17] http://arxiv.org/
[18] http://biorxiv.org/
[19] https://peerj.com
[20] http://f1000research.com/
[21] https://www.overleaf.com/signup

- Opportunity to make a name for yourself
  - R packages
  - Papers
  - Blogs
- Get a job

The first couple of years of graduate school are typically focused on (1) teaching you all the technical skills you need and (2) data dumping as much hard-won practical experience from more experienced people into your head as fast as possible.

After that one of your main focuses should be on establishing your own program of research and reputation. Especially for Ph.D. students it can not be emphasized enough *no one will care about your grades in graduate school but everyone will care what you produced.* See for example, Sherri's excellent guide on CV's for academic positions[22].

I firmly believe that R packages[23] and blog posts can be just as important as papers, but the primary signal to most traditional academic communities still remains published peer-reviewed papers. So you should get started on writing them as soon as you can (definitely before you feel comfortable enough to try to write one).

Even if you aren't going to be in academics, papers are a great way to show off that you can (a) identify a useful project, (b) finish a project, and (c) write well. So the first thing you should be asking when you start a project is "what paper are we working on?"

What is an academic paper?

A scientific paper can be distilled into four parts:

1. A set of methodologies
2. A description of data
3. A set of results
4. A set of claims

When you (or anyone else) writes a paper the goal is to communicate clearly items 1-3 so that they can justify the set of claims you are making. Before you can even write down 4 you have to do 1-3. So that is where you start when writing a paper.

[22] http://drsherrirose.com/academic-cvs-for-statistical-science-faculty-positions

[23] http://simplystatistics.org/2013/01/23/statisticians-and-computer-scientists-if-there-is-no-code-there-is-no-paper/

How do you start a paper?

The first thing you do is you decide on a problem to work on. This can be a problem that your advisor thought of or it can be a problem you are interested in, or a combination of both. Ideally your first project will have the following characteristics:

1. Concrete
2. Solves a scientific problem
3. Gives you an opportunity to learn something new
4. Something you feel ownership of
5. Something you want to work on

Points 4 and 5 can't be emphasized enough. Others can try to help you come up with a problem, but if you don't feel like it is *your* problem it will make writing the first paper a total slog. You want to find an option where you are just insanely curious to know the answer at the end, to the point where you *just have to figure it out* and kind of don't care what the answer is. That doesn't always happen, but that makes the grind of writing papers go down a lot easier.

Once you have a problem the next step is to actually do the research. I'll leave this for another guide, but the basic idea is that you want to follow the usual data analytic process[24]:

1. Define the question
2. Get/tidy the data
3. Explore the data
4. Build/borrow a model
5. Perform the analysis
6. Check/critique results
7. Write things up

The hardest part for the first paper is often knowing where to stop and start writing.

How do you know when to start writing?

Sometimes this is an easy question to answer. If you started with a very concrete question at the beginning then once you have done enough analysis to convince yourself that you have the answer to the question. If the answer to the question is interesting/surprising then it is time to stop and write.

If you started with a question that wasn't so concrete then it gets a little trickier. The basic idea here is that you have convinced yourself you have a result that is worth reporting. Usually this takes the

[24] https://leanpub.com/datastyle/

form of between 1 and 5 figures that show a coherent story that you could explain to someone in your field.

In general one thing you should be working on in graduate school is your own internal timer that tells you, "ok we have done enough, time to write this up". I found this one of the hardest things to learn, but if you are going to stay in academics it is a critical skill. There are rarely deadlines for paper writing (unless you are submitting to CS conferences) so it will eventually be up to you when to start writing. If you don't have a good clock, this can really slow down your ability to get things published and promoted in academics.

One good principle to keep in mind is "the perfect is the enemy of the very good" Another one is that a published paper in a respectable journal beats a paper you just never submit because you want to get it into the "best" journal.

A note on "negative results"

If the answer to your research problem isn't interesting/surprising but you started with a concrete question it is also time to stop and write. But things often get more tricky with this type of paper as most journals when reviewing papers filter for "interest" so sometimes a paper without a really "big" result will be harder to publish. **This is ok!!** Even though it may take longer to publish the paper, it is important to publish even results that aren't surprising/novel. I would much rather that you come to an answer you are comfortable with and we go through a little pain trying to get it published than you keep pushing until you get an "interesting" result, which may or may not be justifiable.

How do you start writing?

1. Once you have a set of results and are ready to start writing up the paper the first thing is *not to write*. The first thing you should do is create a set of 1-4 publication-quality plots (see Chapter 10 here[25]). Show these to someone you trust to make sure they "get" your story before proceeding.
2. Start a project on Overleaf[26] or Google Docs[27].
3. Write up a story around the four plots in the simplest language you feel you can get away with, while still reporting all of the technical details that you can.
4. Go back and add references in only after you have finished the whole first draft.
5. Add in additional technical detail in the supplementary material if you need it.
6. Write up a reproducible version of your code that returns exactly the same numbers/figures in your paper with no input parameters needed.

[25] http://leanpub.com/datastyle
[26] https://www.overleaf.com/
[27] https://www.google.com/docs/about/

What are the sections in a paper?

Keep in mind that most people will read the title of your paper only, a small fraction of those people will read the abstract, a small fraction of those people will read the introduction, and a small fraction of those people will read your whole paper. So make sure you get to the point quickly!

The sections of a paper are always some variation on the following:

1. **Title**: Should be very short, no colons if possible, and state the main result. Example, "A new method for sequencing data that shows how to cure cancer". Here you want to make sure people will read the paper without overselling your results - this is a delicate balance.
2. **Abstract**: In (ideally) 4-5 sentences explain (a) what problem you are solving, (b) why people should care, (c) how you solved the problem, (d) what are the results and (e) a link to any data/resources/software you generated.
3. **Introduction**: A more lengthy (1-3 pages) explanation of the problem you are solving, why people should care, and how you are solving it. Here you also review what other people have done in the area. The most critical thing is never underestimate how little people know or care about what you are working on. It is your job to explain to them why they should.
4. **Methods**: You should state and explain your experimental procedures, how you collected results, your statistical model, and any strengths or weaknesses of your proposed approach.
5. **Comparisons (for methods papers)**: Compare your proposed approach to the state of the art methods. Do this with simulations (where you know the right answer) and data you haven't simulated (where you don't know the right answer). If you can base your simulation on data, even better. Make sure you are simulating both the easy case (where your method should be great) and harder cases where your method might be terrible[28].
6. **Your analysis**: Explain what you did, what data you collected, how you processed it and how you analysed it.
7. **Conclusions**: Summarize what you did and explain why what you did is important one more time.
8. **Supplementary Information**: If there are a lot of technical computational, experimental or statistical details, you can include a supplement that has all of the details so folks can follow along. As far as possible, try to include the detail in the main text but explained clearly.

The length of the paper will depend a lot on which journal you are targeting. In general the shorter/more concise the better. But unless you are shooting for a really glossy journal you should try to include the details in the paper itself. This means most papers will be in the 4-15 page range, but with a huge variance.

Note: Part of this chapter appeared in the Leek group guide to writing your first paper[29]

[28] http://simplystatistics.org/2013/03/06/the-importance-of-simulating-the-extremes/
[29] https://github.com/jtleek/firstpaper

Publishing

Publishing - what should I do and why?

A modern scientific writing process goes as follows.

1. You write a paper
2. You post a preprint
    a. Everyone can read and comment
3. You submit it to a journal
4. It is peer reviewed privately
5. The paper is accepted or rejected
    a. If rejected go back to step 2 and start over
    b. If accepted it will be published

You can take advantage of modern writing and publishing tools to handle several steps in the process.

Post preprints of your work

Once you have finished writing you paper, you want to share it with others. Historically, this involved submitting the paper to a journal, waiting for reviews, revising the paper, resubmitting, and eventually publishing it. There is now very little reason to wait that long for your paper to appear in print. Generally you can post a paper to a preprint server and have it appear in 1-2 days. This is a dramatic improvement on the weeks or months it takes for papers to appear in peer reviewed journals even under optimal conditions. There are several advantages to posting preprints.

- Preprints establish precedence for your work so it reduces your risk of being scooped.
- Preprints allow you to collect feedback on your work and improve it quickly.
- Preprints can help you to get your work published in formal academic journals.
- Preprints can get you attention and press for your work.
- Preprints give junior scientists and other researchers gratification that helps them handle the stress and pressure of their first publications.

The last point is underappreciated and was first pointed out to me by Yoav Gilad[30] It takes a really long time to write a scientific paper. For a student publishing their first paper, the first feedback they get is often (a) delayed by several months and (b) negative and in the form of a referee report. This

[30] http://giladlab.uchicago.edu/

can have a major impact on the motivation of those students to keep working on projects. Preprints allow students to have an immediate product they can point to as an accomplishment, allow them to get positive feedback along with constructive or negative feedback, and can ease the pain of difficult referee reports or rejections.

Choose the journal that maximizes your visibility

You should try to publish your work in the best journals for your field. There are a couple of reasons for this. First, being a scientist is both a calling and a career. To advance your career, you need visibilty among your scientific peers and among the scientists who will be judging you for grants and promotions. The best place to do this is by publishing in the top journals in your field. The important thing is to do your best to do good work and submit it to these journals, even if the results aren't the most "sexy". Don't adapt your workflow to the journal, but don't ignore the career implications either. Do this even if the journals are closed source. There are ways to make your work accessible and you will both raise your profile and disseminate your results to the broadest audience.

Share your work on social media

Academic journals are good for disseminating your work to the appropriate scientific community. As a modern scientist you have other avenues and other communities - like the general public - that you would like to reach with your work. Once your paper has been published in a preprint or in a journal, be sure to share your work through appropriate social media channels. This will also help you develop facility in coming up with one line or one figure that best describes what you think you have published so you can share it on social media sites like Twitter.

Preprints and criticism

See the section on scientific blogging for how to respond to criticism of your preprints online.

Publishing - what tools should I use?

Preprint servers

Here are a few preprint servers you can use.

1. arXiv[31] (free) - primarily takes math/physics/computer science papers. You can submit papers and they are reviewed and posted within a couple of days. It is important to note that once you submit a paper here, you can not take it down. But you can submit revisions to the paper which are tracked over time. This outlet is followed by a large number of journalists and scientists.

[31] http://arxiv.org/

2. biorXiv[32] (free) - primarily takes biology focused papers. They are pretty strict about which categories you can submit to. You can submit papers and they are reviewed and posted within a couple of days. biorxiv also allows different versions of manuscripts, but some folks have had trouble with their versioning system, which can be a bit tricky for keeping your paper coordinated with your publication. bioXiv is pretty carefully followed by the biological and computational biology communities.
3. Peerj[33] (free) - takes a wide range of different types of papers. They will again review your preprint quickly and post it online. You can also post different versions of your manuscript with this system. This system is newer and so has fewer followers, you will need to do your own publicity if you publish your paper here.

Journal preprint policies

This list[34] provides information on which journals accept papers that were first posted as preprints. However, you shouldn't

Publishing - further tips and issues

Open vs. closed access

Once your paper has been posted to a preprint server you need to submit it for publication. There are a number of considerations you should keep in mind when submitting papers. One of these considerations is closed versus open access. Closed access journals do not require you to pay to submit or publish your paper. But then people who want to read your paper either need to pay or have a subscription to the journal in question.

There has been a strong push for open access journals over the last couple of decades. There are some very good reasons justifying this type of publishing including (a) moral arguments based on using public funding for research, (2) each of access to papers, and (3) benefits in terms of people being able to use your research. In general, most modern scientists want their work to be as widely accessible as possible. So modern scientists often opt for open access publishing.

Open access publishing does have a couple of disadvantages. First it is often expensive, with fees for publication ranging between $1,000 and $4,000[35] depending on the journal. Second, while science is often a calling, it is also a career. Sometimes the best journals in your field may be closed access. In general, one of the most important components of an academic career is being able to publish in journals that are read by a lot of people in your field so your work will be recognized and impactful.

However, modern systems make both closed and open access journals reasonable outlets.

[32] http://biorxiv.org/
[33] https://peerj.com/preprints/
[34] https://en.wikipedia.org/wiki/List_of_academic_journals_by_preprint_policy
[35] http://simplystatistics.org/2011/11/03/free-access-publishing-is-awesome-but-expensive-how/

Closed access + preprints

If the top journals in your field are closed access and you are a junior scientist then you should try to submit your papers there. But to make sure your papers are still widely accessible you can use preprints. First, you can submit a preprint before you submit the paper to the journal. Second, you can update the preprint to keep it current with the published version of your paper. This system allows you to make sure that your paper is read widely within your field, but also allows everyone to freely read the same paper. On your website, you can then link to both the published and preprint version of your paper.

Open access

If the top journal in your field is open access you can submit directly to that journal. Even if the journal is open access it makes sense to submit the paper as a preprint during the review process. You can then keep the preprint up-to-date, but this system has the advantage that the formally published version of your paper is also available for everyone to read without constraints.

Responding to referee comments

After your paper has been reviewed at an academic journal you will receive referee reports. If the paper has not been outright rejected, it is important to respond to the referee reports in a timely and direct manner. Referee reports are often maddening. There is little incentive for people to do a good job refereeing and the most qualified reviewers will likely be those with a conflict of interest[36].

The first thing to keep in mind is that the power in the refereeing process lies entirely with the editors and referees. The first thing to do when responding to referee reports is to eliminate the impulse to argue or respond with any kind of emotion. A step-by-step process for responding to referee reports is the following.

1. Create a Google Doc. Put in all referee and editor comments in italics.
2. Break the comments up into each discrete criticism or request.
3. In bold respond to each comment. Begin each response with "On page xx we did yy to address this comment"
4. Perform the analyses and experiments that you need to fill in the yy
5. Edit the document to reflect all of the experiments that you have performed

By actively responding to each comment you will ensure you are responsive to the referees and give your paper the best chance of success. If a comment is incorrect or non-sensical, think about how you can edit the paper to remove this confusion.

[36] http://simplystatistics.org/2015/02/09/the-trouble-with-evaluating-anything/

Finishing

While I have advocated here for using preprints to disseminate your work, it is important to follow the process all the way through to completion. Responding to referee reports is drudgery and no one likes to do it. But in terms of career advancement preprints are almost entirely valueless until they are formally accepted for publication. It is critical to see all papers all the way through to the end of the publication cycle.

You aren't done!

Publication of your paper is only the beginning of successfully disseminating your science. Once you have published the paper, it is important to use your social media, blog, and other resources to disseminate your results to the broadest audience possible. You will also give talks, discuss the paper with colleagues, and respond to requests for data and code. The most successful papers have a long half life and the responsibilities linger long after the paper is published. But the most successful scientists continue to stay on top of requests and respond to critiques long after their papers are published.

Note: Part of this chapter appeared in the Simply Statistics blog post: "Preprints are great, but post publication peer review isn't ready for prime time"[37]

[37] http://simplystatistics.org/2016/02/26/preprints-and-pppr/

Peer review

Peer review - what should I do and why?

If you work in academia you will be asked to referee papers. Your first review will almost certainly be something you are asked to do by your advisor. But as your career goes on and you become recognized as an expert in one (or many!) areas, you will have the ahem…opportunity to review a lot more.

No one ever gave me formal guidelines for writing reviews. I think this experience is pretty similar to most graduate students. The format, tone, and content of these documents is usually learned via an apprenticeship model (if at all). Peer review is also undergoing a revolution as we speak and there are some options that weren't available previously.

Review papers quickly or not at all

The first few times you get asked to peer review a paper you will probably want to do it to build reputation within your community or to help out a mentor. Over time an important strategy is participating actively in peer review without being overwhelmed. Since a typical paper is reviewed by 2-3 people, a good rule of thumb is to review no more than 2-3 times the number of papers that you submit. One very good system is to ask yourself if you can review the paper the week you get the request. If you don't have time that week, politely turn it down as quickly as you can so the associate editor can move on.

Get credit for your reviews

There are two ways to get credit in the peer review system. One is informal - when you review promptly and well editors will remember and potentially be more receptive to your papers when you submit them to that journal. Another, more recent model, is to get credit for your reviews by either posting them openly or submitting them to systems like Publons[38] that aggregate your reviews into a single system.

Review openly post publication or pre-publication

If you find a paper really interesting you should consider writing up your thoughts and posting them to a blog, to a pre-print server, or online. Your comments could be something as simple as what you thought about the paper or as complicated as a re-analysis or re-done experiment. In any case, when posting comments online, be sure to follow the same guidelines of civility and professionalism that you would follow in any formal peer review to avoid accusations of bias or inappropriate behavior.

[38] https://publons.com

Peer review - what tools should I use?

Reviewing template

You can find a template for writing reviews here : reviewing template[39]. More tips for using this template appear in the "Further tips and issues" section below.

Public peer review

As of right now there are a few options for performing reviews online.

- *On a personal blog (free)* - if you have a widely read blog or a reasonable Twitter following this option may be the best for quickly getting your opinion about a scientific paper out in the open. You get very little academic credit for this type of critique but can influence science.
- *Some journals have the option of comments (free)* - on many journal websites you can post a comment about a paper. These are rarely read by scientists, so you will need to promote your review using social media or other means. You get very little academic credit for this type of critique but can influence science.
- *As a comment on Pubmed Commons* - Pubmed Commons[40] is an attempt to aggregate comments on biomedical papers. You need a Pubmed Account to write comments, and you will again need to promote them to gain visibility. You get very little academic credit for this type of critique but can influence science.
- *As an independent paper (publication charges apply)* - if your review is meaty enough to be a paper unto itself, the most credit you can get is probably by publishing it on a platform like f1000research[41]. You may get academic credit for this type of critique, but you will still need to promote it yourself. This format has been used successfully to critique published studies, primarily for pointing our artifacts[42] that the authors missed[43] via re-analysis.

Credit for peer review

- Publons[44] and other companies are trying to develop systems for giving credit for pre and post publication review. If you post a comment on Pubmed Commons, you automatically get this type of credit, but can also do it directly through the site. You can forward any email you get thanking you for a review to Publons and they will verify it and add it to your profile, so you can collect information on your reviewing activities and use it when going up for promotion.

[39] https://raw.githubusercontent.com/jtleek/reviews/master/review-template.Rmd
[40] http://www.ncbi.nlm.nih.gov/pubmedcommons/
[41] http://f1000research.com/
[42] http://f1000research.com/articles/4-121/v1
[43] http://f1000research.com/articles/4-180/v1
[44] https://publons.com/

Peer review - further tips and issues

The key players in peer review

There are a few different players in the peer review process.

The first is the **editor** of the journal, who will do some vetting of papers at the beginning - mostly to screen out really crazy stuff that gets submitted (you'd be amazed at what gets submitted). At most journals, the editor is a senior scientist with a broad knowledge of the field who has a pretty good intuition about what is likely to be interesting to the readership of the journal and what is likely real science. Papers that are uninteresting or obviously wrong usually won't get past the editor.

The editor usually assigns the paper to an **associate editor** who has more expertise in the topics covered in the paper. The associate editor is usually a mid-level faculty member (senior assistant to associate professor). Again, papers that are obviously flawed or make wild claims often don't make it past the associate editor.

If a paper passes these hurdles it does not mean that it is correct or the claims are justified. It only means that on a quick read the paper seems interesting and not outlandish. The associate editor then makes an effort to find **referees** who work in the specific area the paper focuses on but do not have strong conflicts or collaborations with the authors of the paper. As you can imagine, in some areas of science it is hard to find referees that aren't in conflict one way or the other and have time to review. So sometimes they have to find people whose expertise is close, but not perfectly aligned with the theme of the paper.

What is your job in peer review?

A scientific paper can be distilled into four parts:

1. A set of methodologies
2. A description of data
3. A set of results
4. A set of claims

When you (or anyone else) writes a paper the goal is to communicate clearly items 1-3 so that they can justify the set of claims they are making. In the current peer review system there are three tasks you are responsible for as a peer reviewer:

1. Evaluating the quality and accuracy of the methods, data, and results.
2. Determining whether the methods, data and results justify the claims.
3. Determining how important the claims are and whether they belong in the journal the paper was submitted to.

Ideally you would be able to verify every single claim in the paper and test every result. Given the time constraints of the review process and the remuneration you get for reviewing, this is absolutely not feasible. Your goal is instead to do your best to obtain reasonable estimators of each of the three components of the review and "show your work" by providing references, pointing to figures, and putting results in context.

Your prior belief about 1-3 should start with the assumption that the scientists in question are reasonable people who made efforts to be correct, thorough, transparent, and not exaggerate. You may adjust your prior if the paper was submitted to a journal that you and your colleagues have never heard of, or if you are asked to review a paper far outside of your expertise (legit journals try not to do this), or if the claims are so extreme that they would overturn huge areas of science (e.g. a paper claiming to prove evolution isn't true or that vaccines cause autism), or if the journal has the intent of only publishing papers that are groundbreaking.

Structure of a review

Your review will have three parts. The comments to the authors, the comments to the editor, and a recommendation.

Comments to the authors

When you write a review the first part consists of the comments to the authors; it should have the following components:

- A summary of the paper (motivation, methods, results) written in your own words of about a paragraph.
- A list of major issues
- A list of minor issues
- A list of typos you find

I think the summary is critical because if you can't distill the ideas down then you haven't really understood the paper. The summary should absolutely *not* be a restatement of the abstract of the paper, you should find the parts you think are most relevant and include them in the summary.

The major issues should be a bulleted list. Depending on the quality of the paper this list may be longer or shorter. A major issue has to be one of the following: (1) a claim that is not supported by the data, (2) a method or result that appears completely incorrect, (3) a critical missing piece of information, or (4) a paper that is not readable by a person trying their best to understand it. You should point to specific figures, paragraphs, or results for each major issue and be concrete about the problems. Vague criticisms are unacceptable. When possible, use references from previous literature to back up your claims.

Not making all data and code available with a specific link and instructions, is a major issue.

Minor issues should also be a bulleted list. There are a much broader range of minor issues that you may encounter. Some examples include simulations that miss some cases, figures that are missing axis labels, or the paper has extraneous results that aren't relevant to the claims being made.

Typos are not minor issues or major issues. It is not your responsibility to find them. If you do, you should provide them as a bulleted list to the author in the format: "On page x, line z, change … to …".

If there are a huge number of typos then that may be stated as a minor issue. If the paper is completely unreadable then that is a major issue. Completely unreadable means you could not follow the paper even after ignoring all typos.

Here are some things that your comments to the authors should not contain:

- A recommendation of whether to accept or reject the paper
- Requests for citations to a bunch of your papers (this will matter more later in your career)
- Requests for experiments/simulations that are unnecessary to justify the main points in the paper
- Insulting criticism or sarcasm

Remember that this is a professional document. They are typically anonymous (you don't have to sign your name) but the associate editor and editor will see the review and your reputation will be affected by the quality of the work you perform. There is no reason to be rude, competitive, or snarky in a review.

Comments to the editor

If you think you have covered everything in your comments to the authors you may leave this field blank. If you do put any text in, it should be no more than one paragraph. It should not contain any criticism of methods/results that you did not put in your comments to the authors. It may include a statement of how interesting you think the paper is and how appropriate it is for the journal readership if it helps justify your decision. It should be very consistent with the comments for the authors; if you aren't comfortable saying it to the authors directly, you should consider carefully whether it should be said.

Recommendation

You usually have these four options for the decision

- Reject
- Major revisions
- Minor revisions
- Accept

Reject if you think that the methods, results, or claims are blatantly false. Reject if you think the paper has major flaws that could not be corrected. Reject if the paper is clearly not an improvement on the current state of the art. This third category is very hard to judge if you don't have a lot of experience in the field. If you are new to reviewing you should consult your advisor.

You should decide major revisions if you think there are serious problems with the paper but that they can be corrected. If you ask for major revisions your default plan should be that if they can/do correct all of the major issues you pointed out, you would be prepared to accept the paper. Sometimes, in the course of performing the corrections, they will show that their method/results/claims are not actually true. Then you should reject.

Do not ask for major revisions if you think the paper is uninteresting and you wouldn't accept it even if they did everything you said. This is the #1 way to be a jerk reviewer.

You should ask for minor revisions if there are only minor issues with the paper that you are pretty sure the authors can correct and you would be prepared to accept if the authors address those issues.

Do not ask for minor revisions if you think the paper is uninteresting and you wouldn't accept it even if they did everything you said. This is the #1 way to be a jerk reviewer.

It is very atypical for a reviewer to accept a paper outright. However there will be times when you receive a paper that has only minor issues and those issues are only judgement calls on your part, as opposed to things that need to be fixed to justify the claims or to make methods/results/data clear. It is perfectly acceptable in this case to list the minor issues and to suggest acceptance.

Length of a review

The best reviews are bullet pointed, brief, and point out exactly the key issues and nothing more. It is *absolutely not* your responsibility to rewrite the paper, change the message of the paper, or make the authors do something that wasn't in the scope of the original work. If you think the paper isn't appropriate for a journal in its current form, you should explain/justify why and choose reject. But you should not make the authors conform to your opinions.

There is a temptation to write really long reviews to show that you read a paper carefully and show off how expert you are. Do not succumb to this temptation. You get no bonus points for being nitpicky, verbose, or long.

You get big bonus points for the following things: * Being concise - nothing extraneous * Being precise - stating the specific problems with the manuscript * Being constructive - stating how the authors could address the problems you have found * Being polite - this helps focus on real issues rather than pet peeves.

Very good reviews are often 1-2 pages long in bullet pointed format.

Re-review

Unless the paper was outright rejected or accepted, the authors will have a chance to respond to your review. If you have followed the guidelines above, it should make the re-review process more

straightforward:

- If you said minor revision and they addressed your minor issues - accept.
- If you said major revisions and they addressed all your major/minor issues - accept.
- If you said major revisions and they didn't do what you asked - major revisions with the outstanding issues.
- If you said major revisions and their revision showed their method was incorrect/uninteresting - reject.

How long should it take you to review?

You should never take more than a month to review. If you accept a review, you should plan to complete it within a month. Ideally you will do the review in less time than that (think 2 weeks). If you are unable to do the review in that time frame you should politely decline and offer some alternative reviewers.

It is inevitable that you will miss some deadlines for reviews. It is an important component of an academic's professional life but it is not a priority above your own work. If you are going to miss a deadline, you should let the associate editor know and give them a time frame for when you will complete the review.

Remember that someone else put a huge amount of work into this paper and their career/livelihood depends on them getting papers published in a reasonable amount of time. If you think the paper should be rejected, do it quickly! If you think it should be accepted, do it quickly!

Post publication review

Lately there is a push for post-publication review. I used to argue pretty strongly for post-publication peer review[45] but Rafa set me straight[46] and pointed out that at least with peer review every paper that gets submitted gets evaluated by *someone* even if the paper is ultimately rejected.

One of the risks of post publication peer review is that there is no incentive to peer review in the current system. In a paper a few years ago I actually showed that under an economic model for closed peer review the Nash equilibrium is for no one to peer review at all[47]. We showed in that same paper that under open peer review there is an increase in the amount of time spent reviewing, but the effect was relatively small. Moreover the dangers of open peer review are clear (junior people reviewing senior people and being punished for it) while the benefits (potentially being recognized for insightful reviews) are much hazier. Even the most vocal proponents of post publication peer review don't do it that often[48] when given the chance.

[45] http://simplystatistics.org/2012/10/04/should-we-stop-publishing-peer-reviewed-papers/
[46] http://simplystatistics.org/2012/10/08/why-we-should-continue-publishing-peer-reviewed-papers/
[47] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0026895
[48] http://www.ncbi.nlm.nih.gov/myncbi/michael.eisen.1/comments/

The reason is that everyone in academics already have a lot of things they are asked to do. Many review papers either out of a sense of obligation or because they want to be in the good graces of a particular journal. Without this system in place there is a strong chance that peer review rates will drop and only a few papers will get reviewed. This will ultimately decrease the accuracy of science. In our (admittedly contrived/simplified) experiment[49] on peer review accuracy went from 39% to 78% after solutions were reviewed. You might argue that only "important" papers should be peer reviewed but then you are back in the camp of glamour. Say what you want about glamour journals. They are for sure biased by the names of the people submitting the papers there. But it is *possible* for someone to get a paper in no matter who they are. If we go to a system where there is no curation through a journal-like mechanism then popularity/twitter followers/etc. will drive readers. I'm not sure that is better than where we are now.

Post-publication review is theoretically a great idea, but I'm still waiting to see a system that beats pre-publication review for maintaining scientific quality (even though it may just be an impossible problem[50]).

If you do decide to use post-publication review keep in mind a few things. One is that there are power differentials in science and authoring a review can have implications for your career if it is unnecessarily negative. Another is that people will in general draw some conclusions about you from the way you referee. I would suggest following the refereeing model described above and the template[51] when performing post publication review while keeping your tone measured.

Note: Parts of this chapter appeared in the Simply Statistics blog post: "Preprints are great, but post publication peer review isn't ready for prime time"[52] and the Leek group guide to reviewing papers[53].

[49] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.002689

[50] http://simplystatistics.org/2015/02/09/the-trouble-with-evaluating-anything/

[51] https://raw.githubusercontent.com/jtleek/reviews/master/review-template.Rmd

[52] http://simplystatistics.org/2016/02/26/preprints-and-pppr/

[53] https://github.com/jtleek/reviews

Data sharing

Data sharing - what you I do and why?

The reproducibility debate is over. Data should be made available when papers are published. The Potti scandal [54] and the Reinhart/Rogoff scandal[55] have demonstrated the extreme consequences of lack of reproducibility and the reproducibility advocates have taken this one home. The question with reproducibility isn't "if" anymore it is "how".

The transition toward reproducibility is likely to be rough for two reasons. One is that many people who generate data lack training in handling and analyzing data[56], even in a data saturated field like genomics. The story is even more grim in areas that haven't been traditionally considered "data rich" fields.

The second problem is a cultural and economic problem. It involves the fundamental disconnect between (1) the incentives of our system for advancement, grant funding, and promotion and (2) the policies that will benefit science and improve reproducibility. Most of the debate on social media seems to conflate these two issues. I think it is worth breaking the debate down into three main constituencies: journals, data creators, and data analysts.

Data sharing, especially for large data sets, isn't easy and it isn't cheap. Not knowing how to share data is not an excuse - to be a modern scientist this is one of the skills you have to have.

Post your data to a public repository

You should post your data to a public repository. You will eventually change websites and managing your own data can be a hassle. Take advantage of field specific repositories or more general data sharing services like Figshare[57] to make your data available and citable.

Organize and document your data

It is relatively simple to post your data online these days, but often the data are deposited completely haphazardly. The reason is that there has historically been very little encouragement to share data in a thoughtful or maximally useful way. To maximize both the value of your data and your impact (a) post both raw and tidy versions of your data, (b) post relevant metadata about experiments you performed in a README, and (c) post related code that can be used to analyze the data as you did in your paper.

Include links to your data in your publications

[54] http://simplystatistics.org/2012/02/27/the-duke-saga-starter-set/
[55] http://simplystatistics.org/2013/04/19/podcast-7-reinhart-rogoff-reproducibility/
[56] http://simplystatistics.org/2012/04/27/people-in-positions-of-power-that-dont-understand/
[57] http://figshare.com/

It is surprising how often the data relevant to a particular paper are not easily discoverable, even if they are made publicly available. Make sure that you include direct and permanent links to your data sets in your publications.

Data sharing - what tools should I use?

Places you can share data

In general you want to post data to a data hosting website. If you post data on your personal website it is almost always impermanent. When you move institutions or change web servers the data will likely disappear as well. Link rot[58] is one of the most fundamental sources of reproducibility in science.

There are a variety of places you can make data publicly available. General purpose sharing sites include the following.

- Figshare[59] - if you make your data sets publicly available then you can post unlimited data for free. If you want private data sets there are some charges. Figshare accepts all data types and gives you a doi for your data so it can be cited.
- Open Science Framework[60] - can be used to post all types of data sets and also integrates with Figshare.
- Dataverse[61] - hosted by Harard, you can again host data for free.

You can also post data to a variety of field specific data repositories[62]. But be careful to ensure that the websites have (a) been around for a while, (b) have a track record of managing data, and (c) make their pricing structure clear if they charge.

Data sharing and manipulation

When posting data online, it is helpful to post both the raw and processed version of the data, as well as relevant metadata and instructions. I have posted a data sharing guide[63], some of which is also reproduced in the next section. Karl Broman also has an excellent data organization guide[64].

[58] https://en.wikipedia.org/wiki/Link_rot
[59] https://figshare.com/
[60] https://osf.io/
[61] https://dataverse.harvard.edu/
[62] http://oad.simmons.edu/oadwiki/Data_repositories
[63] https://github.com/jtleek/datasharing
[64] http://kbroman.org/dataorg/

Really big data

One major unsolved issue with data sharing is if you want to share very large data sets. Many of the most common repositories have file size limits and sharing limits. An alternative is to host them yourself on something like Dropbox[65] or Amazon S3[66]. Unfortunately storing data with these commercial providers (a) makes data more temporary and (b) incurs ongoing costs. Many fields where large data sets are generated have their own custom repositories you can use and that is still the best route if you have a large data set you need to share.

Private data

Recently there has been an explosion of interest in analyzing data that are kept private from the public for a variety of reasons. These data sets might come from a startup company like Facebook and be highly valuable, or they might be protected medical data in an electronic health record and pose privacy risks. One tricky issue that has arisen in the scientific community is dealing with private data sets. The research on these highly valuable data can be highly valuable. But they pose a risk to the scientific process. In general a paper can not be considered to be scientific if the results are not verifiable by anyone. Papers that (a) define no process for accessing the data used in the paper or (b) provide a mechanism for an outside group to verify their results must be considered suspect until verified through independent sources.

Data sharing - further tips and issues

Data analysts - have sympathy for data generators

Finally, I think that we should be more sympathetic to the career concerns of folks who generate data. I have written methods and made the code available. I have then seen people write very similar papers using my methods and code - then getting credit/citations for producing a very similar method to my own. I've been closely following the fallout from PLoS One's new policy for data sharing[67]. The policy says, basically, that if you publish a paper, all data and code to go with that paper should be made publicly available at the time of publishing and include an explicit data sharing policy in the paper they submit.

I think the reproducibility debate is over. Data should be made available when papers are published. The Potti scandal [68]and the Reinhart/Rogoff scandal[69] have demonstrated the extreme consequences of lack of reproducibility and the reproducibility advocates have taken this one home. The question with reproducibility isn't "if" anymore it is "how".

[65] http://www.dropbox.com/

[66] https://aws.amazon.com/s3/

[67] http://www.plos.org/data-access-for-the-open-access-literature-ploss-data-policy/

[68] http://simplystatistics.org/2012/02/27/the-duke-saga-starter-set/

[69] http://simplystatistics.org/2013/04/19/podcast-7-reinhart-rogoff-reproducibility/

The transition toward reproducibility is likely to be rough for two reasons. One is that many people who generate data lack training in handling and analyzing data[70], even in a data saturated field like genomics. The story is even more grim in areas that haven't been traditionally considered "data rich" fields.

The second problem is a cultural and economic problem. It involves the fundamental disconnect between (1) the incentives of our system for advancement, grant funding, and promotion and (2) the policies that will benefit science and improve reproducibility. Most of the debate on social media seems to conflate these two issues. I think it is worth breaking the debate down into three main constituencies: journals, data creators, and data analysts.

It is really hard to create a serious, research quality data set in almost any scientific discipline. If you are studying humans, it requires careful adherence to rules and procedures for handling human data. If you are in ecology, it may involve extensive field work. If you are in behavioral research, it may involve careful review of thousands of hours of video tape.

The value of one careful, rigorous, and interesting data set is hard to overstate. In my field, the data Leonid Kruglyak's group generated measuring gene expression and genetics[71] in a careful yeast experiment spawned an entirely new discipline within both genomics and statistics.

The problem is that to generate one really good data set can take months or even years. It is definitely possible to publish more than one paper on a really good data set. But after the data are generated, most of these papers will have to do with data analysis, not data generation. If there are ten papers that could be published on your data set and your group publishes the data with the first one, you may get to the second or third, but someone else might publish 4-10.

This may be good for science, but it isn't good for the careers of data generators. Ask anyone in academics whether you'd rather have 6 citations from awesome papers or 6 awesome papers and 100% of them will take the papers.

I'm completely sympathetic to data generators who spend a huge amount of time creating a data set and are worried they may be scooped on later papers. This is a place where the culture of credit hasn't caught up with the culture of science. If you write a grant and generate an amazing data set that 50 different people use - you should absolutely get major credit for that in your next grant. However, you probably shouldn't get authorship unless you intellectually contributed to the next phase of the analysis.

The problem is we don't have an intermediate form of credit for data generators that is weighted more heavily than a citation. In the short term, this lack of a proper system of credit will likely lead data generators to make the following (completely sensible) decision to hold their data close and then publish multiple papers at once - like ENCODE did[72]. This will drive everyone crazy and slow down science - but it is the appropriate career choice for data generators until our system of credit has caught up.

[70] http://simplystatistics.org/2012/04/27/people-in-positions-of-power-that-dont-understand/

[71] http://www.pnas.org/content/102/5/1572.long

[72] http://www.nature.com/encode/#/threads

Data generators - respect data reuse.

The editors of the New England Journal of Medicine posted an editorial[73] showing some moderate level of support for data sharing but also introducing the term "research parasite":

> A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as "research parasites."

While this is obviously the most inflammatory statement in the article, I think that there are several more important and overlooked misconceptions. The biggest problems are:

1. "**The first concern is that someone not involved in the generation and collection of the data may not understand the choices made in defining the parameters.**" This almost certainly would be the fault of the investigators who published the data. If the authors adhere to good data sharing policies[74] and respond to queries from people using their data promptly then this should not be a problem at all.
2. "**… but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited.**" The idea that no one should be able to try to disprove ideas with the authors data has been covered extensively. It is worth considering the concern about credit. I think that the traditional way credit has accrued to authors has been citations. But if you get a major study funded, say for 50 million dollars, run that study carefully, sit on a million conference calls, and end up with a single major paper, that could be frustrating. Which is why I think that a better policy would be to have the people who run massive studies get credit in a way that *is not papers*. They should get some kind of formal administrative credit. But then the data should be immediately and publicly available to anyone to publish on. That allows people who run massive studies to get credit and science to proceed normally.
3. "**The new investigators arrived on the scene with their own ideas and worked symbiotically, rather than parasitically, with the investigators holding the data, moving the field forward in a way that neither group could have done on its own.**" The story that follows about a group of researchers who collaborated with the NSABP to validate their gene expression signature is very encouraging. But it isn't the only way science should work. Researchers shouldn't be constrained to one model or another. Sometimes collaboration is necessary, sometimes it isn't, but in neither case should we label the researchers "symbiotic" or "parasitic", terms that have extreme connotations.

[73] http://www.nejm.org/doi/full/10.1056/NEJMe1516564

[74] https://github.com/jtleek/datasharing

4. **"How would data sharing work best? We think it should happen symbiotically, not parasitically."** I think that it should happen *automatically*. If you generate a data set with public funds, you should be required to immediately make it available to researchers in the community. But you should *get credit for generating the data set and the hypothesis that led to the data set*. The problem is that people who generate data will almost never be as fast at analyzing it as people who know how to analyze data. But both deserve credit, whether they are working together or not.
5. **"Start with a novel idea, one that is not an obvious extension of the reported work. Second, identify potential collaborators whose collected data may be useful in assessing the hypothesis and propose a collaboration. Third, work together to test the new hypothesis. Fourth, report the new findings with relevant coauthorship to acknowledge both the group that proposed the new idea and the investigative group that accrued the data that allowed it to be tested."** The trouble with this framework is that it preferentially accrues credit to data generators and doesn't accurately describe the role of either party. To flip this argument around, you could just as easily say that anyone who uses Steven Salzberg[75]'s software for aligning or assembling short reads should make him a co-author. I think Dr. Drazen would agree that not everyone who aligned reads should add Steven as co-author, despite his contribution being critical for the completion of their work.

How you should format your data

For maximum speed in the analysis this is the information you should share:

1. The raw data.
2. A tidy data set[76]
3. A code book describing each variable and its values in the tidy data set.
4. An explicit and exact recipe you used to go from 1 -> 2,3

Let's look at each part of the data package you will transfer.

The raw data

It is critical that you include the rawest form of the data that you have access to. Here are some examples of the raw form of data:

- The strange binary file[77] your measurement machine spits out
- The unformatted Excel file with 10 worksheets the company you contracted with sent you

[75] http://salzberg-lab.org/
[76] http://vita.had.co.nz/papers/tidy-data.pdf
[77] http://en.wikipedia.org/wiki/Binary_file

- The complicated JSON[78] data you got from scraping the Twitter API[79]
- The hand-entered numbers you collected looking through a microscope

You know the raw data is in the right format if you:

1. Ran no software on the data
2. Did not manipulate any of the numbers in the data
3. You did not remove any data from the data set
4. You did not summarize the data in any way

If you did any manipulation of the data at all it is not the raw form of the data. Reporting manipulated data as raw data is a very common way to slow down the analysis process, since the analyst will often have to do a forensic study of your data to figure out why the raw data looks weird.

The tidy data set

The general principles of tidy data are laid out by Hadley Wickham[80] in this paper[81] and this video[82]. The paper and the video are both focused on the R[83] package, which you may or may not know how to use. Regardless the four general principles you should pay attention to are:

1. Each variable you measure should be in one column
2. Each different observation of that variable should be in a different row
3. There should be one table for each "kind" of variable
4. If you have multiple tables, they should include a column in the table that allows them to be linked

While these are the hard and fast rules, there are a number of other things that will make your data set much easier to handle. First is to include a row at the top of each data table/spreadsheet that contains full row names. So if you measured age at diagnosis for patients, you would head that column with the name `AgeAtDiagnosis` instead of something like `ADx` or another abbreviation that may be hard for another person to understand.

Here is an example of how this would work from genomics. Suppose that for 20 people you have collected gene expression measurements with RNA-sequencing[84]. You have also collected demographic and clinical information about the patients including their age, treatment, and diagnosis. You would

[78] http://en.wikipedia.org/wiki/JSON
[79] https://twitter.com/twitterapi
[80] http://had.co.nz/
[81] http://vita.had.co.nz/papers/tidy-data.pdf
[82] http://vimeo.com/33727555
[83] http://www.r-project.org/
[84] http://en.wikipedia.org/wiki/RNA-Seq

have one table/spreadsheet that contains the clinical/demographic information. It would have four columns (patient id, age, treatment, diagnosis) and 21 rows (a row with variable names, then one row for every patient). You would also have one spreadsheet for the summarized genomic data. Usually this type of data is summarized at the level of the number of counts per exon. Suppose you have 100,000 exons, then you would have a table/spreadsheet that had 21 rows (a row for gene names, and one row for each patient) and 100,001 columns (one row for patient ids and one row for each data type).

If you are sharing your data with the collaborator in Excel, the tidy data should be in one Excel file per table. They should not have multiple worksheets, no macros should be applied to the data, and no columns/cells should be highlighted. Alternatively share the data in a CSV[85] or TAB-delimited[86] text file.

The code book

For almost any data set, the measurements you calculate will need to be described in more detail than you will sneak into the spreadsheet. The code book contains this information. At minimum it should contain:

1. Information about the variables (including units!) in the data set not contained in the tidy data
2. Information about the summary choices you made
3. Information about the experimental study design you used

In our genomics example, the analyst would want to know what the unit of measurement for each clinical/demographic variable is (age in years, treatment by name/dose, level of diagnosis and how heterogeneous). They would also want to know how you picked the exons you used for summarizing the genomic data (UCSC/Ensembl, etc.). They would also want to know any other information about how you did the data collection/study design. For example, are these the first 20 patients that walked into the clinic? Are they 20 highly selected patients by some characteristic like age? Are they randomized to treatments?

A common format for this document is a Word file. There should be a section called "Study design" that has a thorough description of how you collected the data. There is a section called "Code book" that describes each variable and its units.

How to code variables

When you put variables into a spreadsheet there are several main categories you will run into depending on their data type[87]:

[85] http://en.wikipedia.org/wiki/Comma-separated_values
[86] http://en.wikipedia.org/wiki/Tab-separated_values
[87] http://en.wikipedia.org/wiki/Statistical_data_type

1. Continuous
2. Ordinal
3. Categorical
4. Missing
5. Censored

Continuous variables are anything measured on a quantitative scale that could be any fractional number. An example would be something like weight measured in kg. Ordinal data[88] are data that have a fixed, small (< 100) number of levels but are ordered. This could be for example survey responses where the choices are: poor, fair, good. Categorical data[89] are data where there are multiple categories, but they aren't ordered. One example would be sex: male or female. Missing data[90] are data that are missing and you don't know the mechanism. You should code missing values as `NA`. Censored data[91] are data where you know the missingness mechanism on some level. Common examples are a measurement being below a detection limit or a patient being lost to follow-up. They should also be coded as `NA` when you don't have the data. But you should also add a new column to your tidy data called, "VariableNameCensored" which should have values of `TRUE` if censored and `FALSE` if not. In the code book you should explain why those values are missing. It is absolutely critical to report to the analyst if there is a reason you know about that some of the data are missing. You should also not impute[92]/make up/ throw away missing observations.

In general, try to avoid coding categorical or ordinal variables as numbers. When you enter the value for sex in the tidy data, it should be "male" or "female". The ordinal values in the data set should be poor", "fair", and "good" not 1, 2 ,3. This will avoid potential mixups about which direction effects go and will help identify coding errors.

Always encode every piece of information about your observations using text. For example, if you are storing data in Excel and use a form of colored text or cell background formatting to indicate information about an observation ("red variable entries were observed in experiment 1.") then this information will not be exported (and will be lost!) when the data is exported as raw text. Every piece of data should be encoded as actual text that can be exported.

The instruction list/script

You may have heard this before, but reproducibility is kind of a big deal in computational science[93]. That means, when you submit your paper, the reviewers and the rest of the world should be able to exactly replicate the analyses from raw data all the way to final results. If you are trying to be efficient, you will likely perform some summarization/data analysis steps before the data can be considered tidy.

[88] http://en.wikipedia.org/wiki/Ordinal_data
[89] http://en.wikipedia.org/wiki/Categorical_variable
[90] http://en.wikipedia.org/wiki/Missing_data
[91] https://en.wikipedia.org/wiki/Censoring_(statistics)
[92] https://en.wikipedia.org/wiki/Imputation_(statistics)
[93] http://www.sciencemag.org/content/334/6060/1226

The ideal thing for you to do when performing summarization is to create a computer script (in `R`, `Python`, or something else) that takes the raw data as input and produces the tidy data you are sharing as output. You can try running your script a couple of times and see if the code produces the same output.

In many cases, the person who collected the data has incentive to make it tidy for a statistician to speed the process of collaboration. They may not know how to code in a scripting language. In that case, what you should provide the statistician is something called pseudocode[94]. It should look something like:

1. Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3
2. Step 2 - run the software separately for each sample
3. Step 3 - take column three of outputfile.out for each sample and that is the corresponding row in the output data set

You should also include information about which system (Mac/Windows/Linux) you used the software on and whether you tried it more than once to confirm it gave the same results. Ideally, you will run this by a fellow student/labmate to confirm that they can obtain the same output file you did.

Note: Parts of this chapter appeared in the Simply Statistics blog posts "PLoS One, I have an idea for what to do with all your profits: buy hard drives"[95] and "On research parasites and internet mobs - let's try to solve the real problem."[96] as well as the Leek group guide to data sharing[97]

[94] http://en.wikipedia.org/wiki/Pseudocode

[95] http://simplystatistics.org/2014/03/05/plos-one-i-have-an-idea-for-what-to-do-with-all-your-profits-buy-hard-drives/

[96] http://simplystatistics.org/2016/01/25/on-research-parasites-and-internet-mobs-lets-try-to-solve-the-real-problem/

[97] https://github.com/jtleek/datasharing

Scientific blogging

Blogging - what should I do and why?

How much you blog depends a lot on the purpose of the blog. If you are only using it as a place to respond to criticism you don't need to blog very frequently at all. If you are using it to raise your profile, it can be useful to post more frequently so people will start to follow your blog. But be careful because there is an energy tradeoff - if you blog you won't have as much energy for writing papers and doing research.

I think the best way to deal with the energy tradeoff is to start a blog with other people. If you have co-authors then you can keep a more continual stream of content without feeling solely responsible for generating all of those posts. I think having a blogging co-op with 2-5 people is a good balance between distributing energy and ensuring you get maximal credit for the entire blog. But even larger co-ops can be really useful for drawing readership while minimizing your effort.

Blog to respond to criticism

A recent paper by Case and Deaton[98] on death rates for middle class people. was discussed by Andrew Gelman[99] among many others. They noticed a potential bias in the analysis and did some re-analysis. There has been [Roger and a lot of discussion around this paper[100] on death rates for middle class people. Whenever this happens someone points out how academics are often surprised by the speed and ferocity of the criticism of their work[101]. The best way to respond is to have a blog where you can relatively quickly publish long-form responses to any criticism that is raised about your work.

Blog to raise your profile

You can also use your blog for raising your profile in both the academic and non-academic communities. Within the academic community, technical posts that critique papers or propose new ideas may be the most valuable. Outside of the academic community a blog can be even more important. It can be a useful way for people to discover you.

In my own personal experience, a number of companies have approached me based on my blog posts. Several of my students have also used their blogs to help raise their profile and make connections in industry. The kinds of posts that are most useful for raising your profile in terms of job hunting tend to be posts that:

[98] http://www.pnas.org/content/early/2015/10/29/1518393112.full.pdf

[99] http://www.slate.com/articles/health_and_science/science/2015/11/death_rates_for_white_middle_aged_americans_are_not_increasing.html

[100] http://www.pnas.org/content/early/2015/10/29/1518393112.full.pdf

[101] http://noahpinionblog.blogspot.com/2015/11/gelman-vs-case-deaton-academics-vs.html

- Show off your storytelling or communication skills
- Show off your data visualization skills
- Teach people how to do some kind of analysis
- Discuss an important issue that is relevant in industry

Why you are blogging (image via https://xkcd.com/386/)

Blog just for fun/cataloging what you are up to

A blog can be a good place to document the things you are learning or working on. It can also be a nice place to work on projects that aren't your main thesis. In general working on research involves long periods of concentration on a very specific problem. This can be frustrating both because it limits what you are thinking about and because these types of projects take a long time to complete. Sometimes it is nice to have a blog so you can work on a short term and different project for the sense of accomplishment or variety it provides.

Blogging - what tools should I use?

This depends again on how much you want to use your blog to raise your profile versus just periodically. If you are only going to write a post every now and then I would recommend either a very simple blogging platform like

- Medium[102] - widely popular with general audiences and very easy to set up
- Blogger[103] - relatively easy to manage, integrates nicely with Google accounts
- Wordpress.org[104] - slightly less intuitive than Medium or Blogger, but with increased flexibility through plugins

[102] https://medium.com/
[103] http://www.blogger.com/
[104] https://wordpress.org/

If you are going to build a cooperative set of bloggers or are going to blog regularly to try to raise your profile you might consider other platforms like

- A self-hosted Wordpress[105] blog. This requires you to set up hosting and handle more yourself, but gives you some freedom of control and opens up some additional plugins.
- A blog hosted on Github using something like Jekyll[106] which can be used to easily integrate the contributions of many folks. This will require some technical expertise.

In general I would start with the simpler tools and then only move to more complicated tools as they become necessary.

Blogging - further tips and issues

Responding to criticism

Many times that folks have been confronted with blogs or the quick publication process of f1000Research[107]. I think it is pretty scary for folks who aren't used to "internet speed" to see this play out and I thought it would be helpful to make a few points.

1. **Everyone is an internet scientist now**. The internet has arrived as part of academics and if you publish a paper that is of interest (or if you are a Nobel prize winner, or if you dispute a claim, etc.) you will see discussion of that paper within a day or two on the blogs. This is now a fact of life.
2. **The internet loves a fight** The internet responds best to personal/angry blog posts or blog posts about controversial topics like p-values, errors, and bias. Almost certainly if someone writes a blog post about your work or an f1000 paper it will be about an error/bias/correction or something personal.
3. **Takedowns are easier than new research and happen faster**. It is much, much easier to critique a paper than to design an experiment, collect data, figure out what question to ask, ask it quantitatively, analyze the data, and write it up. This doesn't mean the critique won't be good/right it just means it will happen much much faster than it took you to publish the paper because it is easier to do. All it takes is noticing one little bug in the code or one error in the regression model. So be prepared for speed in the response.

In light of these three things, you have a couple of options about how to react if you write an interesting paper and people are discussing it - which they will certainly do (point 1), in a way that will likely make you uncomfortable (point 2), and faster than you'd expect (point 3). The first thing to

[105] https://wordpress.org/
[106] http://jekyllbootstrap.com/
[107] http://f1000research.com/

keep in mind is that the internet wants you to "fight back" and wants to declare a "winner". Reading about amicable disagreements doesn't build audience. That is why there is reality TV. So there will be pressure for you to score points, be clever, be fast, and refute every point or be declared the loser. I have found from my own experience that is what I feel like doing too. I think that resisting this urge is both (a) very very hard and (b) the right thing to do. I find the best solution is to be proud of your work, but be humble, because no paper is perfect and that is ok. If you do the best you can , sensible people will acknowledge that.

I think these are the three ways to respond to rapid internet criticism of your work.

- **Option 1: Respond on internet time**. This means if you publish a big paper that you think might be controversial you should block off a day or two to spend time on the internet responding. You should be ready to do new analysis quickly, be prepared to admit mistakes quickly if they exist, and you should be prepared to make it clear when there aren't. You will need social media accounts and you should probably have a blog so you can post longer form responses. Github/Figshare accounts make it better for quickly sharing quantitative/new analyses. Again your goal is to avoid the personal and stick to facts, so I find that Twitter/Facebook are best for disseminating your more long form responses on blogs/Github/Figshare. If you are going to go this route you should try to respond to as many of the major criticisms as possible, but usually they cluster into one or two specific comments, which you can address all in one.
- **Option2 : Respond in academic time**. You might have spent a year writing a paper to have people respond to it essentially instantaneously. Sometimes they will have good points, but they will rarely have carefully thought out arguments given the internet-speed response (although remember point 3 that good critiques can be faster than good papers). One approach is to collect all the feedback, ignore the pressure for an immediate response, and write a careful, scientific response which you can publish in a journal or in a fast outlet like f1000Research. I think this route can be the most scientific and productive if executed well. But this will be hard because people will treat that like "you didn't have a good answer so you didn't respond immediately". The internet wants a quick winner/loser and that is terrible for science. Even if you choose this route though, you should make sure you have a way of publicizing your well thought out response - through blogs, social media, etc. once it is done.
- __ Option 3: Do not respond.__ This is what a lot of people do and I'm unsure if it is ok or not. Clearly internet facing commentary can have an impact on you/your work/how it is perceived for better or worse. So if you ignore it, you are ignoring those consequences. This may be ok, but depending on the severity of the criticism may be hard to deal with and it may mean that you have a lot of questions to answer later. Honestly, I think as time goes on if you write a big paper under a lot of scrutiny Option 3 is going to go away.

All of this only applies if you write a paper that a ton of people care about/is controversial. Many technical papers won't have this issue and if you keep your claims small, this also probably won't apply. That being said, it is increasingly hard to execute Option 3. Responding to criticism is an important part of working on major scientific problems. Using a blog can be an effective, speedy, and controlled way of responding to issues that come up with your paper.

Controversy

Be aware that blogs are public and can end up widely distributed. A few years ago I wrote a blog post[108] that pointed out some ways that a particular media organization had distorted data. I wasn't trying to be political, just poke fun at the statistical curiosity in the media. But the post quickly went viral and ended up leading to a large amount of interesting email/responses. In general, be aware that if you blog about a controversial topic you may end up having lots of people read it. So if you are going to talk about controversial topics you should do it deliberately. It doesn't mean you shouldn't blog about important/controversial issues but you should be aware of the attention you may get.

Unprofessional content

In general you should reserve your scientific blog for professional writing. Keep in mind that your blog will be a reflection of you as a professional and a scientist. If you write rude, poorly researched, or angry posts - this may be the assumption that people make about you. When you apply for jobs, apply for grants, or apply for promotion you want all of your writing to reflect well on you.

Takedowns

One of the most common reasons to have a blog (and the easiest posts to write) are takedowns of other people's work and blogposts. Keep in mind that just because this is an easy route and may sometimes lead to pageviews it is not always the best thing to do. Keep in mind that easy criticism may mean major trouble for people on the receiving end. That doesn't mean don't criticize, but be aware of your tone and intent when doing it to avoid being cast as "just another angry blogger".

Note: Parts of this chapter appeared in the Simply Statistics blog posts "So you are getting crushed on the internet? The new normal for academics."[109]

[108] http://simplystatistics.org/2012/11/26/the-statisticians-at-fox-news-use-classic-and-novel-graphical-techniques-to-lead-with-data/

[109] http://simplystatistics.org/2015/11/16/so-you-are-getting-crushed-on-the-internet-the-new-normal-for-academics/

Scientific code

Scientific code - what should I do and why?

Code, like data, should be freely distributed. Code can include (a) software for performing analyses specific to a paper or (b) software for performing a more general analysis.

Post your code to a public repository

You should post your data to a public repository. You will eventually change websites and managing your own data can be a hassle. Take advantage of field specific repositories or more general code sharing services like Github[110] to make your code available.

Organize and document your code

It is relatively simple to post your code online these days, but often the code are deposited completely haphazardly. The reason is that there has historically been very little encouragement to share code in a thoughtful or maximally useful way. To maximize both the value of your data and your impact (a) post your code to an open repository, (b) document versions of the code you used for particular analyses, (c) document and explain your code with README files, comments, and literate programming.

Include links to your code in your publications

It is surprising how often the code relevant to a particular paper are not easily discoverable, even if they are made publicly available. Make sure that you include direct and permanent links to your code in your publications, along with relevant version numbers.

When sharing software, use repositories with review

It is very easy to share Python modules or R packages online through Github. But there is no process for vetting these software on general code repositories. It will dramatically improve the credibility and use of your software if you post them to repositories with some level of code review such as CRAN[111] or Bioconductor[112].

[110] github.com/

[111] https://cran.r-project.org/

[112] http://bioconductor.org/

Scientific code - what tools should I use?

Posting general code

For posting general code and software use any system that is public, shareable, and version controlled. The two most common options are Github[113] and Bitbucket[114]. Do not share your code via scripts stored in a non-version controlled directory like Dropbox or in a non-plain text format like pdf.

Posting software

For posting software that will be used by the community you can start with Github[115] and Bitbucket[116] for development and getting off the ground quickly. At some point, to make your software broadly available to the community you should consider distributing it through a curated repository.

- **R** - several curated repositories exist including CRAN[117] and rOpenSci[118] for general purpose software and Bioconductor[119] for genomics software.
- **Python** - there is an index of libraries at PyPI[120]
- _*Galaxy* - you can create Galaxy workflows and share them through the Galaxy Toolshed[121]

Posting analysis code

There are two main approaches to sharing code: through a collection of scripts and through literate programming documents.

Code and scripts

The most basic way to share code is through plain text code files. This is true regardless of what language you use. When writing code and distributing as scripts you should use version control software to keep track of the code you are using in the version of the analysis that you publish. Any text editor will do, but the best software integrate directly with Github/Bitbucket and include code

[113] http://github.com/
[114] https://bitbucket.org/
[115] http://github.com/
[116] https://bitbucket.org/
[117] https://cran.r-project.org/
[118] http://ropensci.org/
[119] http://bioconductor.org/
[120] https://pypi.python.org/pypi
[121] https://wiki.galaxyproject.org/GetInvolved?action=show&redirect=Get+Involved#Develop_Tools_and_Tool_Definitions

highlighting. Emacs[122] is a good general purpose code editor. You can also use online editors such as Cloud9[123]. If you are coding in R, Rstudio[124] is a good option.

Literate programming

Literate programming documents are documents that integrate code and text into a single document so that you can easily explain the results in the same document. The two most widely used literate programming tools are knitr[125] and rmarkdown[126] for R and Jupyter notebooks[127] that can be used with a wide range of programming languages. An advantage of literate programming is that you can make the documents more fully explain the analysis you have done, while leaving your work completely reproducible.

Scientific code - further tips and issues

Analysis code

Commenting and documenting

When distributing analysis code, the critical component is to document the code as extensively as possible. Karl Broman says:

> Your closest collaborator is you, six months ago, but you don't respond to emails.

In flat script files the optimal way is to use comments to describe what the code does. In literate programming documents such as R markdown documents or Jupyter notebooks an improved alternative is to use plain English to describe what you are doing. The key to leaving good and useful comments is to explain: (a) why you wrote this code, (b) how the code works, and (c) what it should produce.

Reporting versions

When you write a document for analysis, you should keep track of the version of the software you used and report it in any of your publications. If you are using Github or Bitbucket you can find this as the commit number from the version of the software that produces the results in your file.

Session information

Analysis code should include a command that prints off all the versions of the hardware, software platform, and software packages or libraries used during the analysis. That way people can go back and reconstruct your analysis even if versions of the software change.

[122] https://www.gnu.org/software/emacs/
[123] https://c9.io/
[124] https://www.rstudio.com/
[125] http://yihui.name/knitr/
[126] http://rmarkdown.rstudio.com/
[127] http://jupyter.org/

Software packages

Naming your software

The first step in creating your software is to give it a name. Hadley Wickham has some ideas[128] about it. Here are some useful rules:

- Make it googleable - check by googling it.
- Make sure there is no library or package with the same name on the repository you will submit to
- No underscores, dashes or any other special characters/numbers
- Make it all lower case - people hate having to figure out caps in names of packages.
- Make it memorable; if you want serious people to use it don't be too cute.
- Make it as short as you possibly can while staying googleable.

Versioning your software

There are a number of potential code conventions that can be used when versioning your software. One useful idea is to think about development and release versions of your software. Release is the stable version that everyone is using and development is where you make all of your potentially experimental changes. Bioconductor[129] has a useful versioning scheme[130] which can be augmented with ith some helpful suggestions from Kasper Hansen[131].

To use this system, the format of the version number will always be `x.y.z`. When you start any new package the version number should be `0.1.0`. Every time you make any change public (e.g., push to GitHub[132]) you should increase `z` in the version number. If you are making local commits but not making them public to other people you don't need to increase `z`. You should stay in version `0.1.z` basically up until you are ready to submit to a repository.

Before release you can increase `y` if you perform a major redesign of how the functions are organized or are used. You should never increase `x` before release.

The first time you release you would bump the version number to `1.0.0`. Immediately after releasing, if you plan to keep working on the project, you should bump your development version to `1.1.0`.

Thereafter, again you should keep increasing `z` every time you make a public change. If you do a major reorganization you should increase `y`.

[128] http://adv-r.had.co.nz/Package-basics.html
[129] http://www.bioconductor.org/
[130] http://www.bioconductor.org/developers/version-numbering/
[131] http://www.biostat.jhsph.edu/~khansen/
[132] https://github.com/

Documentation

This is how I feel about the relative importance of various components of statistical software development:

The relative importance of components of scientific software (image via the Simpsons)

Ideally your software is easy to understand and just works. But this isn't Apple[133] and you don't have a legion of test users to try it out for you. So more likely than not, at least the first several versions of your software will be at least a little hard to use. The first versions will also probably be slower than you would like them to be.

But if your software solves a real problem (it should!) and is well documented (it will be!) then people will use it and you will have a positive impact on the world.

Documentation has two main components. The first component is help files. The second component is vignettes or tutorials.

Help files

These files document each of the functions/methods/classes you will expose to your users. You should document (a) all inputs to the functions, (b) all outputs, and (c) expected behavior of the function under common circumstances. You
should also give examples of how to use them.

Vignettes

Documentation in the help files is important and is the primary way that people will figure out your functions if they get stuck. But it is equally (maybe more) critical that you help people *get started.*

[133] http://www.apple.com/

The way that you do that is to create a vignette. A vignette is a tutorial that includes the following components:

- A short introduction that explains
  - The type of data the package can be used on
  - The general purpose of the functions in the package
- One or more example analyses with
  - A small, real data set
  - An explanation of the key functions
  - An application of these functions to the data
  - A description of the output and how it can be used

Vignettes are a crucial component of making your software widely used. They will get people up and running on your package and give them a jumping off point. Any software that you hope to be widely used should include a Vignette or tutorial.

Good writers borrow from other authors, great authors steal outright

One thing I think a lot of academics (definitely including myself here) struggle with is the need to be "new" and "innovative" with everything they do. There is a strong selective pressure[134] for these qualities in academia.

But when writing software it is very, very important not to reinvent every single wheel you see. One person who is awesome at blending existing tools and exponentially building value is Ramnath Vaidyanathan[135]. He built slidify[136] on top of knitr[137] and rCharts[138] on top of existing D3[139] libraries. They allowed him to create awesome software without having to solve every single problem.

Before writing general purpose functions (say for regression or for calculating p-values or for making plots) make sure you search for functions that already exist.

An important and balancing principle is that you should try to keep the number of dependencies for your software as small as possible. You should also try to use packages that you trust will be maintained. Some ways to tell if a package is "trustworthy" are to check the number of downloads/users (higher is better), check to see if the package is being actively updated (on GitHub[140] or the appropriate repository) and there is a history of updates, and check to see if the authors of the packages routinely maintain important packages (like Hadley[141], Yihui[142], Ramnath[143], Martin

[134] http://en.wikipedia.org/wiki/Evolutionary_pressure
[135] https://github.com/ramnathv
[136] https://github.com/ramnathv/slidify
[137] http://cran.r-project.org/web/packages/knitr/index.html
[138] https://github.com/ramnathv/rCharts
[139] http://d3js.org/
[140] https://github.com/
[141] http://had.co.nz/
[142] http://yihui.name/
[143] https://github.com/ramnathv

Morgan[144], etc.).

Simple >>>> Complex

A major temptation for everyone creating code is to generate a bunch of features they think that users will want without actually testing those features out. The problem is that each new feature you create in your package will monotonically increase the number of dependencies and the amount of code you have to maintain. In general, the principle should be to create exactly enough functions that the users can install your package, perform your analysis, and return the results, **with no extraneous functionality**.

Specifically, be wary of things like GUIs or Shiny[145] apps. Given the heavy emphasis placed on reproducibility these days, it is rarely the case that real/important analyses will be performed in a point and click format.

If you are way into creating products that point-and-click users will be interested in I'm very happy to support you in that, since I think those things are cool, probably the future, and can certainly raise the interest in your work. But they present a potentially major difficulty in maintenance and should be placed in separate packages on your own account.

Note: Part of this chapter appeared in the Leek group guide to developing R packages[146]

[144] http://www.fhcrc.org/en/util/directory.html?q=martin+morgan&short=true#peopleresults
[145] http://www.rstudio.com/shiny/
[146] https://github.com/jtleek/rpackages

Social media in science

Social media - what should I do and why?

Social media can serve a variety of roles for modern scientists. Here I am going to focus on the role of social media for working scientists whose primary focus is not on scientific communication. Something that is often missed by people who are just getting started with social media is that there are two separate components to developing a successful social media presence.

The first is to develop a following and connections to people in your community. This is achieved through being either a content curator, a content generator, or being funny/interesting in some other way. This often has nothing to do with your scientific output.

The second component is using your social media presence to magnify the audience for your scientific work. You can only do this if you have successfully developed a network and community in the first step. Then, when you post about your own scientific papers they will be shared.

To most effectively achieve all of these goals you need to identify relevant communities and develop a network of individuals who follow you and will help to share your ideas and work.

Set up social media accounts and follow relevant people/journals

One of the largest academic communities has developed around Twitter, but some scientists are also using Facebook for professional purposes. If you set up a Twitter account, you should then find as many colleagues in your area of expertise that you can find and also any journals that are in your area.

Use your social media account to promote the work of other people

If you just use your social media account to post links to any papers that you publish, it will be hard to develop much of a following. It is also hard to develop a following by constantly posting long form original content such as blog posts. Alternatively you can gain a large number of followers by being (a) funny, (b) interesting, or (c) being a content curator. This latter approach can be particularly useful for people new to social media. Just follow people and journals you find interesting and share anything that you think is important/creative/exciting.

Share any work that you develop

Any code, publications, data, or blog posts you create you can share from your social media account. This can help raise your profile as people notice your good work. But if you only post your own work it is rarely possible to develop a large following unless you are already famous for another reason.

Social media - what tools should I use?

There are a large number of social media platforms that are available to scientists. Creatively using any new social media platform if it has a large number of users can be a way to quickly jump into the consciousness of more people. That being said the two largest communities of scientists have organized around two of the largest social media platforms.

- Twitter[147] - is a platform where you can post short (less than 140 character) messages. This is a great platform for both discovering science and engaging in conversations about topics at a superficial level. It is not particularly useful for in depth scientific discussions.
- Facebook[148] - some scientists post longer form scientific discussions on Facebook, but the community there is somewhat less organized and people tend to use it less for professional reasons. However, sharing content on Facebook, particularly when it is of interest to a general audience, can lead to a broader engagement in your work.

There are also a large and growing number of academic-specific social networks. For the most part these social networks are not widely used by practicing scientists and so don't represent the best use of your time.

You might also consider short videos on Vine[149], longer videos on Youtube[150], more image intensive social media on Tumblr[151] or Instagram[152] if you have content that regularly fits those outlets. But they tend to have smaller communities of scientists with less opportunity for back and forth.

Social media - further tips and issues

You do not need to develop original content

Social media can be a time suck, particularly if you are spending a lot of time engaging in conversations on your platform of choice. Generating long form content in particular can take up a lot of time. But you don't need to do that to generate a decent following. Finding the right community and then sharing work within that community and adding brief commentary and ideas can often help you develop a large following which can then be useful for other reasons.

[147] https://twitter.com/
[148] https://www.facebook.com/
[149] https://vine.co/
[150] https://www.youtube.com/
[151] https://www.tumblr.com/
[152] https://www.instagram.com

Add your own commentary

Once you are comfortable using the social media platform of your choice you can start to engage with other people in conversation or add comments when you share other people's work. This will increase the interest in your social media account and help you develop followers. This can be as simple as one-liners copied straight from the text of papers or posts that you think are most important.

Make online friends - then meet them offline

One of the biggest advantages of scientific social media is that it levels the playing ground. Don't be afraid to engage with members of your scientific community at all levels, from members of the National Academy (if they are online!) all the way down to junior graduate students. Getting to know a diversity of people can really help you during scientific meetings and visits. If you spend time cultivating online friendships, you'll often meet a "familiar handle" at any conference or meeting you go to.

Include images when you can

If you see a plot from a paper you think is particularly compelling, copy it and attach it when you post/tweet when you link to the paper. On social media, images are often better received than plain text.

Be careful of hot button issues (unless you really care)

One thing to keep in mind on social media is the amplification of opinions. There are a large number of issues that are of extreme interest and generate really strong opinions on multiple sides. Some of these issues are common societal issues (e.g., racism, feminism, economic inequality) and some are specific to science (e.g., open access publishing, open source development). If you are starting a social media account to engage in these topics then you should definitely do that. If you are using your account primarily for scientific purposes you should consider carefully the consequences of wading into these discussions. The debates run very hot on social media and you may post what you consider to be a relatively tangential or light message on one of these topics and find yourself the center of a lot of attention (positive and negative).

Teaching in science

Teaching - what should I do and why?

Teaching used to be an activity that was confined to your local institution. The hyper local nature of teaching had a few implications. The first was that your scientific reputation was less affected by your teaching, both for better and for worse. You can now use the internet to make your teaching material available to the world and even put entire courses online. This means that you can gain a strong, positive reputation in the scientific community for your teaching. It also means that you should exercise some care in deciding which teaching materials to post online.

This chapter is focused on internet-facing teaching as a modern outlet for sharing your educational materials with the world. It has less to do with pedagogy specifically and more to do with how you can get your courses out to a large audience.

Put all of your teaching materials online

Even if you don't organize your teaching materials into a course with assessments, the first step is to just put all of your materials online. You can share these through your personal website or on Figshare if you want people to be able to cite them. Any materials you use including lecture notes, codes, assignments, and tests can help your colleagues, be useful for students around the world, and improve your reputation.

Put teaching videos online

If you teach an in person class you can screen grab your laptop or tablet. Then you can post the video with minimal editing directly to Youtube. Include a link back to your course materials and this is a simple way to get your lectures out to a much broader audience.

(Optional) Build a MOOC

A MOOC is a massive online open course. There are different platforms for offering these classes, complete with assessments and credentials. The largest MOOC platforms can significantly increase the audience you will reach with your materials if you teach on the types of topics (primarily white-collar vocational) that are in large demand. A MOOC requires significantly more effort than posting your materials online and to Youtube.

(Optional) Answer questions on forums

There are now a wide range of forums for answering technical and scientific questions.

- Quora[153] - includes curated answers, often from professionals, primarily to technical questions but also a wide range of questions of interest to academics.

[153] https://www.quora.com/

- Stackoverflow[154] - is a forum for posting answers to questions about computer programming
- Reddit[155] - is a general purpose site where you can find message boards (called subreddits) on a variety of scientific topics.

You can develop your teaching reputation by monitoring sites like these or other field specific sites and providing answers to questions. If you do this often enough, it may boost your teaching reputation both inside and outside of your academic community.

Teaching - what tools should I use?

Tools for sharing lecture notes

- Slideshare[156] and Speakerdeck[157] can be used for sharing lecture slides and come with basic analytics.
- Figshare[158] can be used to share data, code, and slides and gives DOIs which can be cited.

Tools for sharing lecture videos

- Youtube[159] - is the dominant force and one of the easiest ways to share videos online. You can also develop a list of subscribers who will see your videos and potentially be directed back to your course content.
- Vimeo[160] - is an alternative platform to Youtube and is frequently used by artistic types. Particularly if you are making creative or artistic lecture videos this might be a good alternative.

Tools for sharing MOOCs

- Coursera[161] - is the largest MOOC provider. It is a for-profit company and you can only teach a course on the platform if your university has an agreement with Coursera and the university approves.
- Edx[162] - is another large MOOC provider. It is a non-profit, but again you can only teach a course on the main platform if your university has an agreement with Edx and the university approves.

[154] http://stackoverflow.com/
[155] https://www.reddit.com/
[156] http://www.slideshare.net/
[157] https://speakerdeck.com/
[158] https://figshare.com/
[159] https://www.youtube.com/
[160] https://vimeo.com/
[161] https://www.coursera.org/
[162] https://www.edx.org/

- Udemy[163] - this is not a MOOC platform, but allows you to create complete courses including videos and quizzes and sell them directly to consumers. In this case you would not be directly affiliated with your university.

Teaching - further tips and issues

Keep it short

As any Youtube or Vine artist can tell you, attention spans on the internet are very short. It is a good idea to try to keep individual lectures under 10 minutes at most if you can. You should consider breaking up longer material into a series of videos.

Develop a thick skin

Keep in mind that when you are teaching on the internet, it is like anything else on the internet and there will be massive criticism, some very helpful supporters, and a large silent majority. An important component to teaching a successful online course, particularly to a large audience, is to develop a thick skin and a calm demeanor when dealing with the large heterogeneity in response.

Intellectual property

The thinking around intellectual property for online courses is still relatively unformed at many universities. Some universities will claim intellectual property rights to your lecture materials and others won't. This is something you should pay attention to carefully if you plan to use platforms like Udemy where instructors offer a direct to consumer course, bypassing the university infrastructure. It probably will only be a major issue if your courses are hugely popular or generate a lot of revenue.

[163] https://www.udemy.com/

Books

Books - what should I do and why?

Traditional book publishing models involve convincing book editors that you are on to a good idea, writing a draft of a book, getting it reviewed, and then ultimately having it published. The system also means that the largest share of profits from the book go to publishers and most of the consumers are libraries. Modern tools have made it possible to publish books more quickly, with broader feedback, all while keeping a larger share of the profits.

Avoid traditional publishers

Traditional publishing processes are slow, do not provide a huge amount of advertising for authors unless they are famous, and don't return most of the profits to authors. Moreover, the amount of academic credit you get for a book is limited unless it is an absolute classic.

Develop your books on a platform that allows feedback

There are now a number of platforms for writing books and sharing drafts online. As you are working on your book, or when it is in rough draft form, share it online and get feedback from your future readers. This will help you develop an audience and give you strong and immediate feedback.

Sell your book on multiple online platforms

Selling your book across multiple online platforms like Amazon[164] and Leanpub[165] maximizes your audience. It also allows you to shop for the best royalty deals across platforms and drive your users to the platform that is best for you.

Books - what tools should I use?

Writing and selling

There are a number of tools for developing and selling books online including:

- Leanpub[166] - a good platform where you can edit books with Markdown. You get 90% of royalties after a processing fee. This is one of the best royalty rates available. You can also publish books in draft mode and collect a mailing list ot let people know when you have a new book.

[164] http://www.amazon.com/
[165] https://leanpub.com/
[166] https://leanpub.com/

- Gitbook[167] - has a slightly nicer interface for writing books and you can export them to all formats so you can sell them across multiple platforms.
- Amazon Kindle Direct Publishing[168] - probably has the largest audience but not a great royalty rate, as low as 30% for some books sold in some countries. This is your only chance to make a million online publishing - realistically that isn't going to happen though.

Books - further tips and issues

Advertising

Once you have written your book online you will need to do your own advertising. Some useful approaches for advertising books you have written this way include your own social media, blogs, and online courses. It can also be useful to reach out to other bloggers or your friends on social media to see if they will help you promote your work. A little effort in this regard can really magnify your audience.

Collecting emails

If you use a platform like Leanpub, readers can sign up to get alerts when you have new books. This sort of mailing list can steadily build so that you will have an audience for any new books that you decide to write.

Think shorter

As with all internet-facing activities people usually have a relatively short attention span. Keeping your book brief and tightly focused can avoid fatigue from users who don't have a paper copy of your book to read.

Physical copies with on-demand publishing

Here we have focused on internet and web based publishing as the primary outlet for modern scientific books. If you want to get print copies made the ideal situation is to use an on-demand publisher like Lulu[169]. You can upload your books in a slightly modified format from the version you created electronically and then set a price. The books will be printed as people pay for them, so there are no up front printing costs.

[167] https://www.gitbook.com/
[168] https://kdp.amazon.com/
[169] https://www.lulu.com/

Internal scientific communication

Internal communication - what should I do and why?

One of the key components of managing the scientific process is to communicate with your internal scientific team. The difficult thing about science is that this team is always changing and almost always juggling multiple projects simultaneously with overlapping sets of individuals. To effectively communicate you need a way to communicate with people on specific projects, record a history of what is happening and document progress, and ways to quickly bring new team members up to speed.

Use a communication tool that can be open to managed groups

When you have a scientific team that you need to manage on a given project it will be helpful to organize communication in groups. Email is one option, but people need to be able to enter and leave discussions so it is better to use a piece of software where groups can be managed.

Internal communication - what tools should I use?

There are a number of tools that can be used starting with simple message boards like Google Groups[170]. Lately there have been new software platforms developed for team communication that seem particularly useful for scientific groups.

- Slack[171] - is a group messaging app that can be used to manage groups of individuals and keep track of conversations. It does cost money if you want to keep track of old conversations. But it includes integrations for lots of other tools like Github, Google Docs, and Dropbox that make it very useful.
- Hipchat[172]- is an alternative to Slack that has many of the same features. It has a slightly different pricing system and isn't as widely used.

There are a number of open-source and self-hosted alternatives that are being developed to these main two platforms including things like Mattermost[173]. But for the most part, unless you want to manage a lot of the issues with a communication platform Slack or Hipchat are the best options.

[170] https://groups.google.com/forum/#!overview
[171] https://slack.com/
[172] https://hipchat.com/
[173] http://www.mattermost.org/

Internal communication - further tips and issues

Open a discussion for each project

When you are working on multiple projects simultaneously it can be useful to have a discussion for each project you are working on. These discussions should be open to the broadest possible audience you are comfortable with, but should be monitored to make sure that they aren't being taken off topic. When a project is completed the ideal scenario is to move this discussion to an archived section where it can always be referenced, but isn't cluttering the main discussion boards.

Summarize at regular intervals

When using any group communication tool there is a tendency to have long un-interrupted discussions. People who have been directly following the discussion will not have a problem catching up, but new members or people who pay attention irregularly will be left out. For each discussion it is useful to have a running "summary to date" document that can be updated with any crucial points raised by the discussion and can be quickly used to get people up to date.

Regularly onboard and offboard members

One of the things to keep in mind is that if you are using Slack and Hipchat members need to be managed over time. As individuals show up they should be issued an account and as they leave they should be phased out. That way you can (a) keep your costs down since most of the payment is by user/use and (b) make sure that information that is internal to your group isn't being shared with the broader community.

Include a thread for literature curation

One of the best ways for people to discover what is going on with a scientific group is to know the relevant scientific literature. The best way to do that is to have people post internally on a seperate channel or group any paper they think is interesting along with any commentary they think is appropriate. This group can often be the easiest way to get a student up to speed on a group's scientific thinking.

Generate onboarding documents

Turnover is always high in scientific groups. Students and postdocs regularly turnover as they start their new academic careers. To avoid unnecessary interruptions in productivity it is useful to have some onboarding documents that can be used to get students up to speed. Some useful onboarding documents include the following.

A list of relevant papers

This could come from the literature curation thread or be a separate curated list specific for new people.

A list of all relevant tools and how to access them

Whether they are experimental labs, how to order new reagents, how to get a computer, or how to log onto the server it is useful to have all relevant logistical details for the tools the group uses stored in one place.

Lab protocols

It is useful to have a collection of up-to-date protocols for the most common experiments or analysis that a group performs. These can be useful both as a learning exercise and for getting new members ready to help out with projects right from the start.

Scientific talks

This chapter borrows ideas that I got from the following resources:

- Karl Broman's How to give a scientific presentation[174].
- Zach Holman's guide to talks speaking.io[175] and his actual talks[176]

Scientific talks - what should I do and why?

Entertain, don't teach

Because you got invited?

When you are first starting out you should accept pretty much every opportunity to speak about your research you get. In approximate order of importance, the value of talks early in your career are:

1. To meet people
2. To make people excited about your ideas/results
3. To make people understand your ideas/results
4. To practice speaking

Later the reasons evolve, although not be as much as you'd think. The main change is that 4 evolves more into "to show people you are a good speaker".

The importance of point 1 can't be overstated. The primary reason you are giving talks is for people to get to know you. Being well regarded is absolutely not the goal of academia. However, being well known and well regarded can make a huge range of parts of your job easier. So first and foremost make sure you don't forget to talk to people before, after, and during your talk.

Point 2 is more important than point 3. As a scientist, it is hard to accept that the primary purpose of a talk is advertising, not science. See for example Hilary Mason's great presentation Entertain, don't teach[177]. Here are reasons why entertainment is more important:

- People will legit fall asleep on you if you don't keep them awake - this is embarrassing

[174] http://www.biostat.wisc.edu/~kbroman/talks/giving_talks.pdf
[175] http://speaking.io/
[176] http://zachholman.com/talks
[177] http://www.hilarymason.com/speaking/speaking-entertain-dont-teach/

- People will never understand your ideas/software/results if they fell asleep and didn't hear about them
- There is **no way** to convey any reasonably complicated scientific idea completely in an hour
- If you entertain people **they might go read your paper/use your code/download your data** which is the best way to achieve goal 3.

That being said, be very careful to avoid giving a TED talk[178]. If you are giving a scientific presentation the goal is to communicate scientific ideas. So while you are entertaining, don't forget why you are entertaining.

Know your audience

It depends on the event and the goals of the event. Here is a non-comprehensive list:

- **Group meeting**:
  - **Goal**: Update people on what you are doing and get help.
  - **What to talk about**: Short intro on your problem, brief update on what you've tried, long discussion about where you are going/what you need help on.
- **Short talk at conference**:
  - **Goal**: Entertain people, get people to read your paper/blog or use your software.
  - **What to talk about**: Short intro on your problem, brief explanation of solution, links to software
- **Formal seminar**:
  - **Goal**: Entertain people, get people to read your software, make them understand your problem/solution
  - **What to talk about**: Intro to your problem, how you solved it, results, and connection to broader ideas
- **Job talk**:
  - **Goal**: Get a job, entertain people, make them understand your problem/solution
  - **What to talk about**: Brief overview of who you are, intro to your (single) problem, how you solved it, results, summary of what you have done/plan to do.

Scientific talks - what tools should I use?

Making the presentation

In general any of the standard presentation creation tools will work including Powerpoint[179], Google Slides[180], or Keynote[181]. Avoid using software like Latex where you have to adjust the figures through text encoding. When making presentations you want to use software that makes it easy to move and adjust figures.

[178] https://www.ted.com/talks/browse
[179] https://office.live.com/start/PowerPoint.aspx
[180] https://www.google.com/slides/about/
[181] http://www.apple.com/mac/keynote/

Where you should put your talk

If you have weblinks in your talk you need to post it online. There is no way people will write down any of the links in your talk. Two good places to put your talks are https://speakerdeck.com/ or http://www.slideshare.net/. But then link to all the talks from one, single, memorable site you can show people at the very end of your talk. All my talks are at http://jtleek.com/talks/. You are welcome to send a pull request with your talk there if it is Leek group related, or you can put them on your own short and sweet web link.

I personally like Slideshare a little better these days because the slides are much easier to view on mobile phones and iPads, which is the most likely place someone will read your talk.

Scientific talks - further tips and issues

Structure of your talk

The biggest trap in giving a talk is assuming that other people will follow you because you follow the talk. In general it is always better to assume your audience knows less than you think they do. People like to feel smart. I have rarely heard complaints about people who went too basic in their explanations, but frequently hear complaints about people being lost. That being said, here are some structural tips. Your mileage may vary.

- **Always lead with a brief, understandable to everyone statement of your problem.**
- Explain the data and measurement technology before you explain the features you will use
- Explain the features you will use to model data before you explain the model
- When presenting results, make sure you are telling a story.

The last point is particularly important. Usually by the results section people are getting a little antsy. So a completely disparate set of results with little story behind them is going to drive people bonkers. Make sure you explain up front where the results are going (e.g. "Results will show our method is the fastest/most accurate/best ever"), then make sure that your results are divided into sections by what point they are making and organized just like they would be if you were telling a story.

Style of your talk

There are only a few hard and fast rules here. I really like Zach Holman's style guide speaking.io[182] and his actual talks[183]. My suggestion is pick one template/style and go with it for a few consecutive talks to avoid costly overhead. See what you like and what you don't, then edit. Here are the only hard and fast rules.

[182] http://speaking.io/

[183] http://zachholman.com/talks

- Title slide must have a contactable form of you (twitter handle, email address, etc.)
- Fonts can never be too big. Go huge. Small fonts will be met with anger.
- All figures should have big axes in plain English.
- Any figure you borrow off the internet should have a web link to the source
- Any figure you borrow off a paper should have a web link to the paper
- Any time you use someone else's slide you should put "Slide courtesy of So and so" with a link to their page
- Unless you know what you are doing, pick a solid (dark/light) background slide color and an opposite (light/dark) font.
- Pictures beat words by a ratio of 1000 to 1

What you should say out loud about figures in your talk

If you have a figure in your talk you should present it in the following way.

- Explain what the figure is supposed to communicate (e.g. "this figure shows our method has higher accuracy")
- Explain the figure axes (e.g. "the y-axis is sensitivity the x-axis is 1-specificity")
- Explain what trends the audience should look for (e.g. "curves that go straight up at zero and across the top of the plot are best")

About equations in your talks

If you are giving an empirical talk you will probably have some equations. That is ok. But the way you present them is critically important to giving an understandable talk. Here are a few important points:

- Before presenting an equation explain the data
- Whenever possible use words instead of symbols in equations (*Expression = Noise + Signal* is better than $E = N + S$ is better than $Y = X + E$)
- No more than two subscripts
- When explaining an equation
  - First explain the point of the equation (we are trying to model expression as a function of noise and signal)
  - Then explain what each symbol is (E is expression, N is noise, etc.)
  - Then explain how the model relates them (E is a linear function of signal and noise)
- Try to avoid subscripts, and no more than two are allowed.

Job talk specific issues

When giving a job talk you have an additional job on top of the four main goals we talked about above. The goal is to present your complete professional persona to a bunch of people who (mostly) won't know who you are. You should tailor this a little bit to the place you are applying to a job, but don't try to pretend to be something you are not. You can show a little more theory at a theory place and a few more results at an applied place, but don't claim you are proving theorems all the time if you aren't.

You should include both at the beginning and the end of the talk brief summaries of all the stuff you have worked on and plan to work on so they get an idea of who you are in a complete sense. **But only talk about one specific project when giving the talk**. There is nothing more detrimental to your chances of getting a job than going way over time or not getting to give your whole talk.

Two things that I think you want to convey when giving a job talk are:

1. You would be a fun person to work with and can play well with others
2. You have some unique expertise that will strengthen the place you are applying

Traditionally, people demonstrated #2 by showing that they could prove really hard theorems. At some places, that is still a really good thing to do. Other things that might make you unique are the ability to write amazing R packages that get used, the ability to analyze massive data sets other people can't, the ability to teach courses that students will want to take but the department doesn't have, or ideas about a brand new research area (with some data to back them up).

Before giving a job talk ask around and get a feel for the sorts of things that people have heard about the place you are applying so you can be smart about your choices of how/what to present.

Answering hard questions

Inevitably you will get hard questions during your talk. The most important point is not to panic and not to get defensive. It is way better to just say *I don't know*, then to get upset. When you get asked a really hard question you should:

- Take a second and a deep breath. If necessary, ask the person to repeat the question.
- Answer the question the best you can come up with on the spot
- Say *I don't know* if you don't know. If you say I don't know, then you can give your best guess and explain it is a guess.

The key is to distinguish what kind of response you are giving. Are you giving a response where you know the answer because you actually looked into that? Are you giving a complete guess where you have no idea? Or, what is more likely, are you somewhere in between?

Most importantly, don't feel embarrassed! Hard questions happen to everyone and if you are polite, explain how sure you are about your answer, and try your best it won't be a problem.

That one person who keeps going bonkers on you

Almost everywhere you give a talk there will be a person who is upset/intense/aggressive. **Do not fall into the temptation to be aggressive in return**. Answer their first questions politely just like everyone else. If they keep asking really aggressive/lots of questions, you are within your rightst to say: "You are bringing up a lot of good issues, I'd be happy to discuss with you after the presentation more in depth, but in the interest of time I'm going to move on to the next part of my talk". It is ok to keep things moving and to finish on time.

Finishing on time

Do it. People will love you for it. If you are the last speaker in a session and others have gone long, adapt and go shorter. The value you gain by making your audience happy >>>> the extra 5 minutes of details you could have explained.

Note: Parts of this chapter appeared in the Leek group guide to giving talks[184]

[184] https://github.com/jtleek/talkguide

Reading scientific papers

Reading scientific papers - what should I do and why?

The academic paper is still the primary way of distributing new knowledge to the world. There are other ways too, with code, or blogs, or twitter. But academic papers are still the gold standard and where the vast majority of new scientific discoveries are reported.

Well this depends a lot on what you are interested in. There are a few broad categories of journals depending on what you are looking for.

- **Science magazines**: Journals like Nature[185], PNAS[186], and Science[187] publish papers that are supposed to be "breakthroughs" of interest to a "general audience". This means that the papers tend to be written at a slightly less technical level and so are often more readable to people outside of a particular field. It also means that a lot of the most important papers get published in these journals. The flip side is that people often stretch hard in the interpretation of their data to make it seem like a "breakthrough" and get it into one of these journals. This means that the rate that papers are retracted is also very high[188] in these journals. So read what you see there with a healthy grain of salt.
- **Health magazines**: In the health sciences there is a similar set of journals like the New England Journal of Medicine[189] and the Journal of the American Medical Association[190]. They have the same benefits and caveats as the Science magazines, but with a more health flavored bent.
- **Mega journals**: Journals like PLoS One[191] and Peerj[192] also publish in a wide range of areas and a ton of papers. The review criteria here is that it must be "correct" but not necessarily a "breathrough". So the heterogeneity in the papers is high. If a paper seems too good to be true, again it is worth taking with a grain of salt.
- **Field-specific Scientific journals**: Most scientific journals are not megajournals or magazines. These journals tend to be very field specific and tend to be much heavier on the details. This is where most science is published. The papers tend to be less focused on "breakthroughs" but are also less consistently risky to trust in these journals. In my area the journals might be something like Biostatistics[193] or Biometrics[194].

[185] http://www.nature.com/index.html
[186] http://www.pnas.org/
[187] http://www.sciencemag.org/
[188] http://iai.asm.org/content/79/10/3855.long
[189] http://www.nejm.org/
[190] http://jama.jamanetwork.com/journal.aspx
[191] http://www.plosone.org/
[192] https://peerj.com/
[193] http://biostatistics.oxfordjournals.org/
[194] http://www.biometrics.tibs.org/

- **Conference papers**: In some fields, like computer science, people tend to publish in short, peer reviewed conference papers. These papers tend to be quite technical - conferences like NIPS[195] publish similar papers to very technical journals in other fields. Conference papers tend to be lighter on the detail and tend not to come with software/code so they can be a little harder to read and a little harder to use, but they are often talking about the very latest, coolest ideas in their subfield.

Reading scientific papers - what tools should I use?

The best places to find published academic papers are:

- Journal websites are a good place to start. Here[196] is a list of journals.
- You can also read papers in biomedical sciences on aggregator sites like Pubmed Central[197]

One problem with journal websites in particular is that many of the papers are behind a paywall - you have to pay to read them (see the next section). Increasingly you can find the latest papers on something called a pre-print server. These papers aren't peer reviewed yet, but a large fraction of them ultimately end up in peer-reviewed journals. The nice thing about these papers is that they are frequently the latest research and free to read. Two good preprint servers are bioRxiv[198] and arXiv[199].

Open access and #icanhazpdf

One thing that is super frustrating if you aren't at a university or research institute is that many papers you might want to read cost money. They cost money because journals are what's called closed-access and they depend on making their money from readers/subscribers. In general papers will be free in:

- Open access journals (journals that make their money from authors instead of readers) like PLoS journals, Peerj, etc.
- Preprint servers like bioRxiv and arXiv
- Aggregators like Pubmed Central from funders that require papers to be free
- On authors websites where sometimes people post a free version.

If you run into a paper that costs money the first thing to do is check and see if the authors have published a pre-print and then check their website. If you still don't have any luck you could email the authors directly to ask for a copy of their paper (they are usually happy to oblige).

[195] https://nips.cc/
[196] https://en.wikipedia.org/wiki/Lists_of_academic_journals
[197] http://www.ncbi.nlm.nih.gov/pmc/
[198] http://biorxiv.org/
[199] http://arxiv.org/

A more modern approach that has sprung up is something called *#icanhazpdf* which is a way you can crowdsource a pdf of a paper you can't read. If you have a twitter account, post a link to the paper, the hashtag *#icanhazpdf* and your email address and often someone will be willing to find and send you a copy of the paper. When you have the copy, delete your tweet, as this approach is technically a violation and could get you in trouble. Mostly journals won't care if you don't do this over and over with tons of papers, but be warned that journals can be nasty/lawyer up when their interests are being threatened.

Scihub

Currently there are a large number of papers behind paywalls. This is the subject of ongoing debate[200] among scientists, the press, and the public. Currently there is one tool that allows you to access all papers directly through the web. This tool sci-hub.io[201] allows you to insert the link for any paper (closed or open access) and returns a pdf. It is important to note that sci-hub is of questionable legality, but for the moment represents the easiest way to share closed access papers for which there is no preprint available. Ideally in the future, a more acceptable funding model will be created with less legal constraints.

Reading scientific papers - further tips and tricks

How much should you read?

Academic papers come out all the time. Thousands are published every year, including hundreds in any given specific area. Unless you devote yourself full time to reading academic papers you won't be able to keep up with them all. I believe in the idea that you should read papers that you find interesting. Science is awesome and you shouldn't waste your time on the boring parts if you can avoid it.

In general there are two main ways to find papers that I like. The way I used to do it was set up an aggregator with the RSS feeds from journals that I like, then I use the following (approximate) rates of reading parts of papers.

- 100% - read the title
- 20-50% - read the abstract
- 5-10% - look at the figures/captions
- 1-3% - read the whole paper

The new way that I do it is follow bioRxiv and a bunch of other people who have similar interests on Twitter. I use the above percentages for papers tweeted from aggregators and if I see a paper tweeted by 2-3 people I trust I usually end up reading that paper.

[200] http://www.nytimes.com/2016/03/13/opinion/sunday/should-all-research-papers-be-free.html?_r=0
[201] https://sci-hub.io/

Reading a paper - the abstract/title

Different people will have different strategies. First I read the title and the abstract to get a sense for (a) why the paper is interesting according to the authors and (b) what are the main results in the paper. I do this to see if I think the paper is worth spending the time to read any deeper. I don't judge the quality at all from these components, just whether the paper is interesting or not.

Reading a paper - the figures

If I think the paper is interesting based on the title/abstract then the next thing I do is look at the figures and figure captions. As I've mentioned in my guide to writing papers[202] the figures should tell a coherent story and should have figure captions that explain what is going on.

Hopefully the papers you are reading have figures that are this easy to read. I'm usually looking for the "story" that the authors are trying to tell. In the case of statistical or computational papers I'm also looking for comparisons to previous approaches and how this method stacks up.

Reading papers - the introduction

I usually skip the introduction. This is often an extended version of the abstract and often contains more opinion than fact about how awesome a particular result is.

The one exception I have to this rule is if I don't know the scientific area very well. Then I read the part of the introduction that reviews previous work in the area and if I don't understand something, I go chase down the references from the introduction and read through those to "get up to speed".

Reading papers - the methods/supplemental material

If I decide to read a paper carefully I spend the majority of my time reading the methods and supplemental material. This is where most of the real "science" is. It tells you how they did the experiments, how they analyzed the data, and how they support their conclusions. I'm looking for a few things when I read the methods section at a high level including:

- Do they explain clearly exactly which data they collected?
- Do they explain clearly exactly what analysis they performed on those data?
- Do they point to where I can get the data and code so I can verify these things?
- Do they explain every step in a process or skip over steps and reference previous papers?

Unfortunately, after that you sort of have to know the area to judge more critically whether the things they are doing are good or not. This comes with practice or with expertise in an area and can't be summarized very easily into succinct guidelines.

[202] https://github.com/jtleek/firstpaper

Reading papers - the results

I find that if the authors have done their job and made their figures tell the story and clear, then I usually spend less time reading the results section. The key results are usually in the figures, but I still glance over this section to see if there is any claim/idea that I missed from reading the figures. In general, I compare this section very carefully to the methods to make sure that the results seem well justified compared to what they say they did in the methods section.

Reading papers - Conclusions

Just like with the introduction, I often skip the conclusions. It is usually just a recap of what happened in the rest of the paper with a bit of guess work as to how the results might fit into the broader scheme.

Hype

One thing to keep in mind is that science is very often slow, steady, and incremental. But there is a lot of pressure on scientists to come up with "breakthroughs" (sometimes called the "i got a big one here"[203] fallacy). When reading a paper if the authors claim they have cured cancer, discovered life on mars, or unified relativity and quantum theory then you should assume that they are full of it unless conclusively demonstrated otherwise.

Explain it to someone else

Reading academic papers can be a great way to catch up on knowledge. But in general I don't feel like I understand what is going on in a paper until I can explain the paper to someone else. So I try to discuss papers I think are really important with other people. The best way to do this is in a journal club or some other forum where you can put up figures from the paper and try to explain what is going on yourself.

Find out if others have read it

A lot of the papers I find interesting other people also find interesting. One nice way to learn a little more about a piece of scientific research is to see if it has been discussed on blogs. One thing to keep in mind is that blogs often have an agenda, so you should read the posts with a heavy dose of skepticism as well. Still, they can provide useful perspective on papers you've read.

Note: Part of this chapter appeared in the Leek group guide to reading a scientific paper[204].

[203] http://simplystatistics.org/2016/02/01/a-menagerie-of-messed-up-data-analyses-and-how-to-avoid-them/

[204] https://github.com/jtleek/readingpapers

Credit

Credit - what should I do and why?

One of the major issues with a lot of the things I've discussed in this book (data sharing, post-ublication review, blogging, social media) is that they are not well understood by senior scientists. They also do not get as much credit as traditional academic publications. This means a couple of things for scientists seeking to act in the modern way.

Don't forget that the primary thing you will be judged on is publications

In the academic community publications are still the only real way to gain strong and lasting respect. So even if you develop a huge following on social media, share a ton of data sets, and post publication review like crazy you will barely make a blip on the traditional academic community without publishing in good journals. In some ways this is very frustrating, but in other ways it highlights the importance of careful, thorough work in science.

__Magnify your publications with modern scientific principles __

One reason to do many of the things I have talked about in this book is to magnify your presence in the scientific community. If you are publishing regularly in academic journals you can use social media to get more people to read your work, blogs to discuss details that don't fit in the papers, and share your code and data to get more people to cite your papers. At least for the moment this indirect credit is about the most you can hope for in traditional academic communities

Quantify everything

One thing that you can use to start to build a case for your social media, coding, and data sharing skills is to quantify everything that is possible. The number of people who read your blog, the number of downloads for your data, and the number of people who use your code are all important when building a case for academic credit.

Credit - what tools should I use?

The main way to make a case for credit for modern scientific activities is to (a) quantify them and (b) put them into the broader context of your academic trajectory. The first step is collecting the data. These sites can be used to collect valuable information on your research output.

- Google Analytics[205] - for quantifying visitors to websites that you have.

[205] https://www.google.com/analytics/

- Youtube Analytics[206] - for quantifying numbers of views of your videos
- Github stars[207] and Github forks[208] as metrics for code use
- Download statistics from specific curated repositories like on Bioconductor[209] for quantifying code use
- Number of followers on various social media sites
- Alt metrics[210] - for quantifying social media interest in your research outputs
- Publons[211] - for quantifying your refereeing output
- Google Scholar[212] - for quantifying citations to your academic work
- Number of views and downloads to your Figshare repositories for quantifying the impact of your data sharing.

Once the data has been collected you need to try to put it in context. This is an excellent presentation[213] by Lance Waller on how to put your achievements in context. You can also see his talk in this online conference on the Statistics identity crisis[214].

Credit - further tips and issues

Include metrics right on your CV

Include citations, downloads and other metrics directly in the relevant sections in your CV so that they are easily accessible. Consider listing both the numbers themselves and any context you can provide (percentiles, awards, recognition) so that people can easily interpret the metrics that you provide. In some cases, your CV format will be dictated by your school. If that is true, then you should consider compiling all metrics into a single and separate document that outlines your overall impact.

Discuss your approach with a senior colleague

Most of the energy around modern scientific approaches is concentrated in more junior scientists. But more senior scientists make a large number of the important decisions (faculty hiring, grants, faculty promotion). A really good idea is to run your plan by a senior scientist and get their feedback. They can help you understand what are the components of your research profile that will stand out in a positive way and which ones won't matter much. Ideally this will be a mentor who may even disagree with your approach but won't judge you harshly for doing something different. I've been very lucky to have mentors like this.

[206] https://www.youtube.com/analytics
[207] https://help.github.com/articles/about-stars/
[208] https://help.github.com/articles/fork-a-repo/
[209] https://www.bioconductor.org/packages/stats/
[210] https://www.altmetric.com/
[211] https://publons.com/
[212] https://scholar.google.com/
[213] https://github.com/jennybc/2015-08_bryan-jsm-stat-data-sci-talk/blob/master/2015-10_waller-data-science-promotion-tenure.pdf
[214] https://www.youtube.com/watch?v=JLs01Z5baSU&feature=youtu.be

Career planning

This chapter is designed for graduate students and postdocs in science.

Career planning - what should I do and why?

The most common reason that people go into science is altruistic. They loved dinosaurs and spaceships when they were a kid and that never wore off. On some level this is one of the reasons I love this field so much, it is an area where if you can get past all the hard parts can really keep introducing wonder into what you work on every day.

Sometimes I feel like this altruism has negative consequences. For example, I think that there is less emphasis on the career planning and development side in the academic community. I don't think this is malicious, but I do think that sometimes people think of the career part of science as unseemly. But if you have any job that you want people to pay you to do, then there will be parts of that job that will be career oriented. So if you want to be a professional scientist, being brilliant and good at science is not enough. You also need to pay attention to and plan carefully your career trajectory.

A colleague of mine, Ben Langmead, created a really nice guide for his postdocs to thinking about and planning the career side of a postdoc which he has over on Github[215]. I thought it was such a good idea that I immediately modified it and asked all of my graduate students and postdocs to fill it out. It is kind of long so there was no penalty if they didn't finish it, but I think it is an incredibly useful tool for thinking about how to strategize a career in the sciences. I think that the more we are concrete about the career side of graduate school and postdocs, including being honest about all the realistic options available, the better prepared our students will be to succeed on the market.

Have a career plan:

I know that you want to be a scientist because you love science. That is why we all want to be scientists. But if you want to be a professional scientist and get paid money to do it, you will also need a plan for professional development. Having a career in science can be incredibly rewarding - teaching students, doing research you consider important, and communicating those ideas to a big audience are all great. There are also some things about it that are hard. One thing that I think is unnecessarily hard is career planning. If you use this guide it can help you make sense of the opportunities and challenges that come with a data science career starting as a grad student or postdoc in my group.

Figure out what you want to do:

One of the hardest parts of planning as a graduate student or postdoc is figuring out what you want to do next. Here are some relatively rough guidelines of the types of jobs that you could go after.

[215] https://github.com/BenLangmead/langmead-lab/blob/master/postdoc_questionnaire.md

These are by necessity very brief. Rafa's post on hard/soft money jobs[216] and my adendum on liberal arts colleges[217] as well as this excellent post on academia versus industry for a junior person[218] are great places to start reading more.

- **Tenure track faculty/principal investigator research school**: You will primarily be writing lots of papers and to a greater or lesser extent (a) teaching, (b) writing grants, and (c) advising students. This job has some pressure associated with it because you are usually expected to get grants/write methods papers where you are in charge, but in some ways has the most freedom if you are successful.
- **Tenure track liberal arts college professor**: You will write fewer papers and do more teaching. The grant pressure is less but is steadily growing. This is a good option if you are really into teaching and engaging with undergrad students.
- **Research track faculty**: This job varies a ton by institution. It can be anything from a heavy teaching load to pure research and everything in between. Some places tend to treat these faculty really well. But that isn't universally true. These jobs tend to require you to get fewer individual grants/write fewer methods papers but that isn't a rule either.
- **Industry**: I have less experience on this side so I'm hesitant to make too many generalizations. That being said I think there is everything on this side from pretty research-like jobs, to engineering style jobs, to very business oriented jobs. If you are into this area your duties will depend a lot on your exact job.

Career planning - what tools should I use?

The rest of this document is a worksheet for you to fill in and keep updated. Write your answers in blue so they stand out.

What's your goal?

- Where will you go to find relevant job listings? Hints: CRA, Science/Nature jobs,….
- At a high level, academic jobs are a mix of research and teaching. A standard research-oriented "hard money" position is one where 75% of your salary is supported (in return for your teaching and administrative efforts), the other 25% you cover from grants, and you typically teach 2-3 courses a year. A "soft money" position is one where the majority (perhaps all) of your salary you cover from grants. You will usually teach 0 courses a year in this system; maybe you have to give some guest lectures. I don't know as much about pure teaching positions, but I think they tend to be 100% supported but you teach upwards of 5-6 courses a year. I know less about industry jobs, but the skills they value tend to be (a) software engineering, (b) really applied analysis, (c) communication/collaborative skills. You will likely spend very little time teaching and most time either analyzing data or building software. What sort of mix are you looking for?

[216] http://simplystatistics.org/2011/12/19/on-hard-and-soft-money/
[217] http://simplystatistics.org/2011/09/15/another-academic-job-market-option-liberal-arts/
[218] http://www.pgbovine.net/academia-industry-junior-employee.htm

- If you are considering academics what kinds of students would you like to mentor?
- What kinds of colleagues would you like to have?
- Paste links for three jobs that you like. Why do you like those jobs?

Who are you?

- What is the coherent "story of you" that you want your CV to tell?
- What is it that you do better than everyone else?
- What's the 1-paragraph description of what you will do as an independent investigator? (Easy to find examples of such statements on websites of you favorite labs)
- Who are the top 2 or 3 people/labs you admire and want to be like when you are independent

How do you fit?

- How does your story fit with the kind of job you're hoping to get?
- Who (specifically or in general) will advocate for you at a place like that?

How will you be funded?

- What grants will you apply for during your time here? What are the deadlines?
- What grants will you apply for in in the future (if you are going academic)?
- How else can you exhibit your ability to work independently?

How will you exhibit your mentoring and teaching abilities?

- Hints: maybe teach classes, class sessions, post teaching materials, record lectures

Networking

- What conferences / seminars / hackathons / etc should you attend? What are the deadlines?
- How can you expand your network in a way that will open doors? (Hints: lab visits, internships, competitions)
- Do you feel comfortable writing a blog or using social media professionally? (this is absolutely the best way to get an industry job as far as I can tell).

Recommendations

- Who do you want to write your recommendation letters?

- Do you need to meet and/or work with more people before you have all the letters you need? If so, how will you make this happen?
- What do you want those letters to say?

Learning more

- What is the best "job talk" you've seen and what did you like about it?
- What have you learned / would you like to learn about the hiring process from the perspective of the people doing the hiring? For academics I can be a useful resource, for industry we can talk to lab/department alumni.

Where can you go for help?

- Who can you go to for questions about grants/scholarships/awards?
- Do you know anyone who has applied for/got those grants/scholarships/awards?
- Who can you go to for help about jobs (besides your advisor(s))?
- Do you know anyone who had success on the market recently?

Note: Part of this chapter appeared in the Simply Statistics blog post: Science is a calling and a career, here is a career planning guide for students and postdocs[219] and the Leek group guide to career planning[220] which was inspired by Ben Langmead's career planning guide[221].

[219] http://simplystatistics.org/2015/05/28/science-is-a-calling-and-a-career-here-is-a-career-planning-guide-for-students-and-postdocs/
[220] https://github.com/jtleek/careerplanning
[221] https://github.com/BenLangmead/langmead-lab/blob/master/postdoc_questionnaire.md

Your online identity

Your online identity - what should I do and why?

Have a standard handle

Throughout this book we covered a range of tools that can be used to perform research, publish, communicate, and teach online. If you are using these tools to help with your academic career, you want people to be able to find you across all of the different tools we will be using. One way to do this is to identify a consistent identity and use it across different platforms.

For example, whenever possible I use the handle *jtleek*. My Twitter, Github, and Gmail accounts all use that same handle. So when people are looking for me on a new platform, they know what my most likely username will be. This won't always be possible, but when you can, choose a single username/handle and stick with it across platforms.

Use commercial platforms

In general you will be pushed toward using university email systems, data storage systems, and other infrastructure. Keep in mind that very rarely do academics stay at the same institution over time. For example, when you are a student, postdoc, or faculty member you will have an official university email account. This account will be useful for getting education discounts and will be the place that most official email will be sent. But you should, as early as possible, set up an email address on an independent platform like Gmail[222] that you won't lose when you change jobs. When you publish papers, data, or code, you should use this email address as your address for correspondence.

Don't be a jerk

Remember that if you are performing science on the internet, everyone can see it. This includes employers, friends, students, grant reviewers, promotion letter writers and your mom. Having an identity that is associated with positive, constructive contributions can be a massive asset. Having an identity associated with being a jerk can cause you real life problems.

Your online identity - what tools should you use?

Email

Use a commercial email provider rather than your local university email. The commercial email is much more likely to last than your university account. It is also wise to have two emails - one for personal correspondence and one for your professional persona.

[222] https://gmail.com

- Gmail[223] - the most widely used email address, also useful for a variety of other sites
- Yahoo[224] - still widely used, but sometimes perceived as a little past its prime.
- Outlook[225] - recently making a comeback thanks to useful app designs, etc.

Data/file management

You will probably be asked to use a data management system provided by your university. Again, it is better to use a commercial provider so you can ensure permanence and control of your data. The one exception is personally identifiable or regulated data - these must be carefully controlled and so may have to stay on university systems.

- Google Drive[226] - plenty of free storage space and plays nicely with Gmail
- Dropbox[227] - widely known and easy to use
- Box[228] - an alternative to Dropbox that is a little less widely known but easy to use

Your online identity - further tips and tricks

Picking your username

A couple of suggestions about usernames are that they should be (a) unique enough that they will likely be available across platforms, (b) personal to you, but not too cute, and (c) as short as possible so that people will be able to easily remember them. One common option is to use your first two initials and your last name, particularly if your last name is short. Alternative options include using just your first and last name separated by a period, just your first name if it is sufficiently unique, or a nickname that you wouldn't be embarrassed telling your professional colleagues.

[223] https://gmail.com
[224] https://www.yahoo.com/
[225] http://outlook.com/
[226] http://drive.google.com/
[227] http://www.dropbox.com/
[228] https://www.box.com/

About the author

I'm an associate professor of Biostatistics at the Johns Hopkins Bloomberg School of Public Health. I got my Ph.D. in 2007 in Biostatistics from the University of Washington, then did short postdocs at Mt. Sinai School of Medicine in Stem Cell Biology and Johns Hopkins School of Medicine in Computational Biology, then joined the faculty at Johns Hopkins.

When I started my career I planned on doing the usual stuff - publishing papers in journals, teaching courses at Hopkins, and acting like a standard academic. Then in 2013, with my friends Roger[229] and Rafa[230] I started a blog called Simply Statistics[231]. At first this was just a way for the three of us to share links, talk about frustrations with our profession, and point out cool things happening in the world of data science.

Over time I started to add new features to the blog, things like interviews with junior data scientists[232] and online conferences[233]. As our readership grew, the blog ended up being a very powerful force in my career. People who didn't know my papers or research very well still knew me through the blog. Our blog ended up growing a relatively large following on Twitter- @simplystats[234] - which then could be used to promote my scientific work, help my students get jobs, and in general increase my visibility in the scientific community.

Later, with two colleagues, I started to teach data science courses on Coursera. We now teach three large scale programs in data science[235], genomic data science[236], and executive data science[237] on that platform. When I started out teaching these courses, it wasn't clear at all that there was going to be any benefit to my career to teaching online. But the large scale numbers in these courses have again given me an opportunity to raise my profile among an audience that didn't previously know about me.

Should you follow my lead?

People have often asked me whether teaching online, writing a blog, and spending time on Twitter are a good idea for junior scientists. The truth is I'm not entirely sure. There are definitely ways

[229] https://twitter.com/rdpeng
[230] https://twitter.com/rafalab
[231] http://simplystatistics.org/
[232] http://simplystatistics.org/interviews/
[233] http://simplystatistics.org/conferences/
[234] https://twitter.com/simplystats
[235] https://www.coursera.org/specializations/jhu-data-science
[236] https://www.coursera.org/specializations/genomic-data-science
[237] https://www.coursera.org/specializations/executive-data-science

that being a modern scientist has benefited me. But I have had to spend a lot of energy on these endeavors and the payoff isn't always entirely clear.

In general some of the best advice I've received during my academic career came from my PhD advisor John[238] and my first faculty mentor Rafa[239]. They have both encouraged me to pursue things that I'm interested in, to try to show leadership in whatever things I've chosen to do, and to be aware of the important components of advancing my academic career, but not to be driven by them. So I'd pass that advice on.

[238] http://www.genomine.org/
[239] http://rafalab.dfci.harvard.edu/