现代科研指北

于淼

2018-07-16

目录

Chapter 1

前言

才疏学浅，不知何为真，仅通少错之法，故不敢言南，仅指北。或曰：现代科研挖坑 / 跳坑指南

Chapter 2

科研在搞什么鬼

Placeholder

2.1 科研老鸭汤-科学哲学沿革

2.1.1 古希腊

2.1.2 中世纪

2.1.3 1500 年以后

2.1.4 逻辑实证主义

2.1.5 否证主义

2.1.6 历史主义

2.1.7 无政府主义

2.1.8 实用主义

2.1.9 其他

2.2 科研职业化-问题为导向

2.3 科研精细化-体面的博士

2.4 科学知识的五个层次

2.4.1 背景组

2.4.2 已知的已知组

2.4.3 已知的未知组

2.4.4 未知的已知组

2.4.5 未知的未知组

2.5 知识体系的时间构建

Chapter 3

科研现状概览

Placeholder

3.1 国内版

3.2 国际版

3.3 趋势

3.3.1 科学方法

3.3.2 数据驱动的科研

3.3.3 假设检验问题

3.3.3.1 对策

3.3.4 社交网络中的科研

Chapter 4

思维工具篇

Placeholder

4.1 科学思维

4.1.1 规律的失效

4.1.2 哈森奇效应

4.1.3 观察研究的敌人-反馈

4.2 模型思维

4.2.1 可编程

4.2.2 抽象

4.2.3 交互作用

4.3 统计思维

4.4 估算法

4.4.1 费米估计

Chapter 5

实验

5.1 实验设计原则

5.2 定性实验

5.3 定量实验

5.4 思想实验

- 人是否在仿真中

Chapter 6

数据处理

Placeholder

6.1 多重比较

6.2 多重检验

6.3 回归

6.4 预测

6.5 仿真

6.6 可视化

Chapter 7

文献

Placeholder

7.1 文献管理

7.1.1 从无到有

7.1.2 从有到精

7.1.3 从精到用

7.2 信息收集

7.2.1 Zotero

7.2.2 Mendeley

7.2.3 EndNote

7.3 文本挖掘

7.3.1 关键词

7.3.2 作者

7.3.3 时空分布

7.3.4 影响力

7.4 荟萃分析

Chapter 8

学术生活

Placeholder

8.1 项目管理

8.1.1 香肠战术与拖延症

8.1.2 时间管理

8.1.3 笔记管理

8.2 学术出版

8.2.1 期刊论文

8.2.2 会议摘要

8.2.3 专著

8.2.4 专利

8.2.5 软件

8.3 学术会议

8.3.1 口头报告

8.3.2 海报报告

8.3.3 听报告

8.4 审稿

8.5 学术合作

8.5.1 数据共享

8.5.2 社交网络

8.6 讲课

8.7 课题组管理

8.7.1 科研的创业隐喻

8.7.2 基金申请

附录：现代科研兵刃谱

工欲善其事，必先利其器。今天绝大多数知识都是工具生产出来的，也就是想使用知识，肯定要先学工具，而工具又需要知识铺垫，这就成了一个鸡生蛋蛋生鸡的问题。虽然事后总结都有千般道理，但就我人经验而言，工具与知识是相辅相成缺一不可的，过于关注知识会导致脱离实际而沉迷于工具选择则有很高的迁移成本。这里的忠告就是不要想太多，先迈开步子，遵循梯度下降算法来寻优。也就是说，随便找个工具用起来，用实战来丰富需求，根据需求定向选择最适合自己的工具而不做工具的奴隶，如有必要，自己创造工具。另外，尽量选择那些花费百分之二十的精力可以掌握百分之八十的内容或应用场景的工具。同时系统学习那些使用频率高的工具，其余的只要知道其存在即可，不要捡芝麻丢西瓜。

文本编辑

科研用文本编辑工具主要应对排版要求，早期排版系统基本都是通过 TeX 语言来实现的，后来由于个人电脑普及及新兴学科的出现，很多科研人员上手会用的都是可见即可得的文本编辑器。现在期刊投稿一般会支持基于 TeX 的投稿及常见可见即可得文档，这些都是本地编辑。另一个当前流行的可见即可得文本编辑方式是在线协作，例如谷歌文档、石墨文档、腾讯文档等。对于需要协作完成的论文，在线协作文档极大方便了实时交互与版本控制。其实利用基于 Git 的GitHub也可以实现在线协作与修订，不过门槛比较高，但有希望成为一些期刊今后的投稿系统原型。

还有些文本编辑器是基于纯文本的，通过文本中的控制语句来实现排版，TeX 就是其中最流行的。Overleaf支持基于 TeX 的在线文档协作，甚至你可以直接用其向特定期刊投稿，同样的工具还有sharelatex。不过，TeX 的控制语句实在太丰富，学习起来比较困难。Pandoc 的出现方便了其他更简单的标记语言对 Tex 的转换，其中最容易上手的是Markdown。不过 Markdown 存在很多版本，其中基础版支持的排版功能非常有限，Pandoc 对其进行的扩展则支持了更丰富的功能方便排版。所以理论上你可以使用 Markdown 来写论文，不过这需要你的编辑器支持一些额外的功能。

总结一下，作为现代科研工具，理想文本编辑器需要至少有以下功能：

- 支持在线协作、评论与修订
- 支持版本控制
- 支持常见文献管理工具
- 支持期刊样式排版
- 容易上手

文献管理

现在的文献管理工具一般都支持常见文本编辑工具，也就是可以很方便的插入参考文献。然而，文献管理工具要同时具有收集、整理与分析的功能为佳。当前主流文献管理工具都已经支持浏览器层次的文献收集，也就是直接通过快捷键、脚本或浏览器扩展一键自动提取文章页面中参考文献信息并存入用户指定的文献库。要实现这个功能，多数需要知道文献数据库网页结构，当前很多文献数据库都推出了自己的文献收集应用，有的直接收购了文献管理软件。

Endnote是比较老牌的文献管理工具，不同于前面所说的网页采集，其自身就有与常见数据库的搜索接口，国内科研机构图书馆大都提供培训。与之类似的NoteExpress则属于国产软件，据说对中文期刊格式支持更好，类似的还有Mendeley、医学文献王、服务 TeX 里 BibTex 的 JabRef 与 Mac OS 下的Papers。这些工具起步较早，从单机时代就有用户，还有些工具诞生于互联网时代，有着更丰富的功能。

Zotero 属于互联网精神的产物，特别是前者本身就是基于火狐浏览器，其支持的文献格式样式都非常多，而且也有着丰富的文本分析扩展应用。Paperpile则属于基于谷歌文档的应用，可以很方便地管理在谷歌文档中使用到的文献。DOI与crossref的出现则更方便了文献的搜索定位。可以说基于互联网的团队化文献管理正在成为趋势。

总结一下，作为现代科研工具，理想文献管理软件需要至少有以下功能：

- 支持常见文本编辑器
- 支持在线文献采集
- 支持文献库协作与共享
- 支持文献信息学探索
- 容易上手

数据处理与绘图

数据处理方面很多学科只需要电子表格与基本的统计分析就可以了，很多在线服务就可以完成。然而，有些学科需要更丰富的功能例如多元统计分析与假设检验时，电子表格提供的功能可能就不那么明显了，有时需要学习使用电子表格的宏扩展来实现。此时，很多人容易陷入哪个分析一定要用哪个软件做的误区，其实多数数据分析软件的算法都差不多，只不过默认值可能不同，有些功能则藏的比较深，此时请善用搜索引擎。

所见即所得的数据处理与绘图软件有很多，Excel、Origin、SigmaPlot 与SPSS 是科研中用的比较多的。这些软件都是图形界面操作且都收费，其内置很多现成的分析模块应对实际科研问题，但这些简化会导致使用者知其然不知其所以然，在分析方法使用上陷入误区。

编程分析与绘图则属于基础的工具，R 、Python、Matlab 与SAS 都是这类工具的代表，应该说掌握其中任意一个就足够应对科研中需要的数据分析了。不过通常这类工具比较难学，最好是配合数据分析方法的学习同步掌握，而且要通过案例来理解方法，累积经验。如果推荐一个，那么基于 R 的 ggplot2 作图与

其背后的 tidyverse 数据分析套装则是很好的起点。如果更进一步，可以用shiny 来制作交互式数据展示界面。

此外，互联网上也有一些在线应用可以很方便地生成特殊图形例如百度脑图可以用来生成流程图或思维导图、Autodraw可以用来画简笔画、plotly可以在线完成绘图等。甚至网上还有直接上传数据后自动猜测你需要进行分析与制图的Charted。这样的工具只要搜索你所需要的分析然后加上 “online” 作为关键词就可以找到。

总结一下，作为现代科研工具，理想数据分析与绘图软件至少有以下功能：

- 支持科研用统计分析
- 图片默认输出美观大方支持绘图自定义
- 具备可重复性的宏功能或数据处理脚本
- 容易上手

学术交流

学术交流是科研生活中可以说最重要的一环，现代科研体系的分工合作都要通过学术交流来实现。主流趋势包括论文预印本服务器、开放获取与线上学术交流。

预印本指在通过同行评议发表之前事先将论文手稿托管在公开服务器的研究工作。预印本服务器可以加速新思想的交流，接受预印本发表的期刊可以从维基百科上查到。比较知名的预印本服务器包括偏数学物理计算机科学的arxiv、偏生命科学的biorxiv 与偏化学的chemrxiv。国内也有中科院的科技论文预发布[平台来服务国内科研人员。很多期刊出版方也在推广自己的预印本服务器来吸引高水平研究，所以可酌情选择。

开放获取是另一个趋势，要求研究工作可以公开让大众阅读。目前很多科研基金都开始有了这方面的要求及预算。但值得注意的是虽然开放获取期刊可能有更好的阅读数与引用表现，但有很多机构的开放获取期刊属于掠夺性期刊，给钱就发表，对学术评价与学科发展非常不利，可以通过一些网络上的列表来鉴别。要实现开放获取或者说透明科研，f1000research、PeerJ还有Plos都是还不错的先行者，它们在实践一些新理念，不过显然并不便宜。

线上学术交流除了期刊外，实际还要包括学术博客、多媒体展示、学术出版与网络身份。制作学术博客的工具可以直接借助平台例如科学网博客，也可以自己搭建例如使用Wordpress、Blogdown或者Netlify等工具。幻灯片制作也最好使用网页模式方便交流，xaringan、learnr等其他基于 Markdown 语言的幻灯片制作工具可以满足要求。学术出版物则可以通过bookdown或rticles等工具来完成。线上的学术身份识别对于存在大量重名现象的中国科研人员也是很有必要的，ORCID、Researcher ID、Scopus Auther ID、谷歌学术个人主页及国内的百度学术个人主页都是不错的网上学术名片。而在线交流的手段则可通过ResearchGate、Academia、Linkedin及twitter来完成。

审稿也是很重要的学术交流方式，建议使用 Publons 来构建自己的学术审稿记录。当然你可以在博客或微博上评论最新研究，甚至很多网络期刊网站的评论也有很好的思想碰撞，这里最关键的是要搞清楚你所在学科最活跃的网络交流平台，如果没有，自己搭建一个也无妨。

数据分享

数据分享是一个很重要现代科研特征，越来越多的科研成果正在开放自己的原始数据供社区推动学科进步。其中，figshare、Open Science Framework、Dataverse与Zenodo都是这一潮流的引领者。良好的数据分享不仅包含原始数据，还要包括处理后数据、数据收集相关信息与处理代码，另外对于共享数据的使用也要尊重数据生产者。

代码管理

后续我们会看到所有学科都会不可逆引入编程，所以代码管理工具也非常重要。Github与Bitbucket都是非常实用的在线代码管理与版本控制平台。而Rmarkdown与Jupyter Notebook等工具背后提倡的文学化编程也是很重要的代码开发工具。此外应考虑为未来自己做好注释并记录运行环境保证重复性。Docker image等完整的数据分析环境也可能成为现代科研的主流。代码的编写要能站到巨人肩上：

> Good writers borrow from other authors, great authors steal outright

R 包管理

对于 R 包的管理，建议打印相关 Rstudio 出品的小抄作为参考。同时作为 IDE，Rstiduo 提供了包开发的模版，可以使用formatR 与 Rd2roxgen来重新格式化旧代码。同时使用roxygen2来编写开发文档。为了让包更容易使用，可以用 Rmarkdown 来写小品文方便读者上手，另外就是使用testthat来进行代码的单元测试。对于代码的执行效率，可以用Profvis进行可视化而集成在线测试则可以通过travis-ci或appveyor来分别对 R 包进行 Linux 与 Windows 系统下的测试。当然，包完成后可通过 pkgdown来制作网站并通过learnr 来制作交互式教程。