Learn computational biology the ~~hard~~ right way

Ming ‘Tommy’ Tang
Director of computational Biology at Immunitas
Twitter: tangming2005
https://divingintogeneticsandgenomics.com/
03/18/2023

Ming ‘Tommy’ Tang, PhD

Boston, MA
tangming2005@gmail.com

2008 **BS Biotechnology**

2014
PhD. Genetics & Genomics

Outstanding
International student

2014-2018
Postdoc & Research Scientist

Instructor of
Data Carpentries

2018-2020
Senior bioinformatics scientist

Top 75 Bioinformatics
Blogs by feedspot.com

2020 – 2021
Lead Scientist

Lead the Bioinformatics effort for
Cancer Immunologic Data Commons

2021- current
Director of computational biology

40 publications

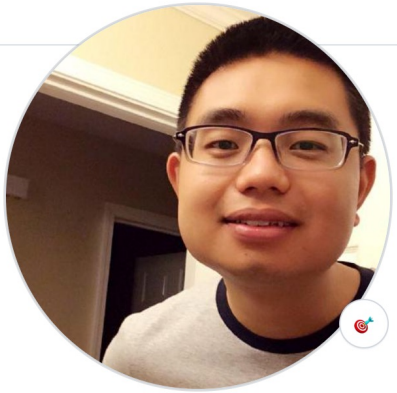
DNA CONFESSES DATA SPEAK

Blog: https://divingintogeneticsandgenomics.com
6000 views per month

@tangming2005 20K followers

Who am I ?

Overview | Repositories 141 | Projects | Packages | Stars 534



Ming Tang
crazyhottommy

Director of Computational Biology at Immunitas working on single-cell RNAseq. Care about reproducible research and open science

Edit profile

1.7k followers · **39** following

Immunitas
Waltham, MA
tangming2005@gmail.com
http://divingintogeneticsandgenomics.r...

Achievements

crazyhottommy / README.md

Hi there 👋

- I am a computational biologist working on (single-cell) genomics, epigenomics and transcriptomics.
- I use machine learning approaches to find new drug targets for cancer patients;
- I use google cloud and Terra for large scale data processing;
- I use R primary for data wrangling and visualization in the tidyverse ecosystem;
- I use python for writing Snakemake workflows and reformatting data;
- I am a unix geek learning shell tricks almost every month; I care about reproducible research and open science.

Learn more about me at my blog

Pinned

Customize your pins

ChIP-seq-analysis Public
ChIP-seq analysis notes from Ming Tang
Python 583 267

RNA-seq-analysis Public
RNAseq analysis notes from Ming Tang
Python 688 262

getting-started-with-genomics-tools-and-resources Public
Unix, R and python tools for genomics and data science
Shell 758 253

pyflow-ChIPseq Public
a snakemake pipeline to process ChIP-seq files from GEO or in-house
Python 89 39

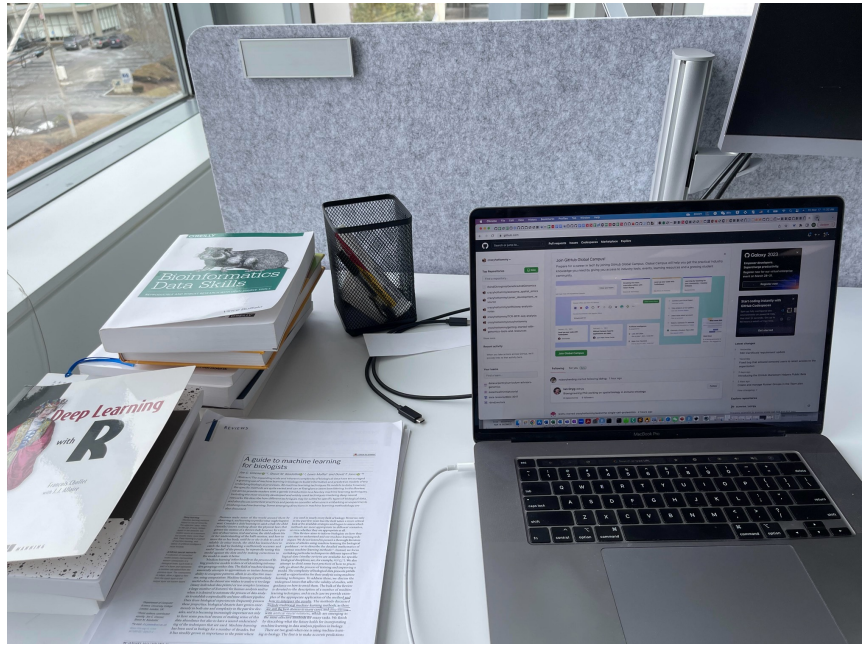
scRNAseq-analysis-notes Public
scRNAseq analysis notes from Ming Tang
373 110

scclusteval Public
Single Cell Cluster Evaluation
R 61 8

https://github.com/crazyhottommy

Make the transformation you want

2013



2023

Data deluge

1.845e+16

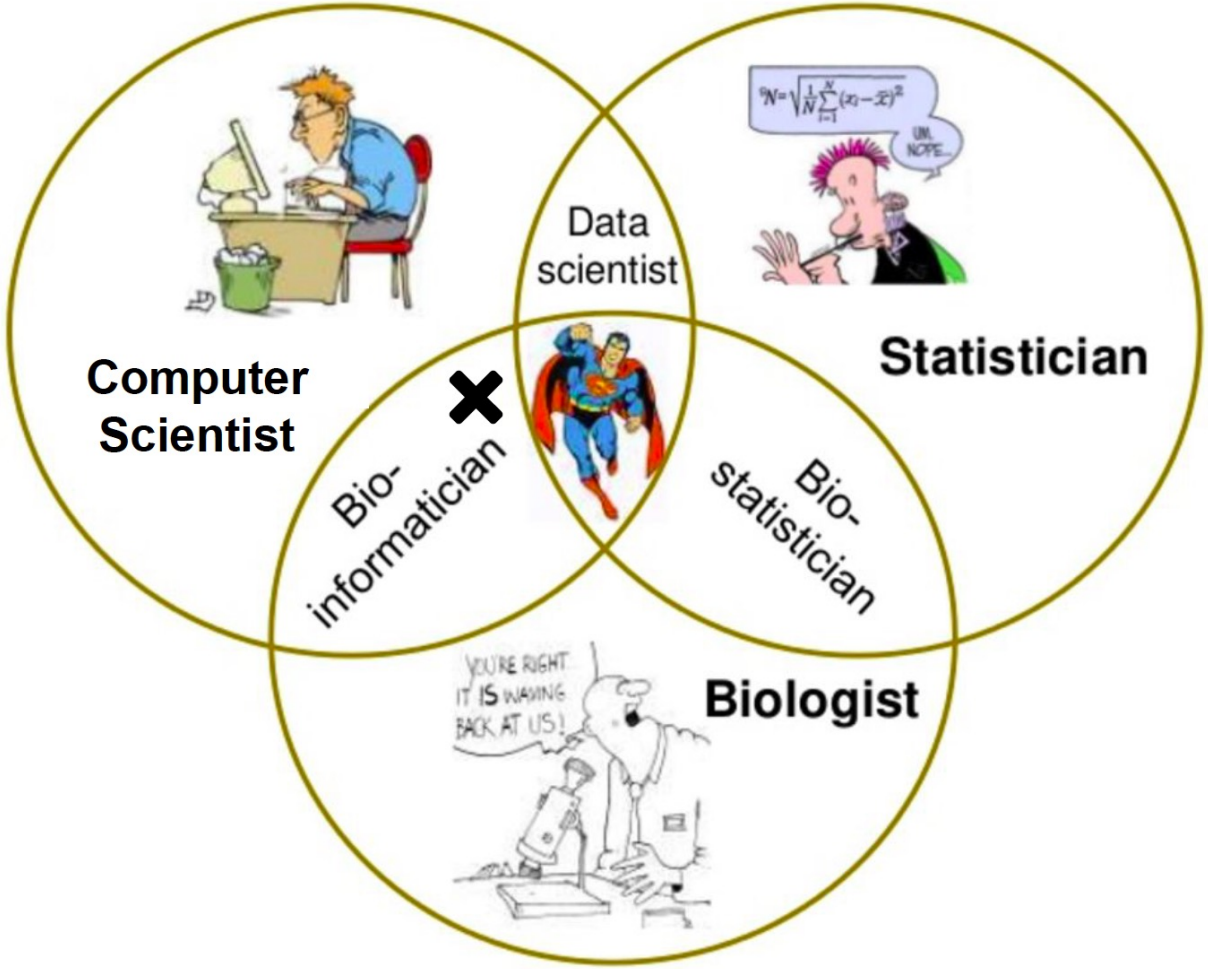
Number of publicly available bases in the NCBI Sequence Read Archive (SRA) as of July 1, 2018. This is the equivalent of 6,153,232 human genomes (which is 3e+9 bases).

30TB

Approximate amount of public sequence data received and processed ***daily*** by the NCBI Sequence Read Archive (SRA).

Sean Davis

Superman/Wonder woman



Bioinformatician
Computational biologist

Credit: Torsten Seemann

What should you learn to tame the data?

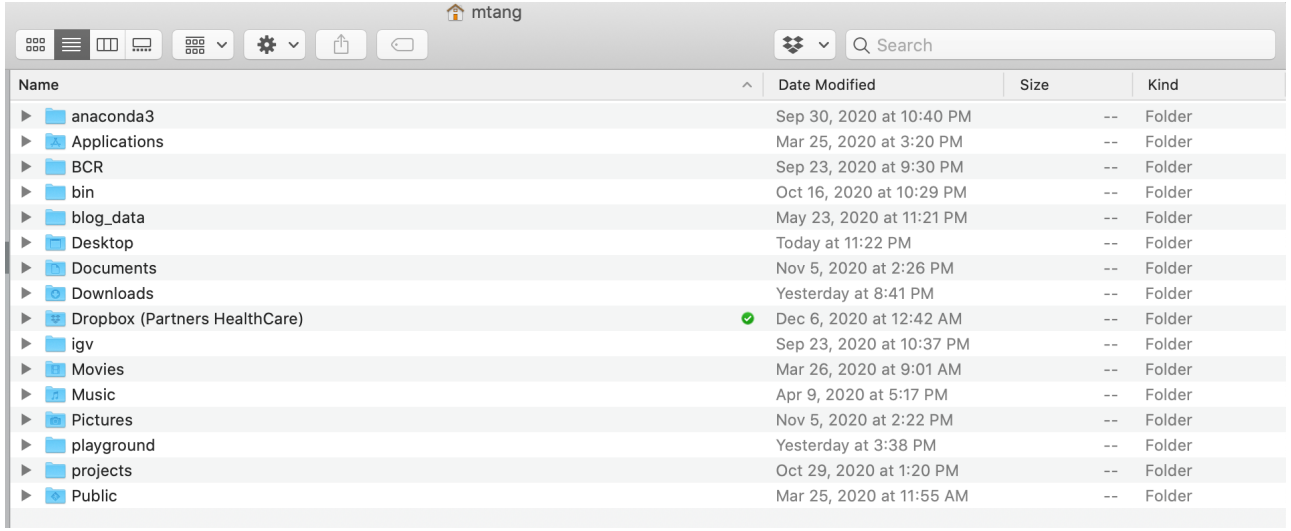
Learn Unix command line

- Why command line?
- The text file is still the ”king” format of bioinformatics. Unix commands are perfect to wrangle files.
- Most bioinformatics tools are run by the command line.
- More efficient/powerful: e.g, cp *png pictures/
- HPC (high-performance computing cluster), cloud computing

Terminal

```
(base) ➜  ~ ls
Applications                    Pictures
BCR                             Public
Desktop                         anaconda3
Documents                       bin
Downloads                       blog_data
Dropbox (Partners HealthCare)   igv
Library                         playground
Movies                          projects
Music
(base) ➜  ~
```

CLI

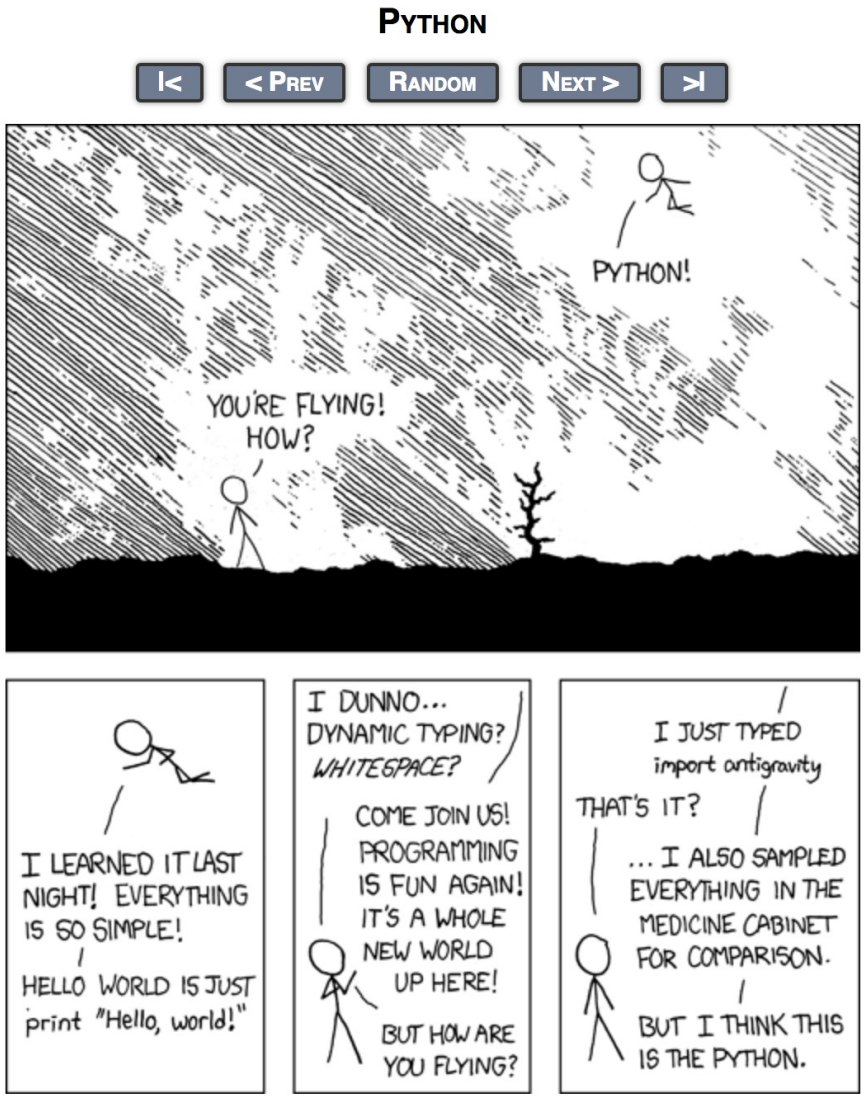


GUI

Use a mac/ubuntu or windows10 has a built-in

http://swcarpentry.github.io/shell-novice/

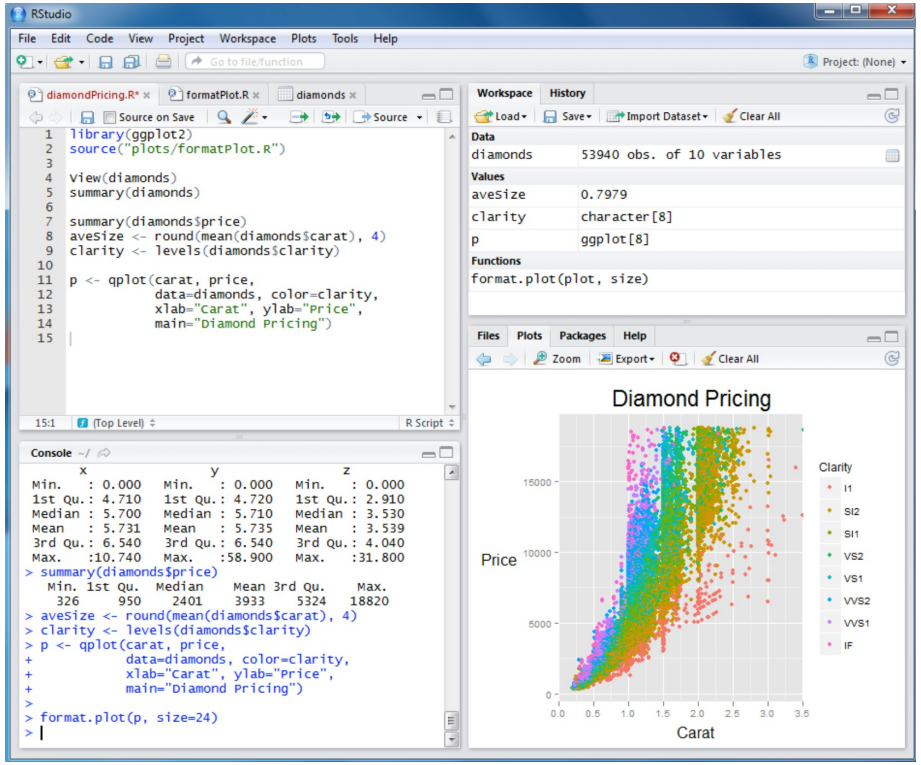
Learn some python

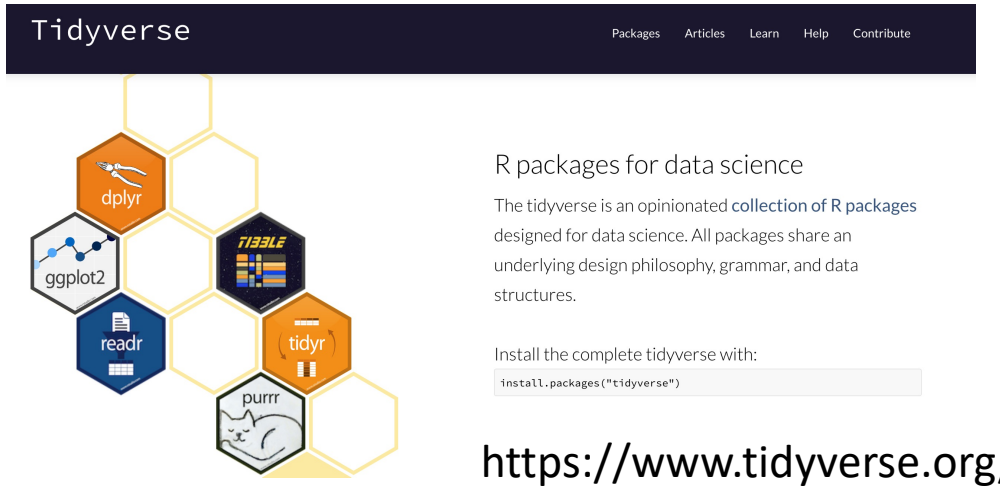


https://xkcd.com/

Learn some R

- Rstudio (IDE)
- Bioconductor
- Tidyverse and ggplot2

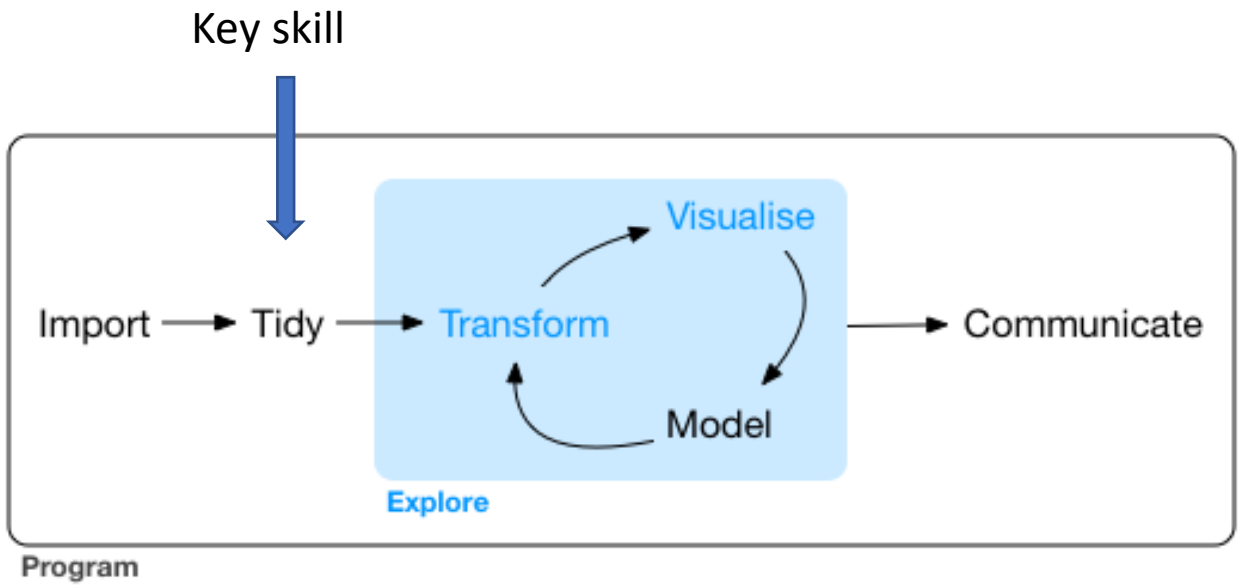




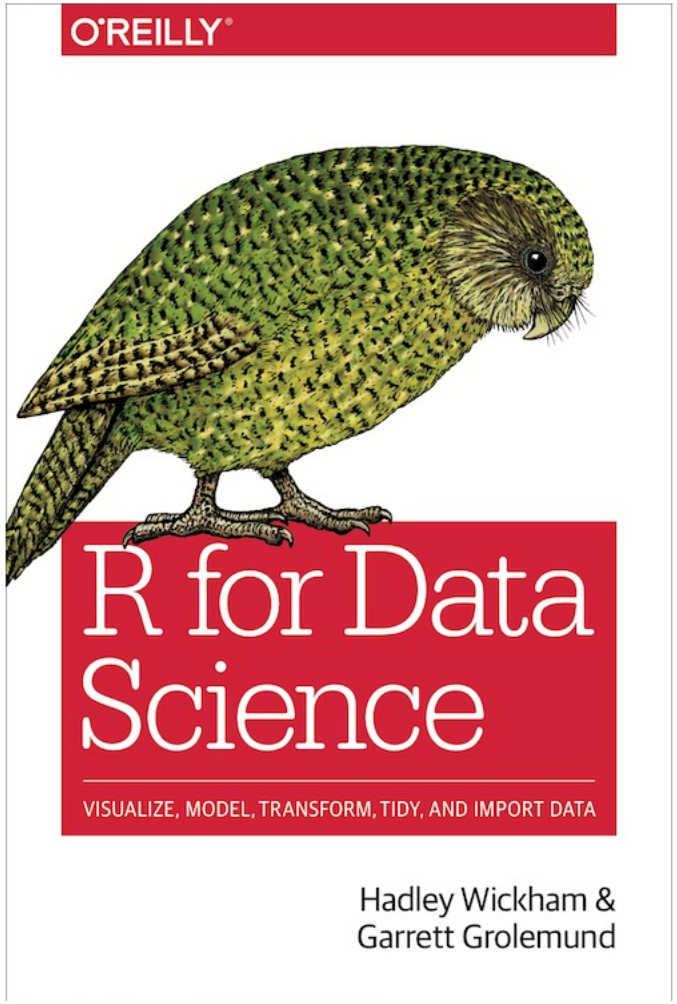
https://www.tidyverse.org/

http://adv-r.had.co.nz/ Advanced R
https://r4ds.had.co.nz/ R for data science

Data analysis workflow

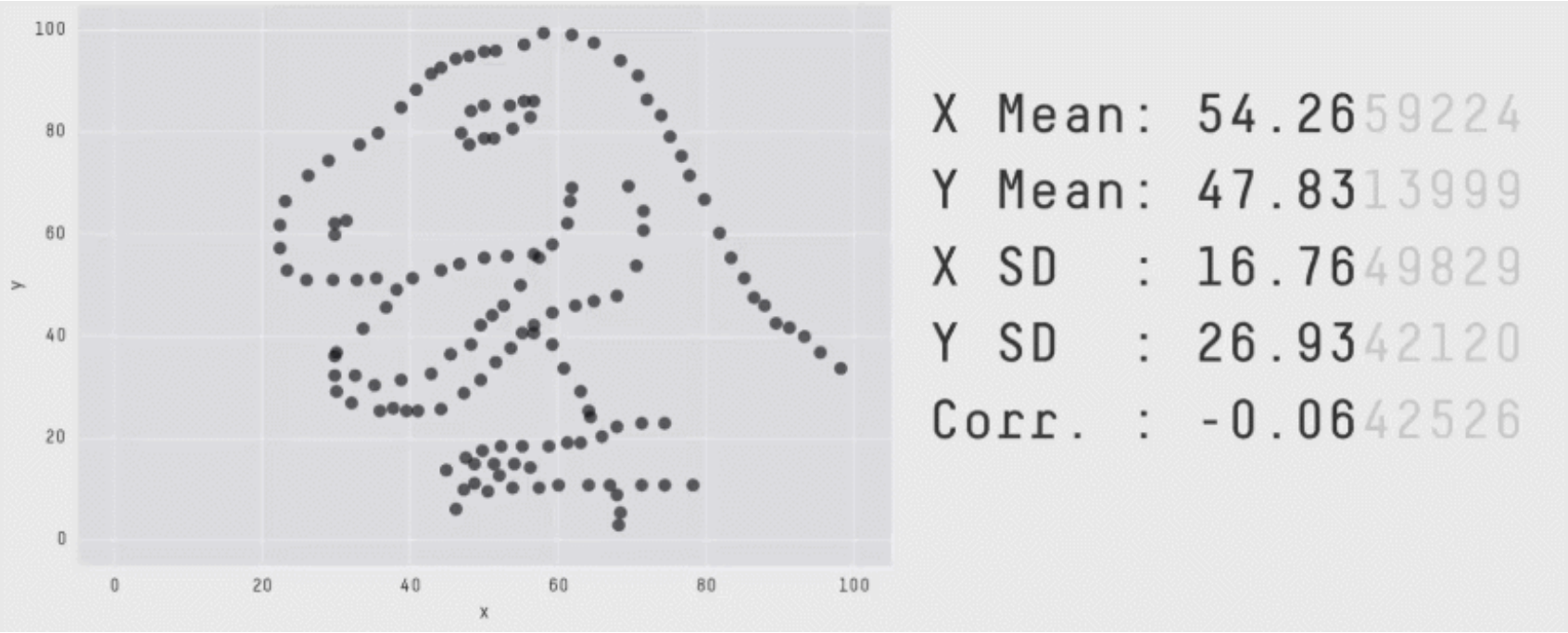


Tidying the data can take 80% of your time



R for data science by Hadley Wickham & Garrett Grolemund
http://r4ds.had.co.nz/

Data visualization



https://www.r-bloggers.com/the-datasaurus-dozen/

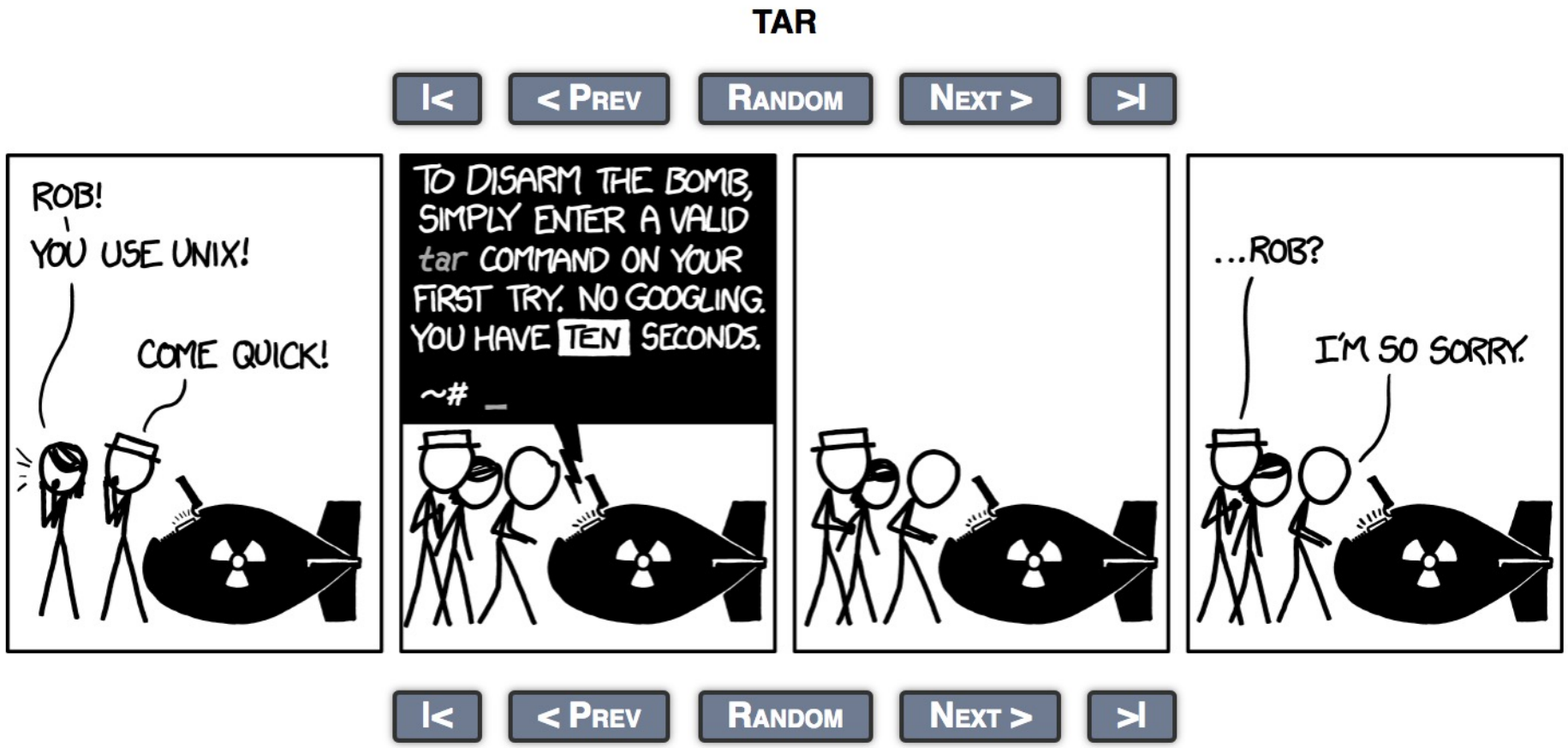
What people think we do

Credit: Torsten Seemann

A typical day of my life as a computational biologist

- Installing software
- Googling (how to and error message etc).
- Read manuals of bioinformatics tools.
- Converting file formats.
- Tidying the data.
- Real analysis (plotting etc) 20%

Google is how we learn and do things

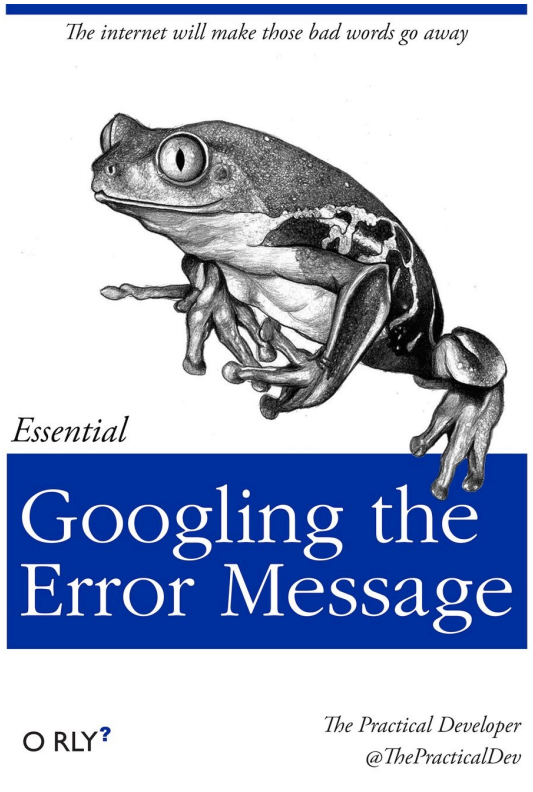


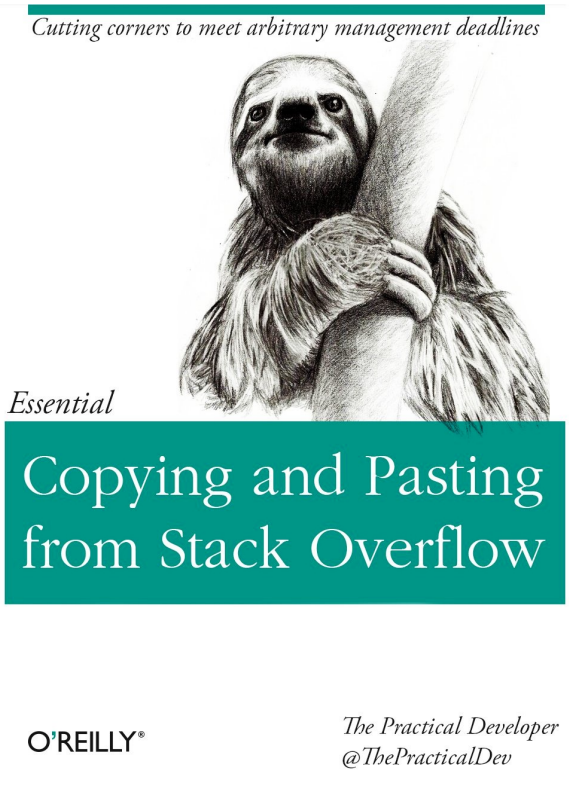
Google tricks

- Add `[r]` to search R programming related pages. e.g., “distance measurement [r].” you can do it with other languages too: “rotate x axis labels [python].” Search “patchwork [r]” will find you the R package.
- Use quotations `" "` to search for the exact phrase.
- Add a tilde `~` in front of a word to find synonyms.
- Exclude terms with a minus `-` symbol.
- Search specific sites with `site:` . “heatmap site:https://support.bioconductor.org” will search heatmap inside the bioconductor support website.
- Define a filetype by: `heatmap filetype:pdf` it will only give you PDF files in the results.

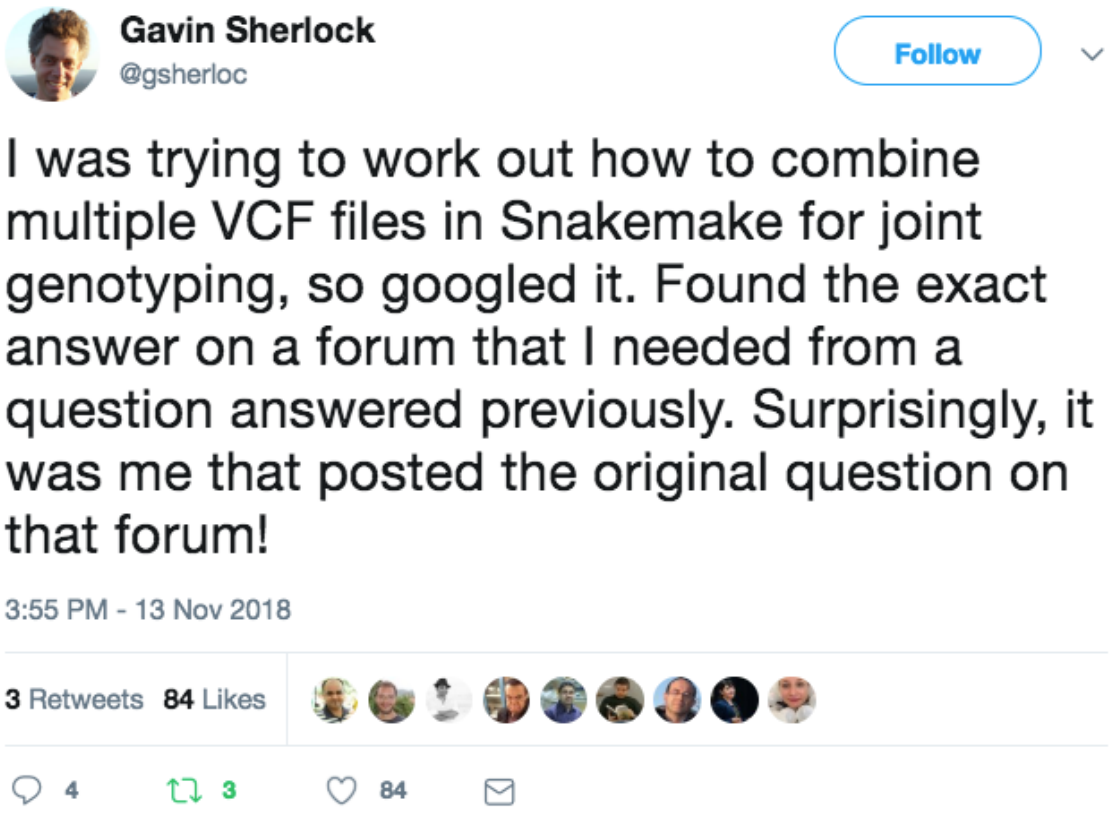
Ask for help

- SeqAnswer
- Biostars
- Stack overflow
- Biconductor help





I am not lying



Use excel with precaution

BBC | Sign in | Home | News | Sport | Reel | Worklife | Travel

NEWS

Home | US Election | Coronavirus | Video | World | US & Canada | UK | Business | Tech | Science | Stories

Tech

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

5 October

Coronavirus pandemic

The problem is that PHE's own developers picked an old file format to do this - known as XLS.

As a consequence, each template could handle only about 65,000 rows of data rather than the one million-plus rows that Excel is actually capable of.

And since each test result created several rows of data, in practice it meant that each template was limited to about 1,400 cases.

When that total was reached, further cases were simply left off.

https://www.bbc.com/news/technology-54423988

Retraction Watch
@RetractionWatch

Follow

An Excel screw-up leads to a retraction. "This technological issue caused rows to shift and the data from the different groups got mixed up." sciencedirect.com/science/articl …

12:27 PM - 6 Aug 2018

17 Retweets **21** Likes

9 17 21

https://www.sciencedirect.com/science/article/pii/S0018506X18302599?via%3Dihub

Excel converts gene names to dates

excel.gene2date.xls

	A (gene names)	B (internal date format)	C (default date format)	D	E (gene names)	F (internal date format)	G (default date format)	H	I (gene names)	J (internal date format)	K (default date format)
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec		OCT2	36069	2-Oct		SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec		OCT3	36070	3-Oct		SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar		OCT4	36071	4-Oct		SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar		OCT6	36073	6-Oct		SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar		OCT7	36074	7-Oct		SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov		SEP-1	36038	1-Sep		SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov		SEP-2	36039	2-Sep		SEPT9	36046	9-Sep
15					SEP1	36038	1-Sep				

Sheet1 Sheet2

Ready Sum=0 SCRL CAPS NUM

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-80

Use R packages:
Readxl, Janitor
To work with
excel sheets

http://blogs.nature.com/naturejobs/2017/02/27/escape-gene-name-mangling-with-escape-excel/

Reproducibility crisis

http://biobungalow.weebly.com/bio-bungalow-blog/everybody-knows-the-scientific-method

Most computational research is not reproducible.

I don’t know of a systematic study, but of papers that I read, approximately 95% fail to include details necessary for replication.

It’s very hard to build off of research like this.

(There’s a lot more to say about repeatability, reproducibility and replicability than I can fit in here…)

An example

- The Importance of Reproducible Research in High-Throughput Biology.
- https://www.youtube.com/watch?v=7gYIs7uYbMo
- By Dr.Keith A. Baggerly from MD Anderson Cancer Center.
- Highly recommend, Keith is very fun.

Flawed Cancer Trial at Duke Sparks Lawsuit

By Jennifer Couzin-Frankel | Sep. 9, 2011 , 3:38 PM

A dozen plaintiffs have filed a **lawsuit** against Duke University and administrators, researchers, and physicians there, alleging that they engaged in fraudulent and negligent behavior when they enrolled cancer patients in a clinical trial compromised by faulty data. The lawsuit, filed Wednesday in a North Carolina court, comes 14 months after a **scandal erupted at Duke** that finally exposed the extent of the trial's problems: in July 2010, Duke oncologist Anil Potti, whose work was central to the trial, admitted that he had embellished his resume and later **resigned**.

Method matters

RESEARCH ARTICLE

Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors

Nathaniel D. Anderson[1,2], Richard de Borja[1,*], Matthew D. Young[3,*], Fabio Fuligni[1,*], Andrej Rosic[1], Nicola D. Roberts[3], Simo…
+ See all authors and affiliations

Science 31 Aug 2018:
Vol. 361, Issue 6405, eaam8419
DOI: 10.1126/science.aam8419

Detection of gene fusions

We detected gene fusions in regions of genomic complexity using an approach that integrates multiple independent fusion algorithms, and then removed those found in normal tissue. Putative fusions were validated by de novo assembly. A total of 1277 normal (nonneoplastic) samples from 43 different tissues were obtained from the NHGRI GTEx consortium (database version 4) and used to remove artifacts. All fusions were visually inspected if one or both genes involved chromoplexy or were adjacent (up to 1 Mbp). Fusions were further filtered by quality of the realigned transcript, breakpoint coverage, and gene expression.

Why reproducibility is hard?

Why reproducibility is hard?

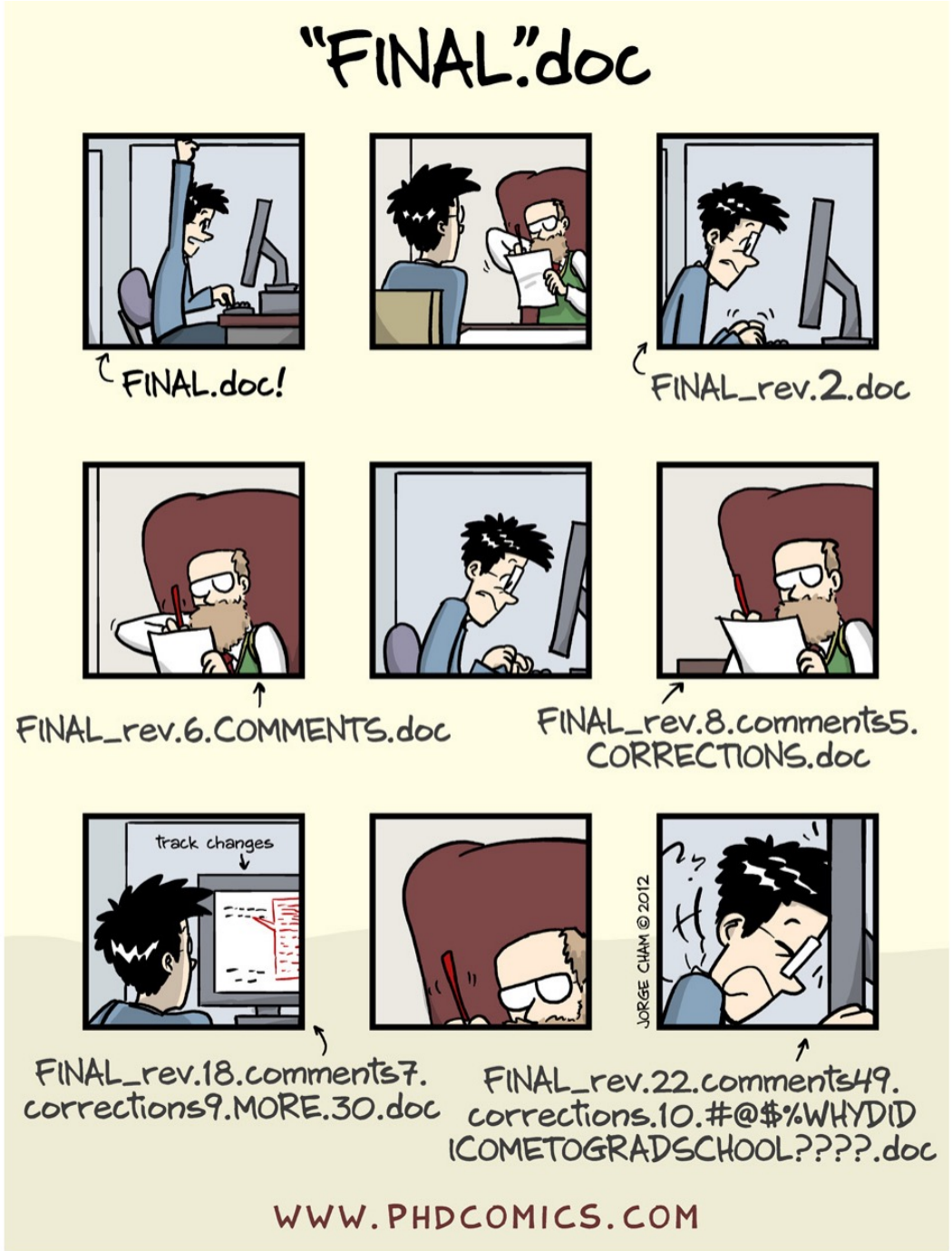
- 1. no raw data are available.
- 2. scripts/data available upon reasonable request ☺
- 3. lack of method description.
- 4. versions of the tools are different. (e.g. R/python/bioinformatics tools)
- 5. different machines (unix vs windows).

If it is so hard, should you care?

- Keep this in mind: You are going to do the same analysis for sure in the future yourself!
- This is for your own benefit.
- I want to make sure my analysis is reproducible because I am discovering drug targets for patients!

How to ensure reproducibility

- Git version control
- Jupyter/R Notebook, documentation
- Containers (docker, singularity, biocontainers https://biocontainers.pro/)
- Unit test
- Continuous Integration/development CI/CD (Travis CI, github action)


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

Version control

- Git
- Github
- Gitlab

https://docs.github.com/en/get-started/quickstart/git-and-github-learning-resources

Jupyter Notebook

jupyter nbviewer

JUPYTER FAQ

notebook / docs / source / examples / Notebook

Running Code

First and foremost, the Jupyter Notebook is an interactive environment for writing and running code. The notebook is capable of running code in a wide range of languages. However, each notebook is associated with a single kernel. This notebook is associated with the IPython kernel, therefor runs Python code.

Code cells allow you to enter and run code

Run a code cell using `Shift-Enter` or pressing the button in the toolbar above:

In [2]:
```
a = 10
```

In [3]:
```
print(a)
```

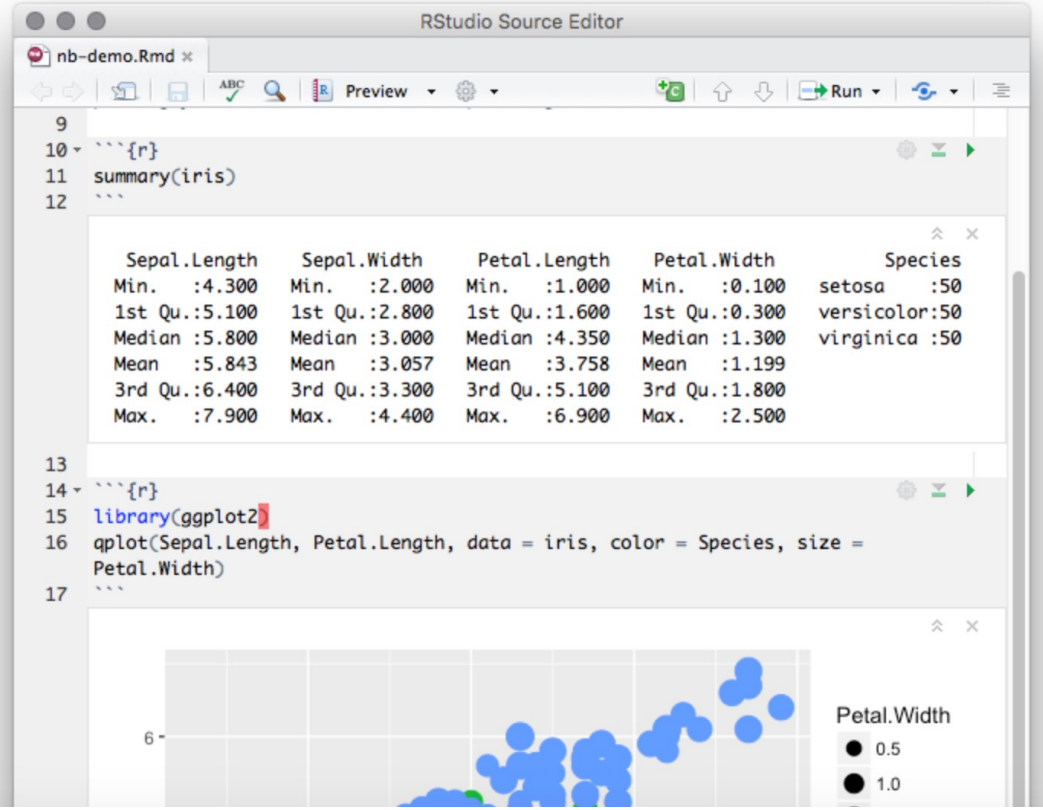
```
10
```

There are two other keyboard shortcuts for running code:

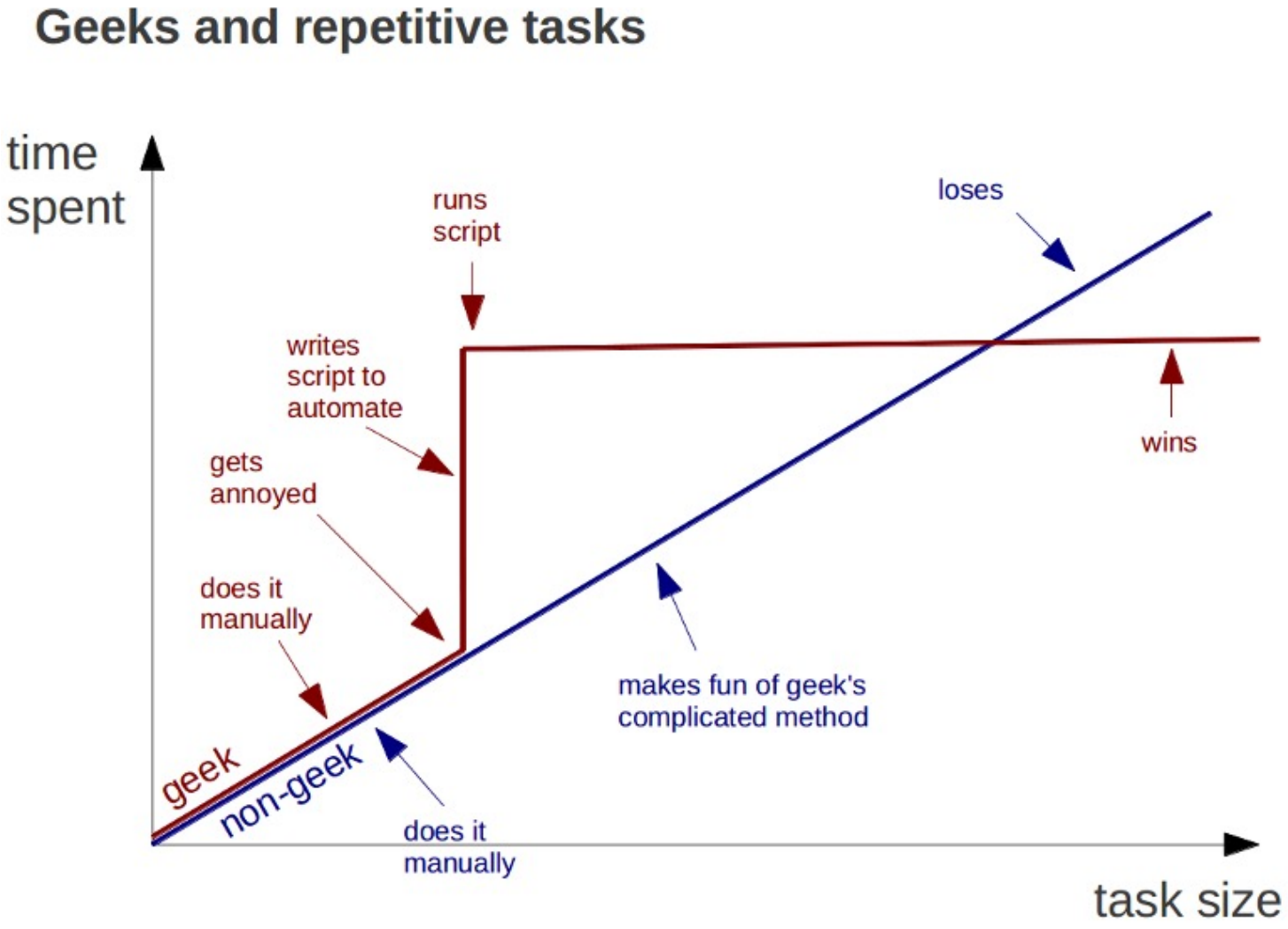
- `Alt-Enter` runs the current cell and inserts a new one below.
- `Ctrl-Enter` run the current cell and enters command mode.

R notebook/markdown

An R Notebook is an R Markdown document with chunks that can be executed independently and interactively, with output visible immediately beneath the input.



Automation makes your research more reproducible AND saves you time in the long run



Computers are good at repetitive work

Good Side effect of automation

- The best documentation is automation
- Write scripts for everything unless it is not possible. (manual editing, document, document, document!)
- Markdown, MKdocs https://www.mkdocs.org/

Credit to someone in the twitter-verse ☺

Tips for automation

- 1. if you have a repetitive simple task, put them in to a shell script: my_routine.sh.
- 2. good old GNU make
- 3. more recent snakemake, nextflow, WDL etc.

Awesome Pipeline

A curated list of awesome pipeline toolkits inspired by Awesome Sysadmin

Pipeline frameworks & libraries

- ActionChain - A workflow system for simple linear success/failure workflows.
- Adage - Small package to describe workflows that are not completely known at definition time.
- Airflow - Python-based workflow system created by AirBnb.
- Anduril - Component-based workflow framework for scientific data analysis.
- Antha - High-level language for biology.
- AWE - Workflow and resource management system with CWL support
- Bds - Scripting language for data pipelines.
- BioMake - GNU-Make-like utility for managing builds and complex workflows.
- BioQueue - Explicit framework with web monitoring and resource estimation.
- Bioshake - Haskell DSL built on shake with strong typing and EDAM support
- Bistro - Library to build and execute typed scientific workflows.

https://github.com/pditommaso/awesome-pipeline

519

Snakemake—a scalable bioinformatics workflow engine

Publication Article in **Bioinformatics**, published October 2012

Authors Johannes Köster, Sven Rahmann

More details

{wdl}

nextflow

Docker

- Why docker?
- Imagine you are working on an analysis in R and you send your code to a friend. Your friend runs exactly this code on exactly the same data set but gets a slightly different result. This can have various reasons such as a different operating system, a different version of an R package, etc. Docker is trying to solve problems like that.
- Think it as a virtual machine!
- This just happened between me and my colleagues who used a different version of R packages!

https://cyverse-cybercarpentry-container-workshop-2018.readthedocs-hosted.com/en/latest/docker/dockerintro.html
https://ropenscilabs.github.io/r-docker-tutorial/01-what-and-why.html

conda and biocoda

Conda

CONDA

Package, dependency and environment management for any language—Python, R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, FORTRAN

MENU

nature methods

Correspondence | Published: 02 July 2018

Bioconda: sustainable and comprehensive software distribution for the life sciences

Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris & Johannes Köster The Bioconda Team

Nature Methods **15**, 475–476 (2018) | Download Citation

PLOS | COMPUTATIONAL BIOLOGY

Browse | Publish | About

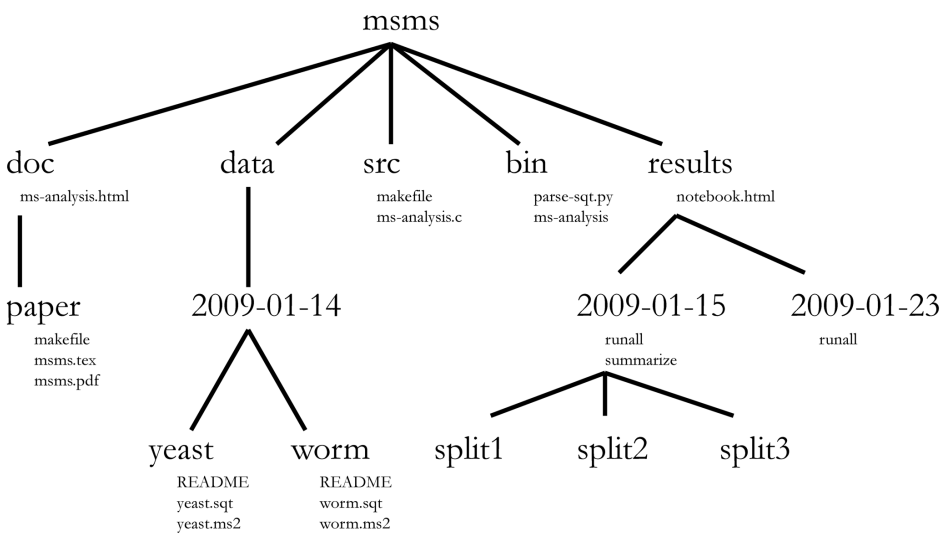
OPEN ACCESS

EDUCATION

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble

Published: July 31, 2009 • https://doi.org/10.1371/journal.pcbi.1000424



PLOS | COMPUTATIONAL BIOLOGY

Browse | Publish | About

OPEN ACCESS

PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

Published: June 22, 2017 • https://doi.org/10.1371/journal.pcbi.1005510

PLOS | BIOLOGY
FIFTEENTH ANNIVERSARY

Browse | Publish | About

OPEN ACCESS

COMMUNITY PAGE

Best Practices for Scientific Computing

Greg Wilson, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, Kathryn D. Huff, Ian M. Mitchell, Mark D. Plumbley, Ben Waugh, Ethan P. White, Paul Wilson

Workflowr for fully reproducible analysis

Belinda Phipson
@BelindaPhipson

Check out this incredibly impressive workflowr analysis website showcasing @JovMaksimovic single cell analysis of paediatric lower airway. A lot of time and effort to ensure the analysis is reproducible. oshlacklab.com/paed-cf-cite-s...

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

biorxiv.org
Multimodal single cell analysis of the paediatric lower airwa...
Respiratory disease is a major cause of morbidity and mortality in children worldwide. Many childhood respiratory...

1:10 AM · Jun 24, 2022 · Twitter Web App

16 Retweets **2** Quote Tweets **60** Likes

paed-cf-cite-seq | Home | About | License

- Abstract
- Authors
- Analysis Overview
- Licenses
- Citations
- Version Information

Multimodal single cell analysis of the paediatric lower airway reveals novel immune cell phenotypes in early life health and disease

Jovana Maksimovic

2022-06-20

workflowr ✔

This site presents the code and results of the analyses described in the pre-print: *"Multimodal single cell analysis of the paediatric lower airway reveals novel immune cell phenotypes in early life health and disease"*.

All the code and results of this analysis are available from GitHub at https://github.com/Oshlack/paed-cf-cite-seq. To reproduce the complete analysis follow the instructions on the getting started page. The raw single cell RNA-seq and CITE-seq count data generated for this study can be downloaded as RDS files from DOI 10.5281/zenodo.6651465.

Follow the links below to view the different parts of the analysis.

Thursday, August 13, 2015

2 cents on coding from a bioinformatics beginner

One needs to be aware that:

1. **Computers make mistakes**. They can give you non-sense results and exit without error, so make extensive tests before running your code.

2. **Share your codes**. Even your codes are correct, you need to share them so that other people can look at them and may improve them.

3. **Make your codes reusable**. Do not hard code your scripts. If it takes a file path as input, make it as an argument in your scripts.

4. **Modulate your scripts.** Data could come in different stage of formats. Take ChIP-sequencing data analysis as an example, if you have a script that starts processing the data from fastq to the final peaks. You may want to modulate your scripts to two modules: one for mapping fastq to bam, and the other for bam to peaks. **Modulate your scripts** so that one can use your script when the data come in a bam format.

5. **Heavily comment your scripts**. It will not only make other people to understand your codes better, but also help the future you to understand what you did.

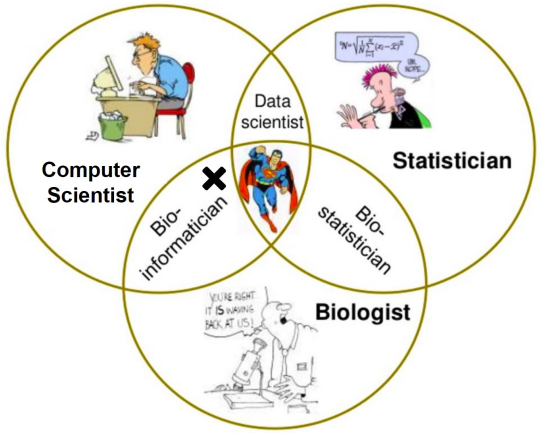
6. **You need to make your analysis reproducible**. Each step of your analysis should be documented in a markdown file. I say every step, yes, every command that you strike in the terminal getting the intermediate files need to be taken down. Moreover, how, when and where did you download the data need to be documented. This will save the future you! Many experienced programmers overlook this point.

http://crazyhottommy.blogspot.com/2015/08/2-cents-on-coding-from-bioinformatics.html

Key take-aways

- Learn Unix commands, python and R
- Google is the way. Now, we have Chat-GPT
- Be cautious with excel
- Git version control your code
- Have a consistent folder structure for projects' reproducible computing
- Learn by doing
- Focus on your strength: biology domain knowledge

Computational ***biologist***.



Acknowledgments

Verhaak Lab
Samir Amin

Titus Brown

Data Carpentry https://datacarpentry.org/

All the people who share their wisdom on the web
Thanks!

What questions do you have?