

# Introduction to Computational Biology

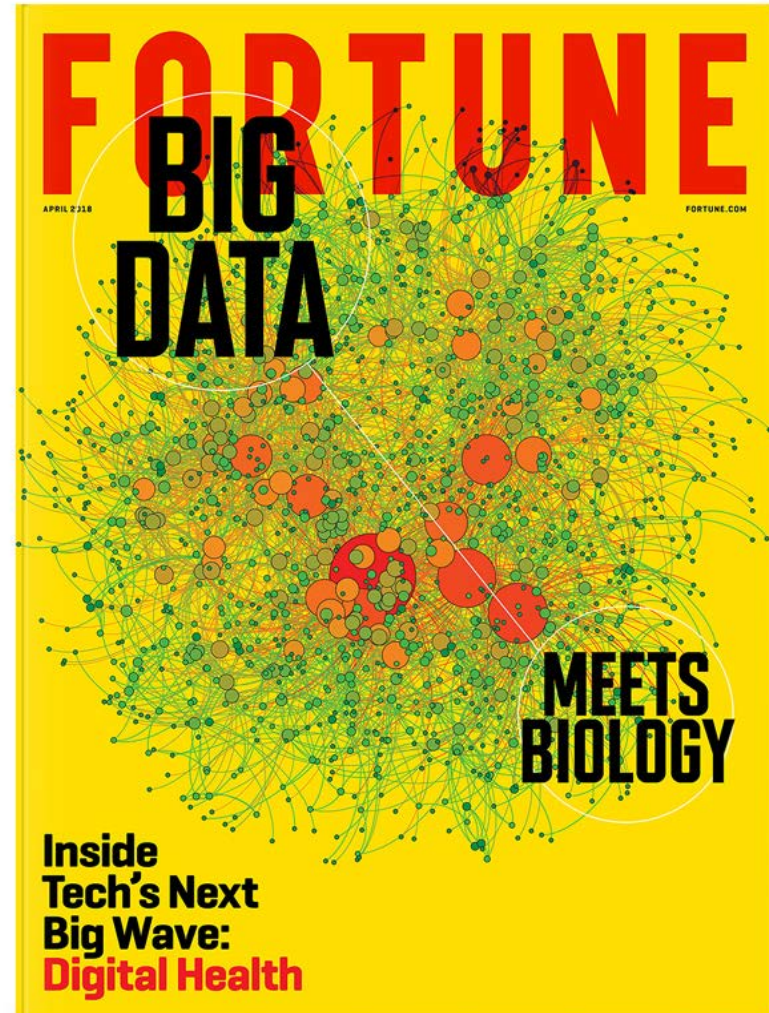
Lecture 0

BIOL 4590, BIOL 5590

Dr. Chris Bird

# Why are computational skills important for biologists?

- Increasing data size and complexity
- Increasing sophistication of statistical and mathematical analyses
- Transparency, reproducibility, and documentation



# Why should biologists be interested in developing your computational kung-fu?

- Automate impossibly tedious, monotonous, and lengthy tasks
- Increased rate and significance of discovery
- Career success
- Maximize potential



# Why am I teaching this class?

- Historical lack of computational courses for biologists
- Steep learning curve
- The days of easily succeeding in biological research without computational knowledge and skill are over



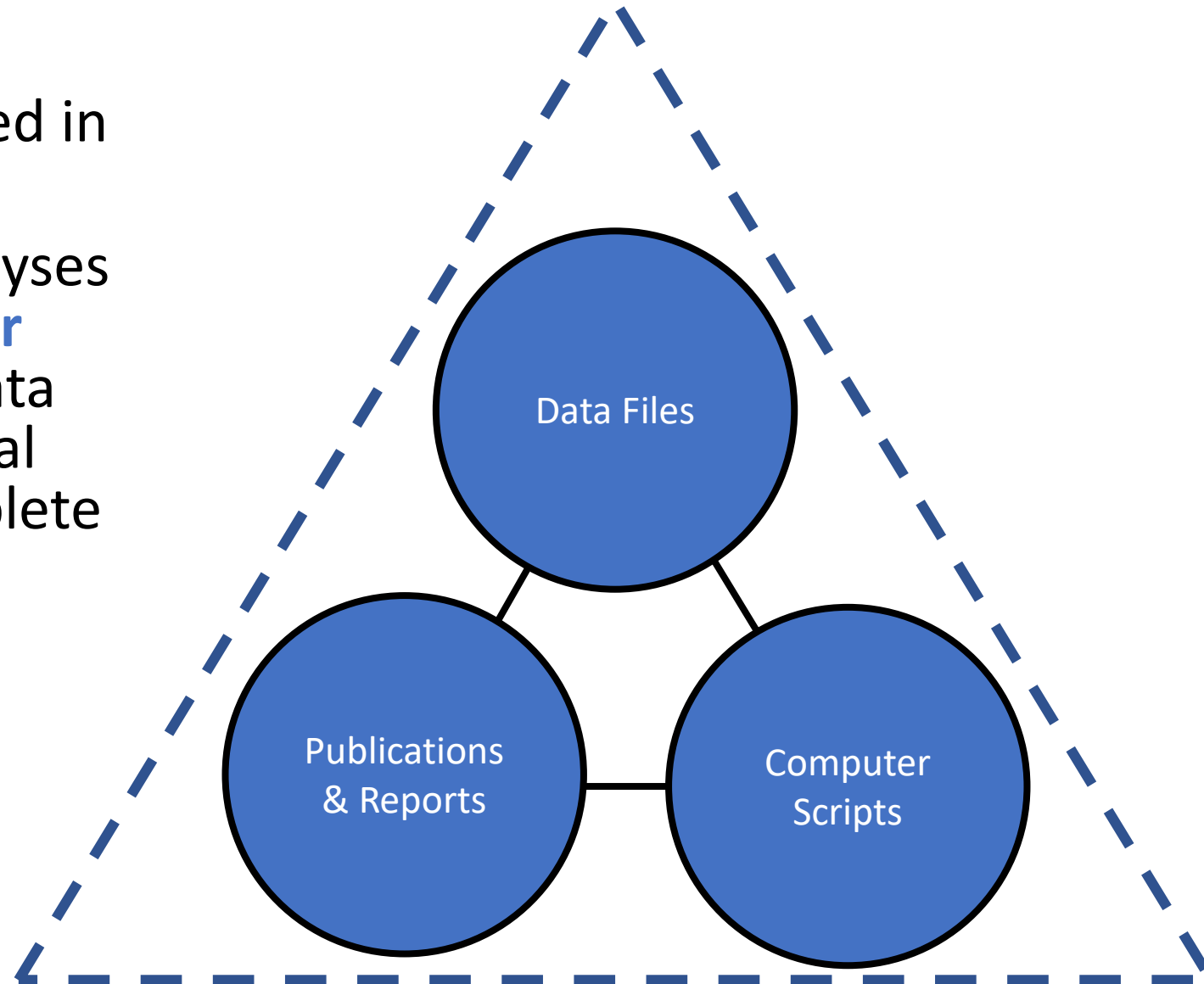
# If you so choose, I will show you the philosophy of data science

- Automation
  - Interconnection
  - Modularity
- Reproducibility
  - Organization
  - Comprehension
- Openness
- Simplicity
- Correctness



# Philosophy of Data Science

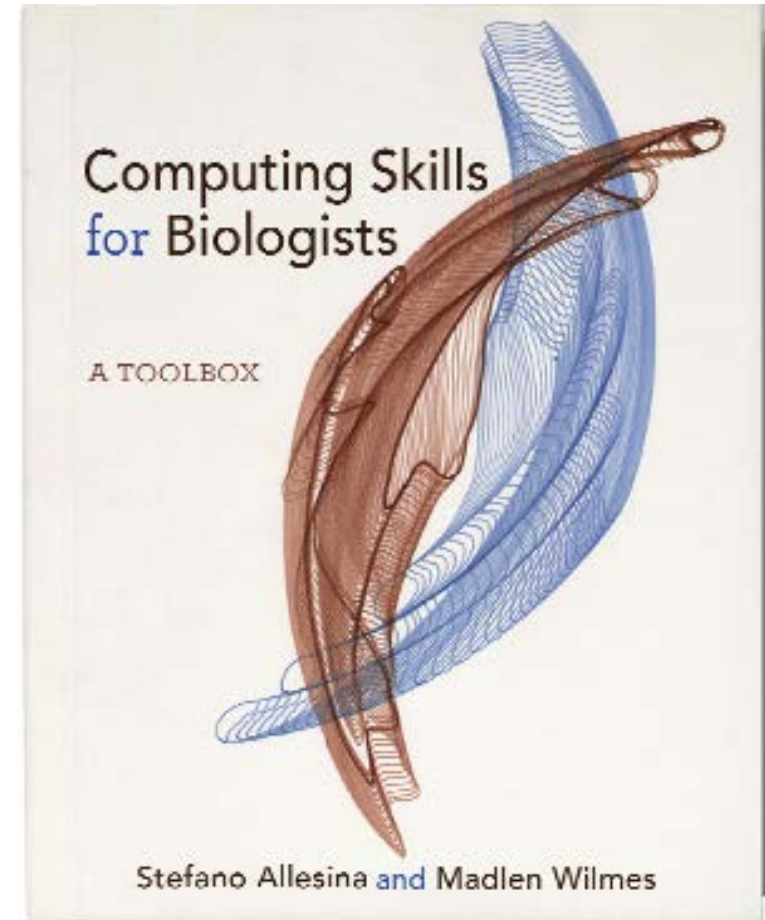
- All **data is digitized** and stored in files
- Data manipulations and analyses are documented in **computer scripts** that interface with data files and require no additional human intervention to complete analysis
- Data & scripts are published with the report and **openly accessible to all**





# We Will Follow The Text Book, Mostly

- Provides you with requisite breadth of tools at the expense of depth
- Showcase of Linux, Python, R
- Organized into 10 chapters, theoretically 1 per lecture
- Goal is to flatten your learning curve



<https://computingskillsforbiologists.com/>

# Learning Objectives

- Recognize, describe, and organize data into standard biological data structures
- Locate scientific data repositories and extract data
- Operate UNIX/LINUX computers from command line
- Construct and modify computer programming/scripting logic structures for processing biological data
- Use version control software (git)
- Describe and use regular expressions to query data
- Typeset with LaTeX or Markdown
- Use the most popular open-source tools for biological data manipulation
  - Shell scripting (bash)
  - Scientific computing (python)
  - Statistical computing (R)
  - Tool repositories



# Syllabus & Course Organization

- Syllabus is on blackboard and github
- 3 Parts of Course
  - Unix, Python, R
- Additional skills
  - Version control with git
  - Typesetting with LaTeX, markdown

## Undergraduates:

ACTIVITY	% of FINAL GRADE
Participation	15
Assignments	40
Exam 1	12.5
Exam 2	12.5
Final Exam	20

## Graduates:

ACTIVITY	% of FINAL GRADE
Participation	10
Assignments	20
Exam 1	10
Exam 2	10
Final Project	50*

# Lectures

- Environment for you to learn new concepts
- Hands-on with computers
- Power-point driven
  - Sometimes remotely delivered
- Independent exercises w/ MS Forms through GitHub

# Assignments

- Generally due each week
  - See syllabus
- Scripts will be submitted through GitHub classroom
- For now, question-answer based work will be conducted with a MS Form “quiz”

# Final Project (Graduate Students)

- Automate the processing and analysis of your data
- Document work on GitHub
- Report written in LaTeX or Markdown
  - State problem/challenge
  - Describe strategy to solve
  - Describe how code works
  - 10 min presentation during Final Exam Period
- Wk 3: Project idea
- Wk 5: Plan/Outline
- Wk 6: GitHub Repo
- Wk 7: Commit working function
- Wk 8: Commit 2 working functions w data input and output
- Wk 11: Draft/ progress report
- Wk 14: Final report, Working code and data on GitHub
- Final Exam: Oral pres

Questions?

# Biological Data

Lecture 0.1

BIOL 4590, BIOL 5590

Dr. Chris Bird



# Big Data Biology

- Massive amounts of data
- Associated tools, processes, procedures
- Volume, velocity, acceleration
- Goal is to tame the data
- Examples: DNA, climate, weather, remote sensing, GIS, all “omics”, populations

EMILY SINGER SCIENCE 10.11.13 09:30 AM

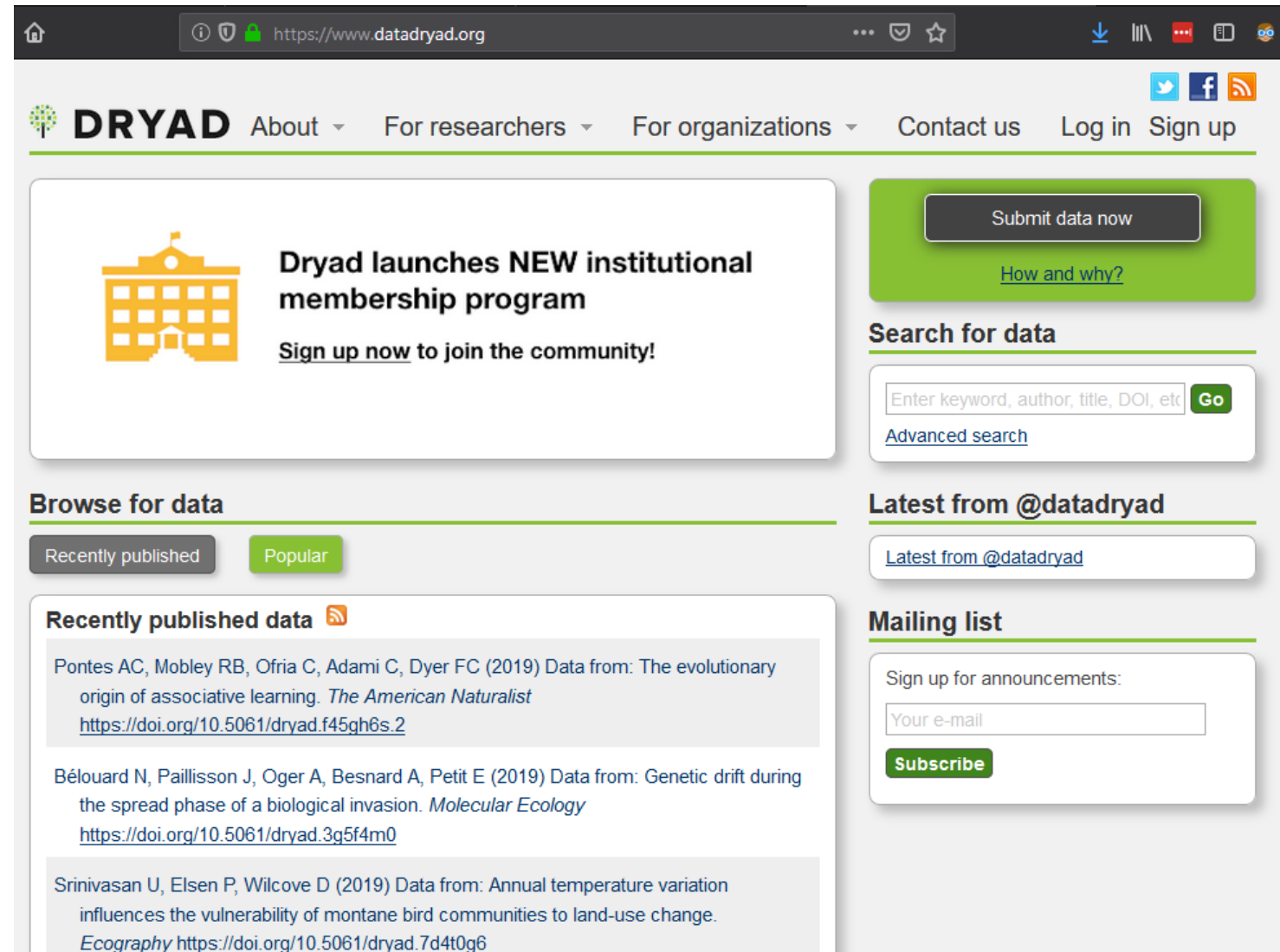
## BIOLOGY'S BIG PROBLEM: THERE'S TOO MUCH DATA TO HANDLE



..and not enough biologists  
with the motivation,  
interest, and/or skill to  
address the issue

# Biological Data Repositories for Data Big and Small

- Data associated with scientific papers should be published
  - Owned by the people
  - Should be freely available
  - Promotes acceleration of knowledge generation
- All Types of Data
  - [www.datadryad.com](http://www.datadryad.com)
- DNA & Proteins
  - <https://www.ncbi.nlm.nih.gov/>
- GIS
  - [https://data.usgs.gov/datacatalog/#fq=dataType%3A\(collection%20R%20non-collection\)&q=%3A\\*](https://data.usgs.gov/datacatalog/#fq=dataType%3A(collection%20R%20non-collection)&q=%3A*)
- Oceanographic
  - <https://data.noaa.gov/datasetsearch/>
- Too many to list



The screenshot shows the Dryad website homepage. The browser address bar displays <https://www.datadryad.org>. The website features a navigation bar with links for 'About', 'For researchers', 'For organizations', 'Contact us', 'Log in', and 'Sign up'. A prominent green button labeled 'Submit data now' is accompanied by a link 'How and why?'. A central announcement states 'Dryad launches NEW institutional membership program' with a 'Sign up now' link. Below this, a 'Browse for data' section offers 'Recently published' and 'Popular' filters. The 'Recently published data' list includes three entries with their titles and DOIs. On the right, there is a 'Search for data' box with a 'Go' button and a link to 'Advanced search'. Below that is a 'Latest from @datadryad' section with a link to the latest tweets. At the bottom right is a 'Mailing list' sign-up form with a 'Subscribe' button.

**DRYAD** About ▾ For researchers ▾ For organizations ▾ Contact us Log in Sign up

**Dryad launches NEW institutional membership program**  
Sign up now to join the community!

**Submit data now**  
[How and why?](#)

**Search for data**  
Enter keyword, author, title, DOI, etc **Go**  
[Advanced search](#)

**Browse for data**  
Recently published Popular

**Recently published data**

- Pontes AC, Mobley RB, Ofria C, Adami C, Dyer FC (2019) Data from: The evolutionary origin of associative learning. *The American Naturalist*  
<https://doi.org/10.5061/dryad.f45gh6s.2>
- Bélouard N, Paillisson J, Oger A, Besnard A, Petit E (2019) Data from: Genetic drift during the spread phase of a biological invasion. *Molecular Ecology*  
<https://doi.org/10.5061/dryad.3g5f4m0>
- Srinivasan U, Elsen P, Wilcove D (2019) Data from: Annual temperature variation influences the vulnerability of montane bird communities to land-use change. *Ecography* <https://doi.org/10.5061/dryad.7d4t0g6>

**Latest from @datadryad**  
[Latest from @datadryad](#)

**Mailing list**  
Sign up for announcements:  
Your e-mail  
**Subscribe**

# Let's Explore a Data Set Published in Dryad

- [www.datadryad.com](http://www.datadryad.com)
- Find Data from:
  - Direct and indirect effects of sexual signal loss on female reproduction in the Pacific field cricket (*Teleogryllus oceanicus*)
- Download the data and view it in MS Excel
  - It is important to look at data and understand how it is organized

Data from: Direct and indirect effects of sexual signal loss on female reproduction in the Pacific field cricket (*Teleogryllus oceanicus*)



Heinen-Kay J, Strub D, Balenger S, Zuk M

Date Published: August 29, 2019

DOI: <https://doi.org/10.5061/dryad.v732vb1>

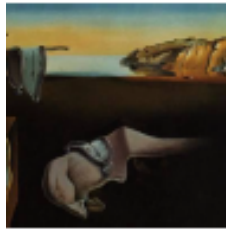


## Files in this package

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the [Dryad Terms of Service](#). To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data.  

Title	Data for Heinen-Kay et al. Sexual signal loss and female reproduction
Downloaded	3 times
Description	Data for (1) comparison of flatwing and normal-wing homozygous female reproductive tissue, (2) offspring production of flatwing and normal-wing females, and (3) reproductive tissue comparison between populations and acoustic treatments
Download	<a href="#">Data for Heinen-Kay et al. Sexual signal loss and female reproduction.xlsx (36.79 Kb)</a>
Details	<a href="#">View File Details</a>

# Tidy Data ([Wickham 2014](#))



---

*Journal of Statistical Software*

August 2014, Volume 59, Issue 10.

<http://www.jstatsoft.org/>

---

- Each row is the “smallest unit of observation”
  - Ex: an individual cricket
- Each column is a variable or dimension of information about the units of observation
  - Ex: somatic mass

## Tidy Data

Hadley Wickham  
RStudio

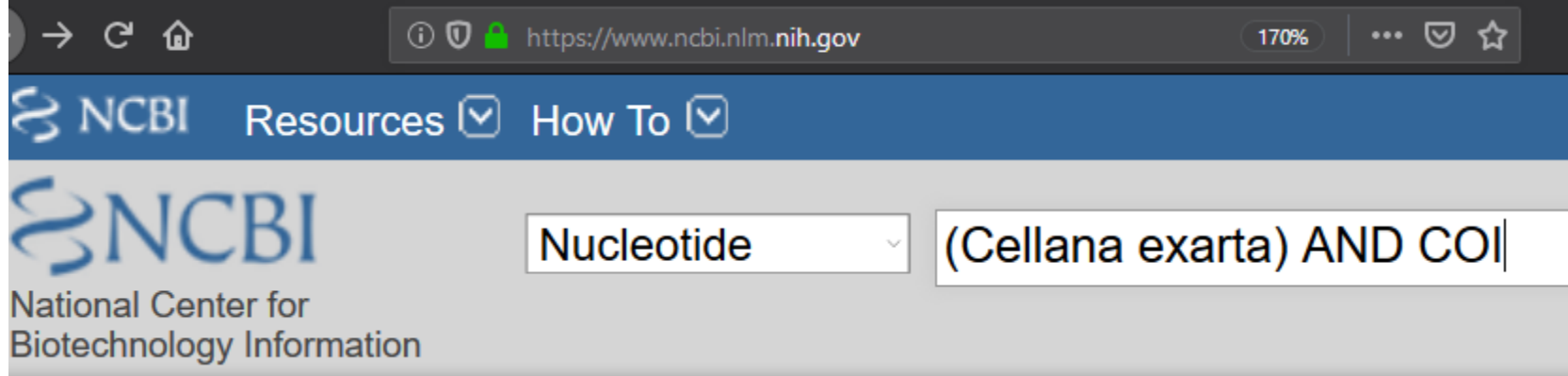
---

### Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

# Common Data Formats & Structures are Not Always Tidy

- <https://www.ncbi.nlm.nih.gov/>
- Conduct the following search



# GenBank Supports Several Formats, None Are Tidy

- <https://www.ncbi.nlm.nih.gov/>
- Switch to FASTA (text)

NCBI Resources ☒ How To ☒

Nucleotide

Nucleotide (Cellana exarta) AND COI

Create alert Advanced

Species Summary 20 per page Sort by Default order

Animals (1,300) Customize ...

Molecule types

genomic

DNA/RNA (1,300) Customize ...

Source databases

Format

- ☒ Summary
- ☐ GenBank
- ☐ GenBank (full)
- ☐ FASTA
- ☐ FASTA (text)
- ☐ ASN.1
- ☐ Revision History
- ☐ Accession List
- ☐ GI List

Send to

306

<< First < Prev Page 1 of 66 Next > La

The following term was not found in Nucleotide: exarta.

```
>AB263731.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen voucher: NUGB-L694 (Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCAGCCTT
TTGTTATGATTTTCTTTTAGTAATACCAATGATAATTGGGGGTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTAAATAATATGAGGTTTGGTTACTGGTTCCTCTCTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTT
CTAGAAATGTGGCCCATTCCTGGTCTTCTGTTGATTGGCTATTTTCTCTTCATTGGCTGGTATTTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTCTCTCTCTGTGTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTAATACCTGTTTTTTGACCTGGAGGAGGAGG
GGACCCCATTTATATCAACATTGTTT

>AB263730.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen voucher: NUGB-L693 (Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCAGCCTT
TTGTTATGATTTTCTTTTAGTAATACCAATGATAATTGGGGGTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTAAATAATATGAGGTTTGGTTACTGGTTCCTCTCTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTT
CTAGAAATGTGGCCCATTCCTGGTCTTCTGTTGATTGGCTATTTTCTCTTCATTGGCTGGTATTTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTCTCTCTCTGTGTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTAATACCTGTTTTTTGACCTGGAGGAGGAGG
GGACCCCATTTATATCAACATTGTTT

>AB263729.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen voucher: NUGB-L692 (Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCAGCCTT
TTGTTATGATTTTCTTTTAGTAATACCAATGATAATTGGGGGTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTAAATAATATGAGGTTTGGTTACTGGTTCCTCTCTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTT
CTAGAAATGTGGCCCATTCCTGGTCTTCTGTTGATTGGCTATTTTCTCTTCATTGGCTGGTATTTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTCTCTCTCTGTGTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTAATACCTGTTTTTTGACCTGGAGGAGGAGG
GGACCCCATTTATATCAACATTGTTT

>AB263728.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen voucher: NUGB-L691 (Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCAGCCTT
TTGTTATGATTTTCTTTTAGTAATACCAATGATAATTGGGGGTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTAAATAATATGAGGTTTGGTTACTGGTTCCTCTCTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTT
CTAGAAATGTGGCCCATTCCTGGTCTTCTGTTGATTGGCTATTTTCTCTTCATTGGCTGGTATTTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTTCGTTGGCGAGGTCTTCAGTTTGAA
CGTCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTCTCTCTCTGTGTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTAATACCTGTTTTTTGACCTGGAGGAGGAGG
GGACCCCATTTATATCAACATTGTTT
```



# Common DNA Data Format

- <https://www.ncbi.nlm.nih.gov/>
- Switch to FASTA (text)
  - Wikipedia is an excellent resource for describing data formats
- FASTA Format
  - DNA
  - Lines beginning with > contain the smallest unit of observation
  - Lines that don't begin with > contain information, each character is a dimension of the units of observation

```
>AB263731.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L694 (Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAAGTATGTTAATTCGGGGCT
GAATTAGGTCAACCTGGTTCCTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTAGTAATACCAATGATAATTGGGGGTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTGGTTACTGGTTCCTCTTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCCTGGTCTTCTGTTGATTGGCTATTTTTCTCTTCATTGGCTGGTATTTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTCTCTTCTCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAAACCTGTTTTTTTACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT
```

```
>AB263730.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L693 (Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAAGTATGTTAATTCGGGGCT
GAATTAGGTCAACCTGGTTCCTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTAGTAATACCAATGATAATTGGGGGTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTGGTTACTGGTTCCTCTTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCCTGGTCTTCTGTTGATTGGCTATTTTTCTCTTCATTGGCTGGTATTTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTCTCTTCTCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAAACCTGTTTTTTTACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT
```

```
>AB263729.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L692 (Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAAGTATGTTAATTCGGGGCT
GAATTAGGTCAACCTGGTTCCTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTAGTAATACCAATGATAATTGGGGGTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTGGTTACTGGTTCCTCTTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCCTGGTCTTCTGTTGATTGGCTATTTTTCTCTTCATTGGCTGGTATTTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTCTCTTCTCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAAACCTGTTTTTTTACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT
```

```
>AB263728.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L691 (Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAAGTATGTTAATTCGGGGCT
GAATTAGGTCAACCTGGTTCCTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTAGTAATACCAATGATAATTGGGGGTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTGGTTACTGGTTCCTCTTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCCTGGTCTTCTGTTGATTGGCTATTTTTCTCTTCATTGGCTGGTATTTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTTCGTTGGCGAGGTCTTCAGTTTGAA
CGTCTACCTTTGTTTGATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTCTCTTCTCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAAACCTGTTTTTTTACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT
```


# Data Formats

- I will emphasize Tidy format
- Many fields of Biology have their own particular data formats
- There are tools available for handing and converting among data formats
- Some data formats are intimidating, at first
  - There will exist published descriptions of these
  - Duckduckgo: sam specification
    - This is a common “big data” format for next generation sequencer data
  - Take a deep breath, it’s never as intimidating as it seems

# Repositories Can Include Scripts for Processing, Analyzing, & Visualizing Data

- [www.datadryad.com](http://www.datadryad.com)
- Find Data from:
  - Meta-analyzing the likely cross-species responses to climate change
- Explore the files
  - \*.xls, \*.txt,
  - The extension indicates file format NOT data format
- R script
  - R is a statistical computer language
  - This file will analyze the data exactly the way it was reported in the publication

## Files in this package

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the [Dryad Terms of Service](#). To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data.  [OPEN DATA](#)

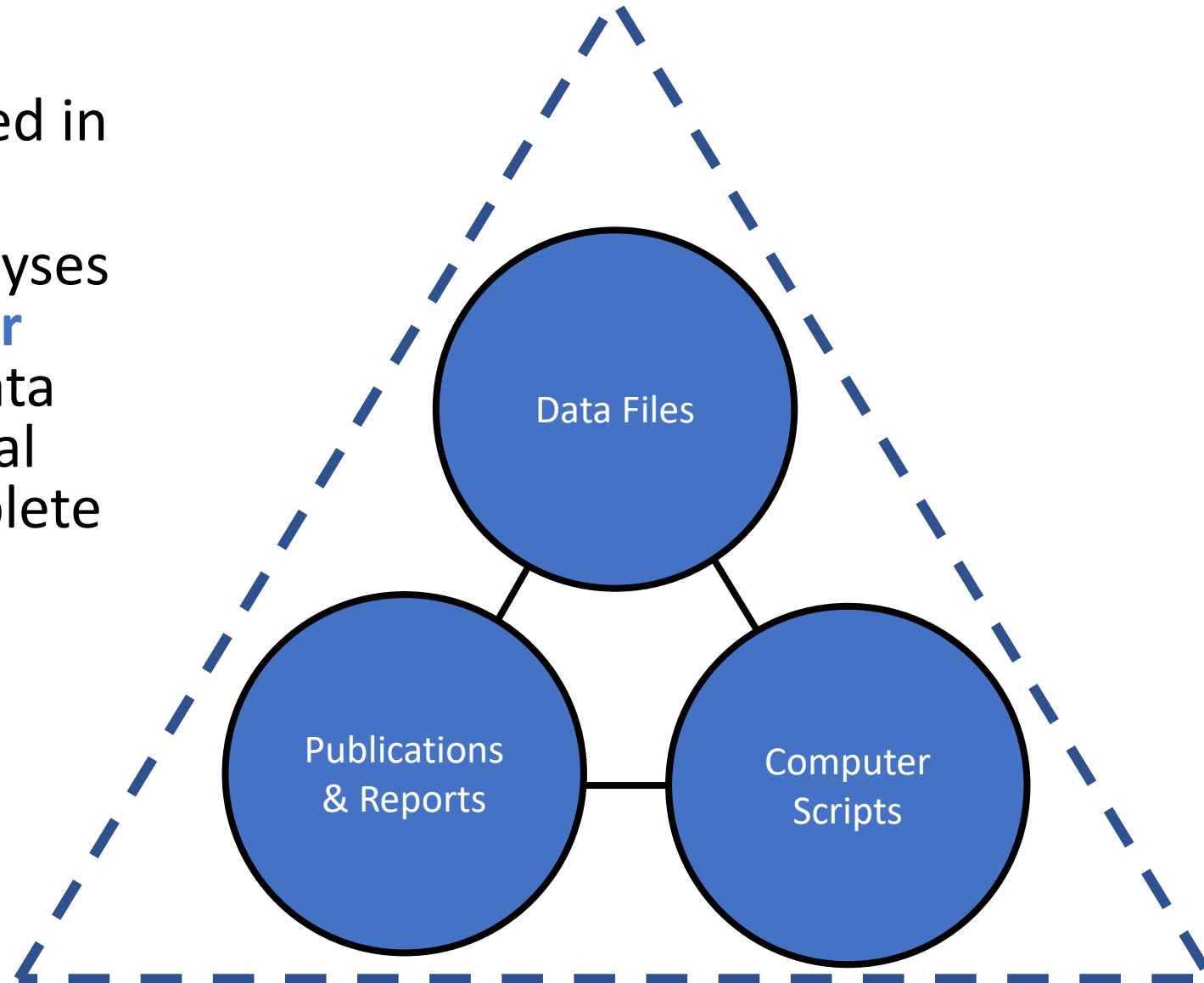
Title	Range extent for bird species
Downloaded	2 times
Description	This file contains the extents of occurrence (range filling) for 1205 Neotropical bird species.
Download	<a href="#">SM1 Metabirds.xls (2.937 Mb)</a>
Download	<a href="#">README.txt (677 bytes)</a>
Details	<a href="#">View File Details</a>

Title	R script
Downloaded	1 time
Description	R script for effect sizes computation and data-analyses built in R version 3.5.1
Download	<a href="#">SM2_script_metabirds.R (5.259 Kb)</a>
Details	<a href="#">View File Details</a>

Title	Neotropical bird consensus phylogeny
Downloaded	1 time
Description	This Neotropical bird consensus phylogeny was estimated from 10,000 random phylogenetic trees with 'Hackett constraint' for the backbone topology from Jetz et al. (2012; Nature, 491, 444–448. <a href="https://doi.org/10.1038/nature11631">https://doi.org/10.1038/nature11631</a> ) available in <a href="https://birdtree.org/">https://birdtree.org/</a> . The function 'consensus.edges' of 'phytools' package (Revell, 2012; Methods in Ecology and Evolution, 3, 217–223. <a href="https://doi.org/10.1111/j.2041-210X.2011.00169.x">https://doi.org/10.1111/j.2041-210X.2011.00169.x</a> ) to build the consensus phylogeny.
Download	<a href="#">phy_consensus.txt (51.13 Kb)</a>
Details	<a href="#">View File Details</a>

# Recall The Philosophy of Data Science

- All **data is digitized** and stored in files
- Data manipulations and analyses are documented in **computer scripts** that interface with data files and require no additional human intervention to complete analysis
- Data & scripts are published with the report and **openly accessible to all**





# GitHub – A Repository of Sorts

- A company
- Website is designed to aid in developing code, like the R script
- It also serves as a repository for code and scripts
- Efficient mechanism to disseminate your code to users
- Can also be used to organize a class

- <https://github.com/comp-bio-fall-2019>

