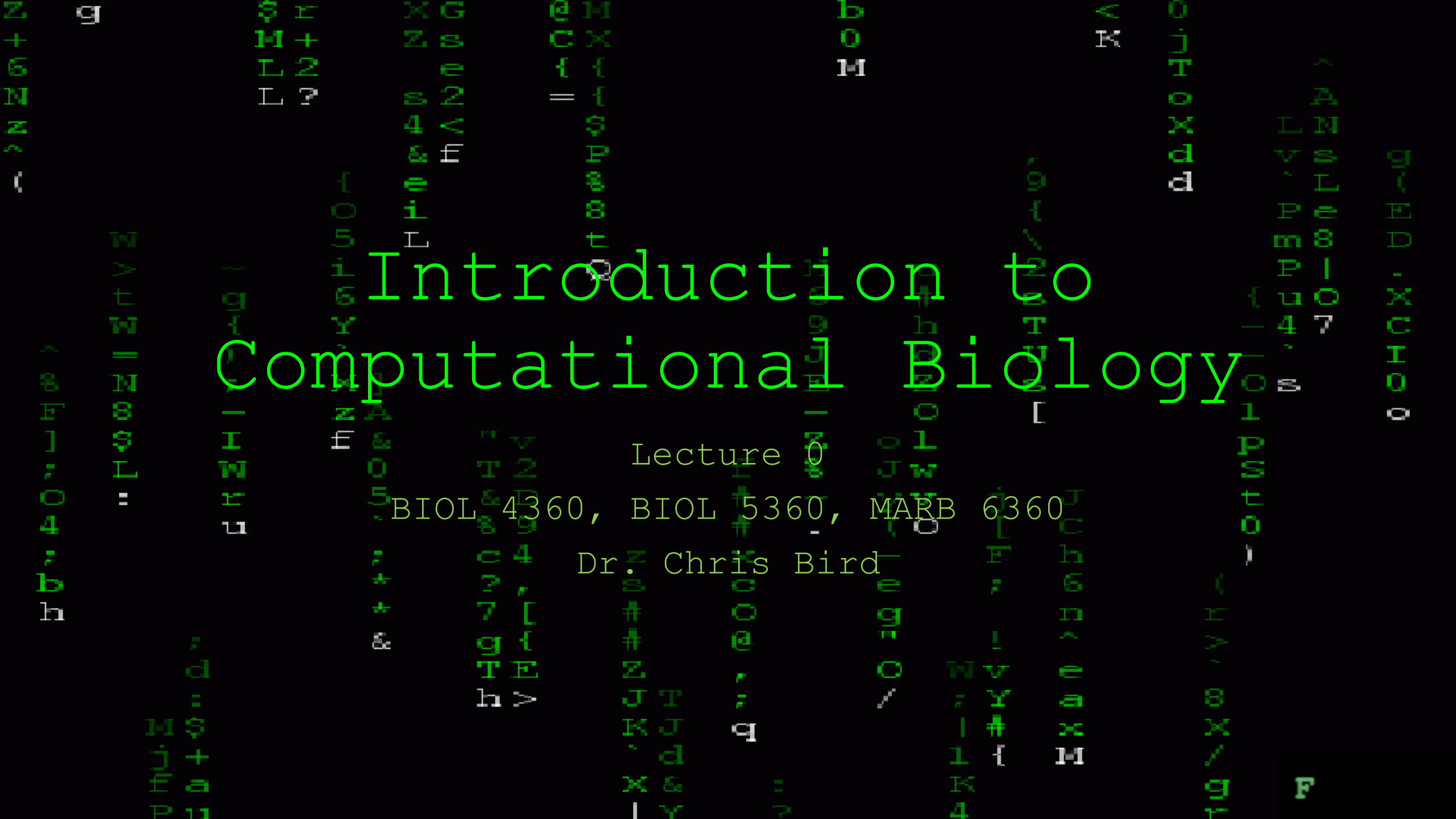


Introduction to Computational Biology

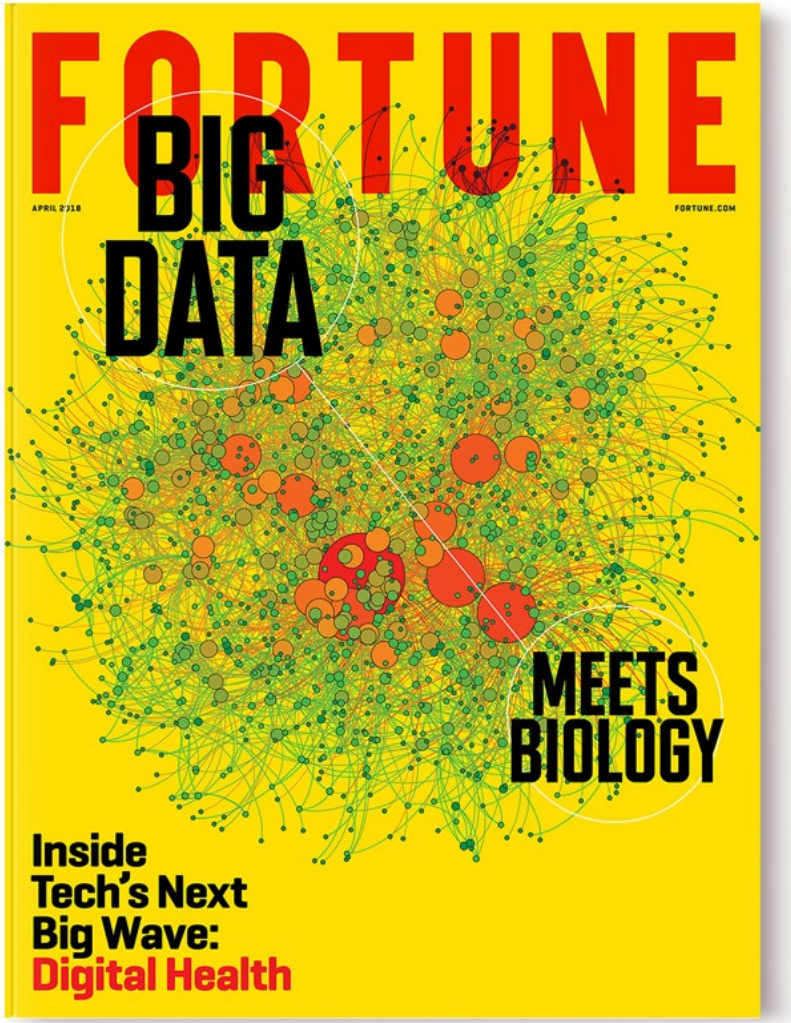
Lecture 0

BIOL 4360, BIOL 5360, MARB 6360

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 computational kung-fu?

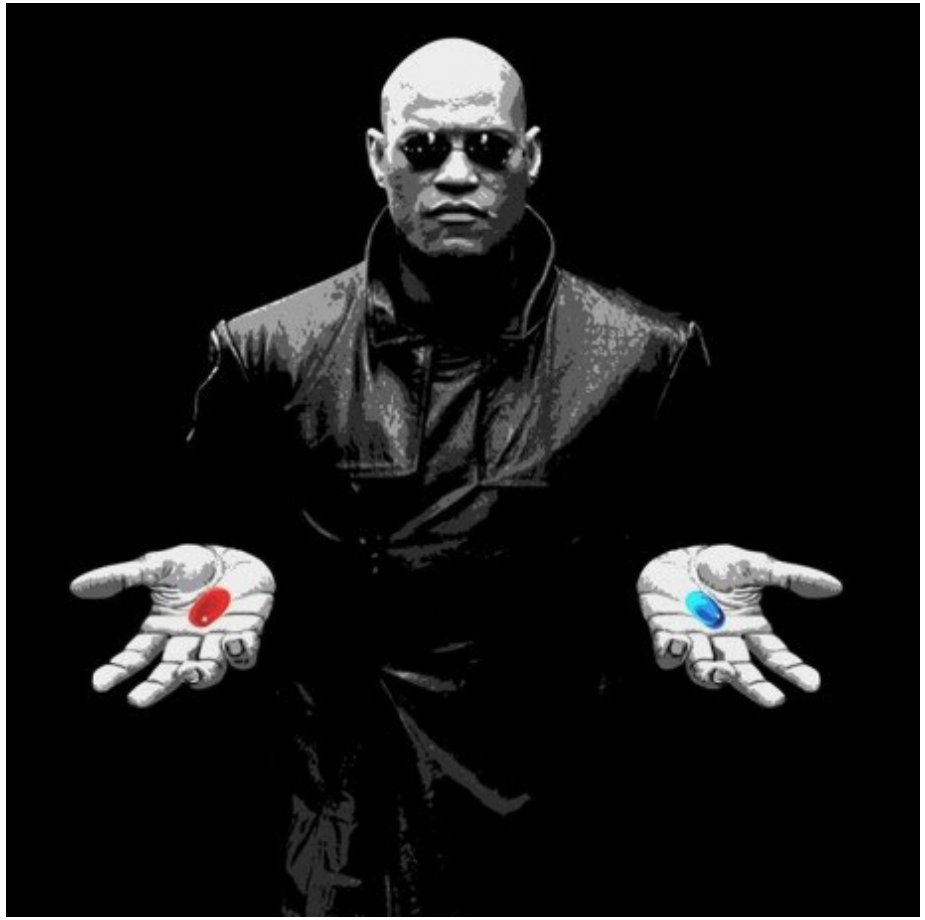
- Automate impossibly tedious, monotonous, and lengthy tasks
- Increased rate and significance of discovery
- Career success
- Maximize potential

Why did I choose to develop this course?

- 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

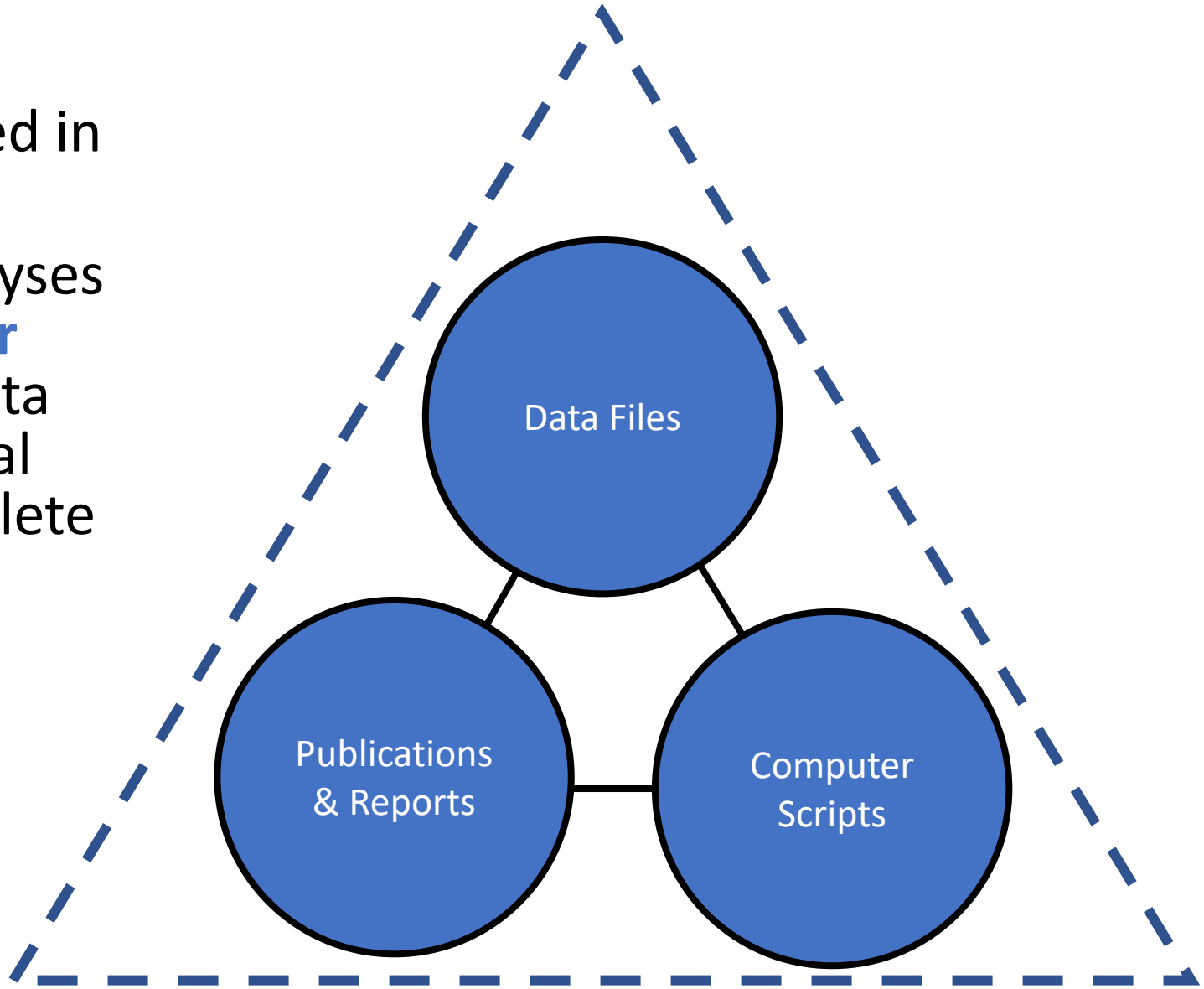
If you so choose, I will show you the philosophy of data science

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



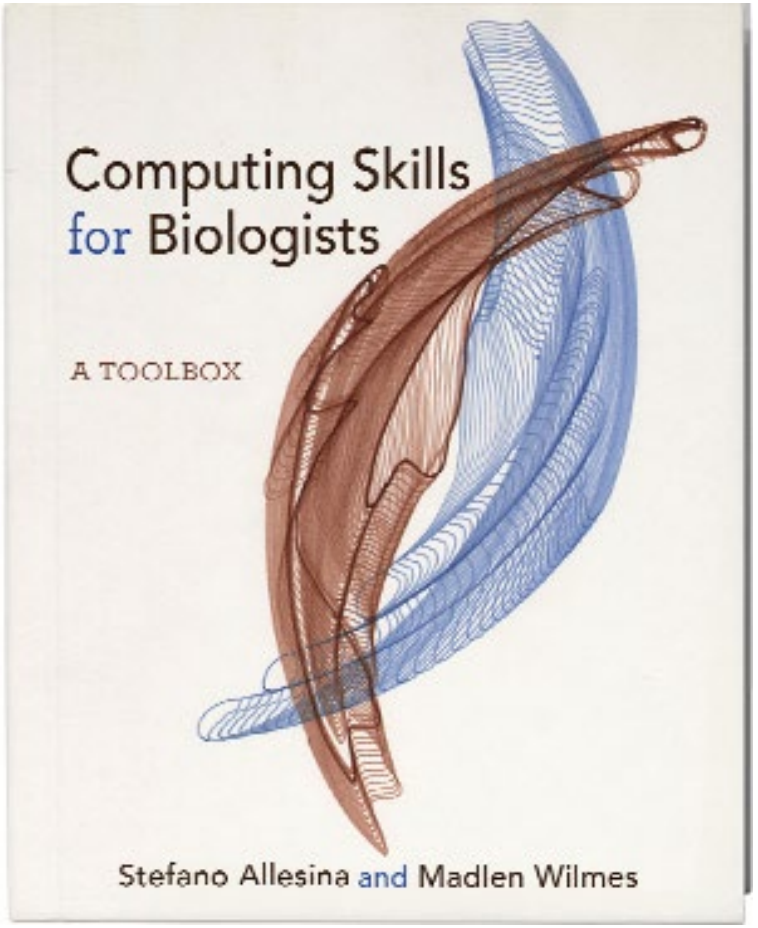
Philosophy of Data Science

- 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 Used to Follow The CSB Text Book, It's a Good Resource But Not Required

- 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

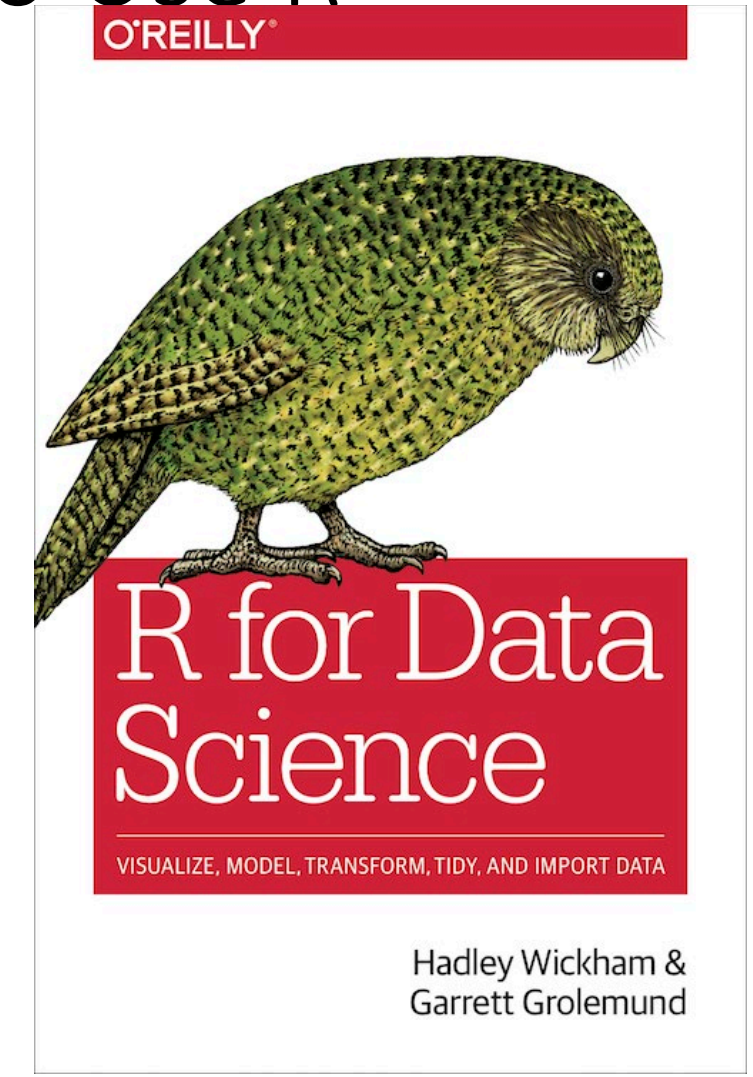


https://computingskillsforbiologists.com/

We are Going to Learn How to Use R

- You will learn core principles of R that aren't taught in other courses that expect you to use R

- You will learn to use `tidyverse`, which was masterminded by the Author of the R for Data Science book, which is free.



https://r4ds.had.co.nz/

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
  - Linux, R, Python
- 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	BIOL 5360: 50 MARB 6360: 40
Final Presentation	MARB 6360: 10

Lectures

- Environment for you to learn new concepts
- Hands-on with computers
- Power-point & github driven
  - On webex
- 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 (MARB 6360)

Questions?

Biological Data

Lecture 01

BIOL 4590, BIOL 5590

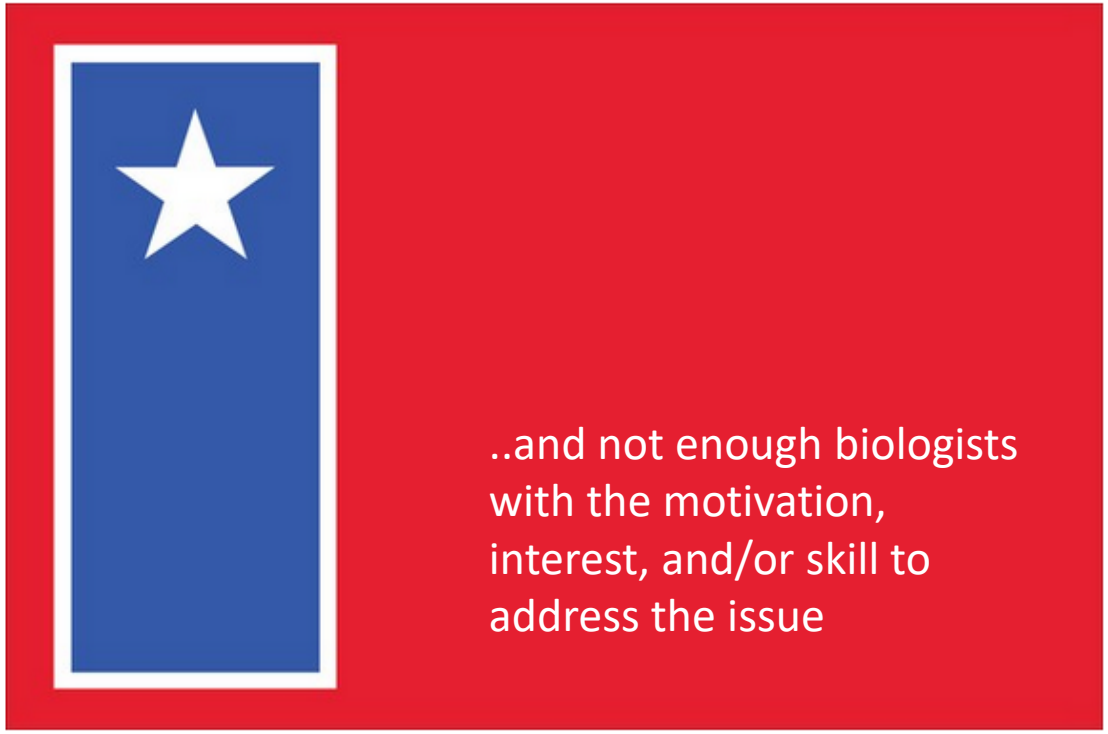
Dr. Chris Bird

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

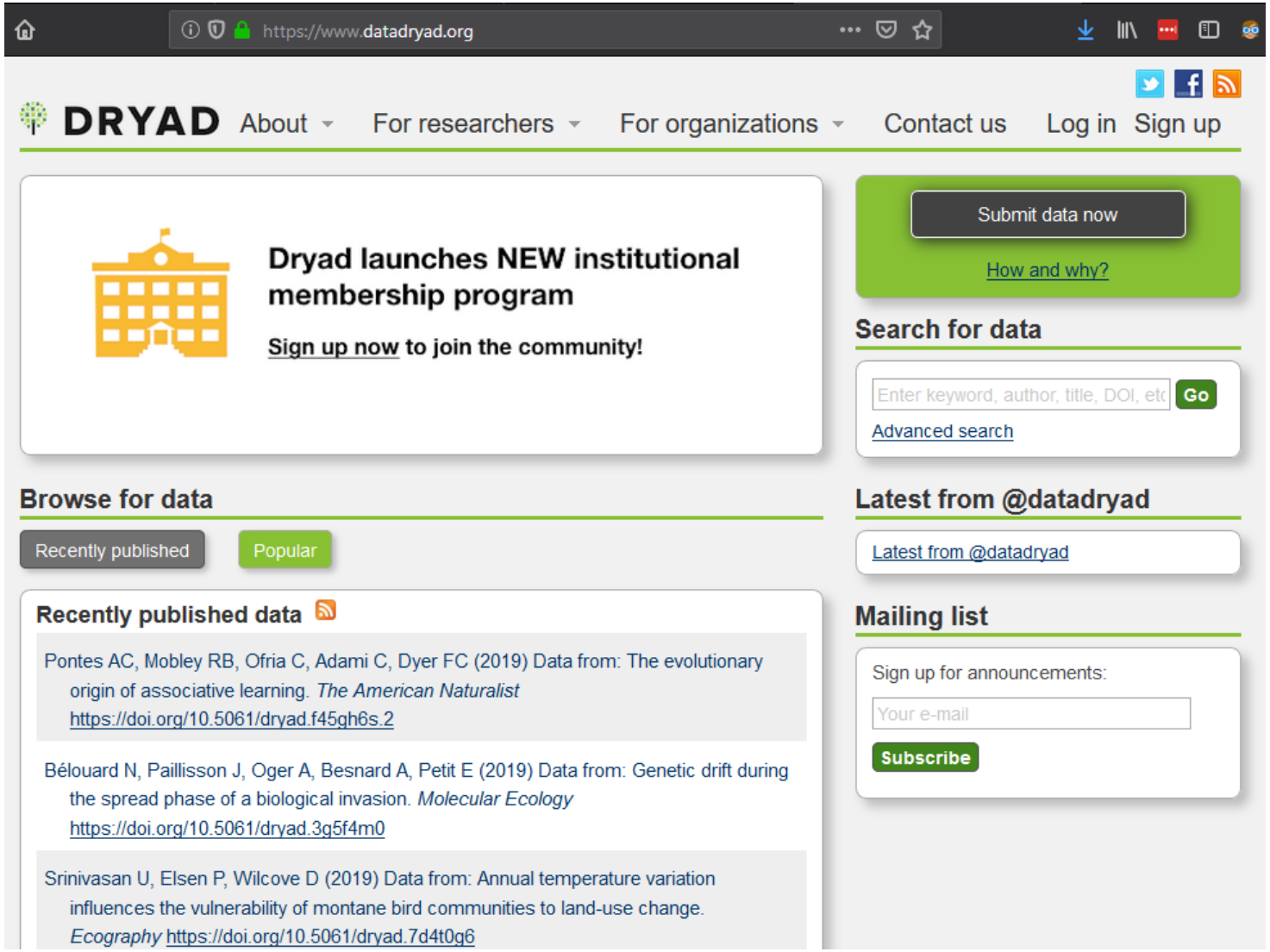
EMILY SINGER SCIENCE 10.11.13 09:30 AM

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



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
- DNA & Proteins
  - https://www.ncbi.nlm.nih.gov/
- GIS
  - https://data.usgs.gov/datacatalog/#fq=dataType%3A(collection%20OR%20non-collection)&q=*%3A*
- Oceanographic
  - https://data.noaa.gov/datasetsearch/
- Too many to list



Let's Explore a Data Set Published in Dryad

- www.datadryad.org
- 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 open files, 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. CC ZERO OPEN 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	Data for Heinen-Kay et al. Sexual signal l....xlsx (36.79 Kb)
Details	View File Details

Tidy Data (Wickham 2014)

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

Journal of Statistical Software

August 2014, Volume 59, Issue 10. *http://www.jstatsoft.org/*

Tidy Data

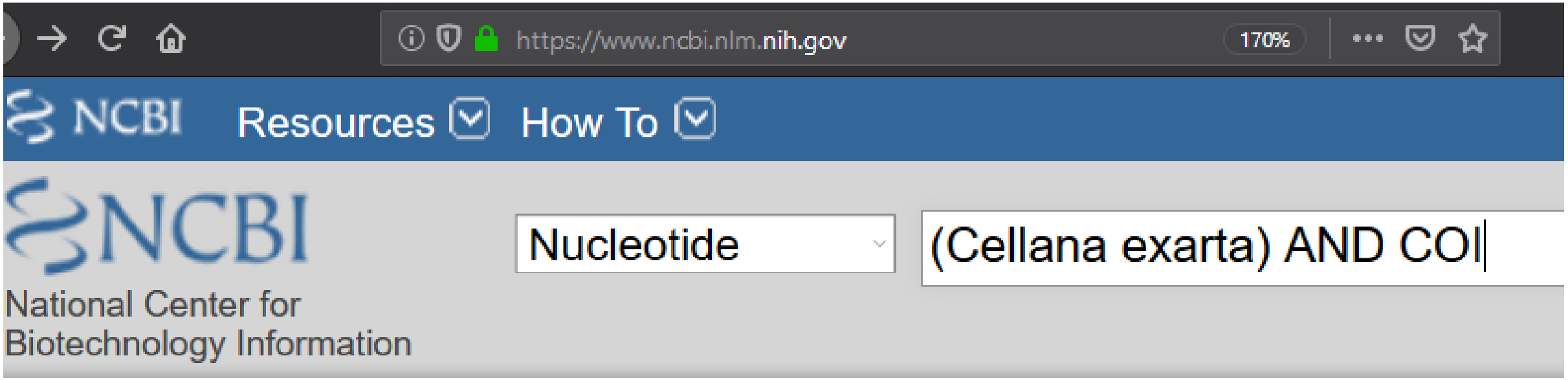
Hadley Wickham
RStudio

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.

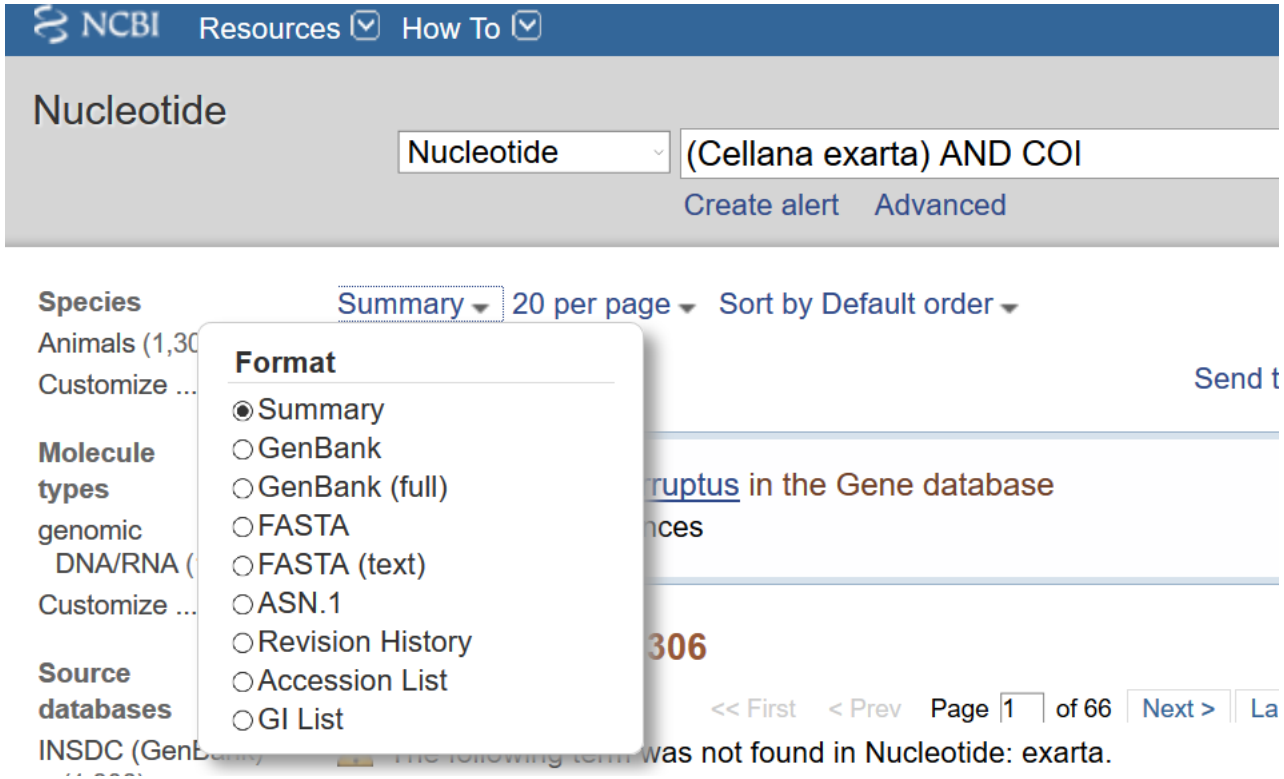
Common Data Formats & Structures are Not Always Tidy

- 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)




```
>AB263731.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L694(Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTTAGTAATACCAATGATAATTGGGGGTTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTTGGTTACTGGTTCCTTCTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCTGGTTCTTCTGTTGATTTGGCTATTTTTTCTCTTCATTTGGCTGGTATTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTTCTCTTCCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAATACCTGTTTTTTTGACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT

>AB263730.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L693(Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTTAGTAATACCAATGATAATTGGGGGTTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTTGGTTACTGGTTCCTTCTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCTGGTTCTTCTGTTGATTTGGCTATTTTTTCTCTTCATTTGGCTGGTATTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTTCTCTTCCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAATACCTGTTTTTTTGACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT

>AB263729.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L692(Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTTAGTAATACCAATGATAATTGGGGGTTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTTGGTTACTGGTTCCTTCTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCTGGTTCTTCTGTTGATTTGGCTATTTTTTCTCTTCATTTGGCTGGTATTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTTCTCTTCCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAATACCTGTTTTTTTGACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT

>AB263728.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L691(Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTTAGTAATACCAATGATAATTGGGGGTTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTTGGTTACTGGTTCCTTCTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCTGGTTCTTCTGTTGATTTGGCTATTTTTTCTCTTCATTTGGCTGGTATTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTCGTTGGCGAGGTCTTCAGTTTGAA
CGTCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTTCTCTTCCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAATACCTGTTTTTTTGACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT
```

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 ID of the unit of observation
  - Lines that don't being with `>` contain information, each character (nucleotide) is a dimension of the unit of observation

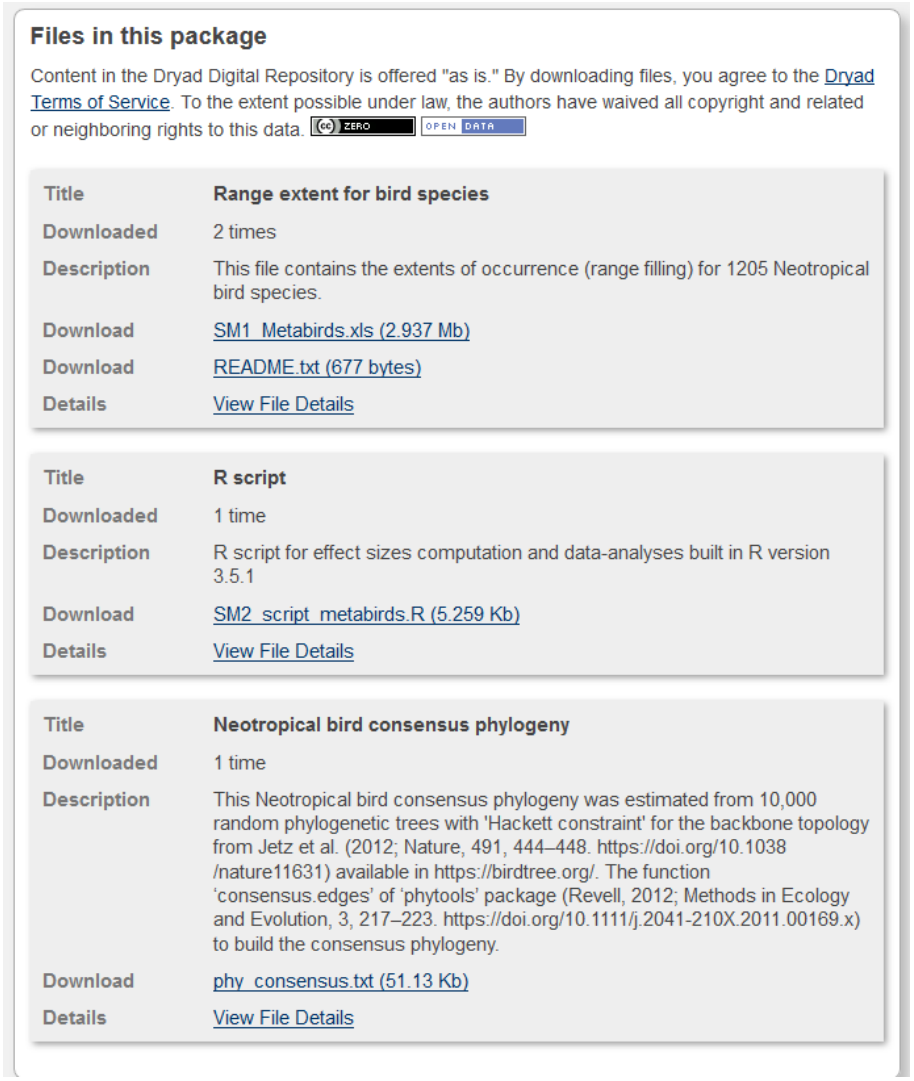
```
>AB263731.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L694(Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTTAGTAATACCAATGATAATTGGGGGTTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTTGGTTACTGGTTCCTTCTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCTGGTTCTTCTGTTGATTTGGCTATTTTTTCTCTTCATTTGGCTGGTATTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTTCTCTTCCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAATACCTGTTTTTTTGACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT

>AB263730.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L693(Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTTAGTAATACCAATGATAATTGGGGGTTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTTGGTTACTGGTTCCTTCTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCTGGTTCTTCTGTTGATTTGGCTATTTTTTCTCTTCATTTGGCTGGTATTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTTCTCTTCCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAATACCTGTTTTTTTGACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT

>AB263729.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L692(Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTTAGTAATACCAATGATAATTGGGGGTTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTTGGTTACTGGTTCCTTCTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCTGGTTCTTCTGTTGATTTGGCTATTTTTTCTCTTCATTTGGCTGGTATTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTCGTTGGCGAGGTCTTCAGTTTGAA
CGGCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTTCTCTTCCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAATACCTGTTTTTTTGACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT

>AB263728.1 Cellana radiata enneagona mitochondrial COI gene for cytochrome c
oxidase subunit I, partial cds, specimen_voucher: NUGB-L691(Nagoya University)
TACATTATACATTATTATAGGAGTTTGATCTGGATTGGCAGGTACTGGTTTAAGTATGTTAATTCGGGCT
GAATTAGGTCAACCTGGTTCTTTGCTAGGAGATGATCAGCTATATAACGTGATTGTTACTGCGCACGCTT
TTGTTATGATTTTCTTTTTAGTAATACCAATGATAATTGGGGGTTTTGGAAATTGGTTGGTTCCTCTTAT
ACTTGGGGCTCCAGATATGGCTTTTCCTCGTTTAAATAATATGAGGTTTTGGTTACTGGTTCCTTCTTTA
TTTTTACTTCTTGCTTCTTCTGCTGTTGAAAGAGGAGTAGGTACAGGTTGGACAGTATACCCCCCTCTTT
CTAGAAATGTGGCCCATTCTGGTTCTTCTGTTGATTTGGCTATTTTTTCTCTTCATTTGGCTGGTATTTC
TTCAATTCTTGGGGCTGTTAATTTTATTACTACAGTGGTAAACATTCGTTGGCGAGGTCTTCAGTTTGAA
CGTCTACCTTTGTTTGTATGATCTGTTAAGATTACAGCTATTTTACTTCTTCTTTCTCTTCCTGTGTTGG
CTGGGGCTATTACTATGCTTTTAACTGACCGTAATTTTAATACCTGTTTTTTTGACCCTGGAGGAGGAGG
GGACCCCATTTTATATCAACATTTGTTT
```

Data Formats

- I will emphasize Tidy format
- Many fields of Biology have their own particular and peculiar 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 not as intimidating as it seems

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

- 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.

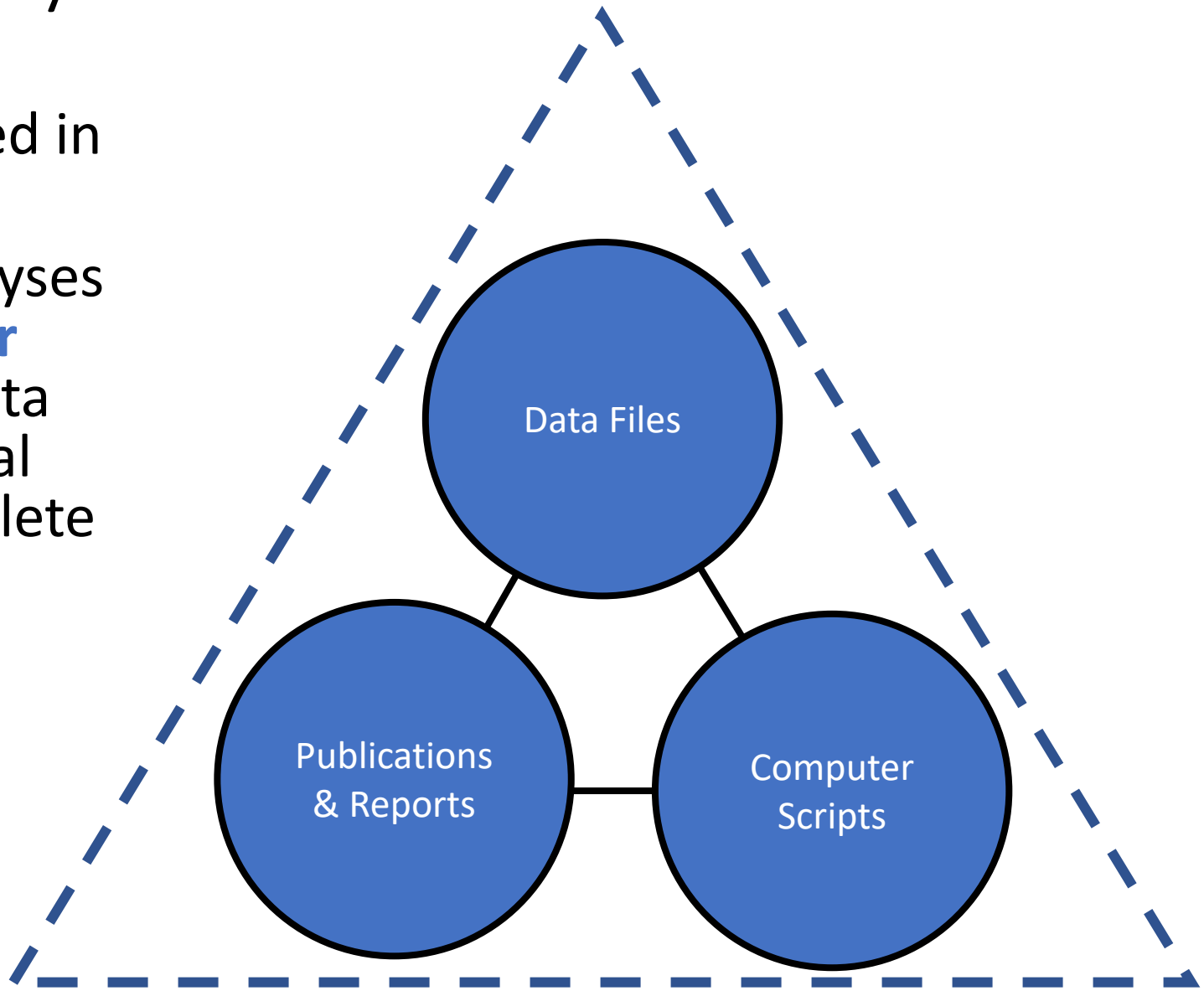
Title	**Range extent for bird species**
Downloaded	2 times
Description	This file contains the extents of occurrence (range filling) for 1205 Neotropical bird species.
Download	SM1_Metabirds.xls (2.937 Mb)
Download	README.txt (677 bytes)
Details	View File Details

Title	**R script**
Downloaded	1 time
Description	R script for effect sizes computation and data-analyses built in R version 3.5.1
Download	SM2_script_metabirds.R (5.259 Kb)
Details	View File Details

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. https://doi.org/10.1038/nature11631) available in https://birdtree.org/. The function 'consensus.edges' of 'phytools' package (Revell, 2012; Methods in Ecology and Evolution, 3, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x) to build the consensus phylogeny.
Download	phy_consensus.txt (51.13 Kb)
Details	View File Details

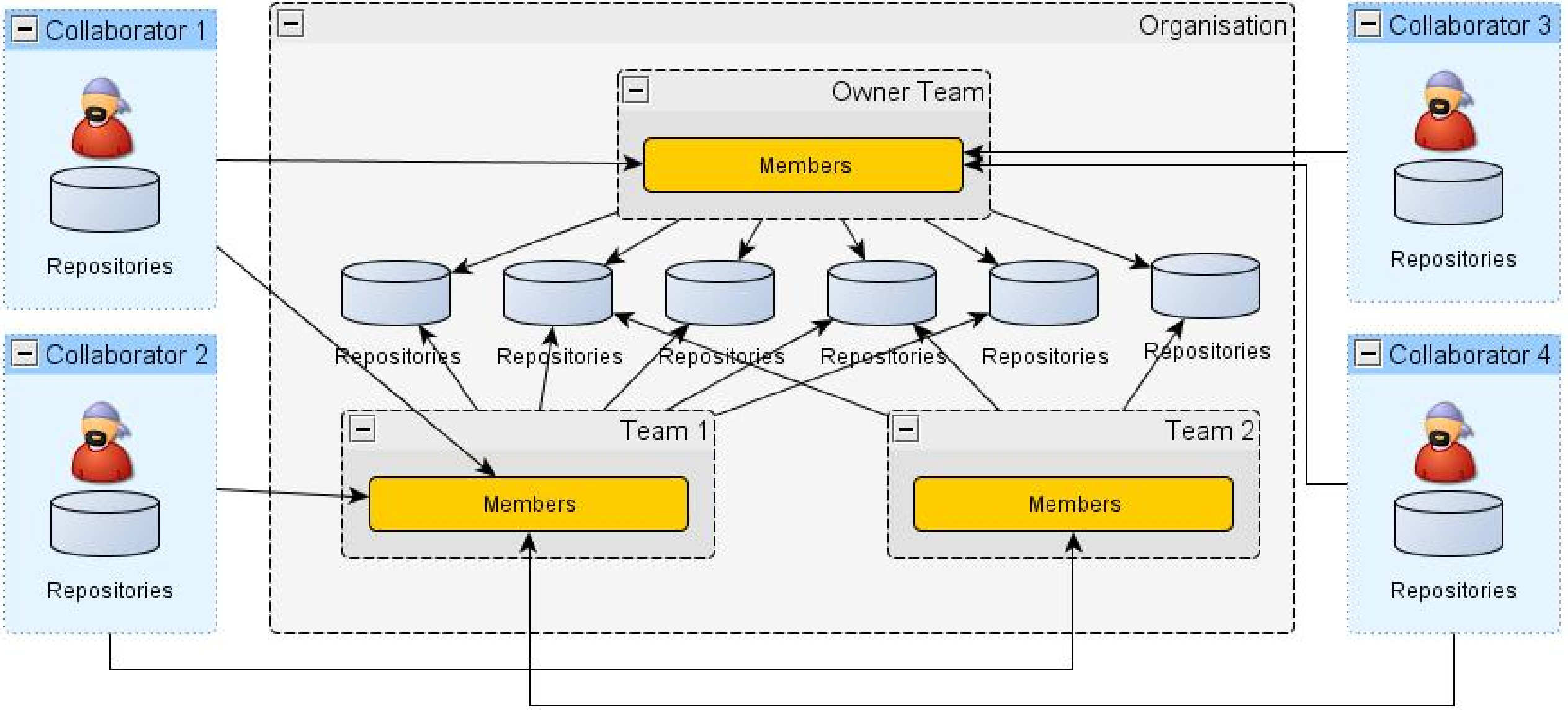
Recall The Philosophy of Data Science

- 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 an R script
- It also serves as a repository for data, code, and scripts
- Efficient mechanism to disseminate your code to users
- Can also be used to organize a class

- https://github.com/orgs/tamucc-comp-bio-2022/repositories

Conceptual Diagram of a GitHub Organization