
EXPLORING ENDANGERED SPECIES
DATA WITH PYTHON

EXPLORING ENDANGERED SPECIES DATA WITH PYTHON

Amanda Devine

Data Wrangler, Global Genome Initiative

27 July 2018

Girls Who Code Summer Immersion Program in Washington DC field trip to the Smithsonian National Museum of Natural History

Slides and Jupyter notebook available at https://github.com/amdevine/gwc-endangered-species


ABOUT ME

BIO

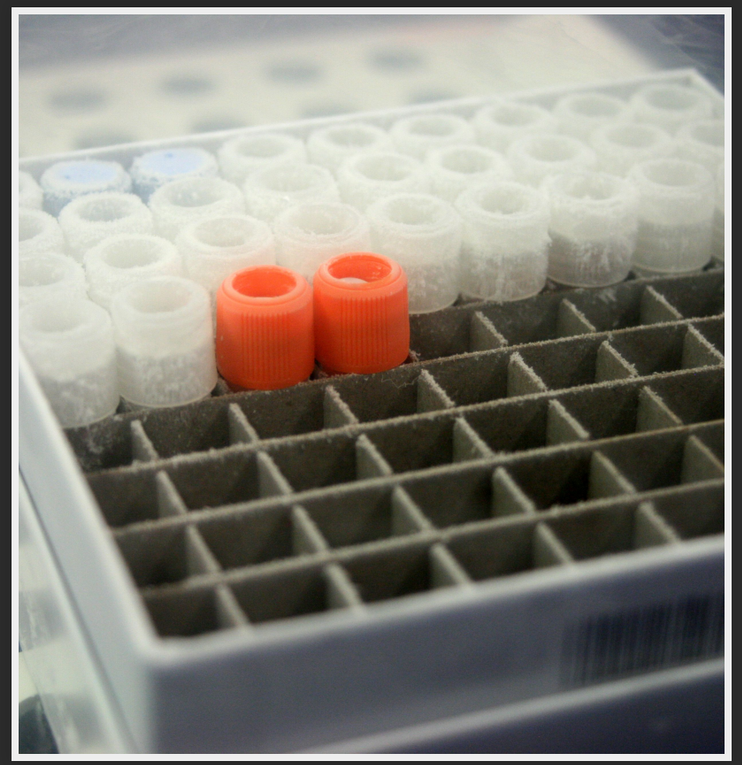
Winston Churchill High School (go Bulldogs!)

Dartmouth College (BA in Biology (Ecology) and Neuroscience

Lab technician (dermatology, infectious disease, coral reefs)

Data wrangler for the Global Genome Initiative

GLOBAL GENOME INITIATIVE (GGI)



Smithsonian initiative

Collect all of life on Earth

Preserve in cryorepositories for genomic research

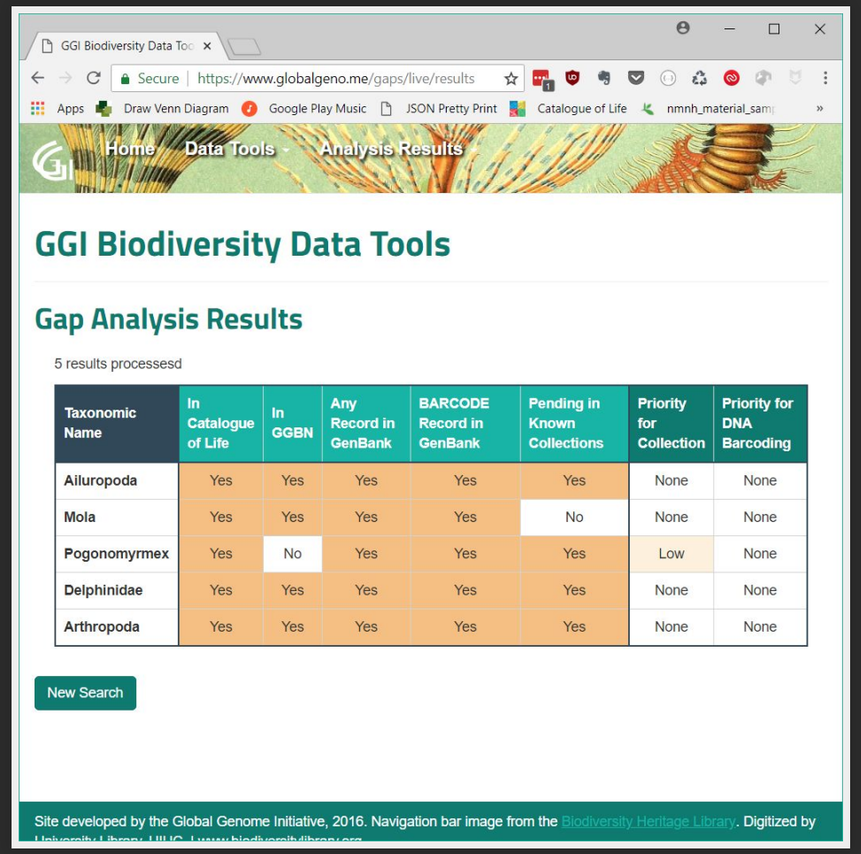
Sample data recorded in the Global Genome Biodiversity Network (GGBN) Data Portal

GGI WEB PROJECTS

GGI Data Tools website (Django; https://www.globalgeno.me)

GGI Gap Analysis app (Shiny; https://ggidata.shinyapps.io/gapanalysis)

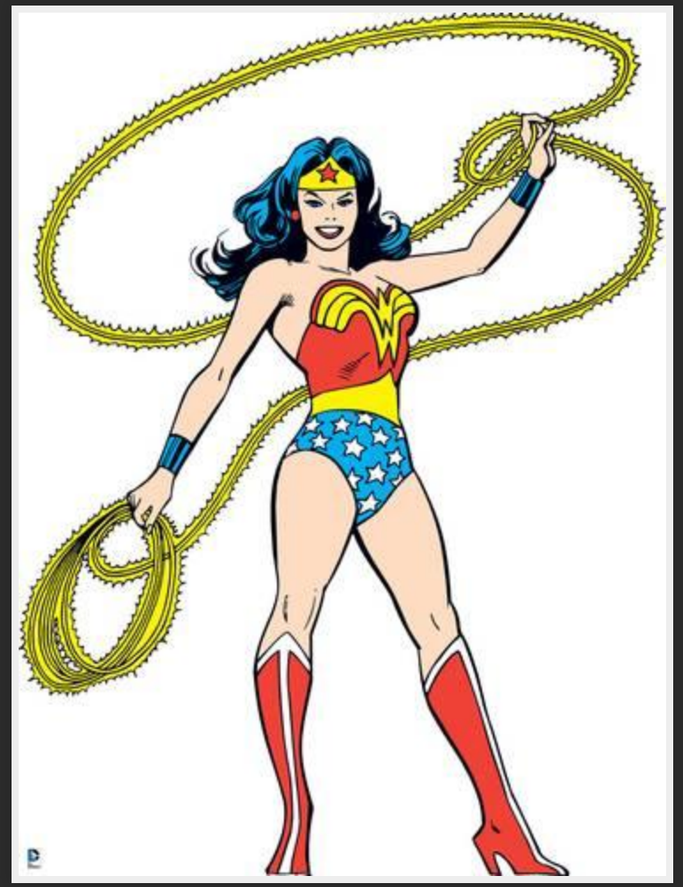
genetic_collections (Python library; https://github.com/MikeTrizna/genetic_collections)



Taxonomic Name	In Catalogue of Life	In GGBN	Any Record in GenBank	BARCODE Record in GenBank	Pending in Known Collections	Priority for Collection	Priority for DNA Barcoding
Ailuropoda	Yes	Yes	Yes	Yes	Yes	None	None
Mola	Yes	Yes	Yes	Yes	No	None	None
Pogonomyrmex	Yes	No	Yes	Yes	Yes	Low	None
Delphinidae	Yes	Yes	Yes	Yes	Yes	None	None
Arthropoda	Yes	Yes	Yes	Yes	Yes	None	None


DATA WRANGLING

WHAT IS DATA WRANGLING?



Per Wikipedia: *Transforming and mapping data from one raw data form into another format with the intent of making it [useful] for a variety of downstream purposes, such as analytics.*

My favorite tools:

- Scripting languages: Python, R, VBA
- Applications: OpenRefine, Jupyter Notebook, Excel, Pandoc
- Version control: GitHub

JUPYTER NOTEBOOK

Document that contains executable Python code and Markdown-formatted text

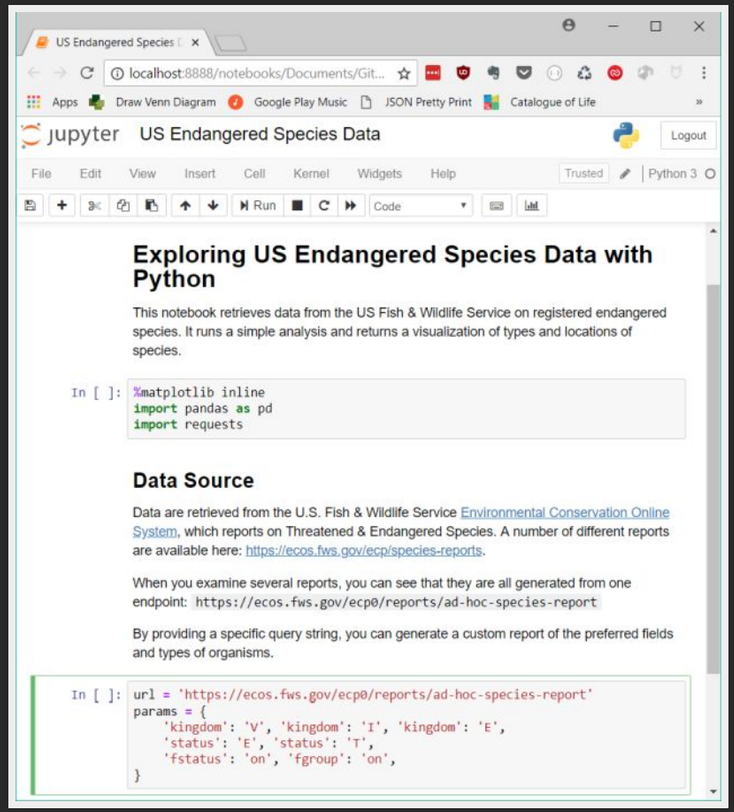
Good for running self-contained analyses

Easily share with others

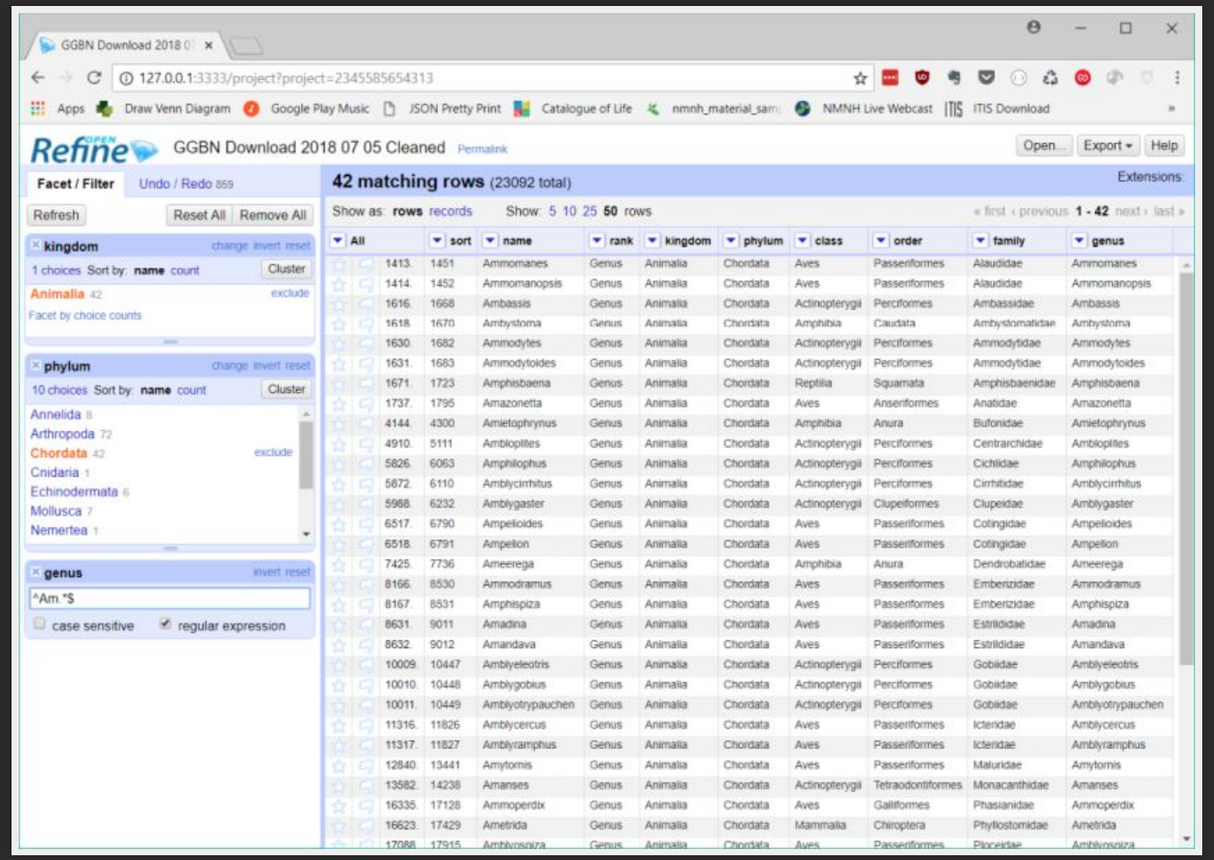
IRkernel: Run notebooks with R instead of Python

nbviewer: Converts notebooks to shareable HTML documents

RISE: Run a Jupyter notebook as a slide show



OPENREFINE



Powerful tool for cleaning messy data

Complex filtering, sorting, and grouping

Mass editing records

Special language (GREL) to filter and edit data with formulas

R

Programming language developed for statistics

Powerful at data manipulation

More intuitive than Python when working with data??

RStudio: popular R development software

Shiny: R library, easily develop web apps to vizualize data


ENDANGERED SPECIES DATA

ENDANGERED SPECIES ACT

Administered by the U.S. Fish & Wildlife Service

Established in 1973 *"to conserve and protect endangered and threatened species and their habitats"*

Species are listed under the ESA in two ways:

1. FWS scientist assessment
2. Petition from the general public

Virginia big-eared bat

WHAT DO WE WANT TO KNOW?

What question are we trying to answer?

> *How have rates of listing species under the Endangered Species Act changed over time?*

What summary or visualization do we want to produce at the end?

> *A bar graph showing the number of species listed by year*

ENVIRONMENTAL CONSERVATION ONLINE SYSTEM (ECOS)

Database that serves reports on threatened and endangered species

Pre-generated reports available online here: https://ecos.fws.gov/ecp/species-reports

Let's look at these data in a Jupyter notebook: https://github.com/amdevine/gwc-endangered-species/blob/master/US%20Endangered%20Species%20Data.ipynb


THANKS!

RESOURCES: WORKING WITH DATA IN PYTHON

- **Automate the Boring Stuff with Python.** https://automatetheboringstuff.com/
- **Python Data Science Handbook.** https://jakevdp.github.io/PythonDataScienceHandbook/
- Coursera: **Using Python to Access Web Data.** https://www.coursera.org/learn/python-network-data (Can choose to audit the course for free.)
- Coursera: **Using Databases with Python.** https://www.coursera.org/learn/python-databases (Can choose to audit the course for free.)

RESOURCES: CODING GROUPS AND ORGANIZATIONS

- **Women Who Code DC**. Meetup group for female-identifying coders in the Washington, DC area. Covers many different tech-related topics, frequent meetups. https://www.meetup.com/Women-Who-Code-DC/
- **Hear Me Code**. Organization that offers beginner coding lessons for women in the Washington, DC area. Also has an excellent Google group that emails out about a lot of professional opportunities. https://hearmecode.com/
- **Data Carpentry**. National organization that offers workshops on data wrangling. The website contains workshop materials if you can't attend a workshop in person. https://datacarpentry.org/

IMAGE CREDITS

Title Slide: Grey Crowned Cranes. Image from Pexels, CC0 License. https://www.pexels.com/photo/nature-bird-love-heart-45853/

About Me: Giant Panda. Photo by Cesar Aguilar from Pexels, Pexels License. https://www.pexels.com/photo/panda-1123765/

Bio: Personal photo.

Global Genome Initative: Tissue samples in the NMNH Biorepository. Photo by Adrian Van Allen, 2015.

Data Wrangling: Whale shark at the Georgia Aquarium. Photo by Zac Wolf; CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=3511009

Data Wrangling: Wonder Woman: Wonder Woman with Lasso. Image from AllPosters. https://www.allposters.ca/-sp/Wonder-Woman-Wonder-Woman-with-Lasso-posters_i13190262_.htm

Endangered Species Data: Bufo periglenes (Golden toad). Photo by Charles H. Smith. Retrieved from Wikipedia: https://commons.wikimedia.org/wiki/File:Bufo_periglenes1.jpg

Thanks: Rafflesia arnoldii. Image from lazypenguins.com, blog post “15 strangely beautiful flowers”. https://lazypenguins.com/15-strangely-beautiful-flowers/

Any Questions: Joes Apartment Cockroach GIF. Image from GIPHY. https://giphy.com/gifs/scarface-when-mtv-was-worth-watching-joes-apartment-CbY83hpLkcrZe

ANY QUESTIONS?