Scraping in Python

WORKSHOP 2 | CREATOR: CHARLOTTE LLOYD

Outline

I. Introduction to Python
II. Python Three Ways
III. Scraping Recap
IV. Workshop Example
V. Verify Data
VI. Celebration, Back-slapping

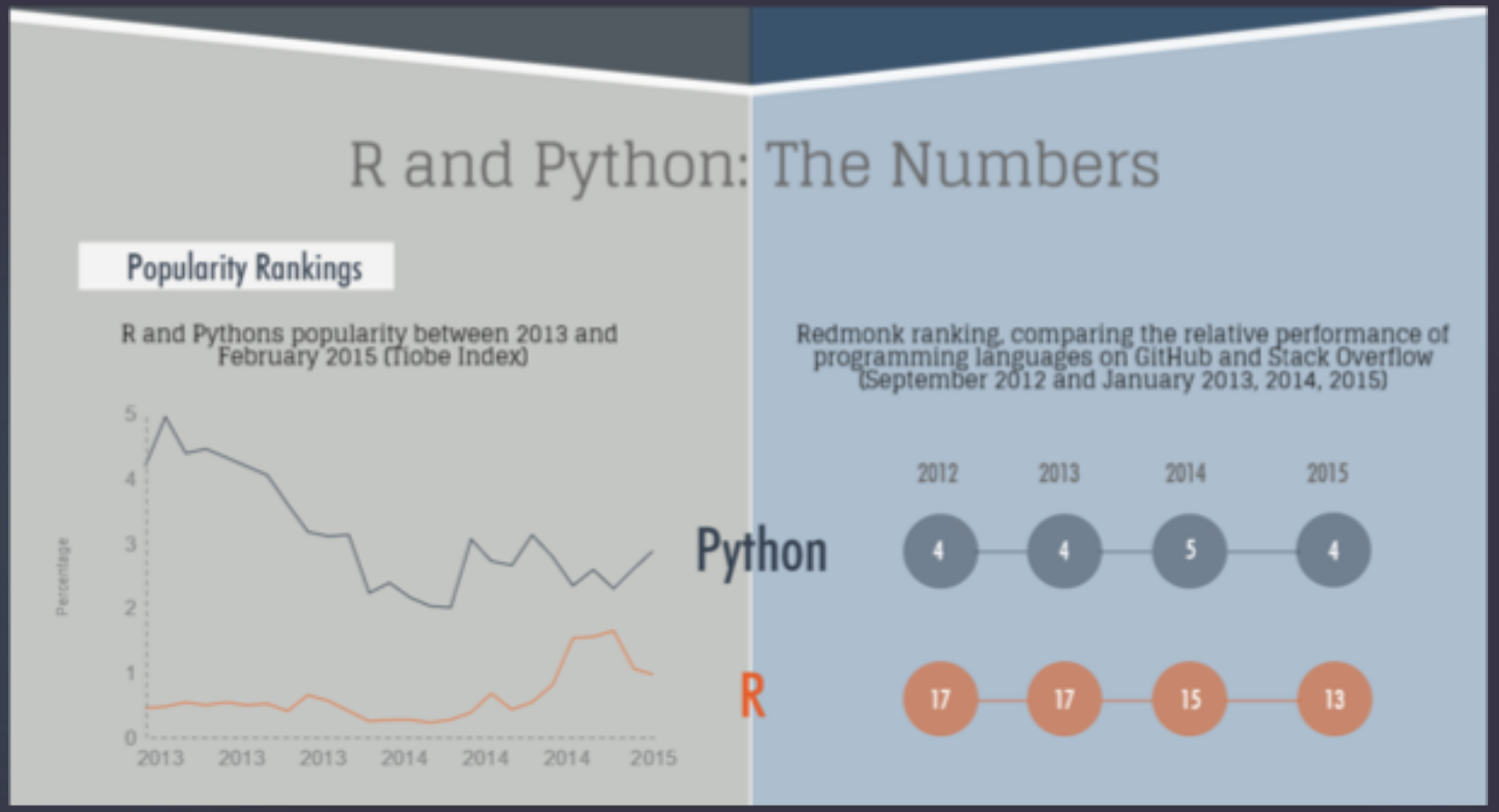
Introduction to Python

PART I

What is Python?

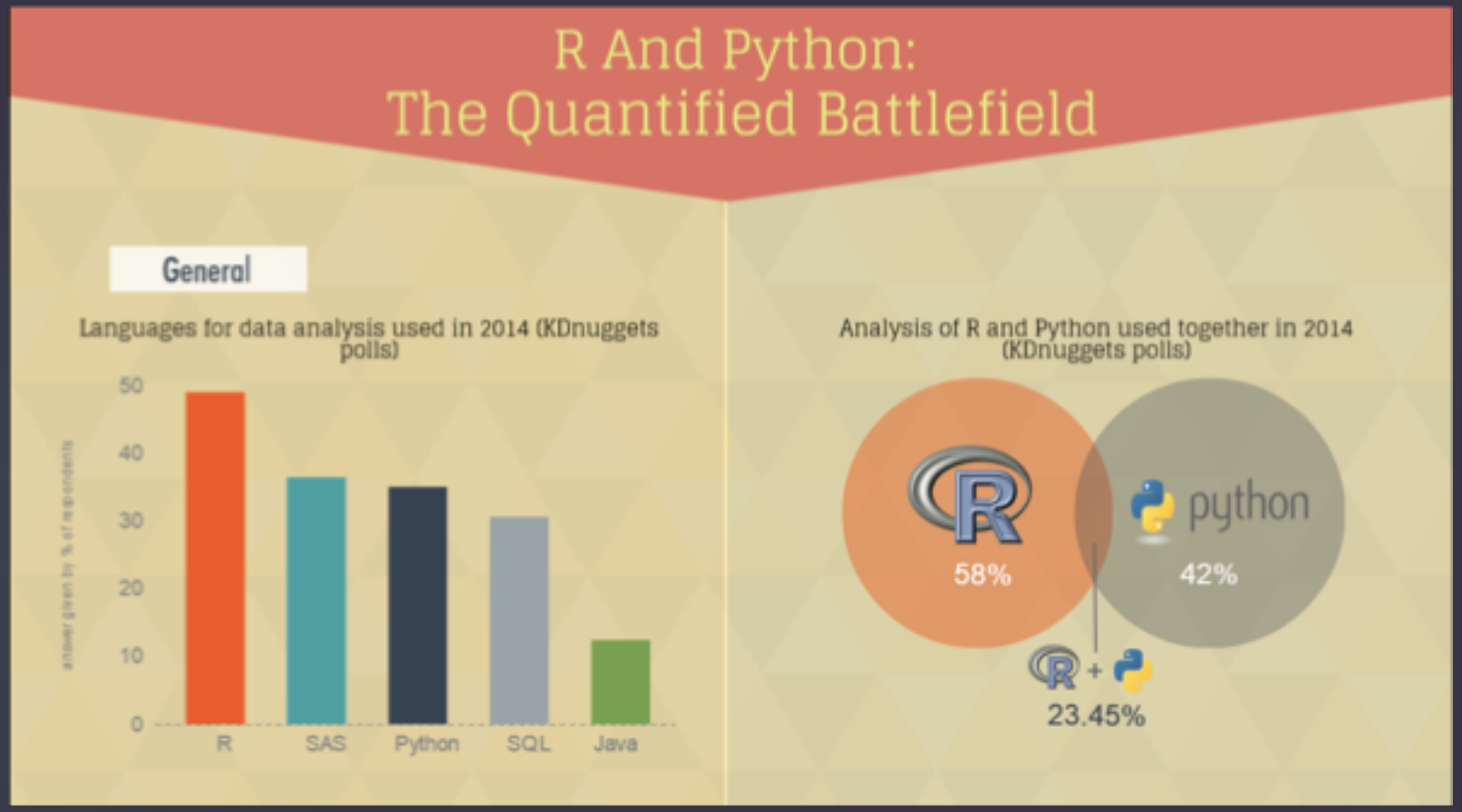
- general purpose
- high-level
- interpreted (not compiled)
- name is related to Monty Python

Very Popular Language



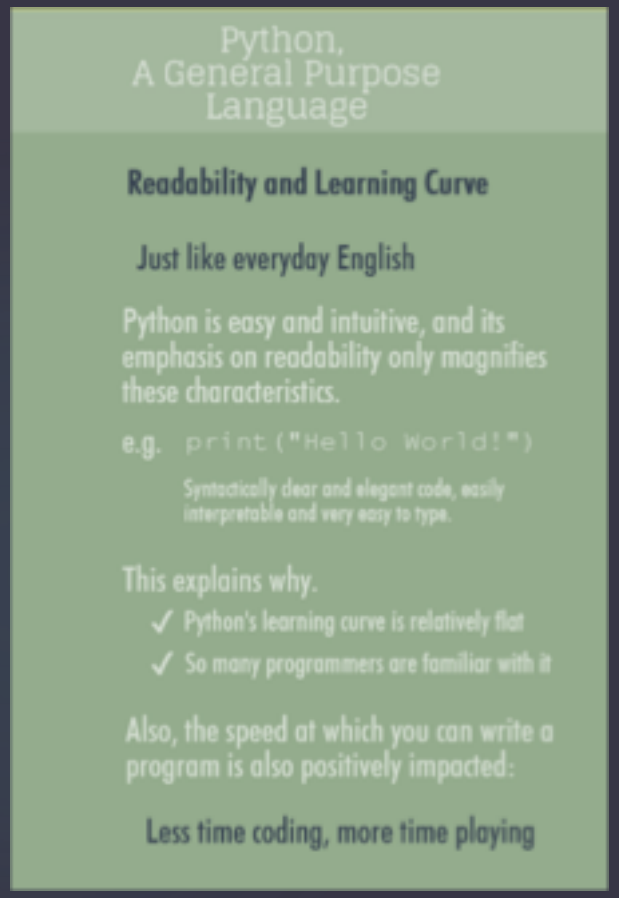
Checkout the full infographic: http://blog.datacamp.com/wp-content/uploads/2015/05/R-vs-Python-216-2.png

Less Popular in Data Analysis



Checkout the full infographic: http://blog.datacamp.com/wp-content/uploads/2015/05/R-vs-Python-216-2.png

Great Beginner Language



Python's focus on readability and simplicity makes that its learning curve is relatively low and gradual.

Python is considered a good language for starting programmers.

Checkout the full infographic: http://blog.datacamp.com/wp-content/uploads/2015/05/R-vs-Python-216-2.png

Packages for Python

- Packages are bits of code that other people have built to extend Python functionality.
  - If you install a package you will be able to use the additional commands that package has defined.

- Over 100,000 publically listed packages famously including:
  - numpy
  - scikit-learn
  - pandas

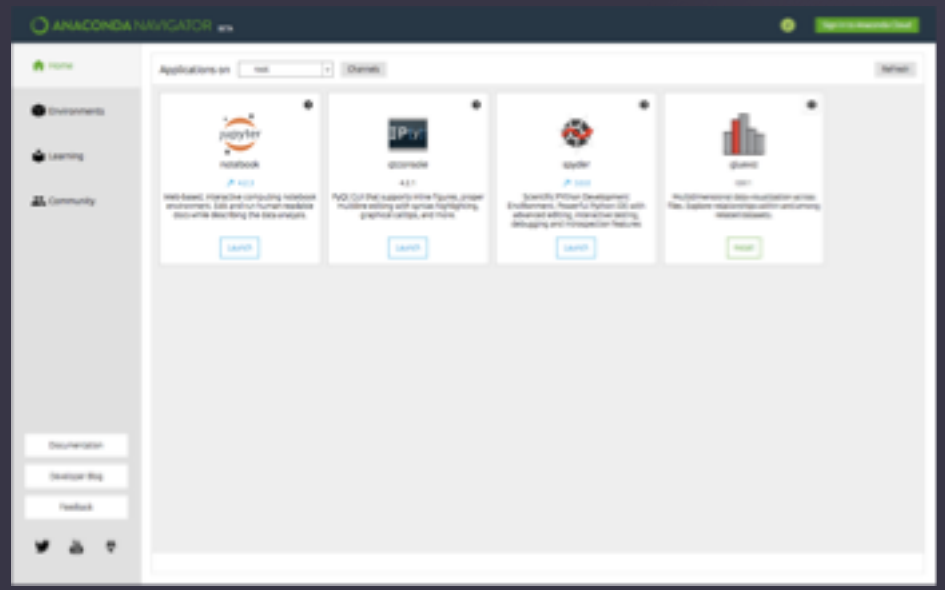
Python Three Ways

PART II

What is Anaconda?

- Anaconda is an “installation” of Python that includes:
  - package management
  - environment management
  - python distribution

- Anaconda pre-installs over 100 packages



Three Major Ways to Use Python

1. Command Line
2. “IDE”
3. Notebook

1. “Command Line” Python

A. **Run an interactive session in a Unix shell**
   1. In Terminal (Mac) or Powershell (PC):
      1. type ”python”
      2. type “2+2”
   2. In qtconsole (Anaconda Navigator): [do nothing]
      - try typing ”2+2”

B. **Run a script (file)**
   1. In Terminal (Mac) or Powershell (PC): type “`python file.py`”
   2. In qtconsole (Anaconda Navigator): type “`%load file.py`”

2. Python in IDEs

- IDE ("integrated development environment")
  - Spyder (provided in Anaconda)
  - PyCharm
  - Xcode (Macs)
- Write code (esp. multiple files) and easily execute within the IDE.
- **Activity: Write a "helloworld" program in Spyder. Execute in both Spyder and Terminal/Powershell.**

3. Python Notebooks

- web-based “interactive computational environment”
- very visual, very cool
- segmented into small cells of executable code

Hands-on Demo

- Open Anaconda Navigator. Open the Jupyter Notebook.
  - Navigate to "**handypy.ipynb**" and open.

- Topics to be covered:
  - integers, floats, and strings
  - lists
  - for and while loops
  - conditionals
  - functions
  - reading and writing csv files

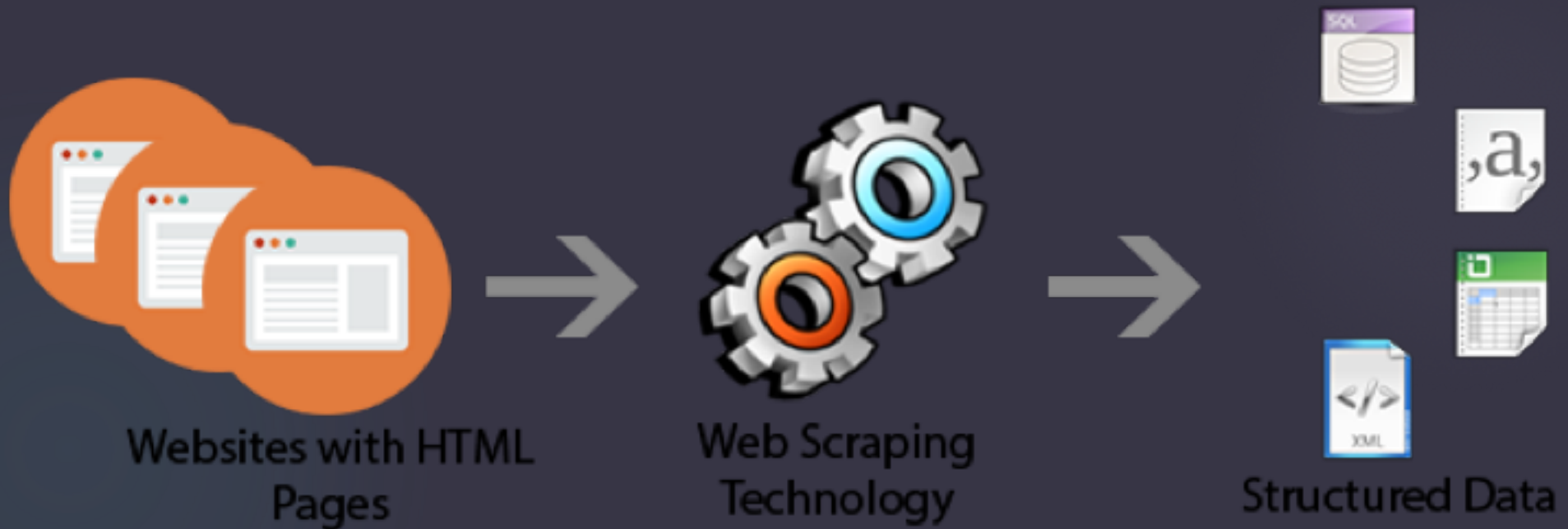
Scraping Recap

PART III

Programming Philosophy

- Concepts are key.
- Syntax is secondary.
- Stackoverflow is your friend.

What is scraping?



Scraping Process // Battle Plan

- 1. Surveillance
  - Evaluate the page, learn the terrain.
- 2. Plan of Attack
  - Brainstorm ways to approach the enemy.
- 3. Write code
  - Be willing to change your strategy if you encounter obstacles or see another "weakness" to exploit.
- 4. Emerge bloodied, yet victorious.
  - Verify the data before all that syntax evaporates from your short term memory.

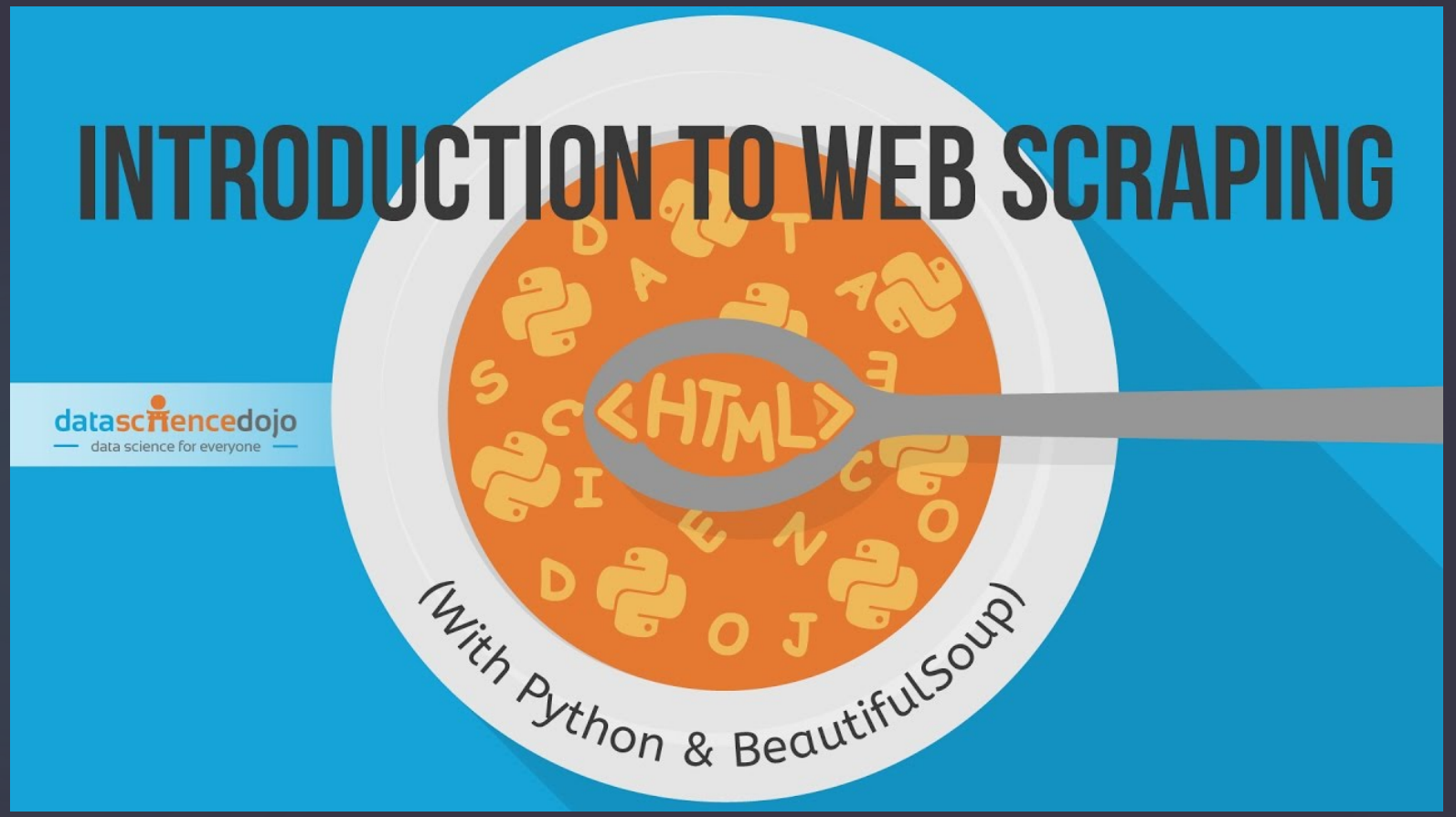
Workshop Example

PART IV

GOAL

- Scrape all text in the table as well as URLs to download files.
  - Save data as a csv file that preserves the table format.
  - Save URLs on separate lines in a txt file.

Package: BeautifulSoup



Hands-on Demo

- Open Anaconda Navigator. Open the Jupyter Notebook.
  - Navigate to "**workshop2.ipynb**" and open.

- http://www.goes-r.gov/users/2016-OCONUS.html

Downloading Files from urls.txt

- **Terminal**
  - `for i in \`cat urls.txt\`; do curl -O $i; done`
- **Powershell** (courtesy of Ryann & Keith!)
  - `foreach ($file in Get-Content url.txt) {echo "downloading $file"; curl -O $file}`

Verify Data

PART V

Make notes, be organized

- This kind of piecemeal code is hard to come back to later, so if you'll need it again, organize it and write yourself notes.
  - SYNTAX WILL LEAK OUT OF YOUR BRAIN FASTER THAN ALL THE OTHER IMPORTANT THINGS THAT YOU HAVE ALREADY FORGOTTEN YOU WERE SUPPOSED TO REMEMBER. PLEASE BELIEVE ME THAT JUST A FEW SHORT WEEKS FROM NOW YOU WILL NOT KNOW WHY YOU DID THAT THING YOU DID OR WHAT THAT VARIABLE MEANS OR WHAT THAT FUNCTION DOES AND WHY YOU APPARENTLY PUT THAT THERE. AND WHY ISN'T THIS PACKAGE UPDATE WORKING WITH MY OLD CODE AND DO I EVEN HAVE THE CORRECT THINGS INSTALLED TO MAKE THIS ALL WORK AGAIN?