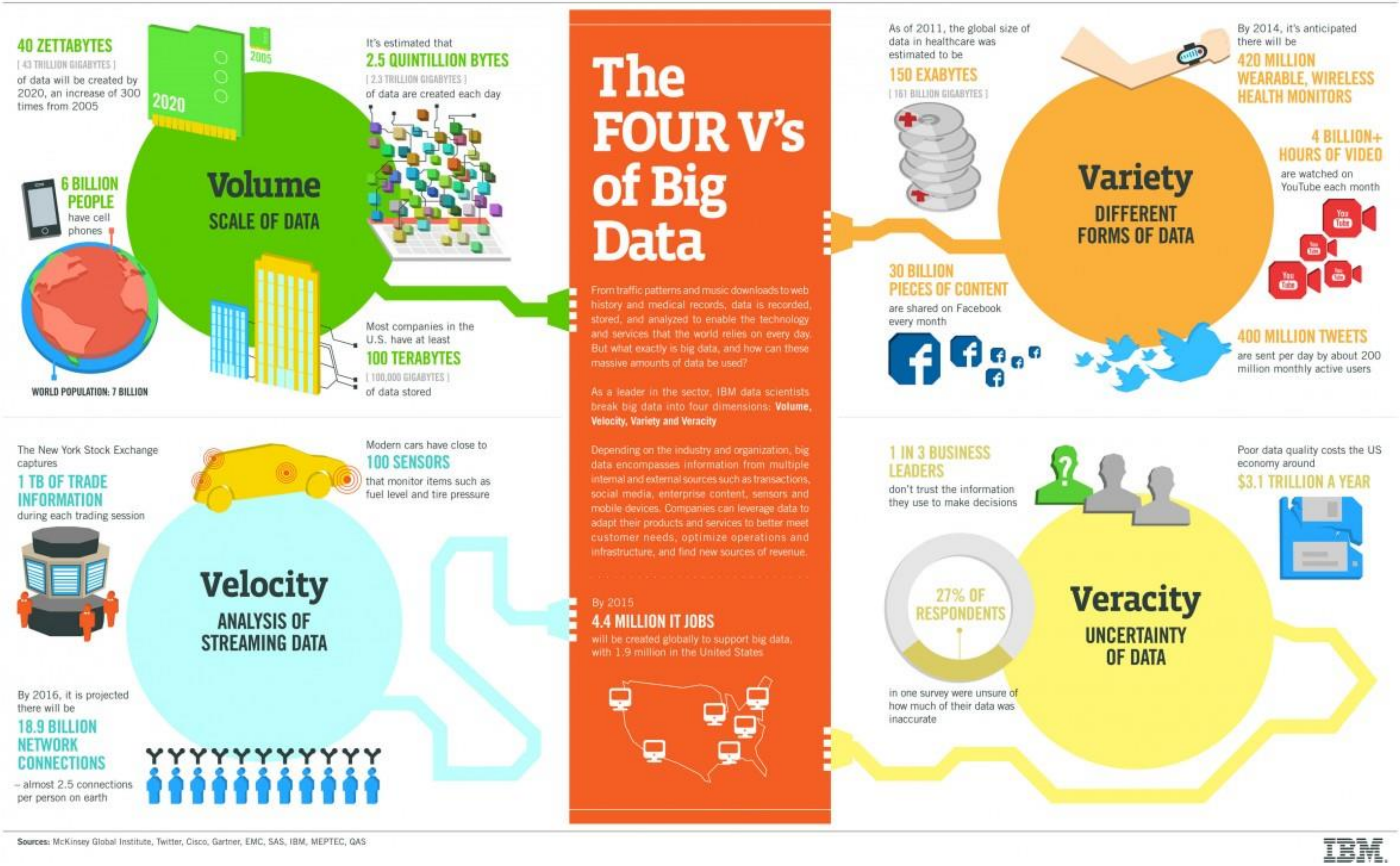
Web GIS 1: Big Data & GIS

ENV 859 – Advanced GIS

What is Big Data?



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

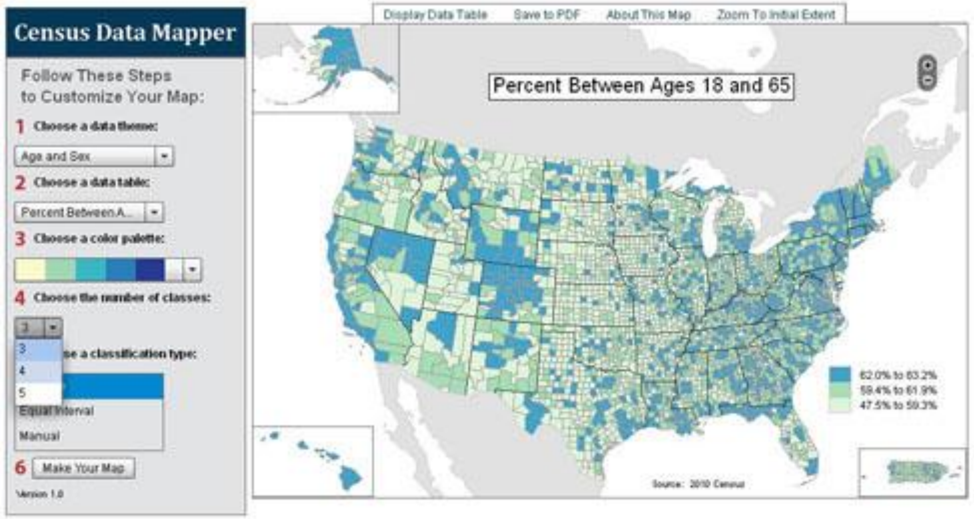
Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Big Data & GIS



Census Data Mapper

Follow These Steps
to Customize Your Map:

1 Choose a data theme: Age and Sex

2 Choose a data table: Percent Between A...

3 Choose a color palette:

4 Choose the number of classes: 3

…se a classification type:

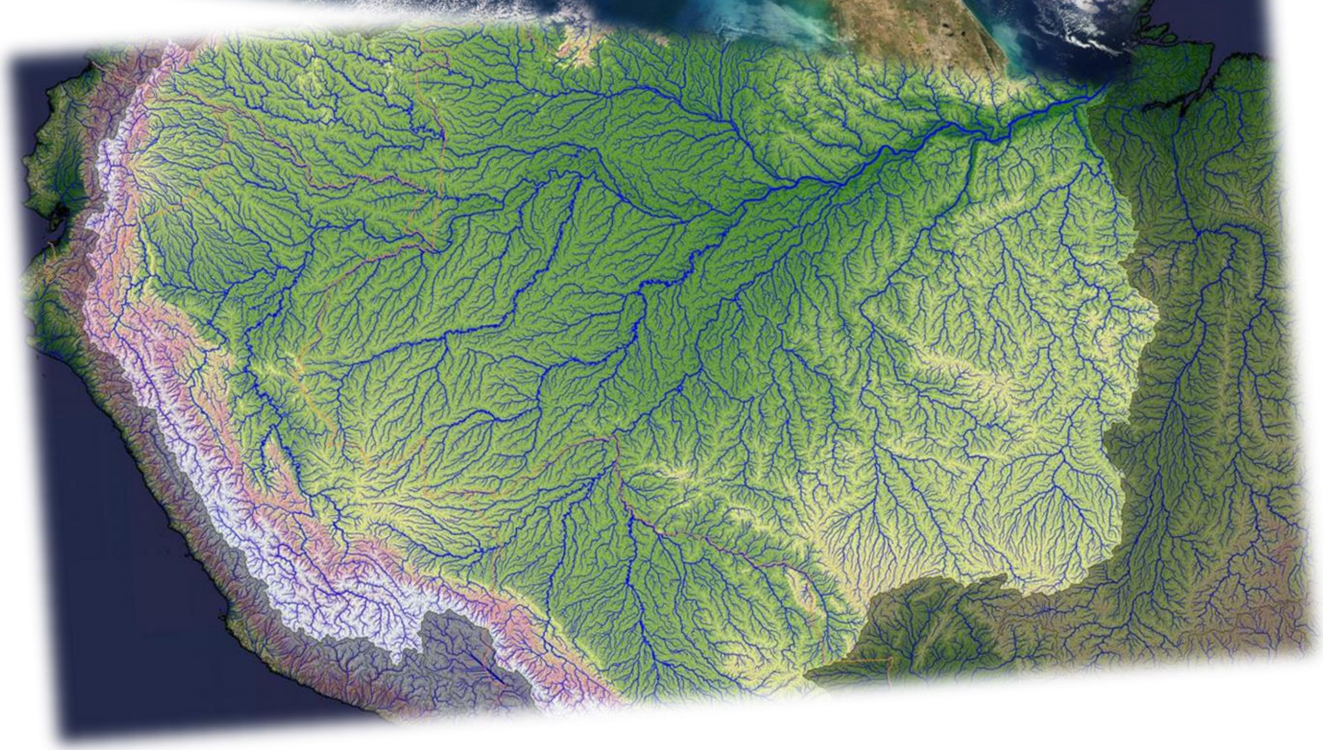
6 Make Your Map

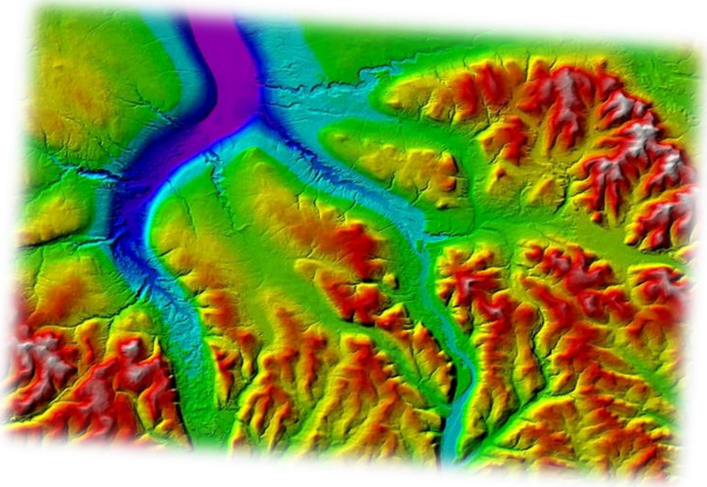
Version 1.0

Percent Between Ages 18 and 65

🟦	62.0% to 63.2%
🟩	59.4% to 61.9%
⬜	47.5% to 59.3%

Source: 2010 Census





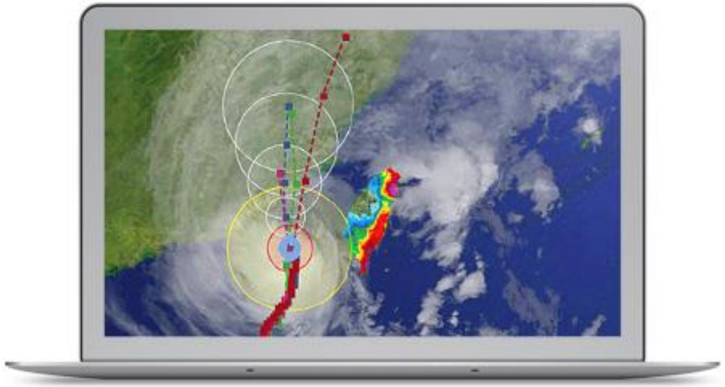
Big Data & GIS

http://www.esri.com/products/arcgis-capabilities/big-data

Big Data & GIS

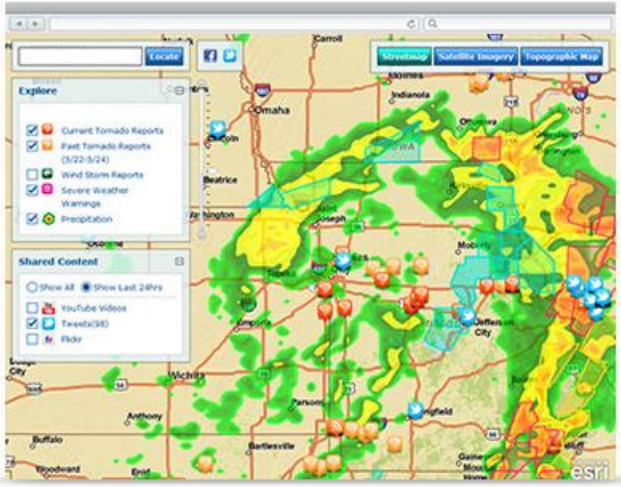
Exposing Patterns

Predictive Modeling



Finding Relationships

Link with Social Media



Explore

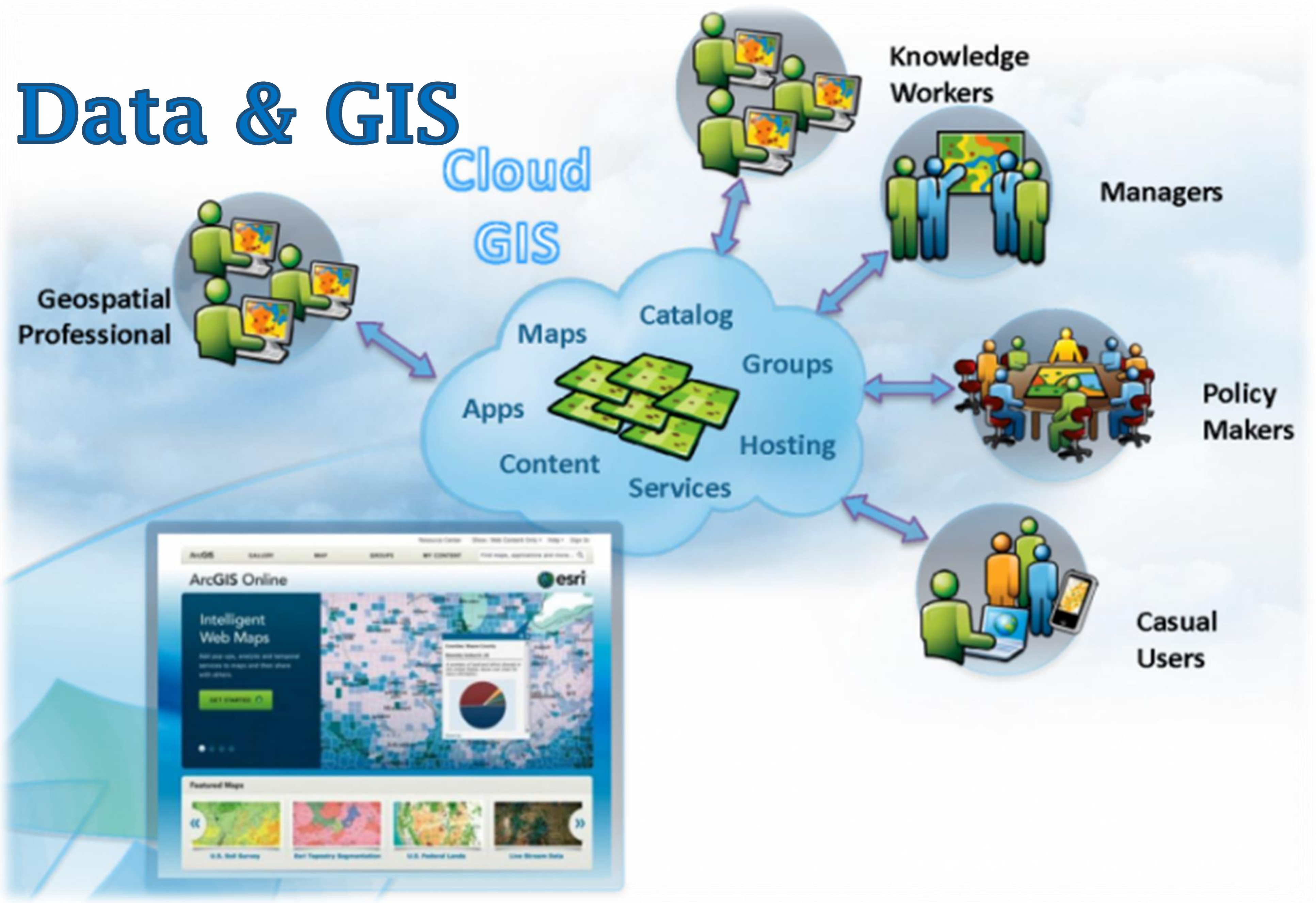
☑ Current Tornado Reports
☑ Past Tornado Reports (3/22-3/24)
☐ Wind Storm Reports
☑ Severe Weather Warnings
☑ Precipitation

Shared Content

○ Show All ● Show Last 24hrs
☐ YouTube Videos
☑ Tweets(98)
☐ Flickr

http://www.esri.com/products/arcgis-capabilities/big-data

Big Data & GIS


Cloud GIS
Geospatial Professional
Knowledge Workers
Managers
Policy Makers
Casual Users
Maps
Catalog
Groups
Apps
Hosting
Content
Services
ArcGIS Online
esri
Intelligent Web Maps

Big Data, GIS, & *US*

How do we, as GIS users, leverage the Big Data revolution?

- **Tapping into Big datasets**

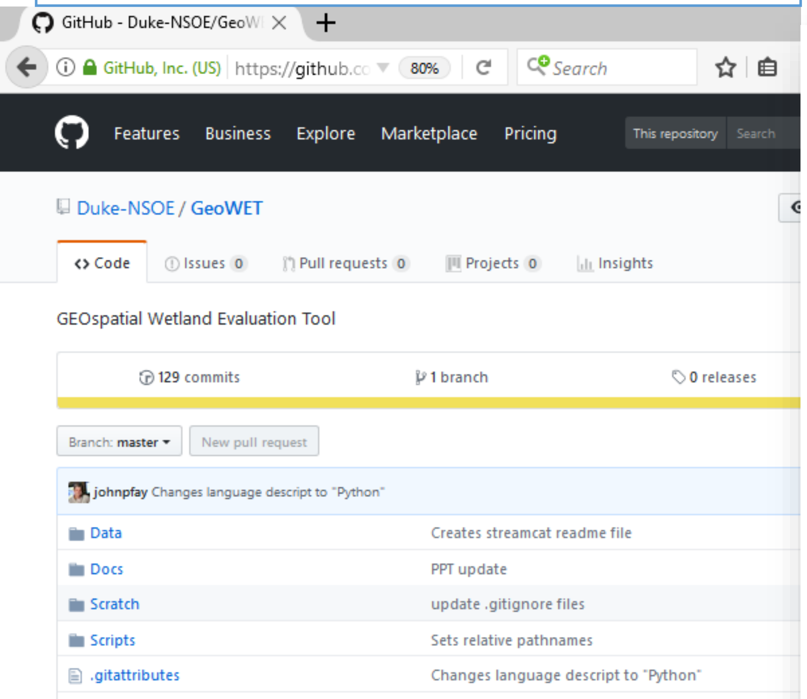
Today

  - Automating data downloads
  - Direct access to Big data

- **Beyond the Desktop**
  - Data processing in the cloud
  - Building lightweight apps

Automating data download w/

https://github.com/Duke-NSOE/GeoWET



Duke-NSOE / GeoWET

<> Code | Issues 0 | Pull requests 0 | Projects 0 | Insights

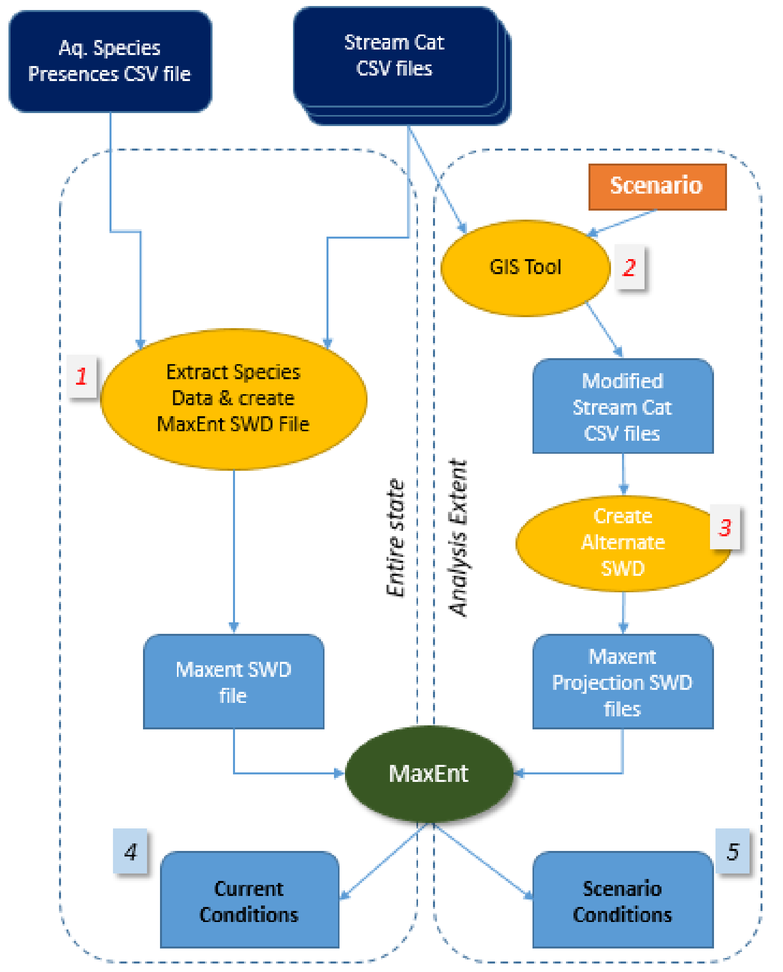
GEOspatial Wetland Evaluation Tool

129 commits | 1 branch | 0 releases

Branch: master ▾ | New pull request

johnpfay Changes language descript to "Python"

Data	Creates streamcat readme file
Docs	PPT update
Scratch	update .gitignore files
Scripts	Sets relative pathnames
.gitattributes	Changes language descript to "Python"



1

- Identifies the HUC8s in which the species was found and extracts all StreamCat catchment records within them.
- Removes any <u>records</u> with missing data where the species was not found and then any <u>attributes</u> with no data where the species was found.
- Removes any <u>attributes</u> with no significant correlation with presence/absence ($p > 0.05$). Then identifies cross-correlated attributes pairs ($r > 0.75$) and removes the one with the least correlation with presence/absence.
- Formats columns and column names to suit the MaxEnt species with data (SWD) format.

2

- Allows user to draw a shape on a map reflecting a change in land cover type.
- The user also designates the analysis extent for projecting uplift. This is usually the HUC 6 in which the modification occurs.
- Based on this change, adjusts values in appropriate StreamCat attribute values for affected records.

3

- Converts each set of StreamCat within the analysis extent (e.g. HUC 6) into its own Maxent SWD files that can be used as MaxEnt projection scenarios.

4

- Listing of each catchment and the estimated percent likelihood of finding the species there based on current conditions (<u>unmodified</u> StreamCat values).

5

- Listing of each catchment and the estimated percent likelihood of finding the species there based on altered conditions (<u>modified</u> StreamCat values).

Automating data download w/

Why automate?

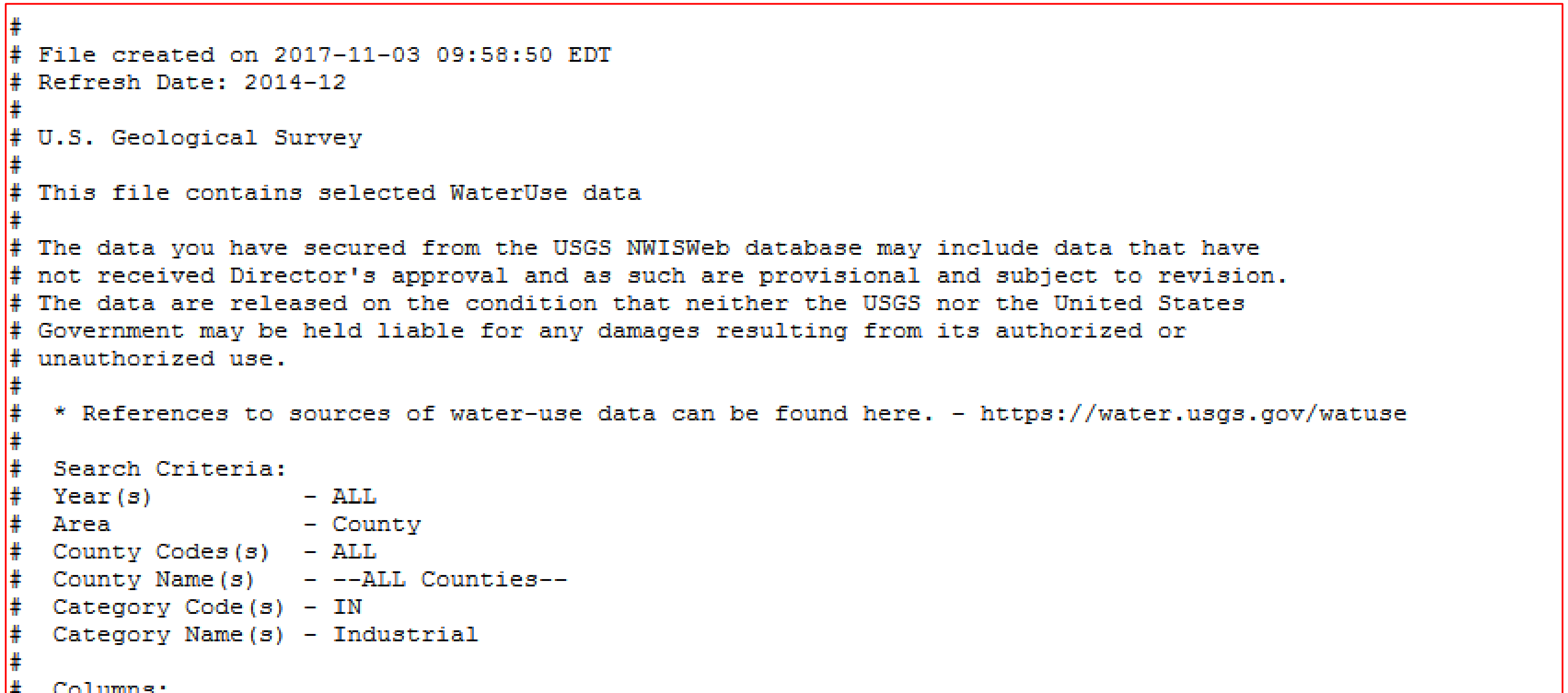
- Too time intensive to acquire manually…
- Update or reuse for new data…
- Reproducibility…

- Some data are only available through an Application Programming Interface (API)…

Tiers of accessing on-line data

- Grabbing static text from a web site via its web address (URL)

https://waterdata.usgs.gov/nc/nwis/water_use?format=rdb&rdb_compression=value&wu_area=County&wu_year=ALL&wu_county=ALL&wu_category=IN&wu_county_nms=--ALL%2BCounties--&wu_category_nms=Industrial

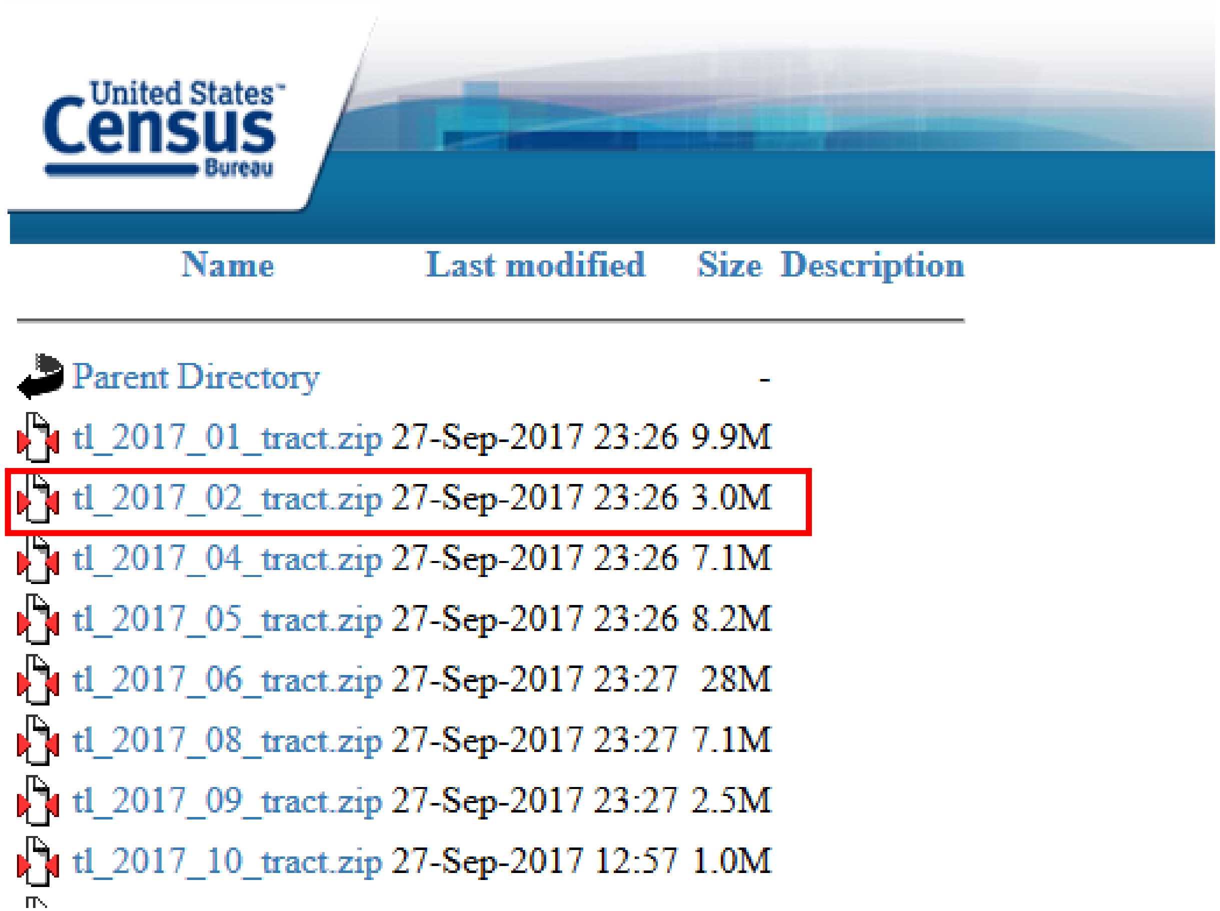


```
#
# File created on 2017-11-03 09:58:50 EDT
# Refresh Date: 2014-12
#
# U.S. Geological Survey
#
# This file contains selected WaterUse data
#
# The data you have secured from the USGS NWISWeb database may include data that have
# not received Director's approval and as such are provisional and subject to revision.
# The data are released on the condition that neither the USGS nor the United States
# Government may be held liable for any damages resulting from its authorized or
# unauthorized use.
#
#  * References to sources of water-use data can be found here. - https://water.usgs.gov/watuse
#
#  Search Criteria:
#  Year(s)           - ALL
#  Area              - County
#  County Codes(s)   - ALL
#  County Name(s)    - --ALL Counties--
#  Category Code(s)  - IN
#  Category Name(s)  - Industrial
#
#  Columns:
```

Tiers of accessing on-line data

- Grabbing hosted binary file(s) from a web address

https://www2.census.gov/geo/tiger/TIGER2017/TRACT/

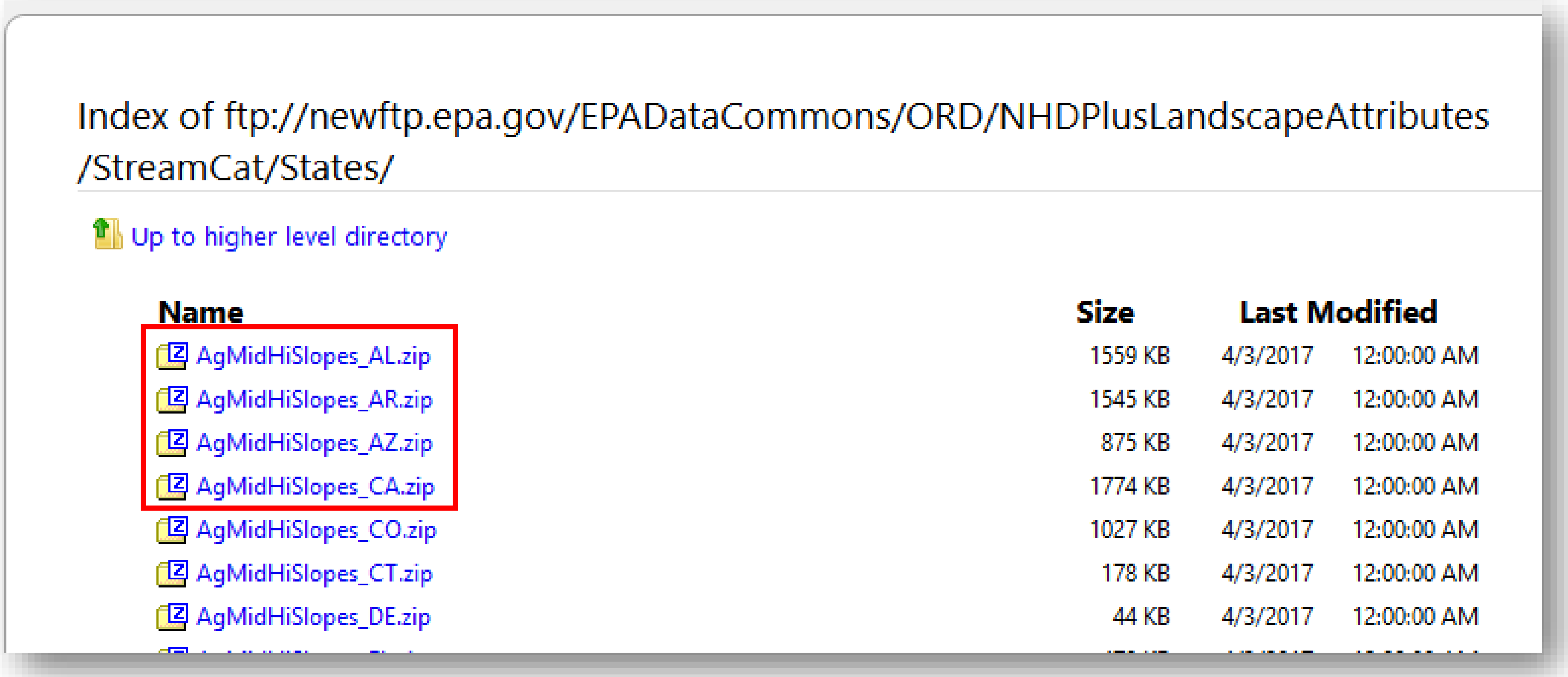


Name	Last modified	Size	Description
Parent Directory		-	
tl_2017_01_tract.zip	27-Sep-2017 23:26	9.9M	
tl_2017_02_tract.zip	27-Sep-2017 23:26	3.0M	
tl_2017_04_tract.zip	27-Sep-2017 23:26	7.1M	
tl_2017_05_tract.zip	27-Sep-2017 23:26	8.2M	
tl_2017_06_tract.zip	27-Sep-2017 23:27	28M	
tl_2017_08_tract.zip	27-Sep-2017 23:27	7.1M	
tl_2017_09_tract.zip	27-Sep-2017 23:27	2.5M	
tl_2017_10_tract.zip	27-Sep-2017 12:57	1.0M	

Tiers of accessing on-line data

- Grabbing hosted binary file(s) from an FTP server

ftp://newftp.epa.gov/EPADataCommons/ORD/NHDPlusLandscapeAttributes/StreamCat/States



Index of ftp://newftp.epa.gov/EPADataCommons/ORD/NHDPlusLandscapeAttributes/StreamCat/States/

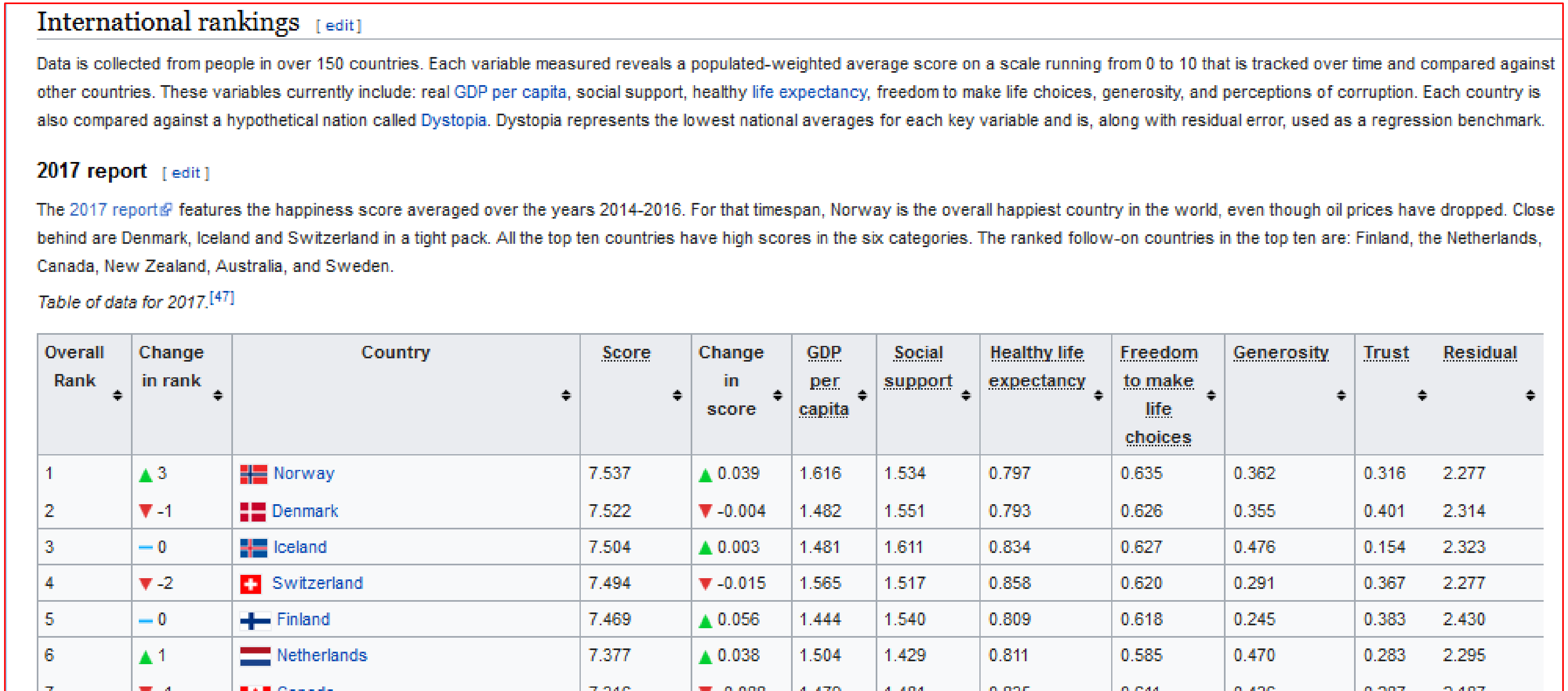
Up to higher level directory

Name	Size	Last Modified	
AgMidHiSlopes_AL.zip	1559 KB	4/3/2017	12:00:00 AM
AgMidHiSlopes_AR.zip	1545 KB	4/3/2017	12:00:00 AM
AgMidHiSlopes_AZ.zip	875 KB	4/3/2017	12:00:00 AM
AgMidHiSlopes_CA.zip	1774 KB	4/3/2017	12:00:00 AM
AgMidHiSlopes_CO.zip	1027 KB	4/3/2017	12:00:00 AM
AgMidHiSlopes_CT.zip	178 KB	4/3/2017	12:00:00 AM
AgMidHiSlopes_DE.zip	44 KB	4/3/2017	12:00:00 AM

Tiers of accessing on-line data

- Grabbing a table seen on a web page

https://en.wikipedia.org/wiki/World_Happiness_Report#International_rankings



International rankings [edit]

Data is collected from people in over 150 countries. Each variable measured reveals a populated-weighted average score on a scale running from 0 to 10 that is tracked over time and compared against other countries. These variables currently include: real GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption. Each country is also compared against a hypothetical nation called Dystopia. Dystopia represents the lowest national averages for each key variable and is, along with residual error, used as a regression benchmark.

2017 report [edit]

The 2017 report features the happiness score averaged over the years 2014-2016. For that timespan, Norway is the overall happiest country in the world, even though oil prices have dropped. Close behind are Denmark, Iceland and Switzerland in a tight pack. All the top ten countries have high scores in the six categories. The ranked follow-on countries in the top ten are: Finland, the Netherlands, Canada, New Zealand, Australia, and Sweden.

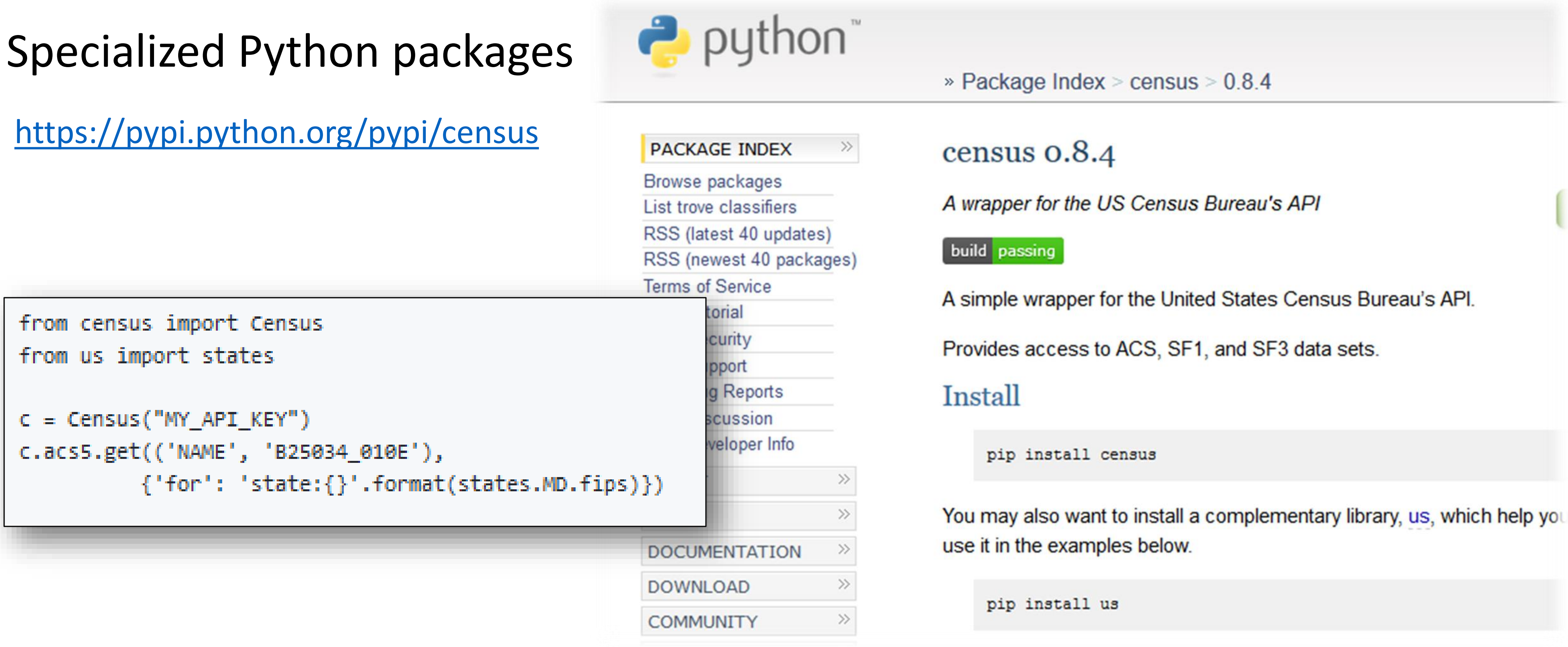
Table of data for 2017.[47]

Overall Rank	Change in rank	Country	Score	Change in score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Trust	Residual
1	▲ 3	Norway	7.537	▲ 0.039	1.616	1.534	0.797	0.635	0.362	0.316	2.277
2	▼ -1	Denmark	7.522	▼ -0.004	1.482	1.551	0.793	0.626	0.355	0.401	2.314
3	— 0	Iceland	7.504	▲ 0.003	1.481	1.611	0.834	0.627	0.476	0.154	2.323
4	▼ -2	Switzerland	7.494	▼ -0.015	1.565	1.517	0.858	0.620	0.291	0.367	2.277
5	— 0	Finland	7.469	▲ 0.056	1.444	1.540	0.809	0.618	0.245	0.383	2.430
6	▲ 1	Netherlands	7.377	▲ 0.038	1.504	1.429	0.811	0.585	0.470	0.283	2.295
7	▼ -1	Canada	7.316	▼ -0.088	1.479	1.481	0.835	0.611	0.436	0.287	2.187

Tiers of accessing on-line data

- Specialized Python packages

https://pypi.python.org/pypi/census



```
from census import Census
from us import states

c = Census("MY_API_KEY")
c.acs5.get(('NAME', 'B25034_010E'),
          {'for': 'state:{}'.format(states.MD.fips)})
```

census 0.8.4

A wrapper for the US Census Bureau's API

A simple wrapper for the United States Census Bureau's API.

Provides access to ACS, SF1, and SF3 data sets.

Install

```
pip install census
```

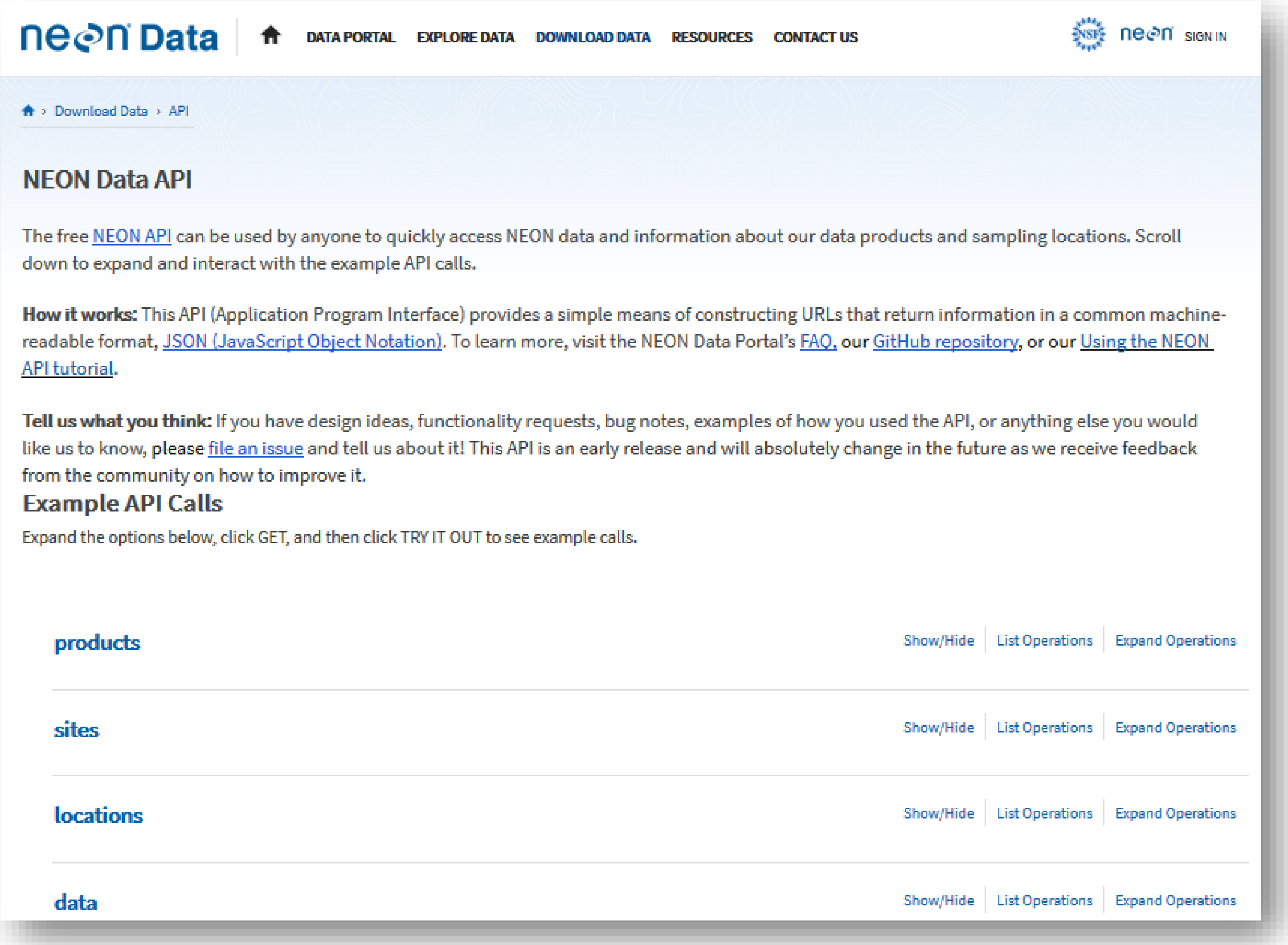
You may also want to install a complementary library, us, which help yo use it in the examples below.

```
pip install us
```

Tiers of accessing on-line data

- Access via API

  http://data.neonscience.org/data-api



NEON Data API

The free NEON API can be used by anyone to quickly access NEON data and information about our data products and sampling locations. Scroll down to expand and interact with the example API calls.

How it works: This API (Application Program Interface) provides a simple means of constructing URLs that return information in a common machine-readable format, JSON (JavaScript Object Notation). To learn more, visit the NEON Data Portal's FAQ, our GitHub repository, or our Using the NEON API tutorial.

Tell us what you think: If you have design ideas, functionality requests, bug notes, examples of how you used the API, or anything else you would like us to know, please file an issue and tell us about it! This API is an early release and will absolutely change in the future as we receive feedback from the community on how to improve it.

Example API Calls

Expand the options below, click GET, and then click TRY IT OUT to see example calls.

products
Show/Hide | List Operations | Expand Operations
sites
Show/Hide | List Operations | Expand Operations
locations
Show/Hide | List Operations | Expand Operations
data
Show/Hide | List Operations | Expand Operations

Tiers of accessing on-line data

- Grabbing static text or a file from a web site via its web address (URL)
- Bulk downloading files from a web or ftp server

- Grabbing (and converting) data seen on a web page – “Scraping”

- Specialized Python modules for accessing on-line data

- Using Application Programming Interfaces (APIs) to pull data

Diving in!

- Download the zip file (or sync from GIT)

- Run through examples & discuss what’s going on…