**Basic principles:**

* Every sector needs both a physical and an economic weightline dataset, as well as a baseline dataset.
* A baseline dataset describes the dependent variable in recent history-- for example, deaths for health or production for agriculture.
* A weightline dataset describes the “denominator” of the variable for use in aggregation-- for example, population for health or planted area for agriculture.
* We want to collect yearly, panel datasets to generate consistent averages now and update them in the future.
* Each baseline or weightline dataset can consist of multiple files, each derived from a single source or process.
* Each baseline or weightline file contains a numerical value for each of a subset of years and a subset of the GCP regions (see the region hierarchy).
* Every weightline file requires meta-data information, including its version, input sources (including their versions), the generating script path, and definitions of the meaning and units of the numerical values.
* The multiple files that make up a baseline or weightline dataset come in an order, where the first is least precise (but typically most comprehensive) and subsequent files override previously defined regions with more precise (but typically country-specific) data
* The generating script should be able to be run with minimal setup (automatically downloading files when possible, and using only relative paths)

**How to contribute baseline and weightline data:**

1. Upload the original data to a subfolder of /shares/gcp/data/sources specific to the collecting agency.
2. Add a .fgh file with the same format as shown in the [Header Specification](https://docs.google.com/document/d/1aCc-erfuu2jLtwWPh2wjUNX0u6buDBwfVJ8bwnrNRxk/edit?usp=sharing), creating a version for the file and units for the variables contained with in.
3. Write a script that translates this data into a baseline or weightline file, with meta-data, capable of being run with minimum setup. The meta-data should include a reference to this script that creates it. See the **Delimited** and **NetCDF** sections below for the standardized format. Only data starting in the year 2000 is important, if it is easier to only process a subset of data.
4. Add your script to the <https://bitbucket.org/globalclimate/socioeconomics/src> bitbucket, under the baselines directory.
5. Run your script on Shackleton, making sure that you need to make no changes to the script itself in order to run it.
6. Add your file to a subfolder of /shares/gcp/data/weightlines or /shares/gcp/data/baselines named <sector>-physical or <sector>-economic.
7. If you are making this folder for the first time, create a file there called index.txt, with just the new file filename; then update the [Master DMAS Information](https://docs.google.com/spreadsheets/d/1lyvAeoUTji-FGH_Fz-hGWQOnEWJb2EMhWgc7LKz2ix0/edit#gid=0) spreadsheet, filling the path to this file relative to the weightlines directory (e.g., violent-crime-physical/index.txt) into the Physical Weightlines or Economic Weightlines column.
8. If you are adding to an existing folder, add the new filename to index.txt, in order with lower files overriding higher ones.

**Standardized Delimited File Format**

If you are generating a .csv or .tsv file, start the file with a meta-data header, formatted as shown in the [Header Specification](https://docs.google.com/document/d/1aCc-erfuu2jLtwWPh2wjUNX0u6buDBwfVJ8bwnrNRxk/edit?usp=sharing).

The next line should be column header, as follows in .csv:

year,region,value,notes

**Standardized NetCDF Format**

Please see the [Header Specification](https://docs.google.com/document/d/1aCc-erfuu2jLtwWPh2wjUNX0u6buDBwfVJ8bwnrNRxk/edit?usp=sharing) for fields to be stored in the NetCDF file.

The dimensions of the NetCDF file should be region and year.

**Generating weightlines from gridded datasets**

The [VersioningExample](https://bitbucket.org/globalclimate/socioeconomics/wiki/VersioningExample) describes how to combine gridded data with national data to generate information at the GCP regions. Although these scripts are not currently in a form that can be run for other gridded and national datasets, James will prepare them for more general use. When a combination of a gridded dataset and a national dataset is prepared, with meta-data, contact James to use as a test case.

**Header Format**

All headers should have the following information:

* File version
* Dependencies and their versions
* Variables defined and their dimension

A source file would have a header like this:

---

oneline: The Global Administrative Areas Database  
version: GADM.2.0

variables:

pop: Population [individuals]

description: Includes all countries. Downloaded from http://www.gadm.org/ between December 2011 (the release of version 2.0) and August 2015.  
...

A processed file might have a header like this:

---

oneline: UN-conforming populations for GADM2 regions  
version: Paris.1  
dependencies: GADM.2.0, UNStats.20150213  
variables:  
 year: Year (CE)  
 region: GCP region code [str]  
 value: Estimated population [people]

notes: Region-specific comments [str]  
sources: Computed by estpop.R, which scales LandScan-augmented GADM population attributes to conform to UNStat country counts.  
...