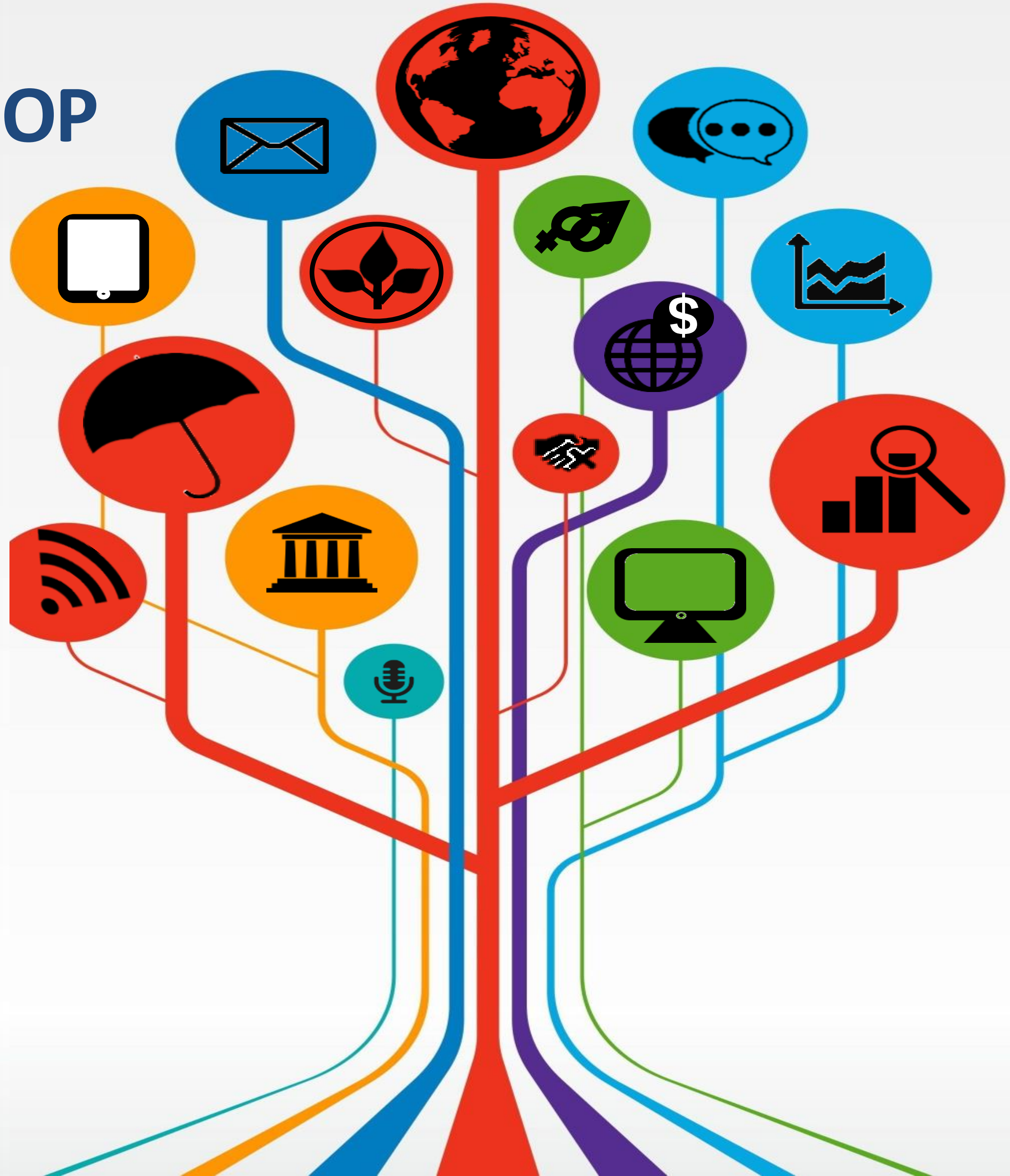
FIELD COORDINATOR WORKSHOP



Manage Successful Impact Evaluations

18 - 22 JUNE 2018
WASHINGTON, DC

i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

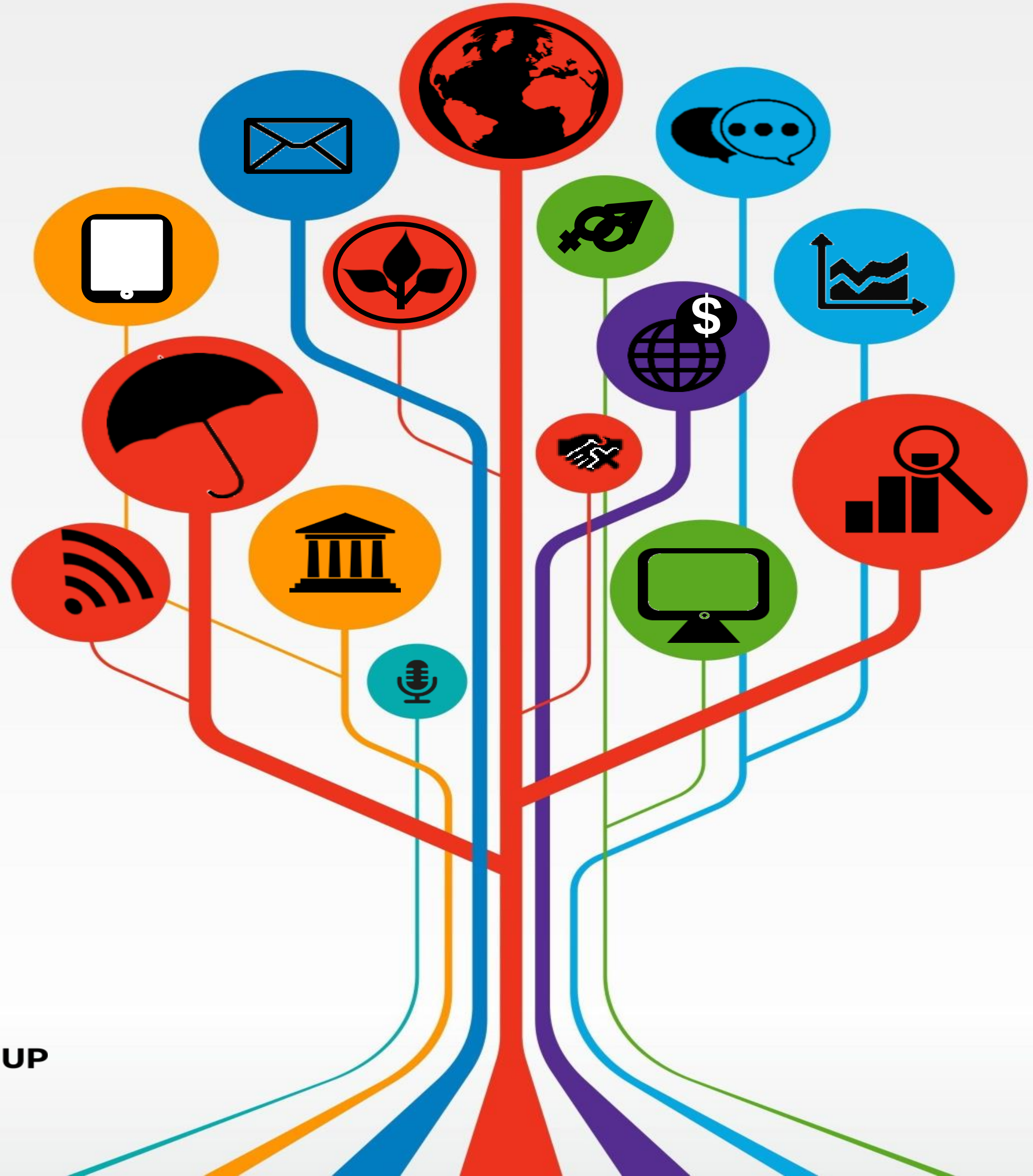
Data Management for Reproducible Research

Stata Track 2

Prepared by **DIME Analytics**
DIMEAnalytics_Internal_Use_Only@worldbank.org

Presented by **Benjamin Daniels and Roshni Khincha**
bdaniels@worldbank.org | rkhincha@worldbank.org

http://www.worldbank.org/en/research/dime
June 19, 2018



i2i DIME TRANSFORM DEVELOPMENT

WORLD BANK GROUP

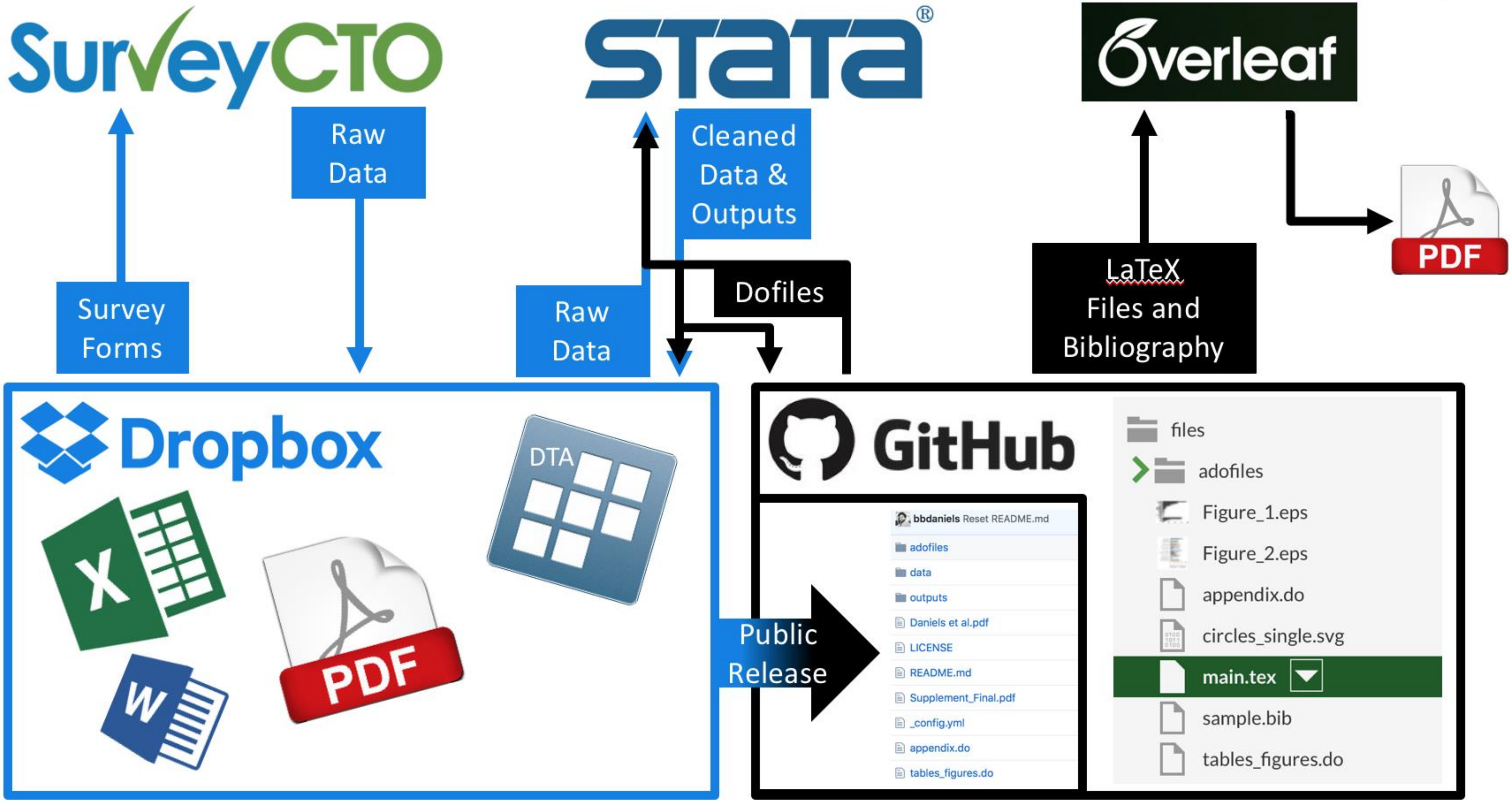
Introduction: Data Management

Data management is part of a reproducible research workflow.

- This presentation will show you best practices to manage data work
- At DIME, we have large teams collaborating on the same codes and data sets
- Long projects easily become complex, with multiple rounds of data collection to organize
- Standardizing organization of documents and code prevents mistakes and reduces the cost of transitioning across projects and teams

Production

Collection → Analysis → Publication



bbdaniels Reset README.md

- adofiles
- data
- outputs
- Daniels et al.pdf
- LICENSE
- README.md
- Supplement_Final.pdf
- _config.yml
- appendix.do
- tables_figures.do

files

- adofiles
- Figure_1.eps
- Figure_2.eps
- appendix.do
- circles_single.svg
- main.tex
- sample.bib
- tables_figures.do

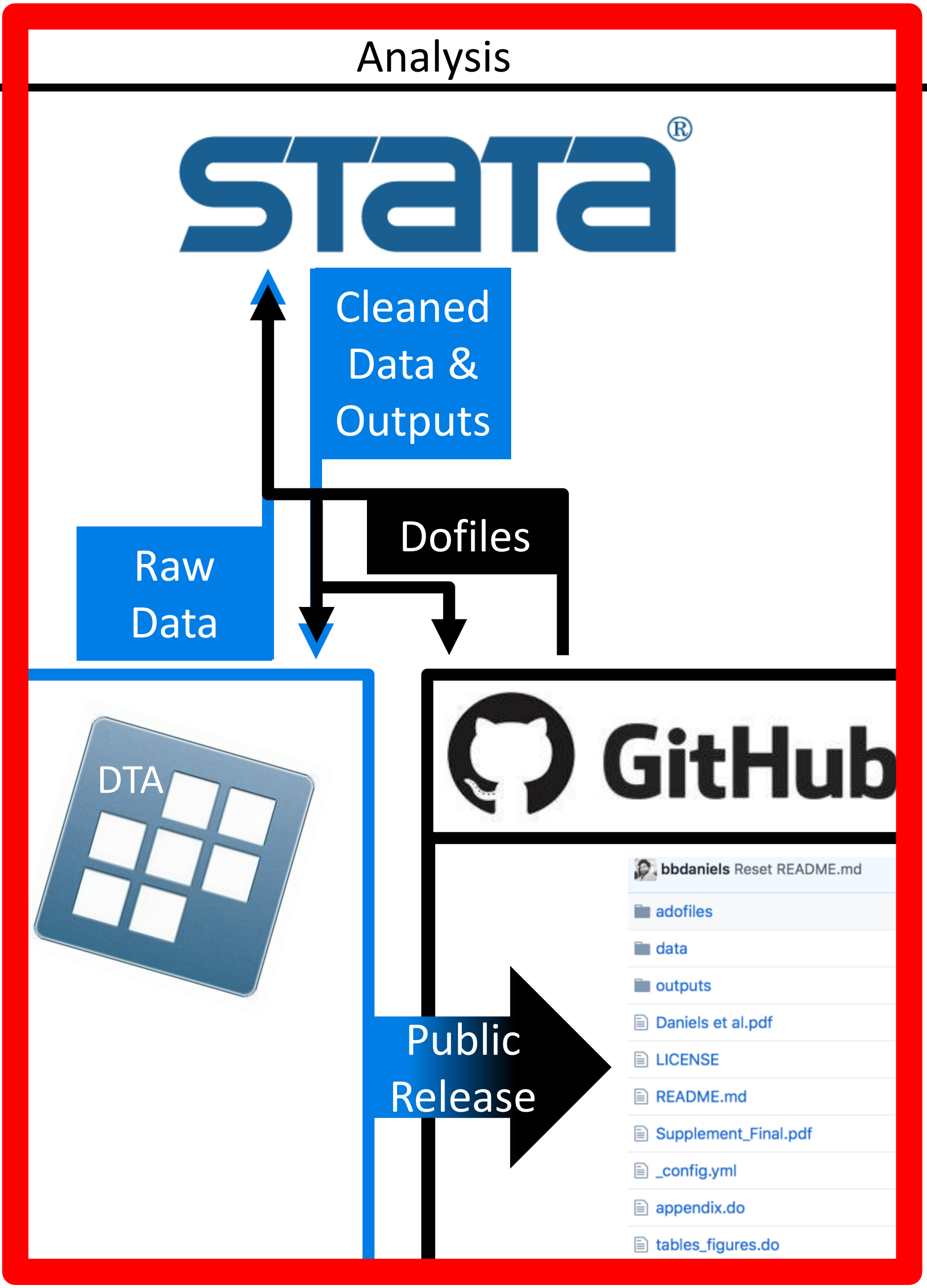
Production

Collection → Analysis → Publication

- **Publishing a paper is not enough!**

- **Code and data to reproduce results is often required by Open Access agreements or journals themselves.**

- **And even if it isn't, others may want or need to use or reuse your code in the future, so it is good academic citizenship.**



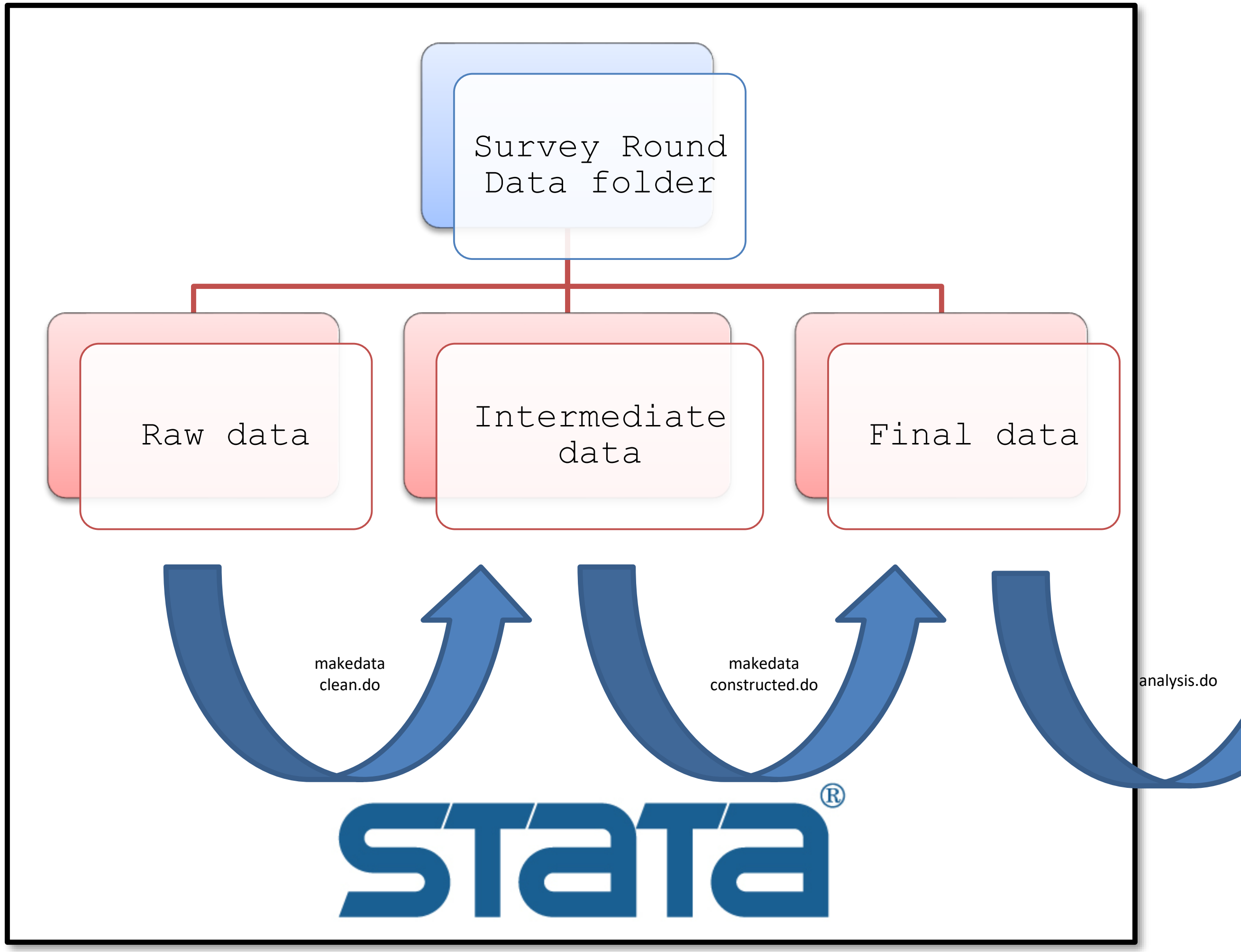
bbdaniels Reset README.md

- adofiles
- data
- outputs
- Daniels et al.pdf
- LICENSE
- README.md
- Supplement_Final.pdf
- _config.yml
- appendix.do
- tables_figures.do

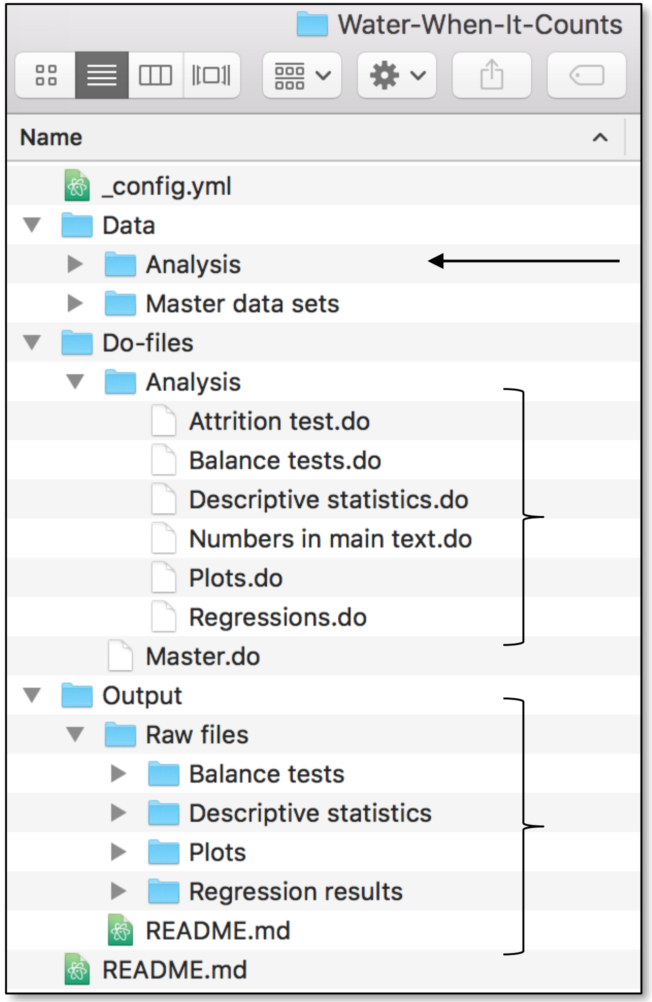
Organized code requires organized data

By the end of this presentation, you should understand how these two are connected in a research workflow:

1. The structure of the **folder** where the data work is stored
2. The **master do-file** for a project's data work



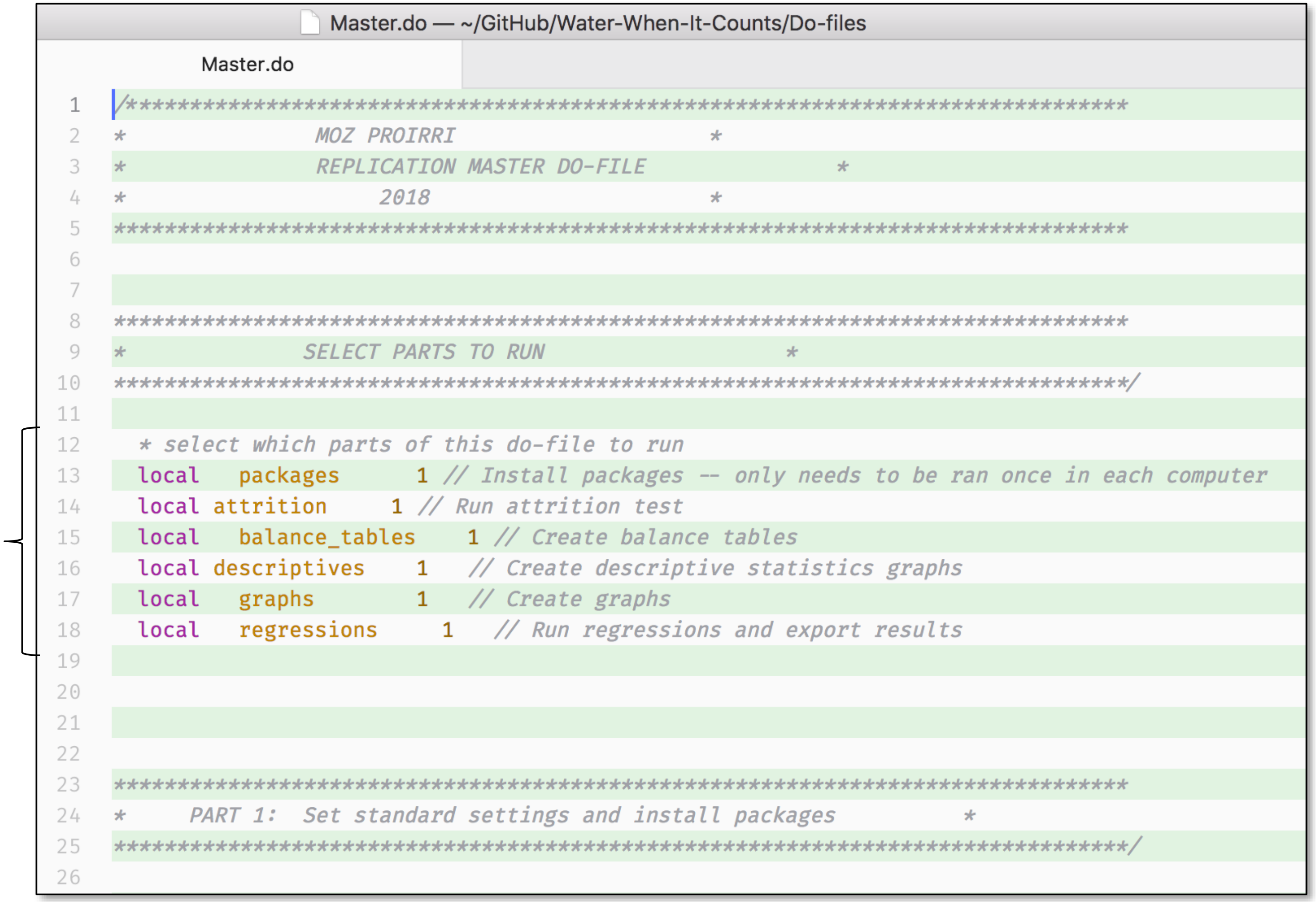
What does this look like in practice?



```
_config.yml
▼ Data
  ▶ Analysis
  ▶ Master data sets
▼ Do-files
  ▼ Analysis
      Attrition test.do
      Balance tests.do
      Descriptive statistics.do
      Numbers in main text.do
      Plots.do
      Regressions.do
    Master.do
▼ Output
  ▼ Raw files
    ▶ Balance tests
    ▶ Descriptive statistics
    ▶ Plots
    ▶ Regression results
    README.md
README.md
```

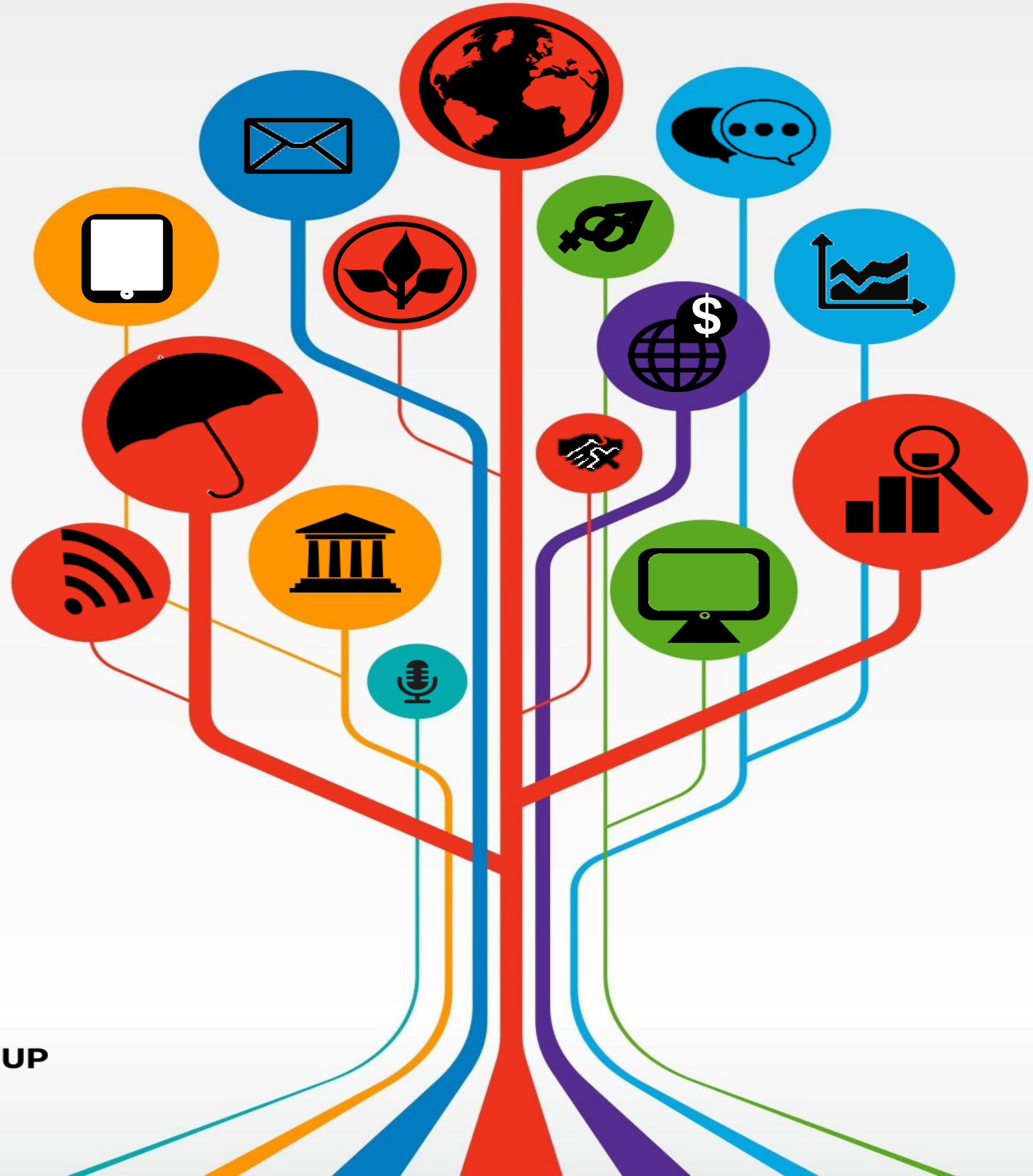
Organized data

Parallel structures



```
/*******************************************************************************
*                 MOZ PROIRRI                                   *
*                 REPLICATION MASTER DO-FILE                              *
*                      2018                                     *
********************************************************************************


********************************************************************************
*                SELECT PARTS TO RUN                                  *
********************************************************************************/

  * select which parts of this do-file to run
  local   packages      1 // Install packages -- only needs to be ran once in each computer
  local attrition      1 // Run attrition test
  local   balance_tables     1 // Create balance tables
  local descriptives     1   // Create descriptive statistics graphs
  local   graphs         1   // Create graphs
  local   regressions      1   // Run regressions and export results




********************************************************************************
*     PART 1:  Set standard settings and install packages            *
********************************************************************************/

```

Structure of the data folder

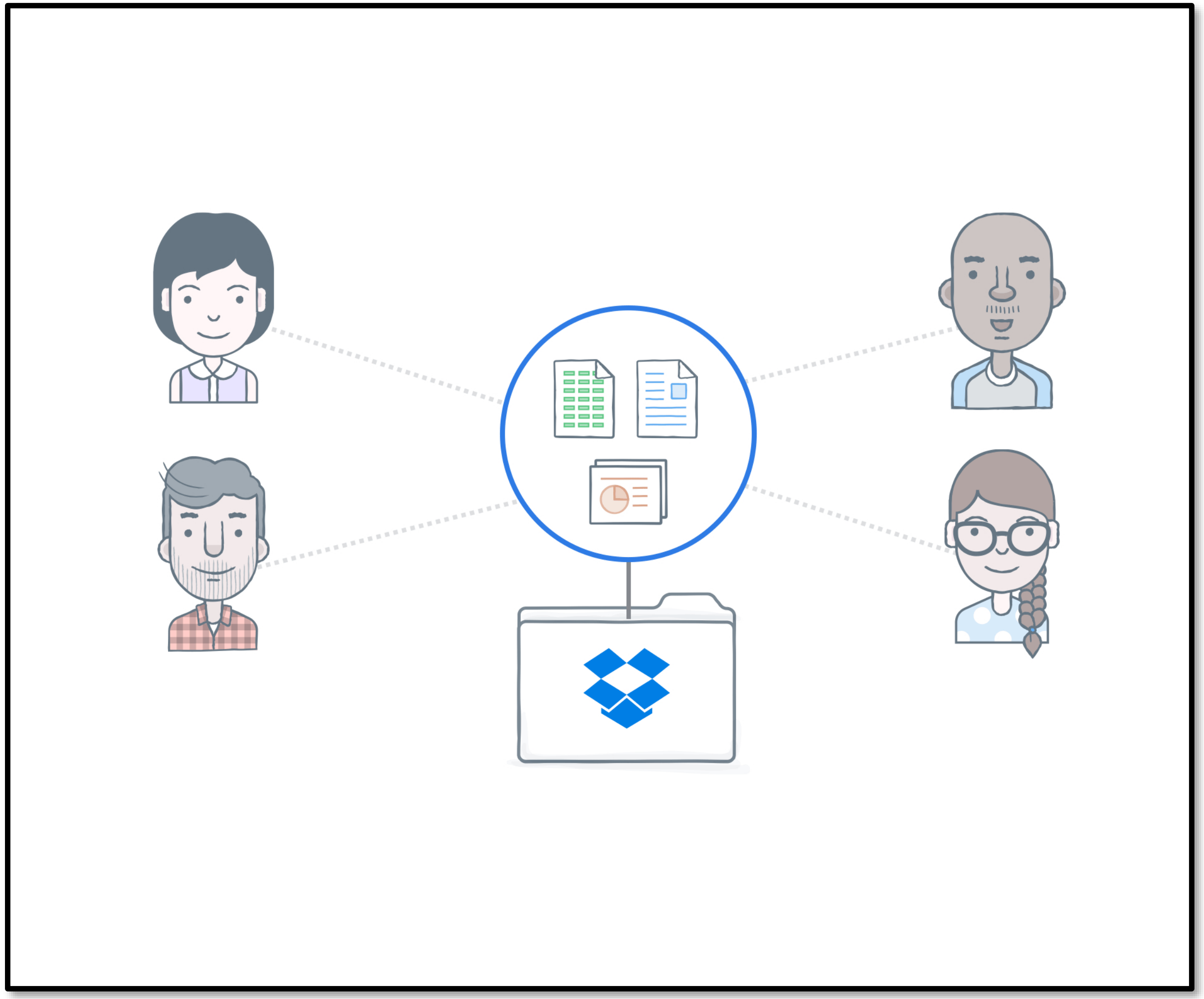


i2i DIME TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Dropbox (or equivalent) as Data Storage

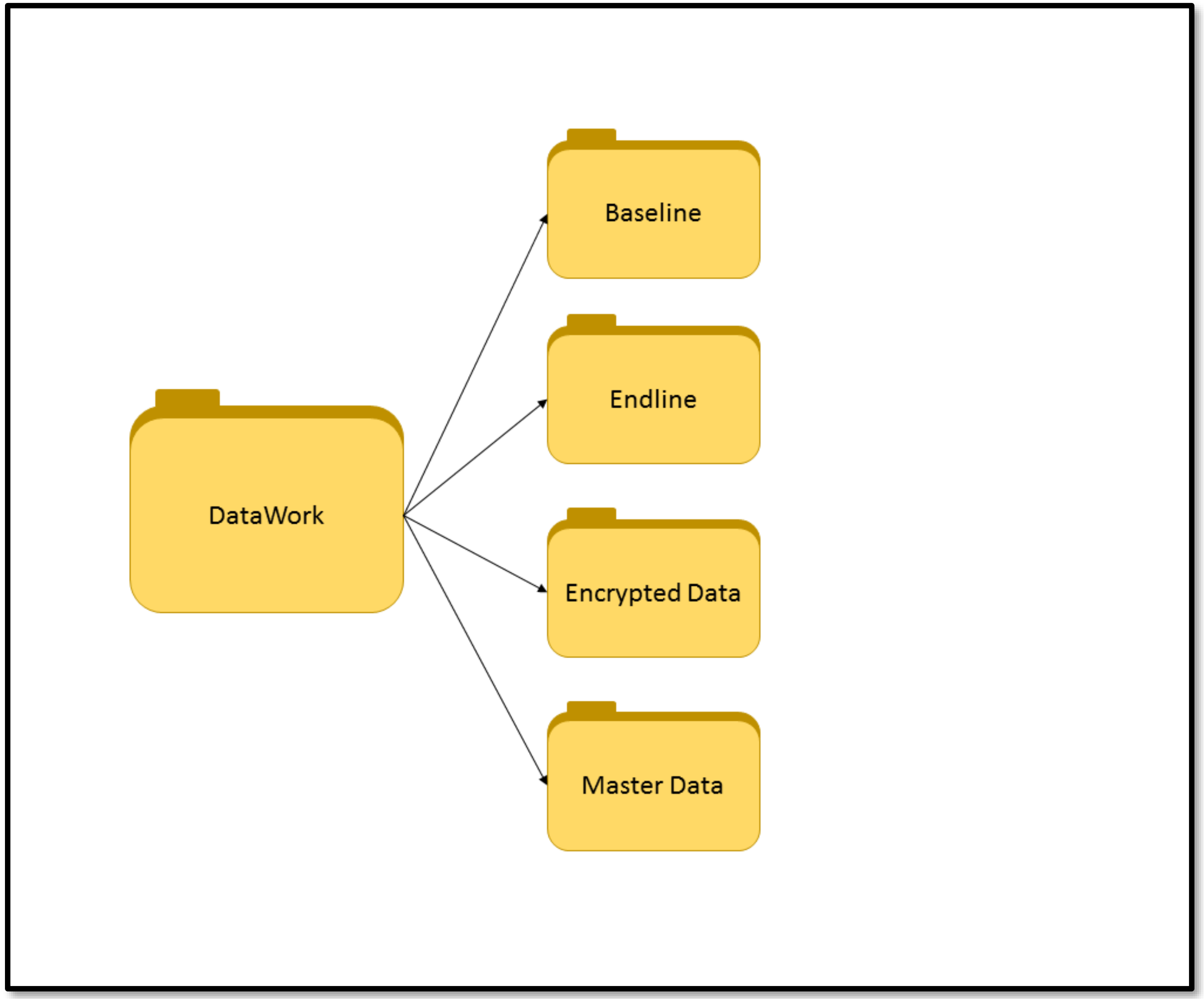
- **Privacy** – Personally-identified data is NEVER available to non-invited people (can be encrypted)
- **Efficiency** – Only the most recent version of files is stored, and individuals can opt out of subfolders
- **Version Control** – Limited version control, but allows mistakes like deletions to be corrected quickly



Structure of the data folder: overview

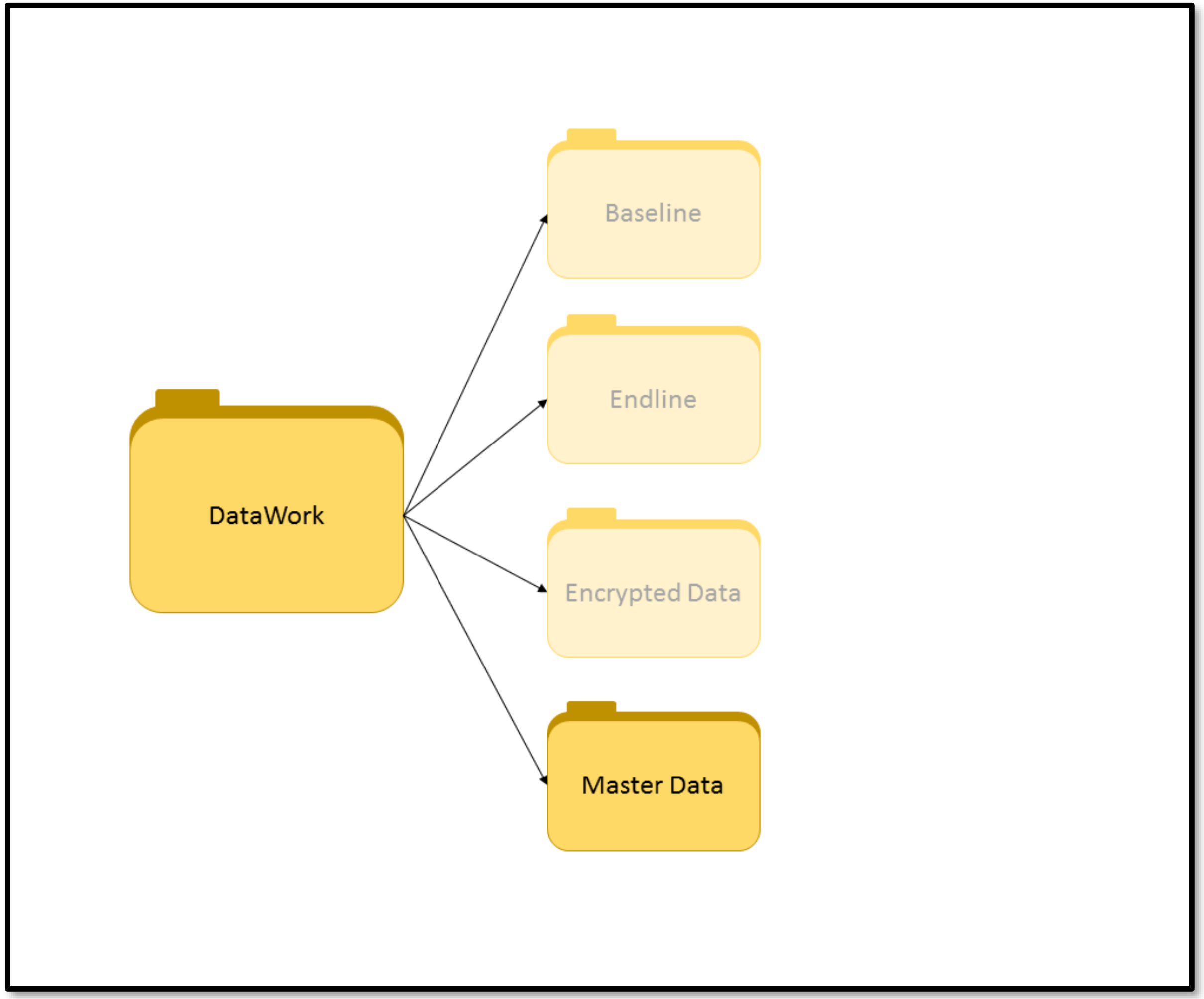
The top-level data folder contains:

1. A folder for each analysis round (in this case, one folder each for baseline and endline)
2. An encrypted folder with keys to PII data (through software like TrueCrypt or BoxCrypt)
3. A “master” file that traces all observations and linkages between data across rounds



Structure of the data folder: the master file

- Master data traces contacts across all rounds for analysis purposes that include analysis of loss to follow-up; differential attrition; and other key reporting elements of research design and execution.
- Every sampled unit that appears in every dataset here should be catalogued here across the whole project.

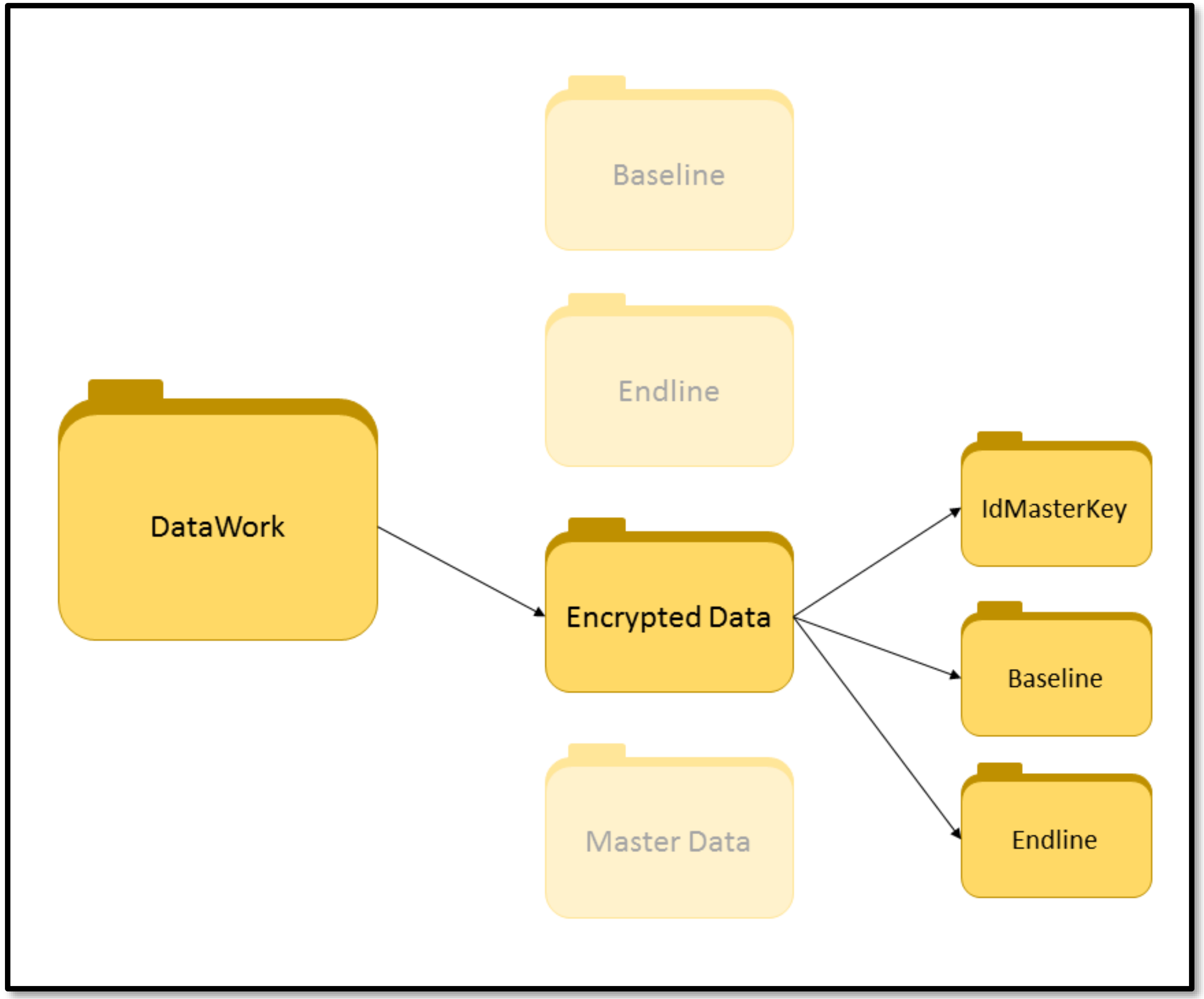


Structure of the data folder: the master file

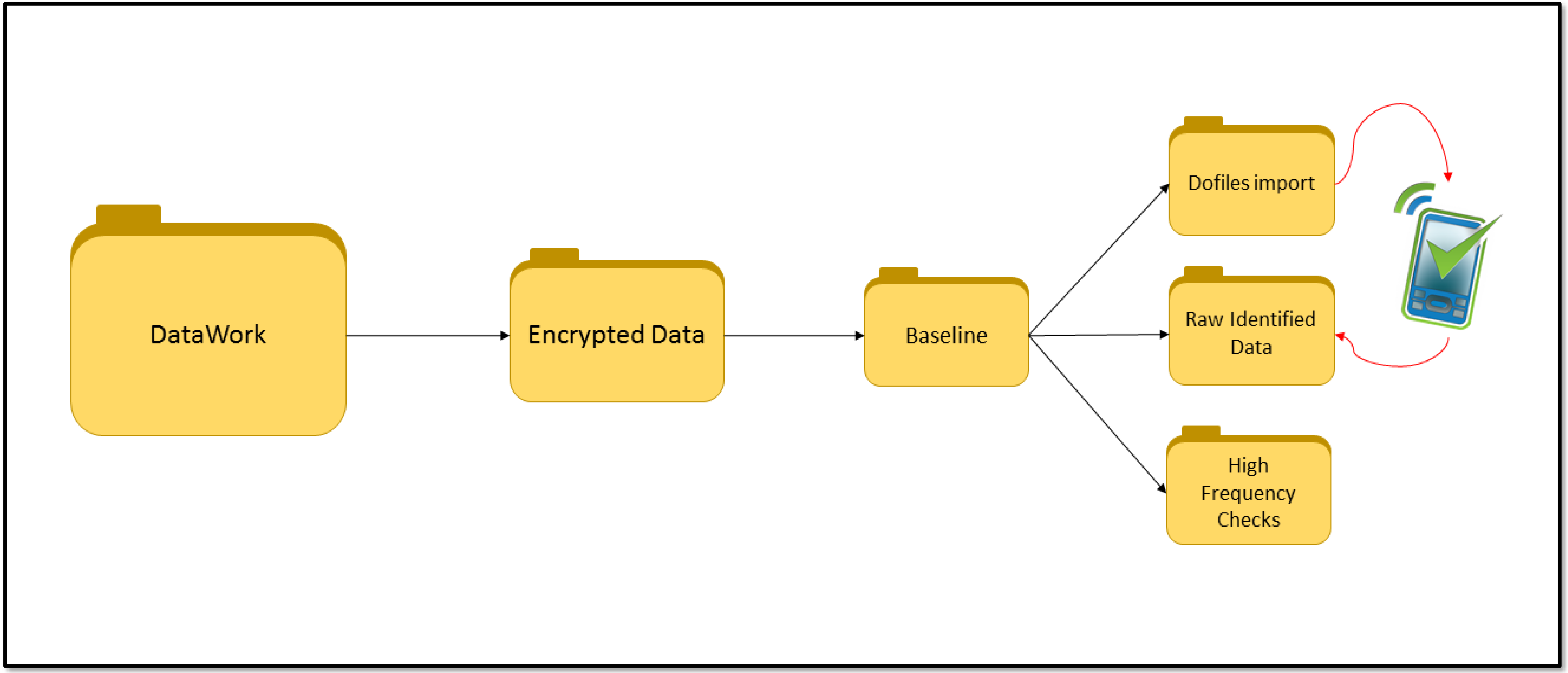
- Include all variables that are constant across the lifetime of a project in master data sets
  - ID variables, treatment status, sampling dummies, monitor outcomes, geo-variables
- One master data set per unit of observation
- Include all observations ever encountered – not just the observations you interview
- If you have discrepancies across data sets, the master data set is the master

Structure of the data folder: encrypted data

- The raw data with identifying information should be stored in the EncryptedData folder
- The do-files used to import your data from SurveyCTO or the equivalent software will also go in this folder



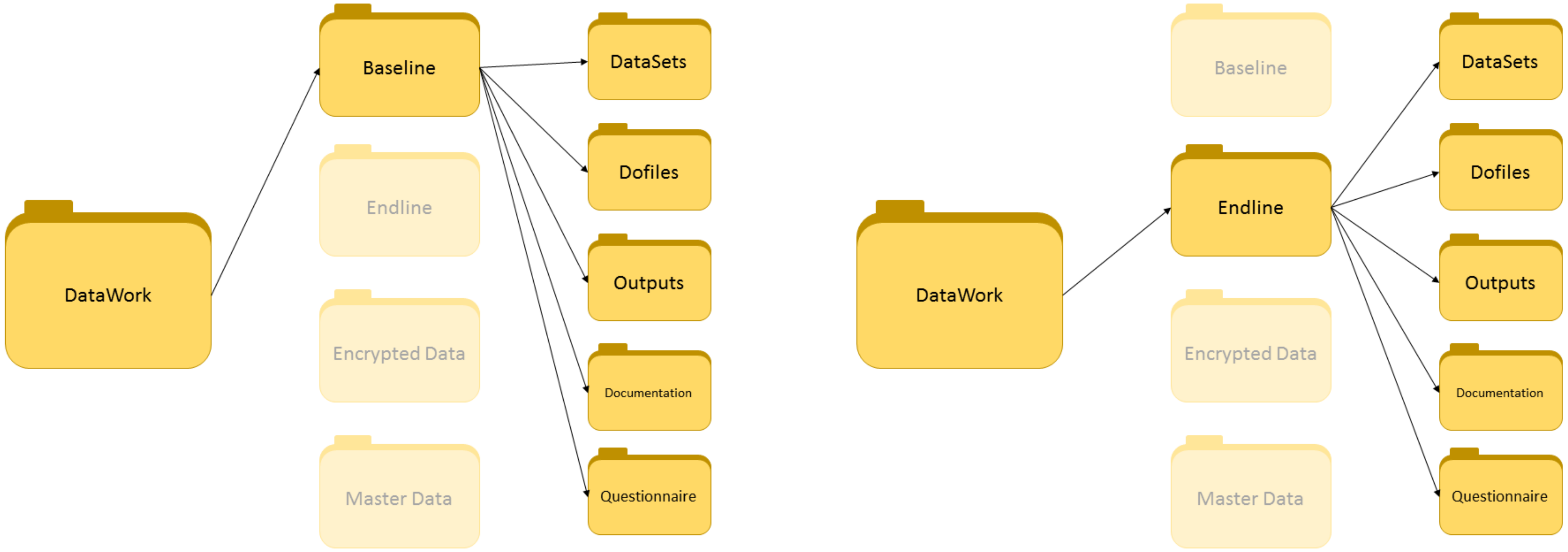
Structure of the data folder: encrypted data



Structure of the data folder: encrypted data

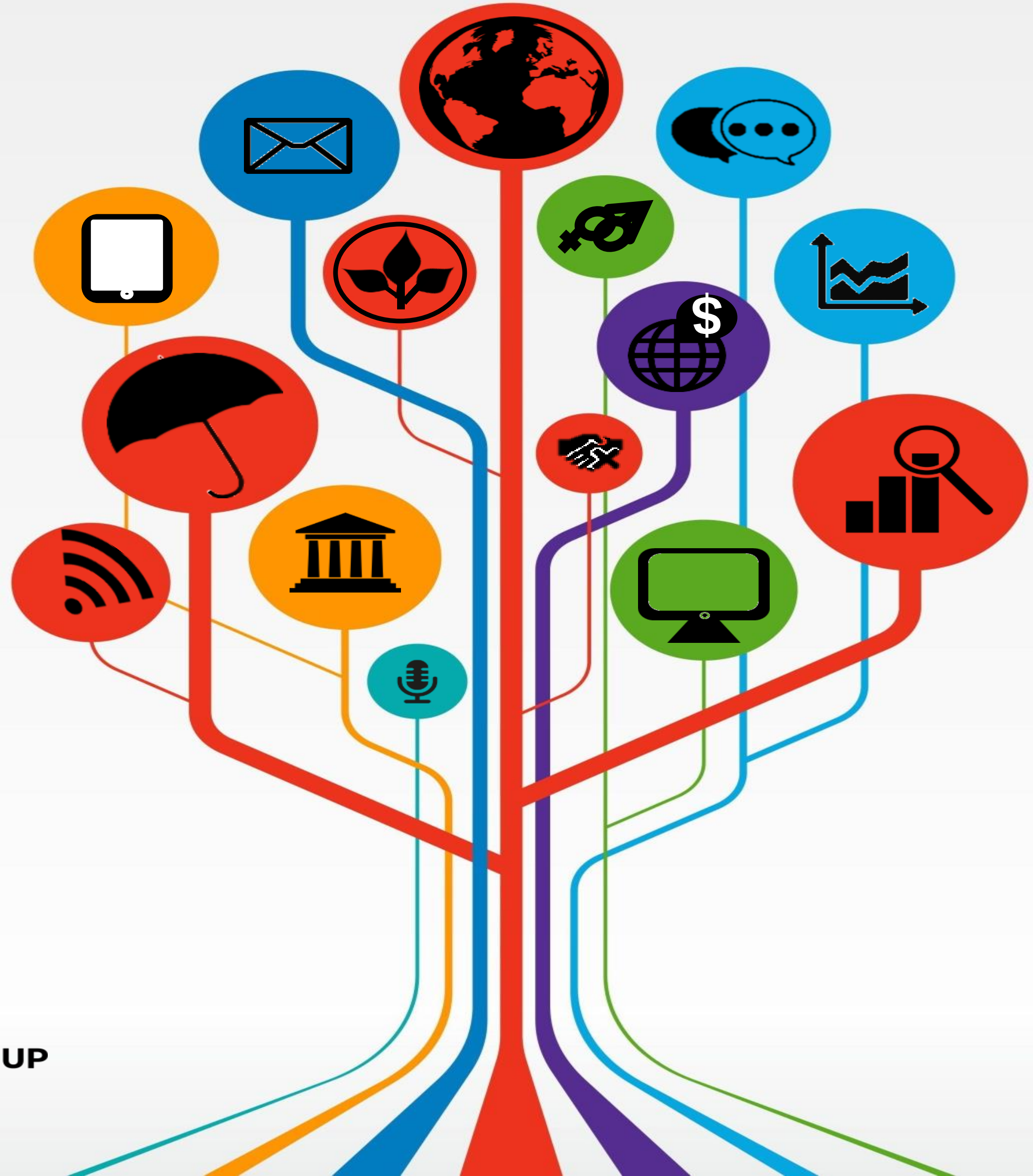
- Leave all files in this folder completely un-altered in the same format as you received them – guidance is forthcoming on a Boxcryptor license
- As soon as you make a change to the data, correct any values, or even import it to a different format, save it somewhere else
- Try to keep even file names unaltered. The exception is if you need to change the file name in order to be able to import them

Structure of the data folder: analysis rounds



No matter how many different types of analysis you do, the basic file structure is totally standardized.

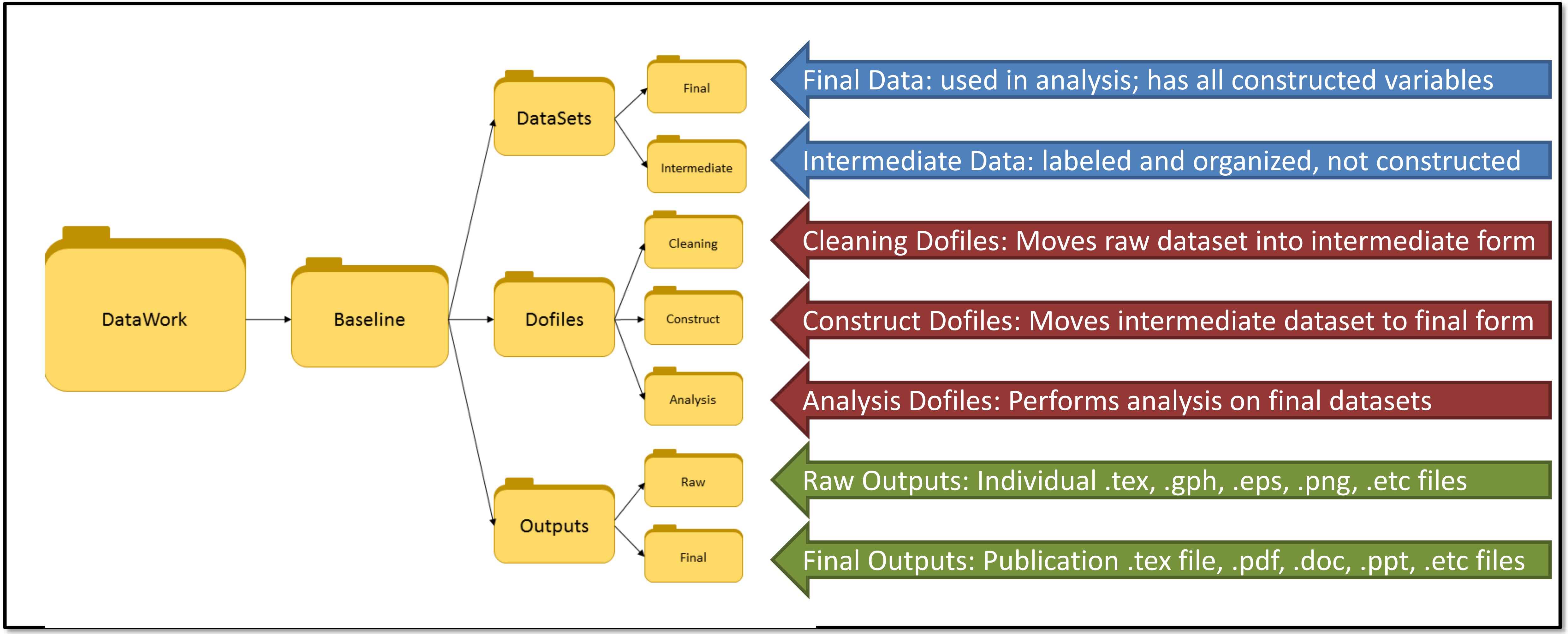
Structure of the analysis folders



i2i DIME TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Structure of the analysis folder: overview



Final Data: used in analysis; has all constructed variables

Intermediate Data: labeled and organized, not constructed

Cleaning Dofiles: Moves raw dataset into intermediate form

Construct Dofiles: Moves intermediate dataset to final form

Analysis Dofiles: Performs analysis on final datasets

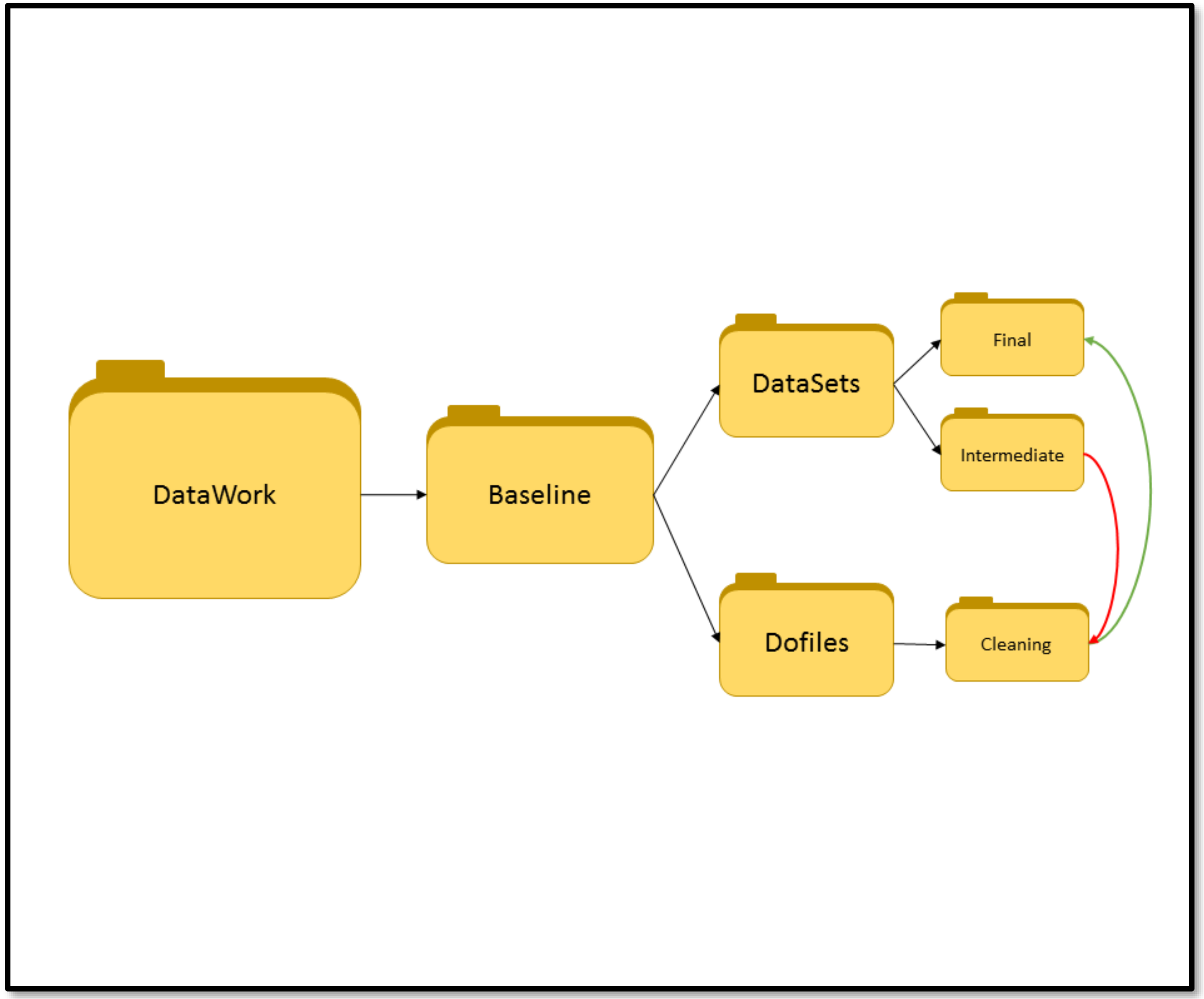
Raw Outputs: Individual .tex, .gph, .eps, .png, .etc files

Final Outputs: Publication .tex file, .pdf, .doc, .ppt, .etc files

Structure of the analysis folder: cleaning data

The cleaning do-files will:

1. Load the raw datasets (not stored in this folder)
2. De-identify and pre-process them (correct known errors, merge sampling information)
3. Save as intermediate datasets

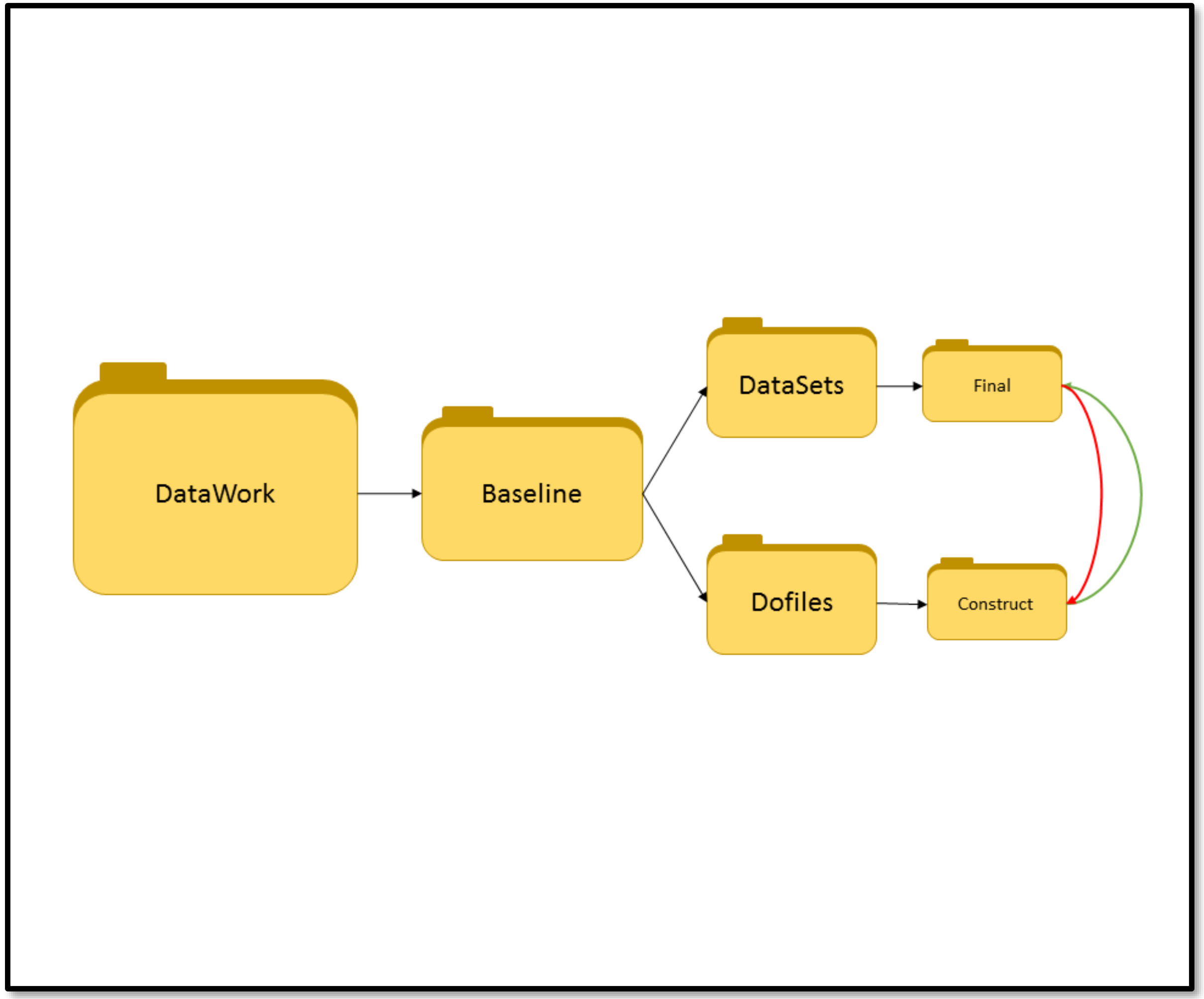


Structure of the analysis folder: cleaning data

The construct do-files will:

1. Load the intermediate data sets
2. Create "constructed" variables for analysis
3. Save the final data sets in the Final DataSets folder

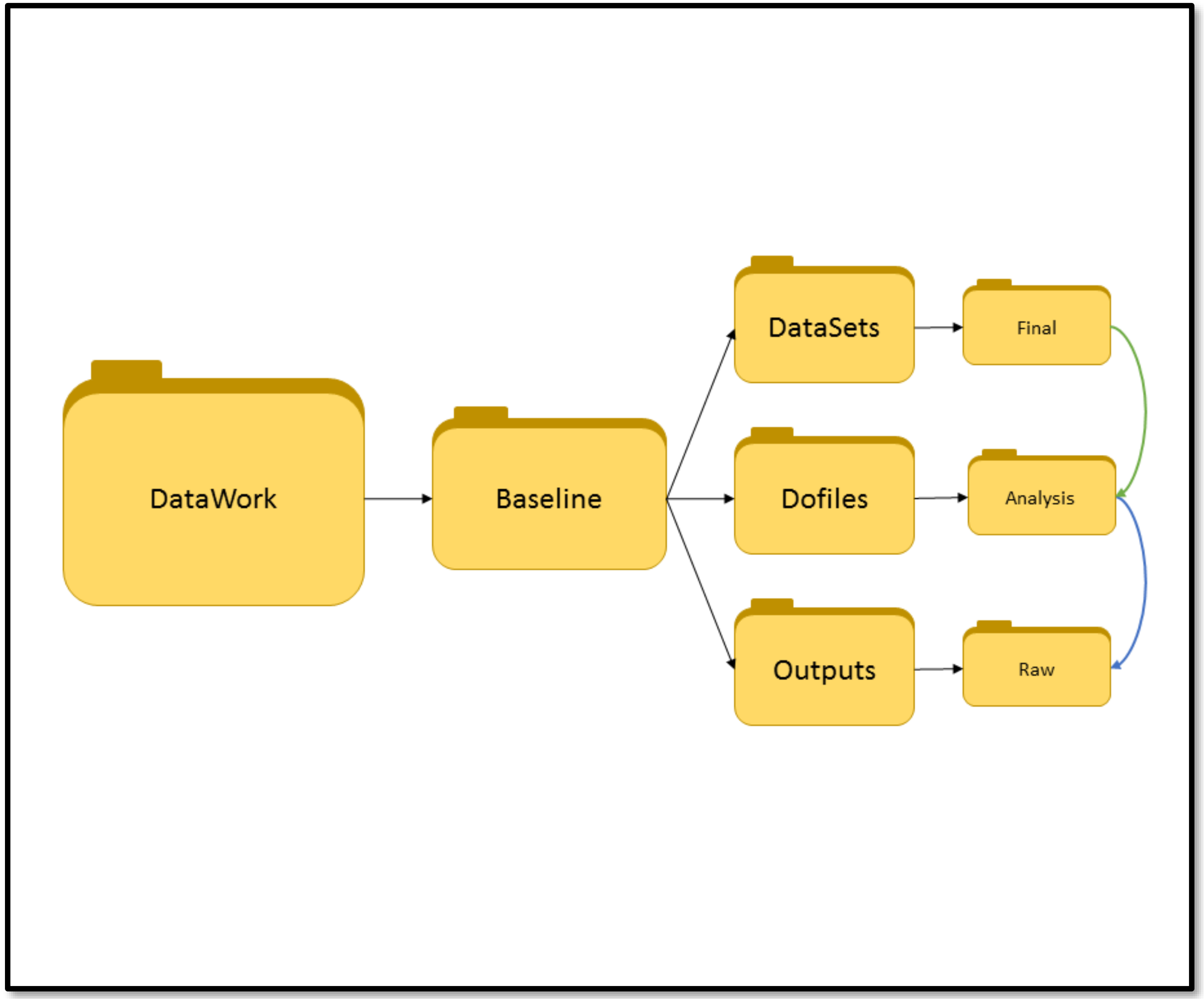
Think of the "final" data as what would be released on Dataverse when the paper is published



Structure of the analysis folder: analysis

The analysis do-files will:

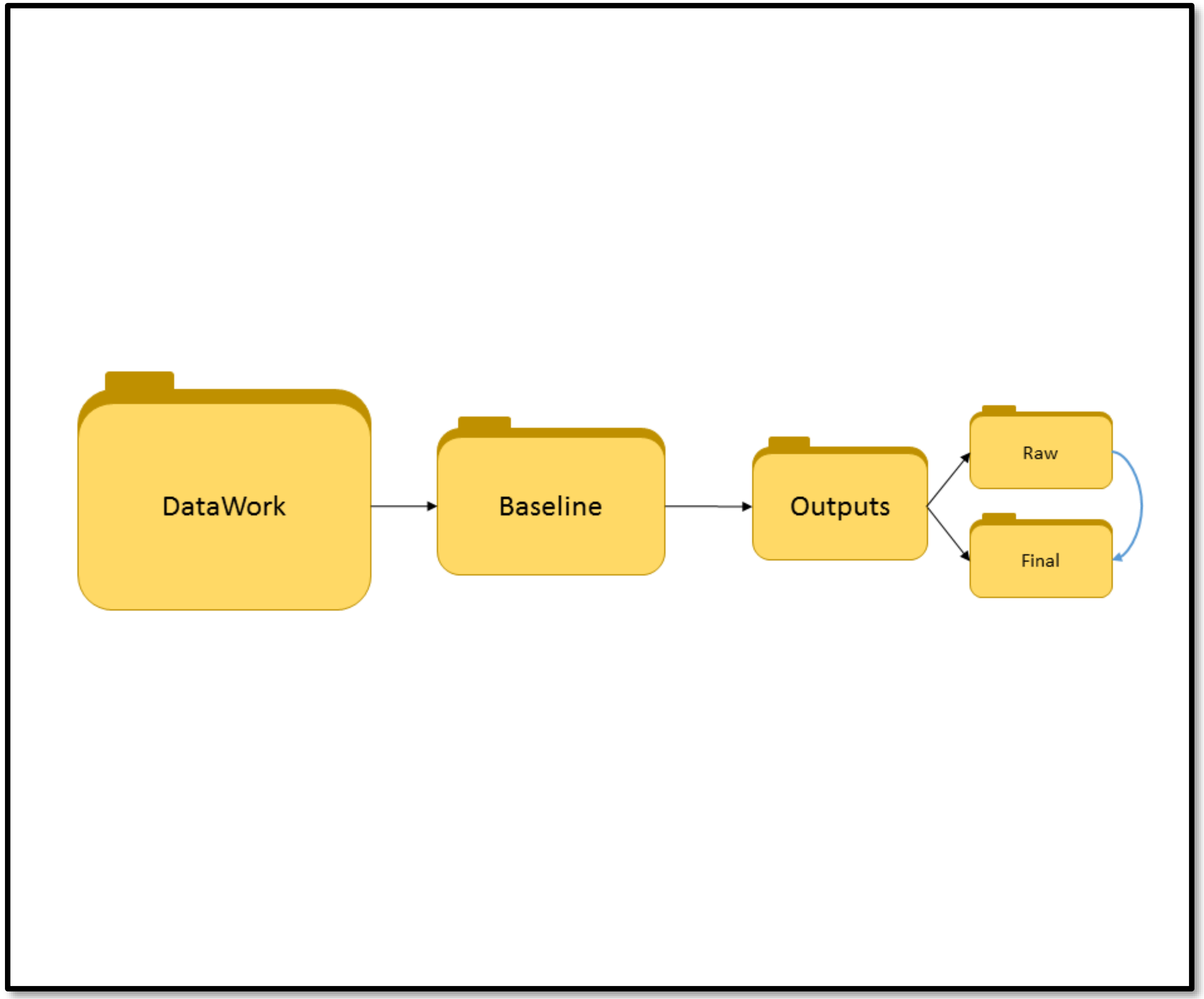
1. Load the constructed data and run the analysis
2. Outputs such as plots and tables generated by these do-files are stored in the Raw Outputs folder
3. Do not create variables in the analysis do-files!
4. Optionally the do-files themselves can be coded to push new outputs to GitHub/Overleaf and/or call pdflatex to compile new drafts



Structure of the analysis folder: outputs

The outputs folder will:

1. Hold “produced” materials that utilize the raw outputs (pdf documents, LaTeX documents, PowerPoints, Word, etc)



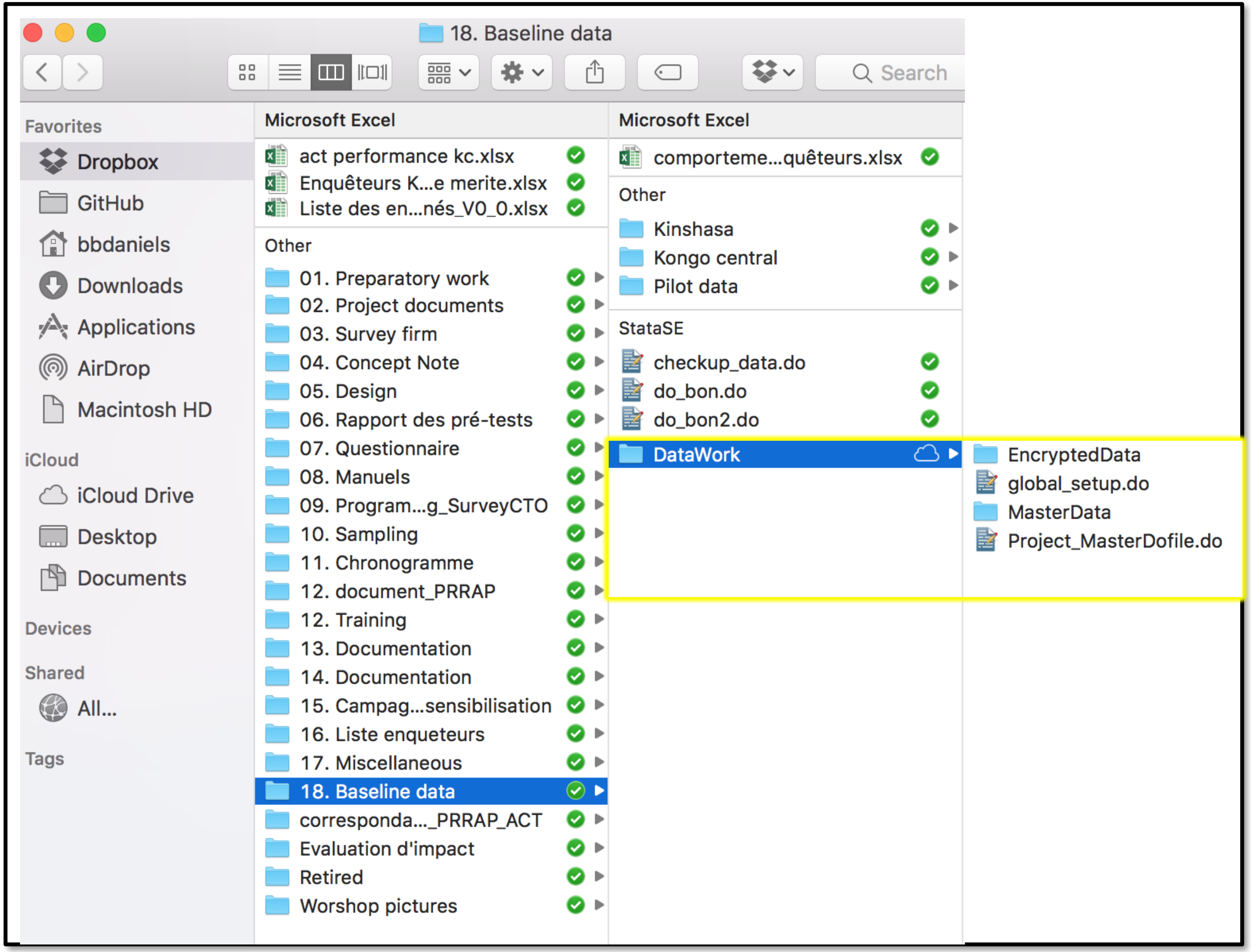
Using [*ietoolkit*] to set this all up

- Once data collection is done, your project looks something like this.

- Head on over to worldbank.github.io/ietoolkit
- Or do [*ssc install ietoolkit*]

- Then, [*iefolder*] will set up your folders.
- Do [*iefolder new project , project(your_project_folder)*]

- ***Voila!***



Microsoft Excel
act performance kc.xlsx
Enquêteurs K...e merite.xlsx
Liste des en...nés_V0_0.xlsx

Other
01. Preparatory work
02. Project documents
03. Survey firm
04. Concept Note
05. Design
06. Rapport des pré-tests
07. Questionnaire
08. Manuels
09. Program...g_SurveyCTO
10. Sampling
11. Chronogramme
12. document_PRRAP
12. Training
13. Documentation
14. Documentation
15. Campag...sensibilisation
16. Liste enqueteurs
17. Miscellaneous
18. Baseline data
corresponda..._PRRAP_ACT
Evaluation d'impact
Retired
Worshop pictures

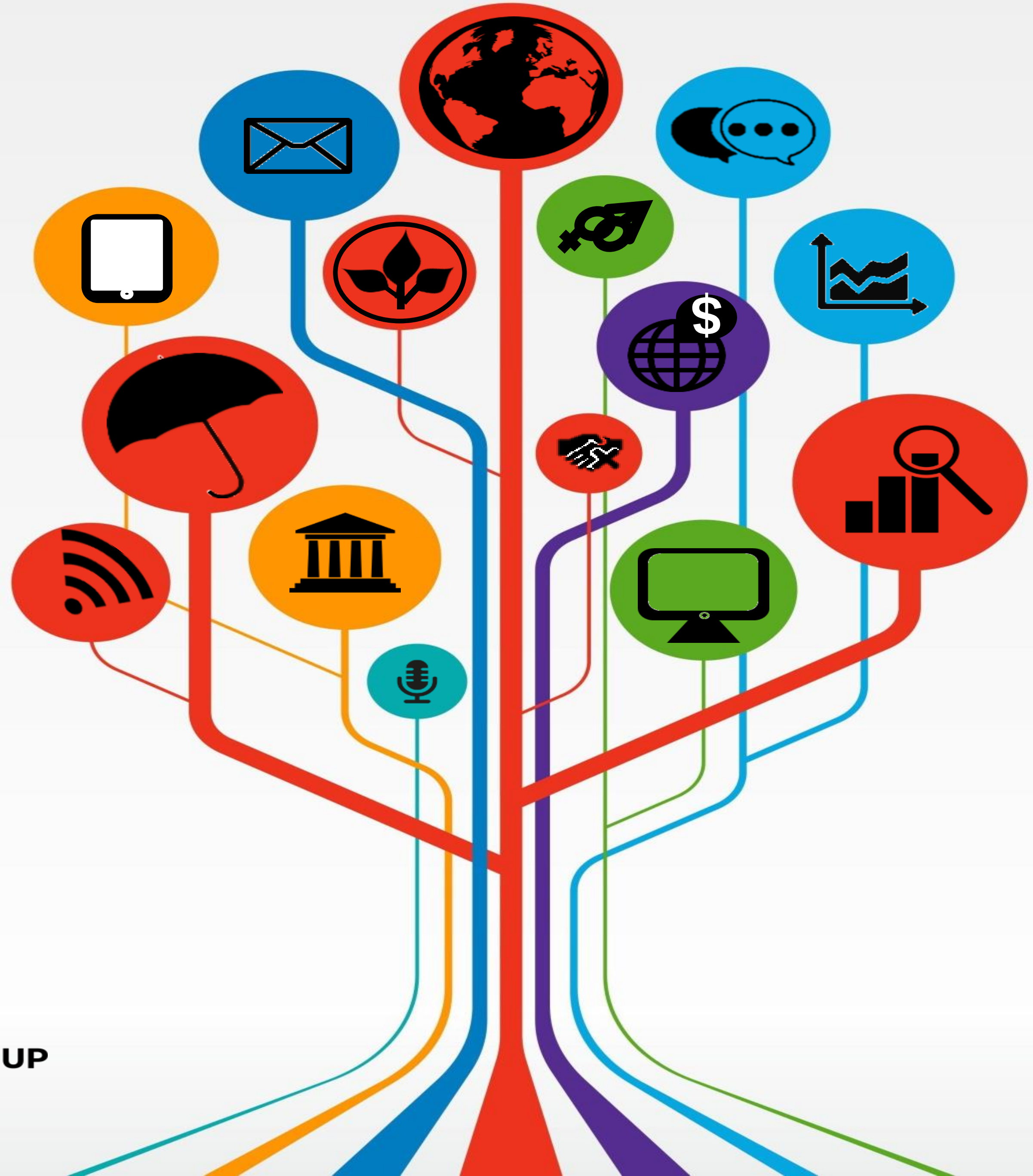
Microsoft Excel
comporteme...quêteurs.xlsx

Other
Kinshasa
Kongo central
Pilot data

StataSE
checkup_data.do
do_bon.do
do_bon2.do
DataWork

EncryptedData
global_setup.do
MasterData
Project_MasterDofile.do

Master do-files provide structure



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

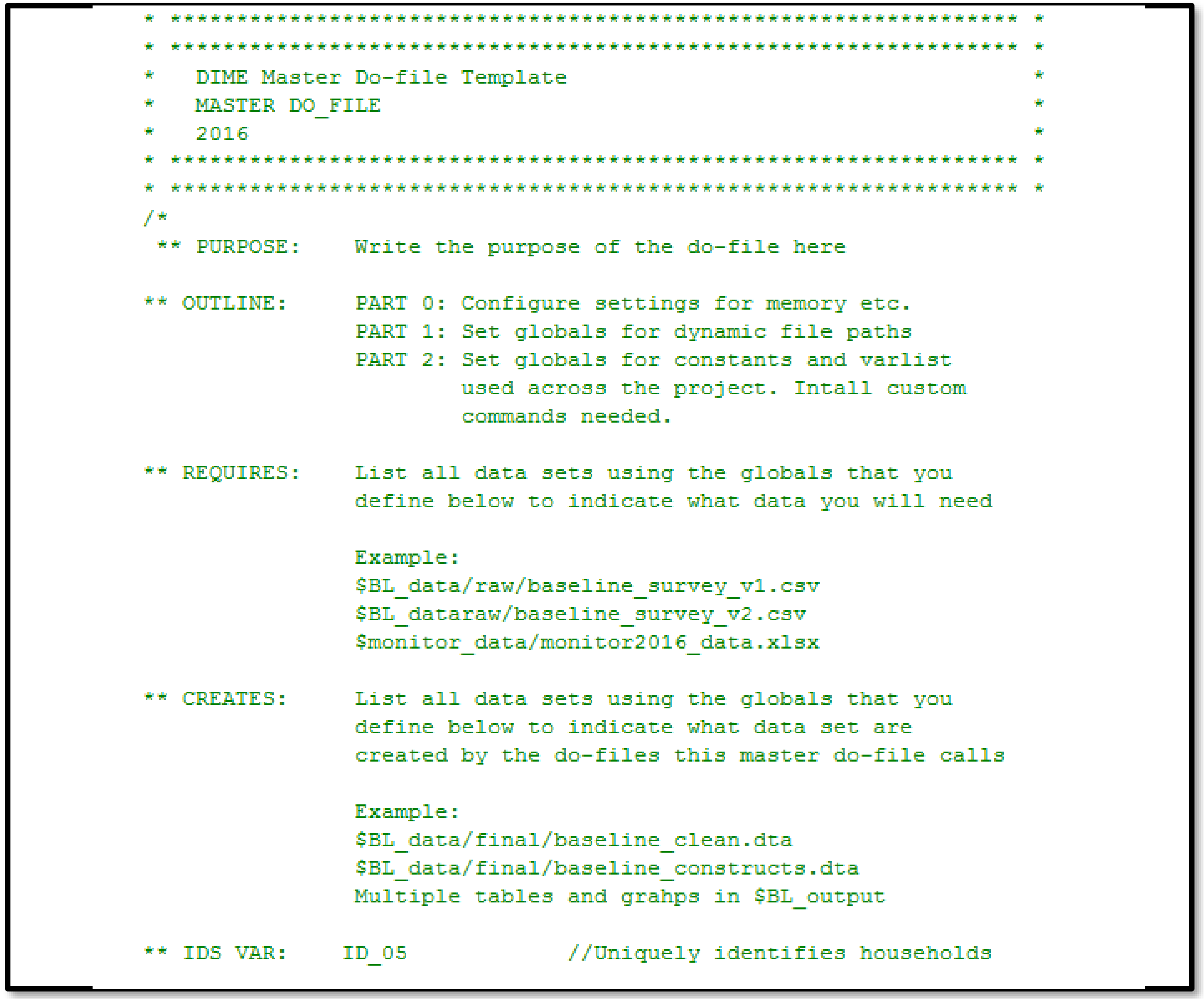
Master do-files: overview

- **As you might have noticed, there are a lot of do-files in the workflow**
  - A big project can become very complex, and do-files need to be run in a certain order to create the right outputs
  - That could mean you'd need to write one extremely long do-file, or a different document with instructions about in which order to run all the do-files
  - Best to keep each dofile self-contained: some projects have one for every exhibit

- **So, you can make a do-file run other do-files using the [*do*] command**
  - This (and a few secondary but still important things) is what a master do-file does
  - The master do-file is the map over all data work in your data folder
  - It's the table of contents for the instructions that you code
  - A total stranger who wants to replicate all your work should only have to open this one short dofile to have a complete roadmap of the analysis

Master do-files: Intro header

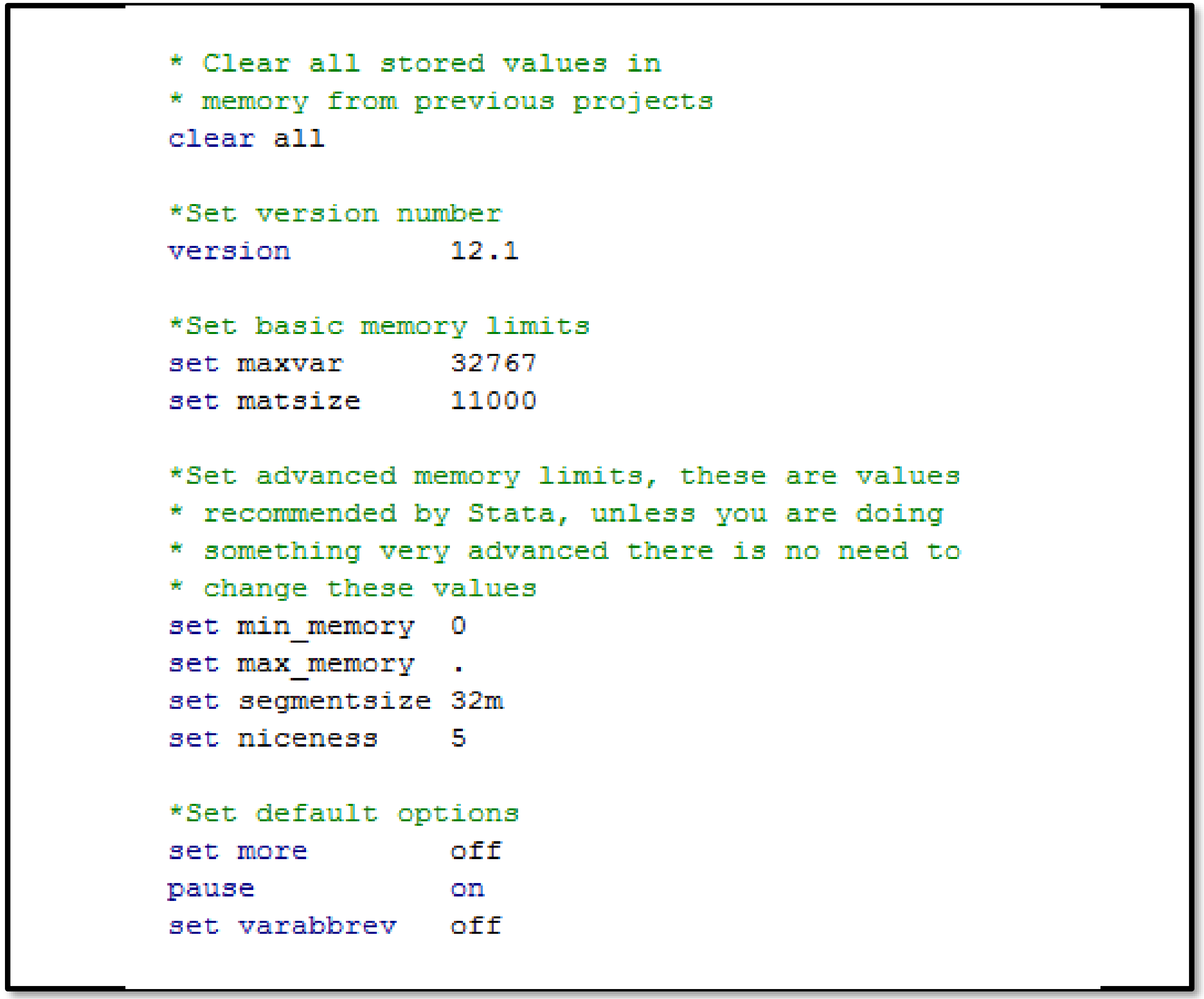
- Short descriptives in the file allow a reader to get the gist of what is accomplished, whether they are reading the right code, and anything else needed
- Highlights key information about the dataset, such as unique ID
- Details outputs created and explains how any outputs not included are created (by hand, with confidential data, etc)



```
* **************************************************************** *
* **************************************************************** *
*   DIME Master Do-file Template                                   *
*   MASTER DO_FILE                                                 *
*   2016                                                           *
* **************************************************************** *
* **************************************************************** *
/*
 ** PURPOSE:    Write the purpose of the do-file here

** OUTLINE:     PART 0: Configure settings for memory etc.
                PART 1: Set globals for dynamic file paths
                PART 2: Set globals for constants and varlist
                        used across the project. Intall custom
                        commands needed.

** REQUIRES:    List all data sets using the globals that you
                define below to indicate what data you will need

                Example:
                $BL_data/raw/baseline_survey_v1.csv
                $BL_dataraw/baseline_survey_v2.csv
                $monitor_data/monitor2016_data.xlsx

** CREATES:     List all data sets using the globals that you
                define below to indicate what data set are
                created by the do-files this master do-file calls

                Example:
                $BL_data/final/baseline_clean.dta
                $BL_data/final/baseline_constructs.dta
                Multiple tables and grahps in $BL_output

** IDS VAR:    ID_05            //Uniquely identifies households
```

Master do-files: Default settings

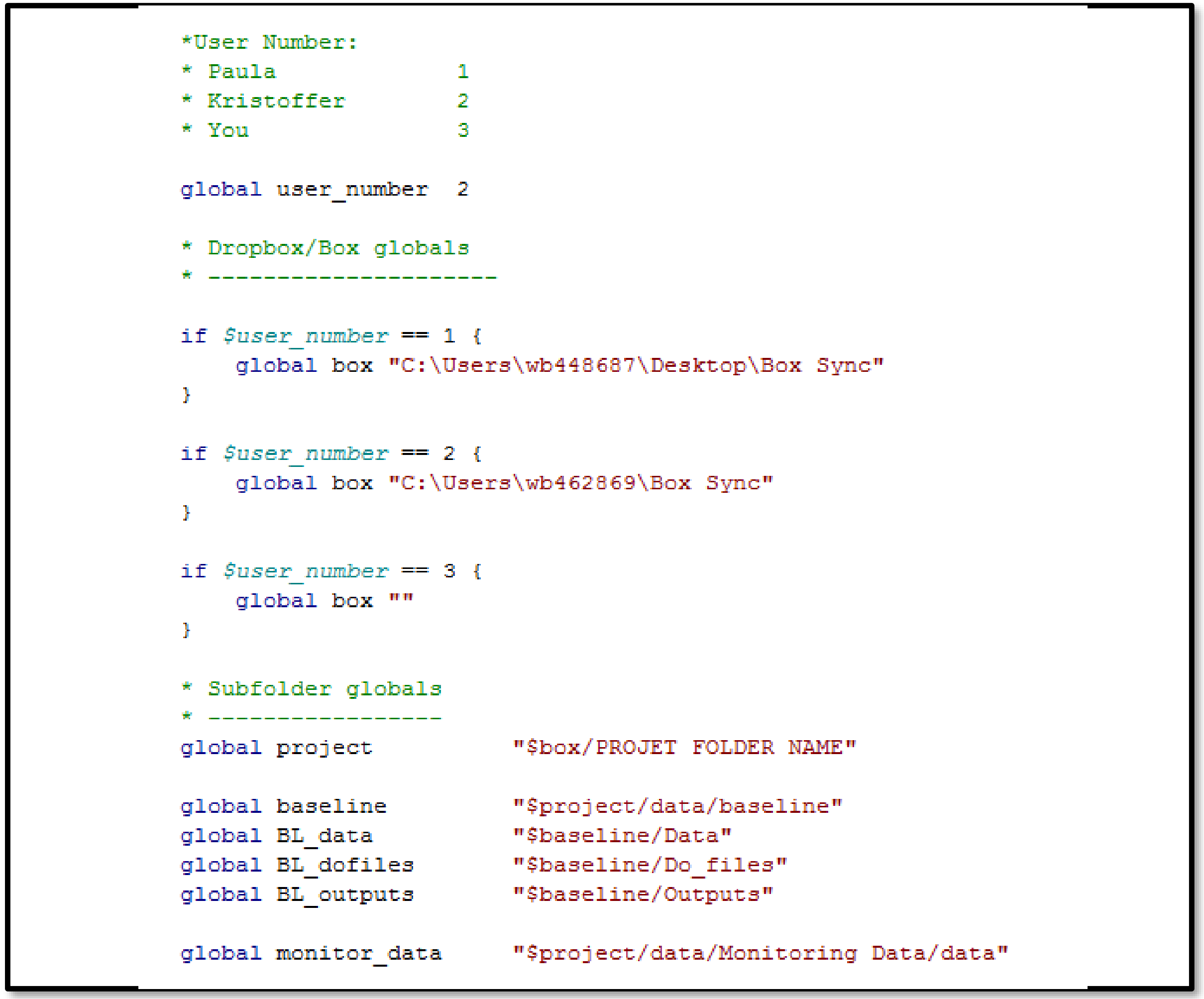
- These commands make everyone’s life easier by disabling some of Stata’s annoying defaults
- ”version” command is essential:
  - Changes to version can alter the results of randomizations
  - Some commands will break on newer/older versions due to changes in syntax



```
* Clear all stored values in
* memory from previous projects
clear all

*Set version number
version         12.1

*Set basic memory limits
set maxvar      32767
set matsize     11000

*Set advanced memory limits, these are values
* recommended by Stata, unless you are doing
* something very advanced there is no need to
* change these values
set min_memory  0
set max_memory  .
set segmentsize 32m
set niceness    5

*Set default options
set more        off
pause           on
set varabbrev   off
```

Master do-files: Folder paths

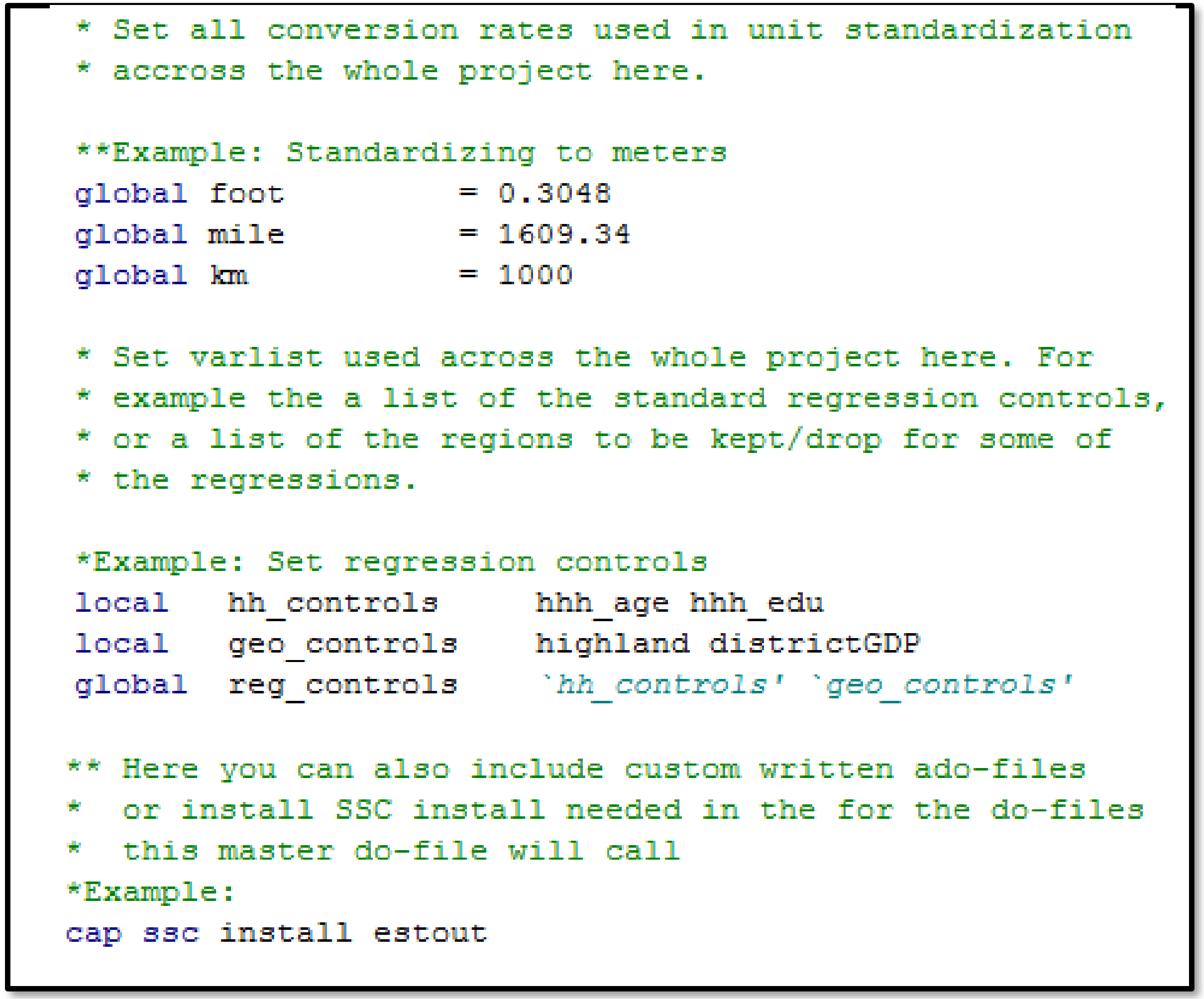
- This crucial component allows a new user to point the entire analysis at the correct (cloned or synced) iteration of the folder
- If this is not available, new users will spend a long time changing filepaths scattered throughout the dofiles.
- ALWAYS USE “/”
- NEVER USE “\”



```
*User Number:
* Paula             1
* Kristoffer        2
* You               3

global user_number  2

* Dropbox/Box globals
* ---------------------

if $user_number == 1 {
    global box "C:\Users\wb448687\Desktop\Box Sync"
}

if $user_number == 2 {
    global box "C:\Users\wb462869\Box Sync"
}

if $user_number == 3 {
    global box ""
}

* Subfolder globals
* -----------------
global project          "$box/PROJET FOLDER NAME"

global baseline         "$project/data/baseline"
global BL_data          "$baseline/Data"
global BL_dofiles       "$baseline/Do_files"
global BL_outputs       "$baseline/Outputs"

global monitor_data     "$project/data/Monitoring Data/data"
```

Master do-files: Globals and programs

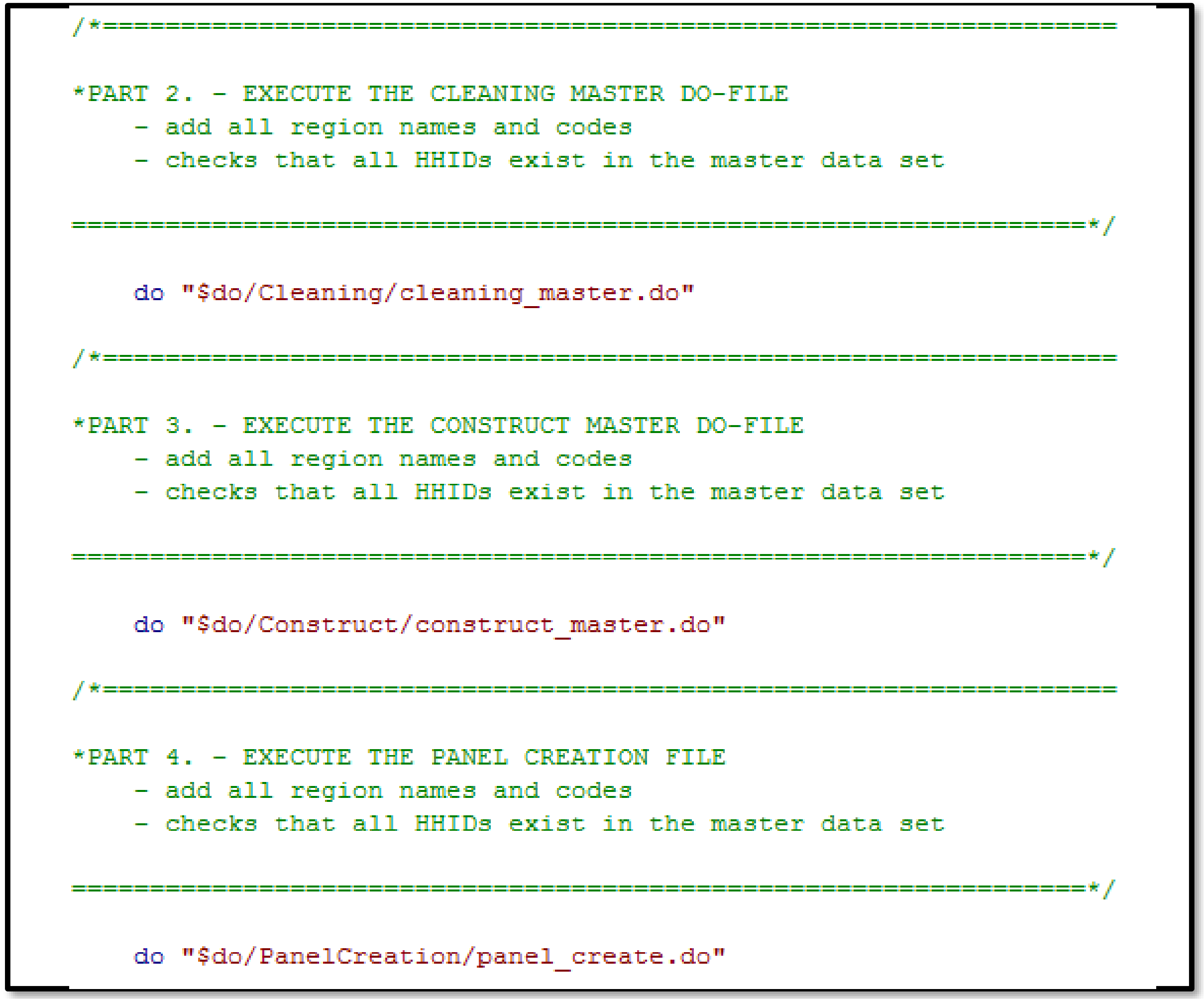
- Add conversions that are needed for the analysis
- Add controls lists that are used throughout the paper, so that all regressions can be altered correctly with one line
- Add options (such as graphing options, clustering, etc)
- Any external programs that need to be installed for code to run



```
* Set all conversion rates used in unit standardization
* accross the whole project here.

**Example: Standardizing to meters
global foot         = 0.3048
global mile         = 1609.34
global km           = 1000

* Set varlist used across the whole project here. For
* example the a list of the standard regression controls,
* or a list of the regions to be kept/drop for some of
* the regressions.

*Example: Set regression controls
local   hh_controls     hhh_age hhh_edu
local   geo_controls    highland districtGDP
global  reg_controls    `hh_controls' `geo_controls'

** Here you can also include custom written ado-files
*  or install SSC install needed in the for the do-files
*  this master do-file will call
*Example:
cap ssc install estout
```

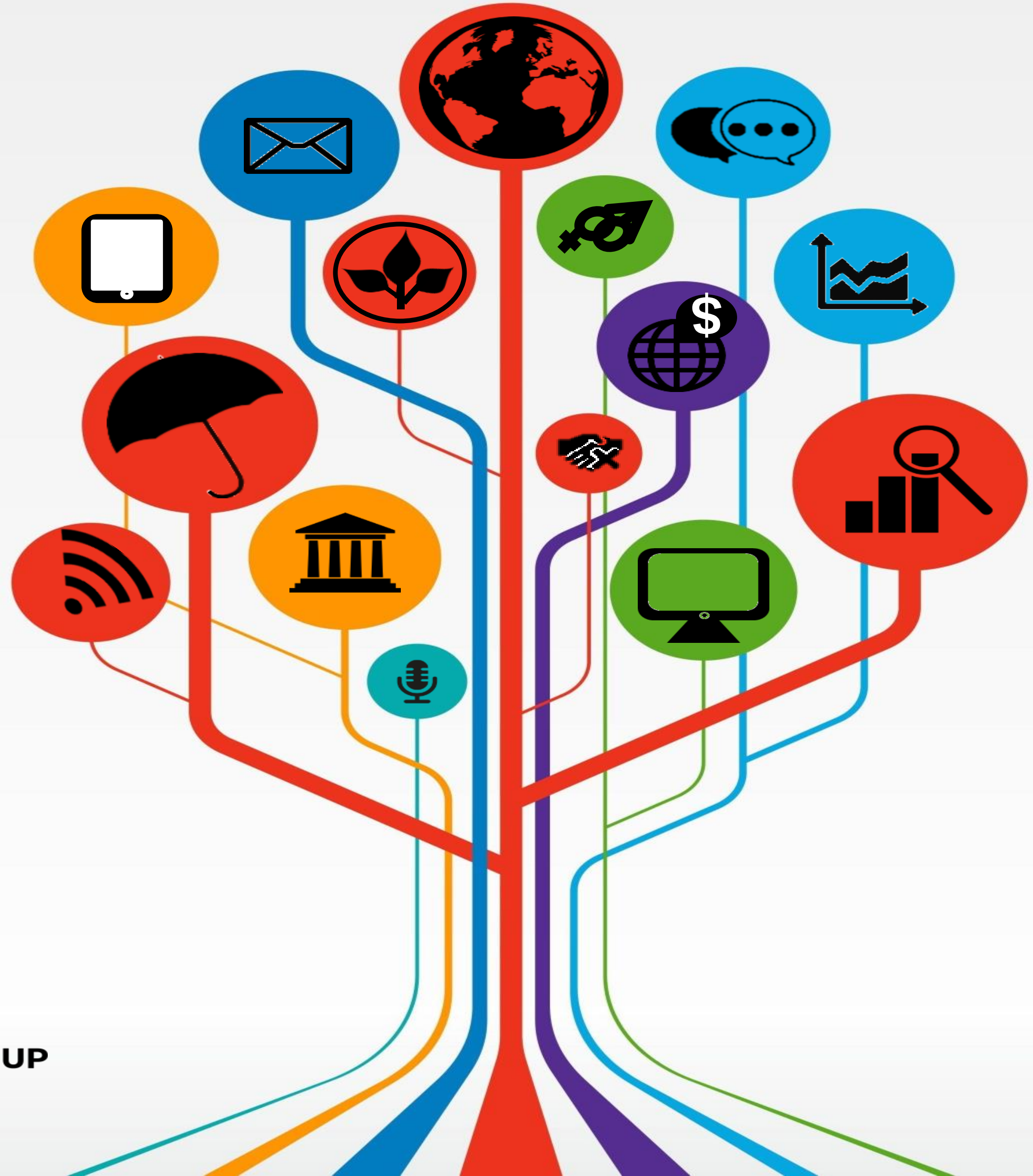
Master do-files: Actually doing stuff

- Detail what each file called in the master do-file is responsible for
- If there are a lot of do-files, create a sub-master for each high level task
  - Example of high level tasks: import, cleaning, construct, analysis etc.
- Use comments that explain where in your code you do what



```
/*====================================================================

*PART 2. - EXECUTE THE CLEANING MASTER DO-FILE
    - add all region names and codes
    - checks that all HHIDs exist in the master data set

====================================================================*/

    do "$do/Cleaning/cleaning_master.do"

/*====================================================================

*PART 3. - EXECUTE THE CONSTRUCT MASTER DO-FILE
    - add all region names and codes
    - checks that all HHIDs exist in the master data set

====================================================================*/

    do "$do/Construct/construct_master.do"

/*====================================================================

*PART 4. - EXECUTE THE PANEL CREATION FILE
    - add all region names and codes
    - checks that all HHIDs exist in the master data set

====================================================================*/

    do "$do/PanelCreation/panel_create.do"
```

Thank you!



i2i
DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP