Data Management

Luiza Andrade & Kristoffer Bjarfekur

DIME - World Bank

October 12, 2017

Overview

Overview

Introduction

- This presentation will show you some best practices developed to manage data work
- At DIME, we frequently have large teams collaborating on the same codes and data sets
- Also, long projects might easily become complex, with multiple rounds of data collection creating a structure that can be hard to navigate
- Having a standardized way to organize documents and code prevents mistakes and reduces the cost of transitioning accross projects and teams

Think reproducible

The end goal of our best practice guidelines is to ensure that all the research we develop is reproducible

- This means transparency, accountability...
- ... and a easier workflow

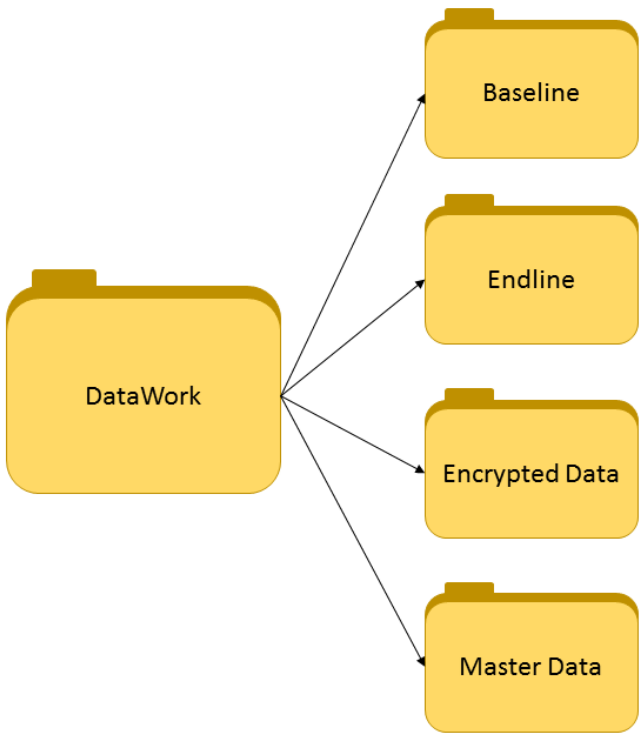
What is this about

- This presentation will take you through two seemingly unconnected subjects:
  1. The structure of the folder where a project's data work is stored
  2. The master do-file for a project's data work
- However, as you should understand by the end of this presentation, these two things are instrisically connected. So stick around even if you think you got lost in a sudden change of subject
- These first two parts of this presentation will introduce you to some data management practices that might seem very advanced and abstract
- In the last section, however, we will show you easy-to-use tools that will set you up to start using them very quickly

Overview

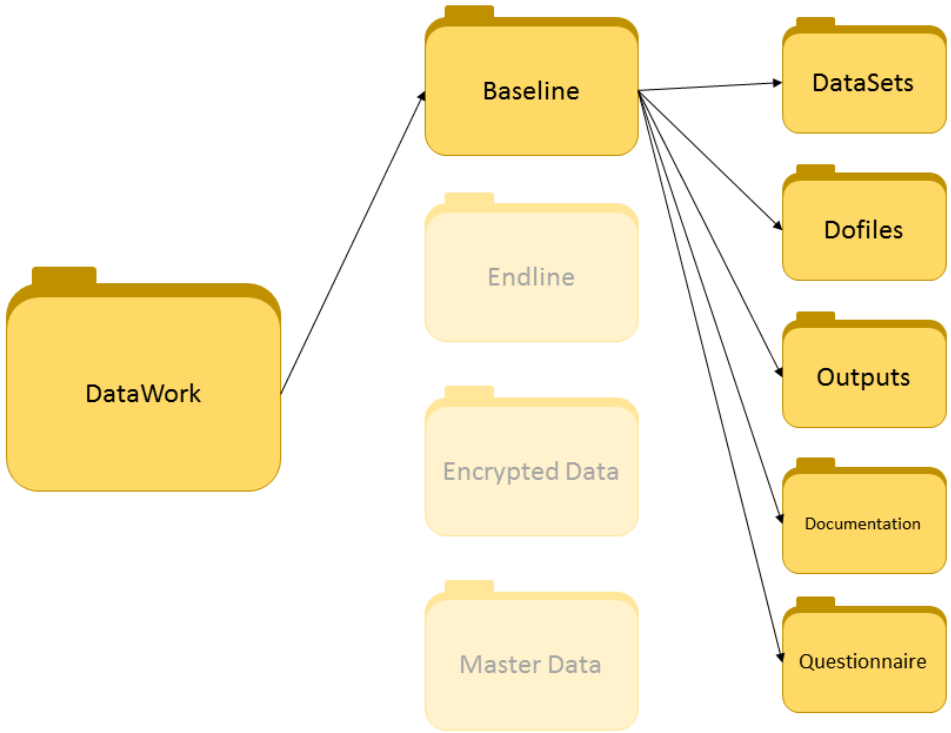
DataWork Folder: Overview

Your project folder probably has a lot of subfolders, for literature, presentations, concept notes and other project documents. All the data work will be stored in a DataWork subfolder.



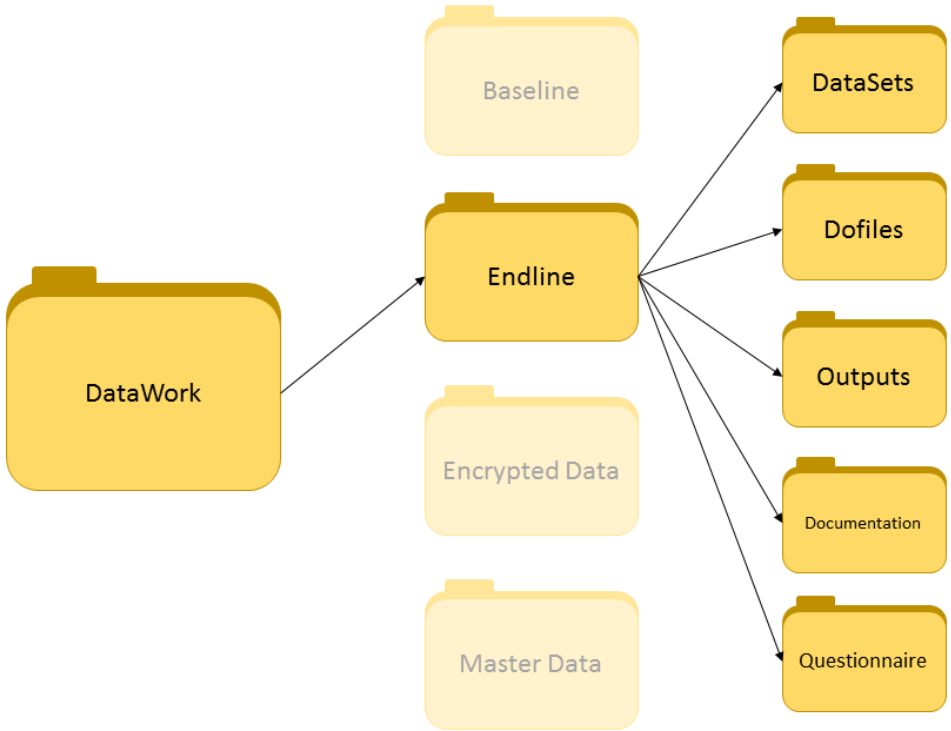
DataWork Folder: Rounds folders

The baseline folder will store all baseline data, as well as do-files and outputs that refer exclusively to this round of data collection.



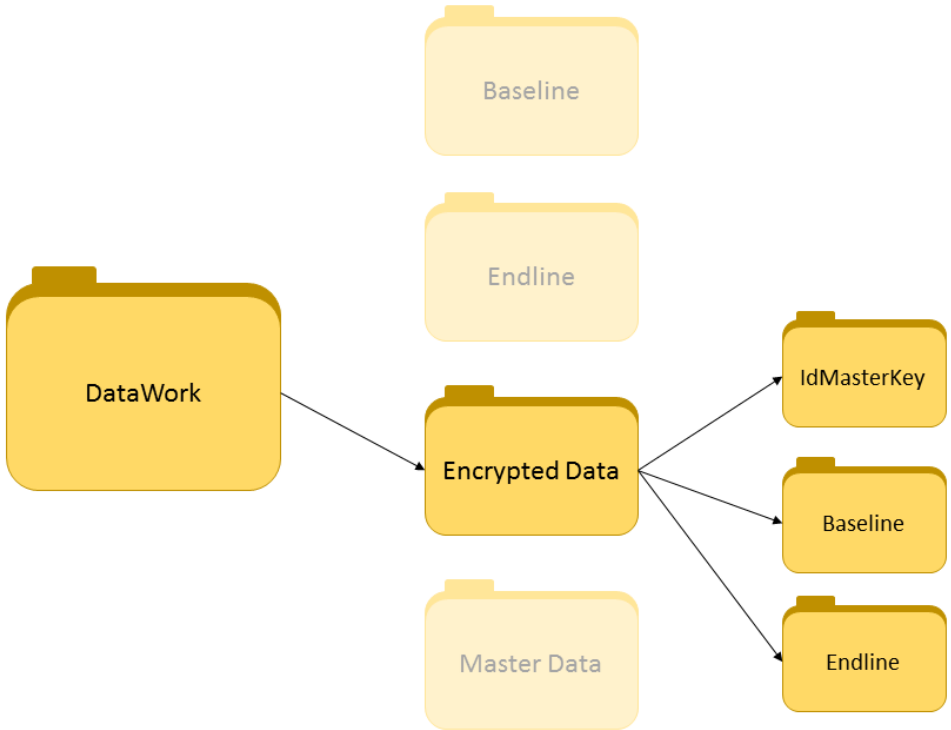
DataWork Folder: Rounds folders

The endline folder will store all endline data, as well as do-files and outputs that refer exclusively to this round of data collection. Note that its sctructure is exactly the same ass the baseline folder's.



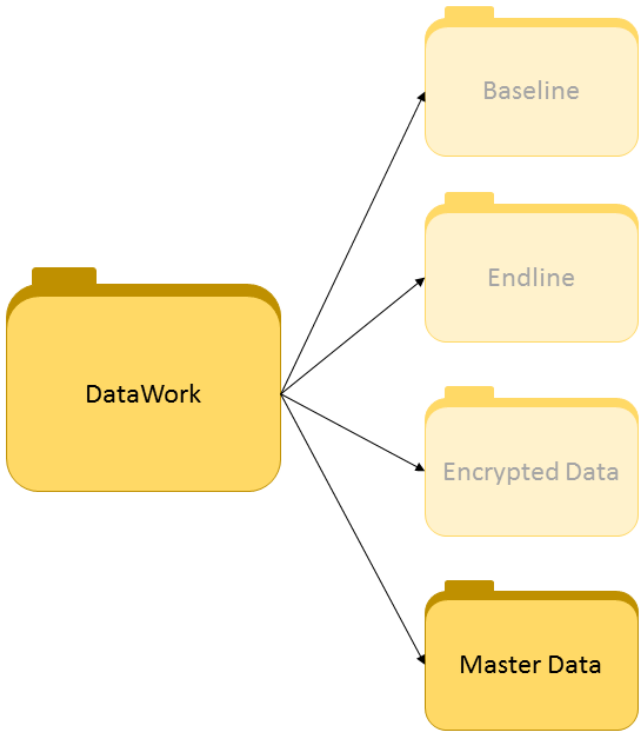
DataWork Folder: Encrypted Data

The encrypted data folder will contain all personally identified data for each round of data collection, as well as a folder with ID master keys linking each unidentified ID to the identified observations. As its name suggests, this folder should be encrypted.



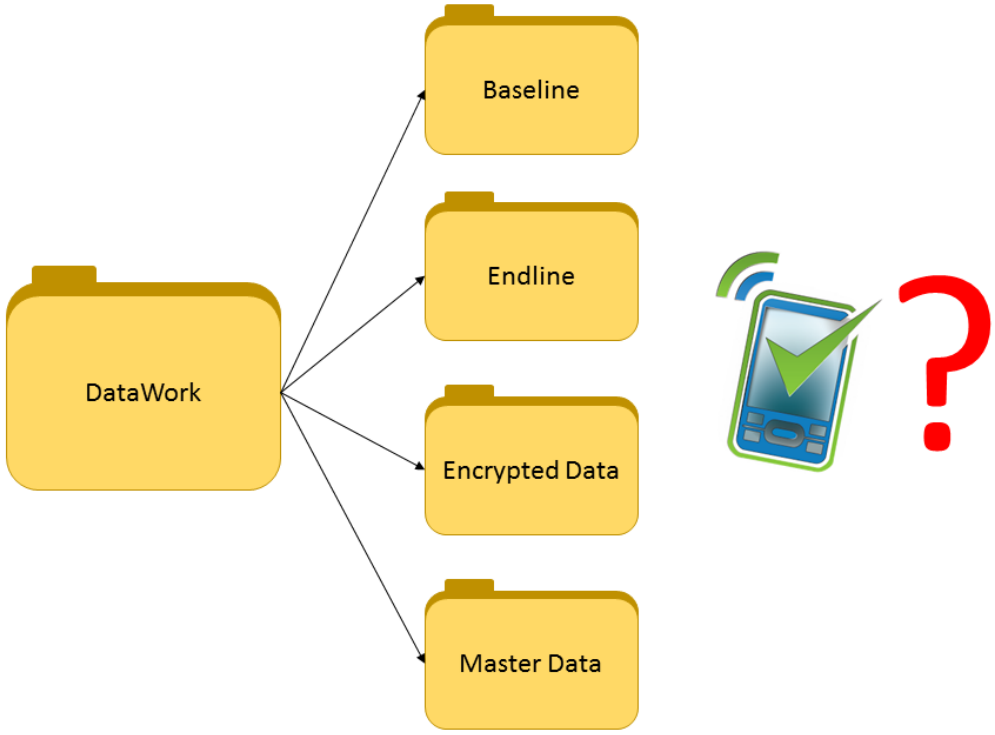
DataWork Folder: Master Data

The master data folder will store the master data sets for each unit of observation in your project. Don't worry if you don't yet know what a master data set is, this is part of another training.

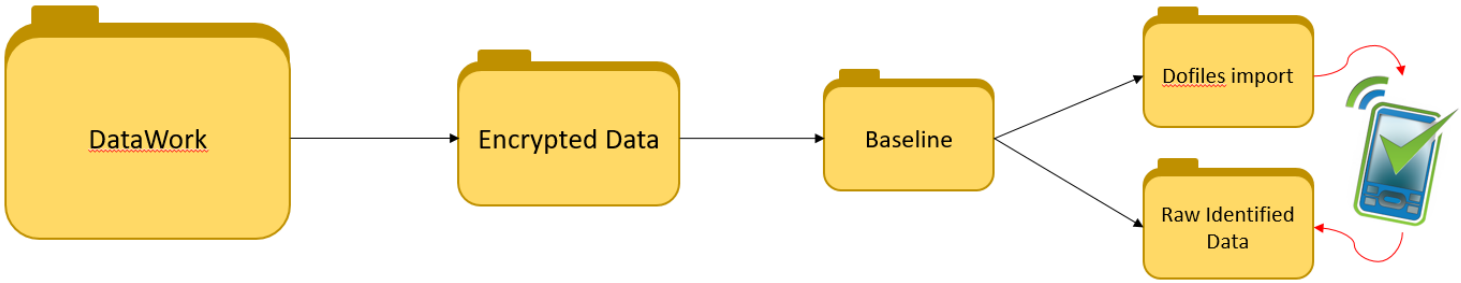


Using the DataWork Folder

So you received data from the field. What now?

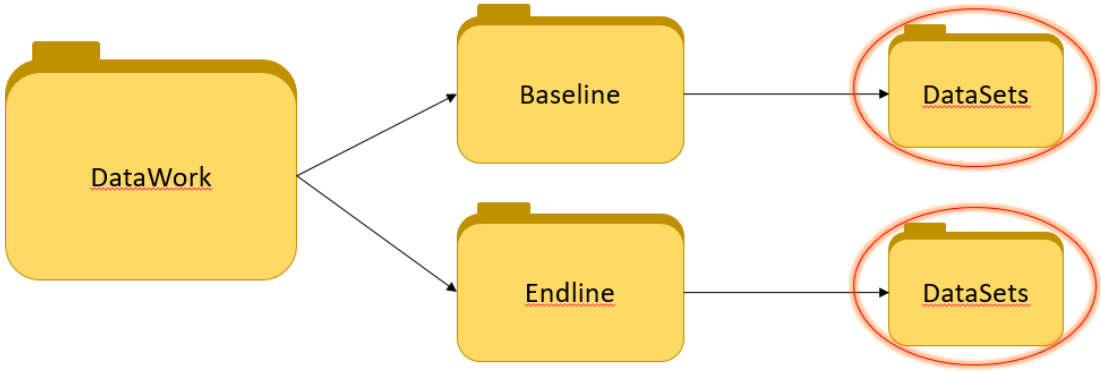


Using the DataWork Folder: EncryptedData



- The raw data with identifying information should be stored in the EncryptedData folder
- The do-files used to import your data from SurveyCTO will also go in this folder
- Don't forget to encrypt the EncryptedData folder!

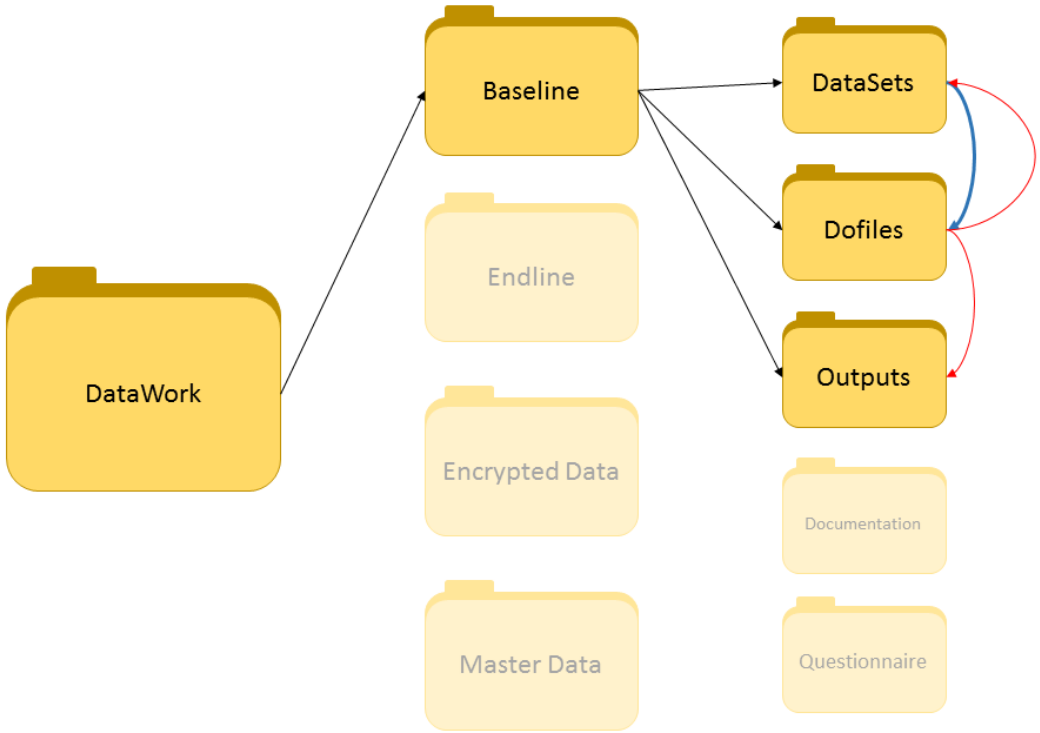
Using the DataWork Folder: DataSets



- All data in the Intermediate or the Final folder should be de-identified

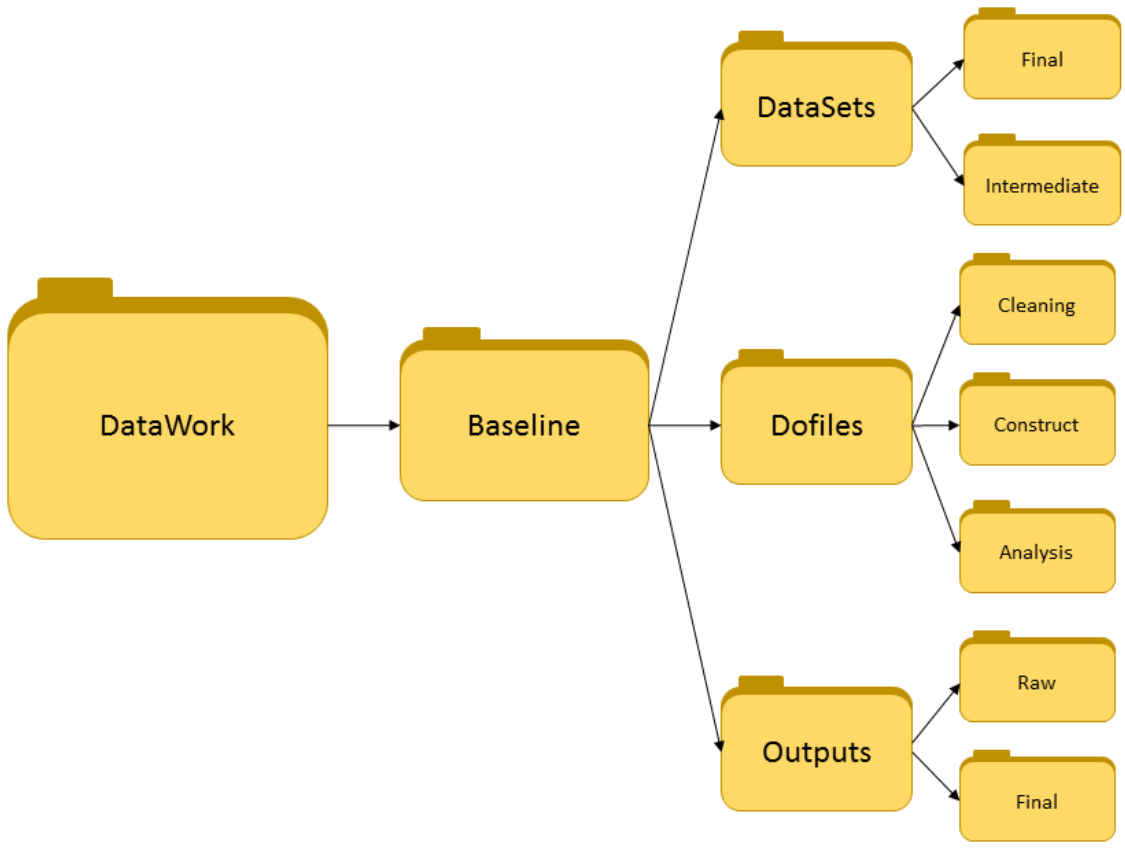
Using the DataWork Folder: Round folders

The do-files in each of you rounds' folders will load data from that round's DataSets folder and store any outputs in the round's Outputs folder



Using the DataWork Folder: Rounds folders

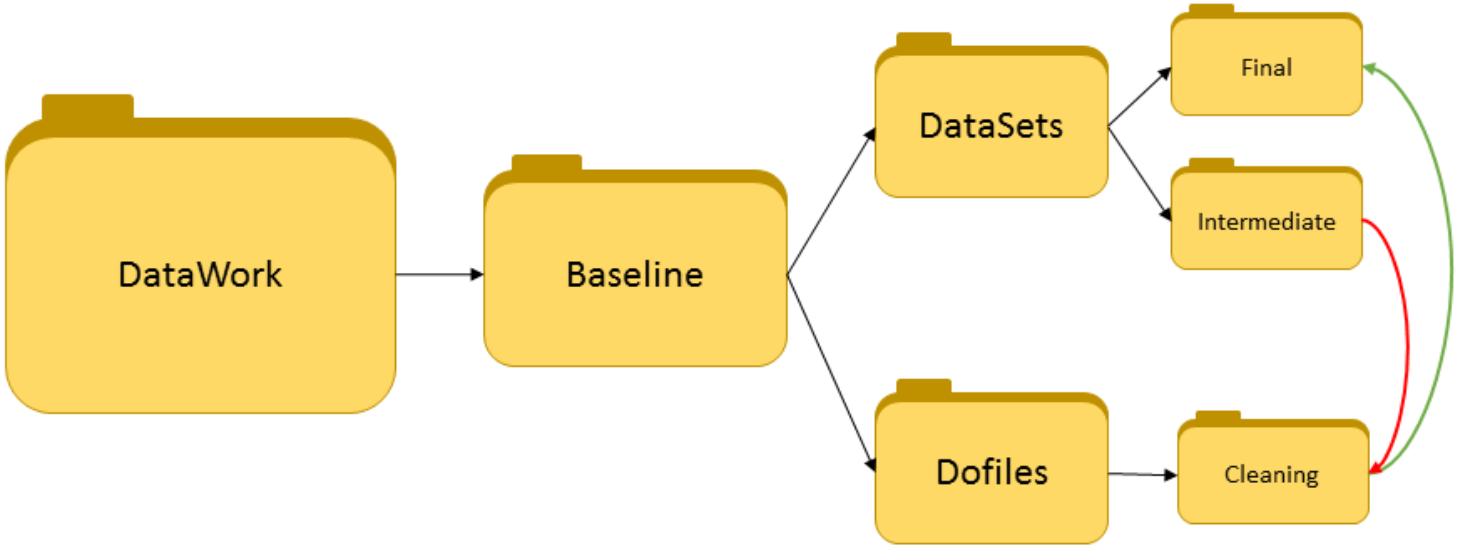
This is what an specific round folder looks like in detail



Using the DataWork Folder: Cleaning

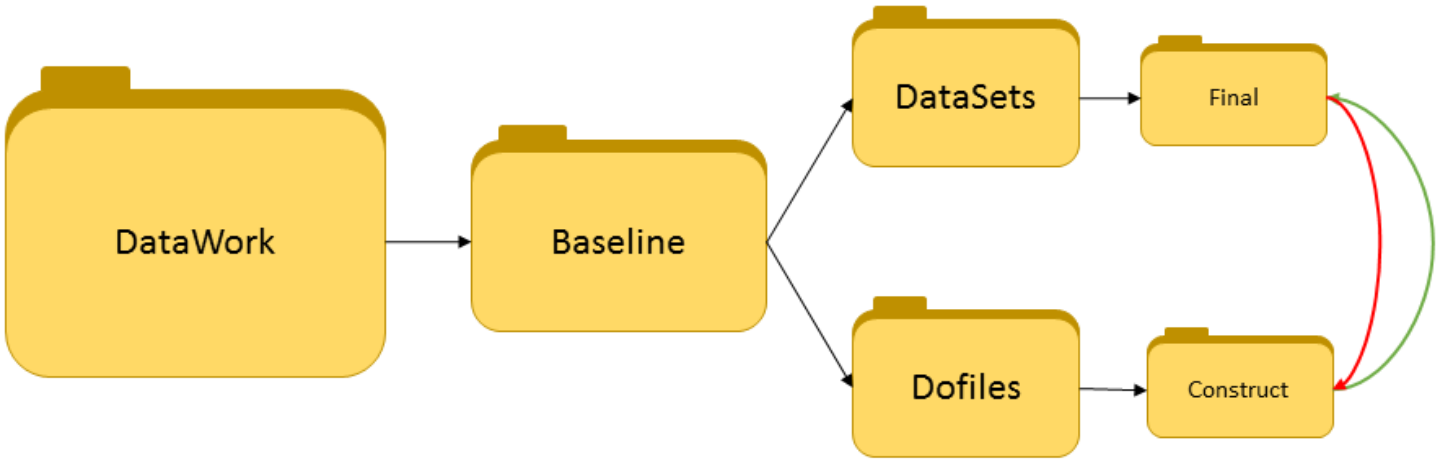
The do-files in the Cleaning Dofiles folder will

1. Load the intermediate data sets (e.g., de-identified survey data)
2. Clean them
3. Save the clean data sets in the Final DataSets folder



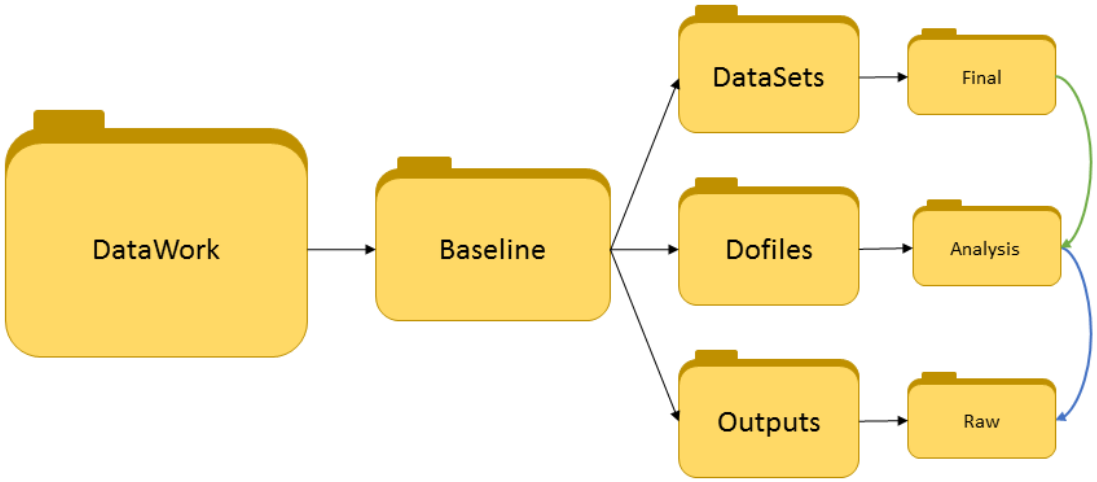
Using the DataWork Folder: Construct

- The do-files in the Construct Dofiles folder load the clean data and use it to construct the indicators that will be used for analysis
- The constructed data set, contaning this indicators, will be stored in the Final DataSets folder



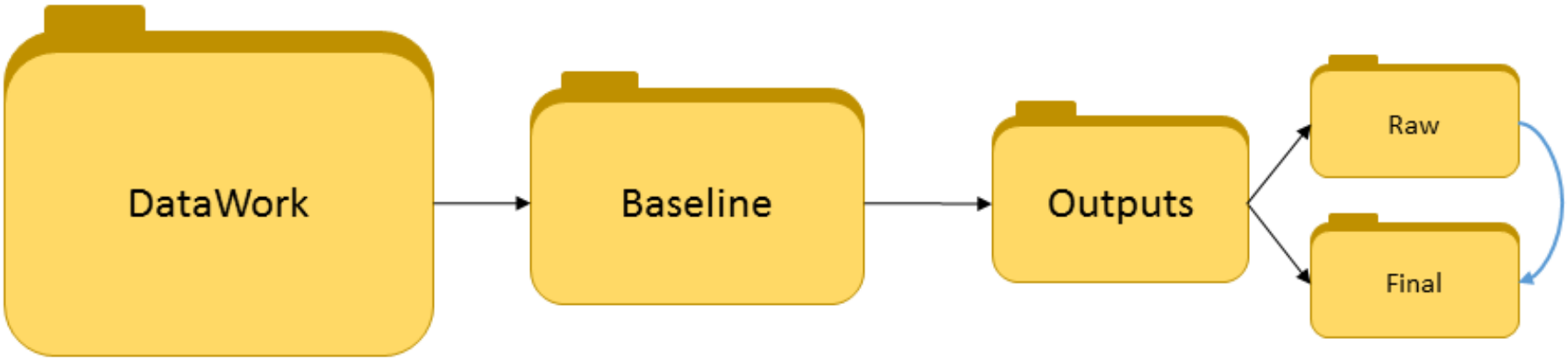
Using the DataWork Folder: Analysis

- The do-files in the Analysis Dofiles folder load the contructed data and run the analysis
- Outputs such as plots and tables generated by these do-files are stored in the Raw Outputs folder



Using the DataWork Folder: Outputs

- The Final Outputs folder stores reports, papers and other documents you create using the files in the Raw Outputs folder



Overview

What is a master do-file?

- As you might have noticed, we mentioned the creation of many do-files in the previous slides
- A big project can become very complex, and do-files need to be ran 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
- **However**, you can make a do-file run other do-files
- This (and a few secondary but still important things) is what a master do-file does

What is a master do-file?

Here's one example of that:

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

*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"
```

Master do-file: the map to all data work

At the end of every round of a project

- You should be able to reproduce all your work from raw data to all outputs with one click in this do-file
- Anyone should be able to follow and to reproduce all your work from raw data to all outputs with one click in this do-file, after adding only their root folder path

Master do-file: allows easy collaboration

- If we share a project over DropBox, all team members have the same folder structure
- A master do-file allows multiple people to set their own global to the project folder
- This way, anyone sharing the project folder can easily run your do-files

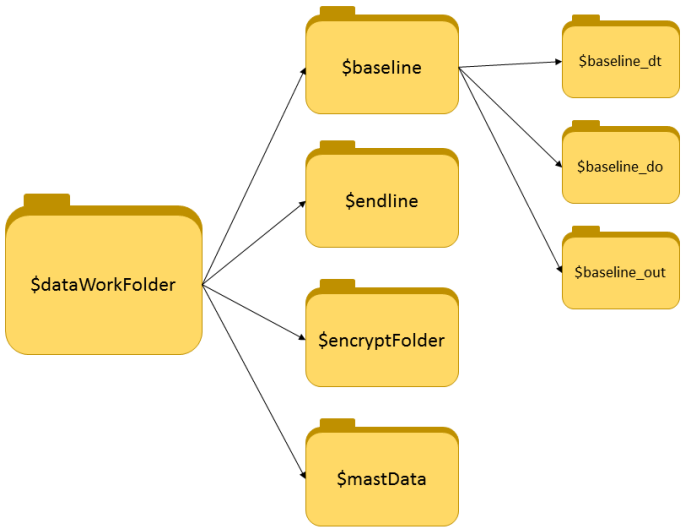
```
*Set this value to the user currently using this file
global user  1    // Luiza
global user  2    // Ben

* Root folder globals
* ---------------------

if $user == 1 {
    global projectfolder "C:\Users\WB501238\Dropbox\DIME\RA survey"
}

if $user == 2 {
    global projectfolder "C:\Users\Benjamin\Dropbox\Work\DIME\RA survey"
}
```

Master do-file: allows easy collaboration

The master do-file contains globals referencing the folders in your dropbox, so you have shortcuts that apply to all users

```
   * Project folder globals
   * ----------------------

   global dataWorkFolder         "$projectfolder/DataWork"

*iefolder*1*FolderGlobals*master************************************************
*iefolder will not work properly if the line above is edited

   global mastData               "$dataWorkFolder/MasterData"

*iefolder*1*FolderGlobals*rawData************************************************
*iefolder will not work properly if the line above is edited

   global encryptFolder          "$dataWorkFolder/EncryptedData"
   global masterIdDataSets       "$encryptFolder/IDMasterKey"


*iefolder*1*RoundGlobals*rounds*baseline*****************************************
*iefolder will not work properly if the line above is edited

   *baseline folder globals
   global baseline                 "$dataWorkFolder/baseline"
   global baseline_dt              "$baseline/DataSets"
   global baseline_do              "$baseline/Dofiles"
   global baseline_out             "$baseline/Output"
```

(a) Master do-file



(b) Dropbox folder

Overview

iefolder

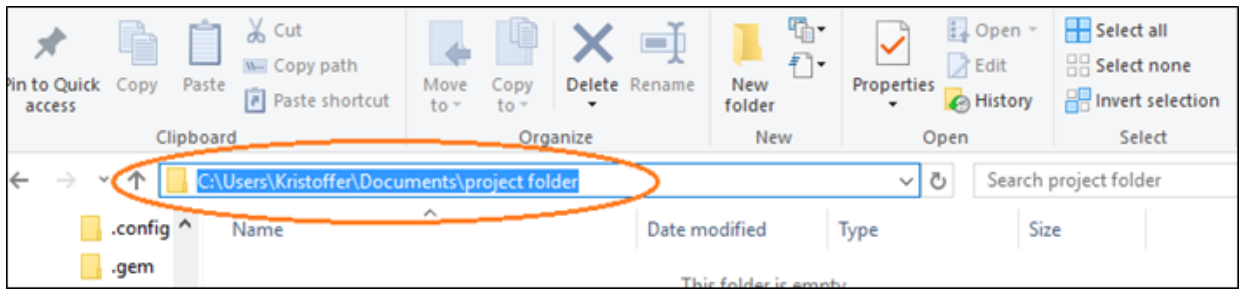
- Sounds complicated to set up? Dont worry, we have a command that sets that up for you
- It is called *iefolder* and it creates the folders and master do-files we have spoken about for you
- *iefolder* is part of *ietoolkit*. If you have not installed that package yet, then go to Stata and type

```
ssc install ietoolkit
```

iefolder: setting up a project folder

In the next few slides, we'll show you how to set up a DataWork folder using *iefolder* and explore some of its features

1. Create a folder on your computer and call it the project folder
2. Create a global called *projectfolder* to that folder. A shortcut to get the folder path is by clicking in the navigation bar of that folder:



3. Now create a new data work folder using iefolder like this:

```
iefolder new project, projectfolder(${projectfolder})
```

iefolder: setting up a round folder

Inside the data work folder you create folders for each round. A round is a data collection exercise. This could be a baseline or endline, for example. If you have simultaneous data collections on different unit of observations, create separate rounds for them as well.

1. To create new round, for example a baseline round, inside your DataWork folder using *iefolder*, type

```
iefolder new round baseline, projectfolder(${projectfolder})
```

iefolder: exploring your folder

- Your data work folder is now ready to be used
- The folder you just created has the structure we described earlier in this presentation
- Master do-files have also been created
  1. For the whole project
  2. For the baseline round
- You can create new rounds and units of observation when new data comes in. Make sure you use iefolder for that, so the new folders have the same structure and the master do-file is updated
- You can also customize your folder, but be sure to reflect any changes made to your folder structure in the master do-file