

Encryption and Data Security for Academic Research

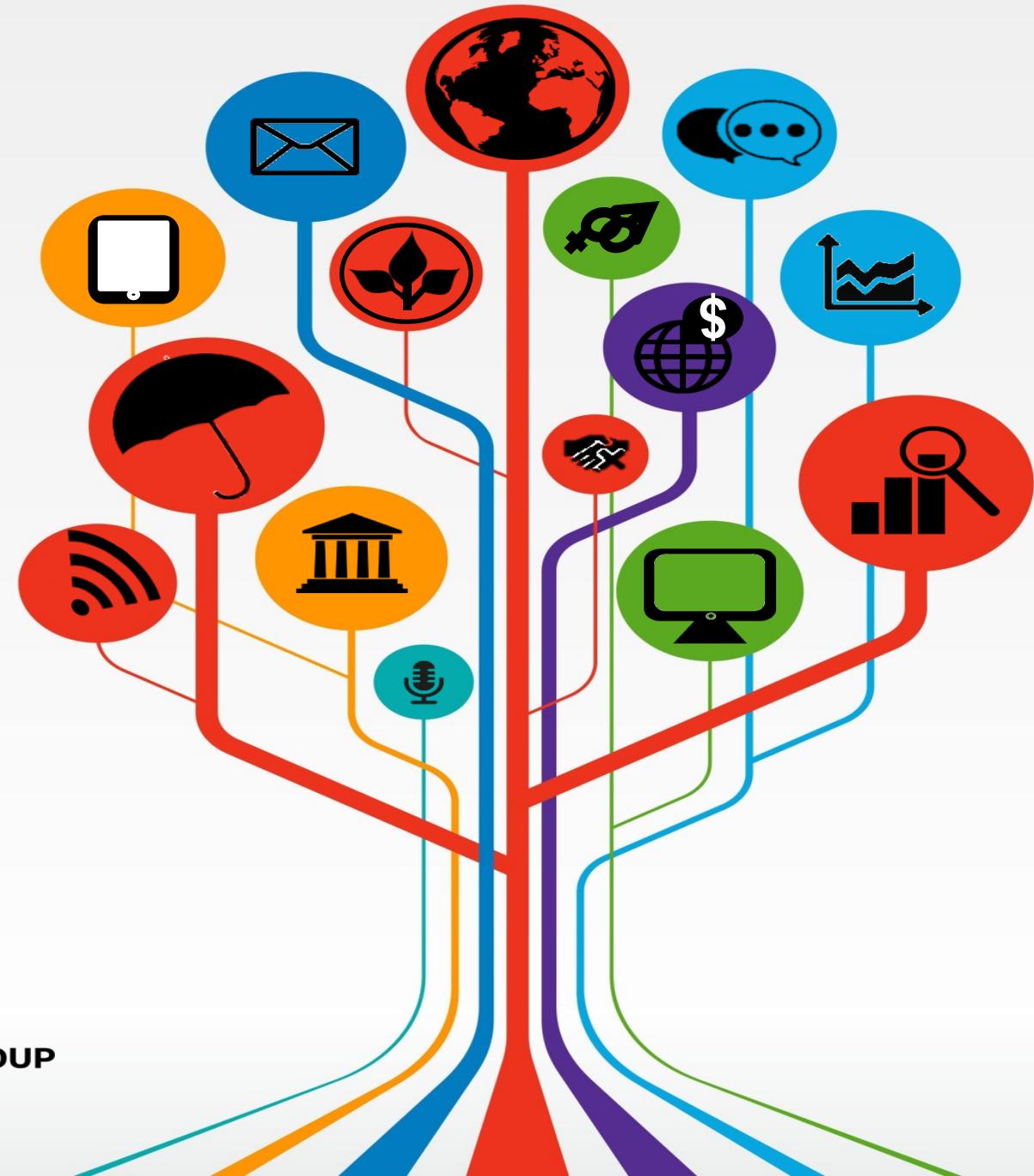
Research Assistant Onboarding

Prepared by DIME Analytics

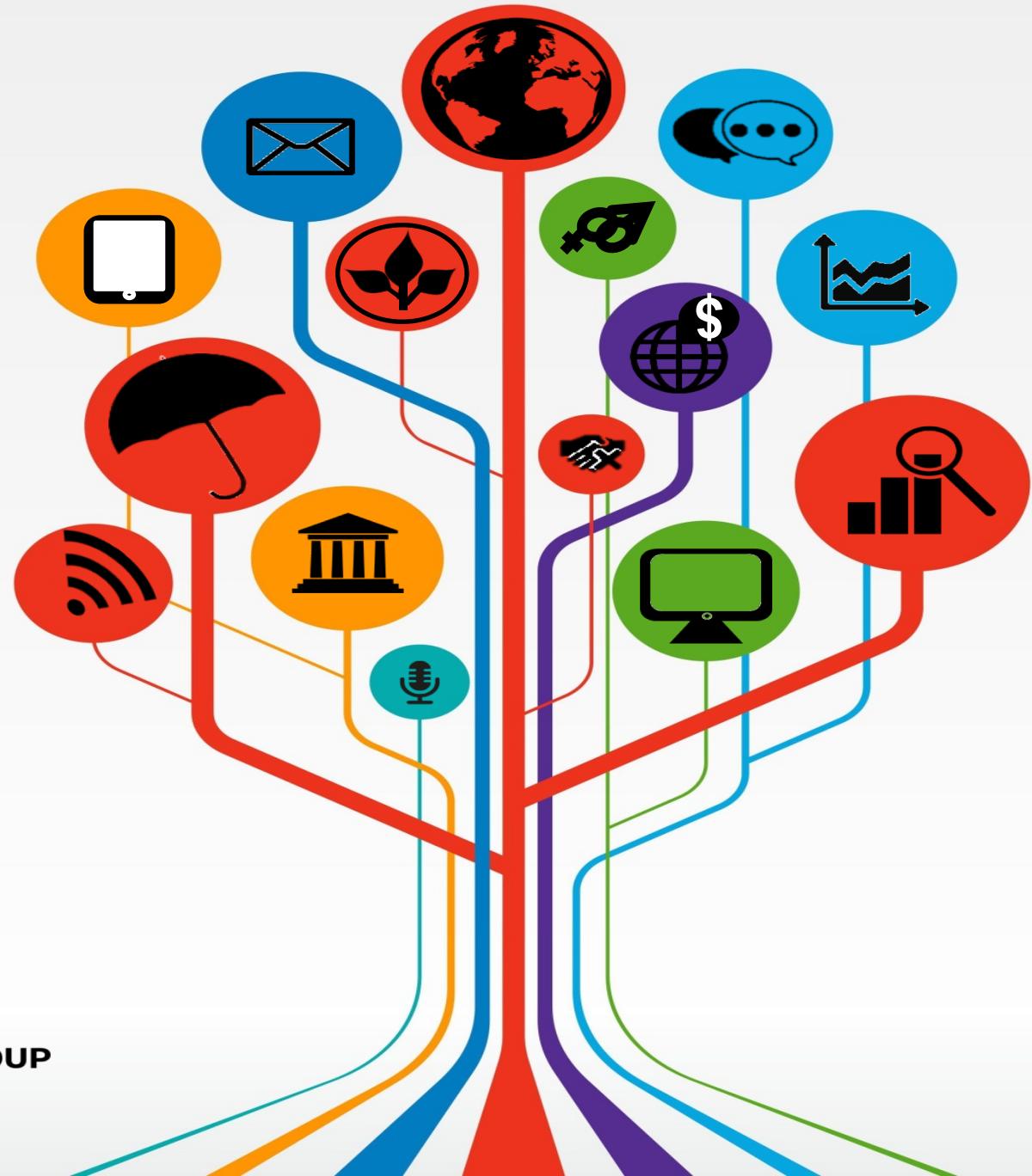
dimeanalytics@worldbank.org

Presented by Benjamin Daniels

bdaniels@worldbank.org



Data Security: Basic Ideas



You have a responsibility to keep data secure

- Ethical – it's the right thing to do
 - Contractual – IRBs, universities, and other partners require it
 - Legal – GDPR legal framework reformed data handling regulations based on a human right to privacy in the EU
-
- Homework: complete the National Institutes of Health course “Protecting Human Research Participants”. You will receive a certificate for this continuing education that is often required for academic submissions and trial registrations.
<https://phrp.nihtraining.com/#/>

Data security should be a core part of your life

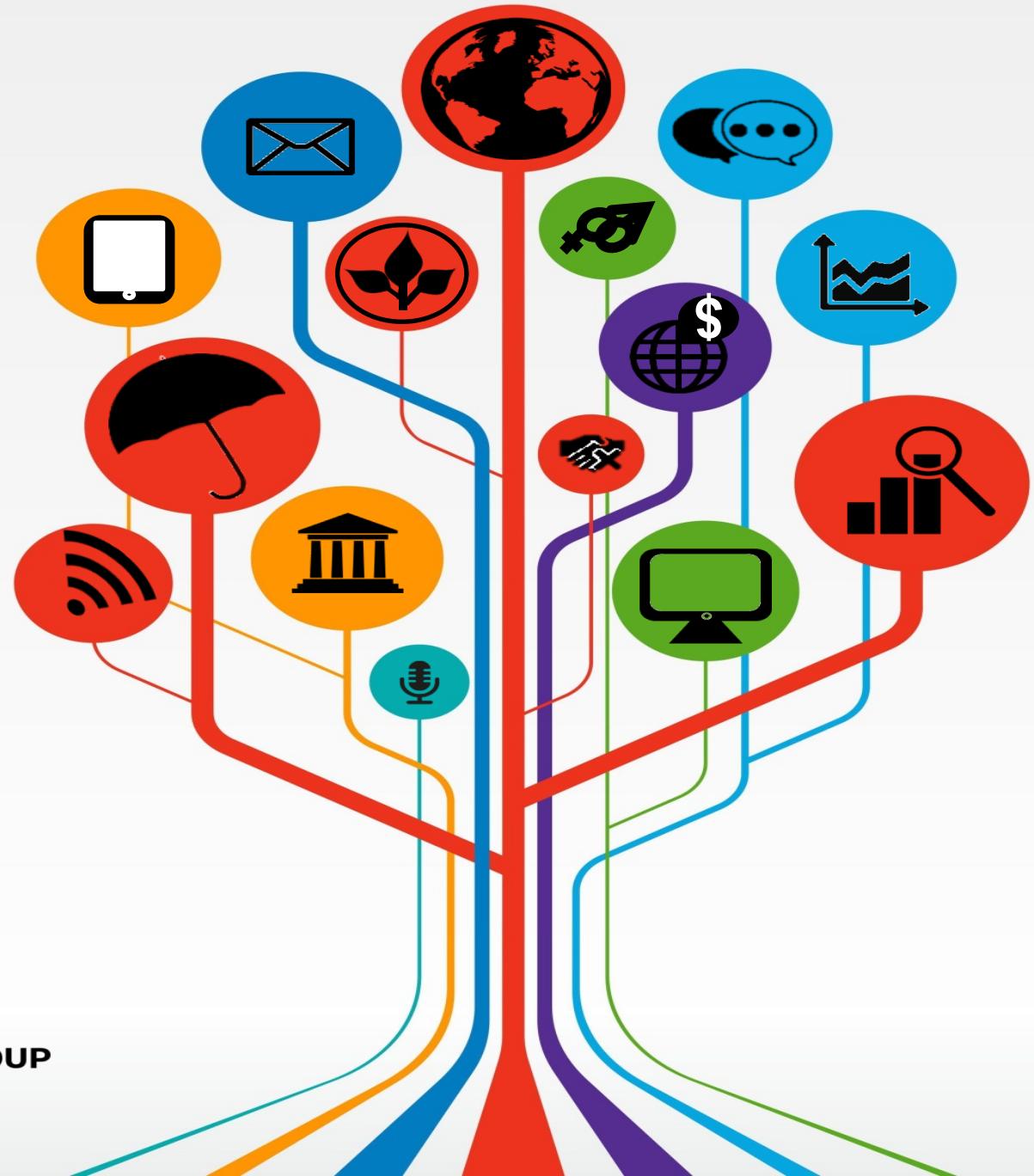
- This is not an “extra” thing
- So we design workflows that make it natural and beneficial to your organizational structure
- In this case, we will leverage our data security obligations to *also ensure*:
 - Easier workflows and naming conventions for surveys and datasets
 - Backed-up and 2FA-protected passwords for *all* personal accounts

Wait, what personal accounts?

- All accounts which have access to financial or personal information should be secured and backed up
- Modern encryption is “strong”, meaning if you lose your passwords, you lose your data – we are going to prevent that!
- Three core tools:
 1. Two-factor authentication (2FA)
 2. Password storage and backups
 3. Encryption

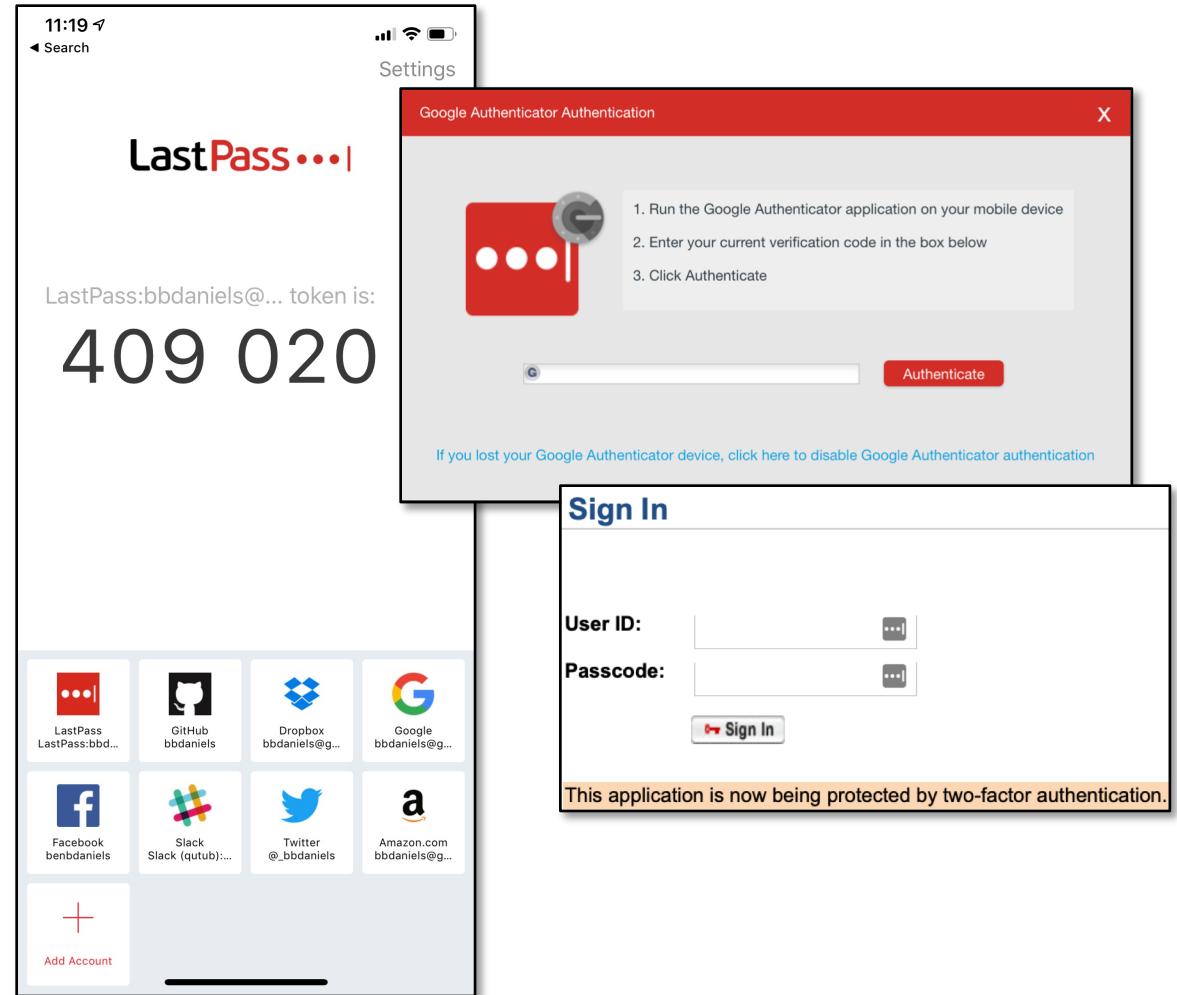


Tools – Outline



Two-factor authentication (2FA) is standard

- Any account you have which has personal or private information should be secured by 2FA
- First you enter your password, then you are prompted for a one-time “token”, which only exists on your phone (in Google Authenticator or Authy)
- This prevents anyone from accessing things like your Facebook account, your credit cards, or your survey data



Password management is critical with many

- LastPass is a free password manager (web and local) that allows you to generate, store, access, and share passwords
- We will use this to create and save passwords to encrypted information
- **Homework: set up Authy and LastPass**

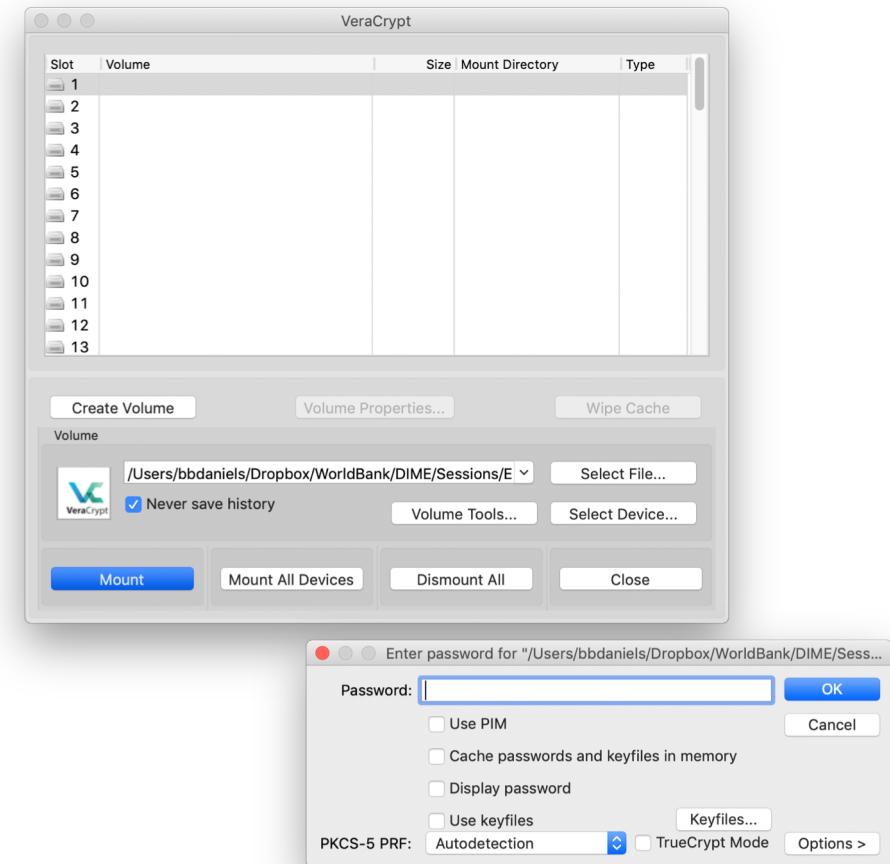
The screenshot displays the LastPass application interface. On the left, a sidebar shows 'SITES' (20), 'SECURE NOTES', and 'FORM FILLS'. The main area lists several saved items:

- amazon.com (bbdaniels@gmail.com)
- auth.lse.ac.uk (danielsb)
- bit.ly (bbdaniels)
- dropbox.com (bbdaniels@gmail.com)

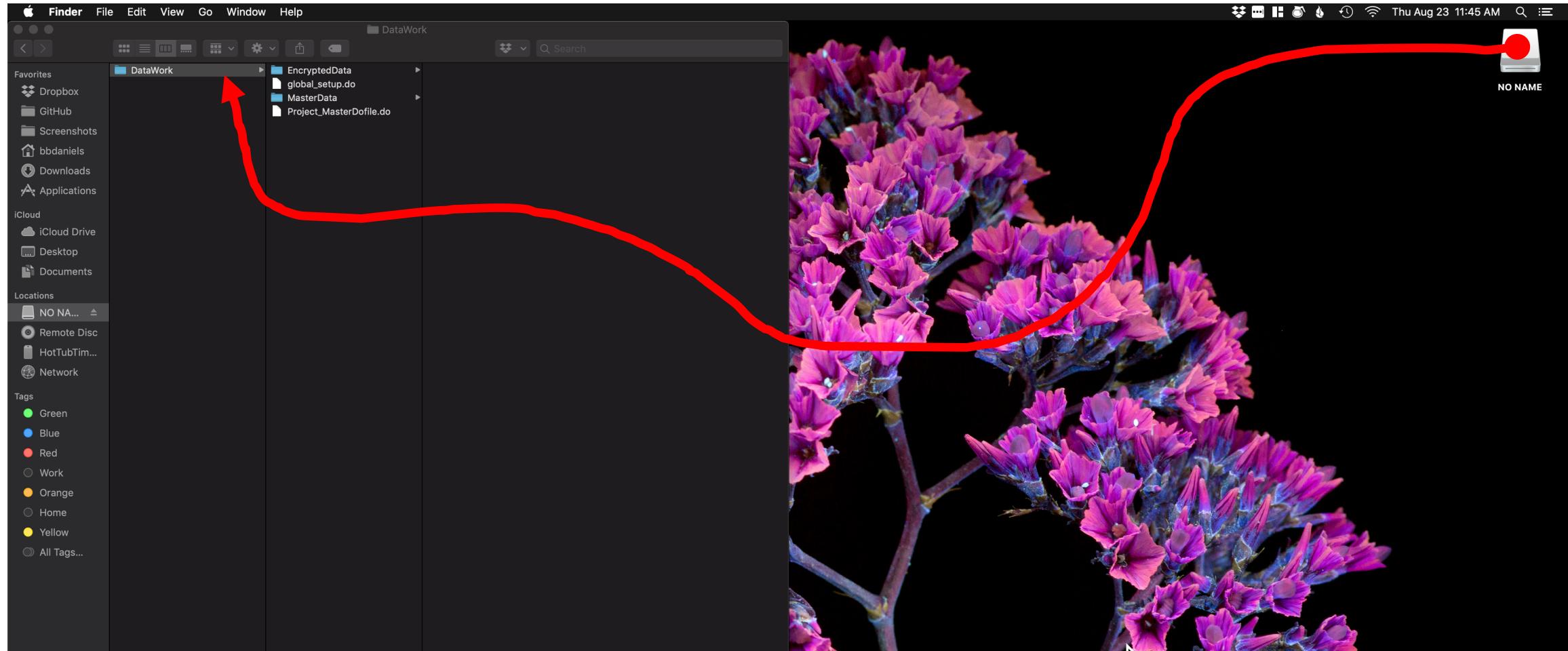
A context menu is open over the 'amazon.com' item, showing options like 'Edit', 'Share', 'Copy', 'Delete', and 'Import'. To the right, a detailed view of a saved item titled 'DIME-veracrypt-test' is shown, including fields for URL, Username, Password, Notes, and Folder. Below this, a 'Share Site' dialog box is open, prompting for 'Recipient Email Addresses' (robmarty3@gmail.com) and a checkbox for 'Allow Recipient to View Password'.

VeraCrypt creates secure objects to contain data

- This works by setting aside a fixed amount of storage space on the hard drive
- This object has only a name, no file extension (similar to a folder)
- It is encrypted using a strong password and randomness generated by moving the mouse
- “Mounting” this object with VC is like plugging in a flash drive

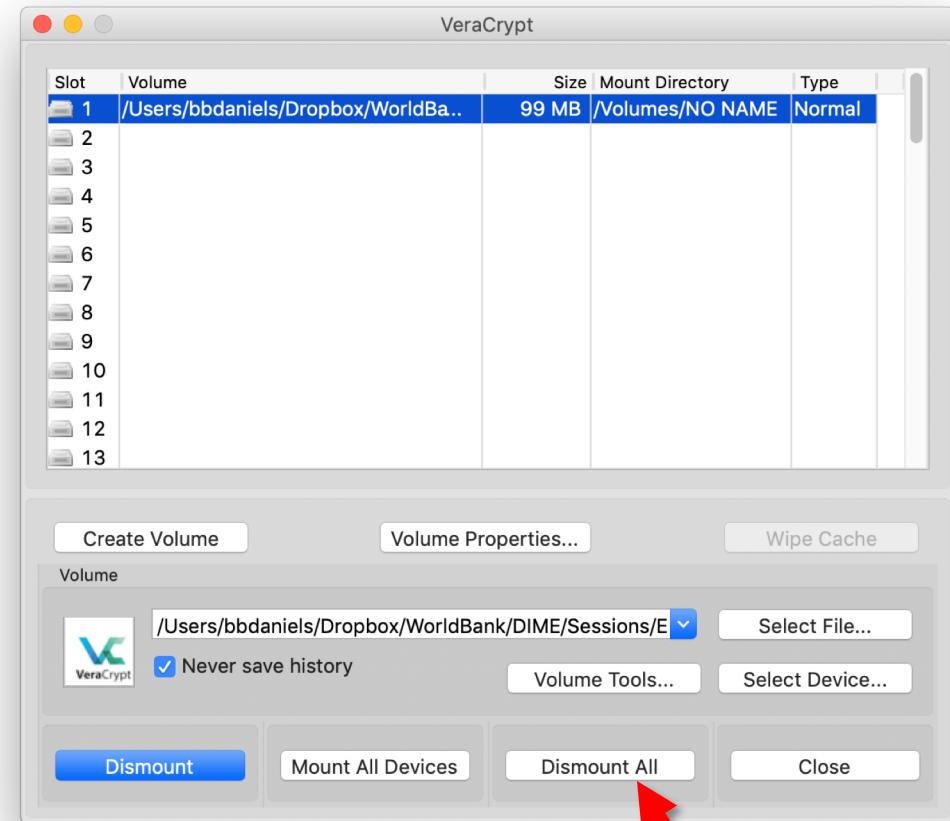


VC Objects are only editable when “mounted”

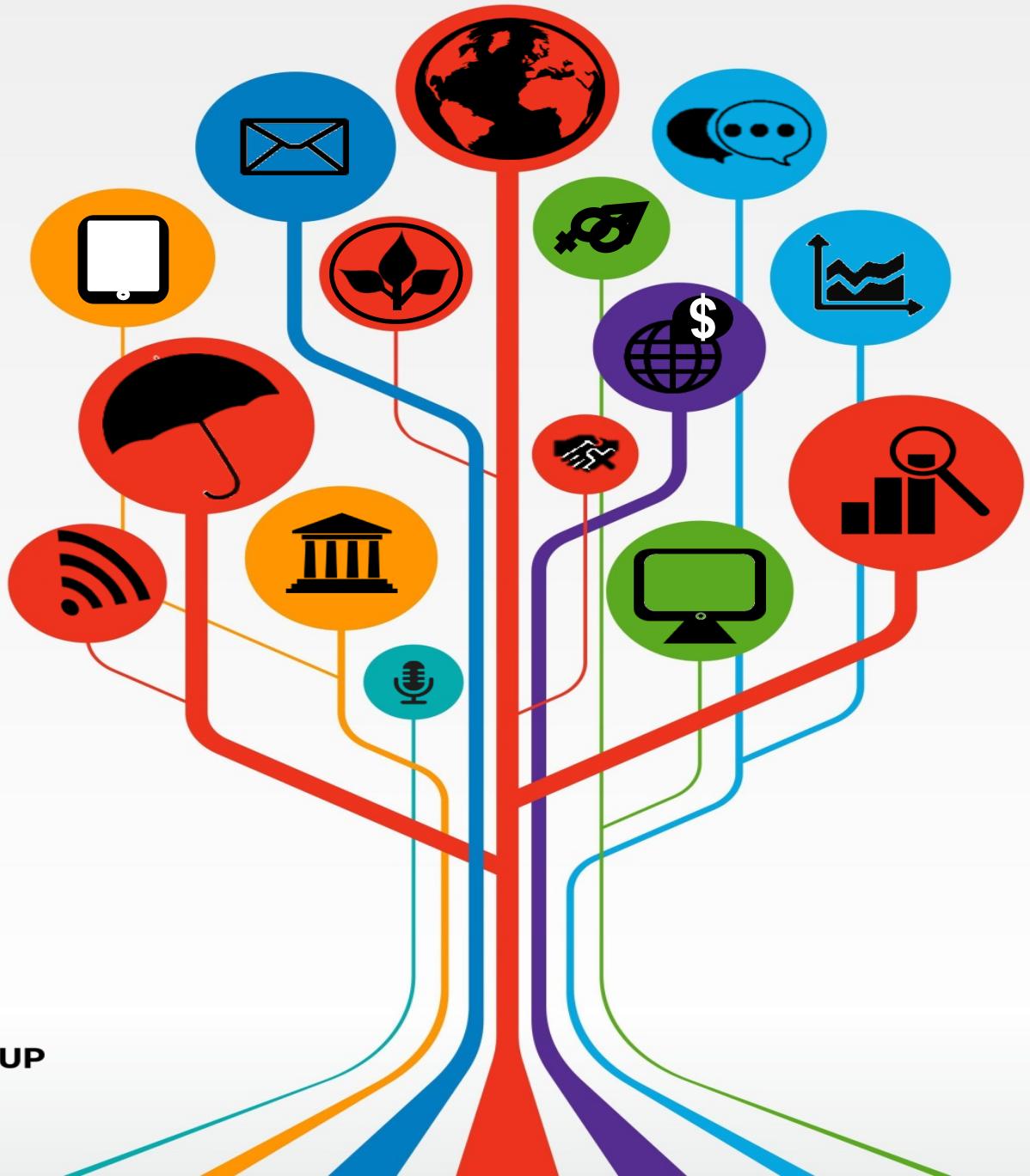


All changes are encrypted when dismounted

- Once the drive is dismounted, it disappears from the file system and then the encrypted object is updated by VeraCrypt
- Your changes are not saved to your hard drive until you complete this step!

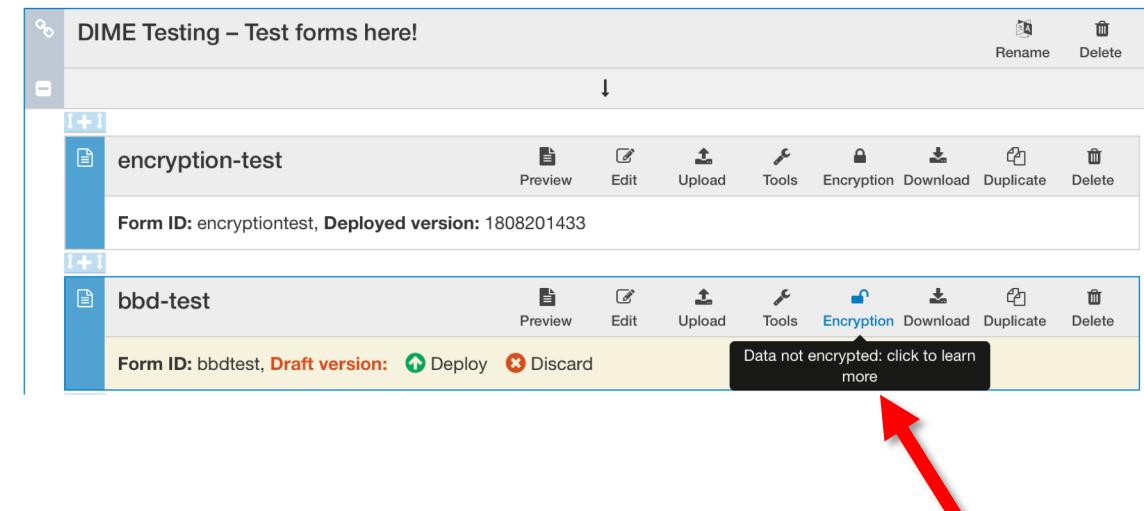


Workflow – Encrypted Data with SurveyCTO



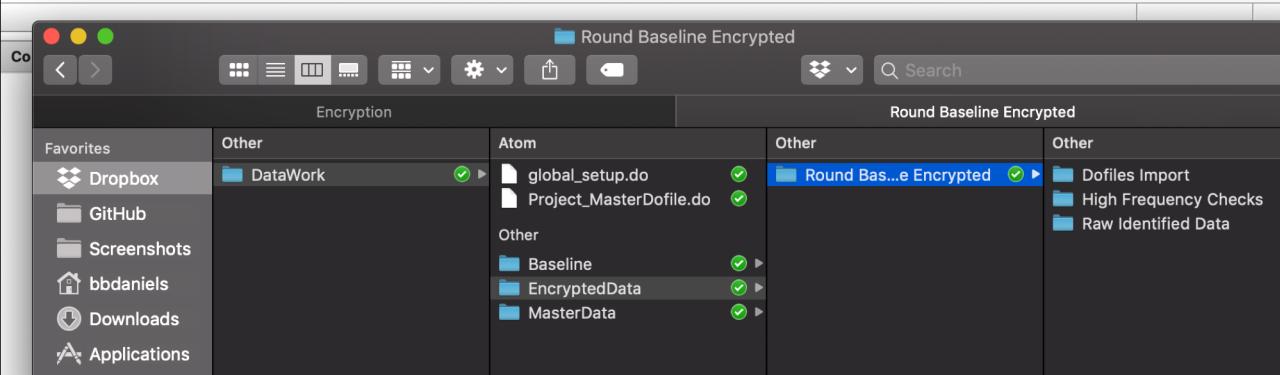
WB SurveyCTO server data *must* be encrypted

- If form data is not encrypted on the server, all server administrators can access and export your data
- SurveyCTO provides easy tools to make sure your new surveys are encrypted
- Since there are several pieces, the key in this workflow is *organization*.



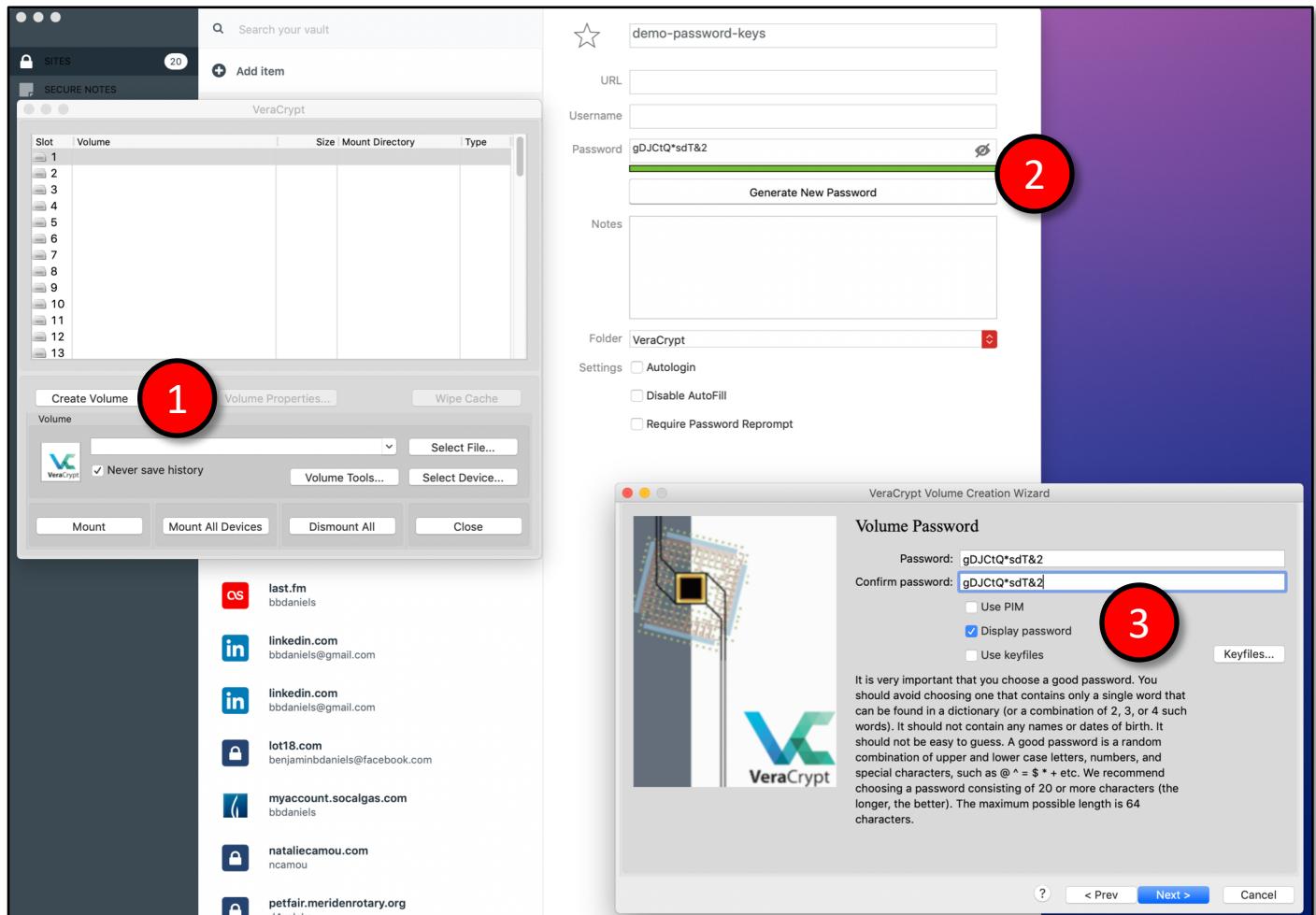
[iefolder] rounds have an EncryptedData folder

- First, we need to create a VeraCrypt object here for our survey keys. 1MB will be large enough since these are small files.

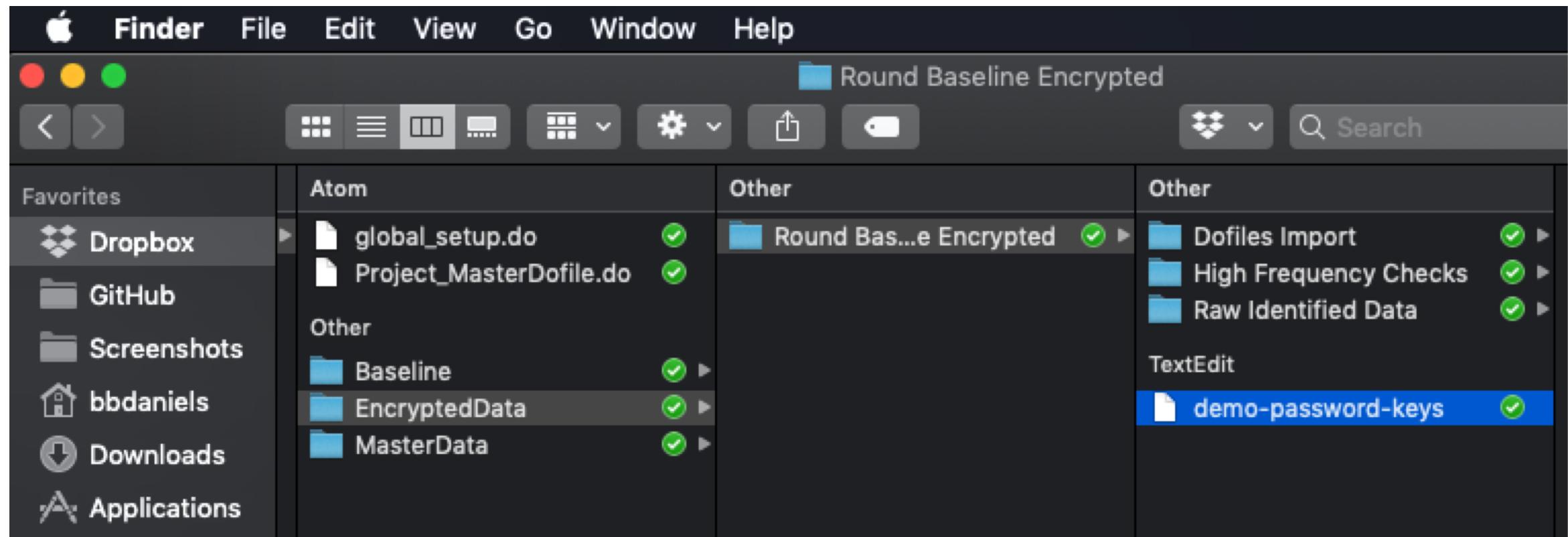
```
. iefolder new project , projectfolder("/Users/bbdaniels/Dropbox/WorldBank/DIME/Sessions/Encryption/Demo")  
  
Command ran successfully, a new DataWork folder was created here:  
1) [/Users/bbdaniels/Dropbox/WorldBank/DIME/Sessions/Encryption/Demo/DataWork]  
  
. iefolder new round Baseline , projectfolder("/Users/bbdaniels/Dropbox/WorldBank/DIME/Sessions/Encryption/Demo")  
  
Command ran successfully, for the round [Baseline] the following folders and master dofile were created:  
1) [/Users/bbdaniels/Dropbox/WorldBank/DIME/Sessions/Encryption/Demo/DataWork/Baseline]  
2) [/Users/bbdaniels/Dropbox/WorldBank/DIME/Sessions/Encryption/Demo/DataWork/EncryptedData/Round Baseline Encrypted]  
3) [/Users/bbdaniels/Dropbox/WorldBank/DIME/Sessions/Encryption/Demo/DataWork/Baseline_MasterDofile.do]  
  
.  
  

```

First, use VeraCrypt to create the volume

- VeraCrypt: “Create Volume” at the location of the Round Encrypted folder.
- LastPass: “add item” with the detailed name of the survey, round, and –survey-keys
- LastPass: “generate password”, then be sure to “Save”
- VeraCrypt: Copy and paste the plaintext password (“display password”, then follow instructions)
- JPAL guide to VeraCrypt:
https://www.povertyactionlab.org/sites/default/files/A%20guide-to-VeraCrypt-Installation-and-Demo_Sep-2016.pdf

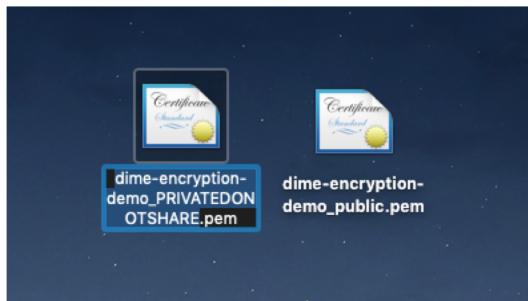


Rename the volume to match the password



Create encryption keys on SurveyCTO

- Use the key generator to create a “public-private key pair” for your data, with the name of your intended survey. These will download two files as below:



The image shows the SurveyCTO software interface. At the top, there are four tabs: "1. Design" (with a dropdown arrow), "2. Collect", "3. Monitor", and "4. Export", followed by a "Configure" button. Below the tabs, a green header bar contains the text "How to design your survey forms". Underneath this, a section titled "Your forms and datasets" features a toolbar with various icons: Tools, Refresh, Search, Organize, Help, Build constraint, Test constraint, Build relevance, Build calculation, Create new key (which is highlighted with a red arrow), and Add sample form. A black button at the bottom right of the toolbar says "Generate encryption keys".

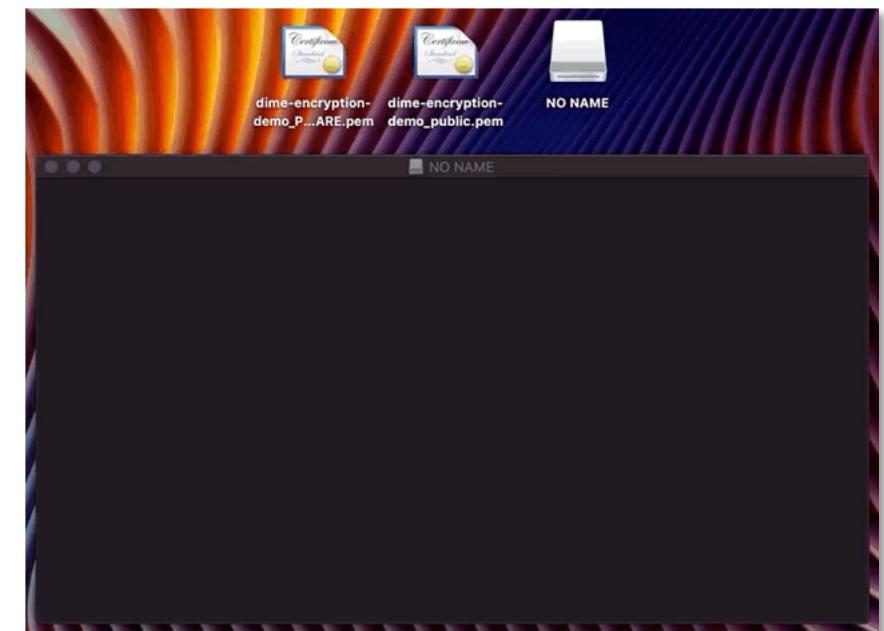
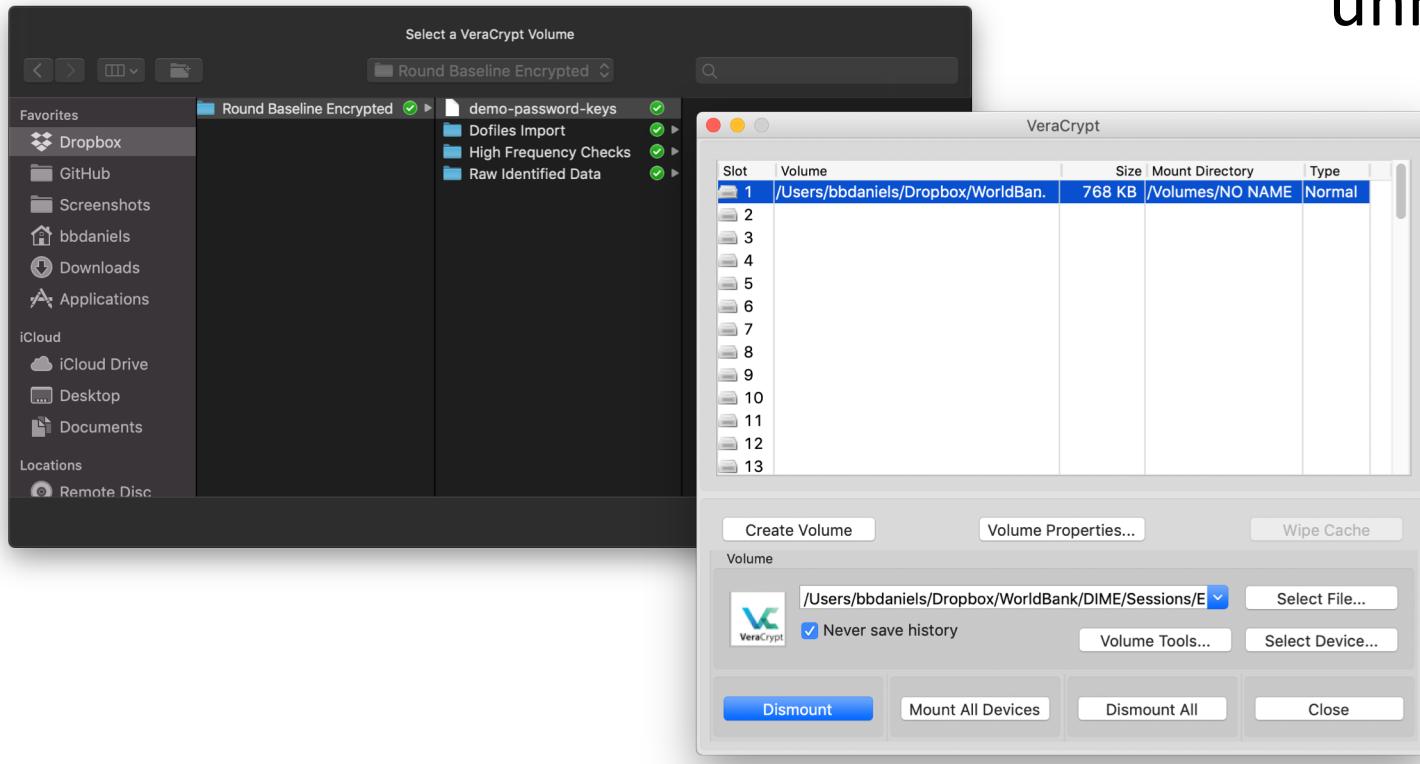
Create an encrypted survey with the *public key*

The screenshot shows the DIME form creation interface with two main steps:

- Step 1: Start new form - Step 1**
 - Form title: dime-encryption-demo
 - Form ID: dimeencryptionsdemo
 - Use a sample form as your starting point? (OFF)
 - Advanced options (ON)
 - Do you want this form's data to be encrypted? (checked)
 - Auto-generate fields necessary for pre-loading data?
- Step 2: Start new form - Step 2: Configure encryption**
 - To encrypt your form data, you will need your own encryption key, which you can get from the [Tools...Create new key](#) option at the very top of this "Your forms" section. [Click here to learn more...](#)
 - Upload public key (radio button selected)
 - Paste public key text
 - Upload public key here:
Choose File dime-encrypt...o_public.pem

Then put the keys in the vault!

- Mount the VC object using VeraCrypt and LastPass
- Then put the keys in the mounted drive, delete them and unmount it with VeraCrypt.



Form data download now requires the private key

dime-encryption-demo

Form ID: dimeencryptiondemo, Complete submissions: 1 (latest Aug. 23, 2018 at 12:26:51PM)

Download your data

You can download and export your data into a .csv format from here in your web browser.

Export to:

- Wide format ([Learn more...](#))
- Long format

Submissions to include:

- All submissions
- A random subset of submissions (1 out of 1)
- Recent submissions only (from the last 7 days)

Fields to include:

- All fields (if you have the private key)
- Publishable fields only (if you don't have the private key)

Note that your encrypted data will be decrypted within your browser memory and then exported to one or more local files. You will then be sure that private data is not shared with anybody who should not have access.

[Download data now](#) [Close](#)

Private key required



To decrypt your data, please find and select the appropriate private key when prompted (hint: its filename probably ends with "PRIVATEDONOTSHARE.pem").

And don't worry: neither your private key nor your decrypted data will be shared on the Internet. Note that your encrypted data will be decrypted within your browser memory and then exported to one or more local files.

[Cancel](#) [Select private key](#)

Data Exporter

Explore [Help](#) [Close](#)

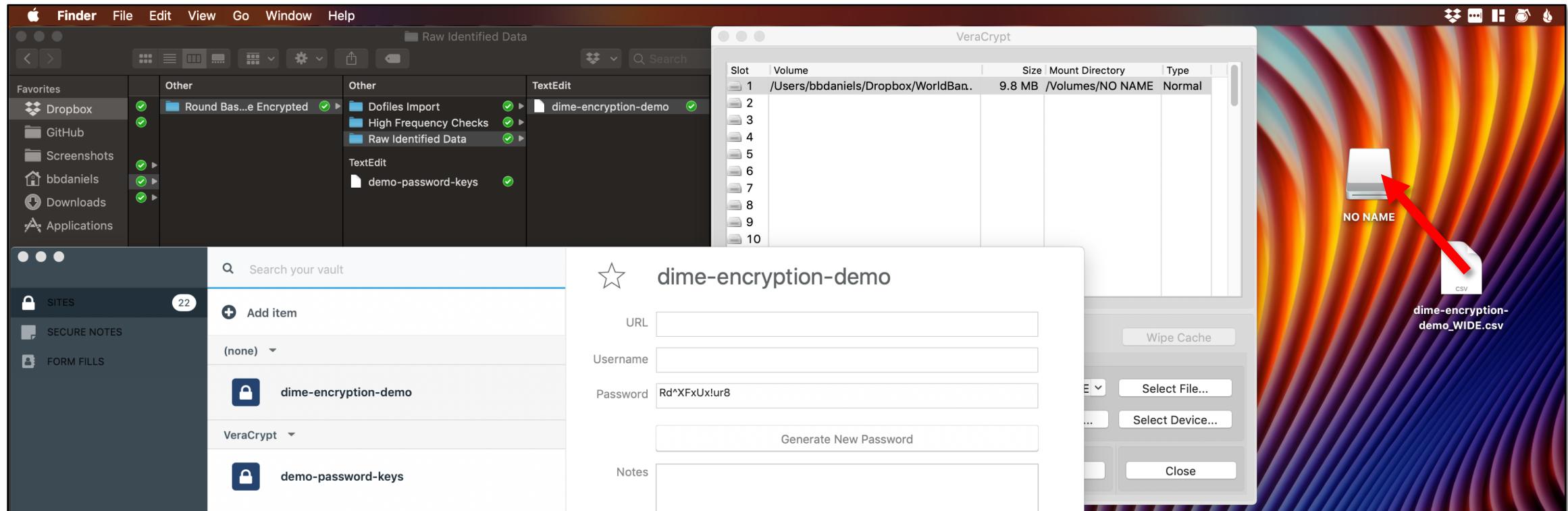
Exported data
dime-encryption-demo_WIDE.csv

Other download options:

[SurveyCTO Sync](#) [Printable version](#) [Mail merge template](#) [Stata .do template](#)

The most powerful and flexible way to download, process, and export your data is to use SurveyCTO Sync, software that runs on your local computer. [Click here](#) to learn more about SurveyCTO Sync, and get started by downloading and installing the appropriate version, based on your computer's operating system: [download for Windows](#), [download for Mac \(OSX\)](#), or [download for other platforms](#).

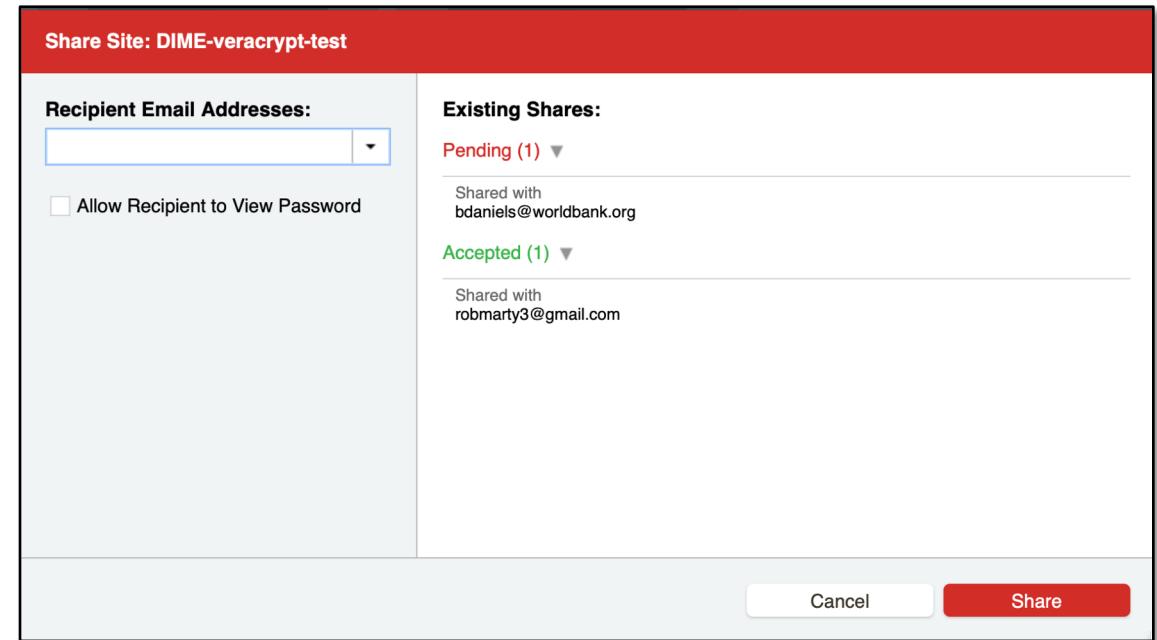
Create a separate container for the data



(This is necessary since you will not know the size of the data when you create the vault for the keys.)

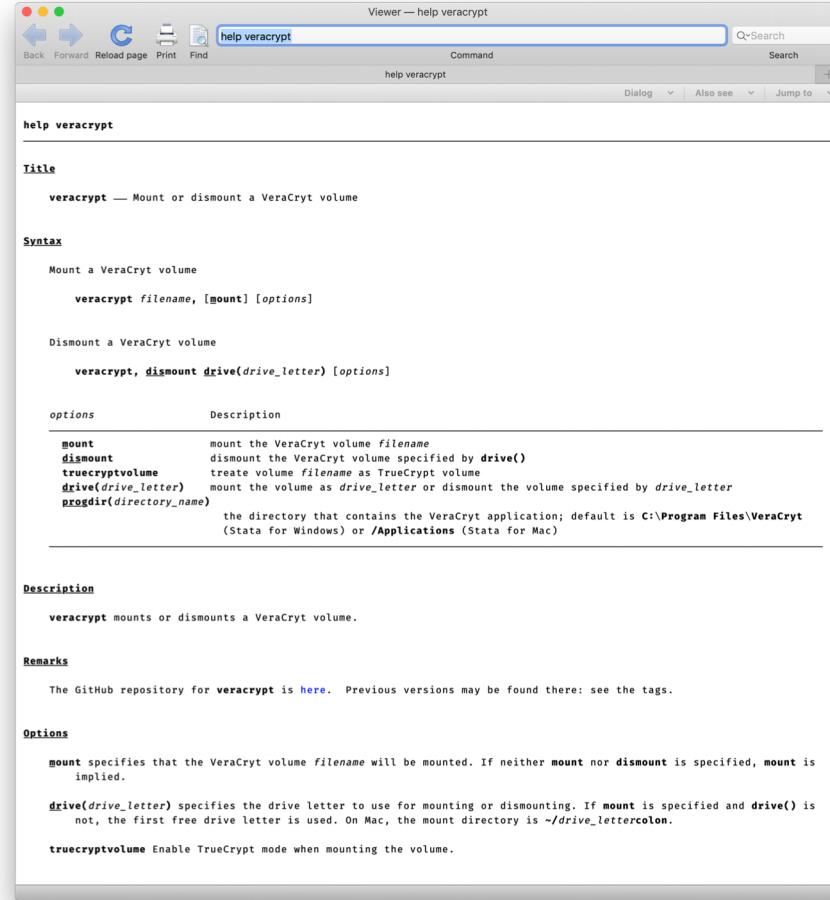
You can share securely using LastPass and Dropbox

- With this method, you never transmit information outside of a secure setting, but all data is synced with the features of Dropbox as you are used to it.



Your first cleaning dofile will create non-PII data

- The [veracrypt] command allows Stata to call for VeraCrypt to mount the drive.
- You have to manually enter the password any time you run the dofile.
- The first cleaning file should move a non-PII version of your data to the regular Data folder.
- *Tip: create a “second language” in the survey called “stata” with the variable labels you want. For variable names, best practice is to use the question codes so that the first dataset matches the survey exactly.*



Rapid cleaning: Codebook commands

Three commands in development:

- **importCodebook**
 - Merges data across multiple survey rounds, useful when labelling and variable names need to be harmonized
- **applyCodebook**
 - Allows you to clean data (rename, recode, value label, variable label) very quickly
- **exportCodebook**
 - Creates a “release” version of your data which only contains the variables used in your analysis dofiles
- Need testing on all three!
 - Hoping to release later this year in iefieldkit as [*iecodebook*] commands.
- <http://worldbank.github.io/stata>

Demos of Codebook commands

importCodebook

importCodebook allows the user to create an Excel-based metadata file, then import one or more .xlsx or .dta files, including harmonizing variable naming and categorical coding and labeling. It can be used to expedite the cleaning of a single file or to combine (append) different surveys or survey rounds, taking the “hard work” out of the dofile.

```
A          B          C          D          E          F          G          H
1 Variable Label  Variable Name  Value Label  kenya      qutub      delhi      china      birbhum
2 Study          study          study          study      study      study      study
3 Observation Type type          type          facilitycode provid      provid      provider_code
4 Facility ID   facilitycode  facilitycode  provid      provid      provider_code
5 SP Case        case          case          case_new    case_new    case_new    sp_case
6 Provider Male  prov_male    cp_18       pro_male    cp_17      pro_age     cp_17      pro_male
7 Provider Age   prov_age    cp_17       pro_age     cp_17      pro_age     provider_age
8 Private Facility private     facility_nch private    private    private    private
```

* Combine SPs

import_metadata ///
"\$directory/Data/Public/kenya_sp.dta" ///
"\$directory/Data/Public/qutub_sp.dta" ///
"\$directory/Data/Public/delhi_sp.dta" ///
"\$directory/Data/Public/china_sp.dta" ///
"\$directory/Data/Public/birbhum_sp.dta" ///
using ///
"\$directory/Data/Metadata/SP_Codebook.xlsx" ///
, old(kenya qutub delhi china birbhum)

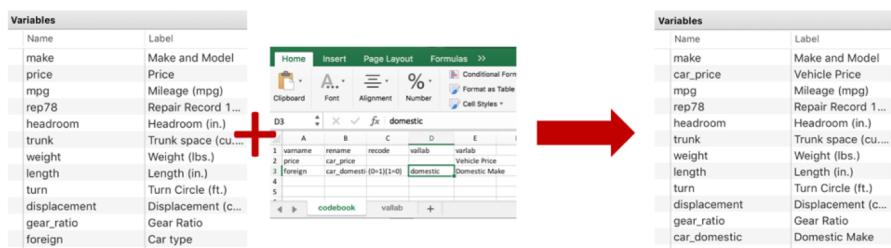
. ta study case

Study	SP Case								Total
	Angina	Asthma	Diarrhea	TB1 Naive	TB2 CXR	TB3 AFB	TB4 MDR	Total	
Birbhum C	392	395	398	0	0	0	0	1,185	
Birbhum T	392	392	395	0	0	0	0	1,179	
China	0	0	0	578	0	0	0	578	
Delhi	0	0	0	168	75	50	50	343	
Kenya	42	42	40	42	0	0	0	166	
Madhya Pradesh	723	850	830	0	0	0	0	2,403	
Mumbai	0	0	0	884	247	204	328	1,583	
Patna	0	0	0	573	138	150	158	1,019	
Total	1,549	1,679	1,663	2,165	460	404	536	8,456	

```
wb_git_install importCodebook  
[see documentation for extensive examples]
```

applyCodebook

applyCodebook allows the user to create an Excel codebook file, which will apply renames, recodes, variable labels, and value labels to the open dataset. It can be used to expedite the cleaning of a single file , taking the “hard work” out of the dofile.

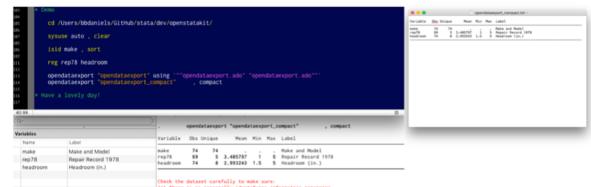


```
. applyCodebook using "applyCodebook_DEMO.xlsx" , rename varlab recode vallab  
Codebook applied!
```

exportCodebook

exportCodebook reads the currently open dataset and either (A) creates a codebook for it in the specified location; or (B) reads a series of .dofiles that reference the data and keeps only the variables that those dofles reference.

```
opendataexport: Data + list of dofiles → trimmed dataset + codebook
```



```
wb_git_install exportCodebook  
sysuse auto , clear  
exportCodebook "exportCodebook_compact" , compact
```

```
wb_git_install applyCodebook  
sysuse auto, clear
```

```
applyCodebook ///  
using "applyCodebook_DEMO.xlsx" ///  
, rename varlab recode vallab
```

Thank you!

