Data cleaning with [*iecodebook*]

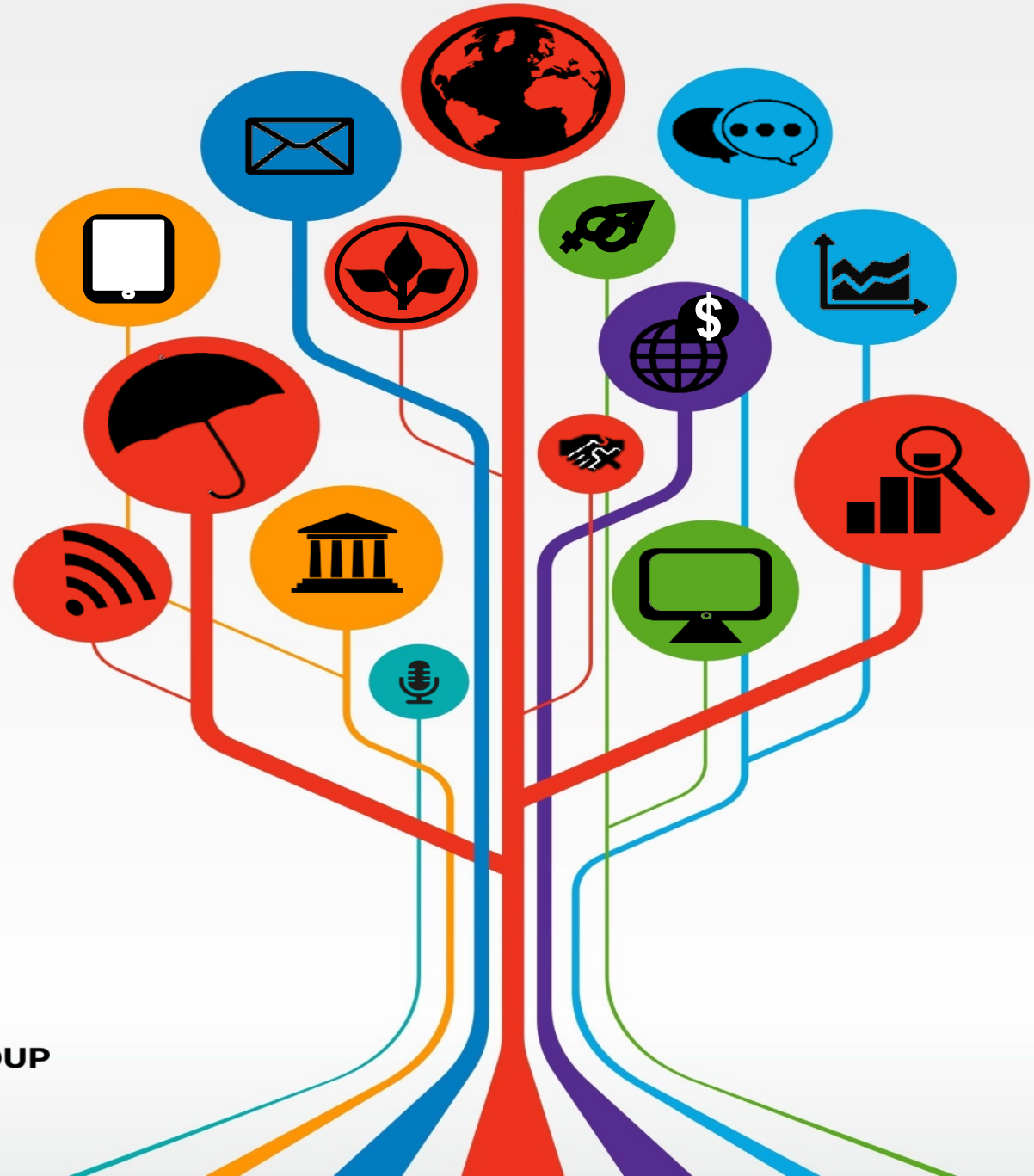
https://dimewiki.worldbank.org/wiki/Iecodebook

Prepared by **DIME Analytics**

dimeanalytics@worldbank.org

worldbank.github.com/dimeanalytics

github.com/worldbank/iefieldkit



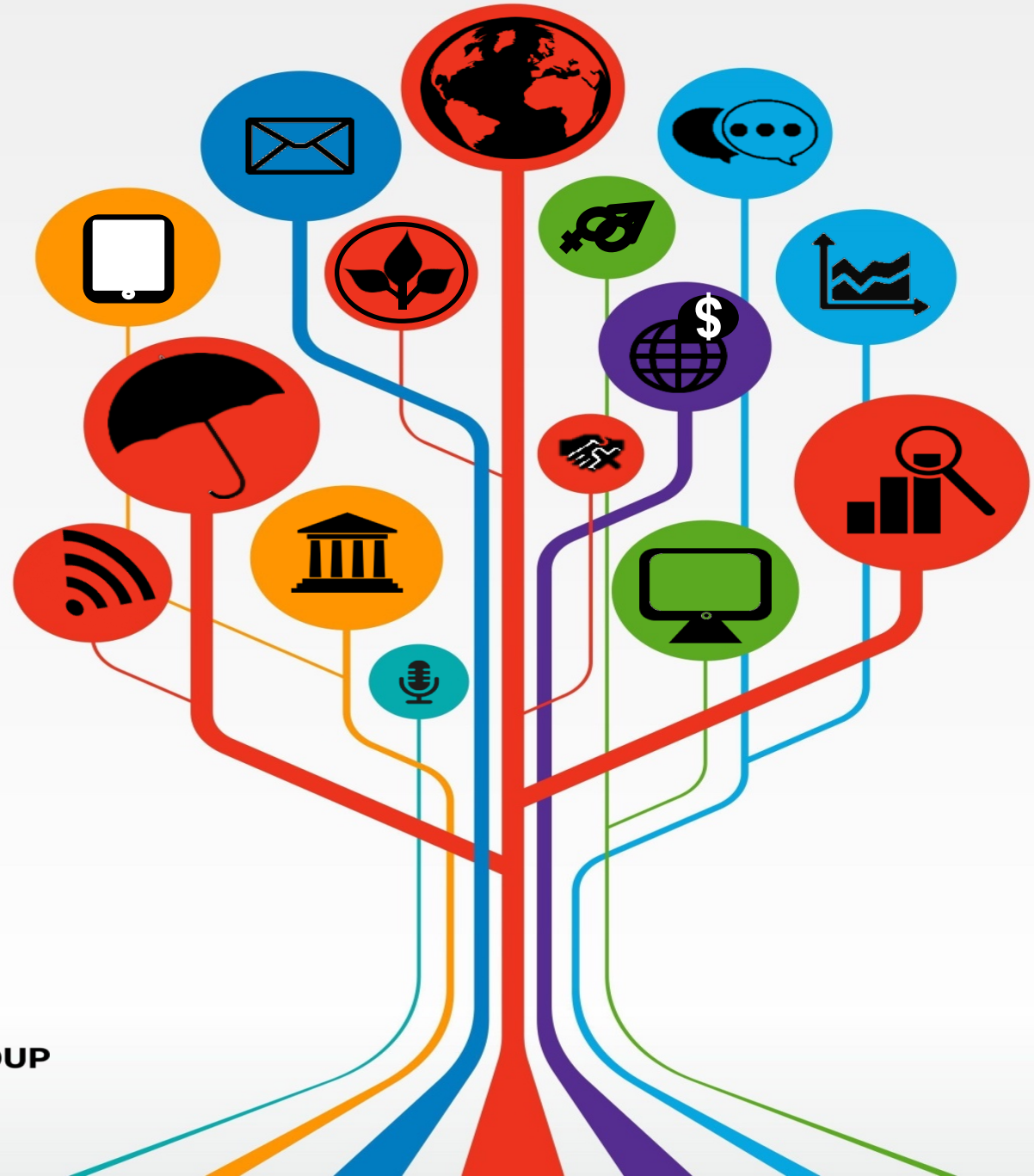
i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Three tasks for reproducible analysis

- **Data cleaning: [*iecodebook apply*]**
  Reads an Excel codebook that specifies renames, recodes, variable labels, and value labels, and applies them to the current dataset.
- **Dataset combination: [*iecodebook append*]**
  Reads an Excel codebook that specifies how variables should be harmonized across two or more datasets - rename, recode, variable labels, and value labels - applies the harmonization, and appends the datasets.
- **Data documentation: [*iecodebook export*]**
  Creates an Excel codebook that describes the current dataset, and optionally produces an export version of the dataset with only variables used in specified dofiles.

Data Cleaning:
[*iecodebook apply*]

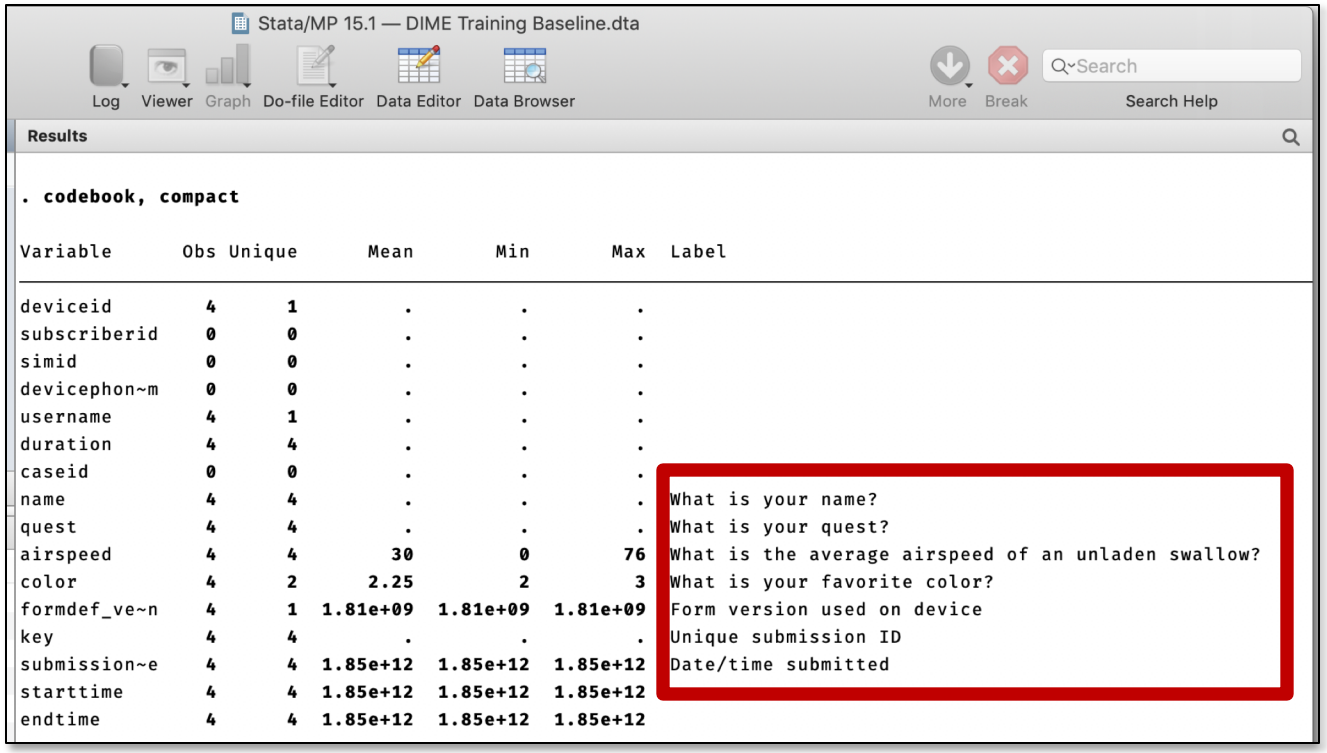


i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Use case: Raw data needs basic cleaning

- PII data can't be stored on Dropbox unencrypted
- The labels that are useful for surveys are terrible in Stata
- De-identification dofile:
  1. Removes all PII variables
  2. Renames and labels variables sensibly for data work (ie, no full questions, English only, no special characters)



[*iecodebook apply*] does both in one command

```
5  *
6
7      use "/Volumes/NO NAME/DIME Training Baseline.dta" , clear
8
9          iecodebook apply ///
10             using "${R1_Baseline_encrypt}/Raw Identified Data/baseline_metadata.xlsx" ///
11             , drop
12
13
14
15 *
16
```

Stata/MP 15.1 — DIME Training Baseline.dta

Log Viewer Graph Do-file Editor Data Editor Data Browser More Break Search Search Help

Results

```
. codebook, compact

Variable      Obs Unique      Mean       Min       Max  Label
-------------------------------------------------------------------------------
deviceid        4      1         .         .         .
subscriberid    0      0         .         .         .
simid           0      0         .         .         .
devicephon~m    0      0         .         .         .
username        4      1         .         .         .
duration        4      4         .         .         .
caseid          0      0         .         .         .
name            4      4         .         .         .  What is your name?
quest           4      4         .         .         .  What is your quest?
airspeed        4      4        30         0        76  What is the average airspeed of an unladen swallow?
color           4      2      2.25         2         3  What is your favorite color?
formdef_ve~n    4      1  1.81e+09  1.81e+09  1.81e+09  Form version used on device
key             4      4         .         .         .  Unique submission ID
submission~e    4      4  1.85e+12  1.85e+12  1.85e+12  Date/time submitted
starttime       4      4  1.85e+12  1.85e+12  1.85e+12
endtime         4      4  1.85e+12  1.85e+12  1.85e+12
```

Variables

Name	Label
your_quest	Quest
airspeed	Swallow Airspeed
color	Favorite Color

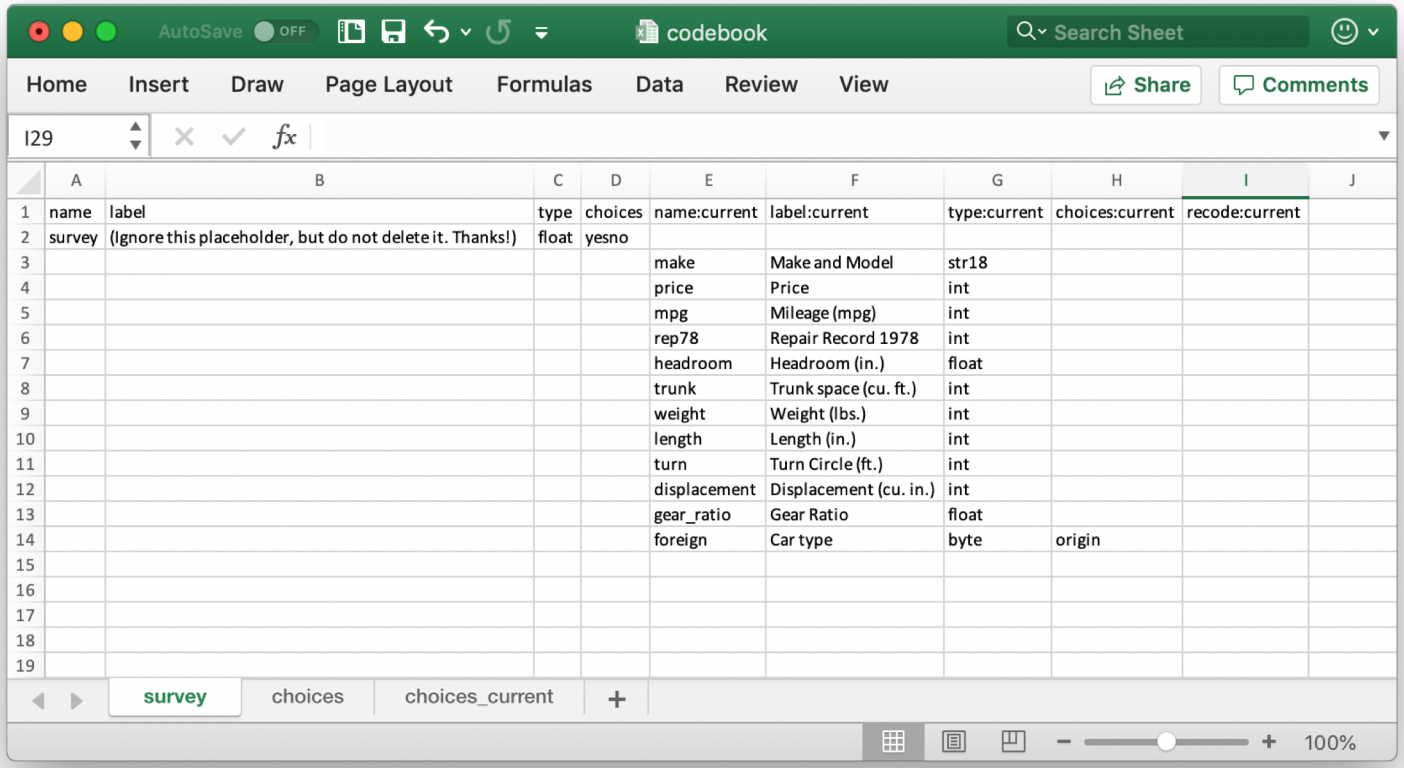
Ready for the Microdata Catalog!

Step 1: Set up Excel codebook template

```
// Load data
sysuse auto.dta , clear


// Create cleaning template
iecodebook template
  using "codebook.xlsx"
```

"Template" command sets up this entire sheet with the current state of the dataset.

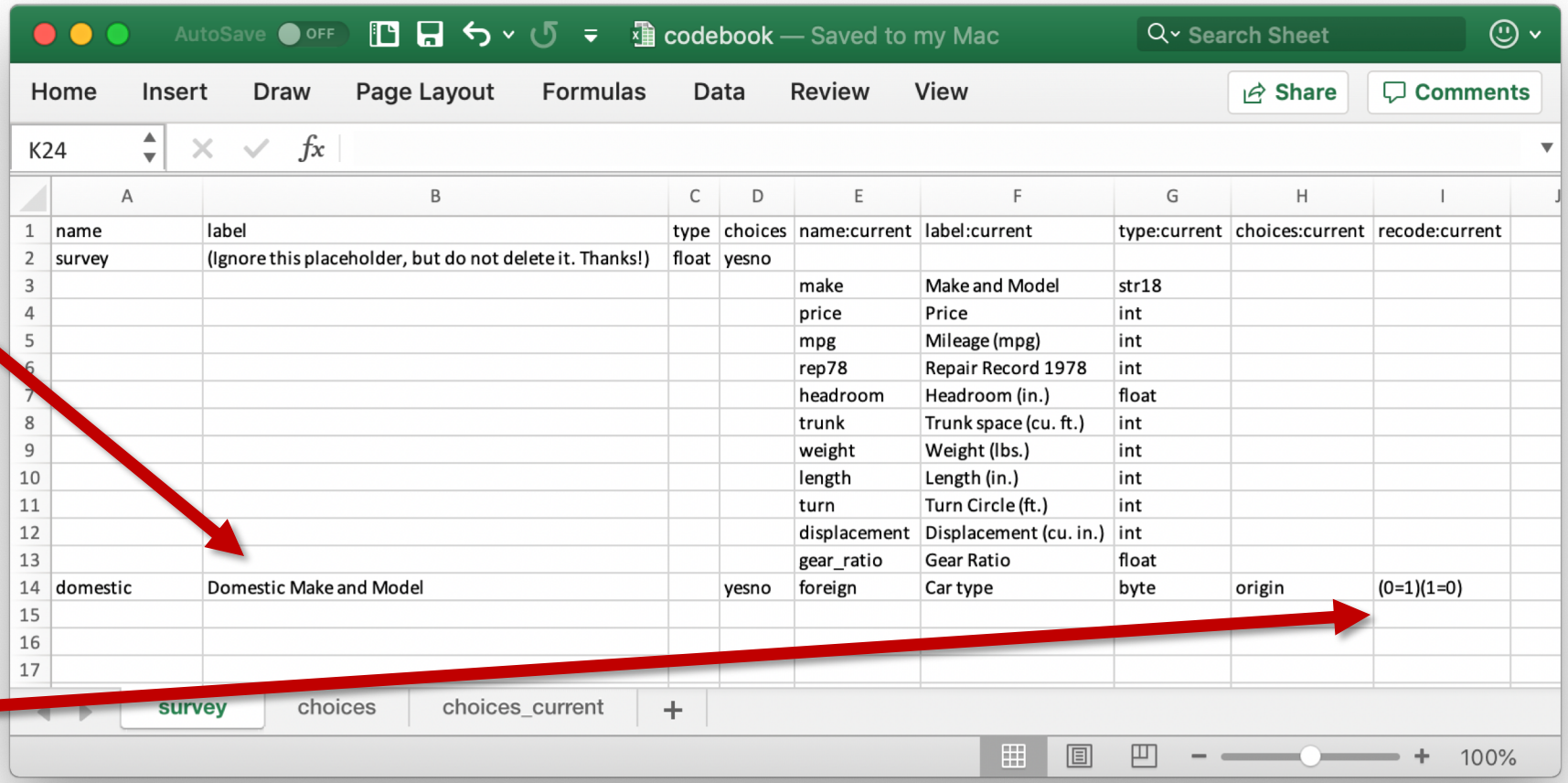


	A	B	C	D	E	F	G	H	I	J
1	name	label	type	choices	name:current	label:current	type:current	choices:current	recode:current	
2	survey	(Ignore this placeholder, but do not delete it. Thanks!)	float	yesno						
3					make	Make and Model	str18			
4					price	Price	int			
5					mpg	Mileage (mpg)	int			
6					rep78	Repair Record 1978	int			
7					headroom	Headroom (in.)	float			
8					trunk	Trunk space (cu. ft.)	int			
9					weight	Weight (lbs.)	int			
10					length	Length (in.)	int			
11					turn	Turn Circle (ft.)	int			
12					displacement	Displacement (cu. in.)	int			
13					gear_ratio	Gear Ratio	float			
14					foreign	Car type	byte	origin		
15										
16										
17										
18										
19										

Step 2: Each row cleans one variable

New variable name, label, value labels ("choices")

Recoding



	A	B	C	D	E	F	G	H	I
1	name	label	type	choices	name:current	label:current	type:current	choices:current	recode:current
2	survey	(Ignore this placeholder, but do not delete it. Thanks!)	float	yesno					
3					make	Make and Model	str18		
4					price	Price	int		
5					mpg	Mileage (mpg)	int		
6					rep78	Repair Record 1978	int		
7					headroom	Headroom (in.)	float		
8					trunk	Trunk space (cu. ft.)	int		
9					weight	Weight (lbs.)	int		
10					length	Length (in.)	int		
11					turn	Turn Circle (ft.)	int		
12					displacement	Displacement (cu. in.)	int		
13					gear_ratio	Gear Ratio	float		
14	domestic	Domestic Make and Model		yesno	foreign	Car type	byte	origin	(0=1)(1=0)
15									
16									
17									

Step 3: Apply Excel codebook to the data

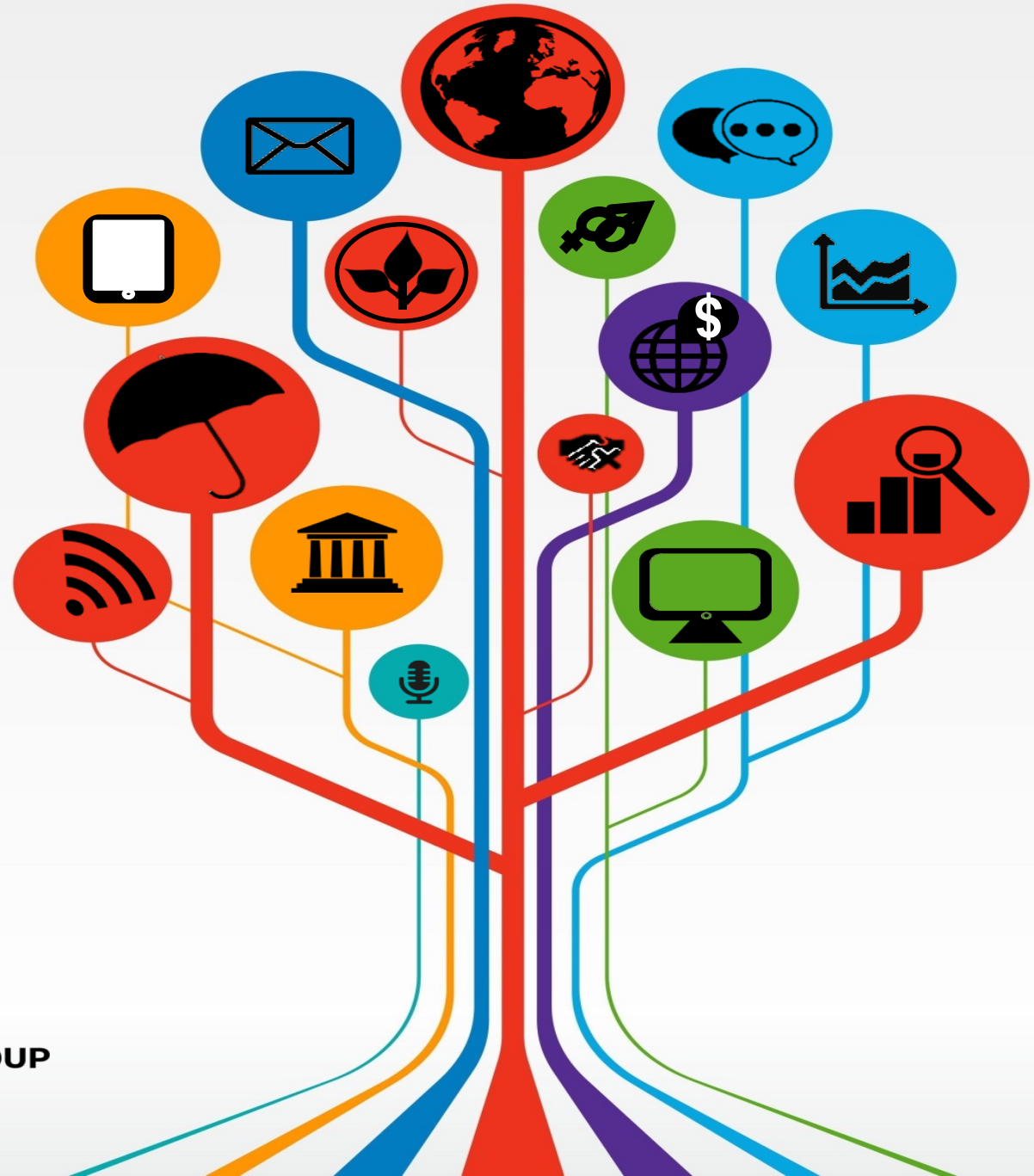
```
// Load data
sysuse auto.dta , clear


// Apply cleaning template
iecodebook apply
  using "codebook.xlsx"
```

Simply changing “template” to “apply” in the command gives the basic syntax.

The [*drop*] option removes all unused variables; the [*missingvalues*()] option lets you specify extended missing value codes for your whole dataset.

Dataset combination:
[*iecodebook append*]



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

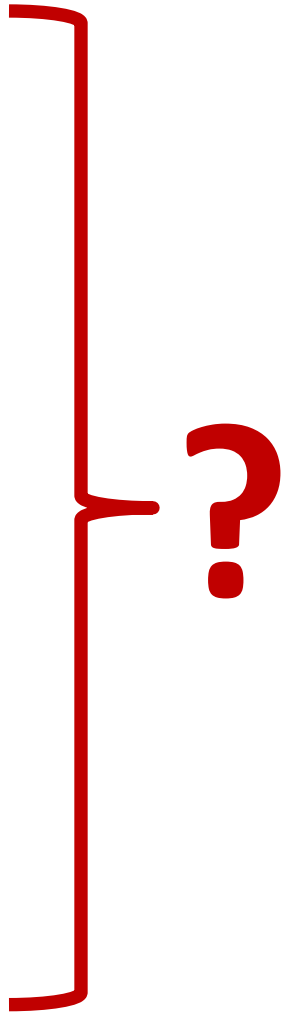
Use case: “Slightly” mismatched survey data

DIME Training Baseline

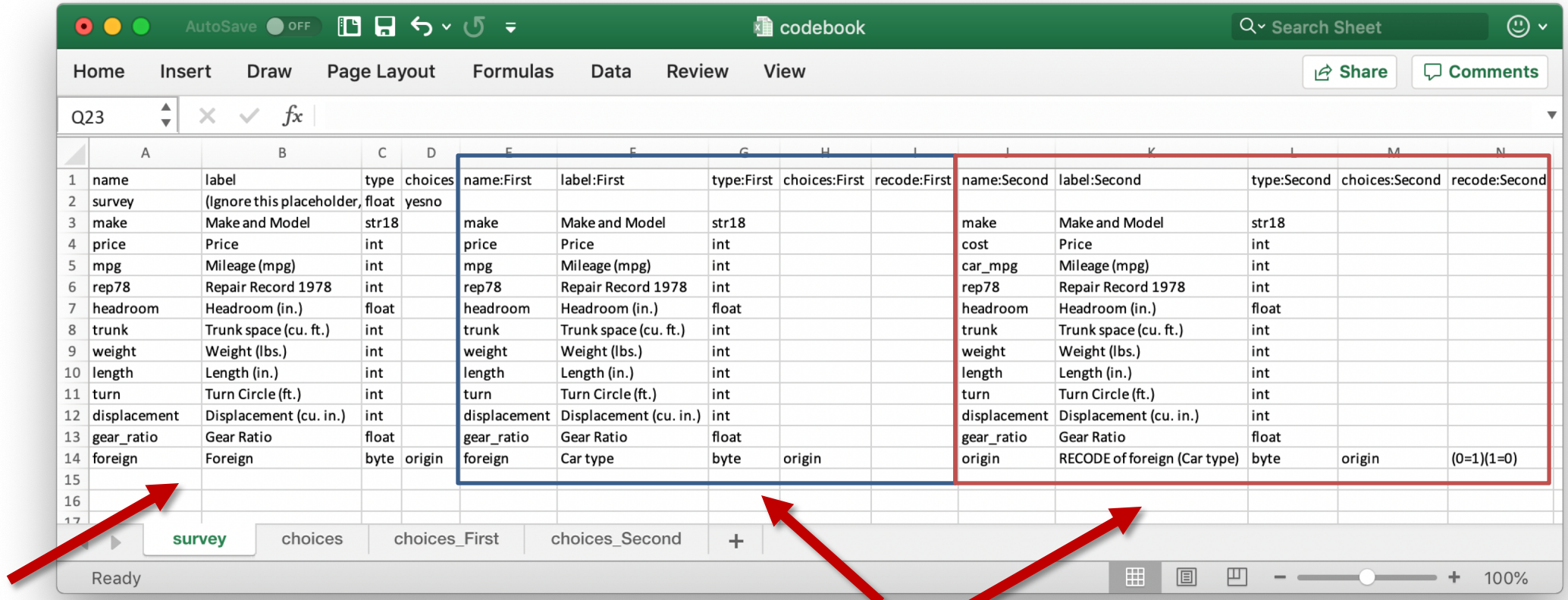
Field	Question	Answer
name *(required)*	What is your name?	
quest *(required)*	What is your quest?	
airspeed *(required)*	What is the average airspeed of an unladen swallow?	
color *(required)*	What is your favorite color?	1 Red 2 Blue 3 Green

DIME Training Endline

Field	Question	Answer
resp_name *(required)*	What is your name?	
resp_quest *(required)*	What quest are you on?	
resp_color *(required)*	Among the following, which is your favorite color:	1 White 2 Blue 3 Red 4 Green
swallow_speed *(required)*	What is the average airspeed of an unladen African swallow?	1 0-25 km/h 2 26-50 km/h 3 51-75 km/h 4 76-100 km/h



[*iecodebook append*] combines datasets



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	name	label	type	choices	name:First	label:First	type:First	choices:First	recode:First	name:Second	label:Second	type:Second	choices:Second	recode:Second
2	survey	(Ignore this placeholder,	float	yesno										
3	make	Make and Model	str18		make	Make and Model	str18			make	Make and Model	str18		
4	price	Price	int		price	Price	int			cost	Price	int		
5	mpg	Mileage (mpg)	int		mpg	Mileage (mpg)	int			car_mpg	Mileage (mpg)	int		
6	rep78	Repair Record 1978	int		rep78	Repair Record 1978	int			rep78	Repair Record 1978	int		
7	headroom	Headroom (in.)	float		headroom	Headroom (in.)	float			headroom	Headroom (in.)	float		
8	trunk	Trunk space (cu. ft.)	int		trunk	Trunk space (cu. ft.)	int			trunk	Trunk space (cu. ft.)	int		
9	weight	Weight (lbs.)	int		weight	Weight (lbs.)	int			weight	Weight (lbs.)	int		
10	length	Length (in.)	int		length	Length (in.)	int			length	Length (in.)	int		
11	turn	Turn Circle (ft.)	int		turn	Turn Circle (ft.)	int			turn	Turn Circle (ft.)	int		
12	displacement	Displacement (cu. in.)	int		displacement	Displacement (cu. in.)	int			displacement	Displacement (cu. in.)	int		
13	gear_ratio	Gear Ratio	float		gear_ratio	Gear Ratio	float			gear_ratio	Gear Ratio	float		
14	foreign	Foreign	byte	origin	foreign	Car type	byte	origin		origin	RECODE of foreign (Car type)	byte	origin	(0=1)(1=0)
15														
16														

Final variable name, label, value labels (“choices”)

Original variable info – think baseline/endline/etc

Step 0: Set up demo datasets

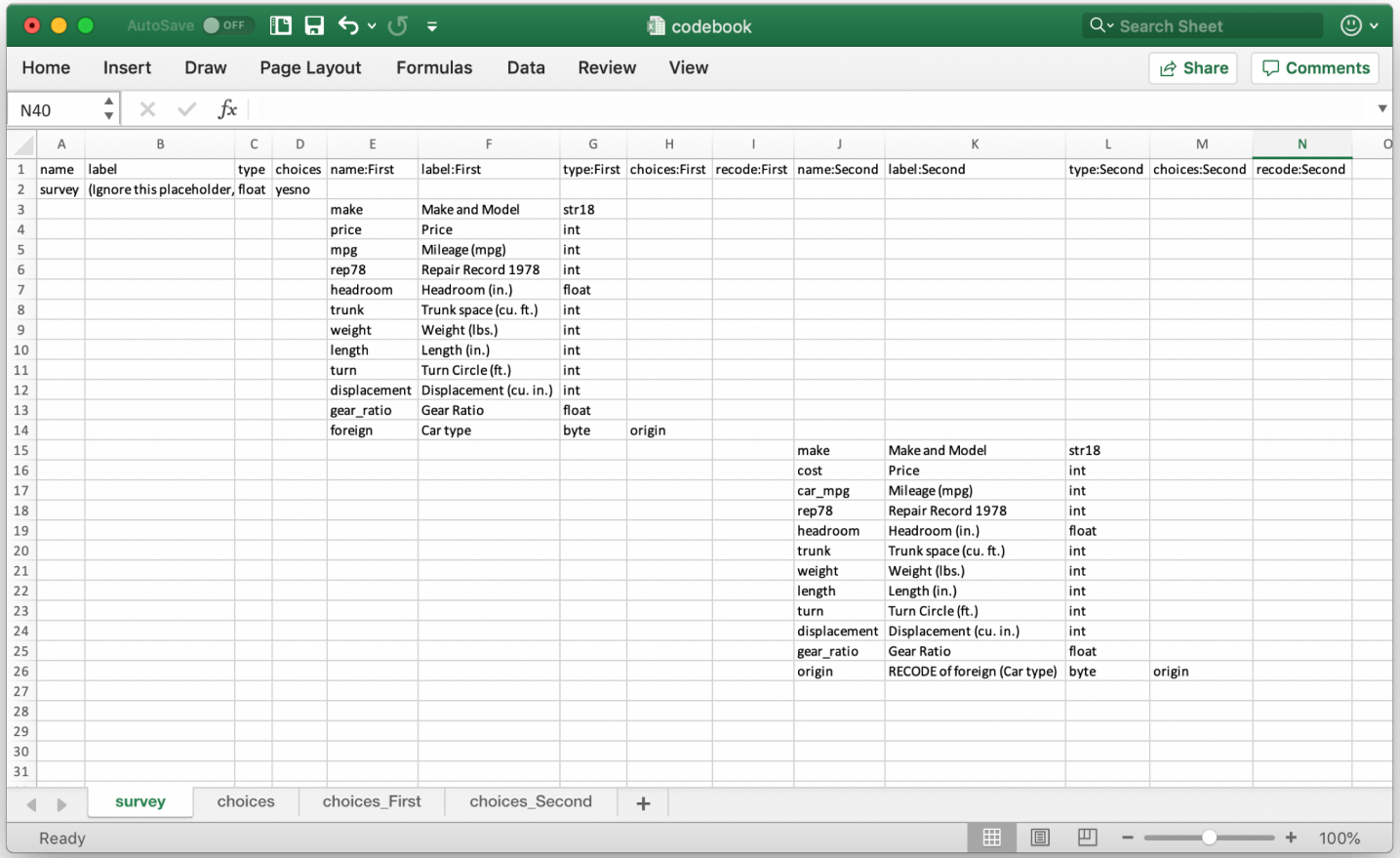
```
// Create demonstration datasets
sysuse auto.dta , clear
save data1.dta , replace

rename (price mpg)(cost car_mpg)
recode foreign ///
  (0=1 "Domestic") ///
  (1=0 "Foreign")  ///
  , gen(origin)
drop foreign
save data2.dta , replace
```

Step 1: Set up Excel codebook template

```
// Create codebook template
iecodebook template       ///
 "data1.dta" "data2.dta" ///
 using "codebook.xlsx"    ///
, surveys(First Second)
```

"Template" command sets up this entire sheet with the current state of all the datasets (two or more).



codebook

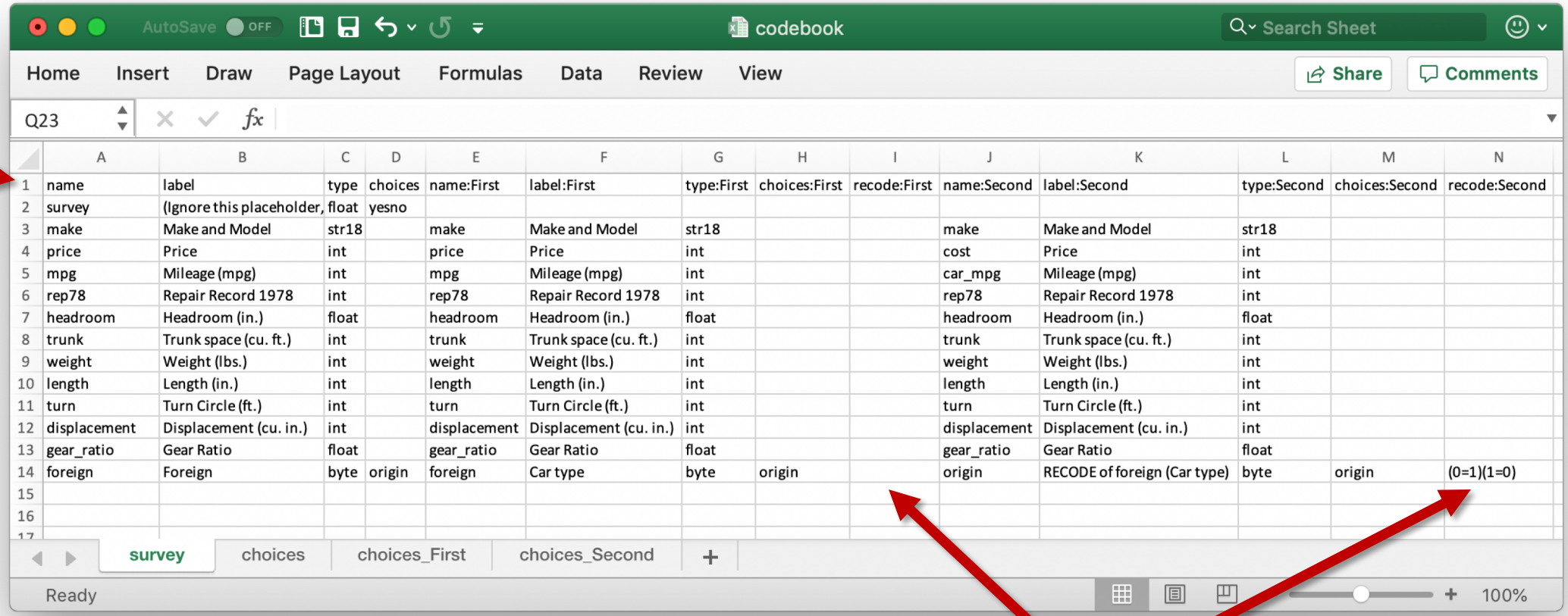
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	name	label	type	choices	name:First	label:First	type:First	choices:First	recode:First	name:Second	label:Second	type:Second	choices:Second	recode:Second
2	survey	(Ignore this placeholder,	float	yesno										
3					make	Make and Model	str18							
4					price	Price	int							
5					mpg	Mileage (mpg)	int							
6					rep78	Repair Record 1978	int							
7					headroom	Headroom (in.)	float							
8					trunk	Trunk space (cu. ft.)	int							
9					weight	Weight (lbs.)	int							
10					length	Length (in.)	int							
11					turn	Turn Circle (ft.)	int							
12					displacement	Displacement (cu. in.)	int							
13					gear_ratio	Gear Ratio	float							
14					foreign	Car type	byte	origin						
15										make	Make and Model	str18		
16										cost	Price	int		
17										car_mpg	Mileage (mpg)	int		
18										rep78	Repair Record 1978	int		
19										headroom	Headroom (in.)	float		
20										trunk	Trunk space (cu. ft.)	int		
21										weight	Weight (lbs.)	int		
22										length	Length (in.)	int		
23										turn	Turn Circle (ft.)	int		
24										displacement	Displacement (cu. in.)	int		
25										gear_ratio	Gear Ratio	float		
26										origin	RECODE of foreign (Car type)	byte	origin	
27														
28														
29														
30														
31														

survey | choices | choices_First | choices_Second

Ready 100%

Step 2: Each row harmonizes one variable

Final variable name, label, value labels ("choices")



codebook

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	name	label	type	choices	name:First	label:First	type:First	choices:First	recode:First	name:Second	label:Second	type:Second	choices:Second	recode:Second
2	survey	(Ignore this placeholder,	float	yesno										
3	make	Make and Model	str18		make	Make and Model	str18			make	Make and Model	str18		
4	price	Price	int		price	Price	int			cost	Price	int		
5	mpg	Mileage (mpg)	int		mpg	Mileage (mpg)	int			car_mpg	Mileage (mpg)	int		
6	rep78	Repair Record 1978	int		rep78	Repair Record 1978	int			rep78	Repair Record 1978	int		
7	headroom	Headroom (in.)	float		headroom	Headroom (in.)	float			headroom	Headroom (in.)	float		
8	trunk	Trunk space (cu. ft.)	int		trunk	Trunk space (cu. ft.)	int			trunk	Trunk space (cu. ft.)	int		
9	weight	Weight (lbs.)	int		weight	Weight (lbs.)	int			weight	Weight (lbs.)	int		
10	length	Length (in.)	int		length	Length (in.)	int			length	Length (in.)	int		
11	turn	Turn Circle (ft.)	int		turn	Turn Circle (ft.)	int			turn	Turn Circle (ft.)	int		
12	displacement	Displacement (cu. in.)	int		displacement	Displacement (cu. in.)	int			displacement	Displacement (cu. in.)	int		
13	gear_ratio	Gear Ratio	float		gear_ratio	Gear Ratio	float			gear_ratio	Gear Ratio	float		
14	foreign	Foreign	byte	origin	foreign	Car type	byte	origin		origin	RECODE of foreign (Car type)	byte	origin	(0=1)(1=0)
15														
16														

Recoding so that value codes match up

Step 3: Apply Excel codebook to the data

```
// Append the datasets
iecodebook append ///
 "data1.dta" "data2.dta" ///
 using "codebook.xlsx" ///
, surveys(First Second)
```

Simply changing “template” to “append” in the command gives the basic syntax.

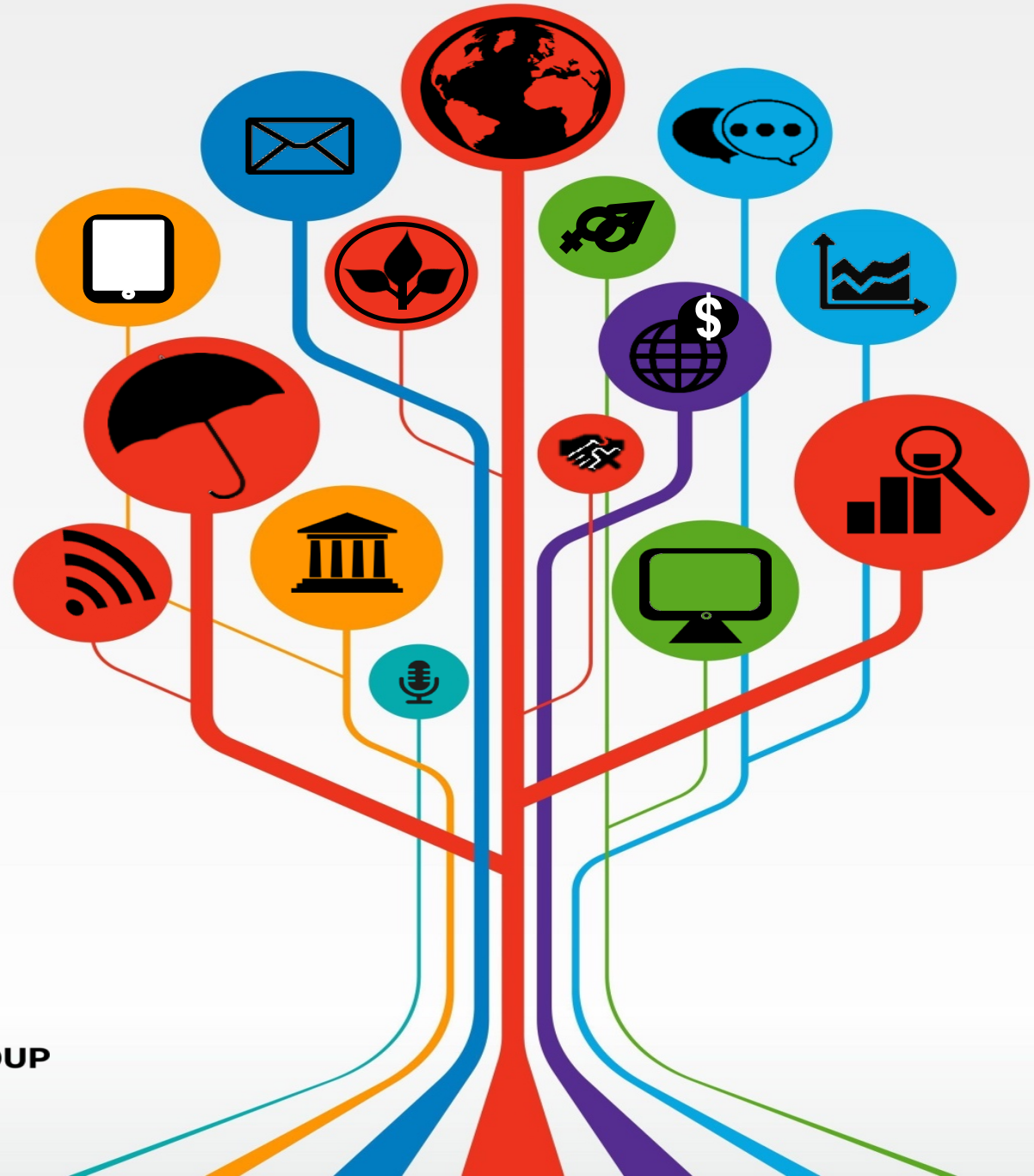
The [*drop*] option is on by default and removes all unused variables; [*nodrop*] cancels this.

The [*missingvalues*()] option lets you specify extended missing value codes for your whole dataset.

```
. ta survey foreign
```

Data Source	Foreign: Domestic	Foreign: Foreign	Total
First	52	22	74
Second	52	22	74
Total	104	44	148

Data documentation:
[*iecodebook export*]

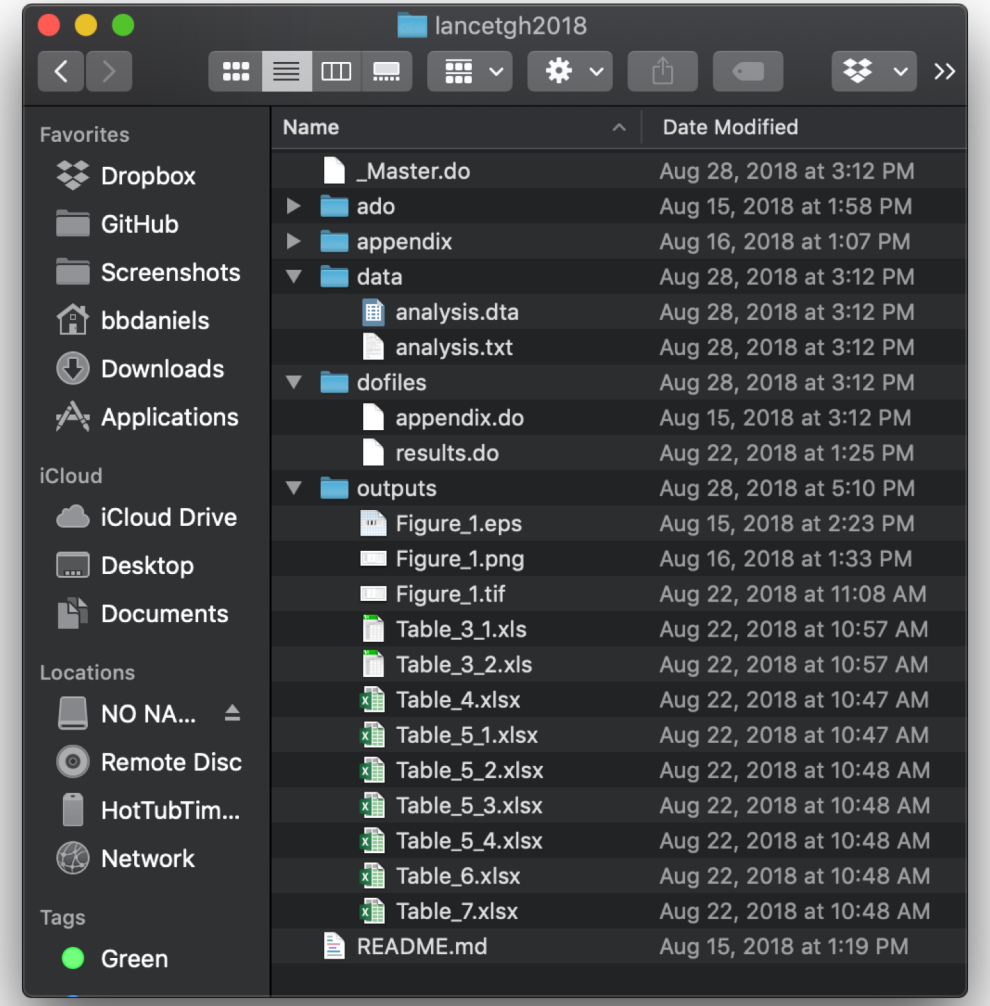


i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Code review and open access needs data

- You don’t want to share your whole constructed dataset with the world when you publish a working paper on a small part
- It’s unwieldy for reproduction
- You have other (unfinished) stuff in there you’re working on
- Solution: subset your analysis dataset for public release



[*iecodebook export*] builds public release data

- Function 1: Just create the codebook for documentation

- Function 2: [*trim*()] dataset:
  - Reads your dofiles
  - Keeps only the variables that are used in analysis
  - Creates a minimal codebook

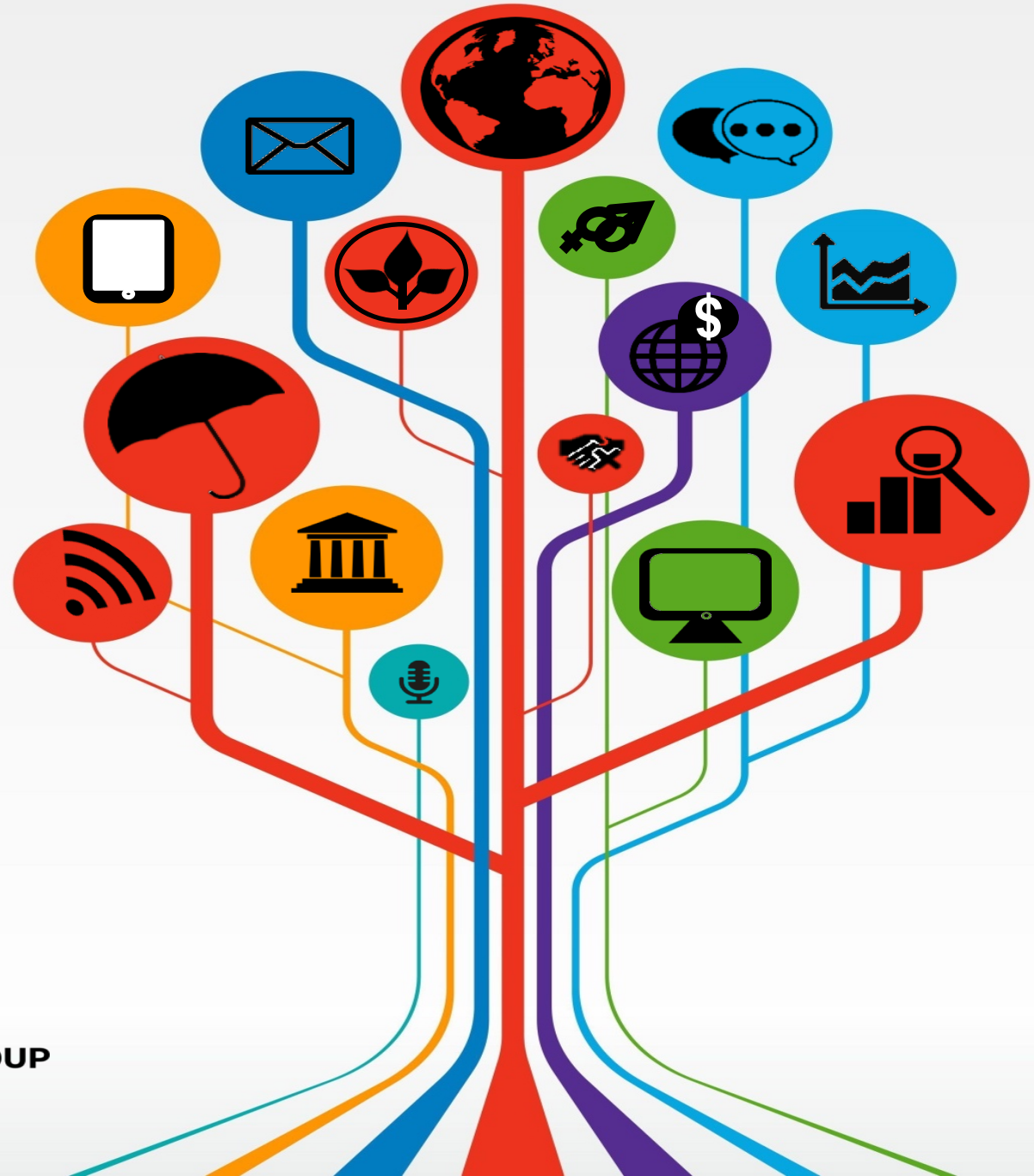
  - Rewards good syntax – you *must*:
    - Spell variable names completely
    - [*set varabbrev off*] ← part of [*ieboilstart*]
    - **No** wildcards or lists: * ? -

```
iecodebook export
 [if] [in] using
 "/path/to/codebook.xlsx"
,
 [trim("/path/dofile1.do"
 ["/path/dofile2.do"]...)]
```

Thank you!

worldbank.github.com/dimeanalytics

github.com/worldbank/iefieldkit



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP