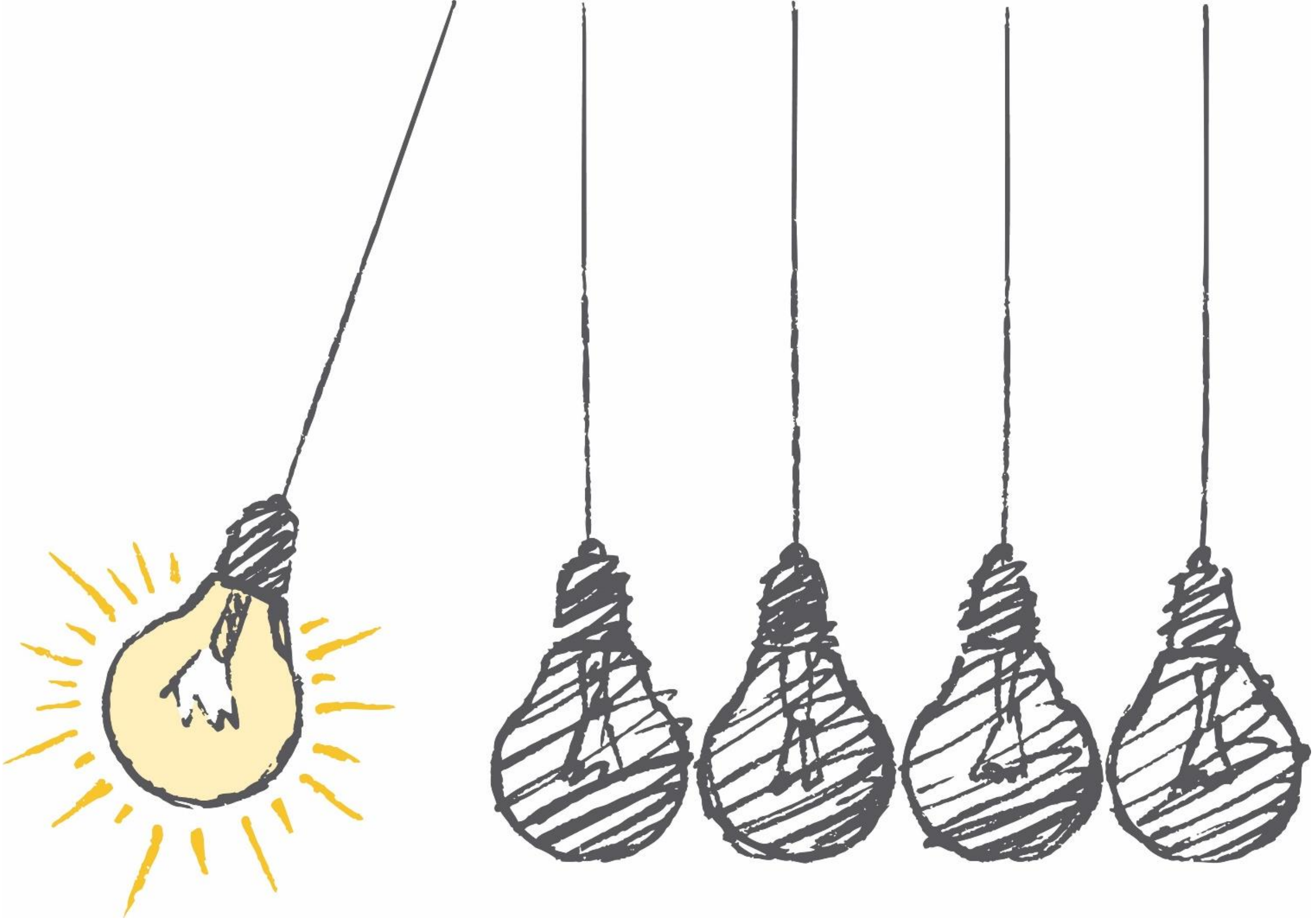
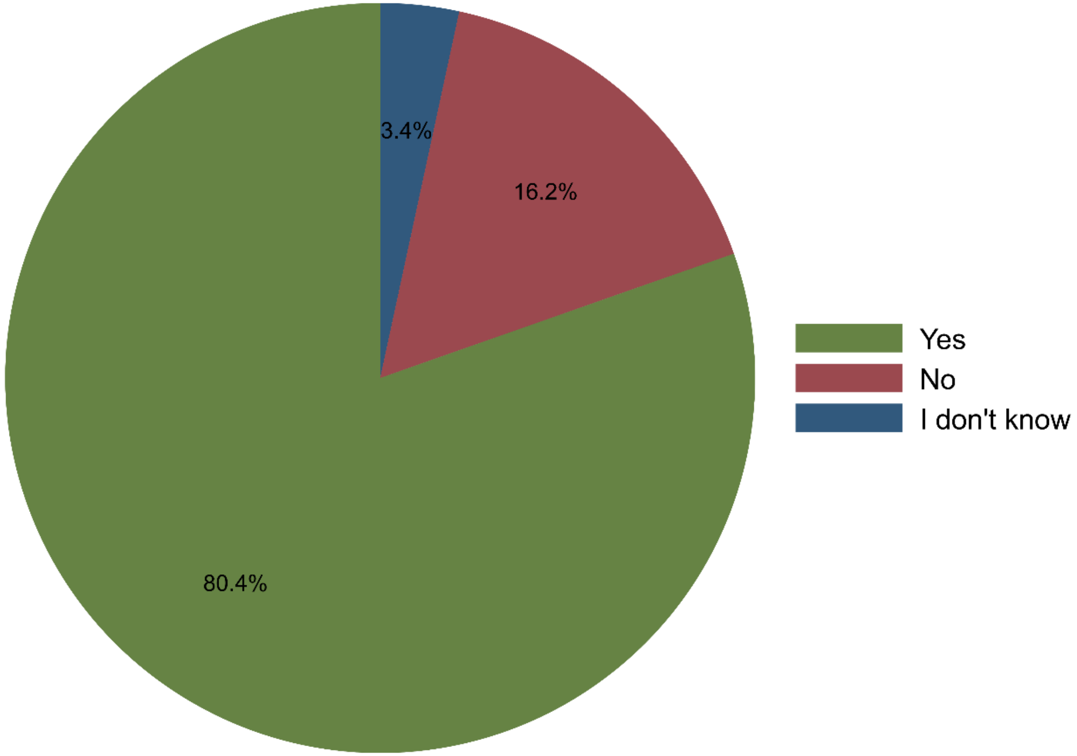
Research ethics:
Data privacy

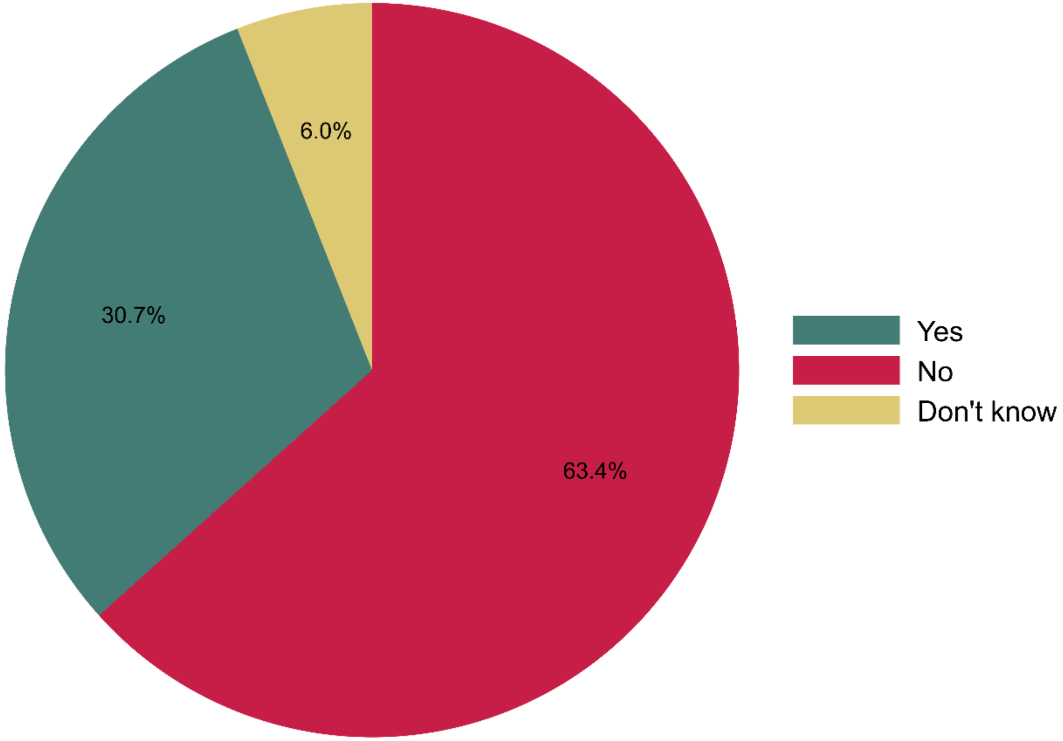


Prepared by **DIME Analytics**

dimeanalytics@worldbank.org

How are we doing now?





Note: Sample of 49 projects that collected sensitive or identifiable data.

Project collects PII data

Analysis data is de-identified

What we’re doing today

1. Quick review: what is PII
2. When to de-identify: workflow
3. How to de-identify: tips and tools
4. Re-identification: weighing disclosure risks

What is PII?

What is “PII” anyway?

Any information that can be used to link survey data with respondents

Direct identifiers

- Name (respondent, household roster, social network, etc)
- Street address, geocoordinates (household, plot, etc)
- Telephone number
- Face photos
- Unique account numbers (national ID, bank account, health insurance)

Implicit or quasi-identifiers (a.k.a. key variables)

- Location + DOB
- Location + outliers

→ 63% of US population uniquely identified by gender + DOB + zip code!

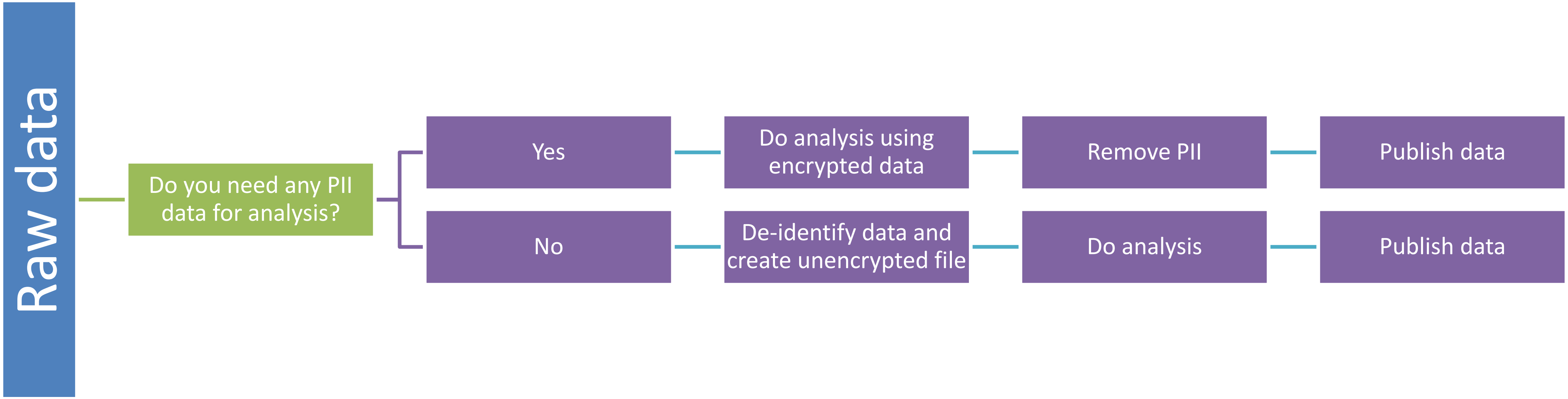
What is “PII” anyway?

- Does my dataset have PII?
  - Almost certainly!

- What’s required to have access to PII?
  - Must be specified in project IRB (PIs + Research Assistants)
  - Must have Human Subjects Research certification

When to de-identify

When should I de-identify my data?



Raw data

Do you need any PII data for analysis?

Yes

Do analysis using encrypted data

Remove PII

Publish data

No

De-identify data and create unencrypted file

Do analysis

Publish data

When should I de-identify my data?

- The earlier the better. Why?
  - – Identified data should always be encrypted.
  - – It is easier to work with unencrypted data

- Data **must be** fully de-identified before data can be published to microdata catalogue
  - – Best to be conservative (remove any potential identifiers)
  - – If PII required for analysis, offer restricted access for replication only
    - Analytics currently working with DECDG to create infrastructure for this

When should I de-identify my data?

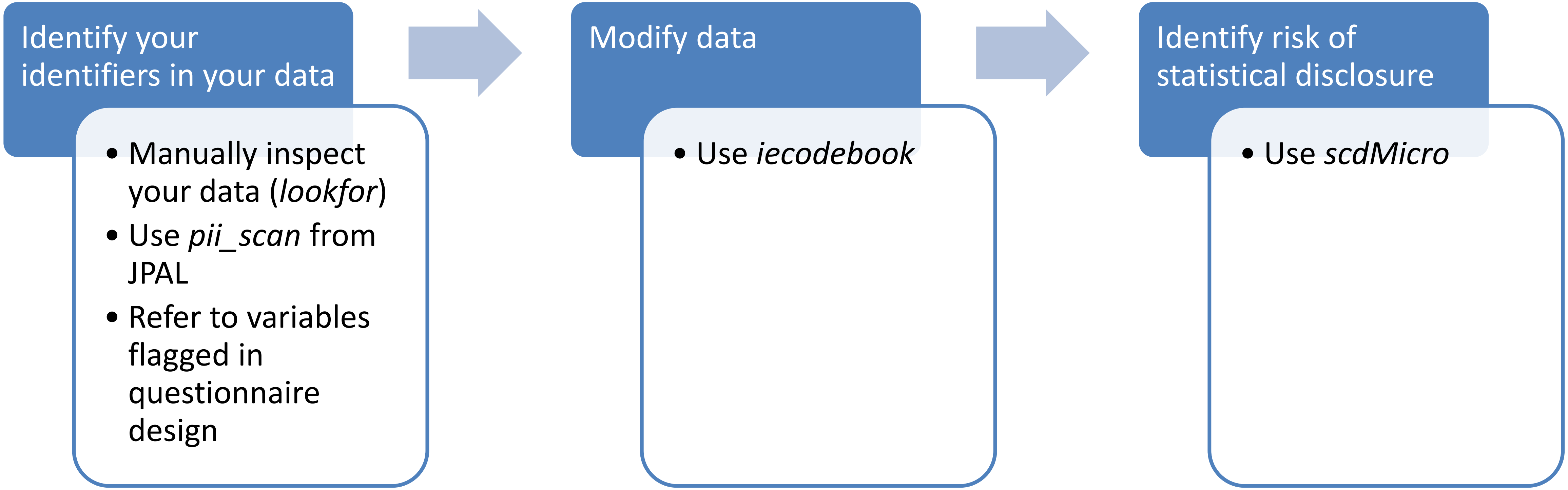
- Best practice: flag all PII questions when designing the survey
  - Some survey software make this straightforward (e.g. publishable field in SurveyCTO)

- Flagging PII questions at design phase saves time later

- Opportunity to ask: is this really needed?
  - If you don’t collect PII, you don’t need to worry about handling it!

How to de-identify

How should I de-identify my data?



Identify your identifiers in your data

- Manually inspect your data (*lookfor*)
- Use *pii_scan* from JPAL
- Refer to variables flagged in questionnaire design

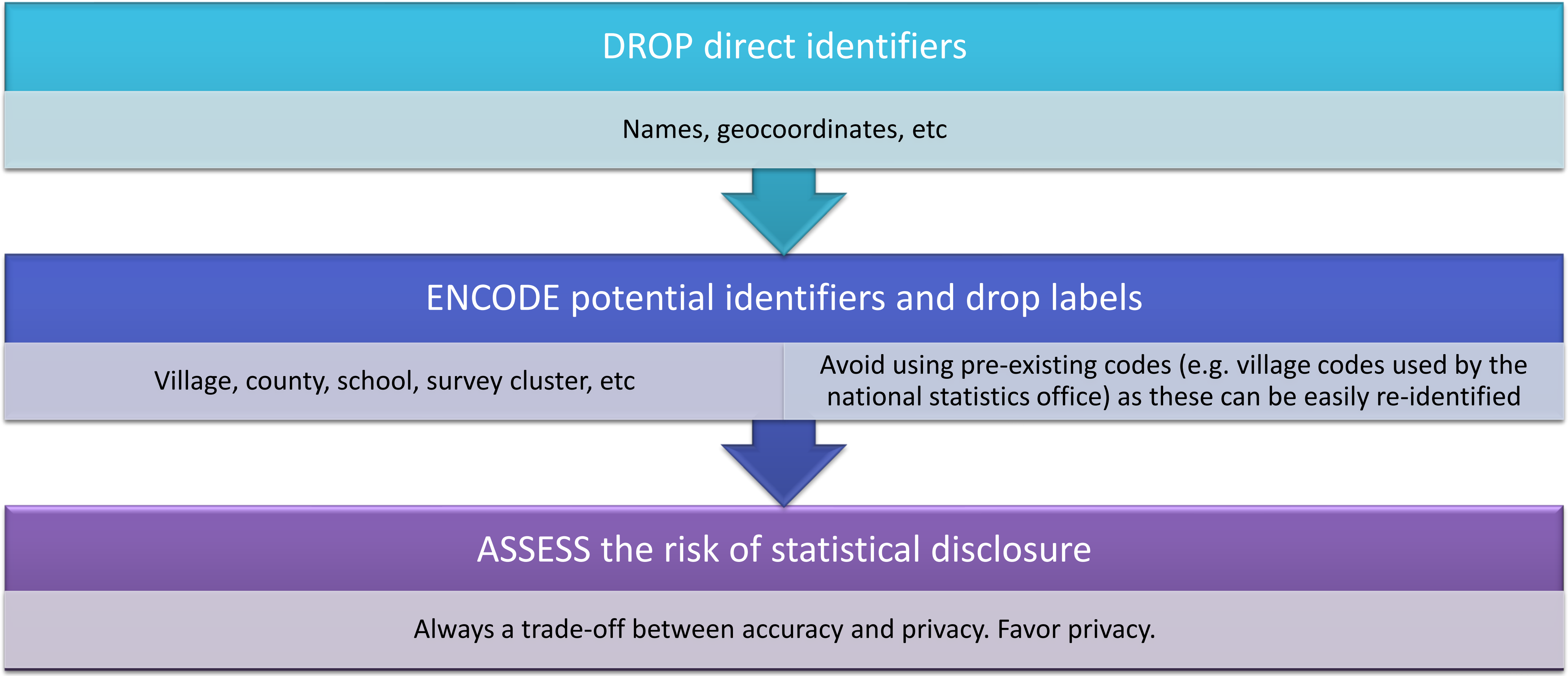
Modify data

- Use *iecodebook*

Identify risk of statistical disclosure

- Use *scdMicro*

How should I de-identify my data?



DROP direct identifiers

Names, geocoordinates, etc

ENCODE potential identifiers and drop labels

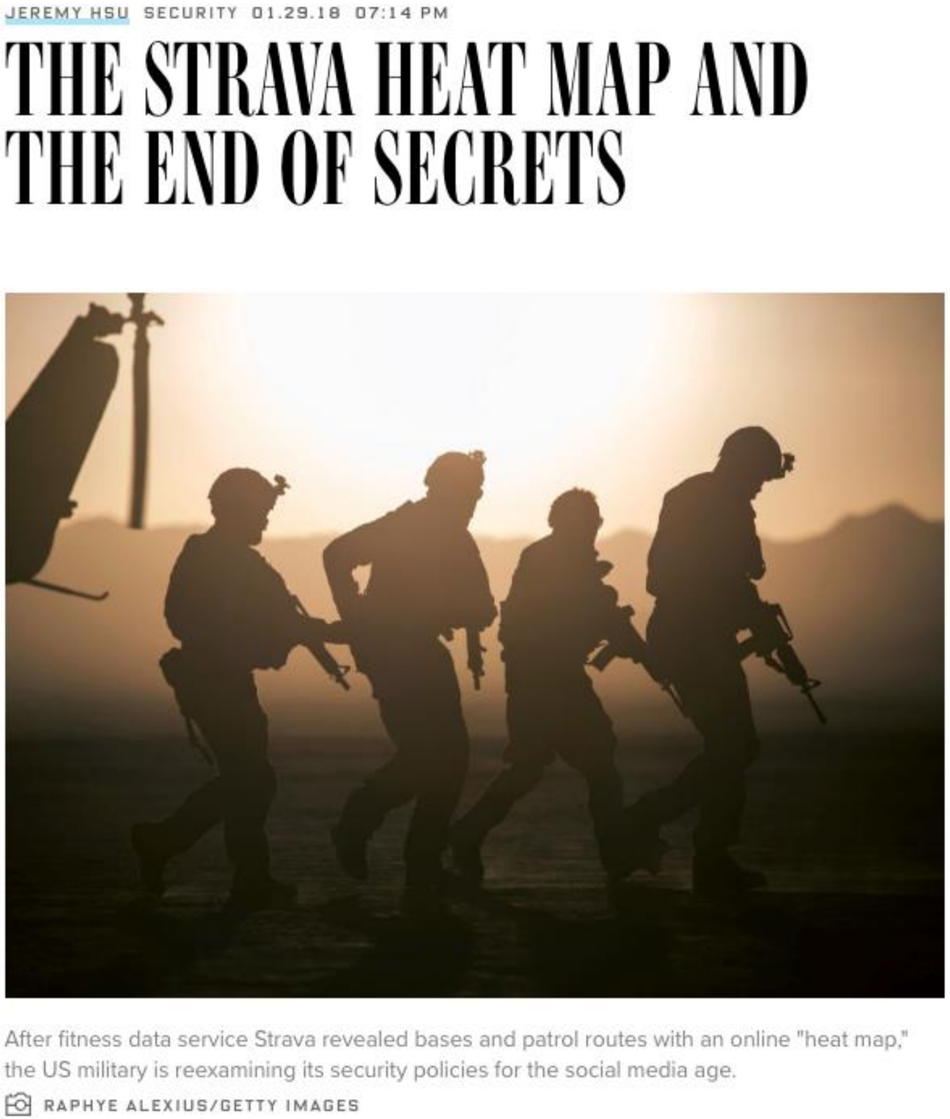
Village, county, school, survey cluster, etc

Avoid using pre-existing codes (e.g. village codes used by the national statistics office) as these can be easily re-identified

ASSESS the risk of statistical disclosure

Always a trade-off between accuracy and privacy. Favor privacy.

Risks of re-identification

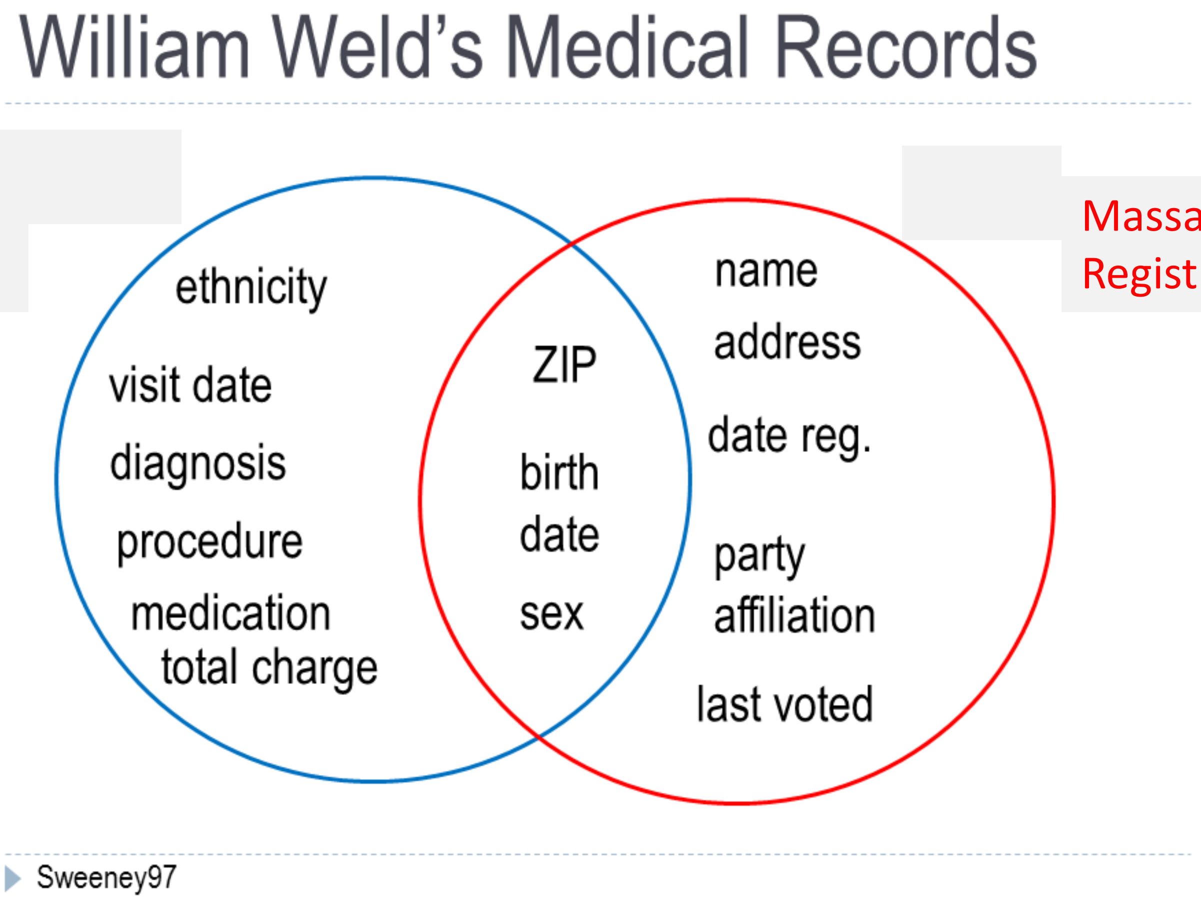


JEREMY HSU SECURITY 01.29.18 07:14 PM

THE STRAVA HEAT MAP AND THE END OF SECRETS

After fitness data service Strava revealed bases and patrol routes with an online "heat map," the US military is reexamining its security policies for the social media age.

RAPHYE ALEXIUS/GETTY IMAGES



William Weld’s Medical Records

Sweeney97

Massachusetts Group Insurance Commission data

Massachusetts Voter Registration data

Disclosure risk

- Disclosure risk = risk that data could be re-identified

- Trade-off between disclosure risk and information loss

- Consider: how difficult it would be to re-identify your data **and** the level of harm that could cause

Disclosure risk

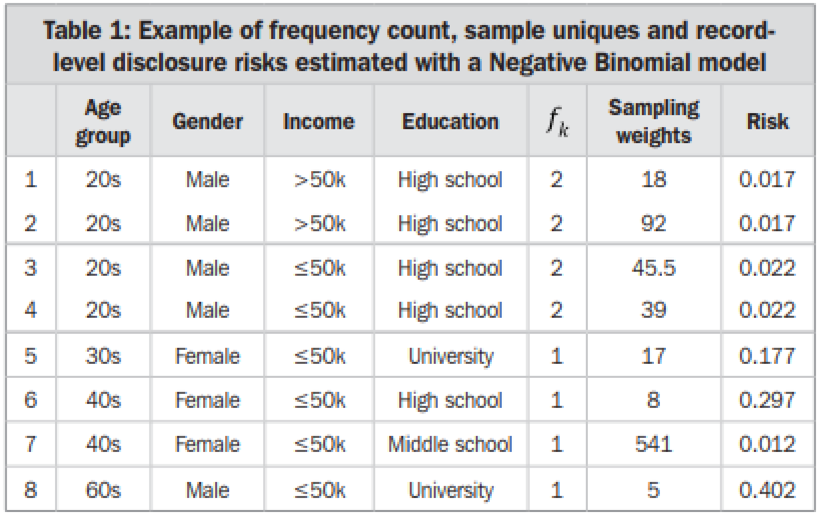


Table 1: Example of frequency count, sample uniques and record-level disclosure risks estimated with a Negative Binomial model

	Age group	Gender	Income	Education	f_k	Sampling weights	Risk
1	20s	Male	>50k	High school	2	18	0.017
2	20s	Male	>50k	High school	2	92	0.017
3	20s	Male	≤50k	High school	2	45.5	0.022
4	20s	Male	≤50k	High school	2	39	0.022
5	30s	Female	≤50k	University	1	17	0.177
6	40s	Female	≤50k	High school	1	8	0.297
7	40s	Female	≤50k	Middle school	1	541	0.012
8	60s	Male	≤50k	University	1	5	0.402

- 4 pre-determined key variables
  - age, gender, income, education
- 6 distinct patterns
- *k*-anonymity
  - Ensuring each pattern has at least *k* records in the sample
  - Rule of thumb: *k≥3*
- For (lots!) more details Statistical Disclosure Control for Microdata: A Theory Guide.

Disclosure risk

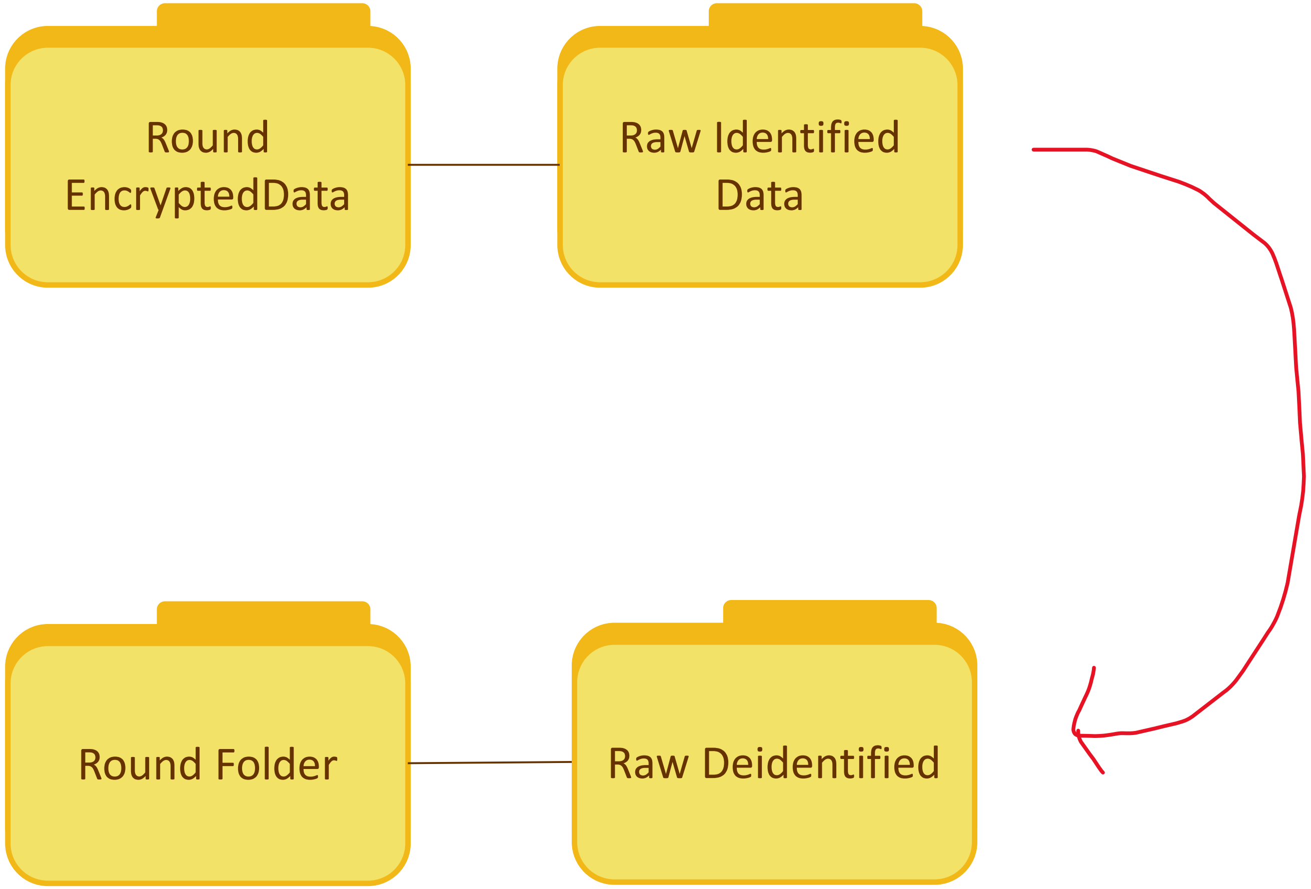
- If potential for harm is high, projects with highly sensitive data may need advanced methods of statistical disclosure control
  - Sensitive data: illegal activity, political activity, voting behavior, medical conditions, financial records

- Best option is differential privacy
  - but no consensus on how to implement this for economics

- Analytics can advise on project-specific basis

Thank you!

De-identifying data with iecodebook

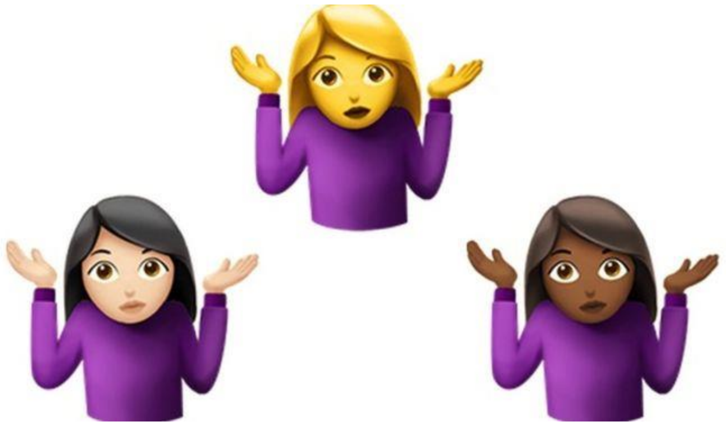


Round EncryptedData

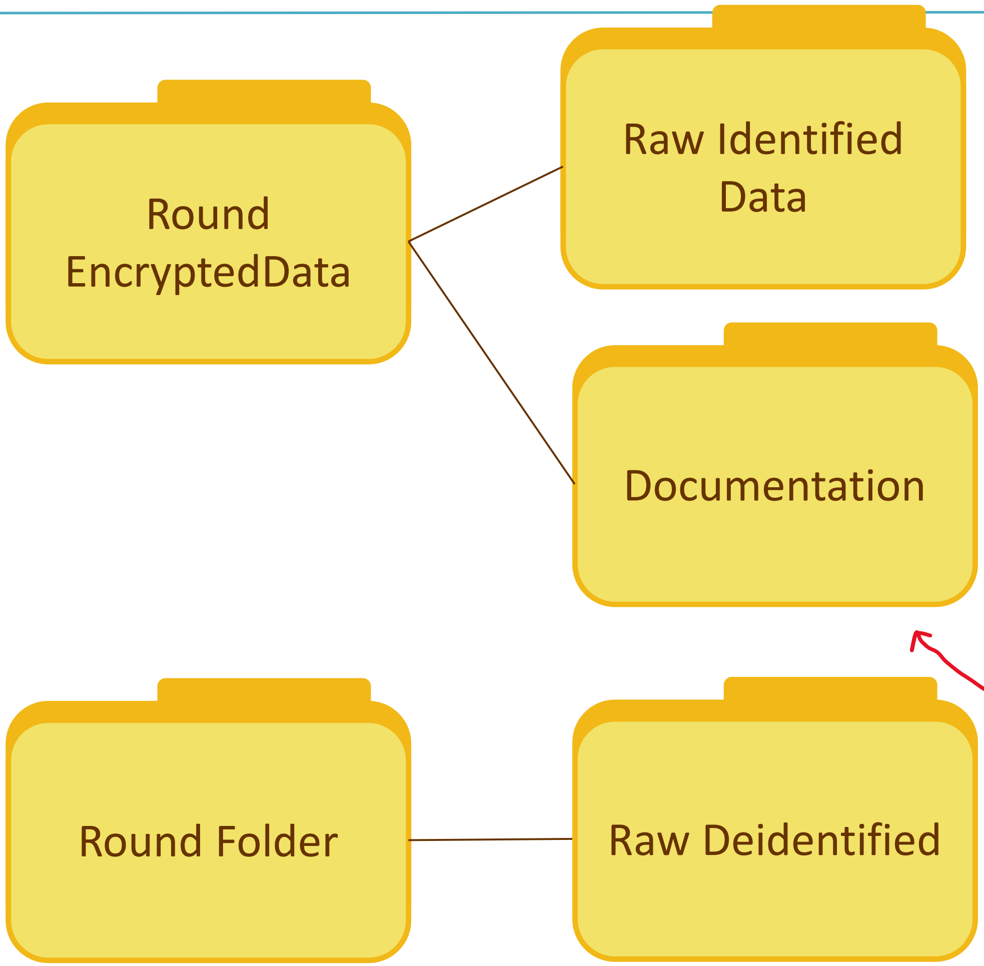
Raw Identified Data

Round Folder

Raw Deidentified



De-identifying data with iecodebook



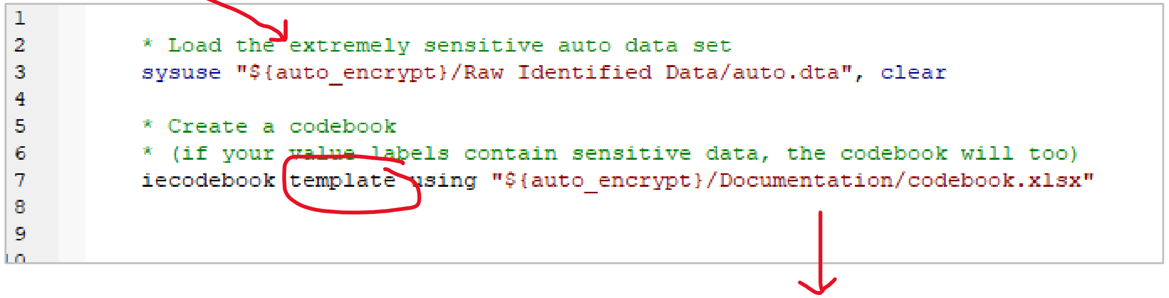
Round EncryptedData

Raw Identified Data

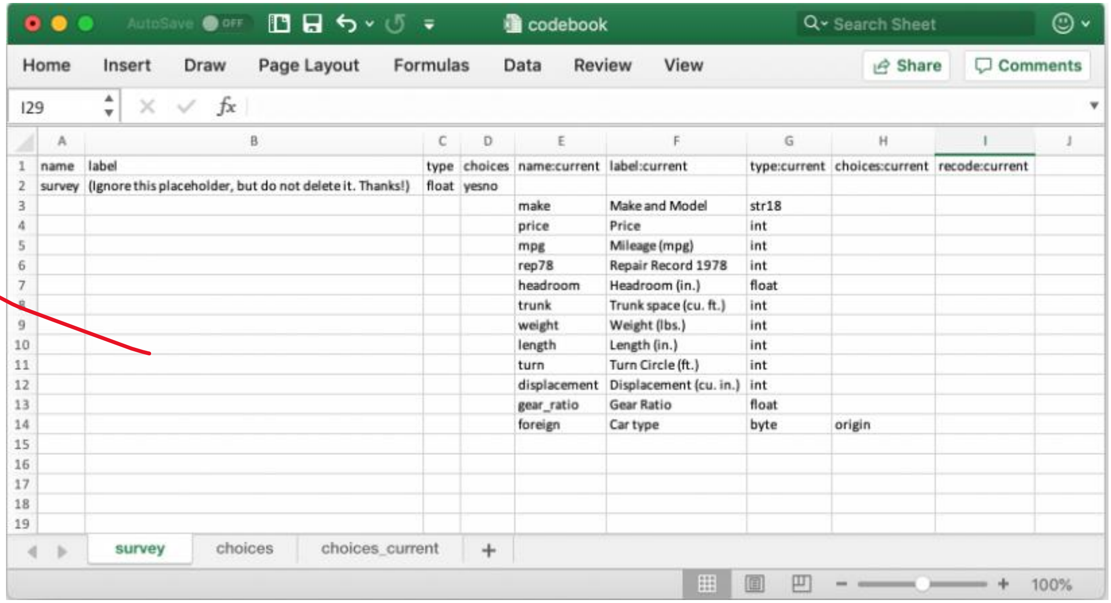
Documentation

Round Folder

Raw Deidentified

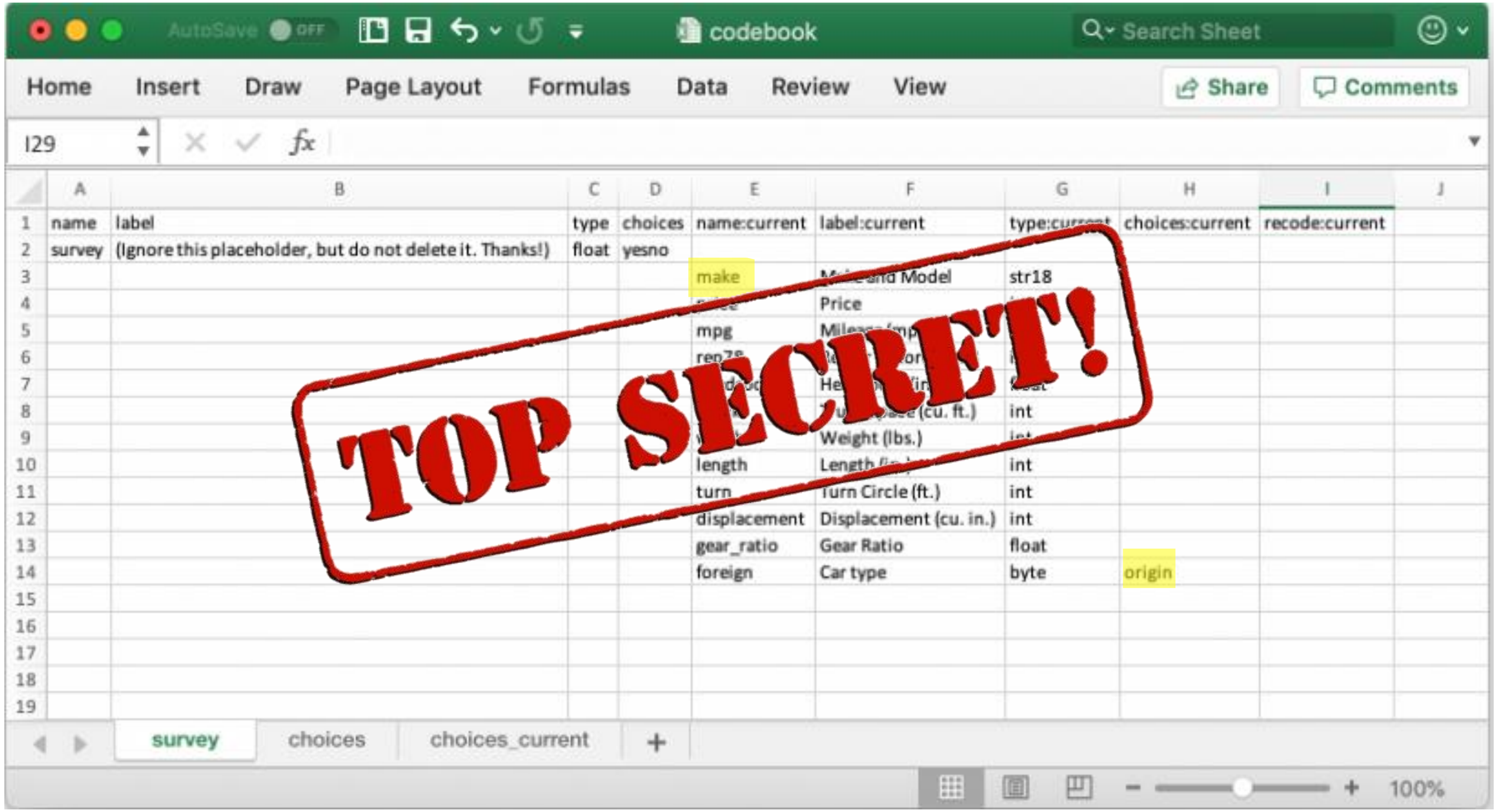


```
    * Load the extremely sensitive auto data set
    sysuse "${auto_encrypt}/Raw Identified Data/auto.dta", clear

    * Create a codebook
    * (if your value labels contain sensitive data, the codebook will too)
    iecodebook template using "${auto_encrypt}/Documentation/codebook.xlsx"
```



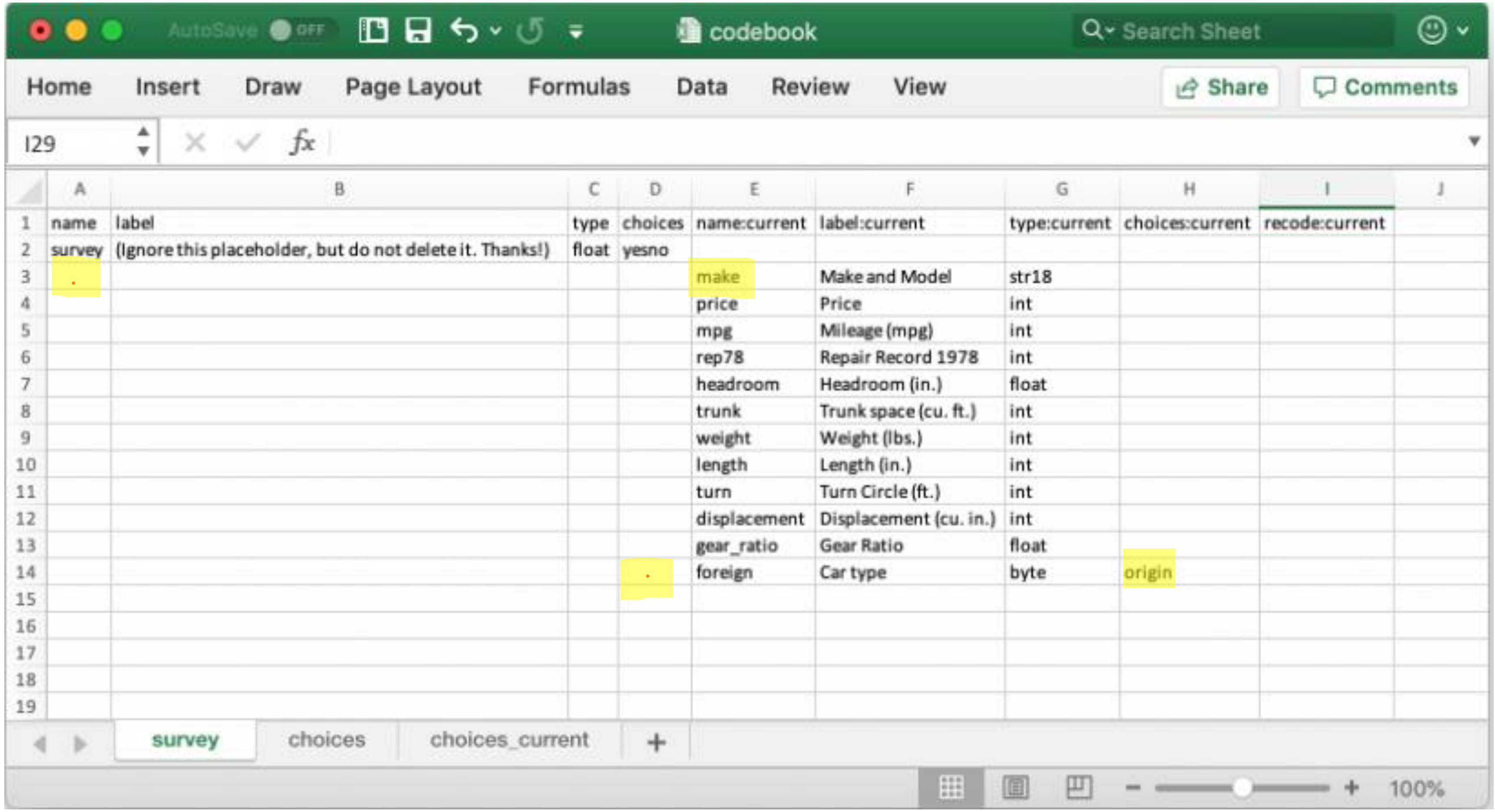
	A	B	C	D	E	F	G	H	I	J
1	name	label	type	choices	name:current	label:current	type:current	choices:current	recode:current	
2	survey	(Ignore this placeholder, but do not delete it. Thanks!)	float	yesno						
3					make	Make and Model	str18			
4					price	Price	int			
5					mpg	Mileage (mpg)	int			
6					rep78	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										

De-identifying data with iecodebook



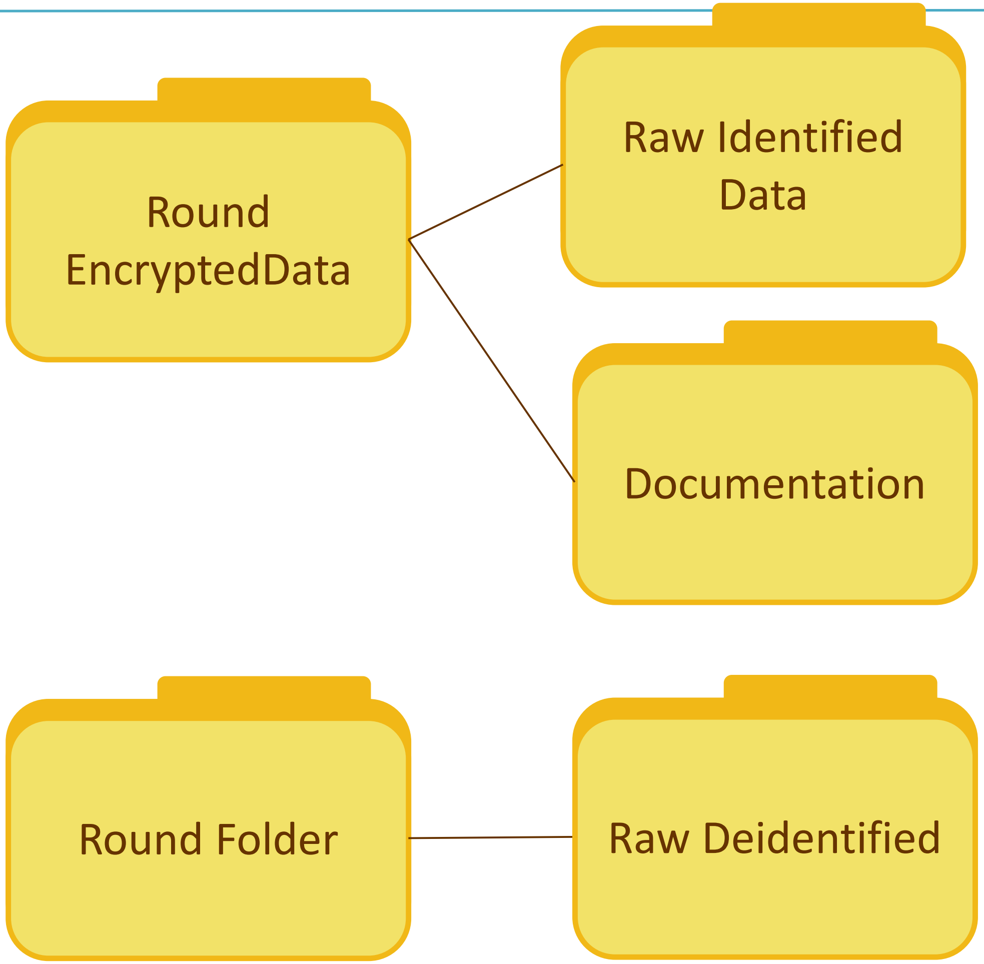
	A	B	C	D	E	F	G	H	I
1	name	label	type	choices	name:current	label:current	type:current	choices:current	recode:current
2	survey	(Ignore this placeholder, but do not delete it. Thanks!)	float	yesno					
3					make	Make and Model	str18		
4					price	Price	int		
5					mpg	Mileage (mpg)	[illegible]		
6					rep78	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	

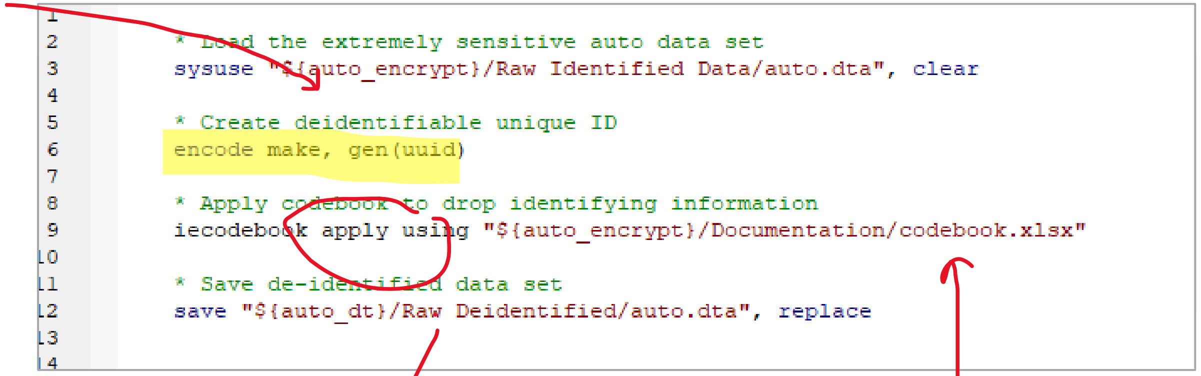
De-identifying data with iecodebook



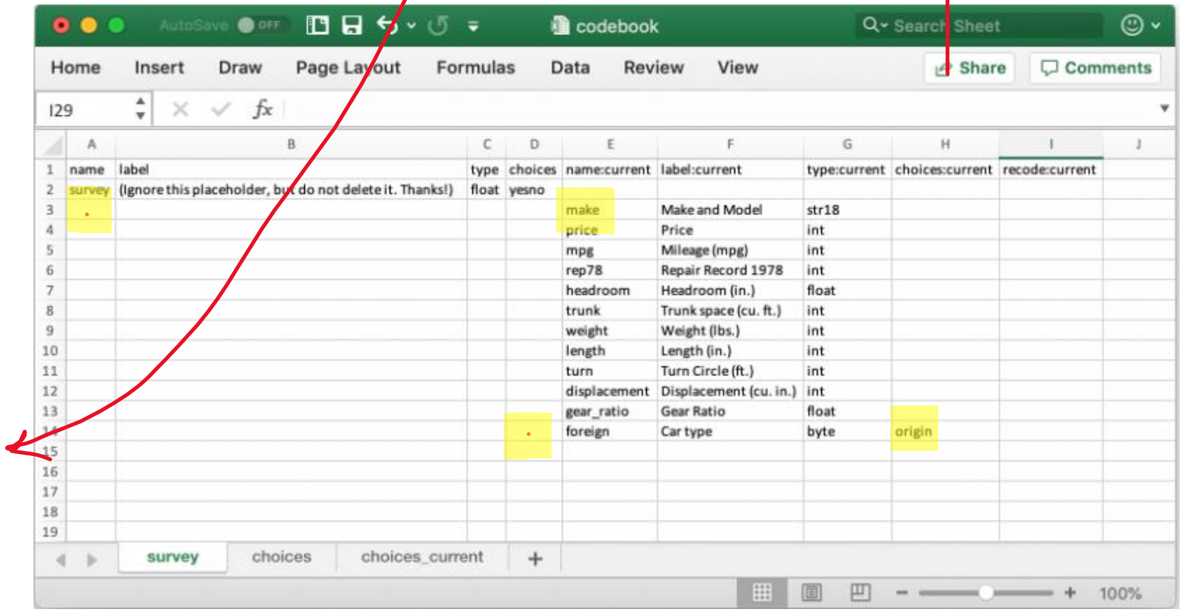
	A	B	C	D	E	F	G	H	I	J
1	name	label	type	choices	name:current	label:current	type:current	choices:current	recode:current	
2	survey	(Ignore this placeholder, but do not delete it. Thanks!)	float	yesno						
3	.				make	Make and Model	str18			
4					price	Price	int			
5					mpg	Mileage (mpg)	int			
6					rep78	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										

De-identifying data with iecodebook





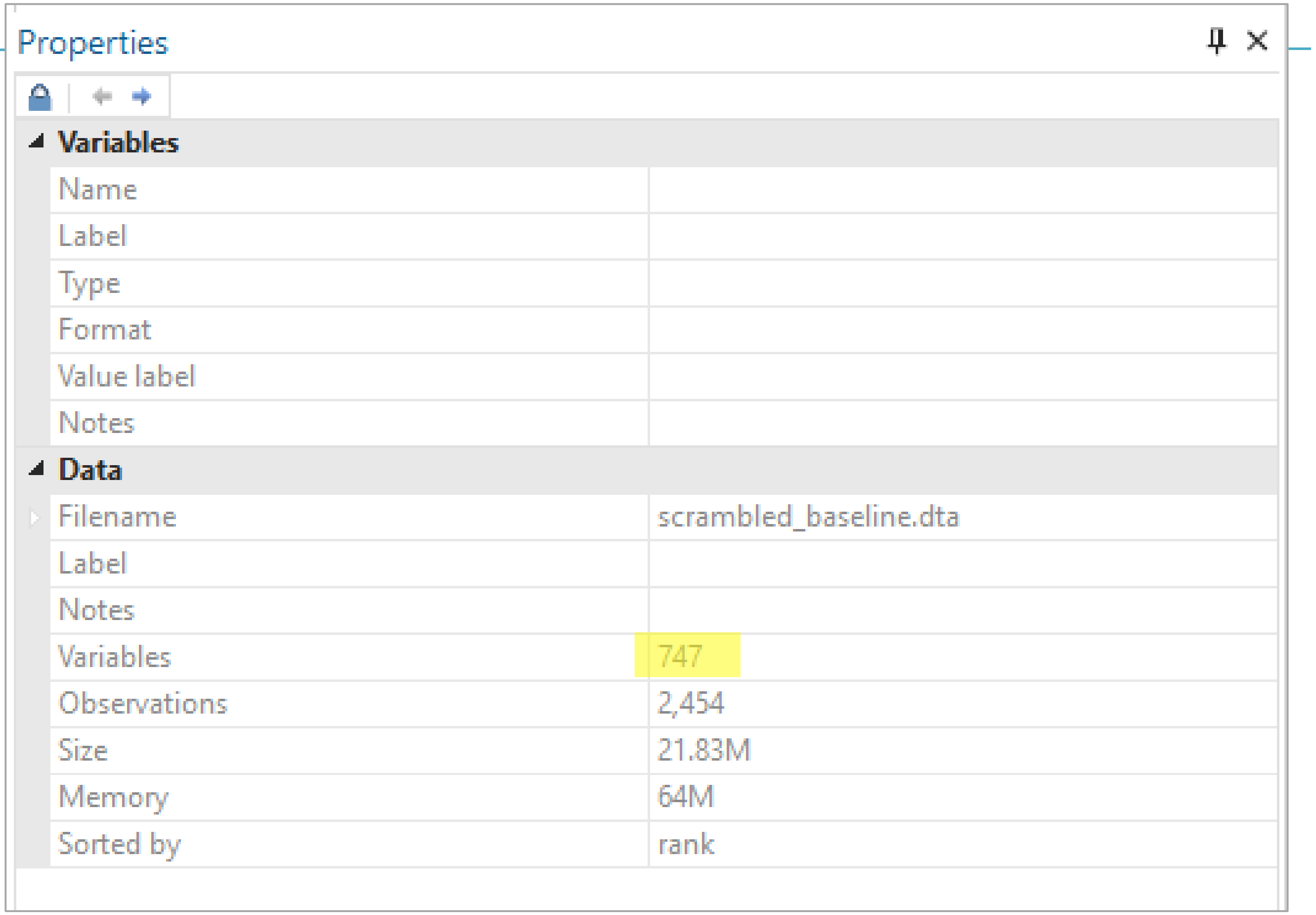
```
* Load the extremely sensitive auto data set
sysuse "${auto_encrypt}/Raw Identified Data/auto.dta", clear

* Create deidentifiable unique ID
encode make, gen(uuid)

* Apply codebook to drop identifying information
iecodebook apply using "${auto_encrypt}/Documentation/codebook.xlsx"

* Save de-identified data set
save "${auto_dt}/Raw Deidentified/auto.dta", replace
```



	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										

De-identifying data with iecodebook

- Easy, right?
- Now it's your turn



Properties

Variables	
Name	
Label	
Type	
Format	
Value label	
Notes	
Data	
Filename	scrambled_baseline.dta
Label	
Notes	
Variables	747
Observations	2,454
Size	21.83M
Memory	64M
Sorted by	rank