LINKING DATA IN THE CVFS AND BEYOND

Emily Treleaven
Adrienne Epstein

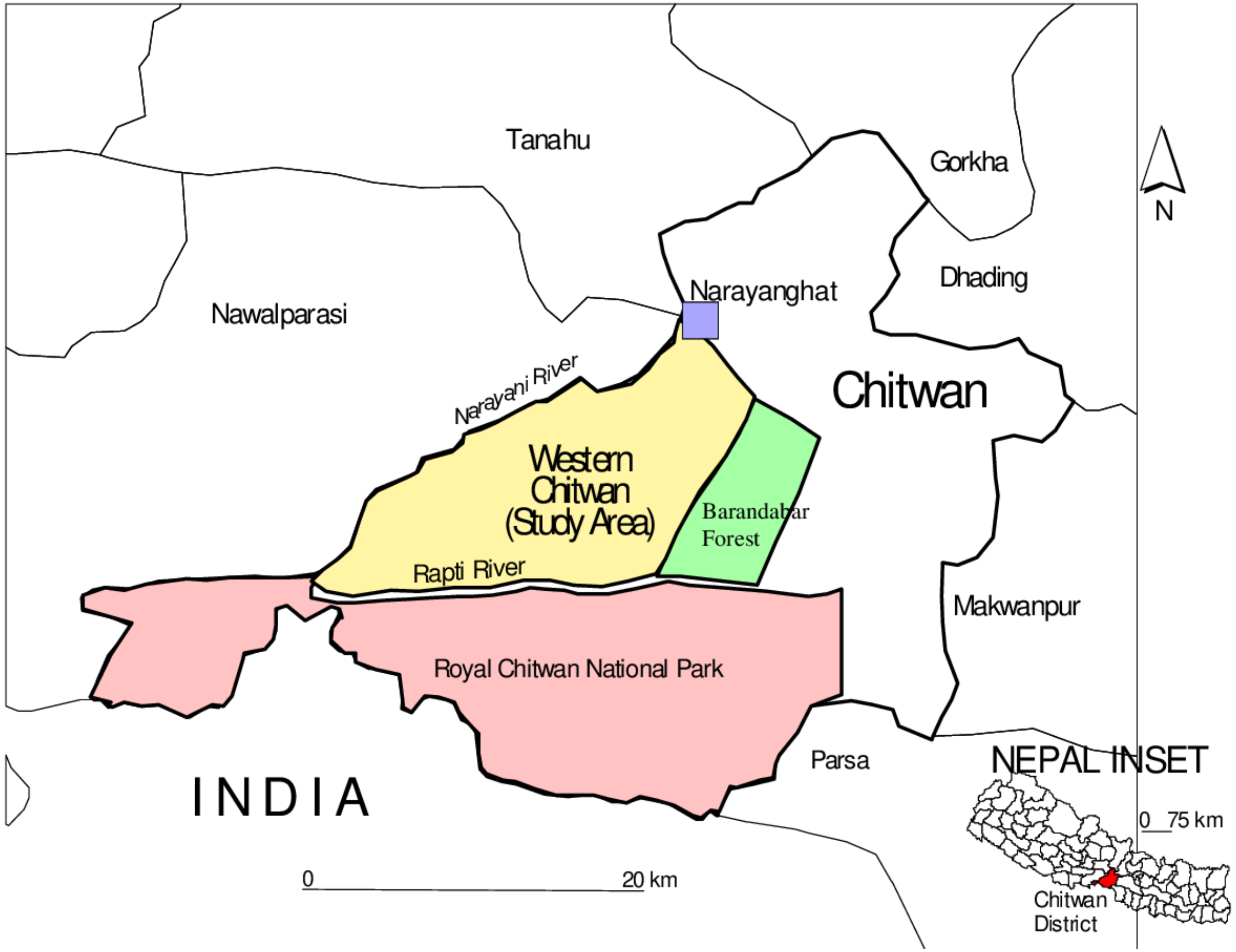
October 13, 2021

OUTLINE

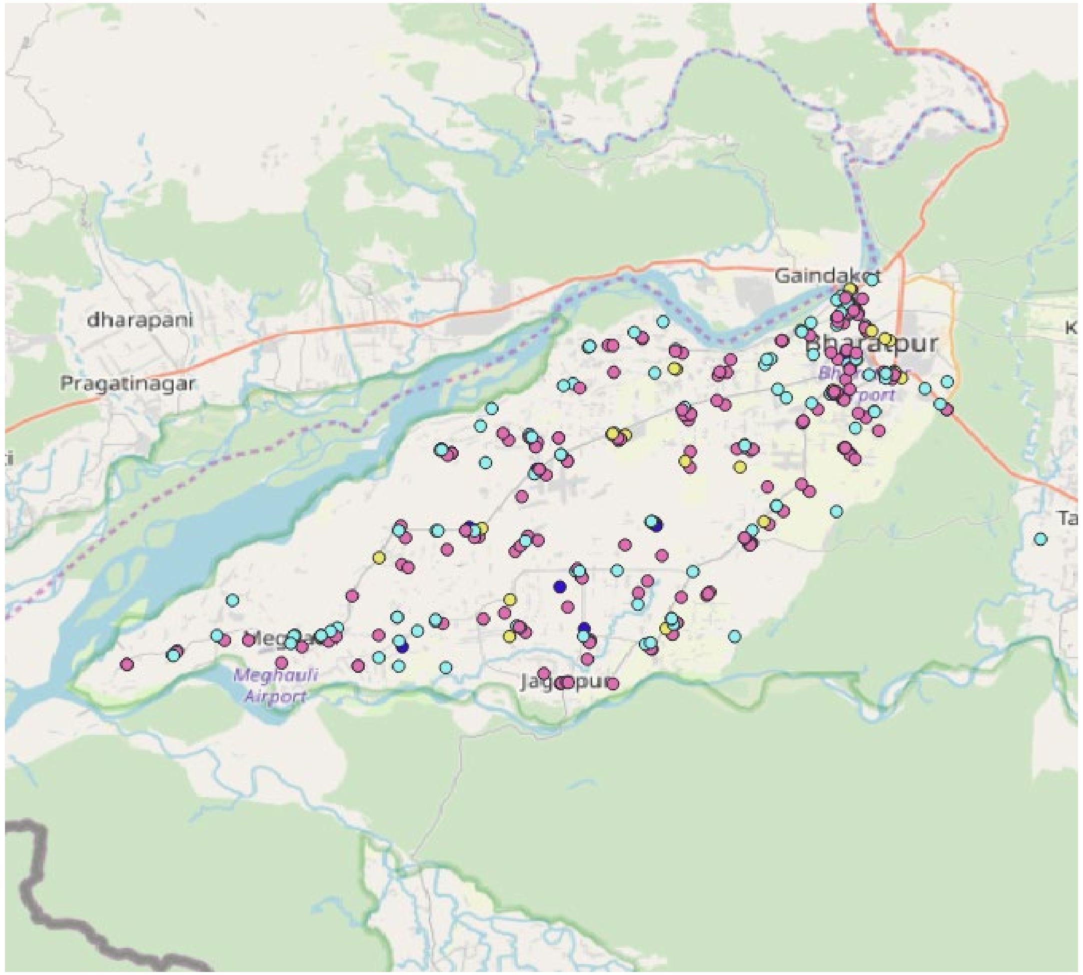
- Overview of CVFS sampling frame
- CVFS file types & organization
- Merging
  - Similar units across files
  - Different units within the same group
  - Spatial data to CVFS
- Example: Using external rainfall data to study drought & migration in CVFS
- Q&A
- Resources

CVFS SAMPLING FRAME

LOCATION OF CVFS STUDY SITE



CVFS STUDY AREA



Map of health facilities in CVFS study area

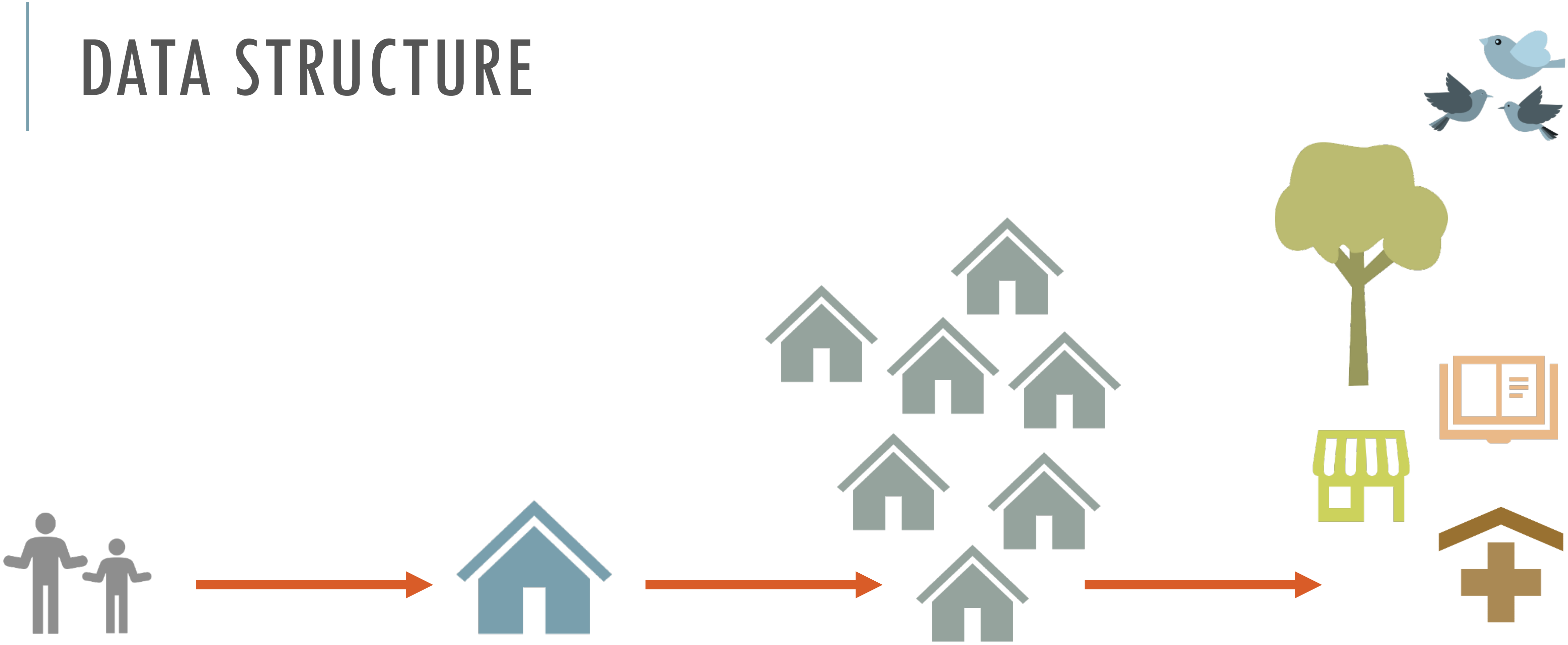
SAMPLED NEIGHBORHOODS

- Neighborhood = geographic cluster of 5-15 households when sample drawn
  - All households in each sampled neighborhood included
- 171 neighborhoods in original 1995-1996 sample
- 151 neighborhoods included in sample from 1997 onwards
- Change in sampled neighborhoods over time
  - Attrition/growth in number of households; may be <5 or >15
  - 2 neighborhoods washed away due to flooding
    - Individuals and households never returned to original neighborhood location
    - Individuals and households from these neighborhoods still included in sample

FEATURES OF CVFS SAMPLE

- Individuals linked to households linked to neighborhoods
- Family members/residents in the same household linked to each other
  - Spouses
  - Parents and children
  - Other household members
- Multiple measures *over time*
- CVFS follows individuals over time *regardless of location*

DATA STRUCTURE



Individuals

Households

Neighborhoods

Context

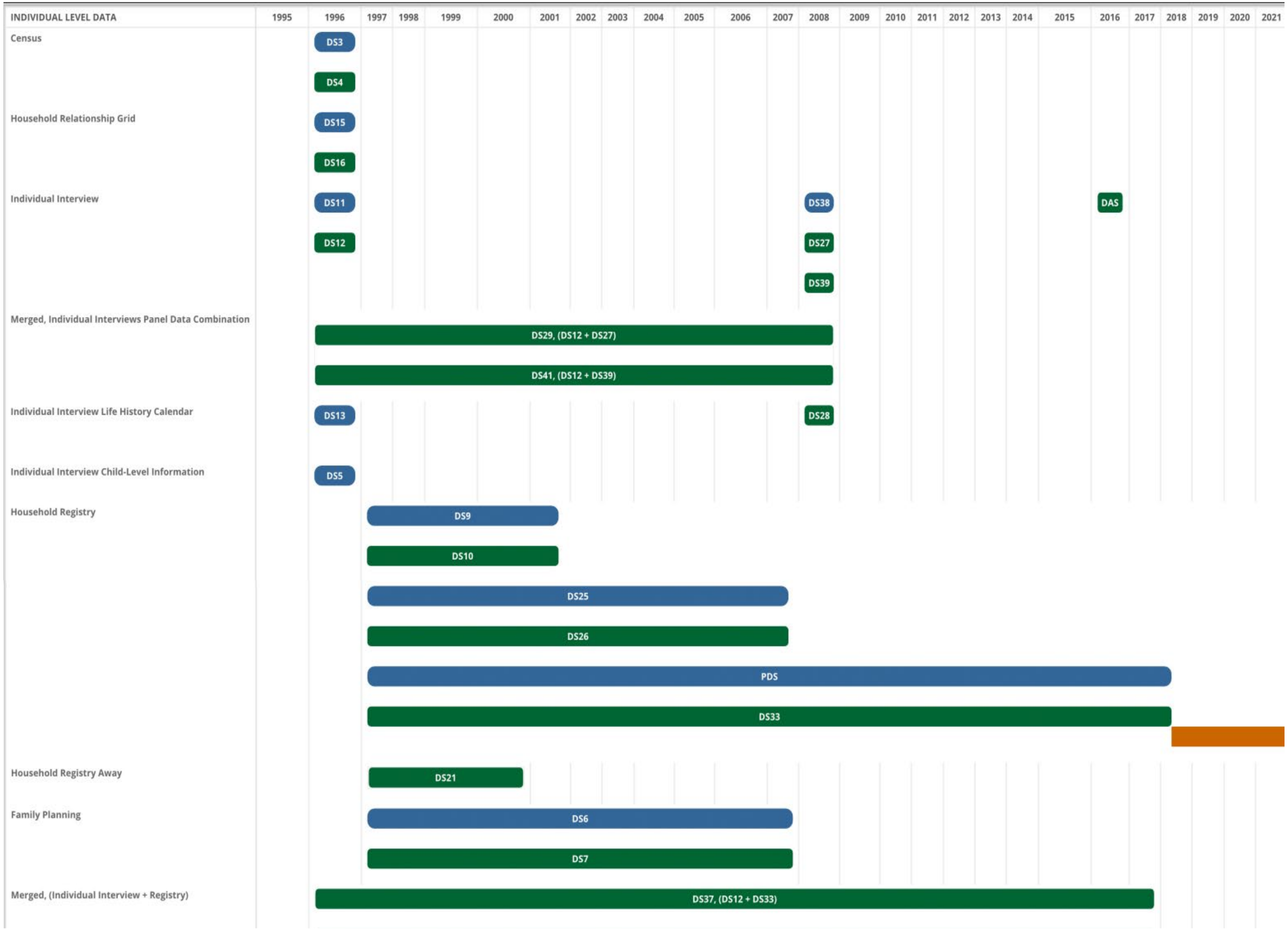
INDIVIDUALS AND HOUSEHOLDS IN CVFS

- Extremely low attrition/loss to follow up rates
  - Households that moved out of study area 1996-2007 were no longer tracked from 2008 onwards
- Individuals can enter sample through birth, marriage, or by joining an existing household
- New households enrolled if they move into a sampled neighborhood
- Enrolled individuals can split off into new household (at least 2 members leave to form new household)
- Individuals leave the sample through death or loss to follow up

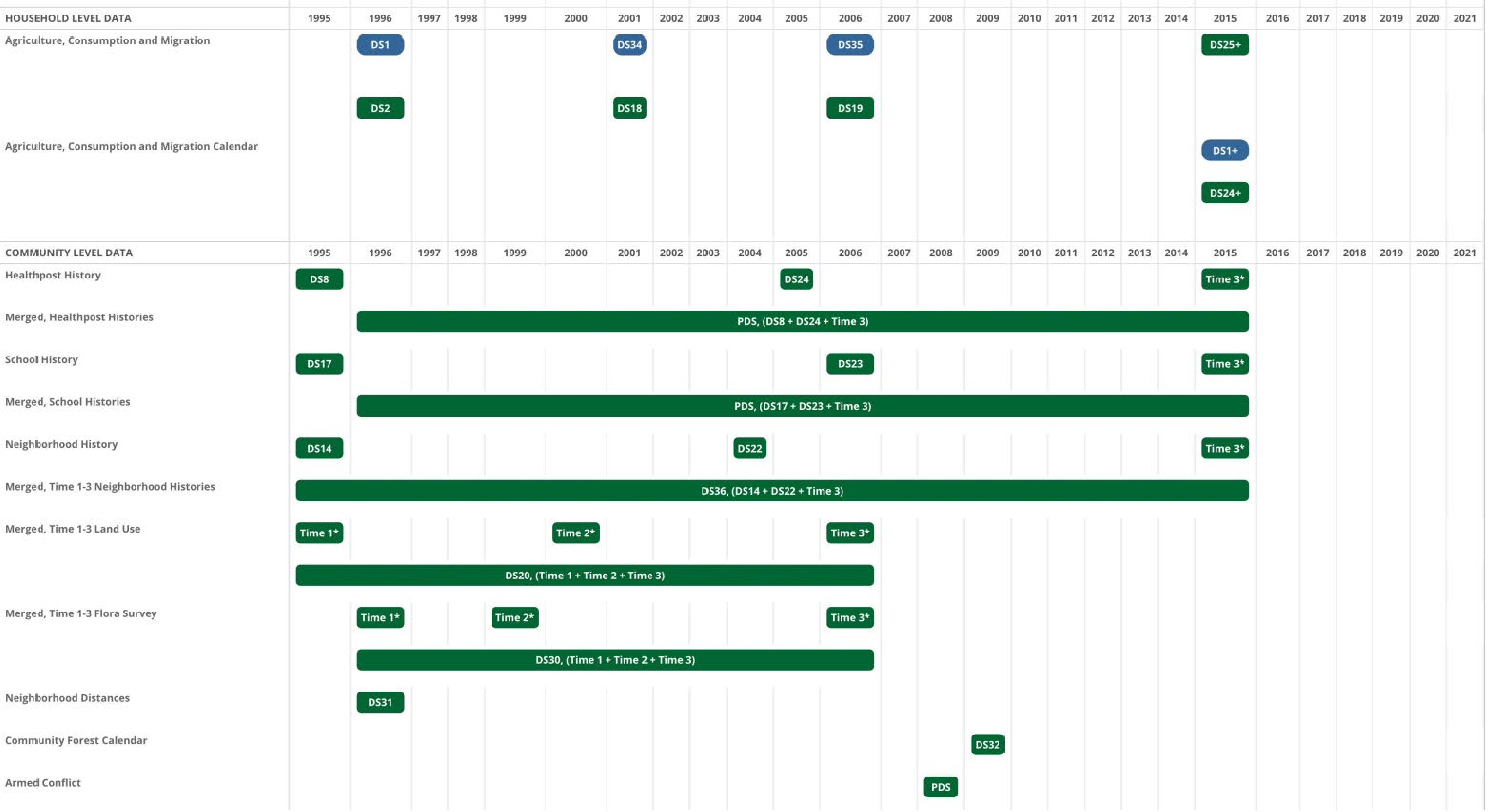
CVFS
FILE TYPES &
ORGANIZATION

WHEN DO I NEED TO MERGE FILES?

- Match information from the same individual, household, neighborhood, or service over multiple time points/data files
- Match household or neighborhood characteristics to an individual
- Match characteristics about one individual to another



CVFS Study Diagram: https://cvfs.isr.umich.edu/wp-content/uploads/2021/09/study_diagram.pdf



HOUSEHOLD LEVEL DATA	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Agriculture, Consumption and Migration		DS1 DS2					DS34 DS18					DS35 DS19									DS25+						
Agriculture, Consumption and Migration Calendar																					DS1+ DS24+						

COMMUNITY LEVEL DATA	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Healthpost History	DS8										DS24										Time 3*						
Merged, Healthpost Histories		PDS, (DS8 + DS24 + Time 3)																									
School History	DS17											DS23									Time 3*						
Merged, School Histories		PDS, (DS17 + DS23 + Time 3)																									
Neighborhood History	DS14									DS22											Time 3*						
Merged, Time 1-3 Neighborhood Histories	DS36, (DS14 + DS22 + Time 3)																										
Merged, Time 1-3 Land Use	Time 1* DS20, (Time 1 + Time 2 + Time 3)					Time 2*						Time 3*															
Merged, Time 1-3 Flora Survey		Time 1* DS30, (Time 1 + Time 2 + Time 3)			Time 2*							Time 3*															
Neighborhood Distances		DS31																									
Community Forest Calendar															DS32												
Armed Conflict														PDS													

CVFS Study Diagram: https://cvfs.isr.umich.edu/wp-content/uploads/2021/09/study_diagram.pdf

INDIVIDUAL-LEVEL FILES

- Household registry
- Individual interviews
- Life history calendars
- Other periodic surveys
  - Child health
  - Time use
  - Health & well-being among older adults
- Relationship grid

HOUSEHOLD-LEVEL FILES

- Agriculture, land use, durable goods, and consumption
- Remittance history calendar

AREA-LEVEL FILES

- Neighborhood history calendars (NHC)
- School history calendars (SHC)
- Health service history calendars (HHC)
- Land use
- Flora & fauna biodiversity
- Community forest characteristics and use

MERGING IN CVFS

7 DIGIT : ID STRUCTURES IN CVFS (1996-2007)

Neighborhood	+	Household	+	Individual
3 digits 001-171		2 digits		2 digits

9 DIGIT ID STRUCTURES IN CVFS (2008 -)

Neighborhood	+	Household	+	Individual
3 digits 001-171		3 digits		3 digits
		0 + 2 digits from older ID		0 + 2 digits from older ID

ICPSR FILES: PUBLIC VS. RESTRICTED USE

- Public-use files released on ICPSR prior to 2016 use consistent IDs
- Public-use files released on ICPSR in 2016 and later have scrambled IDs
  - Cannot match individuals across files
  - No area-level or GIS information
- Restricted-use files use consistent IDs regardless of release date

INDIVIDUALS ACROSS FILES

- Individual to same individual (1:1 merge)
  - Respid (9 digits); 2008 and later
  - Respid or Respid_old (7 digits); up to 2008
- Individual to another individual in same household
  - Spouses
  - Parents & children
  - Siblings
  - Household members

CVFS RELATIONSHIP GRID

- Provides individual IDs by relationship type to every other individual in a household
  - Includes all members listed for a specific household ID in HHR at each time point
  - Did not have to be physically present in household at time of relationship grid to be included
- 4 time points
  - T1 = 1996
  - T2 = 2001
  - T3 = 2008
  - T4 = 2016

CVFS RELATIONSHIP GRID

hhid	subject	Respid	spouse1	spouse2	spouse3	parent1	parent2	child1	child2	child3	child4
212007	1	212007020	212007003					212007011	212007012	212007013	
212007	2	212007003	212007020					212007011	212007012	212007013	
212007	3	212007011				212007003	212007020				
212007	4	212007012				212007003	212007020				
212007	5	212007013				212007003	212007020				
218010	1	218010001	218010012			218010012	218010015	218010003	218010004		
218010	2	218010003				218010012	218010001				
218010	3	218010004				218010012	218010001				
218010	4	218010015	218010016					218010001	218010029	218010019	
218010	5	218010016	218010015					218010001	218010029	218010019	
218010	6	218010007	218010009					218010011	218010013		
218010	7	218010009	218010007			218010016	218010015	218010011	218010013		
218010	8	218010011				218010007	218010009				
218010	9	218010013	218010001					218010003	218010004		

HOUSEHOLDS ACROSS FILES

- Household to same household (1:1 merge)
  - HHID (5 or 6 digits)

- May need to convert old, new HHIDs to merge
  - 5 digit HHID = 3 digit neighborhood + 2 digit household
  - 6 digit HHID = 3 digit neighborhood + 3 digit household
  - To convert from 5 to 6 digit ID, add 0 to 2 digit household ID

NEIGHBORHOODS OR FACILITIES ACROSS FILES

- Neighborhood to neighborhood
  - Neighid, nbhid (1:1 merge)
  - nx, ny (UTM X/Y coordinates in restricted-use data)
- Health facilities
  - hlthid (1:1 merge)
  - hx, hy (UTM X/Y coordinates in restricted-use data)
- Schools
  - schlid (1:1 merge)
  - sx, sy (UTM X/Y coordinates in restricted-use data)

INDIVIDUALS TO HOUSEHOLDS

- Match on household ID variable (m:1 merge)
- Special considerations
  - Household component of individual ID is derived from HHID of household where individual first enrolled in CVFS
  - Household component of individual ID may not match current household
  - Individual IDs are not reassigned: middle 3 digits (household) will not change
  - See Household Registry for household ID of individual in any particular month (hhid*month#*, *e.g.*, hhid180)

INDIVIDUALS TO NEIGHBORHOODS

- Match on neighborhood ID variable (m:1 merge)
- Special considerations
  - Neighborhood component of individual ID may not match current location
  - Individual IDs are not reassigned: first 3 digits (neighborhood) will not change
  - See Household Registry for location of individual in any particular month (loc*month#*, *e.g.*, loc180)

HOUSEHOLDS TO NEIGHBORHOODS

- Match on neighborhood ID variable (m:1 merge)
- Special considerations
  - Neighborhood component of household ID may not match individual members' current location
  - See Household Registry for location of individual in any particular month (loc*month#*, *e.g.*, loc180)

GIS-BASED MERGES

- Neighborhoods and services to non-CVFS data using UTM X/Y coordinates
  - Neighborhoods
  - Health facilities
  - Schools

Q&A

CVFS RESOURCES

CVFS study diagram: https://cvfs.isr.umich.edu/data/data-documentation/study-diagram/

CVFS data access: https://cvfs.isr.umich.edu/data/data-access/data-access-instructions/

CVFS on ICPSR: https://www.icpsr.umich.edu/web/DSDR/series/646

CVFS FAQ: https://cvfs.isr.umich.edu/faq/

RESOURCES: CARRYING OUT MERGES

Stata:

https://www.princeton.edu/~otorres/Merge101.pdf

https://www.ssc.wisc.edu/sscc/pubs/dws/data_wrangling_stata7.htm

SAS: https://stats.idre.ucla.edu/sas/modules/match-merging-data-files-in-sas/

R: https://clayford.github.io/dwir/dwr_05_combine_merge_rehsape_data.html