

**WELCOME
TO THE 2026
NDACAN
SUMMER
TRAINING
SERIES!**

The session will begin at 12pm EST.

This session is being recorded.

Please submit questions to the Q&A box!

NDACAN SUMMER TRAINING SERIES

National Data Archive on Child Abuse and Neglect

Duke University, Cornell University, UC San Francisco, & Mathematica



Children's Bureau

An Office of the Administration for Children & Families



LAYING THE GROUNDWORK

*FOUNDATIONAL SKILLS FOR USING ADMINISTRATIVE
DATA IN CHILD WELFARE RESEARCH*

NDACAN SUMMER TRAINING SERIES SCHEDULE

- **July 1st, 2026**
 - Overview of NDACAN administrative datasets
- **July 8th, 2026**
 - Data cleaning and management
- **July 15th, 2026**
 - Linking NCANDS and AFCARS
- **July 22rd, 2026**
 - Handling missing data
- **July 29th, 2026**
 - Data presentation and visualization

SESSION AGENDA

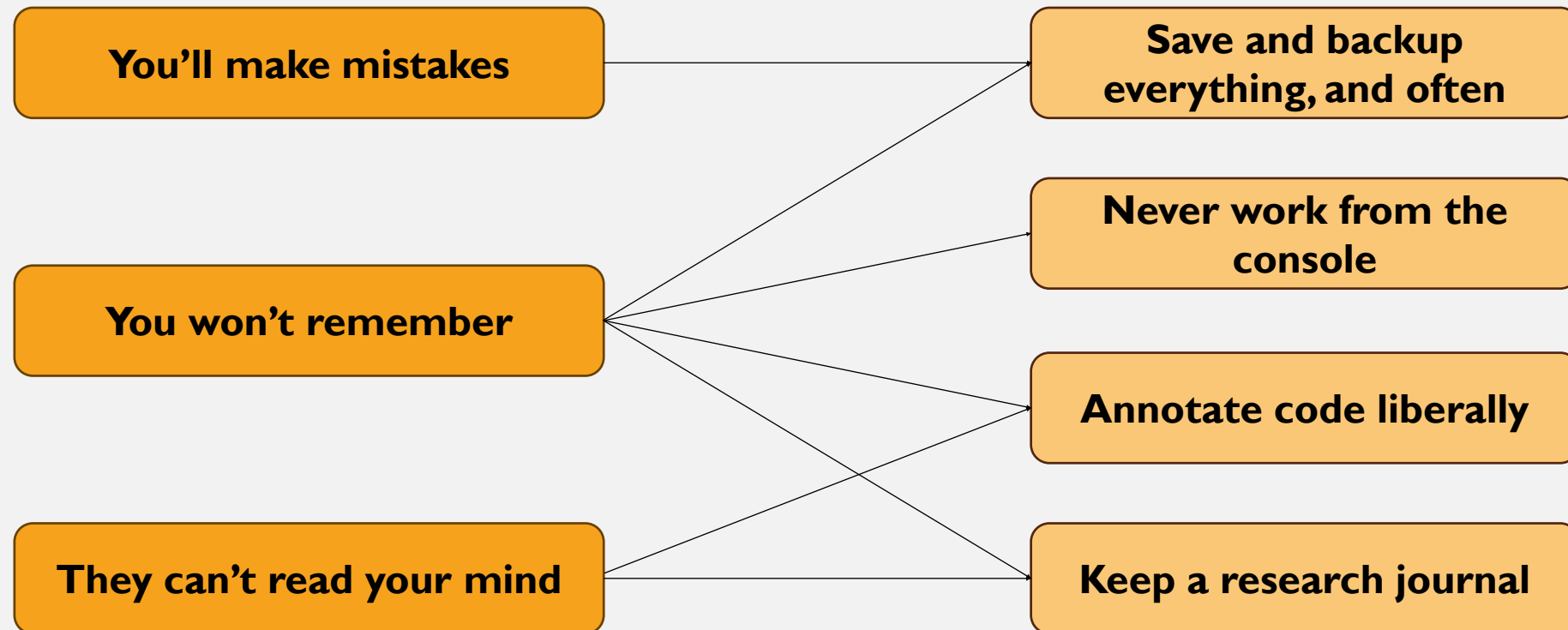
- Data cleaning and management
- Demonstration in R
- Q & A

DATA CLEANING AND MANAGEMENT

GENERAL ADVICE

- Use what's available to you.
 - NDACAN's site is rich with resources for support.
- Don't hesitate to ask for help.
 - NDACAN staff are here to support all aspects of NDACAN-related projects.
- Verify and master all AI-supported work.
 - Non-human support can be generative, but is unreliable.

DATA MANAGEMENT AS CRISIS MANAGEMENT



FILE ORGANIZATION AND WORKFLOW

- Project (local machine with backup to cloud server)
 - Programs (R scripts)
 - Drafts
 - Figures
 - Tables (deidentified data)
- Data (FedRAMP-authorized cloud server, encrypted external drive)
 - Raw
 - Derived (micro-data)

EXAMINING YOUR DATA

- What is the structure of your data? What are the columns? What are the rows?
- What viewpoint on your data do you need to understand it?
 - E.g. selecting columns, filtering rows, or summarizing values

CLEANING YOUR DATA

- How are variables formatted?
 - String, numeric, factor, etc.
- How are values coded?
 - Codes will not always correspond to the code book: always verify.
- What formatting/encoding captures the relevant information most effectively?
 - This depends not only on the data, but also on your analytic goals.

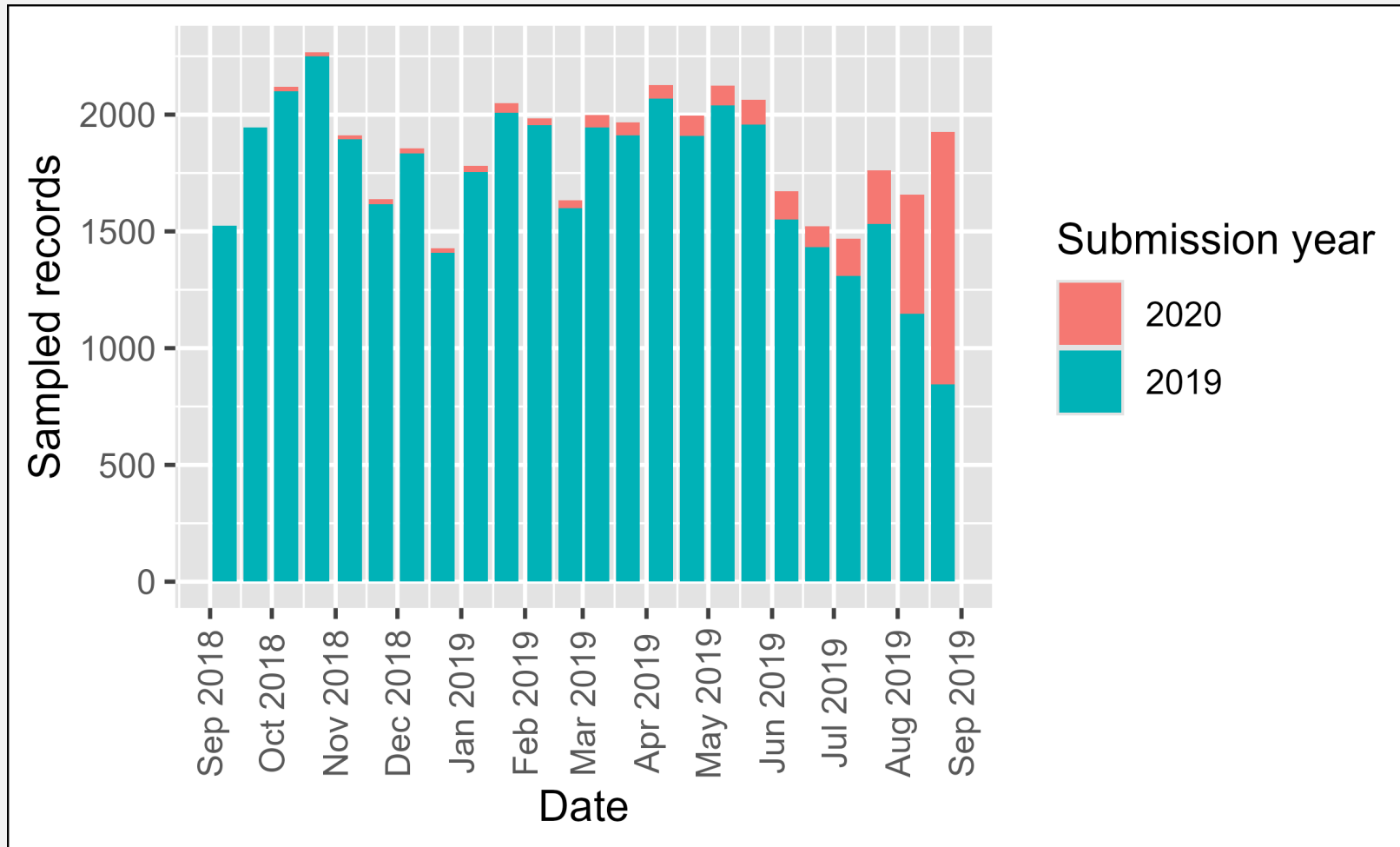
DEFINING YOUR UNIT OF ANALYSIS

- Different NDACAN administrative datasets have different default units of analysis
 - NCANDS: child-report
 - AFCARS: child-year
 - NYTD: child-wave
- Your research question may require you to tidy your data to capture your unit of analysis
 - “How many children experienced a maltreatment investigation in each state annually?”
 - In NCANDS, drop records from subsequent reports experienced by children within the same year, then count records by state and year.
 - “How many maltreatment investigations occurred each half-month?”
 - In NCANDS, drop records from additional children covered in each investigation, then count records by half-month.

CLARIFYING YOUR SAMPLING FRAME

- Different research questions recommend different temporal sampling frames.
 - E.g. fiscal year 2018, calendar year 2010, or 2012 birth cohort through age 6
- It is easy to confuse the timing of event occurrence and event reporting.
 - E.g. a maltreatment investigation included in the 2019 Child File may have occurred in the 2018 or 2020 fiscal years.
- Careful combination and tidying of multi-year data is usually needed to sample validly.

DATA YEAR VS. SUBMISSION YEAR



Note: Results based on 1% random samples of the 2019 and 2020 NCANDS Child Files.

TIDYING YOUR DATA

- Each observation should have its own row.
 - This should correspond to your unit of analysis.
- Each variable should have its own column.
- Each value should have its own cell.
- When they don't, summarizing and pivoting are the most common methods of tidying.

SUMMARIZING AS TIDYING (NCANDS)

Raw data

FY	StFCID	RptDisp
2021	AL12345	1
2021	AL12345	1
2021	KY67890	1
2022	AL12345	1
2022	KY67890	5
2022	KY67890	1



Summarized data

FY	StFCID	NSub
2021	AL12345	2
2022	AL12345	1
2021	KY67890	1
2022	KY67890	1

PIVOTING AS TIDYING (NCANDS)

Long data

FY	StFCID	NSub
2021	AL12345	2
2022	AL12345	1
2021	KY67890	1
2022	KY67890	1



Wide data

StFCID	NSub_2021	NSub_2022
AL12345	2	1
KY67890	1	1

PROGRAMMATIC CODING

- Excessive copying/pasting of code can decrease interpretability, increase risk of error.
- Programmatic coding makes repetitive processes more transparent, malleable.
- Common examples:
 - Within the Tidyverse, functions in the purrr package
 - User-written functions
 - Loops

DEMONSTRATION IN R

GETTING STARTED WITH R AND RSTUDIO

[PRODUCTS](#) ▾[OPEN SOURCE](#) ▾[USE CASES](#) ▾[PARTNERS](#) ▾[LEARN & SUPPORT](#) ▾[ABOUT](#) ▾

DOWNLOAD

RStudio Desktop

Used by millions of people weekly, the RStudio integrated development environment (IDE) is a set of tools built to help you be more productive with R and Python.

Don't want to download or install anything? Get started with RStudio on [Posit Cloud for free](#). If you're a professional data scientist looking to download RStudio and also need common enterprise features, don't hesitate to [book a call with us](#).

Want to learn about core or advanced workflows in RStudio? Explore the [RStudio User Guide](#) or the [Getting Started](#) section.

1: Install R

RStudio requires R 3.6.0+. Choose a version of R that matches your computer's operating system.

R is not a Posit product. By clicking on the link below to download and install R, you are leaving the Posit website. Posit disclaims any obligations and all liability with respect to R and the R website.

[DOWNLOAD AND INSTALL R](#)

2: Install RStudio

[DOWNLOAD RSTUDIO DESKTOP FOR WINDOWS](#)

Size: 281.27 MB | [SHA-256: 9E6F68CA](#) | Version: 2025.05.0+496 | Released: 2025-05-05

ADDITIONAL RESOURCES

- *Hands-On Programming with R*, Garrett Grolemund
- *R Cookbook (2e)*, J.D. Long and Paul Teeter
- *R for Data Science (2e)*, Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund

QUESTIONS?

USER SUPPORT

NDACANSUPPORT@CORNELL.EDU

ALEX ROEHRKASSE

AROHRKASSE@BUTLER.EDU

ALYSSA LINDSEY

ALYSSA.LINDSEY@UCSF.EDU

NEXT WEEK...

Date: July 15th

Topic: Linking NCANDS and AFCARS

Instructor: Alex Roehrkasse

COMMENTS IN THE ZOOM CHAT

- The NDACAN website is <https://www.ndacan.acf.hhs.gov> and contains the following:
 - The R code used in this presentation.
 - Information and instructions for joining the Child-Maltreatment-Research-L (CMRL) list serve.