CHARACTERIZING THE NIH RESEARCH WORKFORCE

NIH Office of the Director, Office of Extramural Research

Katrina Pearson, Director of Division of Statistical Analysis and Reporting

CIVIC DIGITAL FELLOWSHIP

NIH National Institutes of Health
Office of Extramural Research

SAM KIM
Duke University
Biology & Computer Science

MOTIVATION

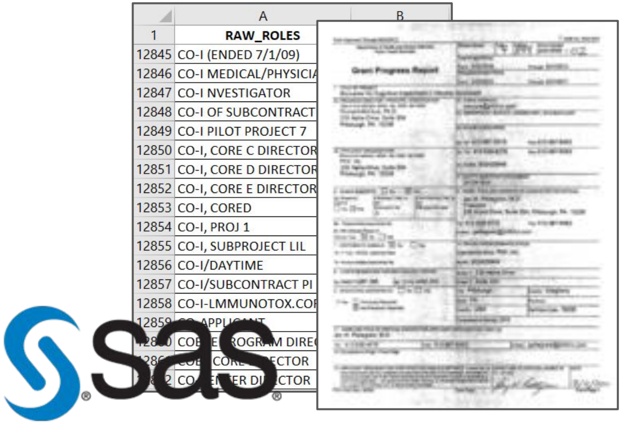
- Annual NIH appropriation exceeds $37 billion*
- More than 80% is awarded to researchers as **extramural grants/contracts**[†]
- Research output can be inferred from publications
- What is the **workforce output**?
    - That is, **how many jobs** does NIH funding directly create?

* NIH Operating Plan for Fiscal Year 2019. NIH Office of Budget, 2019.

[†] NIH Data Book. NIH Division of Statistical Analysis and Reporting, 2019.

MOTIVATION (cont.)

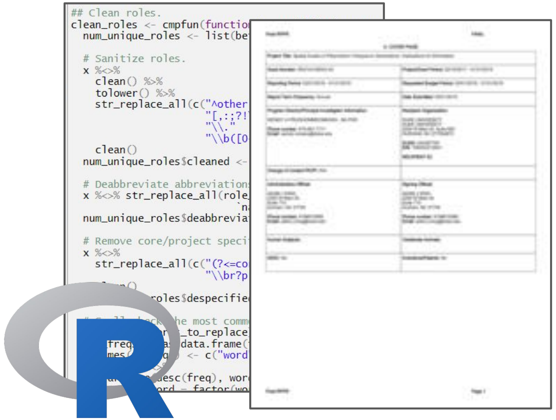
- **Automate and generalize** the workforce analysis



	A	B
1	**RAW_ROLES**	
12845	CO-I (ENDED 7/1/09)	
12846	CO-I MEDICAL/PHYSICIA	
12847	CO-I NVESTIGATOR	
12848	CO-I OF SUBCONTRACT	
12849	CO-I PILOT PROJECT 7	
12850	CO-I, CORE C DIRECTOR	
12851	CO-I, CORE D DIRECTOR	
12852	CO-I, CORE E DIRECTOR	
12853	CO-I, CORED	
12854	CO-I, PROJ 1	
12855	CO-I, SUBPROJECT LIL	
12856	CO-I/DAYTIME	
12857	CO-I/SUBCONTRACT PI	
12858	CO-I-LMMUNOTOX.COR	
12859	CO-APPLICANT	
12860	CORE PROGRAM DIREC	
12861	CORE CORE DIRECTOR	
12862	CO-CENTER DIRECTOR	

Grant Progress Report

Analysis of FY 2009
(Pool et al., 2016)*



```
## Clean roles.
clean_roles <- cmpfun(functio
  num_unique_roles <- list(be

  # Sanitize roles.
  x %<>%
    clean() %>%
    tolower() %>%
    str_replace_all(c("^other
                      "[,:;?!
                      "\\."
                      "\\b([0
    clean()
  num_unique_roles$cleaned <-

  # Deabbreviate abbreviation
  x %<>% str_replace_all(role
                          `na
  num_unique_roles$deabbrevia

  # Remove core/project speci
  x %<>%
    str_replace_all(c("(?<=co
                      "\\br?p
    clean()
        roles$despecifie

  #                he most comm
             _to_replace
   freq      data.frame(
   mes         <- c("word

              desc(freq), wor
              ord - factor(wo
```

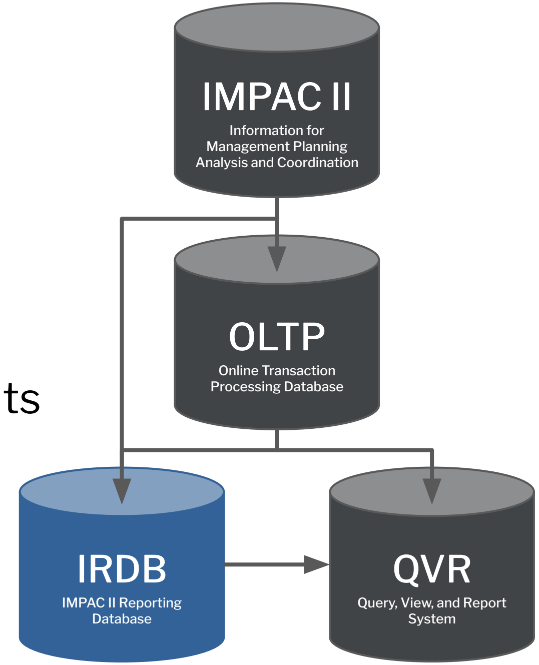
Analysis of FY 2009
(Pool et al., 2016)*

Analysis of FY 2017
(Kim et al., 2019)

* Pool LR, Wagner RM, Scott LL, et al. (2016). Size and characteristics of the biomedical research workforce associated with U.S. National Institutes of Health extramural grants. *FASEB J.* 30(3):1023–1036.

DATA SELECTION

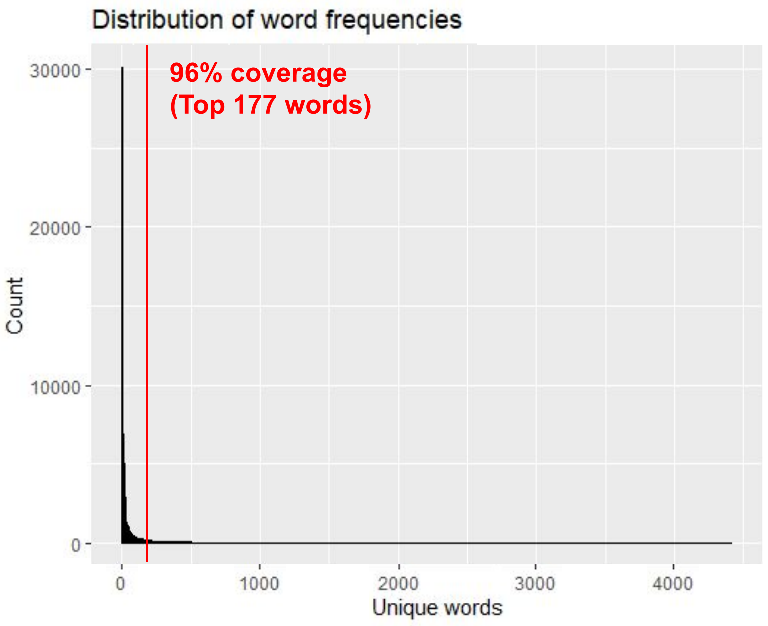
- **All-Personnel Reports** (APRs)
  - Required for RPPRs since 2010
  - Self-reported by awardees
  - Includes all people who devoted 1+ month of effort
- **All awarded and funded NIH extramural grants**
  - No intramural grants or inter-/intra-agency agreements
  - No subprojects or contracts
  - No noncompeting supplements (Type 3's)



IMPAC II
Information for Management Planning Analysis and Coordination
OLTP
Online Transaction Processing Database
IRDB
IMPAC II Reporting Database
QVR
Query, View, and Report System

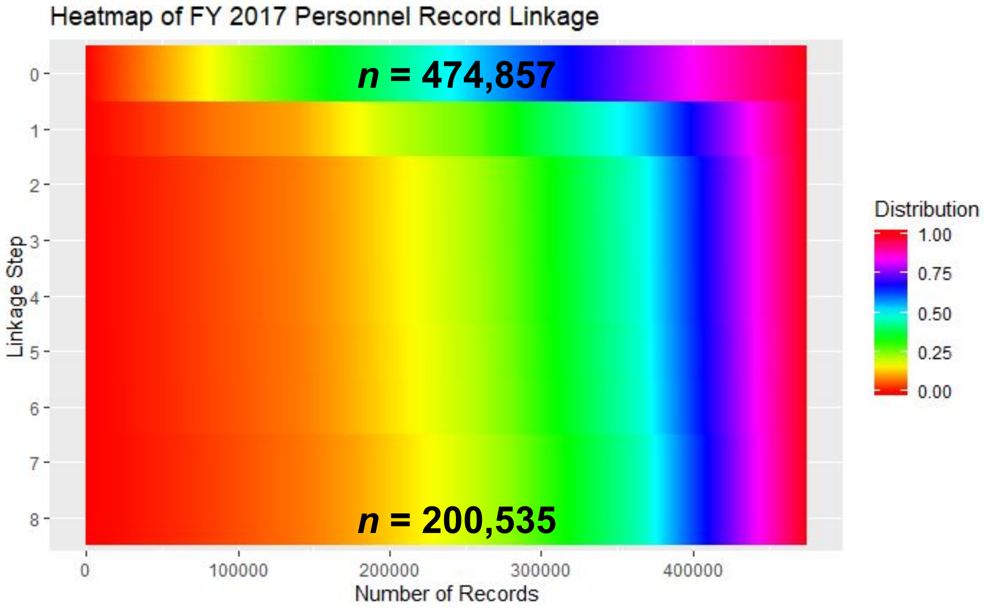
DATA PROCESSING

- **Very poor data quality**
- Misspelled names/words
- Position titles in free text
    - Custom spellchecker using edit distances
- Missing identifiable information
    - Record linkage methods
- **Missing RPPRs** for about 28% of grants
    - Impute personnel data

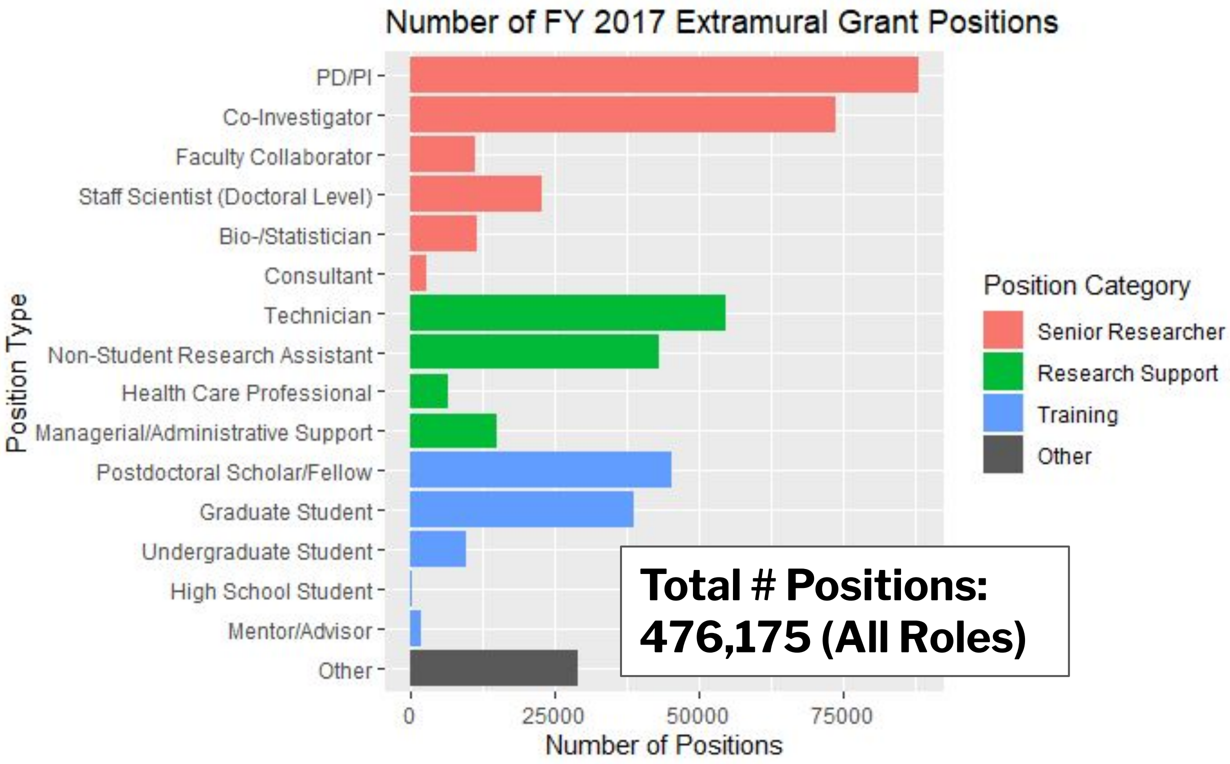


RECORD LINKAGE

- How to resolve duplicated, misspelled or similar names?
  - First and last names (99%)
  - Middle names (33%)
  - Commons profile ID (58%)
  - Institutional ID (100%)
  - SSN, birthdate (13%)
- Machine learning was impractical
  - Clustering: 450k × 450k

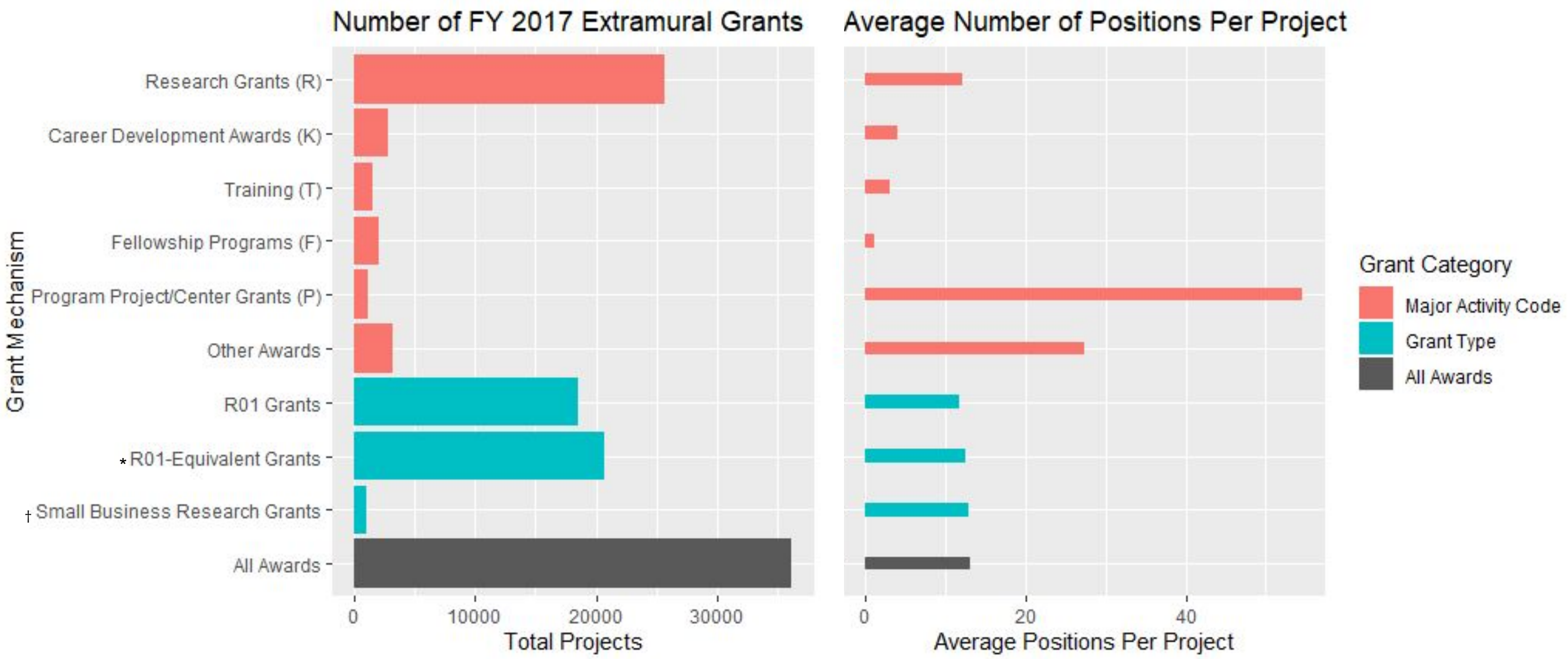


# POSITIONS



Total # Positions: 476,175 (All Roles)

# POSITIONS

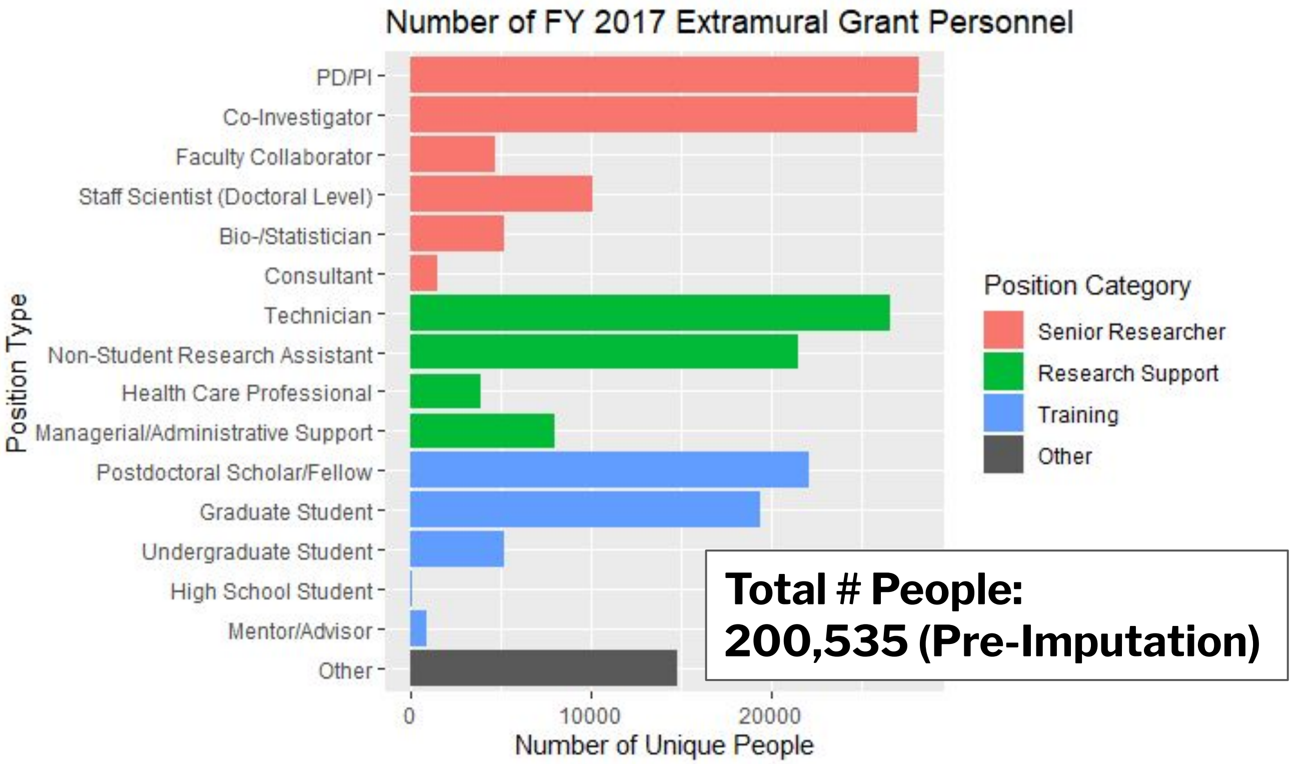


CIVIC DIGITAL FELLOWSHIP | NIH

* Includes R01, R23, R29, R35-MIRA, R37, RF1, RL1, DP1, DP2, DP5 and U01 awards.

† Includes STTR (R41, R42) and SBIR (R43, R44) awards.

# UNIQUE PEOPLE



Total # People: 200,535 (Pre-Imputation)

NEXT STEPS

- Impute missing personnel data, especially for Type 1’s
- Breakdown by geographic location
- Analyze effort and career stage information
- Analyze for FYs 2016, 2015, 2014…
  - Same code, different input

ACKNOWLEDGMENTS

Division of Statistical Analysis and Reporting

- Katrina Pearson
- Deepshikha RoyChowdhury
- Lindsey Scott
- Charles Wu

Civic Digital Fellowship – NIH

- Jessica Mazerik
- Amit Rajesh

Funding Sources

NIH National Institutes of Health
Turning Discovery Into Health

The Washington Center

coding it forward >