Creating the CenHRS

Margaret C. Levenstein
Director, Inter-university Consortium for Political and Social Research

BD2K Workshop
Bethesda, Maryland
March 19, 2018

CenHRS Team

PI Team

- Michigan, Cornell, Census faculty, staff, and graduate students
- John Abowd, Joelle Abramowitz, Margaret Levenstein, Kristin McCue, Dhiren Patki, Ann Rodgers, Matthew Shapiro, Nada Wasi

Supported by a grant from the Sloan Foundation

- Possible because of support from NIA, SSA, and NSF for related work, including HRS itself
- Related research developed in NSF-Census Research Network

HEALTH AND RETIREMENT STUDY

A Longitudinal Study of Health, Retirement, and Aging

Sponsored by the National Institute on Aging

20,000 + Americans over the age of 50

- Surveyed every two years since 1992
- New cohorts added in 1993, 1998, 2004, 2010, 2016
- Includes both spouses
- Follows respondents through death
- Oversamples minorities

What is CenHRS?

- Linking HRS and Census business data

Health and Retirement Study
Panel Survey of 50+ year olds
Links to:
- Medicare and SSA records
- Pension records

Census Business Register
Links to:
- LEHD
- Economic Surveys and Censuses
- Extensive business administrative data

Census-Enhanced HRS
1. Confidential microdata
2. CenHRS

Innovative value of CenHRS

- Most survey locate individuals in households
  - Sometimes neighborhoods or schools
- We spend much of our lives at work
  - CenHRS will allow analysis of impact of work context, including co-workers, on health and well-being of HRS respondents

CenHRS and Big Data

- Enhancing survey data with digital traces of human activity
  - Earnings and employment records of co-workers
- Requires linking disparate data sources
  - Turning “big data” into research data almost always requires linking and classifying

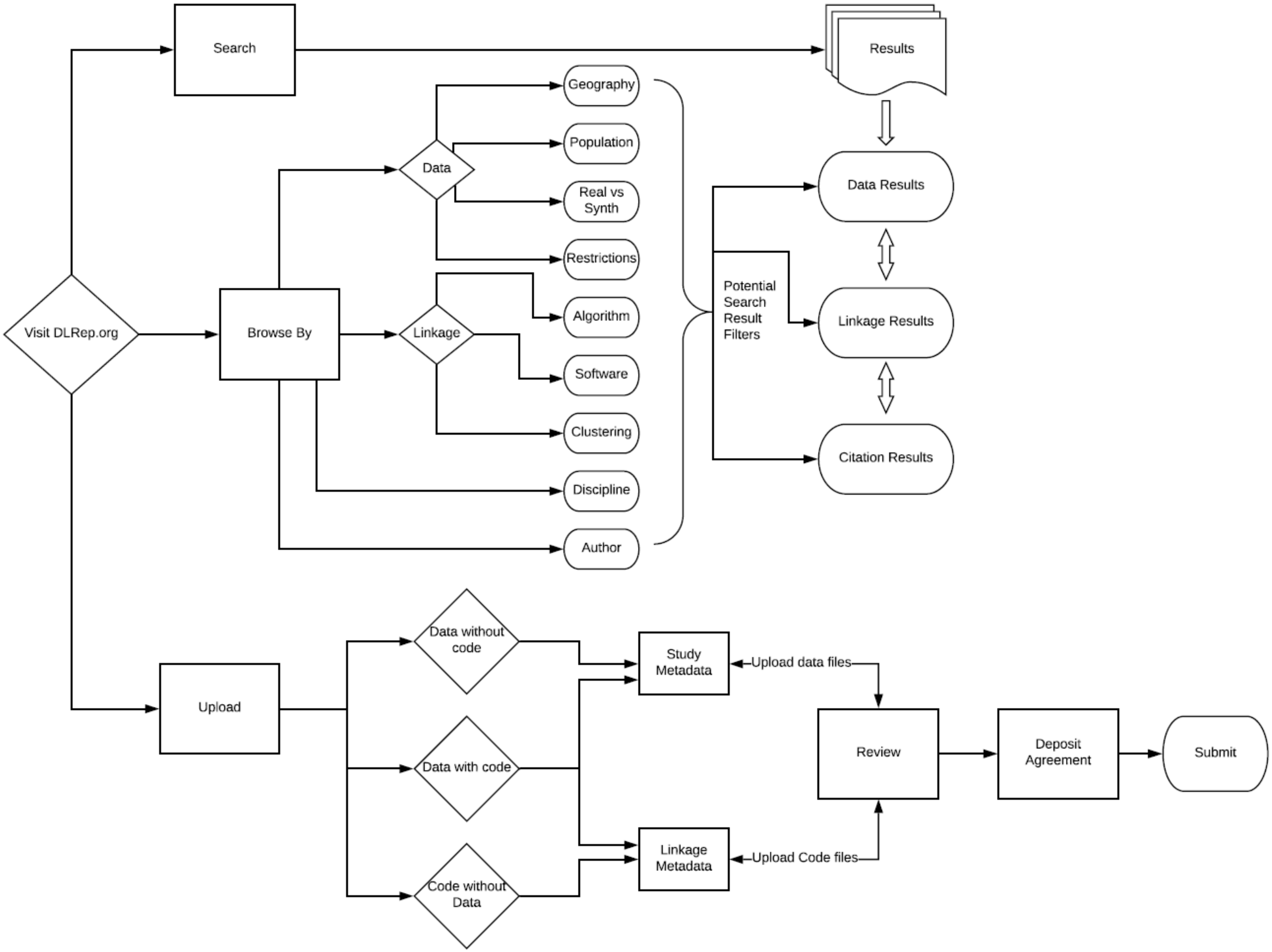
Challenges of linkage

- Rarely trivial, even when we have purportedly unique, direct identifiers
- Important to acknowledge uncertainty
  - Example: Michigan UI fiasco
- Important to acknowledge false positives, not just match rate
  - Example: Treatment effects biased downward when treatment is linked to the untreated
- Important to acknowledge false negatives
  - Often simply dropped, biasing samples
  - Bailey et al. (2017 and 2018) and LIFE-M

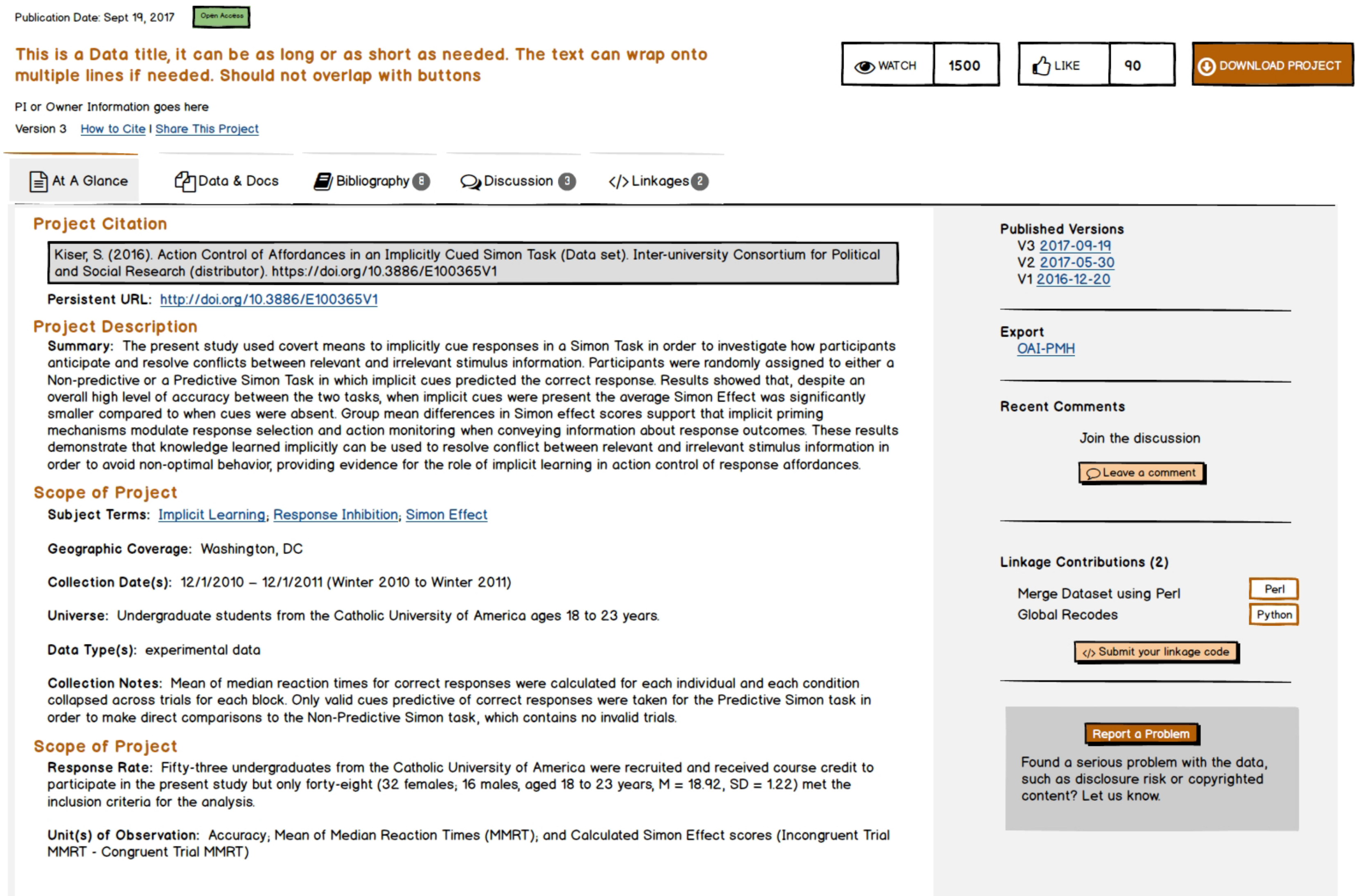
DLRep: Data Linkage Repository

- NSF funded archive at ICPSR
- Bringing together contributions from statisticians, computer scientists, demographers, survey methodologists
- Depositing code and data
- Facilitating comparison of data linkage approaches

DLRep schema wireframe



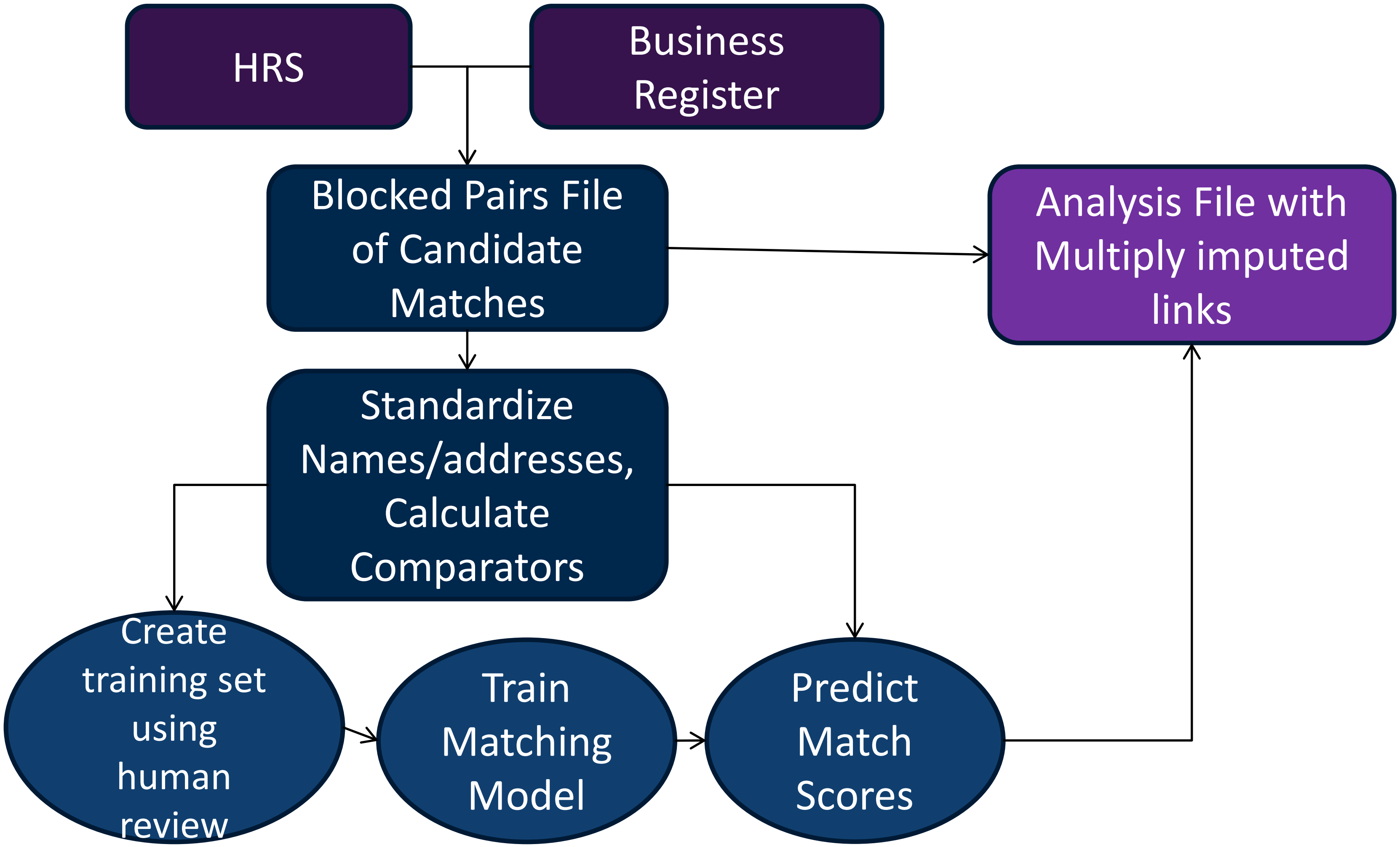
DLRep study home page wireframe



Creating CenHRS

1. Create ground truth (training data)
2. Train model
   - Use machine learning techniques to estimate posterior probability of match of HRS job with BR employer, within block
3. Multiply impute, with cutoffs proportional to block size

Linkage Process Flow



Step 1: Create training data

- Use subset of self-reports of 1992 HRS private-sector jobs, 1992 BR to work out methods
- Block on:
  - 10-digit phone number, where possible
  - 3-digit zip code, otherwise
- Standardize address and name fields, using rules developed specifically for business names
- Compute Jaro-Winkler string comparator scores for names and addresses

Construct set of HRS-BR pairs

- HRS jobs reported in 1998 and 2004
- BR in 1997-1999 and 2003-2006
  - Exclude if missing employer name or state, or missing both zip3 and phone # (10%)
  - <10% of phone numbers successfully blocked
  - Almost always at least 1 BR entry in zip3 block

Initial set of blocked pairs

- All possible within-block pairs > tens of millions
- Calculate JW scores comparing name, address
- Stratify using 4x4 cross-classification of JW scores
- Mean pairs per sampled HRS job=3,100, but varies from 1 to 20,000 across bins.
- Lowest JW scored bin accounts for:
  - 98% of pairs blocked on 3-digit zip
  - 42% of those blocked on 10-digit phone number
- Sample 100 pairs from each bin

Training data

- Each sampled pair reviewed by >=2 reviewers
- Reviewers see 1 pair at a time
  1. Employer name, address, and phone number
  2. Employer single unit/multiple unit status
  3. Employer and establishment size
  4. Employer industry code
- Assign separate scores for firm, establishment
- Score as follows:

1 =	Yes, match
2 =	Probably match
3 =	Maybe-maybe not
4 =	Probably not match
5 =	Not match
6 =	Not enough information

Step 2: Train model

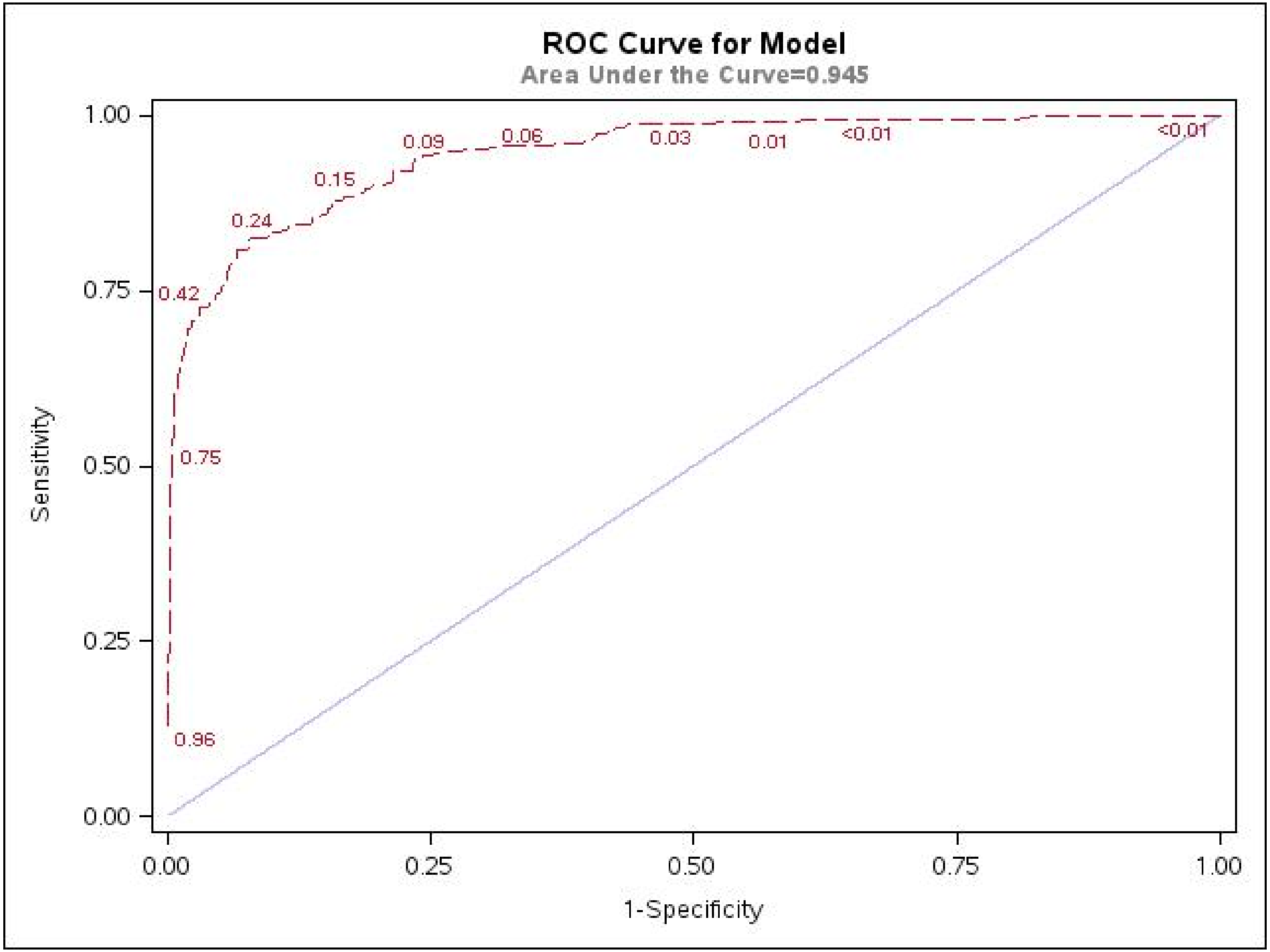
- Logistic model: dependent variable = 1 if pair scored as a match, 0 otherwise
- Regressors cubic splines of continuous variables, indicators, and full set of interactions
  - JW score, share of employment in block, size of employer
  - Agreement or missingness on
    - employer and establishment workforce
    - Single or multi-unit employer
    - seven and ten digit phone number
    - three and four digit zip code
    - SIC code
    - Does HRS job provide health insurance or a retirement plan and, if so, retirement plan type (defined benefit, defined contribution, both, or unknown)

Training matching model

- To limit overfitting and minimize out of sample error, we use elastic net shrinkage (Zou and Hastie, 2005)
  - Elastic net shrinka
    dimensionality of c
  - Optimal set of cov
    minimize cross-val

How well does model work?

- JW score most important determinant
  - Matters most where name and address are very similar
- Employer matches work better than establishment matches
- Checks on model match quality
  - Use EINs from HRS pension project
  - ROC curve



True positive rate

False positive rate

Step 3: Multiply impute linkage

- Unlike Fellegi-Sunter, we do *not* take highest probability match, as long as above threshold
- Rather, estimate probability of match to all employers/establishments in block
  - Drop those below optimal threshold, equally weighting sensitivity and specificity of ROC curve
    - Threshold proportional to size of block
    - Otherwise large mass of probability goes to large number of low probability matches
    - Re-normalize probabilities to sum to one among remaining organizations
- Multiply impute match ten times

Evaluating the MI approach

- Are results reasonable?
  - Concentration across imputations
  - Concordance between employer and establishment
- Is it worthwhile?
  - Employer size
    - Comparison of survey and administrative data
    - Implications for understanding of firm size-wage gradient

Conclusions

- Very cool new data, opens up wide range of research on impact of employment context on health, well-being, and labor-force attachment
- New methods using machine learning models to estimate probabilistic linkage
  - Do a reasonable job
  - Measure uncertainty, rather than throw away households or jobs that are harder to match

Acknowledgements and disclaimers

This research is supported by the Alfred P. Sloan Foundation through the Census-HRS project at the University of Michigan with additional support from the Michigan Node of the NSF-Census Research Network (NCRN) under NSF SES 1131500.

This research uses data from the Census Bureau's Longitudinal Employer-Household Dynamics Program, which was partially supported by the following National Science Foundation Grants SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant AG018854; and grants from the Alfred P. Sloan Foundation.

Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.