Creating a New Data Infrastructure for Foreign-born Scientists and Engineers: Data, Analysis, and Use

Yunie Le
Coleridge Initiative

FedCASIC 2023

Acknowledgement

This project is a part of the America's Datahub Consortium (ADC) collaboration sponsored by the National Center for Science and Engineering Statistics (NCSES) within the National Foundation of Science (NSF)

Disclaimer: The opinions discussed in this presentation are of Coleridge Initiative and do not necessarily reflect those of NCSES.

Outline

- Project Goals
- Data Model
- Data Assessment
- Data Linkage
- Lessons Learned

Foreign-Born Scientists and Engineers and the U.S. Workforce

- ***High-quality data on foreign-born scientists and engineers is important*** for studying inequalities and answering other questions about a population that is crucial to the advancement of U.S. science and engineering.
- Coleridge Initiative teamed up with
  - **New Jersey:** John J. Heldrich Center for Workforce Development, Rutgers the State University of **New Jersey**;
  - **Arkansas** Department of Transformation and Shared Services;
  - **Kentucky** Center for Statistics
- Evidence-building to understand the availability and demand for global science and engineering training and talent: https://www.americasdatahub.org/fbse/

COLERIDGE INITIATIVE ARDATA KYSTATS HELDRICH CENTER FOR WORKFORCE DEVELOPMENT

Project Goals

- National linked-data infrastructure for Foreign-born scientists and engineers.
- Use of state administrative data (NJ, AR, KY) and federal data sources
- Develop a plan for building a national linked-data infrastructure for Foreign-born scientists and engineers: Feasibility (Year 1)
    - Document the key issues such as measurement, data quality, and coverage
    - Technical approaches to addressing the issues
    - What could be done with additional data sources
    - Provide guidance on structuring and linking disparate data

Aspirational Data Model

Foreign-Born Related Demographics

Visa Status
Household Composition
Country of Origin

Education

Secondary Education
SSN
Contact Address
Demographics
Career and Technical Education
Program Participation
Test Scores
Dual Enrollment

Postsecondary Education
Institutional Enrollment
Demographics
Financial Aid
Term Enrollment
Academic Record
CIP
Completion
Award Type
Award Year

Employment and Income

UI Wage
SSN
Year
Quarter
Earnings
Occupation
Employer ID
Employer Industry Code

Income
Annual Earnings

Occupation

Innovation and Productivity

Patents Developed
Licensure
Business Ownership
Employees
Grant Awards
Open Source Program Development

Other Characteristics

Criminal Justice
Human Services
Home Ownership
Family Sponsorship

What We Need to Get There

- Linkage to Federal Sources
- Data Model Documentation
- Panel of Legal Experts
- Template Data Sharing Agreements

Data Sources

- State postsecondary education administrative data from Arkansas, Kentucky, and New Jersey.
- Publicly-available and restricted-use federal survey data:
  - Census Bureau: American Community Survey (ACS)
  - National Center for Education Statistics (NCES)
    - National Postsecondary Student Aid Study (NPSAS)
    - Beginning Postsecondary Students Longitudinal Study (BPS)
    - Baccalaureate and Beyond Longitudinal Study (B&B)
  - National Center for Science and Engineering Statistics (NCSES)
    - Survey of Earned Doctorates (SED)

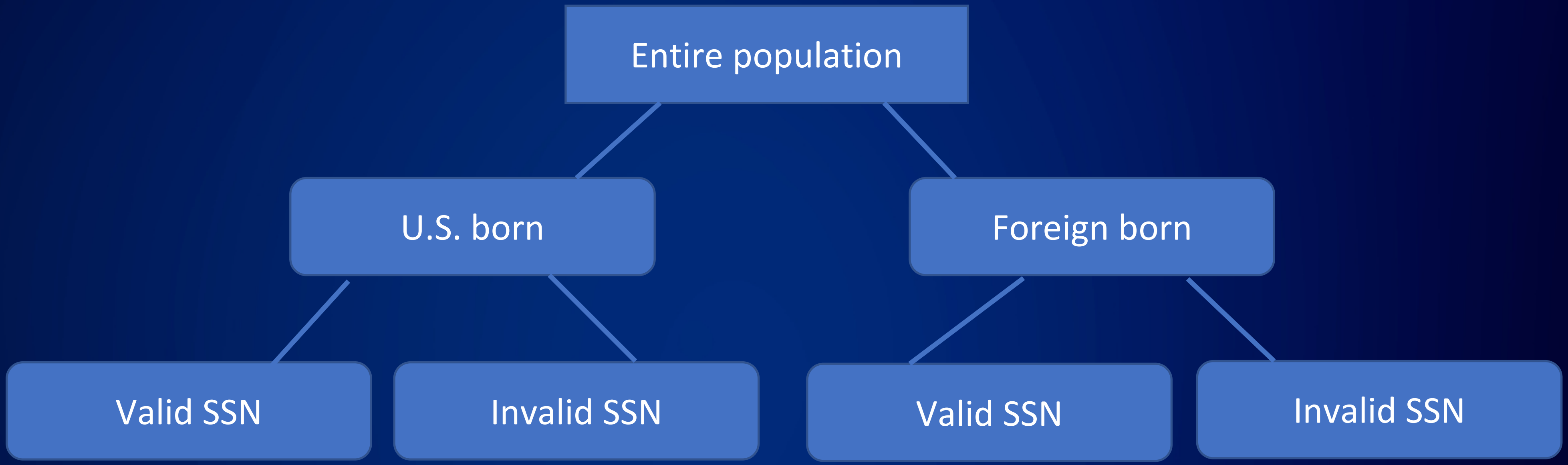
Definitions for Comparison Benchmark

Measure	State Definitions	NCES Definition	NCSES Definition	Census Definition
Population covered	Students who graduated with a U.S. degree	Students who enrolled in a U.S. degree program at the time of the survey	Students who graduated with a U.S. doctoral degree	Individuals who attained a degree
Degree origin	U.S.	Unknown	U.S.	Unknown
Time	Academic year	Academic year	Academic year	Calendar year
Location	State institution	State institution	State institution	State of residence
Foreign born	Citizenship status	Nativity (immigrant status and parent's nativity)	Citizenship status	Citizenship status
Science and Engineering Fields	CIP code of each degree level attained	23 NCES categories of majors of the program enrolled at the time of the survey	PhD field by NSF definition	Bachelor's field of degree (master's, doctoral N/A)
Degree level	Level of the degree attained	Highest degree level ever **expected** (do not know if that degree level ever attained) or the level of the **previously attained** degree	PhD level	Highest level of the degree attained

Linkage Approach

Approach	Attributes
Deterministic	• SSN • Name (First name, Last name) and SSN
Fuzzy Matching	Name (First name, Last name) and SSN
Probabilistic	Name (First name, Last name) and SSN

Linkage Approach



Challenges with Linkage

- Limited accessibility/availability of datasets
- Limited number of similar attributes between datasets
- Records can have differences in representational consistency, including differences in structure or semantics
- Records can have accurate but varying attribute values over time for the same entity

Data Linkage – Intended Benefits

- Facilitate research on the FBSE population, which would have suboptimal coverage with current record linkage approaches due to high incompleteness in Social Security Number
- FBSE linkage research and development could potentially improve the coverage and accuracy of all research and data products in the Administrative Data Research Facility
- Research could benefit other programs linking administrative data such as statewide longitudinal data systems and statistical agencies

Data Linkage – Critical Success Factors

- Performing linkage at the national scale
- Sufficient linkage accuracy for intended uses
- Minimal human intervention required
- Minimal risk of personal reidentification (private and secure)
- Transparent and explainable methods and performance

Lessons Learned

- Data governance is important to establish so that data stewards maintain ownership of their data that contributes to the infrastructure.
- State UI wage records may not provide adequate coverage
- Identifying data beyond standard administrative data to support the infrastructure
    - Consideration of confidentiality and privacy concerns given the population is critical
- Standardization of definitions across states (and inter-state collaboration) could mitigate difficulties in cross-state comparisons of FBSE population

Questions?

Contact: *yunie.le@coleridgeinitiative.org*