Preamble: Importance of Data

- Without supporting "quality" data even the best research idea will become difficult to execute.
- Quality data are necessary for ***evidence based*** and ***reproducible*** research.
- Your research converts data into ***information*** as well as in combination with a ***theoretical basis*** allows us to generalize this information into ***knowledge*** and perhaps ***policy*** suggestions.
- In the social sciences we face the dilemma that our data originate mainly from ***empirical collections*** rather than controlled experiments. Thus, to overcome ***confounding***, our research design becomes highly relevant.
- Caveat: Collecting data is ***expensive*** and bears the risks of induced ***sampling biases***.

> *To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.* – Ronald Fisher

Aggregated Data Sources

- ***Aggregated data*** provide ***summary statistics*** (perhaps including confidence intervals) of individual data records.
  The aggregation is performed ***thematically*** and/or ***geographically*** and/or ***temporally***.
- An analysis of aggregated data always bears the ***risk of committing*** an ecological fallacy, i.e., we cannot deduce based on aggregated findings that these findings hold on average on the individual level.

Freely Obtainable Public Use US Census Data

- Not requiring an ***Institutional Review Board*** (IRB) approval.
- Aside from providing research opportunities on its own, census data can provide ***context*** for many other research projects. See the hierarchical spatial organization of census data here.
- Geo-referenced databases using the GIS software Maptitude. Request a free student license here (make sure to provide enrollment proof)
- Direct downloadable, for instance, with the R-package TidyCensus by Kyle Walker at TCU (see also YouTube videos).
- Census data caveats: Disclosure avoidance and census undercounts.

Freely available Texan Data:

- Texas Demographic Center in San Antonio
- Texas Open Data Portal (see slide 8)
- Texas Natural Resource Information System

Freely Obtainable Health Statistics

- Behavioral Risk Factor Surveillance System at the CDC.
- US Cancer Statistics at the CDC or the National Cancer Institute at the NIH.
- *Many more …*

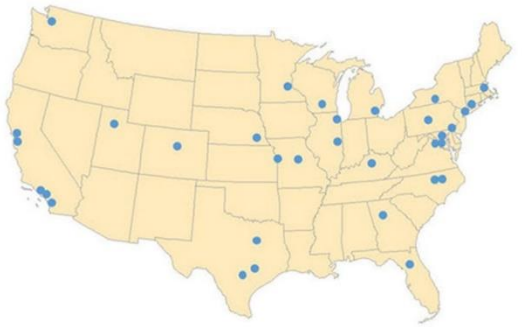
Micro Data Sources at Federal Statistical Research Data Center

- Access to ***confidential microdata*** from Census Bureau [a] censuses and surveys of business and households, [b] linked employer-employee data, and [c] administrative records from federal and state agencies.
- Advantages of microdata:
  - ***Avoidance*** of aggregation biases and ecological fallacy.
  - Access to ***individual*** household, firm, and establishment data.
  - Large scale (***small area***) geographic units.
  - ***Link*** micro data records via unique identifiers.
  - ***Link*** your research data to the secure micro data.
- A full list of demographic and economic ***micro databases*** can be found here.
- The DFW Research Data Center can also provide limited access to some ***local health data*** through its collaboration with The University of Texas Southwestern Medical Centers.

About the DFW Research Data Center

- Research data centers are affiliated either directly with universities or as local consortiums with the Federal Reserve Banks.
- The DFW Research Data Center under the leadership of Executive Director Dr. Anthony Murphy (FRB Dallas) is a consortium of several universities, public organizations in North Texas and Oklahoma.

The Census Bureau operates 33 Federal Statistical Research Data Centers



- Dr. Samuel Bondurant is the center's administrator since its inception and Census Bureau liaison.
- Dr. Michael Tiefelsdorf (EPPS – UTD) is UTD's liaison and Chair of the Consortium.
- UTD manages the administrative side of the consortium.
- Opened in 2018 with the support of an NSF research grant.
- Instrumental DFW Research Data Center founders:
  - Dr. Wenhua Di (formerly Federal Reserve Bank at Dallas)
  - Provost Dr. Inga Musselman (UTD)
  - Dr. Kurt Beron (EPPS – UTD)
  - Several researchers and administrators at consortium institutions.
- Main users at UTD are the *School of Economic, Political & Policy Sciences* and the *Jindal School of Management*.
- Another Federal Statistical Research Data Center in Texas is located at Texas A&M University at College Station with its ***satellite campus*** at The University of Texas Austin.
  Several south Texas universities partner with the Texas Research Data Center.

Restricted Use Data Application Procedure

- A ***must watch*** 2021 video by the DFW Research Data Center administrator Dr. Samuel Bondurant covers the application procedure as well as the available databases in great details.
- For detailed information please visit the FSRDC website here.
- The FSRDCs are transitioning to a standard application process (SAP) since the beginning of 2023 (links currently not active).

Project Proposal

- Proposal idea development:
  - A list of completed and active projects with their abstracts can be found here.
    This list is limited to projects under the direct supervision of the Census Bureau.
  - Center for Economic Studies (CES) workpaper series.
  - Research the available databases at the DFW Research Data Center.
- Work with your supervising professor on a research idea that ***requires*** access to the DFW Research Data Center databases.
- Document your research procedure and needed data in a brief write-up.
- Run your research idea by the administrator for the Dallas Research Data Center Dr. Samuel Bondurant.
  Dr. Samuel Bondurant will guide you through the proposal development and application process.
  A summary of the proposal preparation process can be found here.

Official Proposal Guidelines

- Detailed Census Bureau proposal guidelines.
- Detailed Internal Revenue Service proposal guidelines.

Proposal Review Criteria

- Statistical Purpose: The use of Census Bureau restricted-use data must have a statistical purpose. Data use for commercial, regulatory, or enforcement purposes is prohibited.

- Allowed-Use: The use of Census Bureau restricted-use data must be consistent with U.S. Census Bureau the Privacy Act (Title 13), and data use agreements with other data providers.
- Programmatic Benefit: All proposals must demonstrate one or more benefits to the Census Bureau.
- Statistical Disclosure Limitation: All proposals must adhere to statistically-sound and Census Bureau approved statistical disclosure limitation methodologies to protect the identity of respondents and to adhere to confidentiality provisions.
- Demonstrated Need: Proposals must demonstrate a need for restricted-use data.
- Feasibility: All research must be feasible methodologically, within the timeframe proposed, and given the expertise of the project supporters.
- Maintaining Public Trust: All proposals must support the Census Bureau's ability to maintain public trust and credibility.

Restricted Use Data Proposal Execution Procedure

Security Clearance

Security clearance must be passed to obtain a Special Sworn Status (SSS):

- It takes about 3-4 months to complete the process including background checks and fingerprinting.
- U.S. citizens and foreign national researchers must have lived 36 out of the last 60 months in the U.S.
- The researchers must have affiliations with US-based institution, e.g., a university.

Cost

- Most costs are covered by UTD out of its consortium membership fees.
  Note: For non-members the annual project access cost can exceed $25,000.
- Exceptions may apply to non-census administered databases and special soft- and hardware requirements.
- DFW Research Data Center proposal development seed grants (max. $ 10,000) may be available from UTD's *Office of Research and Innovation*.
  These proposals should lead to external funding; thus, dissertation completions will not be covered.
  However, the Census Bureau provides for U.S. citizens a dissertation mentorship program.

Project Location

- The Federal Reserve Bank in downtown Dallas provides a secure location to conduct the research.
  You will receive 24-hours access to the RDC suite at the Federal Reserve Bank.
- You are connected via terminals to a central remote Linux server to perform your research.
  It may take time to get used to the computing infrastructure and its large databases.
- No data can leave the room.
- Some data can be accessed remotely from home or your office. Consult with the RDC administrator.

Release of Results

- The project research results go through a rigorous disclosure avoidance review. This takes about 6 weeks.
- This process ensures that the project output does not reveal confidential information about individual people, households, or firms.
  Possible ***disclosure concerns*** are small geographies, low cell counts in cross-tabulations or market domination by select few firms.

Final Delivery to the Census Bureau

- Submit a ***working paper*** including an outline of the analysis procedure for ***reproducibility***.

Class Discussion

How shall EPPS graduate students proceed if they want to do thesis research using the unique resources of the DFW Research Data Center?