1-10-2025

Fundamentals of Data Collection
SURV 622/SURVMETH 622
Winter/Spring 2025

Monday 3:30-6:30pm

University of Maryland – 1208 LeFrak Hall
University of Michigan – Perry G300

Instructors:
Frederick Conrad (fconrad@umich.edu)
Brian Kim (kimbrian@umd.edu)
James Wagner (jameswag@umich.edu)

Grader: Alexis Kokoska (alexiskokoska@gmail.com)

Zoom link: https://umich.zoom.us/s/95674676673; passcode: 6222024

Fundamentals of Data Collection is a two-semester course sequence that provides a broad overview of the processes that generate data – data that are *designed*, primarily survey data, and those that are *found*, e.g., administrative data, online transactions, and social media content. The course exposes students to key theoretical ideas about data collection through lectures that are supplemented by readings drawn from the relevant scientific literature. The course introduces the practice of data collection through exercises conducted mostly outside of class. Students will gain a thorough understanding of a wide range of data as well as their relative strengths and weaknesses. The first semester focuses more on data from surveys than alternative sources; the second semester reverses this focus.

The second semester begins with a module on how survey data can be augmented with information from other sources. This is followed by a module on using repurposed data, with an emphasis on linkage with administrative records and the extraction of data from online sources. The third module addresses text classification—the process of turning written documents into usable data. Machine learning techniques are introduced for automated coding of text (open survey responses) and discussed more generally. The fourth module concerns special issues that arise in longitudinal data—repeated data from either survey or non-survey sources for the same households or businesses. The fifth and final module covers methods for evaluating the quality of survey and non-survey data.

Course Requirements:
Reading assignments for each class are available on the course website (https://umich.instructure.com/courses/726436). Please read the assigned material for each class session before class.

Grades for the course will be based on:

- Participation in class demonstrating understanding of the readings and in-class material (10% of grade);
- Practical exercises (60% of grade)
- Final exam (30% of grade)

Attending Class:

We assume that you will attend classes in person. If you would prefer to attend via Zoom for a particular class, it is your responsibility to arrange this with the instructors. JPSM students who work full time may petition the director to attend all classes remotely; once approved, please share the approval with the instructors. You can join the Zoom meeting for Surv/Survmeth 622 using the following link: https://umich.zoom.us/s/95674676673; passcode: 6222025), but please test it before using it. Please be sure your video is turned on and your audio is muted until you speak. You are free to use a virtual background. You will need a good Internet connection and may need a headset including microphone for the experience to be positive and effective. If you need assistance connecting to or testing Zoom, please contact Elisabeth Schneider (dodo@umich.edu).

Accommodations for Students with Disabilities

University of Michigan

If you believe you need an accommodation for a disability, please contact the Services for Students with Disabilities (SSD) office to help us determine appropriate academic accommodations. SSD (734-763-3000; http://ssd.umich.edu) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. Any information you provide is private and will remain confidential.

University of Maryland

To receive service, you must contact the Disability Support Services (DSS) office to register in person for services. Please call the office to set up an appointment to register with a DSS counselor. Contact the DSS office at 301-314-7682; http://www.counseling.umd.edu/DSS/

Academic conduct

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct at the University of Michigan may be found at the Rackham web site (http://www.rackham.umich.edu/policies/academic_policies/section10/) and at the University of Maryland at the web site for the Office of the President (https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf). Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

Schedule and Reading Assignments

WINTER)/SPRING SEMESTER 2025

Module 1: Augmenting survey data

January 13: Working with paradata (Conrad)
Assigned readings:

Kristin Olson. 2013. "Paradata for nonresponse adjustment," *Annals of the American Academy of Political and Social Science*, 645(1): 142-170.

Kristin Olson and Bryan Parkhurst. 2013. "Chapter 3: Collecting paradata for measurement error evaluations," in Frauke Kreuter, ed., *Improving Surveys with Paradata*, John Wiley and Sons: Hoboken, NJ, 43-72.

Optional readings:

Stephanie Eckman. 2013. "Chapter 5: Paradata for coverage research," in Frauke Kreuter, ed., *Improving Surveys with Paradata*, John Wiley and Sons: Hoboken, NJ, 97-120.

Roger Tourangeau, J. Michael Brick, Sharon Lohr and Jane Li. 2017. "Adaptive and responsive survey designs: a review and assessment," *Journal of the Royal Statistical Society A*, 180(1), 203-223.

Frederick G. Conrad, Roger Tourangeau, Mick. P Couper, and Chan Zhang. 2017. "Reducing speeding in web surveys by providing immediate feedback." *Survey Research Methods, 11,* 45-61.

Dana Garbarski, Nora Cate Schaeffer, and Jennifer Dykema. 2011. "Are interactional behaviors exhibited when the self-reported health question is asked associated with health status?." *Social Science Research,* 40(4): 1025-1036.

January 20: NO CLASS (Martin Luther King Jr. Day)

January 27: Passive and situated data collection (Conrad)
Guest speaker: Heidi Guyer, Senior Public Health Research Scientist, RTI

Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology*, *19*(1), 107-131.

Keusch, F., & Conrad, F. G. (2022). Using Smartphones to capture and combine self-reports and passively measured behavior in social research. *Journal of Survey Statistics and Methodology*, *10*(4), 863-885.

Optional readings:

S. Chatzitheochari, K. Fisher, E. Gilbert, L. Calderwood, T. Huskinson, A. Cleary, and J. Gershuny. 2017. "Using new technologies for time diary data collection: instrument design and data quality findings from a mixed-mode pilot survey," *Social Indicators Research*, 1-12.

M. Csikszentmihalyi. 2014. “Ch. 2: The experiencing sampling method,” in *Flow and the Foundations of Positive Psychology*, Springer: Dordrecht, 21-34.

N. Eagle and A.S. Pentland. 2006. “Reality mining: sensing complex social systems,” *Personal and Ubiquitous Computing*, *10*(4), 255-268.

Fingerman, K. L., Huo, M., Charles, S. T., & Umberson, D. J. (2019). Variety is the spice of life: Social integration and activity in late life. *The Journals of Gerontology, Series B: Psychological Sciences and Social Science*, *75*(2), 377-388.

Kapteyn, A., Banks, J., Hamer, M., Smith, J. P., Steptoe, A., Van Soest, A., ... & Wah, S. H. (2018). What they say and what they do: comparing physical activity across the USA, England and the Netherlands. *J Epidemiol Community Health*.

M. Raento, A. Oulasvirta, and N. Eagle. 2009. "Smartphones: An emerging tool for social scientists." *Sociological Methods and Research,* 37(3): 426-454.

S. Shiffman, A. A. Stone, and M. R. Hufford. 2008. "Ecological momentary assessment." *Annu. Rev. Clin. Psychol*. 4: 1-32.

A.A. Stone, J. E. Schwartz, J. M. Neale, S. Shiffman, C. A. Marco, M. Hickcox, J. Paty, L. S. Porter, and L. J. Cruise. 1998. "A comparison of coping assessed by ecological momentary assessment and retrospective recall." *Journal of Personality and Social Psychology*, *74* (6), 1670-1680.

R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, ... and A.T. Campbell. 2014. “Student life: Assessing mental health, academic performance and behavioral trends of college students using smartphones,” in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing,* 3-14.

Module 2: Repurposed data

February 3: Record linkage (Wagner)

Exercise 1 distributed

Robert Groves and Brian Harris-Kojetin, eds. 2017. “Chapter 2: Statistical methods for combining multiple data sources,” in *Federal Statistics, Multiple Data Sources and Privacy Protections, Next Steps*, National Academies Press: Washington, DC, 15-44.

Joshua Tokle and Stefan Bender. 2017. “Chapter 3: Record linkage,” in Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, eds., *Big Data and Social Science: A Practical Guide to Methods and Tools*, *1st Edition*, Chapman and Hall/CRC: Boca Raton, FL, 71-92.

Thomas Herzog, Fritz Scheuren and William Winkler. 2007. *Data Quality and Record Linkage Techniques*, Springer: New York.

Chapter 8, Record linkage methodology, 81-92
Chapter 9, Estimating the parameters of the Felligi-Sunter record linkage model, 93-106
Chapter 10, Standardization and parsing, 107-114
Chapter 11, Phonetic coding systems for names, 115-121
Chapter 12, Blocking, 123-130
Chapter 13, String comparator metrics for typographical error, 131-135

Optional reading:

Peter Christen and Karl Goiser. 2007. "Quality and complexity measures for data linkage and deduplication," in F. Guillet and H. Hamilton, eds. *Quality Measures in Data Mining, Studies in Computational Intelligence*, 43, Springer-Verlag: Berlin, 127–151.

February 10: Record linkage (continued) (Wagner)
Guest speaker: Trent Alexander, Associate Director, ICPSR

Joseph W. Sakshaug and Manfred Antoni. 2017. "Chapter 25: Errors in linking survey and administrative data," in P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker and B. T. West, eds., *Total Survey Error in Practice*, John Wiley and Sons: Hoboken, NJ, 557-573.

February 17: Data privacy (Wagner)

Robert Groves and Brian Harris-Kojetin. 2017. "Chapter 5: Preserving privacy using technology from computer science, statistical methods, and administrative procedures," in *Federal Statistics, Multiple Data Sources and Privacy Protections, Next Steps*, National Academies Press: Washington, DC, 79-107.

Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. "Privacy and Human Behavior in the Age of Information," *Science* 347(6221): 509-514.

Stefan Bender, Ron Jarmin, Frauke Kreuter, and Julia Lane. 2017. "Chapter 12: Privacy and Confidentiality," in Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, eds., *Big Data and Social Science: A Practical Guide to Methods and Tools*, *1st Edition*, Chapman and Hall/CRC: Boca Raton, FL. Access at https://coleridge-initiative.github.io/big-data-and-social-science/chap-privacy.html.

Optional readings:

Sharon L. Lohr, and Trivellore E. Raghunathan. 2017. "Combining survey data with other data sources," *Statistical Science* 32(2), 293--312.

Joseph W. Sakshaug and Frauke Kreuter. 2012. "Assessing the magnitude of non-consent biases in linked survey and administrative data," *Survey Research Methods* 6(2): 113-122.

Paul Ohm. 2010. "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA Law Review*, 57: 1701-1777.

Kobbi Nissim et al. 2018. "Differential privacy: A primer for a non-technical audience," Georgetown University, unpublished working paper.

February 24: Web scraping and APIs (Kim)

Exercise 1 due
Exercise 2 distributed

Cameron Neylon. 2017. "Chapter 2: Working with web data and APIs," in Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, eds., *Big Data and Social Science: A Practical Guide to Methods and Tools*, *1st Edition*, Chapman and Hall/CRC: Boca Raton, FL, 23-70.

March 3: NO CLASS (University of Michigan break)

Module 3: Classification and Text Analysis

March 10: Machine learning for social research (Kim)
Exercise 2 due
Exercise 3 distributed

Rayid Ghani and Malte Schierholz. 2017. "Chapter 6: Machine learning," in Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane, eds., *Big Data and Social Science: A Practical Guide to Methods and Tools*, *1st Edition*, Chapman and Hall/CRC: Boca Raton, FL, 147-186 (especially 161-186).

Christoph Kern, Thomas Klausch, and Frauke Kreuter. 2019. "Tree-based Machine Learning Methods for Survey Research," *Survey Research Methods*. 13(1): 73-93.

Optional readings:

Trent D. Buskirk, Antje Kirchner, Adam Eck, Curtis S. Signorino. 2018. "An Introduction to Machine Learning Methods for Survey Researchers," *Survey Practice*. 11(1): 2718

March 17: NO CLASS (University of Maryland break)

March 24: Text analysis and Large Language Models (Conrad)
Guest speaker: Mao Li

Justin Grimmer and Brandon M. Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21(3): 267-297.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. Sections 1-3.

Video: Intro to Large Language Models

Optional readings:

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. Sections 4-7.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

Küçük, D., & Can, F. (2020). Stance detection: A survey. ACM Computing Surveys (CSUR), 53(1), 1-37.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1-35.

Module 4: Longitudinal data

March 31:

Longitudinal surveys and administrative data (Wagner)

Exercise 3 due

Exercise 4 distributed

Peter Lynn and Peter J. Lugtig. 2017. “Chapter 13: Total survey error for longitudinal surveys,” in P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker and B. T. West, eds., *Total Survey Error in Practice*, John Wiley and Sons: Hoboken, NJ, 279-298.

Optional readings:

Robert Schoeni, Frank Stafford, Katherine A. McGonagle, and Patricia Andreski. 2013. "Response rates in national panel surveys," *Annals of the American Academy of Political and Social Science*, 645, 60-87.

Andrew Halpern-Manners and John Robert Warren. 2012. "Panel conditioning in longitudinal studies: Evidence from labor force items in the Current Population Survey," *Demography* 49:1499-1519.

Mario Callegaro. 2008. "Seam effects in longitudinal surveys," *Journal of Official Statistics* 24(3): 387-409.

Lance J. Rips, Frederick G. Conrad and Scott S. Fricker. 2003. "Straightening the seam effect in longitudinal surveys," *Public Opinion Quarterly* 67:522–554.

Martha Stinson, T. Kirk White and James Lawrence. 2018. "Upcoming Improvements to the Longitudinal Business Database and the Business Dynamics Statistics," unpublished working paper.

April 7: Coding open responses to survey questions (Conrad)
Guest speakers: Brandon Kopp and David Oh, Bureau of Labor Statistics

Frederick G. Conrad, Mick P. Couper and Joseph W. Sakshaug. 2016. "Classifying open-ended reports: Factors affecting the reliability of occupation codes," *Journal of Official Statistics* 32(1): 75-92.

Alexander C. Measure. 2014. "Automated coding of worker injury narratives," paper presented at the Joint Statistical Meetings, Government Statistics Section.

Optional readings:

Alexander C. Measure. 2017. "Neural networks for worker injury autocoding," Bureau of Labor Statistics, unpublished working paper.

Matthias Schonlau and Mick P. Couper. 2016. "Semi-automated categorization of open-ended questions," *Survey Research Methods* 10(2): 143-152.

Module 5: Evaluation methods

April 14: Evaluating survey designs (Wagner)

W.R. Shadish, T.D. Cook, and D.T. Campbell. 2002. "Chapter 8: Randomized experiments," in *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin: Boston, 246-278.

Howard Bloom. 2006. "The core analytics of randomized experiments," MDRC Working Papers on Research Methodology.

April 21: Assessing the quality of survey measurements (Conrad)
Exercise 4 due
Final Exam available

Optional reading:

W.R. Shadish, T.D. Cook, and D.T. Campbell. 2002. "Chapter 3: Construct validity and external validity," in *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin: Boston, 64-102.

April 28: Final exam due