CURRICULUM VITAE

Rebecca C. Steorts

Duke University, Department of Statistical Science
Email: beka@stat.duke.edu

EDUCATION

2012	**Ph.D.**, Statistics, University of Florida, Gainesville, FL Advisor: Malay Ghosh Thesis: *Bayes and and Empirical Bayes Benchmarking for Small Area Estimation,* Honorable Mention (Second Place), **Leonard J. Savage Award** for best thesis in application
2005	**M.Sc.**, Mathematical Sciences, Clemson University, Clemson, SC
2001	**B.Sc.**, Mathematics, Davidson College, Davidson, NC

PROFESSIONAL EXPERIENCE

Academic Appointments

2015 – present	**Assistant Professor** Department of Statistical Science, Duke University Affiliated faculty in the Department of Computer Science, Department of Biostatistics and Bioinformatics, Information Initiative at Duke (iiD), and the Social Science Research Institute (SSRI) at Duke
2012-2015	**Visiting Assistant Professor** Department of Statistics, Carnegie Mellon University Mentor: Stephen E. Fienberg

Other Appointments

2017 – present	Principal Researcher, Research Mathematical Statistician U.S. Census Bureau
2014 – present	Statistical Consultant, Human Rights Data Analysis Group (HRDAG)
2014	Data Science Consultant, Food and Agricultural Organization (FAO) of the United Nations
2014	Visiting Scholar at the University of Trier, Department of Economics and Social Sciences
2014	Visiting Scholar at the University of Rome “La Sapienza”, Department of Methods and Models for Economics, Geography and Finance

2013	Visiting Scientist in Summer at Census Program, U.S. Census Bureau, Washington D.C.

HONORS AND AWARDS

2017	**NSF CAREER Award**
2015	**MIT Review Magazine's 35 Innovators Under 35** Humanitarian for software design in estimating death counts for the Syrian civil war Feature video at EmTech, Boston, MA Feature piece in MIT Review, October, 2015 Feature piece by the Human Rights Data Analysis Group (HRDAG)
2013	Honorable Mention (Second Place) for best thesis in applied methodology, **Leonard J. Savage Award**, International Society for Bayesian Analysis
2010-2012	United States Census Bureau Dissertation Fellowship Program

PUBLICATIONS

Peer-reviewed Publications (all after 2015 published at Duke University)

(* student or postdoctoral fellow supervised by RCS)

1. Datta, G., Ghosh, M., **Steorts, R.** and Maples, J. (2011). Bayesian Benchmarking with Applications to Small Area Estimation, *TEST*, **20**(3) 574–588, doi:10.1007/s11749-010-0218-y.

2. **Steorts, R.** and Ghosh, M. (2013). On Estimation of Mean Squared Errors of Benchmarked Empirical Bayes Estimators, *Statistica Sinica*, **23**(2) 749–767, arxiv:1304.1600, doi:10.5705/ss.2012.053.

3. Ghosh, M. and **Steorts, R.** (2013). Two-stage Bayesian Benchmarking as Applied to Small Area Estimation, *TEST*, **22**(4) 670–687, arxiv:1305.6657, doi: 10.1007/s11749-013-0338-2.

[1] 4. **Steorts, R.**, Hall, R. and Fienberg, S. (2014). SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication, **33** 922–930: *Artificial Intelligence and Statistics (AIStats)*, arxiv:1403.0211.

[5] 5. **Steorts, R.**, Ventura, S., Sadinle, M. and Fienberg, S. (2014). Blocking Comparisons for Record Linkage, *Privacy in Statistical Databases (Lecture Notes in Computer Science 8744)*, ed. J. Domingo-Ferrer, Springer, 253-268; arxiv:1407.3191, doi:10.1007/978-3-319-11257-2_20.

[3] 6. **Steorts, R.** (2015). Entity Resolution using Empirically Motivated Priors, *Bayesian Analysis*, **10**(4) 849–875, arxiv:1409.0643, doi:10.1214/15-BA965SI, **Finalist for Lindley Prize.**

7. Wehbe, L.*, Ramdas, A.*, **Steorts, R.** and Shalizi, C.R. (2015). Regularized Brain Reading with Smoothing and Shrinkage Using Bayesian and Frequentist Methods, *Annals of Applied Statistics*, **9**:4 (1997-2022); arxiv:1401.6595, doi:10.1214/15-AOAS837.

[2] 8. **Steorts, R.**, Hall, R., and Fienberg, S.E. (2016). A Bayesian Approach to Graphical Record Linkage and De-duplication, *Journal of the American Statistical Association*, **111**:516 (1660-1672); arxiv:1312.4645, doi:10.1080/01621459.2015.1105807.

[4] 9. Zanella, G.*, Betancourt, B.*, Wallach, H., Miller, J., Zaidi, A.*, and **Steorts, R.** (2016). Flexible Models for Microclustering with Applications to Entity Resolution, *Neural Information Processing Systems (NIPS)*, 1417–1425, arxiv:1610.09780.

[6] 10. **Steorts. R**, Barnes, M.*, and Neiswanger, M.* (2017). Performance Bounds for Graphical Record Linkage, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AIStats)*, 54:298–306, Editors: Aarti Singh and Jerry Zhu, arxiv:1703.02679.

[10] 11. Durante, D., Mukherjee, N.*, and **Steorts, R.** (2017). Bayesian Learning of Dynamic Multilayer Networks, *Journal of Machine Learning Research*, **18**:43 (1-29); arxiv:608.02209.

[7] 12. Chen, B.*, Shrivastava, A., and **Steorts, R.** (2018), Unique Entity Estimation with Application to the Syrian Conflict, *Annals of Applied Statistics*, **12**:2 (1039-1067); arxiv:1710.02690, doi:10.1214/18-AOAS1163.

13. **Steorts, R.**, Tancredi, A., and Liseo, B. (2018). Generalized Bayesian Record Linkage and Regression with Exact Error Propagation. *Privacy in Statistical Databases (Lecture Notes in Computer Science 11126)*, eds. Domingo-Ferrer J., Montes F., Springer, 297-313; arxiv:1810.04808, doi:10.1007/978-3-319-99771-1_20.

14. **Steorts, R.** and Shrivastava, A. (2018). Probabilistic Blocking with An Application to the Syrian Conflict. *Privacy in Statistical Databases (Lecture Notes in Computer Science 11126)*, eds. Domingo-Ferrer J., Montes F., Springer, 297-313, Springer, 314-327; arxiv:1810.05497, doi:10.1007/978-3-319-99771-1_21.

[9] 15. Bai, L.* , Karwa. V., Slavkovic, A., and **Steorts, R.** (2018). Privacy Preserving Algorithm to Release Sparse High-dimensional Histograms, *Journal of Privacy and Confidentiality*, **8**:1; doi.org/10.29012/jpc.657.

16. Bedoya, A.D., Clement, M.E., Phelan, M., **Steorts, R.**, O'Brien, C., Goldstein, B.A. (2019). Minimal Impact of Implemented Early Warning Score and Best Practice Alert for Patient Deterioration. *Critical Care Medicine*, **47**:1 (149-55); doi.org/10.1097/ccm.0000000000003439.

17. Ghosh, Malay and **Steorts, R.** (2019). Some Variants of Constrained Estimation in Finite Population Sampling. *International Statistical Review*, **87**: 90-103, doi.org/10.1111/insr.12309.

18. Tancredi, A., **Steorts, R.**, and Liseo, B. (2019). A unified framework for de-duplication and population size estimation. *Bayesian Analysis*, In press; doi.org/10.1214/19-BA1146.

Invited Papers and Discussions (all after 2015 published at Duke University)

19. **Steorts, R.** and Ugarte, D.M. (2014). Discussion of "Single and Two-Stage Cross-Sectional and Time Series Benchmarking Procedures for SAE," *TEST*, **23** :680–685, arxiv:1405.6416, doi: 10.1007/s11749-014-0386-2.

20. Fienberg, S. and **Steorts, R.** (2014). Discussion of "Estimating the Distribution of Dietary Consumption Patterns," *Statistical Science*, **29** 1:95–96, arxiv:1403.0566, doi:10.1214/13-STS448.

21. Betancourt, B.* and **Steorts, R.** (2018). Bayesian Decision Making with Application to Resource Allocation, *Wiley StatsRef-Statistics Reference Online*, (eds. N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels). doi:10.1002/9781118445112.stat07856.

Peer-Reviewed Workshop Papers (all after 2015 published at Duke University)

22. Hall, R., **Steorts, R.** and Fienberg, S. (2012). Bayesian Parametric and Nonparametric Inference for High Dimensional Multiple Record Linkage, NIPS *Modern Nonparametric Methods in Machine Learning* workshop paper.

23. Broderick, T. and **Steorts, R.** (2014). Variational Bayes for Merging Noisy Databases, NIPS Workshops in *Advances in Variational Inference*, NIPS, arxiv:1410.4792.

24. Miller, J., Betancourt, B.*, Zaidi, A.*, Wallach, H. and **Steorts, R.** (2015). The Microclustering Problem: When the Cluster Sizes Don't Grow with the Number of Data Points, NIPS *Bayesian Nonparametrics: The Next Generation Workshop Series*, arxiv:1512.00792, **Top Five Best Workshop Papers.**

GRANTS (all after 2015 obtained at Duke University)

1. Posterior Prototyping: Bridging the Gap Between Record Linkage ad Regression, Laboratory for Analytic Sciences at North Carolina State University (NCSU), $99,614 over 01/01/2019–12/31/2019. Role: co-PI.
2. CAREER: Scalable Record Linkage through the Microclustering Property, NSF-SES – 1652431, $449,985, 05/15/17–04/15/22. Role: PI.
3. Record Linkage and Privacy-Preserving Methods for Big Data, NSF-SES – 1534412, $265,579; 07/28/15–08/31/18. Role: co-PI.
4. NCRN-MN:Triangle Census Research Network, NSF-SES SES-1131897, PI: Reiter, $4,087,370 over 10/01/2011– 09/30/2017. Role: Investigator.
5. Big Data Analytics Applied to Tracking and Cybersecurity, Laboratory for Analytic Sciences at North Carolina State University (NCSU), $105,739 over 01/01/2017–12/31/2017. Role: PI.
6. Synthetic Data Release: The Tradeoff Between Privacy and Utility of Big Data, Laboratory for Analytic Sciences at NCSU, $85,000 over 01/01/2015–12/31/2016. Role: PI.

7. Incorporating Dynamic Electronic Health Records Data Into a Model for Patient Deterioration, Collaborative Quantitative Approaches to Problems in the Basic and Clinical Sciences seed funding program, Duke University, $40,000, 2016-2017. Role: co-PI.
8. Computationally Scalable Statistical Methods for High Dimensional Record Linkage, The University of Chicago Metaknowledge Lab and the Templeton Foundation), $135,000; 2015-2016. Role: PI.
9. MIDAS Informatics Services Group—The iSG, NIH, PIs: Wagner, Espino (University of Pittsburgh), Brown (CMU), $234,936 over 8/1/14–06/01/2015. Role: Investigator.
10. Census Research Node: Data Integration, Online Data Collection, and Privacy Protection for Census 2020, NSF SES-1130706, $256,857 over 10/1/11–06/01/2015. PIs: Fienberg/Eddy. Role: Investigator.
11. Statistics and Machine Learning for Scientific Inference, NSF DMS-1043903, $433,261 over 7/15/11–06/01/2015. PI: Kass. Role: Investigator.

SOFTWARE AND PRODUCTS (all after 2015 done at Duke University)

1. `representr` (2018) Record Linkage Package for Selecting the Most Representative Record. Available at representr. Developed by Andee Kaplan, Brenda Betancourt, and Rebecca C. Steorts.

2. `cd, cora, restaurants` (2018) Record Linkage Database R Packages. Storing publicly available record linkage data sets on CRAN for public dissemination, Available at cd, cora, restaurant. Developed by Srini Sunil, Andee Kaplan, and Rebecca C. Steorts.

3. `blink` (Updated 2017) Empirical Bayes Record Linkage and De-duplication `R` software. Available at github and CRAN. Developed and coded by Rebecca C. Steorts.

4. `italy` (Updated 2017). Record Linkage Database R Packages. Storing publicly available record linkage data sets on CRAN for public dissemination, Available at github and CRAN. Developed by Rebecca C. Steorts.

5. (Updated 2017) Microclustering software `R, C++` software. Developed and coded by Brenda Betancourt, Jeffrey W. Miller, and Rebecca C. Steorts.

6. `SMERED` (Updated 2016) Record Linkage and De-duplication `Java` software (*SMERED*) with post-processing software in `R`. Developed and coded by Rebecca C. Steorts and Rob Hall. `https://bitbucket.org/resteorts/smered/`.

7. (Updated 2015) Faculty member of Models of Infectious Disease Agent Study (MIDAS) Group (joint with Pittsburgh Bioinformatics). MIDAS scientists are creating synthetic ecosystems that endeavor to contain multiple populations useful for agent based models. Developed and coded by Shannon Gallagher, Jerzy Wieczorek, Lee Richardson, Rebecca C. Steorts, and William F. Eddy, `http://www.epimodels.org/drupal/?q=node/32`.

MENTORING

Postdoctoral Advisees

Andee Kaplan	2017 – 2019 (Colorado State University)
Brenda Betancourt (Bernstein-Forster Fellowship)	2015 – 2018 (University of Florida)
Nabanita Mukherjee	2015 – 2016 (AbbVie)

Doctoral Advisees

Neil Marchant (University of Melbourne) joint with Ben Rubenstein	2018 – present
Jiurui Tang (Duke Statistical Science) joint with Jerry Reiter	2018 – present

Master's Advisees

Qiaohui Lin	2016 – 2018 (UT Austin Phd student)
Bai Li	2015 – 2017 (Duke University PhD student)
Reuben McCreanor (Statistical Science, MSEM)	2015 – 2017 (Survey Monkey)
Sepideh Mosaferi (CMU Statistics)	2012, joint with Steve Fienberg (ISU PhD student)

Undergraduate advisees

Shrey Gupta	2018 – present
Melody Jiang	2018 – present
Bassim Eledath	2018 – present
Bihan Zhuang	2018 – 2019 (Apple)
Ritika Bharati	2017
Srinivas Sunil	2017 – 2018, joint with Andee Kaplan
Angie Shen	2016 – 2017, joint with Ben Goldstein
Corey Vernot	2016 – 2017
Peter Sadosky	2014 – 2015 (Uber)
Emily Furnish (CMU Statistics)	2013 – 2014, joint with Steve Fienberg (W&M Law)
Stephanie Stern (CMU Statistics)	2013 – 2014 (University of Michican MS student)
Kairavi Chahal (CMU Statistics)	2013 (American Express)
Dahiana Jiminez (CMU Statistics)	2013

Doctoral thesis committee

Sayan Patra	2018 – 2019
Jody Heck Wortman	2016 – 2018
Ye Wang	2018
Matt Barnes (CMU Robotics), External Member	2017– 2018
Mauricio Sadinle (CMU Statistics)	2013–2015
Samuel Ventura (CMU Statistics)	2013–2015
Rafael Stern (CMU Statistics)	2013–2015
Zachary Kurtz (CMU Statistics)	2012–2014

Preliminary oral committee

Jiurui Tang (Statistical Science)	2019
Lindsay Berry (Statistical Science)	2017
Luke Calkins (Pratt School of Engineering)	2017
Jody Heck Wortman (Statistical Science)	2016

Master thesis committee

Sophie Yu (Duke Statistical Science)	2017
Bai Li (Duke Statistical Science)	2017
Reuben McCreanor (Duke Statistical Science)	2017

Undergraduate thesis committee

Bihan Zhuang (Statistical Science)	2019
Lucy Lu (Statistical Science)	2017
Peter Sadosky (CMU Statistics)	2015
Michael Pane (CMU Statistics)	2013

TEACHING EXPERIENCE AT DUKE

STA 325	Machine Learning and Data Mining (Fall 2016 –)
STA 790	Special Topics: Record Linkage (Fall 2017)
STA 360/602	Modern Advancements of Bayesian Methods (Spring 2016 –)
STA 521	Predictive Modeling (Fall 2015)

PROFESSIONAL APPOINTMENTS AND SERVICE

EDITORIAL ACTIVITIES

Editorial Boards

Associate Editor, *Journal of the American Statistical Association (ACS)* (Spring 2019 –)
Associate Editor, *Journal of Survey Statistics and Methodology* (Fall 2018 –)

Peer Review Activities *AIStats*; *American Statistican*; *Annals of Applied Statistics*; *Computational Statistics and Data Analysis*; *International Conference in Machine Learning (ICML)*; *Journal of Agricultural, Biological, and Environmental Statistics*; *Journal of the American Statistical Association*; *Data Mining and Knowledge Discovery*; *Journal of Machine Learning Research*; *Journal of Official Statistics*, *Journal of Privacy and Confidentiality*; *Journal of Multivariate Analysis*; *Journal of the Royal Statistical Society, Series A*; *Journal of Survey Statistics and Methodology*; *Journal*

of Statistical Planning and Inference; *Neural Information Processing Systems (NIPS)*; *Proceedings of the National Academia of Sciences of the United States*; *PLOS ONE*; *Statistical Methods and Applications*; *Statistics in Medicine*; *TEST*; *Springer Publishing, New York*; *Transactions of Knowledge and Data Engineering*.

Grant Review Panels

2019	National Science Foundation Panel, CISE
2018	National Science Foundation Panel, CISE/MMS

Adhoc Review Panels

2015, 2017	National Science Foundation
2018, 2019	National Science Foundation and U.S. Census Bureau

Conference, Workshop, and Professional Service

2018	Area Chair, Women in Machine Learning (WiML) Workshop, Montreal, Canada
2018	Organizing Committee, Workshop on Bayes, Big Data and Social Good, Marseille, France
2017–2018	Organizing Committee, IISA International Conference on Statistics
2017	Session chair, AISTATS
2012, 2014, 2016	Session organizer for topic-contributed sessions at JSM
2014 – 2015	Invited Program Committee Member: IEEE ICDM (International Conference on Data Mining) Workshop on Data Integration and Applications (DINA)
2014 – present	American Statistical Association's Committee on Scientific Freedom and Human Rights
2016 – present	Invited ICML, NIPS Program Committee
2015	Invited Program Committee Member, NIPS, Nonparametric Bayesian Workshop
2014 – present	Invited AISTATS Program Committee
2014	Organizing Committee of the Frontier of Hierarchical Modeling in Observational Studies, Complex Surveys, and Big Data: A Conference Honoring Professor Malay Ghosh; College Park, MD
2013, 2015	Session organizer for invited sessions at JSM

UNIVERSITY SERVICE

2018 – present	Faculty sponsor of MLBytes: Undergraduate seminar series at Duke and iiD
2018 – present	Faculty sponsor of Datathon: Coding event for undergraduates at Duke
2018	MIDS Panel on Data Science
2018 – present	Founder of Duke Undergraduate Machine Learning Day
2017 – present	Faculty Advisor of Duke Undergraduate Machine Learning
2017 – present	Faculty Advisory Board of the Duke Human Rights Center
2016 – 2019	Graduate Admission Committee, Department of Statistical Science
2015 – 2017	Tenure Track Search Committee, Department of Statistical Science

2015 – 2017	Seminar Series Coordinator, Department of Statistical Science
2015 – present	Duke Machine Learning Seminar Committee, IID
2015 – 2016	Departmental Computing Committee: Department of Statistical Science
2013 – 2015	Co-chair of Faculty Seminars, Carnegie Mellon University, Chair 2014–2015
2013 – 2015	Co-organizer of event planning committee, Carnegie Mellon University
2012 – 2013	Committee of Graduate Admissions, Carnegie Mellon University, Department of Statistics

PRESENTATIONS

Invited Short Courses

1. (2018) An Introduction to Modern Record Linkage. Short Course (1-hour short course). Duke Undergraduate Machine Learning Day.

2. (2018) An Introduction to Modern Record Linkage. Short Course (6-hour short course). CIMAT, Guanajuato, Mexico; U.S. Census Bureau. (Joint with Andee Kaplan, Breda Betancourt, and Beidi Chen).

3. (2016) *Teaching Bayes: the Essential Parts*. Short Course (3-hour short course), ISBA World Meeting, Sardinia, Italy. Sardinia, Italy.

4. (2014) *An Introduction to Privacy and Statistical Disclosure*. Short Course at the Universite Paris Dauphine, Mathematiques, Apprentissage et Sciences Humaines (MASH), 12-hour course, Paris, France.

5. (2011) *Fundamentals and Applications of Bayesian Analysis.* Short Course (12-hour course), joint with Malay Ghosh, Novartis Oncology Global Development, Florham Park, NJ.

Invited Seminars and Conference Presentations

1. (2020) Invited Talk, Bayes Comp, University of Florida, Gainesville, Florida (Upcoming)
2. (2019) Invited Talk,
3. (2019) Invited Talk, ISI World Statistics Congress, Kuala Lumpur, Malaysia (Upcoming)
4. (2019) Invited Talk, Conference on Current Trends in Survey Statistics, National University of Singapore, Singapore (Upcoming)
5. (2019) Invited Talk, Workshop on Survey Statistics, National University of Singapore, Singapore (Upcoming)
6. (2019) Invited Talk, La Sapienza, Department of Methods and Models for Economics, Geography and Finance, Rome, Italy
7. (2019) Invited Talk, University of Chicago, Department of Statistics, Chicago, IL
8. (2019) Keynote Talk, Data Sciences in an Academic Health Center setting in the 21st Century, University of Colorado, Denver, CO

9. (2018) Invited Talk, Bayesian Statistics in the Big Data Era, Centre International de Rencontres Mathematiques, Marseille, France
10. (2018) Invited Talk, Workshop in Big Data and Social Good, Centre International de Rencontres Mathematiques, Marseille, France
11. (2018) Invited Talk, Privacy in Statistical Databases (PSD), Valencia, Spain
12. (2018) Invited Talk, JSM, Vancouver, CA
13. (2018) Invited Talk, 2018 International Workshop on Survey Statistics and Big Data, Nancheng, China
14. (2018) Invited Talk, International Small Area Meeting honoring Danny Pfefferman, Shanghai, China
15. (2018) IISA International Conference on Statistics, Gainesville, FL
16. (2018) Keynote Speaker, Undergraduate Machine Learning Day, Duke University, Durham, NC
17. (2018) Faculty Seminar, University of Michigan, Department of Biostatistics
18. (2018) Plenary FCSM/WW Workshop on Quality of Integrated Data, Bureau of Labor Statistics, Washington, DC
19. (2017) IISA International Conference on Statistics, Hyderabad, India
20. (2017) Center for Survey and Research Methodology Seminar, U.S. Census Bureau, Suitland, MD
21. (2017) Faculty Seminar, University of Florida, Gainesville, FL
22. (2017) International Seminar on Data Editing, Imputation and Non Response, Plenary Talk, CIMAT, Guanajuato, Mexico
23. (2017) Faculty Seminar, Harvard University, Department of Statistics and Biostatistics,
24. (2017) INFORMS Healthcare, Rotterdam, Netherlands
25. (2017) BISP, Keynote Talk, Bocconi University, Milano, Italy
26. (2017) Data Plus and Department of Computer Science Undergraduate Seminar, Information Initiative at Duke (iiD), Duke University
27. (2017) AISTATS, Fort Lauderdale, Florida
28. (2017) Weekly Research Meeting, Laboratory for Analytic Sciences, NC State University
29. (2016) Neural Information and Professing Systems, Barcelona, Spain
30. (2016) Faculty Seminar, Center for Language and Speech Processing, Johns Hopkins University
31. (2016) Faculty Seminar, Cambridge Biostatistics Unit
32. (2016) Isaac Newton Programme: Data Linkage and Anonymization, Cambridge, UK
33. (2016) Invited Round Table Discussion, JSM, Chicago, IL
34. (2016) Invited Privacy Talk, JSM, Chicago, IL

35. (2016) Faculty Seminar, Bocconi University, Milano, Italy
36. (2016) Statistical Learning and Data Mining (SLDM) Conference, University of North Carolina at Chapel Hill, Chapel Hill, NC
37. (2016) Faculty Seminar, The University of Chicago, The Computational Institute, Chicago, IL
38. (2016) NISS Affiliates Meeting, Austin, TX
39. (2016) Late Breaking News Session, Sixth IMS-ISBA joint meeting Bayes Comp at MCM-Ski V, Lenzerheide, Switzerland
40. (2015) Spotlight presentation, Neural Information and Professing Systems, Bayesian Nonparametrics: The Next Generation Workshops, Montreal, Canada.
41. (2015) Spotlight presentation, EmTech, MIT Media Lab, Boston, MA
42. (2015) Joint Machine Learning MIT & Microsoft Research Seminar, Boston, MA
43. (2015) FOCUS Interdisciplinary Discussion Course For Duke Freshman, Durham, NC
44. (2015) Invited Session at JSM, Boston, MA
45. (2015) Faculty Seminar, University of Padua, Padua, Italy
46. (2015) ITACOSM 2015, 4th ITAlian Conference on Survey Methodology, Rome, Italy
47. (2015) IMS-Microsoft Research Workshop: Foundations of Data Science, Boston, MA
48. (2015) NCRN Spring Meeting, National Academy of Science, Washington, DC
49. (2015) Discussion of *Doing Data Science, Straight Talk from the Frontline* by Rachel Schutt, Chief Data Scientist and Senior Vice President of *NewsCorp*, Special Invited Session ENAR, Miami, FL
50. (2015) Faculty Seminar, Duke University Computer Science, Durham, NC
51. (2015) Faculty Seminar, University of North Carolina at Chapel Hill, Department of Biostatistics, Chapel Hill, NC
52. (2015) Faculty Seminar, Duke University, Department of Statistical Science, Durham, NC
53. (2015) Faculty Seminar, University of California at Berkeley, Statistics Department, Berkeley, CA
54. (2015) Faculty Seminar, University of Minnesota, Statistics Department, Minneapolis, MN
55. (2015) Faculty Seminar, Pennsylvania State University, Statistics Department, State College, PA
56. (2015) Faculty Seminar, Texas A&M University, Statistics Department, College Station, TX
57. (2015) Faculty Seminar, Florida State University, Statistics Department, Tallahassee, FL
58. (2014) Faculty Seminar, The Hopkins Department of Biostatistics, Balimore, MD
59. (2014) Faculty Seminar, Cornell University, Department of Statistical Science, Ithaca, NY
60. (2014) Faculty Seminar, Bayes in Paris Seminar, Universite Paris Dauphine, Paris, France

61. (2014) Faculty Seminar, University of Minnesota, Department of Biostatistics, Minneapolis, MN
62. (2014) Privacy and Statistical Databases Conference, Ibiza, Spain
63. (2014) Faculty Seminar, University of Trier, Department of Economics and Social Statistics, Trier, Germany
64. (2014) NSF-Census Research Network Annual Meeting, New York, NY
65. (2014) Small Area Estimation Conference, Poznan, Poland
66. (2014) Faculty Seminar, School of Economics, La Spienza, Universita di Roma, Roma, Italy
67. (2014) Computational Methods for Surveys and Census Data in the Social Sciences, University of Montreal, Montreal, Canada
68. (2014) Frontier of Hierarchical Modeling in Observational Studies, Complex Surveys, and Big Data: A Conference Honoring Professor Malay Ghosh, College Park, MD
69. (2014) Joint Conference in Data Mining in Business and Industry, Duke University, Durham, NC
70. (2014) Seventeenth International Conference on Artificial Intelligence and Statistics, Reykjavik, Iceland
71. (2014) Faculty Seminar, Columbia University, Department of Statistics, New York, NY
72. (2014) Faculty Seminar, Columbia University, Department of Applied Mathematics and Physics, New York, NY
73. (2014) Faculty Seminar, Iowa State University, Department of Statistics
74. (2014) Faculty Seminar, Carnegie Mellon University, Department of Statistics, Pittsburgh, PA
75. (2014) SAMSI Workshop: Censuses and Surveys, Washington, DC
76. (2013) Invited Small Area Estimation Talk, JSM, Montreal, Canada
77. (2013) 5th IMS New Researchers Conference, Montreal, Canada
78. (2013) Center for Survey and Research Methodology Seminar, U.S. Census Bureau, Suitland, MD
79. (2013) Faculty Seminar, UCLA, Los Angeles, CA
80. (2013) Faculty Seminar, Duke University, Department of Statistical Science, Durham, NC
81. (2012) Faculty Seminar, Pennsylvania State University, Department of Statistics, State College, PA
82. (2012) Faculty Seminar, Michigan State University, Department of Statistics and Probability, Lansing, MI
83. (2012) Fields Institute Symposium on the Analysis of Survey Data and Small Area Estimation in honour of the 75th Birthday of Professor J.N.K. Rao, Carleton University, Ottawa, Canada

84. (2012) Faculty Seminar, Carnegie Mellon University, Department of Statistics, Pittsburgh, PA
85. (2012) Faculty Seminar, University of Missouri, Department of Statistics, Columbia, MO
86. (2012) Faculty Seminar, Clemson University, Department of Mathematical Sciences, Clemson, SC
87. (2012) Faculty Seminar, Bucknell University, Lewisburg, PA
88. (2012) Faculty Seminar, Wake Forest University, Winston Salem, NC
89. (2012) Faculty Seminar, Williams College, Williams, MA
90. (2011) Center for Survey and Research Methodology Seminar, U.S. Census Bureau, Suitland, MD

Contributed Conference Presentations and Posters

1. (2017) BNP Poster Session, Paris, France
2. (2014) G70: A Celebration of Alan Gelfand's 70th Birthday, Duke University, Durham, NC
3. (2014) Neural Information Processing Systems, Advances in Variational Inference Workshop, Montreal, CA
4. (2014) Bayes 250 and O'Bayes Meeting, Duke University, Department of Statistical Science
5. (2014) The First Asian ISI Satellite Meeting on Small Area Estimation, Bangkok, Thailand
6. (2014) NSF–Census Research Network Poster Session, NISS Headquarters, Research Triangle Park, NC
7. (2013) New Directions in Monte Carlo Methods Workshop, University of Florida, Gainesville, FL
8. (2012) Women in Machine Learning Workshop, NIPS, Lake Tahoe, NV
9. (2012) Poster Presentation and NIPS Workshops Spotlight and Poster Presentation, NIPS, Lake Tahoe, NV
10. (2012) Contributed Small Area Estimation Talk, JSM, San Diego, CA
11. (2012) ISBA 2012 World Meeting, Kyoto, Japan
12. (2012) Faculty Seminar, University of Florida, Gainesville, FL
13. (2011) Contributed Small Area Estimation Talk, JSM, Miami, FL
14. (2010) Contributed Small Area Estimation Talk, JSM, Vancouver, BC
15. (2010) University of Florida, Department of Statistics Graduate Student Seminar, Gainesville, FL

Ten Selected Publications

1. [4] **Steorts, R.**, Hall, R. and Fienberg, S. (2014). SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication, **33** 922–930: *Artificial Intelligence and Statistics (AIStats)*, arxiv:1403.0211.

   Relevance: This paper proposes the first method, to our knowledge, to simultaneously perform entity resolution for more than two databases while propagating the uncertainty associated with the entity resolution process. Records are clustered to similar records using a latent variable model, where the underlying data is assumed to be corrupt, noisy, and distorted; such a distortion process is embedded into the model. The records are clustered using a data structure, called the linkage structure, which along with a Metropolis within Gibbs sampler, allows the posterior to be updated quickly. More specifically, the computational complexity of the algorithm was shown to be $O(N\max S_G S_M)$, where $N\max$ is the total number of observed records, S_G is the total number of Gibbs steps, and S_M is the total number of Metropolis steps. The methodology was shown to work well on an application to longitudinal medical data from the National Long Term Care Survey (NLTCS).

   Contribution: Steorts proposed the methodology and implemented the experimental results that appear in the paper. Hall implemented the first version of the algorithm that was then taken over by Steorts. Steorts wrote the majority of the manuscript, and Fienberg provided feedback periodically. Steorts was the corresponding author. Software can be found at Bitbucket.

2. [8] **Steorts, R.C.**, Hall, R., and Fienberg, S.E. (2016). A Bayesian Approach to Graphical Record Linkage and De-duplication, *Journal of the American Statistical Association*, **111**:516 (1660-1672); arxiv:1312.4645, doi:10.1080/01621459.2015.1105807.

   Relevance: This paper was an extension of Steorts, Hall, Fienberg (2014) and the contributions to the literature are three-fold. First, we frame the entity resolution problem simultaneously, by linking observed records to latent individuals and representing these via a data structure for categorical data. Second, our specific parametric Bayesian model when combined with our data structure allows for efficient inference and exact error rate calculation. Specifically, we lay a general framework for which linkages are most probable with associated posterior probabilities, giving a theoretical justification for the choice of the linkages. All of our methodology is illustrated on both the NLTCS and official statistics data from Italy, with comparisons to Tancredi and Liseo (2011). Third, we have suggested practical guidance to practitioners for doing record linkage using our proposed method, outlining its strengths and its shortcomings.

   The following contributions are in the JASA paper, but not in the AIStats paper: We extend upon the aforementioned framework of latent clustering models, where we provide a general framework to give the user which linkages are most probable with associated posterior probabilities, with a theoretical justification. We provide an extra set of experiments to official statistics data from Italian households, with comparisons to Tancredi and Liseo (2011), showing that our methods outperforms their method in terms of errors rates and significantly reduces the computational run time. We also illustrate a very early connection

with the uniform priors on the linkage structure, which can be represented as a prior of partitions, tying together the work of Tancredi and Liseo (2011) and Sadinle (2014). We illustrate that the uniform provides a biased estimator for the linkage structure and is sub-optimal as, depending on the choice of uniform prior, one will likely have overestimation or underestimation of the resulting sample. This eventually led to the construction of more informative priors that we consider, such as an empirically motivated one in Steorts (2015) and a BNP one in Betancourt et al. (2016). Finally, the paper contains a practitioners guide to record linkage, namely, our successes, and the challenges that are left to be solved.

Contribution: Steorts proposed the methodology and implemented the experimental results that appear in the paper. Steorts wrote the majority of the manuscript, and Fienberg provided feedback periodically. Steorts was the corresponding author. Software can be found at Bitbucket.

3. [6] **Steorts, R.** (2015). Entity Resolution using Empirically Motivated Priors, *Bayesian Analysis*, **10**(4) 849–875, arxiv:1409.0643, doi:10.1214/15-BA965SI, **Finalist for Lindley Prize.**

   Relevance: As in Steorts, Hall, Fienberg (2014, 2016), the entity resolution problem is framed by linking observed records to latent entities and representing the links and non-links via the linkage structure. The contributions of this paper are the following: The model allows one to model categorical data and text data. In order to allow for tractability of the conditional distributions, an empirically motivated approach is taken with regards to several priors. In addition, the distance function for the text data is arbitrary and can be chosen by the user to allow for utmost flexibility. Turning to the experiments, the model was evaluated on both synthetic and real data from Italian household surveys. In the case of the synthetic data, we make comparisons to semi-supervised methods that are commonly used in record linkage, namely logistic regression, Bayesian Adaptive Regression Trees (BART), and random forests, illustrating that our method does as well or better based upon the recall and precision (or false negative rate and false discovery rates). Superiority is also shown to the method of Tancredi and Liseo (2011) in terms of these error rates and also computational speed. A sensitivity analysis is then performed for all assumed hyper-parameters in the model and a discussion is given regarding the sensitivity and robustness, which is a known problem in the entity resolution literature.

   Contribution: Steorts proposed the methodology and implemented the experimental results that appear in the paper. Steorts wrote the paper and was the corresponding author. Steorts also packaged the software for public use with vignettes, which can be found at CRAN and github.

4. [9] Zanella, G., Betancourt, B., Wallach, H., Miller, J., Zaidi, A. and **Steorts, R.** (2016). Flexible Models for Microclustering with Applications to Entity Resolution, Neural Information Processing Systems (NIPS), 1417–1425, arxiv:1610.09780.

   Relevance: Conventional clustering models presume that the goal is to divide the data into a small number of high-probability clusters. Even if this is somewhat loosened to allow for a large number of clusters, each having low probability as in some Bayesian non-parametric models, every cluster still has strictly positive probability. This has two consequences as

the number of observations grows. First, every cluster is observed infinitely often. Indeed, under exchangeability, observing a data point from a cluster generally makes it more probable to observe more data from that cluster in the future. Second, because every cluster is observed infinitely often, the usual asymptotic theory applies to inferring cluster properties or parameters. Uncertainty about what each cluster is like shrinks to zero in the limit. First, we proposed the **microclustering property**, which addresses this problem. Second, we proposed a general set of models — Flexible Microclustering Models (FMMs) — which satisfy the microclustering property, and developed FMMs for record linkage tasks. Third, even the most efficient MCMC approaches are expected to converge slowly in very large and high-dimensional spaces. This is especially problematic for record linkage, since the number of latent variables to be inferred will grow with the number of records. Due to this, a new MCMC algorithm, **the chaperones algorithm**, was proposed and implemented in our paper. Finally, we illustrated our methodology, with comparisons to the Pitman-Yor process and the Dirichlet Process, for four experiments.

Contribution: Miller, Steorts, and Zanella contributed to the methodology of the paper. Betancourt, Steorts, Zaidi, and Zanella contributed in evaluating experiments for the paper. Miller, Steorts, and Wallach did the majority of the writing and Steorts was the corresponding author. Betancourt is preparing an open software package for the algorithm that will be available on `CRAN`.

5. [5] **Steorts, R.**, Ventura, S., Sadinle, M. and Fienberg, S. (2014). Blocking Comparisons for Record Linkage, *Privacy in Statistical Databases (Lecture Notes in Computer Science 8744)*, ed. J. Domingo-Ferrer, Springer, 253-268; arxiv:1407.3191, doi:10.1007/978-3-319-11257-2_20.

   Relevance: The contributions of this paper are three fold. First, we review commonly used blocking techniques in both the computer science and statistics literature. Second, we propose two new blocking techniques, one based upon random projections and another based upon locality sensitive hashing that preserves transitive connections of the graph. Third, we prove the computational complexity of all methods and give comparisons of all methods on simulated methods on errors rates and computational run time. The strengths and weakness of all methods are discussed for potential users in practice.

   Contribution: This was a fruitful collaboration between all authors, where Sadinle focused on traditional blocking methods and constructed the test data for use, Ventura focused on modern blocking methods, and both authors provided experimental results. Steorts proposed two new methods, providing experimental results. All authors contributed to the success of the writing of the manucript; Steorts was the corresponding author.

6. [10] **Steorts, R.**, Barnes, M., and Neiswanger, M. (2017). Performance Bounds for Graphical Record Linkage, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AIStats*, 54:298–306, Editors: Aarti Singh and Jerry Zhu, arxiv:1703.02679; arxiv:1703.02679.

   Relevance: The contributions of the paper are three-fold. The paper presents the first performance bounds for entity resolution, which are critically assessed using the Kullback-Leibler (KL) divergence. Specifically, we provide an upper bound using the KL divergence and a

lower bound on the minimum probability of misclassifying a latent entity in closed form for the work of Steorts, Hall, and Fienberg (2014, 2016) and Steorts (2015). Second, we make general connections to Gibbs or Kolchin partitions models (Pitman, 2006) that include a wide class of priors considered, including the work of Sadinle (2014, 2016), Tancredi and Liseo (2011), Betancourt et al. (2016), and others. Third, we illustrate the tightness of our bounds on simulated data, where the unknown parameters of all models are varied. Then we discuss what can be gained by looking at such bounds in practice.

Contribution: This was another fruitful collaboration between all authors. Steorts provided the building blocks of the methodology and proofs and wrote the majority of the manuscript. Barnes and Neiswanger provided validation of all the experimental results and wrote the experimental sections. Steorts was the corresponding author.

7. [12] Chen, B., Shrivastava, **Steorts, R.** (2018), Unique Entity Estimation with Application to the Syrian Conflict, *Annals of Applied Statistics*, To Appear.

   Relevance: Estimation of death counts and associated standard errors are of great importance in armed conflicts such as the ongoing violence in Syria, as well as historical conflicts in Guatemala, Peru, and elsewhere. Given a data set with records that have duplicated entities, our goal is to estimate the number of unique records with an overall computational cost drastically less than quadratic. To achieve this goal, we formalize it as approximating the number of connected components in a graph with sub-quadratic queries for edges. This proposal is first to describe efficient adaptive locality sensitive hashing on edges to estimate the connected components, where we show under realistic assumptions that our estimator is unbiased and has provably low variance. In addition, we provide empirical results on three real data sets as well as a case study on a subset of the Syrian conflict that the Human Rights Data Analysis Group (HRDAG) analyzed in their 2014 report, where our estimate of 190,309 reported casualties is very close to their estimate of 190,000.

   Contribution: The initial ideas for the paper were jointly done by Shrivastava and Steorts regarding the concept of applying locality sensitive hashing to entity resolution. The main ideas of this paper were jointly done by all three authors. The code and running of the experiments was done by Chen and Steorts, where the code is fully reproducible and this effort was led by Steorts, who met with Chen weekly and sometimes bi-weekly. Chen also visited Duke for a month period to work with Steorts on this collaboration. Steorts wrote the majority of the paper, and Steorts was the corresponding author.

8. [25] **Steorts, R**., Schmid, T., and Tzavdis, N. (2018). Smoothing and Benchmarking for Small Area Estimation with Application to Rental Prices in Berlin, Submitted.

   Relevance: The proposed methodology is motivated by the "rental price break" law in Germany, introduced in summer 2015, which prevents the asking rental price to be in excess of 10% of a fixed amount set by law. The aim is to estimate the average rent per square metre for 447 low geographical areas called *Lebensweltlich orientierte Räume* (LORs). Because access to survey data on rental prices owned by the Berlin Senate is not possible due to confidentiality reasons, we utilize an alternative source of data via web and print media data scraping. This alternative source of data creates the following two new challenges: reliable estimates at LOR level are not available due to the very small or zero sample sizes

and the sample from the alternative data source may fail to capture important parts of the Berlin rental market. Motivated by the aforementioned case-study, we develop constrained Bayesian small area estimation methodology for smoothing and benchmarking for estimating average rent prices at LOR-level in Berlin. In addition, the methods for constrained estimation are developed decision-theoretically and their geometric interpretation is discussed. In particular, the constrained estimators are the solutions to tractable optimization problems and can be obtained in closed-form. Mean squared errors of the constrained estimators are calculated via bootstrapping. The proposed smoothing and benchmarking techniques are free of distributional assumptions and can be applied whether the estimator is linear or non-linear, univariate or multivariate, illustrating the generality of our framework. Finally, returning to our motivating application, we illustrate our proposed methodology and contrast it against unconstrained and direct estimates while providing a discussion of our conclusions and remarks on future work.

Contribution: The motivating example and dataset (confidential and private) was provided by Schmid, while the methodology and code was developed by Steorts as well as the majority of the writing of the manuscript. Schmid analyzed the data set due to privacy and confidentiality issues regarding the data at hand. Steorts wrote all code and tested it. All authors contributed to final comments of the manuscript. Steorts was the corresponding author.

9. [15] Bai, L.* , Karwa. V., Slavkovic, A. **Steorts, R.** (2018). Privacy Preserving Algorithm to Release Sparse High-dimensional Histograms, Submitted.

   Relevance: The contributions of this paper are three-fold. First, despite recent proposals in the literature, the generation of usable DP, high-dimensional and sparse synthetic categorical data still remains a challenge. First, we propose a categorical data synthesizer that is differentially private for the release of sparse, high dimensional histograms that satisfied differential privacy (DP), while maintaining the utility of the data; then we illustrate its ability to overcome the limitations of current data synthesizers. Our proposed method is an (ϵ, δ)-DP algorithm, the *Stability Based Hashed Gibbs Sampler*, for releasing high-dimensional sparse histograms, by combining the (ϵ, δ)-DP Stability Based Algorithm (SBA) with feature hashing and Gibbs sampling. We address the first issue by approximating the empirical distribution, which is a good model with high statistical utility, by a collection of conditional distributions. These are released under DP, and Gibbs sampler is used to generate synthetic datasets from the noisy conditionals. Second, it is well known that support estimation under ϵ-DP is difficult (Wasserman and Zhou, 2010). We address this issue because the SBA allows for release of high-dimensional histograms without destroying their support. Incorporating feature hashing reduces the number and dimensionality of the resulting conditional histograms. As a result, we ensure that the sparsity pattern of the joint distribution of the high-dimensional histogram is partially preserved. These two techniques ensure impossible combinations in the data remain impossible after privacy, and it also ensures that the histograms have enough mass to obtain non-trivial utility. Third, our proposed method is not just a combination of existing methods, but a carefully thought-out solution (and the first of its kind) to the issue of high-dimensional histogram estimation. Since good utility and SBA are central to our objective of obtaining a good trade-off between accuracy loss and privacy, we also propose a utility measure for SBA. We illustrate our results on both simulated and

real data.

Contribution: The building blocks of this paper were a joint collaboration of Karwa and Steorts, while the experimental results were performed by Li and the early versions of the paper were written by Li. All authors have contributed to the writing of the paper, with Slavkovic and Steorts taking the lead. Steorts was the corresponding author.

10. [11] Durante, D., Mukherjee, N.*, and **Steorts, R.** (2018). Bayesian Learning of Dynamic Multilayer Networks, *Journal of Machine Learning Research*, **18**:43 (1-29); arxiv:608.02209.

Relevance: While the analysis of a single network is still of interest, studying more complex multidimensional relationship structures monitored across time and in multiple contexts — called layers — has become a more active area of research. In order to successfully learn and predict wiring structures associated with these dynamic multilayer networks, it is of paramount importance to develop statistical models that can incorporate the complex set of dependencies underlying the observed data, without affecting flexibility. However, the current literature lacks similar strategies, to our knowledge. Motivated by this gap, we develop a tractable Bayesian nonparametric model characterizing the edge probabilities as a function of shared and layer-specific actors' coordinates in a latent space, where these coordinates change in time via Gaussian processes. The model flexibly incorporates information within the network as well as across time and between layers, allowing key improvements in inference and prediction as shown in simulations and applications to face-to-face contacts among individuals collected hourly in multiple days.

Contribution: Durante developed the methodological core of the paper and its validation on real and simulated data. Steorts supervised the collaboration, and refined the organization and overall writing of the paper upon submission. Mukherjee provided proofreading in the revision process and checked the computational part of the paper under the supervision of Steorts and Durante. Durante was the corresponding author.