# STAT 288 Deep Statistics: AI and Earth Observations for Sustainable Development

*A joint course at Harvard University[*] together with Chalmers University of Technology[†] and Linköping[‡] Universities*

Spring 2024

**Course Description** With the aim to enhance concomitantly the rigor and efficiency of data science for scientific inquires, deep statistics emphasizes principled systems thinking throughout the entire data science ecosystem, from data conception to their postmortem examination for scientific reproducibility and replicability. This course introduces the trinity of multi-source, multi-phase, and multi-resolution statistical learning, and invites participants think through their implications and implementations in the context of AI and Earth Observations (EO) for sustainable human development. Theoretically, the course contemplates many trade-offs for 'data science for science' such as data quality vs. quantity, data privacy vs. utility, statistical vs. computational efficiencies, inferential robustness vs. relevance. Practically, it scrutinizes issues such as conceptualizing and collecting complex socioeconomic data, handling messy survey and satellite data, assessing uncertainties with black-box learning, and contemplating causal implications from AI-EO data. High-level overviews of topics such as data collection, messy data, data privacy, causality, uncertainty analysis, and deep learning will be provided on an as-needed basis.

## Course Logistics

**Time:** 9:00-11:45 Tuesdays

**Location:** Science Center Room 706

**Instructor:** Xiao-Li Meng (meng@stat.harvard.edu),
Whipple V. N. Jones Professor of Statistics and Founding Editor-in-Chief of Harvard Data Science Review

**Instructor for IAS and Chalmers course:** Adel Daoud (adel.daoud@liu.se)
Associate Professor in Analytical Sociology, Institute for Analytical Sociology (IAS), Linköping University,
Affiliated Associate Professor in Data Science and AI for the Social Sciences, Chalmers Technical University, Sweden
Fellow at the Center for Advanced Study in the Behavioral Sciences (CASBS) at Stanford University.

**Teaching Fellow:** Kyla Chasalow (kyla_chasalow@g.harvard.edu)
**Office Hours:**
Wednesday 18:00-19:00 Science Center Room 705
Friday 1:30-2:30 Science Center - Virtual Office Hour at `https://harvard.zoom.us/my/kchasalow`

**Course Assistant:** Angela Li (ayli@college.harvard.edu)

[*] Department of Statistics, Harvard University, United States

[†] Division of Data Science and Artificial Intelligence (DSAI) in Department of Computer Science and Engineering, Chalmers University of Technology, Sweden

[‡] Institute for Analytical Sociology (IAS), Linköping University, Sweden

# Summary: Practical and Foundational Issues Explored in the Course

## Motivating Context

*Sustainable development* looks to identify practices that bring balance among economic, social, technological, and environmental interests. While the United Nation's Sustainable Development goals focus on the policy issues of sustainable development, this course will lay out the data scientific issues of sustainable development for global health and living conditions, including how AI and other machine-learning (ML) algorithms are used to produce knowledge on sustainable development.

While AI and ML are experiencing something of a renaissance, their applications have focused mostly on domains in engineering, medicine, autonomous vehicles, and similar topics. Challenges remain before AI and ML methods can be used to analyze topics in sustainable development such as public health, global poverty, and socio-economic inequalities. One of the challenges is a lack of data on populations' health- and material-living conditions. Household data exists, but it is not collected frequently or widely enough for more fine-grained monitoring and prediction. A remedy to this lack of data is earth observation (EO), which is the study of planetary systems and includes the collection of data with the help of satellite technologies. These satellite data images reflect human activities as they appear from space, providing another potential source of information on poverty and living conditions. Currently, more than 80 Terabytes of data are collected every day, and these have been systematically archived by NASA, USGS and others at least since the late 1970s. Combining data from EO with AI tools and more traditional data is proving a promising way to tackle the lack of data and supporting decision making. However, these methods raise several scientific and data analytic questions, including:

1. **Messy data and confidential data:** Under what assumptions can researchers combining AI-EO methods to measure population characteristics (e.g., health and material-living conditions)? As sampling is not done using conventional statistical techniques, what is valid statistical inference? How can we handle the issue of privacy given that EO data can reveal very specific information about individuals and places?

2. **Causal inference:** To what extent can AI-EO data and methods be used for causal inference? High-dimensional analysis is now routinely implemented for other data sources—such as text and tabular data—yet questions remain about the usability of satellite images for causal inference.

3. **Policymaking:** While the primary goal of this course is foundational thinking and substantive understanding, as a secondary goal, the course will explore how scientific insights are translated into policy.

## Examining These Issues via Deep Statistics Lens

Often, scholars and policymakers analyze the three sets of questions above separately, but this can produce sub-optimal results and decisions. The *deep statistics* perspective helps us to deal with these questions jointly. In fact, virtually all data-scientific inferences and learning involve a trinity of challenges: multi-source, multi-phase, and multi-resolution.

*Multi-source* refers to situations where a single study uses data sets from different sources, which may differ spatially and temporally, contain different types of information, and come in different qualities and quantities. Because of these variations, big does not imply better, and large amount of biased data can do far more damage than smaller data sets because they can lead us to be overly confident in erroneous results. In the case of AI-EO methods, data comprises household surveys, satellite images, nightlight data, and other sources (e.g., ImageNet for transfer learning). These data are collected from different continents and years, plagued by sampling variation and biases comprising different satellite technologies, seasonality, and changing survey definitions, etc. How do we quantify the various qualities of such data, and how can we take the quantifications into account when we combine the data sources in a single study to provide as efficient and unbiased learning as possible?

*Multi-phase* refers to the way data are typically collected, preprocessed, and analyzed sequentially by parties with different goals and access to information, and often, limited communications among them. As a result, we encounter the multi-phase inference paradox: even if every party engages in a statistically valid process, the ultimate output from the collective processes can be statistically invalid. For example, household surveys are sampled with the statistical aim of representativeness of a country and satellite images for monitoring the planet, yet surveys and images are combined for training AI-EO methods to measure health and living conditions (i.e., remote surveying). But the survey data suffer from non-response, noise is injected to protect privacy, and satellite images need to be pre-processed before analysis. How can we account for these processes in our analysis, or at least have a sense of their implications and impact, especially the negative ones?

*Multi-resolution* is about the unit of analysis (aggregation) and the resolution of these units (e.g., measurement frequency, granularities of the features). While big data encourage finer resolution analyses (e.g., individualized treatments), there is typically a trade-off between data availability and resolution levels: the higher target resolution the fewer relevant data points. The goal of sustainable development is to create and protect livable societies, economies, and ecologies for all individuals (humans, animals, insects) on this planet, and therefore our analyses need to be of the appropriate resolution to be relevant for informing and evaluating various local and global policies and programs, yet our data often do not have the desired resolution. How can we reliably learn from low-resolution data about high-resolution targets? How can we assess our learning errors due to the mismatch of the resolution levels?

These complex issues make it even more challenging and important for statisticians and data scientists to find effective ways to communicate our results for policymakers. How should we optimally visualize and interpret results? How do we make sensible recommendations based on what we have learned from this "complicated trinity analysis"? We confront a setting where data scientific work matters directly for *policymaking*, which requires translating multi-learning and inference to decisions and actions.

## Course Prerequisites

This course is intended as an advanced Ph.D. level course for statistics, data science, computer science, computational social science, and similar fields. There is a cap on enrolment to ensure most effective classroom discussions. Students in other degree programs who wish to enroll should submit (via email) a maximum 400-word explanation for why they are a good fit for the course and how their knowledge is on par with advanced PhD level. Specifically, enrolled students are expected to have at least one of the following preparations:

- Foundational and theoretical proficiency: at the level of STAT210 and 211 (strong interest in statistical theory and foundational issues)
- Data analytical and computational skills: basic data science skills and being able use image data; strong skills in working with real data, data management, and computational statistics. Preferably these skills are in Python or R

# Assignments Overview

## Course Requirements

1. **Active class participation and discussion** (5%); Questions and discussion are encouraged throughout the course.

2. **Lecture notes and ChatGPT exploration exercise (10%);** Students or pair of students will be responsible for creating lecture notes based on one week's lecture and zoom recording, using ChatGPT as an aid and reflecting on the process.

3. **Coursework** (30%); The course is comprised of three modules, each with one problem set (10% each).

4. **Midterm** (25%); Students will write a 500-word research proposal and discussion. Proposals should address an open challenge chosen and formulated by the student based on a critical reading of the course lectures and materials.

5. **Final** (30%); A final 1500 to 2000-word written project and oral presentation that extends and executes the midterm proposal. With instructor approval, students may also decide to work in a different direction from the midterm proposal.

   *Those final projects that are judged to have a reasonable chance to qualify for publication will be recommended to the "Bits and Bites" column of Harvard Data Science Review (HDSR). Student contributions will undergo the normal peer-review process, and thus only the most qualified contributions will be accepted for publication. While writing a final project is mandatory, students submitting to HDSR is voluntary.*

| Assignment | Assigned | Due | Grade Percentage |
|---|---|---|---|
| Participation | NA | NA | 5% |
| Lecture notes | day of lecture | Two weeks later | 10% |
| Homework 1 | February 6, 2024 | March 5, 2024 | 10% |
| Homework 2 | March 5, 2024 | April 9, 2024 | 10% |
| Midterm Essay | NA | March 21, 2024 | 25% |
| Homework 3 | April 9, 2024 | April 30, 2024 | 10% |
| Final Presentation and Essay | NA | See below | 30% |

**Final Presentation and Essay:** The final presentations will be on Tuesday May 7, 2024 10:00am-1:00pm in Science Center Room 705 and 706. Students with final exams that conflict with the final presentation time should separately arrange to give their presentation to the instructors over zoom sometime on or before May 7. After final presentations, students have until May 9, 2024 at 11:59 pm to submit their final essay. This is so that students have the option to account for questions or comments received during the presentation.

**All assignments should be submitted via Gradescope**
https://www.gradescope.com/courses/700874

## Homework Guidelines

This course draws on several different fields, including statistics, computer science and the social sciences, and students taking this course also have varied backgrounds and interests. We consider the cross-disciplinary nature of this course a strength and we hope that you will find this perspective valuable. At the same time, we recognise that students may not have expertise in every area that the homework questions touch on. Therefore, the homeworks include problems of a different types: some are more mathematical, some emphasize reading and critical interpretation, and others require coding. Although all students will end up doing some of each, students have some flexibility to select which kinds of problems they focus on.

**[IMPORTANT]** Each homework is worth 100 points but has problems totalling to more than 100 points. Students are only expected to do 100 points worth of problems, which may come from either doing some fraction of the total problems in full or from doing sub-components of the problems (point allocations will be clearly marked). If you wish to do more of the assignment, you can gain **up to 20 bonus points** so that the maximum grade is a 120/100. Extra points on one problem set can make-up for a lower score on another. While students are free to do the entire problem set for their own learning, this is not expected, and we will not give more than 120 points. Overall, the final homework score will be out of 300.

Please note also the following:

- **Collaboration Policy:** We encourage you to think deeply both on your own and in discussions with others. However, you must write up your solutions in your own words. dditionally, you must acknowledge people (e.g., other students you conversed with intellectually on the homework) and/or resources from which you received help – this is a well established norm in scholarly research of any kind. Copying someone else's solution, or just making cosmetic changes for the sake of not copying verbatim, is not acceptable, and can be taken as evidence of academic dishonesty.

- **ChatGPT:** With the arrival of ChatGPT, there has been a heated debate about the use of such technological advances in preparing essays and other educational assignments. We believe that you are taking this course to enhance your ability as a data scientist or to gain a broader and deeper appreciation of what data science is or is not. With that in mind, we trust that you will exercise your best judgment in deciding when and how to take advantage of technologies to facilitate and enhance your learning. Our only requirement is that if you use any technology in a substantial way, you will include a note describing how you used it and in what way it helped your learning.

- **Marking:** For some homework problems there are many possible correct solutions. In marking these, we are looking for evidence that you have engaged with core issues underlying the problem at hand. Depth and clarity of thought and exposition are important. Please explain any choices you make and summarize their positive and negative consequences.

- **Late submission:** Late submissions are accepted only with valid reasons, such as a documented health issue.

# Lecture and Reading Schedule

The readings or videos listed in each week are what you should review for the following week's lecture. Although the schedule may still be tweaked a bit, the readings should be fairly fixed but some are still to be announced. We will re-upload the syllabus as TBAs are filled in so **please re-download it regularly**.

## Introduction

1. **Introduction to Deep Statistics** (Jan 23)

   - Course overview
   - Introduction to the AI and Earth Observation (AI/EO) for Poverty Observation
   - The data science big picture

   **For next week**

   - Watch David Gordon's lecture on poverty measurement [Recording here] [Slides here]
   - Read Chapter 1 and Chapter 2 sections 2.1-2.6 of Lohr (1999), available [here]
   - Sign up for lecture notes assignment - due **January 30, 2024**. Anyone not signed up by then will be arbitrarily assigned to an open slot. [Sign-up sheet]

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

## Module I: Multi-source learning and inference (January 30, February 6, 13, 20)

- **Practical issues:** Data for measuring health and material-living standards globally
- **Foundational issues:** data conceptualization and collection, data quality, data minding, issues when combining multiple sources of data (data integration).

1. **Sampling, survey design, and measurement (January 30)**

   - Overview of sampling and survey design
   - How the DHS surveys are collected
   - Discussion of David Gordon's lecture on poverty measurement

   **For next week**

   - *Statistical Paradises and Paradoxes in Big Data* - Meng (2018)
   - Enhancing (publications on) data quality: Deeper data minding and fuller data confession - Meng (2021)

2. **Thinking carefully about data collection I (February 6)**

   - Data Minding
   - The Data Defect Correction (DDC)
   - Data collection in general

   **For next week**

   - *Time series of satellite imagery improve deep learning estimates of neighborhood-level poverty in africa.* - Pettersson et al. (2023)
   - Using satellite imagery to understand and promote sustainable development - Burke et al. (2021)

- (Optional) *A scoping review on the use of machine learning in the research on the social determinants of health: trends and research prospects* - Kino et al. (2021)

3. **Thinking Carefully about data collection II (February 13)**

   - Combining DHS survey data with satellite imagery

   **For next week**

   - Read Chapters 1 and 10 of Bishop and Bishop (2023), available [here]

4. **Combining Data (February 20)**

   - **Guest Lecture**: Mohammad Kakooei (via Zoom) on deep learning with satellite data

   **For next week**

   - Watch short intro on DP big ideas `https://www.youtube.com/watch?v=pT19VwBAqKA&feature=youtu.be`
   - Read chapters 1 and 2 of Dwork and Roth (2013) available [here]

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

## Module II: Multi-phase learning and inference (February 27, March 5, 19)

- **Practical issues:** Pre-processing and managing planetary data for statistical learning and inference
- **Foundational issues:** data pre-processing, data privacy, data imputation, uncertainty quantification

1. **Combining Data, Intro to Multiphase problems (February 27)**

   - Finish talking about deep learning with satellite data
   - Combining DHS survey data with satellite imagery – some discussion of privacy in context of the poverty observation project

   **For next week (and after spring break)**

   - Read Schafer (1999) - overview of multiple imputation
   - Read sections 1 and 2 of Xie and Meng (2017)

2. **Privacy (March 5)**

   - Privacy as multi-phrase inference; introduction to differential privacy

   **For next week**

   - Enjoy your spring break! Read multiple imputation reading from last week if not already

3. **Harvard Spring Break!** No class on March 12

4. **Multiple Imputation (March 19)**

   - Multiple Imputation in theory, congeniality

   **For next week**

   - Draft paper by Kakooei et al. uploaded to canvas (please do not distribute beyond the course)

- Optional: (focus on alternatives to IWI) Daoud et al. (2023) and take a look at Burke et al. (2021) if you have not already

5. **Multiple Imputation and Deep Learning, Intro to Multi-Resolution (March 26)**

    - Multiple Imputation and Deep learning
    - Introducing multi-resolution inference in the context of the poverty observation project (creating maps with deep learning - resolution challenges)

    **For next week**

    - Read Liu and Meng (2016)
    - Optional but relevant for next week: Li and Meng (2021)

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

## Module III: Multi-resolution learning and inference (March 26, April 2, 9)

- **Practical issues:** Spatial and temporal complexities in survey and satellite data
- **Foundational issues:** data and learning resolutions, cross-resolution learning, spatiotemporal modeling, transitional inference.

1. **Multi-resolution thinking (April 2)**

    - Multi-resolution inference, individualized medicine
    - Finish lecture from last time on multi-resolution in context of poverty observation project

    **For next week**

    - First two sections of `https://plato.stanford.edu/entries/paradox-simpson/` (welcome to read more!)
    - Read Liu and Meng (2014) *Comment: A Fruitful Resolution to Simpson's Paradox via Multiresolution Inference*

2. **Simpson's paradox \ The impact of prior resolution on posterior inference (April 9)**

    - Simpson's paradox
    - Delving into the specifics of a multi-resolution problem faced in the poverty observation project

    **For next week**

    - This week's reading is also relevant for next week.
    - Read *Integrating earth observation data into causal inference: Challenges and opportunities.* Jerzak et al. (2023a). Optionally also Jerzak et al. (2023b) on *Image-based Treatment Effect Heterogeneity.*

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

## Conclusions: Causal inference and policymaking (April 16, 23)

1. **Bringing in Causality (April 16)**

   - Overview of causal inference
   - **Guest Lecture**: Connor Jerzak (in person)

   **For next week**

   - Read Meng (2014)
   - Read Daoud and Dubhashi (2023)

2. **Conclusions and Open Discussion (April 23)**

   - Wrap up Connor's lecture from last time
   - Wrap-up from both Adel and Xiao-Li
   - Discussion of major themes and takeaways of course

# Extended Reading List

* = required reading in whole or in part somewhere in the course schedule above

## General References

1. *Survey Sampling: Lohr (1999)
2. *Deep Learning, also some probability review in Chapter 2: Bishop and Bishop (2023)
3. Primer on Neural Networks Derry et al. (2023b) and Primer on CNNs Derry et al. (2023a), Interactive tool for understanding neural networks `https://playground.tensorflow.org/`
4. *Differential Privacy Dwork and Roth (2013)

## Readings on The Trinity of Inference

### Big Picture

1. **A Trio of Inference Problems that Could Win You a Nobel Prize in Statistics (If You Help Fund It)* Meng (2014)
2. **Enhancing (publications on) Data Quality: Deeper Data Minding and Fuller Data Confession* Meng (2021)

### Multi-Source (Module I)

*"Big data" typically involve data sets from multiple and very different sources, with varying qualities and quantities: big does not guarantee better, and can easily mislead.*

1. *I Got More Data, My Model is More Refined, but My Estimator is Getting Worse! Am I Just Dumb?* Meng and Xie (2014)
2. **Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election* Meng (2018)
3. *Unrepresentative Big Surveys Significantly Overestimated US Vaccine Uptake.* Bradley et al. (2021)

### Multi-Phase (Module II)

*Data are often collected, preprocessed, and analyzed sequentially by parties with different goals and with limited communications among them.*

1. *The Potential and Perils of Preprocessing: Building New Foundations* Blocker and Meng (2013)
2. **Dissecting multiple imputation from a multi-phase inference perspective: What happens when god's, imputer's and analyst's models are uncongenial?* Xie and Meng (2017)
3. Further resources on differential privacy include Wood et al. (2018), Bowen and Garfinkel (2021), Near and Darais (2023), and [this blog]

### Multi-Resolution (Module III)

*"Big data" encourage inference at a finer resolution (e.g., individualized treatments), but there are important trade-offs here, including between relevance and robustness.*

1. **A Fruitful Resolution to Simpson's Paradox via Multi-Resolution Inference.*
   Liu and Meng (2014)

2. **There is Individualized Treatment. Why Not Individualized Inference?*
   Liu and Meng (2016)

3. *A Multi-resolution Theory for Approximating Infinite-p-Zero-n: Transitional Inference, Individualized Predictions, and a World Without Bias-Variance Tradeoff.*
   Li and Meng (2021)

## Readings on Sustainable Development

### Modeling Cultures

1. **Statistical Modeling: The Three Cultures.*
   Daoud and Dubhashi (2023)

2. *Melting together prediction and inference*
   Daoud and Dubhashi (2021)

### Overviews

1. **Using satellite imagery to understand and promote sustainable development.*
   Burke et al. (2021)

2. **A scoping review on the use of machine learning in the research on the social determinants of health: trends and research prospects*
   Kino et al. (2021)

### Studies relevant to multi-phase, multi-source, and multi-resolution inference in sustainable development

1. *The impact of austerity on children: Uncovering effect heterogeneity by political, economic, and family factors in low- and middle-income countries.*
   Daoud and Johansson (2024)

2. **Integrating Earth Observation Data into Causal Inference: Challenges and Opportunities*
   Jerzak et al. (2023a)

3. *Image-based Treatment Effect Heterogeneity*
   Jerzak et al. (2023b)

4. **Using satellite images and deep learning to measure health and living standards in India*
   Daoud et al. (2023)

5. *Conceptualizing Treatment Leakage in Text-based Causal Inference*
   Daoud et al. (2022)

6. *A generalizable and accessible approach to machine learning with global satellite imagery*
   Rolf et al. (2021)

7. *Using publicly available satellite imagery and deep learning to understand economic well-being in Africa*
   Yeh et al. (2020)

8. **Time series of satellite imagery improve deep learning estimates of neighborhood-level poverty in Africa*
   Pettersson et al. (2023)

# Further Readings Suggested by Current and Past Students

## Statistics, Data Science, The Big Picture

1. *LANGUAGE: Formulas, Numbers, Words: Statistics in Prose.* Kruskal (1978)

## Sustainable Development

1. *Capitalism and extreme poverty: A global analysis of real wages, human height, and mortality since the long 16th century.* Sullivan and Hickel (2023)

# References

Bishop, C. M. and H. Bishop (2023). *Deep Learning* (1 ed.). Springer Cham.

Blocker, A. W. and X.-L. Meng (2013). The potential and perils of preprocessing: Building new foundations. *Bernoulli 19*(4), 1176 – 1211.

Bowen, C. M. and S. Garfinkel (2021). The philosophy of differential privacy. *Notices of the American Mathematical Society 68*(10), 1727–1739.

Bradley, V. C., S. Kuriwaki, M. Isakov, D. Sejdinovic, X.-L. Meng, and S. Flaxman (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature 600*, 695–700.

Burke, M., A. Driscoll, D. B. Lobell, and S. Ermon (2021). Using satellite imagery to understand and promote sustainable development. *Science 371*(6535), eabe8628.

Daoud, A. and D. Dubhashi (2021). Melting together prediction and inference. *Observational Studies 7*(1), 1–7.

Daoud, A. and D. Dubhashi (2023, jan 26). Statistical Modeling: The Three Cultures. *Harvard Data Science Review 5*(1). https://hdsr.mitpress.mit.edu/pub/uo4hjcx6.

Daoud, A., C. Jerzak, and R. Johansson (2022, July). Conceptualizing treatment leakage in text-based causal inference. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, pp. 5638–5645. Association for Computational Linguistics.

Daoud, A. and F. D. Johansson (2024). The impact of austerity on children: Uncovering effect heterogeneity by political, economic, and family factors in low- and middle-income countries. *Social Science Research 118*, 102973.

Daoud, A., F. Jordán, M. Sharma, F. Johansson, D. Dubhashi, S. Paul, and S. Banerjee (2023). Using satellite images and deep learning to measure health and living standards in india. *Social Indicators Research 167*, 475–505.

Derry, A., M. Krzywinski, and N. Altman (2023a). Convolutional neural networks. *Nature Methods 20*(9), 1269–1270.

Derry, A., M. Krzywinski, and N. Altman (2023b). Neural networks primer. *Nat Methods 20*, 165–167.

Dwork, C. and A. Roth (2013). *The Algorithmic Foundations of Differential Privacy*, Volume 9 of *Foundations and Trends® in Theoretical Computer Science*. Hanover, MA: Now Publishers Inc.

Jerzak, C. T., F. Johansson, and A. Daoud (2023a). Integrating earth observation data into causal inference: Challenges and opportunities. https://arxiv.org/abs/2301.12985.

Jerzak, C. T., F. D. Johansson, and A. Daoud (2023b, 11–14 Apr). Image-based treatment effect heterogeneity. In M. van der Schaar, C. Zhang, and D. Janzing (Eds.), *Proceedings of the Second Conference on Causal Learning and Reasoning*, Volume 213 of *Proceedings of Machine Learning Research*, pp. 531–552. PMLR.

Kino, S., Y.-T. Hsu, K. Shiba, Y.-S. Chien, C. Mita, I. Kawachi, and A. Daoud (2021). A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM - Population Health 15*, 100836.

Kruskal, W. (1978). Language: Formulas, numbers, words: Statistics in prose. *The American Scholar 47*(2), 223–229.

Li, X. and X.-L. Meng (2021). A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association 116*(533), 353–367.

Liu, K. and X.-L. Meng (2014). Comment: A fruitful resolution to simpson's paradox via multiresolution inference. *The American Statistician 68*(1), 17–29.

Liu, K. and X.-L. Meng (2016). There is individualized treatment. why not individualized inference? *Annual Review of Statistics and Its Application 3*(1), 79–111.

Lohr, S. L. (1999). *Sampling : design and analysis.* Pacific Grove, CA: Duxbury Press.

Meng, X.-L. (2014). A trio of inference problems that could win you a nobel prize in statistics (if you help fund it). In X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, and J.-L. Wang (Eds.), *Past, Present, and Future of Statistical Science*, pp. 537–562. Boca Raton, FL: CRC Press.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics 12*(2), 685 – 726.

Meng, X.-L. (2021). Enhancing (publications on) data quality: Deeper data minding and fuller data confession. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 184*(4), 1161–1175.

Meng, X.-L. and X. Xie (2014). I got more data, my model is more refined, but my estimator is getting worse! am i just dumb? *Econometric Reviews 33*(1-4), 218–250.

Near, J. P. and D. Darais (2023, December). Guidelines for evaluating differential privacy guarantees. Special publication, National Institute of Standards and Technology (NIST).

Pettersson, M. B., M. Kakooei, J. Ortheden, F. D. Johansson, and A. Daoud (2023). Time series of satellite imagery improve deep learning estimates of neighborhood-level poverty in africa. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Rolf, E., J. Proctor, T. Carleton, I. Bolliger, V. Shankar, M. Ishihara, B. Recht, and S. Hsiang (2021). A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications 12*, 4392.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research 8*(1), 3–15. PMID: 10347857.

Sullivan, D. and J. Hickel (2023). Capitalism and extreme poverty: A global analysis of real wages, human height, and mortality since the long 16th century. *World Development 161*, 106026.

Wood, A., M. Altman, A. Bembenek, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D. R. OBrien, T. Steinke, and S. Vadhan (2018). Differential privacy: A primer for a non-technical audience. *Vanderbilt Journal of Entertainment & Technology Law 21*(1), 209–275.

Xie, X. and X.-L. Meng (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when god's, imputer's and analyst's models are uncongenial? *Statistica Sinica 27*(4), 1485–1545.

Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications 11*, 2583.