

Special Year on Large Language Models and Transformers, Part 1

Monday, Aug. 26 – Friday, Dec. 13, 2024



Methods innovation introduced



Chris Bail

- **Automated** annotation
- **Synthetic** data
- AI as agent / researcher



Seth Spielman

- New form of **measurement**
- Fusion of data sources
- Shift to model-based information



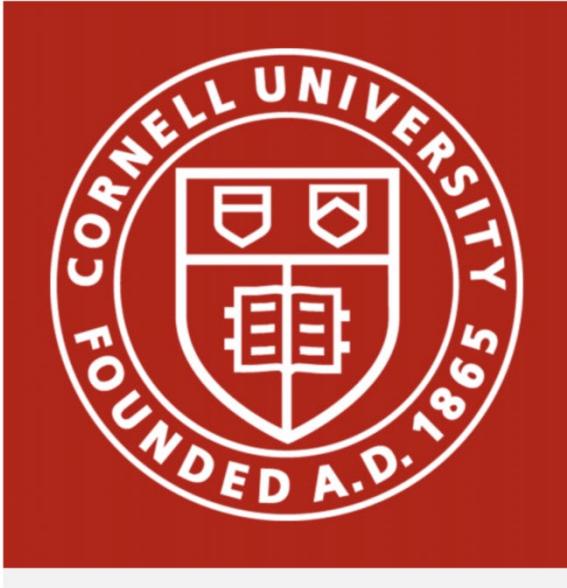
Susan Athey

- Precision treatments
– isolate control
- **New prediction** with context
- Forgo the standardize response

Words of caution

- **Bias**
- Replicability / Junk Science
- Environmental Impact
- **Measurement a moving target**
- AI not a data base
(next token prediction)
- **No understanding** – black box
- Modification for social science
(fine-tuned)
- Data availability

Thought Summit on The Future of Survey Science:



Government, industry, and academic behaviors, technology, and public trust. Thought Summit brings together experts in data science, and AI. Our goal is to develop cost-effective surveys.

Start Date: September 22, 2024

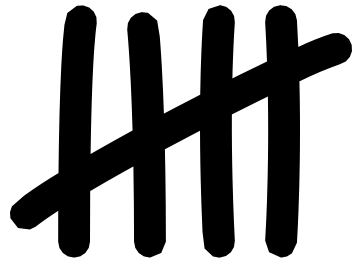
End Date: September 25, 2024

All Day Event? Yes

Location: Cornell University

Event URL:

- **Benchmarks** needed for traditional and new methods.
- **Team science** to get the models right.
- **LLMs for social science** (fine-tuned, instruction tuning on base model).



Benchmarks

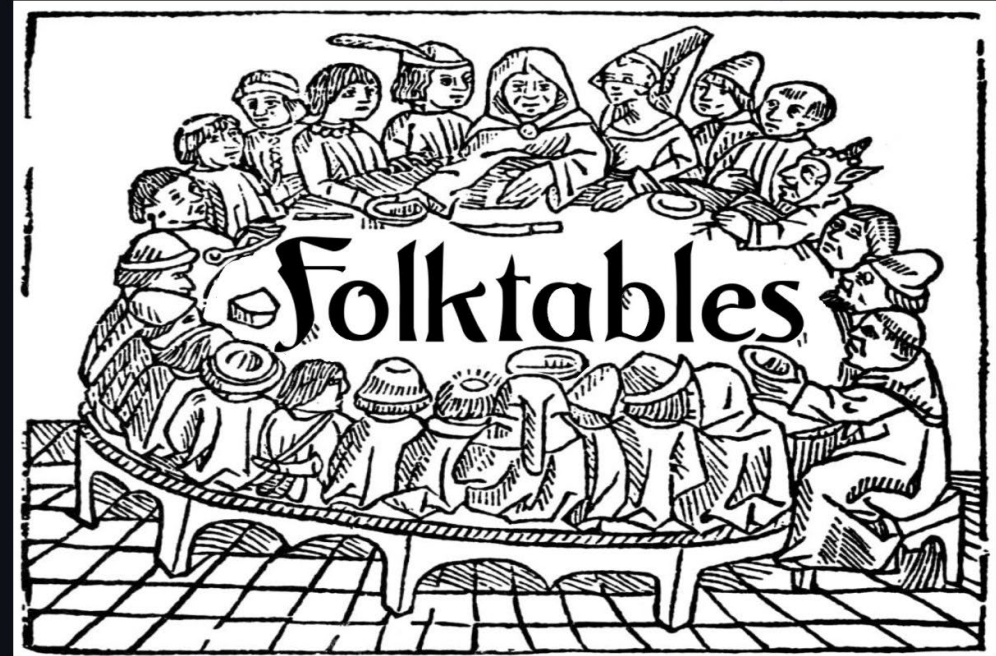


ICPSR

NLST

README.md

License MIT Downloads 18k pypi v0.0.12



Folktables is a Python package that provides access to datasets derived from the US Census, facilitating the benchmarking of machine learning algorithms. The package includes a suite of pre-defined prediction tasks in domains including income, employment, health, transportation, and housing, and also includes tools for creating new prediction tasks of interest in the US Census data ecosystem. The package additionally enables systematic studies of the effect of distribution shift, as each prediction task can be instantiated on datasets spanning multiple years and all states within the US.

Why the name? Folktables is a neologism describing tabular data about individuals. It emphasizes that data has the power to create and shape narratives about populations and challenges us to think carefully about the data we collect and use.

BRINGING SURVEY METHODOLOGY TO MACHINE LEARNING

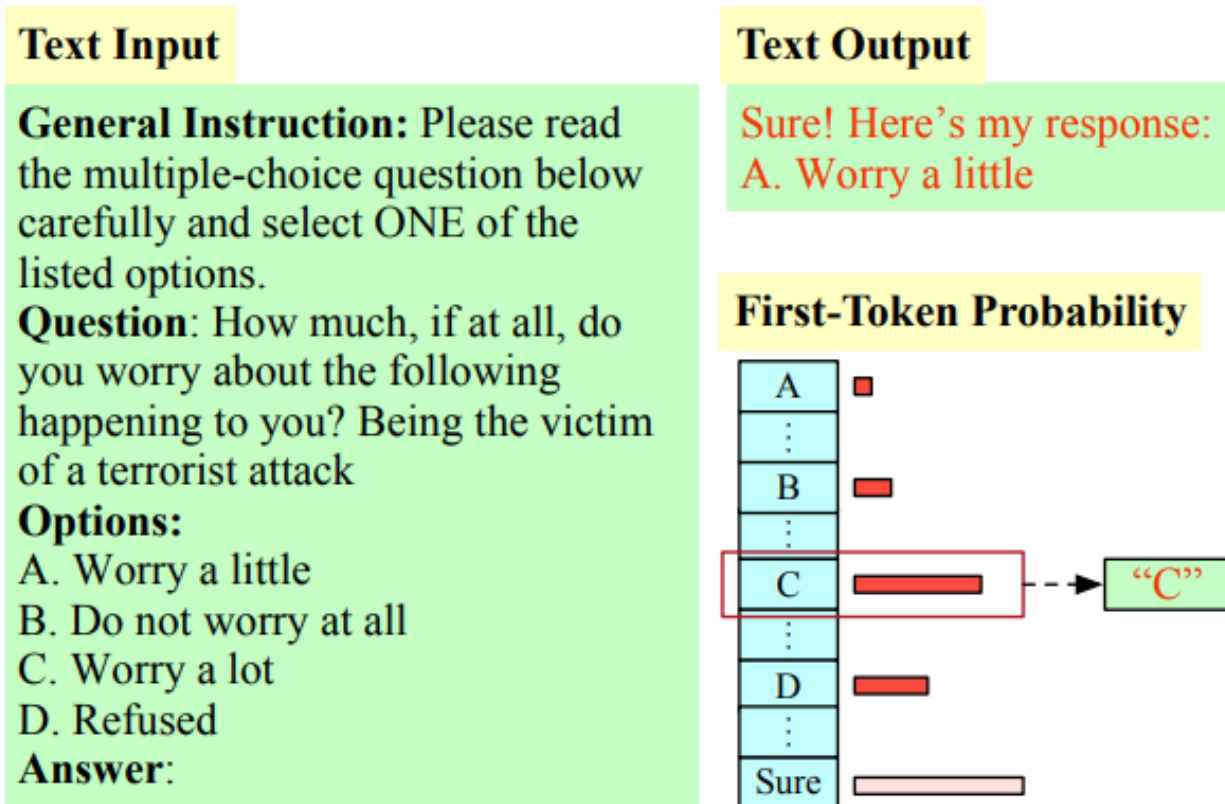
Stephanie Eckman Christoph Kern, Jacob Beck, Bolei Ma, Rob Chew, Frauke Kreuter



“The bias I am most nervous about is the bias of the human feedback raters”

Sam Altman
March 25 2023 “The Lex Fridman Podcast”





**“My Answer is C”: First-Token Probabilities
Do Not Match Text Answers in Instruction-Tuned Language Models**

Xinpeng Wang^{1,2} Bolei Ma^{1,2} Chengzhi Hu¹ Leon Weber-Genzel¹ Paul Röttger³

Frauke Kreuter^{1,2} Dirk Hovy³ Barbara Plank^{1,2}

¹LMU Munich, Munich, Germany

²Munich Center for Machine Learning (MCML), Munich, Germany

³Bocconi University, Milan, Italy

<https://arxiv.org/pdf/2402.14499> or
[Findings of the Association for Computational Linguistics ACL 2024](#)

Figure 1: Example of LLM’s *mismatch* between first-token probability prediction (“C”) and text output (“A”).



- *Who should be funded?* Unclear **what scientific approach** can and should be used to evaluate research proposals using LLMs / of LLMs.
- *What should be reported?* Standards for **transparency** of models and research (reliability / generalizability).

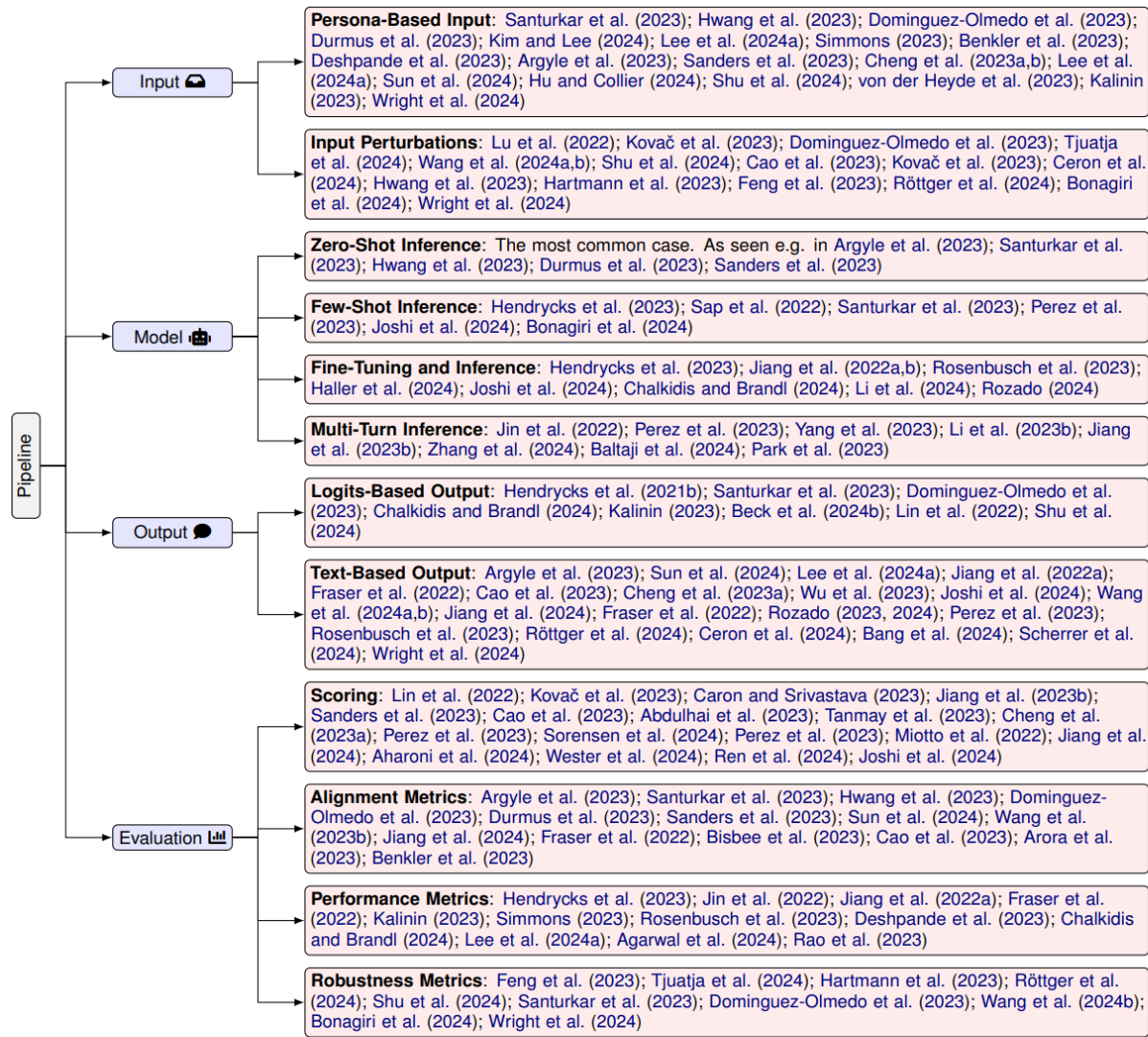


Figure 1: A taxonomy of evaluation pipeline across input 🗨️ → model 🏠 → output 🗨️ → evaluation 📊.

The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models

Bolei Ma* ^{LMU, m} Xinpeng Wang* ^{LMU, m} Tiancheng Hu ^U Anna-Carolina Haensch ^{LMU, U}

Michael A. Hedderich ^{LMU, m} Barbara Plank ^{LMU, m, U} Frauke Kreuter ^{LMU, m, U}

^{LMU}LMU Munich ^mMunich Center for Machine Learning ^UUniversity of Cambridge

^UUniversity of Maryland, College Park ^UITU Copenhagen

bolei.ma@lmu.de, xinpeng@cis.lmu.de

<https://arxiv.org/pdf/2406.11096> or EMNLP 2024

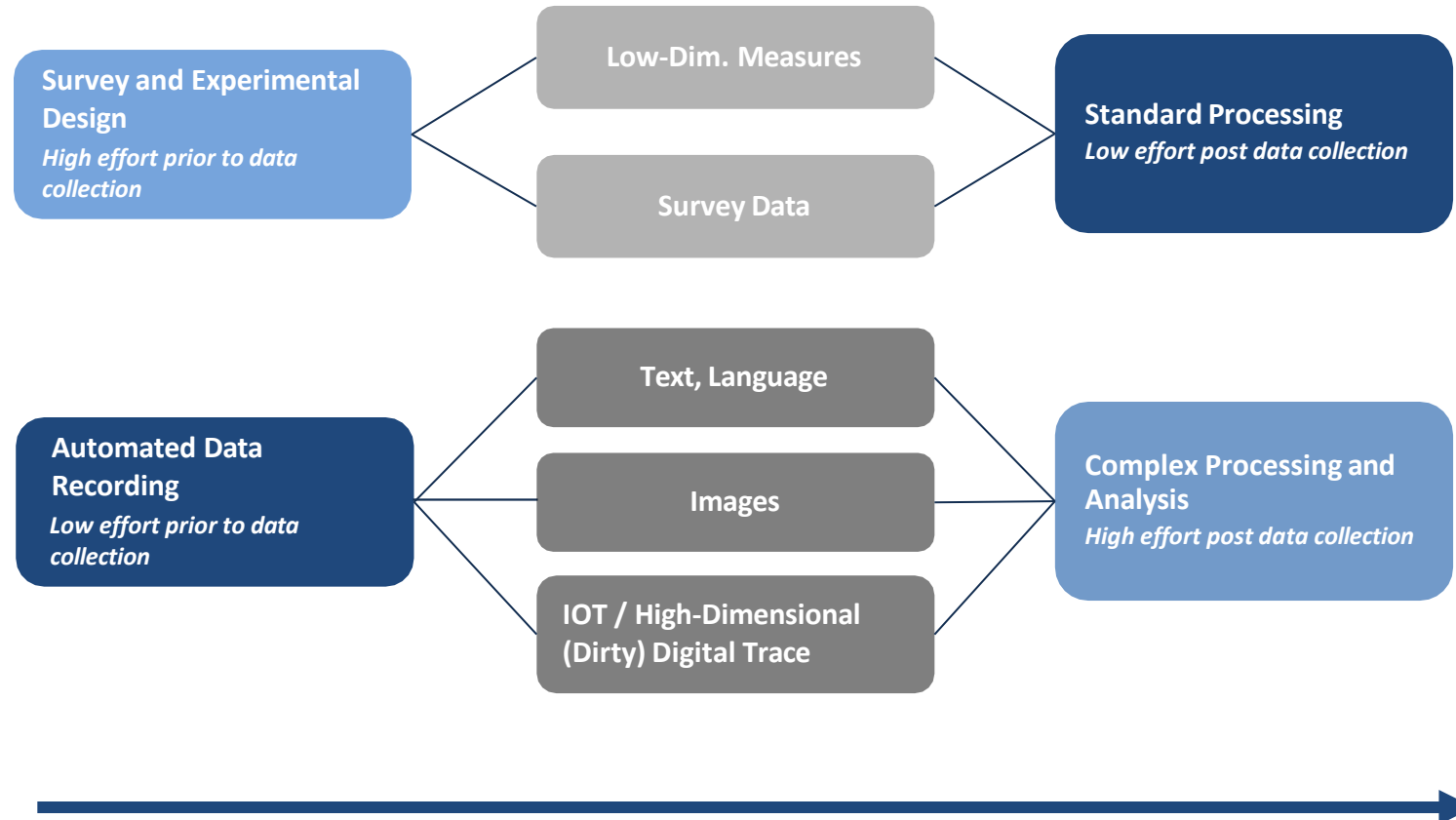
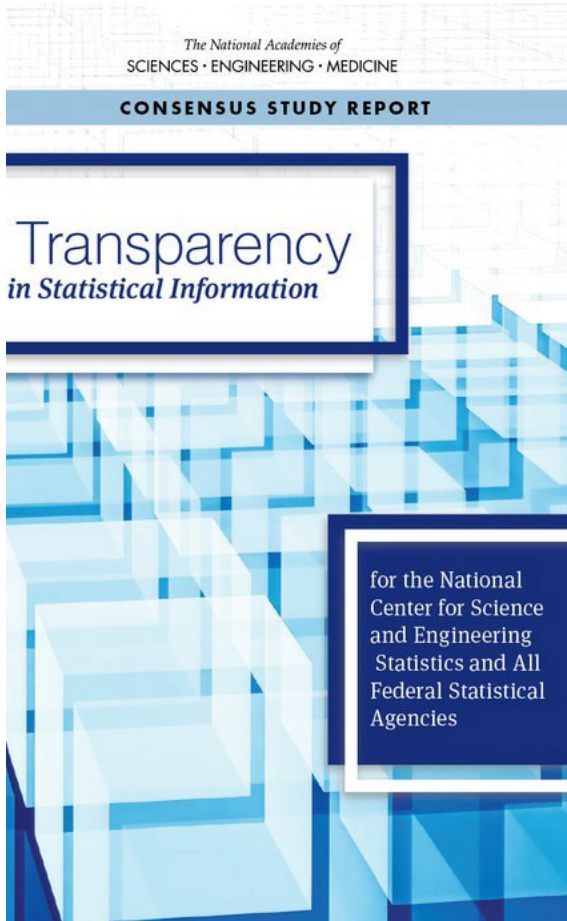
Position: Insights from Survey Methodology can Improve Training Data

Stephanie Eckman ¹ Barbara Plank ^{2,3,4} Frauke Kreuter ^{5,4,1,6}

<https://arxiv.org/pdf/2403.01208> or ICML 2024



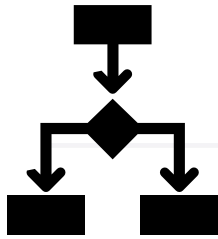
Transparency





Another way to think of AI in analysis

- Generative AI as **assistant** rather than replacement.
- Coding help
- Translation of legacy code
- Ideation
- What is a good data analysis?



Assistants: Job Classification

occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys

Jan Simson ¹, Olga Kononykhina¹, and Malte Schierholz ¹

¹ Department of Statistics, Ludwig-Maximilians-Universität München, Germany  Corresponding author

DOI: [10.21105/joss.05505](https://doi.org/10.21105/joss.05505)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Chris Vernon](#) 

Reviewers:

- [@welch16](#)
- [@danielruss](#)

Submitted: 30 March 2023

Published: 24 August 2023

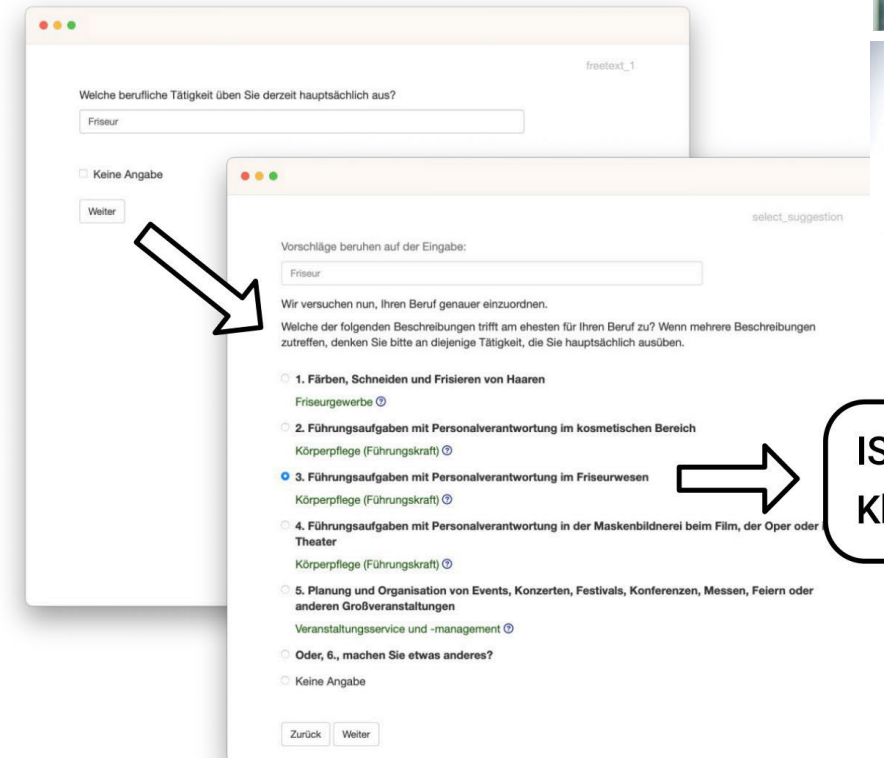
License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

People earn a living a multitude of ways which is why the occupations they pursue are almost as diverse as people themselves. This makes quantitative analyses of free-text occupational responses from surveys hard to impossible, especially since people may refer to the same occupations with different terms. To address this problem, a variety of different classifications have been developed, such as the International Standard Classification of Occupations 2008 (ISCO) ([ILO, 2012](#)) and the German Klassifikation der Berufe 2010 (KldB) ([Bundesagentur für Arbeit, 2011](#)), narrowing down the amount of occupation categories into more manageable numbers in the mid hundreds to low thousands and introducing a hierarchical ordering of categories. This leads to a different problem, however: Coding occupations into these standardized categories is usually expensive, time-intensive and plagued by issues of reliability.

Here we present a new instrument that implements a faster, more convenient and interactive occupation coding workflow where respondents are included in the coding process. Based on the respondent's answer, a novel machine learning algorithm generates a list of suggested occupational categories from the [Auxiliary Classification of Occupations](#) ([Schierholz, 2018](#)), from which one is chosen by the respondent (see [Figure 1](#)). Issues of ambiguity within occupational categories are addressed through clarifying follow-up questions. We provide a comprehensive toolbox including anonymized German training data and pre-trained models without raising privacy issues, something not possible yet with other algorithms due to the difficulties of anonymizing free-text data.



Welche berufliche Tätigkeit üben Sie derzeit hauptsächlich aus?

Friseur

☐ Keine Angabe

Weiter

Vorschläge beruhen auf der Eingabe:

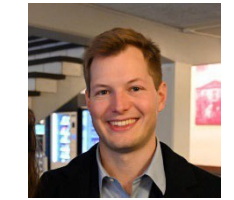
Friseur

Wir versuchen nun, Ihren Beruf genauer einzuordnen.

Welche der folgenden Beschreibungen trifft am ehesten für Ihren Beruf zu? Wenn mehrere Beschreibungen zutreffen, denken Sie bitte an diejenige Tätigkeit, die Sie hauptsächlich ausüben.

- ☐ 1. Färben, Schneiden und Frisieren von Haaren
Friseurgewerbe 
- ☐ 2. Führungsaufgaben mit Personalverantwortung im kosmetischen Bereich
Körperpflege (Führungskraft) 
- ☒ 3. Führungsaufgaben mit Personalverantwortung im Friseurwesen
Körperpflege (Führungskraft) 
- ☐ 4. Führungsaufgaben mit Personalverantwortung in der Maskenbildnerei beim Film, der Oper oder Theater
Körperpflege (Führungskraft) 
- ☐ 5. Planung und Organisation von Events, Konzerten, Festivals, Konferenzen, Messen, Feiern oder anderen Großveranstaltungen
Veranstaltungsservice und -management 
- ☐ Oder, 6., machen Sie etwas anderes?
- ☐ Keine Angabe

Zurück Weiter



ISCO-08: 5141
KldB (2010): 823

Comments to Machine Learning and AI in Econometrics. Susan Athey.



Chris Bail

- Automated annotation
- Synthetic data
- AI as agent / researcher



Seth Spielman

- New form of data – data?
- Fusion of data sources
- Shift to model-based information



Susan Athey

- Precision treatments
– isolate control
- New prediction with context
- Forgo the standardize response

- *Benchmarks*
- *Specialized LLMs*
- *Transparency/Methodology*
- *Team science/Synergies*
- *Assistant as assistant*



October 7 -- 8:30-5pm



University of Maryland / AI Interdisciplinary Institute of Maryland



[LOGIN](#)

[JOIN](#)

[RENEW](#)

[in](#)

[f](#)

[twitter](#)

[MEMBERSHIP](#)

[EVENTS](#)

[STANDARDS AND ETHICS](#)

[PUBLICATIONS & RESOURCES](#)

[MEDIA](#)

Program

5 Short courses

6 Small group discussions

2 Keynotes

2 Panel discussions

Participants

Abt, Adobe, Census Bureau, D3, Deloitte, DOJ, Forbes, Goodwill, Google, IBM, ICF, National Academies, NORC, ORB, Pew, RTI, Rutgers, Sinai, SRRS, Survey 160, World Bank, UMD, RAND, Uni Mannheim, Uni Michigan, Uni Munich, Valoray, WBA

.. you (?)

The leading association of public opinion and survey research professionals.

[Learn More](#)

