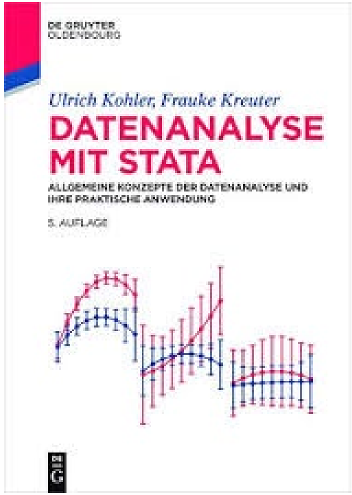
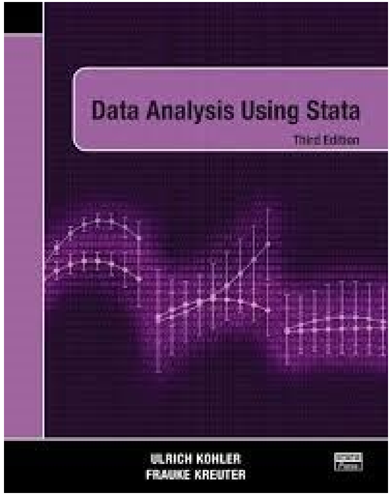
Cocreating with AI: LLMs as data scientist agents





Data Analysis
Using AI

4.4.4 LLMs as programming assistant for Python

As outlined in Chapter 1.3.15, Large Language Models (LLMs) are valuable tools for coding assistance, particularly with languages like Python, where they have shown considerable efficacy. Suppose we are in the Python environment within Stata, having imported the variable 'ybirth' using the sfi module. We seek to perform further computations on this data and are turning to Python programming for assistance. We cannot assure that the prompts and requests demonstrated here will be effective with every LLM or under all circumstances, but we will point out several key features of the prompt that have helped us in the past to get better results:

I am working with the already existing variable ybirth and I want to calculate its mean in Python without any additional modules. The code should all be put into one code section, simple, well-commented, and easy to understand, and reusable. It is important that it is efficient and follows best practices. The goal is to create a solid foundation for analysis that can be easily understood and adapted by others.

ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Perspective on Data Science

Roger D. Peng[1] and Hilary S. Parker[2]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA; email: rdpeng@jhu.edu

[2]Independent Consultant, San Francisco, California 94102, USA

Annu. Rev. Stat. Appl. 2022. 9:1–20

First published as a Review in Advance on September 16, 2021

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040220-013917

Keywords

data analysis, analytic iteration, design thinking, reproducibility, systems engineering

Abstract

The field of data science currently enjoys a broad definition that includes a wide array of activities which borrow from many other established fields of study. Having such a vague characterization of a field in the early stages might be natural, but over time maintaining such a broad definition becomes unwieldy and impedes progress. In particular, the teaching of data science is hampered by the seeming need to cover many different points of interest. Data scientists must ultimately identify the core of the field by determining what makes the field unique and what it means to develop new knowledge in data science. In this review we attempt to distill some core ideas from data science by focusing on the iterative process of data analysis and develop some generalizations from past experience. Generalizations of this nature could form the basis of a theory of data science and would serve to unify and scale the teaching of data science to large audiences.

arXiv:2408.07461v1 [cs.AI] 14 Aug 2024

Problem Solving Through Human-AI Preference-Based Cooperation

Subhabrata Dutta *subhabrata.dutta@tu-darmstadt.de*
TU Darmstadt

Timo Kaufman *timo.kaufmann@lmu.de*
LMU Munich

Goran Glavaš *goran.glavas@uni-wuerzburg.de*
University of Würzburg

Ivan Habernal *ivan.habernal@ruhr-uni-bochum.de*
RU Bochum

Kristian Kersting *kersting@cs.tu-darmstadt.de*
TU Darmstadt

Frauke Kreuter *frauke.kreuter@lmu.de*
LMU Munich

Mira Mezini *mezini@informatik.tu-darmstadt.de*
TU Darmstadt

Iryna Gurevych *iryna.gurevych@tu-darmstadt.de*
TU Darmstadt

Eyke Hüllermeier* *eyke@ifi.lmu.de*
LMU Munich

Hinrich Schuetze* *hhschutze@gmx.de*
LMU Munich

Abstract

While there is a widespread belief that artificial general intelligence (AGI) – or even superhuman AI – is imminent, complex problems in expert domains are far from being solved. We argue that such problems require human-AI cooperation and that the current state of the art in generative AI is unable to play the role of a reliable partner due to a multitude of shortcomings, including inability to keep track of a complex solution artifact (e.g., a software program), limited support for versatile human preference expression and lack of adapting to human preference in an interactive setting. To address these challenges, we propose HAI-Co^2, a novel human-AI co-construction framework. We formalize HAI-Co^2 and discuss the difficult open research problems that it faces. Finally, we present a case study of HAI-Co^2 and demonstrate its efficacy compared to monolithic generative AI models.

1 Introduction

Despite the impressive achievements of generative AI spearheaded by Large Language Models (LLMs), Language Models and code models (Lozhkov et al., 2024; Wang et al., 2021), multiple recent investi

*Shared senior authorship

Tukey defined data analysis as
“procedures for analyzing data,
techniques for interpreting the results of such procedures,
ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data” (**Tukey 1962**, p. 2)

The Excitement

Tukey defined data analysis as
“procedures for analyzing data,
techniques for interpreting the results of such procedures,
ways of planning the gathering of data to make its analysis
easier, more precise or more accurate, and all the
machinery and results of (mathematical) statistics which
apply to analyzing data” (**Tukey 1962**, p. 2)

AI analyzing data

9:17

ChatGPT >

I want to use the Titanic dataset to explore with my students that question whether females and children were prioritized in being saved in the Titanic accident. Can you load the data and show me how to do this in Stata?

Yes! The Titanic dataset is a great resource for exploring whether women and children were prioritized in survival. I'll guide you through loading the dataset and conducting the analysis in Stata.

Step 1: Load the Dataset in Stata

You can download the Titanic dataset from sources like Kaggle or use Stata's built-in datasets (if available). The dataset typically includes variables such as:

9:17

ChatGPT >

- `sex`: Male or Female
- `age`: Age of the passenger

If you have a CSV file (e.g., `titanic.csv`), load it into Stata using:

```
import delimited "titanic.csv", clear
```

Step 2: Check the Data

Verify that your dataset loaded correctly:

```
describe
list survived sex age if _n <= 10  //
Preview first 10 rows
```

If age has missing values, check:

```
sum age, detail
```

Step 3: Analyze Survival Rates

To explore whether women and children were prioritized, we can compare

2. Compare Survival by Age (Defining "Children")

First, create a variable for children (e.g., age < 18):

```
gen child = age < 18
```

Then check survival by gender and child status:

```
tab child survived, col
tab sex child survived, col
```

3. Logistic Regression: Were Women and Children More Likely to Survive?

A logistic regression can help quantify the likelihood of survival:

4. Interaction Between Gender and Being a Child

To check if the effect of being a child differs by gender:

```
logit survived i.sex##i.child i.pclass
```

Step 4: Interpret Results

- If the coefficient for `sex` is positive and significant, women had higher odds of survival.
- If `child` has a positive coefficient, children were prioritized.

AI analyzing data

9:18

ChatGPT >

What refinement should I do?

Good question! Here are some refinements that could improve your analysis:

1. Adjust for Socioeconomic Status (Passenger Class)

- We know that **passenger class (pclass)** played a major role in survival.
- Higher-class passengers had better access to lifeboats.

Refinement: Add pclass as a Control Variable

Modify your logistic regression to adjust for class differences:

```
logit survived i.sex##i.child i.pclass
```

- This helps separate the effects of

9:18

ChatGPT >

Refinement: Visualize Survival by Class

```
graph bar (mean) survived, over(sex)
over(child) over(pclass), blabel(bar)
```

- This helps you see if class influenced survival within each group.

2. Consider Non-Linearity in Age Effects

Instead of a simple **child vs. adult** binary, age could have a more **gradual effect**.

Refinement: Use Age as a Continuous Variable

```
logit survived i.sex c.age i.pclass
```

- Here, `c.age` treats age as a continuous variable.
- Check whether the effect of age on survival is linear

Refinement: Check for a Nonlinear Effect (Quadratic Term)

```
gen age2 = age^2
logit survived i.sex c.age c.age2
i.pclass
```

- If `age2` is significant, age has a **curved** effect on survival.

3. Interaction Between Age and Class

- Were **children in lower classes** at a disadvantage?
- Test if the effect of being a child varies by class.

Refinement: Interaction Between Child and Class

```
logit survived
i.sex##i.child##i.pclass
```

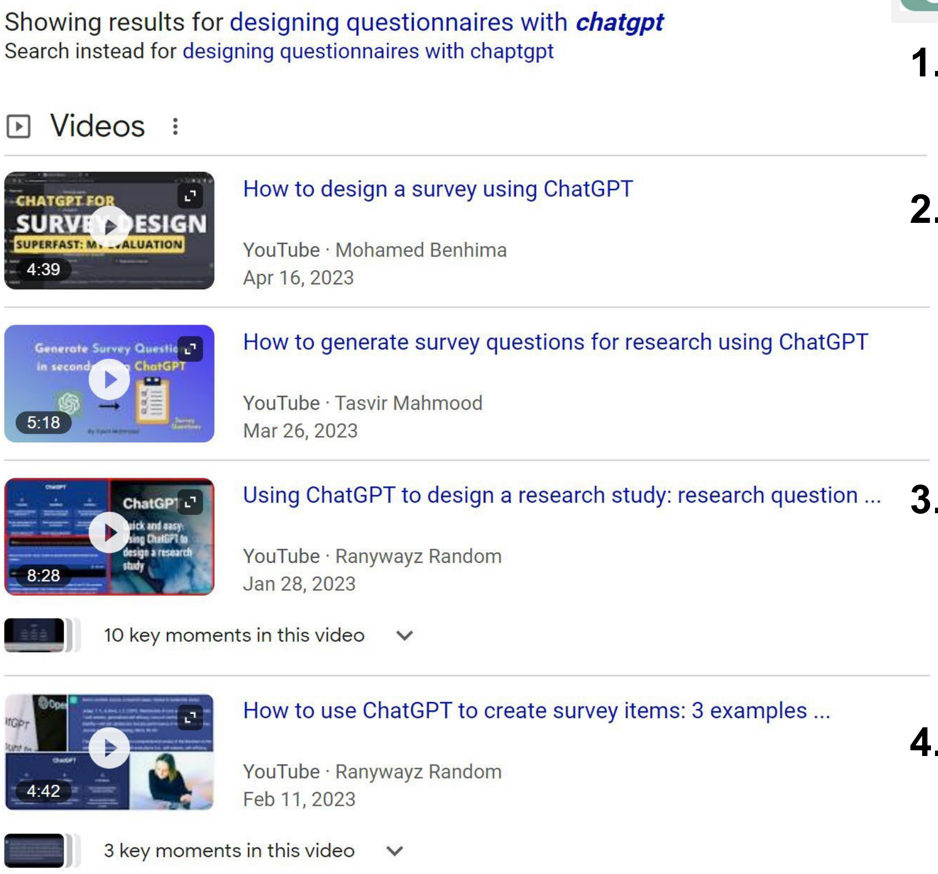
- This tests if **children in 3rd class had lower survival than children in 1st class.**

Refinement: Marginal Effects by Class

```
margins child#pclass
marginsplot, xlabel(1 "1st Class" 2
"2nd Class" 3 "3rd Class")
```

- This shows survival rates for **children vs. adults across classes.**

Tukey defined data analysis as
"procedures for analyzing data,
techniques for interpreting the results of such procedures,
ways of planning the gathering of data to make its analysis
easier, more precise or more accurate, and all the
machinery and results of (mathematical) statistics which
apply to analyzing data" (**Tukey 1962**, p. 2)



1. **On a scale of 1-10, how satisfied were you with this class?**

2. **How relevant and useful did you find the content of the class for your academic and/or professional goals?**
   (1 = Not relevant at all, 10 = Extremely relevant)

3. **How effective was the instructor in delivering the course material?**
   (1 = Not effective at all, 10 = Extremely effective)

4. **How encouraged did you feel to participate and engage in class discussions?**
   (1 = Not encouraged at all, 10 = Highly encouraged)

Improve this question "On a scale of 1-10, how satisfied were you with this class?" Follow the textbook from Schuman and Presser. Structure output "Revised Question:" "Answer Options:" and Explanation

Revised Question: How would you describe your overall satisfaction with this class?
Answer Options: Very dissatisfied, Dissatisfied, Neutral, Satisfied, Very satisfied.
Explanation: A 5-point scale simplifies the choices and captures nuanced feelings. Avoiding numbers ensures that participants don't just choose a 'middle' number without giving it thought.

And this one: "How relevant and useful did you find the content of the class for your academic and/or professional goals? (1 = Not relevant at all, 10 = Extremely relevant)"

How would you rate the relevance of the class content to your academic or professional goals? Answer Options: Not at all relevant, Slightly relevant, Moderately relevant, Very relevant, Extremely relevant. *Explanation: Splitting relevance and usefulness makes it clearer and avoids confusion. Simplified options provide clarity.*

Political Analysis

Article contents

Abstract

Footnotes

References

Out of One, Many: Using Language Models to Simulate Human Samples

Published online by Cambridge University Press: **21 February 2023**

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate

Show author details

Article | Supplementary materials | Metrics

Get access | Share | Cite | Rights & Permissions

Abstract

We propose and explore the possibility that language models can be studied as effective proxies for specific human subpopulations in social science research. Practical and research applications of artificial intelligence tools have sometimes been limited by problematic biases (such as racism or sexism), which are often treated as uniform properties of the models. We show that the "algorithmic bias" within one such tool—the GPT-3 language model—is instead both fine-grained and demographically correlated, meaning that proper conditioning will cause it to accurately emulate response distributions from a wide variety of human subgroups. We term this property *algorithmic fidelity* and explore its extent in GPT-3. We create "silicon samples" by conditioning the model on thousands of sociodemographic backstories from real human participants in multiple large surveys conducted in the United States. We then compare the silicon and human samples to demonstrate that the

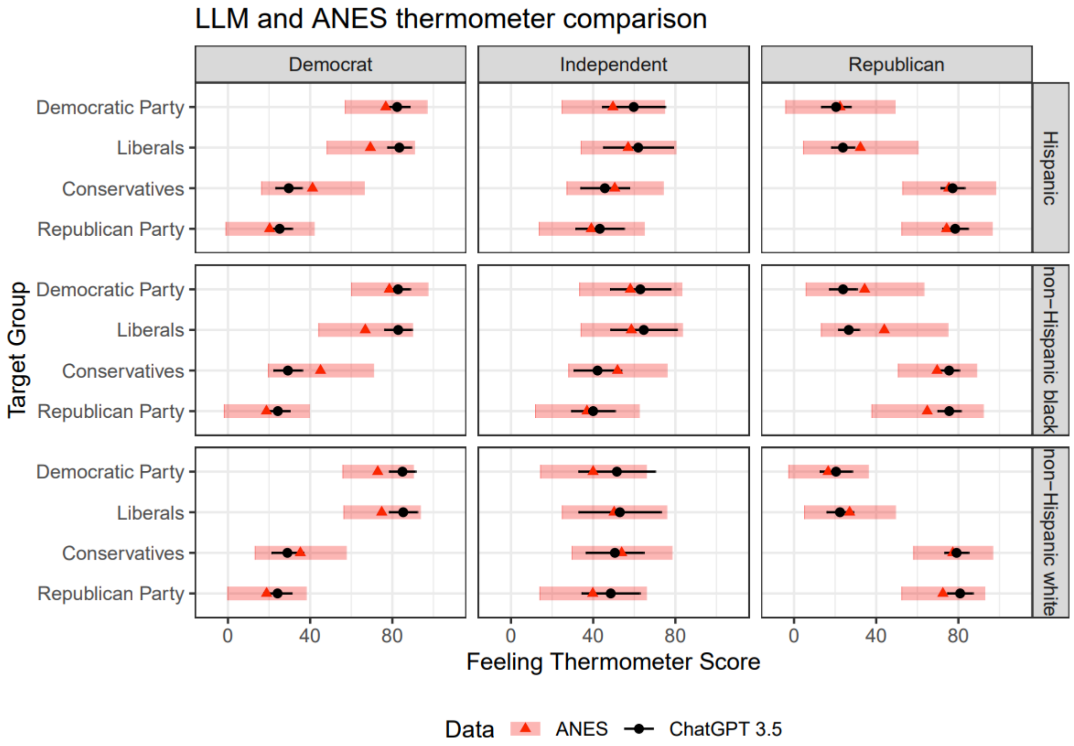
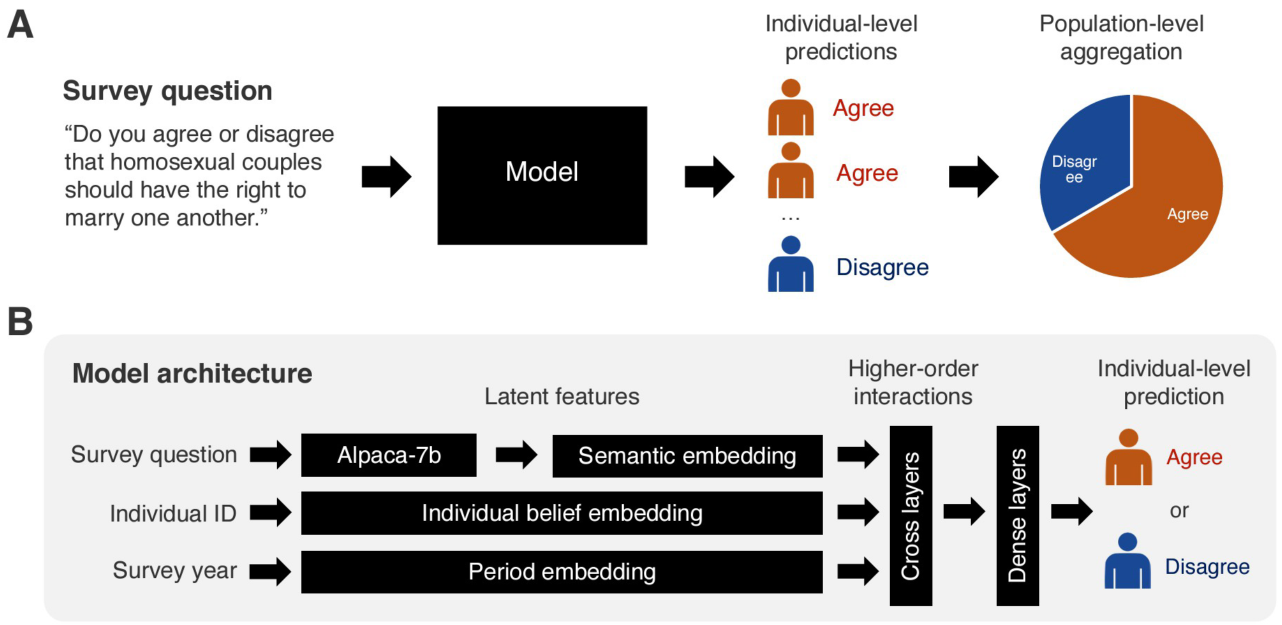


Figure 2: Average feeling thermometer results (x-axis) for different target groups (y-axes) by party ID of respondent (columns). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated by black circles and thin black bars. Sample sizes for each group-wise comparison are identical.

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023, May 4). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. https://doi.org/10.31235/osf.io/5ecfa

Kim, J., Byungkyu, L., (2023, Nov 11). *AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction* https://arxiv.org/abs/2305.09620



DATA: 68,846 individuals' responses to 3,110 questions collected for 33 repeated cross-sectional data between 1972 and 2021 for fine-tuning the LLMs. Retrieved text content of GSS survey questions from GSS data explorer

Figure 2: An overview of our methodological framework. In Panel A, we use survey weights when aggregating individual-level prediction into population-level estimates to account for potential sampling bias. In Panel B, individual belief and period embeddings are initially randomly assigned but optimized during the fine-tuning process using dense and cross layers. Semantic embedding, initially estimated by pre-trained LLMs (e.g., Alpaca-7b), is also optimized during the fine-tuning stage.

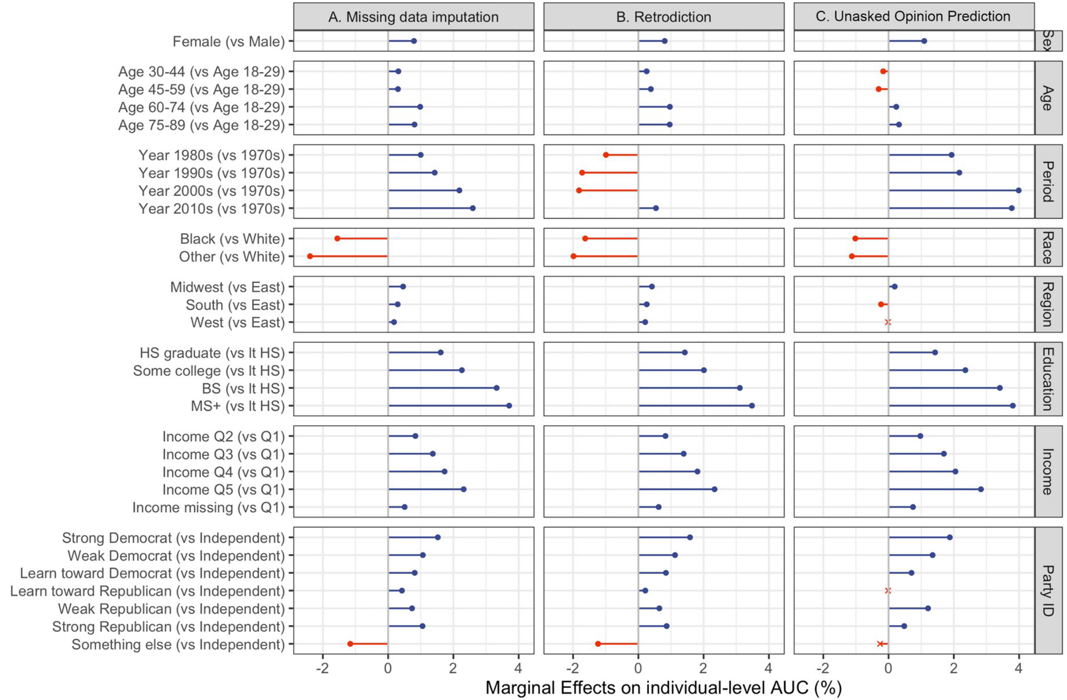


Figure 5: Coefficient plots from OLS regression models predicting individual-level AUC across three different types of missing response prediction. A higher AUC value indicates

For instance, rather than asking the same ten questions to a thousand participants, pollsters can disseminate twenty questions among the same thousand participants, each answering ten questions, and employ the model to infer individual responses to the remaining ten unasked questions. On the other hand, given our model's remarkable ability to mimic human responses, even including biases, researchers can use it to refine their survey questions by systematically examining characteristics of questions that cannot be accurately predicted (e.g., poor question wording).

Kim, J., Byungkyu, L., (2023, Nov 11). *AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction* https://arxiv.org/abs/2305.09620

Tukey defined data analysis as "procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data" (**Tukey 1962**, p. 2)

occupationMeasurement: A Comprehensive Toolbox for Interactive Occupation Coding in Surveys

Jan Simson [1¶], **Olga Kononykhina** [1], **and Malte Schierholz** [1]

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

DOI: 10.21105/joss.05505

Software

- Review
- Repository
- Archive

Editor: Chris Vernon

Reviewers:

- @welch16
- @danielruss

Submitted: 30 March 2023
Published: 24 August 2023

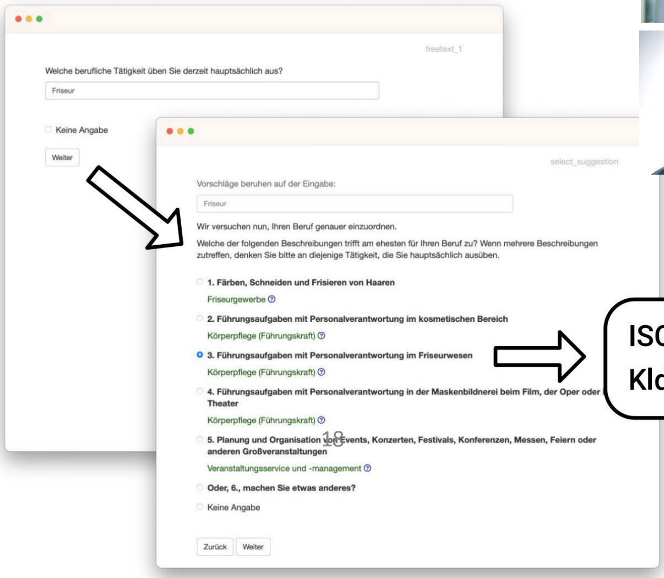
Summary

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.

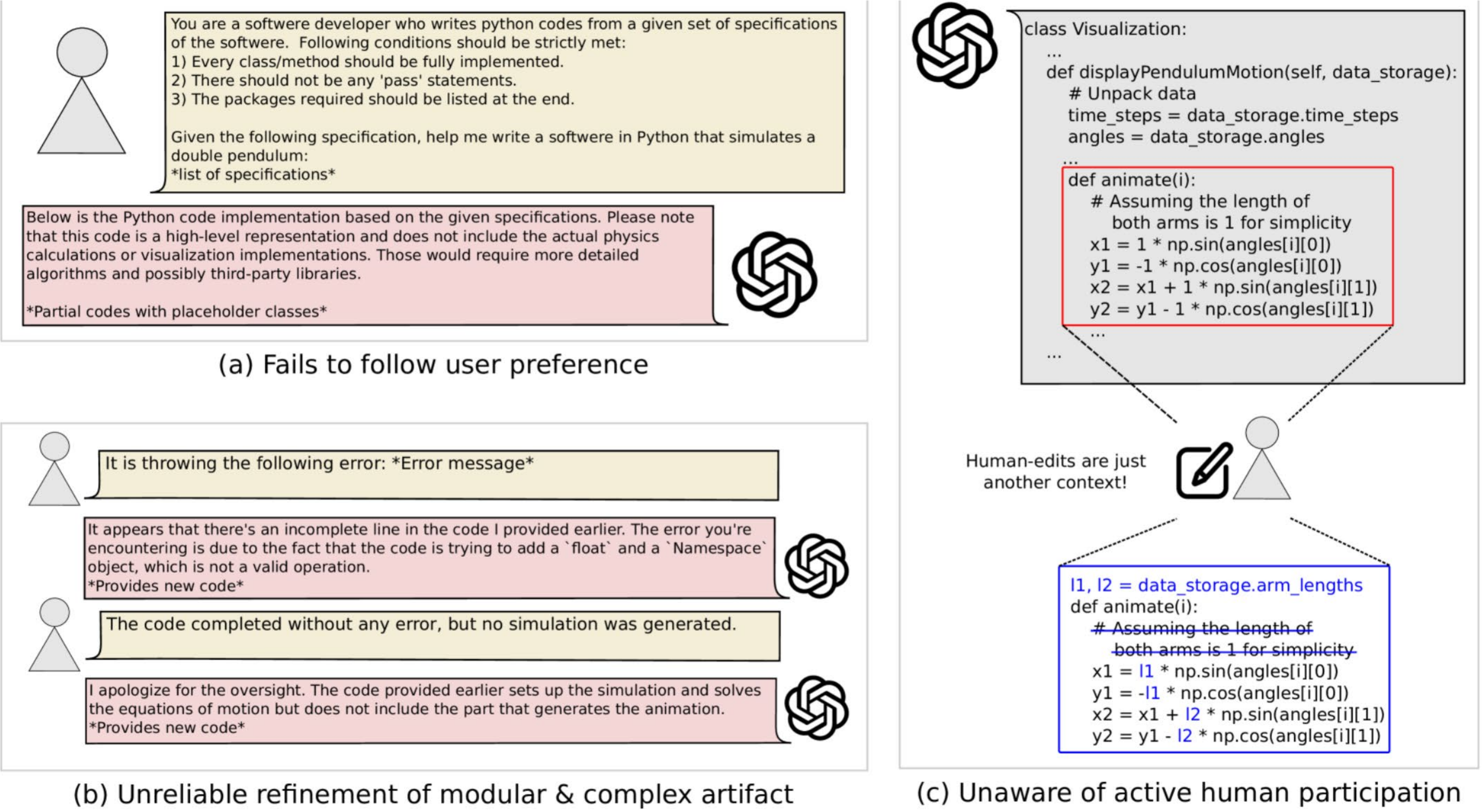
Statement of Need

Assigning occupations to standardized codes is a critical task frequently encountered in research, public administration and beyond: They are used in government censuses (e.g. USA, UK, Germany) and administrative data to better understand economic activity, in epidemiology to estimate exposure to health hazards, and in sociology to obtain a person's socio-economic



ISCO-08: 5141
KldB (2010): 823

The Challenge



(a) Fails to follow user preference

(b) Unreliable refinement of modular & complex artifact

(c) Unaware of active human participation

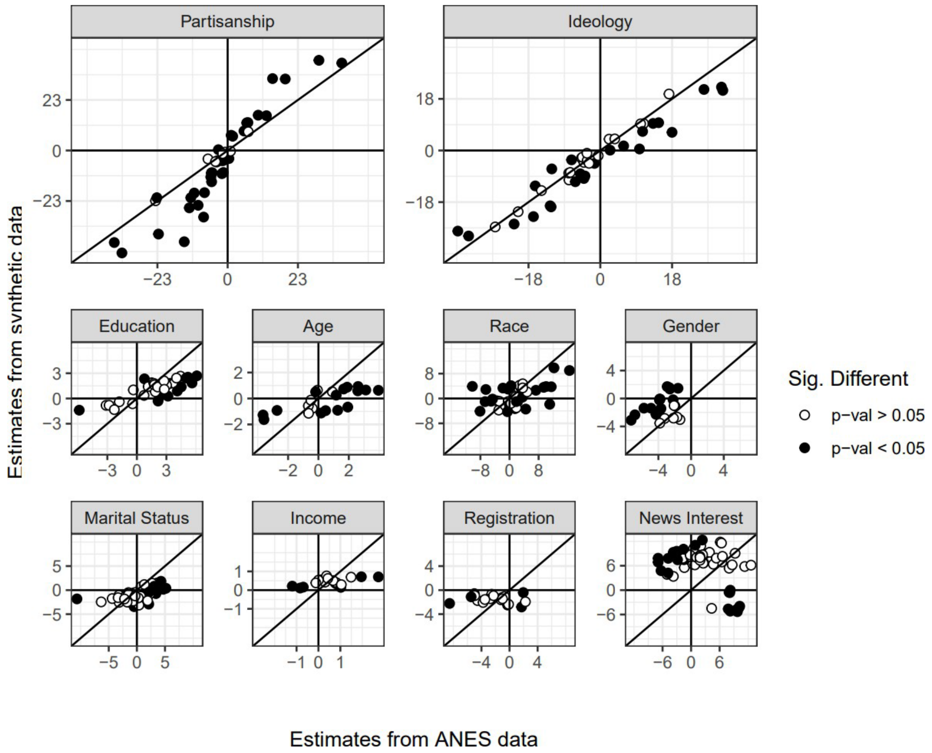


Figure 3: Each point describes the coefficient estimate capturing the partial correlation between a covariate and a feeling thermometer score toward one of the target groups, estimated in either 2016 or 2020. The x-axis position is the coefficient estimated in the ANES data, and the y-axis position is the same coefficient estimated in the synthetic data. Solid points indicate coefficients who are significantly different when estimated in either the ANES or synthetic data, while hollow points are coefficients that are not significantly different. Points in the northeast and southwest quadrants generate the same substantive interpretations, while those in the northwest and southeast quadrants produce differing interpretations. A synthetic dataset that is able to perfectly recover relationships estimated in the ANES data would have all points falling along the 45 degree line.

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023, May 4). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. https://doi.org/10.31235/osf.io/5ecfa

English (translation) I am 28 years old and female. I have a college degree, a medium monthly net household income, and am working. I am not religious. Ideologically, I am leaning center-left. I rather weakly identify with the Green party. I live in West Germany. I think the government should facilitate immigration and take measures to reduce income disparities. Did I vote in the 2017 German parliamentary elections and if so, which party did I vote for? I [INSERT]

Notes: We decided not to include “gewählt” (voted) as a suffix in the prompt, using the [MASK] instead of [INSERT] request, as it might bias the output against non-voters by reducing the likelihood of GPT completing the sentence with “nicht” (not) or “ungültig” (invalid) due to German semantics. We leave the further exploration of these effects to prompt engineering researchers.

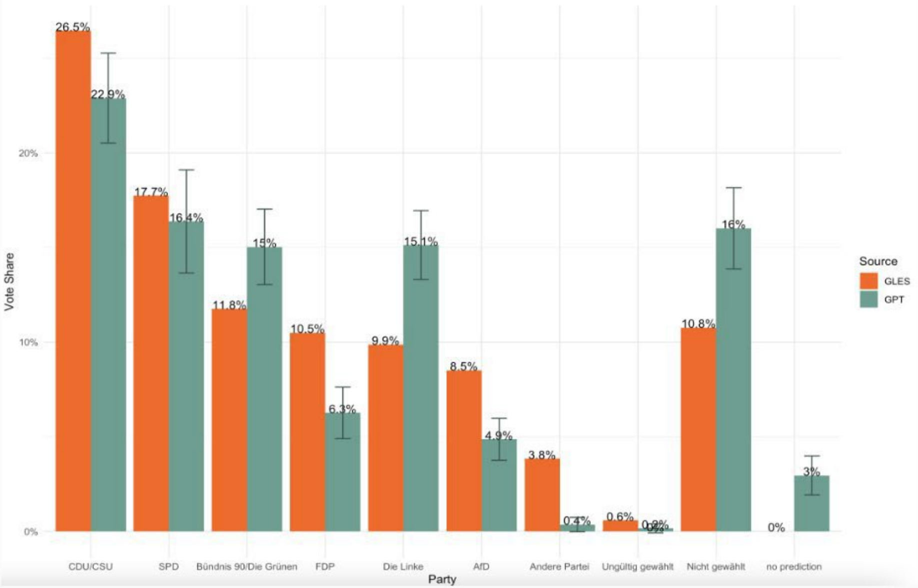


Figure 3: Replicating Argyle et al. for German data (GLES): Current project by Leah von der Heyde, Alexander Wenz and Carolina Haensch

22

Von der Heyde, L., Wenz, A., & Haensch, A.-C. (2024, February 22). Artificial Intelligence, Unbiased Opinions? Assessing GPT’s suitability for estimating public opinion in multi-party systems. https://doi.org/10.17605/OSF.IO/5BRXD

The Possible Solution

ExAIC Framework

The ExAIC framework enables coconstruction of solutions

by an expert and a dynamically adaptive agent

through search in a construction space

via natural language communication

with agent integrity by design.

ExAIC Framework

The ExAIC framework enables coconstruction of solutions

by an expert and a dynamically adaptive agent

through search in a construction space

via natural language communication

with agent integrity by design.

ExAIC Framework

The ExAIC framework enables coconstruction of solutions

by an expert and a dynamically adaptive agent

through search in a construction space

via natural language communication

with agent integrity by design.

ExAIC Framework

The ExAIC framework enables coconstruction of solutions
by an expert and a dynamically adaptive agent
through search in a construction space
via natural language communication
with agent integrity by design.

ExAIC Framework

The ExAIC framework enables coconstruction of solutions

by an expert and a dynamically adaptive agent

through search in a construction space

via natural language communication

with agent integrity by design.

ExAIC Framework

The ExAIC framework enables coconstruction of solutions
by an expert and a dynamically adaptive agent
through search in a construction space
via natural language communication
with agent integrity by design.

DS: Real world example

Inspired by the ongoing collaboration of several ExAICPLs w/ Deutsche Bundesbank

LEGAL ENTITY VALIDATION

DS: Preliminaries

Expert
Data scientist
(= domain expert)

Task
Data analysis and validation

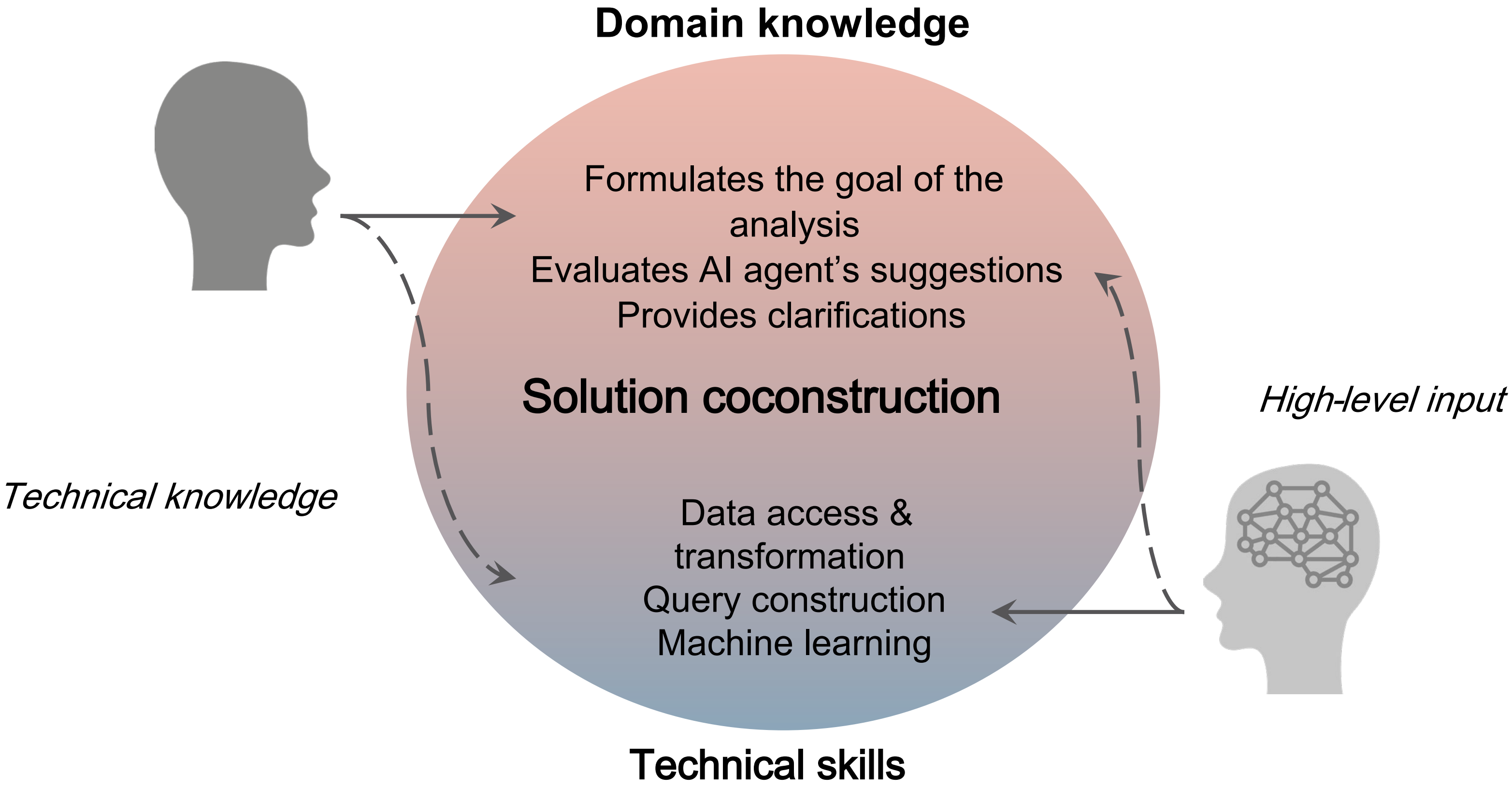
Solution

A data science workflow
(aka *a reasoning chain*) over
several multimodal data sources

Team

Expert and AI *interactively* query,
transform and interpret the data

DS: Expert-AI team



DS: Legal entity validation scenario

Data: private repositories of Deutsche Bundesbank

- **Huge** amounts of data
- Potentially **inconsistent**

Account Holder Name	Company Registration Nr	MS Registry	Installation ID	Installation Name	Activity Type	Permit ID	Permit Expiry	Contact Country	Contact City	Contact PCode	Contact Address L1
Verallia Deutschland AG	HRB 610192	DE	217	Glass production facility Essen	Manufacture of glass	14250-0032	ACTIVE	DE	Bad Wurzach	88410	Oberlandstraße 1-8
Verallia Deutschland AG	HRB 610192	DE	218	Glass production facility Wirges	Manufacture of glass	14250-0033	ACTIVE	DE	Bad Wurzach	88410	Oberlandstraße 1-8
Verallia Deutschland AG	HRB 610192	DE	219	Glass production facility Neuburg	Manufacture of glass	14250-0034	ACTIVE	DE	Bad Wurzach	88410	Oberlandstraße 1-8
Verallia Deutschland AG	HRB 610192	DE	220	Glass production facility Wurzach	Manufacture of glass	14250-0035	ACTIVE	DE	Bad Wurzach	88410	Oberlandstraße 1-8

…

Example data: EU Emissions Trading Scheme (publicly available)

DS: Legal entity validation scenario

Goal: verify that each company exists as a legal entity and its legal data is consistent with the available data sources

Expert can not check it all manually

AI can not solve everything

Expert & AI solution coconstruction !

DS: Legal entity validation scenario

High-level task specification (language)

Task: Find inconsistencies
Area: East Germany
Company types: **Type 1, Type 2, …, Type 10**

Fine-grained types

Find inconsistencies in company type data within east Germany

Please specify which company types you are interested in

Fine-grained; Type 1, .., Type 10

DS: Legal entity validation scenario

Mid-level workflow specification (structured)

W1:

```
For all entries:
text = LOOKUP(company type)
err = SPELLCHECK(text)
REPORT(err)
```

W2:

```
For all entries:
t1 = LOOKUP(company type)
addr = LOOKUP(company address)
img = LOOKUP_SATELLITE(addr)
t2 = ANALYZE_TYPE(img)
err = CHECK(t1, t2)
REPORT(err)
```

Check images

Find inconsistencies in company type data within east Germany

Please specify which company types you are interested in

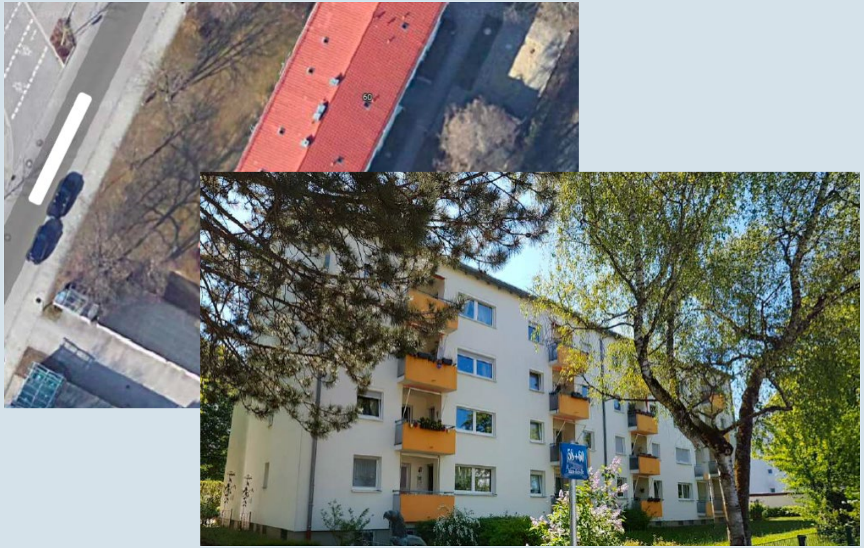
Fine-grained; Type 1, .., Type 10

Alright, which workflow do you prefer, W1 or W2?

W2 is better;

DS: Legal entity validation scenario

Joinery?
(Manufacturing)



Find inconsistencies in company type data within east Germany

Please specify which company types you are interested in

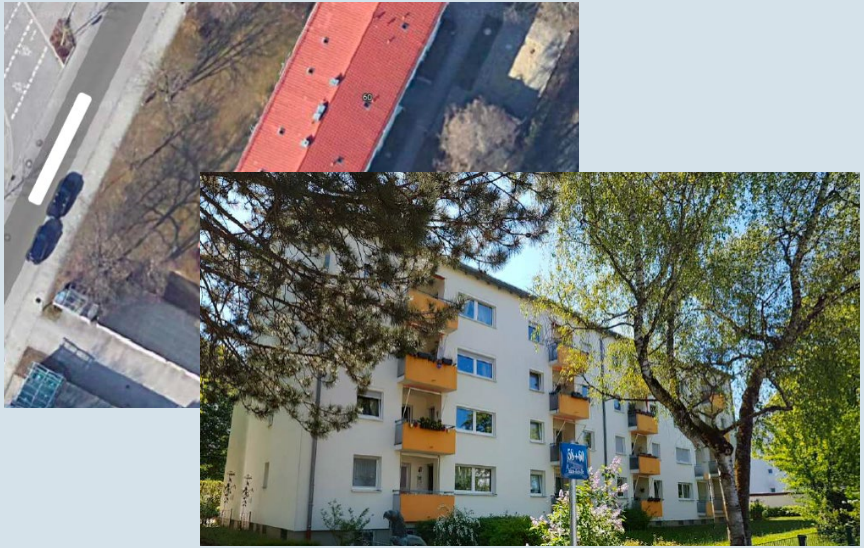
Fine-grained; Type 1, .., Type 10

Alright, which workflow do you prefer, W1 or W2?

W2 is better;

DS: Legal entity validation scenario

Joinery?
(Manufacturing)



Find inconsistencies
in company type data within
east Germany

Please specify which company
types you are interested in

Fine-grained;
Type 1, .., Type 10

Alright, which workflow do you
prefer, W1 or W2?

W2 is better;
however, you should also look up
street view images.

DS: Legal entity validation scenario

Mid-level workflow specification (structured)

```
For all entries:
t1 = LOOKUP(company type)
addr = LOOKUP(company address)
img1 = LOOKUP_SATELLITE(addr)
t2 = ANALYZE_TYPE(img1)
img2 = LOOKUP_STREET(addr)
t3 = ANALYZE_TYPE(img2)
err = CHECK(t1, t2, t3)
REPORT(err)
```

Include street view

Find inconsistencies in company type data within east Germany

Please specify which company types you are interested in

Fine-grained; Type 1, .., Type 10

Alright, which workflow do you prefer, W1 or W2?

W2 is better; however, you should also look up street view images.

Ok; executing.

DS: Legal entity validation scenario

Low-level workflow (executable)

```
Instantiating a solution...
Executing...
157 Errors found
```

Show me a likely error

DS: Legal entity validation scenario

Production of car
input factors?

Show me a likely error

Here you go; the image shows
a residential building.

With scene view
images, you should also look
left and right.

DS: Legal entity validation scenario

Production of car
input factors?

Show me a likely error

Here you go; the image shows a residential building.

With scene view images, you should also look left and right.

Got it; this was a false positive.

DS: Legal entity validation scenario

Mid-level workflow specification (structured)

```
For all entries:
t1 = LOOKUP(company type)
addr = LOOKUP(company address)
img1 = LOOKUP_SATELLITE(addr)
t2 = ANALYZE_TYPE(img1)
[images] = LOOKUP_STREET(addr)
types = ANALYZE_TYPE(images)
err = CHECK(t1, t2, types)
REPORT(err)
```

Multiple views

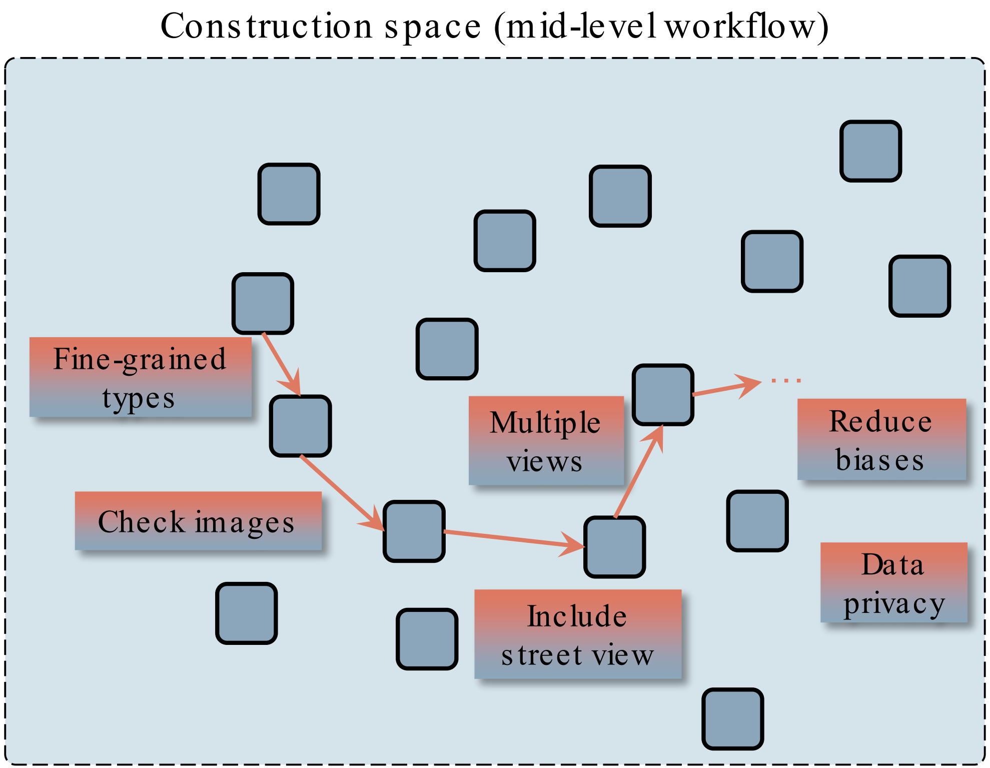
Show me a likely error

Here you go; the image shows a residential building.

With scene view images, you should also look left and right.

Got it; this was a false positive.

DS: Legal entity validation scenario



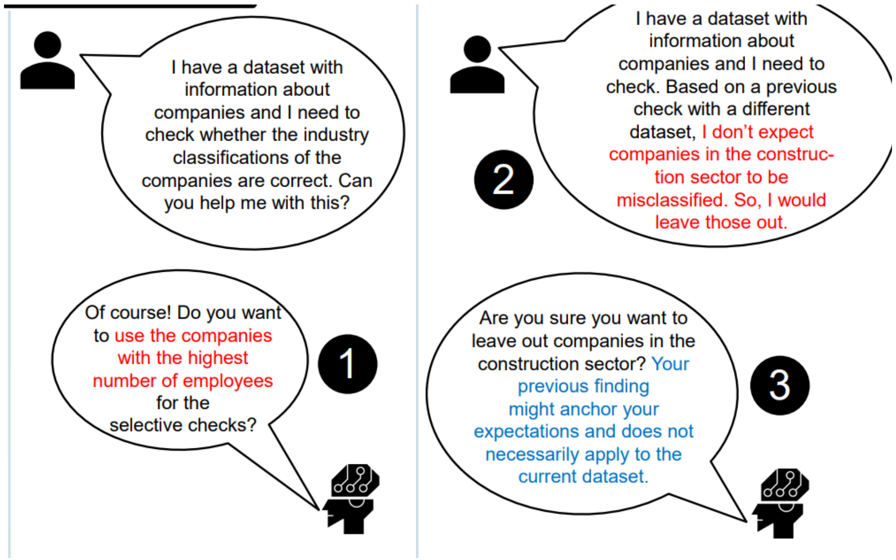
Over time:

- The agent learns the expert's preferences
- The quality/quantity of communication evolves

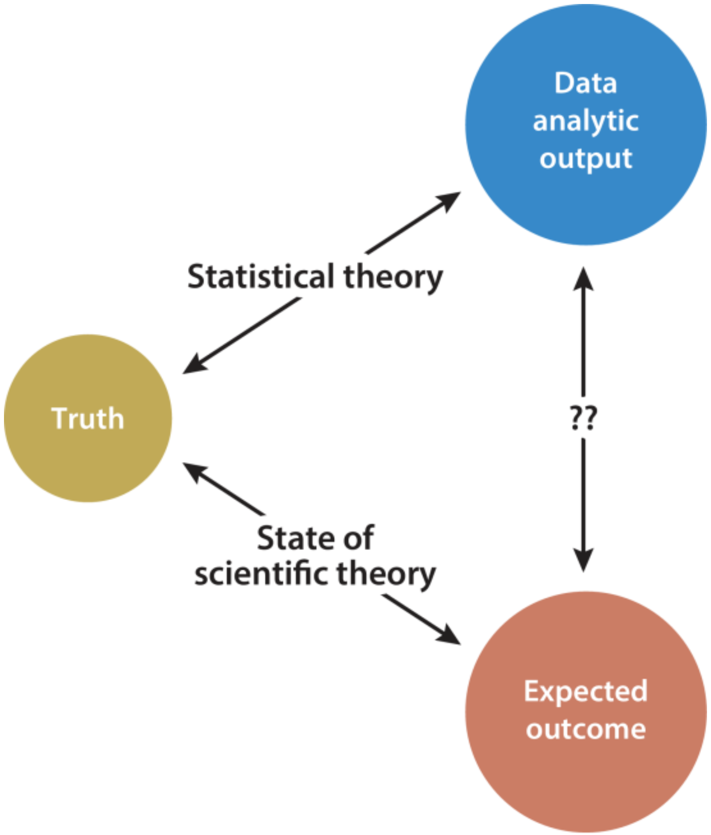
AI agents **need to know how to ask** clarification questions so that the expert provides the right answers

This requires **expert-awareness** and generation of the **right clarification questions**

We develop a method that **incrementally learns from sparse** and **multimodal expert inputs**



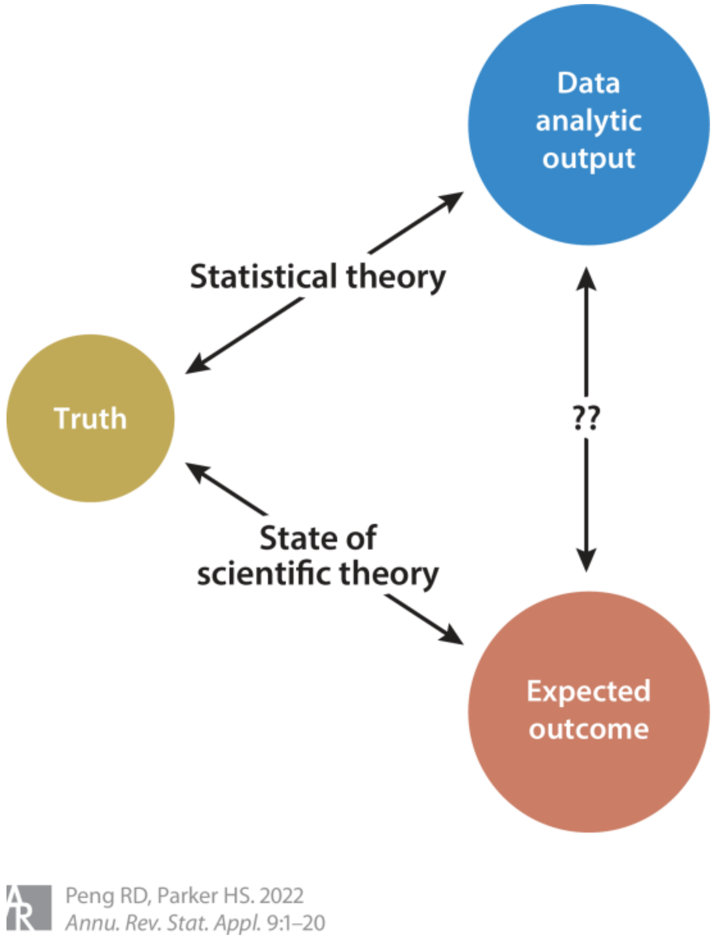
The (missing) Data Science Theory

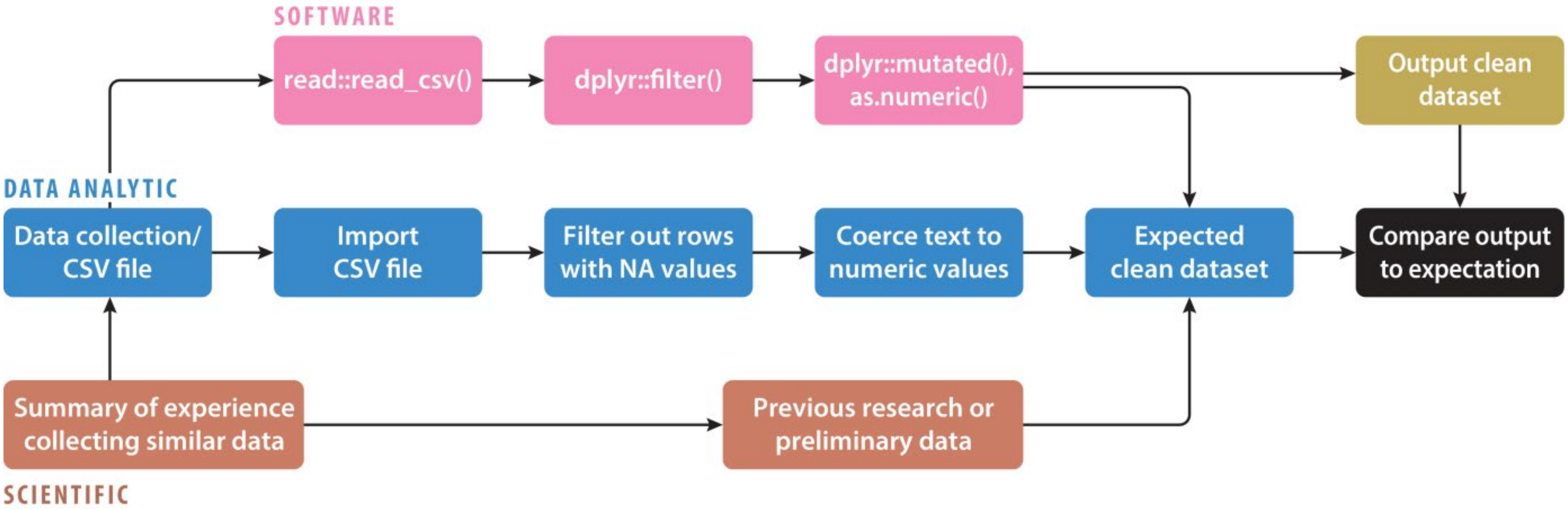


Peng RD, Parker HS. 2022
Annu. Rev. Stat. Appl. 9:1–20

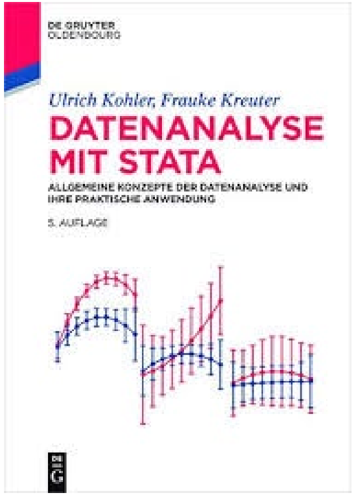
Research Questions

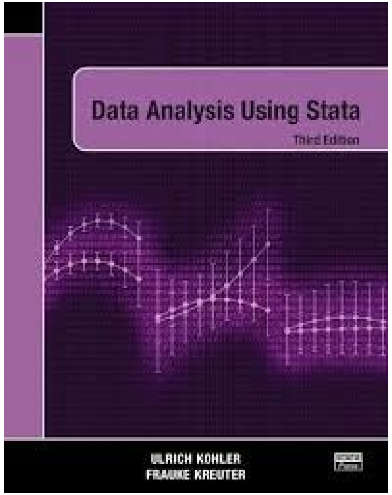
1. What data science steps should the agent **be aware of**?
2. How can we **enable an agent to ask clarification questions** that address the cognitive biases a human expert may have?
3. How can we **evalute the suitability of the agent-expert dialog efficiently** with sparse data?





Peng RD, Parker HS. 2022
Annu. Rev. Stat. Appl. 9:1–20





Data Analysis
With AI
Using Stata