Advancing Perspectivist Ground Truthing with Social Science

Table of contents

# ML/AI relies on reference data

Aroyo & Welty (2015)

* In order to evaluate ML / AI systems, we compare the output of these systems to reference data.
* One method for creating reference data is the collection of human annotations.
* This method typically assumes that, for every piece of content being annotated, there is a single canonical truth
* quality of annotations is assessed using inter-annotator agreement, where more agreement = better annotations

## ML uses human annotations very often

Geiger et al. (2021)

* 200 randomly sampled ML papers from 3 domains:
  + Social Sciences & Humanities
  + Life & Biomedical Sciences
  + Physical & Environmental Sciences
* Out of 141 classification tasks, 103 (73.05%) used human labels
* Out of 103 human labels, 58 (56.31%) used only external labels

i.e. ML re-uses external labels

## Issues using human annotations in ML

Geiger et al. (2021) ML science studies inadequately report ‘ground truth’

Hullman et al. (2022) thus we cannot know what data generating process the resulting model represents

[perhaps cat image parable here?]

### Human annotations aren’t always accurate

Griffin & Brenner (2004) review errors and biases in human judgements[[1]](#footnote-21)

* over/under prediction: confidence score is higher/lower than accuracy
* over/under extremity: confidence is more extreme at ends

also reviews possible reasons:

* optimistic overconfidence
* confirmation bias
* case-based judgment
* ecological probability
* error model (psychometric model)

### sampling and measurement biases in ML research

Hullman et al. (2022) compare claims that ML is facing a reproducibility crisis to the crisis in psychology. Among the issues they note relate to benchmark datasets, which researchers often re-use as they publish on standardized benchmarks, and because they are cost prohibitive to collect. It may also be the case that ML researchers broadly prefer to work on building and evaluating performance, rather than executing ground-truthing projects Sambasivan et al. (2021).

With regards to reference data:

* representation bias / non-representative samples
* measurement bias / unvalidated measurement instruments
* underspecification of portions of input space in training data
* transformation of data to optimize for ‘accuracy’
* lack of or poor dataset documentation

In other words, optimizing for predictive accuracy using very large datasets does not ‘absolve’ researchers from having to consider the data generating process. They note benefits that both machine learning and psychology could gain by borrowing methods from each other, but note the danger if these are misused. For the benefit of machine learning, there are lessons to be learned from social science, and the replication crisis. Among them are 1) collecting samples whose test/evaluation set distributions are drawn from the same deployment distribution, and 2) using valid measurement instruments.

### ML ignores perspectives of annotators

Beyond errors in judgment are questions about the target for the annotations. For at least some phenomena, the assumption that there is a single ground-truth to approximate with annotations doesn’t hold.

Aroyo & Welty (2015)

7 ‘myths’ of human annotation:

* there is one truth
* disagreement is bad
* detailed guidelines help
* experts are better
* one annotator is enough
* all items are created equal
* once done, forever valid

For myths 1 and 2:

* list examples from NLP where the disagreement from annotators is sensible
* they argue that the assumptions of a single ground truth, and that disagreement is indicative of poor annotations are both false.

While it can be indicative of annotation quality, e.g. the annotator is annotating incorrectly, in other cases disagreement indicates that the media being rated is ambiguous.

Beck et al. (2022): we should expect more variance to the degree that tasks measure opinion

* show work on an intuitively perspective-based use-case: hate speech

Aroyo & Welty (2015) operationalize ‘crowd truth’ with an illustration where the ‘gold standard’ is the probability that a sentence contains an element, based on the probability that an annotator annotated that sentence with that element.

* i.e. the label isn’t represented as ‘present’ or ‘not present’, but as a probablility
* thus the ‘crowd truth’ attempts to capture the ‘range of reasonable interpretations’

# Tools from Social Science can help

Recent trends—especially in deep learning—prioritize empirical performance over theoretical assumptions about the data generating process. A systematic analysis of highly cited ML works shows that Performance, Generalization, Quantitative evidence, Efficiency, Building on past work, and Novelty are the primary values in ML work Birhane et al. (2022).

Unlike the social sciences (e.g. psychology), ML work ignores attempts to model the process that gives rise to the data, assuming it cannot be learned, and aims instead at predictors that will within estimable error bounds Hullman et al. (2022). This is problematic as this kind of optimization doesn’t resemble real world deployment.

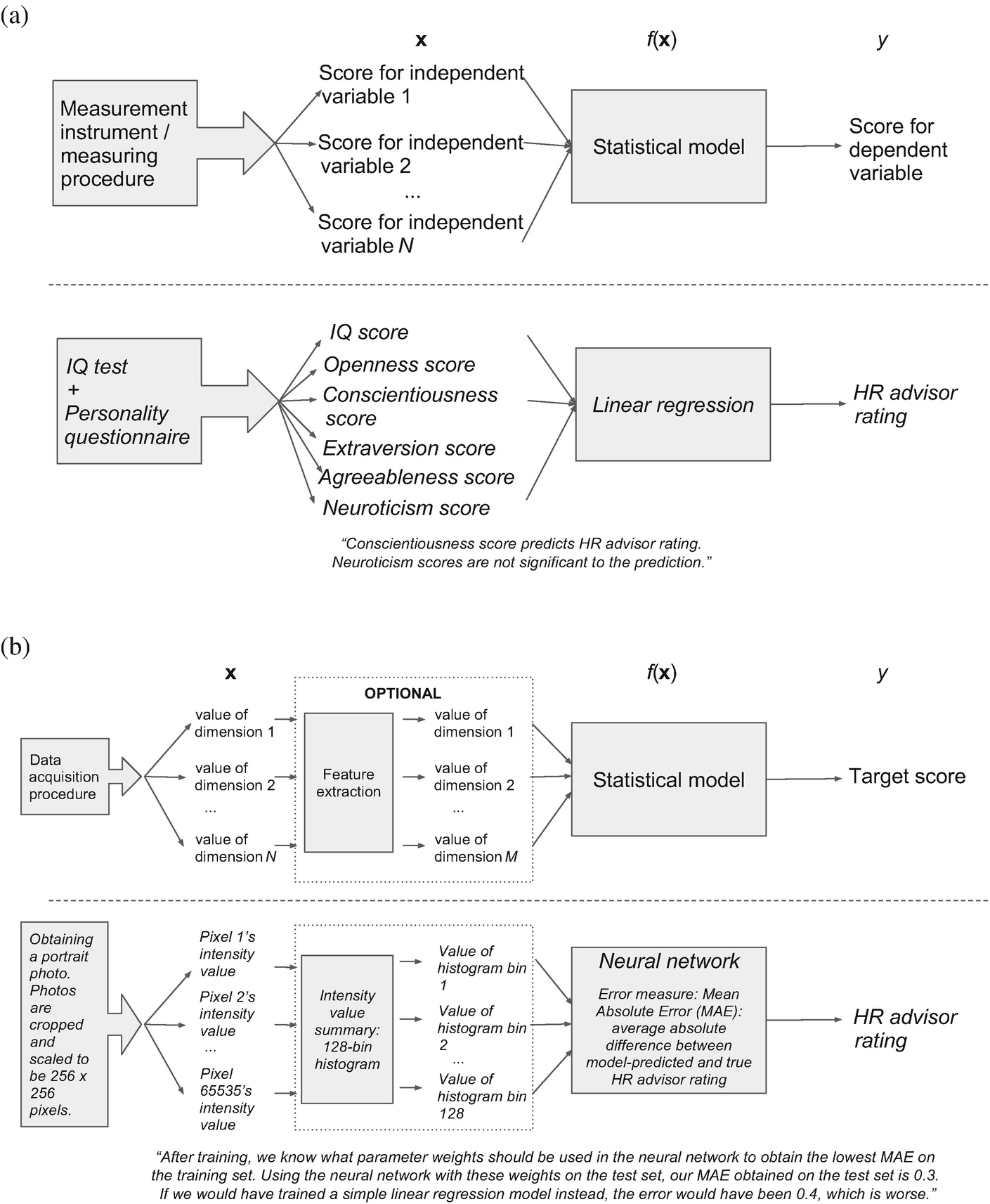
### Social and computational sciences have different focci

social sciences:

* interpretable meaning of x and y
* design is informed by theory

computational sciences:

* learning procedure f(x)

 Fig. 1: Liem et al. (2018)

### Issues with sampling

Solutions to sampling problems can come from sampling theory: Groves et al. (2009)

considerations:

* sampling frame: the elements in from populations that you have access to
* ineligible units: elements in the sampling frame that are not your target
* undercoverage: elements from target population that are not in the frame

solutions:

* stratified sampling

### Issues with instruments

Beck et al. (2022)

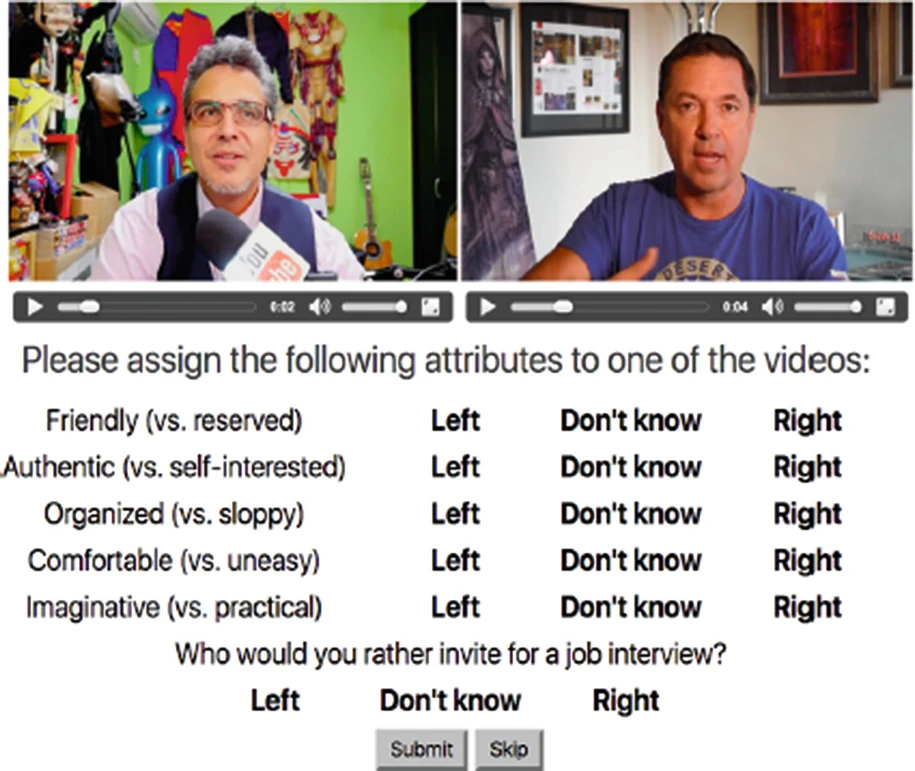
* annotation collection requires design thinking
  + Task Structure: specific wording and response options, including debates over the inclusion of “I don’t Know” option
  + Order Effects: specific judgements are affected by previous perceptions
  + Annotator Effects: backgrounds, opinions, experiences of respondents affect responses

[another ref that describes the target as a latent variable]

### A tale of two studies: the case of personnel selection

Ponce-López et al. (2016)

* efficient way to gather media and annotation data
* BUT no validation of instrument, or ecologically valid media data
* distribution of training /eval data don’t come from the target distribution

 Figure 2: Ponce-López et al. (2016)

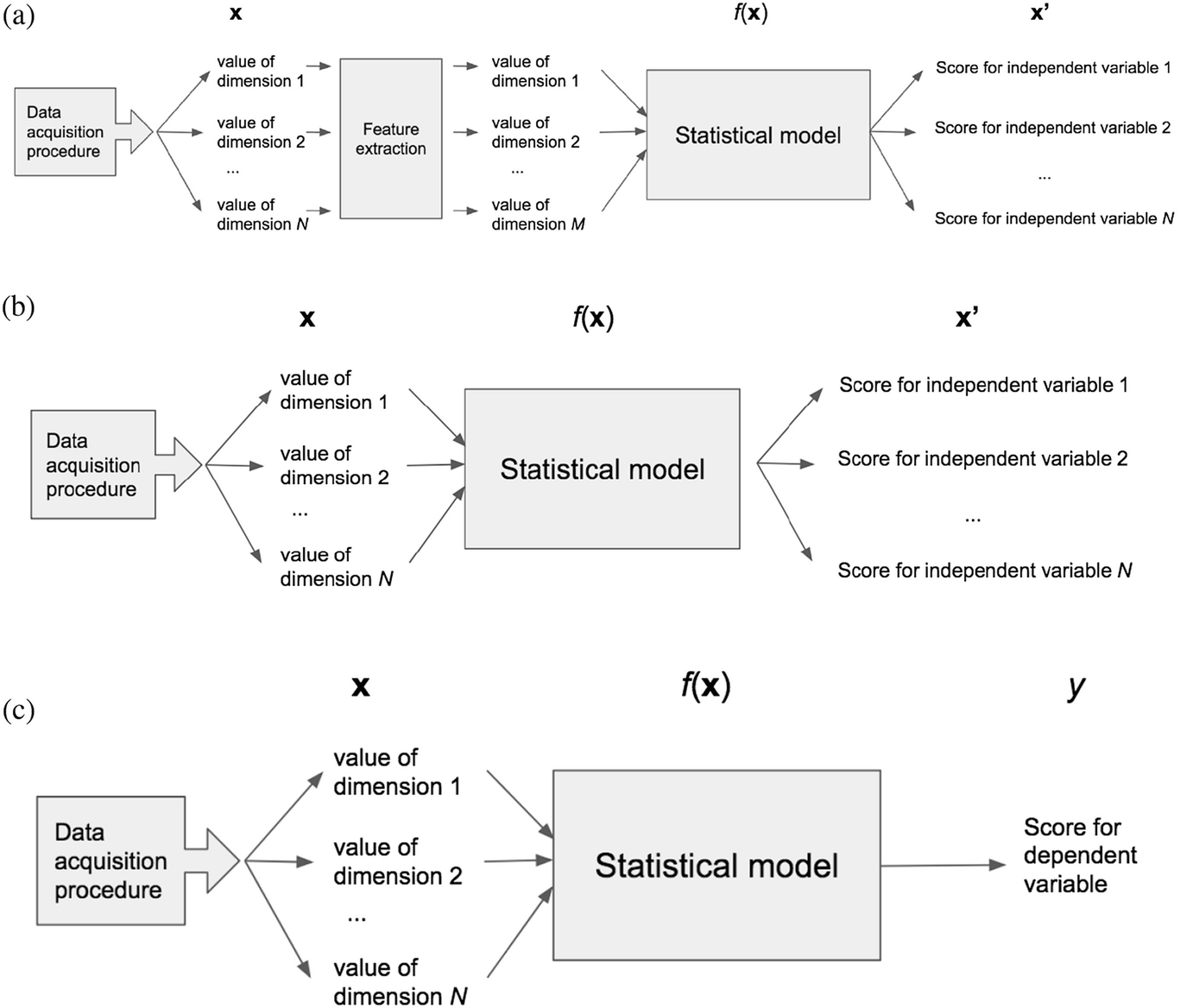
Compared to Koutsoumpis et al. (2024):

* ecological validity: media data was mock asynchronous video interviews
* ecological validity: interview questions designed to activate personality facets
* personality instruments: validated HEXACO scale
* perspectives: self & observer ratings

# Present Work

We incorporate these considerations in the design of our study, and attempt to further the field in the following ways:

* we attempt representative sampling of both media and respondents
* we aim to estimate 10-dimensional psychological construct
* we select media that is ambiguous (i.e. that will result in subjectivity in the ratings) as well as media that we expect not to be ambiguous for comparison
* we estimate a-priori the number of ratings necessary rather than assuming
* we take into account perspectives

case study of this thesis works towards path (b) in Liem et al. (2018) shown in: 

## References

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, *36*(1), 15–24.

Beck, J., Eckman, S., Chew, R., & Kreuter, F. (2022). Improving labeling through social science insights: Results and research agenda. *International Conference on Human-Computer Interaction*, 245–261.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The values encoded in machine learning research. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184.

Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (2021). " garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *arXiv Preprint arXiv:2107.02278*.

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. *Blackwell Handbook of Judgment and Decision Making*, *199*, 158–177.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (Vol. 561). John Wiley & Sons.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 335–348.

Koutsoumpis, A., Ghassemi, S., Oostrom, J. K., Holtrop, D., Breda, W. van, Zhang, T., & Vries, R. E. de. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning. *Computers in Human Behavior*, *154*, 108128.

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Sukma Wicaksana, A., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. *Explainable and Interpretable Models in Computer Vision and Machine Learning*, 197–253.

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Baró, X., Escalante, H. J., & Escalera, S. (2016). Chalearn lap 2016: First round challenge on first impressions-dataset and results. *Computer Vision–ECCV 2016 Workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, 400–418.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

1. Griffin & Brenner (2004) note that much of this work was about people guessing knowledge from an almanac, and then guessing how accurate they were [↑](#footnote-ref-21)