Correcting the Measurement Errors of AI-assisted Labeling in Image Analysis using Design-based Supervised Learning

Alessandra Rister Portinari Maranca* Jihoon Chung† Musashi Hinck‡

Adam D. Wolsky§ Naoki Egami¶ Brandon M. Stewart‖

February 15, 2025

Abstract

Generative artificial intelligence (AI) has shown incredible leaps in performance across data of a variety of modalities including texts, images, audio, and videos. This affords social scientists the ability to annotate variables of interest from unstructured media. While rapidly improving, these methods are far from perfect and, as we show, even ignoring the small amounts of error in high accuracy systems can lead to substantial bias and invalid confidence intervals in downstream analysis. We review how using design-based supervised learning (DSL) guarantees asymptotic unbiasedness and proper confidence interval coverage by making use of a small number of expert annotations. While originally developed for use with large language models in text, we present a series of applications in the context of image analysis, including an investigation of visual predictors of the perceived level of violence in protest images, an analysis of the images shared in the Black Lives Matter movement on Twitter, and a study of US outlets reporting of immigrant caravans. These applications are representative of the type of analysis performed in the visual social science landscape today, and our analyses will exemplify how DSL helps us attain statistical guarantees while using automated methods to reduce human labor.

*Graduate Student in Sociology, Princeton University

†Graduate Student in Computer Science, Princeton University.

‡AI Research Scientist, Intel Labs.

§Senior Research Specialist, Princeton University.

¶Corresponding Author. Assistant Professor, Department of Political Science, Columbia University. Email: naoki.egami@columbia.edu. URL: `https://naokiegami.com`.

‖Corresponding Author. Associate Professor, Department of Sociology and the Office of Population Research, Princeton University. Email: bms4@princeton.edu. URL: `https://brandonstewart.org`.

1 Introduction

Visual media is increasingly an inescapable part of social and political communication. Predominantly visual sites such as Instagram, TikTok, and YouTube, are among the most heavily trafficked social media platforms. Yet many social science studies have constrained themselves to study text—often throwing away the visual information available for mixed environments such as Twitter.[1] Until recently this was a practical concern—methods for annotating documents using 'text as data' approaches were well developed (Grimmer and Stewart, 2013; Grimmer, Roberts and Stewart, 2022)—but annotation techniques for images were in their infancy. As convolutional neural networks (CNNs) became accessible for social scientists, work began to appear introducing 'image as data' pipelines to social scientists (Joo and Steinert-Threlkeld, 2022; Webb Williams, Casas and Wilkerson, 2020). The recent advent of generative artificial intelligence (AI) has only hastened this transformation, allowing an increasing number of annotation tasks on diverse media to be done with little to no training data. These methods are rapidly changing, rapidly improving, and yet are still prone to errors, particularly for certain types of media and tasks. This raises the question that we address in this paper—*how do we retain provable statistical guarantees while using an ever-changing array of methods to annotate our data?*

Consider the example of Casas and Webb Williams (2019) who are interested in whether images are a key factor in political mobilization online. To study this question, they collected almost 9,500 images posted on Twitter from one week in 2015 around a Black Lives Matter protest. They are interested in how the emotional content of the images—for example, whether they evoke anger, enthusiasm, or fear—is related to the attention these messages get on Twitter. To investigate these questions, they manually annotated every image for seven different variables using more than 1200 MTurk workers and 2 undergraduate research assistants.

If Casas and Webb Williams were starting their work today, they would have a wide range of options. They could label a few thousand images by hand and train a convolutional neural network to label the rest (Webb Williams, Casas and Wilkerson, 2020). They could instead prompt a multimodal large language model to label the images without any training data(Ziems et al., 2024; Gilardi, Alizadeh and Kubli, 2023). Depending on the difficulty of the task and the type of data, these annotations might be accurate, and they might not be. The capabilities of the models are often rapidly shifting, and it is unclear what results will look like in the future.

Like most social scientists, Casas and Webb Williams (2019) aren't interested in the annotation on any one image, they are interested in downstream analyses using annotations as variables (e.g., do images that evoke more anger get shared more?). These questions can often be framed as a regression in which the annotation of the image takes on the role of either the outcome, an independent variable, or both. In the case of Casas and Webb Williams (2019), they are regressing an observed outcome (number of retweets) on seven independent variables derived from image annotations and observed characteristics of the poster (e.g., number of followers, number of tweets, etc.). As we will show later, off-the-shelf generative AI systems do a remarkably good job on this image annotation task, producing high-accuracy labels with only 100 examples. A social scientist might try the tool out, sample a subset of images to validate labels, and then—if the accuracy of the automated annotation looks 'good enough'—move on to running their downstream analysis ignoring any prediction error.

While the capacity of generative AI to automate these tasks with little expert supervision

[1] Although now known as X, we refer to the platform as Twitter throughout the paper.

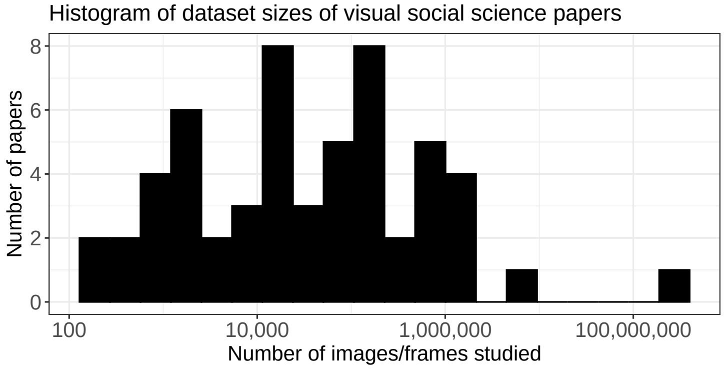
is exciting, it thrusts to the fore two central tensions that animate this piece. First, we don't *know* the annotations are accurate until we evaluate them against some kind of expert coding. This means that *we can never escape* doing some annotation ourselves as researchers. Second, when automated annotations have errors—but particularly when those errors are correlated with other variables in the regression—the estimator will exhibit asymptotic bias (Egami et al., 2024*b*). This, in turn, means that confidence intervals will not attain nominal coverage (i.e., 95% confidence intervals will not contain the truth 95% of the time across samples), and as our sample gets larger and larger, we will become increasingly certain of the *wrong* answer, even when accuracy is very high (unless accuracy is 100%). Given known gender and racial biases in vision models (Buolamwini and Gebru, 2018; Wang et al., 2022; Barocas, Hardt and Narayanan, 2023), we may reasonably assume that errors are often correlated with variables of social science interest. In short, we don't know we have high accuracy until we check, and if the accuracy isn't perfect, we will asymptotically get biased estimates. The first issue destroys the dream of doing computer vision work in the social sciences without expert annotation, and the second means that even our most accurate automated systems are probably not accurate enough to give us the statistical properties we expect.

When the downstream analysis can be framed as a regression, we demonstrate how the recently proposed Design-based Supervised Learning (DSL) framework (Egami et al., 2024*a*,*b*) enables researchers to harness generative AIs without introducing bias from prediction errors in automated AI annotations. Remarkably, by combining a small set of expert annotations with the predictions from computer vision models, we can attain provable statistical guarantees for the downstream regression model including asymptotic unbiasedness and proper coverage of confidence intervals, without needing to assume anything about the computer vision model itself. We argue that because we can't check the accuracy of the computer vision model without annotating some observations by hand anyway, this resolves the two tensions above. We use the same observations we would have used to assess accuracy to instead debias our estimator and regain desirable statistical properties. That is, debiasing should become a standard part of the broader social science framework for responsible use of AI tools along with other elements of research design and validation (Nelson, 2020; Nelson et al., 2021; Egami et al., 2022; Grimmer, Roberts and Stewart, 2022).

In this paper, we will provide a brief introduction to how social scientists are currently using computer vision. We then provide a brief introduction to DSL and show why it is needed in current social science settings. We then demonstrate how to apply DSL through an analysis of three different computer vision applications in the social sciences where predicted variables play different roles. While the paper is primarily concerned with the principled use of computer vision in social science, we emphasize in the conclusion that the techniques we describe are substantially more general and can be used in a broad array of cases.

2 Computer Vision in Social Science

The use of computer vision is only beginning to grow in social science, and many readers may not be familiar with how it can be used in practice. In this section, we briefly survey the literature in the social sciences, best practices for human annotations, and a demonstration of how computer vision can be used in practice.



Visual medium of data source	Number of papers
Social media	25
Television	14
Newspaper/magazine	11
Other	3
Satellite/street	2
Unedited footage	2

Figure 1: **Empirical literature review summary of the sample of 59 papers using visual data annotations for social sciences applications.** (Left) The number of individual data points in each paper plotted on a log-scale. The range of the dataset sizes spans several orders of magnitude. (Right) The data source for the primary application of the 59 papers included in our literature review. The plurality are social media data.

2.1 Current Use of Computer Vision: Empirical Literature Review

We collected a sample of 59 papers using computer vision techniques from across the social sciences and released in the last 20 years.[2] The papers are published in the venues of a variety of fields including political science (14 papers), sociology (4 papers), communications (7 papers), and generalist journals (15 papers). Still more (19 papers) appear in conferences, as working papers, or as chapters in edited volumes.[3] While a sizable number of papers, we note that this is substantially less than the 88 text-analysis papers published in the top 10 political science journals reviewed by Egami et al. (2024*b*) covering the period 2015–2022.

While the papers are predominantly about still images (43 papers; 73%), there is a growing use of video as well (16 papers; 27%). These visual data come from a wide range of sources, the plurality of which are social media with newspapers, magazines, and television comprising another major fraction (see Figure 1, right). Within the twenty-five papers that utilized social media data, the sites used were (in order of prevalence) Instagram, Twitter, Weibo, Facebook, Reddit, and Whatsapp. This reflects the increasingly visual nature of social media content. The content of the studies themselves cast a wide net of social science topics including politicians and political groups (17 papers), online behavior (16 papers), protests (11 papers), journalism (6 papers), human geography (5 papers), elections (2 papers), and other assorted topics (2 papers).

Many of these applications demand an automated annotation approach due to the scale of the data involved (see Figure 1, left) which ranges from a few hundred to well over a hundred

[2] We gathered papers cited by several reviews/applications (Cantú, 2019; Webb Williams, Casas and Wilkerson, 2020; Joo and Steinert-Threlkeld, 2022; Girbau et al., 2024; Zhang, Borch and Pardo-Guerra, 2023) and also those returned by a Google Scholar with the search term "images", "videos", or "computer vision" that were published after 2006. Almost certainly, though, this omits a number of papers. Notably, we did not attempt to exhaustively cover the social science work published in peer-reviewed computer science conference proceedings, although we do include working papers, book chapters, and conference papers that we could find.

[3] If papers had more than one application, we code it as the application the paper spent more time/space dissecting. For downstream analysis, we chose the most "advanced" statistical tool used in the paper (e.g., if the paper presented descriptive statistics and regression analysis, the paper was labeled as presenting a regression analysis).

million units.[4] The median size is 25,000 images which makes an automated approach appealing. As with most social science applications, these papers are predominantly interested in characterizing the 'haystack', i.e., using annotated images and videos as variables in downstream analyses (Hopkins and King, 2010; Egami et al., 2024*a*), rather than annotations of each individual image/video themselves. These annotations are summarized in a variety of downstream analyses including correlations, clustering, and regressions. In about half the cases, the annotations are either the outcome in a downstream regression (13 papers) or the independent variable in a downstream regression (15 papers). The remaining half are also 'haystack' type questions either estimating prevalence (14 papers), performing correlations (11 papers), or other similar analyses (6 papers). Unlike in computer science, where the goal is often to predict the most accurate label for a particular image, social scientists' goal is near-universally to characterize a sample or population as accurately as possible.

When these downstream analyses are done using predicted annotations rather than 'gold-standard' annotations, the statistical properties of them have no guarantees. Only 10 papers that we examined (17%) annotated all their data with human experts. The majority (83%) used a combination of computer vision model predictions trained on human coding (33 papers) or used fully automated computer vision model predictions (16 papers). Without additional correction, those 49 papers would need to assume perfect accuracy in their computer vision models to have statistical guarantees—an unlikely assumption to hold.

2.2 Producing Annotations at Scale

To prove statistical guarantees about downstream 'haystack' analyses, it is necessary to have a notion of the quantity that you want to recover (Lundberg, Johnson and Stewart, 2021). For our purposes, the ideal procedure is to have some expert annotate each data point and then run the downstream analysis (generally a regression) on those annotations. The result of this downstream analysis with complete expert annotation of all data points is the quantity that DSL approximates. In practice, social and computer scientists have relied on human annotators—some experts and some not—to annotate their data (Russakovsky et al., 2015; Won, Steinert-Threlkeld and Joo, 2017). Here we will refer to the 'expert' annotator as the one producing the quality of labels that we would ideally like to produce.

Annotating every data point with experts is generally cost-prohibitive, leading researchers to adopt either scalable human or machine annotation strategies. For scalable human annotation, researchers often use crowd-sourcing resources like Amazon Mechanical Turk, which allows for—at most—limited training and expertise. A strategy for more complex tasks is to recruit undergraduate or graduate research assistants who can be trained by the research team to provide high quality data. Student workers may be more interested and directly invested in the research itself, leading to greater attention.

For all human annotations (expert and non-expert), clear instructions are essential for defining the task. For simpler tasks, such as obtaining a single class from images, providing a website with a definition and clear examples of positive and negative cases of a class that annotators are tasked with labeling can be sufficient to gather high-quality data. For more complex tasks, such as labeling multiple classes in video data, in addition to websites that provide clear definitions

[4] In some cases, n was estimated to the closest order of magnitude based on data provided by the paper, and NA was used if the paper did not provide dataset size and it was impossible to infer. Moreover, training data was not considered as part of the dataset size.

for positive and negative cases of the different classes, supervisors should provide in-person or virtual training for annotators. Throughout the annotation process, having a system to monitor and audit annotations can also improve the accuracy of the annotations. If supervisors observe certain aspects of the data collection that are not working, they should pivot their approach, and retrain annotators or hire new annotators if needed. Oversight can prevent wholesale re-annotation of data that are unreliable and of poor quality. The challenge of labeling with humans is that the cost of scaling to large collections of images is always going to be challenging. Even if there is the budget and supply to hire endless armies of annotators, oversight becomes more challenging as the group grows.

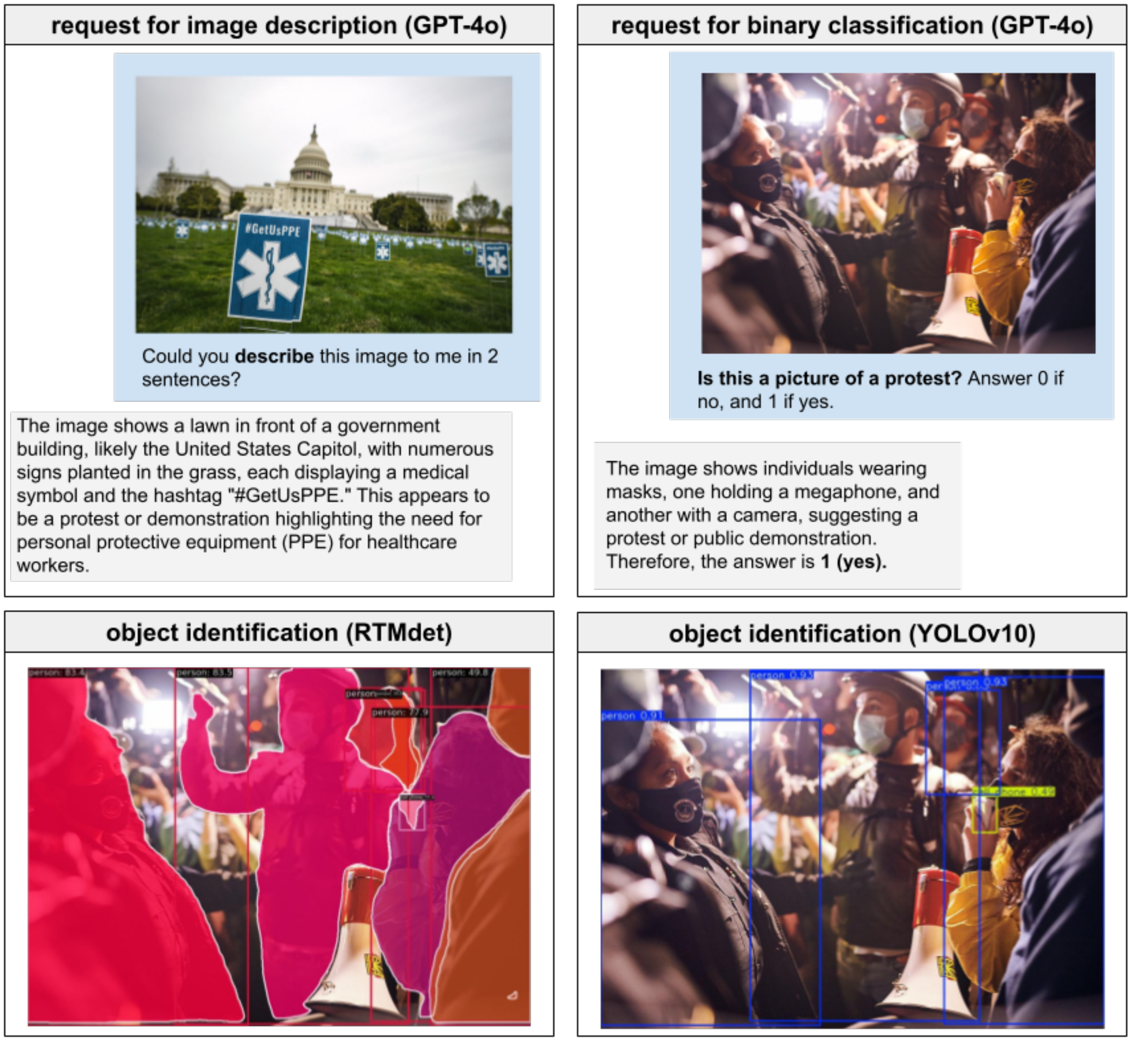
Computer vision provides a powerful toolkit for scaling visual annotations to arbitrarily large datasets (Webb Williams, Casas and Wilkerson, 2020; Joo, Bucy and Seidel, 2019). There are broadly two flavors of approaches: the classical supervised learning pipeline where a classifier is trained to perform a specified task and the few/zero-shot approach where a pretrained model learns the task from very limited cues. The supervised approach involves taking some human annotations to teach a machine learning algorithm—generally a convolutional neural network (CNN)—how to perform a specific task (Webb Williams, Casas and Wilkerson, 2020; Tarr, Hwang and Imai, 2023). The challenge for social scientists is that these models often require enormous training sets. With the median-size image dataset (to this point at least) being 25,000 images, training a CNN is less appealing than simply human annotating all the images.

Given the high-cost of human annotation (expert or not), it is not surprising that approaches which do not require many annotations—such as, few- and zero-shot learning—have quickly become popular. These methods use pre-trained models that are furnished with a handful (few-shot) or no (zero-shot) labeled examples. In case of text analysis, large language models have been shown to outperform crowd-sourced workers (Gilardi, Alizadeh and Kubli, 2023; Ziems et al., 2024; Do, Étienne Ollion and Shen, 2024) creating optimism that off-the-shelf tools might be able to deliver highly accurate annotations with little to no new human annotation. Multi-modal large language models, such as OpenAI's GPT-4o, can create open-ended descriptions of images and answer questions (which can be coding tasks). For example, in Figure 2 (top left) contains an example of an open-ended description of a political images. The remaining three components of Figure 2 show different tasks performed on the same image including answering a text-based question about whether the image contains a protest and two variants of object detection.

When the annotation task matches well with prior tasks in computer vision (e.g., object detection or face detection), these automated systems will perform quite well without further guidance. When the task is more subtle, high-performance can often still be achieved by providing a very small number of examples using few-shot learning. Returning to our example from Casas and Webb Williams (2019), the authors labeled images attached to tweets on a scale of 1–10 based on the emotion they evoke. This task is not-easily replicated by an automated system without any guidance. However, by providing GPT-4o with only 100 images, we generated annotations that produce very similar results in the downstream analysis (see Figure 3).[5]

While this is a promising proof of concept, it is not a panacea. We only know the system

[5] Here and throughout the paper we slightly deviate from the regression specification in Casas and Webb Williams (2019). We use the same variables but instead of a negative-binomial regression we use a linear regression with a logged outcome. As in their paper, the regression results omit some of the observed controls. Rather than standardize the coefficients (as they do) we divide the 0–10 scaled variables by 10 so that all independent variables range from 0 to 1. See appendix for the full specification.


request for image description (GPT-4o)
#GetUsPPE
request for binary classification (GPT-4o)
object identification (RTMdet)
person: 83.4
person: 83.5
person: 49.8
person
person: 77.9
object identification (YOLOv10)
person 0.93
person 0.91
person 0.93
cell phone 0.49

Could you **describe** this image to me in 2 sentences?

The image shows a lawn in front of a government building, likely the United States Capitol, with numerous signs planted in the grass, each displaying a medical symbol and the hashtag "#GetUsPPE." This appears to be a protest or demonstration highlighting the need for personal protective equipment (PPE) for healthcare workers.

Is this a picture of a protest? Answer 0 if no, and 1 if yes.

The image shows individuals wearing masks, one holding a megaphone, and another with a camera, suggesting a protest or public demonstration. Therefore, the answer is **1 (yes).**

Figure 2: **Example annotations from pre-trained computer vision systems.** (Top Left) A response from GPT-4o describing an image. (Top Right) A specific coding task posed as a question with an image, demonstrating how multi-modal large language models can be used to perform custom coding tasks. The model is deciding what a 'protest' is based on it's own world model rather than researcher-coded images (as would be the case in classical supervised learning). (Bottom Left/Right) Two variants of pre-trained object detection systems on the protest image (OpenAI et al., 2024; Wang et al., 2024; Lyu et al., 2022). Images courtesy of the Library of Congress (Public Domain).

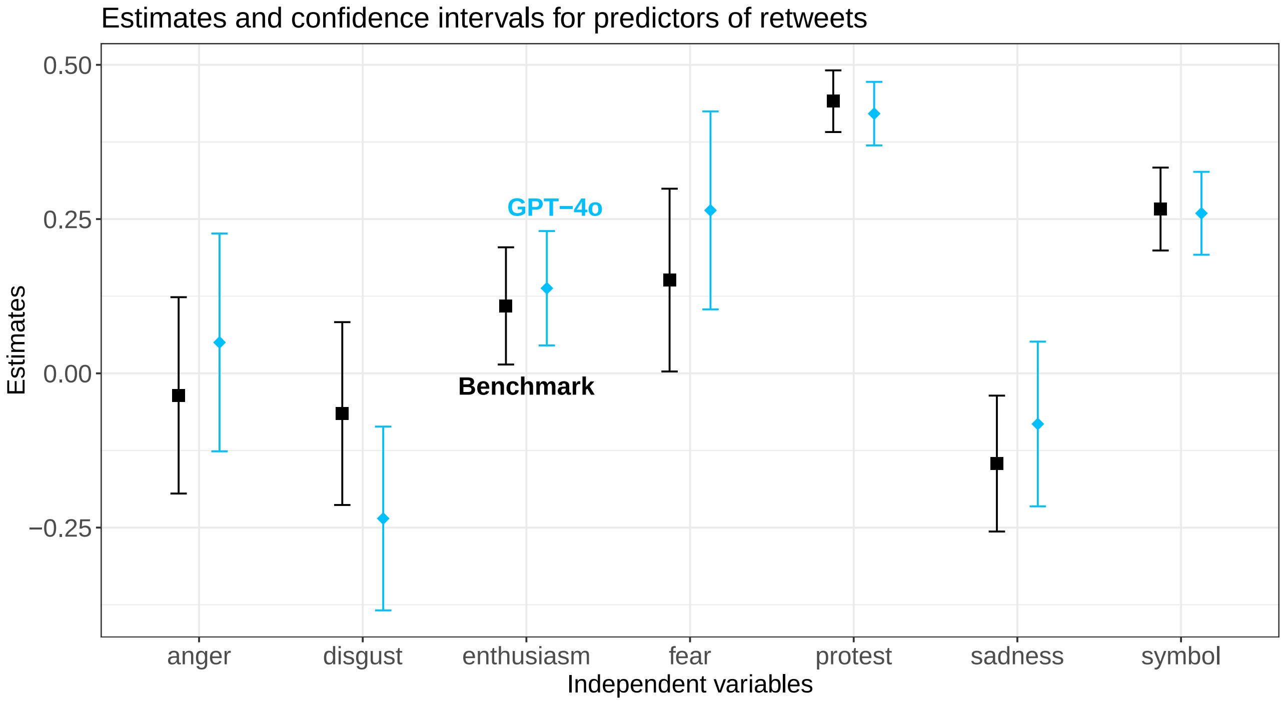


Figure 3: **Comparing downstream regression from expert annotation and GPT-4o.**. The image plots the coefficients for a regression based on Casas and Webb Williams (2019). Shown are coefficients of a linear regression with an outcome of the log of retweets plus one on the variables listed on the vertical axis including the emotions independent variables (*anger*, *fear*, *disgust*, *sadness*, and *enthusiasm*) and an indicator of whether the image was of a *protest* or contained a *symbol*. The "Benchmark" shows the coefficients and uncertainty estimates for the regression when using the human annotations in Casas and Webb Williams (2019). The "GPT-4o" replaces each of the five emotion variables with predictions based on GPT-4o using 100 in-context training examples.

is relatively accurate because we have the human-annotated data to compare against. Also, because the system is not perfectly accurate, there are discrepancies (notably in the emotion variables). While these particular discrepancies may not seem highly empirically relevant, as the sample size gets larger, the uncertainty estimates will shrink leading to ever more confident, but wrong, answers.

3 Design-based Supervised Learning

Design-based Supervised Learning (DSL) is a general framework for using predicted variables in downstream analysis that was developed by Egami et al. (2024*a*) and extended in Egami et al. (2024*b*). DSL builds on the doubly-robust estimation framework of Robins and Rotnitzky (1995) and Chernozhukov et al. (2018) to combine a set of expert annotations with predicted annotations in a way that retains desirable statistical properties.

In this section, we briefly review the terminology and intuition behind DSL, provide simulations to show why it is necessary, and then briefly discuss related work and alternatives.

3.1 Overview

DSL merges the complementary strengths of two annotation approaches: expert annotations are higher quality, but expensive, while automated 'prediction-based' annotations are scalable, but have unknown prediction errors. Using only one of them in the downstream analyses is suboptimal: researchers who only use expert-annotated data can only annotate a small sample of the data, while researchers who only use automated annotation methods will suffer from unknown large biases. DSL allows users to obtain statistically valid estimates, while gaining efficiency from modern computer vision techniques.

The basic steps of the proposed DSL can be summarized as follows.

Design-based Supervised Learning Estimator (DSL)

Step 1: Produce automated 'prediction-based' annotations (called "surrogates") for all images.

Step 2: Sample a subset of images for expert coding (called "expert annotations").

Step 3: Combine expert annotations and surrogates in the DSL regression.

Importantly, the first two steps are the same as what scholars already do when they generate automated annotations and check their accuracy. The third step is the key part of DSL, which we describe below. Researchers can use the software package in `R`, `dsl`, to implement the third step in a single line of code (similar to base models like `lm` and `glm`).

3.2 Intuition

The key intuition for DSL arises from the nature of our goal. We don't need to know the value of the expert annotation for *every* data point, we simply need to be able to approximate the downstream analyses we would conduct if we had all those expert annotations. A common example is the average expert annotation within some defined subgroups or over time.

The core assumption in DSL is Design-based Sampling of Expert Annotations. From Egami et al. (2024*b*),

> **Design-based Sampling for Expert Annotation**
> The probability of sampling observations for expert annotation π_i is known to researchers, and π_i is larger than zero for every observation i.

This assumption ensures that we can reweight the sample of expert annotated observations to look like the population using inverse probability weighting. The advantage of this design is that as long as the researcher has the entire collection of N images that they want to annotate available at the time annotation begins, this assumption is straightforward to *guarantee by design.* Most researchers randomly sample with equal probabilities when annotating, which corresponds to the special case where $\pi_i = 1/N$ for all i. The core insight of DSL is that we can use this random sample of expert annotations to construct an estimate of the bias in the surrogate measure and then simply adjust our estimator by removing that bias.

3.3 A Short Mathematical Introduction

To provide a quick introduction to DSL, we start with a case where the annotated visual component (e.g., whether an image contains a protest) is the outcome variable Y_i. When researchers naively use AI automated annotation systems, they would first predict whether the image is of a protest using a computer vision technique $\widehat{Y}_i$ and then use this predicted outcome variable directly in downstream analyses. To give the simplest example, they might estimate the average of Y, denoted μ, with the surrogate data as,

$$\hat{\mu}_{\text{Surrogate}} \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\widehat{Y}_i. \tag{1}$$

Instead of using $\widehat{Y}_i$, DSL uses the design-adjusted outcome,

$$\widetilde{Y}_i \quad = \quad \underbrace{\widehat{Y}_i}_{\substack{\text{Predicted}\\ \text{Outcome}}} \quad - \quad \underbrace{\frac{R_i}{\pi_i}(\widehat{Y}_i - Y_i)}_{\text{Bias-Correction Term}}, \tag{2}$$

where Y_i is the outcome of interest coded by experts, R_i is a binary variable taking 1 if image i is expert-coded and 0 otherwise, and π_i (defined in Section 3.2) is the probability of sampling image i for expert coding.[6] When researchers use data at hand to generate predictions, they should use cross-fitting such that separate data is used for prediction and de-biasing (Chernozhukov et al.,

[6] The design-adjusted outcome is equal to $\widehat{Y}_i$ when $R_i = 0$ and is equal to $\widehat{Y}_i - (\widehat{Y}_i - Y_i)/\pi_i$ when $R_i = 1$. There are two potentially counterintuitive things here. First, it might seem counter-intuitive to change the outcome for documents where $R_i = 1$ since these are the cases where we observe the "correct" annotation. Second, it might seem counter-intuitive that the design-adjusted outcomes don't obey the ranges of the original variable. For example, for a binary Y_i, $\tilde{Y}_i$ can—and often will—be outside the range of 0–1. The key to understanding the both is that the goal is not to correct the prediction error at each document level, but to correct the aggregate quantity. The design-adjusted outcomes, $\tilde{Y}$, are standing in collectively for both observations where $R = 1$ but also all the bias correction for the observations where $R = 0$. This estimator has deep theoretical connections to doubly robust estimation in the causal inference literature (Robins, Rotnitzky and Zhao, 1994; Chernozhukov et al., 2018), and the bias-correction term is similar to the one in the augmented inverse probability weighting estimator.

2018; Egami et al., 2024*b*). In the simplest case of random sampling with equal probabilities ($\pi = n/N$ where n is the number of expert-coded documents and N is the total number of documents), the DSL estimator simplifies to

$$\widehat{\mu}_{\text{DSL}} = \frac{1}{N}\sum_{i=1}^{N}\widetilde{Y}_i = \overbrace{\underbrace{\frac{1}{N}\sum_{i=1}^{N}\widehat{Y}_i}_{\substack{\text{Mean of all}\\ \textit{Predicted}\text{ Outcomes}}}}^{\widehat{\mu}_{\text{Surrogate}}} - \overbrace{\left(\underbrace{\frac{1}{n}\sum_{i:R_i=1}\widehat{Y}_i}_{\substack{\text{Mean of}\\ \textit{Predicted}\text{ Outcomes}\\ \text{in Expert Data}}} - \underbrace{\frac{1}{n}\sum_{i:R_i=1}Y_i}_{\substack{\text{Mean of}\\ \textit{Expert}\text{ Outcomes}\\ \text{in Expert Data}}}\right)}^{\text{Estimator of the Bias}}. \tag{3}$$

The main idea of DSL is to use the expert-coded data to estimate the bias of surrogate estimator using the prediction errors in the expert-annotated data. Then all we need to do is subtract off the bias to get a debiased estimator. To provide a concrete example, suppose we have $N = 10,000$ images and randomly sampled $n = 100$ for expert annotation of whether show a protest. The first term on the right-hand side estimates the proportion of protest images (i.e. where $Y_i = 1$) by averaging the predicted labels in all $N = 10,000$ documents (suppose it is 20%). The second and third terms on the right-hand side estimate the bias to be subtracted. In particular, the second term estimates the proportion of protest images by averaging the predicted labels in $n = 100$ expert-coded documents (suppose it is 18%), and the third term estimates the proportion of protest images by averaging the expert-coded labels in $n = 100$ expert-coded documents (suppose it is 10%). Because the expert-coded data are randomly sampled, we can estimate the bias by taking the difference between the second and third terms, $18 - 10 = 8$%, which we subtract from the first term. In this simple example, the DSL estimate is $20 - (18 - 10) = 12$%.

Variance of the DSL mean estimator is approximated by

$$\text{Var}\left(\widehat{\mu}_{\text{DSL}}\right) \approx \underbrace{\frac{1}{n}\text{Var}(Y_i - \widehat{Y}_i)}_{\text{Variance of Errors in Expert Data}} + \underbrace{\frac{1}{N}\text{Var}(\widehat{Y}_i)}_{\text{Variance in Unlabeled Data}}, \tag{4}$$

when N is much larger than n (as in most applications) (Angelopoulos et al., 2023).[7] Because N is large in most applications, the variance is mainly determined by the first term $\frac{1}{n}\text{Var}(Y_i - \widehat{Y}_i)$. This simple expression suggests that the variance of the DSL estimator is smaller (a) when prediction $\widehat{Y}_i$ is accurate and close to the observed outcome Y_i and (b) when the number of expert annotations is larger. This expression can be compared to the variance of the estimator that only uses expert annotations $\text{Var}(Y_i)/n$. The variance of the DSL is expected to be much smaller than this alternative estimator in most applications as $\widehat{Y}_i$ can explain variations in Y even though not perfectly.

The same principles also work when both (or any subset of) outcome and independent variables need to be annotated. Rather than construct a design-adjusted outcome, we estimate the bias of the underlying moment (cross-products of X and Y in the case of linear regression) and correct appropriately. The DSL coefficients for linear regression are formally defined as,

$$\widehat{\beta}_{DSL} = \left(\frac{1}{N}\sum_{i=1}^{N} M_i^{XX}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} M_i^{XY}\right) \tag{5}$$

[7] The exact variance is $\frac{1}{n}\text{Var}(Y_i - (1 - n/N)\widehat{Y}_i) + \frac{N-n}{N^2}\text{Var}(\widehat{Y}_i)$.

where

$$M_i^{XY} = \underbrace{\widehat{X}_i\widehat{Y}_i}_{\substack{\text{Predicted}\\\text{Variables}}} - \underbrace{\frac{R_i}{\pi_i}\left(\widehat{X}_i\widehat{Y}_i - X_iY_i\right)}_{\substack{\text{Bias-Correction}\\\text{Term}}}, \quad \text{and} \quad M_i^{XX} = \widehat{X}_i\widehat{X}_i^\top - \frac{R_i}{\pi_i}\left(\widehat{X}_i\widehat{X}_i^\top - X_iX_i^\top\right). \qquad (6)$$

Note that the form for each of the moments is the same as the simple example of averages above: the surrogate-only estimator minus the estimator for the bias in the expert-annotated data.

Here, we only provide examples with linear regression, but the DSL framework accommodates a large class of downstream analyses which can be estimated using the method of moments, including a class of generalized linear models (e.g., logistic, multinomial-logistic, Poisson, and linear fixed-effects regression). See Egami et al. (2024*b*) for more details.

We have also implemented a new technique, power-tuning (described in Appendix E) which guarantees that the DSL estimator will have a smaller asymptotic variance than the estimator that only uses expert annotation.

3.4 Practical Guidance

When the design-based sampling assumption holds, DSL guarantees that the downstream analyses are asymptotically unbiased and asymptotically Normal with a variance that can be estimated in closed form—even if the surrogate model is arbitrarily biased. The result is that confidence intervals will (asymptotically) attain nominal coverage. While this is an asymptotic property, we show below that even a couple of hundred observations is sufficient in real-world settings.

While we refer readers interested in the technical details to Appendix B of Egami et al. (2024*b*), we briefly highlight a few rules of thumb for using the method:

1. **You Can Have More Than One Surrogate:**
   While we mostly focus on the case of a single surrogate created using generative AI, in practice you can have many surrogates. These are all used to predict the expert-annotation so more (non-redundant) predictions can improve accruacy.

2. **Accurate Surrogates Improve Power:**
   The surrogates don't need to be accurate for the DSL properties to hold, but the more accurate they are the more precise your estimates will be. In practice including poor surrogates will simply widen your confidence intervals.

3. **Don't Use the Same Data to Train a Computer Vision Model and Debias:**
   The software package accepts the surrogate measure and the expert annotations for debiasing. Any training data used to create the surrogate measure should not be used for debiasing (because overfitting in the surrogate will make it look too accurate). Researchers can use cross-fitting (Chernozhukov et al., 2018; Egami et al., 2024*b*).

4. **Expert Annotations Can be Sampled with Stratification:**
   Ideally, you want expert annotated data that includes different classes and covers the range of your covariate space. This often means you want to perform stratified sampling conditional on the variables in the downstream regression and the value of the surrogate

itself (particularly when classes are very imbalanced). This is not essential, but it can help with precision. As long as the sampling weights are known, this is okay.

5. **The Required Number of Expert Annotations Does Not Increase As the Size of the Population Increases:**
   Some might think the required number of expert annotations is a certain proportion (e.g., 10%) of the total number of images they analyze, and thus, be worried that they have to expert annotate a large number of images if the total number of images is extremely large (e.g., millions). However, what matters to the standard errors of the DSL is the actual number of expert annotations (not the proportion). This is similar to the standard errors for survey sampling: regardless of whether you want to estimate the average support a particular policy in China or in New Jersey, we get approximately the same standard errors if we randomly sample 1000 subjects (even though 1000 people only constitute a tiny proportion of the entire population in China). When we have an accurate predicted surrogate, an increase in the population size (with a fixed number of expert annotations) will actually improve our standard errors. In practice, we recommend using our proposed data driven power analysis (illustrated in Section 4.1) to determine the number of expert annotations.

3.5 Simulations

We address two common questions researchers have when deciding whether to utilize DSL in their analysis:

1. *"My surrogate is very accurate. Why do I still need DSL?"*

2. *"Human experts are prone to errors. Does DSL still work if there are errors in the expert annotation?"*

To address these questions, we run simulations using a variant of the data-generating process in Wager and Athey (2018) to illustrate the relationship between errors in the surrogates, errors in the expert annotations, and the performance of the surrogate-only and DSL estimators.

The first set of simulations tests how bias, root mean squared error (RMSE) and coverage rates of confidence intervals change as we change the error rate in surrogates and number expert annotations. Figure 4 shows the results, which helps to clarify the answer to the first question. Even highly accurate surrogates (90%) achieve coverage well below 80%. Meanwhile, DSL consistently attains the nominal 95% coverage. Importantly, as the surrogate is more accurate (or the number of labeled examples grows), the RMSE for DSL improves.

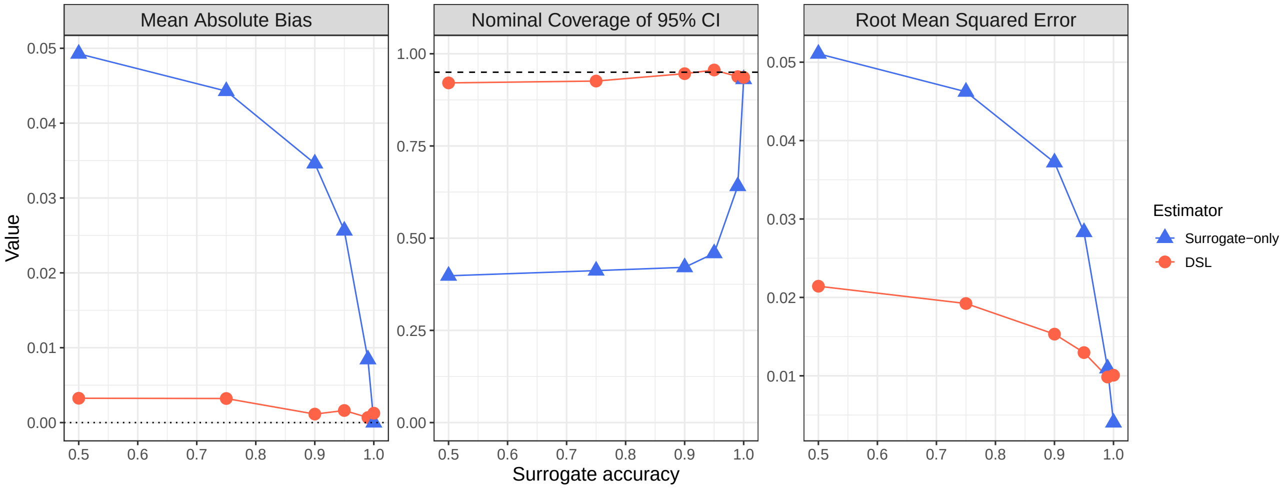


Figure 4: **Effect of surrogate accuracy on bias, RMSE and coverage**. Compares the Surrogate Only (SO, in blue triangles) and DSL (in red squares) estimators over 100 simulations for varying sizes of the expert annotated dataset. Even highly accurate surrogates can lead to poor coverage.

Having established that bias correction is necessary, we tackle the next common question about the accuracy of expert annotations. In the second set of simulations, we fix the surrogate accuracy to 0.75 and introduce errors into the expert labels (accuracy 0.75, 0.8, 0.9, 0.95, 0.99, 1.0). Figure 5 shows these results. As we might expect, when expert label accuracy is exceedingly low, coverage is also approaching the surrogate only baseline. However, mild deviations are not too devastating to bias. When there is any amount of bias, as we include more expert annotations with error, the coverage declines (as the estimator becomes more certain of the biased estimate). The RMSE results suggest that as long as the accuracy exceeds the surrogate accuracy, RMSE is better for DSL than for the surrogate only estimatorfor modest sample sizes. In short, expert annotation accuracy is undoubtedly important, but researchers don't need to ensure it is absolutely perfect. Finally, if researchers want to explicitly incorporate errors in expert annotations as additional uncertainties, they can apply the quasi-Bayesian approach introduced in Egami et al. (2024*b*).

3.6 Alternative Debiasing Approaches

DSL is just one of many approaches that have been developed to address settings where a variable in regression has been measured with error (Wang, McCormick and Leek, 2020; Fong and Tyler, 2021; Zhang, 2021; Knox, Lucas and Cho, 2022). While many other approaches make assumptions about the data-generating process, DSL is notable for only assuming that researchers have control over the sampling process for expert annotations—this makes it a particularly good fit for the rapidly changing world of generative AI.

DSL is directly built off work on doubly robust estimation (Robins, Rotnitzky and Zhao, 1994; Chernozhukov et al., 2018) to handle predicted variables. It is closely related to the contemporaneously-developed prediction-powered inference (Angelopoulos et al., 2023; Angelopoulos, Duchi and Zrnic, 2023), and model-assisted impact analysis (Mozer and Miratrix, N.d.). While these other approaches focus primarily on predicted outcome variables, DSL extends to

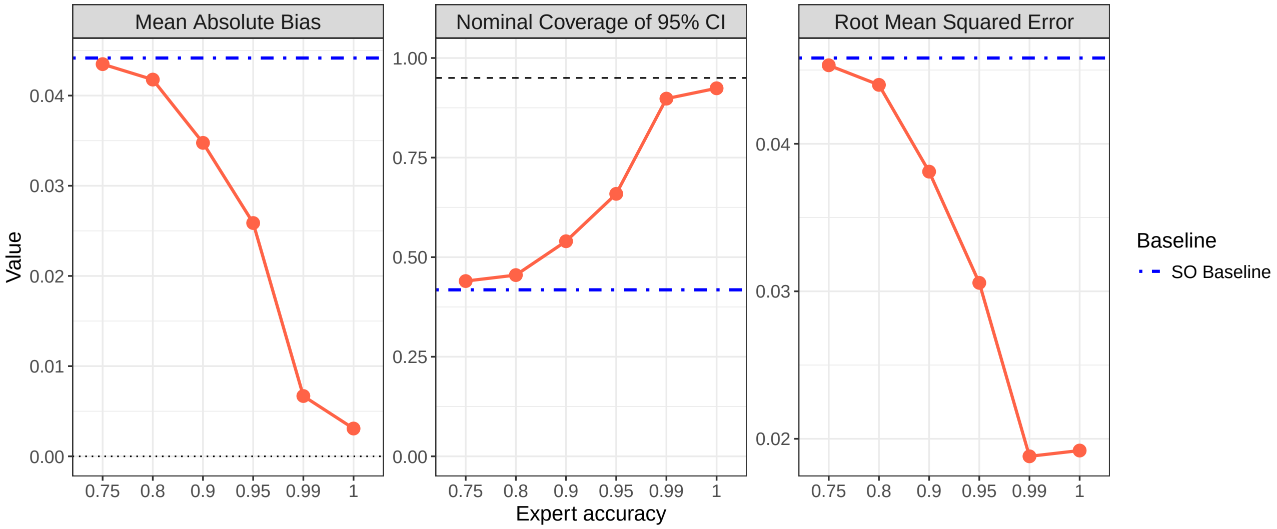


Figure 5: **Effect of expert annotation accuracy on bias, RMSE and coverage of DSL estimator.** Hue indicates the accuracy of the gold-standard annotations. Surrogate only (SO) baseline shows the bias, RMSE and coverage for the SO estimator. Any error in expert annotations will eventually lead to poor coverage, but as long as experts are more accurate than the surrogates, modest sample sizes lead to an improvement in RMSE.

cases where the predicted variables are any combination of the outcomes and independent variables. DSL also has the benefit of providing data-driven power analysis that helps users assess how much improvement they are likely to see in their standard errors from additional expert annotations. While DSL is our framework of choice, our argument is that suitable debiasing should be performed for any downstream analyses using annotations created by generative AI. We provide additional discussion in Appendix **??** about how these results are related.

4 Three Empirical Validations

To investigate the performance of DSL in practice, we conduct three empirical validations using computer vision applications in the social sciences (overview in Table 1). In each study, we use a complete set of human-annotated data as a benchmark. We pretend that we can only sample a subset of images for expert coding and use computer vision models to automatically annotate all the images. We can then assess how well DSL and other methods, which are based on a small number of expert annotations and a large number of automated annotated images, can recover the benchmark estimates, which use all the expert annotations. By doing so, we can showcase the use of DSL, while testing how DSL and other methods perform when the underlying computer vision techniques have non-random prediction errors.

Application	# of images	Downstream analysis	Topic	Original annotations
Won, Steinert-Threlkeld, Joo (2017)	2343	Dependent variable	Estimating the perceived level of violence in protest images	Both fully human annotated and estimated by surrogate
Casas, Webb Williams (2019)	7943	Independent variables	Studying the emotional reactions evoked by images shared on Twitter in the Black Lives Matter movement	All human annotators
Torres (2024)	688	Dependent variable	Detecting the portion of dense crowds in images from news articles on caravans	Surrogate only

Table 1: **Summary of empirical validation studies**. Note that the number of images here is the number of images used in this analysis, which a subset of the total images in (Casas and Webb Williams, 2019) and Won, Steinert-Threlkeld and Joo (2017) since we have removed 100 one-shot training images and non-protest images from the total set, respectively.

4.1 Won et al (2017): Annotation as Outcome

In Won, Steinert-Threlkeld and Joo (2017), the authors released the UCLA Protest Image Dataset. This dataset contains 40,764 images shared on social media including images of protests in Venezuela, Hong Kong, South Korea, and the United States (Women's March and Black Lives Matter protests). The authors included human annotations regarding a variety of visual attributes (whether the photo includes signs, fire, police, children, flags, shouting, a group larger than 20 individuals, and whether it was taken at night), in addition to the perceived level of violence, for 2,343 of these protest images. We will work with this set 2,343 annotated images for our validation study.

A natural question given this dataset is what visual features of a protest photograph are the most predictive of a high level of perceived violence. The question of what influences audiences to believe that a certain protest was violent or not is deeply relevant for social movements and media outlets, since the perceived violence is shown to enhance support for law-and-order policies (Baranauskas, 2022). For simplicity, in this first validation exercise, we treat the independent variables as known. We predict the outcome variable—the continuous level of perceived violence—using a multi-task convolutional neural network (ResNet) based on the proposal in Won, Steinert-Threlkeld and Joo (2017).

We randomly sample 600 images for expert annotations. DSL combines 2343 AI annotations and 600 expert annotations to perform the downstream analysis. Figure 6 (Top) shows the average point estimate and confidence interval across 500 repeated sampling to show the average performance across random sampling of expert coding. The 'Benchmark' represents the benchmark estimate, which uses all the expert annotations. The confidence interval for the 'Benchmark' is the tightest confidence interval that can be achieved and still maintain correct coverage since it is based on all of the expert-annotated data.

As expected by the theory, DSL is asymptotically unbiased and so the DSL confidence intervals are all centered at the same place as the benchmark point estimate. The confidence intervals are slightly wider than the Benchmark even though DSL uses only a fraction of the data. Of course in the real-world, DSL is run once (e.g. as shown in Appendix Figure **??**) and we have no guarantee that our one confidence interval will be centered in the right place (only that it will be on average!). In Appendix Figure **??**, we show a set of fifty DSL confidence intervals across random samples of the expert annotations to give readers a sense of what this

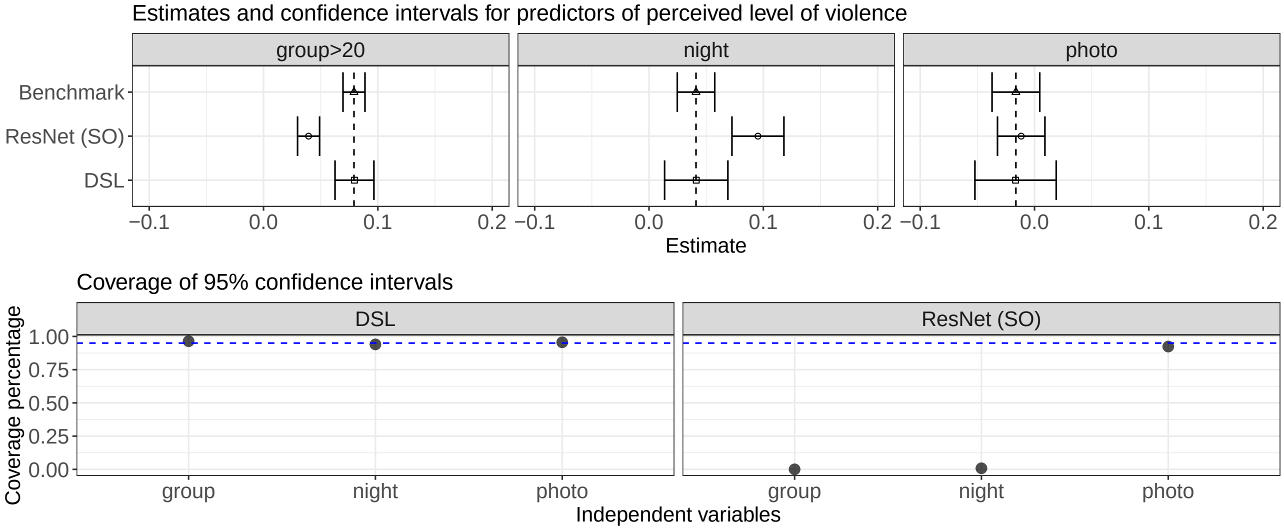


Figure 6: **Empirical validation study of Won et al. (2017).** (Top) We report benchmark estimates, DSL with 600 expert annotations and ResNet predictions, and estimators using the ResNet predictions alone. To show the average performance across random sampling of expert coding, we report the average point estimates and standard errors across 500 repeated sampling. (Bottom) Coverage of the 95% confidence intervals for each estimator across the 500 simulations.

variation looks like.

By contrast to DSL, the estimator directly using predictions from ResNet in downstream regression ("ResNet" in Figure 6) is biased and has a confidence interval that is *too* small (recall that anything tighter than the Benchmark is necessarily misleading). Figure 6 (Bottom) shows what this does to the coverage estimates. DSL slightly over-covers the nominal 95% level while ResNet attains essentially 0 coverage on all variables (DSL also attains nominal coverage with less than 400 annotations). While one variable ('photo') performs well with surrogate labels only, there is no way to know that a variable would perform well ex-ante.

The DSL estimator should always have a wider confidence interval than the (often-infeasible) Benchmark estimator that assumes all images are expert-annotated (because DSL only uses a sample of expert annotations in combination with the surrogates). A natural comparison point for DSL is the sub-sample estimator which *only* uses the sample of expert annotations (with no surrogates).Figure 7 shows the ratio of DSL confidence interval width to sub-sample estimator. Since all confidence intervals shown attain the nominal coverage, values below 1 are strictly better for DSL (since it implies that the confidence intervals are smaller while attaining the same coverage). This improvement can be thought of as what the surrogate predicted labels are doing to improve the estimator.

In practice, the researcher must select the number of expert annotations to complete. Egami et al. (2024*a*) show that you can perform a data-driven power analysis based on an initial set of annotated observations. Figure 8 compares the projected standard error for different numbers of expert annotations based on the first 400 and compares it with the actual standard error. These two track very closely which allows researchers to choose the number of expert annotations on the basis of what level of precision they need to attain to answer their question of interest.

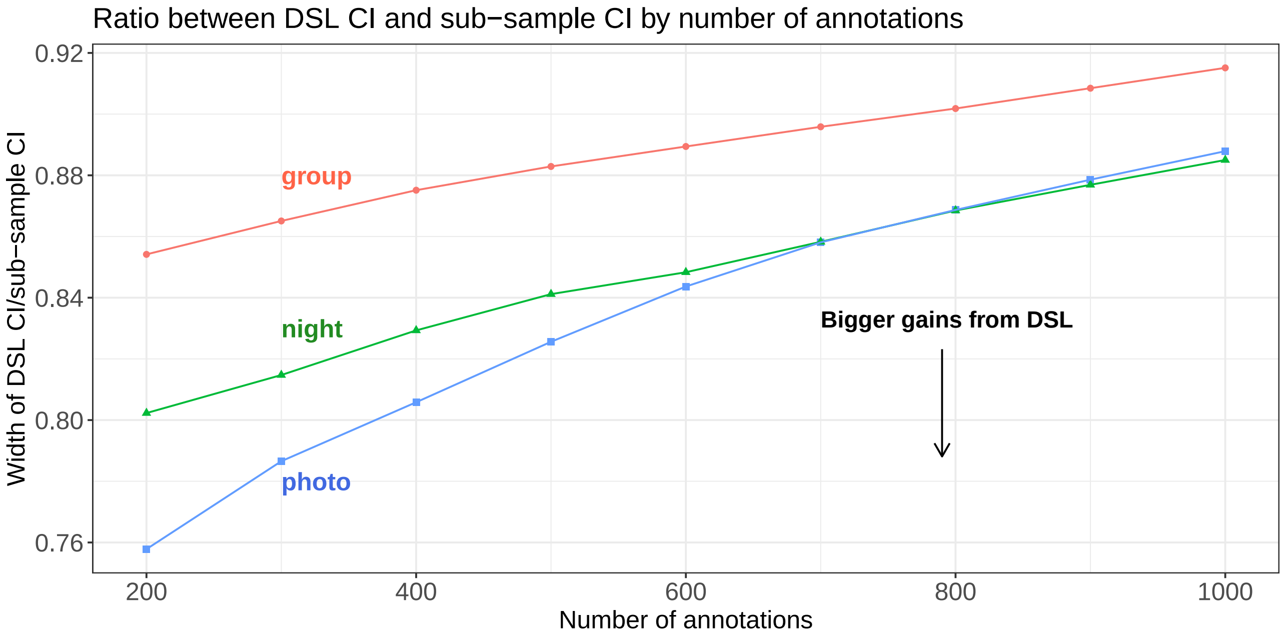


Figure 7: **Comparing the confidence interval width of the DSL and expert-annotation-only estimators**. For each of the coefficients in the Won, Steinert-Threlkeld and Joo (2017) model, we plot the ratio of the DSL confidence interval to the confidence interval of the estimator that assumes the same sub-sample of expert-annotated observations used by DSL. The gains from DSL are largest, when the size of the expert-annotated sample is small.

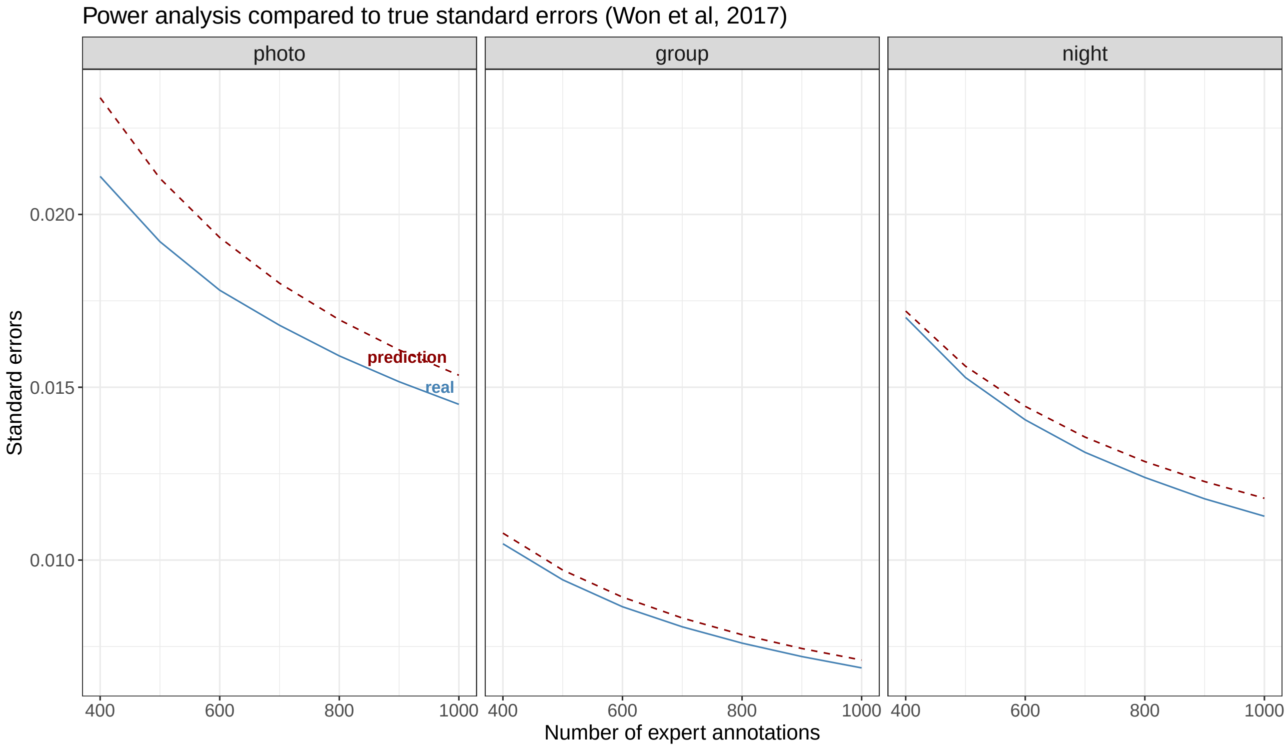


Figure 8: **Validating the data-driven power analysis.** The dashed line shows the projected standard error of the DSL estimator based on randomly sampled (once) 300 expert annotations. The solid line shows the actual standard error for different numbers of annotations (averaged over 500 simulations). The prediction from the data-driven power analysis is highly-accurate.

4.2 Casas and Webb Williams (2019): Annotation as IVs

As described in the introduction, Casas and Webb Williams (2019) investigate the relative attention given to images that provoke different emotions online. In the original article, Casas and Webb Williams (2019) manually annotated all 9,458 images for how much anger, disgust, sadness, fear, and enthusiasm each image evoked on a scale of 0–10 using a total of 1,259 Mechanical Turk annotators. We will treat these labels as the expert annotations.[8]

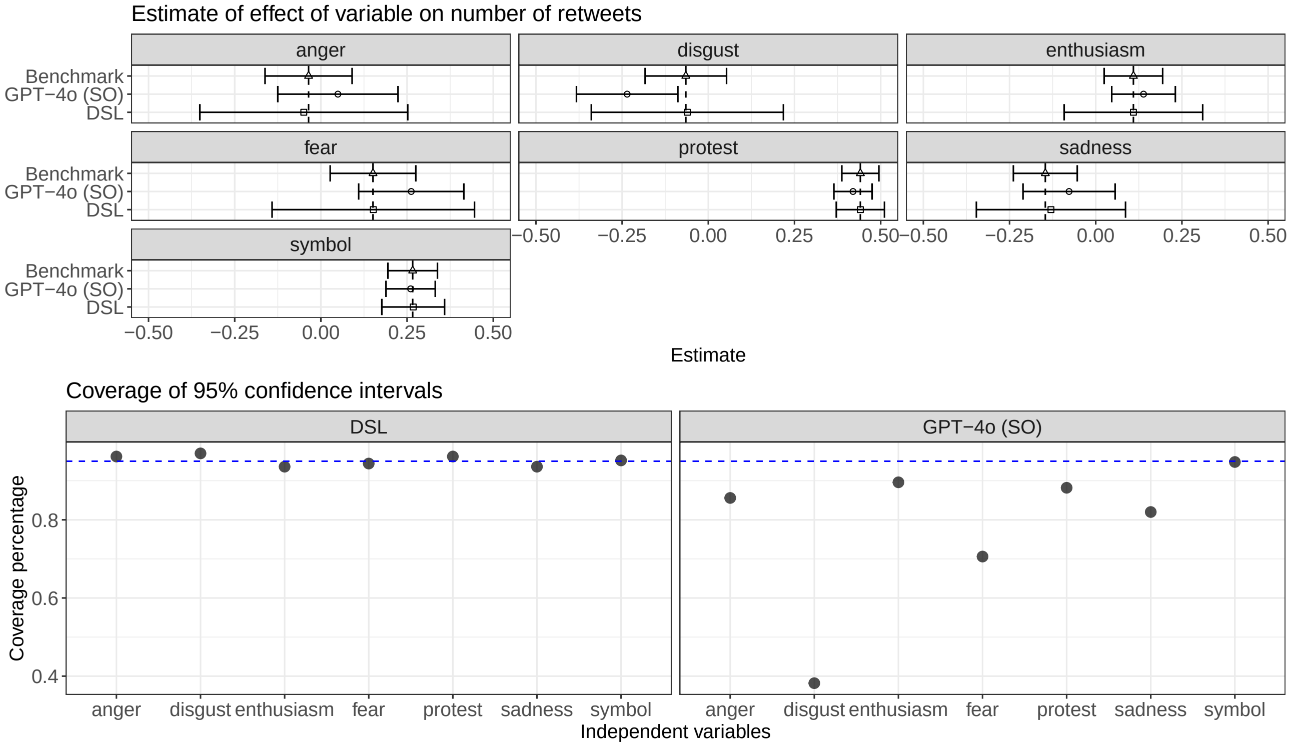


Figure 9: **Empirical validation study of Casas and Webb Williams (2019)**. (Top) We report benchmark estimates, DSL with 1200 expert annotations and GPT-4o predictions, and estimators using the GPT-4o predictions alone. To show the average performance across random sampling of expert coding, we report the average point estimates and standard errors across 500 repeated sampling. (Bottom) Coverage of the 95% confidence intervals for each estimator across the 500 simulations.

Since Casas and Webb Williams (2019) did not use any computer vision techniques to generate automated annotations, we use GPT-4o with 100 example images in the context window to automatically annotate the images. We repeatedly resample 1200 annotations to serve as the expert annotations in DSL and plot the average point estimate and confidence interval in Figure 9. In this example, predictions from GPT-4o are highly accurate, and as a result, estimators directly using GPT-4o predictions in downstream regression have relatively small biases across all coefficients, even though the coverage of its confidence interval for variable “disgust”

[8] In a real application, we would ideally like to have higher quality annotations. Driven by a concern about outliers, Casas and Webb Williams (2019) apply an extra layer of scrutiny for the top 949 unique most-tweeted images by having 5 people label them. Had DSL been available at the time, they could have used the MTurk labels as the surrogates and applied a higher-quality procedure to a randomly selected sample (e.g., the multi-label procedure they adopted).

is essentially zero. Even though this performance is optimistic, in the real-world application, we cannot observe the "Benchmark" estimates (unless they expert-annotate all images), so we do not know whether we are in this "lucky" scenario where prediction errors happen to cause small biases. In contrast, DSL has statistical guarantees: without assuming that prediction errors are small or random, researchers can reliably use DSL, which is asymptotically unbiased, and its confidence intervals have near nominal coverage.

4.3 Torres (2024) Annotation as Outcome

Torres (2024) presents a new unsupervised model of images, akin to a topic model for text. She builds a 'bag of visual words model' which allows her to inductively find a mixed membership over topics for each image. To demonstrate the model, she learns fifteen topic model on a set of 688 images that accompanies news articles on the Central American migrant caravans from October 3 to November 1, 2018. The topics discovered include 'Border/Fence', 'People Walking,' 'Small groups/Individuals' and 'Dense Crowd' among others.

To demonstrate the power of the model, she considers how the use of images containing dense crowds varies by the political leaning of the newspaper. She plots the average proportion of the images that fall into the 'dense crowd' topic by the ideology of the media outlet ('Right', 'Center-Right', 'Center', 'Center-Left', 'Left'). She finds that Right-leaning media sources show dense crowds of migrants at a substantially higher rate than other media sources. This can be re-framed as a regression where the outcome is the proportion of the image that contains a dense-crowd.

Torres (2024) does not provide a continuous expert-coded measure, so we collect our own.[9] In order to expert-code the portion of the images containing a dense crowd, we annotated each of the 688 images by drawing a bounding box to classify dense crowds.[10] We then compute the proportion of the image that is covered by a bounding box and use this as our expert annotation.

We consider two surrogate measures: the topic model measure produced by Torres (2024) and an off-the-shelf label produced by the Contrastive Language–Image Pre-training (CLIP) model (Radford et al., 2021). Figure 10 shows the results. Again, DSL performs well as the theory predicts, with the asymptotic properties of nominal coverage even holding with only 200 expert annotations.[11]

The Torres (2024) example involves a relatively small number of images, but it demonstrates that DSL can be used even in an unsupervised topic model workflow. In topic models we often assign topics with names, but that doesn't necessarily mean that the topic captures exactly the definition we would like it to (Grimmer, Roberts and Stewart, 2022; Ying, Montgomery and Stewart, 2022). DSL allows us to discover a concept of interest (in this case dense crowds) while still giving it an external, falsifiable definition. While the topic measure of dense crowds in Torres (2024) was fairly accurate, we quickly encountered difficult conceptual questions about

[9] Torres (2024) includes numerous other validity checks, including showing representative images. In her Figure 10, she compares her crowd topics to a binary human annotation of whether there is a crowd in the image. We wanted to collect new expert annotations though to have a continuous measure that is similar to her topic model-based measurement.

[10] We hired two undergraduate annotators to complete the coding task and then one of the authors checked all the results.

[11] Because the total set of images here is quite small, the DSL estimates and coverage for Right leaning media are a bit off. There are only 40 images from Right leaning sources in the entire dataset which can make the estimators quite sensitive.

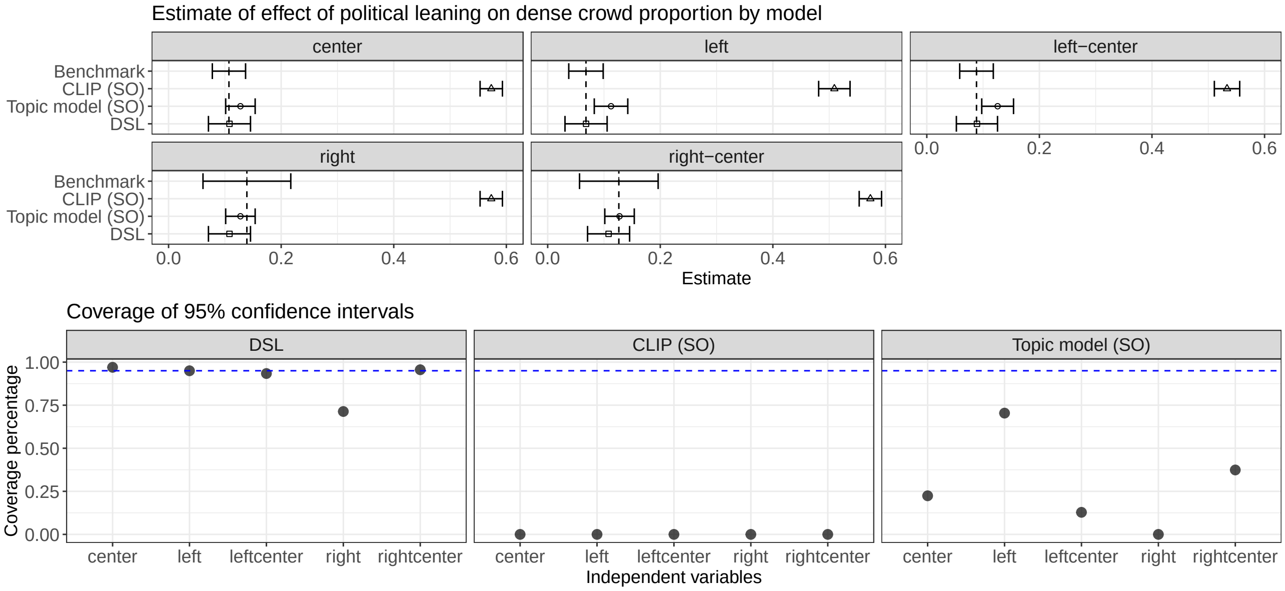


Figure 10: **Empirical validation study of Torres (2024).** (Top) We report benchmark estimates, DSL with 400 expert annotations and Topic Model predictions, and two estimators directly using automated annotations ("Topic Model" and "CLIP"). To show the average performance across random sampling of expert coding, we report the average point estimates and standard errors across 500 repeated sampling. (Bottom) Coverage of the confidence intervals for each estimator across the 500 simulations.

what was and was not a dense crowd—e.g. is the audience at a Trump rally a dense crowd? sort of, but it isn't a dense crowd of *migrants* the relevant point for this study. By annotating the images ourselves we were forced to take positions on these issues, providing an added benefit to the DSL workflow. Figure 11 highlights some of the differences between our expert coding and the Torres (2024) bag-of-visual-words topic model.

5 Conclusion

At the time of writing, the computer vision capabilities of generative AI models are rapidly improving. Multimodal large language models answer ever more complex questions about images with high accuracy, opening the door to more complex annotation schemes. This article is motivated by a core tension in how these generative AI tools interact with social science. The models often perform the tasks we want to do well, but we can't know for sure until we have checked them on a random sample—nullifying the promise of a model that completes our task without expert annotation. Even if the performance is shown to be strong, predicting one of our variables can cause us to lose the statistical guarantees of our downstream analysis. DSL provides a way to use that expert-annotated sample to recover those statistical guarantees.

DSL makes no assumptions about the quality of the surrogate measure and requires only that we choose units to expert code by a sampling scheme with known (non-zero) weights. This is an assumption that, in most research projects, can be guaranteed *by design*. The major limitation of DSL is that an expert must annotate a portion of the data, but this can often be as small as a few hundred. The precision that can be gained from additional annotations is

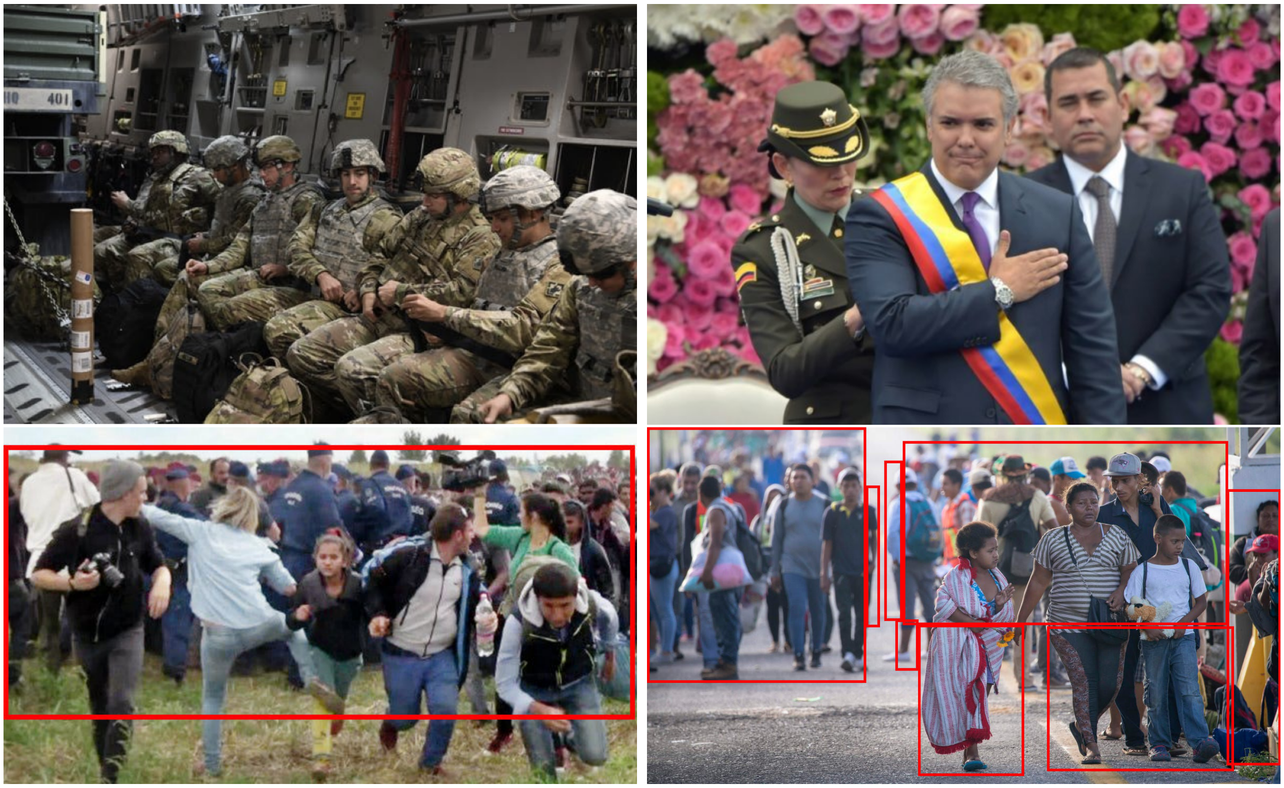


Figure 11: **Example Differences Between the Torres (2024) Topic Measure and Our Gold Standard** (Top Left/Right) The topic model scores both these images as being approximately 75% dense crowd whereas our annotation scheme did not mark any. (Bottom Left) The topic model scores this image as having less than 1% crowd, but our bounding boxes (depicted in red) cover almost 80% of the image. (Bottom Right) The topic model assigns this about 3% to the dense crowd topic, but our bounding boxes cover just over 75% of the image. Credit images: Daniel A. Hernandez/U.S. Air Force (top left), Fernando Vergara/AP (top right), AFP via Getty Images (bottom left), Nick Oza/USA Today (bottom right)

easily estimated using a data-driven power analysis. This process of annotation can also be a great way for the researcher to get more in touch with data—opening up difficult and important questions about how the latent concept is defined.

Most importantly of all, the properties of DSL do not depend on how the surrogate is created. As new generative AI tools are developed, we don't need to create new methods for downstream analyses—if the annotations are getting more accurate, our analyses will keep benefiting. Perhaps more strikingly, in being agnostic to the properties of how the surrogate is created, DSL can be applied to a much wider range of settings than just computer vision. Egami et al. (2024*a*) focus on the use of generative AI for text analysis and Angelopoulos et al. (2023) provide machine learning examples from proteomics to ecology. Our technique is applicable to most settings that involve missing (or mis-measured) data, a mechanism for prediction, and a way to obtain a higher-quality measurement of a randomly selected sample of observations. As machine learning and AI continue to grow, these settings will become ever more abundant!

Data and Code Availability Statement: We make our data and code available at Dataverse at `https://doi.org/10.7910/DVN/D9UGOV`. Some of the images we use cannot be shared publicly but we provide derivative products necessary to replicate our analyses and indicate whom to contact to request the original images.

References

Angelopoulos, Anastasios N, John C Duchi and Tijana Zrnic. 2023. "PPI++: Efficient Prediction-Powered Inference." *arXiv preprint arXiv:2311.01453* .

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan and Tijana Zrnic. 2023. "Prediction-powered inference." *Science* 382(6671):669–674.

Baranauskas, Andrew J. 2022. "News media and public attitudes toward the protests of 2020: An examination of the mediating role of perceived protester violence." *Criminology & Public Policy* 21(1):107–123.

Barocas, S., M. Hardt and A. Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts, USA: MIT Press.

Buolamwini, Joy and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR pp. 77–91.

Cantú, Francisco. 2019. "The fingerprints of fraud: Evidence from Mexico's 1988 presidential election." *American Political Science Review* 113(3):710–726.

Casas, Andreu and Nora Webb Williams. 2019. "Images that matter: Online protests and the mobilizing role of pictures." *Political Research Quarterly* 72(2):360–375.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21:C1 – C68.

Do, Salomé, Étienne Ollion and Rubing Shen. 2024. "The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy." *Sociological Methods & Research* 53(3):1167–1200.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts and Brandon M Stewart. 2022. "How to Make Causal Inferences Using Texts." *Science Advances* 8(42):eabg2652.

Egami, Naoki, Musashi Hinck, Brandon Stewart and Hanying Wei. 2024*a*. "Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models." *Advances in Neural Information Processing Systems* 36.

Egami, Naoki, Musashi Hinck, Brandon Stewart and Hanying Wei. 2024*b*. "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses." Working paper.

Fong, Christian and Matthew Tyler. 2021. "Machine learning predictions as regression covariates." *Political Analysis* 29(4):467–484.

Gilardi, Fabrizio, Meysam Alizadeh and Maël Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks." *Proceedings of the National Academy of Sciences* 120(30):e2305016120.

Girbau, Andreu, Tetsuro Kobayashi, Benjamin Renoust, Yusuke Matsui and Shin'ichi Satoh. 2024. "Face Detection, Tracking, and Classification from Large-Scale News Archives for Analysis of Key Political Figures." *Political Analysis* 32(2):221–239.

Grimmer, J., M.E. Roberts and B.M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, New Jersey, USA: Princeton University Press.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–297.

Hopkins, Daniel J and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1):229–247.

Joo, Jungseock, Erik P Bucy and Claudia Seidel. 2019. "Automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision and deep learning." *International Journal of Communication* pp. 4044–4066.

Joo, Jungseock and Zachary C. Steinert-Threlkeld. 2022. "Image as Data: Automated Content Analysis for Visual Presentations of Political Actors and Events." *Computational Communication Research* 4(1).

Knox, Dean, Christopher Lucas and Wendy K Tam Cho. 2022. "Testing Causal Theories with Learned Proxies." *Annual Review of Political Science* 25:419–441.

Lundberg, Ian, Rebecca Johnson and Brandon M Stewart. 2021. "What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86(3):532–565.

Lyu, Chengqi, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang and Kai Chen. 2022. "Rtmdet: An empirical study of designing real-time object detectors." *arXiv preprint arXiv:2212.07784* .

Mozer, Reagan and Luke Miratrix. N.d. "More power to you: Using machine learning to augment human coding for more efficient inference in text-based randomized trials." *Annals of Applied Statistics*. Forthcoming.

Nelson, Laura K. 2020. "Computational grounded theory: A methodological framework." *Sociological Methods & Research* 49(1):3–42.

Nelson, Laura K, Derek Burk, Marcel Knudsen and Leslie McCall. 2021. "The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods." *Sociological Methods & Research* 50(1):202–237.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt and Sam Altman et al. 2024. "GPT-4 Technical Report." *arXiv:2303.08774*.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR pp. 8748–8763.

Robins, James M and Andrea Rotnitzky. 1995. "Semiparametric efficiency in multivariate regression models with missing data." *Journal of the American Statistical Association* 90(429):122–129.

Robins, James M, Andrea Rotnitzky and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89(427):846–866.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115(3):211–252.

Tarr, Alexander, June Hwang and Kosuke Imai. 2023. "Automated Coding of Political Campaign Advertisement Videos: An Empirical Validation Study." *Political Analysis* 31(4):554–574.

Torres, Michelle. 2024. "A framework for the unsupervised and semi-supervised analysis of visual frames." *Political Analysis* 32(2):199–220.

Wager, Stefan and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113(523):1228–1242.

Wang, Angelina, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan and Olga Russakovsky. 2022. "REVISE: A tool for measuring and mitigating bias in visual datasets." *International Journal of Computer Vision* 130(7):1790–1810.

Wang, Ao, Hui Chen, Lihao Liu, Kai CHEN, Zijia Lin, Jungong Han and Guiguang Ding. 2024. YOLOv10: Real-Time End-to-End Object Detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wang, Siruo, Tyler H McCormick and Jeffrey T Leek. 2020. "Methods for Correcting Inference based on Outcomes Predicted by Machine Learning." *Proceedings of the National Academy of Sciences* 117(48):30266–30275.

Webb Williams, Nora, Andreu Casas and John D Wilkerson. 2020. *Images as data for social science research: An introduction to convolutional neural nets for image classification*. Cambridge, Massachusetts, USA: Cambridge University Press.

Won, Donghyeon, Zachary C Steinert-Threlkeld and Jungseock Joo. 2017. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*. pp. 786–794.

Ying, Luwei, Jacob M Montgomery and Brandon M Stewart. 2022. "Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures." *Political Analysis* 30(4):570–589.

Zhang, Han. 2021. "How Using Machine Learning Classification as a Variable in Regression Leads to Attenuation Bias and What to Do About It." SocArXiv: 453jk.

Zhang, Han, Christian Borch and Juan Pablo Pardo-Guerra. 2023. Analyzing Image Data with Machine Learning. In *The Oxford Handbook of the Sociology of Machine Learning*. Oxford, United Kingdom: Oxford University Press.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang and Diyi Yang. 2024. "Can large language models transform computational social science?" *Computational Linguistics* 50(1):237–291.

ONLINE Appendix to "Correcting the Measurement Errors of AI-assisted Labeling in Image Analysis using Design-based Supervised Learning"

Alessandra Rister Portinari Maranca[*] Jihoon Chung[†] Musashi Hinck[‡]

Adam D. Wolsky[§] Naoki Egami[¶] Brandon M. Stewart[‖]

March 5, 2025

A Model Details

A.1 Surrogate Model specifications

A.1.1 Won et al (2017) application

In this application, we have used the 50-layer ResNet model trained by (Won, Steinert-Threlkeld and Joo, 2017), as a surrogate for the percieved level of violence in the images. This multi-task convolutional neural network was trained on a separate dataset of human annotations on the level of violence of images. Their model architecture consists of 50 convolutional layers with batch normalization and ReLU layers. The authors use mean squared error to train violence dimension.

A.1.2 Casas et al (2019) application

We use GPT-4o in order to generate surrogates for the emotional reactions elicited in the images shared on Twitter OpenAI et al. (2024). We sample 100 images for few-shot training from the human annotations in Casas and Webb Williams (2019) where those 100 images were not used in the final analysis. We used GPT-4o in the 2024-05-13 deployment version. Due to context length constraints, for each query, we randomly sample 20 pairs from the few-shot training set and feed it as context learning. In the style of Maaz et al. (2024), we set the system prompt as:

[*] Graduate Student in Sociology, Princeton University

[†] Graduate Student in Computer Science, Princeton University.

[‡] AI Research Scientist, Intel Labs.

[§] Senior Research Specialist, Princeton University.

[¶] Corresponding Author. Assistant Professor, Department of Political Science, Columbia University. Email: naoki.egami@columbia.edu. URL: `https://naokiegami.com`.

[‖] Corresponding Author. Associate Professor, Department of Sociology and the Office of Population Research, Princeton University. Email: bms4@princeton.edu. URL: `https://brandonstewart.org`.

```
You are an intelligent chatbot designed to predict the human emotion evoked
when looking at an image. Your task is to look at an image and predict the type
and intensity of the human emotion it evokes.:
------
INSTRUCTIONS:
- Closely investigate the image.
- Explain what kind of emotion you will evoke from the given image.
```

And set the user prompt as :

```
Please look at the image and indicate the extent to which an image evoked each
of the five emotions: Anger, Enthusiasm, Fear, Sadness, and Disgust
Provide your amount of emotion only as an integer value between 0 and 10,
with 10 indicating the most evoked. Please generate the response in the form of
a Python dictionary string with keys being the emotion.
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python
dictionary string. For example, your response should look like this:
{’anger’: 5.0, ’enthusiasm’: 4.5, ’fear’: 5, ’sadness’: 0.5, ’disgust’:10.0}.
```

A.1.3 Torres (2024) application

We use the Transformers implementation of CLIP Radford et al. (2021), calculating the score between the image and the text “a photo of a crowd”, where higher score means that the image is close to the text. Since CLIP is highly successful with image/text matching Barraco et al. (2022), we are using the CLIP-generated score as a surrogate for the proportion of the topic “crowd” in a given image (similarly to Torres (2024)’s conceptualization). As CLIP-generated scores are unbounded, we rescale them between 0 and 1 in order to match the original paper (using $f(x) = \frac{x-\min(x)}{\max(x)-\min(x)}$).

A.2 Downstream regression specifications

A.2.1 Won et al (2017) application

Unlike the two other applications, this analysis is not in the results published by (Won, Steinert-Threlkeld and Joo, 2017), however, they represent a simple social scientific question that could be derived from their data. For this application, we ran a linear robust model with the following formula `violence ∼ sign + photo + fire + police + children + group_20 + flag + night + shouting`. Here, `violence` represents the perceived level of violence in a given image, which is both human annotated by (Won, Steinert-Threlkeld and Joo, 2017), and predicted by the author’s application of ResNet. The standard error type utilized here is “HC0”.

A.2.2 Casas et al (2019) application

The analysis of interest here is the robust linear model given by the following formula `log(retweet_n+1) ∼ followers_count + friends_count + prev_tweets + time_control + protest + symbol + anger + fear + disgust + sadness + enthusiasm`. The standard error type utilized here is “HC0”. It is important to note that here we are taking the $\log(x+1)$ of out outcome variable of interest (number of retweets) and then applying a linear model (in opposition to (Casas and

Webb Williams, 2019)'s application, where they use a negative binomial model. Other than that, the regression ran and the variables controlled for are identical. The only variables being predicted by the GPT4-o surrogate in the analysis are the emotional reactions, namely `anger, fear, disgust, sadness, enthusiasm`.

A.2.3 Torres (2024) application

The analysis performed here is a robust linear model with the formula `crow_proportion` $\sim$ `political_leaning -1`, where `crow_proportion` is predicted using CLIP scores, (Torres, 2024)'s bag of visual words model, or given by our ground truth human annotations. Moreover, `political_leaning` is the categorization made by (Torres, 2024) of the political leaning of the news outlet. The standard error type utilized here is "HC0".

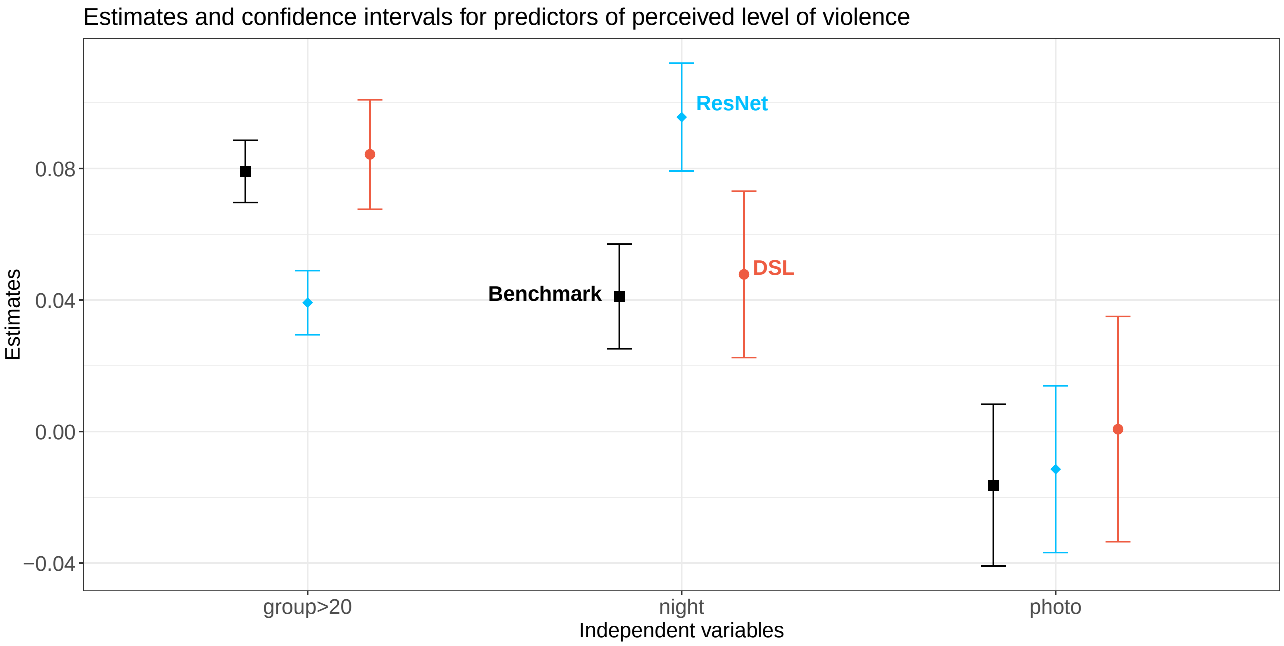


Figure 1: This shows the results of applying DSL once (rather than averaging over many results) to the Won, Steinert-Threlkeld and Joo (2017) application.

B Additional Results

Figures 1–3 contain additional results from the three empirical validations as described in their captions.

Figure 1 shows one application of DSL on the Won et al (2017) empirical validation instead of the avarage of the estimates (in this paper, horizontal estimates represent the result of avaraging 500 applications of DSL, whereas vertical estimates represent the result of applying the method once). We see that the DSL estimate tracks the Benchmark in all three coefficients while the ResNet surrogate underestimates, overestimates, and correctly estimates (overestimating the width of the confidence intervals) respectively for the coefficients showed.

Figure 2 shows the ratio between the DSL confidence intervals and the Oracle (Benchmark) confidence intervals in the Torres (2024) empirical validation for increasing number of labeled annotations used. Above, we see that DSL approximates the ground truth confidence interval width as we add more annotations. Below, we see that as we use more annotations, the marginal effect of using DSL in juxtapoisition to using the labeled annotations only decreases.

Figure 3 shows 50 applications of DSL in comparison to surrogate-only (ResNet), the same sub-sample of labeled annotations used for DSL, and the Benchmark (all labeled annotations) for the variable "police" in the Won et al (2017) empirical validation. We see the trend that DSL approximates the Benchmark better than the sub-sample alone or the surrogate (ResNet) alone.

Table 1 shows the effect of using additional labeled examples to fine-tune a model (and thus perhaps improving predictive capacity of the surrogate) versus using those additional labeled examples directly into DSL for the Casas et al (2019) empirical validation. We have mantained the total number of labeled examples used in the whole process (2000) constant while varying their use. We see that we get the smallest confidence intervals for all the variables in this

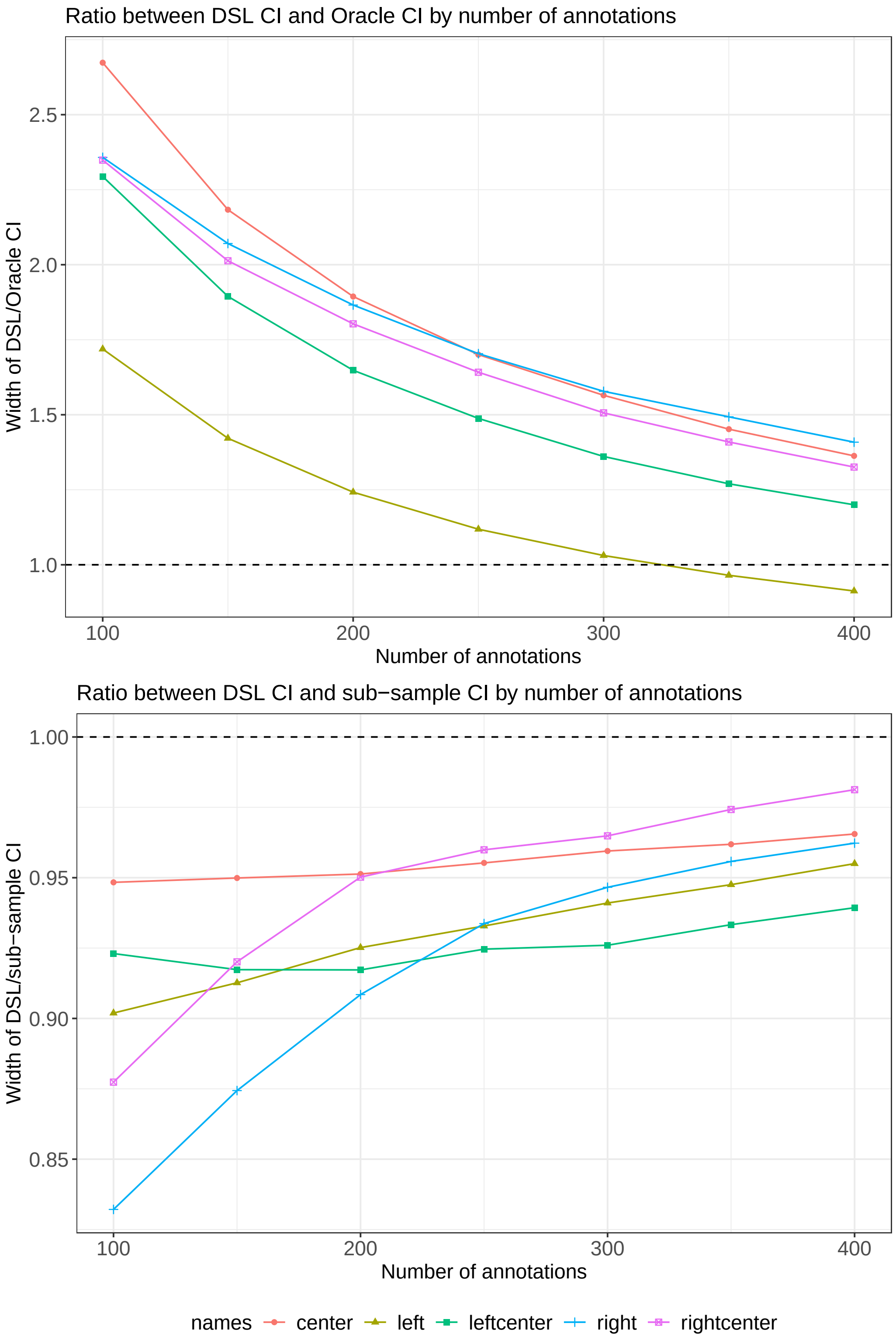


Figure 2: Confidence intervals in Torres (2024) by comparison to the oracle confidence interval and the sub-sample only confidence interval. Over 500 iterations.

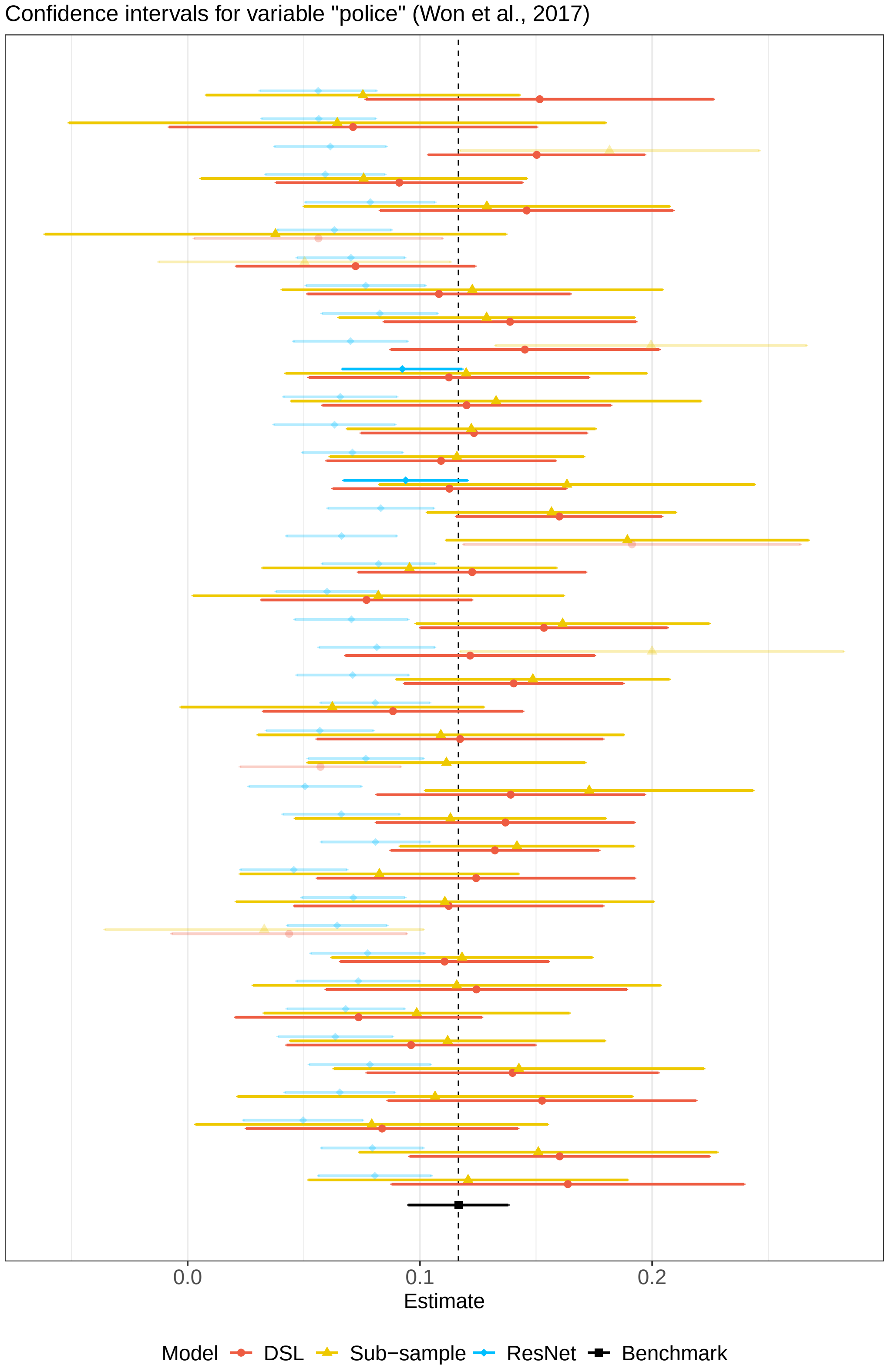


Figure 3: Results of 50 simulations for the variable 'police' in the Won et al empirical validation, for the model using DSL, with the labeled sub-sample observations without using DSL, and compared to ResNet (no labeled observations) and the benchmark (all 2343 labeled observations) applying with 200 labeled observations. In this plot, confidence intervals had their transparency increased if they did not contain the Benchmark estimate.

Type	1000DSL/1000FT	1500DSL/500FT	1900DSL/100FT
average CI width (protest)	0.16	0.13	0.13
average CI width (symbol)	0.21	0.18	0.17
average CI width (anger)	0.66	0.47	0.46
average CI width (fear)	0.70	0.56	0.51
average CI width (disgust)	0.60	0.45	0.43
average CI width (sadness)	0.47	0.34	0.31
average CI width (enthusiasm)	0.42	0.30	0.29
surrogate accuracy (anger)	0.63	0.49	0.70
surrogate accuracy (fear)	0.42	0.41	0.46
surrogate accuracy (disgust)	0.63	0.62	0.68
surrogate accuracy (sadness)	0.63	0.59	0.68
surrogate accuracy (enthusiasm)	0.55	0.72	0.63

Table 1: Results of experiment of maintaining 2000 labeled annotations constant in the Casas et al (2019), and testing whether it is more effective to use them for fine-tuning (FT) to generate more accurate surrogate annotations or to use them directly in DSL.

empirical validation when we minimize the amount of fine-tuning for GPT4-o and maximize the number of labeled examples into DSL (1900 benchmark annotations used for DSL, and 100 benchmark annotations used for fine-tuning). We also see that the surrogate accuracy does not increase linearly with adding more labeled annotations.

C Additional Details on Simulation

C.0.1 Simulation setup

For generating the simulations for Figures **??** and **??** which show the effect of surrogate accuracy on bias, RMSE, and coverage, and the effect of expert annotation accuracy on bias, RMSE, and coverage of DSL estimator (respectively), we have generated synthetic data. This data was generated analogously[1] to Wager and Athey (2018), with $n = 5000$ ('images', in our case), where Y was a smooth function supported on the first two features such as in Equation 1 below, however, we have also added normal error and a linear term of the other covariates.

We ran this simulation with 100 'labeled' examples and 10 covariates[2] $X_1, \ldots, X_{10}$ and introduced the prediction error in the surrogate for X_2. Note our covariates $X_1, \ldots X_{10}$ are initially drawn from a uniform distribution between 0 and 1, and then X_2 is transformed into a binary variable (1 if greater than 0.8, and 0 otherwise). Y, on the other hand, is generated by incorporating non-linear transformations of X_1 and X_2 and added linear contributions from additional covariates, alongside random normal noise. The data generating process of the outcome Y is

[1] To see the simulation this was based on, go to equation (28), p. 1238, in Wager and Athey (2018).

[2] The number of covariates used in the data generating process is 10, but we consider cases when we cannot observe all of them. We only use the first three in the downstream analysis. This shows that we do not need to assume the correct specification of the downstream model.

given by

$$Y = \left(1 + \frac{1}{(1 + e^{-20\cdot(X_1 - 1/3)}}\right) \cdot \left(1 + \frac{1}{1 + e^{-20\cdot(X_2 - 1/3)}}\right) + \sum_{i \in \{4,5,6,7,8,9,10\}} X_i + \varepsilon \tag{1}$$

where $\varepsilon \sim N(0, 1)$.

Then, a surrogate for X_2 is created with specified accuracy (denoted `q_acc` in the figures) by probabilistically altering the original X_2 values. Then, random binary errors in the Benchmark labeled examples are introduced to simulate expert annotation errors. For the downstream analysis showed in the figures, we are fitting the following model: $Y \sim X_1 + X_2 + X_3$ where X_2 is the binary surrogate.

C.0.2 Diagnosands

Denoting the simulation iteration as $i \in N$ and the target coefficient as $\beta_{\text{Estim.}}$, the diagnosands in our simulations are calculated as follows:

- **Mean Absolute Bias**: the average absolute difference between the estimate and true parameter value ($\frac{1}{N}\sum_i^N abs(\beta_{DSL,i} - \beta^*)$, where i denotes simulation iteration, N denotes the number of simulations and $\beta^* = \frac{1}{N}\sum_i^N \beta_{oracle,i}$)
- **Root Mean Squared Error (RMSE)**: the average RMSE
- **Coverage of 95% Intervals**: the proportion of simulations for which the true parameter value is in the confidence interval provided by the estimator

D Connection to Literature

Theoretically, our paper is most closely related to two recent papers that similarly build on the literature on doubly robust methods. Prediction-powered inference Angelopoulos et al. (2023) provides a similar framework to ours, but they have primarily focused on settings where the outcome variable is predicted while providing both asymptotic and non-asymptotic confidence intervals. Mozer and Miratrix (N.d.) focus on settings where the predicted outcome variable is used within randomized experiments. Methodologically, our paper extends these previous results in three ways. First, while these papers only cover cases of text-based outcome variables, we cover cases where any subset of the outcome and independent variables are text-based. Second, we develop a data-driven power analysis to help users determine the required number of expert annotations. Third, we derive DSL estimators for a much wider range of downstream analyses popular in the social sciences, including linear fixed effects regression and the instrumental variable method. In addition, we make practical contributions by providing new statistical software and clarifying detailed guides using two empirical applications. While they all share the same methodological foundation, each paper contains its own unique contributions tailored to different applications they focus on: our paper focusing on the social sciences, prediction-powered inference focusing mostly on the natural sciences, and model-assisted impact analysis focusing on education and randomized experiments. Given the wide applicability of the shared methodological foundation, we expect more exciting methodological developments that address various application-specific problems. Katsumata and Yamauchi (2023) also develop a framework for using predicted variables while building on a different framework of control variates (Chen and Chen, 2000).

E Power-Tuning

Following Angelopoulos, Duchi and Zrnic (2023), we implement the power-tuning that guarantees that the DSL estimator has a smaller asymptotic variance than the estimator that only uses expert annotations.

Using the simplified notation (see more full results in Egami et al. (2024)), we consider the following general moment condition.

$$m_{\mathrm{DSL}}(D_i, \widehat{D}_i, R_i; \beta, \pi) \coloneqq m(\widehat{D}_i; \beta) - \frac{R_i}{\pi_i}\Big(m(\widehat{D}_i; \beta) - m(D_i; \beta)\Big) \tag{2}$$

where D_i is a vector of observed variables, $\widehat{D}_i$ is a vector of predicted variables, R_i is an indicator binary whether a unit is sampled for expert annotations, and π_i is the known probability of being sampled for expert annotations. Note that some of $\widehat{D}_i$ can be observed and might not need to be predicted.

The power-tuning version (Angelopoulos, Duchi and Zrnic, 2023) has an additional tuning parameter λ.

$$m_{\mathrm{DSL},\lambda}(D_i, \widehat{D}_i, R_i; \beta, \pi) \coloneqq \frac{R_i}{\pi_i} m(D_i; \beta) + \lambda\Big(1 - \frac{R_i}{\pi_i}\Big) m(\widehat{D}_i; \beta) \tag{3}$$

The asymptotic variance of this estimator is

$$\lambda^2 \times S_V P_1 S_V - \lambda \times S_V P_2 S_V + P_3 \tag{4}$$

where

$$\begin{aligned}
S_V &= \mathbb{E}\left(\frac{\partial m(D_i; \beta^*)}{\partial \beta}\right)^{-1} \\
P_1 &= \mathbb{E}\left(\frac{1}{\pi_i} m(\widehat{D}_i; \beta^*) m(\widehat{D}_i; \beta^*)^\top\right) - \mathbb{E}\Big(m(\widehat{D}_i; \beta^*) m(\widehat{D}_i; \beta^*)^\top\Big) \\
P_2 &= \mathbb{E}\left(\frac{1}{\pi_i} m(\widehat{D}_i; \beta^*) m(D_i; \beta^*)^\top\right) - \mathbb{E}\Big(m(\widehat{D}_i; \beta^*) m(D_i; \beta^*)^\top\Big) \\
&\quad + \mathbb{E}\left(\frac{1}{\pi_i} m(D_i; \beta^*) m(\widehat{D}_i; \beta^*)^\top\right) - \mathbb{E}\Big(m(D_i; \beta^*) m(\widehat{D}_i; \beta^*)^\top\Big) \\
P_3 &= \mathbb{E}\left(\frac{1}{\pi_i} m(D_i; \beta^*) m(D_i; \beta^*)^\top\right).
\end{aligned}$$

We tune λ to minimize the trace of the asymptotic variance.

$$\lambda^* = \frac{\mathrm{Tr}\{S_V P_2 S_V\}}{2\mathrm{Tr}\{S_V P_1 S_V\}}. \tag{5}$$

References

Angelopoulos, Anastasios N, John C Duchi and Tijana Zrnic. 2023. "PPI++: Efficient Prediction-Powered Inference." *arXiv preprint arXiv:2311.01453* .

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan and Tijana Zrnic. 2023. "Prediction-powered inference." *Science* 382(6671):669–674.

Barraco, Manuele, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi and Rita Cucchiara. 2022. The unreasonable effectiveness of CLIP features for image captioning: an experimental analysis. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4662–4670.

Casas, Andreu and Nora Webb Williams. 2019. "Images that matter: Online protests and the mobilizing role of pictures." *Political Research Quarterly* 72(2):360–375.

Chen, Yi-Hau and Hung Chen. 2000. "A Unified Approach to Regression Analysis under Double-Sampling Designs." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 62(3):449–460.

Egami, Naoki, Musashi Hinck, Brandon Stewart and Hanying Wei. 2024. "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses." Working paper.

Katsumata, Hiroto and Soichiro Yamauchi. 2023. "Statistical Analysis with Machine Learning Predicted Variables." Working Paper.

Maaz, Muhammad, Hanoona Rasheed, Salman Khan and Fahad Shahbaz Khan. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Mozer, Reagan and Luke Miratrix. N.d. "More power to you: Using machine learning to augment human coding for more efficient inference in text-based randomized trials." *Annals of Applied Statistics*. Forthcoming.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt and Sam Altman et al. 2024. "GPT-4 Technical Report." *arXiv:2303.08774*.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR pp. 8748–8763.

Torres, Michelle. 2024. "A framework for the unsupervised and semi-supervised analysis of visual frames." *Political Analysis* 32(2):199–220.

Wager, Stefan and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113(523):1228–1242.

Won, Donghyeon, Zachary C Steinert-Threlkeld and Jungseock Joo. 2017. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*. pp. 786–794.