Sunlight: Fine-grained Targeting Detection at Scale with Statistical Confidence

Mathias Lecuyer, Riley Spahn, Yannis Spiliopoulos,
Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu

(mathias,riley,yannis,augustin,roxana,djhsu@cs.columbia.edu)

ABSTRACT

We present *Sunlight*, a system that detects the causes of targeting phenomena on the web – such as personalized advertisements, recommendations, or content – at large scale and with solid statistical confidence. Today's web is growing increasingly complex and impenetrable as myriad of services collect, analyze, use, and exchange users' personal information. No one can tell who has what data, for what purposes they are using it, and how those uses affect the users. The few studies that exist reveal problematic effects – such as discriminatory pricing and advertising – but they are either too small-scale to generalize or lack formal assessments of confidence in the results, making them difficult to trust or interpret.

Sunlight brings a principled and scalable methodology to personal data measurements by adapting well-established methods from statistics for the specific problem of targeting detection. Our methodology formally separates different operations into four key phases: scalable hypothesis generation, interpretable hypothesis formation, statistical significance testing, and multiple testing correction. Each phase bears instantiations from multiple mechanisms from statistics, each making different assumptions and tradeoffs. Sunlight offers a modular design that allows exploration of this vast design space. We explore a portion of this space, thoroughly evaluating the tradeoffs both analytically and experimentally. Our exploration reveals subtle tensions between scalability and confidence. Sunlight's default functioning strikes a balance to provide the first system that can diagnose targeting at fine granularity, at scale, and with solid statistical justification of its results.

We showcase our system by running two measurement studies of targeting on the web, both the largest of their kind. Our studies – about ad targeting in Gmail and on the web – reveal statistically justifiable evidence that contradicts two Google statements regarding the lack of targeting on sensitive and prohibited topics.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy, Ethics, Use/abuse of power*

Keywords

web transparency; privacy; measurement

CCS'15, October 12 - 16, 2015, Denver, CO, USA.

ACM 978-1-4503-3832-5/15/10$15.00
DOI: http://dx.doi.org/10.1145/2810103.2813614.

1 Introduction

In a 1913 paper [7], Louis Brandeis, the proponent of modern views of individual rights to privacy, stated: *"Sunlight is said to be the best of disinfectants; electric light the most efficient policeman."* Unfortunately, today's Web is a very dark and complex ecosystem driven to a large extent by the massive collection and monetization of personal data. Myriad of Web services, mobile applications, and third parties are collecting large amounts of information from our daily online interactions, such as our website visits, clicks, emails, documents, and pictures. At a surface level, end-users and researchers alike largely understand that these companies may be using this information to target advertisements, customize recommendations, personalize news feeds, and even fine-tune prices. Indeed, the companies' own terms of service often stipulate such uses. But at a concrete level, neither end-users nor researchers – nor individual companies (as we argue) – understand how specific personal data flows through the complex web ecosystem, how it is being used (or abused) in practice by parties that interact with it, and how those uses affect the users.

Questions about the targeting on the web abound: Are our children's online activities being targeted, and if so what kinds of products are they being offered? Are people being targeted because their browsing patterns suggest that they might be vulnerable (e.g., sick, depressed, or in financial difficulty)? Are such inferences being used to increase insurance premiums, deny housing, or place potentially damaging products, such as alcoholic products or risky mortgage deals? In other words, is our data being used without our knowledge or consent in ways that affect us? Today, we lack believable, at-scale answers to such questions.

A good way to shed light on large, complex systems is to measure them at scale using scientific methods and tools. Indeed, a number of measurement studies have emerged in recent years, which attempt to answer questions about how personal data is being used on the web [2, 6, 8, 15, 16, 19–22, 27, 29]. We reviewed 12 of these studies and found a significant gap in scalable experimental methodologies. Generally speaking, prior studies conduct tightly controlled experiments that vary personal data inputs (such as location, search terms, or profile interests) *one at a time* and observe the effect on service outputs (such as ads, recommendations, or prices) compared to a control group. Unfortunately, we find the methodologies employed by prior studies either difficult to scale or lacking formal notions of confidence, which make their results difficult to trust, interpret, and generalize. Consequently, the scale and scope of what we can assert today about data use on the web is limited, and our capacity to exert oversight on this large, complex, and ever-changing ecosystem is virtually non-existent.

This paper argues that shedding light into the web's complex data ecosystem requires the development of robust experimental

methodologies, as well as infrastructures that implement them, that can be used to answer broad classes of questions at large scale and with interpretable, trustworthy, statistical justification of the results. We present *Sunlight*, a new methodology, plus a system that implements it, that achieves these goals in the context of one important class of questions: those that require a fine-grained measurement of the causes of targeting phenomena on the web. All the questions raised at the start of this section can be addressed with Sunlight.

The Sunlight methodology builds upon robust statistical methods to support scalable, trustworthy, and interpretable results. Our key innovation is to formally separate various operations into multiple interacting stages organized in a pipeline and identifying all the right building blocks from statistics and machine learning to leverage at each stage of the pipeline. The Sunlight pipeline analyzes the data collected from an experiment that tries many different inputs *at once*, placing them at random in a small number of user accounts (logarithmic in the number of inputs) and collecting outputs from each account. The Sunlight pipeline analyzes the data to reveal which specific input likely caused which output. The first stage, *scalable hypothesis generation*, creates a set of plausible targeting hypotheses regarding which specific inputs correlate with which outputs. It leverages sparsity properties to support the *simultaneous* estimation of the effect of multiple inputs on the outputs, a consequence of the same phenomenon that underlies compressed sensing [9]. If needed, the second stage, *interpretable hypothesis formation*, converts the targeting hypotheses to an interpretable form that Sunlight users (such as auditors or researchers) can readily understand. The third stage, *hypothesis testing*, establishes the statistical significance of the interpretable, plausible targeting hypotheses by testing their veracity in a separate, testing dataset initially carved out from the collected data but never used until this stage. In some circumstances, specialized tests can establish causation and not just correlation between the inputs and the outputs. Finally, the fourth stage, *multiple testing correction*, accounts for the testing of many hypotheses on the same dataset, which increases the chance of any individual hypothesis being wrong. The end result are validated, interpretable hypotheses about which inputs are targeted by each output, along with a statistical significance score (a *p-value*) for each hypothesis.

Sunlight implements this methodology in a modular way, which supports both the instantiation and the evaluation of each stage based on multiple building blocks from statistics and machine learning. We find that different mechanisms lead to different trade-offs between the scale of and the confidence in the results, hence Sunlight lets its users choose end-to-end pipelines that best fit their needs. Development of effective such pipelines from existing building blocks is surprisingly challenging, as different mechanisms interact in unexpected ways in the pipeline. For example, our detailed evaluation of various Sunlight pipelines reveals counterintuitive inversions of recall near the start of the pipeline and at the end. Indeed, substituting a mechanism for generating hypotheses in Stage 1 with one that has higher recall but lower precision, may ultimately lower the recall at the end of the pipeline. The reason is that the multiple testing correction at Stage 4 tends to favor those mechanisms that generate fewer but more accurate hypotheses.

This paper discusses and evaluates the inherent trade-offs in web transparency measurement designs, bringing the following contributions to this emerging research topic:

1. A review of 12 recent articles on web transparency measurement and tools, which highlights the need for new, principled methodologies for scalable and trustworthy web transparency measurements. (§2)
2. The first methodology for detecting targeting in large-scale experiments with interpretable and statistically justifiable results. While our methodology focuses on our specific problem – fine-grained targeting detection – we believe that its conceptual bearings are relevant to other web transparency problems (e.g., price discrimination studies at scale). (§3)
3. The first system that implements this methodology to detect targeting at fine granularity, at scale, and with solid statistical justification of its results. Sunlight is modular, allows broad design space explorations, and customization of its pipeline to strike varied trade-offs of confidence and scale. (§4)
4. A detailed evaluation of Sunlight with comparisons of multiple design options and prior art. Our evaluation methodology is new in itself, and (we believe) a useful starting point for future transparency infrastructures, an area that currently lacks rigorous evaluations. Our results reveal a trade-off between the statistical confidence and number of targeting hypotheses that can be made. They also show that favoring high precision algorithms can yield better recall at high confidence, and that scaling output numbers may require to accept lower statistical guarantees to find sufficient hypotheses. (§6)
5. Results from analyzing targeting of tens of thousands of ads in two ecosystems: Gmail and the broader Web. Results reveal a large and diverse collection of ads targeting websites across many categories, including ads that appear to contradict explicit statements made by Google about targeting on sensitive topics, as well as advertising network policies about ads facilitating recreational drug use. (§5 and §7).
6. Sunlight's source code and datasets. (`https://columbia.github.io/sunlight/`)

2 Motivation

At an abstract level, our work is motivated by our desire to understand how to build principled, scalable infrastructures that can bring visibility to today's dark data-driven web. Such infrastructures must be able to detect data flows at great scale and in complex, heterogeneous environments, and provide trustworthy assessments about these data flows. We believe there is urgent need for such infrastructures (which we term generically *web transparency infrastructures*), yet we find limited progress in the related literature.

At a concrete level, this paper describes our experience building one such scalable and trustworthy[1] infrastructure, *Sunlight*, which aims to discover data flows in a specific context: detecting the causes of targeting phenomena at fine granularity from controlled experiments with differentiated inputs. This section begins by motivating the need for targeting detection systems in particular, after which it motivates the broader need for scale and confidence in web transparency infrastructures.

2.1 The Targeting Detection Problem

Targeting is a pervasive phenomenon on the web and involves the use of a user's personal data (*inputs*) to tailor some content (*output*), such as an ad, a recommendation, a price, search results, or news. Sunlight aims to identify the likely causes of each targeted output in the context of controlled experiments that test many inputs at once. Numerous use cases exist that could leverage such functionality. For example, researchers could use it to study targeting at larger scale than was possible before. We provide results from our own case studies of ad targeting in Gmail and on the web in §7. Following are two other example use cases that broaden the scope and help underscore Sunlight's design requirements.

[1] In this paper, the term *trustworthy* refers strictly to the level of confidence (in a statistical sense) one can have in the results of an investigation assuming the non-malicious service model in §3.3.

Example 1: Ann, a federal trade commission researcher specializing in COPPA enforcement, plans to investigate whether and how advertisers target children. She hypothesizes that advertisers leverage information amassed by web trackers to bid for users with browsing histories characteristic of children. Ann wants to run a *large-scale study* to both quantify the amount of children-oriented targeting, and find specific instances of what might be deemed as inappropriate or illegal targeting (e.g., targeting pornographic movies at teenagers or promoting unhealthy eating habits to young children). The number of websites dedicated to children is large, and there are even more neutral websites frequented by both children and adults on which targeted ads might appear. Ann fully expects that child-based targeting will be rare events, hence running her experiment at large scale is vital. For any case of inappropriate or illegal targeting, Ann plans to investigate through legal means (e.g., interview the advertiser) to determine whether the targeting was intentional or purely algorithmic. Such investigations are expensive, so Ann requires *high confidence* in an experimental finding to justify her investigative effort.

Example 2: Bob, a tech-savvy investigative journalist, wishes to investigate how coupons are targeted at users. Coupon services aim to provide product discounts to users who are likely to be interested in a particular product but may need some incentive to do so. A wide variety of types of information could feed into the targeting decision, including web history, tweets, and Facebook likes. Bob would like to try many different activities and see which ones are targeted by coupons. In addition to requiring high confidence in the results, Bob needs the results to also be easily *interpretable* so that he and his readers can understand and reason about the implications of the targeting. Ideally, whatever statements he makes in his articles should be directly validated on the datasets and have an associated confidence level that he can understand and potentially communicate to his audience. For example, he imagines statements such as the following to be appropriate for communication with his readers: *"In our experiments, profiles that tweeted about weight loss or diets were much more likely to be offered McDonald's coupons than those without such tweets. This result was very unlikely (0.01% chance) to have been observed if such tweets were not targeted by McDonald's."*

2.2 Limitations of Prior Approaches

The preceding examples illustrate the need for a *generic* system that supports targeting investigations by identifying not only the fact of targeting but also the likely cause of each targeted output at fine granularity (specific inputs). The experiments must run at *large scale* and any results must be *statistically justifiable* and *interpretable*. We know of no prior system that satisfies all these properties. Indeed, when turning to prior literature on measurements and tools for web transparency to guide our own design, we discovered significant mismatches at all levels.

We examined the methodologies used by 12 web transparency measurements and tools to study various aspects of data use, including: personalization in search engines [15, 29], news and product recommendations [16], and online pricing [16, 20, 21, 27]; targeting in advertising on the web [2, 8, 18, 19] and in mobile apps [6, 22]. We make several observations.

• *Generic, reusable methodologies are scarce:* Until 2014 the approach was to investigate specific questions about web targeting and personalization by developing purpose-specific, small-scale experimental methodologies. This resulted in much redundancy between investigations, and typically in small-scale, one-off experiments. In 2014, our team developed XRay [18], the first generic and scalable system design that provides reusable algorithmic building blocks in support of many targeting investigations. Following XRay, AdFisher [8] introduced in 2015 a generic, statistically sound methodology for small-scale targeting investigations. (See §8 for further discussion of XRay and AdFisher.)

• *Scalability is often disregarded:* Most prior works disregard scalability as a core design goal [2, 6, 8, 15, 16, 19–22, 27, 29]. Generally speaking, the approach is to observe data flows by varying *one input at a time* in successive experiments. This independent treatment of inputs limits the forms of personalization (*e.g.,* based on location, cookie profile, or some system-specific attributes) that can be detected by the approach. Extending such approaches to scale to many inputs and hypotheses appears difficult. For example, AdFisher builds a separate classifier for each input and validates its effect with experimental data. To investigate targeting on combinations of multiple inputs, one must build a classifier and run a separate experiment for each such combination – an exponential approach that does not scale. XRay is the only prior system that incorporates scalability with many inputs into its design.

• *Confidence assessments are often missing:* Most prior works lack robust statistical justification for their results [2, 6, 15, 16, 18–22, 27, 29]. Many works use case-by-case, comparative metrics, where the variation in different conditions is compared to that of a control group (*e.g.,* observed price differences [20, 21, 27], fraction of inconsistent search results [29], Jaccard Index and edit distance [15], normalized discount cumulative gain [16]), but do not report any assessments of statistical confidence or reliability. Other works detect targeting by running basic statistical tests, typically to reject that a given input seen conditionally on a given output is distributed uniformly [2, 6, 22]. Running these tests multiple times requires a careful correction step, an aspect that is usually ignored. Finally, our own prior system, XRay [18], provides no confidence on an individual finding basis; its predictions are only shown to become asymptotically accurate overall as XRay is applied to larger and larger systems. This makes individual results hard to trust and interpret. In terms of statistical rigor, the most mature approach is AdFisher [8] which, for a given input, builds a specific classifier and validates its effect with statistical confidence.

• *Limited evaluation and design space exploration:* Most prior work lack a rigorous evaluation of the proposed tools and associated design space. In web transparency, evaluation is extremely challenging because ground truth of targeting effects is unknown. Manual assessment is sometimes used in prior work [8, 18], but it is, in our experience, extremely prone to error (see §6.6). Inability to quantify the accuracy (precision, recall) of an algorithm makes it difficult to explore the design space and understand its trade-offs.

This paper seeks to fill in the preceding gaps by presenting: (1) The first generic web transparency methodology that provides both scalability and robust statistical confidence for individual inferences. (2) An implementation of this methodology in Sunlight. Sunlight's design is inspired by XRay and AdFisher, but improves both in significant ways (see §8 for detailed comparison). (3) An approach for evaluating the design space of a transparency system like Sunlight. We next begin by describing Sunlight's methodology.

3 The Sunlight Methodology

A core contribution in Sunlight is the development of a principled methodology for web targeting investigations, which follows what we believe are important principles to follow when building infrastructures for other types of web transparency investigations.

3.1 Design Principles

• *Design for scale and generality.* The web is big; the number of services and third-parties that could be targeting the users is gigantic. The kinds of personal data they could be using as inputs of their targeting are many and diverse. The number of service outputs that

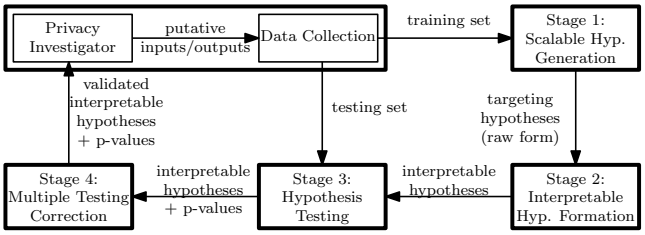


Figure 1: **The Sunlight methodology.** Consists of a four-phase pipeline: (1) scalable hypothesis generation, (2) interpretable hypothesis formation, (3) hypothesis testing, and (4) multiple test correction.

could be targeted at the users is immense. Reliability and trust in an investigation's results depend not only on the methodologies used but also on the scale at which conclusions are reached. Sunlight must thus support large-scale investigations, both in terms of the number of inputs being tracked and in terms of the number of outputs, as well as – to the extent possible – in terms of the services to which it is applied. This last goal requires us to minimize the assumptions we make about the inspected service(s).

• *Provide robust statistical justification for all inferences.* Trustworthiness in the results is key to an investigation, hence Sunlight must provide robust confidence assessments for its inferences. The metrics must be understandable and broadly accepted by the community. Where possible, Sunlight should be able to make causal claims about its inferences, and not simply correlations, which are more difficult to reason about. Enforcing high confidence for all findings may result in missing some. We believe that correct findings are preferable to complete findings. Sunlight hence attempts to limit such effects, but given a choice favors precision over recall.

• *Ensure interpretability of inferences.* A key challenge with many machine learning or statistics mechanisms is that their inferences are often not easily interpretable, and *post hoc* interpretations may not have been statistically validated. Interpretability is a critical aspect of a transparency system as people are the consumers of the system's output. Sunlight explicitly integrates a rudimentary but effective technique to ensure that its inferences are interpretable and statistically validated in these interpretable forms.

To the best of our knowledge, Sunlight is the first web transparency system to closely follow all these principles. It can currently run on hundreds of virtual machines to process data from targeting experiments, and precisely detects targeting of tens of thousands of Gmail and web ads, testing hundreds of inputs simultaneously. It minimizes the assumptions it makes about the services and provides statistically significant and interpretable results.

3.2 Methodology

Fig.1 shows the Sunlight methodology. It consists of a pipeline with four data analysis stages. Those stages depend on experimental data from an initial data collection step determined by the investigator. Data collection begins with the creation of several fictitious user profiles, each with randomly-assigned input attributes (called *inputs*) which are potentially visible to an ad targeting mechanism. For instance, an input may indicate the presence of a particular e-mail in the user profile's webmail inbox, or that the user profile was used to visit a particular website. Then, each user profile is used to measure several potential effects of targeting mechanisms (*outputs*), such as specific ad displays shown on browser visits to a news website or webmail service. The inputs should be specified *a priori*, and for various reasons which we discuss later, it will be desirable that the random assignment of input values be statistically independent (across different inputs and across different user profiles); the outputs may be specified generically (e.g., all possible ads displayed in ten refreshes of `cnn.com`), so that the set of outputs is only determined *post hoc*. The end result is a data set comprised of inputs and output measurements for each profile.

At its core, the Sunlight methodology analyzes the collected data set using a sample-splitting approach (sometimes called the "hold-out method" in machine learning) to generate and evaluate targeting hypotheses. The profiles in the data set are randomly split into a training set and a testing set. In **Stage 1 (Scalable Hypothesis Generation)**, we apply scalable classification and regression methods to the training set to generate prediction functions that can explain the output measurements for a user profile (e.g., indicator of whether a particular ad was displayed to the user) using the profile's input attributes. We focus on scalable methods that generate *simple functions* of a *small number of inputs* so that they are readily interpretable as targeting hypotheses, and take explicit measures in **Stage 2 (Interpretable Hypothesis Formation)** to ensure this if necessary. In addition, we discard any prediction functions that fail some simple sanity checks so as to reduce the number of targeting hypotheses; this again is performed just using the training set. At the end of Stage 2, we have a filtered collection of interpretable targeting hypotheses generated using only the training set.

In **Stage 3 (Hypothesis Testing)**, each such hypothesis is then evaluated on the testing set using a statistical test to generate a measure of confidence in the targeting hypothesis—specifically, a p-value. The p-value computations may make use of the known probability distributions used to assign input values in the test set profiles, and each targeting hypothesis' p-value should be valid under minimal assumptions. Because such statistical tests may be conducted for many targeting hypotheses (e.g., possibly several targeted ads), we finally apply a correction to these confidence scores in **Stage 4 (Multiple Testing Correction)** so that they are *simultaneously* valid. We may then filter the targeting hypotheses to just those with sufficiently high confidence scores, so that the end result is a statistically-validated set of interpretable targeting hypotheses.

3.3 Threat Model and Assumptions

Like all prior transparency systems of which we are aware, we assume that Web services, advertisers, trackers, and any other parties involved in the web data ecosystem, do *not* attempt to frustrate Sunlight's targeting detection. In the future, we believe that robustness against malicious adversaries should become a core design principle, but this paper does not provide such progress. Moreover, we assume that the users leveraging Sunlight are tech-savvy and capable of developing the measurement data collection necessary to collect the data. Sunlight enables targeting detection given the experimental datasets obtained through independent means.

While Sunlight can establish correlation and even causation in some circumstances between particular inputs and targeted outputs (within some confidence level), it *cannot* attribute targeting decisions to particular parties (e.g., advertisers, ad networks, trackers, etc.), nor can it distinguish between *intentional targeting* (e.g., advertisers choosing to target users in a particular category) versus algorithmic decisions (e.g. an unsupervised algorithm decides to target a particular population of users based on patterns of prior ad clicks). Moreover, because Sunlight's correlations and causations are obtained from controlled experiments with synthetic user profiles, its findings are not guaranteed to be representative of the targeting on the real population. Finally, while Sunlight can detect certain combined-input targeting, it cannot detect all forms of targeting, but rather only targeting on disjunctive (OR) combinations of a limited number of controlled inputs.

Given all of these constraints, Sunlight is best used in contexts where its results inform and provide believable justification for subsequent investigations through independent means aimed at establishing the "truth." Our scenarios in §2.1 fall into this category.

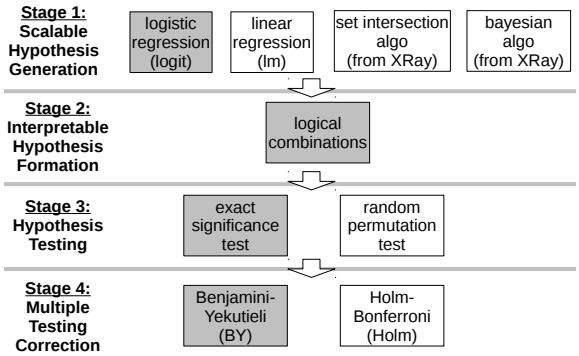


Figure 2: **The Sunlight modular pipeline.** Grey boxes show the default pipeline, which in our experience strikes a good balance between results confidence and interpretability versus support for large-scale investigations.

4 The Sunlight System

Sunlight instantiates the preceding methodology to detect, validate, and report the likely causes of targeting phenomena on the web. This raises three significant challenges. First, at each stage, unique aspects of our domain require careful modeling of the problem to map them onto appropriate statistical mechanisms. Second, across stages, mechanisms may interact in subtle ways and require careful and challenging co-designs. For example, as §6 shows, a design choice to use a permissive classification at Stage 1 (high recall but low precision, as proposed in XRay [18]) results in significant penalty due to correction in Stage 4 and failure to validate many true hypotheses (i.e., poor recall at the end of the Sunlight pipeline). In contrast, a stricter Stage 1 method that we developed for Sunlight (§4.1), which has comparatively lower recall but higher precision in it of itself results in better recall at the end of the Sunlight pipeline. Thus, a second key contribution in this paper is to identify the key requirements that must be met by the mechanisms we use at each stage of the pipeline and combine them in ways that provide scalability, confidence, and interpretability.

To address these challenges, we have designed Sunlight to be modular. It allows both the instantiation of multiple pipelines and the evaluation and comparison at different levels of the pipeline. This provides two benefits. First, it lets us explore the design space and choose the best combination of mechanisms for our problem. Second, it lets our users – researchers and investigators – adapt Sunlight to their own needs. For example, some mechanisms provide confidence at scale while others provide superior statistical guarantees; with Sunlight users can make the choices they prefer. Fig.2 lists the mechanisms we currently support at each stage, some of them which we imported from prior literature, others we developed to address limitations of prior mechanisms. We next describe the mechanisms at each stage.

4.1 Stage 1: Scalable Hypothesis Generation

We generate *interpretable* targeting hypotheses by applying classification and regression methods to the training set; the hypotheses are formulated as interpretable functions of a profiles' input attributes that can explain the corresponding output measurements. In machine learning parlance, we train predictors of the outputs based on the inputs (as input variables). To ensure that the hypotheses are interpretable, we explicitly seek predictors that only depend on a few inputs, and that only have a simple functional form. To this end, we restrict attention to hypotheses about the output that can be represented as a disjunction formula over at most k inputs for some small integer number k (here we also assume for simplicity that inputs and outputs are binary-valued). This class of small disjunction formulae is one of the simplest classes of natural and interpretable hypotheses; our focus on this class here serves as a starting point for building up techniques for other hypothesis classes.

Even with the restriction to the simple class of disjunction formulae, we face formidable computational challenges. Finding the most accurate such disjunction on an arbitrary training set is generally computationally intractable in very strong senses [11], with brute force enumeration requiring $\Omega(d^k)$ time for d inputs. Therefore, we can only hope to find accurate small disjunction hypotheses that have additional special structure. In some cases, we can use a two-step greedy approach to find such disjunction hypotheses: (i) we first use scalable classification and regression methods to order the inputs by some measure of relevance—e.g., by their average correlation with the output across user profiles in the training set (this stage); then (ii) we use this ordering over inputs to greedily construct a disjunction with sufficiently high accuracy on the training set (Stage 2). Under other conditions, it may be possible to directly form small disjunction hypotheses. We discuss some of these approaches in more detail below.

Sparse regression. Our first technique for ordering inputs is based on linear regression. In our application, each of the d inputs is regarded as a (boolean) predictive variable, and our goal is to find a linear combination of these d variables $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ that predicts an associated output measurement. The coefficients $\mathbf{w} = (w_1, w_2, \ldots, w_d)$ used to form the linear combination $\langle \mathbf{w}, \mathbf{x} \rangle = \sum_{i=1}^{d} w_i x_i$ are called the *regression coefficients*, and these can be regarded as a measure of association between the inputs and the output. These coefficients are estimated from a collection of d-dimensional data vectors, which in our setting are the vectors of input attributes for each profile in the training set.

We use a sparse linear regression method called Lasso [26] to estimate the regression coefficients $\mathbf{w}$. Lasso is specifically designed to handle the setting where the number of inputs may exceed the number n of data vectors (i.e., user profiles) in the training set, as long as the number of non-zero regression coefficients is expected to be small—i.e., the coefficient vector is sparse. This sparsity assumption entails that only a few inputs are, in combination, correlated with the output. Under certain conditions on the n data vectors (which we ensure are likely to be satisfied *by construction* of our user profiles), Lasso accurately estimates the coefficients as long as $n \geq O(k \log d)$, where k is the number of non-zero coefficients [4]—i.e., the number of input variables potentially correlated with the output. In fact, this collection of $O(k \log d)$ input vectors supports the *simultaneous* estimation of multiple coefficient vectors for different outputs (e.g., different ads), a consequence of the same phenomenon that underlies compressed sensing [9].

Linear regression permits the use of additional variables for uncontrolled factors (e.g., time-of-day, IP address of machine used to collect data) to help guard against erroneous input/output associations that could otherwise be explained by these factors. For instance, some ads may be shown more during work hours to target office workers, but the time-of-day in which data is collected for certain profiles could inadvertently be correlated with some inputs. Including time-of-day as a variable in the regression model helps suppress these unwanted associations.

We also consider the use of a generalized linear model called *logistic regression*, which is especially suited for binary outputs. This model posits that $\Pr[\text{output} = 1] = g(\langle \mathbf{x}, \mathbf{w} \rangle)$, where $g(z) = 1/(1 + e^{-z})$. To estimate regression coefficients in this model, we use a variant of Lasso called L_1-regularized logistic regression [23], whose effectiveness has been established in several empirical studies across multiple domains (e.g., [5, 28]). As in Lasso, we are able to regard the inputs with large estimated coefficients as likely to be relevant in predicting the output (and this would not be

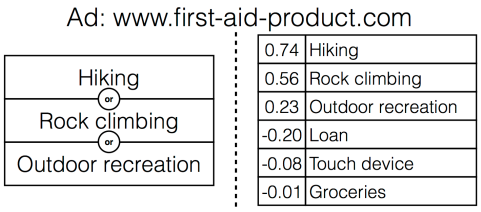


Figure 3: **Interpretable vs. raw hypothesis.** (Left) Targeting hypothesis formulated as disjunction of inputs. (Right) Raw hypothesis based on logistic regression parameters.

the case if we used unregularized or L_2-regularized logistic regression, at least in this $n \ll d$ regime).

Methods from XRay [18]. For direct comparison to prior art, we also implement two algorithms used in XRay [18] to identify inputs that may be triggering targeting outputs. While the algorithms are shown to be asymptotically exact, the system provides no statistical confidence or guarantee for individual inferences. The first algorithm is **Set Intersection**, which orders inputs by the fraction of profiles where the output is present they are covering. In other words, the best candidates for targeted inputs are the one present in the largest fraction of the profiles with the output. The second algorithm from XRay is a **Bayesian algorithm**, which uses a particular generative model for the targeting to compute posterior probabilities that each input is targeted by an output. The algorithm orders the inputs by these probabilities.

4.2 Stage 2: Interpretable Hypothesis Formation

Given an ordering over the inputs, we form disjunctions of inputs in a greedy fashion. Specifically, we first consider a singleton disjunction with the first input on the list, then a disjunction of the first two inputs on the list, and so on. We proceed as long as the training accuracy of the disjunction (in predicting the output for profiles in the training set) is sufficiently high; the criterion we use to determine this threshold is just a heuristic, but is similar to the hypothesis test used in the testing stage (§4.3). The largest sufficiently accurate disjunction is then taken (together with the associated output) as a targeting hypothesis. For some outputs, it is possible that even the singleton disjunction lacks high-enough accuracy; in such cases no hypothesis is formed.

Fig.3 shows an example of this transformation for a hypothesis based on logistic regression. The leftside hypothesis says a profile is targeted if it has at least one of the shown inputs; the rightside hypothesis says a profile is targeted if the sum of coefficients for the inputs in the profile is greater than zero. The latter appears more difficult to interpret than the former.

An alternative is to forgo interpretability and seek out out any kind of potential association between inputs and outputs. For instance, we could look for associations between arbitrary functions of inputs and outputs by using richer classes of prediction functions beyond simple disjunctions (*e.g.,* arbitrary linear threshold functions), as well as by using other flexible measures of association (*e.g.,* [14]). Such associations may be much easier to detect and statistically validate, but they may not be readily interpretable nor easy to reason about in a follow-up studies.

An important and subtle note is that if interpretability is important, then any transformation needed for interpretation should be applied at this stage. For example, an intuitive but *incorrect* way of interpreting a result from Sunlight would be to generate raw hypotheses, validate them in Stages 3 and 4, and then "interpret" those with low enough p-values. That would result in potentially misleading conclusions. For example, just because a hypothesis based on a logistic model can be validated with low p-values, it does not follow that the corresponding disjunctive version of that hypothesis is also statistically significant. For this reason, the Sunlight methodology critically includes this explicitly interpretability stage, which reminds a developer to transform her hypothesis early for interpretability so the p-values can be computed for *that* hypothesis.

4.3 Stage 3: Hypothesis Testing

The second stage of the analysis considers the targeting hypotheses (disjunctions of inputs and an associated output) generated from the first stage and provides a confidence score for each hypothesis. The score, a p-value, comes from an exact statistical test that decides between a null hypothesis H_0 that the disjunction of inputs is independent of the associated output, and an alternative hypothesis H_1 that the disjunction of inputs is positively correlated with the output. A small p-value—≤ 0.05 by convention—implies that our observations on the data (discussed below) are unlikely to be seen under H_0; it lends confidence in rejecting H_0 and accepting H_1 and the validity of the targeting hypothesis.

Computing p-values. The test is based on computing a test statistic on the testing set (i.e., the subset of profiles that were not used to generate targeting hypotheses). A critical assumption here is that the profiles (and specifically, the outputs associated with each profile) are statistically independent, and hence the selected disjunction is also independent of the profiles in the testing set. The specific test statistic T we use is an association measure based on Pearson's correlation: we compute T using the profiles from the testing set, and then determine the probability mass of the interval $\{t \in \mathbb{R} : t \geq T\}$ under H_0. This probability is precisely the p-value we seek. Because the distribution of the inputs for each profile is known (and, in fact, controlled by us), it is straightforward to determine the exact distribution of T under H_0. For example, when the inputs for each profile are determined with independent but identically distributed coin tosses, the p-value computation boils down to a simple binomial tail calculation.

More specifically, suppose the inputs are independent and identically distributed binary random variables with mean $\alpha \in (0, 1)$. Consider a disjunction of k inputs and a particular (binary-valued) output. Let N be the number of profiles for which the output is 1, and let B be the number of profiles for which both the disjunction and the output are 1. If $N = 0$, then the p-value is 1. Otherwise, the p-value is $\sum_{i=B}^{N} \binom{N}{i} \alpha_k^i (1-\alpha_k)^{N-i}$ where $\alpha_k = 1-(1-\alpha)^k$.

Use of p-value in Stage 2. As previously mentioned, we also use this p-value computation in Stage 2 *as a rough heuristic* for deciding which disjunctions to keep and pass on to Stage 3. However, we stress that these Stage 2 p-value computations are not valid because the disjunctions are formed using the profiles from the training set, and hence are not independent of these same training set profiles. The validity of the Stage 3 p-values, which are based on the testing set, relies on the independence of the disjunction formulae and the testing set profiles themselves.

Independence assumption. It is possible to weaken the independence assumption by using different non-parametric statistical tests, as is done in [8]. Such tests are often highly computationally intensive and have lower statistical power to detect associations. We opt to admit the assumption of independence in favor of obtaining more interpretable results under the assumption; gross violations may be identified in follow-up studies by an investigator, which we anyway recommend in all cases.

Causal effects. Under the alternative hypothesis of a positive correlation between a disjunction of inputs and an output, it is possible to draw a conclusion about the *causal effect* of the inputs on the output. Specifically, if the input values are independently assigned for each user profile, then a positive correlation between a given disjunction of inputs and an output translates to a positive *average*

causal effect [24]. This independent assignment of input values can be ensured in the creation of the user profiles.

4.4 Stage 4: Multiple Testing Correction

In the final stage of our analysis methodology, we appropriately adjust the p-values for each of our targeting hypotheses to correct for the *multiple testing problem*. As one simultaneously considers more and more statistical tests, it becomes more and more likely that the p-value for some test will be small just by chance alone even when the null hypothesis H_0 is true. If one simply rejects H_0 whenever the stated p-value is below 0.05 (say), then this effect often leads to erroneous rejections of H_0 (*false rejections*).

This multiple testing problem is well-known and ubiquitous in high-throughput sciences (e.g., genomics [10]), and several statistical methods have been developed to address it. A very conservative correction is the Holm-Bonferroni method [17], which adjusts the p-values (generally making them larger them by some amount) in a way so that the probability of *any* false rejection of H_0 (based on comparing adjusted p-values to 0.05) is indeed bounded above by 0.05. While this strict criterion offers a very strong guarantee on the resulting set of discoveries, it is often overly conservative and has low statistical power to make any discoveries at all. A less conservative correction is the Benjamini-Yekutieli procedure [3], which guarantees that among the adjusted p-values that are less than 0.05, the expected fraction that correspond to false discoveries (i.e., false rejections of H_0) is at most 0.05. Although this guarantee on the expected false discovery rate is weaker than what is provided by the Holm-Bonferroni method, it is widely accepted in applied statistics as an appropriate and preferred correction for exploratory studies.

With either correction method, the adjusted p-values provide a more accurate and calibrated measure of confidence relative to the nominal 0.05 cut-off. We can either return the set of targeting hypotheses whose p-values fall below the cut-off, or simply return the list of targeting hypotheses ordered by the p-values. Either way, the overall analysis produced by this methodology is highly-interpretable and statistically justified.

4.5 Prototype

We implemented Sunlight in Ruby using statistical routines (e.g., Lasso) from R, a programming environment for statistical computing.The analysis is built around a modular pipeline that lists the algorithms to use for each stage, and each algorithm implements a basic protocol to communicate with the next stage.

Default Pipeline. Fig.2 shows the default pipeline used by Sunlight. In Stage 1, we use sparse logistic regression (Logit) to estimate regression coefficients that give an ordering over the inputs. In Stage 2, we select a disjunction (i.e., an "OR" combination) with the best predictive accuracy from ordered inputs from Stage 1, and discard inaccurate hypotheses as determined using heuristic p-value computations. Stage 3 computes the p-values for the statistical test of independence on the test data. Finally, our Stage 4 implementation computes both the Benjamini-Yekutieli (BY) and Holm-Bonferroni (Holm) corrections, though our default recommendation is the BY correction. Finally, we recommend p-values < 0.05 for good confidence.

In §6, we show that these defaults strike a good balance between the scalability and the confidence of these hypotheses. Using these defaults, our targeting experiments on Gmail and on the web produced the largest number of high confidence hypotheses, and we have manually inspected many of these hypotheses to validate the results. We describe these measurements next.

5 Sunlight Use Cases

To showcase Sunlight, we explored targeting in two ad ecosystems with two experiments, on Gmail ads and ads on the web respectively. We used the datasets generated from these experiments for two purposes: (1) to evaluate Sunlight and compare its performance to prior art's (§6) and (2) to study a number of interesting aspects about targeting in these ecosystems (§7). As foreshadowing for our results, both experiments revealed contradictions of separate statements from Google policies or official FAQs. While our use cases refer exclusively to ad targeting detection, we stress that our method is general and (intuitively) should be applicable to other forms of targeting and personalization (e.g., §6.2 shows its effectiveness on Amazon's and YouTube's recommendation systems).

	email subject & text	ads Title, url & text	Results
Race	**Dominican** dominican [...] **(OR)** **Hair** hair cut hair cut	**Shampoo JC** www.shampoojc.com Professional Coloring, Highlights. Make your appointment now!	p-value = 0.0004 1427 impressions in 44 profiles 93% in context
Religion & Spirituality	**Mormon** mormon mormon	**Family History Search** genealogy.com/Family+History 1) Simply enter their name. 2) View their family history now!	p-value = 0.001 237 impressions in 18 profiles 74% in context
	Muslim muslim muslim	**Fine Models & Miniatures Shop** www.1stdibs.com/Modern Our Unique Modern Collection Rare Items Added Every Week.	p-value = 0.007 59 impressions in 11 profiles 100% in context
	Jewish jewish jewish	**Free Ancestor Search** archives.com Looking for Your Family Ancestry? Search for Free [...]	p-value = 0.0007 287 impressions in 15 profiles 100% in context
	Buddhist buddhist buddhist	**Berkshire Retreat** www.eastover.com Holistic Retreat Center Personal Retreat	p-value = 0.008 190 impressions in 17 profiles 43% in context
	Guru guru spiritual guide **(OR)** **Astrology** astrology psychic mystical	**What is Quantum Jumping?** www.quantumjumping.com Discover How Thousands of People are Jumping to Change Their Life [...]	p-value = 0.001 244 impressions In 34 profiles 76% in context
Sexual Orient.	**Gay** gay homosexual lesbian gay [...]	**Men Underwear/Workout** www.gosoftwear.com Underwear, Swimwear Go Natural American [...]	p-value = 0.05 54 impressions in 19 profiles 74% in context
General Health	**Affordable** afforable care [...] **(OR)** **Nursing** nursing home [...]	**Illinois Senior Living** www.cottagesofnewlenox.com Assisted Living for Seniors in New Lenox [...]	p-value = 0.03 103 impressions in 36 profiles 28% in context
	Alzheimer Alzheimer Alzheimer	**1/3 of Seniors 65+ Fall** jacuzzi-walk-in-tubs.com/Safety Help Eliminate the Fear of Falling in the Bathroom [...]	p-value = 0.01 21 impressions in 8 profiles 100% in context
	Depressed depression **(OR)** **Anxious** anxious anxiety	**Is He A Cheater?** spokeo.com/Cheating-Spouse-Search Enter His Email Address. Find Pics & Profiles From 70+ Social Networks.	p-value = 0.03 1179 impressions in 52 profiles 20% in context
	Cancer advice How did you cope with cancer in your family? What an aweful disease!	**The Business of Wellness** healthmediagroup.blogspot.com What my doctor can learn from my Shoe Shine Man [...]	p-value = 0.04 380 impressions in 28 profiles 91% in context
Prohibited	**counterfeit, counterfeit** counterfeit counterfeit	**A&A Global Industries** aaglobal.com Largest Supplier to Bulk Industry Toys, Equipment, Candy, Supplies	p-value = 0.002 66 impressions in 17 profiles 100% in context
	drugs drugs cheap online order	**Eagle Creek Luggage** www.eaglecreek.com/ Extremely Tough & Durable Gear. Luggage, Organizers, Duffels & More	p-value = 0.03 214 impressions in 19 profiles 99% in context
Misc.	**Deregulation** deregulation [...] **(OR)** **Financial Reform** financial reform [...]	**Compliance Audit** unifiedcompliance.com Checklist All IT Compliance You Need to Track In [...]	p-value = 0.0008 1582 impressions in 36 accounts 61% in context
	Unemployed lazy unemployed	**Easy Auto Financing** www.midsouthautoloans.com Need a quick car loan? We work with credit issues	p-value = 0.006 161 impressions in 24 profiles 8% in context
	Payday payday loan	**Fast Cash Loan Online.** www.checkintocash.com Apply Now. Takes Only 5 Minutes. It's as Easy as 1,2,3.	p-value = 0.007 198 impressions in 10 profiles 6% in context
	Veterans war veteran veterans	**Veterans Care Costa Rica** www.veteranscarecostarica.com Receive your proper medical care Tricare, VA, Champ VA	p-value = 0.0006 490 impressions in 15 profiles 84% in context

Figure 4: **Sample targeted ads from the 33-day Gmail experiment.**

5.1 Gmail Ads

As a first example of personal data use, we turn to Gmail which, until November last year, offered personalized advertisements tailored to a user's email content. We selectively placed more than 300 emails containing single keywords or short phrases to encode a variety of topics, including commercial products (e.g. TV, cars, clothes) and sensitive topics (e.g., religion, sexual orientation, health) into 119 profiles. The emails were manually written by us by selecting topics and writing keywords related to this topic. The first column of Figure 4 shows examples of emails we used. The topics were selected from the AdSense categories [12], with other sensitive forbidden by the AdWords policies [13].

The profiles were Gmail accounts created specifically to study Gmail targeting. Because creating Gmail accounts is costly, some accounts were reused from previous studies, and already contained some emails. The emails relevant to this study were different, and assigned independently from previous emails, so our statistical guaranties still hold. To perform the independent assignment each email was sent to each account with a given probability (in this case 0.2). Emails were sent from 30 other Gmail accounts that did not otherwise take part in the study. No account from the study sent an email to another account of the study. Finally we collected targeted ads by calling Google's advertising endpoints the same way Gmail does, looping over each email and account ten times.

Our goal was to study (1) various aspects related to targeted advertisements, such as how frequent they are and how often they appear in the context of the email being targeted (a more obvious form of targeting) versus in the context of another email (a more obscure form of targeting) and (2) whether advertisers are able to target their ads to sensitive situations or special groups defined by race, religion etc. We collected targeting data for 33 days, from Oct. 8 to Nov. 10, 2014 when Google abruptly shut down Gmail ads. One might say that we have the last month of Gmail ads. [2]

Before Google disabled Gmail ads, we collected 24,961,698 impressions created collectively by 19,543 unique ads. As expected, the distribution of impressions per ad is skewed: the median ads were observed 22 times, while the top 25/5/1% of ads were observed 217/4,417/20,516 times. We classify an ad as *targeted* if its statistical confidence is high (corrected p-value$<$ 0.05 with Sunlight's default pipeline). In our experiment, 2890 unique ads (15% of all) were classified as targeted. While we observe that ads classified as targeted are seen more often (1159 impressions for the median targeted ads), this could be an artifact as most ads seen only occasionally present insufficient evidence to form hypotheses.

Figure 4 shows some examples of ads Sunlight identified as targeted, along with the content of the emails they targeted, the corrected p-values, and information about the context where the impressions appeared. Some ads show targeting on single inputs while others show targeting on combinations of emails. We selected these examples by looking at all ads that were detected as targeting the sensitive emails we constructed, and choosing representative ones. When multiple interesting examples were available, we chose those with a lot of data, or that we detected across multiple days, as we are more confident in them.

Notably, the examples show that information about a user's health, race, religious affiliation or religious interest, sexual orientation, or difficult financial situation, all generate targeted advertisements. Our system cannot assign intention of either advertisers or Google for the targeting we found, but this appears to contradict a statement in an official-looking Gmail Help page: [3]

[2]Gmail now has email "promotions;" we did not study those.

[3]The page containing this statement used to be accessible through a user's own account (Gmail - Help - Security & privacy - Privacy policies) and its look and feel until 12/24/2014 was more official than it currently is. The 2014 version is available on archive.org (`https://web.archive.org/web/20141224113252/https://support.google.com/mail/answer/6603`).

	Targeted website	ads Title & text	Results
Drugs	drugs.com	Nasalcrom Proven to Prevent Nasal Alergy Symptoms	p-value = 2.5e-5 374 impressions in 73 profiles 41% in context
	hightimes.com	AquaLab Technologies Bongs, Pipes, and Smoke Accessories	p-value = 2.6e-13 1714 impressions in 76 profiles 99% in context
News	foxnews.com	IsraelBonds.com Invest in Israel	p-value = 0.0041 71 impression in 45 accounts 100% in context
	huffingtonpost.com	Stop The Tea Party Support Patrick Murphy	p-value = 0.010 97 impressions in 37 profiles 100% in context
	economist.com	The Economist Great Minds Like a Think - Introductory Offer	p-value = 0.00066 151 impressions in 77 profiles 0% in context
Misc.	pcgamer.com (games)	Advanced PCs Digital Storm Starting at $699	p-value = 0.035 575 impressions in 129 profiles 66% in context
	soberrecovery.com (rehab)	Elite Rehab Speak w/ a counselor now	p-value = 6.8e-6 5486 impressions 82 profiles 99% in context

Figure 5: **Sample targeted ads from the display-ads experiment (also called Website experiment).**

> Only ads classified as Family-Safe are displayed in Gmail. We are careful about the types of content we serve ads against. For example, Google may block certain ads from running next to an email about catastrophic news. We will also not target ads based on sensitive information, such as *race*, *religion*, *sexual orientation*, *health*, or *sensitive financial categories*.
>
> – `support.google.com/mail/answer/6603`.

While our results do *not* imply that this targeting was intentional or explicitly chosen by any party involved (Google, advertisers, etc.), we believe they demonstrate the need for investigations like the ones Sunlight supports. We also point out that those violations are needles in a haystack. Several topics we included in our experiment (e.g., fatal diseases and loss) generated not a single ad classified as targeted.

§7 presents further results about targeting on Gmail.

5.2 Display Ads on the Web

As a second example of personal data use, we look at targeting of arbitrary ads on the web on users' browsing histories. This experiment is not specifically related to Google, though Google is one of the major ad networks that serve the ads we collect. Similar to the Gmail use case, our goal is to study aspects such as frequency of targeted ad impressions, how often they appear in the context of the website being targeted versus outside, and whether evidence of targeting on sensitive websites (e.g., health, support groups, etc.) exists. We populate 200 browsing profiles with 200 input sites chosen randomly from the top 40 sites across 16 different Alexa categories, such as News, Home, Science, Health, and Children/Teens. Each website is randomly assigned to each profile with a probability 0.5. For each site, we visit the top 10 pages returned from a site-specific search on Google. We use Selenium [25] for browsing automation. We collect ads from the visited pages using a modified version of AdBlockPlus [1] that detects ads instead of blocking them. After collecting data, we use Sunlight's default pipeline and a p-value < 0.05 to assess targeting.

We collect 19,807 distinct ads through 932,612 total impressions. The web display ads we collected skewed to fewer impressions than those we collected in the Gmail experiment. The median ad appears 3 times and we recorded 12/126/584 impressions for the top 25/5/1% of display ads. In this experiment, 931 unique ads (5% of all) were classified as targeted, and collectively they are responsible from 37% of all impressions.

Figure 5 shows a selection of ads from the study, chosen similarly as the ads from the Gmail study. Among the examples are ads targeted on marijuana sites and drug use sites. Many of the ads targeted on drug use sites we saw, such as the "Aqua Lab Technologies" ad, advertise drug paraphernalia and are served from `googlesyndication.com`. This appears to contradict Google's advertising policy, which bans "Products or services marketed as facilitating recreational drug use." – `https://support.google.com/adwordspolicy/answer/6014299`.

§7 presents further results about targeting on the web.

6 Evaluation

We evaluate Sunlight by answering the following questions: (Q1) How accurate is Sunlight's against ground truth, where it is available? (Q2) How do different Stage 1 algorithm's hypotheses compare? (Q3) What is the influence of p-value correction on Sunlight? (Q4) How does scale affect confidence? As foreshadowing, we show that Sunlight's high-confidence hypotheses are precise, and that the Logit (logistic regression) method is best suited among those we evaluated for maximizing hypothesis recall after p-value correction. Somewhat surprisingly, we show that the "winning" inference algorithm at Stage 1 (XRay's) is *not* the winner at the end of the pipeline, after correction is applied. Finally, we show that the same effect in also responsible for a trade-off between confidence and scalability in the number of outputs.

6.1 Methodology

We evaluate Sunlight using the system's split of observations into a training and a testing set, and leveraging the modularity of Sunlight to measure the effectiveness of targeting detection at different stages of the analysis pipeline. We believe that our evaluation methodology, along with the metrics that we developed for it, represents a significant contribution and a useful starting point for the evaluation of future transparency infrastructures, an area that currently lacks rigorous evaluations (see §2.2).

A critical challenge in evaluating Sunlight and its design space is the lack of ground truth for targeting for most experiments. For example, in Gmail, we do not know how ads are targeted; we can take guesses, but that is extremely error prone (see §6.6). In other cases (such as for Amazon and Youtube recommendations), we can obtain the ground truth from the services. For a thorough evaluation, we thus decided to use a multitude of metrics, each designed for a different situation and goal. They are:

1. *hypothesis precision*: proportion of high-confidence hypotheses that are true given some ground truth assessment.
2. *hypothesis recall*: proportion of true hypotheses that are found from some ground truth assessment.
3. *ad prediction precision*, the proportion of success in predicting if an ad will be present in a training set account.[4]
4. *ad prediction recall*: proportion of ads appearances that were correctly guessed when predicting if an ad will be present in a training set account.
5. *algorithm coverage*: proportion of low p-value hypotheses found by an algorithm, out of all low p-value hypotheses found by any of the algorithms.

[4]"Ad" in this section is short for the more generic output.

Workload	Profiles	Inputs	Outputs
Gmail (one day)	119	327	4099
Website	200	84	4867
Website-large	798	263	19808
YouTube	45	64	308
Amazon	51	61	2593

Table 1: **Workloads used to evaluate Sunlight**

We use the first two metrics in cases where ground truth is available (§6.2) and with manual assessments (§6.6). These are typically small scale experiments. We use the next two metrics in cases where ground truth is unavailable; this lets us evaluate at full scale and on interesting targeting. Finally, we use the last metric for comparison of various pipeline instantiations.

Table 1 shows the datasets on which we apply these metrics. The first three datasets come from the experiments described in the preceding section. The Gmail dataset corresponds to one day's worth of ads in the middle of our 33-day experiment. The YouTube and Amazon datasets are from our prior work XRay [18]. They contain targeting observations for the recommendation systems of YouTube and Amazon, for videos and products respectively. They are small (about 60 inputs), and with inputs on very distinct topics, minimizing the chances for targeting on input combinations. On the other hand the Gmail and Websites datasets are larger scale, with up to 327 inputs and thousands outputs. Moreover their inputs are not distinct, containing some redundancy because they include emails or websites on the same topics that are more likely to attract similar outputs. They are thus more representative of experiments that would be conducted by investigators.

In all the evaluation, we use XRay as our baseline comparison with prior art. XRay is Sunlight's most closely related system, inheriting from it many of its design goals, including its focus on scalable, generic, and fine-grained targeting detection. We leave quantitative comparison with other systems for future work and refer the reader to our analytical comparison in §8.

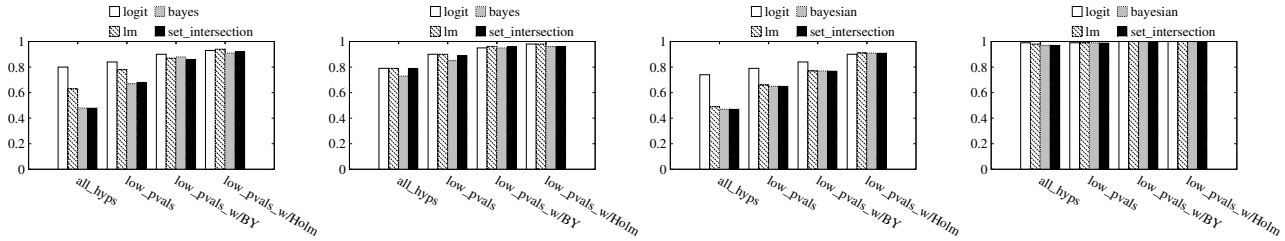
6.2 Q1: Precision and recall on ground truth

Dataset	Precision		Recall		Hyp. count
	Sunlight	XRay	Sunlight	XRay	
Amazon	100%	81%	46%	78%	142
YouTube	100%	93%	52%	68%	1349

Table 2: **Sunlight's hypothesis precision & recall**

Sunlight favors finding reliable, validated targeting hypotheses over finding every potential targeting, so that investigators do not waste time on dead ends. This strategy is characterized by *hypothesis precision* that should be very high, and *hypothesis recall* that we try to keep high without lowering precision. We measure these two metrics on two datasets from YouTube and Amazon from the XRay paper [18], both containing ground truth (Amazon and YouTube inform users why they are shown certain recommendations). This gives us a direct comparison with prior art, as well as an assessment of Sunlight's hypothesis precision and recall on services provided ground truth for recommendation targeting. Table 2 describes the results. We make two observations. First Sunlight's hypothesis precision against ground truth is 100% (with a Logit Stage 1) on both Amazon and YouTube, while XRay's best algorithm reaches only 81% and 93% respectively. This confirms Sunlight's high hypothesis precision that makes a difference even on simple cases.

Second hypothesis recall is higher for XRay. The Bayesian algorithm reaches 68% on YouTube and 78% on Amazon while Logit yields 46% and 52% respectively. This can be explained by the small size of these datasets: when faced with little evidence, Sunlight will return no hypothesis or low confidence hypotheses, favoring precision over recall compared to XRay's algorithms. We believe this is a valuable trade-off when performing large scale ex-



(a) **Ad prediction precision, Gmail** (b) **Ad prediction recall, Gmail** (c) **Ad pred. precision, Websites** (d) **Ad pred. recall, Websites**

Figure 6: **Ad prediction precision and recall.** The x-axis shows different p-value correction methods, and the y-axis shows the proportion of precision and recall. (a) and (b) show ad prediction precision and recall on the Gmail dataset, (c) and (d) on the Websites dataset, for all algorithms. Both metrics increase when using stricter p-values, indicating better hypotheses.

periments. In the absence of ground truth, we need to be able to trust targeting hypotheses even at the cost of some recall.

This confirms Sunlight's focus on precision over recall on datasets with ground truth. We next study more complex targeting with inputs on redundant topics, but that do not provide ground truth.

6.3 Q2: Evaluating the analysis pipeline

We now look inside the analysis pipeline to measure the effects of its different stages and to compare stage 1 algorithms. In order to measure algorithm's performances we use their *ad prediction precision and recall* described in § 6.1. Intuitively if the algorithms detect targeting, they can predict where the ads will be seen in the testing set. Because ads do not always appear in *all* accounts that have the targeted inputs, we do not expect precision to always be 100%. On the other hand, a targeting hypothesis formed using many inputs may easily yield high recall.

Fig. 6 shows the precision and recall of those predictions on the Gmail and Website datasets, first on all hypotheses and then after selecting higher and higher confidence hypotheses. We make three observations. First, the precision is poor if we take every hypotheses into account (see group labeled all_hyps). Precision is below 80% for both datasets, and even less than 60% for most algorithms. Restricting to just the low p-value hypotheses (without correction) somewhat increases ad presence precision (low_pvals group).

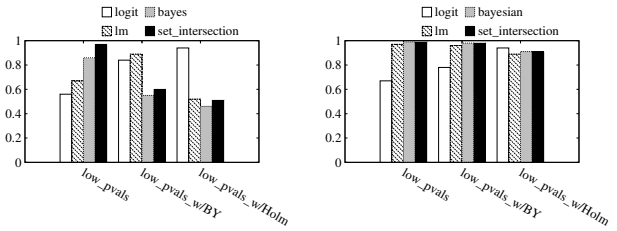
Second, correcting the p-values for multiple testing increases precision as well as recall. The best algorithms on the Gmail and Website datasets, respectively, reach a precision of 90% and 84% after BY correction, and 93% and 91% after Holm correction (low_pvals_w/BY and low_pvals_w/Holm groups). The precision is higher when with Holm because it is more conservative than BY.

Third, the differences introduced by Stage 1 algorithms are reduced by filtering out low-confidence hypotheses. While the precision with all hypotheses (all_hyps group) can vary of up to 40 percentage points, different Stage 1 algorithms vary only by 1 or 2 percentage points after Holm correction (low_pvals_w/Holm group). The exception is with the BY correction (low_pvals_w/BY group), where the precision of Logit is noticeably higher than that of the other algorithms on the Website dataset.

Thus, when selecting only high-confidence hypotheses, Sunlight is able to predict the presence of an ad with high precision and recall. Moreover, all Stage 1 algorithms generally yield accurate high-confidence hypotheses, which suggests that we should maximize the number of hypotheses. We next compare the number of high-confidence hypotheses and how it is affected by correction.

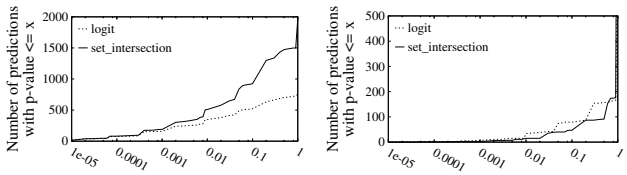
6.4 Q3: The effect of p-value correction

Maximizing the number of high-confidence hypotheses is maximizing *coverage* (see § 6.1), the proportion of all high-confidence hypotheses found by a given algorithm. Fig. 7 shows for each Stage 1 algorithm the coverage on the Gmail and Website datasets for



(a) **Coverage, Gmail** (b) **Coverage, Websites**

Figure 7: **Coverage.** Proportion of ads each algorithm found, out of all ads found by at least one algorithm.

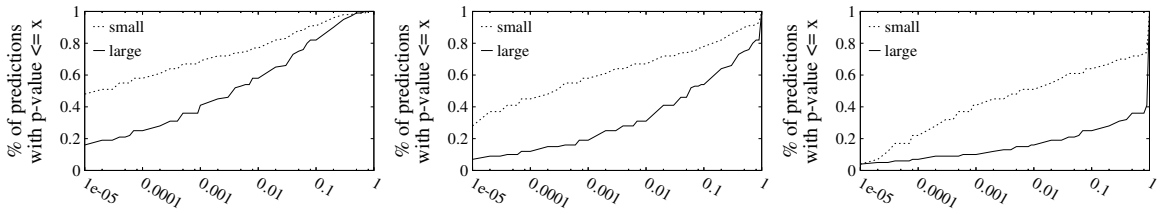


(a) **P-value CDF (Gmail)** (b) **Same as (a) with Holm**

Figure 8: **Effect of p-value correction on the distribution.** The Set Intersection algorithm makes much more hypothesis, and thus has more low p-value hypothesis. After Holm's correction however, the Logit algorithm has more low p-value hypothesis. X is log scale.

low p-values, before and after correction. Set Intersection outputs the most low p-value hypotheses (low_pvals group), but we saw in Fig. 6(a) 6(c) that these hypotheses are poor predictors of the ad presence, with a precision below 50%. After the strictest correction (low_pvals_w/Holm), when all hypotheses have similar predictive power, the Logit Stage 1 gives the best coverage, with 93% on Gmail and 94% on Website, beating Lm, the Bayesian algorithm, and Set Intersection. We can make the same conclusion on Gmail after BY correction, but the picture is not as clear on the Websites dataset, where Logit has a lower coverage (about 80%) but makes hypotheses with a better ad prediction precision (see Fig. 6(c)).

It is interesting to understand why Set Intersection has a much better coverage before correction, but loses this edge to Logit after p-value correction. This can be explained by the fact that the number of hypotheses, and the proportion of high p-value hypotheses play an important role in the correction, both increasing the penalty applied to each p-value. To further demonstrate this effect, Fig. 8 shows the CDF for the distribution of the absolute number of hypotheses per p-value for Logit and Set Intersection. On Fig. 8(a) we observe that the Set Intersection Stage 1 algorithm makes more low p-value hypotheses with 836 hypothesis below 5%, and only 486 for Logit. However we can also see that the total number of



(a) **P-value distribution** (b) **P-value dist. after BY correction** (c) **P-value dist. after Holm correction**

Figure 9: **Effect of p-value correction on the distribution, at different scales.** In this graph, the scale is regarding the number of ads. For our small and large scale Website datasets, for the Logit Stage 1, (a) shows the p-value distribution, (b) the p-value distribution after BY correction, and (c) the p-value distribution after Holm correction. The higher number of hypotheses in the large experiment widens the difference in distributions after correction.

hypothesis is much higher (3282 compared to 928), and that a lot of these hypotheses have a high p-value. After Holm correction however, the Logit algorithm retains 80% more low p-value hypotheses, as shown on Fig. 8(b). BY correction shows the same trend, although with less extreme results, explaining why on the Websites dataset Set Intersection keeps a higher coverage.

We show that making many hypotheses hurts coverage after p-value correction, particularly with the Holm method. This calls for algorithms that favor small number of high quality hypotheses, such as Logit, over algorithms that make hypotheses even on weak signal, such as Set Intersection. Making only a small number of low p-value hypotheses is not always possible however, especially when scaling experiments, which we evaluate next.

6.5 Q4: Confidence at scale

We saw in § 6.4 that favoring fewer, higher quality hypothesis is a winning strategy to retain low p-value hypothesis after correction. Unfortunately this is not always possible when scaling the number of outputs, for instance when experiments have more inputs, more profiles, or just when we collected data for a longer period. In these cases analysis may also be harder, leading to fewer good p-value hypotheses. Fig. 9 shows a CDF of the proportion of inputs below a given p-value, and the same after the two p-value correction techniques we consider. We make three observations. First, it appears that hypotheses are indeed harder to make on the large dataset. Fig. 9(a) shows that the proportion of low p-values is lower even before correction, with 75% of hypotheses with a p-value below 5% for the Website-large, and 88% for Website.

Second, Fig. 9(c) shows that the Holm correction greatly reduces the proportion of low p-values, with the Website experiment going from 88% to 61%. The effect on the many hypotheses of Website-large is much stronger however, with the low p-values dropping from 75% to 21%. We conclude that the Holm correction is very hard on experiments with a lot hypotheses. The larger the proverbial haystack, the harder the needles will be to find.

Third, the BY correction method is milder than Holm's. Even though we can see the large experiment's distribution caving more than the small's, the distinction at scale is smaller. Website-large still has 46% of its p-values below 5%, while the small one has 74%. Despite the weaker guarantees of the BY correction, it can be a useful trade-off to make when dealing with large numbers of outputs. Indeed, it is a well-accepted correction in statistics and machine learning. We hence include it as a default in our system.

6.6 Anecdotal experience with the data

We already saw that high confidence hypotheses give good predictors of the profiles in which an ad will appear. While this *ad prediction precision and recall* reveals that our algorithms are indeed detecting a correlation, we also manually looked at many hypotheses to understand the strengths and weaknesses of our methods. We next describe the results of this experience on large, complex datasets from Gmail and Websites experiments. These services do not provide ground truth at this granularity, so we manually assessed the hypotheses' validity. For this manual assessment we looked at low p-value hypotheses, visited the website targeted by the ad, and decided if it made sense for the website to target the ad. If we did not find a connection for at least one email in the targeting combination, we declared the hypothesis wrong. For instance an ad for a ski resort targeting the "Ski" email was considered right, but the same ad targeting "Ski" and "Cars" was considered a mistake. This labelling is very error prone. On the one hand some associations can be non obvious but still be a real targeting. On the other hand it can be easy to convince ourselves that an association makes sense even when there is no targeting. It is thus an anecdotal experience of Sunlight's performances.

Hypothesis Precision. We examined 100 ads with high-confidence hypotheses (p-value < 0.05 after Holm) from the Gmail and Website experiments, and counted instances where we could not explain the input/output association with high certainty. We found precisions of 95% and 96%, respectively. Hypotheses we classified as false positives were associations of ads and emails that we just could not explain from the topics, such as the luggage company "www.eaglecreek.com" targeting the "Cheap drugs online order" email from Figure 4. In this specific case the ad appears for 3 consecutive days, multiple times (a total of 437 impressions in 15 to 19 different accounts), and almost only in the context of the email, so there does seem to be some targeting, although we cannot semantically explain it. However, in other cases we have less data to gain confidence, and we classify them as a mistake. This example highlights the difficulty and the risk of bias of manual labelling.

Many errors were also combinations with one input that seemed relevant, and other inputs that did not. Intuitively this happens if the targeting detection algorithm adds any other input to a heavily targeted input, because in the specific *training set* this other input correlates with the output (this is a case of over-fitting in the training set). If for instance the ad appears in 10 profiles in the *testing set*, all with the relevant input, and the inputs are assigned to profiles with a 20% probability, the p-value should be $1.0e^{-7}$ with only the relevant input. With the second input added the p-value becomes higher since a combination is more likely to cover profiles. However the new p-value is still $3.6e^{-5}$ for two inputs, which will remain below the 5% accepted error rate after correction.

Hypothesis Recall. Assessing hypothesis recall based on manual inspection is even more challenging than assessing hypothesis pre-

cision. First, there are many more ads to analyze. Second, finding an input among the hundreds we have that is very likely to be targeted is challenging, and the many possibilities make it very easy to invent connections where there is none. For this reason we did not try to quantify hypothesis recall. Instead, we studied low p-value hypotheses that are rejected after correction, a more amenable method that gives information into how many hypotheses we lose due to correction. In our experience, this happens mostly if an ad does not appear enough: the p-value cannot be low enough to be below 5% after correction. For instance if the ad appears in 10 profiles, it will be in about 3 profiles of the testing set, and the p-value cannot be below 0.008 if the inputs are assigned with a 20% probability. After correction this will most likely be over the 5% threshold on big experiments.

This anecdotal experience qualitatively confirms Sunlight's high hypothesis precision on sizeable datasets. It also confirms that manual labelling is unreliable. This is why we conducted our rigorous evaluation with the five objective metrics described in § 6.1. More importantly this experience emphasizes the importance of focusing on quality hypotheses when analyzing a large number of outputs. Indeed, the correction will reject all reasonable hypotheses without a lot of data when the number of hypotheses is too high.

6.7 Summary

We show that Sunlight's high-confidence hypotheses have a good ad prediction precision and recall after p-value correction. This empirically confirms the need to correct for multiple hypothesis testing. The BY correction seems to reach a good trade-off between statistical guarantees and number of low p-value hypotheses.

More surprisingly, we also show an inversion of recall after correction, where algorithms that make fewer, more precise hypotheses end up with better coverage after correction. This makes the case for algorithms that favor precision even at the cost of some recall. Even with such algorithms, recall can become lower after correction when scaling the number of outputs. This represents a fundamental scale/confidence trade-off in targeting detection.

7 Other Targeting Results

Using Sunlight, we found several other interesting aspects about ad targeting in Gmail (now obsolete) and on the web. We next describe those aspects as examples of the kinds of things that could be learned with Sunlight. As before, we consider an ad *targeted* if its corrected p-value is < 0.05 under the Sunlight default pipeline. For the results in this section we use the entire display ad dataset and focus on one day from the Gmail study.

In Context vs. Outside Context. One interesting question one might wonder is how often are ads shown out of the context of the targeted input. Intuitively, if an ad is shown in the email (or on the page) that it targets, its should be more obvious to a user compared to an ad shown with one email (or page) but targeting another email (or page). Figure 10(a) shows a CDF of how often targeted ads appear in their target context for the Gmail and Website-large datasets. The Y axis represents the fraction of all targeted ads in each experiment.

In Gmail, ads are frequently out of context (i.e., alongside emails that they do not target). Approximately 28% of the Gmail ads labeled as targeted appear only in their targeted context and half of targeted Gmail ads appear out of context 48% of the time or more. Thus there is (or rather, was) heavy behavioral targeting in Gmail.

On the web, display ads are rarely shown outside of their targeted context. 73% of ads are only ever shown on the site targeted by the ad. Of the targeted ads that do appear out of context, the majority of them appear on only 1 or 2 other sites. This suggests a very heavy contextual targeting for display ads. That said, we have found convincing examples of behaviorally targeted ads that appear entirely outside of their targeted context. Included in Figure 5 is an ad for *The Economist* encouraging viewers to subscribe to the publication. That ad *never* appeared on the targeted site. We found similar examples for *The New York Times*.

Targeting Per Category. Figures 10(b) and 10(c) show the number of ads targeting emails and websites, respectively in a particular category. For emails, we classify them based on their content. For websites, we use the Alexa categories. It is possible, and common, for Sunlight to detect that an ad targets multiple emails so the cumulative number of guesses represented in the figure may be larger than the total number of ads.

In Gmail, by far the most targeted category (topic) in our dataset was shopping (e.g., emails containing keywords such as clothes, antiques, furniture etc.). The second most popular targeted category was General health (i.e., emails with keywords such as vitamins, yoga, etc.). On the web, we did not observe a single dominant category as we did in Gmail. The News category, containing sites like The Economist and Market, was targeted by the most ads in the study but with only slightly more ads then the Home category.

Overall, these results demonstrate that Sunlight is valuable not only for investigators but also for researchers interested in broader aspects of targeting.

8 Related Work

§2.2 already discusses works closest to ours: web transparency tools and measurements [2, 6, 8, 15, 16, 18–22, 27, 29]. These works aim to quantify various data uses on the web, including targeting, personalization, price tuning, or discrimination. Sunlight is the first system to detect targeting at fine grain (individual inputs), at scale, and with solid statistical justification.

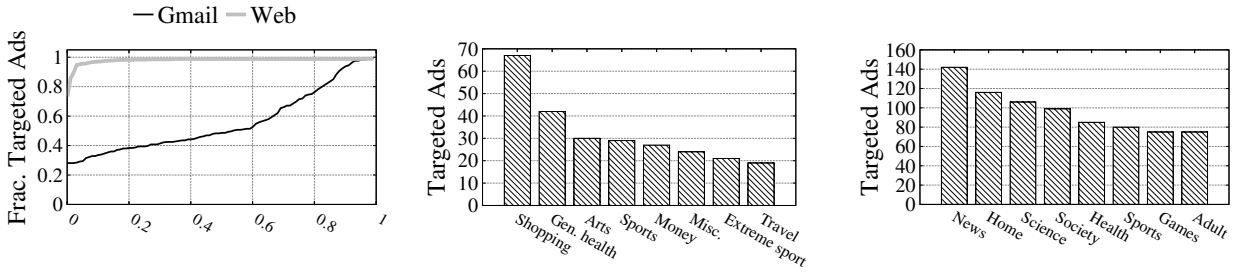
The works closest in spirit to ours are AdFisher [8] and XRay [18]; both of these aim, like us, to create generic, broadly applicable methodologies for various web transparency goals. **AdFisher** shares our goal of providing solid statistical justification for its findings, but, because of scale limitations, makes it hard to simultaneously track many inputs. So far it was applied to relatively coarse targeting (e.g., gender, a specific interest). Since AdFisher grounds its confidence in all outputs simultaneously, its results should be carefully interpreted: it rigorously proved that some targeting is taking place, but does not exhaustively and separately single out the output subject to this targeting. Finally this design disregards scalability with the number of inputs: the effect of each input and each possible combination of inputs needs to be tested separately.

XRay shares our goal of detecting targeting at scale on many inputs, but does not provide any statistical validation of its findings. Because of this lack of statistical confidence, XRay misses the inherent trade-off between scale in number of outputs and confidence in the results, that we evaluate with Sunlight. The effects of multiple hypotheses testing also change the choice of correlation detection algorithms. We found Sunlight's logit-based method to be significantly more accurate than the algorithms from XRay.

Our methods for statistical experimental design and analysis draw from the subjects of *compressed sensing* [9] and *sparse regression* [4, 26]. The experimental setups we consider correspond to sensing matrices that satisfy certain analytic properties that permit robust recovery of sparse signals. In Sunlight, these signals correspond to the hypothesized targeting effects we subsequently test and validate, and they are sparse when the targeting effects only depend on a few variables.

9 Conclusions

This paper argues for the need for scalable and statistically rigorous methodologies, plus infrastructures that implement them, to



(a) **Targeted Ads Out of Context CDF** (b) **Gmail Category Targeting** (c) **Web Category Targeting**

Figure 10: **Measurement Results.** (a) Shows what fraction of targeted ads appear how often in their targeted context in both Gmail and Web experiments. (b) Shows the number of ads are found to target emails in the eight most popular categories in the Gmail experiment. (c) Shows the number of ads are found to target sites in the eight most popular categories in the web experiment.

shed light over today's opaque online data ecosystem. We have presented one such methodology, developed in the context of *Sunlight*, a system designed to detect targeting at fine granularity, at scale, and with statistical justification for all its inferences. The Sunlight methodology consists of a four-stage pipeline, which gradually generates, refines, and validates hypotheses to reveal the likely causes of observed targeting. Sunlight implements this methodology in a modular way, allowing for broad explorations and evaluation of the design space. Our own exploration reveals an interesting trade-off between the statistical confidence and the number of targeting hypotheses that can be made. Our empirical study of this effect suggests that favoring high precision hypothesis generation can yield better recall at high confidence at the end of the Sunlight pipeline, and that scaling the number of outputs of an experiment may require to accept lower statistical semantics. In the future, we plan to break the scaling barrier by developing a reactive architecture that runs additional experiments to obtain the data necessary to confirm plausible hypotheses.

10 Acknowledgements

We thank the anonymous reviewers for their valuable feedback. We also thank Francis Lan for his work on early versions of Sunlight. This work was supported by a Google Research Award, a Microsoft Faculty Fellowship, a Yahoo ACE award, a grant from the Brown Institute for Media Innovation, NSF CNS-1514437, NSF CNS-1351089, NSF 1254035, and DARPA FA8650-11-C-7190.

11 References

[1] ADBLOCKPLUS. https://adblockplus.org/, 2015.

[2] BARFORD, P., CANADI, I., KRUSHEVSKAJA, D., MA, Q., AND MUTHUKRISHNAN, S. Adscape: Harvesting and Analyzing Online Display Ads. *WWW* (Apr. 2014).

[3] BENJAMINI, Y., AND YEKUTIELI, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* (2001), 1165–1188.

[4] BICKEL, P. J., RITOV, Y., AND TSYBAKOV, A. B. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist. 37*, 4 (08 2009), 1705–1732.

[5] BODIK, P., GOLDSZMIDT, M., FOX, A., WOODARD, D. B., AND ANDERSEN, H. Fingerprinting the datacenter: Automated classification of performance crises. In *European Conference on Computer Systems* (2010).

[6] BOOK, T., AND WALLACH, D. S. An Empirical Study of Mobile Ad Targeting. *arXiv.org* (2015).

[7] BRANDEIS, L. What Publicity Can Do. *Harper's Weekly* (Dec. 1913).

[8] DATTA, A., TSCHANTZ, M. C., AND DATTA, A. Automated Experiments on Ad Privacy Settings. In *Proceedings of Privacy Enhancing Technologies* (2015).

[9] DONOHO, D. L. Compressed sensing. *IEEE Transactions on Information Theory 52*, 4 (2006), 1289–1306.

[10] DUDOIT, S., AND VAN DER LAAN, M. *Multiple testing procedures with applications to genomics*. Springer, 2008.

[11] FELDMAN, V. Optimal hardness results for maximizing agreement with monomials. *SIAM Journal on Computing 39*, 2 (2009), 606–645.

[12] GOOGLE. AdSense policy. https://support.google.com/adsense/answer/3016459?hl=en, 2015.

[13] GOOGLE. AdWords policy. https://support.google.com/adwordspolicy/answer/6008942?hl=en, 2015.

[14] GRETTON, A., BOUSQUET, O., SMOLA, A., , AND SCHÖLKOPF, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory* (2005).

[15] HANNAK, A., SAPIEZYNSKI, P., KAKHKI, A. M., KRISHNAMURTHY, B., LAZER, D., MISLOVE, A., AND WILSON, C. Measuring personalization of web search. In *WWW* (May 2013).

[16] HANNAK, A., SOELLER, G., LAZER, D., MISLOVE, A., AND WILSON, C. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *IMC* (Nov. 2014).

[17] HOLM, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 2 (1979), 65–70.

[18] LÉCUYER, M., DUCOFFE, G., LAN, F., PAPANCEA, A., PETSIOS, T., SPAHN, R., CHAINTREAU, A., AND GEAMBASU, R. XRay: Enhancing the Web's Transparency with Differential Correlation. *23rd USENIX Security Symposium (USENIX Security 14)* (2014).

[19] LIU, B., SHETH, A., WEINSBERG, U., CHANDRASHEKAR, J., AND GOVINDAN, R. AdReveal: improving transparency into online targeted advertising. In *HotNets-XII* (Nov. 2013).

[20] MIKIANS, J., GYARMATI, L., ERRAMILLI, V., AND LAOUTARIS, N. Detecting price and search discrimination on the internet. In *HotNets-XI: Proceedings of the 11th ACM Workshop on Hot Topics in Networks* (Oct. 2012), ACM Request Permissions.

[21] MIKIANS, J., GYARMATI, L., ERRAMILLI, V., AND LAOUTARIS, N. Crowd-assisted Search for Price Discrimination in E-Commerce: First results. *arXiv.org* (July 2013).

[22] NATH, S. MAdScope: Characterizing Mobile In-App Targeted Ads. *Proceedings of ACM Mobisys* (2015).

[23] NG, A. Y. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning* (2004).

[24] RUBIN, D. B. Estimating the causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology 66* (1974), 688–701.

[25] SELENIUM. http://www.seleniumhq.org/, 2015.

[26] TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B 58* (1994), 267–288.

[27] VISSERS, T., NIKIFORAKIS, N., BIELOVA, N., AND JOOSEN, W. Crying Wolf? On the Price Discrimination of Online Airline Tickets. *Hot Topics in Privacy Enhancing Technologies* (June 2014), 1–12.

[28] WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E., AND LANGE, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics 25*, 6 (2009), 714–721.

[29] XING, X., MENG, W., DOOZAN, D., FEAMSTER, N., LEE, W., AND SNOEREN, A. C. Exposing Inconsistent Web Search Results with Bobble. In *PAM '14: Proceedings of the Passive and Active Measurements Conference* (2014).