

# LOWERING THE BARRIER FOR WEB ADVERTISEMENT RESEARCH AT SCALE

KAI JI (KEVIN) FENG

ADVISOR: ARVIND NARAYANAN

SUBMITTED IN PARTIAL FULFILLMENT  
OF THE REQUIREMENTS FOR THE DEGREE OF  
BACHELOR OF SCIENCE IN ENGINEERING  
DEPARTMENT OF COMPUTER SCIENCE  
PRINCETON UNIVERSITY

APRIL 2021

© Copyright Kai Ji (Kevin) Feng, 2021.

All rights reserved.

I hereby declare that I am the sole author of this thesis.

This thesis represents my own work in accordance with University regulations.

---

Kai Ji (Kevin) Feng

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

---

Kai Ji (Kevin) Feng

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

---

Kai Ji (Kevin) Feng

# Abstract

Web advertisements are essential to the day-to-day operations on the internet by providing a key channel of revenue to websites that offer content at little to no cost. However, they are also common sources of deception, scams, and privacy violations. Given their significance, ads are of interest to many different groups of experts, including web researchers, communications scholars, and regulators, but their fleeting nature makes them difficult to study systematically and at scale. This thesis presents AdOculus, a technical system comprising a search interface powered by automated visual analysis tools and a continuously updated, large-scale archive of ads crawled from thousands of popular websites. By using the system to uncover novel research questions, dimensions of analysis, and policy recommendations, I demonstrate how AdOculus and its underlying tools enable expanded possibilities in ad research.



# Acknowledgements

Writing may be a solitary activity, but writing a thesis certainly is not. I have many thanks to give to everyone who made this work possible.

First, I am very fortunate to have collaborated with Arunesh Mathur since the beginning of this thesis. Not only did Arunesh build the crawler that was essential to our data collection, he is also a dedicated mentor – generous with his time, expertise, and advice, and constructive with his feedback. Despite his significant research achievements, he still treats me like an equal collaborator, by which I am humbled and grateful.

I would like to thank Prof. Arvind Narayanan for advising me and pointing me in the right directions whenever I felt stuck. Arvind has the remarkable ability to carefully listen to a student and then quickly give insightful and pertinent advice. Through him and his wisdom, I’ve honed my research thinking, presentation skills, and ability to ask and answer hard questions. I’ve learned a lot through our weekly project meetings despite them being completely virtual.

I’d also like to thank members of the Princeton CITP community for their guidance and feedback: Mihir Kshirsagar for being my second reader and helping me navigate the foreign and intricate world of tech policy; Angelina Wang for her computer vision expertise and inspiring work in bias and fairness; Amy Winecoff, Ben Kaiser, Danny Chen, Elena Lucherini, Elizabeth Watkins, Kevin Lee, and Ryan Amos, for engaging questions and discussions during my project talks; and Ben Burgess for showing me how to train a custom model on Google Cloud. Research during a pandemic can feel like it’s happening in a vacuum, and I’m fortunate to have you all to bounce ideas off of. Additionally, Kevin Lee has taught me, through demonstration, that empathy is one of the most important qualities a grad student can have. I will definitely keep this in mind as I start my own journey in grad school.

I've also had the luck and privilege of being influenced by fantastic mentors outside of the immediate Princeton CS community. David Reinfurt from *vis* showed me the world of graphic design and visual reasoning, and how "the most liberal of arts" can be applied to just about anything. Prof. Marshini Chetty from the University of Chicago introduced me to HCI back in sophomore year, which has shaped my thinking, interests, and work in computer science since.

To friends who read or even gave feedback on this thesis – Annie, Bobby, Connie, Doug, Jackie, Rachel, Roger, Rohan, and Victor – thank you for taking time out of your busy days to learn about my work. In particular, thank you Connie for being my thesis fairy and checking in on me to make sure I didn't slide too far off the rails. To the communities on campus with whom I've shared some incredible experiences with – Rehack and the broader E-Club, Sinfonia, Band, Colonial, and Badminton – you are what makes time on this campus memorable. Thank you for transforming Princeton into the school and home that it is. To my roommates – Doug, Rachel, and Roger – our adventures are too numerous, too avant-garde, and too glorious to be put into words. Without you, I would've had no reason to procrastinate on this thesis. Thank you for an unforgettable four years, and for introducing a Canadian to America.

Finally, none of this would be possible without the unconditional support of my parents. They have shown me that no matter what I do, hard work and love will lead to great things, and that nothing being pursued is out of reach. They have opened countless doors for me, and I could not ask for more.

Thank you all.

*To my parents.*

# Contents

Abstract . . . . .	iii
Acknowledgements . . . . .	iv
List of Tables . . . . .	x
List of Figures . . . . .	xi
<b>1 Introduction</b>	<b>1</b>
<b>2 An Overview of Web Advertising</b>	<b>4</b>
2.1 Key Technologies and Agents . . . . .	5
2.2 Ad Design . . . . .	7
2.2.1 Display Ads . . . . .	8
2.2.2 Native Ads . . . . .	12
2.3 Ad Regulations . . . . .	15
<b>3 Literature Review</b>	<b>19</b>
3.1 Visual Analysis . . . . .	20
3.2 Deceptive Practices and Techniques . . . . .	24
3.3 Consumer Perceptions . . . . .	27
<b>4 Implications and Motivations</b>	<b>29</b>
4.1 Searchable Ads . . . . .	30
4.1.1 Dataset . . . . .	31

4.1.2	Search Interface and Visual Analysis . . . . .	33
4.2	Research Implications . . . . .	34
4.2.1	Exploratory Metrics . . . . .	35
4.2.2	Research Questions . . . . .	36
4.3	Policy Implications . . . . .	37
4.3.1	Consumers . . . . .	39
4.3.2	Intermediaries . . . . .	40
4.3.3	Publishers . . . . .	41
<b>5</b>	<b>Implementation</b>	<b>43</b>
5.1	Ad Collector . . . . .	45
5.1.1	Functional Overview and Implementation . . . . .	45
5.1.2	Limitations . . . . .	47
5.2	Ad Search . . . . .	48
5.2.1	Functional Overview and Implementation . . . . .	48
5.2.2	Visual Analysis . . . . .	52
5.2.3	Limitations . . . . .	65
5.3	Tackling Research Questions . . . . .	67
5.3.1	Self-disclosures . . . . .	67
5.3.2	Similar Ads on One Webpage . . . . .	69
<b>6</b>	<b>Evaluation, Findings, and Discussion</b>	<b>71</b>
6.1	Dataset . . . . .	71
6.2	Evaluation of AdOculos . . . . .	75
6.2.1	Visual Analysis Performance . . . . .	76
6.2.2	Feedback from Other Researchers . . . . .	80
6.3	New Research Possibilities . . . . .	81
6.4	Findings from Two Explorations . . . . .	82

6.4.1	Self-disclosures . . . . .	82
6.4.2	Ad Similarity . . . . .	85
6.4.3	Additional Policy Implications . . . . .	87
<b>7</b>	<b>Conclusion</b>	<b>90</b>
	<b>References</b>	<b>93</b>
<b>A</b>	<b>Code</b>	<b>102</b>
<b>B</b>	<b>Advertiser Industries</b>	<b>103</b>
<b>C</b>	<b>Top 30 Publishers by Number of Ads Archived</b>	<b>107</b>

# List of Tables

2.1	Fixed Ad Specifications outlined by the IAB. . . . .	9
3.1	Outline of literature review. . . . .	21
4.1	Metrics and their corresponding ad characteristics. . . . .	35
5.1	Collected ad metadata fields. . . . .	47
5.2	Summary of data associated with each analyzed ad. . . . .	50
5.3	Comparison of differences in common hashing algorithms. . . . .	61
5.4	Variables collected from top 30 publishers in our dataset. . . . .	70
6.1	Exploratory metrics retrieved from the dataset. . . . .	72
6.2	Summary of visual analysis accuracies for 188 ads. . . . .	77
6.3	Potential research questions to explore and groups who may be inter- ested in them. . . . .	83
6.4	Summary of mean and median metrics from self-disclosures. . . . .	85

# List of Figures

2.1	The first banner ad, hosted on HotWired.com . . . . .	4
2.2	A basic model of the ad ecosystem. . . . .	7
2.3	A takeover ad for a Czech bank on the popular Czech site centrum.cz. . . . .	9
2.4	An expanding Verizon iPhone ad on theverge.com. . . . .	10
2.5	An floating ad for health insurance on dictionary.com. . . . .	10
2.6	A variety of interactive display ads from the insurance company Progressive. . . . .	12
2.7	A native ad by Uber Eats on theverge.com. . . . .	13
2.8	A chumbox from Outbrain on wired.com. . . . .	14
2.9	An AdChoices ad not affiliated with Google (left) and one that is (right). . . . .	18
5.1	Manual analysis and clustering by information completeness. . . . .	44
5.2	Screenshot of the AdOculus homepage. . . . .	51
5.3	Overview of our ad collection and archival system. . . . .	53
5.4	Examples of ads with mouseprint indication. . . . .	55
5.5	Examples of self-disclosing ads. . . . .	56
5.6	AdChoices and/or Mute icons found on collected ads. . . . .	57
5.7	A visual example of the icon detection algorithm. . . . .	59
5.8	Ads for hashing comparison. . . . .	60
5.9	Prevalence of text-based brand identities in web ads. . . . .	62



5.10	Dominant colour extraction (a) vs. palette extraction (b) for $n = 3$ colours. . . . .	64
5.11	An M with a font size of 12px in its em-box. . . . .	68
5.12	Capturing placement of the self-disclosure marker through relative positioning. . . . .	68
5.13	Varying levels of RMS contrast. . . . .	69
6.1	Common words aggregated across 8859 ads. . . . .	73
6.2	Top brands (a) and industries (b) across 8859 ads. . . . .	73
6.3	Figure 6.2 broken down by week. . . . .	74
6.4	Prevalence of AdChoices (a) and mute (b), and matrix diagram with the two (c). . . . .	74
6.5	Tag cloud of top 100 objects represented across 8859 ads. . . . .	75
6.6	Top 10 publishers by number of ads archived. . . . .	75
6.7	Example of accuracy analysis of one ad through manual inspection. .	76
6.8	Histograms of font sizes (a) and RMS contrasts (b). . . . .	84
6.9	Visualization of 1048 self-disclosures. . . . .	84
6.10	Histograms of normalized $m_{ind}$ , $m_b$ , and $m_{same}$ across 30 publishers. .	87
6.11	Relationship between normalized $m_b$ and $m_{same}$ across 30 publishers.	87

# Chapter 1

## Introduction

Advertisements are ubiquitous on the modern web. Every time a website is loaded, a social media feed is scrolled, or a query is made on a search engine, hundreds or thousands of advertisers compete in automated auctions for their ad to be shown. These ads come in all shapes and sizes. Some are static images while others are dynamic animations. Some take up the entire screen while others are limited to a small corner. No matter their form, they all have one function: to entice the viewer to click on it.

At its core, the web advertising market is a constant battleground between advertisers and consumers. Advertisers are well aware that the internet is becoming increasingly saturated with ads and employ more eye-catching, click-harvesting ad design strategies. Consumers are becoming increasingly tired of ads' intrusive tracking, opaque data privacy practices, and disruption to web browsing experiences with slow loading times. Indeed, the proportion of internet users in the United States who use an ad blocker increased from 15.7% in 2014 to 25.8% in 2019 [84]. Although it is easy to dismiss web ads as a consumer, they are essential for the operation of the web: many smaller websites use ad revenue to stay afloat and large corporations rake in ad revenue to build more refined products that can better reach individuals who

seek benefit from them.

It would be an understatement to simply describe the ad market as “big”. Digital ad spending in the US was estimated to be \$121 billion in 2020, with a projection of \$153 billion by 2024 [85]. Google’s Display Network, a collection of more than 2 million websites on which web ads can be placed, is estimated to reach more than 90% of internet users worldwide [36]. A market of this size should also come with heavy responsibility. More individuals are joining the internet every day, many of whom are children and older adults who do not have extensive experience with technology. Unfortunately, this responsibility is not always upheld, as some ads employ deceptive techniques to specifically target vulnerable populations and lure them into making undesirable and potentially harmful decisions. Problematic ads such as these should be appropriately monitored to mitigate harm on the internet. Consequently, web ads are of interest to researchers and regulators alike.

However, the automated and sprawling nature of the ad ecosystem make web ads uniquely difficult to study. They appear and disappear from websites unpredictably, making it difficult for problematic ads to be tracked and traced to the relevant perpetrator(s). Ads are also dense with information. To extract the information, many current ad analysis techniques rely on manual inspection and qualitative coding, limiting the scale of conducted investigations. Scale is particularly relevant given the limited bandwidth of regulators and researchers to keep a close eye on this vast and complex space.

With the current climate in mind, I present my thesis, an effort to better facilitate accessible, interdisciplinary ad research and join the push towards higher quality, less harmful ads on the web. Concretely, this thesis contributes to existing work in the following ways:

- Creates a large-scale dataset of modern web ads, the first of its kind
- Builds tools that expose new possibilities in ad research

- Presents research questions in previously inconvenient or unseen dimensions for future work

I begin in Chapter 2 by providing some background to the world of web advertising, introducing key terms and underlying technologies. In Chapter 3, I summarize relevant work already done in this area. I use my literature review to motivate this thesis in Chapter 4 and introduce possibilities in technical, qualitative, and policy work. In Chapter 5, I describe the implementation of the primary technical system as well as methodologies for tackling my research questions. Chapter 6 is where I discuss the collected dataset, evaluate tools built, present my most salient contributions. Finally, I conclude in Chapter 7 by highlighting avenues for future work while emphasizing the importance of ad research.

## Chapter 2

# An Overview of Web Advertising

In 1994, AT&T purchased a small division of space next to the main written content on HotWired.com (now known as Wired Magazine), one of the early commercial web magazines. In that space, they showed a 476px by 56px rectangle that consisted of rainbow text on a dark background asking the viewer “Have you ever clicked your mouse right here? You will.” Upon clicking, the viewer is taken to a landing page where they could learn more about AT&T and its initiatives on the web. The rectangle is widely considered as the first banner ad, and it had a clickthrough rate (CTR) of around 44% [54].



Figure 2.1: The first banner ad, hosted on HotWired.com

Since then, web advertising has expanded at a blistering pace to occupy a unique and prominent space. It changed the way companies operated online and also the appearance of websites themselves. This can be clearly seen in HotWired.com shifting from a simple block layout in 1994 to a three-column design optimized for ad space in 1999 [95]. Ad revenue allowed internet companies to scale to tens of thousands of employees and develop products reaching billions of people around the world. In

turn, ad networks have grown more complex and expansive. As such, web advertising and internet growth can be described as mutually dependent.

In this chapter, I give a high-level overview of modern web advertising. I begin by discussing key technologies and agents in the ad ecosystem, and how they interface with each other to run a massive global market. Then, I outline some design paradigms for web ads, touching on the different types of ads along with their common technical and visual implementations. Finally, I discuss some current regulations on web ads and how they affect the ad content internet users see on a daily basis.

## 2.1 Key Technologies and Agents

The 2000s gave rise to more sophisticated forms of advertising technology. Most notably, in 2006, Right Media launched the first ad exchange featuring Real-Time Bidding (RTB) technology [86]. RTB allowed for automated, dynamic matching of *advertisers* and spaces in which ads can be published, known as ad inventory, hosted by *publishers*. The instant a user opens a webpage, before the page is fully loaded, companies of varying sizes and industries bid for available inventory through a Demand-Side Platform (DSP). DSPs serve as automated auction houses, taking an order from advertisers and bidding on the exchange on their behalf. DSPs then communicate their order to a Supply-Side Platform (SSP), which offers price and inventory selections on behalf of the ad publisher, or website hosting the ads. An RTB exchange facilitates these DSP-SSP auctions and a winning pair is selected milliseconds later [100], after which the advertiser’s ad is loaded and displayed in the purchased inventory. In some special cases, such as close DSP-SSP partnerships, agents can finalize a transaction directly without connecting to an exchange [32]. These cases, however, are far from the majority: soon after it was released, Right Media’s enterprise exchange was recording about 30,000 auctions per second [90].

RTB was revolutionary to web advertising in two major ways. First, it offered something traditional advertisers could not: fine-grained targeting. Advertisers could either collect consumers’ personal information or acquire it from third-party data brokers, known as Data Management Platforms (DMPs), and incorporate it into their bid strategy. This allowed advertisers to engage in the lucrative marketing practice of price discrimination – charging different consumers different prices depending on their willingness to buy a particular product. Publishers with inventory for sale can also use consumer data to package valuable details such as location and demographics of site visitors and charge advertisers plump premiums for prime ad slots. The tantalizing benefits of RTB to both sides resulted in unprecedented efforts to gather more consumer data and develop profiling technologies for hyper-specific, high-accuracy targeting. Second, RTB automated inventory away from control – advertisers could no longer manually choose where their ads appeared, and publishers can no longer dictate which ads from advertising partners appeared on their websites. For example, the same ad may appear simultaneously on the websites of *CNN* and *Playboy*. This drew criticism early on. In 2008, Wendel Millard, then-incoming chair of the Interactive Advertising Bureau (IAB) and former head of ad sales at Yahoo!, argued that it was a foolish shift to scale up the RTB model without firm control of where ads could be placed, resulting in inventory to be traded “like pork bellies” [37]. Despite concerns, RTB was still widely adopted and is the backbone of ad exchanges today [25].

Ad exchanges are one of two major channels through which web ads are distributed. The other is ad networks. Fundamentally, ad networks directly mediate interactions between advertisers and publishers. Their operational model is somewhat similar to that of traditional print or television advertising: inventory is aggregated from publishers and matched with advertisers’ demand. These networks are optimized for a variety of metrics, including impressions (ad views), clicks, and user

acquisitions [64]. They come in a wide variety of flavours. Some take a horizontal approach (distributing ads blindly to a vast selection of inventories), while others are more vertical (distributing ads selectively to a smaller set of inventories). Some specialize in a particular ad format, such as mobile or video ads. Some cater to only premium inventory with well-known publishers and high-quality traffic. Ad networks’ versatility and flexible pricing models can appeal to advertisers and publishers alike, but their targeting capabilities pale in comparison to those of ad exchanges due to the lack of automated, real-time optimization [64]. In the overall ecosystem, ad exchanges and networks can still connect to each other to trade ads and inventory despite operating on different bases. [32].

Figure 2.2 shows how ad agents interact in a basic model of the ad ecosystem. In concert, the interactions are often much more complicated, with multiple interconnected exchanges and networks. There may also be additional agents working with advertisers and publishers, such as ad servers and DMPs.

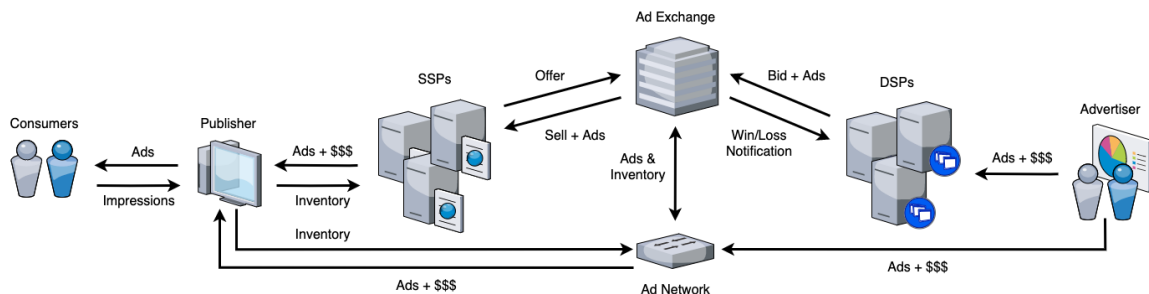


Figure 2.2: A basic model of the ad ecosystem.

For the purposes of this thesis, I will be using “ad intermediaries” as an infrastructure-agnostic term to refer to both ad exchanges and ad networks.

## 2.2 Ad Design

Ads on the web come in a variety of sizes and forms. Some are designed to camouflage with the host site’s content, while others try to stand out from the crowd. Some



attract attention using crude techniques such as gross-out images and clickbait, while others are widely recognizable cultural icons due to their creativity. Ad designs are also dependent on their format. For example, social media ads are often focused more on graphical content compared to search engine ads. New media technologies have allowed for even more diverse designs in experiences such as 360-degree image and video ads, virtual and augmented reality ads, and emoji ads [46]. In this thesis, I focus on digital ads that are displayed on websites, excluding search result pages and social media. This subset can be classified into two broad categories: display ads and native ads.

### 2.2.1 Display Ads

Display ads, also known as banner ads, are often what people think about upon hearing the words “web ad” thanks to both their ubiquity and visibility. These ads consist of an image (.jpg, .png, .gif) or video that when clicked on, will take the viewer to the advertiser’s landing page. Some more sophisticated ads will contain a multimedia wrapper that allows for interactive elements and popups. Display ads usually have a set size at which they are displayed. Fixed size ad units for display ads, specified by the Interactive Advertising Bureau [44], is outlined below:

Not all ads are fixed in the above sizes, however, especially in recent years when web design has become more sophisticated and creative. Additionally, the abundance of display ads has resulted in “banner blindness”: website visitors are so accustomed to seeing display ads that they no longer pay much attention to them. In 2018, the advertising company Outbrain reported a mere 0.05% CTR for display ads, aggregated across all formats and platforms [53]. Some display ads of unconventional sizes attempt to better capture viewer’s attention amongst rising cases of banner blindness. Common examples are as follows.

**Takeover ads** are static images used as a homepage’s background. They may

Ad unit name	Fixed size (px)
Billboard	970x250
Smartphone Banner	300x50 or 320x50
Leaderboard	728x90
Super Leaderboard / Pushdown	970x90
Portrait	300x1050
Skyscraper	160x600
Medium Rectangle	300x250
20x60 / Button	120x60

Table 2.1: Fixed Ad Specifications outlined by the IAB.

or may not have interactive components, such as call-to-action buttons. Unlike most display ads, these ads are usually not delivered in an automated fashion and require direct agreement between the advertiser and publisher to be shown.

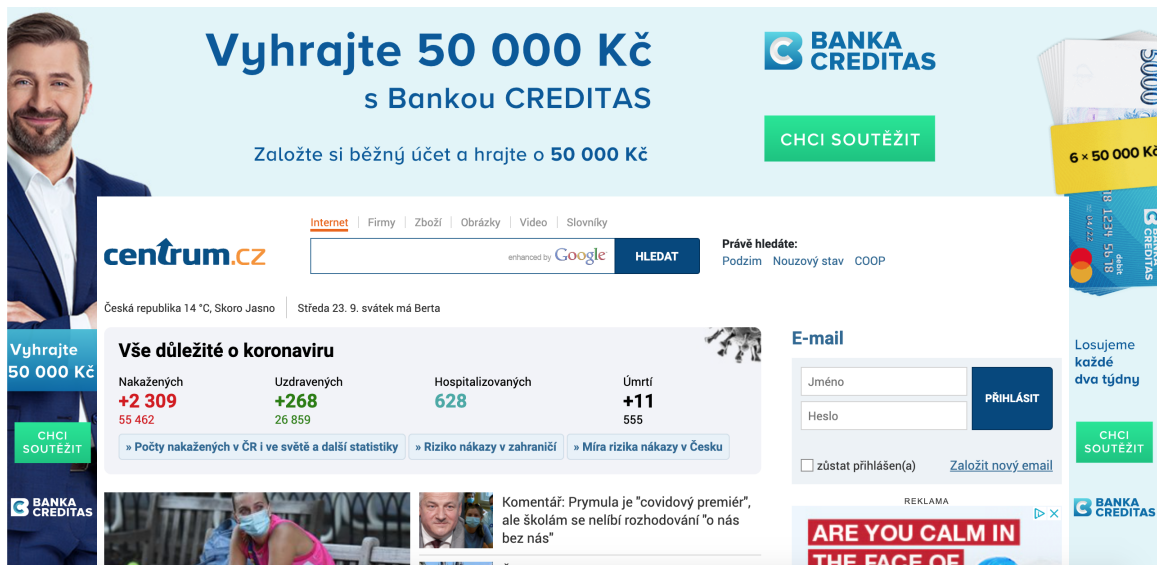


Figure 2.3: A takeover ad for a Czech bank on the popular Czech site centrum.cz.

**Expanding ads**<sup>1</sup> start out at an initial size and expand to a larger size auto-

<sup>1</sup>Expanding ads have also been found to be a point of entry for site hacks. In order to enable expanding capabilities, the ads use scripts that bypass the browser's Same-Origin Policy (SOP) security feature to make changes to the host page. This exposes cross-site request (XSS) vulnerabilities that allow an attacker to run malicious JavaScript on the site's server [9].

matically upon page load or by some viewer-triggered action such as a click or mouse hover. They can be static (image-based) or dynamic (video-based). They can often be dismissed with a close or shrink icon, or if the viewer simply scrolls further down the page. Figure 2.4 shows an expanding ad from before it loads, to its fully loaded state, to a more hidden form as the viewer scrolls away.

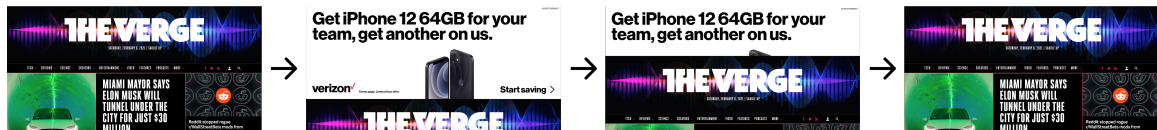


Figure 2.4: An expanding Verizon iPhone ad on theverge.com.

**Floating ads** are perhaps the most intrusive of display ads. They are a type of popup ad that “floats” over the page, usually displayed upon or soon after page load. When displayed, the site’s regular content is obscured and there may or may not be a means of escape on the ad, such as a close button. Mouse input may even be blocked until the ad disappears automatically, which is usually between 5 to 30 seconds [5]. These ads usually include animation and interactive components, and in some cases, even audio.

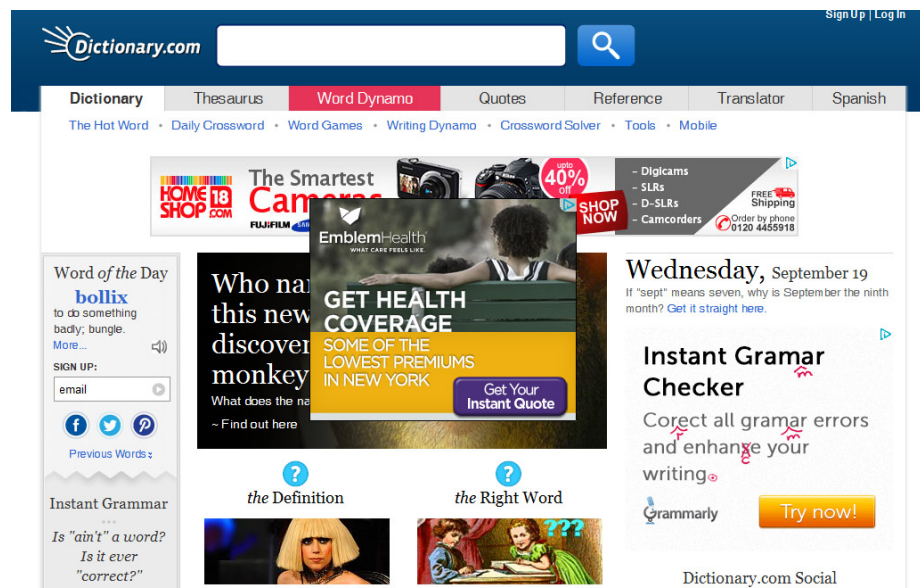


Figure 2.5: An floating ad for health insurance on dictionary.com.

Display ads delivered automatically through ad intermediaries, which includes all of the aforementioned ad types except for takeover ads, are usually implemented in one of two ways: inline JavaScript and iFrames [71]. JavaScript ad tags embedded directly into the website script send requests to an ad intermediary to show an ad in its given place. This technique allows for great flexibility in display size: multiple ad sizes can be shown within a single placement. Moreover, JavaScript tags are the only kind that support expanding ads. However, this technique can also be a disadvantage in that ads can be overly large and intrusive on websites that do not strictly enforce size limits. iFrame ad tags, on the other hand, apply stricter display restrictions to the requested ad. At the same time, iFrames also completely isolate the ad from the publisher’s website code and protect the publisher from improper ad serving, unexpected tampering of site code, and data leakages arising from the ad’s security flaws. It also allows site content to be loaded first and ads to be loaded as the viewer arrives at a certain point on the page, reducing the initial site sluggishness caused by ad service.

While takeover ads have a bit more freedom in the creativity of their visual designs, automated display ads are mass-produced and generally follow a standard template. Common elements of the template include the company logo or logotype, text containing the value proposition, a graphic related to the value proposition, and a call to action (CTA) in the form of text or a button [55]. Accepted design principles for such ads include using contrasting colours, selecting clear and legible fonts, and simplifying the layout whenever possible ([83], [58]).

While some display ads are interactive and others seem like simple static images, they are all built using interactive web languages – HTML, CSS, and JavaScript [40]. This also allows the ad to be responsive and adapt to various size limits without having to be redesigned. Although one can code up an ad as if it were a mini-website, the increasing prevalence of GUI-based ad design tools such as Google’s Web Designer,

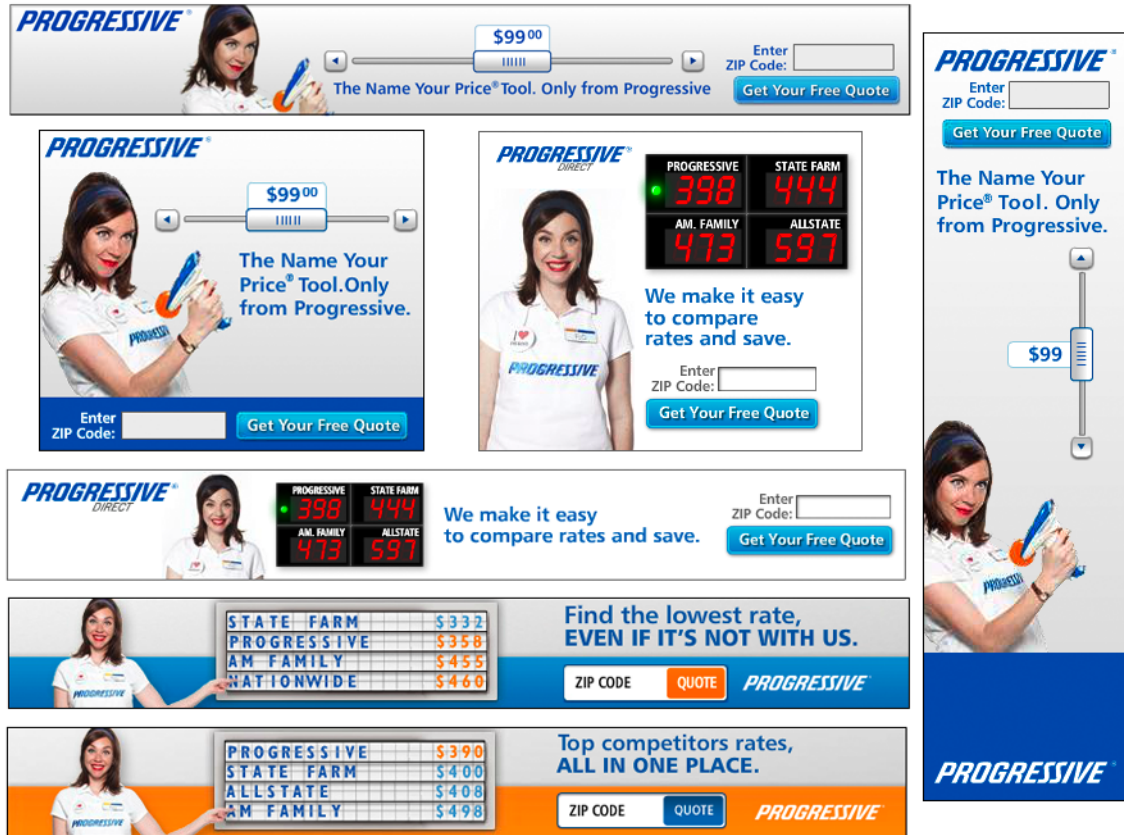


Figure 2.6: A variety of interactive display ads from the insurance company Progressive.

Adobe Spark, and Lucidpress means that marketing professionals don't have to write a single line of code to create and send an ad out for publication.

## 2.2.2 Native Ads

If display ads have flexibility in their design, native ads are the opposite: their design is constrained to that of their publisher. After all, native ads are meant to fit in seamlessly with the publisher's regular site content. Native ads are often found on search and promotional listings, news feeds, and content recommendation blocks. Outbrain reports an average CTR of 0.16% for native ads on desktop and up to 0.38% for premium native ads on mobile [53], both substantially higher than the CTR of display ads.

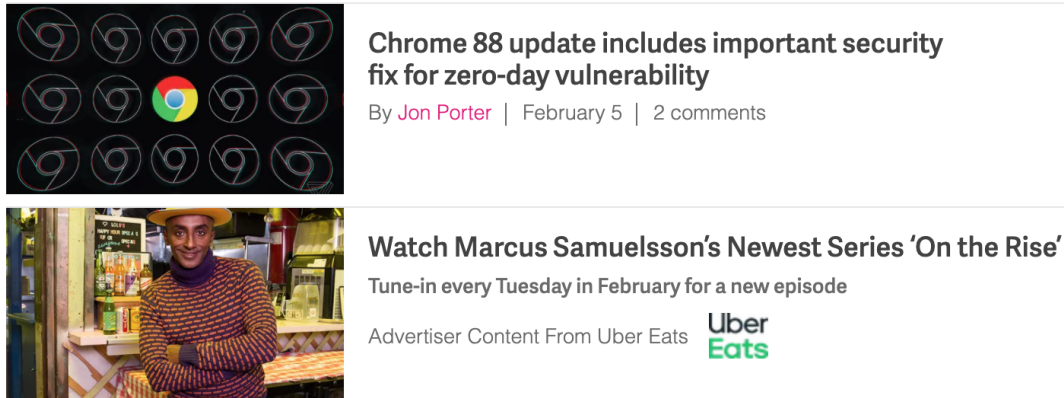


Figure 2.7: A native ad by Uber Eats on theverge.com.

Native ads are implemented as bits of web code wrapped in HTML divs with the same classes as one or more of the publisher’s site components. For example, a native ad on Wired.com may share the same divs and classes as an article listing in order to inherit the same styles. The ads then have additional classes or tag attributes that allow the advertiser’s content to be inserted along with impression and click tracking abilities. Some native ads are a result of a direct agreement between an advertiser and publisher, while others fall into a category called programmatic native advertising: the ads are delivered with RTB through DSPs and SSPs.

A subgenre of programmatic native advertising can be seen with particular prevalence around the web. They appear at the bottom of websites in a grid of thumbnails and captions usually labelled “Promoted stories” or “Around the web”. They mostly use low-level clickbait to capture attention, with headlines such as “You Won’t Believe What These Child Actors Look Like Today!” or “Surgeon Reveals One Weird Trick for Removing Dark Spots!” and images selected specifically to trigger excitement and/or shock from the viewer. Regions containing such ads are commonly known as *chumboxes*.

If regular web advertising were to be an infamous miscreant, the chumbox advertising market would be its brightest protégé. The chumbox market is dominated by only two companies: Outbrain and Taboola [39]. These companies pool together



**SPONSORED STORIES**POWERED BY **Outbrain** | ►**SECURITYSAVERS**  
The 1 Clever Trick Most Mac Owners Don't Know (Do It Today)**CAPITAL ONE SHOPPING**  
Read This Before You Renew Amazon Prime Again**HEALTH GAZETTE™**  
The Face Mask Everyone in America is Talking About**SMARTASSET**  
The Worst Way to Withdraw From Retirement Accounts**PREMIUM NANOTECH SPACE MASK**  
The Only Mask With Over 5,000 Verified 5-Star Reviews. Find Out Why**POST FUN**  
[Photos] Jackie Kennedy's Granddaughter Is A Billionaire

Figure 2.8: A chumbox from Outbrain on wired.com.

content from a vast number of companies specializing in curating viral and clickbait articles, such as LifeBuzz and ViralNova, to create a feed which can then be displayed on a publisher's site. Chumbox ads are especially lucrative due to their higher-than-average CTR compared to other native ads [39] – some publishers make up to 30% of their revenue from chumboxes. In 2020, Taboola reported an estimated \$1.2 billion in gross revenue and \$379 million in net revenue [14]. Noteworthy, companies like LifeBuzz can also make money off of Outbrain and Taboola by publishing chumboxes on their own site, closing the cycle on a profitable, mutualistic relationship.

The content within chumboxes is widely regarded as problematic and distasteful, mostly due to how viral content creation companies such as LifeBuzz write their articles. The companies hire many viral freelance writers, who are often loosely organized in an office with other similar writers, to churn out clickbait article after clickbait article [29]. The writers often take stories and images from around the internet without much regard to usage permissions and add a viral twist to them,

sometimes writing 3 or 4 versions of the same story for different clients. In 2018, a man found a picture of him and his deceased wife, who had passed away 10 years ago, on chumbox ads alongside false, clickbait-saturated titles about their relationship. Freelance writers had supposedly found his blog, which he started briefly after his wife’s passing to grieve and recover, and used the blog’s content and images without permission to spin into viral articles. The ad was eventually taken down, but not without a journalist’s trip to Taboola’s headquarters in Manhattan and a meeting with Adam Singolda, Taboola’s CEO, to bring the man’s case to attention [29].

## 2.3 Ad Regulations

The case of the chumbox ad raises a glaring and simple question: why aren’t there regulations in place to prevent unconsented use of images and falsified statements? Although not as blatant, similar types of deceptive activity can be found in other non-chumbox native ads as well as display ads. Why do such practices and spurious claims seem to run free in the web advertising space?

The structure of the regulatory system may yield some answers. In the United States, the Federal Trade Commission (FTC) is the main national regulator of internet marketing, as well as advertising more broadly [45]. In 1973, the FTC added Section 13(b) to the FTC Act to provide the Commission with the ability to seek, and upon sufficient proof, work with the court to issue a permanent injunction to an offending advertising agency [93]. The FTC issued its first 13(b) permanent conjunction in 1979, and since then, 13(b) has become the linchpin in its enforcement program with dozens of cases – including ones against Uber, Office Depot, and Volkswagen – handled under this section every year [93]. In 2000, in response to the growing trend of web advertising, the FTC published a “Rules of the Road” guidebook for internet advertising and marketing. The guidebook defined a piece of advertising to



be *deceptive* if it is likely to:

- Mislead consumers, and
- Affect consumers’ behaviour or decisions about the product or service

Additionally, a piece of advertising is *unfair* if the injury it causes, or is likely to cause, is:

- Substantial,
- Not outweighed by other benefits, and
- Not reasonably avoidable

Under the FTC Act, advertising that is unfair or deceptive in any medium is prohibited [7].

While this is appealing in theory, it is fiendishly difficult to achieve in practice. First, the guidebook definitions leave lots of room for advertisers to continue employing deceptive techniques. For example, one can argue that an ad telling viewers they can win \$1 million upon buying the company’s product is not misleading unless the probability of each person winning \$1 million is exactly 0. Therefore, headlines such as “Surgeon Reveals One Weird Trick for Removing Dark Spots!” are allowed to be published: they do not mislead consumers because there is a non-zero chance that the “trick” can help at least one person, and it is easily avoidable because viewers are not forced into trying the trick.

Second, the FTC is not responsible for reviewing ads to make sure they abide by guidelines or substantiating ad claims – the publisher is [7]. Since the publisher is motivated by profit, it may run ads that violate FTC and state guidelines if an ad can rake in substantial revenue. Ads that yield higher profits generally have higher CTRs, and higher CTRs can be generated with various low-level clickbait techniques, most of which prioritize grabbing attention over factual correctness. The

reliance on the publisher is further complicated by Section 230 of the Communications Decency Act [94]. For example, if various advertisers publish bogus ads on a shopping platform, Section 230 frees the platform of any legal responsibility for claims made by advertisers. There is then little motivation for the platform to deploy additional resources to police the ads when the ads simply help them earn more revenue.

Third, the sheer scale at which web advertising operates makes overseeing the market difficult. Issuing punishments in a multibillion-dollar market with more than 300 million consumers is unrealistic given the FTC’s limited resources. Moreover, while the FTC can legally deprive a violator of advertising rights, the Commission does not currently have a system in place to expeditiously return ill-retrieved gains to victims. Without monetary punishments and equitable relief, the FTC is left with a decades-old statute that allows wrongdoers to retain funds they took from their victims, ultimately making consumers more vulnerable to harm. A proposed solution to this issue is to establish a civil penalty fund in which the FTC can pass fines from advertisers directly to impacted consumers [93].

While the FTC is the main national regulatory body for web advertising, it is not the only one. The Digital Advertising Alliance (DAA) is a non-profit organization that launched the YourAdChoices program in 2010 to establish more transparency and consumer control around behavioural targeting in online ads [15]. The AdChoices icon is shown automatically on an ad by participating advertisers if web-based consumer data, such as browser history or cookies, is being collected or used. Upon clicking on the icon, which is usually located in the top right corner of an ad, a consumer can learn more about how their data is being used and opt-out of ad targeting. The AdChoices icon is commonly accompanied by the “mute this ad” icon, a measure introduced by Google in 2012 [34]. A consumer can click on this icon to stop seeing ads from a particular campaign. However, muting does not guarantee that a consumer will not see the ad again. The ad can be served to them from a different advertiser or they

clear their cookies and consequently reset their muting record. Figure 2.9 shows the slight variation in icon styles based on the ad intermediary. The left example is not using a Google-based intermediary, while the right is and bears the mute icon.

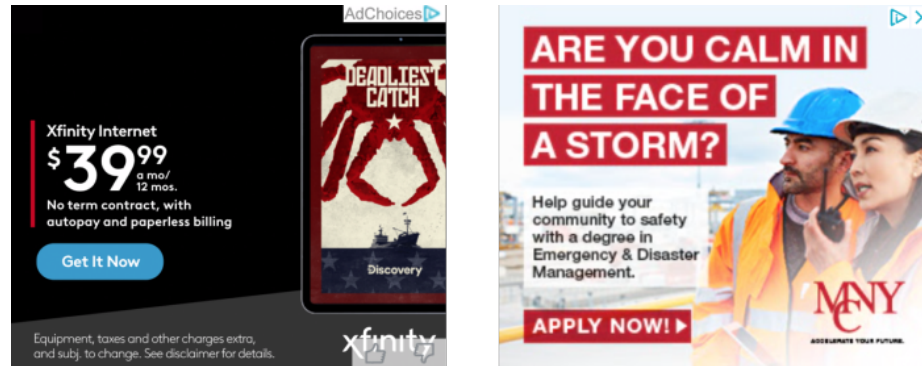


Figure 2.9: An AdChoices ad not affiliated with Google (left) and one that is (right).

Another non-profit organization, the Interactive Advertising Bureau, works with over 650 companies running web ads to develop industry standards (such as the display ad sizes in outlined Table 2.1), conduct research on the advertising market, and liaise between advertisers and legislators. The IAB Tech Lab helps companies implement technical standards and solutions such as Data Transparency Labels and enhanced seller validation to support a healthier market [43]. It is important to note that both the YourAdChoices program and IAB initiatives are self-regulatory: advertisers, publishers, and ad intermediaries are expected to work together to follow standards and best practices. Neither the DAA or IAB has legal power to punish perpetrators [45]. They can, however, accept reports of violations and escalate issues to partner organizations, the FTC, and/or the legal courts.

All in all, the advertising regulation system in the United States is a *reactive* one. Violators may be brought to regulatory attention and punished on a case-by-case basis if they are reported. Otherwise, market activities are assumed to be under control via self-regulation.

# Chapter 3

## Literature Review

In the previous chapter, I outlined how web advertising works on a broad level. Understanding the theory behind the system is an important step forward, but a holistic grasp of the modern online advertising empire would not be possible without understanding how web ads interact with the real world in practice.

Revenue from ads provide a financial lifeline to many internet organizations, and yet negative ad experiences prompt an increase in ad blocker usage [84]. Thus, ads walk the fine line between persistence and annoyance. If an ad is not persistent enough, it may not capture sufficient clicks, especially within a sea of competitors. On the other hand, if it is too persistent, consumers may simply block it out, which not only comes as a cost for that particular ad but also the ad industry as a whole. This may be one of the reasons why web advertising is a well-studied space. For more than a decade, researchers have followed the advertising industry, observed patterns and techniques to understand the impact of ads on the web and people, and developed frameworks to better organize ad-related research. In this chapter, I summarize some of these works. I will first discuss relevant literature in the domain of visual analysis for advertisements, online and otherwise. I then turn to a related topic of problematic content and techniques in web advertising. I conclude by briefly

addressing some findings on consumer perceptions of online ads.

Before I continue, it is perhaps equally important to note what is out of scope for this discussion. This thesis primarily focuses on the collection and automated extraction of visual information from web ads, without diving deep into details on web privacy and user tracking. Although there is important literature on privacy topics such as using microtargeted ads to identify Google personal profiles [11] and tracing information flows between ad exchanges to infer user behaviour [4], I do not discuss them in detail outside of broad consumer perceptions. Furthermore, email ad campaigns make up for a non-trivial portion of web advertising and while their studies help portray a more colourful and holistic picture of web advertising, I do not touch on them because this thesis does not collect or analyse them in any way. In a similar vein, I omit any work focusing exclusively on mobile ads. Table 3.1 provides an outline of my literature review.

### **3.1 Visual Analysis**

The old marketing adage “where the eye stops, the sale begins” certainly rings true when considering the visual properties of ads – they employ a wide range of visual techniques in an effort to capture the viewer’s attention. Visual analysis of ads involves making sense of what and how visual information is communicated, where visual information can include people, objects, textual content, typographic style, colour palette, layout, and more. The desire to better understand these characteristics did not spark recently – psychology and communications researchers began developing methods to formally study the relationship between an ad’s visual properties and marketing effectiveness well before web ads existed [38]. In fact, some of the works I discuss in this domain are rooted in print advertising rather than web advertising, but their findings are relevant across mediums due to similarities in

Table 3.1: Outline of literature review.

Category	Overview	Sample Research Questions	Example Works
Visual analysis	Examination of ads through their visual content, such as textual information, typographic style, colours, graphical content, layout, etc.	<ul style="list-style-type: none"> <li>How does an ad’s visual design affect consumer purchasing intentions?</li> </ul>	<ul style="list-style-type: none"> <li>Hussain et al.</li> </ul>
		<ul style="list-style-type: none"> <li>What classification methods and frameworks can be used to organize ads based on their visual features?</li> </ul>	<ul style="list-style-type: none"> <li>Shaouf et al.</li> <li>Phillips and McQuarrie</li> </ul>
		<ul style="list-style-type: none"> <li>How are visual metaphors used and interpreted in advertising?</li> </ul>	<ul style="list-style-type: none"> <li>Ye and Kovashka</li> </ul>
Deceptive techniques and practices	Investigation of how ads employ deception in the wild	<ul style="list-style-type: none"> <li>How often do people misidentify ads as non-promotional content?</li> </ul>	<ul style="list-style-type: none"> <li>Darke and Ritchie</li> </ul>
		<ul style="list-style-type: none"> <li>How do viewers respond to skeptical claims made by ads?</li> </ul>	<ul style="list-style-type: none"> <li>Johar and Roggeveen</li> </ul>
		<ul style="list-style-type: none"> <li>What qualities and/or practices are considered deceptive?</li> </ul>	<ul style="list-style-type: none"> <li>Swart et al.</li> </ul>
User perceptions	Study of internet users’ attitudes toward web advertising	<ul style="list-style-type: none"> <li>What makes a bad ad “bad”?</li> </ul>	<ul style="list-style-type: none"> <li>Zeng et al.</li> </ul>
		<ul style="list-style-type: none"> <li>How do social media users respond to ads on their platform?</li> </ul>	<ul style="list-style-type: none"> <li>Wei et al.</li> </ul>
		<ul style="list-style-type: none"> <li>What outlook do consumers have on targeted advertising and personal data collection?</li> </ul>	<ul style="list-style-type: none"> <li>Chanchary and Chiasson</li> </ul>

foundational marketing principles.

To measure the effects of visual complexity on ad desirability and comprehension, Pieters et al. posited two distinct classifications of visual complexity: feature and design [77]. Feature complexity is an unstructured metric defined using variation in pixel differences of an ad image to capture details in colour, luminance, and edges, whereas design complexity is a structured metric based on the variation of specific shapes, objects, and their arrangements in the ad. Their study involving 206 college participants found that an ad’s feature complexity had a negative effect on attention towards the advertised brand and attitude towards the ad, but design complexity contributed positively to ad attentiveness as well as overall attitude. Similar results were seen in a study specific to web advertising. Shaouf et al. found that enhanced visual cues in a web ad (e.g. appropriate colours and high-quality images) had direct positive effects to both advertising attitudes and brand attitudes [82]. Interestingly, a direct effect of visual design on purchasing intentions is found to be stronger in male participants, indicating that visual advertising strategies may not be a one-size-fits-all solution across genders.

Phillips and McQuarrie took inspiration from visual metaphors used in advertising and proposed a typology of visual rhetorical figures based on their complexity and ambiguity [76]. They did so to distinguish various graphical strategies advertisers have at their disposal and linked them to consumer response. Based on previous work, they predicted the effects of complexity (visual structure) and richness (meaning operation) on consumer outcomes such as recollection and change in beliefs. The predictions were not tested until 6 years later, when Jeong ran a study with 261 undergraduates and confirmed that metaphorical visual rhetoric contributes positively to cognitive elaboration and perceived source credibility [47]. However, visual metaphors can sometimes fly right over viewers’ heads, as Petridis and Chilton pointed out [75]. Their study recruited 20 participants on Amazon Mechanical Turk and gave them 48

visual metaphors to interpret without any explanatory text, and the metaphors were correctly deciphered only 41.3% of the time.

More recent studies have employed computer vision techniques in an effort to perform automated visual analysis. Hussain et al. collected 64,832 image ads and 3,477 video ads and annotated them such that they encompassed the topic and sentiment of the ads, along descriptions of what actions the ad is prompting the viewer to take [41]. They then trained a classifier to predict ad topic, sentiment, and symbolism, as well as automatically answering viewer action-related questions posed by the ads. Their accuracies ranged from 60.34% in predicting topics to 11.48% in question-answering, both for image ads. Some follow-up work include that of Ye and Kovashka [99], in which they refine automated decoding of visual metaphors, focusing mostly on public service announcements (PSAs). They do this by building a system that when given an ad image and several textual statements, learns an image-text embedding that associates the ad with the most relevant message. Their system outperformed a modified variant of the question-and-answer context retrieval system described in [41] – its accuracy was higher by almost 10%. Another work examined automatic face creation for various ad categories [91]. The authors used work from [41] to predict semantic attributes and inter facial expressions as a supervisory signal when training their model. They found that their models were not only capable of producing visually distinct faces for a collection of fixed ad topic categories, but also outperformed some baseline models developed using generative adversarial networks (GANs) tailored for similar purposes.

Another relevant computer vision challenge is brand recognition. This has many applications, including counterfeit detection, competitor ad monitoring, and ad verification. Brand recognition in advertising presents a particularly formidable problem due to the need for a general purpose model capable of recognizing a large variety of brands. In some mediums, such as print, where the average number of brands



advertised in a newspaper or magazine usually falls below 100, the challenge is not as prominent. However, on the web, where RTB automatically feeds ads from many thousands of advertisers, conventional datasets such as FlickrLogos27 and QMU-LOpenLogo simply do not contain enough categories to train a classifier capable of accurately identifying brands in the wild. As such, many notable contributions made in this area come in the form of datasets. Su et al. released WebLogo-2M in 2017, a logo recognition dataset containing 194 logo classes and over 2 million images [88]. The “web” in its name refers to its creation method of automatically sampling Twitter stream data, not necessarily implying that the logos are commonly found on the web. Wang et al. introduced Logo-2K+ in 2019, with 2341 classes and 167,000+ images [97]. The team also proposed a region-discriminatory network that segments logos’ regions based on their informativeness and augments informative regions for enhanced logo recognition. In late 2020, Jin et al, published the largest known brand detection dataset with rich annotations [48], containing more than 1.4 million images, 1216 logos, and 559 brands. The authors also trained a custom model using their dataset and saw an improved mean average precision (mAP) over state-of-the-art models trained on previous datasets. However, it is important to note that their dataset contained a disproportionate amount of Chinese brands due to collection methods. It is unclear whether the team took that skew into consideration when evaluating their model.

## 3.2 Deceptive Practices and Techniques

As more and more ads fight over viewers’ clicks, their techniques can become increasingly reckless and their content increasingly problematic. This can be deception aimed at seizing attention through false claims made by advertisers or excitement-inducing graphics that mislead viewers, or inversely, deception that evades attention by disguising an ad as regular, non-promotional content. Zeng et al. broke down

problematic content in ads by organizing them into a collection of qualitative codes and found that the level of such content does not vary drastically between ads on sites known for misinformation and ads on mainstream news sites [101]. When uncovered, deceptive techniques harm the advertising industry at large. Darke and Ritchie showed that consumers' awareness of deception lead to defensive behaviour, more distrust of other advertising claims regardless of their veracity, and a worse attitude towards advertising that spans across geographic regions, product categories, and advertisers [13].

A study by Johar and Roggeveen used source credibility (or lack thereof) to examine the effects of false claims on consumers. It revealed that consumers can easily be misled in the multi-channeled advertising world [49]. If a source that previously delivered them credible information started shifting towards less credible content, consumers were unable to re-calibrate their expectations and viewed the information as more favourable than warranted. This is especially pertinent given the complexity of web advertising and difficulty in separating whether an ad message originated from the brand sponsor or a third party. On a related note, Roggeveen and Johar found that a higher variation in sources contributes positively to belief in repeated claims that may be initially perceived as having low plausibility [80]. In web advertising, since the same ad can appear across a wide range of websites, consumers may view the various publishers as distinct sources for ad claims, amplifying the implications of the study. The pair also found that direct refutations of assertions are more likely to be accepted than ones of implications [50]. They hope their research can provide guidance to policymakers and public organizations as to how to best effectively design refutations of false claims.

Another area of deception concerns ad visibility. Previous work has shown that ads are sometimes not properly disclosed [89], especially ads that take the form of affiliate marketing and/or endorsements on social media [102]. Upon studying disclosures

in YouTube and Pinterest, Mathur et al. found that only about 10% of affiliate marketing content contained any disclosures at all [62]. Even if the disclosure could be seen, users may not understand them if they are short and non-explanatory. Given the importance of disclosures for ad recognition [42], the FTC created a guidebook in 2019 that outlines when and how social media influencers should disclose ads [28]. Furthermore, there has also been growing concern over native ads and their ability to hide too well among regular content. A study involving 738 US adults found that fewer than 1 in 10 recognized a sponsored news article as a native ad on a news website, and that recognition was more probable among younger, more educated demographics [2]. Wojdyski and Evans ran a similar study with a different population and benchmarked the native ads with “regular” display and video ads [42]. Averaged across platforms, advertisers, and labeling techniques, 37% of respondents correctly identified examples of native ads as sponsored content, compared to 81% for regular advertising. They also evaluated the effect of label position on ad recognition for native search result ads, and found that the ads were most likely to be correctly identified if they were positioned at the middle and bottom of a search results page, which is not where they are usually placed in practice.

There have also been explorations of misleading user interfaces known as dark patterns, none of which directly consider web ads but are nonetheless relevant due to their prominence on the internet. Mathur et al. developed a taxonomy of dark pattern characteristics to better describe and classify their persuasive harm on consumer decision making, focusing on dark patterns on shopping websites [60]. 11,000 shopping sites were crawled and analyzed, revealing 1800+ instances of dark patterns and 183 websites that engage in deceptive UX practices. Even after the implementation of GDPR, dark patterns found on popular UK websites continued to violate requirements set based on European law [70]. Narayanan et al. points out that designers should be increasingly wary of dark patterns as public awareness of the practice

grows; there is more potential than ever for practicing brands to be disgraced under the scrutiny of regulators, journalists, and academics [68].

### 3.3 Consumer Perceptions

In addition to studying ads themselves, researchers have also gleaned key insights from exploring consumer perceptions of ads. In their most recent work, Zeng et al. assembled a taxonomy of positive and negative consumer reactions to web ads upon surveying 60 internet users and developed a classification of specific ad characteristics that contributes to negative reactions [24]. From their set of 500 ads crawled from popular websites, they found that consumers consider a substantial portion of the set to be clickbait, untrustworthy, or distasteful. While the team does not touch on social media ads, a study of French-speaking Twitter revealed common user criticisms of ads on the platform as “fail”, “cringe”, and “clickbait” [30]. Twitter users in the US and UK also found some lesser-known ad targeting mechanisms used by the platform, including advertiser-uploaded lists of specific users and lookalike audiences, to be among the most privacy invasive [98].

More broadly, O’Donnell and Cramer’s study involving 24 US teens and adults found that although consumers recognize the usefulness of personalized ads, they felt that many data collection practices overstepped boundaries of their private lives [73]. Such targeting can induce discomfort to oneself as well as embarrassment in the presence of others. Chanchary and Chiasson argue that consumers do not inherently have a negative perception of online targeted advertising, but rather the opposite – they appreciated the concept and some would be more willing to share data if they were given finer control mechanisms for tracking protection tools [8].

O’Donnell and Cramer also showed that an inability to recognize an advertised brand drastically reduces the ad’s trustworthiness [73]. With regards to native adver-

tising specifically, Cramer found that native ads that were considered high quality in isolation could still negatively impact credibility and perceived quality of their host site if the ad is too similar to the site's regular content [12]. Therefore, the ideal native ad should strike a fine balance between blending in and standing out.

# Chapter 4

## Implications and Motivations

Chapter 3 shed some light on both researchers' interests as well as the diversity of research challenges in the field of web advertising. Although there exists significant literature in this space, the explorations thus far are by no means comprehensive. In fact, some previous works raised more questions than they set out to answer, as the act of researching ads itself brought about new questions that were not originally on researchers' agendas. One problem of particular interest is making large-scale ad research approachable to experts in a wide range of disciplines, from computer vision to human-computer interaction (HCI) to public policy.

My thesis hopes to add to the domain of web advertising by contributing tools towards ad research as well as exploring technical and societal possibilities with those tools. In this chapter, I provide justification behind key components of my research. I start with motivations for a dataset of web ads collected automatically from popular websites and a search interface for interacting with the dataset. I then discuss two research questions generated from the dataset and reasons for pursuing them. I end with policy implications that can be surfaced by our engineering and research contributions.

## 4.1 Searchable Ads

Although it is well known that digital ads can be problematic, they have, to my knowledge, neither been archived nor studied systematically on a scalable level. Harmful content in ads can range from promoting morally dubious products or services, such as gambling, to images triggering political polarization and headlines brimming with misinformation. Furthermore, many ads with harmful content may be targeted towards vulnerable populations such as children, seniors, and low-income individuals, which calls for deeper investigation into their practices.

However, such investigations are complicated by the fleeting nature of ads. A problematic ad may be shown to a viewer before disappearing minutes or even seconds later back into the complicated depths of the RTB ecosystem. There is no indication of when the ad will appear again, and refreshing the webpage may display a completely new cohort of ads altogether. To capture web ads for detailed study, my collaborators and I propose a continuously updated dataset of web ads along with a search interface to allow insightful interaction with the data.

It is important to note what ads we are and are not archiving, as the term “web ads” encompasses a vast range of digital promotional content. We are interested in observing ads that are not platform-specific and can be seen by diverse demographics. Concretely, this refers to display and native ads that are fed to viewers when they visit a publicly accessible webpage, such as a New York Times article, or a post on someone’s ad-enabled personal blog. This excludes ads on social media platforms, native ads on result pages of search engines, ads on shopping platforms, and email ads. As an initial step, we also limit ourselves to capturing static images of ads as opposed to entire animation sequences or videos, although we have plans to extend our reach to include such ads once we are confident in our capture and analysis pipeline. Finally, we acknowledge that mobile ads may come from different ad networks and we focus on ads in desktop browsers. For the remainder of this thesis, I use the term

“web ads” to refer to platform-agnostic, desktop ads we attempt to archive. Should the need arise, I use “digital ads” to refer to the broader umbrella of ads on the internet.

### 4.1.1 Dataset

A limited number of previous works collected large quantities of ads for study, with varying methodologies. I briefly highlight them to motivate our dataset.

Hussain et al. [41] did not tailor their work specifically to digital ads, so their methodology involved assembling a list of keywords on possible ad topics and downloading images from Google upon querying with each keyword. Each query returned about 600-800 images, resulting in a noisy set of 220,000 images overall. The team performed duplicate removal and then used a classifier to determine whether an image was an ad. They did so by sending a subset of 22,000 images to MTurk for human annotation and using the annotated images to train the classifier. In total, 64,832 images from the initial set (29.5%) were classified as ads.

Zeng et al. ([101],[24]) built a web crawler using Puppeteer [78], an automation and instrumentation library for the Chromium browser. The crawler takes in URLs and visits them, identifying ads on the visited pages by matching on common ad CSS selectors and domains from the EasyList filter list for Adblock Plus [19]. For each ad, their crawler takes a screenshot, stores its HTML content, and then clicks into the ad, and screenshots and scrapes the corresponding landing page. The team ran the crawler on both mainstream news sites and sites known for misinformation for 4 days in January 2020 and collected 55,045 valid HTML elements. However, due to the lack of automated visual analysis tools, the researchers could perform only manual analysis on a small subset of 5413 elements. From that, they found 2995 (55.3%) of elements did not contain a screenshot, were occluded, or did not capture ad content, leaving 2418 ads available for their qualitative coding.



More recently, the NYU Ad Observatory collected political Facebook ads to investigate the role of such ads in US elections [72]. They used a combination of information from Ad Observer, a browser extension that volunteers install to automatically share data on Facebook ads they see with the team, and the Facebook Ad Library API. As of November 2020, the team had about 16,000 browser extension volunteers [20], but it is unclear how many ads were collected.

With these methodologies in mind, we see two missing components in this domain we can address: data accessibility and timing.

**Data accessibility:** while technical researchers can build crawlers, browser extensions, and computer vision models to gather and clean their own datasets, scalable ad collection is not easily accessible to experts in other fields who may wish to study web ads but do not have the resources or relevant skills to code up their own pipeline. Such individuals, including journalists, communications professionals, and regulators, can often directly influence ad policy by bringing them to public attention and should not be left out of research efforts due to technical barriers. Zeng et al. may have collected a reasonable sample of web ads, but to our knowledge, other researchers who would like to use or contribute to it would not be able to do so without establishing some form of direct partnership. Although the NYU Ad Observatory team had their collection of ads available to the public before the 2020 election, their dataset was limited to political ads on Facebook. Hussain et al. also had a publicly available set of collected ad images, but since they were downloaded from Google’s image database, the majority were print ads from magazines and newspapers that did not have high relevance to modern web advertising outside of basic marketing and visual principles.

**Timing:** previous attempts at ad collection have been limited to a single point in time or a small span of time. Resulting datasets do not allow for studies that garner insight from longer time trajectories, such as quantifying the frequency of an ad’s appearance on a given website over a month. Furthermore, operating over

an extended period of time can help scale up collection efforts and simply provide more ads as material for analysis. An exception may be the NYU Ad Observatory’s use of a browser extension to continuously retrieve ad information from volunteers’ Facebook feeds. The problem lies in the retrieval of the ad itself using Facebook’s Ad Library API, which may not be possible as the Ad Library only contains a dataset of ads currently deployed across all Facebook apps and services [26]. The library’s API has also been reported as notoriously tricky to use, even for experienced software engineers [18].

Citing these two gaps, we propose a continually updated dataset of web ads collected and archived automatically from thousands of popular websites. The dataset’s full potential, however, would not be realized without scalable visual analysis and a usable search interface.

### 4.1.2 Search Interface and Visual Analysis

A user-facing layer is crucial in allowing others to browse, understand, and retrieve new insights from our dataset. As such, we propose a search engine for collected web ads. The natural followup question is: what characteristics will our ads be searched and filtered on, and how will we extract those characteristics?

The next chapter will take a deep dive into this question, but our general approach is visual information extraction with computer vision and image processing tools. We realize the importance of automating visual analysis – although Zeng et al. crawled over 55,000 potential web ads, they could only manually analyze less than 10% of it due to the lack of automation. Considering other works, the word “lack” does not necessarily mean the absence of such tools, but *accurate* and *usable* forms of those tools. For example, Hussain et al. employed some automated techniques to predict broad characteristics (whether or not an image was an ad) as well as detailed characteristics (actions prompted by the ad). The broad characteristics identified can

be useful as an early-stage filter in the data cleaning process, but are too broad to generate productive analysis. Attempts at extracting detailed characteristics consistently yielded accuracies of below 50%, making them too unreliable. We believe there is a balance to be sought between granularity, accuracy, and utility, and we target this balance in our approach. We see from past work that qualitative coding can help retrieve fine-grained features that automated techniques tend to miss, but relying purely on manual analysis is simply not a scalable solution. We aim to develop automated tools that are accurate enough to enable analysis at scale while leaving flexibility for researchers to layer on top their own custom analyses.

We have several potential use cases for our dataset and search interface. **Pressure groups** can identify which publishers are running ads from dubious advertisers and try to convince them to stop running those ads. **Advertisers** can see what other ads their publisher is running and discontinue their relationship if needed. **Regulators** can use problematic ads as evidence in investigations and lawsuits. **Computer vision researchers** can easily access weakly labelled ad images to train their models. **Journalists** can quickly identify trends in the advertising space and bring them to public attention. More details on future possibilities can be found in Chapter 6. All in all, we hope that our tools can help mediate more approachable, interdisciplinary ad research.

## 4.2 Research Implications

With a dataset of ads in our arsenal, we seek out visual features we can extract from each web ad that provide us with a higher vantage point to observe the web advertising industry at large. We also hope that our extractions at scale can offer interesting metrics that open doors to unexpected research questions. The following subsections go over some metrics we aim to collect, as well as research questions that

dig deeper into a specific metric or explore a particular phenomenon in detail.

### 4.2.1 Exploratory Metrics

Upon collecting a small portion of about 1000 ads using our prototype crawler, we browsed through their images and qualitatively identified some characteristics that could be fertile material for analysis and also reasonably be extracted given the current state of computer vision technology. We then developed a series of metrics in the form of queries that we could answer with our automated extraction process. The primary intent of these metrics is to use extracted information to set the stage for future research questions and see ads in new ways that are less obvious or even invisible at smaller scales. Some metrics and their corresponding ad characteristic(s) are summarized in Table 4.1.

Metrics Query	Characteristic(s)
What are the most frequently used words in web ads?	Text
What is the average number of words in an ad?	Text
What are some common brands and industries advertised?	Brand, Industry
What are some popular colours used in ads?	Colours
What percentage of ads show participation in the AdChoices program?	Disclosures
What percentage of ads enable muting?	Disclosures
What percentage of ads disclose themselves as ads through keywords such as “Ad” or “Sponsored”?	Disclosures
What are some commonly depicted objects?	Objects
What percentage of ads contain human faces?	Faces
Who are the most prominent publishers by number of ads published?	Publisher

Table 4.1: Metrics and their corresponding ad characteristics.

### 4.2.2 Research Questions

In addition to exploring metrics, we can use our dataset to investigate specific areas in detail. Our work in data collection and visual analysis provide us with a promising start point, but the answers to our research questions are not directly observable from our dataset, so we build additional analytics infrastructure to accomplish our goals. More details on implementation and methodology are available in Chapter 5.

There are many areas we could explore, including attention-grabbing techniques, deceptive language and other dark patterns, and depiction of people in ads. It would be impossible to tackle them all in this thesis; instead, I focus on two questions in two areas:

Area	Question
Disclosures	How are text-based ad disclosures represented across our dataset?
Ad distribution	How prevalent are related ads in ads published on the same webpage?

The first question builds off of our keyword disclosure metric and stems from the observation that textual disclosures in self-disclosing ads – ads that disclose themselves as ads – appear in a wide variety of forms. For example, some have “ADVERTISEMENT” written across the top, some have the same text on the bottom or in a corner, and some have an “Ad” badge in the top left corner. Some are more clear while others (intentionally or unintentionally) blend in with the rest of the ad. I investigate this question with the goal of identifying common text disclosure practices and highlighting ones that may appear as deceptive and/or evasive.

The second question surfaced when I was browsing popular news websites with my ad blocker disabled and noticed that many of the ads from the same publisher were related. This could be a tight relation (most or all ads on a page are identical) or a looser relation (most or all ads on a page are from the same industry). The

relatedness of ads on a page triggers a few questions. Previous work noted that repetition boosts the credibility of a claim, even if the claim was not credible to begin with ([50], [80]). Therefore, can related ads use this to bolster their perceived credibility? Are publishers aware of this phenomenon, and are they concerned about how repeated brands and/or industries can influence consumer perception of their website? What kinds of technological mechanisms do ad intermediaries have in place to support this phenomenon, and is it fair to publishers and advertisers? Using the publisher metadata from our archived ads, I pursue this area to investigate how widespread this phenomenon is and what it means for ad agents and regulators alike.

One further reason to pursue these questions now is reliable accuracy. Text extraction is a well-studied task in computer vision and there is a high probability that text disclosures can be automatically and accurately identified without advanced tools. Publisher information can be gathered on each crawl by noting the website from which the ad was captured from, leaving no room for mistakes. Optimizing for accuracy now also means that we can perform these analyses in tandem with improving our algorithms for some of the more advanced automated tasks such as brand detection. This will prepare us and other researchers to address a broader range of questions in the future.

## 4.3 Policy Implications

As mentioned in the previous chapter, while federal and local governments have regulatory infrastructure set up to penalize advertising violators, the system at large still depends heavily on industry self-regulation. However, many attempts at self-regulation have failed blatantly and somewhat comically. For example, by 2014, companies behind leading advertising platforms such as Google, Facebook, and Twitter had formed the TrustInAds.org consortium to improve practices for combatting prob-

lematic ads [56]. At the time of writing, no activity from them can be found besides a singular blog post with a cursory outline of how online advertising works, while their website still looks incomplete and is littered with filler text [92]. Since federal agencies such as the FTC simply do not have enough resources to keep a close eye on the vast expanse of web ads, they often rely on research and expertise from academic communities to guide policy decisions [51]. After all, such communities tend to critically examine ads from a wide range of perspectives, from their technical roots to social effects, using empirical methods, and advertising laws and regulations should ideally be shaped by empirical observations.

Our ad search tool has the potential to bring those observations closer to regulators and change the way by which they monitor ads. Many ads evade attention due to their automated distribution by ad intermediaries and therefore unpredictability in where and when they appear across the web. By capturing snapshots of this ever-changing space, regulators are not only provided with an elevated platform to better observe ads, but also opportunities to trace paths of problematic ads to relevant publishers, advertisers, and intermediaries. Of course, that is not to say regulators will reduce their collaboration with academic communities upon using this tool. In fact, it may very well be the opposite – now that new insights can be derived from a large corpus of real-world ads, federal and local agencies can engage in more fruitful discussions with researchers and other interested parties on how those insights can impact new policy decisions.

In light of these possibilities, I propose a three-fold policy recommendation for curbing potentially harmful web ads in the wild. I say three-fold because it would be a futile effort to pin down that responsibility on a particular ad agent or legal entity. No, the responsibility should be shared across numerous relevant groups. The fact that billions of people are browsing the internet and potentially consuming web ads confers critical obligations upon advertisers and publishers to output responsible ads.

Additionally, there is more pressure than ever on regulators to monitor this sphere as increasing numbers of new internet users are children and older adults, both widely considered as vulnerable populations on the web due to their lack of experience with technology. Given the market power and reach of ad intermediaries, they should no longer be treated as blackboxes that silently and blindly facilitate ad transactions. We expect stock exchanges to publicly disclose trades, list relevant buyers and sellers, and provide data visualizations of market activity, so why should that be different for ad intermediaries?

The three pillars in the policy recommendations I propose engage three different groups of individuals in the ad market: consumers, intermediaries, and publishers. The following subsections briefly explain how the recommendations interact with the various groups to achieve desirable outcomes.

### **4.3.1 Consumers**

With ad agents occupied in their profit-centric ad market game and regulators unable to bring the entire industry under their watchful supervision, consumers can be a valuable resource for bridging this asymmetry. Zeng et al. [24] discussed the gap between acceptable ads according to consumers and policies of the online ad ecosystem, and how the current system fails to incentivize ad tech to prioritize quality ad experiences. Incorporating consumer feedback can be a powerful mechanism to keep ad agents in check. There are no better real-time, in situ observers of ads than consumers themselves, and passing more regulatory power to them can help exert pressure on ad agents to maintain a higher quality standard. Moreover, the scale and reach of modern ad systems is immensely advantageous when crowdsourcing feedback.

How can consumers acquire more regulatory power? I suggest that regulators oversee the implementation of an industry-wide, standardized reporting system for ads. Similar systems can already be seen on social media platforms, but with one key



difference: the system serves the platform, not regulators. Once an ad is reported by a platform user, the platform decides whether or not to remove the ad, an action that can be biased by profit motives. This essentially brings matters back to square one – platforms still have no real incentive to provide better ad experiences. All of this is considered while putting aside previous records of platforms not taking their reported content seriously [59]. A system overseen by and connects directly with regulators can help tackle this issue head-on. Each ad report should automatically sent to a regulatory entity with a note of the ad’s content, publisher, and if possible, other layers of the programmatic ad ecosystem such as DSP and SSP. With the arrival of more reports, regulators would have a running tally of ad agents associated with frequently reported ads and can reach out to relevant agents for legal action directly.

One catch of this recommendation is that it must be supplemented with a robust pipeline through which ad reports can reach regulators. Just like ad reporting on social media, the recommended system cannot function effectively if reports are left unhandled. A technical system should be built to mitigate this by storing report data and notifying regulators when a new report is submitted. On the receiving end, a team of individuals, either within the FTC or working closely with the FTC, should be trained to monitor the incoming reports and take action as necessary.

### **4.3.2 Intermediaries**

In their current state, ad intermediaries are privately owned entities revealing very little about their practices to the outside world. They undeniably facilitate transactions of problematic ads, but due to their blackboxed nature, it is difficult to pin down their exact actions (or lack thereof) that contribute to the pollution of the ad ecosystem. Some intermediaries are considered “premium” and serve higher-quality traffic [64]; they are more likely to mediate higher quality ads. Some deal specifically with undesirable real estate at the bottom of web pages and are more likely to arbitrate lower

quality ads [39]. It would be unjustified, however, to order certain intermediaries to shut down, partly because there is no proof that they *only* serve problematic ads and partly because many websites who do not acquire premium traffic rely on ad revenue to sustain themselves.

Therefore, I recommend that regulators require ad intermediaries to disclose their activities to all layers of the ad ecosystem. Intermediary activity should ideally be viewed and browsed in a similar fashion as public intermediaries such as stock exchanges and blockchains. The information does not necessarily have to be made available to the general public since unlike the stock market and cryptocurrencies, the average person is not considered a key market player. However, ad agents with stake in the market such as advertisers and publishers should be able to transparently view market activity through, for example, a dashboard with visualizations created by the exchange and/or create their own upon requesting information via an API. Increasing transparency can also help reduce issues indirectly related to ad quality, such as ad fraud and market inefficiencies resulting from untruthful bidding [67].

It is worthwhile to note that this recommendation should not evoke Section 230, as transparency does not equate to liability. Returning to the stock market example, stock exchanges are responsible for ensuring fair pricing practices and transparency in transactions, but are not liable for arbitrating what sorts of trades can cross the trading floor. The same analogy can be applied to ad intermediaries. Even though additional transparency does not directly reduce the number of problematic ads, it can indirectly help by allowing ads to be traced from one agent to another. This way, agents and regulators can better pinpoint the source of the problem.

### **4.3.3 Publishers**

While advertisers dictate content within ads and intermediaries determine which ads will be delivered, publishers are the ultimate gatekeepers between ads and consumers.

An ad will not be shown publicly if a publisher decides to block it, a choice independent of actions from advertisers or intermediaries.

Problematic characteristics in ads have been defined previously by the FTC [28]. While identifying ads with these characteristics is one thing, removing them is another altogether. Highly reputed publishers such as The New York Times may have revenue channels outside of advertising, such as subscriptions, that provide financial freedom for them to run a higher-quality selection of ads. Less privileged publishers may not have the ability to do this and they also occupy a larger portion of the web. Consequently, policies that target such publishers can be crucial to minimizing the distribution of problematic ads.

I recommend that publishers use AdOculos and FTC’s guidelines to publish a set of “community guidelines” for ads displayed on their platforms. They can use these guidelines to identify problematic ads they wish to remove and present evidence of violations to regulators or other publishers with similar guidelines, whom they can then work with to pressure relevant agents to take appropriate actions. Escalating issues directly to the advertiser or intermediary has seen success in the past [29] and did not result in additional financial concerns for the publisher. Regulators also receive corollary benefit as they are prompted to review and update their guidelines through this process.

Note that this policy should not evoke Section 230 because it is an issue of choice over one of liability. If an ad’s messaging conflicts with the publisher’s guidelines, the publisher should be able to remove it just like how Twitter or Facebook bans content that violates its community guidelines. If the publisher decides that no guidelines are necessary, nothing can be done without Section 230 reform.

# Chapter 5

## Implementation

Chapter 4 motivated the development of a web ads dataset interfaced with a search tool. It also highlighted the central research questions in this thesis along with potential policy implications. In this chapter, I describe how the proposed work of myself and my collaborators came into fruition. I dive into the technical implementation of our ad collector, visual analysis tools, and search interface, along with challenges associated with each. I also outline techniques I use to tackle each of my research questions.

Collecting and analyzing ads are remarkably broad tasks. Before we even started automating the two processes, we conducted a manual feasibility study to help narrow down our scope and goals. In September 2020, we retrieved the top 200 websites from Tranco [57], a research-oriented list of 1 million popular websites that uses a ranking strategy that improves upon the shortcomings of current lists, such as Alexa [1]. For each of these websites, we opened an instance of the Chrome browser in Incognito Mode while using a university VPN service to set the browsing location to Princeton, NJ. We turned off Chrome’s third-party cookie blocking to allow for a wider range of ads, visited and scrolled through the entirety of each website’s landing page, and took a screenshot of every ad we saw.

Twenty-five of the sites led to broken links, so we manually browsed 175 sites in total, collecting 127 ads. Through this process, we found that few websites had ads on their landing pages (38/175). Many of the top sites were platform products that required sign-in, and users would presumptively be fed platform-specific ads once on the platform. While these ads are ripe for study, they are notoriously difficult to collect in an automated way across many platforms due to the need for account creation and user verification. The exposure of these ads are also limited to certain user groups on the internet and may not be ideal for our general-purpose dataset. The combination of these two factors steer our focus to platform-agnostic ads, which we found on most news and media sites. We also surveyed the possibilities of visual analysis by uploading our collected ads to Figma, a collaborative design tool that allowed us to view, discuss, and cluster ads together. Figure 5.1 shows an early exploration of clustering by “information completeness” in an ad based on text and graphics.

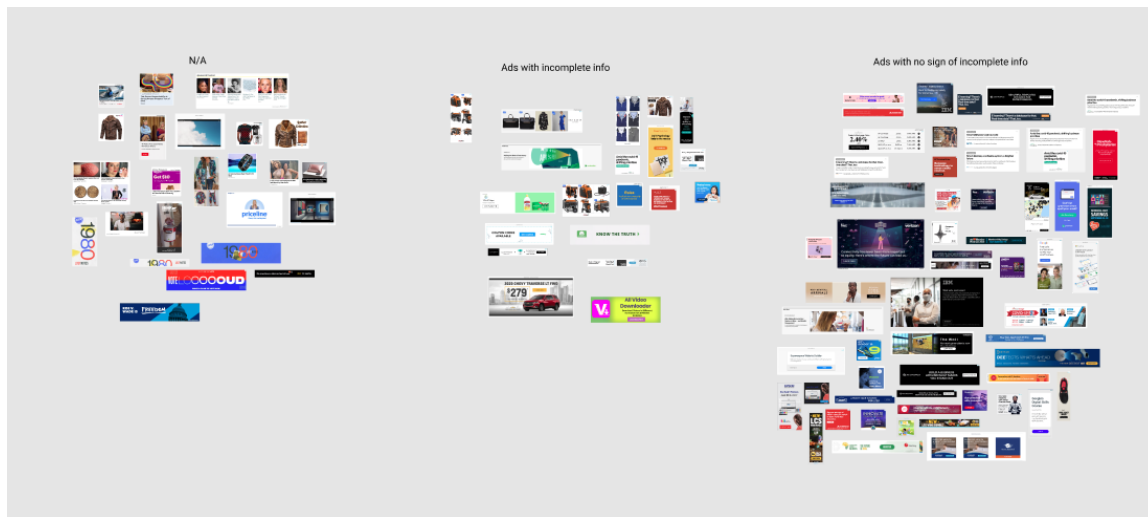


Figure 5.1: Manual analysis and clustering by information completeness.

Upon better grasping our possibilities and limitations, we took an iterative three-step approach to implementing our system. First, we built a web crawler prototype to collect a small sample of ads. We then used the ads to develop and test our visual

analysis tools, while simultaneously improving our ad collector. With the ads and their extracted data, we built AdOculos, our search interface, and populated it with processed ads. By interacting with early versions of AdOculos, we identified areas of improvement, such as new analysis features we think would be helpful and additional metadata to be collected on ad crawls, and iterated on the previous steps accordingly. Throughout all of this, we noted observations and formulated research questions.

## 5.1 Ad Collector

Our ad collector is a web crawler built on OpenWPM [23], an open source and scalable web privacy measurement framework. OpenWPM runs in the Firefox browser, with automation provided by Selenium [81]. We use OpenWPM over other frameworks used in previous studies such as Puppeteer [24] due to its ability to handle large-scale crawls, many included hooks for collection of browsing data such as cookies and JavaScript calls, and ease of integrating browser extensions. A collaborator who is familiar with OpenWPM from previous projects was the primary engineer behind this ad collector, so I will be providing a high-level overview of its functionality and implementation.

### 5.1.1 Functional Overview and Implementation

The foundational idea behind our ad collector is for it to accept a list of websites and for each website:

1. visit the website and locate all ads in the DOM tree
2. capture a screenshot of each ad via its DOM element along with the relevant ad metadata
3. repeat for the rest of the websites

We run the crawls from two machines set up in the Center for Information Technology Policy (CITP). For each crawl, a new, unprofiled instance of Firefox is spun up in a new Docker container with the browsing location set to Princeton University. As we are not yet investigating the effect of profiling or web tracking on ad targeting, this ensures that no browsing data or cookies will influence the ads our crawler sees. Once the crawler visits a page, it waits for  $t_0$  seconds for the page and its ads to load. Through trial and error, we found  $t_0 = 35$  to work well.

To identify ads in the DOM tree, we could match page elements on Easylist [19], a popular list of selectors and domains used by many adblockers such as Adblock Plus. However, this can result in performance issues as the list is long and requires our own parsing pipeline. Instead, we repurpose the Cliqz adblocker extension [10], which already has a built-in parser, for our own use. We use the extension’s standalone JavaScript library to integrate it into our crawler. The library was edited such that upon identifying an ad, the extension takes a screenshot of the element instead of removing it from display.

Once we identify the location of the ad in the DOM, we scroll to it and wait for  $t_1$  seconds for the ad to load. Waiting too little can result in capturing an ad that has not yet fully loaded, and waiting too long can witness a new ad being loaded in and replacing the original. Through trial and error, we found  $t_1 = 4$  to strike a good balance between the two.

In taking the screenshot, we found that simply capturing matching elements wasn’t satisfactory. Areas of the DOM where ads reside are often obfuscated and the structure is unpredictable, so there could be multiple nested matching elements and automatically picking the right one is challenging work. To navigate this, we use Firefox’s `captureVisibleTab()` function from their JavaScript API [65] to screenshot the entire viewport – that is, the entire browser window in view. We then note the pixel coordinates that bound the outermost matching DOM element, which we infer to be

the parent of the ad, and crop the viewport screenshot to those pixel coordinates.

We store our ad screenshot, encoded as a base-64 data URL, in a LevelDB database along with collected metadata in a SQLite database. Table 5.1 elaborates on the metadata fields collected with each ad.

Metadata Field	Description
Publisher	Domain of the website from which the ad was captured
Time	Time at which the ad was captured (Eastern time), in yyyy-mm-dd hh:mm:ss format
Location	Browser location at which the ad was captured
ID	Hash of the ad image
HTML	The ad's parent HTML element from the DOM

Table 5.1: Collected ad metadata fields.

### 5.1.2 Limitations

Our ad collector successfully captures most ads, but it is not perfect. While we identified values for  $t_0$  and  $t_1$  that seemed optimal to us, there are still a limited number of collected ads that have yet to fully load or even loaded at all. Furthermore, some captured ads have been occluded by interstitials. Many websites, upon visit, will display a popup interstitial element containing a CTA, such as cookie banners and sign-up forms. Before being dismissed, these elements cast a dark overlay on the rest of the website content. Since the crawler captures the entire viewport, ads captured on the page in the presence of an interstitial will have drastically lower visibility. To address this, our crawler would need to automatically identify and dismiss any interstitials on the website before beginning the collection process.



Perhaps the largest limitation in our current crawler is the inability to capture dynamic ads. Many ads are more than a static image – they may be GIFs or videos with moving elements, or contain interactive UI elements that respond to mouse hovers, drags, and clicks. By taking a screenshot of the ad, we are capturing it at a single point in time and potentially leaving out some relevant messaging. For example, if the screenshot is taken right when all text has moved off of the ad, we would be missing one important vector of investigation. Being unable to identify interactive ads also limits us from investigating the research question of deceptive UIs: if an ad boasts the appearance of an interactive UI such as a game or calculator, is it actually interactive or is it just a hyperlinked image that takes users to a landing page when they try to interact with it? While interactive ads may be harder to identify, one proposed solution to capture dynamic ads is to take multiple screenshots, compare the “frames” for similarity, and string the frames together into a GIF if they are not identical.

## **5.2 Ad Search**

Enabling searchable ads allows others to dig deep into our dataset and generate insights for their own research. To do this, two main components are established: a search and filtering interface (AdOculos) and a set of visual analysis tools that exposes the search and filter fields. The former delivers the user experience while the latter makes the entire experience possible. As the primary engineer behind both components, I discuss each in detail.

### **5.2.1 Functional Overview and Implementation**

Once users land on AdOculos’s homepage, they are greeted with a search bar, drop-down menu buttons for filters, and a gallery populated with card UIs containing ads.

The default query, upon page load, is a query for all ads in the dataset, sorted by date collected in descending order. Users can then *search* for ads with in-ad text queries using the search bar. This search leaves slight room for fuzziness, meaning that if a user queries on multiple words, ads with exact matches as well as partial matches will be displayed as results. The results are ranked by relevance. For example, if a user queried on the term “hot dog”, an ad with two mentions of “hot dog” will be displayed first, followed by an ad with just one mention, followed by an ad with “dog walker”.

Users can *filter* on 10 different fields, which are summarized in Table 5.2 along with other relevant information such as example values.

Upon entering a search query or setting a filter, an informative tag will appear under the search bar indicating what search term(s) or filter(s) have been applied. This is for users to keep track of their search and filter combinations, and also provides an easy way for them to clear all previous activity with one click.

In the gallery view, results are displayed on card UIs. Each card contains an image of the ad (which can be viewed in its original size in a new window when a user clicks on it), 3 pieces of metadata (publisher, time captured, location) and an Ad Details button that will take the user to a new page displaying all of the ad’s extracted data. The data is summarized in Table 5.2.

I use Elasticsearch [21] as the data storage powering AdOculos. Elasticsearch is a NoSQL, JSON-based information retrieval system based on the Apache Lucene library capable of distributed, full-text search. It supports near real-time search on structured and unstructured text, numerical data, and geospatial data. Compared to SQL databases or other NoSQL options such as MongoDB, Elasticsearch is ideal for not only its optimized search speeds, but also resilience to scale: its distributed nature allows for automatic data re-organization as the dataset grows larger, which inevitably happens in our large-scale crawls. Elasticsearch can be used locally, but

Table 5.2: Summary of data associated with each analyzed ad.

Field	Description	Example Values	JSON Field(s)	Data Type
Text	All text detected in the ad, organized by paragraphs	“Teach your next lesson live with Vimeo” “vimeo” “Go live”	text	string array
Brands	Any brands detected in the ad	Adidas, State Farm, Paramount+	brands	string array
Industries	Industries associated with detected brands (if any)	retail, news & media, insurance	industries	string array
Objects	Any objects detected in the ad	car, outerwear, bicycle	objects	string array
Faces	Whether or not a human face was detected	yes, no	faces	boolean
Disclosures	Whether various forms of disclosures, such as icons or indication of fine print, was detected	AdChoices icon , Mute icon, Ad indicator	adchoices_exists mute_exists text_disclosures asterisk_terms	boolean object
Colour	Colour palette for the ad	blue, lime, fuchsia	dominant_colors_rgb dominant_color_names	string array (filter) / tuple array (ad details)
Size class	Size classification of the ad	billboard, skyscraper, half-page	size_class	string

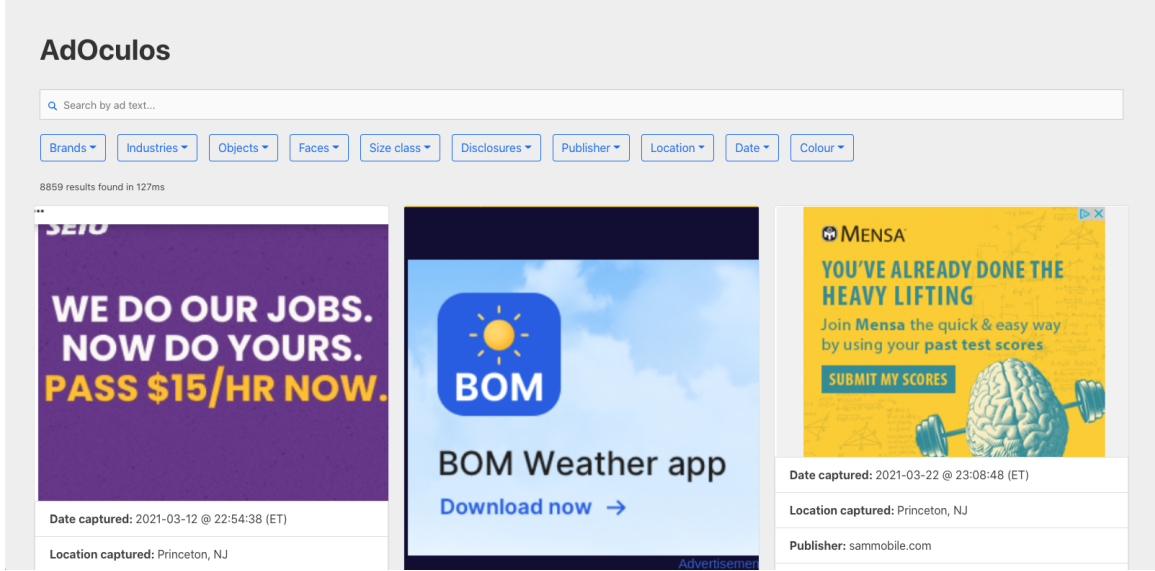


Figure 5.2: Screenshot of the AdOculus homepage.

in order to make our data accessible to the open web, I deployed an instance on Microsoft Azure through Elastic Cloud [22]. I also experimented with deploying on Amazon Web Services; however, the lack of key deployment configurations available on the platform called for the use of a proxy server in order for frontend code to properly communicate with it, which lead to undesirable complications in frontend hosting.

AdOculus’s frontend was built with React. With the help of ReactiveSearch [3], an Elasticsearch-compatible React UI library, I wrote UI components that accept user input, communicate with our Elasticsearch cluster to retrieve requested information via REST API calls, and then display the information in the gallery. I also integrated React-Bootstrap [79] for responsiveness. Our hosted Elasticsearch eliminates the need for our own server, which simplifies the deployment and maintenance process. The frontend is deployed as a React application on Netlify [69], a serverless web application hosting platform, at the time of writing.

Before I even started on AdOculus, however, I built an *ad browser* on the same technological stack to 1) get acquainted with the technical tools and uncover any

limitations, and 2) provide the project team with a convenient way to view all ads collected by our crawler. This was an important checkpoint for us because we could identify potential research questions to pursue by browsing the ads as well as view and debug errors made by our crawler. Since storage space was not a prominent concern, I partitioned the Elasticsearch cluster to dedicate an index (Elasticsearch’s equivalent of a database) with a simple schema for the browser and another with a more complex schema for AdOculos. The benefits of this are twofold. Having the browser index serve as a “waiting room” before the analysis script processes the data and inserts it into the search index meant that the crawler could directly upload data from its local SQL and LevelDB databases to the cloud without worrying about tainting the search index. This was particularly helpful during our early experimental crawls when many collected images were low-resolution, occluded, or badly cropped. We often removed such data or even cleared out the entire index in favour of a fresh crawl before running the analysis script. Additionally, if something goes wrong in our search index, we would have another index that can quickly support it as a hot backup. Eliminating the need to request another upload from the crawling machines’ local storage units can significantly reduce recovery time.

The crawler, Elasticsearch cluster, frontend web applications, and worker scripts all work together to harvest ads from the web to the display case that is AdOculos. A comprehensive view of the entire system can be found in Figure 5.3.

### 5.2.2 Visual Analysis

Ad search would not be possible without the data extracted by our automated visual analysis pipeline. In this subsection, I describe the implementation of each of the analysis capabilities summarized in Table 5.2, plus some additional ones used internally by the system. All are implemented in Python.

Due to the diversity of visual analysis tasks, training my own custom models

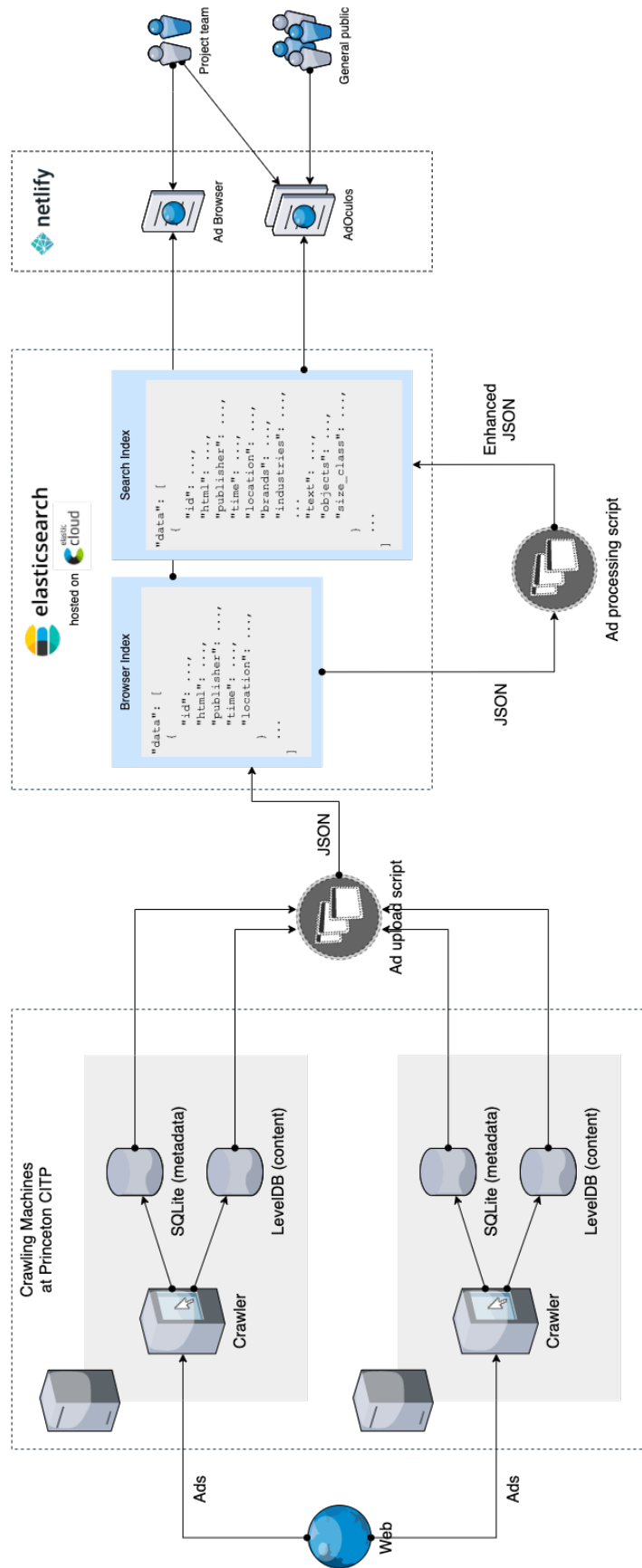


Figure 5.3: Overview of our ad collection and archival system.

for each of the tasks would be immensely time and resource consuming. Even if I manage to do so, the models' accuracies may not be as high as those of some off-the-shelf tools. Moreover, many of these tasks, such as Optical Character Recognition (OCR), are well-trodden areas in computer vision and come with usable tools right out of the box. With this in mind, I use Google's Cloud Vision API [33] to assist in some of the tasks. I selected the Vision API after some experimentation with other cloud APIs (Microsoft Azure Computer Vision service, Clarifai) and found that it offered a feature set better aligned with my goals and higher accuracy in many of those features. The role of the Vision API varies depending on the task. Some do not use the API at all, either because there is no relevant feature in the API or because I write code that accomplishes the task at least equally as well. Some use the API as an initial step, but still rely mainly on custom algorithms. Some see great performance from the API and require little custom code besides some data formatting.

Besides convenience, using a cloud API is an opportunity to survey the usability of modern cloud-based computer vision tools. Significant attention has been devoted to advancements in the field in recent years, particularly from large companies behind many of the APIs such as Google and Microsoft. This translates to higher potential for improvement in these tools over time. If the tools are already robust enough to successfully perform exploratory visual analysis of ads without custom model training and tuning, they can lower the barrier for others who have similar goals. Reflecting on my experience with the Vision API, I discuss the future potential of off-the-shelf computer vision tools in Chapter 6.

## **Text Extraction**

I extract all text in an ad, segmented by paragraphs, using OCR. The Vision API has two OCR features: one for text-dense images such as PDF documents and handwritten notes, and another for sparse text on larger surfaces such as highway signs. I

chose the former due to its superior performance during preliminary testing. While many ads are not packed with text in the same way PDF documents are, the feature optimized for documents was far more successful at identifying smaller text in many ads. This is crucial to our ability to extract and analyze ad mouseprint (fine print describing terms and conditions and other legal details).

The API returns symbol objects with content and bounding box coordinates for each individual character. These objects are deeply nested inside other inter-nested structural objects, including words, paragraphs, blocks, and pages. Symbols are also packaged with symbol properties, with one particularly handy property: break type. To obtain an accurate representation of the text structure in the ad instead of returning a simple text chunk, I use the type of break detected after the symbol (space, newline, etc.) to string together characters into words, and then words into lines. Finally, I use the detected paragraphs' bounding boxes to organize lines into paragraphs. I return the final text extraction result as an array of paragraphs.

## Text-based Disclosure Detection

Now that I have access to all text in the ad, I use it to perform text analysis and identify two types of disclosures: mouseprint indication and self-disclosure. Figures 5.4 and 5.5 show examples of mouseprint indication and self-disclosure, respectively. Note that although the Ram truck ad in 5.4 suggests presence of mouseprint, it does not actually show any.



Figure 5.4: Examples of ads with mouseprint indication.



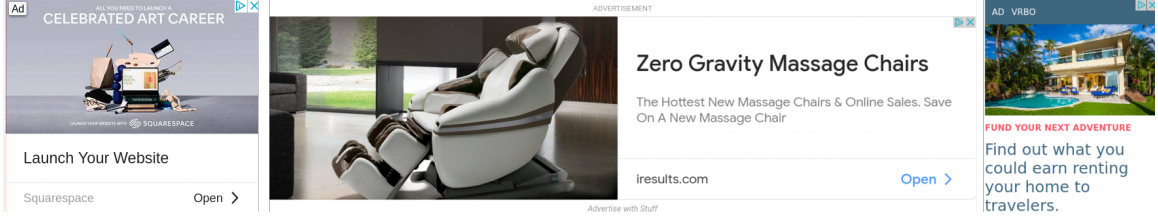


Figure 5.5: Examples of self-disclosing ads.

The presence of an asterisk (\*) in ad text can be used to indicate the presence of mouseprint. However, a naive approach that looks for the asterisk character across all text may run into some false positives with advertised brands such as E\*Trade and other stylized text that uses asterisk(s) in the middle of the word. Therefore, I first look for the asterisk character in a paragraph, and if it is found, I break the paragraph into words and only set the result to true if the asterisk is the last character in the word, or the word consists of one character and is itself the asterisk. I note that the Latin cross character is commonly used as a mouseprint indicator but is often interpreted as an asterisk by Google’s OCR. Besides asterisks, I also look for the presence of key words “terms” or “conditions” accompanying “apply” to catch any mouseprint not disclosed with any particular characters. I take a similar keyword matching approach to detect whether or not an ad is self-disclosing. I match on entire words such as “ad”, “advertisement”, and “sponsored”, as well as a combination of words, such as (“paid” && “content”) or (“paid” && “post”).

### Image-based Disclosure Detection

This task is driven by the desire to detect the AdChoices and Mute icons in our ads. The work of Storey et al. [87] in perceptive ad blocking used a *template matching* technique to classify whether or not an ad contained an AdChoices icon. Template matching locates a template image (AdChoices logo) within a larger input image (ad) by sliding the template across the input and comparing it with the patch of input under it. The comparison can be done using a variety of methods, such as image

hashing and sum of squared differences in pixels. A key drawback to conventional template matching is that it is not background nor size agnostic: two identical icons placed on different backgrounds and/or are of different sizes can still have a low similarity score between them. This is especially problematic since AdChoices and Mute icons can appear on varying backgrounds despite creative guidelines [16]. For example, the DAA condemns icon applications such as the ones shown in Figure 5.6(c), and yet they are still found in the wild. The size of the icons can also vary slightly from ad to ad.

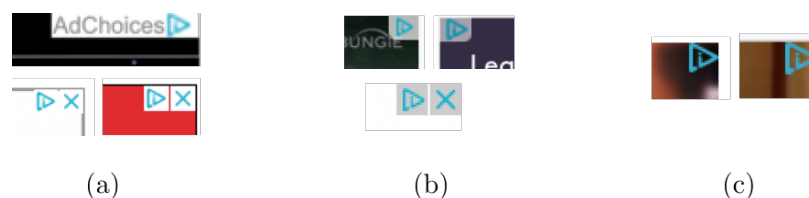


Figure 5.6: AdChoices and/or Mute icons found on collected ads.

Another approach I considered was to train a classifier to detect the icons. A custom model can potentially achieve higher accuracy than template matching if tuned correctly; however, there was, to my knowledge, no dataset for this task when I started on it. Google’s AutoML model trainer recommends at least 1000 images per label to train a classifier with reasonable performance [31]. Time-intensive manual ad collection and labelling was my best option since our crawler was not ready to collect ads yet. Moreover, multiclass classification problems such as this one generally demand more resources than single-class ones such as the one tackled by Storey et al . For example, if we were to build an app that detects whether or not something is or is not a hot dog, it would require one model and a dataset with two labels. Its functionality may be unimpressive compared to an app that can distinguish between hot dogs, pizza, and spaghetti. However, the latter app will require three models (one for each food), more image labelling, and most likely more images to train on. Similarly, detecting whether the AdChoices icon, Mute icon, or both exist in an ad is

a heftier task than locating just the AdChoices icon for ad blocking purposes.

With this in mind, I implemented a background-agnostic, multi-scale template matching algorithm. My strategy takes advantage of the constraints in icon style and placement by the DAA [16], specifically the limited colour palette, placement location, size range, and lack of distortion. Since the icons can only be placed in the corners of an ad, I start by cropping 60px by 24px rectangles from the corners and use OpenCV to apply a colour threshold such that only blue to teal pixels are preserved. I am assured that the edge of the image is at least somewhat close to the edge of the ad because I run a function to crop out any padding space in the ad image before starting any analysis. Then, I eliminate any rectangles that do not have leftover pixels and apply a mask to the remaining ones. This places the thresholded pixels against a black background, hence the background-agnostic nature of this algorithm. I crop the result to the non-black pixels and discard it if it is smaller than 10px by 10px to finish pre-processing.

This result, which I will call  $I_0$ , not only increases the accuracy of template matching, but also its performance, as the algorithm only needs to run along relevant areas of the image. This is especially important in my multi-scale implementation.  $N = 10$  rounds of template matching are performed, with the size of the template image starting out at scale = 1 and shrinking by a constant ratio with every round until scale = 0.5. Each round yields an array of similarity scores, one for each sliding window position. Each score is computed using a normalized Pearson’s correlation coefficient [52], given by

$$R(x, y) = \frac{\sum_{x', y'} T(x', y') \cdot I(x + x', y + y')}{\sqrt{\sum_{x', y'} T(x', y')^2 \cdot \sum_{x', y'} I(x + x', y + y')^2}}$$

where  $(x, y)$  are the upper left-hand coordinates of the sliding window;  $x' = 0, \dots, w - 1, y' = 0, \dots, h - 1$  are coordinates within a window sized  $(w, h)$ ;  $T$  is the mean-shifted

template (the image's mean pixel value is subtracted from every pixel); and  $I$  is the mean-shifted input image. The score of the best matching template across all sizes and rounds can then be retrieved with a two-fold maximization across the two vectors:

$$R_{best} = \max_{n=1\dots N} \left( \max_{\substack{x=1\dots X \\ y=1\dots Y}} R_n(x, y) \right)$$

where  $X$  and  $Y$  are the width and height of  $I_0$ , respectively. I run two rounds of multi-scale template matching on  $I_0$ , once with the AdChoices icon as the template and once with the Mute icon. To determine whether an icon exists based on the score, I apply thresholds  $t_a$  and  $t_m$  on the  $R_{best}$  values of the AdChoices and Mute templates, respectively. The icon is classified to exist in the ad if its  $R_{best}$  exceeds the threshold and not otherwise. If both icons'  $R_{best}$  exceed their thresholds, I estimate whether one or two icons can fit in  $I_0$  using its width to height ratio. If it is one icon, I take the icon that further exceeds its threshold. Based on experimentation and testing, I set  $t_a = 0.45$  and  $t_m = 0.60$  for best results.

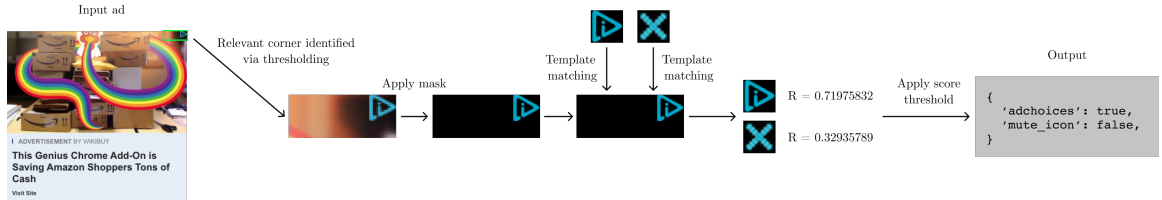


Figure 5.7: A visual example of the icon detection algorithm.

## Image Hashing

I store the image hash of the ad to enable duplicate ad detection as well as quantifying visual similarity through a hash distance. Many hashing algorithms take a similar approach by first compressing the image into an 8px by 8px grayscale (single-channel) image and performing computations on each pixel to assign it a binary value. The 64-bit result is the algorithm's output. The pixel computation method of three of the

most commonly used algorithms implemented by ImageHash Python library [6] are as follows:

- **average hashing (ahash)**: 1 if the pixel value is greater than mean pixel value of the entire image and 0 otherwise.
- **perceptual hashing (phash)**: similar to average hashing but first uses a discrete cosine transform (DCT) and compares based on colour frequencies rather than their values.
- **difference hashing (dhash)**: shrinks image to 9 by 8 instead of 8 by 8. For each pixel not on the right edge, set the pixel to 1 if the pixel value on the right of it is greater and 0 otherwise.

A comparison of hash Hamming distances, or the number of bits that differ between two hash outputs, from images in Figure 5.8 is displayed in Table 5.3.

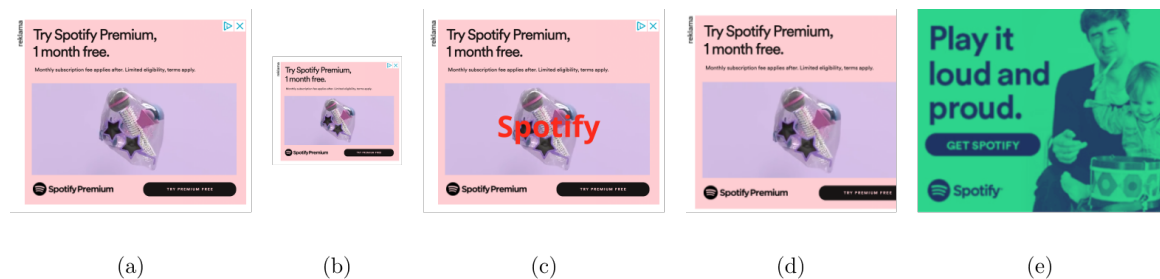


Figure 5.8: Ads for hashing comparison.

I chose **average hash** as my hashing algorithm due to its ability to maintain small distances between identical ads that have been slightly modified with various transformations. In particular, (d) can result due to a miscrop by the crawler, and the small distance computed by ahash (8 vs. 18 vs. 22) allows that ad to be more closely associated to a properly cropped version.

Hash	Image Comparison	Distance
ahash	(a) and (b)	0
	(a) and (c)	1
	(a) and (d)	8
	(a) and (e)	25
phash	(a) and (b)	0
	(a) and (c)	4
	(a) and (d)	18
	(a) and (e)	22
dhash	(a) and (b)	1
	(a) and (c)	3
	(a) and (d)	22
	(a) and (e)	34

Table 5.3: Comparison of differences in common hashing algorithms.

## Brand and Industry Detection

Automatically extracting the brand and its corresponding industry from an ad is a challenging task due to its breadth: with so many advertised brands, how can recognition be robust? An instinctive first approach may be to train a classifier on brand logos, but a similar data bottleneck as the one encountered in image-based disclosure detection also applies here. I say similar because brand detection datasets do exist ([48], [88], [97]), but carry multiple concerns. Most have too few logo classes ( $< 500$ ) for an acceptable general-purpose detector. The classes may not be representative of brands advertised on the internet because they are collected from primarily print ads or photos of brick-and-mortar stores. Some may be regionally skewed. For example, the work of Jin et al. [48] came out of Alibaba Group and sees a substantial lean towards Chinese brands.

Even if I have the perfect dataset of logos, the classifier may still perform poorly.

This is based on the observation that many ads show *logotypes* (text-based logos) instead of *logomarks* (image/symbol-based logos). Some standardized formats even discard the logo for the brand’s name in non-stylized text. Figure 5.9 shows some ads a graphical logo classifier may have trouble with.

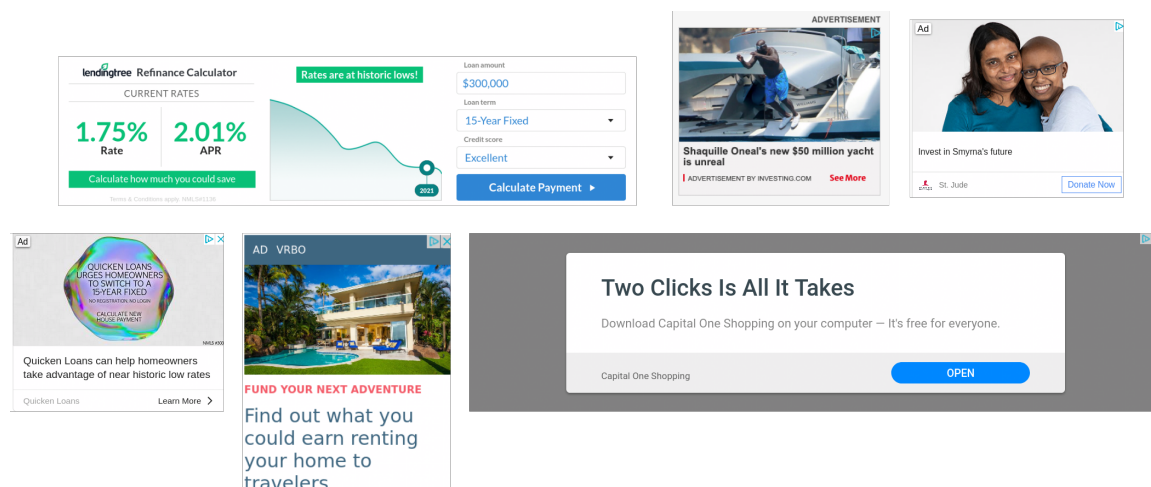


Figure 5.9: Prevalence of text-based brand identities in web ads.

As such, the bulk of my brand detection algorithm is text-based, I use text extracted from the ad and isolate any paragraphs that are likely to contain the name of the brand(s). By observing collected ads, I noticed that the name of the brand usually appeared either by itself or within a very short ( $< 4$  word) text block to draw attention, so I turned my attention to any detected paragraphs that have less than 4 words. I then match words in the paragraphs to a database of around 20,000 brand names and their corresponding industries, using Python’s set intersections for speedy, linear-time performance. I created this dataset from People Data Labs’s 2019 7+ million Global Company Dataset [74], downloaded from Kaggle. Although it contained information on 7+ million organizations and was already sorted by prominence, the data was quite dirty. I cleaned it by removing any organizations whose names contained null values or non-ASCII characters, organizations with invalid websites, and duplicate organization names. I then dropped all organizations after the 20,000th row as most of them were obscure and not based in the US, as well as all columns

except for the organization’s name, industry, and website. The list of industries in the database can be found in Appendix B. If a match in the database was found, the brand and its industry is returned.

Google’s Vision API has a brand detection feature, but it performed poorly in preliminary testing and failed to identify even common brands such as Spotify. It is also heavily prone to misclassifying brands. It is not completely obsolete, though, so I use it as a blind first stab and keep any detected brands returned by the API with a confidence score  $\geq 0.9$ . I then performed an industry lookup for those brands in my database and return the result. Most of the time, however, the API does not detect anything that meets the confidence threshold and I fall back on my own algorithm.

## **Object Detection**

This task primarily relies on the Vision API. Some objects the API detects with lower confidence ( $< 0.7$ ) can often be incorrect, so I filter out any that have a confidence score of less than 0.7. I remove any duplicate objects and return the result.

## **Face Detection**

Like object detection, this task also relies primarily on the API. The API also is capable of estimating the likelihood of 4 emotions (joy, sorrow, surprise, anger) on the face, as well as whether or not the person was wearing headwear. I do not use those features due to accuracy and bias concerns. Choosing to resort the finer details of facial analysis to human annotators, I simply return a boolean indicating whether or not a face was present in an ad.

## **Colour Palette and Colour Name Mapping**

A visually defining characteristic of an ad is its use of colour. Indeed, colour science has been a widely discussed topic in marketing and psychology and has broad appli-



cations in ad design ([83], [58]), thus motivating a colour filter. I first extract RGB colours from an ad and then map the values to colour names for searching.

My initial approach to colour extraction was simple: I determined the pixel frequency for each set of RGB values present in the image, sorted the result by descending frequencies, and took the top  $n$  RGBs. While this worked fine for most ads, it did not return a representative colour sample for ads without solid blocks of colour, such as the ad with a gradient background in Figure 5.10. This called for a new approach: colour palette extraction. I use the colorgram Python package [63] to cluster pixels by their hue, saturation, and lightness values and return the average of each cluster as a member in the palette. This results in a much more accurate colour sample in ads with non-solid backgrounds while retaining accuracy in the rest.

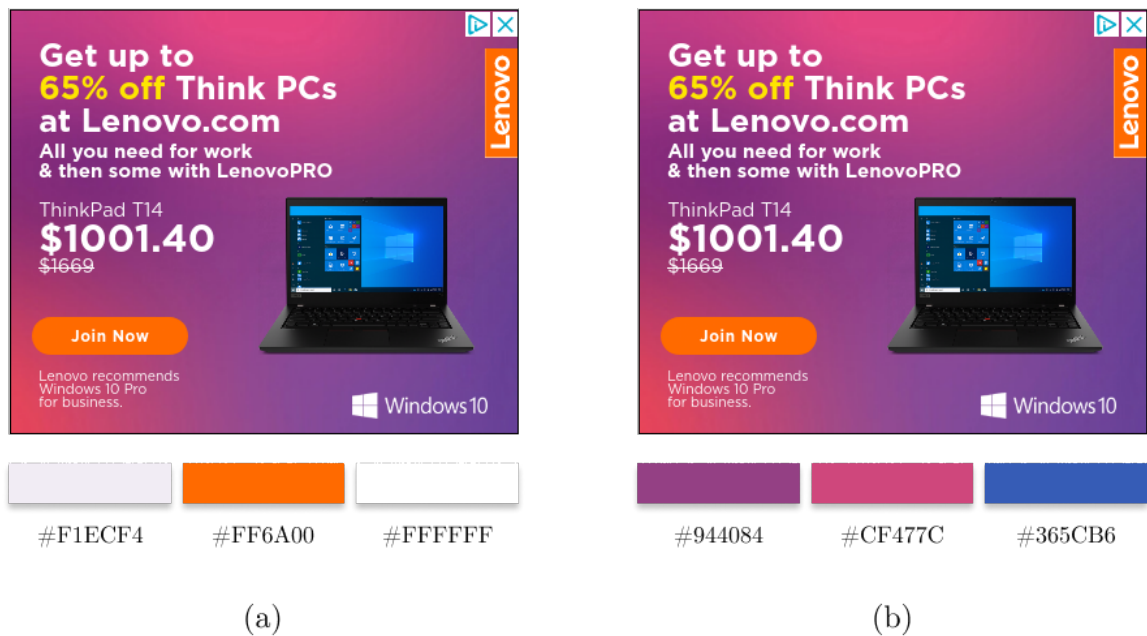


Figure 5.10: Dominant colour extraction (a) vs. palette extraction (b) for  $n = 3$  colours.

Next, to make colours filterable, I map the RGB values to colour names. I first convert them into the Hue-Saturation-Value (HSV) colourspace, which was designed as an alternative to RGB to closer align with the way human vision perceives colours.

For colours in the extracted palette, I minimize the Euclidean distance

$$\sqrt{(h_p - h_w)^2 + (s_p - s_w)^2 + (v_p - v_w)^2}$$

across the 17 web colours outlined in CSS2.1 [96], where  $(h_p, s_p, v_p)$  are the HSV values from the palette<sup>1</sup> and  $(h_w, s_w, v_w)$  are values from the web colour. I use colours from CSS2.1 instead of the modern CSS3 because the ones in CSS3 were too numerous (147 instead of 17) and too granular (names include `lemonchiffon` and `gainsboro`) for a good filtering user experience on the frontend. I retrieve the name of the web colour that corresponds to the minimum distance and assign it to the ad. To make the filtering experience more intuitive and accurate, I only map the top 3 palette colours to names.

### Size Classification

Using the IAB’s ad size specifications and classes outlined in Table 2.1 of Chapter 2, I assign the ad to a size class based on its dimensions. I do so with a bit of additional flexibility due to cropping imperfections in some collected ads.

### 5.2.3 Limitations

The current visual analysis pipeline is robust enough to enable a cohesive ad search experience, but is by no means perfect.

The image-based disclosure detector can sometimes fail to detect icons if the contrast between the icon and its background is too low. Additionally, the template matching thresholds for the AdChoices and Mute icons were tuned based on experimentation and testing, so it is unclear whether they are truly optimal. Some

---

<sup>1</sup>In order for the distance minimization to be effective, I implement a circular variation of  $h_p$ , where  $h_p = 360 - h_{p0}$  if  $h_{p0} > 330$ . This is because hue is measured in degrees from 0 to 360, so  $h = 359$  is much closer to  $h = 0$  (red) than  $h = 300$  (magenta).

parameters within the algorithm can be tweaked, but it quickly becomes a whack-a-mole problem where accuracy improves for a specific type of ad (i.e. ads with icons on gray backgrounds) but decreases for others. A key difference between now and the beginning of this project is that our crawler has collected a dataset of well over 1000 ads on which an icon detection classifier can be trained. It is hard to say for sure that a classifier will yield higher accuracies than the current algorithm, but it is well worth a try.

Brand and industry detection is another task with sizable room for improvement. Some brands in archived ads were not in the database and needed to be manually added. Running separate analyses was then required to update the ad data. The database was (and to a lesser degree, is) littered with incomplete brand names such as “Fitness” and needed to be removed or updated to avoid returning a wrong brand. It may be of interest to run a crowdsourcing task on MTurk or similar platforms where workers look through ads in the ad browser and document any brands not yet in our database before more ads get processed. Further, industry detection is currently dependent on brand detection – if no brand was detected, there would be no industry to lookup in the database. One way to decouple the two would be to train a language classifier and an object detector to estimate an industry based on ad text and graphics, respectively. This way, an industry may be detected regardless of whether a brand is identified.

Lastly, it is important to note that this visual analysis pipeline does not work on GIFs or videos and will need to be updated accordingly for dynamic ads. Perhaps the simplest solution would be to automatically extract the most relevant frame of a dynamic ad (determined by, for example, a frame’s content density) and pass that frame through the current system.

## 5.3 Tackling Research Questions

To dive deeply into the two research questions I outlined in Chapter 4.2.2, I implemented additional scaffolding for data analysis. I relied on automated methods where the task was well-defined and accuracy was consistently high. I employed qualitative methods where I was able to make nuanced connections that may be missed with automation.

### 5.3.1 Self-disclosures

Ads usually declare themselves as ads with key terms, so I used extracted text to look for ads containing the words “ad”, “advertisement”, or “sponsored”, along with combinations of words “paid” & “content” and “paid” & “post”. I took the list and removed visually identical ads based on their hash. I then used the keyword’s bounding box coordinates given by the Vision API to crop the image to those coordinates.

I performed three kinds of analyses on this cropped image. These analyses take inspiration from FTC’s disclosure guidelines for online marketing and advertising [27]. Specifically, they investigate issues of disclosure prominence and placement. I first estimated the text’s font size, in pixels, which is conveniently equal to the height of the text’s em-box (see Figure 5.11). The height of the em-box can be extrapolated from the image’s height through increasing it by approximately 20%, assuming the text is flush against the edges of the image.

Then, I quantified the contrast of the text against its background by computing the image’s Root Mean Squared (RMS) contrast, which is defined as the standard

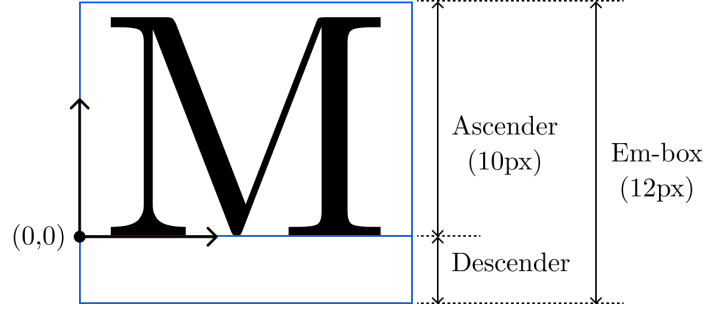


Figure 5.11: An M with a font size of 12px in its em-box.

deviation of pixel intensities<sup>2</sup> and is given by

$$c = \sqrt{\frac{1}{M \cdot N} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} (I_{i,j} - I_{av})^2}$$

where  $I_{i,j}$  is the pixel intensity at  $(i, j)$  in an image of size  $M$  by  $N$ , and  $I_{av}$  is the average pixel intensity across the entire image. Lastly, I computed the  $x, y$  coordinates of the geometric center of the image and expressed that as a percentage of the overall uncropped ad's width and height, respectively, relative to the ad's top left corner. I call this coordinate pair  $(x_r, y_r)$ . Figure 5.13 compares different levels of RMS contrast, while Figure 5.12 provides a visual example of how  $x_r$  and  $y_r$  are computed.

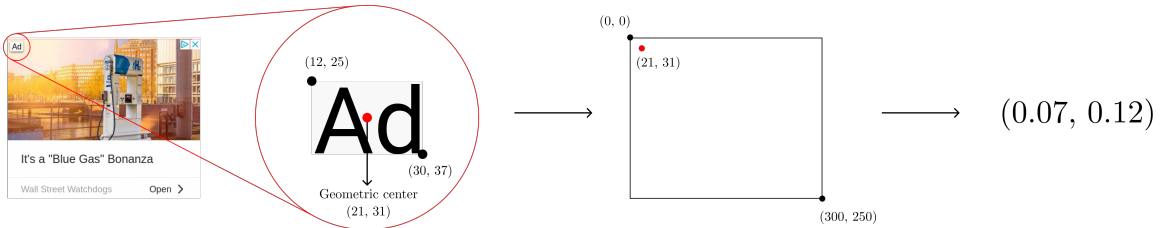


Figure 5.12: Capturing placement of the self-disclosure marker through relative positioning.

The cropped image is encoded into a base-64 data URL and, alongside its analyzed data (bounding box coordinates, font size, RMS contrast, relative position in overall ad), is inserted into an output CSV file.

<sup>2</sup>Pixel intensities are drawn from the single-channel colour values in an image that has been converted to grayscale, normalized in the range  $[0, 1]$ .

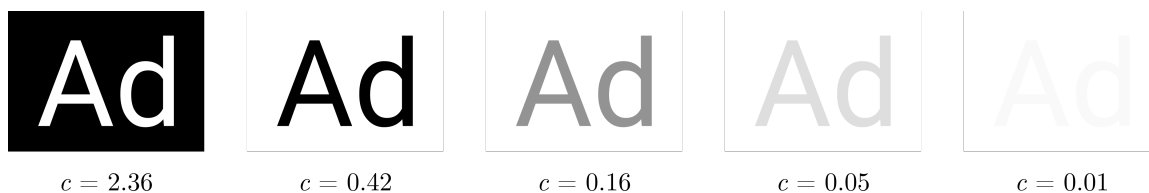


Figure 5.13: Varying levels of RMS contrast.

### 5.3.2 Similar Ads on One Webpage

I tackled this question using qualitative analysis, as some more subtle connections between ads cannot be reliably identified in an automatic way. I used the Publishers filter on AdOculos to identify the 30 top publishers by the number of ads archived. The number of ads associated with each publisher ranged from 29 to 52, with an average of 33.4, and includes websites from non-US but still English-speaking countries. This list of publishers can be found in Appendix C. For each publisher, I evaluated the variables displayed in Table 5.4 using the described methodology.

Variable	Description	Evaluation Methodology
$P$	Domain name of publisher	Retrieve from ad metadata
$n_P$	Number of ads from publisher $P$	Database query via AdOculus
$m_{ind}$	Top # of ads from the same industry	Categorize ads into different industries (according to Appendix B) and return the highest number of ads seen in one industry, or multiple if there is a tie
$m_{same}$	Top # of ads with identical messaging	Count the highest number of ads that advertise the same product from the same brand with the same messaging. They may be identical but can sometimes differ in layout
$m_b$	Top # of ads from same brand	Count the highest number of ads that share the same brand. A superset of $m_{same}$
$b_{ind}$	Binary variable indicating whether the industry/industries corresponding to $m_{ind}$ aligns with the publisher's industry	True if there is alignment, false otherwise

Table 5.4: Variables collected from top 30 publishers in our dataset.

# Chapter 6

## Evaluation, Findings, and Discussion

In this chapter, I step back and analyze my work. To begin, I present the dataset collected and use it to answer metrics posed in Table 4.1. I then evaluate AdOculos by summarizing performance from automated visual analysis as well as feedback we gathered on the tool from other researchers. I discuss our work’s contributions to the future of ad research by outlining some new possibilities and areas of investigation. Finally, I provide examples of those possibilities by presenting findings from my two research questions and walking through their implications on policymaking.

### 6.1 Dataset

As of April 3rd, 2021, there were **8859 ads** archived on AdOculos. This is, to my knowledge, the largest general-purpose, browsable dataset of modern web ads to date. The crawls were performed on the landing pages of **3330 publishers** taken from Tranco’s top 1 million websites [57] over three sessions from March 8 to April 2, 2021: March 8–13 yielded 264 ads, March 23–26 yielded 5079 ads, and March 30–April 2nd yielded 3516 ads. **638 publishers (19.2%)** had at least one ad on



their landing page, with most of the them being news sites, review sites, and blogs.

Using this dataset, I can collect some exploratory metrics by simply referencing the AdOculus interface and occasionally employing some basic data aggregation techniques. The metrics queries from Table 4.1 are listed in Table 6.1 alongside their results. More detailed breakdowns and data visualizations follow the table.

Metric Query	Result
What are the most frequently used words in web ads?	Top 3: “learn”, “now”, “more”
What is the average number of words in an ad?	15.6
What are some common brands and industries advertised?	Top 3 brands: State Farm, Vrbo, Best Buy; top 3 industries (excluding unidentified): internet, financial services, information tech. and services
What are some popular colours used in ads?	Top 3: white, black, navy
What percentage of ads show participation in the AdChoices program?	42.7%
What percentage of ads enable muting?	42.7%
What percentage of ads disclose themselves as ads through keywords such as “Ad” or “Sponsored”?	12.1%
What are some commonly depicted objects?	Top 3: person, packaged goods, top (clothing)
What percentage of ads contain human faces?	18.9%
Who are the most prominent publishers by number of ads published?	Top 3: theadvocate.com, daily-mail.co.uk, mlb.com

Table 6.1: Exploratory metrics retrieved from the dataset.

The most frequently used words were found by aggregating all text across all ads and sorting the words by descending frequency. I then removed functional words such as “a” and “the” and the top 10 words, with their frequency in parentheses, are as





Figure 6.3: Figure 6.2 broken down by week.

4341 ads (49.0%) were detected to bear at least one disclosure icon. 3230 had both the AdChoices and mute icons, indicating that the two icons are often displayed as a package. Additionally, 3787 ads contain at least the mute icon, and since the mute icon is a feature exclusive to Google Ads, we know without foraying into ad data tracing that at least 42.7% of the dataset is a part of Google’s ad ecosystem.

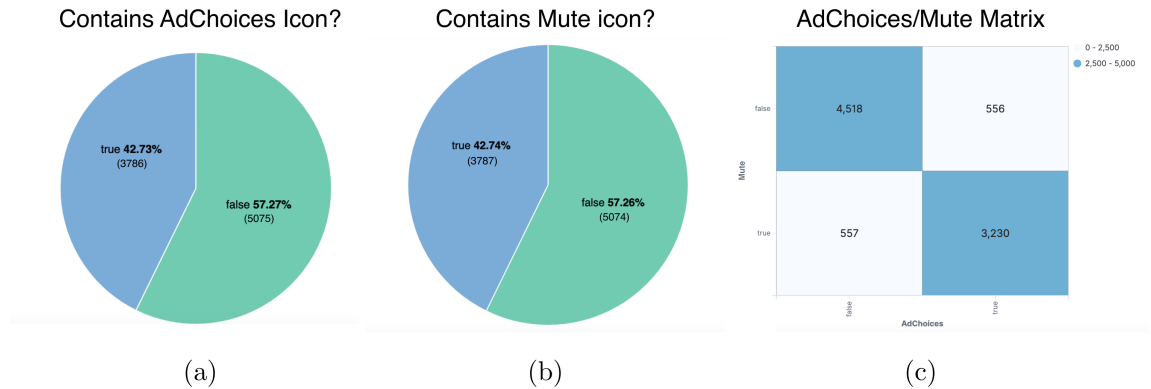


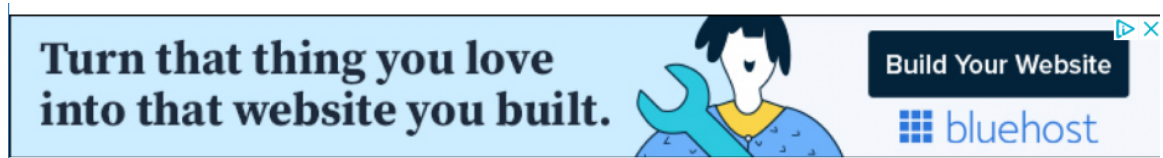
Figure 6.4: Prevalence of AdChoices (a) and mute (b), and matrix diagram with the two (c).

The top 5 most commonly detected objects, with the number of relevant ads in parentheses, are person (1667), packaged goods (575), top (258), wheel (160), and car (147). A tag cloud visualization of the top 100 objects detected across all ads is



### 6.2.1 Visual Analysis Performance

For each of the non-objective visual analysis tasks (that is, every field in Table 5.2 except colour and size class), I manually compared the automated system’s results to the actual image, noting down true and false positives/negatives where applicable and computing precision and recall. I also split disclosure detection into its various types (AdChoices, mute, self-disclosure, terms + conditions). I performed this analysis on a random sample of **188 ads** drawn from the dataset. Note that word-level accuracy was chosen to evaluate text extraction. A visual example of how I evaluated each ad is shown in Figure 6.7. The evaluations are aggregated across all analyzed ads to form the overall rate, and then aggregated by row to calculate precision, recall, and accuracy. The results are summarized in Table 6.2.



	output	reality	TP	TN	FP	FN
text	Build Your Website	Turn that thing you love into that website you built.				
	Turn that thing you love into that website you built.	Build Your Website				
	bluehost	bluehost	14/14 (1.00)	0	0	0
brands	Bluehost	Bluehost	1/1 (1.00)	0	0	0
industries	internet	internet	1/1 (1.00)	0	0	0
objects	none	wrench	0/1(0.00)	0	0	0
human faces	none	none	0	1/1 (1.00)	0	0
adchoices	true	true	1/1 (1.00)	0	0	0
mute	true	true	1/1 (1.00)	0	0	0
self disclosure	false	false	0	1/1 (1.00)	0	0
terms + conditions	false	false	0	1/1 (1.00)	0	0

Figure 6.7: Example of accuracy analysis of one ad through manual inspection.

### Discussion

Overall, the visual analysis pipeline is accurate enough to provide a coherent search experience on the frontend, but it is definitely not perfect.

Table 6.2: Summary of visual analysis accuracies for 188 ads.

Field	TP	TN	FP	FN	Precision	Recall	Accuracy
Text (word-level accuracy)	0.97	0	0.01	0.02	0.99	0.98	0.97
Brands	0.51	0.14	0.01	0.35	0.99	0.59	0.65
Industries	0.50	0.12	0.01	0.37	0.99	0.57	0.62
Objects	0.43	0.33	0.02	0.22	0.96	0.66	0.76
Human faces	0.16	0.81	0	0.03	1.00	0.84	0.97
AdChoices icon	0.51	0.42	0.01	0.06	0.99	0.90	0.94
Mute icon	0.50	0.47	0.01	0.03	0.99	0.95	0.97
Self-disclosure	0.13	0.87	0.01	0	0.96	1.00	0.99
Terms + conditions	0.05	0.92	0	0.04	1.00	0.56	0.96

Brand and industry detection sees the worst performance within the entire task collection. It is not surprising considering the difficulty and wide scope of the task. The relatively low overall accuracy stems primarily from low recall: there are many brands and industries that are missed by the algorithm. One reason for this is that many advertisers do not exist in our database (and are unlikely to in any general-purpose brands database) due to them being relatively small and unknown. Many are also not brands in the conventional, product-selling sense. For example, many chumbox ad “brands” are often marketing and copywriting firms who pool together content from writers specializing in viral and clickbait articles [29]. SmartAsset or world-health-wellness.com are certainly not household names, but may appear on ads as they work directly with Outbrain and Taboola to populate chumboxes at large scales. Such cases may or may not be worth addressing immediately due to the algorithm’s much higher precision. If we were to address them, however, we can run a crowdsourcing task to extract some of the more obscure brands from our archived ads and add them to our brands database. This allows us to identify the brands if we were to encounter them again. Relatedly, although I performed many rounds of cleaning to arrive at the current brands database (see “Brand and Industry Detection” from Chapter 5.2.2), some noise still exists and can lead to false positives. Further cleaning may also be incorporated as part of this crowdsourcing task.

A reason for poor industry detection performance is that currently, brand and industry detection are inextricably linked. Every brand in our database is associated with an industry, which is returned if its corresponding brand is detected. There are a couple issues with this. First, failure to detect a brand means failure to detect an industry. Second, the industry will be set to “unidentified” if the brand is identified by Vision API (as opposed to my custom algorithm) and the API’s name for the brand cannot be found in our database. This may be because the brand simply doesn’t exist in the database, or the same advertiser exists but is under a slightly different name

(e.g. “US Army” vs. “United States Armed Forces”). Both of these problems can be addressed by splitting off industry detection into its own task. A language classifier and object detector can be trained to assign ads to various industries in our database (see Appendix B) independent of brand extraction.

Another popular source of error lies in object detection. The Vision API contains some rather peculiar tags that are frequently detected as objects in ads. These include “1D barcode”, “2D barcode”, and “packaged goods”. While these object tags may be helpful in some other scenarios, such as checking whether a barcode exists on a shipping box, they can hinder accuracy in our task at hand. Balancing the breadth of use cases is the peril of any general-purpose object detector. Therefore, we may want to stray away from such detectors and explore training our own custom model specifically for web ads. This proposition would not have been feasible at the beginning of the project but is now thanks to our archive of more than 8000 ads.

All in all, we see that off-the-shelf computer vision tools are robust for some well-defined tasks. For example, OCR is relatively accurate and quite fast. I also observe that the API is still able to identify faces behind face masks, a particularly useful ability when analyzing ads related to COVID-19. These APIs certainly cannot be relied on for domain-specific use, however. Researchers who wish to perform custom analysis should be prepared to write their own algorithms and/or build upon other pre-existing libraries, as off-the-shelf APIs will either lack the desired feature or show poor performance on the task. Cost is another factor to consider. The Vision API queries are rather inexpensive (most features in the API costs \$1.50 for 1000 queries [33]) but researchers on a tight budget who run large-scale operations using these APIs may find them financially burdensome. Despite their downsides, it is worth noting that cloud API tools have the ability to quickly improve over time as their maintainers refine models and datasets. This makes them worth keeping a close eye on for the future.



### 6.2.2 Feedback from Other Researchers

On March 18, 2021, I presented a version of AdOculus with a smaller dataset of 348 ads to six researchers at Princeton CITP. Three of the six researchers saw potential integration of AdOculus in their work.

One of the researchers, who was analyzing ads on Facebook, expressed interest in both the crawler and the frontend. Although tools exist for browsing Facebook ads, either created by other research teams [72] or even Facebook itself [26], the underlying reliance on the Facebook Ad Library API introduces engineering challenges [18] while limiting the browsable dataset to currently active ad campaigns. A crawler independent of the API allows for the creation of a true archive and more flexible analysis. Being built upon OpenWPM enables our crawler to be repurposed for collecting ads on other platforms such as Facebook while preserving the same metadata collection capabilities. Additionally, the frontend of AdOculus can simply be rewired to another Elasticsearch index to display a separate archive of ads. To facilitate searching, one can archive ads into their own Elasticsearch index using the current visual analysis pipeline and JSON schema. Of course, different platforms have different analysis needs (mute icon detection is not relevant on Facebook, for example), so some modifications to the analysis pipeline will be necessary. However, the backbone of the system will remain consistent, empowering infrastructure modularity in similar types of research and reducing repeated code.

Another researcher was interested in using the publishers filter to aid the examination of the relationship between website appearance and trust. It would be helpful to display the ads in situ on a full-page screenshot of the website, but AdOculus does not currently show the coordinates of an ad on the host page. However, the ad’s HTML is gathered as part of the metadata collection. After some refining and formatting, it could be made available in the interface for this researcher and others who wish to inspect an ad’s wrapper code. This feature has since been added to the

project agenda.

Finally, a researcher wanted to use our tool to examine data privacy through retargeted ads. Our crawler currently runs in an unprofiled, private browser that is closed after each session, and we are working on but have not yet completed intermediary extraction during metadata collection. This limits our ability to assist in research that traces data flows between ad agents. Although we are unable to support the researcher at the moment, we plan on running more diverse crawls using different profiles in the future as well as making ad intermediary information available in the interface upon refining the crawler.

## 6.3 New Research Possibilities

AdOculos, with its underlying dataset, brings about newfound research possibilities to many groups of experts.

For publishers, this means that they can track which ads appear on their site. It may seem absurd that publishers do not have agency over content displayed on their own platform, but the reality of modern ad intermediaries is that once a publisher opts in, any ad in the system can be shown unless they explicitly block a brand or general ad category [35]. However, this presents a chicken-and-egg problem where the publisher needs to first identify problematic brands or categories on their site to set blocking rules. This task, made difficult by the sheer scale on which most intermediaries operate, is easier with AdOculos.

For regulators, evidence of problematic ads can now be directly obtained. Other issues that required manual inspection and may not have been convenient or even possible to investigate at scale, such as claims made in ads or disclosure visibility, is within reach. Regulators can also identify potential partnerships with other experts in fields such as fairness, HCI, privacy, psychology, sociology, and more.

For the computer vision and machine learning community, this dataset provides training data for developing more sophisticated models to extract richer meaning behind text and graphics in advertising. It also provides a benchmark dataset with weakly labelled data so that researchers can further refine and test relevant visual analysis techniques. This dataset also opens up the possibility of new machine learning challenges on platforms such as Kaggle to explore topics such as automated visual metaphor interpretation and perceptual ad blocking.

For HCI and psychology researchers, performing user studies on subsets of this dataset can help enhance understanding of how web ads trigger mental stimuli in viewers. Baseline measurement experiments involving the comparison of web ads to a much more plain, text-only variant, as outlined by Mathur et al. [61], can be used here. User studies can also explore areas where automated algorithms are unreliable. Given the current state of computer vision, some more nuanced visual analyses such as detecting emotion and race in human faces are better left for humans to perform.

A list of potential research questions along with groups who may be interested in them can be found in Table 6.3. There are, of course, many more questions that can be explored than listed. The magnitude of questions and flexibility in future work is perhaps AdOculos’s most important contribution.

## **6.4 Findings from Two Explorations**

To provide an example of just some of the aforementioned possibilities, I tackled two research questions using our work. The questions are listed in Chapter 4.2.2.

### **6.4.1 Self-disclosures**

The first of my questions inquired how self-disclosures are represented in our archived ads. Using the methodology outlined in 5.3.1, I gathered 1069 relevant ads. I then

Question	Relevant Group(s)
How are people depicted in ads of different industries, with respect to gender, race, emotion, etc.?	Fairness researchers, sociologists, journalists
Which ads target vulnerable populations, such as children or older adults?	Regulators, fairness + privacy + HCI researchers, journalists
Which ads mimic the UX of another platform (shopping, games, etc.) and to what extent is that deceptive?	Regulators, HCI researchers, designers
How can one distinguish false claims from exaggerated ones in ads?	Regulators, NLP researchers
How can computer vision systems better understand ads?	CV and ML researchers
How can perceptual ad blockers be refined to increase accuracy?	CV and ML researchers
Are there any publishers who are consistently displaying low-quality/deceptive ads? If so, which ones?	Regulators, publishers, advertisers, intermediaries
To what extent are comparison ads (ads that compare their product to others) truthful?	Regulators, advertisers
What design guidelines advertisers can follow to reduce the number of problematic ads?	Regulators, advertisers, designers
How do ads trigger affective stimuli in viewers?	Psychologists, HCI researchers, designers

Table 6.3: Potential research questions to explore and groups who may be interested in them.

looked through them manually and removed 21 ads that were occluded or miscropped, leaving 1048 ads.

In my analysis, I focused on the relationships between font size, RMS contrast, and relative location of the geometric centers of the self-disclosures. Histogram plots of font size and RMS contrast across all 1048 ads can be found in Figure 6.8. I then used the relative  $(x, y)$  coordinates to plot the self-disclosures as a scatterplot in a unit square, inverting the y-axis so that the square resembles a pixel grid in an ad

image. Each marker has a colour based on the disclosure's RMS contrast and a size based on font size. Figure 6.9 shows the resulting plot.

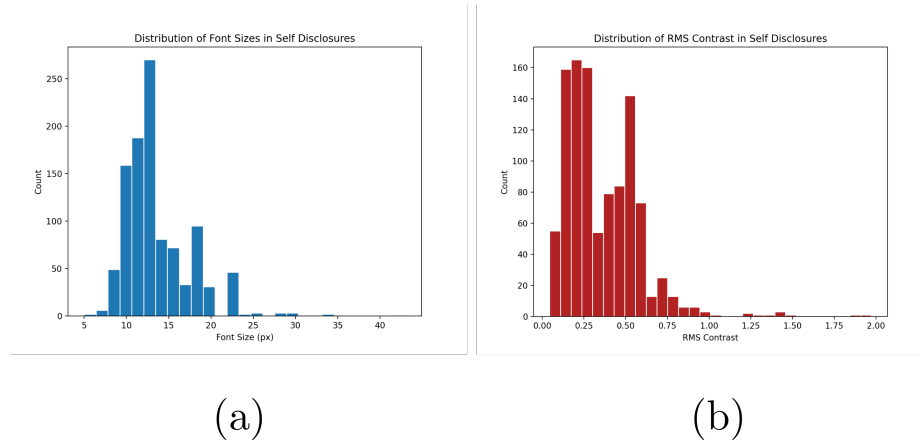


Figure 6.8: Histograms of font sizes (a) and RMS contrasts (b).

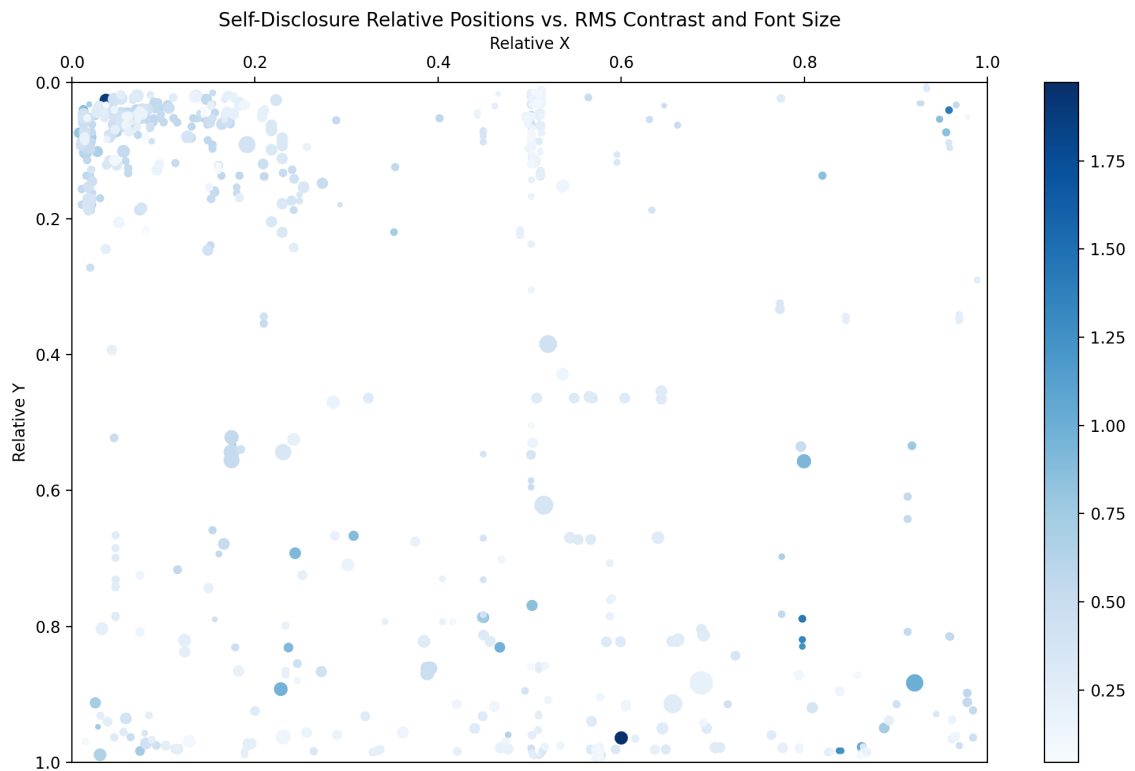


Figure 6.9: Visualization of 1048 self-disclosures.

Figure 6.9 is interesting because it reveals some standardization of disclosure place-

ment and contrast in the top left and top middle of an ad. A collection of smaller, lower-contrast disclosures can be seen near the bottom, with a few exceptions. To concretely examine this phenomenon, I separated the examined ads based on relative y coordinates into the top quartile ( $y_{rel} \leq 0.25$ ) and bottom quartile ( $y_{rel} \geq 0.75$ ). Indeed, upon comparison of mean and median RMS contrast and font size between the top and bottom quartiles, the bottom showed noticeably lower contrast while maintaining a similar font size when compared to the top. Disclosures towards the bottom also varied more stylistically, accounting for the relatively large difference in mean and median values in the bottom quartile. I summarize relevant mean and median statistics for all analyzed disclosures, top quartile, and bottom quartile in Table 6.4.

	<b>n</b>	<b>RMS Contrast</b>	<b>Font Size (px)</b>
<b>all</b>	1048	Mean: 0.36 Median: 0.30	Mean: 13.77 Median: 13.0
<b>top quartile</b>	696	Mean: 0.37 Median: 0.37	Mean: 13.30 Median: 13.0
<b>bottom quartile</b>	260	Mean: 0.32 Median: 0.24	Mean: 14.37 Median: 13.0

Table 6.4: Summary of mean and median metrics from self-disclosures.

### 6.4.2 Ad Similarity

The other research question I tackle revolves around the similarity of ads published on the same webpage. Throughout this section, I will be frequently referencing the variables I defined in Table 5.4.

I qualitatively examined ads from the top 30 publishers on AdOculos and found that on average, 49.7 % of the ads on a publisher’s page belonged to the same industry.

That is,

$$\overline{m_{ind}} = \frac{1}{N} \sum_{i=1}^N \frac{m_{ind_i}}{n_{P_i}} = 0.497$$

when  $N = 30$ . This number can be slightly misleading since it can be broken down into two distinct categories:  $b_{ind} = true$  (the dominant industry is aligned with that of the publisher’s content) and  $b_{ind} = false$  (there is no relationship between the dominant industry and that of the publisher’s content).

In cases where  $b_{ind} = true$  ( $N = 11$ , or 36.7% of publishers examined),  $\overline{m_{ind}} = 0.728$ , whereas the same average was 0.363 for publishers where  $b_{ind} = false$ . It is, after all, not so surprising for the former average to be high. For example, 100% of ads in the dataset published on archdaily.com, an architectural website, are related to architecture and building materials. There is clear agreement between the publisher and SSP to only show ads relevant to the publisher’s regular content. This is not the case in the latter group of publishers. Regardless, an  $\overline{m_{ind}}$  of 36.3% raises some eyebrows since the most common industry across all ads in AdOculus (internet) occupies only 6.87% of the dataset. This phenomenon might not go unexplained, however. Despite the opaque nature of ad intermediaries and lack of accessible technical documentation on SSPs, some researchers have suggested that revenue-maximizing ad selection algorithms for SSPs use k-means clustering to group ads based on common attributes and “package” pre-purposed slots together for sale [66].

Besides similarities in the industry advertised, another relevant metric is the mean normalized  $m_b$ . This represents the mean percentage of ads occupied by the top brand on a page. For publishers where  $b_{ind} = true$ , this number averages out to be 37.9%, and for the rest, it is 26.0%. Going even deeper, the mean  $m_{same}$ , or percentage of ads on a page considered the same (same brand, message, and advertised product but not necessarily dimensions), is 23.8% in publishers where  $b_{ind} = true$  and 17.6% otherwise. Closer definitions of similarity do not show as large of a disparity between the two categories of publishers, but still reveal the effects of supply-side clustering.

Figures 6.10(a), 6.10(b), and 6.10(c) show histograms of normalized  $m_{ind}$ ,  $m_b$ , and  $m_{same}$ , respectively, organized by a publisher's  $b_{ind}$ . Figure 6.11 visualizes the relationship between  $m_b$  and  $m_{same}$ , or how prevalent identical ads are among ads of the top advertiser, across all 30 publishers.

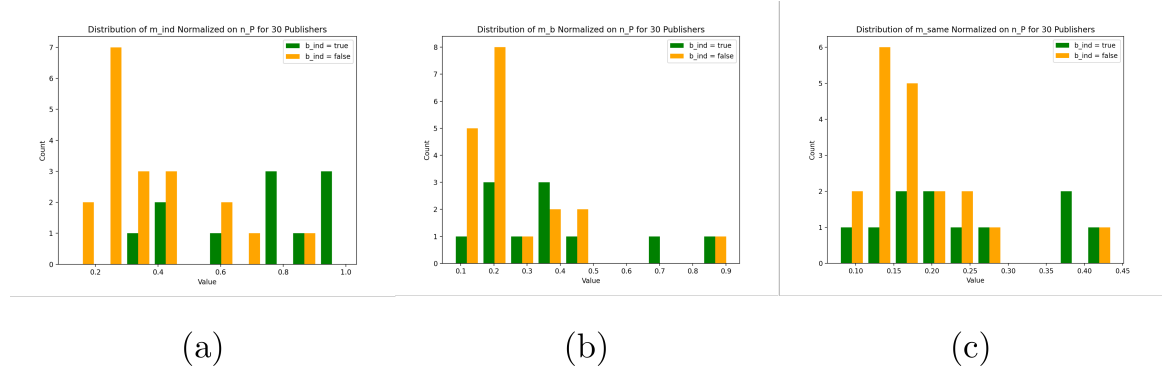


Figure 6.10: Histograms of normalized  $m_{ind}$ ,  $m_b$ , and  $m_{same}$  across 30 publishers.

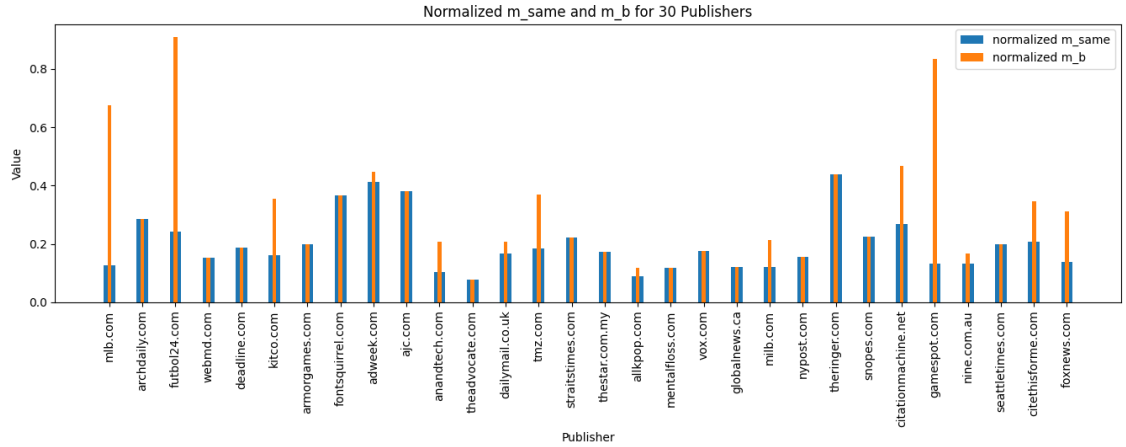


Figure 6.11: Relationship between normalized  $m_b$  and  $m_{same}$  across 30 publishers.

### 6.4.3 Additional Policy Implications

In Chapter 4.3, I outlined some policy recommendations based on the possibilities of AdOculos. After gleaning insights from the dataset and research questions, I propose an additional policy as well as an enhancement to one I proposed previously.



Results to the self-disclosures investigation revealed that there is some standardized disclosure practice already in place. Disclosures in the top geographic quartile of an ad are more consistent in placement and contrast. This contributes positively to the development of a common visual language through which disclosures can be communicated as well as a reliable mind map to help consumers identify disclosures. Beyond the top quartile, however, the area starts becoming more like the wild west of disclosures. As Figure 6.9 shows, highly varied contrasts, font sizes, and placements dominate the lower half of an ad. In the lower quartile specifically, a select few in the area appear to be highly visible, but a large number are small and low-contrast. Such disclosures may be purposefully evading attention.

The FTC emphasizes the importance of geographic placement of disclosures in advertising and discourages placement in areas where consumers will look the least, such as the bottom of a screen or page [27]. To promote clear self-disclosures, I propose a standard disclosure badge that can be used across web ads. Standards along a similar line of execution, such as the AdChoices icon, have seen success in the past. Taking advantage of disclosure standards already in place and drafting a set of creative guidelines, much like the one set out by the DAA [16], may be a promising first step.

Findings from the ad similarity examination indicate that ads are clustered by supply-side platforms during the sale process, and that the results from such clustering are visible to the consumer. To illustrate why this may be problematic, consider the example of [citationmachine.net](http://citationmachine.net). 18 of the 30 ads collected from the publisher are related to job search, with US Army being the advertiser for 15 of those 18 ads. The Army ads contained messaging such as “get your degree without getting into debt”, “serious work, serious benefits”, and “start a solid career by talking to a US Army reserve soldier now”. Moreover, 10 of the 15 Army ads were collected on the same visit. Marketing and psychology studies have shown that repetition of a message

can boost its credibility even if it was not originally perceived to be believable ([50], [80]). While the Army ads are not particularly harmful, the coordinated repetition of messages across a website can be a powerful and manipulative tool for advertisers. In fact, a pernicious use of repetition may be closer than we expect. By inspecting the ads from the news site `nine.com.au`, one can see that 9 of the 30 ads shown by the publisher are low-quality, clickbait ads for blood sugar reduction. 4 of those ads were collected on the same visit.

To prevent potentially harmful messaging from being amplified through web ads, I propose an addition to the aforementioned policy targeting ad intermediaries, this time specifically targeting SSPs: the size of ad clusters should be capped to below a certain proportion of all available slots offered by a publisher. This may also induce a side effect that further contributes to ad diversity. Smaller advertisers who could not afford premium slot “packages” due to the SSP’s revenue-maximizing clustering strategies can now have a higher chance of publishing their ad, leveling the playing field for a fairer market.

# Chapter 7

## Conclusion

In February 2019, Mozilla announced on its company blog that “the online advertising ecosystem is broken” and that it was working with partners to explore alternative funding models for the web [17]. This may be a promising step forward, but with the current scale of ad infrastructure and record levels of ad spending year after year, ads are here to stay, at least for the near future. It is undisputed that ads play a vital role on the internet and will continue doing so. Some platforms that tackle important issues with the modern web, such as DuckDuckGo, rely on ad revenue to do the work they do. With that said, it is equally undeniable that the quality of ads can be improved to reduce consumer frustration and harm.

This thesis strives to improve the ad ecosystem by making ad-related research more accessible and insightful. To do this, my collaborators and I collected and archived ads from thousands of popular websites, creating the largest publicly browsable, general-purpose dataset of modern web ads to date. We combined this with a visual analysis pipeline that automatically extracts ad information at scale and brought everything together in AdOculos, an interface that allows users to browse, search, and filter for ads based on extracted data.

Our methods are not without their limitations. First, we have yet to explore the

world of ads outside of those found on websites’ landing pages. This includes ads on other pages of the same websites we crawled as well as ads on other platforms, such as social media and search engine ads. Our findings and research questions are based on studying the narrow sample of web ads we collected and we may gain many additional insights if we expand our sample. Second, our ads are currently limited to static images due to our collection methodology. Some information may be falling through the cracks since many ads rely on animation and video to convey their full message. Third, we are running crawls on an unprofiled browser with the location set to Princeton, New Jersey. The ads we collect are targeted by location, and it is difficult to say whether or not our sample is representative of ads others see online. Finally, some ads collected are either occluded or have failed to load. Automatically detecting and removing such ads before they are archived into AdOculus will be beneficial for the dataset’s overall presentation.

Our work opens the gate for a plethora of future work. For starters, the AdOculus interface can be improved to support smooth user experiences in browsing large quantities of data. The current gallery view limits how many ads can be seen at once, and it could be helpful, for example, to create an interactive data cloud view in which users can zoom into large clusters of related ads and click into each one for more details. We can add tools for creating interactive data visualizations of the dataset based on searches and filters applied to help other researchers can better communicate their findings. We can also invite others to query our database via an API to enable more granular investigations that cannot easily be conducted by referencing the AdOculus interface alone. I benefited from database API queries when I was tackling the self-disclosures research question in Chapter 6.4.1. Currently, our database API is private, but upon some additional enhancements such as documentation, tutorials, and traffic balancers, we plan to release it to the public.

The crawler also has much potential for improvement. Our short term goals for

it include automatically dismissing popup dialogues, identifying ad intermediaries involved in serving an ad, and clicking into links and buttons to capture ad landing pages. Our longer term goals include running crawls using differently profiled browsers from multiple locations with a VPN to allow for more exploration into areas such as targeting and privacy.

Ultimately, the goal of this thesis is to raise more questions than it was set out to answer. It leaves behind research tools and a cornucopia of research questions ripe for exploration, which can be addressed by multiple groups of experts using more technical methods as well as qualitative user studies. While new funding models for the web may be under development, ads are the current reality we must face. As more people join the internet and ad agents continue to expand their dominance, the stakes for bad ads are higher than ever. Now is the time for experts from diverse fields to work together to reduce problematic web ads, protect consumers, and create a healthier, more sustainable advertising ecosystem.

# References

- [1] Alexa Internet, Inc. The top 500 sites on the web.  
<https://www.alexa.com/topsites>, 2020. Accessed: 2020-06-02.
- [2] M. A. Amazeen and B. W. Wojdyski. Reducing Native Advertising Deception: Revisiting the Antecedents and Consequences of Persuasion Knowledge in Digital News Contexts. *Mass Communication and Society*, 22(2):222–247, 2019.
- [3] AppbaseIO. ReactiveSearch: UI components library for Elasticsearch.  
<https://github.com/appbaseio/reactivesearch>. Accessed: 2021-01-09.
- [4] M. A. Bashir, S. Arshad, W. Robertson, and C. Wilson. Tracing information flows between ad exchanges using retargeted ads. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 481–496, Austin, TX, Aug. 2016. USENIX Association.
- [5] M. Brain. How Web Advertising Works.  
<https://computer.howstuffworks.com/web-advertising.htm#pt6>. Accessed: 2021-02-12.
- [6] J. Buchner. ImageHash. <https://github.com/JohannesBuchner/imagehash>, 2013. Accessed: 2020-09-29.
- [7] Bureau of Consumer Protection. *Advertising and Marketing on the Internet*. Federal Trade Commission , 9 2021.  
<https://www.ftc.gov/system/files/documents/plain-language/bus28-advertising-and-marketing-internet-rules-road2018.pdf>.
- [8] F. Chanchary and S. Chiasson. User Perceptions of Sharing, Advertising, and Tracking. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 53–67, Ottawa, July 2015. USENIX Association.
- [9] C. Cimpanu. Expandable ads can be entry points for site hacks.  
<https://www.zdnet.com/article/expandable-ads-can-be-entry-points-for-site-hacks/>, 2018. Accessed: 2021-02-06.
- [10] Cliqz Open Source. Adblocker. <https://github.com/cliqz-oss/adblocker>, 2017. Accessed: 2021-03-08.

- [11] M. Conti, V. Cozza, M. Petrocchi, and A. Spognardi. Trap: Using targeted ads to unveil google personal profiles. In *2015 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2015.
- [12] H. Cramer. Effects of Ad Quality & Content-Relevance on Perceived Content Quality. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, page 2231–2234, New York, NY, USA, 2015. Association for Computing Machinery.
- [13] P. Darke and R. Ritchie. The Defensive Consumer: Advertising Deception, Defensive Processing, and Distrust. *Journal of Marketing Research*, 44:114–127, 2 2007.
- [14] M. de la Merced and T. Hsu. Taboola, Purveyor of Clickbait Ads, Will Go Public. <https://www.nytimes.com/2021/01/25/business/media/taboola-public-spac.html>, 2021. Accessed: 2021-02-15.
- [15] Digital Advertising Alliance. YourAdChoices Gives You Control. <https://youradchoices.com/>. Accessed: 2021-02-19.
- [16] Digital Marketing Alliance. DAA Icon Ad Marker Creative Guidelines. [https://digitaladvertisingalliance.org/sites/aboutads/files/DAA\\_files/DAA\\_Icon\\_Ad\\_Creative\\_Guidelines.pdf](https://digitaladvertisingalliance.org/sites/aboutads/files/DAA_files/DAA_Icon_Ad_Creative_Guidelines.pdf). Accessed: 2021-01-12.
- [17] P. Dolanjski. Exploring alternative funding models for the web. <https://blog.mozilla.org/futurereleases/2019/02/25/exploring-alternative-funding-models-for-the-web/>, 2019. Accessed: 2021-03-27.
- [18] P. Duke. What it’s like to actually use Facebook’s ad transparency tools. <https://medium.com/online-political-transparency-project/what-its-like-to-actually-use-facebook-s-ad-transparency-tools-7accf22f4ba7>, 2020. Accessed: 2021-03-06.
- [19] EasyList. EasyList. <https://easylist.to/>, 2005. Accessed: 2020-06-01.
- [20] L. Edelson. Publishing Facebook ad data (redux). <https://medium.com/online-political-transparency-project/publishing-facebook-ad-data-redux-ff071c41c12e>, 2020. Accessed: 2021-03-06.
- [21] Elastic. Elasticsearch Reference. <https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html>. Accessed: 2020-12-15.
- [22] Elastic. Enterprise search, observability, and security for the cloud. <https://www.elastic.co/cloud/>. Accessed: 2021-01-17.
- [23] S. Englehardt and A. Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of ACM CCS 2016*, 2016.

- [24] F. R. Eric Zeng, Tadayoshi Kohno. What Makes a “Bad” Ad? User Perceptions of Problematic Online Advertising. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, page 1–24, New York, NY, USA, 2021. Association for Computing Machinery.
- [25] F. Etro. Leadership in Multi-sided Markets and the Dominance in Online Advertising. *Recent Advances in the Analysis of Competition Policy and Regulation*, pages 214–234, 2012.
- [26] Facebook. Ad Library. <https://www.facebook.com/ads/library/>, 2019. Accessed: 2020-10-15.
- [27] L. Fair. Full Disclosure. <https://www.ftc.gov/news-events/blogs/business-blog/2014/09/full-disclosure>, 2014. Accessed: 2021-03-25.
- [28] Federal Trade Commission. *Disclosures 101 for Social Media Influencers*, 2019. Accessed: 2021-02-26.
- [29] A. Goldman and P. Vogt. *An Ad for the Worst Day of Your Life*. Gimlet Media, 6 2018. Podcast audio transcription.
- [30] G. Gomez-Mejia. “fail, clickbait, cringe, cancel, woke”: Vernacular criticisms of digital advertising in social media platforms. In G. Meiselwitz, editor, *Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing*, pages 309–324. Springer International Publishing, 2020.
- [31] Google. Preparing your training data. <https://cloud.google.com/vision/automl/docs/prepare>. Accessed: 2020-11-24.
- [32] Google. The DoubleClick Ad Exchange. <https://static.googleusercontent.com/media/www.google.com/en/adexchange/AdExchangeOverview.pdf>. Accessed: 2020-12-14.
- [33] Google. Vision AI. <https://cloud.google.com/vision>. Accessed: 2020-10-15.
- [34] Google. More control with “mute this ad” [x] icon. <https://adwords.googleblog.com/2012/06/more-control-with-mute-this-ad-x-icon.html>, 2012. Accessed: 2021-02-19.
- [35] Google Ad Manager Help. Block Ads by Type. [https://support.google.com/admanager/topic/2913545?hl=en&ref\\_topic=7505998](https://support.google.com/admanager/topic/2913545?hl=en&ref_topic=7505998). Accessed: 2020-12-13.



- [36] Google Ads. Reach more people in more places online.  
<https://ads.google.com/home/campaigns/display-ads/>. Accessed: 2020-12-12.
- [37] S. Hansell. BITS; If Ads Were Traded Like Pork Bellies.  
<https://archive.nytimes.com/query.nytimes.com/gst/fullpage-9C04EEDF1E3EF930A35750C0A96E9C8B63.html>. Accessed: 2020-12-14.
- [38] J. Hornik. Quantitative analysis of visual perception of printed advertisements. *Journal of Advertising Research*, pages 41–48, 1980.
- [39] T. Hsu. You Will Be Shocked by This Article. <https://www.nytimes.com/2019/10/04/business/media/online-advertising-chumbox-merger.html>, 2019. Accessed: 2021-02-15.
- [40] C. Hudson. Step-By-Step HTML5 Ad Creation With Adobe Animate CC.  
<https://blog.adobe.com/en/2016/05/11/step-by-step-html5-ad-creation-with-adobe-animate-cc.html#gs.tlvq6x>, 2016. Accessed: 2021-02-12.
- [41] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic Understanding of Image and Video Advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110, 07 2017.
- [42] D. A. Hyman, D. Franklyn, C. Yee, and M. Rahmati. Going Native: Can Consumers Recognize Native Advertising? Does it Matter? *Yale Journal of Law & Technology*, 19(2):79–112, 2018.
- [43] IAB Technology Laboratory. Standards.  
<https://iabtechlab.com/standards/>. Accessed: 2021-01-25.
- [44] IAB Technology Laboratory. Fixed Size Ad Specifications.  
[https://www.iab.com/wp-content/uploads/2019/04/IABNewAdPortfolio\\_LW\\_FixedSizeSpec.pdf](https://www.iab.com/wp-content/uploads/2019/04/IABNewAdPortfolio_LW_FixedSizeSpec.pdf), 2017. Accessed: 2021-01-25.
- [45] Interactive Advertising Bureau. Digital Advertising Regulation 101.  
<https://www.iab.com/news/digital-advertising-regulation-101/>, 2014. Accessed: 2021-02-21.
- [46] Interactive Advertising Bureau. IAB New Ad Portfolio: Advertising Creative Guidelines. <https://www.iab.com/guidelines/iab-new-ad-portfolio/>, 2017. Accessed: 2021-01-25.
- [47] S. Jeong. Visual Metaphor in Advertising: Is the Persuasive Effect Attributable to Visual Argumentation or Metaphorical Rhetoric? *Journal of Marketing Communications*, 14(1):59–73, 2008.

- [48] X. Jin, W. Su, R. Zhang, Y. He, and H. Xue. The Open Brands Dataset: Unified Brand Detection and Recognition at Scale. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4387–4391, 2020.
- [49] G. Johar. Mistaken Inferences from Advertising Conversations: A Modest Research Agenda. *Journal of Advertising*, 45:1–8, 08 2016.
- [50] G. V. Johar and A. L. Roggeveen. Changing False Beliefs from Repeated Advertising: The Role of Claim-Refutation Alignment. *Journal of Consumer Psychology*, 17(2):118–127, 2007.
- [51] J. Kees and J. C. Andrews. Research Issues and Needs at the Intersection of Advertising and Public Policy. *Journal of Advertising*, 48(1):126–135, 2019.
- [52] W. Kirch, editor. *Pearson’s Correlation Coefficient*, pages 1090–1091. Springer Netherlands, Dordrecht, 2008.
- [53] L. Kloot. Native Ads Vs. Display Ads: What are the Differences?  
<https://www.outbrain.com/blog/native-ads-vs-display-ads/>, 2018. Accessed: 2021-02-06.
- [54] A. LaFrance. The First-Ever Banner Ad on the Web.  
<https://www.theatlantic.com/technology/archive/2017/04/the-first-ever-banner-ad-on-the-web/523728>, 4 2017. Accessed: 2020-12-13.
- [55] K. Lant. 15 banner ad design tips to get more clicks.  
<https://99designs.com/blog/marketing-advertising/14-design-tips-for-more-clickable-banner-ads/>, 2020. Accessed: 2021-02-11.
- [56] G. Lawton. The Ethics of Ad Platforms.  
<https://torquemag.io/2014/08/ethics-ad-platforms/>, 2014. Accessed: 2021-03-13.
- [57] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczynski, and W. Joosen. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. *Proceedings 2019 Network and Distributed System Security Symposium*, 2019.
- [58] K. Lindsay. I’ve Designed 100+ Display Ads: Here’s What I Learned.  
<https://www.wordstream.com/blog/ws/2019/07/09/ad-design>, 2019. Accessed: 2021-02-10.
- [59] T. Lorenz. Instagram Has a Massive Harassment Problem.  
<https://www.theatlantic.com/technology/archive/2018/10/instagram-has-massive-harassment-problem/572890/>, 2018. Accessed: 2021-03-14.

- [60] A. Mathur, G. Acar, M. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), 2019.
- [61] A. Mathur, J. Mayer, and M. Kshirsagar. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods, 2021.
- [62] A. Mathur, A. Narayanan, and M. Chetty. Endorsements on Social Media: An Empirical Study of Affiliate Marketing Disclosures on YouTube and Pinterest. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), Nov. 2018.
- [63] S. Messner. colorgram.py. <https://github.com/obskyr/colorgram.py>, 2016. Accessed: 2021-03-29.
- [64] Mopub (Twitter Inc.). Understanding ad networks. <https://www.mopub.com/content/dam/mopub-aem-twitter/migration/39639-mopub-understanding-ad-networks.pdf>, 2017. Accessed: 2021-01-22.
- [65] Mozilla. JavaScript APIs. <https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/API>. Accessed: 2021-03-08.
- [66] A. Mukherjee, R. P. Sundarraj, and K. Dutta. Apriori Rule-Based In-App Ad Selection Online Algorithm for Improving Supply-Side Platform Revenues. *ACM Trans. Manage. Inf. Syst.*, 8(2-3), July 2017.
- [67] S. Muthukrishnan. Ad Exchanges: Research Issues. In S. Leonardi, editor, *Internet and Network Economics*, pages 1-12, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
- [68] A. Narayanan, A. Mathur, M. Chetty, and M. Kshirsagar. Dark Patterns: Past, Present, and Future. *ACM Queue*, 18(2):67-91, 2020.
- [69] Netlify. Netlify: Develop & deploy the best web experiences in record time. <https://www.netlify.com/>. Accessed: 2021-02-23.
- [70] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal. Dark Patterns after the GDPR: Scraping Consent Pop-Ups and Demonstrating Their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1-13, New York, NY, USA, 2020. Association for Computing Machinery.
- [71] K. Novatska. What is an Ad Tag and How to Generate It. <https://epom.com/blog/ad-server/what-is-an-ad-tag>, 2020. Accessed: 2021-02-10.
- [72] NYU Cybersecurity for Democracy. NYU Ad Observatory. <https://adobservatory.org/>, 2020. Accessed: 2020-09-14.

- [73] K. O'Donnell and H. Cramer. People's Perceptions of Personalized Ads. In *24th International World Wide Web Conference*, page 1293–1298, New York, NY, USA, 2015. Association for Computing Machinery.
- [74] People Data Labs. 7+ Million Company Dataset. <https://www.kaggle.com/peopledatalabssf/free-7-million-company-dataset>. Accessed: 2021-01-12.
- [75] S. Petridis and L. B. Chilton. Human Errors in Interpreting Visual Metaphor. In *Proceedings of the 2019 on Creativity and Cognition*, page 187–197, New York, NY, USA, 2019. Association for Computing Machinery.
- [76] B. Phillips and E. McQuarrie. Beyond Visual Metaphor: A New Typology of Visual Rhetoric in Advertising. *Marketing Theory*, 4:113 – 136, 2004.
- [77] R. Pieters, M. Wedel, and R. Batra. The Stopping Power of Advertising: Measures and Effects of Visual Complexity. *Journal of Marketing*, 74(5):48–60, 2010.
- [78] Puppeteer. Puppeteer. <https://github.com/puppeteer/puppeteer>, 2017. Accessed: 2021-01-28.
- [79] React-Bootstrap. React-Bootstrap: Bootstrap components built with React. <https://github.com/react-bootstrap/react-bootstrap>. Accessed: 2021-02-03.
- [80] A. L. Roggeveen and G. V. Johar. Perceived Source Variability Versus Familiarity: Testing Competing Explanations for the Truth Effect. *Journal of Consumer Psychology*, 12(2):81–91, 2002.
- [81] Selenium. SeleniumHQ Browser Automation. <https://www.selenium.dev/>. Accessed: 2020-06-03.
- [82] A. Shaouf, K. Lü, and X. Li. The effect of web advertising visual design on online purchase intention: An examination across gender. *Computers in Human Behavior*, 60:622–634, 2016.
- [83] B. Smith. 5 Design Principles to Master for Better Display Ads. <https://www.wordstream.com/blog/ws/2018/07/30/design-principles>, 2018. Accessed: 2021-02-10.
- [84] Statista. Ad blocking user penetration rate in the United States from 2014 to 2021. <https://www.statista.com/statistics/804008/ad-blocking-reach-usage-us/>, 2021. Accessed: 2021-02-22.
- [85] Statista. Internet advertising spending in the United States from 2016-2024. <https://www.statista.com/statistics/183523/online-advertisement-spending-in-the-us/>, 2021. Accessed: 2021-01-11.

- [86] M. Stoller. *Ad Tech and the News*. Center for Journalism and Liberty, 2020. Accessed: 2020-12-14.
- [87] G. Storey, D. Reisman, J. Mayer, and A. Narayanan. The Future of Ad Blocking: An Analytical Framework and New Techniques, 2017.
- [88] H. Su, S. Gong, and X. Zhu. WebLogo-2M: Scalable Logo Detection by Deep Learning from the Web. In *IEEE International Conference on Computer Vision, Workshop on Web-scale Vision and Social Media*, 10 2017.
- [89] M. Swart, Y. Lopez, A. Mathur, and M. Chetty. Is This An Ad?: Automatically Disclosing Online Endorsements On YouTube With AdIntuition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.
- [90] TechCrunch. RMX Direct: alternative ad networks battle for your blog. <https://techcrunch.com/2006/08/12/rmx-direct-alternative-ad-networks-battle-for-your-blog/>. Accessed: 2020-12-13.
- [91] C. Thomas and A. Kovashka. Persuasive Faces: Generating Faces in Advertisements, 2018.
- [92] TrustinAds.org. Trust in Ads. <https://www.trustinads.org/>. Accessed: 2021-03-13.
- [93] Truth in Advertising, Inc. *Prepared Statement of Truth in Advertising, Inc., Before the Committee on Energy and Commerce Subcommittee on Consumer Protection and Commerce House of Representatives*, 2021. <https://docs.house.gov/meetings/IF/IF17/20210204/111139/HHRG-117-IF17-Wstate-PattenB-20210204.pdf>.
- [94] U.S. House of Representatives. Section 230. 230. Protection for private blocking and screening of offensive material. [https://uscode.house.gov/view.xhtml?req=\(title:47%20section:230%20edition:prelim\)](https://uscode.house.gov/view.xhtml?req=(title:47%20section:230%20edition:prelim)). Accessed: 2021-03-14.
- [95] J. Veen. Looking back at Hotwired. <https://veen.com/jeff/archives/000903.html>, 7 2006. Accessed: 2020-12-13.
- [96] W3.org. Syntax and basic data types. <https://www.w3.org/TR/CSS2/syndata.html#color-units>. Accessed: 2021-01-19.
- [97] J. Wang, W. Min, S. Hou, S. Ma, Y. Zheng, H. Wang, and S. Jiang. Logo-2K+: a large-scale logo dataset for scalable logo classification. In *AAAI Conference on Artificial Intelligence*, 2020.

- [98] M. Wei, M. Stamos, S. Veys, N. Reitering, J. Goodman, M. Herman, D. Filipczuk, B. Weinshel, M. L. Mazurek, and B. Ur. What Twitter Knows: Characterizing Ad Targeting Practices, User Perceptions, and Ad Explanations Through Users' Own Twitter Data. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 145–162. USENIX Association, Aug. 2020.
- [99] K. Ye and A. Kovashka. ADVISE: Symbolism and External Knowledge for Decoding Advertisements. In *The European Conference on Computer Vision (ECCV)*, 9 2018.
- [100] M. Zawadziński. How Does Real-Time Bidding (RTB) Work? <https://clearcode.cc/blog/real-time-bidding/>, 2015. Accessed: 2020-12-13.
- [101] E. Zeng, T. Kohno, and F. Roesner. Bad News: Clickbait and Deceptive Ads on News and Misinformation Websites. In *Workshop on Technology and Consumer Protection (ConPro '20)*, 5 2020.
- [102] P. Zialcita. FTC Issues Rules For Disclosure Of Ads By Social Media Influencers. <https://www.npr.org/2019/11/05/776488326/ftc-issues-rules-for-disclosure-of-ads-by-social-media-influencers>. Accessed: 2021-02-26.

# Appendix A

## Code

All code for this project is currently stored in a private Github repository and will be made available upon the public launch of AdOculos. For any code inquiries, please contact `kjfeng@princeton.edu` and `amathur@princeton.edu`.

# Appendix B

## Advertiser Industries

accounting	civic & social organization
airlines/aviation	civil engineering
alternative dispute resolution	commercial real estate
alternative medicine	computer & network security
animation	computer games
apparel & fashion	computer hardware
architecture & planning	computer networking
arts and crafts	computer software
automotive	construction
aviation & aerospace	consumer electronics
banking	consumer goods
biotechnology	consumer services
broadcast media	cosmetics
building materials	dairy
business supplies and equipment	defense & space
capital markets	design
chemicals	e-learning



education management	industrial automation
electrical/electronic manufacturing	information services
entertainment	information technology and services
environmental services	insurance
events services	international affairs
executive office	international trade and development
facilities services	internet
farming	investment banking
financial services	investment management
fine art	judiciary
fishery	law enforcement
food & beverages	law practice
food production	legal services
fund-raising	legislative office
furniture	leisure travel & tourism
gambling & casinos	libraries
glass/ceramics/concrete	logistics and supply chain
government administration	luxury goods & jewelry
government relations	machinery
graphic design	management consulting
health/wellness/fitness	maritime
higher education	market research
hospital & health care	marketing and advertising
hospitality	mechanical or industrial engineering
human resources	media production
import and export	medical devices
individual & family services	medical practice

mental health care	public relations and communications
military	public safety
mining & metals	publishing
motion pictures and film	railroad manufacture
museums and institutions	ranching
music	real estate
nan	recreational facilities and services
nanotechnology	religious institutions
newspapers	renewables & environment
non-profit organization management	research
oil & energy	restaurants
online media	retail
outsourcing/offshoring	security and investigations
package/freight delivery	semiconductors
packaging and containers	shipbuilding
paper & forest products	sporting goods
performing arts	sports
pharmaceuticals	staffing and recruiting
philanthropy	supermarkets
photography	telecommunications
plastics	textiles
political organization	think tanks
primary/secondary education	tobacco
printing	translation and localization
professional training & coaching	transportation/trucking/railroad
program development	utilities
public policy	venture capital & private equity

veterinary

warehousing

wholesale

wine and spirits

wireless

writing and editing

# Appendix C

## Top 30 Publishers by Number of Ads Archived

theadvocate.com

dailymail.co.uk

mlb.com

tmz.com

straitstimes.com

archdaily.com

thestar.com.my

allkpop.com

mentalfloss.com

vox.com

futbol24.com

webmd.com

globalnews.ca

mlb.com

deadline.com

nypost.com

theringer.com

kitco.com

snopes.com

armorgames.com

fontsquirrel.com

citationmachine.net

gamespot.com

nine.com.au

seattletimes.com

adweek.com

ajc.com

anandtech.com

citethisforme.com

foxnews.com