Girls Who Code: A Randomized Field Experiment on Gender-Based Hiring Discrimination

By Marley Finley*

To detect the presence of gender-based hiring discrimination in software engineering roles, this paper analyzes data collected from a correspondence study. Response rates for 1,865 fictitious applications sent to unique firms reveal positive discrimination in favor of female applicants. Specifically, applications randomly assigned a female name exhibit a 41% higher probability of being invited to move forward in the interview process than those with male names, and a null hypothesis of no discrimination is rejected at the 1% significance level. Further analysis suggests that this bias for women may be stronger in younger firms than in older ones.

I. Background

This paper investigates gender discrimination in hiring decisions for software development roles, largely within the tech industry[1] and within so-called "startups."[2] This subset of the labor market has come under scrutiny in recent years, as news outlets (Mundy, 2017) and best-selling books (Szalai, 2018) have accused the technology sector of marginalizing female workers. In parallel with these criticisms, several initiatives exist to promote female representation in the tech industry: internal groups dedicated to supporting female employees (TechPoint, 2017), "Coding Boot Camps"[3] only for women (Bradford, 2020), and networking groups of women in computer science that connect members with mentors and employers (White, 2019). These cultural anecdotes illustrate the conflicting contextual factors that might contribute to discrimination against or for women in the technology sector. Furthermore, hiring practices that select for candidates

* Finley: Yale University, marleyfinley5@gmail.com. Thank you to my advisor, Costas Meghir, to my research team, Iyala Alai, Steve Bekele, Ran Wang, Candy Yang, and Charlotte Zimmer, to Zhengren Zhu, and to Vincent Billaut

[1]According to Investopedia, "The technology sector is the category of stocks relating to the research, development and/or distribution of technologically based goods and services. This sector contains businesses revolving around the manufacturing of electronics, creation of software, computers or products and services relating to information technology." (Frankenfield, 2019)

[2]The soft and hard definitions of a startup company are ambiguous. While business and investment literature tends to focus on companies' size, age, valuation, business model, scalability, and funding (Fontinelle (2020), Wilhelm (2018), Areitio (2018)), culturally, startups are associated with innovative entrepreneurship, focused problem-solving, making it "big" with large growth, and disrupting the status quo (Robehmed (2013), Landau (n.d.)).

[3]Coding Boot Camps teach attendees job-applicable coding languages through intensive lessons and often offer interview guidance and connections to firms.

based on "cultural fit"[4] and diversity[5] reflect these opposing forces and might similarly lead to women being disadvantaged or advantaged in the labor market.

Economic theory states that firms will rationally choose the candidate with the greatest marginal product of labor, optimizing for overall productivity of the firm. This fundamental notion goes back to early literature by Gary Becker on discrimination in labor economics (Becker, 1957), and is followed by the corollary that, under perfect competition, discriminatory employment practices that overlook maximizing firm productivity will lead firms with prejudiced hiring to fail against their rationally behaving competitors (Guryan and Charles, 2013). The type of discrimination described by Becker is referred to as "taste-based," where animus toward a group motivates unequal treatment, and is contrasted in literature to "statistical" discrimination, where firms infer candidates' abilities based on what is known about the groups to which they belong (men/women, black/white, immigrant/native etc.) to compensate for imperfect information about the specific individual (Guryan and Charles (2013), Altonji and Blank (1999)).

Beyond these two categories[6], group productivity dynamics are an important facet to consider as channels of discrimination. Since workers often act within a team setting, their efficacy in communicating and collaborating with other workers affects their overall benefit to a firm. Thus, if demographic attributes are relevant to group work, how a candidate's demographic background relates to that of firms' existing employees might be intrinsically relevant to the candidate's marginal product of labor. Considered in this light, "cultural fit" and "pro-diversity" hiring practices can be understood as rational behaviors under Becker's fundamental theory. Mixed results from research into productivity effects of teams with homogeneous and heterogeneous gender compositions (Bertrand and Duflo, 2017) prevent either practice from being discredited as irrational.

Regardless of these distinctions, just because discrimination might be theoretically explainable does not mean it is fair, and if a minority group is harmed, then the discrimination is illegal by U.S. law (EEOC, n.d.); discrimination, whether

[4]According to a 2012 sociology paper by Lauren Rivera (Rivera (2012)), candidate "leisure activities, experiences, and self-presentation styles," stand out to as "highly salient" factors that "often outweighed concerns about absolute productivity," with 40-75% of law, banking, and consulting firms interviewed citing "fit" to be their top evaluative criteria (Rivera (2012)). As mentioned in Bertrand and Duflo (2017), this "cultural fit" criteria is pervasive in the tech industry.

[5]Some firms actively recruit to increase the diversity of their talent pool, potentially favoring female applicants over otherwise comparable male alternatives (Huet (2017)). By offering larger referral bonuses for women, making employee demographics public, and setting targets for closer-to-equal representation, many tech firms demonstrate a willingness to pay for diversity, in financial incentives to current employees to tap into their female connections, as well as in risking customer backlash from their published statistics (Huet (2017)).

[6]Bertrand and Duflo (2017) raise a compelling point challenging the split categorization of taste-based and statistical discrimination, citing psychological studies that suggest that "the limited information and decision-making model that drives statistical discrimination might be itself endogenous to conscious or unconscious prejudice against the out-group members." They also emphasize that "discrimination, whether it is taste-based or statistical, can create or exacerbate existing differences between groups," and describe how taste-based discrimination "can easily morph into the more 'justifiable' form. 'Valid' stereotypes today could be the product of ambient animus, very much complicating the division between the different theories of discrimination." (Bertrand and Duflo, 2017)

taste-based or statistical, disproportionately harms a group of workers, and in the case of gender-based discrimination, approximately 50% of the overall population is affected. From a social justice perspective, it is important to ensure that the labor market provides all workers with equal opportunities, regardless of their demographic backgrounds. The presence of discrimination can also discourage minority groups from entering sectors of the labor market where they face an unfair disadvantage, depriving the market of their potential contributions and perhaps also diminishing their investment in their own human capital, thus perpetuating socioeconomic inequality. Furthermore, as Becker's theory highlights, firms that act on taste-based biases act contrarily to their own success, diminishing their productivity, shareholder value, and overall market efficiency. Given the great deal of attention on the under-representation of women in the tech industry, specific inquiry into the influence of discrimination within this sector of the labor market is most salient.

To offer insight into how gender-based discrimination influences hiring decisions in software engineering roles, this paper employs a correspondence study. This is a type of field experiment where a large number of fictitious job applications are sent with a single, randomly determined variation, and responses from firms are recorded to measure the effect of this variable of interest on the outcome. With only one variable imposed at random, the experiment effectively controls for other factors that can influence a job applicant's success, such as experience, self-presentation quality, and accolades. The downside of this is that the results are formally valid for the type of person described in the CV. Larger samples would allow varying other characteristics to improve the range of circumstances for which the results are valid.

This approach to studying hiring discrimination was most famously employed by Bertrand and Mullainathan (2004). Though correspondence studies had been carried out for years prior to Bertrand and Mullainathan's publication, their methodology is credited with inciting a slew of similar correspondence studies since (Guryan and Charles (2013), Bertrand and Duflo (2017)). By sending nearly 5,000 resumes to job listings in Chicago and Boston from 2001-2002, randomly varying applicant names to signal race, they recorded that white-sounding applicants had a 50% higher chance of being called-back to proceed with the interview process than their black-sounding counterparts (Bertrand and Mullainathan, 2004). This experimental methodology is arguably the modern standard for hiring discrimination research, although meta-literature on correspondence studies has advocated for more innovative development (Bertrand and Duflo (2017), Guryan and Charles (2013), Azmat and Petrongolo (2014)). Details of the specific implementation used in this paper are discussed in the Experimental Design section.

In general, economics research to detect and measure gender-based discrimination in labor markets has yielded highly variable results. For example, different studies on promotions have found discrimination against both men and women, as well as no detectable discrimination against either group (Blau and DeVaro, 2007). Studies focused on wage increases and hiring decisions have produced similarly inconsistent results. Analyses and experiments observing call-back rates[7] for male and female applicants since the 1970s have also detected statistically significant discrimination against both men and women, and within this literature, the observed relationships between gender discrimination and firm and industry characteristics have fluctuated greatly (Bertrand and Duflo (2017), Baert (2018)).

Economists have paid particular attention to how industries and firms with more or less female representation behave, as well as to the effect of gender compositions of leadership and human resources teams. For both of these characteristics, while there have been mixed results, there does seem to be accumulating evidence that, (a) panels with more women evaluate female applicants more harshly, (b), that women receive lower evaluations than men when genders are known versus when evaluations are gender-blind, and (c), that industries with a skewed gender composition exhibit some favoritism for the majority group, though this last trend is perhaps the least consistent (Azmat and Petrongolo (2014), Bertrand and Duflo (2017), Kübler, Schmid and Stüber (2018)). As many of the firms included in this paper's experiment belong to the technology industry, the influence of these factors is particularly salient; according to Pew Research, women accounted for 25% of the computer science workforce in 2018 (Graf, Fry and Funk, 2018).

This paper directly adds to research by Riach and Rich (2006), where the authors employed correspondence studies to investigate gender-based hiring discrimination within computer analyst programmer positions. Rich and Riach first implemented this in Australia in the mid-1980s and then later in England in 2005. While their initial experiment recorded substantial discrimination against female applicants at a 2% significance level, the subsequent trial in England found even greater discrimination against male applications at a 5% significance level (Riach and Rich, 2006). The geographical and temporal distance between these two iterations of the correspondence study make it difficult to infer much from their opposing results, but the opposition itself supports the relevance of continued investigation into gender discrimination in the computer programming field. In summary, as is the case in economic research broadly, the role of gender discrimination in the tech sector remains inconclusive.

II. Experimental Design

A. Resume

A single resume was created, with the applicant name and its associated email address being the only variables to signal gender. Each company included in the sample received one resume, which had been randomly assigned either a male or female gender identity. A benefit to this approach is the reduced disruption to employers, whose hiring staff spend some time assessing, as well as possibly

[7]The call-back rate refers to the proportion of applications that receive a positive response from a company, versus those that receive a negative or no response, to the total number of applications sent.

documenting and contacting, each applicant. A cost is that more job listings need to be identified to obtain a sufficiently large sample size, and specific variations across submitted resumes were not examined. However, variations in the job requirements, when given the same resume, still enable some investigation as to how a candidate's relative qualification interacts with the candidate's gender.

The resume was designed to be somewhat impressive, though without any stellar accolades or accomplishments (Appendix Figure A1). This responds to a concern that discrimination could be wiped out by overly impressive candidates against whom the cost of discriminating would be too high, while also seeking to obtain enough positive responses from employers to reach statistically significant conclusions. The school, Wesleyan University, is generally well regarded and ranked 40^{th} on Forbes's Top Colleges 2019 list (Forbes, n.d.), but is not particularly known for its computer science department, for which it ranks 167^{th} among other U.S. universities on Computer Science Schools (Computer-Science-Schools.com, n.d.). An average grade at Wesleyan is an A^- (WesleyanArgus.com, n.d.), which translates to a 3.7 GPA, while nationally, the average undergraduate grade is a B (Lindsay, 2020), which translates to a 3.15 GPA. Broadly, the U.S. average GPA for a computer science major is 3.13[8] (Lindsay, 2020). Considering national and school-specific trends, an overall GPA of 3.6 from Wesleyan University was chosen to capture that the candidate was a strong student, though without particularly stellar achievements.

While many correspondence studies focus on jobs for recent college graduates,[9] this experiment creates a fictitious candidate with approximately three years of full-time work experience. A benefit of this is that variability in job listings' minimum experience requirement can be exploited to evaluate gender performances for over and under-qualified applications. When populating the resume with employment and project experiences, the priorities included: believability, encompassing of diverse skills, and maintaining a similar level of impressiveness and modesty as desired from the candidate's educational background. Drafting experiences that included a breath of coding languages, full stack development work, a range of project types, and some data science was decided with the understanding that job listings for software developers tend to require specific skills, and thus in order to achieve a high volume of relevant applications and justify applying to positions with a breadth of requirements, the candidate needed to be versatile. A collection of ~10 real computer science resumes of young professionals were referenced to select companies and craft details regarding work experiences. Consultation with other young professionals working in software development helped refine the resume, to verify its credibility.

[8] Although there may be differences in the average GPA of male and female computer science students, no evidence on this that would be relevant to the fictitious applicant could be found. If such a difference were to be known, it would be important to consider how GPA signals above or below average performance within one's gender group. Factoring in this dimension could be enlightening for revelations on statistical vs. taste-based discrimination in future research.

[9] Bertrand and Duflo (2017) mention this as a criticism.

Several design options were considered that were not used in this experiment. For one, applicant gender could have additionally or alternatively been signaled through the inclusion of other details. For example, participation in a male or female sports league, involvement in organizations like "Girls Who Code," and projects focused on stereotypically masculine or feminine subjects, such as fantasy football or makeup, respectively, could have been included. However, these attributes might inform the "type" of man or woman the applicant was. While this would be highly interesting to examine along with the baseline approach of only varying names, an experiment doing so would need to be able to execute enough applications to reach a sample size conducive to this analysis.

B. Names

Names were chosen to be as similar as possible, while signaling gender clearly. Applications were sent under the names Stephen/Stephanie, Eric/Erica, and Daniel/Danielle.[10] Using the U.S. Social Security Administration's search engine to determine annual popularity of names given to newborns[11], Stephen and Stephanie both ranked in the top 100 for male and female names, respectively, in the early-to-mid 1990s, when the fictitious applicant would have been born. The same is true for Eric and Erica, as well as for Daniel and Danielle. In each case, checking for the popularity of the name for babies of the name's unconventional gender association (e.g. checking for female Erics and male Ericas) yields extremely low rankings and, more often, unavailable data for relevant years. This supports the strongly gendered associations of the chosen applicant names.

In general, these first names are not believed to be particularly associated with any racial group (Tzioumis, 2018). Last names are more correlated with race overall, and ones used in this study included: Nichols, Harris, and Green. Harris and Green, according to census data reviews, are highly represented among both white and black people in the United States (Word et al., 2008). Nichols, however, is most prevalent among white people (NameCensus.com, n.d.). Intersectionality among race and gender is an important research topic that would require a higher sample size than could be obtained in this experiment to properly study.

C. Application Materials Beyond the Resume

A brief cover letter (one paragraph) was submitted along with the resume to each job application.[12] The cover letter was created as a combination of two versions drafted separately by a young man and woman to moderate its tone. Nothing substantive was included in this message, such that it was a rather brief and generic, though personable, introduction and expression of interest.

[10]The names Eric/Erica and Daniel/Danielle were used when application limits were believed to have been reached for the Stephen/Stephanie job listing account in order to continue with the experiment.

[11]See `https://www.ssa.gov/cgi-bin/babyname.cgi`.

[12]In a small number of cases, a cover letter or a resume was not collected by the online application form, such that both of these materials could not be submitted for 100% of job listings.

Some online job application forms required answers to additional questions to be submitted. Answers given were recorded and duplicated on future applications with comparable questions. The content of answers was crafted to be relatively generic and to not provide any substantive information beyond that exhibited on the resume.

In order to apply to jobs through a particular platform, it was required to create an account, which included some brief questions similar to generic supplementary ones on applications. These profiles were made to be precisely identical for the male and female applicants. While adding a profile photo could help to support credibility that the applicant is a real person and signal gender, the decision was made to not do this to avoid complicating variables such as attractiveness, overt racial identity, and clues to the "type" of man or woman the fictitious applicant might be. Exploiting the profile image feature can be used in future research to examine these variations and signal gender through means other than names.

Also in line with the decision to avoid signaling gender through any channel other than applicant name, this experiment did not create online profiles on professional or personal websites for the fictitious applicants. While this could harm believability, it maintained the minimalist gender identification of the resumes and avoided riskier logistics such as fabricating connections to other people on these platforms and publicly expressing fabricated affiliations with real firms.

Email accounts were created for each name, and a single phone number was used for female applications, while another was used for males. Given the norm of hiring process communication happening over email, as well as how email facilitates recording and classifying responses, phone numbers were only included when required and were excluded from resumes.

As can be seen from the overall response rates (see Appendix Table B1), these simplifications did not seem to significantly hinder the credibility of the applications.

D. Job Selections

Positions to apply to were found on two popular online job search platforms. Close to 50% of applications were sent on AngelList, while the others were sent on Indeed. AngelList describes itself as "the world's largest startup community,"[13] such that companies listing jobs on this platform can be considered self-identified startups.

Search criteria on AngelList specified full-time software engineering jobs, ranging from 0-6 years of required experience (plus jobs for which no experience minimum was listed), at U.S.-located companies with 1-200 employees that had been active on the platform within the last 30 days. The experience range outside of the fictitious applicant's approximate three years enabled increased sample size, as well as the possibility of detecting how gender discrimination might vary among

[13]See `https://angel.co/`.

over and under-qualified candidates. The restriction to companies with fewer than 200 employees was imposed to facilitate analysis of how firms' leadership gender representation might affect gender hiring trends, following the assumption that leadership in smaller "startups" would have more influence on company culture.

On Indeed, not all of these criteria were able to be imposed using the site's search engine. Full-time positions requiring a bachelor's degree at US companies were included. Of these, only job postings that did not require an account to apply to were selected to reduce the risk of the experiment being discovered.[14]

Within these search results, discretion was used to identify any results that seemed to be outside of the software development scope that were included. Other than that, so long as a company had not been previously applied to, all openings encountered were included in the experiment.

E. Applicant Gender Randomization

Gender was assigned to a listing only after the job information had been recorded to prevent any unconscious skew of data selection. After a collection of jobs had been selected and their details had been documented, a set of dummy integers (zero, for male; one, for female) was generated,[15] corresponding to the jobs on the spreadsheet in the same row order.

F. Application Procedure

Experimenters submitted applications with the assigned gender to jobs through the job listing websites. A resume and cover letter message were always included, and additional questions were answered in a consistent manner only when required. Following this initial submission, no further action was taken to bring the fictitious candidate to a firm's attention.[16]

G. Response Categorization

Categorizing application responses as positive or negative was not straightforward. While some companies called or emailed to explicitly express interest in moving forward with the interview process, many responses were vague. For example, it was sometimes unclear if a link to a coding assessment was automatically sent to everyone who applied, or if this was itself an indication of having passed the company's initial screening. Furthermore, companies fairly often replied with follow-up or clarifying questions, without expressing interest. These questions

[14] Companies can view profiles, which could lead someone to notice the near-identical nature of the male and female applicants. On AngelList, an account was required to access and apply to job listings.

[15] Randomized integers were sourced from `https://www.random.org/integers/`.

[16] Some companies sent an email asking for the application to be submitted on another website. These requests were ignored because following-up to do so would not have been scalable for the research team. Only once in the first round of applications was there follow-up to a request from a company to email the resume to a specific person, which then resulted in a positive response; this observation is excluded for robustness.

ranged from asking about citizenship status, to location, to motivation. It is possible that an employee responsible for screening decisions asked these with the implicit intention to pursue the candidate in light of a favorable answer, but it is also possible that screeners could have noticed missing key information that was not required in the application submission, which they tried to obtain before making any kind of judgement.

Rejections tended to be less ambiguous. Still, it was not obvious how to classify a response saying that the position applied to had been filled prior to the sending of the application, as well as a case where the candidate passed a preliminary screening before being rejected in a subsequent one. Responses where the employer said that the candidate was not a good fit for the role applied to but expressed interest in the candidate for an alternative position at the firm also challenged the binary classification of "yes" or "no." In any case, the nature of a positive or negative response was recorded and reassessed at the end of the experiment to ensure that these uncertainties were handled consistently.

Responses were ultimately sorted into the following groups: no response ("No Response"), clear rejection ("No"), questionable rejection ("No?"), clear yes ("Yes"), and questionable yes ("Yes?). Combining the clear and questionable groups yields the variables referred to as "No+" and "Yes+" in the analysis.

To reduce the cost of the experiment to employers, favorable responses and those asking for further information were emailed a generic response in a timely manner, indicating that the candidate was no longer looking for a new job. No response was sent to companies that turned down the candidate.

III. Data Overview

The full dataset encompasses 1,865 applications to software engineering jobs at unique firms in the United States. Each observation includes details about the position, the company, and the outcome. Information about the position, such as whether or not it can be performed remotely, its wage,[17] and its experience requirements, comes from the job listing itself. Company data, including year founded, number of employees, and founder gender, was derived from a variety of public websites. Crunchbase, LinkedIn, and the companies' own websites provided most of this information. Although there may be error within these statistics, this is independent of gender treatment, and results reported use heteroskedastic-robust standard errors.

For many observations, not all firm-level data could be found. For example, data on founders was identified for just over two thirds of firms, and the gender of the CEO[18] was collected for approximately 45% of the sample. Combining these

[17]While some job listings included a wage, most had a projected wage range provided by the job listing website, likely derived endogenously from the job's location and experience requirement, using external information on regional wage trends. This variable is not used in the analysis since it is believed to largely be imputed.

[18]For companies where a CEO position did not exist, the President's gender was recorded.

metrics into a "female leader" dummy variable, indicating if a founder or the CEO of a company is female, includes 74% of observations in the overall sample. This composite variable improves the data set's ability to comment on whether or not female leadership affects a firm's treatment of gender in hiring practices.

The "female leader" indicator, and other variables with missing values are handled by, for each variable, creating a new variable that is equal to one when a value is missing (and zero otherwise), and setting missing values in the original column to zero. One benefit of this approach is that more-granular analyses can be performed on the full data set, increasing statistical power. Additionally, contrary to dropping observations with missing values, this approach does not impose the assumption that whether or not certain data was available is independent of outcome.

Comparing summary statistics of the female (treated) and male (untreated) groups supports that the randomization and sample size resulted in similar sets, where the jobs and firms included have similar overall attributes (see Tables 1 and 2). This supports the credibility of inferring that the difference in outcomes is attributable to the independent variable of gender and not to a happenstance skew of either male or female applications being sent to certain types of job openings. For example, it would have been concerning if female-designated jobs had a mean experience requirement of five years, while the male-designated job pool averaged an experience requirement of three years. No such disparities seem to exist.

Table 1—Data Characteristics for Female Applicant

	Mean (SD)	Median
Year Founded	2006 (19)	2013
Firm Size	398 (1,338)	30
Miles from Albany	1,525.36 (1,152.40)	1,299.50
Experience Required	3.37 (1.69)	3.00
Days to Respond	10.95 (13.77)	6.00

Regarding job characteristics, 90% of job listings provided information on remote work. Of these, 40% are categorized as "Yes," and 40% are categorized as "No."[19] Having remote work data for a large portion of the sample, and within that, having relatively large proportions of both "Yes" and "No" is promising for detecting how the remote variable might affect gender-based hiring trends,[20] if it

[19]Twelve percent of jobs are listed as "temporarily remote." This status referred to having employees telework during the COVID-19 pandemic, with the explicit expectation of in-person work in the future.

[20]It is possible that reduced personal contact could attenuate taste-based discrimination.

TABLE 2—DATA CHARACTERISTICS FOR MALE APPLICANT

	Mean (SD)	Median
Year Founded	2006 (22)	2012
Firm Size	296 (1,101)	30
Miles from Albany	1,500.42 (1,137.78)	1,247.50
Experience Required	3.43 (1.93)	3.00
Days to Respond	9.62 (12.14)	5.00

does so at all.

Geographical data exists for 96% of the dataset and includes the distance between the job applied to[21] and the fictitious applicant's most recent work location (Albany, NY).[22] Population data for the job location is also included as a proxy for how urban, suburban, or rural the city is. An online tool[23] was used to compute the population within a 30-mile radius of the precise city. As opposed to a measurement of only the city's population, this approach captures suburban cities' connections to their nearby urban hubs and is not dependent on the size of what could be an arbitrarily large or small city area-wise. This approach thus better describes the urban-ness of the job's location.

Industry data given on job listing sites and inferred by company descriptions illustrates the breadth of firms applied to (Figure 1). These groups are not mutually exclusive; for example, some firms specialize in "IT Consulting," or sell a SaaS[24] product used for education. While the prevalence of IT, consulting (software development-related), security (computer-related), and SaaS are to be expected for software engineering positions, the high number of observations in healthcare, finance, and commerce are noteworthy, since these industries are not inherently linked to software engineering. Furthermore, female workers tend to be overrepresented in healthcare (Berlin et al., 2019), while male workers are overrepresented in finance (Chin, Krivkovich and Nadeau, 2018), such that comparing results within these groups could be enlightening to the question of how gender-based hiring discrimination is affected by industry, beyond the experiment's concentration on software development roles.

Categorizing job titles using key word searches gives a snapshot of the breadth of software engineering roles applied to (Figure 2). As is the case for industries,

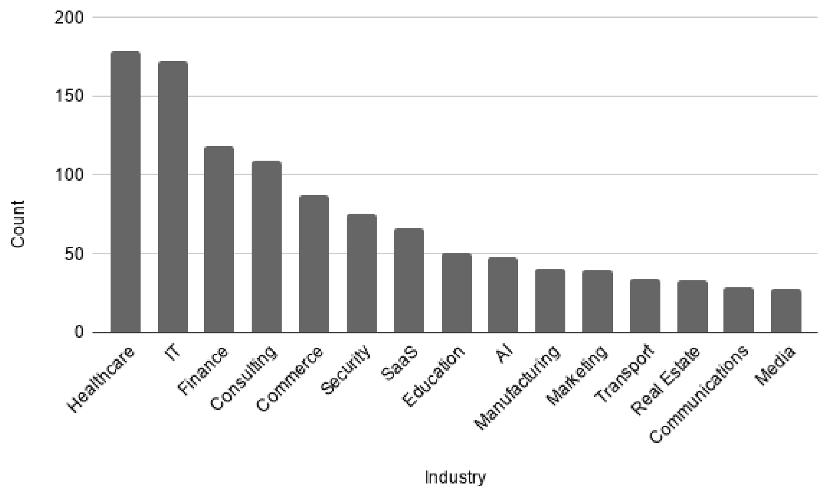
[21]When multiple cities were listed as the location in a job description, the nearest city was used for the analysis.

[22]This distance is specifically reported as the number of miles to drive between the two locations when taking the fastest route.

[23]See `http://www.hoosierdata.in.gov/big_radius/radius.asp` (2019 pop data).

[24]Software as a Service.

FIGURE 1. TOP 15 INDUSTRIES IN THE SAMPLE



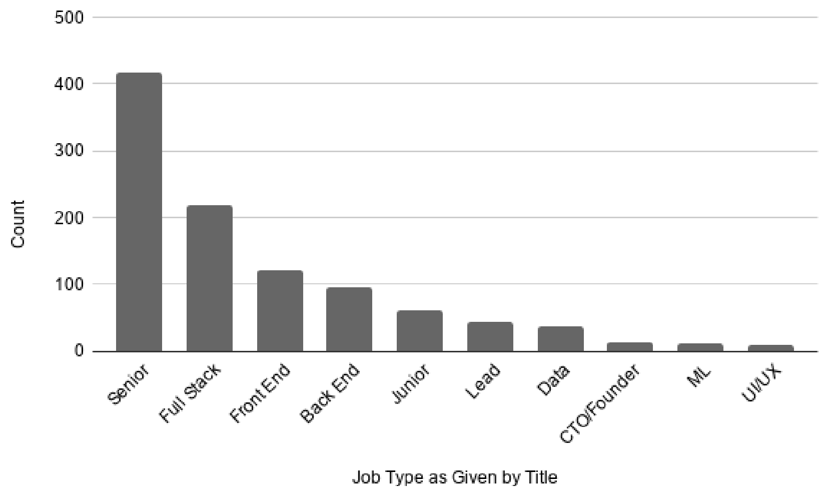
job categories are not mutually exclusive. For example, "Senior Full Stack Developers," and "Lead Data Engineers" would each appear in two groups (Senior/Full Stack, Lead/Data), corresponding to key words in their title. The high volume of "full stack," "front end," and "back end" titles corresponds favorably to the strengths conveyed on the resume. Smaller counts of "Data," "Machine Learning (ML)," and "User Interface/Experience (UI/UX)" positions are appropriate given that the fictitious candidate has limited experience in these areas. Larger amounts of jobs with "senior" in their title than "lead" or "CTO/Founder" is also well-aligned with the experience level of the application. These characteristics support the relevance of the job sample to the fictitious application, and thus support the validity of the job selection process.

IV. Analysis

The data reveal positive discrimination for the female applicant. While male applicants have an average call-back rate of 11.2%, female applicants have an average call-back rate of 15.8% (see Table 3, column 1). This difference implies that the female applications had a 41% greater chance of leading to the next stage of the interview process than the male applications and is statistically significant at the 1% level. Adding controls for firm and job characteristics that might impact call-back rates does not appreciably diminish the size or significance of the benefit to women; the coefficient on an applicant being female remains statistically significant at the 1% level (see Table 3, column 3).

Replicating these regressions on the "Yes+" outcome variable, which includes not only clear positive responses, but also questionably positive ones, female appli-

Figure 2. Types of Jobs in the Sample



cants are found to have a smaller advantage over male ones (see Table 3, columns 2 and 4). Nonetheless, the null hypothesis of no gender-based discrimination can still be rejected at the 5% level when less reliable outcome metric, "Yes+," is the dependent variable.

Outside of gender, the main regression results highlight how applications sent on Indeed performed better than those sent on AngelList. As would be expected, the further from the required years of experience an application was, the less likely it was to receive a favorable response.[25]

Adding interactions to the model allows for some assessment of how gender discrimination might vary across different types of jobs and firms. Interactions included in the specification are those between a female gender assignment and (a) experience mismatch, (b) firm age and size, (d) the presence of a female founder or CEO,[26] (e) whether or not a job is considered a "Senior" position, (f) a firm's number of male and female founders, (g) whether a company was within the healthcare of finance industry, and (h) whether or not the job could be performed remotely.

An applicant's experience mismatch is included to test the hypothesis that firms value (or dislike) women to such an extent that they will hire an applicant who is not a good fit for the job.[27] Firm age and size align with the hypothesis

[25] Testing the effect of an applicant being overqualified or under-qualified, the variables for which were calculated by subtracting the experience requirement from three or three from the experience requirement and zeroing negative differences, produced similar results. Poor experience level fit in either direction appears to harm applicants, though only the detriment to being overqualified produces a significant coefficient on its own.

[26] President, if a firm does not have a CEO.

[27] To get more specific information on this, regressions were performed using over and under qualified

Table 3—Effect of Gender on Positive Firm Response

	Dependent variable:			
	Yes	Yes+	Yes	Yes+
	OLS	*OLS*	*OLS*	*OLS*
	(1)	(2)	(3)	(4)
Female	0.046	0.042	0.043	0.039
(applicant gender)	(0.016)***	(0.018)**	(0.016)***	(0.018)**
Experience Off			−0.015	−0.018
abs(exp. required - 3)			(0.006)***	(0.007)***
Platform			0.054	0.066
(Indeed)			(0.022)**	(0.025)***
Firm Age			−0.0004	−0.0004
(years since founding)			(0.0005)	(0.001)
Firm Size			0.00000	0.00000
(employee count)			(0.00001)	(0.00001)
Distance			0.00001	−0.00000
(job location to Albany, NY)			(0.00001)	(0.00001)
Population			0.000	0.000
(30 mi. radius of job location)			(0.000)**	(0.000)*
Constant	0.112	0.166	0.073	0.145
	(0.010)***	(0.012)***	(0.028)***	(0.032)***
Observations	1,865	1,865	1,865	1,865
Adjusted R^2	0.004	0.002	0.012	0.010

Note: *p<0.1; **p<0.05; ***p<0.01

Standard errors reported in parentheses are hetero-skedastic robust.

"Yes" refers to clear positive firm responses; "Yes+" includes less certain ones.

Dummy variables for missing values are included, but not displayed.

that younger, smaller firms, such as startups, would be more likely to be either pro-diversity or masculine-dominant, given media attention on startup norms. Female leadership as interacted with application gender tests if firms with women in charge might act favorably toward female applicants. An indicator variable for jobs with "Senior" in the title sees if positions internally identified by companies to be of higher status than non-senior positions might exhibit discrimination against women; though expected to be correlated with experience, the "Senior" title reveals the relative rank of a position within a firm. Similarly to the indicator for female leadership, the specific number of male or female founders can highlight how more or less male or female leadership affects discrimination when interacted with applicant gender. Healthcare and finance industries are singled-out due to the aforementioned skewed gender representations within these sectors of the labor market; testing their interaction is in line with research into how gender-based discrimination might vary due to industry-wide female prevalence. Lastly, the ability of a job to be performed remotely tests the hypothesis that taste-based discrimination might be weaker if there is less interpersonal interaction.

The interaction model that takes "Yes" as the dependent variable does not find any of these interactions to be significant, and the significance on the gender coefficient itself is reduced to a 10% level given the dilution of the interaction coefficients (see Table 4, column 1). However, the "Yes+" specification produces more statistically significant findings (see Table 4, column 2). Without correcting for multiple hypothesis testing, a negative interaction between gender and firm age is detected at the 1% level, and a detriment to female-assigned applications to firms in the financial sector is found, but only at a 10% significance level. Adjusting p-values through a Romano-Wolf multiple hypothesis correction (Romano and Wolf, 2005) fails to reject the null of finance and female having no interaction, but maintains the interaction between firm age and applicant gender to be significant at a level of 5%.

This suggests that younger firms have greater discrimination in favor of women. The coefficient can be interpreted as saying that a one-year increase in firm age, holding all else equal, reduces the benefit to females by 0.003 (a 4% loss from the 0.081 coefficient on the female indicator variable). This implies that a ten-year increase in firm age reduces the female benefit by 0.03 (a 37% loss). Furthermore, a 27-year increase in firm age would completely cancel-out the coefficient on the female indicator variable, and increases beyond this threshold are associated with bias *against* females, other gender interactions aside.

However, it is important to keep in mind that this interaction is not identified as significant when the more reliable "Yes" outcome categorization is the dependent variable. This caveat encourages further research into how firms' age affects how they respond to gender in hiring decisions.

metrics in place of the experience mismatch variable. Interactions between over and under qualified measures and gender were not found to be statistically significant.

Table 4—Model with Interactions

	Dependent variable:	
	Yes	Yes+
	OLS	*OLS*
	(1)	(2)
Female	0.051	0.081
(application gender)	(0.030)*	(0.034)**
Female:Exp. Mismatch	−0.011	−0.009
abs(experience required - 3)	(0.011)	(0.013)
Female:Firm Age	−0.001	−0.003
(years since founding)	(0.001)	(0.001)***
Female:Firm Size	−0.00002	0.00001
(employee count)	(0.00001)	(0.00002)
Female:Fem. Leader	0.007	0.058
(female CEO/Founder)	(0.057)	(0.062)
Female:Senior	0.005	0.027
(job title incl. 'Senior')	(0.029)	(0.033)
Female:#M. Founder	0.019	0.006
(male founder count)	(0.013)	(0.014)
Female:#F. Founder	−0.012	−0.029
(female founder count)	(0.055)	(0.057)
Female:Healthcare	0.048	0.025
(industry)	(0.044)	(0.049)
Female:Finance	−0.035	−0.087
(industry)	(0.045)	(0.046)*
Female:Remote	−0.023	−0.034
(remote work allowed)	(0.027)	(0.030)
Constant	0.110	0.189
	(0.031)***	(0.036)***
Female coefficient without interactions	0.043	0.039
	(0.016)***	(0.018)**
Observations	1,865	1,865
Adjusted R^2	0.031	0.036

Note: *p<0.1; **p<0.05; ***p<0.01

Standard errors reported in parentheses are hetero-skedastic robust.
"Yes" refers to clear positive firm responses; "Yes+" includes less certain ones.
Controls not displayed: experience mismatch, platform, firm age, firm size, distance, population, dummies for missing values.

V. External Validity

The study has limited validity beyond detecting statistically significant discrimination in favor of female software engineers among applicants with similar credentials to those of the fictitious one. While correspondence studies that can be implemented at a statistically large enough scale benefit from including variations in resumes, this field experiment is limited by its use of a single resume. Traits such as the applicant's experience level, degree of modesty/prestige, and presentation style are controlled across all results. The extent to which pro-female hiring discrimination exists for different types of applicants within software development roles remains unclear.

Variations in the characteristics of jobs applied to nonetheless provide some heterogeneity across observations that could enable a richer understanding the detected discrimination. However, data analysis shows that for most of these variations, the null hypotheses of no interaction between gender discrimination and the variable cannot be rejected. It is unknown to what extent this is precluded by the sample size, versus by the size of the effects of the interactions, if any such effects exist.

The one interaction found to be significant at a 5% level following a Romano-Wolf multiple hypothesis correction is that which indicates that females are less advantaged at older firms, and that older firms might even be biased against female applicants. However, this result stems from the regression specification taking "Yes+" to be the dependent variable, which includes less certain firm responses as positive results. Regardless of this uncertainty, it remains unclear whether or not this effect is limited to software engineering applicants, or if older firms exhibit a reduced preference for female applicants broadly. This result might be generalizable to support firm age and gender discrimination trends when combined with other experiments that have detected a similar effect.

Another important facet of this experiment to consider is its implementation during the COVID-19 pandemic. Applications were sent from mid-June through October 2020, with $\sim$99% of applications sent in August, September, and October. The total volume of software engineering job listings located in the U.S. on one of the websites used was recorded periodically before and during the pandemic. Unsurprisingly, in line with negative macroeconomic effects of the pandemic, this figure decreased significantly from 2,684 on March 4^{th}, to 1,240 on May 25^{th}, before rebounding some to 2,061 on June 22^{nd}, when applications were first sent.

In all, it is not believed to be likely that the pandemic setting caused an inflated level of gender-based discrimination that would not have been detected in normal economic conditions. Following theory that non-statistical discrimination occurs despite productivity losses, such that discriminating firms are willing to pay a price for their animus, an unfavorable economy should attenuate this behavior, if anything; facing heightened concern for business failure, the cost of taste-based discrimination would be relatively greater. The female bias thus would be expected to remain at least as strong outside of the pandemic. Furthermore, the absence of noteworthy gender-specific movements during this period supports the validity of this study beyond its time of implementation. The same assumption, however, would likely be weaker for correspondence studies on racial discrimination during this timeline, considering the heightened attention on the Black Lives Matter movement (Milligan, 2020).

VI. Conclusion

The results of the field experiment support the hypothesis that female software engineers experience positive discrimination at the screening phase of hiring processes. This is at least found to be true for applicants with moderately impressive backgrounds and a few years of full-time work experience, and this result could be generalized through follow up research. Female applicants were specifically found to have a 41% higher chance of receiving a favorable response from firms than their male counterparts, and the null of there being no gender discrimination is rejected at a 1% significance level. A robustness check categorizing firm responses that were less clearly positive as favorable continued to detect pro-female gender discrimination at the 5% level.

Although most interactions between firm and job characteristics were not found to be significant, a diminished benefit to females when applying to older firms was detected at a 1% level when the dependent variable includes less clearly positive responses as favorable. After correcting p-values for multiple hypothesis testing, the null hypothesis of there being no gender-firm age interaction is still rejected at the 5% level. This finding offers support to the hypothesis that younger firms are making greater efforts to increase their representation of female software developers than older firms are.

While sample size limitations precluded the experiment from including additional applicant variations beyond gender, the experiment succeeded in its goal of detecting gender-based discrimination that might have been present in the market for software engineers. The finding of pro-female bias is consistent with Riach and Rich (2006), which similarly found female computer programmers to benefit in the English labor market. Recalling that Riach and Rich (1987) found male computer programmers to be favored in Australia, the findings of this paper point to a potentially ongoing trend in favor of female coders.

Despite potential forces that could benefit male software engineers, such as people being biased toward those more similar to them, "cultural fit" being a substantial factor in hiring decisions, and women performing relatively worse in evaluations that are not gender-blind, discrimination against female applications was not found. Moreover, the observed high level of discrimination in favor of women suggests that "pro-diversity" efforts to increase the representation of women in software development are significant factors in firms' hiring decisions. The extent to which gender diversity might be desired due to benefits from group productivity dynamics, improving companies' public image, or even statistical or taste-based discrimination against men, is unclear. Follow-up field experiments that include additional applicant variations and interdisciplinary research into the mechanisms through which gender diversity benefits tech firms could help explain the significant female advantage detected in this paper. For now the main takeaway is that girls who code have a reason to be optimistic in the labor market.

REFERENCES

Altonji, Joseph G, and Rebecca M Blank. 1999. "Race and gender in the labor market." *Handbook of labor economics*, 3: 3143–3259.

Areitio, Andy. 2018. "What is a startup and how is it different from other companies (new and old)?" *https://medium.com/theventurecity/what-is-a-startup-and-how-is-it-different-from-other-companies-new-and-old-428875c27c29*.

Azmat, Ghazala, and Barbara Petrongolo. 2014. "Gender and the labor market: What have we learned from field and lab experiments?" *Labour Economics*, 30: 32–40.

Baert, Stijn. 2018. "Hiring discrimination: an overview of (almost) all correspondence experiments since 2005." In *Audit studies: Behind the scenes with theory, method, and nuance*. 63–77. Springer.

Becker, Gary S. 1957. *The Economics of Discrimination.* Economics Research Studies.

Berlin, Gretchen, Lucia Darino, Megan Greenfield, and Irina Starikova. 2019. "Women in the healthcare industry." *https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/women-in-the-healthcare-industry*.

Bertrand, Marianne, and Esther Duflo. 2017. "Field experiments on discrimination." In *Handbook of economic field experiments*. Vol. 1, 309–393. Elsevier.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American economic review*, 94(4): 991–1013.

Blau, Francine D, and Jed DeVaro. 2007. "New evidence on gender differences in promotion rates: An empirical analysis of a sample of new hires." *Industrial Relations: A Journal of Economy and Society*, 46(3): 511–550.

Bradford, Laurence. 2020. "26 Organizations Teaching Women Coders Girls Around The Globe." *https://learntocodewith.me/posts/13-places-women-learn-code/*.

Chin, Stacey, Alexis Krivkovich, and Marie-Claude Nadeau. 2018. "Closing the gap: Leadership perspectives on promoting women in financial services." *https://www.mckinsey.com/industries/financial-services/our-insights/closing-the-gap-leadership-perspectives-on-promoting-women-in-financial-services*.

Computer-Science-Schools.com. n.d.. "Computer Ranking 2018: Wesleyan University." *https://computer-science-schools.com/wesleyan-university#:~:text=Based%20on%2067%20evaluation%20criteria,Best%20Computer%20School%20in%20Connecticut*, Accessed November 2020.

EEOC. n.d.. "Discrimination by Type." *https://www.eeoc.gov/discrimination-type*, Accessed November 2020.

Fontinelle, Amy. 2020. "Startup." *https://www.investopedia.com/ask/answers/12/what-is-a-startup.asp*, Updated Aug 14, 2020.

Forbes. n.d.. "40 Wesleyan University." *https://www.forbes.com/colleges/wesleyan-university/?sh=6e6073fa3b0e*, Accessed November 2020.

Frankenfield, Jake. 2019. "Technology Sector." *https://www.investopedia.com/terms/t/technology_sector.asp*, Updated Jul 13, 2019.

Graf, Nikki, Richard Fry, and Cary Funk. 2018. "7 facts about the STEM workforce." *https://www.pewresearch.org/fact-tank/2018/01/09/7-facts-about-the-stem-workforce/*.

Guryan, Jonathan, and Kerwin Kofi Charles. 2013. "Taste-based or statistical discrimination: the economics of discrimination returns to its roots." *The Economic Journal*, 123(572): F417–F432.

Huet, Ellen. 2017. "Some Tech Companies Are Trying Affirmative Action Hiring—But Don't Call It That." *https://www.bloomberg.com/news/articles/2017-01-10/some-tech-companies-are-trying-affirmative-action-hiring-but-don-t-call-it-that*, Accessed November 2020.

Kübler, Dorothea, Julia Schmid, and Robert Stüber. 2018. "Gender discrimination in hiring across occupations: a nationally-representative vignette study." *Labour Economics*, 55: 215–229.

Landau, Candice. n.d.. "What's the Difference Between a Small Business Venture and a Startup?" *https://articles.bplans.com/whats-difference-small-business-venture-startup/*, Accessed November 2020.

Lindsay, Samantha. 2020. "What's the Average College GPA? By Major?" *https://blog.prepscholar.com/average-college-gpa-by-major*.

Milligan, Susan. 2020. "Pandemic, Recession, Unrest: 2020 and the Confluence of Crises." *https://www.usnews.com/news/national-news/articles/2020-06-02/pandemic-recession-unrest-2020-and-the-confluence-of-crises*.

Mundy, Liza. 2017. "Why Is Silicon Valley So Awful to Women?" *https://www.theatlantic.com/magazine/archive/2017/04/why-is-silicon-valley-so-awful-to-women/517788/*, April 2017 Issue.

NameCensus.com. n.d.. "Most common last names for Whites in the U.S." *https://namecensus.com/data/white.html*, Accessed November 2220.

Riach, Peter A, and Judith Rich. 1987. "Testing for sexual discrimination in the labour market." *Australian Economic Papers*, 26(49): 165–178.

Riach, Peter A, and Judith Rich. 2006. "An experimental investigation of sexual discrimination in hiring in the English labor market." *The BE Journal of Economic Analysis & Policy*, 6(2).

Rivera, Lauren A. 2012. "Hiring as cultural matching: The case of elite professional service firms." *American sociological review*, 77(6): 999–1022.

Robehmed, Natalie. 2013. "What Is A Startup?" *https://www.forbes.com/sites/natalierobehmed/2013/12/16/what-is-a-startup/*.

Romano, Joseph P., and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." */http://www.jstor.org/stable/3598821*.

Szalai, Jennifer. 2018. "In 'Brotopia,' Silicon Valley Disrupts Everything but the Boys' Club." *https://www.nytimes.com/2018/02/07/books/review-brotopia-silicon-valley-emily-chang.html*.

TechPoint. 2017. "Women's groups INSIDE tech companies provide a support system." *https://techpoint.org/2017/07/womens-groups-inside-tech-companies-provide-support-system/*.

Tzioumis, Konstantinos. 2018. *https://www.researchgate.net/figure/First-Names--Distribution-of-proportions-across-race-ethnicity-categories-These-four_fig1_323595981*.

WesleyanArgus.com. n.d.. "The average grade given at the University is an A-minus. So what's the big deal with grade inflation?" *http://wesleyanargus.com/2015/03/05/grade-inflation/*.

White, Sarah K. 2019. "16 organizations for women in tech." *https://www.cio.com/article/3452864/16-organizations-for-women-in-tech.html*.

Wilhelm, Alex. 2018. "The Definition Of A Startup In 2018 (By The Numbers)." *https://news.crunchbase.com/news/the-definition-of-a-startup/*.

Word, David L, Charles D Coleman, Robert Nunziata, and Robert Kominski. 2008. "Demographic aspects of surnames from census 2000." *Unpublished manuscript, Retrieved from http://citeseerx. ist. psu. edu/viewdoc/download*.

Data and Procedure Details

Table A1—Details of Key Variables

Variable	Description
Firm Age	Years since firm's founding
Firm Size	Company's number of employees at time of application, given as a range and simplified as the mid-point of that range
#M/F Founders	Number of male or female founders of the company
Female Leader	Equal to one if company has a female CEO/President, or at least one female founder
Industry	Categorical description of type of business; dummy variables for each industry (e.g. commerce, finance, etc.)
Job Title	Name of position applied to, as given in the job listing; dummy variables for types of job created through key word search on job titles
Experience Mismatch	Minimum years of experience said to be needed for the position minus the applicant's years of full-time experience (three years)
Remote Work Status	Factor of 'no','yes','temporarily', or 'partially', as described in the listing
Distance	Miles from applicant's most recent employment location (Albany, NY) to job location, given for fastest driving route
Application Platform	Factor indicating which job listing website the application was submitted through (AngelList of Indeed)
Yes	Equal to one if received clearly positive response from firm
Yes+	Includes "Yes" and additional likely positive responses from firms

Table A2—Full Sample Firm Characteristics

	Mean or Fraction (Standard Dev.)	Range
Year Founded	2006 (21)	1789 - 2020
Size	346 (1224)	5.5 - 10001
1+ Female Founders	0.10	
1+ Male Founders	0.68	
# Female Founders	0.17 (0.43)	0 - 3
# Male Founders	1.68 (1.00)	0 - 6
Female CEO	0.05	
Male CEO	0.39	

Table A3—Full Sample Job Characteristics

	Mean or Fraction (Standard Dev.)	Range
Experience Required	3.40 (1.81)	0 - 15
Miles from Albany	1512.70 (1144.74)	0 - 4384
Yes to Remote Work	0.37	
Temporarily Remote	0.12	
Not Remote	0.41	
Senior Title	0.22	

Figure A1. Male-Assigned Resume Sent to Firms

Stephen Nichols

nichols.stephenw@gmail.com

Education: Wesleyan University, B.A. in Computer Science, 2017 G.P.A. 3.6

Languages:	C++, C#, Java, Python, Ruby, JavaScript, TypeScript, Go
Font-End:	React, Unity, Node, Swift, PHP, HTML/CSS
Cloud Computing:	Google Cloud Platform, AWS
Infrastructure/Scalability:	Docker, Kubernetes, MongoDB, MySQL, jQuery

EXPERIENCE

Blue Slate Solutions Albany, NY
Software Engineer 2018-Present

- Built a platform to analyze 50K+ Skype call recordings daily using Spark
- Led a 5-person team to create an interactive KPI platform to build & share dashboards
- Developed a cloud app to review performance of IT services daily
- Trained new hires in agile methodology, documentation protocols, & Git procedures
- Built RESTful API endpoints and tests, using Java and SAP HANA SQL database

Astro Technology (acquired by Slack) Palo Alto, CA
Software Development Engineer 2017-2018

- Integrated customer metrics data flow and improved data visualization protocols to give product and marketing teams weekly consumer experience summaries
- Started "Bug Busters," led bi-weekly code review sessions, fixing 100s of bugs
- Deployed TensorFlow jobs to convert Kafka streams of email & voice messages to dynamic calendar events

eBay San Jose, CA
Software Engineer Intern June-August, 2016

- Created a scoring system for advertising unit to evaluate advertiser goals
- Collaborated with data scientists to improve insight reports for 1M+ clients

PROJECTS

Accident Alert | Deep Learning | 4-Person Team 2016

- Engineered product to detect and describe traffic accidents from CCTV video feed, using motion estimation, object detection (LSTM, CNN)
- Designed user interface to notify nearest police with 85% accident classification accuracy

Psychology Research | App Development 2015

- Designed desktop and smartphone app-based strategy game used in psychology research
- Developed front & back end, in consultation with PhD students
- Automated recording, organization, and analysis of gameplay data for 500+ subjects

Languages : English (native), French (fluent), Spanish (proficient)

Results

Table B1—Overall Response Rates and Times

Outcome	Percent	Mean Days to Respond (Standard Deviation)
Yes	0.14	9.07 (11.91)
Yes+	0.19	8.55 (12.24)
No	0.18	12.02 (13.57)
No+	0.18	12.04 (13.56)
No Response	0.64	
No+ or No Response	0.81	

Table B2—Comparison of Response Rates by Gender

	Yes	Yes+	No	No+	No Reply	No+ or No Reply
Female	0.16	0.21	0.16	0.16	0.63	0.79
Male	0.11	0.17	0.19	0.19	0.64	0.83
Difference	0.05	0.04	-0.03	-0.03	-0.01	-0.04
T-Stat	2.92	2.33	1.82	1.81	0.45	2.33
P-Value	0.00	0.02	0.07	0.07	0.65	0.02